CHAPTER

6 Interprocess Communications

Learning Objectives

After reading this chapter, you should be able to:

- Describe the purpose and elements of interprocess communications.
- Discuss the two models of interprocess communications—shared memory region and message passing.
- Explain the various interprocess communication schemes such as signal system, shared memory, message queue, pipe, and FIFO.

6.1 Introduction

Communication is a means of sharing a piece of information between at least two agents. A communication consists of an individual sending some information and another individual receiving that information. Five elements are involved in a communication: (1) the situation (when, where, and what circumstance) that triggers the communication, (2) the information to be communicated, (3) the sender of the information, (4) the medium of transmitting the information, and (5) the receiver of the information. The first element here is not of interest because the operating system only helps processes exchange information and does not create situations for production of information.

In systems involving multiple processes, concurrent processes may interact with one another for many reasons. Two primary objectives of process interactions are: (1) exchange of information to achieve some common goal, and (2) coordination (i.e., synchronization) of activities of the participating processes. Each interaction involves exchanging a finite amount of information between two or more processes. This exchange is accomplished by transferring data from one process to another. Thus, the two flavours comprising process interactions, namely, interprocess communication

≫ Communication is vital for any activity involving two or more agents. The agents involved in communications in operating systems are processes, threads, kernel paths, and processors. Unless stated otherwise, we use the process as the generic agent in this chapter. Communication can be established in many ways but its purpose is principally one, that is, to transfer information from one process to another. The agents "interact" among themselves through communications to achieve this objective.

(IPC) and process synchronization form the two fundamental concepts in designing multiprocess operating systems. We study IPC and related issues in this chapter, and synchronization and related issues in the next.

6.2 Interprocess Communications

Some applications create more than one process to perform application-specific tasks. These processes work together to collaboratively accomplish the tasks. They are called *cooperating processes*. They exchange information among themselves to perform their tasks. Their interaction is direct. Each process is aware of the presence of the other processes in the system, and its execution flow may depend on the executions of the other processes. In addition, the kernel also communicates with application processes in order to coordinate their activities.

At certain points in its execution, a process may decide to send a piece of data to another process. In such an event, the former process is called the *producer* of the data, and the latter the *consumer* of the data, and the data itself a *message*. Producers generate data and send them in the form of finite-size messages to the consumers. The latter receive these messages and consume the data contained in them. For example, a copy program produces blocks of data that are consumed by a disk server. One instance of exchange of data among two or more processes is called an *interprocess communication*, IPC. IPCs are vital in keeping cooperating processes informed about their collective activities.

In a typical communication between two processes, one process sends a message and the other receives it. In many applications, the roles of these two processes are fixed and the communication is repeated many times. Consider a simple example of two processes with a file containing video frames in an encoded form. The objective is to decode the frames and display them on the screen one by one in a specified order. One process is assigned to repeat the activity of producing decoded frames and the other the activity of displaying the decoded frames on the screen. The two processes are allowed to work concurrently. This type of communication commonly occurs in many large interactive applications. Due to its ubiquitous nature and importance, this pattern of communication is abstracted simply as a producer-consumer problem. Other recurring patterns of communications are abstracted with specific names and will be discussed later in this chapter.

Cooperating processes are "loosely" connected in the sense that they each have an independent private address space and proceed at different speeds. Their relative speed is unknown. One process cannot control the speed of another. From time to time, they exchange information among themselves. As noted in Section 1.5.2 on page 20, one process cannot access elements from the private address space of another. This implies that processes need help from the operating system to set up communication facilities among themselves. Modern operating systems support many IPC schemes for this purpose. For each such scheme, the operating system implements a few

communication primitives (i.e., interface operations). Processes perform local computations in their respective private address spaces, and execute these primitives in the kernel space to facilitate the transfer of data across their address space boundaries.

Many IPC schemes have been proposed in literature for various purposes. The schemes are classified into two broad categories: synchronous and asynchronous.

In a communication scheme where each sender waits until a receiver (or a specific receiver) is ready to receive the data is called a *synchronous communication* scheme. When both are ready for the communication, the data exchange takes place between them. Some authors call it a *handshake-* or *rendezvous* scheme.

In contrast, in an *asynchronous communication* scheme, a sender and a receiver may not handshake to exchange data. The sender does not wait until a receiver (or a specific receiver) is ready to receive the data. Instead, the sender stores the data in some a priori known temporary storage (called a *buffer* or *medium of information transmission*),[1] and the receiver obtains the data from that storage later. The buffer is persistent in the sense that until the receiver retrieves the data, it remains in the buffer. There are two alternatives for informing the receiver about the data.

- After putting the data in the buffer, the sender may interrupt the receiver. This is called pushing.
- Alternatively, the receiver may examine the buffer at regular intervals to see if there is any data intended for it. This is called polling.

In this chapter, we will discuss only a few asynchronous IPC schemes that are used in many modern operating systems. Before presenting those schemes, we present two models of IPC.

6.3 Interprocess Communication Models

Processes exchange information either by sending and receiving messages among themselves, or by reading- and writing data in a shared storage space. Thus, there arise two major IPC models: (1) message passing model, and (2) shared memory model. The semantics of communication primitives for these two models are quite different. Both models are widely used. Both are asynchronous communications schemes.

≫ Shared-memory models and message-passing models are functionally equivalent. That is, a program written for a shared-memory model can be transformed to a message-passing model and vice versa without changing the intent of the program.

6.3.1 Message-passing Model

In the message-passing model, an instance of information exchange involves copying data from one process address space into another via the kernel space (see Fig. 1.15 on page 29). The operating system allocates and then manages

[1] A *buffer* is a piece of storage space that is used to temporarily store data that is being transferred between two agents.

the required buffer space in the kernel to hold unreceived messages. A sender-process deposits a message in the buffer, and a receiver-process retrieves the message from the buffer later. There are many cases of this model. A sender-process may send a message to one particular process, to some processes in a group, or to all processes in a group.

Send and receive are the basic communication primitives in this model and they have many variants to emulate various cases. A send-primitive takes a message as its argument, and helps the sender to copy the message from its private address space into the kernel space. A receive-primitive takes a free block in the receiver's private address space as a parameter and copies a message from the kernel space into the block. Each send- or receive operation involves executing a system call as the data transfer takes place between the user space and the kernel space.

Before exchanging messages, processes need to set up communication buffers, often called channels. A channel is a communication medium in which senders deposit messages and receivers retrieve messages. There are many factors that control the behaviour of a channel such as: (1) Who is the owner of the channel? (2) Can the channel be accessed by multiple processes? (3) Is the channel unidirectional or bidirectional? (For a unidirectional channel, senders and receivers are disjoint.) (4) What is the capacity of the channel? (5) Does the channel handle fixed- or variable-size messages? (6) Does the channel support direct- or indirect addressing? (In direct addressing, the sender specifies a particular receiver to whom it wants to send a message, and a receiver specifies a particular sender from whom it wants to receive a message. In indirect addressing, a message sent can be received by any receiver, and a receiver can receive a message from any sender; a *mailbox* kind of buffer stores all unreceived messages.)

6.3.2 Shared-memory Model

In the shared-memory model, parts of process private address spaces are mapped to the same physical memory (see Fig. 1.16 on page 29) called *shared-memory regions*. The operating system helps processes in setting up and destroying shared-memory regions. Processes can directly write data into and read data from shared-memory regions. Read and write are the basic communication primitives in this model and they do not require making system calls. Processes themselves manage the space in shared regions on their own without intervention from the operating system.

6.3.3 Data Representation

All IPC schemes are modelled as presented in the schematic diagram in Fig. 6.1. Ultimately, all communications involve storing and retrieving information in some storage medium that is shared by all agents involved. To manipulate the available information easily, the shared space is organized in various data structures called shared data or shared variables. In general,

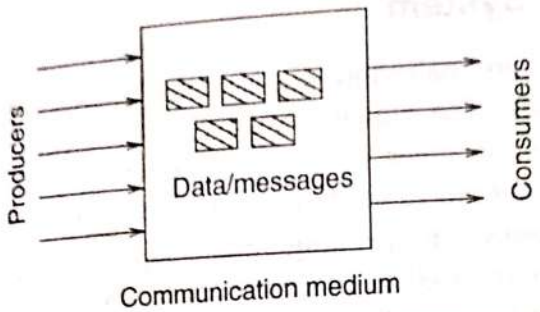

Figure 6.1: *Organizing the medium of information exchange for IPCs.*

a *shared variable* is an abstraction of persistent IPCs. Shared variables are logical units of objects for the purpose of manipulating information by processes. Each shared variable has a unique name or address, and a type. The type defines a finite domain of values, interface operations, and consistency semantics. The variable can store any value from the domain. Interface operations are the only means to access the shared variable. The semantics of the operations describe the permitted behaviour of the operations. Processes communicate among themselves by manipulating values of shared variables by executing operations supported by the variables. A sender process writes new values in the variables that are later read by the receiver processes.

Different IPC schemes use various types of shared variables, and implement different primitives to facilitate IPCs. Senders and receivers invoke these primitives in doing their IPCs. We may need to synchronize accesses to shared variables to ensure integrity of the shared data. (We discuss synchronization in the next chapter.)

If we have an unbounded temporary buffer that can hold an arbitrary number of data items, a producer will never wait to store a data item, but a consumer may need to wait for new data if the buffer is empty. In reality, the buffer is of finite capacity, and can store only a limited number of data items there. Depending on the situation, both consumers and producers may need to wait there. A producer waits only if there is no free space in the buffer, and a consumer waits only if there is no data in the buffer.

6.4 Interprocess Communication Schemes

Various IPC schemes have been proposed in literature (and many of them are implemented in some operating system or other). We discuss here IPC schemes from UNIX systems, and they include signal system, shared memory region, message queue, pipe, FIFO, and socket. One scheme differs from another by the way the temporary storage medium (i.e., buffer) is structured to store information and nature of communication primitives they provide. Except the shared memory region, all other schemes require processes to make system calls to transfer data (even of a single bit) across process address spaces. The above-mentioned six IPC schemes are discussed in the following six subsections.

6.4.1 Signal System

≫ There are three main sources of signals in UNIX. They are: (1) exception from the execution of current instruction, (2) a user action by an input device, and (3) execution of the kill system call.

The signal system, originally introduced in the UNIX system, is the simplest of all IPC schemes. The signal system follows a simpler form of asynchronous producer-consumer model of information exchange. It helps the processes (and the kernel) to send information about occurrences of events to one or more processes. An operating system supports only a limited number of signal types that are used to notify processes about occurrences of a fixed set of predetermined events. Each signal type represents a single event from this set. When an event occurs, a process (or the kernel) notifies its occurrence to a single process or to a group of processes by sending the corresponding signal information to them. For example, the kernel sends one specific signal to a process when one of its children processes exits. A keyboard interrupt can also generate a signal, and the kernel sends it to the process that owns the keyboard.

≫ Unlike other IPC schemes, a process does not need to explicitly set up a signal descriptor by itself; the operating system does so as part of the process creation operation at the time of the process descriptor initialization.

Each process has a reserved space in the kernel (usually, as a part of the process descriptor) to store signal-related information. This space is called the *signal descriptor*. A single bit variable stores information for a specific signal type, as shown in Fig 6.2. The bit pattern in the figure indicates at least two events (of types 1 and 4) occurred and the process is informed of the two events. There are two actions involved for each signal-sending operation, namely delivery and receipt. The former is performed by the kernel, and the latter by the target process. When a specific signal is sent to a process, the signal is considered to have been *delivered* to the process only when the kernel has set the corresponding bit variable in the signal descriptor of the process. When the process receives the signal, the kernel resets the bit variable. The signal subsystem does not record the signal-sender information. A delivered signal may not be received by the target process immediately. There can be at the most one unreceived signal of a given signal type; the other signals of the same type are considered *lost*. Consequently, if the same signal is sent to a process repeatedly without being received by the process, all but the last instance of the signal are considered lost. A signal sent to a process cannot be received by the process more than once.

≫ Signal handlers are procedures for processing (or catching) the signal. The operating system provides handlers to signals that it defines. A process can replace some of these handlers by its own handlers in order to deal with those signals in its customized own way.

A process eventually receives a signal sent to it unless the process terminates. However, if signals of two different types are delivered to a process at the same time, there is no guarantee of the order in which the process will receive the signals. There is no relative priority among signals. On receipt of a signal, the receiving process has one of two possible courses of actions: (1) ignore the signal (and hence the corresponding event), or (2) execute a special function called the *signal handler* to handle the occurrences of the corresponding event. To handle a particular signal in its own way, a process needs

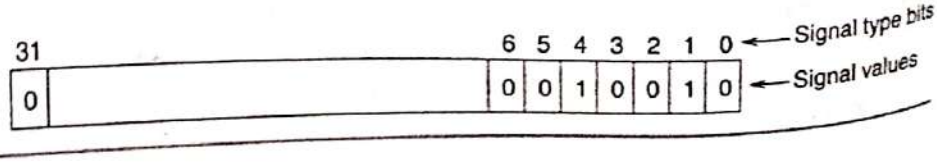

Figure 6.2: *A typical signal descriptor.*

to register a signal handler with the operating system for the said signal type. On receiving the signal, the process stops what it has been doing, and starts executing the signal handler. If a process does not explicitly inform the operating system of one of these two options for a signal and the signal is indeed sent to the process, the operating system performs a default action. Default actions depend on the signal type, and the actions generally are (1) ignore the signal, (2) terminate the process, (3) produce a core dump and terminate the process, (4) suspend the process, and (5) resume the suspended process.

A subtle point to note here is that a process receives signals in the kernel space. But, if there is a registered signal handler, it resides in the user space. The process cannot do a normal return from the kernel space to the user space to execute the signal handler. The kernel makes a kind of upcall in the user space. The process temporarily leaves the kernel space, executes the signal handler, and immediately returns to the kernel space when the handler execution is complete. Note that during the signal handler execution, the process can enter the kernel space by making system calls. (We will revisit Linux way of signal handling in Section 17.6.1.)

A process may temporarily block those signals that are of no interest to it. The process is not interested in these signal types for the present. Note that these signals can be sent to the process, and the kernel does deliver these signals to the process. However, the process will not receive blocked unreceived signals, neither will the kernel perform default actions until the process unblocks the signal types. For every signal type, the operating system maintains another bit variable in the signal descriptor to indicate whether or not the signal type is temporarily blocked. See signal descriptor in Fig. 6.3, signal type-4 is blocked by the process. It will not handle that unreceived signal until it unblocks the signal type.

POSIX standards define about 20 signals. (However, most systems support more than 20 signals. Normally the number of bits available in an integer variable limits the number of signals.) POSIX identifies each signal by a name prefixed with SIG—for example, SIGTERM, SIGSEGV, etc.

Signal sending and handling in single threaded systems is straightforward because all signals are sent and handled by the same thread. In multithreaded systems signals can be handled by different threads in the target process. All threads share the same signal handlers, but they can have different signal masks to block specific signal types. Synchronous signals (those that are generated by the executing thread, for example, division by zero) are delivered to the executing thread. Asynchronous signals come from outside the process, and for such signals we have a choice when the receiving process has multiple threads. Do we deliver an asynchronous signal to any thread, a few

≫ A user process can send signals only to the processes of the same user. A process run by the super- user or the kernel can send signals to any process in the system. Most often, signals are sent by the kernel. Initially, signals were simply binary values (whether received or not) and the reaction to the signal was always terminating the corresponding process (hence the command name **kill** to send signal). Subsequently, signals have been classified into various types and are handled differently.

≫ The operating system reserves certain critical signal types for its own purpose, and may not allow a process to block- or register handlers for those signal types. Linux handles process termination requests from the keyboard by sending the process a SIGTERM signal, and does not allow a process to block that signal. In Linux, exceptions are conveyed to the process through the signal mechanism.

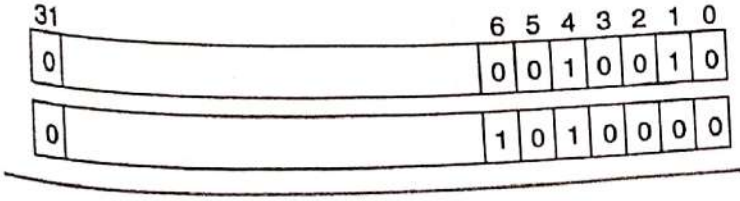

Figure 6.3: *Blocking of signals.*

particular threads, or all threads? Some systems allow threads to specify which signals they are willing to handle, and the system delivers those signals to those threads only. As each signal can only be received once by the process, the system may deliver the signal to the first thread that has not blocked the signal. There are other systems that create a default thread that handles all asynchronous signals.

≫ You may recall from Chapter 4 that threads of one process are not visible to another process. Consequently, one process cannot direct a signal to a particular thread in another process.

6.4.2 Shared Memory Region

A signal conveys only the minimum information about the occurrence of a predefined event, that is, the name of the event. Also, every signal communication is explicitly routed through the kernel. The size, content, and control restrictions are relaxed in the shared memory communication mechanism. A *shared memory* is an abstraction of persistent IPC, and it can hold a collection of related shared variables. Shared variables are used to exchange information among the cooperating processes. Each shared variable is accessed by a predefined set of primitive operations. Processes may execute these operations, often concurrently, to read- and write shared variables. The behaviour of operation executions is required to be "consistent" for effective IPCs.

≫ Shared memory is a generic and one of the most efficient IPC mechanisms. Since it can be accessed like any other memory area, programming using shared memory is straightforward.

Processes in general cannot access each other's private address spaces. The operating system prohibits such access violations. It helps processes in mapping parts of their address spaces to the same physical memory locations. The mapping is set up in the abstraction of memory regions. A *memory region* is a logical entity in an address space, which has a base address and a specified length. As shown in Fig. 6.4, regions (of different processes) of the same length can only be shared, that is, mapped to the same (sized) physical memory region. Once a shared memory region becomes a part of process address spaces, the corresponding processes can access the region without causing address violation exceptions.

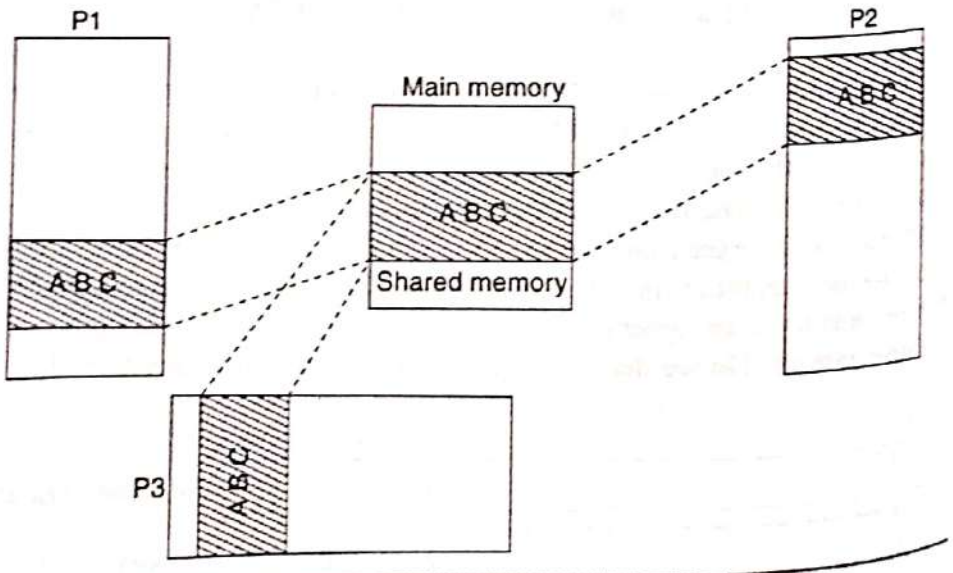

Figure 6.4: *An abstract view of shared memory.*

The operating system implements system call services to create new shared memory regions and to attach existing memory regions to process address spaces. The system associates a unique key with each shared memory region created in the system. Processes use the key to attach (and detach) a shared memory region to (and from) their address spaces. The processes can read from and write into a memory region by executing ordinary memory read and write instructions.

The operating system only creates shared memory regions and attaches them to processes. Granularity of data (i.e., data types) is decided by processes themselves, and not by the operating system. The synchronization of operation executions on shared data is solely done by applications themselves, and not by the operating system. In short, programs for manipulating data in shared memory regions are parts of applications themselves, and not of the operating system. The operating system helps processes in creation, maintenance, and destruction of shared memory regions.

≫ A shared memory region may be mapped to different logical addresses in different processes, see Fig. 6.4. The memory management system translates the different logical addresses to the same physical address pertaining to the shared memory region. We study memory management and runtime address translation schemes in Chapter 8.

6.4.3 Message Queue

A message queue is an abstraction of an asynchronous interprocess message communication scheme with a specified structure and access pattern. A *message queue* is a data structure (usually linked list) that is set up to allow one or more processes to add messages that are removed by other processes. A *message* here is a representation of some structured information, and it is finite in size. Message queues reside in the kernel space, but the operating system is not concerned about message contents. The system does not lose- or alter messages. A process sends a message by enqueuing the message in the message queue. The message is not addressed to a specific process. The message is later dequeued by another process. A message can be dequeued only once, and also, the same message cannot be dequeued by many processes. The message queue service discipline is strictly FIFO.

≫ For other IPC schemes that are discussed in this chapter, the synchronization is provided by the kernel.

The following message queue primitives are supported by some operating systems: (1) creating a new message queue, (2) opening an existing message queue, (3) sending a message to the queue, (4) receiving a message from the queue, (5) closing the queue, and (6) destroying the queue. Before the actual communications begin, a message queue must be created. When a message queue is created, the operating system allocates a finite amount of space in the kernel to temporarily store unreceived messages. The space depends on the size of the message queue, and is referred to as the *message buffer*. Processes open a message queue before exchanging messages through the queue. A producer produces a message and puts it in the message buffer from where a receiver obtains it. The message communication is subject to two resource constraints: (1) producers cannot send more messages than the buffer can hold, and (2) receivers cannot consume messages faster than producers produce them. When a producer attempts to put a message in a full buffer, it may be delayed until a receiver consumes another message from the buffer and makes some space available to the producer. Analogously, when a

≫ A message could be addressed to a receiver process explicitly or implicitly by placing it in a particular queue that the process has access to. In literature, the latter case is called a mailbox and in that sense a UNIX message queue is a mailbox. A message queue simply emulates a bounded size FIFO queue.

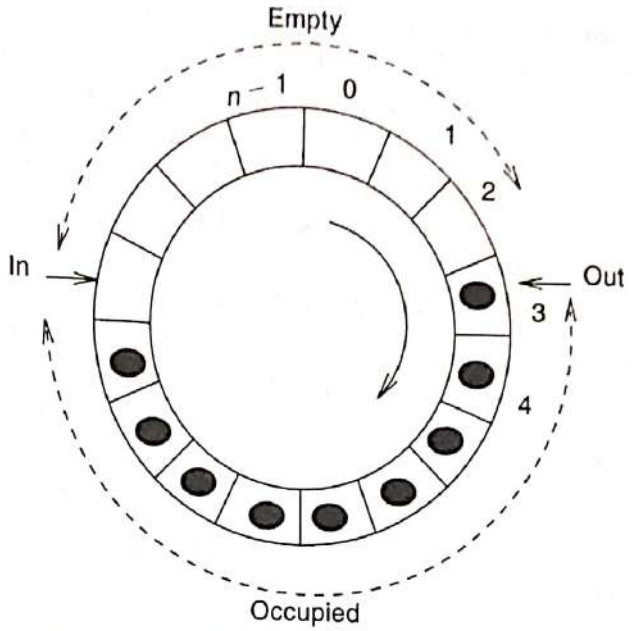

Figure 6.5: *Structure of a bounded message queue.*

receiver attempts to retrieve a message from an empty buffer, the receiver will be delayed until a producer puts a message in the buffer and makes it available to the receiver. To avoid indefinite blocking, a process may opt for the non-blocking mode of operation; when its operation cannot be completed without blocking, it is sent a negative status report instead of blocking it.

In reality, the message buffer of a message queue is treated as a circular storage of a finite number of message entries (see Fig. 6.5). The queue is called a bounded message queue. Figure 6.6 defines **put** and **get** primitives that are executed to send- and receive messages to and from a bounded message queue.[2] (Note that the routines do not take care of any synchronization-related problems; we will discuss them in Chapter 7. For the time being, we assume that processes execute the routines in a mutually exclusive manner and the routine executions are ordered in such a way that no process is blocked in the routines. We will revisit these routines in Section 7.5.9.)

There are n, $n > 0$, entries in the message buffer to hold at the most n unreceived messages. The n entries are manipulated to implement a bounded circular queue. Two integer variables *in* and *out* encompass all entries that contain unreceived messages. The "Occupied" arc in the figure shows this. The shaded ovals are unreceived messages. The integer variable *in* points to the entry where the next message will be put, while *out* points to the entry from where the next message will be received. The integer variable *count* refers to the number of unreceived messages available in the buffer. If *count* is 0, the buffer is empty, and a **get** operation will need to wait. If *count* is n, the buffer is full, and a **put** operation will need to wait. The *count* is incremented by 1 when a new message is put in the buffer, and decremented by 1 when a message is removed from the buffer. A producer produces a new

[2]To conserve print space, we will not present proofs of correctness of any programs in this book. We are primarily interested in learning concepts, and issues involved in designing and developing operating systems rather than finding the best- or absolutely correct solutions to these issues.

```
Constant
   n = maximum number of entries in the message buffer, where n > 0;

Type
   message = structure of buffer entry;

Data structures and initial values
   message buff[n]; /* an array of n message entries */
   int count = 0;      /* number of messages in the buffer; initially empty */
   int in = 0;         /* where the next message will be copied */
   int out = 0;        /* from where the next message will be returned */

void put (message* m)
{
   while (count == n) { wait; }  /* message buffer is full */
   buff[in] = *m;                /* copy message in the buffer */
   in = (in + 1)%n;
   count = count + 1;
}
void get (message *m)
{
   while (count == 0) { wait; }  /* message buffer is empty */
   *m = buff [out];              /* copy message from the buffer */
   out = (out + 1)%n
   count = count - 1;
}
```

Figure 6.6: *A typical implementation of a bounded message queue.*

message, and executes the **put** routine to copy the message in the buffer. A receiver executes the **get** routine to copy the oldest message from the buffer.

6.4.4 Pipe

A *pipe* (also called unnamed pipe) is a "one-way" flow of data between two related processes (see Fig. 6.7). It is a fixed size first-in first-out communication channel, and the size is system-dependent. For each pipe, one process writes data into the pipe, and another process reads the data out of the pipe. A pipe is just like a bounded-message queue. But, unlike a message queue, a pipe is shared only by two related processes, and messages are considered an arbitrary sequence of bytes without any message boundary. Reading from an empty pipe or writing into a full pipe causes a process to be blocked until the state of the pipe changes.

≫ As a pipe is an unnamed object, it cannot be shared by unrelated processes.

In UNIX systems, a pipe is implemented as a sequential file. However, a pipe has no name (i.e., it is anonymous) in the file system. The pipe allows users to funnel the output of one program execution into the input of another program execution, and each program execution sees the pipe as an ordinary (sequential) file. A process creates a pipe by executing the **pipe** system call, and then creates a child process. The two processes can communicate through the pipe. The **pipe** system call actually creates two file descriptors for the calling process using which the pipe is accessed. The system call also allocates space to hold unreceived data. Any process that has access to those file descriptors can communicate through the pipe. (When a parent creates a

≫ In UNIX systems, pipes are normally used as unidirectional byte streams which connect the standard output from one process to the standard input of another process. Neither process is aware of this redirection. They behave just as they would normally do in reading and writing their standard input and output.

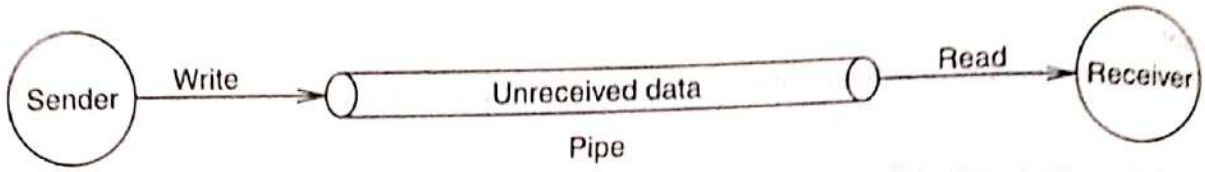

Figure 6.7: *An abstract view of a typical pipe.*

child, UNIX systems automatically duplicate file descriptors of the parent for the child. Thus, the child can access the pipe.) The pipe is closed when all processes close those descriptors. The operating system releases the pipe space. Though not shown in Fig. 6.7, in UNIX a process can both write to and read from the same pipe.

≫ In UNIX, the symbol | is used for pipe in command line. For example, the pipe between two popular commands **ls | more** takes the output of **ls** in one side of the pipe and pumps out as the input to **more**, so producing a paged listing of the current directory. More than two commands can be connected by pipes, one after another.

6.4.5 FIFO

In UNIX systems, a FIFO is a named pipe, operating on the basis of first-in first-out order. A FIFO is actually a special file that is similar to a pipe, except that it is created differently. Instead of being an anonymous communication channel, a FIFO is created in a file system by invoking **mkfifo** system call. A FIFO is an empty file.

Once a FIFO is created, any process can open it for reading or writing, in the same way as for an ordinary pipe. However, it has to be open at both ends before one can proceed to perform any input- or output operations on it. Opening a FIFO for reading normally blocks it until some other process opens the same FIFO for writing, and vice versa.

≫ A FIFO offers the same communication facilities as a pipe and has a name in the file system so that it can be accessed like a file. This allows the named pipe to be used between unrelated processes.

6.4.6 Socket

A socket is somewhat similar to a pipe, but it is a "bidirectional" first-in first-out communication facility. The socket abstraction of IPCs was originally

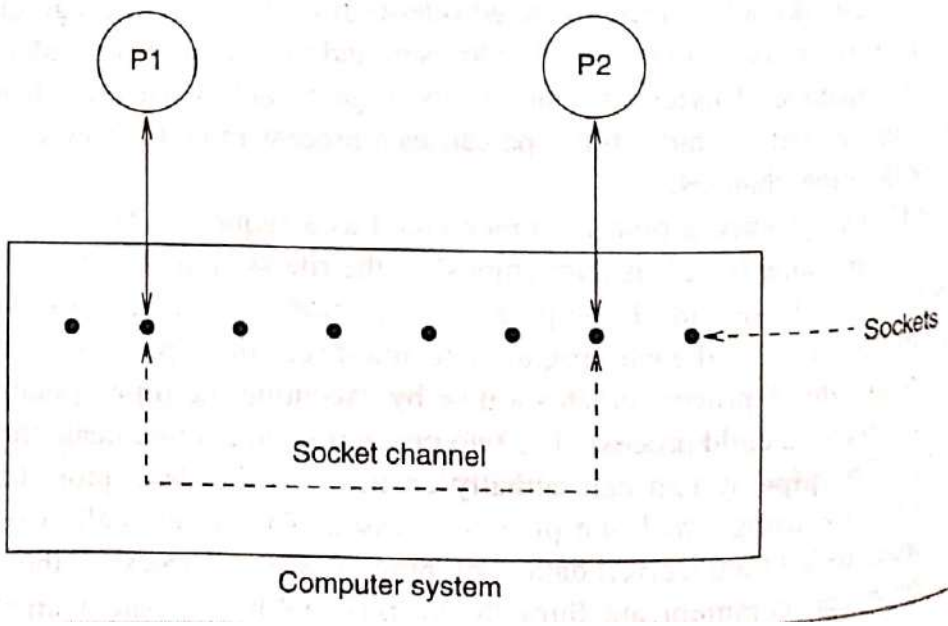

Figure 6.8: *A socket connection between two processes P1 and P2.*

introduced to enable message communications between two processes on two different computer systems, but, as shown in Fig. 6.8, sockets can also be used for communications between processes in the same computer system. A socket is an indirect IPC scheme. Sockets are parts of the communication system proper; each communication channel has two sockets, one at each end of the channel. A process connects to a socket to send and receive data from the socket. Knowing what process is at the other end of the channel may not interest the process. The processes exchange data over the channel by executing read- and write operations on the socket, the same way they access files. We will discuss more about socket in Section 16.5.2, where we talk about IPCs in distributed systems.

≫ IPC sockets are much like electrical sockets. Just as we connect any electric appliance to any electric socket to draw power from the socket, we can connect any process to any IPC socket to exchange data over the socket.

Summary

This chapter discusses the purpose and elements of IPCs. There are two models of IPCs, message passing and shared memory. The chapter presents six IPC schemes from UNIX systems: signal system, shared memory region, message queue, pipe, FIFO, and socket. All these IPC schemes are widely used by application developers.

All schemes except the shared-memory region belong to the message-passing model. They are asynchronous communication schemes. The signal system is the simplest of all. Each signal communication exchanges one bit of information. An application process can send signals to other application processes. The kernel can also send signals to application processes. In fact, in some operating systems, the kernel communicates exception events to processes by sending them signals. The kernel, when it creates an application process, creates a signal descriptor for the process to enable its communication via signals. A process can also block specific signal types, in which case the process is immune to those specific signals. Signal handling in multithread systems is a little problematic. Synchronous signals are always handled by the producing threads. However, asynchronous signals are handled differently in different systems. Some systems create a dedicated thread to handle all asynchronous signals.

For other communication schemes, processes need to set up explicit communication channels. A message queue is a kind of mailbox. All the participating processes have to subscribe to a common message queue. The size of the messages that the message queue can support is arbitrary (subject to some maximum limit). Message **send** and **receive** primitives are atomic, and message queues ensure first-in first-out semantics: messages enqueued first are dequeued first. This chapter presents a circular queue based implementation of message queue. Pipes implement unidirectional flow of data from one process to another "related" process. For each pipe, the operating system reserves a fixed amount of space in the kernel. When a process writes data into the pipe, the kernel copies the data into the reserve space from where another process can read the data out later. FIFOs are named pipes that can be used by any process. They exhibit the same behaviour that pipes do. A socket is somewhat similar to a pipe, but is a bidirectional first-in first-out communication scheme. One socket can only be associated with one process at a time. To set up a communication between two processes, each of them needs to have its individual socket and the kernel set up the communication channel between the two sockets. Each can send- and receive data to and out of the channel through its own socket.

These five types of communications are performed through the kernel space, and synchronization-related issues are taken care of by the kernel. In contrast, in the shared-memory scheme, processes can have shared memory regions that reside in the private address spaces of processes. However, the processes need to create and attach shared memory regions to their address spaces with the help of the kernel. Once a shared memory region is

attached to a process, the process can **read** and **write** data in memory locations in the region without going through the kernel. Applications themselves have to take care of synchronization.

Literature

The UNIX operating system originally introduced signals-based IPC scheme (Ritchie and Thompson 1974). Socket, pipe, and FIFO concepts were introduced in the later versions of UNIX. Hansen (1970, 1973a) discussed IPC schemes in the RC 4000 system. Schlichting and Schneider (1982) discussed message passing primitives. Stevens (1990) elaborately discussed socket programming.

Exercises

1. What is communication? What are the five elements of a communication?
2. Explain the difference between synchronous and asynchronous communications.
3. In operating systems, which agents are involved in communications?
4. Why do processes communicate?
5. There are two models of interprocess communications. What are those models? Compare them. We presented six interprocess communication schemes in this chapter. Identify which models they represent.
6. What is a signal and what are its sources?
7. Explain how signals are handled.
8. What will be consequences if the operating system allows all the signals to be blocked?
9. Where do the user-written signal handlers reside, in the user space or the kernel space?
10. Why do kernel processes not have to have a signal descriptor?
11. Give two examples of how a user can send a signal to a process.
12. Write a simple program illustrating how one process sends signals to another process.
13. What is a synchronous signal? What is an asynchronous signal? How are they different from each other?
14. Can a thread in one process send a signal to a specific thread of another process? Justify your answer.
15. Suppose you have created an application that involves two processes, P and Q, each with exactly two threads, (p1, p2) and (q1, q2), respectively. The application requires p1 and q1 to synchronize via signals. Suggest a scheme by which p1 can reliably send a signal to q1.
16. Consider a multi-threaded program. Give an example of a signal that should be delivered to all threads within the program, and a signal that should be delivered to only a single thread. Explain your answer.
17. How do we differentiate between a user signal and a kernel signal?
18. With suitable examples, explain how shared memory regions can be used for interprocess communication.

CHAPTER 7

Process Synchronization

Learning Objectives

After reading this chapter, you should be able to:

- Explain process synchronization and the need for it.
- Describe some fundamental synchronization problems.
- Discuss some useful synchronization tools.
- Describe solutions to synchronization problems.
- Explain deadlocks and their solutions.

7.1 Introduction

Processes in a computer system execute programs to manipulate data. One process may or may not interact with another process to accomplish its task, and as explained in Section 6.3.3 on page 138 such interactions take place only through shared data. If two processes use two distinct sets of data items, then they are not considered interacting processes and their activities do not affect each other. Their behaviours are independent and do not influence each other. Whereas, if a process A modifies a data item that another process B reads, then the behaviour of B may depend on activities of A. Here we say the two processes are interacting with one another through shared data. The behaviours of these interacting processes then depend on two factors: the relative speeds of the processes, and what they do with shared data. If the activities of interacting processes are not controlled suitably, their behaviour may not be as "consistent" as expected from their specifications.

≫ The operating system or any application execution, may be considered to be a collection of processes in which some interact and some others do not.

The theme of this chapter is the coordination of interacting processes, that is, the orderly execution of their accesses to shared data. Unlike interprocess communication (discussed in Chapter 6), process interaction is somewhat indirect. In many situations, a process may complete its own execution even if other processes are absent. Processes in general are independent, but they

≫ Synchronization is the single most important topic in highly concurrent systems such as operating systems, databases, and networks.

synchronize their relative speeds and accesses to shared data to ensure consistency of the shared data and/or system states. In this chapter, we will study many synchronization problems and their solutions using various synchronization tools.

7.2 Process Synchronization

≫ Concurrency is the notion of more than one related event happening at the same real time.

Modern operating systems are multiprocess systems. They run many processes concurrently. Concurrency among processes is exploited to achieve better utilization of system resources. Two processes are said to be *concurrent* if their program executions overlap in real time. That is, the first operation execution of each process starts before the last operation execution of the other process is complete. Figure 7.1 shows two typical execution scenarios of three concurrent processes. In Fig. 7.1(a), three processes are being executed in parallel (i.e., simultaneously) by three different processors, and in Fig. 7.1(b), they are executed in an interleaved fashion by a single processor. The solid line fragments indicate executions of instructions by a processor. The broken lines represent process durations. The gap between two solid line fragments in a process indicates that the process is waiting for some condition (multiprocessor case) and/or is preempted (uniprocessor case).

≫ Although both concurrent- and parallel executions imply that executions overlap in real time, traditionally parallel executions refer to executions on distinct processors that overlap in real-time while concurrent executions refer to executions that have the potential to be executed in parallel. That is, they can be executed in parallel, but their executions actually may or may not be parallel. When concurrent executions are not executed in parallel (on different processors) they are called pseudo parallel or apparently simultaneous executions.

Depending on the hardware platform configuration, execution of individual instructions from processes may overlap in real time (simultaneity) or interleave. In multiprocessor systems, instruction executions of many processes can proceed simultaneously in real time. In uniprocessor systems, the only way concurrency is achieved is through interleavings of instruction executions. In either case, an instruction execution of one process may affect those of other processes if the instruction executions reference the same piece of shared data. Whether processes execute instructions simultaneously or in an interleaved manner is immaterial to us here. As long as their relative speeds are unknown, they may lead to the occurrences of the same kinds of problems, all due to concurrency.

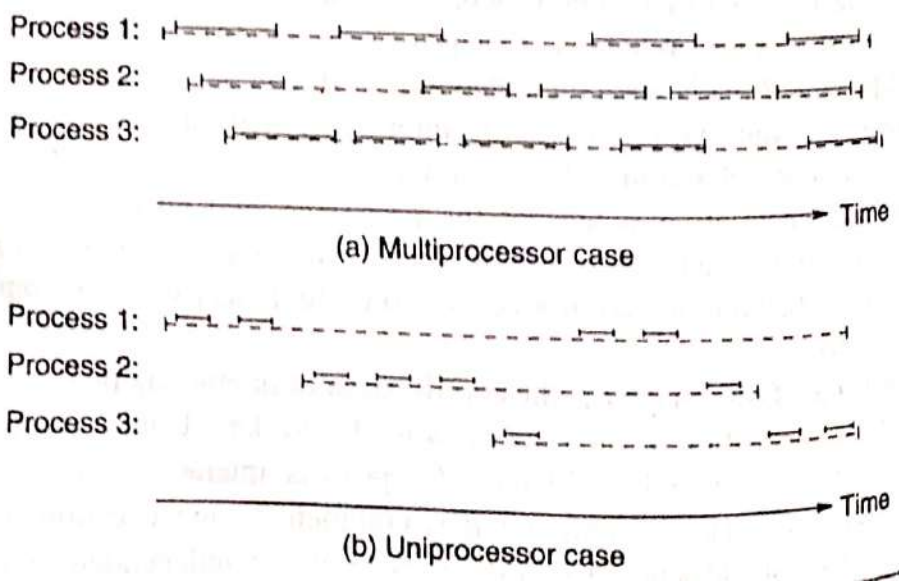

Figure 7.1: *Concurrent process executions.*

In multiprocess systems, different processes are at different states of execution. They normally execute programs and access data from their private address space. From time to time, they execute system call service routines from the kernel space. The system calls occur at unpredictable times, and they may access the same kernel data. They run at different speeds, and one process has no control over the speed of others. Consequently, the behaviour of one process may depend on what other processes do with the shared data and their relative speeds. The operating system must ensure that processes see consistent values of shared data. The data values must satisfy a priori specified integrity constraints.

If executions of concurrent interacting processes are not controlled properly, they may store inconsistent values in shared data and their behaviour may be difficult to predict. Process coordination synchronization (aka) is a fundamental problem in multiprocess operating system design and development. In general, a synchronization problem is to achieve a specified coordination among a set of concurrent processes. Processes themselves must coordinate their activities to maintain a coherent behaviour. They, in general, are independent, and may not be aware of one another. One process becomes aware of another by waiting on a condition that is set by the other process. That is, processes need to communicate (directly or indirectly) information among themselves to coordinate their own activities. (This is different from interprocess communication.) Like interprocess communications, these communications are also done through reading- and writing shared variables.

≫ An integrity constraint is a set of invariants maintained by a system in order to ensure correct system behaviour and to fulfil its interface specification contract.

Kernel space has many shared data structures. Accesses to these shared data are done by executing various kernel paths on behalf of the communicating processes. When a kernel path is modifying a shared variable, no other kernel paths should be allowed to access the shared variable. Otherwise, consistency of the shared variable and behaviours of kernel paths may not be guaranteed. These problems also arise in any shared memory-based communications where processes directly access shared variables in the shared memory. Therefore, whether a shared data is in the user space or in the kernel space, coordination of accesses to the data to ensure its consistency semantics is essential. In a process if there are multiple threads concurrently accessing common global data from the process address space, we have the same kind of synchronization problems as for threads. In multiprocessor systems, we have similar coordination problems for processor actions such as sharing the CPU queue.

In a nutshell, chaos ensues in the system when more than one agent (thread, process, kernel path, processor, etc.) accesses a piece of shared data concurrently, unless such accesses are properly controlled. Coordination or synchronization of the activities of agents is vital to maintain data consistency. In modern high performance computers, a large number of activities (of kernel paths) proceed concurrently in the kernel space. Synchronization problems are almost everywhere in operating systems. Designers and developers of operating systems must understand the intricacies and complexities of these problems before building highly concurrent operating systems.

7.3 A Typical Synchronization Problem

Let us first study a very simple two-process interaction, shown in Fig. 7.2, which illustrates a typical synchronization problem. There are two processes, namely P_1 and P_2. They share an integer variable v that is initialized to 0. It is read and written by executing *atomic* (i.e., indivisible) read- and write operations. In two separate higher-level operations, process P_1 reads and increments v by 1, and P_2 reads and decrements v by 1. The lower level operation executions interleave in the manner shown in the figure. The two reads on v are executed before the two writes on v. Both processes read 0 from v. The final value of v depends upon which write on v is executed last. Then, the final value is either -1 or $+1$, (in the figure, it is -1). If the two higher-level operations were executed sequentially, in either order, the final value of v remains 0. Consequently, for the scenario in the example, we say that the final value of v is inconsistent, that is, the value does not reflect any order in the two higher-level operation executions. The inconsistency has occurred because we allowed both processes to manipulate the shared variable v concurrently in an uncoordinated manner. (This is also the case in the message queue implementation in Fig. 6.6 on page 145 for the *count* shared variable.)

≫ Atomic means that the operation runs to completion or "as if" not run at all. Indivisible means that the operation cannot be stopped in the middle of the run and the shared state cannot be modified by another operation in the interim.

As mentioned previously, a shared data is accessed by a set of operations. When these operations are executed serially, each operation execution is guaranteed to transform the shared data from one consistent state into another. However, data consistency may not be guaranteed when many processes execute those operations concurrently on shared data. In general, when many processes read and write the same shared variable in an uncoordinated manner, the outcomes of these operation executions are unpredictable and would depend on the particular order of execution. A situation where the outcome is unpredictable, like in a race, is called a *race condition*. If the operations are carried out in a different order, the processes may behave differently. Unfortunately, not all outcomes of a race condition may need to be correct. Presence of race conditions in a system is regarded as bad design. Race conditions should be avoided at all costs to ensure predictable system behaviour. Therefore, control of accesses to shared variables is necessary to eliminate race conditions occurring in the system. Orderly execution of operations on shared variables is called a coordination- or synchronization of operation executions.

≫ Anomalies due to race conditions are notorious in concurrent systems, and errors due to them are hard to debug and diagnose. As noted in Section 1.2.2 on page 6, when a race condition is present, a program execution may not complete and even if it does complete, it may not produce correct output. They are race hazards!

Shared data structure and initial value
int $v = 0$;

Process P_1	Process P_2
1. int $v1$ = read(v); /* $v1 == 0$ */	
	2. int $v2$ = read(v); /* $v2 == 0$ */
3. $v1 = v1 + 1$; /* $v1 == +1$ */	
	4. $v2 = v2 - 1$; /* $v2 == -1$ */
5. write(v, $v1$); /* $v == +1$ */	
	6. write(v, $v2$); /* $v == -1$ */

Figure 7.2: *A typical uncoordinated operation executions on a shared variable.*

If a shared variable is of a primitive data type, the machine architecture may provide the basic synchronization for accessing the variable, and it may be treated as an atomic variable. That is, operation executions on the variable happen in some total order respecting their original order. Unfortunately, most shared variables at a higher level are non-atomic data structures. Even for a single atomic variable, a process may execute a sequence of operations on the variable in a row as shown in Fig. 7.2 on page 154, and expect the entire operation sequence to be indivisible. The term *synchronization* is used to specify various constraints on the orderings of operation executions of different processes on shared variables. Until a process modifying a shared data has finished applying all its operations, no other process is allowed to read or write the same data.

Synchronization constraints mostly specify that some operations cannot be executed concurrently. For example, two store operation executions on the same memory address must exclude each other; otherwise some bits in the addressed memory cell may have values from one store operation and other bits from the other store operation. We will study some interesting synchronization constraints in this chapter. The term synchronization is also used to control activities of cooperating processes, for example, their relative speeds.

Shared variables are used to model system resources. Each resource is modelled by a set of related shared variables whose values collectively represent states of the resource; modifying these values reflects changes in the state of the resource. Processes operating on the resource make references to the shared variables. Control of operation executions on shared variables are necessary to keep the resource in consistent states.

Researchers in this field strive to support program designers and developers with tools they can use with relative ease for the synchronization required. As only the memory read and write operations are atomic, constructing a highly concurrent operating system without synchronization tools is a challenging task.

Many synchronization problems are reported in literature that may need different solutions. We will study some important synchronization problems and their solutions in this chapter. We start with defining some well-known synchronization problems first.

7.4 Classical Synchronization Problems

Concurrent processing by several asynchronous processes differs from sequential processing. In concurrent processing, the order of execution of the elementary steps (i.e., operations executions) of different processes is not predetermined. The order may depend on various parameters such as the relative speeds of processes. Interrupts from peripheral devices make things even more unpredictable. Concurrent processes interact with each other accessing shared variables. Synchronizing these accesses avoids race conditions and unwanted time-dependent erroneous behaviour. In the processes where they manipulate shared variables, certain sensitive sections of programs are designated as "critical sections".

A *critical section* is a segment of code where a process accesses a set of related shared variables. For example, a database application updates parts of a database file in a critical section. Different processes have their own critical sections where they manipulate the same set of shared data. For example, in Fig. 7.2 on page 154, the three operations of each process form a critical section. These sections are *critical* in the sense that the processes must not execute the sections without coordination among themselves. Shared data are always accessed inside (and, not outside) critical sections. These critical sections are called *similar* critical sections as they all manipulate the same set of shared variables. For the sake of this chapter, we assume that there is only one critical section shared by all processes, and which we will refer to as *the* critical section. The values of the shared variables determine the state of the critical section. When some process executes (or is in) the critical section, we say the critical section is *busy* or *occupied*; otherwise, the critical section is *free* or *empty*.

The critical section when executed in isolation is assumed to transform itself from one consistent state into another. However when it is executed concurrently by many processes, such a guarantee may not be possible. A specific *synchronization* or *critical section problem* is to order executions of the critical section satisfying some specific constraints on the shared data. So that these constraints are not violated, the processes follow some protocols in their execution. Solutions to synchronization problems entail designing these protocols for processes to be followed before entering and after exiting the critical section. The steps in the protocol are executed in the parts called *entry section* and *exit section*, respectively. The entry section is executed in preparation to enter the critical section, and in the exit section for leaving it.

A process wanting to execute the critical section must first make known its intention to the other processes. In other words, it must acquire permission to enter the critical section. If its immediate entry is not allowed, it must wait until it may enter. A process evinces its interest in the critical section when it starts executing the entry section. Again, at the end of the execution, it must inform other processes by executing the exit section that it has indeed finished.

≫ We may view the critical section as a shared resource. The entry section implements the arbitration logic to "acquire" the critical section and the exit section implements the "release" of the critical section. The entry section implements some non-trivial arbitration logic, but the exit section is normally trivial.

Note that executions of the entry- and exit sections involve communications among the processes, and that the two sections employ a different set of shared variables called *synchronization, coordination*, or *control variables*. These control variables differ from the shared variables proper used within the critical section. (Control variables are solely used to construct the entry- and exit sections.) The processes may, however, concurrently access the control variables. In other words, we allow race conditions to occur for the control variables. These control variables are normally simple shared variables, and the outcomes of the race conditions are somewhat predictable and manageable.

For the sake of this chapter, we logically partition a process address space into four distinct sections, as shown in Fig. 7.3. A process cycles through its remainder, entry, critical, and exit sections until its execution is complete. In the remainder section, a process does something else and does not access the critical section-related shared variables or the synchronization variables. A process can terminate its execution only in the remainder section,

```
while (process is not complete)
{
    <remainder section>;
    <entry section>;
    <critical section>;
    <exit section>;
}
```

Figure 7.3: *Partitions of process address space.*

and not in the other three sections. When a process executes a particular section, we say that the process is "in" that section. A solution to a critical section problem involves designing the codes for entry and exit sections.

We assume that processes are asynchronous, and they execute at a non-zero speed. A process is not stopped when it is supposed to be taking basic steps. That is, we assume that all processes interested in the critical section continue taking steps. However, they proceed at different relative speeds that are not a priori known. Further, in the remainder section, a process may halt its execution or pause for an arbitrary length of time. It is not known which processes are going to request the critical section next or when they will do so. Moreover, we assume that all processes take a finite amount of time to execute the critical section. A process may not stop in the critical, entry, or exit section; otherwise the system as a whole may not be able to make any progress.

A few well-known classical synchronization- or critical section problems are described in the following subsections. They are representatives (or abstractions) of a large number of practical synchronization problems. They differ from one another in the way critical section executions of different processes are ordered. Solutions to these problems are presented in Section 7.5.

7.4.1 The Mutual Exclusion Problem

Mutual exclusion is the most fundamental problem of synchronization—and the most studied—in which the synchronization constraints specify mutually exclusive executions of the critical section. That is, no two processes can execute the critical section concurrently. A process that begins execution of the critical section must finish the execution before another process starts its execution of the critical section.

≫ Mutual exclusion models the problem of managing accesses to an indivisible, non-sharable resource that can only be allocated to one process at a time. For example, a graphics plotter is a physical resource that cannot be used simultaneously by two processes. They must use the plotter mutually exclusively, one after another.

7.4.2 The Producer–Consumer Problem

Producer–consumer is another fundamental synchronization problem. There are two sets of processes—one called *producers* and the other called *consumers*. Producers produce data items that are consumed by consumers. Processes are asynchronous, and they proceed at different non-zero speeds that are not known a priori. Consequently, when a producer produces an item, sometimes no consumer may be ready to receive the item. The producer then has two alternatives to follow:

- **Synchronous communication:** The producer waits until a consumer is ready to receive the produced item. When both are ready, the data exchange takes place.
- **Asynchronous communication:** The producer deposits the item in a known storage space from where a consumer retrieves the item later (see Fig. 7.4).

We consider only asynchronous communications in this book. The storage space used to hold data items is called a *buffer*. The buffer provides a finite storage space to hold unconsumed items. Two primitive operations, namely **put** and **get**, are implemented to access the buffer contents. A producer executes the **put** operation to store an item in the buffer, and a consumer executes the **get** operation to retrieve an item from the buffer. The operation executions by the processes are subject to the following constraints:

- Both the **put** and **get** operations are (or appear to be) atomic.
- Items are received in the order they are put in the buffer.
- When the buffer is full, a **put** execution is blocked until consumers receive some items.
- When the buffer is empty, a **get** execution is blocked until producers put some items.

7.4.3 The Readers–Writers Problem

Readers–writers is another fundamental synchronization problem. It is a special case of the producer–consumer problem described in the previous subsection. The senders are called the *writers*, and the consumers the *readers*. They access a single shared variable. A writer overwrites the previous value of the variable with a new value. A reader returns the current value from the variable without altering the content of the variable. Each writer has an exclusive access right to the variable, but several readers may access the variable simultaneously. That is, a write operation execution is exclusive to other operation executions, but many read operation executions may be concurrent. Unlike the mutual exclusion problem, we allow many readers to concurrently execute the critical section, as they do not change values of the shared variable.

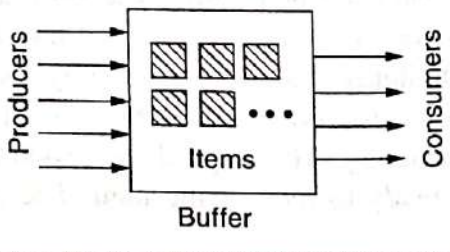

Figure 7.4: *Producer-consumer interactions.*

7.4.4 The Dining-philosophers Problem

The dining-philosophers problem is one of the best-known synchronization problems. Some philosophers are perpetually sitting around a circular table (see Fig. 7.5). They spend their time thinking and eating. A philosopher either thinks or eats, but does not do both at the same time. A philosopher usually thinks, and when she is thinking she does not need any physical resources (chopsticks, in this case). From time to time, each philosopher becomes hungry, and she needs two chopsticks to eat. Eating is a finite time action. As shown in the figure, there is a single chopstick between every adjacent pair of philosophers. A philosopher can pick up the two chopsticks on the table adjacent to her, and not the others. She can however pick up one chopstick at a time. She can eat only when she acquires two chopsticks. If a fellow philosopher holds any of the adjacent chopsticks, she needs to wait until it is available. When she has finished eating her meal, she releases both the chopsticks one by one, and then resumes thinking again. The restriction is that no two neighbouring philosophers can eat at the same time.

7.4.5 The Sleeping-barber Problem

A barbershop has two rooms: a barber's room and a waiting room. The barber's room has a single chair. When there are no customers to serve, the barber goes to sleep on the chair. There are a fixed number of chairs in the waiting room. When the barber is busy serving a customer, new customers sit

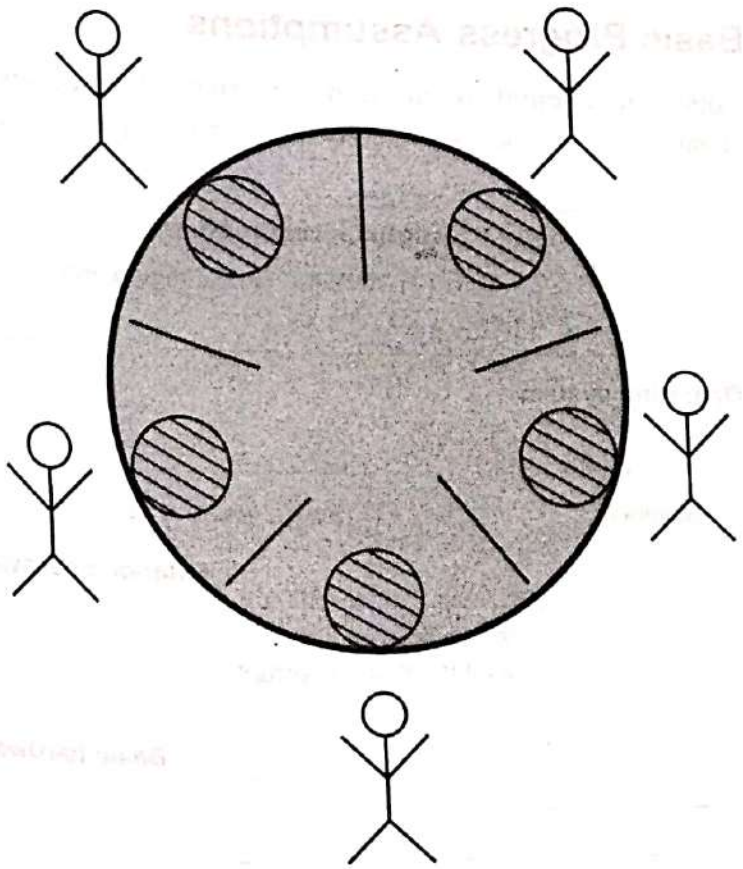

Figure 7.5: *The dining philosophers' round table conference.*

on the chairs in the waiting room, one on each chair, waiting for their turn for the barber. If there are no free chairs, new customers leave the barbershop. If the barber is asleep, the new customer wakes her up. When there is no customer the barber should sleep and when there are customers, the barber should serve them one at a time.

7.5 Synchronization Solutions

In the previous section, we specified several synchronization problems, each requires a different flavour of coordination for executions of critical sections. Different synchronization problems need different types of solutions. Different solutions employ different synchronization primitives. Synchronization primitives are supported at various levels such as basic hardware, extended hardware, operating system, and higher-level programming languages as shown in Fig. 7.6. As operating system experts, we should look at primitives at all levels.

Given a set of synchronization primitives, a solution to a synchronization problem involves designing entry and exit protocols using those primitives, which ensures the desirable properties of the problem. In the following subsections, we study some synchronization solutions based on the read-write, test-and-set, semaphore, spinlock, region, and condition synchronization variables. Before doing so we state a fundamental assumption in Section 7.5.1, and describe various desirable properties of synchronization solutions in Section 7.5.2.

7.5.1 Basic Progress Assumptions

We make one fundamental assumption here that all processes make finite progress. That is, if a process is ready to perform a basic step (aka, one

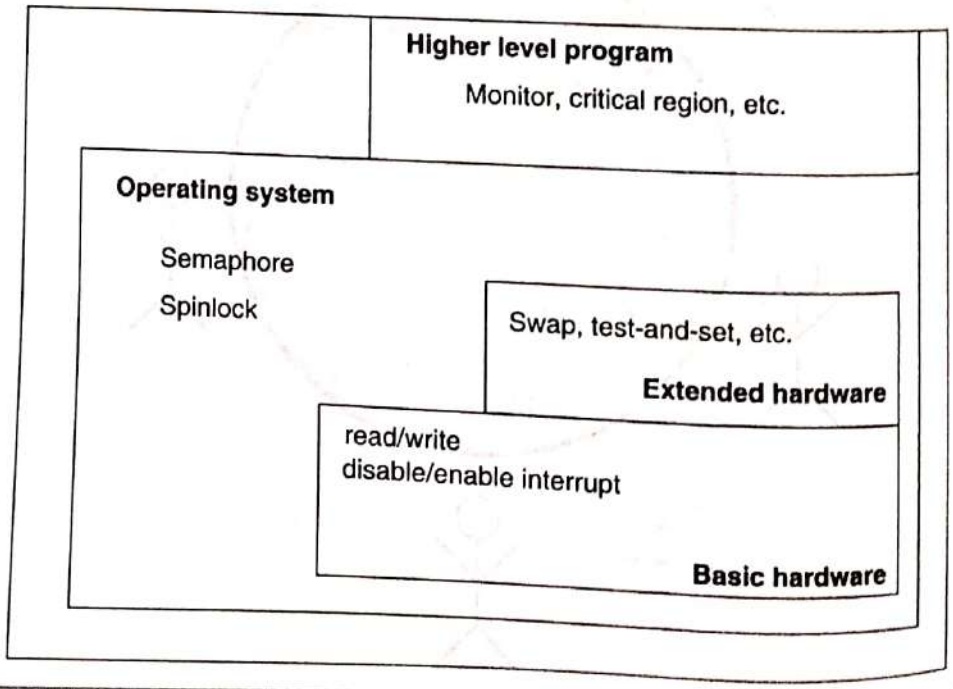

Figure 7.6: *Hierarchy of synchronization primitives.*

machine instruction execution), it does so in a finite time. If a process is infinitely delayed in performing a basic step, then no progress may be ensured for the process. An execution of a set of processes is said to be *fair* if the following two conditions hold for each process P in that set.

- If the execution is finite, then P does not take any basic step at the end of the execution.
- If the execution is infinite, then P performs infinitely many basic steps or P is not interested in taking basic steps infinitely many times.

Without the basic progress assumption, some synchronization properties may not be possible to achieve. We will explicitly point out the use of this basic progress assumption in presenting synchronization solutions.

7.5.2 Solution Characteristics

First, any synchronization solution should correctly satisfy the intent of the problem it is solving. That is, nothing untoward should happen during its execution. This is called a *safety* or *correctness property*. Secondly, the solution should assure some other useful qualities such as making progress, maintaining fairness, etc. We list these properties here.

1. **Safety:** The intended condition will never be violated. For example, a solution to the mutual exclusion problem should never violate the mutual exclusion property. For another example, a solution to the readers-writers problem should never allow a reader to enter the critical section while a writer is engaged in it.
2. **Liveness:** The system should not be blocked forever due to contention for the critical section. That is, contention due to synchronization must be resolved in a finite time so that the system can progress eventually. For example, a solution to the mutual exclusion problem should assure that some process competing for the critical section enters it in a finite time. This property is also called the *finite arbitration-* or *non-blocking* or *progress* property.
3. **Starvation Freedom:** No process should be denied progress forever due to contention. That is, not only some processes, but all processes must eventually make progress. For example, a solution to the mutual exclusion problem should assure that every interested process enters the critical section in a finite time.
4. **k-bounded Bypass:** No process should be allowed to enter the critical section more than k times when another process is already waiting to enter the critical section. This is also called the *k-fairness property*.[1] This property is stronger than starvation freedom property.

≫ *Starvation* is a situation in which one or more processes wait indefinitely in the entry section, and other processes overtake the waiting processes in entering the critical section.

[1] We need to be a bit careful here. We are talking now of some higher-level fairness in critical section executions. Earlier, in Section 7.5.1, we talked about lower-level fairness in basic execution steps. In general, lower-level fairness may not ensure higher-level fairness in critical section executions.

5. **FIFO Fairness:** No process will ever overtake another waiting process. This is the 0-fairness property.
6. **LRU Fairness:** The process that received the service least recently gets the service next.

Among these properties, safety, liveness, and starvation freedom are considered essential properties of any synchronization solution and the remaining are considered desirable properties. Further, an efficient solution needs to assure that a process does not consume CPU time unnecessarily waiting at the entry section.

7.5.3 Solutions to the Mutual Exclusion Problem Using Atomic Variables

We view each memory cell as an ordinary read/write variable. In a single step, a process can either read or write a single cell, but cannot do both; neither can it do so on many cells. Reading and writing each memory cell is viewed as a special case of the readers-writers synchronization problem. The memory hardware arbitration circuit solves this special problem. The arbitration circuit minimally ensures that a read and a write on the same cell exclude each other in time, and also multiple *writes* on the same cell exclude one another in time. When simultaneous conflicting requests are made to the same memory cell, the hardware allows only one of them to proceed and the remaining ones have to wait until the one granted is complete. That is, when two or more conflicting memory operations are executed concurrently, the outcome of the execution is equivalent to their sequential executions in some arbitrary order. One operation execution proceeds while others wait, and these other calls are said to have stalled. That is, this hardware-based solution has a kind of busy waiting property, because the caller remains stalled or blocked until its memory operation (read or write) is complete. The memory hardware though may allow concurrent executions of operations on different cells.

Without loss of generality, we assume, in the rest of this subsection that read and write operation executions on memory cells (primitive variables) are atomic and terminating. We would like to solve various synchronization problems using these two primitives. The race condition, displayed in Fig. 7.2 on page 154, could be avoided if processor architecture supports more powerful atomic instructions than plain read and write instructions. That is, with powerful atomic instructions such as read-increment-write and read-decrement-write, we would be able to solve the synchronization problem stated in Fig. 7.2 trivially: one atomic operation on the shared integer variable replaces the three operations.

All processors support atomic read and write instructions, and some, in addition, support atomic read-modify-write, test-and-set, swap, compare-and-swap, fetch-and-add, and other instructions. The latter operations are made up of many simpler actions, but their atomicity is ensured by the underlying

hardware. For example, the read-modify-write operation on a shared variable x is made up of three actions, in this sequence: (1) read x, (2) evaluate some function using the value of x, and (3) write a new value in x. These three actions must be performed in that sequence in one indivisible execution of the read-modify-write operation, and the underlying hardware ensures this.

A shared variable accessed only by executing atomic operations is called an *atomic variable*. If read and write are the only permitted operations on a variable, we call it an *atomic read-write variable*. If in addition test-and-set operation is permitted, we call it an *atomic test-and-set variable*.

Operating systems normally contain many non-atomic data structures that cannot be manipulated indivisibly by applying atomic operations. Compound operations are needed to manipulate these data structures. It is often not possible to implement atomic compound operations using atomic read-write variables. Consequently, data structures are manipulated in critical sections. We present below a number of solutions to the mutual exclusion problem. Each solution employs atomic variables to implement entry and exit section protocols. Subsections A Naive Two-process Solution to The Lamport Solution present solutions using atomic read-write control variables, and subsections A Solution Using test-and-set Operation to A Solution Using swap Operation using stronger control variables.

If the code of a synchronization solution for the interacting processes is identical except the reference to the process id, which is the typical case for most synchronization algorithms, it is customary to write the code for a generic process with id i. In those subsections, we present solutions for a typical process P_i, $0 \leq i < n$, where n, greater than 1, is the number of concurrent processes.

≫ Intel 80386 processor has atomic inc/dec instructions. They work fine on single CPUs, but are ineffective for multiple CPUs. The processor also supports lock-prefixed instructions, for example, Lock dec or Lock inc. A lock-prefixed instruction blocks all other CPUs from accessing the system bus until the instruction execution is complete.

A Naive Two-process Solution

A very simple solution to the mutual exclusion problem for two processes, say P_0 and P_1, is presented in Fig. 7.7. The solution uses a single atomic boolean control variable *turn*, whose value is initialized to 0.[2] (The variable could also be initialized to 1.) Process P_i uses two local constants *me* and *other*, and *me* is set to i and *other* to $1 - i$. The idea is that a process enters the critical section only when it is its turn. The entry section contains a single while-loop statement, where process P_i repeatedly reads *turn* and tests whether the value indicates its own turn to enter the critical section. If the *turn* value is i, it enters. The exit section contains a single assignment statement where *turn* is set to *other*, indicating that the other process can enter the critical section.

What properties does the solution possess? It ensures the mutual exclusion (aka, the safety) property and 1-bounded bypass property; however, it may not ensure the liveness property. The solution forces processes to take turns alternately to the critical section. If one process completes its execution

[2]The volatile specifier is an indication to compiler not to optimize or reorder accesses to the variable.

```
Data structures and initial values
    shared volatile boolean turn = 0; /* shared by P0 and P1 */
    const local boolean me = i;      /* for process Pi */
    const local boolean other = 1 – i;

Solution
    while turn  me ≠ do; // entry section
    ⟨ Critical section; ⟩
    turn = other;        // exit section
```

Figure 7.7: *A naive solution to the 2-process mutual exclusion problem.*

or becomes uninterested in the critical section, the other process will be blocked forever in the entry section. For example, assume *turn* is 0, P_0 is not interested in the critical section, but P_1 is. P_1, however, cannot enter the critical section until the *turn* becomes 1. The turn value can become 1 only if P_0 becomes interested in entering the critical section (that is not known), enters the critical section, and changes the value of *turn* at the exit section. So we cannot assure that whether P_1 will eventually enter the critical section or not.

The main problem with this solution is that one process does not have sufficient information about the other process for meaningful arbitration. The sufficient information in this case is the status of the other process. Since the boolean variable *turn* can store only information about who can enter the critical section next, if both become interested at the same time, we need more control variables to store the status of the processes.

The Dekker Solution

The previous solution to the two-process mutual exclusion is correct but not satisfactory. The first correct and satisfactory solution to the 2-process mutual exclusion problem is due to Dekker, a Dutch mathematician. The Dekker solution is presented in Fig. 7.8. It is an extension of the naive solution of Section A Naive Two-process Solution. As in the naive solution, we denote two processes by P_0 and P_1. The Dekker solution uses three shared boolean variables: *status[0]*, *status[1]*, and *turn*. The *status* variables are initialized to *false*, and the *turn* to 0. Each *status* variable is written by one process and read by the other process, and they can do so at the same time. However, the *turn* variable is read and written by both the processes. The protocol structure ensures that *turn* is never written by both processes at the same time.

When a process P_i becomes interested in the critical section, it sets *status[i]* to *true* first to indicate its interest in the critical section. P_i then executes an arbitration algorithm to resolve any potential conflict of interests. If the other process is not interested in the critical section, that is, the *status[other]* value is *false*, then P_i enters the critical section directly. Otherwise, there is a conflict, and the tie is broken by allowing the process that did not enter the

```
Data structures and initial values
    shared volatile boolean status[2] = false; /* shared by P0 and P1
    shared volatile boolean turn = 0;          /* shared by P0 and P1 */
    const local boolean me = i;                /* for process Pi */
    const local boolean other = 1 – i;

Solution
    status[me] = true; /* show interest in the critical section */
    while (status[other]) do /* busy wait arbitration */
    {
        if (turn == other) { /* turn of the other process */
            status[me] = false;/* temporary withdrawal */
            while (turn == other) do; /* busy wait until my turn */
            status[me] = true;
        }
        /* else it is my turn, wait until status [other] becomes false */
    }
    ⟨ Critical section; ⟩
    turn = other;
    status[me] = false; /* no more interested in the critical section */
```

Figure 7.8: *The Dekker solution to the 2-process mutual exclusion problem.*

critical section for the longest time (the LRU fairness). This is indicated by the value of *turn*, and whoever has the *turn* wins the tiebreak and enters the critical section.

Consider the case when *turn* value is equal to *0* and both processes have set their *status* value as *true*. Now there are three possible scenarios: (1) both are in the entry section; (2) P_0 has already crossed (i.e., finished) the entry section before P_1 set its *status* as *true*; (3) P_1 has already crossed the entry section before P_0 set its *status* as *true*. There is no way that a process can know in which scenario it is. Although P_0 has the current turn, it cannot simply cross the entry section because safety is violated if it is in scenario (3). Therefore, P_0 must confirm that P_1 has not already crossed the entry section. This information can come only from P_1. For that, P_1 temporarily withdraws from the competition by setting its *status* to *false* so that the process P_0 that has the current turn can go ahead and enter the critical section. After the temporary withdrawal from the competition, P_1 simply waits for its turn by repeatedly checking for it. When P_0 has finished executing the critical section, it sets the value of *turn* to *other* (i.e., to 1), and resets *status[0]* to *false*. When the *turn* value is equal to 1, P_1 restarts its competition by setting its *status* to *true* and continues.

This solution ensures the properties of mutual exclusion, liveness, and starvation freedom. The solution assures the liveness and starvation-freedom properties under the assumption of finite (non-zero) progress in taking basic steps by both processes. However, it does not have the *k*-bounded bypass property; during a temporary withdrawal from the competition, theoretically, there is no bound on how many times the other process can cross the entry code and enter the critical section.

The Peterson Solution

Figure 7.9 presents a new solution to the 2-process mutual exclusion problem. It is substantially simple compared to the Dekker solution, and uses the same number and types of shared variables. It uses three shared boolean variables: *status[0]*, *status[1]*, and *turn*. The *status* variables are initialized to *false*, and the *turn* to 0. Each *status* variable is written by one process and read by the other process, and they can do so at the same time. However, the *turn* variable is read and written by both processes. Unlike the Dekker solution, here both processes can write the *turn* variable at the same time. The outcome of simultaneous accesses is determined by the atomicity property of *turn*. The idea is simple: when both processes are competing, the later process (which writes on *turn* latest) "pushes" the former process (which writes on *turn* first) out of the entry code (the while-loop) by changing the *turn* value. The later process could exit the waiting while-loop as soon as the former process completes and sets its *status* to *false*.

When a process P_i becomes interested in the critical section, it sets *status[i]* to *true* first to indicate the other process its interest in the critical section. P_i then gives the other process (referred to as P_j, $j = 1 - i$) a preferential treatment to enter the critical section if the other process is also interested in executing the critical section. (If both processes attempt to enter the critical section at the same time, it is possible that they both write *turn* simultaneously. The atomicity property of *turn* will determine its final value.) P_i then executes an arbitration algorithm to resolve any potential conflicts. The arbitration logic here is a very simple while-loop statement. P_i repeatedly reads *status[j]* and *turn* values (in arbitrary order), and checks their values. P_i waits on the while-loop until either P_j is not currently involved in the competition to enter the critical section or *turn* = *i*. When P_i has finished executing the critical section, it resets *status[i]* to *false*.

This solution ensures the properties of mutual exclusion, liveness, starvation freedom, and bounded bypass (1-fairness). The solution assures the liveness property under the assumption of finite progress in taking basic steps by the two processes.

```
Data structures and initial values
    shared volatile boolean status[2] = false; /* shared by P0 and P1 */
    shared volatile boolean turn = 0;          /* shared by P0 and P1 */
    const local boolean me = i;                /* for process Pi; */
    const local boolean other = 1 − i;

Solution
    status[me] = true; /* show interest in the critical section */
    turn = other;
    while (status[other] and turn == other) do; /* busy wait arbitration */
    ⟨ Critical section; ⟩
    status[me] = false; /* no more interested in the critical section */
```

Figure 7.9: *The Peterson solution to the 2-process mutual exclusion problem.*

The Lamport Solution

Dekker was the first to solve the mutual exclusion problem satisfactorily. However, his solution is for only two processes. Dijkstra was the first to propose a solution to the general n-process, $n > 1$, mutual exclusion problem. The Dijkstra solution is inefficient, and it does not satisfy the stronger properties such as starvation freedom. Here, we present a simpler solution to the general n-process mutual exclusion problem. The solution, due to Lamport, is known as the *Bakery algorithm*. His solution is presented in Fig. 7.10. It uses n boolean variables *choosing*, all initialized to *false*, and n integer variables *number*, all initialized to 0. Each variable is written by one process and read by others. The values of integer variables are used to form token numbers, the kinds that are used in bakery shops (in USA): A customer on entering the bakery chooses a token number, and within the bakery, customers are served in the order of their token numbers.

When a process becomes interested in the critical section, it chooses a token number greater than those held by other interested processes. The first three statements in the entry section comprise a "doorway" where the process chooses a token number. This is done by reading all *number* variables in arbitrary order. To make other processes aware that it is choosing a token number, a process P_i sets *choosing[i]* to true before reading *number* variables and sets *choosing[i]* to false after reading them and writing its own token number that is one greater than the max of others. After choosing its token number, process P_i enters the arbitration section by a way of executing the for-loop statement where it evaluates the status of other processes. If any other process P_j is choosing its token number, then P_i busy-waits until P_j has obtained its token number. The arbitration logic allows the process with the lowest token number to enter the critical section first. The original process then waits until all processes with lower token numbers are served by the arbitration logic. Because of the race condition at the doorway, it is possible that several processes pick up the same

```
Constants
   n = number of processes, where n > 1;
Data structures and initial values
   shared volatile boolean choosing[n]; /* initialized to false */
   shared volatile int number[n];       /* initialized to 0 */

Solution
   choosing[i] = true; /* show interest in the critical section */
   number[i] = 1 + max{number[0],···, number[n – 1]}; /* get a token number */
   choosing[i] = false;
   for (int j = 0; j < n; j = j + 1) { /* busy wait arbitration */
     while (choosing[j])do; /* wait until Pj gets its token */
     while (number[j]≠0 and ⟨number[j],j⟩ ≺ ⟨number[i],i⟩)do;/ token comparisor.
   }
   ⟨ Critical section; ⟩
   number[i] = 0;
```

Figure 7.10: *Lamport's Bakery algorithm for process P_i.*

token number. The tie is broken by process indexes; the ones with lower indexes are served first. In the solution, we use the relation $(a, b) \prec (c, d)$ to mean $(a < c)$ or $(a = c)$ and $(b < d)$. When a process has finished the critical section execution, it resets its *number* variable to 0.

This solution ensures the properties of mutual exclusion, liveness, and bounded bypass. Actually, the solution ensures the FIFO fairness property after a wait-free[3] doorway: if process P_i completes the doorway before process P_j starts the doorway, then P_j cannot enter the critical section before P_i does. The only shortcoming with this solution, at least theoretically, is that the *number* values may grow without any bound.

A Solution Using test-and-set Operation

In the previous four subsections we studied solutions to the mutual exclusion problem for processor architectures that support only atomic read and write instructions on primitive variables. Some processor architectures implement special machine instructions that allow the CPU to test and modify the content of a single primitive variable, or to swap the contents of two primitive variables, atomically. These operations are more powerful than ordinary read and write operations, and make programming tasks a little easier. In this subsection, we study a solution that uses test-and-set atomic instruction, and in the following subsection, a solution that uses swap instruction. The advantages of these machine instructions are their simplicity and they work for any number of CPUs. They help us developing simple, easy-to-understandable solutions.

There are many variations of test-and-set operation semantics. Figure 7.11 presents one variation in algorithmic form. The operation takes address of a boolean variable, and a boolean value. It writes the new value to the address and returns the old value stored at the address. The reading of the old value from a variable and setting of the new value to the same variable are done in a single *indivisible* operation, and not in two different atomic operations as shown in the figure. The underlying processor and memory hardwares ensure the indivisibility. If two or more CPUs concurrently execute the test-and-set operation on the same variable, the operation executions appear as if they were executed in a total (arbitrary) order.

A solution to the mutual exclusion problem is presented in Fig. 7.12. The solution uses a single shared boolean variable *lock* that is initialized to *false*. (The value *false* indicates that the lock is free, and *true* indicates that the lock is taken.) The entry section contains a single while-loop statement where a process sets the value of *lock* to *true*. If the original value of *lock* is *false*, the process breaks the while-loop and enters the critical section. Otherwise, it keeps trying to obtain the lock. In the exit section, it resets the value of *lock* back to *false*.

This solution is simple compared to those presented in previous subsections, and it scales for arbitrary number of processes, that is, the structure

[3]Wait-freedom means that a process that has started a computation is guaranteed to complete the computation if the process continues to take its own steps, regardless of what other processes do.

```
boolean TestAndSet(boolean* lock, boolean newValue)
{
  boolean oldValue;
  oldValue = *lock;
  *lock = newValue;
  return oldValue;
}
```

Figure 7.11: *Semantics of test and set instruction.*

of the solution is independent of the number of processes. The solution ensures the properties of mutual exclusion and liveness, but does not ensure the starvation freedom property, and hence, does not ensure the bounded bypass property.

≫ The more power primitive operations have, the easier it becomes to construct a synchronization solution.

A Solution Using Swap Operation

The semantics of swap operation is presented in Figure 7.13 in algorithmic form. The operation takes two memory addresses, and exchanges their contents in a single indivisible operation execution. The underlying processor and memory hardwares ensure the atomicity. Normally, one of the addresses is shared, and the other one is local.

A solution to the mutual exclusion problem is presented in Fig. 7.14. (It is similar to the one in Fig. 7.12.) The solution uses a single shared boolean variable *lock* that is initialized to *false*. Each process uses a private boolean variable *key* that is initialized to *true*. The entry section contains a single while-loop statement where a process exchanges the values of *lock* and *key*. If the original value of *lock* is *false*, the process breaks the while-loop and enters the critical section. Otherwise, it repeatedly tries to take the lock. In the exit section, the process sets the value of *lock* to *false*.

This solution is as simple as the one in the previous subsection. It ensures the properties of mutual exclusion and liveness, but does not ensure the starvation freedom property, and hence, does not ensure the bounded bypass property.

7.5.4 Interrupt Disabling

Interrupt disabling is a means to coordinate critical section executions in the kernel space and not in the user space, because application processes do not have

```
Data structure and initial value
    shared volatile boolean lock = false; /* shared by all processes */

Solution
    while(TestAndSet(&lock,true)) do;
    ⟨ Critical section; ⟩
    TestAndSet(&lock, false);
```

Figure 7.12: *Mutual exclusion using a TestAndSet variable.*

```
void Swap(boolean* locA, boolean* locB)
{
   boolean temp;
   temp = *locA;
   *locA = *locB;
   *locB = temp;
}
```

Figure 7.13: *Semantics of swap instruction.*

control over the interrupt system. The kernel controls executions of the critical section by manipulating the processor interrupt facilities. For this scheme, we need instructions to enable and disable the interrupt system. The entry section is a single interrupt disable instruction, and the exit section a single interrupt enable instruction (see Fig. 7.15). That is, the CPU disables the processor interrupt circuit before entering the critical section, and re-enables the circuit after exiting the critical section. As the interrupt circuit is disabled when a process executes the critical section, the CPU cannot be interrupted and preempted from the running process and allocated to another process. Consequently, the critical section is always executed mutually exclusively.

One potential weakness of this scheme is that if the critical section is long, then the interrupt circuit is disabled for a lengthy duration soon. Consequently, many peripheral devices may sit idle waiting to get service from the CPU, and thereby, performance of the system as a whole may degrade. Though this technique of handling critical section executions can be used in uniprocessor systems, it is ineffective in multiprocessor systems, because disabling interrupts by one CPU may not disable interrupts in all CPUs. Disabling and re-enabling interrupts on all CPUs could be time-consuming tasks if not impossible. Many practical operating systems use interrupt disabling on local CPUs in conjunction with other synchronization mechanisms for multiprocessor systems.

7.5.5 Non-preemptive Kernels

In a system with non-preemptive kernel, when a process executes the kernel, the system cannot arbitrarily stop the process and start executing another process. Unless the running process voluntarily releases the CPU, the process will have

```
Data structure and initial value
    shared volatile boolean lock = false; /* shared by all processes */

Solution
    local boolean key = true;
    while (key) Swap(&lock, &key);

    ⟨ Critical section; ⟩

    Swap(&lock, &key);
```

Figure 7.14: *Mutual exclusion using a swap variable.*

```
disable interrupt;  /*entry section */
⟨ critical section; ⟩
enable interrupt;  /* exit section */
```

Figure 7.15: *A solution to the mutual exclusion problem using disable/enable interrupt.*

the CPU until it leaves the kernel. Consequently, when a process executes a system call, it is assured that the CPU will not be taken away from it; but devices can interrupt the process. The data structures that are not modified by interrupt or exception handlers are guaranteed to be free from race conditions. Consequently, critical sections that lie outside the reach of interrupt and exception handlers are always executed mutually exclusively. If a process in the kernel voluntarily releases the CPU, it has to ensure that the data structures are in a consistent state. Critical sections that are executed by the interrupt and exception handlers can be controlled either by using synchronization primitives discussed in other subsections or by preventing the nested interrupt from occuring.

Though this technique is somewhat effective in some uniprocessor systems, it is in general ineffective in multiprocessor systems where many kernel paths (due to system calls) simultaneously execute kernel programs and access kernel data.

≫ There is a related tool called *mutex*. Though it is the same as a semaphore, it is used slightly differently. For mutex, only a token holder can execute an up operation, and not others.

7.5.6 Semaphores

We observed in Section 7.5.3 on page 162 that solutions to the mutual exclusion problem using atomic (read/write) variables are difficult to construct and verify. We did not even attempt to solve other synchronization problems using atomic variables. Atomic variables are not sophisticated enough to be used to solve complex synchronization problems efficiently. In addition, those solutions waste a lot of CPU time as they employ "busy waiting" or "continual retry" when processes cannot immediately enter the critical section: as long as a process is in the critical section, any other process that attempts to enter the critical section loops continuously in the entry section until it can enter the critical section. Busy waiting requires that the values of synchronization variables are repeatedly read and checked until they have the desired values: the CPU loops on synchronization variables. Busy waiting is a huge burden in practical systems where CPUs are shared by many active processes. In uniprocessor systems that prohibit process preemption, busy-waiting solutions become ineffective because when a process loops on a condition the whole system stalls. Even in multiprocessor systems, busy waiting causes a flood of memory requests that may choke the system bus or "processor to memory" network. Using a different synchronization tool called *semaphore* overcomes these shortcomings. Another motivation behind semaphore is to reduce complexity of the programming required to implement solutions to synchronization problems, especially the mutual exclusion problem. It is a very important synchronization tool, usually supported by operating systems. What, then, exactly is a semaphore?

A process needs coordination when its further progress does not guarantee safety. In such situations, tokens are often used as safety certificates. That is, progress of a process can guarantee safety when it holds a token.

```
struct semaphore {
    int         count;/* number of tokens in the semaphore */
    process* waitQ;/* processes that are waiting to receive tokens */
};
```

Figure 7.16: *Semaphore structure definition.*

Here, a semaphore is considered roughly as a token manager. A process requests a semaphore for a token, proceeds further after obtaining one, and returns it to the semaphore when no longer needed. When a process requests a token from a semaphore, it (the semaphore) issues the process a token if it has one. Otherwise, the requesting process is blocked until a token is available. The semaphore accepts the token as soon as a process returns one. Thus, the token request is a blocking operation, but not the token return. Actually, as all tokens are the same, the semaphore manager, instead of keeping them in semaphore storage space, keeps only a count on the number of available tokens, and manipulates their number.

≫ A semaphore is a special case of message-based interprocess communication, where senders send tokens to the semaphore, and receivers retrieve the tokens out of the semaphore. Unreceived tokens remain in the semaphore. Here, no two tokens are different.

Essentially, a *semaphore* is a shared data structure that has two member components (see Fig. 7.16). The first component is an integer variable, referred to as *count*, which takes value from a range of integers to indicate the number of tokens the semaphore has. If the *count* value is zero, the semaphore is said to be empty. The second component is a wait queue, referred to as *waitQ*, to hold the processes waiting to receive tokens from the (empty) semaphore. The initial value of *count* defines the initial number of tokens. Usually, a semaphore is created with a fixed number of tokens.

A semaphore structure, apart from initialization, is accessed only by invoking two "atomic" operations, namely **down** and **up**, respectively, for token request and token release. [Many names are used for these two operations. Some popular names are P and V (first letters of equivalent words for down and up in Dutch), acquire and release, wait and release, wait and signal, etc.]

When a token is available, a decision must be made about which waiting process to select for issuing that token. This decision in essence defines the service discipline of *waitQ*. Some service disciplines such as FIFO may be considered fair and others such as random selection unfair because they may lead to starvation of a process. Based on the *waitQ* service discipline, semaphores are classified into various types. The study of such classification is beyond the scope of this book.

The semantics of the two semaphore operations are sketched in Fig. 7.17 in algorithmic forms. A process takes a token out of a semaphore by invoking the **down** operation on the semaphore, and inserts a token into a semaphore by invoking the **up** operation on the semaphore. In a **down** operation execution, a process attempts to take a token out of the semaphore. If there are tokens in the semaphore (i.e., *count* > 0), the process removes one token by way of decrementing the *count* component by one. Otherwise, (i.e., *count* = 0), the process waits on the semaphore queue until it receives a token. It is not a busy wait. The process blocks itself, and releases the CPU while it is waiting at the *waitQ*. In an **up** operation execution on a semaphore, a process first

```
/* Get a token from semaphore sem */
void down(semaphore* sem)
{
    if (sem-> count > 0) {
      sem-> count = sem - > count - 1;
    }
    else {
      put the calling process in the sem - > waitQ;
      block; /* invoke CPU scheduler to release the CPU */
      /* process returns here when rescheduled and it has got a token */
      remove the calling process from the sem - > waitQ;
    }
    return;
}

/* Return a token to semaphore sem */
void up(semaphore* sem)
{
    if (not empty sem- > waitQ) { /* allocate token to a waiting process */
      select a process from the sem-> waitQ;
      awake the selected process;
    }
    else { /* put the token in the semaphore */
      sem-> count = sem-> cont + 1;
    }
    return;
}
```

Figure 7.17: *A typical semaphore implementation.*

checks if there are any processes waiting at the *waitQ*. If there are some, one of them is awakened. (It is a positive wakeup, and the awakened process now has a token, and is ready to run again.) Otherwise, the semaphore member variable *count* is incremented by one.

A typical problem with some semaphore implementation is starvation. Starvation is a situation in which one or more processes wait indefinitely at the semaphore queue and new **down** operations overtake the waiting process(es). Fairness of a semaphore implementation is determined by the scheduling policy followed in the management of waiting processes. Most semaphore implementations are assumed to exhibit the fairness property, that is, no process while executing the **down** remains delayed forever if **up** operations are performed frequently by other processes. The need for fairness arises when many processes are simultaneously delayed, all attempting to execute the **down** operation on the same semaphore. Clearly, the implementation must choose which process is allowed to proceed when an **up** operation on that semaphore is ultimately performed.

≫ The core semantics of a semaphore is independent of its queue service discipline.

It is of utmost importance that the two semaphore operations, **up** and **down**, are atomic. That is, the operation executions on the same semaphore variable must exclude one another. This situation itself is a mutual exclusion problem; however, the critical section (containing **up** and **down** implementations) is very small. We can solve the problem using a suitable solution discussed in Sections

7.5.3 to 7.5.4; they do involve busy waiting albeit for a shorter duration. As the critical section is very short, busy waiting occurs rarely in practice.

Based on the range of values that *count* can take, semaphores are classified into binary-, general-, and counting semaphores. The binary semaphore takes the count value as either 0 or 1; the general semaphore takes any non-negative value; and the counting semaphore takes any integer value. In the counting semaphore, the positive value of *count* indicates the number of tokens available and the negative value of *count* indicates the number of processes waiting for tokens. These semaphores offer flexibility and convenience, but they are all equivalent when the *waitQ* is the same. That is, one can be implemented using the other. In the case of the counting semaphore, *count* is always updated (decremented for **down** operation and incremented for **up** operation). If the resulting value of *count* is negative, then the process is put into *waitQ* for **down** operation, and a process from *waitQ* is chosen and awakened for **up** operation.

≫ There is a related tool called *mutex*. Though it is the same as a semaphore, it is used slightly differently. For mutex, only a token holder can execute an up operation, and not others.

Semaphores are normally used in synchronizing unconditional ordering of critical section executions. Whenever two or more processes need to synchronize their relative speeds, we can also use a semaphore to block fast processes. A few semaphore-based solutions to critical section problems are discussed in the following subsections.

≫ A binary semaphore acts like a switch: the semaphore is either open or close. It can hold, at most, one token. Counting and general semaphores are useful when there is a need to manage multiple identical copies of a shared resource.

A Solution to the Mutual Exclusion Problem

Semaphores are widely used to control entries to critical sections. The semaphore counter is used as a locking mechanism. A semaphore-based solution to the mutual exclusion problem is presented in Fig. 7.18. This solution is very simple compared to those constructed using atomic variables in Section 7.5.3. It uses a single semaphore variable whose *count* component is initialized to 1 and waitQ is initialized to *NULL*. By initializing the counter value to 1, it is possible to prevent more than one process from entering the critical section at a time. The entry section consists of a single **down** operation on the semaphore, and the exit section a single **up** operation on the semaphore. The solution ensures the properties of mutual exclusion and liveness. The fairness property depends on the implementation of the *waitQ* management.

A Solution to the Producer–Consumer Problem

The producer–consumer problem introduced in Section 7.4.2 on page 157 is solved easily by using semaphores. We have a buffer with n, $n > 0$, slots

```
Data structure and initial value
semaphore sem = {1, NULL};

down(&sem); /* entry section */
⟨ critical section; ⟩
up(&sem);   /* exit section */
```

Figure 7.18: *A solution to the mutual exclusion problem using a semaphore.*

```
Data structures and initial values
    semaphore empty = {n, NULL}; /* n = number of slots in the buffer */
    semaphore full = {0, NULL};
    semaphore mutex = {1, NULL};

void put(item* m)
{
    down(&empty);
    down(&mutex);
    ⟨add item m in the buffer;⟩
    up(&mutex);
    up(&full);
}
void get(item* m)
{
    down(&full);
    down(&mutex);
    ⟨remove an item from the buffer;⟩
    up(&mutex);
    up(&empty);
}
```

Figure 7.19: *Implementation of a bounded buffer using semaphores.*

to hold unconsumed data items. The buffer is accessed by two operations, namely **get** and **put**. A general schematic of **get** and **put** operations is given in Fig. 7.19. The solution uses three semaphores: *empty*, *full*, and *mutex*. The *empty* and *full* semaphores count the number of empty and full slots, respectively, available in the buffer. Their initial values are n and 0, respectively, indicating that initially there is no data item in the buffer and all slots are empty. These two semaphores control the number of simultaneous executions of **put** and **get** operations, respectively. The *mutex* semaphore is used to ensure mutually exclusive access to the buffer; the semaphore is imitalized to 1.

The put operation first removes a token from the *empty* semaphore by way of reserving a free slot in the buffer. Then, it adds an item in the buffer, which is guarded by the *mutex* semaphore. It then inserts a token into the *full* semaphore to inform receivers that there is a new item in the buffer. The implementation for the get operation is symmetric.

A Solution to the Readers–Writers Problem

Solutions to the readers–writers problem of Section 7.4.3 on page 158 are generally asymmetric. To enter the critical section, either readers get priority over writers or vice versa. Therefore, there are a few varieties of solutions to this problem: (1) If a writer is ready to execute the critical section, no new readers may enter the critical section before the writer completes its critical section execution; (2) No reader waits to enter the critical section until a writer has obtained permission to do so. We study a solution for the latter variety in this subsection.

```
Data structures and initial values
    semaphore writeSem = {1, NULL};
    semaphore readSem = {1, NULL};
    int readCount = 0;

void write(Value val)
{
    down(&writeSem); /* exclude other writers */
    ⟨write val in the shared variable proper;⟩
    up(&writeSem);
}
Value read()
{
    Value val;
    down(&readSem); /* exclude other readers */
    readCount = readCount + 1;
    if (readCount == 1) down(&writeSem); /* special reader: exclude writers */
    up(&readSem);
    ⟨read val from the shared variable proper;⟩
    down(&readSem);
    readCount = readCount – 1;
    if(readCount == 0) up(&writeSem); /* special reader */
    up(&readSem);
    return val;
```

Figure 7.20: *A solution to the readers–writers problem using semaphores.*

Figure 7.20 presents a typical solution to the readers-writers problem. The solution uses two semaphores, namely *writeSem* and *readSem*, and an integer variable *readCount*. The semaphores are initialized to 1, and *readCount* to 0. The solution uses the *writeSem* semaphore to exclude writers one another in the **write** operation. The number of *active* readers (that are in the entry section or the critical section) is stored in the *readCount* variable. The *readSem* semaphore controls accesses to this synchronization variable. The **read** operation is simple: it increments the *readCount* under the *readSem* semaphore, performs the actual **read** proper, and finally decrements the *readCount* again under the *readSem* semaphore. A reader that sees *readCount* = 0 at the beginning of the entry section or at the end of the exit section, however, is special and has a special task to perform. In the entry section, a special reader obtains the *writeSem* semaphore to block new writers in entering the critical section, and in the exit section a special reader releases the *writeSem* semaphore. Readers that enter or leave the critical section while other readers are present in the critical section ignore the *writeSem* semaphore. The solution is not free from starvation even with FCFS semaphores as readers keep on entering the critical section may perpetually overtake the writers. See the Literature section at the end of this chapter for starvation-free solutions to the readers–writers problem.

A Solution to the Dining-philosophers Problem

The dining philosophers problem was introduced in Section 7.4.4 on page 159. Here chopsticks are the resources allocated to hungry philosophers. A simple solution is to represent each chopstick by a semaphore initialized to 1. A hungry philosopher first takes the left chopstick, and then the right chopstick. Having finished eating, she releases the left chopstick first, and then the right chopstick. This solution is simple and ensures the safety property that no two neighbouring philosophers can eat simultaneously; it is, however, not free from deadlock. A deadlock occurs when all the philosophers become hungry at the same time, and each picks up her left chopstick and waits indefinitely to get the right chopstick. (The definition of deadlock and a discussion on its issues appear in Section 7.6.)

It has been shown in the synchronization literature that there is no symmetric[4] solution to the dining philosophers problem. Then, to solve the problem we need to break symmetry, and there are several ways to do this: different processes may use different algorithms or randomization. A simple solution is presented in the following paragraph.

Processes are partitioned into two categories, and they execute slightly different algorithms. We assume that the number of philosophers is even, and are numbered consecutively starting from zero, say clockwise. All even numbered processes form one category and odd numbered processes the other. Odd numbered processes pick up the right chopstick first, and the even numbered processes the left. This solution ensures the safety and deadlock freedom properties. If semaphores are FIFO, it also ensures the starvation freedom property.

A Solution to the Sleeping-barber Problem

In Fig. 7.21 we present a solution to the sleeping barber problem described in Section 7.4.5 on page 159. The solution uses three semaphores: (1) **customer** to coordinate customers in the waiting room, (2) **barber** to coordinate activities of the barber, and (3) **mutex** for their general mutual exclusion. The **customer** semaphore keeps track of total number of waiting customers. When a customer arrives, she attempts to acquire the **mutex** first. Then she checks whether she can have a free chair (either in the waiting room or the barber's room). If there is no free chair, she departs the barbershop releasing the **mutex**. Otherwise *she* occupies a chair (statement in the figure is: nFreeChairs = nFreeChairs − 1). She then notifies her presence to the barber, releases the **mutex**, and waits for the barber. The barber algorithm is an infinite-loop type. In each iteration, she waits for some customer to come. When there are customers, she takes the **mutex**, gets up from her chair, and tells customers that she is ready to cut their hair. Finally she releases the **mutex**, and serves one customer.

What properties does the solution provide? It ensures the safety property that the barber cuts the hair of one customer at a time, and that there is no

[4] A solution is said to be symmetric if all processes are identical and may only refer to chopsticks by their names like chopstick(left) and chopstick(right), and if all shared variables have the smae initial values.

```
Data structures and initial values
    semaphore customer = {0, NULL};/* initially no customer */
    semaphore barber = {0, NULL};          /* initially barber is not ready */
    semaphore mutex = {1, NULL};  /* controlling access to nFreeChairs variable */
    int nFreeChairs = init-value;  /* greater than 1 */ /* for controlled
                                        accesses to the nFreeChairs variable */

Customer Algorithm
{
    down(&mutex); /* begin of mutual exclusion */
    if (nFreeChairs > 0) {
      nFreeChairs = nFreeChairs – 1; /* grab a free chair */
      up(&customer); /* tell barber that a new customer is in the shop */
      up(&mutex); /* end of mutual exclusion */
      down(&barber); /* wait for her turn to cut hair */
    } else {
      up(&mutex); /* end of mutual exclusion; she leaves the barber shop dejected */
    }
}
Barber Algorithm
repeat-forever {
    down(&customer); /* if there are customers, choose one; otherwise sleep */
    down(&mutex); /* mutual exclusion with customers to manipulate chairs */
    nFreeChairs = nFreeChairs + 1; /* frees up one chair */
    up(&barber); /* tell customers that she is ready to cut hair */
    up(&mutex); /* end of mutual exclusion */
    ⟨Cut hair of the chosen customer;⟩
}
```

Figure 7.21: *A solution to the sleeping-barber problem using semaphores.*

deadlock in the system. It also satisfies the liveness property. It, however, fails to ensure the starvation freedom property.

7.5.7 Spinlock

In the semaphore implementation (see Fig. 7.17 on page 173), adding (or removing) processes to (or from) the semaphore waiting queue is a costly operation compared to incrementing and decrementing the counter variable. In addition, when there is no token available in the semaphore, a down operation execution on the semaphore blocks the executing process, and calls upon the CPU manager to release the CPU from the process, thereby causing a context switch. We know from Chapter 5 that context switches incur additional overhead. If the execution time of the critical section is short compared to the summation of enqueue, dequeue, and context switch times, semaphore would be an inefficient tool in multiprocessor systems for "processor" synchronization. We use a different flavour of semaphore called spinlock in multiprocessor systems to synchronize the CPU activities.

A *spinlock* is like a semaphore, but there is no waiting queue associated with the spinlock. The spinlock is a special kind of shared non-negative integer variable that is solely manipulated by atomic **up** and **down** operations. A typical implementation of spinlock operations is shown in Fig. 7.22.

```
/*Get a token from spinlock lock*/
void down(spinlock* lock)
{
    while (*lock == 0) do;
    *lock = *lock - 1;
}
/*Return a token to spinlock lock*/
void up(spinlock* lock)
{
    *lock = *lock + 1;
}
```

Figure 7.22: *A typical spinlock implementation.*

Reading and incrementing the lock value in an **up** operation execution (and reading, testing, and decrementing the lock value in a **down** operation execution) must be done indivisibly. When a CPU, in a **down** operation execution, finds the spinlock value is zero, it repeatedly reads and checks the spinlock instead of blocking the running process. That is, it "spins" on the lock until the lock value becomes greater than zero, thereby causing no context switch. Many multiprocessor architectures provide support for spinlock via special machine instructions. Note that spinlocks cannot be used in uniprocessor systems. Spinlocks are primarily used to synchronize CPU activities. For example, it was mentioned in Chapter 5 that in multiprocessor systems many CPUs can simultaneously access the ready queue. Such accesses to the ready queue can be synchronized using spinlocks.

≫ Solaris supports a locking mechanism that is in between spinlock and semaphore. It is called *adaptive mutex*, and is very effective in multiprocessor systems for thread synchronization. The adaptive mutex is like a spinlock. A thread trying to acquire an adaptive mutex spins on the mutex if the thread holding the mutex is in the running state. Otherwise, the requesting thread blocks itself and releases the CPU.

7.5.8 Critical Region

The semaphore and spinlock are very elementary synchronization tools. They may be used to solve almost all synchronization problems. However, if used incorrectly, the system robustness may be compromised. For example, if in a program fragment one mistakenly performs a **down** operation instead of a required **up** operation, it may lead to permanent blocking (or self-locking) of the executing processes. Another example is that a program fragment misses a **down** or **up** operation. Such mistakes are natural and common in software development. However, these kinds of subtle development errors are sometimes difficult to detect, because in many situations time-dependent concurrency errors are irreproducible or extremely difficult to reproduce. To eliminate these kinds of silly, unintentional mistakes in programs, the concept of critical region was invented. Critical region is a programming language construct, and is used to eliminate simple silly programming errors that one may make using semaphores. Note, however, that the critical region construct does not eliminate programming errors of all kinds. It only helps the developers of synchronization solutions in eliminating silly errors and reducing the number of errors.

A *critical region* is defined by a higher-level language statement. The general construct is "region v do S", where v is called a *region variable*. A region variable is a shared data structure that is used only in statements like S under the control of the region variable. Statement S is actually the critical section, and v guards the executions of S. When a process executes S, it has

the exclusive right to the use of v. All statements under the same region variable are guaranteed execution in mutual exclusion. A compiler can check if a region variable is ever used outside critical regions, and signal a compilation error if it is found to be so. It is the duty of the compiler to ensure that all critical regions that use the same region variable v are executed mutually exclusively. The compiler translates the region statements into full proof implementation of critical regions (see Section 7.5.11 on page 185). It is up to that implementation what fairness condition it will follow to schedule waiting processes to execute their critical regions. Critical regions that refer to different region variables can, however, be executed concurrently.

For a solution to the mutual exclusion problem, we put all similar critical sections under the care of a single region variable. The mutual exclusion is guaranteed by the compiler's translation of the programs.

7.5.9 Conditional Critical Region

The synchronization tools discussed so far in the previous subsections help processes to synchronize their activities for special conditions. For example, a binary semaphore acts as a switch. These synchronization tools are very general, and are used to solve almost all synchronization problems. However, using them to synchronize processes for arbitrary conditions becomes a challenging task. For example, suppose we want to make some processes wait until the value of a shared integer variable becomes greater than a specific value. Constructing solutions for these kinds of conditional synchronization problems using the previously mentioned tools is an uphill task, if not wholly impossible. To alleviate the design burden for solutions to arbitrary conditional synchronization problems, the concept of conditional critical region was invented. This new tool enables design of synchronization solutions for arbitrary conditions with lesser effort.

Conditional critical region is another programming language construct, and is a generalization of the unconditional critical region of Section 7.5.8. The general construct is "region v do $S1$ ..., **await**(B), ..., $S2$ done", where v is called a *condition variable* that is a shared data structure, and B is an arbitrary boolean expression that can refer to values of the variable v, and **await** is a new synchronization primitive. The entire sequence of statements "$S1$, ..., **await**(B), ..., $S2$" is a critical region. Like unconditional critical region, all statements under the care of a common condition variable are executed mutually exclusively.

Without condition variables, a process would need to have continual polling (possibly in a critical section) to check if the desired condition B is met. This can be very resource consuming since the process would be continuously busy in this activity, and quite unnecessarily occupy the CPU. A condition variable is a means to achieve the same goal without polling. While a region variable implements synchronization by controlling the accesses of processes to the region variable, a condition variable allows processes to synchronize based on the actual values of the condition variable. Like region variables, a condition variable v also ensures mutually exclusive execution of the regions guarded by

the variable. However, if condition B evaluates to false (inside the critical region), the process temporarily stops executing the critical region and waits in some event queue after releasing the region for other processes. The process would wait in the event queue until the expression is satisfiable. When another process makes changes to the value of v, the condition B may become true. Therefore, when a process exits its critical region, it makes all waiting processes reevaluate their await conditions. (It is a non-positive wakeup.) If the boolean expression B is not satisfied at that point of reevaluation, the process temporarily suspends its execution of the statement, and resumes its execution at a later time.

A Solution to the Producer–Consumer Problem

We now solve the message buffer problem introduced in Section 6.4.3 on page 143. The solution presented there may not work if **put** and **get** operations are executed concurrently. Here we present a solution using the conditional critical region construct. The solution is presented in Fig. 7.23. It is easier to comprehend the solution and prove its correctness. Buffer is a shared data and it acts as a condition variable. The structure of **put** and **get** routines indicate that they are effectively critical regions, and the buffer variable ensures their mutual exclusion. The await statements in the two routines help processes to synchronize until certain conditions are met. For example, in the put routine, if all the message slots are occupied, the sender waits.

A Solution to the Readers–Writers Problem

A simple solution to the readers-writers problem using the conditional critical region construct is presented in Fig. 7.24. The solution uses one region variable *writer* and one condition variable v. Readers, under the condition variable v,

```
Constant
    int n = buffer-size;

Data structures and initial values
    struct T {
        message buff[n];
        int n_message = 0;
        int in = 0;
        int out = 0;
    } buffer;

void put(message* m)
{

region buffer do {
await(n_message < n);
buff[in] = *m;
in = (in + 1)% n;
n_message = n_message + 1;
}
}
```

```
void get(message* m)
    {
    {
    region buffer do
await (n_message > 0);
        *m = buff[out];
    out = (out + 1)% n
n_message = n_message – 1;
    }
    }
```

Figure 7.23: *Implementation of bounded buffer using a condition variable.*

```
Data structures and initial values
    struct T {
        int waitingWriter;
        int readingReader;
    } v = {0, 0};

    shared boolean writer;

Reader_i:
    region v do await(v.waitingWriter == 0);
                    v.readingReader = v.readingReader + 1;

    ⟨Read⟩;
    region v do v.readingReader = v.readingReader – 1;

Writer_i:
    region v do v.waitingWriter = v.waitingWriter + 1;
                    await(v.readingReader == 0);
                done;
    region writer do ⟨Write⟩; /* mutual exclusion of writers */
    region v do v.waitingWriter = v.waitingWriter – 1;
```

Figure 7.24: *A solution to the readers–writers problem using condition and region variables.*

first wait until all writers leave the critical section and then increment the active reader counter *readingReader;* then, they read. Finally, they decrement the counter again under *v*. Writers, under *v*, first increment the active writer counter *waitingWriter* and then wait until the readers have left the critical section. Then, they write under the region variable *writer* to exclude one another. Finally, they decrement the counter *waitingWriter* under *v*.

7.5.10 Monitor

Monitor is another programming language construct for process synchronization. A monitor is an abstract data structure that consists of a set of variables whose values represent states of the monitor. It also has a set of functions that are executed to operate on the monitor to change its state. Resembling C++ or Java programming language, the variables are private to the monitor, and some monitor functions are made public. One cannot access private variables from outside the monitor. The public functions are the sole user interface or entry points to the monitor, and they may take formal parameters. The monitor functions can access private variables and parameters. Monitor is very similar to traditional (C++ or Java) object paradigm. However, process synchronization is a little more involved than just traditional objects. The condition is that at the most one process can be executing at any time inside a monitor. That is, processes access a monitor mutually exclusively. Monitor locks itself when a process begin an execution of a procedure and unlocks when the process completes its execution or blocks for some condition. If another process tries to invoke a procedure of a locked monitor, the process is suspended until the monitor is unlocked. The process is said to

be blocked at the entry to the monitor. The processes do not need to bother about how the synchronization is performed. Normally the compiler supplies these synchronization codes.

A monitor, in addition to ordinary variables, may have synchronization-specific variables called *condition variables*. (One should not confuse this term with conditional critical region variables.) A condition variable supports two interface operations, namely *wait* and *signal*. Inside a monitor, a process suspends its execution by executing the wait operation on a condition variable, and resumes it when another process executes a signal operation on the same condition variable. (Though blocked and not executing the suspended process is still considered to be active inside the monitor.) The signal operation resumes only one suspended process; if there is no suspended process, then the signal operation becomes a no-op. The signal operation does not activate any processes that are waiting at the entry to the monitor. However, only one of the two processes can continue its execution in the monitor while the other has to wait until the former completes its execution or suspend itself by executing another wait operation. In the original implementation of monitor by Hoare, a signalling process always waits until the resumed process leaves the monitor or executes another wait operation. In a variation by Hansen, the signalling process must leave the monitor by executing a return statement in order for the signalled process to continue.

≫ All processes inside a monitor are "active". But only one of them is executing the monitor. The rest are blocked, i.e., waiting on some condition variables.

We use the following syntax to declare a monitor.

```
Monitor name-of-the-monitor {
Variables (global and condition) declarations;
Procedure proc-1(parameters);
Procedure proc-2(parameters);
:
Procedure proc-n(parameters);
{ initialization code, aka, constructor }
};
```

≫ A monitor should allow at the most one process to execute within it. Assume that a process *A* signals on a condition variable on which another process *B* is waiting; the question is, among *A* and *B*, which process will execute a monitor procedure and which process will exit the monitor or wait. There are three popular disciplines: (1) *A* exits the monitor (signal and exit); (2) *A* waits until *B* leaves the monitor (signal and wait); and (3) *B* waits until *A* leaves the monitor (signal and continue). Java implements the signal and continues discipline.

To solve the mutual exclusion problem using a monitor, put the critical section inside the monitor something like the following: The process can invoke mutex.Critical_Section() to execute the critical section.

```
Monitor mutex {
Critical_Section(void);
Procedures(parameters);
};
```

For a specimen example, we solve the dining philosophers problem using a monitor. (It was solved in Section "A Solution to the Dining Philosophers Problem" on page 177 using semaphores.) Assume that there are $n > 1$ philosophers. The chopsticks must be accessed in a mutually exclusive way. The routines that access the chopsticks are put inside the monitor for synchronized access. The philosopher waits on a condition variable when the chopstick

requested is in use. For brevity, in this section we use left and right to refer the ids of the left chopstick and right chopstick, respectively, of each philosopher. This simple solution is given below.

```
repeat-forever {
  think();
  DP.get-chopstick(left);
  DP.get-chopstick(right);
  eat();
  DP.put-chopsticks();
}
Monitor DP {
  const int n = init-number; /* greater than 1 */
  boolean cs[n] = 0; /* representing chop sticks */
  condition csc[n];
  get-chopstick(i)
  {
  if(cs[i] ≠ 0) wait(csc[i]);
  cs[i] = 1;
  }
  put-chopsticks()
  {
  cs[left] = 0;
  cs[right] = 0;
  signal(csc[left]);
  signal(csc[right]);
  }
};
```

This solution assures the safety, but not the liveness property. If every philosopher picks the left chopstick then they form a circular wait and then they are deadlocked. To avoid circular wait, a simpler solution would be just to prohibit all the n philosophers to request chopsticks simultaneously. That is, a philosopher can request a chopstick only if the number of other philosophers requesting chopsticks is less than n. The solution incorporating this idea is given below.

```
repeat-forever {
    think();
    DP.safe();
    DP.get-chopstick(left);
    DP.get-chopstick(right);
    eat();
    DP.put-chopsticks();
}
```

```
Monitor DP {
const int n = init-number; /* greater than 1 */
int count = 0;
boolean cs[n] = 0; /* representing chop sticks */
condition csc[n];
condition notsafe;

safe()
{
  count++;
  if (count == n) wait(notsafe);
}
get-chopstick(i)
{
  if(cs[i] ≠ 0) wait(csc[i]);
  cs[i]= 1;
}
put-chopsticks ()
{
  count–;
  if(count == n–1) signal(notsafe);
  cs[left] = 0;
  cs[right] = 0;
  signal(csc[left]);
  signal(css[right]);
}
};
```

7.5.11 Comparison of Synchronization Primitives

The synchronization tools introduced above are equivalent in the sense that any synchronization problem can be solved using any of the tools. However, for a given problem some tools may lead to complex, difficult to understand solutions, or inefficient solutions. A particular tool is ideal to solve particular kinds of synchronization problems, but leads to complex solutions for other kinds of synchronization problems. A tool is said to support a particular type of synchronization if it provides primitives that make it convenient for constructing solutions to that class of synchronization problems. Each tool solves certain kinds of problems elegantly and efficiently.

For example, we can implement critical region construct using semaphores. For each region variable *v*, which is used in statements such as "region *v* do S", a compiler allocates a new binary semaphore variable, say *v_mutex* whose counter is initialized to 1. The region statement is

transformed by the compiler as "**down**(*v_mutex*); *S*; **up**(*v_mutex*)". We can also implement conditional critical region- and monitor constructs using semaphores and other control variables. See the Literature section of this chapter for appropriate articles. These constructions imply that we can implement language-based synchronization tools using semaphores.

We have assumed that semaphore operations (**up** and **down**) are indivisible, and the indivisibility can be achieved through machine instruction-based constructs from Sections 7.5.3 and 7.5.4. We have assumed that read and write operations on memory cells are executed atomically by the memory hardware. Thus, reading/writing of memory cells is a fundamental synchronization problem solved by the memory hardware.

The irony of the software solutions (presented in previous subsections) to various synchronization problems is that we transfer problems from one level to another: from macro level to micro, from coarse granules to fine, from complex to simpler problems. True synchronization, by atomic executions of read and write operations on memory cells, is provided by the memory hardware. Software solutions elevate the hardware capability to higher levels so that it becomes convenient for users to use synchronization tools. What happens when the memory hardware does not guarantee atomicity of read and write operation executions? Can we develop good synchronization primitives in such systems? These are the fundamental questions in synchronization theory. We will address this question in Section 7.7. Incidentally, the Lamport solution (see Section "The Lamport Solution" on page 167) to the general mutual exclusion problem does not assume the atomicity of read and write operations on primitive shared variables. His solution, however, requires shared variables of unbounded size.

7.6 Deadlock

A computer system has many resources that processes use. One of the major advantages provided by operating systems is the ability to share (i.e., time-multiplex) resources among processes to improve resource utilization and system performance. However, unless done carefully, such sharing may lead to unwanted situations where processes are unable to make progress in their computations. Operating system designers and developers should be well aware of the problems due to resource sharing, and also well acquainted with solutions to those problems.

≫ An example of identical resource units is memory locations.

In this section, we use the term resource in a very generic sense to mean anything that may conditionally block processes. A resource may be a "reusable resource" such as a printer, or a "consumable resource" such as a message. A resource may have many identical units; the units are considered equivalent, and any unit will satisfy a request on the resource. There is a priori known fixed number of units of each reusable resource. Processes compete for reusable resources. A unit of a reusable resource can be assigned at the most to one process at a time. The total number of units of a consumable resource is

not fixed. Producers produce consumable resources. Once a consumable resource unit is allocated to a consumer, the unit is considered destroyed.

A process acquires a resource, uses it, and finally releases it. At the time of acquisition, the process might be blocked if the resource is not immediately available. A process requests a resource, and waits until the resource is granted to it. When granted, the process comes out of the waiting state and uses the resource. Finally, when the resource is no more required, the process releases the resource (reusable case) or destroys it (consumable case). A resource, once allocated to a process, cannot be preempted from the process by force. A process may request as many resources as it needs to accomplish its task, but, not for more reusable resources than the system has.

As mentioned above when a process requests a resource, the process is blocked until a unit of the resource is allocated to it. A blocked process cannot come out of the waiting state on its own and make progress in its execution (i.e., change its state) until it is unblocked by someone else. In such an environment, processes may enter into deadlocks or deadly embrace situations. In any computer system where processes share exclusive resources or exchange messages, we have to handle deadlock related issues.

A *deadlock* is a situation in which there is a group of agents such that each one in the group is waiting for some (resource allocation and release) events to occur, but these events never happen because the events are supposed to occur at other agents in the group. In our case, the agents are processes. The waiting processes never change their states as each has requested a resource that is held by another waiting process. The processes are in a deadly embracing situation, and they cannot come out of the situation on their own. They are blocked forever, one on another, unless an external agent takes some action to unblock them.

In the simplest case a deadlock occurs when two processes, say P_1 and P_2, try to use two exclusive resources r_1 and r_2. Suppose P_1 has r_1, and P_2 has r_2. After a while, P_1 requests for r_2, and P_2 for r_1. See Fig. 7.25; it is called a *resource-allocation graph*, and represents resource holding and unsatisfied requests. The solid arrows represent holding information, and the broken arrows requesting information. Now, P_1 is waiting for P_2 to release r_2, and P_2 is waiting for P_1 to release r_1. Thereby, they form a circular waiting list: P_1 is waiting for P_2 (to release r_2) that in turn is waiting for P_1 (to release r_1). Consequently, none of the two processes can make progress in their respective computations, leading to a deadlock situation.

≫ That a process is blocked does not necessarily imply that it is involved in a deadlock. Also, due to bad system designs some process(es) may be *blocked forever* without involving in deadlocks. The processes are frozen individually and not in an embracing situation. For example, we have a semaphore initialized to 0. A process, however, instead of performing an **up** operation on the semaphore does a **down** and perpetually blocks or freezes or self-locks itself on the semaphore. In this book, we do not study these issues arising out of wrong code development.

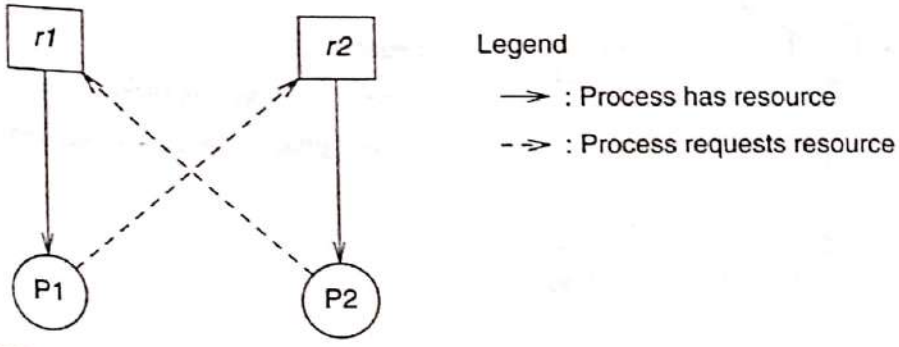

Figure 7.25: *A typical deadlock scenario.*

As mentioned above, a process waiting for a resource does not necessarily mean that it is in a deadlock. There are criteria to declare the existence of deadlocks in a system of processes. The following are the four necessary conditions for deadlock formation.

- **Mutual exclusion:** There are resources that cannot be used by more than one process concurrently. If one process holds an exclusive resource, another requesting process has to wait until the first process releases the resource.
- **Non-preemptive assignment:** Resources allocated to processes cannot be preempted. Only the process that has acquired the resource can release it.
- **Partial allocation of resources:** Resource allocation is incremental. Each process holds some resources and requests additional resources as its computation evolves.
- **Circular waiting:** There exists a circular chain of waiting processes. Each process in the chain is waiting for a resource held by the next process in the chain.

Note that the above four are necessary conditions, but may not be sufficient for some systems. (The first two signify the usage of exclusive resources and the semantics of resource allocation, respectively.) If there is no circular chain of processes, then there is definitely no deadlock. If there is a circular chain, there is however a *possibility* of a deadlock. If a circular chain involves those resources that have one unit each, then there is definitely a deadlock. Otherwise, there may or may not be a deadlock, and we need to derive a correct answer by using additional information. For example, Fig. 7.26 depicts a scenario with a circular chain of processes not involving a deadlock. In the figure, small circles in boxes represent identical resource units. For example, resource r_1 has two identical units that are allocated to processes P_1 and P_3. When P_3 releases the r_1 unit, the unit can be allocated to P_2 to break the cycle.

≫ The presence of a knot is a sufficient condition for a deadlock. A knot is a non-empty set K of nodes such that reachable set of each node in K is the set K ifself.

Deadlock is undesirable. In fact, it is a serious problem as it leads to underutilization of resources and wastage of processing time. Processes

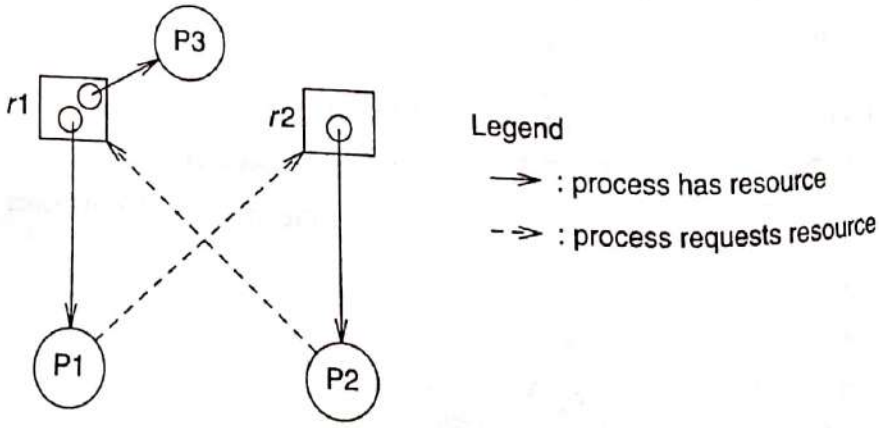

Figure 7.26: *A typical circular chain without a deadlock.*

involved in a deadlock cannot make any progress in their computations but unnecessarily occupy valuable resources. We need to solve this problem.

Deadlock can be dealt with at many levels. We can prevent deadlocks ahead of time (usually by rigid static rules), make careful moves (or aka transitions) to avoid deadlocks by looking at the current state (usually by dynamic or adaptive rules), detect and resolve once deadlocks occurred (by breaking one of the conditions), or simply ignore (popularly called Ostrich algorithm—stick your head in the sand and pretend there is no problem at all). Many solutions to the deadlock problem have been reported in literature. The solutions are classified into three broad categories: (1) prevention, (2) avoidance, and (3) detection and recovery.[5] Solutions in the former two categories ensure that at least one of the four necessary conditions (noted above) never holds, directly or indirectly. Solutions in the last category do allow deadlock formation; they run deadlock detection tests at regular intervals. If a deadlock is found, some processes are forcefully terminated to break the circular list of waiting processes. We discuss a typical solution in each of these three categories in the following three subsections, where we only deal with reusable resources.

≫ Many current operating systems, although consider deadlock issues, appear to restrict concurrency of kernel executions. For example, Linux 2.4 and its earlier releases use a single global kernel lock to synchronize concurrent kernel paths on multiprocessor systems. Before reduced level of kernel concurrency becomes a severe bottleneck, deadlock-related features will have to be incorporated in operating systems. The operating systems we are aware of stipulate manual deadlock prevention.

7.6.1 Deadlock Prevention

A deadlock prevention scheme specifies a set of rules that processes need to follow in acquiring resources. Their strict observance guarantees that no deadlock ever occurs in the system. Most prevention schemes violate at least one necessary condition for formation of deadlocks. In the following paragraph, we discuss a very simple scheme to prevent circular waiting by processes.

All resources are totally ordered. That is, each resource has an ordinal position in the total order. Processes always request resources in this order. If a process holds a resource at ordinal position k, it cannot request resources at ordinal positions k', when $k' \leq k$. If a process P waits for a resource r at ordinal position k and r is held by another process Q, then Q will not further request any resource at ordinal positions less than or equal to k. Hence, Q will never wait (directly or indirectly) for P. Consequently, there will be no circular list of waiting processes, and thereby, no deadlock forms in the system.

7.6.2 Deadlock Avoidance

Deadlock avoidance is a kind of deadlock prevention without any static ordering of resources. Normally, processes can request resources in an arbitrary order. A solution in this category keeps the system always in a safe state. A *safe* state is one in which there exists a way to complete executions of all processes without forming a deadlock. Initially, the system is in a safe state because we can always execute processes sequentially. By definition, a

[5] Levine lately has strenthened the four necessary deadlock conditions, and proposed that prevention and avoidance should be classified into a single category. See the Literature section at the end of this chapter for more information.

sequential execution does not throw the system into a deadlock because there is no competition for resources. A safe state is normally identified by a set of safety conditions. When a process makes a new request on a resource, an avoidance algorithm re-evaluates safety conditions to check whether the system remains in a safe state after the request is granted. If the state becomes unsafe, the request cannot be granted and hence the process needs to wait.

Unlike deadlock prevention schemes, a deadlock avoidance scheme does not force processes to acquire resources in a particular order, but it does require advance knowledge about what resources processes need to accomplish their tasks. Without such advance knowledge, it may not be possible to predict possible future states of the system. With the advance knowledge and that of the current allocation status of all resources, the scheme can always determine whether the current state is safe. Deadlock avoidance solutions differ from one another for information they need about the resource requirements of each process.

Let us study a simple example here. Let R be a resource with $2c + 1$, $c > 0$, identical units, and P and Q be two processes. They may need $2c$ and $c + 1$ units, respectively, to complete their executions. Suppose P already has c units, and Q one unit. They may each request c more units in future. The system is still in a safe state as we can complete the execution of both processes, at least sequentially in arbitrary order from now onwards. If we allocate one more unit to both P and Q, then P will have $c + 1$ and Q two units, and $c - 2$ units remain free. Now, the processes each may need $c - 1$ units to complete their executions, which is impossible to allocate as there are fewer free units. Therefore, the system is not in a safe state: although the processes are not in a deadlock now, they could however be in a deadlock in future. We need to avoid such deadlock formations in the future by carefully allocating resource units to processes. One simple solution is to execute processes mutually exclusively at different times. This, however, would lead to poor utilization of resources and an increase in process response time. A solution that would allow as much concurrency as possible without any possibility of forming deadlocks in the system is the need of the hour.

Initially, the system is in a safe state: no resource is allocated to any process. A resource allocation algorithm must keep the system in a safe state and, at the same time, ensure as much resource utilization as possible. What we need is an algorithm that can determine whether a given system state is safe or not. When a new request comes, we may first pretend to grant the request, and then evaluate the algorithm to determine whether the new state is safe. The request is granted only if it is safe in the new state as well. Otherwise, we shelve the request for future consideration. In the rest of this subsection, we present a deadlock avoidance algorithm that is popularly known as the *banker* algorithm.

Banker Algorithm

Suppose that there are m, $m > 0$, resources $r_1, \ldots, r_m$ that are used by n, $n > 1$, processes $P_1 \ldots, P_n$. Resource r_i has c_i, $c_i \geq 1$, identical units. These units are

equivalent in the sense that when a process requests for a resource r_i, any of the c_i units will satisfy the request.

Each process, at the start of its execution, declares the maximum requirements of all resources it needs to complete its execution. That is, the banker algorithm requires advance knowledge of maximum resource requirements for all processes. A process may not need all its required resources right at the start of its execution. It requests resources one by one in an unknown pattern. If a process's maximum need is satisfied, it is guaranteed that the process will release the resources in a finite time when it has finished using those allocated to it.

The *state* of the system is determined by the state of each resource unit (whether free or not), and the maximum resource requirements of all processes and the resource units they presently hold. A state is called *safe* if the system can enable all its processes complete their executions in a finite time without throwing any of them into deadlocks. In a safe state, it is always possible to satisfy new resource requests from each process in a way that all processes can complete their executions. Otherwise, the state is deemed *unsafe*. An unsafe state does not necessarily mean that there is a deadlock in the system. It only means that there is a possibility of deadlock forming in the future. Some unsafe states do have deadlocks. The banker algorithm takes a pessimistic position, and keeps the system always in a safe state.

The banker algorithm is presented in Fig. 7.27. The constants *units* identify the number of units of all resources. The variables *maxNeed* identify the maximum requirements for all processes, and the variables *allocated* identify the current resource allocation. The assertion that for all processes *i*, and all resources *j*, *unit[j] >= maxNeed[i][j] >= allocated[i][j]* is an invariant to the algorithm. The invariant says that a process cannot ask for more resource units than the system has, and the system never allocates more resources to a process than it needs. The algorithm is quite simple. It repeatedly looks for a process that it can complete with the then available free resources. If the algorithm ends with all processes completed, then the current state is safe.

Further, the system finds a dynamic order of processes such that if they are executed in that order deadlocks will not occur. However, the banker algorithm is quite expensive: each execution may take $O(n^2m)$ time, where n is the number of processes and m the number of resources.

Examples of The Banker Algorithm

Consider the following resource allocation state involving five processes P_0, P_1, P_2, P_3, and P_4, and five resources R_0, R_1, R_2, R_3, and R_4. *Max[i, j]* specifies the maximum number of instances that process P_i may request for resource R_j. *Alloc[i, j]* gives the number of instances of resource R_j currently allocated to process P_i. *Avail[j]* specifies the number of instances of resource R_j are currently available. We need to determine whether the system is in a safe state. Initially, Avail = [7, 7, 7, 7, 10].

```
Constants
    n = number of processes;
    m = number of resources;

Data structures and initial values
   const int unit[m] =          {c1,···,cm};     /* number of copies of m resources */
   const int maxNeed[n][m];                      /* maximum need of processes; initialized to proper values */
   int allocated[n][m];                          /* present allocation; initially all zero */

boolean safe(n,m,unit,maxNeed,allocated) /* check if state is safe */
{
    int available[m];
    boolean completed[n] = {false, false, ···, false};
    for(int j = 0; j < m; j = j + 1) available[j] = unit[j]−∑_{i=0}^{n−1} allocated[i][j];
    repeat
        boolean change = false;
        for (int i = 0; i < n; i = i + 1) {
            if (! completed[i]) { /* Pi is not complete */
                if (completion_possible(maxNeed[i], allocated[i], available)) {
                    for (int j = 0; j < m; j = j + 1) {
                        available[j] = available[j] + allocated[i][j];
                    }
                    completed[i] = true;
                    change = true;
                }
            }
        }
    until (! change);
    return     (∀j)[unit[j] == available[j]]);
}

boolean completion_possible(int maxNeed[m], int allocated[m], int available[m])
{
    for (int j = 0; j < m; j = j + 1) {
        if (maxNeed[j] – allocated[j] > available[j]) {
            return false; /* process needs more resources than currently available */
        }
    }
    return true;
}
```

Figure 7.27: *The banker algorithm.*

Max = [[4,2,3,1,1], [1,2,3,4,5], [3,2,3,1,1], [5,4,3,2,1], [2,0,0,2,2]]
Alloc = [[1,1,1,1,1], [1,0,1,1,1], [0,1,2,0,1], [3,0,2,1,1], [1,0,0,2,1]]
Avail = [1,5,1,2,5]

This is a safe state. We can schedule processes in this sequence to complete their executions: P_4, P_3, P_0, P_1, and P_2. After P_4 is complete Avail = ([1, 5, 1, 2, 5]+[1, 0, 0, 2, 1]) or [2, 5, 1, 4, 6]; after P_3 is complete Avail = ([2, 5, 1, 4, 6]+[3, 0, 2, 1, 1]) or [5, 5, 3, 5, 7]; after P_0 is complete Avail = ([5, 5, 3, 5, 7]+[1, 1, 1, 1, 1]) or [6, 6, 4, 6, 8]; after P_1 is complete Avail = ([6, 6, 4, 6, 8]+[1, 0, 1, 1, 1]) or [7, 6, 5, 7, 9]; and after P_2 is complete Avail = ([7, 6, 5, 7, 9]+[0, 1, 2, 0, 1]) or [7, 7, 7, 7, 10].

Now, suppose process P_0 requests a unit of R_0. If the system grants the request, then we have the following state of the system.

Max = [[4,2,3,1,1], [1,2,3,4,5], [3,2,3,1,1], [5,4,3,2,1], [2,0,0,2,2]]
Alloc = [[2,1,1,1,1], [1,0,1,1,1], [0,1,2,0,1], [3,0,2,1,1], [1,0,0,2,1]]
Avail = [0,5,1,2,5]

This is an unsafe state because requests on R_0 from processes P_0, P_2, P_3, and P_4 and on R_2 from P_1 cannot be satisfied.

7.6.3 Detection and Recovery

The prevention technique (presented in Section 7.6.1) forces one to organize all resources in a total order, and also compels processes to acquire them in that order. It may not be convenient for developers of applications and operating systems to do so. It also leads to poor utilization of resources, as processes are sometimes compelled to acquire resources that they may use only in the remote future or not at all.

The banker algorithm (given in Section "Banker Algorithm") though does not enforce processes acquiring resources in a predetermined order, but it adopts a pessimistic approach to avoid deadlocks. Each process has to declare its maximum requirements of resources when it starts even though it may not acquire them in this incarnation of program execution. The algorithm leads to poorer utilization of resources because the system may keep some resources idle on the assumption that some processes may need them soon. In addition, the banker algorithm is too costly to execute on each request for a resource.

In some practical systems, it may be worth allowing formation of deadlocks, and detecting and resolving deadlocks rather than preventing or avoiding their occurrences. This way of treating deadlocks is suitable in environments where deadlocks are rare and/or recovery is not very expensive. The system may not perform a "safeness" check right when it allocates resources to processes. Consequently, there is a possibility of deadlock formation in the system. The system runs some deadlock detection algorithm, at regular intervals. If a deadlock has indeed occurred, the system takes some corrective actions by forcefully terminating some processes and making their acquired resources available to other processes. The victim processes can be re-started later.

Here, we describe a simple deadlock detection scheme that works for systems having one unit of each resource. (Readers may consult the Literature section of this chapter to know about general solutions.) The system keeps track of which processes use which resources, and which processes wait for which resources. The scheme builds resource-allocation graphs, like the one shown in Fig. 7.25 on page 187. Given a resource-allocation graph, we merge resource nodes with respective processes that have the resources, and we obtain a new graph called *wait-for graph*, (see Fig. 7.28). The broken arcs in the wait-for graph indicate which processes are waiting on which processes. If P_i - - > P_j is an arc in the wait-for graph, then P_i is waiting for a resource that is held by P_j. A deadlock exists if and only if there is a cycle in the wait-for graph. (In Fig. 7.28, the two processes are deadlocked.) We can execute known graph algorithms to find cycles in wait-for graphs. Once a deadlock is found, the system forcefully terminates some

≫ A word of caution. Most real systems do not implement any sophisticated deadlock handling mechanisms. They go the ostrich way: pretend that deadlocks will not be frequent. In case system responsiveness becomes intolerable, reboot the system!

≫ The question is when or how frequently do we run the detection algorithm. A better approach is to associate a timeout with each pending request. On timeout, run the detection algorithm.

Figure 7.28: *The wait-for graph of Fig. 7.25.*

≫ A livelock is similar to starvation. The state of a process involved in a livelock does change, but it is effectively not progressing due to frequent aborts.

process(es) to break the deadlock. Care is necessary in terminating processes, as repeated termination of the same processes may induce livelock in the system.

7.7 The Real Challenge

In Section 7.5, we discussed many solutions to typical synchronization problems using well known tools. How fundamental are these solutions? What assumptions do we make to construct these solutions? Are these assumptions realistic? The most important assumption we made there is that the memory hardware ensures atomic executions of read and write operations on individual memory cells. Conflicting operation executions on the same memory cell exclude each other in real time. Consequently, the read and write operations are the true mutual exclusion problems and which the memory hardware solves. Software solutions elevate the hardware capability to higher levels so that it becomes convenient to use synchronization tools. What happens when multiple processors are connected to the memory through different ports of the memory? The read and write operation executions may truly overlap in real time. What happens if the memory hardware does not ensure atomicity of read and write operation executions? Can we develop good synchronization primitives in such systems? This is the fundamental question in synchronization theory.

Let us study a concrete example to understand this question better. Let v be an integer shared variable with initial value 0. A process P is writing 1 in v. To execute this write operation, a memory port takes some finite amount of time to overwrite the old value of v by the new value. What is guaranteed is that if some other process Q reads v before (respectively, after) the write operation execution starts (respectively, completes) will read 0 (respectively, 1). What value does Q read during the value transition from 0 to 1? This is in general unknown and may depend on the true state of the memory storage medium. What values does Q get if it repeatedly reads v in the writing period? The values read by Q are generally unpredictable if read during the transition period. Figure 7.29 displays a typical scenario of read and write executions on the shared variable v. In reality, the memory hardware arbitration unit solves this synchronization problem. The arbitration unit allows at the most one operation at a time to be performed on a given memory cell. Thus, the fundamental problem of synchronization is solved by the memory hardware. Their solution based on mutual exclusion (or critical region) is implemented in the hardware arbitration unit.

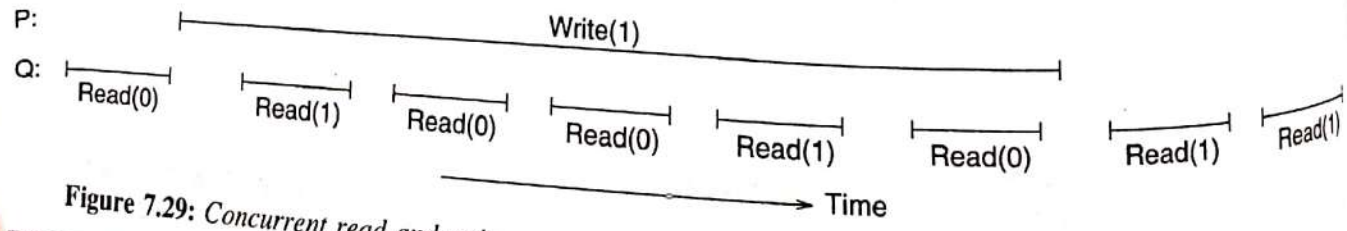

Figure 7.29: *Concurrent read and write on a memory cell.*

Incidentally, Lamport's Bakery algorithm (see Section "The Lamport Solution" on page 167) does not assume atomicity of read and write operations. Nevertheless, the solution needs unbounded size variables. Therefore, we have not really solved the fundamental synchronization problem in software if the memory arbitration unit does not effectively solve the problem at the level of the individual memory cell. The fundamental question is how we can ensure atomicity of (non-blocking, wait-free) read and write operation executions in the absence of the memory arbitration hardware. (Wait-freedom means operation executions of one process are not blocked by activities of other processes.)

Lamport defines the following three types of 1-writer multireader shared variables in which read and write operations can be executed in a *wait-free* manner.

- A *safe* variable is one in which a Read not overlapping any Write returns the most recently written value. A Read that overlaps a Write may return any value from the domain of the variable.
- A *regular* variable is a safe variable in which a Read that overlaps one or more Writes returns either the value of the most recent Write preceding the Read or of one of the overlapping Writes.
- An *atomic* variable is a regular variable in which the Reads and the Writes behave as if they occur in some total order that is an extension of their execution order.

Safe boolean variables come naturally in the form of flip-flops. Other variables are constructed from safe variables. Such constructions have been developed in recent years. However, the constructions are too complicated to present in a book such as this. Interested readers may consult articles referred to in the Literature section of this chapter.

Summary

Concurrency is the notion of doing multiple related activities simultaneously. Concurrency is ubiquitous in modern operating systems, and synchronization (of threads, processes, processors, kernel paths) is becoming increasingly important. This chapter introduces subtleties in process interactions. Race conditions arise in these systems because of the concurrent access of shared data by multiple agents. Not all outcomes of a race condition are correct. Unless race conditions are handled carefully, chaos ensues in these systems.

Synchronization is the coordination of concurrent accesses to shared resources, and it is perhaps the single most challenging aspect in the design and development of multiprocess operating systems, especially for multiprocessor platforms. There are many kinds of synchronization problems. We need to handle them tactfully to promote more concurrency in the kernel without violating any integrity constraints of the operating system.

This chapter introduces many types of synchronization problems such as the mutual exclusion, the producer-consumer, the readers-writers, the dining philosophers, and the sleeping barber, and solutions to these problems using various synchronization tools such as semaphore, critical

the mapping information into blocks of contiguous logical addresses. Entries in MapTables store information pertaining to blocks of addresses instead of individual addresses. The schemes, called segmentation and paging, are discussed in the next two sections. Following the two sections, we present paged segmentation, which is a mix of the two address translation schemes.

8.5 Segmentation

Programmers view an application as a collection of related program units. Each unit contains closely related functions or data, and may obtain services from other units. The units are a higher-level way of structuring an address space. The units are independent as far as compilation is concerned, but are integrated at link time, load time, or runtime. The units may vary in their size (see Fig. 8.10). The figure shows a typical compiler utility. It has four units: main, parser, optimizer, and code generator. In addition, it may have one- or more data units, heap, and stack.

8.5.1 Segment and Segment Block

A *segment* is merely a logical grouping of related information: code or data. In segmentation, a logical address space is organized into a collection of disjoint segments, where the segments are variable size blocks of contiguous logical addresses, (see Fig. 8.11). One may imagine each segment is a sub-logical address space embedded in a full-blown logical address space. Segments are relocatable objects. Entities within a segment are referenced by their relative addresses with respect to the segment base that is assumed to be 0. Each segment then is a linear array of relative addresses and has a limit. Relative addresses in a segment are always less than the value of the segment limit.

Application programmers see a two-dimensional view of the program identifier space: a collection of variable size segments. At runtime, a segment may grow or shrink. Programmers do not concern themselves with the order segments are placed in the logical address space or the distribution of the segments in the memory address space. A unique name, identifier, or number called *segment number* identifies a segment in a logical address space. Any segment knows another segment in the address space only by the latter's segment number. One way to reference an entry in a segment is to

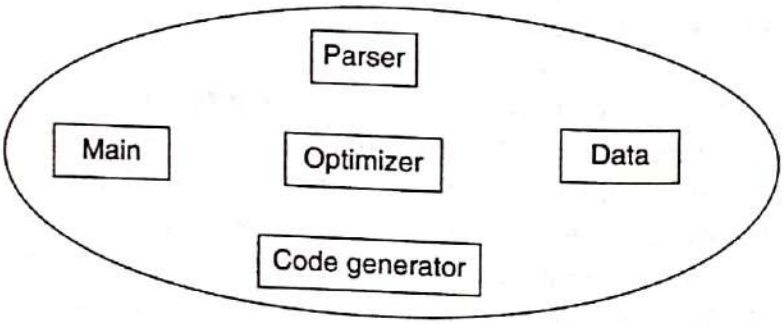

Figure 8.10: *Application as a collection of variable-sized program units.*

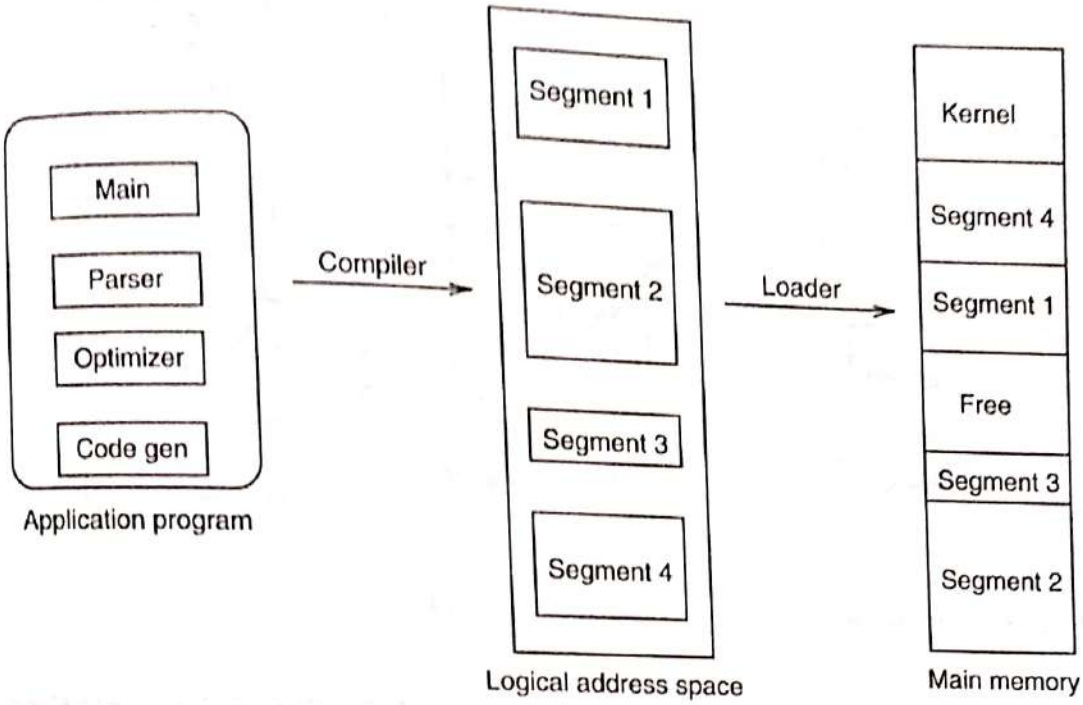

Figure 8.11: *Address binding in segmentation.*

provide a pair $\langle s, o \rangle$, where s is a segment number, and o is a location name within the segment, usually an offset from the beginning of the segment. Offsets are called *segment relative addresses*. Relative addresses in a segment are linearly ordered in the segment, as shown in Fig. 8.12.

Segment number assignment is usually not done by application programmers and definitely not by the operating system, but by compiler and/or loader. Compilers (may be with hints from programmers) do the splitting of an application program into segments and transforming program identifiers to segment relative addresses. (Thereby, applications produce two-dimensional addresses in the segmentation scheme.) Every collection of data or instructions worth assigning a distinct name, a distinct scope of existence, or a distinct protection is placed in a distinct segment. The operating system allocates memory to segments when they are loaded in the main memory. The loader allocates segment numbers while loading segment units unless the compiler or linker has not already allocated segment numbers.

A segment (in entirety) can be placed anywhere in the main memory contiguously starting from a physical address called the segment *base address*, (see Fig. 8.12). In other words, the segment content is linearly mapped into the main memory starting from the base address. The operating system assigns the segment base address when the segment is loaded or relocated in the main memory. The segment also has a *limiting length* that indicates its current size. Addresses for all the identifiers in the segment must be within the segment limit. The region in the main memory that stores a segment is called a *segment block*. (Addresses in a segment and its segment block are in one-to-one correspondence and order preserving.) The operating system makes certain that blocks belonging to two different segments do not overlap in the main memory.

≫ Segments are logical entities and segment blocks are physical entities.

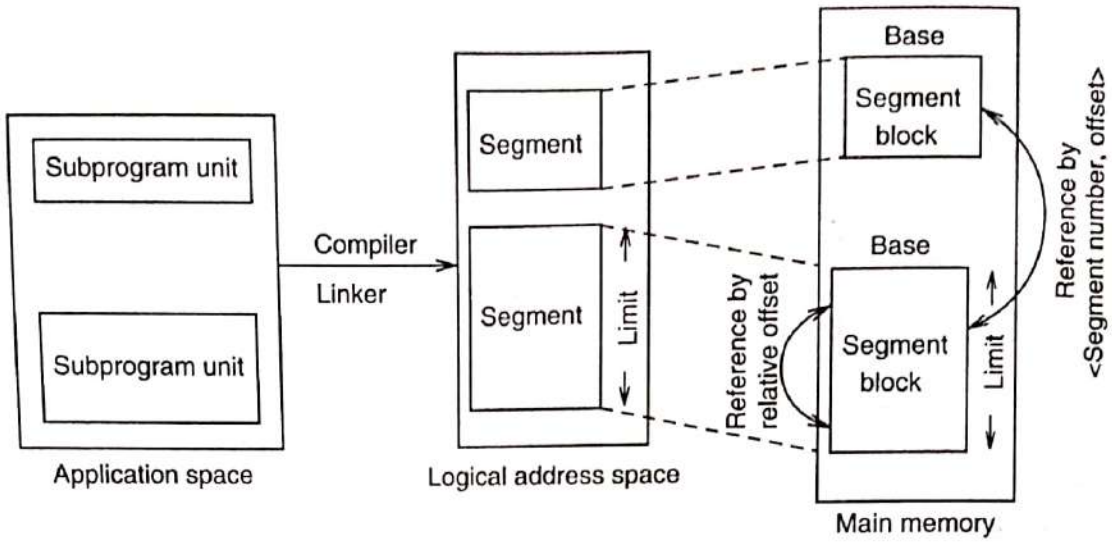

Figure 8.12: *Relationship between segments and segment blocks.*

8.5.2 Segment Table and Address Translation

The system needs to translate all logical addresses that reference a segment to physical memory addresses where the corresponding segment block resides at runtime. This method of address translation is called *segmentation*. As the content of each segment is linearly mapped in a single segment block, the need to store the one-to-one address translation information for all relative addresses in the segment does not arise. An association between a segment and the base address of the segment block is enough. The binding of relative addresses to physical addresses is done at runtime using very simple arithmetic.

Figure 8.13 presents a schematic of address translation in the segmentation scheme. To make address translation and dynamic relocation of segments practical, segment related information of a process is stored in a

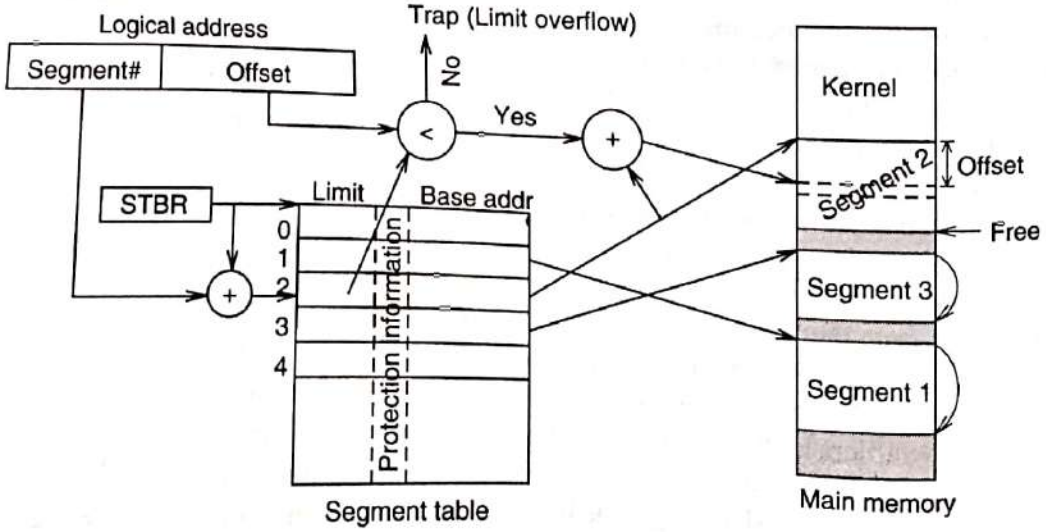

Figure 8.13: *Address translation in segmentation.*

single table called the *segment table*. Each process has its own segment table. The process does not need the table for its computation. It is used solely by the address translation system, and hence is stored as a part of kernel data. The physical address of the segment table base of the running process is stored in a processor register called the *segment table base register* (STBR).

The entries of the segment table are called *segment descriptors*. Each descriptor uniquely identifies a single segment block. A segment descriptor minimally contains the segment's base physical address and the segment limit information, and may also contain segment protection information. A segment table is an array of segment descriptors. For each segment in the logical address space, there is a descriptor in the segment table. To obtain the corresponding segment descriptor, a segment identifier or number is used as an index into the descriptor array.

As shown in Fig. 8.13, each logical address has two components: a *segment number* and an *offset* in the segment. The segment number identifies a segment, and the offset denotes the distance from the start of the segment to the actual address. Mapping from a two dimensional logical address space to the one dimensional memory address space is performed in two steps. In the first step, the segment number is used as an index into the segment table to obtain the segment descriptor that contains the segment's physical base address and limit value. The offset in the logical address is tested against the limit value. If the offset is not less than the limit value, the address translation hardware circuit generates an illegal address violation exception. In the second step, the offset is added to the base address to obtain the actual physical memory address. As different processes may have different numbers of segments in their address spaces, we can associate a length with each segment table. The length is stored in a segment table length register. If the segment number is greater than the limit, the hardware translation circuit raises an illegal address exception.

Let us study a very simple example here. Suppose a process consists of five segments; the second segment is 0x100 bytes long, and it is placed starting at physical address 0x0555. If the process generates a logical address ⟨ 0x2, 0x9 ⟩, the hardware will map it to the physical memory address 0x55E. If the process produces a logical address ⟨ 0x2, 0x10F ⟩, it will cause an address violation exception.

A point to note here is that each segment must be linearly mapped in the physical memory. Some segments (for example, a stack) may grow or shrink at runtime. Consequently, they may need relocation at runtime. When a segment size changes, the operating system updates the corresponding limit value in the segment descriptor. If a segment is relocated to a different memory block, the operating system copies the entire segment in the new block, and at the same time updates the corresponding segment base physical address in the segment descriptor. Such relocation does not involve cooking, that is, fixing of addresses in the code or data.

The system also maintains a symbol table per process, and the table contains information about the map of segment names to segment numbers. The loader/linkage editor uses the table to fix unresolved addresses.

8.5.3 Address Translation Overhead and Its Remedy

Each process has its own segment table, and the table is stored linearly in the main memory as a part of kernel data. When a process is scheduled for execution, the STBR is reinitialized as part of the process context switch-in action. As shown in Fig. 8.13, each memory reference through a logical address incurs two physical memory references: one in the segment table (via the STBR), and the other in the segment block being referenced. Thus, a segmentation scheme slows down the speed of a program execution by a factor of two. If segmentation is to be useful in practice such slowdowns have to be avoided.

The simplest case is that each logical address space is a single segment. We can then have one processor register containing the segment descriptor information of the process currently running. The scheme degenerates to the one presented in Fig. 2.8(b) on page 59, and the programmers' two-dimensional view of applications presented in Figs. 8.10 and 8.11 cannot be supported. In practice, many systems support multi-segment processes. The processor hardware implements an *internal* associative memory of limited size to hold parts of the segment table to speed up address translation. The associative memory accelerates address mapping by bypassing the segment table on repeated accesses to the same segment. Each entry in the associative memory holds a pair $\langle s, d \rangle$, where s is a segment number and d its segment descriptor. When the CPU generates a logical address $\langle s, o \rangle$, s is used as a key to search the associative memory. The search scans all associative memory entities in parallel. If the search succeeds, the descriptor information is obtained from the associative memory without a physical memory reference. Otherwise, the segment descriptor is not in the associative memory, and it is obtained from the segment table, of course by means of a physical memory reference. When there is a context switch, the operating system invalidates the associative memory content. It is found by empirical studies that an associative memory that can hold 8- to 16 segment descriptors is sufficient for well-structured programs to run at nearly their full speed. More about associative memory is discussed in Chapter 14.

≫ Intel X 386 processor has six segmentation register for speeding up the address translation procedure.

8.5.4 Memory Allocation

Segments are units of memory allocation and deallocation. There are three factors that the memory manager needs to consider: (1) allocation and deallocation requests for segments come at unpredictable times, (2) segments have different lengths, (3) segments are placed in blocks of contiguous cells in the main memory (i.e., a segment cannot be partitioned and distributed in the main memory).

In static segmentation systems, many segment blocks (in the main memory) of different sizes are created at bootstrapping time, and they remain fixed. At any given time, an entire segment block is either occupied or free. Management of segment blocks is relatively easy. A segment request can be satisfied by any free block that is large enough to hold the segment; the entire block is allocated. It is not a very effective memory utilization scheme. Most practical system use

dynamic segments blocks: Segment blocks are carved out of free space on demand basis. The main memory is dynamically partitioned into allocated- and free blocks. Free blocks are popularly called *holes*. The memory manager needs to have the complete information pertaining to these blocks. The allocated blocks are parts of programs and processes. The memory manager maintains a list for all holes. Initially, there is a single hole containing the entire available memory. For a new segment-block allocation request, the memory manager runs through the free list, and finds a hole that can hold the segment. It allocates the required amount of memory to the request from the hole, and the remaining part is left as a smaller hole. When an allocated segment-block is deallocated, a new hole is inserted into the list. If its neighbouring memory blocks are free, they are all combined together to form a bigger hole.

There are various strategies in finding a suitable hole for a segment allocation request. Some well known strategies are the following.

- **First fit:** A list of holes is maintained. The holes are placed there in the order of their initial addresses or at the front/tail of the list (or randomly ordered). A segment allocation request runs down the list and is satisfied from the first hole that is large enough to hold the segment.
- **Best fit:** A list of holes is maintained. The holes are placed there in the order of their increasing size. A segment allocation request is satisfied from the smallest hole large enough to hold the segment.
- **Quick fit:** It is based on the observation that most frequent memory requests involve small, but fixed amount of storage. Thus, the holes of most commonly requested sizes are listed separately and served quickly when requested. In some sense, this scheme tries to achieve the benefits of the above two schemes.
- **Buddy system:** Another widely used memory allocation scheme in segmentation is the *buddy system*, where the allocation and deallocation of memory is always in the order of a power of 2. A request for a segment allocation is rounded to the nearest power of 2 that is greater than or equal to the requested amount. The memory manager maintains n, $n \geq 1$, lists of holes. List$_i$, $i = 0, \ldots, n-1$, holds all holes of size 2^i. A hole may be removed from List$_i$, and split into two holes of size 2^{i-1} (called 'buddies' of size 2^i, see Fig. 8.14), and the two holes are entered in List$_{i-1}$. Conversely, a pair of buddies of size 2^i may be removed from List$_i$, coalesced into a single larger hole, and the new hole of size 2^{i+1} is entered in List$_{i+1}$. To allocate a hole of size 2^i, the search is started at List$_i$. If the list is not empty, a hole from the list is allocated. Otherwise, get a hole of size 2^{i+1} from List$_{i+1}$, split the hole into two, put one in List$_i$, and allocate the other one. Hole deallocation is done in reverse fashion: to free a hole of size 2^i, put it in List$_i$; if its buddy is already there, remove both, coalesce, and insert the coalesced hole in List$_{i+1}$. This insertion may cause coalescing of two buddies, their removal from List$_{i+1}$, and a new insertion in List$_{i+2}$, etc.

≫ Although one would expect that the best-fit scheme would improve memory utilization, but, in practice, the first fit leads to better utilization. If the search for the first hole always starts at one end of the free list, small holes concentrate at that end of the list. Consequently, it would increase the search time. In practice, the hole-list is implemented as a circular list. Each time the starting point in the circular list is advanced to the point where the previous search stopped so that small holes are scattered throughout the free list. This is also known as the "next fit" strategy.

≫ The Solaris slab allocator is an implementation of the quick fit memory allocation policy.

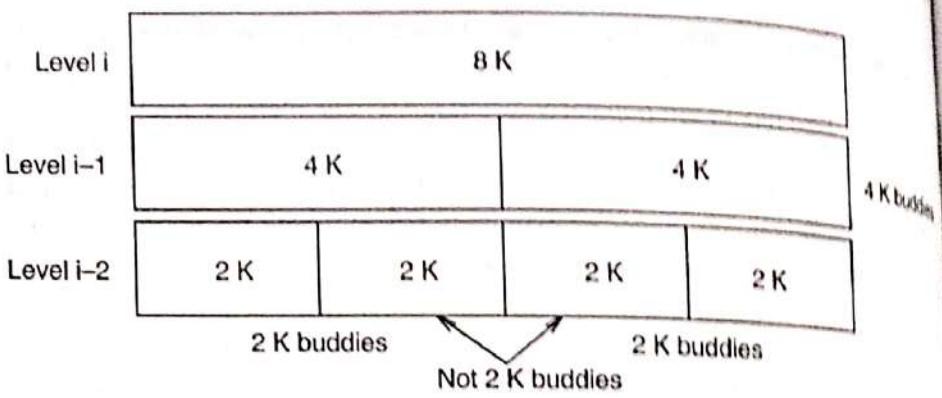

Figure 8.14: *Configuration of a buddy system at three levels.*

8.5.5 Memory Fragmentation and Space Overhead

Continuous allocations and deallocations of segments (of variable size) may lead to many holes scattered throughout the main memory irrespective of any space management scheme used by the memory manager. The holes collectively occupy a substantial amount of free memory. This may lead to *external fragmentation*, where smaller holes individually are not good enough to satisfy any of the pending segment allocation requests, but are collectively able to satisfy some requests. If the memory manager uses the buddy system, we may have *internal fragmentation*, where some allocated space is not needed by requesters. In addition, we need to store a segment table for each process, causing additional memory overhead.

≫ It has been shown that in steady state (equilibrium condition) if there are on the average n segment then there are approximately $n/2$ holes.

Fragmentation is a serious problem as it reduces memory utilization, and may cause serious performance degradation. In the worst case, we have a hole between two allocated blocks. To overcome the external fragmentation problem, the operating system periodically runs through some compaction algorithms to make larger holes from smaller holes. Effect of compaction is shown in Fig. 8.15, where all holes are transferred to one end of the main memory. Note that compaction involves relocating previously allocated

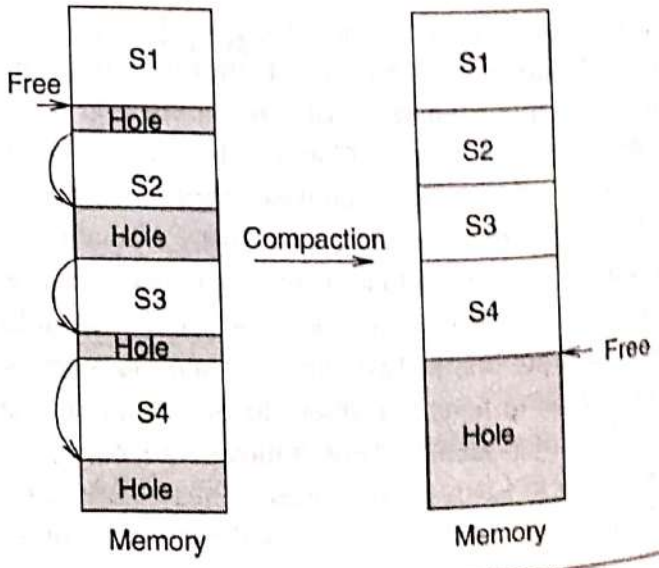

Figure 8.15: *Memory compaction.*

segment blocks in different parts of the main memory. Compaction is an expensive operation as a large number of segments needs transfer, and descriptors updated in segment tables accordingly. Consequently, if there is not much scattered free space, compaction is marginally beneficial.

8.5.6 Segment Relocation

Segments are units of relocation—either the entire segment is relocated or not. Segment relocation is simple—copy the original segment block content, cell by cell, to a new segment block, and then update the corresponding segment descriptor information reflecting the new base address for the segment. The segment limit value is not modified; however, if a segment expands or shrinks, the limit value is changed. Descriptor updates may be needed in all segment tables that refer to the original segment block, (see Section 8.5.8).

8.5.7 Memory Protection

We can associate with each segment a set of access privileges (read, write, execute) information to enforce different rights on the segment. As shown in Fig. 8.13 on page 224, this information in stored in the segment descriptors. Read access permits a process to copy data out of the segment; write access permits to overwrite a data value in the segment; execute access permits to fetch instructions from the segment. For every reference the CPU makes to a segment, the processor hardware checks the protection bits against the type of reference. If the address translation hardware detects a protection violation, it generates a protection violation exception.

≫ A segment block shared by many processes can have different access privileges for different processes.

8.5.8 Memory Sharing

Segments are units of memory shared by multiple processes—either an entire segment is shared or nothing from the segment is shared. In modern computers, compilers produce reentrant codes. Thus, many processes can share code segments. In fact, the operating system takes the initiative to set up code sharing without involving application processes. No processes have write permission on code segments. Normally, data segments are private to processes. Nevertheless, as discussed in section 6.4.2 on page 142, processes themselves may share some data segments with the help of the operating system by creating "shared memory" regions, separate region in a separate segment. The segment tables of these processes will contain segment descriptors with the same physical base address and limit value. If a shared segment is relocated at runtime, all the concerned segment descriptors need to be updated.

Although segment sharing improves memory utilization, sometimes sharing may become difficult. Segment sharing may create some complications in setting up address translation information. You may recall that one segment identifies another segment in the same address space by a segment

number. When all processes refer to the same segment block by the same segment number, then there is no complication. The complication arises when different processes refer to the same segment block by different segment numbers. The question is by what mechanisms does one segment reference another segment. We study two examples here to explain some subtle issues in address translation.

- The first example is about self-referencing segments. Examples include recursive function calls and jumps to different addresses within a segment itself. Consider a typical function in a code segment. The function parameters may come from a different segment (including itself), and the function may reference temporary data at another segment (including itself). The recursive function call creates further complications: a parameter may come from this segment or from a different segment (e.g., parameters passed on from the previous call) that we do not know until runtime. We may need to restructure compiled code: we need to copy parameters into a separate region before making a recursive call. What segment number do we use to refer to entities in the same segment that different processes access by different segment numbers? Normally, a segment refers to entities inside itself by their relative addresses; here, a relative address is specified by an offset relative to the program counter register or some other register holding the segment information. An intra-segment address is always this relative address, but a segment number and an offset pair specify an inter-segment address. We need to solve the inter-segment address related problems.

- Let us study another example here. Figure 8.16 depicts a scenario where two processes share two code segments: the main code and the library. They each have a private data segment. As shown in the figure, process P (respectively, Q) refers to the main code segment by segment number 0 (respectively, 2). Suppose that the code segment is located in a memory block at physical address A. Two processes, P and Q, refer to the block by two different segment numbers s and s', respectively. The segment table of process P (respectively, Q) will have at index s (respectively, s') a segment descriptor containing the physical address A. Both the descriptors will have the same segment limit value, but they may have different protection information. As shown in Fig. 8.16, to call a function in the library segment from the main code, process P makes a call to its segment 1, whereas process Q makes the same call from the same place to its segment 0. How can we write instructions in the shared main code such that executions of the instructions correctly achieve what P and Q want to do? That is, the same main code should work correctly for both of them. The same call instruction should produce two different logical addresses in two processes. Consequently, we may not be able to statically embed segment numbers in codes. How do we resolve such address

≫ Flexibility for processes to refer to a given shared segment by different segment numbers is a great convenience. However, different segment numbers create substantial address translation problems (both at compile/load time and runtime).

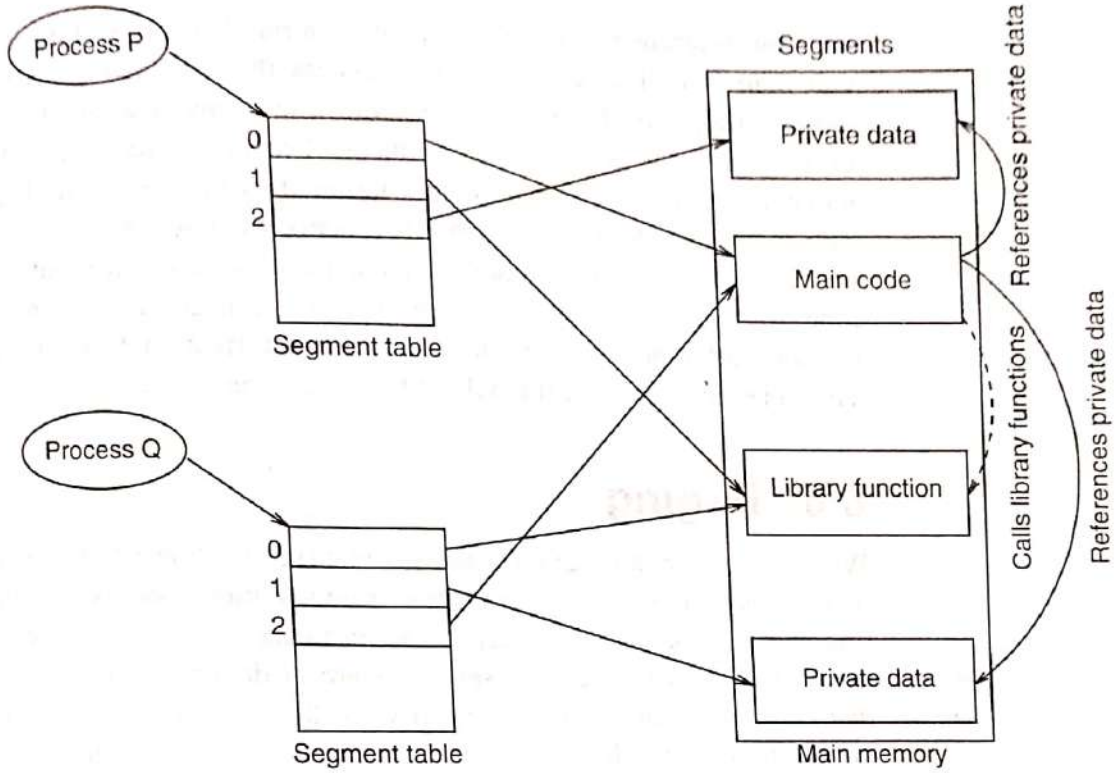

Figure 8.16: *Segment sharing by two processes.*

references made by different processes? It is not possible for the compiler and/or loader to solve this problem themselves. We have to solve the problem at runtime, because we will be able to derive correct segment numbers then. At the time of the invocation of the function, the calling instruction must be able to derive the correct segment number information using some register where the information is held.

There are various solutions to the problem of segment sharing mentioned above. We present two solutions here. Some systems such as Burroughs use uniform address resolution: all processes refer to the same shared segment by the same segment number; also, functionally equivalent (private) segments must have the same segment numbers. This approach may create problems for a compiler/loader to find acceptable segment numbers for all processes. The Burroughs system requires a user to compile all her programs at one stroke so that the compiler can allocate segment numbers at compile time and embed the numbers in the code.

Another solution is to use one linkage table per shared code segment per process. A linkage table contains index to the segment table proper of the process. When a shared code segment generates a segment number (for another shared segment), the number is treated as an index to the corresponding linkage table to obtain another index for the segment table. A linkage table is created when the process executes the shared segment for the first time. The address of a linkage table is held in a CPU register. If a code

segment references another segment by a number, say s, different processes can map s to different segment numbers through their respective linkage tables. Addresses for parameters from one segment to another are treated differently from the addresses embedded in the shared code. The segment numbers are passed as parameters rather than indexes to the linkage table. This is because the segment table is process-wide, whereas linkage tables are not. This solution requires additional space to store linkage tables, additional time to setup indirection information in linkage tables, and additional memory references to obtain indirection information from linkage tables. This scheme is used in the MULTICS operating system.

8.6 Paging

We have seen in the segmentation scheme (see Section 8.5) that management of free space is expensive. External fragmentation reduces memory utilization and may become a source of performance bottlenecks. The fragmentation problem occurs because segments are of different lengths, and some of them can grow and shrink arbitrarily. In this section, we study a radically different memory allocation scheme called *paging* where memory allocation is done in fixed size units. Paging is a widely used memory management scheme in modern operating systems because its space management is quite simple and there is no external fragmentation.

≫ The page number and offset are transparent to the processes.

8.6.1 Page and Frame

A logical address space is partitioned into uniform units called pages. A *page* is a fixed length compact interval of contiguous logical addresses. As we will see shortly, to make address translation easier and memory allocation efficient, the page size is always chosen as a power of 2. A logical address space becomes a linear array of pages (see Fig. 8.17). Unlike segmentation, however, the entire logical address space is treated as a single dimensional space. Any two consecutive addresses in an address space refer to either the same page or two consecutive pages.

A unique name or number called a *page number* identifies each page in an address space. Page numbers are assigned by the operating system (actually, by the processor), and not by the compiler/linker nor by application programmers. A program references each entry in a process address space by providing a logical address l. (In contrast to the segmentation scheme, here applications produce single-dimensional addresses.) As shown in Fig. 8.18, the system treats the logical address as a pair $\langle p, o \rangle$, where p is a page number, and o is a location name within the page, usually an offset from the beginning of the page. For example, consider a logical address 0x0001FF0. If the processor supports 2^{12}-byte pages, then the page number is 1 and offset 0xFF0. If the processor supports 2^{10}-byte pages, then the page number is 7 and offset 0x2F0.

The CPU always generates logical addresses. A processor address translation unit, called the *paging unit*, translates logical addresses into physical

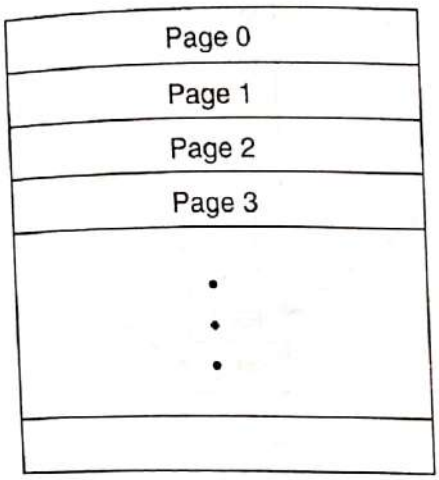

Figure 8.17: *Logical address space as a linear array of pages.*

addresses. The paging unit translates two consecutive logical addresses (that refer to the same page) into two consecutive physical addresses in the same order. That is, each page is linearly mapped into a compact physical address interval block in the main memory. The physical memory is also partitioned into the same size units, called *memory frames* or physical pages or page frames or simply *frames*. That is, frames are fixed-size memory units of contiguous physical addresses. A frame is identified by its *frame address*, that is, the physical address of the first byte in the frame. Addresses in a frame and in a page are in one-to-one correspondence and order preserving.

The entire main memory is partitioned into fixed-sized frames. Frames are considered equivalent resources in the sense that any available frame can satisfy a request for a new frame. A page in a logical address space can be placed in any frame. That is, two consecutive pages in an address space need not be placed in two consecutive frames (see Fig. 8.19); in fact they can be arbitrarily scattered in the main memory.

≫ Please note the difference between these two terminologies, namely page and frame. They refer to the same size compact address intervals. The page is a logical entity, and resides in logical address spaces. The frame is a physical entity, and resides in the memory address space. However, some authors use both terminologies interchangeably.

8.6.2 Page Table and Address Translation

As the content of a page is linearly mapped onto a single frame and the page and the frame are of the same size, we do not have to store the one-to-one address translation information for all addresses falling in the page. We simply make an association between a page and the address of the frame holding the page. The binding of addresses (that refer to the page) to their physical addresses is done using very simple arithmetic at runtime.

Figure 8.20 presents a schematic of address translation in the paging scheme. It is another kind of extension (generalization) to the scheme

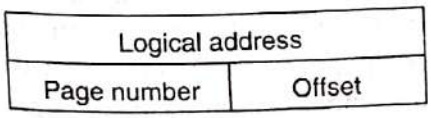

Figure 8.18: *Interpreting a logical address in paging.*

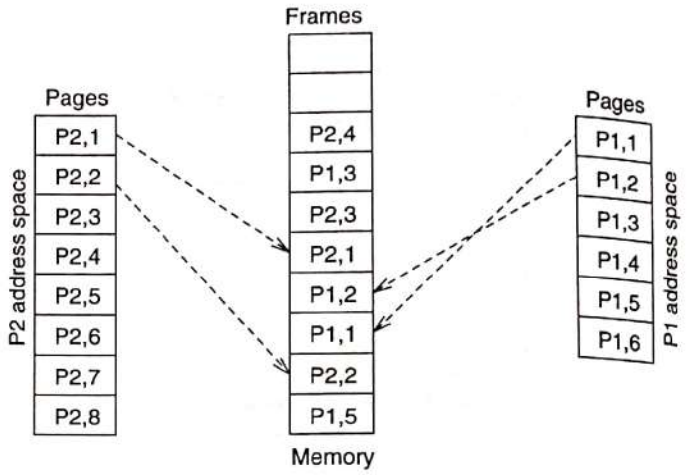

Figure 8.19: *A typical placement of pages in memory frames.*

≫ The CPU is not aware that a process address space is partitioned and scattered in the main memory. The CPU accesses the individual entities in the address space by their logical addresses. The paging unit in the processor maps these logical addresses to appropriate physical addresses at runtime. Although a process address space is scattered throughout the main memory, the CPU sees a contiguous (logical) address space.

presented in Fig. 2.8(b) on page 59. Under the paging scheme, a page may be stored in any frame. The information relating to the location of all the pages of an address space is stored in the main memory in a data structure called the *page table*. The operating system keeps separate page tables for separate processes. The physical address of the page table of the running process is stored in a processor register called the *page table base register* (PTBR). The entries of a page table are called *page descriptors*. A page table is a linear array of page descriptors. A page descriptor minimally contains the physical address of the frame where the page resides in the main memory. The descriptor may also contain some protection information. For each page in the logical address space, there is precisely one entry in the page table.

The page number is used as an index into the page table to obtain the page descriptor containing the corresponding frame address. The offset is added to the frame address to obtain the actual physical address. As mentioned previously in section 8.6.1, the page size is always a power of 2, say 2^i, $i > 0$. Thereby, the offset is always i-bit long, spanning 0,…,0 to 1,…,1. We do not need the lower i bits of the frame address: these bits are always 0. The i-bit offset is actually concatenated with the higher order bits of the frame address instead of doing an addition, see Fig. 8.21. When a page is loaded into a frame, the higher order bits of the frame address are stored in the page descriptor.

8.6.3 Address Translation Overhead and Its Remedy

≫ In many operating systems, the kernel has its own page table.

Each process has its own page table, and the table is stored linearly in the main memory as a part of kernel data. When a process is scheduled for execution, the PTBR is reinitialized as a part of the process context switch-in action. It is shown in Fig. 8.20 that every logical address referenced by the CPU incurs two physical memory references by the processor: one to access the page descriptor from the process page table (via PTBR), and another into the page frame to get the actual content. Thus, the program execution speed

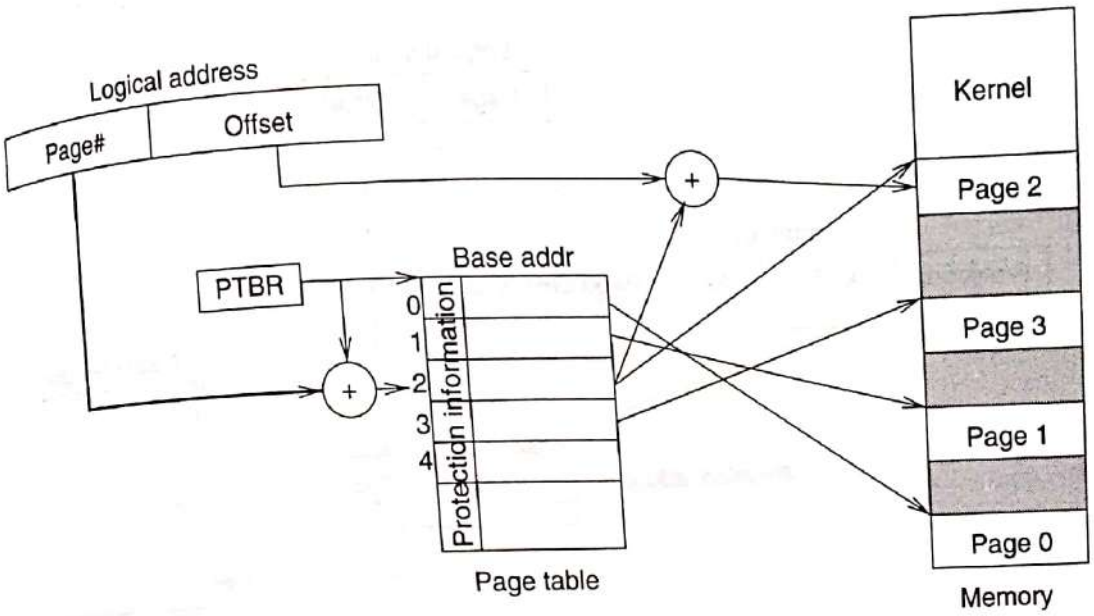

Figure 8.20: *Address translation in paging.*

slows down by a factor of two. Avoiding slowdowns is essential to make paging useful in practice.

Modern processor architectures implement a small but fast associative memory in the processor, called the *translation look-aside buffer* (TLB, in short) to hold parts of the page table of the currently running process. The TLB helps fast translation of logical addresses to physical addresses. The TLB stores pairs $\langle p, d \rangle$, where p is a page number and d its page descriptor. When the CPU generates a logical address $\langle p, o \rangle$, the page number p is used as a key to search the TLB. If there is a match, the page descriptor is obtained from the TLB without a physical memory reference. If there is no match, the page descriptor information is obtained from the original page table kept in the main memory via the PTBR. In this case, a new pair containing the page number and the page descriptor information is inserted into the TLB. When there is a context switch, the operating system invalidates the TLB. (Management schemes for TLBs are discussed in Chapter 14.) Experience and experimental (empirical) studies seem to suggest that 16- to 32 TLB registers are sufficient to cause a well-structured program to run near its full speed.

≫ In Intel 80386, the TLB is a 4-way set-associative cache: eight sets of four registers, a total of 32 TLB registers.

8.6.4 Memory Allocation

Frames are units of memory allocation; fractions of a frame are never allocated. Memory allocation is straightforward: any page can be placed in any available frame unless there are other restrictions. There is no need for frame compaction.

The operating system has to keep track of all frames and their status: which frames are allocated, and which ones are free. The system maintains a data structure for this purpose called the *frame table*. The table has one entry

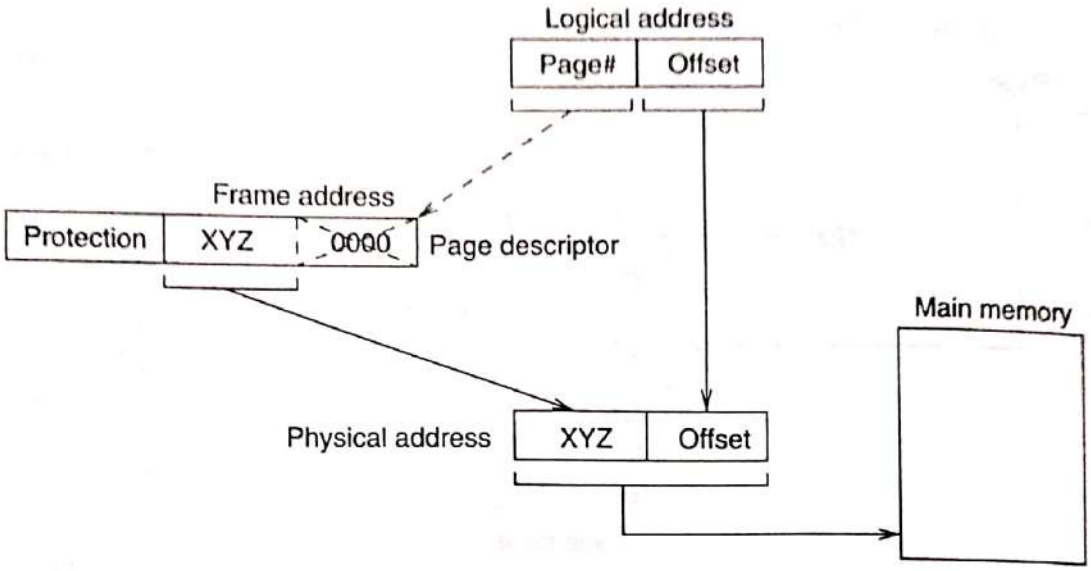

Figure 8.21: *A simpler address translation in paging.*

for each frame. The value of the entry provides the information on the frame status. When a fresh request for frame allocation comes, the memory manager searches the frame table, and finds an available frame. Any free frame will satisfy the request.

Although page allocation is simple, paging creates some problem as far as page tables are concerned. Modern processors support a very large logical address space (2^{32} to 2^{64} bytes). The typical page size varies between 2^9 and 2^{13} bytes. Consequently, page tables are very large. For example, for a page size of 1024 ($= 2^{10}$) bytes, we may end up having as many as 2^{22} to 2^{54} entries in a page table. When the size of the page table becomes massive, allocating contiguous physical memory for page tables becomes a problem by itself. In such a case, page tables themselves are paged, and what we get is a multilevel paging system. Figure 8.22 presents a schematic of a two-level page translation. The PTBR points to the 1st level page table (often called page directory). Each entry in the page directory has its own 2nd level page table. Each entry in this page table stores a frame address. We will discuss more about multilevel paging schemes in Section "Paging" of Chapter 17.

8.6.5 Memory Fragmentation and Space Overhead

In paging, there is no external fragmentation. As long as there is an available frame, it can be used to satisfy a frame allocation request. However, as frames are units of allocation and deallocation, address spaces are always rounded off to an integral number of pages. Thus, a process may not require the entire frame where the last page is mapped. (Unlike in Fig. 4.4. on page 91, here we are assuming that there is no gap inside a process address space.) Part of this frame remains unutilized. This unused, though allocated, space is called *internal fragmentation*. On an average, half of a frame per address space is

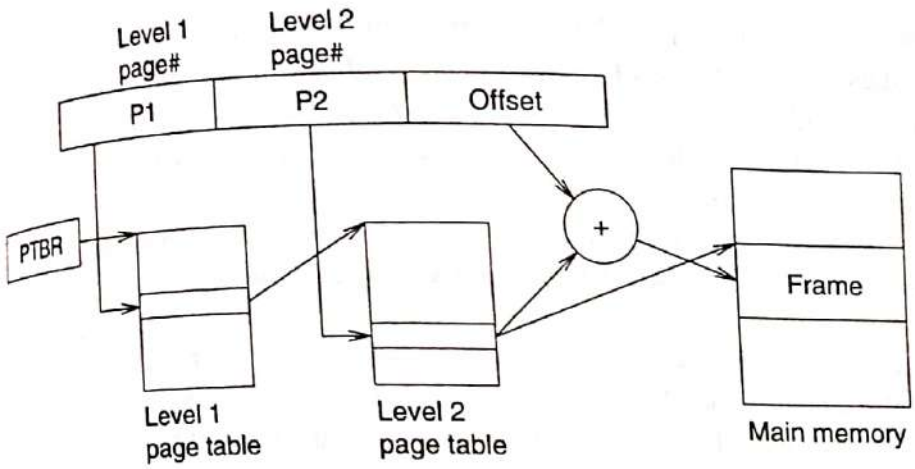

Figure 8.22: *Two-level address translation in paging.*

internally fragmented. In addition, we need to store a page table for each process, causing additional memory overhead. A smaller page size reduces internal fragmentation but increases the page table size. A larger page size has the opposite effect. A system needs to find the optimum page size. Most modern systems use pages of 512, 1024, 2048, or 4096 bytes in length. The current trend is to use larger page sizes. The page size is determined by the processor architecture, and not by the operating system. The operating system is forced to use the same page size.

Internal fragmentation may cause another subtle problem. By definition, an application process must not be able to reference entities outside its address space. The operating system usually does nothing if a process references entities in the internally fragmented memory that, theoretically, lie outside the process address space. Applications that appear normal in one system may show up as address violation exceptions in another system.

≫ If many processes share a frame, their corresponding pages may have different access rights.

8.6.6 Memory Relocation

The frame is the unit of relocation, and the frame relocation is simple. Copy the original page-frame content, cell by cell, to a new frame, and then update the corresponding page descriptor information reflecting the new frame address for the original page. We may need to do page descriptor update in all page tables that refer to the original frame. (See Section 8.6.8.)

8.6.7 Memory Protection

We can associate with each page a set of access privileges (read, write, execute) information to enforce different access rights on the page. This information is stored in the page descriptor. For every reference the CPU makes to a page, the processor hardware checks the page protection bits against the type of reference. If the paging unit detects a protection violation, the unit generates a protection violation exception. In addition, there is a separate bit, called the valid/invalid, that shows whether the page belongs to the process address space. (You may recall from Fig. 4.4 on page 91 that a

process may not span the entire possible maximum address space.) The bit values are set to valid for those pages that belong to the process address space, and set to invalid for other pages. If the process attempts to access entities from invalid pages, the hardware translation unit raises an illegal address exception.

Very often process address spaces do not span the maximum limit supported by the processor. Consequently, it is a waste of memory if we allocate space for full blown, that is, page tables of the maximum possible size. To avoid this, the operating system may associate a limit on the size of a page table. The limit value is loaded in another processor register, called page table length register (PTLR). Every page number is checked against the PTLR value. If the page number crosses the limit, the hardware raises an illegal address exception.

8.6.8 Memory Sharing

Memory sharing in paging is quite simple. Frames are units of memory shared by multiple processes. A frame to be shared among a set of processes is reflected in their respective page tables. Figure 8.23 shows how two processes share the same frame f. Note that the same frame may be mapped at two different logical pages by two different processes.

Modern compilers produce a reentrant code, that is, the code never changes itself during its execution. Many processes can share the code. The operating system takes the initiative to set up code sharing among processes without involving applications. However, processes have different data pages. Thus, one process does not affect the others sharing the same code. But, address sharing leads to the same kind of complications we have seen in Section 8.5.8 on page 229 for segmentation.

8.7 Paged Segmentation

≫ The Intel 80386 processor supports paged segmentation.

In pure paging, each logical address space is considered a single segment that consists of many fixed size pages. Paging does not alter the view of linearity of address spaces; it does not achieve the objectives that motivated segmentation. On the other hand, segmentation requires variable-sized segments of contiguous physical addresses. Loading a segment requires finding a free memory block large enough to contain the entire segment, whereas loading a page requires finding a free frame, and any free frame will do. The former problem is much more difficult than the latter. Application programmers preferred segmentation, while system implementers preferred paging. For example, application processes have arbitrary sized memory regions, different regions in different segments. External fragmentation is a serious problem in segmentation. We would like to have schemes that combine the benefits of both segmentation and paging, but without some of their problems. In this section, we present a mixed address translation scheme, namely paged segmentation.

As in the pure segmentation scheme, an address space in the paged segmentation scheme consists of variable-sized segments. However, unlike in

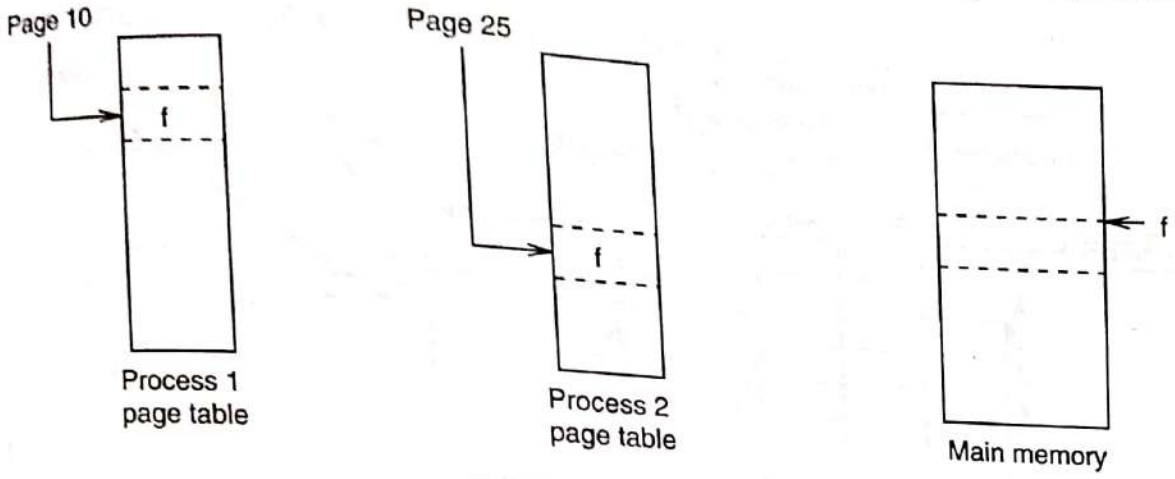

Figure 8.23: *Frame sharing in paging.*

the pure segmentation scheme, a segment may not be allocated to a single memory block. Instead, the segment is paged, and the pages are distributed to memory frames. This avoids both, external fragmentation and time overhead in searching for memory to allocate a segment.

8.7.1 Address Translation

Figure 8.24 presents the address translation mechanism under the paged segmentation scheme. Each segment is treated as a single, small linear address space within the container logical address space proper, and this mini address space is paged. The address translation in the mini address space is done following the pure paging scheme. A page table is associated with the mini address space, which helps in the address translation task.

As in the pure segmentation scheme, segment related information is structured into segment descriptors that are stored in a segment table. Unlike pure segmentation, each segment descriptor contains the physical base address of the segment's page table. The descriptor also has the segment limit and some protection information. For every process, there is a single segment table and one page table for each segment in the address space. The physical address of the segment table base of the running process is stored in the STBR register. When a process is scheduled for execution, the STBR is reinitialized as a part of the process context switch-in action.

Like pure segmentation, a logical address is partitioned into two components $\langle s, r \rangle$, where s is a segment number, and r is a relative address within the segment. The segment number s is used as an index into the segment table to obtain the base of the segment's page table, and the relative address r is checked against the segment limit value. If r is greater than the limit value, the translation hardware circuit generates an illegal address violation exception. Unlike pure segmentation, the relative address r is partitioned into another pair $\langle p, o \rangle$, where p is used as index into the page table to obtain a frame address and o is used as an offset into the frame.

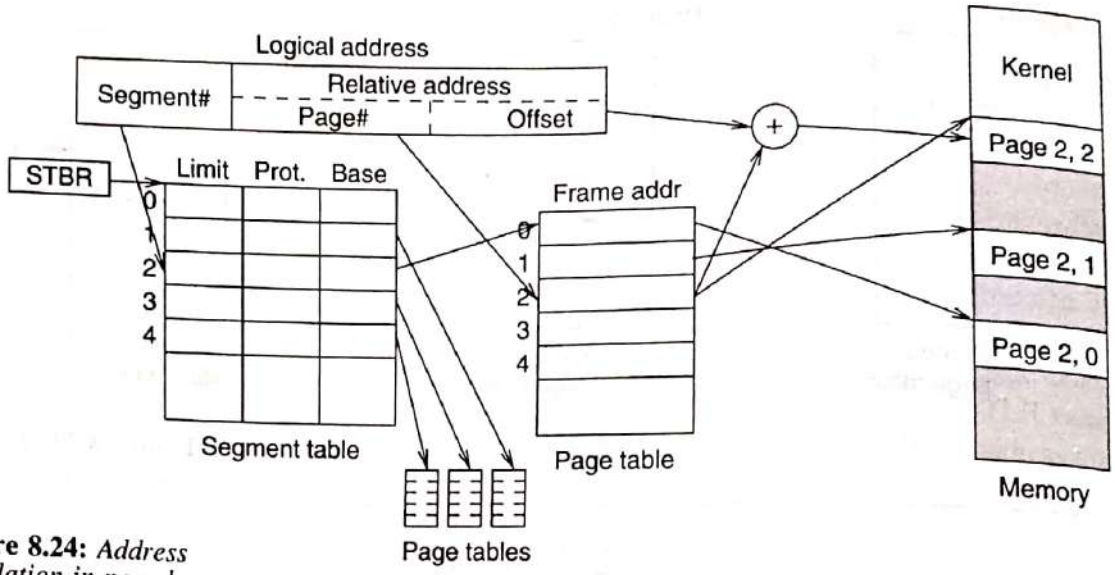

Figure 8.24: *Address translation in paged segmentation.*

8.7.2 Address Translation Overhead and Its Remedy

Each process has its own segment table, and the table is stored in the main memory as a part of the kernel data. As shown in Fig. 8.24, each memory reference through a logical address incurs three physical memory references: (1) the first one into the segment table (via STBR), (2) the second one into the corresponding page table, and (3) the third one into the required page-frame. Thus, a paged segmentation scheme causes the program execution speed slowdown by a factor of three. Like pure segmentation- and pure paging schemes, we can use associative memory and TLB to ameliorate the speed slowdown.

8.7.3 Memory Allocation

Segments are 'logical' units of memory allocation and deallocation, and frames are 'physical' units of memory allocation and deallocation. Segments are carved out in the process address space, and not in the main memory. Under this scheme, it is not very difficult to handle segments of widely different lengths: find the appropriate number of free frames to load a segment in the main memory. Dynamic expansion and contraction of segments at runtime are handled with relative ease without relocating the segments as long as they can grow without the maximum segment-size limit of the processor architecture: simply add- or remove pages to or from the segment.

If a segment table becomes too large to store contiguously in the main memory, we can page the segment table. In this case, we split a segment number into two components: a "segment table page number", and an offset within that page.

8.7.4 Memory Fragmentation and Space Overhead

Each segment has its own page table. Consequently, like pure paging, there is a possibility of internal fragmentation, and on the average, we have half a page of internal fragmentation per segment as all segments are rounded off to an integral number of pages. Though we have eliminated external fragmentation, we have severe internal fragmentation. We also have additional memory overhead to store page tables and segment tables.

8.7.5 Relocation

Page relocation is also quite simple. As in pure paging, copy the original page-frame content, cell by cell, to a new frame, and then update the corresponding page descriptor information reflecting the new base physical address for the frame. We may need to do so in all page tables that refer to the original frame.

Segment relocation, while also simple, is a little more involved than pure segmentation. Copy all the page frames in the segment to new frames and update the page table accordingly. No changes are required for the segment table. As segments are logical objects, they are often not relocated in practice.

8.7.6 Memory Protection

Segments are units of memory protection. We can associate with each segment a set of access privileges (read, write, execute) information to enforce different access rights on the segment. This information in stored in the segment descriptor. For every reference the CPU makes to a segment, the protection bits are checked against the type of reference. If the address translation hardware detects a protection violation, it generates a protection violation exception.

8.7.7 Memory Sharing

Segments are units of memory sharing. Actually, processes can share the segment's page tables to share the segment; there is only one copy of the page table for all processes sharing the segment. However, processes may have different access permissions to shared segments, which is stored in their respective segment tables. As in pure segmentation, different processes may use different segment numbers to identify the same segment.

Summary

The main memory is a hardware device of a limited size. Its space is recycled to hold information required by the processor and the I/O devices. The operating system permanently occupies a part of the memory, and the remaining space is recycled. This space is called the dynamic memory, and it is used to store application code and data, process dynamic data, and kernel dynamic code and data. Memory management implies managing this dynamic memory.