Александр Побегайло

bhv

СИСТЕМНОЕ ПРОГРАММИРОВАНИЕ В WINDOWS

- Синхронизация потоков
- Каналы передачи данных и почтовые ящики
- Виртуальная память и файлы
- Асинхронная обработка данных
- Безопасность доступа к объектам

Наиболее полное руководство

+

В ПОДЛИННИКЕ®

УДК 681.3.06
ББК 32.973.26-018.1
П41

Побегайло А. П.

П41 Системное программирование в Windows. — СПб.: БХВ-Петербург, 2006. — 1056 с.: ил.

ISBN 5-94157-792-3

Подробно рассматриваются вопросы системного программирования с использованием интерфейса Win32 API. Описываются управление потоками и процессами, включая их диспетчеризацию; синхронизация потоков; передача данных между процессами, с использованием анонимных и именованных каналов, а также почтовых ящиков; структурная обработка исключений; управление виртуальной памятью; управление файлами и каталогами; асинхронная обработка данных; создание динамически подключаемых библиотек; разработка сервисов. Отдельная часть книги посвящена управлению безопасностью объектов в Windows. Каждая тема снабжена практическими примерами использования функций Win32 API, которые представлены работающими листингами. Это позволяет использовать книгу в качестве пособия по системному программированию или справочника для системного программиста. Прилагаемый компакт-диск содержит листинги и проекты всех программ, рассмотренных в книге.

Для программистов

УДК 681.3.06
ББК 32.973.26-018.1

Группа подготовки издания:

Главный редактор	*Екатерина Кондукова*
Зам. главного редактора	*Игорь Шишигин*
Зав. редакцией	*Григорий Добин*
Редактор	*Андрей Смышляев*
Компьютерная верстка	*Натальи Караваевой*
Корректор	*Наталия Першакова*
Оформление обложки	*Елены Беляевой*
Зав. производством	*Николай Тверских*

Лицензия ИД № 02429 от 24.07.00. Подписано в печать 15.01.06.
Формат 70×100 1/16. Печать офсетная. Усл. печ. л. 85,14.
Тираж 3000 экз. Заказ №
"БХВ-Петербург", 194354, Санкт-Петербург, ул. Есенина, 5Б.

Отпечатано с готовых диапозитивов
в ОАО "Техническая книга"
190005, Санкт-Петербург, Измайловский пр., 29.

ISBN 5-94157-792-3

Оглавление

Глава 1

Операционные системы и их интерфейсы

1.1. Назначение операционной системы

Физическими или *аппаратными ресурсами компьютера* называются физические устройства, из которых состоит компьютер. К таким устройствам относятся центральный процессор, оперативная память, внешняя память, шины передачи данных и различные устройства ввода-вывода информации. *Логическими* или *информационными ресурсами компьютера* называются данные и программы, которые хранятся в памяти компьютера. Когда говорят обо всех ресурсах компьютера, включая как физические, так и логические ресурсы, то обычно используют термины *ресурсы компьютера* или *системные ресурсы*.

Для выполнения на компьютере какой-либо программы необходимо, чтобы она имела доступ к ресурсам компьютера. Этот доступ обеспечивает операционная система. Можно сказать, что *операционная система* — это комплекс программ, который обеспечивает доступ к ресурсам компьютера и управляет ими. Другими словами, операционная система — это администратор или менеджер ресурсов компьютера. Назначение операционной системы состоит в обеспечении пользователя программными средствами для использования ресурсов компьютера и эффективном разделении этих ресурсов между пользователями. Отсюда следует, что главными функциями операционной системы являются управление ресурсами компьютера и диспетчеризация или планирование этих ресурсов.

1.2. Типы операционных систем

Все программы, которые работают на компьютере под управлением операционной системы, называются *пользовательскими программами*. Совокупность пользовательских программ, которая предназначена для решения определенной задачи, называется *приложением*. Если операционная система одновременно

может исполнять только одну пользовательскую программу, то она называется *однопрограммной* или *однопользовательской*. Если же под управлением операционной системы могут одновременно выполняться несколько пользовательских программ, то такая операционная система называется *мультипрограммной* или *многопользовательской*.

В зависимости от назначения операционной системы и аппаратуры компьютера, на котором она работает, можно определить несколько типов операционных систем. Если операционная система может работать только на компьютере с одним процессором, то такая операционная система называется *однопроцессорной*. Если же операционная система может работать также и на компьютере, который содержит несколько процессоров, то такая операционная система называется *мультипроцессорной*.

Следует делать различие между операционными системами, которые предназначены для обработки информации под управлением пользователя, и операционными системами, которые предназначены для управления объектами при помощи компьютера в реальном времени без участия пользователя. Такими объектами могут быть, например, робот или самолет. Операционная система, предназначенная для работы в режиме реального времени, называется *операционной системой реального времени*. Главное отличие операционных систем реального времени заключается в их быстром реагировании на внешние события и надежности функционирования. Если пользователь, сидя у компьютера, будет только раздражен медленной или ненадежной работой операционной системы, то медленная или ненадежная работа операционной системы реального времени может вызвать поломку оборудования и аварию.

В дальнейшем будут рассматриваться только операционные системы фирмы Microsoft, а именно Windows 98 и Windows 2000, которые предназначены для использования на персональных компьютерах. Эти операционные системы отличаются своей внутренней организацией, но используют один и тот же интерфейс для программирования приложений — Win32 API. Мы не будем рассматривать операционную систему Windows CE, которая предназначена для использования в таких различных устройствах, как, например, устройства бытовой электроники, контроллеры для управления технологическими процессами и устройства управления коммуникационным оборудованием. Но, разобравшись в изложенном материале, вы получите опыт, который поможет вам как в изучении Windows CE, так и других операционных систем.

Относительно операционной системы Windows XP можно сказать следующее. Те приемы системного программирования, которые рассмотрены в этой книге для операционной системы Windows 2000, также работают и в операционной системе Windows XP.

1.3. Интерфейс программирования приложений Win32 API

Интерфейс программирования приложений Win32 API представляет собой набор функций и классов, которые используются для программирования приложений, работающих под управлением операционных систем фирмы Microsoft. Следует отметить, что в работе многих функций Win32 API существуют различия, которые зависят от типа операционной системы. Кроме того, некоторые функции работают только в операционной системе Windows 2000 и не поддерживаются операционной системой Windows 98. Все эти случаи будут отмечаться отдельно. Но все же в работе функций Win32 API в разных версиях операционных систем гораздо больше общего, чем различий. Поэтому чаще всего мы будем говорить, что функции Win32 API предназначены для разработки приложений на платформах операционных систем Windows, не делая различия между операционными системами Windows 98 и Windows 2000. Это соглашение значительно облегчит изложение материала, не загромождая его ненужными подробностями, которые отвлекают от сути рассматриваемых вопросов.

Функционально Win32 API подразделяется на следующие категории:

- Base Services (базовые сервисы);
- Common Control Library (библиотека общих элементов управления);
- Graphics Device Interface (интерфейс графических устройств);
- Network Services (сетевые сервисы);
- User Interface (интерфейс пользователя);
- Windows NT Access Control (управление доступом для Windows NT);
- Windows Shell (оболочка Windows);
- Windows System Information (информация о системе Windows).

Кратко опишем функции, которые выполняются в рамках этих категорий. Функции базовых сервисов обеспечивают приложениям доступ к ресурсам компьютера. Категория Common Control Library содержит классы окон, которые часто используются в приложениях. Интерфейс графических устройств обеспечивает функции для вывода графики на дисплей, принтер и другие графические устройства. Сетевые сервисы используются при работе компьютеров в компьютерных сетях. Интерфейс пользователя обеспечивает функции для взаимодействия пользователя с приложением, используя окна для ввода-вывода информации. Категория Windows NT Access Control содержит функции, которые используются для защиты информации путем контроля и ограничения доступа к защищаемым объектам. Категории Windows Shell и Windows System Information содержат соответственно функции для работы с оболочкой и конфигурацией операционной системы Windows.

В курсе системного программирования главным образом изучается назначение и использование функций из категорий Base Services и Windows NT Access Control. Функции из категорий Common Control Library, Graphics Device Interface и User Interface используются для разработки интерфейса приложений, а курс, который изучает назначение и использование этих функций, как правило, называется "Программирование пользовательских интерфейсов в Windows". Изучив два этих курса и добавив сюда свои знания по программированию на языке C++, вы получите довольно содержательное представление о разработке приложений на платформе Win32 API.

В связи с тем, что программирование графических пользовательских интерфейсов в Windows само по себе является довольно трудоемким занятием, мы будем изучать функции ядра Windows, работая только с консольными приложениями. Это упростит изложение предмета и избавит нас от большого количества кода, не относящегося к существу рассматриваемых вопросов.

1.4. Типы данных в Win32 API

Прежде всего заметим, что интерфейс программирования приложений Win32 API ориентирован на язык программирования C или, в более широком смысле, на процедурные языки программирования. Поэтому в этом интерфейсе, не используются такие возможности языка программирования C++, как классы, ссылки и механизм обработки исключений.

Чтобы сделать интерфейс Win32 API более независимым от конкретного языка программирования или, может быть, более соответствующим аппаратному обеспечению компьютера, разработчики этого интерфейса определили новые простые типы данных. Эти типы данных используются в прототипах функций интерфейса Win32 API.

Новые простые типы данных определены как синонимы простых типов данных языка программирования C. Чтобы отличать эти типы от других типов, их имена определены прописными буквами. Общее количество простых типов данных, определенных в интерфейсе Win32 API, довольно велико. Поэтому ниже приведены определения только тех простых типов данных из этого интерфейса, которые очевидным образом переименовывают простые типы данных языка программирования C.

```
typedef char CHAR;

typedef unsigned char  UCHAR;
typedef UCHAR          *PUCHAR;
typedef unsigned char  BYTE;
typedef BYTE           *PBYTE;
typedef BYTE           *LPBYTE;
```

```
typedef short SHORT;

typedef unsigned short  USHORT;
typedef USHORT          *PUSHORT;
typedef unsigned short  WORD;
typedef WORD            *PWORD;
typedef WORD            *LPWORD;

typedef int   INT;
typedef int   *PINT;
typedef int   *LPINT;
typedef int   BOOL;
typedef BOOL  *PBOOL;
typedef BOOL  *LPBOOL;

typedef unsigned int  UINT;
typedef unsigned int  *PUINT;

typedef long LONG;
typedef long   *LPLONG;

typedef unsigned long  ULONG;
typedef ULONG           *PULONG;
typedef unsigned long   DWORD;
typedef DWORD           *PDWORD;
typedef DWORD           *LPDWORD;

typedef float  FLOAT;
typedef FLOAT  *PFLOAT;

typedef void   *LPVOID;
typedef CONST void  *LPCVOID;
```

Остальные простые типы данных, определенные в интерфейсе Win32 API, имеют, как правило, специфическое назначение и поэтому они будут описаны при их использовании.

Кроме того, в интерфейсе Win32 API определены символические константы `FALSE` и `TRUE` для обозначения соответственно ложного и истинного логических значений. Определения этих констант приведены ниже.

```
#ifndef FALSE
#define FALSE  0
```

```
#endif

#ifndef TRUE
#define TRUE  1
#endif
```

В интерфейсе Win32 API также определено множество сложных типов данных, таких как структуры и перечисления. Как правило, эти типы данных имеют специфическое назначение и поэтому будут описаны при их непосредственном использовании.

1.5. Объекты и их дескрипторы в Windows

Объектом в Windows называется структура данных, которая представляет системный ресурс. Таким ресурсом может быть, например, файл, канал, графический рисунок. Операционные системы Windows предоставляют приложению объекты трех категорий:

- User (объекты интерфейса пользователя);
- Graphics Device Interface (объекты интерфейса графических устройств);
- Kernel (объекты ядра).

Категория User включает объекты, которые используются приложением для интерфейса с пользователем. К таким объектам относятся, например, окна и курсоры. Категория Graphics Device Interface включает объекты, которые используются для вывода информации на графические устройства. К таким объектам относятся, например, кисти и перья. Категория Kernel включает объекты ядра операционной системы Windows. К таким объектам относятся, например, файлы и каналы. При изучении системного программирования подробно рассматриваются только объекты категории Kernel. Объекты двух оставшихся категорий рассматриваются при изучении программирования графических интерфейсов.

Под доступом к объектам понимается возможность приложения выполнять над объектом некоторые функции. Приложение не имеет прямого доступа к объектам, а обращается к ним косвенно. Для этого в операционных системах Windows каждому объекту ставится в соответствие дескриптор (handle). В Win32 API дескриптор имеет тип `HANDLE`. *Дескриптор объекта* представляет собой запись в таблице, которая поддерживается системой и содержит адрес объекта и средства для идентификации типа объекта. Дескрипторы объектов создаются операционной системой и возвращаются функциями Win32 API, которые создают объекты. За редким исключением, эти функции имеют вид `CreateObject`, где слово `Object` заменяется именем конкретного объекта. Например, процесс создается при помощи вызова функции `CreateProcess`. Как правило, такие функции возвращают дескриптор соз-

данного объекта. Если это значение не равно `NULL` (или отрицательному значению), то объект создан успешно.

После завершения работы с объектом его дескриптор нужно закрыть, используя функцию `CloseHandle`, которая имеет следующий прототип:

```
BOOL  CloseHandle(
  HANDLE  hObject    // дескриптор объекта
);
```

При успешном завершении функция `CloseHandle` возвращает ненулевое значение, в противном случае — `FALSE`. Функция `CloseHandle` удаляет дескриптор объекта, но сам объект удаляется не всегда. Дело в том, что в Windows на один и тот же объект могут ссылаться несколько дескрипторов, которые создаются другими функциями для доступа к уже созданному ранее объекту. Функция `CloseHandle` уничтожает объект только в том случае, если на него больше не ссылается ни один дескриптор.

Часть I

Управление потоками и процессами

Глава 2

Потоки и процессы

2.1. Определение потока

Определение потока тесно связано с последовательностью действий процессора во время исполнения программы. Исполняя программу, процессор последовательно выполняет инструкции программы, иногда осуществляя переходы в зависимости от некоторых условий. Такая последовательность выполнения инструкций программы называется *потоком управления* внутри программы. Отметим, что поток управления зависит от начального состояния переменных, которые используются в программе. В общем случае различные исходные данные порождают различные потоки управления. Поток управления можно представить как нить в программе, на которую нанизаны инструкции, выполняемые микропроцессором. Поэтому часто поток управления также называется *нитью* (thread). В русскоязычной литературе за потоком управления закрепилось название *поток*. Для пояснения понятия потока рассмотрим следующую программу, которая выводит минимальное число из двух целых чисел или сообщение о том, что числа равны.

```
#include <iostream.h>
int main()
{
  int a, b;

  cout << "Input two integers: ";
  cin >> a >> b;
  if (a == b)
  {
    cout << "There is no min." << endl;
    return 0;
  }
```

```
  if (a < b)
    cout << "min = " << a << endl;
  else
    cout << "min = " << b << endl;
  return 0;
}
```

Предположим, что перегруженные операторы ввода-вывода не образуют новых потоков. Тогда в зависимости от входных данных эта программа образует один из трех возможных потоков управления. А именно, если выполняется условие `(a == b)`, то образуется поток:

```
  cout << "Input two integers: ";
  cin >> a >> b;
  if (a == b)
  {
    cout << "There is no min." << endl;
    return 0;
  }
```

Если выполняется условие `(a < b)`, то образуется поток:

```
  cout << "Input two integers: ";
  cin >> a >> b;
  if (a == b)
  if (a < b)
    cout << "min = " << a << endl;
  return 0;
```

Если же выполняется условие `(a > b)`, то образуется поток

```
  cout << "Input two integers: ";
  cin >> a >> b;
  if (a == b)
  if (a < b)
    cout << "min = " << b << endl;
  return 0;
```

Теперь перейдем к классификации программ в зависимости от количества определяемых ими параллельных потоков управления. Будем говорить, что программа является *многопоточной*, если в ней может одновременно существовать несколько потоков. Сами потоки в этом случае называются *параллельными*. Если в программе одновременно может существовать только один поток, то такая программа называется *однопоточной*. Например, сле-

дующая программа, которая просто вычисляет сумму двух чисел, является однопоточной:

```
#include <iostream.h>
int sum(int a, int b)
{
  return a + b;
}
int main()
{
  int a, b;
  int c = 0;
  cout << "Input two integers: ";
  cin >> a >> b;
  c = sum(a, b);
  cout << "Sum = " << c << endl;
  return 0;
}
```

Теперь предположим, что после вызова функции `sum` функция `main` не ждет возвращения значения из функции `sum`, а продолжает выполняться. В этом случае получим программу, состоящую из двух потоков, один из которых определяется функцией `main`, а второй — функцией `sum`. Причем эти потоки независимы, т. к. они не имеют доступа к общим или, другими словами, разделяемым переменным. Правда в этом случае не гарантируется, что поток `main` выведет сумму чисел `a` и `b`, т. к. инструкция вывода значения суммы может отработать раньше, чем поток `sum` вычислит эту сумму.

Из этих рассуждений видно, что для того чтобы отметить функцию, которая порождает новый поток в программе, должна использоваться специальная нотация. В операционных системах Windows для обозначения того, что функция образует поток, используются специальные спецификаторы функции. Такая функция обычно также называется потоком.

2.2. Контекст потока

В общем случае содержимое памяти, к которой поток имеет доступ во время своего исполнения, называется *контекстом потока*. Определим, каким ограничениям на доступ к памяти должны удовлетворять функции, чтобы их можно было безопасно вызывать в параллельных потоках. Для этого рассмотрим следующую функцию:

```
int f(int n)
{
```

```
  if (n > 0)
    --n;
  if (n < 0)
    ++n;
  return n;
}
```

Сколько бы раз эта функция не вызывалась параллельно работающими потоками, она будет корректно изменять значение переменной `n`, т. к. эта переменная является локальной в функции `f`. То есть для каждого нового вызова функции `f` будет создан новый локальный экземпляр переменной `n`. Такая функция `f` называется *безопасной для потоков*. Теперь введем глобальную переменную `n` и изменим нашу функцию следующим образом:

```
int  n;
void g()
{
  if (n > 0)
    --n;
  if (n < 0)
    ++n;
}
```

В этом случае параллельный вызов функции `g` несколькими потоками может дать некорректное изменение значения переменной `n`, т. к. значение этой переменной будет изменяться одновременно несколькими функциями `g`. В этом случае функция `g` не является безопасной для потоков.

Та же проблема встречается и в случае, когда функция использует статические переменные. Для разбора этого случая рассмотрим функцию

```
int count()
{
       static int n = 0;
       ++n;
       return n;
}
```

которая возвращает количество своих вызовов. Если эта функция будет вызвана несколькими параллельно исполняемыми потоками, то нельзя точно определить значение переменной `n`, которое вернет эта функция, т. к. это значение изменяется всеми потоками параллельно.

В общем случае функция называется *повторно входимой* или *реентерабельной* (reentrant или reenterable), если она удовлетворяет следующим требованиям:

- не использует глобальные переменные, значения которых изменяются параллельно исполняемыми потоками;

- ❒ не использует статические переменные, определенные внутри функции;
- ❒ не возвращает указатель на статические данные, определенные внутри функции.

В системном программировании часто также рассматриваются программы в кодах микропроцессора, выполнение которых может прерываться и возобновляться в любой момент времени. Причем одна и та же программа может запускаться прежде, чем завершилось исполнение предыдущего экземпляра этой программы. В этом случае также необходимо, чтобы программный код допускал корректное параллельное выполнение нескольких экземпляров программы. Это условие обеспечивается в том случае, если программа не изменяет свой код во время исполнения. Здесь под кодом подразумеваются как команды, так и данные, принадлежащие программе. Программа в кодах микропроцессора, которая не изменяет свой код, также называется *реентерабельной*.

В дополнение к реентерабельным функциям определяют также функции, безопасные для вызова параллельно исполняемыми потоками. Функция называется *безопасной для потоков*, если она обеспечивает блокировку доступа к ресурсам, которые она использует. Как обеспечить блокирование доступа к ресурсам, рассматривается в *гл. 6, 7*, посвященных синхронизации потоков. Сейчас же только скажем, что в этом случае решается задача взаимного исключения доступа к разделяемым ресурсам, используя примитивы синхронизации.

Очевидно, что если функция не является реентерабельной, то она также не является и безопасной для потоков, т. к. в этом случае несколько потоков разделяют общую память, не блокируя доступ к ней. А память, как уже говорилось, также является системным ресурсом.

2.3. Состояния потока

Как видно из определения, поток описывает динамическое поведение всей программы или какой-либо функции в программе. Для удобства обозначений предположим, что программа является однопоточной. Тогда поток можно рассматривать как пару:

$$\text{поток} = (\text{процессор, программа}).$$

Программа может исполняться процессором только в том случае, если она готова к исполнению. То есть все системные ресурсы, которые необходимы для исполнения этой программы, свободны для использования. Кроме того, для исполнения программы необходимо, чтобы и сам процессор был свободен и готов к исполнению этой программы. Для более формального описания этих ситуаций вводятся понятия "состояние процессора" и "состояние

программы". При этом предполагают, что процессор и программа могут находиться в следующих состояниях.

- Состояния процессора:
  - процессор не выделен для исполнения программы;
  - процессор выделен для исполнения программы.
- Состояния программы:
  - программа не готова к исполнению процессором;
  - программа готова к исполнению процессором.

Для краткости записи введем для этих состояний следующие названия:

- Состояния процессора:
  - "не выделен";
  - "выделен".
- Состояния программы:
  - "не готова";
  - "готова".

Тогда мы можем определить *состояние потока* как пару состояний:

состояние потока = (состояние процессора, состояние программы).

Перечислив различные комбинации состояний процессора и программы, можно описать все возможные состояния потока. Введем для состояний потока следующие названия:

- поток блокирован = ("не выделен", "не готова");
- поток готов к выполнению = ("не выделен", "готова");
- поток выполняется = ("выделен", "готова");

Будем считать, что состояние ("выделен", "не готова") является недостижимым для потока. То есть программе, не готовой к исполнению, процессор не выделяется. Более кратко эти состояния потока будем просто обозначать словами: "блокирован", "готов" и "выполняется". Для полноты картины нужно ввести для потоков еще два состояния: "новый" и "завершен", которые описывают соответственно поток, еще не начавший свою работу, и поток, завершивший свою работу. Тогда диаграмма возможных переходов потока из состояния в состояние может быть изображена, как это показано на рис. 2.1.

В результате мы получили простейшую диаграмму переходов потока из состояния в состояние. Сами переходы потока из состояния в состояние, которые на диаграмме обозначаются дугами, описывают некоторые операции

над потоком. Названия этих операций указаны рядом со стрелками. Кратко опишем эти операции.

- Операция `Create` выполняется потоком, который создает новый поток из функции. Эта операция переводит поток из состояния "новый" в состояние "готов".
- Операция `Exit` выполняется самим исполняемым потоком в случае его завершения. Эта операция переводит поток из состояния "выполняется" в состояние "завершен".

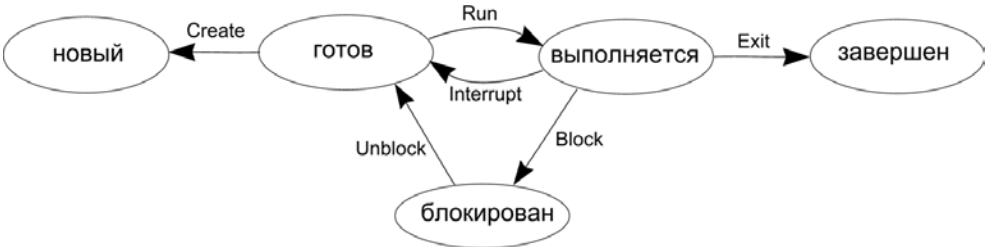

Рис. 2.1. Модель пяти состояний потока.

Оставшиеся четыре операции выполняются операционной системой.

- Операция `Run` запускает готовый поток на выполнение, т. е. выделяет ему процессорное время. Эта операция переводит поток из состояния "готов" в состояние "выполняется". Поток получает процессорное время в том случае, если подошла его очередь к процессору на обслуживание.
- Операция `Interrupt` задерживает исполнение потока и переводит его из состояния "выполняется" в состояние "готов". Эта операция выполняется над потоком в том случае, если истекло процессорное время, выделенное потоку на исполнение, или исполнение потока прервано по каким-либо другим причинам.
- Операция `Block` блокирует исполнение потока, т. е. переводит его из состояния "выполняется" в состояние "блокирован". Эта операция выполняется над потоком в том случае, если он ждет наступления некоторого события, например, завершения операции ввода-вывода или освобождения ресурса.
- Операция `Unblock` разблокирует поток, т. е. переводит его из состояния "блокирован" в состояние "готов". Эта операция выполняется над потоком в том случае, если событие, ожидаемое потоком, наступило.

Разрешим потокам также выполнять операции друг над другом. Для этого введем операции `Suspend` и `Resume`.

- Операция `Suspend` приостанавливает исполнение потока.
- Операция `Resume` возобновляет исполнение потока.

Используя эти операции, один поток может соответственно приостановить или возобновить исполнение другого потока независимо от того, в каком состоянии этот последний поток находится. Впрочем, заметим, что поток может приостановить и свое исполнение. Если над потоком выполнена операция `Suspend`, то будем говорить, что поток находится в *приостановленном* или *подвешенном состоянии*. Кратко будем говорить, что в этом случае поток "подвешен". Дополним диаграмму состояний потока, изображенную на рис. 2.1, этими новыми операциями и состояниями. Получим более полную диаграмму состояний потока, которая показана на рис. 2.2.

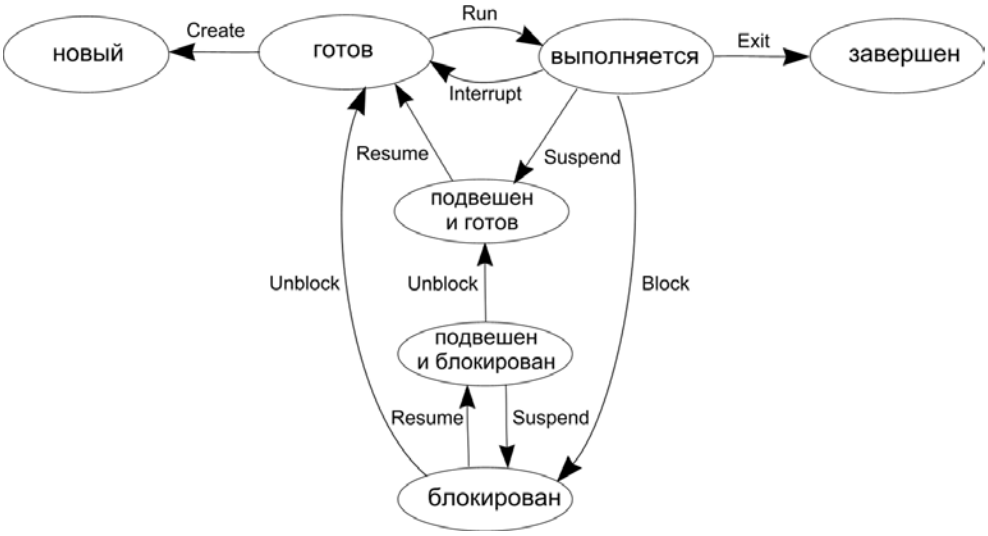

Рис. 2.2. Модель семи состояний потока

Теперь разрешим потоку выполнять операции над самим собой. Для этого введем операцию `Sleep`.

- Операция `Sleep` позволяет потоку приостановить свое исполнение на некоторый интервал времени или, другими словами, заснуть.

Разбудить поток должна операционная система по истечении заданного интервала времени, используя операцию `Wakeup`. Если поток выполнил операцию `Sleep`, то будем говорить, что он перешел в *сонное состояние* или "спит".

- Операция `Wakeup` позволяет операционной системе разбудить поток.

В результате можно построить полную диаграмму состояний потока, которая и приведена на рис. 2.3.

В заключение этого параграфа скажем, что в конкретных операционных системах для работы с потоками могут быть определены и другие состояния,

а также операции, которые переводят потоки в эти состояния. В *гл. 3* будет рассмотрена модель состояний потока в операционной системе Windows 2000.

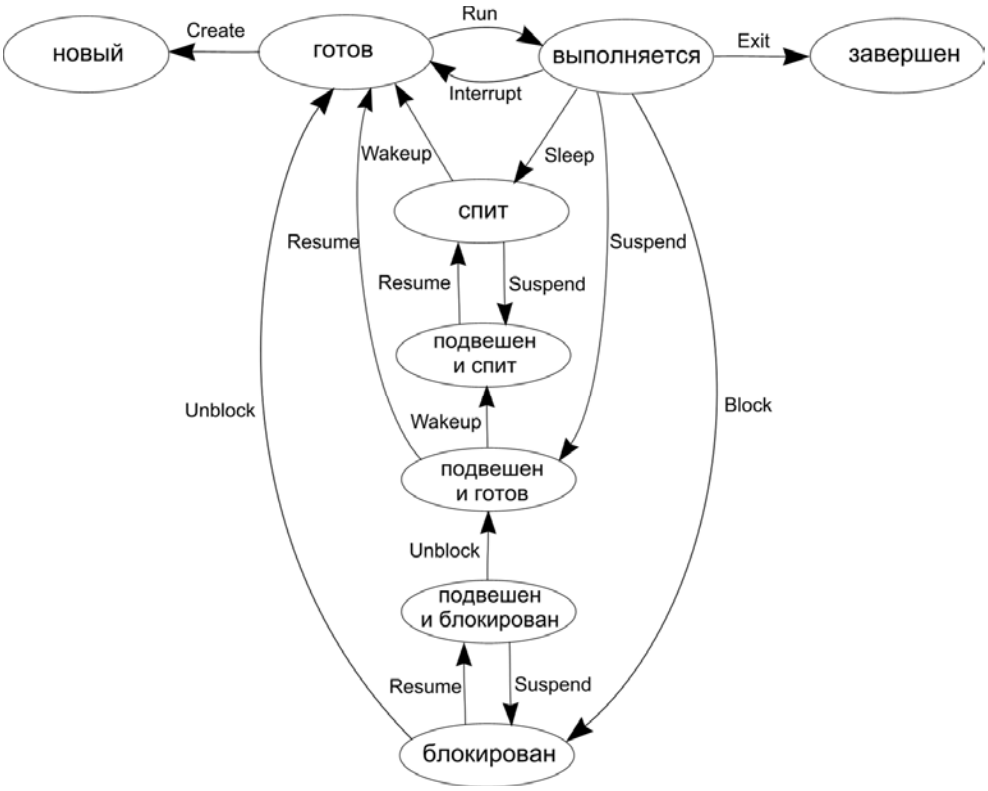

Рис. 2.3. Модель девяти состояний потока

2.4. Диспетчеризация и планирование потоков

В однопрограммной операционной системе одновременно может выполняться только один поток, которому доступны все ресурсы компьютера. Поэтому блокировка потока может происходить только в случаях ожидания этим потоком события, отмечающего завершение операций ввода-вывода. Недостатком однопрограммных операционных систем является их низкая производительность, т. к. процессор простаивает, если поток блокирован.

В мультипрограммных операционных системах одновременно могут существовать несколько потоков, что повышает производительность компьютера. Однако в этом случае требуется некоторая дисциплина обслуживания этих