2

The Process and the Kernel

2.1 Introduction

The principal function of an operating system is to provide an execution environment in which user programs *(applications)* may run. This involves defining a basic framework for program execution, and providing a set of services—such as file management and I/O—and an interface to these services. The UNIX system presents a rich and versatile programming interface [Kern 84] that can efficiently support a variety of applications. This chapter describes the main components of the UNIX systems and how they interact to provide a powerful programming paradigm.

There are several different UNIX variants. Some of the important ones are the System V releases from AT&T (SVR4, the latest System V release, is now owned by Novell), the BSD releases from the University of California at Berkeley, OSF/1 from the Open Software Foundation, and SunOS and Solaris from Sun Microsystems. *This chapter describes the kernel and process architecture of traditional UNIX systems,* that is, those based on SVR2 [Bach 86], SVR3 [AT&T 87], 4.3BSD [Leff 89], or earlier versions. Modern UNIX variants such as SVR4, OSF/1, 4.4BSD, and Solaris 2.*x* differ significantly from this basic model; the subsequent chapters explore the modern releases in detail.

The UNIX application environment contains one fundamental abstraction—the *process*. In traditional UNIX systems, the process executes a single sequence of instructions in an *address space*. The address space of a process comprises the set of memory locations that the process may reference or access. The *control point* of the process tracks the sequence of instructions, using a hardware register typically called the *program counter (PC)*. Many newer UNIX releases support

multiple control points (called *threads* [IEEE 94]), and hence multiple instruction sequences, within a single process.

The UNIX system is a multiprogramming environment, i.e., several processes are active in the system concurrently. To these processes, the system provides some features of a *virtual machine*. In a pure virtual machine architecture the operating system gives each process the illusion that it is the only process on the machine. The programmer writes an application as if only its code were running on the system. In UNIX systems each process has its own registers and memory, but must rely on the operating system for I/O and device control.

The process address space is virtual,[1] and normally only part of it corresponds to locations in physical memory. The kernel stores the contents of the process address space in various storage objects, including physical memory, on-disk files, and specially reserved *swap areas* on local or remote disks. The memory management subsystem of the kernel shuffles *pages* (fixed-size chunks) of process memory between these objects as convenient.

Each process also has a set of registers, which correspond to real, hardware registers. There are many active processes in the system, but only one set of hardware registers. The kernel keeps the registers of the currently running process in the hardware registers and saves those of other processes in per-process data structures.

Processes contend for the various resources of the system, such as the processor (also known as the Central Processing Unit or *CPU*), memory, and peripheral devices. An operating system must act as a resource manager, distributing the system resources optimally. A process that cannot acquire a resource it needs must *block* (suspend execution) until that resource becomes available. Since the CPU is one such resource, only one process can actually run at a time on a uniprocessor system. The rest of the processes are blocked, waiting for either the CPU or other resources. The kernel provides an illusion of concurrency by allowing one process to have the CPU for a brief period of time (called a *quantum*, usually about 10 milliseconds), then switching to another. In this way each process receives some CPU time and makes progress. This method of operation is known as *time-slicing*.

From another perspective, the computer provides several facilities to the user, such as the processor, disks, terminals, and printers. Application programmers do not wish to be concerned with the low-level details of the functionality and architecture of these components. The operating system assumes complete control of these devices and offers a high-level, abstract programming interface that applications can use to access these components. It hides all the details of the hardware, greatly simplifying the work of the programmer.[2] By centralizing all control of the devices, it also provides additional facilities such as access synchronization (if two users want the same device at the same time) and error recovery. The *application programming interface (API)* defines the semantics of all interactions between user programs and the operating system.

---

[1] There are some UNIX systems that do not use virtual memory. These include the earliest UNIX releases (the first virtual memory systems appeared in the late 1970s—see Section 1.1.4) and some real-time UNIX variants. This book deals only with UNIX systems that have virtual memory.

[2] The UNIX system takes this too far in some cases; for example, its treatment of tape drives as character streams makes it difficult for applications to properly handle errors and exceptional cases. The tape interface is inherently record-based and does not fit nicely with the UNIX device framework [Allm 87].

We have already started referring to the operating system as an entity that *does things*. What exactly is this entity? On one hand, an operating system is a program (often called the *kernel*) that controls the hardware and creates, destroys, and controls all processes (see Figure 2-1). From a broader perspective, an operating system includes not just the kernel, but also a host of other programs and utilities (such as the shells, editors, compilers, and programs like *date, ls,* and *who*) that together provide a useful work environment. Obviously, the kernel alone is of limited use, and users purchasing the UNIX system expect many of these other programs to come with it. The kernel, however, is special in many ways. It defines the programming interface to the system. It is the only indispensable program, without which nothing can run. While several shells or editors may run concurrently, only a single kernel may be loaded at a time. This book is devoted to studying the kernel, and when it mentions the *operating system,* or UNIX, it means the kernel, unless specified otherwise.

To rephrase the earlier question, "What exactly is the kernel?" Is it a process, or something distinct from all processes? The kernel is a special program that runs directly on the hardware. It implements the process model and other system services. It resides on disk in a file typically called **/vmunix** or **/unix** (depending on the UNIX vendor). When the system starts up, it loads the kernel from disk using a special procedure called *bootstrapping*. The kernel initializes the system and sets up the environment for running processes. It then creates a few initial processes, which in turn create other processes. Once loaded, the kernel remains in memory until the system is shut down. It manages the processes and provides various services to them.

The UNIX operating system provides functionality in four ways:

- User processes explicitly request services from the kernel through the *system call interface* (see Figure 2-1), the central component of the UNIX API. The kernel executes these requests on behalf of the calling process.
- Some unusual actions of a process, such as attempting to divide by zero, or overflowing the user stack, cause *hardware exceptions*. Exceptions require kernel intervention, and the kernel handles them on behalf of the process.

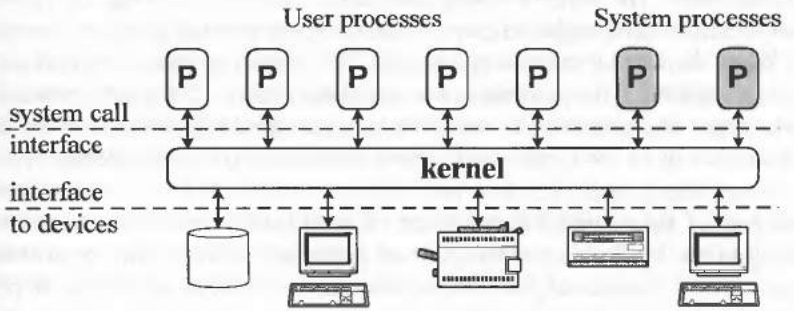

Figure 2-1. The kernel interacts with processes and devices.

- The kernel handles hardware *interrupts* from peripheral devices. Devices use the interrupt mechanism to notify the kernel of I/O completion and status changes. The kernel treats interrupts as global events, unrelated to any specific process.
- A set of special system processes, such as the *swapper* and the *pagedaemon,* perform system-wide tasks such as controlling the number of active processes or maintaining a pool of free memory.

The following sections describe these different mechanisms and define the notion of the execution context of a process.

2.2 Mode, Space, and Context

In order to run UNIX, the computer hardware must provide at least two different *modes of execution*—a more privileged *kernel mode,* and a less privileged *user mode*. As you might expect, user programs execute in user mode, and kernel functions execute in kernel mode. The kernel protects some parts of the address space from user-mode access. Moreover, certain privileged machine instructions, such as those that manipulate memory management registers, may only be executed in kernel mode.

Many computers have more than two execution modes. The Intel 80*x*86 architecture, for example, provides four *rings of execution*—the innermost being the most privileged. *UNIX, however, uses only two of these rings*. The main reason for having different execution modes is for protection. Since user processes run in the less privileged mode, they cannot accidentally or maliciously corrupt another process or the kernel. The damage from program errors is localized, and usually does not affect other activity or processes in the system.

Most UNIX implementations use *virtual memory*. In a virtual memory system the addresses used by a program do not refer directly to locations in physical memory. Each process has its own *virtual address space,* and references to virtual memory addresses are translated to physical memory locations using a set of *address translation maps*. Many systems implement these maps as *page tables,* with one entry for each *page* (a fixed-size unit of memory allocation and protection) of the process address space. The *memory management unit (MMU)* of the computer typically has a set of registers that identifies the translation maps of the currently running process (also called the *current process*). When the current process yields the CPU to another process (a *context switch*), the kernel loads these registers with pointers to the translation maps of the new process. The MMU registers are privileged and may only be accessed in kernel mode. This ensures that a process can only refer to addresses in its own space and cannot access or modify the address space of another process.

A fixed part of the virtual address space of each process maps the kernel text and data structures. This portion, known as *system space* or *kernel space,* may only be accessed in kernel mode. There is only one instance of the kernel running in the system, and hence all processes map a single kernel address space. The kernel maintains some global data structures and some per-process objects. The latter contain information that enables the kernel to access the address space of any process. The kernel can directly access the address space of the current process, since

the MMU registers have the necessary information. Occasionally, the kernel must access the address space of a process other than the current one. It does so indirectly, using special, temporary mappings.

While the kernel is shared by all processes, system space is protected from user-mode access. Processes cannot directly access the kernel, and must instead use the system call interface. When a process makes a *system call,* it executes a special sequence of instructions to put the system in kernel mode (this is called a *mode switch*) and transfer control to the kernel, which handles the operation on behalf of the process. After the system call is complete, the kernel executes another set of instructions that returns the system to user mode (another mode switch) and transfers control back to the process. The system call interface is described further in Section 2.4.1.

There are two important per-process objects that, while managed by the kernel, are often implemented as part of the process address space. These are the *u area* (also called the *user area*) and the *kernel stack*. The u area is a data structure that contains information about a process of interest to the kernel, such as a table of files opened by the process, identification information, and saved values of the process registers when the process is not running. The process should not be allowed to change this information arbitrarily, and hence the u area is protected from user-mode access. (Some implementations allow the process to read, but not modify, the u area.)

The UNIX kernel is *re-entrant,* meaning that several processes may be involved in kernel activity concurrently. In fact, they may even be executing the same routine in parallel. (Of course, only one process can actually run at a time; the others are blocked or waiting to run.) Hence each process needs its own private kernel stack, to keep track of its function call sequence when executing in the kernel. Many UNIX implementations allocate the kernel stack in the address space of each process, but do not allow user-mode access to it. Conceptually, both the u area and the kernel stack, while being per-process entities in the process space, are *owned* by the kernel.

Another important concept is the *execution context*. Kernel functions may execute either in *process context* or in *system context*. In process context, the kernel acts on behalf of the current process (for instance, while executing a system call), and may access and modify the address space, u area, and kernel stack of this process. Moreover, the kernel may block the current process if it must wait for a resource or device activity.

The kernel must also perform certain system-wide tasks such as responding to device interrupts and recomputing process priorities. Such tasks are not performed on behalf of a particular process, and hence are handled in system context (also called *interrupt context*). When running in system context, the kernel may not access the address space, u area, or kernel stack of the current process. The kernel may not block when executing in system context, since that would block an innocent process. In some situations there may not even be a current process, for example, when all processes are blocked awaiting I/O completion.

This far, we have noted the distinctions between user and kernel mode, process and system space, and process and system context. Figure 2-2 summarizes these notions. User code runs in user mode and process context, and can access only the process space. System calls and exceptions are handled in kernel mode but in process context, and may access process and system space. Interrupts are handled in kernel mode and system context, and must only access system space.

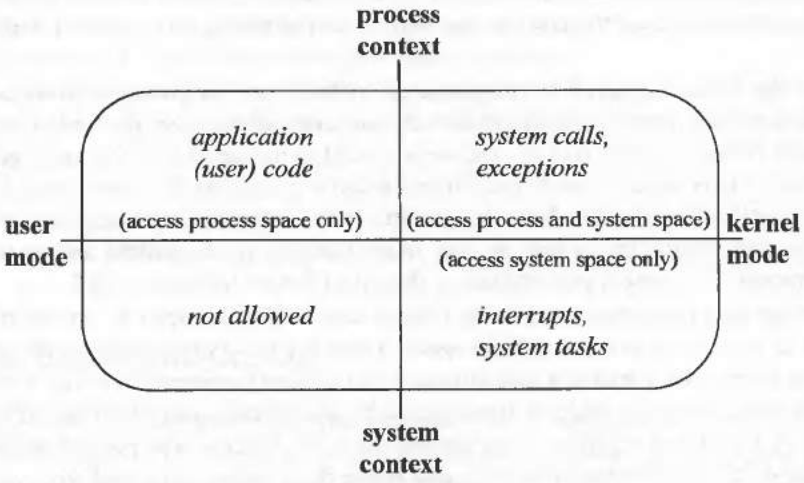

Figure 2-2. Execution mode and context.

2.3 The Process Abstraction

What exactly is a UNIX process? One oft-quoted answer is, "A process is an instance of a running program." Going beyond perfunctory definitions, it is useful to describe various properties of the process. A process is an entity that runs a program and provides an execution environment for it. It comprises an address space and a control point. The process is the fundamental scheduling entity—only one process runs on the CPU at a time. In addition, the process contends for and owns various system resources such as devices and memory. It also requests services from the system, which the kernel performs on its behalf.

The process has a definite lifetime—most processes are created by a *fork* or *vfork* system call and run until they terminate by calling *exit*. During its lifetime, a process may run one or many programs (usually one at a time). It invokes the *exec* system call to run a new program.

UNIX processes have a well-defined hierarchy. Each process has one *parent*, and may have one or more *child* processes. The process hierarchy can be described by an inverted tree, with the *init* process at the top. The *init* process (so named because it executes the program **/etc/init**) is the first user process created when the system boots. It is the ancestor of all user processes. A few system processes, such as the *swapper* and the *pagedaemon* (also called the *pageout daemon*), are created during the bootstrapping sequence and are not descendants of *init*. If, when a process terminates, it has any active child processes, they become *orphans* and are inherited by *init*.

2.3.1 Process State

At all times, UNIX processes are in some well-defined *state*. They move from one state to another in response to various events. Figure 2-3 describes the important process states in UNIX and the events that cause *state transitions*.

The *fork* system call creates a new process, which begins life in the *initial* (also called *idle*) state. When the process is fully created, *fork* moves it to the *ready to run* state, where it must wait to be scheduled. Eventually, the kernel selects it for execution, and initiates a context switch. This invokes a kernel routine (typically called `swtch()`) that loads the hardware context of the process (see Section 2.3.2) into the system registers, and transfers control to the process. From this point, the new process behaves like any other process, and undergoes state transitions as described below.

A process running in user mode enters kernel mode as a result of a system call or an interrupt, and returns to user mode when that completes.[3] While executing a system call, the process may need to wait for an event or for a resource that is currently unavailable. It does so by calling

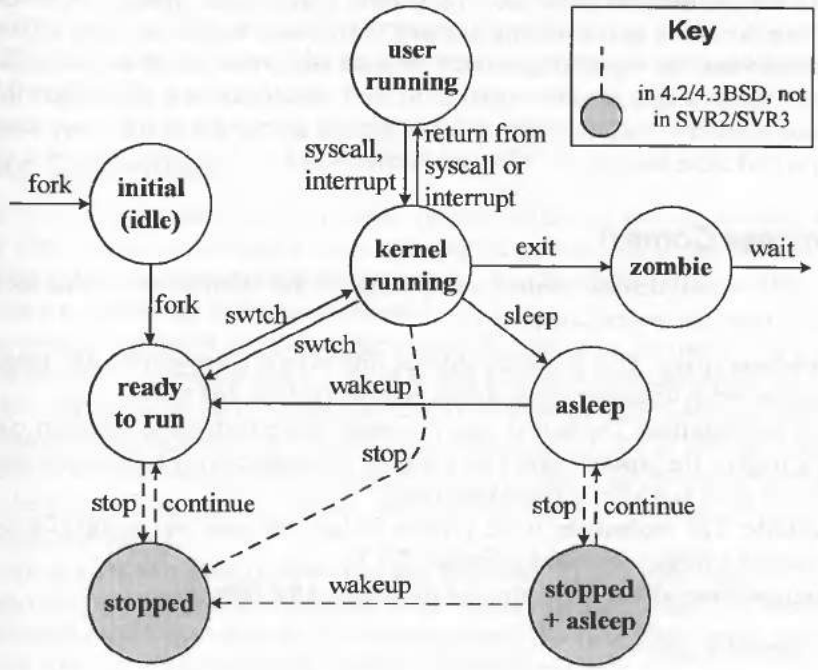

Figure 2-3. Process states and state transitions.

[3] Interrupts can also occur when the system is in kernel mode. In this case, the system will remain in kernel mode after the handler completes.

`sleep()`, which puts the process on a queue of sleeping processes, and changes its state to *asleep*. When the event occurs or the resource becomes available, the kernel wakes up the process, which now becomes *ready to run* and waits to be scheduled.

When a process is scheduled to run, it initially runs in kernel mode (*kernel running* state), where it completes the context switch. Its next transition depends on what it was doing before it was switched out. If the process was newly created or was executing user code (and was descheduled to let a higher priority process run), it returns immediately to user mode. If it was blocked for a resource while executing a system call, it resumes execution of the system call in kernel mode.

Finally, the process terminates by calling the *exit* system call, or because of a *signal* (signals are notifications issued by the kernel—see Chapter 4). In either case, the kernel releases all the resources of the process, except for the *exit status* and *resource usage* information, and leaves the process in the *zombie* state. The process remains in this state until its parent calls *wait* (or one of its variants), which destroys the process and returns the exit status to the parent (see Section 2.8.6).

4BSD defines some additional states that are not supported in SVR2 or SVR3. A process is *stopped*, or *suspended*, by a *stop* signal (`SIGSTOP`, `SIGTSTP`, `SIGTTIN`, or `SIGTTOU`). Unlike other signals, which are handled only when the process runs, a *stop* signal changes the process state immediately. If the process is in the running or ready to run state, its state changes to stopped. If the process is asleep when this signal is generated, its state changes to asleep and stopped. A stopped process may be resumed by a *continue* signal (`SIGCONT`), which returns it to the ready to run state. If the process was stopped as well as asleep, `SIGCONT` returns the process to the asleep state. System V UNIX incorporated these features in SVR4 (see Section 4.5).[4]

2.3.2 Process Context

Each process has a well-defined context, comprising all the information needed to describe the process. This context has several components:

- **User address space:** This is usually divided into several components—the program text (executable code), data, user stack, shared memory regions, and so on.
- **Control information:** The kernel uses two main data structures to maintain control information about the process—the `u` area and the `proc` structure. Each process also has its own kernel stack and address translation maps.
- **Credentials:** The credentials of the process include the user and group IDs associated with it, and are further described in Section 2.3.3.
- **Environment variables:** These are a set of strings of the form

  ```
  variable=value
  ```

  which are inherited from the parent. Most UNIX systems store these strings at the bottom of the user stack. The standard user library provides functions to add, delete, or modify these variables, and to translate the variable and return its value. When invoking a new

[4] SVR3 provides a *stopped* state for the process solely for the purpose of *process tracing* (see Section 6.2.4). When a traced process receives *any* signal, it enters the stopped state, and the kernel awakens its parent.

program, the caller may ask *exec* to retain the original environment or provide a new set of variables to be used instead.

- **Hardware context:** This includes the contents of the general-purpose registers, and of a set of special system registers. The system registers include:
  - The *program counter (PC),* which holds the address of the next instruction to execute.
  - The *stack pointer (SP),* which contains the address of the uppermost element of the stack.[5]
  - The *processor status word (PSW),* which has several status bits containing information about the system state, such as current and previous execution modes, current and previous interrupt priority levels, and overflow and carry bits.
  - *Memory management registers,* which map the address translation tables of the process.
  - Floating point unit (FPU) registers.

The machine registers contain the hardware context of the currently running process. When a context switch occurs, these registers are saved in a special part of the u area (called the *process control block, or PCB*) of the current process. The kernel selects a new process to run and loads the hardware context from its PCB.

2.3.3 User Credentials

Every user in the system is identified by a unique number called the *user ID,* or *UID.* The system administrator also creates several user groups, each with a unique *user group ID,* or *GID.* These identifiers affect file ownership and access permissions, and the ability to signal other processes. These attributes are collectively called the *credentials.*

The system recognizes a privileged user called the *superuser* (normally, this user logs in with the name *root*). The superuser has a UID of 0, and GID of 1. The superuser has many privileges denied to ordinary users. He or she may access files owned by others, regardless of protection settings, and may also execute a number of privileged system calls (such as *mknod,* which creates a special device file). Many modern UNIX systems such as SVR4.1/ES support *enhanced security* mechanisms [Sale 92]. These systems replace the single superuser abstraction with separate privileges for different operations.

Each process has two pairs of IDs—real and effective. When a user logs in, the login program sets both pairs to the UID and GID specified in the *password database* (the **/etc/passwd** file, or some distributed mechanism such as Sun Microsystems' *Network Information Service (NIS)*). When a process *forks,* the child inherits its credentials from the parent.

The *effective UID* and *effective GID* affect file creation and access. During file creation, the kernel sets the owner attributes of the file to the effective UID and GID of the creating process. During file access, the kernel uses the effective UID and GID of the process to determine whether it

[5] Or lowermost, on machines where the stack grows downward. Also, on some systems, the stack pointer contains the address at which the next item can be pushed onto the stack.

can access the file (see Section 8.2.2 for more details). The *real UID* and *real GID* identify the real owner of the process and affect the permissions for sending signals. A process without superuser privileges can signal another process only if the sender's real or effective UID matches the real UID of the receiver.

There are three system calls that can change the credentials. If a process calls *exec* to run a program installed in *suid mode* (see Section 8.2.2), the kernel changes the effective UID of the process to that of the owner of the file. Likewise, if the program is installed in *sgid mode,* the kernel changes the effective GID of the calling process.

UNIX provides this feature to grant special privileges to users for particular tasks. The classic example is the *passwd* program, which allows the user to modify his own password. This program must write to the password database, which users should not be allowed to directly modify (to prevent them from changing passwords of other users). Hence the *passwd* program is owned by the superuser and has its SUID bit set. This allows the user to gain superuser privileges while running the *passwd* program.

A user can also change his credentials by calling *setuid* or *setgid.* The superuser can invoke these system calls to change both the real and effective UID or GID. Ordinary users can use this call only to change their effective UID or GID back to the real ones.

There are some differences in the treatment of credentials in System V and BSD UNIX. SVR3 also maintains a *saved UID* and *saved GID,* which are the values of the effective UID and GID prior to calling *exec.* The *setuid* and *setgid* calls can also restore the effective IDs to the saved values. While 4.3BSD does not support this feature, it allows a user to belong to a set of *supplemental groups* (using the *setgroups* system call). While files created by the user belong to his or her *primary group,* the user can access files belonging either to the principal or to a supplemental group (provided the file allows access to group members).

SVR4 incorporates all the above features. It supports supplemental groups, and maintains the saved UID and GID across *exec.*

2.3.4 The u Area and the proc Structure

The control information about a process is maintained in two per-process data structures—the u area and the `proc` structure. In many implementations, the kernel has a fixed-size array of `proc` structures called the *process table.* The size of this array places a hard limit on the maximum number of processes that can exist at a time. Newer releases such as SVR4 allow dynamic allocation of `proc` structures, but have a fixed-size array of pointers to them. Since the `proc` structure is in system space, it is visible to the kernel at all times, even when the process is not running.

The u area, or user area, is part of the process space, i.e., it is mapped and visible only when the process is running. On many implementations, the u area is always mapped at the same fixed virtual address in each process, which the kernel references simply through the variable `u`. One of the tasks of the context switch is to reset this mapping, so that kernel references to u are translated to the physical location of the new u area.

Occasionally, the kernel may need to access the u area of another process. This is possible, but must be done indirectly using a special set of mappings. These differences in access semantics govern what information is stored in the `proc` structure and what is stored in the u area. The u area

contains data that is needed only when the process is running. The `proc` structure contains information that may be needed even when the process is not running.

The major fields in the u area include:

- The process control block—stores the saved hardware context when the process is not running.
- A pointer to the `proc` structure for this process.
- The real and effective UID and GID.[6]
- Arguments to, and return values or error status from, the current system call.
- Signal handlers and related information (see Chapter 4).
- Information from the program header, such as text, data, and stack sizes and other memory management information.
- Open file descriptor table (see Section 8.2.3). Modern UNIX systems such as SVR4 dynamically extend this table as necessary.
- Pointers to *vnodes* of the *current directory* and the *controlling terminal*. Vnodes represent file system objects and are further described in Section 8.7.
- CPU usage statistics, profiling information, disk quotas, and resource limits.
- In many implementations, the per-process kernel stack is part of the u area.

The major fields in the `proc` structure include:

- Identification: Each process has a unique *process ID (PID)* and belongs to a specific *process group*. Newer releases also assign a *session ID* to each process.
- Location of the kernel address map for the u area of this process.
- The current process state.
- Forward and backward pointers to link the process onto a scheduler queue or, for a blocked process, a sleep queue.
- Sleep channel for blocked processes (see Section 7.2.3).
- Scheduling priority and related information (see Chapter 5).
- Signal handling information: masks of signals that are ignored, blocked, posted, and handled (see Chapter 4).
- Memory management information.
- Pointers to link this structure on lists of active, free, or zombie processes.
- Miscellaneous flags.
- Pointers to keep the structure on a *hash queue* based on its PID.
- Hierarchy information, describing the relationship of this process to others.

Figure 2-4 illustrates the process relationships in 4.3BSD UNIX. The fields that describe the hierarchy are `p_pid` (process ID), `p_ppid` (parent process ID), `p_pptr` (pointer to the parent's `proc` structure), `p_cptr` (pointer to the oldest child), `p_ysptr` (pointer to next younger sibling), and `p_osptr` (pointer to next older sibling).

---

[6] Modern UNIX systems such as SVR4 store user credentials in a dynamically allocated, reference-counted data structure, and keep a pointer to it in the `proc` structure. Section 8.10.7 discusses this arrangement further.

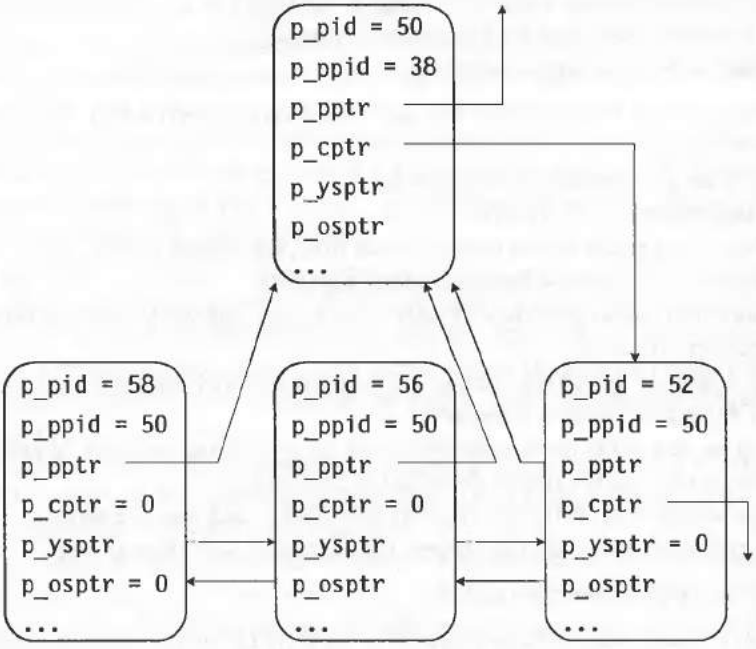

Figure 2-4. A typical process hierarchy in 4.3BSD UNIX.

Many modern UNIX variants have modified the process abstraction to support several threads of control in a single process. This notion is explained in detail in Chapter 3.

2.4 Executing in Kernel Mode

There are three types of events that cause the system to enter kernel mode—device interrupts, exceptions, and traps or software interrupts. In each case, when the kernel receives control, it consults a *dispatch table,* which contains addresses of low-level routines that handle these events. Before calling the appropriate routine, the kernel saves some state of the interrupted process (such as its program counter and the processor status word) on its kernel stack. When the routine completes, the kernel restores the state of the process and changes the execution mode back to its previous value (an interrupt could have occurred when the system was already in kernel mode, in which case it would remain in kernel mode after the handler returns).

It is important to distinguish between interrupts and exceptions. Interrupts are asynchronous events caused by peripheral devices such as disks, terminals, or the hardware clock. Since interrupts are not caused by the current process, they must be serviced in system context and may not access the process address space or u area. For the same reason, they must not block, since that would block an arbitrary process. Exceptions are synchronous to the process and are caused by events related to the process itself, such as attempting to divide by zero or accessing an illegal address. The

exception handler therefore runs in process context; it may access the address space and u area of the process and block if necessary. Software interrupts, or traps, occur when a process executes a special instruction, such as in system calls, and are handled synchronously in process context.

2.4.1 The System Call Interface

The set of system calls defines the programming interface offered by the kernel to user processes. The standard C library, linked by default with all user programs, contains a *wrapper routine* for each system call. When a user program makes a system call, the corresponding wrapper routine is invoked. This routine pushes the system call number (which identifies the particular system call to the kernel) onto the user stack and then invokes a special *trap* instruction. The actual name of the instruction is machine-specific (for example, `syscall` on the MIPS R3000, `chmk` on the VAX-11, or `trap` on the Motorola 680x0). The function of this instruction is to change the execution mode to kernel and to transfer control to the system call handler defined in the dispatch table. This handler, typically called `syscall()`, is the starting point of all system call processing in the kernel.

The system call executes in kernel mode, but in process context. It thus has access to the process address space and the u area. Since it runs in kernel mode, it uses the kernel stack of the calling process. `syscall()` copies the arguments of the system call from the user stack to the u area and saves the hardware context of the process on the kernel stack. It then uses the system call number to index into a system call dispatch vector (usually called `sysent[]`) to determine which kernel function to call to execute that particular system call. When that function returns, `syscall()` sets the return values or error status in the appropriate registers, restores the hardware context, and returns to user mode, transferring control back to the library routine.

2.4.2 Interrupt Handling

The primary function of interrupts on a machine is to allow peripheral devices to interact with the CPU, to inform it of task completion, error conditions, or other events that require urgent attention. Such interrupts are generated asynchronous to regular system activity (i.e., the system does not know at what point in the instruction stream the interrupt will occur) and are usually unrelated to any specific process. The function that is invoked to service an interrupt is called the *interrupt handler* or *interrupt service routine*. The handler runs in kernel mode and system context. Since the process that was interrupted usually bears no relation to the interrupt, the handler must be careful not to access the process context. For the same reason, interrupt handlers are not permitted to block.

There is, however, a small impact on the interrupted process. The time used to service the interrupt is charged to the time slice of this process, even though the activity was unrelated to the process. Also, the clock interrupt handler charges the clock tick (the time between two clock interrupts) to the current process, and thus needs to access its `proc` structure. It is important to note that the process context is not explicitly protected from access by interrupt handlers. An incorrectly written handler has the power to corrupt any part of the process address space.

The kernel also supports the notion of software interrupts or traps, which can be triggered by executing specific instructions. Such interrupts are used, for example, to trigger a context switch or

Table 2-1. Setting the interrupt priority level in 4.3BSD and SVR4

4.3BSD	SVR4	Purpose
spl0	spl0 or splbase	enable all interrupts
splsoftclock	spltimeout	block functions scheduled by timers
splnet		block network protocol processing
	splstr	block STREAMS interrupts
spltty	spltty	block terminal interrupts
splbio	spldisk	block disk interrupts
splimp		block network device interrupts
splclock		block hardware clock interrupt
splhigh	spl7 or splhi	disable all interrupts
splx	splx	restore *ipl* to previously saved value

schedule low-priority clock-related tasks. While these interrupts are synchronous to normal system activity, they are handled just like normal interrupts.

Since there are several different events that may cause interrupts, one interrupt may occur while another is being serviced. UNIX systems recognize the need to prioritize different kinds of interrupts and allow high-priority interrupts to preempt the servicing of low-priority interrupts. For example, the hardware clock interrupt must take precedence over a network interrupt, since the latter may require a large amount of processing, spanning several clock ticks.

UNIX systems assign an *interrupt priority level (ipl)* to each type of interrupt. Early UNIX implementations had *ipl*s in the range 0–7. In BSD, this was expanded to 0–31. The *processor status register* typically has bit-fields that store the current (and perhaps previous) *ipl*.[7] Normal kernel and user processing occurs at the base *ipl*. The number of interrupt priorities varies both across different UNIX variants and across different hardware architectures. On some systems, *ipl 0* is the lowest priority, while on others it is the highest. To make things easier for kernel and device driver developers, UNIX systems provide a set of macros to block and unblock interrupts. However, different UNIX variants use different macros for similar purposes. Table 2-1 lists some of the macros used in 4.3BSD and in SVR4.

When an interrupt occurs, if its *ipl* is higher than the current *ipl*, the current processing is suspended and the handler for the new interrupt is invoked. The handler begins execution at the new *ipl*. When the handler completes, the *ipl* is lowered to its previous value (which is obtained from the old processor status word saved on the interrupt stack), and the kernel resumes execution of the interrupted process. If the kernel receives an interrupt of *ipl* lower than or equal to the current *ipl*, that interrupt is not handled immediately, but is stored in a saved interrupt register. When the *ipl* drops sufficiently, the saved interrupt will be handled. This is described in Figure 2-5.

The *ipl*s are compared and set in hardware in a machine-dependent way. UNIX also provides the kernel with mechanisms to explicitly check or set the *ipl*. For instance, the kernel may raise the *ipl* to block interrupts while executing some critical code. This is discussed further in Section 2.5.2.

---

[7] Some processors, such as the Intel 80x86, do not support interrupt priorities in hardware. On these systems, the operating system must implement *ipl*s in software. The exercises explore this problem further.

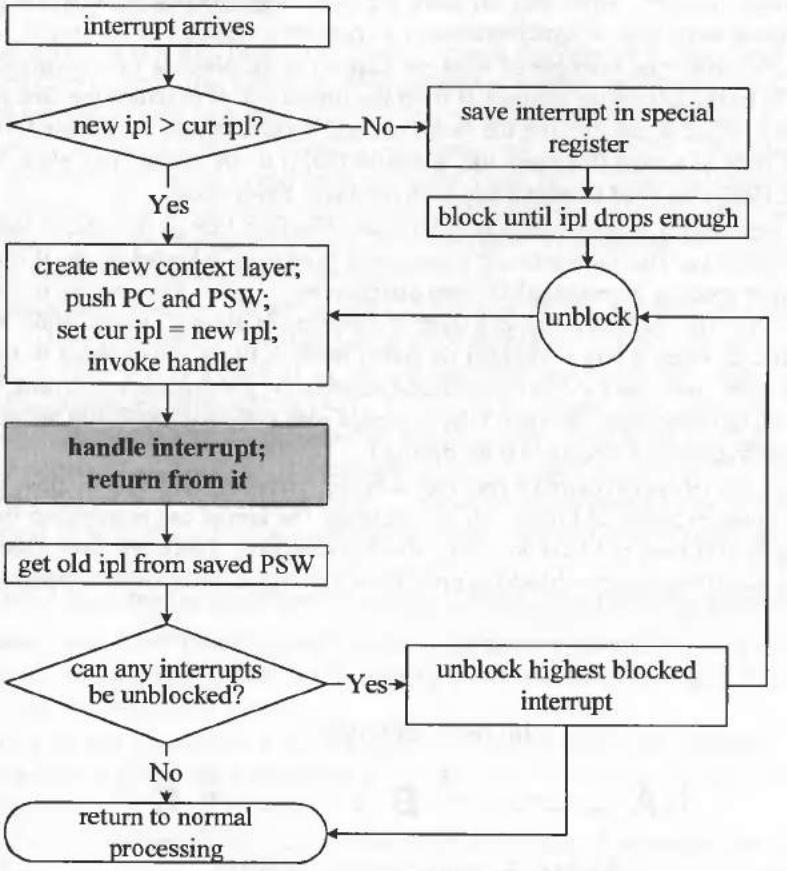

Figure 2-5. Interrupt handling.

Some machines provide a separate global *interrupt stack* used by all the handlers. In machines without an interrupt stack, handlers run on the kernel stack of the current process. They must ensure that the rest of the kernel stack is insulated from the handler. The kernel implements this by pushing a *context layer* on the kernel stack before calling the handler. This context layer, like a stack frame, contains the information needed by the handler to restore the previous execution context upon return.

2.5 Synchronization

The UNIX kernel is re-entrant. At any time, several processes may be active in the kernel. Of these, only one (on a uniprocessor) can be actually running; the others are blocked, waiting either for the

CPU or some other resource. Since they all share the same copy of the kernel data structures, it is necessary to impose some form of synchronization, to prevent corruption of the kernel.

Figure 2-6 shows one example of what can happen in the absence of synchronization. Suppose a process is trying to remove element **B** from the linked list. It executes the first line of code, but is interrupted before it can execute the next line, and another process is allowed to run. If this second process were to access this same list, it would find it in an inconsistent state, as shown in Figure 2-6(b). Clearly, we need to ensure that such problems never occur.

UNIX uses several synchronization techniques. The first line of defense is that the UNIX kernel is nonpreemptive. This means that if a process is executing in kernel mode, it cannot be preempted by another process, *even though its time quantum may expire*. The process must voluntarily relinquish the CPU. This typically happens when the process is about to block while waiting for a resource or event, or when it has completed its kernel mode activity and is about to return to user mode. In either case, since the CPU is relinquished voluntarily, the process can ensure that the kernel remains in a consistent state. (Modern UNIX kernels with real-time capability allow preemption under certain conditions—see Section 5.6 for details.)

Making a kernel nonpreemptive provides a broad, sweeping solution to most synchronization problems. In the example of Figure 2-6, for instance, the kernel can manipulate the linked list without locking it, if it does not have to worry about preemption. There are three situations where synchronization is still necessary—blocking operations, interrupts, and multiprocessor synchronization.

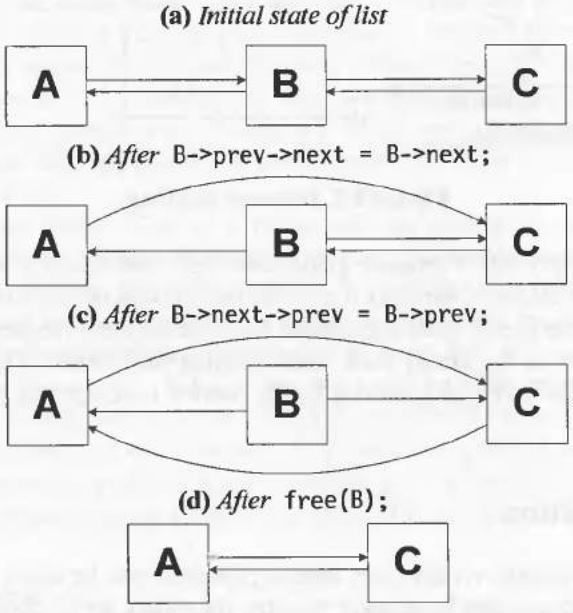

Figure 2-6. Removing an element from a linked list.

2.5.1 Blocking Operations

A *blocking operation* is one that blocks the process (places the process in the *asleep* state until the operation completes). Since the kernel is nonpreemptive, it may manipulate most objects (data structures and resources) with impunity, knowing that no other process will disturb it. Some objects, however, must be protected across a blocking operation, and this requires additional mechanisms. For instance, a process may issue a *read* from a file into a disk block buffer in kernel memory. Since disk I/O is necessary, the process must wait until the I/O completes, allowing other processes to run in the meantime. However, the kernel must ensure that other processes do not access this buffer in any way, since the buffer is in an inconsistent state.

To protect such an object, the kernel associates a *lock* with it. The lock may be as simple as a single bit-flag, which is set when locked and clear when unlocked. Any process that wants to use the object must first check if it is locked. If so, the process must block until the object is unlocked. If not, it locks the object and proceeds to use it. Normally, the kernel also associates a *wanted* flag with the object. This flag is set by a process that wants the object but finds it locked. When a process is ready to release a locked object, it checks the object's *wanted* flag to see if other processes are waiting for the object and, if so, awakens them. This mechanism allows a process to lock a resource for a long period of time, even if it had to block and allow other processes to run while holding the lock.

Figure 2-7 describes the algorithm for resource locking. The following points must be noted:

- A process blocks itself when it cannot obtain a resource, or when it must wait for an event such as I/O completion. It does so by calling a routine called `sleep()`. This is called *blocking on* the resource or event.
- `sleep()` puts the process on a special queue of blocked processes, changes its state to asleep, and calls a function called `swtch()` to initiate a context switch and allow another process to run.
- The process releasing the resource calls `wakeup()`to wake *all* processes that are waiting for this resource.[8] `wakeup()` finds each such process, changes its state to runnable, and puts it on a scheduler queue, where it now waits to be scheduled.
- There can be a substantial delay between the time a process is awakened and the time it is scheduled to run. Other processes may run in the meantime, and may even lock the same resource again.
- Thus upon waking up, the process must check once again if the resource is actually available, and go back to sleep if not.

2.5.2 Interrupts

While the kernel is normally safe from preemption by another process, a process manipulating kernel data structures may be interrupted by devices. If the interrupt handler tries to access those very data structures, they may be in an inconsistent state. This problem is handled by blocking interrupts while accessing such critical data structures. The kernel uses macros such as those in Table 2-1 to

[8] Recent versions of UNIX offer several alternatives to `wakeup()`, such as `wake_one()` and `wakeprocs()`.

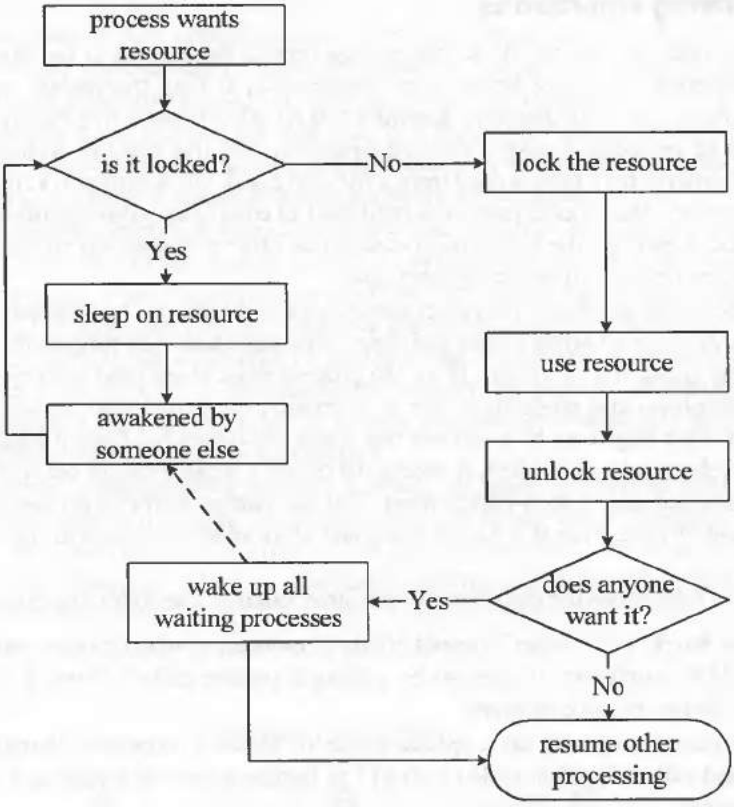

Figure 2-7. Algorithm for resource locking.

explicitly raise the *ipl* and block interrupts. Such a region of code is called a *critical region* (see Example 2-1).

```
int x = splbio();              /* raises ipl, returns previous ipl */
modify disk buffer cache;
splx(x);                       /* restores previous ipl */
```

Example 2-1. Blocking interrupts in a critical region.

There are several important considerations related to interrupt masking:

- Interrupts usually require rapid servicing, and hence should not be interfered with excessively. Thus, critical regions should be few and brief.
- The only interrupts that must be blocked are those that may manipulate the data in the critical region. In the previous example, only the disk interrupts need to be blocked.

- Two different interrupts can have the same priority level. For instance, on many systems both terminal and disk interrupts occur at *ipl* 21.
- Blocking an interrupt also blocks all interrupts at the same or lower *ipl*.

> ***Note:*** *The word* block *is used in many different ways when describing the UNIX subsystem. A process* blocks on *a resource or event when it enters the asleep state waiting for the resource to be available or the event to occur. The kernel* blocks *an interrupt or a signal by temporarily disabling its delivery. Finally, the I/O subsystem transfers data to and from storage devices in fixed-size* blocks.

2.5.3 Multiprocessors

Multiprocessor systems lead to a new class of synchronization problems, since the fundamental protection offered by the nonpreemptive nature of the kernel is no longer present. On a uniprocessor, the kernel manipulates most data structures with impunity, knowing that it cannot be preempted. It only needs to protect data structures that may be accessed by interrupt handlers, or those that need to be consistent across calls to `sleep()`.

On a multiprocessor, two processes may execute in kernel mode on different processors and may also execute the same function concurrently. Thus any time the kernel accesses a global data structure, it must lock that structure to prevent access from other processors. The locking mechanisms themselves must be multiprocessor-safe. If two processes running on different processors attempt to lock an object at the same time, only one must succeed in acquiring the lock.

Protecting against interrupts also is more complicated, because all processors may handle interrupts. It is usually not advisable to block interrupts on every processor, since that might degrade performance considerably. Multiprocessors clearly require more complex synchronization mechanisms. Chapter 7 explores these issues in detail.

2.6 Process Scheduling

The CPU is a resource that must be shared by all processes. The part of the kernel that apportions CPU time between processes is called the *scheduler*. The traditional UNIX scheduler uses *preemptive round-robin scheduling*. Processes of equal priority are scheduled in a round-robin manner, each running for a fixed quantum of time(typically 100 milliseconds). If a higher priority process becomes runnable, it will preempt the current process (unless the current process is running in kernel mode), even if the current process has not used up its time quantum.

In traditional UNIX systems, the process priority is determined by two factors—the *nice value* and the *usage factor*. Users may influence the process priority by modifying its nice value using the *nice* system call (only the superuser may increase the priority of a process). The usage factor is a measure of the recent CPU usage of the process. It allows the kernel to vary the process priority dynamically. While a process is not running, the kernel periodically increases its priority. When a process receives some CPU time, the kernel reduces its priority. This scheme prevents star-

vation of any process, since eventually the priority of any process that is waiting to run will rise high enough for it to be scheduled.

A process executing in kernel mode may relinquish the CPU if it must block for a resource or event. When it becomes runnable again, it is assigned a kernel priority. Kernel priorities are higher than any user priorities. In traditional UNIX kernels, scheduling priorities have integer values between 0 and 127, with smaller numbers meaning higher priorities. (As the UNIX system is written almost entirely in C, it follows the standard convention of beginning all counts and indices at 0). In 4.3BSD, for instance, the kernel priorities range from 0 to 49, and user priorities from 50 to 127. While user priorities vary with CPU usage, kernel priorities are fixed, and depend on the reason for sleeping. Because of this, kernel priorities are also known as *sleep priorities*. Table 2-2 lists the sleep priorities in 4.3BSD UNIX.

Chapter 5 provides further details of the UNIX scheduler.

2.7 Signals

UNIX uses *signals* to inform a process of asynchronous events, and to handle exceptions. For example, when a user types control-C at the terminal, a `SIGINT` signal is sent to the foreground process. Likewise, when a process terminates, a `SIGCHLD` signal is sent to its parent. UNIX defines a number of signals (31 in 4.3BSD and SVR3). Most are reserved for specific purposes, while two (`SIGUSR1` and `SIGUSR2`) are available for applications to use as they wish.

Signals are generated in many ways. A process may explicitly send a signal to one or more processes using the *kill* system call. The terminal driver generates signals to processes connected to it in response to certain keystrokes and events. The kernel generates a signal to notify the process of a hardware exception, or of a condition such as exceeding a quota.

Each signal has a default response, usually process termination. Some signals are ignored by default, and a few others suspend the process. The process may specify another action instead of the default by using the *signal* (System V), *sigvec* (BSD), or *sigaction* (POSIX.1) calls. This other ac-

Table 2-2. Sleep priorities in 4.3BSD UNIX

Priority	Value	Description
`PSWP`	0	swapper
`PSWP + 1`	1	page daemon
`PSWP + 1/2/4`	1/2/4	other memory management activity
`PINOD`	10	waiting for inode to be freed
`PRIBIO`	20	disk I/O wait
`PRIBIO + 1`	21	waiting for buffer to be released
`PZERO`	25	baseline priority
`TTIPRI`	28	terminal input wait
`TTOPRI`	29	terminal output wait
`PWAIT`	30	waiting for child process to terminate
`PLOCK`	35	advisory resource lock wait
`PSLEP`	40	wait for a signal

tion may be to invoke a user-specified signal handler, or it may be to ignore the signal, or even to revert to the default. The process may also choose to block a signal temporarily; such a signal will only be delivered to a process after it is unblocked.

A process does not instantaneously respond to a signal. When the signal is generated, the kernel notifies the process by setting a bit in the pending signals mask in its proc structure. The process must become aware of the signal and respond to it, and that can only happen when it is scheduled to run. Once it runs, the process will handle all pending signals before returning to its normal user-level processing. (This does not include the signal handlers themselves, which run in user mode.)

What should happen if a signal is generated for a sleeping process? Should the signal be kept pending until the process awakens, or should the sleep be interrupted? The answer depends on why the process is sleeping. If the process is sleeping for an event that is certain to occur soon, such as disk I/O completion, there is no need to wake up the process. If, on the other hand, the process is waiting for an event such as terminal input, there is no limit to how long it might block. In such a case, the kernel interrupts the sleep and aborts the system call in which the process had blocked. 4.3BSD provides the *siginterrupt* system call to control how signals should affect system call handling. Using *siginterrupt,* the user can specify whether system calls interrupted by signals should be aborted or restarted. Chapter 4 covers the topic of signals in greater detail.

2.8 New Processes and Programs

UNIX is a multiprogramming environment, and several processes are active in the system at any time. Each process runs a single program at a time, though several processes may run the same program concurrently. Such processes may share a single copy of the program text in memory, but maintain their own individual data and stack regions. Moreover, a process may invoke a new program at any time and may run several programs in its lifetime. UNIX thus makes a sharp distinction between the process and the program it is running.

To support such an environment, UNIX provides several system calls to create and terminate processes, and to invoke new programs. The *fork* and *vfork* system calls create new processes. The *exec* call invokes a new program. The *exit* system call terminates a process. Note that a process may also terminate if it receives a signal.

2.8.1 *fork* and exec

The *fork* system call creates a new process. The process that calls *fork* is the parent, and the new process is its child. The parent-child relationship creates the process hierarchy described in Figure 2-4. The child process is almost an exact clone of the parent. Its address space is a replica of that of the parent, and it also runs the same program initially. In fact, the child begins user mode execution by returning from the *fork.*

Because the parent and child both return from the *fork* and continue to execute the same program, they need a way to distinguish themselves from one another and act accordingly. Otherwise, it would be impossible for different processes to do different things. For this reason, the *fork* system

call returns different values to the parent and child—*fork* returns *0* to the child, and the child's PID to the parent.

Most often, the child process will call *exec* shortly after returning from *fork*, and thus begin executing a new program. The C library provides several alternate forms of *exec*, such as *exece*, *execve*, and *execvp*. Each takes a slightly different set of arguments and, after some preprocessing, calls the same system call. The generic name *exec* refers to any unspecified function of this group. The code that uses *fork* and *exec* looks resembles that in Example 2-2.

```
if ((result = fork()) == 0) {
    /* child code */
    ...
    if (execve ("new_program", ...) < 0)
        perror ("execve failed");
        exit(1);
} else if (result < 0)  {
    perror ("fork");    /* fork failed */
}
/* parent continues here */
...
```

Example 2-2. Using *fork* and *exec*.

Since *exec* overlays a new program on the existing process, the child does not return to the old program unless *exec* fails. Upon successful completion of *exec*, the child's address space is replaced with that of the new program, and the child returns to user mode with its program counter set to the first executable instruction of the new program.

Since *fork* and *exec* are so often used together, it may be argued that a single system call could efficiently accomplish both tasks, resulting in a new process running a new program. Older UNIX systems [Thom 78] also incurred a large overhead in duplicating the parent's address space for the child (during *fork*), only to have the child discard it completely and replace it with that of the new program.

There are many advantages of keeping the calls separate. In many client-server applications, the server program may *fork* numerous processes that continue to execute the same program.[9] In contrast, sometimes a process wants merely to invoke a new program, without creating a new process. Finally, between the *fork* and the *exec*, the child may optionally perform a number of tasks to ensure that the new program is invoked in the desired state. These tasks include:

- Redirecting standard input, output, or error.
- Closing open files inherited from the parent that are not needed by the new program.
- Changing the UID or *process group*.
- Resetting signal handlers.

A single system call that tries to perform all these functions would be unwieldy and inefficient. The existing *fork-exec* framework provides greater flexibility and is clean and modular. In

[9] Modern multi-threaded UNIX systems make this unnecessary – the server simply creates a number of threads.

Section 2.8.3 we will examine ways of minimizing the performance problems associated with this division.

2.8.2 Process Creation

The *fork* system call creates a new process that is almost an exact clone of the parent. The only differences are those necessary to distinguish between the two. Upon return from *fork*, both parent and child are executing the same program, have identical data and stack regions, and resume execution at the instruction immediately following the call to *fork*. The *fork* system call must perform the following actions:

1. Reserve swap space for the child's data and stack.
2. Allocate a new PID and `proc` structure for the child.
3. Initialize the child's `proc` structure. Some fields (such as user and group ID, process group, and signal masks) are copied from the parent, some set to zero (resident time, CPU usage, sleep channel, etc.), and others (such as PID, parent PID, and pointer to the parent `proc` structure) initialized to child-specific values.
4. Allocate address translation maps for the child.
5. Allocate the child's `u` area and copy it from the parent.
6. Update the `u` area to refer to the new address maps and swap space.
7. Add the child to the set of processes sharing the text region of the program that the parent is executing.
8. Duplicate the parent's data and stack regions one page at a time and update the child's address maps to refer to these new pages.
9. Acquire references to shared resources inherited by the child, such as open files and the current working directory.
10. Initialize the child's hardware context by copying a snapshot of the parent's registers.
11. Make the child runnable and put it on a scheduler queue.
12. Arrange for the child to return from *fork* with a value of zero.
13. Return the PID of the child to the parent.

2.8.3 *fork* Optimization

The *fork* system call must give the child a logically distinct copy of the parent's address space. In most cases, the child discards this address space when it calls *exec* or *exit* shortly after the *fork*. It is therefore wasteful to make an actual copy of the address space, as was done in older UNIX systems.

This problem has been addressed in two different ways. The first is the *copy-on-write* approach, first adopted by System V and now used by most UNIX systems. In this method, the data and stack pages of the parent are temporarily made read-only and marked as copy-on-write. The child receives its own copy of the address translation maps, but shares the memory pages with the parent. If either the parent or the child tries to modify a page, a page fault exception occurs (because the page is marked read-only) and the kernel fault handler is invoked. The handler recognizes that this is a copy-on-write page, and makes a new writable copy of that single page. Thus only those

pages that are modified must be copied, not the entire address space. If the child calls *exec* or *exit*, the pages revert to their original protection, and the copy-on-write flag is cleared.

BSD UNIX provided another solution—a new *vfork* system call. A user may call *vfork* instead of *fork* if he or she expects to call *exec* shortly afterward. *vfork* does no copying. Instead, the parent loans its address space to the child and blocks until the child returns it. The child then executes using the parent's address space, until it calls *exec* or *exit*, whereupon the kernel returns the address space to the parent, and awakens it. *vfork* is extremely fast, since not even the address maps are copied. The address space is passed to the child simply by copying the address map registers. *It is, however, a dangerous call, because it permits one process to use and even modify the address space of another process*. Some programs such as *csh* exploit this feature.

2.8.4 Invoking a New Program

The *exec* system call replaces the address space of the calling process with that of a new program. If the process was created by a *vfork*, *exec* returns the old address space to the parent. Otherwise, it frees the old address space. *exec* gives the process a new address space and loads it with the contents of the new program. When *exec* returns, the process resumes execution at the first instruction of the new program.

The process address space has several distinct components:[10]

- **Text:** Contains the executable code, and corresponds to the text section of the program.
- **Initialized data:** Consists of data objects explicitly initialized in the program, and corresponds to the initialized data section of the executable file.
- **Uninitialized data:** Historically called the *block static storage (bss)* region, consists of data variables declared, but not initialized, in the program. Objects in this region are guaranteed to be zero-filled when first accessed. Because it is wasteful to store several pages of zeroes in the executable file, the program header simply records the total size of this region and relies on the operating system to generate zero-filled pages for these addresses.
- **Shared memory:** Many UNIX systems allow processes to share regions of memory.
- **Shared libraries:** If a system supports dynamically linked libraries, the process may have separate regions of memory containing library code and data that may be shared with other processes.
- **Heap:** Source for dynamically allocated memory. A process allocates memory from the heap by making the *brk* or *sbrk* system calls, or using the *malloc()* function in the standard C library. The kernel provides each process with a heap, and extends it when needed.
- **User stack:** The kernel allocates a stack for each process. In most traditional UNIX implementations, the kernel transparently catches stack overflow exceptions and extends the user stack up to a preset maximum.

Shared memory is standard in System V UNIX, but is not available in 4BSD (through release 4.3). Many BSD-based commercial variants support both shared memory and some form of

[10] This division is simply functional in nature; the kernel does not recognize so many different components. SVR4, for instance, views the address space as merely a collection of shared and private mappings.

shared libraries as value-added features. In the following description of *exec*, we will consider a simple program that uses neither of these features.

UNIX supports many executable file formats. The oldest is the *a.out* format, which has a 32-byte header followed by text and data sections and the symbol table. The program header contains the sizes of the *text*, *initialized data*, and *uninitialized data* regions, and the *entry point*, which is the address of the first instruction the program must execute. It also contains a *magic number*, which identifies the file as a valid executable file and gives further information about its format, such as whether the file is demand paged, or whether the data section begins on a page boundary. Each UNIX variant defines the set of magic numbers it supports.

The *exec* system call must perform the following tasks:

1. Parse the pathname and access the executable file.
2. Verify that the caller has execute permission for the file.
3. Read the header and check that it is a valid executable.[11]
4. If the file has SUID or SGID bits set in its mode, change the caller's effective UID or GID respectively to that of the owner of the file.
5. Copy the arguments to *exec* and the *environment variables* into kernel space, since the current user space is going to be destroyed.
6. Allocate swap space for the data and stack regions.
7. Free the old address space and the associated swap space. If the process was created by *vfork*, return the old address space to the parent instead.
8. Allocate address maps for the new text, data, and stack.
9. Set up the new address space. If the text region is already active (some other process is already running the same program), share it with this process. Otherwise, it must be initialized from the executable file. UNIX processes are usually demand paged, meaning that each page is read into memory only when the program needs it.
10. Copy the arguments and environment variables back onto the new user stack.
11. Reset all signal handlers to default actions, because the handler functions do not exist in the new program. Signals that were ignored or blocked before calling *exec* remain ignored or blocked.
12. Initialize the hardware context. Most registers are reset to zero, and the program counter is set to the entry point of the program.

2.8.5 Process Termination

The `exit()` function in the kernel terminates a process. It is called internally when the process is killed by a signal. Alternatively, the program may invoke the *exit* system call, which calls the `exit()` function. The `exit()` function performs the following actions:

---

[11] *exec* can also invoke shell scripts whose first line is

```
#!shell-name
```

in which case, it invokes the program specified by *shell-name* (usually the name of a shell, but could be any executable file) and passes it the name of the script as the first argument. Some systems (such as UnixWare) require a space before the *shell-name*.

1. Turns off all signals.
2. Closes all open files.
3. Releases the text file and other resources such as the current directory.
4. Writes to the accounting log.
5. Saves resource usage statistics and exit status in the `proc` structure.
6. Changes state to `SZOMB` (zombie), and puts the `proc` structure on the zombie process list.
7. Makes the *init* process inherit (become the parent of) any live children of the exiting process.
8. Releases the address space, u area, address translation maps, and swap space.
9. Notifies the parent by sending it a `SIGCHLD` signal. This signal is ignored by default, and thus has an effect only if the parent wants to know about the child's death.
10. Wakes up the parent if it is asleep.
11. Finally, calls `swtch()` to schedule a new process to run.

When `exit()` completes, the process is in the zombie state. *exit* does not free the `proc` structure of the parent, because its parent may want to retrieve the exit status and resource usage information. The parent is responsible for freeing the child's `proc` structure, as described below. When that happens, the `proc` structure is returned to a free list, and the cleanup is complete.

2.8.6 Awaiting Process Termination

Often a parent process needs to know when a child terminates. For instance, when the shell spawns a foreground process to execute a command, it must wait for the child to complete, and then prompt for the next command. When a background process terminates, the shell may want to notify the user by printing an informational message on the terminal. The shell also retrieves the exit status of the child process, so that the user may take different actions based on its success or failure. UNIX systems provide the following calls to await process termination:

```
wait (stat_loc);                        /* System V, BSD, and POSIX.1 */
wait3 (statusp, options, rusagep);      /* BSD */
waitpid (pid, stat_loc, options);       /* POSIX.1 */
waitid (idtype, id, infop, options);    /* SVR4[12] */
```

The *wait* system call allows a process to wait for a child to terminate. Since a child may have terminated before the call, *wait* must also handle that condition. *wait* first checks if the caller has any deceased or suspended children. If so, it returns immediately. If there are no deceased children, *wait* blocks the caller until one of its children dies and returns once that happens. In both cases, *wait* returns the PID of the deceased child, writes the child's exit status into `stat_loc`, and frees its `proc` structure (if more than one child is dead, *wait* acts only on the first one it finds). If the child is being traced, *wait* also returns when the child receives a signal. *wait* returns an error if the caller has no children (dead or alive), or if *wait* is interrupted by a signal.

[12] SVR4 supports *wait3* and *waitpid* as library functions.

4.3BSD provides a *wait3* call (so named because it requires three arguments), which also returns resource usage information about the child (user and system times of the child and all its deceased children). The POSIX.1 standard [IEEE 90] adds the *waitpid* call, which uses the `pid` argument to wait for a child with a specific process ID or process group. Both *wait3* and *waitpid* support two options: `WNOHANG` and `WUNTRACED`. `WNOHANG` causes *wait3* to return immediately if there are no deceased children. `WUNTRACED` also returns if a child is suspended or resumed. The SVR4 *waitid* call provides a superset of all the above features. It allows the caller to specify the process ID or group to wait for and the specific events to trap, and also returns more detailed information about the child process.

2.8.7 Zombie Processes

When a process exits, it remains in zombie state until cleaned up by its parent. In this state, the only resource it holds is a `proc` structure, which retains its exit status and resource usage information.[13] This information may be important to its parent. The parent retrieves this information by calling *wait*, which also frees the `proc` structure. If the parent dies before the child, the *init* process inherits the child. When the child dies, *init* calls *wait* to release the child's `proc` structure.

A problem may arise if a process dies before its parent, and the parent does not call *wait*. The child's `proc` structure is never released, and the child remains in the zombie state until the system is rebooted. This situation is rare, since the shells are written carefully to avoid this problem. It may happen, however, if a carelessly written application does not wait for all child processes. This is an annoyance, because such zombies are visible in the output of *ps* (and users are vexed to find that they cannot be killed—*they are already dead*). Furthermore, they use up a `proc` structure, thereby reducing the maximum number of processes that can be active.

Some newer UNIX variants allow a process to specify that it will not wait for its children. For instance, in SVR4, a process may specify the `SA_NOCLDWAIT` flag to the *sigaction* system call to specify the action for `SIGCHLD` signals. This asks the kernel not to create zombies when the caller's children terminate.

2.9 Summary

We have described the interactions between the kernel and user processes in traditional UNIX kernels. This provides a broad perspective, giving us the context needed to examine specific parts of the system in greater detail. Modern variants such as SVR4 and Solaris 2.*x* introduce several advanced facilities, which will be detailed in the following chapters.

2.10 Exercises

1. What elements of the process context must the kernel explicitly save when handling (a) a context switch, (b) an interrupt, or (c) a system call?

[13] Some implementations use a special `zombie` structure to retain this data.

2. What are the advantages of allocating objects such as proc structures and descriptor table blocks dynamically? What are the drawbacks?
3. How does the kernel know which system call has been made? How does it access the arguments to the call (which are on the user stack)?
4. Compare and contrast the handling of system calls and of exceptions. What are the similarities and differences?
5. Many UNIX systems provide compatibility with another version of UNIX by providing user library functions to implement the system calls of the other version. Why, if at all, does the application developer care if a function is implemented by a library or by a system call?
6. What issues must a library developer be concerned with when choosing to implement a function in the user library instead of as a system call? What if the library must use multiple system calls to implement the function?
7. Why is it important to limit the amount of work an interrupt handler can do?
8. On a system with n distinct interrupt priority levels, what is the maximum number of interrupts that may be nested at a time? What repercussions can this have on the sizes of various stacks?
9. The Intel 80x86 architecture does not support interrupt priorities. It provides two instructions for interrupt management—CLI to disable all interrupts, and STI to enable all interrupts. Write an algorithm to implement interrupt priority levels in software for such a machine.
10. When a resource becomes available, the wakeup() routine wakes up all processes blocked on it. What are the drawbacks of this approach? What are the alternatives?
11. Propose a new system call that combines the functions of *fork* and *exec*. Define its interface and semantics. How would it support features such as I/O redirection, foreground or background execution, and pipes?
12. What is the problem with returning an error from the exec system call? How can the kernel handle this problem?
13. For a UNIX system of your choice, write a function that allows a process to wait for its *parent* to terminate.
14. Suppose a process does not wish to block until its children terminate. How can it ensure that child processes are cleaned up when they terminate?
15. Why does a terminating process wake up its parent?

2.11 References

[Allm 87] Allman, E., "UNIX: The Data Forms," *Proceedings of the Winter 1987 USENIX Technical Conference,* Jan. 1987, pp. 9–15.

[AT&T 87] American Telephone and Telegraph, *The System V Interface Definition (SVID),* Issue 2, 1987.

[Bach 86] Bach, M.J., *The Design of the UNIX Operating System,* Prentice-Hall, Englewood Cliffs, NJ, 1986.

[Kern 84] Kernighan, B.W., and Pike, R., *The UNIX Programming Environment,* Prentice-Hall, Englewood Cliffs, NJ, 1984.

7

Synchronization and Multiprocessors

7.1 Introduction

The desire for more processing power has led to several advances in hardware architectures. One of the major steps in this direction has been the development of multiprocessor systems. These systems consist of two or more processors sharing the main memory and other resources. Such configurations offer several advantages. They provide a flexible growth path for a project, which may start with a single processor and, as its computing needs grow, expand seamlessly by adding extra processors to the machine. Systems used for compute-intensive applications are often CPU-bound. The CPU is the main bottleneck, and other system resources such as the I/O bus and memory are underutilized. Multiprocessors add processing power without duplicating other resources, and hence provide a cost-effective solution for CPU-bound workloads.

Multiprocessors also provide an extra measure of reliability; if one of the processors should fail, the system could still continue to run without interruption. This, however, is a double-edged sword, since there are more potential points of failure. To ensure a high *mean time before failure* (MTBF), multiprocessor systems must be equipped with fault-tolerant hardware and software. In particular, the system should recover from the failure of one processor without crashing.

Several variants of UNIX have evolved to take advantage of such systems. One of the earliest multiprocessing UNIX implementations ran on the AT&T 3B20A and the IBM 370 architectures [Bach 84]. Currently, most major UNIX implementations are either native multiprocessing systems (DECUNIX, Solaris 2.*x*) or have multiprocessing variants (SVR4/MP, SCO/MPX).

Ideally, we would like to see the system performance scale linearly with the number of processors. Real systems fall short of this goal for several reasons. Since the other components of the system are not duplicated, they can become bottlenecks. The need to synchronize when accessing shared data structures, and the extra functionality to support multiple processors, adds CPU overhead and reduces the overall performance gains. The operating system must try to minimize this overhead and allow optimal CPU utilization.

The traditional UNIX kernel assumes a uniprocessor architecture and needs major modifications to run on multiprocessor systems. The three main areas of change are synchronization, parallelization, and scheduling policies. Synchronization involves the basic primitives used to control access to shared data and resources. The traditional primitives of sleep/wakeup combined with interrupt blocking are inadequate in a multiprocessing environment and must be replaced with more powerful facilities.

Parallelization concerns the efficient use of the synchronization primitives to control access to shared resources. This involves decisions regarding lock granularity, lock placement, deadlock avoidance, and so forth. Section 7.10 discusses some of these issues. The scheduling policy also needs to be changed to allow the optimal utilization of all processors. Section 7.4 analyzes some issues related to multiprocessor scheduling.

This chapter first describes the synchronization mechanisms in traditional UNIX systems and analyzes their limitations. It follows with an overview of multiprocessor architectures. Finally the chapter describes synchronization in modern UNIX systems. The methods describe work well on both uniprocessor and multiprocessor platforms.

In traditional UNIX systems, the process is the basic scheduling unit, and it has a single thread of control. As described in Chapter 3, many modern UNIX variants allow multiple threads of control in each process, with full kernel support for these threads. Such multithreaded systems are available both on uniprocessor and multiprocessor architectures. In these systems, individual threads contend for and lock the shared resources. In the rest of this chapter, we refer to a thread as the basic scheduling unit since it is the more general abstraction. For a single-threaded system, a thread is synonymous with a process.

7.2 Synchronization in Traditional UNIX Kernels

The UNIX kernel is reentrant—several processes may be executing in the kernel at the same time, perhaps even in the same routine. On a uniprocessor only one process can actually execute at a time. However, the system rapidly switches from one process to another, providing the illusion that they are all executing concurrently. This feature is usually called *multiprogramming*. Since these processes share the kernel, the kernel must synchronize access to its data structures in order to avoid corrupting them. Section 2.5 provides a detailed discussion of traditional UNIX synchronization techniques. In this section, we summarize the important principles.

The first safeguard is that the traditional UNIX kernel is nonpreemptive. Any thread executing in kernel mode will continue to run, even though its time quantum may expire, until it is ready to leave the kernel or needs to block for some resource. This allows kernel code to manipulate sev-

eral data structures without any locking, knowing that no other thread can access them until the current thread is done with them and is ready to relinquish the kernel in a consistent state.

7.2.1 Interrupt Masking

The no-preemption rule provides a powerful and wide-ranging synchronization tool, but it has certain limitations. Although the current thread may not be preempted, it may be interrupted. Interrupts are an integral part of system activity and usually need to be serviced urgently. The interrupt handler may manipulate the same data structures with which the current thread was working, resulting in corruption of that data. Hence the kernel must synchronize access to data that is used both by normal kernel code and by interrupt handlers.

UNIX solves this problem by providing a mechanism for blocking (masking) interrupts. Associated with each interrupt is an *interrupt priority level* (*ipl*). The system maintains a *current ipl* value and checks it whenever an interrupt occurs. If the interrupt has a higher priority than the *current ipl*, it is handled immediately (preempting the lower-priority interrupt currently being handled). Otherwise, the kernel blocks the interrupt until the *ipl* falls sufficiently. Prior to invoking the handler, the system raises the *ipl* to that of the interrupt; when the handler completes, the system restores the *ipl* to the previous value (which it saves). The kernel can also explicitly set the *ipl* to any value to mask interrupts during certain critical processing.

For example, a kernel routine may want to remove a disk block buffer from a buffer queue it is on; this queue may also be accessed by the disk interrupt handler. The code to manipulate the queue is a *critical region*. Before entering the critical region, the routine will raise the *ipl* high enough to block disk interrupts. After completing the queue manipulation, the routine will set the *ipl* back to its previous value, thus allowing the disk interrupts to be serviced. The *ipl* thus allows effective synchronization of resources shared by the kernel and interrupt handlers.

7.2.2 Sleep and Wakeup

Often a thread wants to guarantee exclusive use of a resource even if it needs to block for some reason. For instance, a thread wants to read a disk block into a block buffer. It allocates a buffer to hold the block and then initiates disk activity. This thread needs to wait for the I/O to complete, which means it must relinquish the processor to some other thread. If the other thread acquires the same buffer and uses it for some different purpose, the contents of the buffer may become indeterminate or corrupted. This means that threads need a way of locking the resource while they are blocked.

UNIX implements this by associating *locked* and *wanted* flags with shared resources. When a thread wants to access a sharable resource, such as a block buffer, it first checks its *locked* flag. If the flag is clear, the thread sets the flag and proceeds to use the resource. If a second thread tries to access the same resource, it finds the *locked* flag set and must block (go to sleep) until the resource becomes available. Before doing so, it sets the associated *wanted* flag. Going to sleep involves linking the thread onto a queue of sleeping threads, changing its state information to show that it is sleeping on this resource, and relinquishing the processor to another thread.

When the first thread is done with the resource, it will clear the *locked* flag and check the *wanted* flag. If the *wanted* flag is set, it means that at least one other thread is waiting for (blocked

on) this resource. In that case, the thread examines the *sleep queue* and wakes up all such threads. Waking a thread involves unlinking it from the sleep queue, changing its state to *runnable*, and putting it on the scheduler queue. When one of these threads is eventually scheduled, it again checks the *locked* flag, finds that it is clear, sets it, and proceeds to use the resource.

7.2.3 Limitations of Traditional Approach

The traditional synchronization model works correctly for a uniprocessor, but has some important performance problems that are described in this section. In a multiprocessor environment, the model breaks down completely, as will be shown in Section 7.4.

Mapping Resources to Sleep Queues

The organization of the sleep queues leads to poor performance in some situations. In UNIX, a thread blocks when waiting for a resource lock or an event. Each resource or event is associated with a *sleep channel*, which is a 32-bit value usually set to the address of the resource. There is a set of sleep queues, and a hash function maps the channel (hence, maps the resource) to one of these queues as shown in Figure 7-1. A thread goes to sleep by enqueuing itself onto the appropriate sleep queue and storing the sleep channel in its `proc` structure.

This approach has two consequences. First, more than one event may map to the same channel. For instance, one thread locks a buffer, initiates I/O activity to it, and sleeps until the I/O completes. Another thread tries to access the same buffer, finds it locked, and must block until it becomes available. Both events map to the same channel, namely the address of that buffer. When the I/O completes, the interrupt handler will wake up both threads, even though the event the second thread was waiting for has not yet occurred.

Second, the number of hash queues is much smaller than the number of different sleep channels (resources or events); hence, multiple channels map to the same hash queue. A queue thus contains threads waiting on several different channels. The `wakeup()` routine must examine each of them and only wake up threads blocked on the correct channel. As a result, the total time taken by `wakeup()` depends not on the number of processes sleeping on that channel, but on the total number

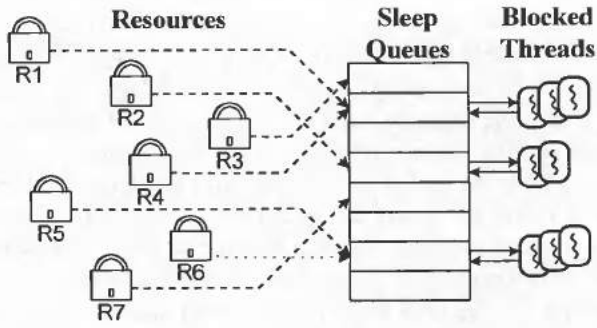

Figure 7-1. Mapping resources to global sleep queues.

of processes sleeping on that queue. This sort of unpredictable delay is usually undesirable and may be unacceptable for kernels that support real-time applications requiring bounded dispatch latency (see Section 5.5.4).

One alternative is to associate a separate sleep queue for each resource or event (Figure 7-2). This approach would optimize the latency of the wakeup algorithm, at the expense of memory overhead for all the extra queues. The typical queue header contains two pointers (forward and backward) as well as other information. The total number of synchronization objects in the system may be quite large and putting a sleep queue on each of them may be wasteful.

Solaris 2.*x* provides a more space-efficient solution [Eykh 92]. Each synchronization object has a two-byte field that locates a *turnstile* structure that contains the sleep queue and some other information (Figure 7-3). The kernel allocates turnstiles only to those resources that have threads blocked on them. To speed up allocation, the kernel maintains a pool of turnstiles, and the size of this pool is greater than the number of active threads. This approach provides more predictable real-time behavior with minimal storage overhead. Section 5.6.7 describes turnstiles in greater detail.

Shared and Exclusive Access

The sleep/wakeup mechanism is adequate when only one thread should use the resource at a time. It does not, however, readily allow for more complex protocols such as readers-writers synchronization. It may be desirable to allow multiple threads to share a resource for reading, but require exclusive access before modifying it. File and directory blocks, for example, can be shared efficiently using such a facility.

7.3 Multiprocessor Systems

There are three important characteristics of a multiprocessor system. The first is its memory model, which defines the way in which the processors share the memory. Second is the hardware support for synchronization. Finally, the software architecture determines the relationships between the processors, kernel subsystems, and user processes.

7.3.1 Memory Model

From a hardware perspective, multiprocessor systems can be divided into three categories (see Figure 7-4), depending on their coupling and memory access semantics:

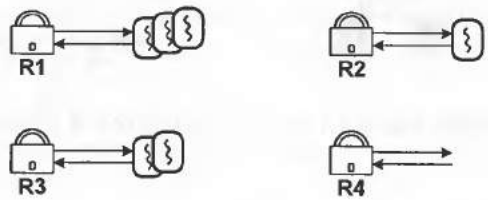

Figure 7-2. Per-resource blocked-thread queues.

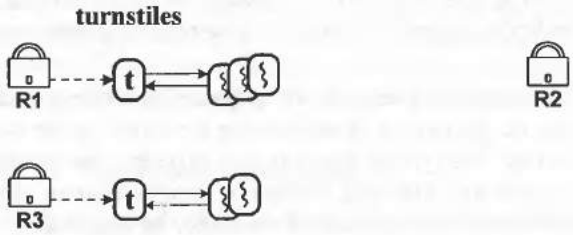

Figure 7-3. Queuing blocked threads on turnstiles.

- *Uniform Memory Access* (UMA)
- *Non-Uniform Memory Access* (NUMA)
- *No Remote Memory Access* (NORMA)

The most common system is the UMA, or shared memory, multiprocessor (Figure 7-4(a)). Such a system allows all CPUs equal access to main memory[1] and to I/O devices, usually by having

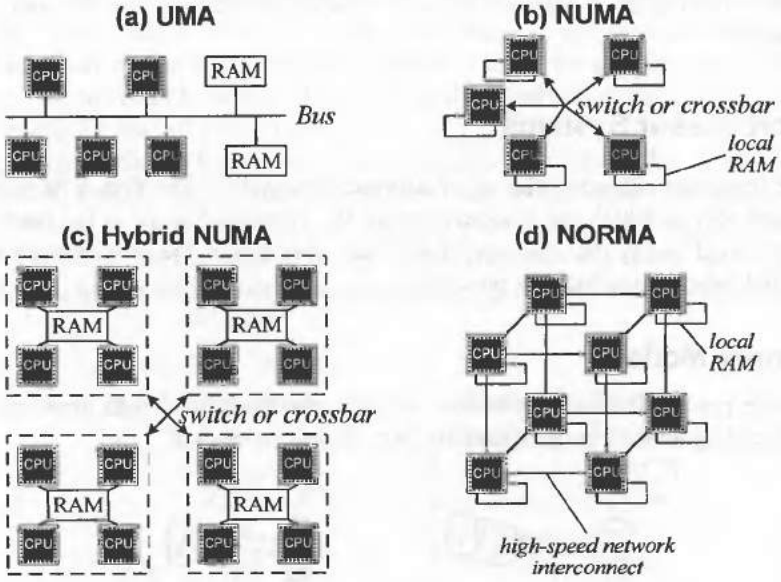

Figure 7-4. UMA, NUMA, and NORMA systems.

[1] However, the data, instruction, and address translation caches are local to each processor.

everything on a single system bus. This is a simple model from the operating system perspective. Its main drawback is scalability. UMA architectures can support only a small number of processors. As the number of processors increases, so does the contention on the bus. One of the largest UMA systems is the SGI Challenge, which supports up to 36 processors on a single bus.

In a NUMA system (Figure 7-4(b)), each CPU has some local memory, but can also access memory local to another processor. The remote access is slower, usually by an order of magnitude, than local access. There are also hybrid systems (Figure 7-4(c)), where a group of processors shares uniform access to its local memory, and has slower access to memory local to another group. The NUMA model is hard to program without exposing the details of the hardware architecture to the applications.

In a NORMA system (Figure 7-4(d)), each CPU has direct access only to its own local memory and may access remote memory only through explicit message passing. The hardware provides a high-speed interconnect that increases the bandwidth for remote memory access. Building a successful system for such an architecture requires cache management and scheduling support in the operating system, as well as compilers that can optimize the code for such hardware.

This chapter restricts itself to UMA systems.

7.3.2 Synchronization Support

Synchronization on a multiprocessor is fundamentally dependent on hardware support. Consider the basic operation of locking a resource for exclusive use by setting a *locked* flag maintained in a shared memory location. This may be accomplished by the following sequence of operations:

1. Read the flag.
2. If the flag is 0 (hence, the resource is unlocked), lock the resource by setting the flag to 1.
3. Return TRUE if the lock was obtained, or else return FALSE.

On a multiprocessor, two threads on two different processors may simultaneously attempt to carry out this sequence of operations. As Figure 7-5 shows, both threads may think they have exclusive access to the resource. To avoid such a disaster, the hardware has to provide a more powerful primitive that can combine the three subtasks into a single indivisible operation. Many architectures solve this problem by providing either an atomic test-and-set or a conditional store instruction.

Atomic Test-and-Set

An atomic test-and-set operation usually acts on a single bit in memory. It tests the bit, sets it to one, and returns its old value. Thus at the completion of the operation the value of the bit is one (locked), and the return value indicates whether it was already set to one prior to this operation. The operation is guaranteed to be atomic, so if two threads on two processors both issue the same instruction on the same bit, one operation will complete before the other starts. Further, the operation is also atomic with respect to interrupts, so that an interrupt can occur only after the operation completes.

Such a primitive is ideally suited for simple locks. If the test-and-set returns one, the calling thread owns the resource. If it returns zero, the resource is locked by another thread. Unlocking the resource is done by simply setting the bit to zero. Some examples of test-and-set instructions are

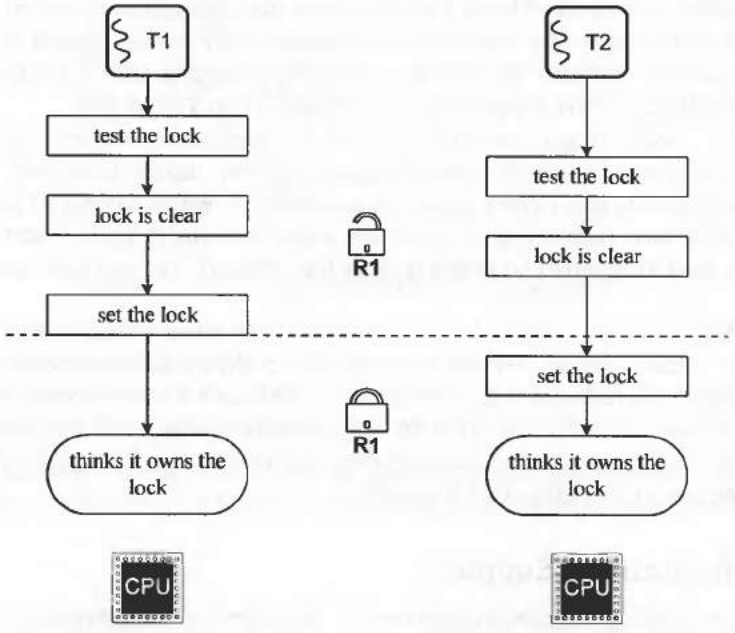

Figure 7-5. Race condition if test-and-set is not atomic.

BBSSI (**B**ranch on **B**it **S**et and **S**et **I**nterlocked) on the VAX-11 [Digi 87] and LDSTUB (**LoaD** and **STore** **U**nsigned **B**yte) on the SPARC.

Load-Linked and Store-Conditional Instructions

Some processors such as the MIPS R4000 and Digital's Alpha AXP use a pair of special load and store instructions to provide an atomic read-modify-write operation. The *load-linked* instruction (also called the *load-locked* instruction) loads a value from memory into a register and sets a flag that causes the hardware to monitor the location. If any processor writes to such a monitored location, the hardware will clear the flag. The *store-conditional* instruction stores a new value into the location provided the flag is still set. In addition, it sets the value of another register to indicate if the store occurred.

Such a primitive may be used to generate an atomic increment operation. The variable is read using load-linked, and its new value is set using store-conditional. This sequence is repeated until it succeeds. Event counters in DG/UX [Kell 89] are based on this facility.

Some systems such as the Motorola MC88100 use a third approach based on a *swap-atomic* instruction. This method is explored further in the exercises at the end of this chapter. Any of these hardware mechanisms becomes the first building block for a powerful and comprehensive synchronization facility. The high-level software abstractions described in the following sections are all built on top of the hardware primitives.

7.3.3 Software Architecture

From a software perspective, again there are three types of multiprocessing systems—master-slave, functionally asymmetric, and symmetric. A *master-slave system* [Gobl 81] is asymmetric: one processor plays the role of a *master processor*, and the rest are *slaves*. The master processor may be the only one allowed to do I/O and receive device interrupts. In some cases, only the master processor runs kernel code, and the slaves run only user-level code. Such constraints may simplify the system design, but reduce the advantage of multiple processors. Benchmark results [Bach 84] have shown that a UNIX system typically spends more than 40% of its time running in kernel mode, and it is desirable to spread this kernel activity among all processors.

Functionally asymmetric multiprocessors run different subsystems on different processors. For instance, one processor may run the networking layer, while another manages I/O. Such an approach is more suitable for a special-purpose system rather than a general-purpose operating system like UNIX. The Auspex NS5000 file server [Hitz 90] is a successful implementation of this model.

Symmetric multiprocessing (SMP) is by far the more popular approach. In an SMP system, all CPUs are equal, share a single copy of the kernel text and data, and compete for system resources such as devices and memory. Each CPU may run the kernel code, and any user process may be scheduled on any processor. This chapter describes SMP systems only, except where explicitly stated otherwise.

The rest of this chapter describes modern synchronization mechanisms, used for uniprocessor and multiprocessor systems.

7.4 Multiprocessor Synchronization Issues

One of the basic assumptions in the traditional synchronization model is that a thread retains exclusive use of the kernel (except for interrupts) until it is ready to leave the kernel or block on a resource. This is no longer valid on a multiprocessor, since each processor could be executing kernel code at the same time. *We now need to protect all kinds of data that did not need protection on a uniprocessor.* Consider for example, access to an *IPC resource table* (see Section 6.3.1). This data structure is not accessed by interrupt handlers and does not support any operations that might block the process. Hence on a uniprocessor, the kernel can manipulate the table without locking it. In the multiprocessor case, two threads on different processors can access the table simultaneously, and hence must lock it in some manner before use.

The locking primitives must be changed as well. In a traditional system, the kernel simply checks the *locked* flag and sets it to lock the object. On a multiprocessor, two threads on different processors can concurrently examine the *locked* flag for the same resource. Both will find it clear and assume that the resource is available. Both threads will then set the flag and proceed to access the resource, with unpredictable results. The system must therefore provide some kind of an atomic *test-and-set* operation to ensure that only one thread can lock the resource.

Another example involves blocking of interrupts. On a multiprocessor, a thread can typically block interrupts only on the processor on which it is running. It is usually not possible to block interrupts across all processors—in fact, some other processor may have already received a conflicting interrupt. The handler running on another processor may corrupt a data structure that the thread is

accessing. This is compounded by the fact that the handler cannot use the sleep/wakeup synchronization model, since most implementations do not permit interrupt handlers to block. The system should provide some mechanism for blocking interrupts on other processors. One possible solution is a global *ipl* managed in software.

7.4.1 The Lost Wakeup Problem

The sleep/wakeup mechanism does not function correctly on a multiprocessor. Figure 7-6 illustrates a potential race condition. Thread **T1** has locked a resource **R1**. Thread **T2**, running on another processor, tries to acquire the resource, and finds it locked. **T2** calls `sleep()` to wait for the resource. Between the time **T2** finds the resource locked and the time it calls `sleep()`, **T1** frees the resource and proceeds to wake up all threads blocked on it. Since **T2** has not yet been put on the sleep queue, it will miss the wakeup. The end result is that the resource is not locked, but **T2** is blocked waiting for it to be unlocked. If no one else tries to access the resource, **T2** could block indefinitely. This is known as the lost wakeup problem, and requires some mechanism to combine the test for the resource and the call to `sleep()` into a single atomic operation.

It is clear then, that we need a whole new set of primitives that will work correctly on a multiprocessor. This gives us a good opportunity to examine other problems with the traditional model and devise better solutions. Most of these issues are performance related.

7.4.2 The Thundering Herd Problem

When a thread releases a resource, it wakes up all threads waiting for it. One of them may now be able to lock the resource; the others will find the resource still locked and will have to go back to sleep. This may lead to extra overhead in wakeups and context switches.

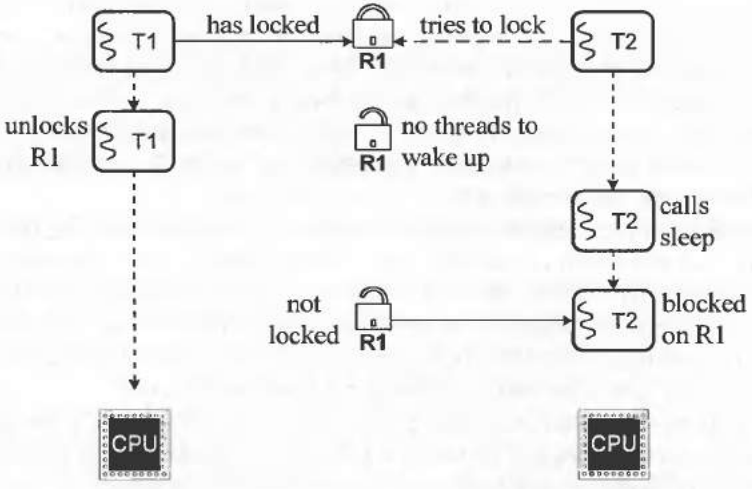

Figure 7-6. The lost wakeup problem.

This problem is not as acute on a uniprocessor, since by the time a thread runs, whoever had locked the resource is likely to have released it. On a multiprocessor, however, if several threads were blocked on a resource, waking them all may cause them to be simultaneously scheduled on different processors, and they would all fight for the same resource again. This is frequently referred to as the *thundering herd* problem.

Even if only one thread was blocked on the resource, there is still a time delay between its waking up and actually running. In this interval, an unrelated thread may grab the resource, causing the awakened thread to block again. If this happens frequently, it could lead to starvation of this thread.

We have examined several problems with the traditional synchronization model that affect correct operation and performance. The rest of this chapter describes several synchronization mechanisms that function well on both uniprocessors and multiprocessors.

7.5 Semaphores

The early implementations of UNIX on multiprocessors relied almost exclusively on *Dijkstra's semaphores* [Dijk 65] (also called *counted semaphores*) for synchronization. A semaphore is an integer-valued variable that supports two basic operations—`P()` and `V()`. `P()` decrements the semaphore and blocks if its new value is less than zero. `V()` increments the semaphore; if the resulting value is less than or equal to zero, it wakes up a thread blocked on it (if any). Example 7-1 describes these functions, plus an initialization function `initsem()` and a `CP()` function, which is a nonblocking version of `P()`:

```
void initsem (semaphore *sem, int val)
{
    *sem = val;
}

void P(semaphore *sem)  /* acquire the semaphore */
{
    *sem -= 1;
    while (*sem < 0)
        sleep;
}

void V(semaphore *sem)  /* release the semaphore */
{
    *sem += 1;
    if (*sem <= 0)
        wakeup a thread blocked on sem;
}
```

```
boolean_t CP(semaphore *sem)    /*  try to acquire semaphore without blocking */
{
    if (*sem > 0)   {
        *sem -= 1;
        return TRUE;
    } else
        return FALSE;
}
```

Example 7-1. Semaphore operations.

The kernel guarantees that the semaphore operations will be atomic, even on a multiprocessor system. Thus if two threads try to operate on the same semaphore, one operation will complete or block before the other starts. The `P()` and `V()` operations are comparable to sleep and wakeup, but with somewhat different semantics. The `CP()` operation allows a way to poll the semaphore without blocking and is used in interrupt handlers and other functions that cannot afford to block. It is also used in deadlock avoidance cases, where a `P()` operation risks a deadlock.

7.5.1 Semaphores to Provide Mutual Exclusion

Example 7-2 shows how a semaphore can provide mutual exclusion on a resource. A semaphore can be associated with a shared resource such as a linked list, and initialized to one. Each thread does a `P()` operation to lock a resource and a `V()` operations to release it. The first `P()` sets the value to zero, causing subsequent `P()` operations to block. When a `V()` is done, the value is incremented and one of the blocked threads is awakened.

```
/* During initialization */
semaphore sem;
initsem (&sem, 1);

/* On each use */
P (&sem);
Use resource;
V (&sem);
```

Example 7-2. Semaphore used to lock resource for exclusive use.

7.5.2 Event-Wait Using Semaphores

Example 7-3 shows how a semaphore can be used to wait for an event by initializing it to zero. Threads doing a `P()` will block. When the event occurs, a `V()` needs to be done for each blocked thread. This can be achieved by calling a single `V()` when the event occurs and having each thread do another `V()` upon waking up, as is shown in Example 7-3.

```
/* During initialization */
semaphore event;
initsem (&event, 0);        /* probably at boot time */

/* Code executed by thread that must wait on event */
P (&event);                 /* Blocks if event has not occurred */
/* Event has occurred */
V (&event);                 /* So that another thread may wake up */
/* Continue processing */

/* Code executed when event occurs */
V (&event);                 /* Wake up one thread */
```

Example 7-3. Semaphores used to wait for an event.

7.5.3 Semaphores to Control Countable Resources

Semaphores are also useful for allocating countable resources, such as message block headers in a STREAMS implementation. As shown in Example 7-4, the semaphore is initialized to the number of available instances of that resource. Threads call `P()` while acquiring an instance of the resource and `V()` while releasing it. Thus the value of the semaphore indicates the number of instances currently available. If the value is negative, then its absolute value is the number of pending requests (blocked threads) for that resource. This is a natural solution to the classic producers-consumers problem.

```
/* During initialization */
semaphore counter;
initsem (&counter, resourceCount);

/* Code executed to use the resource */
P (&counter);    /* Blocks until resource is available */
Use resource;    /* Guaranteed to be available now */
V (&counter);    /* Release the resource */
```

Example 7-4. Semaphore used to count available instances of a resource.

7.5.4 Drawbacks of Semaphores

Although semaphores provide a single abstraction flexible enough to handle several different types of synchronization problems, they suffer from a number of drawbacks that make them unsuitable in several situations. To begin with, a semaphore is a high-level abstraction based on lower-level primitives that provide atomicity and a blocking mechanism. For the `P()` and `V()` operations to be atomic on a multiprocessor system, there must be a lower-level atomic operation to guarantee exclusive access to the semaphore variable itself. Blocking and unblocking require context switches

and manipulation of sleep and scheduler queues, all of which make the operations slow. This expense may be tolerable for some resources that need to be held for a long time, but is unacceptable for locks held for a short time.

The semaphore abstraction also hides information about whether the thread actually had to block in the `P()` operation. This is often unimportant, but in some cases it may be crucial. The UNIX buffer cache, for instance, uses a function called `getblk()` to look for a particular disk block in the buffer cache. If the desired block is found in the cache, `getblk()` attempts to lock it by calling `P()`. If `P()` were to sleep because the buffer was locked, there is no guarantee that, when awakened, the buffer would contain the same block that it originally had. The thread that had locked the buffer may have reassigned it to some other block. Thus after `P()` returns, the thread may have locked the wrong buffer. This problem can be solved within the framework of semaphores, but the solution is cumbersome and inefficient, and indicates that other abstractions might be more suitable [Ruan 90].

7.5.5 Convoys

Compared to the traditional sleep/wakeup mechanism, semaphores offer the advantage that processes do not wake up unnecessarily. When a thread wakes up within a `P()`, it is guaranteed to have the resource. The semantics ensure that the ownership of the semaphore is transferred to the woken up thread before that thread actually runs. If another thread tries to acquire the semaphore in the meantime, it will not be able to do so. This very fact, however, leads to a performance problem called semaphore *convoys* [Lee 87]. A convoy is created when there is frequent contention on a semaphore. Although this can degrade the performance of any locking mechanism, the peculiar semantics of semaphores compound the problem.

Figure 7-7 shows the formation of a convoy. **R1** is a critical region protected by a semaphore. At instant **(a)**, thread **T2** holds the semaphore, while **T3** is waiting to acquire it. **T1** is run-

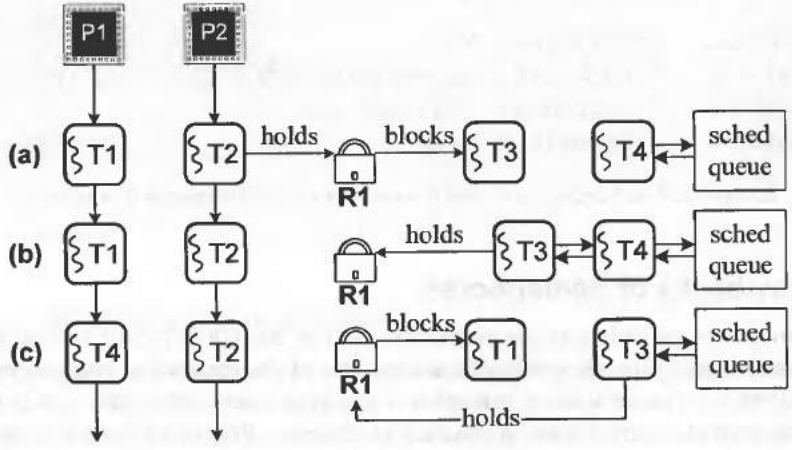

Figure 7-7. Convoy formation.

ning on another processor, and **T4** is waiting to be scheduled. Now suppose **T2** exits the critical region and releases the semaphore. It wakes up **T3** and puts it on the scheduler queue. **T3** now holds the semaphore, as shown in **(b)**.

Now suppose **T1** needs to enter the critical region. Since the semaphore is held by **T3**, **T1** will block, freeing up processor **P1**. The system will schedule thread **T4** to run on **P1**. Hence in **(c)**, **T3** holds the semaphore and **T1** is blocked on it; neither thread can run until **T2** or **T4** yields its processor.

The problem lies in step **(c)**. Although the semaphore has been assigned to **T3**, **T3** is not running and hence is not in the critical region. As a result, **T1** must block on the semaphore even though no thread is in the critical region. The semaphore semantics force allocation in a first-come, first-served order.[2] This forces a number of unnecessary context switches. Suppose the semaphore was replaced by an *exclusive lock,* or *mutex*. Then, in step **(b)**, **T2** would release the lock and wake up **T3**, but **T3** would not own the lock at this point. Consequently, in step **(c)**, **T1** would acquire the lock, eliminating the context switch.

> Note: *Mutex, short for mutual exclusion lock, is a general term that refers to any primitive that enforces exclusive access semantics.*

In general, it is desirable to have a set of inexpensive lower-level primitives instead of a single monolithic higher-level abstraction. This is the trend in the modern multiprocessing kernels, and the following sections examine the low-level mechanisms that together provide a versatile synchronization facility.

7.6 Spin Locks

The simplest locking primitive is a *spin lock*, also called a *simple lock* or a *simple mutex*. If a resource is protected by a spin lock,a thread trying to acquire the resource will *busy-wait* (loop, or spin) until the resource is unlocked. It is usually a scalar variable that is zero if available and one if locked. The variable is manipulated using a busy-wait loop around an atomic test-and-set or similar instruction available on the machine. Example 7-5 shows an implementation of a spin lock. It assumes that `test_and_set()` returns the old value of the object.

```
void spin_lock (spinlock_t *s)  {
    while (test_and_set (s) != 0)    /* already locked */
        ;                            /* loop until successful */
}

void spin_unlock (spinlock_t *s)     { *s = 0; }
```

Example 7-5. Spin lock implementation.

---

[2] Some implementation may choose the thread to wake up based on priority. The effect in this example would be the same.

Even this simple algorithm is flawed. On many processors, `test_and_set()` works by locking the memory bus, so this loop could monopolize the bus and severely degrade system performance. A better approach is to use two loops—if the test fails, the inner loop simply waits for the variable to become zero. The simple test in the inner loop does not require locking the bus. Example 7-6 shows the improved implementation:

```
void spin_lock (spinlock_t *s)
{
    while (test_and_set (s) != 0)    /* already locked */
        while (*s != 0)
            ;                        /* wait until unlocked */
}

void spin_unlock (spinlock_t *s)     { *s = 0; }
```

Example 7-6. Revised spin lock implementation.

7.6.1 Use of Spin Locks

The most important characteristic of spin locks is that a thread ties up a CPU while waiting for the lock to be released. It is essential, then, to hold spin locks only for extremely short durations. In particular, they must not be held across blocking operations. It may also be desirable to block interrupts on the current processor prior to acquiring a spin lock, so as to guarantee low holding time on the lock.

The basic premise of a spin lock is that a thread busy-waits on a resource on one processor while another thread is using the resource on a different processor. This is only possible on a multiprocessor. On a uniprocessor, if a thread tries to acquire a spin lock that is already held, it will loop forever. Multiprocessor algorithms, however, must operate correctly regardless of the number of processors, which means that they should handle the uniprocessor case as well. This requires strict adherence to the rule that threads not relinquish control of the CPU while holding a spin lock. On a uniprocessor, this ensures that a thread will never have to busy-wait on a spin lock.

The major advantage of spin locks is that they are inexpensive. When there is no contention on the lock, both the *lock* and the *unlock* operations typically require only a single instruction each. They are ideal for locking data structures that need to be accessed briefly, such as while removing an item from a doubly linked list or while performing a *load-modify-store* type of operation on a variable. Hence they are used to protect those data structures that do not need protection in a uniprocessor system. They are also used extensively to protect more complex locks, as shown in the following sections. Semaphores, for instance, use a spin lock to guarantee atomicity of their operations, as shown in Example 7-7.

```
spinlock_t list;
spin_lock (&list);
item->forw->back = item->back;
item->back->forw = item->forw;
spin_unlock (&list);
```

Example 7-7. Using a spin lock to access a doubly linked list.

7.7 Condition Variables

A *condition variable* is a more complex mechanism associated with a *predicate* (a logical expression that evaluates to TRUE or FALSE) based on some shared data. It allows threads to block on it and provides facilities to wakeup one or all blocked threads when the result of the predicate changes. It is more useful for waiting on events than for resource locking.

Consider, for example, one or more server threads waiting for client requests. Incoming requests are to be passed to waiting threads or put on a queue if no one is ready to service them. When a server thread is ready to process the next request, it first checks the queue. If there is a pending message, the thread removes it from the queue and services it. If the queue is empty, the thread blocks until a request arrives. This can be implemented by associating a condition variable with this queue. The shared data is the message queue itself, and the predicate is that the queue be nonempty.

The condition variable is similar to a sleep channel in that server threads block on the condition and incoming messages awaken them. On a multiprocessor, however, we need to guard against some race conditions, such as the lost wakeup problem. Suppose a message arrives after a thread checks the queue but before the thread blocks. The thread will block even though a message is available. We therefore need an atomic operation to test the predicate and block the thread if necessary.

Condition variables provide this atomicity by using an additional mutex, usually a spin lock. The mutex protects the shared data, and avoids the lost wakeup problem. The server thread acquires the mutex on the message queue, then checks if the queue is empty. If so, it calls the `wait()` function of the condition with the spin lock held. The `wait()` function takes the mutex as an argument and atomically blocks the thread and releases the mutex. When the message arrives on the queue and the thread is woken up, the `wait()` call reacquires the spin lock before returning. Example 7-8 provides a sample implementation of condition variables:

```
struct condition  {
    proc *next;              /* doubly linked list */
    proc *prev;              /* of blocked threads */
    spinlock_t listLock;     /* protects this list */
};
```

```
void wait (condition *c, spinlock_t *s)
{
    spin_lock (&c->listLock);
    add self to the linked list;
    spin_unlock (&c->listLock);
    spin_unlock (s);          /* release spinlock before blocking */
    swtch();                  /* perform context switch */
    /* When we return from swtch, the event has occurred */
    spin_lock (s);            /* acquire the spin lock again */
    return;
}

void do_signal (condition *c)
/* Wake up one thread waiting on this condition */
{
    spin_lock (&c->listLock);
    remove one thread from linked list, if it is nonempty;
    spin_unlock (&c->listLock);
    if a thread was removed from the list, make it runnable;
    return;
}

void do_broadcast (condition *c)
/* Wake up all threads waiting on this condition */
{
    spin_lock (&c->listLock);
    while (linked list is nonempty)  {
        remove a thread from linked list;
        make it runnable;
    }
    spin_unlock (&c->listLock);
}
```

Example 7-8. Implementation of condition variables.

7.7.1 Implementation Issues

There are a few important points to note. The predicate itself is not part of the condition variable. It must be tested by the calling routine before calling `wait()`. Further, note that the implementation uses two separate mutexes. One is `listLock`, which protects the doubly linked list of threads blocked on the condition. The second mutex protects the tested data itself. It is not a part of the condition variable, but is passed as an argument to the `wait()` function. The `swtch()` function and the code to make blocked threads runnable may use a third mutex to protect the scheduler queues.

We thus have a situation where a thread tries to acquire one spin lock while holding another. This is not disastrous since the restriction on spin locks is only that threads are not allowed to block while holding one. Deadlocks are avoided by maintaining a strict locking order—the lock on the predicate must be acquired before `listLock`.

It is not necessary for the queue of blocked threads to be a part of the condition structure itself. Instead, we may have a global set of sleep queues as in traditional UNIX. In that case, the `listLock` in the condition is replaced by a mutex protecting the appropriate sleep queue. Both methods have their own advantages, as discussed earlier.

One of the major advantages of a condition variable is that it provides two ways to handle event completion. When an event occurs, there is the option of waking up just one thread with `do_signal()` or all threads with `do_broadcast()`. Each may be appropriate in different circumstances. In the case of the server application, waking one thread is sufficient, since each request is handled by a single thread. However, consider several threads running the same program, thus sharing a single copy of the program text. More than one of these threads may try to access the same nonresident page of the text, resulting in page faults in each of them. The first thread to fault initiates a disk access for that page. The other threads notice that the read has already been issued and block waiting for the I/O to complete. When the page is read into memory, it is desirable to call `do_broadcast()` and wake up all the blocked threads, since they can all access the page without conflict.

7.7.2 Events

Frequently, the predicate of the condition is simple. Threads need to wait for a particular task to complete. The completion may be flagged by setting a global variable. This situation may be better expressed by a higher-level abstraction called an *event* that combines a *done* flag, the spin lock protecting it, and the condition variable into a single object. The event object presents a simple interface, allowing two basic operations—`awaitDone()` and `setDone()`. `awaitDone()` blocks until the event occurs, while `setDone()` marks the `event` as having occurred and wakes up all threads blocked on it. In addition, the interface may support a nonblocking `testDone()` function and a `reset()` function, which once again marks the `event` as not done. In some cases, the boolean *done* flag may be replaced by a variable that returns more descriptive completion information when the event occurs.

7.7.3 Blocking Locks

Often, a resource must be locked for a long period of time and the thread holding this lock must be permitted to block on other events. Thus a thread that needs the resource cannot afford to spin until the resource becomes available, and must block instead. This requires a *blocking lock* primitive that offers two basic operations—`lock()` and `unlock()`—and optionally, a `tryLock()`. Again there are two objects to synchronize—the *locked* flag on the resource and the sleep queue—which means that we need a spin lock to guarantee atomicity of the operations. Such locks may be trivially implemented using condition variables, with the predicate being the clearing of the locked flag. For per-

formance reasons, blocking locks might be provided as fundamental primitives. In particular, if each resource has its own sleep queue, a single spin lock might protect both the flag and the queue.

7.8 Read-Write Locks

Although modification of a resource requires exclusive access, it is usually acceptable to allow several threads to simultaneously read the shared data, as long as no one is trying to write to it at that time. This requires a complex lock that permits both shared and exclusive modes of access. Such a facility may be built on top of simple locks and conditions [Birr 89]. Before we look at an implementation, let us examine the desired semantics. A read-write lock may permit either a single writer or multiple readers. The basic operations are `lockShared()`, `lockExclusive()`, `unlockShared()` and `unlockExclusive()`. In addition, there might be `tryLockShared()` and `tryLockExclusive()`, which return `FALSE` instead of blocking, and also `upgrade()` and `downgrade()`, which convert a shared lock to exclusive and vice versa. A `lockShared()` operation must block if there is an exclusive lock present, whereas `lockExclusive()` must block if there is either an exclusive or shared lock on the resource.

7.8.1 Design Considerations

What should a thread do when releasing a lock? The traditional UNIX solution is to wake up all threads waiting for the resource. This is clearly inefficient—if a writer acquires the lock next, other readers and writers will have to go back to sleep; if a reader acquires the lock, other writers will have to go back to sleep. It is preferable to find a protocol that avoids needless wakeups.

If a reader releases a resource, it takes no action if other readers are still active. When the last active reader releases its shared lock, it must wake up a single waiting writer.

When a writer releases its lock, it must choose whether to wake up another writer or the other readers (assuming both readers and writers are waiting). If writers are given preference, the readers could starve indefinitely under heavy contention. The preferred solution is to wake up all waiting readers when releasing an exclusive lock. If there are no waiting readers, we wake up a single waiting writer.

This scheme can lead to writer starvation. If there is a constant stream of readers, they will keep the resource read-locked, and the writer will never acquire the lock. To avoid this situation, a `lockShared()` request must block if there is any waiting writer, even though the resource is currently only read-locked. Such a solution, under heavy contention, will alternate access between individual writers and batches of readers.

The `upgrade()` function must be careful to avoid deadlocks. A deadlock can occur unless the implementation takes care to give preference to upgrade requests over waiting writers. If two threads try to upgrade a lock, each would block since the other holds a shared lock. One way to avoid that is for `upgrade()` to release the shared lock before blocking if it cannot get the exclusive lock immediately. This results in additional problems for the user, since another thread could have modified the object before `upgrade()` returns. Another solution is for `upgrade()` to fail and release the shared lock if there is another pending upgrade.

7.8.2 Implementation

Example 7-9 implements a read-write lock facility:

```
struct rwlock   {
    int nActive;     /* num of active readers, or -1 if a writer is active */
    int nPendingReads;
    int nPendingWrites;
    spinlock_t sl;
    condition canRead;
    condition canWrite;
};

void lockShared (struct rwlock *r)
{
    spin_lock (&r->sl);
    r->nPendingReads++;
    if (r->nPendingWrites > 0)
        wait (&r->canRead, &r->sl); /* don't starve writers */
    while (r->nActive < 0)          /* someone has exclusive lock */
        wait (&r->canRead, &r->sl);
    r->nActive++;
    r->nPendingReads--;
    spin_unlock (&r->sl);
}

void unlockShared (struct rwlock *r)
{
    spin_lock (&r->sl);
    r->nActive--;
    if (r->nActive == 0)    {       /* no other readers */
        spin_unlock (&r->sl);
        do_signal (&r->canWrite);
    } else
        spin_unlock (&r->sl);
}

void lockExclusive (struct rwlock *r)
{
    spin_lock (&r->sl);
    r->nPendingWrites++;
    while (r->nActive)
        wait (&r->canWrite, &r->sl);
    r->nPendingWrites--;
```

```
    r->nActive = -1;
    spin_unlock (&r->sl);
}

void unlockExclusive (struct rwlock *r)
{
    boolean_t wakeReaders;
    spin_lock (&r->sl);
    r->nActive = 0;
    wakeReaders = (r->nPendingReads != 0);
    spin_unlock (&r->sl);
    if (wakeReaders)
        do_broadcast (&r->canRead); /* wake all readers */
    else
        do_signal (&r->canWrite);   /* wake a single writer */
}

void downgrade (struct rwlock *r)
{
    boolean_t wakeReaders;
    spin_lock (&r->sl);
    r->nActive = 1;
    wakeReaders = (r->nPendingReads != 0);
    spin_unlock (&r->sl);
    if (wakeReaders)
        do_broadcast (&r->canRead); /* wake all readers */
}

void upgrade (struct rwlock *r)
{
    spin_lock (&r->sl);
    if (r->nActive == 1) {          /* no other reader */
        r->nActive = -1;
    } else {
        r->nPendingWrites++;
        r->nActive--;               /* release shared lock */
        while (r->nActive)
            wait (&r->canWrite, &r->sl);
        r->nPendingWrites--;
        r->nActive = -1;
    }
    spin_unlock (&r->sl);
}
```

Example 7-9. Implementation of read-write locks.

7.9 Reference Counts

Although a lock may protect the data inside an object, we frequently need another mechanism to protect the object itself. Many kernel objects are dynamically allocated and deallocated. If a thread deallocates such an object, other threads have no way of knowing it and may try to access the object using a direct pointer (which they earlier acquired) to it. In the meantime, the kernel could have reallocated the memory to a different object, leading to severe system corruption.

If a thread has a pointer to an object, it expects the pointer to be valid until the thread relinquishes it. The kernel can guarantee it by associating a reference count with each such object. The kernel sets the count to one when it first allocates the object (thus generating the first pointer). It increments the count each time it generates a new pointer to the object.

This way, when a thread gets a pointer to an object, it really acquires a reference to it. It is the thread's responsibility to release the reference when no longer needed, at which time the kernel decrements the object's reference count. When the count reaches zero, no thread has a valid reference to the object, and the kernel may deallocate the object.

For example, the file system maintains reference counts for *vnodes*, which hold information about active files (see Section 8.7). When a user opens a file, the kernel returns a file descriptor, which constitutes a reference to the vnode. The user passes the descriptor to subsequent *read* and *write* system calls, allowing the kernel to access the file quickly without repeated name translations. When the user closes the file, the reference is released. If several users have the same file open, they reference the same vnode. When the last user closes the file, the kernel can deallocate the vnode.

The previous example shows that reference counts are useful in a uniprocessor system as well. They are even more essential for multiprocessors, since without proper reference counting, a thread may deallocate an object while a thread on another processor is actively accessing it.

7.10 Other Considerations

There are several other factors to consider in the design of a complex locking facility and the way in which the locks are used. This section examines some important issues.

7.10.1 Deadlock Avoidance

It is often necessary for a thread to hold locks on multiple resources. For instance, the implementation of condition variables described in Section 7.7 uses two mutexes: one protects the data and predicate of the condition, while the other protects the linked list of threads blocked on the condition. Trying to acquire multiple locks lead to a deadlock, as illustrated in Figure 7-8. Thread **T1** holds resource **R1** and tries to acquire resource **R2**. At the same time, thread **T2** may be holding **R2** and trying to acquire **R1**. Neither thread can make progress.

The two common deadlock avoidance techniques are *hierarchical locking* and *stochastic locking*. Hierarchical locking imposes an order on related locks and requires that all threads take locks in the same order. In the case of condition variables, for instance, a thread must lock the condition's predicate before locking the linked list. As long as the ordering is strictly followed, deadlock cannot occur.

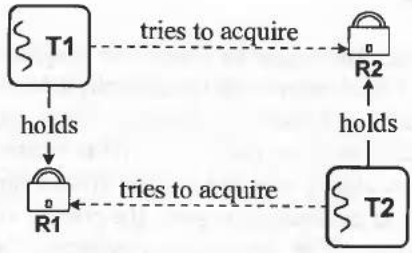

Figure 7-8. Possible deadlock when using spin locks.

There are situations in which the ordering must be violated. Consider a buffer cache implementation that maintains disk block buffers on a doubly linked list, sorted in *least recently used (LRU)* order. All buffers that are not actively in use are on the LRU list. A single spin lock protects both the queue header and the forward and backward pointers in the buffers on that queue. Each buffer also has a spin lock to protect the other information in the buffer. This lock must be held while the buffer is actively in use.

When a thread wants a particular disk block, it locates the buffer (using hash queues or other pointers not relevant to this discussion) and locks it. It then locks the LRU list in order to remove the buffer from it. Thus the normal locking order is *"first the buffer, then the list."*

Sometimes a thread simply wants any free buffer and tries to get it from the head of the LRU list. It first locks the list, then locks the first buffer on the list and removes it from the list. This, however, reverses the locking order, since it locks the list before the buffer.

It is easy to see how a deadlock can occur. One thread locks the buffer at the head of the list and tries to lock the list. At the same time, another thread that has locked the list tries to lock the buffer at the head. Each will block waiting for the other to release the lock.

The kernel uses stochastic locking to handle this situation. When a thread attempts to acquire a lock that would violate the hierarchy, it uses a `try_lock()` operation instead of `lock()`. This function attempts to acquire the lock, but returns failure instead of blocking if the lock is already held. In this example, the thread that wants to get any free buffer will lock the list and then go down the list, using `try_lock()` until it finds a buffer it can lock. Example 7-10 describes an implementation of `try_lock()` for spin locks:

```
int try_lock (spinlock_t *s)
{
    if (test_and_set (s) != 0)  /* already locked */
        return FAILURE;
    else
        return SUCCESS;
}
```

Example 7-10. Implementation of `try_lock()`.

7.10.2 Recursive Locks

A lock is recursive if an attempt by a thread to acquire a lock it already owns would succeed without blocking. Why is this a desirable feature? Why should a thread attempt to lock something it has already locked? The typical scenario has a thread locking a resource, then calling a lower-level routine to perform some operation on it. This lower-level routine may also be used by other higher-level routines that do not lock the resource prior to calling it. Thus the lower-level routine does not know if the resource is already locked. If it tries to lock the resource, a single process deadlock may occur.

Such a situation can, of course, be avoided by explicitly informing the lower-level routine about the lock via an extra argument. This, however, breaks many existing interfaces and is awkward, since sometimes the lower routine may be several function calls down. The resulting interfaces would be extremely nonmodular. An alternative is to allow the locks to be recursive. This adds some overhead, since the lock must now store some sort of owner ID and check it any time it would normally block or deny a request. More important, it allows functions to deal only with their own locking requirements, without worrying about which locks its callers hold, resulting in clean, modular interfaces.

One example when such a lock is used is directory writes in the BSD file system (ufs). The routine `ufs_write()` handles writes to both files and directories. Requests for file writes usually access the file through the file table entry, which directly gives a pointer to the file's *vnode*. Thus the vnode is passed on directly to `ufs_write()`, which is responsible for locking it. For a directory write, however, the directory vnode is acquired by the pathname traversal routine, which returns the vnode in a locked state. When `ufs_write()` is called for this node, it will deadlock if the lock is not recursive.

7.10.3 To Block or to Spin

Most complex locks can be implemented as blocking locks or as complex spin locks, without impacting their functionality or interface. Consider an object that is protected by a complex lock (such as a semaphore or a read-write lock). In most of the implementations described in this chapter, if a thread tries to acquire the object and finds it locked, the thread blocks until the object is released. The thread could just as easily busy-wait and still preserve the semantics of the lock.

The choice between blocking and busy-waiting is often dictated by performance considerations. Since busy-waiting ties up a processor, it is generally frowned upon. However, certain situations mandate busy-waiting. If the thread already holds a simple mutex, it is not allowed to block. If the thread tries to acquire another simple mutex, it will busy-wait; if it tries to acquire a complex lock, it will release the mutex it already holds (such as with conditions).

Sleep and wakeup, however, are expensive operations themselves, involving one context switch at each end and manipulating sleep and scheduler queues. Just as it is preposterous to do a busy-wait on a lock for an extended period of time, it is inefficient to sleep on a resource that is likely to be available soon.

Moreover, some resources may be subject to short-term or long-term locking, depending on the situation. For instance, the kernel may keep a partial list of free disk blocks in memory. When this list becomes empty, it must be replenished from disk. In most instances, the list needs to be

locked briefly while adding or removing entries to it in memory. When disk I/O is required, the list must be locked for a long time. Thus neither a spin lock nor a blocking lock is by itself a good solution. One alternative is to provide two locks, with the blocking lock being used only when the list is being replenished. It is preferable, however, to have a more flexible locking primitive.

These issues can be effectively addressed by storing a hint in the lock that suggests whether contending threads should spin or block. The hint is set by the owner of the lock and examined whenever an attempt to acquire the lock does not immediately succeed. The hint may be either advisory or mandatory.

An alternative solution is provided by the *adaptive locks* of Solaris 2.*x* [Eykh 92]. When a thread **T1** tries to acquire an adaptive lock held by another thread **T2**, it checks to see if **T2** is currently active on any processor. As long as **T2** is active, **T1** executes a busy-wait. If **T2** is blocked, **T1** blocks as well.

7.10.4 What to Lock

A lock can protect several things—data, predicates, invariants, or operations. Reader-writer locks, for example, protect data. Condition variables are associated with predicates. An invariant is similar to a predicate but has slightly different semantics. When a lock protects an invariant, it means that the invariant is TRUE except when the lock is held. For instance, a linked list might use a single lock while adding or removing elements. The invariant protected by this lock is that the list is in a consistent state.

Finally, a lock can control access to an operation or function. This restricts the execution of that code to at most one processor at a time, even if different data structures are involved. The *monitors* model of synchronization [Hoar 74] is based on this approach. Many UNIX variants use a master processor for unparallelized (not multiprocessor-safe) portions of the kernel, thus serializing access to such code. This approach usually leads to severe bottlenecks and should be avoided if possible.

7.10.5 Granularity and Duration

System performance depends greatly on the locking granularity. At one extreme, some asymmetric multiprocessing systems run all kernel code on the master processor, thus having a single lock for the whole kernel.[3] At the other extreme, a system could use extremely fine-grained locking with a separate lock for each data variable. Clearly, that is not the ideal solution either. The locks would consume a large amount of memory, performance would suffer due to the overhead of constantly acquiring and releasing locks, and the chances of deadlock would increase because it is difficult to enforce a locking order for such a large number of objects.

The ideal solution, as usual, lies somewhere in between, and there is no consensus on what it is. Proponents of coarse-granularity locking [Sink 88] suggest starting with a small number of locks to protect major subsystems and adding finer granularity locks only where the system exhibits a

[3] Sometimes, even SMP systems use a master processor to run code that is not multiprocessor-safe. This is known as *funneling*.

bottleneck. Systems such as Mach, however, use a fine-grained locking structure and associate locks with individual data objects.

Locking duration, too, must be carefully examined. It is best to hold the lock for as short a time as possible, so as to minimize contention on it. Sometimes, however, this may result in extra locking and unlocking. Suppose a thread needs to perform two operations on an object, both requiring a lock on it. In between the two operations, the thread needs to do some unrelated work. It could unlock the object after the first operation and lock it again for the second one. It might be better, instead, to keep the object locked the whole time, provided that the unrelated work is fairly short. Such decisions must be made on a case-by-case basis.

7.11 Case Studies

The primitives described in Sections 7.5–7.8 constitute a kind of grab bag from which an operating system can mix and match to provide a comprehensive synchronization interface. This section examines the synchronization facilities in the major multiprocessing variants of UNIX.

7.11.1 SVR4.2/MP

SVR4.2/MP is the multiprocessor version of SVR4.2. It provides four types of locks—basic locks, sleep locks, read-write locks, and synchronization variables [UNIX 92]. Each lock must be explicitly allocated and deallocated through *xxx*_ALLOC and *xxx*_DEALLOC operations, where *xxx*_ is the type-specific prefix. The allocation operation takes arguments that are used for debugging.

Basic Locks

The basic lock is a nonrecursive mutex lock that allows short-term locking of resources. It may not be held across a blocking operation. It is implemented as a variable of type `lock_t`. It is locked and unlocked by the following operations:

```
pl_t LOCK (lock_t *lockp, pl_t new_ipl);
UNLOCK (lock_t *lockp, pl_t old_ipl);
```

The `LOCK` call raises the interrupt priority level to `new_ipl` before acquiring the lock and returns the previous priority level. This value must be passed to the `UNLOCK` operation, so that it may restore the *ipl* to the old level.

Read-Write Locks

A *read-write lock* is a nonrecursive lock that allows short-term locking with single-writer, multiple-reader semantics. It may not be held across a blocking operation. It is implemented as a variable of type `rwlock_t` and provides the following operations:

```
pl_t RW_RDLOCK (rwlock_t *lockp, pl_t new_ipl);
pl_t RW_WRLOCK (rwlock_t *lockp, pl_t new_ipl);
void RW_UNLOCK (rwlock_t *lockp, pl_t old_ipl);
```

The treatment of interrupt priorities is identical to that for basic locks. The locking operations raise the *ipl* to the specified level and return the previous *ipl*. RW_UNLOCK restores the *ipl* to the old level. The lock also provides nonblocking operations RW_TRYRDLOCK and RW_TRYWRLOCK.

Sleep Locks

A *sleep lock* is a nonrecursive mutex lock that permits long-term locking of resources. It may be held across a blocking operation. It is implemented as a variable of type sleep_t, and provides the following operations:

```
void SLEEP_LOCK (sleep_t *lockp, int pri);
bool_t SLEEP_LOCK_SIG (sleep_t *lockp, int pri);
void SLEEP_UNLOCK (sleep_t *lockp);
```

The pri parameter specifies the scheduling priority to assign to the process after it awakens. If a process blocks on a call to SLEEP_LOCK, it will not be interrupted by a signal. If it blocks on a call to SLEEP_LOCK_SIG, a signal will interrupt the process; the call returns TRUE if the lock is acquired, and FALSE if the sleep was interrupted. The lock also provides other operations, such as SLEEP_LOCK_AVAIL (checks if lock is available), SLEEP_LOCKOWNED (checks if caller owns the lock), and SLEEP_TRYLOCK (returns failure instead of blocking if lock cannot be acquired).

Synchronization Variables

A *synchronization variable* is identical to the *condition variables* discussed in Section 7.7. It is implemented as a variable of type sv_t, and its predicate, which is managed separately by users of the lock, must be protected by a basic lock. It supports the following operations:

```
void SV_WAIT (sv_t *svp, int pri, lock_t *lockp);
bool_t SV_WAIT_SIG (sv_t *svp, int pri, lock_t *lockp);
void SV_SIGNAL (sv_t *svp, int flags);
void SV_BROADCAST (sv_t *svp, int flags);
```

As in sleep locks, the pri argument specifies the scheduling priority to assign to the process after it wakes up, and SV_WAIT_SIG allows interruption by a signal. The lockp argument is used to pass a pointer to the basic lock protecting the predicate of the condition. The caller must hold lockp before calling SV_WAIT or SV_WAIT_SIG. The kernel atomically blocks the caller and releases lockp. When the caller returns from SV_WAIT or SV_WAIT_SIG, lockp is not held. The predicate is not guaranteed to be true when the caller runs after blocking. Hence the call to SV_WAIT or SV_WAIT_SIG should be enclosed in a while loop that checks the predicate each time.

7.11.2 Digital UNIX

The Digital UNIX synchronization primitives are derived from those of Mach. There are two types of locks—simple and complex [Denh 94]. A simple lock is a basic spin lock, implemented using the atomic test-and-set instruction of the machine. It must be declared and initialized before being used.

It is initialized to the unlocked state and cannot be held across blocking operations or context switches.

The complex lock is a single high-level abstraction supporting a number of features, such as shared and exclusive access, blocking, and recursive locking. It is a reader-writer lock and provides two options—*sleep* and *recursive*. The sleep option can be enabled or disabled while initializing the lock or at any later time. If set, the kernel will block requesters if the lock cannot be granted immediately. Further, the sleep option must be set if a thread wishes to block while holding the lock. The recursive option can be set only by a thread that has acquired the lock for exclusive use; it can be cleared only by the same thread that set the option.

The interface provides nonblocking versions of various routines, which return failure if the lock cannot be acquired immediately. There are also functions to upgrade (shared to exclusive) or downgrade (exclusive to shared). The upgrade routine will release the shared lock and return failure if there is another pending upgrade request. The nonblocking version of upgrade returns failure but does not drop the shared lock in this situation.

Sleep and Wakeup

Since a lot of code was ported from 4BSD, it was desirable to retain the basic functions of `sleep()` and `wakeup()`. Thus blocked threads were put on global sleep queues rather than per-lock queues. The algorithms needed modification to work correctly on multiprocessors, and the chief issue here was the lost wakeup problem. To fix this while retaining the framework of thread states, the `sleep()` function was rewritten using two lower-level primitives—`assert_wait()` and `thread_block()`, as shown in Figure 7-9.

Suppose a thread needs to wait for an event that is described by a predicate and a spin lock that protects it. The thread acquires the spin lock and then tests the predicate. If the thread needs to

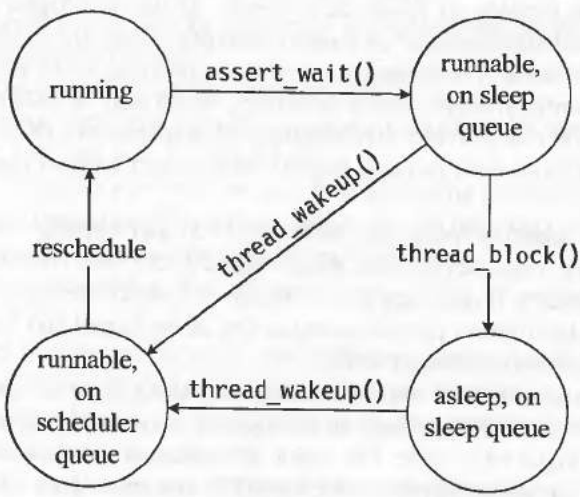

Figure 7-9. Sleep implementation in Digital UNIX.

block, it calls `assert_wait()`, which puts it on the appropriate sleep queue. It then releases the spin lock and calls `thread_block()` to initiate a context switch. If the event occurs between the release of the spin lock and the context switch, the kernel will remove the thread from the sleep queue and put it on the scheduler queue. Thus the thread does not lose the wakeup.

7.11.3 Other Implementations

The initial multiprocessing version of SVR4 was developed at NCR [Camp 91]. This version introduced the concept of *advisory processor locks* (APLs), which are recursive locks containing a *hint* for contending threads. The hint specifies if contending threads should spin or sleep, and whether the hint is advisory or mandatory. The thread owning the lock may change the hint from sleep to spin or vice versa. Such APLs may not be held across calls to sleep and are used mainly to single-thread the access to longer-term locks. The distinctive feature of APLs is that they are automatically released and reacquired across a context switch. This means that the traditional sleep/wakeup interface can be used without change. Further, locks of the same class can avoid deadlock by sleeping, since the sleep releases all previously held locks. The implementation also provided nonrecursive spin locks and read-write APLs.

The NCR version was modified by the Intel Multiprocessor Consortium, formed by a group of companies to develop the official multiprocessing release of SVR4 [Peac 92]. One important change involved the function that acquires an APL lock. This function now also takes an interrupt priority level as an argument. This allows raising of the processor priority around the holding of the lock. If the lock cannot be granted immediately, the busy-wait occurs at the original (lower) priority. The function returns the original priority, which can be passed to the unlocking function.

The lowest-level primitives are a set of atomic arithmetic and logical operations. The arithmetic operations allow atomic incrementing and decrementing of reference counts. The logical functions are used for fine-grained multithreading of bit manipulations of flag fields. They all return the original value of the variable on which they operate. At the next higher level are simple spin locks that are not released automatically on context switches. These are used for simple operations like queue insertion or removal. The highest-level locks are *resource locks*. Resource locks are long-term locks with single-writer, multiple-reader semantics, which may be held across blocking operations. The implementation also provides synchronous and asynchronous *cross-processor interrupts*, which are used for operations such as *clock tick distribution* and *address translation cache coherency* (see Section 15.9).

Solaris 2.*x* uses adaptive locks (see Section 7.10.3) and turnstiles (see Section 7.2.3) for better performance. It provides semaphores, reader-writer locks, and condition variables as high-level synchronization objects. It also uses kernel threads to handle interrupts, so that interrupt handlers use the same synchronization primitives as the rest of the kernel and block if necessary. Section 3.6.5 discusses this feature in greater detail.

Every known multiprocessor implementation uses some form of spin locks for low-level short-term synchronization. The sleep/wakeup mechanism is usually retained, perhaps with some changes, to avoid rewriting a lot of code. The major differences are in the choice of the higher-level abstractions. The early implementations on the IBM/370 and the AT&T 3B20A [Bach 84] relied exclusively on semaphores. Ultrix [Sink 88] uses blocking exclusive locks. Amdahl's UTS kernel

[Ruan 90] is based on conditions. DG/UX [Kell 89] uses *indivisible event counters* to implement *sequenced locks*, which provide a somewhat different way of waking one process at a time.

7.12 Summary

Synchronization problems on a multiprocessor are intrinsically different from and more complex than those on a uniprocessor. There are a number of different solutions, such as sleep/wakeup, conditions, events, read-write locks, and semaphores. These primitives are more similar than different, and it is possible, for example, to implement semaphores on top of conditions and vice-versa. Many of these solutions are not limited to multiprocessors and may also be applied to synchronization problems on uniprocessors or on loosely coupled distributed systems. Many multiprocessing UNIX systems are based on existing uniprocessing variants and, for these, porting considerations strongly influence the decision of which abstraction to use. Mach and Mach-based systems are mainly free of these considerations, which is reflected in their choice of primitives.

7.13 Exercises

1. Many systems have a swap-atomic instruction that swaps the value of a register with that of a memory location. Show how such an instruction may be used to implement an atomic test-and-set.
2. How can an atomic test-and-set be implemented on a machine using load-linked and store-conditional?
3. Suppose a convoy forms due to heavy contention on a critical region that is protected by a semaphore. If the region could be divided into two critical regions, each protected by a separate semaphore, would it reduce the convoy problem?
4. One way to eliminate a convoy is to replace the semaphore with another locking mechanism. Could this risk starvation of threads?
5. How is a reference count different from a shared lock?
6. Implement a blocking lock on a resource, using a spin lock and a condition variable, with a locked flag as the predicate (see Section 7.7.3).
7. In exercise 6, is it necessary to hold the spin lock protecting the predicate while clearing the flag? [Ruan 90] discusses a `waitlock()` operation that can improve this algorithm.
8. How do condition variables avoid the lost wakeup problem?
9. Implement an *event* abstraction that returns a status value to waiting threads upon event completion.
10. Suppose an object is accessed frequently for reading or writing. In what situations is it better to protect it with a simple mutex, rather than with a read-write lock?
11. Does a read-write lock have to be blocking? Implement a read-write lock that causes threads to busy-wait if the resoure is locked.
12. Describe a situation in which a deadlock may be avoided by making the locking granularity *finer*.

13. Describe a situation in which a deadlock may be avoided by making the locking granularity *coarser*.
14. Is it necessary for a multiprocessor kernel to lock each variable or resource before accessing it? Enumerate the kinds of situations where a thread may access or modify an object without locking it.
15. Monitors [Hoar 74] are language-supported constructs providing mutual exclusion to a region of code. For what sort of situations do they form a natural solution?
16. Implement `upgrade()` and `downgrade()` functions to the read-write lock implementation in Section 7.8.2.

7.14 References

[Bach 84] Bach, M., and Buroff, S., "Multiprocessor UNIX Operating Systems," *AT&T Bell Laboratories Technical Journal,* Vol. 63, Oct. 1984, pp. 1733–1749.

[Birr 89] Birrell, A.D., "An Introduction to Programming with Threads," Digital Equipment Corporation Systems Research Center, 1989.

[Camp 91] Campbell, M., Barton, R., Browning, J., Cervenka, D., Curry, B., Davis, T., Edmonds, T., Holt, R., Slice, R., Smith, T., and Wescott, R., "The Parallelization of UNIX System V Release 4.0," *Proceedings of the Winter 1991 USENIX Conference,* Jan. 1991, pp. 307–323.

[Denh 94] Denham, J.M., Long, P., and Woodward, J.A., "DEC OSF/1 Version 3.0 Symmetric Multiprocessing Implementation," *Digital Technical Journal,* Vol. 6, No. 3, Summer 1994, pp. 29–54.

[Digi 87] Digital Equipment Corporation, *VAX Architecture Reference Manual,* 1984.

[Dijk 65] Dijkstra, E.W., "Solution of a Problem in Concurrent Programming Control," *Communications of the ACM,* Vol. 8, Sep. 1965, pp. 569–578.

[Eykh 92] Eykholt, J.R., Kleinman, S.R., Barton, S., Faulkner, R., Shivalingiah, A., Smith, M., Stein, D., Voll, J., Weeks, M., and Williams, D., "Beyond Multiprocessing: Multithreading the SunOS Kernel," *Proceedings of the Summer 1992 USENIX Conference,* Jun. 1992, pp. 11–18.

[Gobl 81] Goble, G.H., "A Dual-Processor VAX 11/780," *USENIX Association Conference Proceedings,* Sep. 1981.

[Hoar 74] Hoare, C. A.R., "Monitors: An Operating System Structuring Concept," *Communications of the ACM,* Vol. 17, Oct. 1974, pp.549–557.

[Hitz 90] Hitz, D., Harris, G., Lau, J.K., and Schwartz, A.M., "Using UNIX as One Component of a Lightweight Distributed Kernel for Multiprocessor File Servers," *Proceedings of the Winter 1990 USENIX Technical Conference,* Jan. 1990, pp. 285–295.

[Kell 89] Kelley, M.H., "Multiprocessor Aspects of the DG/UX Kernel," *Proceedings of the Winter 1989 USENIX Conference,* Jan. 1989, pp. 85–99.

[Lee 87] Lee, T.P., and Luppi, M.W., "Solving Performance Problems on a Multiprocessor UNIX System," *Proceedings of the Summer 1987 USENIX Conference,* Jun. 1987, pp. 399–405.

[Nati 84] National Semiconductor Corporation, *Series 32000 Instruction Set Reference Manual,* 1984.

[Peac 92] Peacock, J.K., Saxena, S., Thomas, D., Yang, F., and Yu, F., "Experiences from Multithreading System V Release 4," *The Third USENIX Symposium of Experiences with Distributed and Multiprocessor Systems (SEDMS III),* Mar. 1992, pp. 77–91.

[Ruan 90] Ruane, L.M., "Process Synchronization in the UTS Kernel," *Computing Systems,* Vol. 3, Summer 1990, pp. 387–421.

[Sink 88] Sinkewicz, U., "A Strategy for SMP ULTRIX," *Proceedings of the Summer 1988 USENIX Technical Conference,* Jun. 1988, pp. 203–212.

[UNIX 92] UNIX System Laboratories, *Device Driver Reference—UNIX SVR4.2,* UNIX Press, Prentice-Hall, Englewood Cliffs, NJ, 1992.