ПРОГРАММИРОВАНИЕ
введение в профессию

А. В. СТОЛЯРОВ

2

СИСТЕМЫ И СЕТИ

Любое использование данного файла означает ваше согласие с условиями лицензии (см. ниже). Электронные версии этой и других книг автора бесплатно доступны на сайте http://www.stolyarov.info

© Андрей Викторович Столяров, 2026

ПУБЛИЧНАЯ ЛИЦЕНЗИЯ

Учебное пособие Андрея Викторовича Столярова «Программирование: введение в профессию» в трёх томах, ранее публиковавшееся в издательстве МАКС Пресс в 2016–2021 гг., а также в издательстве ДМК Пресс в 2021 г., называемое далее «Произведением», защищено действующим российским и международным авторско-правовым законодательством. Все права на Произведение, предусмотренные законом, как имущественные, так и неимущественные, принадлежат его автору.

Настоящая Лицензия устанавливает способы использования электронной версии Произведения, право на которые предоставлено автором и правообладателем неограниченному кругу лиц, при условии безоговорочного принятия этими лицами всех условий данной Лицензии. Любое использование Произведения, не соответствующее условиям данной Лицензии, а равно и использование Произведения лицами, не согласными с условиями Лицензии, возможно только при наличии письменного разрешения автора и правообладателя, а при отсутствии такого разрешения является противозаконным и преследуется в рамках гражданского, административного и уголовного права.

Автор и правообладатель настоящим **разрешает** следующие виды использования данного файла, являющегося электронным представлением Произведения, без уведомления правообладателя и без выплаты авторского вознаграждения:

1) воспроизведение Произведения (полностью или частично) на бумаге путём распечатки с помощью принтера в одном экземпляре для удовлетворения личных бытовых или учебных потребностей, без права передачи воспроизведённого экземпляра другим лицам;

2) копирование и распространение данного файла в электронном виде, в том числе путём записи на физические носители и путём передачи по компьютерным сетям, с соблюдением следующих условий: (1) **все воспроизведённые и передаваемые любым лицам экземпляры файла являются точными копиями оригинального файла** в формате PDF, при копировании не производится никаких изъятий, сокращений, дополнений, искажений и любых других изменений, включая изменение формата представления файла; (2) **распространение и передача копий другим лицам производится исключительно *бесплатно*, то есть при передаче не взимается никакое вознаграждение ни в какой форме**, в том числе в форме просмотра рекламы, в форме платы за носитель или за сам акт копирования и передачи, даже если такая плата оказывается значительно меньше фактической стоимости или себестоимости носителя, акта копирования и т. п.

Любые другие способы распространения данного файла при отсутствии письменного разрешения автора запрещены. В частности, **запрещается**: внесение каких-либо изменений в данный файл, создание и распространение искажённых экземпляров, в том числе экземпляров, содержащих какую-либо часть произведения; распространение данного файла в Сети Интернет через веб-сайты, оказывающие платные услуги, через сайты коммерческих компаний и индивидуальных предпринимателей (включая файлообменные и любые другие сервисы, организованные в Сети Интернет коммерческими компаниями, в том числе бесплатные), а также **через сайты, содержащие рекламу любого рода**; продажа и обмен физических носителей, содержащих данный файл, даже если вознаграждение значительно меньше себестоимости носителя; включение данного файла в состав каких-либо информационных и иных продуктов; распространение данного файла в составе какой-либо платной услуги или в дополнение к такой услуге. С другой стороны, **разрешается** дарение (бесплатная передача) носителей, содержащих данный файл, бесплатная запись данного файла на носители, принадлежащие другим пользователям, распространение данного файла через бесплатные децентрализованные файлообменные P2P-сети и т. п. Ссылки на экземпляр файла, расположенный на официальном сайте автора, разрешены без ограничений.

А. В. Столяров *запрещает* Российскому авторскому обществу и любым другим организациям производить любого рода лицензирование любых его произведений и осуществлять в интересах автора какую бы то ни было иную связанную с авторскими правами деятельность без его письменного разрешения.

А. В. СТОЛЯРОВ

ПРОГРАММИРОВАНИЕ
ВВЕДЕНИЕ В ПРОФЕССИЮ

издание 2,92 **(электронное)**

в трёх томах

Том II: СИСТЕМЫ И СЕТИ

Белград — 2026

УДК 519.683+004.2+004.45
ББК 32.97
С81

Столяров, Андрей Викторович.
С81 Программирование: введение в профессию. – Изд. 2,92 (электронное), с испр. : в 3 томах / А. В. Столяров. – Белград : 2026.

Том II : Системы и сети. – 228 с. : ил.

Учебник «Программирование: введение в профессию» ориентирован на самостоятельное изучение и предполагает использование систем семейства Unix (в т. ч. Linux) в роли сквозной среды для обучения.

Во второй том учебника вошли части IV–VIII, посвящённые языку Си, основным возможностям операционной системы, взаимодействию программ через компьютерные сети, параллельному программированию и устройству ядра операционной системы.

Для школьников, студентов, преподавателей и всех, кто интересуется программированием.

УДК 519.683+004.2+004.45
ББК 32.97

© Столяров А. В., 2017
© Столяров А. В., 2026,
с изменениями

Оглавление

4. Программирование на языке Си **7**
4.1. Феномен языка Си (вместо предисловия) 7
4.2. Первые впечатления 9
4.3. Базовые средства языка Си 14
4.4. Стандартные функции ввода-вывода 37
4.5. Составной тип данных и динамические структуры 47
4.6. Макропроцессор 56
4.7. Раздельная трансляция 62
4.8. Язык Си и стиль кода 65
4.9. «Заковыристые» типы указателей 74
4.10. Ещё о возможностях стандартной библиотеки 79
4.11. (*) Полноэкранные программы на Си 82
4.12. (*) Программа на Си без стандартной библиотеки 87
5. Объекты и услуги операционной системы **91**
5.1. Операционная система: что это и зачем 91
5.2. Ввод-вывод и файловые системы 100
5.3. Процессы 110
5.4. Терминал и сеанс работы 133
6. Сети и протоколы **141**
6.1. Компьютерные сети как явление 141
6.2. Сетевые протоколы 146
6.3. Система сокетов в ОС Unix 156
6.4. Проблема очерёдности действий и её решения 162
7. Параллельные программы и разделяемые данные **171**
7.1. О работе с разделяемыми данными 173
7.2. Классические задачи взаимоисключения 178
7.3. Многопоточное программирование в ОС Unix 186
7.4. Разделяемые данные на диске 190
8. Ядро системы: взгляд за кулисы **195**
8.1. Основные принципы работы ОС 195
8.2. Управление процессами 200
8.3. Управление оперативной памятью 205
8.4. Управление аппаратурой; ввод-вывод 209
Приложение 1. Компилятор gcc 220
Приложение 2. Средства отладки 220
Приложение 3. Автоматическая сборка: утилита make 223
Приложение 4. Редактор Vim: больше возможностей 225

Список литературы 227

Содержание

4. Программирование на языке Си **7**
4.1. Феномен языка Си (вместо предисловия) . . . 7
4.2. Первые впечатления . . . 9
4.2.1. Программа «Hello, world» . . . 9
4.2.2. О завершении программы . . . 11
4.2.3. Квадратное уравнение . . . 11
4.2.4. Как узнать имя нужного заголовочного файла . . . 14
4.3. Базовые средства языка Си . . . 14
4.3.1. Объявления и описания функций . . . 15
4.3.2. Переменные и их описание . . . 15
4.3.3. Встроенные типы . . . 16
4.3.4. Литералы (константы) разных типов . . . 17
4.3.5. Операции и выражения . . . 18
4.3.6. Операторы языка Си . . . 21
4.3.7. Перечислимый тип . . . 25
4.3.8. Введение констант с помощью перечислимого типа . . . 26
4.3.9. Перечислимый тип и оператор выбора . . . 26
4.3.10. Локальные «статические» переменные . . . 27
4.3.11. Адреса, указатели и операции над ними . . . 28
4.3.12. Организация «выходных» параметров функций . . . 29
4.3.13. Массивы . . . 29
4.3.14. Динамическая память . . . 30
4.3.15. Модификатор `const` . . . 30
4.3.16. Понятие леводопустимого выражения (*lvalue*) . . . 31
4.3.17. Инициализаторы для массивов . . . 32
4.3.18. Строки . . . 32
4.3.19. Строковые литералы . . . 34
4.3.20. Аргументы командной строки . . . 35
4.3.21. **(*)** Точки следования (*sequence points*) . . . 35
4.3.22. Избранные примеры программ . . . 36
4.4. Стандартные функции ввода-вывода . . . 37
4.4.1. Посимвольный ввод-вывод . . . 37
4.4.2. Форматированный ввод-вывод . . . 39
4.4.3. Работа с текстовыми потоками . . . 41
4.4.4. Ввод-вывод отдельных строк . . . 43
4.4.5. О буферизации ввода-вывода . . . 45
4.4.6. «Вывод» в строку и «ввод» из строки . . . 46
4.4.7. Блочный ввод-вывод . . . 46
4.5. Составной тип данных и динамические структуры . . . 47
4.5.1. Структуры . . . 47
4.5.2. Односвязные списки . . . 49
4.5.3. Двусвязные списки . . . 52
4.5.4. Простое бинарное дерево поиска . . . 53
4.5.5. Объединения и вариантные структуры . . . 54
4.5.6. Директива `typedef` . . . 55
4.6. Макропроцессор . . . 56
4.6.1. Предварительные сведения . . . 56
4.6.2. Макроопределения и макровызовы . . . 56
4.6.3. Соглашения об именовании . . . 57
4.6.4. Более сложные возможности макросов . . . 57
4.6.5. **(*)** Макросы и конструкция `do { } while(0)` . . . 59
4.6.6. Директивы условной компиляции . . . 59
4.6.7. Ещё несколько полезных директив . . . 60
4.6.8. Директива `#include` . . . 61
4.6.9. Особенности оформления макродиректив . . . 61
4.7. Раздельная трансляция . . . 62
4.7.1. Общая схема раздельной трансляции в Си . . . 62
4.7.2. Видимость объектов из других модулей . . . 62
4.7.3. Заголовочные файлы к модулям . . . 63
4.7.4. Защита от повторного включения . . . 64
4.7.5. Объявления типов; неполные типы . . . 65
4.8. Язык Си и стиль кода . . . 65
4.8.1. Вспоминаем общие принципы . . . 65
4.8.2. Фирменные особенности Си . . . 66
4.8.3. Побочные эффекты . . . 68
4.8.4. Возвращаемые значения и выходные параметры . . . 70
4.8.5. **(*)** Программы, говорящие по-русски . . . 71
4.9. «Заковыристые» типы указателей . . . 74
4.9.1. Многомерные массивы и указатели на массивы . . . 74
4.9.2. Указатели на функции . . . 75
4.9.3. Функции обратного вызова (callback) . . . 76
4.9.4. Сложные описания и общие правила их прочтения . . . 77
4.10. Ещё о возможностях стандартной библиотеки . . . 79
4.10.1. Дополнительные функции работы с кучей . . . 79
4.10.2. Функции обработки строк . . . 79
4.10.3. Генерация псевдослучайных чисел . . . 80
4.10.4. **(*)** Средства создания вариадических функций . . . 81
4.11. **(*)** Полноэкранные программы на Си . . . 82
4.11.1. Простой пример . . . 82
4.11.2. Обработка клавиатурных и других событий . . . 83
4.11.3. Управление цветом и атрибутами символов . . . 84
4.11.4. Клавиатурный ввод с тайм-аутами . . . 86
4.11.5. Обзор остальных возможностей `ncurses` . . . 87
4.12. **(*)** Программа на Си без стандартной библиотеки . . . 87

5. Объекты и услуги операционной системы **91**
5.1. Операционная система: что это и зачем . . . 91
5.1.1. Роль и место операционной системы . . . 91
5.1.2. Управление пользовательскими задачами . . . 92
5.1.3. Управление внешними устройствами . . . 92
5.1.4. Границы зоны ответственности ОС . . . 93
5.1.5. Unix: семейство систем и образ мышления . . . 94
5.1.6. Замечания о системе X Window . . . 96
5.1.7. Системные вызовы и их обёртки . . . 97
5.1.8. О разграничении полномочий . . . 98
5.2. Ввод-вывод и файловые системы . . . 100
5.2.1. Знакомьтесь: файловая система . . . 100
5.2.2. Права доступа к файлам . . . 102
5.2.3. Чтение и запись содержимого файлов . . . 103
5.2.4. Управление объектами файловой системы . . . 106
5.2.5. Файлы устройств и классификация устройств . . . 108
5.2.6. Работа с содержимым каталогов . . . 109
5.2.7. Отображение файлов в память . . . 109
5.3. Процессы . . . 110
5.3.1. Процесс: что это такое . . . 110
5.3.2. Свойства процесса . . . 111
5.3.3. Виртуальное адресное пространство . . . 112
5.3.4. Порождение процесса . . . 114
5.3.5. Замена выполняемой программы . . . 115
5.3.6. Завершение процесса . . . 117
5.3.7. Ожидание завершения; процессы-зомби . . . 118
5.3.8. Пример запуска внешней программы . . . 119
5.3.9. Выполнение процессов и время . . . 119
5.3.10. Перенаправление потоков ввода-вывода . . . 121
5.3.11. Полномочия процесса . . . 122
5.3.12. Количественные ограничения . . . 124
5.3.13. Обзор средств взаимодействия процессов . . . 124
5.3.14. Сигналы . . . 125
5.3.15. Каналы . . . 131
5.3.16. Краткие сведения о трассировке . . . 133
5.4. Терминал и сеанс работы . . . 133
5.4.1. Драйвер терминала и дисциплина линии . . . 133
5.4.2. Сеансы и группы процессов . . . 135
5.4.3. Управление драйвером терминала . . . 137

5.4.4. По ту сторону псевдотерминала . . . 139
5.4.5. Процессы-демоны . . . 140

6. Сети и протоколы 141
6.1. Компьютерные сети как явление . . . 141
6.1.1. Сети и сетевые соединения . . . 141
6.1.2. Шлюзы и маршрутизация . . . 144
6.2. Сетевые протоколы . . . 146
6.2.1. Понятие протокола и модель OSI . . . 146
6.2.2. Физические и канальные протоколы . . . 146
6.2.3. Протокол IP . . . 147
6.2.4. Дейтаграммы и потоки . . . 152
6.2.5. Протоколы прикладного слоя . . . 152
6.2.6. Доменные имена . . . 155
6.3. Система сокетов в ОС Unix . . . 156
6.3.1. Семейства адресации и типы взаимодействия . . . 157
6.3.2. Сокет и его сетевой адрес . . . 157
6.3.3. Приём и передача дейтаграмм . . . 159
6.3.4. Потоковые сокеты. Клиент-серверная модель . . . 160
6.3.5. О «залипании» TCP-порта . . . 161
6.3.6. Сокеты для связи родственных процессов . . . 161
6.4. Проблема очерёдности действий и её решения . . . 162
6.4.1. Суть проблемы . . . 162
6.4.2. Решение на основе обслуживающих процессов . . . 163
6.4.3. Событийно-управляемое программирование . . . 163
6.4.4. Выборка событий в ОС Unix: вызов `select` . . . 164
6.4.5. Сеанс работы как конечный автомат . . . 166
6.4.6. Неблокирующее установление соединения . . . 169
6.4.7. (*) Сигналы в роли событий; вызов `pselect` . . . 169

7. Параллельные программы и разделяемые данные 171
7.1. О работе с разделяемыми данными . . . 173
7.1.1. Устаревание и целостность данных . . . 173
7.1.2. Взаимоисключения. Критические секции . . . 174
7.1.3. Взаимоисключение с помощью переменных . . . 175
7.1.4. Понятие мьютекса . . . 176
7.1.5. О реализации мьютексов . . . 176
7.1.6. Семафоры Дейкстры . . . 177
7.2. Классические задачи взаимоисключения . . . 178
7.2.1. Задача производителей и потребителей . . . 178
7.2.2. Задача о пяти философах и проблема тупиков . . . 179
7.2.3. Граф ожидания . . . 182
7.2.4. Проблема читателей и писателей . . . 183
7.2.5. Задача о спящем парикмахере . . . 184
7.3. Многопоточное программирование в ОС Unix . . . 186
7.3.1. Библиотека Posix Threads . . . 186
7.3.2. Семафоры и мьютексы . . . 187
7.3.3. Демонстрационный пример . . . 188
Заключение . . . 190
7.4. Разделяемые данные на диске . . . 190
7.4.1. Обзор имеющихся возможностей . . . 190
7.4.2. Создание дополнительного файла . . . 191
7.4.3. Системный вызов `flock` . . . 192
7.4.4. Файловые захваты POSIX . . . 193
7.4.5. Некоторые проблемы файловых захватов . . . 194

8. Ядро системы: взгляд за кулисы 195
8.1. Основные принципы работы ОС . . . 195
8.1.1. Ядро как обработчик запросов . . . 195
8.1.2. Загрузка и жизненный цикл ОС UNIX . . . 197
8.1.3. Эмуляция физического компьютера . . . 198
8.1.4. Структура и основные подсистемы ядра . . . 200
8.2. Управление процессами . . . 200
8.2.1. Процесс как объект ядра системы . . . 200
8.2.2. Планирование времени процессора . . . 202
8.2.3. Обработка сигналов . . . 204
8.3. Управление оперативной памятью . . . 205
8.3.1. Проблемы, решаемые менеджером памяти . . . 205
8.3.2. Виртуальная память и подкачка . . . 206
8.3.3. Простейшая модель виртуальной памяти . . . 206
8.3.4. Сегментная организация памяти . . . 206
8.3.5. Страничная организация памяти . . . 207
8.3.6. Сегментно-страничная организация памяти . . . 209
8.4. Управление аппаратурой; ввод-вывод . . . 209
8.4.1. Две точки зрения на ввод-вывод . . . 209
8.4.2. Взаимодействие ОС с аппаратурой . . . 209
8.4.3. Драйверы . . . 210
8.4.4. Ввод-вывод на разных уровнях ВС . . . 212
8.4.5. О роли аппаратных прерываний . . . 213
8.4.6. Буферизация ввода-вывода . . . 213
8.4.7. Планирование дисковых обменов . . . 215
8.4.8. Виртуальная файловая система . . . 216
8.4.9. Файловая система на диске . . . 217
8.4.10. Шина, кеш и DMA . . . 218

Приложение 1. Компилятор gcc . . . 220
Приложение 2. Средства отладки . . . 220
Отладчик gdb . . . 220
Утилита strace . . . 222
Программа valgrind . . . 222
Приложение 3. Автоматическая сборка: утилита make . . . 223
Простейший `Makefile` . . . 224
Переменные и псевдопеременные . . . 224
Обобщённые цели . . . 224
Псевдоцели . . . 225
Автоматическое отслеживание зависимостей . . . 225
Приложение 4. Редактор Vim: больше возможностей . . . 225

Список литературы . . . 227

Список донэйторов, поддержавших проект создания этой книги в ходе подготовки двух первых изданий (кроме пожелавших сохранить анонимность):

Gremlin, Grigoriy Kraynov, Шер Арсений Владимирович, Таранов Василий, Сергей Сетченков, Валерия Шакирзянова, Катерина Галкина, Илья Лобанов, Сюзана Тевдорадзе, Иванова Оксана, Куликова Юлия, Соколов Кирилл Владимирович, jeckep, Кулёва Анна Сергеевна, Ермакова Марина Александровна, Переведенцев Максим Олегович, Костарев Иван Сергеевич, Донцов Евгений, Олег Французов, Степан Холопкин, Попов Артём Сергеевич, Александр Быков, Белобородов И. Б., Ким Максим, artyrian, Игорь Эльман, Илюшкин Никита, Кальсин Сергей Александрович, Евгений Земцов, Шрамов Георгий, Владимир Лазарев, eupharina, Николай Королев, Горошевский Алексей Валерьевич, Леменков Д. Д., Forester, say42, Аня «canja» Ф., Сергей, big_fellow, Волканов Дмитрий Юрьевич, Танечка, Татьяна 'Vikora' Алпатова, Беляев Андрей, Лошкины (Александр и Дарья), Кирилл Алексеев, kopish32, Екатерина Глазкова, Олег «burunduk3» Давыдов, Дмитрий Кронберг, yobibyte, Михаил Аграновский, Александр Шепелёв, G.Nerc=Y.uR, Василий Артемьев, Смирнов Денис, Pavel Korzhenko, Руслан Степаненко, Терешко Григорий Юрьевич 15e65d3d, Lothlorien, vasiliandets, Максим Филиппов, Глеб Семёнов, Павел, unDEFER, kilolife, Арбичев, Рябинин Сергей Анатольевич, Nikolay Ksenev, Кучин Вадим, Мария Трофимова, igneus, Александр Чернов, Roman Kurynin, Власов Андрей, Дергачёв Борис Николаевич, Алексей Алексеевич, Георгий Мошкин, Владимир Руцкий, Федулов Роман Сергеевич, Шадрин Денис, Панфёров Антон Александрович, os80, Зубков Иван, Архипенко Константин Владимирович, Асирян Александр, Дмитрий С. Гуськов, Тойгильдин Владислав, Masutacu, D.A.X., Каганов Владислав, Анастасия Назарова, Гена Иван Евгеньевич, Линара Адылова, Александр, izin, Николай Подонин, Юлия Корухова, Кузьменкова Евгения Анатольевна, Сергей «GDM» Иванов, Андрей Шестимеров, vap, Грацианова Татьяна Юрьевна, Меньшов Юрий Николаевич, nvasil, В. Красных, Огрызков Станислав Анатольевич, Бузов Денис Николаевич, capgelka, Волкович Максим Сергеевич, Владимир Ермоленко, Горячая Илона Владимировна, Полякова Ирина Николаевна, Антон Хван, Иван К., Сальников Алексей, Щеславский Алексей Владимирович, Золотарев Роман Евгеньевич, Константин Глазков, Сергей Черевков, Андрей Литвинов, Шубин М. В., Сыщенко Алексей, Николай Курто, Ковригин Дмитрий Анатольевич, Андрей Кабанец, Юрий Скурский, Дмитрий Беляев, Баранов Виталий, Новиков Сергей Михайлович, maxon86, mishamm, Спиридонов Сергей Вячеславович, Сергей Черевков, Кирилл Филатов, Чаплыгин Андрей, Виктор Николаевич Остроухов, Николай Богданов, Баев Ален, Плосков Александр, Сергей Матвеев a.k.a. stargrave, Илья, aykar, Олег Бартунов, micky_madfree, Алексей Курочкин aka kaa37, Николай Смолин, I, JDZab, Кравчик Роман, Дмитрий Мачнев, bergentroll, Иван А. Фролов, Александр Чащин, Муслимов Я. В., Sedar, Максим Садовников, Яковлев С. Д., Рустам Кадыров, Набиев Марат, Покровский Дмитрий Евгеньевич, Заворин Александр, Павлочев Сергей Юрьевич, Рустам Юсупов, Noko Anna, Андрей Воронов, Лисица Владимир, Алексей Ковура, Чайка Леонид Николаевич, Коробань Дмитрий, Алексей Вересов, suhorez, Ольга Сергеевна Цаун, Боборыкин Сергей Владимирович, Олохтонов Владимир, Александр Смирницкий, Максим Клочков, Анисимов Сергей, Вадим Вадимович Чемодуров, rumiantcev, babera, Артём Коротченко, Евгений Шевкунов, Александр Смирницкий, Артём Шутов, Засеев Заурбек, Konstantin Slobodnyuk, Yan Zaripov, Виталий Бодренков, Александр Сергиенко, Денис Кузаков, Пушистый Шмель, Сергей Спивак, suuuumcaa, Гагарин, Валерий Гайнуллин, Александр Махаев (mankms), VD, А. Б. Лихачёв, Col_Kurtz, Espanka, Дмитрий Сергеевич Х., Анатолий Камчатнов, Табакаев Евгений Владимирович, Александр Трошенков, Малюга Андрей, Сорокин Андрей, Буркин Иван, Александр Логунов, moya_bbf3, Vilnur_Sh, Кипнис Александр, Oleg G. Geier, Владимир Исаев (fBSD), Филимонов Сергей Владимирович, vsudakou, AniMath, Данилов Евгений, Воробьёв В. С., mochalov, Камчатская LUG, Логинов Сергей Владимирович, Чистяков Артём Андреевич, A&A Сулимовы, Денисов Денис Юрьевич, Сутупов Андрей Михайлович, kuddai, Озерицкий Алексей Владимирович, alexz, Владимир Цой, Владимир Бердовщиков, Сергей Дмитриченко, Иванцов Данил Игоревич, Замыслов Д. А., Владимир Халямин, Максим Карасев (begs), ErraticLunatic, А. Е. Артемьев, FriendlyElk, Алексей Спасюк, Андриевский Константин Андреевич, Сукачёв Владислав Вячеславович, Артём Абрамов, maxon86, Соколов Павел Андреевич, Алексей Н, Никита Гуляев, Бодюль Евгений, rebus_x, Рачинский Максим Юрьевич, Мезенцев Андрей Владимирович, Ижмяков Павел Андреевич, Денис К., Умнова Наталья Евгеньевна, Артём Жилеев, Легоцкий Николай Евгеньевич, Виталий Саган, Александр Д., Иванов Павел Николаевич, Роман Викторович Щербинин, Константин Хомутов, barsik, Чистяков Сергей, Дмитрий Нурисламов, Евгений lipklim Климов, Азаров Владимир, Иванов Виталий Сергеевич, Бичекуев Аслан Расулович, Владимир Бахтин, Иванцов Данил Игоревич, Смирнов Иван, lexizz19, Содномов Алексей Эрдэни-Баирович, yshawn, Авдеенков Василий, Андрей, Игорь, Денис Николаевич Ж., Aglaro, aoratos, Шарапов Кирилл Владимирович, Александр, Нестеренко Владислав Валерьевич, Рябчиков Дмитрий Александрович, Андрей Александрович Герман, pashak, Морозов Михаил Павлович, Тихонов Руслан Владимирович, Копкин Максим, ЛАГ-5891, Евгений Феликсович Набоков, Плотников Сергей, Худяев Глеб, Максим Шаталин, Пётр Захаров, Михаил, Тимофей Дубский, Андрей Анатольевич Бородин, gunpi, Зеленин Андрей, Плаксицкая Наталия, enoelia, Юрий Сидоров, Dizzy, Дмитрий Серегин, Михаил К., Павел Егоров, Даниил Шелепанов, Дьяконов Олег, Данилов Роман Викторович, le34, Константин Плахетка, Илья Денисов, Вадим Марченко, Ратников Сергей Валерьевич, Илья Алексеевич Белоусов, Кузьмин Ринат, savtog, Валеев Эдуард Накипович, paq1311, e_matveev, Ефремов Артур, братишка, Иван Михайлович Пузырёв, Jamezzz, Мелкумян Завен, Игорь Романович Шульга, Сергей Спивак, Василий Дзюбенко, sadsnake, Игорь Середов, Pavel Heiden, Богдан Артемьев, A6e3iana, В. М. Сотников, Ярослав Подымский, Андрей Барышев, Лиана Саргсян, Юрий Городничев, Дрожак Александр

Список донэйторов, поддержавших новое (электронное) издание (актуален на 30 апреля 2026 г.):

Даниил Киреев, Александр Шмигидер, posyx, okuznetc, Владислав Беин, Епишин Василий Ильич, user12, Eugene_Palkowski, Шиховцев Евгений, Василий Дзюбенко, Александр Хахмович, vostelmakh, Степан Клюкин, serereg, Пестриков Михаил, Владислав Юсов, Шестаков Егор, Анастасия Назарова, Nick Anoshin, Alex Kotovsky, Александр Морев, Валерий Акуленко, Александр Мукашев, Мария Теплинская, Максим Каланчекаев, Глеб Худяев, Hleb, Павел Зенов, Андрей Столяр, Idel A., Халитов Эльдар, Егор Гавриленко, Maxim Klochkov, Павел Владимирович Савкин (panchous), Арсений Шер, Оськин Александр Николаевич, Якушонок Сергей, Александр Колесниченко, Ansar, Daniil Lemenkov, Morf, Алексей Исмагилов, Дмитрий Макаревич, Имиль, Александр Мокин, Фатеев Дмитрий, Евгений (dovvakkin) К., Fouer, Artem Klimov, Баранов Виталий, Пе-тян, nevill, Андрей, greatEcho, Roman Andreeff, namtyda, А. В. Игнашев, nixan, Дмитрий Червов, Николай Костромитин, Вера Майер, timar07, Пузырёв Иван, Данил Кичай, memeexpert, Дмитрий Кронберг, andry0980, Timur Demyanov, adrianopol, Владимир К., Анатолий, Нецветаев Владимир, Максим Лавров, Ярослав Шальнов, Yury K., Маргарита Каранда, Максим Переведенцев, Дмитрий Володин, sib_coder, Андрей Т., Татьяна Белова, Аристов М.С., pupiz, andliar, Михаил, Юрий Широков, kezhich, Евгений Р., Додевски А.С, Костя, voothi, Евгений, Vasiliy Taranov, asterite, Пузырёв Иван Михайлович, Роман Филиппов, Федотов Максим Андреевич, Иван, Мита Таики, Белобородов И. Б., Миронов Владимир Вячеславович, Артемий Коваленко

Ещё раз спасибо всем тем, кто принял участие в финансировании проекта. Без донэйторов этой книги не было бы вообще ни в каком виде.

Часть 4

Программирование на языке Си

4.1. Феномен языка Си (вместо предисловия)

К настоящему моменту у вас уже есть опыт программирования на Паскале и языке ассемблера. Язык Си во многом напоминает Паскаль: во всяком случае, здесь тоже есть переменные и их типы, есть операторы, в том числе хорошо знакомые нам операторы ветвления и цикла, есть и подпрограммы, которые здесь всегда называются «функциями»; впрочем, аналоги паскалевских процедур в Си тоже присутствуют, хотя их так и не называют. Точно так же, как в Паскале, в Си активно применяется ***составной оператор***, хотя выглядит он на первый взгляд совсем иначе: вместо паскалевских слов `begin` и `end` в Си для группировки операторов используются фигурные скобки.

Интересно, что *визуально* при переходе от Паскаля к Си как раз бросается в глаза вот эта замена `begin` и `end` на скобки; находятся даже люди, которые, если их спросить, в чём разница между Паскалем и Си, тут же вспомнят именно про это, а больше толком ничего сказать не смогут; некоторые ещё добавят, что в Паскале присваивание обозначается символом `:=`, тогда как в Си — простым знаком равенства, как будто это так важно. Как правило, это означает, что минимум одного из двух языков человек не знает. На самом деле совершенно неважно, как именно изображается та или иная сущность; суть, наоборот, в том, что и «там», и «здесь» есть присваивание и есть ***операторные скобки***, с помощью которых создаётся составной оператор. Вспомните, ведь и при обсуждении Паскаля мы всегда называли `begin` и `end` операторными скобками, так что применение в этой роли настоящих скобок (пусть и фигурных) делает синтаксис ближе к используемой терминологии, но *по сути* ничего в этом плане не меняется. Как правило, к этой «разнице» адаптироваться проще всего. Когда визуальные различия перестают резать глаз (а происходит это почти сразу, люди вообще неплохо умеют адаптироваться), становится видна *концептуальная* разница между двумя языками, и чем дальше продвигается изучение языка, тем эта разница серьёзнее. Визуально её ухватить невозможно, тут требуется понимание происходящего и определённый опыт — в частности, если вы не будете активно писать программы на Си, либо если вы в прошлом не дали себе труда плотно попрограммировать на Паскале, идеологические различия этих языков, скорее всего, пройдут мимо вас. Между тем, именно эти различия, будучи, возможно, не столь важны в процессе непосредственного написания кода (в самом деле, мы ведь пишем на каком-то одном языке, что нам за дело до другого), при этом позволяют существенно повысить свой уровень восприятия инструмента, перейти от вопроса «как устроен язык» к вопросу «почему он устроен именно так» и со временем дойти до высших уровней постижения предмета, на которых обсуждается вопрос о том, как должен быть устроен идеальный язык программирования; отметим, что, разумеется, ответа на этот вопрос не знает никто, иначе такой язык уже давно был бы создан. Но даже не имея готового ответа на вопрос, в чём заключается идеал языка программирования, мы вполне можем, обладая достаточным уровнем понимания этих материй, обоснованно показать, каковы *недостатки* каждого конкретного языка программирования, и вот с этим квалифицированному программисту весьма желательно уметь справляться.

Вернёмся, впрочем, с небес на землю. Язык Си может показаться странным; мы не слишком погрешим против истины, если заявим, что это так и есть — язык действительно странный. Адекватной замены этому языку нет и не предвидится, есть такие классы задач, для которых просто нет других подходящих языков; придётся, как следствие, терпеть Си таким, каков он есть. Принять этот язык как феномен вам поможет понимание того, откуда он взялся, почему он именно таков, почему, несмотря на все свои выверты и выкрутасы, этот язык продолжает удерживать первое место по популярности[1] и почему, наконец, программиста, не знающего Си, всегда и везде будут воспринимать как сотрудника второго сорта, *даже если писать на Си от него не требуется*.

История создания языка Си неразрывно связана с возникновением ОС Unix, о чём мы уже рассказывали в первом томе (см. §1.2.2). Первая версия этой системы была написана Кеном Томпсоном на автокоде (ассемблере) для машины PDP-7, которая к тому времени была уже устаревшей и не представляла особого интереса; зато Кена Томпсона с его игрушками с этой машины некому было согнать. Имевшееся для этой машины системное программное обеспечение Томпсона не устроило, так что он написал операционную систему сам; так и появился Unix. В процессе дальнейшей работы Кену Томпсону понадобился язык высокого уровня, достаточно простой, чтобы его можно было реализовать быстро и без существенных трудозатрат. Для этого он придумал очень усечённую версию существовавшего ранее языка BCPL, назвав результат просто B (читается «Би»), и написал интерпретатор этого языка.

Позже Томпсон попытался переписать на Би всю свою систему, с тем чтобы её не нужно было каждый раз переписывать с нуля при переносе на другие машины. К тому времени система уже работала на популярной тогда PDP-11, которая была несовместима с PDP-7; своей очереди ждали другие аппаратные архитектуры, и перспектива каждый раз перелопачивать всю систему Томпсону совершенно не нравилась.

Попытка использовать Би в качестве нового языка реализации ядра системы успехом не увенчалась, слишком уж примитивен был этот язык; достаточно сказать, что *типов данных в нём не было*, вся работа с данными проходила в терминах машинных слов. Положение удалось исправить Дэннису Ритчи, который вовремя присоединился к Томпсону в его экспериментах. Язык, созданный на основе Би, назвали, недолго думая, следующей буквой английского алфавита; так появился язык C (Си).

Дэнниса Ритчи часто называют автором языка Си, поскольку его соавтор по книге «Язык Си» [1] Брайан Керниган заявил в одном из своих поздних интервью, что не принимал участия в создании языка и что всё это результат работы Ритчи; но, судя по всему, правильней считать, что у языка Си было два автора: Дэннис Ритчи и Кен Томпсон; в конце концов, Си был во многом инспирирован языком Би, который придумал Томпсон, и именно Томпсон стал первым активным пользователем Си, переписав на нём свою операционную систему.

Во всей этой истории выделяются три основных фактора, позволяющих понять Си как явление. Во-первых, язык был создан *в качестве заменителя языка ассемблера*; во-вторых, одним из важнейших соображений при его создании была *простота реализации*; в-третьих, язык был, что называется, «слеплен» под конкретную задачу, вставшую здесь и сейчас, и делали его, как говорят в таких случаях, «на коленке». Вряд ли Дэннис Ритчи, в то время ещё очень молодой, мог предполагать, что создаваемый им язык *переживёт своего создателя* и, уже просуществовав больше полувека (а описываемые события происходили в начале 1970-х годов), всё ещё не будет выказывать никаких признаков надвигающейся старости; сейчас можно достаточно смело предсказывать, что ещё по меньшей мере полтора-два десятка лет языку Си ничего не грозит, даже если вдруг кому-то удастся создать ему полноценную замену.

С технологической точки зрения ситуация сейчас выглядит так. Имеются по меньшей мере две области задач, где единственной альтернативой Си оказывается язык ассемблера: это, во-первых, ядра операционных систем, и, во-вторых, прошивки для микроконтроллеров —

[1] По объёму программного кода, находящегося в активном использовании, Си был и остаётся бесспорным мировым лидером.

специализированных компьютеров, используемых для управления техникой (лифтами, стиральными машинами и т. п.); даже билет на метро в Москве представляет собой не что иное, как компьютер на основе микроконтроллера. К вопросу о том, *почему* это так и чем здесь не устраивают другие языки, мы вернёмся позже, ближе к концу этой части; пока просто отметим, что работа на языке ассемблера оказывается в десятки раз более трудоёмкой, причём если когда-то давно ещё можно было ссылаться на высокую эффективность ассемблерных программ, то сейчас оптимизаторы кода, встроенные прямо в компилятор, добиваются таких результатов по быстродействию, что человек, работая вручную, в большинстве случаев просто не может сделать лучше. Поэтому, когда дело доходит до низкоуровневого программирования, язык ассемблера используется лишь для тех редких и незначительных по объёму фрагментов, которые *не могут быть сделаны даже на Си*; примером такого фрагмента может служить точка входа в обработчик прерывания в ядре ОС или, например, обращение к портам ввода-вывода в драйвере, и это обычно несколько строчек. Бóльшую часть низкоуровневой программы пишут именно на Си, что позволяет экономить дорогостоящее время программистов.

В последующих параграфах мы увидим, что Си во многом нелогичен, несуразен и вообще кошмарен, его рождение в виде «наколенной поделки» даёт себя знать. Как говорят в таких случаях, *но любим мы его не за это*: писать на языке ассемблера было бы ещё хуже, а других вариантов может просто не быть. Кроме уже упоминавшихся ядер операционных систем и прошивок для микроконтроллеров, то есть программ, работающих «непосредственно на железе», которые реально больше ни на чём не написать, низкоуровневое программирование часто применяется и в других областях, как правило, ради сокращения *зависимостей от внешних условий*. Известно, что именно на чистом Си написано большое количество *переносимых программ*, таких, которые без существенных изменений запускаются и работают в совершенно разных операционных средах и на разных аппаратных платформах. Само по себе это выглядит несколько неожиданно, ведь программа на низком уровне *учитывает особенности платформы*, на то это и низкий уровень; практика показывает, что учесть особенности разнообразных платформ в программе на Си оказывается во многих случаях проще, нежели обеспечивать единообразие работы интерпретаторов, компиляторов и библиотек, поддерживающих программирование на высоком уровне, абстрагированном от машины; программа, конечно, при этом может вообще никак не учитывать особенности аппаратуры и операционки, вот только библиотеки и трансляторы сами по себе оказываются программами низкого уровня, а их объём и сложность, естественно, могут быть гораздо больше, нежели объём и сложность нашей программы.

Итак, у нас *есть причины терпеть язык Си*, притом именно таким, каков он есть. Главное при этом — не забывать, что мы его именно что *терпим* и не начинать, как это часто происходит с любителями этого языка, на нём *думать*, полностью игнорируя его недостатки и даже воспринимая их как достоинства. К счастью, Си для вас будет не первым языком программирования, так что вы уже избегли части опасностей, которые содержит в себе этот язык для неокрепшего мозга начинающего программиста, но это не повод утрачивать бдительность.

Бдительности от вас потребуют также предпринятые индустрией судорожные попытки *улучшить* язык Си путём принятия новых «стандартов» — сначала C99, потом C11, и совсем недавно — C23. Эти стандарты привели к появлению совершенно нелепого явления, растерявшего большинство основных привлекательных черт языка Си, но сохранившего все его недостатки. При этом комитеты по стандартизации фактически заявляют всему сообществу, что, мол, язык Си, к которому вы привыкли, отныне представляет собой не то, что было раньше, и вам всем придётся с этим смириться. Следует отметить, что так происходит не только в области языков программирования. В последние десятилетия технические стандарты во всём мире всё чаще становятся инструментом крупных корпораций по вытеснению с рынка независимых разработчиков и насаждению «решений», предлагаемых крупными игроками. В довершение картины ISO *запрещает* открытую публикацию своих стандартов и взимает плату за их тексты; если считать одним из основных предназначений технической стандартизации *совместимость* продукции различных производителей, то такая политика ISO очевидным образом напрямую противоречит целям, которым вроде бы должна соответствовать.

Несомненно, совместимость устройств и программ, созданных разными производителями — вещь важная, нужная и при этом не может быть достигнута без формальных *спецификаций*, однако здесь мы сталкиваемся с массовой (и, возможно, в ряде случаев злонамеренной) подменой понятий. Всякий технический стандарт, разумеется, представляет собой спецификацию, но обратное категорически неверно.

Спецификация может описывать нечто уже существующее (реализованное), и когда спецификацию создают технические специалисты, не входящие ни в какие комитеты, обычно именно так и происходит. Такую спецификацию можно рассматривать как заявление вида «я тут что-то сделал, кто хочет — присоединяйтесь». Очевидно, каждый вправе «что-то сделать» (конечно, если это не боевое отравляющее вещество и не ядерная бомба, да и в этих случаях всё не столь очевидно), и, разумеется, *сообщить* об этом широкой публике — тоже. Естественно, это никого ни к чему не обязывает и обязывать не может, но кто-то вполне может посчитать предложенную спецификацию достойной того, чтобы её поддерживать.

Совершенно иное дело — стандарты. Фактически любой «официально принятый» стандарт — это приказание всему миру теперь делать вот так, а не как-то по-другому. Достаточно очевидно, что ни у кого, никогда и нигде не может быть *полномочий приказывать всему миру*; чтобы как-то заретушировать это весьма неприятное для себя обстоятельство, стандартизаторы вынуждены искать некие «источники легитимности». Как правило, в роли таковых выступают, во-первых, полномочия, полученные от правительства или даже многих правительств; во-вторых, личный авторитет отдельных лиц, участвующих в разработке стандарта; и в-третьих, якобы имеющий место учёт мнений всех «основных» заинтересованных сторон, под которыми в большинстве случаев понимаются крупные корпорации.

Неизбежным следствием этого становится образование такой замечательной сущности, как ***комитет***. О том, как комитеты функционируют и какого качества продукт выдают на выходе, сказано и написано уже столько, что, пожалуй, к этому трудно добавить что-то новое. Отметим лишь несколько достаточно очевидных фактов.

Прежде всего, комитет — это всегда сборище фактически случайных людей, которые в большинстве случаев даже не были до той поры друг с другом знакомы. Естественно, комитет принципиально не способен ни к какому поиску консенсуса и конструктивному обсуждению. Чтобы протащить через комитет то или иное новшество, нужно только достаточное количество наглости у того, кто выступает с инициативой; как правило, остальные члены комитета «одобрят» её просто потому, что им всем окажется лень выступать против, спорить, приводить аргументы, прилагать ощутимые усилия. Иногда (вообще-то крайне редко) в комитете оказывается специалист, понимающий, к каким пагубным последствиям ведут предлагаемые новшества, и готовый тратить силы и время на объяснения остальным членам комитета, почему данное конкретное новшество не должно приниматься; такого нонконформиста остальные члены комитета практически всегда воспринимают как досадное недоразумение, из-за которого приходится терять время.

Важно также понимать, что представители разнообразных организаций, откомандированные в комитет — это практически всегда сотрудники второго сорта, поскольку сотрудников первого сорта, естественно, никто не отпустит в какой-то там комитет, для них всегда есть работа в самой организации. Если в комитет приглашают серьёзного (обычно хорошо известного в сообществе) специалиста, то, как правило, исключительно в роли «свадебного генерала», поскольку в любой иной роли такой человек деятельность комитета начисто парализует. С другой стороны, заведомая склонность любых комитетов к попустительству позволяет представителям отдельных коммерческих компаний использовать стандарты в своих интересах — например, для устранения с рынка независимых разработчиков. В целом комитеты любого рода — это всегда (без исключений) явление вредоносное, никакой пользы и никакого конструктива они породить не могут просто в силу своего устройства.

Примером удачных спецификаций можно считать ранние версии соглашений, на которых основан Интернет. Документы серии RFC, написанные в эпоху становления Сети, представляют собой образец простоты, понятности и логичности. Разумеется, никакие комитеты к их созданию не причастны; значительную часть этих спецификаций создал единолично Йон Постел. Увы, заложенные им традиции оказались во многом разрушены после его смерти в 1998 году; современные RFC (многие из которых созданы «рабочими группами», то есть, по сути, теми же комитетами) часто попросту невозможно читать.

Вредоносная сущность комитетских поделий особенно отчётливо видна в области языков программирования. Мы уже упоминали «стандартный Паскаль», который не имеет ничего общего с реальностью и нигде никогда не встречается. Этот стандарт, к счастью, не мешает программистам использовать Паскаль — на практике стандарт просто игнорируется.

Меньше всего повезло в этом плане языку Си++: с принятием первого же стандарта этот некогда уникальный по своим свойствам язык оказался фактически уничтожен, превратившись в заурядный язык высокого уровня. Довершили дело принятые подряд в 2011, 2014, 2017 и 2020 гг. ещё более извращённые версии стандартов, с учётом которых полученный на выходе монстр нежизнеспособен и непригоден не только к использованию, но и к изучению. В одной из следующих частей нашей книги мы рассмотрим Си++, но предметом нашего изучения станет его усечённое подмножество, в которое включены только средства, существовавшие до принятия первого стандарта, и даже эти средства мы будем рассматривать не все.

С чистым Си ситуация несколько лучше — во всяком случае, язык пока не утратил жизнеспособности, несмотря на все усилия стандартизаторов. Тем не менее, такие

одиозные возможности, как массивы переменной длины (см. замечание на стр. 30), комплексные числа, строки из многобайтных символов и тому подобное ничуть не делают язык лучше: использовать эти возможности *недопустимо* в силу массы различных причин, но наличие их в языке вынуждает разумных людей тратить драгоценное время на объяснение другим причин этой недопустимости, причём объяснения часто не достигают цели.

Так или иначе, любой компилятор позволяет отключить возможности, пришедшие из нелепых стандартов, и писать на том языке Си, который существовал до того, как за него всерьёз принялись стандартизаторы. Именно так мы и намерены поступить.

4.2. Первые впечатления

Прежде чем приступить к рассмотрению примеров, сделаем пару важных замечаний. При работе на Паскале мы могли привыкнуть к тому, что компилятор игнорирует различие между верхним и нижним регистром. Для языка Си это не так: ключевые слова обязательно записываются в нижнем регистре, а, к примеру, слова `wordcount`, `WordCount`, `WORDCOUNT` и `WoRdCoUnT` — это четыре **разных** идентификатора. Традиции языка Си предписывают использование идентификаторов, записанных целиком в нижнем регистре; если имя состоит из нескольких слов, их разделяют подчёркиваниями, например, так: `word_count`. Из этого правила есть одно исключение, до которого мы со временем доберёмся.

И ещё одно. Для записи ***комментариев*** в Си используются комбинации «`/*`» (начало комментария) и «`*/`» (конец комментария), причём **вложенные комментарии компилятор не понимает**: если комментарий уже начался, компилятор готов обратить внимание только на «`*/`», а «`/*`» игнорирует, как и всё остальное. Как следствие, комментарии Си нельзя использовать для временного исключения из компиляции кусков кода; для этого следует пользоваться директивами условной компиляции, о которых пойдёт речь в главе о препроцессоре (см. §4.6.6).

Современные компиляторы поддерживают также «строчные» комментарии, начинающиеся с двух слэшей «`//`» и заканчивающиеся переводом строки. Этот стиль комментариев характерен для языка Си++ (именно там такие комментарии изначально появились); использование их в чистом Си часто рассматривается как дурной тон.

4.2.1. Программа «Hello, world»

Обучение языку Си традиционно начинают с примеров программ, пояснение к которым позволяет создать общее впечатление — на что похож этот язык и как с ним следует обращаться. Поступим так и мы, начав с традиционного[2] примера. Вот одна из самых коротких программ на Си; она просто печатает фразу `Hello, world`:

```
#include <stdio.h>
int main()
{
    printf("Hello, world\n");
    return 0;
}
```

Оставим пока что в покое самую первую строчку программы — ***директиву*** `#include`[3] и перейдём сразу к следующей строке, которая представляет собой *заголовок функции*. С функциями мы уже встречались в Паскале; в языке Си функция представляет собой ровно то же самое: обособленный фрагмент программы, имеющий собственное имя, выполнение которого может зависеть от параметров и результатом которого будет вычисленное значение определённого типа. Кстати, в данном случае функция называется `main`, а слово `int` задаёт её тип возвращаемого значения; в языке Си этим словом обозначается тип *целое число*. Пустые круглые скобки означают, что параметров эта функция не принимает.

Если говорить совсем строго, пустые скобки означают, что функция, возможно, принимает произвольное количество параметров, но все их благополучно игнорирует; впрочем, можете не обращать на это внимания.

На этом этапе у нас может возникнуть резонный вопрос, зачем потребовалась функция (то есть *подпрограмма*) в такой простой программе, выполняющей всего одно действие. Дело в том, что в языке Си нет никакого аналога паскалевской *головной программы*. Программа на Си состоит, грубо говоря, *из функций*, то есть любые операторы могут находиться только в телах функций и более нигде. Вне функций могут располагаться описания типов, глобальных переменных и т.п., но они операторами не являются и никаких действий не выполняют; что-то делать могут только функции. В этом плане функция `main` ничем принципиальным от других функций не отличается, за одним маленьким исключением: действует соглашение, что с неё начинается выполнение программы. Иначе говоря, исполняемый файл, полученный из программы на Си, обычно устроен так, что после загрузки его в оперативную память управление получит код, скомпилированный из функции `main`; с некоторой натяжкой можно считать, что функцию `main` вызывает операционная система при старте программы.

В действительности это, конечно же, не так. Изучая ассемблер, мы неоднократно убеждались, что реальной точкой входа служит метка с именем `_start`; впрочем, это имя можно изменить в командной строке редактора связей. Так называемая ***стандартная библиотека*** Си предоставляет фрагмент кода, помеченный меткой `_start`, который как раз и вызывает функцию `main`, а возвращённое ею значение отдаёт системному вызову `_exit`. В главе 4.12 мы напишем программу на Си без использования стандартной библиотеки; там всё это будет наглядно видно.

Коль скоро мы были вынуждены упомянуть термин «стандартная библиотека», отметим, что, вопреки своему названию, эта библиотека называется так отнюдь не потому, что описана в каком-нибудь стандарте. Хотя она там действительно описана, но появилась она задолго до принятия первых стандартов языка Си — фактически вместе с самим языком. Когда речь идёт о стандартной библиотеке (не только языка Си, но и других языков), обычно подразумевается такая библиотека, которая поставляется вместе с компилятором или интерпретатором и, как следствие, доступна якобы «всегда». Чаще всего транслятор подключает стандартную библиотеку к вашим программам, не интересуясь вашим мнением на этот счёт; в этом плане язык Си, как мы позже увидим, интересен именно тем, что от стандартной библиотеки здесь программист может отказаться.

Фигурные скобки мы уже обсуждали, они играют ту же роль, что и слова `begin` и `end` в Паскале: объединяют несколько операторов в один *составной*, а также обрамляют тело подпрограммы (т. е. функции, поскольку это единственный вид подпрограмм в Си). В отличие от Паскаля, в Си операторные скобки — это действительно скобки.

Строчка

```
printf("Hello, world\n");
```

представляет собой *вызов библиотечной функции*, которая называется `printf`. Отметим один крайне важный идеологический момент. **В отличие от Паскаля, где операторы ввода-вывода являются частью языка, в язык Си как таковой никакие средства ввода-вывода не входят**. В этот язык вообще, как мы вскоре убедимся, мало что входит: в следующих параграфах мы перечислим встроенные типы переменных, арифметические операции, из которых строятся выражения, и, наконец, операторы, то есть управляющие конструкции вроде ветвлений и циклов (которых в Си всего одиннадцать), и в итоге практически все средства языка Си исчерпаем.

Функция `printf` не является частью языка Си по меньшей мере в том смысле, что компилятор о ней ничего не знает, а если и знает, то обязан делать вид, что не знает. Более того, **сама функция `printf` написана на Си**.

Мы использовали функцию `printf`, чтобы напечатать строку, но её возможности много шире: она умеет печатать значения всех встроенных типов языка Си, причём целые числа можно печатать в разных системах счисления, можно управлять количеством печатаемых знаков и т. п. Буква `f` в названии функции происходит от слова *formatted*; говорят, что эта функция выполняет ***форматированный вывод***. Основные возможности `printf` мы рассмотрим чуть позже.

Отметим ещё один момент. Функция `printf` — это действительно *функция*; она возвращает целое число, равное количеству напечатанных символов, то есть мы могли бы, например, написать

```
x = printf("Hello, world\n");
```

---

[2] С этого примера начиналась книга «Язык Си» Кернигана и Ритчи как в ранних, так и в более поздних её изданиях; говорят, что эту программу придумал Брайан Керниган. Наш пример отличается от приведённого там; современные компиляторы хотя и готовы скушать тот пример в его исходном виде, всё же выдадут предупреждения в двух местах, тогда как наш вариант скомпилируется молча.

[3] С аналогичной директивой мы уже сталкивались при изучении языка ассемблера NASM; действительно, эта директива делает почти то же самое, но некоторые нюансы всё же есть, и мы поясним их чуть позже.

и в переменную `x` оказалось бы занесено число `13`; но нам это не нужно, поэтому мы просто игнорируем возвращённое функцией значение. В таких случаях говорят, что *функция вызвана ради побочного эффекта*, или, в более общем случае, что ради побочного эффекта было вычислено арифметическое выражение. При работе на языке Си (опять же, в отличие от Паскаля) это происходит очень часто, ведь процедур-то тут нет.

Как мы увидим чуть позже, программы на Си по меньшей мере на две трети, если не больше, состоят из *операторов вычисления выражения ради побочного эффекта*, что касается самих побочных эффектов, то программа на Си попросту только ими и занимается.

Отдельного рассмотрения заслуживает выражение `"Hello, world\n"`, которое представляет собой ***строковую константу*** или, как часто говорят, ***строковый литерал***. Следует обратить внимание на *двойные* кавычки; апострофы в Си тоже используются, но для других целей. Кроме того, внимание привлекает конструкция `\n`; это обозначение ***символа перевода строки***, то есть символа с кодом 10. Можно вспомнить, что в Паскале мы в подобной ситуации применяли оператор `writeln`:

```
writeln('Hello, world');
```

— то есть заставляли нашу программу сначала напечатать данную строку, а затем перевести строку на печати, напечатав для этого ещё и символ перевода строки. В вышеприведённом примере мы просто включили этот символ в состав печатаемой строки.

Отметим, что мы могли бы и в Паскале поступить точно так же; сделать это можно одним из следующих способов:

```
write('Hello, world'#10);
write('Hello, world'^J);
```

или даже вот так:

```
s := 'Hello, world'^J;
write(s);
```

Надо отметить, однако, что обозначение `\n`, в котором буква `n` взята от слова *new [line]*, существенно проще запомнить, нежели код символа и тем более загадочное `^J`.

Прежде чем закончить разбор этой строки программы, обратим внимание на символ точки с запятой. Роль этого символа в языке Си подобна паскалевской, но есть одно важное отличие: **если в Паскале точка с запятой *разделяла операторы*, то в Си точка с запятой *является частью синтаксиса оператора***, так что, в частности, её наличие никак не зависит от положения оператора непосредственно перед закрывающей фигурной скобкой.

Рассмотрим теперь следующую строку, «`return 0;`». Слово `return` в переводе с английского означает «возврат»; в языке Си это ***оператор возврата из функции***. Он делает две вещи: во-первых, работа функции на этом заканчивается, что можно, например, использовать для досрочного её завершения (подобно паскалевскому `exit`); во-вторых, его параметр — выражение, написанное после слова `return`, в данном случае `0` — задаёт то значение, которое ***вернёт*** функция (вспомним, что наша функция `main` описана как возвращающая целое число). В данном случае наша функция `main` вернёт число `0`.

Здесь может естественным образом возникнуть вопрос, *кому и зачем* нужно это число, ведь функцию `main` в нашей программе вроде бы никто не вызывает и её результат не анализирует. Краткий ответ состоит в том, что значение, возвращаемое из `main`, предназначается операционной системе в роли ***кода завершения***. В первом томе мы уже трижды встречались с кодом завершения — во вводной части при изучении скриптов на Bourne Shell (§1.2.15), потом при изучении Паскаля (оператор `halt`, §2.4.2), а затем — при изучении системного вызова `_exit` в части, посвящённой языку ассемблера. Вспомнив этот системный вызов, мы сможем выразиться более точно: значение, которое возвращается из функции `main`, служит аргументом вызова `_exit`; напомним, `0` в данном случае означает, что всё в порядке, т. е. программа, завершаясь, считает, что возложенную на неё миссию успешно выполнила.

Нам осталось рассмотреть, пожалуй, самую заковыристую строку в программе: директиву `#include <stdio.h>`. Как уже, несомненно, догадался читатель, эта директива в тексте программы *заменяет сама себя на полное содержимое файла* `stdio.h`, а нужно это, чтобы компилятор узнал про функцию `printf` — как уже говорилось, сам по себе он о ней не знает. Здесь следует отметить сразу несколько интересных моментов. Во-первых, обращают на себя внимание используемые угловые скобки, которые означают, что включаемый файл следует искать в системных директориях; точнее говоря, считается, что именно так следует включать любой файл, который не является сам по себе частью нашей программы; когда же мы пишем программу, состоящую из многих файлов, то для включения своих собственных файлов мы используем `#include` с параметром в двойных кавычках, например:

```
#include "mymodule.h"
```

Файл `stdio.h` — это вполне реальный файл; скорее всего, он находится в вашей системе в директории `/usr/include`, так что вы можете просмотреть его, например, командой `less`:

```
less /usr/include/stdio.h
```

Важно понимать, что в этом файле содержится только ***заголовок*** функции `printf`, а самой функции там нет; отсюда используемый суффикс «`.h`», от слова *header*, то есть «заголовок» — в английской терминологии это называется *header file*, а по-русски — «заголовочный файл». Всё, что компилятор узнает, увидев заголовок — это что *где-то* (и притом совершенно неважно, где) есть функция с таким-то именем, принимающая столько-то параметров таких-то типов; как функция выглядит, что она делает — этого компилятору знать не нужно. Дело в том, что компилятор генерирует, как мы это уже выяснили при изучении ассемблера, не готовый ***машинный код***, а ***объектный модуль***, в котором пока что не хватает некоторых адресов; в данном случае результатом работы компилятора становится модуль, не содержащий самой функции `printf`. Вместо неё модуль содержит указание на то, что редактору связей (линкеру) следует *откуда-то* добыть функцию с таким именем, а её адрес подставить куда следует. Иначе говоря, реальная функция `printf` появится в нашей программе только на этапе окончательной сборки, когда работа компилятора будет завершена.

Больше того, в реальной ситуации, скорее всего, функция `printf` вообще не будет содержаться в исполняемом файле нашей программы; после её запуска функция будет подгружена из динамической библиотеки. Впрочем, использование динамических библиотек можно отключить.

Обсудив всё это, отметим одну ошибку, столь же грубую, сколь часто встречающуюся, которую делают, к сожалению, не только начинающие программисты, но и некоторые преподаватели. На вопрос о том, что́ же делает директива `#include`, можно очень часто услышать ответ, что-де она «подключает библиотеку», а сам файл `stdio.h` — это якобы, соответственно, библиотека. Так вот, стоит осознать раз и навсегда, что **директива `#include` не имеет никакого отношения к подключению библиотек**, а заголовочные файлы, разумеется, библиотеками не являются — они являются именно заголовочными файлами и ничем иным. Язык Си вообще не предусматривает возможности *подключить библиотеку* из текста программы, потому что компилятор, попросту говоря, этим не занимается, это не его дело. Заявлять противоположное — это грубейшая ошибка, демонстрирующая невежество заявляющего. Кстати, в следующем параграфе мы будем рассматривать пример, который реально потребует подключения библиотеки, заодно и увидим, как на самом деле это делается.

Теперь, когда мы закончили разбор программы «Hello, world», самое время попробовать её запустить. Итак, включаем компьютер, входим в систему, запускаем текстовый редактор; учтите, что ваш файл должен иметь суффикс «`.c`», поскольку вы собираетесь в нём набрать программу на Си. Например, имя `hello.c` для вашего файла вполне подойдёт. Набираем ровно такой текст программы, как показано на стр. 9, и сохраняем его. Теперь нам нужно запустить компилятор, который в данном случае называется `gcc`. Это название образовано от слов *Gnu Compiler Collection*, то есть «коллекция компиляторов Gnu»; компилятор умеет обрабатывать программы не только на Си, но и на языке Си++, а при использовании дополнений, так называемых фронт-эндов, тот же компилятор можно применить для Objective-C, Фортрана, Ады и многих других языков.

В системах семейства Unix доступно много разных компиляторов Си, и возможно даже, что gcc из них сейчас не самый лучший; на момент написания этого текста gcc был просто наиболее распространённым вариантом. Если очень любопытно, попробуйте самостоятельно освоить компилятор `clang`.

Сразу же отметим, что компилятор `gcc` поддерживает множество разнообразных флажков командной строки, из которых нам потребуются по меньшей мере два: `-Wall` (от слов ***w**arnings all*), который включает *все разумные предупреждения*, и `-g`, который заставляет компилятор сгенерировать отладочную информацию. **Эти флажки мы при запуске `gcc` указываем *всегда***, то есть вообще всегда, следует выработать привычку к этому на уровне, как говорят, спинного мозга: отсутствие любого из этих флажков может очень дорого обойтись. Не указывать эти флаги можно разве что тогда, когда компилятор запускает не программист, а конечный пользователь, чтобы из присланных ему исходных текстов получить исполняемую программу. При этом предполагается, что программа уже полностью готова и отлажена, а даже если она и содержит ошибки, то конечный пользователь всё равно не станет их исправлять. Но это не наш случай: мы пользуемся компилятором как инструментом программиста.

Ещё один полезный флаг, `-o`, позволяет задать имя файла, в который будет записан результат компиляции. Если этот флажок не указать, итоговый исполняемый файл будет иметь имя `a.out`, что не всегда удобно. Как мы помним, обычно исполняемые файлы в ОС Unix не имеют суффикса. Именно так мы поступим и сейчас — отправим результат компиляции в файл `hello`:

```
gcc -Wall -g hello.c -o hello
```

Полностью эта команда означает следующее: «возьми исходный файл `hello.c`, откомпилируй его, выдавая при этом все разумные предупреждения, добавь в полученный модуль отладочную информацию, потом вызови редактор связей, скормив ему наш модуль и *стандартную библиотеку языка Си* (!), а результат пусть он запишет в файл `hello`». Промежуточный объектный файл нам в этот раз не нужен, поскольку наша программа состоит из одного модуля; компилятор поместит объектный код во временный файл, а по завершении работы его сотрёт.

Если всё сделать правильно, компилятор отработает полностью молча, не выдав ни слова, а в текущей директории появится файл `hello`, который можно будет запустить:

```
avst@host:~/work$ ./hello
Hello, world
avst@host:~/work$
```

4.2.2. О завершении программы

Как мы отмечали выше, с некоторой натяжкой можно считать, что *в роли вызывающего для функции `main` выступает операционная система*, и именно операционной системе предназначено число, которое эта функция возвращает. Более строго это значение называется ***кодом завершения***; оно используется, чтобы показать операционной системе, считает ли завершающаяся программа, что её выполнение прошло успешно. Если всё было хорошо и программе удалось выполнить ту задачу, ради которой её запускали, она завершается с кодом `0`, как в примере из предыдущего параграфа; если что-то пошло не так, например, она не смогла открыть нужный ей файл, не смогла установить связь с сервером или эта связь неожиданно разорвалась — мало ли ошибок происходит при работе программ — то, чтобы оповестить о неудаче операционную систему, программа завершается с кодом `1`, `2` и так далее. Максимально возможное значение такого кода — 255, поскольку он восьмибитный, но большие значения обычно не используются, в реальной жизни код завершения редко превышает 10.

Между прочим, на этом основаны две *самые короткие* программы на языке Си, которые называются `true` и `false`: обе они ничего не делают, завершаясь немедленно после запуска, при этом первая завершается успешно, а вторая — неуспешно. Самая простая реализация программы `true` будет такой:

```
int main() { return 0; }
```

Программа `false` выглядит почти так же:

```
int main() { return 1; }
```

Подчеркнём, что в обоих случаях написанная строчка — это *текст программы целиком*, никаких директив `#include` и чего-либо ещё не требуется. Эти две программы часто применяются при написании скриптов на командно-скриптовых языках, вроде рассмотренного нами ранее Bourne Shell.

Следует сразу же предостеречь читателя от довольно частого варианта ошибочного понимания происходящего. Поскольку все наши программы пока что состояли из одной функции `main`, оператор `return` в них завершал программу; но на самом деле этот оператор завершает выполнение *одной функции*, и если это будет не `main`, а какая-то другая функция, то, очевидно, `return` к завершению программы не приведёт. Больше того, можно вызвать и саму функцию `main` из какой-нибудь другой функции или даже из неё самой, сделав её рекурсивной; обычно так не делают, но это и не запрещено, и если так сделать, то даже в функции `main` выполнение `return` не станет завершением программы. В наших примерах `return` завершал программу лишь постольку, поскольку он завершал работу того вызова `main`, с которого работа программы началась.

Выполнение программы, написанной на Си, можно завершить не только из функции `main`, но и из любого другого места. Для этого предназначена библиотечная функция `exit`, имеющая один целочисленный параметр. Прототип этой функции описан в заголовочном файле `stdlib.h`; подключив его с помощью `#include`, мы сможем в любой из функций нашей программы написать что-то вроде

```
exit(0);
```

— и выполнение такого оператора программу немедленно завершит. Вместо `0` можно подставить любой другой код, следует только помнить, что для него допустимы значения от `0` до `255`, но обычно используется число не больше 10. Можно обратиться и напрямую к системному вызову `_exit`:

```
_exit(0);
```

— но так поступать следует лишь в случае, если вы точно знаете, что делаете. Конечно, функция `exit` тоже в конечном счёте обратится к тому же системному вызову, но перед этим она «приводит в порядок дела»; в частности, все библиотечные функции вывода, с которыми мы будем работать, *буферизуют* выводимую информацию, и если в их буферах что-то осталось, функция `exit` всю эту информацию из буферов вытеснит, то есть всё, что ваша программа выдала через буферизующие библиотечные функции, будет реально выведено; если завершить программу «грубо», напрямую обратившись к системному вызову, то какие-то из своих сообщений программа может так и не напечатать, а созданные программой файлы могут оказаться короче, чем вы могли бы ожидать.

Правила буферизации вывода мы подробно рассмотрим в §4.4.5; в сочетании с системными вызовами управления процессами буферизация может давать достаточно неожиданные эффекты, примеры которых читатель найдёт в следующей части книги — в §5.3.6.

4.2.3. Квадратное уравнение

Вторая программа, которую мы рассмотрим, будет решать квадратное уравнение. Как известно, квадратное уравнение всегда имеет ровно два корня, но поскольку здесь мы занимаемся программированием, а не математикой, случай комплексных корней мы рассматривать не будем, ограничившись лаконичным школьным «корней нет» — естественно, по-английски. Несмотря на простоту задачи, при её решении мы узнаем довольно много нового.

Чтобы достичь поставленной цели, нам потребуются переменные, способные хранить число с плавающей точкой; в языке Си для этого лучше всего подходит тип `double`[4]. Вычисление дискриминанта мы вынесем в отдельную функцию, заодно увидим, как выглядит функция с параметрами. А вот дальше нас ждут настоящие приключения. Во-первых, нам придётся узнать, как печатать числа (выдавать их представление в поток стандартного вывода) с помощью уже знакомой нам функции `printf` и как читать числа из потока стандартного ввода (с клавиатуры) с помощью другой функции, которая называется `scanf`. Во-вторых, мы обнаружим, что все параметры в Си передаются исключительно как значения, то есть никакого аналога var-параметрам («параметрам-переменным») Паскаля здесь нет; между тем, как-то нужно объяснить `scanf`'у, куда (то есть в какие переменные) девать свежепрочитанные значения. Победив коварный ввод-вывод, мы, возможно, вздохнём с облегчением, но ненадолго: дело в том, что функции, работающие с плавающей

[4] Встроенные типы мы подробно обсудим в §4.3.3.

точкой, не входят в основную библиотеку, которую компилятор подключает по умолчанию, и к функции вычисления квадратного корня сие тоже относится; так что, прежде чем программа заработает, нам придётся ещё выяснить, как же *в действительности* подключаются библиотеки (см. замечание на стр. 10).

Начнём с замечания, что функция вычисления квадратного корня называется `sqrt`, а её заголовок, наряду с заголовками других математических функций (таких как синус, логарифм, экспонента и т. п.), располагается в заголовочном файле `math.h`. Кроме того, поскольку мы собираемся использовать функции ввода-вывода, нам, как и в предыдущей программе, потребуется заголовочник `stdio.h`. Традиционно директивы `#include` располагают в начале программы, поступим так и мы; открываем в редакторе текстов новый файл (например, `qe.c` от слов *quadratic equation*, то есть «квадратное уравнение») и пишем для начала следующие две строки:

```
#include <stdio.h>
#include <math.h>
```

Вспомним теперь, что мы хотели написать отдельную функцию для вычисления дискриминанта. Для этого, как известно, нужны три коэффициента уравнения, а сам дискриминант вычисляется по формуле $D = b^2 - 4ac$. Выше мы уже договорились, что для работы с дробными числами будем использовать тип `double`, то есть и коэффициенты, и сам вычисленный дискриминант должны быть как раз этого типа. Теперь уже написать функцию не составит особого труда:

```
double discrim(double a, double b, double c)
{
    return b*b - 4*a*c;
}
```

Заметим, что описания параметров имеют вид «тип-имя»; в языке Си это традиционный способ описания, причём не только для переменных, но и, например, для функций (в заголовке тоже сначала указывается тип — в данном случае это тип возвращаемого значения, — а потом уже имя функции). Мы могли бы написать и так:

```
double discrim(double a, double b, double c)
{
    double d;
    d = b*b - 4*a*c;
    return d;
}
```

и почему-то начинающие обычно так и пишут. Здесь используется ***локальная*** переменная `d`, в которую сначала заносится вычисленное значение дискриминанта, а затем значение этой переменной возвращается из функции. С технической точки зрения это не имеет никакого смысла, но если вам так понятнее, пишите так.

Начнём теперь писать функцию `main`. Как уже говорилось, она должна возвращать значение типа `int`, которое служит *кодом завершения программы*. Параметров она у нас пока что не принимает. Нам обязательно потребуются локальные переменные — как минимум для хранения коэффициентов, а ещё довольно удобно будет сохранить в отдельной переменной вычисленное значение дискриминанта. Добавим к этому ещё одну переменную типа `int` (для хранения целого числа), зачем она нам нужна — мы поймём чуть позже.

В отличие от Паскаля, в Си можно описать локальную переменную в любом составном операторе, то есть, например, можно завести свои локальные переменные в теле цикла, и видно их будет только в этом теле; нам, впрочем, сейчас потребуется переменная, которую видно во всём теле функции. Описываются локальные переменные сразу после фигурной скобки, открывающей функцию (или составной оператор); с учётом этого начало нашей функции `main` будет выглядеть так:

```
int main()
{
    double p, q, r, d;
    int n;
```

Здесь можно заметить, что мы немного сэкономили, указав тип *один* раз и перечислив имена переменных через запятую. В описании переменных так делать можно, а вот в заголовке функции, к сожалению, нельзя, то есть мы не смогли бы добиться аналогичной экономии, например, в заголовке функции `discrim`. Если бы мы написали что-то вроде `double discrim(double a, b, c)`, то получили бы синтаксическую ошибку.

Ну что же, простая часть решения, собственно говоря, позади, а теперь начнётся самое интересное. Прежде чем идти дальше, давайте посмотрим, какие средства нам понадобятся, чтобы сначала ввести три числа, задающие коэффициенты уравнения, а затем два вычисленных корня напечатать. Начнём со второго. Как уже говорилось, функция `printf` умеет печатать не только строки, но и значения любых встроенных типов. Ранее мы использовали её с одним параметром (строкой), но на самом деле параметров у неё может быть сколько угодно, причём первым параметром в эту функцию передаётся так называемая ***форматная строка***, которая задаёт некий неизменный шаблон того, что должно быть напечатано, а дополнительные параметры, если они есть, задают то, что от случая к случаю может измениться. Функция `printf` просматривает свою форматную строку слева направо по одному символу, и если очередной символ — не «`%`», то функция такой символ просто печатает. Если в форматной строке так и не встретилось ни одного «процента», то вся эта строка будет попросту напечатана как есть, чем мы и воспользовались в программе «Hello, world».

Существенно интереснее функция будет действовать, если очередным символом окажется магический «процент». В большинстве случаев это означает, что нужно взять очередной параметр из списка параметров, преобразовать его в текстовое представление и напечатать. Один или несколько символов, идущих сразу после символа процента в форматной строке, указывают, какого типа будет этот параметр и в каком виде его печатать. Например, комбинация «`%d`» означает, что очередной параметр в списке следует рассматривать как целое число, а напечатать нужно его десятичное представление («`d`» в данном случае происходит от слова *decimal*). Комбинация «`%x`» означает практически то же самое, но напечатано число будет в шестнадцатеричной системе; ну а, к примеру, «`%05d`» означает, что печатаемое целое число будет содержать не менее пяти знаков, а если в числе знаков меньше, то оно будет дополнено слева нужным количеством нулей. Так, если мы напишем в программе

```
x = 17;
printf("%d times %d is %d\n", x, x, x*x);
```

то она напечатает «`17 times 17 is 289`» и переведёт строку; если же мы напишем

```
x = 378;
printf("%d %x %06d\n", x, x, x);
```

— то напечатано будет «`378 17a 000378`» (и перевод строки).

Подробный разговор о возможностях `printf` у нас впереди; пока достаточно заметить, что для печати числа типа `double` (а равно и чисел типа `float`; функция их не различает) можно воспользоваться комбинацией «`%f`» (от слова *float*), но результат будет несколько странно выглядеть; правильнее будет явно указать, сколько знаков после десятичной точки[5] мы хотим видеть, и делается это так: «`%.5f`» (здесь мы указали, что в дробной части должно быть пять цифр).

Для ввода коэффициентов с клавиатуры нам придётся применить функцию `scanf`, у которой есть довольно много общего с `printf`. Эта функция также может принимать произвольное количество параметров, причём первым из них должна быть уже знакомая нам *форматная строка*. Её подробный разбор мы на некоторое время отложим, заметив, что в большинстве случаев форматная строка для `scanf` должна состоять из всё тех же комбинаций, начинающихся с «процента», между которыми ставят пробелы. В частности, строка `"%lf"` будет означать, что следует прочитать из потока стандартного ввода число типа `double` и записать его в переменную того же типа[6]. Переменную, соответственно, нужно *как-то указать в виде параметра*, и

[5] Напомним на всякий случай, что в программировании для отделения дробной части от целой используется десятичная *точка*, а не десятичная запятая.

[6] В отличие от `printf`, здесь `float` и `double` различаются; для `float` мы применяли бы комбинацию `"%f"`, тогда как для `double` вынуждены применять `"%lf"`, от слов *long float*, то есть «длинное с плавающей точкой»; почему это так, мы подробно обсудим в параграфе, посвящённом форматированному вводу-выводу.

вот тут самое время вспомнить, что в Си есть только один способ передачи параметра — а именно, передача по значению. То есть передать «переменную как таковую», как мы делали это в Паскале с помощью var-параметров (см. т. 1, §2.3.4), нельзя — тут нет ничего похожего на var-параметры.

Чтобы понять, как решается эта проблема, вспомним для начала, что ***переменная*** (во всяком случае, в императивных языках программирования) — это не что иное, как *область оперативной памяти*, а к области памяти не обязательно обращаться через имя переменной, достаточно знать её (области) ***адрес***. Иначе говоря, если мы предложим функции `scanf` прочитать с клавиатуры число и дадим ей *адрес области памяти*, куда следует положить прочитанное число, она с этим вполне справится.

Уместно будет отметить, что var-параметры Паскаля на самом деле реализуются в машинном коде точно так же — передачей адреса переменной, но Паскаль от нас всю эту механику скрывает, тогда как Си придуман людьми, привыкшими к языку ассемблера, так что вполне понятно, почему они не сочли нужным ничего скрывать.

Итак, чтобы прочитать три числа с клавиатуры, мы вызовем функцию `scanf` и передадим ей, во-первых, форматную строку, чтобы проинструктировать её читать числа с плавающей точкой в количестве трёх штук; и, во-вторых, мы передадим ей дополнительными параметрами *адреса* трёх переменных типа `double`. Отметим, что ***операция взятия адреса*** обозначается в Си символом «`&`», то есть значением выражения `&t` будет адрес переменной `t`.

Отметим ещё один важный момент. Функция `scanf` *возвращает значение*, и это тот случай, когда возвращаемым значением лучше всё-таки воспользоваться. Дело в том, что пользователь может ввести что-нибудь такое, что наша функция никак не сможет превратить в число — например, произвольную белиберду из букв. В этом случае `scanf` немедленно прекращает чтение и в переданные ей области памяти ничего не записывает, то есть там так и останется мусор. В такой ситуации продолжать решать квадратное уравнение несколько глупо, гораздо правильнее будет сообщить пользователю, что он неправ. Для этого как раз и используется значение, которое `scanf` возвращает как функция, а возвращает она *целое число, равное успешно обслуженным «процентикам»*. Поскольку таких у нас три, то и вернуть `scanf` должна число `3`; иное будет означать, что пользователь ввёл что-то некорректное.

Следующие несколько строк программы будут выглядеть так[7]:

```
n = scanf("%lf %lf %lf", &p, &q, &r);
if(n != 3) {
    printf("Error: wrong input.\n");
    return 1;
}
```

С присваиванием мы уже знакомы, а лексема `!=` обозначает логическую операцию «не равно» (то, что в Паскале мы привыкли обозначать как `<>`). С оператором `if` тоже всё, скорее всего, понятно по аналогии с Паскалем, стоит лишь обратить внимание на круглые скобки вокруг условия (они здесь обязательны) и отсутствие какого-либо аналога паскалевскому слову `then` — при наличии обязательных круглых скобок такое слово не нужно.

Обратите внимание, что, получив некорректный ввод и выдав по этому поводу сообщение об ошибке, наша функция `main` немедленно завершается с помощью уже знакомого нам оператора `return`, но при этом возвращает не ноль, как мы это делали раньше, а единицу. Так она извещает операционную систему, что решить поставленную задачу не удалось; действительно, нельзя же решить квадратное уравнение, не зная его коэффициентов.

Дальше всё уже довольно просто. Если первый коэффициент равен нулю, то это уравнение не квадратное и решать его нужно совсем иначе; но наша программа предназначена для решения квадратных уравнений, а не каких-то других, так что она просто выдаст ошибку:

```
if(p == 0) {
    printf("Error: Not a quadratic equation!\n");
    return 2;
}
```

Как уже, несомненно, догадался читатель, знаком `==` обозначается сравнение, то есть приведённый оператор читается как «если `p` равно нулю, то...». Возвращаем мы в этот раз двойку, это тоже показывает системе наше недовольство, как и любое число, отличное от нуля. Но почему не единица, как в прошлый раз? Что ж, можно было вернуть и единицу тоже, но ведь ошибка-то другая. Возможно, кто-нибудь захочет написать другую программу, которая будет запускать нашу программу для решения какого-нибудь уравнения, и если мы будем выдавать разные коды ошибок, это позволит вызывающему понять, какая из ошибок имела место.

Убедившись, что первый коэффициент отличен от нуля, мы можем с полным на то основанием посчитать дискриминант, проверить его на неотрицательность и извлечь из него квадратный корень, причём сохранить этот корень можно в той же переменной, ведь сам дискриминант нам больше не понадобится:

```
d = discrim(p, q, r);
if(d < 0) {
    printf("No roots\n");
    return 0;
}
d = sqrt(d);
```

Обратите внимание, что в случае, если уравнение не имеет корней, мы завершаем программу с кодом `0` (успех). Дело в том, что отсутствие корней — это, вообще говоря, не ошибка, уравнение вполне может не иметь корней. Иначе говоря, корней мы не нашли, но *задачу как таковую решили*, ведь сообщение об отсутствии корней — это тоже ответ, притом правильный.

Случай с совпадающими корнями мы выделять не будем, так что всё, что осталось сделать — это напечатать вычисленные корни и завершить программу:

```
    printf("%.5f %.5f\n", (-q-d)/(2*p), (-q+d)/(2*p));
    return 0;
}
```

Полностью текст программы получился таким:

```
/* qe.c */
#include <stdio.h>
#include <math.h>

double discrim(double a, double b, double c)
{
    return b*b - 4*a*c;
}
int main()
{
    double p, q, r, d;
    int n;
    n = scanf("%lf %lf %lf", &p, &q, &r);
    if(n != 3) {
        printf("Error: wrong input.\n");
        return 1;
    }
    if(p == 0) {
        printf("Error: Not a quadratic equation!\n");
        return 2;
    }
    d = discrim(p, q, r);
    if(d < 0) {
        printf("No roots\n");
        return 0;
    }
    d = sqrt(d);
    printf("%.5f %.5f\n", (-q-d)/(2*p), (-q+d)/(2*p));
    return 0;
}
```

Как ни странно, это отнюдь не конец истории. Сохраним файл и попробуем откомпилировать полученный результат:

```
avst@host:~/work$ gcc -Wall -g qe.c -o qe
/tmp/ccePjy5n.o: In function ‘main’:
/home/avst/work/qe.c:26: undefined reference to ‘sqrt’
collect2: ld returned 1 exit status
```

[7] Кто-нибудь мог бы заявить нам, что сообщения об ошибках следует выдавать не на стандартный вывод, а в специально предназначенный для этого поток диагностики; терпение, господа, всему своё время!

Ключевым тут является словосочетание `undefined reference`, которое переводится как «неопределённая ссылка». Ошибку нам в этот раз выдал не компилятор, а редактор связей; компиляция-то как раз прошла успешно.

Вообще вышеприведённая диагностика может многое рассказать любопытному читателю. Для начала можно обратить внимание на имя файла `/tmp/ccePjy5n.o`; мы уже отмечали, что компилятор создаёт *временный объектный файл*, чтобы отдать его редактору связей, а потом стереть (см. стр. 11). Здесь мы видим наглядное проявление этого факта: именно этот загадочный `ccePjy5n` и есть тот самый временный файл.

Заслуживает внимания и последняя строчка, где упоминается уже знакомая нам по «ассемблерной» части команда `ld` — напомним, что это и есть редактор связей. Компилятор, завершив свою часть работы, сам вызвал программу `ld`, которая, не сумев выполнить возложенную на неё задачу (в данном случае из-за отсутствия функции `sqrt`), выдала диагностику — первые две строчки того, что мы увидели — и *завершилась с кодом 1*, тем самым сообщив вызвавшемому её компилятору, что произошла ошибка. Последнюю строчку напечатал уже сам компилятор. Как видим, вопрос о том, какое число возвращать из функции `main` в качестве кода завершения — отнюдь не праздный.

Ошибка произошла, потому что функция `sqrt` находится вне той части стандартной библиотеки языка Си, которая подключается по умолчанию. Компилятор узнал из заголовочного файла `math.h`, что функция `sqrt` существует, а также какие она принимает параметры и какого типа возвращает значение; всё это вполне совпало с тем, как мы эту функцию используем в программе, поэтому сам компилятор никаких проблем не заметил, вызвал редактор связей, и ошибка произошла уже на этапе финальной сборки. Чтобы исправить ошибку, придётся *действительно подключить* ***библиотеку***. Нужная библиотека называется «`m`» (от слова *mathematics*), а подключается она указанием в командной строке компилятора флажка `-lm`:

```
gcc -Wall -g qe.c -lm -o qe
```

В этот раз никаких проблем не возникнет; теперь мы на собственном опыте убедились, что заголовочные файлы — это никакие не библиотеки, а кроме них есть ещё и *настоящие* библиотеки.

Версии `gcc` до 4.* включительно позволяли указывать флаги, подключающие библиотеки, в произвольном месте командной строки, но в более поздних версиях это уже не так: **флаги `-l` следует располагать после имён исходных файлов**. Чтобы понять причину этого, стоит вернуться к первому тому и перечитать §3.7.5, посвящённый алгоритму работы редактора связей.

Попробуем потестировать нашу программу на простых данных:

```
avst@host:~/work$ echo "1 -2 1" | ./qe
1.00000 1.00000
avst@host:~/work$ echo "1 2 1" | ./qe
-1.00000 -1.00000
avst@host:~/work$ echo "1 -5 6" | ./qe
2.00000 3.00000
avst@host:~/work$ echo "1 1 1" | ./qe
No roots
avst@host:~/work$ echo "1 -1 -4" | ./qe
-1.56155 2.56155
```

4.2.4. Как узнать имя нужного заголовочного файла

В разобранных выше примерах нам встретилось два заголовочных файла: для использования функций ввода-вывода мы подключали `stdio.h`, а для функции вычисления квадратного корня нам понадобился `math.h`. В дальнейшем нам потребуются и другие заголовочные файлы, а общее число заголовочников, покрывающих одни только возможности стандартной библиотеки, составляет несколько десятков. К счастью, помнить, в каком заголовочнике расположена какая функция, нет ни малейшей необходимости. Если вы забудете подключить заголовочный файл, компилятор выдаст диагностику, начинающуюся с предупреждения примерно такого рода:

```
myprog.c:14:5: warning: implicit declaration of function 'malloc'
```

Конкретная формулировка предупреждения может зависеть от версии компилятора, но ключевые слова *implicit declaration* будут присутствовать обязательно. Поясним, что они переводятся как «неявное объявление»; дело в том, что в ранних версиях Си функцию разрешалось вызывать, не объявив, поэтому компилятор, следуя традициям, не считает это ошибкой. Использовать эту возможность ни в коем случае не следует, ведь если мы не дали компилятору информации о том, каковы параметры нашей функции и какое значение она возвращает, то проверить наш вызов на соответствие типов компилятор уже не сможет — а значит, сделав ошибку, мы об этом не узнаем, или, точнее говоря, узнаем, но уже во время исполнения, и может уйти очень много времени на выяснение, что за ерунда происходит.

Так или иначе, зная имя функции, которая «не понравилась» компилятору, мы можем легко понять, какого заголовочного файла не хватает нашей программе. Для этого достаточно дать команду `man` с именем функции в качестве параметра:

```
man malloc
```

Мы получим на экран текст, описывающий заданную функцию, и в самом начале этого текста будут указаны заголовочные файлы, которые необходимо подключить для её использования. Если этих файлов больше одного, а такое иногда бывает, — это означает, что подключить их нужно все, причём именно в таком порядке, как указано в описании, но при этом можно подключать их не подряд (то есть между любыми из них подключить какой-то ещё заголовочник).

Имена некоторых функций совпадают с именами команд командной строки; именно так обстоят дела, например, с хорошо знакомой нам функцией `printf`. Командный интерпретатор поддерживает одноимённую встроенную команду, которая несколько напоминает функцию Си по своему принципу работы; если мы напишем

```
man printf
```

— то увидим описание *команды* ***`printf`***, а не функции, и это совсем не то, что нам нужно. С этой проблемой легко справиться, если вспомнить, что тексты, выдаваемые командой `man`, сгруппированы в так называемые *секции*, каждая из которых имеет свой номер; мы обсуждали это во вводной части первого тома, см. §1.2.14. Сейчас из всех секций справочника нас интересуют две: секция №3, посвящённая функциям стандартной библиотеки языка Си, и секция №2, которая содержит описания системных вызовов — точнее, даже не их самих, а их «обёрток» в виде функций Си. Получить описание именно библиотечной функции, а не чего-то другого, можно, указав номер секции в явном виде:

```
man 3 printf
```

Если в этой секции нужного вам описания не найдётся, можно попробовать секцию 2 на случай, если функция, которую вы ищете, окажется системным вызовом:

```
man 2 close
```

Таким же точно образом — прочитав текст, выдаваемый командой `man`, — можно узнать и о том, что некая функция находится в библиотеке, требующей явного подключения; например, man-страницы математических функций, в том числе `sqrt`, упоминают флаг `-lm`.

4.3. Базовые средства языка Си

Предыдущая глава позволила нам получить первое впечатление о том, что представляет собой программирование на Си; теперь приступим к его систематическому изучению. При рассказе о Паскале мы вводили его средства постепенно, приводя примеры и решая задачи на каждую описанную возможность. С языком Си мы поступим проще, по крайней мере в самом начале: просто расскажем, что в нём есть.

4.3.1. Объявления и описания функций

Программа на языке Си состоит из отдельных ***единиц трансляции*** — модулей (файлов), каждый из которых обрабатывается компилятором отдельно от других, т. е. компилятор, обрабатывая один модуль, ничего не знает о содержимом других модулей. Результатом трансляции модуля становится файл с объектным кодом; набор таких файлов связывается в готовую программу с помощью системного редактора связей (линкера).

Отдельно взятый файл, написанный на языке Си, состоит из ***глобальных объявлений*** и ***глобальных описаний***; и те, и другие вводят (*объявляют* и *описывают*) глобальные имена — это могут быть имена функций, имена глобальных переменных и констант, имена типов. Слово «глобальный» соответствует области видимости имени и означает, что такое имя видно *как минимум* от его объявления или описания до конца модуля, а в некоторых случаях может быть доступно и в других модулях.

Разница между объявлением и описанием состоит в том, что *объявление* лишь сообщает компилятору, что объект, соответствующий данному имени, *существует где-то* (возможно, в том же модуле, а возможно, что и в другом), тогда как *описание* даёт исчерпывающую информацию, связанную с именем, и предписывает компилятору создать соответствующий объект здесь и сейчас, расположив его в генерируемом объектном коде.

В частности, мы уже видели ***описания функций***, которые состоят из ***заголовка функции*** и ***тела функции***; при этом в заголовке записывается[8] тип значения, которое возвращает функция, затем имя функции и список её параметров, заключённый в круглые скобки, а тело функции, заключённое в фигурные скобки, состоит из необязательной секции описаний локальных имён и последовательности операторов. В предыдущем параграфе мы уже дважды описывали функцию с именем `main` и, кроме того, на стр. 12 была описана функция для вычисления дискриминанта квадратного уравнения. Приведём ещё один пример описания функции:

```
char case_up(char c)
{
    if(c >= 'a' && c <= 'z')
        return c - ('a' - 'A');
    else
        return c;
}
```

Эта функция принимает в качестве параметра код символа, определяет, не соответствует ли этот код строчной латинской букве, и если это так, то возвращает код соответствующей заглавной буквы, в противном случае возвращает параметр без изменения. Здесь мы пользуемся известным фактом, что в таблице ASCII коды заглавных латинских букв идут подряд в соответствии с их алфавитным порядком и то же самое можно сказать о строчных буквах. Для нашего примера всё это неважно, нам просто нужна была какая-нибудь функция.

Обратите внимание, что здесь присутствуют все части описания функции — заголовок, состоящий из типа возвращаемого значения `char`, имени `case_up` и списка формальных параметров `(char c)`, а также тело, состоящее из оператора `if`. Увидев такое описание, компилятор немедленно преобразует тело в объектный код, который поместит в *секцию кода* в формируемом объектном модуле; иначе говоря, компилятор выполняет наше предписание о создании функции `case_up` здесь и сейчас.

Мы могли бы ввести функцию `case_up` иначе — написав только её заголовок, а вместо тела поставив точку с запятой:

```
char case_up(char c);
```

Это уже не описание, а всего лишь *объявление* функции. Увидев его, компилятор не станет формировать какой бы то ни было объектный код и где-то его располагать, тем более что у него для этого просто нет нужной информации — ведь мы же не показали компилятору тело функции, то есть фактически саму функцию. Компилятор просто *примет к сведению*, что *где-то* есть функция с именем `case_up`, принимающая на вход один параметр типа `char` и возвращающая тоже `char`; это означает, что, встретив *вызов* этой функции, компилятор сверит типы параметра и возвращаемого значения в контексте такого вызова с имеющейся информацией и, если всё в порядке, успешно откомпилирует такой вызов. Что касается самой функции, то она может появиться позже в том же модуле, а может не появиться вообще — компилятор в этом случае будет предполагать, что функция находится в другой единице трансляции, а заботу о её поиске оставит редактору связей.

Раз уж речь зашла о функциях, стоит сказать, что, хотя в языке Си нет отдельного понятия для *процедур*, некий аналог паскалевской процедуры можно получить, заявив, что функция ничего не возвращает. Это делается указанием ключевого слова `void` вместо типа функции. Пусть, например, нам нужна подпрограмма для печати заданного символа заданное число раз. В Паскале мы бы, разумеется, оформили такую подпрограмму в виде процедуры, но в Си процедур вроде бы нет; однако это не отменяет того факта, что нам попросту *нечего* возвращать из функции, имеющей такое предназначение, она всегда будет вызываться только ради побочного эффекта. Этот факт мы можем отразить в нашей программе, указав, что функция, хоть и называется функцией, всё-таки никаких значений не возвращает:

```
void print_n_chars(char c, int n)
{
    int k;
    for(k = 0; k < n; k++)
        printf("%c", c);
}
```

Как вы уже, несомненно, догадались, `%c` (от слова *char*) в форматной строке `printf` означает печать символа с заданным кодом; код в данном случае задаётся параметром `c`. С циклом `for` всё несколько сложнее; здесь переменная `k` будет пробегать значения от `0` до `n-1` включительно, примерно как если бы мы на Паскале написали что-то вроде

```
for k := 0 to n - 1 do
```

Полностью с возможностями `for` мы ознакомимся в §4.3.6.

4.3.2. Переменные и их описание

Объявлять и описывать мы можем не только функции. Ранее в примерах мы использовали переменные, вводя их внутри тел функций; это тоже были *описания* — в данном случае *описания локальных переменных*. Если поместить описание переменной *вне тел функций*, например:

```
int dangerous_global_variable;
```

то такая переменная будет *глобальной*, то есть будет видна от места своего описания до конца единицы трансляции, и доступ к ней можно осуществлять из любой функции; больше того, такую переменную можно в принципе «достать» и из других модулей. Си позволяет не только описывать, но и *объявлять* глобальные переменные, а также ограничивать видимость переменных (и функций тоже) текущим модулем, но об этом речь пойдёт в главе, посвящённой раздельной трансляции программ. Пока же считаем уместным напомнить, что **использовать глобальные переменные крайне нежелательно** (см. т. 1, §2.3.5).

Коль скоро речь зашла об описаниях переменных, отметим ещё одну интересную возможность. При описании переменной можно в явном виде задать её начальное значение, например, так:

```
int dangerous_global_variable = 42;


int f()
{
    int local_var = 42;
    /* ... */
```

Такая конструкция называется ***инициализацией***, а выражение справа от знака равенства — ***инициализатором***. Подчеркнём, что **инициализация не имеет ничего общего с присваиванием**, хотя и обозначается тем же знаком. Присваивание — это *деструктивная* операция, она разрушает имеющееся значение, записывая вместо него новое; инициализация ничего не разрушает, ведь переменной до момента её описания просто не было, и, значит, не

[8] В некоторых экзотических случаях тип возвращаемого значения приходится задавать частично перед именем функции, частично — наоборот, уже после; мы столкнёмся с подобной ситуацией ближе к концу этой части книги. Отметим, что подобной экзотики всегда можно избежать.

было значения переменной. Как мы увидим позже, с помощью инициализаторов можно сделать гораздо больше, чем с помощью присваиваний: например, массивы присваивать нельзя, а инициализировать можно. Напомним, что в Паскале инициализация и присваивание обозначаются *разными* знаками; то, что в Си знак один и тот же, не должно сбивать вас с толку.

Если в описании переменной нет инициализатора, то её исходное значение зависит от того, локальная она или глобальная[9]. Глобальные переменные, для которых начальное значение не задано, заполняются нулями; локальные переменные в отсутствие инициализаторов никаких начальных значений не получают, то есть в локальной переменной может оказаться абсолютно произвольный мусор.

Припоминая сведения из части, посвящённой языку ассемблера, поясним, что глобальные переменные, для которых задано начальное значение, компилятор помещает в секцию `.data`, а неинициализированные — в секцию `.bss` (см. т. 1, §3.2.2); при загрузке исполняемого файла в память операционная система формирует из этих секций ***сегмент данных*** и заполняет нулями всю область, соответствующую секции `.bss`, чем и объясняются начальные нулевые значения для неинициализированных глобальных переменных. Что касается локальных переменных, то они располагаются *в стековых фреймах* (см. §3.3.6), которые, как мы помним, создаются при вызове функции и уничтожаются при её завершении. Если очистка (принудительное заполнение нулями) секции `.bss` происходит один раз при загрузке программы и много времени не отнимает (на фоне времени, которое тратится на чтение в память машинного кода из исполняемого файла, на зануление `.bss` можно вообще не обращать внимания), то очистка стековых фреймов, если бы кто-то решил её сделать, потребовала бы лишнего цикла (записи нулей) *при каждом вызове функции*; это происходило бы миллионы, миллиарды, триллионы раз за время выполнения программы, и эффективность пострадала бы весьма заметно. Поэтому стековые фреймы никто не очищает, и в ячейках памяти, отведённых под очередной фрейм, остаются те значения, которые там были (остались от выполнения ранее вызывавшихся функций). Это и есть мусор, который мы наблюдаем в неинициализированных локальных переменных. Использование значения такой переменной до того, как в неё будет какое-то значение занесено — грубейшая ошибка, которую к тому же в общем случае не может «поймать» компилятор, хотя если он всё же смог её обнаружить, он выдаст вам предупреждение.

Отметим, что инициализация глобальных переменных ничего нам не стоит, кроме нескольких байтов в исполняемом файле, поскольку образ секции `.data`, содержащий все эти начальные значения, целиком размещается в исполняемом файле, при старте программы одним махом копируется в оперативную память, так что начальные значения глобальных переменных оказываются ровно там, где нужно; программа не тратит никакого времени на их инициализацию. В противоположность этому, инициализация локальных переменных на уровне машинного кода реализована точно так же, как обыкновенное присваивание, так что если вы, например, в самом начале функции присваиваете локальной переменной некое значение, задавать для этой переменной инициализатор не нужно: экономия двух-трёх тактов процессорного времени, конечно, не слишком серьёзна, но и терять эти такты на пустом месте не стоит.

4.3.3. Встроенные типы

Надо сказать, что по сравнению с Паскалем набор встроенных (базовых) типов языка Си может показаться весьма компактным и даже аскетичным. Фактически здесь есть только целые числа (различающиеся разрядностью и знаковостью) и числа с плавающей точкой (различающиеся разрядностью). И на этом, собственно, всё: нет ни специального типа для символов, ни логического типа, ни строкового.

Достигнуто это довольно просто. Программируя на языке ассемблера, мы на собственном опыте убедились, что для чисел с плавающей точкой у нас есть отдельный «сопроцессор» со своими регистрами, тогда как всё остальное — и символы, и логические значения, и вообще всё, с чем нам взбредёт в голову работать, процессор «перемалывает», используя свои основные регистры, которые хранят не что иное, как *целые числа*. Иначе говоря, если то, с чем мы работаем, не является числом с плавающей точкой (или, возможно, несколькими такими числами), то оно является (на уровне машинных команд) целым числом или набором целых чисел. При создании языка Си никому не пришло в голову этот факт каким-либо образом маскировать, поэтому, в частности, **в качестве логического значения в Си обычно используются целые числа**, причём `0` означает «ложь», а всё остальное — «истину»[10]. Так, если у нас есть целочисленная переменная `a`, то вместо `if(a != 0)` мы могли бы написать просто `if(a)` — работать это будет точно так же.

Как мы знаем, ***символ*** представляется в памяти машины своим ***кодом***, то есть, опять же, целым числом. Как несложно догадаться, авторы Си не стали проводить привычное нам по Паскалю различие между самим символом и его кодом, то есть в Си **символ и его код — это одно и то же**. Для обозначения символов (в отличие от строк!) в Си используются одиночные апострофы, например так: `'a'`; но это для компилятора *абсолютно то же самое*[11], как если бы мы написали просто `97` (именно таков код буквы `a` в таблице ASCII). Мало того, в Си совершенно легитимным оказывается выражение `'a'+'b'`, оно равно сумме кодов букв `a` и `b` (то есть числу `195`), хотя, конечно, смысла в таком выражении никакого нет и писать так не следует.

Начнём с рассмотрения целочисленных типов. Разрядность (то есть, грубо говоря, количество памяти, отводимое под соответствующее значение) этих типов варьируется в довольно широких пределах; для обозначения разрядности используются ключевые слова `char`, `short`, `int` и `long`, а наибольшей возможной разрядности можно достичь, используя тип `long long` («длинное длинное»).

Наименьшую разрядность имеет тип, который называется `char`; в большинстве случаев он имеет размер 1 байт (то есть 8 бит), хотя теоретически возможны аппаратные платформы, где у этого типа разрядность окажется какая-то другая. Строго говоря, тип `char` на любой платформе равен по размеру *минимальному адресуемому*[12], то есть минимальной области памяти, у которой имеется свой собственный адрес. Мы знаем, что ячейка памяти на всех современных машинах (как и на большинстве не очень современных) состоит из восьми бит, так что можем на эту тему особенно не беспокоиться. Интереснее другое: тип `char` может оказаться как знаковым, так и беззнаковым, то есть, грубо говоря, если мы опишем переменную этого типа, то она может оказаться способна принимать значения от 0 до 255 либо от -128 до 127, и зависит это исключительно от компилятора. Впрочем, язык позволяет взять этот момент под контроль, явным образом указав ***знаковость***, для чего предусмотрены слова `signed` и `unsigned`[13]; с учётом этих слов у нас появляются *ещё два типа* — `signed char` (знаковый) и `unsigned char` (беззнаковый), что же касается обычного `char`, то он совпадает с одним из этих двух, но вот с каким именно — неизвестно.

Чтобы понять причину столь странной ситуации, следует вспомнить, что тип `char`, как это следует из его названия, исходно предназначался для хранения кода символа, а таблица ASCII содержит всего 128 позиций, от 0 до 127. В те времена и в том месте, когда и где язык Си обретал более-менее устойчивые очертания, никому не было никакого дела до кодирования символов, не попавших в ASCII, так что «расширенных ASCII-таблиц» вроде koi8-r или cp1251 не было и в проекте, ну а современные многобайтовые кодировки, основанные на Unicode, не могли бы присниться создателям Си даже в самом страшном ночном кошмаре. С учётом этого внезапно оказывается, что знаковость типа `char` не так уж важна: все коды ASCII-таблицы благополучно покрываются обоими вариантами.

Сразу же отметим, что такая ситуация наблюдается *только с типом* ***`char`***, а все остальные целочисленные типы по умолчанию знаковые и становятся беззнаковыми, только если написать слово `unsigned`; применять с ними слово `signed` можно, но бессмысленно, то есть фактически это слово нужно лишь затем, чтобы гарантированно сделать знаковым этот вот «хитрый» `char`. То, что в Си аж целое ключевое слово введено для столь несуразной цели, можно рассматривать как очень характерный пример на тему общей неряшливости построения этого языка; впрочем, это далеко не самый убедительный пример, в дальнейшем мы увидим много выкрутасов подобного рода. Так или иначе, как уже говорилось, мы язык Си любим не за это.

Следующим по разрядности идёт тип `short` и его беззнаковый аналог `unsigned short`. Как водится, требования к языку Си не задают конкретной разрядности для этого типа, известно только, что он не имеет права быть меньше `char`'а (впрочем, трудно быть меньше, нежели минимальное адресуемое); однако есть и хорошая новость: вы можете считать, что переменная

[9] Из этого правила есть одно важное исключение — так называемые локальные статические переменные, которые будут рассмотрены позднее.

[10] Забегая вперёд, отметим, что в качестве логического значения можно также использовать адресное выражение; нулевой указатель в этом случае будет обозначать «ложь», а любой другой — «истину».

[11] Для наиболее «продвинутых» читателей сообщим, что выражение `'a'` даже имеет тип `int`, а не `char`, несмотря на наличие в Си такого типа.

[12] Заметим, что Керниган и Ритчи в своей книге вообще не упоминают никакие «минимальные адресуемые», а говорят просто «байт»; всё остальное — происки комитетов.

[13] Читается приблизительно как «сайнт» и «ансайнт».

этого типа занимает два байта, и это окажется верно на любой машине, с которой вам доведётся столкнуться в реальной жизни; хотя, конечно, стоит помнить, что *теоретически* возможно и иное. Как следствие, если вы пишете программу не «для себя», а для заказчика, лучше всё-таки не делать в ней подобных предположений.

Отметим ещё один интересный казус: исходно предполагалось применять слово `short` в качестве *модификатора*, то есть «прилагательного» к слову `int` подобно словам `signed` и `unsigned`; так можно делать и до сих пор, то есть можно описать переменную, к примеру, вот так:

```
unsigned short int x;
```

От этого стиля в основном отказались из-за довольно простого соображения: простите, а ***что ещё*** тут может быть «коротким»? Оказалось, что ничего. Поэтому сейчас слово `int` можно (и нужно) опускать.

После `short`'ов идёт уже знакомый нам `int` в компании с `unsigned int`'ом. С его разрядностью дела обстоят, пожалуй, самым забавным образом. На старых 16-битных платформах (например, в MS-DOS) `int` был 16-битным, то есть совпадал по своей разрядности с `short`'ом. При переходе к 32-битным архитектурам тип `int` вырос до четырёх байт и стал совпадать с типом `long`; бытовало даже мнение, что тип `int` на любой платформе должен соответствовать размеру машинного слова. Такое предположение не лишено оснований, ведь согласно традициям построения компиляторов Си, например, возврат значения из функции производится через регистр (обычно тот, который по какой-то причине можно считать «главным»; в случае i386 это регистр `EAX`; его часто называют «аккумулятором»), если только возвращаемое значение в этот регистр помещается; при этом до не слишком давнего времени считалось, что если для функции не указан тип возвращаемого значения, то она возвращает `int` (сейчас это тоже так, но компиляторы выдают предупреждение).

Несмотря на это, при переходе от 32-битных архитектур к 64-битным `int` так и остался четырёхбайтным, оно и понятно: если бы его «вырастили» вслед за разрядностью регистров, то целого типа размером в четыре байта в языке бы просто не осталось; можно было бы таким сделать `short`, но тогда не стало бы двухбайтных. Зато на 64-битных платформах «подрос» следующий тип, `long`; он стал 64-битным, в результате чего *совпадает с типом* ***`long long`***. Но здесь мы уже забежали вперёд.

Очередная пара типов одинаковой разрядности — это `long` и `unsigned long`; до наступления эры 64-битных платформ эти типы всегда были четырёхбайтными, а с наступлением этой эры внезапно выросли. Как и слово `short`, изначально слово `long` планировалось использовать как прилагательное, а переменные описывались как-то вроде `long int x`. В отличие от слова `short`, слово `long` имеет ещё одно применение, о котором мы узнаем из обсуждения типов с плавающей точкой; тем не менее, для лучшего единообразия слово `int` разрешили опускать для случая «длинных» точно так же, как и для «коротких».

Пара самых больших целочисленных типов, обозначаемых как «длинное длинное»[14], то есть, соответственно, `long long` и `unsigned long long`, поддерживается не всеми компиляторами и не на всех платформах; но везде, где эти типы присутствуют, их разрядность составляет 64 бита. Даже при появлении 64-битных процессоров `long long` в размерах не вырос.

Отметим ещё один момент: действует соглашение, по которому разрядности числа типа `long` должно на любой архитектуре хватать для представления указателей, хотя это соглашение не является частью официально утверждённых стандартов языка Си.

Говоря о размерах, следует всегда помнить, что любые разговоры о конкретной разрядности целочисленных типов языка Си возможны только с констатационной позиции, то есть по принципу «существующее положение вещей таково», с обязательным учётом того обстоятельства, что завтра может появиться или новая платформа, или новый компилятор под существующую платформу, и там всё будет не так. Гарантировать можно только одно: `short` не будет короче `char`'а, `int` не окажется короче `short`'а, `long` ни за что не станет короче `int`'а, ну и, конечно, `long long` не может стать короче, чем простой `long`. Но и только.

Система типов с плавающей точкой несколько проще, этих типов всего три: `float`, `double` и `long double`, причём беззнаковыми они не бывают. На платформе i386, как правило, `float` и `double` соответствуют числам обычной и двойной точности из стандарта IEEE-754, а тип `long double` предоставляет доступ к числам *повышенной точности* (напомним, таковые занимают 10 байт); при этом в памяти переменная типа `long double` может занимать 12 или даже 16 байт. Если же говорить не о конкретной платформе, а «вообще», то сказать тут можно только одно: `double` не может быть короче, чем `float`, или длиннее, чем `long double`; при этом, например, все три типа могут между собой совпадать, что достаточно часто встречается, когда в процессоре отсутствует аппаратная поддержка арифметики с плавающей точкой.

Отметим ещё один момент: математические функции, такие как уже знакомый нам `sqrt`, а также всевозможные синусы, косинусы, экспоненты и логарифмы по умолчанию работают с типом `double`; больше того, для любых вычислений и даже для простой передачи параметров значения типа `float` всегда преобразуются к типу `double`, что делает использование типа `float` бессмысленным во всех случаях кроме одного: когда таких чисел очень много, а памяти не хватает (а уменьшение её расхода вдвое способно решить проблему). Иначе говоря, пока вы не встретились с нехваткой памяти, `float` применять не надо.

Язык Си предусматривает специальную возможность для определения размеров любых типов — ключевое слово `sizeof`. Выражение вида `sizeof(`*<тип>*`)` компилятор заменяет на целое число, равное размеру данного типа, причём *единицей измерения является* ***`char`***. Выражение `sizeof(char)` по определению равно единице. Кроме имён типов, `sizeof` можно применять также к именам переменных и вообще произвольным выражениям, тип которых компилятор может определить. Например, `sizeof(1)` есть то же самое, что `sizeof(int)`.

Более того, синтаксически `sizeof`, формально говоря, является операцией, а не функцией, так что, применяя его к выражениям (но не к типам), можно не писать скобки; впрочем, ясности программе это не добавит, так что скобки лучше всё-таки поставить.

4.3.4. Литералы (константы) разных типов

Литералами обычно называют значения, записанные в программе в явном виде; простейший пример литерала — целое число, записанное как оно есть.

Язык Си позволяет записывать целые числа в десятичной, шестнадцатеричной и восьмеричной системах счисления. Если написать число обычными арабскими цифрами *без лидирующего нуля*, компилятор будет рассматривать его как десятичное; если же мы *допишем спереди ноль*, это число будет рассматриваться как записанное в *восьмеричной* системе. Например, `035` означает то же, что и 29 ($3 \times 8 + 5 = 29$)[15]. Для записи шестнадцатеричных чисел используется префикс «`0x`», например, всё то же десятичное 29 может быть записано как `0x1d` или `0x1D`, а запись `0x7E5` означает 2021.

Довольно нетривиальным образом обстоят дела с разрядностью и знаковостью целочисленных литералов. Десятичные литералы *в большинстве случаев* рассматриваются как знаковые, тогда как восьмеричные и шестнадцатеричные — всегда беззнаковые; но это только начало истории.

Если десятичный литерал представляет число, «умещающееся» в разрядность числа типа `int`, то он будет считаться имеющим тип `int`; если он перестаёт умещаться в `int`, но пока что помещается в `unsigned int`, то он будет считаться числом типа `unsigned int`. Если для записанного числа не хватает разрядности `unsigned int`, компилятор попытается представить его как `long`, затем как `unsigned long`, как `long long` и как `unsigned long long`, но если и тут не хватит разрядности, то будет наконец-то выдана ошибка.

С литералами восьмеричными и шестнадцатеричными ситуация примерно такая же, только компилятор не пытается делать их знаковыми. Иначе говоря, если хватает разрядности `unsigned int`'а, то он и используется, если же не хватает, компилятор пытается использовать `unsigned long` и `unsigned long long`.

Кроме того, программист может сам явным образом задать тип литерала, добавив в его конец комбинацию из суффиксов `L` (`long`), `LL` (`long long`)[16] и `U` (`unsigned`): например, `15` будет иметь тип «знаковый `int`», тогда как `15ULL` будет

[14] Следует отметить, что адекватное звучание этого термина на русском языке достигается не тогда, когда мы представляем себе что-то вроде *длинное-длинное*, то есть *длинное-предлинное*, а когда мы первое слово воспринимаем как прилагательное, а второе — как существительное; то есть у нас имеется такая сущность «длинное» (в данном случае это существительное), а потом мы это «длинное» ещё сильнее удлинили, и получилось у нас не *простое* длинное, а вот это вот *длинное* длинное. Англоязычным людям в этом плане проще, там одна и та же словоформа почти всегда может использоваться и как существительное, и как прилагательное, и как глагол.

[15] На самом деле это не совсем одно и то же, поскольку десятичные константы по умолчанию знаковые, тогда как восьмеричные и шестнадцатеричные — беззнаковые, но в реальной жизни эта разница обычно не играет роли.

[16] В пресловутые стандарты Си суффикс `LL` вошёл только начиная с C99, но реально этот суффикс поддерживался везде, где присутствовал тип `long long`.

типа `unsigned long long`. В принципе, эти суффиксы можно записывать также и в нижнем регистре (что-нибудь вроде `200ull`), но делать так не рекомендуется, потому что букву `l` очень легко спутать с единицей.

Вопреки распространённому заблуждению, литералов типа `char` в языке Си нет; литералы вида `'a'`, `'F'` и т. п. имеют тип `int`. Литералы типа `short` и `unsigned short` также не предусмотрены.

Числа с плавающей точкой традиционно записываются в виде литералов в десятичной системе счисления. В общем случае такой литерал состоит из целой части, точки, дробной части и *порядка*, причём порядок не обязателен, можно указать одну только целую или одну только дробную часть, если же порядок присутствует, то можно опустить дробную часть вместе с точкой. Порядок записывается, как обычно, в виде буквы `e` или `E`, за которой следует десятичное целое число (возможно, снабжённое знаком), и означает, что исходное число следует домножить на 10^k, где k — число, указанное в порядке. Например, следующие литералы обозначают одно и то же число:

```
1500.0  1500.  15E2  15.E2  15E+2  0.15e4  .15e4  15000e-1
```

По умолчанию такие литералы имеют тип `double`, но вы можете принудительно превратить их во `float`'ы добавлением суффикса `f` или `F`, а также в `long double`'ы добавлением уже знакомого нам суффикса `L` (можно и `l`, но не надо).

Вообще говоря, все эти возможности нужны достаточно редко: для реализации математических вычислений есть более удобные языки программирования, нежели Си.

Как уже говорилось, в языке Си *символ* и *код символа* — это одно и то же, а для записи кодов символов используются ***символьные литералы***, формируемые с помощью апострофов, причём они имеют тип `int`. В частности, `'A'` — это то же самое, что и `65`, а `'1'` — то же, что `49`.

Для представления ***служебных символов*** предусмотрены так называемые ***escape-последовательности***, одну из которых — «`\n`» — мы уже много раз встречали. Поскольку код перевода строки — `10`, запись `'\n'` означает то же, что и `10`. Кроме перевода строки, предусмотрены также `'\r'` — возврат каретки (код `13`), `'\t'` — табуляция (код `9`), `'\a'` — «звонок» (от слова *alarm*, код `7`), `'\b'` — «забой» (от слова *backspace*, код `8`), `'\f'` — так называемый перевод страницы (*form feed*, код `12`) и совсем уже экзотическая `'\v'` — «вертикальная табуляция» (код `11`). Поскольку в escape-последовательностях используется символ обратного слэша «`\`», а сами последовательности применяются внутри символьных и строковых литералов, обрамляемых апострофами и двойными кавычками, предусмотрены вполне логичные способы представления этих трёх символов: `'\\'` обозначает код обратного слэша, `'\''` и `'\"'` — соответственно коды апострофа и кавычки; впрочем, код кавычки можно получить и проще: `'"'`.

Для представления ***строк*** используются ***строковые литералы***, обрамляемые двойными кавычками, и все перечисленные escape-последовательности действуют в них так же, как в символьных литералах, за исключением того, что символ апострофа здесь можно записать как есть (без escape-последовательности), тогда как кавычку, если она потребовалась вам в строке, придётся обязательно снабдить символом слэша, например: `"Never say \"never\""`.

Очень важно уяснить разницу между символьными литералами (в апострофах) и строковыми (в кавычках)! Например, когда компилятор видит литерал `'M'`, он заменяет его на *целое число* 77 (код буквы `M`), тогда как если компилятор увидит литерал `"M"`, он сделает совершенно иное: расположит где-то (на самом деле — в секции кода) массив из двух элементов типа `char` с первым элементом, равным 77, а вторым — равным нулю, при этом сам литерал компилятор заменит на *адрес* той области памяти, где расположен массив. Если это оказалось непонятно, ничего страшного, разговор о массивах и строках у нас ещё впереди; пока просто запомните, что **строки и символы — это совершенно разные сущности**.

Кроме вышеперечисленных escape-последовательностей, язык Си предусматривает возможность задать символ его кодом в восьмеричной или шестнадцатеричной системе; например, `'\115'` и `'\x4d'` означают то же, что и `'M'`. Если такая последовательность встречается в строковом литерале, то для восьмеричного варианта используется от одного до трёх символов, пригодных в качестве восьмеричной цифры (то есть, например, литералы `"a\11522"` и `"aM22"` эквивалентны, как и литералы `"\tX"` и `"\11X"`); для шестнадцатеричного компилятор пытается «съесть» все символы, пригодные в роли шестнадцатеричной цифры, сколько бы их ни было, и выдаёт ошибку, если их слишком много.

В программах на Си часто можно встретить запись `'\0'` для обозначения сущности «символ с кодом ноль». Семантически это, разумеется, абсолютно то же самое, что и просто `0`, но со стилистической точки зрения такая запись имеет смысл — читателю программы сразу станет понятно, что имеется в виду именно символ, а не просто арифметический ноль.

Отметим ещё одну удобную возможность, связанную со строковыми литералами. Если несколько таких литералов записать один за другим, разделив их только пробельными символами, компилятор «склеит» их все в одну строку. Например, следующий фрагмент

```
"This " "is" " a string" " whi" "ch is" " long"
```

есть то же самое, что и

```
"This is a string which is long"
```

Это бывает полезно, если в программе нужна настолько длинная строковая константа, что она не влезает на экран (точнее, в 80 колонок; см. т. 1, §2.12.7) по ширине.

4.3.5. Операции и выражения

Арифметические выражения в языке Си играют существенно более важную роль, чем в привычном нам Паскале. Как мы увидим позже, в Си присутствует такая нетривиальная вещь, как *адресная арифметика*; кроме того, столь привычное нам *присваивание*, которое в Паскале было *оператором*, в Си совершенно неожиданно оказывается *операцией* и может, как следствие, быть частью арифметического выражения. Но — обо всём по порядку.

Арифметика, логика, биты

Начнём мы с того, что над всеми имеющимися в языке встроенными типами (а они, как мы помним, все числовые) определены обычные и ничем особенным не примечательные сложение «`+`», вычитание «`-`» и умножение «`*`». Разумеется, присутствует также и деление, обозначаемое вполне ожидаемой на эту роль дробной чертой «`/`», но здесь уже потребуются пояснения. Дело в том, что семантика операции деления *зависит от типа операндов*: если операнды представляют собой числа с плавающей точкой, то это будет обычное арифметическое деление, тогда как если операнды целые — то деление будет *целочисленным*, подобно паскалевскому `div`. Так, значение выражения `32.0/5.0` *приблизительно* равно `6.4`, тогда как значение выражения `32/5` равно `6`, притом безо всяких «приблизительно» (операции над целыми, как мы знаем, выполняются точно). *Взятие остатка от деления* в языке Си обозначается символом процента «`%`».

Здесь у читателя может возникнуть целый ряд вопросов. Что будет, например, если у операции деления один операнд целый, а другой «плавучий»? Как работает взятие остатка, если один или оба операнда отрицательны? Можно ли применять взятие остатка к дробным числам?

На все эти вопросы можно дать вполне однозначные ответы, однако делать этого мы не будем: не надо засорять мозг информацией подобного свойства. Если вопросы такого рода возникают из чистого любопытства, можно взять любую более-менее серьёзную книжку по Си, можно поискать (и практически мгновенно найти) ответы в Интернете. Важнее другое: *в процессе практического программирования такие вопросы не должны возникать*. В частности, применять деление к операндам разных типов *не надо* — операнды следует привести к одному типу или хотя бы сделать их или оба целыми, или оба «плавающими», а брать остаток от деления следует *только при целых положительных операндах*; если вам кажется, что возникла потребность в ином, нужно проанализировать, откуда такая мысль пришла в голову: знайте, что-то тут пошло не так, в нормальных условиях такое понадобиться не должно.

Набор *операций сравнения* в Си вполне привычен; порядковые операции записываются так же, как в Паскале: `<`, `>`, `<=`, `>=`, что же касается сравнения на равенство и неравенство, то здесь они пишутся иначе: «`==`» (*двойной* знак равенства) означает «равно», а комбинация «`!=`» — «не равно». **Обратите особое внимание на удвоенный знак равенства!** Дело в том, что одиночный знак равенства, как мы уже видели, в Си означает присваивание, причём выше уже упоминалось, что присваивание может встречаться в выражениях. Если, к примеру, вы напишете `if(a = 3)`, программа успешно откомпилируется и запустится, но работать будет совсем не так, как если бы вы написали `if(a == 3)`. Хорошая новость для начинающих состоит здесь в том, что если вы не забываете про флажок `-Wall` при запуске компилятора, то программа, конечно, всё равно откомпилируется, но вы хотя бы получите

предупреждение; плохая новость для лентяев — без флага `-Wall` компилятор просто молча проглотит конструкцию, в большинстве случаев ошибочную.

Если вы *действительно* решите использовать значение операции присваивания в качестве логического — что теоретически может быть осмысленно, если в правой части не константа, а какое-то выражение — то, чтобы успокоить компилятор и избавиться от предупреждения, следует это присваивание заключить в дополнительные круглые скобки, примерно так:

```
if((a = f(x))) {
```

Не надо этого делать, напишите лучше так:

```
a = f(x);
if(a) {
```

Существует целый ряд ситуаций, когда присваивание в роли логического значения оказывается допустимо и даже правильно, но все они имеют отношение к циклам, а отнюдь не к оператору `if`.

Результатом операции сравнения становится *целое число*, причём, как уже говорилось, `0` означает «ложь»; в качестве представления «истины» эти операции всегда возвращают единицу, что позволяет, например, написать вот такую вот белиберду:

```
x = (a < b) + (a > c) - (c != t);
```

Так вот, *не надо так делать*, даже если вы найдёте ситуацию, в которой подобная конструкция вам покажется осмысленной. При чтении программы подобные трюки приходится подолгу анализировать, чтобы понять, что́ имелось в виду.

Введя логические отношения, мы естественным образом приходим к вопросу о ***логических операциях***. В Си они представлены привычным набором «конъюнкция, дизъюнкция, отрицание»: операция «логическое и» записывается знаком «`&&`» (два амперсанда), «логическое или» — знаком «`||`» (две вертикальные черты), унарная операция отрицания обозначается восклицательным знаком «`!`». Сразу же отметим одно интересное свойство бинарных логических операций: **если значение выражения уже определено, второй операнд не вычисляется**. Иначе говоря, если в программе имеется выражение `f() || g()` и функция `f` вернула значение, отличное от нуля (т. е. «истину»), функция `g` вообще не будет вызвана, что особенно важно учитывать, если у этой функции имеется побочный эффект. Аналогично при вычислении выражения `f() && g()` функция `g` не будет вычисляться, если `f` вернула ноль.

Проницательный читатель может спросить, почему используется *два* амперсанда, а не один, и *две* «палочки», а не одна. Дело в том, что *одним* амперсандом обозначается операция «и», но не логическая, а *побитовая*, то есть эта операция производится по отдельности над первыми битами обоих операндов, над вторыми, над третьими и т. д., и результаты коньюнкции отдельных битов образуют результат всей операции; например, значением выражения `25 & 62` будет `24` (чтобы понять, почему это так, переведите 25 и 62 в двоичную систему). Аналогичным образом одной «палочкой» обозначается побитовое «или», так что, например, значением `12 | 56` будет `60`.

В Паскале, напомним, ключевые слова `and` и `or` обозначают как логические, так и побитовые варианты соответствующих операций; компилятор Паскаля различает эти случаи по типам операндов: если это `boolean`'ы, то выполняется логическая операция, если целые числа любой разрядности — то побитовые. В Си так сделать нельзя, поскольку отдельного типа для обозначения логических значений тут нет.

Побитовое отрицание обозначается знаком «`~`» (тильда); например, `~0` будет равно `-1` (все биты — единицы). Кроме того, в Си присутствует операция побитового «исключающего или» (знакомый нам `XOR`), которая обозначается символом «`^`». Логического варианта для этой операции не предусмотрено.

Завершая разговор о побитовых операциях, опишем ***побитовые сдвиги*** влево (знак «`<<`») и вправо («`>>`»); левый операнд обеих операций — исходное сдвигаемое число, правый — количество битов, на которое следует произвести сдвиг. Сдвиг влево выполняется одинаково для всех целочисленных типов: биты сдвигаются влево на заданное число позиций, справа дописываются нули (как если бы число умножили на двойку в соответствующей степени). Со сдвигом вправо всё несколько сложнее, его выполнение зависит от *знаковости* сдвигаемого числа. Биты в представлении числа, естественно, в любом случае сдвигаются вправо на заданное число позиций, но слева (то есть в старшие разряды) для беззнакового числа записываются нули, тогда как для знакового числа — *значение старшего бита исходного числа*; знак числа при этом сохраняется, а сама операция всегда остаётся эквивалентной целочисленному делению на соответствующую степень двойки.

Поскольку с операциями побитовых сдвигов мы уже сталкивались при изучении языка ассемблера, отметим, что для беззнаковых чисел сдвиг выполняется командами `shl` и `shr`, а для знаковых — командами `sal` и `sar`; напомним, что в действительности `shl` и `sal` — это одна и та же команда, тогда как `shr` и `sar` — разные; сдвиг вправо, «размножающий» знаковый бит, называют ***арифметическим***.

Присваиваивания и модификации

Привычные операции на этом кончаются и начинается всевозможная экзотика. Одна из «фирменных» особенностей Си — ***операция присваивания*** (а не оператор присваивания, как в большинстве языков, которые вообще предусматривают присваивание). Присваивание обозначается, как мы уже видели, простым знаком равенства. Хитрость в том, что конструкция вида `a = b` представляет собой *выражение*, и это выражение имеет значение — оно равно тому значению, которое было только что присвоено, то есть занесено в переменную слева от присваивания. Например, в Си допустима такая конструкция:

```
x = (y = 17) + 1;
```

Здесь в переменную `y` заносится значение `17`, это же значение оказывается значением выражения `(y = 17)`, к нему прибавляется единица, получается `18`, что и заносится в `x`.

Другой классический пример присваивания в качестве операции выглядит так:

```
a = b = c = d = e + 5;
```

Здесь будет взято значение переменной `e`, к нему прибавят `5`, результат будет занесён в `d`, причём результатом выражения `d = e + 5` станет только что присвоенное выражение, его в результате занесут в `c`, результатом выражения `c = d = e + 5` станет всё то же значение, и оно таким же образом попадёт в `b`, потом в `a`.

Чтобы понять, почему такие странные возможности проникли в язык, нужно, во-первых, припомнить, что Си изначально придуман как заменитель языка ассемблера, и, во-вторых, принять во внимание, что оптимизаторов в те времена ещё не было. Очевидно, что если в программе встречается присваивание `x = y`, то на выходе сгенерируется что-то вроде

```
mov     eax, [y]
mov     [x], eax
```

В отсутствие оптимизатора каждый оператор транслируется в объектный код по заданной для этого оператора схеме без учёта соседних операторов. Например, для последовательности

```
a = x;
b = x;
c = x;
```

компилятор выдаст

```
mov     eax, [x]
mov     [a], eax
mov     eax, [x]
mov     [b], eax
mov     eax, [x]
mov     [c], eax
```

Очевидно, что две команды тут лишние, ведь присваиваемое значение уже лежит в `EAX`, и загружать его из `x` трижды — совершенно бессмысленно. В более общем смысле можно заметить, что *присвоенное значение остаётся в аккумуляторе*, и если следующая операция использует свежеприсвоенное значение (а так происходит довольно часто: следующая операция может, например, обращаться к только что присвоенной переменной), можно сэкономить объём машинного кода и время исполнения, если заставить компилятор использовать значение, которое уже имеется где надо, вместо того чтобы его туда заносить. Конструкция вида `a = b = c = x` позволяет сократить код:

```
mov     eax, [x]
mov     [c], eax
mov     [b], eax
mov     [a], eax
```

Конечно, в современных условиях это уже не столь важно, оптимизатор такое сокращение кода произведёт без каких-либо подсказок с нашей стороны, но присваивание в роли именно операции (а не оператора) открывает очень интересные возможности при записи циклов. Об этом речь пойдёт, когда мы доберёмся до управляющих конструкций языка.

Кроме операции простого присваивания, в Си предусмотрены операции присваивания переменной значения, которое получено из её же старого значения выполнением некоторого действия. Например, `a += 7` означает «присвоить переменной `a` её же значение, увеличенное на 7».

Часто можно встретить утверждение, что `a += b` — это «сокращённая запись» операции `a = a + b`, но это, вообще говоря, не так. Дело в том, что *переменная* (или, в терминологии, традиционной для языка Си, ***леводопустимое выражение***, то есть то, что стоит или может стоять слева от знака присваивания) зачастую принимает довольно сложные формы. Мы пока не рассматривали работу с массивами, но поскольку операция индексирования в Си выглядит очень похоже на аналогичную операцию Паскаля, следующий пример будет понять несложно. Итак, допустим, у нас есть массив `v` и некая функция `f`, возвращающая целое число. Рассмотрим две записи:

```
v[f(x)] += 10;
v[f(x)] = v[f(x)] + 10;
```

Ясно, что это совершенно не одно и то же, ведь в первом случае `f` будет вычислена один раз, тогда как во втором случае она будет вычисляться дважды. То обстоятельство, что двойное вычисление *неэффективно* (вхолостую расходуется процессорное время) — это ещё полбеды, на это при определённых обстоятельствах можно было бы не обратить внимания. Гораздо серьёзнее всё становится, если *функция `f` имеет побочные эффекты*; в качестве самого простого и безобидного примера такой ситуации можно назвать наличие внутри `f` операции вывода, например простой надписи на экран — в первом случае такая надпись появится один раз, во втором — дважды. А совсем всё станет странно, если окажется, что *возвращаемое значение функции `f` различается от раза к разу даже при одинаковом значении аргумента*; в этом случае значение будет извлечено из одного элемента массива, а занесено — совсем в другой, причём такой эффект оказывается полной неожиданностью для читателя программы (и чаще всего для её автора тоже).

Аналогично операции `+=` работают `-=`, `*=`, `/=`, `%=`, `&=`, `|=`, `^=`, `<<=` и `>>=`: в каждом случае из стоящей слева переменной извлекается значение, над этим значением и правым операндом производится операция, символ которой стоит перед знаком равенства, а результат операции заносится обратно в переменную. Чтобы понять, откуда в Си взялись эти операции, достаточно припомнить язык ассемблера и команды `add`, `sub`, `and`, `or` и т. п.

Слегка особняком в списке операций стоят ***инкремент*** и ***декремент***, которые обозначаются соответственно двумя плюсами и двумя минусами. Это унарные операции, обычно применяемые к целочисленным переменным и предписывающие увеличение (уменьшение) значения переменной на единицу; иначе говоря, `++i` означает «занести в переменную `i` значение, на единицу большее, чем то, что там сейчас», а `--i` — то же самое, только значение заносится на единицу *меньшее*. В такой форме эти операции означают то же самое, что и `i += 1` и `i -= 1`, но, как водится в языке Си, это только начало.

Продолжение состоит в данном случае в том, что у обеих операций наряду с обычной *префиксной* формой (то есть когда символ операции записывается перед переменной) присутствует ещё и *суффиксная форма*, то есть можно написать также `i++` и `i--`. В обоих случаях с переменной `i` происходит *ровно то же самое*, то есть она увеличивается или уменьшается на единицу; на этом месте обычно задают вопрос, зачем же, в таком случае, вообще введены эти суффиксные формы.

Чтобы понять, в чём разница между префиксной и суффиксной формой операций инкремента и декремента, нам придётся в очередной раз вспомнить, что Си — не Паскаль и традиции здесь совершенно другие. В Паскале аналог операций инкремента/декремента присутствует в виде *встроенных процедур* `inc` и `dec`, но, будучи процедурами, они только меняют значение переменной и больше ничего не делают. Иное дело Си, где инкремент и декремент — это *операции*, то есть они могут встречаться в выражениях и, естественно, имеют некоторое значение. Различие заключается как раз в этом значении: если значением операции в префиксной форме оказывается *новое* (уже увеличенное или уменьшенное) значение переменной, то постфиксная форма в качестве своего значения имеет *старое* значение переменной, то, которое было в переменной до выполнения операции.

Рассмотрим для примера два фрагмента:

```
x = 7;              p = 7;
y = ++x;            q = p++;
```

В результате выполнения первого фрагмента обе переменные `x` и `y` получат значение `8`, тогда как после выполнения второго фрагмента `p` тоже будет равна `8`, но `q` будет равна `7`. С декрементом ситуация полностью аналогична.

В системе команд PDP-11, с которой, как мы знаем, началась история ОС Unix и языка Си, присутствовала возможность при работе с косвенной адресацией произвести преинкремент, предекремент, постинкремент или постдекремент того регистра, из которого брался косвенный адрес. Это позволяет лучше понять, откуда взялись в Си операции `++` и `--`.

Ещё несколько операций

Теперь, пожалуй, настало время для настоящей экзотики. Начнём с замечания, что в Си арифметическими операциями считаются (и действительно являются) ***операция вызова функции***, обозначаемая круглыми скобками, и ***операция индексирования*** (то есть обращения к элементу массива), которая, как и в Паскале, обозначается скобками квадратными. Подробное рассмотрение операции индексирования отложим до параграфа, посвящённого массивам; про операцию вызова функции отметим, что первым её операндом является *имя функции* либо *адрес функции*[17], а её ***арность*** (т. е. количество операндов) может быть любым, но не меньше одного. Так, при вызове функции без параметров операция её вызова оказывается унарной, и выглядит это как имя функции, после которого указаны пустые скобки: `f()`. Подчеркнём, что, в отличие от Паскаля, в Си эти скобки обязательно указываются, даже когда параметров нет, ведь именно скобки обозначают для компилятора *вызов функции*; имя функции без скобок воспринимается компилятором как *адрес* функции, а не её вызов.

При вызове функции с параметрами арность операции вызова составит количество параметров функции плюс один; например, при вызове функции от трёх параметров арность операции вызова — четыре: для выражения `f(a, b, c)` можно сказать, что операндами операции `()` стали `f`, `a`, `b` и `c`.

Следующей экзотической операцией мы рассмотрим *тернарную* (то есть имеющую три аргумента) ***условную операцию***. Записывается она в виде трёх выражений, после первого ставится вопросительный знак, после второго — двоеточие, чтобы отделить его от третьего. Первое выражение вычисляется и рассматривается как логическое значение («истина» или «ложь»), и если получилась «истина», то вычисляется второе выражение, а если «ложь» — то третье. Например, мы могли бы написать:

```
x = cond(y) ? pro(y) : contra(y);
```

В этом случае сначала будет вызвана функция `cond(y)`, и если она вернёт «истину» (что угодно кроме нуля) — то в переменную `x` попадёт результат вычисления `pro(y)`, а `contra` вообще не будет вызываться; если же `cond(y)` вернёт «ложь», то есть ноль, то всё будет наоборот: `pro` вызываться не будет, а результат вычисления `contra(y)` будет занесён в `x`. Более популярный пример условной операции выглядит так:

```
max = a > b ? a : b;
```

Здесь в переменную `max` заносится большее из двух значений — `a` и `b`. Этот пример лаконичнее, но он не даёт возможности пояснить, что из второго и третьего операндов выбирается один, а второй при этом не только не выбирается, но и, что очень важно, *не вычисляется*; на предыдущем примере это видно более наглядно.

Чтобы понять, зачем такая операция нужна, следует сообразить, что она, естественно, может применяться не только в правой части присваивания, но и, например, в качестве параметра функции:

```
printf("Maximum value is %d\n", a > b ? a : b);
```

Более сложный пример может оказаться непонятным; в этом нет ничего страшного, его можно сейчас пропустить и вернуться к нему после изучения оператора `for`. Пусть у нас есть массив целых чисел `vec`, а его размер хранится в переменной `vsize`, и от нас требуется выдать его элементы на стандартный вывод, разделив их запятыми и пробелами, а в конце поставить перевод строки. Это можно, например, сделать так:

[17] Есть точка зрения, что в Си имя функции и есть её адрес, но есть и противоположная точка зрения, и самое занятное, что семантика некоторых операций поддерживает *обе* точки зрения, не давая возможности понять, какая из них правильная. Подробное рассмотрение этого момента будет приведено вместе с описанием указателей на функции.

```
for(i = 0; i < vsize; i++)
    printf("%d%s", vec[i], (i < vsize-1) ? ", " : "\n");
```

Пожалуй, самая экзотическая операция языка Си — это операция «запятая». Выражение вида `a,b` вычисляется так: сначала вычисляется `a`, потом вычисляется `b`, а значением всего вместе становится результат вычисления `b`. К тому, как и где эта операция может оказаться полезной, мы ещё вернёмся; но вообще-то этой операции в языке вполне могло бы не быть, поскольку во *всех* случаях, когда её используют, без неё можно легко обойтись. При этом сам символ «запятая», например, разделяет параметры в вызове функции, и там компилятор воспринимает этот символ не как символ операции, а как простой синтаксический разделитель. Остаётся только припомнить нашу мантру: нет, Си мы любим не за это.

Отдельного рассмотрения заслуживает ***операция преобразования типа выражения***. Если `expr` — это некоторое выражение, а `type` — некий тип, то конструкция вида `(type)expr` означает значение `expr`, *тип которого принудительно заменён на* `type`. Пусть, например, у нас имеется целочисленная переменная `n`; выражение `(double)(n+1)/3.0` означает, что к значению `n` будет прибавлена единица, результат будет преобразован в число с плавающей точкой (типа `double`) и поделен на `3.0`, причём деление будет «настоящее» (не целочисленное), так как его операнды — числа с плавающей точкой.

Чаще всего операцию преобразования типа применяют для адресных выражений и указателей, но об этом у нас разговор ещё впереди.

Операции, связанные с адресами и указателями

В завершение обсуждения операций перечислим те из них, для которых время подробного описания ещё не пришло, поскольку мы пока не знакомы ни с указателями, ни со структурными типами. Ко всем этим операциям мы обязательно вернёмся, здесь же мы приводим их исключительно для полноты картины:

- операция взятия адреса обозначается амперсандом «`&`», а от побитового «и» отличается количеством операндов — она унарная, в отличие от бинарного побитового «и»; например, выражение `&x` означает *адрес* переменной `x`;
- обратная операция — операция разыменования (англ. *dereference*), обозначается звёздочкой «`*`»; от операции умножения её тоже отличают по количеству аргументов; например, если `p` — это некий адрес, то `*p` — это то, что расположено в памяти по этому адресу;
- операция выборки поля из структуры или объединения обозначается точкой (что-нибудь вроде `item1.next`);
- операция выборки поля из структуры или объединения по имеющемуся *адресу* обозначается стрелкой `->`; например, `a->b` есть то же самое, что и `(*a).b`.

О приоритетах операций

Большинство книг, посвящённых языку Си, содержат полную таблицу приоритетов операций, и находятся люди, вызубривающие эту таблицу наизусть; в основном это студенты-троечники, которым обязательно надо убедить хотя бы самого себя, что усилий на учёбу потрачено «достаточно», а если ничего не получилось, то виноват предмет, а не студент; кроме того, в вызубривании таблицы приоритетов часто оказываются замечены классические «унылые отличники», у которых в школе выработалась вредная привычка зубрить всё, что попалось в учебнике. Разумеется, для практического программирования знание этой таблицы не только не обязательно, но и, пожалуй, бесполезно. Достаточно помнить всего несколько несложных правил:

- приоритет умножения и деления выше, чем сложения и вычитания — так же, как и в математических формулах; по той же причине приоритет конъюнкции выше приоритета дизъюнкции, ведь конъюнкция аналогична умножению, а дизъюнкция — сложению;
- приоритет сравнений ниже, чем приоритет арифметики, а приоритет логических связок ниже, чем приоритет сравнений; в частности, в выражении `c >= 'a' && c <= 'z'` скобки не нужны (в отличие, кстати, от Паскаля — это тот редкий случай, когда Си оказывается построен логичнее);
- приоритет любой унарной операции выше, чем приоритет любой бинарной; при этом операция индексирования `[]`, как ни странно, считается унарной, несмотря на наличие второго операнда, записываемого между скобками; унарной считается и операция вызова функции `()`, сколько бы параметров ни содержалось внутри скобок;
- приоритет *суффиксных* операций выше, чем приоритет префиксных; например, в выражении `++a[7]` сначала к `a` применяется индексирование, а к его результату (то есть к полученному элементу массива) применяется `++`;
- приоритет тернарной условной операции ниже всего перечисленного, а приоритет всех операций присваивания — ещё ниже, и в самом низу таблицы гордо располагается операция «запятая»; если не обращать внимания на «запятую», то в операндах тернарной операции можно использовать (без скобок) любые выражения, не содержащие присваиваний, а в правой части присваивания, опять же без скобок, поставить вообще любое выражение (кроме содержащих запятую; но выражения, в которых запятая всё-таки содержится, лучше не применять, ясности программе они не добавят).

4.3.6. Операторы языка Си

Количество *операторов* в языке Си гораздо меньше, чем в Паскале, потому что все возможности, которые только могут быть вытеснены из языка в библиотеку, реализованы именно в библиотеке. Как уже говорилось, в Си (самом по себе) отсутствуют какие бы то ни было средства ввода-вывода, а всевозможные `printf`'ы и `scanf`'ы — это библиотечные функции, которые сами написаны на Си. Это позволило создателям языка ограничиться всего *одиннадцатью* операторами, которые мы сейчас перечислим.

Читателям, привыкшим к Паскалю, ещё раз рекомендуется обратить внимание на то, что, в отличие от Паскаля, символ точки с запятой в Си считается не *разделителем* операторов, а *частью синтаксиса самих операторов* (притом не всех). Иначе говоря, если синтаксисом оператора предусмотрена точка с запятой, то она будет нужна там вне всякой зависимости от контекста; если же она не предусмотрена — то её там не будет (это значит, что компилятор без всякой точки с запятой знает, как распознать конец такого оператора).

Ещё одно общее отличие от Паскаля состоит в том, что ключевых слов, которые ставятся между условием и телом (подобно паскалевским `do`, `then` и `of`) в Си не предусмотрено; создателям Си пришлось придумать другой способ обозначения конца условного выражения: это делается путём заключения его в круглые скобки, которые здесь (в отличие от Паскаля) обязательны.

Вычисление выражения

Чаще других в программах на Си встречается ***оператор вычисления выражения ради побочного эффекта***. Его синтаксис выглядит очень просто: это произвольное арифметическое выражение, после которого стоит точка с запятой — она как раз и превращает *выражение* (которое могло бы быть частью другого выражения) в законченную конструкцию — оператор. Роль у этой точки с запятой двойная: во-первых, она сообщает компилятору, что конструкция закончилась и можно больше не ожидать появления знаков операций; во-вторых, она показывает, что для данного выражения нас совершенно не интересует его значение.

В принципе, конструкции вида

```
x + 5;
```

или даже просто «`5;`» с точки зрения языка Си оказываются вполне легитимными, хотя и бессмысленными, ведь здесь нет никакого побочного эффекта. Оптимизатор просто выбросит соответствующий код, а компилятор, если его запустили с флагом `-Wall`, выдаст предупреждение, что данный оператор «не имеет эффекта» (*has no effect*) — он так поступает всегда, когда конструкция не нарушает формальных правил языка, но есть основания сомневаться, что программист имел в виду именно это. Конечно, оператор вычисления выражения придуман для совершенно иных ситуаций. Одна из них — это обыкновенное присваивание:

```
x = 17;
```

Из предыдущего параграфа мы узнали, что присваивание в Си является операцией, так что `x = 17` — это не более чем выражение, которое могло бы встретиться в составе более сложных выражений и которое, кстати говоря, имеет значение; символ точки с запятой указывает, что нас не интересует это значение, а выражение являет собой завершённую конструкцию. Подчеркнём

ещё раз, **это никакой не «оператор присваивания», такого в Си просто нет**. Это оператор вычисления выражения ради побочного эффекта, а само выражение построено на основе бинарной арифметической операции — присваивания.

Вторая часто встречающаяся ситуация — это вызов функции ради побочного эффекта. Такова, например, хорошо знакомая нам строка

```
printf("Hello, world\n");
```

Выглядит это совсем не похоже на вышеприведённое присваивание, и тем не менее это *тот же самый оператор* вычисления выражения. В качестве выражения здесь выступает бинарная версия вызова функции (первый аргумент — `printf`, второй аргумент — строковый литерал `"Hello, world\n"`). Мы даже знаем, что у этого выражения есть значение — функция `printf` в результате такого вызова вернёт число `13`, но нам это число не нужно, о чём мы и сообщаем компилятору, поставив точку с запятой.

Отметим, что выражение может быть пустым, в результате чего получится самый короткий оператор языка Си — «пустой оператор», состоящий из одной только точки с запятой. Его используют, например, в качестве тела цикла, когда всё, что должен делать цикл, оказалось в заголовке; в программах на Си такое встречается гораздо чаще, чем хотелось бы. Впрочем, уж если вам понадобился пустой оператор, правильнее будет сделать его в виде *пустого блока* «`{}`», чем в виде неприметной и сливающейся с окружающим текстом точки с запятой. Пустые фигурные скобки намного лучше видно в тексте программы, и не возникает никаких сомнений, что автор программы имел в виду именно пустой оператор, а не что-то другое.

Составной оператор (блок); локальные переменные

Как и в Паскале, в Си часто (и ровно по тем же причинам) возникает потребность объединить несколько операторов в один; для этого используется ***составной оператор***, который, как мы уже успели понять (и, возможно, даже привыкнуть к этому), обозначается фигурными скобками. Но в сравнении с Паскалем составной оператор языка Си — средство более мощное, поскольку в Си составной оператор является полноценным ***блоком*** — таким фрагментом программы, для которого возможны локальные описания[18].

Локальные переменные (и другие локальные имена, о которых мы узнаем позже) описываются в самом начале блока, сразу после открывающей фигурной скобки[19] и до появления в блоке первого оператора. Напомним, что синтаксис описания переменных в Си в простейшем случае[20] предполагает имя типа, вслед за которым идёт имя переменной, либо несколько таких имён, разделённых запятой. Описания локальных переменных мы уже видели в примере программы, решающей квадратное уравнение, но там переменные описывались в начале функции; теперь мы знаем, что это можно делать в начале *любого составного оператора*.

К примеру, при реализации сортировки массива «пузырьком» нам нужна переменная, через которую мы меняем местами значения в элементах массива, если эти элементы стоят не в том порядке. В Паскале мы эту переменную вынуждены были описывать как локальную переменную соответствующей процедуры. В Си это можно сделать «более локально», прямо в теле `if`'а, например:

```
if(a[i] > a[i+1]) {
    int t;
    t = a[i];
    a[i] = a[i+1];
    a[i+1] = t;
    flag = 1;
}
```

Переменная `t` нигде больше не нужна, но в Паскале нам приходилось терпеть тот факт, что она видна во всём теле процедуры. В приведённом примере мы ограничили область видимости переменной `t` так, что её не видно нигде, кроме того фрагмента, в котором она непосредственно используется.

Ветвление

С оператором `if` мы уже неоднократно встречались в примерах, так что нам осталось сказать про него не так много. Его общий синтаксис таков:

`if (` *<условие>* `)` *<оператор 1>* [`else` *<оператор 2>*]

Как это обычно происходит в Си, условное выражение обязательно должно быть заключено в круглые скобки. Тип условного выражения должен быть либо целочисленным, либо адресным; арифметический ноль и нулевой указатель означают «ложь», всё остальное — «истину». Если результатом вычисления условного выражения стала «истина», выполняется *<оператор 1>*, в противном случае выполняется *<оператор 2>*, если таковой есть; его может не быть, поскольку ветка `else` необязательна.

Следует обратить внимание на то, что **синтаксис оператора `if` сам по себе не предусматривает точки с запятой**, но её могут предусматривать *<оператор 1>* и *<оператор 2>*. Наличие или отсутствие ветки `else` никак не влияет на наличие или отсутствие точки с запятой в *<операторе 1>*, на это влияет только синтаксис самого этого оператора; после Паскаля некоторое время приходится привыкать к тому, что точка с запятой перед `else` — это не ошибка.

Циклы

Простейший вариант цикла — обыкновенный ***цикл с предусловием***, который выглядит в Си почти так же, как в Паскале (даже ключевое слово используется то же самое — `while`), только условие приходится брать в круглые скобки и при этом нет слова `do` или каких-либо его аналогов. Точнее говоря, синтаксис цикла с предусловием в Си такой:

`while (` *<условие>* `)` *<оператор>*

Как и следовало ожидать, этот оператор предполагает вычисление условия, выход, если оно ложно, если же оно истинно — выполнение оператора и снова вычисление условия, и так пока условие не станет ложным.

Цикл с постусловием в Си, который называют обычно просто *do-while*, от паскалевского repeat-until отличается довольно заметно. Формально его синтаксис таков:

`do` *<оператор>* `while (` *<условие>* `) ;`

Пожалуй, самое заметное его отличие от паскалевского варианта — в семантике *<условия>*: здесь это *условие продолжения* работы цикла, тогда как в Паскале цикл с постусловием предполагает выражение для *условия завершения* цикла. Иначе говоря, цикл `do-while` выполняется так: сначала выполняется *<оператор>*, затем вычисляется *<условие>*, и если оно *истинно*, выполнение продолжается с начала, то есть снова выполняется *<оператор>*.

Второе отличие состоит в подходе к телу цикла. В Паскале используется тот факт, что слова, обозначающие начало и конец цикла, *уже есть*, и поэтому между ними можно заключить сколько угодно операторов, не используя операторные скобки. В Си на первый взгляд обозначение начала и конца тоже уже есть, но это не помогает, поскольку в конце используется слово `while`, которое может с тем же успехом обозначать начало другого (вложенного) цикла, на этот раз с предусловием; приходится предполагать, что телом, как и в других сложных операторах, будет *один* оператор. Иной вопрос, что на практике **тело цикла `do-while` всегда записывают в виде составного оператора**, некоторые программисты даже уверены, что это требование синтаксиса языка. Более того, единственная возможность исключить путаницу между двумя случаями употребления ключевого слова `while` состоит в том, что **при записи цикла `do-while` заключительное слово `while` оставляют на одной строке с закрывающей фигурной скобкой от тела**, примерно так:

```
do {
    get_event(&event);
    res = handle_event(&event);
} while(res != EV_QUIT_NOW);
```

[18] Напомним, что в Паскале блоки встречаются только в процедурах и функциях; собственно говоря, паскалевская подпрограмма представляет собой заголовок и блок, а блок, в свою очередь, состоит из секции описаний и секции операторов, последняя обрамляется традиционными «операторными скобками».

[19] Авторы стандарта C99 это ограничение «отменили», а описывать переменные разрешили в любом месте программы; настоятельно не рекомендуем этим пользоваться — эта возможность не стоит того, чтобы ради неё начать использовать нелепый C99 вместо традиционного языка Си.

[20] В общем виде синтаксис описания переменной в Си чрезвычайно сложен — настолько, что мы вообще не будем приводить его полностью, а отдельные его возможности будем вводить постепенно.

Обратите внимание, что оператор `do-while` оканчивается точкой с запятой, она входит в его синтаксис.

Кроме обычных циклов с предусловием и постусловием в Си также присутствует цикл `for`, но, в отличие от Паскаля, где одноимённый цикл просто называют арифметическим, здесь эта конструкция *может, но не обязана* представлять арифметический цикл. Неподготовленному читателю формальное описание цикла `for` может показаться совершенно непонятным; мы пойдём традиционным для таких ситуаций путём — сначала приведём примеры и лишь затем дадим формальное описание.

Допустим, у нас имеется целочисленная переменная `i`. Следующий цикл напечатает квадраты целых чисел от 1 до 25:

```
for(i = 1; i <= 25; i++)
    printf("%d x %d\t = %d\n", i, i, i*i);
```

Это действительно выглядит как арифметический цикл, и, собственно говоря, это он и есть. Однако, как мы сейчас увидим, конструкция `for` никоим образом не заставляет программиста сделать цикл именно арифметическим. Заголовок цикла `for` состоит из трёх секций, разделённых символом точки с запятой; в каждой из них можно записать выражение. Первое из выражений (*инициализация*) вычисляется перед началом работы цикла. Обычно здесь записывается присваивание начального значения переменной цикла, но вообще никто не мешает написать тут что угодно. Второе выражение представляет собой *условие* продолжения цикла (как в цикле `while`). Третье выражение (*итерация*) вычисляется каждый раз *после* выполнения тела цикла, то есть непосредственно перед очередным (но не первым!) вычислением условия. Формально синтаксис цикла `for` можно представить так:

`for (` [*<инициализация>*] `;` [*<условие>*] `;` [*<итерация>*] `)`
<оператор>

Все три выражения, предусмотренные в заголовке `for`, являются необязательными и могут быть пропущены (но точки с запятой должны при этом остаться на месте). Если пропущены инициализация или итерация — считается, что в соответствующей ситуации ничего делать не надо; пропущенное условие означает, что цикл следует выполнять «бесконечно», то есть до тех пор, пока выход из него не будет произведён каким-либо альтернативным способом. В частности, если в заголовке `for` присутствует только условие, такой цикл эквивалентен циклу `while` с таким же условием. Бесконечный цикл программисты на Си традиционно записывают лаконичным `for(;;)` (читается как *forever*; читатели, знакомые с английским, могут оценить изящную игру слов).

Осознать, что цикл `for` в языке Си — это совсем не то же самое, что одноимённый цикл Паскаля и арифметическим быть не обязан, нам поможет следующий пример:

```
for(s = 0; scanf("%lf", &k) == 1; s += k)
    ;
```

На всякий случай подчеркнём, что такой стиль программирования мы никоим образом не приветствуем и привели этот фрагмент лишь в качестве примера того, что можно сделать с циклом `for`, если не проявлять разборчивости в средствах. Как нетрудно догадаться, здесь выполняется чтение чисел типа `double` до наступления ситуации «конец файла» или возникновения ошибки, а в переменной `s` подсчитывается сумма прочитанных чисел; тело цикла оказалось пустым, поскольку всё, что нужно сделать, делается в выражениях заголовка.

Рассказ о цикле `for` будет неполным без упоминания операции «запятая». Описывая её (см. стр. 21), мы обещали рассказать, для чего её можно использовать, но «когда-нибудь потом». Так вот, благодаря наличию операции «запятая», а также тому, что части заголовка `for` разделяются точкой с запятой (отметим, что это *единственный* случай в Си, когда точка с запятой находится не в конце оператора, а в самой его сердцевине), в цикле `for` можно использовать *больше одной переменной цикла*. Например, следующий цикл «переворачивает» массив `arr` длины `arr_len`, состоящий из целочисленных элементов:

```
for(i = 0, j = arr_len - 1; i < j; i++, j--) {
    int t;
    t = arr[i];
    arr[i] = arr[j];
    arr[j] = t;
}
```

Наиболее примечателен здесь тот факт, что, судя по всему, *именно для этого запятую сделали операцией*. Некоторые энтузиасты находят ей другие применения, но всякий раз оказывается, что такие «находки» никакого выигрыша не дают. Иначе говоря, в язык была добавлена арифметическая операция, бесполезная и бессмысленная во всех контекстах, кроме одного, но и в этом одном использующаяся редко, к тому же превращающая синтаксис выражений в сумбур, поскольку во многих случаях запятая используется не как символ операции, а как простой разделитель. Нет, Си мы любим не за это.

Любой из циклов в любой момент может быть досрочно прекращён уже знакомым нам по Паскалю[21] оператором «`break;`» (именно так, используется ключевое слово `break` и точка с запятой, хотя, как это очень легко видеть, точка с запятой тут избыточна — но синтаксис Си именно таков).

Предусмотрен также и второй хорошо знакомый нам оператор, «`continue;`», который досрочно завершает выполнение тела цикла, сам цикл при этом продолжается со следующей итерации. В частности, в циклах `while` и `do-while` выполнение продолжается с вычисления условия (предусловия или постусловия — это здесь неважно), а в цикле `for` выполнение продолжается очередным вычислением выражения «*итерация*», после которого, опять-таки, вычисляется условие. Общее правило тут такое: оператор `continue` предписывает *проигнорировать остаток тела цикла*, но остальные элементы цикла продолжают выполняться в обычном порядке.

Возврат из функции

Как мы уже неоднократно видели в примерах, возврат управления из функции производится оператором `return`. Этот оператор несёт двойную нагрузку: во-первых, он определяет значение, которое вернёт функция; во-вторых, он завершает выполнение функции, в том числе досрочно.

В зависимости от того, предусмотрен ли возврат значения из данной функции, оператор `return` принимает одну из двух форм. Для обычных функций его синтаксис таков:

`return` *<выражение>* `;`

где *<выражение>*, будучи вычисленным, даёт как раз результат выполнения функции. Что же касается функций, для которых вместо типа возвращаемого значения указано слово `void`, то в них, разумеется, никакого выражения нет, ведь они не возвращают значений. В этом случае оператор принимает лаконичную вторую форму:

`return ;`

Следует отметить, что функция вполне способна завершиться без посторонней помощи, просто дойдя до конца, так что **в функциях, которые не возвращают значений, `return` следует использовать только для досрочного выхода**.

Иное дело «настоящие» функции. Поскольку в Си нет иного способа указать возвращаемое значение, а не вернуть никакого значения, когда от нас его ожидают — это явная ошибка, оператор `return` в таких функциях совершенно необходим и неизбежен; в большинстве случаев он ставится в последней строке текста функции. Если этого не сделать, или, точнее говоря, написать текст функции так, что компилятору *покажется*, будто функция может дойти до конца, так и не встретив оператор `return`, и при этом её тип значения не `void`, то будет выдано предупреждение, и такое предупреждение ни в коем случае не следует игнорировать. Пользуясь случаем, ещё раз напомним читателю про флажок `-Wall` в командной строке компилятора. Не будет флажка — не будет и предупреждения, и сколько крови мы на этом попортим, остаётся только гадать.

Чтобы понять, почему определение возвращаемого значения намертво сцеплено с завершением функции, стоит обратиться к конвенции вызовов функций в Си. По традиции возврат значения из функций производится через регистр, причём среди всех регистров выбирается «самый главный»; в частности, на архитектуре i386 целые числа, умещающиеся в 32 бита, возвращаются через регистр `EAX`, а любые числа с плавающей точкой возвращаются через вершину регистрового стека сопроцессора. Вспомнив систему команд i386, мы легко убедимся в том, что после занесения финального значения функции в соответствующий регистр мы больше не сможем производить вычисления: слишком уж неудобно обходиться без «аккумулятора» (а если понадобится умножать и

[21] На самом деле этих операторов не было в оригинальном Паскале; как несложно догадаться, они были заимствованы, причём как раз из языка Си, так что сейчас мы, если можно так выразиться, добрались до их первоисточника.

делить, так и попросту невозможно), ну а вычисления с плавающей точкой в обход вершины стека сопроцессора не получится делать вообще никак. Мало того, если вызвать какую-то ещё функцию, она тоже — в соответствии с той же самой конвенцией — вернёт своё значение через те же «самые главные регистры», тем самым испортив то, что в них лежит. Вот так вот и получается, что занесение возвращаемого значения «куда следует» должно быть *последним* действием перед возвратом управления из функции.

Но почему, можно спросить, такого не происходило в Паскале? Дело в том, что конвенция вызовов подпрограмм, применявшаяся в классических версиях Паскаля, принципиально иная, и, в частности, возврат значения там производился через специально для этого выделенную область памяти в стеке, причём выделял её вызывающий[22]. Ясно, что занести значение в такую область памяти можно в любой момент, и оно будет там лежать, дожидаясь своего часа и никоим образом не мешая нам производить дальнейшие вычисления. Стоит отметить, что для функций, возвращающих нечто, не помещающееся в `EAX` и при этом не являющееся числом с плавающей точкой, компиляторы Си производят передачу возвращаемого значения через область памяти, чей адрес передан в функцию неявным параметром, ну а в самых ранних версиях языка значения таких типов вообще нельзя было вернуть из функции.

Оператор `goto`

Уже знакомый нам по Паскалю ***оператор безусловного перехода*** заслуживает отдельного рассмотрения. Единственное заметное отличие от Паскаля состоит в том, что метку, на которую будет делаться `goto`, не нужно заранее описывать: вы просто используете произвольный идентификатор, не занятый ни подо что другое. Областью видимости такого идентификатора всегда является функция, «прыгать» из одной функции в другую нельзя; но вот внутри функции прыгнуть можно в том числе и *внутрь сложного оператора*, в том смысле, что компилятор не станет вам мешать вытворять подобный идиотизм; разумеется, пользоваться этим ни в коем случае не следует.

Как и в Паскале, меткой можно пометить только *оператор*, так что, если возникает необходимость поставить метку в конце составного оператора, после неё необходимо предусмотреть пустой оператор (обычно просто ставят точку с запятой). Как и в большинстве языков, где присутствуют метки, собственно метка записывается непосредственно перед оператором и отделяется от него двоеточием. Синтаксис оператора `goto` и вовсе ничем на первый взгляд не отличается от паскалевского, только что точка с запятой по традиции входит в состав оператора:

`goto` <*метка*> `;`

Изучая Паскаль, мы уже отмечали (см. т. 1, §2.4.3), что в большинстве случаев `goto` лучше не использовать, но есть два (не один, не три, а ровно два) случая, когда использование `goto` не только не запутывает программу, но, напротив, делает её более ясной. Напомним, что эти ситуации — выход из нескольких вложенных друг в друга конструкций и освобождение локально захваченных ресурсов перед досрочным завершением подпрограммы; наше рассуждение про использовние `goto` в Паскале полностью применимо и к Си.

Оператор выбора

Оператор выбора `switch`, приблизительный аналог паскалевского `case`, пожалуй, самый запутанный в Си. Идея, как и в Паскале, состоит в том, чтобы в зависимости от значения некоторого выражения выполнить одну из нескольких ветвей кода, то есть это своего рода обобщение ветвления на случай N ветвей; однако подход к построению соответствующего оператора в Си кардинально отличается. Общий синтаксис оператора `switch` таков:

`switch (` <*выражение*> `) {` <*метки и операторы*> `}`

При этом <*выражение*> должно иметь произвольный целочисленный тип, а <*метки и операторы*> представляют собой произвольную последовательность операторов языка Си, первый из которых, а также, возможно, *некоторые* другие снабжены так называемыми `case`-метками. Эти метки бывают двух видов; обыкновенные `case`-метки состоят из ключевого слова `case`, *целочисленной константы времени компиляции* и двоеточия, то есть имеют следующий синтаксис:

`case` <*константа*> `:`

Метка второго вида выглядит проще:

`default :`

Такая метка обозначает «все варианты, кроме явно перечисленных»; обычно её ставят в конце оператора `switch`, хотя это и не обязательно. Подчеркнём, что константы в `case`-метках должны быть таковы, чтобы компилятор мог вычислить их во время компиляции; переменные или тем более вызовы функций здесь не годятся. Более того, в языке Си присутствуют константы, описываемые как переменные, но со словом `const`, запрещающим их изменение; так вот, они тоже не считаются ***константами времени компиляции*** и не могут использоваться в `case`-метках; это объясняется тем, что такая константа может, вообще говоря, располагаться в другой единице трансляции, так что компилятор не будет знать её значение во время компиляции.

А теперь попытаемся осознать ключевое отличие оператора `switch` от привычных конструкций выбора из других языков. **Операторы в теле `switch` представляют собой единую секцию операторов, которая выполняется от метки, соответствующей вычисленному значению выражения, и до конца тела `switch`**, если только выполнение не будет прервано с помощью какого-нибудь из операторов, передающих управление куда-нибудь ещё. Например[23], оператор

```
switch(t) {
case 1:
    printf("First\n");
case 2:
    printf("Second\n");
case 3:
    printf("Third\n");
case 4:
    printf("Fourth\n");
case 5:
    printf("Fifth\n");
default:
    printf("More\n");
}
```

при `t`, равном `4`, напечатает не просто `Fourth`, как обычно ожидают начинающие, а гораздо больше:

```
Fourth
Fifth
More
```

Иначе говоря, очередная `case`-метка являет собой точку, откуда выполнение тела `switch` могло бы начаться, но никоим образом не предписывает завершить выполнение операторов, если это выполнение началось выше неё.

Такая экзотическая семантика вполне объяснима, если вспомнить, что Си создан как замена языку ассемблера; в некоторых (достаточно редких) случаях это даже оказывается удобно, но чаще нам в программах, наоборот, хотелось бы разбить код на непересекающиеся ветви и выбрать одну из них в зависимости от значения выражения, то есть, грубо говоря, паскалевский вариант оператора выбора нас в большинстве случаев устроил бы больше. Поэтому перед каждой очередной `case`-меткой нам приходится тем или иным способом явно прекращать выполнение тела `switch`. Обычно для этого используют `break`, который внутри `switch` означает примерно то же, что и внутри циклов: прекратить выполнение тела, передав управление в конец оператора `switch`; например:

```
switch(t) {
case 1:
    printf("First\n");
    break;
case 2:
    printf("Second\n");
    break;
case 3:
    printf("Third\n");
    break;
case 4:
    printf("Fourth\n");
    break;
case 5:
```

[22] Отметим, что Free Pascal, который мы изучали, использует для возврата значения регистры — точно так же, как это делают компиляторы Си; если присваивание, фиксирующее возвращаемое значение, делается в функции раньше её завершения, это значение хранится в области локальных переменных, а перед возвратом переносится в `EAX` или в стек сопроцессора.

[23] Здесь в примерах мы используем в метках `case` обычные целые числа. В реальных программах так делать обычно не следует, у чисел должны быть имена; мы вернёмся к этому вопросу в §4.3.9.

```
        printf("Fifth\n");
        break;
    default:
        printf("More\n");
    }
```

Кроме того, досрочно завершить `switch` можно, сделав `goto` «наружу», вернув управление из функции с помощью `return`, а при наличии объемлющего цикла — выполнив `continue` для досрочного прекращения его итерации. В реальной жизни чаще других из всех этих способов применяется как раз `return`.

Следует сказать несколько слов об оформлении фрагментов кода, содержащих `switch`. Прежде всего отметим, что в вышеприведённых примерах мы не сдвигали метки относительно самого оператора; можно их и сдвигать, например так:

```
switch(t) {
    case 1:
        printf("First\n");
        break;
    case 2:
        printf("Second\n");
        break;
    default:
        printf("More\n");
}
```

Но главное тут в другом. Оператор выбора, будучи применён в реальной практике, обычно имеет тенденцию распухать на несколько десятков строк, делая функцию, в которой он располагается, совершенно нечитаемой. С этим можно бороться несколькими способами:

- старайтесь выносить `switch` целиком в отдельную функцию или хотя бы оставлять вместе с ним в одной функции как можно меньше другого кода;
- если отдельные секции `switch` поддаются оформлению в виде обособленных функций, обязательно сделайте это, оставив в теле самого `switch` только метки, вызовы этих функций и `break`'и;
- **никогда не размещайте один `switch` внутри другого**.

4.3.7. Перечислимый тип

Как мы помним (см. т. 1, §2.6.2), язык Паскаль позволяет ввести тип данных, множество значений которого конечно и явным образом задаётся программистом в виде набора *идентификаторов*; эти идентификаторы, собственно говоря, как раз и являются возможными значениями. Такой тип называется «***перечислимым***», он есть не только в Паскале, но и во многих других языках программирования. В языке Си перечислимый тип тоже присутствует и обозначается ключевым словом `enum` (от слова *enumeration* — «перечисление»), но при внимательном рассмотрении оказывается явлением совершенно иным, нежели перечислимый тип Паскаля.

Приведём для начала пример описания перечислимого типа:

```
enum colors { red, orange, yellow, green, blue, violet };
```

Здесь идентификаторы `red`, `orange` и т. д. представляют набор значений нового типа, `colors` — это ***имя перечисления***, но, как ни странно, *именем типа* этот идентификатор сам по себе не является, хотя, несомненно, новый тип здесь введён. По правилам языка Си имя только что введённого типа состоит из *двух* слов: `enum colors`, то есть мы можем, например, описать переменную такого типа:

```
enum colors one_color;
```

Если при этом забыть слово `enum`, компилятор выдаст ошибку, поскольку типа `colors` он не знает, а знает только тип `enum colors`. По идее, переменная `one_color` теперь должна была бы быть способна принимать только значения, явным образом указанные в описании типа, а сами эти значения — `red`, `green` и прочие — можно было бы заносить только в такие переменные; но не тут-то было, ведь Си не Паскаль.

Такое же точно именование из двух слов — ключевого, указывающего разновидность типа, и идентификатора самого типа — применяется в Си также для составных типов, которые мы будем рассматривать в следующей главе. Почему и зачем создатели языка Си пошли этим довольно неочевидным путём, сказать трудно; в частности, в языке Си++, появившемся на десяток лет позже, идентификаторы перечислений, структур и объединений, а также классов (фирменной особенности Си++) стали полноценными именами типов, но «чистый» Си от такого «двухсловного» именования пользовательских типов так и не отказался.

В отличие от Паскаля, **в Си идентификаторы, введённые в качестве возможных значений в описании перечислимого типа, считаются целочисленными константами**, то есть имеют тип `int`. По умолчанию самая первая константа из описания перечисления, в данном случае `red`, будет иметь значение `0`, а каждая следующая окажется на единицу больше: `orange` будет равна `1`, `yellow` — `2` и т. д., `violet` получит значение `5`. Важно понимать, что эти идентификаторы действительно становятся «честными синонимами» соответствующих чисел, то есть их можно присваивать целочисленным переменным, можно использовать в арифметических выражениях и т. д.: например, выражение `violet * green - blue` компилятор проглотит без звука, получится то же самое, как если бы мы написали `5 * 3 - 4`, то есть `11` (это, разумеется, не значит, что так действительно стоит делать; наоборот, так делать не надо). Больше того, и в переменную `one_color` можно занести любое целое число, поскольку с точки зрения разрядности и знаковости она представляет собой обыкновенный `int`, но здесь компилятор хотя бы выдаст предупреждение.

Многие программисты, пишущие на Си, в особенности те, кто никогда не писал на более строгих языках вроде того же Паскаля, предпочитают вообще не вводить переменные и функции типа `enum`, обходясь обычными `int`'ами, а описания перечислимых типов используют только для задания констант. Такая практика вряд ли заслуживает одобрения, ведь тем самым мы лишаем компилятор возможности выдать нам пусть не ошибку, но хотя бы предупреждение в сомнительной ситуации. Тем не менее этот подход оказался столь популярен, что при описании типа `enum` в Си даже разрешили не указывать имя перечисления, то есть писать что-то вроде

```
enum { alpha, beta, gamma };
```

пропустив при этом идентификатор, относящийся к самому типу. Ясно, что способа описать переменную или функцию такого типа у нас не будет.

Значения констант, вводимых в описании перечислимого типа, можно задать явно; так, при описании цветов мы могли бы поступить хитрее, чем в рассмотренном примере:

```
enum colors {
    red      = 0xff0000,
    green    = 0x00ff00,
    blue     = 0x0000ff,
    yellow   = 0xffff00,
    cyan     = 0x00ffff,
    magenta  = 0xff00ff,
    black    = 0,
    white    = 0xffffff,
    grey     = 0x808080,
    violet   = 0xee82ee
};
```

Здесь мы не просто ввели константы для обозначения цветов, но и снабдили их значениями, соответствующими представлению этих цветов в модели RGB; разумеется, это существенно расширяет область применения введённых констант.

Две константы из одного и того же перечисления могут иметь одинаковые значения, Си это разрешает. Например, мы могли бы добавить в наше перечисление ещё одну константу:

```
    aqua     = 0x00ffff,
```

которая, как можно заметить, имеет такое же значение, как и `cyan`.

В одном описании можно смешивать константы с указанием значения и без такового; константа, значение которой не указано, будет на единицу больше предыдущей. Например, если написать

```
enum numbers { one = 1, two, three, first = 1, second, third };
```

то идентификаторы `two` и `second` будут иметь значение `2`, а `three` и `third` — значение `3`. При указании значения справа от знака равенства можно использовать любую константу времени компиляции, в том числе арифметическое выражение, которое компилятор может вычислить, например:

```
enum example_enum {
    example_first = 100,
    example_second = example_first * 20,
    example_third = example_second + 1000,
    example_last = example_third
};
```

Здесь введённые константы получат значения 100, 2000, 3000 и 3000.

Отдельного замечания заслуживает *точка с запятой после закрывающей фигурной скобки* в описаниях перечислимых типов. Для Си подобное сочетание не вполне характерно и встречается всего в двух случаях: при описании пользовательских типов и при описании переменных со сложными инициализаторами. Мы ещё не рассматривали ни массивы, ни структуры, и, как следствие, не встречались и с их инициализаторами, так что сейчас нам придётся чуть-чуть забежать вперёд, но поскольку с Паскалем (и массивами в нём) мы уже знакомы, проблем возникнуть вроде бы не должно. Итак, примером описания со сложным инициализатором может служить что-нибудь вроде

```
double coeffs[3] = { 0.5, 1.12, 1.92 };
```

Как можно догадаться, здесь описан массив из трёх элементов типа `double`, и для всех трёх заданы *начальные значения* — это числа, помещённые в фигурных скобках. Как видим, описание заканчивается точкой с запятой, причём сразу после закрывающей фигурной скобки — точно так же, как и описание перечислимого типа.

На самом деле это не две различные ситуации, а одна: точкой с запятой в Си должно заканчиваться любое объявление или описание, за исключением *описания функции*, в котором вместо точки с запятой фигурирует тело функции. Больше того, в Си вообще не различаются конструкции описания типа и описания переменной; коль скоро появился тип, можно указать и имена переменных, например:

```
enum states {
    running, blocked, ready
} new_state, last_state;
```

Здесь одно описание вводит сразу тип `enum states`, три целочисленные константы `running`, `blocked` и `ready`, а также две *переменные* `new_state` и `last_state`, обе имеющие тип `enum states`. Точку с запятой, которую мы ставили раньше в описаниях `enum`'ов сразу после закрывающей фигурной скобки, можно рассматривать как наш отказ от возможности немедленно описать переменные для только что введённого типа, но на самом деле она просто обозначает конец конструкции *описания*, которая в общем случае состоит из спецификатора типа и списка описываемых переменных (возможно, пустого). Формальный синтаксис Си позволяет оставить список переменных пустым в любом описании, в том числе и таком, которое никаких новых типов не вводит; например, строка

```
int ;
```

не является ошибочной с формально-синтаксической точки зрения, это обычное *описание* с пустым списком переменных. Конечно, такое описание *бессмысленно*, и большинство компиляторов выдаст здесь предупреждение — но не ошибку.

Так или иначе, если забыть поставить точку с запятой после описания типа, диагностика, которую вы получите от компилятора, скорее всего, будет настолько невразумительной, что у начинающего практически нет шансов с ходу понять, в чём на самом деле состоит проблема. Здесь можно дать сразу два совета. Во-первых, **будьте внимательны и не забывайте точку с запятой в конце описания нового типа**, отнеситесь к этому моменту чуть более серьёзно, чем к остальным вывертам языка Си, требующим внимательности. Во-вторых, попробуйте (можно прямо сейчас) сделать эту ошибку преднамеренно и посмотрите, какова будет диагностика, выданная вашим компилятором. Есть шанс, что после этого, когда то же самое произойдёт нечаянно, вы вспомните, что такую диагностику уже встречали, и потратите меньше времени на выявление источника проблемы.

4.3.8. Введение констант с помощью перечислимого типа

При изучении Паскаля мы отмечали, что наличие в программе явным образом указанных чисел крайне нежелательно, за исключением совершенно очевидных случаев с использованием 0, 1, очень редко 2, иногда ещё -1. Все остальные числовые константы должны быть снабжены именами, и в программе следует использовать имена, а не числа. Паскаль для этого предусматривает секцию описания констант.

В ранних версиях Си для именования константных значений приходилось использовать макропроцессор, никаких иных средств для введения констант не было. Подробный разговор о макросах Си у нас ещё впереди, пока же отметим, что макропроцессор отрабатывает после лексического анализа, но раньше синтаксического, в результате чего идентификаторы, введённые как макросы, не подчиняются областям видимости, в частности, не локализуются в функциях и т. п., и вообще они во многом крайне неудобны.

Введение в язык перечислимого типа, разумеется, не имело самостоятельной целью создание средства для описания констант; факт, однако, состоит в том, что в Си это единственный способ описать корректный идентификатор, подчиняющийся соглашениям об областях видимости, но при этом представляющий собой ***константу времени компиляции***. Вскоре после появления в языке перечислимого типа этот факт был осознан и им начали активно пользоваться. В современных программах на Си можно часто встретить что-то вроде

```
enum { max_buffer_size = 1024 };
```

Очевидно, что в качестве «перечислимого типа» такая конструкция не имеет ни малейшего смысла: никому не нужен тип, выражения которого могут принимать всего одно значение, да к тому же ещё и безымянный. Цель такого описания — совершенно в другом: создать *имя* (в данном случае `max_buffer_size`), которое будет в программе использоваться для обозначения некой константной величины, избавляя программиста от необходимости явно упоминать число 1024, причём, возможно, в нескольких местах, что может обернуться феерическими проблемами в случае принятия решения об изменении константы (в данном случае — об изменении максимального размера какого-то буфера), например, с 1024 на 2048. Читателя программы применение константы избавляет от необходимости каждый раз при виде числа 1024 судорожно пытаться понять, что же имелось в виду под числом 1024 и почему используется именно такое число, а не какое-то другое; при применении имени вместо числа становится более-менее понятно, что здесь имеется в виду какой-то размер буфера, а если непонятно, что это за буфер и почему у него именно такой размер, а не иной, можно найти (обычным текстовым поиском) то место, где этот идентификатор введён; скорее всего, там окажется подобающий комментарий.

Этот способ введения констант имеет одно весьма существенное ограничение: так можно ввести только *целочисленную* константу, и никакую другую; для других приходится по-прежнему использовать макропроцессор. Всё-таки средство (перечислимые типы) здесь применяется не по его исходному назначению, так что тут и жаловаться, в общем, не на что. Но даже с учётом этого ограничения и при наличии чёткого понимания, что сие есть откровенный *хак*, такой способ введения констант имеет несомненные достоинства в сравнении с применением макропроцессора и, естественно, широко применяется.

Отдельно стоит заметить, что в большинстве программ, а равно и во многих книгах, посвящённых Си, включая книгу Кернигана и Ритчи, имена таких констант записывают в верхнем регистре, то есть всеми большими буквами — что-то вроде `MAX_BUFFER_SIZE`. В главе, посвящённой макропроцессору, мы подробно обсудим, откуда взялась такая странная традиция и по каким причинам ей совсем не стоит следовать, когда речь идёт о константах, вводимых через `enum`.

4.3.9. Перечислимый тип и оператор выбора

Поскольку константы-значения перечислимого типа представляют собой, как уже говорилось, *константы времени компиляции*, их можно (и нужно) использовать в составе `case`-меток оператора `switch`. При этом выражение в заголовке оператора может, конечно, иметь обычный целочисленный тип, но если всё-таки постараться и сделать так, чтобы это выражение было типа `enum` — использовать переменную типа `enum` или функцию, возвращающую

значение этого типа, — то мы получим дополнительную поддержку со стороны компилятора: он станет проверять, *все ли возможные значения мы обработали*.

Говоря формально, компилятор предполагает, что если выражение, по которому проводится выбор в операторе `switch`, имеет перечислимый тип, то в теле `switch` должны присутствовать `case`-метки для всех *различных* значений, предусмотренных для данного перечислимого типа, либо, по крайней мере, должна присутствовать метка `default`; если это не так, компилятор выдаёт предупреждение. Пусть, например, у нас описан перечислимый тип

```
enum greek { alpha, beta, gamma = beta, delta };
```

и где-то в программе встречается функция

```
void greek_print(enum greek x)
{
    switch(x) {
    case alpha:
        printf("Alpha\n");
        break;
    case beta:
        printf("Beta\n");
        break;
    }
}
```

При компиляции этой функции компилятор выдаст предупреждение, что значение перечислимого типа `delta` не обработано в операторе `switch`; заметим, про значение `gamma` компилятор ничего не скажет, поскольку оно совпадает с `beta`, а значение `beta` в операторе обработано.

Избавиться от этого предупреждения можно очень просто: поставить метку `default` в самом конце оператора `switch`, а после неё, как этого требует синтаксис языка, предусмотреть оператор — разумеется, пустой (например, просто точку с запятой):

```
      /*   ...    */
    case beta:
        printf("Beta\n");
        break;
    default:
        ;
    }
}
```

Однако намного лучше и правильнее будет не «избавляться от предупреждения», а внимательно изучить ситуацию, чтобы понять, почему же так получилось, что в операторе есть альтернативы для всех вариантов, предусмотренных нашим перечислимым типом, кроме одного. Очень часто это происходит потому, что в описание типа был добавлен новый вариант значения, но при этом в соответствующий оператор `switch` его обработку добавить *забыли*. Предупреждение, выдаваемое компилятором, призвано оповестить программиста об этой ошибке, возможно, избавив от долгой и мучительной отладки в будущем; отмахиваться от таких предупреждений довольно глупо, ведь выйдет в итоге, скорее всего, себе дороже. Больше того, в ситуациях, когда выражение в заголовке `switch` имеет перечислимый тип, рекомендуется вообще воздерживаться от применения метки `default`, перечисляя все возможные альтернативы (которых обычно не так много) в явном виде; если для части альтернатив делать ничего не нужно, их следует сгруппировать в конце тела `switch`, там, где вы в противном случае поставили бы метку `default`. Сгенерированный машинный код ничем отличаться не будет, но читателю вашей программы вы при этом недвусмысленно покажете, что ничего не забыли.

Отметим ещё один важный момент. В программах начинающих программистов часто можно встретить что-то похожее на

```
switch(n) {
case 1:
    /* ... */
case 2:
    /* ... */

    /* ... */

case 13:
    /* ... */

}
```

— такой оператор выбора, в котором метками выступают обычные целые числа. Ни в одном серьёзном проекте такой код не имеет шансов пройти контроль качества; больше того, как только такое увидит ваш более опытный коллега, эффект вам, скорее всего, не понравится. С работы, пожалуй, не уволят (хотя случается всякое), но широкого обсуждения ваших профессиональных качеств, в просторечии именуемого «публичной поркой», вам точно не избежать, и, конечно же, такой код однозначно придётся переделать. В подобных случаях использование `enum` вместо чисел *обязательно*, более того, для таких вещей `enum` как раз и предназначен.

Чтобы понять, почему это так, задайте себе простой вопрос: а что такое, простите, `13`, какую ситуацию это число обозначает, чему соответствует? Будучи автором кода, вы, скорее всего, ответ найдёте почти сразу — даже если вы его не помните, то, вероятно, хорошо помните, куда надо посмотреть, чтобы вспомнить. А теперь поставьте себя на место стороннего читателя вашей программы и прикиньте, сколько у него займёт времени поиск в вашем тексте соответствущей информации. Больше того, вы и сами можете легко перепутать, например, число `13` и число `12`, использовав их в одном месте одним способом, а в другом — другим. Если дело дойдёт до анализа вашего кода на предмет его правильности, придётся в каждом из мест, где встречается загадочная нумерация, проверять, не перепутал ли автор свои собственные номера, на что может уйти масса времени и сил. А если вместо загадочного числа `13` в программе использовать имя константы, описывающее то, чему оно соответствует, все перечисленные проблемы вообще не возникают.

4.3.10. Локальные «статические» переменные

Мы уже знакомы с *глобальными* и *локальными* переменными; напомним, что глобальные переменные описываются вне тел функций и видны во всём тексте исходного файла, начиная от точки их описания (или объявления), тогда как локальные переменные описываются в начале любого блока — тела функции или составного оператора — и видны от точки описания до конца того блока, в котором переменная описана, то есть до закрывающей фигурной скобки.

Кроме ***области видимости***, переменные характеризуются ещё ***временем существования***, то есть временны́м периодом от того момента, когда переменная создаётся (под неё выделяется область памяти), и до момента, когда ранее выделенная под переменную область памяти высвобождается. Для глобальных переменных время существования совпадает с временем выполнения программы, а локальные переменные, располагающиеся в стеке, начинают существовать при входе в блок, в котором они описаны, и исчезают при выходе из этого блока (точнее говоря, они возникают и исчезают в момент соответствующего изменения регистра «указатель стека»).

В языке Си предусмотрен довольно странный гибрид локальных переменных с глобальными — так называемые локальные статические переменные. Как и простые локальные переменные, статические переменные описываются в начале блока и видны в тексте программы от точки описания до конца блока, но их время существования, как и у глобальных переменных, совпадает со временем работы программы, и размещаются они, как и глобальные переменные, в секции `.data` или `.bss` в зависимости от того, задан для них инициализатор или нет; если инициализатора нет, статическая переменная инициализируется нулём. Описание локальной статической переменной отличается от описания обычной локальной переменной наличием слова `static`, например:

```
int my_function(int x, int y)
{
    int tt = 25;
    static int zz = 75;

    /* ... */
}
```

В этом примере `tt` — обыкновенная локальная переменная, она заводится заново каждый раз, когда начинается исполнение функции `my_function`, и получает начальное значение `25`; если функция будет прямо или косвенно вызывать сама себя (напомним, это называется рекурсией), то переменная `tt` может одновременно существовать в нескольких экземплярах — по числу вызовов `my_function`, которые уже состоялись, но ещё не завершились. В момент завершения работы функции соответствующий экземпляр переменной `tt` исчезает вместе со всем стековым фреймом.

Совершенно иначе ведёт себя переменная `zz`: вне всякой зависимости от того, вызывается ли функция `my_function` или нет, переменная `zz` с самого начала работы программы уже существует, имеет значение `75` и готова к работе, а по окончании работы функции она никуда не исчезает; иначе говоря, она *сохраняет своё значение между вызовами функции*. Например, функция

```
void print_next_number()
{
    static int n = 0;
    printf("%d\n", n);
    n++;
}
```

при каждом её вызове печатает следующее целое число, то есть при первом вызове печатает `0`, при втором вызове — `1` и так далее.

Следует отметить, что локальные статические переменные обладают большинством недостатков обычных глобальных переменных, так что перед их использованием следует хорошо подумать. Если вы не совсем уверены, что понимаете, для чего нужно применение локальных статических переменных, то не применяйте их вовсе; ваша программа от этого хуже не станет.

На всякий случай отметим ещё один момент, связанный со словом `static`. Этот модификатор можно применять не только к локальным, но и к глобальным описаниям, причём как к переменным, так и к функциям. При этом смысл слова `static` становится совершенно иным, не имеющим ничего общего с вышеописанными статическими локальными переменными: для глобальных описаний `static` означает, что они не будут видны из других модулей программы. Подробно мы этот момент обсудим в главе, посвящённой раздельной трансляции.

4.3.11. Адреса, указатели и операции над ними

Программировать на языке Си совершенно невозможно без досконального понимания работы с указателями и адресными выражениями; здесь на работу с адресами завязана масса возможностей, без которых вообще трудно себе представить создание мало-мальски сложной программы. Мы уже знаем из §4.2.3, что параметры в подпрограммы (функции) здесь можно передавать только по значению, и если нам требуется, чтобы функция изменила значение переменной, приходится передавать ей *адрес* переменной, как мы это делали при обращении к функции `scanf`. Мало того, работа с массивами в языке Си организована в виде ***арифметики адресов***; её часто называют арифметикой указателей, что не совсем верно. Так или иначе, прежде чем рассматривать массивы, нужно освоить работу с указателями.

Напомним, что наш базовый вычислитель — это машина фон Неймана, одним из основных свойств которой является линейность и однородность оперативной памяти, а отдельные ячейки памяти идентифицируются *адресами*. То, что в языках высокого уровня[24] называют ***переменной***, на уровне машинного кода представляет собой не более чем *область памяти*, то есть несколько ячеек памяти, расположенных подряд (имеющих последовательные адреса). Под ***адресом области памяти*** понимается наименьший из адресов ячеек, составляющих область. Для нас здесь важно то, что любая переменная имеет свой адрес.

Строго говоря, компилятор может расположить локальную переменную в регистре, тогда адреса у неё не будет; но он так поступает лишь в случае, если мы никак не можем этого обнаружить. В частности, если мы применим к переменной операцию взятия адреса, компилятор не станет располагать её в регистре.

Во многих языках программирования, включая Паскаль и Си, адреса считаются информацией, которую можно хранить и обрабатывать; но если при работе на языке ассемблера адреса ничем не отличаются от обычных чисел, то языки высокого уровня вводят для адресов отдельные типы данных. Как и в Паскале, в языке Си адресный тип привязан к типу переменной, адрес которой имеется в виду.

Напомним два базовых принципа, которые мы уже формулировали в первом томе при рассмотрении указателей Паскаля:

Указатель — это переменная,
в которой хранится адрес.

Утверждение вида «A указывает на B»
означает «A содержит адрес B».

Собственно говоря, эти принципы общие, они справедливы не только для Си. А теперь приступим к делу.

Адресный тип описывается в Си с помощью символа звёздочки, которая ставится после типа переменной, адрес которой описывается; например, «`int *`» означает *адрес* ***`int`***, а «`double *`» — *адрес* ***`double`***. Соответствующим образом описываются и переменные-указатели:

```
int *p;
double *q;
```

Отдельно стоит сказать о положении символа «`*`» в описании. Сама «звёздочка» относится к символам-разделителям, то есть выделять её пробелами не обязательно, и с точки зрения компилятора совершенно не важно, с какой стороны мы поставим пробелы и поставим ли их вообще. Собственно говоря, нам было бы совершенно всё равно, как писать объявления указателей, если бы не вариант описания, в котором перечисляются через запятую несколько имён переменных; и здесь внезапно оказывается, что звёздочка, хоть и имеет отношение к типу, воспринимается компилятором как атрибут описываемого имени, а не как часть типа. Например, если нам нужно описать три указателя на `char`, это выглядит так:

```
char *s1, *s2, *s3;
```

Начинающие программисты, присоединяя звёздочку к типу, а не к имени (что, вообще говоря, вполне логично), часто делают довольно характерную ошибку: пишут что-то вроде

```
char* s1, s2, s3;
```

— и ждут, что все три переменные получат тип `char*`; на самом же деле при этом такой тип получит только `s1`, а остальные две окажутся типа `char`. **Если вы чувствуете себя неуверенно, просто описывайте каждый указатель на отдельной строке, не перечисляя их через запятую.** Много места это не займёт, а текст станет яснее.

Основные операции, связанные с адресами — это ***операция взятия адреса***, обозначаемая символом «`&`», и обратная к ней ***операция разыменования***, обозначаемая всё той же «звёздочкой» «`*`». Так, если `t` — произвольная переменная, то `&t` — это её адрес; если `p` — указатель, то `*p` — это то, на что он указывает. Например, если у нас описаны две переменные

```
int x;
int *p;
```

— то мы можем занести адрес `x` в указатель `p`:

```
p = &x;
```

Теперь `*p` — это своего рода «синоним» для `x`, то есть, например,

```
*p = 27;
```

занесёт значение `27` в область памяти типа `int`, расположенную по адресу `p`, то есть в переменную `x`.

Начинающие часто путают ситуации, в которых указатель слева от присваивания снабжается или не снабжается звёздочкой; поэтому мы приведём ещё один пример. Пусть имеются следующие описания:

```
int x, y;
int *p, *q, *r;
```

Пусть теперь мы сделали присваивания:

[24] Во всяком случае, в *императивных* языках, но других мы пока не знаем.

```
x = 25;
y = 36;
p = &x;
q = &y;
```

Рассмотрим следующий оператор:

```
r = p;
```

Здесь мы *переменной* `r` *присваиваем значение переменной* `p`, а поскольку в переменной `p` находится адрес переменной `x`, после присваивания то же самое — адрес переменной `x` — будет находиться также и в переменной `r`, то есть она тоже будет указывать на `x`. Итак, сейчас у нас `p` и `r` указывают на `x`, а `q` указывает на `y`. Рассмотрим теперь оператор

```
*r = *q;
```

Здесь в левой части присваивания — «то, на что указывает `r`», а в правой — «то, на что указывает `q`», а поскольку они указывают (с учётом вышерассмотренного присваивания) на `x` и `y`, этот оператор занесёт в `x` текущее значение `y`, то есть число `36`. Сами указатели при этом не изменятся.

Казалось бы, в приведённом примере всё ясно; к сожалению, опыт показывает, что далеко не все, кто по тем или иным причинам изучает программирование, могут с ходу разобраться в этих присваиваниях. **Если у вас остались здесь хотя бы малейшие неясности, даже не думайте идти дальше!** В дальнейшем тексте вы попросту вообще ничего не поймёте и только зря потратите время. Попробуйте разобрать этот пример ещё раз, если же это не поможет — обратитесь с вопросами к кому-то, кто сможет вам разъяснить происходящее. В языке Си на указателях основано буквально всё; операция взятия адреса потребовалась нам даже в программе, решающей квадратное уравнение. Если сейчас оставить «в тылу» малейшую неуверенность — уже при чтении следующего параграфа такая неуверенность превратится в фундаментальное непонимание происходящего.

Во многих источниках, посвящённых языку Си, можно встретить утверждение, что символ «`*`» в описании переменной-указателя имеет прямое отношение к операции разыменования. Некие зачатки логики здесь можно уловить, если заметить, что после описания

```
int x;
```

`x` имеет тип `int`, а после описания

```
int *p;
```

`p` имеет тип «указатель», но `*p` (то есть то, что занимает место переменной из предыдущего описания) — действительно имеет тип `int`.

Дело вкуса, но на личный взгляд автора этих строк такую логику никак нельзя считать безупречной. Позже, разбирая более сложные описания, мы убедимся, что звёздочка при прочтении описаний превращается в слова «указатель на», то есть она, будучи встреченной в описаниях, не указатель превращает в простую переменную, а, *напротив*, обычный тип превращает в адресный. Иначе говоря, роль звёздочки в описаниях *прямо противоположна* (!) её роли в выражениях.

В Си можно работать с адресами, не уточняя, на переменные какого типа эти адреса будут указывать. Соответствующий тип указателя называется `void*`; если описать указатель такого типа:

```
void *z;
```

— то в переменную `z` можно будет занести совершенно любой адрес, и такое присваивание компилятор рассматривает как легитимное, не выдавая ни ошибок, ни предупреждений. Более того, разрешено также и присваивание в другую сторону, то есть любому типизированному указателю можно присвоить нетипизированный адрес.

Вспомнив указатели Паскаля, мы можем заметить, что Паскаль предусматривал специальное «адресное» значение, которое с гарантией не является адресом какой-либо ячейки или области памяти; это значение в Паскале обозначалось ключевым словом `nil`, его иногда называют «нулевым указателем», хотя это и не совсем верно: это адрес, а не указатель. В языке Си такое тоже есть, но в зависимости от компилятора и архитектуры конкретное значение такого указателя может быть разным; к счастью, библиотека предусматривает специальное обозначение для такого адресного значения: идентификатор `NULL`[25], который всегда равен соответствующему значению для данного компилятора.

Интересно, что значение адреса можно использовать в качестве логического значения везде, где таковое требуется, в том числе в заголовках операторов ветвления и циклов; при этом «нулевой указатель» (то есть значение `NULL`) считается «ложью», а любой другой адрес — «истиной».

4.3.12. Организация «выходных» параметров функций

Почти в самом начале изучения Си, пытаясь написать программу, решающую квадратное уравнение (см. §4.2.3), мы были вынуждены отметить, что в Си есть только один способ передачи параметров — по значению. В Паскале мы могли применить параметр-переменную, то есть передать в подпрограмму не *значение* переменной, а *её саму*, чтобы подпрограмма могла изменить её значение. В Си в таких случаях приходится, как мы уже обсуждали, передавать адрес переменной, получив его с помощью операции `&`.

В теле функции для обращения к такой переменной приходится, что вполне естественно, использовать обратную операцию — разыменование. Например, функция, меняющая местами значения двух переменных типа `int`, на Си будет выглядеть так:

```
void swap(int *a, int *b)
{
    int t;
    t = *a;
    *a = *b;
    *b = t;
}
```

Вызывать её придётся примерно так:

```
swap(&x, &y);
```

На время работы функции в её *локальных* переменных `a` и `b` окажутся *адреса* переменных `x` и `y`, так что выражение `*a` временно станет обозначать переменную `x`, а выражение `*b` — переменную `y`.

4.3.13. Массивы

Рассказ о массивах в языке Си мы начнём с парадоксального, на первый взгляд, утверждения: **в языке Си массивов нет**. Сразу после такого утверждения мы приведём пример описания массива[26]:

```
int m[20];
```

Здесь описан массив из 20 элементов типа `int`, причём эти элементы доступны с помощью операции индексирования под номерами (индексами) от `0` до `19`. Предвидя недоумение читателя (ведь только что мы заявили, что массивов нет), поясним: элементы массива доступны *каждый в отдельности*, тогда как к самому массиву как единому целому обратиться невозможно. Практически во всех ситуациях введённое в описании имя `m` обозначает не сам массив как таковой, а *адрес его первого элемента*. Поскольку элементы массива, в том числе первый из них, в данном случае имеют тип `int`, имя `m` представляет собой *константу* типа `int*`. Например, если мы опишем указатель

```
int *p;
```

присваивание `p = m;` будет не только возможно, но и не потребует от компилятора никакого преобразования типа: ведь здесь и слева, и справа от присваивания мы видим один и тот же тип `int*`. Больше того, после такого присваивания можно будет обращаться к элементам массива `m` через переменную `p` с помощью всё той же операции индексирования: `p[0]`, `p[1]`, ..., `p[19]` обозначают теперь ровно то же самое, что и `m[0]`, `m[1]`, ..., `m[19]`. С другой стороны, поскольку `m` — это адрес, к нему можно применить операцию разыменования, то есть обратиться к первому элементу массива `m` можно не

[25] Этот идентификатор набирается заглавными буквами; почему это так, мы узнаем позднее из главы о макропроцессоре.

[26] Здесь и далее мы в примерах указываем размерности массивов и некоторые другие константы в явном виде, то есть в виде числа. Мы делаем так, чтобы не загромождать примеры описанием констант, что снизило бы их наглядность; следует помнить, что в тексте настоящей программы (в отличие от иллюстративного примера) числа в явном виде встречаться не должны, все их следует описывать константами.

только через индексирование (`m[0]`), но и иначе: `*m` — это абсолютно то же самое. Мы видим, что индексирование в Си — это операция не над массивами (над массивами вообще нет никаких операций), а над *адресами*.

Чтобы понять, что́ на самом деле представляет собой операция индексирования, нужно пояснить ещё один важный момент: в языке Си к типизированному адресному выражению (то есть к адресному выражению любого типа, кроме `void*`) можно *прибавить целое число*, при этом адрес изменяется на размер элемента того типа, на который указывает исходный адрес; например, адрес типа `char*` при добавлении к нему единицы увеличится на единицу, а адрес типа `int*` в таком же случае — на четыре (ведь `int` занимает 4 байта). Таким образом, если в памяти размещён массив элементов одного типа, добавление единицы к адресу одного из этих элементов даст адрес следующего элемента массива. Иначе говоря, `m` — это адрес первого элемента, `m+1` — адрес второго элемента, и так далее; последний элемент нашего массива имеет адрес `m+19`. Это и есть та самая ***адресная арифметика***, которой часто пугают детей и новичков. Ну а если выражения вида `m+`k представляют собой адреса, то, значит, их можно разыменовывать. Получается, что к элементам массива мы можем обращаться через разыменование: `*m`, `*(m+1)`, `*(m+2)`, ..., `*(m+19)`, и это будет совершенно то же самое, что `m[0]`, `m[1]`, ..., `m[19]`.

Скажем больше: выражение `a[b]`, каковы бы ни были эти `a` и `b`, означает *то же самое*, что и `*(a+b)`. Осознать всю глубину экзотики позволяет замечание, что от перемены мест слагаемых сумма не меняется, так что вместо `m[17]` можно написать `17[m]`, и компилятор это благополучно проглотит. **Не надо так делать!** О наличии этой возможности мы рассказываем только для иллюстрации.

Выше упоминалось, что имя массива представляет адрес первого элемента *почти* во всех ситуациях. Единственная *осмысленная* ситуация, в которой имя массива соответствует всему массиву как единому объекту — это применение к имени массива псевдофункции `sizeof`: в этом случае выдаётся размер всего массива целиком. В частности, `sizeof(m)` для массива из нашего примера будет 80. Пресловутыми «стандартами» предусмотрены ещё две ситуации, когда массив выступает как самостоятельный объект, но их достаточно сложно объяснить, а одна из них к тому же совершенно бессмысленна. Можете предполагать, что `sizeof` — единственное исключение из правил; такое предположение вам ничем не грозит.

Одним из самых заметных (и тяжёлых) следствий недоступности массива как единого объекта можно считать невозможность присваивания массивов, которой часто не хватает, особенно при переходе с Паскаля. Как мы увидим чуть позже, строки представляют собой частный случай массивов — и тоже, разумеется, не присваиваются. Для начинающих, привыкших к высокоуровневой работе со строками в других языках, отсутствие возможности присвоить строку превращается в ночной кошмар, по крайней мере на первых порах.

Отметим ещё один интересный момент. В языке Си определена операция ***вычитания типизированных адресов***, результатом которой является целое число. Эта операция имеет смысл только при применении к адресам элементов одного массива; её результат — *расстояние* между двумя элементами, то есть разность их индексов. Например, если мы работаем всё с тем же массивом `m` и присвоили указателю `p` адрес элемента `m[13]` (то есть выполнили `p = &m[13]` или просто `p = m + 13`), то значением выражения `p - m` будет число `13`. Как можно заметить, эта операция *обратна* операции прибавления целого числа к адресу.

4.3.14. Динамическая память

С ***динамической памятью*** мы уже знакомы по Паскалю, но в Си подход к её использованию существенно отличается. Прежде всего отметим, что *в самом языке*, как водится, средств работы с динамической памятью нет: они вытеснены в библиотеку. Основными функциями считаются `malloc`, выделяющая память, и `free`, освобождающая память. Заголовки этих функций выглядят так:

```
void *malloc(int size);
void free(void *p);
```

Функция `malloc` принимает в качестве параметра размер необходимой области памяти, выделяет такую область из «кучи» и возвращает адрес начала выделенной области. Поскольку функция `malloc` не знает и не может знать ничего о том, для каких целей мы затребовали память, она возвращает нетипизированный адрес, который мы сами можем превратить в адрес того, что нам нужно. Например:

```
double *k;
k = malloc(360*sizeof(double));
```

После выполнения такого присваивания указатель `k` будет указывать на свежевыделенную область памяти, размер которой достаточен для размещения 360 элементов типа `double`. Вспомнив, что индексирование есть операция над адресами, мы обнаруживаем, что с этой областью можно работать как с обычным массивом, обращаясь к его элементам по индексам. Например, следующий цикл заполнит элементы этого массива значениями синуса с интервалом один градус:

```
for(i = 0; i < 360; i++)
    k[i] = sin((2*M_PI/360.0) * (double)i);
```

Функция `free` освобождает ранее выделенную память, делая её вновь доступной для выделения. В качестве параметра эта функция принимает **адрес, ранее возвращённый функцией** `malloc`. Например, освободить наш массив можно так:

```
free(k);
```

Размер области памяти библиотека хранит где-то у себя, так что указывать этот размер при освобождении не требуется. С другой стороны, если функции `free` случайно дать параметром что-то отличное от адреса, возвращённого `malloc`'ом, авария вам практически гарантирована, причём если программа «свалится» непосредственно при вызове `free`, можете считать, что вам крупно повезло; скорее всего, программа успеет ещё некоторое время поработать и только потом завершится с аварией, так что определить, где сделана ошибка, может оказаться весьма непросто.

Стандартная библиотека предусматривает ещё две функции для выделения динамической памяти — `calloc` и `realloc`; мы отложим их рассмотрение до главы 4.10. Отметим, что лучше избегать использования `realloc`, пока вы не научитесь уверенно менять размеры динамических массивов без её помощи.

Заголовки функций для работы с динамической памятью расположены в заголовочном файле `stdlib.h`.

4.3.15. Модификатор `const`

Ключевое слово `const`, пришедшее из языка Си++[27], используется в Си (и в Си++), чтобы *запретить изменение некоторой области памяти*. В отличие от Паскаля, в Си для *описания констант*, то есть именования неких значений, слово `const` не годится: переменные, описанные с этим модификатором, всё же остаются переменными, их просто запрещено изменять.

Обычно константные переменные описывают с инициализатором, хотя компилятор этого и не требует; например:

```
const int iteration_count = 78;
```

Подчеркнём, что введённое таким образом имя — не константа в том смысле, что `iteration_count` нельзя использовать там, где требуется *значение времени компиляции* — например, в `case`-метках оператора `switch`, при указании размерностей массивов и т. п. Всё это делает подобные описания не слишком полезными.

Начиная с C99, «стандарты» разрешают использовать такие константы для задания размера массива; размер массива вообще «стало можно» задавать произвольным выражением, вычисляемым во время исполнения. Эта возможность называется *variable length arrays* (VLA). **Никогда, ни при каких условиях не используйте это!** Если до введения VLA имя локальной переменной всегда представляло собой константное смещение в стековом фрейме, то после их введения имена локальных переменных при переводе в машинный код могут превратиться в сколь угодно сложные арифметические выражения, для вычисления которых может не хватить регистров. В итоге VLA оказываются столь сложны в реализации, что многие компиляторы прибегают к услугам runtime-библиотеки, уничтожая таким образом главное и едва ли не единственное достоинство Си — zero runtime. Кроме того, размеры стека обычно ограничены, и бесконтрольное применение VLA резко увеличивает риск аварии, связанной с переполнением стека. Сама возможность VLA превращает Си в высокоуровневый язык, в качестве которого Си не имеет смысла — в мире есть много более простых, приятных и логичных высокоуровневых языков, не обладающих недостатками Си.

Совсем другое дело получается при использовании слова `const` в описаниях (и объявлениях) указателей. Для начала отметим, что

[27] В частности, это слово принципиально не используется даже в последнем издании книги Кернигана и Ритчи.

```
const int *p;
```

описывает не *константный указатель* (то есть указатель, который нельзя изменять; сам указатель в данном случае изменять как раз очень даже можно), а *указатель на константную область памяти*. Иначе говоря, слово `const` относится здесь не к переменной `p`, а к той области памяти, на которую `p` будет указывать.

Описать неизменяемый указатель тоже можно, для этого нужно поставить слово `const` *после звёздочки*. Конечно, такому указателю нужно прямо в описании задать значение, ведь присваивать ему ничего нельзя. Например:

```
int x;
int * const q = &x;
```

Если вам это понадобилось, свяжитесь с автором книги и расскажите, как вы дошли до такой жизни. Отметим, что едва ли не большинство профессиональных программистов, пишущих на Си, об этой конструкции не знают или не помнят, поскольку она практически никогда не нужна.

В этой своей ипостаси модификатор `const` находит очень активное применение, в особенности в описаниях формальных параметров функций. К примеру, если вы видите заголовок функции

```
void suspicious_func(int *a);
```

— то вы вынуждены подозревать, что такая функция что-то сделает с переданной ей областью памяти, тогда как если параметр снабжён модификатором `const`:

```
void trustworthy_func(const int *a);
```

— то тем самым автор функции заявил вам, пользователю, что изменять память, на которую указывает `a`, его функция не собирается.

По правде говоря, слово `const` ничего не гарантирует, ведь всегда можно изменить тип адресного выражения на любой другой, в том числе не предполагающий константности. При работе с низкоуровневыми языками вообще трудно говорить о каких-либо гарантиях. Так или иначе, слово `const` можно воспринимать как *обещание* не менять соответствующую область памяти, а нарушать собственные обещания, как известно, нехорошо; к тому же за соблюдением такого обещания всё-таки худо-бедно следит компилятор, так что нарушить его можно, лишь применив явно «некрасивый» приём.

4.3.16. Понятие леводопустимого выражения (*lvalue*)

Появление адресов, указателей и операций над ними позволяет по-новому осмыслить различие между обычным (произвольным) ***выражением*** и его частным случаем — ***переменной***. Если, к примеру, взять целочисленную переменную `x` и присвоить ей значение 22, то и «x», и «22» будут представлять собой выражения, причём значением и того, и другого будет число 22, но вряд ли кто-то заявит, что `x` и 22 — это одно и то же. Начать с того, что значение `x` можно поменять (на то она и *переменная*), так что `x` *может стоять слева от присваивания*, а также быть аргументом для операций инкремента и декремента (`++` и `--`).

После введения адресов и операций над ними можно отметить ещё одну черту переменных как выражений «особого сорта»: *к ним можно применять операцию взятия адреса* — попросту говоря, у них *есть* адреса, то есть они просто по своей природе являются областями памяти. О произвольном выражении этого сказать никак нельзя. Кстати, во всех остальных случаях — то есть когда к переменной не применяется ни присваивание, ни инкремент/декремент, ни взятие адреса — она в выражении ничем не отличается от своего текущего значения, то есть такое же выражение, в котором переменная заменена, скажем, простой константой (литералом), равным её значению, гарантированно даст тот же результат.

Ещё при изучении Паскаля мы могли заметить, что всеми особыми свойствами, характерными для переменных, обладают не только имена переменных; простейший пример — элемент массива, что уже превращает левую часть присваивания в выражение произвольной сложности, ведь в индексе может быть что угодно, лишь бы результат получался целочисленным (или того типа, который используется в индексе конкретного массива, ведь в Паскале это может быть произвольный порядковый тип). Но в Паскале это явление не столь наглядно. Операции взятия адреса там вообще изначально не было; функции, возвращающие адрес, возможны, но не столь популярны, как в Си; есть ещё поля записей, а сама запись может идентифицироваться её адресом, что порождает выражения вроде `p[f(x)]^.t^.m`, обладающие всеми свойствами переменных — но и на них обычно внимание не особенно акцентируется, и, кстати, пожалуй, зря, ведь их можно наравне с переменными передавать через var-параметры. Так или иначе, при описании Паскаля все сущности такого рода называются просто «переменными»; когда сложные выражения в этой роли встречаются редко, такая терминология вполне адекватна.

Иное дело — язык Си, в котором слева от присваивания простое имя переменной встречается едва ли не реже, чем что-то другое, а ведь всё остальное — это *выражения*, и можно заметить, что у этих выражений имеется одно интересное свойство, которое не все почему-то осознают. Пусть, например, у нас есть некий массив указателей на целочисленные переменные, а ещё есть простая целочисленная переменная, и мы в эту переменную занесли число 27, а в один из элементов массива поместили её адрес:

```
int *ptrvec[20];
int x;
x = 27;
ptrvec[5] = &x;
```

Рассмотрим теперь выражение `*ptrvec[5]`. Если мы захотим его *вычислить*, то есть получить его *значение*, то на уровне машинного кода произойдёт следующее. Имени `ptrvec` может соответствовать либо константный адрес, если массив объявлен как глобальный или как локально-статический, либо смещение относительно реперной точки стекового фрейма; к этому адресу будет прибавлено число 20 или 40 — размер пяти элементов массива (это в нашем случае указатели, они могут быть 4-байтными или 8-байтными в зависимости от разрядности процессора), по полученному адресу процессор должен будет обратиться к памяти, извлечь оттуда адресное значение (то есть, опять же, 4 байта или 8 — это будет в нашем примере адрес переменной `x`), и уже по этому адресу снова обратиться к памяти, чтобы оттуда добыть искомое значение 27 — это и есть результат вычисления.

Но что если *то же самое выражение* окажется либо слева от присваивания, либо в роли операнда для инкремента, декремента или взятия адреса? В этих случаях узнать, что значением нашего выражения является число 27, явно недостаточно для дальнейшей работы — а то и вовсе бесполезно: простой операции присваивания это значение не нужно ни с какого боку, она его должна *затереть*; операции взятия адреса тоже совершенно всё равно, какое конкретно число лежит в том месте памяти, адрес которого она должна выдать; прочие операции — инкремент, декремент и «сложные» присваивания вроде `+=` или `>>=` — значение 27, конечно, используют, но его одного им будет мало, нужно ещё знать, куда потом положить новое значение. Попросту говоря, этим операциям нужно прежде всего знать, *где находится* их операнд, а не чему он равен. Поэтому компилятор, формируя код, вычисляющий операнд этих «хитрых» операций, *завершит вычисление на шаг раньше* — не будет извлекать из памяти значение по вычисленному адресу, а оставит в качестве результата вычисления сам этот адрес (в нашем случае — адрес переменной `x`).

Далеко не все выражения способны к такому «укороченному» вычислению; скажем, выражение `x+1` так не умеет, ведь в памяти нет такого места, где лежало бы его значение. Но это выражение, очевидно, и не может служить ни левым операндом присваивания, ни единственным операндом инкремента, декремента и взятия адреса. Так мы естественным образом приходим к разделению всех возможных выражений на те, что могут стоять слева от присваивания, и все остальные, которые этого не могут.

Устоявшийся русскоязычный термин для обозначения выражений первого вида — ***леводопустимые выражения***. Интересно, что в английской литературе принят лаконичный термин *lvalue* или (реже) *l-value*, причём буква *l* исходно означала слово *left* из громоздкого словосочетания *left-hand side operand of assignment*. Все остальные выражения — такие, которые в присваивании могут быть только справа — обозначаются столь же коротким термином *rvalue*, в котором, понятное дело, буква *r* соответствует слову *right*.

Проблемы с этими терминами — точнее, с термином *lvalue* — начались после появления в языке Си модификатора `const`. Выражение, в типе которого есть `const`, стоять слева от присваивания уже не может; но такое выражение по-прежнему проявляет едва ли не основное свойство *lvalue* — умение вычисляться не до значения, а до адреса, где значение находится, пусть даже теперь этим свойством может воспользоваться лишь операция

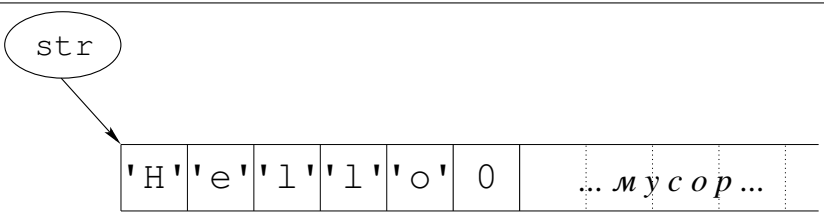

Рис. 4.1. Размещение в памяти строки "Hello"

взятия адреса. К тому же вообще-то защитный механизм компилятора, не позволяющий выполнять присваивание в константные области памяти, можно легко обойти, но реальное присваивание провести можно будет лишь в случае, если есть куда присваивать, и вот это вот «есть куда присваивать» нуждается в обозначении. Поэтому термин *lvalue* решили сохранить для обозначения всех выражений, последним шагом вычисления которых является извлечение итогового значения из памяти — иначе говоря, таких, к которым применима операция взятия адреса, ну а букве *l* «сменили смысл» — теперь она «официально» соответствует не слову *left*, а слову *locator*.

Если вы попытаетесь поискать термины *lvalue* и *rvalue* в Интернете, то неизбежно наткнётесь на довольно многословные и мутные рассуждения, в которых будет фигурировать ещё и *xvalue*. Не пугайтесь, вся эта муть не имеет отношения к Си, речь идёт о языке Си++, причём о его «версиях», придуманных сумасшедшими стандартизаторами; *xvalue* в Си++ появились начиная со «стандарта» C++11. В третьем томе нашей книги мы будем изучать язык Си++, но поделья стандартизационных комитетов при этом проигнорируем, так что термин *xvalue* нам не потребуется.

4.3.17. Инициализаторы для массивов

При описании массива язык Си позволяет задать начальные значения его элементов. Массив при этом изображается в виде перечисления в фигурных скобках, а сами элементы разделяются запятыми. Например:

```
int m[10] = { 2, 3, 5, 7, 11, 13, 17, 19, 21, 23 };
```

Обратите внимание на точку с запятой после закрывающей фигурной скобки, она здесь обязательна — это часть синтаксиса описания. Кроме того, крайне важно понимать ещё один момент: **все элементы инициализатора обязаны быть константами времени компиляции**, то есть среди перечисленных через запятую элементов не должно быть ни переменных, ни вызовов функций, ни таких выражений, которые компилятор не сможет вычислить прямо во время компиляции.

Если для массива предусмотрен инициализатор, то можно не указывать в явном виде его размерность, компилятор вычислит её сам:

```
int m[] = { 2, 3, 5, 7, 11, 13, 17, 19, 21, 23 };
```

Здесь размерность массива будет, как и в прошлом случае, составлять десять элементов, поскольку именно столько элементов предусмотрено в инициализаторе. Так сделано, чтобы избавить программиста от необходимости подсчитывать количество элементов вручную; узнать, какой размер получился, можно, поделив размер массива на размер его элемента, то есть написав `sizeof(m)/sizeof(m[0])` или `sizeof(m)/sizeof(*m)`. Например, следующий цикл распечатает все элементы массива:

```
for(i = 0; i < sizeof(m)/sizeof(*m); i++)
    printf("[%d] = %d\n", i, m[i]);
```

На примере инициализаторов для массивов наглядно видно, что **инициализация и присваивание — это совершенно разные вещи**. Массиву *нельзя* таким способом ничего присвоить, когда он уже существует, то есть конструкция из фигурных скобок и списка значений не является в Си выражением и не может встречаться в других выражениях, в том числе и в присваивании; она может встречаться только в *описании* массива — в роли инициализатора.

Инициализировать можно не только глобальные, но и локальные массивы, то есть можно написать что-то вроде

```
int f(int x)
{
    int m[] = { 2, 3, 5, 7, 11, 13, 17, 19, 21, 23 };

    /* ... */

}
```

Прежде чем воспользоваться этой возможностью, следует осознать, что всё это отнюдь не бесплатно: *данные будут копироваться в область памяти в стековом фрейме каждый раз в начале выполнения такой функции*. В некоторых случаях это не страшно, но чаще всего так делать не стоит.

4.3.18. Строки

Как мы уже знаем, ***строки*** (или, говоря шире, ***данные в текстовом представлении***) знамениты непредсказуемостью своего размера: в большинстве случаев во время написания программы мы можем лишь приблизительно прикинуть, какого количества памяти «уж точно хватит» для размещения строки, и в итоге с хорошей вероятностью обнаружим, что ошиблись. Память обычно выделяется с некоторым запасом, а во время исполнения оказывается, что строка занимает меньше памяти, чем под неё выделено; как следствие, при обработке строк требуется какой-то способ указания текущей длины строки.

В частности, при работе со строками на Паскале под хранение длины строки выделяется особый байт; если рассматривать паскалевскую строку как массив символов, то её длина хранится в элементе этого массива, имеющем индекс `0`. Из-за этого паскалевские строки не могут превышать в длину 255 символов.

В Си применяется другой подход. Строки здесь тоже представляются массивами символов, то есть массивами элементов типа `char`, при этом *корректной строкой* считается только такой массив типа `char`, в котором хотя бы один элемент равен нулю; он рассматривается как *ограничитель*, то есть все элементы, предшествующие нулевому, считаются составляющими строку, а всё, что находится в массиве после нулевого элемента, игнорируется (см. рис. 4.1). Конечно, в массиве может оказаться больше одного элемента, равного нулю; в этом случае в качестве ограничителя рассматривается тот из них, индекс которого меньше, чем у других, или, иначе говоря, ближний к началу массива. Длина строки, таким образом, есть разница между положением ограничителя (нулевого элемента) и начала строки.

Если сравнивать этот подход с принятым в Паскале, в качестве его несомненного достоинства следует отметить отсутствие априорного ограничения на длину строк: можно работать со строками длиной хоть в миллион символов, а если очень надо, то и больше, лишь бы хватило оперативной памяти. С другой стороны, очевидны и недостатки этого подхода: во-первых, для определения длины строки её приходится просматривать всю в поисках нулевого элемента, тогда как для паскалевской строки достаточно было извлечь байт из её начала; во-вторых, сама строка не может, очевидно, содержать символ с кодом `0`. Впрочем, считается, что «символ» с таким кодом не может встречаться в текстовых данных; в частности, если в файле встретился нулевой байт, то этот файл заведомо не текстовый.

Рассмотрим несколько примеров. Для начала напишем функцию, которая получает на вход строку и определяет её длину. Начнём с заголовка. С типом возвращаемого значения всё более-менее понятно, это ведь длина строки в символах, то есть обыкновенное целое число. Интереснее обстоят дела с параметром. На вход по условию задачи необходимо получить строку, но как это сделать? Строка — это частный случай массива символов, а мы уже знаем, что никакого способа обратиться к массиву как единому объекту язык Си не предусматривает; мы ведь по этой причине даже заявляли, что массивов в языке Си вообще нет. Тем не менее, с массивами можно работать благодаря адресной арифметике, частным проявлением которой является операция индексирования, что же до самого массива, то доступ к нему осуществляется через *адрес первого элемента*. Строка у нас состоит из элементов типа `char`, так что параметр, через который она будет передаваться в функцию, должен быть типа «адрес `char`'а». Давайте, однако, не спешить с описанием параметра типа `char*`; есть ещё одно соображение, влияющее на этот тип. Очевидно, что для вычисления длины строки не нужно вносить в эту строку какие-либо изменения; этот факт стоит отразить в типе параметра функции, воспользовавшись модификатором `const` (см. §4.3.15).

Применять `const` в таких случаях следует обязательно, этот момент намного важнее, чем может показаться, ведь та строка, длину которой нам нужно посчитать (и вообще произвольная область памяти, где содержатся какие-то данные), может сама оказаться доступна только на чтение. Как правило, такая ситуация возникает, когда нужно применить нашу функцию в какой-то другой функции, которая сама получает «константный» параметр.

Если же говорить именно о строках, стоит припомнить, что **компилятор размещает строковые литералы в неизменяемой памяти**. Язык Си++, до которого мы рано или поздно обязательно доберёмся, рассматривает строковые литералы — строки, записанные в двойных кавычках — как выражения типа `const char*`, так что, если мы *в программе на Си++* напишем без `const` функцию, получающую на вход строку, её нельзя будет применять к строкам, заданным явно; это касается в том числе и случая, когда модуль, экспортирующий такую функцию, написан целиком на чистом Си, но кому-то (скорее всего, нам самим) потребовалось задействовать его в программе на Си++.

В чистом Си модификатора `const` изначально не было, он пришёл из Си++; чтобы не нарушать совместимости с исходными текстами, написанными до появления `const`, компиляторы чистого Си считают строковые литералы имеющими тип `char*` без всякого `const`. Конечно, это никоим образом не означает, что внутрь строковых литералов можно что-то записывать во время работы программы; к этому вопросу мы чуть позже вернёмся. В действительности такая «неконстантность» строковых литералов значит ровно одно: ошибки, которые компилятор Си++ может поймать во время компиляции, при работе на Си практически неизбежно переносятся на время выполнения программы. Впрочем, компилятор `gcc` поддерживает параметр командной строки `-Wwrite-strings`, при наличии которого строковые литералы получают `const` в свой тип, так что при присваивании строковых литералов указателям без `const`, инициализации таких указателей строковыми литералами, а также при передаче строковых литералов в функции через параметры, не имеющие `const`, компилятор всё-таки выдаёт предупреждение.

Так или иначе, заведите себе привычку всегда использовать `const` при описании параметров адресного типа, когда указуемый объект не предполагается изменять внутри функции.

Итак, нам нужно отразить в заголовке функции, что она возвращает целое число, а на вход принимает адрес `char`'а, и притом не собирается изменять область памяти по этому адресу. Назвав нашу функцию `string_length`, мы сможем начать её описание примерно так:

```
int string_length(const char *str)
```

Теперь нам нужно просмотреть область памяти, начало которой располагается по адресу `str`, в поисках нулевого элемента типа `char`, который, заметим, обязан там присутствовать по условию задачи, иначе это была бы не строка. Наш первый вариант решения будет выглядеть совершенно по-паскалевски: мы просто вспомним, что, коль скоро указатель `str` указывает на массив, то с этим указателем можно и работать как с массивом, то есть применять к нему операцию индексирования; напомним ещё раз, что индексирование в Си — это операция над адресом, а не над массивом. Выглядеть это всё может примерно так:

```
int string_length(const char *str)
{
    int i;
    i = 0;
    while(str[i] != '\0')
        i++;
    return i;
}
```

Текущая рассматриваемая позиция здесь хранится в переменной `i`, мы начинаем рассмотрение с нулевой позиции, если в текущей позиции в массиве оказался ноль — возвращаем номер этой позиции в качестве результата: нетрудно видеть, что, поскольку нумерация позиций начинается с нуля, длина строки в точности равна первой позиции, которая в строку не входит: например, если ограничитель встретился в нулевой позиции, то строка пустая и её длина, соответственно, равна нулю, а если, скажем, в третьей — то строка состоит из трёх символов, находящихся в нулевом, первом и втором элементах массива.

Впрочем, опытные программисты на Си напишут то же самое совсем иначе. Во-первых, выражение `str[i]` никто не мешает использовать в качестве логического, а не сравнивать его зачем-то с нулём. Во-вторых, если внимательно посмотреть на цикл `while`, можно заметить, что непосредственно перед ним записано присваивание переменной цикла начального значения, а последним (и единственным) оператором в теле цикла является изменение значения переменной цикла на то, которое будет использовано в следующей итерации. В подобных ситуациях в Си используется цикл `for`. Исправив оба указанных недочёта, мы получим следующее:

```
int string_length(const char *str)
{
    int i;
    for(i = 0; str[i]; i++)
        {}
    return i;
}
```

Тело цикла здесь получилось пустым, что для Си вполне типично. Однако наверняка найдутся и те, кто заявит, что здесь слишком много операций сложения: на каждой итерации, во-первых, увеличивается на единицу переменная `i`, а во-вторых, её значение прибавляется к указателю `str`: в самом деле, `str[i]` есть, как мы уже знаем, строго то же самое, что и `*(str+i)`. Сэкономить одно сложение можно, если в качестве переменной цикла использовать не индекс текущего элемента строки, а указатель на него; вычислить результат в конце мы сможем благодаря операции вычитания адресов. Наша функция теперь примет следующий вид:

```
int string_length(const char *str)
{
    const char *p;
    p = str;
    while(*p)
        p++;
    return p - str;
}
```

Можно написать и так:

```
int string_length(const char *str)
{
    const char *p;
    for(p = str; *p; p++)
        {}
    return p - str;
}
```

Среди «прожжёных сишников» кто-то наверняка предпочтёт следующий вариант:

```
int string_length(const char *str)
{
    const char *p = str;
    while(*p++);
    return p - str;
}
```

Никогда так не делайте! Во-первых, тело цикла, даже если оно пустое, следует записать на отдельной строке; во-вторых, пустое тело правильнее обозначать пустым блоком `{}`, а не одинокой точкой с запятой, которую на фоне окружающих операторов просто не видно; в-третьих, **прежде чем загнать *побочный эффект* в условное выражение, подумайте трижды**; в данном случае сделано именно это, операция `++` меняет переменную `p`, это и есть побочный эффект. Чтобы стало понятно, почему так делать не следует, попробуйте обнаружить в этом коде ошибку и засеките, сколько времени у вас на это уйдёт. На всякий случай подчеркнём, что ошибка тут есть.

В стандартной библиотеке языка Си присутствует функция `strlen`, решающая ровно эту задачу. Объявление (заголовок) этой функции находится в заголовочном файле `string.h`. Если определение длины строки потребуется вам в «настоящей» (не учебной) программе, лучше использовать функцию из библиотеки — так ваша программа будет понятнее стороннему читателю; но пока вы только учитесь, от использования функций из `string.h` вам настоятельно рекомендуется воздержаться, иначе вы так и не поймёте, что такое строка в Си.

В качестве следующего примера рассмотрим функцию копирования строки в заранее отведённую область памяти; будем предполагать, что памяти отведено достаточно, то есть за это несёт ответственность вызывающий. Мы назовём её `string_copy`. Возвращать этой функции вроде бы нечего, поэтому в качестве типа возвращаемого значения укажем `void`. В качестве параметров функция будет получать адрес области памяти, *куда* следует скопировать строку, а также, собственно, адрес строки; параметры мы по традиции назовём `dest` и `src` от слов *destination* и *source*. Порядок параметров выберем по аналогии с операцией присваивания: сначала *куда*, потом *откуда*. Заметим, что оригинал строки функция не меняет, а вот область памяти, куда производится копирование — очевидно, меняет; с учётом этого заголовок получается таким:

```
void string_copy(char *dest, const char *src)
```

Начнём, как и в предыдущем примере, с варианта «по-паскалевски»; объявим переменную для текущей позиции и применим к обоим массивам операцию индексирования, но вариант с `while` пропустим и перейдём сразу к использованию `for`:

```
void string_copy(char *dest, const char *src)
{
    int i;
    for(i = 0; src[i]; i++)
        dest[i] = src[i];
    dest[i] = 0;
}
```

Обратите внимание на последнюю строчку; она нужна, чтобы превратить массив `dest` в корректную строку. Дело в том, что для значения `i`, соответствующего позиции ограничивающего нуля в исходной строке, тело цикла выполнено уже не будет, так что ноль не будет скопирован.

Следующая версия той же функции будет использовать указатели, чтобы сэкономить на сложениях; при этом мы воспользуемся тем, что параметры функции внутри неё представляют собой простые локальные переменные[28], которые вполне можно менять. В результате мы сможем обойтись без введения дополнительных локальных переменных:

```
void string_copy(char *dest, const char *src)
{
    while(*src) {
        *dest = *src;
        dest++;
        src++;
    }
    *dest = 0;
}
```

Или даже так:

```
void string_copy(char *dest, const char *src)
{
    for(; *src; dest++, src++)
        *dest = *src;
    *dest = 0;
}
```

Вообще-то так делать не очень хорошо; многие программисты, включая автора книги, придерживаются мнения, что имена формальных параметров следует использовать только для коммуникации с вызывающим, но никак не для локальных нужд. Другие программисты с этим не соглашаются по принципу «а *чо* плохого-то». Так или иначе, шанс нарваться на подобное в чужом коде достаточно велик, будьте к этому готовы.

Рассмотрение задачи копирования строки не будет полным без примера, представляющего подлинный образчик «истинно сишного» программирования. Оставляем сей пример для самостоятельного анализа в надежде, что время, потраченное на распутывание этого ребуса, позволит читателю в будущем не увлекаться подобными трюками:

```
void string_copy(char *dest, const char *src)
{
    while((*dest++ = *src++));
}
```

Отметим, что вторые круглые скобки мы здесь поставили, чтобы компилятор не выдавал нам предупреждение об использовании присваивания в качестве логического значения, что в большинстве случаев (но не в этом) свидетельствует об ошибочном использовании знака присваивания вместо знака сравнения (двойного равенства).

Функция стандартной библиотеки, выполняющая копирование строки, называется `strcpy`; всё, что в примечании на стр. 33 сказано про функцию `strlen`, справедливо также и для `strcpy`, и вообще для всех функций, описанных в заголовочном файле `string.h`.

Следует заметить, что все вышеприведённые версии `string_copy` написаны в предположении, что области памяти `dest` и `src` не пересекаются. Для пересекающихся областей памяти функция побайтового копирования будет чуть сложнее: в зависимости от их взаимного расположения копирование может потребоваться в прямом или обратном порядке.

4.3.19. Строковые литералы

Разобравшись, что представляют собой строки в языке Си, мы сможем теперь понять, что конкретно делает компилятор при виде строки, заключённой в двойные кавычки. Прежде всего компилятор отводит где-то в *неизменяемой области памяти* (чаще всего — в сегменте кода, то есть в секции `.text`) нужное количество памяти, по одному байту на символ и один байт на завершающий ноль, и соответствующим образом эту область памяти заполняет. Например, видя литерал `"Hello"`, компилятор отведёт *шесть* ячеек памяти, в первую занесёт ASCII-код буквы `H`, во вторую — код буквы `e`, а в последнюю — ноль.

Неизменяемая область памяти выбирается из соображений эффективности. Как правило, изменение строкового литерала во время работы программы не нужно, а при одновременном запуске нескольких экземпляров одной и той же программы современные операционные системы создают в памяти только один экземпляр неизменяемой части программы (в том числе сегмента кода). Как следствие, размещение строковых литералов в неизменяемой области памяти позволяет при работе нескольких копий одной программы иметь в памяти один экземпляр всех её строковых констант. В системах семейства Unix это особенно актуально, поскольку новые процессы здесь создаются путём копирования существующего процесса.

Разместив содержимое строкового литерала в генерируемом объектном модуле, компилятор вместо него подставляет, как можно догадаться, *адрес первого элемента*. Иначе говоря, в контексте выражения, где встречен строковый литерал, он имеет тип `char*` и представляет адрес той области (неизменяемой) памяти, где расположено его содержимое.

Как уже отмечалось в предыдущем параграфе, в языке Си++ в аналогичной ситуации тип адреса снабжается модификатором `const`, поскольку попытки изменения внутренностей строкового литерала ни к чему хорошему не приведут, а возможные ошибки правильнее будет выявить при компиляции, чем во время работы. К сожалению, в чистом Си из соображений совместимости со старыми диалектами строковые литералы сохраняют тип `char*` без всякого `const`.

Между прочим, возможно даже такое выражение: `"Abrakadabra"[4]`; оно будет иметь тип `char`, а его значением будет код буквы `k`. В самом деле, литерал `"Abrakadabra"` есть адрес, а к адресу можно применить операцию индексирования. Некоторые программисты пытаются применять выражения подобного рода (конечно, с переменной величиной в качестве индекса, иначе это было бы вовсе лишено смысла), но лучше так не делать, поскольку ясности вашей программе подобные трюки не добавят. Также возможно и выражение `*"Abrakadabra"`, которое равно коду буквы `A`; его применение может преследовать только одну цель: специально запутать программу.

Стоит упомянуть один важный особый случай. **Строковый литерал может применяться в качестве инициализатора массива** элементов типа `char`, `signed char` и `unsigned char`, и при этом, естественно, компилятор не станет размещать массив в неизменяемой памяти (кстати, даже в том случае, если массив будет объявлен со словом `const`). Например, описание

```
char str[6] = "Hello";
```

создаст массив из шести элементов — точно так же, как если бы мы написали

```
char str[6] = { 'H', 'e', 'l', 'l', 'o', 0 };
```

Число, обозначающее размерность массива, можно, как обычно при наличии инициализатора, опустить:

```
char str[] = "Hello";
```

Здесь размерность массива составит шесть элементов, так как именно столько подразумевается строковым литералом (пять — длина строки, плюс один на ограничительный ноль). Подчеркнём, что это совершенно не то же самое, что написать

```
char *ptr = "Hello";
```

[28] Вспомните структуру стекового фрейма, которую мы подробно рассматривали при изучении программирования на языке ассемблера.

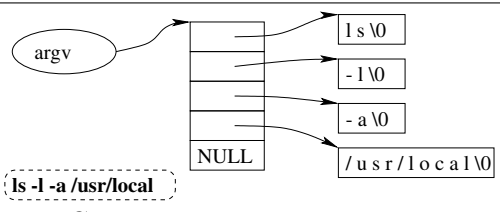

Рис. 4.2. Структура данных командной строки

В этом случае описывается не массив, а указатель, который инициализируется адресом строкового литерала, тогда как сам литерал при этом воспринимается компилятором согласно общим правилам, и всё сказанное про неизменяемые области памяти снова обретает силу.

Самое очевидное проявление различия между вышеописанными `str` и `ptr` состоит в том, что `str` — не переменная в том смысле, что ей нельзя ничего присваивать, тогда как `ptr` — обыкновенная переменная типа «указатель». Кроме того, `sizeof(ptr)` будет равен размеру указателя (то есть 4 на 32-битных системах и 8 на 64-битных), тогда как `sizeof(str)` будет равен размеру области памяти, занятой массивом, в данном случае `6`. Наконец, массив `str` можно изменять, то есть, например, присваивание `str[4]=0` превратит слово `Hello` в слово `Hell`; если же попытаться сделать `ptr[4]=0`, это приведёт к попытке записи внутрь строкового литерала, так что ваша программа завершится аварийно. Настоятельно рекомендуется попробовать так сделать, много времени это не займёт и компьютер от такой ошибки не взорвётся, а вы будете знать, как всё это выглядит на практике.

4.3.20. Аргументы командной строки

Как мы видели ещё при изучении языка ассемблера (см. т. 1, §3.6.8), при запуске программы операционная система отводит в её адресном пространстве специальную область памяти, в которой располагает слова, составляющие командную строку. Получить доступ к этой области памяти из программы, написанной на Си, можно, описав функцию `main` как *имеющую параметры*, причём первый из этих параметров должен быть типа `int` — он задаёт количество элементов командной строки. Второй параметр представляет собой адрес начала массива из указателей на сами строки (слова), составляющие командную строку; ясно, что каждый элемент этого массива имеет тип `char*`, ну а адрес первого (как и любого) элемента этого массива будет уже типа `char**`. Обычно эти аргументы функции `main` называются `argc` и `argv` от слов *argument count* и *argument vector*. Заголовок функции `main` с учётом этого принимает следующий вид:

```
int main(int argc, char **argv)
```

На рис. 4.2 показана структура данных, доступная через указатель `argv`, на примере команды `ls -l -a /usr/local`. Параметр `argc` в этом случае будет равен четырём по количеству значащих элементов командной строки, но в массиве, как можно заметить, элементов на один больше: последний из них (в данном случае пятый) равен «нулевому адресу» `NULL`. Например, следующая программа напечатает свои аргументы командной строки, взяв каждый из них для наглядности в квадратные скобки:

```
#include <stdio.h>
int main(int argc, char **argv)
{
    int i;
    for(i = 1; i < argc; i++)
        printf("[%s]\n", argv[i]);
    return 0;
}
```

Комбинация `%s` в форматной строке `printf` (`s` от слова *string*) означает печать строки, расположенной в памяти по адресу, который берётся из очередного аргумента. Результат работы программы будет выглядеть, например, так:

```
avst@host:~$ ./cmdprint abra kadabra shvabra
[abra]
[kadabra]
[shvabra]
avst@host:~$ ./cmdprint 1       2    " 3   4 " 5
[1]
[2]
[ 3   4 ]
[5]
avst@host:~$
```

Написать программу `cmdprint` можно и иначе, не используя `argc`, а вместо этого воспользовавшись наличием `NULL` в конце массива указателей, например:

```
#include <stdio.h>
int main(int argc, char **argv)
{
    char **p = argv + 1;
    while(*p) {
        printf("[%s]\n", *p);
        p++;
    }
    return 0;
}
```

Следующая программа напечатает имя, по которому её вызвали:

```
#include <stdio.h>
int main(int argc, char **argv)
{
    printf("My name is %s\n", argv[0]);
    return 0;
}
```

Например:

```
avst@host:~$ ./printname
My name is ./printname
avst@host:~$
```

4.3.21. (*) Точки следования (*sequence points*)

Возможность наличия в одном выражении нескольких побочных эффектов поднимает вопрос о том, в какой конкретно последовательности эти эффекты произойдут. Проще всего представить, что все побочные эффекты случатся ровно в той последовательности, в которой (вроде бы) должны вычисляться части выражения; но не тут-то было. В Си побочные эффекты операций, входящих в одно арифметическое выражение, могут произойти в такой хитрой последовательности, что, если не знать, как всё устроено, можно случайно поверить в демонов.

Самый простой пример этого — последовательность вычисления аргументов функции. Рассмотрим для примера выражение `f(h1(), h2(), h3())`. Как ни странно, здесь про последовательность вызовов функций можно достоверно сказать только одно: функция `f` будет вызвана никак не раньше, чем отработают `h1`, `h2` и `h3`, что и понятно — для вызова функции `f` нужно, чтобы все её аргументы уже вычислились. Но на этом всё. Новички часто предполагают, что `h1`, `h2` и `h3` будут вызваны прямо так, слева направо; более опытные люди вспоминают, что вроде бы аргументы помещаются в стек справа налево, так что последовательность будет противоположной; но и те, и другие, увы, заблуждаются. Последовательность вызовов `h1`, `h2` и `h3` здесь *не определена*; больше того, один и тот же компилятор, транслируя одну и ту же программу, может в разных её местах применить разную последовательность вычислений, исходя из своих «тайных» соображений, обычно обусловленных оптимизацией; кстати, при изменении уровня оптимизации последовательность может измениться.

Это отнюдь не конец истории. Последовательность вычисления операндов арифметических операций тоже в большинстве случаев не определена (например, в выражении `f()+g()+h()` последовательность выполнения функций `f`, `g` и `h` может быть любой); есть несколько «особенных» операций, для которых последовательность зафиксирована, к этому вопросу мы вернёмся чуть позже. Что самое неприятное — это то, что **для операций, изменяющих значение переменной** — то есть для обычного присваивания, для всех «хитрых» присваиваний вроде `+=` или `&=`, а также для операций инкремента и декремента — **не определён тот момент, когда новое значение переменной будет записано в память**. Классический пример на эту тему, приведённый в книге Кернигана и Ритчи, выглядит так:

```
a[i] = i++;
```

Если, скажем, переменная `i` имела значение `7`, то это значение как раз и будет записано в элемент массива (в самом деле, ведь значением `i++` как выражения является *старое* значение переменной), только при этом *неизвестно, в какой* элемент массива оно будет записано — в седьмой или в восьмой. Компилятор может сделать и так, и так; точнее говоря, компилятор волен физически записать в память, занятую переменной `i`, её новое значение как *до* вычисления `a[i]`, так и *после*.

Рассмотрим ещё пример:

```
int a, b;
a = 1;
b = (a+=5) + (a*=2);
```

Здесь есть прямо-таки масса возможностей для значений переменных `a` и `b` после выполнения второго оператора. Самый простой и понятный сценарий таков: сначала переменная `a` увеличивается на `5`, результат (и её собственное значение, и значение, возвращённое операцией `+=`) получается `6`, потом она увеличивается вдвое с результатом `12`, этот результат в ней и остаётся, а операция сложения складывает `6` и `12`, так что в `b` помещается значение `18`. Можете быть уверены, так ваш компилятор не сделает — это было бы слишком просто.

Начнём с того, что операнды сложения могут быть вычислены в обратном порядке, так что сначала `a` увеличится вдвое, а потом уже на пять, и получит итоговое значение `7`, ну а `b` тогда, наверное, должно стать равным девяти. Но это тоже выглядит слишком просто. Обе операции присваивания, работающие с `a`, имеют полное право отложить запись её значения «до лучших времён», так что каждая из них начнёт со значения `1`; переменная `a` в итоге станет равна то ли двум, то ли шести, в зависимости от того, какое из двух значений будет записано последним, что до `b` — то, по-видимому, в ней окажется `8`.

Автор должен признать, что компилятор `gcc` (той версии, что была под рукой; никто, разумеется, не поручится за *все* версии) превзошёл его ожидания: в переменной `a` оказалось `12` (даже странно — это самое ожидаемое значение для неё, если не задумываться о всевозможных оптимизациях), а вот `b` оказалась равна `24`. Чтобы понять, как такое может быть, пришлось припомнить, что *формально* результат операции присваивания равен значению переменной, стоявшей в нём слева. Присваивания переменной `a` прошли последовательно одно за другим в «естественном» порядке, но вот *значения* этих присваиваний как выражений компилятор никуда, кроме переменной `a`, заносить не стал, а перед сложением просто обратился к ней дважды. Там, естественно, оба раза оказалось одно и то же, так что результатом всей правой части стало `12+12`. К чести `gcc` следует заметить, что при компиляции он выдал предупреждение.

Несколько снизить остроту проблемы позволяют ***точки следования***; английский оригинал термина — *sequence points*, иногда встречается перевод «точки гарантированных вычислений». В каждой такой «точке» компилятор доводит до конца все побочные эффекты, которые он мог ранее «отложить». Самая простая из них — *конец выражения*. Это касается прежде всего выражения в ***операторе вычисления выражения***: дошли до точки с запятой — всё, что должно было случиться, случилось; но точно так же законченными считаются все выражения, предполагаемые синтаксисом *любых* операторов, то есть это условные выражения в `if`, `while` и `do-while`, выражение в заголовке `switch`, выражение, задающее возвращаемое значение в операторе `return`, а также все три выражения в заголовке `for` (каждое по отдельности).

Бывают точки следования и в рамках одного выражения. Во-первых, это логические связки `&&` и `||`: в них всегда сначала вычисляется выражение, стоящее слева, его вычисление доводится до конца со всеми его побочными эффектами, и только потом, *если нужно*[29], вычисляется правый аргумент. Тернарная ***условная операция*** (`?:`) тоже предполагает точку следования: прежде чем будет вычисляться её второй или третий операнд, компилятор доведёт до завершения вычисление первого. Аналогично действует и операция «запятая»: её операнды всегда вычисляются слева направо, и перед началом вычисления правого операнда завершаются все побочные эффекты левого. Наконец, перед вызовом функции компилятор всегда завершает вычисление всех её аргументов (хотя конкретная последовательность их вычисления неизвестна).

Есть и другие ситуации, в которых компилятор гарантирует финализацию всех вычислений, но мы их обсуждение опустим; вместо этого предложим читателю осознать один простой момент: **если в вашей программе в каждом выражении будет не больше одного побочного эффекта, весь текст этого параграфа можно с чистой совестью выкинуть из головы**. К рекомендациям о том, как следует обращаться с побочными эффектами, мы вернёмся в §4.8.3.

4.3.22. Избранные примеры программ

Прежде чем идти дальше, рассмотрим два интересных примера. Начнём с программы, которую мы обещали написать ещё во вводной части (см. т. 1, §1.3.5).

Ханойские башни

Подробно мы эту задачу уже исследовали при изучении Паскаля (см. т. 1, §2.11.2) и убедились в том, что её лучше не пытаться решать без применения рекурсии — получается сложно и некрасиво. Здесь мы ограничимся только рекурсивным решением, которое будет практически дословно повторять решение на Паскале, приведённое в первом томе в §2.11.2. Выглядеть программа будет так:

```
#include <stdio.h>                                  /* hanoi.c */
#include <stdlib.h> /* for atoi */
void solve(int source, int target, int interm, int n)
{
    if(n == 0)
        return;
    solve(source, interm, target, n-1);
    printf("%d: %d -> %d\n", n, source, target);
    solve(interm, target, source, n-1);
}

int main(int argc, char **argv)
{
    int n;
    if(argc < 2) {
        fprintf(stderr, "No parameter given\n");
        return 1;
    }
    n = atoi(argv[1]);
    if(n < 1) {
        fprintf(stderr, "Incorrect token count\n");
        return 2;
    }
    solve(1, 3, 2, n);
    return 0;
}
```

Подробные пояснения мы опускаем, поскольку уже привели их для программы на Паскале; обратите внимание, насколько две программы получились схожими.

Читатель, несомненно, догадался, что `fprintf(stderr, ...)` производит вывод в поток диагностики, тем более что слово `stderr` мы упоминали в части о Паскале как синоним для `ErrOutput`; функция `fprintf` работает примерно так же, как и обычный `printf`, но принимает дополнительный (первый) аргумент, задающий поток вывода. Позже мы обсудим всё это подробнее; в этом примере нам не хотелось выдавать диагностические сообщения в поток стандартного вывода, чтобы не возникало на ровном месте отличия этой программы от её паскалевского аналога, да ещё и в худшую сторону.

Сопоставление строки с образцом

С задачей сопоставления строки с образцом мы дважды сталкивались в первом томе: в §2.11.3 мы решили эту задачу на Паскале, в §3.3.9 решение той же задачи на языке ассемблера использовалось в качестве примера организации рекурсии на уровне команд процессора. Решение на Си благодаря использованию арифметики указателей оказывается, пожалуй, несколько элегантнее, чем на Паскале; что касается ассемблерного решения, то следующий код на Си повторяет его практически слово в слово:

[29] Если у вас вызывает неуверенность словосочетание «если нужно», перечитайте рассказ о логических связках на стр. 19.

```
/* match_c.c */
int match(const char *str, const char *pat)
{
    int i;
    for(;; str++, pat++) {
        switch(*pat) {
        case 0:
            return *str == 0;
        case '*':
            for(i=0; ; i++) {
                if(match(str+i, pat+1))
                    return 1;
                if(!str[i])
                    return 0;
            }
        case '?':
            if(!*str)
                return 0;
            break;
        default:
            if(*str != *pat)
                return 0;
        }
    }
}
```

Впрочем, для большей ясности решения его можно разбить на две взаимно рекурсивные функции. Основная по-прежнему будет называться `match`, а цикл, организуемый при нахождении в образце звёздочки, мы выделим в отдельную функцию, которую назовём `starmatch`. Поскольку функции вызывают друг друга, для одной из них придётся сделать прототип; мы напишем прототип для `starmatch`, затем реализуем `match`, и лишь после этого напишем, наконец, тело `starmatch`:

```
/* match_c2.c */
int starmatch(const char *str, const char *pat);
int match(const char *str, const char *pat)
{
    switch(*pat) {
        case 0:
            return *str == 0;
        case '?':
            if(!*str)
                return 0;
            return match(str+1, pat+1);
        case '*':
            return starmatch(str, pat+1);
        default:
            if(*str != *pat)
                return 0;
            return
                match(str+1, pat+1);
    }
}
int starmatch(const char *str, const char *pat)
{
    int i;
    for(i=0; ; i++) {
        int res = match(str+i, pat);
        if(res)
            return 1;
        if(!str[i])
            return 0;
    }
}
```

4.4. Стандартные функции ввода-вывода

Отвлечёмся на некоторое время от возможностей языка Си и попробуем освоить небольшую часть стандартной библиотеки, которую обычно называют ***средствами высокоуровневого ввода-вывода***. Все возможности, которые мы рассмотрим в этой главе, покрываются всё тем же заголовочным файлом `stdio.h`.

4.4.1. Посимвольный ввод-вывод

Напомним, что постоянно попадающиеся нам «ввод с клавиатуры» и «вывод на экран» правильнее было бы называть вводом *из стандартного потока ввода* и выводом *в стандартный поток вывода*, ведь никто не может нам гарантировать, что по ту сторону стандартных потоков находятся действительно клавиатура и экран; в общем случае мы даже не можем с уверенностью это определить[30]. Именно так мы и будем говорить; «вывод на экран и ввод с клавиатуры» оставим тем, кто не собирается становиться программистом. Больше того, поскольку этот и следующий параграфы посвящены только стандартным потокам, мы не будем этот момент уточнять, пока не придёт время работы с произвольными потоками, а не только стандартными; пока же под словами «ввод» и «вывод», а равно «чтение» и «запись» мы будем понимать соответствующие операции над стандартными потоками ввода-вывода.

Начнём с *ввода символа* или, точнее говоря, ввода очередной порции данных, соответствующей минимальному адресуемому на данной машине; в реальной жизни это всегда один восьмибитный байт. Функция, которая выполняет эту операцию, называется `getchar`; её профиль выглядит так:

```
int getchar();
```

Как видим, она не получает параметров, а возвращает значение типа `int`, то есть целое число. Но почему не `char`, или `signed char`, или `unsigned char`? По смыслу ведь они, как кажется, подходят лучше? Действительно, если читается один байт, разрядности любого из `char`'ов хватило бы для представления прочитанного значения, но только при условии, что *чтение прошло успешно* и очередное значение действительно прочитано. Проблема здесь не только и не столько в том, что может произойти ошибка: на стандартном вводе ошибки происходят редко. Мы знаем одну *абсолютно штатную* ситуацию, когда никаких ошибок не происходит, но очередного байта в потоке ввода нет и прочитать оттуда нечего: это ***ситуация конца файла***. Если пользователь, запустивший нашу программу, перенаправил нам стандартный ввод, «подсунув» вместо клавиатуры обыкновенный файл, ситуация конца файла возникнет (спасибо Капитану Очевидность), когда файл будет весь прочитан, то есть он кончится; но даже если на ввод нашей программе поступают данные, которые живой пользователь набивает на обычной клавиатуре, ситуация конца файла всё равно может случиться: драйвер терминала *имитирует* её при нажатии комбинации клавиш `Ctrl-D`.

Если функция `getchar` успешно прочитала один байт, она именно его и вернёт, причём прочитанный байт будет проинтерпретирован как беззнаковое число соответствующей разрядности; иначе говоря, возвращено будет значение от `0` до `255`. Если бы так происходило всегда, функция действительно могла бы возвращать значение типа `unsigned char`, но есть ведь ещё ситуация конца файла. Обозначить эту ситуацию одним из возможных значений типа `char` нельзя, поскольку в потоке ввода может встретиться абсолютно любой байт — никто ведь не гарантирует нам, что эти данные текстовые, бинарный файл тоже может быть прочитан из стандартного потока ввода; кстати, целый ряд утилит командной строки именно так и делает — в среде ОС Unix так поступают, например, практически все архиваторы. Каково бы ни было значение из диапазона `0..255`, избранное нами для обозначения ситуации конца файла, это привело бы к неразличимости ситуации «файл кончился» и ситуации «из файла прочитан байт, равный значению, обозначающему ситуацию конца файла». Ясно, что это абсолютно разные ситуации, так что специальное значение, возвращаемое функцией, если очередной байт прочитан не был, не должно находиться в диапазоне значений прочитанных байтов, то есть должно лежать *за пределами диапазона* `0..255`. Множества возможных значений типа `unsigned char` оказывается *недостаточно*, нужно ещё одно значение.

В качестве такого значения авторы стандартной библиотеки Си избрали обыкновенную минус-единицу (`-1`), но для большей ясности её обозначают идентификатором `EOF` (*end of file*), причём именно так, заглавными буквами.

[30] В ОС Unix мы можем задать системе вопрос, связан ли данный файловый дескриптор с ***терминальным устройством*** или нет; но терминальное устройство может быть (а в современных условиях практически всегда является) результатом программной эмуляции, так что экрана и клавиатуры по ту сторону дескриптора может не оказаться, даже если на вопрос о терминальном устройстве система ответит утвердительно; хотя, конечно, в большинстве случаев «терминальное устройство» предполагает, что «с той стороны» присутствует живой пользователь.

Поскольку функция возвращает одно из 257 возможных значений, её тип возвращаемого значения должен иметь бóльшую разрядность, чем у типа `char`.

Для примера рассмотрим программу, которая читает текст из стандартного потока ввода, пока не возникнет ситуация конца файла, и на каждую прочитанную строку отвечает лаконичным `OK`. Это можно сделать, например, так:

```
#include <stdio.h>

int main()
{
    int c;
    c = getchar();
    while(c != EOF) {
        if(c == '\n')
            printf("OK\n");
        c = getchar();
    }
    return 0;
}
```

Профессиональные программисты, пишущие на Си, практически никогда так не делают, поскольку считают, что нехорошо дублировать строчку «`c = getchar();`» перед циклом и в конце тела цикла. В отличие от Паскаля, где такого дублирования можно избежать разве что с помощью оператора `break`, в Си можно воспользоваться тем, что присваивание является операцией, и «загнать» вызов функции `getchar` в заголовок цикла, а результат присваивания сравнить с константой `EOF`. Поскольку приоритет присваивания ниже, чем приоритет сравнения, операцию присваивания придётся взять в скобки. Всё вместе будет выглядеть так:

```
#include <stdio.h>

int main()
{
    int c;
    while((c = getchar()) != EOF) {
        if(c == '\n')
            printf("OK\n");
    }
    return 0;
}
```

Здесь мы имеем тот редкий случай, когда побочный эффект в условном выражении оказывается уместен и оправдан. Этот вопрос мы подробно обсудим в §4.8.3.

Теперь давайте слегка усложним нашу программу: пусть она вместо `OK` печатает длину только что прочитанной строки. Для этого мы введём счётчик, то есть целочисленную переменную, в которой будет накапливаться количество прочитанных символов, а когда строка кончится, программа напечатает значение этой переменной и обнулит её, чтобы подсчёт символов следующей строки начался с нуля. Всё вместе будет выглядеть так:

```
#include <stdio.h>

int main()
{
    int c, n;
    n = 0;
    while((c = getchar()) != EOF) {
        if(c == '\n') {
            printf("%d\n", n);
            n = 0;
        } else {
            n++;
        }
    }
    return 0;
}
```

Подчеркнём, что в обоих примерах переменная, в которую заносится значение, возвращённое функцией `getchar`, имеет тип `int`. Начинающие программисты, не учитывая вышеприведённых рассуждений о типе значения `getchar`, часто делают характерную ошибку, описывая эту переменную как `char`, в результате чего их программа не различает ситуации, когда `getchar` вернула `EOF` (т. е. `-1`) и когда она вернула число `255`, прочитав байт с таким значением. При обработке текстовой информации видимый результат такой ошибки зависит от применяемой кодировки; к примеру, если наша программа работает в ОС Windows и обрабатывает текст в кодировке cp1251, она «свалится» по концу файла, прочитав маленькую букву «я», поскольку её код в cp1251 равен как раз `255`; при использовании кодировки koi8-r, которая до сих пор достаточно часто встречается в системах семейства Unix, тот же эффект будет наблюдаться при прочтении заглавного твёрдого знака. Если текст представлен в кодировке utf8, то байт со значением `255` там встретиться не должен, но если он в потоке данных всё-таки случайно окажется, то программа, опять-таки, примет его за конец файла. Если же на стандартный поток ввода поданы бинарные данные, то байт со значением `255` в них может встречаться довольно часто: достаточно записать в файл в двоичном виде небольшое по модулю отрицательное целое число, и его представление будет состоять из одного значащего байта и всех остальных, равных как раз `255`.

На всякий случай повторим ещё раз: **значение, возвращённое функцией `getchar`, нельзя сразу присваивать переменной типа `char`**, это приведёт к потере информации; такое присваивание можно сделать лишь после того, как мы убедились, что возвращено значение, отличное от `EOF`.

Операцию вывода символа в стандартный поток вывода производит функция `putchar`; её заголовок выглядит так:

```
int putchar(int c);
```

Параметр задаёт код выводимого символа; формально он имеет тип `int`, но на самом деле числа за пределами диапазона значений `unsigned char` (в реальной жизни это всегда диапазон `0..255`) будут превращены в числа этого диапазона путём отбрасывания «лишних» битов, так что давать их этой функции ни к чему.

Функция возвращает код выведенного символа, если всё в порядке, а если произошла ошибка, то уже знакомую нам константу `EOF`. Впрочем, обычно в программах это значение не проверяется, точно так же, как не проверяется значение `printf`: ошибки при выводе в поток стандартного вывода встречаются довольно редко.

Например, следующая программа принимает на ввод произвольный текст, а выводит начало каждой строки до первого пробела. Для этого предусматривается флажок `pr`, который равен «истине», если в текущей строке ещё не было пробелов, и «лжи», если пробелы уже встречались.

```
#include <stdio.h>                              /* untilspace.c */

int main()
{
    int c, pr;
    pr = 1;   /* true */
    while((c = getchar()) != EOF) {
        switch(c) {
        case '\n':
            putchar('\n');
            pr = 1;
            break;
        case ' ':
            pr = 0;
            break;
        default:
            if(pr)
                putchar(c);
        }
    }
    return 0;
}
```

Таблица 4.1. Спецификаторы преобразований функции `printf`

символ	что и в каком виде печатается
`d`, `i`	знаковое целое число в десятичной системе счисления
`o`	беззнаковое целое число в восьмеричной системе счисления
`u`	беззнаковое целое число в десятичной системе счисления
`x`, `X`	беззнаковое целое число в шестнадцатеричной системе счисления, причём `x` использует буквы `abcdef`, а `X` использует буквы `ABCDEF`
`f`, `F`	число с плавающей точкой в виде десятичной дроби
`e`, `E`	число с плавающей точкой в экспоненциальной форме, например, `6.6234e-34` (буква `e` или `E`, отделяющая порядок от мантиссы, выбирается та же, что и в спецификаторе)
`g`, `G`	число с плавающей точкой в экспоненциальной форме, если значение порядка меньше `-4` или больше либо равно точности, заданной перед спецификатором (по умолчанию 6); в противном случае — в виде обычной десятичной дроби, причём если дробная часть нулевая, то десятичная точка не печатается
`c`	параметр — целое число, это число воспринимается как однобайтный код символа, печатается в итоге символ
`s`	параметр — адрес строки; строка печатается
`p`	параметр — адрес типа `void*`, печатается в шестнадцатеричном виде
`%`	печатается символ `%`, из стека ничего не извлекается

4.4.2. Форматированный ввод-вывод

Функции ***форматированного ввода-вывода*** `printf` и `scanf` мы уже активно использовали в примерах, каждый раз поясняя очередную их возможность. Настало время навести в этом деле порядок, и начнём мы с функции `printf`. Формальный её прототип выглядит так:

```
int printf(const char *format, ...);
```

Многоточие означает, что в этом месте может быть *ещё сколько угодно параметров*; такие функции называются обычно ***вариадическими***. Здесь присутствуют некоторые ограничения на типы таких параметров, которые мы будем рассматривать, когда дело дойдёт до *написания* вариадических функций, а пока просто примем как данность, что параметры всех типов, которые умеет печатать `printf`, через загадочное многоточие передать можно.

Работа функции `printf` *управляется* форматной строкой; можно сказать, что функция представляет собой *интерпретатор* форматной строки. Формально говоря, форматная строка состоит из ***форматных директив***, которые делятся на *обычные символы* (любые байты, кроме нулевого и символа «`%`») и *директивы преобразования*, которые начинаются с символа «`%`» и заканчиваются *спецификатором преобразования*, в роли которого могут выступать символы `diouxXeEfFgGcsp%`; некоторые реализации поддерживают и другие символы в дополнение к этому набору. Спецификатор преобразования указывает, какого типа значение следует извлечь из списка фактических параметров функции (то есть, попросту говоря, из очередной позиции стека) и в каком виде представить это значение на печати. Например, мы уже встречали спецификатор «`d`» (от слова *decimal*) и знаем, что «`%d`» означает целый тип и его десятичное представление; спецификаторы «`o`» и «`x`» тоже означают целый тип, но требуют представить его в восьмеричной и шестнадцатеричной системах соответственно. Описание всех перечисленных спецификаторов приведено в табл. 4.1.

Внутри директивы преобразования, то есть между символом `%` и символом спецификации формата, могут быть указаны дополнительные параметры форматирования (каждый из них не обязателен, то есть может присутствовать, а может и отсутствовать):

- флаги: набор (возможно, пустой) из символов «`-`», «`+`», «`0`», «`#`» и пробел;
- целое число, задающее ширину поля, то есть количество знакомест, отведённых на вывод данного параметра;
- десятичная точка и целое число, задающее так называемую ***точность представления***;
- так называемый ***модификатор разрядности***.

Начнём с ширины и флагов. Число, задающее ширину, означает минимальное количество знакомест, которое будет использовано для выдачи представления очередного параметра; если представление занимает меньше знаков, оно будет (в простейшем случае) дополнено пробелами слева. Например, `printf("[%5d]", 12)` напечатает «`[   12]`». Флаги позволяют изменить это поведение:

- знак «`-`» означает, что выравнивание печатаемой информации следует проводить по левому краю, а не по правому; например, `printf("[%-5d]", 12)` напечатает «`[12   ]`»;
- цифра «`0`» означает, что число следует дополнять слева не пробелами, а нулями; например, `printf("[%05d]", 12)` напечатает «`[00012]`»;
- знак «`+`» означает, что число должно обязательно печататься со знаком, даже если оно положительное; например, `printf("%+d,%+d", 12, -3)` напечатает «`+12,-3`»; без этого знака отрицательные числа всё равно печатаются с минусом, но положительные знаком не снабжаются;
- символ «пробел» означает, что для знака нужно оставить место, но печатать его только для отрицательных чисел, а положительные предварять пробелом; например, `printf("[% d],[% d]", 12, -3)` напечатает «`[ 12],[-3]`».

Флаг «`#`», хотя и входит в число флагов, задаёт не дополнительное требование к ширине поля, а требование модифицировать сам формат. Для восьмеричных и шестнадцатеричных чисел этот флаг требует вывести соответственно лидирующий ноль или комбинацию `0x`/`0X` (в зависимости от того, используется ли `%x` или `%X`); для спецификаторов `e`, `E`, `f` и `F` флаг требует обязательной выдачи десятичной точки, даже если дробная часть равна нулю; для `g` и `G` флаг запрещает удалять незначащие нули в дробной части.

Отметим, что ширина задаёт минимальное, но отнюдь не максимальное количество используемых знакомест. Например, `printf("%3d", 1234)` напечатает «`1234`», никакие цифры отброшены не будут, несмотря на то, что число `1234` в указанные три позиции не помещается.

После ширины (или вместо неё, если её нет) можно задать *точность*, записанную с точки. Проще всего с точностью обстоят дела для чисел с плавающей точкой: точность задаёт количество цифр после запятой (для `g` и `G` — количество значащих цифр в мантиссе, т. е. общее количество значащих цифр до и после десятичной точки, что влияет также на решение, применять ли при выводе экспоненциальную форму). Однако работает эта часть директивы преобразования не только для чисел с плавающей точкой, но и для целых и даже для строк, и в этих случаях её семантика не столь очевидна. Для строк «точность» означает *максимальное* количество символов строки, остальные будут просто отброшены. Например, `printf("[%7.5s]", "abrakadabra")` напечатает «`[  abrak]`»: от строки будет взято только пять символов в соответствии с заданной «точностью», при этом на всю директиву будет отведено семь позиций, то есть произойдёт дополнение слева двумя пробелами; `printf("[%-7.5s]", "abrakadabra")` напечатает «`[abrak  ]`».

Для целого числа «точность» означает *минимальное количество цифр*; обычно это применяется совместно с «шириной», то есть число сначала дополняется незначащими нулями до требуемой «точности», а затем — пробелами до нужной «ширины». Например, `printf("[%6.4d]", 12)` напечатает «`[  0012]`».

Последняя часть директивы преобразования, записываемая непосредственно перед символом-спецификатором, представляет собой модификатор разрядности, который может быть:

- буквой «`l`» («эл»), означающей, что целочисленный параметр имеет тип `long` или `unsigned long`;
- двумя буквами «`ll`» («эл-эл»), означающими, что целочисленный параметр имеет тип `long long` или `unsigned long long`;
- заглавной буквой «`L`», означающей, что параметр с плавающей точкой имеет тип `long double`.

Отдельной буквы, чтобы различать `float` и `double`, не предусмотрено, но этого и не требуется, поскольку при вызове любой функции значение типа `float`, указанное в списке фактических параметров, всегда преобразуется к типу `double`. Надо сказать, что аналогичная ситуация имеется и с целыми — **при передаче в функции целочисленных параметров разрядности меньше, чем `int`, компилятор преобразует их к типу `int`**, так что получить `char` или `short` функция `printf` не может (как и любая другая функция). При этом, как ни странно, модификатор разрядности для «коротких» целых всё-таки предусмотрен — в этой роли выступает буква «`h`». Зачем она нужна и что делает — не вполне понятно, но функция `printf` по крайней мере не считает появление «`h`» ошибкой. Для целых типа `char` модификатора не предусмотрено.

Приведём пример: «`% #7.4Lg`» — это директива преобразования, в которой присутствуют все возможные части: два флага (пробел и «`#`»), ширина `7`,

Таблица 4.2. Спецификаторы преобразований функции `scanf`

символ	что и в каком виде читается
`d`	целое число в десятичной системе счисления
`i`	целое число в восьмеричной, десятичной или шестнадцатеричной системе счисления (с префиксами в соответствии с правилами Си)
`o`	целое число в восьмеричной системе счисления
`u`	беззнаковое целое число в десятичной системе счисления
`x`	беззнаковое целое число в шестнадцатеричной системе счисления (допускается префикс `0x` или `0X`, но он не обязателен)
`e`, `f`, `g`	число с плавающей точкой (по умолчанию `float`)
`p`	адресное выражение в том виде, как его печатает `printf`
`c`	символ (записывается код символа); по умолчанию читается один символ, но если задана ширина поля, то будет прочитано соответствующее количество символов в элементы массива, на начало которого указывает очередной параметр (массив не будет строкой, поскольку «нулевой» символ в его конец не добавляется)
`s`	строка; читается от текущей позиции максимально длинная цепочка, не содержащая пробельных символов (и не превышающая заданную ширину поля)
`[...]`	строка, состоящая из заданных символов; читается от текущей позиции максимально длинная цепочка, содержащая только символы из перечисленных между скобками (и не превышающая заданную ширину поля); если нужно включить в набор символ `]`, его указывают первым, например `[]abc]`
`[^...]`	строка, состоящая из любых символов, кроме заданных; читается от текущей позиции максимально длинная цепочка, содержащая любые символы, кроме перечисленных между скобками (и не превышающая заданную ширину поля); если нужно включить в набор символ `]`, его указывают первым, например `[^]abc]`

точность 4, модификатор разрядности `L` и, наконец, символ-спецификатор преобразования `g`.

Отметим ещё одну интересную возможность. Как ширину, так и точность можно не писать в явном виде, а взять из очередного целочисленного параметра. Для этого в соответствующем месте директивы вместо числа пишется звёздочка «`*`». Например, `printf("%*.*f", w, p, x)` напечатает число `x` с использованием ширины, взятой из `w`, и точности, взятой из `p`.

Для форматированного *ввода* из стандартного потока ввода используется функция `scanf`; её заголовок несколько напоминает заголовок функции `printf`:

```
int scanf(const char *format, ...);
```

Эта функция, как и функция `printf`, *интерпретирует форматную строку*, однако неверно было бы предполагать, что, запомнив правила записи форматной строки `printf`, мы сможем грамотно воспользоваться также и `scanf`'ом; правила интерпретации их форматных строк весьма существенно различаются.

Отметим прежде всего, что видов форматных директив тут не два, а три. Во-первых, есть *директивы преобразования*, формируемые, как и для `printf`, из символа процента и символа-спецификатора, между которыми могут быть необязательные дополнительные настройки в виде числа, обозначающего максимальную ширину поля (здесь надо отметить, что у `scanf` отсутствует аналог задания «точности», а сама «ширина» имеет совершенно иную семантику), и, возможно, буквы, задающей/изменяющей разрядность читаемого числа. Каждая такая директива означает, что нужно прочитать из потока значение в определённом формате и занести его в область памяти, *адрес которой передан очередным параметром*. Кроме перечисленных настроек, в директиве сразу после символа `%` можно указать единственный *флаг*, понимаемый `scanf` — символ «`*`», который *запрещает присваивание*, то есть значение читается, но никакой параметр из стека не берётся и никакая информация ни в какую память не заносится: например, «`%*i`» предписывает прочитать целое число, но никуда его не заносить.

Во-вторых, для `scanf` в качестве директив особого рода рассматриваются последовательности *пробельных символов* произвольной длины, причём пробельными считаются собственно пробел, а также табуляция, перевод строки и т. д.; любая такая последовательность, встреченная в форматной строке (чаще всего это просто один пробел, во всяком случае, нет никаких оснований использовать что-то другое) рассматривается как указание выбирать из потока ввода пробельные символы и игнорировать их, и так до тех пор, пока очередной символ не окажется непробельным.

Наконец, как и `printf`, `scanf` рассматривает *все остальные символы* как директивы особого рода, но если `printf` их просто печатал, то `scanf` требует, чтобы, если он дошел в интерпретации форматной строки до обычного символа, то в потоке ввода в этот момент встретился *точно такой же символ*; если символы в потоке и в форматной строке не совпадают, `scanf` прекращает работу, оставив «неправильный» символ в потоке. То же самое происходит, если `scanf` не смогла трактовать последовательность символов как текстовую запись значения, которое от неё ожидается (целое число, число с плавающей точкой и т. п.)

Отметим ещё одно ключевое отличие. Для `printf` можно не различать, например, `float` и `double` (поскольку все значения типа `float` при передаче в качестве параметров приводятся к `double`), а также можно не отличать числа типа `char` и `short` от обычных `int`'ов. При использовании `scanf` такой номер не проходит, ведь в функцию передаются не значения, а *адреса* областей памяти, и не принимать во внимание их разрядность — заведомая ошибка.

Набор используемых символов-спецификаторов и их роли для `scanf` несколько напоминают спецификаторы для `printf`, но эти наборы отнюдь не одинаковы. Например, для `printf` «`%d`» и «`%i`» означают одно и то же, тогда как поведение `scanf` для этих спецификаторов различается: если «`%d`» предписывает прочитать из потока целое число, записанное в десятичной системе, то «`%i`» предписывает прочитать целое число, которое может быть записано в десятичной, восьмеричной или шестнадцатеричной системах в соответствии с правилами языка Си: лидирующий ноль означает восьмеричную систему, а префикс «`0x`» или «`0X`» — шестнадцатеричную. С другой стороны, при работе с `printf` мы не различаем числа типа `float` и `double`, тогда как для `scanf` мы вынуждены обозначать их по-разному: например, «`%f`» означает прочтение числа типа `float` (иначе говоря, прочтение числа с плавающей точкой и занесение результата в память по адресу, заданному в параметре, в предположении, что по этому адресу расположена переменная типа `float`), тогда как для `double` мы используем комбинацию «`%lf`»; функция `printf` такого спецификатора вообще не знает.

Спецификаторы, используемые `scanf`, перечислены в табл. 4.2. При чтении целых чисел `scanf` предполагает, что очередной параметр имеет тип `int*` (для «`%i`» и «`%d`») или `unsigned int*` (для «`%o`», «`%x`» и «`%u`»); это можно изменить, добавив непосредственно перед спецификатором модификатор разрядности: `h` для `short` (или `unsigned short`), `l` для `long` (соответственно, `unsigned long`), `L` (а не `ll`, как можно было бы ожидать) для `long long` или `unsigned long long`. При чтении чисел с плавающей точкой предполагается тип `float`, для `double` нужно указать модификатор `l`, для `long double` — модификатор `L`. Ширина поля для чисел обычно не используется, но в принципе её тоже можно указать; тогда для формирования числа будет использовано не более чем заданное количество символов из потока ввода.

Совершенно иначе обстоят дела при чтении строк, то есть при использовании «`%s`», «`%[...]`» и «`%[^...]`». **Если не указать ширину поля при чтении строки с помощью `scanf`, то каков бы ни был размер массива, который вы создали для размещения этой строки, этого размера может не хватить и произойдёт переполнение**. При этом если ваша программа просто «упадёт» (то есть аварийно завершится), вы можете рассматривать это как невероятное везение; чем на самом деле опасны переполнения массивов данными, читаемыми извне, будет подробно рассказано на стр. 44. Подчеркнём, что **за такое увольняют с работы**, так что игнорировать это замечание крайне не рекомендуется. Короче говоря, **указание ширины поля при чтении строк абсолютно обязательно**, причём следует учитывать, что массив, в который будет прочитана строка, должен быть хотя бы на один элемент больше, чтобы осталось место для оканчивающего строку «нулевого» символа.

Отметим, что возможности, как для `printf`, взять параметр ширины поля из списка аргументов у `scanf` нет, что резко сужает нашу свободу манёвра: либо мы фиксируем ширину поля и размер массива во время написания программы, либо, как вариант, можем *сгенерировать форматную строку* во время исполнения, но это уже оказывается сложнее, чем устроить обыкновенное посимвольное чтение без всякого `scanf`.

Многие программисты вообще считают, что осмысленное применение `scanf` возможно только в маленьких вспомогательных программах, которые пишутся ради одного-двух запусков и не передаются конечным пользователям, то есть единственным пользователем которых является их автор. Во всех остальных случаях возможностей `scanf` недостаточно для организации полноценной диагностики ошибок.

Так или иначе, полезно знать, что `scanf` возвращает количество успешно преобразованных и помещённых в память значений, то есть, иначе говоря, количество успешно обслуженных директив преобразования, предполагающих запись по адресам, заданным в списке параметров, либо `-1`, если достигнут конец файла или произошла ошибка при чтении. Например, `scanf("%d %d %d", &a, &b, &c)` может вернуть `-1`, если, не успев прочитать ни одного числа, она «упёрлась» в конец файла; она вернёт `0`, если при попытке проанализировать первое же число она встретила ошибочный ввод (буквы, знаки препинания или любую другую белиберду, которая не может быть истолкована как число); значение `1` будет возвращено, если первое число было успешно прочитано, а белиберда встретилась при чтении второго, и так далее; в случае полного успеха (то есть когда все три предполагаемых числа успешно прочитаны) будет возвращено значение `3`. Учтите, что программа, использующая `scanf`, и так мало на что годится, но если она ещё и не анализирует значение, которое `scanf` возвращает, то такую программу можно разве что выбросить.

4.4.3. Работа с текстовыми потоками

Ясно, что кроме стандартных потоков ввода-вывода нам могут понадобиться и другие потоки, к которым относятся, кроме всего прочего, обыкновенные файлы; точнее говоря, для работы с файлом его, как известно, требуется *открыть*, в результате чего как раз и образуется новый поток ввода-вывода.

Для идентификации потоков ввода-вывода библиотека предусматривает своеобразный аналог паскалевских файловых переменных — тип `FILE*`. Судя по наличию звёздочки, это некий *адрес*, и можно даже сказать более определённо — это адрес чего-то, что имеет тип `FILE`. Больше на эту тему сказать ничего нельзя, разные версии библиотеки могут описывать это «нечто» совершенно по-разному, и как оно на самом деле устроено, нас не волнует; то есть мы можем, разумеется, заглянуть как в заголовочный файл (в данном случае `stdio.h`), так и в исходные тексты стандартной библиотеки и в результате узнать, что за сущность в действительности скрывается за словом `FILE`, но делать это стоит разве что из любопытства. В самом деле, если мы попытаемся как-то воспользоваться полученным знанием, то наша программа с хорошей вероятностью не откомпилируется при смене версии стандартной библиотеки из-за того, что там внутренности типа `FILE` могут оказаться совершенно иными.

Можно заметить, что подход к *файловым переменным* в Си радикально отличается от принятого в Паскале. Там переменная соответствующего типа содержит всё необходимое для работы с файлом, в результате чего, например, переменные файлового типа в Паскале нельзя передавать в подпрограммы по значению. Библиотека языка Си, напротив, *сама* заботится о хранении всей нужной для работы с файлом информации, а пользователю (программисту) показывает лишь некий *адрес*, служащий для целей идентификации потока ввода-вывода (то есть отличия конкретного потока от всех остальных), но не более того. Указатели типа `FILE*` можно присваивать, можно передавать их в функции по значению, можно возвращать из функций, ведь передаётся/копируется в каждом случае не информация, связанная с файлом, а простое адресное значение.

Второе весьма заметное отличие от Паскаля состоит в том, что для работы с «нестандартными» потоками ввода-вывода предусмотрен отдельный набор функций, тогда как в Паскале мы использовали те же операторы (псевдопроцедуры) `read` и `write`, что и для стандартных потоков, только первым параметром указывали файловую переменную.

Операция открытия файла для высокоуровневого ввода-вывода выполняется функцией `fopen`, которая имеет следующий заголовок:

```
FILE *fopen(const char *name, const char *mode);
```

Как можно заметить, оба параметра — это *строки*, причём неизменяемые, то есть функция `fopen` ничего не пытается делать с их содержимым, так что в роли значений одного или обоих параметров могут выступать строковые литералы. Первым параметром в функцию передаётся ***имя файла***, про которое нужно только знать, что если оно начинается с символа «/», то это ***абсолютный путь***, начинающийся от корневой директории и не зависящий от того, в какой директории мы сейчас находимся; в противном случае имя считается заданным относительно текущей директории. В частности, если в имени вообще нет ни одного слэша, имеется в виду файл в текущей директории.

Второй параметр задаёт режим открытия файла, то есть то, что мы собираемся с этим файлом делать. Вариантов здесь не так много: `"r"` означает, что файл следует открыть только для чтения; `"r+"` показывает, что файл нужно открыть на чтение и запись, при этом работа начнётся с начала файла; `"w"` предписывает открытие файла на запись, при этом если файла не было, он создаётся (пустым), а если он уже существовал, его старое содержимое уничтожается; `"w+"` делает практически то же самое, только файл при этом открывается на запись и на чтение, то есть можно, записав что-то в файл, после этого принудительно спозиционироваться на уже записанные данные и прочитать их; `"a"` открывает файл на добавление информации в конец, то есть файл открывается на запись, текущее содержимое сохраняется, а запись начинается с первой позиции после старого конца файла; если файла не существовало, он создаётся; `"a+"` работает довольно хитро: файл открывается на чтение и запись, причём чтение начинается с начала файла, а запись происходит в его конец.

Реализация функции `fopen` в системах семейства Unix допускает ещё одну букву — «`b`» от слова *binary*, которая по идее должна означать, что открываемый файл следует рассматривать как двоичный, то есть не текстовый. Эту букву можно добавить в конец строчки `mode` или поставить перед символом `+` (что-то вроде `"rb"`, `"a+b"`, `"wb+"` и т. п.). В unix-системах эта буква игнорируется, то есть совершенно ничего не делает и ни на что не влияет; однако если вам в голову придёт перенести вашу программу под Windows, вы можете с удивлением обнаружить, что работает ваша программа «как-то не так». Автор книги в своё время потратил часа полтора на поиск «ошибки», пока не разобрался, что вся проблема состоит в злосчастной буковке.

Функция `fopen` возвращает адрес типа `FILE*`, который потом можно будет использовать при вызове других функций, если только этот адрес — не `NULL`; если же `fopen` вернула `NULL` — это свидетельствует о произошедшей ошибке. Поводов ошибиться у `fopen` достаточно: файла может не существовать, у нас может не хватать прав на его чтение и/или модификацию, либо на создание файла в указанной директории, и так далее; поэтому проверка на `NULL` после `fopen` строго обязательна.

Итак, мы добрались до такой функции, вызов которой не просто «теоретически» может кончиться ошибкой, но и в действительности настолько часто кончается ошибкой, что не проверять на ошибку просто нельзя. Коль скоро это так, самое время узнать, что же делать, если ошибка произошла, и, прежде всего, как узнать, в чём причина ошибки.

Ответ на вопрос «что делать» зависит от того, какую программу мы пишем и как далеко находимся от функции `main`. Ясно, что надо бы сообщить пользователю о произошедшем, но не во всякой программе мы можем просто взять и напечатать сообщение. Кроме того, возникает неизбежный вопрос «что дальше». Если ошибка проявилась прямо в функции `main` или в какой-то из функций, написанных в качестве вспомогательных для `main`, можно просто завершить программу с ошибочным кодом; если же ошибка возникла в недрах какого-нибудь периферийного модуля или вообще в написанной нами библиотеке, про которую мы даже не знаем, кто и как её будет использовать — то завершать программу здесь будет, разумеется, неуместно; в такой ситуации надо проинформировать вызывающего об ошибке и дать ему возможность самому решить что делать.

Так или иначе, представляется необходимой возможность узнать, в чём причина неуспешного вызова. В языке Си (точнее, в его стандартной библиотеке) этот вопрос для функций, предполагающих обращение к операционной системе[31], решается единообразно: код ошибки в виде обычного целого числа заносится в глобальную переменную `errno`. Чтобы получить доступ к этой переменной, нужно подключить заголовочный файл `errno.h`. Возможные значения переменной `errno` имеют символические имена; например, константа `ENOENT` означает, что ошибка произошла из-за отсутствия либо самого файла, который вы пытались открыть (на чтение), либо директории, в которой этот файл должен был быть расположен; `EACCES` означает, что для открытия файла вашей программе не хватило полномочий; `EROFS` означает, что файл находится на таком диске, который можно только читать (например, на CDROM'е), а вы пытаетесь открыть его на запись, и так далее.

Для каждой библиотечной функции или системного вызова все возможные коды ошибок, которые могут произойти, перечислены в соответствующих

[31] Сама функция `fopen` не является системным вызовом, но открыть файл без участия ОС она, разумеется, не может; свою задачу она решает, применяя системный вызов `open`, знакомый нам из части, посвящённой языку ассемблера.

описаниях, доступных по команде `man`; однако в большинстве случаев этих кодов оказывается слишком много, чтобы по отдельности отрабатывать их все. В частности, для `fopen` список возможных ошибок содержит 21 различный код, большинство из которых вам никогда на практике не встретится. Обычно в программах либо вообще не анализируют значение `errno`, либо проверяют её на одно-два особых значения, а все остальные ошибки обрабатывают одним махом, используя предназначенные для этого средства. Так, функция `strerror` принимает на вход код ошибки, а возвращает адрес строки (то есть выражение типа `char*`), содержащей соответствующее этому коду диагностическое сообщение; на вход этой функции мы можем подать прямо саму переменную `errno`, а полученную строку, например, напечатать.

Впрочем, если мы всерьёз собрались выдать диагностическое сообщение, то лучше выдавать его не на стандартный вывод, а в стандартный поток диагностики (`stderr`), и проще всего это сделать с помощью функции `perror`. Эта функция имеет один параметр — строку, описывающую то, с чем возникла проблема (для файлов обычно указывают имя файла, если же проблема не с файлом, указывают, например, имя функции, вызов которой привёл к ошибке); в поток диагностики выдаётся параметр `perror`, затем двоеточие, пробел и диагностическое сообщение, выбираемое в зависимости от значения `errno`. Например, открытие файла, имеющего имя `file.txt`, мы могли бы оформить так:

```
FILE *f;
f = fopen("file.txt", "r");
if(!f) {
    perror("file.txt");
    exit(1);
}
```

Если файла с таким именем в текущей директории не окажется, в диагностический поток будет выдано:

```
file.txt: no such file or directory
```

Если в вашей системе сконфигурирована локализация, выданное сообщение может оказаться русскоязычным или, в общем случае, может быть сформулировано на том языке, который установлен в вашей системе. Не пугайтесь. Отключить непрошеную русификацию в текущем сеансе работы вы можете командой

```
export LC_ALL= LANG=
```

К сожалению, совсем «открутить» русскоязычные сообщения может оказаться несколько сложнее — это потребует редактирования системных конфигурационных файлов, конкретные имена и содержание которых зависят от используемого вами дистрибутива.

Так или иначе, попытки вашей системы общаться с вами на языке, отличном от английского, полезно сразу же пресечь. Русификации редко бывают удачными, при этом если конечный пользователь может себе позволить не знать английского, программист себе такого позволить не может: если программист не владеет английским, он заведомо профессионально непригоден. Если же вы английским владеете, русификация может доставить вам только неудобства.

Забегая вперёд, отметим, что в стандартный поток диагностики можно, разумеется, выдать произвольное сообщение, например, так (функция `fprintf` и глобальная переменная `stderr` будут рассмотрены позже):

```
fprintf(stderr, "I think there's an error\n");
```

Операция, обратная открытию файла, то есть его *закрытие*, выполняется функцией `fclose`:

```
int fclose(FILE *f);
```

На вход эта функция получает поток ввода-вывода и, собственно говоря, закрывает его, делая дальнейшие операции с ним невозможными; при этом высвобождаются ресурсы операционной системы, обеспечивавшие работу данного потока ввода-вывода[32]. Функция возвращает ноль в случае успеха и `-1`, если произошла ошибка; впрочем, поток закрывается в любом случае, остаться открытым после применения `fclose` поток не может.

Как мы увидим чуть позже, высокоуровневый ввод-вывод работает с ***буферизацией***, что, в частности, подразумевает временное хранение данных, записанных в поток, в буфере, организованном библиотекой, то есть в памяти вашей программы. Вполне возможна, например, такая ситуация: вызванная вами функция записи возвращает управление с признаком успеха (то есть не заявляет ни о каких ошибках), но данные при этом не отданы операционной системе, а всего лишь помещены в буфер. Естественно, никто не гарантирует, что позже, когда библиотечные функции всё же попытаются «слить» системе полученную информацию, не произойдёт ошибка. Если на момент вызова `fclose` в буфере имеются непереданные данные, функция, естественно, попытается их передать, прежде чем окончательно закрыть поток, и при этом вполне возможно, что произойдёт ошибка.

Любопытно будет отметить, забегая вперёд, что ровно такая же ситуация складывается при работе с вводом-выводом низкого уровня, то есть когда ввод-вывод осуществляется непосредственно системными вызовами; в этом случае свою роль играют внутренние буфера операционной системы, а ошибку может выдать системный вызов `close`.

Коль скоро поток ввода-вывода открыт, над ним можно выполнять тот же набор действий, что и над стандартными потоками ввода-вывода, то есть библиотека предусматривает для потоков типа `FILE*` набор функций, аналогичный уже рассмотренному нами для стандартных потоков. В частности, для ввода и вывода отдельных символов (`char`'ов) имеются функции `fgetc` и `fputc`:

```
int fgetc(FILE *stream);
int fputc(int c, FILE *stream);
```

— совершенно аналогичные ранее рассмотренным `getchar` и `putchar`, но принимающие на один аргумент больше; этот аргумент задаёт поток, через который следует работать. Например, следующая программа, получив через аргументы командной строки имена двух файлов, откроет первый на чтение, второй на запись и запишет во второй первые десять строк из первого (если строк окажется меньше, файл окажется просто скопирован):

```
#include <stdio.h>                              /* fgetcputc.c */

int main(int argc, char **argv)
{
    FILE *from, *to;
    int c, lcnt;
    if(argc < 3) {
        fprintf(stderr, "Too few arguments\n");
        return 1;
    }
    from = fopen(argv[1], "r");
    if(!from) {
        perror(argv[1]);
        return 2;
    }
    to = fopen(argv[2], "w");
    if(!to) {
        perror(argv[2]);
        fclose(from);
        return 3;
    }
    lcnt = 0;
    while((c = fgetc(from)) != EOF) {
        fputc(c, to);
        if(c == '\n')
            lcnt++;
        if(lcnt >= 10)
            break;
    }
    fclose(to);
    fclose(from);
    return 0;
}
```

К посимвольному вводу-выводу относится ещё одна функция, которая оказывается неожиданно полезной при анализе текстов, хотя без неё всегда можно обойтись. Это функция ungetc, которая позволяет *вернуть только что прочитанный символ обратно в поток*. Её заголовок выглядит так:

```
int ungetc(int c, FILE *stream);
```

[32] Как мы узнаем из следующей части нашей книги, возможна ситуация, когда с одним потоком ввода-вывода связано больше одного ***дескриптора***; ресурсы ОС в этом случае освобождаются только после закрытия последнего из них. К языку Си это прямого отношения не имеет, это особенности модели ввода-вывода в системах семейства Unix.

Следующая операция чтения получит этот символ первым, как если бы он ещё не был извлечён из потока. Впрочем, тут действует целый ряд ограничений. Во-первых, вернуть можно только один символ; во-вторых, это реально должен быть символ, только что прочитанный из этого же потока. При попытке сделать что-то иное результат оказывается неопределён, то есть, возможно, где-то оно даже сработает, а где-то — нет.

Эту функцию иногда применяют при ***лексическом анализе***, когда невозможно понять, не прочитав очередной символ, является он частью текущей лексемы или нет. Например, читая из потока выражение 756+27, мы поймём, что число 756 кончилось, лишь *прочитав* символ плюс. Вполне возможно, что обработка числа у нас выделена в отдельную функцию, в которой совершенно не хочется устраивать ещё и обработку всех доступных разделителей, тем более что разделители обрабатывать приходится, естественно, не только после чисел.

Ясно, что возврат значения в поток — не более чем хак, тем более что при грамотной реализации лексического анализатора он вообще не должен знать, откуда берутся символы для анализа. Просто в данном конкретном случае этот хак часто экономит значительные усилия, в результате чего функция `ungetc` даже вошла, как мы видим, в состав стандартной библиотеки. Использовать её или нет — решайте сами.

Над потоками возможен также и форматированный ввод-вывод. Соответствующие функции называются `fprintf` и `fscanf`, и от рассмотренных выше функций `printf` и `scanf` они отличаются только наличием дополнительного параметра типа `FILE*`, который указывается первым (то есть перед форматной строкой). Например, следующая программа запишет в файл `sincos.txt` значения синуса и косинуса для углов от 0 до 359 градусов (не забудьте подключить математическую библиотеку при запуске компилятора! Напомним, это делается флагом `-lm`):

```
#include <stdio.h>                                  /* sin360.c */
#include <math.h>

int main()
{
    FILE *f;
    int angledeg;
    f = fopen("sincos.txt", "w");
    if(!f) {
        perror("sincos.txt");
        return 1;
    }
    for(angledeg = 0; angledeg < 360; angledeg++) {
        double rads, s, c;
        rads = (double)angledeg * M_PI / 180.0;
        s = sin(rads);
        c = cos(rads);
        fprintf(f, "%03d % 7.5f % 7.5f\n", angledeg, s, c);
    }
    fclose(f);
    return 0;
}
```

В качестве следующего примера рассмотрим функцию, которая получает на вход уже открытый поток ввода и, рассматривая его как последовательность представлений целых чисел, подсчитывает, во-первых, сумму этих чисел, и, во-вторых, их количество; чтение выполняется до тех пор, пока попытка чтения очередного числа по тем или иным причинам не оканчивается неудачно. Поскольку вернуть нужно два числа, возврат производится через параметры, а сама функция имеет тип возвращаемого значения `void`:

```
void intfilesum(FILE *f, int *sum, int *count)
{
    int n;
    *count = 0;
    *sum = 0;
    while(1 == fscanf(f, "%d", &n)) {
        *sum += n;
        (*count)++;
    }
}
```

Отметим ещё один момент: **любую функцию (как библиотечную, так и вашу собственную), которая рассчитана на работу с потоком типа `FILE*`, можно применить к одному из стандартных потоков ввода-вывода**. Для этого в библиотеке описаны, а в заголовочном файле `stdio.h` объявлены три глобальные переменные типа `FILE*`: `stdin` (поток стандартного ввода), `stdout` (поток стандартного вывода) и `stderr` (диагностический поток).

Например, для функции `ungetc` нет аналога, работающего со стандартным вводом, но благодаря наличию переменной `stdin` вернуть символ в поток стандартного ввода можно, написав что-то вроде `ungetc(c, stdin)`.

4.4.4. Ввод-вывод отдельных строк

Как мы уже неоднократно видели, при обработке текстовых данных одной из основных единиц информации оказывается *строка*, причём текстовый файл состоит из строк, разделённых ***символом перевода строки***. Паскаль, как мы помним, предлагает даже специальные операторы (псевдопроцедуры) для записи и чтения строк целиком. Имеются аналогичные возможности и в стандартной библиотеке Си.

Начнём с функций, требующих явного указания потока ввода-вывода; они называются `fputs` и `fgets`, и, как можно догадаться, первая из них выдаёт заданную строку в заданный поток вывода, а вторая прочитывает строку из заданного потока ввода и размещает прочитанное в указанной области памяти:

```
int fputs(const char *s, FILE *stream);
char *fgets(char *s, int size, FILE *stream);
```

С функцией `fputs` всё более-менее понятно; в частности, мы могли бы вывести надпись `"Hello, world\n"` не с помощью `printf`, как мы делали до сих пор, а с помощью `fputs`:

```
fputs("Hello, world\n", stdout);
```

Фундаментальное отличие `fputs` от `printf` и `fprintf` состоит в том, что содержимое строки *не интерпретируется*, `fputs` никак не обрабатывает никакие специальные символы (вроде символа «`%`», который специальным образом обрабатывается функциями форматного ввода-вывода); заданная строка просто выводится как она есть, символ за символом, пока в очередном элементе массива не окажется ноль, означающий, что строка закончилась. Если стоит задача просто вывести строку, не вставляя в вывод никакие сконвертированные в текст значения выражений, использование `fputs` оказывается эффективнее, то есть, попросту говоря, она работает быстрее, ведь ей не нужно анализировать строку; впрочем, на практике выигрыш совершенно незаметен. Возвращает функция неотрицательное число при успехе и `-1` в случае ошибки; обычно число, возвращаемое при успехе, равно длине выданной строки, но в описаниях этой функции (включая стандарты) это не является требованием.

Функция `fgets` заслуживает более тщательного обсуждения. Параметр `s` задаёт адрес области памяти, в которую вы хотите поместить прочитанную строку. Обычно в качестве такой области памяти используют обыкновенный массив `char`'ов, причём он может быть как описан статически (то есть именно в виде массива), так и выделен динамически, это не важно. Основное, что тут необходимо заметить — это что **достоверно предсказать длину строки, которая будет введена, *невозможно***, так что, сколь бы большой массив мы ни описали, его всё равно может не хватить для размещения прочитанного; если бы функция не знала, *сколько* памяти имеется в её распоряжении, то, получив слишком длинную строку, она продолжила бы записывать её символы в память, находящуюся за пределами массива, и в лучшем случае что-нибудь бы испортила (*худший* случай будет рассмотрен чуть позже). Именно для решения этой проблемы предназначен второй параметр функции, который называется `size`: через него мы сообщаем функции `fgets`, сколько элементов в том массиве, адрес начала которого передан первым параметром.

Получив свои параметры, `fgets` читает из заданного потока ввода *не более чем* ***`size-1`*** символов. Чтение прекращается, если очередной символ оказался символом перевода строки, либо если очередной символ прочитать не удалось (то есть произошла ошибка или возникла ситуация «конец файла»), либо если уже прочитано максимально возможное (т. е. `size-1`) количество символов. Если причиной окончания чтения стал перевод строки, он тоже заносится в массив как обыкновенный символ. Если хотя бы один символ был прочитан, функция заносит ноль в элемент массива, непосредственно следующий за тем, в который был помещён последний из успешно прочитанных символов, превращая массив в корректную строку; при этом функция в качестве своего

значения возвращает тот же адрес, который она получила через параметр `s`. Если ни одного символа прочитать не удалось, то функция никак не изменяет содержимое массива `s`, а возвращает `NULL`.

Например, следующая функция читает строки из потока `f1`, после чего записывает в поток `f2` эти же строки, обрамлённые квадратными скобками; при этом строкам позволяется иметь длину не более 50 символов, если же очередная строка оказывается длиннее, в поток `f2` записывается пустая строка (без скобок).

```
void string50(FILE *f1, FILE *f2)
{
    char buf[51];
    while(fgets(buf, sizeof(buf), f1)) {
        int i;
        int nlpos = -1;
        for(i = 0; i < sizeof(buf) && buf[i]; i++)
            if(buf[i] == '\n') {
                nlpos = i;
                break;
            }
        if(nlpos == -1) {
            int c;
            fputc('\n', f2);
            while((c = fgetc(f1)) != EOF)
                if(c == '\n')
                    break;
        } else {
            buf[nlpos] = '\0';
            fprintf(f2, "[%s]\n", buf);
        }
    }
}
```

Поясним, что после успешного прочтения очередной строки мы здесь пытаемся в прочитанной строке отыскать символ перевода строки `'\n'`, наличие которого свидетельствует о том, что строка была прочитана полностью. Если этого символа в строке нет (то есть переменная `nlpos` так и осталась равна `-1`) — это значит, что произошло одно из двух: либо очередная строка оказалась слишком длинной, либо ситуация «конец файла» возникла в процессе чтения строки, то есть последняя строка в потоке `f1` оказалась неполной. В приведённом примере эти две ситуации не различаются, хотя одну от другой легко отличить, например, по позиции, в которой кончается строка (то есть встречен «нулевой символ»), а ещё лучше — по наличию или отсутствию ситуации «конца файла».

В случае, если прочитанная строка не заканчивается символом перевода строки, мы предполагаем, что это была слишком длинная строка, и пытаемся выбрать из потока остаток этой строки (цикл `while` после `fputc`). Если причиной отсутствия `'\n'` стала неполная последняя строка в потоке, этот цикл просто завершится на первой же итерации из-за ситуации «конец файла». Так или иначе, при этом мы по условию задачи выводим в поток `f2` пустую строку, то есть просто символ перевода строки.

В противном случае (то есть если символ `'\n'` найден) мы выводим в поток `f2` прочитанную строку, заключённую в квадратные скобки, для чего нам сначала нужно отсечь от неё символ перевода строки; это делается занесением в ту позицию, где до этого располагался символ перевода строки, нулевого символа.

Для обеих рассмотренных библиотечных функций имеются их «аналоги», не требующие указания потока и работающие со стандартным вводом-выводом, но эти «аналоги» на поверку оказываются весьма сомнительными. Вот их профили:

```
int puts(const char *s);
char *gets(char *buf);
```

Начнём с `puts`; совершенно неожиданно оказывается, что эта функция работает не так, как её «файловый» аналог: после вывода строки она зачем-то выдаёт ещё и символ перевода строки, то есть если бы нам пришло в голову выдать всё то же `"Hello, world"` с её помощью, написать надо было бы так:

```
puts("Hello, world");
```

Обратите внимание на отсутствие в строке комбинации «`\n`».

Если про `puts` можно сказать, что она несколько *странная*, но всё же иногда может быть полезна, то **за использование `gets` увольняют с работы** (!), и это ничуть не преувеличение — автор своими глазами наблюдал такое увольнение на одном из своих коммерческих мест работы.

Дело тут вот в чём. Единственный параметр, который `gets` получает на вход, задаёт адрес массива для размещения строки; не предусмотрено никакого способа сообщить, каков же у этого массива размер. Как мы уже отмечали, предсказать длину строки *невозможно*, как невозможно (никакими способами!) и гарантировать, что при очередном обращении к `gets` в потоке ввода не окажется строка, превышающая размер отведённой памяти во сколько угодно раз.

В самом лучшем случае программа, «напоровшись» на такую строку, просто «свалится»; если так получилось — считайте, вам очень крупно повезло. Чтобы понять, каков может быть *наихудший* сценарий, придётся вспомнить структуру стекового фрейма, знакомую нам с тех пор, когда мы изучали программирование на языке ассемблера (см. т. 1, §3.3.6). Локальные переменные размещаются в стеке, а сам стек растёт в направлении уменьшения адресов, то есть в сторону, противоположную той, куда мы перемещаемся, увеличивая, например, индекс в массиве. Если мы будем перемещаться вдоль массива, созданного локально в функции, от его начала к концу, но не будем знать, где находится конец, и в результате продолжим перемещаться по памяти уже за границей массива, то рано или поздно мы доберёмся до места, где в стеке располагается адрес возврата из текущей функции.

Дальнейшее — дело техники, пусть и филигранной. Допустим, наша программа вводит в одной из функций локальный массив и выполняет в него чтение с помощью `gets`. Некий злоумышленник, зная об этом, подсовывает нашей программе на вход строку такой длины, что адреса возврата сразу из нескольких функций оказываются содержимым этой строки затёрты, причём не просто так, а с тщательно рассчитанным эффектом. При завершении выполнения текущей функции вместо возврата управления вызывающему произойдёт передача управления туда, куда захотел злоумышленник, что позволяет ему вызвать в нашей программе любую функцию с такими значениями параметров, с какими он пожелает — ведь функция возьмёт параметры из стека, а его содержимое затёрто прочитанной «строкой». Больше того, если в нашей программе не найдётся функции, подходящей для исполнения коварных замыслов злоумышленника, он может поступить ещё хитрее: вставить в подсунутую `gets`'у строку произвольный машинный код (главное, чтобы в этом коде не встречался байт `10`, иначе функция прекратит чтение; но обойтись без такого байта проще простого), и путём затирания адреса возврата обеспечить передачу управления прямо на этот код.

Иначе говоря, программу, в которой имеется вызов `gets`, можно заставить делать практически что угодно, в том числе не имеющее ничего общего с её исходным предназначением. Осталось представить, что ваша программа работает на каком-нибудь сервере или, скажем, управляет банкоматом.

Описанная техника называется ***buffer overflow exploit*** и представляет собой один из простейших методов взлома компьютерных систем; обычно именно эту технику рассказывают «для начала» начинающим «хакерам», и её же описывают в учебниках по информационной безопасности, если только автор такого учебника сам хоть что-то понимает в этой самой безопасности.

Вообще говоря, применение слова *хакер* для обозначения компьютерных взломщиков исходно некорректно. Хакерами называют себя высококвалифицированные программисты определённого класса, в том числе Линус Торвальдс, Ричард Столлман, Эрик Реймонд и многие другие знаменитости. Журналистское понимание слова «хакер» являет собой не более чем следствие некомпетентности подавляющего большинства журналистов практически в любом предмете, о котором они пытаются писать. С книгами по «безопасности» тоже имеются сложности. Большинство книг, в заглавии которых упоминается информационная безопасность, на деле целиком посвящены криптографии, а другие аспекты безопасности полностью игнорируют. Криптография, конечно, вещь полезная, но ни от взлома, ни от угроз, связанных с человеческим фактором, она нас не спасёт.

Так или иначе, **использовать функцию `gets` категорически недопустимо — никогда, ни в каких программах, ни для каких целей**. В библиотеку языка Си эта функция попала, когда ещё не было ни компьютерных сетей, ни компьютерных взломщиков, ни самого понятия компьютерной безопасности. Те времена давно прошли.

Почему-то, даже услышав всё сказанное, некоторые студенты пытаются использовать `gets`, надеясь, по-видимому, на чудо; часто можно увидеть программу, где на вход `gets` подаётся указатель, который автор программы даже

не потрудился инициализировать. Разумеется, такая программа не работает, просто падает; некоторых любителей чудес даже это не останавливает. При инициализации массива нужно указать размерность, но ведь мы не знаем, какой длины будет строка. При вызове `malloc` нужно указать размер выделяемой памяти, но ведь мы не знаем, какова будет строка. При вызове `fgets` требуется указать размер буфера, но ведь мы не знаем, какой длины будет строка! И только `gets` не требует указывать никаких размеров.

На самом деле массив в любом случае нужно расположить в памяти и при этом, разумеется, придётся указать его размер — *до* того, как строка будет считана. Указания размера избежать попросту *невозможно*, в стандартной библиотеке Си нет средств, которые бы это позволяли. Всё, чем отличается `gets` — это тем, что она совершенно спокойно допускает выход за границы массива, каков бы он ни был, и никак этому не препятствует, то есть позволяет программе завершаться аварийно, позволяет злоумышленникам затереть адрес возврата в стеке и всё такое. *Помочь* эта функция не может *ничем*; кстати, если вдруг это утверждение оставило у вас ощущение какой-то недосказанности или подозрение, что «всё на самом деле не совсем так» — значит, вы до сих пор не понимаете, о чём идёт речь, и вам срочно требуется посторонняя помощь для приведения мыслей в порядок. Не поленитесь найти кого-нибудь, к кому можно за такой помощью обратиться.

Подчеркнём, что **штатного способа прочитать строку, никак не ограничивая её длину, в библиотеке Си нет ни в каком виде**. Если требуется такое средство, его придётся изготовить самостоятельно — например, с помощью посимвольного чтения в элементы динамически выделенного массива, с увеличением (ручным!) размера этого массива по мере надобности. Также можно использовать другие динамические структуры данных, например, списки; что заведомо никак не получится — так это найти что-нибудь «волшебное», которое выполнит эту работу за вас. Здесь таких чудес не предусмотрено.

4.4.5. О буферизации ввода-вывода

Как уже говорилось, общая особенность библиотечных функций высокоуровневого ввода-вывода состоит в том, что они, помимо прочего, обеспечивают ***буферизацию*** данных. Так, если мы вызываем функцию `getchar`, то реально прочитан будет не один байт, как мы могли бы предположить, а столько, сколько прямо сейчас можно прочитать из потока стандартного ввода, но не больше определённого количества (например, четырёх килобайт). В частности, если чтение реально выполняется с клавиатуры, за которой сидит живой пользователь, и этот пользователь успел уже набрать некоторое количество данных, то с хорошей степенью вероятности эти данные будут прочитаны все; иначе говоря, когда мы вызовем `getchar`, она прочитает всё, что пользователь уже набрал, при этом нам она отдаст только один байт, а остальное поместит в буфер; при следующем вызове `getchar` она вовсе не станет обращаться к операционной системе, вместо этого она отдаст нам очередной байт из буфера.

То же самое происходит при записи в поток вывода. Данные, которые основная программа передаёт на вывод через высокоуровневые функции — `fputc`, `printf` и т. п. — сохраняются в буфере до наступления одного из условий, при которых буфер должен быть вытеснен. Например, если мы решим сформировать более-менее крупный файл с помощью `fputc`, то операционной системе данные будут передаваться не по одному символу, а сравнительно большими порциями (чаще всего — по 4 KB), что позволяет заметно ускорить выполнение программы, ведь каждое обращение к операционной системе — это изрядная потеря процессорного времени на переключении контекстов и т. п.

Заметим, что буферизация *ввода* в большинстве случаев ни в каком управлении не нуждается, и хотя некоторые возможности такого управления существуют, они применяются очень редко. Дело в том, что буферизация ввода обычно просто экономит время, не давая никакого видимого эффекта.

Совершенно иначе обстоят дела с буферизацией *вывода*. Обнаружить её наличие — проще пареной репы; начинающие программисты при отладке программ очень часто допускают весьма характерный просчёт: вставив в программу отладочную печать, забывают при этом перевести строку. В результате их отладочные надписи, естественно, не достигают экрана, а скапливаются в буфере. Если программа при этом завершается аварийно, вытеснить буфер оказывается некому, и ожидавшаяся надпись на экране так и не появляется, из чего незадачливый программист заключает, что до оператора с отладочной печатью управление не дошло; заключение, понятное дело, оказывается целиком и полностью ошибочным.

Самый надёжный способ заставить библиотеку отдать, наконец, содержимое буфера операционной системе — это потребовать вытеснения буфера в явном виде; для этого предназначена функция `fflush`:

```
int fflush(FILE *stream);
```

Применив эту функцию к потоку *вывода*, вы тем самым потребуете немедленно очистить буфер вывода, связанный с этим потоком; данные, естественно, будут отправлены по назначению.

Интересно, что `fflush` можно применить и к потоку ввода; при этом *в большинстве реализаций* информация, хранящаяся в буфере, будет попросту сброшена. Официально такое поведение никак не закреплено, так что вполне возможны реализации, не делающие этого. Впрочем, в одном из новомодных «стандартов» появилось утверждение, что применение `fflush` к потоку ввода — это undefined behaviour, так что на всякий случай теперь этого действительно лучше не делать: люди, рассматривающие стандарты как руководство к действию (в том числе вопреки здравому смыслу) встречаются намного чаще, чем хотелось бы.

Естественно, буфер вывода будет очищен, а данные переданы системе, если этот буфер заполнен и больше в него ничего не помещается. Кроме того, очистка всех буферов вывода происходит при *корректном* завершении вашей программы.

Существуют ещё две ситуации, когда буфер вывода очищается, хотя об этом никто не просил; но чтобы описать эти ситуации, нам понадобится припомнить, что поток вывода может быть связан с различными сущностями: он может идти в дисковый файл, или в канал (обычно на ввод другой программе), или в сокет для отправки по компьютерной сети, или в псевдоустройство, которое может, например, отдавать данные на принтер или ещё в какое-нибудь периферийное устройство. Среди всех этих случаев следует особо выделить один, когда вывод идёт на ***терминал*** — устройство (возможно, виртуальное), предполагающее, что по ту сторону его находится живой пользователь. Операционная система позволяет узнать, так ли это; в библиотеке для этого предусмотрена функция `isatty`. Конечно, это не даёт стопроцентной гарантии наличия «по ту сторону» живого пользователя, но это и не требуется: если какая-либо программа эмулирует работу терминала, то, следовательно, по каким-то причинам ей требуется, чтобы её партнёры по взаимодействию вели себя так, как будто работают с пользователем.

Так или иначе, **если вывод происходит на терминал, буфер вывода вытесняется ещё в двух важных случаях: при выводе символа перевода строки и при выполнении операции ввода** (что весьма логично: представьте себе программу, выдавшую некое приглашение к вводу, не содержащее перевода строки, просто чтобы вводимый пользователем текст появлялся в той же строке экрана; вытеснение буфера на выводе при операции ввода позволит пользователю увидеть приглашение и понять, чего, собственно говоря, от него ожидает программа). Такой режим называется ***построчной буферизацией*** (англ. *line buffering*). Для потоков, не связанных с терминалом, ничего подобного не происходит.

Вопрос о том, *какие* операции ввода приводят к вытеснению буферов вывода при построчной буферизации, внезапно оказывается историей совершенно отдельной и запутанной: в одних источниках утверждается, что операция ввода должна происходить с потоком, для которого тоже установлен режим построчной буферизации (т. е. в нормальных условиях — при чтении с терминала), авторы других источников пытаются нас убедить, что достаточно вообще любой операции ввода, но такой, чтобы при этом реально происходило чтение, т. е. обращение к операционной системе (а не выдача информации, ранее сохранённой в буфере), третьи источники уточняют, что какой попало ввод, конечно, никак на вытеснении вывода не скажется, но бывают ещё небуферизуемые потоки ввода (в реальной жизни не встречающиеся, хотя потоки вывода без буферизации очень даже бывают), и вот на них операция ввода тоже приведёт к искомому эффекту. Можете не обращать внимания на весь этот бардак: для практических целей достаточно, что операция ввода с терминала обязательно приведёт к вытеснению буфера вывода на терминал, если только вы сами не поменяете что-нибудь в настройках режимов буферизации.

Стоит отметить ещё один интересный момент: в современных реализациях поток диагностики (`stderr`) обычно не буферизуется, т. е. каждое обращение к любой из высокоуровневых функций, предполагающее вывод в `stderr` — даже если выводится всего один символ — обязательно приводит к системному вызову `write`. Так сделано, чтобы понизить вероятность потери или несвоевременной выдачи диагностических сообщений; по-видимому, всему виной неожиданно большое количество программистов, пишущих на чистом Си, не

имея для этого сколько-нибудь достаточной квалификации. **Полагаться на эту небуферизуемость не следует!** Вообще говоря, нет никакой гарантии, что в очередной версии стандартной библиотеки поток `stderr` не окажется внезапно буферизуемым, так что, если вам по каким-то причинам важно, чтобы очередное диагностическое сообщение было выдано немедленно, а не «когда-нибудь», обязательно воспользуйтесь функцией `fflush`, или хотя бы убедитесь, что не забыли закончить сообщение переводом строки (а если по смыслу перевод строки не нужен — тем более используйте `fflush`).

Режимом буферизации для каждого из потоков можно управлять с помощью функций `setbuf` и `setvbuf`. На практике это почти никогда не нужно, но полезно иметь в виду, что такая возможность есть: для любого потока, включая `stderr`, куда бы он ни шёл, на терминал или куда-то ещё, можно установить любой из трёх режимов: блочной буферизации (буфер вытесняется, когда окажется заполнен), строчной буферизации (буфер вытесняется при выдаче конца строки) и небуферизуемый. Можно, более того, заставить библиотеку использовать в роли буфера область памяти, которую вы почему-то хотите видеть в этой роли, и задать такой размер буфера, какой вам кажется правильным. Если что-то из этого вам потребовалось, обязательно расскажите, как вам удалось до такого докатиться.

Примечательна ситуация с размерами самих буферов, а также объёмами операций чтения и записи, которые стандартная библиотека запрашивает у операционной системы. Если вы работаете с ОС Linux на x86-совместимой машине (что в большинстве случаев так и есть), скорее всего константа `BUFSIZ`, соответствующая размерам буферов, используемых вашей версией стандартной библиотеки, окажется равна 8192, т. е. сами буферы — если, конечно, вы в эту кухню не вмешались — имеют размер 8 KB. При этом чтение с терминала библиотека пытается выполнять блоками по 1024 байта, а чтение из любых других потоков — блоками по 4096 байт. Автор книги вынужден признать, что не готов объяснять стоящую за всем этим логику; некоторые подозрения на этот счёт есть, но для книги таковые не годятся.

4.4.6. «Вывод» в строку и «ввод» из строки

Возможности описанных ранее функций форматного ввода-вывода можно задействовать для формирования отформатированного текстового представления в строковом буфере в памяти, не выводя его никуда, и наоборот, для анализа информации, в таком буфере содержащейся. Функция `sprintf`, как и `printf`, анализирует форматную строку, по мере необходимости извлекая значения из списка своих параметров (то есть из стека), но результат преобразования в текстовое представление никуда не выводится, а вместо этого складывается в предоставленный строковый буфер. Профиль этой функции такой:

```
int sprintf(char *buf, const char *format, ...);
```

Буфер, то есть массив, в котором следует сформировать результирующую строку, передаётся функции через первый параметр; остальные параметры такие же, как у `printf`: форматная строка, которая анализируется в точности по тем же правилам, что и для обычного `printf`, и дополнительные параметры, которые должны соответствовать присутствующим в строке директивам преобразования. Следует, разумеется, помнить, что этой функции неоткуда знать размер буфера, так что никаких проверок на эту тему она не производит; обеспечить буфер нужного размера — обязанность вызывающего. К примеру, если вы применяете директиву «`%s`» для вставки в ваш результат некой строки, относительно которой точно не знаете, какого размера она может оказаться, следует обязательно ограничить пространство, которое она может занять в вашем результате, указав перед спецификатором «точность»; например, использовать что-нибудь вроде «`%.20s`»: это гарантирует, что больше 20 элементов из буфера использовано не будет.

В современных версиях библиотеки, как правило, присутствует функция `snprintf`, принимающая ещё один параметр — размер буфера. К сожалению, эта функция в библиотеках Си появилась достаточно поздно; в стандартах она закреплена только начиная с C99. В некоторых старых версиях библиотеки эта функция присутствовала, но игнорировала свой второй параметр.

Если параметрами командной строки компилятора зафиксировать версию языка Си, имевшую место до принятия C99 (например, указать в командной строке `-ansi -pedantic`), то функция `snprintf` окажется недоступна: системные заголовочные файлы написаны так, чтобы опции компилятора влияли не только на сам язык, но и на возможности библиотеки. Это означает, что в проектах, в которых использование ANSI C является требованием, применение `snprintf` может оказаться невозможным.

Функцию `sprintf` часто используют, чтобы перевести число того или иного типа в его текстовое представление, например:

```
char str[32];
int n;
/* ... */
sprintf(str, "%d", n);
```

Здесь вызов функции `sprintf` занесёт в массив `str` текстовое представление числа, хранящегося в переменной `n`. Очень полезна функция `sprintf`, когда возникает необходимость передать некие текстовые данные с использованием ввода-вывода *низкого уровня*, то есть непосредственно через системные вызовы и дескрипторы потоков, минуя высокоуровневую библиотеку; в этом случае данные сначала формируют в массиве типа `char` при помощи `sprintf`, а затем уже передают операционной системе.

Полезно знать, что функция `sprintf`, как и `printf`, и `fprintf`, возвращает количество «выведенных» символов, то есть, попросту говоря, длину строки, сформированной в буфере. «Нулевой» байт, помещённый в конец строки, здесь не учитывается.

Функция `sscanf` проделывает противоположную операцию, то есть анализирует строку в соответствии с заданным форматом и раскладывает полученные результаты по адресам из списка параметров. Её профиль таков:

```
int sscanf(const char *buf, const char *format, ...);
```

Правила интерпретации форматной строки в точности совпадают с таковыми для `scanf`; роль ситуации конца файла здесь играет преждевременное окончание строки `buf`, то есть если эта строка кончилась раньше, чем функция успешно проанализировала хотя бы одну директиву преобразования, возвращается `-1`.

Есть мнение, что пользы от `sscanf` ещё меньше, чем от `scanf`; впрочем, мы не претендуем на истину в последней инстанции.

4.4.7. Блочный ввод-вывод

Ещё во вводной части (см. т. 1, §1.4.6) мы объяснили различия между текстовым и бинарным представлением данных, а позже, обсуждая Паскаль, сами научились работать с файлами, имеющими как текстовый, так и бинарный формат. Естественно, программы, написанные на Си, тоже могут работать не только с текстовыми файлами.

Средства посимвольного ввода и вывода (рассмотренные ранее функции `fgetc` и `fputc`) прекрасно справляются не только с потоками текста, но и с произвольными потоками байтов, в том числе не имеющими ничего общего с текстом. Определяющей причиной, требующей применения для работы с бинарными файлами специальных средств, отличных от рассмотренных «потоковых» функций, оказывается скорее не само нетекстовое представление данных, а то, что такие данные отнюдь не всегда рассматриваются как *поток*. Работа с бинарными файлами часто организуется с использованием попеременно операций чтения и записи в разные места файла; с текстовым файлом такая работа невозможна из-за непостоянства размеров строк и текстовых фрагментов, представляющих числа и другие данные, но файл, который мы не считаем текстовым, никто не мешает рассматривать как последовательность неких записей, которые никогда не изменяют свой размер. Если такие записи будут расположены в «настоящем» файле, то для каждой записи будет известен не только размер, но и её местонахождение (ведь размеры всех предыдущих записей тоже известны), так что можно в любой момент любую из записей как прочитать, так и перезаписать. Очевидно, что при такой работе мы рассматриваем файл не как поток или последовательность (байтов или чего бы там ни было), а скорее как некий аналог массива, только расположенный на диске. Конечно, бинарные данные можно не только хранить в файлах, но и передавать через всевозможные каналы и сокеты, и в этом случае о «местонахождении» говорить бессмысленно, но это не отменяет других отличий бинарных данных от текстовых.

Рассказ о стандартных функциях ввода-вывода был бы неполным без упоминания средств работы с бинарными файлами, и именно им мы посвятим остаток параграфа; но прежде чем продолжить, отметим, что эти средства среди программистов, пишущих на Си, не слишком популярны. **Чаще всего для работы с файлами, имеющими тот или иной нетекстовый формат, используются функции низкоуровневого ввода-вывода**; в системах семейства Unix эти функции представляют собой попросту *системные вызовы*, точнее — их обёртки, то есть полностью соответствуют модели

ввода-вывода, реализуемой ядром операционной системы; получается, что работать с бинарными файлами программисты предпочитают без помощи со стороны библиотеки, обращаясь к системе напрямую.

Причина тут вот в чём. Высокоуровневые средства ввода-вывода, работающие с файловыми переменными типа `FILE*`, имеют два несомненных достоинства. Во-первых, они позволяют переводить числа из внутреннего представления в текстовое (при записи в потоки ввода-вывода) и обратно (при чтении из потоков). Во-вторых, ввод-вывод через высокоуровневые потоки, как уже говорилось, *буферизуется*, что позволяет экономить на количестве дорогостоящих обращений к ядру операционной системы. Первое оказывается ценно при работе с текстовыми файлами, второе — при работе с потоками (неважно, текстовыми или бинарными); но если мы рассматриваем файл как массив неких (бинарных) записей, чередуя при этом операции чтения, записи и позиционирования, то нам оказывается не нужен ни перевод в текстовое представление (ведь формат у нас бинарный), ни буферизация, которая рассчитана на работу с потоками и не приносит пользы, если прыгать по файлу туда-сюда.

Тем не менее, высокоуровневая библиотека всё-таки содержит средства для работы с файлами как массивами произвольных данных — функции блочного чтения и блочной записи. Вы можете пропустить их описание, которому посвящён остаток этого параграфа; при этом вы ничего не потеряете. О том, как работать с файлами на уровне системных вызовов, мы расскажем в §5.2.3.

Заголовочный файл `stdio.h` содержит прототипы нескольких функций, позволяющих работать с файлом как с массивом произвольных (нетекстовых) данных. Открыть файл следует, как обычно, с помощью функции `fopen`; для чтения и записи предназначены функции `fread` и `fwrite`, имеющие следующие прототипы:

```
int fread(void *ptr, int size, int n, FILE *stream);
int fwrite(const void *ptr, int size, int n, FILE *stream);
```

Для обеих функций параметр `stream` задаёт открытый файл, с которым нужно работать; параметр `n` указывает *количество элементов*, которые нужно прочитать или записать, параметр `size` задаёт *размер каждого такого элемента*. Через параметр `ptr` в функции передаётся адрес области памяти: для `fwrite` — *содержащей информацию*, которая должна быть записана в файл, а для `fread` — той, *куда следует поместить* прочитанную информацию. В обоих случаях эта область памяти обязана иметь размер не меньше чем `size`·`n` байт.

Функции возвращают количество успешно прочитанных или записанных *элементов* (не байтов!). Если произошла ошибка, функции возвращают ноль. То же самое происходит, если функция `fread`, не прочитав ни одного элемента (это важно), обнаружила ситуацию «конец файла». Ситуация «конец файла» может возникнуть, когда несколько элементов уже прочитано; в этом случае `fread` возвращает число, мéньшее, чем указанный ей параметр `n` (равное количеству успешно прочитанных элементов), а следующее обращение к `fread` уже вернёт ноль.

Отличить ошибочную ситуацию от штатной (то есть вполне нормальной, не ошибочной) ситуации «конец файла» можно с помощью функций `feof` и `ferror`:

```
int feof(FILE *stream);
int ferror(FILE *stream);
```

Функции возвращают ненулевое значение («истину») в случае, если в заданном потоке возникла ситуация «конец файла» (`feof`) или ошибка (`ferror`), в противном случае возвращается ноль («ложь»). Например, если функция `fread` вернула ноль, можно сразу же вызвать `feof`, и если она вернёт истину — то всё в порядке, просто файл кончился, если же она вернёт ложь — то что-то пошло фундаментально не так.

Изменить текущую позицию в файле можно с помощью функции `fseek`:

```
int fseek(FILE *stream, long offset, int whence);
```

Параметр `stream` задаёт открытый файл, с которым нужно работать. Параметр `offset` указывает, на сколько байтов следует сместиться, а параметр `whence` определяет, от какого места эти байты следует отсчитывать. При значении `whence`, равном константе `SEEK_SET`, отсчёт пойдёт от начала файла (значение `offset` при этом должно быть неотрицательным), при значении `SEEK_CUR` — от текущей позиции (`offset` может быть как положительным, так и отрицательным, и нулевым), при значении `SEEK_END` — от конца файла (самое забавное, что для этого случая значение `offset` тоже может быть любым, как отрицательным или нулевым, так и положительным). Функция возвращает `0`, если всё в порядке, и `-1`, если произошла ошибка. Стоит сразу же оговориться, что не все потоки ввода-вывода допускают изменение текущей позиции — например, потоки, связанные с терминалом (как на ввод, так и на вывод) по вполне очевидным причинам такой возможности не дают: данные, пропускаемые через эти потоки, нигде не хранятся, так что понятия «текущей позиции» для них просто нет. Примером открытого потока, заведомо поддерживающего ***позиционирование***, можно считать обычный дисковый файл, открытый на чтение или запись.

Возвращаясь к функциям `fread` и `fwrite`, отметим, что наиболее универсальный размер «элемента» — один байт, и чаще всего именно так с этими функциями и работают, передавая параметром `size` единицу.

4.5. Составной тип данных и динамические структуры

При изучении Паскаля мы уже сталкивались со связными динамическими структурами данных, такими как односвязный и двусвязный список, двоичное дерево поиска и хеш-таблица. Для их построения, кроме указателей, нужен *составной тип данных*, который в Паскале называется «запись» (`record`); в этой главе мы введём аналог паскалевской записи, который в языке Си называется «структурой» и обозначается ключевым словом `struct`.

4.5.1. Структуры

Как мы уже знаем, ***составной тип данных*** предполагает, что *переменная* этого типа состоит из нескольких переменных других типов, причём в общем случае различных — в этом (безотносительно используемого языка программирования) состоит ключевое отличие переменных составного типа от массивов. Элементы составного типа, то есть те переменные, из которых состоит переменная составного типа, называются ***полями***[33].

В языке Си для создания классических составных типов применяется понятие ***структуры***, вводимое ключевым словом `struct`. Общий подход к синтаксису описания здесь тот же, что и для уже знакомых нам перечислимых типов: сначала ключевое слово (здесь `struct`, для перечислимых типов это было слово `enum`) извещает компилятор, какого рода тип будет описываться, затем следует *необязательное* имя — идентификатор, который вместе с предшествующим ключевым словом составит имя типа; после этого в фигурных скобках следует информация о типе, для структурного типа это перечисление полей, составляющих структуру; завершает описание список вводимых переменных, возможно, пустой:

`struct` [<*имя*>] `{` <*список_полей*> `}` [<*список_перем.*>] `;`

Как и в Паскале, для доступа к полям структуры в Си используется точка, то есть если `s1` — это имя переменной структурного типа и у этой структуры есть поле `f`, то `s1.f` обозначает соответствующую часть переменной `s1`.

Например, создавая базу данных для учебного заведения, мы могли бы заметить, что относительно каждого *студента* нас интересует его имя, пол, год рождения, код специальности, номер курса, номер группы (который на самом деле лучше хранить в виде небольшой строки, а не числа, потому что такие номера часто включают в себя дополнительные буквы), а также его средний балл, который можно хранить в виде числа с плавающей точкой. Структура, хранящая всю перечисленную информацию, могла бы быть описана, например, так:

```
enum { max_name_len = 64, max_group_len = 8 };
/* ... */
struct student {
    char name[max_name_len];
    char sex;   /* 'm' or 'f' */
    int year_of_birth;
    int major_code;
    int year;
    char group[max_group_len];
    float average;
};
```

Подчеркнём, что `student` здесь — так называемое ***имя структуры***, которое само по себе не является именем типа; введённый тип имеет имя, состоящее из двух слов `struct student`. Например, описать переменную этого типа мы можем так:

```
struct student st1;
```

[33] Соответствующий английский термин — *fields*; как видим, перевод в данном случае получился буквальный.

Теперь `st1` — это переменная, которая *состоит* из переменных меньшего размера — тех самых ***полей***. Воспользовавшись одной из версий функции `string_copy`, описанных ранее (см. стр. 34), мы могли бы заполнить эту переменную информацией:

```
string_copy(st1.name, "Otlichnikov Vasiliy Sergeevich");
st1.sex = 'm';
st1.year_of_birth = 1995;
st1.major_code = 51311;
st1.year = 3;
string_copy(st1.group, "312");
st1.average = 4.792;
```

Переменные структурного типа можно инициализировать при их описании, например:

```
struct student st1 = {
    "Otlichnikov Vasiliy Sergeevich",
    'm', 1995, 51311, 3, "312", 4.792
};
```

Отметим, что подобные конструкции допустимы при *инициализации*, но не при *присваивании* (см. рассуждение на стр. 32); с другой стороны, присваивать структурные переменные друг другу, в отличие от массивов, можно, к этому вопросу мы чуть позже вернёмся.

Сами по себе имена структур, не снабжённые словом `struct`, образуют в языке Си отдельное ***пространство имён***, так что программист может использовать такой же идентификатор для чего-нибудь ещё — как имя переменной, функции, какого-то другого типа и т. п., и конфликта имён не возникнет. Пользоваться этим не рекомендуется, поскольку в языке Си++ ситуация прямо противоположна, и ваш код может оказаться непригоден для проектов, в которых используется Си++.

Интересно, что итоговый размер переменной структурного типа может оказаться больше, чем суммарный размер всех её элементов. Дело тут в том, что переменные (т. е. области памяти), занимающие больше одного байта (в общем случае — больше одной ячейки памяти) на многих аппаратных архитектурах могут начинаться в памяти только с определённых адресов, обычно чётных. Например, процессоры Sparc вообще не поддерживают операции извлечения из памяти двухбайтных и четырёхбайтных чисел по нечётным адресам; процессоры интеловской архитектуры, с которыми мы чаще всего работаем, такое извлечение допускают, но оно происходит заметно медленнее, нежели при использовании чётных адресов. Может оказаться, что четырёхбайтные целые в силу тех или иных причин должны располагаться по адресам, кратным четырём, и так далее.

Компилятор всё это учитывает и вставляет в переменную структурного типа так называемые ***выравнивающие байты*** (англ. *padding bytes*), в которых не будет храниться никакая информация; единственное предназначение этих байтов — *занимать место* так, чтобы очередное поле начиналось с адреса, кратного двум или четырём. Так, в структуре `student` из нашего примера компилятор обязательно вставит выравнивающий байт после поля `sex`, причём в зависимости от конкретной архитектуры и версии компилятора такой байт там может оказаться один (следующее поле начнётся с чётного адреса) или их может быть три (следующее поле начнётся с адреса, кратного четырём).

Это явление называется ***выравниванием*** (англ. *alignment*, читается «элайнмент»), и у него есть два очень важных следствия. Во-первых, мы в общем случае не можем знать, сколько ячеек памяти будет занимать переменная структурного типа; во-вторых, невозможно точно сказать, *в каком месте* области памяти, занятой переменной структурного типа, будет располагаться каждое из полей.

Единственный достоверный источник информации о размере структуры — уже знакомая нам псевдофункция `sizeof`, которая позволяет узнать, какой размер приписывает структуре сам компилятор. Так, `sizeof(struct student)` будет при компиляции заменено целым числом, означающим размер этой структуры в памяти; если, скажем, нам нужно выделить динамическую память под размещение такой структуры, правильно это сделать так:

```
struct student *ptr;
ptr = malloc(sizeof(struct student));
```

Если такое выражение кажется слишком громоздким, можно воспользоваться тем, что `sizeof` работает не только для имён типов, но и для произвольных выражений (значение выражения в этом случае игнорируется), а выражение `*ptr` как раз имеет нужный нам тип `struct student`; с учётом этого строчку с вызовом `malloc` можно переписать так:

```
ptr = malloc(sizeof(*ptr));
```

Если вы не поняли, что произошло, не паникуйте и пишите так, как понятнее.

С расположением полей внутри структуры тоже связаны определённые соображения, о которых нельзя забывать. Достоверно об их расположении известно только то, что поля будут располагаться в памяти в том же порядке, в котором они описаны в структуре, что самое первое поле будет расположено точно в начале области памяти, отведённой для структуры, и что поля размером точно в один `char` (т. е. такие, для которых `sizeof` равен единице) всегда располагаются сразу после предыдущего поля без всяких выравнивающих байтов. Из этого следует, в частности, что стоит трижды подумать, прежде чем записывать данные структурных типов в файлы или посылать такие данные куда-нибудь по компьютерным сетям. Формально говоря, даже одну и ту же программу на одной и той же системе два разных компилятора Си «имеют право» откомпилировать с разным выравниванием одной и той же структуры, так что если наша программа запишет в файл содержимое переменной структурного типа, а потом мы программу перекомпилируем, то этот же самый файл (свой собственный!) программа вполне может не суметь корректно прочитать обратно в память.

Справедливости ради скажем, что в рамках одной системы эта возможность представляется скорее теоретической; но вот отправка структурных данных по сети с учётом того, что получатель может работать на совсем другой системе — идея совсем сомнительная. Надёжнее всего пересылать информацию в форме текста, но если это для конкретной задачи совсем не подходит, то с отправляемыми и получаемыми данными следует работать как с однородными массивами байтов, выделяя из такого массива нужные группы байтов, расположенные в заранее оговорённых местах, и преобразуя их в данные, предусмотренные соглашениями (например, в целые числа), не забывая при этом, что и порядок байтов в целых числах может на разных компьютерах отличаться.

Выше мы уже отметили, что переменные структурного типа, в отличие от массивов, разрешено друг другу присваивать — конечно, если тип у них одинаковый. Более того, структуры можно передавать в функции в качестве параметров — прямо по значению, так что структура окажется скопирована в локальную переменную внутри функции. Возвращать их из функций по значению тоже возможно; это, впрочем, не означает, что так действительно нужно делать, за исключением случаев совсем небольших структур. Если ваша структура имеет более-менее существенный объём, передавать в функцию лучше всё-таки её адрес. Что касается возврата из функции, то обычно делают иначе: в функцию передают адрес структурной переменной, и функция заполняет её поля нужной информацией.

Многие программисты, пишущие на Си, стараются избегать не только передачи структур по значению в функции и обратно, но и присваивания структур, даже совсем небольших. Вместо операции присваивания они применяют вызов библиотечной функции `memcpy`, выглядит это примерно так:

```
struct mystr s1, s2;

/* ... */

memcpy(&s1, &s2, sizeof(s1));
```

Если придерживаться такого стиля, получится, что для переменных структурного типа остаётся всего две операции: выборка поля (через точку) и взятие адреса (амперсандом), а больше с ними делать ничего нельзя. В пользу такого самоограничения обычно выдвигается тезис, что структура может содержать поля указательного типа, указывающие на какие-нибудь объекты в динамической памяти, и если структуры то и дело присваивать, очень легко забыть о существовании таких полей и утратить контроль за тем, кто *владеет* такими объектами. Вызов `memcpy` отличается от присваивания определённой громоздкостью, так что его сложнее написать «машинально», не задумавшись о возможных последствиях. Следовать этому принципу или нет — решайте сами.

Так или иначе, указатели на переменные структурного типа (и, в общем случае, выражения типа «адрес структуры») используются в Си настолько часто, что для работы с ними даже придумана специальная операция. Пусть у нас некий указатель `p` указывает на структуру типа `struct student`, то есть он описан как

```
struct student *p;
```

Если мы хотим теперь, используя этот указатель, обратиться к полю структуры (например, к полю `year`), то прежде чем применять операцию «точка», нам нужно сделать из указателя на структуру, собственно говоря, саму структуру; это несложно, достаточно применить операцию разыменования: `*p` и есть нужная нам структура. Но вот прежде чем применять точку к выражению `*p`, стоит вспомнить, что точка представляет собой *унарную операцию*, то есть операцию, у которой всего один операнд. В самом деле, имя поля считать операндом невозможно, ведь оно не является выражением какого-либо типа, его нельзя вычислить, его нельзя вернуть из функции и т. п.; правильнее всего считать, что символ точки — это ещё не операция, а вот `.year` — это операция (извлечения поля `year`), причём она, очевидно, унарная и имеет суффиксную форму. Ранее мы обсуждали, что *суффиксные унарные операции имеют приоритет более высокий, нежели префиксные* (см. стр. 21). Следовательно, если написать (как это часто пытаются сделать начинающие) `*p.year`, мы получим ошибку: операция «`.year`» будет применена не к `*p`, а к самому `p`, но ведь `p` — это просто указатель, никаких полей у него нет. Приходится вмешаться в приоритетность операций, применив круглые скобки, и выглядит это так: `(*p).year`.

Никто не мешает прямо так и писать в программах, но, поскольку в серьёзных программах на Си такая ситуация встречается очень часто, а выражение такого вида выглядит громоздко и непонятно, в язык была добавлена операция «стрелка» («`->`»). **Выражение вида `a->b` означает то же самое, что и `(*a).b`**; при этом, разумеется, «`b`» должно быть именем поля, а «`a`» — адресом структуры, в которой такое поле есть. Вместо `(*p).year` мы можем, следовательно, написать `p->year`, что выглядит не так ужасно.

Синтаксис структур используется также в довольно неожиданной роли — для описания ***битовых полей***, а точнее — для анализа и синтеза целых чисел, состоящих из отдельно рассматриваемых битов и их групп. Для целочисленных полей структуры язык Си позволяет в явном виде указать количество битов, которые данное поле будет занимать. Как правило, такие поля описывают как беззнаковые, а точнее — с использованием типа `unsigned int` (напомним, что слово `int` можно опустить); количество битов указывается сразу после имени поля через двоеточие. Например, структура

```
struct myflags {
    unsigned io_error:1;
    unsigned seen_a_digit:1;
    unsigned had_a_eol:1;
    unsigned signaled:1;
    unsigned count:4;
};
```

состоит из четырёх именованных *флагов*, каждый из которых, будучи однобитовым, может быть равен либо нулю, либо единице, и некоего «счётчика», на который отведено четыре бита, то есть он может принимать значение от 0 до 15. Важно понимать, что подобное имеет смысл, только если несколько битовых полей расположены в структуре подряд друг за другом, поскольку любое поле, не являющееся битовым, естественно, будет начинаться с границы ячейки памяти, возможно, даже не ближайшей.

Изначально введённые в языке битовые поля вообще не предполагалось сочетать в одной структуре с полями других типов. В книге Кернигана и Ритчи говорится, что битовые поля предназначены, чтобы разбить на отдельные биты *машинное слово*, причём порядок битов в слове зависит от архитектуры (порядка байтов в целых числах; см. т. 1, §1.4.2), описывать эти поля можно только типа `int` или `unsigned int`, и лучше (для совместимости) простые `int`'ы не применять, только беззнаковые, а в том, как это всё будет обрабатываться, многое зависит от реализации компилятора. Отметим, что группа идущих подряд битовых полей всегда занимает *целое число слов*, но, впрочем, под «словом» понимается столько байт, сколько занимает `int`; в частности, структура из нашего примера займёт 4 байта как на 32-битной, так на 64-битной машине. Может ли одно битовое поле располагаться частично в одном «слове», а частично — в другом, тоже зависит от реализации компилятора.

Компилятор `gcc`, который мы используем, позволяет описывать битовые поля других типов — например, типа `char`. При этом идущие подряд битовые поля «упаковываются» (физически) в неявное число того типа, какой использован при описании полей. Если структуру из нашего примера модифицировать вот так:

```
struct myflags_c {
    unsigned char io_error:1;
    unsigned char seen_a_digit:1;
    unsigned char had_a_eol:1;
    unsigned char signaled:1;
    unsigned char count:4;
};
```

— то она займёт в памяти один байт, то есть `sizeof(struct myflags_c)` окажется равно 1. Эта возможность относится к числу расширений языка, предложенных GNU, и в других компиляторах не поддерживается — там группы битовых полей всегда занимают целые «слова».

В целом полезно знать о существовании битовых полей (например, чтобы не чувствовать себя идиотом, когда они вам встретятся в чужом коде), но стоит ли их применять на практике в своих программах — вопрос открытый. Многие программисты считают, что с наборами отдельных битов правильнее работать, используя побитовые операции, так что битовые поля просто не нужны.

4.5.2. Односвязные списки

С простейшей из всех «связных» динамических структур данных — ***односвязным списком*** — мы уже знакомы, на Паскале мы с этой сущностью работали. Для экономии места мы опустим здесь подробное описание приёмов работы со списками, всевозможные диаграммы и т. п., и построим рассказ в предположении, что читатель уже умеет работать со списками. Если вы чувствуете себя неуверенно, вернитесь к первому тому и заново проработайте §2.10.4.

Напомним, что под односвязным списком понимается связная структура данных, состоящая из отдельных ***звеньев***, имеющих составной тип, одно из полей которого представляет собой указатель на следующий элемент списка; в последнем элементе списка этот указатель имеет нулевое значение (`nil` для Паскаля, `NULL` для Си). Для работы с односвязным списком используют *указатель на его первый элемент* и дополнительные указатели, где это требуется; нулевое значение в указателе на первый элемент означает, что список пуст. При работе на Си в роли звеньев используются, естественно, структуры; в частности, список целых чисел можно построить из таких звеньев:

```
struct item {
    int data;
    struct item *next;
};
```

Для примера рассмотрим задачу построения списка целых чисел по заданному массиву целых чисел; решение оформим в виде функции, которая будет принимать на вход адрес массива и его длину, а возвращать адрес первого элемента построенного списка. Начнём с варианта решения, при котором новые элементы заносятся *в конец* списка, для чего придётся хранить два указателя: на первый элемент списка и на последний; их мы назовём `first` и `last`. Кроме того, нам потребуется временный указатель, который мы назовём `tmp` (от слова *temporary*), и целочисленная переменная для организации цикла по элементам массива. Пишем:

```
struct item *int_array_to_list(const int *arr, int len)
{
    struct item *first = NULL, *last = NULL, *tmp;
    int i;
    for(i = 0; i < len; i++) {
        tmp = malloc(sizeof(struct item));
        tmp->data = arr[i];
        tmp->next = NULL;
        if(last)
            last->next = tmp;
        else
            first = tmp;
        last = tmp;
    }
    return first;
}
```

На каждом шаге цикла мы сначала создаём очередной элемент списка и заполняем его поля; заполнять поля свежесозданной структуры данных вообще всегда желательно сразу после её создания. Затем в зависимости от того, первый это элемент или нет, мы либо добавляем элемент «в хвост» имеющемуся списку (при этом его начало никак не затрагивается), либо, если список пустой, делаем его состоящим из единственного элемента.

Решение можно существенно упростить, если вспомнить, что добавление *в начало* односвязного списка намного проще, чем добавление в его конец; в частности, рассматривать отдельно случай пустого списка при этом не нужно. Элементы массива перебирать в обратном порядке ничуть не труднее, чем в прямом, так что для нашей задачи добавление в начало вполне подходит. Выглядеть это будет примерно так:

```
struct item *int_array_to_list(const int *arr, int len)
{
    struct item *first = NULL, *tmp;
    int i;
    for(i = len-1; i >= 0; i--) {
        tmp = malloc(sizeof(struct item));
        tmp->data = arr[i];
        tmp->next = first;
        first = tmp;
    }
    return first;
}
```

Наконец, можно вспомнить, что задачи, связанные со списочно-древесными структурами данных, обычно легко решаются с помощью рекурсии. На Си такое решение будет выглядеть более естественно, чем на Паскале, поскольку любой фрагмент (сегмент) массива в Си может рассматриваться как самостоятельный массив; в Паскале такого нет. Базис рекурсии оказывается тривиальным: если массив пустой, следует вернуть пустой список. Если же массив не пуст, то следует создать элемент списка для хранения числа из *первого* элемента массива, после чего, обратившись рекурсивно к своей же функции, построить остаток списка, рассматривая все элементы массива, кроме первого, как массив на один элемент короче. Выглядит это так:

```
struct item *int_array_to_list(const int *arr, int len)
{
    struct item *tmp;
    if(!len)
        return NULL;
    tmp = malloc(sizeof(struct item));
    tmp->data = *arr;
    tmp->next = int_array_to_list(arr + 1, len - 1);
    return tmp;
}
```

На всякий случай для читателей, всё ещё не обретших уверенности с адресной арифметикой, поясним, что `arr + 1` — это адрес того же массива, только начиная с его второго элемента.

Не менее интересна с иллюстративной точки зрения другая задача — перебор всех элементов списка. Для примера напишем функцию, вычисляющую *сумму всех элементов* списка, для которого известен адрес первого элемента. Если применять подход, уже знакомый нам по Паскалю, эта функция может выглядеть, например, так:

```
int int_list_sum(const struct item *lst)
{
    int sum = 0;
    const struct item *tmp = lst;
    while(tmp) {
        sum += tmp->data;
        tmp = tmp->next;
    }
    return sum;
}
```

— однако в программах на Си то же самое обычно записывают иначе, с помощью цикла `for`:

```
int int_list_sum(const struct item *lst)
{
    int sum = 0;
    const struct item *tmp;
    for(tmp = lst; tmp; tmp = tmp->next)
        sum += tmp->data;
    return sum;
}
```

Учитывая, что параметр `lst` — это тоже локальная переменная и менять её (формально) никто не запрещает, можно сделать функцию ещё короче, хотя это уже весьма сомнительное «усовершенствование»:

```
int int_list_sum(const struct item *lst)
{
    int sum = 0;
    for(; lst; lst = lst->next)
        sum += lst->data;
    return sum;
}
```

Впрочем, даже здесь остаётся пространство для дальнейшего манёвра. Вспомнив старую добрую рекурсию, мы можем вообще обойтись без локальных переменных и циклов:

```
int int_list_sum(const struct item *lst)
{
    if(lst)
        return lst->data + int_list_sum(lst->next);
    else
        return 0;
}
```

Наконец, применив условную операцию, мы благополучно превращаем нашу функцию в так называемый «one-liner» (однострочник):

```
int int_list_sum(const struct item *lst)
{
    return lst ? lst->data + int_list_sum(lst->next) : 0;
}
```

Почти всегда, когда мы создали список, возникает вопрос, как его удалить, то есть как убрать из памяти все его элементы. Рассмотрим два очевидных решения, первое:

```
void delete_int_list(struct item *lst)
{
    while(lst) {
        struct item *tmp = lst;
        lst = lst->next;
        free(tmp);
    }
}
```

— и очень похожее второе:

```
void delete_int_list(struct item *lst)
{
    while(lst) {
        struct item *tmp = lst->next;
        free(lst);
        lst = tmp;
    }
}
```

В обоих случаях временная переменная помогает нам решить проблему нарушения связности списка при удалении его первого элемента: в самом деле, если удалить первый элемент, то обращаться к нему больше нельзя, а адрес второго элемента есть только в одном месте — в поле первого элемента; с другой стороны, если сразу переставить указатель на второй элемент, то мы потеряем адрес первого. Первое из вышеприведённых решений состоит в том, что во временном указателе сохраняется адрес первого элемента, основной указатель переставляется на второй элемент, после чего первый удаляется, для чего используется его адрес, сохранённый во временном указателе. Второе решение работает иначе: во временном указателе сохраняется адрес

второго элемента, первый удаляется с использованием основного указателя, затем в основной указатель заносится адрес следующего элемента, ранее сохранённый во временном.

Как водится, с использованием рекурсии то же самое делается несколько изящнее. В качестве базиса выбирается случай пустого списка: тогда нам делать нечего, возвращаем управление сразу же; в противном случае удаляем «хвост» списка, вызвав сами себя, а затем отдельно удаляем «голову» списка:

```
void delete_int_list(struct item *lst)
{
    if(!lst)
        return;
    delete_int_list(lst->next);
    free(lst);
}
```

Можно это записать и иначе, если не ставить целью обязательное текстуальное выделение базиса рекурсии:

```
void delete_int_list(struct item *lst)
{
    if(lst) {
        delete_int_list(lst->next);
        free(lst);
    }
}
```

В заключение рассмотрим задачу *удаления из заданного списка всех элементов, удовлетворяющих определённому условию*; к примеру, попытаемся удалить из нашего списка целых чисел все элементы, в которых хранятся отрицательные числа. Для начала не будем оформлять решение в виде функции; условимся, что у нас есть указатель `first`, хранящий адрес первого элемента, и напишем фрагмент кода, который выкинет все отрицательные числа из такого списка.

Проблема здесь в том, что для исключения элемента из списка *нужно модифицировать тот указатель, который указывает на удаляемый элемент*, причём это может оказаться как поле `next` одного из элементов (предшествующего удаляемому), так и указатель `first` (в случае, если нужно удалить самый первый элемент). Изучая работу со списками на Паскале, мы решению этой проблемы посвятили отдельный параграф (см. т. 1, §2.10.6). Напомним, решение состоит в том, чтобы хранить *адрес указателя на текущий элемент*. В этом случае, если нам придётся уничтожать текущий элемент, мы будем точно знать, какой указатель следует изменить, заставив указывать туда же, куда указывает поле `next` удаляемого элемента. В роли «указателя на текущий» у нас будут выступать последовательно указатель `first`, поле `next` первого элемента, поле `next` второго элемента и так далее; иначе говоря, нам нужен такой указатель, в котором сначала будет лежать *адрес указателя* `first`, затем *адрес поля* `next` *из первого элемента списка*, затем из второго и так пока список не кончится, то есть не окажется, что указатель, адрес которого мы храним, равен `NULL`.

Нашу переменную цикла мы назовём `pcur` и опишем так:

```
struct item **pcur;
```

Сам цикл будет выглядеть следующим образом:

```
pcur = &first;          /* сначала pcur указывает на first */
while(*pcur) {    /* пока то, на что он указывает, не NULL */
    if((*pcur)->data < 0) {   /* если элемент надо удалить */
        struct item *tmp = *pcur;  /* запоминаем его адрес */
        *pcur = (*pcur)->next;      /* исключаем из списка */
        free(tmp);                              /* удаляем */
                  /* при этом следующий уже стал текущим! */
    } else {  /* в противном случае переходим к следующему */
        pcur = &(*pcur)->next;
    }
}
```

Попробуем теперь оформить наше решение в виде отдельной функции. Поскольку указатель `first` может потребоваться изменить, нам придётся передавать в функцию не его значение, а его *адрес*, но это в целом не проблема, так даже проще: мы просто назовём этот параметр `pcur` и используем в качестве переменной цикла. Выглядеть это будет так:

```
void delete_negatives(struct item **pcur)
{
    while(*pcur) {
        if((*pcur)->data < 0) {
            struct item *tmp = *pcur;
            *pcur = (*pcur)->next;
            free(tmp);
        } else {
            pcur = &(*pcur)->next;
        }
    }
}
```

Вызывать эту функцию придётся так:

```
delete_negatives(&first);
```

Как обычно, можно и для этой задачи применить рекурсию:

```
void delete_negatives(struct item **pcur)
{
    if(!*pcur)
        return;
    delete_negatives(&(*pcur)->next);
    if((*pcur)->data < 0) {
        struct item *tmp = *pcur;
        *pcur = (*pcur)->next;
        free(tmp);
    }
}
```

Некоторые программисты, пишущие на Си, предпочитают иной подход к использованию параметра: вместо адреса указателя, который будет меняться, передают его значение, а *новое* значение возвращают из функции обычным способом и заносят обратно в указатель. Вызов функции при этом будет выглядеть так:

```
first = delete_negatives(first);
```

Рекурсивное решение при таком подходе становится примерно таким:

```
struct item *delete_negatives(struct item *pcur)
{
    struct item *res = pcur;
    if(pcur) {
        pcur->next = delete_negatives(pcur->next);
        if(pcur->data < 0) {
            res = pcur->next;
            free(pcur);
        }
    }
    return res;
}
```

Конечно, функцию с таким профилем можно реализовать и без рекурсии, тем же циклом по указателям; принципиальна тут не рекурсия, а подход к обмену информацией между функцией и вызывающим.

Что-то подобное вы наверняка не раз встретите в чужом коде; популярность этого подхода — не только при работе со списками, но и вообще в любой ситуации, когда по итогам работы вызванной функции некий указатель должен по смыслу происходящего поменять значение — к сожалению, намного выше, чем подход того заслуживает. В действительности этот трюк, позволяющий (признаем это) сэкономить несколько тактов процессорного времени, довольно жёстко искажает восприятие функций как явления.

Подробное обсуждение этого момента оставим до специально посвящённого таким ситуациям §4.8.4. Прямо сейчас можно припомнить сформулированные ещё в первом томе принципы ***декомпозиции*** (см. §2.4.4) и заметить, что с *правилом одной фразы* при таком решении справиться намного труднее, ответ на вопрос «что делает функция» то ли получится корявым, то ли слишком длинным, то ли из него не будет понятно, почему у функции именно такие параметры и как с ней работать; ну а *правило обособления* вообще невозможно считать соблюдённым, поскольку реализация функции исходит из предположения, что вызывающий будет обращаться к нашей функции неким предписанным (и притом неочевидным) способом.

4.5.3. Двусвязные списки

В ***двусвязном списке*** каждое звено хранит адреса предыдущего звена и следующего звена, что позволяет, зная адрес любого из звеньев списка, добраться до всех остальных звеньев. Обычно при работе с двусвязным списком используют два указателя — на первое звено и на последнее; в таком списке можно легко вносить новые элементы и удалять существующие в начале списка, в его конце и в произвольной позиции, причём никаких хитрых техник вроде применения указателя на указатель для этого не требуется. Работу с двусвязными списками мы подробно обсуждали в первом томе (см. §2.10.7); ограничимся тем, что адаптируем рассмотренные там решения к реалиям языка Си.

Рассмотрим для примера двусвязный список чисел типа `double`, который можно составить из таких звеньев:

```
struct dbl_item {
    double data;
    struct dbl_item *prev, *next;
};
```

Для работы с этим списком опишем указатели на его начало, конец, текущий элемент, а также временный (вспомогательный) указатель:

```
struct dbl_item *first = NULL, *last = NULL;
struct dbl_item *current = NULL, *tmp;
```

Операции внесения элемента в начало и конец списка вынужденно обрабатывают специальную ситуацию — случай пустого списка. Допустим, у нас имеется переменная `x` типа `double` и число, которое в ней хранится, нужно внести в список. Внесение в начало выглядит так:

```
tmp = malloc(sizeof(struct dbl_item));
tmp->data = x;
tmp->prev = NULL;
tmp->next = first;
if(first)
    first->prev = tmp;
else
    last = tmp;
first = tmp;
```

Внесение в конец выглядит практически так же с точностью до «зеркальной симметрии»:

```
tmp = malloc(sizeof(struct dbl_item));
tmp->data = x;
tmp->prev = last;
tmp->next = NULL;
if(last)
    last->next = tmp;
else
    first = tmp;
last = tmp;
```

Между прочим, в обоих случаях оператор `if` можно заменить однострочным присваиванием. В первом случае это будет выглядеть так:

```
*(first ? &first->prev : &last) = tmp;
```

Во втором случае — так:

```
*(last ? &last->next : &first) = tmp;
```

Разгадать эти ребусы предложим читателю самостоятельно. Разумеется, мы не предлагаем в действительности так писать программы; но если вы с подобным столкнётесь в чужом коде, этот опыт, по крайней мере, поможет быстрее справиться с шоком.

Изъятие первого элемента двусвязного списка выглядит так:

```
if(first) {  /* проверка обязательна! */
    tmp = first;
    first = first->next;
    if(first)
        first->prev = NULL;
    else            /* список опустел */
        last = NULL;
    free(tmp);
}
```

Аналогичным образом выполняется изъятие последнего элемента:

```
if(last) {
    tmp = last;
    last = last->prev;
    if(last)
        last->next = NULL;
    else
        first = NULL;
    free(tmp);
}
```

Заметим, что если в любом из этих двух фрагментов заменить `if` на `while`, мы получим код, удаляющий все элементы списка; впрочем, сделать это можно и проще:

```
if(first) {
    first = first->next;
    while(first) {
        free(first->prev);
        first = first->next;
    }
    free(last);
    last = NULL;
}
```

Поясним, что здесь мы, прежде чем что-то делать, сдвигаем указатель `first` на следующую позицию, и если он после этого всё ещё указывает на очередное звено списка (то есть не стал нулевым), мы уничтожаем предыдущее звено, пользуясь указателем на него из нового «первого» звена. Последний оставшийся элемент мы так уничтожить не можем, потому что нет звена, указатель `prev` которого указывал бы на него; сдвинувшись на следующую позицию с последнего звена, мы получим указатель `first`, равный нулевому адресу; но это не страшно, ведь у нас есть ещё и `last`, через который мы благополучно ликвидируем последнее оставшееся звено списка. Примечательно, что здесь мы обошлись без вспомогательного указателя.

Рассмотрим теперь операции по модификации списка в произвольном месте с использованием указателя на *текущий элемент*. Работа с текущим элементом, естественно, предполагает, что он есть; в дальнейшем мы это не проверяем, предполагая, что проверка уже сделана. Пусть, как и раньше, значение, которое нужно вставить, находится в переменной `x`. Для начала выполним вставку *перед текущим элементом*:

```
tmp = malloc(sizeof(struct dbl_item));
tmp->data = x;                 /* заполняем новый элемент */
tmp->next = current;
tmp->prev = current->prev;
current->prev = tmp; /* он будет предыдущим для текущего */
if(tmp->prev)                /* если есть предыдущий, то */
    tmp->prev->next = tmp;   /* новый для него следующий */
else               /* в противном случае новый элемент - */
    first = tmp;                          /* теперь первый */
```

Аналогично выполняется вставка *после* текущего элемента:

```
tmp = malloc(sizeof(struct dbl_item));
tmp->data = x;
tmp->prev = current;
tmp->next = current->next;
current->next = tmp;
if(tmp->next)
    tmp->next->prev = tmp;
else
    last = tmp;
```

Рассмотрим теперь *удаление текущего элемента*. Для начала мы исключим его из списка, модифицировав указатели в предыдущем и следующем элементах, если они есть, либо указатели `first` и `last`:

```
if(current->prev)
    current->prev->next = current->next;
else
    first = current->next;
if(current->next)
    current->next->prev = current->prev;
else
    last = current->prev;
```

Теперь наш текущий элемент находится вне списка, несмотря на то, что его указатели `prev` и `next` всё ещё указывают на соседние элементы или равны `NULL`, если соответствующих элементов не было (то есть текущий элемент располагался в начале и/или конце списка). Если мы согласны потерять место в списке, где находится сейчас текущий элемент, то операция завершается совсем просто:

```
free(current);
current = NULL;
```

В зависимости от решаемой задачи мы можем захотеть, чтобы новым текущим элементом стал левый или правый сосед удаляемого элемента (обычно это бывает нужно, если мы идём вдоль списка из конца в начало или из начала в конец соответственно). В этом случае нам придётся воспользоваться вспомогательным указателем. Если мы шли слева направо и новым текущим элементом хотим объявить тот элемент, что был справа от удаляемого, вместо двух «простых» строчек, приведённых выше, пишем так:

```
tmp = current;
current = current->next;
free(tmp);
```

Если мы шли справа налево и новым текущим должен стать тот, что был слева, просто меняем `next` на `prev`.

4.5.4. Простое бинарное дерево поиска

Бинарное (то есть двоичное, по числу возможных потомков у каждого узла) дерево поиска состоит из узлов, в каждом из которых присутствуют указатели на левое и правое поддеревья, а также «полезная информация», для хранения которой, собственно говоря, и создаётся дерево. В первом томе мы лишь упомянули дерево поиска (см. §2.10.8), но не стали с ними работать; попробуем частично восполнить этот пробел.

Дерево для хранения целых чисел можно было бы составить из таких узлов:

```
struct node {
    int val;
    struct node *left, *right;
};
```

Для работы с такими деревьями в большинстве случаев применяют рекурсию. Чтобы понять, почему это так, приведём простенький пример. Допустим, у нас есть дерево из узлов типа `struct node` и нам нужно напечатать с помощью `printf` все числа, хранящиеся в дереве. Рекурсивное решение будет тривиальным; если дерево пустое, возвращаем управление (это базис рекурсии), в противном случае обращаемся сами к себе, чтобы напечатать левое поддерево, затем печатаем число в «корне» (то есть текущем узле) дерева и снова обращаемся сами к себе, но на этот раз печатаем правое поддерево:

```
void int_bin_tree_print_rec(struct node *r)
{
    if(!r)
        return;
    int_bin_tree_print_rec(r->left);
    printf("%d ", r->val);
    int_bin_tree_print_rec(r->right);
}
```

Попытаемся теперь сделать то же самое без использования рекурсии. Для начала заметим, что нам, рассматривая любой из узлов дерева, нужно будет помнить, какие узлы расположены выше, чтобы иметь возможность вернуться на верхние уровни дерева. Глубина дерева заранее не известна, так что придётся хранить соответствующую информацию то ли в массиве, то ли в списке. Кроме того, в один и тот же узел мы попадаем (и вынуждены его рассматривать) в общем случае трижды: «сверху» (когда этот узел оказался корневым для очередного рассматриваемого поддерева), при возврате из рассмотрения левого поддерева и при возврате из рассмотрения правого поддерева. Чтобы знать, что ещё осталось сделать в данном конкретном узле дерева и куда следует двигаться после, нам придётся для каждого узла, работу с которым мы начали, но ещё не закончили, хранить *состояние*; мы обозначим вариант «ещё ничего не сделано» словом `start`, ситуацию «мы уже рассмотрели левое поддерево» словом `left_visited` и, наконец, состояние «всё сделано, пора на выход» словом `completed`. Коль скоро для каждого узла приходится хранить больше одного значения (по меньшей мере указатель на сам узел плюс состояние дел с этим узлом), это естественно приводит нас к идее использовать структуры, ну а *стековая* по сути природа этих данных — к идее использования обыкновенного односвязного списка. Последний элемент этого списка будет всегда соответствовать корню всего дерева, тогда как элемент в начале списка — тому узлу, который рассматривается прямо сейчас. Полностью реализация обхода дерева будет выглядеть, например, так:

```
/* bintree.c */
void int_bin_tree_print_loop(struct node *r)
{
    enum state { start, left_visited, completed };
    struct backpath {
        struct node *p;
        enum state st;
        struct backpath *next;
    };
    struct backpath *bp, *t;
    if(!r)
        return;
    bp = malloc(sizeof(*bp));
    bp->p = r;
    bp->st = start;
    bp->next = NULL;
    while(bp) {
        switch(bp->st) {
        case start:
            bp->st = left_visited;
            if(bp->p->left) {
                t = malloc(sizeof(*t));
                t->p = bp->p->left;
                t->st = start;
                t->next = bp;
                bp = t;
                continue;
            }
            /* no break here */
        case left_visited:
            printf("%d ", bp->p->val);
            bp->st = completed;
            if(bp->p->right) {
                t = malloc(sizeof(*t));
                t->p = bp->p->right;
                t->st = start;
                t->next = bp;
                bp = t;
                continue;
            }
            /* no break here */
        case completed:
            t = bp;
            bp = bp->next;
            free(t);
        }
    }
}
```

Даже на самый первый взгляд эта реализация много сложнее, чем приведённый выше вариант с использованием рекурсии. Впрочем, никакой магии здесь нет. В рекурсивном варианте рассмотрению каждого узла дерева соответствует очередной вызов функции, а значит — стековый фрейм; в нём

имеется значение фактического параметра (переменная `r`), которая указывает на текущий узел. Кроме того, в тело функции мы попадаем трижды: при её вызове (попадаем в начало), после возврата из первого рекурсивного вызова (попадаем на ту строчку, где находится `printf`) и после возврата из второго рекурсивного вызова (попадаем в самый конец, в результате выходим из функции). Это в точности соответствует нашим трём *состояниям*, то есть получается, что при рекурсивном обходе дерева состояние тоже хранится (неявно) — в стековых фреймах в виде *адреса возврата*. Больше того, наш стек, реализованный в виде односвязного списка из структур `struct backpath` — есть не что иное, как модель аппаратного стека, используемого в рекурсивном решении.

Такое «принудительное развёртывание» рекурсии находит применение, когда дерево нужно обойти не за один приём, а пошагово: например, в дереве могут храниться некие значения, которые наш модуль должен отдавать по одному по мере того, как их у нас запрашивают. Однако пока такая ситуация не возникла, следует, разумеется, использовать рекурсию: как можно легко догадаться, циклическое решение здесь оказывается не только сложнее по трудозатратам и приложению интеллекта, оно вдобавок использует больше памяти и работает заметно медленнее.

Двоичное дерево обычно используют для ускорения поиска нужного элемента. Для этого элементы, ме́ньшие данного, располагают в левом поддереве, а элементы, бо́льшие данного — в правом. Добавление нового элемента в дерево с сохранением его упорядоченности можно реализовать, например, так:

```
void int_bin_tree_add(struct node **root, int n)
{
    if(!*root) {
        *root = malloc(sizeof(**root));
        (*root)->val = n;
        (*root)->left = NULL;
        (*root)->right = NULL;
        return;
    }
    if((*root)->val == n)
        return;
    if(n < (*root)->val)
        int_bin_tree_add(&(*root)->left, n);
    else
        int_bin_tree_add(&(*root)->right, n);
}
```

Поясним, что в функцию мы передаём *адрес* указателя на корневой узел поддерева; сам этот указатель может быть нулевым, что означает, что поддерево пусто, и ко всему дереву целиком это тоже относится: пока соответствующий указатель содержит нулевой адрес, дерево считается пустым. При первоначальном вызове функции она получает адрес того указателя, который используется в вызывающей программе для работы с деревом; обращаясь рекурсивно к себе самой, она передаёт параметром адрес одного из указателей `left` или `right` из текущего узла дерева.

В самой функции прежде всего (в качестве базиса рекурсии) рассматривается как раз случай пустого дерева/поддерева: в этом и только в этом случае создаётся новый элемент дерева, его поле `val` заполняется добавляемым в дерево числом, а указатели на поддеревья получают нулевые значения, что соответствует пустоте обоих поддеревьев. На этом задача считается решённой и управление возвращается вызвавшему.

Если дерево оказалось непустым, прежде всего производится проверка, не *равно* ли новое число текущему. В этом случае не делается ничего, поскольку такое число уже внесено в дерево; управление немедленно возвращается. В зависимости от решаемой задачи здесь можно было бы как-то оповестить главную программу о том, что на самом деле в дерево ничего не внесено — например, вернуть «ложь», а в других случаях возвращать «истину», и т. д. Можно было бы поступить и хитрее, например, в каждом узле дерева хранить не только само число, но и счётчик, показывающий, *сколько* раз это число было добавлено в дерево.

Наконец, если текущее поддерево не пусто, но при этом добавляемое число не равно текущему, оно может быть либо меньше, либо больше текущего; в этих случаях его следует попытаться добавить соответственно в левое или в правое поддерево. Это делается рекурсивным вызовом, при этом в качестве адреса указателя на корень поддерева используется адрес указателя `left` или `right` из текущего элемента. Дальнейшее — забота рекурсивного вызова; например, если поддерево оказалось пустым, то узел, содержащий добавляемое число, станет непосредственным потомком текущего узла, и так далее.

Как известно, с двоичными деревьями поиска связана одна очень серьёзная проблема: их топология оказывается зависима от порядка, в котором значения заносились в дерево. Так, если в пустое дерево начать заносить последовательность целых чисел, уже упорядоченных по возрастанию или по убыванию, новые числа всегда будут заноситься в самую правую (или самую левую) позицию дерева, а само дерево по сути превратится в односвязный список, официально именуемый вырожденным деревом. Ясно, что поиск в таком дереве будет происходить очень медленно. Для решения этой проблемы применяются так называемые *сбалансированные деревья*. Рассказ о них опустим, тем более что способы их построения и известные алгоритмы балансировки подробно расписаны в литературе и множестве источников в Интернете. Отметим только, что подходов к поддержанию дерева в сбалансированном состоянии существует довольно много, и среди них нет ни одного *простого*; читателям, заинтересовавшимся этой проблематикой, посоветуем начать с так называемых *красно-чёрных деревьев*, поскольку среди всех подходов к балансировке деревьев этот, пожалуй, проще других. Впрочем, другой популярный подход, известный как *AVL-деревья*, если и сложнее, то не сильно. Примечательно, что AVL-деревья были изобретены в России, точнее — в СССР.

Как ни странно, во многих задачах удаётся обойтись без балансировки — если входные данные в силу своей природы оказываются хорошо распределены, дерево изначально получается достаточно «развесистым». Если же это не ваш случай, то часто оказывается проще применить какой-нибудь другой вид дерева, нежели строить двоичное дерево и пытаться его сбалансировать. Например, для хранения строк часто применяются ***префиксные деревья*** (английский термин — *trie*), в которых у каждого узла может быть столько потомков, сколько существует разных букв/символов; продвигаясь от корня к листьям, мы одновременно сдвигаемся вдоль строки, рассматривая каждый раз следующий её символ, то есть на n-ном уровне дерева мы принимаем решение, в какое поддерево пойти, используя n-ный символ из строки. «Высота» такого дерева в точности равна длине самой длинной из хранимых в нём строк, а поиск нужной строки требует стольких сравнений, сколько в строке есть символов.

4.5.5. Объединения и вариантные структуры

Под *объединением* (`union`) в Си понимается такой тип данных, переменная которого может хранить значение одного из нескольких типов, но только одно в каждый момент. Синтаксически объединения очень похожи на структуры: они описываются точно так же и по тем же правилам, только вместо слова `struct` используется слово `union`; у объединений точно так же есть *поля*, к которым можно обращаться через точку, а если используется *адрес* объединения — то через «стрелочку». Отличие от структур, говоря технически, в том, что все поля объединения расположены в памяти начиная с одного и того же адреса — с того адреса, с которого начинается само объединение. Ясно, что присваивание значения любому из полей объединения попросту *затирает* остальные его поля.

Название «объединение» оправдывается соображениями из теории множеств. В самом деле, если рассматривать *тип данных* как *множество возможных значений*, то `union` действительно оказывается теоретико-множественным *объединением* типов как множеств.

Рассмотрим для примера объединение

```
union sample_un {
    int i;
    double k;
    char str[16];
};
```

Как и в случае структур, идентификатор `sample_un` здесь представляет собой *имя объединения*, которое не является именем типа; имя типа здесь состоит из двух слов `union sample_un`. Если теперь описать переменную этого типа:

```
    union sample_un su;
```

— то в нашем распоряжении окажутся: переменная целого типа `su.i`; переменная типа `double su.k`; массив `char`'ов, доступный по адресу `su.str`. Все они расположены в памяти в одном и том же месте, иначе говоря, численное значение адресов `&su`, `&su.i`, `&su.k` и `su.str` одно и то же. Общий объём такого объединения, скорее всего, составит 16 байт, поскольку элемент `str` у него самый большой (хотя теоретически возможны архитектуры, на которых тип `double` окажется занимающим больше 16 байт). В общем случае размер объединения равен наибольшему из размеров его полей.

Изначально объединения были предназначены для экономии памяти, но в такой роли их давно уже никто не использует, поскольку память нынче дёшева, чего никак нельзя сказать о времени, затрачиваемом программистами на создание и отладку программ; между тем, экономия памяти с помощью объединений чревата неочевидными ошибками: можно очень легко перепутать, какое из полей мы присваивали последним, а ведь все остальные при этом содержат откровенную абракадабру.

Одно из частных применений объединения состоит в расчленении того или иного значения на составляющие его байты. Так, объединение

```
union split_int {
    int integer;
    unsigned char bytes[sizeof(int)];
};
```

позволит узнать, из каких байтов состоит представление произвольного целого числа, например:

```
int i;
union split_int si;
printf("Please enter an integer number: ");
scanf("%d", &si.integer);
for(i = 0; i < sizeof(int); i++)
    printf("byte #%d is %02x\n", i, si.bytes[i]);
```

Более интересные возможности открываются, если объединение используется в качестве поля структуры. В таких случаях поля объемлющей структуры обычно содержат в том или ином виде информацию о том, какое из полей объединения следует использовать.

К примеру, если бы мы писали какой-нибудь «продвинутый» калькулятор или другую программу, работающую с формулами, нам, скорее всего, потребовался бы список элементов формулы, в качестве которого могут выступать целочисленные и «плавающие» константы, имена переменных, а также символы действий и скобок. Если, скажем, мы согласимся ограничить длину имени переменной семью символами, чтобы соответствующая строка не занимала больше места, чем число типа `double`, то для представления таких элементов и их хранения в виде списка мы могли бы использовать следующую структуру:

```
struct expression_item {
    char op_type;
    union un_data {
        int i;
        double d;
        char var[sizeof(double)];
    } data;
    struct expression_item *next;
};
```

При этом можно, например, условиться, что если значение поля `op_type` превышает число 32 (ASCII-код пробела), то оно расматривается как код символа операции или другого «формульного» символа (например, скобки), при этом поля объединения `data` не используются; в противном случае значение `op_type` (например, 0, 1 и 2) определяет, какой тип имеет наш элемент формулы и, соответственно, какое из полей объединения следует использовать. Для этих значений `op_type` следует, по-видимому, описать перечислимый тип:

```
enum expr_item_types {
    eit_int = 0, eit_dbl = 1, eit_var = 2, eit_min_op = ' ' + 1
};
```

Теперь если поле `op_type` равно значению `eit_int`, то это целочисленная константа и используется `data.i`, если `eit_dbl` — это дробная константа и используется `data.d`, если `eit_var` — это имя переменной и используется `data.var`; наконец, если значение `op_type` больше либо равно `eit_min_op`, то `data` не используется вообще, а элемент списка представляет собой символ операции или скобку.

Практически все компиляторы ещё в середине девяностых поддерживали так называемые *анонимные объединения*, хотя в стандарт языка эта возможность вошла только начиная с C11. Состоит этот механизм в том, что внутри структуры описывается объединение, причём ему ни в каком виде не даётся имени: не указывается ни имя самого объединения, ни имя поля для него. Тогда поля такого объединения становятся полями самой структуры, но располагаются, как и положено полям объединения, «одно на другом». В частности, с использованием этой возможности вышеприведённую структуру можно было бы описать так:

```
struct expression_item {
    char op_type;
    union {
        int i;
        double d;
        char var[sizeof(double)];
    };
    struct expression_item *next;
};
```

Теперь для структуры типа `struct expression_item` доступны поля `op_type`, `i`, `d`, `var` и `next`, при этом из трёх полей `i`, `d` и `var` можно использовать только какое-то одно, поскольку они занимают одну и ту же память.

Использовать эту возможность, пожалуй, всё же не стоит, несмотря на её довольно очевидное удобство, но полезно помнить о её существовании.

4.5.6. Директива `typedef`

Язык Си позволяет ввести *пользовательское имя типа* — идентификатор, обозначающий тот или иной тип значений и переменных. Для этого используется директива `typedef`; синтаксически описание имени типа получается из *описания одной переменной произвольного типа*: добавление слова `typedef` превращает его в описание нового имени этого типа вместо переменной. Например, описание

```
typedef int *intptr;
```

вводит имя типа `intptr`; переменные такого типа будут указателями на `int`. С помощью `typedef` можно задать синоним для уже существующего типа, в том числе встроенного:

```
typedef int integral_number;
```

Такие синонимы широко используются библиотеками для введения типов, описание которых зависит от архитектуры. Например, тип `intptr_t`, обозначающий целое, разрядности которого хватает для хранения адресов на данной архитектуре, на 32-битных системах обычно вводится как синоним типа `int`, а на 64-битных — как синоним типа `long`.

Очень часто программисты, не желая каждый раз писать имя структурного типа из двух слов, вводят для него синоним с помощью `typedef`, например:

```
struct tag_mystruct {
    int i;
    double d;
};
typedef struct tag_mystruct mystruct;
```

или просто

```
typedef struct tag_mystruct {
    int i;
    double d;
} mystruct;
```

Это позволяет описывать переменные такого типа без использования слова `struct`:

```
mystruct s1, s2, s3;
mystruct *ptr;
```

Начинающие часто оказываются недовольны тем, что с помощью такого имени типа нельзя описать поле-указатель на структуру того же типа, как это требуется при построении списков, деревьев и прочих подобных структур данных. В самом деле, имя, введённое с помощью `typedef`, начинает действовать только *после* директивы; приходится поэтому по-прежнему использовать имя структуры, например

```
typedef struct tag_item {
    int data;
    struct tag_item *next;
} item;
```

— хотя во всём дальнейшем тексте можно уже использовать введённое имя типа:

```
item *first = NULL;
```

Иногда в программах на Си можно встретить описания структур с `typedef`, в которых имя структуры и имя вводимого типа совпадают. Компилятор позволяет это, поскольку имена типов и имена структур относятся в Си к разным областям видимости и не конфликтуют. Следующее описание компилятор вполне нормально «скушает»:

```
typedef struct item {
    int data;
    struct item *next;
} item;
```

Несмотря на кажущуюся привлекательность такого варианта, делать так не рекомендуется, во всяком случае, если вы выносите всё это в заголовочный файл. Дело в том, что в языке Си++, в отличие от Си, «имя структуры» как отдельная сущность отсутствует, там имя структуры считается именем типа само по себе, без добавления слова `struct`, и описание, подобное этому, вызовет конфликт имён, то есть ваш модуль нельзя будет использовать в проектах на Си++, только на чистом Си.

4.6. Макропроцессор

4.6.1. Предварительные сведения

С концепцией макропроцессора мы уже знакомы по языку ассемблера; в языке Си он тоже присутствует и, более того, мы с самого начала используем его во всех наших примерах программ, хотя и не обращаем на это особого внимания. Дело в том, что хорошо знакомая нам директива `#include` как раз является частью макропроцессора. Общая идея макропроцессирования, напомним, состоит в том, что текст (в данном случае текст программы) в какой-то момент подаётся на вход специальной программе или подсистеме компилятора, которая называется ***макропроцессором***. Текст, подаваемый на вход макропроцессору, как правило, содержит определённые указания по его преобразованию, выраженные в виде ***макродирективов*** и ***макровызовов***; в первом приближении можно считать, что макродирективы встроены в макропроцессор, тогда как макровызовы основаны на ***макросах***, определяемых пользователем. Результатом работы макропроцессора становится опять-таки текст, но уже не содержащий ни макровызовов, ни макродирективов, и не требующий дальнейшего преобразования.

Все макродирективы препроцессора языка Си имеют общую особенность: они записываются на отдельной строке, и первым значащим (непробельным) символом такой строчки должен быть символ «#». Обычно перед макродирективой пробелов не оставляют, то есть символ «#» располагают в первой позиции. Макродирективы представляют собой прямой приказ макропроцессору; по окончании его работы никаких следов от макродирективов в анализируемом тексте не остаётся.

Кроме макродирективов, препроцессор обрабатывает также ***макроимена***, то есть идентификаторы, связанные с макросами, введёнными пользователем.

Макропроцессор языка Си обладает одним довольно неприятным свойством: он работает *до начала синтаксического анализа*, хотя и после анализа лексического. При компиляции текст программы на Си сначала разбивается на так называемые ***лексемы***: ключевые слова, идентификаторы, знаки арифметических операций и другие знаки препинания, такие как, например, фигурные скобки, запятые, точки с запятой и двоеточия, а также числовые, символьные и строковые литералы. Например, фрагмент

```
while(count < 10)
    printf("[%d] Hello\n", count++);
```

превратится в результате анализа в следующую последовательность лексем: «`while`», «`(`», «`count`», «`<`», «`10`», «`)`», «`printf`», «`(`», «`"[%d] Hello\n"`», «`,`», «`count`», «`++`», «`)`», «`;`». Комментарии и пробельные символы после лексического анализа исчезают без следа. На всех последующих этапах трансляции вместо изначального текста программы, состоящего из отдельных символов, рассматривается последовательность таких вот лексем, при этом каждая лексема считается единым неделимым целым. Именно на этом этапе, то есть когда текст уже разбит на лексемы, но ещё не проанализирован, переменным не приписаны типы, не выстроено синтаксическое дерево операторов и выражений, не определены области видимости — как раз и запускается макропроцессор. Из этого следует, во-первых, что имена, введённые макропроцессором, представляют собой отдельные лексемы-идентификаторы, то есть макровызовы не могут выполняться внутри комментариев (собственно, их к этому времени уже нет), идентификаторов (что довольно очевидно) и строковых констант; во-вторых, **макроимена не подчиняются областям видимости**, поскольку во время работы макропроцессора текст программы на области видимости ещё не разбит.

4.6.2. Макроопределения и макровызовы

Макроопределение представляет собой фрагмент исходного текста программы, в котором вводится новое имя (идентификатор), предназначенное к обработке макропроцессором. В языке Си макроопределения делаются с помощью директивы `#define`; в простейшем случае она содержит идентификатор, а следом за ним и до конца строки идёт текст, на который этот идентификатор должен быть заменён. Вот несколько примеров:

```
#define BUFFERSIZE 1024
#define HELLOMSG "Hello, world\n"
#define MALLOCITEM malloc(sizeof(struct item))
```

Здесь `BUFFERSIZE`, `HELLOMSG` и `MALLOCITEM` — ***макроимена*** или просто макросы; встретив любой из этих идентификаторов в дальнейшем тексте программы, компилятор заменит первый из них на число `1024`, второй — на строковую константу `"Hello, world\n"`, третий — на вызов функции `malloc`. Можно ожидать теперь появления в программе чего-то вроде следующего:

```
char buffer[BUFFERSIZE];
  /* ... */
puts(HELLOMSG);
  /* ... */
struct item *p = MALLOCITEM;
```

и так далее. Это, собственно говоря, и есть ***макровызовы***.

Тело макроса не обязано быть законченным выражением или даже законченной конструкцией; вполне можно описать, например, такое:

```
#define IF if(
#define THEN ) {
#define ELSE } else {
#define FI }
```

и затем в программе изобразить что-то вроде

```
IF a > b THEN
    printf("the first was greater than the second\n");
    b = a;
ELSE
    printf("the second was greater or equal\n");
    a = b;
FI
```

Конечно, конкретно вот так делать не стоит (хотя язык это и допускает), но в более сложных случаях подобный синтез языковых конструкций из макросов вполне может себя оправдать.

4.6.3. Соглашения об именовании

Читатель наверняка обратил внимание, что все имена макросов в наших примерах набраны заглавными буквами, тогда как до сих пор в программах на Си мы заглавные буквы в идентификаторах не использовали. Это не случайно. Как уже говорилось, макроимена не подчиняются правилам видимости. Что это значит на практике, нам поможет понять следующий пример. Допустим, где-то в одном из заголовочных файлов нашего проекта кто-то из участников определил макрос с именем `count`:

```
#define count 35
```

Мы, не зная об этом или попросту забыв, решаем в какой-нибудь функции объявить локальную переменную с таким же именем:

```
int f(int n, const char *str)
{
    int count;
    /* ... */
```

Поскольку макропроцессор отрабатывает до начала синтаксического анализа, он не может ничего знать ни о функциях, ни об их границах, ни о локализации переменных; как следствие, он просто заменит слово `count` на число `35`, в результате чего синтаксический анализатор «увидит» замечательную конструкцию

```
    int 35;
```

— и, разумеется, выдаст сообщение об ошибке, в котором ключевыми словами будет что-то вроде *identifier expected*, что буквально переводится как «ожидается идентификатор», а означает это, если не делать подстрочного перевода, «здесь должен стоять идентификатор». При этом позиция произошедшей ошибки будет как раз там, где стоит идентификатор — слово `count`. Возможна ещё более пикантная ситуация: мы уже давно написали функцию с локальной переменной `count`, после чего кто-то из наших коллег в одном из заголовочных файлов решил ввести это слово в качестве макроса, и весь наш код разом перестал компилироваться.

Именно по этой причине программисты, работающие на Си, придерживаются традиции использовать для макросов имена, состоящие из букв верхнего регистра. Во-первых, такие имена хорошо видно в тексте программы, они как будто говорят: «Осторожно, тут может возникнуть какая-нибудь нежданная проблема» (и, что характерно, говорят весьма по делу). Во-вторых, если все следуют этому соглашению, то макросы и обычные переменные никогда не смогут вступить в конфликт и создать ситуацию, подобную вышеприведённой: в самом деле, если все макроимена набраны большими буквами, а все прочие идентификаторы — маленькими, то ни одно макроимя не совпадёт ни с одним из «прочих идентификаторов» — имён переменных, типов и т. д.

Подчеркнём, что с точки зрения компилятора Си именем макроса может быть любой идентификатор; соглашение насчёт заглавных букв — это традиция, введённая программистами, а не требование языка.

С этой традицией связан достаточно забавный казус. Как мы уже упоминали при обсуждении перечислимого типа (см. §4.3.8), до появления `enum`'ов макросы были в языке Си единственным способом введения именованных констант; затем, когда перечислимый тип в языке появился, программисты довольно быстро сообразили, что его (а точнее, вводимые им идентификаторы-значения) можно использовать для именования целочисленных констант; но к этому времени привычка именовать константы заглавными буквами пустила настолько глубокие корни, что об исходной причине такого именования (в качестве которой исходно выступала опасность макросов, связанная с их неподчинением областям видимости) люди уже успели позабыть и принялись именовать так *любые* константы, в том числе введённые с помощью `enum`. Закреплению этого нонсенса в немалой степени способствовало то, что Керниган и Ритчи в своей книге использовали заглавные буквы для именования макросов, *никак этого не поясняя*, и вдобавок стали использовать заглавные буквы для именования значений в перечислимых типах, опять-таки не давая на эту тему никаких пояснений.

Итогом всей этой ахинеи стало удивительно сильное распространение среди программистов убеждённости в том, что-де *заглавными буквами следует именовать константы*, при этом никто из разделяющих этот принцип не может дать внятного объяснения, чем же столь «интересны» именно константы, а не, например, имена типов или функций. Оно и понятно, ведь изначально под «константами» в Си можно было понимать только макросы, а с ними, как мы уже видели, всё настолько плохо, что их действительно весьма желательно делать «заметными» в тексте:

```
#define VERY_DANGEROUS_MACRO 756
```

Но вот значения перечислимых типов — это, на первый взгляд, обыкновенные идентификаторы, и в них нет ровным счётом ничего особенного. Отметим, что и *на второй взгляд* они тоже являются обычными идентификаторами, так что если вы используете `enum`'ы по назначению — для обозначения одной ситуации из предопределённого множества — то идентификаторы можно (и, пожалуй, нужно) писать, как обычно, в нижнем регистре с использованием подчёркивания. Здравый смысл подсказывает, что при использовании `enum`'а не по назначению, то есть для введения именованных констант, следует поступать точно так же:

```
enum { absolutely_safe_constant = 756 };
```

Идентификатор, подчиняющийся областям видимости, в любом случае лучше, нежели такой идентификатор, который ведёт себя как слон в посудной лавке (а именно таковы макросы в Си), поэтому метод введения констант как идентификаторов для перечислимых типов получил широкое распространение. Здесь всё ещё как будто нет повода для беспокойства, но он появится, как только мы сделаем следующий — вполне логичный — шаг: решим переделать таким способом все уже имеющиеся константы, которые были ранее объявлены с помощью `#define` и в соответствии с вышеописанными соглашениями названы именами, состоящими из заглавных букв. Возможно, *некоторые* из них мы даже переименуем. Но рано или поздно мы столкнёмся со случаем, когда переименование константы недопустимо: например, константа заявлена как часть интерфейса популярной библиотеки, включена в документацию и используется в десятке (а то и в десятке тысяч) программ. Итак, мы обнаружили константу, которую нельзя переименовать; что же теперь, так и оставить её в виде макроса? Иногда, что характерно, именно так и поступают, в современных программах на Си константы, введённые `#define`'ом, всё ещё встречаются чаще, нежели `enum`'овые. Но, если подумать, валидных причин для такого решения попросту нет: в самом деле, ну *чем хуже* станет код, если очередной макрос заменить на безопасную константу с тем же именем? Результатом становится массовое использование в качестве имён констант перечислимого типа идентификаторов, более привычных в роли макроимён, то есть состоящих из заглавных латинских букв:

```
enum { VERY_DANGEROUS_MACRO = 756 };
```

Здесь наша константа уже и не dangerous, и не macro, но если переименовать её мы не смогли, то что же теперь поделать?

К сожалению, при этом несколько нарушается единообразие имён идентификаторов: одни константы именуются в нижнем регистре, а другие в верхнем. Поэтому требование именовать с использованием букв верхнего регистра *любые* константы, в том числе и такие, которые никогда не были макросами:

```
enum { ABSOLUTELY_SAFE_CONSTANT = 756 };
```

не получается объявить заведомо неправильным: как мы видели, *некоторые* константы, пусть и не являющиеся макросами, всё же приходится так именовать, а единообразие имён ценно само по себе.

Как это часто бывает, однозначных указаний здесь дать нельзя. Если вас пригласили в существующий проект, следуйте имеющимся соглашениям; если вы начинаете проект с нуля, подумайте сами, как вам будет удобнее именовать константы, но не забывайте, откуда *на самом деле* взялось именование заглавными буквами. Если бы не макросы, не было бы никаких оснований «выделять» в тексте программы именно константы, а не что-то другое.

4.6.4. Более сложные возможности макросов

Препроцессор языка Си позволяет задавать макросы *с параметрами*. В качестве примера приведём своеобразную «классику жанра»:

```
#define MAX(A, B) ((A) > (B) ? (A) : (B))
```

Здесь вводится макрос `MAX`, принимающий два параметра, которые в теле макроса обозначаются как `A` и `B`. Вызов этого макроса может выглядеть, например, так: `MAX(x, 15)`, при этом вместо макровызова будет подставлено тело макроса, в котором параметр `A` будет заменён на `x`, а параметр `B` — на `15`. **Перед открывающей круглой скобкой, с которой начинается список параметров, не должно быть пробела**, иначе ваш список параметров будет воспринят как часть тела макроса, а сам макрос станет обычной макроконстантой без параметров.

Вы наверняка обратили внимание на неожиданно большое количество круглых скобок; интересно, что Керниган и Ритчи, приводя этот пример в своей книге, не тратят времени на объяснения, а вместо этого дают ещё один пример:

```
#define square(x) x * x
```

и предлагают вызвать его так: `square(z+1)`, надеясь, что этот ребус окажется «по зубам» всем без исключения читателям. Мы не будем столь оптимистичны и поясним, что произойдёт: вместо макровызова `square(z+1)` препроцессор сгенерирует последовательность лексем «`z + 1 * z + 1`», что, как можно заметить, на ожидавшийся квадрат числа совершенно не похоже. Аналогично, если бы мы определили наш макрос для вычисления максимума без применения скобок:

```
#define BADMAX(A, B)  A > B ? A : B
```

— а потом использовали бы его, например, в составе выражения `BADMAX(100, 10) * 5`, то вместо ожидавшегося `500` мы бы получили результат `100`: в самом деле, чему ещё должно быть равно выражение `100 > 10 ? 100 : 10 * 5`, полученное после макроподстановки? Чтобы избежать подобных вывертов, программисты применяют простое правило: **в параметризованном макросе каждое вхождение макропараметра следует обязательно заключить в круглые скобки, а всё тело макроса — ещё в одни**. Отметим, что это правило ни в коей мере не является требованием со стороны компилятора: ему всё равно; скорее это правило следует отнести к категории придуманных программистами способов преодоления последствий одного из многих кошмарных недочётов языка Си. Заметим, что круглые скобки, даже если их правильно расставить, не способны побороть все возможные последствия применения макросов; в частности, если выражение, заданное в роли одного из параметров нашего макроса `MAX`, окажется имеющим побочный эффект, то этот эффект может проявиться дважды; к примеру, в выражении `MAX(f(x),g(x))` та функция (`f` или `g`), результат которой окажется больше, будет вызвана второй раз.

Вывод оказывается достаточно очевиден: использования макропроцессора следует по возможности избегать. К сожалению, язык Си сравнительно беден, поэтому часто экономия времени и сил, достигаемая с помощью макросов, перевешивает возможные проблемы. Язык Си позволяет писать макроопределения из нескольких строк, и даже эта возможность, несмотря на все неудобства, довольно часто используется. Строки «склеиваются» с помощью символа обратной косой черты «\», который ставится в последней позиции строки (то есть непосредственно перед переводом строки), например, так:

```
#define HELP_TEXT \
    "This is a very good program that displays file size.\n" \
    "To use it you should specify the file name\n" \
    "as a command line parameter, such as\n" \
    "   fsize file.txt\n"
```

Вспомнив, что несколько идущих подряд строковых констант компилятор «склеивает» в одну строку, мы можем догадаться, что здесь определена большая строковая константа, содержащая в себе четыре строки текста, разделённые символами перевода строки; её можно, например, напечатать в случае, если пользователь забыл дать нашей программе нужный ей параметр командной строки:

```
    if(argc < 2) {
        fputs(HELP_TEXT, stderr);
        return 1;
    }
```

Возможность описания многострочных макросов часто используют для описания макросов с параметрами, которые потом разворачиваются в целые функции. Пусть, например, нам потребовалась возможность суммирования элементов заданного массива. Если тип элементов не вызывает сомнения, то всё довольно просто; в частности, массив из `int`'ов можно просуммировать так:

```
int int_array_sum(const int *a, int n)
{
    int s = 0;
    while(n > 0) {
        s += *a;
        a++;
        n--;
    }
    return s;
}
```

Если теперь нам потребуется сумма элементов массива какого-то другого типа, в голову может прийти идея скопировать эту функцию и заменить в копии слово `int` на имя нужного типа, например, `double`:

```
double double_array_sum(const double *a, int n)
{
    double s = 0;
          /* ... */
```

Однако от реализации таких идей следует воздерживаться: копирование фрагментов кода, как известно, до добра не доводит. С помощью механизма параметризованных многострочных макросов можно поступить чуть-чуть лучше:

```
#define MAKE_ARRAY_SUM_FUNCTION(FUNNAME, TYPE) \
    TYPE FUNNAME(const TYPE *a, int n) \
    { \
        TYPE s = 0; \
        while(n > 0) { \
            s += *a; \
            a++; \
            n--; \
        } \
        return s; \
    }
```

Теперь ввести ещё одну такую функцию для суммирования массивов очередного понадобившегося нам типа можно без всякого копирования, достаточно сделать соответствующий макровызов:

```
MAKE_ARRAY_SUM_FUNCTION(int_array_sum, int)
MAKE_ARRAY_SUM_FUNCTION(double_array_sum, double)
MAKE_ARRAY_SUM_FUNCTION(long_array_sum, long)
```

и так далее. Можно пойти ещё дальше и воспользоваться возможностью «склеивания токенов» через псевдооперацию «##»; поясним, что в ходе макроподстановки препроцессор «склеивает» в одну лексему две разные лексемы, между которыми обнаружилось это самое «##», при этом в роли склеиваемых лексем обычно выступают идентификаторы. Например, мы могли бы описать наш макрос с одним параметром, а не с двумя:

```
#define MAKE_ARRAY_SUM_FUNCTION(TYPE) \
    TYPE TYPE ## _array_sum(const TYPE *a, int n) { \
        TYPE s = 0; \
        /* ... */
```

При макровызове, например, с параметром `int` жутковатое `TYPE ## _array_sum` превратится в идентификатор `int_array_sum`, что вполне соответствует нашим целям. Правда, такое решение имеет довольно очевидный недостаток: существуют имена типов из более чем одного слова, и для таких случаев новая версия макроса не годится (как минимум, придётся предварительно ввести новое имя для такого типа с помощью директивы `typedef`, см. §4.5.6).

Неудобства работы с многострочными макросами начинаются с того, что символ обратной косой черты, «экранирующий» перевод строки, обязан быть именно *последним* символом строки; **если случайно оставить после него пробел, текст перестанет компилироваться**, при этом пробел, который стал причиной ошибки, *не будет видно*. Впрочем, к этой особенности компилятора довольно легко приноровиться; но вот если вы допустили в тексте макроса ошибку, результатом которой становится ошибка компиляции, то **компилятор выдаст ошибку в том месте, где макрос раскрывается, а не там, где он описан**, и выискивать причину ошибки в теле макроса вам придётся без помощи со стороны компилятора; фактически в вашем распоряжении останется только небезызвестный «метод пристального взгляда». К этой особенности многострочных макросов «приноровиться» невозможно, поскольку это вообще не вопрос сноровки.

Небольшую помощь в деле отладки макросов вам может оказать ключ компилятора `-E`, предписывающий прекратить компиляцию сразу после стадии макропроцессирования, а результат выдать в поток стандартного вывода;

будьте готовы при этом увидеть не только вашу программу, но и всё, что получилось из всех подключённых к ней заголовочных файлов (а это обычно много).

Коль скоро мы упомянули о склеивании токенов через ##, расскажем заодно об ещё одной своеобразной возможности макропроцессора: о превращении токена в строковую константу. Достигается это с помощью одиночной «решётки», поставленной непосредственно перед именем макропараметра. Например:

```
#define VAR_PRINT(x) printf("%s = %d\n", #x, x)
```

Если теперь этот вызов использовать, скажем, так:

```
int abrakadabra = 13;
VAR_PRINT(abrakadabra);
```

— то раскроется наш макрос в выражение

```
printf("%s = %d\n", "abrakadabra", abrakadabra);
```

ну а напечатано будет, как легко видеть, что-то вроде

```
abrakadabra = 13
```

Вспомнив, что встреченные в тексте два или более идущих подряд строковых литерала «склеиваются», мы можем того же эффекта достичь несколько более интересным способом:

```
#define VAR_PRINT(x) printf(#x " = %d\n", x)
```

4.6.5. (*) Макросы и конструкция `do { } while(0)`

При активном использовании макросов вы рано или поздно наткнётесь на ситуацию, когда в теле макроса по той или иной причине хочется использовать *несколько операторов* или другую сложную конструкцию вроде оператора `if` с веткой `else` или блока с локальными переменными. Бывает и так, что один и тот же макрос вы в зависимости от сложившейся ситуации определяете по-разному, то с одним действием внутри, то с двумя, то вообще как «пустышку», не делающую ничего. При этом хочется, чтобы макровызов в программе выглядел так же, как обычный оператор с вызовом функции, вроде «`MYMACRO("argument");`», с точкой с запятой на конце, и чтобы эта конструкция могла быть без опаски использована в любом контексте, где допустим *оператор вычисления выражения*, например, в качестве тела `if`. Если не принять специальных мер и использовать один из следующих вариантов:

```
#define MYMACRO(arg) f(arg); g();
#define MYMACRO(arg) f(arg); g()
#define MYMACRO(arg) { f(arg); g(); }
```

— то можно с ходу указать конструкцию, которая хоть и выглядит в тексте программы совершенно правильной, тем не менее не откомпилируется:

```
if(cond)
    MYMACRO("argument");
else
    /* ... */
```

Можно придумать и другие контексты, в которых все или некоторые из вышеприведённых определений «сломаются». Как это часто бывает с языком Си, программисты довольно быстро нашли для возникшей проблемы работающее решение, основанное на конструкции, исходно предназначенной совершенно не для этого; проблема решается путём заключения тела макроса в вырожденный цикл `do-while`:

```
#define MYMACRO(arg) do { f(arg); g(); } while(0)
```

Поскольку условие в `while` изначально установлено ложным, такой цикл, будучи *циклом с постусловием*, всегда выполняется ровно один раз; с другой стороны, синтаксис языка Си требует после него ставить точку с запятой, что полностью соответствует нашему желанию относительно вызова `MYMACRO` с точкой с запятой на конце, подобно обыкновенной функции.

Подчеркнём, что подходящая конструкция нашлась в языке случайно; в частности, синтаксис мог бы и не требовать точки с запятой в конце конструкции `do-while`, она там явно избыточна. Программисты на Си вынуждены были стать большими мастерами на подобные выдумки.

4.6.6. Директивы условной компиляции

Препроцессор языка Си содержит средства, позволяющие временно исключать из компиляции фрагменты исходного кода, а также выбирать один из нескольких фрагментов в зависимости от заданных условий; это позволяет быстро создавать различные (например, отличающиеся набором возможностей, применёнными алгоритмами и т. д.) варианты одной и той же программы, причём организовать работу можно так, что для переключения с одного варианта на другой не потребуется вносить никаких изменений в исходные тексты, что может оказаться очень важно.

С условной компиляцией мы встречались уже дважды (оба раза в первом томе) — сначала в части о Паскале при обсуждении отладочной печати (§2.13.3), а затем в части, посвящённой ассемблеру — при обсуждении макропроцессора (§3.5.4). Надо сказать, что и во Free Pascal, и в NASM макродирективы, связанные с условной компиляцией, пришли из языка Си, и это сейчас станет очевидно.

Самый простой пример применения условной компиляции — исключение фрагмента, который больше не нужен (например, вы уже написали другой вариант реализации той же задачи), но удалить его рука не поднимается. При работе на других языках программирования, в том числе на Паскале, для этого обычно используются знаки комментариев, соответствующее действие даже получило жаргонное название: *закомментарить* кусок текста; в английском языке используется лаконичное *to comment out*. В программах на Си такой вариант не проходит, поскольку вложенные комментарии компилятором не поддерживаются; этот факт уже упоминался на стр. 9. Вместо знаков комментария в Си для временного исключения фрагмента кода используют макродирективы `#if 0` и `#endif`, которые, как и любые другие макродирективы, должны быть записаны на отдельных строках:

```
#if 0  /* this implementation was too slow */
void sort(int *a, int n)
{
    /* ... */
}
#endif
```

Всё, что написано между строками `#if 0` и `#endif`, компилятор просто проигнорирует; там даже можно оставлять синтаксически неполные конструкции, такие как непарные скобки в выражениях и операторах, отсутствующие условия и т. п., нельзя только допускать лексические ошибки вроде незакрытых комментариев или незавершённых строковых констант. Если потребуется снова включить такой фрагмент в компиляцию, убирать директивы `#if` и `#endif` не обязательно, достаточно заменить `#if 0` на `#if 1`.

Добавление макродирективы `#else` позволяет (опять-таки в простейшем случае) организовать выбор между разными вариантами кода. Пусть, например, мы никак не можем для себя определить, какой из двух вариантов функции `string_copy`, рассмотренных в §4.3.18, лучше использовать; применение условной компиляции позволит нам иметь в коде обе реализации, оперативно переключаться между ними и сравнивать, с какой из них программа получается лучше (например, быстрее работает):

```
#if 1
void string_copy(char *dest, const char *src)
{
    int i;
    for(i = 0; src[i]; i++)
        dest[i] = src[i];
    dest[i] = 0;
}
#else
void string_copy(char *dest, const char *src)
{
    for(; *src; dest++, src++)
        *dest = *src;
    *dest = 0;
}
#endif
```

Заменяя в макродирективе `#if` единицу на ноль и обратно, мы тем самым прикажем компилятору обрабатывать первую либо вторую версию функции, например, в зависимости от нашего настроения сегодняшним утром. Можно поступить и хитрее. В директиве `#if` на самом деле можно указать любое целочисленное константное выражение, которое компилятор сможет вычислить на этапе макропроцессирования, то есть до начала собственно компиляции; это означает, что ни к переменным, ни к функциям, ни даже к константам, введённым с помощью `enum`, мы в таких выражениях обратиться не сможем (всего этого попросту *ещё нет*, когда работает макропроцессор), но ведь существуют ещё и константы, вводимые директивой `#define`, то есть обыкновенные макросы. Итак, для начала придумаем себе имя макросимвола для выбора между реализациями; как несложно заметить, первая отличается от второй использованием индексирования, соответственно наш символ мы так и назовём: `USE_INDEX_IN_STRING_COPY`. Строчку с директивой `#if` заменим на вот такую:

```
#if USE_INDEX_IN_STRING_COPY
```

Если соответствующий макросимвол в нашей программе не определён, то в директиве `#if` препроцессор заменит его на `0` (точнее, на `0L`, но это в данном случае неважно), и компилироваться будет, соответственно, вторая версия функции, то есть та, которая не использует индексы. Изменить это можно, вставив в программу (например, куда-нибудь в самое начало) соответствующую директиву `#define`, например,

```
#define USE_INDEX_IN_STRING_COPY 1
```

Можно действовать ещё интереснее. Мы видели, что и Free Pascal, и NASM позволяют *определять макросимволы прямо в командной строке компилятора, не трогая исходные тексты*. Естественно, такая возможность предусмотрена и в любом компиляторе Си; в частности, `gcc` для этого использует ключ `-D`, например:

```
gcc -Wall -g -D USE_INDEX_IN_STRING_COPY=1 prog.c -o prog
```

Это позволяет выбирать, какой вариант программы мы хотим собрать, *без внесения изменений в исходные тексты программы*, то есть без редактирования её исходных файлов. Этот момент часто недооценивается начинающими, но опытные программисты знают, насколько важно обходиться без «временных» изменений в файлах, составляющих программу. В серьёзных проектах любые изменения, внесённые в исходные тексты, автоматически фиксируются системами контроля версий, которые позволяют восстановить любую версию исходного текста, существовавшую в любой момент времени на протяжении разработки программы. В ряде случаев журналы изменений файлов приходится анализировать людям, время которых очень дорого стоит; наводнять историю изменений бесконечными правками в стиле «туда-обратно» значит отнимать время у своих коллег, а часто и у самого себя.

Вернёмся к нашей условной компиляции. Мы можем не полагаться на то, что компилятор заменит неопределённый макросимвол нулём; можно проверить, определён ли уже этот символ, и если нет, то определить его в тексте программы. Для этого можно воспользоваться директивами `#ifdef` и `#ifndef`, последняя в данном случае удобнее:

```
#ifndef USE_INDEX_IN_STRING_COPY
#define USE_INDEX_IN_STRING_COPY 1
#endif
```

Теперь если символ был определён в командной строке компилятора или, например, в каком-нибудь из подключённых заголовочных файлов, то он сохранит своё значение, если же его ещё никто не определил, то он будет определён и станет равным единице. К помощи `#ifdef` и `#ifndef` часто прибегают для оформления отладочной печати (см. т. 1 §2.13.3, §3.5.4; соответствующие конструкции в Си будут выглядеть совершенно аналогично, что и понятно, ведь в Си всё это появилось изначально, а Free Pascal и NASM *заимствовали* условную компиляцию из Си).

Остановимся подробнее на выражениях, допустимых для директивы `#if`. Эти выражения, как уже было сказано, должны быть целочисленными и вычислимыми во время работы препроцессора, то есть единственный вид идентификаторов, который в них допускается — это макроимена. Подчёркивается явным образом, что тип чисел, используемых в директивах условной компиляции — `long`. Спектр используемых операций ограничен, хотя и довольно широк: разрешается использовать все пять действий арифметики (включая взятие остатка от деления), операции сравнения, логические связки и побитовые операции, допустима также условная операция «`?:`». С другой стороны, запрещены все операции, имеющие или потенциально способные иметь побочный эффект (присваивания, инкремент/декремент, вызов функции), операции с адресами (взятие адреса, разыменование и индексирование), операция приведения типа, операция «запятая» и `sizeof`.

Наконец, в этих выражениях допустима одна операция, специфичная для препроцессора — это операция проверки определённости макросимвола. Она записывается с помощью слова `defined`, причём обычно её используют как *вызов псевдофункции* с круглыми скобками, хотя это и не обязательно (то есть скобки можно и не писать). Например:

```
#if defined(DEBUG_PRINT) && DEBUG_PRINT > 7
```

Упомянутые выше `#ifdef` и `#ifndef` представляют собой сокращённую запись для `#if defined` и `#if !defined`.

Наряду с `#if`, `#else` и `#endif` предусмотрена также директива `#elif`, означающая *else if* и предполагающая условное выражение, аналогичное тем, которые предполагаются для `#if`. При изучении макропроцессора ассемблера NASM мы рассматривали пример, когда у нас есть два «особенных» заказчика Петров и Сидоров, для которых нужно включать в программу специфические версии какого-то кода. Допустим, для всех остальных заказчиков нужно компилировать некий третий вариант; тогда мы могли бы предусмотреть символы `FOR_PETROV` и `FOR_SIDOROV`, определяемые в командной строке компилятора, а в самой программе написать что-то вроде следующего:

```
#if defined(FOR_PETROV)
   /* код для Петрова */
#elif defined(FOR_SIDOROV)
   /* код для Сидорова */
#else
   /* код для всех остальных */
#endif
```

Отметим, что определить символ можно, не задавая ему никакого значения (оставив его значение пустым). С помощью директивы `#define` это делается просто:

```
#define MYSYMBOL
```

В командной строке компилятора это тоже можно сделать с помощью всё того же флага `-D`:

```
gcc -Wall -g -D MYSYMBOL prog.c -o prog
```

Обычно такие макросимволы, не имеющие значений, предназначаются для управления директивами условной компиляции, то есть для использования во всевозможных `#if defined`, `#ifndef` и прочих директивах, зависящих не от значения символа, а только от факта его определённости или неопределённости.

4.6.7. Ещё несколько полезных директив

С помощью макродирективы `#undef` можно заставить препроцессор «забыть» макросимвол: после того, как препроцессор обработает строку

```
#undef MYMACRO
```

символ `MYMACRO` становится неопределённым.

Директива `#error` позволяет прервать компиляцию и выдать сообщение об ошибке, например:

```
#if !defined(FOR_PETROV) && !defined(FOR_SIDOROV)
#error Please define either FOR_PETROV or FOR_SIDOROV
#endif

#ifndef BUFFER_SIZE
#error Please specify the buffer size
#endif
```

Для случаев, когда текст на Си представляет собой результат компиляции программы с какого-то другого языка, предусмотрена директива `#line`, позволяющая заставить компилятор выдавать любую диагностику, как если бы он обрабатывал строку с заданным номером из файла с заданным именем; в обычных программах, изначально написанных на Си, эта директива никогда не используется.

Наконец, стоит упомянуть несколько макросимволов, которые определены изначально (компилятором) и не могут быть ни переопределены, ни отменены. Это символы `__LINE__`, `__FILE__`, `__DATE__` и `__TIME__`, значения которых соответствуют номеру текущей строки, имени текущего файла, текущей дате и текущему времени; номер строки представляется как целочисленная константа, остальные величины — как строковые литералы. Ещё один предопределённый макросимвол, `__STDC__`, равен единице (якобы это позволяет проверить, соответствует ли компилятор стандарту).

4.6.8. Директива `#include`

До сих пор мы сталкивались только с одним вариантом директивы `#include`, в которой имя включаемого файла обрамлялось угловыми скобками, например:

```
#include <stdio.h>
```

Препроцессор предусматривает ещё одну форму этой директивы, с двойными кавычками вместо угловых скобок:

```
#include "mymodule.h"
```

С этими двумя вариантами в наше время связана изрядная путаница, причём бóльшую часть этой путаницы, как водится, внесли всё те же особо опасные международные террористические организации, называемые по недоразумению «стандартизационными комитетами».

Изначально вариант с кавычками был предназначен для «своих» заголовочных файлов, то есть таких, которые написаны для той же программы, что и файл, из которого происходит включение, тогда как вариант с угловыми скобками — для заголовочных файлов, описывающих возможности внешних библиотек. Это простое и понятное правило почему-то не давало покоя членам комитетов по разработке стандартов, которые с упорством, достойным лучшего применения, на протяжении вот уже третьего десятка лет пытаются убедить публику, что так называемая «стандартная библиотека» вопреки здравому смыслу и техническим реалиям является частью языка Си.

Если вас уже успели в этом убедить, попробуйте ответить на один простой вопрос: если функция `printf` является *частью языка Си*, то на каком языке написана сама функция `printf`? Если признать, что она написана именно на Си, то входить в него составной частью она никак не может. Более того, ядра операционных систем пишутся без использования стандартной библиотеки; находятся люди, утверждающие на этом основании, что ядра, стало быть, написаны *не на Си*. Если учесть, что изначально Си был предназначен именно для написания ядра Unix, можно ответить этим мракобесам, что уж ядра-то точно написаны на Си, а вот то, что вы, господа, подразумеваете под «языком Си» — это не Си, а некий плод вашего чрезмерно воспалившегося воображения.

Одним из проявлений комитетского неприятия чёткой границы между языком и библиотеками стала фиксация на уровне стандарта совершенно безумных нормативов, согласно которым вариант `#include` с угловыми скобками предназначен для «указания наборов возможностей из стандартной библиотеки», причём создатели компиляторов имеют право реализовывать этот вариант безо всяких файлов, т. е., например, заголовочный файл `stdio.h` в соответствии с этими новыми веяниями имеет право вообще не существовать; компилятор, построенный согласно этой идее, при виде хорошо знакомой нам строки `#include <stdio.h>` вместо поиска и включения заголовочного файла должен каким-то образом «включить» встроенное в него знание об объектах (функциях, типах и глобальных переменных), которые стандарт полагает описанными в рамках `stdio.h`.

К счастью, ни один компилятор не использует такой подход к реализации, поскольку это сделало бы его полностью непригодным для серьёзной работы. Но определённые негативные последствия от неуёмной творческой активности стандартизаторов всё же наблюдаются; так, в результате всех этих нововведений в воздухе повис вопрос о том, каким же образом следует подключать заголовочные файлы *от сторонних библиотек*. В самом деле, для «своих» файлов остаётся несомненным использование `#include "..."`, для «системных» заголовочников (то есть относящихся к стандартной библиотеке) — `#include <...>`, но как быть с третьим вариантом, с заголовочными файлами, которые не являются ни частью нашего проекта, ни принадлежностью стандартной библиотеки? Ещё совсем недавно ответ на этот вопрос был очевиден: следует использовать вариант с угловыми скобками; сейчас этот ответ приходится снабжать оговоркой, что работоспособность такого решения обеспечивается исключительно здравым смыслом разработчиков компиляторов, которым хватает ума игнорировать наиболее одиозные комитетские рекомендации.

Если рассмотреть вопрос с сугубо технических позиций, можно заметить, что компилятор `gcc` и многие другие компиляторы применяют один и тот же подход к обработке двух вариантов директивы `#include`. Во время сборки самого компилятора в него зашивается некий фиксированный список «системных» директорий; кроме того, дополнительные директории для поиска заголовочных файлов можно указать через параметры командной строки, обычно с помощью ключа `-I`. Для директивы `#include <...>` компилятор выполняет поиск заданного файла сначала в директориях, указанных через командную строку, а затем в системных директориях; для директивы `#include "..."` сначала проверяется наличие указанного файла в **одной директории с файлом, из которого он включается** (а не в текущей директории, из которой запустили компилятор, как это часто неверно полагают), или, точнее, наличие включаемого файла с заданным путём относительно директории, в которой находится включающий файл. Если же такого файла нет, то далее компилятор делает то же самое, что и для варианта с угловыми скобками — проверяет сначала директории, заданные в командной строке, а затем системные директории.

Очевидно, что заголовочный файл, принадлежащий сторонней библиотеке, искать в одной директории с нашими исходниками бессмысленно; больше того, если библиотека установлена в системе, то, как правило, её заголовочники доступны в тех директориях, которые компилятор считает «системными»; в частности, практически всегда для библиотек, доступных в виде пакетов в дистрибутивах Linux, заголовочные файлы устанавливаются в директорию `/usr/include`, то есть туда же, где находится `stdio.h` и прочие файлы, нежно любимые стандартизаторами.

Итак, несмотря на все потуги стандартизаторов испортить окружающий мир, рабочее решение остаётся прежним: **для включения своих собственных заголовочных файлов следует использовать вариант `#include` с двойными кавычками, а для включения любых заголовочников, не являющихся частью нашего проекта — вариант с угловыми скобками, причём вне всякой зависимости от «системности» включаемого заголовочника**.

4.6.9. Особенности оформления макродирективов

С макродирективами связан достаточно важный момент, касающийся оформления текста. Дело в том, что макродирективы, вообще говоря, *находятся вне структуры текста программы*. Во время работы макропроцессора текст ещё не до конца сформирован, даже сама структура текста программы, подчёркиваемая структурными отступами, может ещё измениться. Исходя из этого, приходится признать некорректным рассмотрение макродирективов наравне с остальным текстом. Например, фрагмент, подобный следующему, можно встретить разве что в программах новичков:

```
void f( /* ... */ )
{
    /* ... */
    if(!key_already_known) {
        #ifdef DEBUG_PRINT
        fprintf(stderr, "DEBUG: figuring out the key\n");
        #endif
        find_the_key();
    }
    /* ... */
}
```

Хотя компилятор позволяет в строках макродирективов помещать любое количество пробелов перед символом «решётки», так обычно не делают и пишут макродирективы в крайней левой позиции, не обращая внимания на окружающий текст программы — ведь во время работы макропроцессора этот текст ещё не проанализирован и на макроподстановки повлиять не может:

```
void f( /* ... */ )
{
    /* ... */
    if(!key_already_known) {
#ifdef DEBUG_PRINT
        fprintf(stderr, "DEBUG: figuring out the key\n");
#endif
        find_the_key();
    }
    /* ... */
}
```

Пока макродирективы сохраняют простую структуру, их вообще не снабжают отступами ни в каком виде, но бывают и такие ситуации, когда макродирективы во избежание путаницы могут потребовать *своих собственных* структурных отступов. В таких случаях можно воспользоваться тем, что между символом «решётки» и самой макродирективой тоже допускаются пробелы. «Решётка» ставится по-прежнему в крайней левой позиции, но макродирективы, *вложенные* в конструкцию из других макродиректив, сдвигаются вправо. Размер отступов для макродиректив может отличаться от основного размера отступов в программе. Например, возможна такая ситуация:

```
#define MAX_ARRAY_FOR_BUBBLE 30
#ifdef FIXED_ARRAY_SIZE
    int array[ARRAY_SIZE];
#else
    int *array = malloc(sizeof(*array) * arrsize);
#endif
        /* ... */
#ifdef FIXED_ARRAY_SIZE
#  if ARRAY_SIZE > MAX_ARRAY_FOR_BUBBLE
    quick_sort_int(array, ARRAY_SIZE);
#  else
    bubble_sort_int(array, ARRAY_SIZE);
#  endif
#else
    if(arrsize > MAX_ARRAY_FOR_BUBBLE)
        quick_sort_int(array, arrsize);
    else
        bubble_sort_int(array, arrsize);
#endif
```

Здесь макродирективы условной компиляции, составляющие внутренний `#if`, сдвинуты на два пробела относительно объемлющих макродиректив. Вообще говоря, подобные ситуации возникают редко, к тому же их стоит по возможности избегать, но если что-то подобное всё-таки возникло в вашей программе, желательно знать, что делать.

4.7. Раздельная трансляция

4.7.1. Общая схема раздельной трансляции в Си

Мы уже знакомы с построением программ из отдельных модулей на примере Паскаля и языка ассемблера. Средства разбиения на модули, присутствующие в Си, существенно ближе к той *раздельной трансляции*, которую мы использовали при работе на языке ассемблера, нежели к настоящей *модульности*, примером которой можно считать версии Паскаля, поддерживающие создание `unit`'ов. В сущности, поддержки модулей в Си нет вообще; всё, что есть — это средства, позволяющие делать те или иные *символы* (функции и глобальные переменные) видимыми или невидимыми для редактора связей, а также ссылаться на символы, отсутствующие в текущей единице трансляции, которые редактор связей должен потом откуда-то добыть. Всё остальное достигается использованием препроцессора и целой системы *хаков*, причём многие программисты, пишущие на Си, ухитряются в упор не видеть, что имеют дело именно с хаками, а не с чем-то иным.

Общая идея раздельной трансляции в Си такова: мы можем разбросать наши функции и глобальные переменные по разным исходным файлам, каждый из которых представляет собой текст на языке Си и обычно имеет соответствующий суффикс («`.c`») в имени. Функция с именем `main` должна при этом присутствовать только в одном из файлов, и этот файл обычно называют «главным». Далее каждый из файлов компилируется отдельным запуском компилятора, при этом компилятору сообщается, что в данном случае от него требуется только сгенерировать объектный файл, а попыток сборки финальной программы мы от него не ждём; например, компилятор `gcc` для этого поддерживает флаг командной строки «`-c`». Результатом такой компиляции становится объектный модуль, причём имя для него компилятор обычно выбирает сам, заменяя суффикс «`.c`» на суффикс «`.o`». Например, если один из наших модулей называется `mod.c`, то откомпилировать его мы можем командой

```
gcc -Wall -g -c mod.c
```

— в результате чего, если не произойдёт ошибок, в текущей директории появится файл `mod.o`. Далее у нас есть два варианта, как поступить с «главным» исходным файлом: мы можем совместить его компиляцию с вызовом редактора связей для сборки исполняемого файла, либо можно откомпилировать его точно так же, как все остальные модули, а для вызова редактора связей запустить компилятор `gcc` лишний раз. В обоих случаях в командной строке компилятора необходимо указать все объектные модули, которые следует использовать для построения исполняемой программы. К примеру, если наша программа состоит из главного модуля `prog.c` и двух дополнительных модулей `mod1.c` и `mod2.c`, мы можем собрать её так:

```
gcc -Wall -g -c mod1.c
gcc -Wall -g -c mod2.c
gcc -Wall -g prog.c mod1.o mod2.o -o prog
```

или так:

```
gcc -Wall -g -c mod1.c
gcc -Wall -g -c mod2.c
gcc -Wall -g -c prog.c
gcc prog.o mod1.o mod2.o -o prog
```

Теоретически без последнего запуска компилятора можно обойтись, запустив сразу редактор связей (напомним, он называется `ld`), но для этого нам надо знать, где находится файл, содержащий стандартную библиотеку Си. Компилятор этим знанием уже обладает, так что проще поручить запуск редактора связей ему, а не делать это самим.

4.7.2. Видимость объектов из других модулей

Напомним, что при построении раздельно транслируемой программы на языке ассемблера NASM мы использовали директивы `global` и `extern`; первая делает метку из текущего модуля видимой для редактора связей (и, как следствие, для других модулей), а вторая объявляет метку, которой в текущем модуле нет, но наличие которой предполагается в каком-то другом модуле той же программы. Все остальные метки, не попавшие под действие ни одной из этих директив, NASM считал локальными для текущей единицы трансляции и оставлял их невидимыми для редактора связей.

Подход, принятый в Си, прямо противоположен: по умолчанию компилятор делает все «глобальные объекты» (то есть функции, а также глобальные переменные) видимыми для редактора связей; аналога директивы `global` в Си, соответственно, нет, но зато есть противоположное по смыслу ключевое слово `static`: глобальные объекты, помеченные этим словом, для редактора связей останутся невидимы, так что добраться до них из других единиц трансляции невозможно. Например, если мы уверены, что какая-то функция из написанных нами настолько специфична, что её вызов из других модулей никогда не понадобится, мы можем сделать её полностью невидимой извне нашего модуля:

```
static int very_local_subroutine(int x, const char *n)
{
   /* ... */
}
```

Точно так же мы можем поступить и с глобальной переменной; конечно, лучше вообще их не использовать, но если всё же пришлось, стоит подумать на тему ограничения доступа к ней: глобальная переменная, видимая только в одном модуле, всё же лучше, чем глобальная переменная, к которой могут обращаться функции из других модулей. Достигается это с помощью того же слова `static`:

```
int very_bad_global_var;     /* видно отовсюду, очень плохо */
static int a_bit_better_var; /* только из текущего модуля */
```

Для доступа к функциям и переменным, которые описаны в других модулях, в текущем модуле помещают их ***объявления***; объявление, как мы помним, сообщает компилятору, что соответствующий объект *где-то существует* и можно не пугаться его отсутствия, оставив решение проблемы редактору связей. Объявлением для функции, как мы уже знаем, служит её заголовок; так, если в одном из исходных файлов описана функция

```
int cube(int x)
{
    return x * x * x;
}
```

— то в любом другом из наших исходных файлов мы можем поместить её заголовок:

```
int cube(int x);
```

и использовать её обычным образом, не опасаясь ошибок. Компилятор, обрабатывая вызов такой функции, проверит соответствие типов фактических параметров заявленным в заголовке, а также соответствие типа возвращаемого значения контексту вызова, после чего, если всё хорошо, спокойно откомпилирует такой вызов, оставив в объектном коде требование к редактору связей добыть где-нибудь функцию с таким именем и подставить её адрес в инструкцию вызова. Видеть тело функции компилятору для этого не нужно; собственно говоря, в этом и состоит преимущество раздельной трансляции: при обработке одного исходного файла компилятор не тратит время на анализ кода, содержащегося в других файлах. Конечно, остальные файлы тоже придётся откомпилировать, но если компиляция уже проходила и у вас есть объектные файлы для всех ваших модулей, то при внесении изменений только в один или два из них вам не придётся компилировать их все: достаточно будет перекомпилировать те модули, которые изменились, а затем выполнить финальную сборку, которая происходит достаточно быстро даже для очень больших программ.

Несколько сложнее обстоят дела с объявлением (в отличие от описания) глобальных переменных. Если в *другом* модуле есть глобальная переменная, к которой вы хотите получить доступ из текущего исходного файла, нужно в этом файле повторить описание этой переменной, добавив к нему ключевое слово `extern`:

```
extern int very_bad_global_variable;
```

После этого её можно будет использовать, как если бы она была описана в текущем модуле. Собственно говоря, это и есть *объявление*.

Пользуясь случаем, напомним, что глобальные переменные есть частный случай абсолютного вселенского зла, а сцепленность модулей по глобальным переменным — это самый худший вариант взаимодействия между модулями. Это, впрочем, не относится к константам, то есть таким переменным, значение которых никогда не меняется. Например, если в одном модуле есть глобальный константный массив

```
const int prime_numbers[] = { 2, 3, 5, 7, 11, 13, 17 /* ... */
```

— то в любом другом модуле можно написать

```
extern const int prime_numbers[];
```

и спокойно использовать уже готовую таблицу простых чисел из другого модуля, вместо того чтобы делать свою. Такое использование ничем не плохо, ведь переменная, которая не меняется, не может *накапливать состояние* и оказывать неочевидное влияние на работу функций — как раз всё то, из-за чего глобальные переменные нежелательны.

4.7.3. Заголовочные файлы к модулям

Допустим, у нас есть модуль `m1.c`, в котором описан десяток-другой функций; этими функциями активно пользуются ещё пять-шесть модулей. Если буквально следовать инструкциям, предложенным в предыдущем параграфе, то в каждом из этих пяти-шести модулей нам придётся поместить заголовки всех функций из `m1.c`, которые в них используются. По мере увеличения проекта будет расти и общее количество таких заголовков, и расти оно будет *нелинейно*; в какой-то момент мы можем обнаружить, что такие заголовки составляют уже процентов двадцать всего объёма кода. Само по себе это ещё не очень страшно, но ведь функции иногда меняются, причём может измениться в том числе и информация, представленная в заголовке: мы можем принять решение добавить или убрать параметр, изменить тип параметра или даже тип возвращаемого значения. Каждый раз, когда мы вносим подобное изменение, нам придётся найти все файлы нашей программы, в которых имеется копия изменившегося заголовка, и соответствующим образом их отредактировать. Если забыть это сделать, *программа благополучно откомпилируется и соберётся*, но работать, разумеется, будет неправильно; в таких случаях возникают настолько феерические ситуации, что неопытные программисты, оторвавшись от экрана с отладчиком, недоумённо восклицают: «Но ведь такого не может быть!».

В самом деле, пусть у нас была функция `int func(int x, int y);`, и именно такой заголовок фигурировал в нескольких модулях. Мы приняли решение добавить в эту функцию третий параметр, `int z`, и отредактировали все эти модули, что само по себе нетривиально — про существование некоторых из них мы можем даже не догадываться, ведь над серьёзными программами обычно работает несколько человек, и кто-то вполне мог воспользоваться нашей функцией, не сообщив нам об этом, а если и сообщил, мы об этом, скорее всего, тут же забыли. Но, так или иначе, для поиска модулей, использующих функцию, есть много довольно простых инструментов, начиная с программы `grep`, которая позволяет произвести простой текстуальный поиск по всему нашему набору исходников. Итак, мы отредактировали все модули — кроме одного, про который случайно забыли. Во всех модулях, кроме этого одного, компилятор точно знает, что функция `func` должна вызываться от трёх параметров, так что, пока мы не исправим все её вызовы, добавив третий параметр, компилятор будет выдавать ошибку; но вот при компиляции последнего оставшегося модуля компилятор по-прежнему уверен, что параметров там всего два, и именно так в этом модуле функция `func` и вызывается. Сама функция, понятное дело, ожидает наличие в стеке трёх параметров, а не двух. Третий параметр она из стека в любом случае извлечёт, но вряд ли кто возьмётся предсказать, что там окажется, когда функцию вызовут с двумя параметрами вместо трёх.

Ещё интереснее будет картина, если мы решим изменить тип одного из параметров с `int` на `double` и при этом где-то забудем подредактировать заголовок; ну а при смене типа *возвращаемого* значения (опять же, если забыть где-то это исправить) можно ожидать полного хаоса. Впрочем, степень полноты хаоса — вопрос философский; любой из перечисленных вариантов ошибки может стоить вам нескольких потерянных дней.

Во избежание таких ошибок программисты обычно не пишут в файлах модулей заголовки функций из других модулей. Вместо этого они применяют ***заголовочные файлы***; в большинстве случаев отдельным заголовочным файлом снабжают каждый модуль, кроме главного, то есть на каждый файл с суффиксом `.c` за исключением файла, содержащего функцию `main`, заводят файл с таким же именем, но с суффиксом `.h` (например, для модуля `m1.c` создают заголовочный файл `m1.h`). **В заголовочный файл выносят объявления всех объектов, описанных в модуле, к которым предполагаются обращения из других модулей**; иначе говоря, если в вашем модуле имеется функция `func` и вы предполагаете, что в других модулях к этой функции будут обращения, то её заголовок нужно вынести в заголовочный файл вашего модуля. Аналогично, если в вашем модуле описана глобальная переменная и вы предполагаете её доступность из других модулей (вы хорошо подумали?), то в заголовочном файле следует поместить её *объявление* со словом `extern` (в отличие от *описания* без слова `extern`, которое остаётся в самом модуле).

Теперь достаточно включить этот заголовочный файл с помощью директивы `#include` (с кавычками) в каждый из модулей, использующих возможности вашего модуля, и дело, как говорится, в шляпе: при трансляции всех этих модулей компилятор будет видеть ваши объявления, но сами эти объявления не «размножаются», как описывалось в начале параграфа, а существуют в одном экземпляре — в вашем заголовочном файле. Если теперь вам придёт в голову поменять параметры какой-то из ваших функций, вам не придётся искать её заголовки по всем модулям программы, достаточно будет отредактировать ваш собственный заголовочник — то есть *один* файл

(конечно, не считая самого вашего модуля). Все случаи несоответствия параметров вызовов параметрам нового заголовка вам укажет компилятор в процессе пересборки программы.

Отметим ещё два очень важных момента. Во-первых, **в модуль обязательно нужно включать (с помощью `#include`) *свой собственный* заголовочный файл**. В этом случае компилятор при обработке вашего модуля сначала увидит объявления функций и переменных, содержащиеся в заголовочнике, а потом их же *описания*, находящиеся в тексте самого модуля; это позволит компилятору проверить их соответствие и выдать ошибку, например, если вы изменили профиль какой-нибудь функции в модуле, но отразить изменения в заголовочном файле забыли.

Во-вторых, **все функции и глобальные переменные, которые вы решили не выносить в заголовочник, нужно обязательно пометить модификатором `static`** (см. стр. 62), тем самым сделав их невидимыми для редактора связей. Глобальное пространство имён у вас одно на всю программу, при этом программисты, как правило, обращают внимание только на содержимое заголовочников, не вникая в то, что написано в модулях, особенно если эти модули написаны другими участниками работы; имя, которое вы не вынесли в заголовочник, но оставили видимым, может стать причиной неожиданного конфликта имён, ведь кто-то из ваших коллег вполне может задействовать такое же имя в одном из своих модулей. Вообще говоря, все глобальные имена, описанные в вашем модуле, делятся на две категории: те, которые вы экспортируете, и те, которые вы описали для использования в самом модуле; первые выносятся в заголовочник (для функций — прототипы, для переменных — описания со словом `extern`), вторые помечаются словом `static`. Если не сделать ни того, ни другого, можно будет предположить, что вы сами не знаете, для чего ввели то или иное имя.

Единственным исключением из этого правила является функция `main`: в заголовочник её, как правило, не выносят, но редактор связей должен её видеть, иначе он не сможет собрать исполняемый файл.

4.7.4. Защита от повторного включения

Пока ваши модули экспортируют только функции и глобальные переменные, всё сравнительно просто; к сожалению, в реальной жизни так не бывает. Скорее всего, уже на ранних стадиях работы вам потребуется ввести какой-нибудь *пользовательский тип*, будь то структура, объединение или перечисление, с которым будут работать ваши функции и/или который будет использован при описании глобальных переменных. Такой тип потребуется сделать доступным более чем в одном модуле; как следствие, его описание придётся вынести в заголовочный файл. Больше того, если в *другом* модуле появятся экспортируемые функции, работающие с вашим типом, вам может потребоваться включить *один заголовочный файл из другого*: в том заголовочнике, где объявлены функции, использующие ваш тип, придётся предусмотреть директиву `#include`, включающую заголовочник, содержащий описание этого типа.

Кроме пользовательских типов, существуют ещё макросы; их тоже приходится с целью экспорта выносить в заголовочные файлы, и они тоже (хотя и существенно реже) могут оказаться причиной для включения одного заголовочника из другого.

Можно выделить общее правило для выноса тех или иных сущностей в заголовочные файлы:

- **всё, что *занимает память* в какой-либо из секций будущей исполняемой программы, *описывается* в файле реализации модуля (файл с суффиксом `.c`) и либо *объявляется* в заголовочном файле (если экспортируется), либо помечается словом `static` (если объект предназначен только для внутреннего использования в модуле);**
- **всё, что *не занимает памяти*, или, иначе говоря, всё, что используется только во время компиляции и затем бесследно исчезает, *описывается* в заголовочном файле (либо, если предназначено только для внутреннего использования в модуле, описывается в файле реализации)**.

К первой категории сущностей относятся функции и глобальные переменные, ко второй — типы и макросы. Функции объявляются своими заголовками-прототипами, переменные объявляются с помощью слова `extern`, что до описаний типов и макросов, то они в заголовочном файле выглядят точно так же, как и везде.

Теперь представьте, что вам потребовалась какая-то функция, написанная вашим коллегой в другом модуле. Заглянув в документацию или просто задав вопрос коллеге, вы узнали, что эта функция находится в модуле `ddd.c`, который снабжён заголовочным файлом `ddd.h`. Ваш логичный следующий шаг — вставить в свой модуль директиву `#include "ddd.h"` и, воспользовавшись нужной функцией, продолжить работу. При этом вряд ли вы станете подробно изучать файл `ddd.h`, а в нём, вполне возможно, тоже содержится директива `#include`, включающая ещё какой-нибудь заголовочник, например, `bbb.h`, а вам ещё месяц назад потребовались какие-то функции из модуля `bbb`, так что в вашем собственном тексте тоже есть `#include "bbb.h"`. В результате компилятор в процессе обработки вашего модуля «увидит» `bbb.h` дважды; это называется ***повторным включением*** заголовочного файла.

Если бы `bbb.h` содержал только объявления переменных и функций, ни к чему страшному такое повторное включение не привело бы, только чуть-чуть замедлило бы компиляцию. Но это вряд ли наш случай: заголовочники, в которых содержатся только функции и переменные, нет никакого смысла включать из других заголовочников. Так что если повторное включение произошло — видимо, в повторно включённом заголовочном файле описаны пользовательские типы и (или) макросы. В случае с типом компилятор попросту выдаст ошибку, ведь он увидел второе *описание* того же самого имени, а описывать одно и то же имя дважды нельзя (в отличие от объявлений; объявлять одно и то же имя можно сколько угодно раз, главное, чтобы эти объявления друг другу не противоречили). В случае с макросом компилятор выдаст предупреждение, а не ошибку, но это тоже не слишком хорошо; в большинстве проектов действует требование полного отсутствия предупреждений при компиляции.

Первое, что приходит в голову — разобраться, какой из файлов включён повторно (это сделать довольно просто: компилятор при выдаче ошибки или предупреждения сам скажет, в каком файле и какой строке возникла проблема) и убрать из своего модуля соответствующую директиву `#include`, в данном случае ту, которая включает `bbb.h`; в самом деле, коль скоро этот файл всё равно включается через `ddd.h`, то этого нам уже достаточно и нет повода включать его самим. Однако и тут реальность оказывается несколько сложнее, нежели на первый взгляд: представьте себе ситуацию, когда вы включили в свой модуль файлы `bbb.h` и `ddd.h` и никакого повторного включения не произошло, а уже позже, возможно, через полгода или год, кто-то решил усовершенствовать файл `ddd.h` и в процессе совершенствования добавил там включение `bbb.h`, в результате чего ваш ни в чём не повинный модуль, который вы уже полгода как не трогали, вдруг перестал компилироваться. Возможны и более заковыристые ситуации; придумать их мы предложим читателю в качестве головоломки.

Как показывает практика, бороться вручную с каждым случаем повторного включения — идея совершенно неработоспособная, так что программистам пришлось изобрести некую технику, автоматически не позволяющую заголовочникам включаться в одну и ту же единицу трансляции больше одного раза. Для этого используется препроцессор, а точнее — его возможности *условной компиляции*, описанные в §4.6.6. Для каждого заголовочного файла выбирается некое уникальное имя макросимвола, обычно включающее в себя само имя файла, приведённое к верхнему регистру; так, автор этих строк обычно использует суффикс «`_SENTRY`», так что для заголовочника `mymodule.h` получается что-то вроде `MYMODULE_H_SENTRY`, но возможны и другие подходы. Главное — позаботиться о том, чтобы вероятность появления где-то символа с таким же именем была достаточно близка к нулю. Далее в начале заголовочного файла (то есть прямо первыми двумя его строками; возможно, после комментария, в котором описывается, что́ это за файл и для чего он нужен) пишутся директивы

```
#ifndef MYMODULE_H_SENTRY
#define MYMODULE_H_SENTRY
```

а последней строкой того же заголовочного файла ставится

```
#endif
```

Тем самым всё «значащее» (то есть как-то влияющее на компиляцию) содержимое нашего заголовочника оказывается обрамлено директивой условной

компиляции. При обработке *первого* случая включения данного файла в отдельно взятой единице трансляции символ `MYMODULE_H_SENTRY` ещё не определён, так что содержимое заголовочника компилятором обрабатывается, причём, в числе прочего, срабатывает и директива `#define`, которая определяет символ. Если *в той же единице трансляции* файл `mymodule.h` включается второй раз (а равно и третий, и четвёртый, и последующие) — к этому времени символ `MYMODULE_H_SENTRY` уже определён, так что директива `#ifndef` исключает содержимое файла из компиляции, и компилятор его повторно не увидит. Иначе говоря, повторное включение (физически) происходит, но никак не влияет на ход компиляции.

Интересно, что эта техника защиты от повторного включения, известная под названием «защитные макросы» (*sentry macros*), многими программистами на Си воспринимается как должное, то есть люди не видят подлинной сущности этой техники; между тем это не более чем очередной *хак*, использование подвернувшегося под руку инструмента совершенно не по его прямому назначению, но приводящее при этом к нужному результату.

Как мог заметить читатель, использование защиты от повторного включения не всегда оказывается обязательным; если очередной заголовочный файл содержит только объявления функций и переменных, то его повторное включение ни к ошибкам, ни даже к предупреждениям не приведёт. Однако исходные файлы нашей программы, в том числе заголовочные — это отнюдь не мраморные скрижали, они достаточно часто и активно меняются; если мы будем снабжать защитой от повторного включения только такие заголовочники, в которых присутствуют типы или макросы, то каждый раз, вставляя в какой-нибудь заголовочный файл описание очередного нового типа или макроса, мы будем вынуждены проверять, есть ли уже в этом заголовочнике защита от повторного включения. Проще (и дешевле по трудозатратам) снабжать *каждый* заголовочный файл защитой от повторного включения сразу, как только этот файл создан. Написать три короткие строки текста в пустом файле — вопрос нескольких секунд, при этом *каждая* проверка наличия этих строчек в файле более-менее осмысленного размера может занять столько же и даже больше.

Итак, **каждый заголовочный файл следует снабжать директивами защиты от повторного включения сразу же после его создания**; это один из тех редких случаев, когда думать вредно. Ещё одно действие, выполняемое всегда сразу после создания файла модуля – это добавление директивы `#include`, включающей в модуль его собственный заголовочный файл. Вообще, процедуру создания в программе нового модуля (назовём его `newmod`) можно описать так:

- создайте пустые файлы `newmod.c` и `newmod.h`;
- вставьте в файл `newmod.c` директиву
  ```
  #include "newmod.h"
  ```
- вставьте в файл `newmod.h` директивы
  ```
  #ifndef NEWMOD_H_SENTRY
  #define NEWMOD_H_SENTRY


  #endif
  ```
  — оставив между ними пустое пространство для содержимого заголовочного файла;
- зарегистрируйте созданные файлы в системе контроля версий и системе сборки;
- отразите появление нового модуля в документации.

4.7.5. Объявления типов; неполные типы

До сих пор мы сталкивались с *объявлениями* переменных и функций; для пользовательских типов, включая структуры, объединения и перечисления, мы использовали только *описания*, то есть конструкции, задающие исчерпывающую информацию о типе. Как ни странно, структуры в Си также можно *объявлять*, то есть сообщать компилятору о существовании типа с заданным именем, не показывая собственно сам тип. Например, можно написать:

```
struct item;
```

С этого момента компилятор будет знать, что словосочетание `struct item` обозначает некий структурный тип, но больше ничего про этот тип знать не будет; такой тип называется ***неполным*** (англ. *incomplete type*). Точно так же можно описать неполный тип-объединение, но это обычно не используется.

Конечно, описать переменную неполного типа компилятор не позволит, ведь он даже не знает, сколько памяти под такую переменную необходимо выделить; однако неполный тип можно вполне успешно использовать для описания *указателей*: в самом деле, размер указателя обычно не зависит от того, на переменную какого типа он указывает.

Основное назначение таких объявлений состоит в том, чтобы позволить, например, двум структурам иметь в качестве полей указатели друг на друга. Более того, мы уже использовали неполные типы, когда для работы со списками и деревьями описывали структуры с полями-указателями на структуру того же типа; например, в знакомом нам описании

```
struct item {
    int data;
    struct item *next;
};
```

на тот момент, когда компилятор видит описание поля `next`, сам тип `struct item` ещё не описан, ведь его описание ещё не закончилось; с другой стороны, поскольку словосочетание `struct item` компилятор уже видел, этот тип считается *объявленным* (хотя и неполным), и компилятор полагает, что для описания указателя этого достаточно.

Вообще говоря, писать объявление структуры отдельной конструкцией особого смысла нет, поскольку использовать такой неполный тип можно лишь указав его полное имя, включающее слово `struct` (в данном случае `struct item`), а это уже само по себе послужит для компилятора объявлением типа. Как следствие, тип `struct item*` можно использовать, даже если до сей поры компилятор вообще ничего не слышал о слове `item`.

Кроме очевидного применения для создания структур со ссылками на себя или друг на друга, неполные типы дают ещё одну важную возможность: *не показывать компилятору описание типа, если нужен только указатель на такой тип*. Это часто позволяет написать объявление структуры (неполной), вместо того чтобы включать с помощью `#include` тот файл, где находится её полное описание. Так можно достичь изрядной экономии времени компиляции, выкинув лишние включения, в особенности если речь идёт о включении одного заголовочного файла из другого; кроме того, никакие два заголовочных файла по понятным причинам не могут взаимно включать друг друга, так что объявления типов оказываются просто незаменимы.

Иногда описание структуры вообще не выносят в заголовочный файл, несмотря на то, что в нём содержатся прототипы функций, работающие с указателями на эту структуру. Такой вариант применяется, если автор модуля предлагает пользователю хранить указатели на какую-то его структуру, например, для целей идентификации, но при этом предпочитает, чтобы пользователь не обращался к полям этой структуры напрямую. Кстати, примером такой ситуации является хорошо знакомый нам тип `FILE*`.

Отметим, что большинство компиляторов позволяют также объявить тип `enum`, но это в явном виде запрещено спецификациями и к тому же совершенно бессмысленно: проще тогда использовать обыкновенный `int`.

4.8. Язык Си и стиль кода

Во всём предшествующем тексте, начиная с первого тома, мы уделяли правилам оформления текста программы самое пристальное внимание, но тем не менее нам по-прежнему есть что обсудить в этой области. Мы надеемся, что к настоящему времени читатель уже приобрёл достаточный опыт в написании программ и лучше подготовлен к восприятию некоторых тонкостей.

4.8.1. Вспоминаем общие принципы

Перечислим основополагающие правила, которые нам известны (в основном ещё из первого тома). Прежде всего напомним, что **текст программы предназначен в первую очередь для прочтения человеком, и лишь во вторую — для обработки компьютером**. Обеспечение хорошей читаемости текста — задача первоочередная; программа, в тексте которой нарушены правила оформления, вообще не может рассматриваться в качестве правильной или неправильной, работающей или неработающей, полезной или бесполезной. До тех пор, пока нарушения в оформлении не будут устранены, они представляют собой *единственное* свойство программы, которое можно обсуждать, а их устранение — это единственное, что с такой программой допустимо делать.

К тексту программы на любом языке программирования предъявляется несколько универсальных требований. Текст должен содержать только символы из набора ASCII; комментарии, если они есть, должны быть написаны по-английски (не по-немецки, не по-французски и тем более не по-русски транслитом, а именно по-английски); при выборе имён для идентификаторов

```
while(p)
{
    s += p->data;
    p = p->next;
}
```

```
while(p) {
    s += p->data;
    p = p->next;
}
```

```
while(p)
    {
        s += p->data;
        p = p->next;
    }
```

Рис. 4.3. Три стиля расположения операторных скобок

также должны использоваться именно английские слова, адекватно отражающие предназначение идентификатора. Короткие имена можно (и нужно) использовать только для локальных идентификаторов в случаях, если их предназначение очевидно из контекста. Комментариями лучше не злоупотреблять; вместо этого следует писать текст самой программы (в частности, выбирать идентификаторы) так, чтобы он пояснял сам себя. Выбору идентификаторов был посвящён §2.12.10.

При написании программы на языке, имеющем структурный синтаксис, то есть подразумевающем вложенные конструкции (а именно таковы и Паскаль, и Си, и большинство других языков программирования) **необходимо визуально выделять вложенные фрагменты, используя пробельные символы в начале каждой строки**. Вложенный текст должен сдвигаться вправо относительно текста, в который он вложен, на фиксированный размер отступа — два, три, четыре пробела или один символ табуляции. Размер отступа выбирается один раз и сохраняется одинаковым во всей программе. Строки, которые начинают и заканчивают цельную конструкцию, должны быть сдвинуты одинаково, то есть их первые непробельные символы обязаны находиться точно друг над другом. Изучая Паскаль, мы начали обсуждать отступы в §2.1, затем по мере введения новых операторов показывали, как их следует оформлять. В частности, введя составной оператор в §2.2.8, мы сразу же продемонстрировали три возможных стиля оформления операторов, в которых в роли тела выступает составной оператор; к этому вопросу мы ещё раз вернулись в §2.12.2. Напомним, что открывающую скобку можно сносить на следующую строку после заголовка оператора, располагая и её, и закрывающую скобку в той же горизонтальной позиции, в которой расположен заголовок оператора — так мы поступали в программах на Паскале. Можно оставлять открывающую скобку на одной строке с заголовком оператора, а закрывающую располагать под началом заголовка — именно так мы действовали в программах на Си. Существует третий стиль, который мы не рекомендуем, но имеем в виду его допустимость: операторные скобки в этом стиле снабжаются дополнительным отступом (см. рис. 4.3).

Ещё одна важная подробность: **операторную скобку, обозначающую начало тела подпрограммы (в терминах Си — функции), всегда сносят на отдельную строку**, даже если при написании сложных операторов (ветвлений и циклов) скобку решено оставлять на одной строке с заголовком; более того, **операторные скобки, обрамляющие тело функции, никогда не сдвигают** — они всегда пишутся в крайней левой колонке текста вне всякой зависимости от избранного стиля. Для Паскаля это требование очевидно, достаточно вспомнить о секциях локальных описаний; для Си, возможно, это требование покажется не столь очевидным, но оно и здесь очень просто объясняется. Дело в том, что, в отличие от заголовка сложного оператора, **заголовок функции имеет самостоятельное значение**, то есть часто встречается отдельно от тела функции — при этом он служит *объявлением* соответствующей функции, в противоположность *описанию*, в котором присутствует тело. Разнесение заголовка и тела на разные строки позволяет подчеркнуть факт самостоятельности заголовка. Что касается дополнительного отступа для всего тела, то он был бы попросту излишним: тело функции может прилагаться к её заголовку, но оно не вложено в заголовок. Со сложными операторами ситуация иная, там имеет место как раз вложение одного оператора в другой, поэтому, несмотря на то, что мы такой стиль не рекомендуем, в пользу дополнительного сдвига составного оператора в составе другого оператора можно найти определённые аргументы.

Итак, следующие варианты оформления будут неправильными:

```
float cube(float a)
    {
        return a * a * a;
    }
```

```
float cube(float a) {
    return a * a * a;
}
```

```
float cube(float a)
{ return a * a * a; }
```

Правильно будет написать так:

```
float cube(float a)
{
    return a * a * a;
}
```

4.8.2. Фирменные особенности Си

Начнём с соглашений об именах идентификаторов. В отличие от Паскаля, язык Си чувствителен к регистру букв, так что никакой свободы в написании ключевых слов здесь нет; например, `if` — это название оператора, тогда как `If`, `iF` и `IF` — это обыкновенные идентификаторы, притом различные.

Из §4.6.3 мы знаем, что в программах на Си имена макросов обычно набираются целиком заглавными буквами, чтобы их было лучше видно, а нужно это по причине опасности макросов, обусловленной их неподчинением общим правилам видимости. Как можно догадаться по приводившимся примерам, все остальные идентификаторы в языке Си обычно набирают целиком в нижнем регистре, т. е. маленькими буквами, а отдельные слова в составе идентификаторов отделяют друг от друга символом подчёркивания. **Идентификаторы, сочетающие в себе буквы обоих регистров, при работе на языке Си обычно не используются**. Если вы увидели программу на Си, в которой имеются идентификаторы «смешанного регистра», такие как `MixedCase`, `isItGood`, `CamelIsAnAnimal`, `exGF` и прочее в таком духе — скорее всего, автор программы редко пишет на Си. Подчеркнём, что сказанное верно лишь для чистого Си; традиции языка Си++ совершенно иные, но о них пока речи не идёт.

Иногда можно встретить, в том числе в системных библиотеках и заголовочных файлах, применение «смешанного регистра» для именования структурных типов, а в некоторых случаях — и для других целей. Было бы неправильно утверждать, что это *недопустимо*, коль скоро ясности программ такой стиль не вредит. Тем не менее мы не рекомендуем использовать такие отступления от традиционного именования.

Несколько особняком стоят идентификаторы, имена которых начинаются с подчёркивания. Традиционно считается, что эти имена *зарезервированы за системой программирования*, то есть, например, разработчики системных заголовочных файлов могут использовать такие имена для своих внутренних нужд, не внося их в документацию. В связи с этим не следует пользоваться именами, начинающимися с подчёркивания, в своих программах: такие имена (например, `_counter`, `_error_code`, `__LINES_NUM` и т. п.) могут вступить в конфликт с заголовочными файлами, поставляемыми вместе с компилятором. Следует учитывать, что, даже если одно конкретное имя ни к каким конфликтам не привело, это не является основанием для дальнейшего его использования: одно из основных достоинств языка Си — *переносимость программ*, а использование потенциально «конфликтных» имён эту переносимость, естественно, снижает, ведь если в вашей версии библиотечных заголовочников соответствующего имени нет, то это не значит, что такое же имя не появится при работе с другим компилятором, на другой платформе, да и просто в одной из будущих версий вашей системы программирования. Итак, **не начинайте имена ваших идентификаторов с символа подчёркивания**, оставьте такое именование системным заголовочным файлам.

Оформление сложных объявлений, описаний и инициализаторов тоже имеет свои особенности. В языке Си часто встречаются описания, имеющие сложный синтаксис: это описания структурных и перечислимых типов, а также снабжённые инициализаторами описания массивов и переменных-структур.

С описанием типа, предполагающим использование фигурных скобок (`struct`, `union`, `enum`) всё довольно просто: единственная степень свободы — положение открывающей фигурной скобки (её, как водится, можно оставить на одной строке с именем структуры, а можно снести на следующую строку). Чего точно не следует делать — это сдвигать фигурные скобки относительно начала описания типа, то есть вот так писать можно:

```
struct item {
    const char *str;
    struct item *next;
};
```

```
struct item
{
    const char *str;
    struct item *next;
};
```

а вот так уже не стоит:

```
struct item
    {
        const char *str;
        struct item *next;
    };
```

Ситуация здесь подобна ситуации с началом и концом тела функции: всё тело целиком не сдвигают даже при использовании стиля, предполагающего сдвиг составного оператора относительно заголовка сложного оператора. Не стоит также и вытягивать описание структуры в одну строку:

```
struct item { const char *str; struct item *next; };
```

А вот описание перечислимого типа вытянуть можно, и смотреться это будет вполне логично, но только в случае, если оно умещается в одну строку и не содержит явно заданных значений констант:

```
enum state { home, whitespace, stringconst, ident, end };
```

Если в одну строку описание не поместилось или если имеется потребность в явном задании целочисленных значений, то лучше будет каждую константу описывать на отдельной строке. Обратите внимание, что для удобства чтения знаки равенства рекомендуется расположить в одну колонку (используйте пробелы, если пользуетесь ими для структурных отступов; если в качестве отступа у вас табуляция, используйте её же для выравнивания знаков равенства):

```
enum state {
    st_home,
    st_whitespace,
    st_stringconst,
    st_ident,
    st_end
};
```

```
enum state {
    st_home         = 0,
    st_whitespace   = 32,
    st_stringconst  = 12,
    st_ident        = 77,
    st_end          = -1
};
```

Как обычно, положение открывающей фигурной скобки определяется избранным стилем. Если вы сносите её в описаниях структур, сделайте то же самое и при описании `enum`'ов:

```
enum state
{
    st_home,
```

```
enum state
{
    st_home         = 0,
```

Аналогично обстоят дела с длинными инициализаторами (например, при описании инициализированных массивов). Если инициализатор полностью умещается в одну строку, лучше его так и записать:

```
const int some_primes[] = { 11, 17, 37, 67, 131, 257 };
```

Если инициализатор в одну строку не поместился, или если к значениям требуются комментарии, можно записать каждый элемент инициализатора на отдельной строке:

```
const unsigned long hash_sizes[] = {
    11,     /* > 8 */
    17,     /* > 16 */
    37,     /* > 32 */
    67,     /* > 64 */
    131,    /* > 128 */
    257,    /* > 256 */
    521,    /* > 512 */
    1031,   /* > 1024 */
    2053    /* > 2048 */
};
```

Нет ничего страшного и в разбиении инициализатора на несколько строк из соображений равномерности:

```
const unsigned long hash_sizes[] = {
    11,       17,       37,       67,       131,     257,
    521,      1031,     2053,     4099,     8209,    16411,
    32771,    65537,    131101,   262147,   524309,  1048583,
    2097169,  4194319,  8388617,  16777259, 33554467
};
```

Положение открывающей фигурной скобки в этих случаях также зависит от избранного стиля, причём, как ни странно, часто её сдвигают, то есть следующий вариант не только допустим, но и довольно популярен:

```
const unsigned long hash_sizes[] =
    {
        11,     /* > 8 */
        17,     /* > 16 */
        37,     /* > 32 */
        67,     /* > 64 */
        131,    /* > 128 */
        257,    /* > 256 */
        521     /* > 512 */
        1031,   /* > 1024 */
        2053    /* > 2048 */
    };
```

Забегая вперёд, отметим, что при инициализации двумерного массива обычно каждый элемент (то есть одномерный массив) стараются записать в одну строку, например:

```
const int change_level_table[5][5] = {
    { 4, 4, 2, 1, 1 },
    { 3, 4, 3, 1, 1 },
    { 1, 3, 4, 3, 1 },
    { 1, 1, 3, 4, 3 },
    { 1, 1, 2, 4, 4 }
};
```

Если же в одну строку кода каждая строка вашей матрицы не помещается (случай сам по себе довольно редкий), придётся их равномерно разбить на несколько строк.

Оформление оператора постусловия в Си тоже имеет свои особенности. Ключевое слово `while` в этом языке используется в двух ролях: в заголовке цикла с предусловием и в эпилоге цикла с постусловием (`do-while`). Сразу после слова `while` в обеих ситуациях следует условное выражение в круглых скобках; в конструкции `do-while` следом за этим выражением идёт точка с запятой, но такая же точка с запятой может обнаружиться и после заголовка обычного `while` — там она будет изображать пустое тело цикла. Итак, глядя на конец цикла `do-while`, можно (и очень легко) перепутать его с циклом `while`, имеющим пустое тело.

Решение этой проблемы довольно очевидно для стиля расстановки фигурных скобок, при котором открывающая скобка не сносится на следующую строку. Прежде всего отметим, что **использование фигурных скобок в цикле `do-while` строго обязательно вне всякой зависимости от количества операторов в этом теле**. Некоторые программисты даже пребывают в уверенности, что это требование языка Си; на самом деле это не так, телом `do-while` может быть любой оператор, не только составной; но эта возможность Си никогда не используется. Проблема со смыслом `while` теперь решается тем, что слово `while` попросту остаётся на той же строке, что и предшествующая ему закрывающая скобка, примерно так, как мы поступали со словом `else`:

```
do {
    get_event(&event);
    res = handle_event(&event);
} while(res != EV_QUIT_NOW);
```

Такое же решение мы можем применить и для стиля, при котором открывающая фигурная скобка сносится на следующую строку, но не сдвигается:

```
do
{
    get_event(&event);
    res = handle_event(&event);
} while(res != EV_QUIT_NOW);
```

Это не слишком изящно, поскольку вынуждает нас вводить исключение из общего правила, однако так всё же лучше, чем путать конец цикла с началом нового цикла. Но вот что делать при использовании стиля, где скобки не только сносятся, но и сдвигаются, оказывается непонятно. Ричард Столлман в GNU Coding Style Guide предлагает оформлять `do-while` так:

```
do
  {
    get_event(&event);
    res = handle_event(&event);
  }
while(res != EV_QUIT_NOW);
```

При всём уважении к Столлману, эта идея крайне неудачна. Если цикл занимает хотя бы десяток строк, при взгляде на слово `while` мы совершенно однозначно не заметим соответствующее ему `do`. Циклы с предусловием встречаются гораздо чаще циклов с постусловием, так что при чтении программы мы подсознательно ожидаем именно обычный `while`. После этого мы примем `while` за заголовок цикла, дальше, возможно, не заметим точку с запятой, а возможно, наоборот, заметим, решим, что это ошибка, «споткнёмся», после чего, просматривая фрагмент уже более внимательным взглядом, найдём, наконец, пресловутое `do`, при этом потратив не меньше секунды на размышления в неправильном направлении и ещё секунду-другую на то, чтобы вернуться к тем мыслям, из которых нас выбил проклятый `while`. Можно определённо сказать, что такие `while`'ы обходятся читателю программы слишком дорого. Поэтому, сколь бы противно сие ни выглядело, мы возьмём на себя смелость рекомендовать что-то вроде следующего:

```
do
  {
    get_event(&event);
    res = handle_event(&event);
  } while(res != EV_QUIT_NOW);
```

Одной этой сложности может быть вполне достаточно, чтобы отказаться от такого стиля расстановки операторных скобок, по крайней мере, для языка Си (для Паскаля этот аргумент не действует, там такой проблемы нет).

Рассмотрим теперь *характерные для Си ошибки в оформлении функции*. Достаточно часто можно видеть программы, в которых описания локальных переменных в функциях, а также заключительный оператор `return` почему-то не сдвинуты на размер отступа, а написаны с крайней левой позиции экрана, примерно так:

```
int main()
{
int i;
const char *hello = "Hello";
const char *goodbye = "Good Bye";
    for (i = 0; i < 10; i++) {
        printf("%s\n", hello);
    }
    for (i = 0; i < 10; i++) {
        printf("%s\n", goodbye);
    }
return 0;
}
```

Такой стиль, разумеется, недопустим. Прежде всего, оператор `return` — это обычный оператор, встречаться он может не только в самом конце функции, но и в её середине, будучи при этом вложенным в циклы и ветвления; нет никаких оснований считать, что случай возврата значения в последней строчке функции чем-то принципиально отличается от других ситуаций, когда `return` встречается в коде.

Что касается переменных, то, конечно, их описания *операторами* не являются[34], но при таком стиле форматирования фигурная скобка, открывающая функцию, начинает сливаться с окружающим кодом; совершенно неясно, какого преимущества хотят добиться сторонники такого странного форматирования.

4.8.3. Побочные эффекты

Изучая Паскаль, мы всегда явным образом отмечали каждый из редких случаев возникновения ***побочного эффекта***. Напомним, что о побочных эффектах говорят, когда *при вычислении выражения* произошло что-то ещё, кроме собственно создания результата; это может быть изменение значений каких-то переменных, операции ввода-вывода, вообще любое изменение, которое может быть позже обнаружено (см. т. 1, §2.3.6).

Хотелось бы на всякий случай ещё раз предостеречь читателя от весьма распространённой в наше время ошибки, связанной с этим понятием: многие программисты свято уверены, что «побочным эффектом» следует называть вообще любое действие, изменяющее хоть что-нибудь в среде выполнения программы — в частности, любое присваивание, ввод или вывод информации и т. п. Мы уже отмечали это, изучая Паскаль, а ранее упоминали в предисловиях, но всё же повторим ещё раз: нет, если изменение произошло не в ходе вычисления выражения, то побочным эффектом такое изменение считать нельзя; в частности, в программах на том же Паскале получить побочный эффект можно только одним способом — написать *функцию* (не процедуру!), которая будет что-то изменять за пределами своих локальных переменных. Некоторые функции, встроенные в Паскаль или входящие в его библиотеку, сами по себе обладают побочными эффектами (примеры этого — функция `random`, возвращающая псевдослучайное число, но при этом изменяющая глобальную переменную `randseed`, чтобы следующее число, выданное функцией, было уже другим, а также функция `readkey` из модуля `crt`, побочный эффект которой состоит в изъятии очередного кода из входящего буфера), но в целом побочные эффекты в Паскале — явление довольно редкое.

Совсем другое дело — язык Си. Мало того, что процедур тут нет, есть только функции — так ещё, в довершение, все способы изменить значение переменной сделали арифметическими операциями. Присваивание стало операцией — и теперь то, ради чего мы его обычно делаем, то есть собственно занесение нового значения в переменную, оказывается *побочным эффектом*. Нет процедур — и побочными эффектами оказывается вообще всё, что делается внутри подпрограмм. Мы уже убедились на собственном опыте, что самый популярный оператор в Си — это оператор вычисления выражения ради побочного эффекта. Да что там оператор, ведь **выполнение программы, написанной на Си, *состоит* из побочных эффектов** — не «допускает побочные эффекты», не «использует побочные эффекты», а *состоит* из них, то есть программа, попросту говоря, не делает вообще ничего, кроме побочных эффектов.

Теперь можно легко догадаться, откуда берётся железобетонная уверенность, с какой большинство программистов *неправильно* трактуют побочный эффект как вообще любое изменение, какое может произвести программа: дело в том, что *в языке Си это так и есть*, при этом если взять сам язык Си и множество его концептуальных потомков, начиная с Си++, то окажется, что на этих языках написано процентов этак 99, если не больше, всех программ в мире. Немудрено, что люди попросту забывают исходный смысл слова «побочный» в словосочетании «побочный эффект» — и, надо сказать, напрасно забывают, поскольку побочные эффекты представляют собой эффективный инструмент для превращения текста программы в запутанный ребус.

Примеры таких ребусов мы уже приводили. «Хрестоматийное» копирование строки с тремя побочными эффектами в заголовке `while` при пустом теле цикла (см. стр. 34) работает правильно (при условии, что исходная строка корректна, а в массиве, куда её копируют, достаточно места), но если не знать этот пример заранее, то разобраться, как эта штука работает и понять, корректно она написана или нет, довольно непросто. А вот пример на ту же тему, который мы привели чуть раньше — на стр. 33, — напротив, содержит ошибку, и если не знать, что она там есть, заметить её вообще невозможно[35].

Чтобы запутать программу, побочным эффектам не обязательно находиться в заголовках циклов и ветвлений. Пусть, к примеру, у нас есть массив и к нему прилагается переменная, хранящая количество элементов, заполненных полезной информацией:

```
double array[ARRAY_SIZE];
int array_filled = 0;
```

Если теперь требуется занести значение (скажем, из переменной `x`) в *очередной* элемент массива, большинство программистов, пишущих на Си, поступит примерно так:

```
array[array_filled++] = x;
```

[34] Речь идёт о чистом Си; в Си++ описание переменной является оператором.

[35] Автор без малейшего смущения и даже с определённым удовольствием готов признаться, что в первом издании книги этот пример был приведён как якобы правильный, а ошибку в нём нашли читатели, причём не сразу.

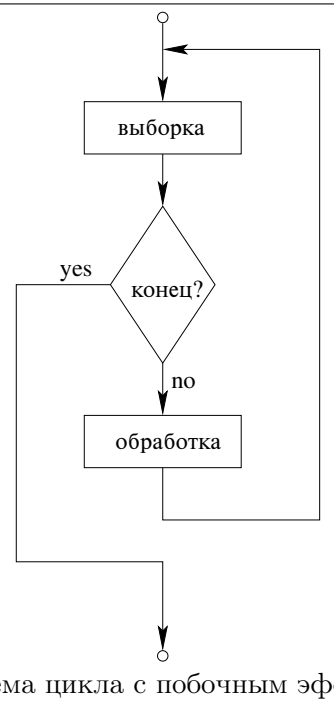

Рис. 4.4. Блок-схема цикла с побочным эффектом в условии

Чтобы понять, правильно этот оператор будет работать или нет, придётся вспомнить разницу между `i++` и `++i`, прикинуть, что при нумерации элементов с нуля их *текущее количество* равно *номеру первого свободного*, так что, по-видимому, как раз значение `array_filled`, предшествующее выполнению нашего оператора, содержит нужный номер, а поскольку постфиксная форма его и возвращает, то, наверное, тут всё правильно. При этом фрагмент

```
array[array_filled] = x;
array_filled++;
```

делает абсолютно то же самое, и даже требует примерно тех же рассуждений — только без судорожного (и дающего лишнюю возможность ошибиться) припоминания, чтó там возвращает операция инкремента в той или другой её форме. Запись этого в одну строку, а не в две, когда-то давно была осмысленной из соображений эффективности — но только на машине с соответствующей системой команд (напомним, PDP-11 при использовании косвенной адресации предусматривала преинкремент, предекремент, постинкремент или постдекремент адреса) и вдобавок при использовании компилятора, не оснащённого подсистемой оптимизации.

Здесь напрашивается вывод, что побочные эффекты — зло и использовать их нельзя, но он оказывается слишком поспешным. Во-первых, на Си вообще, формально говоря, невозможно написать программу без побочных эффектов, так что запретить их все — всё равно что запретить сам язык Си; как минимум нужны какие-то критерии, чтобы отличать допустимые побочные эффекты от недопустимых. Во-вторых, вообще-то даже такая ребусоопасная штука, как побочный эффект в условном выражении в заголовке цикла, в некоторых случаях оказывается приёмом вполне оправданным, и мы этот приём неоднократно применяли, начиная с §4.4.1, где, рассказав о функции `getchar`, мы привели простейший пример программы, читающей текст посимвольно до наступления конца файла. Вспомним наши два варианта решения:

```
c = getchar();
while(c != EOF) {
    if(c == '\n')
        printf("OK\n");
    c = getchar();
}
```

```
while((c = getchar()) != EOF) {
    if(c == '\n')
        printf("OK\n");
}
```

Можно ли здесь с уверенностью заявить, что первое решение лучше второго? Вроде бы во втором имеется побочный эффект в заголовке `while`, и это, признаемся честно, несколько затрудняет восприятие программы; но зато в первом варианте *дублируется код*. Здесь это всего лишь вызов `getchar`, в других случаях может потребоваться что-нибудь посложнее.

Сама эта ситуация в программировании довольно типична; очень часто в самых разных задачах требуется циклически что-то откуда-то извлекать, проверять, успешно ли прошло извлечение и не кончились ли те объекты, которые мы извлекаем, и если очередной объект успешно добыт, то обрабатывать его и продолжать работу, а если они кончились — завершать цикл. Важно тут то, что каждая итерация цикла состоит из двух стадий — *выборки* и *обработки*, а проверка по смыслу должна делаться *между* ними (см. рис. 4.4). Но языки программирования нам такого оператора цикла не предоставляют, вот и приходится что-то изобретать, как говорится, из подручных материалов.

Кроме рассмотренных вариантов с побочным эффектом в заголовке и с дублированием кода, известен ещё третий вариант — с «бесконечным» циклом: в заголовке пишется `for(;;)` или `while(1)`, а проверка и выход делаются в середине цикла комбинацией `if`/`break`. Но и это решение не слишком удачно: видя такой заголовок, читатель программы подсознательно ожидает, что цикл имеется в виду действительно бесконечный, а чтобы понять, какой он *на самом деле*, нужно отыскать взглядом условное выражение, которое, вместо того чтобы красоваться на видном месте в заголовке, оказывается закопано куда-то внутрь тела цикла, где его совершенно не видно.

Коль скоро альтернативные варианты решения проблемы тоже обладают определённой *кривизной*, в такой ситуации возникает внятное оправдание побочному эффекту в заголовке: без него получится ещё хуже, ну или, во всяком случае, *не факт, что будет лучше*.

Важно понимать, что сказанное касается только циклов, и только таких, которые нужно (в силу особенностей решаемой задачи) выстроить по схеме с выборкой и обработкой. В частности, **побочный эффект в заголовке оператора `if` оправданий не имеет — никогда и никаких**. То же самое можно сказать про всевозможные конструкции, использующие *значение* операций инкремента и декремента, вроде «`a[i++]=b`» — когда-то давно их можно было оправдать повышением быстродействия, но со времени появления оптимизирующих компиляторов подобный код перешёл в разряд неуместных трюков.

Обратим внимание читателя ещё на один приём, который в программах на Си используется столь часто, что воспринимается как нечто само собой разумеющееся; во всяком случае, автору книги стоило определённого труда перестать так делать в программах на Паскале, где эта техника откровенно чужеродна. Речь идёт о функциях, которые по смыслу должны были бы быть процедурами, если бы процедуры в Си вообще были — то есть о таких функциях, *основной* задачей которых является их побочный эффект. Часто такие функции объявляют со словом `void` в качестве возвращаемого значения (между прочим, в оригинальном языке Си не было слова `void`, а функции по умолчанию возвращали `int` — но это так, к слову). Но среди всех таких функций выделяются те, работа которых может пройти успешно или неуспешно — и вот их в Си обычно делают возвращающими некое значение, из которого можно заключить, насколько успешно всё прошло. Хрестоматийным примером такой функции можно считать `scanf`; но вообще-то в стандартной библиотеке Си именно так построены *все* функции, работающие с вводом-выводом. В следующих частях книги мы убедимся, что практически все системные вызовы ОС Unix — точнее, их обёртки в виде функций для Си — тоже укладываются в ту же схему: они что-то делают, и часто что-то довольно сложное, а возвращают (почти все!) некое целое число, причём `-1` означает, что произошла ошибка. Довольно много системных вызовов кроме этой «ошибочной» минус-единицы умеют возвращать лишь одно число — ноль, который означает, что вызов отработал успешно.

Если эту схему подвергнуть вдумчивому анализу, можно заметить, что сие есть очередной хак, ведь возвращаемое значение функции изначально придумано не для этого. Но этот хак столь привычен и настолько часто применяется, что даже просто заставить себя понять, что это хак, а не «так и должно быть», может оказаться несколько затруднительно. Так или иначе, бороться с этим бессмысленно, не переписывать же ради этого всю стандартную библиотеку и все обёртки системных вызовов.

С учётом всего сказанного можно попытаться сформулировать некую инструкцию, отделяющую допустимые случаи (и способы применения) побочных эффектов от недопустимых. Получится примерно так:

- функции могут обладать (формальным) побочным эффектом в том и только в том случае, если этот побочный эффект представляет собой по смыслу именно то, ради чего функция написана;
- операции присваивания, включая операции вида `+=`, `<<=` и т. п., а также операции инкремента и декремента должны всегда выполняться ради

побочного эффекта, то есть быть операцией верхнего уровня в *операторе вычисления выражения*[36]; использовать *значение* этих операций не следует;

- функции, имеющие побочный эффект, должны вызываться из оператора вычисления выражения, причём вызов функции должен быть либо операцией верхнего уровня в этом выражении, либо операцией верхнего уровня в правой части *простого* присваивания (под простым присваиванием понимается операция `=` и только она); иначе говоря, оператор, содержащий вызов такой функции, должен иметь вид «`f(`<*параметры*>`);`» или «`res = f(`<*параметры*>`);`» — и никакой другой;

- из двух предыдущих правил имеется ровно одно исключение: при построении цикла по схеме «выборка-проверка-обработка» в условное выражение цикла `while` (и только его) может быть включён побочный эффект (как правило, вызов функции, в том числе в сочетании с присваиванием), если этот побочный эффект представляет собой этап выборки.

Так, согласно этой инструкции следующие операторы допустимы, какова бы ни была функция `f`:

```
x = 17;          f(x);          f((x+5)*(z-2));
i++;             t = f(x);      t = f(100-x);
```

Допустимо также и следующее:

```
while((c = getchar()) != EOF) {
    /* ... */
}
```

— и вообще любые побочные эффекты в условном выражении в заголовке цикла, если речь идёт о построении цикла по схеме «выборка-проверка-обработка». В то же время следующие операторы запрещены:


```
a = b = c;        v[i++] = t;    t = a>b ? (a=b) : (b=a);

if((f = fopen(fname, "r"))) {       if((c = getchar()) != EOF) {
     /* ... */                           /* ... */
}                                   }

          t = f(x) + 1;       m[f(x)] += 7;
t = g(f(x));          g(f(x));              g(f(x) + 1);
```

4.8.4. Возвращаемые значения и выходные параметры

Напомним, что ***выходными*** называют такие параметры подпрограмм, которые предназначены не для получения подпрограммой информации, а, напротив, для передачи информации вызывающему. Если ваша подпрограмма должна вернуть больше одного значения, то иного выхода, кроме использования выходных параметров, собственно говоря, нет; в языке Си выходные параметры организуются путём передачи в подпрограмму адреса переменной, в которую следует записать вычисленное значение (см. §4.3.12).

Среди выходных параметров можно выделить важный частный случай — параметры, работающие «в обе стороны» — и на вход, и на выход. По-английски такие параметры обозначаются лаконичным словом *inout*, но устоявшегося русского перевода нет, словоформы вроде *входновыходной* всерьёз воспринимать невозможно. Суть параметра этого вида в том, что некая переменная до вызова функции содержит значение, которым функция должна воспользоваться, изменить его и отдать вызывающему новое значение. Очевидным решением тут будет просто отдать функции саму эту переменную, с тем чтобы функция пользовалась ею, как хотела, а перед возвратом управления оставила в переменной то самое новое значение; технически такой параметр ничем от обычного выходного не отличается, мы точно так же передаём в функцию адрес переменной.

Интересно, что если такой параметр у функции предполагается *один*, программисты часто предпочитают вообще не применять выходной параметр; вместо этого исходное значение (переменной) передаётся в функцию *обычным* параметром, а итоговое значение функция *возвращает* (как своё обычное возвращаемое значение), и вызывающий его благополучно присваивает той же переменной. Мы уже упоминали такой подход; в §4.5.2 мы рассматривали в качестве одного из примеров удаление из списка целых чисел всех звеньев, содержащих отрицательные числа (см. стр. 51). Указатель на первый элемент списка при этом может измениться (хотя может и не измениться), так что очевидное решение состоит в том, чтобы в функцию передать адрес этого указателя, т. е. параметр типа указатель на указатель. Заголовок функции получался таким:

```
void delete_negatives(struct item **pfirst)
{
```

Передавать такой функции нужно *адрес* указателя на первый элемент списка, так что, если указатель называется `first`, вызов функции будет выглядеть так:

```
delete_negatives(&first);
```

Там же мы привели и другое решение, при котором исходный указатель на первый элемент передаётся функции по значению, то есть она не знает, *где* в памяти находится сам этот указатель; новое значение для него она возвращает обычным способом, так что присваивание возлагается на вызывающего. Заголовок функции при этом становится таким:

```
struct item *delete_negatives(struct item *first)
{
```

— а её вызов таким:

```
first = delete_negatives(first);
```

У функции могут быть другие *входные* (обычные) параметры, то есть такие, которые она использует только для получения информации; их наличие, собственно говоря, ничего не меняет. Так, функцию добавления элемента в тот же список целых чисел мы можем оформить как с выходным параметром:

```
void add_to_int_list(struct item **pfirst, int val)
{     /* ... */
```

— так и без такового, возвращая новое значение адреса первого элемента:

```
struct item *add_to_int_list(struct item *first, int val)
{     /* ... */
```

В первом случае вызов будет таким:

```
add_to_int_list(&first, 27);
```

Во втором случае — таким:

```
first = add_to_int_list(first, 27);
```

Можно встретить утверждение, что второй подход *более гибок*; на операциях со списками это может не быть столь заметным, но ведь бывают и другие задачи. Пусть, к примеру, вам в какой-то момент потребовалось взять от значения некоторой переменной логарифм по заданному основанию, а результат поместить в ту же самую переменную. Припомнив, что $\log_a b \equiv \frac{\ln b}{\ln a}$ и что в стандартную библиотеку входит функция `log`, вычисляющая натуральный логарифм, с самой задачей вычисления логарифма по произвольному основанию мы справимся мгновенно:

```
double logbase(double arg, double base)
{
    return log(arg) / log(base);
}
```

[36] На всякий случай напомним, что полное название этого оператора — *оператор вычисления выражения ради побочного эффекта.*

Оформить решение как-то иначе нам, скорее всего, и в голову не придёт; между тем, согласно исходному условию требуется результат занести в ту же переменную, где до того находился аргумент. Разумеется, никто не мешает сделать так: `a = logbase(a, bs);` но если мы припомним то, с чего начинали, можно заметить возможность другого решения — вызывать функцию как-то так: `logbase(&a, bs)`, для чего саму функцию переделать следующим образом:

```
void logbase(double *arg, double base)
{
    *arg = log(*arg) / log(base);
}
```

Достаточно очевидно, что делать так не надо: вполне возможно, что нам потребуется вычислить логарифм по произвольному основанию где-нибудь ещё, и там нам не только не будет нужно заносить результат в ту же переменную, где был исходный аргумент, но и самой этой переменной может не быть — скажем, если аргументом логарифма должно служить некоторое выражение, а результат вычисления нужен в качестве аргумента для какой-то ещё функции или простой операции вроде сложения. Очевидно, что подпрограмма, оформленная фактически в виде процедуры и требующая для своей работы завести переменную, окажется в такой ситуации элементарно неудобна. Кстати, такое неудобство можно было предвидеть, поскольку здесь мы нарушили старый и очень важный архитектурный принцип: **подпрограмма должна делать что-то одно**. Если мы пишем функцию, вычисляющую логарифм, то она должна отвечать за вычисление логарифма и только за него; заносить что-то в какую-то переменную — это уже не её дело.

Но как так получается, что *одно и то же решение* нас явно не устраивает, когда речь идёт об арифметике, но почему-то приходит в голову первым и кажется правильным, когда мы имеем дело со списками? Ответ на этот вопрос требует довольно нетривиального рассуждения. Дело в том, что *список как сущность не исчерпывается значением адреса его первого элемента*, т. е., попросту говоря, тот указатель `first`, с которым мы работаем — это не сам список, это только некая информация для доступа к нему. При присваивании значения типа `double`, а равно и при передаче его по значению в функцию и при возврате из функции мы получаем полноценную копию исходного значения, с которой можно делать что угодно, не оглядываясь на факт существования исходной переменной (или откуда там взялось это значение); делать всё то же самое с указателем на первый элемент списка, конечно, можно, но *копии списка* при этом никакой не появится и вообще нужно хорошо понимать, что́ конкретно мы делаем — уж во всяком случае работать с копией `first` так же, как с оригиналом, не оглядываясь на существование оригинала, точно не получится, ведь список-то остаётся один.

Проговорив всё это, мы, можно надеяться, подобрались к сути принципиальной разницы между рассмотренными только что `add_to_int_list` и `logbase`. В силу особенностей решаемой задачи `logbase` представляет собой *функцию* не только в «программистском», но и в *математическом* смысле: она берёт некие значения и из них получает результат; заметим, её первая реализация *не имела побочных эффектов*, что лишний раз подтверждает её «функциональную» сущность. Иное дело `add_to_int_list`. Она вносит в список новый элемент, но она при этом *не создаёт никакого нового списка* — вместо этого она модифицирует список существующий, так что рассматривать её как функцию в математическом смысле не получится; чтобы такое рассмотрение было корректно, пришлось бы при каждом добавлении элемента строить *буквально* новый список — копию предыдущего, но с одним новым элементом. Но копировать весь список при каждом добавлении элемента мы, разумеется, не собираемся; ну а функция, модифицирующая существующий список — это по сути вообще не функция, а процедура, и на Паскале нам бы и в голову не пришло описывать её как функцию, но Си нам никакой другой возможности не даёт. Так или иначе, второй вариант `add_to_int_list` — тот, который *возвращает* новое значение адреса первого элемента списка — при внимательном рассмотрении оказывается не только ничем не лучше первого, но даже ощутимо хуже: мало того, что он «притворяется функцией», хотя по сути ею не является, так его реализация ещё и построена *исходя из предположения о том, что вызывающий сделает присваивание той же переменной*; между тем, чем меньше будет таких предположений, тем лучше.

Отметим, забегая вперёд, что интерфейс функции `realloc` из стандартной библиотеки Си построен именно так, как мы только что рекомендовали не делать. Впрочем, странно было бы ожидать, что создатели Си и его библиотеки сами как-то избежали сишности головного мозга.

Ситуация резко усложняется, если подпрограмма (в терминах Си — «функция», хотя это уже точно никакая не функция) должна вернуть *больше одного значения*. При этом, разумеется, возникает соблазн задействовать возвращаемое значение функции — хотя бы даже из соображений эффективности, ведь возвращаемое значение передаётся через регистр, тогда как выходной параметр — это всегда область памяти, да ещё её адрес, ко всему, сам по себе должен быть как-то передан в функцию. Вот только здесь возникает одна проблема, на которую «истинные сишники» привыкли не обращать внимания: *а которое* из нескольких значений вернуть как результат функции? Как его выбрать — одно из нескольких?

Что ж, здесь мы рискнём дать определённую рекомендацию, с которой, опять-таки, не все согласны. Если одно из значений, которые ваша подпрограмма должна вернуть, можно (по смыслу) рассматривать как *признак успешности* выполнения подпрограммы (например, одно из его значений, либо какое-то подмножество возможных значений свидетельствует о возникшей ошибке) — то из функции в качестве её «основного» результата следует вернуть именно это значение, а все остальные вернуть через выходные параметры; если же ни одно из возвращаемых значений не играет роли индикатора успешности (например, если функция всегда завершается успешно) — наплюйте на эффективность и сделайте функцию `void`'овой, вернув всё, что нужно, через параметры. Искусственное превращение процедур в функции не добавляет ясности.

Читатель легко найдёт в стандартной библиотеке Си примеры, противоречащие и этой рекомендации тоже — например, функции `strtol`, `strtoll` и `strtof`. Ваш покорный слуга прекрасно знает об этих функциях и сам их часто использует, но их интерфейс находит спроектированным отвратительно.

Отметим ещё один крайне важный момент. **Если параметр имеет адресный тип, но не является *выходным*, обязательно поставьте модификатор** `const`. Наличие `const` покажет читателю программы, что через этот параметр вы не собирались никакой информации отдавать наружу, и тем самым изрядно облегчит понимание вашего текста. Напротив, если некий параметр функции, имеющий адресный тип (указатель) написан без `const`, большинство программистов попытается выяснить, что́ и зачем ваша функция записывает в эту область памяти; убедившись, что она туда в действительности ничего не пишет, читатель вашей программы, скорее всего, помянет вас *словом*, и добрым это слово не будет.

4.8.5. (*) Программы, говорящие по-русски

Коль скоро в тексте программы нельзя употреблять языки, отличные от английского, возникает резонный вопрос, как быть, если программа в соответствии с поставленной задачей должна общаться с пользователем по-русски (или по-немецки, или по-китайски — это не важно).

Изоляционизм в программировании, к счастью, ушёл в далекое прошлое, и в наши дни над одной программой могут работать программисты из нескольких десятков разных стран, а пользоваться одной и той же программой могут пользователи всего мира — во всяком случае, всех частей мира, где есть компьютеры. В такой обстановке постоянно возникает ситуация, когда программу необходимо «научить» общаться с конечным пользователем на языке, которого не знает никто из её авторов, причём эта ситуация в наше время представляет собой скорее правило, нежели исключение. Адаптация программы к использованию другого языка оказывается достаточно простой, если её автор следовал определённым соглашениям; общая идея тут такова, что исходные строки, которые должен увидеть (или ввести) пользователь, прямо во время работы программы заменяются строками, написанными на другом языке, которые загружаются из внешнего файла (то есть *не являются частью программы*).

Обычно для этой цели используются готовые библиотеки, специально предназначенные для «интернационализации» сообщений в программах. В ОС Unix наиболее популярна библиотека `gettext`. При работе на языке Си, чтобы сделать возможной загрузку строк во время исполнения, все строки в программе традиционно обрамляются знаком подчёркивания и скобками — например, вместо

```
printf("Hello, world\n");
```

пишут

```
printf(_("Hello, world\n"));
```

Макрос с именем «`_`» разворачивается в вызов функции `gettext`, которая пытается найти файл с переводом сообщений на используемый язык интерфейса, а в нём, в

свою очередь — перевод для строки "Hello, world\n", причём сама эта строка используется как ключ для поиска. Если не удалось найти такой перевод (или сам файл с переводами), в качестве результата возвращается аргумент, то есть если gettext не знает, как перевести строку "Hello, world\n" на нужный язык, она оставит эту строку нетронутой.

Если вы по каким-то причинам не хотите использовать gettext, достаточно переопределить макрос _, чтобы он всегда возвращал свой аргумент; переделывать программу при этом не придётся.

Больше того, если даже вы не сочли нужным подготовить свою программу к использованию с библиотекой gettext, это может сделать другой программист, просто заключив все строковые константы в макровызов _(), что достаточно просто — при известной ловкости это можно сделать одной командой в текстовом редакторе. Есть, однако, одно крайне важное условие, выполнение которого необходимо для превращения моноязычной программы в «международную»: все сообщения в её тексте должны быть английскими. Двойной перевод системами интернационализации не предусмотрен, ну а переводчиков с *русского* на другие языки, будь то китайский или немецкий, найти гораздо сложнее, чем переводчиков с *английского* — особенно если учесть, что речь идёт не о профессиональных переводчиках, а, как правило, о программистах, которым пришло в голову перевести сообщения очередной программы на свой родной язык; английский знают программисты во всём мире, но вот найти нерусского программиста, при этом знающего русский — практически нереально.

Итак, программа, даже не адаптированная исходно под нужды многоязычности, вполне имеет шансы стать когда-нибудь «международной» — но только если исходно все сообщения в её тексте английские. Из этого очевидным образом следует сформулированное выше утверждение: ваша программа должна быть либо «международной», либо англоязычной. Если вы не чувствуете себя в силах или не имеете желания учитывать возможный перевод интерфейса программы на другие языки, по крайней мере не лишайте *других* программистов возможности сделать это за вас. Если же русскоязычность входит в постановку задачи, не поленитесь сделать всё *правильно* — то есть в соответствии с требованиями «международности».

На случай, если вам потребуется срочно сделать русскоговорящую программу, мы приведём простой пример работы с gettext в надежде, что вы не станете ради сиюминутной потребности нарушать фундаментальные требования к оформлению исходных текстов (а наличие русских букв в тексте программы — это именно фундаментальное нарушение).

Начнём с простого текста программы, печатающей два сообщения:

```
#include <stdio.h>

int main()
{
    printf("Hello, world!\n");
    printf("Good bye, world!\n");
    return 0;
}
```

Эта программа, понятное дело, никоим образом не может считаться *интернационализированной*, но у неё есть все шансы такой стать, поскольку выполнено главное условие: все сообщения в программе написаны по-английски. Первый шаг в направлении интернационализации состоит в обрамлении всех сообщений вызовом макроса _(); сам макрос мы введём в начале программы самым простым из возможных способов:

```
#include <stdio.h>

#define _(STR) (STR)

int main()
{
    printf(_("Hello, world!\n"));
    printf(_("Good bye, world!\n"));
    return 0;
}
```

Как можно догадаться, эта программа делает то же самое, что и предыдущая, то есть по сути программа не изменилась; просто она стала «чуть более подготовленной» к превращению в «международную». Здесь есть ещё один момент: иногда в программах встречаются строки, которые либо вообще не нужно переводить, либо они будут переводиться не в том месте, где встречены. Например, если строка расположена в инициализаторе какого-нибудь массива строк, вызывать функцию gettext из такого инициализатора, пожалуй, всё же не стоит; возможны и другие ситуации. Строки, не подлежащие переводу, можно просто оставить как есть, но у читателя программы это может вызвать подозрение, что про данную конкретную строку просто забыли. Поэтому рекомендуется ввести ещё один макрос:

```
#define N_(STR) (STR)
```

и в программе *все* строки оформить с вызовом одного из макросов: те, строки, что должны будут переводиться, заключить в макровызов _(), а те, что переводить не нужно — в макровызов N_(). Это можно сделать сразу же, даже если вы пока не собираетесь превращать вашу программу в международную. Когда дело дойдёт до непосредственного задействования gettext, вы сэкономите время, поскольку все строковые литералы в программе уже будут подготовлены к переводу.

Когда вам потребуется превратить программу в международную, придётся сделать несколько модификаций её исходного текста. Во-первых, нужно будет добавить включение заголовочных файлов locale.h ради функции setlocale (к сожалению, библиотека gettext намертво завязана на использование механизма так называемых системных *локалей* и без настройки локали работать не будет) и libintl.h, в котором объявлены функция gettext и вспомогательные bindtextdomain и textdomain, которые нам тоже потребуются.

Библиотека gettext требует первичной настройки, вся цель которой, собственно говоря, в том, чтобы сообщить, из какого файла брать переводы строк. Как это, к сожалению, часто водится в проектах группы GNU, да и вообще во многих современных проектах, просто указать имя файла библиотека не позволяет. Вместо этого требуется задать некую *базовую директорию*, где хранятся данные, имеющие отношение к локалям, и совсем непонятный *текстовый домен*, который при ближайшем рассмотрении оказывается просто именем нашей программы (или пакета программ, к которому она относится — но это не наш случай).

В качестве имени программы (того самого «текстового домена») мы воспользуемся словом «hellobye», поскольку именно так называется наша программа. С базовой директорией всё несколько сложнее. Если бы мы собирались инсталлировать нашу программу в системные директории, в этой роли выступала бы директория /usr/share/locale или /usr/local/share/locale (в зависимости от того, каким способом производится установка). Поскольку инсталляция программы в системные директории вряд ли входит в наши планы, мы поступим проще: укажем в качестве «базовой» текущую директорию (попросту строку "."). Это позволит запустить наш простенький пример, но для более сложных задач может оказаться непригодно, поскольку программы чаще всего запускаются не из той директории, в которой находятся принадлежащие им файлы; в настоящей «боевой» программе, возможно, имя базовой директории локализационных данных придётся получить через параметры командной строки, прочитать из конфигурационного файла или каким-то образом синтезировать из имени домашней директории пользователя. Так или иначе, нам для демонстрации работы gettext вполне подойдёт текущая директория.

Имя базовой директории и текстового домена вынесем в макроконстанты в начале программы:

```
#define LOCALEBASEDIR "."
#define TEXTDOMAIN "hellobye"
```

Туда же добавим новые определения для макросов _() и N_() (в нашем примере второй из них не будет задействован, но лучше сразу добавлять оба, чтобы, когда нам встретится строка, не подлежащая переводу, у нас не возникало соблазна оставить её как есть):

```
#define _(STR) gettext(STR)
#define N_(STR) (STR)
```

В начало функции main вставим настройку локали и самой библиотеки gettext:

```
    setlocale(LC_CTYPE, "");
    setlocale(LC_MESSAGES, "");
    bindtextdomain(TEXTDOMAIN, LOCALEBASEDIR);
    textdomain(TEXTDOMAIN);
```

Окончательный текст программы примет следующий вид:

```
#include <stdio.h>                    /* gettext/hellobye.c */
#include <locale.h>
#include <libintl.h>

#define LOCALEBASEDIR "."
#define TEXTDOMAIN "hellobye"

#define _(STR) gettext(STR)
#define N_(STR) (STR)

int main()
{
    setlocale(LC_CTYPE, "");
    setlocale(LC_MESSAGES, "");
    bindtextdomain(TEXTDOMAIN, LOCALEBASEDIR);
    textdomain(TEXTDOMAIN);
    printf(_("Hello, world!\n"));
    printf(_("Good bye, world!\n"));
    return 0;
}
```

Подготовив исходник, приступим к созданию его «переводов», то есть файлов, задающих для английских сообщений их переводы на другие языки. Для начала нам потребуется запустить программу xgettext, которая автоматически извлечёт из нашего текста программы все строки, заключённые в макровызов _(), и на их основе создаст шаблонный файл, который можно будет отредактировать, создав файл перевода. Запускаем команду:

```
xgettext --keyword="_" hellobye.c -o hellobye.pot
```

Здесь параметр --keyword задаёт имя того вызова (в нашем случае — макровызова), которым помечены в тексте программы строки для перевода, ключ -o указывает имя файла для записи результата. Программа xgettext создаст в рабочей директории файл с именем hellobye.pot (суффикс .pot означает *portable object template*, то есть «шаблон переносимого объекта») примерно с таким содержимым:

```
# SOME DESCRIPTIVE TITLE.
# Copyright (C) YEAR THE PACKAGE'S COPYRIGHT HOLDER
# This file is distributed under the same license as the PACKAGE package.
# FIRST AUTHOR <EMAIL@ADDRESS>, YEAR.
#
#, fuzzy
msgid ""
msgstr ""
"Project-Id-Version: PACKAGE VERSION\n"
"Report-Msgid-Bugs-To: \n"
"POT-Creation-Date: 2016-03-21 14:19+0300\n"
"PO-Revision-Date: YEAR-MO-DA HO:MI+ZONE\n"
"Last-Translator: FULL NAME <EMAIL@ADDRESS>\n"
"Language-Team: LANGUAGE <LL@li.org>\n"
"MIME-Version: 1.0\n"
"Content-Type: text/plain; charset=CHARSET\n"
"Content-Transfer-Encoding: 8bit\n"

#: hellobye.c:17
#, c-format
msgid "Hello, world!\n"
msgstr ""

#: hellobye.c:18
#, c-format
msgid "Good bye, world!\n"
msgstr ""
```

Этот шаблон отражает то, как, по убеждениям авторов gettext, должен выглядеть файл перевода. Скопируем этот файл под именем ru.po (ru здесь означает русский язык, po — *portable object*) и отредактируем. Прежде всего модифицируем комментарий в начале файла, указав, что́ это за файл и какие на него накладываются лицензионные ограничения (в нашем примере таких нет, что мы и напишем). Затем в строке Project-Id-Version укажем имя и версию нашей программы. В строке Report-Msgid-Bugs-To следует указать, к кому обращаться, если обнаружена ошибка в оригинальных сообщениях (не в переводе). Строку PO-Revision-Date следует оставить как есть, с ней потом разберётся программа, создающая на основе нашего файла перевода индексированный файл, в котором поиск сообщений происходит быстрее. В строке Last-Translator надо указать своё имя или псевдоним, а вот адрес электронной почты лучше не указывать, сколь бы мы ни уважали мнение группы GNU на эту тему: вряд ли вы будете рады увеличению количества поступающего вам спама. Строку Language-Team мы просто выкинем, поскольку не входим ни в какую команду по переводу; в строке Content-Type заменим слово CHARSET на используемую кодировку (в нашем примере это будет KOI8-R, но в вашей системе, скорее всего, используется UTF-8). Строки MIME-Version и Content-Transfer-Encoding оставим как есть.

Наконец, сделаем самое главное: вставим русский перевод для обеих имеющихся текстовых строк. Результат будет выглядеть примерно так:

```
# gettext/ru.po
# Russian translation for the HelloBye program
# No copyright is claimed!
# Andrey V. Stolyarov, 2016.
#
#, fuzzy
msgid ""
msgstr ""
"Project-Id-Version: HelloBye 1.0\n"
"POT-Creation-Date: 2016-03-21 14:19+0300\n"
"PO-Revision-Date: YEAR-MO-DA HO:MI+ZONE\n"
"Last-Translator: Andrey V. Stolyarov\n"
"MIME-Version: 1.0\n"
"Content-Type: text/plain; charset=KOI8-R\n"
"Content-Transfer-Encoding: 8bit\n"

#: hellobye.c:17
#, c-format
msgid "Hello, world!\n"
msgstr "Здравствуй, мир!\n"

#: hellobye.c:18
#, c-format
msgid "Good bye, world!\n"
msgstr "До свиданья, мир!\n"
```

Теперь этот файл следует откомпилировать, получив собственно тот (бинарный) файл, который gettext будет использовать в работе. От своего исходника результат такой компиляции отличается возможностью быстрого поиска нужного сообщения. Компиляцию произведёт программа msgfmt:

```
msgfmt ru.po -o ru.mo
```

Результатом станет файл ru.mo (суффикс здесь означает *machine object*). Для наглядности снабдим нашу программу переводом ещё и на немецкий, для чего создадим копию ru.po под именем de.po и отредактируем, заменив кодировку на ASCII, а русские фразы — на соответствующие немецкие:

```
# German translation for the HelloBye program
# No copyright is claimed!
# Andrey V. Stolyarov, 2016.
#
#, fuzzy
msgid ""
msgstr ""
"Project-Id-Version: HelloBye 1.0\n"
"POT-Creation-Date: 2016-03-21 14:19+0300\n"
"PO-Revision-Date: YEAR-MO-DA HO:MI+ZONE\n"
"Last-Translator: Andrey V. Stolyarov\n"
"MIME-Version: 1.0\n"
"Content-Type: text/plain; charset=ASCII\n"
"Content-Transfer-Encoding: 8bit\n"

#: hellobye.c:17
#, c-format
msgid "Hello, world!\n"
msgstr "Hallo Welt!\n"

#: hellobye.c:18
#, c-format
msgid "Good bye, world!\n"
msgstr "Auf Wiedersehen Welt!\n"
```

Запустив msgfmt, получим файл de.mo. Теперь эти два файла с суффиксом .mo нужно разместить так, чтобы привередливая gettext их нашла во время работы нашей программы:

```
mkdir -p ru/LC_MESSAGES
cp ru.mo ru/LC_MESSAGES/hellobye.mo
mkdir -p de/LC_MESSAGES
cp de.mo de/LC_MESSAGES/hellobye.mo
```

Поясним, что полное имя файла складывается из четырёх компонентов. Первый — обсуждавшаяся выше *базовая директория*, но в этой роли мы задали текущую директорию, что несколько облегчает дело. Второй компонент — идентификатор языка, в нашем случае ru для русского и de для немецкого. Третий компонент задаёт *подсистему локали*, в нашем случае это LC_MESSAGES; последний — это собственно имя файла, название которого должно совпадать с *текстовым доменом* (мы использовали hellobye), снабжённым суффиксом .mo, отражающим формат файла. Судя по всему, мы готовы. Компилируем программу как обычно (в системах, отличных от Linux, может потребоваться явное подключение библиотеки intl, то есть флажок -lintl, но в Linux это не нужно):

```
gcc -Wall -g hellobye.c -o hellobye
```

И пробуем запустить. Если в переменных окружения задана русская локализация, программа выдаст нам сообщения, заданные русским переводом:

```
avst@host:~/gettext$ ./hellobye
Здравствуй, мир!
До свиданья, мир!
avst@host:~/gettext$
```

Язык, на котором программа разговаривает, можно оперативно менять, задавая соответствующие значения для переменной окружения LANGUAGE:

```
avst@host:~/gettext$ LANGUAGE=en ./hellobye
Hello, world!
Good bye, world!
avst@host:~/gettext$ LANGUAGE=de ./hellobye
Hallo Welt
Auf Wiedersehen Welt!
avst@host:~/gettext$ LANGUAGE=ru ./hellobye
Здравствуй, мир!
До свиданья, мир!
avst@host:~/gettext$
```

Если соответствующий язык не найден или локаль настроена неправильно, скорее всего, программа выдаст английские сообщения.

Вскоре у вас наверняка возникнет вопрос, что делать, если в программе появились новые сообщения — не переводить же всё с нуля. Для таких случаев создатели gettext предусмотрели программу msgmerge, которая позволяет построить новый .po-файл по имеющемуся (устаревшему) .po-файлу и обновлённому .pot-файлу. Попробуем для примера вставить в нашу программу ещё один вызов printf между двумя уже имеющимися:

```
printf(_("We are speaking\n"));
```

Теперь построим обновлённый файл hellobye.pot с помощью xgettext точно так же, как мы делали это в прошлый раз, после чего обновим имеющиеся у нас файлы перевода с помощью msgmerge. Эту программу можно запускать в разных режимах; один из удобных вариантов выглядит так:

```
msgmerge -U --backup=numbered ru.po hellobye.pot
msgmerge -U --backup=numbered de.po hellobye.pot
```

При таких параметрах msgmerge запишет результаты сразу в файлы ru.po и de.po, а старые версии на всякий случай сохранит под именами ru.po.~1~ и de.po.~1~. Новые версии будут отличаться от старых датой/временем в строке POT-Creation-Date, номерами строк в комментариях к ранее переведённым сообщениям (сами переводы никуда не денутся) и наличием новых (пока ещё не переведённых) сообщений. Остаётся вписать туда переводы новых строк (в нашем примере — строки "We are speaking"), запустить для каждого из файлов программу msgfmt и скопировать полученные файлы с суффиксом .mo куда следует — точно так же, как мы делали это раньше.

Как обычно в таких случаях, приходится признать, что мы рассмотрели только самые тривиальные возможности библиотеки gettext; за более подробной информацией следует обратиться к документации, размещённой на web-странице http://www.gnu.org/software/gettext/ — официальной странице проекта gettext. Сразу же хотелось бы предостеречь читателя от одной ошибки. В Интернете можно найти достаточно много текстов, в которых рассказывается, как работать с gettext через пакеты autoconf/autotools, причём в некоторых случаях авторы таких текстов безапелляционно утверждают, что последние для работы с gettext *необходимы*. К счастью, это не так; в приведённых выше примерах мы никак не задействовали autotools, не требует их задействования и текст официальной документации. Что касается самих этих пакетов, то они относятся к числу таких инструментов, которые не следует применять ни для чего, никогда и ни за что — ну разве что за очень большие деньги. Никаких проблем они в действительности не решают, что бы вам об этом ни говорили, зато создают проблемы на ровном месте, что называется, пачками.

Отметим один довольно важный идейный момент. **Библиотека gettext — это далеко не единственное средство создания «международных» программ.** Нет ничего плохого даже в том, чтобы сделать программу многоязычной «вручную», вообще без использования готовых библиотек. Серьёзное ограничение только одно: все сообщения на языках, отличных от английского, должны находиться вне исходных текстов вашей программы. Чаще всего их размещают в отдельных файлах, которые читаются уже во время работы программы; есть и другой вариант — с помощью какого-нибудь простенького скрипта превратить файлы со строками на неанглийских языках в (сгенерированный) файл на языке Си, содержащий описания константных массивов с инициализаторами, причём в этих инициализаторах использовать, естественно, не сами символы, из которых состоят строки, а их числовые коды. Такие сгенерированные файлы, хоть и содержат фрагмент программы на Си, не являются, строго говоря, *исходными текстами*, ведь их не писал программист; впрочем, и неанглийских строк такие файлы не содержат. Дальнейшее становится делом техники.

4.9. «Заковыристые» типы указателей

В этой главе мы введём такие типы указателей, «благодаря» которым обыкновенный заголовок функции или описание переменной может превратиться в нелепое и совершенно нечитаемое нагромождение скобочек и звёздочек. Прежде чем это будет сделано, отметим, что директива typedef (см. §4.5.6) всегда позволяет обойтись без таких кошмарных конструкций, так что её использование настоятельно рекомендуется во всех случаях, когда описание кажется вам сложным.

4.9.1. Многомерные массивы и указатели на массивы

С многомерными массивами мы до сих пор не сталкивались, но в Си они поддерживаются — во всяком случае, в том же смысле, в каком поддерживаются массивы одномерные. Например, матрицу целых чисел из трёх строк по четыре столбца в каждой можно ввести так:

```
int m[3][4];
```

Это описание можно прочитать следующим образом: *m есть массив из трёх массивов из четырёх элементов типа int*. В памяти массив располагается построчно; в данном случае сначала идут четыре элемента первой строки (с m[0][0] по m[0][3]), затем четыре элемента второй строки (m[1][0] ... m[1][3]), последними — четыре элемента третьей строки (см. рис. 4.5).

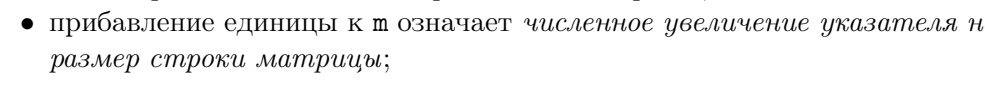

Рис. 4.5. Расположение в памяти массива int m[3][4]

Обращение к элементам многомерного массива производится через несколько (по количеству измерений, в данном случае два) применений операции индексирования, например m[2][1]. **Каждый индекс записывается в отдельных квадратных скобках**. Перечислять индексы через запятую, как в Паскале, нельзя; оно и понятно, ведь квадратные скобки представляют собой, как мы видели, обыкновенную арифметическую операцию от двух аргументов, одним из которых должен быть адрес. Например, следующий фрагмент заполнит наш массив значениями, полученными как произведение индексов (в частности, m[0][2] получит значение 0, m[2][3] — значение 6):

```
for(i = 0; i < 3; i++)
    for(j = 0; j < 4; j++)
        m[i][j] = i * j;
```

Описание многомерного массива можно снабдить инициализатором, например:

```
int m[3][4] = {{0, 0, 0, 0}, {0, 1, 2, 3}, {0, 2, 4, 6}};
```

Как оформлять инициализаторы подобного рода, мы уже рассказывали (см. стр. 67).

Теперь мы плавно подбираемся к самому интересному аспекту многомерных массивов. Если ваши мозги категорически откажутся воспринять остаток параграфа, не паникуйте: всего этого порой не знают некоторые профессиональные программисты, пишущие на Си за деньги. Впрочем, даже если с ходу вы ничего не поймёте, постарайтесь в будущем вернуться к этому параграфу и всё-таки постичь всю эту механику, которую можно считать одним из самых любопытных изобретений в области семантики языков программирования.

Итак, вернёмся к нашему массиву int m[3][4] и зададим себе незатейливый, на первый взгляд, вопрос: а что собой представляет имя m?

Напомним, что по правилам языка Си во всех случаях, кроме применения sizeof и ещё двух экзотических ситуаций, про которые мы даже не стали говорить, имя массива есть адрес его начала. Кроме того, заметим, что выражение m[1] должно быть таково, чтобы к нему можно было применять вторую операцию индексирования, результатом которой становились бы элементы второй строки матрицы. Следовательно, выражение m[1] должно удовлетворять двум свойствам: это, во-первых, должен быть адрес начала второй строки, и, во-вторых, применение индексирования к этому выражению должно давать, собственно говоря, сами элементы второй строки. Иначе говоря, m[1] должно представлять собой адрес типа int*, указывающий на начало второй строки матрицы.

Вспомним теперь, что m[1] есть не что иное, как сокращённая запись выражения *(m+1). Получается, что:

- имя m представляет собой адрес начала матрицы;
- прибавление единицы к m означает *численное увеличение указателя на размер строки матрицы*;

- разыменование выражений `m`, `m+1`, `m+2` даёт значение типа `int*`.

Ни один из ранее рассмотренных нами типов не обладает всеми этими свойствами одновременно. В самом деле, единственный известный нам тип, такой, что для переменной `p` такого типа выражение `*p` будет иметь тип `int*` — это, как несложно догадаться, тип `int**`, но он нам здесь явно не подходит: прибавление к нему единицы сдвинет нас по адресному пространству на размер указателя (4 или 8 байт), а не на размер строки матрицы (в нашем примере 16 байт, но может быть сколько угодно).

Создатели языка Си решили этот вопрос, введя особенный вид типов — ***указатель на массив*** (или, строго говоря, *адрес массива*, ведь указатель — это переменная, хранящая адрес, а не сам адрес), причём это совершенно не то же самое, что хорошо знакомый нам *указатель на первый элемент массива*. Особенную пикантность этому придаёт то обстоятельство, что **типа «массив» в языке Си нет**, во всяком случае, полноценного — а вот указатели на него, как видим, есть. Для нашего примера — массива `int m[3][4]` — нужным типом будет «адрес массива из четырёх элементов типа `int`» в соответствии с параметрами строки массива. Переменная этого типа описывается так:

```
int (*p)[4];
```

(читается как «`p` есть указатель на массив из четырёх элементов типа `int`»). Теперь можно, например, сделать присваивание `p = m` и работать с этим `p` точно так же, как мы работали с `m`, в том числе применяя двойное индексирование. Такое описание ни в коем случае не следует путать с похожим на него

```
int *q[4];
```

Здесь `q` — это обыкновенный массив из четырёх указателей на `int`, то есть массив из четырёх элементов типа `int*`; с указателями на массивы это не имеет ничего общего.

Аналогичным образом для трёхмерного массива

```
int z[10][15][20];
```

нам потребовался бы *указатель на двумерный массив*:

```
int (*zptr)[15][20];
```

Прибавление единицы к адресу такого типа перемещает нас в адресном пространстве сразу на $15 \times 20 \times 4 = 1200$ ячеек — ровно на размер «среза» массива, представляющего собой матрицу 15x20 элементов типа `int`.

4.9.2. Указатели на функции

Откомпилированные подпрограммы, в том числе функции языка Си — это, в сущности, фрагменты машинного кода; ясно, что этот код хранится где-то в оперативной памяти, то есть *занимает область памяти* и, как следствие, можно говорить об ***адресе функции***, который равен адресу первой из ячеек, занятых этим кодом. Адрес, понятное дело, можно где-то хранить; так появляется ***указатель на функцию***. Язык Си позволяет вызывать функцию, используя её адрес.

Для примера рассмотрим несколько функций, принимающих на вход массив элементов типа `double` (в виде адреса первого элемента и целого числа, обозначающего длину массива) и возвращающих число типа `double`. Самая простая из них будет находить сумму элементов массива:

```
double dbl_sum(const double *a, int size)
{
    return size > 0 ? *a + dbl_sum(a+1, size-1) : 0;
}
```

Вторая функция будет отыскивать минимальное число в массиве:

```
double dbl_min(const double *a, int size)
{
    double d;
    if(size == 1)
        return *a;
    d = dbl_min(a+1, size-1);
    return *a < d ? *a : d;
}
```

Третью мы заставим посчитать среднее арифметическое:

```
double dbl_average(const double *a, int size)
{
    return dbl_sum(a, size) / (double) size;
}
```

Можно придумать и ещё подобных функций; заметим, что их заголовки будут различаться только именем, тогда как количество и типы параметров, а также тип возвращаемого значения у них одинаковый. Говорят, что эти функции *имеют одинаковый профиль*. Для нашего рассуждения это важно, поскольку указатели на функции в Си являются типизированными в том смысле, что при их описании требуется указывать типы параметров функции и тип её возвращаемого значения, то есть всю информацию, входящую в упомянутый *профиль*. В частности, указатель, способный хранить адрес любой из трёх перечисленных функций, описывается так:

```
double (*fptr)(const double *, int);
```

Как и в случае с указателями на массивы, здесь очень важны круглые скобки вокруг `*fptr`; если бы их не было, мы бы получили что-то вроде

```
double *fn(const double *, int);
```

— но ведь это вовсе не описание переменной; это *объявление (заголовок) функции*, принимающей на вход два параметра и возвращающей адрес типа `double*`.

Вернёмся к указателю `fptr`. Ему можно присвоить адрес любой из наших функций. Сделать это можно как с указанием операции взятия адреса, так и без; оба следующих присваивания правомерны:

```
fptr = &dbl_min;
fptr = dbl_min;
```

Вызов функции через указатель (или другое адресное выражение) можно, опять-таки, произвести непосредственно либо после применения операции разыменования. К примеру, если у нас есть массив и переменная:

```
double arr[100];
double res;
```

— то оба следующих вызова компилятор сочтёт правильными:

```
res = (*fptr)(arr, sizeof(arr)/sizeof(*arr));
res = fptr(arr, sizeof(arr)/sizeof(*arr));
```

Больше того, если `f` — имя функции или выражение типа «адрес функции», то компилятор примет и `**f`, и `*******f`, и `&f`, и `& & & &f` — правда, тут всё же потребуются пробелы, чтобы амперсанды не «сливались» в символ логического «и» (`&&`). Никакой разницы между перечисленными выражениями компилятор не усматривает.

Глядя на эти примеры, можно заключить, что создатели Си так и не решили, существует ли в семантике этого языка *функция как таковая*; иначе говоря, представляет ли имя функции сразу её адрес или же имя представляет *саму функцию*, а её адрес можно получить соответствующей операцией. На самом деле в оригинальном языке Си, предложенном его создателями, функция рассматривалась как самостоятельная сущность, а вызов функции через указатель (или другое адресное выражение) всегда делался с применением операции разыменования («`*`»). При этом, правда, взятие адреса функции (например, для присваивания указателю) можно было сделать без операции «`&`» — по аналогии с именами массивов.

При внимательном рассмотрении оказалось, что над «функцией как самостоятельной сущностью» нет никаких других операций, кроме её вызова, обозначаемого круглыми скобками, и взятия адреса, а с адресом функции нельзя сделать ничего, кроме его разыменования с последующим применением операции вызова функции (ну а что ещё можно сделать с результатом разыменования адреса функции, то есть с самой функцией? взятие адреса здесь смысла уже не имеет, мы ведь с адреса начали). Авторы некоторых компиляторов решили, что операция разыменования для адресов функций избыточна, и сделали её использование необязательным, как и использование операции взятия адреса; такая практика была окончательно легитимизирована в стандарте ANSI, который, судя по весьма сдержанным, но от этого даже более красноречивым комментариям в книге Кернигана и Ритчи, создателям языка не понравился.

Остаётся естественный вопрос: так как же следует писать? Однозначного ответа нет, можно найти достаточно много программистов, делающих и так, и эдак. Программисты «старой закваски», пишущие на чистом Си, часто предпочитают при вызове функции через указатель использовать операцию разыменования этого указателя, чтобы вызов через указатель зрительно отличался от вызова функции по имени. В мире Си++ от этой традиции давно уже ничего не осталось, ведь там специально сделано всё возможное, чтобы введённый пользователем объект можно было использовать «как массив», «как функцию» и т. д.

Чтобы понять, зачем могут потребоваться указатели на функции, вернёмся к примеру, который мы привели на стр. 51: там мы рассматривали функцию, которая удаляет из заданного списка целых чисел все элементы, хранящие отрицательное число. Функция получилась, как можно заметить, нетривиальная, хотя и не очень сложная. Представьте себе теперь, что вам потребовалась функция, удаляющая из такого же списка не отрицательные числа, а, скажем, чётные, или делящиеся без остатка на семь, или ещё какие-нибудь; что же, скопировать функцию под другим именем и, изменив в ней всего одно условие, так и оставить эту копию существовать? А потом, возможно, ещё одну, и ещё, и ещё? Но мы уже знаем, что копирование фрагментов кода — это крайне порочная практика, во многих случаях даже *запрещённая*.

Адреса функций позволяют нам выйти из положения, передав одним из параметров *критерий удаления из списка*. Такой критерий оформляется в виде функции, принимающей на вход число, хранящееся в списке (в данном случае это `int`) и возвращающей «истину», если число следует удалить, и «ложь», если его следует оставить. Например, можно в качестве критериев написать следующие функции:

```
int is_negative(int x) { return x < 0; }
int is_even(int x) { return x % 2 == 0; }
int is_div7(int x) { return x % 7 == 0; }
```

Теперь мы можем переписать функцию `delete_negatives`, текст которой был приведён на стр. 51, приспособив её для удаления элементов, отвечающих *произвольному* критерию; для этого добавим в её заголовок второй параметр — указатель на функцию, задающую критерий, то есть, попросту, на функцию, которая принимает число типа `int` на вход и возвращает логическое значение (то есть тоже `int`); вместо сравнения с нулём в операторе `if` подставим вызов этой функции от числа из текущего элемента списка:

```
void delete_from_int_list(struct item **pcur, int (*crit)(int))
{
    while(*pcur) {
        if((*crit)((*pcur)->data)) {
            struct item *tmp = *pcur;
            *pcur = (*pcur)->next;
            free(tmp);
        } else {
            pcur = &(*pcur)->next;
        }
    }
}
```

Если для работы со списком используется, как и раньше, указатель с именем `first`, то теперь удалить из этого списка все отрицательные числа можно так:

```
delete_from_int_list(&first, &is_negative);
```

а удалить все числа, делящиеся без остатка на семь — вот так:

```
delete_from_int_list(&first, &is_div7);
```

Ещё одному важному примеру применения указателей на функции мы посвятим следующий параграф.

4.9.3. Функции обратного вызова (callback)

Появление в нашем инструментарии указателей на функции даёт нам возможность применить классический подход к взаимодействию подсистем в программе, называемый *callback function*[37]. Подход состоит в том, что один модуль (пользователь) поручает другому модулю (серверу) отыскать (найти в памяти, загрузить по сети, прочитать из файла, получить из базы данных и т. п.) некие данные; по мере того, как подходящие данные будут появляться, сервер должен вызывать некую функцию пользователя[38], передавая этой функции в качестве параметра найденные данные.

Поскольку сервер, как правило, хочется сделать по возможности универсальным, об устройстве пользовательского модуля он должен знать как можно меньше; поэтому сервер, естественно, не знает[39], какую именно функцию пользователя надо вызывать; ему передают указатель на нужную функцию.

Ещё один нетривиальный момент состоит в том, что для пользователя обработка найденных объектов данных может быть единым процессом, то есть на обработку последующих объектов влияют результаты обработки предыдущих; для хранения этих результатов пользователь вынужден завести некую память (переменную, массив и т. п.), причём к этой памяти необходимо обеспечить доступ для callback-функции; теоретически это можно сделать через глобальную переменную, но глобальные переменные, во-первых, зло сами по себе, а во-вторых, таким образом мы лишили бы себя возможности запустить два и более процесса обработки одновременно — например, обращаясь к ним по очереди. Но если не использовать глобальные переменные, то остаётся только один способ передачи информации в callback-функцию — через её параметр; при этом для разных пользователей может потребоваться применение совершенно разных данных, и сервер ничего об этом знать не должен.

Все проблемы одним махом снимаются путём передачи в callback-функцию нетипизированного указателя, то есть адреса типа `void*`. Соответствующий параметр обычно называется *пользовательскими данными* (англ. *user data*). Заведя у себя нужную структуру данных, пользователь передаёт серверу адрес callback-функции и нетипизированный адрес созданных данных; по мере нахождения искомых объектов сервер вызывает callback-функцию, используя переданный адрес; в качестве параметров сервер передаёт ей, во-первых, очередной найденный объект, и, во-вторых, нетипизированный указатель на пользовательские данные.

За основу нашего примера мы возьмём обход двоичного дерева поиска, реализация которого приведена на стр. 53. Напомним, что мы рассматривали дерево, в котором хранятся целые числа. Функция, которую мы написали, умеет, обходя дерево, делать только одну операцию: печатать числа, хранящиеся в узлах дерева. Мы изменим её так, чтобы она *для каждого найденного в дереве числа вызывала некую «пользовательскую» функцию*, передавая ей, во-первых, найденное число, и, во-вторых, нетипизированный адрес пользовательских данных, который она сама получила в качестве параметра.

Наша callback-функция должна, следовательно, получать два параметра (`int` и `void*`), а возвращать ничего не должна. Её профиль будет примерно таким:

```
void callback_function(int num, void *userdata);
```

Сама функция обхода дерева станет при этом такой:

```
void int_bin_tree_traverse(struct node *r,
                           void (*callback)(int, void*),
                           void *userdata)
{
    if(!r)
        return;
    int_bin_tree_traverse(r->left, callback, userdata);
    (*callback)(r->val, userdata);
    int_bin_tree_traverse(r->right, callback, userdata);
}
```

Чтобы решить с помощью этой функции прежнюю задачу — напечатать все элементы дерева — нам придётся ввести дополнительную функцию в качестве callback'а:

---

[37] На русский язык это буквально переводится как *функция для обратного вызова*, но программисты не используют ни этот, ни какой-либо другой перевод, а применяют английское слово *callback*, которое читается примерно как *колбэ́к*. Это, конечно, жаргонизм, но, какой бы перевод вы ни пытались использовать, вы рискуете остаться непонятыми.

[38] Подчеркнём ещё раз, что под *пользователем* в данном случае понимается не человек, сидящий за компьютером, а тот *модуль нашей программы*, который использует услуги, предоставленные модулем-сервером.

[39] Возможно, программист, пишущий модуль, прекрасно знает, как устроены все модули, которые будут использовать его подсистему, но даже в этом случае он не должен своим знанием пользоваться — следует исходить из предположения, что кто-нибудь когда-нибудь допишет другие модули.

```
void int_callback_print(int data, void *userdata)
{
    printf("%d ", data);
}
```

Обратите внимание, что параметр `userdata` здесь не используется; в самом деле, печать очередного элемента никак не зависит от «результатов» печати предыдущих элементов, так что передавать в качестве пользовательских данных здесь нечего. Если адрес корневого элемента дерева находится в переменной `root`, то напечатать содержимое дерева мы теперь сможем, например, так:

```
int_bin_tree_traverse(root, int_callback_print, NULL);
```

В качестве адреса пользовательских данных мы передали `NULL`; можно было, в принципе, передать что угодно, всё равно это значение игнорируется нашим callback'ом, но написать в подобной ситуации именно `NULL` — значит облегчить потенциальному читателю понимание нашей программы.

Пока назначение «пользовательских данных» осталось непонятным, но это лишь потому, что мы рассмотрели такой «особенный» пример. Всё станет ясно, если мы попробуем что-то сделать не с каждым элементом дерева в отдельности, а со всем деревом целиком. Для начала попробуем просуммировать элементы дерева; для этого нам потребуется переменная типа `int`, в которой будет накапливаться сумма; именно её адрес мы и передадим в качестве пользовательских данных в callback-функцию, которая для этой задачи примет следующий вид:

```
void int_callback_sum(int data, void *userdata)
{
    int *sum = userdata;
    *sum += data;
}
```

Зная, что через параметр `userdata` нам передали адрес целочисленной переменной, мы приводим этот адрес к типу `int*` и, используя его, прибавляем к текущему значению сумматора значение числа из текущего элемента дерева. Это можно записать и короче, без использования локальной переменной:

```
void int_callback_sum(int data, void *userdata)
{
    *(int*)userdata += data;
}
```

Для подсчёта суммы мы теперь можем сделать следующее:

```
int sum;
sum = 0;
int_bin_tree_traverse(root, int_callback_sum, &sum);
```

После выполнения этого фрагмента в переменной `sum` окажется сумма всех элементов дерева.

Рассмотрим более сложную задачу: пусть нам за один проход дерева требуется найти минимальное число, максимальное число и общее количество чисел. Пользовательские данные при этом будут представлять собой *структуру* из трёх полей:

```
struct minmaxcount {
    int count, min, max;
};
```

Напишем следующую callback-функцию:

```
void int_callback_minmaxcount(int data, void *userdata)
{
    struct minmaxcount *mmc = userdata;
    if(mmc->count == 0) {
        mmc->min = data;
        mmc->max = data;
    } else {
        if(mmc->min > data)
            mmc->min = data;
        if(mmc->max < data)
            mmc->max = data;
    }
    mmc->count++;
}
```

Вызов теперь будет выглядеть так:

```
struct minmaxcount mmc;
mmc.count = 0;
int_bin_tree_traverse(root, int_callback_minmaxcount, &mmc);
```

После выполнения этого фрагмента структура окажется заполнена значениями, хотя поля `min` и `max` получат значения только в случае, если хотя бы один элемент в дереве присутствовал; это проверяется по значению поля `count`: если он отличен от нуля, имеет смысл рассматривать минимум и максимум, в противном случае их следует игнорировать, осмысленной информации они не содержат. В самом деле, каков наибольший или наименьший элемент пустого множества? Очевидно, ответ не определён.

4.9.4. Сложные описания и общие правила их прочтения

Указатели на массивы и функции резко усложняют общую структуру объявлений и описаний функций и переменных. Происходит это из-за наличия в описаниях *префиксных* и *суффиксных* символов; к префиксным относятся имена типов, символы «`*`» и модификаторы, включая `const`, а к суффиксным — квадратные скобки с обозначением размерности массива (или без неё, если она может быть восстановлена по структуре инициализатора или вообще не важна в данном контексте) и круглые скобки со списками параметров для вызовов функций. Хаоса добавляет ещё и то, что круглые скобки, помимо обозначения функций, используются также для изменения порядка применения символов в описателе, как мы это видели для указателей на массивы и функции.

Для создания общего впечатления о масштабах катастрофы приведём несколько примеров. Пусть имеется функция

```
int f(int x, int y) { return x + y; }
```

и мы решили написать такую функцию, которая принимает на вход целое число и возвращает адрес этой `f` (или другой функции с таким же профилем) в качестве своего значения. Назовём эту функцию, например, `funret`. Выглядеть это будет так:

```
int (*funret(int x))(int, int)
{
   /* ... */
}
```

Пусть теперь у нас есть десяток функций с профилем как у `f` и мы решили поместить указатели на эти функции в массив с именем, скажем, `funvec`. Попытавшись описать такой массив, мы получим следующее:

```
int (*funvec[10])(int, int);
```

Допустим теперь, что у нас есть двумерный массив чисел типа `double`, например, такой:

```
double matrix[100][10];
```

— причём мы в разных случаях используем из него группы по десять идущих подряд строк, чтобы сформировать матрицу 10x10. Для этого нам может потребоваться функция, возвращающая значение выражения вида `matrix+n`, где `n` — некоторое целое число, означающее строку, начиная с которой мы хотим выделить очередную группу строк. Мы помним, что выражение такого вида, как и сама адресная константа `matrix`, имеет тип `double (*t)[10]` (указатель на массив из десяти элементов типа `double`). Функция, которая возвращает такое выражение, будет выглядеть вот так:

```
double (*select_segment(int a, int b))[10]
{
    /* ... */
}
```

А теперь представьте, что вам зачем-то потребовался *указатель* на такую функцию. Выглядеть этот монстр будет так:

```
double (*(*selptr)(int, int))[10];
```

Если кто-нибудь скажет вам, что «это всё на самом деле совсем просто», не верьте: даже самые опытные программисты на Си при виде таких построений вынуждены остановиться хотя бы на секунду, чтобы понять, что́ же здесь написано.

Всё становится ещё хуже, если указатели на функции и массивы приходится передавать в функцию через параметр, а на саму эту функцию делать указатель. Вернёмся к функции `int f(int a, int b)`, с которой мы начали этот параграф. Пусть нам потребовалась функция, которую мы назовём `replace_f` и которая получает в качестве параметра адрес функции с профилем, как у `f`, и адрес такого же типа возвращает. Например, такая функция может нам потребоваться, если какой-нибудь модуль использует внутри себя указатель на функцию, подобную `f`, и значение этого указателя меняется через вызов `replace_f`, при этом, установив новое значение такого указателя, `replace_f` возвращает его старое значение, например, чтобы вызывающий мог его сохранить, а в будущем вернуть на место. Функция `replace_f` сама по себе будет выглядеть не очень страшно, по крайней мере, после всего, что мы уже видели:

```
int (*replace_f(int (*func)(int, int)))(int, int)
{
    /* ... */
}
```

— но вот описание указателя на неё способно вогнать в ступор кого угодно. Второй идентификатор в описании *одной* переменной выглядел бы несколько странно, поэтому параметр функции, который в описании самой функции имеет имя `func`, в описании указателя на эту функцию желательно оставить безымянным. При этом тип параметра `int (*func)(int, int)` всё-таки как-то нужно указать, и результатом становится совершенно неудобоваримая последовательность символов «`(*)`», означающая «здесь мог бы быть идентификатор, но его тут нет, так что начинайте отсюда». Тип параметра в итоге выглядит так: `int (*)(int, int)`, а полностью описание указателя на функцию `replace_f` (или другую с тем же профилем) окажется таким:

```
int (*(*replace_f_ptr)(int (*)(int, int)))(int, int);
```

Теоретически имя параметра можно было бы и не опускать, что спасло бы нас от звёздочки в скобках, но ясности бы уж точно не добавило. Смотрите сами:

```
int (*(*mptr)(int (*func)(int, int)))(int, int);
```

Вы сможете без раздумий сказать, который из двух идентификаторов — `mptr` или `func` — здесь описывается? Пока идентификатор был один, мы хотя бы точно знали, откуда начинать распутывать это безобразие.

Понятно, что и это не предел, но мы здесь всё же остановимся — с происходящим пора что-то делать. Вспомнив о существовании директивы `typedef`, введём для начала имя для того типа, который у нас оказался на входе и выходе из функции:

```
typedef int (*fptr)(int, int);
```

Теперь описание указателя `replace_f_ptr` можно будет написать гораздо проще:

```
fptr (*replace_f_ptr)(fptr);
```

Ясно, что именно так с самого начала и следовало действовать; к сожалению, несмотря ни на что, *это очевидно далеко не всем*, поэтому при чтении чьих-нибудь программ вы можете в любой момент нарваться на нагромождение скобочек и звёздочек, подобное чему-то из приведённого выше. Поэтому желательно уметь подобные вещи читать, сколь бы нелепым ни выглядело маниакальное нежелание отдельных программистов применять `typedef`. Тем более что правила такого чтения довольно просты, нужно только не запутаться в скобочках.

Чтобы прочитать сложное описание или объявление, нужно сначала найти описываемый идентификатор; всё начинается именно с него. Затем мы двигаемся от него «наружу» — вправо к концу описания и влево к его началу. Суффиксные символы, стоящие справа от нас, обозначающие функции и массивы, имеют приоритет над префиксными, стоящими от нас слева. Если вокруг нас оказалась пара круглых скобок, нужно исчерпать символы внутри этих скобок, и только потом выходить за их пределы. При этом начинаем мы чтение с того, что произносим описываемое имя-идентификатор и добавляем союз «это»; двигаясь «наружу», мы, увидев квадратные скобки, произносим «массив из стольких-то элементов типа...»; увидев круглые скобки справа от нас, произносим «функция(-ю), которая получает на вход...», читаем описания типов параметров, затем продолжаем словами «и возвращает...»; видя звёздочку, произносим «указатель на...»[40]; видя слово `const`, произносим слово «константный(-ая)»; наконец, добравшись до стоящего в самом начале имени типа, завершаем чтение подходящей концовкой вроде «переменная(-ую) типа `int`», «значение типа `int`», «область памяти типа `int`» (последнее приходится применять, если у нас вырисовывается константный указатель на начало массива). Конечно, не следует забывать про согласование падежей и прочую косметику, позволяющую остаться в рамках правил русского языка.

Например, уже знакомое нам

```
int (*(*replace_f_ptr)(int (*)(int, int)))(int, int);
```

можно прочитать следующим образом: `replace_f_ptr` — *это* (справа мы упёрлись в закрывающую круглую скобку, так что сначала надо посмотреть, что у нас слева) *указатель на* (содержимое скобок кончилось, выходим наружу, справа что-то есть, поэтому начинаем оттуда) *функцию, которая получает на вход* (читаем тип параметра, для чего находим «исполняющую обязанности идентификатора» конструкцию «`(*)`») *указатель на функцию, принимающую на вход два целых числа и возвращающую значение типа* ***`int`*** (список формальных параметров закончился, завершаем фразу про функцию) *и возвращает* (справа опять закрывающая круглая скобка, приходится посмотреть налево) *указатель на* (опять кончилось содержимое круглых скобок, выходим наружу) *функцию, принимающую на вход два целых числа и возвращающую число типа* ***`int`***.

Аналогичным образом описание

```
int (*funvec[10])(int, int);
```

читается так: ***`funvec`*** *— это массив из десяти указателей на функции, принимающие на вход два целых числа и возвращающие значение типа* ***`int`***.

Любопытно, что если случайно забыть звёздочку и скобки, можно придать описанию такой смысл, будто мы имеем в виду не указатель на функцию или массив, а «сам по себе» массив или «саму по себе» функцию. Легко можно убедиться, что бдительный компилятор не позволит нам подобных вольностей: функция «сама по себе» в Си бывает только если это действительно прототип или описание функции (а не какой-либо переменной, типа и тому подобного), и точно так же массив «сам по себе» бывает только при описании или объявлении массива, ну и ещё при применении `sizeof`. Функцию «саму по себе» нельзя ни получать как параметр, ни возвращать как значение; массив «сам по себе», как ни странно, можно передавать как параметр (по правде говоря, это совершенно то же самое, как и передавать указатель), но вот возвращать массив «сам по себе» нельзя.

Например, следующие описания ошибочны и компилятор их обрабатывать откажется:

```
void (*f15)(int)(int);  /* ОШИБКА! */
   /* указатель на функцию, возвращающую функцию */

void *m[5](int);        /* ОШИБКА! */
   /* массив функций, а не указателей на них */

int (*f100)()[15];      /* ОШИБКА! */
   /* указатель на функцию, возвращающую массив */
```

[40] Это не всегда правильно, поскольку это не всегда именно указатель, а может быть просто адрес; по-английски можно было бы произнести что-то вроде *address of...*, но в русском языке нет соответствующего предлога, вместо этого следующую сущность нужно поставить в родительный падеж, но это уже сложнее описать; в любом случае большинство программистов не делает различия между понятиями «адрес» и «указатель», а зря.

4.10. Ещё о возможностях стандартной библиотеки

Стандартная библиотека Си содержит достаточно большое количество разнообразных инструментов для работы; конечно, рассмотреть их все мы не сможем, да это и не нужно. Материала этой главы вам хватит, чтобы составить общее представление о возможностях стандартной библиотеки; дальнейшее — дело самостоятельного изучения. Учтите, что **большинство возможностей, которые рассматриваются в этой главе, лучше не использовать, пока вы не научитесь уверенно программировать на Си**.

4.10.1. Дополнительные функции работы с кучей

Ранее мы рассмотрели всего две функции для работы с динамической памятью — `malloc` для выделения и `free` для освобождения. С освобождением всё довольно просто, ничего кроме `free` для этого не требуется, а вот для выделения памяти есть ещё одна функция — `calloc`:

```
void *calloc(int nmemb, int size);
```

Эта функция предназначена для выделения памяти под массив, содержащий `nmemb` элементов, каждый из которых имеет размер `size`, но это с таким же успехом можно проделать, перемножив эти два числа и вызвав `malloc`; фундаментальное отличие `calloc` от `malloc` состоит в том, что она забивает выделяемую память нулями. Некоторые программисты отдают ей предпочтение именно по этой причине.

Нельзя не упомянуть также функцию `realloc`, которая исходно предназначена для изменения размера уже выделенной области памяти. Профиль её таков:

```
void *realloc(void *ptr, int size);
```

Параметр `ptr` указывает на имеющуюся область памяти, для которой нужно поменять размер, а `size` задаёт этот новый размер. В большинстве случаев функция вынуждена создавать новую область динамической памяти, копировать туда всё содержимое старой области (которую вы задаёте адресом `ptr`), саму эту старую область освобождать, а возвращать адрес новой. После этого адрес, переданный через `ptr`, становится невалидным, и вместо него нужно использовать адрес, который вернул `realloc`, поэтому чаще всего результат `realloc`'а присваивают той же переменной, значение которой передают его первым параметром, примерно так: `p = realloc(p, sz);`. Несомненное достоинство `realloc` состоит в том, что иногда, когда сразу после данной области памяти достаточное количество памяти свободно, она может сэкономить процессорное время, не копируя информацию — то есть реально увеличив размер имеющейся области памяти. Есть у `realloc` и недостаток. Ясно, что она не может знать, *как* в вызывающей программе используется изменяющаяся область памяти; если всё же приходится копировать, функция вынуждена копировать всю область целиком, даже если программа реально использует из неё один байт.

Довольно интересны два специальных случая вызова `realloc`: если первым параметром передан нулевой адрес, то функция работает в точности как `malloc`, если же в качестве второго параметра передан ноль, она работает в точности как `free`. Иногда попадаются программы, написанные на Си, в которых для работы с динамической памятью используется один только `realloc`; пожалуй, так лучше всё же не делать, ясности это не прибавит.

4.10.2. Функции обработки строк

Все функции, перечисленные в этом параграфе, можно очень легко написать самому. Бытует мнение, что библиотечные функции работают быстрее за счёт того, что в их реализации используются архитектурно-зависимые ассемблерные вставки; но на самом деле это не так, при современном состоянии оптимизаторов использование ассемблерных вставок никакого выигрыша не даёт. Единственный (но, заметим, вполне достаточный) смысл применения этих функций — в том, чтобы *облегчить сторонним людям чтение вашей программы*. В самом деле, операции, реализуемые этими[41] функциями, бывают нужны очень часто; если использовать свою реализацию таких операций, читатель вашей программы будет вынужден потратить время, пусть и небольшое, на то, чтобы понять, о чём идёт речь, тогда как при использовании стандартных функций любой, кто знает Си, сразу поймёт, что здесь делается.

Как следствие, в серьёзной разработке следует пользоваться для операций над строками именно этими функциями; но это не значит, что ими следует пользоваться с самого начала обучения программированию. Автор книги много раз видел студентов, легко жонглирующих всякими `strlen`, `strcmp`, `strcpy` и прочими, при этом не понимающих, как устроена строка и не способных, как следствие, сделать со строкой ничего сколько-нибудь нестандартного. **Пока вашей основной целью является обучение и пока вы не почувствовали полной уверенности в обращении со строками, лучше вообще не прикасаться к стандартным строковым функциям.**

Прототипы всех функций из этого параграфа находятся в заголовочном файле `string.h`. Подключайте его только тогда, когда уже достигнете уровня свободного и уверенного владения преобразованиями строк на Си, и не раньше; до той поры реализуйте все эти операции сами, как это делалось в примерах в нашей книжке. Для самопроверки обязательно напишите свои собственные реализации всех перечисленных функций, и только когда всё получится, можете переходить к использованию их библиотечных версий.

Начнём с функции `strlen`, определяющей длину строки:

```
int strlen(const char *str);
```

Здесь вряд ли требуется много пояснений; ранее в примерах мы использовали свою собственную функцию `string_length`, которая делала абсолютно то же самое.

Функция `strcpy` предназначена для копирования строки в заранее подготовленный массив:

```
char *strcpy(char *destination, const char *source);
```

Порядок аргументов тут такой же, как в присваивании: сначала «куда», потом «откуда». Иначе говоря, параметр `destination` задаёт адрес начала массива, *в который* производится копирование, а параметр `source` указывает на начало *строки, которая должна быть скопирована*. **В массиве должно быть достаточно места, и забота об этом лежит на пользователе!** У вас должны быть веские основания считать, что вы выделили достаточно места или описали массив достаточного размера; если присутствует хотя бы малейшая неуверенность, нужно сначала измерить длину копируемой строки. О возможных последствиях переполнения массива мы писали в §4.4.4 (см. стр. 44); напомним, что за такие вещи увольняют с работы и правильно делают.

Когда с длиной строки могут, как говорится, «быть варианты», существенно правильнее использовать другую функцию, `strncpy`:

```
char *strncpy(char *destination, const char *source, int size);
```

Первые два аргумента используются точно так же, как и для `strcpy`, а третий задаёт размер массива `destination`; если задать его правильно, функция не допустит переполнения, хотя при этом может быть скопирована не вся строка. У этой функции есть другой крайне неприятный недостаток: если не хватило места для копирования, то нулевой байт она в конец массива не записывает, так что в вашем массиве окажется нечто, не являющееся корректной строкой с точки зрения Си. Впрочем, это легко обойти, записав такой байт принудительно; более того, если в массиве осталось свободное место, `strncpy` заполняет его нулями, так что по содержимому последнего байта массива можно понять, хватило места или нет: если там ноль — значит, копия строки полностью уместилась в отведённом пространстве, а если не ноль — то строка была обрезана; если вас это не беспокоит, остаётся занести в последний байт ноль и продолжить работу.

Функции `strcpy` и `strncpy` в качестве своего значения зачем-то возвращают параметр `destination`, причём всегда. Как следствие, их обычно вызывают ради побочного эффекта, игнорируя возвращаемое значение.

Если вам не хочется возиться с самостоятельным выделением памяти под копию строки, можно воспользоваться функцией `strdup` (эта функция пришла из спецификаций, имеющих отношение к Unix; в системе, отличной от Unix, её может не найтись):

[41] Во всяком случае, теми, которые будут рассмотрены; библиотека содержит много других функций, в том числе и для работы со строками, которые применяются гораздо реже и польза которых сомнительна.

```
char *strdup(const char *s);
```

Она сама измерит длину вашей строки, создаст с помощью `malloc` область динамической памяти нужной длины, скопирует туда заданную строку и вернёт адрес начала созданной копии. Не забудьте только потом освободить память с помощью `free`. Многие программисты не любят эту функцию, поскольку она затрудняет проверку правильности программы, ведь выделение памяти происходит не там, где её освобождение: память выделяется где-то в недрах библиотечной функции, а освобождать вы её вынуждены в своей программе.

Для сравнения строк применяются функции `strcmp` и `strncmp`:

```
int strcmp(const char *s1, const char *s2);
int strncmp(const char *s1, const char *s2, int n);
```

Обе функции посимвольно сравнивают строки, начиная с адресов `s1` и `s2`; работа заканчивается, если строки различаются символом в очередной позиции, если одна из строк кончилась, а `strncmp` дополнительно заканчивает работу, если просмотрено уже `n` позиций. Если очередной символ в строках различается, та строка, в которой в очередной позиции оказался символ с меньшим кодом, считается «меньше», чем другая. Если одна из строк кончилась, а вторая ещё нет, кончившаяся считается «меньше» второй. Функции возвращают отрицательное число, если `s1` оказалась «меньше», чем `s2`, положительное — если `s2` оказалась «меньше», чем `s1`, и ноль, если различий обнаружено не было. Правило упорядочивания строк можно выразить одной фразой: строки сравниваются *в лексикографическом (словарном) порядке*.

Иногда в программах можно встретить конструкцию вроде


```
if(!strcmp(str1, str2)) {
    /* ... */
}
```

Как можно догадаться, имеется в виду условие «если строки равны», ведь в этом случае `strcmp` возвращает ноль. Но так писать не надо. Подсознательно операция отрицания в условии воспринимается как «что-то не получилось», а в данном случае всё, наоборот, замечательно получилось: строки оказались равны. Лучше написать так:

```
if(0 == strcmp(str1, str2)) {
    /* ... */
}
```

— причём ноль оставить слева от операции сравнения: так вы избавляете читателя от необходимости искать глазами правый край условного выражения.

С помощью следующих функций можно найти в заданной строке заданный символ:

```
char *strchr(const char *s, int c);
char *strrchr(const char *s, int c);
```

Обе функции отыскивают символ с кодом `c` в строке, начинающейся с адреса `s`. Различие между ними проявляется только в случае, если таких символов больше одного: функция `strchr` находит *первый* такой символ, а `strrchr` — *последний* (`r` от слова *right*, то есть «правый»). Обе функции возвращают адрес того места, где в строке располагается искомый символ; если символа с заданным кодом в строке не нашлось, возвращается `NULL`.

Функция `strstr` позволяет найти вхождение подстроки в строку:

```
char *strstr(const char *haystack, const char *needle);
```

Этот заголовок заимствован из официальной man-страницы по этой функции; *haystack* по-английски «стог сена», *needle* — иголка. Как можно догадаться, первый параметр задаёт строку, где нужно искать, второй — подстроку, которую нужно искать. Возвращается адрес того места в строке, где найдена искомая подстрока, либо `NULL`, если ничего не нашлось. Например,

```
char ak[] = "abrakadabra";
/* ... */
p = strstr(ak, "kada");
```

занесёт в `p` адрес `ak+4`.

Следующие три функции не имеют прямого отношения к строкам в том смысле, что они не рассматривают нулевой байт в качестве признака конца полезной информации; их объектом работы являются произвольные области памяти. Функция `memset` позволяет заполнить область памяти заданного размера, начиная с заданного адреса, заданным значением байта:

```
void *memset(void *memory, int value, int size);
```

Первый параметр задаёт адрес начала, второй — значение, которым следует заполнять ячейки памяти, и третий — сколько ячеек надо заполнить. От значения `value` используются только младшие восемь бит, остальные отбрасываются. Функция возвращает свой первый параметр; зачем она это делает — неизвестно.

Следующие две функции делают практически одно и то же — копируют содержимое заданного размера из одной области памяти в другую:

```
void *memcpy(void *dest, const void *src, int size);
void *memmove(void *dest, const void *src, int size);
```

Для обеих функций первый параметр задаёт, куда производится копирование, второй — откуда, а третий — размер копируемой области. Различие между функциями в том, что `memcpy` предназначена для копирования информации между областями, которые не могут накладываться друг на друга; `memmove` в этом плане хитрее, она делает соответствующие проверки и если области памяти `dest` и `src` перекрываются, производит копирование от начала к концу, когда `dest` находится левее `src`, и от конца к началу, когда правее. Результат копирования, таким образом, всегда остаётся корректным. Эту функцию можно использовать, например, когда в имеющуюся строку нужно вставить несколько символов (в середину) или, наоборот, изъять из строки несколько символов; в обоих случаях остаток строки приходится сдвигать — соответственно вправо и влево. Если по смыслу происходящего области памяти перекрываться не могут (например, `dest` указывает на только что выделенную память, или они относятся к заведомо разным объектам, и т. д.), лучше использовать `memcpy`: она работает быстрее за счёт отсутствия лишних проверок (кто левее, кто правее).

4.10.3. Генерация псевдослучайных чисел

Мы уже обсуждали случайные и псевдослучайные числа при изучении Паскаля (см. т. 1, §2.8.5). Потребность внести в выполнение программы какое-то разнообразие чаще всего возникает при создании компьютерных игр, хотя, конечно, не только; случайные числа используются и в математических расчётах (небезызвестный «метод Монте-Карло»), и в области обеспечения безопасности (например, можно узнать, что мы имеем дело с тем, за кого он себя выдаёт, если он сможет подписать своей электронной подписью предложенное нами число, но для этого нужно, чтобы заранее этого числа никто не знал). Получение *действительно случайных* чисел, таких, которые невозможно предугадать со сколь бы то ни было приемлемой вероятностью, представляет собой серьёзную техническую проблему.

Современные операционные системы способны генерировать случайные числа на основе случайных событий, таких как временны́е промежутки между приходящими по сети пакетами или нажатиями клавиш на клавиатуре, длительность дисковых обменов (точнее, не вся длительность, которая довольно хорошо предсказывается, а её отклонения от прогнозируемых величин) и т. п. Такой случайной информации, называемой в приложениях словом «энтропия», в систему поступает сравнительно немного, поэтому *настоящих* случайных чисел можно сгенерировать довольно ограниченное количество.

Получить «настоящие» случайные числа в ОС Linux можно, открыв на чтение файл псевдоустройства `/dev/random`; это можно сделать обычной функцией `fopen`, а само чтение выполнить с помощью `fread`, либо воспользоваться напрямую системными вызовами `open` и `read`, которые мы подробно рассмотрим позже. Этим методом стоит пользоваться только в случаях, когда вам действительно нужны хорошо распределённые и непредсказуемые числа — например, при генерации всевозможных паролей и криптографических ключей.

Во многих случаях, как правило, не связанных с безопасностью — например, в тех же компьютерных играх — непредсказуемость случайных чисел не слишком важна, поскольку никто, скорее всего, не станет пытаться их предсказывать. В такой ситуации случайные числа обычно заменяют ***псевдослучайными***, получить которые намного проще.

Стандартная библиотека языка Си предусматривает для генерации псевдослучайных чисел две функции, прототипы которых описаны в заголовочном файле `stdlib.h`:

```
int rand();
void srand(unsigned int seed);
```

Функция `rand`, не принимающая параметров, возвращает случайное (точнее, псевдослучайное) число от `0` до значения `RAND_MAX` включительно. Это значение в разных реализациях библиотеки может отличаться; например, на машине автора книги оно равно максимально возможному значению для 32-битного знакового целого. Если `rand` использовать одну, без `srand`, то получающаяся последовательность выдаваемых чисел будет одна и та же при каждом запуске программы — а именно, такая же, как если бы мы в начале программы написали `srand(1)`.

Чтобы последовательность чисел получалась каждый раз другая, нужно *один раз в начале программы* вызвать функцию `srand`, дав ей в качестве параметра какое-то число, которое будет каждый раз новым. Эта функция инициализирует датчик псевдослучайных чисел, задав начальное число последовательности (само это число в последовательность псевдослучайных чисел, конечно, не войдёт).

Естественным образом возникает вопрос, откуда взять число для `srand`. Пожалуй, самый простой вариант — это «скормить» функции текущее время. Например, если подключить заголовочный файл `time.h`, вам станет доступна функция `time`, выдающая *текущее время как число секунд, прошедшее с 1 января 1970 года*. Функция принимает некий параметр, в качестве которого практически всегда передают `NULL`. С учётом этого вы можете инициализировать генератор случайных чисел примерно так:

```
srand(time(NULL));
```

Напомним, что это надо сделать *один раз* в начале выполнения программы! Начинающие иногда делают ошибку, вызывая `srand` (например, так, как показано выше) перед каждым обращением к `rand`; в такой ситуации `rand`, например, будет возвращать одинаковые числа, если к ней обратиться несколько раз на протяжении одной секунды; это наверняка не то, чего вы хотите.

Ни в коем случае не пользуйтесь этим способом инициализации датчика случайных чисел, если хоть что-то зависит от секретности получаемой последовательности! Например, если вам потребуется сгенерировать случайный пароль и вы при этом сделаете `srand(time(NULL))`, а некий злоумышленник будет хотя бы с точностью до дня знать, *когда* с помощью вашей программы был сгенерирован некий пароль, ему нужно будет перебрать всего 86400 вариантов, ведь в сутках именно столько секунд. Устроить такой перебор — дело техники, притом несложной. Впрочем, если вы имеете дело с паролями, пин-кодами и прочей информацией, связанной с безопасностью, генератор случайных чисел из стандартной библиотеки вам в любом случае не подходит.

Часто требуется не просто случайное число, а число из заданного диапазона. Проще всего этого добиться с помощью взятия остатка от деления; например, случайное число r от 1 до 12 можно получить, вычислив выражение `rand()%12+1`. Но так делать не рекомендуется, поскольку распределение младших разрядов псевдослучайных чисел может быть неравномерным, а именно младшие биты при таком применении `rand` оказывают влияние на получаемые числа. Одна из версий страницы справочника `man` по функции `rand`, ссылаясь на книгу [2], рекомендует применять в подобных ситуациях следующее выражение:

```
1 + (int)(12.0*rand()/(RAND_MAX+1.0))
```

Это позволяет использовать в равной степени все составляющие псевдослучайного числа, а в целом псевдослучайные числа имеют распределение, достаточно близкое к равномерному.

4.10.4. (*) Средства создания вариадических функций

Под ***вариадическими функциями*** понимаются функции, способные принимать переменное число аргументов. Мы уже встречались с ними: именно таковы `printf`, `scanf` и все их «родственники», и все они сами написаны на Си. Любопытно отметить, что средства создания вариадических функций, как и многое другое, вытеснены из языка в библиотеку. Чтобы их задействовать, нужно подключить заголовочный файл `stdarg.h`, в котором описан тип `va_list` (нас не должно волновать, что́ он собой представляет; в разных системах реализации этого типа могут быть совершенно различны) и три макроса — `va_start`, `va_end` и `va_arg`.

Сама вариадическая функция должна иметь хотя бы один параметр, снабжённый именем; имя последнего из именованных параметров используется макросом `va_start` для инициализации переменной типа `va_list`. Последующие параметры извлекаются из стека макросом `va_arg`, которому нужно указать, параметр какого типа требуется извлечь; в конце работы (обязательно в той же самой функции!) мы должны поместить макровызов `va_end`, который приведёт дела в порядок.

Припомнив свои знания архитектуры i386 и конвенции CDECL, мы можем прикинуть, что простейшая реализация всей этой механики состоит во взятии адреса именованного параметра, что позволит с помощью арифметики указателей двигаться вдоль стекового фрейма, извлекая последующие параметры; однако не во всех случаях это делается именно так, на некоторых архитектурах и при использовании некоторых компиляторов параметры полностью или частично передаются через регистры, да и при использовании стека существуют другие возможные решения — например, создать «массив из одного элемента» и обращаться к элементам этого массива, и тому подобное.

Имеющиеся средства не позволяют понять, сколько параметров было фактически передано в функцию; это предмет соглашения между вызывающим и вызываемым. Например, в функцию `printf` должно быть, помимо форматной строки, передано ровно столько параметров, сколько форматных директив имеется в форматной строке.

Приведём простой пример. Следующая функция принимает произвольное (но не менее одного) количество параметров типа `int`, не равных нулю, и возвращает их сумму; последний параметр в списке параметров должен быть, напротив, равен нулю — это позволяет функции понять, что список параметров кончился:

```
int sum(int c, ...)
{
    va_list vl;
    int s = c, k;

    va_start(vl, c);
    while((k = va_arg(vl, int)) != 0)
        s += k;
    va_end(vl);

    return s;
}
```

Как видим, все три макроса в качестве первого параметра принимают переменную типа `va_list`, причём, поскольку это макросы, нет необходимости с этой переменной делать что-то дополнительное, брать её адрес и т. п. Для макроса `va_arg` вторым параметром выступает тип очередного безымянного параметра нашей функции, и этот макрос разворачивается в выражение, имеющее именно такой тип. У макроса `va_end` всего один параметр — всё та же переменная типа `va_list`.

Приведём более сложный пример. Следующая функция принимает на вход попеременно указатель на строку и целое число, и так сколько угодно раз, пока очередной указатель не окажется нулевым; каждая строка печатается столько раз, сколько указано в следующем за ней параметре (числе):

```
void print_times(const char *str, ...)     /* variadic.c */
{
    va_list vl;
    const char *p;

    va_start(vl, str);
    for(p = str; p; p = va_arg(vl, const char *)) {
        int n, i;
        n = va_arg(vl, int);
        for(i = 0; i < n; i++)
            printf("%s ", p);
        printf("\n");
    }
    va_end(vl);
}
```

Пример вызова этой функции приведём такой:

```
print_times("once", 1, "twice", 2, "seven times", 7, NULL);
```

В заключение рассказа о вариадических функциях обратим внимание читателя на функции

```
int vprintf(const char *fmt, va_list ap);
int vfprintf(FILE *stream, const char *fmt, va_list ap);
int vsprintf(char *str, const char *fmt, va_list ap);
int vsnprintf(char *str, int size, const char *fmt, va_list ap);
int vscanf(const char *fmt, va_list ap);
int vsscanf(const char *str, const char *fmt, va_list ap);
int vfscanf(FILE *stream, const char *fmt, va_list ap);
```

Все эти функции работают точно так же, как и их аналоги без буквы `v` в имени (то есть `printf`, `fprintf` и т. д.), за исключением того, что вместо списка аргументов переменной длины все эти функции принимают параметр типа `va_list`; это позволяет создавать функции, аналогичные функциям семейств `printf`/`scanf`, то есть принимающие на вход форматную строку и параметры, но при этом производящие какие-то дополнительные действия, а затем вызывающие соответствующие библиотечные функции из приведённого списка для выполнения основной работы.

4.11. (*) Полноэкранные программы на Си

При изучении языка Паскаль мы рассматривали библиотечный модуль `crt`, позволяющий создавать полноэкранные терминальные программы, то есть такие программы, которые работают в окне терминала, используя все его возможности — выводя текст в произвольные места, изменяя цвет фона и букв, заставляя буквы мигать, реагируя на нажатия клавиш, не дожидаясь, пока пользователь нажмёт Enter, и т. д. (см. т. 1, гл. 2.8). Естественно, аналогичные программы можно писать и на Си; больше того, наши возможности при этом не ограничиваются интерфейсом, созданным во времена MS-DOS для работы в текстовом режиме персональных компьютеров той эпохи.

Пожалуй, самый простой и в то же время надёжный подход к созданию полноэкранных программ для терминала заключается в использовании библиотеки `ncurses`, которая берёт на себя заботу об особенностях низкоуровневой «кухни», связанной с многообразием терминалов и их эмуляторов. Программа, написанная с использованием `ncurses`, будет корректно работать практически на любом терминале — как в окне вашего `xterm`'а, так и, например, на терминале DEC VT52 1974 года производства, если вы такой, конечно, найдёте — несмотря на то, что для VT52 требуются совершенно иные управляющие последовательности. В современных условиях, когда алфавитно-цифровые терминалы больше не выпускаются, может сложиться обманчивое впечатление, что подобная универсальность (основанная на специальной базе данных, описывающей особенности разных терминалов) более не актуальна, но это не так. Кроме виртуального терминала, эмулируемого в окошке `xterm`'а, ваша программа может столкнуться с текстовой консолью Linux, с той же текстовой консолью для FreeBSD, с запущенным на Windows клиентом удалённого доступа (например, Putty), с эмулятором терминала для последовательного порта (самый популярный из них — Minicom, но есть и другие), с различными программами из менее популярных версий Unix (таких как AIX или Solaris). Все эти эмуляторы терминалов хотя и похожи друг на друга, но всё-таки различаются.

4.11.1. Простой пример

Мы начнём с того же примера, с которого начали при изучении паскалевского модуля `crt` — выведем фразу «`Hello, world!`» в центр пустого терминального окна, подождём пять секунд и завершимся. При этом мы сразу же заметим, что наши возможности оказались шире, чем при использовании `crt`: программа после завершения полностью восстановит состояние терминала, вернув на место весь текст, который там был перед её запуском.

Надо сказать, что такое восстановление возможно не на любом терминале. Так, если мы заставим нашу программу поверить, что она работает на классическом DEC vt100, текст на экране она уже не восстановит; чтобы проверить это, достаточно перед запуском нашей программы дать команду «`export TERM=vt100`»; различия в поведении программы будут видны невооружённым глазом. Для программ, использующих управление цветом текста, разница будет ещё заметнее: vt100 был чёрно-белым, так что раскрашивать текст `ncurses` тоже откажется. При этом все возможности, которые терминал vt100 поддерживал, работают и в `xterm`'е, поэтому программы в целом продолжают корректно функционировать, несмотря на то, что тип терминала им сообщён заведомо неправильный.

Для нашей простейшей программы потребуется, во-первых, умение *инициализировать* библиотеку `ncurses`, то есть сообщать ей, что пора принять на себя управление терминалом. Это делается вызовом функции `initscr` без параметров. Дальше нам потребуется узнать, сколько строк имеет наше окно терминала и сколько знакомест содержится в каждой строке. Для этого мы опишем две целочисленные переменные `row` и `col` и напишем такую строку:

```
getmaxyx(stdscr, row, col);
```

Поясним, что слово `stdscr` (*standard screen*) означает *окно* на весь терминал; дело в том, что библиотека `ncurses` позволяет объявлять отдельные «окна» в разных местах терминала и вывод в каждое такое окно выполнять независимо от других окон, очищать окно, производить в нём скроллинг текста; это бывает удобно при построении сложных пользовательских интерфейсов. В нашей простой программе работа с окнами не нужна, так что мы ограничимся окном `stdscr`. Ещё одна особенность `ncurses` состоит в том, что при работе с экранными позициями сначала всегда указывается координата по вертикали (номер строки), а координата по горизонтали (номер столбца) идёт второй; буквы `yx` в названиях макросов и функций помогут об этом не забыть.

Внимательный читатель может удивиться, почему перед именами переменных мы не поставили операцию взятия адреса, ведь `getmaxyx`, очевидно, должна *записать* значения в эти переменные. Дело в том, что `getmaxyx` — это на самом деле не функция, а *макрос*, который разворачивается во фрагмент текста, содержащий обычные присваивания. Возможно, создатели `ncurses` (точнее, создатели её предшественницы `curses`, с которой поддерживается совместимость) избрали здесь далеко не лучшее решение: строка действительно выглядит странно и требует пространного пояснения. Впрочем, у такого решения есть и несомненное достоинство: переменные `row` и `col` могут иметь произвольный целочисленный тип, лишь бы его разрядности хватило для хранения соответствующих чисел.

Для перемещения курсора в нужную позицию мы воспользуемся функцией `move`, которая получает два параметра: номер строки и номер столбца. В отличие от паскалевского `GotoXY`, здесь, как и везде в `ncurses`, первой указывается координата по вертикали; кроме того, стоит запомнить, что координаты отсчитываются с нуля, а не с единицы, то есть координаты верхнего левого угла экрана задаются парой $(0, 0)$. Координаты для выдачи сообщения мы вычислим точно так же, как когда-то делали это для аналогичной паскалевской программы: вертикальный размер экрана просто поделим пополам, а из горизонтального вычтем длину нашего сообщения и пополам поделим уже то, что останется.

В отличие от Паскаля, для вывода на экран в наших полноэкранных программах придётся воспользоваться специально предназначенными для этого функциями, предоставленными библиотекой `ncurses`; стандартные функции вроде `printf` или `putchar` здесь не подойдут. Кроме того, вывод на экран в программах на основе `ncurses` имеет одну особенность, которая поначалу кажется странной и неудобной, но к которой можно довольно быстро привыкнуть: **операции вывода не оказывают никакого влияния на реальное содержимое экрана до тех пор, пока мы не потребуем привести экран в актуальное состояние**.

Простейшая функция, которая выводит на экран заданную строку в текущей позиции курсора, называется `addstr` и принимает один параметр — собственно выводимую строку; эту функцию мы и используем. Аналогичная функция для вывода одного символа называется `addch`, и её мы тоже будем использовать в последующих примерах. Проделав операцию вывода, мы спрячем курсор, чтобы он не отвлекал внимание от выданного сообщения; в программе на Паскале мы делали это, перемещая курсор в верхний левый угол, но `ncurses` позволяет сделать правильнее: вызвав `curs_set(0)`, мы вообще уберём курсор с экрана. Что касается приведения экрана в соответствие с нашим выводом, то это делается с помощью функции `refresh`, которая вызывается без параметров.

Закончив с выводом, мы подождём пять секунд. Надо сказать, что функция, позволяющая сделать такую пятисекундную задержку, не имеет никакого отношения к `ncurses` (в отличие от Паскаля, где процедура `delay` предоставляется модулем `crt`); мы воспользуемся стандартной функцией `sleep`, описанной в заголовочном файле `unistd.h`. После этого останется только завершить работу с `ncurses`, вызвав функцию `endwin` (тоже без параметров) и

на этом закончить нашу программу. Полностью текст программы получается такой:

```
#include <curses.h>                          /* curses_hello.c */
#include <unistd.h>

const char message[] = "Hello, world!";
enum { delay_duration = 5 };

int main()
{
    int row, col;
    initscr();
    getmaxyx(stdscr, row, col);
    move(row/2, (col-(sizeof(message)-1))/2);
    addstr(message);
    curs_set(0);
    refresh();
    sleep(delay_duration);
    endwin();
    return 0;
}
```

На всякий случай напомним, что `ncurses` — это самостоятельная библиотека, её функции в число «стандартных» уж точно не входят, так что для сборки нашей программы потребуется указать флаг «`-lcurses`»:

```
gcc -Wall -g curses_hello.c -lcurses -o curses_hello
```

У этой программы есть одно важное отличие от её паскалевской предшественницы: если мы не хотим ждать пять секунд, мы можем выполнение этой программы прекратить так же, как прекращали выполнение других программ — нажатием `Ctrl-C`. В принципе `ncurses` умеет действовать так же, как модуль `crt` — перепрограммировать терминал, чтобы он обрабатывал всевозможные комбинации клавиш как обычные символы, но об этом мы должны библиотеку попросить сами. Как это делается, мы расскажем в следующем параграфе.

4.11.2. Обработка клавиатурных и других событий

В обычных условиях терминал работает в так называемом ***каноническом режиме ввода***, при котором активная программа получает введённые пользователем символы только после нажатия Enter (сразу всю строку); некоторые события, поступающие от клавиатуры, драйвер терминала обрабатывает сам — например, нажатие Backspace приводит к удалению последнего введённого символа, то есть активная программа в итоге не получает ни этот удалённый символ, ни спецсимвол, сгенерированный клавишей Backspace; комбинации `Ctrl-C`, `Ctrl-D`, `Ctrl-Z` и некоторые другие имеют специальный смысл. Вводимые символы драйвер терминала отображает на экране.

Как правило, в полноэкранных интерактивных программах такой режим ввода неудобен. Мы видели, что паскалевский модуль `crt` перенастраивает драйвер терминала: активная программа начинает получать информацию о нажатых клавишах сразу после их нажатия и может обрабатывать не только алфавитно-цифровые клавиши, но и всевозможные стрелочки, функциональные клавиши и т. п., комбинации клавиш разом утрачивают особый смысл (так что мы не можем прервать зациклившуюся программу с помощью `Ctrl-C`), вводимые символы на экране не отображаются, если только программа сама их не выведет.

Библиотека `ncurses` тоже предоставляет аналогичные возможности, но поскольку, в отличие от модуля `crt`, она изначально была ориентирована на управление терминалами в ОС Unix, её подход к установке режима терминала существенно более гибок: отдельные особенности режима работы терминала могут включаться и выключаться независимо друг от друга, так что мы можем, например, выключить «канонический» режим и начать обрабатывать нажатия на клавиши немедленно, не дожидаясь конца строки, но при этом сохранить в действии комбинации `Ctrl-C`, `Ctrl-D` и прочие (хотя при желании можем выключить и их; тогда, например, нажатие `Ctrl-C` приведёт к получению нашей программой псевдосимвола с кодом `3`). Отдельно мы можем указать, нужно ли отображать на экране вводимые символы; обычно, впрочем, нам это не нужно — весь вывод лучше производить самим.

Для начала следует решить, хотим ли мы, чтобы комбинации клавиш вроде `Ctrl-C` и `Ctrl-D` продолжали работать как обычно. Если да, то следует вызвать функцию `cbreak`, в противном случае — функцию `raw`; обе вызываются без параметров. При желании мы можем вернуть терминал в исходный («канонический») режим, вызвав соответственно функцию `nocbreak` или `noraw`. Далее следует выключить автоматическое отображение вводимых символов на экране, вызвав функцию `noecho` (отменить её эффект можно с помощью функции `echo`). Наконец, в большинстве случаев следует поручить библиотеке `ncurses` обработку escape-последовательностей, генерируемых «специфическими» клавишами вроде стрелок; это делается так: `keypad(stdscr, 1)` (если указать `0` вместо `1`, режим обработки последовательностей будет выключен).

После такой настройки терминала мы можем узнавать о нажатиях клавиш на клавиатуре, а также о некоторых других происходящих событиях, вызывая функцию `getch`, которая, не принимая параметров, возвращает значение типа `int`. При нажатии на обычные «символьные» клавиши `getch` возвращает код введённого символа. Если же пользователь нажимает «специальные» клавиши, `getch` (при условии, что мы не забыли включить режим `keypad`, см. выше) возвращает значения, находящиеся за пределами диапазона кодов символов; библиотека `ncurses` предоставляет программисту символические имена для этих значений, такие как `KEY_UP`, `KEY_DOWN`, `KEY_LEFT` и `KEY_RIGHT` для «стрелочек», `KEY_F(1)`, `KEY_F(2)`, `KEY_F(10)` и т. д. для функциональных клавиш. В действительности все эти имена обозначают целые числа, например, `KEY_RIGHT` в версии библиотеки, имевшейся у автора, соответствует числу `261`, но другие версии могут предусматривать для этих целей другие численные значения, так что в программе следует, конечно же, использовать имена, в том числе и из соображений наглядности.

При нажатии на Enter и Backspace — клавиши, традиционно генерирующие управляющие псевдосимволы — функция `getch` может повести себя по-разному в зависимости от реализации: для Enter может быть возвращено как значение `10` (оно же `'\n'`, то есть код символа перевода строки), так и значение `KEY_ENTER` (в версии, имеющейся у автора, этим именем обозначено число `343`); для Backspace — значение `8` (`'\b'`) или `KEY_BACKSPACE`. Полезно знать, что клавиша Insert обозначается `KEY_IC`, клавиши PgUp и PgDn — соответственно `KEY_PPAGE` и `KEY_NPAGE`.

Выше мы упоминали, что функция `getch` может сообщить нам не только о нажатиях на клавиши, но и о других событиях; к таким событиям относятся действия с мышью (её перемещения, нажатия и отпускания кнопок), а также *изменение размера экрана*. Обработку мыши мы рассматривать не будем (заинтересованные читатели могут это сделать сами, в Интернете много хороших текстов на эту тему), а вот обрабатывать изменения размера экрана оказывается, с одной стороны, очень просто, а с другой — весьма полезно. Каждый раз, когда пользователь меняет размер окна терминала, функция `getch` в активной программе возвращает специальное значение `KEY_RESIZE`. Получив это значение, следует воспользоваться уже знакомым нам макросом `getmaxyx`, чтобы узнать новые максимально допустимые значения для координат по вертикали и горизонтали, после чего при необходимости «перерисовать» изображение.

Для примера рассмотрим программу, очень похожую на ту, что мы писали в аналогичной ситуации на Паскале (см. т. 1, §2.8.3, пример `movehello.pas`): программа выведет в середине экрана сообщение «`Hello, World!`», которое затем можно будет перемещать по экрану «стрелочными» клавишами. Выйти из программы можно будет, нажав Escape либо «убив» её с помощью `Ctrl-C` (мы воспользуемся режимом терминала, при котором специальные комбинации клавиш работают).

Для начала мы напишем несколько вспомогательных функций. Само сообщение будет вынесено в константу в начале программы; функция `show_message` будет выдавать сообщение в указанном месте экрана (получая координаты через параметры); функция `hide_message` будет печатать в указанном месте соответствующее количество пробелов, чтобы убрать ранее выданное сообщение с экрана (для вывода отдельного пробела мы воспользуемся функцией `addch`). Функция `move_message` будет перемещать сообщение в новую позицию экрана, одновременно изменяя переменные, в которых хранятся текущие координаты; для этого ей будут передаваться адреса этих двух переменных, а также максимально допустимые значения координат и то, *на сколько* позиций следует сдвинуть сообщение по горизонтали и по вертикали.

Обработку события, связанного с изменением размера терминала, мы вынесем в функцию `handle_resize`, которая, узнав с помощью `getmaxyx` новый размер, вычислит новые максимально допустимые значения координат сообщения и при необходимости переместит сообщение так, чтобы оно по-прежнему находилось в пределах экрана.

Программа целиком будет выглядеть так:

```
#include <curses.h>                             /* movehello.c */

static const char message[] = "Hello, world!";
enum { key_escape = 27 };

static void show_message(int x, int y)
{
    move(y, x);
    addstr(message);
    refresh();
}
static void hide_message(int x, int y)
{
    int i;
    move(y, x);
    for(i = 0; i < sizeof(message)-1; i++)
        addch(' ');
    refresh();
}
static void check(int *coord, int max)
{
    if(*coord < 0)
        *coord = 0;
    else
    if(*coord > max)
        *coord = max;
}
static void
move_message(int *x, int *y, int mx, int my, int dx, int dy)
{
    hide_message(*x, *y);
    *x += dx;
    check(x, mx);
    *y += dy;
    check(y, my);
    show_message(*x, *y);
}
static void handle_resize(int *x, int *y, int *mx, int *my)
{
    int row, col;
    getmaxyx(stdscr, row, col);
    *mx = col - sizeof(message) + 1;
    *my = row - 1;
    hide_message(*x, *y);
    check(x, *mx);
    check(y, *my);
    show_message(*x, *y);
}
int main()
{
    int row, col, x, y, max_x, max_y, key;
    initscr();
    cbreak();
    keypad(stdscr, 1);
    noecho();
    curs_set(0);
    getmaxyx(stdscr, row, col);
    x = (col-(sizeof(message)-1))/2;
    y = row/2;
    max_x = col - sizeof(message) + 1;
    max_y = row - 1;
    show_message(x, y);
    while((key = getch()) != key_escape) {
        switch(key) {
        case KEY_UP:
            move_message(&x, &y, max_x, max_y, 0, -1);
            break;
        case KEY_DOWN:
            move_message(&x, &y, max_x, max_y, 0, 1);
            break;
        case KEY_LEFT:
            move_message(&x, &y, max_x, max_y, -1, 0);
            break;
        case KEY_RIGHT:
            move_message(&x, &y, max_x, max_y, 1, 0);
            break;
        case KEY_RESIZE:
            handle_resize(&x, &y, &max_x, &max_y);
            break;
        }
    }
    endwin();
    return 0;
}
```

Существуют и другие удобные функции для вывода. Одна из них называется `printw`; она принимает такие же аргументы, как привычная нам `printf`, и используется для *форматированного вывода*. Например, если в вышеприведённой программе в функцию `show_message` перед вызовом `refresh` добавить следующие две строки:

```
    move(0, 0);
    printw("(%d,%d)    ", x, y);
```

— то в левом верхнем углу экрана программа будет печатать текущие координаты сообщения. Обратите внимание на четыре пробела в конце нашей форматной строки — они добавлены, чтобы с гарантией затереть всё, что осталось от предыдущих координат (например, на случай, если координаты только что выражались трёхзначными числами, а теперь стали однозначными). Такого же эффекта можно достичь и в одну строку с помощью функции `mvprintw`:

```
    mvprintw(0, 0, "(%d,%d)    ", x, y);
```

(нужно только не забывать, что координаты для вывода указываются в последовательности «`y`, `x`», то есть сначала вертикальная, потом горизонтальная).

4.11.3. Управление цветом и атрибутами символов

Изучая Паскаль, мы узнали, что при выводе текста на экран можно с каждым символом связать те или иные *атрибуты*, влияющие на его внешний вид. Библиотека `ncurses` позволяет управлять атрибутами выводимых символов, в том числе цветом самого символа и его фона, если только терминал, с которым мы работаем, цветной[42]. Сразу же отметим, что возможности, которые мы рассмотрим в этом параграфе, нужно сначала *включить*, вызвав функцию `start_color`; обычно это делают сразу после вызова `initscr`. Функция не принимает параметров и не возвращает значения.

Некоторые атрибуты не зависят от «цветности» терминала. С их помощью символ можно сделать жирным, инверсным (когда цвета фона и символа меняются ролями), подчёркнутым, мигающим и т. п. Каждый такой атрибут обозначается своим идентификатором: например, `A_BLINK` означает «мигающий», `A_UNDERLINE` — «подчёркнутый», `A_BOLD` — «жирный», `A_DIM` — «тусклый», `A_REVERSE` — «инверсный» и т. д., полный список можно посмотреть, например, в тексте страницы системного справочника, которая называется `attr` (для этого дайте команду `man attr`). Каждый такой атрибут — это отдельный бит, так что их можно сочетать, используя операцию побитового «или». Пожалуй, проще всего управлять атрибутами символов с помощью функций `attron` и `attroff`; обе принимают по одному параметру, первая *включает* указанные атрибуты, вторая их, наоборот, *выключает*. Например, фрагмент

```
    attron(A_BOLD);
    addstr("Hello, ");
    attron(A_UNDERLINE);
    addstr("wonderful");
    attroff(A_BOLD|A_UNDERLINE);
    addstr(" world!");
```

[42] Конечно, все эмуляторы терминалов цветные, но пользователь по каким-то причинам может захотеть, чтобы запускаемые программы не использовали цвет; для этого достаточно установить соответствующий тип терминала, см. замечание на стр. 82. Кроме того, в частных коллекциях попадаются работающие (и даже используемые) настоящие терминалы, выпущенные в 1980-е, а иногда и в 1970-е годы. Интересы владельцев таких терминалов можно, конечно, проигнорировать, но нужно ли?

Таблица 4.3. Обозначения цветов в библиотеке `ncurses`

`COLOR_BLACK`	чёрный	`COLOR_BLUE`	синий
`COLOR_RED`	красный	`COLOR_MAGENTA`	фиолетовый
`COLOR_GREEN`	зелёный	`COLOR_CYAN`	голубой
`COLOR_YELLOW`	жёлтый	`COLOR_WHITE`	белый

напечатает фразу «`Hello, wonderful world!`», причём первые два слова будут напечатаны жирными, второе к тому же подчёркнуто, а третье будет напечатано обычным — ни жирности, ни подчёркивания. Есть и другие способы работы с атрибутами текста; функций, предназначенных для этого, `ncurses` предоставляет больше двух десятков, но рассматривать их мы для экономии места не будем. Отметим только, что при выводе по одному символу с помощью `addchar` атрибуты можно «присоединить» к коду символа с помощью всё той же операции побитового «или». Например, выдать мигающую жирную звёздочку можно так:

```
addch('*' | A_BOLD | A_BLINK);
```

Рассмотрим теперь возможности по управлению цветом. Прежде чем пытаться с ним работать, следует проверить, возможно ли это на имеющемся терминале. Это делается вызовом функции `has_colors`; если она вернула «ложь» (то есть `0`), то цвета нам недоступны; что делать в этом случае — зависит от решаемой задачи, но лучше всего написать программу так, чтобы на чёрно-белом терминале она сохраняла работоспособность, не делая попыток задействовать цвет. Например, мы можем завести переменную `work_bw` (`bw` — это обычное сокращение для слов *black&white*, обозначающих чёрно-белое, будь то фото, изображение или режим работы) и во всех наших функциях, осуществляющих вывод, прибегать к управлению цветом лишь в том случае, если эта переменная хранит значение «ложь», в противном случае обходиться атрибутами символов, не зависящими от цветности. В начале нашей главной функции следует предусмотреть примерно такой фрагмент:

```
initscr();
work_bw = !has_colors();
if(!work_bw)
    start_color();
```

Заметим, мы можем при необходимости занести в переменную `work_bw` значение «истины», даже если терминал у нас цветной, но пользователь каким-то образом (например, через параметры командной строки или как-то ещё) потребовал от нас работать в чёрно-белом режиме; это вполне обычная ситуация, ведь при этом гораздо меньше устают глаза.

В дальнейших примерах мы для краткости предполагаем, что `has_colors` вернула «истину» и работать в чёрно-белом режиме нас никто не заставляет. Наша цель сейчас — показать, как *работать* с цветом, а не как *обходиться* без него.

Библиотека `ncurses` поддерживает восемь основных цветов, перечисленных в табл. 4.3. Спецификация интерфейса библиотеки предусматривает средства для создания других цветов, но в реальной жизни эти средства никогда не работают — соответствующие возможности не поддерживаются ни терминалами, ни их эмуляторами, а библиотечные функции, предназначенные для этого, просто не реализованы (всегда возвращают ошибку).

Цвета символов в `ncurses` всегда устанавливаются так называемыми *цветовыми парами*; пара состоит из цвета символа и цвета фона. С цветовыми парами связано довольно много путаницы. Сами пары различаются по номерам, причём пара номер `0` соответствует цветам «по умолчанию» и не должна изменяться. Чтобы вывести на экран цветную надпись, нужно выбрать номер цветовой пары, с помощью функции `init_pair` назначить этой паре цвет текста и цвет фона, после чего использовать в качестве атрибута выводимого текста выражение `COLOR_PAIR(n)`, где `n` — номер цветовой пары. Функция `init_pair` принимает три параметра: номер пары, номер цвета для текста и номер цвета для фона. Например,

```
init_pair(1, COLOR_WHITE, COLOR_BLUE);
attrset(COLOR_PAIR(1) | A_UNDERLINE);
addstr("White on blue");
refresh();
```

выведет надпись «`White on blue`» белыми буквами по синему фону, используя для этого цветовую пару №1, причём надпись будет подчёркнута. Если вы будете применять несколько разных сочетаний цветов для символа и его фона (а так, скорее всего, и будет), то для каждого сочетания нужно будет использовать свой собственный номер цветовой пары. Использовать один и тот же номер для разных сочетаний не получится: **если изменить цвета для существующей цветовой пары, то на экране соответствующим образом изменятся цвета всех символов, выданных с использованием этого номера цветовой пары**. Кстати, это свойство само по себе можно использовать для создания интересных визуальных эффектов.

Заслуживает внимания вопрос, *какие* числа можно использовать в качестве номеров цветовых пар. Как ни странно, эту тему документация и всевозможные руководства тщательно обходят стороной. Можно определённо сказать, что значения номеров цветовых пар, превышающие `255`, создадут проблемы, обусловленные распределением отдельных разрядов в атрибуте символа. Проблемы могут возникнуть и с мéньшими номерами, причём вопрос о максимально допустимом номере цветовой пары, судя по всему, зависит от реализации библиотеки. В некоторых описаниях упоминается константа `COLOR_PAIRS` (хотя при этом не говорится, чему она должна быть равна), но в наиболее популярной реализации `ncurses` этот идентификатор представляет собой не константу, а переменную.

Вообще-то ясно, что при наличии всего восьми различных цветов можно обойтись не более чем 64 цветовыми парами. Конечно, если бы синтез дополнительных цветов работал, это соображение ничем бы помочь не могло — но он не работает, так что каким-то ещё цветам попросту неоткуда взяться. Поскольку пару с нулевым номером трогать нельзя, нам могут потребоваться для представления всех возможных цветовых комбинаций номера от `1` до `64`; как показывает практика, в существующих реализациях `ncurses` использование всех этих номеров проблем не создаёт.

Можно себе представить программу, в которой некоторые цветовые пары повторяются, если они исходно соответствуют не цветовым сочетаниям, а, например, элементам визуального интерфейса, чтобы можно было синхронно «перекрасить» элементы одного типа; впрочем, даже в этом случае трудно себе представить больше 64 цветовых пар в одной программе.

Отметим ещё один момент. Добавление атрибута `A_BOLD` фактически изменяет цвет символа; это как раз та причина, по которой паскалевский модуль `crt` вводит восемь цветов для фона и целых шестнадцать — для самих символов.

Для демонстрации возможностей управления цветом мы напишем программу, подобную той, что писали для той же цели на Паскале (см. т. 1, §2.8.4): она тоже будет выводить на экран звёздочки всеми возможными цветами. Номера используемых цветовых пар мы жёстко привяжем к составляющим их цветам по формуле $8 \cdot b + f + 1$, где b — номер цвета для фона, f — для символа. Единицу мы прибавляем, чтобы не использовать «запрещённую» цветовую пару №0.

Константы `COLOR_BLACK`, `COLOR_RED` и т. д. равны числам от `0` до `7` включительно, но поскольку этот момент нигде в документации не зафиксирован, мы не будем это использовать; вместо этого опишем массив из восьми элементов и занесём в его элементы значения для разных цветов; идентифицировать цвета будем индексами этого массива, что гарантирует нам использование именно номеров от 0 до 7, даже если в какой-нибудь реализации библиотеки цветовые константы будут равны каким-нибудь другим числам.

Привязку номеров цветовых пар к номерам самих цветов будет обеспечивать функция `setpair`, которая, получив номера цветов для символа и фона, будет вычислять по вышеприведённой формуле нужный номер цветовой пары, устанавливать для этой пары соответствующие цвета, а сам номер пары возвращать в качестве значения.

За вывод одной строки будет отвечать функция `make_line`; символы в каждой строке будут печататься одним и тем же цветом (самого символа) поверх различных цветов фона, причём каждый второй символ получит дополнительно атрибут `A_BOLD`, а каждая вторая пара символов будет сделана ещё и мигающей. Всю картинку на экране будет формировать из отдельных строк функция `make_screen`.

Когда картинка будет сформирована, в главной функции программы мы запустим цикл чтения с клавиатуры, который будет прекращаться при получении символа `Escape` (код `27`), а при нажатии любых других клавиш из тела цикла будет вызываться функция `shift_pairs`, которая для всех задействованных 64 цветовых пар будет изменять значения цветов, что позволит наблюдать интересный визуальный эффект.

Полностью текст программы получается таким:

```
#include <stdio.h>                          /* curses_col.c */
#include <curses.h>

enum { key_escape = 27 };
enum { color_count = 8 };
static const int all_colors[color_count] = {
    COLOR_BLACK, COLOR_RED, COLOR_GREEN, COLOR_YELLOW,
    COLOR_BLUE, COLOR_MAGENTA, COLOR_CYAN, COLOR_WHITE
};

static int setpair(int fg, int bg)
{
    int n = bg*8 + fg + 1;
    init_pair(n, fg, bg);
    return n;
}
static void make_line(int line, int width, int fgcolor)
{
    int i, j, w, pn, att;
    w = width / color_count;
    for(i = 0; i < color_count; i++) {
        move(line, i*w);
        pn = setpair(fgcolor, all_colors[i]);
        for(j = 0; j < w; j++) {
            att = COLOR_PAIR(pn);
            if(j % 2 == 0)
                att |= A_BOLD;
            if((j / 2) % 2 == 1)
                att |= A_BLINK;
            attrset(att);
            addch('*');
            refresh();
        }
    }
}
static void make_screen(int h, int w)
{
    int i;
    clear();
    for(i = 0; i < h; i++)
        make_line(i, w, all_colors[i % color_count]);
}
static void shift_pairs(int shift)
{
    int i;
    for(i = 1; i <= color_count * color_count; i++) {
        int fg = (i + shift) % color_count;
        int bg = ((i + shift) / color_count) % color_count;
        init_pair(i, all_colors[fg], all_colors[bg]);
    }
}
int main()
{
    int row, col, ch, shift;
    initscr();
    if(!has_colors()) {
        endwin();
        fprintf(stderr, "I can't show colors on a BW screen\n");
        return 1;
    }
    cbreak();
    keypad(stdscr, 1);
    noecho();
    curs_set(0);
    start_color();
    getmaxyx(stdscr, row, col);
    make_screen(row, col);
    shift = 2;
    while((ch = getch()) != key_escape) {
        shift_pairs(shift);
        shift++;
        refresh();
    }
    endwin();
    return 0;
}
```

4.11.4. Клавиатурный ввод с тайм-аутами

До сих пор в наших примерах функция `getch` ждала, пока пользователь не нажмёт какую-нибудь клавишу, то есть в ожидании очередного события программа ничего не делала; но как быть, если мы хотим всё время выполнять какие-то действия, позволяя пользователю вмешаться в происходящее путём нажатия клавиш, но не тратя время на ожидание? В программах на Паскале мы решали эту проблему с помощью функции `keypressed`, вызывая её время от времени и выполняя чтение символа с клавиатуры лишь если `keypressed` вернула «истину». В `ncurses` аналога для `keypressed` нет, зато есть возможность гибкой настройки режима ввода. По умолчанию, как мы видели, функция `getch` ждёт наступления очередного события, если оно ещё не произошло на момент её вызова; но режим работы можно изменить так, чтобы `getch` всегда немедленно возвращала управление, причём если никаких событий не произошло, возвращаемое значение будет равно константе `ERR`. Более того, можно задать для `getch` ***тайм-аут*** (в миллисекундах); в этом случае функция будет ожидать наступления события, но не более чем в течение заданного тайм-аутом временно́го периода, после чего, если никакое событие так и не произойдёт, благополучно вернёт `ERR`.

Управлять режимом тайм-аута можно с помощью функции, которая так и называется `timeout`; функция принимает один целочисленный параметр, означающий величину тайм-аута в миллисекундах, то есть, например, значение параметра `100` предписывает функции `getch` ждать наступления события не более чем в течение 0,1 секунды, и если ничего за это время не произойдёт, вернуть значение `ERR`. Ноль в качестве аргумента для `timeout` означает, что ждать не следует вовсе: если к моменту вызова `getch` никаких событий не произошло, она при этом вернёт `ERR` немедленно (такой режим называется неблокирующим). Любое отрицательное значение аргумента `timeout` (например, `-1`) установит обычный блокирующий режим, при котором `getch` ждёт наступления события без ограничения по времени.

Совместную работу функций `timeout` и `getch` мы продемонстрируем на примере программы, которая похожа на программу `MovingStar`, написанную нами на Паскале (см. т. 1, §2.8.3): на экран выводится неподвижная звёздочка, которую можно заставить двигаться по экрану, нажав одну из «стрелочных» клавиш. Звёздочка при этом начнёт перемещаться, не дожидаясь нажатий на другие клавиши, но её можно будет в любой момент запустить в другом направлении, нажав соответствующую «стрелку», или остановить, нажав пробел.

Для удобства всю информацию о текущем состоянии звёздочки (текущие координаты и вектор скорости) мы соберём в одну структурную переменную, для работы с ней напишем несколько вспомогательных функций. В ходе инициализации мы вызовем `timeout` с константой, имеющей значение `100`, в качестве аргумента, так что `getch`, используемая нами в главном цикле программы, будет возвращать обычные коды клавиш при их нажатии, а при отсутствии событий — значение `ERR`, которое можно рассматривать как признак того, что прошли очередные `100` миллисекунд и звёздочку пора перемещать в следующую позицию. Полностью текст программы выглядит так:

```
#include <curses.h>                 /* movingstar.c */

enum { delay_duration = 100 };
enum { key_escape = 27 };
struct star {
    int cur_x, cur_y, dx, dy;
};
static void show_star(const struct star *s)
{
    move(s->cur_y, s->cur_x);
    addch('*');
    refresh();
}
static void hide_star(const struct star *s)
{
    move(s->cur_y, s->cur_x);
    addch(' ');
    refresh();
}
static void check(int *coord, int max)
{
    if(*coord < 0)
        *coord += max;
    else
    if(*coord > max)
        *coord -= max;
}
static void move_star(struct star *s, int max_x, int max_y)
{
    hide_star(s);
```

```
    s->cur_x += s->dx;
    check(&s->cur_x, max_x);
    s->cur_y += s->dy;
    check(&s->cur_y, max_y);
    show_star(s);
}
static void set_direction(struct star *s, int dx, int dy)
{
    s->dx = dx;
    s->dy = dy;
}
static void handle_resize(struct star *s, int *col, int *row)
{
    getmaxyx(stdscr, *row, *col);
    if(s->cur_x > *col)
        s->cur_x = *col;
    if(s->cur_y > *row)
        s->cur_y = *row;
}
int main()
{
    int row, col, key;
    struct star s;
    initscr();
    cbreak();
    timeout(delay_duration);
    keypad(stdscr, 1);
    noecho();
    curs_set(0);
    getmaxyx(stdscr, row, col);
    s.cur_x = col/2;
    s.cur_y = row/2;
    set_direction(&s, 0, 0);
    while((key = getch()) != key_escape) {
        switch(key) {
        case ' ':
            set_direction(&s, 0, 0);
            break;
        case KEY_UP:
            set_direction(&s, 0, -1);
            break;
        case KEY_DOWN:
            set_direction(&s, 0, 1);
            break;
        case KEY_LEFT:
            set_direction(&s, -1, 0);
            break;
        case KEY_RIGHT:
            set_direction(&s, 1, 0);
            break;
        case ERR:
            move_star(&s, col-1, row-1);
            break;
        case KEY_RESIZE:
            handle_resize(&s, &col, &row);
        }
    }
    endwin();
    return 0;
}
```

4.11.5. Обзор остальных возможностей `ncurses`

Как обычно в таких случаях, мы не рассмотрели и десятой доли возможностей библиотеки `ncurses`; их полное изучение потребовало бы отдельной книги. Наша цель состояла в том, чтобы продемонстрировать основные возможности и более-менее осознать концепцию в целом. Так, мы рассмотрели только несколько функций из двух десятков, предназначенных для управления атрибутами выводимого текста; мы не стали рассматривать возможности по созданию так называемых *окон* — прямоугольных областей на экране, позволяющих выводить текст с использованием своих координат в каждом окне; мы оставили за рамками нашего текста бóльшую часть функций инициализации и настройки режима терминала, совсем не упомянули возможности скроллинга (как всего экрана, так и отдельных его частей).

Кроме прочего, вместе с библиотекой `ncurses` распространяются дополнительные библиотеки: `menu`, `panel` и `cdk`. Первая из них, `menu`, позволяет составлять и выдавать на экран текстовые меню, в которых в зависимости от настроек пользователь может выбрать один или несколько пунктов. Библиотека `panel` предназначена для работы с прямоугольными окнами, которые могут частично перекрываться, накладываться одно на другое, окна можно перемещать по экрану, убирать и снова показывать и т. д. Наконец, `cdk` (*curses development kit*) предоставляет целый ряд ***виджетов*** — готовых элементов для построения диалоговых окон: поля ввода текста, кнопки, переключатели и тому подобное.

4.12. (*) Программа на Си без стандартной библиотеки

В процессе изучения языка Си мы неоднократно повторяли, как мантру, фразу «Си мы любим не за это». Теперь, когда практически все возможности этого языка нам известны, настала пора понять, *за что же* мы его всё-таки любим. Эту небольшую главу мы посвятим эксперименту, который наглядно продемонстрирует одно уникальное свойство языка Си, делающее его во многих случаях незаменимым.

Начнём с того, что напишем простенькую программу, которая принимает ровно один параметр командной строки, рассматривает его как имя и здоровается с человеком, чьё имя указано, фразой `Hello, dear NNN!` (имя подставляется вместо `NNN`). Для начала попробуем написать эту программу самым очевидным способом:

```
#include <stdio.h>                    /* no_libc/greet.c */

int main(int argc, char **argv)
{
    if(argc < 2) {
        printf("I don't know how to greet you\n");
        return 1;
    }
    printf("Hello, dear %s!\n", argv[1]);
    return 0;
}
```

Откомпилировав эту программу, мы получим сравнительно небольшой исполняемый файл; в системе, имевшейся в распоряжении автора этих строк, полученный «бинарник» занимал около 8,5 KB при включённой отладочной информации и меньше семи KB — при её отключении, то есть при компиляции без флага `-g`. Однако тут имеет место некоторое жульничество: в этот исполняемый файл не включён код функции `printf` и всего, что нужно для её использования (она сама вызывает много других библиотечных функций); вместо этого компилятор построил нам так называемый *динамически загружаемый исполняемый файл*, что означает, что в процессе выполнения в память сначала будет загружен сам этот файл, потом код, помещённый в него, потребует загрузки в память динамически загружаемой версии стандартной библиотеки языка Си (если интересно, это обычно файл `/lib/libc.so.X`, где `X` — номер версии библиотеки). Недостаток такого подхода легко обнаружить, если запустить нашу программу под управлением утилиты `strace`, позволяющей увидеть список всех выполненных программой ***системных вызовов***; нас, впрочем, будет интересовать только их количество — чтобы узнать его, достаточно посчитать строки в полученном файле: количество системных вызовов будет на два меньше.

```
avst@host:~$ strace -o LOG ./greet Andrey
Hello, dear Andrey!
avst@host:~$ wc -l LOG
29 LOG
avst@host:~$
```

Итак, наша программа сделала 27 системных вызовов, причём полезны из них только два последних — `write`, который напечатал строчку, и `exit_group`, используемый современными версиями стандартной библиотеки вместо традиционного `_exit` для завершения программы. Все остальные 25 вызовов — подготовительная работа, которая в основном состоит как раз в загрузке динамической библиотеки.

Попробуем запретить компилятору жульничать — потребуем от него создать статически собранный исполняемый файл, не зависящий ни от каких динамических библиотек. Это достигается ключом `-static`:

```
gcc -Wall -g -static greet.c -o greet
```

Посмотрев на размер полученного исполняемого файла, мы сразу, как говорится, *почувствуем разницу*: размер «бинарника» у нас вырос раз этак в семьдесят; у автора на его системе получилось около 565 KB при включённой отладочной информации и около 510 KB при выключенной. С другой стороны, снова сосчитав системные вызовы, мы обнаружим, что вместо 27 их стало всего 10 — что, впрочем, тоже не так мало.

Поскольку в наши планы не входит реализация функции `printf`, попробуем обойтись без неё, обращаясь напрямую к операционной системе, для чего нам потребуется системный вызов `write`. С этим вызовом мы уже знакомы по программам на языке ассемблера (см. т. 1, § 3.6.5); в программах на языке Си можно воспользоваться «обёрткой» этого вызова, которая выглядит как обычная функция. Напомним, что вызов принимает три параметра, которые мы для обращения к ядру Linux раскладывали по регистрам, а для обращения к ядру FreeBSD — помещали в стек; эти параметры — номер дескриптора потока, адрес области памяти, где находятся данные для записи, и количество данных в байтах. Обёртка вызова `write` вполне естественно оказывается имеющей следующий профиль:

```
int write(int fd, const void *data, int len);
```

Функция объявлена в заголовочном файле `unistd.h`.

Сразу же отметим, что, отказавшись от `printf` в пользу вызова `write`, на большинстве систем (хотя и не на всех) мы не достигнем совершенно никакой экономии в размере исполняемого файла, а системных вызовов можем сэкономить «целых» два (опять же, результаты на вашей системе могут отличаться), и то при условии, если уложимся в один вызов `write`; но для этого понадобилось бы писать функцию конкатенации строк и множество всяческих проверок на переполнение буфера, что существенно загромоздило бы наш пример. Поэтому мы поступим проще: выводить надпись будем в три приёма, то есть в три обращения к ядру; всю экономию это, увы, сведёт на нет, вызовов станет даже на один больше. Выглядеть программа будет так:

```
#include <unistd.h>                     /* no_libc/greet2.c */

static const char dunno[] = "I don't know how to greet you\n";
static const char hello[] = "Hello, dear ";

static int string_length(const char *s)
{
    int i = 0;
    while(s[i])
        i++;
    return i;
}

int main(int argc, char **argv)
{
    if(argc < 2) {
        write(1, dunno, sizeof(dunno)-1);
        return 1;
    }
    write(1, hello, sizeof(hello)-1);
    write(1, argv[1], string_length(argv[1]));
    write(1, "!\n", 2);
    return 0;
}
```

Попробовав откомпилировать эту программу и с разочарованием убедившись, что мы ничего толком не добились, приступим собственно к тому, ради чего всё затеяли — попытаемся избавиться от использования стандартной библиотеки Си. Для этого нам для начала потребуется заменить ту её часть, в которой реализована небезызвестная «невидимая функция» (а точнее, подпрограмма в ассемблерном смысле) `_start`, хорошо знакомая нам со времён изучения языка ассемблера. Именно эта подпрограмма служит *настоящей* точкой входа в программу, то есть операционная система *на самом деле* вызывает её, а вовсе не `main`; как можно легко догадаться, функцию `main` вызывает как раз подпрограмма `_start`, именно на неё возлагаются обязанности по подготовке аргументов `main`, а равно и вызов `_exit` (или что там используется вместо него) после того, как `main` завершит свою работу.

Припомнив, что и как расположено в стеке на момент вызова `_start`, мы легко убедимся, что для функции `main` аргументы придётся готовить, так как то, что в стеке уже есть, хотя и представляет собой именно ту информацию, которая нам нужна, но *не совсем в том виде*; напомним, что по адресу `[ESP]` лежит количество параметров командной строки, то есть прямо-таки значение `argc`, но вот дальше лежат указатели на сами параметры, то есть фактически прямо в стеке лежит массив `argv`, но нам ведь нужно, чтобы в стеке лежал адрес этого массива (параметр `argv`). Поэтому мы поступим проще: загрузим из стека в регистр `ECX` значение `argc`, потом в регистре `EAX` сформируем значение `argv`, затем, не трогая имеющуюся в стеке информацию, занесём туда в соответствии с конвенцией вызовов сначала `argv`, потом `argc` и, наконец, вызовем функцию `main`. После её завершения — если, конечно, до этого дело дойдёт, ведь кто-то вполне мог вызвать `_exit` без нашего участия — нам останется извлечь её возвращаемое значение из регистра `EAX` и вызвать `_exit`.

Полностью файл, содержащий подпрограмму `_start`, с учётом конвенции вызовов ОС Linux будет выглядеть так:

```
global _start                           ; no_libc/start.asm
extern main
section         .text
_start:         mov     ecx, [esp]      ; argc in ecx
                mov     eax, esp
                add     eax, 4          ; argv in eax
                push    eax
                push    ecx
                call    main
                add     esp, 8          ; clean the stack
                mov     ebx, eax        ; now call _exit
                mov     eax, 1
                int     80h
```

Для ОС FreeBSD придётся заменить последние три строчки на следующие:

```
                push    eax             ; now call _exit
                push    eax
                mov     eax, 1
                int     80h
```

Следующим этапом нашего эксперимента станет написание своей реализации «обёрток» для системных вызовов. По большому счёту, нам достаточно одного лишь вызова `write`, но, чтобы показать, как всё выглядело бы для более серьёзной программы, мы напишем обёртки для двух вызовов: `write` и `read` (его профиль выглядит так же, как и профиль `write`, только без слова `const`, ведь в эту область памяти данные *записываются*); во избежание лишней путаницы назовём наши версии обёрток `sys_write` и `sys_read`. Как и для «настоящих» обёрток системных вызовов, мы предусмотрим глобальную переменную для хранения кода ошибки, которую, опять-таки во избежание путаницы, назовём `sys_errno`.

Отметим сразу же, что реализации этих двух вызовов, как и вообще всех системных вызовов, имеющих три параметра, разрядность которых не превышает 32 бит, требуют совершенно одинаковой последовательности команд и отличаются только номером системного вызова. Поэтому мы создадим общую для всех «трёхместных» вызовов реализацию, а сами обёртки системных вызовов будут представлять собой лишь точки входа (а не целые подпрограммы), которые будут помещать нужный номер системного вызова в регистр `EAX` и делать простой (безусловный) переход на общую реализацию трёхместного вызова. Что касается этой реализации, то она будет действовать в соответствии с конвенцией CDECL: сохранять старое значение `EBP`, создавать классический стековый фрейм, уже в нём сохранять старое значение `EBX` (согласно правилам CDECL, вызванная подпрограмма имеет право испортить `EAX`, `ECX` и `EDX`, а наша реализация обёрток вынуждена также использовать `EBP` и `EBX`, поэтому эти два регистра она сохраняет в стеке и потом восстанавливает), затем, обратившись к соответствующим позициям стека, извлечь оттуда параметры системного вызова, разложить их в соответствии с конвенцией ОС Linux по регистрам `EBX`, `ECX` и `EDX`, затем, наконец, отдать управление ядру операционной системы; по возвращении выяснить (опять же, в соответствии с конвенцией Linux), не находится ли возвращённое значение в диапазоне от `0xfffff000` до `0xffffffff`, если да — занести его в `sys_errno`, а в `EAX` оставить значение `-1`, если же нет, то просто вернуть управление вызвавшему, предварительно восстановив состояние регистров и стека. Полностью наш модуль будет выглядеть так:

```
global          sys_read                ; no_libc/calls.asm
global          sys_write
global          sys_errno

section         .text

generic_syscall_3:
                push    ebp
                mov     ebp, esp
                push    ebx
                mov     ebx, [ebp+8]
                mov     ecx, [ebp+12]
                mov     edx, [ebp+16]
                int     80h
                mov     edx, eax
                and     edx, 0ffff000h
                cmp     edx, 0ffff000h
                jne     .okay
                mov     [sys_errno], eax
                mov     eax, -1
.okay:          pop     ebx
                mov     esp, ebp
                pop     ebp
                ret

sys_read:       mov     eax, 3
                jmp     generic_syscall_3
sys_write:      mov     eax, 4
                jmp     generic_syscall_3

section         .bss
sys_errno       resd    1
```

Для FreeBSD подпрограмма `generic_syscall_3` будет выглядеть намного проще, ведь параметры системного вызова передаются через стек, а в стеке они у нас уже лежат, притом в нужном порядке. Здесь мы сможем себе позволить не оформлять полноценный стековый фрейм, поскольку никаких локальных переменных использовать не будем, сохранять и восстанавливать регистры нам тоже не нужно. Выглядеть это будет так:

```
generic_syscall_3:
                int     80h
                jnc     .okay
                mov     [sys_errno], eax
                mov     eax, -1
.okay:          ret
```

Больше различий между Linux и FreeBSD у нас не предвидится, то есть весь оставшийся текст нашего примера подойдёт для обеих систем.

Вернёмся к нашей главной программе и перепишем её с расчётом на использование только что написанных обёрток вместо функции `write` из стандартной библиотеки. Для этого нам придётся убрать директиву `#include`, добавить прототип функции `sys_write` и заменить все вызовы `write` вызовами `sys_write`. Получится следующее:

```
/* no_libc/greet3.c */
int sys_write(int fd, const void *buf, int size);

static const char dunno[] = "I don't know how to greet you\n";
static const char hello[] = "Hello, dear ";

static int string_length(const char *s)
{
    int i = 0;
    while(s[i])
        i++;
    return i;
}

int main(int argc, char **argv)
{
    if(argc < 2) {
        sys_write(1, dunno, sizeof(dunno)-1);
        return 1;
    }
    sys_write(1, hello, sizeof(hello)-1);
    sys_write(1, argv[1], string_length(argv[1]));
    sys_write(1, "!\n", 2);
    return 0;
}
```

Обратите внимание, что здесь нет ни одной директивы `#include`, что выглядит несколько непривычно для текста на Си; но включения заголовочников тут действительно не нужны, ведь единственная внешняя функция, которую мы используем — это `sys_write`, её прототип мы написали прямо в тексте. Мы, впрочем, могли бы снабдить наш модуль `calls.asm` заголовочным файлом, назвав его, к примеру, `calls.h`:

```
#ifndef CALLS_H_SENTRY
#define CALLS_H_SENTRY

int sys_read(int fd, void *buf, int size);
int sys_write(int fd, const void *buf, int size);
extern int sys_errno;

#endif
```

Если так сделать, то в тексте самой программы стоит, по-видимому, убрать прототип `sys_write` и подключить заголовочник директивой

```
#include "calls.h"
```

Сборку программы произведём так:

```
nasm -f elf start.asm
nasm -f elf calls.asm
gcc -Wall -c greet3.c
ld start.o calls.o greet3.o -o greet3
```

Если вы работаете в 64-битной системе, потребуется явным образом сообщить компилятору `gcc` и редактору связей, что целью является архитектура 32-битная:

```
gcc -m32 -Wall -c greet3.c
ld -m elf_i386 start.o calls.o greet3.o -o greet3
```

Обратим внимание, что для финальной сборки мы вызвали редактор связей вручную, не используя возможности `gcc`; в самом деле, ведь мы отказались от использования библиотеки Си, так что знания компилятора относительно того, где брать эту библиотеку, нам больше не нужны.

Убедившись, что программа `greet3` работает в соответствии с поставленной задачей, мы можем посмотреть, наконец, на размер исполняемого файла, и увидеть, что он занимает *меньше одного килобайта*; на машине автора этих строк размер файла составил 816 байт, что *в семьсот с лишним раз меньше*, чем размер статического исполняемого файла, получавшийся до этого. Но дело даже не в этой экономии — в конце концов, если сравнивать результат не с размером статического «бинарника», а с размером исполняемого файла, получаемого штатным способом с использованием возможностей динамической подгрузки библиотек, разница будет меньше чем в десять раз; тоже, конечно, много, но не настолько. Намного важнее сам принцип: **язык Си позволяет полностью отказаться от возможностей стандартной библиотеки**. Кроме Си, таким свойством — абсолютной независимостью от библиотечного кода, также иногда называемым *zero runtime*[43] — обладают на сегодняшний день только языки ассемблеров; ни один язык высокого уровня не предоставляет такой возможности. В частности, суперпопулярный язык Си++ утратил независимость от библиотек времени исполнения ещё в середине 1980-х годов, когда в него был встроен механизм обработки исключений. **Чтобы обладать свойством zero runtime, язык не должен включать никаких средств, реализация которых на уровне машинного кода столь сложна, чтобы имело смысл выносить её в подпрограмму**; иначе говоря, свойство zero runtime достигается благодаря *отсутствию* в языке определённых возможностей, а не благодаря их наличию.

[43] Этим термином подчёркивается нулевой размер библиотеки времени исполнения, то есть той части библиотеки, которая обязательно присутствует во всех исполняемых файлах, создаваемых конкретным компилятором.

Именно это свойство — zero runtime — делает Си единственным и безальтернативным кандидатом на роль языка для реализации ядер операционных систем и прошивок для микроконтроллеров. Тем удивительнее, насколько мало людей в мире этот момент осознают; и стократ удивительнее то, что людей, понимающих это, судя по всему, вообще нет среди членов комитетов по стандартизации, которые, начиная ещё с конца прошлого века, упорно пытаются «срастить» Си с его стандартной библиотекой, а сам язык утяжеляют возможностями, какие в некоторых компиляторах уже реализуются подпрограммами; иначе говоря, тот бастард, который вышел из-под пера стандартизаторов, более не обладает главным, да и едва ли не единственным достоинством языка Си, зато обладает при этом всеми его недостатками в виде антилогичной семантики, высокой трудоёмкости работы и тому подобного.

К сожалению, даже с имеющимися на сегодняшний день реализациями Си есть определённые проблемы. Например, при использовании `gcc` наша программа могла бы не пройти финальную сборку, при этом компоновщик пожаловался бы на отсутствие каких-то неведомых функций, чьи имена начинаются с двух подчёркиваний. Это означало бы, что в нашей программе мы использовали некую возможность, которую компилятор реализует через вызов «своей собственной» функции. Примером такой возможности служит умножение чисел типа `long long` на 32-битной системе. К счастью, в такой ситуации требуется подключать не обычную «стандартную библиотеку», а сравнительно небольшую библиотеку, содержащую как раз такие вот «хитрые» функции. Для `gcc` эта библиотека называется `libgcc` и подключается добавлением флага `-lgcc` при вызове линкера:

```
ld start.o calls.o greet3.o -lgcc -o greet3
```

Кроме того, «современные» версии `gcc` иногда якобы «для оптимизации» заменяют некоторые фрагменты кода вызовами функций `memset`, `memcpy` и некоторых других. С этим можно в большинстве случаев справиться, добавив флажок `-ffreestanding`; ещё иногда помогает отключение оптимизации (`-O0`).

Так или иначе, язык Си допускает свойство zero runtime, но не все его реализации этим свойством обладают: как видим, `gcc` в некоторых случаях требует библиотечных функций для реализации возможностей, встроенных в язык, и даже (при включённой оптимизации) может заменить фрагмент нашего кода на вызов функции из стандартной библиотеки. Нам остаётся утешаться тем, что это свойство конкретного компилятора, а не языка в целом.

Часть 5

Объекты и услуги операционной системы

5.1. Операционная система: что это и зачем

5.1.1. Роль и место операционной системы

Операционная система неоднократно упоминалась в предыдущих частях нашей книги. Мы уже знаем, что **операционная система — это программа**, которая загружается на компьютере первой (на самом деле не совсем первой, но от тех, кто загружается раньше, к моменту начала обычной работы не остаётся никаких следов) и что все остальные программы запускаются и работают *под управлением* операционной системы. Сам по себе термин «операционная система» употребляется в литературе и лексиконе специалистов в двух существенно различных значениях; ранее в тексте нашей книги под «операционной системой» всегда понималась именно та *программа*, которая загружается первой и управляет аппаратурой, обрабатывает прерывания и обслуживает системные вызовы. Эту программу также называют ***ядром операционной системы***.

В некоторых случаях под словосочетанием «операционная система» понимают большой набор программ, включающий, кроме ядра, разнообразные системные утилиты, программы для управления и настройки, иногда даже компиляторы и интерпретаторы языков программирования. Как видим, термин «операционная система» может использоваться в разных значениях, тогда как использование термина «ядро операционной системы» никакой неопределённости не оставляет. В дальнейшем мы будем использовать оба термина, считая их синонимами; для обозначения «большого набора программ» мы будем применять другие термины, такие как «дистрибутив», «окружение» и т. п., наиболее подходящие по контексту.

Изучая язык ассемблера, мы выяснили, что наши программы самостоятельно могут только преобразовывать информацию в отведённой им памяти, любое взаимодействие с внешним миром возможно исключительно через операционную систему, а обращение пользовательской задачи к операционной системе за услугами называется ***системным вызовом***.

Соответствующий материал в первом томе располагается в §3.1.2 и главе §3.6 (§§3.6.1–3.6.4). Если при чтении этого параграфа вы почувствуете себя неуверенно — попытайтесь перечитать указанный материал, возможно, это решит проблему. Если нет — найдите кого-нибудь, кто сможет вам помочь. Изучать операционные системы как предмет, не понимая основ их функционирования, невозможно.

Нам также известно, что команды, которые умеет выполнять центральный процессор, делятся на *привилегированные* и *непривилегированные*, а сам процессор может работать в *привилегированном режиме* или в *ограниченном режиме*; в первом процессор готов выполнять любые команды, во втором — только непривилегированные. Операционная система стартует в привилегированном режиме, а когда ей приходится отдать управление пользовательской задаче, она меняет режим процессора на ограниченный.

Процессор может перейти обратно в привилегированный режим только при условии одновременной передачи управления на так называемые *обработчики* — фрагменты программного кода, адреса которых заданы заранее, во время работы в привилегированном режиме, и эти адреса нельзя изменить, находясь в режиме ограниченном; естественно, операционная система настраивает процессор так, чтобы в роли обработчиков использовались её собственные фрагменты. Как следствие, одновременно с переключением режима процессора на привилегированный управление всегда возвращается операционной системе, так что только её код может выполнять привилегированные команды, больше ни у кого такой возможности нет. Возврат управления системе происходит в трёх основных случаях: когда внимание операционной системы требуется внешнему устройству (***аппаратные/внешние прерывания*** или просто ***прерывания***); когда процессор не может исполнить очередную команду из-за её некорректности — несуществующего кода операции, деления на ноль, обращения за пределы отведённой задаче памяти и т. п. (***внутренние прерывания*** или ***исключения***); и при системном вызове (иногда говорят, что системные вызовы реализуются с помощью ***программных прерываний***).

Подчеркнём, что с момента загрузки **ядро является единственной программой, код которой выполняется в привилегированном режиме центрального процессора**. Все остальные программы, вне зависимости от уровня их полномочий, выполняются в ограниченном режиме в виде пользовательских задач (рис. 5.1). Вынужденная обходиться только непривилегированными командами, пользовательская задача сама по себе не может сделать ничего такого, что повлияло бы на всю вычислительную систему[1] в целом, хотя она может попросить о таких действиях операционную систему.

Мы также успели обсудить, что́ на самом деле скрывается под термином «одновременное выполнение задач»; мы уже знаем, что ***мультизадачность*** бывает ***пакетной***, где задача перестаёт выполняться лишь дойдя до конца, либо когда ей требуется дождаться какого-нибудь события, чаще всего — окончания операции ввода; кроме того, существует ***режим разделения времени***, при котором задачи дополнительно могут сменять друг друга, если одна из них успела отработать определённое количество времени, а в системе есть другие задачи, готовые к выполнению. Пакетный режим применяется на всевозможных суперкомпьютерах, где конкретная последовательность обработки задач не столь важна, важнее их общее количество, которое система успеет обработать за единицу времени, и нет смысла терять драгоценное процессорное время на лишние переключения между задачами. Для машин, ведущих непосредственный диалог с пользователями, а также для серверных машин, обслуживающих множество клиентов одновременно, пакетный режим не подходит и приходится применять режим разделения времени, обеспечивающий каждой задаче, которая готова к выполнению, определённую долю процессорного времени. При этом создаётся впечатление, что каждая задача работает на своей собственной машине, возможно, более медленной. Кроме этих двух режимов, выделяют также промежуточный ***невытесняющий режим мультизадачности***, сейчас вышедший из употребления, и ***режим реального времени***, где задача планировщика не в том, чтобы дать всем поработать хоть сколько-нибудь, а в том, чтобы обеспечить предсказуемость «грязного» времени выполнения задачи, т. е. возможность заранее понять, успеет ли та или иная программа завершиться к заданному моменту времени.

Наконец, нам известно, что для полноценного обеспечения мультизадачного режима аппаратура компьютера должна поддерживать по меньшей мере четыре возможности: аппарат прерываний, защиту памяти, разделение машинных команд на обычные и привилегированные (с поддержкой соответствующих режимов процессора) и таймер — простейшее устройство, генерирующее запросы прерывания через равные промежутки времени.

Все эти сведения об операционных системах мы были вынуждены приводить в предыдущих частях нашей книги, поскольку без них не могли двигаться дальше; но задачи, стоявшие перед нами ранее, были достаточно сложны сами по себе, так что операционные системы пришлось обсуждать довольно поверхностно.

Подробное рассмотрение операционных систем логично разделяется на два аспекта: как выглядят услуги, которые операционная система предоставляет программам, выполняющимся под её управлением, и каково устройство самой системы. В этой и двух последующих частях нашей книги мы постараемся осветить первый аспект, то есть показать, как выглядит операционная

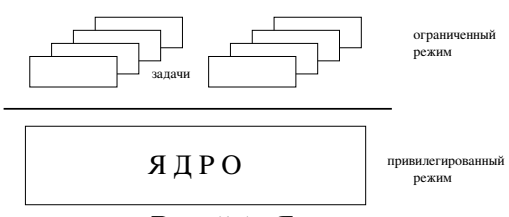

Рис. 5.1. Ядро и задачи

[1] Под ***вычислительной системой*** понимается, грубо говоря, компьютер со всеми программами, которые на нём используются.

система с точки зрения программиста, непосредственно использующего системные вызовы. Заключительная часть этого тома, восьмая в общей нумерации, посвящена тому, как операционная система устроена и каким образом она функционирует.

Если задаться вопросом, зачем нужна операционная система и что полезного она делает, то ответ можно свести всего к двум пунктам. Во-первых, операционная система *управляет пользовательскими задачами*, то есть умеет запускать программы (причём так, чтобы не потерять при этом контроля за машиной в целом), распределяет между ними процессорное время и оперативную память, даёт возможность запущенным программам взаимодействовать между собой при возникновении такой потребности. Во-вторых, она полностью берёт на себя управление аппаратной частью компьютера. Фактически операционная система — это своеобразная прослойка между прикладными программами и «железом», координирующая их взаимодействие. Каждый из этих аспектов заслуживает отдельного обсуждения, чем мы сейчас и займёмся.

5.1.2. Управление пользовательскими задачами

К управлению пользовательскими задачами относится прежде всего *запуск* задач и их *останов*. В самом деле, сам факт появления задачи в системе означает, что её кто-то запустил; как было замечено выше, операционная система, стартовав на компьютере, полностью «захватывает власть», так что запустить задачу может только она. С другой стороны, задачи имеют свойство завершаться, и действие, обратное запуску, то есть останов задачи — тоже, естественно, выполняет именно операционная система: с каждой задачей у неё связаны достаточно сложные структуры данных, и только она знает, как их нужно правильно модифицировать, чтобы исключить завершённую задачу из очереди на выполнение, высвободить занятую ею память и т. п.

Кроме того, к управлению задачами следует отнести *планирование времени центрального процессора*; коль скоро у нас в системе может одновременно существовать больше одной задачи, нужно как-то распределять между ними время центрального процессора (или центральных процессоров, если их больше одного, а в наше время это обычно так и есть), временно приостанавливать одни задачи и ставить на исполнение другие. В зависимости от используемого типа мультизадачности задаче может быть предоставлена возможность работать, пока ей есть что делать (пакетная мультизадачность), либо задачам могут выделяться ***кванты времени***[2], по истечении которых планировщик снимает текущую задачу с выполнения и даёт поработать другим задачам (конечно, если в системе есть другие задачи, готовые к выполнению).

Управление пользовательскими задачами также подразумевает *управление оперативной памятью* в широком смысле, включающем, например, распределение имеющейся памяти между задачами, управление настройками виртуальной памяти, защитой памяти и прочими возможностями аппаратуры, связанными с оперативной памятью. Временная ***откачка***[3] данных из оперативной памяти на диск с целью её высвобождения тоже относится к управлению памятью.

Отдельным запущенным в системе программам (так называемым ***процессам***) нужно дать возможность при необходимости взаимодействовать между собой; но поскольку повлиять друг на друга напрямую они не могут, средства для такого взаимодействия должна предоставлять им операционная система; иначе говоря, к управлению пользовательскими задачами относится также *организация межпроцессного взаимодействия*.

Наконец, к управлению пользовательскими задачами мы отнесём *разграничение полномочий*. Сразу отметим, что речь обычно идёт о полномочиях *пользователей*, но реально *полномочиями* в системе обладает не человек, а запущенная программа — всё тот же *процесс*, который может представлять собой пользовательскую задачу или её часть.

В системе, к которой имеет доступ много пользователей — например, если это серверный компьютер, обслуживающий группу людей или целое предприятие — необходимость регламентирования доступа пользователей к имеющимся ресурсам очевидна: даже если исключить злонамеренные действия (что, вообще говоря, уже само по себе далеко не всегда возможно), простые (не злонамеренные) ошибки пользователей в такой системе могут приводить к слишком серьёзным последствиям, если у каждого будет возможность сделать в системе что угодно. Но многопользовательские компьютеры встречаются сравнительно редко, так что такой случай, *на первый взгляд*, можно было бы считать исключением из правил, а для машин, не предполагающих работы большого числа разных людей — например, для домашних компьютеров, ноутбуков, офисных рабочих мест — пользователю компьютера можно предоставить неограниченные полномочия в системе, ведь он в системе один.

Именно так решили поступить в последней четверти прошлого века специалисты одной известной компании, выпускающей коммерческое программное обеспечение, в том числе операционные системы. Последствия этой ошибки не заставили себя долго ждать. Как мы уже отмечали, реально доступ к ресурсам и возможностям компьютера осуществляет не *пользователь*, а запускаемые им *программы*; фактически при запуске любой программы пользователь передаёт управление своим компьютером автору этой программы. Между тем, **далеко не всякой программе можно полностью доверять.** Отсутствие разграничения полномочий в некоторых популярных операционных системах стало причиной эпидемий вирусов и массового распространения всевозможных «троянских» программ, делающих на компьютере что-то такое, что напрямую противоречит интересам его владельца, но при этом отвечает интересам «хозяина» такой программы. Кроме того, при отсутствии разграничения полномочий зачастую ошибка в одной из программ, пусть и не злонамеренная, способна привести к уничтожению всей системы вместе с важными пользовательскими данными.

В системах семейства Unix вот уже без малого тридцать лет никто не видел живых и реально размножающихся вирусных программ; тому есть несколько причин.

Во-первых, в мире Unix никто не работает в системе с полномочиями системного администратора. Сами администраторы заводят себе учётные записи обычных пользователей и работают под ними, пока не потребуется сделать что-то такое, на что у обычного пользователя нет полномочий. В таких случаях администраторы запускают сеанс работы с неограниченными полномочиями, выполняют нужные действия, после чего немедленно выходят из системы. Программы, запускаемые в ходе повседневной работы, полномочий администратора не получают, возможности же, предоставляемые системой простому пользователю, настолько ограничены, что делают размножение вирусов практически неосуществимым.

Во-вторых, на проектирование систем семейства Unix в большинстве случаев почти не влияют маркетинговые соображения; поэтому, например, в пользовательских интерфейсах в среде Unix нет понятия «открыть файл произвольного типа», что в некоторых случаях оказалось бы связано с запуском файла как программы. Говоря обобщённо, в unix-системах от пользователя не стараются из маркетинговых соображений скрыть технические детали, в особенности такие, от незнания которых может пострадать безопасность. В частности, если по электронной почте отправить приложенный исполняемый файл, то в клиентских программах, работающих в системах семейства Unix, не найдётся никаких пунктов меню, кнопок и т. п., чтобы его запустить; в других системах клиентские программы такую возможность предоставляют, хотя и выдают предупреждение, а ещё несколько лет назад «случайно» (по незнанию) запустить приехавшую не пойми откуда и непонятно что делающую программу пользователь мог, не встретив ни малейших препятствий.

Кроме того, и само техническое устройство систем семейства Unix определяется в первую очередь соображениями инженерного, а не маркетингового характера, так что выбор между безопасностью и неведомым «удобством для пользователя», как правило, делается в пользу безопасности; к сожалению, в последние годы возник целый ряд исключений из этого правила. Подчеркнём, что в действительности никто толком не знает, что́ на самом деле было бы удобно пользователю; то, что пропаганда крупных коммерческих компаний называет «удобством пользователя», на самом деле представляет собой удобство для продавца по впариванию соответствующих «продуктов» покупателю. Красивые картинки, которые могут понравиться неискушённому пользователю при просмотре рекламного буклета, на выставке или на витрине магазина, совершенно не обязаны быть сколько-нибудь удобными в повседневной работе.

5.1.3. Управление внешними устройствами

Управление внешними устройствами складывается из двух основных составляющих: **абстрагирования** и **координации**. Координация доступа к устройствам нужна в мультизадачных системах, чтобы избежать конфликтов между разными задачами, которым понадобилось одно и то же устройство. Если этот момент вызывает сомнения, достаточно представить себе ситуацию, когда одной задаче понадобилось выполнить операцию чтения с диска, причём нужные ей данные находятся в его начале, а другой задаче требуется операция записи, и нужное ей место расположено в конце диска. Первая задача потребует от контроллера переместить рабочую головку диска в его начало; поскольку это требует времени, задача после этого должна «заснуть»

[2] Отметим, что исходный англоязычный термин — *time slice*; слово *quantum* в этом контексте используется реже — обычно в случаях, когда квант времени имеет переменную длину и речь идёт о вычислении этой длины.

[3] Английский термин — *swapping*; данные, *откачанные на диск*, по-английски называются *swapped out*.

(заблокироваться) в ожидании окончания запрошенного действия. Пока первая задача спит, вторая задача обратится к тому же контроллеру и потребует переместить рабочую головку в конец. Две задачи ещё имеют какие-то шансы между собой «договориться», но если таких задач с десяток, заниматься они будут в основном «перетягиванием контроллера», а не полезной работой, и в особо неудачных случаях могут вообще испортить друг другу результаты. Чтобы такого не происходило, всю работу с внешними устройствами операционная система берёт на себя. Задачи обращаются к ней через системные вызовы с просьбами о выполнении тех или иных действий с устройствами, а система решает, в какой последовательности эти просьбы исполнять; это мы и называем координацией.

С абстрагированием дела обстоят чуть иначе, здесь определяющим является не то, что в системе много задач, а то, что в мире много разнообразных внешних устройств. Работа с разными устройствами часто ведётся по совершенно различным принципам, их контроллеры поддерживают разные команды и разные форматы данных. Если бы программистам, создающим прикладные программы, приходилось в каждой такой программе учитывать особенности всех устройств, с которыми их программам, возможно, придётся работать, то на такую поддержку уходило бы гораздо больше времени, чем собственно на реализацию прикладной логики, и всё равно при установке очередной программы на случайно выбранный компьютер оставалась бы высокой вероятность того, что эта программа окажется несовместима с аппаратурой, установленной в этом компьютере, просто потому что авторы программы не учли существования какого-то из устройств, или даже, возможно, устройство ещё не существовало в природе, когда программисты писали эту программу.

Операционная система избавляет программистов от необходимости всё время думать о многообразии существующих устройств, а пользователей — от опасений за совместимость программ с аппаратурой их компьютеров. Частные особенности каждого устройства операционная система учитывает сама, а с точки зрения программ, выполняющихся под её управлением (то есть на уровне системных вызовов) большинство устройств выглядят значительно проще, чем они есть на самом деле, причём многие совершенно различные устройства выглядят так, как если бы они по своей конструкции были одинаковы. Иначе говоря, задачи вместо реальных физических устройств со всеми их особенностями видят некие *упрощённые абстракции*; поэтому мы ведём речь об *абстрагировании* как о функции операционной системы из области управления внешними устройствами.

Пожалуй, самой наглядной иллюстрацией на эту тему будет абстрактное понятие ***диска***. Читатели этой книги, возможно, видели пятидюймовые дискеты и, скорее всего, встречались с дискетами трёхдюймовыми; любители самостоятельной сборки компьютеров наверняка знают, как выглядит встроенный в компьютер жёсткий диск, а некоторые, возможно, из любопытства разбирали неисправные жёсткие диски[4] и знают, что находящаяся внутри «начинка» выглядит совсем не похожей на дискеты, несмотря на то, что принцип работы у них одинаковый — хранить информацию в виде намагниченных участков поверхности, расположенных концентрическими дорожками на вращающемся дискообразном носителе. Оптические диски (CD и DVD) читатель наверняка не только видел, но и активно использовал; принцип хранения информации на них иной, здесь под воздействием лазерного луча меняются *оптические* свойства поверхности диска, а затем другой лазер, используя эти свойства, восстанавливает исходную информацию. Дорожка присутствует и здесь, но представляет собой не семейство концентрических окружностей, как на магнитных дисках, а одну непрерывную спираль, раскручивающуюся из центра диска к краю. Для полноты картины прибавим сюда ещё флеш-брелки, «карточки памяти»[5], используемые в портативных устройствах вроде фотоаппаратов и смартфонов, и сравнительно недавно появившиеся «твердотельные накопители» (англ. *solid-state drive*, SSD), построенные на основе всё той же flash-памяти. По своему физическому устройству они вообще не похожи на диски, в них нет ничего круглого и ничего вращающегося.

На первый взгляд все перечисленные носители информации не имеют между собой ничего общего. Тем не менее, всё это многообразие устройств операционная система представляет для пользовательских задач в виде одной и той же абстракции — *диска*, то есть некоторого объекта, про который известно, что он состоит из пронумерованных *секторов* одинакового размера (обычно 512 байт каждый), хранящих некую информацию. Диски отличаются друг от друга, во-первых, количеством таких секторов; во-вторых, тем, возможно ли в них записывать данные или только читать; и, в-третьих, можно ли ожидать внезапного появления или исчезновения такого диска во время работы, то есть, грубо говоря, может ли пользователь (физически) неожиданно выдернуть носитель из компьютера. Большинство пользовательских задач, впрочем, на этот уровень не спускается, а работает с абстракцией ***файлов***, у которых есть *имена* и которые можно, используя их имена, *открывать* для чтения или записи информации, производить чтение и запись, а потом *закрывать*. При работе с файлами задача может никак не учитывать (и, как правило, не учитывает) никакие особенности дисков, на которых эти файлы расположены, особенно если говорить о физической конструкции внешних запоминающих устройств: чтение файла с оптического диска и с флеш-брелка с точки зрения пользовательской задачи выглядит абсолютно одинаково.

5.1.4. Границы зоны ответственности ОС

Итак, задачи, решаемые операционными системами, покрываются следующим списком:

- **управление пользовательскими задачами:**
  - запуск и останов процессов;
  - планирование времени центрального процессора;
  - распределение оперативной памяти;
  - организация взаимодействия между процессами;
  - разграничение полномочий;
- **управление внешними устройствами:**
  - абстрагирование от особенностей устройств разных моделей;
  - координация доступа нескольких задач к одному устройству.

Ядро операционной системы занимает особое положение среди других программ. В частности, память, занятая ядром, как правило, не может быть откачана на диск, так что чем больше будет занимать памяти ядро, тем меньше *физической* памяти останется пользовательским задачам; с самими пользовательскими задачами такой проблемы не возникает, поскольку их можно в любой момент как по частям, так и целиком откачать, освободив память для других задач. Кроме того, код ядра выполняется в привилегированном режиме, что накладывает весьма суровые дополнительные требования на его качество: ошибка в коде ядра может слишком дорого стоить, а обнаружить и исправить её гораздо сложнее, чем ошибку в обычной пользовательской программе.

Всё это ведёт нас к довольно простому принципу: ядро операционной системы не должно заниматься ничем таким, чем может без серьёзных потерь заниматься программа, запущенная как обычная пользовательская задача. В некоторых операционных системах ядро стараются минимизировать даже ценой потери производительности; например, ***драйверы***[6] физических устройств могут быть перемещены из ядра в обычные программы, но это приведёт к появлению активного обмена информацией между таким драйвером и ядром, этот обмен пойдёт через системные вызовы, так что ощутимое количество времени будет потеряно из-за частого переключения контекстов между ядром и программой, реализующей драйвер. Но если вытеснение из ядра драйверов сомнительно в силу их «системной» сущности, то всё, что связано с *прикладными* задачами, то есть с непосредственным обслуживанием интересов конечного пользователя, попросту *не имеет права* находиться в ядре.

[4] Если вас заинтересовало, что́ в процессе такой разборки можно увидеть, ни в коем случае не разбирайте рабочий жёсткий диск: после нарушения герметичности корпуса он, как правило, приходит в негодность, а информация на нём теряется. Найдите для экспериментов испорченное устройство, благо это не так уж трудно.

[5] Слово «память» здесь вносит изрядную путаницу; помните, что формально памятью следует называть только оперативную и постоянную память компьютера, с которой процессор может работать через шину без применения контроллеров, а все flash-накопители представляют собой не память, а внешние запоминающие устройства.

[6] Драйвер — это набор подпрограмм, отвечающих за работу с конкретным физическим устройством; как правило, драйвер действует как часть ядра операционной системы. Подробное обсуждение драйверов нам предстоит позднее; если вы не уверены в своём понимании этого термина — ничего страшного, здесь глубокое понимание пока не требуется.

В частности, можно совершенно определённо сказать, что **операционная система не должна содержать никаких средств, направленных на организацию общения с пользователем**. Ни оконных, ни каких-либо других пользовательских интерфейсов операционная система поддерживать не должна, это прерогатива прикладных программ; именно прикладные программы должны «разговаривать» с пользователем и обращаться к операционной системе за услугами, которые нужны, чтобы исполнить желания пользователя; иначе говоря, всё, что непосредственно видит пользователь, должно быть результатом работы прикладных программ, а не системы. В идеальной ситуации *конечный пользователь вообще должен иметь возможность не видеть её*, примерно так же, как пассажир автобуса, скорее всего, догадывается о существовании двигателя, но может совершенно не представлять, как этот двигатель устроен и даже как он выглядит.

Когда речь идёт о таких «операционных системах», как Android или Mac OS X, в первом приближении кажется, что как раз построением оконного интерфейса эти системы и заняты, но это не совсем точно. Графические надстройки в обеих этих системах не являются частью ядра, то есть, формально говоря, ни Android, ни Mac OS X нельзя вообще называть операционными системами; в роли операционной системы в первом случае выступает Linux, во втором — Darwin, ядро из семейства BSD. Программирование «под Android» или «под Mac OS» означает в основном использование высокоуровневых библиотек для построения графического интерфейса в соответствии с требованиями высокоуровневых надстроек, но это происходит в пользовательских задачах, а не в ядре.

5.1.5. Unix: семейство систем и образ мышления

Рассказывая историю ОС Unix (см. т. 1, §1.2.2), мы отметили, что сейчас этим словом обозначается не одна конкретная операционная система, а целое семейство систем, объединённых общими принципами построения. За десятилетия истории ОС Unix вокруг неё выросла целая философия; одно из наилучших и наиболее полных изложений этой философии можно найти в книге Эрика Реймонда «Искусство программирования для Unix» [6]. Мы здесь отметим только наиболее важные моменты, связанные с ответом на вопрос, что же такое Unix.

Пожалуй, главный принцип построения операционной среды ОС Unix состоит в том, что **каждая программа должна решать ровно одну задачу, и делать это хорошо**. Этот принцип в такой формулировке приписывают Дугу Макилрою (Doug McIlroy) вместе с двумя другими тезисами: **программы должны работать совместно друг с другом; программы должны использовать текстовые потоки как универсальное представление информации**.

В качестве иллюстрации этих принципов в действии автор этих строк обычно приводил своим студентам следующий пример. В одном из зарубежных филиалов МГУ приходилось вести занятия по программированию в компьютерном классе, предназначенном для студентов нескольких разных факультетов, так что на всех машинах была установлена какая-то из версий Windows; для обеспечения занятий по программированию там же был установлен сервер под управлением ОС Linux, на который студенты заходили в режиме удалённого доступа. В какой-то момент потребовалось узнать, сколько из присутствующих студентов успешно вошли в систему.

За основу была взята программа `w` (от слова *who*), выдающая список открытых в системе сеансов работы с указанием, какому пользователю принадлежит каждый из сеансов. Подвергнув её выдачу нескольким последовательным преобразованиям, удалось в итоге получить искомое число; каждое из преобразований выполнялось отдельной программой-фильтром[7], причём весь набор этих программ запускался на одновременное исполнение *конвейером*: данные, выведенные первой программой, поступали на ввод второй, выдача второй отдавалась на вход третьей и т. д.

Первые две строки в выдаче команды `w` — служебные (своего рода заголовок), они в дальнейшем анализе не нужны, поэтому их следует пропустить; это было сделано с помощью программы `tail` с параметрами `-n +3`, которая принимает на вход (читает из потока стандартного ввода) произвольный текст, первые несколько строк отбрасывает, а остальные печатает (в данном случае — начиная с третьей по счёту). Из полученной информации следовало выделить входные имена, поскольку всё остальное (адрес, с которого выполнен заход в систему, время работы, текущая запущенная программа и прочее) нас в данном случае не интересует. Входное имя в каждой строке выдачи `w` стоит в начале, от остальной строки оно отделено пробелом; чтобы оставить в каждой строке одно только входное имя, можно воспользоваться командой `cut -d' ' -f 1` (параметр `-d` задаёт символ-разделитель, параметр `-f` — номер нужного поля). Остаётся колонка из входных имён.

Самого себя, естественно, присутствующего в системе, автор учитывать не собирался; для этого пришлось с помощью программы `grep` убрать из выдачи строки, соответствующие его собственному входному имени (`avst`): `grep -v ^avst$`. После этого в обрабатываемом тексте остались только входные имена студентов, поскольку больше в системе никто на тот момент не работал. Однако некоторые студенты уже успели создать несколько сеансов работы одновременно; так сделать очень легко, достаточно запустить несколько экземпляров программы `putty` или, если используется доступ по X-протоколу, просто открыть несколько эмуляторов терминала. Для исключения повторов полученные строки пришлось отсортировать и оставить только уникальные; это было сделано командой `sort -u`. После этого осталось лишь посчитать оставшиеся строки с помощью `wc -l`. Полностью команда выглядела так:

```
w | tail -n +3 | cut -d' ' -f 1 | grep -v ^avst$ | sort -u | wc -l
```

Этот своеобразный рекорд — шесть программ в одном конвейере, решающем реально вставшую задачу — вашему покорному слуге пока побить не удалось.

Текстовое представление данных (см. т. 1, §1.4.6) в некоторых случаях требует бóльшего расхода памяти и дискового пространства, нежели компактное двоичное, и заставляет программы проводить анализ текста, который сам по себе может оказаться относительно трудоёмким делом, но при этом использование текстовых данных в качестве универсального представления информации имеет совершенно очевидные преимущества. Человеку не требуются никакие специальные программы, чтобы прочитать и понять такое представление, чтобы изменить его, а при необходимости — и создать данные в таком представлении. При наличии нескольких десятков разнообразных программ, тем или иным способом преобразующих текст, таких как `tail`, `cut`, `grep`, `sort` и `wc`, задействованных в вышеприведённом примере, работа с текстовыми данными становится на порядки легче (по трудозатратам), нежели с любыми другими.

В мире Unix текстовым данным отдаётся явное предпочтение; исключения составляют разве что исполняемые файлы, основанные на машинном коде, а также цифровая информация, полученная измерением исходно аналоговых явлений — фотографические изображения, видео- и звуковую информацию теоретически тоже можно перевести в текст, но человек с этим непосредственно работать всё равно не сможет, а места итоговый файл займёт в несколько раз больше. Следует подчеркнуть, что сказанное не распространяется на искусственно созданные изображения (рисунки, диаграммы и т. п.), на музыку, синтезированную в виде программы для виртуального музыкального инструмента (в противоположность записанной путём оцифровки сигнала с микрофонов), а также на компьютерную анимацию — в этих случаях текстовое представление остаётся оправданным и предпочтительным. Кроме оцифрованной аналоговой информации и машинного кода можно назвать ещё два случая использования нетекстовых данных: данные, запакованные архиваторами для уменьшения хранимого объёма, и данные, зашифрованные разнообразными криптографическими средствами; но на этом всё. Если обрабатываемая информация не попадает ни в одну из четырёх перечисленных категорий, в мире Unix она, скорее всего, будет представлена в виде текста.

Отметим, что многие протоколы, используемые в сети Интернет, также основаны на текстовых потоках — именно так работают, в частности, электронная почта и Web; справедливости ради надо сказать, что бинарные протоколы в Интернете тоже часто встречаются, самый популярный пример — система DNS.

Ещё один принцип, которому следует большинство программ в ОС Unix, выражен фразой «**молчание — золото**». В случае, если работа проходит успешно, программы выдают лишь ту информацию, которая от них требуется, и никоим образом не больше. Читатель мог уже заметить, что ассемблер NASM и компилятор gcc в случае успеха завершаются молча, не выдавая на терминал ни единой буквы; к сожалению, этого нельзя сказать о Free Pascal, создатели которого, по-видимому, не разделяют философию Unix.

Говоря об особенностях общения ОС Unix с пользователем, можно выделить ещё два ключевых момента. Во-первых, пользовательский интерфейс обычно строится в предположении, что пользователь знает, что делает. Большинство программ в ОС Unix не пытается делать вид, что они умнее пользователя или лучше знают, как надо работать; любой отданный пользователем приказ просто молча выполняется. Во-вторых, к любому функционалу, имеющемуся в системе, возможен доступ через команды командной строки; альтернативные интерфейсы, такие как графические оболочки или полноэкранные терминальные программы, могут существовать, но они не заменяют командную строку и не вытесняют её, они лишь предоставляют пользователю

[7] Напомним, что *фильтр* — это такая программа, которая читает некие данные из потока стандартного ввода и выводит результаты в поток стандартного вывода.

альтернативу. Во многих случаях альтернативные интерфейсы реализованы через обращение к утилитам командной строки.

Средства графического пользовательского интерфейса в ОС Unix проектируются по принципу «**метод, а не политика**». Это означает, что средства работы с графикой дают приложениям возможность создания графических интерфейсов без уточнения, как конкретно такие интерфейсы должны выглядеть и функционировать. В результате пользователь может, например, выбрать любой из нескольких десятков существующих оконных менеджеров, отвечающих за оформление окошек (рамок, заголовков и прочего), полностью изменив внешний вид своего рабочего экрана; авторы этих оконных менеджеров не ограничены в построении внешнего вида никакими правилами, навязанными со стороны.

Ещё один принцип ОС Unix можно выразить фразой «**всё есть файл**»; имена файлов в файловой системе — это самый популярный способ именования не только файлов как таковых, но и периферийных устройств (а равно и так называемых ***виртуальных устройств***, которые выглядят с точки зрения пользователя как устройства, но физически не существуют, их работу эмулирует ядро операционной системы), каналов связи программ друг с другом. Более того, через привычный файловый интерфейс в современных системах можно «добраться» до настройки режимов работы многих подсистем ядра, узнать об имеющихся в системе процессах и т. п. (это позволяют делать так называемые *искусственные* файловые системы; к ним мы вернёмся в заключительной части тома). К сожалению, из этого принципа тоже есть исключения: например, файловому принципу именования объектов не соответствуют объекты, относящиеся к так называемому System V IPC; в ОС Linux отсутствуют файлы устройств, отображающие сетевые карты, и так далее; но разнообразные девиации, даже многочисленные, не отменяют общего принципа.

Философия Unix, естественно, распространяется не только на программы, выполняемые под управлением операционной системы, но и на неё саму, то есть на ядро и в особенности на набор его системных вызовов. Как мы увидим позже, большинство системных вызовов обходится очень небольшим количеством параметров: многие системные вызовы вообще не принимают параметров (`getpid`, `fork` и прочие), часто встречаются системные вызовы с одним (`chdir`, `alarm`, `wait`, `close`), двумя (`signal`, `kill`, `getcwd`, `dup2`) или тремя (`read`, `write`, `execve`) параметрами. Реже встречаются вызовы с четырьмя параметрами (`recv`, `send`, `wait4`) и пятью (`select`, `setsockopt`, `getsockopt`), очень редко — с шестью (`recvfrom`, `sendto`, `mmap`, `pselect`); системных вызовов с бóльшим числом параметров в ОС Unix нет. Сложные действия обычно выполняются последовательным обращением к нескольким разным системным вызовам.

Также следует отметить широкое использование в интерфейсе системных вызовов ОС Unix универсальной парадигмы «**всё есть поток байтов**»: именно в виде таких потоков представлены ввод информации из файла и вывод информации в файл, чтение с клавиатуры и выдача на терминал, передача данных другому процессу через канал, во многих случаях — обмен данными по компьютерной сети. Как мы увидим позже, все так называемые ***потоки ввода-вывода*** делятся на *позиционируемые* (обычные файлы, в которых текущая рабочая позиция может быть в любой момент изменена) и *непозиционируемые* — все остальные, где байты передаются один за другим и на последовательность их передачи или приёма никак нельзя повлиять. В эту несложную модель укладываются почти все случаи приёма и передачи данных, за немногочисленными исключениями, которые обусловлены техническими причинами (например, обмен дейтаграммами по компьютерной сети не представляется в виде потока, потому что имеет совершенно иную природу).

Общий подход к созданию программного обеспечения, принятый в ОС Unix, можно выразить одним лозунгом, который, как считается, изначально появился на американском военно-морском флоте: «***Keep it simple, Stupid!***», сокращённо *KISS*; русскоязычные программисты часто называют этот лозунг «принципом поцелуя», буквально переводя слово «*kiss*» на русский.

Наконец, в ОС Unix имеются определённые традиции, связанные с разработкой и распространением программ. Как правило, программы создаются ***переносимыми***; это означает, что программа, написанная под одну unix-систему на одной аппаратной архитектуре, может быть (во всяком случае, если она написана грамотно) без изменения исходных текстов откомпилирована на любой другой unix-системе и другой аппаратной платформе. При этом конвенции системных вызовов изменяются от системы к системе, а машинный код, естественно, привязан к конкретной аппаратной платформе, так что двоичные исполняемые файлы свойством переносимости не обладают (и не могут им обладать). Как следствие, основным способом распространения программ в ОС Unix является передача их исходных текстов. Большинство программ, включая и ядра самих систем, распространяются *свободно*, то есть условия лицензии таких программ разрешают всем желающим копировать эти программы, передавать их другим лицам, изучать, модифицировать и распространять модифицированные версии, при этом не требуется кому-либо за это платить. Большинство дистрибутивов того же Linux изначально содержит тысячи полезных программ, с помощью которых можно решать практически любые повседневные задачи, и всё это может быть совершенно свободно скачано из Интернета.

Свободное распространение программ играет чрезвычайно важную роль в культуре Unix, при этом неискушённые люди часто путают понятия «бесплатная программа», «программа с открытым исходным кодом» (open source) и «свободная программа». Между тем это совершенно разные понятия. В мире Windows встречаются программы, которые можно назвать бесплатными (причём зачастую с большой натяжкой), но никак нельзя считать ни открытыми, ни тем более свободными. Наиболее распространённое определение ***свободного программного обеспечения*** (free software) предложено Ричардом Столлманом и Фондом свободного программного обеспечения; согласно этому определению, любому пользователю программы принадлежат четыре фундаментальные свободы:

0. свобода запускать программу как будет угодно пользователю, для любых целей;
1. свобода изучать, как программа работает, и изменять её сообразно со своими потребностями; для этого необходима доступность исходных текстов;
2. свобода распространять копии, чтобы можно было помочь своему соседу;
3. свобода распространять (передавать другим людям) модифицированные версии программы, давая тем самым возможность всему сообществу извлекать пользу из этих модификаций; для этого также необходима доступность исходных текстов.

Сторонники концепции свободного программного обеспечения отстаивают эти свободы в качестве неотъемлемого права каждого пользователя и составляют тексты своих лицензий таким образом, чтобы лишить пользователей этих свобод было как можно труднее. В частности, одна из самых популярных лицензий на программное обеспечение, GNU GPL, фиксируя все четыре перечисленные свободы, при этом запрещает распространение копий (как точных, так и модифицированных), а равно и производных работ под какими-либо иными лицензиями, кроме GNU GPL. Самим фактом использования любой программы, распространяемой под GNU GPL, пользователь соглашается с условиями этой лицензии. При этом производной работой считается любая программа, в которую вошёл какой-либо фрагмент кода, распространяемого под GNU GPL, так что если вы хотите заимствовать такой фрагмент для своей программы, то и свою программу вы сможете распространять только под GNU GPL — либо не распространять вовсе.

Вопреки расхожему мифу, GNU GPL не требует *бесплатности* распространения как таковой. Напротив, сам Столлман неоднократно разъяснял, что каждый волен при желании брать деньги в обмен на программу, распространяемую под GNU GPL. Иной вопрос, что коль скоро кто-то купил у вас программу, распространяемую под GNU GPL, этот кто-то получает все права, предусмотренные GNU GPL, и вы, согласно условиям лицензии, не имеете права даже просить его воздержаться от использования полученных таким образом прав. Например, если кто-то купит у вас за миллион долларов программу, на которую распространяется лицензия GNU GPL, а затем выложит её в Интернет в режиме свободного доступа, вы никак не сможете (и не будете вправе!) на это повлиять; любые «дополнительные соглашения», ограничивающие свободу, данную лицензией GNU GPL, являются нарушением этой лицензии.

Отметим, что понятие программного обеспечения с открытым кодом отличается от понятия свободного программного обеспечения: для свободного программного обеспечения открытость исходного кода необходима, но обратное неверно. Во-первых, существуют другие лицензии на программное обеспечение, опубликованное в виде исходных кодов; такие лицензии могут не заботиться, как GNU GPL, о сохранении свобод для *всех* пользователей. Например, так называемая BSD license позволяет делать с программой что угодно, в том числе и распространять производные работы без их исходных текстов; получатель таких программ обладает всеми вышеперечисленными свободами, но не обязан заботиться о том, чтобы ими обладал кто-то ещё. С другой стороны, известны примеры программного обеспечения, которое хотя и распространяется в исходных текстах, но при этом не является свободным ни с какой точки зрения: лицензионный договор не позволяет распространять его дальше ни в виде точных копий, ни тем более в виде копий модифицированных.

5.1.6. Замечания о системе X Window

Мы уже отмечали, что ОС Unix как таковая не претерпела существенных архитектурных изменений даже при таком серьёзном шаге, как введение графических оболочек, поскольку, не будучи частью операционной системы, они не смогли оказать на неё заметного влияния. Давайте попробуем разобраться, как это стало возможным.

Средства, используемые в большинстве unix-систем для работы с графическими (оконными) приложениями, получили общее название ***X Window System***. В качестве основного определяющего свойства этих средств можно назвать то, что реализуются они целиком в виде обычных пользовательских программ; поддержка со стороны ядра операционной системы ограничивается специальными средствами для доступа к видеокарте, но эти средства никак не учитывают особенности X Window System и могут использоваться любыми другими графическими оболочками, если таковые появятся. В проприетарных системах на основе Unix действительно часто встречаются альтернативные графические оболочки; к таким системам относятся, например, упоминавшиеся выше Mac OS X и Android. Созданные для них графические средства не имеют никакого отношения к X Window, за исключением разве что основного принципа: они тоже реализованы в виде пользовательских программ и не затрагивают ядро.

Возвращаясь к X Window System, отметим, что приложение, использующее оконный графический интерфейс, продолжает при этом быть обычным Unix-процессом. В частности, как и у любого процесса, у оконного приложения есть параметры командной строки, а также дескрипторы стандартного ввода, вывода и сообщений об ошибках (`stdin`, `stdout` и `stderr`). Чтобы отобразить окно, приложение должно установить соединение с системой X Window и отправить запрос в соответствии с определёнными соглашениями, известными как ***X-протокол***.

«По ту сторону» соединения находится программа, называемая X-сервером. Именно эта программа производит непосредственное отображение графических объектов на экране пользователя. По собственной инициативе она ничего не рисует; чтобы на экране что-то появилось, необходим запрос на отрисовку того или иного изображения. Таким образом, в основном отображающая программа выполняет действия в ответ на запросы других программ (клиентов), что оправдывает название «X-сервер». Услугой (сервисом), которую оказывает клиентам этот сервер, оказывается отрисовка изображений на экране. В некоторых случаях X-сервер сам проявляет инициативу при общении с клиентом. Это происходит при возникновении тех или иных событий, относящихся к области экрана, в которой отображено окно клиента; к таким событиям относятся нажатия клавиш на клавиатуре, движения и щелчки мыши, перемещения окон, требующие полной или частичной повторной отрисовки изображения в окне, и т. д.

Система спроектирована так, что соединение с X-сервером можно установить, используя как локальные средства взаимодействия внутри одного компьютера, так и инструменты работы через компьютерную сеть. С точки зрения пользователя это значит, что при работе с X Window можно запускать оконные приложения на удалённых компьютерах, при этом графические окна этих приложений видеть локально, то есть на своём экране.

Забегая вперёд, скажем, что для соединения с X-сервером используется потоковый сокет (сокет типа `SOCK_STREAM`), причём обычно X-сервер заводит слушающие сокеты как в семействе `AF_UNIX` для локальных подключений, так и в семействах `AF_INET` и `AF_INET6`, что позволяет связываться с X-сервером по сети. После установления соединения такие сокеты выглядят совершенно одинаково — как двунаправленные каналы последовательной передачи данных. Во время работы X-клиенту и X-серверу безразлично, какую природу имеет установленное между ними соединение. Подробности отложим до следующей части нашей книги.

Важно отметить, что X-сервер, как и X-клиенты, представляет собой обыкновенный процесс, т. е. это просто программа, которую нужно запустить обычным путём. Поддержка графического интерфейса со стороны ядра ограничивается предоставлением доступа к видеокарте, например, путем отображения видеопамяти на виртуальное адресное пространство X-сервера. Это позволяет X-серверу не быть частью операционной системы.

Интересно, что X-сервер не обязан для отображения графики использовать доступ к видеокарте; он, собственно говоря, обязан поддерживать X-протокол — и всё. Так, широко известен X-сервер `Xvfb`, который построенное изображение нигде не показывает, а просто записывает в файл. Кроме того, существуют X-сервера, работающие в системах семейства Windows и позволяющие пользователю рабочей станции под MS Windows запускать удалённо (на Unix-машине) оконные приложения и взаимодействовать с ними; наиболее широкое распространение среди них получили Xming и Cygwin/X.

Для лучшего понимания происходящего полезно провести небольшой наглядный эксперимент, в ходе которого потребуется запустить X-сервер без обычной прикладной обвески. Раньше это делалось командой `X` (файл `/usr/bin/X` обычно представляет собой символическую ссылку на установленный в вашей системе X-сервер); видимо, авторы X.org (а в вашей системе, скорее всего, установлен именно он, если только вы не выбрали дистрибутив с новомодным и, как водится, ни на что не годным Wayland) в какой-то момент решили, что так теперь не модно, и благополучно сломали привычное поведение X-сервера по умолчанию, но, к счастью, сохранили возможность получить классический режим работы, только для этого теперь нужно указать флаг `-retro`.

Если вы захотите провести описываемый эксперимент самостоятельно, для начала нужно понять, будут ли в это время на вашей машине запущены другие X-сервера. Во многих «современных» дистрибутивах Linux от работающего X-сервера довольно непросто избавиться, но это не страшно: мы можем приказать запускаемому X-серверу работать вторым дисплеем машины (используйте команду «`X -retro :1`» или «`X -retro :1.0`»). Если же у вас есть возможность провести эксперимент на «чистой» машине, достаточно будет просто команды «`X -retro`»; при этом будет использоваться идентификатор дисплея по умолчанию (`:0.0`).

Запустив X-сервер, мы увидим пустой экран с курсором мыши, напоминающим очень жирную букву X, и фоном, выглядящим, как увеличенный рисунок грубой ткани — это обычное фоновое изображение X-сервера, которое прикладные программы, как правило, тут же меняют на какое-то своё.

Всё, что мы можем сделать с запущенным таким вот образом сервером — это подвигать курсор с помощью мыши. Чтобы получить более интересные результаты, нужно запустить хотя бы одну прикладную программу. Для этого следует, нажав комбинацию `Ctrl-Alt-F1`[8], вернуться в текстовую консоль, с которой мы запустили X-сервер, нажатием `Ctrl-Z` и командой `bg` убрать работающую программу `X` в фоновый режим. К сожалению, в современных условиях может не получиться даже это: в зависимости от конкретной сборки X-сервер может заблокировать реакцию на `Ctrl-Z` (не спрашивайте, зачем это сделано; скорее всего, этого не знают толком даже те, кому пришла в голову такая «замечательная» идея). В этом случае придётся воспользоваться ещё одной текстовой консолью. Напомним, что переключение между ними делается нажатием `Alt-F1`, `Alt-F2`, `Alt-F3` и т. д., а из запущенного X-сервера можно перейти сразу на нужную консоль, нажав `Ctrl-Alt-F2`, `Ctrl-Alt-F3` и др. Естественно, на второй консоли придётся тоже ввести ваше имя и пароль.

Итак, тем или иным способом мы получили в своё распоряжение приглашение командного интерпретатора. Теперь мы можем запустить какую-нибудь прикладную программу, например `xterm`. Поскольку мы не воспользовались обычной «обвеской» для запуска X-сервера, нам придётся самостоятельно указать программе, с каким X-сервером пытаться установить соединение. С учётом этого команда (при использовании Bourne Shell) будет выглядеть так:

```
DISPLAY=:0.0 xterm
```

Если вы запустили свой экспериментальный экземпляр X-сервера одновременно с другим работающим X-сервером, вместо «`:0.0`» укажите соответствующий идентификатор дисплея, например «`:1.0`».

Вернёмся теперь к нашему X-серверу; в зависимости от обстоятельств нам для этого понадобится комбинация клавиш `Alt-F7`, `Alt-F8` и т. п. Если всё прошло успешно, мы увидим в левом верхнем углу экрана окно программы `xterm` и, переместив в него курсор мыши, сможем убедиться, что командный интерпретатор в нём работает. Однако целью нашего эксперимента было не это. Главный факт, который сейчас можно констатировать — это полное отсутствие у окна каких-либо элементов оформления. Нет ни рамки, ни заголовка, ни привычных кнопок в уголках окна, служащих для свёртывания в иконку, развёртывания на весь экран и закрытия — ничего! В такой ситуации мы не можем, например, переместить имеющееся окошко или изменить его размер.

[8] Предполагается, что эксперимент проводится на машине под управлением ОС Linux или FreeBSD, а запуск произведён с первой виртуальной консоли.

Итак, наш нехитрый эксперимент дал нам возможность узнать, что в системе X Window за обрамление окон не отвечают ни X-сервер, ни клиентское приложение. Так сделано не случайно. Возложив ответственность за декор окон на X-сервер, мы навязали бы один и тот же внешний вид (и возможности управления) всем пользователям данного сервера. Несмотря на то, что известные коммерческие компании именно так и поступают с пользователями своих операционных систем, этот подход невозможно назвать правильным. Если же заставить каждого X-клиента отвечать за стандартные элементы его собственного окна, это резко увеличит сложность оконных приложений, причём изрядная часть их функциональности будет в программах дублироваться. Кроме того, это снизит возможности пользователя по выбору удобного ему декора окон. В системе X Window за стандартные элементы оконного интерфейса отвечают специальные программы, называемые ***оконными менеджерами***.

Интересно, что в мире unix-систем нашлась целая команда странных людей, решивших разломать даже это. Программы, написанные для «современных» версий Gnome, зачем-то берут на себя функции, традиционно относящиеся к зоне ответственности оконного менеджера. В итоге с этими программами крайне неудобно работать под управлением оконников, отличных от Gnome.

Продолжая наш эксперимент, запустим какой-нибудь простой оконный менеджер, например `twm`, обычно входящий в поставку X Window. Это можно сделать, не покидая X-сервер, ведь у нас уже запущена программа `xterm`, и в её окне работает интерпретатор командной строки. Итак, переместите курсор мыши в область окна и дайте команду `twm`. Если программы с таким именем в вашей системе не нашлось, вы можете её поставить средствами вашего пакетного менеджера, а можете воспользоваться каким-то ещё из «лёгких» оконных менеджеров — `fvwm2`, `icewm` и т. д.; для нужд нашего эксперимента подходит любой «оконник» из тех, что не пытаются притворяться «рабочими столами» (Desktop Environment).

После запуска оконного менеджера вокруг всех имеющихся окон появятся элементы оформления. Теперь окна можно двигать, закрывать, менять их размер и т. д. Для получения более интересного эффекта можно сначала запустить одно-два небольших оконных приложения, например, `xeyes`, и только после этого запускать `twm`. Теперь попробуйте подвигать окна по экрану, после чего убейте процесс `twm`; элементы декора со всех окон тут же исчезнут, но сами окна никуда не денутся. После этого можно снова запустить `twm` или другой оконный менеджер.

Оконный менеджер в X Window представляет собой клиентскую программу, которая с точки зрения самой системы ничем не отличается от других приложений. С X-сервером оконный менеджер общается с помощью того же X-протокола, что и остальные клиенты.

Как уже говорилось, оконное приложение может выполняться на той же машине, где запущен X-сервер, а может и на совсем другой, связываясь с X-сервером по сети. Пользуясь этим свойством X-протокола, можно создавать специализированные компьютеры, единственным назначением которых будет поддержка X-сервера. Такие компьютеры называются ***X-терминалами***. При работе с X-терминалом все пользовательские программы выполняются где-то в другом месте, как правило, на специально предназначенной для этого мощной машине. Такую машину обычно называют ***сервером приложений***. Отметим, что сервер приложений может вообще не иметь собственных устройств отображения графической информации, что не мешает ему выполнять графические программы.

При работе с X-терминалом пользователю нужен доступ к его домашнему каталогу и другим ресурсам, которые находятся, естественно, на удалённой машине (на самом сервере приложений или на файловом сервере), ведь X-терминал никаких задач, кроме отображения графики (то есть выполнения программы X-сервера), не решает. Как следствие, требуется обеспечить возможность аутентификации пользователя на удалённой машине, создание сеанса работы, включающего, например, программу оконного менеджера, которая уже управляет X-сервером и позволяет запускать программы пользователя. Для проведения такой удалённой аутентификации система X Window предусматривает специальные средства. На сервере приложений запускается процесс-демон, называемый обычно `xdm` (X Display Manager). Запущенный на терминале пользователя X-сервер обращается к программе `xdm` с использованием протокола XDMCP (X Display Manager Control Protocol). Функционирование `xdm` несколько напоминает традиционную схему `getty`: на графический экран пользователя выдается приглашение к вводу входного имени и пароля, после чего (уже с правами аутентифицировавшегося пользователя) запускается головной процесс нового сеанса. В традиционной схеме работы текстовых терминалов таким главным процессом выступает интерпретатор командной строки, а для сеансов работы с X-терминалом в качестве главного процесса запускается обычно либо оконный менеджер, либо (чаще) некий командный файл, который производит подготовительные действия, а затем уже запускает оконный менеджер.

Со схемой взаимодействия X-терминала, сервера приложений, программы `xdm` и пользовательских программ (X-клиентов) связана, к сожалению, изрядная терминологическая путаница. X-терминал — это *клиентская* рабочая станция, в конечном счёте обращающаяся к *серверной* (обслуживающей) машине — серверу приложений. Более того, на сервере приложений запускается программа `xdm`, представляющая собой *сервер* протокола XDMCP. И в то же время программа, выполняющаяся на X-терминале (клиентской машине!), называется ***X-сервером***, а пользовательские программы, выполняющиеся на сервере приложений (!), называются ***X-клиентами***.

Дело в том, что с точки зрения X Window System сервером, предоставляющим *услугу по отображению графических объектов*, является как раз X-терминал, а обращающиеся за такой услугой программы (пользовательские приложения) оказываются, соответственно, *клиентами*. Попросту говоря, используемая терминология зависит от уровня, на котором мы рассматриваем участников взаимодействия. На уровне X Window System X-терминал является сервером, на уровне пользовательских услуг — безусловно, клиентом.

Достоинством схемы с использованием нескольких мощных серверов приложений и множества X-терминалов является крайняя дешевизна администрирования и обслуживания такой сети. X-терминалы обычно не имеют дисковых подсистем; некоторые из них способны обходиться без вентиляторов за счёт использования сравнительно медленных процессоров. Как следствие, в них попросту нечему ломаться. Никакой настройки большинство X-терминалов не требует, все необходимые параметры они получают при подключении к сети. В обслуживании и администрировании нуждаются только серверные машины. В организации, имеющей несколько сот рабочих мест, в качестве серверов приложений приходится использовать мощные и дорогостоящие компьютеры, однако вложения в них быстро окупаются за счёт экономии расходов на текущий ремонт и прочее обслуживание пользовательских рабочих станций. Такие расходы при использовании X-терминала могут быть в десятки раз ниже, чем при использовании обычных персональных компьютеров.

До недавнего времени функциональность X-терминалов ограничивалась только передачей в одну сторону событий от клавиатуры и мыши, а в другую — графической информации. Для обычной работы с компьютером в офисе этого было достаточно, но в современных условиях тяжело говорить о полноценной работе без воспроизведения звука. Этот вопрос при работе с X-терминалами также легко решается с помощью специальных программ, таких как JACK и PulseAudio; большинство современных приложений для unix-систем, предполагающих работу со звуком, поддерживают взаимодействие с одной из этих программ или с обеими. С точки зрения пользователя это означает возможность услышать средствами локальной машины (в том числе X-терминала) звук, воспроизводимый программой, запущенной на удалённой машине (сервере приложений).

5.1.7. Системные вызовы и их обёртки

Как мы уже знаем, ядро операционной системы предоставляет свои услуги пользовательским программам через интерфейс системных вызовов; программируя на языке ассемблера, мы видели этот интерфейс в действии, инициируя вручную так называемые программные прерывания, характерные для процессоров семейства i386, имеющих 32-битную архитектуру. Другие процессоры не используют концепцию программных прерываний, и даже на 64-битных «наследниках» архитектуры i386 команда `int` больше не используется для обращения к ядру операционной системы, поскольку появились более «правильные» механизмы, но общая идея остаётся прежней: пользовательская программа для выполнения системного вызова должна произвести некие (зависящие от процессора) действия, в результате которых ядро операционной системы получит управление и сможет узнать, чего от него хочет пользовательская программа.

Обсуждая язык Си, мы убедились, что стандартная библиотека предоставляет программисту обёртки системных вызовов в виде обычных (по

внешнему виду) функций языка Си, которые называются так же, как и сами системные вызовы. Кроме того, мы узнали, что в принципе, применив фрагменты на языке ассемблера, мы можем обойтись без библиотечных «обёрток» и взаимодействовать с ядром напрямую, но так обычно всё же не делают. Интерфейсы системных вызовов для ядер различных систем друг от друга заметно отличаются — например, мы видели, что для ядер Linux и FreeBSD некоторые системные вызовы имеют различные номера, плюс к тому один и тот же системный вызов (например, `open`) может в разных системах требовать различных значений констант. Интересно, что при переходе к 64-битной архитектуре в ядре Linux были изменены номера всех системных вызовов; их объединили в группы по восемь штук так, чтобы вызовы, для которых выше вероятность их использования вместе, по возможности оказывались в одной «восьмёрке». Сделано это было из соображений оптимизации: указатели на соответствующие функции образуют массив, который в памяти ядра специально располагается так, чтобы каждые восемь элементов начинались с адреса, кратного 64, в результате указатели из каждой «восьмёрки», занимая как раз 64 байта, оказываются в одной странице кеша. Откровенно говоря, выигрыш от этого новшества достаточно сомнителен.

Так или иначе, библиотека языка Си сглаживает различия между интерфейсами различных ядер, делая программы переносимыми. Дело здесь не только и не столько в проблемах нумерации и константах; часто бывает так, что некая функция, будучи системным вызовом в одной системе, оказывается в другой системе простой библиотечной функцией, эмулирующей соответствующую функциональность через другие системные вызовы.

Обёртки системных вызовов в системах семейства Unix используют общий механизм обработки ошибок. Как правило, если системный вызов в принципе может завершиться ошибочно, его обёртка возвращает значение `-1`, при этом в глобальную переменную `errno` заносится код ошибки. Этот код можно использовать напрямую, сравнивая его с макроконстантами, определёнными той же библиотекой, такими как `ENOENT`, `EINVAL`, `EPERM`, `EACCES` и т. п.; можно также воспользоваться библиотечными функциями, такими как `strerror` (возвращает текстовое описание ошибки по её коду) или `perror` (выдаёт сообщение о последней ошибке в диагностический поток вывода). В документации на каждый системный вызов перечисляются все коды ошибок, которые может установить данный системный вызов.

Полезно знать, что с массовым распространением многопоточного программирования `errno` перестала быть, собственно говоря, переменной; теперь это, как правило, макрос, который вызывает некую функцию, которая возвращает указатель на целочисленную переменную, макрос разыменовывает полученное значение, превращая его в «переменную» — точнее, в леводопустимое выражение. Все эти пляски и приседания сделаны затем, чтобы у каждого независимого потока управления был свой собственный экземпляр `errno`. В целом со словом `errno` вы можете работать так же, как и с любой другой целочисленной переменной — обращаться к ней, присваивать ей значение, брать её адрес; единственное, чего точно не следует делать — это пытаться назвать таким именем свою собственную переменную или другой объект, поскольку при этом прелесть макропроцессора языка Си проявится во всей красе, ваша программа, разумеется, не пройдёт компиляцию, причём выданная компилятором диагностика будет такова, что разобраться в ней окажется не слишком просто даже опытным программистам.

5.1.8. О разграничении полномочий

Как мы знаем, возможности пользовательской задачи (самой по себе) ничтожны: всё, что она может без обращения к системе — это преобразовывать данные в отведённой ей памяти. Возможности ядра операционной системы, напротив, безграничны — или, точнее, ограничены только возможностями аппаратуры компьютера. Обслуживая системные вызовы, ядро задействует часть своего всемогущества, чтобы исполнить просьбы, поступившие от пользовательских задач; без этого выполнение пользовательских задач не имело бы смысла, ведь все результаты работы так и оставались бы в отведённой им памяти, где их никто не видит.

С другой стороны, как мы уже отмечали, одна из важнейших функций операционной системы — это *разграничение полномочий*. Если бы ядро было готово исполнить *любую* просьбу пользовательской задачи, то само ядро как отдельная сущность было бы не нужно, оно не смогло бы достигнуть большей части поставленных перед ним целей. Можно придумать множество потенциальных просьб со стороны пользовательских задач, которые ядро не станет выполнять, сколько его ни проси. Так, никакая пользовательская задача не может получить доступ к памяти ядра, перепрограммировать защиту памяти, запретить прерывания или перенастроить их обработку; для таких просьб ядро попросту не предусматривает системных вызовов. Но даже если то или иное действие по инициативе пользовательской задачи может быть выполнено, это не означает, что оно может быть выполнено по просьбе *любой* пользовательской задачи. Технически именно в этом и состоит разграничение полномочий в системе: прежде чем исполнить тот или иной запрос пользовательской задачи (т. е. системный вызов), ядро операционной системы проверяет, *имеет ли право* эта задача требовать выполнения такого запроса.

Можно привести примеры запросов к ядру, на которые имеет право вообще любая задача. Так, задача всегда может потребовать её немедленно завершить: знакомый нам по предыдущему тому системный вызов `_exit` всегда отрабатывает успешно. Также любая задача имеет право узнать текущее время (значение системных часов), получить целый ряд сведений о самой себе, закрыть любой из своих открытых потоков ввода-вывода и т. п. Во всех этих случаях ядро исполняет просьбу задачи без каких-либо проверок допустимости такой просьбы.

Во многих случаях ядро соглашается исполнить просьбу задачи лишь при условии, что задача работает от имени конкретного пользователя или же от имени одного из пользователей, входящих в ту или иную *группу*. К примеру, задача может обратиться к ядру с просьбой о прекращении выполнения *другой* задачи. Такая просьба будет удовлетворена только в случае, если у обеих задач один и тот же владелец; иначе говоря, пользователь имеет право снять свою собственную задачу, но не чужую. Кроме того, именно принадлежность задачи конкретному пользователю лежит в основе разрешения или запрещения тех или иных операций с файлами в файловой системе: как мы знаем из вводной части первого тома, у файла тоже есть владелец, а также *права доступа*, устанавливаемые отдельно для владельца, группы пользователей и всех остальных пользователей. Когда пользовательская задача пытается открыть файл с помощью системного вызова `open`, ядро проверяет, кто является владельцем задачи и достаточно ли у него полномочий в отношении данного файла, чтобы открыть этот файл в том режиме, в котором просит задача; если полномочий не хватает, система отказывается открывать файл, вызов `open` возвращает ошибку. Аналогичные проверки выполняются и в других случаях; например, изменить права доступа к файлу может только его владелец.

Среди всех пользователей системы выделяется один особый пользователь, называемый системным администратором или ***суперпользователем*** (англ. *superuser*); в системах семейства Unix этот пользователь по традиции называется `root`[9]. На задачи, выполняемые от имени суперпользователя, ограничения по владельцу не действуют. В частности, такие задачи могут открывать любые файлы в системе как на чтение, так и на запись, не обращая внимания на права доступа; также они могут потребовать от ядра приостановить или вообще уничтожить любую другую задачу, кому бы она ни принадлежала, и т. д. Кроме того, существует множество действий, которые ядро готово исполнить только для суперпользователя, т. е. если соответствующий системный вызов выполнила задача, имеющая полномочия суперпользователя. Простейший пример такого действия — изменение текущего времени в системе, но, конечно, этим дело не ограничивается; позже мы столкнёмся с целым рядом системных вызовов, которые доступны только пользователю `root` (его задачам).

Наконец, в системе могут быть установлены разнообразные количественные ограничения, также во многих случаях связанные с пользователями, хотя и не всегда. Например, можно ограничить количество оперативной (на самом деле виртуальной) памяти, которую имеет право затребовать отдельно взятая задача, или количество одновременно открытых файлов (потоков ввода-вывода); можно также ограничить количество одновременно работающих задач для отдельно взятого пользователя и т. д. Бóльшая часть таких ограничений тоже реализована на уровне системных вызовов: ядро отказывается исполнять просьбу, поступившую от пользовательской задачи, если это приведёт к превышению действующих ограничений.

В системах семейства Unix реализация разграничения полномочий основана на концепции ***идентификатора пользователя*** и ***идентификатора группы пользователей*** (англ. *user identifier* (`uid`), *group idendifier* (`gid`)). Эти идентификаторы представляют собой неотрицательные целые

[9] Основное значение этого английского слова — *корень* или *основа*; как это часто бывает, слово *root* в английском языке имеет много значений и переносных смыслов; точно перевести, чтó в данном случае имели в виду создатели Unix, затруднительно.

числа; в некоторых системах разрядность параметров `uid` и `gid` составляет 16 бит, в большинстве современных систем — 32 бита, хотя обычно в системе не требуется заводить так много (свыше 65 тысяч) пользователей.

Следует сразу же оговориться, что под словом «пользователь» далеко не всегда подразумевается живой человек; для того, что в системах семейства Unix традиционно именуют «пользователем» (англ. *user*), в современных условиях точнее подходит другой термин — английское слово *account*, которое на русский язык в этом контексте лучше всего переводить словосочетанием *учётная запись*. В частности, никто не мешает пользователю-человеку завести на компьютере несколько учётных записей для разных видов работы; с точки зрения системы это ничем не будет отличаться от ситуации, когда с компьютером работают разные люди. Более того, многие учётные записи (как правило, даже большинство) вообще не имеют ничего общего с людьми и заводятся для изоляции в системе тех или иных программ, то есть чтобы при запуске некоторой программы она получала только строго определённые возможности; часто такие программы запускаются автоматически, без участия пользователя-человека. Учётные записи, не соответствующие реальным пользователям системы, по-английски иногда называют *dummy users* или *dummy accounts*. Несмотря на всё сказанное, мы будем по-прежнему употреблять термин «пользователь», но при этом важно помнить, что пользователь-человек и «пользователь» с точки зрения ядра системы — это совершенно разные сущности; по большому счёту, с точки зрения ядра «пользователь» — это просто номер, тот самый `uid`.

Идентификатор пользователя, равный нулю, имеет в ОС Unix особый смысл: это идентификатор суперпользователя — вышеупомянутого пользователя `root`. Как мы уже говорили, суперпользователь обладает в системе гораздо более широкими возможностями, нежели обычный пользователь; ядро отличает задачи, работающие от имени суперпользователя, как раз по нулевому значению их пользовательского идентификатора, а про имена пользователей ядро вообще ничего не знает, ему это не нужно.

Пользователи в системах семейства Unix могут объединяться в ***группы***, причём каждый пользователь входит минимум в одну группу, называемую *основной* для данного пользователя. В некоторых дистрибутивах Linux при создании в системе новой пользовательской учётной записи одновременно создаётся также и группа, а создаваемый пользователь становится в этой группе единственным членом; в других дистрибутивах по умолчанию создаётся одна группа для всех пользователей, обычно она так и называется *users*. Следует отметить, что группа пользователей — это самостоятельный объект, она может существовать, даже если в неё не входит ни один пользователь. С другой стороны, отдельно взятый пользователь может входить в несколько групп, но только одна из них будет для него основной; остальные группы, в которые входит пользователь, называются *дополнительными* (англ. *supplementary*).

Исходно группы были придуманы, чтобы пользователи, имеющие доступ к многопользовательской вычислительной машине, могли организовать совместную работу — например, делать те или иные файлы и директории доступными членам своей группы, но оставлять их недоступными для всех остальных. В современных условиях, когда у большинства компьютеров всего один живой пользователь, группы чаще применяются для более гибкой настройки полномочий, ассоциированных с разными учётными пользовательскими записями в системе. Если, к примеру, нужно дать возможность печатать на принтере не всем пользователям в системе, а только некоторым из них (в том числе, возможно, лишить этой возможности часть автоматически запускаемых программ, для которых созданы отдельные учётные записи), то можно создать специальную группу, дать этой группе доступ к файлам соответствующих устройств или каналов, имеющих отношение к печати на принтере, и включить в эту группу тех и только тех пользователей системы, которым печать на принтере разрешена.

Чтобы получить более наглядное представление о пользователях и группах, вы можете посмотреть, как обстоят с ними дела в вашей собственной системе. В системах семейства Unix для описания пользователей и групп используются файлы `/etc/passwd` и `/etc/group`. Это обычные текстовые файлы, доступные на чтение всем пользователям системы, так что просмотреть их (например, с помощью команды `less`) вы можете в своём обычном сеансе работы, администраторские полномочия для этого не нужны. Каждая строка в файле `/etc/passwd` соответствует учётной записи пользователя, строка в `/etc/group` — группе пользователей. Отдельные поля в строках обоих файлов разделены двоеточиями. Файл `passwd` может выглядеть примерно так:

```
root:x:0:0:root:/root:/bin/bash
daemon:x:1:1:daemon:/usr/sbin:/bin/sh
bin:x:2:2:bin:/bin:/bin/sh
sys:x:3:3:sys:/dev:/bin/sh
sync:x:4:65534:sync:/bin:/bin/sync
```

...

```
avst:x:1000:1000:Andrey V. Stolyarov:/home/avst:/bin/bash
```

Традиционно этот файл начинается с описания пользователя `root`, за ним следуют системные учётные записи, не имеющие отношения к живым людям — упоминавшиеся выше *dummy accounts*. Свою собственную учётную запись вы, скорее всего, обнаружите в самом конце файла. Поля в каждой строке-записи означают (слева направо) входное имя пользователя (*login name*), пароль, идентификатор пользователя (`uid`), идентификатор его основной группы (`gid`), «настоящее имя» (*real name*), путь к домашней директории пользователя и, наконец, командный интерпретатор, который нужно запускать для данного пользователя при его входе в систему. Так, последняя строчка в примере выше, которая соответствует учётной записи автора книги на его личном компьютере, указывает, что входное имя для этой учётной записи — `avst`, идентификаторы (`uid` и `gid`) равны `1000`, домашняя директория — `/home/avst`, командный интерпретатор — `/bin/bash`, а в качестве «настоящего имени» указано «`Andrey V. Stolyarov`» (конечно, здесь можно написать вообще что угодно, ведь система не умеет проверять документы).

Отметим, что поле, предназначенное для пароля, практически во всех системах содержит букву «`x`»; это означает, что никакого пароля там на самом деле нет. Примерно до середины 1990-х годов пароли действительно хранили в файле `/etc/passwd` в виде хеша, т. е. в зашифрованной форме, не допускающей расшифровки; правильность введённого пароля в такой ситуации проверяется путём его зашифровки и сравнения уже зашифрованного варианта с хешем, имеющимся в системе. Практика показала, что такой подход не удовлетворяет требованиям безопасности. Дело в том, что файл `/etc/passwd` обязательно должен быть доступен на чтение всем пользователям системы — в противном случае, например, команда `ls` не сможет показать, кто является владельцем того или иного файла, и т. д. Имея хеши паролей всех пользователей системы, злоумышленник вроде бы не может их расшифровать, но зато он может *подобрать* пароли с помощью некоего оптимизированного перебора. Так, пароли длиной до шести символов включительно подбираются полным перебором; более длинные пароли можно попытаться подобрать, используя словарные слова и их комбинации с цифрами; известны и более изощрённые методы подбора. Поэтому в современных системах пароли в `/etc/passwd` не включаются ни в каком виде. В большинстве случаев пароль хранится (в виде всё того же хеша) в отдельном файле `/etc/shadow`, который доступен на чтение только суперпользователю; в некоторых системах нет и его, а пароли хранятся в специальной директории `/etc/tcb`, где для каждой учётной записи имеется отдельная поддиректория, а в ней — файл, содержащий пароль. Существуют и другие подходы.

Структура файла `/etc/group` несколько проще. Выглядит он примерно так:

```
root:x:0:
daemon:x:1:
bin:x:2:
sys:x:3:
adm:x:4:avst,olga
```

...

```
avst:x:1000:
```

Полей здесь, как видим, всего четыре: имя группы, пароль (причём никто толком не знает, зачем он здесь нужен; автору никогда не встречались системы, где хоть как-то мог использоваться пароль группы), идентификатор (`gid`) и список пользователей, входящих в группу; пользователи в списке перечисляются через запятую. Отметим, что в свою *основную* группу — ту, что указана в четвёртом поле в `/etc/passwd` — пользователь входит в любом случае, этот факт не требует отражения в `/etc/group`. Любопытно, что обычно в системе присутствует группа `root` с идентификатором `0`, но, в отличие

от такого же идентификатора *пользователя*, нулевой идентификатор группы никакого специального смысла не имеет.

Важно понимать, что **ядро системы ничего не знает о файлах `/etc/passwd` и `/etc/group`**. Информацию из этих файлов используют программы, работающие под управлением системы, но не сама система; с точки зрения ядра учётная запись пользователя представляет собой только `uid` (т. е. целое число) и больше ничего, группа пользователей — `gid` и больше ничего. В частности, когда пользователь входит в систему, диалог с ним ведёт одна из программ, выполняющихся с полномочиями системного администратора; эта программа запрашивает у пользователя входное имя и пароль, проверяет их, используя информацию в файлах `/etc/passwd` и `/etc/shadow`, затем присваивает себе (как задаче) соответствующие идентификаторы пользователя и основной группы, а также идентификаторы дополнительных групп (уже на основе информации из `/etc/group`); от прав администратора программа при этом отказывается. Установив себе такие полномочия в системе, которые, согласно информации из конфигурационных файлов, положены входящему в систему пользователю, программа вместо себя запускает пользовательский сеанс работы; при входе в систему через текстовую консоль это командный интерпретатор, при входе в графическом режиме — специальная программа (лидер сеанса), например, оконный менеджер. Роль ядра в этой процедуре сводится к обработке системных вызовов, изменяющих полномочия задачи по её просьбе (естественно, такие вызовы доступны только задаче, исходно выполнявшейся с правами администратора). Позже (см. §5.3.11) мы подробно рассмотрим соответствующие системные вызовы.

5.2. Ввод-вывод и файловые системы

С вводом-выводом мы уже давно и прочно знакомы, ведь без него не обходится ни одна программа. Напомним, что к вводу-выводу относится как работа с терминалом (пресловутые «ввод с клавиатуры» и «вывод на экран»), так и чтение/запись дисковых файлов, причём мы уже успели понять, что и на стандартных потоках ввода-вывода у нас вместо клавиатуры и экрана могут оказаться текстовые файлы, а наша программа этого не заметит.

Осваивая начальные навыки программирования на примере Паскаля, мы могли себе позволить не задумываться, как в действительности устроен ввод-вывод. Добравшись до языка ассемблера, мы были вынуждены обратить внимание, что программы, работающие под управлением мультизадачных операционных систем, все свои операции ввода-вывода могут проделать только через обращения к операционной системе, называемые системными вызовами. Рассматривая ввод-вывод с помощью стандартной библиотеки языка Си, мы отметили, что, хотя высокоуровневые механизмы этой библиотеки позволяют при необходимости производить «блочный» ввод и вывод произвольных порций данных, удобнее для этого пользоваться непосредственно системными вызовами.

Несмотря на весь наш опыт работы с файлами и потоками, многие возможности и особенности так и остались «за кадром». В этой главе мы попытаемся наверстать упущенное.

5.2.1. Знакомьтесь: файловая система

Под ***файлом*** в терминологии, связанной с компьютерами, обычно понимается некий набор хранимых на внешнем запоминающем устройстве данных, имеющий имя[10]. Часть операционной системы, отвечающая за хранение информации на внешних запоминающих устройствах в виде именованных файлов, называется ***файловой системой***.

Термин «*файловая система*» часто используется в совершенно ином значении, а именно для обозначения структур данных, создаваемых на внешнем запоминающем устройстве (то есть, попросту говоря, в секторах диска) для организации хранения на этом устройстве данных в виде именованных файлов. Такая структура данных обычно включает некий заголовок, хранящий наиболее общую информацию о файловой системе в целом (так называемый ***суперблок***), а также тем или иным способом выстроенное хранилище информации, какие сектора диска свободны, какие заняты, как называются файлы, хранящиеся в системе, какой набор секторов занимает каждый из файлов. Кроме того, для каждого файла файловая система может хранить разнообразную дополнительную информацию вроде типа файла, прав доступа, даты последней модификации и т. п.

Обычно из контекста можно однозначно определить, какая из двух «файловых систем» имеется в виду, так что путаницы такая перегрузка термина не создаёт. Тем не менее, стоит помнить о том, что под термином «файловая система» может пониматься как определённая структура данных на диске, так и набор подпрограмм в ядре операционной системы, и очевидно, что это совершенно разные вещи.

Ранние файловые системы позволяли давать файлам имена только ограниченной длины, причём эти имена находились в одном общем пространстве имён; естественно, от имён требовалась уникальность. Конечно, такой подход годится лишь до тех пор, пока файлов на одном устройстве сравнительно немного. С ростом объёма дисковых накопителей потребовалась более гибкая схема именования. Чтобы получить искомую гибкость, ввели понятие ***директории***, или ***каталога***[11].

Как мы уже упоминали (см. т. 1, сноску 15 на стр. 29), заменять термины «директория» и «каталог» словом «папка» (англ. *folder*) нельзя, для профессиональных программистов такая лексика неприемлема. **Папки — в шкафу.**

Каталог представляет собой особый тип файла, хранящий имена файлов, среди которых, возможно, тоже есть каталоги. При создании на диске файловой системы формируется один каталог, называемый ***корневым каталогом***. При необходимости можно создать дополнительные каталоги, а их имена записать в корневой. Такие каталоги иногда называют *каталогами первого уровня (вложенности)*. В них, в свою очередь, тоже можно создавать каталоги (*второго уровня*) и т. д.

Мы знаем, что при работе с операционной системой всегда некий каталог считается ***текущим***[12], в связи с чем операционная система позволяет использовать имена файлов различного вида — ***абсолютные***, ***относительные*** и ***краткие*** (как частный случай относительных); абсолютные имена начинаются с символа «`/`» (слэш), обозначающего корневой каталог, и от него же отсчитываются, так что никак не зависят от того, какой каталог сейчас текущий; краткие имена не содержат в себе ни одного слэша, то есть не упоминают никаких каталогов и подразумевают имя файла в текущем каталоге; относительные имена *отсчитываются* от текущего каталога, но могут в своём составе содержать одно и больше имён каталогов, включая ссылку на *каталог уровнем выше* (`..`). Если с именами файлов вы ощущаете малейшую неуверенность, обязательно вернитесь к первому тому, перечитайте §1.2.5 и попрактикуйтесь в работе с командной строкой, иначе дальнейший материал может вам не покориться.

В отличие от некоторых других операционных систем, в ОС Unix имена файлов организованы в виде единого дерева каталогов. В имя файла ни в каком виде не входит имя устройства, на котором этот файл находится, то есть ничего похожего на привычные для пользователей Windows обозначения `A:`, `C:` и т. п. в ОС Unix нет. Когда в системе имеется несколько дисков, файловая система одного из них объявляется ***корневой***, а остальные ***монтируются*** в тот или иной каталог, называемый ***точкой монтирования*** (англ. *mount point*), при этом для указания полных путей к файлам на этом диске нужно к полному имени файла в рамках диска добавить спереди полный путь точки монтирования. К примеру, если у нас есть флеш-брелок, на нём создан каталог `work`, в этом каталоге — файл `prog.c`, а сам брелок смонтирован с использованием каталога `/mnt/flash` в качестве точки монтирования, то полный путь к нашему файлу будет выглядеть так: `/mnt/flash/work/prog.c`.

В ОС Unix ***каталоги*** хранят только имя файла и некоторый номер, позволяющий идентифицировать соответствующий файл. Вся прочая информация о файле, включая его тип, размер, расположение на диске, даты создания, модификации и последнего обращения, данные о владельце файла и о правах доступа к нему связываются не с именем файла (как это делается в некоторых других операционных системах), а с вышеупомянутым номером. **Хранимая на внешнем запоминающем устройстве (диске) структура данных, содержащая всю информацию о файле, исключая его имя, называется *индексным дескриптором*** (англ. *index node*, или *i-node*). Индексные дескрипторы имеют номера, уникальные в рамках

[10] Ясно, что эта фраза не может претендовать на роль строгого определения; она дана лишь для того, чтобы зафиксировать основные нужные нам свойства файлов.

[11] В англоязычной литературе используется термин *directory* (читается «дайре́ктри»); слово «каталог» представляет собой более-менее точный перевод этого термина на русский, однако ещё в девяностые годы прошлого века слово «директория» вошло в действующий словарный состав русского языка, перестав быть жаргонным. Отметим, что слово «каталог» тоже имеет иностранное происхождение, просто было заимствовано раньше.

[12] Как мы увидим позже, текущий каталог — это свойство, которым обладает в системе каждый процесс независимо от других процессов.

файловой системы данного диска; как раз эти номера и записываются в каталоги вместе с именами файлов. Итак, в файловых системах, «родных» для ОС Unix, **каждая запись в каталоге состоит из двух полей: имя файла и номер индексного дескриптора**.

С некоторой натяжкой можно заявить, что индексный дескриптор — это и есть файл как таковой, но это будет верно лишь для некоторых специфических типов файлов, а в общем случае файл состоит из индексного дескриптора и некоторого количества *блоков данных*, расположенных где-то на диске. Информация об этих блоках тоже хранится в индексном дескрипторе. Можно также считать, что индексный дескриптор — это некая служебная часть файла, а сам файл представляет собой набор блоков на диске; но это тоже не всегда верно, ведь количество блоков, занимаемых файлом, может быть равно нулю, и тогда такой файл действительно состоит из одного только индексного дескриптора. Никто не мешает нам завести *пустой* файл (файл размера ноль); ни одного блока данных этот файл занимать не будет. Существуют такие типы файлов, которые никогда не занимают ни одного блока. Зато можно сказать совершенно определённо, что **имя файла в ОС Unix не является ни в каком виде частью или принадлежностью самого файла** — имя файла не указывается ни в его индексном дескрипторе, ни тем более в блоках данных, относящихся к файлу; единственное место, где мы встречаем имена файлов — это каталоги.

Отметим, что имя файла в ОС Unix может быть достаточно длинным (обычно ограничение составляет 255 символов) и содержать, вообще говоря, любые символы, кроме нулевого и символа-разделителя. Так, имя файла из пятнадцати точек является в Unix вполне допустимым. Тем не менее, настоятельно не рекомендуется использование в именах файлов таких символов, как пробел, звёздочка, восклицательный и вопросительный знаки, хотя это и возможно. Также рекомендуется воздержаться от использования в именах файлов спецсимволов, таких как перевод строки, табуляция, звонок, backspace и пр., и символов с кодом, превышающим 127, таких как русские буквы. Наконец, имя файла не стоит начинать с символа «-» (минус). Несоблюдение этих рекомендаций приводит к возникновению проблем в работе. Эти проблемы всегда могут быть преодолены, однако преодолимость трудностей не следует считать поводом для их создания.

В ОС Unix допускается, чтобы два или более имён файлов, расположенных как в разных каталогах, так и в одном, ссылались на один и тот же номер индексного дескриптора. Конечно, файл всегда создаётся под каким-то определённым именем, но позже он может получить дополнительные имена — записи в каталогах, ссылающиеся на тот же самый индексный дескриптор. Такие имена называются ***жёсткими ссылками*** (англ. *hardlinks*). Отличить жёсткую ссылку от оригинального имени файла невозможно: эти имена совершенно равноправны, то есть на самом деле изначальное имя файла — это *тоже* жёсткая ссылка. Очевидно, что жёсткая ссылка может быть установлена только в пределах одного диска — точнее, в пределах одной ***файловой системы***, если под таковой понимать структуру данных на диске; в самом деле, нумерация индексных дескрипторов у каждого диска своя, так что сослаться на индексный дескриптор другого диска не представляется возможным, для этого в структуре данных файла-каталога просто не предусмотрено полей. В индексном дескрипторе содержится, кроме всего прочего, *счётчик количества ссылок* на данный дескриптор. При создании файла этот счётчик устанавливается в единицу, при создании новой жёсткой ссылки — увеличивается на единицу, при удалении ссылки — уменьшается.

Интересно, что в ОС Unix не предусмотрено функции *удаления файла*; вместо этого имеется системный вызов `unlink`, *удаляющий ссылку*. При выполнении этого вызова имя удаляется из каталога, а счётчик ссылок в соответствующем индексном дескрипторе уменьшается. **Сам файл удаляется, только если удалённая ссылка была последней (счётчик обратился в нуль) и при этом файл не был ни одним из процессов открыт на запись или чтение.** Если счётчик обратился в нуль, но файл кем-то открыт, удалён он будет только после закрытия. Название `unlink` распространилось и на другие системы — функция, удаляющая файл, обычно называется именно так, что часто вызывает удивление у программистов, плохо знакомых с ОС Unix. Подробно этот и другие системные вызовы для работы с файлами мы рассмотрим в §5.2.4.

Для создания жёсткой ссылки средствами командной строки можно воспользоваться командой `ln`. Она похожа на хорошо знакомую нам команду `cp`, но не копирует файл, а создаёт для него новое имя.

Каталоги (они же — ***директории***) в файловой системе ОС Unix представляют собой не более чем *файлы*, пусть и специального типа. Информацию, содержащуюся в каталоге — имена файлов и номера индексных дескрипторов — нужно где-то хранить, так что файлу типа каталог, как и обычному файлу, должны принадлежать дисковые блоки, у него должен быть размер и т. п. На низком уровне каталог отличается от обычного файла только значением признака типа в индексном дескрипторе; в остальном хранение на диске каталога организовано точно так же, как и хранение обычного файла. Обычными файлами и каталогами многообразие типов файлов в ОС Unix не исчерпывается, постепенно мы познакомимся со всеми остальными типами файлов. Забегая вперёд, скажем, что всего их существует семь, но обычный файл и каталог встречаются чаще других.

Особая роль каталогов накладывает определённые ограничения на то, что с ними можно делать. Во-первых, каталог невозможно удалить, как обычный файл, для этого применяется особый системный вызов `rmdir`, который сработает успешно лишь в случае, если удаляемый каталог пуст. Существует также одноимённая команда командной строки. Во-вторых, **система запрещает создание жёстких ссылок на каталоги**. Дело в том, что создание таких жёстких ссылок может привести к возникновению ориентированных циклов в дереве каталогов: например, такой цикл получился бы после выполнение команд

```
mkdir a; cd a; mkdir b; cd b; ln ../../a ./c
```

если бы, конечно, система позволила их выполнить. В этой ситуации попытка рекурсивно пройти директорию `a`, скажем, с целью подсчёта количества файлов в ней, привела бы к зацикливанию. Кроме того, оказалось бы, что директорию `a` невозможно удалить, ведь она всегда что-то содержит (косвенно она содержит сама себя). По этой причине жёсткие ссылки на директории запрещены на уровне ядра операционной системы, и никакие права доступа не позволяют этот запрет обойти.

Символическая ссылка (англ. *symbolic link*) представляет собой файл специального типа, содержащий *имя* другого файла. Операция открытия символической ссылки на чтение или запись приводит на самом деле к открытию файла, на который она ссылается, а не её самой. В отличие от жёсткой ссылки, символическая ссылка легко отличима от основного имени файла, поскольку представляет собой самостоятельный файл особого типа; это позволяет не накладывать ограничений на установление символических ссылок на каталоги. Будучи файлом, символическая ссылка имеет свой собственный номер индексного дескриптора. Создание и удаление символической ссылки никак не затрагивает ни файл, на который она ссылается, ни его индексный дескриптор. Более того, файл, на который указывает ссылка, может вообще не существовать в момент ее создания, а также может быть удалён позднее, что никак не повлияет на ссылку. Символическая ссылка, содержащая имя несуществующего файла, называется ***висячей*** (англ. *dangling*, более точный буквальный перевод — «болтающаяся»). Висячие ссылки считаются вполне корректным явлением и в некоторых случаях целенаправленно используются. Символические ссылки не ограничены рамками одного диска, поскольку реализуются через хранение имён, а дерево имён, как мы знаем, в ОС Unix общее для всех имеющихся в системе дисковых устройств.

Для создания символической ссылки средствами командной строки следует использовать уже знакомую нам команду `ln`, указав ей флаг `-s`:

```
ln -s /path/to/old/name new_name
```

С символическими ссылками связан один любопытный момент. В индексном дескрипторе довольно ощутимое пространство выделено под информацию о номерах блоков, занимаемых данным файлом; конкретный размер этого пространства различается от системы к системе, но он вряд ли где-то будет меньше 52 байт[13], а на современных дисках с их огромными размерами может быть гораздо больше. В то же время на символическую ссылку никогда не нужно больше одного блока, имена просто не бывают такими большими. Даже один блок расходовать несколько жаль, ведь из всего блока (например, 4-килобайтного) будет использоваться в большинстве случаев несколько десятков байт. Поэтому символические ссылки реализованы довольно экзотическим способом. Если длина имени, хранимого в ссылке, такова, что её можно

[13] Десять номеров обычных блоков и три номера *косвенных* блоков, каждый номер занимает четыре байта; косвенные блоки мы обсудим в последней части тома.

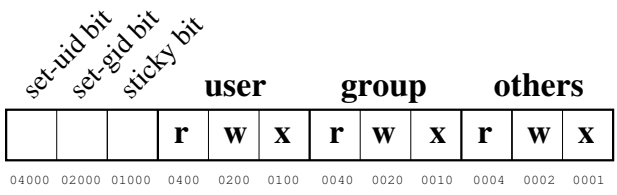

Рис. 5.2. Права доступа к файлу

уместить в индексный дескриптор, заняв пространство, обычно используемое для информации о номерах блоков, то именно там это имя и располагается, не занимая лишнего блока, то есть файл такой ссылки состоит из одного лишь индексного дескриптора. Если же имя в дескриптор не помещается, система выделяет под такую ссылку отдельный блок.

Позже мы познакомимся с другими типами файлов, которые вообще никогда не занимают дискового пространства за пределами своего индексного узла. Символическая ссылка в этом плане интересна тем, что она может как занимать, так и не занимать отдельный блок.

5.2.2. Права доступа к файлам

Права доступа к файлу (англ. *access permissions*) определяют, кто из пользователей (точнее, процессов, выполняющихся от имени пользователей) какие операции может с данным файлом произвести; мы уже обсуждали права доступа в первом томе (см. §1.2.13), но многие подробности оставили за кадром, а сейчас постараемся это исправить.

Для каждого файла в его индексном дескрипторе хранятся идентификатор пользователя-владельца (`uid`, *user id*) и идентификатор группы пользователей (`gid`, *group id*). Сами по себе права доступа к файлу, также хранящиеся в индексном дескрипторе, представляются в виде 12-битного слова (см. рис. 5.2). Младшие девять бит этого слова объединены в три группы по три бита; каждая группа задаёт права доступа для владельца файла, для группы владельца и для всех остальных пользователей. Три бита в каждой группе отвечают за право чтения файла, право записи в файл и право исполнения файла.

Поскольку права доступа состоят из групп битов по три в каждой, логично применять для их записи восьмеричную систему счисления. В большинстве случаев для записи прав доступа хватает трёх восьмеричных цифр, старшая обозначает права для владельца, следующая — для группы, младшая — для всех остальных. Значения отдельных двоичных разрядов в таком числе соответствуют показанным на рисунке: `4` — права на чтение, `2` — права на запись, `1` — права на исполнение. Например, число $\texttt{644}_8$ означает права на чтение и запись для владельца файла, права на чтение для группы и всех остальных; число $\texttt{751}_8$ означает, что для владельца установлены права на чтение, запись и исполнение, для группы — на чтение и исполнение, для всех остальных — только на исполнение.

В случае директории биты исполнения указывают, кто (владелец, член группы, обычный пользователь) имеет право доступа к файлам из директории — например, может открыть файл, используя имя из этой директории (при условии, что ему также хватит прав на сам файл). Так, если у нас есть право на чтение директории, но нет права на её «исполнение», то мы сможем узнать имена файлов, содержащиеся в директории, но ни одним из них не сможем воспользоваться; напротив, если у нас есть права на «исполнение», но нет прав на чтение, то прочитать имена файлов в этой директории мы не можем, а воспользоваться сможем лишь такими файлами, имена которых нам стали известны каким-то другим путём, например, их нам просто сказали (при условии, что у нас хватит прав также и на сами файлы). Чаще всего права на чтение и «исполнение» на директорию устанавливаются одинаковыми; реже можно встретить директории, к которым у тех или иных категорий пользователей есть права на «исполнение», но нет прав на чтение; этим вариантом можно воспользоваться, например, чтобы из определённой директории наши коллеги могли забрать (прочитать и скопировать) те файлы, имена которых мы им скажем. Обратная ситуация (чтение разрешено, исполнение запрещено) на практике бессмысленна и не используется.

Оставшиеся три старших разряда прав доступа называются SetUid bit (04000), SetGid bit (02000) и Sticky bit (01000); они имеют достаточно специфическое назначение, различное для исполняемых файлов и директорий; для прочих файлов эти биты обычно не устанавливаются, поскольку ни на что не влияют. Для исполняемых файлов установленный бит SetUid означает, что при запуске такого файла программа, записанная в нём, будет выполняться не как обычно, с правами того, кто её запустил, а *с правами владельца файла*; аналогичным образом программа, для исполняемого файла которой установлен бит SetGid, при выполнении получает «групповые полномочия» в соответствии с пользовательской группой, которой принадлежит файл.

Чаще всего SetUid bit устанавливают на исполняемые файлы, принадлежащие администратору. К числу таких программ относится, например, `passwd` — хорошо знакомая вам программа для смены пароля; впрочем, на некоторых системах `passwd` уже перестала быть suid'ной.

Довольно интересна история Sticky bit. Само слово *sticky* переводится как «липкий»; в очень старых версиях ОС Unix этим битом помечали часто используемые исполняемые файлы, чтобы сообщить ядру, что после окончания выполнения такой программы её сегмент кода не следует удалять из оперативной памяти, поскольку, скорее всего, кто-нибудь в ближайшее время запустит её снова. Машинный код такой программы, будучи единожды загружен в оперативную память, как бы «прилипал» к ней. В современных системах такая оптимизация лишена смысла, поскольку ядро буферизует весь файловый ввод-вывод, так что если некоторый исполняемый файл будет использоваться достаточно часто, его содержимое в любом случае окажется в буферах ядра; поэтому Sticky bit, даже если его установить, будет системой проигнорирован. В наше время он используется только для директорий и имеет совершенно иной смысл, который мы опишем чуть позже.

Ещё одно важное замечание связано с тем, что в системах Unix исполняемые файлы бывают двух видов: обычные, содержащие машинный код в определяемом системой формате, и *скриптовые* — текстовые, содержащие программу на некотором интерпретируемом языке программирования. Со скриптами на языке Bourne Shell мы уже встречались в первом томе (см. §1.2.15) и знаем, что такой файл должен начинаться со строки вроде «`#!/bin/sh`», которая указывает системе, что для исполнения этого файла нужно запустить программу `/bin/sh` (командный интерпретатор) и передать ему имя файла-скрипта через аргумент командной строки. В роли интерпретатора может выступать любая программа, способная обработать скрипт, имя которого ей передали параметром командной строки — даже `/bin/cat`; такой «скрипт», естественно, просто напечатает собственное содержимое. Можно даже указать для интерпретатора аргумент командной строки, правда только один, больше система передавать не умеет (ещё одним параметром будет имя скрипта). Достаточно «навесить» на скриптовой файл бит исполнения, и его можно будет запускать в командной строке как обычную программу.

Так вот, **биты SetUid и SetGid для скриптовых исполняемых файлов не работают**; это ограничение введено искусственно, из соображений безопасности.

Чтобы понять, в чём заключается проблема с безопасностью suid-скриптов, представьте себе, что в вашей системе есть скрипт (неважно, на каком языке), который должен выполняться с правами его владельца — например, пользователя `root` (системного администратора). При этом некто получил (легитимно или нет) доступ к вашей системе как обычный пользователь, имеющий право исполнять этот скрипт, и хочет расширить свои полномочия.

При запуске скрипта на исполнение система вынуждена сначала обратиться к файлу скрипта, открыть его на чтение, чтобы получить из него информацию об имени используемого интерпретатора; затем она закрывает файл скрипта и запускает командный интерпретатор, который, в свою очередь, откроет файл скрипта на чтение.

Пользователь, имеющий право на исполнение скрипта, для которого установлен бит SetUid, может атаковать систему, написав свой собственный скрипт на том же языке, делающий что-то совершенно другое — например, открывающий какой-нибудь канал для связи и исполняющий произвольные команды, присланные через этот канал. Дальше атакующий создаёт символическую ссылку на ваш скрипт, а затем запускает (естественно, не вручную, а с помощью им же написанной программы) ваш скрипт через эту ссылку и тут же меняет ссылку так, чтобы она указывала на его собственный скрипт. После достаточного числа попыток он может поймать такой момент, когда система уже успела прочитать имя командного интерпретатора и запустить этот интерпретатор, но сам он ещё не успел открыть на чтение файл скрипта; в этом случае с правами администратора (или другого пользователя — того, который числится владельцем скрипта) будет выполняться не ваш скрипт, а тот, который написал атакующий, и таким образом он получит возможность выполнять произвольные действия с правами, которых вы ему не давали. Именно поэтому система игнорирует биты SetUid и SetGid при исполнении скриптов.

Для директорий старшие биты прав доступа имеют иное значение, нежели для исполняемых файлов, причём применять их имеет смысл только для директорий, публично доступных на запись. Чаще всего используется `Sticky bit`: если он установлен, то в такой директории пользователи имеют право удалять *только свои собственные файлы*, несмотря на наличие прав на запись в директорию. Обычно для создания «общественной» директории

применяются права `01777` — всем разрешается всё, но при этом взведённый `Sticky bit` не позволяет удалять чужие файлы. Именно такие права устанавливаются на директорию `/tmp`, предназначенную для временных файлов и доступную всем пользователям системы.

`SetGid bit`, установленный на директорию, приводит к тому, что для файлов, создаваемых в такой директории, в качестве *группы пользователей* принудительно устанавливается та же группа, что и для самой директории, вне зависимости от того, в какие группы входит создатель файла. Такие права применяются, например, для директории `/var/spool/mail`, содержащей почтовые ящики пользователей системы; это позволяет почтовой подсистеме, для которой создана специальная группа (обычно называемая просто `mail`), осуществлять доступ ко всем почтовым ящикам, кто бы их ни создал. Если в такой директории создать поддиректорию, она унаследует не только группу пользователей своей родительской директории, но и её `SetGid bit`, так что внутри поддиректорий новые файлы тоже будут наследовать ту же самую группу.

`SetUid bit` для директорий в большинстве систем (в том числе в Linux) игнорируется, но, например, в системе FreeBSD имеет эффект, аналогичный вышеописанному эффекту `SetGid bit`, то есть все файлы в такой директории создаются принадлежащими её владельцу вне зависимости от того, кто из пользователей создаёт этот файл.

Некоторые системы поддерживают специальную комбинацию битов: если на файле установлен `SetGid bit`, но при этом сброшен бит исполнения для группы, считается, что для такого файла должны применяться так называемые *обязательные захваты* (англ. *mandatory locking*). Вопросы, связанные с «захватом» доступа к файлам, мы будем рассматривать в части, посвящённой работе с разделяемыми данными; впрочем, обязательные захваты всё равно никто не использует.

Узнать права доступа к конкретному файлу можно знакомой нам командой `ls -l`, которая для их получения использует системный вызов `stat`, а изменить права — командой `chmod` (сокращение английских слов *change mode* — смена режима), системный вызов называется так же. Сменить владельца и группу для файла в командной строке можно командами `chown` (*change owner*) и `chgrp` (*change group*), а системный вызов для этих целей используется один — `chown`. Системные вызовы `stat`, `chmod` и `chown` мы рассмотрим в §5.2.4.

Отметим ещё один довольно неочевидный момент. Для символических ссылок права обычно игнорируются, то есть ни на что не влияют; при выполнении всех операций используются права доступа того файла, на который ссылка ссылается. Этот момент станет более понятен, если мы заметим, что *изменить* права доступа для символической ссылки мы не смогли бы, даже если бы очень хотели: системный вызов `chmod` и одноимённая команда командной строки, если их применить к символической ссылке, изменят права не для неё самой, а для файла, на который она ссылается.

5.2.3. Чтение и запись содержимого файлов

Как отмечалось в §5.1.5, одно из центральных мест в принципах (если угодно, в философии) систем семейства Unix занимает понятие *потока байтов*; именно в виде таких потоков оформляется едва ли не любая[14] передача массивов информации. Изучая язык Си, мы познакомились с тем, как потоки ввода-вывода выглядят на уровне «высокоуровневых» функций стандартной библиотеки, использующих для идентификации потоков переменные типа `FILE*`; в §4.4.7 мы отметили, что, хотя библиотека и содержит функции блочного ввода-вывода, их применение не имеет большого смысла, поскольку, в отличие от работы с текстами и побайтовой работы с потоками произвольной природы, не имеет никаких преимуществ в сравнении с моделью ввода-вывода, реализованной на уровне системных вызовов.

С системными вызовами для ввода-вывода мы уже сталкивались при изучении языка ассемблера (см. т. 1, §3.6.7). В программах на Си эти вызовы доступны в виде функций-обёрток, которые так и называются `open`, `read`, `write`, `lseek` и `close`. В системах, отличных от Unix, эти функции могут не быть системными вызовами, но работать будут точно так же.

Системные вызовы не используют никаких специальных типов для файловых переменных; вместо этого они различают потоки ввода-вывода *по номерам*, называемым ***файловыми дескрипторами***. Это небольшие неотрицательные целые числа: `0`, `1`, `2`, `3` и так далее. Пусть вас не смущает постоянно используемое и устоявшееся словосочетание «файловый дескриптор»; дескрипторы могут быть связаны не только с открытыми дисковыми файлами, но и с потоками ввода-вывода произвольной природы, многие из которых вообще не имеют к файлам никакого отношения.

Дескрипторы открытых файлов ни в коем случае не следует путать с *индексными дескрипторами*, это совершенно разные и никак между собой не связанные сущности. В английском языке слово *descriptor* применяется только для обозначения дескрипторов открытых файлов, термин же *индексный дескриптор* представляет собой пример неудачного (но, к сожалению, прижившегося) перевода: оригинальный англоязычный термин *index node* вообще не содержит слова *descriptor* и буквально может быть переведен как *индексный узел*.

Как мы уже знаем (см. т. 1, §3.6.7), дескрипторы `0`, `1` и `2` соответствуют стандартным потокам (соответственно ввода, вывода и диагностики); как правило, при запуске программы эти дескрипторы уже открыты и готовы к работе, но кроме них, никаких открытых потоков ввода-вывода исходно нет. Особая роль дескрипторов `0`, `1` и `2` никоим образом не означает, что мы не можем использовать их для своих целей или связать с другими потоками ввода-вывода, в частности, с файлами. В действительности ядро системы вообще не приписывает этим дескрипторам никакой особой роли, вся их «стандартность» — это лишь соглашение, которому следуют авторы программ и библиотек, работающих в пользовательском режиме. Подробнее к этому вопросу мы ещё вернёмся.

Конечно, все эти целые числа — лишь номера; «настоящий» файловый дескриптор располагается в ядре операционной системы и представляет собой достаточно сложную структуру данных, зависящую от конкретной реализации ядра и содержащую всю информацию, нужную ядру для поддержки работы с данным потоком ввода-вывода. Помнить, какому из них соответствует тот или иной номер (то самое целое число, больше нам ничего не показывают) — тоже забота ядра. Заметим, что номер дескриптора локален по отношению к процессу: скажем, дескриптор №5 может в одном процессе быть связан с одним файлом, в другом — с совсем другим, а в третьем и вовсе не соответствовать никакому потоку ввода-вывода.

Один из наиболее очевидных (хотя, как мы позже увидим, далеко не единственный) способ создать новый поток ввода-вывода — открыть файл вызовом `open`. Приготовьтесь к тому, что рассказ об этом вызове будет долгим, нам даже придётся ввести ещё один системный вызов, а потом уже завершать рассказ про `open`. Итак, в зависимости от ситуации `open` вызывается с двумя или тремя параметрами, то есть в соответствии с одним из двух возможных профилей:

```
int open(const char *name, int mode);
int open(const char *name, int mode, int perms);
```

Параметр `name` задаёт имя файла, который мы хотим открыть. Как обычно, это может быть короткое имя файла, находящегося в текущем каталоге, или же путь к файлу, как полный, так и относительный (см. обсуждение в §5.2.1). Параметр `mode` задаёт режим, в котором мы намерены работать с файлом. Режим задаётся в виде целого числа, отдельные биты которого обозначают те или иные особенности предстоящей работы с файлом; каждый такой бит представлен предопределённой целочисленной константой, а комбинации из нескольких таких битов мы получаем, применяя операцию побитового «или» («`|`»). Основными режимами считаются `O_RDONLY` (только чтение), `O_WRONLY` (только запись) и `O_RDWR` (чтение и запись). Из этих трёх констант нужно указать одну и только одну.

Кроме трёх основных констант, существуют еще и *модифицирующие* константы, которые можно добавить к основным, используя операцию побитового «или». К ним относятся:

- `O_APPEND` — открыть файл на добавление в конец: каждая операция записи будет осуществлять запись в конец файла;
- `O_CREAT` — если файла не существует, разрешить операционной системе его создание;
- `O_TRUNC` — если файл существует, перед началом работы уничтожить его старое содержимое, в результате чего длина файла станет нулевой; запись начнется с начала файла;
- `O_EXCL` (используется только в сочетании с `O_CREAT`) — потребовать от операционной системы создания нового файла; если файл уже существует, не открывать его и выдать ошибку;

[14] Существуют исключения и из этого правила; так, передача данных из одного процесса в другой через очереди сообщений System V и через области разделяемой памяти совсем не похожа на потоки байтов. В unix-системах эти способы применяются во много раз реже, чем традиционные потоки ввода-вывода.

- `O_NONBLOCK` — открыть файл в ***неблокирующем режиме***, при котором системные вызовы чтения и записи никогда не ждут готовности, а возвращают управление сразу же — возможно, при этом выдают ошибку, если немедленной готовности не было.

Существуют также другие флаги; некоторые из них мы рассмотрим позже.

Третий параметр (`perms`) задаёт права доступа на случай, если при открытии будет создан новый файл. Этот параметр следует указывать только тогда, когда значение второго параметра предполагает возможность создания файла, то есть если используется `O_CREAT`; при открытии существующих файлов параметр `perms` не имеет никакого смысла и просто игнорируется, так что, если создание нового файла не предполагается, функцию `open` нужно вызывать от двух параметров, а третий просто опустить. Напротив, при наличии `O_CREAT` параметр `perms` надо обязательно указать, иначе результат вам может не понравиться.

Права доступа к файлам в ОС Unix мы подробно рассмотрели в предыдущем параграфе, но чтобы понять, какие числа следует указывать в качестве значений параметра `perms`, нужно знать ещё одну важную особенность системы. Для каждого процесса ядро системы поддерживает некое хитрое число, называемое `umask`. При порождении новых процессов и запуске одних программ другими это число обычно наследуется, так что если пользователь (или системный администратор) установит для этого числа то или иное значение, обычно все запускаемые программы будут работать с тем же самым значением `umask`. **При создании новых файлов биты, установленные в параметре `umask`, *удаляются* из прав доступа к файлу, которые процесс, создающий файл, указал в параметре своего обращения к системному вызову**; как мы увидим позже, это касается не только вызова `open`, то же самое происходит при создании директорий и других файловых объектов специального вида.

Например, в системе, предназначенной «для своих», пользователи чаще всего устанавливают `umask` равным `022`, так что для создаваемых файлов права на запись для всех, кроме владельца, сбрасываются, а возможность читать и выполнять предоставляется всем желающим. Более осторожные пользователи могут установить `umask` равным `027`, запретив любой доступ к новосоздаваемым файлам для всех пользователей, не входящих в ту же группу, или вообще `077`, что соответствует сбрасыванию всех прав доступа для всех, кроме владельца. Вы можете сами поэкспериментировать с разными значениями `umask`; в командной строке этот параметр меняется командой, которая так и называется `umask` (дав её без параметров, вы узнаете текущее значение; указав параметр — установите новое), но действует это только на текущий сеанс работы с данным конкретным экземпляром командного интерпретатора, поскольку он устанавливает этот параметр для себя, а все запускаемые из него программы установленное значение наследуют. Чтобы изменить `umask` для всей системы, придётся редактировать конфигурационные файлы, конкретика зависит от используемого дистрибутива. Заодно скажем, что любой процесс может изменить свой собственный `umask` с помощью системного вызова, который, естественно, тоже называется `umask`:

```
int umask(int mask);
```

В качестве своего значения вызов возвращает значение `umask`, установленное ранее; ошибочных ситуаций для него не предусмотрено. Отметим, что во внимание принимаются только девять младших битов параметра; биты SetUid, SetGid и Sticky «замаскировать» невозможно.

Вернёмся к вызову `open` и его третьему параметру (`perms`). Файлы, которые мы создаём, обычно не нуждаются в правах на исполнение, то есть нам остаются лишь права на запись и чтение. Как мы только что сказали, пользователь может, устанавливая параметр `umask`, сам выбрать, какой режим защиты для его файлов будет ему наиболее удобен; поэтому **при создании обычного файла данных следует задавать значение прав доступа** `0666` — права на чтение и запись (но не на исполнение) для владельца, группы и всех остальных (напомним, что в языке Си лидирующий ноль означает восьмеричную константу). Обычно права на запись для группы и остальных пользователей автоматически исчезают после вычитания `umask`, если же этого не происходит — значит, так хотел пользователь. Наряду со значением `0666` достаточно часто используется `0600`, запрещающее (независимо от значения `umask`) любой доступ к создаваемому файлу для кого-либо, кроме его владельца. Так следует поступать, если создаваемый файл предназначен для хранения конфиденциальной информации — паролей, номеров банковских карточек и т. п. **Значения параметра `perms`, отличные от `0666` и `0600`, практически никогда не встречаются**.

Вызов `open` возвращает значение `-1` в случае ошибки, а в случае успеха — тот самый ***файловый дескриптор***, связанный с только что созданным в системе (по просьбе нашего процесса) потоком ввода-вывода.

Рассказывая в §4.4.3 о функции `fopen`, мы отметили, что у неё существует очень много разнообразных поводов ошибиться, и большинство из них автору программы никак не подконтрольны. В действительности это всё относится скорее к вызову `open`, к которому `fopen`, естественно, обращается для создания нового потока. Повторим ещё раз: **открытие файла — это такое действие, для которого проверка его успешности (и обработка ошибочных ситуаций) абсолютно обязательна**. Если после вызова `open` в программе нет проверки возвращённого им значения на равенство `-1`, то эта программа заведомо неправильна.

Чтение из потока ввода и *запись* в поток вывода производится системными вызовами `read` и `write`:

```
int read(int fd, void *mem, int len);
int write(int fd, const void *data, int len);
```

Первый параметр задаёт файловый дескриптор; второй параметр указывает на область памяти, в которой следует разместить прочитанные данные (для `read`) либо из которой нужно взять данные для передачи в поток (для `write`); последний параметр указывает размер этой области памяти, то есть задаёт количество данных, которые нужно прочитать или записать.

Вызов `read` пытается прочитать из заданного потока заданное количество данных. Если в указанном потоке отсутствуют данные, готовые к прочтению, вызов заблокирует вызвавший процесс до тех пор, пока данные не появятся, и только после их прочтения вернёт управление. Если данные присутствуют, но их меньше, чем требует вызывающий, вызов сохранит их в области памяти по адресу `mem` и вернёт управление; подчеркнём, что **вызов `read` немедленно возвращает управление, если может отдать вызвавшей программе хотя бы один байт данных**, то есть он не станет ждать, пока данных накопится столько, сколько просили. Если бы это было устроено иначе, мы не могли бы писать интерактивные программы.

Вызов `write` пытается записать в поток заданное количество информации, расположенной по заданному адресу. При работе с некоторыми «особенными» потоками вывода вызов `write` может заблокировать вызвавший процесс до тех пор, пока в потоке (точнее, в буфере внутри ядра) не появится свободное место. Например, так может случиться, если поток предполагает отправку данных через сеть на другой компьютер, но скорость, с которой вы пытаетесь передавать данные в поток, превышает скорость работы сети. То же самое произойдёт, если ваш поток вывода представляет собой канал связи с другим процессом и этот другой процесс, занятый своими делами, не читает данные из канала. При записи в обычные дисковые файлы блокировки почти никогда не случается, диски для этого достаточно быстры.

В случае ошибки `read` и `write` возвращают `-1`. В случае успешного выполнения чтения или записи возвращается положительное число, означающее количество переданных байтов информации. Естественно, это число не может быть больше `len`. **При наступлении ситуации «конец файла» `read` возвращает ноль**. Для обычного файла это означает, что мы дочитали его до конца и больше там читать нечего; при чтении с клавиатуры ситуация конца файла означает, что пользователь сымитировал её нажатием `Ctrl-D`; для потоков иной природы будет иной и природа ситуации «конец файла» — например, на сетевом сокете она наступает, если наш партнёр разорвал соединение.

Напомним, что анализ значения, возвращаемого `read` как функцией, строго обязателен. В программе не должно быть операторов вроде

```
read(fd, buf, sizeof(buf));
```

Правильный вызов `read` выглядит, например, так:

```
count = read(fd, buf, sizeof(buf));
```

Для вызова `write` всё не так строго, ведь не проверяли же мы в наших программах результат работы `printf`. Впрочем, вывод в стандартный поток вывода — это скорее исключение из правил, в остальных же случаях значение,

возвращённое функцией `write`, лучше проверить. Чаще всего число записанных байтов в точности равняется значению `len`, однако полагаться на это опасно. Можно назвать две ситуации, при которых `write` заведомо вернёт число, мéньшее значения параметра `len`. Первая — когда мы пишем в простой дисковый файл, и на диске заканчивается свободное место (очередной вызов запишет меньше, чем мы просили, а следующий после него вернёт ошибку); вторая — когда наш поток связан с сетевым соединением, *объявлен неблокирующим* и мы попытаемся за один вызов отправить в сеть столь крупную порцию данных, что она не поместится в буфер ядра операционной системы (например, несколько мегабайт). Система при этом примет у нас для последующей отправки в сеть столько информации, сколько поместится в её буферной памяти.

Надо сказать, что если не перевести поток в неблокирующий режим, то вызов `write` в большинстве систем предпочтёт отправить в поток все данные, которые вы ему предоставили, и вернуть управление только после этого. Напомним, что поток будет неблокирующим, если при его открытии вызовом `open` воспользоваться флагом `O_NONBLOCK`; система также позволяет перевести в неблокирующий режим уже открытый поток с помощью системного вызова `fcntl`, и для потоков, создаваемых иными способами, нежели `open`, это может оказаться единственным вариантом. Вызов `fcntl` мы подробно рассмотрим чуть позже.

Вызов `write` в современных системах не возвращает ноль, если только параметр `len` не был равен нулю (но так делать не следует в любом случае). В прошлом существовали системы семейства Unix, в которых `write` мог вернуть ноль в неких специфических обстоятельствах; современные системы в таких случаях всегда возвращают `-1`.

Для изменения *текущей позиции* в открытом файле используется вызов `lseek`, очень похожий на функцию `fseek`, которую мы обсуждали на стр. 47.

```
long lseek(int fd, long offset, int whence);
```

Параметр `fd` здесь задаёт номер файлового дескриптора; как и для `fseek`, параметр `offset` указывает, на сколько байтов следует сместиться, а параметр `whence` определяет, от какого места эти байты следует отсчитывать: от начала файла (`SEEK_SET`), от текущей позиции (`SEEK_CUR`) или от конца файла (`SEEK_END`). Вызов возвращает, как и `fseek`, новое значение текущей позиции, считая от начала файла. Например, `lseek(fd, 0, SEEK_SET)` установит текущую позицию на начало файла, `lseek(fd, 0, SEEK_END)` — на конец файла. Вызов `lseek(fd, 0, SEEK_CUR)` никак позицию не изменит, но его можно использовать, чтобы узнать текущее значение позиции. Прочитать последние 100 байт файла можно так:

```
int rc;
char buf[100];
/* ... */
lseek(fd, -100, SEEK_END);
rc = read(fd, buf, 100);
```

При смене позиции с помощью `lseek` можно зайти за конец файла. Само по себе это не приводит к изменению размера файла, но если после этого произвести запись, размер файла увеличится; конечно, файл при этом должен быть открыт в режиме, допускающем запись. При этом возможно образование «дырки» (англ. *hole*) между последними данными перед старым концом файла и первыми данными, записанными в новой позиции. Ничего страшного в этом нет, «дырки» считаются штатной возможностью файловых систем. При чтении из таких мест файла мы получаем массивы нулевых байтов, так что если нам по каким-то причинам нужен файл, значительная часть которого забита нулями, можно не тратить физическое место на диске для хранения таких «нулевых областей».

Напомним, что не все потоки ввода-вывода поддерживают понятие *текущей позиции*; потоки, к которым можно применить вызов `lseek` — в частности, потоки, связанные с обычными файлами на диске — называются ***позиционируемыми*** (соответствующий английский термин — *seekable*). К этому свойству потоков мы ещё раз вернёмся в §5.2.5.

Отдельного замечания заслуживает тип параметра `offset`, а также тип значения, возвращаемого самим вызовом. В официальной документации используется имя типа `off_t`, специально предназначенное для этой цели (имя образовано от слов *offset type*). На 32-битных системах этот тип представляет собой 32-битное знаковое целое — обычно он вводится как синоним типа `long`, который, как мы помним, на 32-битных машинах совпадает с `int`. Это означает, что вызов `lseek` в 32-битных системах не мог работать с файлами размером свыше 2 GB; до не слишком давних пор такие файлы не встречались на практике, поскольку не существовало столь «огромных» дисков. Впрочем, ещё в середине девяностых появление DVD, имеющих ёмкость 4,7 GB (при том, что жёсткие диски на тот момент по своей ёмкости были скромнее) сделало этот лимит неадекватным реальности.

Изменять устоявшийся интерфейс системного вызова, к счастью, не стали, вместо этого был введён другой интерфейс:

```
long long lseek64(int fd, long long offset, int whence);
```

(официальная документация использует идентификатор `off64_t` для обозначения `long long`). В этой своей версии вызов работает точно так же, как и обычный `lseek`, отличается он только типом второго параметра и возвращаемого значения. Примечательно, что на 64-битных платформах `long` совпадает с `long long`, так что в 64-битном случае эти два интерфейса *вообще ничем не отличаются*.

После окончания работы с файлом его следует закрыть. Это особенно важно, поскольку дескрипторы — ресурс ограниченный: их общее количество в системе не может превышать некоторого числа, как и количество дескрипторов, открытых одним процессом. Закрытие файла производится вызовом `close`:

```
int close(int fd);
```

— где `fd` — дескриптор, подлежащий закрытию. Вызов возвращает ноль в случае успеха, `-1` в случае ошибки, но даже если произошла ошибка, **дескриптор после вызова `close` никогда не остаётся открытым**.

Многие программисты привыкли игнорировать значение, возвращаемое вызовом `close`; такая практика имеет под собой определённые основания, ведь дескриптор закрывается даже в том случае, если вызов вернул ошибку. Несмотря на это, игнорировать значение, возвращаемое вызовом `close`, не следует, особенно если мы *записывали* данные в поток, открытый на вывод или на ввод и вывод одновременно. Дело тут в том, что операционная система может ради повышения производительности применять свою собственную буферизацию: получив от нашей программы данные, которые нужно записать в файл, она сразу же вернёт управление, делая вид, что запись уже произведена, тогда как на самом деле запись на диск произойдёт гораздо позже. Такой режим вывода называется ***асинхронным***; наша программа о применении этого режима может не знать. Если теперь какие-то проблемы с реальной записью данных произойдут уже *после последнего вызова `write`*, то момент закрытия потока (вызов `close`) останется для операционной системы последней возможностью нам об этих проблемах сообщить.

Забегая вперёд, отметим ещё один нетривиальный момент. С одним и тем же потоком ввода-вывода (как объектом ядра) может быть связано больше одного номера дескриптора, причём как в одном процессе, так и в разных. Вызов `close` ликвидирует только номер дескриптора; объект в ядре при этом исчезает лишь в том случае, если в системе не осталось других дескрипторов, связанных с этим же потоком.

Раз уж речь зашла об асинхронном режиме вывода, отметим, что в современных условиях в большинстве случаев (хотя и не во всех) применяется именно он: система сразу же принимает у нас информацию, переданную через вызов `write`, копирует её в свою внутреннюю память, возвращает нашему процессу управление (в нашей программе это выглядит как успешно завершившийся вызов `write`), а физическую операцию записи производит позже — например, когда окажется свободен нужный контроллер и для него не будет более срочных дел. Асинхронный режим позволяет процессам продолжать работу, не дожидаясь, пока будет физически завершена операция вывода, и это в большинстве случаев полезно, поскольку от результата записи (в отличие от результатов чтения) дальнейшая работа программы обычно не зависит.

С другой стороны, у асинхронного режима вывода имеется очевидный недостаток: если работа компьютера будет неожиданно прервана (например, пропадёт электропитание), некоторые из операций записи, завершившиеся успешно с точки зрения процессов, физически так и не будут исполнены. Если речь идёт о данных, только что сгенерированных нашим процессом, то в принципе ничего страшного не произойдёт — можно представить, что работа компьютера прервалась чуть раньше, так что процесс просто не успел создать эти данные. Совершенно иначе выглядит ситуация с данными, перемещаемыми из одного места в другое — с диска на диск и т. п. Например, почтовый сервер при выполнении локальной доставки электронного письма читает его содержимое из почтовой очереди, которая хранится в виде файлов на диске, записывает письмо в почтовый ящик пользователя, который

также представляет собой файл, а затем, убедившись, что запись успешно прошла, удаляет письмо из очереди. Если компьютер выключить во время выполнения такой операции, может получиться так, что письма не окажется ни в очереди, ни в почтовом ящике, то есть оно потеряется.

Для таких случаев система предусматривает возможность в явном виде дождаться завершения операции записи; делается это системным вызовом `fsync`:

```
int fsync(int fd);
```

Параметром служит номер файлового дескриптора. Вызов сообщает операционной системе, что все данные, связанные с файлом, должны быть как можно скорее физически записаны на диск; точнее говоря, должна быть как можно скорее выполнена физическая операция записи, поскольку речь не всегда идёт именно о дисковых файлах. Управление из этого вызова возвращается не раньше, чем система реально выполнит все нужные операции записи. Вызов возвращает `-1` в случае ошибки, `0` в случае успеха.

Во многих случаях вместо вызова `fsync` имеет смысл применять вызов `fdatasync`, имеющий точно такой же профиль и работающий аналогично, но гарантирующий запись на диск только той информации, которая необходима для последующего корректного прочтения данных; например, информация о времени последней модификации при этом на диск принудительно не записывается, поскольку для последующего чтения она не важна.

Кроме того, система поддерживает вызов `sync`, не принимающий параметров и не возвращающий значений. Этот вызов блокирует вызвавший процесс (то есть не возвращает ему управление) до тех пор, пока вся информация, которая пока что содержится только в памяти системы, но предназначена для записи на диски, не окажется физически записана. Подчеркнём, что речь идёт о *всей* информации для *всех* дисковых файлов, находящихся в работе, для *всех пользователей системы*. Вызов `sync` всегда заканчивается успешно, то есть система ни при каких условиях никому не отказывает в запросе на общую синхронизацию. Впрочем, не следует думать, что, получив этот вызов, ядро бросит все свои дела и побежит писать данные на диск; запрос будет принят к сведению, соответствующие дисковые операции запланированы, дальше ядро их *когда-то* выполнит в своём обычном режиме. Если из соображений эффективности ядро сочтёт, что какие-то другие действия следует выполнить раньше, то оно выполнит их раньше. Иначе говоря, речь тут идёт не о том, что ядро будет сломя голову исполнять вашу волю, а о том, что ваш процесс будет ждать столько, сколько потребуется ядру на физическую запись всей информации, которая ещё не была записана в момент вызова `sync`. Это объясняет, почему вызов `sync` доступен всем, а не только администратору.

Ядро операционной системы поддерживает гибкую настройку поведения потоков, связанных с дескрипторами; доступ практически ко всем возможностям такой настройки предоставляет вызов `fcntl`:

```
int fcntl(int fd, int cmd, ...);
```

Первым параметром вызов получает номер дескриптора, вторым — некий «код команды», задаваемый одной из предопределённых целочисленных констант, предназначенных специально для этого. Многоточие означает, что потенциально вызов может получить ещё сколько угодно параметров любых типов; конкретные параметры и их типы зависят от значения второго параметра (команды).

Для описания большинства команд `fcntl` у нас пока недостаточно знаний, но две мы можем рассмотреть прямо сейчас. При описании вызова `open` (см. стр. 103) мы перечислили несколько флагов, передаваемых в `open` вторым параметром. Эти флаги можно разделить на *флаги режима работы* (`O_RDONLY`, `O_WRONLY`, `O_RDWR`), *флаги создания файла* (`O_CREAT`, `O_EXCL`, `O_TRUNC` и пока не рассматривавшийся нами `O_NOCTTY`), а также *флаги статуса* — все остальные, из которых мы пока обсуждали только `O_APPEND` и `O_NONBLOCK`. Флаги статуса отличаются от всех остальных тем, что их возможно и даже иногда осмысленно менять уже после начала работы с файлом. Текущее значение флагов статуса для открытого потока можно узнать, выполнив вызов `fcntl` с командой `F_GETFL`. Дополнительные параметры в этом случае не предусмотрены, вызов выполняется от двух параметров:

```
flags = fcntl(fd, F_GETFL);
```

Установлен ли или нет конкретный флаг, можно понять с помощью операции побитового «и»:

```
if(flags & O_APPEND) { /* флаг O_APPEND установлен */
    /* ... */
```

Изменить текущий набор флагов статуса позволяет команда `F_SETFL`; в этом случае вызов `fcntl` получает третий параметр, имеющий тип `long` и представляющий собой побитовую дизъюнкцию нужных флагов. Например, следующая последовательность действий позволяет установить для файла неблокирующий режим, сохранив остальные флаги неизменными:

```
flags = fcntl(fd, F_GETFL);
fcntl(fd, F_SETFL, flags | O_NONBLOCK);
```

Впрочем, можно просто выполнить `fcntl(fd, F_SETFL, O_NONBLOCK)`, если вы уверены, что никакие другие флаги статуса на вашем потоке ввода-вывода не используются. Позже, по мере появления нужных знаний, мы рассмотрим другие команды вызова `fcntl`.

5.2.4. Управление объектами файловой системы

Кроме чтения информации из файлов и записи в файлы, с объектами файловой системы приходится производить целый ряд других действий. Рассмотрим некоторые из предназначенных для этого системных вызовов. Сразу же отметим, что все эти системные вызовы возвращают значение `-1` в случае ошибки, а в случае успеха возвращают какое-то другое значение (чаще всего `0`), но традиционно результат проверяют именно на равенство `-1`.

Жёсткие ссылки (см. стр. 101) создаются с помощью системного вызова `link`:

```
int link(const char *oldpath, const char *newpath);
```

Здесь `oldpath` — существующее имя файла, `newpath` — новое имя. Как обычно, имя считается абсолютным, если оно начинается с символа «`/`», в противном случае оно отсчитывается от текущей директории; в дальнейшем мы не будем возвращаться к этой оговорке, поскольку она без изменений применяется ко всем системным вызовам и библиотечным функциям, в которых упоминаются имена файлов, за одним исключением, которое мы, когда придёт время, оговорим явно. Отметим, что оба имени должны быть именами файлов, а не чем-то другим. Читатель может заметить, что команда командной строки `ln`, позволяющая создавать жёсткие ссылки, принимает в качестве второго аргумента как имя файла, так и имя директории, при этом ссылка создаётся в этой (новой) директории под тем же именем, что и в старой; но этот вариант команда `ln` реализует своими средствами, а системный вызов `link` такого не умеет.

Системный вызов, предназначенный для удаления файла или, точнее, для удаления *ссылки* на файл из каталога, называется `unlink`:

```
int unlink(const char *name);
```

Напомним, что файл реально удаляется лишь тогда, когда удаляется последняя жёсткая ссылка на него, и то только в том случае, если этот файл никем не открыт, а иначе файл продолжает существовать в качестве дискового объекта, не имея при этом ни одного имени; счётчик ссылок в его индексном дескрипторе при этом равен нулю. Такой файл удаляется с диска после того, как его закроет последний из работавших с ним процессов. Может получиться так, что система окажется некорректно остановлена (например, при аварии электропитания) и файл с нулевым счётчиком ссылок так и останется в файловой системе; в этом случае он, как правило, удаляется при следующей загрузке системы. Отметим, что вызов `unlink` не позволяет удалять директории.

Переименование файла производится системным вызовом `rename`:

```
int rename(const char *oldpath, const char *newpath);
```

Параметры задают старое имя, то есть то имя, под которым файл уже существует в системе, и новое имя, которое этому файлу следует дать вместо старого. Имена могут относиться как к одной директории, так и к разным, то есть файл можно с помощью `rename` переместить из одной директории в другую, но только если эти директории находятся на одном диске; перемещать файл между разными файловыми системами этот вызов не умеет.

У вызова `rename` есть одно важное свойство, которое часто используется на практике: если файл с именем, заданным параметром `newpath`, существует, то вызов *заменяет* его файлом, заданным параметром `oldpath`, причём это происходит за одно атомарное (неделимое) действие. На директории это свойство не распространяется: если `oldpath` задаёт директорию, то `newpath` должен либо не существовать вовсе, либо быть пустой директорией (в этом случае замена всё же произойдёт). Кроме того, атомарность переименования не всегда может быть обеспечена для сетевых дисков (физически находящихся на других компьютерах).

Как мы знаем, для создания ***директории*** в командной строке применяется команда `mkdir`, а для её удаления — команда `rmdir`; точно так же называются и системные вызовы, с помощью которых можно попросить операционную систему создать или удалить директорию:

```
int mkdir(const char *name, int perms);
int rmdir(const char *name);
```

Параметр `name` для обоих вызовов задаёт имя директории (создаваемой или удаляемой); что касается параметра `perms`, то этот параметр задаёт ***права доступа*** к создаваемой директории, подобно одноимённому параметру для вызова `open`, рассмотренному в предыдущем параграфе. Как и при создании файлов, из значения `perms` удаляются биты, которые установлены в `umask`, что позволяет в большинстве случаев не беспокоиться о защите создаваемой директории: пользователь может позаботиться об этом самостоятельно, установив соответствующее значение параметра `umask`. С другой стороны, в данном случае мы создаём директорию, а не файл, а для директории очень важен бит `x` (права на исполнение), без которого использовать содержимое директории невозможно; поэтому в большинстве случаев мы указываем параметром `perms` число `0777`, а в достаточно редких случаях, когда создаваемая директория будет заведомо критична по безопасности — число `0700`, чтобы никто, кроме владельца, не имел к ней доступа вне зависимости от значения `umask`.

Символические ссылки создаются вызовом `symlink`:

```
int symlink(const char *oldpath, const char *newpath);
```

Профиль этого вызова напоминает профиль вызова `link`, который мы рассмотрели раньше, но работает он совершенно иначе. Система требует, чтобы строка `oldpath` не была пустой, но больше никаких проверок этого параметра не делает. Символическая ссылка (то есть файл специального типа) создаётся под именем `newpath` и ссылается на `oldpath` вне зависимости от того, существует ли в действительности файл с таким именем; если его не существует, созданная ссылка оказывается «висячей», но в этом, как уже отмечалось, нет ничего страшного.

Удаление символической ссылки, как и обычного файла, производится с помощью вызова `unlink`. Стоит обратить внимание, что этот вызов, в отличие от большинства других, будучи применённым к символической ссылке, работает с ней самой, а не с файлом, на который она ссылается.

Для изменения прав доступа к файлам используется системный вызов, называющийся так же, как и соответствующая команда командной строки, `chmod`:

```
int chmod(const char *path, int perms);
```

Первым параметром вызова указывается имя файла, вторым — новые права доступа к нему в виде целого числа, биты которого соответствуют битам прав доступа. Чаще всего число задаётся явным образом в виде восьмеричной константы. Например, `chmod("file.txt", 0600);` установит для файла `file.txt` в текущей директории права на запись и чтение для его владельца, а все остальные права сбросит. Обычный пользователь может изменить права доступа только для файлов, принадлежащих ему, но на системного администратора это ограничение, как обычно, не распространяется. Вызов `chmod` возвращает `-1` в случае ошибки, `0` в случае успеха. Отметим, что **параметр `umask` на работу вызова `chmod` никакого влияния не оказывает**.

Для изменения владельца файла служат вызовы `chown` и `lchown`:

```
int chown(const char *path, int owner, int group);
int lchown(const char *path, int owner, int group);
```

Различие между ними в том, что если их применить к символической ссылке, то `chown` будет работать с файлом, на который эта ссылка ссылается, а `lchown` — с самой ссылкой, хотя в большинстве случаев это бессмысленно; единственный хорошо известный случай, когда для символической ссылки важен её владелец — при нахождении ссылки в директории, имеющей установленный Sticky bit, поскольку в такой директории пользователь может удалять только свои файлы. Параметры `owner` и `group` задают числовые идентификаторы пользователя и группы пользователей. Если в качестве любого из этих параметров указать значение `-1`, соответствующий идентификатор для файла сохраняется без изменений. Отметим, что изменить владельца для любого файла может только суперпользователь (т. е. для этого процесс, выполняющий вызов, должен иметь `euid`, равный `0`), а изменить группу может, кроме суперпользователя, также владелец файла, но только на такую группу, членом которой он является (то есть указанный третьим параметром `gid` должен либо быть равен `egid`'у вызывающего процесса, либо находиться среди его «дополнительных групп»). Попытка выполнить неразрешённые изменения приведёт к тому, что вызов завершится ошибкой. Как обычно, в случае ошибки вызов возвращает значение `-1`; в случае успеха возвращается ноль.

Вызовы `stat`, `lstat` и `fstat` позволяют узнать подробную информацию о файле. Профили этих вызовов выглядят так:

```
int stat(const char *path, struct stat *buf);
int lstat(const char *path, struct stat *buf);
int fstat(int fd, struct stat *buf);
```

Различие между `stat` и `lstat` такое же, как между `chown` и `lchown`: если первым аргументом будет имя символической ссылки, то `stat` пройдёт по ссылке и выдаст информацию по файлу, а `lstat` выдаст информацию по самой ссылке. Вызов `fstat` позволяет узнать информацию об открытом файле, не зная его имя: первым параметром он получает открытый файловый дескриптор. Извлечённую информацию вызовы записывают в структуру, адрес которой передан параметром `buf`. Используемая для этого структура `struct stat`, описанная в системных заголовочных файлах, содержит следующие поля:

- `st_dev`: идентификатор устройства, на котором находится файл;
- `st_ino`: номер индексного дескриптора (*i-node*);
- `st_mode`: тип файла и права доступа к файлу;
- `st_nlink`: количество ссылок на файл;
- `st_uid`: идентификатор пользователя-владельца файла;
- `st_gid`: идентификатор группы пользователей;
- `st_rdev`: идентификатор устройства (для специальных файлов, соответствующих устройствам в системе);
- `st_size`: размер файла в байтах;
- `st_blksize`: размер блока ввода-вывода на данной файловой системе;
- `st_blocks`: количество блоков по 512 байт (*sic!*), занятых данным файлом;
- `st_atime`: дата и время последнего доступа к файлу;
- `st_mtime`: дата и время последней модификации содержимого файла;
- `st_ctime`: дата и время последнего изменения свойств файла (например, владельца, группы, прав доступа; запись информации в файл также изменяет этот параметр).

Отметим, что битовая строка, записываемая в поле `st_mode`, содержит как права доступа к файлу, так и его тип; для извлечения прав доступа как таковых следует произвести побитовую конъюнкцию значения из этого поля с маской `07777` (или, что то же самое, `0x0fff`), выделив из него 12 младших бит.

Существующий *обыкновенный* файл можно укоротить, уничтожив при этом всё его содержимое, начиная с определённой позиции. Это можно сделать с помощью системного вызова `truncate`, не открывая файл; если же файл уже открыт, можно воспользоваться вызовом `ftruncate`:

```
int truncate(const char *path, int length);
int ftruncate(int fd, int length);
```

Первым параметром для `truncate` служит имя файла, для `ftruncate` — открытый файловый дескриптор; вторым параметром обоих вызовов задаётся новая длина файла. Если файл до этого имел бóльшую длину, он укорачивается, а его содержимое, начиная с позиции `length`, теряется; если же файл был короче, чем указано в параметре, то он становится длиннее, а при чтении байтов, добавленных таким способом, мы получаем нули. Система, насколько возможно, старается в таких случаях использовать уже знакомые нам по вызову `lseek` «дырки». Операция изменения длины считается операцией записи в файл, так что при использовании `truncate` у нас должны быть права на запись в данный файл, а при использовании `ftruncate` файл должен быть открыт в режиме записи (`O_WRONLY` или `O_RDWR`).

5.2.5. Файлы устройств и классификация устройств

Одно из замечательных свойств ОС Unix — обобщённая концепция ***файла*** как универсальной абстракции. Большинство внешних устройств представлено на пользовательском уровне как файлы специального ***типа***, которые так и называются ***файлами устройств***. Это касается и жёстких дисков, и всевозможных последовательных и параллельных портов, и виртуальных терминалов, и т. п.

Так, часто требуется записать в файл образ компакт-диска (CD), например, для последующего создания его копии. В некоторых других операционных системах для проведения такой операции потребуется специальное программное обеспечение. В системах семейства Unix достаточно вставить диск в привод и дать команду

```
cat /dev/cdrom > image.iso
```

Точно так же, чтобы отправить файл на печать, достаточно команды

```
cat myfile.ps > /dev/lp0
```

Конечно, обычно так не делают, полагаясь на подсистему печати, однако нам в данном случае важнее сам факт такой возможности.

Такой подход позволяет работать с устройствами в основном с помощью тех же системных вызовов, что и для обычных файлов. Так, чтобы записать информацию в определённый сектор жёсткого диска, в других операционных системах требуется обратиться к системному вызову, специально предназначенному для записи секторов физического диска. В ОС Unix достаточно открыть на чтение специальный файл, соответствующий нужному диску, с помощью вызова `lseek` позиционироваться на нужный сектор и выдать обычный `write`. Именно так происходит, например, форматирование диска, т. е. создание файловой системы.

В старых версиях Linux при вводе звукового сигнала с микрофона можно было открыть на чтение файл, соответствующий звуковому устройству, и произвести чтение. Прочитанную таким образом информацию можно было снова записать в звуковое устройство, в результате чего звук воспроизводился. К сожалению, нынешняя звуковая подсистема стала слишком навороченной для таких наглядных действий.

Надо отметить, что некоторые периферийные устройства могут не иметь файлового представления. Например, не во всех ОС семейства Unix существуют файлы, связанные с сетевыми интерфейсами; Linux таких файлов не поддерживает.

Устройства, имеющие представление в виде файла, делятся на два типа: ***символьные*** (или ***потоковые***, они же ***байт-ориентированные***) и ***блочные*** (***блок-ориентированные***). Основные операции над байт-ориентированными устройствами — запись и чтение одного символа (байта) или их последовательности. В противоположность этому блок-ориентированные устройства воспринимаются как хранилище данных, разделённых на блоки фиксированного размера; основными операциями, соответственно, становятся чтение и запись заданного блока. В качестве примеров байт-ориентированных устройств можно назвать терминал (клавиатура и устройство отображения), принтер (точнее говоря, *принтерный порт*), манипулятор «мышь», звуковую карту.

Существует целый ряд символьных ***псевдоустройств***, не имеющих физического воплощения. Так, в устройство `/dev/null` можно записать любую информацию, которая попросту игнорируется; попытка читать из этого устройства вызывает ситуацию «конец файла». Устройство `/dev/zero` позволяет прочитать любой объём данных, причём все прочитанные байты будут равны нулю; записывать в него, как и в `/dev/null`, можно что угодно, вся эта информация игнорируется. Устройство `/dev/full` похоже на `/dev/zero` тем, что при чтении выдаёт бесконечный поток нулевых байтов, а вот запись вызывает такую же ошибку, как при отсутствии места на диске: вызов `write` возвращает `-1`, переменная `errno` принимает значение `ENOSPC`. Устройство `/dev/random` выдаёт читающему процессу последовательность случайных чисел, и т. п.

В виде блок-ориентированных устройств обычно представляются ***диски*** и другие подобные устройства. Надо сказать, что дисковые устройства обычно позволяют чтение или запись только целыми секторами, что обусловлено особенностями их физической реализации, но «блок-ориентированность» файлов устройств к этому прямого отношения не имеет: при работе с таким устройством пользовательская задача не обязана соблюдать границы секторов и может, к примеру, запросить чтение 1000 байт, первые 100 из которых лежат в одном секторе, следующие 512 — в следующем, а оставшиеся 388 — в третьем. О том, чтобы прочитать все три сектора и поместить в память задачи только нужные байты, позаботится драйвер диска. Можно встретить утверждение, что основная особенность блок-ориентированных устройств состоит в наличии кеша в ядре, то есть содержимое некоторых блоков ядро может держать в памяти, что позволяет экономить на операциях чтения, а операции записи откладывать до момента, когда для контроллера диска не будет более важных дел.

Блок-ориентированные и байт-ориентированные устройства представляют собой ещё два специальных типа файлов наряду с директориями и символическими ссылками. Файлы устройств можно открывать на чтение и запись, а к полученным дескрипторам применять вызовы `read` и `write`. Кроме того, блок-ориентированные устройства поддерживают позиционирование с помощью вызова `lseek`. **Байт-ориентированные устройства операцию позиционирования не поддерживают.** Это, пожалуй, наиболее заметное отличие байт-ориентированных устройств от блок-ориентированных с точки зрения прикладного программиста. Заметим кстати, что вместе с обычными дисковыми файлами блок-ориентированные устройства исчерпывают множество ***позиционируемых*** потоков ввода-вывода; больше таких нет, специальные файлы всех остальных типов либо вообще не могут быть открыты на чтение и запись, либо не поддерживают `lseek`.

В системе FreeBSD в какой-то момент отказались от блок-ориентированных устройств, мотивируя это неуправляемостью кеша; теперь там все файлы устройств имеют тип байт-ориентированных, но некоторые из них — те, что связаны с дисками — поддерживают вызов `lseek`. В ОС Linux, а равно и в других системах семейства BSD блочные устройства поддерживаются по-прежнему.

Ясно, что управление устройствами не может ограничиваться только операциями чтения и записи. Для многих устройств определены также дополнительные операции, специфические для данного устройства, такие как открытие и закрытие лотка привода CD-ROM, установка скорости обмена последовательного порта, низкоуровневая разметка гибких дисков, управление громкостью воспроизведения звука звуковой картой и т. п. Все операции такого рода выполняются с помощью системного вызова `ioctl`:

```
int ioctl(int fd, int request, ...);
```

Параметр `fd` задаёт дескриптор открытого файла устройства, параметр `request` — код нужной операции. Вызов может получать дополнительные параметры, требуемые для выполнения данной операции. Например, следующий код

```
int fd = open("/dev/cdrom", O_RDONLY|O_NONBLOCK);
ioctl(fd, CDROMEJECT);
ioctl(fd, CDROMCLOSETRAY);
```

откроет, а затем закроет лоток привода CD-ROM. Параметр `O_NONBLOCK` задаётся, чтобы избежать поиска диска в устройстве и ошибки в случае, если диск в устройство не вставлен. Если же вставить в то же устройство музыкальный диск (audio CD), код

```
struct cdrom_ti cti;
cti.cdti_trk0 = 2;
cti.cdti_ind0 = 0;
cti.cdti_trk1 = 2;
cti.cdti_ind1 = 0;
ioctl(fd, CDROMPLAYTRAKIND, &cti);
```

заставит ваше устройство воспроизвести вторую дорожку диска.

5.2.6. Работа с содержимым каталогов

Как мы уже отмечали, ***каталог*** представляет собой дисковый файл специального типа, в котором содержится информация об именах файлов и соответствующих номерах индексных дескрипторов.

Любые изменения в содержимом каталогов система производит только сама — при создании, удалении и переименовании файлов (точнее, жёстких ссылок на файлы). Любые попытки что-то записать в каталог в обход этих операций, очевидно, привели бы к нарушению целостности файловой системы, так что ядро их не допускает. С другой стороны, довольно часто требуется узнать, что содержится в каталоге — например, это нужно команде `ls`, но, конечно, не только ей. Поскольку чтение информации не может само по себе нарушить её целостность, а всевозможные проверки прав доступа и прочие действия при извлечении информации из каталога во многом похожи на аналогичные действия при чтении обычного файла, **система позволяет открывать каталоги на чтение с помощью вызова** `open` (но только на чтение). Впрочем, вызов `read` на полученном дескрипторе всё равно работать не будет; название системного вызова, специально предназначенного для чтения из директорий, зависит от конкретной системы; в ОС Linux и FreeBSD этот вызов называется `getdents`, причём работает он в обеих системах одинаково, но формальные профили — типы принимаемых параметров — различаются. Использовать его напрямую в любом случае не следует, поскольку это требует знания внутренней структуры каталогов и к тому же существенно ограничивает переносимость программы на другие системы.

Библиотека языка Си предусматривает для анализа содержимого каталогов (директорий) функции `opendir`, `readdir` и `closedir`:

```
DIR *opendir(const char *name);
struct dirent *readdir(DIR *dirp);
int closedir(DIR *dirp);
```

Загадочное значение типа `DIR*` играет здесь приблизительно ту же роль, что и `FILE*` для высокоуровневой работы с файлами: это указатель на некую структуру, создаваемую библиотекой в своей памяти, причём содержимое этой структуры нас волновать не должно, нам она интересна лишь как «нечто, обеспечивающее работу с директорией».

Функция `readdir` читает очередную запись из открытого с помощью `opendir` каталога и возвращает указатель на некую структуру, имеющую тип `struct dirent`. Сама эта структура находится там же, где и созданное библиотекой *нечто*, именуемое `DIR`, и существует в одном экземпляре для каждого открытого на чтение каталога, то есть, скорее всего, читая из одного и того же потока типа `DIR*`, вы каждый раз будете получать один и тот же адрес; это нужно учитывать, ведь содержимое структуры `dirent`, возвращённое при предыдущем обращении к `readdir`, окажется перезаписано следующим обращением.

Когда записи в каталоге заканчиваются, `readdir` возвращает `NULL`, после чего поток следует закрыть с помощью `closedir`, чтобы высвободить ненужный теперь файловый дескриптор.

Довольно своеобразно обстоят дела с полями структуры `dirent`. Можно совершенно точно сказать только одно: в структуре присутствует поле `d_name`, представляющее собой массив элементов типа `char`, но каков размер этого массива — в общем случае неизвестно. Библиотека поддерживает константу `NAME_MAX`, равную максимальной длине имени файла, хранимого в директории; в большинстве случаев это число 255, но никакой гарантии этого не даётся. Кроме того, обычно структура `dirent` содержит поле `d_ino`, имеющее *какой-то* целочисленный тип (например, в системе автора книги это оказался `unsigned long`; системные заголовочные файлы обозначают этот тип именем `ino_t`, вводимым с помощью `typedef`). В поле `d_ino` записывается номер индексного дескриптора (*i-node*). Как правило, структура `dirent` имеет и другие поля, но они специфичны для конкретной системы.

Например, следующая программа принимает в качестве параметра командной строки имя каталога (если параметр не задан, используется текущий каталог) и выдаёт список имён файлов в этом каталоге, по одному имени в строке:

```
#include <stdio.h>                              /* poor_ls.c */
#include <sys/types.h>
#include <dirent.h>
int main(int argc, char **argv)
{
    DIR *dir;
    struct dirent *dent;
    const char *name = ".";
    if(argc > 1)
        name = argv[1];
    dir = opendir(name);
    if(!dir) {
        perror(name);
        return 1;
    }
    while((dent = readdir(dir)) != NULL) {
        printf("%s\n", dent->d_name);
    }
    closedir(dir);
    return 0;
}
```

5.2.7. Отображение файлов в память

В ОС Unix предусмотрена возможность отображения содержимого некоторого файла в виртуальное адресное пространство процесса. В результате такого отображения появляется возможность работы с данными в файле как с обычными переменными в оперативной памяти, то есть, например, с помощью присваиваний. Отображение осуществляется системным вызовом `mmap`:

```
void *mmap(void *start, int length, int protection,
           int flags, int fd, int offset);
```

Перед обращением к `mmap` файл должен быть открыт с помощью `open`; вызов принимает дескриптор файла, подлежащего отображению, через параметр `fd`. Параметры `offset` и `length` задают соответственно позицию начала отображаемого участка в файле и его длину. Здесь нужно заметить, что отображения на самом деле всегда делаются *целыми страницами*, так что и длина отображения, и позиция его начала относительно начала файла, и численное значение его адреса всегда кратны некоторому предопределённому числу, называемому *размером страницы*. Его можно узнать с помощью функции

```
int getpagesize();
```

Несмотря на это, параметр `length` мы можем задать равным любому строго положительному числу; до целой страницы система округлит этот параметр самостоятельно. А вот если некратным размеру страницы окажется параметр `offset`, вызов завершится с ошибкой.

Размер страницы для `mmap` в общем случае может не совпадать с размером страницы виртуальной памяти; так, на некоторых старых версиях Linux вызов `getpagesize` возвращал число 16, хотя таких виртуальных страниц нет ни на одной архитектуре.

Параметр `protection` вызова `mmap` задаёт режим доступа к получаемому участку виртуальной памяти. Для этого служат константы `PROT_READ`, `PROT_WRITE` и `PROT_EXEC`, которые можно объединять операцией побитового «или». Как ясно из названия, первые две константы соответствуют доступу на запись и чтение. Третья позволяет передавать управление в область отображения, то есть исполнять там код; это используется, например, при подгрузке динамических библиотек. Существует также константа `PROT_NONE`, соответствующая запрету доступа любого вида. Задаваемый параметром `protection` доступ должен быть совместим с режимом, в котором был открыт файл: так, если файл открыт в режиме «только чтение», то есть в вызове `open` был использован флажок `O_RDONLY`, то попытка отобразить файл в память с режимом, допускающим запись, вызовет ошибку.

В качестве параметра `flags` указывают либо `MAP_SHARED`, либо `MAP_PRIVATE` (в этом случае изменения, производимые в виртуальном адресном пространстве, никак на файле не отразятся). Кроме того, к любому из этих двух флагов можно добавить через операцию побитового «или» флажки дополнительных опций. Среди этих опций есть `MAP_ANONYMOUS`, позволяющая обойтись без файла, то есть создать просто область виртуальной памяти; в этом случае параметры `fd` и `offset` игнорируются.

Память, выделенная с помощью `mmap` с одновременным указанием `MAP_ANONYMOUS` и `MAP_SHARED`, отличается от обычной памяти тем, что при создании нового процесса она становится доступна из обоих процессов, то есть изменения, сделанные в такой памяти порождённым процессом, будут

доступны родительскому и наоборот; это называется ***разделяемой памятью*** (англ. *shared memory*). Разделяемая память иногда используется для организации взаимодействия между процессами; надо сказать, что работа с разделяемой памятью порождает целый ряд неочевидных проблем, которым будет посвящена отдельная (седьмая) часть нашей книги.

Параметр `start` позволяет указать системе, в каком месте нашего адресного пространства нам хотелось бы видеть новую область памяти. Обычно пользовательские программы не используют эту возможность; в качестве параметра `start` можно передать `NULL`, тогда система сама выберет свободную область виртуального адресного пространства. Случаи, когда параметр `start` задаётся ненулевым, мы рассматривать не будем; при желании читатель легко найдёт соответствующую информацию в других источниках, в том числе в технической документации и на разнообразных сайтах в Интернете.

Вызов `mmap` возвращает адрес созданной области виртуальной памяти. В случае ошибки `mmap` возвращает значение `MAP_FAILED`, равное `-1`, преобразованной к типу `void*`.

Приведём пример:

```
int fd, size;
char *p;
fd = open("file.dat", O_RDWR);
if(fd == -1) {
    /* ... обработка ошибки ... */
}
size = lseek(fd, 0, SEEK_END);
p = mmap(NULL, size, PROT_READ|PROT_WRITE, MAP_SHARED, fd, 0);
if(p == MAP_FAILED) {
    /* ... обработка ошибки ... */
}
```

После выполнения этих действий файл `"file.dat"` будет отображён в память целиком, так что, например, выражение `p[25]` будет равно значению 26-го байта в файле, причём операция присваивания `p[25] = 'a'` занесёт в этот байт символ `'a'`.

Размер файла, как мы понимаем, может быть совершенно произвольным и уж точно не обязан быть кратным каким-то там страницам, так что в конце последней страницы отображения может возникнуть область памяти, не соответствующая файлу. Система при создании отображения заполняет такую область нулями; запись в эту область памяти не приводит ни к каким изменениям содержания файла. Но вот если длина отображения превышает длину файла более чем на страницу, возможны весьма разнообразные эффекты. Во-первых, сам `mmap` может (хотя и не обязан) отказаться создавать отображение, в этом случае он вернёт `MAP_FAILED`. Во-вторых, если ваша программа всё-таки обратится к такой странице отображения, которая находится полностью за пределами отображаемого файла, ядро может прибить её как аварийную (сигналом `SIGBUS`[15]), причём в документации подчёркивается, что это может произойти уже после успешного обращения к `mmap`, например, если другой процесс уменьшит размер файла.

Отменить отображение, созданное вызовом `mmap`, можно с помощью вызова `munmap`:

```
int munmap(void *start, int length);
```

Интересно, что аргументы `munmap` не обязаны совпадать с началом и длиной какого-то из отображений, ранее созданного с помощью `mmap`. Повинуясь вызову `munmap`, система прекратит отображение всех страниц, полностью или частично пересекающихся с заданной областью адресного пространства, так что при желании мы можем отменить как часть ранее созданного отображения (естественно, состоящую из целых страниц), так и сразу несколько отображений, стоящих в нашем адресном пространстве подряд. Численное значение адреса, передаваемого параметром `start`, должно быть кратно размеру страницы, тогда как `length` может быть произвольным положительным числом.

Физическую запись в файл изменений, сделанных в области отображения, система может произвести не сразу. С помощью вызова `msync` можно заставить систему произвести запись немедленно; изучение этого вызова оставляем читателю для самостоятельной работы.

Отметим, что вызов `mmap`, изначально предназначавшийся для работы с содержимым дисковых файлов, в наше время воспринимается скорее как универсальный интерфейс для создания областей виртуальной памяти. Так, в системах семейства BSD именно этот вызов использует функция `malloc`, чтобы затребовать у системы очередную порцию памяти. Например, после выполнения

```
p = mmap(NULL, size, PROT_READ|PROT_WRITE,
         MAP_PRIVATE|MAP_ANONYMOUS, -1, 0);
```

указатель `p` будет содержать адрес новой области виртуальной памяти, полученной непосредственно от операционной системы.

5.3. Процессы

5.3.1. Процесс: что это такое

Как уже говорилось, пользовательские программы выполняются под управлением операционной системы в виде ***процессов***; иначе говоря, ***процесс*** — это некая сущность, создаваемая ядром операционной системы, чтобы выполнять пользовательскую программу.

Как это часто бывает в технических дисциплинах, то или иное понятие спонтанно формируется в среде специалистов, обретает устойчивые очертания, а когда кто-то спохватывается, что новый термин не имеет строгого определения, оказывается уже поздно: термин живёт своей жизнью и совершенно не желает укладываться в рамки любых определений. Именно так произошло и с понятием процесса: каждый программист понимает, что это такое, но никто не может дать такое определение, для которого нельзя было бы с ходу предъявить контрпример. Какое бы определение процесса ни давалось, всегда можно придумать что-то, что процессом не является, но под определение подходит, или что-то, что не подходит, но при этом является процессом.

В самом первом приближении можно считать, что процесс — это *программа, которую запустили на выполнение*; выполняемая прямо сейчас программа, кроме собственно машинного кода, её составляющего, обладает также текущими значениями переменных, текущей позицией исполнения и прочими особенностями, образующими её ***состояние***. Кроме того, запущенная программа может владеть какими-то объектами, находящимися вне её — например, открытыми файлами; как уже говорилось, с процессом связаны определённые *полномочия*. Дойдя до этого момента, мы можем заметить, что наше «определение в первом приближении» уже опровергнуто: ядро операционной системы, вне всякого сомнения, представляет собой программу и даже выполняется, но *процессом не является*, с ядром не связаны ни полномочия (они у него безграничны, к тому же если спросить, кто следит за полномочиями процесса, правильным ответом будет как раз «ядро операционной системы», а за полномочиями самого ядра следить некому), ни состояние выполнения (в зависимости от подхода к реализации ядра в нём может существовать несколько контекстов выполнения одновременно). Впрочем, это даже не столь важно: просто любой программист, имеющий дело с процессами, скажет, что исполняющееся на машине ядро операционной системы — это не процесс.

Между прочим, то же самое можно сказать про специфические ситуации, когда операционной системы вообще нет; именно так, на «голом железе» (соответствующий английский термин — *on bare metal* или *on bare bones*), выполняются прошивки микроконтроллеров. В такой вычислительной системе программа вообще одна, ей не к кому обращаться, её некому контролировать. Несмотря на то, что программа, очевидно, запущена и выполняется, то есть обладает состоянием, называть это явление процессом никому в голову не приходит.

Исправить ситуацию позволяет небольшое уточнение: *процесс — это программа, которую запустили на выполнение под управлением операционной системы*. В таком виде определение можно принять в качестве рабочего, но и оно окажется опровергнуто, когда мы обнаружим, что в системах семейства Unix *можно в рамках одного процесса заменить одну выполняющуюся программу на другую*. Впрочем, это не страшно; следует, по-видимому, помнить, что строгого определения процесса нет (и не надо), но при этом рассмотрение процесса как «программы, которая выполняется под управлением ОС» позволяет нам приблизиться к пониманию действительной сущности процесса.

К сущности процессов можно подойти с другой стороны. Для того, чтобы запустить в системе на выполнение некую программу, ядро должно, очевидно, создать в своей памяти определённую структуру данных, или, говоря

[15] Подробный разговор о сигналах у нас пока впереди, но сами сигналы как явление нам уже встречались в первом томе, см. §1.2.9 и §3.6.7.

более «научно», некий объект. В этом объекте (структуре данных) будут содержаться, например, сведения о памяти, которая выделена запущенной программе, о её полномочиях, об открытых ею файлах (потоках ввода-вывода) и многое другое. В этой же структуре данных в те периоды, когда программа временно снята с выполнения, будет храниться содержимое регистров центрального процессора: при снятии программы с выполнения ядро скопирует содержимое регистров в наш объект, а при возобновлении выполнения — заполнит регистры информацией, ранее сохранённой в объекте. Программисты, пишущие части ядра операционной системы, часто говорят, что этот объект ядра и есть процесс; иначе говоря, процесс — это структура данных, создаваемая в ядре операционной системы для поддержки выполнения пользовательской программы. Конечно, это тоже никоим образом не определение, это скорее отражение определённой точки зрения.

5.3.2. Свойства процесса

Не имея возможности дать определение процесса, мы восполним этот пробел, обсудив, какими процесс обладает *свойствами*. Прежде всего отметим, что каждый процесс в системе имеет уникальный идентификатор — целое число, называемое «`pid`» от слов *process identifier*. В ОС Unix с процессом связан также параметр `ppid`, равный идентификатору родительского процесса, т. е. процесса, породившего данный, если этот процесс ещё существует; если родительский процесс завершается раньше порождённого, `ppid` потомка становится равен `1`. Идентификация процесса не ограничивается его `pid`'ом и `ppid`'ом: каждый процесс в ОС Unix относится к некой *группе* в рамках *сеанса*, которые обозначаются *идентификатором группы процессов* (`pgid`, *process group identifier*) и *идентификатором сеанса* (`sid`, *session identifier*). Сеансы и группы процессов мы подробно рассмотрим в §5.4.2.

При обсуждении задач, решаемых операционными системами, мы в §5.1.2 специально подчеркнули важность *разграничения полномочий*, упомянув при этом, что полномочиями в системе на самом деле обладает не пользователь, а процесс, запущенный от его имени. В системах семейства Unix *информация, определяющая полномочия процесса*, включает *идентификатор пользователя* (`uid`, *user identifier*), *идентификатор группы пользователей* (`gid`, *group identifier*), а также *эффективные* идентификаторы пользователя и группы (`euid` и `egid`, от слова *effective*). В большинстве случаев эффективные идентификаторы совпадают с обычными; примером случая, когда это не так, служат так называемые suid-программы, то есть программы, выполняемые с правами пользователя, владеющего исполняемым файлом данной программы, а не того пользователя, который программу запустил (см. §5.2.2).

Узнать свои значения `uid`, `gid`, `euid`, `egid`, `pid`, `ppid`, `sid` и `pgid` процесс может с помощью системных вызовов, которые так и называются `getpid`, `getuid`, `getgid` и т. д. Все эти функции вызываются без параметров и возвращают значение соответствующего свойства процесса; ошибочные ситуации для них не предусмотрены. Параметры `pid` и `ppid` (идентификатор процесса и его предка) изменить нельзя. Манипуляция параметрами `sid` и `pgid` будет рассмотрена при обсуждении сеансов и групп процессов (§5.4.2). Параметры `uid`, `gid`, `euid` и `egid` в некоторых случаях могут меняться; об этом речь пойдёт при рассмотрении полномочий процессов (§5.3.11).

Продолжая рассмотрение свойств процессов, мы можем догадаться, что у каждого процесса имеется выделенная ему память, ведь иначе он не мог бы выполняться; как мы уже знаем из опыта работы на языке ассемблера, пространство памяти выполняемой задачи (т. е. процесса) состоит из нескольких *сегментов* — непрерывных адресных областей, обладающих определёнными свойствами; иногда их ещё называют *секциями*, что не вполне корректно, хотя и понятно, или даже *отображениями* (англ. *mapping*), что, напротив, совершенно корректно, но не очень понятно (позже мы узнаем, при чём тут «отображения», и всё с ними станет вполне понятно, но всему своё время). Пока будем использовать термин «сегмент».

В сегменте кода хранится в виде машинного кода собственно программа, выполняющаяся в данном процессе. Изменять содержимое этой области памяти процесс не может, что позволяет при запуске нескольких экземпляров одной и той же программы держать в памяти только одну копию её кода. В сегменте данных располагаются глобальные переменные и небезызвестная *куча*, из которой выделяются области памяти для динамических переменных. Наконец, сегмент стека используется для организации хорошо знакомого нам аппаратного стека, хранящего локальные переменные подпрограмм и адреса возврата из них. В пространстве памяти, принадлежащем процессу, могут присутствовать и другие сегменты, точный список которых зависит от конкретной системы.

Очень важной характеристикой процесса является *состояние регистров центрального процессора*, в том числе счётчика команд (он же указатель инструкции), регистра флагов, указателя стека и всех регистров общего назначения. Когда процесс по тем или иным причинам находится вне состояния выполнения, то есть он либо заблокирован, либо готов к выполнению, но процессор занят чем-то другим, содержимое регистров ЦП для этого процесса хранится в специальной структуре данных в ядре; когда процесс снова ставится на выполнение, данные из этой структуры копируются обратно в регистры.

За обеспечение иллюзии одновременного исполнения процессов отвечает подсистема ядра, которая называется ***планировщиком времени центрального процессора*** или, для краткости, просто планировщиком[16]. С точки зрения планировщика процесс обладает, во-первых, логическим значением *готовности*, показывающим, нужно ли данный процесс вообще ставить на выполнение (иными словами, *готов* ли он к дальнейшему исполнению); во-вторых, процессу приписывается некий *приоритет*, используемый планировщиком для выбора процесса из нескольких готовых к выполнению, а в некоторых системах ещё и для определения величины кванта времени. Например, приоритет может состоять из двух чисел, называемых ***статическим приоритетом*** и ***динамическим приоритетом***. Сложение этих двух чисел даёт общий приоритет процесса. Статический приоритет процессу задаётся извне, тогда как значение динамического приоритета планировщик меняет в зависимости от поведения процесса; пока процесс ожидает выполнения в состоянии готовности, его динамический приоритет повышается, когда же процесс находится в состоянии выполнения на центральном процессоре, его динамический приоритет снижается. Каковы бы ни были значения статической составляющей приоритета двух процессов, если один из них выполняется, а другой ожидает в очереди, за счёт изменения динамического приоритета рано или поздно приоритет второго окажется выше, чем приоритет первого, что приведёт к их «рокировке».

Динамическая составляющая приоритета остаётся внутренним делом планировщика и извне не контролируется, а вот статическая, которая в ОС Unix обычно называется *nice value*, может быть изменена с помощью системных вызовов `nice` и `setpriority`, что позволяет управлять планировщиком. Например, процессам, выполняющим долгие расчёты, можно снизить приоритет до минимального, чтобы они не мешали другим процессам, а занимали процессор только тогда, когда он всё равно свободен; в то же время процессам, обслуживающим внешние запросы, критичные по времени отклика, можно назначить высокий приоритет и т. д. Подробности мы приведём в §5.3.9.

Начиная ещё с вводной части, мы неоднократно обсуждали ***аргументы командной строки*** и ***переменные окружения*** (см. т. 1, соотв. §1.2.6 и §1.2.16). И то, и другое можно (и нужно) отнести к свойствам процесса; в ОС Linux через псевдофайлы, находящиеся в директории `/proc`, можно узнать и то, и другое для любого процесса в системе, если, конечно, хватит полномочий. Как добраться до аргументов командной строки, мы обсуждали в §4.3.20. Для хранения переменных окружения применяется точно такая же структура данных — массив указателей на строки, заканчивающийся элементом `NULL`; в каждой из строк хранится имя и значение одной переменной окружения, причём значение отделяется от имени знаком равенства.

Обычно новые запускаемые программы наследуют окружение от тех, кто их запустил, что позволяет передавать определённые настройки системы всем имеющимся процессам с возможностью некоторые процессы запустить с изменёнными настройками. Забегая вперёд, отметим, что наследование окружения обусловлено не свойством операционной системы, а библиотечными функциями, через которые обычно запускаются новые программы; с точки зрения ядра окружение для каждой запускаемой программы задаётся своё. Окружение доступно в программах на Си через глобальную переменную

```
extern char **environ;
```

Переменная содержит адрес начала массива указателей, каждый из которых хранит адрес строки, представляющей одну переменную окружения; конец массива, как сказано выше, помечен нулевым элементом. Довольно забавен

[16] Отметим, что соответствующий англоязычный термин — *CPU scheduler*, а не «planner», что иногда встречается в английских текстах русских авторов.

тот факт, что системные заголовочные файлы не содержат объявления этой переменной, так что, если она нужна, придётся объявить её самостоятельно — ровно так, как написано выше, со словом `extern`.

Существует другой способ доступа к окружению: можно в заголовке функции `main` добавить третий параметр, который тоже обычно называют `environ` или `envir`. Тип этого параметра такой же, как у хорошо знакомого нам параметра `argv` — `char**` или `const char * const *`.

Поскольку окружение располагается в памяти процесса, системные вызовы для работы с окружением не нужны. Для манипуляции переменными окружения служат библиотечные функции `getenv`, `setenv` и `unsetenv`:

```
char *getenv(const char *name);
int setenv(const char *name, const char *value, int overwrite);
void unsetenv(const char *name);
```

Функция `getenv` возвращает (в виде указателя на строку) значение переменной окружения, имя которой задаётся аргументом `name`. Если такой переменной в окружении нет, возвращается значение `NULL`. Очень важно произвести проверку на `NULL` перед анализом возвращённой строки: никто не может гарантировать наличие какой бы то ни было переменной в окружении процесса, даже если речь идет о, казалось бы, всегда определённых переменных, таких как `PATH` или `HOME`. Функция `setenv` устанавливает новое значение переменной, причём если такой переменной не было, значение устанавливается в любом случае, если же соответствующая переменная в окружении уже есть, новое значение устанавливается только при ненулевом значении параметра `overwrite`; иначе говоря, параметр `overwrite` разрешает или запрещает изменять значение переменной окружения, если она есть. Функция `unsetenv` удаляет переменную с заданным именем из окружения.

Работа функций `setenv` и `unsetenv` (по крайней мере в ОС Linux) никак не влияет на «настоящее» содержимое окружения, то, которое считает свойством процесса сама система; что бы вы ни делали с помощью этих функций, содержимое псевдофайла `/proc/`NNN`/environ` (где NNN — номер процесса) не изменится. Но изменить его процесс всё-таки может — путём записи в эту область памяти напрямую. Например, можно при старте программы запомнить значение указателя `environ[0]`, извлечь из окружения значения всех переменных, которые вам по какой-то причине интересны (или даже просто сформировать копию окружения и занести соответствующие адреса в элементы массива `environ`), после чего записать в память по ранее запомненному адресу всё, что вашей душе угодно, и полюбоваться на новое содержимое `/proc/`NNN`/environ`. Правда, зачем это делать — не вполне понятно, поскольку это содержимое ни на что не влияет, все библиотечные функции используют массив `environ` и строки, на которые его элементы указывают (а то, что эти строки теперь не в «настоящем» окружении, а в куче — никто толком и не знает).

Интереснее ситуация с командной строкой. Документированного способа её изменить не существует, но на вопрос, *зачем* её менять, ответ есть, и очень простой: чтобы некие данные о текущем состоянии программы можно было показать в выдаче команды `ps`; так поступают многие системные программы. В принципе, если просто записать какую-нибудь строку в память, на которую указывает `argv[0]`, вы именно эту строку и получите в выдаче `ps`, только прежде чем это делать, подумайте тридцать раз. Конкретно в ОС Linux вы при этом, скорее всего, затрёте часть «настоящего окружения», поскольку следом за командной строкой в памяти (на самом дне стека) располагается именно оно; что произойдёт в других системах — ведает разве что Ктулху.

К свойствам процесса в ОС Unix относятся также ***текущий каталог*** и ***корневой каталог***. С текущим каталогом всё просто: каждый процесс считается «находящимся» в одном из каталогов файловой системы; от текущего каталога, как мы видели ранее, отсчитываются «относительные пути» — имена файлов, начинающиеся с любого символа, кроме «/»; значение текущего каталога, таким образом, влияет на работу всех системных вызовов, в которые через параметры передаются имена файлов. Процесс может в любой момент сменить свой текущий каталог.

С корневым каталогом дела обстоят несколько сложнее. Unix позволяет ограничить файловую систему, видимую процессу и всем его потомкам, частью дерева каталогов, имеющей общий корень. Например, если установить процессу корневой каталог `/foo`, то под именем `/` процесс будет видеть каталог `/foo`, а под именем `/bar` — каталог `/foo/bar`. Каталоги за пределами `/foo` процессу и всем его потомкам вообще не будут видны ни под какими именами. Это используется для запуска отдельных программ в безопасном варианте — так, чтобы они могли получить доступ только к файлам, предназначенным специально для них.

Впрочем, полагаться на такие меры следует с осторожностью, особенно если вы не обладаете уверенными знаниями в этой области. Так, если за пределами нового корневого каталога есть работающие процессы, запущенные от имени того же пользователя, что и внутри, то выбраться из «клетки» — дело техники; что до процессов с полномочиями администратора — то для них (точнее, для тех людей, кто их контролирует) получение неограниченного доступа ко всей системе вообще не представляет никаких сложностей, изменённый корневой каталог этому помешать не способен.

Процесс может сменить текущий каталог с помощью вызова

```
int chdir(const char *path);
```

— подав в качестве параметра полный путь нового каталога либо путь относительно текущего каталога. Строка `".."` означает каталог уровнем выше: например, `/usr/local/share/..` — это то же самое, что `/usr/local`. Для смены корневого каталога предусмотрен вызов

```
int chroot(const char *path);
```

После выполнения этого вызова каталоги за пределом нового корневого перестают быть видны или каким-либо образом доступны процессу и всем его потомкам. Операция смены корневого каталога необратима. Вызов `chroot` могут выполнять только процессы, имеющие права администратора системы (пользователя `root`), то есть имеющие нулевой `uid`. Отметим, что **как текущий, так и корневой каталог нельзя изменить для другого процесса; процесс может сменить их только для самого себя**.

Ещё одно важное свойство процесса связано с открытыми потоками ввода-вывода: это ***таблица файловых дескрипторов***. Каждый поток ввода-вывода на самом деле представляет собой довольно сложную структуру данных в памяти ядра, причём конкретный вид этой структуры существенно зависит от природы потока — для потока, связанного с обычным дисковым файлом и для какого-нибудь сетевого сокета ядру приходится помнить совершенно разные сведения. Как мы уже неоднократно видели, с точки зрения процесса файловый дескриптор — это просто номер, причём это даже не номер потока ввода-вывода в масштабах всей системы, а номер открытого потока для данного процесса: так, дескрипторы `0`, `1` и `2` есть практически в каждом процессе, но связаны с совершенно разными потоками.

Таблица файловых дескрипторов процесса задаёт соответствие собственно файловых дескрипторов (то есть небольших целых чисел, с помощью которых различаются между собой открытые потоки ввода-вывода) объектам внутри ядра ОС, содержащим всё необходимое для функционирования этих потоков. Система предоставляет процессу широкий набор инструментов, влияющих на его таблицу дескрипторов. Два наиболее очевидных способа изменения содержимого этой таблицы — это открытие файла с вызовом `open` и закрытие его вызовом `close`. Позже мы узнаем, во-первых, другие способы создания потоков ввода-вывода, при которых новые дескрипторы появляются без использования `open`; во-вторых, в §5.3.10 мы рассмотрим системные вызовы, напрямую манипулирующие содержимым таблицы дескрипторов.

Обсуждая системный вызов `open`, мы упоминали параметр `umask`; он тоже является в ОС Unix свойством процесса. Напомним, что изменить его можно с помощью системного вызова, который так и называется `umask` (см. стр. 104).

Несколько позже, изучая так называемые ***сигналы***, мы увидим, что к свойствам процесса относится также ***диспозиция сигналов***, то есть набор указаний операционной системе, что следует делать, когда данному процессу приходит тот или иной сигнал. Также к свойствам процесса относятся *счётчики потреблённых ресурсов* (процессорного времени, памяти и т. п.) и некоторые другие свойства, рассмотрение которых мы оставим за рамками нашей книги.

5.3.3. Виртуальное адресное пространство

В первом томе, изучая ассемблер, мы были вынуждены обсудить ***секции***, из которых формируется исполняемый файл (поскольку в программах на языке ассемблера секции указываются явно), а с ними — и память пользовательской задачи, состоящую из отдельных ***сегментов*** (см. §3.2.2), но постарались сделать это по возможности кратко, поскольку там хватало других проблем. Сейчас, коль скоро на повестке дня оказались процессы и их свойства, самое время вернуться к структуре адресного пространства, каким его видит программа, выполняющаяся в пользовательском режиме под управлением операционной системы.

В большинстве случаев — в частности, на всех компьютерах общего назначения, начиная примерно с конца 1980-х, — в распоряжении операционной

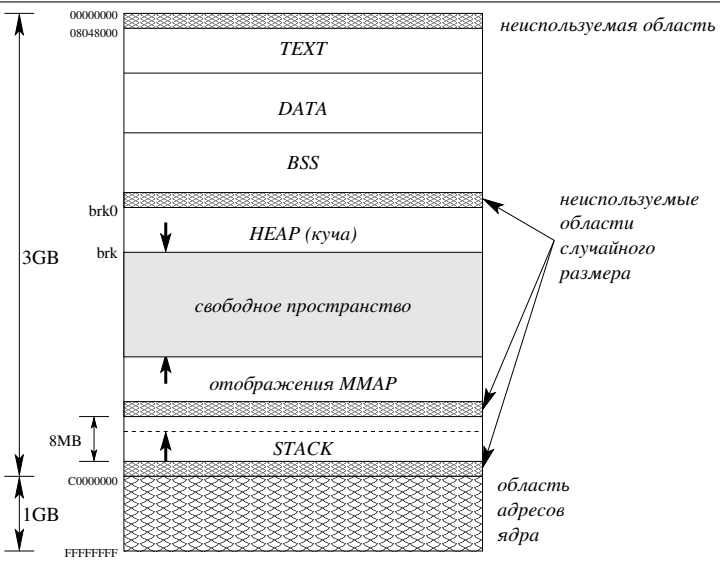

Рис. 5.3. Адресное пространство процесса

системы имеется механизм ***виртуальной памяти***, о котором мы тоже упоминали при изучении ассемблерного программирования (см. §3.1.2). Основная идея виртуальной памяти, напомним, состоит в том, что исполнительные адреса, фигурирующие в машинных командах, считаются не адресами физических ячеек памяти, а некими абстрактными ***виртуальными адресами***. Всё множество виртуальных адресов называется ***виртуальным адресным пространством***. Виртуальные адреса преобразуются центральным процессором — а точнее, специальной схемой в составе процессора, которая называется MMU, *memory management unit* — в адреса ячеек памяти (***физические адреса***) по некоторым правилам, причём эти правила могут динамически изменяться. В современных условиях преобразование виртуальных адресов в физические практически всегда происходит в соответствии с так называемой ***страничной моделью***. Саму эту модель мы обсудим в §8.3.5; пока нас интересует только два её базовых свойства: каждая задача в системе имеет своё собственное (виртуальное) адресное пространство, причём одна и та же физическая память может при необходимости соответствовать как разным адресам в виртуальном пространстве одной задачи, так и адресам из разных пространств.

Конечно, в том, как организовано адресное пространство процесса, многое зависит и от аппаратной платформы, и от конкретной системы. Например, бывают компьютеры без виртуальной памяти (то есть построенные на процессорах без MMU); в наши дни такими бывают только специализированные компьютеры, предназначенные для управления различной техникой, производственными процессами и т. п., но на некоторых из них можно запустить ядро Linux, и, очевидно, бóльшая часть того, что мы тут сейчас наговорим на тему организации адресного пространства процесса, не будет иметь ни малейшего отношения к тому, как это всё устроено на столь «хитрых» (на самом деле наоборот — слишком упрощённых) аппаратных архитектурах.

От одной операционной системы к другой организация адресного пространства тоже может различаться, в чём-то сильно, в чём-то едва заметно. Для примера мы рассмотрим структуру адресного пространства задачи в ОС Linux на всё той же нашей «любимой» (будь она трижды неладна) платформе i386. Схематически и, к сожалению, с некоторыми упрощениями эта структура показана на рис. 5.3. На всякий случай подчеркнём ещё раз, что речь идёт о *виртуальном* адресном пространстве; это значит, что неиспользуемые области, показанные на рисунке — это лишь области абстрактных виртуальных адресов, никакой настоящей (физической) памяти за этими адресами нет, то есть они не стоят ничего, кроме небольшого уменьшения доступного отдельной задаче множества виртуальных адресов — но этих адресов в любом случае обычно больше, чем нужно задаче.

Первая неиспользуемая область адресов начинается с нулевого адреса и простирается до «магического» `0x08048000`. Зачем она нужна и почему имеет именно такой размер — вопрос, конечно, интересный; судя по всему, точного ответа не знает вообще никто. На других архитектурах, в том числе на 64-битном Linux для x86_64, адрес нижней границы используется другой, да и на Linux/i386 конкретно такой адрес задаётся не ядром, а параметрами, указанными в исполняемом файле; ну а в исполняемый файл именно такое число попадает благодаря линкеру, и его при большом желании можно изменить. Надо признать, что некая (не очень большая) область неиспользуемых адресов в окрестностях «адресного нуля» довольно полезна — она позволяет отловить допущенные в программах ошибки, связанные с использованием нулевого указателя, в том числе для случаев, когда адрес, формально уже не нулевой, получен из `NULL` арифметическими действиями (например, делается попытка через `NULL` обратиться к элементу массива или полю структуры); конкретного значения это, конечно, не объясняет, но оно и не столь важно.

Начиная с адреса `0x08048000` в пространстве процесса располагаются хорошо знакомые нам сегменты кода и данных. В применении к Linux (да и к другим современным системам тоже) понятие «сегмента» здесь условно, ядро оперирует другим термином — ***отображение*** (англ. *mapping*). Название оправдано тем, что **многие (а в некоторых системах, например во FreeBSD — вообще *все*) участки виртуального адресного пространства процесса образуются с помощью механизма, реализованного в системном вызове `mmap`**; рассматривая его в §5.2.7, мы отметили, что этот вызов, изначально предназначенный для отображения в виртуальную память дисковых файлов, постепенно стал основным инструментом для работы с виртуальной памятью как таковой; самое время об этом вспомнить. Кстати, действующие отображения для любого процесса в Linux можно увидеть, выдав на печать (например, командой `cat`) файл `/proc/`NNN`/maps`, где NNN — `pid` процесса. Конечно, для каждого конкретного процесса эта информация доступна только его владельцу и администратору, то есть простой пользователь не может таким способом узнать информацию о процессах других пользователей.

Возвращаясь к диаграмме на рис. 5.3, заметим, что первые три отображения обозначены хорошо знакомыми нам словами `text`, `data` и `bss` — именно так назывались три ***секции***, которые мы использовали в программах на языке ассемблера. Надо сказать, что первые две области реально создаются как *отображения*, то есть при подготовке программы к запуску ядро отображает в новое виртуальное адресное пространство фрагменты исполняемого файла, используя механизм `mmap`. Дальнейшее во многом зависит от настроек линкера, использованных при создании этого файла; так, если в секции `.data` остаётся незадействованной часть последней (или единственной) страницы и туда помещается `.bss`, линкер может объединить эти две секции, и т. д., но в общем случае под `.bss` создаётся отдельное отображение, не связанное с файлом (анонимное). Также в виде анонимного отображения создаётся сегмент стека, а остальные области виртуальной памяти возникают уже во время работы процесса по его требованию.

Как уже говорилось, во многих системах для процесса остался лишь один способ потребовать от системы больше памяти — вызов `mmap`; но в ОС Linux наряду с этим механизмом сохранился системный вызов `brk`, который используется для создания и последующего увеличения памяти под ***кучу***, в которой, напомним, располагаются динамические переменные. Изначальная суть `brk` состояла в том, чтобы увеличить размер сегмента данных; этот вызов принимает один параметр — собственно новый адрес, где, с точки зрения процесса, должен теперь заканчиваться сегмент; если в качестве параметра указать `0`, вызов возвращает текущий адрес конца сегмента. Работа с `brk` обычно выглядит так: вызываем его от нуля, к полученному адресу добавляем число, кратное размеру страницы, и снова вызываем `brk`, в этот раз указав параметром полученный новый адрес.

В старых версиях Linux адрес, который возвращал вызов `brk(0)` сразу после старта программы, реально представлял собой адрес, непосредственно следующий за концом имеющегося сегмента данных, обычно содержащего переменные из секций `.data` и `.bss`. Современные версии ядра вставляют перед этим адресом случайное (то есть реально выбираемое случайным образом при каждом старте программы) количество неиспользуемых виртуальных страниц, так называемый *random brk offset*. Аналогичные случайные пустоты вставляются между «дном» стека и концом доступного адресного пространства (адресом `0xC0000000`), а также между сегментом стека и началом области адресов, выбираемых системой для новых отображений, создаваемых вызовом `mmap`. Наличие всех этих «дыр», имеющих случайно выбранный размер, вносит существенный элемент непредсказуемости в адреса любых объектов в памяти, создаваемых во время работы программы — локальных и динамических переменных, а равно и `mmap`-отображений, для чего бы они ни использовались. Зачастую эта непредсказуемость существенно затрудняет взлом системы через разнообразные программные уязвимости.

Размер стека определяется при старте программы[17], чаще всего это 8 MB. В большинстве unix-систем размер стека для запускаемых программ можно изменить, в том числе прямо в командной строке, но делают это редко. Важно понимать, что речь тут идёт именно об области виртуальных адресов, а не о настоящем выделении памяти; система изначально выделяет под стек ровно столько физической памяти, чтобы там можно было разместить переменные окружения, слова командной строки и массив указателей на них; дополнительные страницы выделяются по мере надобности, то есть лишь тогда, когда стек перестаёт помещаться в страницах, которые уже выделены: при обращении процесса к несуществующей пока части стека происходит исключительная ситуация («внутреннее прерывание»), что позволяет ядру вмешаться в происходящее, выделить очередную страницу физической памяти и продолжить выполнение процесса с того же места, а сам процесс, как водится, при этом ничего не замечает. Но при этом рост стека подчиняется сразу двум ограничениям — на размер самого стека и на общий объём областей виртуальной памяти, выделенных данному процессу; при выходе за любую из этих границ обращение к несуществующей странице стека приведёт к тому, что ядро уничтожит процесс как аварийный. Мораль здесь, кстати говоря, проста: экономьте стек!

В отличие от стека, чей предельный размер задан заранее и даже не слишком велик, области кучи и отображений `mmap` ограничены только объёмом адресного пространства и, возможно, уже упоминавшимся предельным объёмом памяти, выделенной одному процессу. К счастью, этих областей только две, что избавляет систему от необходимости предугадывать, сколько пространства потребуется каждой из них: им просто позволяется расти навстречу друг другу, постепенно захватывая свободное адресное пространство. Столкнуться они не могут: такое столкновение будет означать, что адресное пространство исчерпано, а лимит на количество выделенных процессу страниц, скорее всего, кончится раньше; зато любая из этих двух областей может вырасти на всё доступное пространство, если вторая этому не помешает.

Последняя область адресов, на которую стоит обратить внимание, располагается в старшей четверти пространства — с `0xC0000000` до `0xFFFFFFFF`. Процессу, то есть программе, которая в нём выполняется, эти адреса недоступны, любое обращение к ним приведёт к аварии. Эта часть адресного пространства задействуется, когда управление находится в ядре ОС. Дело в том, что ядро — это пусть и особенная, но всё же не более чем программа, и она точно так же, как и пользовательский процесс, вынуждена работать в виртуальном адресном пространстве. Конечно, ядро, в отличие от процесса, может это пространство перекраивать по своему усмотрению, но это не отменяет того факта, что любой исполнительный адрес, используемый ядром, процессор точно так же пропускает через MMU, как и при выполнении пользовательских задач. А ещё ядру часто требуется работать с памятью процессов — например, вызовы `read` и `write` вынуждают ядро копировать данные между памятью вызывающего процесса и своими буферами ввода-вывода; любые системные вызовы, принимающие параметрами *текстовые строки* (чаще всего — имена файлов, но не только), в действительности передают в ядро *адреса* этих строк, а сам текст из строки ядро вынуждено добывать из памяти процесса, пользуясь, как можно догадаться, *виртуальным* адресом, причём это ведь виртуальный адрес из пространства процесса, а не самого ядра.

Чтобы избежать постоянных перенастроек MMU, ядро поступает проще. Само оно для своей работы использует виртуальные адреса как раз из этой вот «старшей четверти» — `0xC0000000` до `0xFFFFFFFF`, а процессам, напротив, использовать виртуальные адреса из этого диапазона не позволяет. Когда же дело доходит до обслуживания конкретного процесса, ядро настраивает MMU, чтобы виртуальные адреса из младших трёх четвертей отображались на физические адреса, как это происходит в «окучиваемом» процессе, а адреса из старшей четверти продолжали отображаться на физические страницы в соответствии с правилами самого ядра — и в результате код ядра получает в своё распоряжение и память ядра, и память процесса (правда, только одного, так что между пространствами разных процессов перегонять данные намного сложнее).

В действительности там даже перестраивать ничего не надо, ядро обычно включает «свои» страничные отображения в страничные таблицы каждого из процессов; процессу они остаются недоступны из-за соответствующим образом настроенных атрибутов страниц, содержащих страничные таблицы, но на ядро эти ограничения не действуют, так что, когда дело доходит до обслуживания конкретного процесса, ядро просто *активирует его контекст* и продолжает работу, пользуясь теми же настройками MMU, что и процесс. Собственно говоря, пресловутое ***выполнение процесса в режиме ядра*** с точки зрения настроек виртуальной памяти выглядит именно так. Как физически всё это достигается, станет ясно после рассмотрения тонкостей страничной модели виртуальной памяти, которое мы отложим до §8.3.5.

Рассмотренная нами структура адресного пространства пользовательской задачи не учитывает некоторых совсем специфических моментов, которые у каждой операционной системы свои; скажем, ядро того же Linux отображает на адресное пространство процесса свою собственную страницу, которая в зависимости от версии называется *vsyscall* или *vdso*, и процесс, обращаясь к коду и данным, находящимся в этой странице, может в некоторых случаях избежать выполнения системных вызовов (например, через vdso можно узнать текущее системное время, не передавая для этого управление в ядро). Аналогичные нетривиальные хаки практикуются и в других системах. Увы, эти механизмы столь специфичны, запутанны и столь часто меняются, что мы предпочтём воздержаться от их рассмотрения; читатель без труда найдёт материалы об этом в Интернете по ключевым словам.

5.3.4. Порождение процесса

Единственный способ порождения процесса в ОС Unix — это создание копии существующего процесса[18]. Для этого используется системный вызов `fork`:

```
int fork();
```

Если произойдёт ошибка, `fork`, как и другие системные вызовы, вернёт значение `-1`; возможны ещё два варианта значения, возвращаемого из `fork`, и оба свидетельствуют об успешном создании копии. Новый процесс, созданный ядром в ответ на `fork`, будет точной копией родительского, за исключением следующих различий:

- порождённый процесс имеет свой идентификатор (`pid`), естественно, отличающийся от идентификатора родителя;
- параметр `ppid` порождённого процесса равен `pid`'у родительского процесса;
- счётчики потреблённых ресурсов порождённого процесса сразу после `fork` равны нулю;
- выполнение обоих процессов (родительского и порождённого) продолжается с первой инструкции, следующей сразу за вызовом функции `fork` (чаще всего это присваивание возвращаемого ею значения какой-либо переменной), причём в родительском процессе `fork` возвращает `pid` порождённого процесса, а в порождённом — «возвращает»[19] число `0`.

С формальной точки зрения можно отметить ещё целый ряд отличий нового процесса от родительского. Так, забегая вперёд, мы можем отметить, что порождённый процесс не наследует «заказ на доставку сигнала `SIGALRM`», сделанный вызовом `alarm` (см. §5.3.14; сам вызов `alarm` описан на стр. 129). Есть и другие различия, но они проявляются в таких свойствах процесса, которые мы рассматривать не будем.

Например, программа

```
#include <stdio.h>
#include <unistd.h>
int main()
{
    fork();
    fork();
    fork();
    printf("Hello\n");
    return 0;
}
```

напечатает восемь строк `"Hello"`. В самом деле, после первого вызова `fork` процессов станет два, каждый из них продолжит выполнение со следующей после вызова операции, в данном случае — со второго вызова `fork`, то есть каждый из двух процессов снова породит свою копию; полученные четыре

[17] Предельный размер стека устанавливается лимитом `RLIMIT_STACK`; работу с этими лимитами мы рассмотрим в §5.3.12.

[18] Некоторые системы семейства Unix имеют альтернативные возможности, такие как `clone` в ОС Linux, но эти возможности специфичны для каждой системы и их прямое использование не рекомендуется.

[19] Слово «возвращает» здесь взято в кавычки, поскольку на самом деле возвращать там нечего, ведь процесс только что появился; но всё выглядит именно так, как будто функция `fork` вернула значение.

процесса (два старых, два новых) продолжат выполнение с третьего `fork`, так что их станет восемь, и каждый из них выполнит `printf`.

В этом примере все процессы делали одно и то же. На практике обычно требуется, чтобы порождённый процесс делал не то же самое, что его родитель, а что-то другое. Для этого используется обычный `if`, проверяющий возвращённое вызовом `fork` значение, например, так:

```
p = fork();
if(p == 0) {
    /* действия порождённого процесса */
} else {
    /* действия родительского процесса */
}
```

или так:

```
p = fork();
if(p == 0) {
    /* действия порождённого процесса */
    exit(0);
}
/* действия родительского процесса */
```

В реальных программах, особенно таких, которые предполагается передавать пользователям, стоит добавить также проверку на ошибку, то есть на равенство `-1`; чаще всего неудача происходит из-за превышения установленного лимита на количество процессов (такие лимиты мы рассмотрим в §5.3.12), но бывают и другие ситуации — например, в ядре может исчерпаться доступная память.

После обращения к `fork` оба процесса работают параллельно («одновременно»); например, если в обоих процессах программа предусматривает выполнение какого-то действия, то без специальных ухищрений нельзя будет предсказать, какой из процессов успеет своё действие сделать раньше. Рассмотрим фрагмент программы, который порождает процесс-потомок и выдаёт два сообщения, одно из порождённого процесса, второе из родительского:

```
p = fork();
if(p == 0) {
    printf("I'm the child\n");
} else {
    printf("I'm the parent\n");
}
```

Возможны две ситуации. В первой порождённый процесс не успевает начать исполнение прежде, чем родительский дойдёт до вызова функции `printf`. Например, системе могло не хватить памяти и порождённый процесс был сослан в откачанном состоянии. В этом случае сначала будет напечатана фраза «I'm the parent», затем — фраза «I'm the child».

Вторая ситуация противоположна. К примеру, у родительского процесса может сразу после выполнения вызова `fork` истечь квант времени. При этом порождённый процесс, которому на сей раз хватило памяти, получит управление и успеет за свой квант времени выполнить печать. Тогда фразы появятся на экране в обратном порядке.

Ситуации, в которых результат зависит от конкретной последовательности независимых событий (обычно событий из работающих параллельно процессов) называются *ситуациями гонок* (соответствующий англоязычный термин — *race condition*; в русских переводах встречается также «ситуация состязаний»). Ситуации гонок часто возникают в параллельном программировании, то есть при наличии более чем одного потока управления. К их возникновению следует относиться очень внимательно, поскольку в некоторых случаях неучтённые ситуации гонок могут приводить к ошибкам и даже проблемам в безопасности. К этому вопросу мы вернёмся в седьмой части нашей книги, которая будет посвящена параллельному программированию.

После вызова `fork` оба процесса используют один и тот же сегмент кода; это возможно, т. к. сегмент кода не может быть модифицирован. Остальная память процесса, за исключением некоторых особых случаев, копируется[20]. Это означает, в частности, что в порождённом процессе присутствуют все переменные, существовавшие в родительском процессе, причём изначально они имеют те же значения, но изменения переменных в родительском процессе никак не отражаются на порождённом и наоборот. Копированию подвергаются открытые дескрипторы файлов, установленные обработчики сигналов и т. п.

К особым случаям, когда для части памяти процесса не создаётся копия при выполнении `fork`, относятся фрагменты виртуальной памяти, созданные вызовом `mmap` (см. §5.2.7) с использованием флага `MAP_SHARED`. Рассмотрим пример:

```
int pgs, pid;
int *ptr;
int size = 4096;
pgs = getpagesize();
size = ((size-1) / pgs + 1) * pgs;
ptr = mmap(NULL, size, PROT_READ|PROT_WRITE,
           MAP_SHARED|MAP_ANONYMOUS, 0, 0);
if(ptr == MAP_FAILED) {
    /* ... обработка ошибки ... */
}
pid = fork();
if(pid == 0) {
    /* ... порождённый процесс ... */
} else {
    /* ... родительский процесс ... */
}
```

В этом примере родительский и порождённый процессы имеют доступ к одному и тому же массиву целых чисел длиной 1024 элемента, если считать, что `int` занимает 4 байта. Массив доступен через указатель `ptr`, так что если один из процессов сделает присваивание `ptr[77] = 120`, то в *обоих* процессах выражение `ptr[77]` будет иметь значение 120.

Корректная работа с разделяемыми данными требует определённой квалификации. Если не соблюдать при этом целый ряд весьма неочевидных правил, ваша программа будет демонстрировать неожиданное поведение в самые неподходящие моменты. Работе с разделяемыми данными будет посвящена отдельная (седьмая) часть нашей книги; пока просто помните, что разделяемая память — это отнюдь не так просто, как может показаться на первый взгляд.

Отдельного замечания заслуживают дескрипторы открытых потоков ввода-вывода. Как мы отмечали в §5.2.3, поток ввода-вывода — это объект ядра ОС, представляющий собой достаточно сложную структуру данных, а те дескрипторы, которые мы видим в пользовательских программах — небольшие целые числа — представляют собой не более чем «номерки», про которые ядро помнит, какой из них соответствует какому объекту самого ядра. Все эти соответствия хранятся в ***таблице дескрипторов***, которая, как было сказано в §5.3.2, представляет собой одно из свойств процесса, так что номера дескрипторов оказываются для процесса локальными. Таблица дескрипторов при порождении процесса копируется, как и все остальные его свойства, так что в порождённом процессе оказывается точно такой же набор открытых потоков (и их дескрипторов), какой был в родительском. При этом важно понимать, что **потоки ввода-вывода как объекты ядра при порождении процесса не копируются**. Соответствующие дескрипторы в родительском и порождённом процессах ссылаются на *одни и те же объекты потоков*.

Проявляется это, например, в том, что указатель текущей позиции в файле (для потоков, у которых такой указатель есть — например, для обычных дисковых файлов) оказывается для родительского и порождённого процесса общим. Так, если один из них выполнит вызов `lseek`, а второй после этого попытается читать или писать, то такая операция чтения/записи пойдёт с той позиции, которую первый процесс установил своим `lseek`'ом. В §5.3.10 мы встретимся с другими ситуациями, когда для одного и того же потока ввода-вывода возникает больше одного номера дескриптора в пользовательских процессах.

5.3.5. Замена выполняемой программы

Запустить на выполнение программу в ОС Unix можно путём *замены выполняемой программы* в рамках одного процесса. Концепция «замены программы в рамках процесса» может создать несколько удручающее впечатление, поскольку, как мы обсуждали в начале этой главы, процесс — это нечто создаваемое операционной системой, чтобы выполнять программы, и при таком подходе к объяснению вполне естественным оказывается восприятие *программы* как «главной сущности», а процесса — как чего-то вторичного

[20] В современных системах обычно процессы продолжают разделять страницы памяти до тех пор, пока один из них не попытается ту или иную страницу модифицировать; копия страницы создаётся только после этого. Такая техника называется *copy-on-write*, COW.

по отношению к программе. В принципе это так и есть: процессы созданы для программ, а не наоборот. Несмотря на это, как мы сейчас увидим, работающую программу можно заменить на другую программу, которая начнёт своё выполнение с её начала (например, с функции `main` для программ, написанных на Си), а *процесс* — то есть тот объект, который создала для выполнения программы операционная система — останется при этом тот же, который был. Больше того, в системах Unix это *единственный* способ запуска программ, другого просто нет.

Такой подход на первый взгляд может показаться (и действительно многим кажется) нелогичным, но у него есть несомненные достоинства. Дело в том, что при замене программы процесс не просто «формально» остаётся тем же самым — это действительно продолжает существовать тот же объект ядра операционной системы, и он сохраняет целый ряд свойств, таких как полномочия процесса, открытые потоки ввода-вывода и многое другое. Прежде чем запускать внешнюю программу, мы можем подвергнуть процесс предварительной настройке, затрагивая лишь те свойства, которые нужно поменять, и не отвлекаясь на постороннее.

Если бы процесс порождался непосредственно при запуске внешней программы, пришлось бы задать *все* его свойства непосредственно в том месте нашего кода, где мы его порождаем. Например, WinAPI содержит для порождения процессов целое семейство функций, каждая из которых требует указания от 9 до 11 параметров, причём некоторые параметры представляют собой указатели на довольно развесистые структуры данных.

Замена программы, выполняющейся в рамках процесса, на новую программу производится с помощью системного вызова `execve`:

```
int execve(const char *name, char * const *argv,
                             char * const *envir);
```

Параметр `name` задаёт исполняемый файл программы, которую нужно запустить на выполнение вместо текущей; файл можно задать как полным путём, так и относительно текущего каталога — как и во всех системных вызовах, использующих имена файлов. Через оставшиеся два параметра передаются (в виде массивов указателей на строки) аргументы командной строки и набор переменных окружения, то есть эти два параметра определяют, что́ получит запускаемая программа через свои аргументы функции `main`[21] и глобальную переменную `environ` (см. стр. 111), либо их аналоги, если программа написана не на Си.

К системному вызову `execve` напрямую обращаются редко; обычно намного удобнее воспользоваться одной из библиотечных функций, реализованных через тот же `execve` и образующих так называемое «семейство `exec`». Начнём его рассмотрение с функции `execv`:

```
int execv(const char *name, char * const *argv);
```

От вызова `execve`, как можно заметить, эта функция отличается отсутствием параметра `envir`. Окружение для запускаемой программы она берёт такое же, как было; говорят, что окружение *наследуется*.

Важную роль играет функция `execvp`; профиль у неё такой же, как у `execv`:

```
int execvp(const char *name, char * const *argv);
```

Эта функция отличается от предыдущих подходом к обработке первого параметра: она умеет *отыскивать запускаемую программу по её имени в «системных» директориях*, то есть в директориях, перечисленных в ***переменной окружения*** `PATH`[22]. Поиск производится, если в строке, переданной параметром `name`, нет ни одного символа «/»; так, если значение переменной `PATH` содержит директорию `/bin`, то вызвать программу `/bin/ls` можно просто по имени «`ls`», не указывая полный путь. **Если в первом параметре есть хотя бы один символ «/», функция `execvp` работает точно так же, как и `execv`**. Между прочим, именно эта особенность `execvp` — причина того, что для запуска свежеоткомпилированных программ в ходе наших упражнений мы вынуждены пользоваться «хитрыми» командами вроде `./prog` вместо просто `prog`. Командный интерпретатор сам написан на Си и использует для запуска внешних программ именно функцию `execvp`, ну а префикс «`./`» обеспечивает наличие слэша в параметре `name`, отключая тем самым поиск по директориям из `PATH` и превращая имя программы в обычное относительное имя файла.

Часто встречаются случаи, когда уже на этапе написания исходной программы нам известно точное количество параметров командной строки для программы, которую мы собираемся запустить с помощью `exec`. Если это так, нам нет необходимости самим формировать массив указателей на элементы командной строки, как это требуется для функций семейства `exec`, рассмотренных выше. Вместо этого можно использовать функцию `execl` или `execlp`, которые сформируют этот массив за нас:

```
int execl(const char *name, const char *argv0, ...);
int execlp(const char *name, const char *argv0, ...);
```

Эти функции получают произвольное число аргументов, первый из которых задаёт исполняемый файл, остальные — аргументы командной строки. Чтобы функция «знала», где остановиться, после последнего слова командной строки нужно добавить ещё один параметр со значением `NULL`[23]. Напомним, что командная строка включает нулевой элемент, под которым подразумевается имя самой программы; таким образом, аргумент `argv0` — это не первый аргумент командной строки, а нечто имеющее отношение к имени программы, в большинстве случаев значение `argv0` попросту совпадает с `name`. Различие между `execl` и `execlp` в том, что первая требует указания явного пути к исполняемому файлу, тогда как вторая выполняет поиск по переменной `PATH`, подобно тому, как это делает `execvp`. Всё, что было сказано выше относительно первого аргумента `execvp`, точно так же применимо и к первому аргументу `execlp`.

Допустим, требуется выполнить команду `ls -l -a /var`. Это можно сделать, например, так:

```
char *cmdline[] = { "ls", "-l", "-a", "/var", NULL };
execvp("ls", cmdline);
```

либо так:

```
execlp("ls", "ls", "-l", "-a", "/var", NULL);
```

Повторим, что все функции семейства `exec` *заменяют* в памяти процесса выполнявшуюся (и вызвавшую `exec`) программу на другую, указанную в параметрах вызова. Поэтому в случае успеха эти функции управление уже не возвращают: в самом деле, программы, в которую можно было бы вернуть управление, уже нет, вместо неё работает новая программа. В случае ошибки возвращается значение `-1`, но проверять его бессмысленно, ведь сам факт возврата управления свидетельствует об ошибке. Чаще всего после любой из функций `exec` следующая строка программы представляет собой вызов знакомой нам функции `perror`. Поскольку в большинстве случаев `exec` выполняется в процессе, созданном специально для этого, а после ошибки этот процесс уже больше не нужен, после `perror` ставится вызов функции `exit`, завершающей процесс. **Если после вызова функции семейства `exec` в тексте программы мы видим что-то отличное от связки `perror`/`exit`, в большинстве случаев здесь что-то не так;** ситуации, когда после `exec` не ставятся вызовы этих двух функций, редки и достаточно необычны.

Строго говоря, следует учитывать, что функция `exit` перед завершением процесса может сделать что-то ещё; в подавляющем большинстве случаев это «что-то ещё» сводится к вытеснению буферов вывода в функциях из `stdio.h`. Если вы ухитритесь вызвать `fork`, когда некоторые из ваших высокоуровневых потоков вывода содержат неотправленные данные, эти данные окажутся скопированы вместе со всеми данными родительского процесса, а функция `exit`, вызванная из порождённого процесса, произведёт их вытеснение. Родительский процесс об этом, разумеется, ничего знать не будет и рано или поздно, скорее всего, вытеснит свой экземпляр скопированных данных. Результат может получиться довольно своеобразный.

Основываясь на этом соображении, некоторые авторы рекомендуют после `exec` использовать `_exit`, а не обычный `exit`, то есть обращаться сразу же к ядру системы с просьбой немедленно завершить ваш процесс без всякой предварительной подготовки. Рекомендация сама по себе правильная, но здесь следует учесть один важный момент.

[21] По идее к этому моменту у читателя уже не должно быть проблем с аргументами командной строки. Если они всё же есть, перечитайте §4.3.20; возможно, по ссылкам из него придётся вернуться к первому тому, и если вы чувствуете малейшую неуверенность — обязательно сделайте это.

[22] Мы уже обсуждали и переменную `PATH`, и само окружение в целом; если всё-таки остались какие-то проблемы с пониманием, что такое эта `PATH` и как с ней бороться, *обязательно* вернитесь к первому тому и перечитайте §1.2.16, иначе дальше станет хуже.

[23] В результате придирчивого анализа текстов разнообразных стандартов многие программисты приходят к выводу, что использовать следует не `NULL`, а выражение `((char*)0)`. Судя по всему, в реально существующих системах никакой разницы нет.

Диагностический поток, в который выводит сообщение функция `perror`, тоже *может* оказаться буферизуемым: несмотря на то, что в современных реализациях стандартной библиотеки буферизация для `stderr` обычно выключена (см. §4.4.5), полагаться на «добрую» библиотеку всё-таки не стоит. Если `stderr` по каким-то причинам окажется буферизуемым по общим правилам и при этом пользователь его куда-нибудь перенаправит, так что он не будет связан с терминалом, то при завершении процесса с помощью `_exit` данные из его буфера вытеснены не будут — попросту говоря, сообщение, выданное `perror`'ом, так и не покинет пределы своего процесса и сгинет вместе с ним (напомним, что выдача перевода строки приводит к вытеснению буфера, только если вывод идёт на терминал; поток `stderr` служит исключением из общего правила, причём это исключение введено явочным порядком и нигде не закреплено).

По-видимому, «совсем правильной» будет примерно такая последовательность вызовов:

```
execlp("ls", "ls", "-l", NULL);
perror("ls");
fflush(stderr);
_exit(1);
```

К этому можно добавить ещё вытеснение буфера потока `stderr` *перед* порождением нового процесса, то есть непосредственно перед вызовом `fork` тоже стоит вставить `fflush(stderr)`, в противном случае (особенно когда `stderr` куда-то перенаправлен) вы рискуете некоторые диагностические сообщения, выданные родительским процессом, увидеть дважды. Впрочем, это уже не столь серьёзный риск, а `stderr` всё-таки обычно не буферизуется.

Подробное обсуждение различных вариантов завершения процесса вы найдёте в следующем параграфе.

Отметим, что **открытые файловые дескрипторы при выполнении `exec` остаются открытыми**, что позволяет перед запуском внешней программы произвести манипуляции с дескрипторами. Это свойство `exec` используется для реализации хорошо знакомых нам перенаправлений ввода-вывода.

Из этого правила есть одно исключение. На файловом дескрипторе может быть установлен флаг `FD_CLOEXEC` (*close-on-exec*), и в этом случае файл — точнее, один отдельно взятый дескриптор — будет закрыт при успешно отработавшей замене программы.

Наиболее общепринятым способом установки этого флага можно считать вызов `fcntl` (см. стр. 106). В отличие от флагов статуса, таких как уже известные нам `O_APPEND` или `O_NONBLOCK`, флаг *close-on-exec* относится не к потоку ввода-вывода, а к конкретному файловому дескриптору, связанному с этим потоком. Таких дескрипторов, как мы уже понимаем, может стать больше одного, например, после выполнения вызова `fork`; при обсуждении перенаправления ввода-вывода мы узнаем о других способах «размножения» дескрипторов, относящихся к одному потоку ввода-вывода. Так или иначе, флаги статуса, которые мы обсуждали раньше и которые можно изменить с помощью команды `F_SETFL` вызова `fcntl`, относятся к потоку ввода-вывода и действуют вне зависимости от того, через какой из, возможно, нескольких дескрипторов этого потока мы работаем; флаг *close-on-exec*, напротив, относится к конкретному дескриптору.

В связи с этим для работы с флагом *close-on-exec* вызов `fcntl` предусматривает отдельные команды — `F_SETFD` и `F_GETFD`. Изначально предполагалось, что эти команды будут работать с флагами, относящимися к дескриптору (в отличие от потока), но в итоге был введён всего один такой флаг — `FD_CLOEXEC`. Установить его можно, выполнив

```
fcntl(fd, F_SETFD, FD_CLOEXEC);
```

Узнать, установлен флаг или нет, можно с помощью `fcntl(fd, F_GETFD)` (такой вызов вернёт значение `FD_CLOEXEC`, если флаг установлен, и `0` — если нет), но это почти никогда не нужно.

С некоторых пор в наиболее новых версиях unix-систем флаг *close-on-exec* можно установить в момент открытия файла, добавив во второй параметр вызова `open` константу `O_CLOEXEC`, но это поддерживается не везде и не факт, что будет хорошей идеей эту возможность использовать.

5.3.6. Завершение процесса

В системах семейства Unix предусмотрено два способа завершения процесса: самостоятельное и принудительное. В первом случае процесс сам обращается к ядру операционной системы с просьбой завершить его, и эта просьба немедленно удовлетворяется. Во втором случае кто-то — другой процесс либо сама операционная система — присылает процессу ***сигнал***, и процесс завершается, получив этот сигнал.

Программируя на языке ассемблера, мы узнали о существовании системного вызова `_exit` (см. т. 1, §3.6.5), который предусматривает один параметр — ***код завершения процесса***. Сам этот код мы упоминали ещё при программировании на Паскале, ведь оператор `halt` позволяет указать значение кода завершения. Изучая язык Си, мы завершали программы, возвращая управление из функции `main` оператором `return`, а для досрочного завершения программы использовали библиотечную функцию `exit` (см. §4.2.2). Написав на Си программу без стандартной библиотеки, мы на собственном опыте убедились, что функцию `main` «кто-то» (на самом деле — невидимая подпрограмма `_start`) вызывает в составе выражения `exit(main())`, то есть возврат значения из «главного» экземпляра `main` и прямой вызов `exit` есть в сущности одно и то же. Как несложно догадаться, паскалевский `halt` делает то же самое; все эти способы завершения в конечном счёте приводят к системному вызову `_exit`[24], который из программы на Си можно вызвать и напрямую. Профиль его обёртки таков:

```
void _exit(int code);
```

Параметр `code` задаёт всё тот же *код завершения процесса*. Напомним, что значение 0 означает успешное завершение, значения 1, 2, 3 и т. д. — что произошла та или иная ошибка или неудача. Обычно используются значения, не превышающие 10, хотя это не обязательно; следует только учитывать, что сам по себе код завершения процесса — это беззнаковое восьмибитное целое, т. е. в терминах Си — значение типа `unsigned char`; поэтому не следует пытаться использовать отрицательные значения или значения, превосходящие 255, это всё равно не получится. Прямое обращение к системе с требованием завершить текущий процесс, как можно догадаться, немедленно завершает процесс, никаких иных вариантов тут не предусмотрено. В этом плане вызов библиотечной функции `exit`, а равно и возврат значения из функции `main` (который тоже приводит к вызову именно библиотечной функции `exit`, а не к прямому обращению к ОС) существенно отличаются: конечно, функция `exit` тоже рано или поздно обратится к операционной системе, причём с помощью того же `_exit`, но перед этим она может «привести дела в порядок»: в частности, при использовании библиотечного («высокоуровневого») интерфейса ввода-вывода, т. е. функций, вводимых заголовочником `stdio.h`, функция `exit`, прежде чем завершить программу, вытеснит всю информацию из буферов вывода (см. §4.4.5), тогда как `_exit`, будучи системным вызовом, этого не сделает. Эту разницу легко обнаружить на примере следующей программы (обратите внимание, что перевод строки в `printf` отсутствует, это сделано специально, иначе эффекта не возникнет):

```
#include <stdio.h>
#include <unistd.h>
int main()
{
    printf("Hello, world");
    _exit(0);
}
```

Если эту программу запустить, она *ничего не напечатает*. Дело в том, что `printf` поместит строку в промежуточный буфер вывода и на этом успокоится; поскольку в строке нет символа перевода строки, вытеснения буфера не произойдёт. Далее вызов `_exit` завершит программу немедленно, не оставив библиотеке шанса вытеснить буфер. Если заменить `_exit` на обычный `exit` (то есть вызвать библиотечную функцию вместо прямого обращения к системному вызову), фраза «`Hello, world`» будет благополучно напечатана, и то же самое произойдёт, если завершить `main` традиционным «`return 0;`».

Чуть более интересные результаты получатся, если добавить к печатаемому сообщению символ перевода строки, как мы это делали раньше, а для завершения использовать `_exit`:

```
#include <stdio.h>
#include <unistd.h>
int main()
{
    printf("Hello, world\n");
    _exit(0);
}
```

[24] Или его аналогу; современные реализации стандартной библиотеки Си в ОС Linux обычно используют вызов `exit_group`.

Если в таком виде программу запустить без перенаправления вывода, фраза окажется напечатана, а вот если вывод перенаправить (в файл, в канал, в сокет, куда угодно, лишь бы не на терминал) — напечатано ничего не будет:

```
avst@host:~/work$ ./h
Hello, world
avst@host:~/work$ ./h | cat
avst@host:~/work$ ./h > file
avst@host:~/work$ cat file
avst@host:~/work$
```

Причина в том, что вытеснение буфера по символу перевода строки производится только при выводе на терминал. Ещё более загадочной выглядит на первый взгляд работа следующей программы:

```
#include <stdio.h>
#include <unistd.h>
int main()
{
    printf("Hello, world\n");
    fork();
    return 0;
}
```

Если её вывод не перенаправлять, `Hello, world` печатается, как обычно; но если вывод перенаправить — опять же, куда угодно, кроме как на терминал — сообщение окажется напечатано *дважды*:

```
avst@host:~/work$ ./h
Hello, world
avst@host:~/work$ ./h | cat
Hello, world
Hello, world
avst@host:~/work$ ./h > file
avst@host:~/work$ cat file
Hello, world
Hello, world
avst@host:~/work$
```

Эту загадку оставим читателю; информации, чтоб её разгадать, было дано уже более чем достаточно.

Так или иначе, и `_exit`, и `exit`, и возврат управления из `main` — это случаи, когда процесс завершается самостоятельно, указав значение кода завершения. Как было сказано выше, возможен другой случай — когда процессу приходит ***сигнал***. Подробное рассмотрение сигналов оставим для отдельного параграфа; сейчас отметим лишь два момента.

Во-первых, о *коде завершения* при завершении по сигналу речи идти не может, поскольку процесс не дошёл до своего вызова `_exit` и, как следствие, никакого кода завершения не указал; с другой стороны, имеется информация о *номере сигнала*, который стал причиной досрочного завершения процесса. Обычно номер сигнала — это целое число от `1` до `31`.

Во-вторых, все варианты аварийного завершения программы вследствие её собственных некорректных действий, таких как нарушение защиты памяти, деление на ноль и тому подобное, представляют собой частный случай завершения по сигналу. Когда процесс, работающий под управлением ОС Unix, становится причиной возникновения внутреннего прерывания (исключения) и ядро видит, что к этому привели некорректные действия самого процесса, оно отправляет процессу один из сигналов, предусмотренных специально для этих случаев. О том, что́ же произошло, мы можем узнать по номеру сигнала.

Отметим ещё один важный момент. Как мы видим, при завершении процесса возникает некоторое количество информации об обстоятельствах его завершения: самостоятельно он завершился или был убит сигналом, если самостоятельно — то с каким кодом завершения, если сигналом, то с каким номером. К этому следует добавить ещё информацию из счётчиков потреблённых процессом ресурсов системы, и всё это нужно где-то хранить, а идентификатор (`pid`) процесса нужно защитить от повторного использования до тех пор, пока кто-то не востребует всю эту информацию — ведь она связана с конкретным процессом, а процесс можно надёжно идентифицировать только его `pid`'ом. Если под тем же `pid`'ом в системе появится другой процесс, путаницы не избежать. Поэтому **в ОС Unix процесс при его завершении не исчезает**, а переходит в так называемое ***состояние процесса-зомби***. Тому, как с этими зомби обращаться, мы посвятим следующий параграф.

5.3.7. Ожидание завершения; процессы-зомби

Итак, после завершения процесса в системе остаётся информация о том, при каких обстоятельствах он завершился (сам ли он завершился, если да — то с каким кодом завершения, если нет — то каким сигналом он уничтожен) и значения счётчиков потреблённых ресурсов. Эту информацию должен затребовать родительский процесс; если родительский процесс завершается раньше своего непосредственного потомка, функции родительского берёт на себя процесс `init` (процесс номер `1`), при этом даже значение `ppid` (идентификатора родительского процесса) для «осиротевшего» процесса становится равным `1`. Отметим, что больше никто информацию о том, как процесс завершился, получить не может — это прерогатива его непосредственного предка либо «исполняющего обязанности предка» — процесса № `1`.

Завершённый процесс продолжает существовать в системе в виде процесса-зомби, то есть занимает место в таблице процессов до тех пор, пока находящаяся в нём информация об обстоятельствах завершения не будет затребована родительским процессом. Затребовать информацию и убрать зомби-процесс из системы позволяют системные вызовы семейства `wait`. Простейший из них имеет следующий прототип:

```
int wait(int *status);
```

Если у процесса нет ни одного непосредственного потомка, вызов завершается с ошибкой (возвращает `-1`). Когда порождённые процессы есть, но ни один из них ещё не завершился, то есть нет ни одного зомби, которого можно снять прямо сейчас, вызов *ждёт завершения любого из порождённых процессов*; отсюда название вызова — *wait* (англ. *ждать*). Дождавшись появления зомби (либо если зомби среди потомков данного процесса уже присутствовали на момент обращения к вызову), `wait` изымает из процесса-зомби хранящуюся в нём информацию об обстоятельствах завершения процесса, а сам процесс окончательно отправляет в небытие, освобождая слот таблицы процессов. При этом вызов возвращает `pid` завершившегося процесса, то есть того зомби, который только что окончательно исчез из системы. Если параметр представлял собой ненулевой указатель, то в целочисленную переменную, на которую он указывал, записывается информация о коде завершения процесса или о номере сигнала, по которому процесс был снят.

Для анализа информации, занесённой в такую переменную, используются макросы `WIFEXITED`, `WIFSIGNALED` (был ли процесс завершён обычным способом или по сигналу), `WEXITSTATUS` (если завершён обычным образом, то каков код завершения), `WTERMSIG` (если по сигналу, то каков номер сигнала). Например:

```
int status, wr;
wr = wait(&status);
if(wr == -1) {
    printf("There are no child processes at all\n");
} else {
    printf("Process with pid=%d finished.\n", wr);
    if(WIFEXITED(status)) {
        printf("It has exited with code=%d.\n",
            WEXITSTATUS(status));
    } else {
        printf("It was killed by signal %d.\n",
            WTERMSIG(status));
    }
}
```

Более гибкие возможности предоставляет системный вызов `wait4`:

```
int wait4(int pid, int *status, int opt, struct rusage *usage);
```

В качестве первого параметра вызова `wait4` можно указать идентификатор конкретного процесса, либо `-1`, если требуется дождаться любого из порождённых процессов. Учтите, что при наличии нескольких потомков ждать какого-то одного из них рискованно: остальные могут завершиться раньше, перейти в статус зомби, но убирать их будет некому.

Есть также возможность дождаться процесса из определённой группы процессов. Группы процессов мы рассмотрим позднее, но на всякий случай отметим, что нулевое

значение параметра pid соответствует ожиданию любого из непосредственных потомков, оставшихся в одной группе с родительским процессом, тогда как отрицательное значение, ме́ньшее -1, означает ожидание завершения процесса из группы с заданным номером (например, -2735 означает группу 2735). Естественно, речь по-прежнему идёт только о непосредственных потомках, всех остальных «дождаться» нельзя.

Параметр `status` используется так же, как для предыдущего вызова: через него передаётся адрес переменной типа `int`, в которую вызов записывает основную информацию об обстоятельствах завершения процесса. В качестве значения параметра `opt` можно указать число `0` или константу `WNOHANG`; в этом случае вызов не ждёт завершения процессов: если ни одного подходящего зомби нет, вызов немедленно возвращает значение `0`. Что касается параметра `usage`, то если он ненулевой, в указанную этим параметром область памяти записываются значения счётчиков ресурсов завершившегося процесса. Как выглядит структура `struct rusage`, можно узнать из документации.

Вызов `wait4` возвращает `-1` в случае ошибки, `0` в случае, если использовалась опция `WNOHANG` и завершившихся процессов не было, и `pid` завершившегося процесса, если вызов успешно получил информацию из зомби.

В ОС Unix предусмотрены также системные вызовы `waitpid` и `wait3`, оба от трёх параметров:

```
int wait3(int *status, int opt, struct rusage *usage);
int waitpid(int pid, int *status, int opt);
```

Первый эквивалентен `wait4` с `-1` в качестве параметра `pid`, второй — `wait4` с `NULL` в качестве параметра `usage`. На вопрос «зачем их столько», к сожалению, невозможно ответить иначе как сакраментальной фразой об «исторических причинах».

Предостережём читателя от одной довольно популярной ошибки. Часто можно наблюдать, как начинающие программисты и сисадмины пытаются «убить зомби» с помощью команды `kill` и удивляются, что ничего не выходит. Так вот, **зомби нельзя убить, поскольку он уже мёртвый**. Если в вашей системе наблюдаются процессы-зомби, ни команда `kill`, ни какие-то иные ухищрения вам не помогут, убрать зомби из системы может лишь тот процесс, который его породил. Впрочем, вы можете попробовать «пристрелить» нерадивого «родителя», забывшего про свои обязанности, и тогда зомби, скорее всего, исчезнет — его уберёт из системы процесс № `1`.

Отметим ещё один момент. **Подумайте двести раз, прежде чем использовать `wait4` или `waitpid` с параметром `pid`, отличным от `-1`, и при этом без опции `WNOHANG`.** Скорее всего, у вашего процесса при этом есть или как минимум *может быть* больше одного потомка (в противном случае зачем бы вам указывать в явном виде, кого ждать). Так вот, заблокировав свой процесс в ожидании завершения одного конкретного потомка, вы предоставите всем остальным своим потомкам замечательную возможность превращаться в зомби и оставаться в этом состоянии сколь угодно долго — а точнее, пока не завершится, наконец, тот единственный потомок, которого вы ждёте. Такое поведение программы часто расценивается как моветон; вообще зомби довольно редко попадаются в выдаче команды `ps ax`, но если пользователь смог одного и того же зомби увидеть дважды — значит, с программой, которая этого зомби изначально запустила, что-то не так.

Если требуется дождаться завершения конкретного процесса, намного правильнее будет организовать цикл вызовов `wait` (причём, возможно, простого `wait` с одним параметром) и каждый раз проверять, дождались вы кого надо или кого-то другого, сравнивая с нужным `pid`'ом значение, которое вернул `wait`.

5.3.8. Пример запуска внешней программы

Приведём пример запуска внешней программы с помощью связки `fork+exec`. Наша программа сначала запустит на выполнение (в качестве внешней программы) команду `ls` с параметрами «`-l -a /var`», а затем, дождавшись её завершения, выдаст сообщение «`Ok`».

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/types.h>
#include <sys/wait.h>
int main()
{
    int pid;
    pid = fork();
    if(pid == -1) {             /* ошибка порождения процесса */
        perror("fork");
        exit(1);
    }
    if(pid == 0) {                      /* порождённый процесс */
        execlp("ls", "ls", "-l", "-a", "/var", NULL);
        perror("ls");     /* exec вернул управление -> ошибка */
        exit(1);        /* завершаем процесс с кодом неуспеха */
    }
    /* родительский процесс */
    wait(NULL); /* дожидаемся завершения порождённого процесса,
                                        заодно убираем зомби */
    printf("Ok\n");
    return 0;
}
```

5.3.9. Выполнение процессов и время

Очевидно, что ни одна программа не может быть выполнена мгновенно; любое выполнение происходит *с течением времени*. Ядро операционной системы хранит текущее время и поддерживает ***системные часы***; это позволяет не только узнать, который час, но и при необходимости привязать выполнение программ к определённым моментам времени или временны́м промежуткам.

Узнать текущее системное время можно с помощью системного вызова `time`:

```
time_t time(time_t *t);
```

Этот вызов возвращает текущее время в виде *количества секунд, прошедших с начала 1 января 1970 года, причём по Гринвичу*. По-английски такой способ измерения времени называется «*since Epoch*».

Примечательно, что третий том первого издания нашей книги, как раз содержавший в том числе и эту главу, был подписан к печати в день, когда число секунд since Epoch перевалило за полтора миллиарда, о чём автор с удовольствием написал, прежде чем отправить рукопись в издательство. Это было 14 июля 2017 года; полтора миллиарда секунд с 1 января 1970 г. исполнилось в 5:40 по московскому времени. Два миллиарда исполнится 18 мая 2033 г., а за первый миллиард это число перевалило в 2001 году.

То же число вызов записывает в переменную типа `time_t`, адрес которой передан ему параметром, если только параметр не равен `NULL`; в большинстве случаев его оставляют нулевым и используют значение, возвращаемое `time` как функцией.

Тип `time_t` — это в большинстве случаев обыкновенный `long`, но если вы собираетесь описать переменную, чтобы её адрес передать параметром вызову `time`, лучше описать её как имеющую тип `time_t`, чтобы не возникало ошибки, если вдруг в какой-то из версий библиотеки это окажется не так. В Си все целые типы совместимы по присваиванию, но указатель по понятным причинам должен указывать именно на такую переменную, которую от него ожидают.

Конечно, работать с числом секунд, прошедших с далёкого 1970 года, не слишком удобно, но, к счастью, в библиотеке содержатся функции, несколько облегчающие нам жизнь. Функции `gmtime` и `localtime`, а также их «осовремененные» версии `gmtime_r` и `localtime_r`:

```
struct tm *gmtime(const time_t *timep);
struct tm *gmtime_r(const time_t *timep, struct tm *result);
struct tm *localtime(const time_t *timep);
struct tm *localtime_r(const time_t *timep, struct tm *result);
```

— позволяют, имея время в виде секунд *since Epoch*, разложить его на составляющие — год, месяц, число, часы, минуты и секунды. Результат представляется в виде структуры `struct tm`, имеющей поля `tm_year` (номер года за вычетом 1900 — например, 2016 год представляется числом `116`), `tm_mon` (месяц, от `0` до `11`), `tm_mday` (число, от `1` до `31`), `tm_hour` (часы, от `0` до `23`), `tm_min` и `tm_sec` (минуты и секунды, от `0` до `59`); кроме того, структура содержит поля `tm_wday` (день недели, от `0` до `6`, причём `0` соответствует воскресенью), `tm_yday` (день в году, от `0` до `365`) и `tm_isdst`, показывающее, действует ли *daylight saving time* (в России это явление было известно как «летнее время»). Функции `gmtime`/`gmtime_r` возвращают результат для Гринвича, `localtime`/`localtime_r` — для часового пояса, установленного в настройках системы.

Кроме того, в библиотеке присутствуют функции, позволяющие создать текстовое представление заданной даты и времени:

```
char *asctime(const struct tm *tm);
char *asctime_r(const struct tm *tm, char *buf);
char *ctime(const time_t *timep);
char *ctime_r(const time_t *timep, char *buf);
```

Как можно заметить, исходными данными для `ctime` и `ctime_r` служит уже знакомое нам число секунд since Epoch в виде переменной типа `time_t`, а для `asctime` и `asctime_r` — структура `tm`, которую мы могли бы получить с помощью рассмотренных выше `gmtime` и `localtime`, а при необходимости — скомпоновать любым другим способом. Результат выполнения функции — строка вроде `"Sat Oct 22 00:57:53 2016\n"`.

Функции с суффиксом `_r` используют для сохранения результата области памяти, предоставленные вызывающим, тогда как функции без этого суффикса возвращают указатель на принадлежащую им структуру данных (обычно описанную внутри в виде статической локальной переменной). Это требует определённого внимания, поскольку следующий вызов такой функции затирает результат, созданный предыдущим вызовом.

Если вам потребуется текстовое представление даты и/или времени в форме, отличной от предлагаемой функциями `gmtime`/`localtime`, можно воспользоваться также функцией `strftime`:

```
int strftime(char *s, int max, const char *format,
             const struct tm *tm);
```

Полное описание этой функции довольно громоздко, так что мы оставим её читателю для самостоятельного изучения.

Время с точностью до секунды нас во многих случаях не устраивает; в программах, особенно имеющих дело с коммуникацией через компьютерную сеть или по другим каналам связи, зачастую требуется отследить, истёк или не истёк интервал времени, составляющий несколько миллисекунд. Узнать текущее время с бóльшей точностью позволяет системный вызов `gettimeofday`, имеющий довольно странный профиль:

```
int gettimeofday(struct timeval *tv, struct timezone *tz);
```

Структура `timeval` состоит из двух полей: `tv_sec` для целого числа секунд (всё тех же «since Epoch») и `tv_usec` для *микросекунд*, т. е. миллионных долей секунды. Чтобы узнать текущее время, мы должны в программе описать переменную типа `struct timeval` и передать её адрес первым аргументом в вызов `gettimeofday`; про второй аргумент в документации прямо сказано, что его использование нежелательно и будет лучше, если там указать `NULL`.

В ОС Unix, как и в других многозадачных операционных системах, предусмотрены возможности *повлиять на ход выполнения процесса*; в частности, достаточно широк ассортимент способов, которыми процесс может повлиять на ход *своего собственного* выполнения. Простейший способ такого влияния — заявить операционной системе, что в течение некоторого периода времени наш процесс ничего делать не собирается и не нуждается в выделении квантов времени; иначе говоря, процесс хочет приостановить своё собственное выполнение. Процесс в таком состоянии называют ***спящим*** (англ. *sleeping*). Самая простая функция, «отправляющая спать» вызывающий процесс, так и называется `sleep`:

```
int sleep(unsigned int seconds);
```

Единственным параметром этой функции указывается число секунд, которые процесс намерен «проспать»; например, `sleep(60)` «усыпит» наш процесс на минуту, то есть в течение минуты процесс не будет ничего делать — операционная система переведёт его в состояние блокировки, а когда указанный интервал времени закончится — разблокирует.

Когда-то давно функция `sleep` представляла собой отдельный системный вызов. Сейчас это не так: система поддерживает несколько системных вызовов, через каждый из которых может быть реализована функция `sleep`, так что в специальном системном вызове нет нужды.

Отдельного комментария заслуживает возвращаемое значение функции `sleep`. Если процесс благополучно «проспал» указанное время, функция вернёт `0`; но, как мы увидим позже, в некоторых случаях (именно — при доставке процессу обрабатываемого ***сигнала***, см. §5.3.14) функция `sleep` может завершиться досрочно. При этом она вернёт положительное целое число, равное количеству секунд (полных или неполных), которые ей оставалось проспать, когда её прервали.

Поскольку секунда — это по компьютерным меркам целая вечность, часто задание времени «сна» в секундах оказывается чрезмерно грубым. Функция `usleep` принимает в качестве параметра число микросекунд:

```
int usleep(long usec);
```

К сожалению, у этой функции имеется довольно серьёзный недостаток: на большинстве unix-систем она отказывается принимать значения параметра от миллиона и выше, то есть можно попросить «усыпить» наш процесс только на интервал времени, мéньший одной секунды. Конечно, никто не мешает скомбинировать `sleep` и `usleep` — например, если нам нужно заснуть на 3,5 с, мы можем сделать

```
sleep(3);
usleep(500000);
```

Но так лучше не делать. Обе функции не гарантируют точности измерения временнóго промежутка хотя бы в силу того, что на передачу управления в ядро с целью блокировки процесса, а затем на возврат управления процессу тоже тратится время, а поскольку в системе могут работать и другие процессы, после выхода из блокировки процесс какое-то время может простоять в очереди, ожидая выделения очередного кванта времени. Естественно, использование двух вызовов функций вместо одного ещё сильнее ухудшит точность.

Нельзя не отметить и ещё один момент: функция `usleep` может вернуть либо `0`, если она благополучно «проспала» сколько её просили, либо `-1`, если «проспать» нужное количество времени ей не дали. В этом случае узнать, сколько времени в действительности прошло, можно только одним способом: перед «отходом ко сну» спросить у системы текущее время, а после «пробуждения» снова узнать текущее время и вычислить их разность.

Наиболее универсальной функцией для «сна» можно считать `nanosleep`, причём это обычно системный вызов; в частности, современные версии Linux именно через `nanosleep` реализуют библиотечные функции `sleep` и `usleep`. Профиль `nanosleep` таков:

```
int nanosleep(const struct timespec *req, struct timespec *rem);
```

Структура `timespec` состоит из двух полей: `tv_sec` для секунд и `tv_nsec` для *наносекунд*, т. е. миллиардных долей секунды; значение второго поля должно быть заключено между 0 и $10^9 - 1$. Через первый параметр мы указываем нужный нам отрезок времени; вторым параметром мы подаём адрес структуры того же типа, в который функция `nanosleep` запишет, сколько времени осталось до изначально заданного момента пробуждения, если ей не дали «проспать» сколько было указано.

Использование наносекунд может создать впечатление, будто время в системе измеряется с точностью до миллиардных долей секунды. На самом деле это не так, интерфейс системного вызова `nanosleep` спроектирован с большим запасом. Время в системе измеряется хорошо если с точностью до тысячной доли секунды; ядро в качестве основного источника данных о времени использует прерывания таймера, а их в современных системах как раз происходит 1000 в секунду.

Ещё один важный способ влияния на выполнение процесса во времени — это изменение его ***приоритета***. В §5.3.2 мы уже упоминали две составляющие приоритета — статическую и динамическую; повлиять мы можем лишь на статическую составляющую, поскольку динамическую планировщик вычисляет сам, исходя из фактического времени выполнения и ожидания для каждого процесса. Ядро Linux использует в качестве приоритета число от `-20` (наивысший приоритет) до `19` (самый низкий приоритет). В некоторых других unix-системах этот интервал составляет `-20..20`; само число 20 здесь представляет собой некую исторически сложившуюся реальность и никакими серьёзными причинами не обусловлено. По умолчанию, т. е. до тех пор, пока ядро не получит на этот счёт явных указаний, все процессы в системе выполняются с приоритетом `0`.

Процесс может *понизить* свой собственный приоритет с помощью системного вызова `nice`:

```
int nice(int inc);
```

Параметром передаётся небольшое целое число, которое *прибавляется* к текущему значению статического приоритета, тем самым *понижая* приоритет процесса. Опустить свой приоритет ниже минимального, конечно, не получится, так что если указать параметром `nice`, например, число `1000`, ваш процесс получит минимальный возможный приоритет — `19` (в некоторых системах — `20`). Вызов `nice` возвращает новое значение приоритета, так что, например, процесс может узнать свой текущий приоритет, выполнив `nice(0)`.

Процессы, имеющие полномочия системного администратора, могут указывать *отрицательное* значение параметра при вызове `nice`, тем самым увеличивая свой приоритет в системе. Для обычных процессов это не проходит, вызов заканчивается ошибкой.

Интересно, кстати, что в случае ошибки `nice` возвращает, как и другие системные вызовы, значение `-1`, но то же самое значение вызов может вернуть, если новый приоритет стал равен `-1` — например, если процесс, выполнявшийся в системе с наивысшим приоритетом `-20`, обратился к ядру с вызовом `nice(19)`. В документации говорится, что единственный надёжный способ отличить эту ситуацию от ошибочной — заранее присвоить значение `0` переменной `errno`, а после вызова проверить, не поменялось ли оно. Впрочем, единственная ошибка, возвращаемая вызовом `nice` — это `EPERM`, и происходит она лишь в одном случае: если указано отрицательное значение параметра, а процесс не имеет полномочий суперпользователя. Очевидно, что этой ситуации легко избежать.

Значение приоритета наследуется при создании нового процесса вызовом `fork` и не меняется при вызове `execve`. Это позволяет, в частности, запустить внешнюю программу как с пониженным, так и (при наличии полномочий) с повышенным приоритетом. Более того, теоретически возможно запустить с заданным приоритетом целый сеанс работы, если это сделает программа, запрашивающая пароль пользователя при входе в систему, ведь все процессы сеанса являются её потомками. Сама эта программа обычно обладает полномочиями суперпользователя, так что она может установить как пониженный приоритет, так и повышенный.

Впрочем, для управления *чьим-то* (не своим) приоритетом система предлагает более удобный механизм — вызовы `getpriority` и `setpriority`, которые позволяют манипулировать приоритетом конкретного процесса, группы процессов[25], либо всех процессов, принадлежащих заданному пользователю:

```
int getpriority(int which, int who);
int setpriority(int which, int who, int prio);
```

В обоих вызовах параметр `which` задаёт способ, которым идентифицируются нужные процессы: `PRIO_PROCESS` означает, что будет задан номер (`pid`) конкретного процесса, `PRIO_USER` предполагает, что речь пойдёт обо всех процессах, принадлежащих заданному пользователю, а `PRIO_PGRP` позволяет задать группу процессов. Значение `who` зависит от того, какой из вариантов мы избрали в первом параметре, и равняется соответственно `pid`'у нужного процесса, `uid`'у пользователя или идентификатору (`pgid`'у) группы процессов. Вызов `getpriority` возвращает текущее значение приоритета заданных процессов; если среди процессов, идентифицируемых параметрами вызова, есть процессы с различными приоритетами, возвращается значение самого высокого из этих приоритетов (то есть наименьшее численное значение приоритета). Вызов `setpriority` устанавливает для всех заданных процессов значение приоритета, заданное третьим параметром. Как уже говорилось, численное значение приоритета может составлять от `-20` до `19` (в некоторых системах — до `20`); понижать это значение может только системный администратор; кроме того, обычный пользователь, разумеется, может изменять приоритет только для своих процессов.

Как и для вызова `nice`, для `getpriority` возвращаемое значение `-1` может означать и приоритет, равный `-1`, и ошибку. Как с этим бороться, вы уже знаете.

5.3.10. Перенаправление потоков ввода-вывода

Новые файловые дескрипторы создаются при успешном выполнении вызова `open`, а также во многих других случаях; как мы увидим позже, дескрипторы используются также для каналов, сокетов и т. п., и вообще часто бывают связаны с объектами ядра, не имеющими отношения к файловой системе. **При создании нового файлового дескриптора система всегда выбирает наименьший свободный номер**; так, если закрыть нулевой дескриптор, следующий успешный вызов `open` вернёт ноль. Для закрытия дескриптора используется уже рассматривавшийся вызов `close`, и это единственный способ избавиться от дескриптора вне зависимости от его природы.

С точки зрения ядра все дескрипторы обрабатываются одинаково, независимо от их номеров, но, как мы знаем, программы, работающие под управлением системы, обычно считают поток с номером `0` стандартным потоком ввода, поток номер `1` — стандартным потоком вывода, а поток номер `2` — стандартным потоком для сообщений об ошибках. Если говорить точнее, **библиотечные функции считают потоки `0`, `1` и `2` стандартными, что́ бы ни было под этими номерами открыто**. Следовательно, если мы тем или иным способом добьёмся открытия под этими номерами какого-нибудь файла, именно этот файл станет для нашего процесса соответствующим стандартным потоком. Например, если последовательно выполнить

```
close(1);
fd = open("file.txt", O_WRONLY|O_CREAT|O_TRUNC, 0666);
```

и файл `file.txt` успешно откроется на запись, то он откроется под номером `1` (именно единица будет занесена в переменную `fd`) и станет стандартным потоком вывода, то есть с этого момента функции вывода, не предполагающие указания потока, такие как `printf`, `putchar` и т. д., будут выводить информацию в файл `file.txt`. Собственно говоря, все эти функции даже не узнают, что у них сменился стандартный поток.

Недостаток такого способа изменения стандартного потока очевиден: если файл по тем или иным причинам не откроется и `open` вернёт значение `-1`, указывающее на происшедшую ошибку, то закрытый нами стандартный дескриптор так и останется закрытым, и после этого очень легко ошибиться и случайно открыть под тем же номером что-то такое, что совершенно не предназначалось для работы в роли стандартного потока. Поэтому обычно для манипуляций с таблицей файловых дескрипторов используют специально предназначенные для этого системные вызовы `dup` и `dup2`:

```
int dup(int fd);
int dup2(int fd, int new_fd);
```

Вызов `dup` создаёт новый файловый дескриптор, связанный *с тем же самым* потоком ввода-вывода, что и `fd`. Следует понимать, что нового потока ввода-вывода при этом не создаётся, оба дескриптора оказываются связаны с одним и тем же объектом ядра, реализующим поток ввода-вывода. Новый и старый дескрипторы, в частности, используют один и тот же указатель текущей позиции в файле: если на одном из них сменить позицию с помощью `lseek`, позиция на втором из них тоже изменится. С подобной ситуацией мы уже знакомы: при создании нового процесса вызовом `fork` все файловые дескрипторы родительского процесса копируются, но потоки ввода-вывода остаются прежними, так что порождённый процесс своими дескрипторами ссылается на те же самые потоки.

Вызов `dup2` отличается тем, что новый дескриптор создаётся под заданным номером (параметр `new_fd`). Если на момент выполнения вызова у процесса был поток ввода-вывода, связанный с дескриптором `new_fd`, то есть этот дескриптор был открыт, он закрывается.

Рассмотрим для примера ситуацию, когда некоторая библиотека, которую мы используем, производит вывод нужной нам информации всегда в стандартный поток вывода, а нам желательно соответствующую информацию вывести в файл. Это можно сделать с помощью такого фрагмента кода:

```
int save1, fd;
fd = open("file.dat", O_CREAT|O_WRONLY|O_TRUNC, 0666);
                    /* открыли файл */
if(fd == -1) {      /*   ... обработка ошибки ...   */
                    /* в этом случае дальше не идём */
}
fflush(stdout);     /* на всякий случай очищаем буфер
                       стандартного вывода */
save1 = dup(1);     /* сохраняем наш стандартный вывод */
dup2(fd, 1);        /* сделали открытый файл
                       стандартным потоком вывода */
close(fd);          /* закрыли "лишний" дескриптор */

/* ... производим действия с нашей библиотекой ...
   всё это время вызовы функций, работающих со стандартным
   выводом (таких как printf, puts и т.п.), будут выводить
   информацию в наш файл
```

[25] Группы процессов мы уже упоминали в §5.3.2, а подробному рассмотрению этого явления будет посвящён отдельный параграф в главе о работе с терминалом.

```
*/

dup2(save1, 1);     /* восстановили старый стандартный
                       поток вывода */
                    /* файл при этом закрылся автоматически */
close(save1);       /* лишняя копия нам не нужна */
```

Обычно (хотя и не всегда) стандартные потоки ввода-вывода перенаправляют при запуске внешней программы; например, перенаправления выполняет командный интерпретатор, когда пользователь даёт команду, содержащую соответствующие символы (<, > или >>). При этом используется свойство вызова `execve` сохранять таблицу файловых дескрипторов неизменной. Чаще всего перенаправления производятся с помощью `dup2` в порождённом процессе непосредственно перед вызовом одной из функций семейства `exec`, после чего ненужный уже исходный дескриптор закрывается.

Рассмотрим пример. Допустим, у нас возникла потребность в программе на Си смоделировать функционирование команды

```
ls -l -a -R / > flist
```

— т. е., используя возможности программы `ls`, сгенерировать файл `flist`, содержащий список всех файлов в системе с расширенной информацией по каждому из них. Это можно сделать с помощью следующего фрагмента:

```
int pid, status;
pid = fork();
if(pid == -1) { /* ... обработка ошибки ... */ }
if(pid == 0) {       /* порождённый процесс */
    int fd = open("flist", O_CREAT|O_WRONLY|O_TRUNC, 0666);
    if(fd == -1) {
        perror("flist");
        exit(1);
    }
    dup2(fd, 1);
    close(fd);
    execlp("ls", "ls", "-l", "-a", "-R", "/", NULL);
    perror("ls");
    exit(1);
}
/* родительский процесс */
wait(&status);
if(!WIFEXITED(status) || WEXITSTATUS(status)!=0) {
    /* ... обработка ошибки ... */
}
```

В этом примере вызов `open` для открытия файла мы тоже выполнили в порождённом процессе, но часто файлы открывают в родительском процессе — до выполнения `fork`; это позволяет в родительском же процессе обработать возможные ошибки, связанные с открытием файла, и, например, вообще не запускать новый процесс, если файл не открылся. Нужно только не забыть закрыть дескриптор также и в родительском процессе:

```
int pid, status, fd;
fd = open("flist", O_CREAT|O_WRONLY|O_TRUNC, 0666);
if(fd == -1) {
    perror("flist");
    exit(1);
}
pid = fork();
if(pid == -1) { /* ... обработка ошибки ... */ }
if(pid == 0) {       /* порождённый процесс */
    dup2(fd, 1);
    close(fd);
    execlp("ls", "ls", "-l", "-a", "-R", "/", NULL);
    perror("ls");
    exit(1);
}
/* родительский процесс */
close(fd);           /* про это важно не забыть */
wait(&status);
```

Сдублировать дескриптор можно также с помощью знакомого нам `fcntl` (см. стр. 106), используя команду `F_DUPFD`; в этом случае вызов принимает третий параметр (почему-то типа `long`), задающий *наименьший желаемый номер дублирующего дескриптора*. Вызов отыскивает первый свободный номер дескриптора, не меньший, чем значение этого параметра, и использует его в качестве номера для нового дескриптора. Так, следующие два вызова эквивалентны:

```
nfd = dup(fd);
nfd = fcntl(fd, F_DUPFD, 0);
```

Вызов `fcntl(fd, F_DUPFD, 20)` создаст дескриптор, дублирующий `fd`, под номером 20, если этот номер свободен, в противном случае найдёт первый свободный номер дескриптора, бóльший, чем 20.

5.3.11. Полномочия процесса

Ранее (см., напр., §§5.1.8, 5.3.2) мы многократно подчёркивали, что именно процесс — это тот, кто обладает в системе теми или иными полномочиями, и упомянули такие параметры процесса, как `uid`, `gid`, `euid` и `egid`. Как водится, в реальности всё несколько сложнее. Процессу в системе приписываются:

- «настоящие» (англ. *real*) `uid` и `gid`, как правило, соответствующие идентификаторам пользователя, запустившего процесс;
- «эффективные» или «действующие» (англ. *effective*) `euid` и `egid`, которые как раз используются в большинстве случаев для проверки полномочий процесса;
- «сохранённые» идентификаторы (*saved set-user-ID, saved set-group-ID*);
- массив *дополнительных* идентификаторов групп (*supplementary group IDs*).

Во многих системах этим дело не ограничивается; так, в ОС Linux процесс имеет ещё отдельные «файловые» идентификаторы `fsuid` и `fsgid`, в большинстве случаев равные «эффективным», но допускающие отдельную установку; кроме того, ядро Linux поддерживает механизм *capabilities*, позволяющий разрешать или запрещать процессам определённые привилегированные действия независимо от наличия или отсутствия у них суперпользовательских полномочий. Всё это мы оставим за рамками нашей книги; при желании читатель может обратиться к технической документации.

При старте системы ядро запускает процесс с номером `1` (обычно это программа `/sbin/init`); все шесть основных идентификаторов для этого процесса — «настоящие», «эффективные» и «сохранённые» — равны нулю, массив дополнительных групп пуст. Все остальные процессы в системе являются потомками процесса № `1`, хотя, возможно, очень дальними; иное невозможно, поскольку ядро, запустив первый процесс, больше по своей инициативе никаких процессов не создаёт. При порождении нового процесса вызовом `fork` (см. §5.3.4) все идентификаторы наследуются новым процессом без изменений, так что непосредственные потомки `init`'а обычно также имеют `0` в качестве значения всех шести идентификаторов и пустой список дополнительных групп.

Процесс, обладающий полномочиями суперпользователя (т. е. `euid` которого равен нулю), имеет право с помощью системных вызовов сделать со своими полномочиями что угодно, в том числе, что важно, полностью или частично отказаться от своих привилегий. Именно так поступают программы, создающие в системе сеанс работы пользователя: прежде чем запустить командный интерпретатор или оконный менеджер, такие программы устанавливают себе полномочия в соответствии с настройками для данного пользователя. Вернуть себе полномочия суперпользователя после этого уже нельзя, и это правильно: пользователь, работающий в системе, должен обладать полномочиями, которые ему положены в соответствии с конфигурацией системы, и не больше. Чаще всего при таком «сбросе полномочий» все три пары идентификаторов изменяются синхронно, то есть одно и то же число, соответствующее `uid`'у выбранного пользователя, заносится и в «настоящий» `uid`, и в `euid`, и в *saved set-user-ID*, а идентификатор основной группы данного пользователя заносится синхронно в «настоящий» `gid`, в `egid` и в *saved set-group-ID*. Перед этим (не после, после будет уже поздно) при необходимости формируется список дополнительных групп.

Наиболее очевидный случай, когда `uid` и `gid` могут отличаться от `euid` и `egid` — это запуск с помощью вызова `execve` (или любой библиотечной функции из семейства `exec`, ведь все они в конечном итоге вызывают именно `execve`) программы, имеющей установленные флаги `SetUid` и/или `SetGid` (см. стр. 102). В этом случае «настоящие» идентификаторы остаются те же,

что были у запускающего, т. е. у процесса, сделавшего `exec`, на момент обращения к `exec`; в то же время, если у исполняемого файла (запускаемой программы) был установлен бит `SetUid`, то в `euid` и в *saved set-user-ID* записывается идентификатор владельца исполняемого файла; аналогично при установленном бите `SetGid` идентификатор группы исполняемого файла заносится в `egid` и в *saved set-group-ID*. Массив дополнительных групп не изменяется.

Идея всех этих танцев с тремя парами идентификаторов состоит в том, что процесс может (т. е. имеет право с точки зрения ядра) менять свой `euid` туда-сюда между «настоящим» и «сохранённым» значениями, то есть (для случая выполняющейся SUID'ной программы) при необходимости часть работы выполнять с правами владельца программы, а часть — с правами пользователя, запустившего программу. Надо сказать, что это ограничение, естественно, перестаёт действовать, если процесс выполняется с правами суперпользователя, т. е. когда его `euid` равен нулю: в этом случае он имеет право делать со своими идентификаторами всё что угодно.

Пусть, например, пользователь Вася с входным именем `vasya` и `uid`'ом `1003` создал в системе исполняемый файл `vasyaprog` и поставил на него атрибут `SetUid`, а пользователь Петя с входным именем `petya` и `uid`'ом `1012` этот файл запустил. В этом случае программа `vasyaprog` сразу после запуска обнаружит, что её «настоящий» `uid` равен `1012` (это отражает тот факт, что запустил её Петя), тогда как «эффективный» и «сохранённый» `uid`'ы оба равны `1003` — это результат действия атрибута `SetUid`. Если программа не будет предпринимать действий по изменению своего `uid`'а, то она будет работать в системе от имени Васи и с его полномочиями. В то же время система позволит процессу, в котором работает `vasyaprog`, изменить свой «эффективный» `uid`, установив в качестве такового `uid` Пети (`1012`): если это проделать, «настоящий» и «эффективный» идентификаторы будут равны `1012` и программа будет продолжать работу уже от имени Пети и с полномочиями Пети, а не Васи. «Сохранённый» идентификатор останется равен `1003`, так что программа сможет снова поменять свой «эффективный» `uid` обратно и продолжить работу от имени Васи, и так далее.

Процесс может узнать свои «настоящие» и «эффективные» идентификаторы с помощью системных вызовов, имеющих очевидные названия:

```
int getuid();
int geteuid();
int getgid();
int getegid();
```

Эти вызовы всегда работают успешно, возвращая численное значение запрошенного идентификатора.

На самом деле тип возвращаемого значения этих системных вызовов — не `int`, а `uid_t` и `gid_t`; в современных версиях того же Linux эти типы определены как `unsigned int`. Использование `int` может создать вам проблемы, если в системе найдётся пользователь или группа, идентификаторы которых превышают $2^{31} - 1$, т. е. 2147483647. Если вы найдёте такую систему, настоятельно рекомендуем вам держаться подальше от её администраторов — они опасны для окружающих.

Стандартного способа узнать «сохранённые» идентификаторы не предусмотрено, но во многих системах, включая Linux, существуют нестандартные вызовы `getresuid` и `getresgid` (буквы `res` обозначают *real, effective, saved*), которые позволяют запросить у ядра сразу все три идентификатора пользователя или группы. Впрочем, если вам это зачем-то понадобилось, скорее всего вы делаете что-то не то.

Узнать свои идентификаторы дополнительных групп процесс может с помощью системного вызова `getgroups`:

```
int getgroups(int size, gid_t *arr);
```

К сожалению, поскольку здесь задействован указатель (и массив), игнорировать разницу между `int` и `gid_t` уже не получится. Вторым параметром в вызов нужно передать указатель на начало массива (из элементов типа `gid_t`), первым — размер этого массива, чтобы вызов «знал», сколько идентификаторов можно записать в указанный вами массив. Первым параметром можно передать ноль, в этом случае второй параметр игнорируется, а вызов возвращает количество дополнительных групп. Если места в массиве не хватает, вызов завершается ошибкой, но этого можно избежать, заранее узнав, какой длины должен быть массив. Вы можете применить примерно такой код:

```
int n;
gid_t *p;
n = getgroups(0, NULL);
p = malloc(n * sizeof(gid_t));
getgroups(n, p);
```

После этого в переменной `n` будет находиться общее количество дополнительных групп, а в массиве по адресу `p` — сами идентификаторы этих групп. Спецификация вызова `getgroups` не указывает, будет ли в число групп включена основная группа, т. е. значение вашего `egid`; это значит, что исчерпывающую информацию о группах пользователей, права которых даны вашему процессу, вы можете узнать только с помощью комбинации вызовов `getgroups` и `getegid`.

Для манипуляции полномочиями процесса предусмотрены следующие системные вызовы:

```
int setuid(int uid);
int setgid(int gid);
int seteuid(int euid);
int setegid(int egid);
int setgroups(int size, const gid_t *list);
```

С первыми четырьмя из них часто возникает изрядная путаница. Проще всего дело обстоит с вызовами `seteuid` и `setegid`: они в полном соответствии со своими названиями предназначены для изменения «эффективных» идентификаторов пользователя и группы вызвавшего процесса (параметров `euid` и `egid`). Если текущий `euid` процесса отличен от нуля, то процесс может потребовать от системы изменить свой «эффективный» идентификатор (пользователя или группы) на соответствующий «настоящий», «эффективный» или «сохранённый» идентификатор. Иначе говоря, непривилегированный процесс имеет право в качестве параметра вызова `seteuid` указывать любое из трёх значений своего пользовательского идентификатора, в противном случае возникнет ошибка; указывать текущее значение `euid`'а может показаться бессмысленным, но система это ошибкой не считает, просто `euid` останется тем же, что и был. То же касается вызова `setegid` и идентификатора группы: здесь непривилегированный процесс может в качестве параметра указать свой «настоящий», «эффективный» или «сохранённый» идентификатор группы, иное приведёт к ошибке. На процессы с `euid == 0` ограничение, как обычно, не распространяется.

Теперь, собственно, источник путаницы: **вызовы `setuid` и `setgid`, вопреки своим названиям, тоже меняют прежде всего именно «эффективные» идентификаторы пользователя и группы**, а вовсе не «настоящие», как можно было бы ожидать, и меняют их по совершенно тем же правилам, что и `seteuid`/`setegid`: непривилегированный процесс имеет право указать в качестве аргумента свой «настоящий», «эффективный» или «сохранённый» идентификатор пользователя или группы. Единственное отличие этих двух вызовов от двух предыдущих проявляется, когда процесс обладает полномочиями суперпользователя, то есть его `euid` равен нулю. В этом случае вызовы `setuid` и `setgid` изменяют синхронно все три значения соответствующего идентификатора — и «эффективное», и «настоящее», и «сохранённое», тогда как вызовы `seteuid` и `setegid` изменяют только значения «эффективных» идентификаторов.

С вызовом `setgroups` всё, к счастью, довольно просто: второй параметр должен указывать на массив идентификаторов дополнительных групп, первый параметр задаёт длину массива. Вызов, что вполне естественно, доступен только процессам, работающим с полномочиями суперпользователя. Старый список дополнительных групп, каков бы он ни был, бесследно исчезает, вместо него процессу приписывается новый. Следует отметить, что на значения «эффективного», «настоящего» и «сохранённого» идентификатора группы этот вызов никакого влияния не оказывает.

В некоторых случаях более удобными могут оказаться вызовы `setreuid`, `setregid`, `setresuid`, `setresgid`, `getresuid` и `getresgid`, которые мы оставим читателю для самостоятельного изучения.

5.3.12. Количественные ограничения

Помимо ограничений вида «можно-нельзя», которые мы рассматривали в предыдущем параграфе, ядро операционной системы может накладывать на процессы ограничения *количественные*, при которых процесс вправе использовать те или иные ресурсы системы, но только до определённого предела. В §5.1.8 мы приводили примеры таких ограничений: предельное количество виртуальной памяти для отдельно взятого процесса, количество одновременно открытых (одним процессом) файлов, количество самих процессов, запущенных от имени данного пользователя и т. д.

Чтобы получить представление о действующих лимитах, можно воспользоваться командой `ulimit` в командной строке. В частности, `ulimit -a` выдаст список всех возможных в данной системе количественных ограничений и их текущие значения:

```
avst@host:~$ ulimit -a
core file size          (blocks, -c) unlimited
data seg size           (kbytes, -d) unlimited
scheduling priority             (-e) 20
file size               (blocks, -f) unlimited
pending signals                 (-i) 16382
max locked memory       (kbytes, -l) 64
max memory size         (kbytes, -m) unlimited
open files                      (-n) 1024
pipe size            (512 bytes, -p) 8
POSIX message queues     (bytes, -q) 819200
real-time priority              (-r) 0
stack size              (kbytes, -s) 8192
cpu time               (seconds, -t) unlimited
max user processes              (-u) unlimited
virtual memory          (kbytes, -v) unlimited
file locks                      (-x) unlimited
avst@host:~$
```

Все эти ограничения представляют собой *свойства отдельно взятого процесса* — как обычно в таких случаях, наследуемые при создании новых процессов и сохраняющиеся при замене выполняемой программы с помощью `exec`. Иначе говоря, процесс может установить то или иное ограничение, но действовать оно будет только на сам этот процесс и на его потомков, а всю систему в целом не затронет. Важно понимать, что этот принцип распространяется в том числе и на такое (вроде бы) «глобальное» ограничение, как количество процессов, разрешённых для отдельно взятого пользователя (`uid`'а): для процесса, установившего ограничение, и для его потомков ядро при обработке вызова `fork` будет проверять, не превышен ли лимит процессов, учитывая при этом *все* процессы, принадлежащие данному пользователю, и если лимит превышен, `fork` вернёт `-1`, а в `errno` будет занесено значение `EAGAIN`. В то же самое время другие процессы, в том числе принадлежащие тому же пользователю, но не являющиеся потомками процесса, установившего ограничение, будут по-прежнему беспрепятственно порождать новые процессы.

Для каждого вида количественного ограничения система позволяет установить два предельных уровня — мягкий и жёсткий (*soft/hard*). Ядро в своих проверках использует мягкий уровень, но любой процесс может изменить этот уровень (как уменьшить, так и увеличить) для любого из своих ограничений в пределах от нуля до установленного жёсткого уровня. Что касается жёсткого уровня, то любой процесс может его уменьшить, но увеличивать его разрешается только процессам, работающим с правами суперпользователя; для всех прочих процессов уменьшение жёсткого уровня любого из лимитов необратимо.

Для установки количественных ограничений используется системный вызов `setrlimit`:

```
int setrlimit(int resource, const struct rlimit *rlim);
```

Структура `rlimit` имеет два поля: `rlim_cur` (мягкий уровень) и `rlim_max` (жёсткий уровень); оба поля имеют некий целочисленный тип. Имена полей образованы от слов *current* и *maximum*, т. е. имеется в виду «текущее» и «максимальное» значение ограничения. Для обозначения ситуации отсутствия ограничения используется константа `RLIM_INFINITY`; обычно это число `-1`, преобразованное к типу, используемому для полей `rlim_cur` и `rlim_max`, а поскольку эти поля в большинстве реализаций беззнаковые, `RLIM_INFINITY` оказывается равно наибольшему возможному беззнаковому целому соответствующей разрядности.

Первый параметр вызова (`resource`) указывает, какое именно ограничение следует установить (изменить). Перечислим некоторые возможные значения этого параметра:

- `RLIMIT_AS` — максимальное количество виртуальной памяти для отдельно взятого процесса (в байтах);
- `RLIMIT_CORE` — максимальный возможный размер core-файла[26], создаваемого при авариях (в байтах); если установить нулевой лимит, core-файлы создаваться не будут;
- `RLIMIT_CPU` — максимальное количество процессорного времени, которое может использовать процесс (в секундах); забегая вперёд, отметим, что при достижении мягкого лимита система отправит процессу ***сигнал*** `SIGXCPU`, который по умолчанию убивает процесс, но если процесс перехватит или проигнорирует этот сигнал, то система продолжит посылать процессу сигнал `SIGXCPU` каждую секунду, пока не будет достигнут жёсткий лимит, после чего уничтожит процесс сигналом `SIGKILL`, который нельзя ни перехватить, ни проигнорировать; подробное рассмотрение сигналов ждёт нас в §5.3.14;
- `RLIMIT_FSIZE` — максимальный размер создаваемых дисковых файлов (в блоках по 512 байт); если при записи в какой-либо файл этот размер окажется превышен, процесс получит сигнал `SIGXFSZ`, если же этот сигнал не убьёт процесс, то соответствующий системный вызов (например, `write`) вернёт `-1`, а `errno` получит значение `EFBIG`;
- `RLIMIT_NOFILE` — предельное количество одновременно открытых потоков ввода-вывода для одного процесса; если говорить точнее, процесс не может создать (в том числе с помощью `dup2`, см. стр. 121) файловый дескриптор с номером, равным или превосходящим установленный лимит, даже если открытых потоков на этот момент гораздо меньше;
- `RLIMIT_NPROC` — предельное количество процессов, разрешённых для создания от имени конкретного пользователя (т. е. с тем же `uid`'ом); как уже отмечалось, вызов `fork` учитывает при этом все процессы с данным `uid`'ом, имеющиеся в системе, но проверка производится только когда `fork` сделал процесс, для которого установлен соответствующий лимит;
- `RLIMIT_STACK` — размер виртуального адресного пространства, выделяемого под сегмент стека (см. §5.3.3).

Существуют и другие лимиты; в наш список мы включили только те, которые проще всего объяснить. Полный список вы найдёте в технической документации, в том числе в тексте справочной страницы по вызову `setrlimit` (команда `man 2 setrlimit`).

Узнать текущие значения ограничений позволяет вызов `getrlimit`:

```
int getrlimit(int resource, struct rlimit *rlim);
```

Помимо количественных ограничений, налагаемых вызовом `setrlimit`, большинство unix-систем поддерживает также ограничения на использование дискового пространства (так называемое *квотирование*); такие ограничения устанавливаются отдельно по каждому диску для заданных пользователей и/или пользовательских групп. В Linux и FreeBSD интерфейсом для этого механизма служит системный вызов `quotactl`. Оставляем его заинтересованным читателям для самостоятельного изучения.

5.3.13. Обзор средств взаимодействия процессов

В рамках одной unix-системы процессы могут так или иначе взаимодействовать между собой. Вообще говоря, один процесс может повлиять на работу другого, не прибегая к специализированным средствам; например, процесс может модифицировать файл, читаемый другим процессом, и для этого не потребуется ничего, кроме уже знакомых нам вызовов, позволяющих работать с файлами. Кроме того, системный вызов `mmap`, как уже говорилось, позволяет создать область памяти, доступную нескольким процессам. Такая область памяти называется ***разделяемой***, а процессы, работающие с ней, считаются *взаимодействующими через разделяемую память*, поскольку действия одного из них очевидным образом влияют на работу других.

[26] Core-файл — это файл с именем `core` или `prog.core` (где `prog` — имя программы), создаваемый операционной системой в текущем каталоге при аварийном завершении процесса. В этот файл полностью записывается содержимое сегментов данных и стека на момент аварии. Core-файлы позволяют с помощью отладчика проанализировать причины аварии, в том числе узнать точку кода, в которой произошла авария, посмотреть значения переменных на момент аварии и т. д.; мы рассмотрим работу с ними в приложении, посвящённом отладчику.

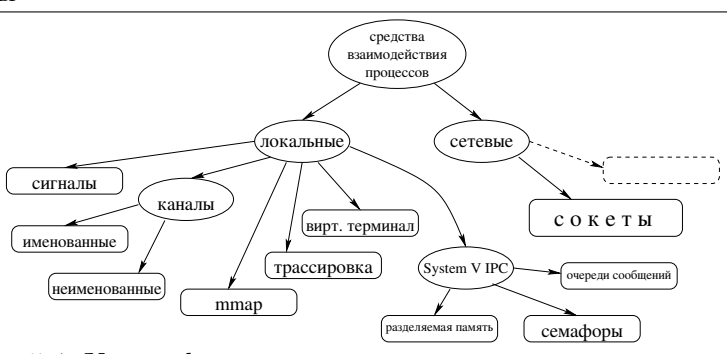

Рис. 5.4. Классификация средств взаимодействия процессов

Конечно, удобнее для взаимодействия процессов использовать возможности системы, которые для этого специально предназначены. Наиболее примитивным из них можно считать уже встречавшиеся нам ***сигналы***. Сигнал не несёт в себе никакой информации, кроме ***номера сигнала*** — целого числа из предопределённого множества.

Для передачи данных между процессами можно использовать однонаправленные ***каналы***, различающиеся на ***именованные*** (FIFO) и ***неименованные*** (pipe). С каналом связываются два файловых дескриптора, один из которых открыт на запись, другой — на чтение, так что данные, записанные в канал одним процессом, могут быть прочитаны другим.

При отладке программ используется режим ***трассировки***, когда один процесс (обычно отладчик) контролирует выполнение другого (отлаживаемой программы).

Как уже говорилось ранее, важную роль в системах семейства Unix играет понятие ***терминала***. При необходимости функциональность терминала как устройства может имитировать пользовательский процесс: так работает, например, программа `xterm`, а также серверы, отвечающие за удалённый доступ к машине, такие как `sshd` или `telnetd`. Взаимодействие процесса, эмулирующего терминал, с процессами, для которых эмулируемый (то есть программно реализованный) терминал является управляющим, называется взаимодействием через ***виртуальный терминал***.

Несколько особое место в классификации занимают средства, объединённые общим названием ***System V IPC***[27]. Эти средства включают механизмы создания разделяемой памяти, массивов семафоров и очередей сообщений. В практическом программировании System V IPC используется редко, если вообще используется. Эрик Реймонд в книге [6] называет эти средства устаревшими, но это не совсем верно, правильнее было бы назвать их мертворождёнными.

Основным средством *взаимодействия через компьютерную сеть* (то есть взаимодействия процессов, находящихся в разных системах) можно считать ***сокеты*** (*sockets*). Сокеты представляют собой универсальный интерфейс, пригодный для работы с широким спектром протоколов; это означает, что область применения сокетов не ограничена сетями на основе TCP/IP или какого-либо другого стандарта; более того, при добавлении в систему поддержки новых протоколов нет нужды расширять набор системных вызовов. Системы семейства Unix поддерживают также специальный вид сокетов, который можно использовать внутри одной системы, даже если поддержка компьютерных сетей в системе отсутствует.

5.3.14. Сигналы

Здесь и далее мы полностью проигнорируем появившиеся относительно недавно так называемые *сигналы реального времени*, которые, в отличие от классических сигналов, способны нести на себе дополнительную информацию и выстраиваться в очереди. Обычно в таких случаях мы отсылаем читателя к технической документации, но в этот раз дадим дополнительную рекомендацию: вообще не использовать (и не изучать) этот механизм; если сигналы реального времени вам зачем-то понадобились, то вы, скорее всего, делаете что-то не то.

Один из простейших способов повлиять на работу процесса — это отправить ему ***сигнал*** из некоторого предопределённого множества. Изначально сигналы были предназначены для снятия процессов с выполнения, но с развитием системы Unix приобрели другие функции. Перечислим некоторые наиболее употребительные сигналы.

Сигнал `SIGTERM` предписывает процессу завершиться; процесс может перехватить или игнорировать этот сигнал. `SIGKILL` уничтожает процесс; в отличие от `SIGTERM`, этот сигнал ни перехватить, ни игнорировать нельзя. Не лишним будет запомнить, что сигнал `SIGKILL` имеет номер 9, а `SIGTERM` — 15; некоторые «продвинутые» пользователи Unix могут ожидать от вас такого знания, используя фразы вроде «сначала попробуй пятнадцатым, если через секунду не сдохнет — бей девятым». К счастью, номера всех остальных сигналов не помнит большинство даже самых «продвинутых» любителей Unix.

Два «уничтожающих» сигнала — перехватываемый и неперехватываемый — введены, чтобы можно было снять процесс более гибко. Так, при перезагрузке системы обычно всем процессам рассылается сначала `SIGTERM`, а затем через пять секунд — `SIGKILL`. Это позволяет процессам «привести дела в порядок»: например, редактор текстов может сохранить редактируемый текст во временном файле, чтобы потом (в начале следующего сеанса редактирования) предложить пользователю восстановить пропавшие изменения.

Сигналы `SIGILL`, `SIGSEGV`, `SIGFPE` и `SIGBUS` система отправляет процессам, чьи действия привели к возникновению исключения (внутреннего прерывания): попытка выполнить несуществующую или недопустимую команду процессора, нарушение защиты памяти, деление на ноль и обращение к памяти по некорректному адресу соответственно. По умолчанию любой из этих сигналов уничтожает процесс с созданием core-файла[28] для последующего анализа причин происшествия. Однако любой из этих сигналов можно перехватить, например, чтобы попытаться перед завершением записать в файл результаты работы; впрочем, в большинстве случаев из подобной отчаянной попытки всё равно ничего не выходит — например, если наш процесс перехватил сигнал `SIGSEGV`, то структуры данных в его памяти с хорошей вероятностью уже безнадёжно разрушены, и обращение к ним с целью что-то спасти, увы, приведёт лишь к новому сигналу с тем же номером.

Сигналы `SIGSTOP` и `SIGCONT` позволяют соответственно приостановить и продолжить выполнение процесса. Отметим, что `SIGSTOP`, как и `SIGKILL`, нельзя ни перехватить, ни игнорировать. `SIGCONT` перехватить можно, но свою основную роль — продолжить выполнение процесса — он в любом случае сыграет.

Сигналы `SIGINT` и `SIGQUIT` отправляются текущей группе процессов данного терминала[29] при нажатии на клавиатуре комбинаций `Ctrl-C` и `Ctrl-\` соответственно. По умолчанию оба сигнала приводят к завершению процесса, причём `SIGQUIT` ещё и создаёт core-файл.

Сигнал `SIGCHLD` система присылает родительскому процессу при завершении его непосредственного потомка; по умолчанию с родительским процессом при получении этого сигнала ничего не происходит.

Сигнал `SIGALRM` присылается по истечении заданного интервала времени после вызова `alarm`. Процессы используют этот вызов и этот сигнал в роли своеобразного «будильника», например, на случай чрезмерно долгого выполнения тех или иных действий. Обычно `SIGALRM` отправляет операционная система. По умолчанию его получение приводит к завершению процесса, так что следует заранее позаботиться о настройке реакции на этот сигнал.

Сигналы `SIGUSR1` и `SIGUSR2` предназначены для использования программистом для своих целей. По умолчанию эти сигналы также завершают процесс.

Отметим один очень важный момент: все названия сигналов, такие как `SIGKILL` или `SIGCHLD`, представляют собой целочисленные константы, описанные в библиотечных заголовочных файлах и соответствующие номерам сигналов; например, идентификатор `SIGKILL` вводится примерно так:

```
#define SIGKILL 9
```

На всякий случай подчеркнём, что вам в своих программах ничего подобного писать не надо, соответствующие макроопределения уже есть в системных заголовочных файлах.

Отправителем сигнала может быть как процесс, так и операционная система, получатель — всегда процесс. Для отправки сигнала служит системный вызов `kill`:

```
int kill(int target_pid, int sig_no);
```

---

[27] Символ «V» в данном случае означает римское «пять»; термин читается как «си́стэм файв ай-пи-си».

[28] См. сноску 26 на стр. 124.

[29] Сеансы и группы процессов будут рассмотрены позже. Пока можно считать, что сигналы `SIGINT` и `SIGQUIT` при нажатии соответствующих клавиш получает тот процесс, который вы запустили, набрав команду, а также его потомки (если они не приняли мер против этого).

Параметр `sig_no` задаёт номер сигнала, который следует отправить. Для лучшей ясности программы рекомендуется использовать вместо чисел макроконстанты с префиксом `SIG`, такие как `SIGINT`, `SIGUSR1` и т. п. Параметр `target_pid` задает процесс(ы), которому (которым) следует отправить сигнал. Если в качестве этого параметра передать положительное число, это число будет использоваться как номер процесса, которому следует послать сигнал. Если передать число `-1`, сигнал будет послан всем процессам, кроме самого вызвавшего `kill`, а также процесса № `1` (init). Отрицательное число, бóльшее единицы по модулю, означает передачу сигнала группе процессов с соответствующим номером, а ноль — всем процессам своей группы (в обоих случаях — за исключением самого себя). Напомним, что с таким способом идентификации процессов и их групп мы уже сталкивались при рассмотрении вызовов `wait4` и `waitpid` в §5.3.7. Процессы с администраторскими полномочиями (имеющие нулевой `euid`), могут отправлять сигналы любым процессам; все прочие процессы имеют право отправлять сигналы только процессам, принадлежащим тому же пользователю. Как следствие, для непривилегированного процесса вызов `kill(-1, SIGTERM)` означает отправку сигнала `SIGTERM` всем процессам того же пользователя, кроме самого себя. Вызов `kill` возвращает `-1` в случае ошибки, `0` в случае успеха.

Как мы уже знаем, большинство сигналов *по умолчанию* завершают процесс, причём некоторые из них ещё и создают core-файл; некоторые сигналы (например, `SIGCHLD`) по умолчанию игнорируются. Этим «по умолчанию» возможности не ограничиваются, поскольку процесс может для любого сигнала, кроме `SIGKILL` и `SIGSTOP`, установить свой режим обработки: вызов функции-обработчика, игнорирование или обработка по умолчанию. Говорят, что процесс может установить ***диспозицию сигнала***.

Функция-обработчик должна принимать один целочисленный параметр и иметь тип возвращаемого значения `void`, т. е. это должна быть функция вида

```
void handler(int s)
{
    /* ... */
}
```

Для изменения диспозиции сигнала, в том числе для установки обработчика, можно использовать системный вызов `signal`:

```
typedef void (*sighandler_t)(int);
sighandler_t signal(int signo, sighandler_t hdl);
```

Параметр `signo` задаёт номер сигнала, параметр `hdl` — новую диспозицию для этого сигнала; это может быть адрес функции-обработчика, которая должна быть вызвана при получении соответствующего сигнала; также можно использовать специальные значения диспозиции `SIG_IGN` (игнорировать сигнал) и `SIG_DFL` (вернуть диспозицию по умолчанию). Вызов `signal` возвращает значение, соответствующее предыдущей диспозиции данного сигнала, либо специальное значение `SIG_ERR` в случае ошибки. `SIG_ERR` на самом деле представляет собой число `-1`, преобразованное к типу `sighandler_t`.

Следует подчеркнуть, что **установка диспозиции сигнала влияет на то, что будет, если/когда кто-то пришлёт процессу соответствующий сигнал**; иначе говоря, никаких немедленных последствий от изменения диспозиции не наступает, и если сигнал с данным номером не придёт процессу никогда, то от изменения его диспозиции вообще не будет видимых эффектов. В частности, установка функции-обработчика не приводит к вызову этой функции. Обработчик будет вызван лишь в том случае, если кто-либо отправит нашему процессу сигнал; вполне возможно, что этого так и не случится до завершения процесса.

Когда процесс получает сигнал, для которого в качестве диспозиции назначена функция-обработчик, эта функция вызывается *асинхронно* по отношению к выполнению остального процесса, то есть, попросту говоря, в самый неожиданный момент. При этом параметр функции будет равен номеру сигнала, что позволяет один и тот же обработчик использовать для нескольких разных сигналов.

Дальнейшее поведение процесса после получения первого экземпляра обрабатываемого сигнала зависит от версии операционной системы (а иногда, как в случае Linux, и от версии системных библиотек). В классических версиях Unix режим обработки сигнала при получении такового (непосредственно перед передачей управления функции-обработчику) сбрасывался в режим по умолчанию. Обычно обработчики первым своим действием снова обращались к вызову `signal`, чтобы переустановить себя в качестве диспозиции, то есть чтобы следующий сигнал с тем же номером тоже привёл к вызову этой функции. Формально говоря, следующий сигнал мог успеть прийти за тот промежуток времени, пока действует диспозиция по умолчанию. К примеру, процесс перехватывает какой-нибудь `SIGUSR1`, чтобы при его получении выполнить какие-то действия, время от времени требующиеся пользователю (скажем, сбросить в файл промежуточные результаты вычислений). Затем первый пришедший сигнал с этим номером приводит к вызову функции-обработчика, обработчик пытается переустановить сигнал, но не успевает, поскольку сразу же за первым `SIGUSR1` приходит второй, и этот второй благополучно убивает процесс, ведь именно такова для этого сигнала диспозиция по умолчанию.

Разработчики ранних версий системы сигналов не беспокоились по этому поводу, поскольку в реальной жизни, если сделать переустановку сигнала самым первым действием функции-обработчика, то времени для этого хватит заведомо. Дело в том, что сигнал всегда обрабатывается в начале очередного кванта времени, и следующий сигнал не может быть обработан (и прийти, собственно, тоже не может) раньше, чем текущий квант времени истечёт и начнётся следующий, ну а на *одно обращение к системному вызову* кванта времени достаточно, более короткие кванты выделять бессмысленно, да и невозможно — не хватит частоты таймера. Однако всё это обусловлено одним конкретным подходом к реализации сигналов в ядре системы; включить соответствующие гарантии в спецификацию интерфейса сигналов означало бы фактически запретить любые другие подходы к реализации, что само по себе плохо. Кроме того, возникли бы неприятные вопросы на тему, *сколько же* времени есть у обработчика сигнала, чтобы успеть себя переустановить. Допустим, переустановка диспозиции первым действием успевает всегда; а если она будет не первым, а *вторым* действием? А третьим? Конечно, программист, досконально понимающий реализацию сигналов и устройство планировщика в данной конкретной системе, сможет ответить на такие вопросы достаточно достоверно; но ведь суть интерфейсов именно в том, чтобы их пользователь (программист, обращающийся к системным функциям) мог позволить себе не думать, как это всё реализовано «по ту сторону». Кроме того, отличительной особенностью программ на Си считается их *переносимость*, то есть неизменность свойств программы при её компиляции и запуске на совершенно другой системе, и в этом плане намертво завязывать поведение программы на недокументированные свойства одной конкретной реализации ядра тоже не вполне правильно.

К тому, как именно реализована обработка сигналов, мы вернёмся при обсуждении принципов реализации ядра системы.

Ещё один довольно ощутимый недостаток классической семантики обработки сигналов состоит в том, что (и это уже отнюдь не теоретическая возможность) в случае повторного прихода сигнала с тем же номером обработчик, переустанавливающий диспозицию сигнала снова на себя, но при этом (уже после переустановки диспозиции) работающий достаточно долго, мог быть прерван вызовом самого себя. Это не составляет проблемы, если не использовать глобальные переменные, но ведь для общения обработчика со всей остальной программой просто нет никаких средств, кроме глобальных переменных.

Классическую семантику сигналов часто называют «семантикой System V». Как мы уже знаем из вводной части, в конце 1970-х семейство систем Unix разделилось на две основные ветви — точнее, от «основной» линии развития unix-систем отпочковалась ветвь университета Беркли, известная как BSD. Господствующей версией Unix «основной» линии на тот момент была System V, так что программисты при обозначении наиболее заметных различий между «основной» линией и линией Беркли упоминают с одной стороны BSD (что и понятно), а с другой — обычно как раз System V.

Создатели BSD ликвидировали «шероховатость» семантики сигналов, введя сразу два новшества. Во-первых, в BSD-системах диспозиция сигнала могла быть изменена только явным образом, то есть пока процесс сам не обратится к системе с просьбой об изменении диспозиции, никто другой её не изменит. Иначе говоря, при вызове функции-обработчика диспозиция в BSD-системах остаётся прежней, то есть повторный приход сигнала снова вызовет тот же обработчик, переустанавливать его не нужно. Во-вторых, на время работы функции-обработчика сигнал с данным номером блокируется, то есть функция не может быть вызвана второй раз из-за пришедшего сигнала, пока первый её вызов не закончился. К сожалению, в качестве интерфейса к системе сигналов в BSD оставили системный вызов с тем же именем и параметрами, что привело к одновременному существованию систем как со старой («одноразовой» и без блокировки), так и с новой (постоянной и с блокировкой) семантикой.

Интересно, что ядро Linux обрабатывает системный вызов `signal`, имеющий классическую семантику System V, но большинство версий стандартной библиотеки скрывает этот факт от пользователя, реализуя библиотечную функцию с таким же именем, но имеющую семантику BSD. Обращение к ядру при этом производится через другой системный вызов, `sigaction`.

В современных программах для изменения диспозиции сигналов обычно рекомендуют использовать именно вызов `sigaction`, а не `signal`; отметим, что при переходе от 32-битной архитектуры к 64-битной набор системных вызовов ядра Linux слегка изменился, и в новый набор вызов `signal` не вошёл; впрочем, 64-битные ядра Linux продолжают поддерживать полный набор 32-битных системных вызовов, куда `signal`, естественно, как входил, так и входит; кроме того, функция `signal` (реализованная через `sigaction`) по-прежнему предоставляется стандартной библиотекой, так что изъятие вызова `signal` из интерфейса ядра не должно было привести ни к каким нарушениям совместимости.

Вызов `sigaction` появился позже, имеет одинаковую семантику во всех unix-системах и существенно более гибок. В частности, он позволяет задействовать сигналы, несущие на себе дополнительную информацию, а также очереди сигналов (для этого потребуется вызов `sigqueue` вместо привычного `kill`); вся эта функциональность появилась в системах семейства Unix относительно недавно — в начале XXI века. К сожалению, подробное описание вызова `sigaction` оказывается громоздким и перегруженным техническими деталями, поэтому мы оставляем его читателю для самостоятельного изучения.

Чтобы написанная программа вела себя более-менее одинаково при любом из двух вариантов семантики вызова `signal`, следует переустанавливать режим обработки каждый раз в начале функции-обработчика, либо, наоборот, сбрасывать режим обработки в `SIG_DFL`, если требуется перехватить только один сигнал.

Подчеркнём, что функция-обработчик не вызывается из основной программы и не возвращает ей управление; в определённом смысле можно сказать, что её вызывает ядро операционной системы. Так или иначе, не имея ни возможности передать информацию обработчику через параметры, ни получить информацию от неё через возвращаемое значение, основная программа вынуждена взаимодействовать с функцией-обработчиком через глобальные переменные, ведь других способов связи с ней не остаётся.

Сигнал может, как мы уже отмечали, прийти в самый неожиданный момент, когда программа находится в произвольной точке выполнения, и это также порождает определённые сложности. Большинство функций стандартной библиотеки изнутри обработчиков сигналов вызывать нельзя, ведь сигнал может прийти процессу во время выполнения той же самой функции или какой-то другой функции, работающей с теми же данными; функция, выполняемая в основном процессе, может привести некие «глобальные» данные в нецелостное состояние, которое никогда не наблюдается и не должно наблюдаться извне, возникает лишь на короткий промежуток времени, и которое ни сама эта функция, ни другие функции не рассчитывают «увидеть», когда бы их ни вызвали. Поскольку из обработчика никак нельзя определить, в какой момент была прервана основная программа, дальнейшее вмешательство в нецелостные структуры данных приведёт к непредсказуемым последствиям.

Прежде всего это касается динамической памяти, ведь динамическая память как целое — так называемая ***куча*** — это довольно сложная структура данных. Вызывая из обработчика сигнала `malloc` или `free`, мы рискуем попасть ровно на тот момент, когда основная программа тоже находится внутри `malloc`'а или `free`, так что, не зная о присутствии друг друга, экземпляр, вызванный из основной программы, и экземпляр, вызванный из обработчика сигнала, вполне могут разрушить кучу. Наша программа после этого, скорее всего, «свалится», причём не сразу. Такие ошибки проявляются с достаточно низкой вероятностью, ведь для этого нужно, чтобы обработчик сигнала оказался вызван именно тогда, когда основная программа что-то делает с кучей; как известно, ошибка, которую тяжело или вообще невозможно воспроизвести — это самый кошмарный вариант с точки зрения трудоёмкости и утомительности отладки.

По тем же причинам нельзя использовать в обработчиках сигналов функции высокоуровневого ввода-вывода. Причина здесь кроется в буферизации потоков; для её организации с каждым потоком ввода-вывода библиотека вынуждена связывать буфер, который тоже представляет собой нетривиальную структуру данных.

С другой стороны, ввод-вывод изнутри обработчиков сигналов можно произвести с помощью системных вызовов, но здесь тоже имеется своя сложность, заключающаяся в наличии глобальной переменной `errno`. Представьте себе, что основная программа выполнила системный вызов, который завершился с ошибкой, но воспользоваться значением `errno` не успела, и тут как раз процессу пришёл сигнал, запустился обработчик, и в нём тоже выполнился какой-то системный вызов, закончившийся ошибкой. Естественно, он запишет результат своего выполнения в `errno`, так что диагностика, которую потом выдаст основная программа, не будет иметь ничего общего с действительностью.

Даже простые обращения к глобальным переменным из обработчика сигнала могут оказаться неприемлемыми. Так, если мы работаем с 32-битным процессором, присваивание значения переменной типа `long long` будет выполняться в две команды, как и извлечение значения из такой переменной. Если представить, что основная программа начала извлекать значение из переменной, но успела выполнить только одну машинную команду из двух, а в это время функция-обработчик занесла в переменную новое значение, то значение, которое в итоге извлечёт из переменной основная программа, окажется состоящим из обрывков двух значений: четыре байта от старого значения, четыре байта от нового. Правильных результатов после этого уже не получить.

Мало того, в дело могут вмешаться ещё и особенности компилятора Си. Как известно, современные компиляторы предполагают оптимизацию машинного кода; в частности, компилятор может предположить, что, коль скоро в некую переменную никто ничего не записывает «прямо здесь», то её значение измениться не может, и если несколько команд назад это значение было прочитано из памяти в регистр, оптимизатор вполне может выбросить из кода повторное чтение той же памяти, воспользовавшись значением, всё ещё находящимся в регистре. Чтобы избежать подобных эффектов, язык Си предусматривает ключевое слово `volatile` для пометки переменных, значение которых может «неожиданно» измениться.

Современные спецификации системы сигналов предлагают программисту воспользоваться специальным типом `sig_atomic_t`; предполагается, что обращение к переменным этого типа (если совсем строго — чтение значения из памяти и запись значения в память) гарантированно проходит *атомарно*, то есть за одно неделимое действие. В действительности речь идёт не о настоящей атомарности, а об *атомарности относительно сигналов*: требуется, чтобы обработка очередного сигнала не могла вклиниться между частями одной операции чтения или записи переменной. Заниматься многопоточным программированием (мультитредингом) мы с вами не собираемся, во всяком случае, всерьёз[30], и уж точно не будем обсуждать обработку сигналов в многопоточных программах; в отсутствие же многопоточности — точнее, до тех пор, пока рассматриваемая переменная не находится в памяти, разделяемой между независимо выполняющимися «действующими лицами» — **для обеспечения искомой «сигнальной атомарности» достаточно, чтобы арифметические операции над этой переменной выполнялись за одну машинную команду**. В самом деле, мы уже знаем, что при наступлении прерывания процессор либо доводит выполнение текущей машинной команды до конца, либо откатывает её, как если бы она выполняться и не начинала; ну а никакие иные силы, кроме прерываний, в принципе не способны ничего «вклинить» в обычную работу с переменными, если только эти переменные не располагаются в разделяемой памяти.

Припомнив наши познания в ассемблерном программировании, мы можем заметить, что даже на 32-битных процессорах все операции с четырёхбайтными целыми происходят за одну команду, причём коль скоро речь идёт о процессорах семейства i386, за одну команду выполняются не только чтение и запись, но и все арифметические операции. Можно заметить, что в наше время работающий компьютер с 32-битным процессором уже приходится специально искать, а на 64-битных процессорах, что вполне естественно, восьмибайтные (64-битные) целые тоже читаются и записываются за одну команду; но если для коммуникации с обработчиком сигнала вам потребовалось 64-битное целое, вы совершенно точно делаете что-то не так.

Раскроем теперь небольшой секрет: вы вряд ли сможете найти такую версию заголовочных файлов стандартной библиотеки, где тип `sig_atomic_t` определён как что-то отличное от простого `int`. Именно этот тип — `int` — мы и будем использовать во всех наших примерах; комитетский корявый `sig_atomic_t`, как водится, просто ни для чего не нужен. Важно только не

[30] Вся следующая (седьмая) часть нашей книги посвящена параллельному программированию, и там, естественно, будут рассматриваться примеры многопоточных программ, но там же содержатся объяснения, почему всем этим не стоит заниматься иначе как под дулом пистолета.

забывать, что с типами размера *бóльшего*, чем `int`, могут возникнуть проблемы.

Ещё один важный момент, который необходимо учитывать при обработке сигналов — это поведение блокирующих системных вызовов. Если процесс выполнил системный вызов, который перевёл его в состояние блокировки — например, запросил с помощью `read` чтение данных из потока, в котором сейчас нет данных, или с помощью функции `sleep` сообщил операционной системе, что в ближайшее время ничего делать не собирается, и т. п — то при получении процессом обрабатываемого сигнала такой системный вызов немедленно вернёт управление. В большинстве случаев системный вызов, прерванный сигналом, вернёт значение `-1`, то есть заявит о происшедшей ошибке. Отличить эту ситуацию от настоящих ошибок можно по значению переменной `errno`: она будет равна константе `EINTR`.

Дело в том, что для запуска функции-обработчика процессу нужно быть в состоянии обычного выполнения, а реализованы сигналы так, что после завершения функции-обработчика неизбежно будет продолжено выполнение кода основной программы — с того места, где оно было прервано пришедшим сигналом. Если при этом процесс находился в состоянии «выполнения в ядре», в том числе в состоянии блокировки, то вернуться обратно в ядро он может только по своей инициативе. Ядро не может (во всяком случае, надёжно) отследить момент завершения выполнения обработчика и отнять у процесса инициативу, принудительно вернув его «в ядро»; почему это так, станет ясно при рассмотрении механизма реализации вызова обработчиков сигналов.

С неблокирующими вызовами проблем не возникает: ядро завершает обработку такого вызова, что обычно не занимает много времени, после чего возвращает управление коду процесса, и в этот момент как раз отрабатывает функция-обработчик сигнала. С блокирующими вызовами этот номер не проходит: в блокировке процесс может «провисеть» сколь угодно долго, и если допустить, чтобы всё это время процесс не реагировал на сигналы, это сделало бы всю систему сигналов бессмысленной. Поэтому операционная система предпочитает немедленно прекратить обработку имеющегося системного вызова, чтобы дать возможность процессу получить и обработать сигнал, но о том, что вызов не был завершён, естественно, процессу сообщает — вызов сигнализирует об «ошибке» `EINTR`. Как правило, программы пишут так, чтобы в этом случае снова обратиться к «неожиданно прерванному» системному вызову.

На самом деле, как ни странно, это происходит не всегда. Уже упоминавшийся вызов `sigaction` позволяет при установке диспозиции сигнала указать ряд настроек, в число которых входит флажок `SA_RESTART`; если он указан, то при поступлении обрабатываемого сигнала процессу, который находится в состоянии блокировки *в одном из явно перечисленных в документации блокирующих системных вызовов*, система должна вывести процесс из состояния блокировки, дать ему квант времени для обработки сигнала, а затем — сразу после выхода из обработчика — вернуть процесс в состояние блокировки. Картина осложняется тем, что так происходит далеко не для всех блокирующих системных вызовов; например, `wait` и `write` перезапускаются, тогда как `usleep` и `nanosleep` при поступлении обрабатываемого сигнала всегда возвращают ошибку `EINTR`; с вызовом `read` всё ещё интереснее: в большинстве случаев он перезапускается, *но не всегда*. К тому же процесс может «обмануть» ядро, не дав ему шанса что-то перезапустить.

Некоторые (опять же, не все) версии библиотеки, которые эмулируют вызов `signal` через обращение к `sigaction`, вдобавок зачем-то устанавливают флажок `SA_RESTART`, что окончательно запутывает дело.

Правильнее всего, конечно, было бы отказаться от `signal` в пользу `sigaction`, а `SA_RESTART` либо не использовать вовсе, либо перед его использованием проштудировать списки системных вызовов, умеющих и не умеющих перезапускаться при поступлении сигнала; но всё это влечёт заметные трудозатраты, а роль сигналов в жизни вашей программы обычно достаточно скромна. Поэтому можно посоветовать другой вариант: если в своей программе вы обрабатываете сигналы, для каждого блокирующего системного вызова будьте готовы к тому, что он завершится с «ошибкой» `EINTR` (которая на самом деле вовсе никакая не ошибка), но при этом не полагайтесь на то, что так обязательно произойдёт.

Напишем для примера программу, которая при нажатии `Ctrl-C` сначала выдаёт сообщение и лишь на 25-й раз завершается. Для начала сделаем это так, как обычно такие программы пишут начинающие:

```
#include <stdio.h>
#include <signal.h>
int n = 0;
void handler(int s)
{
    n++;
    printf("Press it again, I like it\n");
}
int main()
{
    signal(SIGINT, handler);
    while(n < 25)
        {}
    return 0;
}
```

Несмотря на то, что эта программа успешно пройдёт компиляцию и даже будет работать в соответствии с условиями задачи, правильной она не является. Попробуем исправить ошибки. Переменную `n` пометим словом `volatile`, чтобы исключить сюрпризы со стороны оптимизатора. В начало обработчика вставим переустановку диспозиции. Поскольку применение `printf` в обработчике сигнала неуместно, заменим его на системный вызов `write`; для этого, во-первых, опишем сообщение в виде глобальной строковой константы, и, во-вторых, снабдим обработчик действиями по сохранению и восстановлению значения `errno`. Наконец, обратим внимание, что наша программа непроизводительно расходует процессорное время в бесконечном цикле и вставим туда вызов `sleep(1)`, так что программа будет просыпаться один раз в секунду, чтобы снова заснуть; сколько-нибудь заметной нагрузки на процессор она уже не создаст.

```
#include <signal.h>                          /* pressagain.c */
#include <unistd.h>
#include <errno.h>
volatile static int n = 0;
const char message[] = "Press it again, I like it\n";
static void handler(int s)
{
    int save_errno = errno;
    signal(SIGINT, handler);
    n++;
    write(1, message, sizeof(message)-1);
    errno = save_errno;
}
int main()
{
    signal(SIGINT, handler);
    while(n < 25)
        sleep(1);
    return 0;
}
```

Разумеется, мы могли бы использовать и другой аргумент для `sleep`, например, при использовании `sleep(3600)` программа просыпалась бы раз в час, но и один раз в секунду — это для наших целей уже достаточно редко. Напомним, что при получении процессом обрабатываемого сигнала `sleep` всегда возвращает управление, так что, сколь бы долгий период сна мы ни задали, сразу после обработки сигнала произойдёт очередная проверка условия `while`.

ОС Unix предусматривает специальный системный вызов `pause` как раз для таких случаев. Этот вызов, не принимающий параметров, блокирует процесс до тех пор, пока не будет получен обрабатываемый сигнал. Интересно, что единственный вариант возвращаемого значения для вызова `pause` — это `-1`, то есть «ошибка», при этом в `errno` заносится значение `EINTR`. Если заменить в нашей программе строку «`sleep(1);`» на «`pause();`», она станет «ещё правильнее».

К сожалению, полностью правильной программа не станет даже после этого. Ближе к концу этого параграфа мы попытаемся прояснить один хитрый момент, связанный с природой сигналов, и покажем, какими внешними действиями можно заставить нашу программу работать совсем не так, как мы от неё ожидали. Но прямо сейчас мы в эти дебри не полезем во избежание перегрузки мозгов.

Отметим ещё одно странное свойство сигналов: **если одному и тому же процессу отправить несколько сигналов с одним и тем же номером, они могут «склеиться», то есть процесс получит только один сигнал** (или не один, но меньше, чем их было отправлено). Почему такое может случаться, мы поймём при обсуждении реализации сигналов в последней части этого тома.

С помощью вызова `alarm` можно затребовать от ядра отправки вызвавшему процессу сигнала `SIGALRM` через определённое количество секунд времени. Прототип вызова таков:

```
int alarm(unsigned int seconds);
```

Параметр задаёт количество секунд, через которое следует прислать сигнал. Когда заказанный период времени истекает, система присылает процессу сигнал, а сам активный «заказ» сбрасывается. Возвращаемое вызовом `alarm` значение зависит от того, имеется ли уже для данного процесса активный заказ на отправку `SIGALRM`. Если нет, вызов возвращает ноль. Если же активный заказ уже есть, возвращено будет количество секунд (полных или неполных, то есть всегда не менее 1), оставшееся до момента его исполнения.

Система может помнить только об одном заказанном сигнале `SIGALRM` для каждого процесса. Если по результатам предыдущего вызова сигнал процессу ещё не прислали, новый вызов отменит старый заказ и установит новый. Нулевое значение параметра `seconds` отменит активный заказ, не установив новый.

Следует учитывать, что по умолчанию сигнал `SIGALRM` *убивает* процесс, так что в большинстве случаев нужно *сначала* изменить диспозицию `SIGALRM`, установив функцию-обработчик, и лишь после этого обращаться к вызову `alarm`.

В описании функции `sleep` говорится, что она может быть реализована через вызов `alarm` и обработку `SIGALRM`, так что использование этих двух механизмов не следует смешивать в одной программе. В реальности, конечно, ничего подобного не происходит; например, в большинстве версий Linux функция `sleep` реализуется через обращение к системному вызову `nanosleep`, но с таким же успехом её можно было бы реализовать через системный вызов `select`, который мы будем рассматривать позже; есть и другие способы.

Про `usleep` ничего подобного в спецификации не говорится, а про `nanosleep` в явном виде сказано, что она никак не пересекается с обработкой сигналов. Кроме того, стоит заметить, что `usleep` и `nanosleep`, будучи прерваны пришедшим сигналом, ведут себя как обычные блокирующие системные вызовы: возвращают `-1`, занося в `errno` значение `EINTR`. Функция `sleep` в такой ситуации *возвращает число секунд (полных или неполных), оставшихся до предполагавшегося момента пробуждения*.

В целом можно сделать вывод, что лучше вообще не использовать `sleep`, отдавая предпочтение `usleep` и `nanosleep`; впрочем, это скорее дело вкуса.

Несмотря на кажущуюся простоту, активная работа с сигналами требует высокой квалификации. При использовании сигналов часто возникают ситуации гонок, сами сигналы ненадёжны, при отправке двух одинаковых сигналов прийти может только один и т. д. Если ваша программа обрабатывает сигналы, нужно помнить об этом при обращении ко всем блокирующим системным вызовам, ведь каждый из них может «ошибиться» просто из-за того, что во время его выполнения процесс получил сигнал. Написание корректной программы, активно использующей сигналы, может оказаться весьма нетривиальным делом.

Чтобы понять, о чём идёт речь, попробуем прикинуть, с чем мы можем столкнуться, если попытаемся задействовать сигналы для общения двух процессов между собой по довольно простому сценарию. Пусть процесс `A` для какой-то своей надобности порождает (с помощью обычного `fork`) процесс `B`. Порождённый процесс должен выполнить некую подготовительную работу, до завершения которой он к решению своей основной задачи не готов. О достижении готовности он извещает своего родителя сигналом `SIGUSR1`, после чего (но не раньше) его можно «эксплуатировать». Родительскому процессу по каким-то причинам нечего делать, пока порождённый не будет готов к работе, так что нужно дождаться прихода `SIGUSR1` и тогда уже работать. Дело осложняется тем, что родительский процесс обрабатывает и другие сигналы, так что ему недостаточно, например, просто вызвать `pause`, а нужно ещё проверить, тем ли сигналом вызов был прерван.

На первый взгляд кажется, что тут всё очень просто: создадим глобальную переменную (пусть она, например, называется `child_ready`), перед созданием процесса-потомка занесём в неё ноль, а диспозицию сигнала `SIGUSR1` изменим, установив обработчик, который в ту же переменную заносит единицу. Затем после `fork` устроим цикл обращений к вызову `pause` до тех пор, пока в `child_ready` не окажется единица, и тогда уже начнём активную работу с потомком, который теперь готов нас обслуживать. Общая схема получается такой:

```
volatile int child_ready;
void usr1hdl(int n)
{
    child_ready = 1;
}

int main()
{
    int pid;
    /* ... подготовка к работе главного процесса ... */

    child_ready = 0;
    signal(SIGUSR1, usr1hdl);
    pid = fork();
    if(pid == 0) {  /* порождённый процесс */
        /* ... подготовка ... */
        kill(getppid(), SIGUSR1);
        /* ... основная работа ... */
        exit(0);
    }
    /* продолжение родительского процесса */
    while(!child_ready)
        pause();
    /* ... считаем, что теперь потомок готов к работе ... */
    /* ... */
}
```

Увы, программа, построенная по такой схеме, заведомо неправильна, и вот почему. Допустим, у родительского процесса истёк квант процессорного времени сразу после первой проверки значения `child_ready` — то есть когда процесс уже знает, что в этой переменной всё ещё ноль, но уйти в вызов `pause` не успел. Пока родитель ожидает следующего кванта, потомок успевает завершить свою подготовительную работу и отправить предку сигнал. В начале очередного кванта времени предок, разумеется, этот сигнал обработает, переменная `child_ready` получит значение 1, вот только будет поздно, проверка-то уже позади! Так что сразу по завершении обработчика сигнала наш родительский процесс окажется в блокировке внутри вызова `pause`, и сколько он там провисит — зависит разве что от благосклонности небесных светил; поскольку закон подлости никто не отменял, разумеется, никто не станет присылать этому процессу никаких сигналов в течение времени, достаточного, чтобы «всё сломалось» (конкретное значение тут зависит от решаемой задачи).

Чтобы наше решение можно было превратить в правильное, ядро предусматривает системные вызовы `sigprocmask` и `sigsuspend`. Общая идея тут в том, что нужный нам сигнал (в данном случае `SIGUSR1`) мы должны с помощью `sigprocmask` заблокировать, а вместо `pause` использовать `sigsuspend`, которому указать, что на время его выполнения сигнал `SIGUSR1` нужно, наоборот, разблокировать. В этом случае, сколь бы рано сигнал ни пришёл, родительский процесс его получит только после захода в `sigsuspend`.

Если к текущему моменту у вас сложилось ощущение, что мозг норовит взорваться или вскипеть, не отчаивайтесь. Дело в том, что сигналы в действительности вообще не следует использовать для взаимодействия процессов между собой. Основных областей применения сигналов существует две. Первая — завершать процессы; для этого сигналы может применять как пользователь-человек, так и ядро ОС (для завершения аварийных процессов). Ясно, что, когда мы собираемся *уничтожить* процесс, нам более-менее всё равно, «правильно» он сдохнет или неправильно, прибьём-то мы его в любом случае. Вторая область применения сигналов — дать пользователю (человеку!) возможность что-то сообщить процессу, который в силу своих задач выполняется достаточно долго и который по каким-то причинам не хочется прерывать и запускать заново; например, мы можем с помощью сигнала сообщить какому-нибудь серверу, что нужно перечитать его конфигурационные файлы, или какую-нибудь программу, перемалывающую большой объём данных, попросить таким способом, чтобы она выдала свои промежуточные результаты. В этом случае от правильности обработки сигналов практически ничего не зависит: если мы отправим процессу сигнал и умудримся попасть именно в тот хитрый момент между проверкой и входом в вызов `pause` (или любой другой блокирующий вызов), что само по себе практически невероятно — то ничего страшного всё равно не произойдёт, мы ведь увидим, что процесс не отреагировал на наш сигнал, и, скорее всего, пошлём ему тот же сигнал снова.

Остаток параграфа мы посвятим *правильной* организации ожидания сигнала с помощью `sigsuspend`; если этот материал покажется вам слишком

заумным, вы можете его безболезненно пропустить и вернуться к нему позднее, когда окажетесь к этому готовы.

Для начала отметим, что система позволяет процессу временно заблокировать доставку некоторых сигналов; как раз для этого предназначен вызов sigprocmask. Если процессу послать сигнал, который у него сейчас заблокирован, то доставка этого сигнала (иначе говоря, выполнение действий, предписанных диспозицией сигнала) будет отложена до тех пор, пока процесс этот сигнал не разблокирует. Естественно, заблокировать сигналы SIGKILL и SIGSTOP нельзя — с ними вообще ничего нельзя сделать.

Профиль системного вызова sigprocmask выглядит так:

```
int sigprocmask(int how, const sigset_t *set, sigset_t *oldset);
```

Второй и третий параметры используют переменные специального типа sigset_t; такая переменная представляет *множество сигналов*. Первый параметр (how) указывает, какое действие следует произвести с множеством сигналов, заблокированных на текущий момент, и может принимать одно из трёх значений: SIG_BLOCK, SIG_UNBLOCK и SIG_SETMASK; в первом случае сигналы, входящие в множество set, блокируются, во втором, наоборот, разблокируются, а в третьем — блокируются все сигналы, входящие в set, а все остальные разблокируются. Если адрес, переданный третьим параметром (oldset), не нулевой, то в переменную, находящуюся по этому адресу, вызов запишет множество сигналов, которые были заблокированы на момент обращения к sigprocmask.

Вызовом sigprocmask можно также воспользоваться, чтобы просто узнать, какие сигналы сейчас заблокированы. Для этого вторым параметром передают NULL; первый параметр при этом игнорируется (обычно его указывают равным нулю), а в третий, как обычно, записывается множество заблокированных сигналов, но вызов при этом ничего не блокирует и не разблокирует.

Для работы с множествами сигналов предусмотрены специальные функции:

```
int sigemptyset(sigset_t *set);
int sigfillset(sigset_t *set);
int sigaddset(sigset_t *set, int signum);
int sigdelset(sigset_t *set, int signum);
int sigismember(const sigset_t *set, int signum);
```

Функция sigemptyset очищает заданное множество, т. е. делает его пустым; sigfillset вносит в множество все существующие сигналы; sigaddset добавляет заданный сигнал в заданное множество, sigdelset удаляет из множества заданный сигнал; sigismember позволяет узнать, входит ли данный сигнал в множество — она возвращает 1, если сигнал в множество входит, 0 — если не входит, -1 в случае ошибки. Остальные функции возвращают 0, если всё в порядке, и -1 в случае ошибки. Отметим, что ошибка тут может быть только одна: параметр signum не является номером существующего сигнала; переменная errno получает при этом значение EINVAL.

Профиль вызова sigsuspend, который следует использовать для ожидания прихода сигнала вместо pause, таков:

```
int sigsuspend(const sigset_t *mask);
```

Как видим, параметром ему служит переменная всё того же типа sigset_t; параметр задаёт множество сигналов, которые будут заблокированы на время работы вызова. Точнее говоря, вызов на время своей работы *заменяет* действовавшее множество заблокированных сигналов на то, которое указано в его параметре, а перед возвратом управления меняет их обратно, т. е. восстанавливает множество заблокированных сигналов в том виде, в котором оно было на момент обращения к вызову.

Процесс, обратившийся к вызову sigsuspend, блокируется до тех пор, пока не придёт сигнал, *разрешённый* параметром mask (то есть не входящий в это множество), причём такой, диспозиция которого — либо завершение процесса, либо вызов функции-обработчика. Сигналы, для которых установлена диспозиция SIG_IGN, а равно и такие, которые не собираются убивать процесс по какой-то другой причине (например, SIGCHLD в диспозиции по умолчанию), никакого влияния на процесс, находящийся в вызове sigsuspend, не оказывают. Если пришедший сигнал должен процесс уничтожить, вызов sigsuspend управление уже не вернёт, а процесс будет убит обычным для данного сигнала образом. Если же процессу, «висящему» в sigsuspend, отправить сигнал, диспозиция которого — вызов функции-обработчика, то сначала отработает эта функция, а затем уже sigsuspend вернёт управление; примечательно, что возвращает он при этом -1 (как будто бы при ошибке, хотя это совершенно не ошибка), а errno получает значение EINTR. Сделано это для единообразия: как мы помним, именно так ведут себя *все* блокирующие системные вызовы, когда их работу прерывают сигналом. Как следствие, проверять возвращённое вызовом значение бессмысленно: ничего, кроме -1, он вернуть всё равно не может.

Описание sigsuspend вроде бы не содержит ничего сложного, но вопрос «а как с этим работать» возникает у начинающих системщиков с завидной регулярностью. Дело тут в том, что множества сигналов («маски») предназначены, чтобы *блокировать* сигналы, и сей факт подспудно создаёт ощущение, что в параметре вызова sigsuspend следует перечислить сигналы, которые на время работы вызова нужно заблокировать, то есть как будто бы должно быть так: программа некоторое время работает, потом с помощью sigsuspend запрещает доставку некоторых сигналов, ждёт прихода, по-видимому, какого-то другого сигнала (и зачем, спрашивается, тогда какие-то сигналы блокировать?), а когда дождётся — обратно разблокирует всё, что раньше заблокировала, и продолжает работу.

В действиях по такой схеме, конечно же, никакого смысла нет. На самом деле всё должно быть совсем наоборот: **параметр** mask **вызова** sigsuspend **нужен не для того, чтобы какие-то сигналы временно заблокировать, а как раз напротив — чтобы их *разблокировать* точно на время работы вызова, а затем заблокировать обратно**.

Общая идея тут в том, чтобы позволить нужному нам сигналу прийти (быть обработанным) лишь тогда, когда мы этого ожидаем, а ожидать мы этого, понятное дело, будем внутри вызова sigsuspend. Всё остальное время сигнал должен быть заблокирован. В этом случае сигнал, пришедший слишком рано (как в примере, разобранном выше), останется заблокирован до тех пор, пока мы не войдём в sigsuspend; но как только мы в него войдём, он сигнал разблокирует, обработает и, как следствие, тут же вернёт управление. Если же сигнал придёт слишком поздно, т. е. уже после выхода из sigsuspend, то мы обработаем его при следующем обращении к sigsuspend (если, конечно, такое предусмотрено — но если повторное обращение не предусмотрено, то, по-видимому, нашей программе больше сигналы просто не нужны, что, как водится, зависит от решаемой задачи). Например, правильная реализация сценария, при котором родительский процесс дожидается сигнала от потомка, и только после этого продолжает работу, будет примерно такой (здесь мы приводим только текст функции main, поскольку обработчик сигнала будет точно такой же, как и на стр. 129):

```
int main()
{
    int pid;
    sigset_t mask_usr1, mask_empty;
    sigemptyset(&mask_usr1); /* множество из одного SIGUSR1 */
    sigaddset(&mask_usr1, SIGUSR1);
    sigemptyset(&mask_empty); /* пустое множество */

    /* ... подготовка к работе главного процесса ... */

    child_ready = 0;
    sigprocmask(SIG_SETMASK, &mask_usr1, NULL);
    signal(SIGUSR1, usr1hdl);
    pid = fork();
    if(pid == 0) {  /* порождённый процесс */
        /* ... подготовка ... */
        kill(getppid(), SIGUSR1);
        /* ... основная работа ... */
        exit(0);
    }
    /* продолжение родительского процесса */
    while(!child_ready)
        sigsuspend(&mask_empty);
    /* ... теперь можно работать с потомком, он готов ... */
    /* ... */
}
```

Теперь мы, хочется надеяться, готовы к объяснению, по какой такой причине наша первая программа с сигналами, та, что должна выдавать 25 сообщений по нажатию Ctrl-C, так и осталась формально неправильной даже после всех внесённых в неё усовершенствований.

Для начала отметим довольно очевидный момент, аналогичный тому, с чем мы столкнулись, ожидая от порождённого процесса «доклада» о его готовности. Последний, 25-й по счёту сигнал может прийти в тот момент, когда в функции main проверка условия цикла уже пройдена, но вызвать pause программа ещё не успела. В этом случае наша программа так и будет «висеть», пока не придёт ещё один сигнал, а поскольку обрабатываемый сигнал у нас только один, это должен быть именно он, т. е. пользователю придётся для завершения программы нажать Ctrl-C не 25, а 26 раз. *Сущий кошмар, не так ли*.

Шутки шутками, но с формальной точки зрения такая возможность делает программу неправильной. Впрочем, программу в том виде, в котором мы её оставили, можно заставить делать вещи намного более странные. В архиве примеров к нашей книге вы найдёте программу int_grill.c, которая получает через аргументы командной строки имя (и, возможно, аргументы) для запуска внешней программы; запустив указанную внешнюю программу обычным путём через комбинацию fork/execlp, программа int_grill выжидает «для приличия» десятую долю секунды, после чего в цикле шлёт своему дочернему процессу сигнал SIGINT, проверяет с помощью waitpid с WNOHANG, жив ли ещё дочерний процесс, и если да, то продолжает цикл, то есть снова шлёт сигнал — и так пока внешняя программа не завершится тем или иным способом. Текст программы int_grill мы для экономии места не приводим, там всё достаточно очевидно (впрочем, если что-то в её устройстве непонятно — архив примеров никто не отменял). Если запустить наш исходный пример (напомним, что он называется pressagain) через int_grill — мы немедленно обнаружим, что строк Press it

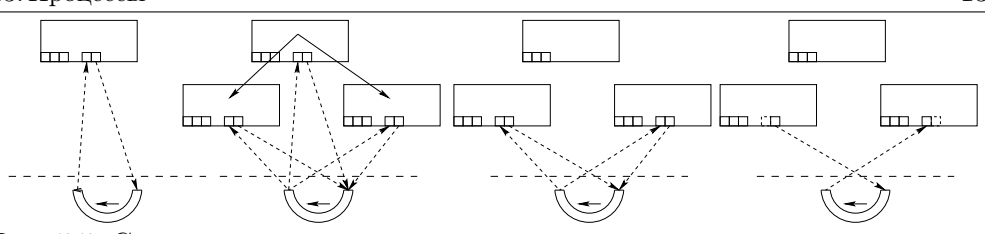

Рис. 5.5. Связывание двух процессов-потомков через неименованный канал

`again, I like it` печатается намного больше, чем 25, хотя (судя по тому, что завершается такой эксперимент всё же довольно быстро) и не так чтоб строк печаталось очень много. Но вот если, поленившись считать выданные строки, мы перенаправим поток вывода конвейером на вход программе `wc -l` (напомним, программа Word Count с этим ключом как раз подсчитывает в своём потоке стандартного ввода строки текста), то есть применим команду вроде

```
./int_grill ./pressagain | wc -l
```

— можно будет наблюдать совершенно неожиданные числа от нескольких сотен до *сотен тысяч*. Лишившись естественного тормоза в виде сравнительно медленной перерисовки текста в окошке терминала, наша связка программ оказывается, как видим, намного шустрее.

Всё дело в том, что программа `int_grill`, которая не занимается ничем, кроме посылания сигнала своему потомку (в данном случае — программе `pressagain`), бóльшую часть времени успевает этих сигналов отправить больше, чем дочерний процесс успевает их обработать. Лишние сигналы при этом «склеиваются», но того, что остаётся, хватает, чтобы к моменту завершения функции-обработчика очередной сигнал уже ждал обработки. Здесь не помогает даже то, что функция `signal` из библиотеки `glibc` реализует семантику BSD, то есть, кроме прочего, блокирует доставку очередного сигнала с тем же номером на время выполнения обработчика. Стековый фрейм, созданный ядром ОС в стеке нашей задачи для выполнения функции-обработчика, устроен несколько хитро: когда обработчик возвращает управление, передача управления в действительности происходит на «подсунутый» ядром кусочек кода, который обращается к системному вызову `sigreturn`. Этот системный вызов, в частности, разблокирует повторную доставку сигнала, который мы только что обрабатывали, а поскольку в большинстве случаев очередной `SIGINT` к этому времени уже пришёл, на возврате управления от ядра после `sigreturn` снова срабатывает установленная диспозиция и управление опять отдаётся обработчику. До цикла, расположенного в главной программе, дело просто не доходит.

К счастью для нашей программы, так происходит всё же не всегда: у предка выполнение сначала `kill`, а потом ещё `waitpid` тоже занимает какое-то время, так что иногда потомку «везёт»: предок, замешкавшись, не успевает прислать очередной сигнал, так что цикл в главной программе потомка всё-таки ухитряется прокрутиться и вывалиться. Именно поэтому наша связка, прогнав через конвейер пару сотен тысяч строк, всё-таки довольно быстро завершается — обычно мы даже не успеваем заметить задержку. Но получить сотню тысяч строк вместо ожидавшихся двадцати пяти, да ещё по причине, которую приходится так долго объяснять — уже в достаточной степени неприятно.

Решить исходно поставленную задачу полностью корректно можно лишь одним способом: исходно (ещё до изменения диспозиции) заблокировать сигнал `SIGINT` с помощью `sigprocmask`, а в цикле заменить `pause` на `sigsuspend`, *разблокирующий* `SIGINT` на время своей работы. В этом случае ни в какие «неожиданные моменты» сигнал уже не придёт.

Некоторые программисты считают, что использовать `pause` нельзя никогда и ни для чего, и вообще следует считать некорректной и ни на что не годной любую программу, работающую с сигналами и не применяющую `sigsuspend` и его аналоги (например, вызов `pselect`, который мы рассмотрим в части, посвящённой компьютерным сетям). Автор книги рискнёт с ними не согласиться.

Припомним, как была исходно поставлена задача, решаемая программой `pressagain`. Мы хотели написать *программу, которая при нажатии `Ctrl-C` сначала выдаёт сообщение и лишь на 25-й раз завершается*. Отметим, что речь шла о нажатии `Ctrl-C`, а не о получении сигнала `SIGINT`, и если задачу рассматривать именно в такой формулировке, программа `pressagain` решает её совершенно правильно — во всяком случае, никто и никогда не сможет проявить хитрые особенности её поведения, *нажимая комбинацию на клавиатуре*. Дело даже не в том, что человек физически не способен нажимать клавиши с нужной для этого скоростью. Скорость тут требуется порядка миллионов, если не десятков миллионов нажатий в секунду; обеспечить такую скорость не сможет ни механическая часть клавиатуры, ни даже, что особенно забавно, связка программ, обеспечивающих передачу клавиатурных событий активной программе: драйвер терминала, дисциплина линии, эмулятор терминала, каналы передачи данных между ними — всё это рассчитано на пользователя-человека и просто не умеет функционировать на таких скоростях.

Гонясь за чистотой и каноничностью применяемых решений, не следует утрачивать элементарный здравый смысл.

5.3.15. Каналы

Канал — это объект ядра операционной системы, представляющий собой средство однонаправленной передачи данных от одного процесса другому. Можно представлять себе, что канал имеет два конца, один для записи, другой для чтения. С каждым концом канала могут быть связаны файловые дескрипторы, принадлежащие, возможно, разным процессам. Иначе говоря, канал — это такой объект ядра, с которым связываются поток ввода и поток вывода, причём информация, переданная каким-то процессом в канал через поток вывода, поступает в его поток ввода, откуда может быть считана другим процессом.

В ОС Unix различают *именованные* и *неименованные* каналы. Для работы с именованными каналами в файловой системе создаются файлы специального типа, открытие которых с помощью вызова `open` приводит к *подключению* к каналу; сам канал как объект ядра создаётся в тот момент, когда какой-то из процессов пытается открыть на запись или чтение файл именованного канала, который пока не открыт ни в каком другом процессе. С неименованными каналами ситуация выглядит иначе: они сразу создаются открытыми с обоих концов, при этом ни с какими файлами они не связаны, так что «подключиться» к уже созданному неименованному каналу невозможно.

Неименованный канал создаётся системным вызовом `pipe`:

```
int pipe(int fd[2]);
```

На вход ему подаётся адрес начала массива из двух элементов типа `int`; в этот массив вызов `pipe` записывает дескрипторы, связанные с созданным каналом: `fd[0]` — для чтения, `fd[1]` — для записи. В случае ошибки вызов возвращает `-1`; в правильно написанной программе `pipe` может завершиться ошибкой только при исчерпании предельного числа доступных дескрипторов. С объектом канала, созданного вызовом `pipe`, не связано ни имени, ни какой-либо иной идентифицирующей информации, есть лишь два дескриптора и всё. Единственный способ добиться того, чтобы к одному и тому же каналу оказались подключены разные процессы — это породить копию процесса, создавшего канал, с помощью вызова `fork`, так что использовать неименованный канал для взаимодействия между собой могут только процесс, создавший канал, и его потомки.

На рис. 5.5 показано связывание двух порождённых процессов с помощью неименованного канала. На первом шаге (до порождения новых процессов) родительский процесс создаёт канал. Затем родительский процесс создаёт двух потомков, в результате чего во всех трёх процессах оказываются дескрипторы как одного, так и второго концов канала. После этого родительский процесс закрывает свои экземпляры дескрипторов, чтобы не мешать потомкам использовать канал. В свою очередь, в первом потомке закрывается дескриптор, предназначенный для чтения, во втором — дескриптор, предназначенный для записи. В результате первый потомок может передавать второму данные через канал. Соответствующий код на языке Си будет выглядеть приблизительно так:

```
int fd[2];
int p1, p2;
pipe(fd);
p1 = fork();
if(p1 == 0) { /* child #1 */
    close(fd[0]);
    /* ... */
    write(fd[1], /* ... */);
    /* ... */
    exit(0);
}
p2 = fork();
if(p2 == 0) { /* child #2 */
    close(fd[1]);
    /* ... */
    rc = read(fd[0], /* ... */);
    /* ... */
    exit(0);
}
/* parent */
close(fd[0]);
close(fd[1]);
/* ... */
```

Рассмотрим ситуацию, когда в системе присутствуют открытые дескрипторы обоих концов канала. При попытке чтения из канала, в который пока ничего не записано, читающий процесс будет заблокирован (то есть `read` не вернёт управление) до тех пор, пока либо кто-нибудь не запишет данные в канал, либо все дескрипторы, открытые на запись в этот канал, не окажутся закрыты. Отметим, что, если в канале доступны для чтения данные (независимо от их количества, хотя бы один байт), функция `read` при попытке чтения из канала вернёт управление немедленно; если третий параметр `read` (количество запрашиваемых байтов) больше, чем на момент вызова оказалось доступно данных, прочитаны будут все доступные данные и `read` вернёт их количество, которое при этом будет меньше заказанного.

Попытки записи в канал, из которого никто не читает (при условии, что в системе имеются дескрипторы, связанные с противоположным концом канала), некоторое время будут успешными. Дело в том, что канал имеет внутренний буфер, размер которого зависит от реализации; так, в старых версиях ядра Linux он обычно составлял 4096 байт, а в современных его зачем-то расширили до 65536 байт. Когда буфер окажется заполнен, очередной вызов `write` заблокирует процесс до тех пор, пока кто-нибудь не начнёт из канала читать, тем самым освобождая место в буфере.

Рассмотрим теперь случаи, когда все дескрипторы, связанные с одним из концов канала, оказались закрыты. Ясно, что данный конкретный канал более никогда не удастся использовать, поскольку способа вновь связать дескриптор с одним из концов неименованного канала в системе нет. **Если оказались закрыты все дескрипторы, через которые можно было записывать данные в канал, операции чтения (вызовы `read`) сначала опустошат внутренний буфер канала, а затем будут возвращать `0`, сигнализируя о наступлении ситуации «конец файла».**

Если, наоборот, оказались закрыты все дескрипторы, через которые можно было из канала читать, то процесс, попытавшийся писать в канал, получит сигнал `SIGPIPE`. По умолчанию этот сигнал завершает процесс. Вызов `write` при этом возвращает `-1`, но это может быть обнаружено, только если процесс перехватывает или игнорирует сигнал `SIGPIPE`.

Наиболее активно неименованные каналы используются для реализации ***конвейеров*** — групп программ, которые запускаются на одновременное выполнение, причём информация, выдаваемая первой программой на стандартный вывод, поступает второй программе на стандартный ввод, вывод второй программы — на ввод третьей программе и т. д. (см. пример на стр. 94).

Для построения конвейера следует создать нужное число каналов (на один меньше, чем в конвейере будет элементов) и соответствующим образом связать потоки стандартного ввода и вывода (то есть дескрипторы `0` и `1`) в процессах, составляющих конвейер, с концами каналов. Очень важно при этом закрыть все лишние дескрипторы, связанные с данным каналом, во всех процессах, вовлечённых в решение задачи. Программы в ОС Unix часто пишутся так, чтобы работать до возникновения ситуации «конец файла» в потоке стандартного ввода; такая ситуация может возникнуть на канале только в случае, если *все* дескрипторы записи окажутся закрыты. Наличие лишнего открытого дескриптора записи нарушит нормальную работу конвейера.

После исчезновения процесса, для которого предназначены генерируемые программой данные, продолжать выполнение программы обычно бессмысленно. Если с исчезновением следующего элемента конвейера закроется *последний* дескриптор, открытый на чтение из канала, то пишущий процесс будет снят сигналом `SIGPIPE`. Если же где-то останется ещё хотя бы один открытый дескриптор для чтения, процесс будет просто заблокирован; возможно, это блокирует выполнение ещё каких-то задач, которые дожидаются завершения этого процесса.

Рассмотрим для примера конвейер

```
ls -lR  |  grep '^d'
```

Программа на Си, выполняющая те же действия (то есть запускающая те же программы с теми же параметрами и в таком же режиме взаимодействия), будет выглядеть так:

```
int main()
{
    int fd[2];
    int pid;
    pipe(fd);                       /* создаем канал для связи */
    pid = fork();
    if(pid == 0) {           /* процесс для выполнения ls -lR */
        close(fd[0]);            /* читать из канала не нужно */
        dup2(fd[1], 1);             /* станд. вывод - в канал */
        close(fd[1]);                /* fd[1] больше не нужен */
        execlp("ls", "ls", "-lR", NULL);  /* запускаем ls -lR */
        perror("ls");    /* не получилось, сообщаем об ошибке */
        exit(1);
    }
    pid = fork();
    if(pid == 0) {             /* процесс для выполнения grep */
        close(fd[1]);              /* писать в канал не нужно */
        dup2(fd[0], 0);            /* станд. ввод - из канала */
        close(fd[0]);                /* fd[0] больше не нужен */
        execlp("grep", "grep", "^d", NULL); /* запускаем grep */
        perror("grep");  /* не получилось, сообщаем об ошибке */
        exit(1);
    }
        /* в родительском процессе закрываем оба конца канала */
    close(fd[0]);
    close(fd[1]);
    wait(NULL);             /* дожидаемся завершения потомков */
    wait(NULL);
    return 0;
}
```

Каналы, создаваемые с помощью `pipe`, не позволяют организовать передачу данных между неродственными процессами. Этого недостатка лишены ***именованные*** каналы, которые в англоязычных источниках называются просто FIFO (*first in, first out*, т. е. «очередь»). Англоязычное название оправдывается тем, что байты, переданные в такой канал (как, впрочем, и в любой поток ввода-вывода, имеющий «другой конец») при передаче сохраняют свой порядок, так что байт, переданный раньше других, извлечён (прочитан) из канала будет также раньше других.

Именованные каналы подобны неименованным, с той разницей, что именованному каналу соответствует размещаемый в файловой системе файл специального типа; этот тип так и называется FIFO. К именованному каналу могут присоединяться процессы, не имеющие родственных связей; более того, закрытие всех дескрипторов, отвечающих за чтение из такого канала или за запись в такой канал, ещё не означает, что канал более не пригоден для работы, т. к. в любой момент такие дескрипторы могут появиться вновь.

Для создания файла типа FIFO используется функция `mkfifo`:

```
int mkfifo(const char *name, int permissions);
```

Первый параметр задаёт имя файла, второй — права доступа к нему (аналогично вызовам `open` и `mkdir`). Права, естественно, модифицируются параметром `umask`. Функция возвращает `-1` в случае ошибки, ноль — в случае успеха.

При создании файла FIFO система не создаёт сам объект канала в ядре; это происходит только тогда, когда какой-нибудь процесс открывает файл FIFO с помощью вызова `open` на чтение или запись; от того, в каком режиме (`O_RDONLY` или `O_WRONLY`) процесс попытается открыть файл типа FIFO, зависит, к какому концу канала будет присоединён полученный файловый дескриптор. Открывать файл FIFO в режиме `O_RDWR` не следует, несмотря на то, что в некоторых системах (включая Linux) это вполне сработает, то есть созданный файловый дескриптор будет ассоциирован с обоими концами канала; но так, во-первых, происходит не во всех существующих системах, и, во-вторых, само по себе такое действие выглядит несколько глупо.

Канал как объект ядра продолжает существовать до тех пор, пока существует хотя бы один связанный с ним дескриптор, после чего уничтожается. Уничтожение канала не означает уничтожения файла FIFO: после закрытия всех дескрипторов файл остаётся на месте и может быть снова открыт каким-нибудь процессом, после чего в ядре системы снова появится объект канала.

Прежде чем начать передачу данных, канал надо открыть с обоих концов. Обычно попытка открыть канал с одной из сторон блокируется до тех пор, пока кто-то не откроет второй конец канала; иначе говоря, вызов `open`, применённый к одному концу канала, не возвращает управление до тех пор, пока кто-то не откроет второй конец. Это поведение можно изменить, добавив во

втором параметре вызова `open` флаг `O_NONBLOCK`; в этом случае открытие на чтение пройдёт успешно без блокировки вне зависимости от того, открыт ли кем-нибудь противоположный конец канала, но вот попытка неблокирующего открытия на запись может привести к неудаче: если канал никем не открыт на чтение, то вызов `open`, выполненный с флагами `O_WRONLY|O_NONBLOCK`, вернёт `-1`, а в переменную `errno` будет записан код `ENXIO`.

Поведение именованного канала при закрытии последнего из дескрипторов, отвечающих за один из концов, полностью аналогично поведению неименованного канала в таких же случаях, то есть попытка читать из канала, у которого закрылся последний пишущий дескриптор, даёт ситуацию «конец файла», а попытка писать в канал, у которого закрылся последний читающий дескриптор, приводит к получению сигнала `SIGPIPE`. Отличие в том, что здесь оба случая не являются фатальными; так, после получения ситуации «конец файла», вообще говоря, возможно, что один из следующих вызовов `read` прочитает из того же потока какие-то данные. Это произойдёт, если какой-то другой процесс снова откроет тот же канал на запись. При этом всё время, пока ни одного пишущего дескриптора в системе нет, `read` будет продолжать возвращать `0`, сигнализируя о ситуации конца файла. С записью и `SIGPIPE` ситуация аналогична, хотя использовать это свойство на практике трудно.

5.3.16. Краткие сведения о трассировке

Трассировка обычно применяется при отладке программ. В режиме трассировки один процесс (отладчик) контролирует выполнение другого процесса (отлаживаемой программы), может остановить его, просмотреть и изменить содержимое его памяти, выполнить в пошаговом режиме, установить точки останова, продолжить выполнение до точки останова или до системного вызова и т. п.

В ОС Unix трассировка поддерживается системным вызовом `ptrace`:

```
int ptrace(int request, int pid, void *addr, void *data);
```

В качестве параметра `request` вызов получает одну из возможных команд (какие конкретно действия, связанные с трассировкой, требуются). Интерпретация остальных параметров зависит от команды.

Начать трассировку можно двумя способами: запустить трассируемую программу с начала или присоединиться к существующему (работающему) процессу. В первом случае отладчик делает `fork`, порождённый процесс сначала устанавливает режим отладки (вызывает `ptrace` с командой `PTRACE_TRACEME`), затем делает `exec` программы, подлежащей трассировке. Сразу после `exec` система останавливает трассируемый процесс и отправляет родительскому процессу (отладчику) сигнал `SIGCHLD`; собственно говоря, для этого и нужна команда `PTRACE_TRACEME`, она предписывает системе приостановить процесс после `exec`. Отладчик должен дождаться нужного момента с помощью вызовов семейства `wait`, которые в этом случае будут ожидать не окончания порождённого процесса, а его останова для трассировки. Далее отладчик может заставить отлаживаемый процесс выполнить один шаг с помощью команды `PTRACE_SINGLESTEP`, продолжить его выполнение с помощью `PTRACE_CONT`, узнать содержимое регистров с помощью `PTRACE_GETREGS` и т. п.

Для присоединения к существующему процессу используется вызов `ptrace` с командой `PTRACE_ATTACH`. При этом трассирующий во многих смыслах начинает выполнять роль родительского процесса по отношению к отлаживаемому; в частности, сигналы `SIGCHLD` посылаются теперь трассирующему, а не исходному родительскому процессу, хотя функция `getppid` в отлаживаемом процессе продолжает возвращать идентификатор настоящего родительского процесса.

Факт существования вызова `ptrace` и набор его основных возможностей (в особенности команду `PTRACE_ATTACH`) следует всегда иметь в виду при оценке мер, принимаемых для безопасности. Так, злоумышленник, получивший в вашей системе возможность запускать свои программы с полномочиями того или иного пользователя, с помощью `ptrace` сможет контролировать любые программы, запущенные от имени того же пользователя — даже если, например, собственные программы злоумышленника запускаются в какой-нибудь обособленной области дискового пространства, созданной вызовом `chroot` (см. стр. 112), что для начинающих сисадминов часто становится полной неожиданностью. В частности, если получить полномочия системного администратора, пусть даже внутри `chroot`, то подчинить себе всю систему — дело техники: достаточно помощью `ptrace` «присоединиться» к какому-нибудь процессу, работающему за пределами `chroot`, и заставить его сделать всё, что нужно злоумышленнику.

5.4. Терминал и сеанс работы

5.4.1. Драйвер терминала и дисциплина линии

Понятие ***терминала*** в unix-системах было и остаётся одним из центральных, в особенности когда речь идёт об управлении процессами. Изначально терминал представлял собой устройство, к которому подключаются клавиатура, экран и линия связи; текст, набираемый на клавиатуре, терминал отправлял в линию, а информацию, полученную «с того конца» линии, отображал на экране. Соединив линией связи два терминала, их операторы могли бы между собой пообщаться, но обычно между собой связывались не два терминала, а терминал и компьютер, причём один компьютер мог обслуживать несколько десятков терминалов одновременно; клавиатура терминала использовалась для набора команд командной строки, экран — для получения результатов.

Потоки ввода-вывода, связанные с терминалом, часто воспринимаются как некий особый случай, что и понятно: работа с терминалом подразумевает наличие «по ту сторону» живого пользователя. В частности, функции высокоуровневого ввода-вывода из стандартной библиотеки языка Си применяют для терминалов совершенно иную стратегию буферизации ввода-вывода, нежели для любых других потоков (см. §4.4.5). Более того, некоторые хорошо знакомые нам команды командной строки работают по-разному в зависимости от того, связан ли их поток вывода с терминалом или перенаправлен куда-то ещё; так, команда `ls` при работе с терминалом выстраивает имена файлов в несколько аккуратных колонок и (при соответствующих настройках) может сделать свой вывод разноцветным, вставив в него управляющие Esc-последовательности, тогда как при выводе в файл, канал или любой другой поток, не связанный с терминалом, та же команда выдаст все имена файлов в один столбик (собственно говоря, это будет просто список имён, разделённых символом перевода строки) и, конечно же, никаких спецсимволов, меняющих цвета, в свой вывод вставлять не будет. Чтобы увидеть эту разницу, перенаправьте вывод `ls` на вход программе `cat`, дав команду «`ls | cat`»; как известно, `cat` без параметров просто копирует стандартный ввод на стандартный вывод, но поскольку канал, связывающий эти две программы, терминалом не является, вы увидите выдачу команды `ls` в том варианте, который она применяет, работая не с терминалом.

Узнать, связан ли данный конкретный поток ввода-вывода с терминалом, позволяет функция `isatty`:

```
int isatty(int fd);
```

Параметром этой функции служит файловый дескриптор; если он связан с терминалом, функция возвращает `1`, в противном случае — `0`. Это библиотечная функция; работает она через знакомый нам системный вызов `ioctl` — пытается выполнить на переданном дескрипторе такую операцию, которая предусмотрена только для терминалов, и если это удаётся — считает, что имеет дело с терминалом.

Как мы уже обсуждали (см. т. 1, §1.2.1), алфавитно-цифровые терминалы в наше время не производятся, поскольку при необходимости с их ролью может справиться любой ноутбук, и это обойдётся намного дешевле специального устройства, выпускаемого малыми сериями; встретить аппаратные терминалы сегодня можно разве что в музеях и частных коллекциях. С логической точки зрения это ничего не меняет, функциональность терминала эмулируется программно, но терминал как сущность (пусть и виртуальная) при этом сохраняется практически в неизменном виде. Чаще всего в современных условиях отсутствует и линия связи; точнее говоря, она существует только в виде виртуального, программно эмулируемого объекта. Так, ядра Linux и FreeBSD эмулируют набор виртуальных терминалов, используя видеокарту и клавиатуру компьютера, на котором запущена система. Это те самые «консоли», доступные до запуска графической подсистемы; между ними можно переключаться нажатием комбинаций `Alt-F1`, `Alt-F2` и т. д., и если вам не требуется графика (например, для просмотра фотографий), то возможностей этих консолей вполне может хватить для работы с компьютером.

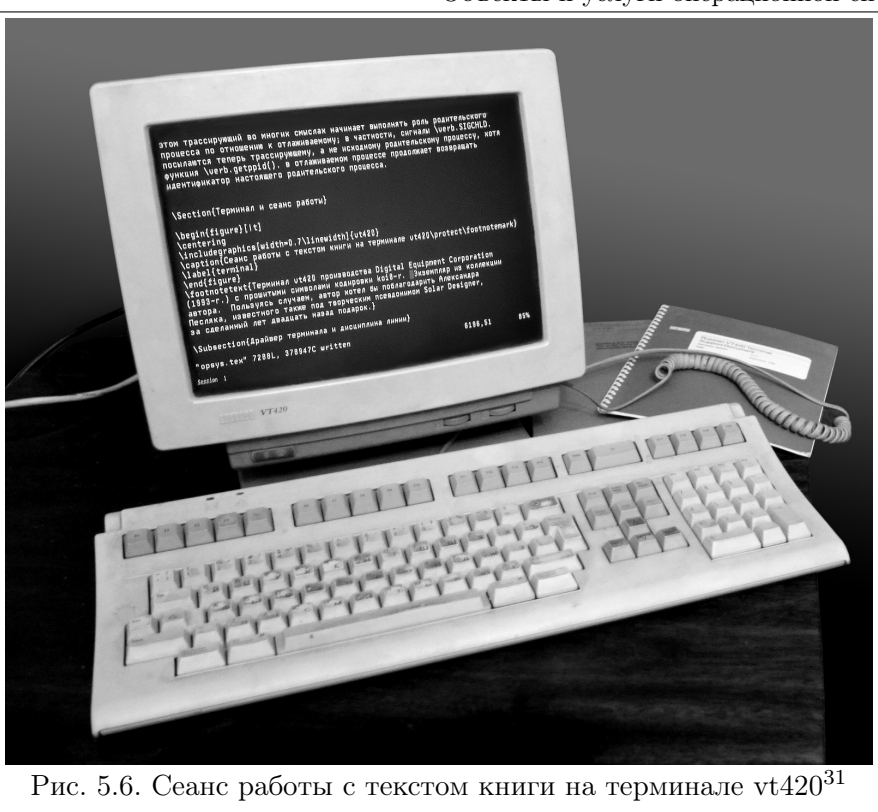

Рис. 5.6. Сеанс работы с текстом книги на терминале vt420[31]

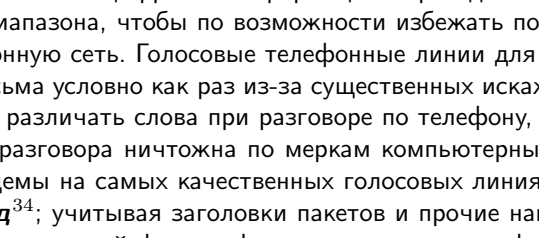

Рис. 5.7. Порт RS-232: разъём (слева[32]) и штекер (справа)

При работе с серверами системные администраторы часто предпочитают не запускать на них X Window — бóльшую часть работы выполняют в режиме удалённого доступа через сеть, а в тех редких случаях, когда с серверной машиной приходится иметь дело непосредственно, довольствуются текстовыми виртуальными консолями.

После запуска X Window виртуальные консоли никуда не деваются, перейти на них можно нажатием `Ctrl-Alt-F1`, `Ctrl-Alt-F2` и т. д, а вернуться в графику — нажатием `Alt-F7`, `Alt-F8` или чего-то вроде, конкретная комбинация зависит от конфигурации вашей системы (точнее, от того, какую виртуальную консоль использует X Window), но вместо них становится удобнее использовать специальные оконные приложения, эмулирующие терминал, такие как `xterm`, `rxvt`, `konsole` и т. п. В отличие от виртуальных консолей, где за эмуляцию терминала отвечает ядро системы, здесь функциональность терминала эмулирует обычная программа, выполняющаяся в виде процесса; но, как и в случае с виртуальными консолями, линия связи между терминалом и компьютером оказывается воображаемой.

Сравнительно редко возникает ситуация, требующая наличия реальной линии связи. Так, компьютеры, предназначенные для использования в качестве серверов, иногда выпускаются без видеокарты и порта для подключения клавиатуры; всё, что есть у такого компьютера для связи с внешним миром — это сетевые интерфейсы и консольный порт для подключения терминала. С аппаратной точки зрения это обычный COM-порт (известный также под названием RS-232), так что вместо терминала можно использовать обычный компьютер (чаще всего ноутбук), подключённый к такому порту трёхпроводным шнуром (раньше такой шнур называли «нуль-модемом»); на большинстве современных ноутбуков встроенного COM-порта уже нет, так что приходится применять переходник USB2COM, а функциональность терминала реализуется одной из существующих «терминальных программ», умеющих работать с COM-портом, таких как `minicom`, `hyperterm` или `termite`. Обычно через консольный порт выполняют только первичную настройку системы, а с того момента, когда начинает работать доступ к сети, дальнейшие действия производят удалённо.

Подробный рассказ о COM-портах мы опустим, но некоторые пояснения всё же будут уместны. Для передачи информации здесь используется три провода — по одному на каждое направление и ещё один в качестве общей «земли». Информация передаётся последовательно, то есть по одному биту; биты объединяются в группы (чаще всего — 8-битные байты, хотя это зависит от настроек), каждая группа снабжается служебными битами для обозначения начала и окончания передачи, а также, возможно, контрольным битом чётности, который служит для обнаружения ошибок, происходящих из-за электрических помех. Изначально COM-порты предназначались для передачи данных через аналоговые (голосовые) телефонные линии, для чего с помощью специальных устройств — ***модемов***[33] — цифровая информация переводилась в звуковой сигнал такого частотного диапазона, чтобы по возможности избежать потерь данных при передаче через телефонную сеть. Голосовые телефонные линии для передачи цифровых данных подходят весьма условно как раз из-за существенных искажений звукового сигнала; это не мешает различать слова при разговоре по телефону, но информационная ёмкость голосового разговора ничтожна по меркам компьютерных сетей. Самые лучшие аналоговые модемы на самых качественных голосовых линиях могли работать со скоростью 38400 ***бод***[34]; учитывая заголовки пакетов и прочие накладные расходы, на такой скорости передача одной фотографии с современного цифрового фотоаппарата занимает 15-20 минут, а то и больше, а на более низких скоростях, например, 9600 или 19200 бод, и вовсе растягивается на несколько часов, ну а о передаче видеофайлов можно было даже не думать. Несмотря на это, модемный доступ по коммутируемым телефонным линиям (диал-ап) на рубеже веков благодаря доступности именно этого вида линий связи был основным способом подключения к Интернету для широкой публики, его популярность пошла на спад лишь с появлением дешёвого и сравнительно быстрого доступа через сети мобильной связи.

Поскольку COM-порты предназначались в основном для подключения телефонных модемов, а скорость передачи данных через такие модемы никогда не была высокой, не было особого смысла бороться за увеличение скорости самих COM-портов. Большинство аппаратных реализаций COM-портов не поддерживает скорости выше, чем 115200 бод, что, конечно, очень мало для передачи данных при их современных объёмах, но при этом многократно превосходит потребности в скорости при подключении терминала. В самом деле, мало кто умеет набирать текст со скоростью, превышающей 400 знаков в минуту, а для передачи текста с такой скоростью хватило бы (даже с некоторым запасом) скорости в 100 бод. Текст, передаваемый на терминал со стороны компьютера, имеет существенно бóльшие объёмы, поскольку может включать элементы оформления, управляющие escape-последовательности и т. п., к тому же работающие программы часто генерируют потоки текста, из которых человек выхватывает только самые важные фрагменты; но и здесь скорость в 9600 бод обеспечивает вполне комфортные условия работы.

Отметим заодно, что при предоставлении удалённого доступа к машине также используется программная эмуляция терминала; в роли эмулятора выступает серверная программа (так называемый ***демон***), предоставляющая доступ — `sshd`, `telnetd` или их аналог.

Какова бы ни была природа терминального устройства, будь это физический терминал или эмулируемый (любым из имеющихся способов), между ним и программами, выполняющими операции ввода-вывода, находится специфический слой программного управления, называемый ***дисциплиной линии*** (этот термин — не слишком удачная калька английского *line discipline*, но более удачного или хотя бы устойчивого перевода так и не появилось). Чтобы понять, в чем заключается её роль, рассмотрим простейший пример — нажатие комбинации клавиш `Ctrl-C`. Мы знаем, что при этом активная программа получает сигнал `SIGINT`; вопрос в том, откуда этот сигнал берётся. Ясно, что обычный терминал — устройство, передающее и принимающее данные, — ничего не знает о сигналах ОС Unix и вообще может работать с разными операционными системами. Поэтому логично предположить, что сигнал генерирует сама система, получив от терминала некий специальный символ. Это действительно так: по нажатию `Ctrl-C` терминал просто генерирует символ с кодом 3. Кстати, это верно не только для `Ctrl-C`: `Ctrl-A` генерирует 1, `Ctrl-B` — 2 и т. д., комбинации `Ctrl-Z` соответствует код 26, далее идут

[31] Терминал vt420 производства Digital Equipment Corporation (1993 г.) с прошитыми символами кодировки koi8-r. Экземпляр из коллекции автора. Пользуясь случаем, автор хотел бы поблагодарить Александра Песляка, известного также под творческим псевдонимом Solar Designer, за сделанный в конце прошлого века подарок.

[32] Фото с сайта Викимедиа Коммонс; `https://commons.wikimedia.org/wiki/File:Serial_port.jpg`.

[33] Слово *modem* образовано как сокращение от *modulator and demodulator*, что буквальным образом отражает функцию этого устройства.

[34] *Бод* (англ. *baud*) — единица измерения скорости, на которой работает линия передачи информации; эта единица соответствует одному «символу» в секунду, при этом под символом понимается некая неделимая единица информации; для линий с двоичным кодированием, а почти все компьютерные каналы связи именно таковы, таким символом будет бит, но при измерении скорости COM-порта в итоговое число (в данном случае 38400) входят служебные биты, так что реальная скорость передачи (в битах в секунду) несколько меньше.

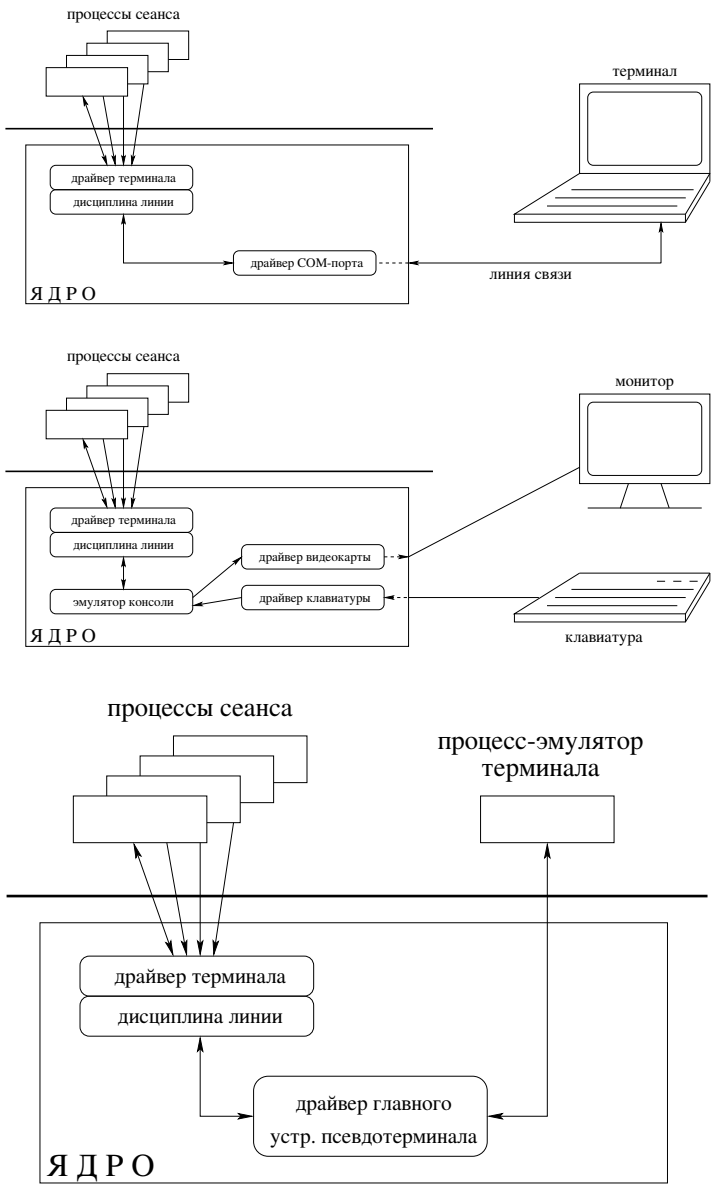

Рис. 5.8. Сеанс под управлением реального терминала (вверху), виртуальной консоли (в середине) и псевдотерминала (внизу)

`Ctrl-[`, `Ctrl-\` и `Ctrl-]`, порождающие коды 27, 28 и 29[35]. Получив символ с кодом 3, драйвер терминала, подчиняясь настройкам дисциплины линии, рассылает процессам, работающим под управлением данного терминала, сигнал `SIGINT`.

Точнее говоря, сигнал отправляется процессам текущей группы в сеансе, связанном с данным терминалом; разговор о группах процессов и сеансах у нас ещё впереди.

Функции дисциплины линии не ограничиваются рассылкой сигналов пользовательским процессам. Например, когда активная программа читает текстовую информацию, вводимую пользователем, терминал рассматривает некоторые коды символов как команды для редактирования вводимой строки: нажатие клавиши Backspace приводит к удалению последнего введённого символа, и то же самое произойдёт, если вы нажмёте `Ctrl-Shift-?` (и то, и другое генерирует символ с кодом `127`), нажатие `Ctrl-W` удаляет целиком последнее введённое слово, `Ctrl-R` приводит к тому, что вводимая строка повторно отображается на экране (этим можно воспользоваться, если вывод какой-то другой программы испортил изображение текущей строки), `Ctrl-U` очищает вводимую строку, так и не передав её активной программе. Всё это происходит без участия активной программы, все необходимые действия выполняет дисциплина линии.

Как мы вскоре увидим, дисциплину линии можно перепрограммировать, чтобы драйвер терминала отправлял `SIGINT` по какой-либо другой комбинации клавиш или не отправлял его вовсе; после такого перепрограммирования символ с кодом `3` будет просто передан работающей программе на стандартный ввод. Точно так же могут быть изменены или отменены особые роли символов, редактирующих вводимую строку, и этим пользуются «продвинутые» программы, чтобы сделать редактирование строки удобнее. Вы уже наверняка привыкли к возможности использовать клавиши «стрелок», автодописыванию команд и имён файлов, повторению ранее введённых команд «стрелкой вверх»; программы, предоставляющие все эти возможности, такие как командный интерпретатор `bash` и отладчик `gdb`, вынуждены делать всё это сами, поскольку дисциплина линии так не умеет; зато для всего этого имеются библиотеки, наиболее известная из них называется GNU Readline.

Рассмотрим ситуацию с программой `xterm`. Ясно, что при нажатии `Ctrl-C` получить `SIGINT` должна не сама программа `xterm`, а те процессы, которые запущены в её окошке. Собственно говоря, сама по себе программа `xterm`, будучи оконным приложением, и не получает никакого `SIGINT`, по крайней мере, когда активно именно её окно и мы нажали `Ctrl-C`. Вместо этого она получает ***клавиатурное событие*** от системы X Window, свидетельствующее о нажатии комбинации клавиш. Но генерировать сигнал для запущенных под её управлением процессов ей не нужно, достаточно передать символ с кодом `3` драйверу псевдотерминала, и драйвер (благодаря настройкам дисциплины линии) поступит так же, как если бы на месте программы был настоящий терминал — то есть перехватит символ и вместо него выдаст сигнал `SIGINT`.

Аналогично обстоят дела и при нажатии `Ctrl-D` (имитация конца файла). Программе `xterm` не нужно закрывать канал связи с активной программой, выполняющейся в её окошке, тем более что этого и нельзя делать, ведь сеанс одним `EOF`'ом не заканчивается, в нём могут быть и другие запущенные программы: представьте, что вы в окошке `xterm`'а запустили команду `cat`, потом нажали `Ctrl-D`, но вместо того, чтобы вернуться в командный интерпретатор, `xterm` закрылся; вряд ли вам такое понравится. Программа `xterm` действует иначе: она просто передаёт драйверу терминала символ, соответствующий комбинации `Ctrl-D`, то есть символ с кодом `4`. Получив его, драйвер терминала — та его часть, которую называют дисциплиной линии — обеспечит, чтобы ближайший вызов `read`, выполненный на этом логическом терминале, вернул `0`, то есть сигнализировал о ситуации «конец файла».

5.4.2. Сеансы и группы процессов

Для объединения процессов, запущенных пользователем в ходе работы с одного терминала, в unix-системах введено понятие ***сеанса***. Реализуется эта сущность очень просто: у каждого процесса есть ***идентификатор сеанса*** (`sid`, *session id*), а сам сеанс (буквально) состоит из процессов, имеющих одинаковый идентификатор сеанса, то есть сеанс представляет собой определённое множество процессов. Сеансу может принадлежать ***управляющий терминал***, не более чем один (возможно, ни одного); с другой стороны, терминал может служить управляющим для сеанса, но не более чем одного (или ни для одного не служить). Сеанс обычно создаётся в тот момент, когда пользователь вошёл в систему через виртуальную консоль или с использованием удалённого доступа (или с помощью настоящего терминала, хотя в наше время такое встречается редко). Менее очевидно, что новый сеанс создаётся, когда пользователь запускает программу, эмулирующую терминал, например, `xterm`; иначе говоря, при работе в оболочке X Window у вас будет по меньшей мере столько сеансов, сколько открыто окошек `xterm` и других подобных программ, плюс ещё один, созданный при запуске оконного менеджера.

В рамках одного сеанса процессы могут быть разбиты на отдельные ***группы***. Группа процессов входит в сеанс целиком, то есть в одну группу не могут входить процессы из разных сеансов. Минимум одна группа в сеансе всегда присутствует, она возникает при создании сеанса. В каждом сеансе, имеющем управляющий терминал, одна группа процессов называется ***текущей группой***[36], а остальные группы того же сеанса — ***фоновыми группами***[37]. Любая из групп сеанса, имеющего управляющий терминал, может быть в любой момент назначена текущей, тогда бывшая текущая становится фоновой. С некоторой натяжкой можно считать, что в сеансе, не имеющем управляющего терминала, все группы являются фоновыми, но на самом деле текущая группа — это, как мы вскоре увидим, свойство терминала, а не сеанса, так что в сеансе без терминала понятие текущей группы просто некому поддерживать.

Группы процессов изначально задуманы для объединения процессов, работающих над общей задачей. Командный интерпретатор обычно создаёт новую группу процессов для выполнения каждой поданной команды, так что все процессы, порождённые в ответ на каждую конкретную команду пользователя, оказываются объединены в одну группу; в частности, при запуске

[35] С этой возможностью терминала мы уже встречались, когда изучали Паскаль; см. т. I, §2.5.1.

[36] Английский оригинал этого термина — *foreground group*, что можно буквально перевести как «группа переднего плана».

[37] Англ. *background group*.

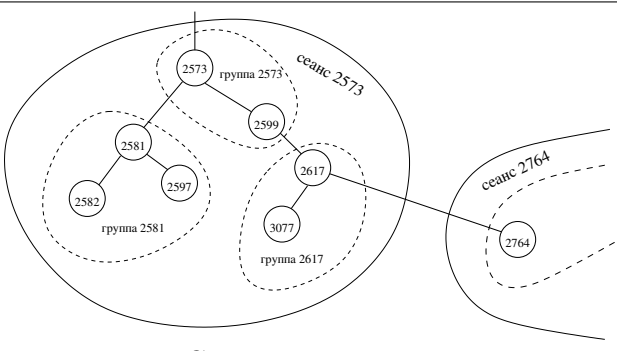

Рис. 5.9. Сеансы и группы процессов

конвейера все его элементы объединяются в группу. Группа процессов имеет свой идентификатор, который называют `pgid` (*process group id*).

Процесс наследует принадлежность к сеансу и группе от своего непосредственного предка — при вызове `fork` параметры `sid` и `pgid` не изменяются. Однако процесс может при желании уйти в новую группу или даже в новый сеанс. Говоря конкретнее, процесс может:

- создать сеанс; одновременно создаётся и группа, причём процесс, создавший сеанс, автоматически становится лидером и сеанса, и группы; идентификатор сеанса и идентификатор группы всегда равны идентификатору процесса, их создавшего;
- создать группу в рамках того же сеанса, при этом процесс становится лидером группы;
- перейти в другую группу того же сеанса.

Отметим ещё один момент: процесс не может создать новый сеанс, если он уже является лидером сеанса, то есть его `sid` совпадает с `pid`'ом, и не может создать новую группу, если уже является лидером группы, то есть его `pgid` совпадает с `pid`'ом. Это утверждение можно усилить: если процесс является лидером группы (или тем более лидером сеанса), то он не может создавать ни групп, ни сеансов. В самом деле, всякий лидер сеанса является также и лидером группы, возникшей в момент создания сеанса; как следствие, он не может создавать ни групп, ни сеансов, ведь идентификатор того и другого уже занят (им же самим). С другой стороны, лидер группы, не являющийся лидером сеанса, не может создать новый сеанс, ведь при этом должна возникнуть группа с тем же идентификатором, а этот идентификатор уже занят.

Все эти механизмы введены в ОС Unix для облегчения управления процессами. Например, при нажатии комбинации `Ctrl-C`: сигнал `SIGINT` получают только процессы текущей группы, а фоновые продолжают работу. Более того, сам по себе ввод с терминала разрешён только процессам текущей группы. Фоновый процесс при попытке чтения с терминала приостанавливается сигналом `SIGTTIN`. Можно (но не обязательно) запретить фоновым группам также и вывод на терминал. Если пользовательская сессия работы с терминалом по той или иной причине завершилась — например, пользователь выключил терминал, разорвал соединение удалённого доступа или закрыл окно программы `xterm`, то процессы текущей группы получат сигнал `SIGHUP`, причём если интерпретатор командной строки был в текущей группе (как правило, это означает, что он ждал ввода очередной команды), то он сам, получив `SIGHUP`, разошлёт этот же сигнал всем процессам своего сеанса; если текущей группой были другие процессы, интерпретатор поймёт, что они завершились по `SIGHUP`, и в этом случае тоже разошлёт всем остальным процессам сеанса `SIGHUP`; примечательно, что если интерпретатор завершить командой `exit` или нажатием `Ctrl-D` (создав ситуацию «конец файла»), рассылать `SIGHUP` он не станет, так что фоновые процессы, если такие в сеансе были, останутся работать. Вообще механизм сеансов и групп создан в основном для того, чтобы можно было установить, каких процессов непосредственно касаются действия, выполняемые пользователем терминала.

Рассмотрим кратко системные вызовы и функции, имеющие отношение к управлению сеансами и группами. Узнать для процесса параметры `sid` и `pgid` можно с помощью вызовов `getsid` и `getpgid`:

```
int getsid(int pid);
int getpgid(int pid);
```

где `pid` — идентификатор интересующего нас процесса. Специальное значение `0` означает вызывающий процесс. Отметим, что идентификатор сеанса совпадает с идентификатором (`pid`) процесса, создавшего этот сеанс; обычно идентификатор группы также совпадает с `pid`'ом создавшего группу процесса. Соответствующие процессы называются ***лидерами*** соответственно сеанса и группы.

Создание нового сеанса производится вызовом `setsid`:

```
int setsid();
```

Вызов не проходит, если данный процесс уже является лидером сеанса или хотя бы группы. Чтобы гарантировать успешное создание сеанса, можно вызвать `fork` и завершить родительский процесс, сменив, таким образом, свой `pid`:

```
pid = fork();
if(pid > 0)
    exit(0);
setsid();
```

При успешном выполнении `setsid` будут созданы одновременно новый сеанс и новая группа, идентификаторы которых будут совпадать с `pid`'ом процесса, выполнившего `setsid`, причём вызвавший процесс окажется единственным членом и того, и другого.

Как уже говорилось, каждый терминал может быть управляющим для не более чем одного сеанса и каждый сеанс может иметь не более одного управляющего терминала. При успешном выполнении `setsid` процесс теряет управляющий терминал, даже если у него остаются связанные с терминалом открытые дескрипторы. Чтобы снова получить управляющий терминал, процессу следует открыть вызовом `open` файл терминального устройства, не управляющего другим сеансом; интересно, что получить управляющий терминал может только процесс, создавший сеанс (лидер сеанса). Ясно, что лидер сеанса, не обладающего управляющим терминалом, должен соблюдать известную осторожность при открытии файлов, чтобы случайно не захватить себе свободный терминал, когда это не требуется; в частности, можно во всех вызовах `open` добавлять во втором параметре флаг `O_NOCTTY`, который блокирует превращение открываемого терминала в управляющий, даже если все остальные условия выполнены (вызывающий процесс является лидером сеанса, сеанс не обладает управляющим терминалом, открываемый файл является терминальным устройством, с которым не связан существующий сеанс).

Для смены группы предназначен вызов `setpgid`:

```
int setpgid(int pid, int pgid);
```

Параметр `pid` задаёт номер процесса, который нужно перевести в другую группу; `pgid` — номер этой группы. Сменить группу процесс может либо самому себе, либо своему непосредственному потомку, если только этот потомок ещё не выполнил вызов `execve`. Группу можно менять либо на уже существующую в рамках сеанса, либо можно задать параметр `pgid` равным параметру `pid`, в этом случае создаётся новая группа, а процесс становится её лидером.

Если у сеанса есть управляющий терминал, можно указать драйверу терминала, какую группу процессов считать текущей. Это делается с помощью библиотечной функции `tcsetpgrp`:

```
int tcsetpgrp(int fd, int pgrp);
```

Здесь `fd` — файловый дескриптор, который должен быть связан с управляющим терминалом; он нужен, потому что смена основной группы реализуется через вызов `ioctl` для терминала как логического устройства. Дело в том, что за понятие *текущей группы* в сеансе отвечает драйвер управляющего терминала, что и понятно, ведь именно ему нужно знать, каким процессам, например, разослать тот же `SIGINT`, если пользователь нажал `Ctrl-C`.

Здесь может возникнуть вполне законный вопрос, где, собственно говоря, такой дескриптор взять. Если вы по каким-то причинам уверены, что ваши стандартные потоки (или хотя бы какой-то один из них) не перенаправлены, можно использовать соответствующий дескриптор (`0`, `1` или `2`); вот только не вполне понятно, что́ может послужить основанием для такой уверенности. Можно, конечно, проверить, терминал ли там, с помощью функции `isatty`, и если да — *скорее всего* это ваш управляющий терминал; но, во-первых, это совершенно не факт, ведь процессу одного сеанса в общем случае никто

не мешает работать с терминалом другого сеанса, лишь бы прав хватило; а во-вторых, остаётся вопрос, что делать, если все три стандартных потока всё-таки окажутся перенаправлены.

Ключом к решению этой проблемы служит файл псевдоустройства `/dev/tty`; этот файл соответствует управляющему устройству сеанса вызывающего процесса, если у этого сеанса есть управляющий терминал; в противном случае попытка открыть этот файл закончится ошибкой. Итак, если вам потребовалось изменить текущую группу процессов в вашем сеансе, проще всего будет открыть файл `/dev/tty`, и если он успешно откроется, применить `tcsetpgrp` к полученному дескриптору, ну а если нет — считать, что управляющего терминала у вас нет, и, как следствие, в вашем сеансе никакой «текущей группы процессов» нет и быть не может, так что и менять нечего.

Остаётся ещё одна хитрость. Если процесс, обратившийся к `tcsetpgrp`, сам не входит на момент вызова в текущую группу, драйвер терминала, ничтоже сумняшеся, пришлёт сигнал `SIGTTOU` — причём сразу всей группе, в которую этот процесс входит. К счастью, это происходит только в случае, если процесс, вызвавший `tcsetpgrp`, не игнорирует и не блокирует сигнал `SIGTTOU`; достаточно ему (одному! всей группе это делать не требуется) установить для `SIGTTOU` диспозицию `SIG_IGN` (см. §5.3.14) или хотя бы заблокировать этот сигнал с помощью `sigprocmask`, и можно вызывать `tcsetpgrp` без опасений за судьбу своей группы.

5.4.3. Управление драйвером терминала

Из опыта нам уже известны некоторые особенности работы терминала. Так, обычно активная программа не получает символы пользовательского ввода до тех пор, пока пользователь не завершит ввод строки, после чего программе передаётся сразу вся строка; при вводе символы, набираемые на клавиатуре, отображаются на экране; нажатие определённых комбинаций клавиш приводит к специфическим эффектам и т. д.

Создавая полноэкранные программы с помощью паскалевского модуля `crt`, а позже — на Си с использованием библиотеки `ncurses`, мы много раз упоминали возможность *перепрограммирования* драйвера терминала. Попробуем разобраться, как это делается.

Как и для любого устройства, представленного в системе файлом специального типа, для управления терминалом — а точнее, для управления *линией связи с терминалом* — используется системный вызов `ioctl`. Для удобства в библиотеку языка Си включена целая группа функций под общим названием `termios`, специально предназначенных для управления терминалом; кстати, в эту группу входит упоминавшаяся в предыдущем параграфе функция `tcsetpgrp`. Создавался интерфейс `termios` ещё в те годы, когда аппаратные терминалы встречались чаще виртуальных, а поскольку подключались они в большинстве случаев через линии связи RS-232, функции модуля `termios` предоставляют целый ряд возможностей, которые сегодня могут показаться рудиментарными, как, например, установка скорости передачи данных. На самом деле все эти возможности по-прежнему используются в тех не столь уж редких случаях, когда к компьютеру всё-таки подключают физическую линию RS-232 или аналогичную (например, интерфейс RS-485 в наши дни довольно часто используется для управления промышленной автоматикой). На другом конце такой линии связи сейчас вряд ли окажется терминал, но ещё 20–25 лет назад именно терминалы были самым популярным видом оборудования, подключаемого к unix-машинам через RS-232, и это хорошо знают специалисты, работающие с последовательными интерфейсами, поскольку весь интерфейс функций `termios` построен в предположении, что скорее всего по ту сторону COM-порта находится терминал, а если нет — то это исключение из общего правила.

Взаимодействие с физическим последовательным портом — предмет достаточно увлекательный, но требуется в наши дни редко и в специфических случаях, поэтому рассматривать его мы не будем. Совсем другое дело — управление режимом работы терминала (дисциплиной линии). Конечно, при написании полноэкранных программ всё общение с драйвером терминала берут на себя соответствующие библиотеки, но иногда требуется решить совсем простенькую задачу вроде ввода пароля или какого-нибудь пин-кода (так, чтобы вводимые символы при этом не отображались на экране), и ради этого попросту глупо подключать к проекту монстров вроде `ncurses`.

При управлении терминалом следует помнить, что основных режимов работы терминала существует ровно два: ***канонический*** и ***неканонический***. Основное различие между ними в том, будет ли драйвер терминала накапливать вводимую строку символов, чтобы отдать её активной программе сразу всю, или же каждый введённый символ окажется доступен активной программе немедленно. Кроме основных режимов, драйвер терминала поддерживает несколько десятков менее значимых опций, многие из которых в современных условиях неприменимы, но некоторые (вроде отображения вводимых символов, *local echo*) достаточно важны и умение ими управлять может оказаться полезно. Наконец, драйвер терминала (точнее, дисциплина линии) позволяет полностью переназначить роли всех спецсимволов; например, при желании вы можете заставить дисциплину линии имитировать «конец файла» не по `Ctrl-D`, а по `Ctrl-N`, сигнал `SIGINT` посылать по нажатию клавиши Escape, удалять последний введённый символ по `Ctrl-B` и т. п. — иной вопрос, *нужно ли* так делать.

Все настройки дисциплины линии собраны в одну структуру `struct termios`; вы можете получить текущие настройки терминала с помощью функции `tcgetattr` и установить новые с помощью `tcsetattr`:

```
int tcgetattr(int fd, struct termios *tp);
int tcsetattr(int fd, int options, const struct termios *tp);
```

Первым параметром обе функции получают файловый дескриптор потока ввода-вывода, связанного с настраиваемым терминалом. Скорее всего, вам подойдёт дескриптор `0`, если только пользователь, запустивший вашу программу, не перенаправил ей поток стандартного ввода, но это можно проверить с помощью `isatty` (см. стр. 133). Если там окажется не терминал — скорее всего, пытаться что-то перепрограммировать бессмысленно, а продолжать ли работу вообще — зависит от решаемой задачи (например, программа `passwd` обычно при перенаправленном вводе работать отказывается).

При установке нового режима работы терминала с помощью `tcsetattr` нужно её вторым параметром (`options`) указать, в какой момент драйверу следует действительно перейти на новый режим работы. Здесь имеется три варианта: `TCSANOW` (применить новые настройки немедленно), `TCSADRAIN` (дождаться, пока все данные, записанные в поток вывода, связанный с `fd`, не будут физически переданы в линию связи) и `TCSAFLUSH` (дождаться, пока все записанные данные уйдут в линию, *сбросить все данные, пришедшие из линии, но пока не прочитанные*, и после этого применить новые настройки). В большинстве случаев годится `TCSANOW`.

Структура `termios` содержит как минимум следующие поля:

```
int c_iflag;
int c_oflag;
int c_cflag;
int c_lflag;
char c_cc[NCCS];
```

На самом деле полей больше, но функции, с которыми мы будем работать, их не используют. Кроме того, иногда может оказаться важно, что поля флагов в системных заголовочных файлах описаны как имеющие тип `tcflag_t`, а не `int`, а элементы массива `c_cc` имеют тип `cc_t`, а не `char`; например, в системе автора этих строк `tcflag_t` был определён как синоним для `unsigned int`, `cc_t` — как синоним `unsigned char`. В большинстве случаев это несущественно.

Поле `c_iflag` содержит флаги, отвечающие за преобразование входящего потока информации (буква `i` означает *input*), то есть информации, пришедшей со стороны терминала. Здесь единственный флаг, эффект от которого понятен и полезен — это `IXON`; если он установлен, то вывод на терминал можно в любой момент приостановить нажатием комбинации `Ctrl-S`, а затем продолжить нажатием `Ctrl-Q` (если, конечно, мы не переназначили коды для этих двух операций). Эта возможность полезна, если какая-то программа выдаёт текст на экран слишком быстро, а пользователю нужно внимательно прочитать фрагмент выдачи, который норовит исчезнуть за верхним краем экрана.

Остальные флаги из набора `c_iflag` выполняют не столь очевидные функции. Например, установка флага `IGNCR` позволяет игнорировать символы «возврат каретки» (`CR`, код `13`), т. е. если установить этот флаг, драйвер терминала перехватит все символы с кодом `13`, поступающие с терминала, и ни один из них активной программе не отдаст. В современных условиях этот флаг довольно бесполезен; он позволял работать с терминалами (естественно, аппаратными), которые при нажатии Enter передавали в линию связи два байта — возврат каретки и перевод строки. Флаг `ISTRIP` позволяет от

всех входящих символов «откусить» старший бит; между прочим, это сделает нечитаемым текст на русском языке в любой кодировке, кроме koi8-r (см. т. 1, §1.4.5), но вот как извлечь из этого практическую пользу, остаётся неясным. Подчеркнём, что эти два флага мы привели лишь для примера, чтобы создать некоторое впечатление относительно возможностей поля `c_iflag`. Назначение большинства других флагов из этого поля вообще невозможно понять без досконального анализа принципов работы с асинхронной последовательной линией связи, и вряд ли хотя бы один из них вам пригодится.

Флаги, содержащиеся в поле `c_oflag` (`o` от слова *output*), позволяют управлять обработкой потока, *выдаваемого* на терминал. К примеру, `ONLRET` запрещает вывод символа возврата каретки, то есть драйвер терминала «скушает» все символы с кодом `13`, выдаваемые активными программами, и в линию связи ни один из них не попадёт; флаг `OLCUC` все строчные латинские буквы превратит в заглавные; от остальных флагов из этого набора, пожалуй, пользы ещё меньше.

Среди флагов из поля `c_cflag` (`c` от слова *control*) вообще нет ни одного такого, который можно было бы описать в отрыве от тонкостей работы с физическими последовательными линиями связи и модемами. Когда линия связи с терминалом существует лишь в нашем воображении, ничего интересного с этими флагами сделать не удастся.

Совершенно иначе обстоят дела с полем `c_lflag` (`l` от слова *local*); всё самое интересное, что можно сделать с драйвером терминала, сосредоточено именно здесь. Пожалуй, самым «сильным» можно считать флаг `ICANON`: пока он взведён, терминал работает в каноническом режиме, то есть дожидается ввода строки целиком, прежде чем отдать её активной программе, обрабатывает специальные символы, предназначенные для удаления последнего символа и очистки всей строки (по умолчанию Backspace и `Ctrl-U`). Если добавить флаг `IEXTEN`, терминал будет обрабатывать также спецсимволы для удаления последнего введённого слова (`Ctrl-W`) и для восстановления (повторной печати) введённой части строки (`Ctrl-R`); следует учесть, что флаг `IEXTEN` без `ICANON` не работает.

Флаг `ISIG` указывает драйверу терминала, следует ли обрабатывать специальные символы, предполагающие отправку сигнала текущей группе процессов. Это хорошо знакомые нам комбинации `Ctrl-C`, `Ctrl-\` и `Ctrl-Z`, отправляющие соответственно сигналы `SIGINT`, `SIGQUIT` и `SIGTSTP`.

Довольно интересен флаг `TOSTOP`: он определяет, разрешено ли неактивным (фоновым) процессам выдавать данные на терминал. Если этот флаг сброшен, вывод на терминал могут осуществлять все процессы, у которых для этого достаточно полномочий; иногда из-за этого вывод разных программ создаёт путаницу. Если же флаг `TOSTOP` взвести, то попытка фонового процесса что-то напечатать приведёт к отправке всей его группе сигнала `SIGTTOU`; по умолчанию это приостановит работу процессов. Аналогичного флага для *ввода* данных с терминала не предусмотрено, делать это имеет право только текущая группа процессов, а фоновые группы при попытке что-то прочитать со своего управляющего терминала немедленно получают сигнал `SIGTTIN`, что также приводит к приостановке работы. Возобновить работу остановленных процессов можно сигналом `SIGCONT`.

Пожалуй, один из самых употребимых флагов здесь — `ECHO`; он определяет, следует ли выдавать на экран вводимые символы по мере их ввода (так называемый режим *local echo*). В каноническом режиме этот флаг обычно установлен, так что пользователь видит, что́ он вводит; но иногда желательно, чтобы какая-то информация вводилась «вслепую» — например, если это пароль. В такой ситуации мы можем сбросить флаг `ECHO` на время ввода секретного текста, а затем включить его обратно:

```
#include <stdio.h>                                  /* blindpsw.c */
#include <termios.h>
#include <unistd.h>  /* for isatty */
#include <string.h>  /* for memcpy */

enum { bufsize = 128 };

int main()
{
    struct termios ts1, ts2;
    char buf[bufsize];
    char *r;
    if(!isatty(0)) {
        fprintf(stderr, "Not a terminal, sorry\n");
        return 1;
    }
    tcgetattr(0, &ts1);         /* получаем текущие настройки */
    memcpy(&ts2, &ts1, sizeof(ts1));         /* создаём копию */
    ts1.c_lflag &= ~ECHO;             /* сбрасываем флаг ECHO */
    tcsetattr(0, TCSANOW, &ts1);      /* выключаем local echo */
    printf("Please blind-type the code: ");
    r = fgets(buf, sizeof(buf), stdin);
    if(!r) {
        fprintf(stderr, "Unexpected end of file\n");
        return 1;
    }
    printf("\nThe code you entered is [%s]\n", buf);
    tcsetattr(0, TCSANOW, &ts2); /* восстанавливаем настройки */
    return 0;
}
```

В этой программе мы сначала получаем текущие настройки терминала, затем модифицируем их и устанавливаем модифицированную версию, а позже восстанавливаем исходные настройки. Этот подход типичен при «хитрых» операциях с терминалом, поскольку заполнить структуру `termios` «с чистого листа» не так просто. На всякий случай напомним, что отдельные флаги, относящиеся к одному полю структуры, следует объединять операцией побитового «или» («`|`»). Установить в заданном поле заданные флаги можно с помощью операции «`|=`» (побитовое «или», совмещённое с присваиванием); например, установить несколько флагов в поле `c_lflag`, можно примерно так:

```
ts1.c_lflag |= (ICANON | IEXTEN | ISIG | ECHO);
```

Для принудительного сброса заданных флагов в заданном поле поступают следующим образом. Сформированную комбинацию флагов побитово инвертируют с помощью операции «`~`», так что биты, соответствующие нужным флагам, оказываются равны нулю, а все остальные — единице, и полученное значение используют в качестве «маски» для побитового «и»; обычно маску накладывают операцией «`&=`»:

```
ts1.c_lflag &= ~(ICANON | IEXTEN | ISIG | ECHO);
```

Именно так мы и поступили с флагом `ECHO` в примере, приведённом выше. На всякий случай подчеркнём, что все эти операции только изменяют поля вашей структуры, но никак не влияют на рабочие настройки терминала; чтобы их изменить, нужно вызвать `tcsetattr`.

Последнее документированное поле структуры `termios` называется `c_cc` (`cc` образовано от слов *control codes* — управляющие коды). Элементы этого массива имеют тип `char`; большинство из них (но не все!) содержат коды символов, которые дисциплина линии должна обрабатывать особым образом. Чтобы не нужно было помнить, какой из элементов массива содержит тот или иной код, заголовочные файлы определяют целый ряд констант, значения которых соответствуют индексам в массиве `c_cc`. Например, `c_cc[VEOF]` обычно содержит значение `4`, что соответствует использованию `Ctrl-D` для имитации ситуации «конец файла» (напомним, что комбинации вида `Ctrl-`X генерируют «символ» с кодом, соответствующим номеру буквы X в латинском алфавите; в частности, `Ctrl-D` генерирует код `4`). Изменив значение этого поля, можно заставить дисциплину линии использовать для имитации конца файла какой-то другой код — конечно, при условии, что в поле `c_lflag` установлен флаг `ICANON`.

Аналогично `c_cc[VINTR]`, `c_cc[VQUIT]` и `c_cc[VSUSP]` задают коды, используемые для порождения сигналов `SIGINT`, `SIGQUIT` и для приостановки активной задачи сигналом `SIGTSTP`; обычно в этих элементах содержатся коды `3`, `28` и `26`, соответствующие комбинациям `Ctrl-C`, `Ctrl-\` и `Ctrl-Z`. Естественно, эти комбинации работают только при установленном флаге `ISIG`. Коды из `c_cc[VSTOP]` и `c_cc[VSTART]` используются для приостановки и возобновления вывода на экран терминала (при установленном `IXON` в поле `c_iflag`), по умолчанию это `19` (`Ctrl-S`) и `17` (`Ctrl-Q`). Коды `c_cc[VERASE]`, `c_cc[VKILL]`, `c_cc[VWERASE]` и `c_cc[VREPRINT]` используются соответственно для удаления последнего символа, удаления всей текущей строки, удаления последнего слова и повторной печати вводимой строки. Первые два из них требуют только включённого флага `ICANON`, для двух остальных нужен ещё

и IEXTEN. По умолчанию это коды 127 (Backspace), 21 (Ctrl-U), 23 (Ctrl-W) и 18 (Ctrl-R).

Интересно, что код, используемый для перевода строки, перепрограммировать нельзя, это всегда 10. Присутствующие среди констант VEOL и VEOL2 обозначают элементы массива c_cc, задающие *дополнительные* символы для перевода строки; они обычно не задействуются.

Если обработку того или иного специального кода нужно отключить, в соответствующий элемент массива c_cc заносится ноль; например, если занести ноль в c_cc[VINTR], комбинация Ctrl-C работать перестанет, причём сигнал SIGINT не будет генерироваться вообще никакой комбинацией клавиш. Авторы POSIX для обозначения этого нуля ввели константу _POSIX_VDISABLE, но вероятность того, что когда-то где-то эта константа окажется отличной от нуля, пожалуй, не стоит рассмотрения.

Если отключить канонический режим — а его, как правило, отключают вместе с ISIG — то все вышеперечисленные коды работать перестанут; зато при этом в работу включатся c_cc[VMIN] и c_cc[VTIME], позволяющие управлять тем, когда (при каких условиях) будет возвращать управление вызов read в активной программе. Значение c_cc[VMIN] обозначает минимальное количество прочитанных из порта символов, а значение c_cc[VTIME] — время ожидания (в десятых долях секунды). В большинстве случаев c_cc[VMIN] устанавливают равным единице, а c_cc[VTIME] — равным нулю; при этом read будет работать так, как мы привыкли — немедленно возвращать управление, если доступен (введён на клавиатуре) хотя бы один байт, в противном случае блокироваться и ждать, пока этот байт не появится.

Второй достойный упоминания вариант — нули в обоих элементах; в этом случае read всегда будет возвращать управление немедленно, причём если ни одного байта прочитать не удалось, он вернёт 0. Заметим, что практически такого же эффекта можно достичь, переведя дескриптор ввода в неблокирующий режим с помощью fcntl (см. стр. 106), только при этом read будет возвращать -1, занося в errno значение EAGAIN.

Варианты с отличным от нуля c_cc[VTIME] обычно применяются при коммуникации по COM-порту с устройством, отличным от терминала. На всякий случай отметим, что семантика этого параметра зависит от значения c_cc[VMIN]: если там ноль, то установленный отрезок времени отсчитывается от входа в вызов read, и он возвращает управление не позднее чем через заданное время; если же в c_cc[VMIN] не ноль, отсчёт начинается после поступления первого символа и сбрасывается при поступлении каждого следующего, а возврат управления происходит, если превышен заданный лимит времени *между* двумя символами.

Читателю, заинтересованному в дополнительных подробностях, мы можем порекомендовать статью «The TTY demystified» [12], а также справочную страницу, доступную по команде man 3 termios.

5.4.4. По ту сторону псевдотерминала

Итак, линия связи с терминалом в современных условиях — вещь чаще всего сугубо виртуальная, а сам терминал эмулируется теми или иными программами, причём все случаи такой эмуляции можно разделить на два класса: виртуальные консоли, предоставляемые ядром операционной системы, и, собственно говоря, *всё остальное* — когда за эмуляцию терминала отвечает обычная пользовательская программа. Как система создаёт виртуальные консоли — внутреннее дело самой системы, а вот для эмуляции терминала пользовательскими программами предусмотрен довольно интересный механизм ***псевдотерминалов*** — специальных объектов ядра, имитирующих как раз ту самую воображаемую линию связи с терминалом.

Псевдотерминал как объект ядра имеет два двусторонних канала связи, один — для программы, эмулирующей функционирование терминала (например, xterm), второй — для программ, выполняющихся под управлением терминала. Программа, эмулирующая терминал, в этом виде взаимодействия называется *главной* (master), а работающие под управлением терминала — *подчинёнными* (slaves).

Чтобы создать псевдотерминал, главная программа должна вызвать специально предназначенную для этого функцию. Большинство версий стандартной библиотеки, следуя за glibc, предоставляют для этого функцию getpt:

```
int getpt();
```

— но в документации к ней сказано, что она специфична для glibc и в переносимых программах следует использовать функцию с корявым именем posix_openpt:

```
int posix_openpt(int flags);
```

В качестве параметра эта функция может принимать комбинацию флагов O_RDWR и O_NOCTTY (см. стр. 136), и обычно (а точнее, судя по всему, просто всегда) при вызове posix_openpt указывают оба этих флага: posix_openpt(O_RDWR|O_NOCTTY). Такой вызов эквивалентен вызову getpt без параметров, и в обоих случаях на самом деле происходит системный вызов open, применённый к специальному файлу виртуального устройства /dev/ptmx. При этом ядро создаёт пару связанных между собой устройств, которые называются ***главное устройство псевдотерминала*** (англ. *pseudoterminal master*, ptm) и ***подчинённое устройство псевдотерминала*** (*pseudoterminal slave*, pts). Первое из них ядро открывает в виде потока ввода-вывода и возвращает файловый дескриптор этого потока. Одновременно в файловой системе появляется файл подчинённого устройства, открытие которого позволит присоединиться к тому же псевдотерминалу уже со стороны программ, для которых он будет управляющим.

Подчинённый псевдотерминал сразу открыть не получится, поскольку для него ещё не настроены права доступа — во всяком случае, эти права могут не соответствовать задаче, которую пытается решить создатель псевдотерминала. Чтобы сделать подчинённое устройство доступным для открытия, нужно применить к *дескриптору главного устройства* (то есть тому дескриптору, который нам вернула функция getpt, posix_openpt или просто open, применённый к /dev/ptmx) последовательно функции grantpt и unlockpt:

```
int grantpt(int fd);
int unlockpt(int fd);
```

Первая из них изменяет принадлежность файла устройства подчинённого псевдотерминала так, что он становится доступен владельцу текущего процесса. Если программа, эмулирующая терминал, изначально работала с правами нужного пользователя (как, например, xterm), она может вызвать grantpt сразу после получения дескриптора ведущего терминального устройства. Часто бывает и так, что программа, создающая псевдотерминал, исходно работала в системе с полномочиями администратора (например, программа-сервер для удалённого доступа к системе); такие программы, убедившись, что имеют дело с определённым пользователем системы — например, проверив пароль или криптографическую подпись — отказываются от привилегий, установив себе uid, euid, gid, egid того пользователя, для которого запускается сеанс работы, и уже потом вызывают grantpt.

Функция unlockpt разрешает открыть файл псевдотерминала с помощью вызова open; до этого он недоступен к открытию. Прежде чем вызвать эту функцию, главная программа может установить права доступа к подчинённому псевдотерминалу с помощью вызова chmod. После выполнения unlockpt псевдотерминал готов к работе и его можно открыть с помощью open. Например, можно создать для этого отдельный процесс, там закрыть потоки стандартного ввода, вывода и ошибок, создать новый сеанс вызовом setsid, после чего открыть псевдотерминал и связать его дескриптор со всеми тремя потоками, а потом выполнить exec для запуска программы, которая будет лидером нового сеанса. Узнать имя файла устройства подчинённого псевдотерминала можно с помощью функции

```
char *ptsname(int master_fd);
```

где master_fd — дескриптор, полученный от getpt или posix_openpt.

Всю работу, связанную с созданием описанной связки главный-подчинённый, можно выполнить и проще, с помощью одной функции openpty:

```
int openpty(int *master, int *slave, char *name,
            const struct termios *termp,
            const struct winsize *winp);
```

Параметры master и slave задают адреса переменных, в которые следует записать дескрипторы, связанные соответственно с главным и подчинённым каналами связи с псевдотерминалом. Параметр name указывает на буфер, куда следует записать имя подчинённого псевдотерминала, параметры termp и winp задают режим работы псевдотерминала. В качестве любого из последних трёх параметров можно передать нулевой указатель.

Функция openpty считается особенностью систем семейства BSD; в литературе можно встретить рекомендации воздержаться от её использования.

5.4.5. Процессы-демоны

Под ***демонами***[38] понимаются процессы, не предназначенные для непосредственного взаимодействия с пользователями системы. Примерами демонов могут служить серверные программы, обслуживающие WWW или электронную почту. Существуют демоны и для выполнения внутрисистемных задач: так, демон `crond` позволяет автоматически запускать различные программы в заданные моменты времени, а демон системы печати собирает от пользователей задания на печать и отправляет их на принтеры. Демоны обычно расчитаны на длительное функционирование; в некоторых системах демоны могут годами работать без перезапуска. При старте демона принимаются меры, чтобы его функционирование не мешало работе и администрированию системы. Так, текущий каталог обычно меняется на корневой, чтобы не мешать системному администратору при необходимости удалять каталоги, монтировать и отмонтировать диски и т. п.

Хотя демону не нужен управляющий терминал (и вообще дескрипторы стандартного ввода, вывода и выдачи сообщений об ошибках), желательно, чтобы дескрипторы `0`, `1` и `2` оставались открыты, потому что демоны, естественно, работают с файлами, и если тот или иной файл будет открыт с номером дескриптора `0`, `1` или `2` (а это обязательно произойдёт, если дескрипторы просто закрыть), какая-нибудь процедура может случайно испортить файл, попытавшись выполнить ввод-вывод на стандартных дескрипторах. Поэтому все три стандартных дескриптора обычно связывают с устройством `/dev/null` (см. §5.2.5).

Чтобы действия, производимые с каким-либо терминалом, не сказались на функционировании демона (например, было бы нежелательно получить в некий момент `SIGHUP` из-за того, что пользователь прекратил работу с терминалом), он обычно работает в отдельном сеансе. Наконец, во избежание ошибочного получения управляющего терминала демон обычно делает лишний `fork`, чтобы перестать быть лидером сеанса. В целом процедура старта процесса-демона (так называемая «демонизация», англ. *daemonization*) может выглядеть примерно так:

```
close(0);
close(1);
close(2);
open("/dev/null", O_RDONLY); /* stdin  */
open("/dev/null", O_WRONLY); /* stdout */
open("/dev/null", O_WRONLY); /* stderr */
chdir("/");
pid = fork();
if(pid > 0)
    exit(0);
setsid();
pid = fork();
if(pid > 0)
    exit(0);
```

Обратите внимание, что первая связка `fork-exit` выполняется перед вызовом `setsid`, чтобы ничто не мешало процессу создавать новый сеанс. Если этого не сделать, имеется риск, что `setsid` не пройдёт: например, процесс может уже быть лидером сеанса или хотя бы группы. Вторая такая же связка выполняется уже после `setsid`, чтобы процесс, породив сеанс, перестал быть его лидером.

Поскольку с терминалами процесс-демон не связан, всевозможные сообщения об ошибках, предупреждения, информационные сообщения и т. п., адресованные системному администратору, приходится передавать неким альтернативным способом. Обычно это делается через инфраструктуру ***системной журнализации***. Для этого используются библиотечные функции `openlog`, `syslog` и `closelog`:

```
void openlog(const char *ident, int options, int facility);
void syslog(int priority, const char *format, ...);
void closelog();
```

Реализация этих функций зависит от конкретной системы. Например, демон, отвечающий за системную журнализацию, может держать открытый общедоступный сокет, в который функции журнализации будут выдавать свои сообщения.

Чтобы начать работу с системой журнализации, программа вызывает `openlog`, передавая первым параметром своё название или иную идентифицирующую строку, указывая некоторые дополнительные опции в параметре `options` (в большинстве случаев достаточно передать число `0`) и указывая, к какой подсистеме относится данная программа, через параметр `facility`. Наиболее популярные подсистемы, такие как почтовый сервер, имеют специальные значения этого параметра, прочим программам следует использовать константу `LOG_USER`.

Функция `syslog` похожа на функцию `printf`. Первым параметром указывается *степень важности* сообщения; например, `LOG_ERR` используется для сообщений о возникших ошибках, `LOG_WARNING` — для предупреждений, `LOG_INFO` — для простых информационных сообщений. Системный администратор может настроить систему журнализации так, чтобы в файлы журналов попадали только сообщения с уровнем важности не ниже определённого. Второй параметр — это форматная строка, аналогичная используемой в функции `printf`. Например, вызов функции `syslog` может выглядеть так:

```
syslog(LOG_INFO, "Daemon started, pid == %d", getpid());
```

Функция `closelog` завершает работу с системой журнализации, закрывая открытые файлы и т. п.

Управлять процессами-демонами можно, например, через сигналы; так, многие демоны в ответ на сигнал `SIGHUP` перечитывают конфигурационные файлы и при необходимости меняют режим работы. В более сложных случаях возможны и другие схемы управления.

[38] Слово «демон» в данном случае представляет собой прямой перевод английского *daemon*. Такой вариант перевода не совсем удачен, но другого всё равно нет.

Часть 6

Сети и протоколы

6.1. Компьютерные сети как явление

Строго говоря, ***компьютерной сетью*** считается любое соединение компьютеров (хотя бы двух), позволяющее передавать между ними данные. В истории известны примеры, когда с некоторой натяжкой сетевым соединением считалась даже организованная на регулярной основе передача файлов через внешние носители — попросту говоря, дискеты. Компьютеры, участвовавшие в таком «сетевом взаимодействии», вообще не были друг с другом соединены, но при этом пользователи этих «сетей» посылали друг другу электронную почту, участвовали в текстовых телеконференциях, обменивались файлами. Конечно, прибегать к обмену данными через дискеты приходилось только от недостатка технических возможностей по организации соединений между компьютерами; в наше время, когда цифровые сети связи доступны практически везде и всегда, пользователям нет нужды идти на подобные ухищрения, так что мы будем рассматривать только сети с соединениями.

6.1.1. Сети и сетевые соединения

Если рассмотреть все существующие сети связи безотносительно компьютерного контекста, то среди них можно выделить сети с фиксированными соединениями, сети коммутации соединений и сети коммутации пакетов. Фиксированные соединения в наше время встречаются редко; примером такой сети можно считать системы громкой связи на вокзалах и в аэропортах. Распространённые до недавнего времени у военных полевые телефоны, где для соединения двух аппаратов прокладывался телефонный кабель, тоже представляют собой сеть связи с фиксированным соединением. Суть этого вида сетей связи в том, что узлы соединяются друг с другом каналами, схема включения которых не предусматривает динамического изменения; любые поправки можно внести только вручную. Каждый абонент в такой сети постоянно связан с одним или несколькими другими, всегда одними и теми же.

Более гибкое устройство имеют сети коммутации соединений; хрестоматийный пример такой сети — хорошо известная нам сеть проводных («городских») телефонов. В сети коммутации соединений все конечные абоненты подключены к одному коммутатору либо (чаще) к одному из многих соединённых между собой коммутаторов, образующих *коммутирующую среду*. Посылая управляющие команды коммутаторам, абоненты таких сетей могут по своему усмотрению устанавливать соединения между собой и разрывать их.

В компьютерных сетях используется совершенно иной принцип работы — ***коммутация пакетов***. Следует отметить, что коммутация пакетов возможна только в сетях, рассчитанных на передачу цифровых данных. Информация, которую нужно передать в сеть, делится на ***пакеты***[1], каждый пакет снабжается адресом получателя, после чего оборудование, составляющее сеть связи, передаёт пакет нужному абоненту. По одному каналу связи могут за одну секунду пройти пакеты, адресованные миллионам разных участников сетевого взаимодействия. Внешне сеть коммутации пакетов напоминает сеть с фиксированными соединениями, поскольку физически каналы связи никто не переключает, но суть её, как видим, совершенно другая.

Если хорошо поискать, то можно найти действующие компьютерные сети, относящиеся как к сетям коммутации соединений, так и к сетям с фиксированными соединениями (и без всякой коммутации пакетов!), но рассматривать такие сети мы не будем: их области применения настолько специфичны, что вы вряд ли когда-нибудь с ними столкнётесь, а если всё-таки столкнётесь, то вам в любом случае придётся изучать документацию на конкретную сеть. Итак, **в компьютерных сетях, которые мы будем рассматривать, вся информация передаётся небольшими порциями (пакетами данных); каждый такой пакет содержит идентификатор (адрес) получателя, для которого он предназначен**; оборудование и программное обеспечение компьютерных сетей обеспечивает передачу каждого пакета по назначению, хотя и не даёт гарантии успеха такой передачи.

Существует множество разнообразных способов соединения компьютеров и всевозможных технологий передачи цифровой информации между ними. Чаще всего данные передаются по проводам в виде электрических сигналов, либо в виде радиосигналов — без всяких проводов; некоторые авторы заявляют, что наличие или отсутствие «материального» соединения (провода, кабеля и т. п.) представляет собой главный признак, по которому следует классифицировать сети. Мы позволим себе с этим не согласиться, ведь, например, технология передачи данных *световыми* сигналами через оптоволоконный кабель отличается от любой из множества технологий, подразумевающих электрические провода, ничуть не меньше, чем они, в свою очередь, отличаются от соединений через радиоволны; беспроводные соединения тоже не всегда основаны именно на радиоволнах, существуют методы передачи информации через инфракрасное излучение, через сфокусированные световые лучи и так далее. Исследуя вопрос об электрических соединениях, мы обнаружим широчайший ассортимент подходов, различающихся количеством проводов (от двух до нескольких десятков), применяемыми способами модуляции (представления цифровой информации электрическими сигналами), методами синхронизации, допустимыми типами кабелей и их длиной. Попытавшись вникнуть в построение радиоканалов для передачи компьютерной информации, мы и там обнаружим существенное технологическое разнообразие.

Самое интересное, что разнообразие обитателей всего этого зоопарка нас может не волновать: не всё ли равно, как именно происходит передача данных с одного компьютера на другой? Пользователи мобильных устройств иногда забывают, каким из двух способов — через сотовую сеть или через WiFi — они в настоящий момент соединены с Интернетом; пользователь домашнего фиксированного канала вполне может не иметь никакого понятия о разнице между ADSL, Ethernet и прочими видами физического соединения «последней мили».

Конечно, разные способы соединения могут обладать особенностями, которые придётся учитывать; так, если все компьютеры в вашей квартире или офисе соединены проводами, вы можете более-менее достоверно предполагать, что кто попало к вашей сети не подключится, и, например, разрешить сетевому принтеру принимать задания на печать от любых компьютеров, находящихся с ним в одном сегменте. С WiFi такой номер не проходит, даже если сеть защищена паролем: как правило, пароль там один на всех, то есть его заведомо знает больше одного человека, а раз так — можете считать, что пароль вообще не представляет собой никакого секрета, так что потенциально к вашей сетке может «присосаться» кто угодно, проходивший мимо по коридору или сидящий в соседнем помещении, а в некоторых случаях даже прохожий с улицы. В такой ситуации доступ к принтеру придётся как-то ограничить, если вы не хотите однажды обнаружить на полу груду бумаги, на которой напечатано непонятно что, а в принтере — пустой картридж. Но в целом разницу между технологиями физического уровня трудно считать определяющей для целей классификации сетей.

В учебных пособиях можно встретить долгие рассуждения о возможных видах *топологии* компьютерной сети; обычно при этом рассказывают про «шину», «звезду», «кольцо», «соединение каждого с каждым» и — в качестве белого флага — «смешанную топологию»; последнюю, как правило, описывают некими наукообразными эпитетами, смысл которых сводится к сакраментальному «как попало». Отметим, что топология «кольцо», непременно упоминаемая в таких случаях, в реальности не встречается, как и соединение каждого с каждым. Ещё интереснее обстоят дела с «шинами» и «звёздами»: одна и та же сеть может оказаться «звездой», если нарисовать схему соединения проводов, но при этом «шиной», если рассматривать логику взаимодействия компьютеров.

В первой половине 1990-х годов самым популярным способом соединения компьютеров можно было считать Ethernet, основанный на ***коаксиальном кабеле***[2]. В таком кабеле предусмотрены два проводника — проходящий по центру *сердечник* и отделённая от него слоем изоляции *оплётка*, служащая

[1] Английский оригинал этого термина — *packet*. Не следует путать термин «пакет», используемый в контексте передачи данных по сети, с термином, звучащим по-русски точно так же, но относящимся к типу планирования времени центрального процессора в многозадачных операционных системах; там оригинальный английский термин — *batch* (буквально «колода»).

[2] Слово «коаксиальный» происходит от *co-axis*; имеются в виду геометрические *оси* центральной жилы и оплётки, которые в таком кабеле совпадают. Здесь вполне подошло

Рис. 6.1. Коаксиальные соединения: кабель с соединителем, Т-коннектор, терминатор, узел для подключения компьютера в сборе

Рис. 6.2. Витая пара: кабель, концентратор, коммутатор

одновременно вторым проводом и экраном, предотвращающим потери на излучение и защищающим от электромагнитных помех. Коаксиальные кабели применяются в технике для передачи высокочастотных электрических сигналов — например, для подключения антенн к приёмопередающим устройствам; если у вас дома есть телевизор, принимающий эфирные телеканалы, то антенна к нему подведена коаксиальным кабелем, хотя и не такой конструкции, как у кабелей, применявшихся для соединения компьютеров. В области компьютерных сетей коаксиальный кабель практически вышел из употребления, но произошло это не так давно: в первой половине 2000-х годов сети на коаксиальном кабеле ещё довольно часто встречались.

Сетевые интерфейсы компьютеров подключались к коаксиальному кабелю параллельно с помощью специальных Т-образных соединителей, так называемых *Т-коннекторов*, так что сеть, основанная на коаксиальном кабеле, представляла собой *шину* в чистом виде. Скорость передачи данных по такой сети составляла 10 Мбит в секунду, что по современным меркам не так много, но основной недостаток коаксиальной сети состоял не в абсолютном значении скорости, а в неизбежности *коллизий*, когда один из участников взаимодействия начинает передачу данных, а другой участник, не успев на это среагировать, тоже начинает передачу и портит пакет данных, передаваемый первым участником. Чем плотнее загружена сеть, организованная в виде шины, тем выше вероятность коллизий и тем бóльшая часть пропускной способности сети на этих коллизиях теряется.

Был у коаксиального соединения и другой серьёзный недостаток. Коаксиальный кабель должен был представлять собой замкнутую среду передачи сигнала, завершающуюся с обеих сторон *терминаторами* — специальными насадками, в которых стояли резисторы по 50 Ом. Без терминатора сигнал, проходящий по кабелю, отражается от его оконечности; отражённый сигнал накладывается на основной сигнал и искажает его, так что сеть теряет работоспособность. Нетрудно догадаться, что такая сеть перестанет работать при любом обрыве кабеля или разъединении коннекторов, ведь в месте разрыва терминатору взяться неоткуда. По закону подлости обрывы и разъединения чаще всего происходили в запертой комнате, единственный ключ от которой находится у сотрудника, уехавшего в командировку. При этом нельзя было защитить кабель от случайных нарушений целостности, убрав его куда-нибудь в стену, короб и т. п., ведь каждый компьютер нужно было подключить непосредственно к кабелю[3]; иначе говоря, кабель приходилось прокладывать так, чтобы он проходил через каждое из рабочих мест, подключённых к сети, и окрестности всех этих рабочих мест неизбежно оказывались зоной риска.

Естественно, столь неудобному способу соединения машин между собой специалисты довольно быстро нашли замену — так называемую ***витую пару***. Связь здесь производится по четырём проводам, попарно перевитым между собой для избавления от электромагнитных помех (отсюда название); по одной паре происходит передача, по другой — приём информации. Два компьютера можно соединить кабелем витой пары, подключив пары, используемые для приёма и передачи, крест-накрест; по-английски кабель, специально предназначенный для такого соединения, называется *crossover*. Когда компьютеров больше двух, ситуация становится существенно более сложной: нужно специальное устройство, которое по-английски называется *hub*; русскоговорящие специалисты обычно называли это устройство попросту «хаб», используя прямую транслитерацию английского термина, но такое именование следует отнести к разряду профессионального жаргона; имеется устоявшийся официальный перевод термина *hub* на русский язык — ***концентратор***. Каждый компьютер локальной сети при таком способе её построения соединяется с концентратором (хабом) отдельным кабелем, подключаемым к так называемому *порту*; встречались концентраторы, имеющие от 3 до 24 портов. Сигнал, получаемый от любого из компьютеров (или, правильнее говоря, на любом из своих портов) по проводам, ответственным за передачу, концентратор транслирует всем остальным компьютерам, подключённым к нему (на все остальные порты), но уже через провода, ответственные за приём.

В наше время концентраторы считаются устаревшим видом оборудования, вместо них используются ***коммутаторы*** (англ. *switch*; чаще всего для обозначения этих устройств в русском языке используется жаргонизм «свитч»). Коммутатор отличается от концентратора тем, что транслирует получаемый пакет не всем компьютерам сети, а только тем из них, которым этот пакет предназначен (то есть чаще всего — только одному компьютеру). Это естественным образом снижает нагрузку на сеть: например, если в сети одновременно обмениваются информацией две пары компьютеров, то при использовании концентратора им придётся делить между собой общую пропускную способность сети, тогда как при работе через коммутатор компьютеры одной пары могут вообще не обращать внимания на существование другой пары.

Конечно, коммутатор технически представляет собой устройство намного более сложное, нежели концентратор, ведь ему приходится анализировать содержимое каждого передаваемого по сети пакета, чтобы узнать, кому этот пакет предназначается, и плюс к тому нужно помнить, к какому порту подключён какой компьютер; только так можно понять, через какой порт транслировать каждый отдельно взятый пакет. К тому же и концентраторы, и коммутаторы можно соединять между собой для получения сетей большого размера; с точки зрения коммутатора это значит, что к одному порту может быть подсоединён не один компьютер, а (в общем случае) произвольное их число, и приходится «помнить», на каком порту доступны какие компьютеры. Всё это приводит к необходимости использования процессора, памяти и программы, управляющей работой коммутатора, то есть технически коммутатор — это компьютер специального назначения. Наиболее «продвинутые» коммутаторы имеют в сети свой собственный адрес, как у обычного компьютера, и позволяют в режиме удалённого доступа управлять своей работой — например, задавать явным образом, какие компьютеры могут подключаться к отдельным портам, разбивать множество портов на группы, логически представляющие собой отдельные сети (англ. *VLAN*, *virtual local area network*; устоявшегося русского перевода пока нет) и т. п.

Благодаря совершенствованию производственных технологий в наше время простейшие коммутаторы — неуправляемые и имеющие обычно от 4 до 8 портов — стоят достаточно дёшево[4], что делает полностью бессмысленным использование концентраторов; дело в том, что корпус, разъёмы портов, блок питания, основная монтажная плата нужны как в коммутаторе, так и в концентраторе, а разница в сложности электроники (технически очень высокая, ведь концентратору не нужен ни процессор, ни память, ни программа) с производственной точки зрения почти не даёт различия в себестоимости изделия. Никто не отдаст предпочтения концентратору ради двух-трёх процентов денежной экономии.

В те годы, когда сети на основе витой пары только начинали развиваться, ситуация выглядела совершенно иначе. Идея коммутатора возникла, естественно, практически сразу в силу своей очевидности (даже раньше перехода с коаксиального кабеля на витую пару; известны коммутаторы для коаксиальных сегментов), но цены на первые такие устройства были совершенно заоблачными. Даже концентраторы («хабы») вплоть до конца 1990-х годов

бы русское слово «соосный», часто используемое в механике, но исторически утвердилась именно такая вот транслитерация английского *coaxial*.

[3] Спецификация коаксиального Ethernet'а допускала подключение сетевого интерфейса к основному кабелю с помощью отводного провода, но длина его не могла превышать 4 см; очевидно, это не решало никаких проблем.

[4] На момент написания текста первого издания книги в 2017 году стоимость самого дешёвого коммутатора составляла около 500 рублей; во время подготовки второго издания в конце 2020 года в магазинах удалось найти коммутатор меньше чем за 400 руб.; забавно, что почти столько же стоили банальные флеш-брелки на 8 GB — мéньшего из имеющихся размеров.

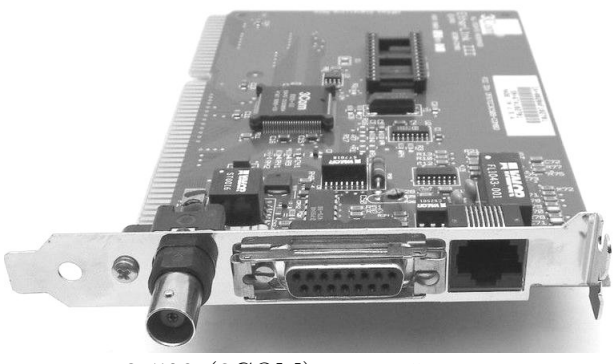

Рис. 6.3. Сетевая карта 3c509 (3COM) с разъёмами для подключения коаксиального кабеля (слева), внешнего адаптера (в центре) и кабеля витой пары (справа).

стоили достаточно дорого и вытеснить коаксиальный кабель попросту не могли: несмотря на все неудобства эксплуатации сетей на коаксиале, концентратор для таких сетей не требуется, и это обстоятельство оставалось во многих случаях определяющим.

Возвращаясь к вопросу о классификации сетей, отметим один интересный момент. Сети на основе коаксиального кабеля очевидным образом относятся к топологии «шина»; при переходе к сетям на витой паре, использующим концентраторы, топология физических соединений превращается в «звезду»: имеется центральный узел (концентратор), к которому подключены все остальные (в данном случае компьютеры). Но с точки зрения передачи информации при этом *ничего не меняется*, то есть, если можно так выразиться, «логическая» топология сети так и остаётся шиной! В самом деле, при передаче в сеть пакета данных этот пакет через концентратор получают *все* абоненты — точно так же, как это происходило в коаксиальной сети.

Неизменность «логической топологии» сети при переходе от коаксиальной шины к витой паре с концентратором лучше всего иллюстрирует такой факт. Существовало довольно много сетевых карт (печатных плат, вставляемых в компьютер и обеспечивающих аппаратное подключение компьютера к сети), имевших одновременно разъёмы для подсоединения витой пары и коаксиального кабеля (рис. 6.3). Зачастую администратор сети, устав от постоянных неполадок коаксиального кабеля, покупал концентратор и переводил свою сеть на витую пару. Для этого было достаточно переключить сетевые карты на работу с портом витой пары вместо коаксиального порта и заменить провода, а *программное обеспечение компьютеров никакой перенастройки не требовало*, то есть программы, использующие выход в сеть, вообще не замечали смены физического носителя соединения.

С учётом всего сказанного можно заключить, что топология физического соединения — это тоже совершенно не тот признак, который должен нас волновать при попытках классифицировать сети. Если цель, стоящая перед нами — научиться писать программы, то в нашу область интересов входят скорее такие особенности сетей, которые наша программа может обнаружить и которые, следовательно, необходимо учитывать при программировании.

Лет двадцать назад нас ещё могла бы с этой точки зрения заинтересовать разница между локальными, глобальными и прочими сетями в зависимости от их размера. Сейчас, когда фактически единственным используемым ***стеком протоколов***[5] остался TCP/IP, на котором основана работа сети Интернет, в большинстве случаев программам, работающим через сеть, оказывается безразлично, взаимодействуют ли они с компьютером, находящимся в другом полушарии Земли, в соседней комнате, на соседнем столе или даже их партнёр — другая программа, работающая на том же самом компьютере.

С повсеместным проникновением Интернета становится не всегда понятно, где кончается одна сеть и начинается другая; любая локальная сеть, если она имеет выход в Интернет, в определённом смысле представляет собой часть Интернета, причём в большинстве случаев можно выделить ещё несколько уровней, на каждом из которых найдётся компьютерная сеть, состоящая из других сетей и сама по себе входящая составной частью в более крупную сеть. Один из основных критериев отнесения компьютеров к единой сети базируется не на технических, а на организационных критериях: отдельной сетью считаются компьютеры, имеющие между собой соединения для передачи данных и при этом находящиеся под единой политикой управления, то есть, проще говоря, эксплуатируемые одной командой системных администраторов. Отметим, что, несмотря на нетехнический характер этого критерия, нам как будущим программистам он весьма интересен, поскольку программы, работающие внутри сети, и программы, работающие за её пределами, оказываются с нашей точки зрения в существенно различных категориях: первым в некоторых случаях можно доверять (хотя и с ограничениями), вторым же доверять a priori нельзя, ведь мы не знаем и не можем знать, кто и с какой целью их написал и запустил. Далеко не все пользователи компьютерных сетей соответствуют нашим представлениям об ангелах с крылышками. Конечно, злоумышленник может оказаться и внутри периметра нашей сети, и такую возможность тоже приходится учитывать, но это всё же скорее чрезвычайное происшествие, тогда как наличие злоумышленников в Интернете «по ту сторону» нашего периметра — это обычная рабочая обстановка.

Возвращаясь к техническим аспектам построения компьютерных сетей, отметим, что простейшая (наименьшая) компьютерная сеть — это такая сеть, в которой имеется всего одно сетевое соединение; такую сеть при всём желании не получится разделить на сети меньшего размера. И именно здесь мы неожиданно натыкаемся на самый, пожалуй, интересный (с точки зрения программистов и сетевых администраторов) критерий классификации, относящийся, правда, не к сетям как таковым, а как раз к отдельно взятым сетевым соединениям. Дело в том, что, как мы уже видели, одно сетевое соединение может быть рассчитано на достаточно большое количество компьютеров — во всяком случае, более чем на два. Именно так работали и работают сети Ethernet, и совершенно не важно, используется ли для соединения компьютеров коаксиальный кабель (теоретически можно себе представить, что где-то коаксиальные сети ещё работают) или витая пара, применяются ли концентраторы или коммутаторы. Любые два компьютера, подключённые к одной локальной сети Ethernet, могут устанавливать между собой сетевые соединения и обмениваться информацией, не прибегая к помощи посредников; в сетях на коаксиальном кабеле такие посредники действительно отсутствовали, в сетях на витой паре посредники, формально говоря, есть — это так называемое активное оборудование в виде «хабов» или «свитчей», но участники сетевого взаимодействия их не видят и могут никак не учитывать их существование.

С другой стороны, достаточно часто нам приходится иметь дело с сетевыми соединениями, в которых задействовано только два компьютера. Такое соединение можно построить, естественно, на основе технологий семейства Ethernet (выше мы уже упоминали вариант соединения двух компьютеров кабелем витой пары без использования активного оборудования), но чаще такие соединения возникают при использовании технологий, по самой своей сути не предполагающих появления «третьего лишнего». Именно так в большинстве случаев организованы «нелокальные» каналы передачи данных; иначе говоря, если вы рассмотрите любой цифровой канал, выходящий за пределы отдельного здания (например, канал, связывающий ваш компьютер или вашу локальную сеть с провайдером доступа в Интернет), то с хорошей вероятностью этот канал окажется соединением вида «точка-точка» — именно так называют сетевые соединения «на двоих» (соответствующий англоязычный термин *point-to-point* обычно сокращают до *PtP*).

Различие между двухточечным и многоточечным вариантами сетевого соединения оказывается неожиданно принципиальным. При передаче данных через соединение вида «точка-точка» сразу понятно, кому предназначены передаваемые данные; но если данные передаются через сетевой интерфейс, к которому непосредственно подключено (кроме нас самих) больше одного потенциального получателя информации, нам приходится уточнять, кому конкретно предназначен тот или иной пакет данных. Как мы увидим позже, признаки, по которым отдельные сетевые интерфейсы идентифицируют друг друга при работе через многоточечное соединение, отличаются от адресов, используемых взаимодействующими по сети программами для идентификации своих партнёров по соединениям; для установления соответствия одних адресов другим в многоточечных соединениях используется специальный протокол, в котором нет надобности при работе через соединение «точка-точка».

Отметим ещё один момент. Мы уже встречались с понятием *сетевой карты*; так вот, строго говоря, её наличие не обязательно для работы с компьютерной сетью, поскольку передачу данных между двумя компьютерами можно организовать через имеющиеся порты едва ли не любого вида. Так, в предыдущей части книги при обсуждении терминалов мы упоминали по-

[5] Будучи центральным в области компьютерных сетей, понятие протокола вызывает неожиданные сложности у многих новичков. Мы подробно рассмотрим это понятие чуть позже, а пока ограничимся замечанием, что *протокол* — это некий набор соглашений, которым следуют участники сетевого взаимодействия, чтобы понять друг друга.

следовательные порты (COM-порты) и модемы для подключения к телефонным линиям; этот вид оборудования вышел из употребления не так давно, а в конце 1990-х и начале 2000-х годов модемное соединение оставалось основным способом доступа в Интернет для большинства пользователей. Если же два компьютера находятся физически недалеко друг от друга — например, в соседних комнатах,— то можно соединить их COM-порты шнуром из трёх проводов, который обычно называли «нуль-модемом», и организовать через него сетевое соединение, единственным недостатком которого будет низкая скорость передачи данных — как правило, не выше 115200 ***бод*** (см. стр. 134), что почти в сто раз медленнее старых типов Ethernet-сетей и в тысячу раз медленнее, чем наиболее распространённые в наше время 100-мегабитные соединения. Существует программное обеспечение для организации сетевого соединения через порт USB, хотя при этом на одном из двух соединяемых компьютеров должен присутствовать «ведомый» порт, что для обычных компьютеров редкость, но совершенно не редкость, например, для планшетников. В те времена, когда оборудование Ethernet стоило дорого, но практически каждый компьютер был оснащён так называемым параллельным портом, использовавшимся для подключения принтера, можно было встретить сетевые соединения с использованием параллельных портов. Шина SCSI, часто встречающаяся на серверных компьютерах и в основном предназначенная для подключения внешних устройств хранения данных — жёстких дисков, CD, ленточных накопителей (стриммеров) и т. п. — физически позволяет связать два компьютера, и энтузиасты немедленно воспользовались этим для организации сетевого соединения (например, IP over SCSI).

Даже аппаратура, специально предназначенная для соединения с компьютерной сетью, будь то Ethernet или WiFi, в наше время далеко не всегда имеет вид карты. Большинство современных компьютеров имеет порт для подключения витой пары на материнской плате, если же говорить о ноутбуках и нетбуках, то там — опять-таки, на материнской плате — обычно имеется также и WiFi-адаптер. По правде говоря, со схематической точки зрения это всё те же сетевые карты, просто смонтированные непосредственно на основной плате, а не в виде платы-расширения; но бывают ведь ещё и адаптеры Ethernet и WiFi, подключаемые к USB-порту, и тут уже ни о какой «карте» речи не идёт.

С учётом всего сказанного явно требуется какой-то другой термин, и мы его сейчас введём. **Аппаратное устройство (в составе компьютера или подключённое к нему), используемое для соединения с одним или несколькими другими компьютерами для передачи данных, обычно называют *сетевым интерфейсом***.

Сразу же предостережём читателя от неправильного восприятия сказанного. Фразу, выделенную в предыдущем абзаце, не следует воспринимать как *определение* сетевого интерфейса, поскольку во многих контекстах тот же термин имеет несколько иной смысл. Так, часто встречаются *виртуальные сетевые интерфейсы*, которые с точки зрения программного обеспечения практически ничем не отличаются от обычных (физических), но при этом в физическом смысле не существуют; такие интерфейсы возникают при создании *туннелей* и *виртуальных частных сетей* (VPN'ов, от английского *virtual private network*). Кроме того, один физический сетевой интерфейс может в некоторых случаях быть настроен как *транковый*; в этом случае с программной точки зрения мы увидим не один интерфейс, а несколько, каждый из которых участвует в своём сетевом соединении[6], хотя с аппаратной точки зрения интерфейс останется один. Как обычно, наша задача сейчас не в том, чтобы дать строгое определение термина, а в том, чтобы понять, о чём идёт речь.

6.1.2. Шлюзы и маршрутизация

Отдельное сетевое соединение может объединять несколько десятков, а иногда даже пару сотен компьютеров, но наращивать это количество дальше оказывается нецелесообразно, причём как технически, так и организационно. В частности, с технической точки зрения коммутаторы, образующие одну локальную сеть, не допускают циклических соединений, то есть от каждого компьютера к каждому другому в графе соединений должен быть ровно один путь; это значит, что выход из строя любого элемента инфраструктуры нарушит работу сети, а обеспечивать бесперебойное функционирование кабелей (в том числе соединяющих между собой коммутаторы) тем труднее, чем больше их общая длина. Кроме того, чем больше к сети подключено компьютеров, тем больше в ней должно быть коммутаторов и тем труднее спроектировать сеть так, чтобы бóльшая часть информации передавалась между компьютерами, соединёнными одним коммутатором; как следствие, с ростом числа компьютеров в одной сети растёт и нагрузка на кабели, соединяющие между собой отдельные коммутаторы, так что рано или поздно их пропускной способности становится недостаточно. Наконец, в едином сетевом соединении в определённых случаях неизбежно использование широковещательных сообщений — таких, которые должен получить и как минимум проанализировать каждый из компьютеров; чем больше компьютеров в локальной сети, тем больше будет таких пакетов.

С организационной точки зрения слишком большие локальные сети тоже порождают определённые проблемы. В рамках одного сетевого соединения, как правило, невозможно ввести никакие ограничения по видам и количеству информации, передаваемой между компьютерами: все доступны всем без ограничений. Из этого вытекает необходимость мер по недопущению в сеть злоумышленников, и чем сеть больше, тем эти меры сложнее; если в одну локальную сеть (одно сетевое соединение) объединить компьютеры разных организаций, обеспечить безопасность в такой сети будет нереально.

Обратим теперь внимание на то, что на отдельно взятый компьютер никто не мешает поставить больше одного сетевого интерфейса; как следствие, один и тот же компьютер может одновременно быть участником нескольких сетевых соединений. С такой ситуацией вполне можно столкнуться в сугубо бытовых условиях, ведь большинство современных ноутбуков имеют в своём составе как обычный сетевой интерфейс, рассчитанный на подключение витой пары, так и беспроводной адаптер для работы с WiFi-сетями. Конфигурация систем, предназначенных для конечного пользователя, обычно построена так, что из двух (и более) сетевых интерфейсов компьютер реально использует только один, даже когда из них активно больше одного, но такое ограничение не имеет никаких технических причин.

Ситуация становится интереснее, если компьютер, участвующий в нескольких сетевых соединениях одновременно, настроить так, чтобы он мог перебрасывать между ними пакеты. Такой компьютер обычно называют ***сетевым шлюзом*** или ***маршрутизатором***; часто можно услышать жаргонное слово *роутер*, транслитерированное с английского *router*[7]. Полезно помнить, что английский оригинал словосочетания «сетевой шлюз» звучит как *network gateway*, что в разговорной речи часто превращается в краткое «*gate*» (буквально «ворота»; читается «гейт»). В настройках, имеющих отношение к компьютерным сетям, часто встречается аббревиатура «GW», означающая, как нетрудно догадаться, «gateway»; если в какой-нибудь сети вам попался компьютер, имеющий такое сетевое имя — скорее всего, именно через него происходит связь с внешним миром.

С хорошей степенью вероятности читатель уже сталкивался с маршрутизаторами. Если ваш дом или квартира подключены к Интернету, то в наши дни в большинстве случаев линия связи, которую к вам привёл провайдер, подключается к устройству, которое затем «раздаёт» подключение к Интернету через WiFi, а при желании позволяет также подключить несколько компьютеров проводами. Это невзрачное устройство, выглядящее как небольшая коробочка, как раз и есть маршрутизатор. Пусть вас не обманывает его скромный вид: внутри это полноценный компьютер со своей операционной системой, причём в роли этой системы довольно часто выступает Linux, хотя со стороны этого может быть и не видно.

Будучи частью нескольких сетей одновременно, компьютер-шлюз служит своеобразным «пропускным пунктом» для пакетов, направляющихся из одной сети в другую (отсюда название). В простейшем случае шлюз просто принимает из одной сети пакеты, адресованные компьютерам другой сети, и передаёт их по назначению; конечно, при этом все компьютеры в сетях, имеющих соединение через шлюз, должны знать о существовании компьютеров, доступных через шлюз, а также и о самом шлюзе. В реальной жизни большинство шлюзов не ограничивается простым перебрасыванием пакетов туда-сюда: на компьютере, выполняющем роль шлюза, можно настроить разнообразные фильтры, пропускающие только часть пакетов, можно ограничить количество передаваемых пакетов, можно организовать всевозможный

[6] Такой режим работы интерфейса становится возможен благодаря использованию «продвинутых» коммутаторов, которые позволяют сгруппировать их порты в две или более виртуальные Ethernet-сети. Один из портов может быть объявлен как *транковый*, то есть обслуживающий несколько виртуальных интерфейсов в рамках одного физического, причём каждый из виртуальных интерфейсов оказывается «подключён» к своей виртуальной сети.

[7] Буквально слово *route* означает «маршрут»; поскольку словоформа «маршрутер» по-русски звучит несколько диковато, слово «маршрутизатор» следует считать буквальным переводом английского слова *router*.

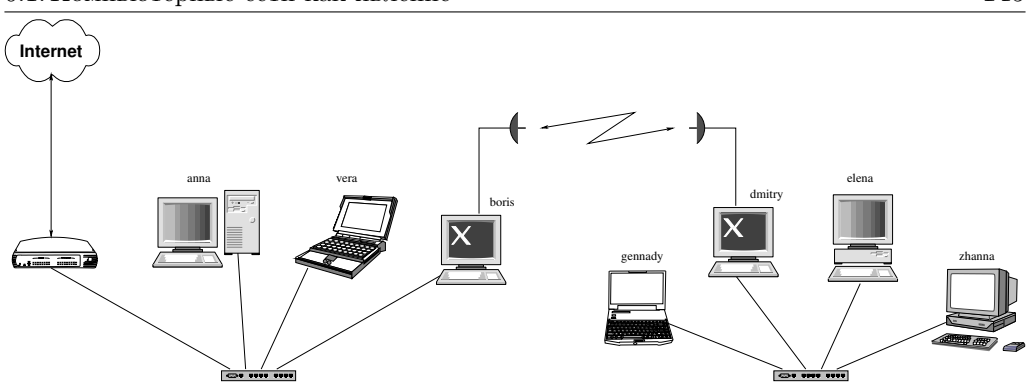

Рис. 6.4. Пример компьютерной сети

мониторинг того, какая информация в какое время проходила через шлюз, и т. д. Достаточно часто шлюзы даже изменяют содержащуюся в пакетах адресную информацию. Кроме того, в одном сетевом соединении могут участвовать несколько шлюзов, так что пакетам из одной сети, чтобы достичь другой сети, придётся совершить путешествие от шлюза к шлюзу, прежде чем они попадут куда нужно. Схемы передачи пакетов через шлюзы как раз и называют «маршрутами», отсюда второе название для шлюзов — «маршрутизаторы».

Попробуем проиллюстрировать сказанное. Пусть в некоем доме (представьте себе для наглядности, что это коттедж в коттеджном посёлке) живут Анна, Борис и Вера; их родственные отношения нас совершенно не волнуют, так что можете сами придумать, кем они друг другу приходятся. У каждого из них есть свой компьютер, причём эти компьютеры соединены в локальную сеть — например, витой парой через коммутатор. В другом доме того же посёлка живут Геннадий, Дмитрий, Елена и Жанна, у них тоже есть компьютеры, и эти компьютеры объединены в свою локальную сеть.

В некий момент Борис и Дмитрий решают, что им было бы интересно связать свои компьютеры, например, чтобы играть в многопользовательские игры. Кроме того, Борис вспоминает, что у Геннадия есть большая коллекция музыки и фильмов в виде файлового архива и Геннадий с удовольствием делится своими файлами со всеми знакомыми[8], но при отсутствии связи между сетями таскать файлы приходится на флешках, что не очень удобно. Жанна, узнав о планах Бориса и Дмитрия, заявляет, что Вера — её лучшая подруга и они хотят общаться между собой, используя микрофоны и видеокамеры. После этого становится очевидно, что обитателям обоих домов просто необходима связь между их локальными сетями.

Если бы все компьютеры находились в одном здании, можно было бы просто соединить проводом коммутаторы и, возможно, поменять сетевые адреса некоторых компьютеров, чтобы исключить совпадения; в результате вместо двух разных локальных сетей наши персонажи обнаружили бы себя в одной большой сети. Но тут в дело вступают вполне ожидаемые технические и организационные соображения. С технической точки зрения кабель витой пары по улице лучше не прокладывать: при первой же грозе есть риск лишиться компьютеров, да и чисто механически протянуть провод не всегда возможно — например, если дома расположены не рядом, а через улицу или между ними находятся чужие участки. Объединению сетей могут помешать и сугубо организационные причины: например, Елена когда-то успела в дым рассориться с Анной и заявляет, что не желает позволять «всей её семейке» даже потенциально иметь какой бы то ни было доступ к своему компьютеру.

Взвесив все имеющиеся соображения, Борис и Дмитрий решают купить пару узконаправленных антенн, установить сетевые интерфейсы для WiFi и создать между своими компьютерами радиоканал, а сами компьютеры настроить как шлюзы: компьютер Бориса будет передавать на компьютер Дмитрия через радиоканал пакеты, адресованные компьютерам самого Дмитрия, а также Геннадия, Елены и Жанны, тогда как компьютер Дмитрия будет передавать по радиоканалу на компьютер Бориса пакеты, предназначенные для компьютеров Анны, Веры и самого Бориса.

После этого наши друзья приступают к настройке компьютеров своих домочадцев. На компьютерах Анны и Веры указывается, что пакеты для компьютеров Геннадия, Дмитрия, Елены и Жанны нужно передавать через компьютер Бориса; компьютеры Геннадия и Жанны настраиваются на передачу пакетов для компьютеров Анны, Бориса и Веры через компьютер Дмитрия. Елена от аналогичной настройки наотрез отказывается; следуя её настойчивым требованиям, Дмитрий на своём компьютере устанавливает специальное правило для шлюза, чтобы пакеты, пришедшие извне (через сетевой интерфейс радиоканала) и адресованные компьютеру Елены, сбрасывались, никуда не передаваясь. После этого Геннадий у себя на компьютере запускает сервер игры Battle For Wesnoth, и наши персонажи — все, кроме Елены — проводят вечер, разыгрывая сражения между эльфами и орками; Елена могла бы принять участие в игре, поскольку сервер находится в одной сети с её компьютером, но она играть не хочет и вообще недовольна происходящим.

До сих пор мы никак не упоминали связь с Интернетом; предположим, до сего момента компьютеры наших друзей вообще не имели выхода в Интернет — хотя, конечно, такое предположение в наши дни выглядит несколько странно, но в качестве иллюстрации наш пример более интересен, если всю инфраструктуру приходится строить именно «с нуля». Пусть теперь Борис заключает договор с провайдером и к нему в дом вводят канал для связи с Интернетом, который, как водится, заканчивается маршрутизатором. Здесь можно не изобретать лишних сущностей, а просто подключить этот маршрутизатор к уже имеющемуся в доме коммутатору, сделав его ещё одним (четвёртым) компьютером в локальной сети. На компьютерах Анны, Бориса и Веры добавляется ещё одно маршрутное указание: пакеты, предназначенные для «всех остальных» компьютеров (читай — для всех компьютеров Интернета, за исключением тех, про которые есть другие указания) направлять на новый маршрутизатор. Такое правило маршрутизации называется «шлюз по умолчанию». К примеру, на компьютерах Анны и Веры теперь по три правила маршрутизации: пакеты для компьютеров своей локальной сети отдавать в локальное соединение (без указания шлюза), пакеты для локальной сети соседей отдавать на компьютер Бориса, а пакеты для всех остальных компьютеров — на маршрутизатор от провайдера.

Дмитрий предлагает Борису поучаствовать в оплате интернет-канала, после чего наши друзья добавляют несложные настройки: на компьютере Дмитрия шлюзом по умолчанию объявляется компьютер Бориса, на компьютерах Геннадия, Елены и Жанны — компьютер Дмитрия. Кстати, ранее прописанное у Геннадия и Жанны правило насчёт соседской локальной сети теперь можно убрать, ведь для них шлюзом выступает тот же самый компьютер Дмитрия. А вот с компьютером Елены имеется проблема: ограничение, которое по её просьбе ввёл Дмитрий, не позволяет ей выходить в Интернет, ведь пакеты, предназначенные для её компьютера, фильтруются. Подумав, Дмитрий модифицирует фильтр так, чтобы из пакетов, адресованных компьютеру Елены, отфильтровывались только те, в которых обратным адресом обозначен какой-нибудь из адресов локальной сети соседей.

Для нашего примера возможны и более сложные варианты. Например, Дмитрий тоже мог бы заключить договор с провайдером (желательно — с другим, не тем, что у Бориса) и иметь свой собственный канал в Интернет, но при этом наши друзья могли бы настроить маршрутизацию так, чтобы при выходе из строя одного из каналов обе локальные сети начинали бы использовать то соединение с Интернетом, которое осталось рабочим; кроме того, в какой-то момент обоим «администрам-любителям» могло бы прийти в голову, что использовать в качестве маршрутизаторов личные компьютеры не очень удобно (например, их приходится всегда держать включёнными и нельзя лишний раз перезагрузить), и в обеих локальных сетях появились бы выделенные маршрутизаторы — либо специализированные, либо в виде обычных компьютеров, на которые в таком случае можно возложить функции серверов, и так далее. Позже мы подробно разберём, как выглядят сетевые адреса компьютеров (точнее, их сетевых интерфейсов) при использовании протоколов TCP/IP и как могли бы выглядеть адреса и маршрутные правила на компьютерах из нашего примера. Узнаем мы и о том, что́ обычно происходит внутри маршрутизатора, соединяющего локальную сеть с Интернетом, почему компьютеры локальной сети из Интернета не видны и как получается, что локальные сетевые адреса не нужно менять при переходе на другой интернет-канал; но всему своё время.

[8] Если у вас на этом месте возникли какие-нибудь странные мысли, связанные с так называемым авторским правом, вам стоит обратить внимание, что (во всяком случае, на момент написания этого текста) в России создание копий в цифровой форме для личного использования является абсолютно законным, если только копируемые файлы не являются программами для ЭВМ; см. ст. 1273 ГК РФ. Кроме того, настоятельно рекомендуем вернуться к первому тому нашей книги и перечитать самое первое («философское») предисловие.

6.2. Сетевые протоколы

6.2.1. Понятие протокола и модель OSI

Под ***протоколом обмена*** (или, для краткости, просто «протоколом») понимается набор соглашений, которым должны следовать участники обмена информацией[9], чтобы понять друг друга. При любом осмысленном взаимодействии по компьютерной сети задействуется сразу несколько протоколов, относящихся к разным *уровням*. Так, сетевая карта, через которую наш компьютер подключён к локальной сети, следует протоколу, фиксирующему правила перевода цифровых данных в аналоговый сигнал, передающийся по проводам, и обратно. Одновременно запущенный нами браузер связывается с сайтом в сети Internet, используя транспортный протокол TCP. Сервер и браузер обмениваются информацией, используя протокол HTTP (*hypertext transfer protocol*).

Понятие протокола почему-то вызывает у значительного числа новичков неожиданные сложности, и мы попробуем ещё раз заострить внимание на том, что́ же такое «протокол». Говоря, что протокол есть набор соглашений, мы имеем в виду *буквально* именно это. Иначе говоря, протокол — это некий *текст*, написанный на естественном языке (в применении к компьютерам язык обычно оказывается английским). Этот текст содержит указания для разработчиков программного обеспечения, а в некоторых случаях — и аппаратуры: *что нужно сделать, чтобы разработанные вами программы или устройства могли взаимодействовать с другими программами/устройствами, декларирующими поддержку данного протокола.*

Широко известна «стандартная» ***модель ISO OSI***, предполагающая разделение всех сетевых протоколов на семь уровней. ISO расшифровывается как International Standard Organization (организация, утвердившая соответствующий стандарт), OSI означает Open Systems Interconnection (буквально переводится как «взаимосоединение открытых систем», но обычно при переводе используется слово «взаимодействие»). Модель включает семь уровней:

- **физический** — соглашения об использовании физического соединения между машинами, включая количество проводов в кабеле, частоту и другие характеристики сигнала и т. п.;
- **канальный** (в оригинале *datalink*) — соглашения о том, как будет использоваться физическая среда для передачи данных, включая, например, размеры пакетов и способы коррекции ошибок;
- **сетевой** — соглашения о том, как данные будут передаваться по сети от шлюза к шлюзу; именно на этом уровне определяется, как выглядят сетевые адреса и как настраивается маршрутизация;
- **транспортный**; пакеты, передаваемые по сети с помощью протоколов сетевого уровня, обычно ограничены в размерах и, кроме того, могут доставляться не в том порядке, в котором были отправлены, теряться или, наоборот, дублироваться (приходить в двух и более экземплярах); обычно прикладным программам требуется более высокий уровень сервиса, обеспечивающий надёжность доставки данных и простоту работы, и за это как раз отвечают протоколы транспортного уровня; реализующие их программы сами следят за доставкой пакетов, отправляя и анализируя соответствующие подтверждения, нумеруют пакеты и расставляют их в нужном порядке после получения;
- **сеансовый** — определяет порядок проведения сеанса связи, очерёдность запросов и т. п.;
- **представительный**; на этом уровне определяются правила представления данных, в частности, кодировка, способы представления двоичных данных текстом и т. п.;
- **прикладной**; протоколы этого уровня определяют, как прикладные программы будут использовать сеть для решения конкретных задач конечного пользователя.

Для упрощения запоминания английских названий уровней модели ISO OSI существует мнемоническая фраза «All People Seem To Need Data Processing» («всем людям, похоже, нужна обработка данных»). Первые буквы слов этой фразы соответствуют первым буквам названий уровней: Application, Presentation, Session, Transport, Network, Datalink и Physical. Аналогичной русской фразы автору, к сожалению, не встречалось. В реальной жизни модель ISO OSI не используется; существовавшие когда-то буквальные её реализации, поддерживавшие ровно семь слоёв, распространения не получили. Для нас эта модель представляет интерес скорее как иллюстрация того, как *может быть* организовано взаимодействие между программами, работающими на разных компьютерах.

Специалисты иногда шутят, что модель OSI получилась семиуровневой, потому что в соответствующем комитете образовалось семь подкомитетов и каждый предложил что-то своё. В каждой шутке есть доля шутки, но всё остальное — чистая правда; модель OSI служит прекрасной иллюстрацией того, сколь «полезны» результаты деятельности комитетов, в особенности когда речь идёт о технической стандартизации. Отметим, что набор протоколов TCP/IP, на котором построена сеть Интернет и который в итоге вытеснил все остальные протоколы из активного употребления, был создан узкой группой людей по принципу ad hoc, то есть чтобы решить задачи, вставшие здесь и сейчас. Никаких комитетов в создании TCP/IP задействовано не было.

6.2.2. Физические и канальные протоколы

Хотя в явном виде семиуровневая модель OSI не используется, можно, проанализировав обычный сеанс повседневной работы с Интернетом, обнаружить задействованные в этом сеансе протоколы, *приблизительно* соответствующие уровням OSI. Начнём мы, естественно, с самого «низа» — физического уровня. Какой из существующих физических протоколов используется в вашем сеансе работы — зависит от способа, которым ваш компьютер[10] подключается к Интернету. Самый простой способ — обычная локальная сеть на основе витой пары; соответствующий протокол определяется спецификацией, известной под именем 100BASE-TX, она же известна как «стандарт IEEE 802.3u». Протокол описывает, в частности, какие кабели, штекеры и разъёмы должны использоваться для соединения, каковы должны быть технические характеристики проводов, какую следует использовать частоту для передачи сигнала по проводам, какое электрическое напряжение, как конкретно будут кодироваться двоичные цифры электрическими сигналами. Надо сказать, что спецификация 100BASE-TX этим не ограничивается, она определяет ещё и то, какой длины будут отдельные передаваемые порции информации («фреймы», или кадры), какие в этих фреймах предусмотрены служебные данные, включая аппаратный идентификатор (mac-адрес) сетевого интерфейса, для которого предназначен конкретный фрейм. В модели OSI это уже следующий, «канальный» уровень.

Если для подключения к сети вы используете WiFi, протокол физического уровня оказывается сложнее, ведь используется радио, а эфир, в отличие от проводов, один на всех. Приходится учитывать, что в одном месте могут работать разные сети WiFi, и они не должны друг другу мешать; в разных странах действуют различные правила использования радиочастот, и те же самые частоты могут использоваться устройствами, не имеющими отношения к WiFi. Спецификация WiFi предполагает использование 14 различных частотных каналов, в каждом из которых производится деление на временны́е срезы (*time slices*); выбор каналов и распределение временных срезов между разными сетевыми соединениями производится автоматически, без участия пользователя — так, чтобы обеспечить, насколько это возможно, успешную работу всех соединений, оказавшихся в зоне действия друг друга (возможно, принадлежащих совершенно разным сетям). Принципы и правила взаимодействия WiFi-устройств между собой описаны в наборе спецификаций, известных под общим названием IEEE 802.11. Эти спецификации, как и Ethernet, полностью покрывают физический уровень модели ISO OSI и часть канального уровня.

Ещё сложнее обстоят дела, если вы выходите в Интернет через сеть сотовой телефонной связи; здесь роль физических и канальных протоколов исполняют спецификации GSM, включающие специальные протоколы для передачи цифровых данных, такие как GPRS и более новые. Спецификации GSM-сетей затрагивают не только физический и канальный, но и сетевой уровень.

Протоколы физического уровня бывают не только сложными, но и совсем простыми; например, если вам придёт в голову соединить два компьютера нуль-модемным кабелем через COM-порты, то в роли протокола физического уровня будет выступать RS232, спецификация которого достаточно проста даже для «самопальной» реализации радиолюбителями. К сожалению, из-за

[9] Это не обязательно должны быть компьютеры. Скажем, азбука Морзе также является своего рода протоколом; более сложный пример некомпьютерного протокола — правила радиообмена между пилотами самолетов и авиадиспетчерами.

[10] Каково бы ни было устройство, с помощью которого вы просматриваете странички в Интернете — обычный настольный компьютер, ноутбук, планшет, мобильный телефон (смартфон) или даже «умные часы» — строго говоря, всё это компьютеры; мы не делаем между ними различий, обсуждая работу в Сети.

низкой скорости передачи данных RS232 в наше время вряд ли можно считать актуальным, когда речь идёт об организации компьютерной сети.

Поднимаясь на следующий уровень, канальный, мы обнаружим, что проблемы, которые по замыслу создателей OSI должны решаться протоколами этого уровня, частично решены спецификациями Ethernet и WiFi, но остаётся ещё кое-что. Сетевые интерфейсы, участвующие в одном многоточечном соединении, опознают друг друга по так называемым mac-адресам — уникальным шестибайтовым идентификаторам, которые приписываются каждому сетевому адаптеру прямо на заводе-изготовителе. Например, на одном из старых компьютеров автора этих строк основная сетевая карта имела mac-адрес `00:1F:C6:65:42:48`, причём первые три байта (`00:1F:C6`) здесь представляют собой префикс, закреплённый за компанией-производителем (ASUS), а остальные байты назначает производитель, свои для каждого экземпляра устройства, так что получается что-то вроде серийного номера.

Не все существующие mac-адреса имеют такую структуру; существуют блоки с префиксами длиной 28, 32 и 36 бит. Подробности читатель может самостоятельно отыскать в Интернете, хотя в большинстве случаев эта информация интересна только членам комитетов и прочим бессмысленным бюрократам; технической ценности она не представляет.

Современные операционные системы позволяют отключить функции сетевой карты, отвечающие за формирование и распознавание Ethernet-кадров, содержащих именно такой mac-адрес. Пользователь может вручную задать произвольные шесть байт, которые и будут использоваться в качестве mac-адреса, причём формировать кадры будет не сетевой адаптер, а его драйвер, находящийся в ядре операционной системы. В стационарных сетях в большинстве случаев в этом нет нужды, а mac-адрес, зашитый в адаптер при его изготовлении, нас вполне устраивает; однако для мобильных устройств, подключающихся к точкам доступа по Wi-Fi, фиксированный mac-адрес становится инструментом отслеживания их перемещений, так что операционные системы предпочитают видимый MAC-адрес генерировать случайным образом.

Так или иначе, адреса, используемые для сетевых интерфейсов при настройке сети — так называемые ip-адреса, по которым, в частности, производится маршрутизация пакетов — с mac-адресами не имеют ничего общего, но при этом сетевые адаптеры должны знать, кто из них отвечает за какой сетевой адрес, чтобы в рамках многоточечного сетевого соединения отправлять пакеты кому следует, а не кому попало. Для автоматического построения таблицы соответствия ip-адресов и mac-адресов служит протокол ARP (*address resolution protocol*). Когда операционной системе требуется отправить пакет через многоточечное сетевое соединение, драйвер сетевого адаптера должен указать mac-адрес получателя, но система может его ещё не знать. В этом случае как раз и задействуется протокол ARP. Через сетевое соединение отправляется широковещательный (то есть предназначенный для всех, кто здесь есть) пакет, содержащий оформленный по правилам ARP запрос вида «у кого здесь такой-то ip-адрес?». Если кто-то из участников соединения опознаёт этот ip-адрес как свой, он формирует (опять-таки в соответствии с соглашениями ARP) ответный кадр, означающий «такой-то ip-адрес у меня, а мой mac-адрес такой-то». Получив этот ответ, участник, задавший вопрос, сохраняет полученную информацию в своих таблицах, так что для передачи следующего пакета на тот же адрес задействовать ARP уже не придётся.

Протокол ARP используется в многоточечных сетевых соединениях, таких как Ethernet и WiFi. В соединениях «точка-точка» он не нужен, но физические протоколы, используемые в таких соединениях, чаще всего не определяют никаких соглашений канального уровня (в отличие от физических протоколов для многоточечных соединений, которые вынужденно фиксируют соглашения о длине и структуре передаваемых кадров — то есть соглашения канального уровня). В роли канального протокола в соединениях «точка-точка» часто выступает протокол PPP (*Point to Point Protocol*). Кроме структуры кадра и способа контроля его целостности (через контрольную сумму), этот протокол содержит правила для установления соединения и проверки состояния линии связи (при отсутствии трафика по линии время от времени передаются специальные пакеты, чтобы удостовериться, что связь всё ещё установлена), для аутентификации участников соединения (это может быть нужно, если к линии связи физически может подключиться посторонний; аутентификационная составляющая PPP активно использовалась в эпоху доступа в Интернет через телефонные модемы).

6.2.3. Протокол IP

Следующий слой в модели OSI — «сетевой». Часто можно встретить утверждение, что в Интернете в роли этого уровня выступает протокол IP[11]; это утверждение, как часто бывает, в принципе верно, но не совсем. С одной стороны, именно IP определяет структуру отдельного пакета, передающегося по сети через шлюзы, систему сетевых адресов и некоторые дополнительные функции; все остальные протоколы в Интернете работают «поверх» протокола IP, то есть «заворачивают» всю свою информацию, как служебную, так и прикладную, в IP-пакеты и с их доставкой полагаются на программы, реализующие IP. С другой стороны, в Интернете используется ещё несколько протоколов, которые, если классифицировать их в соответствии с моделью OSI, приходится тоже отнести к сетевому уровню, хотя они и работают поверх IP, то есть вроде бы должны относиться к более высокому слою. Наиболее важным из этих протоколов следует считать ICMP (*internet control message protocol*), без которого сети на основе IP вообще, скорее всего, не могли бы работать, ведь именно в соответствии с соглашениями ICMP участники сетевых соединений и шлюзы обмениваются между собой сообщениями о всяческих штатных и нештатных ситуациях, таких как обрыв соединения, отсутствие подходящего маршрута и другие случаи невозможности доставки пакета; в некоторых случаях сообщения ICMP содержат рекомендации по использованию альтернативного маршрута. Можно обнаружить и другие протоколы, работающие поверх IP, но очевидным образом относящиеся всё к тому же «сетевому» слою OSI.

Протокол IP изначально разрабатывался для американских военных и был предназначен для создания *гетерогенной* компьютерной сети — такой, которую невозможно вывести из строя, уничтожив какой-нибудь «самый главный узел». Предполагалось, что каналы, связывающие между собой отдельные узлы и подсети, могут иметь какую угодно природу, лишь бы по ним можно было передавать пакеты с данными, а схема соединения этих каналов между собой не обязана иметь никакой чёткой структуры, лишь бы только для любых двух компьютеров, входящих в сеть, существовала цепочка каналов (возможно, не единственная), по которой они могли бы друг с другом связаться. Модель, расчитанная на быстрое задействование резервных каналов при разрушении основных, была придумана для сохранения работоспособности в условиях войны, когда узлы сети и каналы связи могут быть в любой момент разрушены. Сеть, получившая название ARPANET, начала работу в 1969 году; протокол IP в его нынешнем виде (IPv4) был опубликован в 1981 году и полностью заменил в ARPANET другие протоколы сетевого слоя 1 января 1983 года; эту дату иногда называют днём рождения Интернета.

На первый взгляд может показаться странным, что именно военная модель и разработанный для неё сетевой протокол оказались самыми подходящими для построения всемирной компьютерной сети — а между тем всё именно так и есть. Явление, которое мы сейчас знаем под именем «Интернет», возникло на основе самого подходящего протокола из многих десятков разнообразных сетевых протоколов, существовавших в те времена. Этому легко найти объяснение. Явление такого масштаба не может возникнуть по указке сверху и тем более на основе каких-либо международных договорённостей, поскольку любые попытки централизованного построения подобной сети совершенно неизбежно утонут в бюрократической волоките. Интернет, каким мы его знаем, возник «снизу» — из частной инициативы, проявляемой разными, никак не связанными друг с другом людьми. Но в таких условиях сложно полагаться на какой-то «порядок», ведь в Интернете свой владелец у каждого канала, у каждой небольшой локальной сетки, у каждого компьютера; Интернет как целое — это некое социальное явление, никому не принадлежащее и никем не контролируемое. Функционирование сети обеспечивает глобальная инфраструктура, элементы которой находятся в частных руках — в десятках тысяч частных рук.

Побочным эффектом этого становится заведомая ненадёжность каждого отдельно взятого элемента глобальной инфраструктуры. Любой канал может перестать работать даже не вследствие аварии или другого ЧП, а просто потому, что его владелец потерял к нему интерес; сложность глобальной инфраструктуры Интернета такова, что каналы связи вводятся в работу и

[11] Считается, что аббревиатура IP означает *Internet Protocol*, хотя протокол появился задолго до возникновения слова *Internet* как имени собственного; изначально первая буква в названии протокола означала *internetworking*, то есть буквально «межсетевой».

выходят из работы буквально ежеминутно, своей динамичностью сеть напоминает муравейник с его вечной суетой. Аналогию с муравейником можно продолжить: надёжность одного отдельно взятого муравья ничтожна, но их много и они взаимозаменяемы, поэтому надёжность муравейника как целого достаточно высока: уничтожить его можно разве что огнём, и то не сразу.

Очевидно, что такой способ существования сети прекрасно ложится на исходную модель протокола IP. Каналы здесь постоянно выходят из строя именно так, как это предполагалось создателями ARPANET, пусть и не в результате военных действий, а по сугубо мирным причинам; сеть при этом сохраняет работоспособность, поскольку информация сразу же начинает передаваться по другим каналам «в обход» прекратившего работу.

Прежде чем двигаться дальше, введём важное понятие ***сетевого хоста***[12]. Под хостом понимается компьютер, подключённый к компьютерной сети и (что важно) имеющий свой собственный сетевой адрес, что позволяет ему получать разнообразные запросы от других компьютеров. Как мы увидим позже, часто встречается ситуация, когда компьютер подключён к сети, но хостом в ней не является и может лишь направлять запросы к другим компьютерам, но сам получать их не может.

В качестве сетевых адресов в протоколе IP используются так называемые ***ip-адреса***, состоящие из четырёх восьмибитных байтов. По традиции ip-адреса записывают в виде четырёх *десятичных* чисел, разделённых точками: `192.168.215.17`, `203.0.113.171`, `10.10.15.0` и т. п. В двоичном (побитовом) представлении эти адреса будут выглядеть так: «1100 0000 1010 1000 1101 0111 0001 0001»; «1100 1011 0000 0000 0111 0001 1010 1011»; «0000 1010 0000 1010 0000 1111 0000 0000». Отметим, что в компьютерных сетях принят порядок байтов в целых числах, соответствующий архитектурам *big endian* (см. т. 1, §1.4.2), то есть старший байт всегда записывается первым.

Как мы уже видели в примере, приведённом в §6.1.2, при настройке маршрутов для передачи пакетов через шлюзы постоянно приходится перечислять компьютеры, для которых будет использоваться тот или иной маршрут, но при этом, как правило, все хосты, относящиеся к одному сетевому соединению (или, в более общем случае, к одной сети, что́ бы под сетью ни понималось), достижимы с помощью одного и того же маршрута. Необходимость как-то обозначить разом все хосты той или иной сети возникает не только при настройке маршрутов, но и во многих других случаях; поскольку в одну сеть может входить несколько тысяч хостов, перечислять все их адреса по одному не годится.

При работе с сетями на основе IP принято соглашение, что адреса можно объединять в ***подсети*** (англ. *subnet*). Размер подсети (т. е. количество адресов в ней) всегда представляет собой степень двойки, при этом в неё входят те и только те адреса, для которых совпадают сколько-то первых битов. Оставшиеся биты в пределах подсети могут принимать произвольные значения, что как раз и даёт степень двойки. Например, если нам нужна подсеть на восемь адресов, то варьироваться в такой подсети будут три последних бита ($8 = 2^3$); поскольку всего в ip-адресе 32 бита, остальные 29 фиксируются и образуют *адрес подсети*, также называемый *ip-префиксом*. Точно так же, если нам нужна подсеть на 256 адресов, то варьироваться будут 8 младших разрядов адреса, а остальные 24 составят адрес подсети (префикс); для подсети на 1024 адреса варьироваться будут уже 10 разрядов, а длина префикса будет равна 22. Адрес подсети записывают так же, как обычный ip-адрес, четырьмя десятичными числами через точку; варьируемые биты при этом принимают равными нулю. Иначе говоря, адрес подсети записывается так же, как наименьший из входящих в эту подсеть ip-адресов.

Подсеть обычно обозначают в виде её адреса подсети, после которого через дробную черту записывают длину префикса. Например, рассмотрим подсеть `203.0.113.32/28`; длина префикса здесь — 28 бит, так что на варьируемую часть адреса остаётся 4 бита; следовательно, эта подсеть содержит $2^4 = 16$ адресов. Нетрудно сообразить, что это ip-адреса с `203.0.113.32` по `203.0.113.47`. В двоичном виде все адреса этой подсети имеют вид «`11001011 00000000 01110001 0010xxxx`», где `xxxx` — варьируемые биты.

В учебниках часто встречается утверждение, что первый и последний адреса любой подсети нельзя использовать для хостов, поскольку самый первый адрес обозначает всю подсеть целиком, а самый последний считается «широковещательным» и предназначен для пакетов, адресованных всем хостам подсети. На самом деле это не совсем так. При использовании ip-подсети для нумерации хостов, например, в локальной сети на основе Ethernet действительно обычно самый первый адрес не используют (хотя при желании использовать его можно), а самый последний настраивают как широковещательный, но так делают, только если адресов достаточно. При нехватке ip-адресов вполне можно задействовать первый и последний адреса подсети для нумерации хостов, нужно только соответствующим образом настроить все хосты данной сети, чтобы, в частности, они не реагировали на последний адрес подсети в качестве широковещательного (практически любая система это позволяет). Бывает и так, что ip-подсеть обозначает не хосты, включённые в какое-то сетевое соединение, а просто абстрактный набор адресов. В такой ситуации первый и последний адрес тем более не имеют никакого особого смысла.

Часто можно увидеть обозначение вроде `203.0.113.41/32`. На первый взгляд понятие «подсети с префиксом в 32 бита» не имеет смысла, ведь в такой подсети не остаётся ни одного бита на вариативную часть; но *один* адрес в эту подсеть всё же входит — в нашем примере это адрес `203.0.113.41`. Иначе говоря, длина префикса 32 бита используется для обозначения *одного отдельно взятого ip-адреса*.

Очевидно, что для каждого адреса можно указать подсеть *любого размера*, в которую он входит. Например, адрес `198.51.100.115` входит в следующие подсети:

`198.51.100.115/32`	`198.51.100.0/24`	`198.51.0.0/16`	`198.0.0.0/8`
`198.51.100.114/31`	`198.51.100.0/23`	`198.50.0.0/15`	`198.0.0.0/7`
`198.51.100.112/30`	`198.51.100.0/22`	`198.48.0.0/14`	`196.0.0.0/6`
`198.51.100.112/29`	`198.51.96.0/21`	`198.48.0.0/13`	`192.0.0.0/5`
`198.51.100.112/28`	`198.51.96.0/20`	`198.48.0.0/12`	`192.0.0.0/4`
`198.51.100.96/27`	`198.51.96.0/19`	`198.32.0.0/11`	`192.0.0.0/3`
`198.51.100.64/26`	`198.51.64.0/18`	`198.0.0.0/10`	`192.0.0.0/2`
`198.51.100.0/25`	`198.51.0.0/17`	`198.0.0.0/9`	`128.0.0.0/1`
		`0.0.0.0/0`	

Последняя «подсеть» обозначает весь Интернет. **Если здесь что-то непонятно, попробуйте представить адреса перечисленных подсетей в двоичном виде**.

Чтобы минимизировать трудозатраты при настройке сетей, системные администраторы всегда стараются выбрать ip-адреса и подсети так, чтобы ip-адреса, используемые в каждой сети, что́ бы ни понималось под «сетью», можно было указать в виде одной ip-подсети. Из этого естественным образом следует, что компьютер, входящий больше чем в одну сеть, должен иметь собственный адрес в *каждой* из сетей, в которые он входит. Поскольку каждое сетевое соединение обычно рассматривается как отдельная сеть (возможно, являющаяся частью большей сети), компьютеру, выполняющему роль шлюза, нужны адреса в каждом из соединений, в которых он участвует. Поэтому **в сетях на основе протокола IP сетевой адрес приписывается не компьютеру, а сетевому интерфейсу**, так что шлюзы обычно имеют больше одного адреса — точнее, столько адресов, сколько у данного шлюза интерфейсов.

Из всего пространства ip-адресов некоторые ip-подсети выделены для особых нужд. В частности, подсеть `0.0.0.0/8` (то есть все ip-адреса, первый байт которых равен нулю) зарезервирована для обозначения хостов в этой же сети; по правде говоря, в такой роли эти адреса нигде не используются и операционные системы такое их использование не поддерживают, но использовать их для чего-то ещё всё равно нельзя. Подсеть `127.0.0.0/8` (все адреса, у которых старший байт равен 127) используется внутри одной системы для обозначения её же самой, когда нужно заставить программу, обычно работающую через сеть, установить связь с другой программой, запущенной на том же компьютере. Опять-таки, в реальности используется всего один адрес — `127.0.0.1`, этот адрес действительно присутствует в любой системе, поддерживающей TCP/IP (а в наши дни это значит — практически в любой системе мира) и обозначает «сам этот компьютер». Остальные адреса, начинающиеся с байта 127, просто не используются. Точно так же не используются адреса, начинающиеся с байтов 240–255. Кто-то когда-то отнёс их к «зарезервированным»; отменить этот их статус никто так и не решился.

[12] Этот термин происходит от английского *network host*; придумать русский перевод никто не потрудился. На всякий случай отметим, что английское слово *host* буквально переводится как «хозяин, принимающий гостей»; в старину этим словом обозначали, например, хозяина постоялого двора или трактира. Читатель наверняка встречался с этим словом и его производными как в области гостиничного бизнеса, так и в области телекоммуникаций (например, в словосочетании «услуги хостинга»).

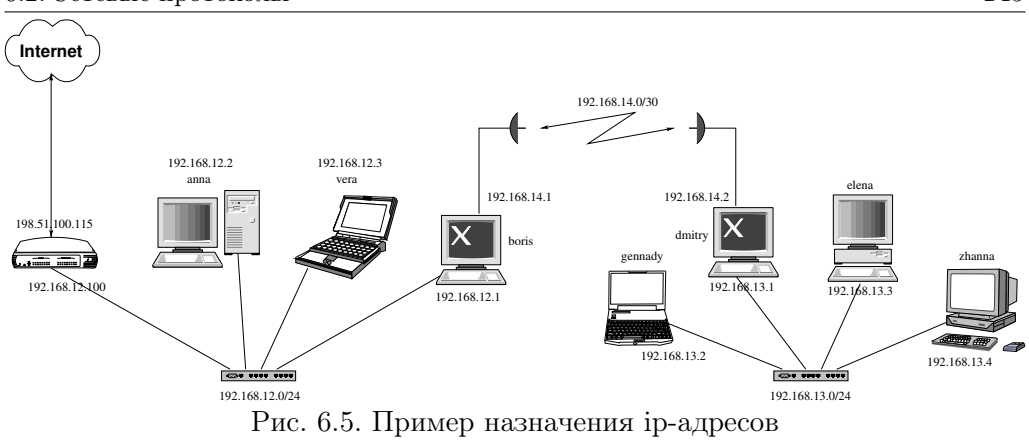

Рис. 6.5. Пример назначения ip-адресов

Адреса, начинающиеся с байтов 224-239, формально говоря, используются — они были выделены для протоколов «группового вещания» (или *мультикастных*, от английского *multicast*). Общая идея «группового вещания» состоит в том, что для одного источника передаваемой информации имеется неопределённое число её получателей, примерно так, как это происходит при передаче теле- и радиопрограмм. Отдельный ip-адрес из подсети `224.0.0.0/4` представляет собой некий аналог *канала*.

Исходно предполагалось, что передающий «мультивещательную» информацию просто начинает передавать в сеть пакеты, адресованные на выбранный ip-адрес; все, кто желает получать информацию, отправленную на этот ip-адрес, сообщают об этом ближайшему маршрутизатору по протоколу IGMP, маршрутизаторы также передают друг другу информацию о собственной заинтересованности в получении данного канала, в результате чего рано или поздно между отправителем и каждым из получателей возникает связная цепочка маршрутизаторов, каждый из которых знает, каким ещё маршрутизаторам следует передавать пакеты, направленные на данный ip-адрес. При этом, по идее, должна достигаться серьёзная экономия на загрузке каналов: например, отправитель посылает ближайшему маршрутизатору только один экземпляр каждого пакета, сколько бы при этом ни было получателей; если на очередном шаге нужно передать «мультикастный» пакет нескольким получателям через одно и то же сетевое соединение, то такой пакет тоже передаётся всего один.

Некоторые из мультикастных ip-адресов зарезервированы для общения между собой маршрутизаторов в крупных локальных сетях, где шлюзы сами динамически выстраивают таблицу маршрутизации; впрочем, изначально ip multicast предназначался для передачи через Интернет аудио- и видеопотоков, аналогичных традиционному радио и телевидению. Проблема в том, что протокол IGMP поддерживается далеко не везде; даже если ближайший к вам маршрутизатор готов принять от вас запрос на «присоединение к группе» (т. е. пожелание принимать мультикастный трафик с заданным адресом), до отправителя пакетов на этот адрес, скорее всего, цепочка так и не выстроится.

Существуют и другие области адресов, зарезервированных под специальные нужды. Для нас могут представлять интерес подсети `10.0.0.0/8`, `172.16.0.0/12` и `192.168.0.0/16`, которые предназначены для использования *в частных сетях*: кто угодно, выстраивая свою компьютерную сеть, может использовать в ней любые из этих адресов по своему усмотрению, но при этом не должен выпускать пакеты с такими адресами за пределы своей сети, то есть компьютеры, имеющие адреса из этих трёх блоков, не видны из Интернета (в нашей терминологии — не являются хостами Интернета), хотя благодаря хитрым преобразованиям адресов сами могут осуществлять доступ к серверам через Интернет.

Так, герои нашего примера, приведённого в §6.1.2, могли бы договориться, что Борис в своём доме будет использовать сеть `192.168.12.0/24` (то есть любые адреса, начинающиеся с `192.168.12.`), а Дмитрий — сеть `192.168.13.0/24`. После этого Борис настроил бы на своём собственном компьютере адрес `192.168.12.1`, на компьютере Анны — адрес `192.168.12.2`, на компьютере Веры — `192.168.12.3`; Дмитрий для своего компьютера мог бы взять адрес `192.168.13.1`, для компьютеров Геннадия, Елены и Жанны — адреса `192.168.13.2`, `192.168.13.3` и `192.168.13.4` (см. рис. 6.5). Впрочем, с тем же успехом можно было бы использовать и другие значения младшего байта адреса, например, `192.168.13.17`, `192.168.13.49`, `192.168.13.121` и `192.168.13.197`, лишь бы все они входили в подсеть. Экономить адреса в частных сетях смысла нет, поэтому наши админы-любители могли бы считать адреса `192.168.12.0` и `192.168.13.0` сетевыми, а адреса `192.168.12.255` и `192.168.13.255` — широковещательными.

Радиоканал между компьютерами Бориса и Дмитрия — это отдельное сетевое соединение, для которого тоже можно выделить подсеть, но использовать подсеть на 256 адресов тут несколько глупо, поскольку в соединении «точка-точка» не может быть больше двух участников; поэтому для своего радиоканала Борис и Дмитрий могли бы, например, воспользоваться подсетью `192.168.14.0/30`, установив на радиоканальном интерфейсе компьютера Бориса адрес `192.168.14.1`, а у Дмитрия — `192.168.14.2`; в роли сетевого и широковещательного тут выступали бы адреса `192.168.14.0` и `192.168.14.3`.

После подключения к Интернету Борис мог бы дать маршрутизатору провайдера адрес `192.168.12.100` (или любой другой из своей подсети). Именно этот адрес должен быть объявлен «шлюзом по умолчанию» для компьютеров, находящихся с ним в одной подсети; а для компьютеров из сети Дмитрия в этой роли будет выступать адрес `192.168.13.1` (компьютер Дмитрия) — именно через него Геннадий, Елена и Жанна будут «видеть» весь Интернет.

Прежде чем показать, как будут выглядеть таблицы маршрутизации на компьютерах из нашего примера, отметим ещё один очень важный момент. Правила маршрутизации бывают двух видов: «пакеты для адресов из данной подсети пересылать через данный сетевой интерфейс» и «пакеты для данной подсети передавать заданному шлюзу для дальнейшей доставки». В обоих случаях требуется указать подсеть (т. е. адрес сети и маску), для которой действует данное правило; после этого в первом варианте указывается обозначение интерфейса, во втором — ip-адрес шлюза. Несложно догадаться, что маршрутизационное правило вида «через интерфейс» для многоточечных соединений может (и должно) использоваться только для указания подсети, используемой для данного соединения, то есть включающей *непосредственно* подключённые к нашему интерфейсу компьютеры; соединения «точка-точка» такого ограничения не имеют. Естественно, для успешной работы правила «через шлюз» нужно, чтобы другое правило определяло, как до этого шлюза добраться.

Допустим, интерфейс для подключения к локальной сети Ethernet на всех компьютерах, участвующих в нашем примере, называется `eth0`, интерфейсы радиоканала — `wlan0` (от слов *ethernet* и *wireless LAN*, именно так обозначаются сетевые интерфейсы в ОС Linux; аббревиатура «LAN» означает *local area network*, т. е. «локальная сеть», слово *wireless* буквально переводится как «беспроводной»). Самыми простыми окажутся маршрутизационные таблицы на компьютерах Геннадия, Елены и Жанны, они будут содержать всего по два правила:

```
192.168.13.0/24   ->   eth0
0.0.0.0/0         ->   192.168.13.1
```

Чуть сложнее будет ситуация на машинах Анны и Веры, поскольку весь Интернет эти машины «видят» через один шлюз, а локальную сеть соседей — через другой. Выглядеть это будет так:

```
192.168.12.0/24   ->   eth0
192.168.13.0/24   ->   192.168.12.1
0.0.0.0/0         ->   192.168.12.100
```

На машинах Бориса и Дмитрия таблица окажется ещё более «заковыристой», ведь каждая из них участвует в двух сетевых соединениях. На машине Бориса таблица маршрутизации будет такой:

```
192.168.12.0/24   ->   eth0
192.168.14.0/30   ->   wlan0
192.168.13.0/24   ->   192.168.14.2
0.0.0.0/0         ->   192.168.12.100
```

Компьютер Дмитрия потребует следующих маршрутных правил:

```
192.168.13.0/24   ->   eth0
192.168.14.0/30   ->   wlan0
0.0.0.0/0         ->   192.168.14.1
```

Сложнее всего будут настройки маршрутизатора, через который в наших сетях осуществляется выход в Интернет. Предположим, что его внешний сетевой интерфейс, обращённый в сеть провайдера, называется `wan0` (*wide area network*; адекватного русского перевода нет) и работает в режиме «точка-точка», а ip-адрес, выделенный нам провайдером — `198.51.100.115`[13]. Непосредственно наш маршрутизатор видит только сеть `192.168.12.0/24`, а сеть `192.168.13.0/24` ему доступна через шлюз, тогда как весь Интернет — через интерфейс `wan0`. Таблица маршрутизации получается такая:

[13] На самом деле такой адрес нам не мог бы выделить ни один провайдер в мире, поскольку подсеть `198.51.100.0/24` зарезервирована для примеров; именно в этом качестве — для примера — мы её и используем. В реальной жизни провайдер должен дать нам адрес из тех, которые выделены ему из общего адресного пространства и больше нигде в мире не используются.

```
192.168.12.0/24   ->   eth0
192.168.13.0/24   ->   192.168.12.1
0.0.0.0/0         ->   wan0
```

Сама по себе эта таблица кажется довольно простой, но на сей раз это лишь начало истории. Дело в том, что адреса, начинающиеся на «192.168.», как мы уже отмечали, *частные*. С одной стороны, это удобно, поскольку каждый волен использовать такие адреса в своей сети как ему будет угодно — именно это обстоятельство позволило нашим персонажам организовать свои домашние сети. С другой стороны, очевидное неудобство этих адресов состоит в том, что компьютеры за пределами отдельно взятой локальной сети не знают и не могут знать, что в этой локальной сети использованы такие адреса. Иначе говоря, маршрутизатор с его внешним адресом является *хостом* в Интернете, а вот компьютеры Бориса, Дмитрия и их домочадцев — нет.

Тем не менее благодаря небольшой хитрости все эти компьютеры могут получить доступ ко всему, что есть в Интернете, практически наравне с полноценными хостами. Это потребует простой дополнительной настройки маршрутизатора, которая включит так называемую ***трансляцию сетевых адресов*** (англ. *network address translation*, NAT). В данном конкретном случае можно заметить, что все адреса, используемые в нашей примерной сети, «накрываются» подсетью 192.168.12.0/22 — эта подсеть, как нетрудно видеть, включает все адреса с 192.168.12.0 по 192.168.15.255; остаётся лишь сообщить маршрутизатору, что все пакеты, исходящие из этой подсети и адресованные хостам за её пределами, подлежат трансляции через внешний адрес маршрутизатора — адрес 198.51.100.115. После этого маршрутизатор, видя такой пакет, подменяет в нём обратный адрес на свой собственный и в таком виде отправляет в Интернет — как будто от своего имени. При получении *ответного* пакета маршрутизатор производит обратную замену и отправляет полученный пакет в локальную сеть. Адрес 198.51.100.115 в такой ситуации называется *внешним*, а адреса из подсети 192.168.12.0/22 — *внутренними*.

Хосты в Интернете, к которым обращаются машины из нашей локальной сети, воспринимают это так, как если бы к ним приходили запросы только от внешнего адреса, про существование внутренних адресов они могут не догадываться. Подчеркнём ещё раз, что машины, работающие на внутренних адресах, вообще не видны из Интернета (не являются хостами Интернета), то есть они могут посылать запросы другим машинам и получать ответы, но сами принимать запросы не могут.

На самом деле, конечно, всё далеко не так просто, ведь с одного и того же «внутреннего» адреса в сеть отправляются пакеты, относящиеся к самым разным соединениям и прочим сеансам работы. Для корректного выполнения трансляции адресов маршрутизатор должен знать, как устроены протоколы транспортного уровня, а в ряде случаев (когда на транспортном уровне отсутствует соединение как таковое, например, при использовании UDP) — учитывать также и протоколы более высоких уровней, вплоть до прикладного. Кроме прочего, маршрутизатор должен помнить, какие соединения от чьего имени в настоящий момент находятся в работе, чтобы не перепутать, какой ответ кому отдать. NAT как явление появился во второй половине 1990-х годов, когда начала ощущаться нехватка ip-адресов, то есть намного позже, чем большинство используемых в Интернете протоколов; исходно TCP/IP и протоколы прикладного уровня не были рассчитаны на такой режим работы. Поэтому, например, протокол FTP, предназначенный для передачи файлов, сначала вообще не мог работать из локальных сетей, скрытых за NAT'ом, а позже этот протокол модифицировали, добавив специальный «пассивный» режим.

Возвращаясь к настройкам маршрутизации в нашем примере, мы можем заметить, что адреса из подсети 192.168.14.0/30, используемые на радиоканале между двумя локальными сетями, не упоминаются в настройках нигде, кроме самих компьютеров Бориса и Дмитрия. В принципе ничто не мешает указать эти адреса в таблицах маршрутизации на всех компьютерах; например, на компьютерах Анны и Веры таблица маршрутизации станет такой:

```
192.168.12.0/24   ->   eth0
192.168.13.0/24   ->   192.168.12.1
192.168.14.0/30   ->   192.168.12.1
0.0.0.0/0         ->   192.168.12.100
```

После этого компьютеры Бориса и Дмитрия будут доступны как по своим исходно установленным адресам (192.168.12.1 и 192.168.13.1), так и по адресам интерфейсов радиоканала — 192.168.14.1 и 192.168.14.2. В большинстве случаев так и поступают, но есть ещё один вариант, который, как ни странно, часто упускают из вида даже опытные системные администраторы: радиоканал (и вообще любое соединение вида «точка-точка») можно вообще не снабжать ip-адресами, ведь, как мы уже отмечали, такие соединения позволяют указывать маршрутные правила «через интерфейс» для произвольных подсетей, а не только непосредственно подключенных, как в случае многоточечных соединений.

Если избавиться от отдельной подсети на радиоканале, то на машине Бориса таблица маршрутизации станет такой:

```
192.168.12.0/24   ->   eth0
192.168.13.0/24   ->   wlan0
0.0.0.0/0         ->   192.168.12.100
```

а на компьютере Дмитрия — такой:

```
192.168.13.0/24   ->   eth0
0.0.0.0/0         ->   wlan0
```

Как можно заметить, маршрутизация при этом стала даже проще.

Есть ещё один важный факт, который следует помнить при работе с таблицами маршрутизации. **Если ip-адрес подходит под несколько разных правил маршрутизации, то маршрутизатор всегда использует то правило, в котором указана подсеть наименьшего размера**. Этот принцип по-английски звучит как *use most specific rule*. Вспомним, например, приведённую чуть выше таблицу маршрутизации для компьютеров Анны и Веры (при условии использования отдельной подсети для радиоканала); применив most-specific rule, её можно несколько упростить:

```
192.168.12.0/24   ->   eth0
192.168.12.0/22   ->   192.168.12.1
0.0.0.0/0         ->   192.168.12.100
```

При таких настройках все пакеты, адресованные в подсеть 192.168.12.0/22, будут передаваться через шлюз 192.168.12.1, *за исключением* пакетов, адресованных в подсеть 192.168.12.0/24, ведь эта подсеть вчетверо меньше по размеру и для неё указано отдельное правило — передавать пакеты через интерфейс eth0, используя непосредственно подключённое соединение.

В нашем примере подсети в 256 адресов используются для сетевых соединений на три и четыре компьютера. Полезно иметь в виду, что адреса можно использовать более экономно. Конечно, экономить частные адреса, в том числе адреса 192.168.*.*, в большинстве случаев нет никакого смысла, но иногда приходится иметь дело с глобально-уникальными адресами, которых вечно не хватает; встречаются, хотя и редко, ситуации, когда экономить приходится и частные адреса тоже — например, когда наша сеть должна быть частью бóльшей сети, использующей частные адреса.

Для примера допустим, что нашим друзьям зачем-то потребовалось, с одной стороны, обойтись по возможности мéньшей подсетью, а с другой стороны — обязательно использовать ip-адреса для интерфейсов радиоканала, плюс к тому во всех подсетях иметь как широковещательные, так и сетевые адреса, не задействуя их для конкретных компьютеров. Ясно, что для каждой из двух локальных сетей можно воспользоваться ip-подсетью на восемь адресов (/29), ведь такая подсеть позволяет обслуживать до шести компьютеров, а вот подсеть следующего размера — /30 — содержит всего четыре адреса, из которых лишь два могут быть приписаны компьютерам. Но ведь есть ещё и радиоканал, для которого нужна как раз подсеть на четыре адреса, и плюс к тому стоит задуматься о размере подсети, которая «накроет» все используемые адреса разом. Если использовать две ip-подсети по восемь адресов и одну на четыре адреса, то их общая ёмкость составит 20 адресов. Общую подсеть придётся сделать на 32 адреса (/27), ведь в подсеть меньшего размера 20 адресов уже не поместятся. Но при этом пропадёт целых 12 адресов, и этим можно воспользоваться, чтобы расширить одну из используемых подсетей. Расширять подсеть, выделенную для радиоканала, нет никакого смысла, поскольку это соединение «точка-точка» и в нём больше двух интерфейсов быть не может, что называется, по определению; иное дело — подсети, используемые для локальных домашних сетей. Сеть в доме Дмитрия больше (пусть и всего на один компьютер), так что логично было бы расширить именно её, воспользовавшись следующим возможным размером ip-подсети — 16 адресов (/28).

Итак, для начала выбираем произвольную ip-подсеть на 32 адреса; пусть это будет 192.168.45.64/27, то есть диапазон от 192.168.45.64 до

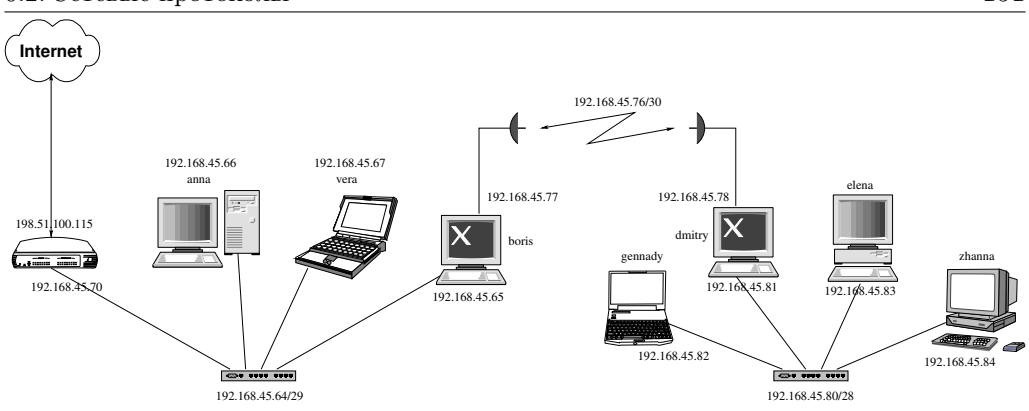

Рис. 6.6. Более компактное использование ip-адресов

Таблица 6.1. Маски для ip-подсетей разных размеров

префикс	маска	префикс	маска
/32	255.255.255.255	/16	255.255.0.0
/31	255.255.255.254	/15	255.254.0.0
/30	255.255.255.252	/14	255.252.0.0
/29	255.255.255.248	/13	255.248.0.0
/28	255.255.255.240	/12	255.240.0.0
/27	255.255.255.224	/11	255.224.0.0
/26	255.255.255.192	/10	255.192.0.0
/25	255.255.255.128	/9	255.128.0.0
/24	255.255.255.0	/8	255.0.0.0
/23	255.255.254.0	/7	254.0.0.0
/22	255.255.252.0	/6	252.0.0.0
/21	255.255.248.0	/5	248.0.0.0
/20	255.255.240.0	/4	240.0.0.0
/19	255.255.224.0	/3	224.0.0.0
/18	255.255.192.0	/2	192.0.0.0
/17	255.255.128.0	/1	128.0.0.0
		/0	0.0.0.0

192.168.45.95. Первые восемь адресов, то есть подсеть 192.168.45.64/29, отведём под локальную сеть Бориса, назначив его компьютеру адрес 192.168.45.65, а компьютерам Анны и Веры — соответственно 192.168.45.66 и 192.168.45.67. Адрес 192.168.45.71 в этой сети будет широковещательным; маршрутизатору, соединяющему нас с Интернетом, выделим адрес 192.168.45.70, а адреса 68 и 69 оставим «на вырост».

Поскольку мы договорились, что для локальной сети Дмитрия задействуем подсеть на 16 адресов, для выделения этой сети у нас не остаётся никаких иных вариантов, кроме как использовать вторую половину нашей «общей» ip-подсети. Читатель, несомненно, догадался, что это будет подсеть 192.168.45.80/28. В ней можно выделить адрес 192.168.45.81 для Дмитрия, адреса 192.168.45.82, 192.168.45.83 и 192.168.45.84 — для Геннадия, Елены и Жанны; широковещательным здесь будет адрес 192.168.45.95, а адреса с 85 по 94 останутся в резерве.

Остаётся ещё радиоканал. Здесь у нас имеется два варианта: подсети 192.168.45.72/30 и 192.168.45.76/30. Мы выберем второй вариант, установив на радиоинтерфейсе компьютера Бориса адрес 192.168.45.77, а на компьютере Дмитрия — адрес 192.168.45.78. Адреса с 192.168.45.72 по 192.168.45.75 так и останутся не входящими ни в одну используемую подсеть, но теперь их всего четыре, а не 12. Новое распределение адресов в сети Бориса и Дмитрия показано на рис. 6.6; написать таблицы маршрутизации для этого варианта предоставим читателю в качестве упражнения.

Во всех наших примерах мы задавали размер ip-подсети, указывая через дробную черту длину её префикса. Достаточно часто можно столкнуться с совершенно иным обозначением — в виде так называемой ***маски подсети***. Такая маска представляет собой 32-битное число, в котором сначала идёт N единичных разрядов (где N — длина префикса соответствующей подсети), а остальные разряды сброшены в ноль. Записывается маска точно так же, как и ip-адрес — в виде четырёх десятичных чисел, разделённых точками, каждое число соответствует одному байту. Пусть, например, нам нужна маска для подсети /27. В двоичном виде это будут байты 11111111 11111111 11111111 11100000, первые три из них представляют собой двоичную запись числа 255, последний — числа 224, так что подсети /27 соответствует маска 255.255.255.224. Маски для всех возможных размеров подсетей приведены в таблице 6.1.

Кроме адресации и принципов передачи пакетов, протокол IP предусматривает для шлюзов некоторые дополнительные обязанности. В частности, каждый пакет имеет *ограниченный срок жизни* (англ. *time to live*, TTL), устанавливаемый при его создании; шлюз при передаче пакета должен уменьшить это значение, и когда оно становится равно нулю, пакет уничтожается. Это делается, чтобы избежать засорения каналов «бесхозными» пакетами при случайных зацикливаниях маршрутных правил. Разные системы устанавливают различные значения TTL по умолчанию: большинство версий Linux использует значение 64, Windows — 128, встречаются системы, позволяющие своим пакетам «пережить» всего 30 переходов через шлюзы (что характерно, этого почти всегда хватает).

Кроме того, обычно каждое сетевое соединение определяет некий максимальный размер пакета, который можно через него передать (*maximal transfer unit*, MTU; чаще всего 1500 байт), и в соответствии с протоколом IP, если пакет не может быть передан через очередное соединение из-за превышения размера, шлюз должен этот пакет разбить на части и передать каждую часть в виде отдельного пакета; обратная сборка пакета в этом случае производится получателем. IP предусматривает контроль целостности заголовка пакета (но не всего пакета) через подсчёт контрольной суммы.

Поскольку ip-адрес представляет собой 32-разрядное целое число, всего их существует 2^{32}, то есть чуть больше 4 миллиардов, но, как уже говорилось, в силу тех или иных причин многие из этих адресов не могут использоваться. Кроме того, на начальных этапах развития сети Интернет некоторым организациям в пользование выдавались адресные блоки гигантского размера, например, с префиксом /8 или /16, а подсети размера /24 (так называемые «сети класса C») одно время считались вообще минимально возможными для предоставления отдельно взятой организации.

Уже в начале 1990-х годов, когда широкая публика ещё не успела толком узнать о существовании Интернета, стало понятно, что ip-адресов скоро перестанет хватать. Раздача адресов направо и налево, характерная для 1980-х, сменилась более бережным отношением к ip-пространству, а затем, ближе к рубежу веков, и вовсе превратилась в режим жёсткой экономии, когда получение уникальных адресов для конечного пользователя стало проблемой. С 1992 года начали появляться различные предложения о замене имеющегося протокола IP новой версией; к 1996 году создание нового протокола, получившего название «IPv6», было более-менее завершено.

В IPv6 адрес по-прежнему имеет фиксированный размер, но этот размер составляет 128 бит, так что пространство адресов IPv6 включает 2^{128} адресов. Чтобы понять, насколько это много, достаточно вспомнить легенды о числе 2^{64} (одна из них связана с изобретением шахмат, вторая — с «ханойскими башнями»), а затем прикинуть, что 2^{128} — это, ни много ни мало, 2^{64} *в квадрате*. Записываются такие адреса в шестнадцатеричной системе счисления, но всё равно получается довольно громоздко, примерно так: 2001:0db8:21b1:a742:0000:ff00:0042:8326. В настоящее время при выдаче блоков адресов IPv6 обычно провайдеры получают подсети с префиксом /32, то есть блоки, содержащие $2^{128-32} = 2^{96}$, или примерно $6{,}5 \cdot 10^{27}$ адресов, а конечным пользователям, запросившим адреса IPv6, выделяется блок /64, содержащий 2^{64} уникальных адресов (отметим, что это число можно получить, если *возвести в квадрат* общий размер адресного пространства протокола IPv4 — и всю эту прорву адресов предоставляют одному пользователю в единоличное распоряжение).

С повсеместным внедрением IPv6, что вполне закономерно, возникли сложности. Дело в том, что, в отличие от 1983 года, когда переход к использованию протокола IPv4 произошёл одномоментно в соответствии с указаниями администрации ARPANET, у Интернета никакой администрации нет, а если бы она и была, требованию «с такого-то момента всем перейти на новый протокол» никто бы не подчинился, Интернет для этого слишком большой и сложный. Поддержка IPv6 сейчас есть практически во всех используемых операционных системах и на всех устройствах, вообще способных работать в Интернете, но не помогает даже это. Дело в том, что, если разместить некий общедоступный ресурс (например, web-сайт) на адресе IPv6, не дав ему адреса IPv4, то для пользователей, не имеющих поддержки IPv6, этот ресурс будет недоступен. Для владельцев сайтов это неприемлемо, поскольку ведёт к потере большей части аудитории (а если называть вещи своими именами — то аудитория теряется вообще вся, ведь пока что с IPv6 никто реально не работает, несмотря на декларируемую готовность); поэтому все создаваемые сайты по-прежнему доступны (и, судя по всему, будут доступны в сколько-нибудь обозримом будущем) через адреса IPv4. Что касается пользователей Интернета, не предоставляющих от своего имени никаких доступных кому-то ещё ресурсов (а таких, к сожалению, подавляющее большинство), то им вполне нормально живётся вообще без ip-адресов — точнее говоря, на «приватных» ip-адресах, невидимых из Интернета, но позволяющих получить доступ ко всем имеющимся в Интернете сайтам и прочим серверам.

Энтузиасты предсказывали массовый переход на IPv6, когда «старые» ip-адреса кончатся; но когда в 2011 году между пятью региональными регистратурами ip-адресов были распределены последние пять блоков /8, на IPv6 так никто и не перешёл. Можно было встретить утверждение, что переход на IPv6 пойдёт быстрее, когда эти пять блоков подойдут к концу, но и этого не произошло. Региональные регистратуры, достигнув определённого предела исчерпания свободного ip-пространства, окончательно урезали возможности получения ip-адресов для новых провайдеров и автономных сетей: теперь, создав новый провайдер доступа в Интернет или хостинговый оператор, можно получить для него подсеть /22, содержащую всего 1024 адреса, больше никто не даст.

Конечным пользователям практически нереально получить в своё распоряжение больше *одного* адреса, да и один адрес дают не во всех провайдерах; такое положение жесточайшей экономии адресов существует уже несколько лет и (предположительно) при нынешней скорости раздачи адресов IPv4 может протянуть ещё не одно десятилетие. Формально организации, ответственные за распределение ip-адресов, называют нынешнюю эпоху «периодом перехода к IPv6», но можно усомниться, что всеобщий переход на IPv6 когда-нибудь в действительности произойдёт. К примеру, если всё-таки начать использование адресов с первым байтом 240-255 (а технические сложности, которые при этом возникнут, не идут ни в какое сравнение со сложностями перехода на IPv6), то это даст ещё 16 блоков размера /8, которых при нынешней черепашьей скорости раздачи адресов может хватить лет на сто. Кроме того, большое количество уже выданных адресов сейчас не используется, поскольку когда-то они были выделены организациям, не нуждающимся в таком количестве адресов; просто «взять и отобрать» у них адреса вряд ли получится, но, вполне возможно, когда-то ip-адреса станут обычным предметом купли-продажи. Ничего хорошего это не принесёт, возникнет огромное количество проблем как технического (взрывной рост размера мировой таблицы маршрутизации из-за дробления подсетей), так и правового характера (в самом деле, как может быть объектом купли-продажи *32-битное целое число*?!), но проблема исчерпания адресного пространства при этом, скорее всего, исчезнет надолго.

Конечно, если не рассматривать сценарии мировой катастрофы, то рано или поздно адреса протокола IPv4 всё же закончатся, и закончатся совсем, полностью, когда никакая экономия уже не поможет, поскольку экономить будет нечего; но к тому времени, скорее всего, появится что-то более удачное, нежели протокол IPv6, который даже сейчас, спустя всего двадцать лет после своего появления, выглядит довольно странно. Может так случиться, что этот протокол будет признан морально устаревшим, так и не дождавшись внедрения.

6.2.4. Дейтаграммы и потоки

Следующий слой модели OSI, как мы помним, называется *транспортным*. Чтобы понять, в чём состоит его роль, нужно для начала заметить, что пакеты, передаваемые на уровне IP — это инструмент, с которым достаточно тяжело работать. Как мы уже отмечали выше, протоколы сетевого уровня, в том числе IP, не гарантируют доставку пакета. Отправленный пакет может потеряться или, наоборот, прийти в двух экземплярах, а пакеты, отправленные раньше других, могут прийти к получателю позже. Содержимое такого пакета обычно называют ***дейтаграммой***. С точки зрения процесса, который пытается обменяться информацией с партнёром через компьютерную сеть, дейтаграммное взаимодействие выглядит так: берём некую порцию данных (например, содержимое какого-то массива или структуры, либо часть такого содержимого), надписываем адрес получателя и просим операционную систему позаботиться о пересылке. Поскольку пакет может не дойти, адресат в ответ на полученный пакет обязательно должен отправить подтверждение; в свою очередь, отправитель должен некоторое время ожидать получения подтверждения, а если такового не поступит, снова отправить ту же самую дейтаграмму, предполагая, что первая потерялась. Для экономии пропускной способности сети подтверждения можно отправлять не на каждый пакет, а сразу на некоторую их последовательность, причём в подтверждении можно указать, какие пакеты были получены, а какие потерялись. Всю эту механику приходится реализовывать прямо в программах, которые будут взаимодействовать друг с другом через сеть; всё, что предоставляет нам операционная система (содержащая реализацию протокола передачи дейтаграмм) — это возможность послать дейтаграмму, указав адрес её получателя.

Ситуация существенно осложняется, когда дейтаграммы на самом деле представляют собой фрагменты одного информационного объекта — например, файла. Поскольку при доставке дейтаграмм может нарушиться их порядок, нужно, чтобы отправитель в каждом пакете указывал его номер или как-то иначе обозначал место этой дейтаграммы в составе целого, а получатель на своей стороне должен восстановить мозаику фрагментов, при необходимости запросив у отправителя повторную отправку тех пакетов, которые потерялись по дороге.

Если программистов устраивает такой режим работы, то от протокола транспортного уровня им нужно совсем немного. Когда в качестве сетевого протокола используется IP, как мы уже знаем, контрольная сумма, содержащаяся в заголовке его пакета, позволяет проверить целостность заголовка, но не всего пакета; для проверки правильности передачи «полезной нагрузки» (в данном случае — дейтаграммы) нужно где-то хранить контрольную сумму для дейтаграммы. Кроме того, на одном компьютере может работать одновременно много разных программ, желающих отправлять и принимать дейтаграммы, и их нужно как-то между собой различать; сам компьютер можно идентифицировать по сетевому адресу.

Соглашения о том, как решить эти вопросы, содержатся в протоколе UDP (*user datagram protocol*). Это один из самых простых протоколов в Интернете, поскольку почти всю работу уже сделал за него протокол IP. UDP предусматривает, что участники сетевого взаимодействия, работающие на одном компьютере, будут для идентификации использовать двухбайтные беззнаковые целые числа, называемые ***номерами портов*** или просто ***портами***. К данным, которые нужно передать через сеть, протокол UDP предписывает добавить заголовок из 8 байт, содержащий номер порта отправителя, номер порта получателя, длину дейтаграммы (данные вместе с заголовком) и контрольную сумму; полученная структура (дейтаграмма) «заворачивается» в пакет в соответствии с соглашениями протокола IP. Сетевые адреса отправителя и получателя указываются уже в заголовке этого пакета.

Как видим, дейтаграммное взаимодействие оказывается очень простым в реализации; при этом его довольно сложно использовать, ведь все заботы о подтверждениях, повторных отправках, о делении информации на небольшие порции и о последующем восстановлении исходной последовательности приходится брать на себя автору программы. Со вторым основным видом транспортного взаимодействия — *потоковым* — дела обстоят прямо противоположным образом. С точки зрения программ, взаимодействующих таким способом, сетевое соединение выглядит как обычный поток ввода-вывода, к которому можно применять хорошо знакомые нам системные вызовы `read` и `write`. Информацию, записанную в один конец сетевого соединения, можно после доставки её через сеть прочитать из другого конца, и обратно, то есть поток в данном случае оказывается двухсторонним. Операционная система гарантирует, что все байты, записанные в поток, будут затем доступны для чтения на другом конце потока, причём их порядок будет сохранён; при невозможности соблюдения этой гарантии соединение окажется разорвано, о чём узнают оба партнёра — очередные системные вызовы вернут им ошибки.

Очевидно, что обмениваться информацией через такое соединение намного проще и удобнее, чем с помощью дейтаграмм. Решение всех проблем, связанных с потерями пакетов, восстановлением данных из отдельных фрагментов и т. д. берёт на себя реализация транспортного протокола, которая, конечно же, оказывается гораздо сложнее реализации протокола дейтаграммного. Соответствующий транспортный протокол, используемый в Интернете, называется TCP (*transmission control protocol*, т. е. «протокол управления передачей [данных]»). Точно так же, как и UDP, протокол TCP использует *номера портов* для идентификации участников взаимодействия, работающих на одном компьютере, но сходство на этом заканчивается; достаточно сказать, что заголовок пакета для TCP имеет переменную длину и может занимать от 16 до 56 байт. В отличие от работы с дейтаграммами, при потоковой работе необходимо сначала *установить соединение*, для чего один из будущих партнёров должен находиться в состоянии ожидания запроса на соединение, а второй должен этот запрос послать; естественно, партнёры должны уметь подтверждать получение отдельных пакетов, а при отсутствии такого подтверждения — посылать недостающие пакеты повторно; кроме того, в зависимости от целого ряда условий партнёры по соединению выбирают *размер окна передачи*, то есть количество данных, передаваемых в каждом пакете, и этот размер может изменяться в ходе работы. Протокол TCP предусматривает, кроме перечисленного, ещё целый ряд нетривиальных соглашений, пояснение которых заняло бы слишком много места. К счастью для прикладных программистов, всю реализацию соглашений протокола TCP берёт на себя операционная система.

6.2.5. Протоколы прикладного слоя

Оставшиеся три уровня модели OSI — сеансовый, представительный и прикладной — вычленить из реально существующих протоколов будет несколько сложнее.

Интереснее всего обстоят дела с сеансовым уровнем. Наиболее популярно мнение, что в Интернете протоколы этого уровня попросту отсутствуют; это звучит несколько неожиданно, поскольку сеансы работы как таковые, вне всякого сомнения, не только присутствуют, но и постоянно используются. Проблема лишь в том, что организуются они либо средствами протоколов прикладного уровня, либо соглашениями, имеющими *ещё более высокий уровень*. В модели OSI не предусмотрено слоя, находящегося выше прикладного; ну а в Интернете невозможно найти такие протоколы, которые занимались

бы организацией сеансов и при этом находились *между* транспортным слоем (TCP/UDP) и слоем прикладным.

Хрестоматийным примером сеанса, организованного поверх протоколов прикладного уровня, можно считать сеанс работы с веб-сайтом, предусматривающим аутентификацию пользователя. Протоколы HTTP и HTTPS, используемые для взаимодействия браузера и сервера, предполагают загрузку каждой новой версии страницы, а также дополнительных элементов — изображений, стилевых файлов и т. п. — с использованием новых соединений; одно соединение может использоваться, а может и не использоваться для обслуживания нескольких запросов, идущих подряд, но, так или иначе, сеанс работы с информационной системой через web-интерфейс заведомо состоит более чем из одного соединения. Логическое объединение этих соединений в один сеанс работы с пользователем достигается путём сохранения идентификатора сеанса либо в файле cookie на стороне браузера, либо в виде параметра в адресной строке; очевидно, что эти механизмы работают *поверх* протоколов HTTP/HTTPS, то есть находятся на уровне более высоком, а не менее высоком, как того требует модель OSI.

Поскольку протокол есть набор соглашений, мы можем отнести к числу протоколов кодировки текста (ASCII, koi8-r, utf-8 и другие), а также форматы данных и правила их интерпретации — например, спецификации разнообразных версий HTML/XHTML, графических форматов JPEG, GIF, PNG и прочее в таком духе (ведь всё это тоже наборы соглашений). Кроме того, стоит упомянуть используемый в Интернете набор спецификаций под общим названием MIME, который исходно предназначался для электронной почты, но со временем проник и в другие коммуникационные среды. MIME позволяет представлять в виде ASCII-текста файлы различных типов и наборы файлов, задаёт классификацию форматов и правила их идентификации; когда мы отправляем по электронной почте письма с вложенными файлами, почтовые программы используют для этого как раз соглашения MIME. Если считать соглашения о форматах данных протоколами, то они будут, очевидно, относиться к представительному уровню, хотя создатели модели OSI имели в виду не совсем это.

С прикладным уровнем всё оказывается достаточно просто: на этом уровне организуется взаимодействие программ, непосредственно обеспечивающих решение информационных задач для конечного пользователя, в том числе, например, таких хорошо известных систем, как электронная почта, «всемирная паутина» (World Wide Web, или WWW), службы передачи моментальных сообщений и т. п.

Большинство сервисов Интернета построено в соответствии с так называемой ***клиент-серверной моделью***, что, по-видимому, нуждается в дополнительных пояснениях. **Под *сервером* понимается программа (или, в широком смысле слова, любое «действующее лицо» — набор взаимосвязанных программ, компьютер, система в целом и т. п.), ожидающая запросов и производящая какие-либо действия исключительно в ответ на запросы, а при отсутствии запросов не делающая ничего**. Как несложно догадаться, **под *клиентом* понимается программа (или другой участник взаимодействия), обращающаяся с запросом к серверу**. Вообще говоря, одна и та же программа может выполнять функции сервера по отношению к одним программам и клиента — по отношению к другим, если серверу для удовлетворения запроса клиента приходится пользоваться услугами другого сервера.

Многие протоколы прикладного уровня (хотя, конечно, не все) основываются на обмене *текстовыми* репликами через двунаправленное потоковое соединение. Именно таковы, в частности, протоколы SMTP (*simple mail transfer protocol*, протокол передачи электронной почты) и HTTP (*hypertext transfer protocol* — протокол, по которому браузер связывается с web-сайтами). Текстовый формат этих протоколов позволяет при желании связаться с сервером «вручную» и «поговорить» с ним, используя любую программу, позволяющую установить потоковое соединение и передавать туда-обратно простой текст; например, для этого вполне подойдёт программа `telnet`, входящая в стандартную комплектацию практически любой unix-системы, а до недавних пор (вплоть до Windows XP) входившая также и в состав систем линейки Windows.

Приведём пример сеанса отправки электронного письма от имени пользователя `test@stolyarov.info` получателю, имеющему адрес `test123@unicontrollers.com`:

```
=   bash-3.1$ telnet 195.42.170.131 25
=   Trying 195.42.170.131...
=   Connected to 195.42.170.131.
=   Escape character is '^]'.
    220 pluto.unicontrollers.com ESMTP Postfix
>   EHLO reptile.croco.net
    250-pluto.unicontrollers.com
    250-PIPELINING
    250-SIZE 10240000
    250-ETRN
    250-ENHANCEDSTATUSCODES
    250-8BITMIME
    250 DSN
>   MAIL FROM: <test@stolyarov.info>
    250 2.1.0 Ok
>   RCPT TO: <test123@unicontrollers.com>
    250 2.1.5 Ok
>   DATA
    354 End data with <CR><LF>.<CR><LF>
>   From: Andrey Stolyarov <test@stolyarov.info>
>   To: Test Account <test123@unicontrollers.com>
>   Subject: test message
>
>   This is a test
>   .
    250 2.0.0 Ok: queued as 1549E481E12
>   QUIT
    221 2.0.0 Bye
=   Connection closed by foreign host.
=   bash-3.1$
```

Для наглядности мы пометили знаком «=» строки, напечатанные локальными программами — интерпретатором командной строки и программой `telnet`; эти строки не являются частью сеанса и не имеют никакого отношения к протоколу. Строки, переданные серверу (напечатанные на клавиатуре в ходе сеанса), помечены знаком «>»; все остальные строки — это ответы сервера. Попробуем разобрать, что здесь к чему.

Сразу после установления соединения сервер «представился» нам, прислав строку

```
220 pluto.unicontrollers.com ESMTP Postfix
```

Здесь `pluto.unicontrollers.com` — доменное имя компьютера в сети, аббревиатура `ESMTP` указывает на то, что сервер поддерживает протокол ESMTP (расширенную версию SMTP, буква `E` означает слово *extended*), а слово `Postfix` — это название используемой «по ту сторону» серверной программы. Некоторые программы выдают в этой строке номер своей версии, так может поступить и Postfix, если его по-другому настроить. Число `220` — это так называемый *код ответа* (*response code*); к этому моменту мы ещё вернёмся.

Первая команда, которую мы дали серверу — это команда `EHLO` с параметром, который означает доменное имя нашей (передающей) машины, сейчас выступающей в роли клиента. В ранних версиях протокола команда называлась `HELO` и была образована от английского слова *Hello*; используя команду `EHLO` вместо `HELO`, мы извещаем сервер, что готовы к использованию поздних расширений протокола. Сервер ответил на эту команду, перечислив кодовые имена расширений, которые он поддерживает; если бы мы использовали `HELO`, ответ сервера состоял бы из одной строки:

```
250 pluto.unicontrollers.com
```

— поскольку именно такой ответ предусмотрен ранними спецификациями SMTP. Впрочем, если говорить совсем строго, то после кода `250` и пробела протокол допускает произвольную строку, предназначенную скорее для человека, нежели для компьютерных программ.

Так или иначе, сейчас мы используем расширенную версию протокола, и многострочный ответ сервера позволяет клиенту (в роли которого обычно выступает другая программа, а не человек, как в нашем случае) узнать, какие из расширений протокола сервер способен обработать; от этого зависит, какие команды стоит посылать, а какие нет. Впрочем, мы расширениями протокола не воспользовались; оставшиеся команды, использованные в нашем сеансе, присутствовали в SMTP с самого начала.

Сразу после обмена приветствиями с сервером мы послали ему команду `MAIL FROM:`, которая указывает, что мы сейчас намерены передать электронное письмо, отправителем которого является владелец адреса `test@stolyarov.info`:

```
MAIL FROM: <test@stolyarov.info>
```

Ответ сервера, как видим, весьма лаконичен:

```
250 2.1.0 Ok
```

Большинство программ, работающих с протоколом SMTP, из всей этой строки использует только код 250, означающий, что последняя команда принята сервером без каких-либо замечаний. Несколько реже используется так называемый *усовершенствованный код* — цифры 2.1.0; это уже особенность расширенной версии SMTP, точнее говоря — одного конкретного расширения, которое сервер поддерживает, о чём он заявил в ответ на наше EHLO (идентификатор этого расширения — ENHANCEDSTATUSCODES). Что касается слова Ok, то это уже сугубо декоративный комментарий; некоторые сервера могут в таких случаях быть более многословны, выдать, например, что-нибудь вроде «Sender address test@stolyarov.info is welcome». Postfix, как видим, предпочитает краткость. Надо сказать, что команда MAIL FROM: допускается только одна, то есть мы не имеем права (согласно протоколу) давать вторую такую команду, пока не закончим с передачей одного письма.

Следующая команда — RCPT TO: — задаёт адрес *получателя*. Здесь интересны сразу два момента. Во-первых, этих команд мы можем дать сколько угодно; именно так и происходит, если наше письмо адресовано сразу нескольким получателям. Во-вторых, если команду MAIL FROM: сервер чаще всего пропускает (лишь некоторые сервера производят различные проверки адреса отправителя и, например, отказываются принимать почту, если домен отправителя не существует), то команда RCPT TO: будет успешной лишь при соблюдении довольно жёстких условий. Почтовые сервера в наше время настраиваются так, чтобы принимать почту от любого пользователя для своего пользователя либо от своего пользователя (то есть, например, пришедшего из локальной сети, обслуживаемой этим сервером) для кого угодно. Сервер pluto.unicontrollers.com, с которым мы установили соединение, настроен на обслуживание нескольких почтовых доменов, в том числе домена unicontrollers.com, и знает, что адрес test123@unicontrollers.com в этом домене действительно существует (на самом деле он был создан специально для нашего эксперимента).

Примерно до 1994 года практически все почтовые сервера сети Интернет были настроены существенно либеральнее — они были готовы от кого угодно принять почту для доставки кому угодно, а затем уже сами определяли, на какой сервер следует отправить письмо, чтобы оно достигло своего адресата. К сожалению, этим начали активно пользоваться спамеры — любители засорять чужие почтовые ящики рекламным мусором. Сообществу системных администраторов сети Интернет пришлось поменять своё отношение к настройкам серверов. Уже к концу 1995 года в Интернете почти не осталось серверов, готовых принимать и передавать почту для кого угодно — так называемых *открытых релеев*. В наше время открытые релеи изредка появляются, но тут же оказываются внесены в разнообразные блокирующие списки, в результате чего от них отказывается принимать почту большая часть Интернета. Надо сказать, что попасть в такой «чёрный список» обычно гораздо проще, чем потом из него выбраться; если вам когда-нибудь доведётся настраивать почтовый сервер, обратите особое внимание на запрет открытой ретрансляции.

Конечно, сервер, с которым мы экспериментировали, не допускает открытой ретрансляции. Если бы в команде RCPT TO: мы указали адрес в домене, не входящем в число обслуживаемых этим сервером, мы бы получили отказ:

```
>  RCPT TO: <test17@croco.net>
   554 5.7.1 <test17@croco.net>: Relay access denied
```

Отказ мы бы получили также и в случае, если бы указанный нами адрес находился в «правильном» домене, но при этом сервер обнаружил, что такого адреса на самом деле нет:

```
>  RCPT TO: <abrakadabra@unicontrollers.com>
   550 5.1.1 <abrakadabra@unicontrollers.com>: Recipient address
   rejected: User unknown in virtual alias table
```

(на самом деле весь ответ, начиная с кода 550, исходно был записан в одну строчку, но нам пришлось его разбить).

Сообщив серверу адрес отправителя, адреса всех получателей и получив положительные ответы на все команды, мы можем, наконец, приступить к отправке письма; делается это командой DATA. Сервер нам на неё ответил немного хитро:

```
354 End data with <CR><LF>.<CR><LF>
```

Здесь для клиентской программы интересен только код 354, означающий, что следует продолжить начатое, а комментарий предназначается для человека вроде нас и переводится так: «закончите данные последовательностью <CR><LF>.<CR><LF>». Поясним, что под <LF> понимается хорошо знакомый нам символ перевода строки (код 10), а под <CR> — символ *возврата каретки* (код 13). Как видим, протокол SMTP предполагает, что строки в текстовом потоке разделяются двумя байтами примерно так, как в текстовых файлах в системах линейки Windows; впрочем, автору не попадалось ни одной программы, которая не принимала бы поток со строками, разделёнными одним только переводом строки. Смысл этого ответа от сервера сводится к тому, что текст письма следует начать сразу после команды DATA, а когда он закончится, сообщить об этом серверу, передав строку, состоящую из одного символа точки. Именно так мы и поступили; нужно только заметить, что текст электронного письма состоит из *заголовка* и *тела*, разделённых пустой строкой. Заголовок мы сформировали из трёх строк:

```
From: Andrey Stolyarov <test@stolyarov.info>
To: Test Account <test123@unicontrollers.com>
Subject: test message
```

— а в качестве тела передали одну короткую фразу «This is a test», отделив её пустой строкой от заголовка. После этого мы сообщили серверу, что письмо окончено, отправив строку, состоящую из одной точки. Сервер ответил нам, что письмо принято для дальнейшей обработки:

```
250 2.0.0 Ok: queued as 1549E481E12
```

(поясним, что 1549E481E12 — это присвоенный нашему письму идентификатор в локальной очереди сообщений, ожидающих дальнейшей доставки). Интересно, что в ходе дальнейшей обработки сервер добавил к нашему письму ещё несколько строк заголовка. Вот во что превратилось наше письмо, когда оно дошло до почтового ящика:

```
From test@stolyarov.info  Wed Mar 29 21:35:15 2017
Return-Path: <test@stolyarov.info>
Delivered-To: test123@unicontrollers.com
Received: from reptile.croco.net (reptile.croco.net [195.42.160.101])
        by pluto.unicontrollers.com (Postfix) with ESMTP id 1549E481E12
        for <test123@unicontrollers.com>; Wed, 29 Mar 2017 14:11:29 +0000
        (UTC)
From: Andrey Stolyarov <test@stolyarov.info>
To: Test Account <test123@unicontrollers.com>
Subject: test message
Message-Id: <20170329141142.1549E481E12@pluto.unicontrollers.com>
Date: Wed, 29 Mar 2017 14:11:29 +0000 (UTC)

This is a test
```

Вы и сами можете провести аналогичный эксперимент, отправив «вручную» письмо самому себе; нужно только узнать, какая машина обслуживает ваш домен. Для этого достаточно обратиться к глобальной распределённой базе данных DNS с помощью, например, команды host, которая, скорее всего, уже имеется в вашей системе. Так, машину, обрабатывающую почту для unicontrollers.com, можно установить с помощью следующей команды:

```
avst@host:~$ host -t MX unicontrollers.com
unicontrollers.com mail is handled by 5 pluto.unicontrollers.com.
```

С протоколом HTTP всё обстоит ещё проще. Для примера запросим главную страницу сайта www.stolyarov.info:

```
=  avst@host:~$ telnet www.stolyarov.info 80
=  Trying 195.42.170.129...
=  Connected to varan103.croco.net.
=  Escape character is '^]'.
>  GET / HTTP/1.1
>  Host: www.stolyarov.info
>
```

В первой строчке здесь мы видим приглашение командной строки на машине, на которой мы работаем, и команду `telnet`. Следующие три строчки нам выдала сама программа `telnet`, после чего мы ввели *заголовок запроса*, состоящий в нашем случае из двух строк:

```
GET / HTTP/1.1
Host: www.stolyarov.info
```

Чтобы показать, что заголовок запроса окончен, мы передали серверу пустую строку. Ответ сервера при этом довольно многословен:

```
HTTP/1.1 200 OK
Date: Wed, 29 Mar 2017 18:07:52 GMT
```

и так далее; весь его мы приводить не будем. Сначала в ответе идёт заголовок, содержащий разнообразную служебную информацию, затем пустая строка, а после неё — собственно HTML-код главной страницы сайта. В отличие от SMTP, где сервер присылает «приветственную» строку сразу после установления соединения, протокол HTTP предполагает, что обмен по установленному каналу начнёт клиент; именно это мы и сделали, послав сначала команду `GET` с указанием локального пути к нужной странице и версии протокола, а потом дополнительную строку `Host`, позволяющую серверу понять, к какому из обслуживаемых сайтов мы хотим обратиться. Получив заголовок целиком, сервер ответил на запрос, прислав нам сначала служебную информацию, а затем и текст нужного нам документа — главной страницы сайта.

У двух рассмотренных нами протоколов совершенно неожиданно обнаруживается общая (точнее, похожая) система трёхзначных *кодов результата*. Поясним, что коды, начинающиеся с двойки, означают успешное выполнение запроса; коды, начинающиеся с четвёрки, обозначают «устранимые» ошибки (запрос невозможно выполнить прямо сейчас, но, возможно, через некоторое время ситуация станет более благоприятной); а коды, начинающиеся с пятёрки, сигнализируют о фатальных ошибках, таких, которые точно никуда не денутся, во всяком случае без вмешательства администраторов сервера. Коды, начинающиеся с единицы, используются только в HTTP и означают, что сервер предлагает клиенту продолжать начатое; коды, начинающиеся с тройки, в HTTP означают, что запрошенную информацию следует искать где-то в другом месте; в SMTP используется только один код, начинающийся с тройки (`354`), он выдаётся в ответ на команду `DATA` и означает, что сервер теперь ожидает получения тела электронного письма.

6.2.6. Доменные имена

Ранее мы обсуждали ip-адреса, записываемые цифрами, которые используются для идентификации хостов в Интернете. Несомненно, читатель знает из своего опыта, что для конечного пользователя Интернета необходимости работы с такими адресами практически никогда не возникает, вместо них используются ***доменные имена***, такие как `www.stolyarov.info`, которые существенно проще запомнить. Такие же имена мы использовали в примерах работы по протоколам SMTP и HTTP в предыдущем параграфе.

Использование доменных имён становится возможно благодаря функционированию в Интернете так называемой ***системы доменных имён*** (*domain name system*, DNS). Часто можно услышать, что её функция состоит в преобразовании символических доменных имён в ip-адреса, что в принципе верно (можно даже считать, что это *основная* функция DNS), но реальность, как водится, устроена несколько сложнее. **Система доменных имён представляет собой глобальную распределённую базу данных, в которой иерархия доменных имён, записываемых через точку, используется как ключ для поиска; информация, связанная с конкретным доменным именем, совершенно не обязана ограничиваться ip-адресом**, структура базы данных позволяет хранить записи различных типов, в том числе простые текстовые сообщения.

Само по себе слово «домен» (*domain*) переводится как «множество» или «область», но в достаточно специфических контекстах. Например, область определения и область значений для математической функции по-английски называются *domain* и *codomain*. Биологи для систематизации живых организмов используют самую верхнюю ступень классификации — делят все живые организмы на три *домена*: археи, бактерии и эукариоты; к последним относятся все организмы, клетки которых имеют клеточное ядро, в том числе растения, животные и грибы (это уже так называемые *царства*). Слово *домен* в данном случае прямо заимствовано из английского, где эта ступень классификации называется точно так же — *domain*.

В системе доменных имён собственно *доменное имя* — это последовательность разделённых точками *токенов*, каждый из которых представляет собой непустую последовательность латинских[14] букв, цифр и символа дефиса, причём начинаться и заканчиваться дефисом токен не может. Например, доменное имя `www.stolyarov.info` состоит из трёх токенов: `www`, `stolyarov` и `info`. В качестве других примеров токенов перечислим `zx19`, `25ab`, `25-a-b`, `a`, `zzzzzz`, `777` и т. п. Количество токенов в доменном имени называется *уровнем* этого имени: так, `www.stolyarov.info` — это имя третьего уровня, `stolyarov.info` — второго, а `info` — первого («верхнего», англ. *top level*).

С каждым доменным именем DNS позволяет связать произвольное количество *записей*, каждая из которых имеет служебный параметр *time to live* (количество секунд, в течение которого с момента получения запись можно продолжать считать актуальной), а также *класс*, *тип* и собственно содержимое. Класс записей в DNS реально используется всего один — `IN` (от слова *Internet*), хотя в спецификациях упоминается ещё по меньшей мере два других класса и два служебных значения, которые классами не являются, но якобы должны использоваться вместо класса. С типами записей ситуация интереснее. Самым очевидным типом DNS-записи можно считать тип `A` (от слова *address*): такая запись содержит адрес протокола IPv4, который должен быть поставлен в соответствие данному доменному имени. Для ip-адресов протокола IPv6 используются записи типа `AAAA`, для указания серверов электронной почты — записи `MX` (*mail exchanger*), что, кстати, позволяет создавать доменные имена, используемые только для почты. Для указания местонахождения серверов некоторых других протоколов можно использовать записи типа `SRV`, для указания обычных текстовых данных — записи типа `TXT`. Есть и другие типы записей; с некоторыми из них мы столкнёмся позже.

Суффиксами доменного имени называют доменные имена, полученные из исходного отбрасыванием нуля или более токенов слева. Так, доменное имя `www.stolyarov.info` имеет три суффикса: `www.stolyarov.info`, `stolyarov.info` и просто `info`. **Доменные имена, имеющие общий заданный суффикс, как раз и образуют домен**; иначе говоря, ***домен*** — это (бесконечное) множество доменных имён, имеющих некий заданный общий суффикс. Например, в домен `stolyarov.info` входят реально используемые имена `stolyarov.info` и `www.stolyarov.info`, но могут также входить и другие имена, например, `test.stolyarov.info`, `mars.stolyarov.info`, `z12.class.stolyarov.info` и т. п. Следует отметить, что домены в полном соответствии с приведённым определением *вкладываются друг в друга*: например, имя `z12.class.stolyarov.info` входит как в домен `class.stolyarov.info`, так и в домен `stolyarov.info`, и в домен первого уровня `info`. Можно сказать, что домен `class.stolyarov.info` является подмножеством (*поддоменом*) домена `stolyarov.info`, при этом оба они представляют собой поддомены домена `info`, а он, в свою очередь, считается поддоменом ***корневого домена***. Корневой домен обозначается точкой «`.`» и объединяет все существующие домены и доменные имена.

Иерархичность доменов — это как раз то свойство, которое позволяет глобальной базе данных DNS быть полностью распределённой, обслуживаться сотнями тысяч DNS-серверов, находящихся в совершенно разных руках, и при этом избегать появления «бутылочных горлышек» — таких мест, через которые проходило бы слишком много запросов и которые было бы тяжело «расширить». Обслуживание конкретного домена может быть поручено двум или более[15] серверам; такие сервера называются *авторитативными* для данного домена. Обычно один из авторитативных серверов выполняет функции «главного» (*master*), а остальные — функции «подчинённых» или «ведомых» (*slaves*). Время от времени ведомые запрашивают у главного актуальную версию информации, относящейся к обслуживаемому домену. В случаях, когда информация изменилась на главном сервере, он сам оповещает об этом подчинённых, чтобы они не ждали следующего планового момента обновления, но даже если оповещение (представляющее собой одну дейтаграмму) до адресата не дойдёт — это не так страшно, просто информация будет обновлена чуть позже.

[14] Читатель может заметить, что в наше время домены бывают и кириллическими. Это не совсем верно: русскоязычные домены в «зоне .рф» на самом деле тоже представляют собой последовательности латинских букв, цифр и дефиса, получаемые из русских слов по хитрым правилам, известным как IDN.

[15] Теоретически можно использовать и один сервер, но из соображений надёжности так делать не рекомендуется, ведь любой компьютер хотя бы иногда приходится перезагружать.

```
$TTL 86400
@       IN      SOA     ns2.croco.net. netop.croco.net. (
                        4
                        28800
                        7200
                        604800
                        86400   )
        IN      NS      ns2.croco.net.
        IN      NS      ns3.croco.net.
        IN      NS      ns.intelib.org.
@       IN      MX 5    reptile.croco.net.
@       IN      A       195.42.170.129
www     IN      CNAME   varan103.croco.net.
```

Рис. 6.7. Файл доменной зоны `stolyarov.info`

Администратор домена может принять решение *делегировать* обслуживание поддомена своего домена другой группе серверов. В этом случае за обслуживание запросов по доменным именам, входящим в такой поддомен, будет отвечать уже новая группа серверов, а не та, которая обслуживает исходный домен. Домен за вычетом поддоменов, делегированных на другие сервера, называется ***доменной зоной***. Правильнее говорить, что авторитативные DNS-сервера обслуживают не домен как таковой, а *доменную зону*.

Доменная зона обычно формируется в виде текстового файла, содержащего *записи*. Из этих записей состоит база данных DNS, а функционирование DNS как системы состоит в том, чтобы по заданному доменному имени найти нужный авторитативный сервер и получить от него запись нужного типа либо несколько таких записей — большинство типов допускают множественность. Пример файла доменной зоны приведён на рис. 6.7. Поясним, что символом «`@`» обозначается сам домен, для которого написан файл зоны, а все прочие доменные имена в файле зоны считаются относительными (то есть к ним добавляется имя текущего домена), если только в конце не поставить точку. Собственно говоря, в рассматриваемом файле точка стоит в конце всех имён, кроме «`www`» (поскольку здесь имеется в виду `www.stolyarov.info`). Загадочное `IN` в каждой строке указывает на *класс* записей (напомним, что класс `IN` — единственный, который реально используется в DNS).

Самая первая строка нашего файла, начинающаяся с совсем уж странно выглядящего `$TTL`, задаёт для записей из этой зоны *время жизни* (англ. *time to live*) по умолчанию (в секундах, т. е. в нашем примере это время равно суткам). Выше мы упоминали, что задать время жизни можно при желании для каждой записи отдельно, но такими тонкостями обычно никто не заморачивается, довольствуясь одним значением для всей зоны.

Запись `SOA` (*start of authority*) означает, что для данного имени имеется отдельный файл зоны (в нашем случае это так и есть). Записи `NS` (*name server*) указывают доменные сервера, обслуживающие данный домен; на самом деле набор серверов, обслуживающих домен, задаётся записями в вышестоящей зоне (в данном случае — в зоне `info`), и технически используются именно записи `NS` из вышестоящей зоны, но считается хорошим тоном поддерживать в своей собственной зоне тот же набор NS-записей. Запись `MX` задаёт машину, обрабатывающую электронную почту для данного доменного имени, запись `A` (для имени `@`, что означает `stolyarov.info`) указывает ip-адрес, который соответствует доменному имени `stolyarov.info`, а запись `CNAME` для имени `www` (то есть `www.stolyarov.info`) предписывает рассматривать это имя как *синоним* другого имени (на самом деле это другое имя — `varan103.croco.net` — тоже преобразуется в тот же самый ip-адрес).

Упомянем ещё несколько интересных возможностей файлов доменных зон. Вместо младшего токена в доменном имени можно указать звёздочку «`*`»; это будет означать, что указанную запись следует отдавать в ответ на запросы, в которых вместо звёздочки может стоять что угодно. Так, если бы мы внесли в наш примерный файл зоны запись вроде

```
*.camp      IN    A    198.51.100.72
```

— то ip-адрес `198.51.100.72` система DNS стала бы отдавать в ответ на запросы `www.camp.stolyarov.info`, `vasya.camp.stolyarov.info`, `abrakadabra.camp.stolyarov.info` и т. п.

Для делегирования поддомена достаточно занести в зону исходного домена соответствующие записи. Например, чтобы делегировать домен третьего уровня `example.stolyarov.info` серверам `ns1.example.com` и `ns2.example.com`, достаточно в доменную зону `stolyarov.info` внести такие строки:

```
example     IN    NS   ns1.example.com.
example     IN    NS   ns2.example.com.
```

С делегированием поддоменов связана довольно забавная проблема, неизменно вызывающая трудности у начинающих сисадминов. Если доменное имя (домен $(N+1)$-го уровня) делегируется на DNS-сервера, имена которых сами располагаются в делегируемом поддомене, то для этих серверов требуется наличие записей типа `A` в вышестоящей зоне. Так, домен `stolyarov.info` делегирован на DNS-сервера, имена которых расположены за его пределами; если бы его нужно было делегировать на сервера, имена которых сами находятся в этом домене — например, на сервера с именами `ns1.stolyarov.info` и `ns2.stolyarov.info` — то для этих имён пришлось бы в доменной зоне `info` расположить записи типа `A`. Объясняется это очень просто: *чтобы обратиться к DNS-серверам, нужно знать, на каких ip-адресах они находятся*; если бы в вышестоящей доменной зоне не было соответствующих записей, обращение к авторитативным серверам для делегированного поддомена оказалось бы невозможным. Такие записи называются *склеивающими* (*glue records*).

Регистрируя доменные имена, мы чаще всего имеем дело с доменами *второго уровня*. Домены первого уровня, такие как `info`, `com`, `net`, `ru` и т. п, *тоже делегируются*; их делегирование выполняется на так называемых *корневых серверах*, обслуживающих *корневой домен* (точнее, *корневую зону*). Текущий список корневых серверов — это некая общедоступная информация, необходимая для первичной настройки любого DNS-сервера. К счастью, изменения в список корневых серверов вносятся достаточно редко, причём если один из серверов меняет свой адрес, остальные ещё несколько лет остаются работать на прежних адресах, так что даже если администратор DNS-сервера не уследил за новостями, его сервер долгое время не утратит работоспособности.

Сейчас корневой домен обслуживают 13 «логических» серверов: `a.root-servers.net`, `b.root-servers.net`, ..., `m.root-servers.net`. Каждому из этих имён приписан как адрес протокола IPv4, так и адрес протокола IPv6. При этом все эти сервера, за исключением одного (под буквой `b`), на самом деле представляют собой целое множество «реплик» — физических компьютеров, размещённых в самых разных точках мира. Так, на момент написания первой версии этого текста (в 2017 году) в Москве были размещены реплики корневых серверов «`F`», «`J`», «`K`» и «`L`», в Санкт-Петербурге — реплики «`I`», «`J`» и «`K`», а больше всего реплик по всему миру — 158 — поддерживалось для сервера «`L`».

Конечно, если бы каждый DNS-запрос проходил через корневые сервера, Интернет не спасли бы никакие ухищрения — с такой нагрузкой не справились бы и десятки тысяч машин. К счастью, запросы, обращённые к корневым серверам — на самом деле не такое уж частое событие. Дело в том, что каждый DNS-сервер, обслуживающий пользователей, *запоминает* все записи, которые через него проходят, и помнит их, пока для них не истечёт установленное время жизни *time-to-live*. Если DNS-серверу потребовалась информация в каком-нибудь домене второго уровня, кончающемся на `com`, то ему придётся обратиться к корневым серверам, чтобы узнать, какие сервера обслуживают домен `com`; время жизни для этих записей установлено в 48 часов, так что эта информация останется в памяти сервера надолго: если не перезапускать его, то следующее обращение к корневым серверам за списком серверов домена `com` раньше чем через двое суток не произойдёт. То же самое можно сказать и про другие домены первого уровня, ну а самих этих доменов не так много — всего несколько сотен, и обычно пользователи одного отдельно взятого DNS-сервера к большинству доменов первого уровня не обращаются.

6.3. Система сокетов в ОС Unix

Две предыдущие главы позволили нам составить общее впечатление о компьютерных сетях; теперь настало время вернуться к интерфейсу системных вызовов ОС Unix и научиться писать программы, использующие сеть для взаимодействия друг с другом и с другими программами. Системы семейства Unix (а в современных условиях — вообще все существующие операционные системы) предоставляют пользовательским задачам доступ к возможностям сети через так называемые ***сокеты*** или, говоря строже, через системные вызовы, спецификация которых использует понятие сокета. Изучением таких системных вызовов мы и займёмся в этой главе.

6.3.1. Семейства адресации и типы взаимодействия

Дать строгое определение сокета, как водится, совершенно невозможно. Ограничимся замечанием, что сокет (англ. *socket*) — это объект ядра операционной системы, через который происходит сетевое взаимодействие[16]. На сокет как абстрактный объект имеются две совершенно разные точки зрения. Процесс, создавший (или использующий) сокет для связи с другими процессами, видит его как нечто, «живущее» в ядре ОС и определённым образом реагирующее на системные вызовы. Все остальные участники сетевого взаимодействия воспринимают сокет как этакую (естественно, сугубо абстрактную) «амбразуру», видимую извне системы, на которой «написан» определённый адрес и через которую можно взаимодействовать с кем-то, работающим внутри системы.

В ОС Unix с сокетом связывается файловый дескриптор; в частности, работа с сокетом может быть завершена обычным вызовом `close`. Для идентификации сокетов (или, точнее, абонентов связи) в сетях используются ***адреса***. В зависимости от используемых протоколов адреса могут выглядеть совершенно по-разному. Так, обсуждая в §6.2.4 идентификацию абонентов сетевого взаимодействия при использовании протоколов UDP и TCP, мы ввели понятие *сетевого порта*, который позволяет идентифицировать участников взаимодействия внутри одной системы, тогда как сама система идентифицируется ip-адресом. **Эта пара — ip-адрес плюс номер порта — как раз и составляет сетевой адрес для сокета при использовании протоколов семейства IP**.

Очевидно, что если произойдёт чудо и Интернет всё же перейдёт на протокол IPv6, то для представления сетевых адресов потребуются совершенно иные структуры данных: в самом деле, для ip-адреса нужно будет уже не 4 байта, а 16. С другой стороны, в сетях, построенных по технологии компании Novell и вышедших из употребления сравнительно недавно, использовался стек протоколов IPX/SPX. В рамках этих протоколов адрес сокета состоит из трех частей: 4-байтного номера сети, 6-байтного номера машины (хоста) и 2-байтного номера сокета. Существуют и другие семейства протоколов. Кроме того, отдельный специальный вид сокетов предназначен для связи процессов в рамках одной машины; в качестве адресов такие сокеты используют имена специальных файлов в файловой системе.

Несмотря на такие различия, между разными видами взаимодействия очень много общего. В любом случае было бы категорически неприемлемо модифицировать интерфейс ядра операционной системы и переделывать всё прикладное программное обеспечение при добавлении поддержки очередного семейства протоколов. Именно поэтому введена подсистема сокетов — своего рода общий знаменатель для всех возможных видов сетевого взаимодействия процессов. При создании сокета указывается, к какому ***семейству адресации*** (англ. *address family*)[17] новый сокет будет принадлежать. Набор поддерживаемых семейств может быть расширен добавлением соответствующих модулей в ядро и написанием прикладных программ, работающих с новыми адресами; при этом системные вызовы останутся прежними, а значит, не придётся переделывать системные библиотеки и программы, не использующие новые протоколы.

Кроме используемого семейства адресации, при создании сокета нужно указать ***тип взаимодействия***. Обсуждая в §6.2.4 протоколы транспортного уровня, мы упомянули два основных типа взаимодействия — *дейтаграммный* и *потоковый*; их мы и будем рассматривать. Ядра операционных систем (в том числе Linux и FreeBSD) обычно поддерживают сокеты с другими типами взаимодействия, но некоторые из них недоступны в семействе протоколов TCP/IP (то есть в большинстве современных компьютерных сетей), другие оказываются особенностью конкретного ядра, а их рассмотрение потребовало бы рассказа о достаточно специфических свойствах системы.

Напомним, что при *дейтаграммном взаимодействии* на сокете доступны две основные операции: передача и приём пакета данных (*дейтаграммы*), причём размер пакета, вообще говоря, ограничен; в частности, для UDP ограничение максимальной длины дейтаграммы обусловлено максимальной длиной пакета, передаваемого через сетевые соединения (так называемый *maximal transfer unit*, MTU), которое в большинстве случаев составляет 1500 байт, что, за вычетом длины заголовков IP и UDP, оставляет на саму дейтаграмму 1432 байта. Никто, с другой стороны, не гарантирует, что MTU на всех сетевых соединениях, которые встретятся на пути вашего пакета, будет составлять именно 1500 байт. В RFC791, определяющем протокол IPv4, содержится рекомендация для всех сетевых хостов принимать пакеты длиной хотя бы 576 байт («октетов», т. е. групп по восемь бит), что оставляет на дейтаграмму 508 байт; дейтаграммы бóльшего размера рекомендовано отправлять только при наличии веских оснований для предположения о достаточности размера MTU по всему пути следования пакета. Переданный пакет может быть потерян или, наоборот, случайно сдублирован, то есть получено будет два или более одинаковых пакета; два переданных пакета могут прийти получателю в обратном порядке. Как уже говорилось, при таком режиме работы обеспечение надёжности ложится на пользовательскую программу (приложение).

Потоковый тип взаимодействия предоставляет прикладному программисту иллюзию надёжного двунаправленного канала передачи данных. Данные могут быть записаны в канал порциями любого размера; гарантируется, что на другом конце данные либо будут получены без потерь и в том же порядке, либо не будут получены вообще: соединение в этом случае будет разорвано с фиксацией ошибки. При использовании потокового взаимодействия заботу о передаче подтверждений, о расстановке пакетов в исходном порядке, о повторной передаче потерянных пакетов и т. п. берёт на себя операционная система.

Сокет в ОС Unix создаётся с помощью системного вызова, который так и называется `socket`:

```
int socket(int family, int type, int protocol);
```

Параметр `family` задаёт используемое семейство адресации. Мы будем рассматривать два из них: `AF_INET` для взаимодействия по сети посредством протоколов TCP/IP (адрес сокета в этом случае представляет собой пару ip-адрес/порт) и `AF_UNIX` для взаимодействия в рамках одной машины (в этом случае адрес сокета представляет собой имя файла). Параметр `type` задаёт тип взаимодействия; константа `SOCK_STREAM` соответствует потоковому взаимодействию, константа `SOCK_DGRAM` — дейтаграммному. Наконец, последний параметр задаёт конкретный используемый протокол. Для рассматриваемых нами двух семейств адресации и двух типов взаимодействия протокол однозначно определяется значениями первых двух параметров, так что в качестве этого параметра всегда можно указать число `0`. Можно, впрочем, указать и значение конкретного протокола: `IPPROTO_TCP` для TCP и `IPPROTO_UDP` для UDP.

Вызов возвращает `-1` в случае ошибки; в случае успеха возвращается номер файлового дескриптора, связанного с созданным сокетом.

6.3.2. Сокет и его сетевой адрес

Для того, чтобы с сокетом мог установить соединение (или отправить ему дейтаграмму) другой участник сетевого взаимодействия, нужно, очевидно, некое средство идентификации — ***адрес сокета***, который, как уже говорилось в предыдущем параграфе, в зависимости от используемого семейства адресации может иметь существенно разный вид. Снабдить сокет адресом можно с помощью системного вызова `bind`:

```
int bind(int sockfd, const struct sockaddr *addr, int addrlen);
```

— где `sockfd` — дескриптор сокета, полученный в результате выполнения вызова `socket`; `addr` — указатель на структуру, содержащую адрес; наконец, `addrlen` — размер структуры адреса в байтах.

Вызов `bind` возвращает `0` в случае успеха, `-1` в случае ошибки. Учтите, что существует множество ситуаций, в которых вызов `bind` может не пройти; например, ошибка произойдет при попытке использования привилегированного номера порта (от `1` до `1023`) или порта, который на данной машине уже кем-то занят (возможно, другой вашей программой). Поэтому обработка ошибок при вызове `bind` особенно важна.

Реально в качестве параметра `addr` используется не структура типа `sockaddr`, а структура другого типа, который зависит от используемого семейства адресации. В семействе `AF_INET` используется структура `struct sockaddr_in`, умеющая хранить пару «ip-адрес + порт». Эта структура имеет следующие поля:

[16] Это нельзя считать определением сокета хотя бы по той причине, что основанное на сокетах взаимодействие не обязательно происходит по сети, а взаимодействие по сети бывает основано не только на сокетах — во всяком случае, раньше существовали и другие программные интерфейсы для работы с сетью, такие как STREAMS.

[17] Часто используется также термин *семейство протоколов* (англ. *protocol family*). Это синонимы.

- `sin_family` — обозначает семейство адресации;
- `sin_port` — задаёт номер порта в *сетевом порядке байтов*, который может отличаться от порядка байтов, используемого на данной машине;
- `sin_addr` — задаёт ip-адрес; поле `sin_addr` само является структурой, имеющей лишь одно поле с именем `s_addr`, которое хранит ip-адрес в виде беззнакового четырёхбайтного целого, причём, опять же, в сетевом порядке байтов.

Попробуем понять, что тут к чему. Первое поле, которое в структуре `sockaddr_in` называется `sin_family` — это единственное общее поле для всех существующих структур, используемых для задания адреса сокета. Называется оно, впрочем, по-разному в разных структурах; это отголоски далёкой эпохи, когда в языке Си в разных структурах нельзя было использовать одинаковые имена полей. Так или иначе, в это поле всегда заносится идентификатор используемого семейства адресации; в данном случае это `AF_INET`.

Предназначение оставшихся двух полей как будто бы достаточно очевидно, ведь мы уже знаем, что при работе в сети Интернет участник сетевого соединения идентифицируется ip-адресом и ***портом*** (см. §§6.2.3, 6.2.4); мы даже знаем, что ip-адрес обычно приписывается *сетевому интерфейсу*, так что некоторые компьютеры обладают более чем одним ip-адресом (хотя большинство компьютеров — именно что одним), но в любом случае количество процессов, участвующих в сетевых взаимодействиях, может превышать количество ip-адресов на данном конкретном компьютере; отсюда потребность в использовании портов — абстрактных беззнаковых двухбайтных чисел, дополняющих ip-адреса. Но что за загадочный «сетевой порядок байтов»?

Как мы знаем из вводной части (см. т. 1, §1.4.2), порядок байтов в представлении целых чисел в памяти может варьироваться от одной архитектуры к другой. Архитектуры, в которых старший байт числа имеет наименьший адрес (иначе говоря, байты расположены в *прямом* порядке), в англоязычной литературе обозначаются термином *big-endian*, а архитектуры, в которых наименьший адрес имеет младший байт, то есть порядок байтов *обратный*, — *little-endian*. Чтобы сделать возможным взаимодействие по сети между машинами, имеющими разные архитектуры, принято соглашение, что передача целых чисел по сети всегда производится в прямом (*big-endian*) порядке байтов, т. е. старший байт передается первым. Чтобы обеспечить переносимость программ на уровне исходного кода, в операционных системах семейства Unix введены стандартные библиотечные функции для преобразования целых чисел из формата данной машины (*host byte order*) в сетевой формат (*network byte order*) и обратно. На машинах, архитектура которых предполагает порядок байтов, совпадающий с сетевым, эти функции просто возвращают свой аргумент, в ином случае они производят нужные преобразования. Вот эти функции:

```
unsigned int htonl(unsigned int hostlong);
unsigned short int htons(unsigned short int hostshort);
unsigned int ntohl(unsigned int netlong);
unsigned short int ntohs(unsigned short int netshort);
```

Буква `n` в названиях этих функций означает *network* (т. е. сетевой порядок байтов), буква `h` — *host* (порядок байтов данной машины); `s` соответствует «коротким», а `l` — «длинным» целым числам, причём под короткими понимаются действительно числа типа `short` (16-битные), а вот под «длинными» — так уж сложилось — 32-битные числа; и те и другие — беззнаковые. Например, функция `ntohl` используется для преобразования 32-битного целого из сетевого порядка байтов в порядок байтов, используемый на данной машине. Номер порта представляет собой двухбайтное целое — такое целое, которое большинство компиляторов Си (если не все) называют «коротким целым» (`short int` или просто `short`). Итак, для заполнения поля `sin_port` нам потребуется функция `htons`; например, если мы решили запустить сервер, использующий порт `7654`, то выглядеть это будет так:

```
struct sockaddr_in addr;
/* ... */
addr.sin_port = htons(7654);
```

Отметим ещё один немаловажный момент, который нужно будет учесть, если вы решите поэкспериментировать с серверными сокетами. Как для протокола TCP, так и для протокола UDP **номера портов от 1 до 1023 включительно считаются привилегированными**; такой порт может забрать себе только процесс, имеющий полномочия администратора (`euid == 0`). Если обычный процесс попытается задать сокету адрес с привилегированным номером порта, вызов `bind` вернёт `-1`, а `errno` получит значение `EACCES`.

Что касается ip-адреса, то, имея его текстовое представление (например, строку `"192.168.10.12"`), можно воспользоваться функцией `inet_aton` для формирования структуры типа `struct in_addr` (поле `sin_addr` структуры `sockaddr_in` имеет именно этот тип):

```
int inet_aton(const char *cp, struct in_addr *inp);
```

Здесь `cp` — строка, содержащая текстовое представление ip-адреса, а `inp` указывает на структуру, подлежащую заполнению. Функция возвращает ненулевое значение, если заданная строка является допустимой текстовой записью ip-адреса, и `0` в противном случае. Допустим, нужный нам ip-адрес содержится в строке `char *serv_ip`, а порт — в переменной `port` в виде целого числа. Тогда заполнение структуры `sockaddr_in` может выглядеть так:

```
int ok;
struct sockaddr_in addr;
addr.sin_family = AF_INET;
addr.sin_port = htons(port);
ok = inet_aton(serv_ip, &(addr.sin_addr));
if(!ok) {
    /* Ошибка - невалидный ip-адрес */
}
```

Можно также воспользоваться функцией `inet_addr`, которая принимает один аргумент — строковое представление адреса, а возвращает четырёхбайтное целое число, которое можно занести в поле `sin_addr.s_addr`:

```
unsigned int inet_addr(const char *cp);
```

Эта функция возвращает адрес уже в сетевом порядке байтов, так что никаких дополнительных преобразований не нужно. К сожалению, с ней имеется небольшая проблема: если подать ей на вход белиберду, она вернёт значение `(unsigned) -1`, но то же самое она вернёт, если подать ей строку `"255.255.255.255"` — а этот «адрес» в некоторых специфических случаях всё-таки встречается.

В большинстве случаев задавать созданному нами сокету конкретный ip-адрес не обязательно, да это и не так легко сделать: как ни странно, в системах семейства Unix нет простого и переносимого способа узнать, какие ip-адреса имеются в системе. Когда нам нужен серверный сокет, проще проинструктировать систему принимать соединения (или дейтаграммы) *на заданный порт на любом из имеющихся в системе ip-адресов*, а при отправке исходящих дейтаграмм или запросов на установление соединения использовать ip-адрес того сетевого интерфейса, через который происходит работа. Для этого поле `sin_addr` следует заполнить специальным значением `INADDR_ANY`:

```
struct sockaddr_in addr;
addr.sin_family = AF_INET;
addr.sin_port = htons(port);
addr.sin_addr.s_addr = htonl(INADDR_ANY);
```

На самом деле `INADDR_ANY` — это просто арифметический ноль, или ip-адрес `0.0.0.0`. При подготовке к вызову `bind` в большинстве случаев ip-адрес, т. е. значение поля `sin_addr`, как раз и заполняют с помощью константы `INADDR_ANY`; функции `inet_addr` и `inet_aton` чаще используют при формировании адреса *получателя* дейтаграммы или запроса. Вообще-то применение функции `htonl` к арифметическому нулю не обязательно, поскольку все байты этого числа нулевые и совершенно всё равно, как их переставлять; но системные заголовочные файлы предоставляют целый набор констант с префиксом `INADDR_`, и многие из этих констант зависят от порядка байтов, так что лучше выработать привычку всегда применять к ним `htonl`.

Полезно помнить, что структура `sockaddr_in` описана в заголовочном файле `netinet/in.h`. Как ни странно, эту информацию довольно тяжело отыскать в тексте страниц `man`'а.

Поскольку тип структуры `struct sockaddr_in` формально отличается от типа второго параметра вызова `bind`, при обращении к этому вызову приходится прибегать к явному приведению типа, например:

```
res = bind(sockfd, (struct sockaddr*)&addr, sizeof(addr));
if(res == -1) {
    /* обработка ошибки */
}
```

В семействе `AF_UNIX` используется структура `struct sockaddr_un`, в которой можно хранить имя файла. Эта структура описана в заголовочном файле `sys/un.h`; она состоит из двух полей:

- `sun_family` — обозначает семейство адресации (в данном случае значение этого поля должно быть установлено в `AF_UNIX`);
- `sun_path` — массив на 108 символов, в который непосредственно записывается строка имени файла.

Файл, имя которого используется в качестве адреса сокета, должен иметь специальный тип «сокет»; это отдельный ***тип файла*** наряду с обычными файлами, директориями, символическими ссылками, символьными и блочными устройствами и именованными каналами; поскольку всего типов, как уже говорилось, семь, можно заметить, что теперь мы знаем их все. Если файла с заданным именем не существует, вызов `bind` попытается его создать; при этом у вашего процесса может не хватить полномочий для создания файла (например, у вас может не быть прав на запись в директорию, в которой вы пытаетесь создать сокет), что приведёт к ошибке. Следует также учитывать, что файл сокета сам по себе не исчезает при завершении работы с сокетом; его нужно удалить явным образом с помощью системного вызова `unlink`.

Ещё один интересный момент, связанный с сокетами, касается прав доступа к ним. В ОС Linux права доступа к сокету используются, хотя и не совсем так, как для файлов других типов: для того, чтобы связаться с сокетом, процесс должен иметь полномочия на запись в него. При этом некоторые другие системы, в том числе BSD, игнорируют права доступа, установленные на файле сокета.

Для других семейств адресации системные заголовочные файлы предоставляют свои структуры, имена которых тоже имеют вид `sockaddr_XX`. Структуры этих типов нужны не только для вызова `bind`, они используются везде, где происходит работа с адресом сокета. С некоторыми такими случаями мы познакомимся позже.

6.3.3. Приём и передача дейтаграмм

Рассмотрим работу с сокетами дейтаграммного типа. Сокет создаётся вызовом `socket` с указанием константы `SOCK_DGRAM` в качестве второго параметра. Желательно связать сокет с конкретным адресом с помощью `bind`, при этом ip-адрес можно заменить константой `INADDR_ANY`, но порт всё же задать. Если этого не сделать, то при отправке первой дейтаграммы система сама выберет один из своих адресов и портов в качестве адреса отправителя. Это не проблема, если мы собираемся сначала отправить кому-то дейтаграмму (например, в качестве запроса), а потом уже пытаться получить дейтаграмму, которая будет послана нам в качестве ответа; но если мы предполагаем, что кто-то другой может послать нам дейтаграмму по своей инициативе, а не в ответ на наш запрос, то, очевидно, наш сокет должен быть связан с конкретным адресом, и этот адрес должен быть известен нашим партнёрам.

Для передачи и приёма данных предназначены системные вызовы `sendto` и `recvfrom`:

```
int sendto(int sockfd, const void *buf, int len, int flags,
           const struct sockaddr *to, socklen_t tolen);
int recvfrom(int sockfd, void *buf, int len, int flags,
           struct sockaddr *from, socklen_t *fromlen);
```

В обоих вызовах параметр `sockfd` задаёт дескриптор сокета, `buf` указывает на буфер, содержащий данные для передачи либо предназначенный для размещения принятых данных, `len` устанавливает размер этого буфера (соответственно, количество данных, подлежащих приему или передаче). Параметр `flags` используется для указания дополнительных опций; для нормальной работы обычно достаточно указать значение `0`.

В вызове `sendto` параметр `to` указывает на структуру, содержащую адрес сокета, на который нужно отправить данные (то есть адрес получателя сообщения). Ясно, что используется при этом структура типа, соответствующего избранному семейству адресации: `sockaddr_in` для `AF_INET`, `sockaddr_un` для `AF_UNIX` и т. д. Параметр `tolen` должен быть равен размеру этой структуры. Тип `socklen_t` может быть описан в системных заголовочных файлах как синоним типа `int` или `unsigned int`; при использовании `sendto` можно не обращать внимания на идентификатор `socklen_t`, но вызов `recvfrom` требует *адреса* переменной такого типа, и если не угадать с его знаковостью, получится ошибка несоответствия типов. Во избежание лишних проблем проще описать нужную переменную с использованием типа `socklen_t`.

Заодно можно за весь этот маразм «поблагодарить» комитет, создававший POSIX, который, кстати, чуть было не сломал всё, что можно, первоначально установив, что параметр `fromlen` должен иметь тип «указатель на `size_t`»; проблема тут в том, что в силу ряда причин переменная, используемая в этой роли, обязана совпадать по своей разрядности с `int`, а `size_t` на многих архитектурах имеет бóльшую разрядность. Линус Торвальдс по этому поводу утверждал, что лишь после крайне «громких жалоб», исходивших в том числе и от него самого, комитет отказался от мысли использовать здесь `size_t`, но зато придумал совершенно никому не нужный `socklen_t`, просто чтобы сохранить лицо.

В вызове `recvfrom` параметр `from` указывает на структуру, в которую вызову следует записать адрес отправителя полученного пакета, т. е. этот параметр позволяет узнать, откуда пакет пришёл. Параметр `fromlen` представляет собой указатель на переменную типа `socklen_t`, причём перед вызовом `recvfrom` в эту переменную следует занести размер адресной структуры, на которую указывает предыдущий параметр. После возврата из `recvfrom` переменная будет содержать количество байтов, которые вызов в итоге в эту структуру записал.

Как уже отмечалось, при использовании протокола UDP передан может быть только пакет ограниченного размера. Конкретный предельный размер дейтаграммы, вообще говоря, может оказаться различным для различных адресов получателей; если особенности решаемой задачи не позволяют удовлетвориться «гарантированным» размером дейтаграммы в 508 байт, придётся применить достаточно сложные процедуры динамического определения допустимого размера пакета. Их мы описывать не будем; при желании читатель может самостоятельно изучить эту область, обратившись, например, к книге [5].

При работе с дейтаграммами следует учитывать один важный момент. Можно считать типичным сценарий работы, при котором наша программа отправляет кому-то дейтаграмму и затем ожидает ответа (тоже в виде дейтаграммы). «Наивная» реализация этого довольно очевидна: создать сокет, сформировать дейтаграмму, отправить её с помощью `sendto`, после чего вызвать `recvfrom`. К сожалению, такая реализация никуда не годится. Дело в том, что дейтаграммы, как уже неоднократно отмечалось, ненадёжны. Наша исходная дейтаграмма может потеряться, так и не дойдя до нашего партнёра, так что он не будет знать, что мы ждём от него ответа; либо может потеряться ответ. В обоих случаях наша программа так и останется в вызове `recvfrom`; внешне это будет выглядеть так, будто она зависла.

Правильным здесь будет более сложное решение, подразумевающее установку (тем или иным способом) ограничения на время ожидания ответной дейтаграммы (тайм-аута). Самый простой способ введения такого ограничения нам уже известен: мы можем предусмотреть в программе обработчик сигнала `SIGALRM`, после чего попросить операционную систему прислать нам этот сигнал, скажем, через пять секунд (см. §5.3.14), и лишь после этого обращаться к `recvfrom`.

Существуют и другие способы ограничить время ожидания ответной дейтаграммы. В §6.4.3 мы будем рассматривать вызов `select`, который позволяет дождаться одного из нескольких предопределённых событий; в нашем случае в качестве таких событий могут выступать получение дейтаграммы на заданном сокете и истечение установленного тайм-аута. Кроме того, возможность установки тайм-аута для `recvfrom` предусмотрена в самой подсистеме сокетов. Для этого нужно, используя системный вызов `setsockopt`, установить на уровне `SOL_SOCKET` опцию `SO_RCVTIMEO`, например, так:

```
struct timeval tv;
tv.tv_sec = 3;            /* полные секунды */
tv.tv_usec = 500000;      /* микросекунды */
setsockopt(sockfd, SOL_SOCKET, SO_RCVTIMEO, &tv, sizeof(tv));
```

Здесь мы установили для сокета `sockfd` тайм-аут на чтение (получение дейтаграммы) в 3,5 секунды. Подробности об этом способе можно найти, например, дав команду «`man 7 socket`». С вызовом `setsockopt` мы снова встретимся в §6.3.5.

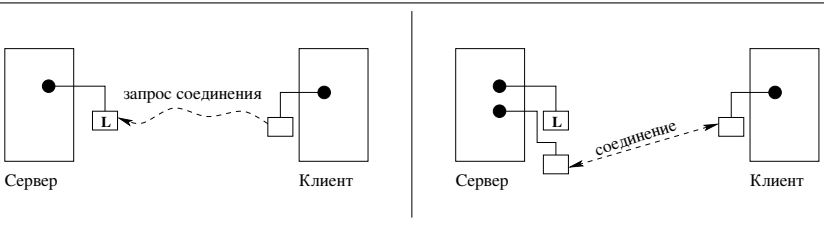

Рис. 6.8. Установление соединения между потоковыми сокетами

6.3.4. Потоковые сокеты. Клиент-серверная модель

При взаимодействии с помощью потоковых сокетов нужно перед началом взаимодействия установить ***соединение***. Ясно, что если речь идет о взаимодействии процессов не только не родственных, но и, возможно, находящихся на разных машинах, один из участников взаимодействия должен быть инициатором соединения, а второй — принять соединение (согласиться на его установление). Здесь мы вновь сталкиваемся с уже знакомыми нам понятиями ***клиента*** и ***сервера***. При установлении соединения между потоковыми сокетами один процесс ожидает запроса на установление соединения, а другой инициирует такой запрос. Эти процессы с точки зрения установления соединения и называются сервером и клиентом; в частности, при работе по сети Интернет для организации взаимодействия потоковых сокетов используется протокол TCP, а соответствующие программы называются TCP-сервером и TCP-клиентом.

Чтобы начать ожидание запросов на соединение, сервер создаёт сокет соответствующего типа, связывает его с адресом и переводит в специальное состояние, называемое ***слушающим*** (англ. *listening*). На сокете, находящемся в слушающем состоянии, может быть выполнена только одна операция — ***приём соединения***. При установлении соединения ядро операционной системы, которая обслуживает программу-сервер, создаёт еще один сокет, который и будет использоваться для передачи данных по только что установленному соединению (рис. 6.8).

Итак, на стороне сервера сокет нужно создать вызовом `socket` с соответствующими параметрами и связать его вызовом `bind` с конкретным адресом, на котором будут приниматься соединения. Затем сокет следует перевести в слушающий режим (на рис. 6.8 слушающий сокет обозначен буквой `L`) с помощью вызова `listen`:

```
int listen(int sockfd, int qlen);
```

Параметр `sockfd` — связанный с сокетом файловый дескриптор. Параметр `qlen` задаёт размер очереди непринятых запросов на соединение. Поясним это. Допустим, мы перевели сокет в слушающий режим, и несколько клиентов уже отправили нам запросы на соединение, но в силу тех или иных причин мы некоторое время не принимаем эти запросы, то есть не выполняем соответствующую операцию со слушающим сокетом. Возможности системы по хранению необработанных запросов ограничены. Первые `qlen` запросов будут ожидать принятия соединения, если же система получит ещё запросы, они будут отклонены. Следует особо подчеркнуть, что параметр `qlen` не имеет никакого отношения к общему количеству соединений с клиентами.

Сказанное, впрочем, не позволяет дать ответ на очень простой и весьма актуальный вопрос: так *какое же* число следует передать вторым параметром вызова `listen`? В примерах часто встречается значение `5`, имеющее исторические причины: ядра некоторых операционных систем не позволяли создавать очередь бóльшего размера. В современных условиях такое значение может оказаться недостаточным. К счастью, если использовать «слишком большое» число, ничего страшного не произойдёт: система «молча» заменит его на наибольшее поддерживаемое. Но слишком усердствовать с этим не стоит. Многие современные системы, включая Linux, поддерживают максимально возможное значение `128`; если вашему серверу этого стало не хватать, то дальнейшее увеличение очереди вас уже не спасёт, ваш сервер работает слишком медленно и подкручиванием размеров очереди вы это не исправите. Всё сказанное останется верно и для существенно меньших значений: так, если вы укажете в качестве `qlen` число `16` и его в какой-то момент не хватит, это должно стать скорее поводом для поиска путей возможной оптимизации, а не для увеличения очереди.

Принятие соединения производится вызовом `accept`:

```
int accept(int sockfd, struct sockaddr *addr, socklen_t *addrlen);
```

Параметр `sockfd` задаёт дескриптор слушающего сокета. Параметр `addr` указывает на структуру для записи адреса сокета, с которым установлено соединение (иначе говоря, сюда будет записан адрес другого конца соединения). Параметр `addrlen` представляет собой адрес переменной типа `socklen_t`, причём перед вызовом `accept` в эту переменную следует занести размер адресной структуры, на которую указывает предыдущий параметр; после возврата из `accept` переменная будет содержать количество байт, которые вызов в итоге в эту структуру записал. Это аналогично параметрам `from` и `fromlen` в вызове `recvfrom` (см. стр. 159).

Вызов `accept` возвращает файловый дескриптор нового сокета, созданного специально для обслуживания вновь установленного соединения (либо `-1` в случае ошибки). Если на момент выполнения `accept` запросов на соединение ещё не поступило, вызов блокирует вызвавший процесс и ожидает поступления запроса на соединение, возвращая управление только после того, как такой запрос поступит и соединение будет установлено. Подчеркнём, что **с момента принятия первого соединения в программе-сервере будет два дескриптора, требующих обработки**: дескриптор слушающего сокета, на котором можно принимать новые запросы на соединения, и сокет, соответствующий принятому соединению, с которого требуется читать пришедшие от клиента данные (например, текст запроса). Это создаёт ***проблему очерёдности*** дальнейших действий, которой будет посвящена следующая глава.

Клиентская программа должна, как и сервер, создать сокет вызовом `socket`. Связывать сокет с конкретным адресом в этом случае не обязательно, хотя и возможно; если этого не сделать, система выберет адрес автоматически. Запрос на соединение формируется вызовом `connect`:

```
int connect(int sockfd, const struct sockaddr *addr, int addrlen);
```

Параметр `sockfd` — связанный с сокетом файловый дескриптор. Параметр `addr` указывает на структуру, содержащую адрес сервера, т. е. адрес слушающего сокета, с которым мы хотим установить соединение. Естественно, используется при этом структура типа, соответствующего избранному семейству адресации: `sockaddr_in` для `AF_INET`, `sockaddr_un` для `AF_UNIX`. Параметр `addrlen` должен быть равен размеру этой структуры. Вызов возвращает `0` в случае успеха, `-1` в случае ошибки.

После успешного установления соединения для передачи по нему данных можно использовать уже известные нам вызовы `read` и `write`, считая дескрипторы соединённых сокетов обычными файловыми дескрипторами; на стороне сервера это дескриптор, возвращённый вызовом `accept`, на стороне клиента — дескриптор сокета, к которому применялся вызов `connect`. Для более гибкого управления обменом данными существуют также вызовы `recv` и `send`, отличающиеся от `read` и `write` только наличием дополнительного параметра `flags`. При желании читатель может ознакомиться с этими вызовами самостоятельно, воспользовавшись командой `man` или книгами [3] и [5].

Вызов `connect` можно применять не только к потоковым, но и к дейтаграммным сокетам. Настоящего соединения при этом не устанавливается, вызов влияет только на сам объект сокета в ядре; иначе говоря, в результате такого вызова `connect` никакие пакеты по сети не передаются и никакие удалённые машины никакой информации о возникшем «соединении» не получают. Зато поведение нашего собственного сокета меняется при этом весьма заметно. Сокет намертво «привязывается» к единственному (указанному в вызове `connect`) адресу партнёра по коммуникации. При чтении из сокета мы теперь получаем только дейтаграммы, обратным адресом в которых указан наш партнёр; все прочие полученные дейтаграммы ядро молча сбрасывает. Отправлять дейтаграммы мы теперь тоже можем только на один адрес. Обычно для приёма и передачи дейтаграмм через «соединённые» сокеты используются вызовы `read` и `write` или `send`/`recv`, а не `sendto`/`recvfrom`; их, впрочем, тоже можно использовать, но это довольно бессмысленно.

Завершить работу с дескриптором сокета можно, как уже говорилось, с помощью хорошо знакомого нам вызова `close`. Следует понимать, что в общем случае закрытие дескриптора не означает закрытия соединения и вообще прекращения использования сокета, ведь с одним и тем же сокетом (объектом ядра) может быть связано больше одного дескриптора, причём как в одном процессе, так и в разных. В разных процессах дескрипторы, связанные с одним и тем же потоком ввода-вывода (объектом ядра), появляются при создании нового процесса с помощью `fork`, а в одном процессе такие дескрипторы могут появиться в результате манипуляций с вызовами `dup` и `dup2` (см. §§5.3.4, 5.3.10). В обеих ситуациях в роли потоков ввода-вывода могут,

естественно, оказаться сокеты; если с помощью `close` мы закрываем лишь один дескриптор из нескольких связанных с данным сокетом, то никакого влияния на сокет и его соединение наш вызов `close` не окажет, просто на него будет ссылаться на один дескриптор меньше.

Конечно, если мы закроем *все* дескрипторы, связанные с сокетом (или, что бывает гораздо чаще, закроем *единственный* такой дескриптор), системе всё-таки придётся закрыть соединение и уничтожить объект сокета. Точнее говоря, вызов `close` вернёт управление сразу же, но сокет может исчезнуть не сразу: система сначала передаст «на тот конец» соединения все данные, которые ещё не переданы, и лишь затем, отправив партнёру извещение о закрытии соединения, окончательно ликвидирует сокет как объект ядра. **Когда наш партнёр по соединению закрыл свой сокет, очередной вызов чтения (`read` или `recv`) вернёт нам значение `0`, соответствующее ситуации «конец файла»**.

Несколько иной подход к завершению обмена информацией через установленное соединение реализует системный вызов `shutdown`:

```
int shutdown(int sockfd, int how);
```

Параметр `sockfd` задаёт дескриптор сокета, `how` — что́ именно следует прекратить. Для параметра `how` есть три возможных значения: `SHUT_RD` прекращает приём данных через сокет, `SHUT_WR` — передачу данных, `SHUT_RDWR` закрывает обмен данными в обоих направлениях. Полезно знать, что эти три константы были введены сравнительно недавно, а в более старых программах можно встретить обращения к `shutdown`, в которых параметр `how` задан явным образом в виде целого числа (соответственно `0`, `1` и `2`).

В отличие от `close` **вызов `shutdown` относится к сокету как объекту ядра, а не к отдельному дескриптору**. Это значит, что, если уж мы прекратили передачу информации через сокет в одном или в обоих направлениях, то передавать информацию в соответствующих направлениях не удастся ни через один из дескрипторов, связанных с данным сокетом, сколько бы их ни было. Есть и ещё один крайне важный момент: **вызов `shutdown` не уничтожает ни объект сокета, ни связанные с ним дескрипторы**, так что применение этого вызова не отменяет необходимости последующего закрытия сокета с помощью `close`. Помните, что количество одновременно открытых файловых дескрипторов в системе ограничено.

Когда с помощью `shutdown` прекращается *передача* данных через сокет, система отправляет «на тот конец» все данные, которые были ей переданы до вызова `shutdown`, после чего извещает партнёра о прекращении передачи данных. «На том конце» результатом такого прекращения становится ситуация «конец файла». При этом мы сохраняем возможность получать данные и, по идее, должны их получать до тех пор, пока «конец файла» не возникнет уже на нашем конце. Это позволяет реализовать типичный сценарий обмена информацией: после установления соединения клиент отправляет на сервер некий запрос, после чего делает `shutdown`; сервер получает «конец файла», из чего делает вывод, что запрос кончился, и только после этого анализирует полученный запрос и отвечает на него.

Прекращение *приёма* данных через сокет означает, что в дальнейшем получении данных мы не заинтересованы. Данные, полученные нашей системой «с того конца» до того, как мы обратились к `shutdown`, но которые мы ещё не прочитали, будут при этом сброшены; если система получит новые данные уже после запрета на их получение, то все такие данные будут «тихо проигнорированы».

Обычно при закрытии сокета вызовом `close` система возвращает управление немедленно; при этом, возможно, в буферной памяти ядра всё ещё находятся данные для отправки. По умолчанию в этом случае система передаёт все оставшиеся данные и лишь затем закрывает соединение. Если в это время произойдёт ошибка, программа об этом уже не узнает.

Такое поведение вызова `close` в отношении потоковых сокетов можно изменить, модифицировав с помощью системного вызова `setsockopt` параметр `SO_LINGER`. Помимо варианта «по умолчанию» система предоставляет ещё две возможности: либо при вызове `close` немедленно сбросить все неотправленные данные и разорвать соединение, либо, напротив, заблокировать вызов `close` до тех пор, пока все данные, предназначенные к отправке, не окажутся отправлены и их доставка адресату не будет подтверждена. Для такого варианта система требует указания временно́го периода, в течение которого вызов `close` может оставаться заблокированным; если за это время система не успеет передать все данные и получить подтверждение об их доставке, соединение будет разорвано, а `close` вернёт ошибку.

Технические подробности работы с параметром `SO_LINGER` можно найти в книге [5] и в технической документации — в частности, дав команду «`man 7 socket`».

В статье [13] приведён придирчивый анализ текстов спецификации протокола TCP, на основании чего автор статьи делает вывод, что использования `SO_LINGER` недостаточно для достижения стопроцентной надёжности (либо все данные благополучно переданы и получены «на том конце», либо система сообщает об ошибке). Там же описаны некоторые трюки для повышения степени уверенности в корректной работе TCP. Текст статьи довольно сложен, но при должном приложении усилий понять его вполне возможно.

6.3.5. О «залипании» TCP-порта

Часто при работе с сервером можно заметить, что после завершения программы-сервера её некоторое время невозможно запустить с тем же значением номера порта. Это происходит обычно при некорректном завершении программы-сервера либо если программа-сервер завершается при активных клиентских соединениях. В этих случаях ядро операционной системы некоторое время продолжает считать адрес занятым (находящимся в так называемом состоянии `TIME_WAIT`). Делается это для того, чтобы все пакеты, которые (теоретически) могут быть в пути от вашего сокета к партнёру по соединению или обратно, достигли своей цели или исчезли, так и не дойдя по назначению; считается, что в противном случае такие пакеты, придя на наш компьютер в неподходящий момент, могут помешать работе соединений, которые установит новый процесс, использующий тот же самый адрес сокета.

На практике это выглядит довольно неприятно: в течение ощутимого времени после завершения старого процесса-сервера при попытке запустить его снова вызов `bind` выдаёт ошибку, не позволяя получить порт. Период `TIME_WAIT` может составлять от 30 до 120 секунд; если вы обнаружили ошибку в вашей серверной программе, то за это время вы можете успеть найти ошибочный фрагмент в исходном тексте, исправить его, перекомпилировать программу и попытаться запустить исправленную версию — но не тут-то было; система всё ещё считает порт занятым. Кстати, если вы пренебрежёте проверкой значения, возвращаемого вызовом `bind`, то происходящее с вашей программой будет выглядеть как совершеннейшая мистика (хотя в действительности всё очень просто: начиная с `bind`, почти все вызовы работы с сокетами будут «ошибаться» и, как следствие, ничего не делать).

Эффект «залипшего» порта может превратить отладку серверных программ в натуральную пытку. К счастью, система сокетов позволяет изменить поведение ядра в отношении адресов, находящихся в состоянии `TIME_WAIT`. Для этого нужно перед вызовом `bind` выставить на будущем слушающем сокете опцию `SO_REUSEADDR`. Это делается с помощью системного вызова `setsockopt`:

```
int setsockopt(int sockfd, int level, int optname,
               const void *optval, int optlen);
```

Параметр `sockfd` задаёт дескриптор сокета, `level` обозначает уровень (слой) стека протоколов, к которому имеет отношение устанавливаемая опция (в данном случае это уровень сокетов, обозначаемый константой `SOL_SOCKET`). Параметр `optname` задаёт «имя» (на самом деле, конечно, это номер, или числовой идентификатор) устанавливаемой опции; в данном случае нам нужна опция `SO_REUSEADDR`. Поскольку информация, связанная с нужной опцией, может иметь произвольную сложность, вызов принимает нетипизированный указатель на значение опции и длину опции (параметры `optval` и `optlen` соответственно). Значением опции в данном случае будет целое число `1`, так что следует завести переменную типа `int`, присвоить ей значение `1` и передать в качестве `optval` адрес этой переменной, а в качестве `optlen` — выражение `sizeof(int)`. В итоге наш вызов будет выглядеть так:

```
int opt = 1;
setsockopt(sockfd, SOL_SOCKET, SO_REUSEADDR, &opt, sizeof(opt));
```

6.3.6. Сокеты для связи родственных процессов

В отличие от неименованных каналов, потоковые сокеты подразумевают двунаправленную связь, что делает их в некоторых случаях более удобным средством даже при взаимодействии родственных процессов, но процедура установления соединения, рассмотренная выше, представляется для такой ситуации слишком сложной. В связи с этим в ОС Unix предусмотрен вызов `socketpair`, создающий сразу два сокета, между которыми уже установлено соединение:

```
int socketpair(int family, int type, int protocol, int sv[2]);
```

Параметры `family`, `type` и `protocol` задают соответственно семейство адресации, тип и протокол для создаваемых сокетов. Многие реализации допускают для первого параметра только одно значение — `AF_UNIX`. Что касается параметров `type` и `protocol`, то для них обычно используются значения `SOCK_STREAM` и `0`. Параметр `sv` должен указывать на массив из двух элементов типа `int`, в которые вызов занесёт файловые дескрипторы двух созданных сокетов. Сокеты создаются уже связанными друг с другом, причём оба конца открыты как на чтение, так и на запись. Вызов возвращает `0` в случае успеха, `-1` в случае ошибки. От классического `pipe` результат работы `socketpair` отличается тем, что создаваемое соединение оказывается двухсторонним, в остальном принципы работы с ним совершенно такие же: предполагается, что после создания такой пары сокетов процесс породит одного или более потомков, а передача информации через созданное соединение будет производиться либо между потомком и предком, либо между потомками одного предка (процесса, обратившегося к `socketpair`).

6.4. Проблема очерёдности действий и её решения

6.4.1. Суть проблемы

Ранее мы упоминали, что при работе с сокетами потокового типа после принятия первого соединения в программе-сервере возникает проблема очередности действий. Поясним, в чём она заключается. После первого успешного обращения к вызову `accept` у нашего сервера имеется два дескриптора, а после установления соединения с новыми клиентами их станет ещё больше. На один из дескрипторов может в любой момент прийти запрос на соединение от нового клиента, так что вроде бы нужно вызвать `accept`, чтобы такой запрос принять. Но если мы сделаем `accept`, а никакого запроса на соединение не было, наша программа так и останется внутри вызова `accept`, причём, может быть, надолго: никто ведь не гарантирует, что кто-либо захочет установить с нами новое соединение. В это время на любой из клиентских сокетов, то есть сокетов, созданных предыдущими обращениями к `accept` и отвечающих за соединение с каждым из клиентов, могут прийти данные, требующие обработки; ясно, что пока мы висим внутри вызова `accept`, никакой обработки данных не произойдёт, т. к. мы их даже не прочитаем.

Если, напротив, попытаться произвести чтение из какого-то клиентского сокета, есть риск, что клиент по каким-либо причинам долго не будет нам ничего присылать. Всё это время наша программа будет находиться внутри вызова `read` или `recv`, не выполняя никакой работы, в том числе не принимая новых соединений и не обрабатывая информацию, поступившую от других (уже присоединённых) клиентов.

В ходе активной работы сервера к нежелательной блокировке может привести не только попытка приёма данных (или запросов), но и попытка что-то отправить. В самом деле, пропускная способность сети ограничена; следовательно, заблокировать наш процесс могут не только вызовы `read` и `accept`, но и `write`. К примеру, если один из клиентов прислал нам запрос, в ответ на который мы должны передать ему файл размером в сотню мегабайт, у нас есть вполне реальная возможность растерять всех остальных клиентов за то время, пока этот файл будет передаваться, особенно если пытаться записывать данные в сокет большими порциями.

Проиллюстрировать проблему можно на примере из совсем некомпьютерной области. Представим себе ресторан, в котором каждый столик размещён в отдельной кабинке, так что в каком бы месте мы ни стояли, нельзя одновременно увидеть все столики и входные двери. Представим себе теперь, что на весь ресторан есть только один официант, исполняющий еще и обязанности метрдотеля. В нашей аналогии официант будет играть роль процесса. Сразу после открытия ресторана официант, естественно, подойдет к входным дверям и будет поджидать новых клиентов (вызов `accept`), чтобы встретить их и проводить к столику. Однако как только первые клиенты займут столик и примутся изучать меню, официант окажется перед проблемой: следует ли ему стоять около столика в ожидании, пока клиенты определятся с заказом, или же пойти к дверям проверить, не пришёл ли ещё кто-нибудь. Когда же клиентов за столиками станет много, официанту и вовсе придется тяжело. Ведь каждому клиенту за любым из столиков в любой момент может что-то понадобиться — он может решить заказать ещё что-нибудь, может уронить вилку и попросить другую, может, наконец, решить, что ему пора идти, и потребовать счёт. Кроме того, в любой момент могут прийти новые клиенты, и если их не встретить у дверей, они могут уйти, а ресторан недополучит денег. Напомним, что по условиям нашей аналогии в ресторане нет такого места, откуда было бы видно и столики, и входные двери; точнее говоря, из любого места видно либо один столик (но не больше), либо двери, либо вообще ни то, ни другое.

Самое простое решение, приходящее в голову — это бегать по всему ресторану, подбегая по очереди к каждому из столиков, а также и к дверям, узнавать, что внимание официанта там не требуется и идти, вернее, бежать на следующий круг. Аналогичное решение возможно и для нашего процесса. Можно с помощью вызова `fcntl` перевести все сокеты (и слушающий, и клиентские) в ***неблокирующий режим***; напомним, как это делается (здесь `sd` — дескриптор сокета; с переводом в неблокирующий режим мы уже встречались в §5.2.3, где делали это для обычных дескрипторов):

```
flags = fcntl(sd, F_GETFL);
fcntl(sd, F_SETFL, flags | O_NONBLOCK);
```

Для сокетов в неблокирующем режиме вызовы `read` и `accept` всегда возвращают управление немедленно, ничего не ожидая; если не было данных или входящего соединения, возвращается ошибка. Вызов `write` на неблокирующем сокете также возвращает управление немедленно; если ему не удалось записать ни одного байта, он завершается с ошибкой (т. е. возвращает `-1`), если же хотя бы один байт был записан, вызов считается прошедшим успешно; возвращает он при этом, как всегда, количество записанных байтов. Когда все сокеты настроены как неблокирующие, можно их опрашивать по очереди в бесконечном цикле. Это называется ***активным ожиданием***.

В многозадачных системах такой вариант считается совершенно неприемлемым. Официант из нашей аналогии будет понапрасну уставать, нарезая круги по ресторану (вполне возможно, что за несколько десятков таких кругов от него ничего никому так и не понадобится) и в итоге попросту упадёт без сил. Процесс, бесконечно опрашивающий набор сокетов с помощью системных вызовов, тоже будет вхолостую тратить процессорное время. Конечно, в отличие от официанта процесс не устанет, но ведь процессорное время, расходуемое без пользы, могло бы пригодиться другим задачам. Если в системе установлено ограничение на потребляемое процессом время, сервер может исчерпать свой лимит и будет уничтожен ядром системы. Кроме того, современные компьютеры умеют менять тактовую частоту процессора — и, как следствие, энергопотребление — в зависимости от текущей вычислительной нагрузки, так что жечь процессор вхолостую может оказаться ещё и дорого.

Проблема действительно серьёзна. Существуют серверные программы, которые сутками, а иногда и месяцами находятся в бездействии, и было бы очень странно, если бы всё это время программа, которая совершенно ничего не делает, занимала бы процессор; но даже если сервер активно работает, это нельзя считать поводом для создания бесполезной нагрузки на центральный процессор.

Другое решение (для ресторана) состоит в том, чтобы нанять в ресторан адекватное количество персонала. Во-первых, разумеется, у дверей должен стоять метрдотель, встречая приходящих клиентов и проводя их к свободным столикам. Во-вторых, дорогой ресторан вполне может себе позволить прикрепить к каждому столику своего официанта. Аналогичным образом при написании серверной программы мы можем *создать отдельный процесс* для обслуживания каждого пришедшего клиента.

Если же этот вариант неприемлем — например, если бóльшую часть времени весь персонал все равно простаивает, — можно сделать и иначе. Официант может, нарезав несколько кругов, сообразить, что так дело не пойдёт и, например, подвесить колокольчик к входным дверям, а каждый столик снабдить кнопкой звонка. После этого можно будет спокойно усесться в укромный угол и спать или, например, разгадывать кроссворд, пока один из колокольчиков или звонков не зазвонит. В продолжение аналогии стоит вспомнить, что официанту достаточно часто приходится отрываться от своего кроссворда не потому, что его куда-то позвали, а просто потому что прошло некоторое время; например, если клиент заказал блюдо, которое будет готовиться двадцать минут, официанту следует передать заказ на кухню, а через двадцать минут прийти за готовым блюдом, чтобы отнести его клиенту. Поэтому в дополнение к звонкам следует, видимо, запастись ещё и будильником, который можно будет ставить на определённое время. Аналогичный вариант в программировании называется ***мультиплексированием ввода-вывода***.

Рассмотрим два последних варианта подробнее.

6.4.2. Решение на основе обслуживающих процессов

В этом варианте наш главный процесс выполняет обязанности метрдотеля, находясь бóльшую часть времени в вызове `accept`. Приняв очередное соединение, он порождает процесс для обслуживания этого соединения (аналог официанта, прикреплённого к столику). После этого родительский процесс закрывает сокет клиентского соединения, а порождённый процесс закрывает слушающий сокет. Все обязанности по обслуживанию пришедшего клиента возлагаются на порождённый процесс; после завершения сеанса связи с клиентом этот процесс завершается. Всё это время родительский процесс беспрепятственно продолжает исполнять обязанности «метрдотеля», вызывая `accept`. Соответствующий код будет выглядеть примерно так:

```
enum { listen_queue_len = 16 };
int ls, res, opt;
struct sockaddr_in addr;
ls = socket(AF_INET, SOCK_STREAM, 0);
if(ls == -1) {
    /* ... ошибка ... */
}
opt = 1;
setsockopt(ls, SOL_SOCKET, SO_REUSEADDR, &opt, sizeof(opt));
addr.sin_family = AF_INET;
addr.sin_port = htons(port);
addr.sin_addr.s_addr = htonl(INADDR_ANY);
res = bind(ls, (struct sockaddr*)&addr, sizeof(addr));
if(res == -1) {
    /* ... ошибка ... */
}
listen(ls, listen_queue_len);
for(;;) {
    int cls, pid;
    socklen_t slen = sizeof(addr);
    cls = accept(ls, (struct sockaddr*)&addr, &slen);
    pid = fork();
    if(pid == 0) { /* обслуживающий процесс */
        close(ls);
        /* ...
           работаем с клиентом через сокет cls
           Клиент пришёл с адреса, хранящегося
           теперь в структуре addr
           ...
         */
        exit(0);
    }
    /* родительский процесс */
    close(cls);
    /* проверить, не завершились ли какие-либо
       процессы-потомки (убрать зомби) */
    do {
        pid = wait4(-1, NULL, WNOHANG, NULL);
    } while(pid > 0);
}
```

В этом примере зомби-процессы убираются только в момент запуска нового обслуживающего процесса, и это не совсем хорошо: наш «метрдотель» может пробыть в вызове `accept` в ожидании новых входящих соединений сколь угодно долго, и всё это время зомби, которые могли образоваться из ранее запущенных обслуживающих процессов, так и будут «светиться» в таблице процессов. Правильнее было бы перехватывать в программе сигнал `SIGCHLD` (см. §5.3.14) и очистку от зомби производить в его обработчике, но тогда программа должна быть готова ещё и к тому, что `accept` будет прерван сигналом `SIGCHLD`, вернёт `-1`, в `errno` окажется значение `EINTR` — и это не будет ошибкой. Если все эти изменения применить к фрагменту кода, который мы привели выше, его наглядность, к сожалению, пострадает; поэтому мы оставляем всё это читателю для самостоятельных экспериментов.

6.4.3. Событийно-управляемое программирование

Рассмотрим теперь случай, когда порождение отдельного процесса на каждое клиентское соединение неприемлемо. Так может получиться, если сервер достаточно серьёзно загружен: операция порождения процесса сравнительно дорога, так что при загрузках порядка тысяч соединений в секунду затраты на порождение процессов могут оказаться чрезмерными. Кроме того, между сеансами обслуживания разных клиентов может происходить активное взаимодействие. Например, сервер может поддерживать сетевую компьютерную игру, где действия каждого игрока влияют на ситуацию в общем игровом пространстве и сказываются на других игроках. В этом случае разделение серверной программы на отдельные процессы потребует высоких накладных расходов на взаимодействие между этими процессами; к тому же проблема очерёдности действий встанет снова, только уже в каждом из процессов: действительно, следует ли процессу в конкретный момент времени анализировать изменения в игре или же принимать данные с клиентского сокета?

В обоих случаях возникает потребность обслуживать всех клиентов силами одного процесса, причём активное ожидание, естественно, приемлемым не считается. На проблему можно взглянуть шире. Есть некоторое количество типов ***событий***, каждый из которых требует своей обработки. Некоторые системные вызовы, предназначенные для обработки событий, являются *блокирующими*, то есть будучи вызваны раньше наступления события, они этого события ожидают, блокируя вызвавший процесс и делая невозможной обработку других событий. Может случиться и так, что ни одно из событий в течение долгого периода времени не произойдёт, и тогда нужно исключить холостой расход процессорного времени. Идеальной была бы ситуация, когда мы, каким-то образом узнав о наступлении события, реагируем на него соответствующим системным вызовом, точно зная, что он нас не заблокирует и при этом сделает что-то полезное (примет или передаст данные, примет запрос на соединение и т. п.). Этим полезным действием наша «идеальная» ситуация отличается от активного ожидания с циклическим опросом неблокирующих сокетов: не гарантировав никакого полезного действия, мы обрекаем большинство наших обращений к ядру на холостой расход процессорного времени.

Способ построения программ, при котором программа имеет главный цикл, одна итерация которого соответствует наступлению некоторого события из определённого множества, а все действия программы построены как реакция на событие, называется *событийно-управляемым программированием* (англ. *event-driven programming*). Кроме серверных программ, рассчитанных на одновременное обслуживание неопределённого количества клиентов, существуют и другие области применения этого подхода; так, на принципе обработки событий обычно основываются программы, снабжённые графическим интерфейсом пользователя. В роли событий в них выступают действия пользователя — нажатия на кнопки и «горячие» клавиши, а также перемещения и «щелчки» мыши.

Итерация главного цикла в событийно-ориентированной программе делится на две фазы: ***выборка события*** и ***обработка события***. Чаще всего обработка выносится в отдельные функции, которые вызываются из главного цикла при наступлении соответствующего события. Эти функции так и называются ***обработчиками событий***.

Одна из основных особенностей событийно-управляемого программирования состоит в том, что в обработчике события нельзя оставаться надолго. Что такое «надолго» — зависит от задачи; в программе с графическим интерфейсом «надолго» обычно означает «так, что пользователь заметит паузу», то есть десятую долю секунды в обработчике провисеть ещё можно, а вот полсекунды — уже нельзя, поскольку программа, то и дело «подвисающая» на полсекунды, среднестатистического пользователя может буквально взбесить. В случае TCP-сервера всё, как правило, не так критично, в крайнем случае даже пауза на две-три секунды может остаться незамеченной, нужно только не допускать, чтобы клиент «отвалился по тайм-ауту» (конкретная величина тайм-аута может оказаться разной, но довольно редко бывает меньше 20–30 секунд); впрочем, действовать по принципу «авось прокатит» — тоже идея не слишком удачная.

Так или иначе, можно определённо назвать одну вещь, которая в обработчиках событий заведомо недопустима: это обращения к блокирующим системным вызовам, когда имеется вероятность, что блокировка реально произойдёт. Длительность блокировки в большинстве случаев непредсказуема; допустив блокировку, мы уже не можем гарантировать своевременный возврат управления в главный цикл, то есть получается, что мы можем застрять в обработчике на сколько угодно — в том числе «надолго».

Недопустимость блокировок в обработчиках событий часто вынуждает нас отказываться от очевидного и привычного в пользу таких методов программирования, которые, как это ни странно, некоторым программистам оказываются неподвластны. Когда логика задачи предполагает ту или иную *последовательность действий*, обычно мы без раздумий пишем в программе *последовательность операторов*, выполняющих эти действия. Но в событийно-ориентированной программе такая последовательность операторов окажется недопустимой, если хотя бы один из них может заблокировать наш процесс. Приходится для каждого клиента помнить, на чём мы с ним остановились, а вместо последовательностей операторов писать такой код, который, получив управление, вспоминает, какое действие для данного конкретного клиента должно стать следующим, делает это действие, запоминает, на чём *теперь* закончилась работа с этим клиентом, и возвращает управление, так нигде и не заблокировавшись. К этому вопросу мы вернёмся в §6.4.5.

6.4.4. Выборка событий в ОС Unix: вызов `select`

Для построения серверной программы в виде событийно-ориентированного приложения, очевидно, требуется поддержка со стороны операционной системы; в самом деле, нам нужно *дождаться* наступления события — иначе говоря, *заблокироваться* до момента, когда событие произойдёт, но ведь блокировать процессы умеет только операционная система.

В большинстве своём блокирующие системные вызовы (включая все, которые мы до сих пор обсуждали) ориентированы на ожидание одного конкретного события, но нам в данном случае нужна возможность отдать управление операционной системе, отказавшись от выполнения (т. е. от квантов времени) до тех пор, пока не произойдет одно из *многих* интересующих нас событий. Наступление события операционная система должна отследить сама и вернуть управление процессу, при этом, желательно, сообщив ему, какое именно событие наступило: пришли ли данные по одному из клиентских соединений, получен ли очередной запрос на создание нового соединения, освободилось ли место в буфере для передачи данных в сеть, прошло ли некое заданное время, пришёл ли сигнал. Получив информацию о событии, мы сможем среагировать на него соответствующим образом, то есть принять данные или запрос соединения, отправить очередную порцию данных и т. п., и при должной сноровке избежим как холостого опроса источников событий, так и блокировок, мешающих обрабатывать новые события. Когда в роли событий рассматривается готовность нескольких потоков ввода-вывода к очередным операциям, а сами операции ввода-вывода становятся реакцией на эти события, говорят о ***мультиплексировании ввода-вывода***.

В системах семейства Unix выборку событий можно организовать с помощью системных вызовов `select` и `poll`, которые, вообще говоря, предназначены для одних и тех же действий; `select` несколько проще в работе, `poll` несколько более универсален. В некоторых системах ядро реализует только один вариант интерфейса, при этом второй эмулируется через него в виде библиотечной функции. Так, в системах линейки Solaris присутствовал системный вызов `poll`, а `select` был библиотечной функцией. Кроме того, в некоторых современных системах, в частности в FreeBSD, присутствует также вызов `kqueue`, реализующий альтернативный подход к выборке события. В Linux тоже имеется свой вариант событийного интерфейса — набор вызовов под общим названием `epoll`. Мы будем рассматривать вызов `select` как более простой; изучить `poll`, а также `kqueue` и `epoll` читатель при желании может самостоятельно, прибегнув к технической документации.

Вызов `select` позволяет обрабатывать ***события*** трёх типов:

- изменение состояния файлового дескриптора (появление данных, доступных на чтение, или входящего запроса на соединение; освобождение места в буфере исходящей информации; исключительная ситуация);
- истечение заданного количества времени с момента входа в вызов;
- получение процессом неигнорируемого сигнала.

Профиль вызова выглядит так:

```
int select(int n, fd_set *readfds, fd_set *writefds, fd_set *exfds,
           struct timeval *timeout);
```

Параметры `readfds`, `writefds` и `exfds` обозначают множества файловых дескрипторов, для которых нас интересует соответственно возможность немедленного чтения, возможность немедленной записи и наличие исключительной ситуации. Параметр `n` указывает количество значащих элементов в этих множествах. Этот параметр нужно установить равным `max_d+1`, где `max_d` — максимальный номер дескриптора среди подлежащих обработке. Параметр `timeout` задаёт промежуток времени, спустя который следует вернуть управление, даже если никаких событий, связанных с дескрипторами, не произошло.

Объект «множество дескрипторов» задаётся переменной типа `fd_set`. Внутренняя реализация переменных этого типа для различных систем теоретически может оказаться разной, но проще всего её представлять себе как битовую строку, где каждому дескриптору соответствует один бит. Для работы с переменными этого типа система предоставляет в наше распоряжение следующие макросы:

```
FD_ZERO(fd_set *set);                    /* очистить множество */
FD_CLR(int fd, fd_set *set);   /* убрать дескриптор из мн-ва */
FD_SET(int fd, fd_set *set);  /* добавить дескриптор к мн-ву */
FD_ISSET(int fd, fd_set *set); /* входит ли дескр-р в мн-во? */
```

С формальной точки зрения тип `fd_set` имеет ограничение на максимальный номер дескриптора, входящего в множество. Константа `FD_SETSIZE`, определённая в одном заголовочном файле с вызовом `select` и типом `fd_set`, задаёт *количество* потенциальных элементов множества, а поскольку дескрипторы нумеруются с нуля, *максимальный номер* оказывается на единицу меньше.

Впрочем, везде, где автору книги доводилось этот момент проверять, `FD_SETSIZE` имеет значение 1024, что в точности соответствует максимально возможному количеству файловых дескрипторов на один процесс практически во всех unix-системах. Нужно ли тратить силы, чтобы учесть гипотетически возможное расхождение этих величин на *какой-нибудь* системе, какой никто никогда не видел — вопрос открытый.

Структура `timeval`, служащая для задания последнего параметра, имеет два поля типа `long`. Поле `tv_sec` задаёт количество секунд, поле `tv_usec` — количество микросекунд, т. е. миллионных долей секунды, и его значение не должно превышать `999999`. Например, задать тайм-аут в 5.3 секунды можно следующим образом:

```
struct timeval t;
t.tv_sec = 5;
t.tv_usec = 300000;
```

Отметим, что в разных системах вызов `select` по-разному ведёт себя по отношению к этому параметру. Одни реализации оставляют структуру `*timeout` нетронутой, другие записывают в неё время, оставшееся до истечения заданного тайм-аута. Такая неопределённость приводит к тому, что мы вообще вынуждены не строить никаких предположений относительно содержимого этой структуры после выхода из `select`: использовать её для отслеживания остатков временны́х интервалов нельзя, поскольку не все системы этот остаток туда записывают, но и предполагать её неизменность мы тоже не можем, ведь другие системы её изменяют. Приходится считать, что вызов `select` просто *портит* параметр `timeout`.

В качестве любого из параметров, кроме первого (количества дескрипторов), можно передать нулевой указатель, если задание этого параметра нам не требуется. Так, если нужно просто некоторое время подождать, можно указать `NULL` вместо всех трёх множеств дескрипторов; можно обойтись и без последнего параметра, задающего тайм-аут. Вызов `select(0,NULL,NULL,NULL,NULL)` эквивалентен вызову `pause` (см. стр. 128); если же оставить нулевыми все параметры, кроме последнего, то можно сымитировать `sleep` и `usleep` (§5.3.9).

Вызов `select` возвращает управление в следующих случаях:

- произошла ошибка (в частности, в одном из множеств дескрипторов оказалось число, не соответствующее ни одному из открытых дескрипторов); в этом случае вызов возвращает `-1`;
- наш процесс получил обрабатываемый сигнал; в этом случае также возвращается `-1`, а отличить эту ситуацию от ошибочной можно по значению глобальной переменной `errno`, которая будет равна константе `EINTR`;
- истёк тайм-аут, то есть с момента входа в вызов прошло больше времени, чем указано в параметре `timeout` (если, конечно, этот параметр не был нулевым указателем); в этом случае `select` возвращает `0`;
- на один из дескрипторов, входящих в множество `readfds`, пришли данные, которые можно прочитать, или запрос, который можно принять, либо возникла ситуация «конец файла» (это называют ***готовностью к чтению***, хотя «чтение» в буквальном смысле подразумевается отнюдь не всегда, те же слушающие сокеты тому пример);

- один из дескрипторов, входящих в `writefds`, готов к немедленной записи, то есть если применить к нему вызов `write`, `send` или ещё какой-то подобный, то он сможет *немедленно* записать *хоть сколько-нибудь* данных (***готовность к записи***);
- на одном из дескрипторов, входящих во множество `exfds`, возникла исключительная ситуация;
- с одним из потоков данных, чьи дескрипторы включены во множества `readfds`, `writefds` или `exfds`, случилась неприятность (произошла ошибка).

В последних четырёх случаях вызов `select` возвращает количество дескрипторов, на которых произошли события, причём если дескриптор входит в разные множества и на нём одновременно наступили разные события, он может быть посчитан больше одного раза. Все множества дескрипторов, переданные вызову `select`, при этом перезаписываются: в них остаются только те дескрипторы, на которых что-то произошло (готовность, ошибка и т. д.). Проверив с помощью макроса `FD_ISSET` интересующие нас дескрипторы, мы можем узнать, на каком из них требуется выполнить операцию чтения, записи или принятия соединения.

Сразу же отметим, что исключительные ситуации бывают только на сетевых сокетах и только при использовании механизма *внеполосных данных* (англ. *out-of-band*, OOB), а он используется сравнительно редко, так что и сам параметр `exfds` используется редко; в большинстве случаев в качестве четвёртого параметра `select` указывается `NULL`. Рассматривать внеполосные данные мы в нашей книге не будем.

Готовность на чтение для обычных потоков, т. е. файловых дескрипторов, допускающих чтение вызовом `read`, означает, что во входном буфере есть хотя бы один байт данных, так что вызов `read` (**один**) не заблокируется. Следует обратить внимание, что **ситуация «конец файла» также истолковывается как готовность сокета на чтение**, поскольку в этой ситуации вызов `read` не блокируется — он немедленно возвращает `0`.

Во множество `readfds` можно включать не только дескрипторы, из которых мы собираемся читать `read`'ом. Так, к слушающим сокетам, очевидно, `read` применять нельзя, а к дейтаграммным сокетам — теоретически можно, но обычно так не делают. При этом в `readfds` и слушающие, и дейтаграммные сокеты можно включить с совершенно аналогичной целью: чтобы узнать, когда с ними пора работать. **Для слушающего сокета в роли «пришедших данных» выступают запросы на соединение**, то есть если `select` заявил о готовности слушающего сокета на «чтение», это гарантирует, что вызов `accept` не заблокируется. Для дейтаграммных сокетов гарантируется, что не будет заблокирован вызов приёма дейтаграммы, поскольку минимум одна дейтаграмма пришла.

С готовностью к записи дела обстоят несколько сложнее. При обсуждении системного вызова `write` в §5.2.3 мы отметили, что в ситуации, когда все данные не могут быть записаны немедленно, реализация `write` в большинстве систем предпочитает заблокироваться до тех пор, пока все данные не будут переданы, и лишь после этого вернуть управление. Готовность к записи, о которой нам сообщает `select`, означает, что *сколько-то* данных в сокет (или другой поток) можно записать немедленно, но само по себе это никоим образом не гарантирует, что `write` пройдёт без блокировки: мы ведь не знаем, *сколько* места свободно в буфере ядра, который связан с нашим потоком, и можем потребовать от `write` записи бóльшего количества данных. Вызов примется передавать всё, что мы ему дали, и не вернёт управление нашему процессу, пока не справится с поставленной задачей полностью. Ясно, что это может затянуться, то есть произойдёт та самая блокировка, которая нам совершенно не годится, ведь у нас могут быть другие клиенты со своими запросами.

Часто в силу специфики задачи эту проблему попросту игнорируют. В самом деле, если порции данных, которые вы отправляете своим клиентам, не превышают одного-двух килобайт, вы можете считать, что блокировки на `write` у вас никогда не случится. Даже если она произойдёт, с такими объёмами данных её длительность будет пренебрежимо мала и никак не повлияет на всю вашу программу. В этой ситуации можно вообще не задействовать в вызове `select` параметр `writefds`. Если же объёмы данных, отправляемых клиентам, могут быть существенными, то игнорировать проблему блокировки при записи уже не получится. Правильный выход здесь — перевести ваши сокеты в неблокирующий режим (см. стр. 162), в противном случае никто не может гарантировать вам, что `write` не заблокируется. Значение, возвращаемое вызовом `write`, нужно будет анализировать всегда, предполагая, что оно может оказаться меньше, чем вы ожидали. Данные, подлежащие отправке, вам придётся хранить в своих собственных буферах до тех пор, когда сначала `select` сообщит вам, что можно (осмысленно) попытаться записать данные в сокет, а затем уже `write` подтвердит, что данные отправлены. Когда `write` отправляет только часть данных, эту часть следует изъять из своего «исходящего» буфера, а остальное оставить в буфере до следующего раза, когда `select` снова покажет готовность сокета к записи.

Отметим, что **большинство дескрипторов, открытых на запись, к записи готовы в любой момент**; дескриптор потока вывода может быть не готов к записи, только если буфер исходящих данных, который ядро поддерживает для данного потока, заполнен до отказа. Если случайно внести какой-то из дескрипторов, исходящий буфер которых не переполнен, в множество `writefds`, вызов `select` вернёт управление немедленно; если при этом у вас не найдётся данных, которые в этот дескриптор можно записать, фактически ваша программа, несмотря на использование `select`, будет работать в режиме активного ожидания. Поэтому в множество `writefds` следует вносить только те сокеты, для которых прямо сейчас есть данные, требующие отправки.

Работу с вызовом `select` можно построить по нижеприведённой схеме. Предполагается, что номер слушающего сокета хранится в переменной `ls`; организовать хранение дескрипторов клиентских сокетов можно самыми разными способами, в зависимости от задачи: в списке, в массиве, в дереве и т. п. Кроме того, предполагается, что out-of-band data не используется, так что параметр `exfds` оставлен нулевым.

```
for(;;) { /* главный цикл */
    int fd, res;
    fd_set readfds, writefds;
    int max_d = ls;
        /* изначально полагаем, что максимальным является
           номер слушающего сокета */
    FD_ZERO(&readfds);    /* очищаем множества */
    FD_ZERO(&writefds);
    FD_SET(ls, &readfds); /* вводим в множество
                             дескриптор слушающего сокета */
     /* организуем цикл по сокетам клиентов */
    for(fd= /* дескриптор первого клиента */  ;
        /* клиенты ещё не исчерпаны? */       ;
        fd= /* дескриптор следующего клиента */)
    {
        /* здесь fd - очередной клиентский дескриптор */
        /* вносим его в множества */
        FD_SET(fd, &readfds);
        if(/* для клиента есть данные к отправке? */)
            FD_SET(fd, &writefds);
        /* проверяем, не больше ли дескриптор,
           нежели текущий максимум */
        if(fd > max_d)
            max_d = fd;
    }
    timeout.tv_sec =  /*    заполняем    */;
    timeout.tv_usec = /*    тайм-аут     */;

    res = select(max_d+1, &readfds, &writefds, NULL, &timeout);
    if(res == -1) {
        if(errno == EINTR) {
            /* обработка события "пришедший сигнал" */
        } else {
            /* обработка ошибки, происшедшей в select'е */
        }
        continue; /* дескрипторы проверять бесполезно */
    }
    if(res == 0) {
        /* обработка события "тайм-аут" */
        continue; /* дескрипторы проверять бесполезно */
    }
    if(FD_ISSET(ls, &readfds)) {
        /* пришёл новый запрос на соединение */
        /* здесь его надо принять вызовом accept
           и запомнить дескриптор нового клиента;
           кроме того, не забудьте про перевод сокета
           в неблокирующий режим с помощью fcntl,
```

```
                если используете writefds */
        }
            /* теперь перебираем все клиентские дескрипторы */
        for(fd= /* дескриптор первого клиента */  ;
            /* клиенты ещё не исчерпаны? */       ;
            fd= /* дескриптор следующего клиента */)
        {
            if(FD_ISSET(fd, &readfds)) {
                /* пришли данные от клиента с сокетом fd */
                /* читаем их вызовом read или
                   recv и обрабатываем */
                /* не забываем отдельно обработать
                   ситуацию "конец файла"! */
            }
            if(FD_ISSET(fd, &writefds)) {
                /* готовность на запись: пробуем отправить
                   очередную порцию данных */
            }
        }
        /* конец главного цикла */
    }
```

Как отмечалось в предыдущем параграфе, в событийно-ориентированных программах тело главного цикла делится на две фазы — *выборку события* и *обработку события*. В нашей схематической программе фаза выборки заканчивается вызовом `select`, а всё остальное до конца цикла — это обработка полученного события.

6.4.5. Сеанс работы как конечный автомат

Как уже отмечалось в §6.4.3, недопустимость действий, способных привести к блокировке процесса, вынуждает программистов отказываться от привычных методов работы, заменяя их неочевидными конструкциями: так, реализация *простой последовательности действий*, если хотя бы одно из этих действий может привести к блокировке, в событийно-ориентированной программе не будет иметь ничего общего с привычной *последовательностью операторов*.

В частности, системные вызовы `read` или `recv` могут отдать вам лишь часть запроса, отправленного клиентом; в иной ситуации вы могли бы просто «дочитать» остаток запроса, повторно вызвав `read`/`recv`, но в событийно-ориентированной программе так делать нельзя, ведь гарантия отсутствия блокировки, которую вам дал вызов `select`, распространяется только на *один* вызов чтения. Поэтому, коль скоро запрос прочитан не полностью, приходится запомнить ту его часть, которая была считана, после чего вернуть управление в главный цикл. Вполне возможно, что ваша программа успеет пообщаться с другими клиентами, принять новые соединения и т. п., прежде чем вы наконец получите остаток недочитанного запроса; возможно, для этого потребуется больше одного чтения, и каждый раз придётся возвращать управление в главный цикл, а после каждого чтения «склеивать» новую порцию данных с принятыми от того же клиента ранее и проверять, получен ли теперь весь запрос целиком или придётся подождать следующей порции.

Из этого следует, что **при событийно-ориентированном построении TCP-сервера совершенно необходимо предусмотреть для каждого из клиентов отдельный накопительный буфер**, в котором прочитанная часть запроса будет храниться до тех пор, пока запрос не окажется принят целиком. Следует также учитывать, что если ваш протокол общения с клиентом позволяет клиенту направить вам (серверу) следующий запрос, не дожидаясь ответа на предыдущий, то в вашем буфере в какой-то момент может оказаться запрос целиком, а следом за ним — начало следующего запроса. Если запросы не слишком велики, то вы вполне можете обнаружить в буфере несколько запросов сразу. Всё это приходится учитывать при анализе содержимого вашего накопительного буфера.

Склеиванием запроса из отдельных кусочков, прочитанных в разных вызовах обработчика события, проблемы не ограничиваются. Логика взаимодействия сервера с клиентом в большинстве случаев предполагает некие цепочки, составленные из отдельных обменов данными; такие цепочки как раз и образуют протоколы прикладного уровня. Если бы клиент был один, всю логику прикладного протокола можно было бы напрямую, практически один в один, перевести в операторы языка программирования, но в событийно-ориентированных программах такой подход не годится.

Для примера вспомним школьную задачу по информатике — программу, которая «знакомится» с пользователем, задавая ему один за другим несколько вопросов, например, «как вас зовут», «сколько вам лет», «из какого вы города» и «назовите вашу любимую музыкальную группу». Школьник, изучающий Паскаль, ни на секунду не задумываясь, напишет для этого примерно такой фрагмент:

```
write('What is your name? ');
readln(name);
write('How old are you? ');
readln(age);
write('What city/town are you from? ');
readln(town);
write('What is your favorite band? ');
readln(band);
```

Если подобный диалог потребуется организовать с пользователем, находящимся «по ту сторону» сетевого соединения, мы могли бы написать аналогичную последовательность действий на Си, заменив паскалевский оператор печати на системный вызов `write`, отправляющий вопросы в сетевой сокет, а оператор `readln` — на цикл обращений к вызову `read`, работающий, пока не будет прочитана строка целиком. Но поступить так мы можем лишь когда клиент у нас один и нам не нужно отвлекаться на обслуживание других сеансов связи — например, если мы применяем подход с обслуживающими процессами.

В событийно-ориентированной программе ни о чём подобном не может идти речи, ведь каждый `read` — это источник блокировки; стоит только пользователю отвлечься от общения с нашим сервером, оставив какой-то из вопросов без немедленного ответа, и все остальные пользователи, подключившиеся к серверу, обнаружат, что сервер перестал реагировать на передаваемую ими информацию, а новые клиенты и вовсе не смогут подключиться; все будут вынуждены ждать, когда же единственный «рассеянный» пользователь вспомнит о своём диалоге с сервером. Сервер, построенный подобным образом, ни на что не годится. В конце концов, кто-нибудь может *намеренно* подключиться, начать диалог и не завершить его, оставив соединение открытым, чтобы нарушить работу нашего сервера — из вредности или же, возможно, с целью устранения конкурента в нашем лице.

Для построения последовательного диалога с пользователем в событийно-ориентированной программе приходится *для каждого подключённого клиента* помнить, на какой стадии диалога мы с ним находимся (сакраментальное «на чём мы остановились»). После каждого выполненного чтения из сокета мы вне всякой зависимости от достигнутых результатов обязаны немедленно вернуть управление в главный цикл. Если очередная строка, отправленная пользователем, пришла полностью, то есть в накопительном буфере, связанном с данным клиентом, нашёлся символ перевода строки, мы изымаем строку из буфера, обрабатываем полученный ответ (например, сохраняем информацию в переменных, тоже связанных с конкретным клиентом, но это уже зависит от задачи), после чего запоминаем, что эту стадию диалога с данным клиентом мы уже прошли, отправляем ему очередной вопрос и после этого обработку события заканчиваем, то есть возвращаемся в главный цикл. Если строка после очередного обращения к `read` пока ещё прочитана не полностью — вообще не меняем ничего, кроме накопительного буфера, и немедленно выходим из обработчика события. Продолжить диалог с этим клиентом мы сможем, когда `select` снова сообщит нам о поступлении данных от него — и не раньше.

Итак, в событийно-ориентированной программе нам нужно помнить, на какой стадии коммуникации мы находимся с каждым из клиентов, или, иначе говоря, *помнить* ***состояние*** *сеанса взаимодействия* для каждого подключённого клиента. Это состояние меняется при получении от клиента очередного запроса, ответа на наш запрос или другой информации.

Из курса дискретной математики читателю может быть известно понятие ***конечного автомата***[18] — абстрактного устройства, которое может находиться в одном из нескольких предопределённых состояний и переходить из одного состояния в другое при получении символов из входного потока (или, в более общем случае, при наступлении тех или иных событий). То, как приходится описывать сеанс работы с клиентом в событийно-ориентированных программах, весьма напоминает конечные автоматы.

[18] Соответствующий английский термин — *finite state machine* (FSM); в англоязычных текстах, посвящённых программированию, эта аббревиатура встречается довольно часто, так что полезно знать, что это такое.

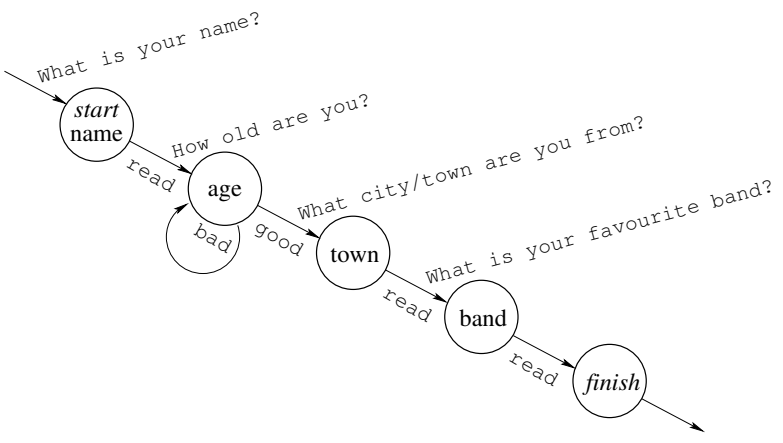

Рис. 6.9. Диаграмма состояний сеанса «знакомства» с пользователем

Диаграмма состояний для диалога с пользователем из нашего примера приведена на рис. 6.9; отметим, что на подобных диаграммах, обычно называемых ***диаграммами Мура***, дуги (стрелки), соединяющие разные состояния, подписываются, во-первых, условием, при котором происходит переход из одного состояния в другое, и, во-вторых, (хотя и не всегда) — *действием*, которое автомат должен выполнить в качестве своего рода побочного эффекта. Наш автомат при создании сеанса работы с клиентом (например, сразу после выполнения вызова `accept`) отправляет в сеть свой первый вопрос (`What is your name?`) и принимает начальное состояние; переход в новое состояние происходит, когда получен ответ, под которым мы подразумеваем строку, заканчивающуюся символом перевода строки. При переходе в новое состояние мы отправляем пользователю текст очередного вопроса. Состояние «start», оно же «name», означает, что мы ещё ни одного ответа не получили, в состояние «age» мы переходим, когда пользователь ответит нам на вопрос о его имени, а мы спросим его о возрасте, в состояние «town» — после валидного ответа на вопрос о возрасте и т. д. Состояние «finish» означает, что мы успешно завершили диалог с пользователем и больше никаких данных «с той стороны» не ожидаем.

Обычно вводят ещё ошибочное состояние (например, «error»), в которое автомат переходит, если произошла ошибка, не позволяющая продолжить работу с данным клиентом. К примеру, в такое состояние мы могли бы перейти, если при чтении у нас кончился буфер, отведённый под накопление входящих данных, а строка полностью так и не пришла: корректное восстановление после такой ошибки в принципе возможно, но довольно проблематично, а буфер можно сделать такого размера, чтобы по смыслу нашей задачи он заведомо (и многократно) превосходил размер любой осмысленной строки, какую может прислать клиент. Например, буфер в 2048 байт будет точно больше, чем привычные нам человеческие имена, названия городов, музыкальных групп и тем более — чем десятичная запись чисел, которые могли бы означать возраст. Если принимаемая строка всё-таки не поместится в такой буфер, то в отказе от дальнейшей работы ничего криминального не будет. Но такую ситуацию нужно уметь отличить от ситуации успешного завершения диалога, и в этом нам как раз поможет введение специального состояния для обозначения ошибки.

Наш пример с диалогом — очень простой, в нём из каждого состояния воображаемый конечный автомат может перейти только в одно (следующее) состояние, причём такой переход во всех случаях, кроме одного, происходит без проверки дополнительных условий — достаточно того, что пришла *какая-то* строка. В более сложных случаях возможны различные варианты переходов из одного и того же состояния в зависимости от наступления тех или иных событий; мы могли бы спросить пользователя о каком-нибудь аспекте его взглядов на жизнь и в зависимости от этого скорректировать дальнейшие вопросы (например, принципиальным пешеходам бессмысленно задавать вопрос о любимой марке автомобиля, а вегетарианцам — о предпочтениях в области охотничьего оружия); можно было бы варьировать происходящее также в зависимости от указанного пользователем возраста и т. п. Диаграммы состояний часто оказываются достаточно сложными и перестают помещаться на бумаге и экране; в таких ситуациях приходится применять хорошо знакомые нам методы борьбы со сложностью — объединять состояния в группы и составлять для каждой группы отдельную диаграмму, а ещё одной диаграммой описывать переходы между группами.

Относительно конечных автоматов следует уяснить два момента. Во-первых, **при реализации конечных автоматов в программах состояниям соответствуют значения переменных**. В простейшем случае можно завести специальный перечислимый тип, имеющий столько же возможных значений, сколько состояний предполагается реализовать в автомате; текущее состояние автомата будет в этом случае храниться в переменной этого типа. Для диалога из нашего примера можно было бы применить следующий перечислимый тип:

```
enum fsm_states {
    fsm_start,
    fsm_name = fsm_start,
    fsm_age,
    fsm_town,
    fsm_band,
    fsm_finish,
    fsm_error
};
```

В более сложных случаях дело может не ограничиться одной переменной. Никто не запрещает делать состояние сколь угодно сложным, включать в него всевозможные счётчики и другие параметры; конечный автомат от этого не перестанет быть конечным автоматом.

Второй момент несколько более сложен для понимания. **Конечный автомат не может управлять поступлением событий**, он над ними не властен. При переходе из одного состояния в другое автомат может выполнять какие-то дополнительные действия (так, в нашем примере он отправлял пользователю текст очередного вопроса), но вот *выборка события* должна оставаться для автомата чем-то внешним и неподконтрольным. Автомат, в частности, не может решать, читать ему очередную строку или не читать: его просто *извещают* о поступлении очередной строки (в общем случае — о наступлении события), ставят перед фактом, а его дело — среагировать, изменив соответствующим образом своё состояние.

Продолжим обсуждение программы, ведущей диалог с пользователем через TCP-соединение. Сеанс работы с клиентом можно описать структурой данных, содержащей дескриптор сокета, состояние автомата, накопительный буфер и поля для размещения сведений, полученных в ходе опроса:

```
#define INBUFSIZE 2048

/* ... */

struct session {
    int fd;                     /* дескриптор сокета */
    char buf[INBUFSIZE];      /* накопительный буфер */
    int buf_used;          /* кол-во данных в буфере */
    enum fsm_states state;       /* состояние сеанса */
    char *name, *town, *band; /* полученные сведения */
    int age;
};
```

Инициализацию сеанса, то есть действия, выполняемые сразу после вызова `accept`, можно совместить с созданием в динамической памяти экземпляра такой структуры. Кроме выделения памяти под структуру и заполнения её полей начальными значениями, в инициализацию сеанса придётся включить отправку клиенту первого вопроса, поскольку согласно нашему простенькому «протоколу» начать диалог (и в дальнейшем вести его) должен именно сервер. У нас получится что-то вроде следующего:

```
static struct session *make_new_session(int fd)
{
    struct session *sess = malloc(sizeof(*sess));
    sess->fd = fd;
    sess->buf_used = 0;
    sess->state = fsm_start;
    sess->name = NULL;
    sess->town = NULL;
    sess->band = NULL;
    sess->age = -1;
    session_send_string(sess, "What is your name?\n");
    return sess;
}
```

Для отправки строки мы здесь воспользовались отдельной функцией `session_send_string`, поскольку это действие нам ещё потребуется, и не раз. Текст этой функции совсем простой:

```
static void session_send_string(struct session *sess, const char *s)
{
    write(sess->fd, s, strlen(s));
}
```

Он мог бы быть гораздо сложнее, если бы мы использовали наряду с буфером для входящих данных также и буфер для отправки в сочетании с неблокирующим режимом сокета и отслеживанием его готовности к записи; но в задаче, которую мы сейчас решаем, общий объём данных, передаваемых в ходе отдельно взятого сеанса работы, запросто поместится в один пакет, так что возможностью блокировки вызова `write` мы можем пренебречь.

Шаг автомата будет реализован в другой функции, которую вызовут, когда в накопительном буфере окажется полная строка. Вызывающий создаст для нас копию этой строки, а из буфера её уберёт, сдвинув в нём данные в начало и уменьшив значение поля `buf_used`; функция, таким образом, становится *владельцем* копии строки с момента её вызова. Это удобно в большинстве состояний, поскольку полученный от клиента ответ сохраняется в структуре сеанса, и если бы переданная нам строка не отдавалась нам в единоличное пользование, нам пришлось бы её скопировать ещё раз; получив её «в собственность», мы, в свою очередь, делаем её «владельцем» структуру сеанса работы. Единственное исключение — состояние `fsm_age`, в котором строка переводится в число, а сама она после этого не нужна и её приходится удалять. Отметим, что обработка состояния `fsm_age` несколько сложнее, чем для остальных состояний, так что мы вынесли эту обработку во вспомогательную функцию, чтобы не загромождать код второстепенными деталями. В целом шаг автомата получается таким:

```
static void session_fsm_step(struct session *sess, char *line)
{
    switch(sess->state) {
    case fsm_name:
        sess->name = line;
        session_send_string(sess, "How old are you?\n");
        sess->state = fsm_age;
        break;
    case fsm_age:
        session_handle_age(sess, line);
        free(line); /* because we still own it... */
        break;
    case fsm_town:
        sess->town = line;
        session_send_string(sess,
            "What is your favorite band?\n");
        sess->state = fsm_band;
        break;
    case fsm_band:
        sess->band = line;
        session_send_string(sess, "Thank you, bye!\n");
        sess->state = fsm_finish;
        break;
    case fsm_finish:
    case fsm_error:
        free(line);   /* this should never happen */
    }
}
```

Оператор выбора (`switch`), выполняющий разные фрагменты кода в зависимости от текущего состояния — это классическая идиома при реализации конечных автоматов в программах. В вышеприведённом фрагменте наглядно видно, как меняется состояние и какая из веток будет выполнена при следующем вызове функции, то есть на следующем шаге автомата. Вспомогательная функция для обработки ответа о возрасте выглядит так:

```
static void session_handle_age(struct session *sess, const char *line)
{
    char *err;
    int age;
    age = strtol(line, &err, 10);
    if(!*line || *err || age < 0 || age > 150) {
        session_send_string(sess, "Please try again!\n");
        session_send_string(sess, "How old are you?\n");
    } else {
        sess->age = age;
        session_send_string(sess,
            "What city/town are you from?\n");
        sess->state = fsm_town;
    }
}
```

Конечно, все эти функции не предназначены для вызова напрямую из главного цикла. Главный цикл, вообще говоря, не должен ничего знать о том, как устроен сеанс работы; здесь мы продолжаем соблюдать принцип борьбы со сложностью через разделение зон ответственности. Так или иначе, получив от вызова `select` информацию о готовности сокета к чтению, главный цикл должен вызвать (напрямую или как-либо косвенно) функцию, входящую в реализацию сеанса, которая выполнит чтение из сокета в накопительный буфер. Эту функцию мы назовём `session_do_read`. Она будет возвращать значение, которое укажет вызывающему, следует ли продолжать работу с данным сеансом или можно его закрыть: `1` будет означать. что сеанс ещё не завершён и должен продолжаться, `0` — что продолжение не требуется. В свою очередь функция `session_do_read`, выполнив чтение, дальнейшие действия — проверку, пришла ли строка, и вызов шага автомата — возложит на следующую функцию, которая будет называться `session_check_lf` (`lf` образовано от *line feed* — одно из названий символа перевода строки). Заодно (уже после возможного изъятия из буфера очередной строки) функция проверит, не переполнился ли накопительный буфер, а затем — не перешёл ли автомат в заключительное состояние. Код функции получается таким:

```
static int session_do_read(struct session *sess)
{
    int rc, bufp = sess->buf_used;
    rc = read(sess->fd, sess->buf + bufp, INBUFSIZE - bufp);
    if(rc <= 0) {
        sess->state = fsm_error;
        return 0;
    }
    sess->buf_used += rc;
    session_check_lf(sess);
    if(sess->buf_used >= INBUFSIZE) {
        /* we can't read further,
           no room in the buffer, no whole line yet */
        session_send_string(sess, "Line too long! Bye...\n");
        sess->state = fsm_error;
        return 0;
    }
    if(sess->state == fsm_finish)
        return 0;
    return 1;
}
```

Функция `session_check_lf` должна связать между собой «системную» обвеску и «прикладной» автомат. Для этого она попытается найти в «занятой» части накопительного буфера символ перевода строки; если его там нет, то ей ничего не останется, кроме как вернуть управление в надежде, что в следующий раз всё будет по-другому. Если же перевод строки в буфере найдётся, функция создаст копию данных из буфера (от начала до перевода строки) в виде отдельной строки в памяти, причём эта строка не будет включать символ перевода строки. Данные в буфере будут сдвинуты к началу, чтобы затереть изъятую строку, а саму эту строку функция передаст уже знакомой нам функции `session_fsm_step`, реализующей шаг автомата. В качестве последнего штриха наша функция проверит, не оканчивается ли полученная строка символом возврата каретки (`'\r'`), и если да — уберёт его из строки, заменив нулевым байтом. Код функции будет выглядеть так:

```
static void session_check_lf(struct session *sess)
{
    int pos = -1;
    int i;
    char *line;
    for(i = 0; i < sess->buf_used; i++) {
        if(sess->buf[i] == '\n') {
```

```
                pos = i;
                break;
            }
        }
        if(pos == -1)
            return;
        line = malloc(pos+1);
        memcpy(line, sess->buf, pos);
        line[pos] = 0;
        sess->buf_used -= (pos+1);
        memmove(sess->buf, sess->buf+pos+1, sess->buf_used);
        if(line[pos-1] == '\r')
            line[pos-1] = 0;
        session_fsm_step(sess, line); /* we transfer ownership! */
    }
```

Мы привели здесь лишь часть реализации сервера, проводящего опрос пользователей — ту, которая относится к сеансу работы, построенному в виде конечного автомата. Полностью текст программы, организующей с пользователями вышеописанный диалог через TCP-соединения и сохраняющей результаты в текстовый файл, читатель найдёт в архиве примеров к книге; соответствующий файл называется `tcp_questn.c`.

Любопытно заметить, что при переходе от программы, построенной как последовательность действий, к программе, основанной на состояниях, в действительности не возникает ничего нового. Если вернуться к фрагменту на Паскале, приведённому на стр. 166, то *текущее состояние* диалога с пользователем в этой программе тоже присутствует, просто выражено оно не значением переменной, а *текущей позицией исполнения* самой программы. Иногда говорят, что состояние в обычных императивных (то есть построенных на приказаниях) программах присутствует *неявно* — в отличие от случая, когда состояние *явным образом* определяется значениями переменных. Сам этот стиль написания программ называют ***программированием в терминах явных состояний***.

6.4.6. Неблокирующее установление соединения

Достаточно часто одна и та же программа выступает как в роли сервера, так и в роли клиента. Например, в серверной программе может потребоваться запросить ту или иную информацию у другого сервера, для чего придётся инициировать создание соединения, то есть стать клиентом другого сервера.

Если серверная программа написана в виде событийно-ориентированного приложения, то, как мы неоднократно подчёркивали, в ней нельзя допускать блокировку ни на каких системных вызовах; естественно, к вызову `connect` это тоже относится. Между тем обычно вызов `connect` блокирует вызвавший процесс до тех пор, пока попытка установления соединения не завершится успехом или неудачей. Время такой блокировки может составить (в наихудших случаях) несколько десятков секунд, что для событийно-ориентированных программ совершенно неприемлемо.

Техника установления соединения с помощью вызова `connect`, гарантирующая отсутствие блокировки, довольно нетривиальна, но освоить её вполне возможно. Итак, первое, что нужно сделать — это перевести сокет в неблокирующий режим (см. стр. 162). После этого можно сделать вызов `connect`, как это обсуждалось в §6.3.4; поскольку сокет находится в неблокирующем режиме, вызов вернёт управление немедленно, и здесь возможны три варианта:

- если вызов `connect` вернул `0` — значит, соединение уже благополучно установлено; так бывает сравнительно редко и обычно только при соединении с сервером, работающим на одной с нами машине, но игнорировать эту возможность ни в коем случае не следует;
- если вызов вернул `-1`, но при этом в переменной `errno` содержится значение `EINPROGRESS` — значит, запрос на соединение отправлен, но результатов пока нет; это самый частый и самый интересный случай;
- если вызов вернул `-1` и переменная `errno` содержит какое-то другое значение, отличное от `EINPROGRESS` — значит, произошла ошибка, то есть соединение заведомо не будет установлено; ждать в этом случае нечего.

С первым и третьим случаями всё понятно; интереснее всего, когда `connect` завершился с «ошибкой» `EINPROGRESS`, которая на самом деле вовсе не ошибка — это значение в `errno` указывает, что установление соединения началось, но пока не завершилось. В этой ситуации следует внести дескриптор сокета в множество дескрипторов, для которых нас интересует *готовность к записи* (в терминах §6.4.4 — множество `writefds`); именно готовность к записи (не к чтению!) покажет, что попытка установления соединения так или иначе закончилась — либо успехом, либо неудачей.

Остаётся узнать, удалось установить соединение или нет, и здесь нам придётся упомянуть ещё один системный вызов — `getsockopt`:

```
int getsockopt(int sockfd, int level, int optname,
               void *optval, socklen_t *optlen);
```

Никакой логики в том, что придётся использовать именно этот вызов, вообще-то нет: `getsockopt` изначально предназначен, чтобы узнавать текущее значение параметров сокета, устанавливаемых с помощью уже знакомого нам `setsockopt` (см. §6.3.5); очевидно, что статус «ошибки» к настраиваемым параметрам сокета («опциям») никакого отношения не имеет. Так или иначе, логично это или нет, но узнать, чем кончилась попытка установления соединения, мы можем только так: в качестве «уровня» указываем уже знакомое нам `SOL_SOCKET`, в качестве «имени опции» — `SO_ERROR`; параметр `optval` должен задавать адрес области памяти, куда следует записать значение «опции» — в данном случае эта область памяти представляет собой обычную переменную типа `int`; наконец, параметр `optlen` работает (в общем случае) как на вход, так и на выход — на входе через него передаётся размер области памяти, адрес которой передан через `optval`, а на выходе эта переменная будет содержать количество байтов, которое вызов реально использовал для записи значения «опции». В нашем случае мы точно знаем, что это значение не изменится, но это знание нам, увы, ничем не поможет; придётся честно завести переменную типа `socklen_t`, присвоить ей значение `sizeof(int)` и передать её адрес последним параметром вызова. Итоговое обращение к `getsockopt` будет выглядеть так:

```
int opt;
socklen_t optlen = sizeof(opt);
getsockopt(sd, SOL_SOCKET, SO_ERROR, &opt, &optlen);
```

После этого значение переменной `opt` укажет нам статус нашего соединения: ноль будет означать, что соединение установлено, а любое другое значение — что произошла ошибка.

6.4.7. (*) Сигналы в роли событий; вызов `pselect`

Материал этого параграфа не имеет прямого отношения к программированию сетевого взаимодействия; но если вы будете писать событийно-ориентированные программы, построенные на основе вызова `select`, рано или поздно вам встретятся *обрабатываемые сигналы* в роли событий, а вместе с ними — утверждение, что для их корректной обработки вызов `select` «не годится» и нужен загадочный вызов `pselect`. Вопросы вроде «как им пользоваться и зачем он вообще нужен» возникают с завидной регулярностью, причём не только у начинающих; в этом параграфе мы попытаемся прояснить, что тут в действительности к чему. Если у вас нет железобетонной уверенности в собственных навыках обращения с сигналами, будет правильно сначала вернуться к §5.3.14, и только после этого разбираться с `pselect`.

Вызов `select` относится к числу таких блокирующих вызовов, которые при поступлении обрабатываемого сигнала обязательно возвращают управление с ошибкой `EINTR`, что в принципе позволяет рассматривать сигналы как разновидность *событий* и реагировать на них, как и на другие события, в главном цикле программы. Проблема в том, что главный цикл состоит отнюдь не только из вызова `select`.

Допустим, наша программа обрабатывает сигнал `SIGUSR1`, для чего предусмотрена глобальная переменная `usr1_caught` и функция-обработчик:

```
volatile int usr1_caught = 0;
void usr1_handler(int n)
{
    signal(n, usr1_handler);
    usr1_caught++;
}
```

При поступлении сигнала нам нужно выполнить некие действия, объединённые функцией `actions_on_usr1`, которую надлежит вызвать из главного цикла. Отметим, что в такой функции — в отличие от функции-обработчика сигнала — мы можем себе позволить использование любых средств, лишь бы они не выполнялись «долго»; дело в том, что `actions_on_usr1` вызывается из главной программы и не может (в отличие от «настоящего» обработчика сигнала) оказаться вызвана «в неожиданный момент». Сам главный цикл схематически можно представить следующим образом:

```
for(;;) {
    /* A */
    res = select( /* ... */ );
    /* B */
    while(usr1_caught > 0) {
        actions_on_usr1();
        usr1_caught--;
    }
    /* C */
}
```

— где A, B и C — некие фрагменты кода; теперь мы уже можем догадаться, какого рода проблема связана с этим решением. Если сигнал придёт во время выполнения select, то есть когда наш процесс будет находиться в состоянии блокировки, а равно во время выполнения фрагмента B или даже во время выполнения цикла по переменной usr1_caught, всё будет в порядке: управление дойдёт до actions_on_usr1 на текущей итерации главного цикла, причём разнообразные варианты обработки других событий, в том числе, возможно, наступивших на той же итерации, погоды не сделают — мы ведь помним, что в обработке отдельно взятого события нельзя (по условиям событийно-ориентированной парадигмы) задерживаться «надолго».

Картина существенным образом меняется, если наш процесс получит сигнал в ходе выполнения фрагментов A или C. В этом случае вызов select не будет ничего знать о том, что сигнал уже пришёл, и заблокирует наш процесс до тех пор, пока не наступит *ещё какое-нибудь событие*; лишь после этого, то есть уже на следующей итерации главного цикла, функция actions_on_usr1 наконец получит управление. Проблема в том, что очередное событие может не происходить очень долго, и именно на этот (в общем случае неопределённый) срок реакция нашей программы на пришедший сигнал окажется «отложена».

Чаще всего на сигналы «вешают» какое-нибудь изменение режима работы программы, например, повторное считывание конфигурационных файлов и т. п.; системный администратор, когда надо, отправляет нашей программе сигнал с помощью команды kill. Если ему «не повезёт» и сигнал придётся на неудачную фазу главного цикла, а никаких новых событий происходить не будет, выглядеть это будет так, как будто наша программа вообще не желает откликаться на сигнал.

Начинающие программисты часто, не задумываясь, применяют «очевидное» решение, вставляя проверку переменной и обработку события-сигнала прямо перед вызовом select:


```
for(;;) {
    /* ... */
    while(usr1_caught > 0) {
        actions_on_usr1();
        usr1_caught--;
    }
    res = select( /* ... */ );
    /* ... */
}
```

Программа при этом становится «почти правильной», но ключевое слово тут «почти»: по закону подлости сигнал может прийти как раз между циклом while и вызовом select. Вероятность этого довольно низкая, но всё же не нулевая.

Прежде чем рассказать, как выглядит полностью правильное решение для обработки сигнала в качестве события, припомним, что с совершенно аналогичной проблемой мы уже сталкивались, пытаясь дождаться сигнала от потомка с помощью функции pause (см. рассуждение на стр. 129–130); там тоже сигнал мог прийти «слишком рано» — до того, как процесс войдёт в состояние блокировки, из которого сигнал должен был его вывести, только там процесс блокировался в функции pause, а не select. Решение, к которому мы в итоге пришли, состояло в том, чтобы заменить pause вызовом sigsuspend, который на время своего выполнения устанавливает альтернативное множество заблокированных сигналов, а перед возвратом управления восстанавливает исходное; основная идея решения — чтобы нужный нам сигнал был всегда заблокирован и разблокировался только на время нахождения нашего процесса в блокировке, что как раз и обеспечивает вызов sigsuspend (вместо pause).

Полностью аналогичным «заменителем» вызова select как раз и выступает вызов pselect. Он отличается от select наличием ещё одного, шестого по счёту параметра, имеющего тип sigset_t; этот параметр указывает, какие сигналы должны быть заблокированы на время работы вызова. Сигналы, не входящие в указанное множество, на время работы pselect разблокируются. При возврате управления pselect восстанавливает множество заблокированных сигналов в том виде, в котором оно было на момент входа в вызов. Ещё одно отличие состоит в том, что pselect для задания тайм-аута использует структуру типа struct timespec (а не timeval, как обычный select), причём согласно спецификации не изменяет её содержимое этой структуры, передаваемой пятым параметром (напомним, что select в некоторых системах оставляет аналогичный параметр нетронутым, но в других — записывает в него время, оставшееся до истечения тайм-аута, так что для обеспечения переносимости программы приходится предполагать, что содержимое структуры просто портится). Профиль pselect выглядит так:

```
int pselect(int n, fd_set *readfds, fd_set *writefds, fd_set *exfds,
            const struct timespec *timeout, const sigset_t *sigmask);
```

Структура timespec состоит из двух полей, первое из которых называется tv_sec (точно так же, как и в структуре timeval, используемой обычным select'ом), но второе поле называется tv_nsec и должно содержать *наносекунды*, а не микросекунды, как было в структуре timeval; значение этого поля, естественно, должно находиться в диапазоне от 0 до 999999999 (миллиард без единицы).

Начинающие программисты часто не понимают, как пользоваться вызовом pselect, поскольку его профиль и описание создают (ложное) впечатление, будто бы параметр sigmask предназначен, чтобы *заблокировать* некоторые сигналы на время работы вызова. В действительности роль этого параметра прямо противоположна: он используется, чтобы на время работы pselect некоторые сигналы *разблокировать*, а потом заблокировать обратно. Впрочем, точно такие же трудности возникают и с восприятием вызова sigsuspend, о чём мы уже рассказывали (см. всё тот же §5.3.14).

Если сигнал, который надлежит обрабатывать в качестве события, будет заблокирован в течение всего времени выполнения, *кроме* времени, которое программа проводит в вызове pselect, то доставка процессу этого сигнала будет возможна только во время работы вызова и, как следствие, обязательно приведёт к выходу из него и выполнению итерации главного цикла. Например, если наша программа обрабатывает в качестве события сигнал SIGUSR1, схема главного цикла будет выглядеть так:

```
sigset_t mask, orig_mask;
/* ... */
sigemptyset(&mask);
sigaddset(&mask, SIGUSR1);
sigprocmask(SIG_BLOCK, &mask, &orig_mask);
signal(SIGUSR1, usr1_handler);
/* ... */
for(;;) {
    /* ... */
    res = pselect( /* ... */ , &orig_mask);
    /* ... */
    while(usr1_caught > 0) {
        actions_on_usr1();
        usr1_caught--;
    }
    /* ... */
}
```

Справедливости ради отметим, что во многих случаях все эти сложности оказываются не нужны. К примеру, если ваша программа, построенная на обычном вызове select, в силу особенностей решаемой задачи «просыпается» через каждую десятую (или какую-то другую) долю секунды по тайм-ауту, то время, которое пройдёт между доставкой сигнала и его обработкой, даже в самом неудачном случае будет разве что чуть-чуть (на время работы других обработчиков) превышать эту долю секунды; скорее всего, это окажется приемлемо. Тем не менее, в любом случае полезно понимать суть проблемы сигналов, приходящих «не вовремя», и принцип предложенного решения на основе pselect и sigprocmask; во-первых, это поможет понять другие подобные проблемы и то, как их решать (как мы увидим позже, таких проблем не так уж мало), а во-вторых, понимание описанной в этом параграфе механики и умение рассказать всё это на собеседовании может сильно повысить вашу привлекательность для потенциальных работодателей.

Часть 7

Параллельные программы и разделяемые данные

Эта часть книги будет посвящена *параллельному программированию*, то есть такому подходу к написанию программы, при котором задача разбивается на подзадачи, выполняющиеся *одновременно*. Такие подзадачи часто оформляются в виде уже упоминавшихся ***легковесных процессов*** (***тредов***); с точки зрения программиста механизм тредов выглядит как возможность написать некую функцию и запустить её на выполнение параллельно с основной программой, причём с сохранением доступа ко всему, к чему имеет доступ основная программа, в том числе, например, к глобальным переменным или к структурам данных, которые сформированы в основной программе и в других тредах.

Программирование с использованием тредов в английском языке обозначается словом *multithreading*. С переводом этого слова на русский имеются определённые проблемы. Английское слово *thread* в программистском контексте чаще всего так и «переводят» путём транслитерации, то есть говорят попросту о *тредах*, не утруждая себя подбором адекватного русского термина; именно так поступаем и мы в нашей книге. Но со словом *mutlithreading* этот номер проходит несколько хуже — точнее, можно сказать, что не проходит совсем из-за его тяжеловесности. Наиболее часто встречающийся русскоязычный термин — ***многопоточное программирование***. Подразумевается, что легковесный процесс обозначается словом *поток* (управления), что частично оправдывается существованием в английском языке словосочетания *control flow*, которое с некоторой натяжкой можно перевести как «поток управления»; хорошо знакомые нам блок-схемы по-английски называются *flow charts*.

Всё бы хорошо, но попытка перевести слово *thread* как «поток» натыкается на другую лингвистическую проблему. Английское слово *stream* тоже переводится на русский язык как «поток», причём если мы попытаемся передать смысловую разницу между *stream* и *flow*, то *flow* придётся переводить скорее не как «поток», а как «течение». При этом слово *stream*, совершенно законно переведённое именно как *поток*, в программировании активно используется в значении «поток данных» — так, мы уже хорошо знакомы с понятием *потока ввода-вывода*. Попросту говоря, русское слово *поток* в программистской терминологии давно и прочно занято.

Несмотря на это, термин «многопоточное программирование» в русском языке вполне прижился; а вот термин «поток» для обозначения отдельно взятого треда приживаться не хочет, и это вполне можно понять.

С появлением и массовым распространением *многоядерных процессоров*, представляющих собой фактически несколько независимых процессоров в одной микросхеме, тематика параллельных вычислений стала настолько популярной, что возможности, ориентированные исключительно на распараллеливание программы, проникли даже в стандарты языков программирования, включая Си и Си++, окончательно всё там испортив. На самом деле в подавляющем большинстве задач, решаемых с помощью компьютеров, видимая скорость выполнения определяется в основном такими факторами, как время отклика в компьютерной сети, скорость обмена с внешними запоминающими устройствами, пропускная способность шины компьютера, но никак не количество доступной вычислительной мощности центральных процессоров, которой как раз обычно хватает (и даже с избытком). От многоядерных процессоров вообще оказывается на удивление мало пользы. Задачи, в принципе допускающие распараллеливание на независимые последовательности вычисления, встречаются редко, поскольку при решении большинства задач последующие шаги зависят от результатов предыдущих шагов и не могут начаться раньше, нежели эти результаты будут получены. Ещё реже встречаются программисты, умеющие эффективно выделять независимые подпоследовательности из общего алгоритма и разносить их на отдельные потоки управления так, чтобы от этого был какой-то ощутимый положительный эффект. Но даже когда задача попалась как раз такая, как надо, а программист, который её решает, прекрасно знает своё дело, итоговый эффект может несколько обескуражить.

Ещё в 1967 году был сформулирован ***закон Амдала***: «общее повышение производительности, полученное за счёт оптимизации отдельной части системы, ограничено долей времени, используемой этой частью»; в русскоязычных источниках встречается более прямолинейная формулировка: **когда задача разделяется на несколько частей, суммарное время её выполнения на параллельной системе не может быть меньше времени выполнения самого медленного фрагмента**. При всей их очевидности словесные формулировки не дают, как говорится, полноты ощущений; иное дело, если закон Амдала выразить в виде математической формулы, отражающей возможный выигрыш в быстродействии в зависимости от количества процессоров. Если обозначить за α долю вычислений, оставшихся нераспараллеленными (α при этом соответствует числу от 0 до 1, причём единица соответствует обычной последовательной программе, ну а ноль, можно считать, недостижим), и считать, что все остальные вычисления $(1-\alpha)$ распараллелены *идеально*, то в системе, имеющей p процессоров, теоретический предел ускорения будет выражаться формулой

$$S = \frac{1}{\alpha + \frac{1-\alpha}{p}}$$

Например, если всего лишь четверть от общего количества вычислений не может быть распараллелена, или, иначе говоря, минимальное «расстояние», которое придётся пройти по последовательным инструкциям от старта программы до её завершения, не обращая внимания на те инструкции, которые выполняются параллельно в то же самое время — вот это вот расстояние составляет четверть от исходного, т. е. $\alpha = \frac{1}{4}$ (что вообще-то весьма и весьма неплохой результат), то в системе с десятью процессорами мы получим ускорение чуть более чем втрое, ускорение ровно в 3.5 раза получится при 21 процессоре, при 100 процессорах будет что-то около 3.88, ну а ускорения вчетверо мы не получим вообще никогда, будь у нас этих процессоров хоть миллион. Если же некие гениальные программисты смогут какую-то сферическую задачу в вакууме раздербанить на параллельные ветки так, что от старта до финиша она станет короче в 20 раз ($\alpha = 0.05$; вообще-то обычно это нереально), то теоретический предел составит те самые 20 раз, причём десятикратное ускорение будет теоретически возможно уже при 21 процессоре, для пятнадцатикратного ускорения процессоров потребуется уже 57, а, скажем, для 19-кратного — 361. На этом месте стоит припомнить, что закон Амдала указывает теоретический (идеальный) предел возможного ускорения, который, разумеется, недостижим; на практике в дело вмешиваются издержки на коммуникацию между потоками управления, на передачу между ними данных (даже если процессоры работают с общей памятью, остаётся проблема так называемой когерентности кешей, которая решается аппаратурой без участия программиста — но отнюдь не бесплатно). Плюс к тому сколько-нибудь «достойные» значения для α достигаются только на весьма специфических задачах.

Едва ли не единственная область программистской деятельности, в которой параллельные вычисления действительно нужны, а наша пресловутая α принимает значения, при которых хотя бы не обидно бороться за возможное ускорение работы — это числовые расчёты, в большинстве случаев связанные с математическим моделированием физических явлений; с такими задачами мы обычно сталкиваемся в суперкомпьютерных вычислительных центрах, принадлежащих научным организациям. Обычный пользователь с объёмными расчётами может встретиться разве что при обработке собственноручно отснятого цифрового видео, но этим занимаются далеко не все; писать же *программы*, предполагающие преобразование видеоданных, приходится совсем небольшому количеству программистов.

Условно (*весьма* условно) «удачным» примером использования легковесных процессов может служить интерактивная программа, работающая со сжатыми данными (например, музыкальный редактор, способный работать с mp3-файлами, или видеоплеер). С одной стороны, операции упаковки/распаковки данных требуют больших объемов вычислений и могут длиться по несколько минут и даже десятков минут. С другой стороны, интерактивная программа не может себе позволить в течение нескольких минут не реагировать на действия пользователя: например, пользователь может переместить окно программы на экране (что потребует его отрисовки) или даже принять решение об отмене текущей операции. В такой ситуации логично

использовать один легковесный процесс для выполнения вычислений (упаковки/распаковки) и другой — для обработки интерфейсных событий, то есть для реакции на действия пользователя. Впрочем, даже здесь в действительности следует использовать не треды, а хорошо знакомые нам процессы, обладающие изолированным друг от друга пространством памяти.

Примеров, когда легковесные процессы (и, говоря шире, параллельные вычисления как таковые) вроде бы применяются «по делу», очень мало; привести второй пример в дополнение к вышеописанному не так просто, как кажется. А вот примеры использования многопоточности *не по делу*, напротив, встречаются постоянно. Зачастую программы, написанные с использованием легковесных процессов, запускают их лишь затем, чтобы они почти тут же остановились в состоянии блокировки и ждали результатов работы друг друга.

В §6.4.3 мы рассматривали пример ситуации, в которой имеются *независимые источники внешних событий*, и эти события нуждаются в своевременной обработке. При этом внешние события могут выстраиваться в логические цепочки, такие, например, как вопросы, заданные пользователю-человеку, и получение ответов на эти вопросы; если такая цепочка у нас одна, мы можем написать последовательность операторов, реализующих, например, логику диалога: задать вопрос, дождаться ответа, обработать ответ, задать следующий вопрос, дождаться ответа и так далее. При работе с несколькими независимыми источниками событий мы не можем себе позволить роскошь *дожидаться* чего-то конкретного, поскольку если мы будем так делать, то все наши клиенты кроме одного — того, ответа от которого мы ждём — будут вынуждены ждать уже нас.

Решение, которое мы предложили в предыдущей части книги (см. §6.4.3), состоит в том, чтобы построить приложение в виде *цикла обработки событий*, на каждой итерации которого мы сначала производим выборку события, то есть узнаём, из какого источника пришло очередное событие и принимаем это событие (например, читаем данные из потока ввода), после чего обрабатываем полученное событие. Мы назвали такой стиль работы ***событийно-управляемым программированием***. При этом во время обработки событий мы не имеем права делать ничего такого, что может нашу программу заблокировать; в частности, производить чтение из потоков ввода мы можем лишь тогда, когда уверены, что данные туда уже пришли. Не можем мы также выполнять и «большие» секции вычислений, занимающие ощутимое время — при необходимости их приходится разбивать на отдельные части и выполнять порциями, при условии, что новых событий не поступило. Для каждой логической цепочки связанных внешних событий — например, для каждого сеанса обслуживания отдельно взятого клиента — мы при этом должны помнить в явном виде состояние дел, всё то же «на чём мы остановились». Часто при получении очередного события мы только меняем соответствующим образом состояние и больше ничего не делаем, поскольку для действий время ещё не пришло. Событийно-ориентированное програмирование требует от программиста определённого мастерства, поскольку мышление в терминах явного выделения состояния несколько отличается от традиционного мышления в терминах последовательностей действий.

Довольно часто встречаются программы, авторы которых то ли не умеют работать в терминах явных состояний, то ли вообще не знают о возможности событийно-ориентированного построения программы, зато, к несчастью, знают о возможности запуска тредов. Проблему с независимыми источниками событий они решают очень просто: на каждый такой источник запускают свой собственный тред. «Параллельная» программа, написанная таким образом, обычно имеет весьма характерную особенность: *все её треды, сколько бы их ни было, дружно «стоят»*, заблокировавшись на очередной операции чтения или в ожидании другого события. Естественно, такое построение программы не даёт и не может дать никакого выигрыша в быстродействии. Квалифицированный программист не станет применять легковесные процессы в качестве хранилища состояния, поскольку на это тратится слишком много памяти, причём как памяти задачи, так и памяти ядра системы. Кроме того, кажущаяся простота многопоточной обработки событий оборачивается неожиданно серьёзными проблемами, имманентно присущими параллельному программированию.

Важнейшей особенностью параллельного программирования является возникновение ***разделяемых данных***, то есть таких данных, к которым одновременно имеют доступ два и более независимых «действующих лица», будь то процессы или треды. Обычно это переменные в памяти нашего процесса, но с таким же успехом это могут быть данные, хранящиеся в файлах на диске; проблема разделяемых данных при этом никуда не исчезает.

Многопоточное программирование требует определённого мастерства и аккуратности, как и любое программирование с использованием разделяемых данных. При появлении разделяемых данных приходится принимать меры, чтобы разные процессы никогда не обращались к разделяемым данным одновременно; для этого вводится понятие ***критической секции*** (участка программы, где происходит работа с разделяемыми данными) и создаются специальные средства ***взаимного исключения***, благодаря которым процессы не попадают в критическую секцию, когда в этой же секции находится другой процесс.

Проблемы, которые возникают при работе с разделяемыми данными, примечательны уже тем, что их *довольно трудно осознать*; многие программисты, использующие многопоточную модель программирования, при этом не понимают, зачем нужны средства взаимоисключений и прочие нетривиальные концепции, сопровождающие поддержку тредов.

Здесь автор вынужден подчеркнуть, что это ничуть не преувеличение. Чуть позже мы подробно рассмотрим несколько примеров, наглядно демонстрирующих возможные последствия одновременного обращения двух процессов или тредов к одной несчастной целочисленной переменной; эти примеры, можно надеяться, убедят читателя в необходимости взаимоисключения при работе с разделяемыми данными. Вашему покорному слуге неоднократно встречались программисты, работающие с применением тредов и получающие за это деньги, и при этом не способные ничего толком ответить на просьбу вроде «убеди меня в том, что здесь нужен мьютекс» — даже после многочисленных наводящих вопросов.

Другая (увы, даже более серьёзная) особенность проблем с разделяемыми данными состоит в сравнительно низкой вероятности их проявления. Неправильно написанная многопоточная программа может довольно долго работать без ошибок, а подведёт, естественно, в самый неподходящий момент. Пытаться выявить ошибки такого рода с помощью привычной нам ***отладки*** столь же бесполезно, как, к примеру, носиться с сачком по подвалу, кишащему крысами: даже если нам повезёт поймать несколько грызунов, меньше их от этого не станет. Отлаживать многопоточные программы вообще довольно сомнительное занятие: воспроизвести конкретную конфигурацию состояний всех имеющихся потоков обычно нереально, это превышает возможности человека.

В статье Википедии, посвящённой многопоточному программированию, приводится удачная цитата из работы [14] профессора университета Беркли Эдварда Ли:

> Хотя потоки выполнения кажутся небольшим шагом [в сторону] от последовательных вычислений, по сути этот шаг огромен. Они отказываются от наиболее важных и привлекательных свойств последовательных вычислений: понятности, предсказуемости и детерминированности. Потоки выполнения как модель вычислений страшно недетерминированы; работа программиста превращается в отстригание этого недетерминизма.

Впрочем, даже если принять решение об отказе от многопоточного программирования, квалифицированный программист в любом случае должен понимать, что за проблемы возникают при работе с разделяемыми данными — хотя бы затем, чтобы точно знать, *почему* не следует применять многопоточное программирование без крайней необходимости, и уметь аргументированно показать неправильность чужой программы. Кроме того, иногда можно встретить ситуации, в которых появление разделяемых данных неизбежно. Например, именно так обстоят дела внутри ядер операционных систем, если, конечно, система способна использовать больше одного процессора (а таковы *все* современные системы).

Вообще говоря, для появления разделяемых данных даже не обязательно сознательно пытаться написать ту или иную программу в «параллельном» виде. Нередко бывает, что к одним и тем же данным вынуждены обращаться совершенно разные программы. Хрестоматийный пример — файл почтового ящика на сервере электронной почты, куда почтовый сервер должен дописывать новые письма, а клиентским программам и серверам финальной доставки почты (pop3/imap) в то же самое время нужен доступ к этому файлу для чтения писем, а также для их удаления по требованию пользователя. Но здесь всё довольно просто: каждая из программ, обращающаяся к файлу почтового ящика, использует системные средства для извещения остальных

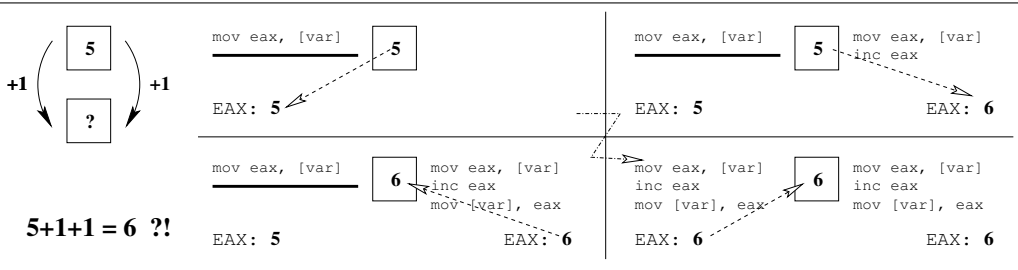

Рис. 7.1. Одновременное увеличение разделяемой переменной

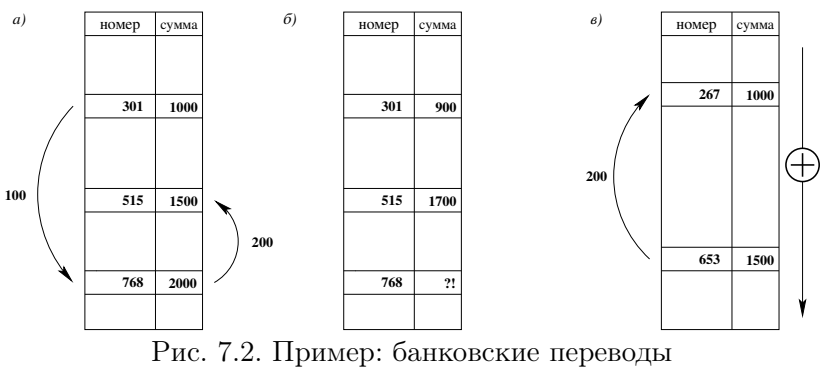

Рис. 7.2. Пример: банковские переводы

о том, что файл сейчас находится в работе и лезть к нему не надо. Намного жёстче ситуация с базой данных, доступ к которой нужен многим людям одновременно; количество таких людей варьируется от двух-трёх (например, несколько бухгалтеров, совместно обрабатывающих поступающие документы на предприятии) до сотен тысяч (система бронирования авиабилетов). Очевидно, что здесь проблема уже не в том, какие средства мы выбираем для работы, а в самой задаче: разделяемые данные могут быть свойством предметной области. Если это так, то избежать работы с разделяемыми данными мы никак не сможем.

7.1. О работе с разделяемыми данными

7.1.1. Устаревание и целостность данных

Рассмотрим простой пример, подразумевающий доступ к разделяемым данным. Два различных процесса имеют доступ к разделяемой памяти, в которой содержится целочисленная переменная. Пусть её начальное значение равно 5. Оба процесса в некий момент пытаются увеличить значение переменной на 1, так что в итоге переменная должна получить значение 7. Для этого нужно загрузить значение из памяти в регистр, увеличить значение регистра на 1 и выгрузить его значение обратно в память:

```
mov eax, [var]
inc eax
mov [var], eax
```

Как мы знаем, процессоры семейства i386 могут выполнять арифметические операции непосредственно в памяти, а с помощью префикса `lock` (см. т. 1, §3.2.14) такие команды можно сделать атомарными ценой некоторого снижения скорости их работы. Это, однако, не означает неправильности примера: представьте себе, что число хранится в памяти в виде строкового представления — в этом случае операция его приращения никак не сможет быть атомарной. Кроме того, i386 — не единственная в мире архитектура; существуют процессоры, имеющие лишь две операции над областями памяти — загрузку значения из памяти в регистр и выгрузку значения из регистра в память.

Представим, что первый процесс выполнил команду `mov` для загрузки значения переменной в регистр, то есть в регистре теперь содержится число 5, и в этот момент произошло прерывание по таймеру, в результате которого процесс оказался снят с исполнения и помещён в очередь процессов, готовых к выполнению (см. рис. 7.1). В это время второй процесс также выполняет команду `mov` (в его регистре `EAX` теперь находится значение 5), потом команду `inc` для увеличения значения в регистре (получается 6) и команду `mov` для выгрузки содержимого регистра в память. Переменная теперь равна 6.

Между тем первый процесс дождался своей очереди и снова поставлен на выполнение. Первую команду `mov` он уже выполнил, и в его регистре сейчас число 5. Он выполняет команду `inc` (в регистре теперь 6) и команду `mov` (значение 6 выгружается в память). Получается, что в итоге переменная получила значение 6, а не 7. Заметим, что значением было бы именно 7, если бы первый процесс не оказался прерван после первой команды. Узнать, был процесс прерван или нет, вообще говоря, невозможно, как и запретить прерывать процесс. Получается, что **конечный результат зависит от того, в какой конкретно последовательности произойдут события в независимых процессах**. С подобным мы уже сталкивались ранее (см. §5.3.4) и знаем, что это называется ***ситуацией гонок*** или ***состязаний*** (англ. *race condition*).

Рассмотрим более серьёзный пример. Имеется база данных, содержащая остатки денег на банковских счетах. Допустим, на счетах под номерами 301, 515 и 768 содержится соответственно \$1000, \$1500 и \$2000. Один оператор проводит транзакцию по переводу суммы в \$100 со счёта 301 на счёт 768, а другой — транзакцию по переводу \$200 со счёта 768 на счёт 515 (см. рис. 7.2, *а, б*). Если эти действия провести в разное время, остатки на счетах составят, понятно, \$900, \$1700 и \$1900.

Теперь предположим, что процесс, запущенный вторым оператором, был прерван после операции чтения остатков на счетах 515 и 768. Прочитанные остатки (соответственно \$1500 и \$2000) процесс сохранил в своих переменных, и именно в этот момент случилось прерывание таймера, в результате которого управление получил процесс, запущенный первым оператором. Этот процесс произвёл чтение остатков счетов 301 и 768 (\$1000 и \$2000), вычислил новые значения остатков (\$900 и \$2100) и записал их в базу данных, после чего завершился. Затем в системе снова был запущен на выполнение первый процесс, уже прочитавший остатки; он вычисляет новые остатки для счетов 515 и 768 на основе ранее прочитанных (\$1500 и \$2000), получая \$1700 и \$1800. Именно эти остатки он и записывает в базу данных. В итоге владелец счёта 768 обнаружит на счету недостаток суммы в \$100 (остаток окажется равен \$1800 вместо \$1900). Как нетрудно убедиться, общая сумма остатков на счетах 301, 515 и 768 уменьшилась на \$100, то есть деньги клиента попросту бесследно исчезли. Могло получиться и наоборот: если бы после чтения был прерван первый процесс, а второй в это время выполнил свою операцию, клиент обнаружил бы на своём счету лишние \$200 (\$2100 вместо \$1900). Клиент, возможно, порадуется, а акционеры банка — вряд ли.

Представим теперь третий процесс, вычисляющий общую сумму остатков на счетах всех клиентов банка. Очевидное решение — прочитать в цикле остатки на каждом счёте и просуммировать их. Однако если во время работы этого процесса какой-нибудь другой процесс произведёт перевод денег с одного из счетов в конце таблицы, сумма которого ещё не учтена, на один из счетов в начале, то есть на такой, баланс которого уже попал в суммирование (см. рис. 7.2, *в*), полученная в результате итоговая общая сумма будет отличаться от реальной как раз на сумму транзакции по переводу, поскольку переведённые деньги окажутся не попавшими в подсчёт ни на новом, ни на старом месте — на новый счёт они пришли уже после того, как он был учтён, а со старого сняты до того, как он был учтён. Как видим, **нежелательные эффекты возможны даже тогда, когда один из участников не производит никаких изменений, а только читает данные**.

Причина получения неправильных результатов в наших примерах в том, что **вычисленный результат основывается на исходных данных, которые к моменту окончания вычислений уже устарели**. В примере с целочисленной переменной (стр. 173) второй процесс записывает в разделяемую переменную значение, вычисленное прибавлением единицы к значению, которое само по себе уже нельзя считать правильным, поскольку его изменил другой процесс. В примере с двумя «встречными» переводами денег (стр. 173) та транзакция, которая завершается позднее, вычисляет новый остаток на счёте 768, пользуясь устаревшими сведениями об исходном количестве денег на нём. Наконец, в примере с сумматором к моменту окончания суммирования успевает устареть информация об остатке на том счёте, который, находясь ближе к началу таблицы, успел поучаствовать в посторонней транзакции между моментом, когда его остаток был учтён в суммировании (то есть повлиял на результат), и моментом окончания суммирования.

Способность данных устаревать — это их фундаментальное свойство, и, как ни странно, в большинстве случаев нет ничего страшного в том, что некий результат основан на данных, которые уже устарели; зачастую данные успевают устареть, пока их передают с компьютера, где их вычисляли, на компьютер, где их будет просматривать пользователь. Проблемы начинаются лишь в тех случаях, когда либо новое значение для каких-то данных вычислено на основании устаревшего значения *тех же самых* данных, либо некое вычисление основывается на данных, часть которых устарела к моменту появления другой части, то есть исходные данные одного вычисления ни в какой момент не были актуальны одновременно; в последнем случае говорят, что была нарушена ***целостность данных***.

7.1.2. Взаимоисключения. Критические секции

Проблем с устареванием данных можно избежать, если сделать так, чтобы процессы, обращающиеся к разделяемым данным, не могли работать одновременно: например, пока работает один процесс, обращающийся к базе данных, другой процесс, которому также нужна эта база данных, блокируется при запуске до тех пор, пока первый не завершится. Такое решение представляется неудачным, поскольку внесение изменений в базу данных может быть не единственной задачей процесса. Вполне возможно, что процесс бóльшую часть времени занимается работой, никак с разделяемыми данными не связанной, причем время жизни процесса может оказаться слишком большим, чтобы на всё это время запрещать доступ к данным кому бы то ни было ещё. Вполне можно представить себе постоянно работающий сервер, с помощью которого клиенты банка могут узнавать остатки на своих счетах. Такой сервер сравнительно редко обращается к базе данных, но при этом работать может месяцами без перерыва. Ясно, что блокировать на всё это время доступ к базе данных других процессов нельзя, ведь это парализовало бы работу банка. В связи с этим вводится понятие ***критической секции***, которое нуждается в дополнительных пояснениях.

Проблема, которую мы решаем, состоит в возможности получения и записи обратно в разделяемую область данных такого значения, которое вычислено на основе других значений, уже устаревших к моменту окончания вычисления. Следовательно, избежать проблем мы можем, если каким-то образом предотвратим устаревание разделяемых данных, которые мы уже начали использовать в своих вычислениях — до окончания вычислений. Интересно, что момент окончания вычислений в разных случаях определяется по-разному. Если вычисляемое нашей программой значение должно быть записано обратно в разделяемое хранилище, где его могут использовать для дальнейших вычислений другие процессы, то вычисления мы можем считать оконченными не раньше, чем будет завершена запись в хранилище последнего из вычисленных результатов. Если же вычисляемый результат предназначен для передачи куда-то «наружу», то есть на нём не будут основываться дальнейшие вычисления в нашей системе — как это подразумевается в примере с вычислением суммы по всем счетам, — то такой результат «имеет право» устареть даже раньше, чем мы предъявим его пользователю, и пользователь, как правило, об этом осведомлён; скорость передачи данных, как и скорость их восприятия человеком — отнюдь не бесконечны, поэтому вычисленные результаты в любом случае могут устареть к моменту их использования, и возможность устаревания исходных сведений во время вычислений здесь ничего качественно нового не добавляет. Единственное ограничение состоит в том, чтобы результат не основывался на устаревших сведениях *на момент получения последних из них*, то есть чтобы исходные сведения не успели устареть по отношению друг к другу. Иначе говоря, результат должен соответствовать состоянию исходных данных на *хотя бы какой-то* момент времени. Время окончания вычислений в этом случае чаще всего определяется как момент, когда последнее из вычисляемых значений вычислено до конца; но можно подойти к вопросу более аккуратно и заметить, что коль скоро само по себе вычисляемое значение «имеет право устареть», можно считать критическую секцию оконченной, когда *прочитано* последнее из значений разделяемых данных, участвующее в вычислении результата.

Можно сказать, что **под критической секцией понимается такая часть программы, в которой производятся логически связанные манипуляции с разделяемыми данными**. Иначе говоря, критическая секция начинается в момент, когда прочитано первое из значений разделяемых данных, способных повлиять на результаты вычисления, и заканчивается, когда вычислены все значения, которые предполагалось вычислить (либо как минимум для них уже получены все исходные данные), и те из значений, которые должны быть записаны обратно в разделяемое хранилище, уже не только вычислены, но и записаны. Так, в примерах предыдущего параграфа критическая секция в процессах, увеличивающих переменную на единицу, начинается перед чтением исходного значения переменной и заканчивается после записи нового значения; критическими секциями в процессах, переводящих деньги со счёта на счёт, будут действия с момента чтения первого остатка до момента записи последнего остатка; в процессе, вычисляющем сумму всех остатков, критическая секция начинается в момент начала просмотра таблицы и заканчивается с его окончанием.

Понятие критической секции очевидным образом дуалистично: речь может идти о *фрагменте программного кода* (от этого места в программе до вот этого), а может — о *временно́м интервале* (с момента начала выполнения вот этого действия до момента окончания вот этого). Конечно, эти две ипостаси понятия критической секции прочно связаны друг с другом: критической секцией во втором смысле оказывается тот временной интервал, в течение которого выполняется критическая секция в первом смысле. Из контекста обычно понятно, о чём идёт речь.

Отметим, что о критической секции можно говорить только в применении к конкретным разделяемым данным. Так, если два процесса обращаются к *разным* данным, пусть и разделяемым с кем-то ещё, друг другу они при этом помешать не могут. Поэтому часто можно встретить выражение *критическая секция по переменной* x или *критическая секция по файлу* f и т. п. Кроме того, важно отличать критические секции, в которых разделяемые данные только читаются, от критических секций, предполагающих изменения разделяемых данных. Последние обычно называют ***модифицирующими критическими секциями***.

Говорят, что две критические секции *конфликтуют* между собой, если среди разделяемых данных, которые в них затрагиваются, есть такие, которые затрагиваются в обеих критических секциях, и при этом минимум одна по этим данным является модифицирующей. Если в двух критических секциях осуществляется доступ к одним и тем же данным, но только на чтение, то между ними никакого конфликта нет и они могут выполняться одновременно.

Такая организация работы процессов, при которой два (и более) процесса не могут одновременно находиться в критических секциях, конфликтующих между собой, называется ***взаимным исключением*** (англ. *mutual exclusion*). Следует учитывать, что взаимные исключения могут породить дополнительные проблемы, так что при выборе метода их реализации нужно соблюдать известную осторожность.

Остановимся на одном моменте, который часто вызывает сложности у студентов. **Процессы, участвующие во взаимоисключении, следуют соглашениям добровольно**. Средства взаимоисключения не создают никаких физических препятствий в обращении к разделяемым переменным; напротив, процессы *самостоятельно* проверяют, разрешён ли им доступ, и лишь после этого такой доступ осуществляют. Иначе говоря, процесс, имеющий доступ к разделяемым данным, *может* обратиться к ним в любой момент, ему ничто (внешнее) не мешает это сделать; просто программы, исполняемые в этих процессах, написаны так, чтобы (сугубо добровольно!) выполнять мероприятия для взаимного исключения и не лезть в критическую секцию, покуда это не окажется позволено в соответствии с используемой моделью взаимоисключения.

Механизм взаимного исключения должен соответствовать определённым требованиям:

- два и более процесса не должны ни при каких условиях находиться одновременно в критических секциях, конфликтующих между собой;
- в программе не должно быть никаких предположений о скорости выполнения процессов и о количестве процессоров в системе (если такие предположения есть, в условиях мультизадачной системы всегда может получиться ситуация, которая в эти предположения не впишется: например, один из процессов может оказаться приостановлен сразу после начала очередного кванта времени из-за наличия в системе процесса с более высоким приоритетом, а другому процессу при этом «повезёт» больше, в итоге он будет выполняться в разы (или даже на несколько порядков) быстрее);
- процесс, находящийся вне критических секций, не должен при этом быть причиной блокировки других процессов;
- недопустима ситуация «вечного ожидания», при которой некоторый процесс никогда не получит доступ в нужную ему критическую секцию (такая ситуация обычно называется ***голоданием***, англ. *starvation*);
- процесс, заблокированный в ожидании разрешения на вход в критическую секцию, не должен расходовать процессорное время, то есть не должно использоваться активное ожидание (в реальности это правило иногда нарушается, если есть уверенность в небольшом времени ожидания; по возможности, однако, активного ожидания следует избегать).

```
for(;;) {
    while(turn != 0) {}
    section();
    turn = 1;
    noncritical_job();
}
```

```
for(;;) {
    while(turn != 1) {}
    section();
    turn = 0;
    noncritical_job();
}
```

Рис. 7.3. Взаимоисключение на основе чередования

```
void enter_section() {
    interested[0] = TRUE;
    who_waits = 0;
    while(who_waits==0 &&
          interested[1])
    {}
}
void leave_section() {
    interested[0] = FALSE;
}
```

```
void enter_section() {
    interested[1] = TRUE;
    who_waits = 1;
    while(who_waits==1 &&
          interested[0])
    {}
}
void leave_section() {
    interested[1] = FALSE;
}
```

Рис. 7.4. Алгоритм Петерсона

7.1.3. Взаимоисключение с помощью переменных

Все подходы, перечисленные в этом параграфе, используют активное ожидание. Специально акцентировать на этом внимание мы не будем, тем более что у этих подходов есть и другие недостатки. Есть у них, впрочем, и достоинство, хоть и всего одно, зато несомненное: для их работы не требуется ничего, кроме обычных переменных, расположенных в разделяемой памяти.

Начнём с подхода, который кажется работающим лишь на первый взгляд, а в действительности взаимоисключения не обеспечивает. Подход называется «***блокировочная переменная***». Пусть имеются некоторые данные, доступные нескольким процессам. Заведём в разделяемой памяти целочисленную переменную (будем называть её `s`); значение переменной `1` будет означать, что с разделяемыми данными никто не работает, а значение `0` — что один из процессов в настоящее время работает с разделяемыми данными и надо подождать, пока он не закончит. При запуске системы присвоим переменной `s` значение `1`. Доступ к данным организуем следующим образом:

```
while(s == 0) {} /* пустой цикл, пока нельзя входить
                                 в критическую секцию */
s = 0;          /* запретили доступ другим процессам */

section(); /* ... работа с разделяемыми данными ... */

s = 1;                              /* разрешили доступ */
```

Предположим, что все процессы, которым нужен доступ к этим данным, будут следовать такой схеме. Возникает ощущение, что оказаться одновременно в критической секции (в примере она показана вызовом функции `section`) два процесса не могут — ведь если один из процессов собрался войти в секцию, он предварительно заносит `0` в переменную `s`, что заставит любой другой процесс, имеющий намерение зайти в критическую секцию, подождать, пока первый процесс не закончит работу в секции и не занесет в `s` снова значение `1`.

К сожалению, всё не так просто. Выполнение процесса может быть прервано (например, прерыванием таймера, в результате которого планировщик сменит текущий процесс) точно в тот момент, когда он уже «увидел» число `1` в переменной `s` и вышел из цикла `while`, но присвоить переменной значение `0` не успел. В этом случае другой процесс может также «увидеть» значение `1`, присвоить `0` и войти в секцию; затем управление будет возвращено первому процессу, но ведь проверку значения он уже выполнил, так что он также произведёт присваивание нуля и войдёт в секцию. В результате оба процесса окажутся в секции одновременно, то есть произойдет то, чего мы пытались избежать.

Пример блокировочной переменной иллюстрирует потребность в ***атомарности*** некоторых действий при организации взаимоисключения. В самом деле, если бы цикл `while` и присваивание `s = 0` в приведённом примере выполнялись как одна неделимая операция без возможности прервать её на середине, то проблема бы не возникла.

Логично приходит в голову идея о *запрете внешних (аппаратных) прерываний* на время выполнения критической секции, но этот вариант неприемлем по целому ряду причин. Во-первых, запрет прерываний годится лишь для кратковременных критических секций: длительное запрещение прерываний нарушит работу аппаратуры — перестанут приниматься и передаваться данные по сети, прекратится запись и чтение файлов и т. д. Во-вторых, запрет прерываний пригоден только для работы с данными, находящимися в оперативной памяти, поскольку для любого обмена с внешними устройствами (в том числе для чтения и записи файлов) аппарат прерываний должен работать. Заметим, при этом нужно тем или иным способом гарантировать, что данные действительно находятся в памяти, ведь если процесс запретит прерывания, а потом обратится к области (виртуальной) памяти, в настоящее время откачанной на диск, операция подкачки либо просто не будет выполнена (что приведёт к аварии), либо операционная система всё-таки разрешит прерывания, что может, в свою очередь, привести к получению управления другим процессом и нарушению взаимного исключения. В-третьих, запрет прерываний обычно касается только одного процессора. В системе с несколькими процессорами (в том числе если это не процессоры, а вычислительные ядра одного процессора) это эффекта не даст.

Далее, в комбинации с активным ожиданием запрет прерываний (в однопроцессорной системе) приведёт к зависанию системы, ведь при запрещённых прерываниях ни один другой процесс (в том числе и виновник блокировки) не может получить управление и снять блокировку, исчезновения которой мы активно ожидаем.

Наконец (и это, пожалуй, самое важное соображение), допускать запрет прерываний пользовательскими процессами слишком опасно. Даже если исключить злой умысел со стороны пользователей, что уже само по себе странно для многопользовательской системы, остаётся возможность случайных ошибок. Если, к примеру, процесс, запретивший прерывания, случайно зациклится, система в результате этого «повиснет» и её придется перезагружать. Поэтому запрет прерываний считается привилегированным действием и для пользовательских процессов недоступен. Отметим, что ядро ОС при этом само достаточно часто использует кратковременные запреты прерываний для обеспечения атомарности некоторых операций; как мы увидим позже, в современных системах обеспечение взаимоисключения иногда возлагается на ядро ОС.

Следующий любопытный способ взаимоисключения, называемый ***чередованием***, заключается в том, что процессы по очереди передают друг другу право работы с разделяемыми данными на манер эстафетной палочки. На рис. 7.3 показаны два процесса, обращающиеся к разделяемым данным (функция `section`) в соответствии с *маркером чередования*, который хранится в переменной `turn`. Значение `0` показывает, что право на доступ к разделяемым данным имеет первый процесс, значение `1` соответствует праву второго процесса. Завершив работу в критической секции, процесс «передаёт ход» другому процессу и приступает к выполнению действий, не требующих доступа к разделяемым данным (функция `noncritical_job`).

Такой способ действительно не даёт процессам оказаться в критической секции одновременно, но имеет, к сожалению, другой недостаток. Если один из процессов, передав ход другому, быстро выполнит все некритические действия и снова попытается войти в критическую секцию, может получиться так, что второй процесс в это время до своей критической секции так и не дошёл (и, соответственно, не передал ход первому процессу). В результате второй процесс, не нуждаясь в доступе к разделяемым данным, тем не менее будет мешать работать первому процессу, то есть одно из сформулированных выше базовых условий окажется напрямую нарушено.

От недостатков предыдущих подходов избавлен ***алгоритм Петерсона***, рассчитанный на случай двух процессов; эти процессы у нас будут фигурировать под номерами `0` и `1`. Аналогичные алгоритмы существуют и для произвольного числа процессов — например, к таким относится известный «алгоритм булочной» (*bakery algorithm*), придуманный Лесли Лампортом; рассматривать его мы не будем, ограничившись алгоритмом Петерсона — он позволяет почувствовать общий принцип. Для взаимоисключения нам в этот раз потребуется создать в разделяемой памяти массив из двух логических переменных `interested[2]`, показывающих, нуждается ли соответствующий (нулевой или первый) процесс в выполнении критической секции; во время выполнения секции соответствующее логическое значение также будет истинным. Кроме того, введём (также в разделяемой памяти) переменную `who_waits`, которая будет показывать, который из процессов в случае столкновения должен подождать завершения критической секции второго.

Показав свою заинтересованность во входе в критическую секцию, то есть присвоив значение «истина» соответствующему элементу массива `interested`, процесс заявляет о своей готовности подождать, если нужно, занеся в переменную `who_waits` свой номер. Затем он будет ждать до тех пор, пока либо не изменится номер `who_waits` (это означает, что второй процесс проявил готовность подождать), либо второй процесс не окажется *не заинтересован* во вхождении в секцию.

На рис. 7.4 алгоритм Петерсона показан в виде двух процедур: `enter_section` (вход в критическую секцию) и `leave_section` (выход из критической секции). Единственным недостатком алгоритма Петерсона и более сложных алгоритмов, построенных на этой идее, таких как алгоритм фильтровый и алгоритм булочной, оказывается активное ожидание. К сожалению, этого вполне достаточно, чтобы считать эти решения неприемлемыми для большинства возникающих ситуаций.

7.1.4. Понятие мьютекса

Подходы к построению взаимного исключения, перечисленные в предыдущем параграфе, характерны наличием ***активного ожидания*** — такого состояния процесса, при котором он в ожидании момента, когда можно будет войти в критическую секцию, вынужден постоянно опрашивать определённые переменные в разделяемой памяти, при этом не выполняя никаких полезных действий, но занимая время центрального процессора. Чтобы процесс, ожидающий входа в критическую секцию, не расходовал попусту процессорное время, следует, очевидно, заблокировать его до тех пор, пока нужные ему разделяемые данные не окажутся свободны, то есть не выделять ему квантов времени, пока не будет известно, что все остальные процессы критическую секцию покинули. С другой стороны, в нужный момент процесс требуется «разбудить», то есть перевести из состояния блокировки в состояние готовности; желательно при этом сразу пометить разделяемые данные как занятые, чтобы процессу не пришлось снова выдерживать конкурентный поединок с другими процессами за вход в критическую секцию. Всего этого можно добиться, если ввести специальные объекты, среди которых особенно примечательны *мьютексы* и *семафоры*.

Под ***мьютексом***[1] в общем случае понимается объект, имеющий два состояния (открыт/заперт) и, соответственно, две операции: `lock` (запереть) и `unlock` (открыть). Обычно мьютекс изначально открыт (создаётся открытым). Операция `lock` в её исходном варианте, если применить её к открытому мьютексу, закрывает его и возвращает управление; при применении её к закрытому мьютексу она блокирует вызвавший процесс до тех пор, пока мьютекс не окажется открыт, после чего закрывает его и возвращает управление. Операция `lock` также может быть реализована как булевская функция; при применении её к открытому мьютексу она закрывает его и возвращает значение «истина» (успех), при применении к закрытому мьютексу функция возвращает значение «ложь» (неудача). Такой вариант реализации называется *неблокирующим*. Обычно реализация мьютекса предоставляет оба варианта операции `lock`.

С момента закрытия мьютекса процесс или тред, выполнивший закрытие, становится «владельцем» мьютекса; открыть мьютекс имеет право только его владелец: кто закрыл, тому и открывать. Операция `unlock`, выполненная владельцем мьютекса, всегда проходит успешно (и немедленно возвращает управление), переводя объект в состояние «открыт».

В действительности многие реализации мьютексов из соображений эффективности никак не проверяют соблюдение правила владения, позволяя открывать мьютекс кому угодно, а не только тому процессу или треду, который его закрыл; из-за этого мьютексы часто путают с ***двоичными семафорами***. Автор книги вынужден признать, что совсем недавно сам входил в число тех, кто путает эти два объекта, и узнал, что был неправ, только после выхода третьего тома первого издания книги; увы, теперь число запутавшихся пополняют читатели первого издания. Даже если реализация мьютексов, с которой вам пришлось работать, не выполняет никаких проверок, всё равно нарушать правило владельца мьютекса не стоит: следующая версия той же библиотеки функций может внезапно оказаться оснащена соответствующей проверкой.

Важнейшее свойство операций `lock` и `unlock` — их *атомарность*. Это означает, что обе операции выполняются как единое целое и не могут быть прерваны (кроме случая, когда операция `lock` блокирует вызвавший процесс — тогда на время блокировки прерывание операции разрешается). В частности, операция `lock` не может быть прервана между проверкой текущего значения мьютекса и изменением этого значения на «закрыто». Вопрос о том, как такой атомарности достичь, мы рассмотрим чуть позже.

Вспомним теперь нашу попытку выполнить взаимоисключение с помощью блокировочной переменной (см. стр. 175). Заменим обычную переменную на мьютекс, обозначив его, как и блокировочную переменную, буквой `s`; вместо присваиваний `s = 0` и `s = 1` применим соответственно операции `lock` и `unlock`. Рассмотрим для начала неблокирующую реализацию `lock`:

```
do {
    locked = lock(s);
} while(!locked); /* повторяем, пока не получим мьютекс */
section();                /* ... критическая секция ... */
unlock(s);          /* разрешаем другим процессам доступ */
```

В отличие от варианта с использованием блокировочной переменной, данный вариант полностью корректен. В самом деле, операция `lock` атомарна, так что выход из цикла (по истинному значению функции `lock`) означает, что в некий момент мьютекс оказался открыт, то есть никто в это время не работал с разделяемыми данными, и нашему процессу удалось его закрыть, причём никто другой вклиниться между проверкой и закрытием не мог. Если применить блокирующий тип реализации операции `lock`, наш код станет ещё проще и из него исчезнет цикл активного ожидания:

```
lock(s);                    /* получаем мьютекс */
section();        /* ... критическая секция ... */
unlock(s); /* разрешаем другим процессам доступ */
```

7.1.5. О реализации мьютексов

Заблокировать процесс может только операционная система. Если бы процесс точно знал, через какой промежуток времени нужный ресурс окажется освобождён, он мог бы выполнить системный вызов, подобный функции `sleep`, чтобы отказаться от выполнения на заданный период. Однако момент освобождения нужного ресурса процессу неизвестен, поскольку зависит от функционирования других процессов. Здесь вполне логично возникает идея возложить управление пометками занятости/освобождения ресурсов на операционную систему, создав ещё один способ взаимодействия процессов. У такого подхода, кроме избавления от активного ожидания, есть и другое важное преимущество. Операционная система, в отличие от процесса, может запретить прерывания на время исполнения определённых действий внутри ядра, обеспечив атомарность сколь угодно сложных операций. При этом исчезает необходимость в ухищрениях, подобных алгоритму Петерсона.

В то же время реализация взаимоисключения полностью средствами ядра имеет недостаток, почти всегда фатальный: системный вызов, как мы знаем, — довольно дорогостоящее действие. К примеру, если в роли разделяемых данных выступает простая целочисленная переменная, а для любых операций с ней нам приходится создавать критическую секцию, обращаясь в начале и в конце секции к операционной системе, то вся затея начисто лишается смысла: намного проще и дешевле будет отказаться от разделяемой переменной, а для взаимодействия процессов использовать какой-нибудь канал или сокет.

К счастью, если процессор поддерживает команду `xchg` (см. т. 1, §3.2.14) или подобную, мьютекс можно реализовать так, чтобы в большинстве случаев обращение к ядру не требовалось. Напомним, что команда `xchg` имеет два операнда, один из которых регистровый, а второй может быть либо регистровым, либо операндом типа «память»; нам потребуется как раз такой вариант. Команда за одно неделимое (атомарное) действие меняет местами содержимое своих операндов, и эта неделимость позволяет реализовать мьютекс; посмотрим, как это делается.

Несмотря на то, что команда `xchg` может работать с операндами любой длины — байтами, словами и двойными словами, — при работе в 32-битном режиме эффективнее всего будет использовать в качестве мьютекса четырёхбайтное целое. Поэтому в роли мьютекса мы будем использовать четырёхбайтную переменную, которую пометим меткой `mutex`; договоримся, что значение `1` в мьютексе соответствует состоянию «открыт», а значение `0` — состоянию «закрыт» (можно и наоборот, конкретные значения не так важны). Операция *открытия* мьютекса реализуется простым занесением единицы в переменную, с этим проблем нет. Если же нужно *закрыть* мьютекс, мы сначала заносим в регистр (например, `EAX`) значение `0`, затем выполняем команду `xchg` между этим регистром и мьютексом и смотрим, какое значение в итоге

[1] Англ. *mutex* — сокращение от слов *mutual exclusion*.

оказалось в регистре. Если там по-прежнему 0, то, стало быть, мьютекс уже был закрыт и наша команда ничего не изменила, он так и остался закрытым, а наша операция закрытия должна быть признана неуспешной и нам надо тем или иным способом подождать, не заходя в критическую секцию. Если же после `xchg` в регистре оказалась единица — значит, мьютекс был открыт, а теперь в результате нашей команды он закрыт.

В самом примитивном варианте операции `lock` и `unlock` можно представить в виде подпрограмм без параметров, а «блокировку» вызывающего процесса реализовать через активное ожидание:

```
section         .data

mutex           dd 1    ; изначально мьютекс открыт

section         .text

mutex_lock:     mov eax, 0
.again:         xchg eax, [mutex]
                cmp eax, 0
                je .again
                ret

mutex_unlock:   mov dword [mutex], 1
                ret
```

Такой вариант будет работать, но, конечно, активное ожидание в большинстве практических случаев неприемлемо. Если представить себе, что у нас есть процедура переключения управления на другой тред того же процесса, можно будет перед переходом на начало процедуры `mutex_lock` вставить обращение к процедуре переключения тредов, чтобы уменьшить потери от активного ожидания:

```
mutex_lock:     mov eax, 0
.again:         xchg eax, [mutex]
                cmp eax, 0
                jne .ok
                call switch_threads
                jmp .again
.ok             ret
```

В практических реализациях всё происходит существенно сложнее. Треды, не получившие нужный мьютекс, помечаются как заблокированные, при этом для каждого мьютекса поддерживается список тредов, ожидающих его открытия, и когда какой-то из тредов открывает этот мьютекс, один из ожидающих этого события тредов разблокируется. Никакой «процедуры переключения тредов» обычно нет: в большинстве случаев тред представляет собой единицу планирования с точки зрения операционной системы, так что передать управление от одного треда другому можно только через планировщик, то есть для этого придётся обратиться к ядру. Но если к ядру всё равно придётся обращаться, проще будет попросить ядро не о переключении треда, а о блокировке данного конкретного треда до тех пор, пока нужный мьютекс не окажется доступен; с другой стороны, тред, «отпускающий» мьютекс, может (что очень важно, сам, без обращения к ядру) проверить, заблокирован ли кто-нибудь из других тредов на этом мьютексе, и если такие есть (и только в этом случае) — попросить ядро кого-нибудь из них разблокировать.

Операционные системы обычно предоставляют специальные системные вызовы для таких случаев; так, в ОС Linux поддерживается вызов `futex`. Описание этого вызова само по себе достаточно сложно, а программирование с его использованием представляет собой сочетание довольно нетривиальных трюков. Библиотека Си даже не предоставляет обёртки для этого вызова; соответствующая страница `man` содержит замечание о том, что этот вызов не предназначен для обычных пользователей (программистов). Технические подробности организации работы с вызовом `futex` любопытствующий читатель может найти, например, в статье [15]; не стоит огорчаться, если текст статьи покажется вам слишком сложным: даже профессиональные программисты зачастую не могут понять, что там к чему.

Следует обратить внимание, что **подавляющее большинство операций блокировки и разблокировки мьютексов не требует обращения к ядру**. Если говорить точнее, то системный вызов потребуется лишь в случаях, когда операция захвата мьютекса прошла неудачно: сначала тому треду или процессу, который захватывал мьютекс, придётся обратиться к ядру с просьбой его усыпить, а потом тред или процесс, который будет мьютекс освобождать, будет вынужден тоже обратиться к ядру, чтобы разбудить кого-нибудь из тех, кто ранее на этом мьютексе заблокировался. Грамотно написанные параллельные программы в такую ситуацию попадают относительно редко, поскольку — если, подчеркнём, делать всё правильно — длительность критических секций невелика в сравнении с общим временем работы программы; ну а до тех пор, пока все операции захвата мьютексов проходят успешно, обращения к ядру не потребуется ни при захвате мьютексов (поскольку блокироваться не нужно, можно сразу работать дальше), ни при их освобождении (поскольку на них никто не блокирован). Важно понимать, что **если ваши треды вынуждены блокироваться и разблокироваться на мьютексах сколько-нибудь часто, ни о какой эффективности речи уже не пойдёт**; такая «распараллеленная» программа, скорее всего, окажется заметно *медленнее*, чем аналогичная ей последовательная, и попасть в эту ситуацию намного проще, чем может показаться.

7.1.6. Семафоры Дейкстры

Так называемые ***семафоры*** на первый взгляд очень похожи на мьютексы: это тоже некие «хитрые» объекты, над которыми определены две атомарные операции, противоположные по смыслу, причём одна из этих операций может заблокировать вызывающий процесс до тех пор, пока какой-то другой процесс не произведёт над тем же объектом противоположную операцию. Подобие настолько очевидно, что, как уже говорилось, эти два термина — мьютекс и семафор — часто путают.

В действительности семафоры и мьютексы различаются как технически, так и по своему предназначению. Если ***мьютекс*** — это механизм корректного захвата разделяемого ресурса, используемый для исключения одновременных обращений к нему, то семафоры — это скорее инструменты *сигнализации*, то есть с их помощью один процесс может сообщить другому о наступлении некоторого события — например, о собственной готовности к какому-нибудь совместному действию. В отличие от мьютексов, для семафоров абсолютно корректна ситуация, когда один процесс закрывает семафор, а другой его открывает.

Строго говоря, семафор представляет собой целочисленную переменную, про которую известно, что она принимает только неотрицательные значения и над которой определены две операции: `up` и `down`. Операция `up` всегда проходит успешно, увеличивая значение переменной на 1, и немедленно возвращает управление. Операция `down` должна, наоборот, уменьшать значение на 1, но сделать это она может только когда текущее значение строго больше нуля, ведь значение семафора не имеет права становиться отрицательным. При положительном значении семафора операция `down` уменьшает его значение на 1 и немедленно возвращает управление. В случае же нулевого значения семафора операция блокирует вызвавший процесс до тех пор, пока значение не станет положительным, после чего уменьшает значение и возвращает управление. Как и в случае с мьютексом, операции над семафором обязательно должны быть реализованы *атомарно*: необходимо полностью исключить ситуации, когда результат нескольких независимых операций над семафором может зависеть от времени вызова операций независимыми процессами.

Ясно, что с помощью семафора можно имитировать мьютекс, если считать значение 0 состоянием «закрыт», значение 1 — состоянием «открыт», а операции `lock` и `unlock` заменить на `down` и `up`. Нужно только следить, чтобы операция `up` никогда не применялась к семафору, уже имеющему значение 1. Больше того, часто можно встретить термин ***двоичный семафор*** (англ. *binary semaphore*): под таковым понимают семафор, способный принимать только значения 0 и 1. От мьютекса он отличается ровно одним: для него не действует «правило владельца», т. е. процесс, закрывший двоичный семафор, не получает в отношении этого семафора никаких исключительных полномочий; открыть его может (и имеет право) любой процесс, как тот же самый, так и другой.

Семафоры, не ограниченные двумя значениями, по-английски называются *counting semaphore*; устоявшегося русского термина пока не появилось. В теории такие семафоры рассматриваются как способные принимать значение от нуля до бесконечности, на практике верхний предел обычно существует (обычно значение семафора хранится в четырёхбайтной переменной, причём даже знаковой), но это ничему не мешает, поскольку в программах просто *не*

бывает разделяемых ресурсов, по своему количеству сколько-нибудь сравнимых с максимально возможным значением четырёхбайтного целого.

Изобретение семафоров обычно приписывают Эдсгеру Дейкстре. Согласно одной из версий, упоминаемых значительно реже, сам Дейкстра предложил объект, который мы называли двоичным семафором, а на произвольные целочисленные значения абстракцию обобщил позже сослуживец Дейкстры и его соавтор по нескольким работам Карел Шолтен (Carel S. Scholten).

Как было сказано в §7.1.4, реализации мьютексов часто не проверяют соблюдение правила владельца. Если предполагать, что они так и не будут ничего проверять, такие мьютексы (технически) вроде бы можно будет использовать вместо двоичных семафоров, и, больше того, можно встретить много программ, где именно так всё и сделано; тому есть причина: например, до не столь давних пор[2] в документации по мьютексам, реализованным в ОС Linux, про правило владельца не было ни слова. Тем не менее, использовать мьютексы в роли семафоров не стоит по двум причинам. Во-первых, в наше безумное время никто не гарантирует, что разработчики библиотек, реализующих мьютексы, не плюнут разом и на эффективность, и на обратную совместимость, и не вставят в реализацию соответствующие проверки; во-вторых, идеологически мьютекс и семафор имеют разное предназначение, так что использование мьютекса «не в той роли» может сбить с толку читателя программы. Чтобы понять, насколько это в действительности разные сущности, можно заметить, что **мьютекс всегда создаётся открытым, тогда как семафоры в большинстве случаев исходно равны нулю**.

В отличие от мьютекса, который следует воспринимать как этакий флажок «занято», семафор представляет собой скорее *счётчик доступных ресурсов*, которых может быть больше одного; именно так его обычно и используют. Чтобы понять, о чём идёт речь, представьте себе, что в критической секции производится *распределение* между работающими процессами чего-то такого, чего не хватает на всех. В этом случае процессу нет смысла заходить в критическую секцию, когда дефицитный ресурс исчерпан: нужно подождать, пока он снова появится. Пример такой ситуации мы рассмотрим в §7.2.1 при обсуждении задачи производителей и потребителей. Что касается двоичных семафоров, то их удобнее воспринимать как способ одному процессу дождаться, пока другой не даст отмашку, означающую «можно работать». С этим мы столкнёмся в одном из решений задачи о пяти философах в §7.2.2, в задаче о читателях и писателях в §7.2.4 и в задаче о спящем парикмахере в §7.2.5.

Некоторые реализации обобщают операции над семафорами, вводя дополнительный целочисленный параметр. Операция `up(sem, n)` увеличивает значение семафора на `n`, операция `down(sem, n)` ждёт, пока значение не окажется бóльшим либо равным `n`, и после этого уменьшает значение на `n`. Кроме того, реализации обычно имеют неблокирующий вариант операции `down`, аналогичный нашей операции `lock` для мьютексов, представленный в виде логической функции; этот вариант вместо ожидания сразу возвращает управление, указывая, что операция прошла неудачно. Большинство существующих реализаций позволяет узнать текущее значение семафора или даже установить его без всяких блокировок и проверок, что бывает удобно при инициализации программы.

Известная реализация семафоров из System V IPC (см. [4]) предоставляет *массивы* семафоров, над которыми можно производить сложные атомарные операции, например, увеличить третий семафор на два, а четвертый уменьшить на три за одно (атомарное!) действие. Кроме того, в этой реализации есть дополнительная операция «блокироваться, пока семафор не станет равен нулю».

Что касается классического семафора без дополнительных операций, то реализовать его можно теми средствами, которые у нас уже есть. В самом деле, возьмём обыкновенный канал (см. §5.3.15) — именованный или неименованный, неважно. Операцию up реализуем в виде записи в канал одного байта с помощью `write` (содержание байта нас в данном случае не волнует), а операцию `down` — в виде чтения байта из канала с помощью `read`. Когда во внутреннем буфере канала нет ни одного байта, вызов `read` заблокируется в ожидании поступления данных, то есть до тех пор, пока кто-нибудь не выполнит запись в канал (наш аналог операции up). Нетрудно видеть, что такое взаимодействие полностью соответствует классическому определению семафора. Единственный нюанс здесь в том, что буфер канала отнюдь не бесконечен, так что у семафора, реализованного таким способом, будет довольно скромная верхняя граница принимаемых значений (для современных версий Linux это будет $2^{16}-1$); но аналогичное ограничение есть у всех реально существующих семафоров, и хотя обычно граница пролегает выше — чаще всего это $2^{32}-1$ — такое различие скорее количественное, нежели качественное, и в любом случае для большинства задач это неважно, поскольку на практике значения семафоров редко превышают несколько десятков. Проблема с такой реализацией, как и с реализацией из System V IPC, в том, что для любой операции над семафором требуется системный вызов, а это, как мы уже отмечали, может свести на нет все преимущества по быстродействию, с таким трудом полученные за счёт применения разделяемой памяти.

7.2. Классические задачи взаимоисключения

В этой главе мы рассмотрим несколько классических примеров программирования с использованием мьютексов и семафоров. Примеры мы постараемся сделать максимально абстрактными, не уточняя, настоящие процессы имеются в виду или треды, какая конкретно используется реализация семафоров и мьютексов и т. п. Приводимые примеры будут даже написаны не на Си, а скорее на си-подобном псевдокоде: обозначив мьютекс буквой `m`, мы будем записывать операции над ним как `lock(m)` и `unlock(m)`, а операции над семафором `s` обозначим как `up(s)` и `down(s)`, не обращая внимания на то, что передача параметров в Си происходит только по значению[3]. Примеры, корректные с технической точки зрения, т. е. представляющие собой реально работающие программы, мы приведём в следующей главе.

7.2.1. Задача производителей и потребителей

Пусть имеется несколько процессов, *производящих* данные, и эти данные нуждаются в дальнейшей обработке. Это могут быть, например, процессы, опрашивающие какие-то датчики; также это могут быть процессы, получающие некую информацию по сети с других машин и преобразующие её во внутреннее представление; возможно, что данные получаются в результате интерактивного взаимодействия с пользователем или, наоборот, вычисляются в ходе математических расчётов. Назовём эти процессы *производителями*.

С другой стороны, есть несколько процессов, обрабатывающих (*потребляющих*) данные, подготовленные процессами-производителями. Эти процессы мы назовём *потребителями*.

Задача производителей и потребителей, предназначенная для иллюстрации применения семафоров, рассматривает следующий вариант передачи информации от производителей потребителям. Пусть каждая порция информации, создаваемая производителем и обрабатываемая потребителем, имеет фиксированный размер. В разделяемой памяти организуем буфер, способный хранить N таких порций информации. Ясно, что работа с буфером требует взаимоисключения. Для упрощения работы будем считать, что буфер представляет собой единое целое[4], и на время работы любого процесса с буфером исключать обращения к буферу других процессов.

Легко видеть, что проблема взаимоисключения в данном случае оказывается не единственной. При взаимодействии потребителей и производителей возможны две (симметричные) ситуации, требующие блокировки:

- потребитель готов к получению порции данных, но в буфере ни одной порции данных нет (буфер пуст);
- производитель подготовил к записи в буфер порцию данных, но записывать ее некуда (все слоты буфера заняты).

В обоих случаях процессу нужно дождаться, когда другой процесс изменит состояние буфера: в первом случае потребителю — когда производитель запишет новые данные; во втором случае, наоборот, производителю — когда потребитель заберёт какие-то уже имеющиеся данные.

Простейшая возможная стратегия — блокировать операции над буфером (войти в критическую секцию), проверить, не изменилось ли соответствующим образом состояние буфера, и если нет, выйти из критической секции, ничего в буфере не поменяв, а затем опять войти, и т. д., то есть просто выполнять циклически проверку изменений. Иначе говоря, такая стратегия представляет собой активное ожидание, только не на входе в критическую секцию, а в процессе работы. Мы уже знаем, что активное ожидание — идея крайне неудачная, и описываемая ситуация — не исключение: в самом деле, *через какое время процессу следует снова заходить в критическую секцию?* Через секунду? Через десятую долю секунды? Или, наоборот, через десять секунд? Никаких сведений для вычисления этого интервала у нас нет, и если

[2] Автор не может точно сказать, в какой момент текст документации претерпел существенные изменения; но в версии man-страницы от 2009 года ограничение «кто закрыл — тому и открывать» не упоминалось.

[3] Заметим, в Си++ имеется передача параметров по ссылке, так что наиболее педантичным читателям мы можем предложить считать наш псевдокод написанным на Си++; кроме того, вы можете считать, что переменные `m` и `s` — это не сами семафоры и мьютексы, а указатели на них или какие-нибудь дескрипторы наподобие файловых, либо что `lock` и `unlock` — не функции, а макросы. Подчеркнём, что для иллюстративных целей, которые мы перед собой ставим в этой главе, всё это неважно.

[4] Это так и есть, если в буфере, например, имеются глобальные данные о том, какие из слотов свободны, а какие заняты.

```
void producer()
{
  /* ... подготовить
         данные ...
   */
  down(empty);
  lock(m);
  put_data();
  unlock(m);
  up(full);
}
```

```
void consumer()
{
  down(full);
  lock(m);
  get_data();
  unlock(m);
  up(empty);
  /* ... обработать
         данные ...
   */
}
```

Рис. 7.5. Производители и потребители

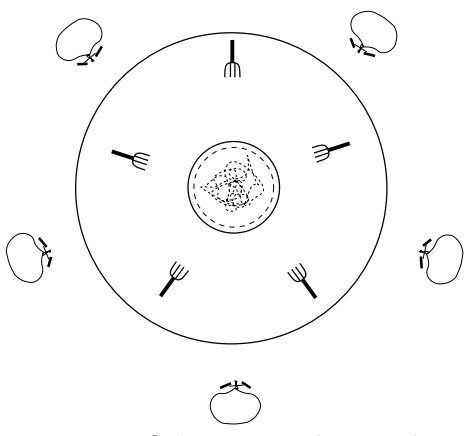

Рис. 7.6. Обедающие философы

мы выберем его слишком большим, это повлечёт задержки в работе взаимодействующих процессов: производители могут продолжать ждать, когда в буфере уже давно появились свободные слоты, потребители могут «висеть», когда в буфере уже лежит для них информация (может быть, даже весь буфер заполнен). Если же временной интервал выбрать слишком коротким, процессы будут в основном заняты «пиханием локтями» на входе в критическую секцию, заходя туда и выходя обратно, возможно, по много тысяч раз, прежде чем удастся сделать что-то полезное; при этом они будут не только расходовать процессорное время, но и мешать друг другу.

Было бы, по-видимому, правильно придумать такой механизм, при котором процессы в такой ситуации вообще не будут заходить в критическую секцию, пока в буфере не появятся данные (для потребителя) либо свободные слоты (для производителя); это и будет решением задачи о производителях и потребителях. Такой механизм можно создать на основе семафоров Дейкстры в качестве счётчиков доступных ресурсов. Задействуем два семафора: один будет считать свободные слоты, и на нём будут блокироваться процессы-производители, если свободных слотов нет; второй будет считать слоты, заполненные данными, и на нём будут блокироваться процессы-потребители, если нет готовых данных. Назовём эти семафоры соответственно `empty` и `full`; в начале работы одновременно с созданием буфера выполним операцию `up(empty)` столько раз, сколько в буфере имеется свободных слотов. Наличие этих семафоров не отменяет необходимости взаимного исключения операций с разделяемым буфером. Для этой цели мы введём мьютекс, который назовём `m`.

Будем считать, что для помещения данных в буфер и извлечения данных из буфера у нас есть процедуры `put_data` и `get_data`, которые не делают никаких операций с семафорами и мьютексами: они написаны в предположении, что процесс уже находится в критической секции и что наличие в буфере нужного ресурса (свободного или занятого слота) кто-то уже гарантировал, и выполняют только рутинную работу вроде «найти нужный слот, скопировать информацию в него или из него, пометить его как занятый или свободный».

Итоговое решение показано на рис. 7.5. Подчеркнём ещё раз, что семафоры в данном случае используются не для взаимоисключения критических секций, а для блокирования процессов, которые всё равно не могут (прямо сейчас) сделать ничего полезного. Взаимоисключение обеспечивает только мьютекс.

7.2.2. Задача о пяти философах и проблема тупиков

При совместной работе нескольких процессов с несколькими разделяемыми объектами возможна ситуация, при которой два или больше участников взаимодействия оказываются в состоянии блокировки, из которого каждый мог бы выйти, если бы другой освободил какой-то из объектов, но этот другой тоже находится в блокированном состоянии и освободить ничего не может. Это называется ***тупиком***.

Полезно помнить, что соответствующий английский термин — *deadlock*, буквально означающий что-то вроде «заклинивания намертво»; точного перевода на русский для этого слова нет. «Тупик» в некомпьютерном смысле по-английски будет *dead end* или вовсе *cul-de-sac*; если попытаться обозначить этими словами ситуацию с заблокировавшими друг друга процессами, вас, скорее всего, не поймут.

Для иллюстрации проблемы тупиков Эдсгер Дейкстра предложил шуточную ***задачу о пяти обедающих философах***. За круглым обеденным столом сидят пять философов, размышляющих о высоких философских материях. В середине стола стоит большая тарелка спагетти. Между каждыми двумя философами на столе располагается вилка, т. е. вилок тоже пять, причём каждый философ может взять вилку слева и вилку справа, если, конечно, вилкой в данный момент не пользуется сосед. Поскольку спагетти отличаются изрядной длиной, для еды каждому философу нужно две вилки[5]. Поэтому каждый философ, поразмыслив некоторое время о непреходящих категориях и решив подкрепиться, пытается сначала взять в левую руку вилку, находящуюся от него слева; если вилка занята, философ с поистине философским спокойствием ожидает, пока вилка не освободится. Завладев левой вилкой, философ точно так же пытается правой рукой взять вилку справа, не выпуская при этом первую вилку из левой руки. Философы, как известно, отличаются философским отношением к житейским трудностям, так что каждый философ готов ждать появления нужной ему вилки хоть до скончания веков — точнее, до наступления голодной смерти, ибо бренные тела, в отличие от просветлённых умов, требуют иногда пищи отнюдь не духовной.

Завладев двумя вилками, философ некоторое время утоляет голод, затем кладёт вилки обратно на стол и продолжает размышлять о вечных вопросах, пока снова не проголодается. Проблема заключается в том, что все пять философов могут проголодаться одновременно (с точностью до времени, затрачиваемого на процедуру завладения вилкой). В этом случае все философы успеют взять вилки в левые руки, да так и замрут с этими вилками в ожидании, когда же появится вилка справа. Однако справа вилка может появиться только тогда, когда правый сосед утолит голод, а этого не происходит, ведь у него тоже только одна вилка. В результате наши достойнейшие мудрые мужи безвременно покинут сей мир, так и не дождавшись вторых вилок. Экая неприятность!

Именно такие ситуации и называются ***тупиками***. Для возникновения тупика, вообще говоря, достаточно двух процессов и двух ресурсов. Пусть имеются два мьютекса `m1` и `m2`. Если первый процесс выполняет код, содержащий вызовы

```
lock(m1);
lock(m2);
```

а второй в это же время выполняет код, содержащий те же вызовы в обратном порядке:

```
lock(m2);
lock(m1);
```

— то при неудачном стечении обстоятельств оба процесса успеют сделать по одному вызову и войдут во взаимную блокировку на вторых, попав таким образом в тупик.

Более того, ситуации взаимоблокировки возможны не только с участием семафоров и мьютексов. Рассмотрим для примера одну такую ситуацию. Пусть нам понадобилось запустить команду `ls` для получения в текстовом

[5] Студенты по разные стороны океана с завидной регулярностью задают вопрос, как же (технически) использовать две вилки для поедания спагетти. Можете как-нибудь на досуге попробовать: одной вилкой накалываете несколько спагеттин, а второй вилкой наматываете их на первую и в таком виде употребляете. Впрочем, позже был предложен другой вариант условий задачи: за столом сидят *восточные* философы, перед ними блюдо с рисом, а на столе лежат не вилки, а палочки для еды. Всем известно, что этих палочек нужно две; правда, держат их всё же одной рукой. Автор этих строк точно помнит, что в каком-то из источников наткнулся на вариант задачи, в котором философы едят креветок, а вторая вилка нужна, судя по всему, чтобы этих креветок разделывать; повторно отыскать тот источник не удалось.

виде списка файлов в текущем каталоге. Начинающие программисты часто делают характерную ошибку, применяя примерно такой код:

```
char buf[100];
int pid, rc;
int fd[2];
pipe(fd);
pid = fork();
if(pid == 0) {
    dup2(fd[1], 1);
    close(fd[1]);
    close(fd[0]);
    execlp("ls", "ls", NULL);
    perror("ls");
    exit(1);
}
close(fd[1]);
wait(NULL);         /* !!! так делать не стоило */
while((rc = read(fd[0], buf, sizeof(buf))) > 0) {
    /* ... */
}
```

Любопытно, что такая программа, вообще говоря, может и заработать, однако может и «зависнуть». Экспериментируя с ней, мы обнаружим, что программа корректно работает в каталогах со сравнительно небольшим количеством файлов, а на больших каталогах «зависает». Прежде чем читать дальше, рекомендуем читателю попытаться самостоятельно догадаться о причинах.

Итак, рассмотрим программу подробнее. Запускаемая нами в порождённом процессе программа `ls` в качестве дескриптора стандартного вывода получает входной дескриптор канала, и прежде чем завершиться, она будет записывать в канал имена файлов из текущего каталога. Между тем родительский процесс, движимый благородной целью недопущения засорения системной таблицы зомби-процессами, выполняет вызов `wait`, в результате чего блокируется до тех пор, пока порождённый процесс не завершится. Лишь после этого родительский процесс выполняет чтение из канала.

В результате получается, что во время работы порождённого процесса (программы `ls`) никто из канала не читает. Как нам известно из §5.3.15, размер буфера канала ограничен[6], и когда буфер заполнится, очередной вызов `write`, выполненный программой `ls`, заблокируется в ожидании освобождения места в буфере. Однако буфер освобождать некому, поскольку родительский процесс, блокированный на вызове `wait`, до первого вызова `read` не дошёл и не дойдет, пока порождённый процесс не завершится.

Как видим, здесь возникает замкнутый круг: предок ожидает, что потомок завершится, и не выполняет чтение из канала, а потомку, чтобы завершиться, нужно, в свою очередь, чтобы предок начал читать. Это и есть, собственно говоря, ***тупик***. Как уже, несомненно, догадался читатель, в данном случае взаимоблокировка возникнет только тогда, когда выдача `ls` для данного каталога превысит размер буфера. Ясно, что приведённое решение очень просто превратить в правильное: достаточно перенести вызов `wait` на несколько строк ниже, чтобы он выполнялся уже *после* выполнения вызовов `read`.

Вернёмся к задаче о пяти философах и попробуем её решить. Для начала рассмотрим вариант, никак не защищённый от вышеописанной тупиковой ситуации. Заведём массив из пяти мьютексов, каждый из которых связан с соответствующей вилкой; обозначим этот массив идентификатором `forks`[7]. И философов, и вилки занумеруем числами от `0` до `4`. Опишем две вспомогательные функции, позволяющие вычислить номер соседа справа и слева:

```
int left(int n) { return (n - 1 + 5) % 5; }
int right(int n) { return (n + 1) % 5; }
```

Будем считать, что номер вилки, лежащей слева от философа, совпадает с номером самого философа. Жизненный цикл философа тогда можно будет представить следующей процедурой:

```
void philosopher(int n)
{
    for(;;) {
        think();
        lock(forks[n]);         /* ! */
        lock(forks[right(n)]);
        eat();
        unlock(forks[n]);
        unlock(forks[right(n)]);
    }
}
```

Ясно, что при одновременном выполнении таких процедур для `n` от `0` до `4` возможна ситуация, когда все они успеют выполнить блокировку, помеченную в листинге восклицательным знаком. При этом все пять «вилок» (мьютексов) окажутся блокированы, так что все процессы также заблокируются на следующей строке процедуры при попытке получить вторую вилку.

Избежать тупика можно, например, применив семафор, не позволяющий философам приступать к трапезе всем одновременно. Заведём семафор и назовём его `sem`. Тогда жизненный цикл философа примет следующий вид:

```
void philosopher(int n)
{
    for(;;) {
        think();
        down(sem);
        lock(forks[n]);
        lock(forks[right(n)]);
        eat();
        unlock(forks[n]);
        unlock(forks[right(n)]);
        up(sem);
    }
}
```

Тупик теперь невозможен, но полученное решение трудно назвать удачным. В самом деле, какое значение присвоить семафору перед началом работы? Если присвоить ему значение `1`, употреблять спагетти в любой момент сможет лишь один философ, остальным придётся ждать. Мы получим нерациональный простой ресурсов, ведь условия позволяют есть двум философам одновременно, не мешая друг другу. Начальное значение, равное двум, не спасёт ситуацию, ведь «по закону подлости» за семафор обязательно пройдут философы, сидящие рядом, так что пока один философ будет кушать, никто другой подкрепиться не сможет: соседу, прошедшему за семафор, не достанется вилки, а остальные за семафор не пройдут.

Очевидно, максимальное возможное значение семафора — четыре, в противном случае теряется его смысл. При таком значении возможна ситуация, когда три философа успели взять по одной вилке и лишь один взял две. Пока он не поест, остальные будут ждать. «Правильного» начального значения здесь просто нет, все возможные значения имеют те или иные недостатки. Впрочем, как мы сейчас увидим, для задачи возможны гораздо более простые решения, хотя с каждым из них связаны определённые проблемы.

Довольно красивое решение — посадить за стол одного левшу и предложить четверым философам по-прежнему сначала брать левую вилку, а затем правую, тогда как левше предложить сначала брать вилку справа, а затем слева. Тупика в этом случае не получится. Официальное название этого метода — «иерархия ресурсов», и это решение предложил сам автор задачи про философов — Дейкстра. В простейшем изложении этого метода предлагается занумеровать разделяемые объекты и оформить все процессы так, чтобы они захватывали объекты в порядке возрастания их номеров. Если использовать нашу нумерацию вилок, то в роли «левши» окажется философ, сидящий между вилками №0 и №4: вилка №0 лежит справа от него, но поскольку `0` меньше, чем `4`, ему в соответствии с предложенным правилом придётся брать её первой.

В общем случае для разделяемых объектов достаточно установить так называемое *отношение частичного порядка* при условии, что *несоразмеримые* (в соответствии с этим отношением) объекты никогда не используются одним процессом. Как ни странно,

[6] До недавних пор размер этого буфера в ОС Linux составлял 4096 байт, но в современных ядрах под буфер канала по умолчанию выделяется 64 КВ; для большинства директорий этого достаточно, что несколько затрудняет возможность увидеть вышеприведённую программу в состоянии тупика. Впрочем, в других версиях ОС Unix буфер может иметь любой размер, возможна даже реализация вовсе без буфера. Но загнать нашу программу в состояние тупика можно и на современных системах: попробуйте добавить команде `ls` ключик `-l` и запустить программу, например, в директории `/usr/bin`.

[7] Английское слово *fork* буквально переводится как «вилка» (которой едят) или «развилка» (на дороге). Системный вызов `fork` назван с использованием второго значения, а здесь мы используем первое.

иногда это оказывается проще, чем перенумеровывать объекты, в особенности когда множество доступных объектов динамически меняется в ходе работы системы.

Решение на основе иерархии ресурсов позволяет избежать тупика, но при внимательном рассмотрении оказывается не слишком удачным с точки зрения эффективности. В применении к философам легко заметить, что возможна ситуация, когда философ № 3 кушает, а все остальные его ждут: философы №№ 0, 1 и 2 при этом держат вилки в левых руках, философ № 4 ничего не держит — он ждёт освобождения вилки № 0. Что касается задач из реального мира, процессам часто требуется не два разделяемых объекта, а больше, причём узнать, какой объект понадобится следующим, процесс может уже после захвата других объектов; такие ситуации — обычное дело при работе с базами данных. Если номер очередного нужного объекта окажется меньше, чем номера уже захваченных, процессу придётся сначала эти объекты освободить, захватить объект с меньшим номером, а затем снова попытаться захватить только что освобождённые объекты; вся эта канитель ведёт к существенным потерям в быстродействии.

Ещё одно достаточно изящное решение задачи о философах состоит в том, чтобы только одному из них разрешать брать вилки. Описывая это решение, обычно говорят о некоем «слуге» или «официанте», который в каждый момент времени может стоять за спиной одного из философов. Ощутив желание подкрепиться, философ поднимает руку и ждёт, когда слуга окажется за его спиной, после чего действует стандартным способом: берёт сначала одну вилку, затем вторую. Слуга, увидев чью-то поднятую руку, подходит к этому философу и не отходит от него до тех пор, пока философ не возьмёт две вилки; лишь когда обе вилки окажутся в руках философа, слуга отходит от него и смотрит, не поднимет ли руку кто-то ещё. Кушать спагетти и класть вилки обратно на стол философы могут без оглядки на слугу. В применении к процессам слуга превращается в простой мьютекс. Если назвать этот мьютекс `steward`, жизненный цикл философа примет следующий вид:

```
void philosopher(int n)
{
    for(;;) {
        think();
        lock(steward);
        lock(forks[n]);
        lock(forks[right(n)]);
        unlock(steward);
        eat();
        unlock(forks[n]);
        unlock(forks[right(n)]);
    }
}
```

К сожалению, это решение тоже не вполне годится: если один из философов кушает, а любой из его соседей проголодался, позвал слугу и, дождавшись его, взял одну из вилок, никто больше не сможет кушать, пока первый философ не насытится — ведь до этого момента его сосед так и будет сидеть с одной вилкой, а слуга — стоять за его спиной.

Задача о философах позволяет также проиллюстрировать ***проблему ресурсного голодания*** (*resource starvation problem*), когда процесс бесконечно повторяет попытки захвата нужных ему ресурсов, но из-за неправильного построения системы взаимоисключений или планирования никогда их не получает. Допустим, философы, не теряя философского отношения к жизни, обрели, тем не менее, толику здравого смысла и решили, что каждому, кто просидит минуту с одной вилкой, следует положить её обратно на стол, подождать десять минут и тогда уже повторить попытку. Такое «простое и очевидное» решение приводит вместо тупика к ресурсному голоданию: если все философы одновременно возьмут по вилке, то через минуту они — опять же одновременно — положат вилки обратно, дружно подождут десять минут и снова — увы, одновременно — возьмут левые вилки. В английском языке для такой ситуации имеется термин *livelock* (в противоположность *deadlock*'у). Проблемы подобного рода чаще всего решаются введением в систему датчика случайных чисел, определяющего, через какой интервал времени следует повторить попытку; в применении к философам можно предложить каждому из них, положив вилку после минуты безуспешного ожидания, бросить игральную кость и повторить попытку взятия вилок через столько минут, сколько очков выпадет на кости. Надо сказать, что современные реализации мьютексов обычно позволяют ограничить время блокировки вызвавшего процесса неким заданным тайм-аутом, но использовать эту возможность следует аккуратно, чтобы не устроить ситуацию голодания.

В книге [7] Э. Таненбаум приводит решение[8], в котором философы, подкрепляясь, при этом не мешают никому, кроме своих соседей. В этом решении каждому философу соответствует переменная, хранящая его *состояние*: `hungry`, `thinking` или `eating`[9]; массив этих переменных назовём `state`. Кроме того, каждому философу соответствует ***двоичный семафор***, на котором он блокируется до того момента, когда ему будет можно приступить к трапезе, чтобы при этом никому не мешать. Таковым считается момент, когда ни один из его соседей (ни слева, ни справа) не приступил к еде и не принял решение приступать к еде. Если в тот момент, когда философ проголодался, оба соседа размышляли, философ сам себе поднимает свой семафор, позволяя самому себе начать трапезу, то есть выполняет операцию `up`; если же один из соседей в этот момент ел, философ свой семафор оставляет нулевым. Несколькими шагами позже философ пытается опустить собственный семафор, что удаётся ему, только если перед этим он его взвёл. В противном случае философ будет ждать в режиме блокировки на семафоре до тех пор, пока сосед, утолив голод, не предложит ему подкрепиться. При этом философ приступит к трапезе только в том случае, если второй его сосед также в настоящий момент не ест; в противном случае он продолжит ждать, уповая на то, что уже второй сосед, насытившись, напомнит нашему мудрецу, что пришло время утолить голод.

Отметим, что в этом решении мьютексы, связанные с вилками, оказываются не нужны: алгоритм и так гарантирует, что два философа не попытаются схватить одну вилку одновременно. Зато нам потребуется один общий мьютекс для защиты массива `state`. Соответствующий код приведён ниже. Центральное место в нём занимает функция `test`. С её помощью каждый философ, проголодавшись, определяет, следует ли ему прямо сейчас приступать к трапезе. Утолив голод, философ вызывает функцию `test` для соседей (это и есть наше любезное предложение подкрепиться), в результате чего, если соответствующий сосед голоден, а соседи соседа в этот момент не едят, происходит «увеличение» (взведение) двоичного семафора и философ, находившийся в состоянии блокировки на нём, приступает к трапезе.

```
enum possible_states { hungry, eating, thinking };
int state[5] =
    { thinking, thinking, thinking, thinking, thinking };
bin_semaphore can_eat[5]; /* начальное значение -- 0 */
mutex state_mut;          /* изначально открыт */
void philosopher(int n)
{
    for(;;) {
        think();
        take_forks(n);
        eat();
        put_forks(n);
    }
}
void take_forks(int i)
{
    lock(state_mut);
    state[i] = hungry;
    test(i);
    unlock(state_mut);
    down(can_eat[i]);
    /* если философ не разрешил сам себе начать трапезу, здесь
       он будет ждать, пока ему о трапезе не напомнят соседи;
       если мы прошли за этот down, то обе вилки заведомо
       доступны, можно их обе брать и приступать к трапезе */
}
void put_forks(int i)
{
    lock(state_mut);
    state[i] = thinking;
    /* теперь любезно поинтересуемся,
       не хотят ли наши соседи кушать */
    test(left(i));
    test(right(i));
    unlock(state_mut);
}
```

[8] Наш текст, приведённый ниже, от решения Таненбаума несколько отличается.
[9] Голоден, думает, кушает (англ.).

```
void test(int i)
{
    if(state[i] == hungry &&
       state[left(i)] != eating && state[right(i)] != eating)
    {   /* настал черед i-го философа поесть */
        state[i] = eating;
        up(can_eat[i]);
    }
}
```

Самое интересное здесь — это то, что даже такое сложное и по-своему красивое решение оказывается не совсем правильным. Несмотря на то, что здесь вроде бы нет никаких тайм-аутов, наш алгоритм оказывается подвержен проблеме голодания: если, к примеру, философы с номерами 1 и 3 окажутся изрядными чревоугодниками и будут наслаждаться поеданием спагетти столь долго и столь часто, что вдвоём «закроют» всю ось времени, сидящий между ними философ № 2 будет вынужден положить зубы на полку, и, в отличие от голодания, возникающего при захватах с тайм-аутами, здесь никакой датчик случайных чисел не поможет.

Ещё одно решение, при котором ни один из философов не мешает кушать никому, кроме своих соседей, было предложено американскими учёными К. М. Чанди и Дж. Мисрой (K. Mani Chandy, Jayadev Misra). Это решение не подвержено проблеме голодания; иной вопрос, что вместо мьютексов для синхронизации используется обмен сообщениями. Вилки в решении Чанди-Мисры вообще никогда не кладутся на стол: философы держат вилки в руках и передают их соседу (левую — левому, правую — правому) по их просьбе, то есть каждая вилка в каждый момент времени находится в руке одного из двух философов. Вилка может быть *чистой* и *грязной*; вилка становится грязной, когда её используют для еды, но каждый философ, протерев вилку салфеткой, может сделать её чистой, и именно так наши вежливые философы и поступают, прежде чем передать вилку соседу. Изначально все вилки считаются грязными; это не совсем логично в наших декорациях, но для решения задачи нужно считать именно так.

Решив подкрепиться, философ проверяет, есть ли у него обе вилки. Для каждой недостающей вилки он направляет соответствующему соседу просьбу о передаче вилки, после чего, если хотя бы одну просьбу пришлось передать (то есть хотя бы одной вилки у него нет), ждёт, пока ему дадут вилки, и тогда начинает кушать. Если у философа есть обе вилки, он может начать кушать немедленно.

Действия философа при получении от соседа просьбы о предоставлении вилки зависят от того, чистая эта вилка или грязная. Если вилка грязная, философ протирает её, делая чистой, и передаёт соседу; интересно, что философ при этом может быть голоден; в этом случае вилку он всё равно отдаёт, но сопровождает передачу просьбой вернуть вилку, когда сосед подкрепится. Если же вилка чистая, философ оставляет её при себе, но запоминает, что сосед просил ему эту вилку отдать. Завершая трапезу, философ вспоминает, не просил ли кто-то из соседей вилки, и если просил — протирает соответствующую вилку, делая её чистой, и отдаёт соседу. Теперь мы можем догадаться, почему изначально все вилки считаются грязными: если исходно считать их чистыми, ни один философ не отдаст свою вилку соседу — и, следовательно, ни один не сможет приступить к еде.

Проблема голодания тут снимается как раз через смену статуса вилок (чистые-грязные). Если бы «чистота» вилок не учитывалась, не в меру активные соседи могли бы то и дело выдёргивать вилки из рук отдельно взятого философа, которому не повезло. Конечно, отдавая вилку, голодный философ всегда просил бы её вернуть, и соседи ему бы её даже возвращали — лишь для того, чтобы едва ли не сразу опять потребовать её назад; такой момент, когда у философа есть обе вилки сразу, мог бы не настать никогда. Но право философа оставить себе вилку, когда она чистая, означает, что философ, *готовый* воспользоваться вилкой (как только получит вторую), но ещё не воспользовавшийся ею, не отдаёт вилку до тех пор, пока *один раз* не поест. В такой системе время ожидания до начала трапезы всегда будет конечным.

Завершим обсуждение задачи о философах вопросом о том, почему их именно пять. Многие специалисты полагают, что это число Дейкстра выбрал случайно, из эстетических или каких-то других внетехнических соображений, но это не так; в действительности именно при пяти действующих лицах задача приобретает какой-то смысл, и сейчас мы в этом убедимся.

При наличии у нас всего одного философа задача вырождается: никаких взаимоисключений больше нет, ситуация тупика невозможна, философ кушает, когда ему вздумается (конечно, вилок ему придётся всё же выдать две) и никто ему в этом не мешает. Этот случай нам попросту неинтересен. Если философов двое, то формально устроить тупиковую ситуацию уже возможно, но при более внимательном взгляде на задачу мы обнаружим, что реализация, при которой каждый из философов (процессов) захватывает каждый из ресурсов (вилок) независимо, изначально идиотична: в самом деле, вилок всё равно хватит только на одного, почему бы в таком случае не рассматривать эти вилки как *один* ресурс, а не два разных? Например, философы могли бы договориться класть вилки в футляр; решив подкрепиться, философ, если нужно, ждёт, пока футляр не окажется на столе, берёт его себе, вынимает из него вилки, кушает, после чего кладёт вилки обратно и возвращает футляр на стол. В применении к реальным программистским задачам это означает, что два имеющихся процесса взаимоисключают доступ к разделяемым данным с помощью одного мьютекса; никакого тупика здесь, понятное дело, не получится.

То же самое можно сказать про случай трёх философов: вилок на двоих всё равно не хватит, так что в любой момент времени кушать может только один из них. Если по какой-то причине нужны именно три вилки, философы могут договориться, к примеру, придвигать к себе блюдо со спагетти, и лишь после этого брать вилки, а после трапезы блюдо отодвигать; опять-таки, взаимоисключение здесь обеспечивается одним мьютексом, и никаких тупиков не возникает.

При наличии четырёх философов ситуация усложняется, поскольку кушать могут уже двое; но и здесь можно обнаружить, что задача имеет упрощённое решение. Когда один из философов подкрепляется, кушать одновременно с ним может лишь тот, кто сидит напротив. Иначе говоря, философы могут есть спагетти исключительно фиксированными парами; внутри каждой из двух пар конфликта доступа к вилкам произойти не может. В терминах философов, вилок и блюда можно предложить, например, такое решение: блюдо, стоящее на столе, сделать не круглым, а вытянутым на манер селёдочницы, и поставить его так, чтобы исходно оно своими концами смотрело между сидящих. Решив поесть, философ смотрит на положение блюда; если оно стоит одним концом к нему, философ просто начинает трапезу, если блюдо стоит в «нейтральном положении», философ его поворачивает одним из концов к себе, и только если блюдо развёрнуто поперёк (что может иметь место лишь в случае, если один или оба философа из противоположной пары сейчас как раз кушают), философ ждёт, пока блюдо не вернётся в нейтральное положение. В реальной программистской задаче можно применить, например, уже известный нам алгоритм чередования; можно также завести переменную, в которой будет храниться в виде четырёх логических флагов список философов, принимающих пищу либо намеренных это делать, и т. д. Так или иначе, решение для двух пар можно сделать качественно проще, нежели для пяти действующих лиц.

Как видим, именно число пять оказывается минимальным, при котором задача об обедающих философах не допускает упрощённых решений. В то же время для бóльшего числа философов — шести, семи, двадцати, сколько угодно, лишь бы не меньше пяти — решение концептуально будет точно таким же, как для пяти, отличаясь чисто количественно.

7.2.3. Граф ожидания

Существуют различные подходы к автоматическому отслеживанию наступления тупиковых ситуаций, и, пожалуй, самый простой из них основан на анализе так называемого ***графа ожидания*** — двудольного ориентированного графа, вершины которого соответствуют процессам (первая доля) и ресурсам (вторая доля). Ситуация «процесс монопольно владеет ресурсом» изображается ориентированной дугой от ресурса к процессу; ситуация «процесс заблокирован в ожидании освобождения ресурса» изображается дугой от процесса к ресурсу. **Появление в графе ожидания ориентированных циклов означает, что система зашла в тупик**.

На рис. 7.7 даны примеры графа ожидания. Слева показан граф ожидания с четырьмя процессами и шестью ресурсами; в этой системе тупиковой ситуации пока нет. В середине приведён пример простейшей тупиковой ситуации (этот пример нами уже рассматривался на стр. 179). Справа показан граф ожидания для задачи о пяти философах в тот момент, когда все пятеро взяли по одной (левой) вилке.

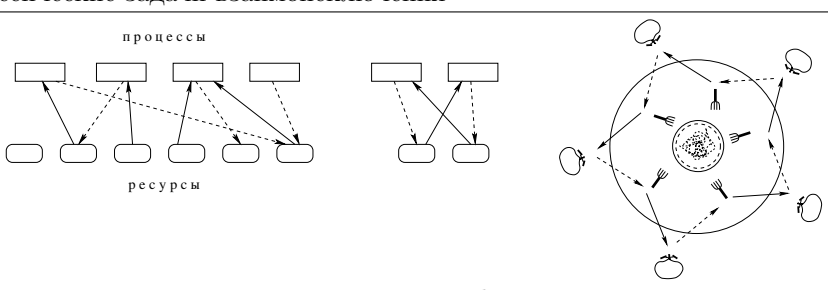

Рис. 7.7. Примеры графа ожидания

Редукция графа ожидания состоит в пошаговом отбрасывании от него дуг, имеющих начало в вершине, в которую ни одна дуга не входит, или конец в вершине, из которой ни одна дуга не выходит. Если в какой-то момент в графе ещё остаются дуги, но ни одна из них не может быть отброшена, констатируется наступление ситуации тупика.

В системах, где количество разделяемых данных и обращающихся к ним процессов делает тупиковые ситуации практически неизбежными, обращения к разделяемым данным объединяют в ***транзакции***, каждая из которых до своего завершения помнит, какие из данных поменяла, и может быть подвергнута ***откату***, при котором все изменённые значения восстанавливаются к состоянию, имевшему место на момент начала транзакции. В графе ожидания в этом случае фигурируют не процессы, а транзакции. Система периодически производит редукцию графа ожидания, а при обнаружении ориентированного цикла сравнивает стоимость отката каждой из транзакций, участвующих в цикле, и, откатив самую дешёвую, продолжает редукцию графа ожидания, пока он не опустеет. Такой транзакционный доступ к данным применяется в системах управления базами данных (СУБД). Способность транзакции откатываться к началу существенно отличает её от критической секции; впрочем, критические секции при транзакционном доступе тоже приходится применять, но они сводятся к тому, чтобы прочитать или обновить одну запись в базе данных и пометить эту запись как закреплённую за транзакцией, так что транзакция за время своей работы может побывать в изрядном количестве критических секций. Транзакционная модель работы позволяет не заводить по мьютексу для каждой записи, что при наличии нескольких миллиардов записей попросту технически невозможно.

7.2.4. Проблема читателей и писателей

Ещё один интересный пример связан с базой данных, к которой одни процессы («читатели») нуждаются только в доступе на чтение, а другие («писатели») могут производить запись (и чтение, разумеется, тоже). Как мы неоднократно убеждались, доступ нескольких процессов на запись одних и тех же данных приводит к проблемам (ситуациям гонок). Даже если модификацией разделяемых данных занимается только один из двух или нескольких процессов, а остальные довольствуются их чтением, это всё равно может привести к ситуации гонок, мы видели такое в примере с подсчётом остатков денег на банковских счетах (см. стр. 173). В то же время процессы, обращающиеся к данным одновременно, но только на чтение (без вмешательства пишущих процессов), помешать друг другу не могут. Задача читателей и писателей состоит в том, чтобы позволить одновременный доступ к данным произвольному числу читателей, но при этом так, чтобы наличие хотя бы одного читателя исключало доступ писателей, а наличие хотя бы одного писателя исключало доступ вообще кого бы то ни было, включая и читателей.

Для решения задачи введём общую переменную, которая будет показывать текущее количество читателей (назовём её `rc` от слов *readers count*). Это позволит первому пришедшему читателю узнать, что он первый, и блокировать доступ к базе для писателей, а последнему уходящему читателю — узнать, что он последний, и разблокировать доступ. Здесь придётся обратить внимание, что снимать блокировку доступа во многих случаях будет вынужден не тот, кто её установил, так что мьютекс здесь не годится; поэтому для блокировки доступа к данным — тем, которые пишутся писателями и читаются читателями — мы воспользуемся двоичным семафором, который назовём `db_access`. Кроме того, для защиты целостности переменной `rc` нам потребуется мьютекс, который мы назовём `rc_mutex`.

Процедура записи в область общих данных («писатель») будет достаточно простой:

```
void writer(...)
{
    down(db_access);
    /* ... пишем данные в общую память ... */
    up(db_access);
}
```

Процедура «читателя» окажется существенно сложнее, поскольку требует манипуляций с переменной `rc`. Читатель прежде всего проверяет, не первый ли он среди читателей. Если есть другие читатели, работающие в настоящий момент с общей памятью, читатель просто присоединяется к ним, отразив факт своего присутствия в переменной `rc`; если же других читателей нет, первый пришедший читатель должен будет получить доступ в секцию (для всех читателей сразу), возможно, дождавшись, пока секцию не освободит писатель. Для этого первый читатель сначала дожидается, пока критическую секцию не покинет писатель (если, конечно, он там есть) — это делается опусканием семафора `db_access`. Если в это время к входу в критическую секцию подойдут другие читатели, они блокируются на мьютексе, защищающем переменную `rc` (`rc_mutex` в этот момент всё ещё удерживает читатель, пришедший первым). Когда вход в критическую секцию окажется открыт, первый из пришедших читателей сумеет опустить семафор, после чего отпустит мьютекс `rc_mutex`, что позволит заходить в секцию и другим читателям. Выходя из критической секции, каждый читатель будет отражать факт своего выхода в переменной `rc`, проверять, не последний ли он, и если последний — поднимать семафор `db_access`, разрешая доступ писателям.

```
void reader(...)
{
    lock(rc_mutex);
    rc++;
    if(rc == 1)                    /* первый! */
        down(db_access);
    unlock(rc_mutex);
    /* ... читаем данные из общей памяти ... */
    lock(rc_mutex);
    rc--;
    if(rc == 0)        /* уходя, гасите свет */
        up(db_access);
    unlock(rc_mutex);
}
```

Приведённое решение имеет серьёзный недостаток, неизменно ускользающий от внимания теоретиков. Если читателей много, их число может очень долго не достигать нуля: одни читатели будут уходить, выполнив свою задачу, но на смену им будут появляться другие. Для иллюстрации рассмотрим какой-нибудь телефонный справочный колл-центр линий на пятьсот. Если линий и операторов достаточно для обслуживания потока входящих звонков, то дождаться, пока хотя бы одна линия окажется свободна, будет несложно; но попробуйте дождаться такого момента, когда *ни одна* линия не будет занята! Такой момент может не наступить ни разу в течение нескольких лет. Представим себе, что операторы, приняв звонок клиента, зачитывают ему какой-нибудь текст, записанный на большой доске на стене колл-центра, то есть выступают в роли *читателей*, а нам нужно изменить текст на доске — часть его стереть и записать новый на его место (исполнить роль *писателя*). Чтобы не сбивать операторов с толку, мы можем решить подождать такого момента, когда ни один из операторов не будет разговаривать с клиентом. Долго же нам придётся ждать!

В параграфе, посвящённом пяти философам, мы уже встречали эту проблему и знаем, что она называется *голоданием* (*starvation*). Решить эту проблему оказывается неожиданно просто, добавив в систему ещё один мьютекс, который мы назовём `barrier`. Все читатели, прежде чем хотя бы попытаться зайти в критическую секцию, будут этот мьютекс закрывать и сразу же открывать; если мьютекс изначально открыт, никаких препятствий читателям он не создаст. Что касается писателя, то он, намереваясь войти в критическую секцию, заблокирует этот новый мьютекс, обозначив таким образом свои намерения. Заметим, что читатели никоим образом не могут помешать писателю это сделать, поскольку закрывают мьютекс лишь на очень краткое время, после чего сразу же его открывают; если же мьютекс уже закрыт другим писателем, то это никак не ухудшает нашу работу, ведь писатели всё равно не могут работать одновременно. После закрытия `barrier` писатель,

скорее всего, заблокируется на попытке опустить `db_access` (этого не произойдёт лишь в случае, если ни одного читателя в секции не было).

С момента блокировки мьютекса `barrier` ни один новый читатель в критическую секцию попасть более не сможет, все они будут блокироваться на том же мьютексе, при этом читатели, уже прошедшие за барьер, беспрепятственно продолжат работу, но рано или поздно все они её завершат и покинут критическую секцию. Последний из них, уходя, откроет семафор `db_access`, разрешив работу писателю. Писатель, выполнив свою задачу, откроет и `db_access`, и `barrier`, и читатели, в том числе те, что ждали на закрытии `barrier`, снова смогут заходить в критическую секцию; фрагмент с закрытием и открытием `barrier` они преодолеют по очереди, но поскольку каждый из них сделает это достаточно быстро, суммарное время, которое потребуется всем ранее заблокированным на «барьере» читателям, чтобы пройти за «барьер», тоже будет невелико. Обновлённый вариант решения будет выглядеть так:

```
void reader(...)
{
    lock(barrier);        /* барьер на входе */
    unlock(barrier);
    lock(rc_mutex);
    rc++;
    if(rc == 1)
        down(db_access);
    unlock(rc_mutex);
    /* ... читаем данные из общей памяти ... */
    lock(rc_mutex);
    rc--;
    if(rc == 0)
        up(db_access);
    unlock(rc_mutex);
}

void writer(...)
{
    lock(barrier); /* закрываем барьер для читателей */
    down(db_access);
    /* ... пишем данные в общую память ... */
    up(db_access);
    unlock(barrier);              /* отпираем барьер */
}
```

7.2.5. Задача о спящем парикмахере

Задачу о спящем парикмахере, как и задачу о пяти философах, часто приписывают Эдсгеру Дейкстре. Приведём классическую формулировку этой задачи. Имеется парикмахерская с одним парикмахером, у которого есть кресло для работы. Кроме того, в парикмахерской есть холл, в котором стоит N стульев для клиентов, ждущих своей очереди.

Закончив стричь и отпустив очередного клиента, парикмахер идёт в холл, и если там есть ждущие клиенты, провожает одного из них в рабочее кресло и принимается стричь. Когда клиентов нет, парикмахер возвращается к своему креслу, усаживается в него и засыпает.

Клиент, зайдя в парикмахерскую, должен посмотреть, занят ли парикмахер, и если нет — то разбудить его; проснувшись, парикмахер усадит разбудившего его клиента в кресло и будет стричь. Если же парикмахер занят, клиент направляется в холл и либо занимает там один из стульев и ждёт, пока его позовут стричься, либо, если свободных стульев нет, уходит из парикмахерской нестриженым.

В реальной парикмахерской с этим сценарием работы вроде бы не должно возникнуть никаких проблем, но у нас ведь речь на самом деле идёт о взаимодействующих процессах, выполняющих определённые действия: проверить, есть ли клиенты, посмотреть, не спит ли парикмахер, занять один из стульев и т. д. Каждое такое действие в системе взаимодействующих процессов будет занимать какое-то время, причём это время, вообще говоря, заранее не известно, так что проблемы нам, как водится, обеспечены. Представим себе, к примеру, что парикмахер кого-то стрижёт, при этом холл забит почти под завязку — там есть всего одно свободное место. В парикмахерскую заходят одновременно два новых клиента, видят, что парикмахер занят, идут в холл, видят там одно свободное место и оба одновременно пытаются на него сесть; ничего хорошего из этого не выйдет.

Пусть теперь парикмахер стрижёт единственного клиента, в комнате ожидания при этом никого нет. В какой-то момент в парикмахерскую заходит клиент, заглядывает к парикмахеру, видит, что тот занят, и направляется в холл, чтобы сесть там на стул. В это же самое время парикмахер заканчивает стрижку, отпускает клиента и тоже направляется в холл за очередным клиентом, причём действует достаточно быстро, так что *опережает* только что пришедшего клиента, видит пустой холл (потому что клиент до стула ещё не дошёл), возвращается к себе и засыпает. Уже после этого клиент всё-таки добирается до холла, усаживается на свободный стул и ждёт. В терминах процессов получается, что они оба заблокировались, каждый в ожидании действия второго, и вывести из этого состояния их может только кто-то третий (в парикмахерскую должен прийти ещё один клиент), а это может случиться нескоро.

Задача о спящем парикмахере допускает разные решения в зависимости от используемых средств синхронизации. Чаще всего в литературе встречается решение, в котором используется целочисленная переменная для хранения количества свободных мест в холле, мьютекс для защиты этой переменной, двоичный семафор, показывающий, занят ли парикмахер (именно на нём блокируются клиенты в ожидании, когда их будут стричь; поднятием этого семафора парикмахер приглашает на стрижку очередного клиента) и обычный семафор (*counting semaphore*), равный количеству ожидающих клиентов — на нём засыпает парикмахер, когда клиентов нет. Как ни странно, это наиболее популярное решение оказывается неполным. Очень редко, но всё же можно встретить решение, учитывающее, что клиент в кресле окажется не мгновенно и не следует позволять парикмахеру стричь пустое кресло; кроме того, сама стрижка тоже происходит не моментально, так что надо позаботиться и о том, чтобы клиент не сбежал из кресла прямо во время стрижки. Такое решение мы (для начала) и приведём; у него тоже есть недостатки, но об этом позже.

Переменную, равную количеству свободных стульев, мы назовём `seats`, мьютекс для её защиты — `seats_mutex`; семафор, равный количеству ожидающих клиентов, будет называться `customers`. Кроме того, нам потребуются три двоичных семафора: `barber` для блокировки клиентов в ожидании стрижки или, если угодно, для приглашения клиентов на стрижку (название оправдывается тем, что значение этого семафора равно количеству свободных парикмахеров); `client` для блокировки парикмахера в ожидании, пока приглашённый клиент не усядется со всеми удобствами в парикмахерское кресло; и `seat_belt`[10] для блокировки клиента в ожидании завершения стрижки. Начальные значения всех наших семафоров будут нулевые; мьютекс, напротив, изначально открыт (но это общее свойство всех мьютексов).

```
int seats = TOTAL_SEATS_COUNT;
mutex seats_mutex;
semaphore customers;
bin_semaphore barber;
bin_semaphore client;
bin_semaphore seat_belt;
```

Парикмахер будет работать по следующей схеме:

```
for(;;) {
    down(customers);      /* засыпаем, если стричь некого */
    lock(seats_mutex);
    up(barber);           /* приглашаем клиента */
    seats++;
    unlock(seats_mutex);
    down(client);         /* ждём, пока он усядется */
    BARBER_WORK();        /* стрижём */
    up(seat_belt);        /* отпускаем */
}
```

Здесь стоит обратить внимание, что семафор `barber`, на котором «спят» клиенты, поднимается внутри критической секции по переменной `seats`. Если поднять его до начала критической секции, то клиент, зашедший в парикмахерскую в это же самое время, может зайти в критическую секцию, увидеть, что свободных стульев нет и уйти, хотя на самом деле один из стульев только что освободился. Напротив, если его разблокировать после критической

[10] Словосочетанием *seat belt* в английском языке обозначаются страховочные ремни — в машине или, например, в самолёте.

секции, то такой (зашедший в парикмахерскую «в неподходящий момент») клиент может попытаться занять *несуществующий* свободный стул: в самом деле, между критической секцией и приглашением стричься возникает момент, когда счётчик свободных стульев уже увеличен, но клиента пока никто не стрижёт, так что реально он всё ещё занимает стул в холле (количество клиентов, заблокированных на семафоре `barber`, равно количеству стульев). Парикмахер избегает обеих некорректных ситуаций, соединив два действия — приглашение очередного клиента на стрижку и освобождение стула в холле — в одно неделимое мероприятие, когда никто другой не может «влезть» в происходящее.

Поведение клиента будет чуть более сложным:

```
lock(seats_mutex);
if(seats > 0) {
    up(customers);      /* если парикмахер спал --
                           это его разбудит */
    seats--;
    unlock(seats_mutex);
    down(barber);       /* если парикмахер занят -- ждём */
    CUSTOMER_PREPARE(); /* усаживаемся в кресло */
    up(client);         /* разрешаем начать стрижку */
    down(seat_belt);    /* ждём окончания стрижки */
} else {
    /* мест нет -- придётся уйти нестриженым */
    unlock(seats_mutex);
}
```

Клиент начинает с того, что входит в критическую секцию — для доступа к переменной `seats`, но не только. Если в холле нет свободных мест, клиент сразу же освобождает мьютекс, выходя тем самым из критической секции, и покидает парикмахерскую, не получив желаемого. Если же свободные места в холле есть, клиент для начала поднимает на единичку семафор (возможно, разбудив при этом парикмахера), затем занимает одно из мест в холле (уменьшает переменную `seats` на единицу), выходит из критической секции и ждёт, когда его пригласят стричься — для этого он пытается опустить семафор `barber`, в результате чего блокируется до тех пор, пока парикмахер этот семафор не поднимет; если есть несколько клиентов, заблокировавшихся на семафоре `barber`, то согласно определению семафора каждый раз, когда парикмахер его поднимает, разблокируется при этом только один из клиентов — именно он и идёт стричься.

Можно заметить, что *клиент будит парикмахера* внутри критической секции по той же переменной. Этому тоже есть причина, но, что довольно забавно, наше «полное» решение, включающее семафоры `client` и `seat_belt`, не позволяет эту причину проиллюстрировать. Выше мы упоминали, что у этого решения есть недостатки; самое время их рассмотреть.

Основной (и, если можно так выразиться, фундаментальный) недочёт тут состоит в том, что введение семафоров `client` и `seat_belt` заметно сужает задачу: оно предполагает, что во взаимодействии парикмахера и клиента (в ходе стрижки) роль клиента полностью пассивна, то есть ему нужно только дойти до кресла, сесть — и больше никаких активных действий до окончания стрижки от него не потребуется. В реальной жизни такое бывает, но редко (например, «парикмахер» может что-то делать с областью памяти, в другое время закреплённой за «клиентом»); чаще задача стоит так, что и парикмахер, и клиент в ходе стрижки должны что-то делать, так что клиент не может всё время стрижки провисеть на семафоре. Видимо, поэтому решение задачи о спящем парикмахере в большинстве источников выглядит иначе — вот так:

```
/* парикмахер */
for(;;) {
    down(customers);
    lock(seats_mutex);
    up(barber);
    seats++;
    unlock(seats_mutex);
    BARBER_WORK();
}
```

```
/* клиент */
lock(seats_mutex);
if(seats > 0) {
    up(customers);
    seats--;
    unlock(seats_mutex);
    down(barber);
    CUSTOMER_WORK();
} else {
    unlock(seats_mutex);
}
```

Впрочем, своя шероховатость есть и у этого решения — недаром мы выше назвали его «неполным». Здесь подразумевается (но не оговаривается) наличие внутри процедур `BARBER_WORK` и `CUSTOMER_WORK` «чего-то ещё», каких-то дополнительных синхронизирующих действий, не позволяющих, например, парикмахеру махнуть ножницами над пустым креслом и считать, что стрижка окончена, а клиенту — приземлиться в кресло, тут же с него соскочить и убежать в полной уверенности, что его уже подстригли. Проблема в том, что свойства этого вот «чего-то» явным образом не проговариваются, во всяком случае, ни в одном из тех источников, которые удалось найти автору книги; если попытаться их проговорить, решение просто станет другим, например, таким, какое мы привели выше (с семафорами `client` и `seat_belt`) или каким-то ещё, но другим.

Предположим, однако, что мы не знаем, как внутри устроены `BARBER_WORK` и `CUSTOMER_WORK`, но при этом кто-то очень добрый заверил нас, что наша задача — корректно довести парикмахера и клиента до входа в эти процедуры, чтобы они оказались внутри одновременно (неважно, кто первый, кто второй), а дальше они там сами разберутся. С учётом такой оговорки решение вроде бы становится правильным и даже позволяет понять, почему клиент обязательно должен разбудить парикмахера внутри критической секции по переменной `seats`. Согласно исходной формулировке задачи клиент должен сначала посмотреть, занят ли парикмахер, и если нет, то разбудить его, а уже потом идти занимать стул в холле. Если, соблюдая эту последовательность, клиенты станут будить парикмахера до входа в критическую секцию, то два клиента одновременно могут увеличить семафор (тем самым они заявят парикмахеру, что уже ждут, когда их постригут), но потом один из них обнаружит, что ему не хватает стула, и будет вынужден уйти; казалось бы, он при этом может уменьшить семафор, и всё будет в порядке, но это не так: если клиент попадётся достаточно нерасторопный, то за время, прошедшее между обнаружением отсутствия стула и уменьшением семафора, парикмахер может успеть перестричь всю имеющуюся очередь клиентов и попытаться пригласить стричься клиента, который вообще-то на это не рассчитывает: он уже принял решение уйти из парикмахерской нестриженым, то есть его программа уже выполняется по ветке, не предполагающей ни стрижку, ни ожидание. В терминах наших мьютексов и семафоров это будет означать, что парикмахер повторно поднимет семафор `barber` (в зависимости от реализации двоичного семафора здесь либо произойдёт ошибка, либо `barber` получит непредусмотренное значение 2, что ещё хуже) и попытается начать стрижку, не имея для этого клиента.

Если же клиенты будут сначала занимать для себя стул, выходить из критической секции, а потом уже пытаться будить парикмахера, то понять, что́, собственно, не так, оказывается сложнее.

Можно придраться к тому, что несколько клиентов могут начать тормошить уснувшего парикмахера одновременно. Исходным условиям задачи это несколько противоречит, но в терминах наших семафоров тут ничего страшного не произойдёт, поскольку действие «разбудить парикмахера» у нас никак явно не обозначено, ведь клиенты просто выполняют `up(customers)`, и их дальнейшие действия никак не зависят от того, разблокировало это парикмахера или нет: сами клиенты в любом случае блокируются на строчке `down(barber)`, а парикмахер, выполнив `up(barber)`, по определению семафора разблокирует кого-то одного, что в целом и требуется. Мало того, если мы заявим, что наш клиент проделывает `up(customers)` внутри критической секции именно по такой причине, то кто-нибудь очень внимательный может нам ответить, что возможна ведь и такая ситуация, что клиент займёт стул, но дальше его что-то отвлечёт и он долго не будет выполнять свою строчку `up(customers)`; в этом случае парикмахер может благополучно заснуть на своём `down(customers)`, несмотря на наличие клиента. Здесь тоже получается так, что формально это противоречит условиям задачи, хотя в терминах используемых средств синхронизации всё в итоге окажется в порядке: клиент рано или поздно доберётся до строчки `up(customers)`, парикмахер снова проснётся и радостно примется этого клиента стричь. Но самое интересное, что от этого нас выполнение `up(customers)` внутри критической секции никак не спасёт; и чем, спрашивается, эта ситуация «лучше» предыдущей? Да и вообще, если посмотреть на ситуацию внимательнее, окажется, что для буквального соблюдения исходных условий задачи придётся изрядно усложнить реализацию и для клиента, и для парикмахера, чтобы, например, парикмахер, прежде чем опустить семафор, проверял, приведёт ли это к блокировке, и если да, то выдавал какое-нибудь сообщение, что он отправляется спать, а клиент, поднимая тот же семафор, проверял, разблокировал ли он при этом

парикмахера, и тоже выдавал сообщение, что только что парикмахера разбудил; кроме того, клиент, разбудивший парикмахера, вовсе не должен занимать стул, а парикмахер, если он только что проснулся, вроде бы не должен идти за очередным клиентом в холл, забирать оттуда клиента и увеличивать количество свободных стульев, ведь клиент-то к нему уже пришёл. От попытки реально написать такое решение мы, пожалуй, воздержимся.

Содержание этого параграфа вообще успело обрасти довольно долгой историей. В первом издании книги было приведено только «неполное» решение, и его неполнота никак не была оговорена. Во *втором* издании, как и в том, которое вы сейчас читаете, были приведены оба решения, но что ускользнуло от внимания автора — это то, что введение семафора `client` заодно снимает ту проблему, которую вообще-то должно решать как раз пробуждение парикмахера клиентом внутри его критической секции по переменной `seats`. На тот факт, что объяснение, почему якобы нельзя будить парикмахера вне этой секции, с введением дополнительных семафоров больше не работает, обратил внимание один из читателей; нам всем, надо сказать, повезло, что этот внимательный читатель — один за почти четыре года — всё-таки нашёлся. Что здесь обескураживает больше всего — это что во всех источниках, какие автору удалось отыскать, воспроизводится (с незначительными вариациями) одно из двух решений, при этом не приводится никаких ссылок на возможность второго решения, не делается никаких попыток их сравнить, не даётся никаких (!) пояснений по конкретной последовательности действий, а когда решение показывается без семафоров `client` и `seat_belt` — никак не оговариваются внутренние свойства процедур `BARBER_WORK` и `CUSTOMER_WORK`. Складывается ощущение, что с этой «классической» задачей что-то глобально пошло не так.

7.3. Многопоточное программирование в ОС Unix

В области программирования, как и в других инженерных дисциплинах, никакая теория не стоит ломаного гроша, если она не подкреплена практикой. Наш рассказ о критических секциях, мьютексах, семафорах и классических задачах останется пустопорожней болтовнёй, если всё это будет не на чем попробовать. Автор этих строк, будучи жёстким противником применения многопоточного программирования, вынужден, тем не менее, признать, что работу с разделяемыми данными проще всего осваивать именно на тредах. Фактически единственная доступная в unix-системах альтернатива — семафоры и разделяемая память System V, но эти инструменты, объединённые общим названием System V IPC и включающие ещё очереди сообщений, производят впечатление мертворожденных.

Кроме того, как уже говорилось, даже если мы отказываемся от использования многопоточности, делать это нужно осознанно и при этом точно знать, от чего конкретно мы отказались. В частности, к сожалению, у вашего будущего работодателя может возникнуть странное подозрение, что вы, мол, просто не знаете, как работать с тредами и потому не хотите с ними работать; чтобы опровергнуть подобные домыслы, вам нужно будет знать, как работать с тредами, и уметь это продемонстрировать.

Прежде чем приступить к рассказу о функциях работы с тредами, напомним некоторые базисные моменты. **Легковесные процессы, они же «треды», запускаются в рамках одного обычного процесса**, работают в его адресном пространстве (попросту говоря, в его памяти) и, как следствие, имеют ничем не ограниченный доступ ко всем переменным своего основного процесса, а равно и других тредов, работающих в нём; естественно, основной процесс, называемый также «основным тредом», тоже имеет ничем не ограниченный доступ ко всем переменным всех своих «дополнительных» тредов. Каждый тред представляет собой самостоятельную *единицу планирования* с точки зрения планировщика времени центрального процессора; кроме того, у каждого треда имеется свой собственный стек, что и понятно, ведь нужно же ему где-то располагать параметры функций, локальные переменные и адреса возврата; но на этом «самостоятельность» треда заканчивается, даже небезызвестный идентификатор процесса (`pid`) у тредов «один на всех» — вызовы `getpid` и `getppid` во всех тредах одного процесса возвращают одни и те же значения.

Вообще, если речь идёт о тредах, то можно обнаружить, что практически ничего *действительно своего* у них нет. Даже стек, который заводится специально для каждого запускаемого треда, в действительности располагается в общем адресном пространстве, так что тред этим стеком пользуется, но нельзя в полной мере сказать, что он им *владеет*.

7.3.1. Библиотека Posix Threads

В 1995 году в состав стандарта POSIX вошло описание функций управления тредами под общим названием ***Posix Threads*** (pthread). Как водится, интерфейс выглядит отвратительно, но статус стандарта фактически лишил конкурирующие спецификации шансов на успех; сейчас pthread — единственный вариант тредов, на наличие которого в системе можно рассчитывать. К счастью, в наши планы не входит серьёзное использование этого интерфейса; наша задача — *попробовать* работу с тредами, для этого нужна хотя бы какая-то их реализация, так что сойдёт и эта. Впрочем, возможна ли вообще сколько-нибудь красивая реализация многопоточного программирования — вопрос открытый.

Согласно спецификации Posix Threads, у треда должна быть главная функция (аналогично тому, как у процесса есть функция `main`) следующего вида:

```
void *my_thread_main(void *arg)
{
    /* ... */
}
```

Как мы видим, треду в качестве стартового параметра можно передать указатель на произвольную область памяти (`void*`); тред может при завершении сообщить другим потокам результат своей работы в виде опять-таки произвольного указателя. Здесь прослеживается некоторая аналогия с обычными процессами, которые при запуске получают в качестве аргумента главной функции командную строку, а при завершении формируют числовой код возврата (см. §5.3).

Каждый тред имеет свой уникальный идентификатор, который можно сохранить в переменной типа `pthread_t`. Для создания (запуска) треда используется функция `pthread_create`:

```
int pthread_create(pthread_t *thr, pthread_attr_t *attr,
                   void*(*start_routine)(void*), void *arg);
```

Параметр `thr` указывает, в какую переменную следует записать идентификатор нового треда. Аргумент `attr` позволяет задать специфические параметры; в большинстве случаев такие параметры не нужны, так что можно в качестве этого аргумента передать нулевой указатель. Параметр `start_routine` указывает на главную функцию треда. Именно эта функция будет запущена в новом треде, причём на вход ей будет передан указатель, который при вызове `pthread_create` мы указали в качестве параметра `arg`. Функция `pthread_create`, как и все функции этого семейства, возвращает `0` в случае успеха либо код ошибки, если создать новый тред не удалось; для `pthread_create` это может быть только код `EAGAIN`, который означает, что для создания треда не хватило системных ресурсов или было достигнуто предельное количество тредов для одного процесса.

Тред может завершиться двумя способами: вернув управление из своей главной функции подобно тому, как процесс возвращает управление из функции `main`, или вызвав функцию `pthread_exit`:

```
void pthread_exit(void *retval);
```

В первом случае результатом работы треда станет значение, возвращённое из главной функции (напомним, оно имеет тип `void*`), во втором случае — значение аргумента `retval`. Здесь снова прослеживаются аналогии с управлением процессами, на сей раз — с функцией `exit`.

Тред может дождаться завершения другого треда с помощью функции `pthread_join`:

```
int pthread_join(pthread_t th, void **result);
```

Аргумент `th` задаёт идентификатор треда, завершения которого мы хотим подождать. Через параметр `result` передаётся *адрес* указателя типа `void*`, в который следует записать результат работы треда. Это несколько напоминает функционирование вызова `waitpid` для обычных процессов.

Результат выполнения завершённого треда должен где-то храниться; если его не востребовать вызовом `pthread_join`, он будет впустую занимать системные ресурсы, как это происходит с процессами (зомби). Однако для тредов этого можно избежать, переведя тред в ***«отсоединённый» режим*** (англ. *detached mode*). Это делается функцией `pthread_detach`:

```
int pthread_detach(pthread_t th);
```

Недостаток «отсоединённых» тредов в том, что их невозможно дождаться с помощью `pthread_join`, так что нет штатного способа проанализировать результат их работы. Функции `pthread_detach` и `pthread_join` возвращают, как и `pthread_create`, `0` в случае успеха или код ошибки, если выполнить действие не удалось.

Узнать свой собственный идентификатор тред может с помощью функции `pthread_self`:

```
pthread_t pthread_self();
```

Например, тред может перевести сам себя в «отсоединённый режим», выполнив вызов

```
pthread_detach(pthread_self());
```

Тред может досрочно завершить другой тред, вызвав функцию

```
int pthread_cancel(pthread_t th);
```

В этом случае результатом работы треда `th` будет специальное значение `PTHREAD_CANCELED`. Следует отметить, что вызов `pthread_cancel` не уничтожает тред, а *отменяет* его, что, вообще говоря, не всегда приводит к немедленному прекращению выполнения другого треда: возможно, что тред завершится, только дойдя до вызова одной из функций библиотеки pthread, входящей в число ***точек отмены*** (англ. *cancellation points*). Такие функции, кроме основных действий, производят проверку наличия запроса на отмену данного треда. Список функций, являющихся точками отмены, можно узнать из документации на `pthread_cancel`.

7.3.2. Семафоры и мьютексы

Библиотека pthread включает как семафоры, так и мьютексы, но мьютексы из pthread были изначально предназначены для взаимоисключений между тредами, тогда как интерфейс семафоров исходно был задуман как допускающий взаимодействие между обычными процессами; поэтому функции для работы с семафорами, как мы увидим чуть позже, не имеют префикса `pthread_` в именах.

В качестве мьютексов pthread использует переменные типа `pthread_mutex_t`, определение которого в большинстве реализаций достаточно сложное, но нас оно, вообще говоря, волновать не должно — это проблемы создателей конкретной реализации библиотеки. Начальное значение такой переменной, соответствующее состоянию «мьютекс открыт», следует задать инициализатором `PTHREAD_MUTEX_INITIALIZER`, например:

```
pthread_mutex_t my_mutex = PTHREAD_MUTEX_INITIALIZER;
```

Возможны и другие, более сложные варианты инициализации мьютекса, в том числе с помощью специальной функции, однако для наших иллюстративных целей достаточно одного. Следует обратить внимание, что `PTHREAD_MUTEX_INITIALIZER` представляет собой именно *инициализатор*, т. е., вообще говоря, попытка *присвоить* мьютексу это значение, скорее всего, приведёт к ошибке при компиляции: результат макроподстановки этого макроса может содержать (и практически всегда содержит) фигурные скобки, допустимые в инициализаторе, но не предусмотренные в обычных выражениях.

Напомним, что под мьютексом понимается объект, способный находиться в одном из двух состояний (открытом и закрытом), над которым определены две операции: открытие (`unlock`) и закрытие (`lock`), причём первая всегда переводит мьютекс в открытое состояние и возвращает управление, вторая же, если ее применить к открытому мьютексу, закрывает его и возвращает управление, если же её применить к закрытому мьютексу, может либо вернуть управление, сигнализируя о неудаче (неблокирующий вариант), либо блокировать вызвавший процесс (или, в данном случае, поток), дождаться, пока кто-то не откроет мьютекс, закрыть его и только после этого вернуть управление (блокирующий вариант). В pthread основные операции над мьютексами производятся с помощью функций

```
int pthread_mutex_unlock(pthread_mutex_t *mutex);
int pthread_mutex_lock(pthread_mutex_t *mutex);
int pthread_mutex_trylock(pthread_mutex_t *mutex);
```

Эти функции осуществляют открытие мьютекса (`unlock`), блокирующее закрытие мьютекса (`lock`) и неблокирующее закрытие мьютекса (`trylock`). Все функции возвращают `0` в случае успеха или ненулевой код ошибки, причём в случае, если `pthread_mutex_trylock` применяется к закрытому мьютексу, она возвращает код `EAGAIN`. Мьютекс можно уничтожить вызовом функции `pthread_mutex_destroy`:

```
int pthread_mutex_destroy(pthread_mutex_t *mutex);
```

которая высвободит используемые мьютексом ресурсы, если таковые есть; в реализации мьютексов в ОС Linux весь мьютекс целиком умещается в переменной `pthread_mutex_t`, так что высвобождать оказывается нечего. На момент уничтожения мьютекса он должен находиться в состоянии «открыт», иначе функция вернёт ошибку, и если говорить о реализации мьютексов, имеющейся в ОС Linux, то проверкой этого условия действия `pthread_mutex_destroy`, собственно говоря, и ограничиваются. В других реализациях это может быть не так, поэтому, когда очередной мьютекс становится вам не нужен, к нему, прежде чем соответствующая переменная исчезнет, нужно обязательно применить `pthread_mutex_destroy` и проверить, что она вернула значение «успех» (т. е. ноль). Впрочем, это следует сделать, даже если ваша программа предназначена для работы исключительно в Linux, поскольку физическое уничтожение переменной-мьютекса, на которой кто-то из ваших тредов всё ещё заблокирован, обычно приводит к аварии, но не сразу: программа может ещё некоторое время проработать и лишь затем «свалиться», так что найти причину может оказаться очень сложно. Ненулевое значение, возвращённое при попытке уничтожить мьютекс, позволит понять, что ваша программа работает некорректно, и исправить ошибку, не тратя лишнее время на поиски.

Семафоры POSIX представляются переменными типа `sem_t`; ещё раз отметим, что префикс `pthread_` тут не используется. Дело в том, что спецификация семафоров POSIX исходно предназначалась для работы не только с тредами, но и с обычными процессами; хотя мы такой вариант рассматривать не будем, всё же придётся иметь в виду, что он был предусмотрен.

Некоторые реализации семафоров POSIX не предусматривали возможности взаимодействия через них для обычных процессов, выдавая ошибку при попытке такого их использования; в частности, так до сравнительно недавних пор обстояли дела в Linux. Современные реализации, в том числе и в Linux, такое использование допускают; для этого переменную-семафор нужно разместить в разделяемой памяти и при её инициализации указать, что к ней предполагается разделяемый доступ (т. е. доступ из разных процессов).

Инициализация семафора производится функцией `sem_init`:

```
int sem_init(sem_t *sem, int pshared, unsigned int value);
```

Параметр `sem` задаёт адрес инициализируемого семафора. Параметр `pshared` указывает, будет ли семафор доступен для других процессов; если наша переменная типа `sem_t` описана как обычная переменная (не размещена в разделяемой памяти), то такое использование в любом случае невозможно, так что параметр `pshared` должен быть равен нулю. Наконец, параметр `value` задаёт начальное значение семафора. Функция `sem_init` возвращает `0` в случае успеха, `-1` в случае ошибки, а сам код ошибки заносится в переменную `errno`. Как видим, даже подход к возвращаемому значению для «семафорных» функций отличается от принятого в `pthread`; здесь используется конвенция, характерная для системных вызовов. В дальнейшем изложении мы не будем делать на этот счёт специальных оговорок, поскольку *все* функции, составляющие интерфейс к семафорам POSIX, ведут себя именно так: возвращают `0` при успехе, а при неудаче возвращают `-1` и заносят код ошибки в `errno`. Пусть вас, впрочем, не обманывает используемая конвенция; ни одна из этих функций не является системным вызовом и вообще не обращается к ядру, кроме случая, когда им не остаётся иного выхода, кроме как «заснуть», либо когда нужно «разбудить» кого-то ещё.

Как мы помним, семафор есть по определению объект, внутреннее состояние которого представляет собой неотрицательное целое число, и над которым определены две операции: `up` и `down`. Первая из них увеличивает значение семафора на `1` и немедленно возвращает управление. Вторая, если значение семафора равно нулю, блокирует вызвавший процесс или поток до тех пор, пока кто-то другой не увеличит значение семафора (если значение изначально ненулевое, блокировки не происходит), после чего уменьшает значение семафора на `1` и возвращает управление. Для семафоров POSIX соответствующие операции выполняются функциями

```
int sem_post(sem_t *sem);    /* up   */
int sem_wait(sem_t *sem);    /* down */
```

Также имеется неблокирующий вариант операции `down`:

```
int sem_trywait(sem_t *sem);
```

Если семафор на момент вызова имеет значение `0`, эта функция, вместо того чтобы блокировать вызвавший процесс, немедленно завершается, возвратив значение `-1` и установив `errno` в значение `EAGAIN`.

Текущее значение семафора можно узнать с помощью функции `sem_getvalue`:

```
int sem_getvalue(sem_t *sem, int *sval);
```

Значение семафора возвращается через параметр `sval`.

Если семафор больше не нужен, его следует ликвидировать с помощью функции `sem_destroy`:

```
int sem_destroy(sem_t *sem);
```

При этом не должно быть ни одного треда (а для случая «разделяемых» семафоров — ни одного процесса), находящегося в состоянии ожидания на этом семафоре, то есть выполняющего в настоящий момент `sem_wait` с тем же параметром, что и `sem_destroy`. Надо сказать, что с семафорами в этом плане дело обстоит гораздо жёстче, чем с мьютексами: `sem_destroy` не проверяет, есть ли или нет такие треды или процессы, а просто приводит к неопределённому поведению (*undefined behaviour*).

7.3.3. Демонстрационный пример

Читатель мог обратить внимание, что раньше мы всегда старались приводить примеры, имеющие хотя бы намёк на какую-нибудь задачу, способную возникнуть на практике. Увы, такой пример для демонстрации Posix Threads придумать крайне сложно; как уже говорилось, задач, которые реально могли бы оправдать применение многопоточного программирования за пределами ядра ОС, попросту не существует в природе, а задачи, в которых мультитрединг хотя бы не выглядит проявлением идиотизма, в принципе существуют, но слишком сложны для учебной иллюстрации. Поэтому мы рискнём подойти к демонстрации мультитрединга с другой стороны. Программа, которую мы напишем, будет на практике абсолютно бесполезна, но она, можно надеяться, позволит нам преодолеть остаточное неверие в то, что с разделяемыми данными всё столь серьёзно.

В нашем примере головная программа создаст массив из некоторого количества (например, из десяти) целочисленных элементов и заполнит этот массив нулями, после чего запустит несколько (неважно сколько, это количество можно варьировать от запуска к запуску; двух уже достаточно) «рабочих» тредов, каждый из которых будет в бесконечном цикле изменять случайно выбранные элементы массива, причём делать это всегда парами — один элемент увеличивать на единицу, а другой на единицу уменьшать. Ясно, что общая сумма элементов массива при этом должна оставаться той же, что и была — нулевой. Головная программа раз в секунду будет вычислять и печатать сумму элементов массива, а также и значения самих элементов.

Столь простенькой демонстрации нам хватит, чтобы показать категорическую необходимость взаимоисключения доступа к разделяемым данным. Для этого мы сделаем механизм критических секций *отключаемым* — всё, что относится к работе с мьютексами и семафорами, обвесим директивами условной компиляции (см. §4.6.6) так, чтобы в командной строке компилятора можно было одним флажком устранить из нашей программы все «хитрые» объекты с атомарными операциями. Запустив нашу программу, откомпилированную с корректной организацией критических секций, можно будет убедиться, что всё работает именно так, как должно: элементы массива в каждый момент времени отличаются от нуля в обе стороны (возможно, не все; любой элемент может случайно оказаться равным нулю), но сумма остаётся всегда нулевой. После этого можно будеть перекомпилировать ту же программу, «вырубив» механизмы взаимоисключения, и, запустив её, понаблюдать, сколь быстро сумма, которая вроде бы должна была оставаться нулевой, «улетает в космос» — обычно семизначных значений она достигает за считанные секунды.

Интересно, что даже такой примитивный вроде бы пример содержит серьёзный подводный камень, из-за которого нам придётся воспользоваться ещё и семафором, правда, только двоичным, не счётным — или, лучше будет сказать, *благодаря* этому подводному камню мы сможем продемонстрировать много чего интересного, включая работу с семафором. Но обо всём по порядку.

По условию задачи каждый из тредов на каждой итерации своего цикла должен сгенерировать два случайных числа, чтобы выбрать два элемента массива, с которыми он произведёт свои операции. Функции `srand` и `rand`, которые мы рассматривали в §4.10.3, вроде бы позволяют получать последовательности псевдослучайных чисел, но не тут-то было: применять эти функции в многопоточной программе нельзя. Дело в том, что, как мы обсуждали ещё в первом томе (см. §2.8.5), при генерации псевдослучайных чисел используется глобальная переменная; своё начальное значение она получает при инициализации генератора (вызовом процедуры `randomize` в Паскале, функции `srand` в программах на Си), а функция, возвращающая очередное случайное число (в стандартной библиотеке Си это `rand`), заодно *изменяет* значение этой переменной. Следовательно, вызовы функции `rand` в тредах нам пришлось бы объявить критическими секциями и обвесить мьютексами, ещё больше замедлив и так не слишком быструю программу и ещё сильнее увеличив количество коллизий. Кроме того, такое решение имеет ещё один недостаток: обычно последовательность псевдослучайных чисел можно повторить, используя от запуска к запуску одно и то же начальное значение для генератора; это может оказаться очень полезно при отладке. Но ведь треды могут обращаться к общему для них генератору в непредсказуемом порядке, так что конкретная последовательность случайных чисел, полученных одним отдельно взятым тредом, окажется принципиально невоспроизводимой.

К счастью, в библиотеке предусмотрена специально для таких случаев ещё одна функция — `rand_r`:

```
int rand_r(unsigned int *seedp);
```

Эта функция не обращается ни к каким глобальным переменным; вместо этого вызывающий должен сам указать, какую конкретно переменную (имеющую тип `unsigned int`) следует использовать в роли «затравки», так что можно завести свою переменную в каждом из тредов. Но это вовсе не конец истории: сразу же возникает вопрос, а каким начальным значением инициализировать эту переменную, чтобы в разных тредах последовательности случайных чисел не повторялись? Функцию `srand` мы, как несложно вспомнить, вызывали, передавая ей аргументом текущее время (результат вызова `time`), но здесь такой номер не пройдёт: все треды стартуют настолько быстро, что в них во всех, скорее всего, `time` вернёт одно и то же число.

Придётся поступить хитрее. Начальное значение «затравки» для генерации случайных чисел для каждого треда будет вычислять головная программа. Чтобы обеспечить разнообразие от запуска к запуску, она вызовет функцию `time`, но только один раз; ну а различие последовательностей в разных тредах будет обеспечено простым прибавлением единицы к полученному значению перед запуском каждого следующего треда.

Итак, у нас появляется некая информация, которую нужно передать в тред при старте — начальное значение для генерации псевдослучайных чисел. Можно легко заметить, что это не единственная такая информация. По условиям задачи тред должен работать с неким массивом, а доступ к массиву выполнять в критической секции, защищённой мьютексом. Конечно, мы могли бы сделать и массив, и мьютекс глобальными переменными, но мы ведь помним, что глобальные переменные — вселенское зло, а задача у нас явно не настолько сложна, чтобы это зло соглашаться терпеть.

Всё, что мы можем передать в тред — это указатель типа `void*`, но его, по идее, должно хватить, ведь указывать-то он может на что угодно, в том числе, например, на структуру, содержащую сколько угодно полей. Так что мы можем описать структуру вроде следующей:

```
struct thread_start_data {
    unsigned int randseed;
    int *array;
    pthread_mutex_t *arr_mutex;
};
```

— и, заполнив перед запуском очередного треда все её поля, указать адрес самой структуры (точнее, переменной этого типа) четвёртым параметром функции `pthread_create`, так что головная функция треда получит этот адрес через свой единственный параметр.

На этом месте стоит немного сбавить обороты и задуматься. Вот у нас стартует тред, его главная функция, по-видимому, начинает свою работу с анализа полей структуры, указатель на которую передан в тред через параметр... но что при этом — в это же самое время! — происходит в головной программе? По-видимому, она, запустив наш тред, готовится к запуску следующего треда — то есть заполняет поля структуры и вызывает функцию `pthread_create` уже для запуска следующего треда, и этот следующий тред получает указатель на структуру — только ведь физически это, скорее всего, *та же самая* структура, то есть ровно та же область памяти, адрес которой был сообщён и нашему треду тоже. Как говорится, опаньки: кто первый успеет, функция нашего треда — проанализировать содержимое структуры, или же главная программа — изменить это содержимое?

Можно, конечно, завести для каждого треда свой экземпляр структуры, которая будет ему передаваться параметром, но стоит ли? В главной программе тогда появится массив структур, размер которого будет соответствовать количеству тредов; элементы массива, заметим, будут различаться значением только одного поля — значения для генератора псевдослучайных чисел, остальные поля там указывают на одни и те же объекты. Главной программе придётся следить, чтобы этот массив существовал, пока существуют треды (в нашей задаче для этого ничего делать не нужно, но не все задачи столь просты), а если ей придёт в голову запустить (по какой-то причине) ещё несколько тредов, возникнет уже совсем интересная проблема: размер массива нужно будет как-то увеличить, но ведь перемещать его в другую область памяти нельзя, адреса-то тредам уже переданы, как они узнают, что предназначенные для них структуры куда-то «уползли»?

Более правильным представляется другой способ решения проблемы. Пускай тред, получив управление, копирует все данные из полученной структуры в свои локальные переменные, после чего сообщает головной программе, что структурой более не пользуется, так что головная программа теперь вольна делать с этой структурой что ей заблагорассудится. Головная же программа, запустив тред, дождётся, пока он закончит работу со структурой и сообщит об этом, и только после этого приступит к запуску следующего треда.

Для такой сигнализации идеально подходят двоичные семафоры, но специального объекта для двоичного семафора библиотека не предусматривает, так что придётся использовать обыкновенный семафор. Его начальным значением будет ноль; запустив очередной тред, головная программа будет выполнять на семафоре операцию `down` (в терминах pthread — `sem_wait`), и эта операция, что вполне естественно, головную программу заблокирует до тех пор, пока кто-то не сделает операцию противоположную (`up`, т. е. `sem_post`). Именно это и сделает тред, когда закончит извлечение информации из переданной ему структуры — тем самым разрешив головной программе двигаться дальше.

Обсудив всё это, мы готовы приступить к написанию текста программы. Для начала введём несколько макроконстант:

```
#define ARRAY_SIZE 10     /* размер массива */
#define WORKERS_COUNT 3   /* количество тредов */
#define PAUSE_LENGTH 1
     /* пауза между итерациями печати значений */

#ifndef PROTECTION
#define PROTECTION 1      /* 1 означает использовать
        семафоры и мьютексы, 0 -- не использовать их */
#endif
```

В структуру, которую мы уже пытались описать выше, придётся добавить ещё одно поле — указатель на семафор; кроме того, указатели на мьютекс и семафор в структуру будут включаться только в случае, если их использование вообще предполагается (то есть `PROTECTION` имеет значение `1`). Получится следующее:

```
struct thread_start_data {
    unsigned int randseed;
    int *array;
#if PROTECTION == 1
    pthread_mutex_t *arr_mutex;
    sem_t *data_sem;
#endif
};
```

Напишем теперь главную функцию для треда. Она будет копировать сведения из переданной ей структуры в свои локальные переменные, сообщать (через семафор) главной программе, что более в структуре не нуждается, и приступать к работе: в бесконечном цикле получать два случайных индекса элементов массива, если вдруг они выпали одинаковые — на этом итерацию заканчивать; если же индексы разные — входить в критическую секцию, уменьшать на единицу первый из выбранных элементов массива, увеличивать на единицу второй, выходить из секции и возвращаться к началу цикла. Несмотря на то, что цикл у нас получился бесконечный и выхода из него не предусмотрено, в конце функции придётся поставить оператор `return`, чтобы компилятор не выдавал предупреждения. Всё вместе будет выглядеть так:

```
static void *worker_thread(void *v_data)
{
    struct thread_start_data *data = v_data;
    unsigned int randseed = data->randseed;
    int *array = data->array;
#if PROTECTION == 1
    pthread_mutex_t *arr_mutex = data->arr_mutex;
    sem_t *data_sem = data->data_sem;
    sem_post(data_sem);  /* структура больше не нужна */
#endif
    for(;;) {
        int idx1, idx2;
        idx1 = rand_r(&randseed) % ARRAY_SIZE;
        idx2 = rand_r(&randseed) % ARRAY_SIZE;
        if(idx1 == idx2)
            continue;
#if PROTECTION == 1
        pthread_mutex_lock(arr_mutex);
#endif
        array[idx1]++;
        array[idx2]--;
#if PROTECTION == 1
        pthread_mutex_unlock(arr_mutex);
#endif
    }
    return NULL;  /* сюда мы никогда не попадём */
}
```

Очередь за функцией `main`. Массив чисел, которые будут меняться в тредах, мьютекс для организации критических секций по этому массиву, структура данных для передачи в треды стартовой информации и семафор для сигнализации о прекращении использования этой структуры мы создадим в виде локальных переменных функции `main`; кроме того, нам потребуется вспомогательная переменная типа `pthread_t` и счётчик для цикла запуска тредов. Начало функции `main` получится таким:

```
int main()
{
    int i;
    int the_array[ARRAY_SIZE];
    struct thread_start_data tsdata;
    pthread_t thr;
#if PROTECTION == 1
    pthread_mutex_t arr_mutex = PTHREAD_MUTEX_INITIALIZER;
    sem_t tsd_sem;
#endif
    for(i = 0; i < ARRAY_SIZE; i++)
        the_array[i] = 0;
```

Подготовим структуру `tsdata` к работе (в дальнейшем меняться будет только поле `randseed`, остальные значения одинаковы для всех тредов):

```
    tsdata.randseed = time(NULL);
    tsdata.array = the_array;
#if PROTECTION == 1
    tsdata.arr_mutex = &arr_mutex;
    tsdata.data_sem = &tsd_sem;
    sem_init(&tsd_sem, 0, 0);
#endif
```

Для запуска очередного треда нужно увеличить `randseed` на единицу (можно, впрочем, было бы изменить это значение как-то ещё, лишь бы оно не повторялось; увеличение на единицу просто выглядит наиболее простым), вызвать `pthread_create` и дождаться, пока новый тред закончит свои дела с переданной ему структурой. Полностью цикл запуска тредов получится таким:

```
    for(i = 0; i < WORKERS_COUNT; i++) {
        tsdata.randseed++;
        pthread_create(&thr, NULL, worker_thread, &tsdata);
#if PROTECTION == 1
        sem_wait(&tsd_sem);
#endif
    }
```

Осталось написать главный цикл. Все взаимосвязанные действия с разделяемыми данными нужно производить в критической секции, в противном случае, например, напечатанная сумма может оказаться (да что там «может», непременно окажется!) не равна сумме напечатанных значений элементов. Поэтому на каждой итерации цикла блокируем мьютекс, вычисляем и печатаем всё что надо, после этого отпускаем мьютекс и «спим»:

```
    for(;;) {
        int sum;
#if PROTECTION == 1
        pthread_mutex_lock(&arr_mutex);
#endif
        sum = 0;
        for(i = 0; i < ARRAY_SIZE; i++)
            sum += the_array[i];
        printf("%d ", sum);
        for(i = 0; i < ARRAY_SIZE; i++)
            printf("%c%d", i ? ',' : '(', the_array[i]);
        printf(")\n");
#if PROTECTION == 1
        pthread_mutex_unlock(&arr_mutex);
#endif
        sleep(PAUSE_LENGTH);
    }
```

Остаётся только написать, как и в функции треда, оператор `return`, до которого управление никогда не дойдёт, и завершить функцию:

```
    return 0;
}
```

Полностью текст нашей программы вы найдёте в файле `balance.c`. Получить две версии исполняемого файла — одну «правильную», другую с отключённой защитой разделяемых данных — можно следующими командами:

```
gcc -Wall -g balance.c -lpthread -o balance
gcc -DPROTECTION=0 -Wall -g balance.c -lpthread -o bal_noprot
```

Работа этих двух исполняемых файлов будет выглядеть примерно так (обратите внимание, что сумма печатается первым числом в строке):

```
avst@host:~/work$ ./balance
0 (229,-238,229,252,180,-230,123,-329,-26,-190)
0 (37,794,-1099,691,1314,1522,843,-93,-1566,-2443)
0 (1195,-62,-1045,803,501,1471,1566,-1547,-1417,-1465)
0 (1830,-973,-709,403,-433,1523,2546,-1012,-1117,-2058)
0 (3296,243,-712,475,-2456,284,4561,-2943,-112,-2636)
^C
avst@host:~/work$ ./bal_noprot
0 (-3,-21,36,-14,-16,-24,25,20,-30,24)
12262 (5228,5522,-11410,466,-4974,-2277,700,16762,-1474,3723)
15357 (4755,-3842,-11539,-626,-12064,1315,10849,25307,1678,-473)
17517 (-324,-1442,-8564,1972,-12759,-6501,11932,21774,9060,2374)
36603 (-10590,4993,-1237,22182,-11303,-873,13228,7811,18242,-5850)
^C
avst@host:~/work$
```

Как говорится, почувствуйте разницу.

В первом издании книги, а также в книжке «Введение в операционные системы» 2006 года издания был приведён пример реализации средствами pthread задачи о производителях и потребителях. Пример этот был, во-первых, притянут за уши, поскольку взаимодействие тредов давало заведомо больше потерь, нежели выигрыша от распараллеливания; и, во-вторых, содержал серьёзные ошибки. Исправленный вариант того примера по-прежнему доступен в архиве примеров под именем `prod_cons.c`.

Заключение

Подчеркнём ещё раз, что с механизмом тредов можно поэкспериментировать, но прежде чем его использовать на практике, следует очень серьёзно подумать. Обратиться к разделяемым данным можно случайно, не заметив этого, а в некоторых случаях даже не подозревая, что что-то подобное происходит: например, ваш коллега по проекту или автор какой-нибудь библиотеки может использовать в своих функциях глобальную или статическую переменную, но забыть вам об этом сказать, и вы, вызвав его функции из разных тредов, устроите себе ситуацию гонок; что самое противное, такая ситуация гонок может никак себя не проявлять, пока вы не закончите тестирование, а затем «выстрелит» в самый неподходящий момент, например, у заказчика в другом городе.

Применение многопоточного программирования оправдано только в одной ситуации: когда, разбросав независимые части одного вычисления по разным тредам, вы можете добиться видимого выигрыша по времени исполнения за счёт использования нескольких ядер центрального процессора. Отметим, что автор книги не встретил ни одной такой задачи за всю свою практику. Впрочем, даже в таких задачах в подавляющем большинстве случаев будет правильнее применять обычные процессы.

7.4. Разделяемые данные на диске

Избежать возникновения разделяемых данных в оперативной памяти обычно можно (а с определённых точек зрения — и нужно); но, как мы уже отмечали на стр. 172, данные на внешних запоминающих устройствах (т. е. на дисках) во многих случаях оказываются разделяемыми в силу природы поставленной задачи, и поделать здесь ничего нельзя. Прежде чем начать рассказ об имеющихся инструментах — которые, к сожалению, все как один кривые — позволим себе один совет: покуда вы *можете* избежать ситуации доступа к одному файлу нескольких программ или процессов, избегайте её. Средства, рассмотренные в этой главе, следует использовать лишь тогда, когда иного выхода нет.

7.4.1. Обзор имеющихся возможностей

Для начала нам придётся условиться о терминах. Операционная система предоставляет нам (т. е. программисту; или, с другой точки зрения, *процессу*) средства, с помощью которых мы можем известить другие процессы в системе, что данный файл у нас прямо сейчас находится в работе, так что доступ к нему может привести к нежелательным последствиям. Эта сущность по-английски называется *file lock*. В большинстве русскоязычных источников слово *lock* в этом контексте переводят русским словом «блокировка», но такой вариант неудачен: мы уже активно использовали термин «блокировка», оригиналом которого послужило слово *blocking*.

В повседневной работе программисты обычно вообще не переводят слово *lock* с английского, так что вы вполне можете услышать выражения вроде «залочить файл», «сбросить лок» и т. п. Конечно, это откровенный сленг, но использование сленга в некоторых случаях можно считать допустимым; вот только проблема в том, что слово «лок», будучи напечатанным в тексте, смотрится совершенно кошмарно, тут даже можно не сразу догадаться, о чём идёт речь — в отличие от ситуации, когда оно произносится, но не пишется. Поскольку нам предстоит активное использование соответствующего термина в тексте книги, придётся подыскать что-то более подходящее.

Не найдя ничего лучшего, мы воспользуемся для перевода термина *lock* словом «захват». Этот вариант тоже имеет свои недостатки, и мы это с готовностью признаём, но другого, более удачного перевода нам придумать не удалось. Если вам известен лучший вариант, свяжитесь с автором книги.

Отметим, что файловые захваты, имеющиеся в unix-системах, в большинстве своём никого ни к чему не обязывают: процесс, имеющий достаточно прав на чтение файла, может его читать, а имеющий права на запись — может файл модифицировать, и установленные другими процессами захваты

никак ему не помешают. По-английски такие «ни к чему не обязывающие» захваты называются *advisory locks*; чтобы не усугублять и без того плачевную ситуацию с терминами, мы не будем пытаться перевести в этом контексте слово *advisory* на русский язык, а любителям терминов вроде «консультационный захват» или «консультативная блокировка» посоветуем прекратить издеваться над родным языком. Поскольку файловые захваты практически всегда именно таковы, мы просто не будем это каждый раз уточнять.

Поскольку установление захвата на файл реально не создаёт препятствий к чтению или записи его содержимого, ясно, что для успешной работы все процессы, использующие данный файл, должны *добровольно* придерживаться установленного порядка работы. Само по себе это не так уж страшно — достаточно вспомнить, что мьютексы и семафоры тоже сами по себе ничему не препятствуют. Здесь важно другое: если какая-нибудь из программ, работающих с файлом, не будет ничего знать про захваты, то все наши усилия по обеспечению корректной работы пойдут прахом; между тем такую программу может — просто по незнанию — запустить сам пользователь, и с этим мы ничего не сделаем.

В литературе и Интернете вам могут встретиться упоминания *обязательных захватов* (*mandatory locks*) — таких, установка которых реально исключает выполнение другими процессами операций чтения и записи. Этот механизм так и не получил широкого распространения. В системах семейства BSD вообще отсутствует поддержка обязательных захватов; в ОС Linux они поддерживаются ядром, но для этого нужно для начала смонтировать файловую систему с параметром, указывающим на их применение, и, кроме того, установить «хитрые» права доступа к файлу, на котором захваты должны иметь «обязательный» характер (см. комментарий на стр. 103). Автор этих строк ни разу за всю свою практику не видел, чтобы файловые системы монтировались таким образом; сам механизм, введённый в ядро в 1996 году, изначально считался и продолжает считаться проблемным [16]. Предположив, что обязательные захваты в реальной жизни не применяются, мы, по-видимому, окажемся недалеки от истины; поэтому рассматривать этот вариант файловых захватов мы не станем.

Основных подходов к захвату доступа к файлу существует три, причём один из них вообще не требует поддержки со стороны операционной системы: он заключается в том, что рядом с захватываемым файлом создаётся ещё один файл, который показывает другим процессам, что в настоящее время с файлом лучше ничего не делать. Два других механизма, известные как *BSD locks* (системный вызов `flock`) и *POSIX locks* (функция `lockf`, работающая через системный вызов `fcntl`) реализованы в ядре операционной системы; коль скоро один процесс установил файловый захват, другим процессам ядро откажет в предоставлении захвата, конфликтующего с уже установленным.

Отметим одно неочевидное обстоятельство, которое следует учитывать в работе. Среди всех файловых систем несколько особое место занимают такие, которые физически находятся на другом компьютере (файловом сервере); доступ к ним производится через компьютерную сеть. **Файловые захваты, поддерживаемые на уровне операционной системы, не работают для файлов, находящихся на сетевых дисках**. Если написанные вами программы в своей работе полагаются на эти механизмы, нужно обязательно отразить в документации, что соответствующие файлы должны находиться на локальных файловых системах, в противном случае всё перестанет работать. Прежде чем вы это сделаете, учтите, что в реальной практике часто встречаются компьютеры, не имеющие никаких дисков, кроме сетевых; декларируя свою зависимость от файловых захватов, поддержанных операционной системой, вы тем самым заявляете, что пользователи таких компьютеров вообще не смогут применять ваши программы.

Наиболее правильным представляется гибкий подход, при котором программа умеет применять все три возможных механизма захватов, в том числе одновременно, а то, какие из них следует использовать в конкретной ситуации, указывает пользователь с помощью настроек.

7.4.2. Создание дополнительного файла

Пожалуй, самый простой способ сообщить другим «действующим лицам», что в настоящий момент с файлом работаете вы — это создать рядом другой файл (чаще всего пустой); факт существования этого дополнительного файла рассматривается как извещение о захвате основного файла. Например, если вам нужно работать с файлом, который называется `foobar.dat`, то вполне логично воспользоваться именем `foobar.dat.lock`, расположенным в той же директории, для его захвата. Для этого, прежде чем начинать работу с `foobar.dat`, нужно попытаться создать файл `foobar.dat.lock`, воспользовавшись флажком `O_EXCL` (см. §5.2.3), т. е. потребовать от системы именно что создать новый файл, а если это невозможно — сообщить об ошибке. Если вызов `open` вернёт `-1` и переменная `errno` примет значение `EEXIST` — значит, файл `foobar.dat.lock` уже кем-то создан и нужно некоторое время подождать — например, одну секунду. По окончании работы с `foobar.dat` следует удалить файл `foobar.dat.lock` с помощью вызова `unlink` (см. §5.2.4). Выглядеть это всё будет примерно так:

```
for(;;) {
    int lckd;
    lckd = open("foobar.dat.lock", O_WRONLY|O_CREAT|O_EXCL, 0666);
    if(lckd != -1) {
        close(lckd);
        break;  /* всё в порядке, мы установили захват */
    }
    if(errno != EEXIST) {
        /* Ошибка, НЕ связанная с тем, что кто-то сейчас работает
           с нашим файлом; например, у нас могло не хватить
           полномочий для создания foobar.dat.lock, или произошло
           что-то ещё.  Обрабатываем как обычную ошибку.  */
    }
    /* файл уже есть; следовательно, надо подождать */
    sleep(1);
}
/* работаем с foobar.dat */
/* ... */
unlink("foobar.dat.lock");    /* снимаем захват */
```

Здесь можно заметить аналогию с мьютексом: создание «захватного» файла[11] играет роль закрытия мьютекса, а его удаление — роль открытия мьютекса. Как и в случае с мьютексом, здесь крайне важна *атомарность* операций. Во всех современных системах семейства Unix операция «эксклюзивного» открытия файла атомарна, так что программа, построенная по приведённой схеме, будет работать корректно; но так было не всегда. Например, для всё тех же сетевых файловых систем поддержка флага `O_EXCL` появилась только начиная с протокола `NFSv3`; при её отсутствии ядро системы обычно эмулировало эффект от `O_EXCL` за два обращения к файловому серверу, так что никакой атомарности не обеспечивалось. В частности, ядра Linux начали полноценно (атомарно) поддерживать `O_EXCL` для сетевых файловых систем (естественно, при условии, что соответствующая поддержка есть в протоколе) только с версий `2.6.*`, то есть сравнительно недавно.

Для ситуаций, в которых атомарность открытия файла с флагом `O_EXCL` гарантировать невозможно, файловые захваты с использованием дополнительного файла всё ещё можно организовать, но уже по существенно более сложной схеме. Так, в man-странице по вызову open для Linux предлагается создать (обязательно на том же диске, где и предполагаемый «захватный» файл; проще всего делать это в одной и той же директории) временный файл с произвольным именем, таким, чтобы минимизировать вероятность совпадения выбранного имени с каким-то ещё; например, можно использовать в имени такого файла некую комбинацию из своего номера процесса и текущего времени. Этот файл не будет использоваться в качестве индикации захвата сам по себе; вместо этого при попытке захватить доступ мы потребуем от системы создать *жёсткую ссылку* на него с помощью вызова `link`; как раз эта ссылка будет играть роль индикатора действующего захвата.

Если вызов `link` пройдёт успешно — всё в порядке, мы создали жёсткую ссылку; но это ещё не конец истории. Поскольку обычно все эти пляски устраивают в расчёте на сетевые диски, придётся припомнить, что именно на сетевых дисках `link` может вернуть ошибку, но при этом всё равно создать требуемую ссылку: например, ядро нашей системы обратилось к файловому серверу, он создал жёсткую ссылку и послал нашему ядру ответ с сообщением об успехе, но ответ по каким-то причинам не дошёл, так что наше ядро сообщает нам об ошибке. Поэтому в случае, когда `link` вернул ошибку, нам придётся, не доверяя ему, проверить, не создалась ли ссылка. Для этого нужно применить к нашему временному файлу системный вызов `stat` и проверить, не стало ли количество жёстких ссылок на него равно двум; соответствующая информация будет содержаться в поле `st_nlink` структуры `struct stat`.

По окончании работы с разделяемым файлом нужно будет удалить жёсткую ссылку, тем самым сообщив другим программам, что мы сняли свой захват. Временный файл тоже можно удалить, а можно оставить, чтобы снова воспользоваться им через некоторое время; важно только не забыть удалить его перед завершением программы.

[11] Оригиналом для нашего словосочетания «захватный файл» послужил английский термин *lock file*; как видим, трудности с переводом продолжаются.

Подход с использованием «захватного» файла обладает несомненным достоинством — своей простотой. При этом подходе вообще не требуется никакой специальной поддержки со стороны операционной системы, и (с учётом всего вышесказанного) его можно применять даже на «заковыристых» сетевых файловых системах. С другой стороны, подход также не лишён и недостатков.

Прежде всего заметим, что программы, взаимоисключающие свой доступ к разделяемому файлу с использованием этого подхода, должны не только знать о существовании друг друга и о том, что используется именно подход с «захватным» файлом, но и знать, как этот «захватный» файл называется. Пока мы взаимодействуем только с другими экземплярами той же программы или только с программами, написанными тем же коллективом программистов, всё, скорее всего, будет в порядке — мы как-нибудь договоримся; но если будет нужно обеспечить работу с разделяемым файлом для программ от разных авторов или производителей, соглашение об имени «захватного» файла вполне может стать камнем преткновения. Между прочим, такие файлы даже не всегда создают в одной директории с файлом, доступ к которому захватывают; некоторые программы предпочитают использовать для этой цели системную директорию `/var/lock`.

Есть и ещё один недостаток «захватных» файлов, не менее очевидный: при некорректном завершении программы она может не успеть стереть ранее созданный «захватный» файл, так что остальные участники взаимодействия будут считать, что захват всё ещё установлен. Это тоже приходится учитывать, усложняя программы и делая их менее надёжными. Так, мы могли бы в вышеприведённом примере ограничить количество безуспешных попыток установления захвата, после чего считать, что тот, кто его установил, благополучно помер. Идея в большинстве случаев работающая, но явно не слишком удачная: а вдруг он на самом деле не помер, а просто слишком долго «тормозит»? Некоторые программы записывают в «захватный» файл свой идентификатор процесса, а при обнаружении чужого захвата пытаются тем или иным способом связаться с процессом, установившим захват, и если это не удаётся — считают, что процесса, установившего захват, уже нет. Конечно, такой вариант обычно работает лишь для экземпляров одной и той же программы. Наконец, можно честно известить пользователя о проблеме и попросить его «разрулить» ситуацию, сказав ему что-то вроде «если в действительности никакие другие программы не работают с таким-то файлом, удалите такой-то («захватный») файл». У системных администраторов и прочих профессионалов такие ситуации трудностей не вызывают, а вот конечный пользователь вполне может от таких «заявочек» впасть в депрессию.

7.4.3. Системный вызов `flock`

Так называемые *BSD locks* (файловые захваты в стиле BSD) предполагают, что программы, работающие с разделяемым файлом, оповещают друг друга о входе в критическую секцию и о выходе из неё через системный вызов `flock`:

```
int flock(int fd, int operation);
```

Файл у нас должен быть уже открыт, его дескриптор передаётся в вызов первым параметром. Второй параметр может принимать одно из трёх основных значений: `LOCK_SH` — установить «разделяемый» захват (*shared lock*), `LOCK_EX` — установить «эксклюзивный» захват (*exclusive lock*) и `LOCK_UN` — убрать ранее установленный захват. На одном и том же файле могут быть одновременно установлены несколько «разделяемых» захватов, то есть если один процесс установил такой захват, то это не помешает сделать то же самое другому процессу. «Эксклюзивный» захват может быть установлен только один, причём он не может существовать одновременно ни с другими эксклюзивными, ни с разделяемыми захватами.

Если требуемый захват установить нельзя из-за захвата, установленного другим процессом, по умолчанию вызов `flock` блокирует вызвавший процесс до тех пор, пока конкурирующий захват не будет снят. Это поведение можно изменить, добавив ко второму параметру — как обычно, через операцию побитового «или» — флаг `LOCK_NB` (*non-blocking lock*), то есть указать значение `LOCK_SH|LOCK_NB` или `LOCK_EX|LOCK_NB`. Совместно с `LOCK_UN` флаг использовать бессмысленно, эта операция и так никогда не блокируется.

Назначение двух разных видов захватов — эксклюзивного и разделяемого — становится понятно, если мы припомним, что читать разделяемые данные можно сразу многим «действующим лицам», тогда как наличие одной *модифицирующей* критической секции, то есть такой, в которой предполагается изменение разделяемых данных, требует исключить на время её работы как запись, так и чтение данных для всех остальных участников взаимодействия. В §7.2.4 мы обсуждали одну из классических задач синхронизации — проблему читателей и писателей, решение которой направлено на то, чтобы «читатели» (то есть немодифицирующие критические секции) друг другу не мешали, а на время работы одного «писателя» (модифицирующей критической секции) все другие критические секции были исключены. Можно сказать, что задачу читателей и писателей вызов `flock` решает сам, то есть если такая задача перед нами встанет в применении к файлу, то `flock` даст нам уже готовое решение.

Здесь есть, впрочем, одна шероховатость. Обсуждая проблему читателей и писателей, мы отметили, что если читателей достаточно много, писатель может никогда не дождаться момента, когда в критической секции не будет ни одного читателя; это, как мы уже говорили, называется ***голоданием***. Реализация вызова `flock` вполне могла бы решать эту проблему тоже — например, не выдавать никому новых разделяемых захватов, если есть хотя бы один процесс, заблокированный в ожидании эксклюзивного захвата; но, увы, авторы спецификации этим вопросом не озаботились, так что ничего подобного `flock` не делает. Продемонстрировать этот факт нам поможет простенькая демонстрационная программа:

```
#include <stdio.h>                              /* flock_starve.c */
#include <unistd.h>
#include <sys/types.h>
#include <sys/stat.h>
#include <sys/file.h>
#include <fcntl.h>

#ifdef SHARED
#define LOCK_MODE LOCK_SH
#define LOCK_NAME "shared"
#else
#define LOCK_MODE LOCK_EX
#define LOCK_NAME "exclusive"
#endif
#define LOCKED_SLEEP  1900000
#define UNLOCKED_SLEEP 100000

int main(int argc, char **argv)
{
    int fd, i;
    if(argc < 2) {
        fprintf(stderr, "No argument specified\n");
        return 1;
    }
    fd = open(argv[1], O_RDWR|O_CREAT, 0666);
    if(fd == -1) {
        perror(argv[1]);
        return 2;
    }
    for(i = 0; ; i++) {
        int res;
        printf("%d Getting %s lock... ", i, LOCK_NAME);
        fflush(stdout);
        res = flock(fd, LOCK_MODE);
        printf("got it\nNow sleeping\n");
        usleep(LOCKED_SLEEP);
        printf("%d Releasing %s lock... ", i, LOCK_NAME);
        fflush(stdout);
        res = flock(fd, LOCK_UN);
        printf("done\nNow sleeping\n");
        usleep(UNLOCKED_SLEEP);
    }
}
```

Для демонстрации эффекта ресурсного голодания нам потребуются два исполняемых файла, полученных из этого исходного текста: первый, откомпилированный с флагом `-D SHARED`, будет устанавливать разделяемый захват, второй мы откомпилируем без дополнительных флагов, так что захват он будет устанавливать эксклюзивный:

```
gcc -Wall -g -D SHARED flock_starve.c -o flock_sh
gcc -Wall -g flock_starve.c -o flock_ex
```

В обоих вариантах программа принимает аргументом командной строки имя файла, открывает его на чтение/запись с возможностью создания, после чего циклически устанавливает на нём захват соответствующего типа, спит 1,9 секунды, снимает захват, спит 0,1 секунды и начинает цикл сначала. Запустив в разных окнах несколько экземпляров `flock_sh`, вы можете убедиться, что они друг другу совершенно не мешают. Убрав все экземпляры `flock_sh`, кроме одного, и запустив один экземпляр `flock_ex`, вы увидите, как они работают «в противофазе». Но вот если после этого запустить ещё один экземпляр `flock_sh`, скорее всего — а точнее, с вероятностью 90% — писатель `flock_ex` окажется заблокирован навсегда, поскольку у двух ваших читателей отсутствует пересечение временны́х промежутков нахождения вне критической секции. Для трёх читателей вероятность «голодной смерти» писателя окажется уже больше 99%, то есть вариант, что вас «случайно пронесёт», можно всерьёз не рассматривать; ну а запустить четырёх читателей так, чтобы писатель иногда мог проникнуть в критическую секцию, вам, скорее всего, просто не удастся (хотя тут есть одна хитрость: если сначала запустить писателя, а всех читателей запускать во время его нахождения в критической секции, то все они станут работать синхронно; но это уже немножко жульничество, и оно, конечно же, не пройдёт, если сделать длительности периодов работы и сна непостоянными).

Проблему можно решить примерно так, как мы решили её в §7.2.4. Напомним, там мы ввели дополнительный мьютекс. При работе с файловыми захватами нам тоже никто не мешает завести дополнительный файл (скорее всего, пустой), на котором и читатели, и писатели будут устанавливать исключительный захват, но читатели будут это делать перед входом в критическую секцию, после чего тут же снимать захват, тогда как писатели будут этот захват устанавливать на всё время нахождения в критической секции.

Довольно нетривиально поведение вызова `flock`, если его применить к потоку ввода-вывода, на котором вы уже установили захват. В этом случае вызов может заменить эксклюзивный захват на разделяемый и наоборот, причём замена разделяемого захвата на эксклюзивный может оказаться невозможна прямо сейчас (если кто-то другой тоже обладает разделяемым захватом на том же файле), и вызов при этом заблокируется в ожидании, пока требуемая операция не окажется возможна. Учтите, что **смена типа захвата не атомарна**, и это прямо отражено в спецификации вызова: сначала снимается имеющийся захват, затем устанавливается новый, и между этими операциями может теоретически «вклиниться» другой процесс. Следовательно, нельзя — недопустимо! — производить смену типа захвата внутри критической секции, то есть до тех пор, пока не будут доведены до «логической точки» все вычисления, основанные на прочитанных данных, и не будут записаны все данные, которые должны быть записаны.

Теперь самое время обратить внимание на некоторые тонкости. Прежде всего подчеркнём, что **захват, установленный вызовом `flock`, связан с потоком ввода-вывода как объектом ядра операционной системы**. Напомним, что с одним и тем же объектом ядра может быть связано несколько файловых дескрипторов, причём как в разных процессах, так и в одном. При создании процесса с помощью `fork` новый процесс обладает всеми теми же дескрипторами потоков, но *объект потока ввода-вывода в ядре ОС при этом не копируется*; точно так же не копируются объекты ядра ни при выполнении вызовов `dup` и `dup2`, ни при дублировании дескриптора с помощью `fcntl` (командой `F_DUPFD`). С другой стороны, если ваш процесс откроет один файл дважды, в ядре появятся два независимых объекта, с каждым из которых может быть связан захват.

Эта особенность работы `flock` может проявиться довольно неожиданным образом. Представьте себе, что ваш процесс установил захват, после чего выполнил `fork`, а порождённый процесс затребовал снятие захвата. В этом случае захват будет снят с объекта потока ввода-вывода в ядре, то есть обладать захватом не будет ни порождённый процесс, ни родительский — ведь объект потока у них общий. С другой стороны, можно представить, что вы открыли некий файл, установили на нём захват, после чего вызвали какую-нибудь библиотечную функцию, которая открыла тот же самый файл и тоже попыталась установить на нём захват. Если этот захват несовместим с вашим исходным (то есть хотя бы один из них эксклюзивный), ваша библиотечная функция благополучно заблокируется, так что вы попадёте в хорошо известную нам тупиковую ситуацию (deadlock) *с самим собой*.

Установленные на потоке ввода-вывода захваты сохраняют своё действие при замене выполняемой программы с помощью функций `exec*`. Это тоже следует учитывать: в большинстве случаев программа, которую вы запускаете, не имеет ни малейшего понятия о наличии захвата на каком-то из полученных ею «в наследство» дескрипторов, так что ей не придёт в голову этот захват снять. Как следствие, захват сохранит своё действие до тех пор, пока запущенная программа не закроет этот поток или не завершится; при невнимательном отношении к работе это может привести к тому, что другие участники взаимодействия будут вынуждены ждать не пойми чего.

Ещё раз напомним, что **захваты, устанавливаемые с помощью `flock`, не работают на сетевых дисках**. С другой стороны, у захватов `flock` имеется одно несомненное преимущество перед рассмотренными в предыдущем параграфе «захватными файлами»: если ваша программа завершится аварийно, то все её дескрипторы автоматически закроются, и если с потоком, на котором она поставила захват, больше не связано дескрипторов (например, в порождённых `fork`'ом процессах), то захват исчезнет вместе с объектом потока ввода-вывода.

7.4.4. Файловые захваты POSIX

Файловые захваты, описанные в спецификации POSIX, как и рассмотренные в предыдущем параграфе захваты BSD, реализованы в ядре операционной системы; по своим свойствам эти два вида захватов достаточно сильно различаются. Прежде всего, захват POSIX устанавливается не на файл целиком, а на определённый его фрагмент.

Второй момент не столь очевиден. В отличие от захватов BSD, которые связаны с объектом потока ввода-вывода в ядре ОС, **захваты POSIX устанавливают отношение между файлом (дисковым объектом, а не объектом открытого потока) и конкретным процессом**. Так, если в вашем процессе есть несколько дескрипторов, связанных с одним и тем же файлом, в том числе если они открыты независимо друг от друга вызовом `open` (то есть связаны с разными объектами потоков ввода-вывода), то установку и снятие захвата можно производить через любой из них, это в любом случае будет один и тот же захват. При этом **захваты POSIX не наследуются порождённым процессом (!)**, то есть если вы установили захват на тот или иной файл, после чего выполнили `fork`, система будет считать, что ваш родительский процесс обладает захватом этого файла, тогда как порождённый — не обладает.

Для работы с захватами POSIX можно использовать два разных интерфейса: системный вызов `fcntl` (см. §5.2.3) с командами `F_SETLK`, `F_SETLKW` и `F_GETLK`, а также функцию `lockf`:

```
int lockf(int fd, int cmd, int len);
```

В большинстве систем, в том числе в Linux и FreeBSD, эта функция реализована через системный вызов `fcntl`; при этом возможности системного вызова оказываются *шире* — в частности, `fcntl` позволяет устанавливать как захваты для чтения (аналог «разделяемых» захватов из предыдущего параграфа), так и захваты для записи, тогда как `lockf` устанавливает только эксклюзивные захваты.

Первый параметр функции `lockf` — это открытый файловый дескриптор, связанный с нужным файлом; файл должен быть открыт на запись — либо в режиме `O_WRONLY`, либо в режиме `O_RDWR` (но не «только чтение», задаваемое константой `O_RDONLY`; это тоже отличает захваты POSIX от захватов BSD, где файл может быть открыт в произвольном режиме). Вторым параметром передаётся «команда», в роли которой может выступать одно из значений `F_LOCK` (установить захват; если прямо сейчас этого сделать нельзя, функция блокируется до тех пор, пока такая возможность не появится), `F_TLOCK` (*test and lock*: попытаться установить захват, но не блокироваться — немедленно вернуть управление), `F_UNLOCK` (снять захват) и `F_TEST` (не устанавливать никаких новых захватов, только проверить, возможно ли это). Третьим параметром функция получает длину захватываемого фрагмента файла, которая всегда отсчитывается от *текущей позиции* — той, с которой началась бы следующая операция чтения или записи. Можно в качестве этого параметра передать отрицательное число, тогда длина будет отсчитываться от текущей позиции *назад* — в направлении начала файла.

Функция возвращает `0` в случае успеха и `-1` в случае неудачи; в частности, при использовании команды `F_TEST` возвращается `0`, если указанный фрагмент файла можно захватить, и `-1`, если этому мешает захват, установленный кем-то ещё.

Работа через вызов `fcntl` выглядит несколько сложнее. Когда вторым параметром вызова указана одна из команд `F_SETLK`, `F_SETLKW` или `F_GETLK`, третьим параметром нужно передать адрес структуры типа `struct flock` (при том, что никакого отношения к функции `flock` всё это не имеет); выглядит это примерно так:

```
struct flock s1;
int res;
/* ... */
res = fcntl(fd, F_SETLKW, &s1);
```

Структура `struct flock` имеет следующие поля (в каком порядке они расположены — не уточняется): `l_type`, `l_whence`, `l_start`, `l_len` и `l_pid`. Поле `l_type` может принимать значения `F_RDLCK`, `F_WRLCK` или `F_UNLCK`, которые означают соответственно установку захвата на чтение (аналог «разделяемого» захвата), установку захвата на запись (аналог «эксклюзивного» захвата) и снятие захвата. Поле `l_start` задаёт позицию, с которой начинается нужный нам фрагмент файла. Число, находящееся в `l_start`, может отсчитываться от начала файла, от текущей позиции в файле или от конца файла в зависимости от того, какое значение будет помещено в `l_whence`; это может быть `SEEK_SET`, `SEEK_END` или `SEEK_CUR`, имеющие в точности такой же смысл, как для системного вызова `lseek` (см. §5.2.3). Поле `l_len` задаёт длину захватываемого фрагмента, и, как и для функции `lockf`, может быть отрицательным: в этом случае позиция «начала» рассматривается как конец фрагмента, а длина отсчитывается «влево» (в сторону начала файла). Кроме того, в поле `l_len` можно занести ноль, что будет означать захват от заданной точки начала до конца файла.

Поле `l_pid` используется только для команды `F_GETLK`, которая предназначена, чтобы узнать, имеется ли в настоящий момент на данном файле захват, конфликтующий с тем, который мы хотели бы установить. В него вызов записывает `pid` процесса, установившего этот захват. При использовании этой команды нужно предварительно заполнить все перечисленные поля структуры `struct flock`, описав захват, который мы хотели бы установить. Если такой захват может быть установлен, то вызов запишет значение `F_UNLCK` в поле `l_type` переданной ему структуры, а остальные поля оставит без изменений; если же в настоящий момент в системе имеется захват, установленный кем-то другим и конфликтующий с нашим, вызов заполнит поля `l_type` (в этом случае тут будет значение `F_RDLCK` или `F_WRLCK`), `l_whence`, `l_start`, `l_len` и `l_pid` в соответствии с тем, какой захват нам мешает и какой процесс его установил.

Команды `F_SETLK` и `F_SETLKW` предназначены для установки и снятия захвата; первая пытается выполнить запрошенную операцию, и если это не удалось — немедленно возвращает управление (естественно, вызов `fcntl` при этом вернёт `-1`); вторая же блокируется, если запрошенная операция не может быть выполнена прямо сейчас, то есть в системе имеется конфликтующий захват. Буква `W` в названии команды, как несложно догадаться, означает *wait*. Для обеих команд нужно заполнить поля `l_type`, `l_whence`, `l_start` и `l_len`.

Любопытно будет отметить, что снять захват можно с фрагмента, лишь частично совпадающего с тем, который ранее был захвачен. В этом случае остаток фрагмента продолжает быть захваченным, причём этих захваченных фрагментов может стать два: например, если вы захватите в файле `1000` байт, начиная с позиции `500`, а затем освободите `100` байт с позиции `700`, то у вас останется два захвата: с позиции `500` длиной `200` и с позиции `800` длиной `700`. Точно так же возможны слияния и разделения имеющихся захватов, если вы попытаетесь установить новый захват (другого типа) на фрагмент, который уже полностью или частично захвачен.

Захваты POSIX могут работать с сетевыми дисками, если используются достаточно новые версии протокола и ядра системы; это выгодно отличает их от захватов BSD, которые на сетевых дисках просто не работают, и это никак не зависит от версий.

7.4.5. Некоторые проблемы файловых захватов

Файловые захваты POSIX имеют одну очень странную особенность: если ваш процесс закрывает с помощью `close` любой из дескрипторов, связанных с некоторым файлом, то при этом автоматически снимаются все захваты, которыми данный процесс обладал в отношении данного файла. Понять, откуда взялась эта особенность, не так трудно: если процесс тем или иным способом завершается, все его захваты должны исчезнуть, но в отличие от захватов BSD, которые связаны с объектом потока ввода-вывода, захваты POSIX связаны с парой «процесс-файл»; по-видимому, создатели исходной реализации захватов POSIX решили не дублировать код ядра и «повесили» сброс всех захватов на закрытие потока, которое, как известно, происходит при завершении процесса автоматически. Проблема в том, что из-за этой особенности вы можете потерять все свои захваты совершенно неожиданно для себя — если какая-нибудь библиотека, функции которой вы вызываете, откроет какой-то из файлов, с которыми вы работаете в основной программе, и закроет его. Аналогичные проблемы возникнут, если в основной программе вы установите эксклюзивный захват (на запись), а некая подсистема вашей программы откроет тот же файл второй раз и попытается захватить его для чтения. В результате ваш эксклюзивный захват превратится в разделяемый. В некоторых случаях такие вещи довольно сложно предугадать, и особенно жёстко эта непредсказуемость проявляется в многопоточных программах; пожалуй, это ещё одна причина (хотя, надо признать, и не слишком серьёзная) отказаться от использования тредов.

Ещё одна проблема связана с обоими вариантами захватов, поддерживаемыми ядром (то есть как с захватами POSIX, так и с захватами BSD). Если некий файл доступен на запись только процессам определённого пользователя или группы, а на чтение — либо всем пользователям системы, либо, по крайней мере, более широкому их кругу, нежели на запись (например, записывать может один пользователь, а читать — некая группа), и при этом процессы, записывающие информацию в этот файл, используют взаимоисключение через файловые захваты, то (несколько неожиданно для нас) любой процесс, имеющий доступ на чтение, может нарушить работу всех процессов, осуществляющих запись и использующих для этого захваты — достаточно попросту установить разделяемый захват; больше того, доступа на чтение достаточно, чтобы установить даже эксклюзивный захват, если только это будет захват в стиле BSD.

Небезынтересен также вопрос о взаимоотношении различных типов файловых захватов. Как уже говорилось, функция `lockf` практически всегда реализована через `fcntl`, так что захваты через эти два интерфейса — это, собственно говоря, одни и те же захваты. Иное дело — вызов `flock`. В современных версиях ОС Linux захваты, устанавливаемые через `flock` и `fcntl`, друг друга вообще не видят и никак друг на друга не влияют; но в той же FreeBSD это не так, там ядро отслеживает оба типа захватов в связке. Пикантности ситуации добавляет ещё и то, что в некоторых системах функция `flock` реализована через `fcntl`, как и `lockf`, причём в ОС Linux достаточно старых версий дело тоже обстояло именно так.

Если рассматривать имеющуюся картину целиком, получится, что относительно использования функций `flock`, `lockf` и `fcntl` (в части файловых захватов) вообще невозможно указать никаких способов их использования, которые бы были заведомо корректны. Поэтому мы просто повторим рекомендацию, которую вы уже видели: избегайте появления разделяемого доступа к файлам, покуда это хоть как-то возможно.

Часть 8

Ядро системы: взгляд за кулисы

Читатель, скорее всего, уже представляет, зачем нужна операционная система и каковы её роль и место в происходящем. Мы активно использовали системные вызовы, чтобы обращаться к системе за услугами и управлять объектами ядра. В этой части книги мы попытаемся показать отдельные аспекты того, что происходит внутри ядра и обычно остаётся за кадром для всех, кроме программистов, создающих само ядро и его отдельные компоненты — драйверы устройств и другие «ядерные» модули.

Пытаться самостоятельно писать модули ядра мы не будем — это довольно специфическая область, для которой высок порог вхождения. Книги, посвящённые этому, вынужденно обрушивают на читателя поток разнообразных сведений, без которых невозможно обойтись при работе внутри ядра — чтобы в конце позволить написать явно искусственный и не имеющий никакой, даже потенциальной ценности модуль вроде убогого драйвера клавиатуры или какого-нибудь совершенно бесполезного виртуального устройства. Что самое неприятное здесь — это то, что все полученные сведения применимы обычно только к одной конкретной версии одного конкретного ядра; с выходом следующей версии того же ядра значительную часть полученных сведений приходится выбросить.

Программирование внутри ядра ОС можно назвать одним из самых увлекательных, но и самых сложных аспектов системного программирования. Если вы захотите этим заниматься, к вашим услугам масса специальной литературы и сайтов в Интернете; надо сказать, что здесь будет лучше воспользоваться англоязычными источниками, поскольку к моменту перевода таких текстов на русский язык они успевают изрядно устареть.

С другой стороны, в функционировании ядра системы остаётся целый ряд интересных аспектов, которые желательно представлять хотя бы на качественном уровне, то есть понимать в общих чертах, как всё это устроено. Этому мы и посвятим заключительную часть второго тома книги.

8.1. Основные принципы работы ОС

Прежде чем двигаться дальше, кратко напомним основы. Ядро операционной системы загружается в память раньше других программ, благодаря чему выполняется в *привилегированном режиме* центрального процессора; оно запускает пользовательские программы в виде *процессов*, которые работают в *ограниченном режиме*, в котором процессор «забывает» все машинные команды, способные повлиять на систему в целом. До передачи управления процессам ядро настраивает *обработчики* всех «особых» событий — аппаратных прерываний, исключений/ловушек (внутренних прерываний) и системных вызовов (программных прерываний) так, чтобы при возникновении любого из этих событий управление снова передавалось коду ядра, а поскольку режим процессора переключается с ограниченного обратно в привилегированный только при наступлении таких событий, всё это гарантирует, что только код ядра и будет выполняться на компьютере в привилегированном режиме ЦП. Ограниченный режим ЦП позволяет процессу только преобразовывать данные в отведённой ему памяти — и больше ничего; для всего остального, в том числе чтобы выдать какое-нибудь сообщение и даже чтобы просто завершить выполнение, процесс вынужден обращаться к ядру через механизм системных вызовов.

На всякий случай повторим ещё раз, что **никакой код никакого пользовательского процесса ни при каких условиях не может быть исполнен в привилегированном режиме процессора**. Часто в литературе можно встретить фразы вроде «процесс выполняется в режиме ядра», и мы тоже будем использовать это выражение; но пусть вас не обманывает такая терминология — на самом деле речь идёт о выполнении машинного кода *самого ядра*, а не процесса, просто этот код выполняется *в рамках процесса как единицы планирования* — то есть планировщик выделяет ему кванты времени так же, как и другим процессам, а сам этот код может «заснуть» (заблокироваться) в ожидании наступления какого-нибудь события. Кроме того, обычно при этом сохраняются настройки отображения виртуальных адресов в физические, если, конечно, компьютер вообще предусматривает виртуальную память (процессор оснащён MMU)[1]. Такое «исполнение в режиме ядра» происходит при обработке системного вызова, исключения/ловушки (внутреннего прерывания), а также непосредственно перед возобновлением работы процесса, ранее снятого с процессора из-за истечения кванта времени.

Ядро операционной системы берёт на себя две основные функции: *управление задачами* и *управление внешними устройствами*. Первое включает запуск и останов процессов, планирование времени центрального процессора, распределение оперативной памяти, организацию взаимодействия между процессами и разграничение их полномочий; второе сводится к *абстрагированию* от особенностей конкретных устройств и *координации* запросов к ним (см. §§5.1.2, 5.1.3).

Отметим, что ядро не обязано быть монолитной программой. В некоторых операционных системах ядро представляет собой набор взаимодействующих программ (архитектуры с микроядром и экзоядром); почти во всех современных системах в ядро во время работы можно добавлять дополнительные модули. Некоторые свои части ядро может оформить в виде процессов, работающих наравне с пользовательскими задачами в ограниченном режиме. Могут существовать и такие процессы, которые выполняются *только* в ядре, то есть не имеющие той части, которая должна выполняться в пользовательском режиме — в действительности это части ядра, работающие как единицы планирования с точки зрения планировщика; обычно они применяются в ядре для выполнения работы, требующей длительного времени.

8.1.1. Ядро как обработчик запросов

Интересно отметить, что, загрузившись в память после старта системы, **ядро ничего не делает по своей инициативе**. На первый взгляд это может показаться странным, но всё становится на свои места, если вспомнить, что компьютер в основном предназначен для решения проблем *конечных пользователей*, а эти проблемы решают прикладные программы, которые создаются и работают в виде пользовательских задач. При отсутствии в системе активных пользовательских задач ядру просто оказывается нечего делать; в этом случае оно может *остановить* центральный процессор в ожидании, пока не произойдёт что-нибудь интересное. В таких случаях говорят, что система перешла в ***состояние покоя*** (англ. *idle*). Отметим, что состояние покоя возможно и осмысленно, когда задачи в системе есть, но все они находятся в блокировке, то есть ждут наступления каких-то событий, чтобы после этого продолжить работу. Если же задачи в системе вообще исчезнут, ядру не останется ничего иного, кроме как прекратить работу — либо отправить компьютер на перезагрузку, либо вообще «повиснуть» за неимением лучших идей. Сообщение об этом — лаконичное «`System halted`» — часто снится сисадминам в ночных кошмарах.

Пользовательским задачам для работы часто требуется что-то такое, что может сделать только ядро (чаще всего это ввод-вывод информации), и они через системные вызовы обращаются к ядру за услугами. В этом случае ядро, естественно, начинает активно работать, но, заметим, оно при этом работает не по своей инициативе, а в ответ на запросы пользовательских задач. Обращаясь к контроллерам устройств, ядро может получать ответы от них не сразу, а по мере готовности — через механизм аппаратных прерываний; через них же ядро узнаёт о наступлении внешнего события, требующего каких-то действий — например, о приходе пакета данных по локальной сети, о нажатии клавиши на клавиатуре и т. п., и в этом случае ядро тоже включается в работу, но инициатива и здесь, как видим, исходит не от него.

В первом томе в части, посвящённой программированию на языке ассемблера, мы уже обсуждали термин «прерывание» и связанную с ним путаницу (см. т. 1, §3.6.3). Напомним, что по инициативе внешних устройств у нас могут

[1] Виртуальную память и MMU мы упоминали в первом томе (§3.1.2), а также в части 5 (§5.3.3); подробный разговор о виртуальной памяти у нас впереди.

происходить собственно ***прерывания***; в результате ошибочных действий выполняющейся программы процессор может генерировать ***исключения***, называемые также ***внутренними прерываниями***. В обоих случаях режим центрального процессора становится привилегированным, а управление передаётся на процедуру-обработчик, адрес которой настроен заранее. Сама такая процедура, естественно, является частью ядра операционной системы. Частным случаем внутреннего прерывания можно считать ***программное прерывание***, которое задача инициирует не по ошибке, а намеренно, чтобы отдать управление ядру для выполнения ***системного вызова***. Мы также отмечали, что терминология с тремя разными видами прерываний не совсем удачна, поскольку реально что-то *прерывают* только «настоящие» прерывания, они же ***аппаратные*** или ***внешние***, а остальные — внутренние и программные — на самом деле ничего и никого не прерывают; при описании архитектур, отличных от Intel, термин «прерывание» обычно используется только для обозначения аппаратных прерываний, внутренние прерывания называют исключениями, а между системным вызовом и его реализацией не делают различия, то есть вместо фразы «программа инициировала программное прерывание для выполнения системного вызова» говорят просто «программа выполнила системный вызов».

Вне зависимости от того, какой из стилей терминологии использовать, ядро операционной системы после окончания загрузки может получить управление только по запросу внешнего устройства, по запросу пользовательской задачи или в результате ошибочных действий пользовательской задачи. Всё это мы тоже уже обсуждали, но не лишним будет напомнить ещё раз, что обработка аппаратных прерываний на уровне ядра существенно отличается от обработки событий, ставших результатом действий пользовательской задачи — неважно, случайных или намеренных. При получении управления по инициативе пользовательской задачи ядро выполняет свой обработчик в контексте этой задачи, то есть MMU отображает виртуальные адреса в физические по правилам задачи, только дополнительно становятся доступны адреса самого ядра, а планировщик продолжает рассматривать такую задачу как единицу планирования (это как раз и называют выполнением задачи в режиме ядра); задача может заблокироваться, заснуть, отдать управление другим частям ядра, получить управление обратно и т. д.; всё это никак не мешает работать остальным частям ядра.

С аппаратными прерываниями дела обстоят сложнее. Исключения и системные вызовы — это своего рода продолжение выполнения задачи, они возникают в ходе обычного выполнения машинных команд; аппаратные прерывания, напротив, возникают в произвольные моменты времени, в том числе и в самые неподходящие. Уже на аппаратном уровне процессору приходится это учитывать: если запрос на прерывание пришёл в середине выполнения очередной машинной инструкции, то эту инструкцию приходится либо доводить до конца, либо откатывать к началу. Когда обработчик аппаратного прерывания получает управление, никакого контекста пользовательской задачи у него нет, ведь прерывание может возникнуть в том числе и тогда, когда никакая пользовательская задача не выполнялась — либо когда работало ядро, либо когда система вообще находилась в состоянии покоя.

Больше того, ничто не исключает возможности возникновения следующего прерывания сразу после предыдущего — например, запрос прерывания может прийти от другого устройства. Как правило, обработчик прерывания к этому не готов, поскольку он мог ещё не успеть выяснить, что́ же произошло и стало причиной прерывания, и при возникновении нового прерывания эта информация окажется потеряна; поэтому центральный процессор при получении запроса прерывания *блокирует* другие прерывания. В этом режиме поступающие запросы прерываний накапливаются, но никакого эффекта не имеют; процессор как бы *откладывает* их обработку до тех пор, пока обработчик текущего прерывания не снимет блокировку. Надо отметить, что возможности хранения отложенных запросов у процессора весьма ограниченны: как правило, он может «помнить» не более чем об одном отложенном прерывании от каждой из нескольких групп внешних устройств. Если в режиме блокировки прерываний система пробудет достаточно долго, это может привести к потере запросов прерываний и в конечном итоге к сбоям в работе системы. Поэтому обработчик обязан как можно скорее достичь такого состояния, когда вся информация о поступившем прерывании будет надлежащим образом сохранена, и разрешить (разблокировать) прерывания.

Из всех случаев обработки аппаратных прерываний можно выделить важный вариант, при котором обработчик успевает сделать всё необходимое за столь короткий период времени, что в разрешении аппаратных прерываний необходимости не возникает: управление практически сразу возвращается той программе, выполнение которой было прервано, при этом блокировка прерываний автоматически снимается. Самый простой пример такой ситуации — это прерывание таймера, поступившее, когда по тем или иным причинам вызывать планировщик не нужно: если нет других задач, готовых к выполнению, или если квант времени текущей задачи ещё не истёк. Обычно в ядре для таких ситуаций предусматриваются специальные флажки и счётчики, проверить и изменить которые можно за несколько десятков тактов центрального процессора, не сохраняя нигде большинство регистров; если обработчик прерывания таймера видит, что больше от него ничего не требуется, он сразу же завершается.

Если аппаратное прерывание требует от ядра более сложных действий, картина резко усложняется. Прежде чем разрешить другие прерывания, обработчик должен позаботиться о том, чтобы неожиданный вызов другого (или, возможно, того же самого) обработчика ничего не испортил. Кроме того, обработчик аппаратного прерывания, работая вне контекста какой-либо пользовательской задачи, не является единицей планирования, что накладывает серьёзные ограничения на ассортимент используемых средств. В частности, обработчик прерывания не может «заснуть» (заблокироваться), поскольку разблокировать его будет некому.

Если бы центральный процессор в системе был один, обработчику прерывания было бы вообще не нужно ничего ждать, да и *нельзя*, ведь при запрещённых прерываниях ничего «внешнего» по отношению к активной программе, собственно говоря, произойти не может; но в современных условиях процессоров в системе обычно больше одного, ведь с программной точки зрения «многоядерные»[2] процессоры представляют собой несколько независимых центральных процессоров, выполненных в виде одной микросхемы. В таких условиях неизбежно возникают *критические секции*[3] по данным, к которым могут обратиться компоненты ядра, выполняемые одновременно на разных процессорах, и эти критические секции требуют взаимоисключения; но поскольку блокировка здесь невозможна (ведь, напомним ещё раз, обработчик прерывания не является единицей планирования, в отличие от процесса), приходится применять активное ожидание, вхолостую расходуя процессорное время.

Ядро операционной системы обычно реализует целый ассортимент средств взаимоисключения для разных случаев; в частности, в ядре Linux их больше десятка. Программисту приходится тщательно следить, чтобы применяемый механизм соответствовал контексту; так, средства взимоисключения, предназначенные для обработчиков аппаратных прерываний, нельзя применять в контексте процесса, и наоборот.

Итогом обработки аппаратного прерывания может стать констатация необходимости смены текущей задачи, например, в силу истечения отведённого ей кванта времени, либо если событие, о котором сигнализирует полученное прерывание, становится причиной разблокировки задачи, имеющей приоритет более высокий, чем текущая. Может случиться и так, что система находилась в состоянии покоя, но происшедшее событие разблокировало какую-то задачу из ранее заблокированных. В этом случае всё сравнительно просто: ядро сохраняет контекст текущей задачи, если таковая была, затем восстанавливает контекст задачи, которую нужно поставить на выполнение, и дальнейшая подготовка этой задачи к выполнению происходит уже в её контексте.

В сравнительно редких случаях при получении прерывания оказывается нужно выполнить какие-то действия, которые невозможно отнести к выполнению той или иной задачи, но при этом они слишком сложны, чтобы их можно было выполнить вне контекста задачи. Для таких случаев ядра некоторых систем предусматривают специальные ***псевдопроцессы***, порождаемые ядром и выполняющиеся исключительно в режиме ядра, но имеющие

[2]Здесь имеется очередная терминологическая сложность. Студенты, не понявшие в должной мере, о чём идёт речь, иногда путаются со словом «ядро», которое используется в русском языке для обозначения, с одной стороны, основной программы операционной системы, а с другой — для отдельных независимых процессоров в составе одной «многоядерной» микросхемы. В английском такой путаницы нет: ядро операционной системы называется *kernel*, тогда как часть микросхемы называется *core*. Вообще-то слово *core* следовало бы переводить как «сердечник», а не «ядро», но словосочетание «многосердечниковый процессор», что называется, не звучит.

[3]Если словосочетание «критическая секция» вызывает какие-то сложности с пониманием, самое время вернуться к предыдущей части и перечитать §7.1.2.

свой контекст в качестве единиц планирования. Обработчик прерывания может поручить выполнение оставшейся части работы такому псевдопроцессу.

В частности, авторы ядра Linux используют для разных стадий реакции на прерывание термины ***верхняя половина*** и ***нижняя половина*** (англ. *top half*, *bottom half*). Под верхней половиной понимается обработчик аппаратного прерывания как таковой, то есть та часть кода ядра, на которую процессор передаёт управление при наступлении прерывания и которая выполняется при заблокированных аппаратных прерываниях. Очевидно, она должна завершиться как можно скорее; основную работу следует перепоручить нижней половине.

Для организации нижней половины обработки прерывания Linux предусматривает две принципиально различные сущности: ***тасклеты*** (*tasklets*) и «рабочие очереди» (*workqueues*). Тасклет представляет собой сравнительно короткую функцию, выполнение которой предполагается уже без блокировки аппаратных прерываний, то есть не столь жёстко критично по времени; с другой стороны, тасклет должен выполниться «за один приём», поскольку единицей планирования он не является и «заснуть», как следствие, не может. Обработчик аппаратного прерывания, решив возложить часть работы на тасклет, сообщает об этом планировщику, после чего завершается; планировщик выполняет тасклет всегда на том же процессоре, на котором его запланировали, так что он точно не может начать выполняться раньше, чем завершится запланировавший его обработчик. Выполнение тасклета может начаться немедленно, как только обработчик прерывания вернёт управление, если процессор, на котором это происходит, свободен; если же на процессоре выполняется пользовательская задача, то тасклет будет выполнен после прихода следующего прерывания таймера, временно вытеснив задачу с процессора.

Рабочая очередь отличается от тасклета тем, что при выполнении «работы», поставленной в очередь, сама эта «работа» обладает всеми преимуществами единицы планирования, то есть ей, как обычному процессу, выделяются кванты времени, она может «заснуть» — заблокироваться в ожидании события, и т. д. Рабочей очереди, в отличие от тасклета, можно поручить сколь угодно длинное задание; с другой стороны, никто не гарантирует (опять же в отличие от тасклета), что выполнение задания, поставленного в очередь, начнётся не позже какого-то момента. С точки зрения планировщика рабочая очередь — это почти обыкновенный процесс, у него даже есть свой `pid`; в выдаче команды `ps ax` рабочие очереди хорошо заметны — их названия выдаются заключёнными в квадратные скобки, примерно так:

```
  8 ?        S<     0:00 [cpuset]
  9 ?        S<     0:00 [khelper]
 10 ?        S      0:00 [kdevtmpfs]
```

Как и в случае с тасклетом, обработчик прерывания («верхняя половина») может потребовать выполнить очередную «работу» в рамках заданной рабочей очереди, после чего с чистой совестью вернуть управление; всё, что осталось, сделает за него «работа».

Вернёмся к обсуждению пассивной роли ядра в системе; как мы отметили выше, если в системе нет пользовательских задач, ядру будет нечего делать, но если оно не будет ничего делать, то откуда возьмутся задачи? Выйти из порочного круга позволяет введение специфической пользовательской задачи, которая выполняется в течение всего времени работы системы, начиная от момента её загрузки и заканчивая моментом останова или перезагрузки. В ОС Unix такая задача называется ***процессом init***, который всегда имеет номер `1`. По сути это обыкновенная программа, написанная, как правило, на языке Си и не имеющая принципиальных отличий от других программ. Единственная существенная особенность `init` состоит в том, что ему приходится выполнять обязанности предка для тех процессов, «настоящие» предки которых уже завершились; мы упоминали эту его особую роль при обсуждении процессов-зомби и системных вызовов семейства `wait` в §5.3.7.

Настройки самого ядра или (чаще) программы-загрузчика определяют, какую программу ядро должно загрузить и запустить в качестве `init`; последним действием в ходе загрузки операционной системы ядро запускает `init`, после чего считает загрузку завершённой и более ничего не делает иначе как в ответ на запросы. Первоначально такие запросы поступают от самого `init`: он должен прочитать свои конфигурационные файлы и решить, какие ещё программы запустить. В итоге система оказывается «населена» пользовательскими задачами, которые запускают друг друга.

Следует отметить, что завершение процесса `init` означает завершение работы системы; ядро при этом считает, что от него больше ничего не требуется, и в зависимости от настроек либо останавливается, либо (реже) отправляет компьютер на перезагрузку.

8.1.2. Загрузка и жизненный цикл ОС UNIX

Сразу после включения компьютера центральному процессору нужно начать что-то делать, в противном случае ничего никогда не произойдёт и компьютер останется для нас бесполезен. Как мы знаем, процессор умеет только одно — *выполнять программы*; но, с другой стороны, процессор не знает и не может знать ничего о внешних устройствах — для работы с ними нужны ***драйверы***, которые тоже представляют собой программы, причём довольно сложные.

Всё, что процессор умеет сам — это общаться через шину с оперативной памятью, но мы знаем, что оперативная память не сохраняет информацию при выключении питания. Состояние оперативной памяти сразу после включения питания *не определено*, то есть в каждой ячейке памяти может оказаться совершенно произвольное число. Очевидно, что никакой программы там быть не может — ей неоткуда взяться.

Итак, процессор получил электропитание и «ожил», оперативная память содержит заведомый мусор, про внешние устройства процессор ничего не знает; что́ же, в таком случае, ему выполнять?

Специально на этот случай в конструкции компьютера предусмотрено так называемое ***постоянное запоминающее устройство*** (ПЗУ; англ. *ROM, read-only memory*). Микросхемы, составляющие ПЗУ, с точки зрения шины и центрального процессора «выглядят» в точности как обыкновенная оперативная память, состоящая из ячеек, за исключением того, что они поддерживают только операцию чтения, а попытки записи в свои ячейки молча игнорируют. По своей конструкции ПЗУ резко отличается от оперативной памяти (ОЗУ): каждая из ячеек постоянной памяти содержит одно фиксированное, раз и навсегда определённое значение, которое и выдаёт в шину в ответ на запрос чтения.

В разное время для создания ячеек ПЗУ использовались разные технологии. Самая старая и «кондовая» из них состояла в том, что каждая ячейка выполнялась в виде восьми микроскопических перемычек; замкнутая перемычка обозначала единицу в соответствующем разряде ячейки. При записи информации в микросхему с помощью специального устройства, называемого *программатором*, некоторые из перемычек попросту выжигались — на них подавалось высокое напряжение, под воздействием которого тонкий медный проводок плавился и переставал существовать; выжженные перемычки, как можно догадаться, обозначали ноль. Микросхемы такого типа назывались ППЗУ — программируемое постоянное запоминающее устройство (англ. *PROM — programmable read-only memory*). Очевидно, что такая микросхема была одноразовой: возможность изменить значения, записанные в её ячейки, физически отсутствовала, при любой ошибке микросхему приходилось выбрасывать.

Позже появилась более «аккуратная», хотя и дорогая технология, в которой вместо перемычек использовались специальным образом оформленные полевые транзисторы. Состояние такого транзистора исходно соответствовало «единице», но, подав на него достаточно сильное напряжение, можно было перевести его в состояние «ноль» — практически так же, как и с микросхемами типа PROM, только напряжение требовалось более скромное; в этом состоянии транзистор оставался вне зависимости от наличия или отсутствия питания. Все транзисторы такой микросхемы можно было вернуть в исходное состояние, подвергнув её облучению сильным источником ультрафиолета. Такие микросхемы получили название СППЗУ — стираемое программируемое постоянное запоминающее устройство (англ. *EPROM — erasable programmable read-only memory*).

Технология, используемая в наше время, ультрафиолетовой лампы для стирания уже не требует; соответствующие микросхемы по-английски называются *EEPROM — electrically erasable programmable read-only memory*; изредка можно встретить также русскую аббревиатуру ЭСППЗУ. Программатор такой памяти может быть встроен в общую схему компьютера или другого устройства, использующего EEPROM, что позволяет изменить содержимое ПЗУ, не имея для этого специальных устройств — например, в домашних условиях сменить версию программы, «зашитой» в ПЗУ.

Программа, которую процессор должен начать выполнять после включения питания, размещается в микросхемах ПЗУ; именно эта программа получает управление самой первой — как только процессор начнёт работу. Адрес её начала (точки входа) для конкретного типа процессора всегда один и тот же; например, i386 начинает выполнять инструкции с адреса $\texttt{FFFFFFF0}_{16}$, причём процессор стартует в так называемом «реальном режиме» (*real mode*), совместимом с предыдущими процессорами линейки. В этом режиме процессор выполняет команды точно так же, как это делали его 16-битные «предки», отсутствует защита памяти и деление команд на привилегированные и

непривилегированные. Переключение процессора в «защищённый» режим, в котором работают современные операционные системы, производится позже, это может сделать уже ядро операционной системы — либо (как это происходит при использовании UEFI) некая «продвинутая» часть загрузчика.

Программа, зашитая в ПЗУ для выполнения при старте процессора, в IBM-совместимых компьютерах, выпущенных до недавнего времени, называлась BIOS (*basic input-output system*), но в последние годы её сменили разнообразные версии UEFI (*unified extensible firmware interface*). На самом деле BIOS и UEFI — это даже не сами программы, зашитые в ПЗУ, а определённые наборы соглашений о том, как эти программы должны работать и как они производят дальнейшую загрузку компьютера. После выполнения некоторых действий по проверке оборудования программа, находящаяся в ПЗУ, определяет загрузочное устройство (загрузочный диск), а дальнейшие действия зависят от её типа: BIOS считывает в память первый сектор этого диска, называемый также ***загрузочным сектором***, и передаёт управление машинному коду, прочитанному из загрузочного сектора, тогда как UEFI отыскивает на диске небольшой раздел специального вида, находит на этом разделе исполняемый файл, загружает его в память и, опять-таки, передаёт управление прочитанному коду. Впрочем, если такого раздела не нашлось, UEFI обычно может произвести загрузку так же, как это делали BIOSы.

Поскольку размер загрузочного сектора сравнительно невелик (512 байт, причём начальный сектор может содержать ещё данные помимо загрузочного кода), программа, записанная в загрузочный сектор, не может выполнить никаких сложных действий, она для этого слишком коротка. Поэтому её роль заключается в загрузке в память более сложной программы, записанной на диске в специальных областях; эти области различны в разных операционных системах и для разных версий загрузочных программ. Новая программа называется ***загрузчиком операционной системы*** и может быть уже сравнительно сложной. Так, некоторые загрузчики имеют собственную командную строку, позволяющую выбрать, какой раздел считать корневым, из какого файла загружать ядро и т. п., при этом возможен даже просмотр каталогов на дисках. Классический загрузчик ОС Linux (LILO) не столь гибок в возможностях: он обладает способностью загружать альтернативные операционные системы, выбирать одно из предопределённых ядер для загрузки и передавать ядру параметры, но формат файловой системы не понимает и просматривать диски не позволяет. Ядро он загружает из фиксированного набора физических секторов диска. Какой конкретно загрузчик будет использоваться — зависит от дистрибутива; LILO в последние годы встречается довольно редко.

UEFI в этом плане отличается от BIOS тем, что сразу же читает с диска код загрузчика (тот самый исполняемый файл со специального раздела диска), минуя стадию загрузочного сектора. Дальнейшая загрузка системы уже никак не зависит от того, каким из двух способов загрузчик попал в память.

Загрузчик загружает в память выбранное ядро и передает управление его коду. Ядро инициализирует свои подсистемы, в том числе драйверы устройств, что включает, естественно, проверку наличия в системе соответствующего оборудования; при отсутствии нужного устройства драйвер обычно отказывается инициализироваться, после чего ядро удаляет его из памяти. Затем ядро монтирует файловую систему корневого дискового устройства в качестве корневого каталога системы. Обычно корневой каталог монтируется в режиме «только для чтения». После того как ядро готово к работе и смонтировало корневое устройство, оно формирует процесс с номером `0`. Этот процесс существует только на этапе загрузки, и единственная его роль — создать с помощью `fork` процесс с номером `1`. После этого нулевой процесс прекращает существование, так что в системе с этого момента есть только процессы, созданные с помощью вызова `fork`.

Процесс номер `1` выполняет вызов `execve`, чтобы загрузить в память программу `init`. Обычно это файл `/sbin/init`; ясно, что он должен находиться на корневом дисковом устройстве. Это, как мы уже говорили, обычная программа, написанная на Си или (теоретически) на любом другом компилируемом языке программирования. Как правило, загрузчик позволяет передать ядру специальный параметр, указывающий, какую программу следует загрузить вместо `init`; например, так можно запустить интерпретатор командной строки для выполнения действий по обслуживанию системы.

Процесс `init` работает в течение всего времени работы ОС; его завершение влечёт останов системы. Именно процесс `init` выполняет проверку дисков, перемонтирует корневой раздел в режим «чтение/запись» и монтирует остальные файловые системы. Затем процесс `init` должен инициализировать подсистемы ОС, например, сконфигурировать интерфейсы работы с локальной сетью, запустить системные процессы-демоны и, наконец, запустить на имеющихся терминальных линиях программы `getty`, отвечающие за запрос входного имени и пароля пользователей. Программа `getty` создаёт сеанс, связанный с её терминалом, и после успешной аутентификации пользователя запускает с помощью `exec` интерпретатор командной строки. Когда процесс интерпретатора завершается, программа `init` снова запускает `getty` на освободившейся терминальной линии.

Функциональность программы `init`, как видим, оказывается достаточно сложной. Чтобы немного «разгрузить» её, выполнение большинства действий, связанных с инициализацией системы, обычно возлагается на ***стартовые скрипты*** (англ. *startup scripts*); в зависимости от системы это может быть файл `/etc/rc`, `/etc/rc.d/rc` и т. п. Программе `init` тогда достаточно указать через конфигурационный файл или в качестве параметра компиляции, где находится соответствующий скрипт. В классическом варианте скрипты обычно писали на языке стандартного командного интерпретатора (Bourne Shell).

Процедура корректного останова системы для перезагрузки или выключения компьютера также возлагается на `init`, который запускает для этого специально предназначенный скрипт. В этом скрипте содержатся команды по уничтожению работающих процессов (сначала с помощью сигнала `SIGTERM`, затем, после паузы, — сигналом `SIGKILL`), размонтирования всех файловых систем, кроме корневой (её размонтировать невозможно), перевод корневой системы в режим «только чтение», синхронизация корневой файловой системы, т. е. запись недозаписанных данных из буферов. После этого выполняется собственно останов — прекращение работы ядра, которое можно совместить с отправкой компьютера на перезагрузку или с выключением питания.

8.1.3. Эмуляция физического компьютера

Как мы знаем, при попытке процесса выполнить некорректную инструкцию, в том числе привилегированную, возникает исключение («внутреннее прерывание»), в результате которого управление отдаётся ядру системы. Это позволяет сымитировать действия физической машины таким образом, чтобы у программы, работающей в рамках пользовательского процесса, «создалось впечатление», что эта программа работает в привилегированном режиме на машине, на которой никого, кроме неё, нет. Действия, которые на физической машине выполнял бы процессор в ответ на привилегированную команду, в режиме эмуляции производят обработчики прерываний по некорректной инструкции и нарушению защиты памяти.

В режиме такой эмуляции можно запустить в виде пользовательского процесса ядро другой операционной системы или даже второй экземпляр той же самой. Впервые такая эмуляция была реализована на мейнфреймах IBM/360 операционной системой CP/CMS в конце 1960-х годов. Под управлением этой системы можно было запустить несколько операционных систем OS/360, причём каждая из них была уверена, что весь мейнфрейм находится в её полном распоряжении. Под CP/CMS можно было даже загрузить в режиме эмуляции её саму. Бóльшую известность получила вышедшая в 1972 году система VM/370; в литературе часто именно ей приписывают первенство в этом классе систем.

Подобные системы обычно называют ***гипервизорами***. Гипервизор может сам играть роль операционной системы, то есть загружаться на компьютере первым, а затем запускать «настоящие» операционные системы в качестве своих «задач»; кроме того, гипервизор может быть выполнен как подсистема обычной операционной системы — в этом случае такая система (работающая на физическом компьютере) называется *хостовой*[4], а системы, запущенные в виртуальных машинах под управлением гипервизора — *гостевыми*. Среди используемых в наши дни можно назвать «отдельно стоящий» гипервизор Xen[5], а также Kernel Virtual Machine (KVM) для Linux и *bhyve*[6] для FreeBSD, которые позволяют этим системам выполнять роль хостовых.

Технологию виртуализации, при которой гипервизор перехватывает исключения, возникающие при попытке исполнить привилегированные коман-

[4] С английским термином *host* мы уже встречались при обсуждении компьютерных сетей, см. сноску на стр. 148; напомним, что это слово переводится как *хозяин, принимающий гостей*.

[5] Как ни странно, это название произносится примерно как *зен*.

[6] Читается примерно как *бихáйв*.

ды, и эмулирует их, так и называют — *trap and emulate*. При всей очевидности этого подхода отнюдь не любой процессор позволяет эту технологию использовать, и i386, который мы изучали во втором томе, как раз из тех, для которых такой вариант не годится. В частности, некоторые команды i386 не входят в число привилегированных, так что их выполнение никаких исключений не вызывает, но работают они при этом по-разному в зависимости от установленного режима процессора. Чаще всего в качестве такого примера приводят команду `POPF`, которая извлекает из стека значение и записывает его в регистр флагов; её обычно применяют в паре с командой `PUSHF`, которая, наоборот, помещает значение регистра флагов в стек. Проблема в том, что некоторые флаги (но не все) считаются привилегированными; например, флаг `IF` (*interrupt flag*) указывает, разрешены ли в данный момент аппаратные прерывания, и изменять этот флаг в ограниченном режиме, естественно, нельзя. Команда `POPF`, будучи исполнена в привилегированном режиме, устанавливает *все* имеющиеся флаги в соответствии с извлечённым из стека значением, а в ограниченном режиме затрагивает только флаги, изменение которых допустимо для пользовательской задачи, «втихую» игнорируя при этом привилегированные флаги. Кроме того, анализируя значения в сегментных регистрах, выполняющаяся на i386 программа может узнать, в каком кольце защиты она работает, то есть, грубо говоря, ядро «гостевой» операционной системы имеет возможность понять, что оно в настоящий момент работает не на настоящей машине, а на виртуальной.

Подход с эмуляцией привилегированных команд осложняется ещё и тем, что работа процессора может напрямую зависеть от структур данных, находящихся в обычной памяти. В основном это касается MMU, той части процессора, которая отвечает за преобразование виртуальных адресов в физические; немного позже мы увидим, что в этих преобразованиях участвует информация, которую операционная система должна сформировать в обычной оперативной памяти — так называемые страничные таблицы и дескрипторы сегментов. Изменение содержания этих страниц при работе на настоящем процессоре должно немедленно повлиять на то, к каким физическим адресам будет обращаться процессор; но когда «гостевая» система что-то записывает в свои таблицы, выглядит это как обычное обращение к памяти, то есть привилегированным действием не является. Для отслеживания таких обращений «хостовая» система вынуждена прибегать к хитростям — например, частично изменять машинный код «гостевого» ядра. Подробности о некоторых таких методах можно найти в статье [17].

Виртуализация может быть поддержана центральным процессором на аппаратном уровне; так, в 2006 году почти одновременно Intel и AMD включили в свои очередные процессоры линейки x86 расширения Intel VT-x и AMD-V, специально предназначенные для организации виртуальных машин.

Существуют и другие способы одновременного запуска нескольких систем на одной машине. Наиболее универсальным можно считать подход, предполагающий *интерпретацию машинного кода*, когда программа-эмулятор фактически выполняет (эмулирует) работу центрального процессора. Несомненным достоинством такого варианта является возможность эмулировать любой процессор на любой машине и отсутствие необходимости поддержки со стороны операционной системы; совершенно очевиден и фундаментальный недостаток такой схемы — чрезвычайно низкая эффективность: на обработку *одной* машинной команды эмулируемой программы уходят *десятки* реальных машинных команд. Несмотря на это, именно такой режим эмуляции находит применение, во-первых, для запуска программ, написанных для очень старых компьютеров — IBM-PC-совместимых машин эпохи MS-DOS, игровых компьютеров восьмидесятых годов вроде Atari, Commodore-64 и т. п.; современные машины превосходят их по быстродействию в сотни раз, так что потери на эмуляцию перестают быть серьёзной проблемой. Так работает, например, широко известный эмулятор DOSBox.

Встречаются и промежуточные варианты. Так, известный эмулятор Wine, позволяющий запускать под управлением систем семейства Unix программы, написанные для Windows, на самом деле, строго говоря, не является эмулятором, что отражено в его названии (*Wine Is Not Emulator*). Дело в том, что программы, написанные под Windows, за вводом-выводом и другими привилегированными услугами обращаются не напрямую к операционной системе, а к слою системных библиотек, вызывая их функции, составляющие так называемый WinAPI[7]; все эти функции системными вызовами не являются, а «настоящие» системные вызовы в мире Windows вообще не документированы, и прямое их использование в программах не предполагается. Wine подменяет библиотеки Windows своими, реализованными через системные вызовы той системы, на которой он работает; загрузка программы, предназначенной для Windows, в качестве процесса на unix-системах оказывается делом довольно сложным, поскольку адресное пространство процесса на Windows устроено не так, как на unix-системах, но с этой проблемой Wine ухитряется справляться, частично модифицируя код, прочитанный из исполняемого файла.

Ещё один известный эмулятор, Qemu, применяет гибридный подход. Qemu способен имитировать работу разных процессоров, при этом он сам может исполняться на разных архитектурах; при необходимости он применяет интерпретацию машинного кода, но когда эмулируемая система команд совпадает с той, на которой запущен сам Qemu, он для увеличения скорости работы частично выполняет код непосредственно на процессоре, используя для этого поддержку виртуализации, встроенную в ядро операционной системы.

В современных условиях встречаются операционные системы, специально предназначенные для работы на виртуальных машинах; управлять физическим компьютером они не могут. Кроме того, широко распространены специальные модификации ядер систем общего назначения, таких как Linux и FreeBSD, предназначенные для более эффективной работы этих ядер в роли гостевых систем. Такой подход к виртуализации, предполагающий поддержку со стороны ядра гостевой системы, получил название ***паравиртуализации***. В некоторых случаях гостевая операционная система работает в виртуальной машине без модификаций — например, модифицировать ядро Windows не представляется возможным, поскольку его исходные тексты недоступны — но при этом в ней устанавливается набор драйверов устройств специально для работы под управлением конкретного гипервизора. Этот вариант также относят к паравиртуализации.

Рассказ о виртуальных машинах будет неполным без упоминания варианта, когда одно ядро обслуживает как хостовую, так и гостевые системы. Строго говоря, операционная система в этом случае одна, разделяются только пространства пользовательских процессов, причём процессы основной системы (её называют *hardware node*) могут видеть все другие процессы, в том числе и запущенные в рамках виртуальных систем — так называемых *контейнеров* (англ. *container*), или ***виртуальных окружений*** (англ. *virtual environment, VE*). Процессы, работающие внутри контейнера, видят только друг друга; ни процессы основной системы, ни процессы других контейнеров с их точки зрения не существуют. Аналогичная ситуация создаётся также и с файлами: процессы основной системы видят всё дерево каталогов целиком, тогда как процессы внутри контейнера в качестве корневого каталога воспринимают некий каталог, специально созданный для данного контейнера, и видят только поддерево, начинающееся с этого каталога. При старте контейнера основная система запускает в нём его собственный процесс `init`, после чего контейнер работает практически как обычная unix-машина.

Такие виртуальные машины — контейнеры — часто применяются хостинговыми провайдерами. Пользователю для запуска серверных программ предоставляется отдельный контейнер с администраторским доступом к нему, то есть в контейнере существует свой собственный пользователь `root`, не способный влиять на основную систему, но обладающий полной властью над системой внутри контейнера. Пользователь может установить любое нужное ему программное обеспечение (в основном обычно серверное, но ограничиваться этим не обязательно) и настроить его в соответствии со своими пожеланиями. Такой вид услуги называется VPS (*virtual private server*); стоимость VPS может быть в десятки раз ниже, чем аренда физического сервера. Иногда VPS создаётся с использованием гипервизоров (например, Linux KVM), но чаще — в виде контейнеров; это позволяет провайдеру хостинга более эффективно использовать аппаратуру своих серверов, ведь здесь не нужно запускать отдельное ядро ОС для каждой виртуальной машины.

Среди популярных реализаций этого подхода для Linux можно назвать OpenVZ и LXC (*linux containers*). Для работы OpenVZ требуется специально модифицированная версия ядра, LXC работает на обычных ядрах и присутствует в репозиториях многих дистрибутивов. При описании ядра Linux механизм, изолирующий процессы отдельных контейнеров друг от друга, по-английски называют *namespaces*, что переводится как «пространства

[7] Напомним, что аббревиатура API означает *application programming interface*, т. е. *интерфейс прикладного программирования*.

имён» (хотя о каких именах идёт речь, не вполне понятно). Полезно знать, что аналогичная возможность ядра FreeBSD называется *jail*[8].

Наряду с термином «VPS» можно встретить аббревиатуру «VDS», которая расшифровывается как *virtual dedicated server*, но разные хостинговые компании придают этому термину различный смысл. Иногда под VDS понимается виртуальная машина, созданная с помощью гипервизора, то есть имеющая своё собственное ядро, в отличие от контейнеров, которые работают под управлением одного ядра. В других случаях внимание акцентируется на фиксированном объёме памяти и процессорного времени, выделяемых конкретной виртуальной машине, и подчёркивается, что на каждом физическом сервере работает ровно столько виртуальных, чтобы сумма выделенных им ресурсов была равна количеству ресурсов физического сервера; понятное дело, что за такую услугу требуется заплатить больше денег, только проблема в том, что проверить, действительно ли цифры сходятся, заказчик никак не может.

Очень редко под VDS понимают такой виртуальный сервер, который целиком (один) занимает всю физическую машину. Смысл здесь в том, что в таком виде его проще перенести на другую машину, если возникнет такая потребность. Впрочем, опять же, если вам предлагают такое оплатить в качестве услуги, учтите, что вы никак не сможете проверить, нет ли на одном с вами физическом сервере других виртуалок.

На просторах Интернета можно встретить и такое утверждение, будто VPS и VDS — вообще одно и то же. В общем, сохраняйте бдительность и всегда выясняйте, что конкретно имеется в виду.

8.1.4. Структура и основные подсистемы ядра

Ядра операционных систем обычно проектируются по «слоёной» схеме снизу, от непосредственного управления аппаратурой, вверх — к интерфейсу системных вызовов. Каждый следующий слой наращивает степень абстрагированности, то есть предоставляет следующему слою более обобщённые и удобные функции, чем те, на основе которых он реализован сам.

Самый «нижний» слой включает в себя драйверы физических внешних устройств, обработчики аппаратных прерываний и другие компоненты, предназначенные, чтобы справляться с особенностями конкретной аппаратной платформы, в том числе процессора и шины. Именно на этом слое обычно присутствует сравнительно небольшое, но всё же заметное количество программного кода, написанного на языке ассемблера.

Остальные слои не имеют столь же чётких границ; в зависимости от того, какую из подсистем мы обсуждаем, может меняться даже их общее количество. Например, если мы посмотрим на подсистему управления памятью, то в аппаратно-зависимом слое обнаружим функции, работающие с MMU конкретного процессора. «Этажом выше» мы увидим функции, отвечающие за учёт, выделение и освобождение кадров физической памяти; реализация этих функций почти не зависит от того, для какого процессора (аппаратной платформы) компилируется код ядра — в расчёт здесь берётся только то, что на разных платформах размер кадра может отличаться. Имея в своём распоряжении физические кадры, следующий слой реализует более удобную абстракцию *выделения физической памяти* — как для нужд самого ядра, так и для поддержки виртуальных адресных пространств процессов. Ещё выше располагается подсистема, отвечающая за отображение в память дисковых файлов (см. §5.2.7, системный вызов `mmap`), на которую вынуждены полагаться не только подсистема, отвечающая за управление виртуальной памятью, но и подсистема, обеспечивающая работу с файлами (так называемая ***виртуальная файловая система***, англ. *virtual file system, VFS*). Это уже «самый верхний этаж» ядра; соответствующие системные вызовы обращаются непосредственно к этим подсистемам. Надо сказать, что обе эти подсистемы — и подсистема управления виртуальной памятью, и виртуальная файловая система — обращаются также и к другим подсистемам, расположенным в более низких уровнях. Например, подсистеме виртуальной памяти нужны возможности подсистемы, отвечающей за откачку и подкачку (своппинг), а виртуальная файловая система должна, очевидно, обращаться к подсистемам, отвечающим за диски, но не только — для подключения «удалённых» или «сетевых» файловых систем ей нужно взаимодействовать с подсистемой работы с сетью, и т. д.

Если мы попытаемся предпринять аналогичную экскурсию по подсистеме внешних запоминающих устройств (англ. *storage*), то на нижнем слое увидим подпрограммы, обеспечивающие взаимодействие с конкретными типами дисковых контроллеров (так называемые ***драйверы*** физических устройств) и обработчики аппаратных прерываний от дисковых устройств, слоем выше — функции, создающие абстракцию знакомого нам по §5.2.5 *блочного устройства*; здесь же расположены функции, отвечающие за планирование дисковых операций. На блочное устройство полагаются в своей работе уже знакомые нам виртуальная файловая система и подсистема своппинга. Что касается виртуальной файловой системы, то ей для работы требуются также реализации конкретных файловых систем; более подробный разговор о виртуальной файловой системе у нас ещё впереди.

Изучая часть ядра, обеспечивающую взаимодействие по компьютерным сетям, мы на нижнем её слое увидим драйверы (физических) сетевых карт и, как можно догадаться, обработчики их прерываний; уровнем выше реализуется единая абстракция «сетевого интерфейса», ещё выше расположены реализации конкретных сетевых протоколов, затем — модули, поддерживающие семейства протоколов, и, наконец, подсистема, реализующая хорошо знакомые нам *сокеты* (см. гл. 6.3). Ещё одна подобная «башня» из подсистем прослеживается в окрестностях планировщика времени центрального процессора: на нижнем слое имеется подсистема, отвечающая за программирование обработчиков аппаратных прерываний, и сами эти обработчики, в том числе обработчик небезызвестного прерывания таймера; выше расположен собственно планировщик, обслуживающий (в соответствии с данными о приоритетах) так называемые *потоки ядра* (*kernel threads*), на основе которых реализуются в том числе и обычные процессы. Непосредственно над планировщиком располагаются функции, реализующие блокирующие примитивы взаимоисключения (мьютексы ядра и некоторые другие механизмы, предполагающие возможность блокировки), над ними — подсистема, отвечающая за единицы планирования (треды) в том виде, в котором они известны пользователю, а на самом верхнем уровне — подсистема, отвечающая за формирование и ликвидацию пользовательских процессов, включающая, в частности, системные вызовы `fork`, `execve`, `_exit` и `wait`.

Если говорить о неких неведомых «основных подсистемах» ядра, то обычно к ним относят, во-первых, аппаратно-зависимый слой, включающий, помимо прочего, обработчики прерываний и драйверы физических устройств; во-вторых, подсистему управления процессами, включая планирование времени ЦП; в-третьих, менеджер оперативной памяти; в-четвёртых, подсистему внешнего хранилища (*storage*), в которую входит ***виртуальная файловая система*** (*VFS*); в-пятых, сетевую подсистему. Этот список, естественно, можно продолжить, детализировать и т. д., но общее представление о ядре он уже даёт.

Более детально архитектуру ядра операционной системы имеет смысл рассматривать на примере какого-то конкретного ядра, поскольку различия между построением ядра Linux и ядра той же FreeBSD существенны в достаточной мере, чтобы не позволять детальных рассуждений о «ядре вообще». Например, архитектура ядра Linux вплоть до роли отдельных функций разобрана в книге [11]. Столь глубокое проникновение в проблематику ядер операционных систем может быть весьма интересно, но требует большой траты сил и времени; мы посвятим остаток этой части книги сравнительно краткому обсуждению общих принципов функционирования отдельных подсистем ядра, а читателю, заинтересовавшемуся «ядерным программированием», порекомендуем обратиться к специальной литературе и, конечно, к сети Интернет.

8.2. Управление процессами

8.2.1. Процесс как объект ядра системы

Как мы уже знаем, ***процесс*** — это некая сущность, создаваемая ядром операционной системы, чтобы выполнять пользовательскую программу. До сих пор мы рассматривали процессы с точки зрения программиста, пишущего пользовательские программы; процесс при этом воспринимается как нечто такое, что существует *вне* ядра: это прежде всего области памяти, содержащие машинный код программы, её глобальные структуры данных (в том числе кучу) и стек; иначе говоря, *программа и её состояние исполнения.*

В §5.3.1 мы упоминали, что процесс можно рассматривать совершенно иначе — не как нечто внешнее по отношению к ядру, а как *объект ядра.* В самом деле, ядро должно где-то хранить информацию о выполняющейся пользовательской задаче: список выделенных ей областей памяти, приписанные задаче полномочия, таблицу файловых дескрипторов и многие другие сведения. Кроме того, как мы уже хорошо знаем, процесс выполняется не всегда:

[8] Английское слово *jail* формально переводится как «тюрьма», но на самом деле это скорее изолятор, помещение, оборудованное для содержания узников в течение короткого времени, что-то вроде существующих в России СИЗО и спецприёмников; небезызвестные «обезьянники» в отделениях полиции тоже по-английски будут обозначаться как jails.

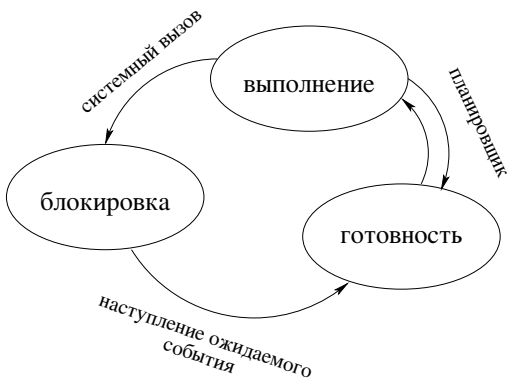

Рис. 8.1. Упрощённая диаграмма состояний процесса

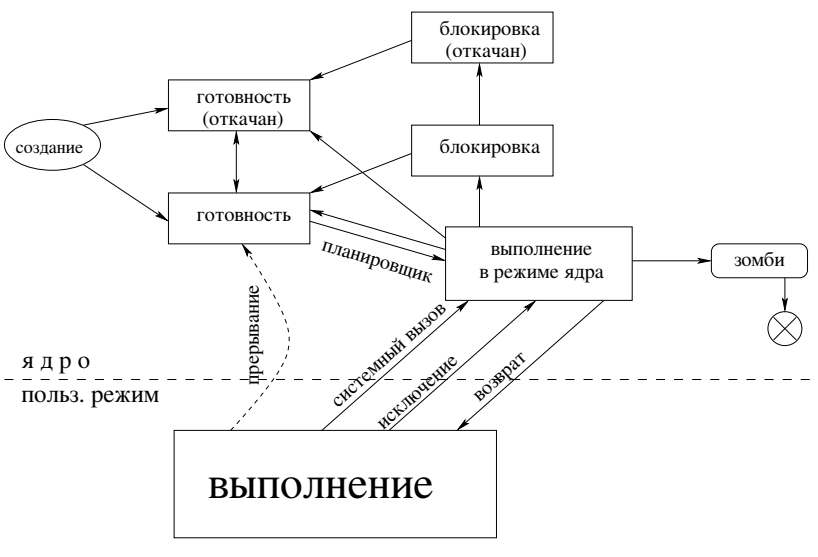

Рис. 8.2. Жизненный цикл процесса

он может быть блокирован или просто ждать своей очереди. Операционной системе нужна возможность снова продолжить исполнять этот процесс, для чего следует помнить значения всех регистров центрального процессора на момент снятия процесса с исполнения. Эту информацию называют ***контекстом выполнения процесса***, и её тоже надо где-то хранить. Естественно, в памяти ядра для всего этого создаётся структура данных, которая как раз и представляет собой процесс как объект.

Для запуска пользовательской программы операционная система должна выделить нужное количество памяти под код, под инициализированные данные, под неинициализированные данные и под стек, после чего первые две области памяти заполнить информацией, хранящейся в исполняемом файле программы, сформировать вышеупомянутый *контекст выполнения*, состоящий из начальных значений регистров. В частности, в этот контекст нужно записать адрес точки входа в программу — вспомните метку `_start` в наших ассемблерных программах — в качестве исходного значения «счётчика команд», а также начальное значение указателя стека; как правило, находятся и другие регистры, начальное значение которых по каким-то причинам важно. Существуют и другие подготовительные действия, зависящие от конкретной системы; например, ОС Unix должна расположить в памяти запускаемой программы слова, из которых состоит командная строка, и занести в стек указатели на эти слова, а также сделать ещё некоторые приготовления подобного рода.

Интересно, что все эти действия могут быть связаны с возникновением нового процесса, а могут и не быть; например, в ОС Unix, как мы знаем, процесс возникает как копия какого-то из существующих процессов, а затем в уже созданном процессе производится замена одной выполняемой программы на другую, то есть действия по запуску программы, перечисленные в предыдущем абзаце, производятся, когда процесс как таковой уже есть.

Вне зависимости от того, в какой момент создаётся процесс и происходит ли при этом загрузка новой программы, для нового процесса должна быть выделена и так или иначе заполнена память, нужно создать настройки MMU для преобразования его виртуальных адресов в физические и т. д.; результатом всей этой подготовительной работы становится полностью готовая к работе структура данных (объект) процесса. Затем новый процесс включается в очередь процессов, готовых к выполнению и ожидающих своей очереди, а управление процессу потом передаст планировщик — подсистема ядра, отвечающая за планирование времени центрального процессора.

В конкретный момент времени процесс может как исполняться, так и не исполняться, причём процесс, который сейчас не исполняется, может быть как готов к возобновлению исполнения, так и не готов: например, процесс может ожидать результатов операции ввода-вывода. Говорят, что процесс может находиться в одном из трёх состояний: ***выполнение***, ***блокировка*** (в ожидании события) и ***готовность*** (см. рис. 8.1). Между состояниями выполнения и готовности процесс переходит при вмешательстве планировщика времени центрального процессора; один процесс может быть снят с выполнения, то есть переведён из состояния выполнения в состояние готовности, а другой при этом, наоборот, поставлен на выполнение. Отметим, что наша диаграмма соответствует положению вещей в системах разделения времени. В системах, реализующих пакетную мультизадачность, переход из состояния выполнения в состояние готовности никогда не производится, то есть аналогичная диаграмма для пакетного режима содержит на одну стрелку меньше.

Если не учитывать возможность откачки памяти процессов на диск (а наша упрощённая диаграмма этого не учитывает), то в состояние блокировки процесс может попасть только одним способом: выполнив такой системный вызов, после которого немедленное продолжение выполнения невозможно. Например, процесс может затребовать чтение данных с диска; в этом случае продолжать выполнение имеет смысл не раньше, чем данные будут прочитаны, что требует времени. Также процесс может в явном виде потребовать приостановить его выполнение на несколько секунд или до поступления внешнего сигнала, вызвав `sleep`, `nanosleep`, `pause` и т. п. Ранее нам встречались и другие случаи блокировки.

Когда компьютер, на котором запущена операционная система, обладает механизмом виртуальной памяти (то есть в процессоре есть ***MMU***), а среди периферийной аппаратуры присутствует внешнее запоминающее устройство (попросту говоря, диск), обычно операционная система при нехватке оперативной памяти может временно ***откачать*** на диск содержимое отдельных областей памяти, принадлежащих пользовательским задачам. Как правило, для этого выбираются такие области памяти, которые давно не использовались их владельцами. При этом у нас появляется ещё одна причина блокировки: если процесс попытался обратиться к области своей виртуальной памяти, которая в настоящий момент откачана, то система запланирует операцию обратной подкачки нужных областей, а сам процесс заблокирует до тех пор, пока подкачка не будет завершена. Как мы увидим несколько позже, виртуальное адресное пространство делится на ***страницы***, и каждая такая страница откачивается и подкачивается как единое целое; при обращении процесса к своей странице, которой нет в памяти, операционная система обычно подкачивает только одну эту страницу.

Кроме того, с самим по себе «выполнением» тоже не всё так просто, мы ведь уже неоднократно были вынуждены упоминать ***выполнение процесса в режиме ядра***, когда в контексте процесса (в его виртуальном адресном пространстве и рамках его единицы планирования) выполняется код ядра в привилегированном режиме процессора. Именно это происходит при обработке системных вызовов и особых ситуаций, вызывающих исключения ЦП (внутренние прерывания), а также перед любой передачей управления коду процесса — перед стартом, при возврате из системного вызова, при постановке на выполнение в начале очередного кванта времени.

С учётом возможности откачки и выполнения в режиме ядра жизненный цикл процесса принимает вид, показанный на рис. 8.2. Под «откачанным» здесь понимается процесс, который не может продолжить работу, поскольку в оперативной памяти отсутствует хотя бы одна из страниц, нужных ему «прямо сейчас» — для выполнения той инструкции, на которую указывает счётчик команд. Это с равным успехом может быть страница из сегмента данных или стека, содержащая ячейки, к которым обратилась очередная машинная инструкция, а также страница из сегмента кода, содержащая саму очередную инструкцию. Если некоторые из страниц, принадлежащих процессу, откачаны, но он при этом может продолжать работу, весь процесс как целое откачанным не считается.

Процесс может оказаться снят с исполнения (перейти в состояние готовности), минуя стадию выполнения в режиме ядра. Это может произойти, если ядро примет решение о смене активного процесса во время обработки аппаратного прерывания (например, прерывания таймера). Возобновление выполнения процесса в любом случае требует подготовительных действий в режиме ядра.

Стоит обратить внимание на то, что процесс, попавший в режим блокировки (например, ожидающий результатов ввода-вывода), может как быть откачан, так и оставаться в памяти, если откачка системе не потребовалась; но если процесс всё же был откачан, обратная подкачка будет осуществлена не раньше, чем процесс окажется готов к выполнению (то есть исчезнет причина блокировки). Если говорить точнее, сначала исчезает основная причина блокировки — наступает событие, которого процесс ждал; после этого процесс остаётся заблокированным, но уже по причине отсутствия в памяти нужных ему страниц, и лишь тогда система принимает меры к их подкачке.

В предыдущей части книги мы обсуждали ***многопоточное программирование***, в основе которого лежит запуск ***легковесных процессов*** (***тредов***). Легковесный процесс представляет собой дополнительную единицу планирования в рамках одного обычного процесса; иначе говоря, обычный процесс в таких системах можно представить как группу легковесных процессов, работающих одновременно с одними и теми же кодом, данными и ресурсами. Как мы видели, с точки зрения программиста это выглядит как возможность запуска некоторых подпрограмм (функций) *параллельно* основной программе — одновременно с продолжением её выполнения. Для каждого легковесного процесса создаётся свой отдельный сегмент стека, чтобы хранить локальные переменные и параметры той функции, которая была запущена параллельно остальной работе, а также функций, вызываемых из неё. Все остальные сегменты у легковесных процессов общие с главным, и даже эти «личные» стеки никак друг от друга не защищены, поскольку находятся в общем адресном пространстве.

Когда в рамках процесса выполняются легковесные процессы, о состоянии (выполнение, готовность, блокировка) имеет смысл говорить в отношении каждого из легковесных процессов отдельно; вообще, каждый из них имеет свой собственный жизненный цикл, практически такой же, как у обычного процесса; единственное различие обнаруживается в порядке завершения легковесных процессов, ведь вызов `wait` для их окончательного снятия не требуется. Естественно, для каждого из них операционная система вынуждена хранить свой контекст выполнения (значения регистров ЦП); собственно говоря, для ядра системы каждый тред — это почти настоящий процесс; одно на всех пространство памяти — это очень заметный фактор с точки зрения программиста, работающего в пользовательском пространстве (пишущего обычные программы), но не слишком значительное обстоятельство при взгляде со стороны ядра.

8.2.2. Планирование времени процессора

Планировщик времени центрального процессора — это важнейшая часть любой многозадачной операционной системы; как мы знаем ещё из первого тома (см. §3.6.1), основные типы мультизадачных операционных систем различаются как раз по принципам планирования времени ЦП.

При описании планировщика иногда выделяют *три уровня планирования*: долгосрочное, среднесрочное и краткосрочное. На долгосрочном уровне принимается принципиальное решение о допуске задачи к выполнению; иначе говоря, долгосрочный планировщик решает, какие задачи будут выполняться в системе одновременно, а какие будут отложены до тех пор, пока не завершится часть уже выполняемых программ. Введение понятия долгосрочного планирования может показаться противоречащим нашему опыту, ведь при работе с ОС Unix все запускаемые программы всегда начинали выполнение немедленно; это действительно так, в системах с разделением времени долгосрочное планирование либо не применяется вовсе, либо применяется, но исключительно для распределения процессов между несколькими процессорами. С другой стороны, долгосрочное планирование крайне важно в системах реального времени, где чрезмерная нагрузка на систему может привести к недостижению её главной цели — обеспечения успешного завершения определённых задач к указанному моменту. Кроме того, долгосрочное планирование применяется в системах пакетной мультизадачности (вроде суперкомпьютеров), где обычно присутствует входящая очередь заданий и нужно принимать решение о запуске очередного задания из этой очереди с учётом текущей загруженности системы.

О среднесрочном планировании обычно говорят при обсуждении откачки процессов. По мере исчерпания физической оперативной памяти ядро вытесняет (откачивает на диск) принадлежащие процессам виртуальные страницы, которые давно не использовались; но рано или поздно может сложиться ситуация, когда оперативную память потребуется освободить, а «ненужных» страниц не найдётся. В этом случае системе придётся принять решение об откачке страниц, без которых тот или иной процесс выполняться дальше не может; процесс при этом перейдёт в состояние «откачан» (см. рис. 8.2 на стр. 201). Принятие решения о том, какой именно из процессов будет откачан, а также о том, какой из откачанных процессов следует подкачать (точнее, подкачать те из его страниц, которые ему нужны для исполнения прямо сейчас), как раз и называется среднесрочным планированием. В системах без своппинга этот уровень планирования отсутствует.

Наконец, ***краткосрочное планирование*** или ***диспетчеризация процессов*** состоит в принятии решений о том, какие из выполняемых процессов должны быть принудительно переведены в режим готовности («вытеснены», англ. *preempted*), какие из готовых к выполнению процессов должны быть поставлены на выполнение при наличии возможности и какой длины кванты времени им следует выделить. Надо сказать, что в системах с разделением времени обычно задействуется только этот уровень планирования: долгосрочный планировщик отсутствует вовсе, а среднесрочный в системе, как правило, есть, но его вынужденное включение в работу свидетельствует о предаварийной перегруженности системы, чего стараются не допускать. В системах с пакетной мультизадачностью, напротив, краткосрочное планирование почти отсутствует, или, лучше сказать, имеет рудиментарную форму. Задачи в пакетном режиме никогда не переводятся из состояния выполнения в состояние готовности, поскольку, напомним, снимаются с выполнения только при завершении и при уходе в блокировку; при наличии свободных процессоров диспетчер просто выбирает ту из готовых к выполнению задач, у которой выше значение приоритета (если в системе вообще есть приоритеты), либо просто выбирает первую задачу, стоящую в очереди.

Вне зависимости от того, с какой системой мы имеем дело, задачи можно условно поделить на те, скорость выполнения которых в основном зависит от доступного процессорного времени, и те, которые, напротив, бóльшую часть времени проводят в блокировках (обычно на операциях ввода). Потребности последних в процессорном времени невелики. По-английски такие задачи называются соответственно *CPU-bound* и *I/O-bound*. Считается, что хороший планировщик должен учитывать разницу между этими задачами и их потребностями. Так, краткосрочный планировщик может заметно повысить общую производительность системы, если будет выделять часто блокирующимся задачам пусть и небольшие кванты времени, но делать это быстро, как только задача вышла из очередной блокировки. Видимое время отклика системы на внешние события это снизит, а на «вычислительные» задачи практически не повлияет. Если при этом в системе присутствует долгосрочный планировщик, то в его задачу обычно входит соблюдение некоего баланса между «вычислительными» и «блокирующимися» задачами, чтобы предоставить краткосрочному планировщику некое пространство для манёвра.

Алгоритмы краткосрочного планирования бывают очень разными, от совсем простых до весьма заковыристых, особенно в системах реального времени. Планирование в реальном времени — отдельная область научных исследований, этому вопросу посвящены целые книги — как популярные, так и монографии; мы эту проблематику затрагивать не будем. Что касается систем разделения времени, то самый простой вариант алгоритма планирования — ***циклический***[9], при котором диспетчер выстраивает все имеющиеся задачи в кольцевой список и выбирает каждый раз следующую задачу, готовую к выполнению, то есть блокированные задачи пропускает при просмотре списка, но не исключает из него; кванты времени выделяются одинаковые.

Чуть сложнее устроен алгоритм очереди (FIFO): диспетчер выстраивает готовые к выполнению задачи в очередь, выделяет задачам одинаковые кванты времени, выбирает для выполнения всегда первую задачу из очереди, а задачи, снятые с выполнения, как и задачи, вышедшие из состояния блокировки, ставит в конец очереди. Алгоритм можно модифицировать: на выходе из блокировки задачу ставить в начало очереди, а не в её конец, чтобы задаче, долго ждавшей какого-то события, не приходилось ждать ещё и своего первого кванта времени; при этом, впрочем, есть риск, что задачи специально будут обращаться к системе и блокироваться на короткое время, чтобы почти сразу получить следующий квант.

Циклическое планирование и планирование на основе очереди не подразумевают никаких приоритетов для выполняемых задач; между тем в реальности часто требуется одним задачам назначить повышенный приоритет,

[9] Английский термин здесь — *round-robin*. Попытки перевести слово «циклический» обратно на английский — всевозможные *cyclic*, *loop* и прочее — ведут к полному непониманию.

чтобы, например, ускорить реакцию на действия пользователя (или клиента, если речь идёт о сервере), тогда как другим установить низкий приоритет, чтобы они выполнялись, когда системе больше «нечего делать» — например, долгие математические расчёты, результат которых требуется не срочно.

В §5.3.2 мы упоминали, что процесс может обладать двумя составляющими его приоритета — статической, задаваемой извне, и динамической, пересчитываемой планировщиком по мере выполнения задач в системе. Простейший вариант планирования с двумя составляющими приоритета выглядит так: динамический приоритет может меняться от нуля до некоторого максимального значения; при постановке задачи на исполнение её динамический приоритет обнуляется, а динамический приоритет всех остальных готовых к выполнению задач увеличивается на единицу. Для постановки на выполнение планировщик выбирает задачу, имеющую максимальную сумму приоритетов — статического и динамического. Некоторые системы позволяли задавать процессам значения статических приоритетов, превышающие максимально возможное динамическое значение; в этом случае такие процессы могут захватить процессор и долгое время не давать выполняться процессам с более низким приоритетом.

В современных системах подобное тоже возможно, но в более гибком варианте. Например, в ОС Linux с помощью системного вызова `sched_setscheduler` для процесса можно задать тип используемого планировщика или, точнее, *режим планирования* (англ. *scheduling policy*): `SCHED_RR`, `SCHED_FIFO`, `SCHED_OTHER`, `SCHED_BATCH` и `SCHED_IDLE`. Обычные процессы используют `SCHED_OTHER`. Режимы `SCHED_RR` и `SCHED_FIFO` называются режимами «реального времени» и отличаются друг от друга применяемым алгоритмом планирования (Round-Robin или FIFO), но при этом имеют всегда заведомо более высокий приоритет, чем все остальные процессы; иначе говоря, при наличии в системе готового к выполнению процесса, имеющего режим `SCHED_RR` и `SCHED_FIFO`, любой другой процесс будет снят с исполнения, чтобы дать возможность работать процессам реального времени. Сами процессы реального времени различаются значением статического приоритета, который задаётся отдельным параметром. Динамическая составляющая приоритета для этих процессов отсутствует, так что процессы с более высоким приоритетом всегда вытесняют процессы, приоритет которых ниже.

Относительно процессов, находящихся в режиме `SCHED_BATCH`, ядро предполагает, что это «счётные» задачи (*CPU-bound*), которым нужно много процессорного времени, но для которых не критично время реакции. При принятии решения ядро отдаёт предпочтение процессам, имеющим режим `SCHED_OTHER`, предполагая, что они займут меньше времени; в то же время этот вариант предпочтения не является абсолютным, то есть рано или поздно процесс в режиме `SCHED_BATCH` всё же получит управление, даже если в системе имеются готовые к выполнению процессы в режиме `SCHED_OTHER`. Наконец, процессы в режиме `SCHED_IDLE` выполняются лишь тогда, когда других задач нет; как только появляется готовый к выполнению процесс, имеющий любой другой режим, процесс в режиме `SCHED_IDLE` немедленно вытесняется.

Описанное выше решение на основе статической и динамической составляющих приоритета даёт возможность понять, как в общих чертах устроен планировщик, но при «лобовой» реализации такой планировщик оказывается неэффективен *сам по себе*, особенно когда в системе много готовых к выполнению процессов. В самом деле, планировщик получает управление, как мы знаем, по прерыванию таймера, и это происходит довольно часто; если слова об увеличении динамического приоритета воспринять буквально, то планировщику придётся при каждом получении управления пройти по всему списку процессов, готовых к выполнению, увеличить каждому из них значение динамического приоритета, найти процесс с максимальной суммой приоритетов, и если она достаточно высока — поставить этот процесс на выполнение, вытеснив один из выполняющихся (либо единственный выполняющийся, если в системе только один процессор или планирование производится для каждого процессора отдельно). Цикл по сравнительно небольшому (в пределах нескольких сотен элементов) списку занимает не так много времени, но поскольку его приходится прогонять, например, тысячу раз в секунду, потери могут оказаться заметными. Кроме того, не следует забывать, что всё это происходит в обработчике прерывания, то есть при запрещённых аппаратных прерываниях. Плюс к тому едва ли не во всех современных компьютерах имеется больше одного процессора, так что доступ к списку процессов приходится считать критической секцией и организовывать взаимоисключение, которое, в довершение всего, не может воспользоваться блокировкой; в самом деле, блокировки обеспечиваются планировщиком, но не может же он обеспечить их сам для себя, не говоря уже о том, что в обработчиках прерываний у нас нет контекста процесса. В целом приходится признать, что устройство планировщика нуждается в усовершенствовании: буквальным образом пересчитывать значение динамического приоритета для всех имеющихся процессов слишком долго.

Реально существующие в современных системах планировщики применяют алгоритмы, позволяющие производить любые пересчёты только при постановке задачи на выполнение и при снятии её с выполнения, а какие бы то ни было циклы по всем имеющимся задачам при этом полностью исключаются. Например, в современных версиях ядра Linux используется так называемый *Completely Fair Scheduler* (CFS; название буквально переводится как «полностью справедливый планировщик»), основанный на принципе «справедливого деления» времени между процессами. Организуется такое «честное деление» довольно просто: для каждого процесса планировщик помнит, сколько времени тот успел отработать, и выбирает для постановки на выполнение те задачи, которые на текущий момент получили времени меньше других. Чтобы объяснить, как этот планировщик работает, представим себе для начала, что все процессы в системе существуют с момента начала работы самой системы. Конечно, такое предположение не соответствует действительности ни в каком виде, но мы вскоре увидим, как от него избавиться.

Заведём для каждого процесса переменную, которая будет хранить время (*буквально* время, измеряемое в наносекундах), предоставленное данному процессу до настоящего момента. В коде ядра Linux эта переменная — точнее, поле структуры, связанной с процессом — называется `vruntime`. Продолжая считать, что все процессы стартовали вместе с системой, отметим, что исходно `vruntime` для всех процессов должна быть равна нулю, ведь они ещё не успели начать работу. Планировщик выбирает процесс с наименьшим значением `vruntime` и ставит его на выполнение, предоставив ему квант времени, длина которого, что важно, определяется значением приоритета — хорошо знакомым нам *nice value* (см. §5.3.9). По истечении этого кванта планировщик снимает процесс с выполнения и возвращает его в очередь, увеличив его `vruntime` (буквально!) на число наносекунд, соответствующее кванту времени, который только что был процессу предоставлен и им использован, после чего снова выбирает процесс с наименьшим значением `vruntime`.

Поскольку кванты времени, как мы уже поняли, могут различаться по своей длине, в очереди на выполнение могут быстро скопиться процессы с довольно большим разбросом значений отработанного времени, так что планировщику нужно содержать очередь в виде, упорядоченном по значению `vruntime`, и уметь эффективно вставлять в эту очередь элементы с сохранением упорядоченности. Планировщик CFS для этого использует структуру данных, известную как *красно-чёрное дерево* — разновидность самобалансирующегося двоичного дерева поиска.

Читатель может обратить внимание, что мы не разбирали работу со сложными структурами данных; в первом томе, рассказывая о двоичных деревьях поиска, мы обошли вниманием алгоритмы их балансировки, ограничившись замечанием, что такие существуют. Это было сделано вполне намеренно: тот уровень знаний и навыков, на который рассчитан текст первого тома, для освоения тех же красно-чёрных деревьев явно недостаточен.

Мы надеемся, что сейчас, подходя к концу второго тома, читатель уже готов к восприятию хитросплетений сложных структур данных; удовлетворить своё любопытство относительно красно-чёрных деревьев и других методов построения сбалансированных деревьев поиска читатель может, воспользовавшись книгами [8], [9] и [10]. Все эти книги упоминались и в первом томе, но сейчас, возможно, у вас получится лучше. Впрочем, для понимания того, как именно работает планировщик CFS, не обязательно в деталях изучать красно-чёрные деревья, достаточно знать, что операции вставки, поиска и удаления элемента для такого дерева выполняются достаточно быстро.

В дереве поиска процессы, успевшие отработать меньше других, оказываются слева; планировщику остаётся только выбрать крайний левый элемент, изъять его из дерева и поставить соответствующий процесс на выполнение, а затем вернуть элемент в дерево уже с новым (увеличенным) значением `vruntime`, так что его новая позиция окажется где-то близко к правому краю дерева. Получается, что процессы в дереве постепенно продвигаются справа налево, так что рано или поздно каждый из готовых к выполнению процессов (а в дереве находятся только такие) получит свой квант времени.

Вернёмся теперь к нашему (довольно безумному) предположению, что все процессы стартовали вместе с системой. Мы прекрасно знаем, что такого быть не может, тем более что процессы приходится помещать в очередь на выполнение не только при их старте, но и при выходе из блокировки, и как раз с этим у описанного принципа планирования имеются явные проблемы. В самом деле, пусть наша система загрузилась час назад, и через небольшое время после загрузки в системе стартовал процесс, активно использующий процессорное время. К текущему моменту этот процесс мог отработать, к примеру, 1000 секунд процессорного времени, т. е. его значение `vruntime` будет равно 10^{12} (напомним, что единица измерения тут — наносекунды, то есть

миллиардные доли секунды). Если теперь новый процесс, только что запущеный на выполнение, поставить в очередь с нулевым значением `vruntime` — что вроде бы логично на первый взгляд, ведь он ещё не успел поработать — то планировщик надолго забудет про старую задачу, виновную лишь в том, что она успела начать работу раньше. Серверным задачам, работающим годами, в такой системе вообще перестанет доставаться время.

Описанная проблема имеет очень простое решение: планировщик помнит *наименьшее* значение `vruntime` среди всех элементов очереди, и именно его (а не ноль!) присваивает полям `vruntime` для новых задач. Получается, что `vruntime` теперь не равна, строго говоря, времени, которое в действительности успел отработать процесс; про эту переменную можно сказать только то, что она *растёт* в соответствии с временем, выделяемым процессу, но в её текущей величине складываются отработанное время и некое начальное значение, определяемое моментом создания задачи.

Остаётся понять, что делать с процессами, выходящими из режима блокировки. Простейшим решением было бы поступать с ними так же, как поступают с новыми процессами: заносить в `vruntime` текущее минимальное значение и помещать процесс в начало очереди (левый край дерева), но это не совсем правильно, поскольку в блокировке процесс мог провести меньше времени, чем другие (готовые к выполнению) процессы проводят в ожидании. Поэтому планировщик поступает чуть хитрее: если при выходе из блокировки значение `vruntime` меньше текущего минимального, то присваивается минимальное значение (чтобы не давать процессу несправедливого преимущества), в противном случае значение `vruntime` сохраняется без изменений. Есть и другие хитрости, связанные с планировщиком CFS; подробности можно узнать в книге [11], а также из посвящённого этому планировщику текста [18] из документации по ядру Linux.

8.2.3. Обработка сигналов

Сигналы, которые мы рассматривали в §5.3.14, при обсуждении устройства ядра интересны нетривиальностью реализации их доставки. Точнее, нетривиальностью отличается лишь один вариант ***диспозиции сигнала*** — выполнение функции-обработчика.

Обсуждая работу с сигналами, мы отметили, что функция-обработчик вызывается «в самый неожиданный момент», что требует определённой осторожности: из обработчиков сигналов нельзя вызывать функции, модифицирующие сколько-нибудь сложные структуры данных за пределами самого обработчика. С точки зрения работающего процесса момент вызова обработчика действительно наступает неожиданно, поскольку, как мы знаем, снятие с выполнения и обратная постановка на выполнение происходят для процесса (точнее, для его пользовательской части — той пользовательской программы, которая в нём выполняется) совершенно прозрачно, то есть процесс не может отследить эти моменты. Но, обсуждая устройство ядра системы, мы можем указать вполне определённый момент, когда (конкретно!) вызывается и выполняется обработчик сигнала.

В §5.3.14 мы отметили, что в определённом смысле обработчик сигнала вызывается прямо из ядра. На самом деле это, конечно же, не так. При прямом вызове функции из кода ядра она сама тоже будет выполняться в режиме ядра, т. е. в привилегированном режиме ЦП, но это, как мы знаем, категорически недопустимо для функции, находящейся в пользовательской программе. Поэтому ядро поступает хитрее.

Прежде всего отметим, что в терминах ядра можно определённо сказать, в какой момент будет вызван обработчик сигнала. Если процессу был (в любой момент времени) отправлен сигнал, диспозиция которого предполагает вызов обработчика, то этот обработчик сработает, когда процесс в очередной раз будет переходить из состояния выполнения в ядре в состояние выполнения в ограниченном режиме (см. рис. 8.2 на стр. 201). Таких ситуаций можно перечислить три: постановка процесса на выполнение после ожидания в режиме готовности, возврат из системного вызова (неважно, блокировался процесс в этом вызове или нет) и возврат к выполнению после возникновения исключительной ситуации (внутреннего прерывания), если соответствующий сигнал (`SIGSEGV`, `SIGFPE`, `SIGBUS` или `SIGILL`) процессом перехвачен или игнорируется.

Во всех случаях перед передачей управления в пользовательский режим ядро проверяет, нет ли для процесса поступивших сигналов, диспозиция которых предполагает выполнение обработчика, и если такой сигнал найден, происходит нечто, лучше всего характеризуемое словосочетанием *black magic*: ядро напрямую модифицирует стек процесса так, чтобы на вершине стека возник стековый фрейм[10] для выполнения обработчика сигнала — такой же, как если бы обработчик был вызван самой программой, за исключением того, что адрес возврата в этом стековом фрейме тоже «хитрый» — он указывает на заранее размещённый ядром в виртуальном адресном пространстве процесса кусочек машинного кода, реализующий одно простое действие: обращение к системному вызову `sigreturn`. Этот вызов приводит стек процесса в исходное состояние, восстанавливает прежнее состояние *маски сигналов* (см. стр. 130) и выполняет некоторую другую работу по ликвидации последствий столь *необычного* вызова функции.

Сам по себе вызов обработчика делается совсем просто: в структуре состояния процесса в поле, хранящее значение регистра `EIP` (на других платформах — в соответствующий регистр, играющий роль указателя на текущую машинную команду), заносится адрес входа в тело функции-обработчика, после чего производится возврат управления коду процесса — самым обычным образом; но после всех описанных нами танцев управление при этом получает функция-обработчик, отрабатывает, выполняет свою инструкцию `RET` («возврат управления»), что приводит, как мы уже сказали, к вызову `sigreturn`. Между прочим, возврат из `sigreturn` тоже представляет собой ситуацию перехода процесса «из ядра» к обычному выполнению, так что здесь ядро вполне может вызвать обработчик следующего сигнала, если он успел прийти.

Понимание того, как реализованы вызовы обработчиков сигналов, проливает свет на многие аспекты обработки сигналов, которые до сей поры могли оставаться непонятными. Так, мы упоминали ранее, что сигналы обладают свойством «склеиваться»: если отправить процессу подряд много сигналов с одним и тем же номером, процесс в итоге вполне может получить только один такой сигнал. Дело в том, что факт получения процессом сигнала ядро отмечает взведением флага, то есть занесением единицы в соответствующий разряд обычной четырёхбайтной целой переменной, связанной с процессом. Перед вызовом обработчика сигнала этот флаг сбрасывается в ноль. Очевидно, что если продолжать «бомбить» процесс сигналами с тем же номером на протяжении временно́го периода до ближайшего перехода процесса «из ядра в юзерспейс», то каждый такой сигнал приведёт к повторному взведению флага, который и так уже взведён.

Кроме того, становится понятно, почему обрабатываемые сигналы прерывают работу блокирующих системных вызовов: для обработки сигнала нужно вернуться в ограниченный режим, то есть *прекратить* выполнение системного вызова, вернуться из него. Впрочем, как мы знаем, современные ядра можно попросить автоматически возвращаться после этого к выполнению прерванных системных вызовов — такой возврат тоже организует вызов `sigreturn`; работает это не для всех блокирующих системных вызовов, что добавляет ещё больше путаницы.

Чтобы лучше показать всю глубину извращённости происходящего, попытаемся описать ещё одну возможность, связанную с обработчиками сигналов, но для этого сначала придётся рассказать о паре стандартных библиотечных функций, которые, возможно, уместнее было бы описать раньше — в главах, посвящённых стандартной библиотеке языка Си; вот только делать этого очень не хотелось, и сейчас станет ясно, почему. Это функции `setjmp` и `longjmp`:

```
int setjmp(jmp_buf env);
void longjmp(jmp_buf env, int val);
```

На самом деле `setjmp` обычно макрос, а не функция, что же касается типа `jmp_buf`, то он в большинстве случаев представляет собой массив, что и позволяет передавать его в обе функции без применения операции взятия адреса, даже если это действительно функции. Вызвав `setjmp`, мы при этом «сохраняем» текущее *состояние выполнения* — попросту говоря, текущую позицию в программе и текущее положение указателя стека. Сама `setjmp` при этом возвращает `0`.

Переменная `env` «действительна» до тех пор, пока продолжает существовать стековый фрейм, из которого вызвали `setjmp`, то есть пока вызвавшая её функция не вернёт управление. В течение всего этого времени можно — из той же функции или из любой функции, вызванной *из неё* прямо или косвенно — вернуться к состоянию выполнения, запомненному с помощью `setjmp`. Для этого вызывается функция `longjmp`; первым параметром она получает переменную, содержащую «законсервированное» состояние, а вторым — некое целочисленное значение, которое должно быть ненулевым. Выглядит «возврат в прошлое» так, как будто функция `setjmp` в том же самом месте снова вернула управление, только на этот раз она возвращает не ноль, а то значение,

[10] Если вы почувствовали какую-то неуверенность, увидев термин «стековый фрейм», вернитесь к первому тому и перечитайте главу 3.3.

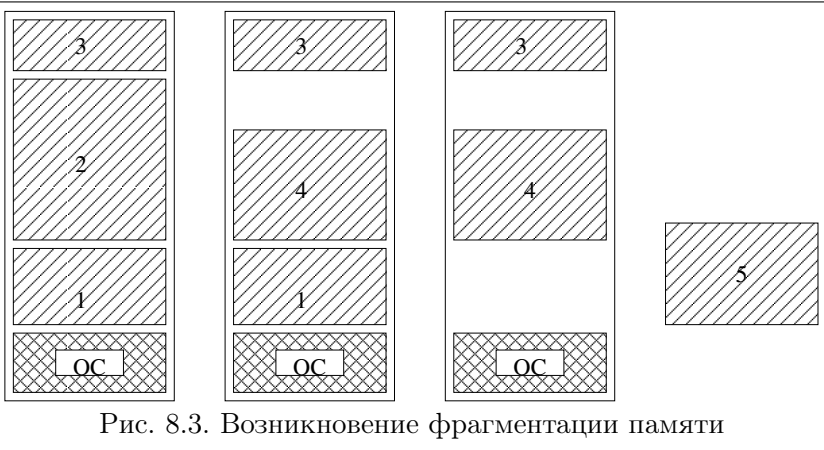

Рис. 8.3. Возникновение фрагментации памяти

которое `longjmp` получила вторым параметром. Отметим, что, конечно, при этом восстанавливается только позиция выполнения и принудительно ликвидируются все более поздние стековые фреймы, но никакие ***побочные эффекты***, в том числе присваивания глобальным переменным, назад не откатываются.

Примечательно сказанное в документации на эти функции для случая, если программисту придёт в голову вызвать `longjmp` для сохранённого состояния *после* того, как функция, в которой состояние было «упаковано», уже вернула управление. В английском оригинале это звучит лаконично и на удивление понятно: *total chaos is guaranteed*.

Реализация этих функций достаточно очевидна, хотя и требует применения ассемблерных вставок. Если говорить об изучавшейся нами системе регистров i386, то `setjmp` записывает в `EAX` значение 0, записывает в `env` значения регистров `ESP` и `EIP` и выполняет `RET`; `longjmp` копирует свой параметр `val` в `EAX`, после чего восстанавливает значения `ESP` и `EIP` из параметра `env`, что приводит к повторному выполнению `RET` в теле `setjmp`, стек при этом находится в том же виде, в котором был на момент её первого вызова, но в `EAX` теперь не ноль, а новое значение — его-то функция `setjmp` и «возвращает» на сей раз.

Зачем нужны эти функции, читатель без труда может догадаться сам, если же с этим возникнут сложности — лучше будет отложить этот вопрос до последнего тома нашей книги, где, помимо прочего, будет описан механизм *обработки исключений* в языке Си++ и некоторых других языках; для тех, кто знает, о чём речь, скажем, что `setjmp` и `longjmp` нужны примерно для того же самого. Дадим только один совет: если вы можете без них обойтись, лучше обойдитесь. Мы отнюдь не случайно решили не рассматривать их в части, посвящённой языку Си.

Теперь, зная о существовании «длинных прыжков» (буквальный перевод слов *long jump*) и представляя, как они реализованы, отметим, что *длинный прыжок можно выполнить изнутри обработчика сигнала*. Автор вынужден признать, что испытал изрядный шок, когда много лет назад впервые узнал об этой возможности. Тем не менее, такая возможность не только есть, но для неё даже предусмотрены специальные варианты функций:

```
int sigsetjmp(sigjmp_buf env, int savesigs);
void siglongjmp(sigjmp_buf env, int val);
```

Ненулевое значение параметра `savesigs` предписывает сохранить текущую маску сигналов; если это было сделано, то `siglongjmp` перед выполнением «прыжка» восстанавливает сохранённую маску. Больше эти две функции от предыдущей пары ничем не отличаются; в частности, «выпрыгнув» из обработчика сигнала, мы *перепрыгнем через вызов `sigreturn`* (!). Впрочем, как легко догадаться, ничего страшного при этом не произойдёт: восстановление стека мы произвели сами, вызвав `siglongjmp`, он же восстановил маску сигналов, так что теперь мы спокойно обойдёмся без услуг ядра по зачистке последствий «магического» вызова обработчика сигнала.

При всей уродливости описанных инструментов автору известен как минимум один случай, когда весьма известная и часто используемая библиотека предполагает выпрыгивание из обработчика сигнала в качестве штатного способа досрочного завершения некоего действия. Речь идёт о библиотеке GNU Readline, которая во многих интерактивных программах вроде командных интерпретаторов или отладчика `gdb` обеспечивает редактирование вводимой команды, дополнение слов по нажатию `Tab`, хранение и поиск истории команд с помощью стрелок вверх/вниз и по нажатию `Ctrl-R`. Если читатель, вняв нашим рекомендациям, сделал командную строку своим основным рабочим инструментом (а мы надеемся, что это именно так, ведь в противном случае читатель вряд ли добрался бы почти до конца второго тома), то он, несомненно, знает, что при нажатии `Ctrl-C` в тех же командных интерпретаторах вводимая строка сбрасывается и её ввод начинается сначала. Чтобы так сделать, нужно, находясь *где-то внутри* функции, собственно осуществляющей чтение строки с клавиатуры со всеми подобающими «примочками», заставить эту функцию завершиться. Столкнувшись с этим, автор попытался понять, как же сообщить библиотеке Readline, что активную функцию следует завершить прямо сейчас. Такого средства не нашлось, зато в документации обнаружилось подробное описание того, как можно перехватить сигнал `SIGINT` (мы помним, что именно он присылается по нажатию `Ctrl-C`) и как из него, а заодно и из пресловутой функции чтения строки выпрыгнуть с помощью `siglongjmp`.

Воистину, сон разума рождает чудовищ.

8.3. Управление оперативной памятью

В ситуации, когда пользовательских задач запущено больше одной, возникает вопрос, как их расположить в памяти вычислительной машины и как разделить память между ними. Естественно, решение возлагается на операционную систему, в которой специально для этого создаётся подсистема, известная как *менеджер оперативной памяти*. Как мы уже знаем, для запуска полноценной мультизадачности требуется, чтобы аппаратура компьютера обладала определённым минимальным набором свойств, в который входит *защита памяти*. В реальности, как мы обсуждали ранее, центральный процессор, предназначенный для многозадачных компьютеров, обычно содержит схему, называемую *устройством управления памятью* (*Memory Management Unit*, MMU), и это устройство производит преобразование адресов перед обращением к физической памяти, создавая возможность для запуска программ в *виртуальном адресном пространстве* (или просто «в виртуальной памяти»). Управление защитой памяти, настройка MMU для преобразования виртуальных адресов в физические — всё это возлагается на менеджер оперативной памяти.

8.3.1. Проблемы, решаемые менеджером памяти

Перечислим проблемы, с которыми сталкивается операционная система при управлении оперативной памятью:

- управление аппаратной защитой памяти;
- недостаток объёма памяти;
- дублирование данных;
- перемещение кода;
- фрагментация.

Остановимся на каждом из пунктов подробнее; начнём по порядку с управления защитой памяти. Цель здесь — в защите процессов друг от друга и операционной системы от процессов. В мультизадачной системе, как мы уже не раз отмечали, необходимы аппаратные механизмы защиты памяти. Управление ими возлагается, естественно, на операционную систему, ведь это требует привилегированных действий. На неё же ложится и обязанность по распределению доступной памяти между процессами.

Объёма оперативной памяти может не хватить для размещения ядра и всех процессов. В такой ситуации современные операционные системы высвобождают физическую память, сбрасывая (***откачивая***) давно не использовавшиеся данные на диск. Когда процессу снова требуются данные из откачанной области памяти, операционная система производит обратную операцию, так что процесс ничего не замечает — происходящее проявляется только в некотором замедлении работы.

Дублирование данных может возникнуть, например, при запуске нескольких копий одной программы: хотя при этом их данные могут различаться, содержимое сегментов кода будет одинаковым. Естественно, такого дублирования желательно избегать.

Код программ может быть привязан к конкретным значениям адресов памяти, в которые загружается программа; например, код может использовать переходы по абсолютным адресам. В мультизадачной ситуации заранее не известно, в какое конкретно место физической памяти придется загружать конкретную программу; если привязать машинный код к физическим адресам, именно это место в памяти может оказаться занято другой программой.

Наконец, при постоянном выделении и освобождении блоков памяти разного размера может возникнуть ситуация, при которой очередной блок размещать негде, хотя общее количество свободной памяти превышает его размер. Пример такой ситуации показан на рис. 8.3. В некоторый момент мы не можем разместить в памяти задачу №5, потому что нет подходящего свободного блока адресов, при том что общее количество свободной памяти превышает размер новой задачи. С проблемой фрагментации связана проблема увеличения размеров существующей задачи в случае, если ей потребовалась дополнительная память: может оказаться, что память за верхней границей задачи занята и расширять её некуда.

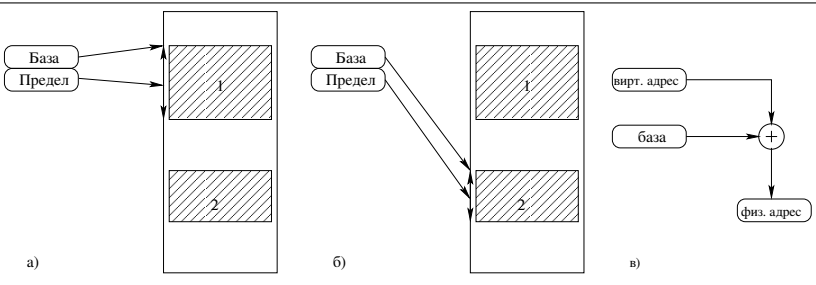

Рис. 8.4. Модель база/предел

Заметим, что большинство перечисленных проблем возникает лишь в мультизадачных системах. В самом деле, если система однозадачна, то защищать, вообще говоря, некого и не от чего, дублирование возникнуть не может (задача всего одна), проблемы с перемещением кода не возникают, так как все программы можно грузить в одну и ту же область памяти; по той же причине отсутствует и фрагментация. Остаётся только проблема объёма (для случая одной задачи, не умещающейся в памяти целиком), но и эта проблема оказывается решаема с помощью оверлейных структур — частей кода, загружаемых и выгружаемых под контролем основной программы, — хотя это и усложняет программирование. В мультизадачной же системе управление оперативной памятью превращается в целый комплекс технических решений, требующих как аппаратной, так и программной поддержки.

8.3.2. Виртуальная память и подкачка

Строго говоря, ***виртуальная память*** не входит в список обязательных условий для реализации мультизадачности; мы много раз упоминали специализированные компьютеры, не имеющие MMU. Но если MMU в процессоре всё-таки есть, работа в виртуальных адресах позволяет эффективно преодолевать проблему перемещения кода, а при использовании достаточно гибкой модели преобразования адресов — облегчает решение проблем дублирования данных, фрагментации и защиты. Обычно для каждой задачи устанавливаются свои правила вычисления физических адресов по виртуальным; перед передачей управления коду задачи операционная система соответствующим образом настраивает MMU. В распоряжении задачи оказывается своё собственное виртуальное адресное пространство, при использовании которого можно никак не учитывать существование других задач. При необходимости задачу (а в некоторых случаях и отдельную её часть) можно перенести в другое место физической памяти, одновременно изменив для этой задачи правила преобразования адресов так, чтобы те же самые виртуальные адреса соответствовали новым физическим. Кроме того, наличие виртуальной памяти, причём обязательно обладающей определёнными свойствами — это необходимое условие для организации ***откачки*** на диск содержимого областей физической памяти, которые временно не используются; без MMU откачивать можно разве что задачи целиком.

Некоторые виртуальные адреса могут не соответствовать физическим; при попытке задачи обратиться к таким адресам возникает исключение (внутреннее прерывание), так что управление получает операционная система; что она будет делать, зависит от конкретной ситуации. Если задача попыталась обратиться к памяти, которой у неё нет и никогда не было, это рассматривается как ошибка (нарушение защиты памяти); задача снимается как аварийная. В то же время обращение задачи на запись к памяти по адресу, непосредственно примыкающему к стеку (с той стороны, в которую растёт стек на данной архитектуре), указывает на необходимость увеличения сегмента стека. Наконец, область, к которой обратилась задача, может отсутствовать в физической памяти из-за того, что ранее система откачала её на диск. В этом случае задачу следует заблокировать, выделить физическую память (или запланировать такое выделение, если прямо сейчас свободной памяти нет), запланировать чтение с диска откачанных данных, а когда всё будет готово — разблокировать задачу и продолжить её выполнение. Подчеркнём ещё раз, что всё это — обязанности операционной системы, и выполнить их она может благодаря тому, что MMU, не сумев преобразовать виртуальный адрес в физический, инициирует исключение, в результате которого ядро системы как раз и получает управление.

8.3.3. Простейшая модель виртуальной памяти

Снабдим процессор двумя регистрами специального назначения, которые будем называть *базой* и *пределом* (соответствующие английские термины — *base* и *limit*). Для простоты будем считать, что в привилегированном режиме значения этих регистров игнорируются. После перехода процессора в ограниченный режим при выполнении любой команды процессор (на аппаратном уровне) к любому заданному командой исполнительному адресу прибавляет значение базы и уже результат этого сложения использует в качестве адреса в физической памяти (рис. 8.4). Одновременно с этим процессор сравнивает значение исполнительного адреса с содержимым регистра «предел»; если обнаруживается превышение, процессор отрабатывает исключение (внутреннее прерывание) «нарушение защиты памяти». Модификация базы и предела при работе процессора в ограниченном режиме запрещена.

Как видим, регистр «база» задаёт адрес, начиная с которого в памяти располагается текущая задача. Исполнительные адреса, задаваемые инструкциями в коде задачи, не совпадают с «настоящими» адресами ячеек памяти, к которым в итоге производится обращение, так что эти адреса можно считать виртуальными. Регистр «предел» задаёт размер блока памяти, доступного текущей задаче. Это позволяет защитить память, находящуюся вне области, выделенной данной задаче, от случайных обращений со стороны текущей задачи.

Адреса в коде задачи теперь формируются в предположении, что задача будет работать в адресном пространстве, начинающемся с нуля. Операционная система может загрузить задачу в любой свободный участок памяти: проблема адаптации программы к адресам решается установкой соответствующих значений базового и предельного регистров. Более того, при необходимости задачу можно переместить в другое место памяти — для этого достаточно скопировать содержимое её области памяти в память по новым (физическим) адресам и изменить соответствующим образом значение базы. Операционная система для передачи управления задаче заполняет базовый и предельный регистры, после чего переключает режим исполнения в ограниченный и передаёт управление коду задачи. При возникновении прерывания, исключения или при выполнении активной задачей системного вызова ограниченный режим снимается и управление вновь получает код операционной системы, который может принять решение о смене активной задачи; для этого придётся в регистры базы и предела занести значения, соответствующие другой задаче, и передать управление на текущую точку этой другой задачи, одновременно включив ограниченный режим ЦП.

Ясно, что проблемы защиты и адаптации к адресам таким образом решены. С проблемой объёма памяти дела обстоят далеко не так гладко: на диск можно сбросить только целиком всю память той или иной задачи. Кроме того, если объёма физической памяти не хватает даже для одной задачи (то есть нашлась такая задача, потребности которой превышают объём физической памяти), описанная модель выйти из положения не позволяет. Проблема фрагментации решается только путём перемещения задач в физической памяти, что влечёт относительно дорогостоящие операции копирования существенных объёмов данных. Проблема дублирования не решена вообще.

8.3.4. Сегментная организация памяти

Усовершенствуем модель «база-предел», для чего введём понятие ***сегмента***. Под сегментом будем понимать область физической памяти, имеющую начало (по аналогии с базой) и длину (по аналогии с пределом). В отличие от предыдущей модели, позволим каждой задаче иметь *несколько* пар база-предел, то есть несколько сегментов.

Конечно, такая смена модели существенно затронет как устройство процессора, так и программное обеспечение, причём если в модели «база-предел» потребовалась поддержка со стороны операционной системы, то в сегментной модели новую архитектуру придётся учитывать при написании кода пользовательских программ — по крайней мере, если мы захотим писать на языке ассемблера; трансляторы языков высокого уровня все нужные моменты учтут автоматически при переводе нашей программы в исполняемый код.

Итак, первое и наиболее видимое новшество состоит в том, что в исполнительном адресе, то есть в адресе, формируемом тем или иным способом инструкциями процессора, появляется специфическая часть, называемая ***селектором сегмента***. Под него можно отвести несколько старших разрядов адреса или использовать отдельный специально предназначенный для этого регистр процессора. Уместно будет вспомнить о существовании в архитектуре i386 «загадочных» регистров `CS`, `DS`, `ES`, `SS`, `FS` и `GS`, которые мы в первом

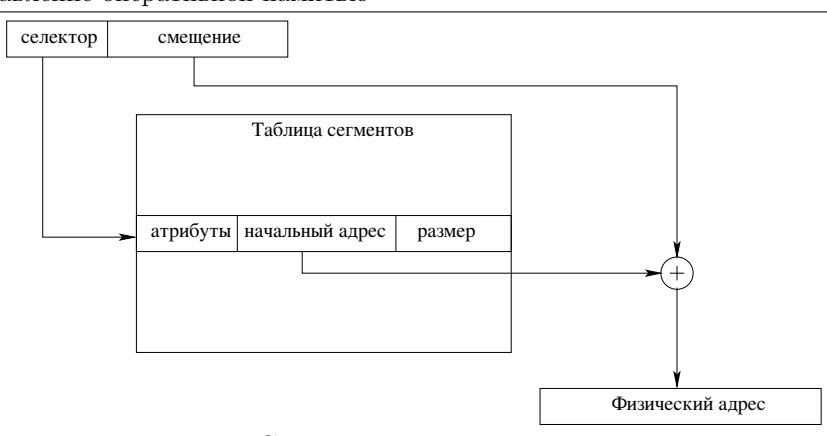

Рис. 8.5. Сегментная организация памяти

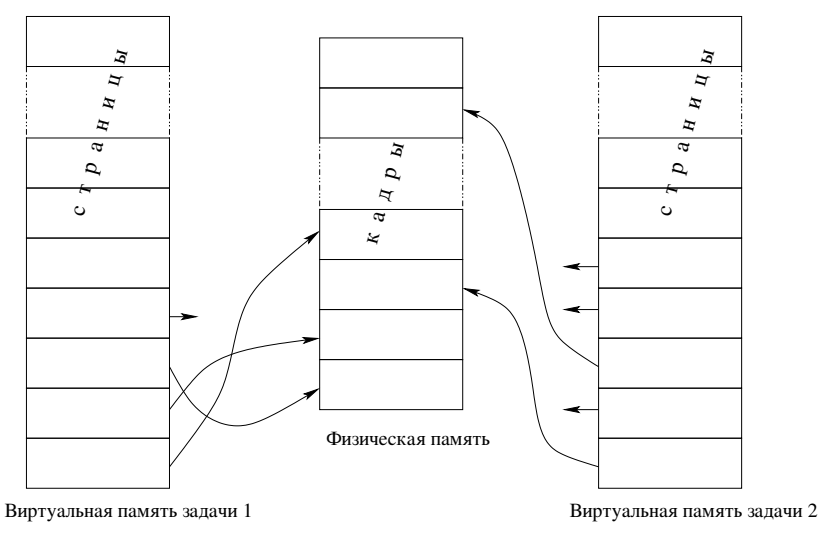

Рис. 8.6. Страничная организация памяти

томе упомянули, но работать с ними не пытались; при использовании сегментной составляющей MMU процессора i386 именно эти (сегментные) регистры используются в роли селекторов сегментов.

Селектор сегмента содержит число, представляющее собой порядковый номер сегмента в ***таблице дескрипторов сегментов***. Для каждого сегмента эта таблица содержит физический адрес его начала и его длину (то есть базу и предел), плюс к этому некоторые служебные параметры — например, флаги, разрешающие или запрещающие запись в этот сегмент, исполнение его содержимого в качестве кода и т. п.; вся эта информация как раз и составляет ***дескриптор сегмента***. Физический адрес ячейки памяти вычисляется путем прибавления адреса начала сегмента, взятого из строки таблицы, выбранной селектором, к смещению, взятому из исполнительного адреса (рис. 8.5).

Возможность завести несколько сегментов для одной задачи позволяет, например, сделать некоторый сегмент общим для двух и более задач, так что теперь мы можем решить проблему дублирования.

Некоторые преимущества можно разглядеть и в отношении проблем объёма и фрагментации. Откачивать на диск по-прежнему требуется сегмент целиком, но при наличии у каждой задачи нескольких сегментов это всё же проще, чем откачивать всю задачу; при перемещении данных с целью дефрагментации можно перемещать не задачу целиком, а лишь некоторые из её сегментов, и, кстати, найти свободную область подходящего размера для размещения сегмента в среднем проще, чем для всей задачи. Впрочем, в действительности эти преимущества всерьёз на ситуацию не влияют; откачку в сегментной модели реализовывать по-прежнему практически бессмысленно, а с фрагментацией выигрыш чисто количественный и не слишком большой.

В определённых случаях сегменты удобны сами по себе. Представьте задачу, требующую нескольких больших таблиц, каждая из которых может увеличиваться в размерах. Если использовать обычную (плоскую) модель памяти, не исключено, что одну из таблиц потребуется перемещать, чтобы дать возможность расширить другую; при этом придётся не только проводить копирование, но и пересчитывать все указатели, содержавшие адреса элементов перемещённой таблицы. Всех этих трудностей можно избежать, если под каждую таблицу выделить отдельный сегмент.

Таблица дескрипторов сегментов хранится в оперативной памяти, так что, если не предпринять специальных мер, на каждое обращение активной программы к памяти потребовалось бы ещё одно — за информацией из этой таблицы, что снизило бы производительность системы практически вдвое. Поэтому разработчики архитектуры процессоров организуют хранение информации об используемых сегментах непосредственно в процессоре. Например, процессоры серии Intel загружают информацию из таблицы дескрипторов каждый раз при изменении содержимого сегментного регистра; информация из соответствующей строки таблицы дескрипторов загружается в «невидимую» часть сегментного регистра и хранится там до следующего его изменения.

8.3.5. Страничная организация памяти

Более эффективно решить проблемы объёма и фрагментации позволяет *страничная* организация памяти. Отметим сразу, что в этой модели обычно каждая задача имеет свое собственное пространство виртуальных адресов и свои таблицы для перевода их в адреса физические.

Разделим физическую память на ***кадры***[11] фиксированного размера, а пространство виртуальных адресов — на ***страницы*** того же размера (рис. 8.6). Если, к примеру, наш процессор использует 32-битные виртуальные адреса, а физической памяти в компьютере установлено 8 GB (2^{33} байт), мы можем разделить физическую память на $2^{21} = 2097152$ кадров по $2^{12} = 4096$ байт, или 4 KB каждый; виртуальное адресное пространство тогда разделится на $2^{20} = 1048576$ страниц такого же размера. Размер страницы и кадра всегда составляет степень двойки (2^n; в нашем примере $n = 12$); младшие n разрядов адреса — как виртуального, так и физического — задают смещение относительно начала страницы или кадра, а остальные биты — номер страницы или кадра.

Остаётся каким-то образом для каждой задачи сопоставить *некоторым* (не всем!) её страницам номера физических кадров. Важно, что сопоставление производится в произвольном порядке, то есть двум соседним страницам могут соответствовать кадры из совершенно разных областей физической памяти. Обычно отображение страниц на физические кадры осуществляется через ***таблицу страниц***, принадлежащую активной задаче (рис. 8.7). Для каждой страницы таблица содержит запись, состоящую из номера кадра и служебных атрибутов. **В число атрибутов страницы обычно входит так называемый *признак присутствия*, означающий, находится ли данная страница в настоящее время в оперативной памяти или нет. При попытке обращения к странице, для которой признак присутствия сброшен, процессор инициирует исключение, называемое *страничным*; получив управление, операционная система производит, если возможно, подкачку соответствующей страницы с диска**.

Существует проблема выбора размера страницы. Чем больше страница, тем больше памяти пропадает впустую в последних страницах задач. Так, если выбрать размер страницы 1 MB, то задача, занимающая чуть больше 1 MB, займёт два физических кадра, причём второй почти весь (то есть почти 1 MB) не будет использоваться. С другой стороны, чем меньше размер страницы, тем больше количество самих страниц и тем, соответственно, больше размеры страничных таблиц. Процессоры i386 работали со страницами размером 4 KB; более поздние процессоры линейки x86, начиная с Pentium,

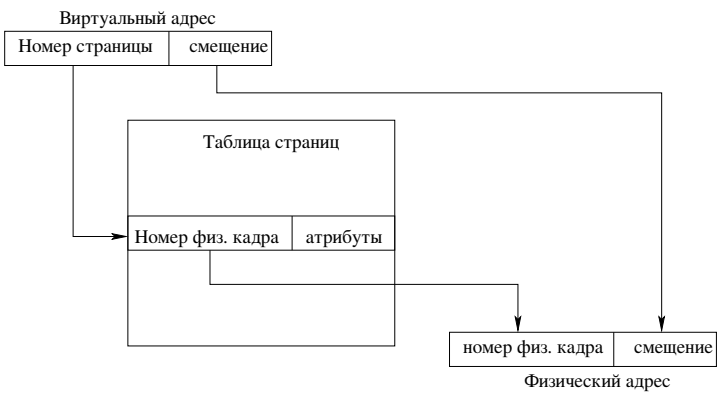

Рис. 8.7. Страничное преобразование адреса

[11] Английский оригинал термина «кадр» в данном случае — *frame*; некоторые русскоязычные авторы пользуются вместо слова «кадр» словом «фрейм», либо используют слово «страница» как для обозначения страниц виртуальных адресов, так и для обозначения кадров физической памяти.

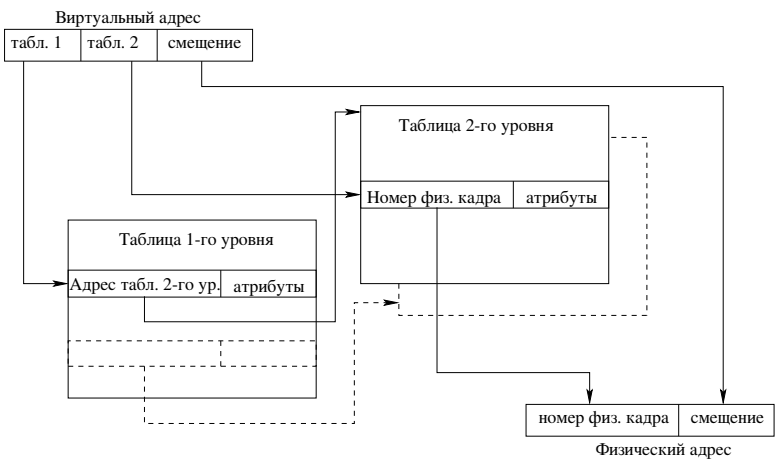

Рис. 8.8. Двухуровневая схема страничного преобразования

умеют работать со страницами размером 4 KB, 2 MB и 4 MB. Размер страницы 4 KB (2^{12} байт) можно считать наиболее популярным среди различных архитектур. Обратим внимание, что при 32-битной адресации (4 GB адресуемого пространства) количество страниц составит 2^{20}, то есть больше миллиона. Если предположить, что каждая строка в таблице страниц занимает 4 байта (для i386 это так и есть), получим, что на таблицу страниц каждого процесса требуется 4 MB памяти. Безусловно, это неприемлемо: большинство задач занимает в памяти существенно меньше одного мегабайта, так что на таблицы страниц придётся потратить больше памяти, чем на сами задачи. Одним из возможных решений могло бы стать искусственное ограничение количества страниц, но это решение далеко не лучшее, поскольку приводит к невозможности выполнения задач, требующих бóльшего количества памяти.

Более элегантное решение состоит в применении ***многоуровневых страничных таблиц***. В этом случае номер страницы делится на группы битов, задающих номер строки в таблицах соответствующих уровней; таблица первого уровня содержит адреса таблиц второго уровня, и т. д., а таблица последнего уровня — номера физических кадров (рис. 8.8). При описании двухуровневой схемы таблицу первого уровня часто называют ***каталогом страничных таблиц***, а таблицы второго уровня — просто страничными таблицами, без уточнения насчёт уровня[12]. Для каждой задачи при использовании такой схемы придётся завести таблицу верхнего уровня и по одной таблице для всех остальных уровней; остальные таблицы будут добавляться по мере надобности. Поскольку схема, как правило, содержит не более трёх уровней[13], для небольшой задачи требуется хранить всего две или три сравнительно небольшие таблицы.

Рассмотрим для примера двухуровневую страничную схему для случая 32-битных адресов и размера страницы 4 KB. Смещение при таком размере страницы занимает 12 бит, на номер страницы остаётся 20 бит, из которых 10 используется для выбора строки в таблице первого уровня, остальные 10 — для выбора строки в таблице второго уровня. Именно так обстоят дела в i386-совместимых процессорах; в документации обычно таблица первого уровня называется каталогом таблиц (или страничным каталогом, англ. *page directory*), а таблицы второго уровня — собственно таблицами страниц. Тогда таблицы обоих уровней могут содержать $2^{10} = 1024$ строки, что при длине строки в 4 байта требует 4096 байт, то есть каждая таблица занимает ровно один кадр в памяти.

Как можно заметить, страничная организация памяти снимает проблемы объёма и фрагментации. Действительно, в рассмотренной модели физическая память выделяется кадрами, причём совершенно безразлично, будут ли кадры соседними или нет; любой свободный кадр может быть использован для любой задачи. Фрагментация здесь просто не может возникнуть, ей неоткуда взяться. Каждая страница независимо от других может быть в любой момент откачана на диск. При обращении к ней произойдёт страничное прерывание, по которому операционная система может подкачать страницу обратно в память, после чего продолжить выполнение задачи в точности с инструкции, вызвавшей прерывание. Возможно, страница будет подкачана в совершенно другой физический кадр, но пользовательская задача этого не заметит, как и вообще самого факта подкачки. Благодаря независимой обработке отдельных страниц можно выделить для отдельно взятой задачи больше памяти, чем физически есть на машине; в этом мы ограничены только объёмом виртуального адресного пространства и, конечно, дисковым пространством, выделенным под откачку (своппинг).

Отметим, что разрядности физического и виртуального адресов не обязаны совпадать, причём физический адрес по своему размеру может быть как больше, так и меньше адреса виртуального. Разрядность физического адреса совпадает с количеством дорожек на шине адресов (см. т. 1, §1.1.2) и определяет максимально возможное количество физической памяти, которое можно к этой шине (и к этому процессору) подключить. Виртуальное адресное пространство к характеристикам шины никак не привязано и определяется обычно разрядностью соответствующих регистров процессора.

К недостаткам страничной модели можно отнести, во-первых, сравнительно большое количество неиспользуемой памяти в конце последней страницы каждой задачи, особенно при больших размерах страниц, и, во-вторых, большие объёмы служебной информации, которую процессор вынужден загружать из страничных таблиц, находящихся в оперативной памяти. Остановимся на этом подробнее. Если не принять специальных мер, при работе по двухуровневой табличной схеме каждое обращение к памяти за командой или данными потребует ещё двух обращений: к таблице первого уровня и к таблице второго уровня. Ситуация во многом аналогична возникшей в сегментной модели, за исключением того, что страница — не сегмент, её объём существенно меньше, а количество страниц — гораздо больше; программа может работать с сотнями, тысячами, даже миллионами страниц. Ясно, что хранить информацию о страничном соответствии в регистровой памяти (как это делается для сегментных дескрипторов) не получится из-за большого объёма.

Решить проблему позволяет встроенное в процессор специальное устройство, называемое ***ассоциативной памятью***[14]; оно представляет собой таблицу из двух полей — номер виртуальной страницы и информация о соответствующем ей кадре (номер кадра и его атрибуты). Электронная схема ассоциативной памяти устроена так, что запрос к ней формируется не по номеру строки таблицы, а по *значению первого поля*, то есть по ключу. Сличение предъявленного значения со значениями ключевого поля производится одновременно во всех строках таблицы (параллельными схемами). Если в одной из строк значение совпадает, в качестве результата выдаётся второе поле той же строки, то есть номер кадра и его атрибуты.

При обращении к странице, о которой в ассоциативной памяти нет информации, фиксируется так называемый ***промах*** и производится преобразование через страничные таблицы, то есть к страницам всё-таки приходится обратиться, но после этого информация, полученная из таблиц, записывается в ассоциативную память. Если теперь программа обратится к той же самой странице, читать содержимое таблиц уже не потребуется. Когда все строки таблицы оказываются заполнены, очередное обращение к неизвестной странице приводит к тому, что информация об одной из известных страниц из ассоциативной памяти удаляется. Количество строк TLB варьируется в достаточно широких пределах: можно встретить процессоры с TLB на восемь строк и на несколько тысяч. В любом случае разработчики процессоров стараются сделать TLB достаточно большим, чтобы удерживать вероятность промаха в пределах долей процента.

Следует отметить, что информация об отображении страниц локальна для каждой задачи, так что при смене активной задачи содержимое ассоциативной памяти приходится сбрасывать, и некоторое время новая задача тратит на заполнение ассоциативной памяти своими данными. Это ещё сильнее увеличивает и без того высокую стоимость переключения контекста, но выбора у нас нет.

Отметим также, что управление ассоциативной памятью можно возложить на операционную систему. В этом случае процессор вообще не делает никаких предположений о том, как устроены структуры данных, отвечающие за отображение страниц; это полностью отдано на откуп операционной системе. Когда в ассоциативной памяти отсутствует строка для требуемого номера страницы, процессор генерирует исключение, в ответ на которое операционная система должна программно произвести поиск или иное вычисление соответствующей информации и занести эту информацию в ассоциативную

[12] В литературе можно встретить противоположный подход к нумерации уровней: таблицами первого уровня называют таблицы, содержащие номера физических кадров, таблицами второго уровня — таблицы, содержащие адреса таблиц первого уровня и т. д. Всё это не более чем вопрос принятой терминологии.

[13] Обычно два уровня для 32-битных архитектур и три — для 64-битных.

[14] В англоязычной литературе обычно используется термин «*Translation Lookaside Buffer*», TLB, а в русскоязычной можно встретить термин «буфер быстрого преобразования адреса».

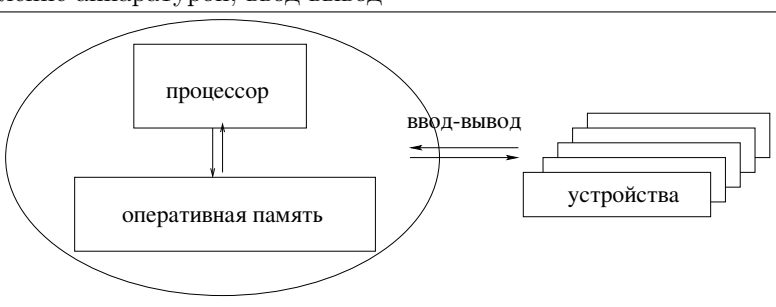

Рис. 8.9. Ввод-вывод с точки зрения аппаратуры

память, после чего возобновить исполнение активной задачи. Достоинство такого подхода состоит в его очевидной гибкости: операционная система может применять для хранения сведений о виртуальных страницах любые структуры данных, какие сочтёт нужным — например, самостоятельно определять, сколько будет уровней страничных таблиц и каков будет их размер; с другой стороны, цена «промаха» оказывается много выше из-за вынужденного переключения контекста при передаче управления в ядро ОС.

8.3.6. Сегментно-страничная организация памяти

Как говорилось ранее, сегментная модель организации памяти имеет свои специфические достоинства, делая адресное пространство задачи многомерным. Сегментно-страничная модель представляет собой попытку объединить достоинства сегментной и страничной организации памяти. В этой модели виртуальный исполнительный адрес претерпевает двойное преобразование: сначала из адреса выделяется сегментная часть, которая с использованием таблицы дескрипторов сегментов превращается в «плоский» (но всё ещё виртуальный) адрес; полученный адрес подвергается страничному преобразованию, результатом которого становится физический адрес. Иначе говоря, в сегментно-страничной модели виртуальной памяти каждая задача имеет своё виртуальное страничное адресное пространство, внутри которого могут быть организованы сегменты.

Сегментно-страничная организация реализована на i386-совместимых процессорах, но больше, собственно говоря, нигде не встречается. Основной недостаток этой модели — её чрезмерная сложность; большинство операционных систем этой моделью не пользуется. Свою роль здесь играет и то обстоятельство, что операционную систему, которая задействовала бы сегментно-страничную модель виртуальной памяти во всём её великолепии, нельзя было бы перенести на другие процессоры, где страничное преобразование адресов присутствует (оно в наше время присутствует практически на всех процессорах, имеющих MMU), а сегментное отсутствует. Изучая программирование на языке ассемблера, мы видели, что пользовательские задачи под управлением ОС Linux и FreeBSD «живут» в плоской модели памяти; операционная система создает несколько «сегментов», «накрывающих» все страничное пространство целиком, и дескрипторы этих сегментов заносит в сегментные регистры, на чём и успокаивается. Трогать содержимое сегментных регистров задаче не рекомендуется, ни к чему хорошему это не приведёт.

8.4. Управление аппаратурой; ввод-вывод

8.4.1. Две точки зрения на ввод-вывод

Схематически ***ввод-вывод*** (или, говоря шире, ***управление устройствами***) показан на рис. 8.9. Центральный процессор и оперативная память занимают в этой схеме несколько особое место: несмотря на то, что и процессор, и память, несомненно, являются техническими устройствами, они не входят в число «устройств», об управлении которыми идёт речь. Иногда во избежание путаницы говорят об управлении ***внешними устройствами***, при этом подразумевается, что процессор и память — устройства «внутренние», хотя так их и не называют. Считается, что взаимодействие между процессором и памятью не входит в понятие ввода-вывода; с точки зрения, принятой при обсуждении аппаратного обеспечения (в том числе среди программистов, создающих ядра операционных систем), ЦП и память представляют собой единый конструктив, а вводом-выводом считается обмен информацией между этим конструктивом и всем остальным миром.

На то, что считать, а что не считать вводом-выводом, существует и иная точка зрения, которая принята среди прикладных программистов. Эта точка зрения могла бы полностью совпадать с предыдущей, если бы не тот факт, что одна и та же программа, выполняя одни и те же действия, в зависимости от обстоятельств, в которых её запустили, может как осуществлять аппаратный ввод-вывод, так и не осуществлять его[15]. Рассмотрим для примера хорошо знакомую нам программу «Hello, world». Если вызвать её командой

```
./hello > file1
```

то строка `"Hello, world!"` будет выведена в файл `file1` на диске, то есть произойдёт аппаратный ввод-вывод. Если же *та же самая программа* будет запущена командой

```
./hello | ./prog2
```

то никакого ввода-вывода (с аппаратной точки зрения) не последует: строка, выведенная программой `hello`, будет подана на вход программе `prog2`, так что всё взаимодействие, скорее всего, останется между процессором и памятью.

Если использовать «аппаратную» терминологию, программистам придётся в каждой такой ситуации уточнять, что речь идёт не о вводе-выводе, а о его *возможности* или *высокой вероятности* (или низкой, зависит от задачи), и такое уточнение потребуется едва ли не в каждом месте, где речь идёт о вводе-выводе. Очевидно, что такой вариант терминологии оказывается неудобен в достаточной степени, чтобы его никто не использовал. Поэтому прикладные программисты обычно под вводом-выводом понимают любой обмен информацией с внешним миром через потоки ввода-вывода, то есть через вызовы `read`, `write`, `send`/`recv` и т. п. Конечно, такое «определение» ввода-вывода очевидным образом порочно, ведь оно фактически сводится к тому, что «ввод-вывод — это то, что делается через вызовы для ввода-вывода»; с таким же успехом можно было бы заявить, что ввод-вывод — это ввод-вывод.

Можно было бы сказать, что вводом-выводом с программистской точки зрения считается вообще любой обмен информацией между процессом и внешним миром, но это не так: процесс может как принимать, так и получать информацию, например, через сигналы или через разделяемую память, и программисты традиционно не рассматривают всё это как ввод-вывод. Остаётся лишь констатировать, что техническая предметная область в очередной раз демонстрирует своё нежелание соответствовать каким бы то ни было определениям.

8.4.2. Взаимодействие ОС с аппаратурой

Мы уже обсуждали (см. т. 1, §1.1.4), что центральный процессор взаимодействует с остальными частями компьютера через общую шину, а для подключения к этой шине разных внешних устройств используются ***контроллеры***[16]. Контроллеры, будучи своего рода посредниками между процессором и внешними устройствами, позволяют до определённой степени унифицировать способы подключения периферийных устройств к компьютеру. Очевидно, что никакой из выводов («ножек») процессора не может соответствовать конкретным действиям с внешними устройствами, таким как перемещение читающих головок, запуск или останов моторов и т. п.: во-первых, различных действий такого рода слишком много, а выводов у процессора — ограниченное количество, и, во-вторых, такое построение архитектуры ограничило бы ассортимент устройств, которыми может управлять данный конкретный процессор — например, к процессору нельзя было бы подключить никакое устройство, созданное позже данного процессора и, как следствие, не предусмотренное его конструкцией. Контроллеры эту проблему решают: благодаря им от процессора оказываются скрыты схематические особенности различных устройств, то есть с точки зрения процессора устройства различаются между собой разве что номерами.

Контроллер представляет собой электронную логическую схему, подключаемую с одной стороны к общей шине компьютера, что позволяет обмениваться данными с процессором, а с другой стороны — к физическому устройству, которым нужно управлять. Поскольку контроллер создается всегда для

[15] Имеется в виду, естественно, ввод-вывод через системные вызовы.

[16] Здесь будет уместно напомнить, что основной перевод английского глагола *to control* на русский язык — *управлять*, а вовсе не «контролировать», как часто ошибочно полагают.

управления конкретным устройством, для него не составляет проблемы генерировать именно такие электрические сигналы, которые нужны для управления данным устройством. С другой стороны, поскольку любой контроллер взаимодействует с процессором путём приёма и передачи неких числовых значений по шине, процессор оказывается способен работать с любым контроллером, который можно физически подключить к общей шине данного компьютера; особенности работы с конкретным контроллером задаются программой (драйвером устройства), а не самим процессором. Благодаря этому процессор может, в частности, работать с устройствами, которых на момент выпуска данного процессора ещё не существовало.

Поскольку к общей шине может быть подключено одновременно большое количество различных контроллеров, нужно каким-то образом различать, кому предназначены передаваемые данные или, наоборот, кто должен ответить на запрос данных. Для этого вводится понятие ***порта ввода-вывода***. Порт ввода-вывода представляет собой абстракцию, имеющую адрес, представимый на данной шине. Для каждого порта возможны (с точки зрения процессора) операции *записи* и *чтения*, то есть соответственно передачи значения по заданному адресу и запроса значения с заданного адреса. Диапазон возможных значений, как и диапазон адресов портов, зависит от архитектуры шины (определяется разрядностью соответственно шины данных и шины адресов). За каждым контроллером числится один или несколько (а иногда целая область) портов ввода-вывода, причём возможно, что некоторые из них могут быть только прочитаны, а некоторые — только записаны.

Можно заметить, что порт ввода-вывода похож на обыкновенную ячейку памяти. Более того, в некоторых архитектурах, как мы увидим позже, порты ввода-вывода располагаются в том же адресном пространстве, что и обычная оперативная память (адреса памяти в данном случае имеются в виду, естественно, ***физические***), и читаются/записываются теми же командами процессора. Конечно, на самом деле порт ввода-вывода ячейкой памяти не является, поскольку ничего не хранит, а значения, читаемые из порта, обычно отличаются от значений, туда заносимых, даже для портов, доступных и на чтение, и на запись. Значение, записываемое в порт, может представлять собой, например, код команды включения мотора жёсткого диска, а значение, из того же порта читаемое,— код, показывающий, завершена ли для этого диска последняя операция чтения.

Кроме портов, некоторые контроллеры имеют ещё и *буферы ввода-вывода*, представляющие собой оперативную память, конструктивно интегрированную в контроллер и предназначенную для обмена массивами информации между контроллером и центральным процессором. Например, данные, предназначенные для записи на диск, нужно скопировать в буфер контроллера этого диска, а затем дать через порт вывода команду на запись; наоборот, при чтении информации с диска она помещается в буфер контроллера, откуда её можно потом скопировать в основную память. Другим примером буфера ввода-вывода можно считать видеопамять, то есть память, в которой хранится изображение, видимое на экране дисплея.

На некоторых процессорах порты ввода-вывода имеют отдельное адресное пространство (как правило, меньшей разрядности, чем пространство адресов оперативной памяти). В этом случае для работы с портами используются отдельные инструкции процессора (например, `IN` и `OUT` вместо `MOV`).

Альтернативное решение — разместить контроллеры устройств в том же адресном пространстве, что и оперативную память. В этом случае процессору не нужны отдельные инструкции для работы с портами, все делается обычной командой `MOV`. Такая схема была, например, реализована на PDP-11. Одно из достоинств этой схемы — возможность сопоставить области портов ввода-вывода виртуальным адресам некоторых процессов, что позволит им взаимодействовать с устройствами напрямую, без обращений к операционной системе (системных вызовов), что называется, «на каждый чих». Это, в частности, упрощает вынесение драйверов отдельных устройств из ядра в пользовательские процессы. Конечно, можно обойтись и без прямого доступа к портам — существуют операционные системы, в которых драйверы аппаратуры, работая как обычные пользовательские процессы, обращаются к портам через системные вызовы; но работает такая конструкция не слишком эффективно. Больше того, отсутствие специальных инструкций для взаимодействия с портами делает возможным написание драйверов на языках высокого уровня (таких как Си, Си++ или даже Паскаль — лишь бы в языке поддерживались указатели) без применения вставок на языке ассемблера.

К недостаткам единого адресного пространства можно отнести, например, тот факт, что на порты расходуется основное адресное пространство, которого может быть не так много. На PDP-11 эта проблема была достаточно актуальна, т. к. адреса на этой машине были 16-разрядные. Кроме того, на современных процессорах применяется кеш-память, так что участки адресного пространства, соответствующие буферам и портам, из процесса кеширования приходится искусственно исключать. Наконец, оперативная память и контроллеры устройств представляют собой сущности весьма различные, и поместить их на одну общую шину может оказаться затруднительно (хотя бы даже в силу того, что оптимальная разрядность шины данных для памяти и контроллеров различна), а построение двух разных шин при использовании одного общего пространства адресов приводит к необходимости на том или ином уровне принимать решение, по какой шине следует работать с конкретным адресом.

Ещё один возможный подход, применяемый в хорошо знакомой нам архитектуре i386 — комбинированный. Порты ввода-вывода в этом случае размещаются в отдельном адресном пространстве, а буферы — в общем.

8.4.3. Драйверы

Под ***драйвером*** исходно понималось некое *программное средство*, берущее на себя заботу об особенностях управления определённым внешним устройством. Напомним, что в задачи операционной системы входит *управление аппаратурой*, которое подразумевает *абстрагирование* и *координацию*; можно считать, что абстрагирование как раз и обеспечивается драйверами — они предоставляют некий унифицированный набор программных функций, позволяющих управлять внешним устройством, не зная, как это на самом деле делается: соответствующее знание разработчики драйвера вложили в реализацию его функций.

Стоит обратить внимание, что мы назвали драйвер каким-то «программным средством», а не просто «программой». Этому есть вполне внятная причина: строго говоря, драйвер не является программой, он представляет собой набор подпрограмм (функций), пусть даже связанных между собой, но именно подпрограмм, рассчитанных на то, что вызывать их будет кто-то ещё — в большинстве случаев имеются в виду другие подсистемы ядра ОС. В этом плане драйвер своей структурой напоминает скорее библиотеку, чем программу, хотя по назначению драйвер, конечно, далеко не библиотека. Если говорить о *программе*, то программой, работающей с внешними устройствами, как можно догадаться, является ядро ОС, ну а драйвер с момента его загрузки становится (обычно) частью ядра.

Зная, как нужно работать с данным конкретным аппаратным устройством, драйвер скрывает особенности этого устройства от всей остальной системы. Другие части ядра обращаются к драйверу, вызывая его внешние функции — так называемые ***точки входа***, которые имеют одинаковую структуру для всех драйверов определённой категории. Так, обращение к драйверу жёсткого диска с целью записать сектор №55 и такое же обращение к драйверу flash-брелка[17] будет (с точки зрения любой подсистемы ядра, кроме самих этих драйверов) выглядеть одинаково, несмотря на то, что эти устройства на аппаратном уровне управляются совершенно по-разному. Детали процесса управления конкретными устройствами оказываются локализованы в коде драйвера. При написании остальных частей операционной системы, а также при создании пользовательского программного обеспечения становится возможно их не учитывать; именно так и достигается искомое *абстрагирование*.

В большинстве случаев драйвер рассчитан на выполнение в привилегированном режиме, то есть должен быть частью ядра; привилегированные команды нужны драйверу для обращения к контроллеру (то есть к портам ввода-вывода) через шину. Теоретически возможно выполнение драйвера и в пользовательском режиме в виде обычного процесса; при этом драйвер должен сообщать операционной системе, какое число в какой порт следует отправить и из какого порта прочитать ответ, а ядро уже выполняет эти действия и сообщает драйверу о результатах. Необходимость переключения контекста между процессом драйвера и ядром может существенно сказаться на эффективности работы, так что применяется такой подход достаточно редко. Отметим, что при этом драйвер всё-таки превращается в обычную программу, хотя и имеющую специфическое назначение.

[17] Правильнее будет сказать — к драйверу USB-устройств класса mass storage.

Вернёмся к наиболее распространённой ситуации, когда драйвер работает в виде части ядра операционной системы. Можно выделить три основных способа помещения драйвера в ядро: включение кода драйвера в качестве модуля на этапе сборки ядра, подключение драйвера на этапе загрузки операционной системы и динамическая загрузка модулей в ядро.

Если код драйвера включается в ядро в качестве модуля прямо при его сборке, то добавление или удаление драйвера требует пересборки (перекомпиляции) ядра и перезагрузки операционной системы. При этом пользователю должны быть доступны исходные тексты ядра либо, как минимум, его объектные модули и соответствующие им интерфейсные файлы. Такой подход считается традиционным для операционных систем семейства Unix, хотя в последние 10-15 лет он едва ли не полностью вытеснен из реальной практики динамической загрузкой модулей.

Подключение драйвера на этапе загрузки операционной системы предполагает, что в системе имеется файл, содержащий список драйверов, и сами драйверы в виде отдельных файлов. При загрузке ядро анализирует список и подключает драйверы, после чего начинает работу. В этом случае для подключения дополнительного драйвера достаточно перезагрузить систему; перекомпиляция ядра не требуется. Традиционно этот подход использовали системы семейства Windows, а ранее — и MS-DOS.

Динамическая загрузка модулей в ядро позволяет добавлять драйверы в уже работающую систему без её перезагрузки. Для этого достаточно подготовить файл драйвера, соблюдающий определенные соглашения об именах, и выдать ядру соответствующий системный вызов, после чего модуль становится частью ядра. Ненужные модули обычно можно из ядра изъять. Такая возможность присутствует практически на всех современных unix-системах, включая Linux и FreeBSD. Стоит отметить, что при высокой гибкости эта модель считается небезопасной, т. к. фактически позволяет при наличии прав системного администратора запустить произвольный код в привилегированном режиме.

Конкретный набор «внешних функций» (точек входа), которые драйвер предоставляет остальной системе, составляет ***интерфейс драйвера***. Абстрагирование достигается тем, что драйверы совершенно разных устройств могут реализовывать один и тот же интерфейс, то есть иметь (и предоставлять остальной системе) внешне полностью одинаковые наборы точек входа. Остальным подсистемам ядра при этом становится всё равно, с каким устройством в действительности происходит работа, ведь эта работа для разных устройств требует вызова одинаковых функций. Обращение к точкам входа обычно производится через указатели на функции (см. §4.9.2). Драйвер *экспортирует* свои точки входа, то есть сообщает остальному ядру об их существовании, предоставляя некую таблицу их адресов; это может быть или массив указателей на функции, или (чаще) структура с полями-указателями.

Конечно, всё многообразие внешних устройств не позволяет придумать единый интерфейс драйвера, который подходил бы вообще для любого устройства. Обычно операционная система поддерживает несколько ***типов устройств***; каждый такой тип есть не что иное, как фиксированный набор точек входа, которые должен предоставлять соответствующий драйвер.

Наличие у драйверов унифицированного интерфейса — фиксированного набора точек входа — открывает одну интересную возможность. Если в систему поместить драйвер, реализующий все положенные точки входа, то с точки зрения всего остального ядра, а также, естественно, с точки зрения пользовательских программ в системе появится новое устройство; как именно оно будет реализовано — никого, кроме собственно драйвера, не касается. Это позволяет заводить в системе ***виртуальные устройства*** — такие, за которыми на самом деле вообще нет ничего аппаратного; их функциональность полностью реализуется драйвером без обращений к каким-либо реальным устройствам. Классическим примером такого устройства можно считать ***виртуальный диск*** (англ. *virtual disk* или, чаще, *RAM disk*). Драйвер такого «диска» при старте выделяет себе определённый объём оперативной памяти и эмулирует операцию записи сектора путём размещения записываемой информации в этой памяти, а операцию чтения сектора — путём копирования информации из памяти. Виртуальный диск со всех возможных точек зрения выглядит как обычный диск, его можно отформатировать (то есть создать на нём файловую систему) точно так же, как это делается с обычными дисками, и смонтировать на выбранную точку монтирования.

Виртуальные устройства не следует путать с ***логическими устройствами***, несмотря на несомненную схожесть этих двух понятий. Как виртуальные, так и логические устройства представляют собой реализованную программно абстракцию «устройства, которого на самом деле нет»; но если виртуальное устройство — это имитация физического устройства программными средствами (с помощью драйвера, который на самом деле не управляет никаким физическим устройством, а вместо этого занимается программной эмуляцией), то логические устройства — это дополнительный уровень абстрагирования, реализованный *поверх* «настоящих» устройств вместе с их драйверами. Например, физический диск может быть поделен (*размечен*) на несколько ***разделов*** (англ. *partitions*), каждый из которых будет представлен в системе как отдельный *логический* диск; в то же самое время физический диск можно не делить на разделы, используя его как единое целое, и в этом случае можно будет сказать, что логическое представление диска совпадает с его физическими параметрами.

При эксплуатации серверных компьютерных систем часто используют так называемые RAID-массивы[18]; в этом случае ситуация противоположна — один логический диск оказывается реализован на основе (*поверх*) двух или более (вплоть до нескольких десятков) физических дисков. Наконец, многие системы (включая Linux и FreeBSD) позволяют представить в виде дискового устройства обыкновенный файл; обычно эта возможность называется *loopback device*. Это довольно удобно, когда у вас есть файл образа диска, предназначенный для записи чаще всего на CD, DVD или флеш-брелок и включающий всю информацию, содержащуюся на физическом диске, в том числе в его системных областях; может так случиться, что вам потребуется извлечь из этого образа отдельные файлы, при этом реально записывать образ на «болванку» или брелок не захочется. Здесь тоже речь идёт о логическом диске, а не о виртуальном, поскольку нельзя сказать, что такого диска «на самом деле нет»: он есть, просто он на самом деле не диск, но при этом программный слой, отвечающий за логические устройства, представил его в виде диска.

Отметим, что *все* устройства, видимые пользовательским программам, следует считать именно логическими. Даже если созданная ядром системы абстракция полностью отвечает физическим свойствам «настоящего» устройства, всё равно через системные вызовы нам доступна именно логическая абстракция, которая может быть (в очень редких случаях) «прозрачной», то есть проявлять ровно такие свойства, которыми обладает реальное физическое устройство.

Некоторую путаницу создаёт тот факт, что термин «драйвер устройства» применяют как к драйверам физических устройств (их мы уже обсуждали), так и к компонентам «логического» слоя ядра, которые отвечают за формирование абстракций логических устройств. Драйверы логических устройств по своей сути очень похожи на драйверы устройств физических: это тоже некие наборы функций, включающие унифицированные наборы ***точек входа*** — функций, вызываемых из других частей ядра. Набор точек входа, поддерживаемых драйвером, составляет ***интерфейс драйвера***. Мы уже обсуждали (см. §5.2.5), что файлы устройств в системах семейства Unix делятся на два типа: ***символьные*** (или ***потоковые***) и ***блочные***; эти два типа файлов устройств как раз и соответствуют двум «классическим» типам драйверов *логических устройств* в Unix. Наборы точек входа у драйверов символьных устройств и у драйверов блочных устройств различны; при этом любой драйвер символьного устройства должен иметь такой набор точек входа, какой предусмотрен в данной системе для символьных устройств; то же самое будет верно для блочных устройств и вообще для драйверов устройств того или иного типа. Можно сказать, что сами понятия «символьного» и «блочного» устройства обозначают просто два вида интерфейса драйвера.

Помимо символьных и блочных устройств, система может поддерживать и другие типы драйверов логических устройств. Так, в ОС Linux поддерживается третий тип логического устройства — сетевые интерфейсы. Как мы знаем, сетевые карты в Linux не имеют представления в виде файлов, так что к ним нельзя обратиться с помощью обычных системных вызовов, таких как `read`, `write` или `ioctl`; естественно, интерфейс драйвера сетевой карты или другого сетевого интерфейса «выглядит» с точки зрения всей остальной системы совершенно не так, как диск или простой коммуникационный порт, то есть набор точек входа — функций, вызываемых из других частей ядра — для драйвера сетевого интерфейса не такой, как для драйверов других типов.

[18] Название образовано как сокращение от *Redundant Array of Independent Disks* — избыточный массив независимых дисков.

Рис. 8.10. Ввод-вывод на разных уровнях вычислительной системы

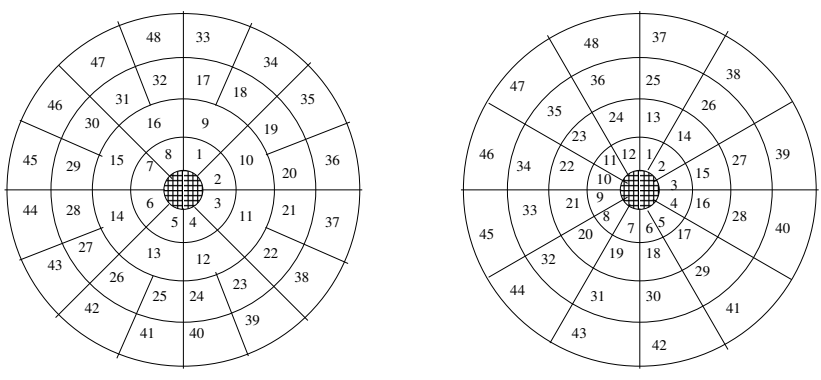

Рис. 8.11. Физическая и виртуальная геометрия диска

Интересно отметить, что создатели ядра FreeBSD в какой-то момент решили отказаться от поддержки блочных устройств, исходя из несколько неожиданных соображений. При обсуждении файлов устройств в §5.2.5 мы назвали основным отличием между символьными и блочными устройствами возможность позиционирования, то есть применимость вызова `lseek`. Авторы ядра FreeBSD к этому вопросу подходят иначе: с их точки зрения основной особенностью блочных устройств следует считать их буферизацию в памяти ядра, а это, в частности, означает, что данные, полученные ядром системы через вызов `write` для записи на блочное устройство, могут неопределённо долго находиться в памяти ядра, не попадая при этом на физический диск, и этот аспект никак не контролируется системными вызовами. К обсуждению буферизации мы вернёмся позже. Так или иначе, в ядре FreeBSD блочные устройства как отдельная сущность отсутствуют, а для драйверов символьных устройств предусмотрена опциональная возможность отработки вызова `lseek` — естественно, не для всех существующих устройств. Создатели Linux пошли другим путём, предусмотрев при открытии файла устройства дополнительный флаг для вызова `open` — `O_DIRECT`; этот флаг позволяет обойти слой буферизации.

Программный слой, создающий логические абстракции, не всегда представляет их в форме «устройства». Так, ***виртуальная файловая система***, которой мы позже посвятим отдельный параграф, позволяет монтировать файловые системы, не связанные ни с каким устройством — ни физическим, ни виртуальным, ни даже логическим. Дело в том, что реализация файловой системы в ядре тоже представляет собой фиксированный набор функций, подобных точкам входа драйвера; обычные файловые системы реализуют свои функции через обращения к структурам данных на диске, но никто не мешает реализовать их каким-то иным способом.

Поскольку виртуальные устройства представляют собой имитацию физических устройств, их место в архитектуре системы расположено рядом с физическими устройствами; это позволяет выстраивать логические устройства и другие логические абстракции как поверх физических, так и поверх виртуальных устройств.

Выше мы приводили примеры, связанные с дисками; системы семейства Unix поддерживают также и потоковые (символьные) устройства, не имеющие под собой физического содержания. С некоторыми из них мы уже знакомы (см. §5.2.5, стр. 108): это `/dev/null`, `/dev/zero`, `/dev/full` и `/dev/random`. Вопрос, считать ли эти устройства «логическими» или «виртуальными», относится скорее к области философии; в литературе встречаются как оба этих термина, так и использованный нами ранее термин ***псевдоустройство***. Если следовать терминологии, введённой нами выше, перечисленные устройства правильнее будет отнести к логическим, поскольку они реализуются в том слое ядра системы, который отвечает за логические устройства.

8.4.4. Ввод-вывод на разных уровнях ВС

Как известно, под ***вычислительной системой*** понимается компьютер вместе со всеми программами, которые на нём выполняются. Припомнив «слоёное» описание ядра операционной системы, начатое в §8.1.4, мы можем заметить, что аналогичным образом на отдельные слои можно разделить не только ядро, но и всю вычислительную систему — и аппаратуру, и пользовательские программы. Конечно, любое такое деление остаётся условным, поскольку реальность обычно ни в какие абстрактные схемы не укладывается; но если некая абстракция позволяет нам лучше понять происходящее, то она полезна хотя бы этим, несмотря на неполное соответствие реальности.

Аппаратуру мы делить на слои не будем, хотя такое деление, безусловно, тоже возможно. Для нужд нашего изложения важнее структура программного обеспечения — как внутри ядра системы, так и за его пределами. Из пользовательского кода мы выделим в отдельный слой библиотеки — ту их часть, которая включает функции-обёртки системных вызовов, а также функции, так или иначе к системным вызовам обращающиеся. Поскольку мы намерены рассматривать ввод-вывод, внутри ядра отметим ***виртуальную файловую систему***, которая, помимо прочего, поддерживает *потоки ввода-вывода* как объекты ядра; кроме того, нас, естественно, заинтересует уровень, создающий абстракции логических устройств, и уровень драйверов физических устройств.

Компоненты каждого слоя, получив «сверху» соответствующее обращение, переводят его в термины следующего уровня и передают ниже, а полученный ответ переводят, наоборот, в термины уровня предыдущего и отправляют наверх. По мере движения от аппаратуры к пользователю нарастает уровень абстрагирования, а сложность описания падает.

Рассмотрим для примера запуск программы с выводом информации в файл (рис. 8.10). Пользователь подаёт пользовательской программе (в данном случае — командному интерпретатору) команду на понятном пользователю языке. Программа использует для вывода библиотечную функцию высокого уровня. Библиотека функций переводит полученный запрос на язык, понятный более низкому уровню (ядру операционной системы); таким языком будет интерфейс системных вызовов, а переведённый запрос становится системным вызовом (например, `write`) с соответствующими параметрами.

Со стороны операционной системы вызов `write` обрабатывает виртуальная файловая система, поскольку объект открытого файла находится именно в ней. Параметры системного вызова `write` здесь преобразуются в последовательность действий, необходимых для записи полученной информации в файл, на открытый дескриптор которого сослался вызвавший процесс. Говоря конкретнее, модуль файловой системы переводит запрос в последовательность операций по модификации секторов логического дискового устройства.

Драйвер логического диска, в свою очередь, должен выполнить последовательность операций над секторами диска физического. Соответствующую последовательность запросов он передаёт низкоуровневым подпрограммам ядра, включающим драйвер физического диска. Этот драйвер уже осуществляет действия, нужные для выполнения поступившего запроса с учётом особенностей конкретного физического дискового устройства, то есть переводит полученные запросы на язык команд контроллера.

Современные контроллеры периферийных устройств сами представляют собой достаточно сложные устройства, позволяющие в ряде случаев абстрагироваться от некоторых особенностей аппаратуры, что снимает часть нагрузки с драйверов. Так, большинство современных жёстких дисков на самом деле имеет геометрию (расположение секторов), отличающуюся от видимой для операционной системы. Например, на дорожках, расположенных ближе к центру диска, для сохранения более-менее постоянной плотности записи данных размещают меньше секторов, чем на внешних дорожках (рис. 8.11); но с точки зрения всей остальной системы геометрия диска состоит из N_h сторон, N_c дорожек на каждой стороне (цилиндров), N_s секторов на каждой дорожке, причём все три параметра постоянны и не зависят от значения других. Функцию отображения виртуальной геометрии на физическую берёт на себя контроллер диска. Таким образом, внутри слоя, обозначенного нами как «аппаратура», также имеются различные уровни абстрагирования.

Если говорить совсем строго, то в современных условиях контроллер дисков, подключённый к шине компьютера, на самом деле работает не с диском, а с неким абстрагированным аппаратным интерфейсом; читатель наверняка не раз встречал обозначения таких интерфейсов — SATA, SCSI или уже практически вышедший из употребления IDE. Каждый из них тоже представляет собой шину, «по ту сторону» которой находится очередной контроллер. Физически этот контроллер размещён прямо на самом диске — это электронные печатные платы, расположенные на нижней стороне его корпуса.

Рис. 8.12. Уровни организации ввода-вывода

8.4.5. О роли аппаратных прерываний

Чтобы понять, как функционируют драйверы физических устройств, важно обратить внимание на то, как организовано *ожидание момента, когда устройство завершает требуемую операцию*. Допустим, драйвер устройства выполнил последовательность команд записи в порты, инициировав тем самым некоторую операцию ввода-вывода. Поскольку скорость работы внешних устройств конечна и в большинстве случаев сравнительно невысока, до завершения операции (то есть до момента, когда можно будет воспользоваться результатом) должно пройти некоторое время. В первом томе мы обсуждали (см. §3.6.2) два подхода к ожиданию этого момента: через циклический опрос порта (активное ожидание) и через обработчик аппаратного прерывания. В первом случае, отправив контроллеру запрос на выполнение какого-то действия, драйвер в цикле выполняет чтение из порта контроллера, чтобы узнать, завершило ли устройство выполнять запрошенное действие. Всё время от начала работы внешнего устройства до её окончания центральный процессор не занят ничем полезным, то есть вычислительное время попросту теряется.

Во втором случае драйвер ничего не ждёт; вместо этого он настраивает обработчик соответствующего аппаратного прерывания и на этом завершает работу, освобождая процессор для более важных дел. Другие подсистемы ядра, получив управление, возвращённое драйвером, при необходимости переводят процесс, затребовавший (прямо или косвенно) данную операцию, в режим блокировки и ставят на выполнение другой процесс, если готовые к выполнению процессы в системе есть, если же их нет — переводят отдельно взятый процессор или всю систему в «состояние покоя» (*idle state*), в котором аппаратура, отвечающая за вычисления (прежде всего центральный процессор) может потреблять меньше электричества. Устройство, завершив заданную драйвером операцию, выдаёт запрос прерывания, в результате которого драйвер устройства снова получает управление (на сей раз — от обработчика прерывания) и выполняет действия, нужные для завершения операции и получения ее результатов. О результатах драйвер сообщает вышестоящей подсистеме. При этом возможно, что процесс, ожидавший результата операции, будет переведён из режима блокировки в режим готовности.

В некоторых случаях оказывается выгоднее (с точки зрения затрат процессорного времени) применить активное ожидание, а не обработку прерывания; обычно так бывает, если ожидаемое время отклика устройства столь незначительно, что потери времени центрального процессора на активное ожидание оказываются меньше, чем потенциальные затраты времени на настройку обработчика прерывания, а затем на обработку самого прерывания. Именно такова ситуация, если драйверу потребовалось узнать некую информацию о текущем состоянии устройства, причём заведомо известно, что ответ на запрос не требует от устройства выполнения каких-либо действий: например, это может быть запрос к сетевой карте относительно того, подключён к ней (физически) провод для передачи данных или нет.

Место обработчиков прерываний в системе можно проиллюстрировать, выделив их в отдельный слой (рис. 8.12). Разумеется, такое деление остаётся во многом условным. Слой аппаратно-независимых компонентов здесь символизирует всё, что есть в операционной системе выше уровня драйверов физических устройств, включая и реализацию логических дисков, и виртуальную файловую систему. Слой обработчиков прерываний не является слоем в полном смысле слова. Драйверы физических устройств отдают распоряжения устройствам о проведении тех или иных операций напрямую, и так же напрямую могут извлекать результаты; в обоих случаях драйвер, являясь программой, выполняется на центральном процессоре и использует команды чтения/записи портов ввода-вывода, взаимодействуя с контроллером устройства. Однако момент, когда следует начать извлечение результатов операции из контроллера, может выбираться с помощью обработчиков прерываний, т. к. именно определённое прерывание сигнализирует о том, что данные для этого готовы. Обработка прерываний может сыграть свою роль и при выборе момента для инициирования операции ввода-вывода. Действительно, в момент, когда драйвер получил на вход запрос на проведение операции, устройство, которым управляет этот драйвер, может быть ещё занято обработкой предыдущей операции. В этом случае драйверу придётся подождать её завершения, о котором сообщит прерывание; после извлечения результатов можно будет приступить к началу следующей операции с устройством.

8.4.6. Буферизация ввода-вывода

При выполнении операции ввода-вывода бывает так, что соответствующую процедуру на физическом уровне начать в данный момент невозможно. Допустим, процесс A инициировал запись в дисковый файл; в результате цепочки вызовов запрос на эту операцию трансформировался в запрос на запись определённого сектора диска и был передан контроллеру в виде инструкции на позиционирование головки, ожидание нужной фазы поворота диска и запись сектора. Эти физические операции требуют времени, так что процесс A переводится в режим блокировки, а на выполнение запускается процесс B. Допустим теперь, что процесс B тоже потребовал записи в файл, причём этот файл находится на том же физическом диске. Операционная система может трансформировать и его запрос в набор физических операций, однако передать их контроллеру диска не представляется возможным, ведь контроллер всё ещё занят выполнением заказа процесса A.

Логично было бы блокировать процесс B до освобождения контроллера, а затем уже приступить к выполнению его операции. Представим теперь следующую ситуацию. Контроллер по заданию процесса A занят выполнением операции в области последних цилиндров диска. В это время операционной системе пришлось сначала блокировать процесс B, требующий операции в области первых цилиндров, а затем — процесс C, требующий снова операции с последними цилиндрами. Если рассматривать соответствующие запросы в порядке поступления, контроллеру придётся сначала перевести головку из конечной в начальную позицию, а затем снова в конечную. Если процессы A, B и C будут продолжать активно использовать диск, такие переводы головки туда и обратно могут весьма негативно сказаться на общем быстродействии.

Попытки оптимизации (сначала разбудить процесс C, затем уже B) потребуют от планировщика знаний о том, как следует оптимизировать последовательности запросов к данному конкретному устройству. Давать планировщику такие знания нежелательно, ведь они специфичны для разных типов устройств, а такую специфичную информацию не хотелось бы выпускать за пределы драйверов устройств. Кроме того, процессу в его дальнейшей работе, возможно, и не требуется дожидаться результатов операции вывода, в противоположность операции ввода, результат которой, скорее всего, необходим в вычислениях. Поэтому блокирование процесса в ожидании доступности контроллера для операции может оказаться идеей неудачной.

Рассмотрим теперь операцию ввода данных. Пусть процесс A запросил чтение данных из файла и был заблокирован в ожидании поступления этих данных. Через некоторое время обработчик прерывания сообщил драйверу устройства о готовности запрошенных данных. Теперь драйверу нужно скопировать данные из буфера контроллера куда-то в память. Логично было бы копировать непосредственно в пространство процесса A, однако здесь можно столкнуться ещё с одной проблемой: соответствующая область памяти может оказаться откачана на диск. В этом случае для разгрузки буфера контроллера от поступивших данных драйверу придётся сначала инициировать загрузку в память нужных страниц из области процесса A. Чтобы понять недопустимость такого варианта, достаточно представить себе, что область подкачки располагается на том же физическом диске, что и прочитанные данные. В этой ситуации драйвер загонит систему в порочный круг: чтобы освободить контроллер, необходимо сначала подкачать в память страницы процесса A, но чтобы это сделать, надо сначала освободить контроллер.

Как видим, ядру часто желательно, а иногда и просто необходимо при проведении операций ввода-вывода сохранять данные в неких областях памяти для промежуточного хранения. Такие области памяти называются ***буферами ввода-вывода***; их не следует, разумеется, путать с буферами контроллеров устройств.

Для дисковых устройств буферы организуются как набор порций данных, соответствующих тому, что должно находиться (но, возможно, пока не

находится) в секторах физического диска. В итоге при операции записи вместо непосредственного обращения к драйверу устройства верхний слой ядра просто заносит переданную ему информацию в буфер. Драйвер устройства, имеющий доступ к буферам, самостоятельно определит очерёдность, в которой содержимое буферов будет реально записано на диск; например, при наличии нескольких буферов, соответствующих смежным областям диска, драйвер может записать их все, прежде чем переходить к другим буферам, вне зависимости от того, сколько времени назад соответствующие буферы были созданы.

Как можно догадаться, буферизация часто приводит к уменьшению числа физических операций. Так, если сначала процесс A потребовал записи некоторого сектора, затем процесс B потребовал чтения того же сектора, модифицировал полученные данные и тоже потребовал записи, то физически операция с диском, возможно, произойдёт всего одна. Действительно, информация от процесса A окажется сохранена в буфере; операция чтения, запрошенная процессом B, ни к каким физическим действиям не приведёт: ему просто выдадут информацию из буфера. Наконец, запрос процесса B на запись модифицирует содержимое уже существующего буфера, и, если к этому времени операция записи всё ещё не была выполнена, вместо двух операций теперь потребуется только одна: запись последней версии информации. Такие ситуации действительно нередки: системные области диска, отвечающие за глобальные параметры файловой системы, подвергаются модификации очень часто, и экономия может достигать тысяч логических операций на одну физическую.

Как мы знаем, внешние устройства бывают не только дисковые (блочные); обмен данными с устройствами других видов обычно организуется в виде потоков — последовательностей байтов. Именно такие потоки используются при передаче данных на печать, при передаче информации по локальной сети или по модемному каналу и т. п. Здесь также возникают определённые трудности, в основном обусловленные ограничением скорости обмена с периферийными устройствами; частично справиться с ними позволяет опять-таки буферная память.

Буферизация потокового вывода позволяет процессам не ждать результатов выполнения операции вывода. Что касается буферизации ввода, то с её помощью можно накапливать информацию, полученную от внешнего источника, чтобы выдать её читающему процессу в один приём (за один системный вызов). Поскольку системный вызов, как мы знаем, представляет собой операцию дорогостоящую в плане использования процессорного времени, а информация от внешнего устройства может приходить небольшими порциями и даже отдельными символами (байтами), буферизация здесь также приводит к экономии времени. Кроме того, в некоторых случаях буферизация последовательного ввода оказывается необходима по тем же причинам, что и буферизация ввода дискового: системе может потребоваться очистить контроллер, когда процесс, для которого предназначены полученные данные, либо находится в откачанном состоянии, либо занят другими действиями и не выполняет вызов чтения.

Поскольку буфер может понадобиться в самый неподходящий момент и времени на его подкачку не будет, буферы ввода-вывода всегда размещаются в памяти ядра, которая не подлежит откачке и подкачке. Каждый дополнительный буфер уменьшает количество доступной физической памяти. Ясно, что общий объём буферов в системе оказывается ограничен. При достижении предельного совокупного объёма дисковых буферов некоторые из них могут быть ликвидированы, чтобы освободить память. Ликвидировать, однако, можно не всякий буфер: так, если в буфере содержатся данные, подлежащие записи на диск, но ещё на диск не записанные, уничтожать такой буфер нельзя.

Если предельный объём достигнут, а буферов, допускающих ликвидацию, нет, очередная операция дискового вывода будет заблокирована до тех пор, пока драйвер не запишет на диск информацию из некоторых существующих буферов и не освободит буферную память. Аналогично происходит и работа с буферами последовательного вывода. При их переполнении очередная операция вывода блокируется до тех пор, пока в буфере не освободится место. Наиболее тяжёлая ситуация складывается при переполнении буфера последовательного ввода. Такое может произойти, например, если по каналу связи поступают данные, а процесс, отвечающий за их получение и обработку, по каким-то причинам чтения данных не производит — например, просто не успевает. В этом случае возможна потеря входящих данных, разрыв соединения по каналу связи и т. п. Ясно, что таких ситуаций следует по возможности избегать.

Отдельного внимания заслуживает вопрос, на какую из подсистем ядра следует возложить буферизацию. Традиционный подход — назначить держателями буферов драйверы логических устройств, как символьных (потоковых), так и блочных. С потоковыми устройствами всё до сих пор обычно так и есть, что же касается дискового ввода-вывода, то методы его буферизации изрядно развились. Так, в современных ядрах Linux присутствует подсистема, именуемая Page Cache (страничный кеш), которая изначально была предназначена вовсе не для буферизации ввода-вывода, а для отслеживания статуса имеющихся страниц физической памяти, в том числе их использования процессами, когда физические страницы соответствуют виртуальным страницам процессов. Учитывая, что система поддерживает вызов `mmap` (см. §5.2.7), подсистема Page Cache была вынуждена также хранить информацию о том, что та или иная физическая страница отображает данные, хранимые на диске (в файле), так что вносимые в неё изменения следует рано или поздно отразить на диске. До некоторых пор в Linux существовал отдельный от страничного кеша «буферный кеш», применявшийся при работе с файлами через вызовы `read` и `write` и поддерживавшийся драйверами блочных устройств; одни и те же данные — точнее, один и тот же фрагмент какого-нибудь файла — могли оказаться одновременно отражены и в страничном, и в буферном кеше, что порождало дополнительные сложности. Начиная с версий 2.4.*, в ядре Linux от этого дуализма начали избавляться, а в современных версиях ядра от буферного кеша никаких следов не осталось. Подсистема страничного кеша отвечает за распределение физической памяти, буферизацию дискового ввода-вывода и откачку/подкачку (своппинг), которая тоже, естественно, связана с дисковым вводом-выводом. Применение единой системы для учёта использования физических страниц, своппинга и буферизации файловых операций имеет ряд преимуществ; в частности, поскольку этой подсистеме известно количество свободных страниц, она может задействовать их для временного хранения информации, прочитанной с диска, а как только эти страницы кому-то потребуются, отдавать их; общий размер дисковых буферов, таким образом, становится динамическим — он всегда равен количеству свободной физической памяти.

Создатели FreeBSD пошли другим путём. Мы уже упоминали (см. стр. 212), что от блочных устройств в ядре FreeBSD отказались; буферизация файлового ввода-вывода там возложена на виртуальную файловую систему.

Наличие в ядре буферной памяти логично приводит нас к вопросу о том, в какой момент следует вернуть управление процессу, запросившему операцию вывода. Благодаря буферизации управление можно вернуть *сразу же*, записав данные в буфер и не дожидаясь каких-либо результатов. Можно, напротив, записать информацию в буфер, но управление пользовательскому процессу не возвращать, пока операция не будет физически завершена. Эти два подхода, которые мы уже упоминали в §5.2.3, называются соответственно ***асинхронным*** и ***синхронным***. Асинхронный подход более эффективен, поскольку позволяет процессу продолжать работу, не дожидаясь окончания медленной физической операции. С другой стороны, синхронный подход более надёжен, т. к. если во время выполнения физических действий с устройством произойдет ошибка, об этом можно будет сразу же сообщить процессу, тогда как при асинхронном построении вывода процесс может завершиться до того, как результаты операции станут известны. Выбор подхода зависит от стоящих перед нами задач. К примеру, дисковые файловые системы в Unix можно использовать как в синхронном, так и в асинхронном режиме; это указывается при монтировании файловой системы. Иногда используется подход, при котором диски, постоянно установленные в компьютер, работают в асинхронном режиме, а съёмные (дискеты, флеш-карты и пр.) — в синхронном.

Отметим, что **именно применением асинхронного режима обусловлена крайняя нежелательность выключения питания компьютера без подготовки системы к такому выключению.** Результатом неожиданного отключения питания может стать, как известно, нарушение целостности файловой системы и потеря данных, а иногда и полное разрушение файловой системы на логическом уровне, несмотря на то, что физически аппаратура может при этом нисколько не пострадать.

Иногда нужно точно удостовериться, что информация, переданная в операции вывода, была физически записана на диск. В качестве простейшего примера можно назвать операцию вывода на съёмный диск (флеш-брелок,

внешний жёсткий диск, старотипную дискету и т. п) перед физическим удалением этого диска из системы. Можно также привести пример с банкоматом, выдающим деньги: прежде чем выдать пачку банкнот, необходимо удостовериться, что эта операция реально зафиксирована в долговременной памяти, то есть что со счёта клиента соответствующая сумма будет списана. При работе в асинхронном режиме операционная система обычно предоставляет возможность принудительного сброса содержимого буферов на диск. Это называется ***принудительной синхронизацией***. В §5.2.3 мы обсуждали системные вызовы, позволяющие её затребовать — `fsync`, `fdatasync` и `sync`, а в §5.2.7 упоминали также вызов `msync`, работающий с отображениями, созданными с помощью `mmap`.

Отметим ещё один довольно печальный момент. Никакие системные вызовы не могут на 100% гарантировать, что записанные данные достигли поверхности диска. Дело в том, что контроллеры современных жёстких дисков, которые, как уже говорилось, расположены прямо на корпусе устройства и представляют с ним единое целое, обладают своей собственной кеш-памятью, которая может быть не подконтрольна программному обеспечению.

8.4.7. Планирование дисковых обменов

Потребность чтения с диска или записи на диск возникает у системы в двух основных случаях: когда соответствующее действие запрошено пользовательской задачей и когда пользовательский процесс обращается к странице виртуальной памяти, которая в настоящий момент откачана. Если в момент возникновения потребности в чтении или записи контроллер диска свободен, то система, в принципе, может начать нужную операцию немедленно; но контроллер с таким же успехом может быть занят выполнением предыдущей операции. В этом случае сформированный запрос на дисковую операцию помещается в очередь. К тому моменту, когда контроллер завершит выполнение текущего запроса и вновь окажется готов к работе, в очереди может находиться больше одного запроса на работу с тем же диском; на загруженных системах очередь запросов может вообще никогда не пустеть.

На первый взгляд наиболее естественным кажется решение выбирать из очереди те запросы, которые находятся в ней дольше всего, то есть считать очередь действительно очередью, работающей по принципу «кто первый пришёл, того первым обслужили» (*first come first served, FCFS*). Для виртуальных и сетевых дисков, а также дисковых устройств, не являющихся в физическом смысле дисками (например, flash-брелков и накопителей SSD) всё именно так и происходит: поступившие запросы драйвер обрабатывает в порядке поступления.

С устройствами, которые в буквальном смысле физически представляют собой диски, так тоже можно поступить, но если немного изменить подход к выбору следующего запроса из очереди, можно получить заметное повышение производительности системы. К примеру, если для обработки текущего запроса головку пришлось поставить на дорожку с номером 735, а после завершения обработки этого запроса в очереди обнаруживаются два других запроса, причём первый из них требует чтения с дорожки номер 114, а второй — записи на дорожку с номером 738, то обработка этих запросов в «естественном» порядке потребует сначала переместить головку с дорожки 735 на дорожку 114, а потом обратно и чуть дальше — на дорожку 738; но если запросы поменять местами, длинный путь головка проделает лишь один раз.

Принцип, в соответствии с которыми драйвер выбирает следующий запрос из очереди, называется ***стратегией планирования*** или — не совсем корректно — ***алгоритмом планирования***. Таких стратегий известно довольно много; одна из простейших, но эффективных — на каждом шаге выбирать запрос, выполнение которого требует наименьших затрат времени на перемещение головки. Английское название этой стратегии — *shortest seek time first, SSTF*; она относится к числу так называемых «жадных алгоритмов» (*greedy algorithms*), которые на каждом шаге из всех возможных действий выбирают позволяющее достичь наибольшей выгоды. Применив SSTF, можно добиться высокой производительности, но у этой стратегии есть фатальный недостаток — возможность уже знакомого нам ***ресурсного голодания*** (*starvation*). Если два или три процесса будут активно работать с файлами, находящимися в одной области диска, то процесс, чьи потребности относятся к другой области диска, может никогда не дождаться своей очереди, ведь до дорожек, нужных первым процессам, всегда будет ближе.

Другая широко известная стратегия получила название ***лифтового алгоритма*** (англ. *elevator algorithm*). В этом случае планировщик перемещает головку в одном направлении — например, от меньших номеров к бóльшим, по пути обслуживая все имеющиеся в очереди запросы, относящиеся к дорожкам, которые он проходит; дойдя до края — например, до запроса с наибольшим номером дорожки среди всех, которые были в очереди, и обслужив этот запрос, планировщик меняет направление движения головки; теперь она перемещается в сторону меньших номеров дорожек, опять-таки обслуживая все запросы, попадающиеся по пути, пока не будет обслужен запрос с наименьшим (среди имеющихся запросов) номером дорожки, и так далее. Название стратегии объясняется тем, что автоматические лифты работают примерно так же: кабина движется вверх, пока не достигнет наибольшего запрошенного этажа, затем движется вниз, пока не достигнет наименьшего этажа среди запрошенных. Проблема ресурсного голодания здесь не возникает, рано или поздно любой запрос будет обслужен.

В англоязычных источниках часто рассматривают в качестве двух разных стратегий нечто, именуемое SCAN и LOOK. Тот вариант, который только что ввели мы — это как раз LOOK; что касается SCAN, то он отличается тем, что головка перемещается в заданном направлении, пока не достигнет *края диска* (а не дорожки с максимальным/минимальным номером среди требуемых имеющимися запросами). В действительности такой вариант никогда не применяется, поскольку он заведомо хуже; кто первым придумал рассматривать SCAN как отдельную стратегию, непонятно, зато на этом примере мы можем наглядно видеть, сколь охотно люди готовы копировать ошибки других людей.

В качестве проблемы лифтового алгоритма обычно называют сравнительно большой разброс времени ожидания для поступившего запроса. Например, если на диске есть 1000 дорожек, головка начала движение в направлении уменьшения номеров и только что обслужила запрос на чтение или запись на дорожке 995, а в это время поступил запрос на работу с дорожкой 996, но в очереди также присутствует запрос на дорожку с номером 12, то время ожидания для запроса 996 составит почти двойное время перемещения головки диска от края до края. Отметим, что такое возможно только для дорожек, расположенных близко к краям диска; для дорожек в середине диска разброс времени ожидания оказывается вдвое меньше. Вероятность того, что очередной запрос будет вынужден долго ждать, существенно повышается при движении головки от конца диска к началу, ведь при чтении и записи больших объёмов данных возникает последовательность запросов, затрагивающих области диска, идущие подряд — естественно, в направлении возрастания номеров.

Простая модификация лифтового алгоритма позволяет сократить и выровнять разброс времени ожидания, хотя среднее время ожидания при этом не меняется. В изменённом варианте головка движется от меньших номеров дорожек к бóльшим, по пути обслуживая запросы; исполнив запрос с максимальным номером дорожки, головка возвращается к началу диска и снова проходит диск в направлении от меньших номеров к бóльшим. Эта стратегия, называемая ***круговым лифтовым алгоритмом***, может быть почти столь же эффективной (по среднему времени обслуживания запроса), как и обычный лифтовый алгоритм, благодаря тому, что большинство физических дисков способно перемещать головку на край диска с бóльшей скоростью, чем когда требуется найти конкретную дорожку. Возврат головки в начало диска с позиции в его конце происходит почти так же быстро, как перемещение на соседнюю дорожку. Разброс времени ожидания для этой стратегии вдвое меньше и не зависит от номера дорожки.

Существует множество модификаций описанных стратегий. Например, можно применять «жадную» стратегию не более чем для N запросов, после чего, даже если есть ещё запросы на близлежащие дорожки, очередной запрос выбирать с помощью лифтового алгоритма. В таком варианте эффективность близка к показателям чистой «жадной» стратегии, но при этом имеется очевидная гарантия недопущения ресурсного голодания.

Довольно интересен так называемый ***предчувствующий алгоритм*** (англ. *anticipation algorithm*). Если после обслуживания очередного запроса требуется большое перемещение головки, то прежде чем приступить к нему, планировщик делает небольшую паузу в надежде, что за это время поступит ещё запрос на ту же или следующую дорожку, и это во многих случаях действительно происходит (отсюда название).

Планирование обменов в реально существующих ядрах операционных систем в наши дни производится по ещё более изощрённым стратегиям. Так, в ядре Linux используется так называемый *Completely Fair Queuing (CFQ)*,

который разделяет запросы на синхронные и асинхронные, отдавая предпочтение синхронным (например, запросам на чтение), поскольку до их окончания соответствующий процесс не может продолжать выполнение; каждому процессу выделяется некое *время*, в течение которого планировщик дискового обмена готов выполнять его запросы, не обращая внимание на запросы, поступающие от других процессов. Подробности об устройстве этого планировщика читатель без труда найдёт в Интернете.

В §8.4.4 мы обсуждали виртуализацию геометрии физических дисков на уровне контроллера самого диска; с этим связано очевидное ограничение применимости (и осмысленности) изощрённых стратегий планирования дисковых обменов, ведь те воображаемые «головки», которыми позволяет «управлять» контроллер диска, теперь напоминают реально существующие механические головки только издали; попросту говоря, операционная система не знает и не может знать, какое расстояние в действительности придётся преодолеть физическим головкам диска, она даже не знает, *сколько* этих головок у диска реально есть. Создателям жёстких дисков приходится подбирать такое отображение физической геометрии на виртуальную, чтобы минимизация пути виртуальных головок в большинстве случаев уменьшала также и путь головок реальных; надо сказать, что они с этой задачей справляются, несмотря на её нетривиальность.

8.4.8. Виртуальная файловая система

Мы уже не раз упоминали ***виртуальную файловую систему*** — подсистему ядра, отвечающую за поддержку файлов как абстрактных объектов, обладающих определёнными свойствами. Как мы знаем, файл можно *открыть* (на запись, чтение или и то и другое), можно читать из него или записывать в него информацию, если это позволяет установленный режим работы, можно закрыть его, когда работа окончена. К открытому файлу можно, хотя и не всегда, применить отображение в память (вызов `mmap`), позиционирование (`lseek`), есть и другие операции. Файл в системах семейства Unix обладает информацией о владельце и группе, правами доступа, датой создания, модификации и последнего доступа, можно узнать количество установленных на него ***жёстких ссылок*** — то есть, попросту говоря, файловых имён, которые с данным файлом связаны. Более того, мы обсуждали, что вся служебная информация о файле, за исключением его имени, сосредоточена в ***индексном дескрипторе***, а для некоторых типов файлов только индексный дескриптор и есть, то есть файл может не занимать на диске никакого места, кроме одного дескриптора в массиве индексных дескрипторов. Коль скоро мы вспомнили про типы файлов, можно заодно отметить, что в unix-системах их поддерживается семь: обычные файлы, директории (каталоги), символические ссылки, каналы (FIFO), сокеты и два типа файлов устройств — потоковые и блочные. Имена файлов хранятся в директориях, один файл может иметь больше одного имени (жёсткой ссылки), причём как в разных директориях, так и в одной; и так далее. Свойства файлов в unix-системах и операции, которые над ними можно производить, мы уже обсуждали (см. главу 5.2). Все эти свойства и операции как раз и образуют ***абстракцию файла***, характерную именно для систем семейства Unix.

Напомним, что термин ***файловая система*** может обозначать две совершенно разные, хотя и связанные между собой сущности: структуру данных на диске и набор программных функций, которые с этой структурой данных умеют работать. Программные функции, которые в случае Unix обычно находятся в ядре системы (хотя и не всегда), определяют то, *как* должны быть организованы данные на диске. Надо сказать, что unix-системы работают с довольно разными файловыми системами; для одного только Linux характерны системы `ext2`, `ext3`, `ext4`, `ReiserFS`/`reiser4`, а кроме них Linux поддерживает также `minix`, `xia`, `jfs`, `xfs` и другие. Для FreeBSD основной файловой системой считается `ffs` (она же `ufs`), несколько реже используется `zfs`, пришедшая из системы Solaris. Для обеих этих файловых систем есть поддержка и в Linux, а FreeBSD можно при желании «научить» работать с `ext2-4` и другими файловыми системами, используемыми в Linux. Все перечисленные файловые системы поддерживают общий набор понятий, характерных для файловых систем семейства Unix. Файлы в них основаны на индексных дескрипторах, имеются файлы семи известных нам типов, в том числе, что особенно важно, директории и символические ссылки; индексные дескрипторы содержат привычные для Unix права доступа, идентификаторы владельца и группы, информацию о датах и прочее.

Работающая операционная система использует один из своих дисков в роли *корневого*; организованная на нём файловая система считается ***корневой файловой системой***, её корневой каталог совпадает с общесистемным — каталогом «/». Корневые каталоги остальных файловых систем отображаются в общее дерево каталогов в так называемых ***точках монтирования*** (англ. *mount points*) — директориях, которые на момент подключения (***монтирования***) новой файловой системы уже есть в общем дереве. Как правило, в роли точек монтирования используют пустые директории; если там ранее содержались файлы, то они никуда не денутся, но видно их уже не будет — до тех пор, пока новая система не будет размонтирована.

Общее дерево каталогов операционной системы может состоять из файловых систем разных типов, так что ядру придётся задействовать несколько разных наборов функций, причём делать это прозрачно для пользователя в том смысле, что работа с файлами с точки зрения пользовательских программ протекает совершенно одинаково вне зависимости от того, какой тип имеет файловая система.

Дело резко осложняется тем, что ядро вынуждено в некоторых случаях поддерживать диски, организованные по правилам других операционных систем. Как Linux, так и FreeBSD (и другие unix-системы) без особых проблем справляются с носителями, отформатированными на системах линейки Windows, для чего их ядра включают поддержку таких файловых систем, как `msdos`/`fat`, `fat32`, `vfat` и `NTFS`. Обе операционные системы поддерживают также файловую систему `iso9660`, используемую для оптических дисков (CD и DVD). Сложность здесь в том, что во всех этих файловых системах нет никаких индексных дескрипторов, права доступа либо отсутствуют, как в `fat`, либо организованы совершенно не так, как в unix-системах (`NTFS`).

Картина становится ещё более интересной, если упомянуть файловые системы, не предполагающие использования диска. Речь в данном случае идёт не о том, что файловая система создаётся на виртуальном диске в памяти, как это обсуждалось в §8.4.3; в этом случае всё как раз достаточно просто, ведь с точки зрения аппаратно-независимых компонентов ядра виртуальный диск почти ничем от реального не отличается. Более хитро устроены файловые системы, вообще не обращающиеся ни к каким дискам — ни реальным, ни виртуальным; можно сказать, что такая файловая система существует только в одной своей ипостаси — программной, а дисковых структур данных нет за неимением диска. Тем не менее, в такой файловой системе располагаются файлы, их можно открывать, читать, записывать и т. п.

Именно такова, например, файловая система `tmpfs`, поддерживаемая большинством современных unix-систем и специально предназначенная для создания временных файлов, не предполагающих длительного хранения. Потеря временных файлов при, скажем, аварии питания — совершенно не проблема, поскольку никто обычно и не ожидает их сохранности; при этом расположение хранилища временных файлов в оперативной памяти, а не на «настоящем» физическом диске, может заметно улучшить показатели скорости работы системы. Система `tmpfs` существенно удобнее для таких целей, нежели обычная файловая система на виртуальном диске. Дело в том, что при создании файловой системы на виртуальном диске нужно указать и зафиксировать его размер, поскольку структуры данных практически любой дисковой файловой системы существенно зависят от размера диска; `tmpfs` этого ограничения лишена, на неё можно записывать файлы до тех пор, пока не исчерпается виртуальная память, то есть вся физическая память и все возможности своппинга. При удалении файлов с такой файловой системы память немедленно освобождается и может быть использована ядром для других целей — например, выделена какому-нибудь процессу.

Ещё больше путаницы возникает благодаря файловым системам, которые мы будем называть ***искусственными***, хотя этот термин и не вполне удачен. В английском оригинале их обычно называют *pseudo file systems*, но прямой перевод на русский оказывается невозможен из-за того, что «псевдо» в русском языке — приставка, а не отдельное слово, и применить её не к слову, а к словосочетанию (в данном случае — к термину «файловая система») не получается. Несколько реже в англоязычных источниках та же самая сущность обозначается термином *synthetic file system*, чем мы и воспользуемся, переведя слово *synthetic* как «искусственный».

Искусственные файловые системы не имеют отношения не только к дискам, но и вообще к хранению информации. Выглядят они более-менее как

обычные файловые системы с деревом каталогов и файлами, но все эти каталоги и файлы на самом деле представляют собой отражение состояния каких-то подсистем внутри ядра. Читая эти файлы, мы можем узнать, что творится в ядре; запись в некоторые из них позволяет менять настройки ядра и режимы работы отдельных его подсистем. Иначе говоря, такая файловая система — это *интерфейс* для управления ядром, который только выглядит как файловое поддерево. Такой стиль построения интерфейсов к подсистемам ядра имеет несомненное достоинство: не нужно добавлять лишние системные вызовы, которых и так очень много.

Классическим примером искусственной файловой системы может служить `procfs`, которая обычно монтируется на директорию `/proc`. Для каждого процесса, существующего в системе, появляется поддиректория, соответствующая его `pid`'у; например, для процесса `2715` это будет директория `/proc/2715`. Она содержит несколько десятков «файлов», отражающих различные аспекты работы процесса. Так, `/proc/2715/exe` — это символическая ссылка на исполняемый файл, `/proc/2715/cwd` — на текущую директорию (*current working directory*) процесса `2715`, а поддиректория `/proc/2715/fd` содержит символические ссылки на имеющиеся у процесса открытые потоки ввода-вывода; файл `/proc/2715/cmdline` «содержит» аргументы его командной строки, разделённые нулевым байтом; `/proc/2715/maps`, если его прочитать (например, выдать на экран с помощью `cat`), выглядит как текстовый файл, содержащий информацию об областях виртуальной памяти процесса. Между прочим, настоятельно рекомендуем поэкспериментировать с этими файлами; особенно интересным может оказаться анализ выполнения своей собственной программы, написанной на языке ассемблера или на Си.

Помимо информации о процессах, `/proc` содержит также целый ряд других псевдофайлов. Например, файл `/proc/sys/net/ipv4/ip_forward` может «содержать» `0` или `1`; `0` означает, что ядро не передаёт пакеты между сетевыми интерфейсами, то есть система не работает в роли маршрутизатора, `1` показывает, что передача пакетов включена. Более того, чтобы её включить или выключить, надо занести в этот файл соответствующее значение, как в обычный файл, например:

```
echo "1" > /proc/sys/net/ipv4/ip_forward
```

В ОС Linux традиционно многие системные программы опираются на информацию из `/proc`; так, команда `ps` всю нужную ей информацию получает именно оттуда. В мире FreeBSD отношение к этой искусственной файловой системе иное: для работы собственных утилит FreeBSD она не нужна, так что не на каждом компьютере её монтируют, но поддержка для неё в ядре есть — на случай запуска программ, пришедших из Linux.

Ещё один пример искусственной файловой системы в Linux — `sysfs`, которая обычно монтируется на директорию `/sys`. Ввели её сравнительно недавно, чтобы разгрузить `/proc`, переполненную информацией, не имеющей отношения к процессам. Исследуя «содержимое» `/sys`, можно узнать, какие устройства подключены к вашему компьютеру (в том числе, например, к портам USB), какие модули загружены в ядро и многое другое. В ОС FreeBSD аналога этой системы нет.

Наконец, файловая система `devfs` позволяет сэкономить индексные дескрипторы на физических дисках, не создавая файлы устройств в директории `/dev`. Вместо этого `/dev` используется как точка монтирования для `devfs`, которая уже делает вид, что содержит файлы для всех устройств, поддерживаемых ядром. Создатели многих современных дистрибутивов Linux предпочитают идти другим путём: на `/dev` монтируется обычная `tmpfs`, которую наполняет файлами работающий в системе демон `udev` или его аналог; в этом случае обычно на `/dev/pts` монтируется другая искусственная файловая система, `devpts`, содержащая только файлы для главных и подчинённых устройств псевдотерминалов.

Рассказ о файловых системах, не подразумевающих использования дискового устройства, был бы неполным без упоминания `NFS` (*network file system*) и `smbfs`, которые представляют в виде локальных файловых деревьев диски, расположенные на других компьютерах. `NFS` работает по протоколу, который так и называется NFS и изначально появился в системах семейства Unix, тогда как `smbfs` использует протокол SMB/CIFS, пришедший из Windows, что означает неизбежное применение «чужой» абстракции файла (другое устройство прав доступа и т. п.).

Всё это великолепие должно обслуживаться единым набором системных вызовов, большинство которых нам уже знакомо. Для файловых систем, характерных для семейства Unix, всё ещё сравнительно просто, ведь они различаются в основном только форматом представления информации о файлах на диске, а сущности вроде индексного дескриптора, суперблока (дискового блока, содержащего «глобальную» информацию о файловой системе), директории и прочего остаются общими для них всех. Для «чужих» файловых систем, таких как `vfat`, `NTFS` или `smbfs`, абстракции, ожидаемые unix-системой, приходится эмулировать; для искусственных файловых систем эмулировать приходится вообще всё, что в них есть. За обработку особенностей каждой файловой системы отвечает соответствующий модуль ядра, который, как мы уже неоднократно отмечали, тоже называется файловой системой; эти модули, как обычно в подобных случаях, экспортируют некий фиксированный набор интерфейсных функций (точек входа) — примерно так же, как это делают драйверы.

Обязанности координатора всех файловых систем выполняет виртуальная файловая система; она отвечает за монтирование и размонтирование отдельных файловых систем, за формирование и обслуживание общего дерева каталогов, за выбор подходящего модуля для поддержки каждой конкретной файловой системы и за перевод запросов с языка системных вызовов на язык точек входа модулей файловых систем. С другой стороны, именно виртуальная файловая система обрабатывает системные вызовы файлового ввода-вывода, создаёт и поддерживает потоки ввода-вывода (как объекты ядра).

8.4.9. Файловая система на диске

Структуры данных, создаваемые на диске для поддержки файловой абстракции, могут быть очень разными. Даже файловые системы, выдержанные в стиле Unix (то есть основанные на индексных дескрипторах), друг от друга сильно отличаются; но выделить некие общие принципы всё же можно.

Для построения файловой системы физические сектора диска (имеющие обычно размер 512 байт) группируются в ***блоки***, размер которых зависит от файловой системы; более того, один формат файловой системы может допускать разные размеры блоков, в этом случае размер задаётся при форматировании диска. Каждый блок имеет свой номер. Обычно самый первый блок на диске содержит глобальную информацию обо всей файловой системе (так называемый ***суперблок***); как правило, в нескольких местах диска располагают копии суперблока на случай его потери.

Часть блоков отводится под ***области индексных дескрипторов***; таких областей может быть от нескольких десятков до нескольких тысяч. В каждый блок области помещается несколько индексных дескрипторов (всегда степень двойки); в рамках одной файловой системы дескрипторы имеют сквозную нумерацию. Например, системы `ext2`–`ext4`, применяемые в ОС Linux, делят блоки на группы блоков, каждая такая группа имеет свою копию суперблока, свой массив дескрипторов и свою карту свободных блоков — массив данных, занимающий ровно один блок, в котором каждый бит соответствует блоку из этой же группы, ноль означает, что блок свободен, единица — что он занят.

Размер индексного дескриптора тоже различается от системы к системе; в наши дни типичный размер — 256 байт. Как мы знаем, индексный дескриптор хранит всю информацию о файле, кроме его имени и содержимого; имена файлов хранятся в директориях (напомним, что это файлы специального типа), а содержимое файлов располагается в блоках, не занятых служебной информацией. Если файлу принадлежат блоки диска, индексный дескриптор содержит информацию о размещении, то есть о номерах этих блоков. Обычно структура индексного дескриптора предусматривает массив для хранения номеров блоков, причём первые N элементов массива содержат непосредственно номера блоков, в которых размещено содержимое файла; в конце массива три элемента содержат номера ***косвенных блоков*** (англ. *indirect block*) — таких, которые сами содержат номера блоков, причём косвенные блоки различаются по уровням: косвенный блок первого уровня содержит номера обычных блоков, косвенный блок второго уровня — номера косвенных блоков первого уровня, косвенный блок третьего уровня — номера косвенных блоков второго уровня. Элементы массива в индексном дескрипторе содержат номер одного косвенного блока каждого уровня.

Например, в системе `ext2` в индексном дескрипторе массив номеров блоков имеет 15 элементов, из которых первые 12 содержат номера простых блоков; если размер блока составляет 4 KB (это самый типичный случай), то файлы до 48 KB (12×4) обходятся без косвенных блоков. Для файлов

бóльшего размера приходится завести косвенный блок первого уровня, который (при размере номера блока в 4 байта) может содержать до 1024 номеров блоков; это позволяет организовать файлы размером до 1036 блоков, т. е. до 4144 KB. Когда не хватает и этого, система заводит косвенный блок второго уровня; это позволяет применить ещё 1024 косвенных блока первого уровня, а общее число блоков, занимаемых файлом (не считая косвенных), увеличить до $1024 \times 1024 + 1024 + 12 = 1049612$; сам файл при этом может увеличиться до 4 GB с небольшим. Если не хватит и этого, в бой будет брошена тяжёлая артиллерия — косвенный блок третьего уровня; после этого резервов больше не остаётся, но максимальный возможный размер файла вырастает до $(1024^3 + 1024^2 + 1024 + 12) \times 4096$ байт, т. е. свыше 4 TB (терабайт). Впрочем, конкретно `ext2` имеет другое ограничение — 32-разрядное число в структуре индексного дескриптора, означающее размер в 512-байтных секторах, так что вырастить отдельно взятый файл больше чем до 2 TB не получится.

8.4.10. Шина, кеш и DMA

Материал этого параграфа не имеет прямого отношения к операционным системам, речь здесь пойдёт о возможностях аппаратуры. При соблюдении «справочной» логики изложения уместнее было бы рассказать об этом в первом томе — в главе 1.1.2, где шёл разговор об основах устройства компьютера, или в части III, где говорилось о возможностях процессора; но наша книга — не справочник. Во вводной части первого тома объяснять тонкости, связанные с работой шины, было явно рано, ведь она написана в расчёте на читателя, пока что не знающего о компьютерах вообще ничего. В части, посвящённой процессору, мы намеренно не полезли в архитектурные дебри. Если там рассказывать о кеш-памяти, DMA и прочих изысках из этой области, то пришлось бы перетащить туда же, например, MMU с моделями виртуальной памяти; но первый том и без того получился довольно объёмным, к тому же хотелось дать читателю возможность быстрее проскочить ассемблер и приступить к изучению Си.

Кеш-память поэтому лишь вскользь упомянута в первом томе; прямой доступ контроллеров устройств к оперативной памяти (*direct memory access, DMA*) мы пока не упоминали вообще. Такое положение вещей вполне могло нас устраивать до поры до времени, но не сейчас, когда в своих изысканиях мы спустились на уровень взаимодействия драйверов с аппаратурой. В любых (на что-то годных) источниках на эту тему, которые читатель пожелает изучить в дополнение к нашей книге, обе сущности наверняка будут постоянно упоминаться, причём, скорее всего, без пояснений — в предположении, что читатель знает, о чём идёт речь.

Углубляться в тонкости устройства кеш-памяти и DMA мы не будем, для этого есть специальная (в основном справочная) литература, к тому же в этой области всё довольно быстро меняется и сильно зависит от конкретной аппаратной платформы. В частности, мы не станем рассматривать алгоритмы вытеснения страниц кеша и принципы поддержки когерентности кешей в многопроцессорных машинах — это тоже очень долгий разговор, без знания этих тонкостей вполне можно обойтись, если же вдруг станет интересно — то ни в коем случае не отказывайте себе в знаниях, тем более что с поиском информации в наше время никаких проблем нет, надо только точно знать, что искать.

Приблизительно до середины 1980-х годов процессоры были довольно медленными, так что шина и микросхемы оперативной памяти, будучи устройствами гораздо более простыми, легко «поспевали» за темпом работы процессора. По мере развития технологий наметился определённый дисбаланс в скоростях; дело в том, что скорость работы цифровой электроники невозможно наращивать бесконечно, этому мешает скорость света. Если сравнить линейные размеры кристалла микропроцессора, диагональ которого редко превышает 3 см, с длиной дорожек шины, мы обнаружим разницу более чем на порядок; приняв во внимание, что даже непосредственно на кристалле схема самого процессора (точнее, вычислительного ядра[19], которое нас сейчас и интересует) обычно занимает лишь малую часть площади, мы получим ещё один порядок разницы.

Если принять для ровного счёта длину шины равной 25 см, получится, что за один такт её работы электрический сигнал должен преодолеть 0,5 м, чтобы успеть дойти от запрашивающего к запрашиваемому и обратно, как это и происходит при операции чтения — неважно, из памяти или порта ввода. Скорость света, как мы знаем, составляет приблизительно 300 тысяч км/с, или 300 миллионов м/с, или 600 миллионов «полуметров» в секунду; следовательно, для шины длиной 25 см тактовая частота 600 МГц представляет абсолютный теоретический предел. В реальной жизни такт должен длиться дольше, оставляя запас времени, чтобы запрашиваемый успел среагировать на запрос, а запрашивающий — «снять» с шины пришедшую информацию. С учётом всего этого можно заметить, что шины современных материнских плат, работающие на частотах около 450 МГц, «разгонять» совершенно некуда. Тактовая частота современных процессоров достигает 3 ГГц, что почти в семь раз больше. Разгонять их ещё больше мешают физические ограничения — всё та же скорость света и проблемы с отводом тепла; но и имеющаяся разница достаточна, чтобы процессор, коль скоро ему придётся всё время взаимодействовать с оперативной памятью, бóльшую часть времени простаивал.

Если рассматривать хорошо знакомую нам линейку процессоров x86 (см. т. 1, §3.1.3), то для процессоров, предшествовавших 80286, проблема дисбаланса скоростей не проявлялась вообще. На 80286 эта проблема уже возникла: выполняя операцию чтения из памяти, процессор был вынужден пропускать такт, давая микросхемам памяти и шине достаточно времени, чтобы они успели сработать. Если бы всё по-прежнему работало так, как в те времена, современным процессорам приходилось бы пропускать больше десяти тактов. Появившуюся проблему нужно было как-то решать, поэтому на некоторых (хотя и не на всех) материнских платах, предназначенных для i386, появились специальные микросхемы памяти, расположенные рядом с процессором. Скорость их работы ненамного превосходила скорость обычной памяти, но для обращения к этим микросхемам не нужно было задействовать шину.

В общих чертах работа с кешем протекает так. Вся физическая память делится на блоки одинакового размера. В кеш-памяти содержатся *копии* некоторых блоков физической памяти вместе с *метками*, показывающими, копия какого блока сохранена в этой области кеша, а при некоторых способах организации кеша — ещё и информация о том, когда в последний раз к этому блоку происходило обращение. Процессор, прежде чем обратиться к памяти, должен проверить, не содержится ли в кеше копия нужного блока; если нет — весь блок целиком копируется в кеш, при этом из кеша может быть удалён какой-то другой блок, чтобы расчистить место. Обмен с памятью по шине теперь протекает в основном целыми блоками, что позволяет его ускорить, увеличив разрядность шины данных.

К выполнению операций записи есть два подхода. Можно записывать информацию в ячейки кеша, помечая соответствующий блок как «грязный», и при вытеснении этого блока из кеша записывать его в память целиком; можно, напротив, каждую операцию записи выполнять немедленно, то есть записывать сразу и в кеш, и в память, благо процессору совершенно не нужно дожидаться окончания операции записи. Эти подходы называются ***отложенная запись*** и ***сквозная запись*** (англ. *write-back* и *write-through*); кстати, такие же термины часто применяются при обсуждении буферизации ввода-вывода.

Отдельного внимания заслуживает вопрос, как реализовать метки, как искать в кеше нужный блок и как принимать решение, какой из блоков должен быть заменён, когда места в кеше больше нет. Подходов к этой проблеме существует достаточно много, подробное их рассмотрение оставим за рамками книги; отметим только, что самый простой и довольно остроумный способ состоит в том, чтобы каждый блок физической памяти мог отображаться только на один из блоков кеша: если кеш может содержать одновременно M блоков, то блок памяти с номером n может быть отображён только в блок кеша с номером, равным остатку от деления n на M. Поскольку число M обычно выбирается равным 2^k, деление с остатком сводится к выделению k младших битов. Такой подход может показаться неэффективным, но стоит учесть, что при этом не нужно ни запоминать, к каким блокам производились обращения, ни искать что бы то ни было где бы то ни было (поскольку блок памяти либо находится в «своей» части кеша, либо нет).

Начиная с процессоров i486, кеш стали размещать на одном кристалле с процессором. Конечно, этот кристалл не резиновый, поэтому много кеш-памяти на нём поместиться не может; процессоры i486 имели 8 KB кеш-памяти на одном кристалле с самим процессором, и, кроме этого, некоторое количество кеша (чаще всего 256 KB) предусматривалось на материнской плате рядом с процессором. Такая организация кеша называется многоуровневой; первый уровень кеша (L1) располагается ближе всего к процессору и имеет самую высокую скорость работы, но при этом обладает незначительным объёмом, по мере роста номера уровня объём кеша растёт, скорость падает. Кеш последнего уровня (в наши дни — обычно четвёртого, L4) представляет собой отдельную микросхему, кеши остальных уровней располагаются на одном кристалле с процессором. Многоядерные процессоры обычно предусматривают отдельный кеш L1 для каждого ядра и общий для всех ядер кеш L3, а L2 может быть как общий, так и свой у каждого ядра — это зависит от модели процессора. Например, четырёхъядерный Intel Core i7

[19] См. сноску 2 на стр. 196.

имеет на каждое ядро по 32 KB кеша L1 отдельно для команд и для данных, смешанный (для инструкций и данных) кеш L2 размером 256 KB (тоже свой для каждого из ядер) и общий для всех ядер кеш L3 на 8 MB.

Применение кеш-памяти имеет любопытный побочный эффект: если всё происходит правильно, то процессор будет сравнительно редко обращаться к общей шине. Время, когда шина не нужна процессору, можно использовать, если вспомнить, что достаточно большие объёмы данных часто приходится перекачивать между памятью и контроллерами внешних устройств. Раньше этим приходилось заниматься центральному процессору, и мы описывали именно этот вариант взаимодействия между драйвером и контроллером; например, при чтении с диска драйвер задавал контроллеру выполнение операции и на этом заканчивал работу, освобождая процессор до тех пор, пока не произойдёт прерывание от контроллера; вновь получив управление, на сей раз от обработчика прерываний, драйвер выполнял копирование считанных с диска данных из буфера контроллера в оперативную память. Но благодаря использованию кеш-памяти шина бóльшую часть времени всё равно свободна, а копирование данных — операция достаточно примитивная, и задействовать для неё центральный процессор не обязательно. Из этого логично вытекает идея возложить копирование данных на кого-то ещё, причём так, чтобы это происходило, когда центральному процессору шина не требуется.

Позволить управлять шиной кому-то кроме процессора пытались и раньше, но по противоположной причине: ранние процессоры работали не быстрее, а, наоборот, медленнее шины. По своему устройству и шина, и микросхемы памяти намного проще процессора, а до ограничений из-за скорости света в те времена было ещё очень далеко, так что скорость работы более простых компонентов системы росла быстрее.

Уже в составе IBM PC, основанного на процессоре Intel 8088, присутствовал ***контроллер прямого доступа к памяти*** (контроллер DMA) — микросхема Intel 8237. Это было сравнительно простое устройство, подключённое к общей шине и имеющее свои собственные порты ввода, позволявшие сформулировать ему задание: откуда копировать информацию, куда и в каком количестве. Пусть нам, к примеру, нужно выдать некие данные в потоковое устройство, то есть записать порцию информации, расположенную в памяти, в однобайтовый порт вывода. Мы можем с помощью обычного цикла последовательно извлекать содержимое из каждой ячейки (например, командой `MOV`) и записывать полученное значение в порт (командой `OUT`), после чего увеличивать адрес ячейки, уменьшать счётчик, и если он всё ещё больше нуля, возвращаться к началу цикла. В исполнении ранних процессоров всё это происходило, как мы уже отметили, довольно медленно.

Наличие контроллера DMA позволяло драйверу устройства поступить иначе: занести в порты контроллера DMA адрес начала области памяти, содержащей данные, и количество этих данных, после чего дать ему команду на проведение операции вывода. Шина ISA, использовавшаяся в компьютерах линейки IBM PC вплоть до некоторых компьютеров на процессоре Pentium-1, предусматривала (в составе шины управления) специальные дорожки для управления прямым доступом к памяти, составлявшие так называемые *каналы DMA*, по две дорожки на каждый канал (запрос DMA — *DRQ, DMA request* и подтверждение DMA — *DACK, DMA acknowledged*). В ранних версиях шины ISA таких каналов было три[20], в более поздних — семь. Контроллеры внешних устройств, поддерживавшие прямой доступ к памяти, выдавали очередную порцию (байт) данных в шину (или, наоборот, считывали порцию данных) по сигналу «подтверждение DMA» с тем номером, который им был приписан. Это позволяло контроллеру DMA не обрабатывать данные самому: он давал порту ввода сигнал «чтение» через канал DMA, а ячейке памяти — обычный сигнал «запись», сопровождаемый выставлением её адреса на шине адресов, в результате чего порт устанавливал шину данных в состояние, соответствующее читаемым из него данным, а ячейка памяти снимала эту информацию с шины данных и запоминала, и всё это за один такт шины; естественно, можно было передать данные и в обратном направлении. Такой режим передачи данных по-английски называется *fly-by mode*. Всё, что требуется при этом от контроллера DMA — это после каждого такта увеличивать адрес ячейки памяти, в нужные моменты выдавать управляющие сигналы в шину управления и подсчитывать, сколько ещё осталось передать байтов[21], чтобы вовремя остановиться.

Передача блока данных между контроллером периферийного устройства и памятью в таком режиме происходила в несколько раз быстрее, чем если то же самое делать силами центрального процессора, что уже само по себе оправдывало существование контроллера DMA; но, кроме передачи блоков данных за один раз, контроллер DMA умел работать и по-другому. Например, внешнее устройство могло отдавать данные через порт со скоростью, меньшей скорости шины; в этом случае по мере готовности очередных байтов данных контроллер внешнего устройства взводил на шине сигнал DRQ; получив этот сигнал, контроллер DMA отвечал сигналом DACK и производил один цикл передачи. Благодаря наличию независимых каналов DMA в таком режиме можно было обслуживать несколько устройств одновременно, позволяя при этом работать и центральному процессору тоже — если шина была занята контроллером DMA, а процессору нужно было выполнить операцию с шиной, он попросту пропускал нужное количество тактов или, наоборот, заставлял контроллер DMA подождать.

Кроме режима *fly-by* контроллер DMA Intel 8237 поддерживал также и режим «получить-записать», требовавший двух тактов на пересылку одного байта. В этом режиме его можно было использовать в том числе и для копирования данных между двумя областями памяти. В компьютерах линейки x86 этот режим никогда не применялся.

С появлением новой шины, получившей обозначение PCI, подход к прямому доступу к памяти изменился. Контроллер DMA уступил место другой микросхеме — ***арбитру шины***, который не участвует в передаче данных; его функция сводится к тому, чтобы устройства, подключённые к шине, могли договориться, кто из них сейчас играет роль «главного» (*bus master*). К этому времени процессоры обзавелись кеш-памятью, так что доступ к шине стал им требоваться реже. Что касается обмена данными между памятью и внешними устройствами, то управление таким обменом в режиме прямого доступа (DMA) взяли на себя контроллеры внешних устройств. Теперь, например, драйвер жёсткого диска сообщает контроллеру, в какую область памяти нужно записать данные, считанные с диска, или из какой области памяти нужно взять данные, чтобы затем записать их на диск, и на этом успокаивается. Контроллер жёсткого диска, получив шину (через арбитра) в своё единоличное распоряжение, сам производит запись данных в память или чтение их из памяти — точно так же, как это умеет делать центральный процессор.

Довольно интересную проблему порождает здесь использование виртуальной адресации памяти. Если с диска нужно будет прочитать, скажем, мегабайт данных, то операционной системе, скорее всего, не удастся найти 256 свободных страниц, идущих *подряд*; между тем контроллеры периферийных устройств не знают никаких адресов, кроме физических, так что область оперативной памяти, участвующая в работе DMA, должна быть непрерывной. Это несколько ограничивает возможности DMA и вынуждает ядра операционных систем по возможности избегать фрагментации физической памяти.

[20] Хотя контроллер DMA поддерживал четыре канала, один из них — нулевой — в совсем ранних версиях архитектуры использовался для обновления содержимого микросхем памяти; в расширенной версии шины ISA он стал использоваться для каскадирования со вторым контроллером DMA.

[21] На IBM PC это действительно были байты, а на IBM PC AT, построенной на основе процессора 80286, за один приём передавалось 16 бит, несмотря на то, что контроллер DMA остался тот же самый, и он был восьмибитным (хотя их стало два ради увеличения числа каналов DMA). Для программирования самого контроллера DMA этих восьми бит хватало, а разрядность проходящих по шине данных его никоим образом не трогала.

Приложение 1. Компилятор gcc

Компиляторы семейства GCC (Gnu Compiler Collection) представляют собой утилиты командной строки, т. е. все необходимые действия задаются при запуске компилятора и выполняются уже без непосредственного участия пользователя. Это, в частности, позволяет использовать компилятор в командных файлах (скриптах).

Команда `gcc` предназначена для компиляции программ на языке Си, а команда `g++` — на языке Си++. На самом деле используется один и тот же компилятор; оба имени обычно представляют собой символические ссылки на исполняемый файл компилятора. Поведение компилятора зависит от того, по какому имени его вызвали; прежде всего различие выражается в наборе стандартных библиотек, подключаемых по умолчанию при сборке исполняемого файла. Имена файлов, подлежащих компиляции и сборке, компилятор принимает через параметры командной строки. Кроме того, компилятор воспринимает большое количество разнообразных ключей, часть из которых мы уже знаем; вот несколько более широкий их список:

- `-o` <*имя-файла*> задаёт имя исполняемого файла, в который будет записан результат компиляции; если не указать этот флаг, результат компиляции будет помещен в файл `a.out`;
- `-Wall` (от слов *warnings all*) приказывает компилятору выдавать все разумные предупреждения; **обязательно всегда указывайте этот флаг**, он поможет вам сэкономить немало времени и нервов;
- `-ggdb` и `-g` используются для включения в результирующие файлы отладочной информации, т. е. информации, используемой отладчиком, включая имена переменных и функций, номера строк исходных файлов и т. п.; флаг `-ggdb` снабжает файлы расширенной отладочной информацией, понятной только отладчику `gdb`; если вам кажется, что что-то не в порядке с отладчиком, попробуйте использовать `-g`;
- `-c` указывает компилятору, что результатом должна быть не вся программа, а отдельный её модуль; в этом случае имя файла для объектного модуля можно не задавать, оно будет сгенерировано автоматически заменой суффикса на `.o`;
- `-O`*n* задаёт уровень оптимизации; *n*=0 означает отсутствие оптимизации (значение по умолчанию); для получения более эффективного объектного кода рекомендуется использовать флаг `-O2`; кроме того, поддерживается оптимизация *размера кода* (в противоположность скорости выполнения), этот режим включается флагом `-Os`; учтите, что оптимизация может затруднить работу с отладчиком, а уровни выше второго могут привести к неприятным неожиданностям, поэтому на период отладки оптимизацию лучше отключить;
- `-ansi` приказывает компилятору работать в соответствии со стандартом ANSI C;
- `-pedantic` запрещает все расширения языка, не входящие в выбранный стандарт;
- `-E` останавливает компилятор после проведения стадии макропроцессирования; результат макропроцессирования выдаётся на стандартный вывод; этот режим работы может быть полезен, если ваши макроопределения повели себя не так, как вы ожидали, и хочется понять, что на самом деле происходит;
- `-S` останавливает процесс компиляции после стадии генерации ассемблерного кода; получившийся текст на языке ассемблера записывается в файл с суффиксом `.S`; по умолчанию генерируемый ассемблерный код имеет синтаксис AT&T, что сильно отличается от изучавшегося нами языка ассемблера NASM, следующего синтаксическим традициям Intel; это можно исправить, добавив опцию `-masm=intel`, что несколько снизит остроту проблемы, хотя элементы непривычности всё-таки останутся; следует также принять во внимание, что в современных условиях генерируемый код, скорее всего, будет соответствовать 64-битной системе команд, тогда как мы изучали её 32-битный вариант; но разобраться в том, что получится, вполне реально, а посмотреть, что конкретно сделал компилятор из вашего исходника, иногда бывает крайне любопытно, особенно на разных уровнях оптимизации;
- `-D` позволяет из командной строки (т. е. без изменения исходных файлов) определить в программе некий макросимвол; это полезно, если в вашей программе используются директивы условной компиляции и требуется, не изменяя исходных файлов, быстро откомпилировать альтернативную версию программы; например, `-DDEBUG=2` имеет такой же эффект, какой дала бы директива «`#define DEBUG 2`» в начале исходного файла; значение символа задавать не обязательно: флаг `-DDEBUG` даст тот же эффект, что и «`#define DEBUG`», что осмысленно, если в программе вы используете `#ifdef` и `#ifndef` без проверки значения символа;
- `-include` включает в вашу программу текстовый файл, как если бы он был подключён директивой `#include`;
- `-I` добавляет к рассмотрению директорию, в которой следует искать файлы, подключаемые с помощью `#include`; этот флаг оказывает влияние как на `#include "..."`, так и на `#include <...>`;
- `-isystem` делает примерно то же самое, но добавленная директория рассматривается как «системная»; это сравнительно недавнее новшество в компиляторах семейства `gcc`, которое может, в частности, оказаться важным при использовании флага `-MM`;
- `-l` позволяет подключить к программе библиотеку функций; так, если в вашей программе используются математические функции (`sin`, `exp` и другие), необходимо при компиляции задать ключ `-lm`; в некоторых вариантах ОС Unix (например, в SunOS/Solaris) при использовании сокетов вам понадобится ключ `-lnsl`; учтите, что, начиная с версий `gcc` 5.*, этот флаг должен обязательно располагаться *после* имён исходных или объектных файлов, нуждающихся в соответствующей библиотеке — точно так же, как это требуется для редактора связей (линкера; подробности см. в первом томе, §3.7.5;
- `-L` добавляет к рассмотрению директорию, в которой следует искать файлы библиотек, подключённых с помощью `-l`;
- `-MM` анализирует заданные исходные файлы и строит информацию об их взаимозависимостях; о том, как использовать полученную информацию, рассказывается в приложении 3;
- `-s` предписывает компилятору и линкеру выкинуть из генерируемого кода всё, что оттуда возможно выкинуть; в частности, если использовать `-s` при финальной сборке многомодульной программы, то из итогового исполняемого файла будет исключена отладочная информация — даже из тех модулей, которые были откомпилированы с флагом `-g`.

Примеры использования основных флагов компилятора мы приводили в части 4, см. стр. 11, 14, 14, 60, 60, 62 и §4.7.1.

Приложение 2. Средства отладки

Отладчик gdb

С отладчиком `gdb` мы уже встречались, изучая Паскаль (см. т. 1, §2.13.4); исходно `gdb` ориентирован именно на язык Си, так что в определённом смысле работать с ним теперь будет проще, чем когда приходилось отлаживать программы на Паскале. С другой стороны, в первом томе мы рассмотрели далеко не все возможности отладчика. В частности, там не упоминается режим анализа `core`-файла, который полезен в случае аварийного завершения программы; исполняемые файлы, создаваемые компилятором Free Pascal, никогда не порождают `core`-файлов, так что этот режим отладчика при изучении Паскаля был для нас бесполезен. Напротив, при работе на Си анализ `core`-файлов требуется то и дело, так что мы его подробно разберём. Кроме того, мы расскажем про некоторые команды отладчика, оставшиеся «за кадром» в паскалевской части, и дадим кое-какие дополнительные рекомендации.

Для нормальной работы отладчика нужно, чтобы все модули вашей программы были откомпилированы с ключом `-ggdb` или `-g` (см. приложение 1). В некоторых случаях работе отладчика может помешать включённый режим оптимизации, так что перед отладкой оптимизацию лучше отключить.

Напомним, что запустить отладчик для программы, исполняемый файл которой называется `prog`, можно командой

```
gdb ./prog
```

Отладчик сообщит свою версию и некоторую другую информацию, после чего выдаст приглашение своей командной строки, обычно выглядящее так:

```
(gdb)
```

Перечислим основные команды отладчика:

- `run` запускает программу в отладочном режиме; перед запуском целесообразно задать точки останова (см. ниже);
- `start` запускает программу в отладочном режиме, остановив её в начале функции `main`;
- `list` показывает на экране несколько строк программы, предшествующих текущей и идущих непосредственно после текущей;
- `inspect` позволяет просмотреть значение переменной (в том числе и заданной сложным выражением вроде `*(a[i+1].p)`);
- `backtrace` или `bt` показывает текущее содержимое стека, что позволяет узнать последовательность вызовов функций, приведшую к текущему состоянию программы;
- `frame` *<номер>* объявляет текущим один из фреймов, показанных командой `backtrace`, чтобы исследовать значения переменных в этом фрейме, просмотреть текст нужной функции и т. п.;
- `step` выполняет одну строку программы; если в строке содержится вызов функции, текущей строкой станет первая строка этой функции, т. е. трассировка зайдёт внутрь функции;
- `next` подобна команде `step`, с тем отличием, что в тела вызываемых функций не заходит;
- `until` *<номер-строки>* позволяет выполнять программу до тех пор, пока текущей не окажется строка с указанным номером;
- `call` позволяет выполнить вызов произвольной функции;
- `set var` *<присваивание>* позволяет занести заданное значение в переменную; например, `set var i=17` занесёт в переменную `i` в вашей программе значение 17;
- `break` позволяет задать точку приостановки выполнения программы (breakpoint); точка останова может быть задана именем функции, номером строки в текущем файле или выражением <имя-файла>:<номер-строки>, например `file1.c:73`;
- `disable` *<номер-точки-останова>* позволяет временно отменить точку останова; отладчик продолжает помнить о ней, но программа в этом месте не останавливается;
- `enable` *<номер-точки-останова>* снова активирует точку останова, ранее выключенную командой `disable`;
- `ignore` *<номер-точки-останова>* *<число>* предписывает отладчику проигнорировать заданную точку останова указанное число раз; например, `ignore 5 200` означает, что точка останова № 5 должна быть 200 раз проигнорирована, а при попадании в эту точку в 201-й раз программу следует приостановить;
- `cond` *<номер-точки-останова>* *<условие>* задаёт условие, при выполнении которого следует приостановить выполнение программы в данной точке останова; например, `cond 5 i<100` означает, что в точке останова № 5 следует остановиться только в случае, если значение переменной `i` в программе будет меньше 100;
- `cont` позволяет продолжить прерванное выполнение программы;
- `help` позволит узнать подробнее об этих и других командах отладчика;
- `quit` завершает работу отладчика (можно также воспользоваться комбинацией клавиш `Ctrl-D`).

Эти команды позволяют организовать уже знакомое нам *выполнение программы в пошаговом режиме*; именно такой режим подразумевается, если мы запустим отладчик простой командой `gdb ./prog`, не указав никаких дополнительных параметров. Если отлаживаемой программе нужно задать аргументы командной строки, это можно сделать одним из двух способов. Во-первых, можно при запуске указать флаг `--args` примерно так:

```
gdb --args ./prog abra schwabra kadabra
```

(здесь слова `abra`, `schwabra` и `kadabra` выступают в роли аргументов командной строки). Во-вторых, можно запустить отладчик без указания аргументов командной строки, а при запуске программы на пошаговое выполнение указать эти аргументы в команде `run`:

```
(gdb) run abra schwabra kadabra
```

Полезно помнить, что при выполнении программы в режиме отладки вы можете в любой момент нажать комбинацию `Ctrl-C`; отладчик при этом остановит вашу программу и вы сможете, используя команды, проанализировать её текущее состояние, а затем продолжить её выполнение командной `cont`, начать пошаговое выполнение с помощью `step` и `next` и т. д.

Рассмотрим теперь режим анализа причин аварийного завершения по core-файлу. Когда ошибки в программе приводят к её аварийному завершению, операционная система создаёт (если это не запрещено настройками; см. §5.3.12) так называемый core-файл. При этом выдаётся сообщение

```
Segmentation fault (core dumped)
```

Это означает, что в текущей директории аварийно завершённого процесса создан файл с именем `core` (или `prog.core`, где `prog` — имя вашей программы), в который система записала содержимое сегмента данных и стека программы на момент её аварийного завершения. Сегмент кода в core-файл не записывается, поскольку его можно взять из исполняемого файла. С помощью отладчика `gdb` можно проанализировать core-файл, узнав, в частности, при выполнении какой строки программы произошла авария, откуда и с какими параметрами была вызвана функция, содержащая эту строку, каковы были значения переменных на момент аварии и т. д. Отладчик в режиме анализа core-файла запускается командой

```
gdb ./prog core
```

— где `prog` — имя исполняемого файла вашей программы, а `core` — имя созданного системой core-файла. Очень важно при этом, чтобы в качестве исполняемого файла выступал именно тот файл, при исполнении которого получен core-файл. Так, если уже после получения core-файла вы перекомпилируете свою программу, анализ core-файла в большинстве случаев приведёт к ошибочным результатам.

Сразу после запуска отладчика в режиме анализа core-файла рекомендуется дать команду `backtrace` или просто `bt`; в большинстве случаев по одной только её выдаче вы уже поймёте, что и где произошло, в остальных случаях вам помогут команды `frame`, `list` и `inspect`. Следует учитывать, что при анализе core-файла ваша программа уже не работает, она аварийно завершилась; как следствие, вам доступны только команды, показывающие текущее состояние программы. Команды, подразумевающие выполнение программы (обычное или пошаговое), такие как `step`, `next`, `cont`, `break` и прочие, в этом режиме применять нельзя.

В заключение напомним о режиме анализа причин зацикливания. Если ваш процесс «завис», не торопитесь его убивать. С помощью отладчика можно понять, какой фрагмент кода выполняется в настоящий момент, и проанализировать причины зацикливания — либо узнать, что никакого зацикливания на самом деле не произошло, так тоже бывает. Для этого нужно *присоединить* отладчик к существующему процессу. Определите номер процесса с помощью команды `ps`. Допустим, имя исполняемого файла вашей программы — `prog`, и она выполняется как процесс с номером `12345`. Тогда команда запуска отладчика должна выглядеть так:

```
gdb ./prog 12345
```

При подключении отладчика выполнение программы будет приостановлено; вы сможете продолжить его командой `cont`. В случае повторного зацикливания можно приказать отладчику вновь приостановить выполнение нажатием комбинации `Ctrl-C`. Остальные команды (за исключением команд запуска, ведь процесс уже запущен) в этом режиме тоже прекрасно работают, то есть вы можете выполнять программу пошагово, устанавливать и убирать точки останова, просматривать и модифицировать значения переменных и т. д.

Если выйти из отладчика, находящегося в этом режиме, процесс продолжит обычное выполнение, если, конечно, за время отладки вы его не убили.

Возможности отладчика `gdb` не ограничиваются перечисленными; на первых порах вам этого, скорее всего, будет достаточно, но в перспективе мы рекомендуем обратиться к документации и другим источникам, чтобы составить впечатление о возможностях `gdb`, которые остались за рамками нашей книги.

Утилита strace

Утилита `strace` позволяет узнать, к каким системным вызовам (и с какими аргументами) обращается исследуемая программа. Очень часто запуск `strace` оказывается самым быстрым способом понять, что происходит, если что-то пошло не так; её любят использовать не только программисты, но и системные администраторы, поскольку, когда какая-нибудь очередная программа «капризничает» и при этом не желает говорить, в чём дело (увы, очень частая ситуация — не все программисты относятся к обработке ошибок и диагностике с должной внимательностью), как правило, из выдачи `strace` оказывается понятно, на чём программа «споткнулась» — какой файл она пыталась открыть и не смогла, какие данные передавала или получала, после какого события и каким способом «слетела».

Самый простой способ запуска `strace` — указать ей единственным параметром имя программы, работа которой вас интересует, например:

```
strace ./myprogram
```

Информацию, которую выдаст `strace`, вы увидите на экране вперемешку с тем, что будет печатать ваша программа, и иногда это, как ни странно, даже удобно — видно, в какой момент какое сообщение было выдано и это часто помогает понять, что происходит. Впрочем, `strace` может свою выдачу направить в указанный вами файл, но об этом позже.

Если запускаемой программе нужны аргументы командной строки — укажите их как обычно:

```
strace ./myprogram abra schwabra kadabra
```

Сама `strace` как программа тоже принимает через командную строку целый ряд аргументов, но все они должны располагаться *до* имени запускаемой программы, а всё, что после, рассматривается как аргументы командной строки для запускаемой программы, а не для самой `strace`. Кстати, об аргументах. Пожалуй, самый нужный и часто употребляемый из них — `-o`, он позволяет задать имя файла, в котором вы хотите видеть выдаваемый `strace` список системных вызовов, например:

```
strace -o LOG ./myprogram abra schwabra kadabra
```

Здесь вы на экране увидите только то, что напечатает `myprogram`, а информация о её системных вызовах окажется в файле `LOG`. Кстати, если программа `myprogram` будет работать долго и вам захочется узнать, что она делает прямо сейчас, не дожидаясь её завершения, вы можете в соседнем окошке терминала запустить команду `tail -f LOG`, которая будет печатать строки, попадающие в `LOG`, по мере его роста. Можно также использовать `less LOG` (если нажать `Shift-F`, это приведёт примерно к такому же эффекту, как `tail -f`, но можно из этого режима выйти по `Ctrl-C` и, как обычно в `less`, полистать содержимое `LOG` стрелочками, воспользоваться поиском по подстроке и т. п.)

Если ваша программа порождает процессы и вы хотите знать, что делает каждый из них, укажите `strace` флаг `-f` (от слова *fork* — напомним, именно так называется системный вызов, порождающий процесс). В этом случае в каждой строчке своей выдачи `strace` напечатает сначала `pid` процесса, а потом уже системный вызов, значения его параметров и то, что он вернул.

Наконец, если у вас уже есть запущенная программа и вы хотите узнать, что в ней происходит, `strace` может стать прекрасной альтернативой отладчику `gdb`. В отличие от него, `strace` не останавливает отлаживаемый процесс, никак не мешает ему работать, не пытается вести с вами диалог — она просто делает своё дело, то есть выдаёт информацию о системных вызовах, в том числе о том из них, в котором заинтересовавший вас процесс «висел» в блокировке (конечно, если он именно висел в блокировке, а не, скажем, гонял бесконечный цикл). Подключиться к имеющемуся процессу можно с помощью ключа `-p`, параметром ему указывается `pid` процесса:

```
strace -p 2713
```

Ключи `-f` и `-o` в таком режиме тоже работают.

Автор вынужден признать, что, хотя `strace` он использует едва ли не каждый день аж с 1994 года, никаких её ключей, кроме перечисленных, он не помнит. Разумеется, это не значит, что их нет — `strace` обрабатывает несколько десятков опций; заинтересованный читатель найдёт их в выдаче команды `man strace`.

К величайшему прискорбию, `strace` доступна только для Linux. Заставить её работать под FreeBSD и другими BSD-системами можно, но это требует нетривиальных телодвижений, а их собственные программы аналогичного назначения — `truss` и `ktrace` — далеко не столь удобны.

Программа valgrind

Утилита `valgrind` позволяет в автоматическом режиме обнаружить в поведении вашей программы такие особенности, которые свидетельствуют о допущенных ошибках и могут рано или поздно привести к тяжёлым последствиям. Многие ошибки, выявляемые при помощи `valgrind`, носят характер труднообнаружимых, поскольку никак себя не проявляют при выполнении ошибочного кода и либо приводят к неправильному функционированию программы позже, либо, что самое неприятное, могут оставаться незамеченными в отладочных запусках, а проявиться (причём, возможно, с тяжёлыми последствиями) уже на этапе активной эксплуатации программы.

Сама утилита `valgrind` представляет собой *интерпретатор машинного кода*, то есть она реализует возможности центрального процессора программно; отлаживаемая программа запускается на выполнение в этом эмуляторе процессора, что позволяет тщательно анализировать каждое её действие. Конечно, программа, выполняемая таким способом, работает во много раз медленнее, чем при обычном запуске, но результат того стоит. Контролируемое исполнение машинного кода в режиме интерпретации позволяет выявить такие заведомо ошибочные действия, как чтение из памяти, в которую программа не занесла никаких значений (использование неинициализированных переменных), обращение к несуществующим элементам массива, использование освобождённых областей динамической памяти, потерю связи с фрагментами динамической памяти и многие другие.

Учтите, что `valgrind` **полагается на собственные версии многих функций из стандартной библиотеки, для чего при загрузке вашей программы в память подменяет системные разделяемые библиотеки своими**; к сожалению, в результате `valgrind` оказывается совершенно беспомощна со статически собранными исполняемыми файлами — выдаёт бессмысленные сообщения об ошибках едва ли не на каждый вызов `malloc` и `free`. Если вы используете статическую сборку (что само по себе абсолютно правильно), то для отладки с помощью `valgrind` придётся временно перейти на сборку динамическую, убрав из командной строки компилятора или линкера ключ `-static`. Разумеется, ничто не мешает потом вернуть всё как было.

Для примера напишем программу, содержащую заведомую ошибку:

```
#include <stdio.h>
int main()
{
    int x, i;
    for(i = 0; i < 10; i++) {
        if(x < 10)
            printf("First\n");
        else
            printf("Second\n");
        x = i*i;
    }
    return 0;
}
```

Здесь в переменной `x` изначально содержится мусор, а осмысленное значение заносится в неё лишь в конце тела цикла; следовательно, на *первой* итерации цикла значение `x` не определено, то есть работа оператора `if` зависит от неопределённого значения. Сохраним эту программу в файл `badcode.c`, откомпилируем и попробуем выполнить её под контролем `valgrind`. Команда запуска будет выглядеть так:

```
valgrind --tool=memcheck ./badcode
```

На самом деле единственный параметр, который мы здесь указали (`--tool=memcheck`), можно было и не указывать, поскольку именно `memcheck` (*memory checking tool*, инструмент для проверки корректности работы с памятью) `valgrind` использует по умолчанию; тем не менее мы рекомендуем всегда указывать используемый инструмент в явном виде. В некоторых версиях `valgrind` «умолчание» может отличаться.

Запустив `valgrind`, мы увидим довольно пространный текст, в котором строки, которые выдал сам `valgrind`, будут перемешаны со строками, печатаемыми нашей программой. Отметим, что всю свою выдачу `valgrind` направляет в поток диагностики (`stderr`), так что можно, например, перенаправить этот поток в файл:

```
valgrind --tool=memcheck ./badcode 2> VG_LOG
```

В этом случае вы увидите на экране только строки, выдаваемые вашей программой, а информацию от `valgrind` найдёте в файле `VG_LOG`. Впрочем, даже если всё будет выдано на экран вперемешку, отличить строки, которые напечатал `valgrind`, очень просто: они начинаются с `==NNNN==` или `--NNNN--`, где `NNNN` — номер процесса. Например, последняя строка, выданная утилитой `valgrind` при запуске нашего примера неправильной программы, будет выглядеть так:

```
==6503== ERROR SUMMARY: 1 errors from 1 contexts (suppressed: 12 from 7)
```

(здесь номер процесса — `6503`). Как видим, найдена одна ошибка; пролистав полученную выдачу вверх, мы обнаружим текст, подобный следующему:

```
==6503== Conditional jump or move depends on uninitialised value(s)
==6503==    at 0x80483FC: main (badcode.c:6)
```

Английская фраза в первой строке переводится как *условный переход или копирование зависит от неинициализированного значения (значений)*; вторая строка сообщает, что ошибка содержится в функции `main`, в строке 6 файла `badcode.c`. Адрес `0x80483FC` позволяет найти соответствующее место в машинном коде, что может быть полезно, например, при использовании отладчика.

Более того, `valgrind` умеет сам взаимодействовать с отладчиком `gdb`: в этом режиме работы он останавливает исполнение программы при обнаружении ошибки, а запущенный в соседнем окне терминала отладчик даёт возможность с помощью своих обычных команд выполнение продолжить (как пошагово, так и до точки останова), а также посмотреть (в том числе непосредственно в месте обнаружения ошибки) текущую строку программы и её окрестности, стек вызовов функций и значения переменных, что, конечно, позволяет быстрее понять, что́, собственно, такое произошло и по какой причине.

Для задействования этого режима нужно воспользоваться опцией `--vgdb-error=0` или `--vgdb-error=1`. Число после знака равенства означает порядковый номер обнаруженной ошибки, когда `valgrind`'у следует остановиться и ждать подключения отладчика. Если указать `--vgdb-error=0`, `valgrind` остановит вашу программу непосредственно на точке входа (т. е. на метке `_start`), тогда как `--vgdb-error=1` приведёт к остановке на первой обнаруженной ошибке. В руководствах, встречающихся в Интернете, почему-то обычно отдают предпочтение первому варианту (с нулём); рискнём сделать утверждение, что это неудобно, и порекомендовать вариант с единицей.

Когда `valgrind`, повинуясь указанным опциям, остановит выполнение программы, следует в другом окне или на другой консоли запустить `gdb` обычным способом с указанием исполняемого файла программы, в нашем примере это будет

```
gdb ./badcode
```

и, попав в отладчик, дать ему следующую «хитрую» команду:

```
target remote | vgdb
```

Подробно объяснять, что́ тут происходит, мы не будем, отметим только, что `vgdb` — это поставляемый вместе с `valgrind`'ом «сервер» для `gdb`. Если в вашей системе эта программа почему-то оказалась за пределами директорий, указаных в `PATH`, может потребоваться указать полный путь к ней; кроме того, если в вашей системе одновременно работает больше одного `valgrind`'а, например, если это многопользовательская машина и кто-то из ваших коллег отлаживает свои программы одновременно с вами, может потребоваться указать ещё и номер процесса вашего экземпляра `valgrind`. Впрочем, `valgrind` сам, остановившись где сказано, покажет на экране, как это всё должно выглядеть в самом полном варианте.

После успешного подключения к соответствующему экземпляру `valgrind`'а отладчик получает полный контроль над происходящим, как если бы программа выполнялась непосредственно под управлением отладчика в режиме трассировки. Основные команды отладчика, доступные (и полезные) в этом режиме, мы уже перечисляли раньше (см. стр. 220). Отметим только один интересный момент. **С момента предписанной остановки для дальнейшего взаимодействия с отладчиком сам `valgrind` считает, что всё управление происходящим отныне и до завершения работы должно происходить исключительно через отладчик**. Собственно говоря, даже прибить `valgrind` не получится — все привычные комбинации клавиш работать перестают. **Самым быстрым способом прекратить это всё оказывается команда отладчика `kill`**. Проблема тут в том, что про эту команду мало кто помнит — в обычных режимах использования `gdb` она практически никогда не нужна.

Между прочим, даже выход из отладчика (что вообще-то очень хочется сделать чисто «по привычке») никак не поможет угомонить упрямый `valgrind`, так что, если выполнение отлаживаемой программы не завершилось, но вы не хотите его продолжать (что практически всегда именно так и бывает), **не забудьте сначала дать команду `kill`, и только потом выходить из отладчика**. Конечно, всегда остаётся вариант уничтожить процесс `valgrind`'а из другого окна, просто послав ему сигнал, но это требует больше времени.

Между прочим, в очень старых версиях `valgrind` взаимодействие с отладчиком было организовано в разы проще: `valgrind` мог сам запустить для нас отладчик при обнаружении ошибочных действий нашей программы. Эта возможность активировалась одним-единственным ключом командной строки `--db-attach=yes`, и больше можно было ни о чём не беспокоиться. Текст, рассказывающий о таком варианте взаимодействия, впервые написанный для старой методички и воспроизведённый в первых двух изданиях этой книги, был раз в пять короче.

Разумеется, «современные» версии `valgrind` не поддерживают этот ключ, наплевав и на обратную совместимость, и заодно на пользователя. Вы всё ещё верите в существование технического прогресса?

Подробное описание всех основных возможностей `valgrind` заняло бы целую книгу, причём такие книги, собственно говоря, уже есть, некоторые из них даже можно найти в Интернете. Дальнейшее изучение этой программы оставим читателю. Чтобы почувствовать себя увереннее в обращении с `valgrind`, лучше всего будет написать ещё несколько маленьких программ с очевидными ошибками, такими как неосвобождение выделенной динамической памяти, затирание единственного указателя, хранящего адрес того или иного объекта в динамической памяти и т. п., и попробовать их все позапускать под управлением `valgrind`; в этом случае его диагностика не станет для вас сюрпризом в более сложных случаях. Обязательно освойте `valgrind`! Это займёт у вас от силы полтора часа, которые окупятся, скорее всего, не одну сотню раз.

Приложение 3. Автоматическая сборка: утилита make

Рано или поздно сложность вашей программы заходит за невидимый рубеж, когда вы просто-таки вынуждены разбить её на отдельные модули (см. §4.7). Практически сразу вы можете заметить, что при сборке такой программы давать все нужные команды вручную довольно неудобно и здесь напрашивается некая автоматизация; именно такую автоматизацию предоставит вам `make`. Кратко говоря, эта утилита позволяет автоматически строить одни файлы на основании других (например, исполняемые файлы на основании исходных текстов программы) в соответствии с заданными правилами. При этом make отслеживает даты последней модификации файлов и производит перестроение только тех целевых файлов, для которых исходные файлы претерпели изменения.

Существует несколько различных версий утилиты make. Изложенное в данном параграфе соответствует варианту Gnu Make; именно этот вариант используется в большинстве дистрибутивов Linux. **При работе с FreeBSD для вызова Gnu Make вместо команды `make` применяется `gmake`.** Если в системе такой команды нет, обратитесь к системному администратору с просьбой её установить.

Правила для утилиты make задаются в файле `Makefile`, который утилита ищет в текущей директории.

Простейший Makefile

Допустим, ваша программа состоит из главного модуля `main.c`, содержащего функцию `main`, а также из дополнительных модулей `mod1.c` и `mod2.c`, имеющих заголовочные файлы `mod1.h` и `mod2.h`. Для сборки исполняемого файла (назовем его `prog`) нужно дать следующие команды:

```
gcc -g -Wall -c mod1.c -o mod1.o
gcc -g -Wall -c mod2.c -o mod2.o
gcc -g -Wall main.c mod1.o mod2.o -o prog
```

Первые две команды даются для компиляции дополнительных модулей. Полученные в результате файлы `mod1.o` и `mod2.o` используются в третьей команде для сборки исполняемого файла.

Допустим, мы уже произвели компиляцию программы, после чего внесли изменения в файл `mod1.c` и хотим получить исполняемый файл с учетом внесённых изменений. При этом нам надо будет дать только две команды (первую и третью), поскольку перекомпиляции модуля `mod2` не требуется. Чтобы подобные ситуации отслеживались автоматически, мы можем использовать утилиту `make`. Для этого напишем следующий `Makefile`:

```
mod1.o: mod1.c mod1.h
        gcc -g -Wall -c mod1.c -o mod1.o

mod2.o: mod2.c mod2.h
        gcc -g -Wall -c mod2.c -o mod2.o

prog: main.c mod1.o mod2.o
        gcc -g -Wall main.c mod1.o mod2.o -o prog
```

Поясним, что файл состоит из так называемых *целей* (в нашем случае таких целей три: `mod1.o`, `mod2.o` и `prog`). Описание каждой цели состоит из заголовка и списка команд. Заголовок цели — это **одна** строка, начинающаяся всегда с первой позиции (т. е. перед ней не допускаются пробелы и т. п.). В начале строки пишется имя цели (обычно это просто имя файла, который мы хотим построить). Оставшаяся часть заголовка отделяется от имени цели двоеточием. После двоеточия перечисляется, от каких файлов (или, в более общем случае, от каких целей) зависит построение файла. В данном случае мы указали, что модули зависят от их исходных текстов и заголовочных файлов, а исполняемый файл — от основного исходного файла и от двух объектных файлов.

После строки заголовка идёт список команд (в нашем случае все три списка имеют по одной команде). **Строка команды всегда начинается с символа табуляции**, замена табуляции пробелами недопустима и ведёт к ошибке. Утилита `make` считает признаком конца списка команд первую строку, начинающуюся с символа, отличного от табуляции.

Имея в текущей директории вышеописанный `Makefile`, мы можем для сборки нашей программы дать команду `make prog`.

Переменные и псевдопеременные

В предыдущем параграфе описан `Makefile`, в котором можно обнаружить несколько повторяющихся фрагментов. Так, строка параметров компилятора «`-g -Wall`» встречается во всех трех целях. Помимо необходимости повторения одного и того же текста, мы можем столкнуться с проблемами при модификации. Предположим, нам понадобилось задать компилятору второй уровень оптимизации кода по скорости (флаг `-O2`). Для этого нам пришлось бы внести совершенно одинаковые изменения в три разные строки файла. В более сложном случае таких строк может оказаться несколько десятков и даже сотен. Аналогичная проблема встанет, например, если мы захотим произвести сборку другим компилятором.

Решить проблему позволяет введение *make-переменных*. Обозначим имя компилятора Си переменной `CC`, а общие параметры компиляции — переменной `CFLAGS` (причины выбора именно таких обозначений станут ясны чуть позже), и перепишем наш `Makefile`:

```
CC = gcc
CFLAGS = -g -Wall -ansi -pedantic

mod1.o: mod1.c mod1.h
        $(CC) $(CFLAGS) -c mod1.c -o mod1.o

mod2.o: mod2.c mod2.h
        $(CC) $(CFLAGS) -c mod2.c -o mod2.o

prog: main.c mod1.o mod2.o
        $(CC) $(CFLAGS) main.c mod1.o mod2.o -o prog
```

Существуют соглашения об именах переменных, причём некоторым переменным утилита `make` присваивает значения сама, если соответствующие значения не заданы явно. Вот некоторые традиционные имена переменных:

- `CC` — команда вызова компилятора языка Си;
- `CFLAGS` — параметры для компилятора языка Си;
- `CXX` — команда вызова компилятора языка Си++;
- `CXXFLAGS` — параметры для компилятора языка Си++;
- `CPPFLAGS` — параметры препроцессора (обычно сюда помещают предопределённые макропеременные);
- `LD` — команда вызова системного линкера (редактора связей);
- `MAKE` — команда вызова утилиты `make`.

По умолчанию переменные `CC`, `CXX`, `LD` и `MAKE` имеют соответствующие значения, справедливые для данной системы и в данной ситуации. Значения остальных перечисленных переменных по умолчанию пусты. Так, при написании нашего примерного `Makefile` мы могли бы пропустить строку, в которой задаётся значение переменной `CC`, в надежде, что соответствующее значение переменная получит без нашей помощи.

Кроме таких переменных общего назначения, в каждой цели могут использоваться так называемые *псевдопеременные*:

- `$@` — имя текущей цели;
- `$<` — имя первой цели из списка зависимостей;
- `$^` — весь список зависимостей.

Используя их, мы можем переписать наш `Makefile` следующим образом:

```
CFLAGS = -g -Wall -ansi -pedantic

mod1.o: mod1.c mod1.h
        $(CC) $(CFLAGS) -c $< -o $@

mod2.o: mod2.c mod2.h
        $(CC) $(CFLAGS) -c $< -o $@

prog: main.c mod1.o mod2.o
        $(CC) $(CFLAGS) $^ -o $@
```

Обобщённые цели

Как можно заметить, в том варианте `Makefile`, который мы написали в конце предыдущего параграфа, правила для сборки обоих дополнительных модулей оказались совершенно одинаковыми. Можно пойти дальше и задать одно обобщённое правило построения объектного файла для любого модуля, написанного на языке Си, исходный файл которого имеет имя с суффиксом `.c`, а заголовочный файл — имя с суффиксом `.h`:

```
%.o: %.c %.h
        $(CC) $(CFLAGS) -c $< -o $@
```

Если теперь задать список дополнительных модулей с помощью переменной, получим следующий вариант `Makefile`:

```
OBJMODULES = mod1.o mod2.o
CFLAGS = -g -Wall -ansi -pedantic

%.o: %.c %.h
        $(CC) $(CFLAGS) -c $< -o $@

prog: main.c $(OBJMODULES)
        $(CC) $(CFLAGS) $^ -o $@
```

Теперь для добавления к программе нового модуля нам достаточно добавить имя его объектного файла к значению переменной `OBJMODULES`. Если перечисление модулей через имена объектных файлов представляется неестественным, можно заменить первую строку `Makefile` следующими двумя строками:

```
SRCMODULES = mod1.c mod2.c
OBJMODULES = $(SRCMODULES:.c=.o)
```

Запись `$(SRCMODULES:.c=.o)` означает, что нужно взять значение переменной `SRCMODULES` и в каждом входящем в это значение слове заменить суффикс `.c` на `.o`.

Псевдоцели

Утилиту `make` можно использовать не только для построения файлов, но и для выполнения произвольных действий. Добавим к нашему файлу две дополнительные цели:

```
run: prog
        ./prog

clean:
        rm -f *.o prog
```

Теперь по команде `make run` утилита `make` произведёт, если нужно, сборку нашей программы и запустит её, а командой `make clean` мы сможем очистить рабочую директорию от объектных и исполняемых файлов (например, если нам понадобится произвести сборку программы с нуля или подготовить исходные тексты к архивированию). Такие цели обычно называют *псевдоцелями*, поскольку их имена не обозначают имён создаваемых файлов.

Автоматическое отслеживание зависимостей

В более сложных проектах модули могут использовать заголовочные файлы других модулей, что делает необходимой перекомпиляцию модуля при изменении заголовочного файла, не относящегося к этому модулю. Информацию о том, какой модуль от каких файлов зависит, можно задать вручную, однако в больших программах этот способ приведет к трудностям, поскольку программист при модификации исходных файлов может случайно забыть внести изменения в `Makefile`.

Разумнее будет поручить отслеживание зависимостей компьютеру. Утилита `make` позволяет наряду с обобщённым правилом указать список зависимостей для построения конкретных модулей. Например:

```
%.o: %.c
        $(CC) $(CFLAGS) -c $< -o $@

mod1.o: mod1.c mod1.h mod2.h mod3.h
```

В этом случае для построения файла `mod1.o` будет использовано обобщённое правило (поскольку никаких команд в цели `mod1.o` мы не указали), но список зависимостей будет использован из цели `mod1.o`.

Списки зависимостей можно построить с помощью компилятора. Получить строку, подобную последней строке вышеприведённого примера, можно, дав команду «`gcc -MM mod1.c`». Если результат выполнения такой команды перенаправить в файл, то этот файл можно будет включить в наш `Makefile` директивой `include`. Эта директива имеет специальную форму со знаком «`-`», при использовании которой `make` не выдаёт ошибок, если файл не найден. Если использовать для файла зависимостей имя `deps.mk`, директива его включения будет выглядеть так:

```
-include deps.mk
```

Более того, если предусмотреть цель для генерации файла, включаемого такой директивой, например:

```
deps.mk: $(SRCMODULES)
        $(CC) -MM $^ > $@
```

— утилита `make`, прежде чем начать построение любых других целей, будет пытаться построить включаемый файл.

Отметим, что такое поведение нежелательно для псевдоцели clean, поскольку для очистки рабочей директории от мусора построение файлов зависимостей не нужно и только отнимает время. Чтобы избежать этого, следует снабдить директиву `-include` условной конструкцией, исключающей эту строку из рассмотрения, если единственной целью, заданной в командной строке, является цель clean. Это делается с помощью директивы `ifneq` и встроенной переменной `MAKECMDGOALS`:

```
ifneq (clean, $(MAKECMDGOALS))
-include deps.mk
endif
```

Окончательно `Makefile` будет выглядеть так:

```
SRCMODULES = mod1.c mod2.c
OBJMODULES = $(SRCMODULES:.c=.o)
CFLAGS = -g -Wall -ansi -pedantic

%.o: %.c %.h
        $(CC) $(CFLAGS) -c $< -o $@

prog: main.c $(OBJMODULES)
        $(CC) $(CFLAGS) $^ -o $@

ifneq (clean, $(MAKECMDGOALS))
-include deps.mk
endif

deps.mk: $(SRCMODULES)
        $(CC) -MM $^ > $@

clean:
        rm -f *.o prog deps.mk
```

Относительно флага `gcc -MM` следует сделать одно важное замечание. В ранних (вплоть до 3.*) версиях `gcc` этот флаг означал, что анализировать надо все заголовочные файлы, подключаемые с помощью `#include "..."`, при этом файлы, подключаемые через `#include <...>`, следует игнорировать. К сожалению, это простое и понятное поведение было «модифицировано» (точнее, просто сломано) под давлением безответственных лиц из стандартизационного комитета, которым не даёт покоя классическая семантика `#include`, завязанная на файлы файловой системы. В современных версиях `gcc` флаг `-MM` не различает виды директив `#include`; вместо этого `gcc -MM` игнорирует файлы, включённые из «системных директорий», вне зависимости от того, какой директивой они включаются.

Всё это может оказаться важно, если вы используете возможности сторонней библиотеки, которая в системе не установлена, так что приходится в явном виде указывать пути к её заголовочным файлам. Поскольку эта библиотека не является частью вашей программы, использовать для подключения её заголовочников следует вариант `#include <...>`; но само по себе это не спасёт вас от генерации зависимостей от файлов, не являющихся частью вашего проекта и, следовательно, не изменяющихся и не требующих внимания со стороны `make`. Проблема решается использованием флага `-isystem` вместо классического `-I`; благодарите стандартизаторов за необходимость всех этих нелепых танцев с бубнами.

Приложение 4. Редактор Vim: больше возможностей

Редактор Vim мы уже описывали в первом томе (см. §1.2.12). Теперь, когда вы изучили язык Си и, можно надеяться, начали работать с программами, имеющими заметный объём, самое время обратить внимание на функции Vim, которые во вводной части книги были оставлены за кадром.

Для начала обратим внимание на типичную ситуацию в рабочем цикле программиста: компилятор при очередном его запуске вывалил несколько экранов ошибок и нам теперь предстоит их исправлять. Найти нужное место в тексте, в принципе, особых проблем не составляет: запомнить выданный компилятором номер строки, открыть нужный файл в редакторе текстов, и если это Vim, то ввести символ двоеточия и номер строки; но всё это требует времени, пусть и не очень большого. Между тем Vim вполне способен выполнить эту работу за нас.

Чтобы задействовать эту его способность, программу нужно (обязательно) компилировать и компоновать с помощью системы `make` (см. приложение 3), так что если вы ещё не написали себе `Makefile`, самое время это сделать; поверьте, делать это всё равно придётся. Утилиту `make` можно запустить изнутри Vim, и именно так следует поступить, если вы хотите, чтобы Vim потом сам открыл вам нужный файл и поставил курсор на нужную строку. Например, если ваша программа собирается командой `make prog`, в командном режиме Vim следует набрать `:make prog`. Редактор при этом перехватит поток диагностических сообщений, выдаваемых самим `make` и всеми программами, которые из него запускаются. Как, возможно,

заметил читатель, компиляторы выдают сообщения о предупреждениях и ошибках, следуя определённому формату: сначала печатается имя файла, потом (через двоеточие) номер строки в нём, потом, возможно, ещё и номер позиции в строке (тоже через двоеточие), и лишь потом, после ещё одного двоеточия и пробела — само текстовое сообщение. Именно этот формат позволяет Vim'у понять, какое место вашей программы «не понравилось» компилятору.

Сообщение, выданное компилятором, вы при этом увидите в нижней строке экрана, причём оно может быть выдано в сокращённой форме — редактор старается уместить его в одну строчку. Чтобы увидеть сообщение компилятора целиком, используйте команду `:cc`, для перехода к следующей ошибке или предупреждению — команду `:cn`, для возврата к предыдущему сообщению — команду `:cp` (от слов *next* и *previous*). Здесь настоятельно рекомендуется не увлекаться — помните, что последующие сообщения об ошибках могут быть следствием того, что компилятор запутался на предыдущей ошибке и не смог корректно восстановиться; правильнее будет после каждого исправления снова запускать `make`, если только вы не уверены, что отчётливо видите следующую ошибку в тексте и знаете, как её надо исправить.

Возможности Vim возрастут, если вы создадите тэг-файл для ваших исходных текстов. Это делается программой `ctags`, например:

```
ctags *.c *.h
```

В результате в текущей директории появится файл `tags`, содержащий информацию о расположении в вашей программе объявлений и описаний. Наиболее очевидное использование этой информации — автоматический поиск в вашей программе описания функции, переменной или типа, имя которых находится в настоящий момент под курсором. Для использования этой возможности нажмите комбинацию `Ctrl-]`; вернуться обратно можно нажатием `Ctrl-T`.

Возможности Vim намного шире, чем мы можем описать в книге; в действительности они настолько широки, что вряд ли в мире найдётся человек, использующий их хотя бы наполовину. Какие из них понравятся лично вам — зависит от многих причин; в Интернете можно найти руководства по этому редактору буквально на любой вкус.

Литература

[1] Керниган Б., Ритчи Д. Язык программирования Си. 3-е издание. СПб.: Невский диалект, 2000.

[2] William H. Press, Brian P. Flannery, Saul A. Teukolsky, William T. Vetterling. Numerical Recipes in C: The Art of Scientific Computing. New York: Cambridge University Press, 1992 (2nd ed.)

[3] А. М. Робачевский. Операционная система Unix. Изд-во «BHV—Санкт-Петербург», Санкт-Петербург, 1997.

[4] Уильям Стивенс. UNIX: Взаимодействие процессов. СПб.: Питер, 2002.

[5] У. Р. Стивенс. UNIX: Разработка сетевых приложений. СПб.: Питер, 2004.

[6] Эрик С. Реймонд. Искусство программирования для Unix. М.: изд-во Вильямс, 2005. *В оригинале на английском языке книга доступна в сети Интернет по адресу* `http://www.faqs.org/docs/artu/`

[7] Э. Танненбаум. Современные операционные системы. 2-е издание. СПб.: Питер, 2002.

[8] Вирт Н. Алгоритмы и структуры данных. Пер. с англ. 2-е изд. СПб.: Невский Диалект, 2001. 352 с. ISBN 5-7940-0065-1

[9] Кормен Т., Лейзерсон Ч., Ривест Р. Алгоритмы: построение и анализ. М.: МЦНМО, 2000. 960 с. ISBN 5-900916-37-5.

[10] Кнут Д. Искусство программирования, т. 3. Сортировка и поиск, 2-е изд. Пер. с англ. М.: Издательский дом «Вильямс», 2000. 832 с. ISBN 5-8459-0082-4

[11] Wolfgang Mauerer. Professional Linux Kernel Architecture. Wiley Publishing, Inc., Indianapolis, 2008. *Полный текст книги доступен в сети Интернет; используйте поисковые машины.*

[12] Linus Åkesson. The TTY demystified. 25-Jul-2008, `http://www.linusakesson.net/programming/tty/`

[13] Bert Hubert. The ultimate `SO_LINGER` page, or: why is my tcp not reliable // Bert Hubert finally blogs, 18-Jan-2009. `https://blog.netherlabs.nl/articles/2009/01/18/the-ultimate-so_linger-page-or-why-is-my-tcp-not-reliable`

[14] Edward A. Lee. The Problem with Threads. UC Berkeley, January 10, 2006, Technical Report No. UCB/EECS-2006-1 `http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-1.html`

[15] Ulrich Drepper. Futexes Are Tricky. November 5, 2011. `http://www.akkadia.org/drepper/futex.pdf`

[16] Andy Walker. Mandatory File Locking For The Linux Operating System. 15 April 1996 (Updated September 2007). `https://www.kernel.org/doc/Documentation/filesystems/mandatory-locking.txt`

[17] Keith Adams, Ole Agesen. Comparison of Software and Hardware Techniques for x86 Virtualization // ASPLOS'06 October 21–25, 2006, San Jose, California, USA.

[18] CFS Scheduler // Linux kernel documentation, file `scheduler/sched-design-CFS.txt`. `https://www.kernel.org/doc/Documentation/scheduler/sched-design-CFS.txt`

http://www.stolyarov.info

Учебное издание

СТОЛЯРОВ Андрей Викторович

ПРОГРАММИРОВАНИЕ: ВВЕДЕНИЕ В ПРОФЕССИЮ
в трёх томах
Издание 2,92 (электронное), с исправлениями
Том II: СИСТЕМЫ И СЕТИ

Рисунок и дизайн обложки *Елены Доменновой*

Официально опубликовано 30 апреля 2026 г.
на сайте http://www.stolyarov.info