1.3. Operating System Concepts

The interface between the operating system and the user programs is defined by the set of "extended instructions" that the operating system provides. These extended instructions have been traditionally known as **system calls**, although they can be implemented in several ways. To really understand what operating systems do, we must examine this interface closely. The calls available in the interface vary from operating system to operating system (although the underlying concepts tend to be similar).

We are thus forced to make a choice between (1) vague generalities ("operating systems have system calls for reading files") and (2) some specific system ("MINIX 3 has a `read` system call with three parameters: one to specify the file, one to tell where the data are to be put, and one to tell how many bytes to read").

We have chosen the latter approach. It's more work that way, but it gives more insight into what operating systems really do. In Sec. 1.4 we will look closely at the basic system calls present in UNIX (including the various versions of BSD), Linux, and MINIX 3. For simplicity's sake, we will refer only to MINI 3, but the corresponding UNIX and Linux system calls are based on POSIX in most cases. Before we look at the actual system calls, however, it is worth taking a bird's-eye view of MINIX 3, to get a general feel for what an operating system is all about. This overview applies equally well to UNIX and Linux, as mentioned above.

The MINIX 3 system calls fall roughly in two broad categories: those dealing with processes and those dealing with the file system. We will now examine each of these in turn.

1.3.1. Processes

A key concept in MINIX 3, and in all operating systems, is the **process**. A process is basically a program in execution. Associated with each process is its **address space**, a list of memory locations from some minimum (usually 0) to some maximum, which the process can read and write. The address space contains the executable program, the program's data, and its stack. Also associated with each process is some set of registers, including the program counter, stack pointer, and other hardware registers, and all the other information needed to run the program.

We will come back to the process concept in much more detail in Chap. 2, but for the time being, the easiest way to get a good intuitive feel for a process is to think about multiprogramming systems. Periodically, the operating system decides to stop running one process and start running another, for example, because the first one has had more than its share of CPU time in the past second.

When a process is suspended temporarily like this, it must later be restarted in exactly the same state it had when it was stopped. This means that all information about the process must be explicitly saved somewhere during the suspension. For example, the process may have several files open for reading at once. Associated with each of these files is a pointer giving the current position (i.e., the number of the byte or record to be read next). When a process is temporarily suspended, all these pointers must be saved so that a `read` call executed after the process is restarted will read the proper data. In many operating systems, all the information about each process, other than the contents of its own address space, is stored in an operating system table called the **process table**, which is an array (or linked list) of structures, one for each process currently in existence.

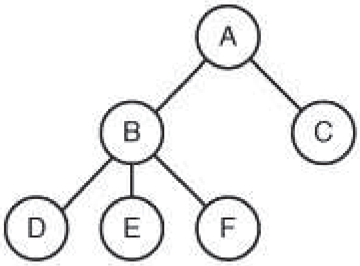
Thus, a (suspended) process consists of its address space, usually called the **core image** (in honor of the magnetic core memories used in days of yore), and its process table entry, which contains its registers, among other things.

The key process management system calls are those dealing with the creation and termination of processes. Consider a typical example. A process called the **command interpreter** or **shell** reads commands from a terminal. The user has just typed a command requesting that a program be compiled. The shell must now create a new process that will run the compiler. When that process has finished the compilation, it executes a system call to terminate itself.

On Windows and other operating systems that have a GUI, (double) clicking on a desktop icon launches a program in much the same way as typing its name at the command prompt. Although we will not discuss GUIs much, they are really simple command interpreters.

If a process can create one or more other processes (usually referred to as **child processes**) and these processes in turn can create child processes, we quickly arrive at the process tree structure of Fig. 1-5. Related processes that are cooperating to get some job done often need to communicate with one another and synchronize their activities. This communication is called **interprocess communication**, and will be addressed in detail in Chap. 2.

Figure 1-5. A process tree. Process *A* created two child processes, *B* and *C*. Process *B* created three child processes, *D*, *E*, and *F*.



Other process system calls are available to request more memory (or release unused memory), wait for a child process to terminate, and overlay its program with a different one.

Occasionally, there is a need to convey information to a running process that is not sitting around waiting for it. For example, a process that is communicating with another process on a different computer does so by sending messages to the remote process over a network. To guard against the possibility that a message or its reply is lost, the sender may request that its own operating system notify it after a specified number of seconds, so that it can retransmit the message if no acknowledgement has been received yet. After setting this timer, the program may continue doing other work.

When the specified number of seconds has elapsed, the operating system sends an **alarm signal** to the process. The signal causes the process to temporarily suspend whatever it was doing, save its registers on the stack, and start running a special signal handling procedure, for example, to retransmit a presumably lost message. When the signal handler is done, the running process is restarted in the state it was in just before the signal. Signals are the software analog of hardware interrupts. They are generated by a variety of causes in addition to timers expiring. Many traps detected by hardware, such as executing an illegal instruction or using an invalid address, are also converted into signals to the guilty process.

Each person authorized to use a MINIX 3 system is assigned a **UID** (User IDentification) by the system administrator. Every process started has the UID of the person who started it. A child process has the same UID as its parent. Users can be members of groups, each of which has a **GID** (Group IDentification).

One UID, called the **superuser** (in UNIX), has special power and may violate many of the protection rules. In large installations, only the system administrator knows the password needed to become superuser, but many of the ordinary users (especially students) devote considerable effort to trying to find flaws in the system that allow them to become superuser without the password.

We will study processes, interprocess communication, and related issues in Chap. 2.

1.3.2. Files

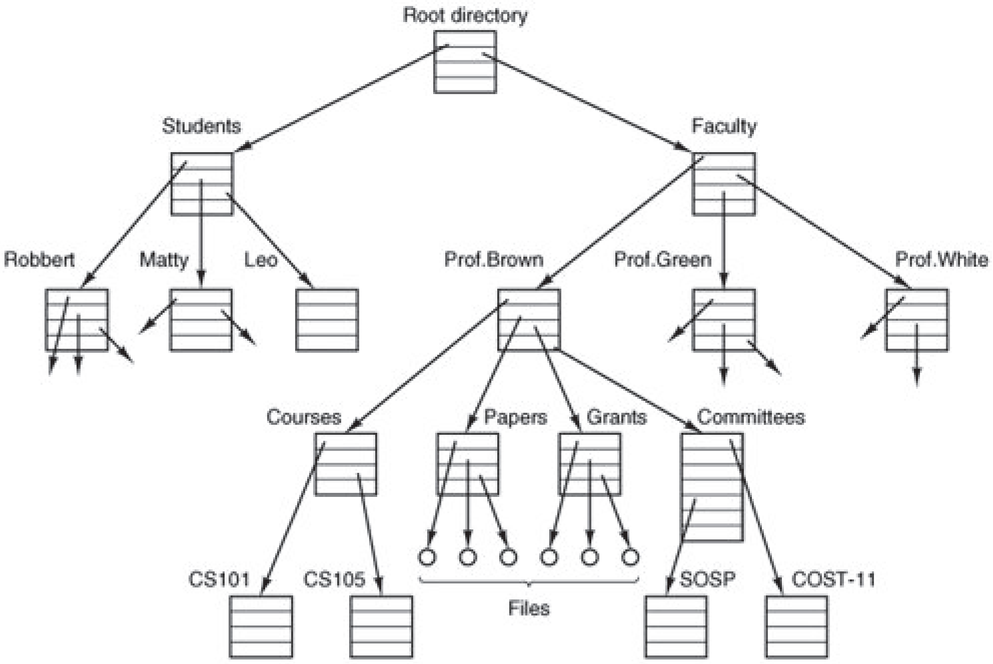
The other broad category of system calls relates to the file system. As noted before, a major function of the operating system is to hide the peculiarities of the disks and other I/O devices and present the programmer with a nice, clean abstract model of device-independent files. System calls are obviously needed to create files, remove files, read files, and write files. Before a file can be read, it must be opened, and after it has been read it should be closed, so calls are provided to do these things.

To provide a place to keep files, MINIX 3 has the concept of a **directory** as a way of grouping files together. A student, for example, might have one directory for each course he is taking (for the programs needed for that course), another directory for his electronic mail, and still another directory for his World Wide Web home page. System calls are then needed to create and remove directories. Calls are also provided to put an existing file into a directory, and to remove a file from a directory. Directory entries may be either files or other directories. This model also gives rise to a hierarchythe file systemas shown in Fig. 1-6.

Figure 1-6. A file system for a university department.

(This item is displayed on page 23 in the print version)

[View full size image]



The process and file hierarchies both are organized as trees, but the similarity stops there. Process hierarchies usually are not very deep (more than three levels is unusual), whereas file hierarchies are commonly four, five, or even more levels deep. Process hierarchies are typically short-lived, generally a few minutes at most, whereas the directory hierarchy may exist for years. Ownership and protection also differ for processes and files. Typically, only a parent process may control or even access a child process, but mechanisms nearly always exist to allow files and directories to be read by a wider group than just the owner.

Every file within the directory hierarchy can be specified by giving its **path name** from the top of the directory hierarchy, the **root directory**. Such absolute path names consist of the list of directories that must be traversed from the root directory to get to the file, with slashes separating the components. In Fig. 1-6, the path for file *CS101* is */Faculty/Prof.Brown/Courses/CS101*. The leading slash indicates that the path is absolute, that is, starting at the root directory. As an aside, in Windows, the backslash (\) character is used as the separator instead of the slash (/) character, so the file path given above would be written as *\Faculty\Prof.Brown\Courses\CS101*. Throughout this book we will use the UNIX convention for paths.

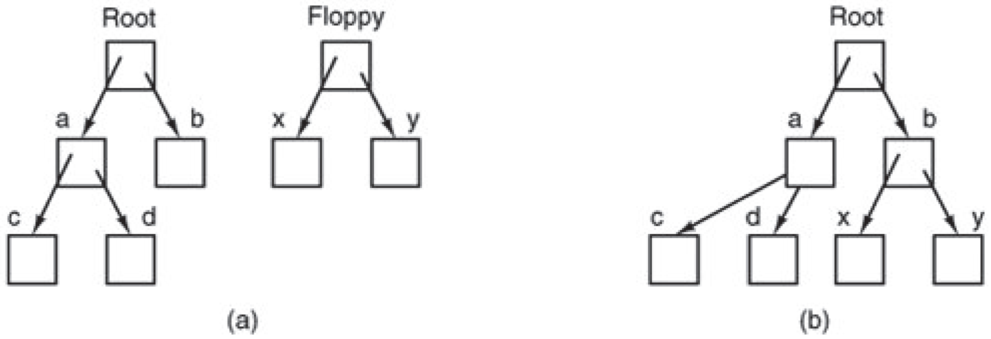
At every instant, each process has a current **working directory**, in which path names not beginning with a slash are looked for. As an example, in Fig. 1-6, if */Faculty/Prof.Brown* were the working directory, then use of the path name *Courses/CS101* would yield the same file as the absolute path name given above. Processes can change their working directory by issuing a system call specifying the new working directory.

Files and directories in MINIX 3 are protected by assigning each one an 11-bit binary protection code. The protection code consists of three 3-bit fields: one for the owner, one for other members of the owner's group (users are divided into groups by the system administrator), one for everyone else, and 2 bits we will discuss later. Each field has a bit for read access, a bit for write access, and a bit for execute access. These 3 bits are known as the **rwx bits**. For example, the protection code *rwxr-x--x* means that the owner can read, write, or execute the file, other group members can read or execute (but not write) the file, and everyone else can execute (but not read or write) the file. For a directory (as opposed to a file), *x* indicates search permission. A dash means that the corresponding permission is absent (the bit is zero).

Before a file can be read or written, it must be opened, at which time the permissions are checked. If access is permitted, the system returns a small integer called a **file descriptor** to use in subsequent operations. If the access is prohibited, an error code (1) is returned.

Another important concept in MINIX 3 is the mounted file system. Nearly all personal computers have one or more CD-ROM drives into which CD-ROMs can be inserted and removed. To provide a clean way to deal with removable media (CD-ROMs, DVDs, floppies, Zip drives, etc.), MINIX 3 allows the file system on a CD-ROM to be attached to the main tree. Consider the situation of Fig. 1-7(a). Before the `mount` call, the **root file system**, on the hard disk, and a second file system, on a CD-ROM, are separate and unrelated.

Figure 1-7. (a) Before mounting, the files on drive 0 are not accessible. (b) After mounting, they are part of the file hierarchy.

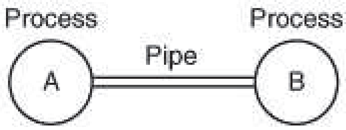


However, the file system on the CD-ROM cannot be used, because there is no way to specify path names on it. MINIX 3 does not allow path names to be prefixed by a drive name or number; that is precisely the kind of device dependence that operating systems ought to eliminate. Instead, the `mount` system call allows the file system on the CD-ROM to be attached to the root file system wherever the program wants it to be. In Fig. 1-7(b) the file system on drive 0 has been mounted on directory *b*, thus allowing access to files */b/x* and */b/y*. If directory *b* had originally contained any files they would not be accessible while the CD-ROM was mounted, since */b* would refer to the root directory of drive 0. (Not being able to access these files is not as serious as it at first seems: file systems are nearly always mounted on empty directories.) If a system contains multiple hard disks, they can all be mounted into a single tree as well.

Another important concept in MINIX 3 is the **special file**. Special files are provided in order to make I/O devices look like files. That way, they can be read and written using the same system calls as are used for reading and writing files. Two kinds of special files exist: **block special files** and **character special files**. Block special files are normally used to model devices that consist of a collection of randomly addressable blocks, such as disks. By opening a block special file and reading, say, block 4, a program can directly access the fourth block on the device, without regard to the structure of the file system contained on it. Similarly, character special files are used to model printers, modems, and other devices that accept or output a character stream. By convention, the special files are kept in the */dev* directory. For example, */dev/lp* might be the line printer.

The last feature we will discuss in this overview is one that relates to both processes and files: pipes. A **pipe** is a sort of pseudofile that can be used to connect two processes, as shown in Fig. 1-8. If processes *A* and *B* wish to talk using a pipe, they must set it up in advance. When process *A* wants to send data to process *B*, it writes on the pipe as though it were an output file. Process *B* can read the data by reading from the pipe as though it were an input file. Thus, communication between processes in MINIX 3 looks very much like ordinary file reads and writes. Stronger yet, the only way a process can discover that the output file it is writing on is not really a file, but a pipe, is by making a special system call.

Figure 1-8. Two processes connected by a pipe.



1.3.3. The Shell

The operating system is the code that carries out the system calls. Editors, compilers, assemblers, linkers, and command interpreters definitely are not part of the operating system, even though they are important and useful. At the risk of confusing things somewhat, in this section we will look briefly at the MINIX 3 command interpreter, called the **shell**. Although it is not part of the operating system, it makes heavy use of many operating system features and thus serves as a good example of how the system calls can be used. It is also the primary interface between a user sitting at his terminal and the operating system, unless the user is using a graphical user interface. Many shells exist, including *csh*, *ksh*, *zsh*, and *bash*. All of them support the functionality described below, which derives from the original shell (*sh*).

When any user logs in, a shell is started up. The shell has the terminal as standard input and standard output. It starts out by typing the **prompt**, a character such as a dollar sign, which tells the user that the shell is waiting to accept a command. If the user now types

```
date
```

for example, the shell creates a child process and runs the *date* program as the child. While the child process is running, the shell waits for it to terminate. When the child finishes, the shell types the prompt again and tries to read the next input line.

The user can specify that standard output be redirected to a file, for example,

```
date >file
```

Similarly, standard input can be redirected, as in

```
sort <file1 >file2
```

which invokes the sort program with input taken from *file1* and output sent to *file2*.

The output of one program can be used as the input for another program by connecting them with a pipe. Thus

```
cat file1 file2 file3 | sort >/dev/lp
```

invokes the *cat* program to con*cat*enate three files and send the output to *sort* to arrange all the lines in alphabetical order. The output of *sort* is redirected to the file */dev/lp*, typically the printer.

If a user puts an ampersand after a command, the shell does not wait for it to complete. Instead it just gives a prompt immediately. Consequently,

```
cat file1 file2 file3 | sort >/dev/lp &
```

starts up the sort as a background job, allowing the user to continue working normally while the sort is going on. The shell has a number of other interesting features, which we do not have space to discuss here. Most books for UNIX beginners are useful for MINIX 3 users who want to learn more about using the system. Examples are Ray and Ray (2003) and Herborth (2005).

PREV NEXT

PREV
NEXT

[Page 26 (continued)]

1.4. System Calls

Armed with our general knowledge of how MINIX 3 deals with processes and files, we can now begin to look at the interface between the operating system and its application programs, that is, the set of system calls. Although this discussion specifically refers to POSIX (International Standard 9945-1), hence also to MINI 3, UNIX, and Linux, most other modern operating systems have system calls that perform the same functions, even if the details differ. Since the actual mechanics of issuing a system call are highly machine dependent, and often must be expressed in assembly code, a procedure library is provided to make it possible to make system calls from C programs.

It is useful to keep the following in mind: any single-CPU computer can execute only one instruction at a time. If a process is running a user program in user mode and needs a system service, such as reading data from a file, it has to execute a trap or system call instruction to transfer control to the operating system. The operating system then figures out what the calling process wants by inspecting the parameters. Then it carries out the system call and returns control to the instruction following the system call. In a sense, making a system call is like making a special kind of procedure call, only system calls enter the kernel or other privileged operating system components and procedure calls do not.

[Page 27]

To make the system call mechanism clearer, let us take a quick look at `read` . It has three parameters: the first one specifying the file, the second one specifying the buffer, and the third one specifying the number of bytes to read. A call to `read` from a C program might look like this:

```
count = read(fd, buffer, nbytes);
```

The system call (and the library procedure) return the number of bytes actually read in *count* . This value is normally the same as *nbytes* , but may be smaller, if, for example, end-of-file is encountered while reading.

If the system call cannot be carried out, either due to an invalid parameter or a disk error, *count* is set to 1, and the error number is put in a global variable, *errno* . Programs should always check the results of a system call to see if an error occurred.

MINIX 3 has a total of 53 main system calls. These are listed in Fig. 1-9 , grouped for convenience in six categories. A few other calls exist, but they have very specialized uses so we will omit them here. In the following sections we will briefly examine each of the calls of Fig. 1-9 to see what it does. To a large extent, the services offered by these calls determine most of what the operating system has to do, since the resource management on personal computers is minimal (at least compared to big machines with many users).

Process management

pid = **`fork`** ()

Create a child process identical to the parent

pid = `waitpid` (pid, &statloc, opts)

Wait for a child to terminate

s = `wait` (&status)

Old version of waitpid

s = `execve` (name, argv, envp)

Replace a process core image

`exit` (status)

Terminate process execution and return status

size = `brk` (addr)

Set the size of the data segment

pid = `getpid` ()

Return the caller's process id

pid = `getpgrp` ()

Return the id of the caller's process group

pid = `setsid` ()

Create a new session and return its proc. group id

l = `ptrace` (req, pid, addr, data)

Used for debugging

Signals

s = `sigaction` (sig, &act, &oldact)

Define action to take on signals

s = `sigreturn` (&context)

Return from a signal

s = `sigprocmask` (how, &set, &old)

Examine or change the signal mask

s = `sigpending` (set)

Get the set of blocked signals

s = `sigsuspend` (sigmask)

Replace the signal mask and suspend the process

s = `kill` (pid, sig)

Send a signal to a process

residual = `alarm` (seconds)

Set the alarm clock

s = `pause` ()

Suspend the caller until the next signal

File Management

fd = `creat` (name, mode)

Obsolete way to create a new file

fd = `mknod` (name, mode, addr)

Create a regular, special, or directory i-node fd = `open` (file, how, ...)

Open a file for reading, writing or both

s = `close` (fd)

Close an open file

n = `read` (fd, buffer, nbytes)

Read data from a file into a buffer

n = `write` (fd, buffer, nbytes)

Write data from a buffer into a file

pos = `lseek` (fd, offset, whence)

Move the file pointer

s = `stat` (name, &buf)

Get a file's status information

s = `fstat` (fd, &buf)

Get a file's status information

fd = `dup` (fd)

Allocate a new file descriptor for an open file

s = `pipe` (&fd[0])

Create a pipe

s = `ioctl` (fd, request, argp)

Perform special operations on a file

s = `access` (name, amode)

Check a file's accessibility

s = `rename` (old, new)

Give a file a new name

s = `fcntl` (fd, cmd, ...)

File locking and other operations

Dir. & File System Mgt .

s = `mkdir` (name, mode)

Create a new directory

s = `rmdir` (name)

Remove an empty directory

s = `link` (name1, name2)

Create a new entry, name2, pointing to name1

s = `unlink` (name)

Remove a directory entry

s = `mount` (special, name, flag)

Mount a file system

s = `umount` (special)

Unmount a file system

s = `sync` ()

Flush all cached blocks to the disk

s = `chdir` (dirname)

Change the working directory

s = `chroot` (dirname)

Change the root directory

Protection

s = `chmod` (name, mode)

Change a file's protection bits

uid = `getuid` ()

Get the caller's uid

gid = `getgid` ()

Get the caller's gid

s = `setuid` (uid)

Set the caller's uid

s = `setgid` (gid)

Set the caller's gid

s = `chown` (name, owner, group)

Change a file's owner and group

oldmask = `umask` (complmode)

Change the mode mask

Time Management

seconds = `time` (&seconds)

Get the elapsed time since Jan. 1, 1970

s = `stime` (tp)

Set the elapsed time since Jan. 1, 1970

s = `utime` (file, timep)

Set a file's "last access" time

s = `times` (buffer)

Get the user and system times used so far

Figure 1-9. The main MINIX system calls. *fd* is a file descriptor; *n* is a byte count.

(This item is displayed on page 28 in the print version)

This is a good place to point out that the mapping of POSIX procedure calls onto system calls is not necessarily one-to-one. The POSIX standard specifies a number of procedures that a conformant system must supply, but it does not specify whether they are system calls, library calls, or something else. In some cases, the POSIX procedures are supported as library routines in MINIX 3. In others, several required procedures are only minor variations of one another, and one system call handles all of them.

1.4.1. System Calls for Process Management

The first group of calls in Fig. 1-9 deals with process management. `Fork` is a good place to start the discussion. `Fork` is the only way to create a new process in MINIX 3. It creates an exact duplicate of the original process, including all the file descriptors, registerseverything. After the `fork` , the original process and the copy (the parent and child) go their separate ways. All the variables have identical values at the time of the `fork` , but since the parent's data are copied to create the child, subsequent changes in one of them do not affect the other one. (The program text, which is unchangeable, is shared between parent and child.) The `fork` call returns a value, which is zero in the child and equal to the child's process identifier or **PID** in the parent. Using the returned PID, the two processes can see which one is the parent process and which one is the child process.

In most cases, after a `fork` , the child will need to execute different code from the parent. Consider the shell. It reads a command from the terminal, forks off a child process, waits for the child to execute the command, and then reads the next command when the child terminates. To wait for the child to finish, the parent executes a `waitpid` system call, which just waits until the child terminates (any child if more than one exists). `Waitpid` can wait for a specific child, or for any old child by setting the first parameter to 1. When `waitpid` completes, the address pointed to by the second parameter, *statloc* , will be set to the child's exit status (normal or abnormal termination and exit value). Various options are also provided, specified by the third parameter. The `waitpid` call replaces the previous `wait` call, which is now obsolete but is provided for reasons of backward compatibility.

Now consider how `fork` is used by the shell. When a command is typed, the shell forks off a new process. This child process must execute the user command. It does this by using the `execve` system call, which causes its entire core image to be replaced by the file named in its first parameter. (Actually, the system call itself is `exec` , but several different library procedures call it with different parameters and slightly different names. We will treat these as system calls here.)A highly simplified shell illustrating the use of `fork` , `waitpid` , and `execve` is shown in Fig. 1-10 .

Figure 1-10. A stripped-down shell. Throughout this book, *TRUE* is assumed to be defined as 1.

```
#define TRUE 1

while (TRUE){                              /* repeat forever */
     type_prompt();                        /* display prompt on the screen */
     read_command(command, parameters);    /* read input from terminal */

     if (fork() != 0){                     /* fork off child process */
          /* Parent code. */
          waitpid(1, &status, 0);         /* wait for child to exit */
     } else {
         /* Child code. */
         execve(command, parameters, 0);   /* execute command */
     }
}
```

In the most general case, `execve` has three parameters: the name of the file to be executed, a pointer to the argument array, and a pointer to the environment array. These will be described shortly. Various library routines, including *execl* , *execv* , *execle* , and *execve* , are provided to allow the parameters to be omitted or specified in various ways. Throughout this book we will use the name `exec` to represent the system call invoked by all of these.

Let us consider the case of a command such as

```
cp file1 file2
```

used to copy *file1* to *file2* . After the shell has forked, the child process locates and executes the file *cp* and passes to it the names of the source and target files.

The main program of *cp* (and main program of most other C programs) contains the declaration

```
main(argc, argv, envp)
```

where *argc* is a count of the number of items on the command line, including the program name. For the example above, *argc* is 3.

The second parameter, *argv* , is a pointer to an array. Element *i* of that array is a pointer to the *i* -th string on the command line. In our example, *argv* [0] would point to the string "cp", *argv* [1] would point to the string "file1", and *argv* [2] would point to the string "file2".

The third parameter of *main* , *envp* , is a pointer to the environment, an array of strings containing assignments of the form *name=value* used to pass information such as the terminal type and home directory name to a program. In Fig. 1-10 , no environment is passed to the child, so the third parameter of *execve* is a zero.

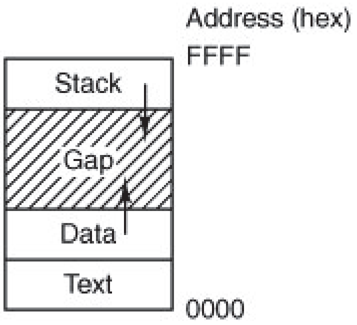
If `exec` seems complicated, do not despair; it is (semantically) the most complex of all the POSIX system calls. All the other ones are much simpler. As an example of a simple one, consider `exit` , which processes should use when they are finished executing. It has one parameter, the exit status (0 to 255), which is returned to the parent via *statloc* in the `waitpid` system call. The low-order byte of *status* contains the termination status, with 0 being normal termination and the other values being various error conditions. The high-order byte contains the child's exit status (0 to 255). For example, if a parent process executes the statement

```
n = waitpid(1, &statloc, options);
```

it will be suspended until some child process terminates. If the child exits with, say, 4 as the parameter to *exit* , the parent will be awakened with *n* set to the child's PID and *statloc* set to 0x0400 (the C convention of prefixing hexadecimal constants with 0x will be used throughout this book).

Processes in MINIX 3 have their memory divided up into three segments: the **text segment** (i.e., the program code), the **data segment** (i.e., the variables), and the **stack segment** . The data segment grows upward and the stack grows downward, as shown in Fig. 1-11 . Between them is a gap of unused address space. The stack grows into the gap automatically, as needed, but expansion of the data segment is done explicitly by using a system call, `brk` , which specifies the new address where the data segment is to end. This address may be more than the current value (data segment is growing) or less than the current value (data segment is shrinking). The parameter must, of course, be less than the stack pointer or the data and stack segments would overlap, which is forbidden.

Figure 1-11. Processes have three segments: text, data, and stack. In this example, all three are in one address space, but separate instruction and data space is also supported.



As a convenience for programmers, a library routine *sbrk* is provided that also changes the size of the data segment, only its parameter is the number of bytes to add to the data segment (negative parameters make the data segment smaller). It works by keeping track of the current size of the data segment, which is the value returned by `brk` , computing the new size, and making a call asking for that number of bytes. The `brk` and `sbrk` calls, however, are not defined by the POSIX standard. Programmers are encouraged to use the *malloc* library procedure for dynamically allocating storage, and the underlying implementation of *malloc* was not thought to be a suitable subject for standardization since few programmers use it directly.

The next process system call is also the simplest, `getpid` . It just returns the caller's PID. Remember that in `fork` , only the parent was given the child's PID. If the child wants to find out its own PID, it must use `getpid` . The `getpgrp` call returns the PID of the caller's process group. `setsid` creates a new session and sets the process group's PID to the caller's. Sessions are related to an optional feature of POSIX, **job control** , which is not supported by MINIX 3 and which will not concern us further.

The last process management system call, `ptrace` , is used by debugging programs to control the program being debugged. It allows the debugger to read and write the controlled process' memory and manage it in other ways.

1.4.2. System Calls for Signaling

Although most forms of interprocess communication are planned, situations exist in which unexpected communication is needed. For example, if a user accidently tells a text editor to list the entire contents of a very long file, and then realizes the error, some way is needed to interrupt the editor. In MINIX 3, the user can hit the CTRL-C key on the keyboard, which sends a **signal** to the editor. The editor catches the signal and stops the print-out. Signals can also be used to report certain traps detected by the hardware, such as illegal instruction or floating point overflow. Timeouts are also implemented as signals.

When a signal is sent to a process that has not announced its willingness to accept that signal, the process is simply killed without further ado. To avoid this fate, a process can use the `sigaction` system call to announce that it is prepared to accept some signal type, and to provide the address of the signal handling procedure and a place to store the address of the current one. After a `sigaction` call, if a signal of the relevant type is generated (e.g., by pressing CTRL-C), the state of the process is pushed onto its own stack, and then the signal handler is called. It may run for as long as it wants to and perform any system calls it wants to. In practice, though, signal handlers are usually fairly short. When the signal handling procedure is done, it calls `sigreturn` to continue where it left off before the signal. The `sigaction` call replaces the older `signal` call, which is now provided as a library procedure, however, for backward compatibility.

Signals can be blocked in MINIX 3. A blocked signal is held pending until it is unblocked. It is not delivered, but also not lost. The `sigprocmask` call allows a process to define the set of blocked signals by presenting the kernel with a bitmap. It is also possible for a process to ask for the set of signals currently pending but not allowed to be delivered due to their being blocked. The `sigpending` call returns this set as a bitmap. Finally, the `sigsuspend` call allows a process to atomically set the bitmap of blocked signals and suspend itself.

Instead of providing a function to catch a signal, the program may also specify the constant SIG_IGN to have all subsequent signals of the specified type ignored, or SIG_DFL to restore the default action of the signal when it occurs. The default action is either to kill the process or ignore the signal, depending upon the signal. As an example of how SIG_IGN is used, consider what happens when the shell forks off a background process as a result of

```
command &
```

It would be undesirable for a SIGINT signal (generated by pressing CTRL-C) to affect the background process, so after the `fork` but before the `exec` , the shell does

```
sigaction(SIGINT, SIG_IGN, NULL);
```

and

```
sigaction(SIGQUIT, SIG_IGN, NULL);
```

to disable the SIGINT and SIGQUIT signals. (SIGQUIT is generated by CTRL-\; it is the same as SIGINT generated by CTRL-C except that if it is not caught or ignored it makes a core dump of the process killed.) For foreground processes (no ampersand), these signals are not ignored.

Hitting CTRL-C is not the only way to send a signal. The `kill` system call allows a process to signal another process (provided they have the same UID unrelated processes cannot signal each other). Getting back to the example of background processes used above, suppose a background process is started up, but later it is decided that the process should be terminated. SIGINT and SIGQUIT have been disabled, so something else is needed. The solution is to use the *kill* program, which uses the `kill` system call to send a signal to any process. By sending signal 9 (SIGKILL), to a background process, that process can be killed. SIGKILL cannot be caught or ignored.

For many real-time applications, a process needs to be interrupted after a specific time interval to do something, such as to retransmit a potentially lost packet over an unreliable communication line. To handle this situation, the `alarm` system call has been provided. The parameter specifies an interval, in seconds, after which a SIGALRM signal is sent to the process. A process may only have one alarm outstanding at any instant. If an `alarm` call is made with a parameter of 10 seconds, and then 3 seconds later another `alarm` call is made with a parameter of 20 seconds, only one signal will be generated, 20 seconds after the second call. The first signal is canceled by the second call to `alarm` . If the parameter to `alarm` is zero, any pending alarm signal is canceled. If an alarm signal is not caught, the default action is taken and the signaled process is killed.

It sometimes occurs that a process has nothing to do until a signal arrives. For example, consider a computer-aided-instruction program that is testing reading speed and comprehension. It displays some text on the screen and then calls `alarm` to signal it after 30 seconds. While the student is reading the text, the program has nothing to do. It could sit in a tight loop doing nothing, but that would waste CPU time that another process or user might need. A better idea is to use `pause` , which tells MINIX 3 to suspend the process until the next signal.

1.4.3. System Calls for File Management

Many system calls relate to the file system. In this section we will look at calls that operate on individual files; in the next one we will examine those that involve directories or the file system as a whole. To create a new file, the `creat` call is used (why the call is `creat` and not `create` has been lost in the mists of time). Its parameters provide the name of the file and the protection mode. Thus

```
fd = creat("abc", 0751);
```

creates a file called *abc* with mode 0751 octal (in C, a leading zero means that a constant is in octal). The low-order 9 bits of 0751 specify the *rwx* bits for the owner (7 means read-write-execute permission), his group (5 means read-execute), and others (1 means execute only).

`Creat` not only creates a new file but also opens it for writing, regardless of the file's mode. The file descriptor returned, *fd* , can be used to write the file. If a `creat` is done on an existing file, that file is truncated to length 0, provided, of course, that the permissions are all right. The `creat` call is obsolete, as `open` can now create new files, but it has been included for backward compatibility.

Special files are created using `mknod` rather than `creat` . A typical call is

```
fd = mknod("/dev/ttyc2", 020744, 0x0402);
```

which creates a file named */dev/ttyc2* (the usual name for console 2) and gives it mode 020744 octal (a character special file with protection bits *rwxr--r--*). The third parameter contains the major device (4) in the high-order byte and the minor device (2) in the low-order byte. The major device could have been anything, but a file named */dev/ttyc2* ought to be minor device 2. Calls to `mknod` fail unless the caller is the superuser.

To read or write an existing file, the file must first be opened using `open` . This call specifies the file name to be opened, either as an absolute path name or relative to the working directory, and a code of *O_RDONLY* , *O_WRONLY* , or *O_RDWR* , meaning open for reading, writing, or both. The file descriptor returned can then be used for reading or writing. Afterward, the file can be closed by `close` , which makes the file descriptor available for reuse on a subsequent `creat` or `open` .

The most heavily used calls are undoubtedly `read` and `write` . We saw `read` earlier; `write` has the same parameters.

Although most programs read and write files sequentially, for some applications programs need to be able to access any part of a file at random. Associated with each file is a pointer that indicates the current position in the file. When reading (writing) sequentially, it normally points to the next byte to be read (written). The `lseek` call changes the value of the position pointer, so that subsequent calls to `read` or `write` can begin anywhere in the file, or even beyond the end.

`lseek` has three parameters: the first is the file descriptor for the file, the second is a file position, and the third tells whether the file position is relative to the beginning of the file, the current position, or the end of the file. The value returned by `lseek` is the absolute position in the file after changing the pointer.

For each file, MINIX 3 keeps track of the file mode (regular file, special file, directory, and so on), size, time of last modification, and other information. Programs can ask to see this information via the `stat` and `fstat` system calls. These differ only in that the former specifies the file by name, whereas the latter takes a file descriptor, making it useful for open files, especially standard input and standard output, whose names may not be known. Both calls provide as the second parameter a pointer to a structure where the information is to be put. The structure is shown in Fig. 1-12 .

Figure 1-12. The structure used to return information for the `stat` and `fstat` system calls. In the actual code, symbolic names are used for some of the types.

(This item is displayed on page 35 in the print version)

```
struct stat{
   short st_dev;                       /* device where i-node belongs */
   unsigned short st_ino;              /* i-node number */
   unsigned short st_mode;             /* mode word */
   short st_nlink;                     /* number of links */
   short st_uid;                       /* user id */
   short st_gid;                       /* group id */
   short st_rdev;                      /* major/minor device for special files */
   long st_size;                       /* file size */
   long st_atime;                      /* time of last access */
   long st_mtime;                      /* time of last modification */
   long st_ctime;                      /* time of last change to i-node */
};
```

When manipulating file descriptors, the `dup` call is occasionally helpful. Consider, for example, a program that needs to close standard output (file descriptor 1), substitute another file as standard output, call a function that writes some output onto standard output, and then restore the original situation. Just closing file descriptor 1 and then opening a new file will make the new file standard output (assuming standard input, file descriptor 0, is in use), but it will be impossible to restore the original situation later.

[Page 35]

The solution is first to execute the statement

```
fd = dup(1);
```

which uses the dup system call to allocate a new file descriptor, *fd* , and arrange for it to correspond to the same file as standard output. Then standard output can be closed and a new file opened and used. When it is time to restore the original situation, file descriptor 1 can be closed, and then

```
n = dup(fd);
```

executed to assign the lowest file descriptor, namely, 1, to the same file as *fd* . Finally, *fd* can be closed and we are back where we started.

The dup call has a variant that allows an arbitrary unassigned file descriptor to be made to refer to a given open file. It is called by

```
dup2(fd, fd2);
```

where *fd* refers to an open file and *fd2* is the unassigned file descriptor that is to be made to refer to the same file as *fd* . Thus if *fd* refers to standard input (file descriptor 0) and *fd2* is 4, after the call, file descriptors 0 and 4 will both refer to standard input.

Interprocess communication in MINIX 3 uses pipes, as described earlier. When a user types

```
cat file1 file2 | sort
```

the shell creates a pipe and arranges for standard output of the first process to write to the pipe, so standard input of the second process can read from it. The pipe system call creates a pipe and returns two file descriptors, one for writing and one for reading. The call is

```
pipe(&fd[0]);
```

where *fd* is an array of two integers and *fd* [0] is the file descriptor for reading and *fd* [1] is the one for writing. Typically, a fork comes next, and the parent closes the file descriptor for reading and the child closes the file descriptor for writing (or vice versa), so when they are done, one process can read the pipe and the other can write on it.

Figure 1-13 depicts a skeleton procedure that creates two processes, with the output of the first one piped into the second one. (A more realistic example would do error checking and handle arguments.) First a pipe is created, and then the procedure forks, with the parent eventually becoming the first process in the pipeline and the child process becoming the second one. Since the files to be executed, *process1* and *process2* , do not know that they are part of a pipeline, it is essential that the file descriptors be manipulated so that the first process' standard output be the pipe and the second one's standard input be the pipe. The parent first closes off the file descriptor for reading from the pipe. Then it closes standard output and does a DUP call that allows file descriptor 1 to write on the pipe. It is important to realize that dup always returns the lowest available file descriptor, in this case, 1. Then the program closes the other pipe file descriptor.

Figure 1-13. A skeleton for setting up a two-process pipeline.

(This item is displayed on page 37 in the print version)

```
#define STD_INPUT0                        /* file descriptor for standard input */
#define STD_OUTPUT1                       /* file descriptor for standard output */
pipeline(process1, process2)
char *process1, *process2;                /* pointers to program names */
{
 int fd[2];

 pipe(&fd[0]);                            /* create a pipe */
 if (fork() != 0) {
      /* The parent process executes these statements. */
      close(fd[0]);                       /* process 1 does not need to read from pipe
      close(STD_OUTPUT);                  /* prepare for new standard output */
      dup(fd[1]);                         /* set standard output to fd[1] */
      close(fd[1]);                       /* this file descriptor not needed any more
      execl(process1, process1, 0);
 } else {
      /* The child process executes these statements. */
      close(fd[1]);                       /* process 2 does not need to write to pipe
      close(STD_INPUT);                   /* prepare for new standard input */
      dup(fd[0]);                         /* set standard input to fd[0] */
      close(fd[0]);                       /* this file descriptor not needed any more
      execl(process2, process2, 0);
 }
}
```

After the exec call, the process started will have file descriptors 0 and 2 be unchanged, and file descriptor 1 for writing on the pipe. The child code is analogous. The parameter to *execl* is repeated because the first one is the file to be executed and the second one is the first parameter, which most programs expect to be the file name.

The next system call, ioctl , is potentially applicable to all special files. It is, for instance, used by block device drivers like the SCSI driver to control tape and CD-ROM devices. Its main use, however, is with special character files, primarily terminals. POSIX defines a number of functions which the library translates into ioctl calls. The *tcgetattr* and *tcsetattr* library functions use ioctl to change the characters used for correcting typing errors on the terminal, changing the **terminal mode** , and so forth.

Traditionally, there are three terminal modes, cooked, raw, and cbreak. **Cooked mode** is the normal terminal mode, in which the erase and kill characters work normally, CTRL-S and CTRL-Q can be used for stopping and starting terminal output, CTRL-D means end of file, CTRL-C generates an interrupt signal, and CTRL-\ generates a quit signal to force a core dump.

In **raw mode** , all of these functions are disabled; consequently, every character is passed directly to the program with no special processing. Furthermore, in raw mode, a read from the terminal will give the program any characters that have been typed, even a partial line, rather than waiting for a complete line to be typed, as in cooked mode. Screen editors often use this mode.

Cbreak mode is in between. The erase and kill characters for editing are disabled, as is CTRL-D, but CTRL-S, CTRL-Q, CTRL-C, and CTRL-\ are enabled. Like raw mode, partial lines can be returned to programs (if intraline editing is turned off there is no need to wait until a whole line has been receivedthe user cannot change his mind and delete it, as he can in cooked mode).

POSIX does not use the terms cooked, raw, and cbreak. In POSIX terminology **canonical mode** corresponds to cooked mode. In this mode there are eleven special characters defined, and input is by lines. In **noncanonical mode** a minimum number of characters to accept and a time, specified in units of 1/10th of a second, determine how a read will be satisfied. Under POSIX there is a great deal of flexibility, and various flags can be set to make noncanonical mode behave like either cbreak or raw mode. The older terms are more descriptive, and we will continue to use them informally.

`Ioctl` has three parameters, for example a call to *tcsetattr* to set terminal parameters will result in

```
ioctl(fd, TCSETS, &termios);
```

The first parameter specifies a file, the second one specifies an operation, and the third one is the address of the POSIX structure that contains flags and the array of control characters. Other operation codes instruct the system to postpone the changes until all output has been sent, cause unread input to be discarded, and return the current values.

The `access` system call is used to determine whether a certain file access is permitted by the protection system. It is needed because some programs can run using a different user's UID. This SETUID mechanism will be described later.

The `rename` system call is used to give a file a new name. The parameters specify the old and new names.

Finally, the `fcntl` call is used to control files, somewhat analogous to `ioctl` (i.e., both of them are horrible hacks). It has several options, the most important of which is for advisory file locking. Using `fcntl` , it is possible for a process to lock and unlock parts of files and test part of a file to see if it is locked. The call does not enforce any lock semantics. Programs must do this themselves.

1.4.4. System Calls for Directory Management

In this section we will look at some system calls that relate more to directories or the file system as a whole, rather than just to one specific file as in the previous section. The first two calls, `mkdir` and `rmdir` , create and remove empty directories, respectively. The next call is `link` . Its purpose is to allow the same file to appear under two or more names, often in different directories. A typical use is to allow several members of the same programming team to share a common file, with each of them having the file appear in his own directory, possibly under different names. Sharing a file is not the same as giving every team member a private copy, because having a shared file means that changes that any member of the team makes are instantly visible to the other membersthere is only one file. When copies are made of a file, subsequent changes made to one copy do not affect the other ones.

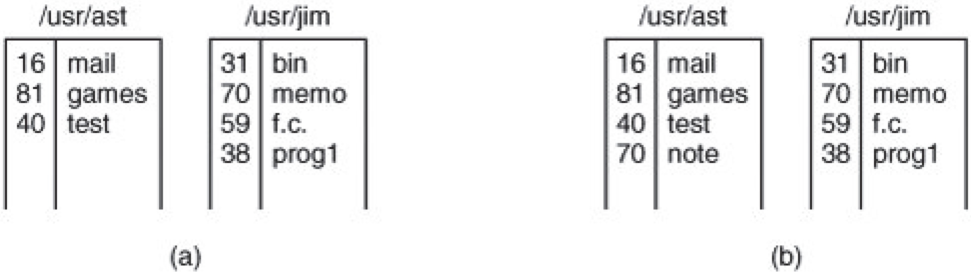
To see how `link` works, consider the situation of Fig. 1-14(a) . Here are two users, *ast* and *jim* , each having their own directories with some files. If *ast* now executes a program containing the system call

```
link("/usr/jim/memo", "/usr/ast/note");
```

the file *memo* in *jim* 's directory is now entered into *ast* 's directory under the name *note* . Thereafter, */usr/jim/memo* and */usr/ast/note* refer to the same file.

Figure 1-14. (a) Two directories before linking */usr/jim/memo* to ast's directory. (b) The same directories after linking.

(This item is displayed on page 39 in the print version)



(a)

/usr/ast	
16	mail
81	games
40	test

/usr/jim	
31	bin
70	memo
59	f.c.
38	prog1

(b)

/usr/ast	
16	mail
81	games
40	test
70	note

/usr/jim	
31	bin
70	memo
59	f.c.
38	prog1

Understanding how `link` works will probably make it clearer what it does. Every file in UNIX has a unique number, its i-number, that identifies it. This inumber is an index into a table of **i-nodes,** one per file, telling who owns the file, where its disk blocks are, and so on. A directory is simply a file containing a set of (i-number, ASCII name) pairs. In the first versions of UNIX, each directory entry was 16 bytes2 bytes for the i-number and 14 bytes for the name. A more complicated structure is needed to support long file names, but conceptually a directory is still a set of (i-number, ASCII name) pairs. In Fig. 1-14 , *mail* has inumber 16, and so on. What `link` does is simply create a new directory entry with a (possibly new) name, using the i-number of an existing file. In Fig. 1-14(b) , two entries have the same i-number (70) and thus refer to the same file. If either one is later removed, using the `unlink` system call, the other one remains. If both are removed, UNIX sees that no entries to the file exist (a field in the i-node keeps track of the number of directory entries pointing to the file), so the file is removed from the disk.

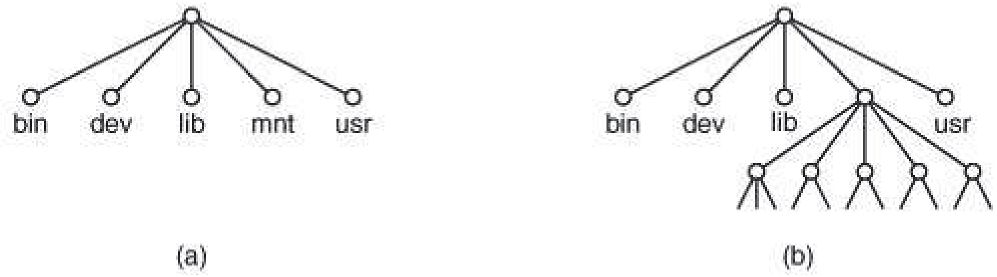
As we have mentioned earlier, the `mount` system call allows two file systems to be merged into one. A common situation is to have the root file system containing the binary (executable) versions of the common commands and other heavily used files, on a hard disk. The user can then insert a CD-ROM with files to be read into the CD-ROM drive.

By executing the `mount` system call, the CD-ROM file system can be attached to the root file system, as shown in Fig. 1-15 . A typical statement in C to perform the mount is

```
mount("/dev/cdrom0", "/mnt", 0);
```

where the first parameter is the name of a block special file for CD-ROM drive 0, the second parameter is the place in the tree where it is to be mounted, and the third one tells whether the file system is to be mounted read-write or read-only.

Figure 1-15. (a) File system before the mount. (b) File system after the mount.



After the `mount` call, a file on CD-ROM drive 0 can be accessed by just using its path from the root directory or the working directory, without regard to which drive it is on. In fact, second, third, and fourth drives can also be mounted anywhere in the tree. The `mount` call makes it possible to integrate removable media into a single integrated file hierarchy, without having to worry about which device a file is on. Although this example involves CD-ROMs, hard disks or portions of hard disks (often called **partitions** or **minor devices**) can also be mounted this way. When a file system is no longer needed, it can be unmounted with the `umount` system call.

MINIX 3 maintains a **block cache** cache of recently used blocks in main memory to avoid having to read them from the disk if they are used again quickly. If a block in the cache is modified (by a `write` on a file) and the system crashes before the modified block is written out to disk, the file system will be damaged. To limit the potential damage, it is important to flush the cache periodically, so that the amount of data lost by a crash will be small. The system call `sync` tells MINIX 3 to write out all the cache blocks that have been modified since being read in. When MINIX 3 is started up, a program called *update* is started as a background process to do a `sync` every 30 seconds, to keep flushing the cache.

Two other calls that relate to directories are `chdir` and `chroot` . The former changes the working directory and the latter changes the root directory. After the call

```
chdir("/usr/ast/test");
```

an open on the file *xyz* will open */usr/ast/test/xyz* . `chroot` works in an analogous way. Once a process has told the system to change its root directory, all absolute path names (path names beginning with a "/") will start at the new root. Why would you want to do that? For securityserver programs for protocols such as **FTP** (File Transfer Protocol) and **HTTP** (HyperText Transfer Protocol) do this so remote users of these services can access only the portions of a file system below the new root. Only superusers may execute `chroot` , and even superusers do not do it very often.

1.4.5. System Calls for Protection

In MINIX 3 every file has an 11-bit mode used for protection. Nine of these bits are the read-write-execute bits for the owner, group, and others. The `chmod` system call makes it possible to change the mode of a file. For example, to make a file read-only by everyone except the owner, one could execute

```
chmod("file", 0644);
```

The other two protection bits, 02000 and 04000, are the SETGID (set-group-id) and SETUID (set-user-id) bits, respectively. When any user executes a program with the SETUID bit on, for the duration of that process the user's effective UID is changed to that of the file's owner. This feature is heavily used to allow users to execute programs that perform superuser only functions, such as creating directories. Creating a directory uses `mknod` , which is for the superuser only. By arranging for the *mkdir* program to be owned by the superuser and have mode 04755, ordinary users can be given the power to execute `mknod` but in a highly restricted way.

When a process executes a file that has the SETUID or SETGID bit on in its mode, it acquires an effective UID or GID different from its real UID or GID. It is sometimes important for a process to find out what its real and effective UID or GID is. The system calls `getuid` and `getgid` have been provided to supply this information. Each call returns both the real and effective UID or GID, so four library routines are needed to extract the proper information: *getuid* , *getgid* , *geteuid* , and *getegid* . The first two get the real UID/GID, and the last two the effective ones.

Ordinary users cannot change their UID, except by executing programs with the SETUID bit on, but the superuser has another possibility: the `setuid` system call, which sets both the effective and real UIDs. `setgid` sets both GIDs. The superuser can also change the owner of a file with the `chown` system call. In short, the superuser has plenty of opportunity for violating all the protection rules, which explains why so many students devote so much of their time to trying to become superuser.

The last two system calls in this category can be executed by ordinary user processes. The first one, `umask` , sets an internal bit mask within the system, which is used to mask off mode bits when a file is created. After the call

```
umask(022);
```

the mode supplied by `creat` and `mknod` will have the 022 bits masked off before being used. Thus the call

```
creat("file", 0777);
```

will set the mode to 0755 rather than 0777. Since the bit mask is inherited by child processes, if the shell does a `umask` just after login, none of the user's processes in that session will accidently create files that other people can write on.

When a program owned by the root has the SETUID bit on, it can access any file, because its effective UID is the superuser. Frequently it is useful for the program to know if the person who called the program has permission to access a given file. If the program just tries the access, it will always succeed, and thus learn nothing.

What is needed is a way to see if the access is permitted for the real UID. The `access` system call provides a way to find out. The *mode* parameter is 4 to check for read access, 2 for write access, and 1 for execute access. Combinations of these values are also allowed. For example, with *mode* equal to 6, the call returns 0 if both read and write access are allowed for the real ID; otherwise1 is returned. With *mode* equal to 0, a check is made to see if the file exists and the directories leading up to it can be searched.

Although the protection mechanisms of all UNIX-like operating systems are generally similar, there are some differences and inconsistencies that lead to security vulnerabilities. See Chen et al. (2002) for a discussion.

1.4.6. System Calls for Time Management

MINIX 3 has four system calls that involve the time-of-day clock. `Time` just returns the current time in seconds, with 0 corresponding to Jan. 1, 1970 at midnight (just as the day was starting, not ending). Of course, the system clock must be set at some point in order to allow it to be read later, so `stime` has been provided to let the clock be set (by the superuser). The third time call is `utime` , which allows the owner of a file (or the superuser) to change the time stored in a file's i-node. Application of this system call is fairly limited, but a few programs need it, for example, *touch* , which sets the file's time to the current time.

Finally, we have `times` , which returns the accounting information to a process, so it can see how much CPU time it has used directly, and how much CPU time the system itself has expended on its behalf (handling its system calls). The total user and system times used by all of its children combined are also returned.

PREV
NEXT

PREV NEXT

3. Input/Output

One of the main functions of an operating system is to control all the computer's I/O (Input/Output) devices. It must issue commands to the devices, catch interrupts, and handle errors. It should also provide an interface between the devices and the rest of the system that is simple and easy to use. To the extent possible, the interface should be the same for all devices (device independence). The I/O code represents a significant fraction of the total operating system. Thus to really understand what an operating system does, you have to understand how I/O works. How the operating system manages I/O is the main subject of this chapter.

This chapter is organized as follows. First we will look at some of the principles of how I/O hardware is organized. Then we will look at I/O software in general. I/O software can be structured in layers, with each layer having a well-defined task to perform. We will look at these layers to see what they do and how they fit together.

After that comes a section on deadlocks. We will define deadlocks precisely, show how they are caused, give two models for analyzing them, and discuss some algorithms for preventing their occurrence.

Then we will move on to look at MINIX 3 We will start with a bird's-eye view of I/O in MINIX 3, including interrupts, device drivers, device-dependent I/O and device-independent I/O. Following that introduction, we will look at several I/O devices in detail: disks, keyboards, and displays. For each device we will look at its hardware and software.

PREV NEXT

3.1. Principles of I/O Hardware

Different people look at I/O hardware in different ways. Electrical engineers look at it in terms of chips, wires, power supplies, motors, and all the other physical components that make up the hardware. Programmers look at the interface presented to the softwarethe commands the hardware accepts, the functions it carries out, and the errors that can be reported back. In this book we are concerned with programming I/O devices, not designing, building, or maintaining them, so our interest will be restricted to how the hardware is programmed, not how it works inside. Nevertheless, the programming of many I/O devices is often intimately connected with their internal operation. In the next three subsections we will provide a little general background on I/O hardware as it relates to programming.

3.1.1. I/O Devices

I/O devices can be roughly divided into two categories: **block devices** and **character devices**. A block device is one that stores information in fixed-size blocks, each one with its own address. Common block sizes range from 512 bytes to 32,768 bytes. The essential property of a block device is that it is possible to read or write each block independently of all the other ones. Disks are the most common block devices.

If you look closely, the boundary between devices that are block addressable and those that are not is not well defined. Everyone agrees that a disk is a block addressable device because no matter where the arm currently is, it is always possible to seek to another cylinder and then wait for the required block to rotate under the head. Now consider a tape drive used for making disk backups. Tapes contain a sequence of blocks. If the tape drive is given a command to read block *N*, it can always rewind the tape and go forward until it comes to block *N*. This operation is analogous to a disk doing a seek, except that it takes much longer. Also, it may or may not be possible to rewrite one block in the middle of a tape. Even if it were possible to use tapes as random access block devices, that is stretching the point somewhat: they are not normally used that way.

The other type of I/O device is the character device. A character device delivers or accepts a stream of characters, without regard to any block structure. It is not addressable and does not have any seek operation. Printers, network interfaces, mice (for pointing), rats (for psychology lab experiments), and most other devices that are not disk-like can be seen as character devices.

This classification scheme is not perfect. Some devices just do not fit in. Clocks, for example, are not block addressable. Nor do they generate or accept character streams. All they do is cause interrupts at well-defined intervals. Still, the model of block and character devices is general enough that it can be used as a basis for making some of the operating system software dealing with I/O device independent. The file system, for example, deals only with abstract block devices and leaves the device-dependent part to lower-level software called **device drivers**.

I/O devices cover a huge range in speeds, which puts considerable pressure on the software to perform well over many orders of magnitude in data rates. Fig. 3-1 shows the data rates of some common devices. Most of these devices tend to get faster as time goes on.

Figure 3-1. Some typical device, network, and bus data rates.

Device	Data rate
Keyboard	10 bytes/sec
Mouse	100 bytes/sec
56K modem	7 KB/sec
Scanner	400 KB/sec
Digital camcorder	4 MB/sec
52x CD-ROM	8 MB/sec
FireWire (IEEE 1394)	50 MB/sec
USB 2.0	60 MB/sec
XGA Monitor	60 MB/sec
SONET OC-12 network	78 MB/sec
Gigabit Ethernet	125 MB/sec
Serial ATA disk	200 MB/sec
SCSI Ultrawide 4 disk	320 MB/sec
PCI bus	528 MB/sec

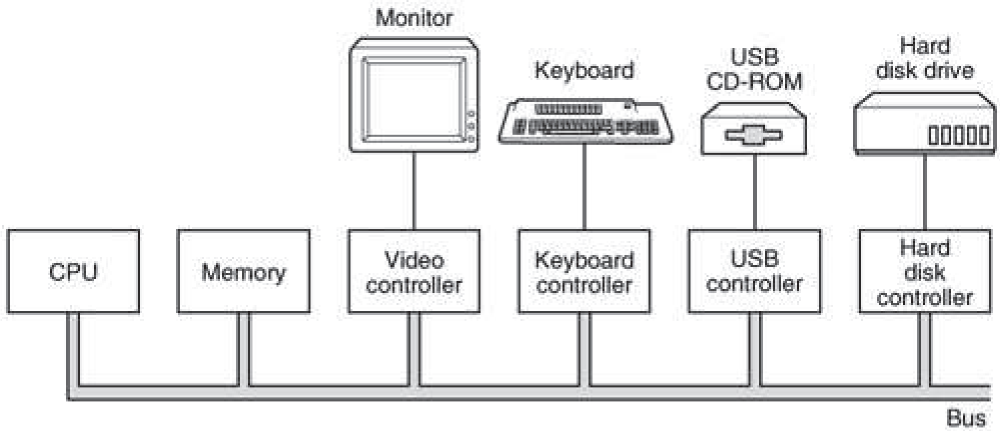
3.1.2. Device Controllers

I/O units typically consist of a mechanical component and an electronic component. It is often possible to separate the two portions to provide a more modular and general design. The electronic component is called the **device controller** or **adapter**. On personal computers, it often takes the form of a printed circuit card that can be inserted into an expansion slot. The mechanical component is the device itself. This arrangement is shown in Fig. 3-2

Figure 3-2. A model for connecting the CPU, memory, controllers, and I/O devices.

(This item is displayed on page 224 in the print version)

[View full size image]



The controller card usually has a connector on it, into which a cable leading to the device itself can be plugged. Many controllers can handle two, four, or even eight identical devices. If the interface between the controller and device is a standard interface, either an official ANSI, IEEE, or ISO standard or a de facto one, then companies can make controllers or devices that fit that interface. Many companies, for example, make disk drives that match the IDE (Integrated Drive Electronics) and SCSI (Small Computer System Interface) interfaces.

We mention this distinction between controller and device because the operating system nearly always deals with the controller, not the device. Most personal computers and servers use the bus model of Fig. 3-2 for communication between the CPU and the controllers. Large mainframes often use a different model, with specialized I/O computers called **I/O channels** taking some of the load off the main CPU.

The interface between the controller and the device is often low-level. A disk, for example, might be formatted with 1024 sectors of 512 bytes per track. What actually comes off the drive, however, is a serial bit stream, starting with a **preamble**, then the 4096 bits in a sector, and finally a checksum, also called an **Error-Correcting Code** (**ECC**). The preamble is written when the disk is formatted and contains the cylinder and sector number, the sector size, and similar data.

The controller's job is to convert the serial bit stream into a block of bytes and perform any error correction necessary. The block of bytes is typically first assembled, bit by bit, in a buffer inside the controller. After its checksum has been verified and the block declared to be free of errors, it can then be copied to main memory.

The controller for a monitor also works as a bit serial device at an equally low level. It reads bytes containing the characters to be displayed from memory and generates the signals used to modulate the CRT beam. The controller also generates the signals for making a CRT beam do a horizontal retrace after it has finished a scan line, as well as the signals for making it do a vertical retrace after the entire screen has been scanned. On an LCD screen these signals select individual pixels and control their brightness, simulating the effect of the electron beam in a CRT. If it were not for the video controller, the operating system programmer would have to program the scanning explicitly. With the controller, the operating system initializes the controller with a few parameters, such as the number of characters or pixels per line and number of lines per screen, and lets the controller take care of actually driving the display.

Controllers for some devices, especially disks, are becoming extremely sophisticated. For example, modern disk controllers often have many megabytes of memory inside the controller. As a result, when a read is being processed, as soon as the arm gets to the correct cylinder, the controller begins reading and storing data, even if it has not yet reached the sector it needs. This cached data may come in handy for satisfying subsequent requests. Furthermore, even after the requested data has been obtained, the controller may continue to cache data from subsequent sectors, since they are likely to be needed later. In this manner, many disk reads can be handled without any disk activity at all.

3.1.3. Memory-Mapped I/O

Each controller has a few registers that are used for communicating with the CPU. By writing into these registers, the operating system can command the device to deliver data, accept data, switch itself on or off, or otherwise perform some action. By reading from these registers, the operating system can learn what the device's state is, whether it is prepared to accept a new command, and so on.

In addition to the control registers, many devices have a data buffer that the operating system can read and write. For example, a common way for computers to display pixels on the screen is to have a video RAM, which is basically just a data buffer, available for programs or the operating system to write into.

The issue thus arises of how the CPU communicates with the control registers and the device data buffers. Two alternatives exist. In the first approach, each control register is assigned an **I/O port** number, an 8- or 16-bit integer. Using a special I/O instruction such as

```
IN REG,PORT
```

the CPU can read in control register `PORT` and store the result in CPU register `REG`. Similarly, using

```
OUT PORT,REG
```

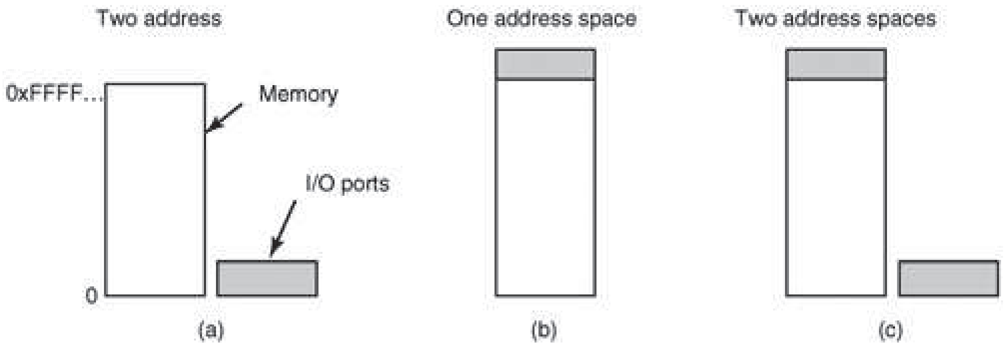
the CPU can write the contents of `REG` to a control register. Most early computers, including nearly all mainframes, such as the IBM 360 and all of its successors, worked this way.

In this scheme, the address spaces for memory and I/O are different, as shown in Fig. 3-3(a).

Figure 3-3. (a) Separate I/O and memory space. (b) Memory-mapped I/O. (c) Hybrid.

(This item is displayed on page 226 in the print version)

[View full size image]



On other computers, I/O registers are part of the regular memory address space, as shown in Fig. 3-3(b). This scheme is called **memory-mapped I/O,** and was introduced with the PDP-11 minicomputer. Each control register is assigned a unique memory address to which no memory is assigned. Usually, the assigned addresses are at the top of the address space. A hybrid scheme, with memory-mapped I/O data buffers and separate I/O ports for the control registers is shown in Fig. 3-3(c). The Pentium uses this architecture, with addresses 640K to 1M being reserved for device data buffers in IBM PC compatibles, in addition to I/O ports 0 through 64K.

How do these schemes work? In all cases, when the CPU wants to read a word, either from memory or from an I/O port, it puts the address it needs on the address lines of the bus and then asserts a `READ` signal on a bus control line. A second signal line is used to tell whether I/O space or memory space is needed. If it is memory space, the memory responds to the request. If it is I/O space, the I/O device responds to the request. If there is only memory space [as in Fig. 3-3(b)], every memory module and every I/O device compares the address lines to the range of addresses that it services. If the address falls in its range, it responds to the request. Since no address is ever assigned to both memory and an I/O device, there is no ambiguity and no conflict.

3.1.4. Interrupts

Usually, controller registers have one or more **status bits** that can be tested to determine if an output operation is complete or if new data is available from an input device. A CPU can execute a loop, testing a status bit each time until a device is ready to accept or provide new data. This is called **polling** or **busy waiting**. We saw this concept in Sec. 2.2.3 as a possible method to deal with critical sections, and in that context it was dismissed as something to be avoided in most circumstances. In the realm of I/O, where you might have to wait a very long time for the outside world to accept or produce data, polling is not acceptable except for very small dedicated systems not running multiple processes.

In addition to status bits, many controllers use interrupts to tell the CPU when they are ready to have their registers read or written. We saw how interrupts are handled by the CPU in Sec. 2.1.6. In the context of I/O, all you need to know is that most interface devices provide an output which is logically the same as the "operation complete" or "data ready" status bit of a register, but which is meant to be used to drive one of the IRQ (Interrupt ReQuest) lines of the system bus. Thus when an interrupt-enabled operation completes, it interrupts the CPU and starts the interrupt handler running. This piece of code informs the operating system that I/O is complete. The operating system may then check the status bits to verify that all went well, and either harvest the resulting data or initiate a retry.

The number of inputs to the interrupt controller may be limited; Pentium-class PCs have only 15 available for I/O devices. Some controllers are hard-wired onto the system parentboard, for example, the disk and keyboard controllers of an IBM PC. On older systems, the IRQ used by the device was set by a switch or jumper associated with the controller. If a user bought a new plug-in board, he had to manually set the IRQ to avoid conflicts with existing IRQs. Few users could do this correctly, which led the industry to develop **Plug 'n Play**, in which the BIOS can automatically assign IRQs to devices at boot time to avoid conflicts.

3.1.5. Direct Memory Access (DMA)

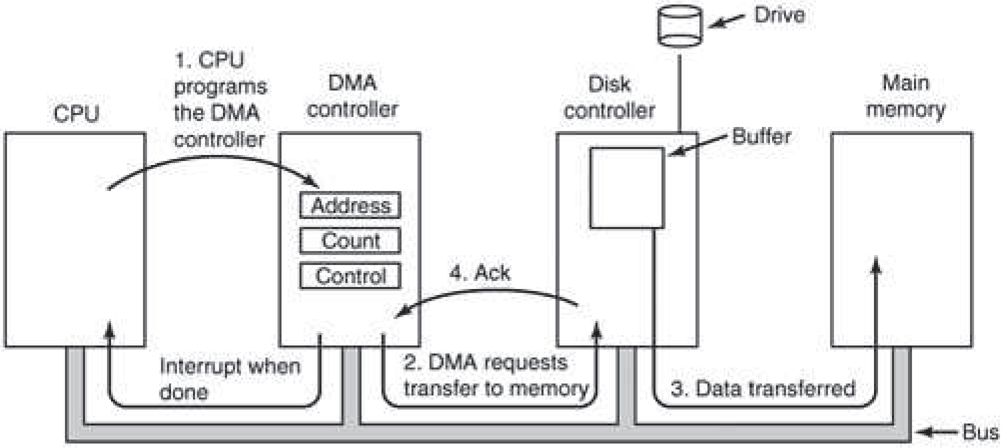
Whether or not a system has memory-mapped I/O, its CPU needs to address the device controllers to exchange data with them. The CPU can request data from an I/O controller one byte at a time but doing so for a device like a disk that produces a large block of data wastes the CPU's time, so a different scheme, called **DMA** (**Direct Memory Access**) is often used. The operating system can only use DMA if the hardware has a DMA controller, which most systems do. Sometimes this controller is integrated into disk controllers and other controllers, but such a design requires a separate DMA controller for each device. More commonly, a single DMA controller is available (e.g., on the parentboard) for regulating transfers to multiple devices, often concurrently.

No matter where it is physically located, the DMA controller has access to the system bus independent of the CPU, as shown in Fig. 3-4. It contains several registers that can be written and read by the CPU. These include a memory address register, a byte count register, and one or more control registers. The control registers specify the I/O port to use, the direction of the transfer (reading from the I/O device or writing to the I/O device), the transfer unit (byte at a time or word at a time), and the number of bytes to transfer in one burst.

Figure 3-4. Operation of a DMA transfer.

(This item is displayed on page 228 in the print version)

[View full size image]



To explain how DMA works, let us first look at how disk reads occur when DMA is not used. First the controller reads the block (one or more sectors) from the drive serially, bit by bit, until the entire block is in the controller's internal buffer. Next, it computes the checksum to verify that no read errors have occurred. Then the controller causes an interrupt. When the operating system starts running, it can read the disk block from the controller's buffer a byte or a word at a time by executing a loop, with each iteration reading one byte or word from a controller device register, storing it in main memory, incrementing the memory address, and decrementing the count of items to be read until it reaches zero.

When DMA is used, the procedure is different. First the CPU programs the DMA controller by setting its registers so it knows what to transfer where (step 1 in Fig. 3-4). It also issues a command to the disk controller telling it to read data from the disk into its internal buffer and verify the checksum. When valid data are in the disk controller's buffer, DMA can begin.

The DMA controller initiates the transfer by issuing a read request over the bus to the disk controller (step 2). This read request looks like any other read request, and the disk controller does not know or care whether it came from the CPU or from a DMA controller. Typically, the memory address to write to is on the address lines of the bus so when the disk controller fetches the next word from its internal buffer, it knows where to write it. The write to memory is another standard bus cycle (step 3). When the write is complete, the disk controller sends an acknowledgement signal to the disk controller, also over the bus (step 4). The DMA controller then increments the memory address to use and decrements the byte count. If the byte count is still greater than 0, steps 2 through 4 are repeated until the count reaches 0. At this point the controller causes an interrupt. When the operating system starts up, it does not have to copy the block to memory; it is already there.

You may be wondering why the controller does not just store the bytes in main memory as soon as it gets them from the disk. In other words, why does it need an internal buffer? There are two reasons. First, by doing internal buffering, the disk controller can verify the checksum before starting a transfer. If the checksum is incorrect, an error is signaled and no transfer to memory is done.

The second reason is that once a disk transfer has started, the bits keep arriving from the disk at a constant rate, whether the controller is ready for them or not. If the controller tried to write data directly to memory, it would have to go over the system bus for each word transferred. If the bus were busy due to some other device using it, the controller would have to wait. If the next disk word arrived before the previous one had been stored, the controller would have to store it somewhere. If the bus were very busy, the controller might end up storing quite a few words and having a lot of administration to do as well. When the block is buffered internally, the bus is not needed until the DMA begins, so the design of the controller is much simpler because the DMA transfer to memory is not time critical.

Not all computers use DMA. The argument against it is that the main CPU is often far faster than the DMA controller and can do the job much faster (when the limiting factor is not the speed of the I/O device). If there is no other work for it to do, having the (fast) CPU wait for the (slow)

PREV NEXT

[Page 229 (continued)]

3.2. Principles of I/O Software

Let us now turn away from the I/O hardware and look at the I/O software. First we will look at the goals of the I/O software and then at the different ways I/O can be done from the point of view of the operating system.

3.2.1. Goals of the I/O Software

A key concept in the design of I/O software is **device independence**. What this means is that it should be possible to write programs that can access any I/O device without having to specify the device in advance. For example, a program that reads a file as input should be able to read a file on a floppy disk, on a hard disk, or on a CD-ROM, without having to modify the program for each different device. Similarly, one should be able to type a command such as

```
sort <input >output
```

and have it work with input coming from a floppy disk, an IDE disk, a SCSI disk, or the keyboard, and the output going to any kind of disk or the screen. It is up to the operating system to take care of the problems caused by the fact that these devices really are different and require very different command sequences to read or write.

Closely related to device independence is the goal of **uniform naming**. The name of a file or a device should simply be a string or an integer and not depend on the device in any way. In UNIX and MINIX 3, all disks can be integrated into the file system hierarchy in arbitrary ways so the user need not be aware of which name corresponds to which device. For example, a floppy disk can be **mounted** on top of the directory */usr/ast/backup* so that copying a file to that directory copies the file to the diskette. In this way, all files and devices are addressed the same way: by a path name.

[Page 230]

Another important issue for I/O software is **error handling**. In general, errors should be handled as close to the hardware as possible. If the controller discovers a read error, it should try to correct the error itself if it can. If it cannot, then the device driver should handle it, perhaps by just trying to read the block again. Many errors are transient, such as read errors caused by specks of dust on the read head, and will go away if the operation is repeated. Only if the lower layers are not able to deal with the problem should the upper layers be told about it. In many cases, error recovery can be done transparently at a low level without the upper levels even knowing about the error.

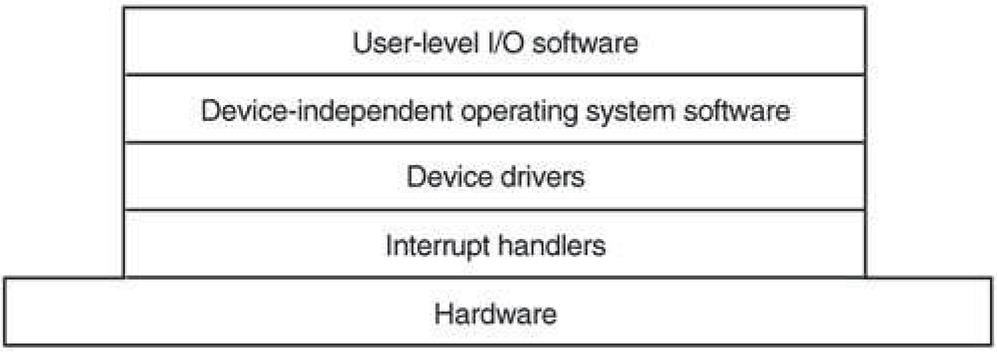
Still another key issue is **synchronous** (blocking) versus **asynchronous** (interrupt-driven) transfers. Most physical I/O is asynchronousthe CPU starts the transfer and goes off to do something else until the interrupt arrives. User programs are much easier to write if the I/O operations are blockingafter a `receive` system call the program is automatically suspended until the data are available in the buffer. It is up to the operating system to make operations that are actually interrupt-driven look blocking to the user programs.

Another issue for the I/O software is **buffering**. Often data that come off a device cannot be stored directly in its final destination. For example, when a packet comes in off the network, the operating system does not know where to put it until it has stored the packet somewhere and examined it. Also, some devices have severe real-time constraints (for example, digital audio devices), so the data must be put into an output buffer in advance to decouple the rate at which the buffer is filled from the rate at which it is emptied, in order to avoid buffer under-runs. Buffering involves considerable copying and often has a major impact on I/O performance.

The final concept that we will mention here is sharable versus dedicated devices. Some I/O devices, such as disks, can be used by many users at the same time. No problems are caused by multiple users having open files on the same disk at the same time. Other devices, such as tape drives, have to be dedicated to a single user until that user is finished. Then another user can have the tape drive. Having two or more users writing blocks intermixed at random to the same tape will definitely not work. Introducing dedicated (unshared) devices also introduces a variety of problems, such as deadlocks. Again, the operating system must be able to handle both shared and dedicated devices in a way that avoids problems.

I/O software is often organized in four layers, as shown in Fig. 3-5. In the following subsections we will look at each in turn, starting at the bottom. The emphasis in this chapter is on the device drivers (layer 2), but we will summarize the rest of the I/O software to show how the pieces of the I/O system fit together.

Figure 3-5. Layers of the I/O software system.



3.2.2. Interrupt Handlers

Interrupts are an unpleasant fact of life; although they cannot be avoided, they should be hidden away, deep in the bowels of the operating system, so that as little of the operating system as possible knows about them. The best way to hide them is to have the driver starting an I/O operation block until the I/O has completed and the interrupt occurs. The driver can block itself by doing a `down` on a semaphore, a `wait` on a condition variable, a `receive` on a message, or something similar, for example.

When the interrupt happens, the interrupt procedure does whatever it has to in order to handle the interrupt. Then it can unblock the driver that started it. In some cases it will just complete `up` on a semaphore. In others it will do a `signal` on a condition variable in a monitor. In still others, it will send a message to the blocked driver. In all cases the net effect of the interrupt will be that a driver that was previously blocked will now be able to run. This model works best if drivers are structured as independent processes, with their own states, stacks, and program counters.

3.2.3. Device Drivers

Earlier in this chapter we saw that each device controller has registers used to give it commands or to read out its status or both. The number of registers and the nature of the commands vary radically from device to device. For example, a mouse driver has to accept information from the mouse telling how far it has moved and which buttons are currently depressed. In contrast, a disk driver has to know about sectors, tracks, cylinders, heads, arm motion, motor drives, head settling times, and all the other mechanics of making the disk work properly. Obviously, these drivers will be very different.

Thus, each I/O device attached to a computer needs some device-specific code for controlling it. This code, called the **device driver**, is generally written by the device's manufacturer and delivered along with the device on a CD-ROM. Since each operating system needs its own drivers, device manufacturers commonly supply drivers for several popular operating systems.

Each device driver normally handles one device type, or one class of closely related devices. For example, it would probably be a good idea to have a single mouse driver, even if the system supports several different brands of mice. As another example, a disk driver can usually handle multiple disks of different sizes and different speeds, and perhaps a CD-ROM as well. On the other hand, a mouse and a disk are so different that different drivers are necessary.

In order to access the device's hardware, meaning the controller's registers, the device driver traditionally has been part of the system kernel. This approach gives the best performance and the worst reliability since a bug in any device driver can crash the entire system. MINIX 3 departs from this model in order to enhance reliability. As we shall see, in MINIX 3 each device driver is now a separate user-mode process.

As we mentioned earlier, operating systems usually classify drivers as **block devices**, such as disks, or **character devices**, such as keyboards and printers. Most operating systems define a standard interface that all block drivers must support and a second standard interface that all character drivers must support. These interfaces consist of a number of procedures that the rest of the operating system can call to get the driver to do work for it.

In general terms, the job of a device driver is to accept abstract requests from the device-independent software above it and see to it that the request is executed. A typical request to a disk driver is to read block *n*. If the driver is idle at the time a request comes in, it starts carrying out the request immediately. If, however, it is already busy with a request, it will normally enter the new request into a queue of pending requests to be dealt with as soon as possible.

The first step in actually carrying out an I/O request is to check that the input parameters are valid and to return an error if they are not. If the request is valid the next step is to translate it from abstract to concrete terms. For a disk driver, this means figuring out where on the disk the requested block actually is, checking to see if the drive's motor is running, determining if the arm is positioned on the proper cylinder, and so on. In short, the driver must decide which controller operations are required and in what sequence.

Once the driver has determined which commands to issue to the controller, it starts issuing them by writing into the controller's device registers. Simple controllers can handle only one command at a time. More sophisticated controllers are willing to accept a linked list of commands, which they then carry out by themselves without further help from the operating system.

After the command or commands have been issued, one of two situations will apply. In many cases the device driver must wait until the controller does some work for it, so it blocks itself until the interrupt comes in to unblock it. In other cases, however, the operation finishes without delay, so the driver need not block. As an example of the latter situation, scrolling the screen on some graphics cards requires just writing a few bytes into the controller's registers. No mechanical motion is needed, so the entire operation can be completed in a few microseconds.

In the former case, the blocked driver will be awakened by the interrupt. In the latter case, it will never go to sleep. Either way, after the operation has been completed, it must check for errors. If everything is all right, the driver may have data to pass to the device-independent software (e.g., a block just read). Finally, it returns some status information for error reporting back to its caller. If any other requests are queued, one of them can now be selected and started. If nothing is queued, the driver blocks waiting for the next request.

Dealing with requests for reading and writing is the main function of a driver, but there may be other requirements. For instance, the driver may need to initialize a device at system startup or the first time it is used. Also, there may be a need to manage power requirements, handle Plug 'n Play, or log events.

3.2.4. Device-Independent I/O Software

Although some of the I/O software is device specific, a large fraction of it is device independent. The exact boundary between the drivers and the device-independent software is system dependent, because some functions that could be done in a device-independent way may actually be done in the drivers, for efficiency or other reasons. The functions shown in Fig. 3-6 are typically done in the device-independent software. In MINIX 3, most of the device-independent software is part of the file system. Although we will study the file system in Chap. 5, we will take a quick look at the device-independent software here, to provide some perspective on I/O and show better where the drivers fit in.

Figure 3-6. Functions of the device-independent I/O software.

Uniform interfacing for device drivers
Buffering
Error reporting
Allocating and releasing dedicated devices
Providing a device-independent block size

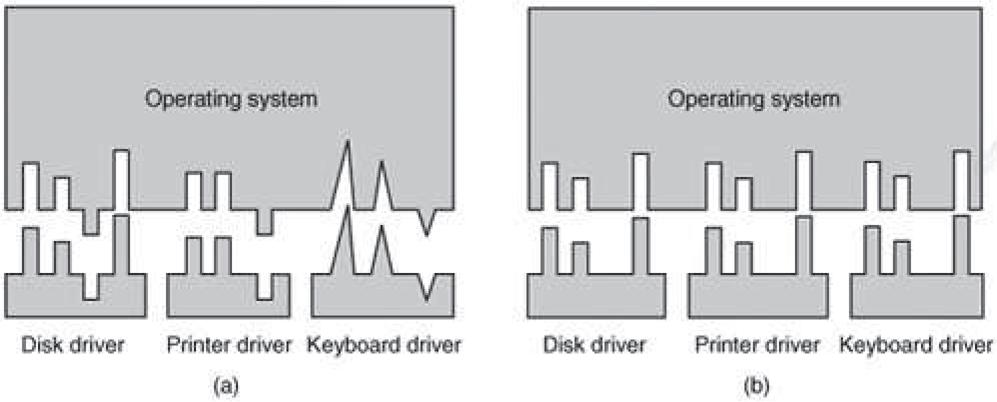
The basic function of the device-independent software is to perform the I/O functions that are common to all devices and to provide a uniform interface to the user-level software. Below we will look at the above issues in more detail.

Uniform Interfacing for Device Drivers

A major issue in an operating system is how to make all I/O devices and drivers look more-or-less the same. If disks, printers, monitors, keyboards, etc., are all interfaced in different ways, every time a new peripheral device comes along, the operating system must be modified for the new device. In Fig. 3-7(a) we illustrate symbolically a situation in which each device driver has a different interface to the operating system. In contrast, in Fig. 3-7(b), we show a different design in which all drivers have the same interface.

Figure 3-7. (a) Without a standard driver interface. (b) With a standard driver interface.

[View full size image]



With a standard interface it is much easier to plug in a new driver, providing it conforms to the driver interface. It also means that driver writers know what is expected of them (e.g., what functions they must provide and what kernel functions they may call). In practice, not all devices are absolutely identical, but usually there are only a small number of device types and even these are generally almost the same. For example, even block and character devices have many functions in common.

Another aspect of having a uniform interface is how I/O devices are named. The device-independent software takes care of mapping symbolic device names onto the proper driver. For example, in UNIX and MINIX 3 a device name, such as */dev/disk0,* uniquely specifies the i-node for a special file, and this i-node contains the **major device number**, which is used to locate the appropriate driver. The inode also contains the **minor device number**, which is passed as a parameter to the driver in order to specify the unit to be read or written. All devices have major and minor numbers, and all drivers are accessed by using the major device number to select the driver.

Closely related to naming is protection. How does the system prevent users from accessing devices that they are not entitled to access? In UNIX, MINIX 3, and also in later Windows versions such as Windows 2000 and Windows XP, devices appear in the file system as named objects, which means that the usual protection rules for files also apply to I/O devices. The system administrator can then set the proper permissions (i.e., in UNIX the *rwx* bits) for each device.

Buffering

Buffering is also an issue for both block and character devices. For block devices, the hardware generally insists upon reading and writing entire blocks at once, but user processes are free to read and write in arbitrary units. If a user process writes half a block, the operating system will normally keep the data around internally until the rest of the data are written, at which time the block can go out to the disk. For character devices, users can write data to the system faster than it can be output, necessitating buffering. Keyboard input that arrives before it is needed also requires buffering.

Error Reporting

Errors are far more common in the context of I/O than in any other context. When they occur, the operating system must handle them as best it can. Many errors are device-specific, so only the driver knows what to do (e.g., retry, ignore, or panic). A typical error is caused by a disk block that has been damaged and cannot be read any more. After the driver has tried to read the block a certain number of times, it gives up and informs the device-independent software. How the error is treated from here on is device independent. If the error occurred while reading a user file, it may be sufficient to report the error back to the caller. However, if it occurred while reading a critical system data structure, such as the block containing the bitmap showing which blocks are free, the operating system may have to display an error message and terminate.

Allocating and Releasing Dedicated Devices

Some devices, such as CD-ROM recorders, can be used only by a single process at any given moment. It is up to the operating system to examine requests for device usage and accept or reject them, depending on whether the requested device is available or not. A simple way to handle these requests is to require processes to perform `open`s on the special files for devices directly. If the device is unavailable, the `open` fails. Closing such a dedicated device then releases it.

Device-Independent Block Size

Not all disks have the same sector size. It is up to the device-independent software to hide this fact and provide a uniform block size to higher layers, for example, by treating several sectors as a single logical block. In this way, the higher layers only deal with abstract devices that all use the same logical block size, independent of the physical sector size. Similarly, some character devices deliver their data one byte at a time (e.g., modems), while others deliver theirs in larger units (e.g., network interfaces). These differences may also be hidden.

3.2.5. User-Space I/O Software

Although most of the I/O software is within the operating system, a small portion of it consists of libraries linked together with user programs, and even whole programs running outside the kernel. System calls, including the I/O system calls, are normally made by library procedures. When a C program contains the call

```
count = write(fd, buffer, nbytes);
```

the library procedure *write* will be linked with the program and contained in the binary program present in memory at run time. The collection of all these library procedures is clearly part of the I/O system.

While these procedures do little more than put their parameters in the appropriate place for the system call, there are other I/O procedures that actually do real work. In particular, formatting of input and output is done by library procedures. One example from C is *printf*, which takes a format string and possibly some variables as input, builds an ASCII string, and then calls `write` to output the string. As an example of *printf*, consider the statement

```
printf("The square of %3d is %6d\n", i, i*i);
```

It formats a string consisting of the 14-character string "The square of" followed by the value *i* as a 3-character string, then the 4-character string "is", then i^2 as six characters, and finally a line feed.

An example of a similar procedure for input is *scanf* which reads input and stores it into variables described in a format string using the same syntax as *printf*. The standard I/O library contains a number of procedures that involve I/O and all run as part of user programs.

Not all user-level I/O software consists of library procedures. Another important category is the spooling system. **Spooling** is a way of dealing with dedicated I/O devices in a multiprogramming system. Consider a typical spooled device:a printer. Although it would be technically simple to let any user process open the character special file for the printer, suppose a process opened it and then did nothing for hours? No other process could print anything.

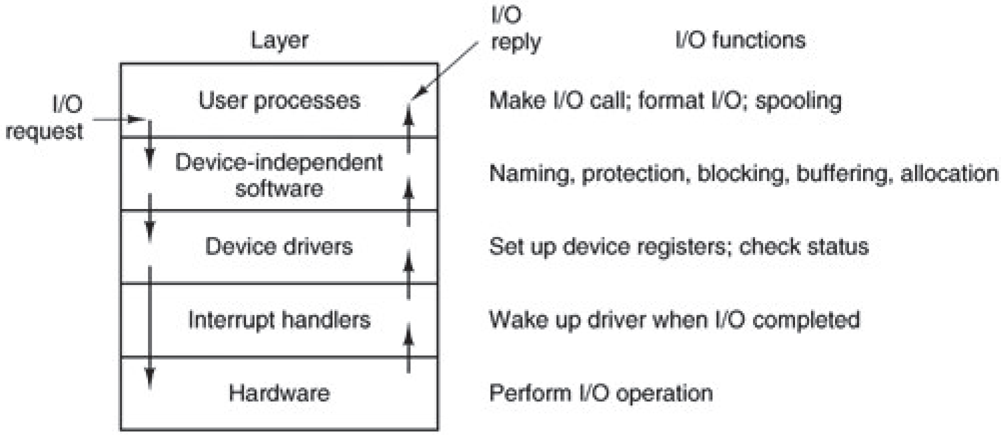
Instead what is done is to create a special process, called a **daemon**, and a special directory, called a **spooling directory**. To print a file, a process first generates the entire file to be printed and puts it in the spooling directory. It is up to the daemon, which is the only process having permission to use the printer's special file, to print the files in the directory. By protecting the special file against direct use by users, the problem of having someone keeping it open unnecessarily long is eliminated.

Spooling is used not only for printers, but also in various other situations. For example, electronic mail usually uses a daemon. When a message is submitted it is put in a mail spool directory. Later on the mail daemon tries to send it. At any given instant of time a particular destination may be temporarily unreachable, so the daemon leaves the message in the spool with status information indicating it should be tried again in a while. The daemon may also send a message back to the sender saying delivery is delayed, or, after a delay of hours or days, saying the message cannot be delivered. All of this is outside the operating system.

Figure 3-8 summarizes the I/O system, showing the layers and principal functions of each layer. Starting at the bottom, the layers are the hardware, interrupt handlers, device drivers, device-independent software, and the user processes.

Figure 3-8. Layers of the I/O system and the main functions of each layer.

[View full size image]


I/O reply
Layer
I/O functions
I/O request
User processes
Make I/O call; format I/O; spooling
Device-independent software
Naming, protection, blocking, buffering, allocation
Device drivers
Set up device registers; check status
Interrupt handlers
Wake up driver when I/O completed
Hardware
Perform I/O operation

The arrows in Fig. 3-8 show the flow of control. When a user program tries to read a block from a file, for example, the operating system is invoked to carry out the call. The device-independent software looks for it in the buffer cache, for example. If the needed block is not there, it calls the device driver to issue the request to the hardware to go get it from the disk. The process is then blocked until the disk operation has been completed.

When the disk is finished, the hardware generates an interrupt. The interrupt handler is run to discover what has happened, that is, which device wants attention right now. It then extracts the status from the device and wakes up the sleeping process to finish off the I/O request and let the

3.3. Deadlocks

Computer systems are full of resources that can only be used by one process at a time. Common examples include printers, tape drives, and slots in the system's internal tables. Having two processes simultaneously writing to the printer leads to gibberish. Having two processes using the same file system table slot will invariably lead to a corrupted file system. Consequently, all operating systems have the ability to (temporarily) grant a process exclusive access to certain resources, both hardware and software.

[Page 238]

For many applications, a process needs exclusive access to not one resource, but several. Suppose, for example, two processes each want to record a scanned document on a CD. Process *A* requests permission to use the scanner and is granted it. Process *B* is programmed differently and requests the CD recorder first and is also granted it. Now *A* asks for the CD recorder, but the request is denied until *B* releases it. Unfortunately, instead of releasing the CD recorder *B* asks for the scanner. At this point both processes are blocked and will remain so forever. This situation is called a **deadlock**.

Deadlocks can occur in a variety of situations besides requesting dedicated I/O devices. In a database system, for example, a program may have to lock several records it is using, to avoid race conditions. If process *A* locks record *R1* and process *B* locks record *R2*, and then each process tries to lock the other one's record, we also have a deadlock. Thus deadlocks can occur on hardware resources or on software resources.

In this section, we will look at deadlocks more closely, see how they arise, and study some ways of preventing or avoiding them. Although this material is about deadlocks in the context of operating systems, they also occur in database systems and many other contexts in computer science, so this material is actually applicable to a wide variety of multiprocess systems.

3.3.1. Resources

Deadlocks can occur when processes have been granted exclusive access to devices, files, and so forth. To make the discussion of deadlocks as general as possible, we will refer to the objects granted as **resources**. A resource can be a hardware device (e.g., a tape drive) or a piece of information (e.g., a locked record in a database). A computer will normally have many different resources that can be acquired. For some resources, several identical instances may be available, such as three tape drives. When interchangeable copies of a resource are available, called **fungible resources**[ ], any one of them can be used to satisfy any request for the resource. In short, a resource is anything that can be used by only a single process at any instant of time.

[ ] This is a legal and financial term. Gold is fungible: one gram of gold is as good as any other.

Resources come in two types: preemptable and nonpreemptable.A **preemptable resource** is one that can be taken away from the process owning it with no ill effects. Memory is an example of a preemptable resource. Consider, for example, a system with 64 MB of user memory, one printer, and two 64-MB processes that each want to print something. Process *A* requests and gets the printer, then starts to compute the values to print. Before it has finished with the computation, it exceeds its time quantum and is swapped or paged out.

Process *B* now runs and tries, unsuccessfully, to acquire the printer. Potentially, we now have a deadlock situation, because *A* has the printer and *B* has the memory, and neither can proceed without the resource held by the other. Fortunately, it is possible to preempt (take away) the memory from *B* by swapping it out and swapping *A* in. Now *A* can run, do its printing, and then release the printer. No deadlock occurs.

A **nonpreemptable resource**, in contrast, is one that cannot be taken away from its current owner without causing the computation to fail. If a process has begun to burn a CD-ROM, suddenly taking the CD recorder away from it and giving it to another process will result in a garbled CD. CD recorders are not preemptable at an arbitrary moment.

In general, deadlocks involve nonpreemptable resources. Potential deadlocks that involve preemptable resources can usually be resolved by reallocating resources from one process to another. Thus our treatment will focus on nonpreemptable resources.

The sequence of events required to use a resource is given below in an abstract form.

1. Request the resource.
2. Use the resource.
3. Release the resource.

If the resource is not available when it is requested, the requesting process is forced to wait. In some operating systems, the process is automatically blocked when a resource request fails, and awakened when it becomes available. In other systems, the request fails with an error code, and it is up to the calling process to wait a little while and try again.

3.3.2. Principles of Deadlock

Deadlock can be defined formally as follows:

> *A set of processes is deadlocked if each process in the set is waiting for an event that only another process in the set can cause.*

Because all the processes are waiting, none of them will ever cause any of the events that could wake up any of the other members of the set, and all the processes continue to wait forever. For this model, we assume that processes have only a single thread and that there are no interrupts possible to wake up a blocked process. The no-interrupts condition is needed to prevent an otherwise deadlocked process from being awakened by, say, an alarm, and then causing events that release other processes in the set.

In most cases, the event that each process is waiting for is the release of some resource currently possessed by another member of the set. In other words, each member of the set of deadlocked processes is waiting for a resource that is owned by a deadlocked process. None of the processes can run, none of them can release any resources, and none of them can be awakened. The number of processes and the number and kind of resources possessed and requested are unimportant. This result holds for any kind of resource, including both hardware and software.

Conditions for Deadlock

Coffman et al. (1971) showed that four conditions must hold for there to be a deadlock:

1. Mutual exclusion condition. Each resource is either currently assigned to exactly one process or is available.

2. Hold and wait condition. Processes currently holding resources that were granted earlier can request new resources.

3. No preemption condition. Resources previously granted cannot be forcibly taken away from a process. They must be explicitly released by the process holding them.

4. Circular wait condition. There must be a circular chain of two or more processes, each of which is waiting for a resource held by the next member of the chain.

All four of these conditions must be present for a deadlock to occur. If one of them is absent, no deadlock is possible.

In a series of papers, Levine (2003a, 2003b, 2005) points out there are various situations called deadlock in the literature, and that Coffman et al.'s conditions apply only to what should properly be called **resource deadlock**. The literature contains examples of "deadlock" that do not really meet all of these conditions. For instance, if four vehicles arrive simultaneously at a crossroad and try to obey the rule that each should yield to the vehicle on the right, none can proceed, but this is not a case where one process already has possession of a unique resource. Rather, this problem is a "scheduling deadlock" which can be resolved by a decision about priorities imposed from outside by a policeman.

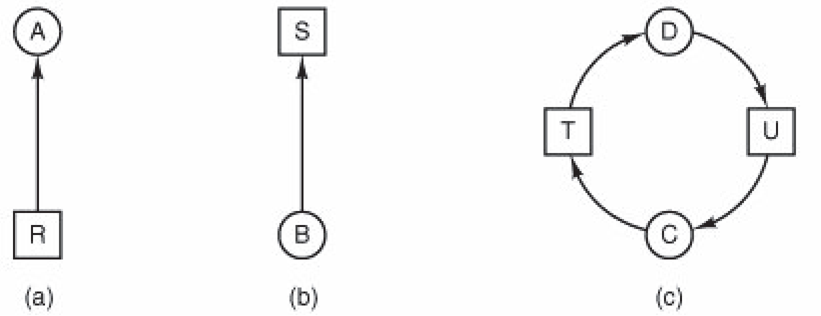
It is worth noting that each condition relates to a policy that a system can have or not have. Can a given resource be assigned to more than one process at once? Can a process hold a resource and ask for another? Can resources be preempted? Can circular waits exist? Later on we will see how deadlocks can be attacked by trying to negate some of these conditions.

Deadlock Modeling

Holt (1972) showed how these four conditions can be modeled using directed graphs. The graphs have two kinds of nodes: processes, shown as circles, and resources, shown as squares. An arc from a resource node (square) to a process node (circle) means that the resource has previously been requested by, granted to, and is currently held by that process. In Fig. 3-9(a), resource *R* is currently assigned to process *A*.

Figure 3-9. Resource allocation graphs. (a) Holding a resource. (b)

Requesting a resource. (c) Deadlock.



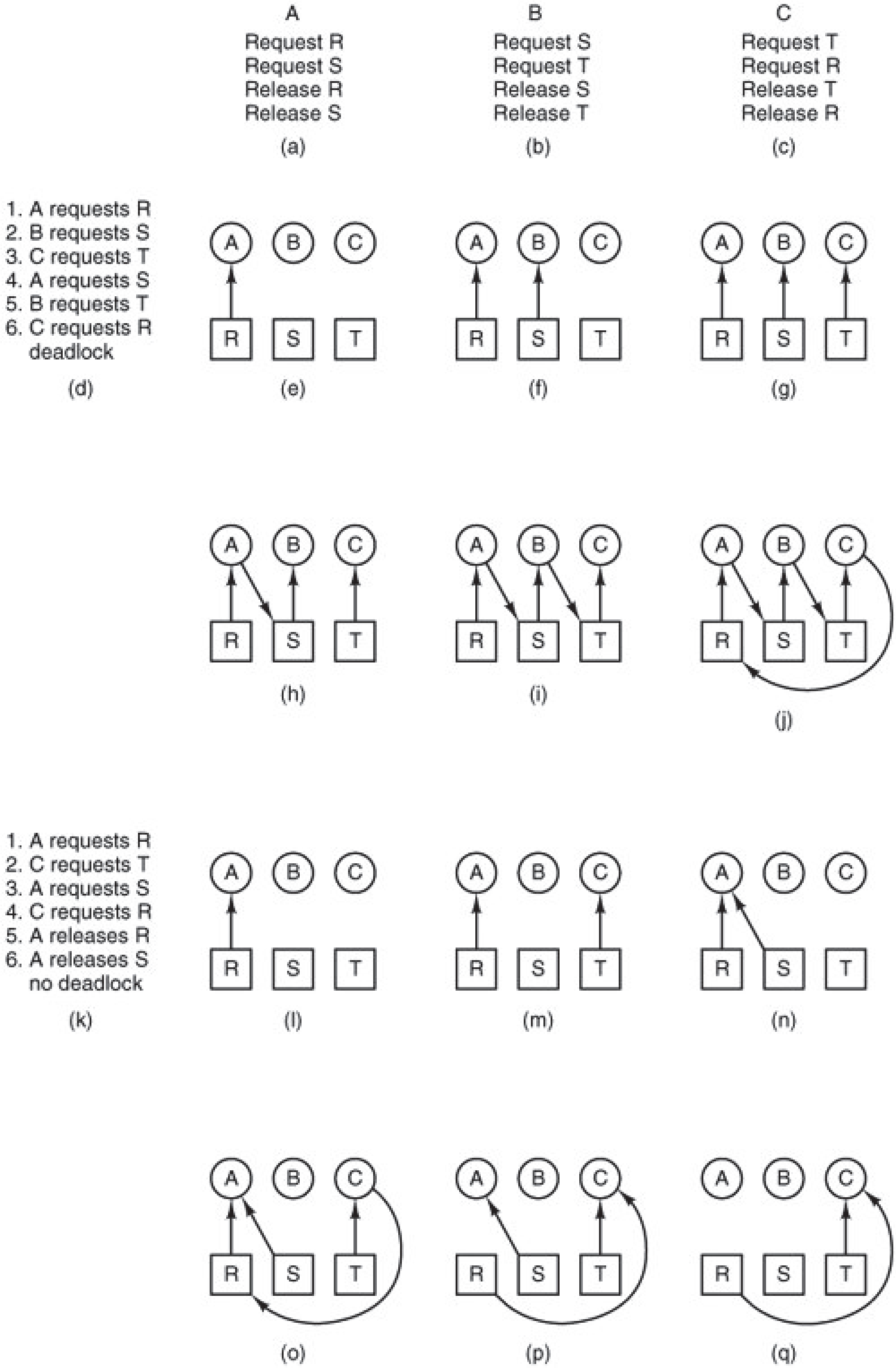
An arc from a process to a resource means that the process is currently blocked waiting for that resource. In Fig. 3-9(b), process *B* is waiting for resource *S*. In Fig. 3-9(c) we see a deadlock: process *C* is waiting for resource *T*, which is currently held by process *D*. Process *D* is not about to release resource *T* because it is waiting for resource *U*, held by *C*. Both processes will wait forever. A cycle in the graph means that there is a deadlock involving the processes and resources in the cycle (assuming that there is one resource of each kind). In this example, the cycle is *CTDUC*.

Now let us see how resource graphs can be used. Imagine that we have three processes, *A*, *B*, and *C*, and three resources, *R*, *S*, and *T*. The requests and releases of the three processes are given in Fig. 3-10(a)-(c). The operating system is free to run any unblocked process at any instant, so it could decide to run *A* until *A* finished all its work, then run *B* to completion, and finally run *C*.

Figure 3-10. An example of how deadlock occurs and how it can be avoided.

(This item is displayed on page 243 in the print version)

[View full size image]



This ordering does not lead to any deadlocks (because there is no competition for resources) but it also has no parallelism at all. In addition to requesting and releasing resources, processes compute and do I/O. When the processes are run sequentially, there is no possibility that while one process is waiting for I/O, another can use the CPU. Thus running the processes strictly sequentially may not be optimal. On the other hand, if none of the processes do any I/O at all, shortest job first is better than round robin, so under some circumstances running all processes sequentially may be the best way.

Let us now suppose that the processes do both I/O and computing, so that round robin is a reasonable scheduling algorithm. The resource requests might occur in the order of Fig. 3-10(d). If these six requests are carried out in that order, the six resulting resource graphs are shown in Fig. 3-10(e)-(j). After request 4 has been made, *A* blocks waiting for *S,* as shown in Fig. 3-10(h). In the next two steps *B* and *C* also block, ultimately leading to a cycle and the deadlock of Fig. 3-10(j). From this point on, the system is frozen.

However, as we have already mentioned, the operating system is not required to run the processes in any special order. In particular, if granting a particular request might lead to deadlock, the operating system can simply suspend the process without granting the request (i.e., just not schedule the process) until it is safe. In Fig. 3-10, if the operating system knew about the impending deadlock, it could suspend *B* instead of granting it *S*. By running only *A* and *C*, we would get the requests and releases of Fig. 3-10(k) instead of Fig. 3-10(d). This sequence leads to the resource graphs of Fig. 3-10(l)-(q), which do not lead to deadlock.

After step (q), process *B* can be granted *S* because *A* is finished and *C* has everything it needs. Even if *B* should eventually block when requesting *T*, no deadlock can occur. *B* will just wait until *C* is finished.

Later in this chapter we will study a detailed algorithm for making allocation decisions that do not lead to deadlock. For the moment, the point to understand is that resource graphs are a tool that let us see if a given request/release sequence leads to deadlock. We just carry out the requests and releases step by step, and after every step check the graph to see if it contains any cycles. If so, we have a deadlock; if not, there is no deadlock. Although our treatment of resource graphs has been for the case of a single resource of each type, resource graphs can also be generalized to handle multiple resources of the same type (Holt, 1972). However, Levine (2003a, 2003b) points out that with fungible resources this can get very complicated indeed. If even one branch of the graph is not part of a cycle, that is, if one process which is not deadlocked holds a copy of one of the resources, then deadlock may not occur.

In general, four strategies are used for dealing with deadlocks.

1. Just ignore the problem altogether. Maybe if you ignore it, it will ignore you.
2. Detection and recovery. Let deadlocks occur, detect them, and take action.
3. Dynamic avoidance by careful resource allocation.
4. Prevention, by structurally negating one of the four conditions necessary to cause a deadlock.

We will examine each of these methods in turn in the next four sections.

3.3.3. The Ostrich Algorithm

The simplest approach is the ostrich algorithm: stick your head in the sand and pretend there is no problem at all.[ ] Different people react to this strategy in very different ways. Mathematicians find it completely unacceptable and say that deadlocks must be prevented at all costs. Engineers ask how often the problem is expected, how often the system crashes for other reasons, and how serious a deadlock is. If deadlocks occur on the average once every five years, but system crashes due to hardware failures, compiler errors, and operating system bugs occur once a week, most engineers would not be willing to pay a large penalty in performance or convenience to eliminate deadlocks.

[ ] Actually, this bit of folklore is nonsense. Ostriches can run at 60 km/hour and their kick is powerful enough to kill any lion with visions of a big chicken dinner.

To make this contrast more specific, UNIX (and MINIX 3) potentially suffer from deadlocks that are not even detected, let alone automatically broken. The total number of processes in a system is determined by the number of entries in the process table. Thus process table slots are finite resources. If a `fork` fails because the table is full, a reasonable approach for the program doing the `fork` is to wait a random time and try again.

Now suppose that a MINIX 3 system has 100 process slots. Ten programs are running, each of which needs to create 12 (sub)processes. After each process has created 9 processes, the 10 original processes and the 90 new processes have exhausted the table. Each of the 10 original processes now sits in an endless loop forking and failinga deadlock. The probability of this happening is minuscule, but it *could* happen. Should we abandon processes and the `fork` call to eliminate the problem?

The maximum number of open files is similarly restricted by the size of the inode table, so a similar problem occurs when it fills up. Swap space on the disk is another limited resource. In fact, almost every table in the operating system represents a finite resource. Should we abolish all of these because it might happen that a collection of *n* processes might each claim 1 /*n* of the total, and then each try to claim another one?

Most operating systems, including UNIX, MINIX 3, and Windows, just ignore the problem on the assumption that most users would prefer an occasional deadlock to a rule restricting all users to one process, one open file, and one of everything. If deadlocks could be eliminated for free, there would not be much discussion. The problem is that the price is high, mostly in terms of putting inconvenient restrictions on processes, as we will see shortly. Thus we are faced with an unpleasant trade-off between convenience and correctness, and a great deal of discussion about which is more important, and to whom. Under these conditions, general solutions are hard to find.

3.3.4. Detection and Recovery

A second technique is detection and recovery. When this technique is used, the system does not do anything except monitor the requests and releases of resources. Every time a resource is requested or released, the resource graph is updated, and a check is made to see if any cycles exist. If a cycle exists, one of the processes in the cycle is killed. If this does not break the deadlock, another process is killed, and so on until the cycle is broken.

A somewhat cruder method is not even to maintain the resource graph but instead periodically to check to see if there are any processes that have been continuously blocked for more than say, 1 hour. Such processes are then killed.

Detection and recovery is the strategy often used on large mainframe computers, especially batch systems in which killing a process and then restarting it is usually acceptable. Care must be taken to restore any modified files to their original state, however, and undo any other side effects that may have occurred.

3.3.5. Deadlock Prevention

The third deadlock strategy is to impose suitable restrictions on processes so that deadlocks are structurally impossible. The four conditions stated by Coffman et al. (1971) provide a clue to some possible solutions.

First let us attack the mutual exclusion condition. If no resource were ever assigned exclusively to a single process, we would never have deadlocks. However, it is equally clear that allowing two processes to write on the printer at the same time will lead to chaos. By spooling printer output, several processes can generate output at the same time. In this model, the only process that actually requests the physical printer is the printer daemon. Since the daemon never requests any other resources, we can eliminate deadlock for the printer.

Unfortunately, not all devices can be spooled (the process table does not lend itself well to being spooled). Furthermore, competition for disk space for spooling can itself lead to deadlock. What would happen if two processes each filled up half of the available spooling space with output and neither was finished producing output? If the daemon was programmed to begin printing even before all the output was spooled, the printer might lie idle if an output process decided to wait several hours after the first burst of output. For this reason, daemons are normally programmed to print only after the complete output file is available. In this case we have two processes that have each finished part, but not all, of their output, and cannot continue. Neither process will ever finish, so we have a deadlock on the disk.

The second of the conditions stated by Coffman et al. looks slightly more promising. If we can prevent processes that hold resources from waiting for more resources, we can eliminate deadlocks. One way to achieve this goal is to require all processes to request all their resources before starting execution. If everything is available, the process will be allocated whatever it needs and can run to completion. If one or more resources are busy, nothing will be allocated and the process would just wait.

An immediate problem with this approach is that many processes do not know how many resources they will need until after they have started running. Another problem is that resources will not be used optimally with this approach. Take, as an example, a process that reads data from an input tape, analyzes it for an hour, and then writes an output tape as well as plotting the results. If all resources must be requested in advance, the process will tie up the output tape drive and the plotter for an hour.

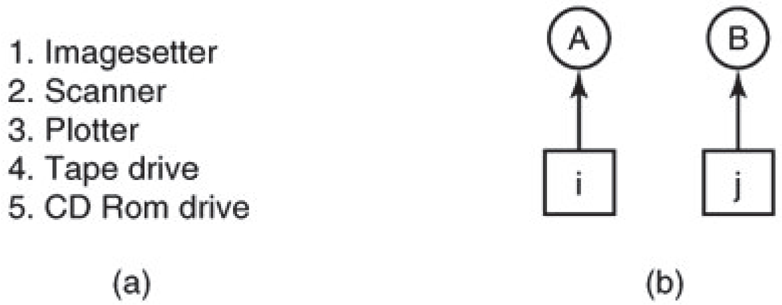
A slightly different way to break the hold-and-wait condition is to require a process requesting a resource to first temporarily release all the resources it currently holds. Then it tries to get everything it needs all at once.

Attacking the third condition (no preemption) is even less promising than attacking the second one. If a process has been assigned the printer and is in the middle of printing its output, forcibly taking away the printer because a needed plotter is not available is tricky at best and impossible at worst.

Only one condition is left. The circular wait can be eliminated in several ways. One way is simply to have a rule saying that a process is entitled only to a single resource at any moment. If it needs a second one, it must release the first one. For a process that needs to copy a huge file from a tape to a printer, this restriction is unacceptable.

Another way to avoid the circular wait is to provide a global numbering of all the resources, as shown in Fig. 3-11(a). Now the rule is this: processes can request resources whenever they want to, but all requests must be made in numerical order. A process may request first a scanner and then a tape drive, but it may not request first a plotter and then a scanner.

Figure 3-11. (a) Numerically ordered resources. (b) A resource graph.



With this rule, the resource allocation graph can never have cycles. Let us see why this is true for the case of two processes, in Fig. 3-11(b). We can get a deadlock only if *A* requests resource *j* and *B* requests resource *i*. Assuming *i* and *j* are distinct resources, they will have different numbers. If *i* > *j*, then *A* is not allowed to request *j* because that is lower than what it already has. If *i* < *j*, then *B* is not allowed to request *i* because that is lower than what it already has. Either way, deadlock is impossible.

With multiple processes, the same logic holds. At every instant, one of the assigned resources will be highest. The process holding that resource will never ask for a resource already assigned. It will either finish, or at worst, request even higher numbered resources, all of which are available. Eventually, it will finish and free its resources. At this point, some other process will hold the highest resource and can also finish. In short, there exists a scenario in which all processes finish, so no deadlock is present.

A minor variation of this algorithm is to drop the requirement that resources be acquired in strictly increasing sequence and merely insist that no process request a resource lower than what it is already holding. If a process initially requests 9 and 10, and then releases both of them, it is effectively starting all over, so there is no reason to prohibit it from now requesting resource 1.

Although numerically ordering the resources eliminates the problem of deadlocks, it may be impossible to find an ordering that satisfies everyone. When the resources include process table slots, disk spooler space, locked database records, and other abstract resources, the number of potential resources and different uses may be so large that no ordering could possibly work. Also, as Levine (2005) points out, ordering resources negates fungibilitya perfectly good and available copy of a resource could be inaccessible with such a rule.

The various approaches to deadlock prevention are summarized in Fig. 3-12.

Figure 3-12. Summary of approaches to deadlock prevention.

Condition	Approach
Mutual exclusion	Spool everything
Hold and wait	Request all resources initially
No preemption	Take resources away
Circular wait	Order resources numerically

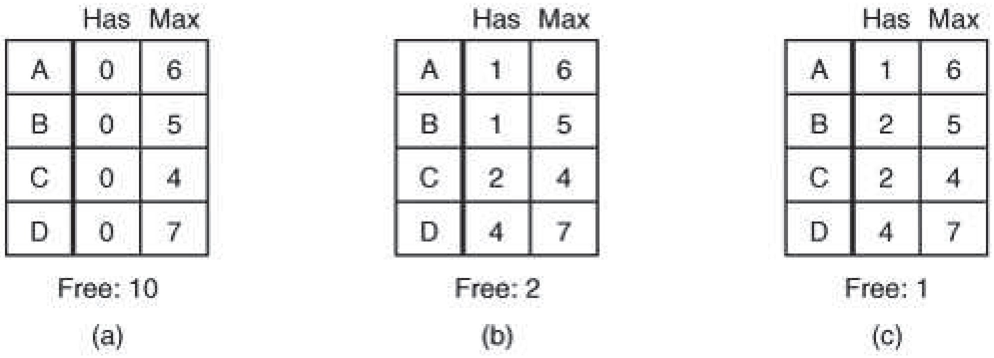
3.3.6. Deadlock Avoidance

In Fig. 3-10 we saw that deadlock was avoided not by imposing arbitrary rules on processes but by carefully analyzing each resource request to see if it could be safely granted. The question arises: is there an algorithm that can always avoid deadlock by making the right choice all the time? The answer is a qualified yes we can avoid deadlocks, but only if certain information is available in advance. In this section we examine ways to avoid deadlock by careful resource allocation.

The Banker's Algorithm for a Single Resource

A scheduling algorithm that can avoid deadlocks is due to Dijkstra (1965) and is known as the **banker's algorithm**. It is modeled on the way a small-town banker might deal with a group of customers to whom he has granted lines of credit. The banker does not necessarily have enough cash on hand to lend every customer the full amount of each one's line of credit at the same time. In Fig. 3-13(a) we see four customers, *A*, *B*, *C*, and *D*, each of whom has been granted a certain number of credit units (e.g., 1 unit is 1K dollars). The banker knows that not all customers will need their maximum credit immediately, so he has reserved only 10 units rather than 22 to service them. He also trusts every customer to be able to repay his loan soon after receiving his total line of credit (it is a small town), so he knows eventually he can service all the requests. (In this analogy, customers are processes, units are, say, tape drives, and the banker is the operating system.)

Figure 3-13. Three resource allocation states: (a) Safe. (b) Safe. (c) Unsafe.



	Has	Max
A	0	6
B	0	5
C	0	4
D	0	7

Free: 10

(a)

	Has	Max
A	1	6
B	1	5
C	2	4
D	4	7

Free: 2

(b)

	Has	Max
A	1	6
B	2	5
C	2	4
D	4	7

Free: 1

(c)

Each part of the figure shows a **state** of the system with respect to resource allocation, that is, a list of customers showing the money already loaned (tape drives already assigned) and the maximum credit available (maximum number of tape drives needed at once later). A state is **safe** if there exists a sequence of other states that leads to all customers getting loans up to their credit limits (all processes getting all their resources and terminating).

The customers go about their respective businesses, making loan requests from time to time (i.e., asking for resources). At a certain moment, the situation is as shown in Fig. 3-13(b). This state is safe because with two units left, the banker can delay any requests except *C*'s, thus letting *C* finish and release all four of his resources. With four units in hand, the banker can let either *D* or *B* have the necessary units, and so on.

Consider what would happen if a request from *B* for one more unit were granted in Fig. 3-13(b). We would have situation Fig. 3-13(c), which is unsafe. If all the customers suddenly asked for their maximum loans, the banker could not satisfy any of them, and we would have a deadlock. An unsafe state does not *have* to lead to deadlock, since a customer might not need the entire credit line available, but the banker cannot count on this behavior.

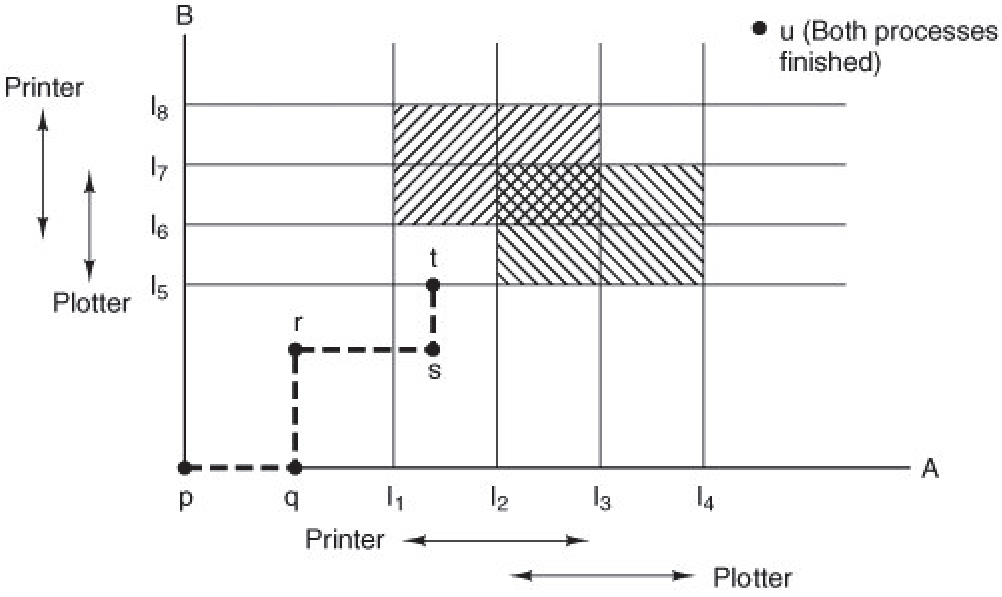
The banker's algorithm considers each request as it occurs, and sees if granting it leads to a safe state. If it does, the request is granted; otherwise, it is postponed until later. To see if a state is safe, the banker checks to see if he has enough resources to satisfy some customer. If so, those loans are assumed to be repaid, and the customer now closest to the limit is checked, and so on. If all loans can eventually be repaid, the state is safe and the initial request can be granted.

Resource Trajectories

The above algorithm was described in terms of a single resource class (e.g., only tape drives or only printers, but not some of each). In Fig. 3-14 we see a model for dealing with two processes and two resources, for example, a printer and a plotter. The horizontal axis represents the number of instructions executed by process *A*. The vertical axis represents the number of instructions executed by process *B*. At I_1 *A* requests a printer; at I_2 it needs a plotter. The printer and plotter are released at I_3 and I_4, respectively. Process *B* needs the plotter from I_5 to I_7 and the printer from I_6 to I_8.

Figure 3-14. Two process resource trajectories.

[View full size image]



Every point in the diagram represents a joint state of the two processes. Initially, the state is at *p*, with neither process having executed any instructions. If the scheduler chooses to run *A* first, we get to the point *q*, in which *A* has executed some number of instructions, but *B* has executed none. At point *q* the trajectory becomes vertical, indicating that the scheduler has chosen to run *B*. With a single processor, all paths must be horizontal or vertical, never diagonal. Furthermore, motion is always to the north or east, never to the south or west (processes cannot run backward).

When *A* crosses the I_1 line on the path from *r* to *s*, it requests and is granted the printer. When *B* reaches point *t*, it requests the plotter.

The regions that are shaded are especially interesting. The region with lines slanting from southwest to northeast represents both processes having the printer. The mutual exclusion rule makes it impossible to enter this region. Similarly, the region shaded the other way represents both processes having the plotter, and is equally impossible. Under no conditions can the system enter the shaded regions.

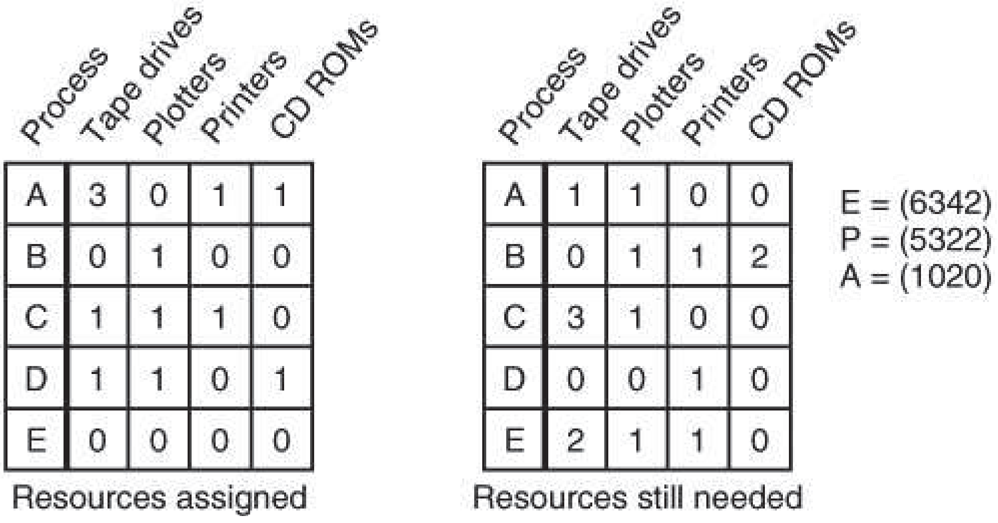
If the system ever enters the box bounded by I_1 and I_2 on the sides and I_5 and I_6 top and bottom, it will eventually deadlock when it gets to the intersection of I_2 and I_6. At this point, *A* is requesting the plotter and *B* is requesting the printer, and both are already assigned. The entire box is unsafe and must not be entered. At point *t* the only safe thing to do is run process *A* until it gets to I_4 *.* Beyond that, any trajectory to *u* will do.

The important thing to see here is at point *t B* is requesting a resource. The system must decide whether to grant it or not. If the grant is made, the system will enter an unsafe region and eventually deadlock. To avoid the deadlock, *B* should be suspended until *A* has requested and released the plotter.

The Banker's Algorithm for Multiple Resources

This graphical model is difficult to apply to the general case of an arbitrary number of processes and an arbitrary number of resource classes, each with multiple instances (e.g., two plotters, three tape drives). However, the banker's algorithm can be generalized to do the job. Figure 3-15 shows how it works.

Figure 3-15. The banker's algorithm with multiple resources.



Process	Tape drives	Plotters	Printers	CD ROMs
A	3	0	1	1
B	0	1	0	0
C	1	1	1	0
D	1	1	0	1
E	0	0	0	0

Resources assigned

Process	Tape drives	Plotters	Printers	CD ROMs
A	1	1	0	0
B	0	1	1	2
C	3	1	0	0
D	0	0	1	0
E	2	1	1	0

Resources still needed

E = (6342)
P = (5322)
A = (1020)

In Fig. 3-15 we see two matrices. The one on the left shows how many of each resource are currently assigned to each of the five processes. The matrix on the right shows how many resources each process still needs in order to complete. As in the single resource case, processes must state their total resource needs before executing, so that the system can compute the right-hand matrix at each instant.

The three vectors at the right of the figure show the existing resources, *E*, the possessed resources, *P*, and the available resources, *A*, respectively. From *E* we see that the system has six tape drives, three plotters, four printers, and two CD-ROM drives. Of these, five tape drives, three plotters, two printers, and two CD-ROM drives are currently assigned. This fact can be seen by adding up the four resource columns in the left-hand matrix. The available resource vector is simply the difference between what the system has and what is currently in use.

The algorithm for checking to see if a state is safe can now be stated.

1. Look for a row, *R*, whose unmet resource needs are all smaller than or equal to *A.* If no such row exists, the system will eventually deadlock since no process can run to completion.

2. Assume the process of the row chosen requests all the resources it needs (which is guaranteed to be possible) and finishes. Mark that process as terminated and add all its resources to the *A* vector.

3. Repeat steps 1 and 2 until either all processes are marked terminated, in which case the initial state was safe, or until a deadlock occurs, in which case it was not.

If several processes are eligible to be chosen in step 1, it does not matter which one is selected: the pool of available resources either gets larger or stays the same.

Now let us get back to the example of Fig. 3-15. The current state is safe. Suppose that process *B* now requests a printer. This request can be granted because the resulting state is still safe (process *D* can finish, and then processes *A* or *E*, followed by the rest).

Now imagine that after giving *B* one of the two remaining printers, *E* wants the last printer. Granting that request would reduce the vector of available resources to (1 0 0 0), which leads to deadlock. Clearly *E*'s request must be deferred for a while.

The banker's algorithm was first published by Dijkstra in 1965. Since that time, nearly every book on operating systems has described it in detail. Innumerable papers have been written about various aspects of it. Unfortunately, few authors have had the audacity to point out that although in theory the algorithm is wonderful, in practice it is essentially useless because processes rarely know in advance what their maximum resource needs will be. In addition, the number of processes is not fixed, but dynamically varying as new users log in and out. Furthermore, resources that were thought to be available can suddenly vanish (tape drives can break). Thus in practice, few, if any, existing systems use the banker's algorithm for avoiding deadlocks.

In summary, the schemes described earlier under the name "prevention" are overly restrictive, and the algorithm described here as "avoidance" requires information that is usually not available. If you can think of a general-purpose algorithm that does the job in practice as well as in theory, write it up and send it to your local computer science journal.

Although both avoidance and prevention are not terribly promising in the general case, for specific applications, many excellent special-purpose algorithms are known. As an example, in many database systems, an operation that occurs frequently is requesting locks on several records and then updating all the locked records. When multiple processes are running at the same time, there is a real danger of deadlock. To eliminate this problem, special techniques are used.

The approach most often used is called **two-phase locking**. In the first phase, the process tries to lock all the records it needs, one at a time. If it succeeds, it begins the second phase, performing its updates and releasing the locks. No real work is done in the first phase.

If during the first phase, some record is needed that is already locked, the process just releases all its locks and starts the first phase all over. In a certain sense, this approach is similar to requesting all the resources needed in advance, or at least before anything irreversible is done. In some versions of two-phase locking, there is no release and restart if a lock is encountered during the first phase. In these versions, deadlock can occur.

However, this strategy is not applicable in general. In real-time systems and process control systems, for example, it is not acceptable to just terminate a process partway through because a resource is not available and start all over again. Neither is it acceptable to start over if the process has read or written messages to the network, updated files, or anything else that cannot

PREV NEXT

PREV
NEXT

3.4. Overview of I/O in MINIX 3

MINIX 3 I/O is structured as shown in Fig. 3-8 . The top four layers of that figure correspond to the four-layered structure of MINIX 3 shown in Fig. 2-29 . In the following sections we will look briefly at each of the layers, with an emphasis on the device drivers. Interrupt handling was covered in Chap. 2 and the device-independent I/O will be discussed when we come to the file system, in Chap. 5 .

3.4.1. Interrupt Handlers and I/O Access in MINIX 3

Many device drivers start some I/O device and then block, waiting for a message to arrive. That message is usually generated by the interrupt handler for the device. Other device drivers do not start any physical I/O (e.g., reading from RAM disk and writing to a memory-mapped display), do not use interrupts, and do not wait for a message from an I/O device. In the previous chapter the mechanisms in the kernel by which interrupts generate messages and cause task switches has been presented in great detail, and we will say no more about it here. Here we will discuss in a general way interrupts and I/O in device drivers. We will return to the details when we look at the code for various devices.

For disk devices, input and output is generally a matter of commanding a device to perform its operation, and then waiting until the operation is complete. The disk controller does most of the work, and very little is required of the interrupt handler. Life would be simple if all interrupts could be handled so easily.

[Page 253]

However, there is sometimes more for the low-level handler to do. The message passing mechanism has a cost. When an interrupt may occur frequently but the amount of I/O handled per interrupt is small, it may pay to make the handler itself do somewhat more work and to postpone sending a message to the driver until a subsequent interrupt, when there is more for the driver to do. In MINIX 3 this is not possible for most I/O, because the low level handler in the kernel is a general purpose routine used for almost all devices.

In the last chapter we saw that the clock is an exception. Because it is compiled with the kernel the clock can have its own handler that does extra work. On many clock ticks there is very little to be done, except for maintaining the time. This is done without sending a message to the clock task itself. The clock's interrupt handler increments a variable, appropriately named *realtime* , possibly adding a correction for ticks counted during a BIOS call. The handler does some additional very simple arithmeticit increments counters for user time and billing time, decrements the *ticks _left* counter for the current process, and tests to see if a timer has expired. A message is sent to the clock task only if the current process has used up its quantum or a timer has expired.

The clock interrupt handler is unique in MINIX 3, because the clock is the only interrupt driven device that runs in kernel space. The clock hardware is integral to the PCin fact, the clock interrupt line does not connect to any pin on the sockets where add-on I/O controllers can be plugged inso it is impossible to install a clock upgrade package with replacement clock hardware and a driver provided by the manufacturer. It is reasonable, then, for the clock driver to be compiled into the kernel and have access to any variable in kernel space. But a key design goal of MINIX 3 is to make it unnecessary for any other device driver to have that kind of access.

Device drivers that run in user space cannot directly access kernel memory or I/O ports. Although possible, it would also violate the design principles of MINIX 3 to allow an interrupt service routine to make a far call to execute a service routine within the text segment of a user process. This would be even more dangerous than letting a user space process call a function within kernel space. In that case we would at least be sure the function was written by a competent, security-aware operating system designer, possibly one who had read this book. But the kernel should not trust code provided by a user program.

There are several different levels of I/O access that might be needed by a user-space device driver.

1. A driver might need access to memory outside its normal data space. The memory driver, which manages the RAM disk, is an example of a driver which needs only this kind of access.
2. A driver may need to read and write to I/O ports. The machine-level instructions for these operations are available only in kernel mode. As we will soon see, the hard disk driver needs this kind of access.

3. A driver may need to respond to predictable interrupts. For example, the hard disk driver writes commands to the disk controller, which causes an interrupt to occur when the desired operation is complete.
4. A driver may need to respond to unpredictable interrupts. The keyboard driver is in this category. This could be considered a subclass of the preceding item, but unpredictability complicates things.

All of these cases are supported by kernel calls handled by the system task.

The first case, access to extra memory segments, takes advantage of the hardware segmentation support provided by Intel processors. Although a normal process has access only to its own text, data, and stack segments, the system task allows other segments to be defined and accessed by user-space processes. Thus the memory driver can access a memory region reserved for use as a RAM disk, as well as other regions designated for special access. The console driver accesses memory on a video display adapter in the same way.

For the second case, MINIX 3 provides kernel calls to use I/O instructions. The system task does the actual I/O on behalf of a less-privileged process. Later in this chapter we will see how the hard disk driver uses this service. We will present a preview here. The disk driver may have to write to a single output port to select a disk, then read from another port in order to verify the device is ready. If response is normally expected to be very quick, polling can be done. There are kernel calls to specify a port and data to be written or a location for receipt of data read. This requires that a call to read a port be nonblocking, and in fact, kernel calls do not block.

Some insurance against device failure is useful. A polling loop could include a counter that terminates the loop if the device does not become ready after a certain number of iterations. This is not a good idea in general because the loop execution time will depend upon the CPU speed. One way around this is to start the counter with a value that is related to CPU time, possibly using a global variable initialized when the system starts. A better way is provided by the MINIX 3 system library, which provides a *getuptime* function. This uses a kernel call to retrieve a counter of clock ticks since system startup maintained by the clock task. The cost of using this information to keep track of time spent in a loop is the overhead of an additional kernel call on each iteration. Another possibility is to ask the system task to set a watchdog timer. But to receive a notification from a timer a `receive` operation, which will block, is required. This is not a good solution if a fast response is expected.

The hard disk also makes use of variants of the kernel calls for I/O that make it possible to send a list of ports and data to write or variables to be altered to the system task. This is very usefulthe hard disk driver we will examine requires writing a sequence of byte values to seven output ports to initiate an operation. The last byte in the sequence is a command, and the disk controller generates an interrupt when it completes a command. All this can be accomplished with a single kernel call, greatly reducing the number of messages needed.

This brings us to the third item in the list: responding to an expected interrupt. As noted in the discussion of the system task, when an interrupt is initialized on behalf of a user space program (using a `sys_irqctl` kernel call), the handler routine for the interrupt is always *generic _handler* , a function defined as part of the system task. This routine converts the interrupt into a notification message to the process on whose behalf the interrupt was set. The device driver therefore must initiate a `receive` operation after the kernel call that issues the command to the controller. When the notification is received the device driver can proceed to do what must be done to service the interrupt.

Although in this case an interrupt is expected, it is prudent to hedge against the possibility that something might go wrong sometime. To prepare for the possibility that the interrupt might fail to be triggered, a process can request the system task to set up a watchdog timer. Watchdog timers also generate notification messages, and thus the `receive` operation could get a notification either because an interrupt occurred or because a timer expired. This is not a problem because, although a notification does not convey much information, the notification message indicates its origin. Although both notifications are generated by the system task, notification of an interrupt will appear to come from *HARDWARE* , and notification of a timer expiring will appear to come from *CLOCK* .

There is another problem. If an interrupt is received in a timely way and a watchdog timer has been set, expiration of the timer at some future time will be detected by another `receive` operation, possibly in the main loop of the driver. One solution is to make a kernel call to disable the timer when the notification from *HARDWARE* is received. Alternatively, if it is likely that the next `receive` operation will be one where a message from *CLOCK* is not expected, such a message could be ignored and `receive` called again. Although less likely, it is conceivable that a disk operation could occur after an unexpectedly long delay, generating the interrupt only after the watchdog has timed out. The same solutions apply here. When a timeout occurs a kernel call can be made to disable an interrupt, or a `receive` operation that does not expect an interrupt could ignore any message from *HARDWARE* .

This is a good time to mention that when an interrupt is first enabled, a kernel call can be made to set a "policy" for the interrupt. The policy is simply a flag that determines whether the interrupt should be automatically reenabled or whether it should remain disabled until the device driver it serves makes a kernel call to reenable it. For the disk driver there may be a substantial amount of work to be done after an interrupt, and thus it may be best to leave the interrupt disabled until all data has been copied.

The fourth item in our list is the most problematic. Keyboard support is part of the tty driver, which provides output as well as input. Furthermore, multiple devices may be supported. So input may come from a local keyboard, but it can also come from a remote user connected by a serial line or a network connection. And several processes may be running, each producing output for a different local or remote terminal. When you do not know when, if ever, an interrupt might occur, you cannot just make a blocking `receive` call to accept input from a single source if the same process may need to respond to other input and output sources.

MINIX 3 uses several techniques to deal with this problem. The principal technique used by the terminal driver for dealing with keyboard input is to make the interrupt response as fast as possible, so characters will not be lost. The minimum possible amount of work is done to get characters from the keyboard hardware to a buffer. Additionally, when data has been fetched from the keyboard in response to an interrupt, as soon as the data is buffered the keyboard is read again before returning from the interrupt. Interrupts generate notification messages, which do not block the sender; this helps to prevent loss of input. A nonblocking `receive` operation is available, too, although it is only used to handle messages during a system crash. Watchdog timers are also used to activate the routine that checks the keyboard.

3.4.2. Device Drivers in MINIX 3

For each class of I/O device present in a MINIX 3 system, a separate I/O device driver is present. These drivers are full-fledged processes, each one with its own state, registers, stack, and so on. Device drivers communicate with the file system using the standard message passing mechanism used by all MINIX 3 processes. A simple device driver may be written as a single source file. For the RAM disk, hard disk, and floppy disk there is a source file to support each type of device, as well as a set of common routines in *driver.c* and *drvlib.c* to support all blcok device types. This separation of the hardware-dependent and hardware-independent parts of the software makes for easy adaptation to a variety of different hardware configurations. Although some common source code is used, the driver for each disk type runs as a separate process, in order to support rapid data transfers and isolate drivers from each other.

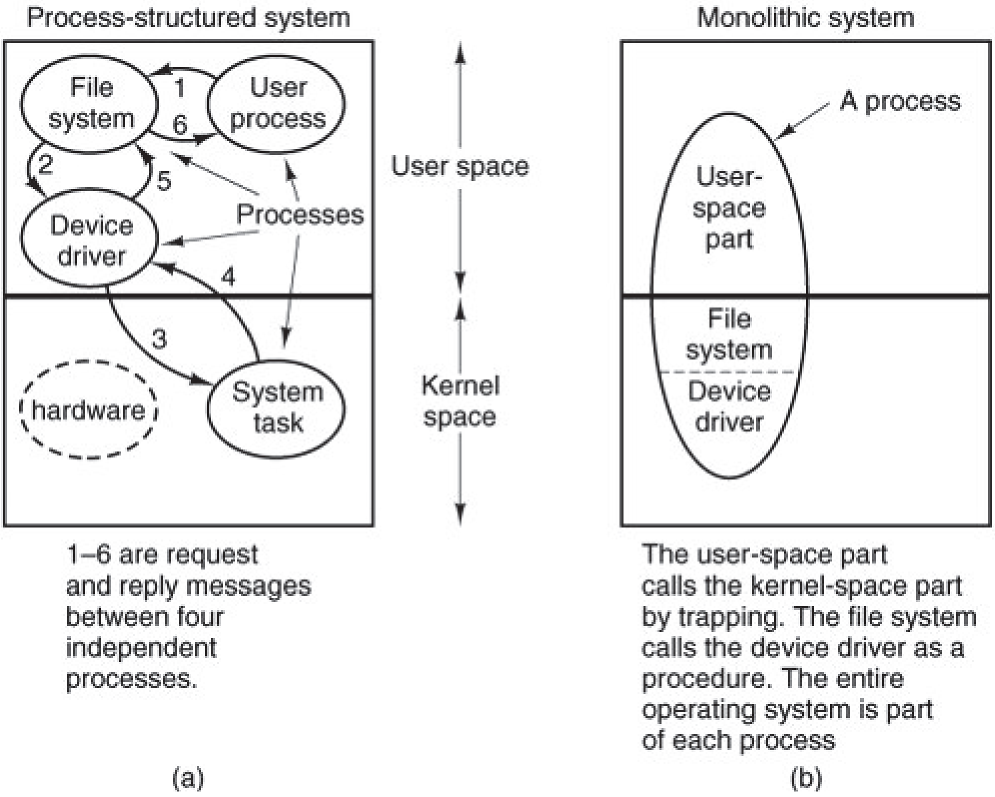
The terminal driver source code is organized in a similar way, with the hardware-independent code in *tty.c* and source code to support different devices, such as memory-mapped consoles, the keyboard, serial lines, and pseudo terminals in separate files. In this case, however, a single process supports all of the different device types.

For groups of devices such as disk devices and terminals, for which there are several source files, there are also header files. *Driver.h* supports all the block device drivers. *Tty.h* provides common definitions for all the terminal devices.

The MINIX 3 design principle of running components of the operating system as completely separate processes in user space is highly modular and moderately efficient. It is also one of the few places where MINIX 3 differs from UNIX in an essential way. In MINIX 3 a process reads a file by sending a message to the file system process. The file system, in turn, may send a message to the disk driver asking it to read the needed block. The disk driver uses kernel calls to ask the system task to do the actual I/O and to copy data between processes. This sequence (slightly simplified from reality) is shown in Fig. 3-16(a) . By making these interactions via the message mechanism, we force various parts of the system to interface in standard ways with other parts.

Figure 3-16. Two ways of structuring user-system communication.

[View full size image]



Process-structured system
File system
User process
Device driver
Processes
System task
hardware
1
2
3
4
5
6
User space
Kernel space
Monolithic system
A process
User-space part
File system
Device driver
1–6 are request and reply messages between four independent processes.
(a)
The user-space part calls the kernel-space part by trapping. The file system calls the device driver as a procedure. The entire operating system is part of each process
(b)

In UNIX all processes have two parts: a user-space part and a kernel-space part, as shown in Fig. 3-16(b) . When a system call is made, the operating system switches from the user-space part to the kernel-space part in a somewhat magical way. This structure is a remnant of the MULTICS design, in which the switch was just an ordinary procedure call, rather than a trap followed by saving the state of the user-part, as it is in UNIX.

Device drivers in UNIX are simply kernel procedures that are called by the kernel-space part of the process. When a driver needs to wait for an interrupt, it calls a kernel procedure that puts it to sleep until some interrupt handler wakes it up. Note that it is the user process itself that is being put to sleep here, because the kernel and user parts are really different parts of the same process.

Among operating system designers, arguments about the merits of monolithic systems, as in UNIX, versus process-structured systems, as in MINIX 3, are endless. The MINIX 3 approach is better structured (more modular), has cleaner interfaces between the pieces, and extends easily to distributed systems in which the various processes run on different computers. The UNIX approach is more efficient, because procedure calls are much faster than sending messages. MINIX 3 was split into many processes because we believe that with increasingly powerful personal computers available, cleaner software structure was worth making the system slightly slower. The performance loss due to having most of the operating system run in user space is typically in the range of 510%. Be warned that some operating system designers do not share the belief that it is worth sacrificing a little speed for a more modular and more reliable system.

In this chapter, drivers for RAM disk, hard disk, clock, and terminal are discussed. The standard MINIX 3 configuration also includes drivers for the floppy disk and the printer, which are not discussed in detail. The MINIX 3 software distribution contains source code for additional drivers for RS-232 serial lines, CD-ROMs, various Ethernet adapter, and sound cards. These may be compiled separately and started on the fly at any time.

All of these drivers interface with other parts of the MINIX 3 system in the same way: request messages are sent to the drivers. The messages contain a variety of fields used to hold the operation code (e.g., *READ* or *WRITE*) and its parameters. A driver attempts to fulfill a request and returns a reply message.

For block devices, the fields of the request and reply messages are shown in Fig. 3-17 . The request message includes the address of a buffer area containing data to be transmitted or in which received data are expected. The reply includes status information so the requesting process can verify that its request was properly carried out. The fields for the character devices are basically similar but can vary slightly from driver to driver. Messages to the terminal driver can contain the address of a data structure which specifies all of the many configurable aspects of a terminal, such as the characters to use for the intraline editing functions erase-character and kill-line.

Figure 3-17. Fields of the messages sent by the file system to the block device drivers and fields of the replies sent back.

(This item is displayed on page 259 in the print version)

Field	**Type**	**Meaning**
m.m_type	int	Operation requested
m.DEVICE	int	Minor device to use
m.PROC_NR	int	Process requesting the I/O
m.COUNT	int	Byte count or ioctl code
m.POSITION	long	Position on device

m.ADDRESS	char*	Address within requesting process

Requests

Field	**Type**	**Meaning**
m.m_type	int	Always DRIVER_REPLY
m.REP_PROC_NR	int	Same as PROC_NR in request
m.REP_STATUS	int	Bytes transferred or error number

Replies

The function of each driver is to accept requests from other processes, normally the file system, and carry them out. All the block device drivers have been written to get a message, carry it out, and send a reply. Among other things, this decision means that these drivers are strictly sequential and do not contain any internal multiprogramming, to keep them simple. When a hardware request has been issued, the driver does a `receive` operation specifying that it is interested only in accepting interrupt messages, not new requests for work. Any new request messages are just kept waiting until the current work has been done (rendezvous principle). The terminal driver is slightly different, since a single driver services several devices. Thus, it is possible to accept a new request for input from the keyboard while a request to read from a serial line is still being fulfilled. Nevertheless, for each device a request must be completed before beginning a new one.

The main program for each block device driver is structurally the same and is outlined in Fig. 3-18 . When the system first comes up, each one of the drivers is started up in turn to give each a chance to initialize internal tables and similar things. Then each device driver blocks by trying to get a message. When a message comes in, the identity of the caller is saved, and a procedure is called to carry out the work, with a different procedure invoked for each operation available. After the work has been finished, a reply is sent back to the caller, and the driver then goes back to the top of the loop to wait for the next request.

Figure 3-18. Outline of the main procedure of an I/O device driver.

(This item is displayed on page 260 in the print version)

```
message mess;                           /* message buffer*/

void io_driver() {
  initialize();                         /* only done once, during system init.*/
  while (TRUE)  {
        receive(ANY, &mess);            /* wait for a request for work*/
        caller = mess.source;           /* process from whom message came*/
        switch(mess.type) {
            case READ:       rcode = dev_read(&mess); break;
            case WRITE:      rcode = dev_write(&mess); break;
            /* Other cases go here, including OPEN, CLOSE, and IOCTL*/
             default:        rcode = ERROR;
        }
        mess.type = DRIVER_REPLY;
        mess.status = rcode;            /* result code*/
        send(caller,&mess);             /* send reply message back to caller*/
  }
}
```

Each of the *dev _XXX* procedures handles one of the operations of which the driver is capable. It returns a status code telling what happened. The status code, which is included in the reply message as the field *REP _STATUS* , is the count of bytes transferred (zero or positive) if all went well, or the error number (negative) if something went wrong. This count may differ from the number of bytes requested. When the end of a file is reached, the number of bytes available may be less than number requested. On terminals at most one line is returned (except in raw mode), even if the count requested is larger.

3.4.3. Device-Independent I/O Software in MINIX 3

In MINIX 3 the file system process contains all the device-independent I/O code. The I/O system is so closely related to the file system that they were merged into one process. The functions performed by the file system are those shown in Fig. 3-6 , except for requesting and releasing dedicated devices, which do not exist in MINIX 3 as it is presently configured. They could, however, easily be added to the relevant device drivers should the need arise in the future.

In addition to handling the interface with the drivers, buffering, and block allocation, the file system also handles protection and the management of i-nodes, directories, and mounted file systems. This will be covered in detail in Chap. 5 .

3.4.4. User-Level I/O Software in MINIX 3

The general model outlined earlier in this chapter also applies here. Library procedures are available for making system calls and for all the C functions required by the POSIX standard, such as the formatted input and output functions *printf* and *scanf* . The standard MINIX 3 configuration contains one spooler daemon, `lpd` , which spools and prints files passed to it by the `lp` command. The standard MINIX 3 software distribution also provides a number of daemons that support various network functions. The MINIX 3 configuration described in this book supports most network operations, all that is needed is to enable the network server and drivers for ethernet adapters at startup time. Recompiling the terminal driver with pseudo terminals and serial line support will add support for logins from remote terminals and networking over serial lines (including modems). The network server runs at the same priority as the memory manager and the file system, and like them, it runs as a user process.

3.4.5. Deadlock Handling in MINIX 3

True to its heritage, MINIX 3 follows the same path as UNIX with respect to deadlocks of the types described earlier in this chapter: it just ignores the problem. Normally, MINIX 3 does not contain any dedicated I/O devices, although if someone wanted to hang an industry standard DAT tape drive on a PC, making the software for it would not pose any special problems. In short, the only place deadlocks can occur are with the implicit shared resources, such as process table slots, i-node table slots, and so on. None of the known deadlock algorithms can deal with resources like these that are not requested explicitly.

Actually, the above is not strictly true. Accepting the risk that user processes could deadlock is one thing, but within the operating system itself a few places do exist where considerable care has been taken to avoid problems. The main one is the message-passing interaction between processes. For instance, user processes are only allowed to use the `sendrec` messaging method, so a user process should never lock up because it did a `receive` when there was no process with an interest in `send` ing to it. Servers only use `send` or `sendrec` to communicate with device drivers, and device drivers only use `send` or `sendrec` to communicate with the system task in the kernel layer. In the rare case where servers must communicate between themselves, such as exchanges between the process manager and the file system as they initialize their parts of the process table, the order of communication is very carefully designed to avoid deadlock. Also, at the very lowest level of the message passing system there is a check to make sure that when a process is about to do a send that the destination process is not trying to the same thing.

In addition to the above restrictions, in MINIX 3 the new `notify` message primitive is provided to handle those situations in which a message must be sent in the "upstream" direction. `Notify` is nonblocking, and notifications are stored when a recipient is not immediately available. As we examine the implementation of MINIX 3 device drivers in this chapter we will see that `notify` is used extensively.

Locks are another mechanism that can prevent deadlocks. It is possible to lock devices and files even without operating system support. A file name can serve as a truly global variable, whose presence or absence can be noted by all other processes. A special directory, */usr/spool/locks/,* is usually present on MINIX 3 systems, as on most UNIX-like systems, where processes can create **lock files** , to mark any resources they are using. The MINIX 3 file system also supports POSIX-style advisory file locking. But neither of these mechanisms is enforceable. They depend upon the good behavior of processes, and there is nothing to prevent a program from trying to use a resource that is locked by another process. This is not exactly the same thing as preemption of the resource, because it does not prevent the first process from attempting to continue its use of the resource. In other words, there is no mutual exclusion. The result of such an action by an ill-behaved process is likely to be a mess, but no deadlock results.

PREV

NEXT

PREV NEXT

5. File Systems

All computer applications need to store and retrieve information. While a process is running, it can store a limited amount of information within its own address space. However, the storage capacity is restricted to the size of the virtual address space. For some applications this size is adequate, but for others, such as airline reservations, banking, or corporate record keeping, it is far too small.

A second problem with keeping information within a process' address space is that when the process terminates, the information is lost. For many applications, (e.g., for databases), the information must be retained for weeks, months, or even forever. Having it vanish when the process using it terminates is unacceptable. Furthermore, it must not go away when a computer crash kills the process.

A third problem is that it is frequently necessary for multiple processes to access (parts of) the information at the same time. If we have an online telephone directory stored inside the address space of a single process, only that process can access it. The way to solve this problem is to make the information itself independent of any one process.

Thus we have three essential requirements for long-term information storage:

1. It must be possible to store a very large amount of information.

2. The information must survive the termination of the process using it.

3. Multiple processes must be able to access the information concurrently.

The usual solution to all these problems is to store information on disks and other external media in units called **files**. Processes can then read them and write new ones if need be. Information stored in files must be **persistent**, that is, not be affected by process creation and termination. A file should only disappear when its owner explicitly removes it.

Files are managed by the operating system. How they are structured, named, accessed, used, protected, and implemented are major topics in operating system design. As a whole, that part of the operating system dealing with files is known as the **file system** and is the subject of this chapter.

From the users' standpoint, the most important aspect of a file system is how it appears to them, that is, what constitutes a file, how files are named and protected, what operations are allowed on files, and so on. The details of whether linked lists or bitmaps are used to keep track of free storage and how many sectors there are in a logical block are of less interest, although they are of great importance to the designers of the file system. For this reason, we have structured the chapter as several sections. The first two are concerned with the user interface to files and directories, respectively. Then comes a discussion of alternative ways a file system can be implemented. Following a discussion of security and protection mechanisms, we conclude with a

5.1. Files

In the following pages we will look at files from the user's point of view, that is, how they are used and what properties they have.

5.1.1. File Naming

Files are an abstraction mechanism. They provide a way to store information on the disk and read it back later. This must be done in such a way as to shield the user from the details of how and where the information is stored, and how the disks actually work.

Probably the most important characteristic of any abstraction mechanism is the way the objects being managed are named, so we will start our examination of file systems with the subject of file naming. When a process creates a file, it gives the file a name. When the process terminates, the file continues to exist and can be accessed by other processes using its name.

The exact rules for file naming vary somewhat from system to system, but all current operating systems allow strings of one to eight letters as legal file names. Thus *andrea*, *bruce*, and *cathy* are possible file names. Frequently digits and special characters are also permitted, so names like *2*, *urgent!*, and *Fig. 2-14* are often valid as well. Many file systems support names as long as 255 characters.

Some file systems distinguish between upper- and lower-case letters, whereas others do not. UNIX (including all its variants) falls in the first category; MS-DOS falls in the second. Thus a UNIX system can have all of the following as three distinct files: *maria*, *Maria*, and *MARIA*. In MS-DOS, all these names refer to the same file.

Windows falls in between these extremes. The Windows 95 and Windows 98 file systems are both based upon the MS-DOS file system, and thus inherit many of its properties, such as how file names are constructed. With each new version improvements were added but the features we will discuss are mostly common to MS-DOS and "classic" Windows versions. In addition, Windows NT, Windows 2000, and Windows XP support the MS-DOS file system. However, the latter systems also have a native file system (**NTFS**) that has different properties (such as file names in Unicode). This file system also has seen changes in successive versions. In this chapter, we will refer to the older systems as the Windows 98 file system. If a feature does not apply to the MS-DOS or Windows 95 versions we will say so. Likewise, we will refer to the newer system as either NTFS or the Windows XP file system, and we will point it out if an aspect under discussion does not also apply to the file systems of Windows NT or Windows 2000. When we say just Windows, we mean all Windows file systems since Windows 95.

Many operating systems support two-part file names, with the two parts separated by a period, as in *prog.c.* The part following the period is called the **file extension** and usually indicates something about the file, in this example that it is a C programming language source file. In MS-DOS, for example, file names are 1 to 8 characters, plus an optional extension of 1 to 3 characters. In UNIX, the size of the extension, if any, is up to the user, and a file may even have two or more extensions, as in *prog.c.bz2,* where *.bz2* is commonly used to indicate that the file (*prog.c)* has been compressed using the bzip2 compression algorithm. Some of the more common file extensions and their meanings are shown in Fig. 5-1

Figure 5-1. Some typical file extensions.

(This item is displayed on page 484 in the print version)

Extension	Meaning
file.bak	Backup file
file.c	C source program
file.gif	Graphical Interchange Format image
file.html	World Wide Web HyperText Markup Language document
file.iso	ISO image of a CD-ROM (for burning to CD)
file.jpg	Still picture encoded with the JPEG standard
file.mp3	Music encoded in MPEG layer 3 audio format
file.mpg	Movie encoded with the MPEG standard
file.o	Object file (compiler output, not yet linked)
file.pdf	Portable Document Format file
file.ps	PostScript file
file.tex	Input for the TEX formatting program
file.txt	General text file
file.zip	Compressed archive

In some systems (e.g., UNIX), file extensions are just conventions and are not enforced by the operating system. A file named *file.txt* might be some kind of text file, but that name is more to remind the owner than to convey any actual information to the computer. On the other hand, a C compiler may actually insist that files it is to compile end in *.c,* and it may refuse to compile them if they do not.

Conventions like this are especially useful when the same program can handle several different kinds of files. The C compiler, for example, can be given a list of files to compile and link together, some of them C files (e.g., *foo.c),* some of them assembly language files (e.g., *bar.s),* and some of them object files (e.g., *other.o).* The extension then becomes essential for the compiler to tell which are C files, which are assembly files, and which are object files.

In contrast, Windows is very much aware of the extensions and assigns meaning to them. Users (or processes) can register extensions with the operating system and specify which program "owns" which one. When a user double clicks on a file name, the program assigned to its file extension is launched and given the name of the file as parameter. For example, double clicking on *file.doc* starts Microsoft *Word* with *file.doc* as the initial file to edit.

Some might think it odd that Microsoft chose to make common extensions invisible by default since they are so important. Fortunately most of the "wrong by default" settings of Windows can be changed by a sophisticated user who knows where to look.

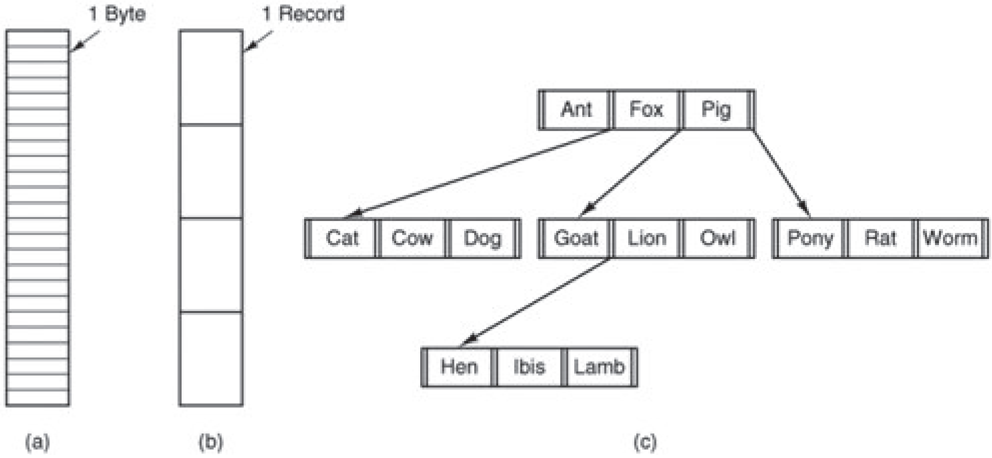
5.1.2. File Structure

Files can be structured in any one of several ways. Three common possibilities are depicted in Fig. 5-2. The file in Fig. 5-2(a) is just an unstructured sequence of bytes. In effect, the operating system does not know or care what is in the file. All it sees are bytes. Any meaning must be imposed by user-level programs. Both UNIX and Windows 98 use this approach.

Figure 5-2. Three kinds of files. (a) Byte sequence. (b) Record sequence. (c) Tree.

(This item is displayed on page 485 in the print version)

[View full size image]



Having the operating system regard files as nothing more than byte sequences provides the maximum flexibility. User programs can put anything they want in their files and name them any way that is convenient. The operating system does not help, but it also does not get in the way. For users who want to do unusual things, the latter can be very important.

The first step up in structure is shown in Fig. 5-2(b). In this model, a file is a sequence of fixed-length records, each with some internal structure. Central to the idea of a file being a sequence of records is the idea that the read operation returns one record and the write operation overwrites or appends one record. As a historical note, when the 80-column punched card was king many (mainframe) operating systems based their file systems on files consisting of 80-character records, in effect, card images. These systems also supported files of 132-character records, which were intended for the line printer (which in those days were big chain printers having 132 columns). Programs read input in units of 80 characters and wrote it in units of 132 characters, although the final 52 could be spaces, of course. No current general-purpose system works this way.

The third kind of file structure is shown in Fig. 5-2(c). In this organization, a file consists of a tree of records, not necessarily all the same length, each containing a **key** field in a fixed position in the record. The tree is sorted on the key field, to allow rapid searching for a particular key.

The basic operation here is not to get the "next" record, although that is also possible, but to get the record with a specific key. For the zoo file of Fig. 5-2(c), one could ask the system to get the record whose key is *pony*, for example, without worrying about its exact position in the file. Furthermore, new records can be added to the file, with the operating system, and not the user, deciding where to place them. This type of file is clearly quite different from the unstructured byte streams used in UNIX and Windows 98 but is widely used on the large mainframe computers still used in some commercial data processing.

5.1.3. File Types

Many operating systems support several types of files. UNIX and Windows, for example, have regular files and directories. UNIX also has character and block special files. Windows XP also uses **metadata** files, which we will mention later. **Regular files** are the ones that contain user information. All the files of Fig. 5-2 are regular files. **Directories** are system files for maintaining the structure of the file system. We will study directories below. **Character special files** are related to input/output and used to model serial I/O devices such as terminals, printers, and networks. **Block special files** are used to model disks. In this chapter we will be primarily interested in regular files.

Regular files are generally either ASCII files or binary files. ASCII files consist of lines of text. In some systems each line is terminated by a carriage return character. In others, the line feed character is used. Some systems (e.g., Windows) use both. Lines need not all be of the same length.

The great advantage of ASCII files is that they can be displayed and printed as is, and they can be edited with any text editor. Furthermore, if large numbers of programs use ASCII files for input and output, it is easy to connect the output of one program to the input of another, as in shell pipelines. (The interprocess plumbing is not any easier, but interpreting the information certainly is if a standard convention, such as ASCII, is used for expressing it.)

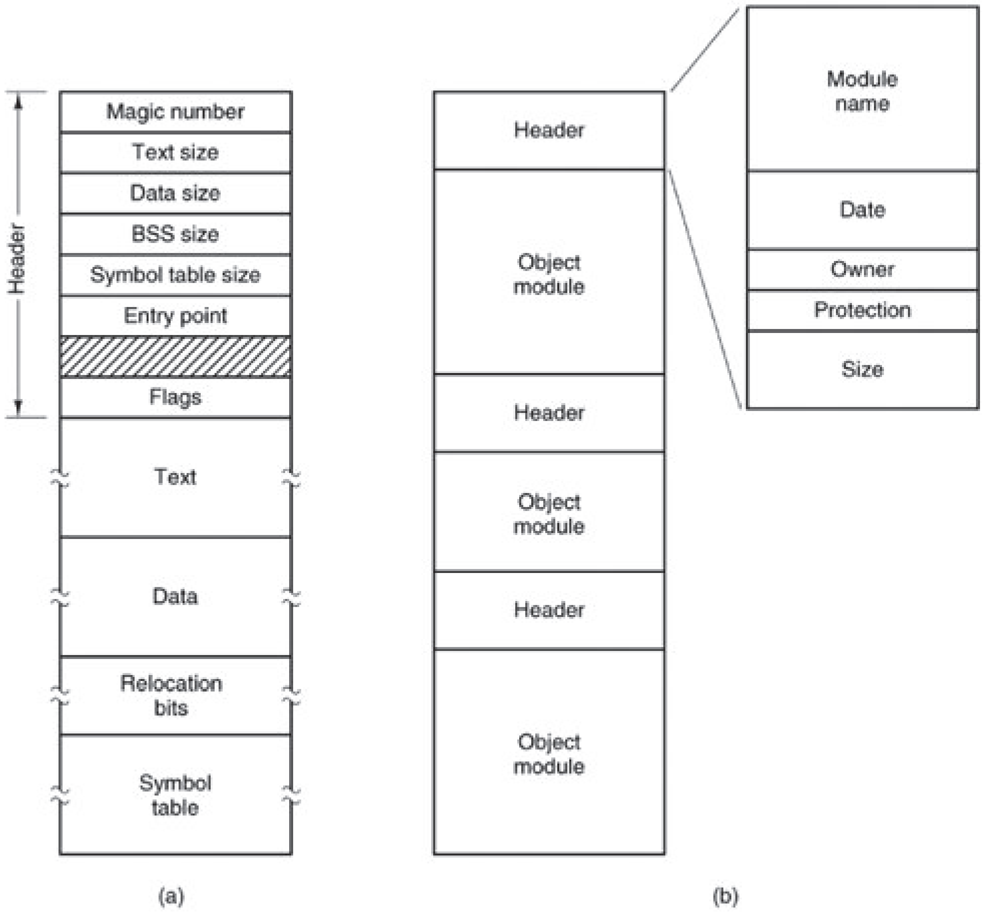
Other files are binary files, which just means that they are not ASCII files. Listing them on the printer gives an incomprehensible listing full of what is apparently random junk. Usually, they have some internal structure known to programs that use them.

For example, in Fig. 5-3(a) we see a simple executable binary file taken from an early version of UNIX. Although technically the file is just a sequence of bytes, the operating system will only execute a file if it has the proper format. It has five sections: header, text, data, relocation bits, and symbol table. The header starts with a so-called **magic number**, identifying the file as an executable file (to prevent the accidental execution of a file not in this format). Then come the sizes of the various pieces of the file, the address at which execution starts, and some flag bits. Following the header are the text and data of the program itself. These are loaded into memory and relocated using the relocation bits. The symbol table is used for debugging.

Figure 5-3. (a) An executable file. (b) An archive.

(This item is displayed on page 487 in the print version)

[View full size image]



Our second example of a binary file is an archive, also from UNIX. It consists of a collection of library procedures (modules) compiled but not linked. Each one is prefaced by a header telling its name, creation date, owner, protection code, and size. Just as with the executable file, the module headers are full of binary numbers. Copying them to the printer would produce complete gibberish.

Every operating system must recognize at least one file type: its own executable file, but some operating systems recognize more. The old TOPS-20 system (for the DECsystem 20) went so far as to examine the creation time of any file to be executed. Then it located the source file and saw if the source had been modified since the binary was made. If it had been, it automatically recompiled the source. In UNIX terms, the *make* program had been built into the shell. The file extensions were mandatory so the operating system could tell which binary program was derived from which source.

Having strongly typed files like this causes problems whenever the user does anything that the system designers did not expect. Consider, as an example, a system in which program output files have extension *.dat* (data files). If a user writes a program formatter that reads a *.c* file (C program), transforms it (e.g., by converting it to a standard indentation layout), and then writes the transformed file as output, the output file will be of type *.dat.* If the user tries to offer this to the C compiler to compile it, the system will refuse because it has the wrong extension. Attempts to copy *file.dat* to *file.c* will be rejected by the system as invalid (to protect the user against mistakes).

While this kind of "user friendliness" may help novices, it drives experienced users up the wall since they have to devote considerable effort to circumventing the operating system's idea of what is reasonable and what is not.

5.1.4. File Access

Early operating systems provided only a single kind of file access: **sequential access**. In these systems, a process could read all the bytes or records in a file in order, starting at the beginning, but could not skip around and read them out of order. Sequential files could be rewound, however, so they could be read as often as needed. Sequential files were convenient when the storage medium was magnetic tape, rather than disk.

When disks came into use for storing files, it became possible to read the bytes or records of a file out of order, or to access records by key, rather than by position. Files whose bytes or records can be read in any order are called **random access files**. They are required by many applications.

Random access files are essential for many applications, for example, database systems. If an airline customer calls up and wants to reserve a seat on a particular flight, the reservation program must be able to access the record for that flight without having to read the records for thousands of other flights first.

Two methods are used for specifying where to start reading. In the first one, every `read` operation gives the position in the file to start reading at. In the second one, a special operation, `seek`, is provided to set the current position. After a `seek`, the file can be read sequentially from the now-current position.

In some older mainframe operating systems, files are classified as being either sequential or random access at the time they are created. This allows the system to use different storage techniques for the two classes. Modern operating systems do not make this distinction. All their files are automatically random access.

5.1.5. File Attributes

Every file has a name and its data. In addition, all operating systems associate other information with each file, for example, the date and time the file was created and the file's size. We will call these extra items the file's **attributes** although some people called them **metadata**. The list of attributes varies considerably from system to system. The table of Fig. 5-4 shows some of the possibilities, but others also exist. No existing system has all of these, but each is present in some system.

Figure 5-4. Some possible file attributes.

(This item is displayed on page 489 in the print version)

Attribute	**Meaning**
Protection	Who can access the file and in what way
Password	Password needed to access the file
Creator	ID of the person who created the file
Owner	Current owner
Read-only flag	0 for read/write; 1 for read only
Hidden flag	0 for normal; 1 for do not display in listings
System flag	0 for normal files; 1 for system file
Archive flag	0 for has been backed up; 1 for needs to be backed up
ASCII/binary flag	0 for ASCII file; 1 for binary file
Random access flag	0 for sequential access only; 1 for random access
Temporary flag	0 for normal; 1 for delete file on process exit
Lock flags	0 for unlocked; nonzero for locked
Record length	Number of bytes in a record
Key position	Offset of the key within each record
Key length	Number of bytes in the key field
Creation time	Date and time the file was created
Time of last access	Date and time the file was last accessed
Time of last change	Date and time the file has last changed
Current size	Number of bytes in the file
Maximum size	Number of bytes the file may grow to

The first four attributes relate to the file's protection and tell who may access it and who may not. All kinds of schemes are possible, some of which we will study later. In some systems the user must present a password to access a file, in which case the password must be one of the attributes.

The flags are bits or short fields that control or enable some specific property. Hidden files, for example, do not appear in listings of the files. The archive flag is a bit that keeps track of whether the file has been backed up. The backup program clears it, and the operating system sets it whenever a file is changed. In this way, the backup program can tell which files need backing up. The temporary flag allows a file to be marked for automatic deletion when the process that created it terminates.

The record length, key position, and key length fields are only present in files whose records can be looked up using a key. They provide the information required to find the keys.

The various times keep track of when the file was created, most recently accessed and most recently modified. These are useful for a variety of purposes. For example, a source file that has been modified after the creation of the corresponding object file needs to be recompiled. These fields provide the necessary information.

The current size tells how big the file is at present. Some old mainframe operating systems require the maximum size to be specified when the file is created, in order to let the operating system reserve the maximum amount of storage in advance. Modern operating systems are clever enough to do without this feature.

5.1.6. File Operations

Files exist to store information and allow it to be retrieved later. Different systems provide different operations to allow storage and retrieval. Below is a discussion of the most common system calls relating to files.

1. `Create`. The file is created with no data. The purpose of the call is to announce that the file is coming and to set some of the attributes.

2. `Delete`. When the file is no longer needed, it has to be deleted to free up disk space. A system call for this purpose is always provided.

3. `Open`. Before using a file, a process must open it. The purpose of the `open` call is to allow the system to fetch the attributes and list of disk addresses into main memory for rapid access on later calls.

4. `Close`. When all the accesses are finished, the attributes and disk addresses are no longer needed, so the file should be closed to free up some internal table space. Many systems encourage this by imposing a maximum number of open files on processes. A disk is written in blocks, and closing a file forces writing of the file's last block, even though that block may not be entirely full yet.

5. `Read`. Data are read from file. Usually, the bytes come from the current position. The caller must specify how much data are needed and must also provide a buffer to put them in.

6. `Write`. Data are written to the file, again, usually at the current position. If the current position is the end of the file, the file's size increases. If the current position is in the middle of the file, existing data are overwritten and lost forever.

7. `Append`. This call is a restricted form of `write`. It can only add data to the end of the file. Systems that provide a minimal set of system calls do not generally have `append`, but many systems provide multiple ways of doing the same thing, and these systems sometimes have `append`.

8. `Seek`. For random access files, a method is needed to specify from where to take the data. One common approach is a system call, `seek`, that repositions the file pointer to a specific place in the file. After this call has completed, data can be read from, or written to, that position.

9. `Get attributes`. Processes often need to read file attributes to do their work. For example, the UNIX *make* program is commonly used to manage software development projects consisting of many source files. When *make* is called, it examines the modification times of all the source and object files and arranges for the minimum number of compilations required to bring everything up to date. To do its job, it must look at the attributes, namely,

the modification times.

10. `Set attributes`. Some of the attributes are user settable and can be changed after the file has been created. This system call makes that possible. The protection mode information is an obvious example. Most of the flags also fall in this category.

11. `Rename`. It frequently happens that a user needs to change the name of an existing file. This system call makes that possible. It is not always strictly necessary, because the file can usually be copied to a new file with the new name, and the old file then deleted.

12. `Lock`. Locking a file or a part of a file prevents multiple simultaneous access by different process. For an airline reservation system, for instance, locking the database while making a

PREV NEXT

5.2. Directories

To keep track of files, file systems normally have **directories** or **folders**, which, in many systems, are themselves files. In this section we will discuss directories, their organization, their properties, and the operations that can be performed on them.

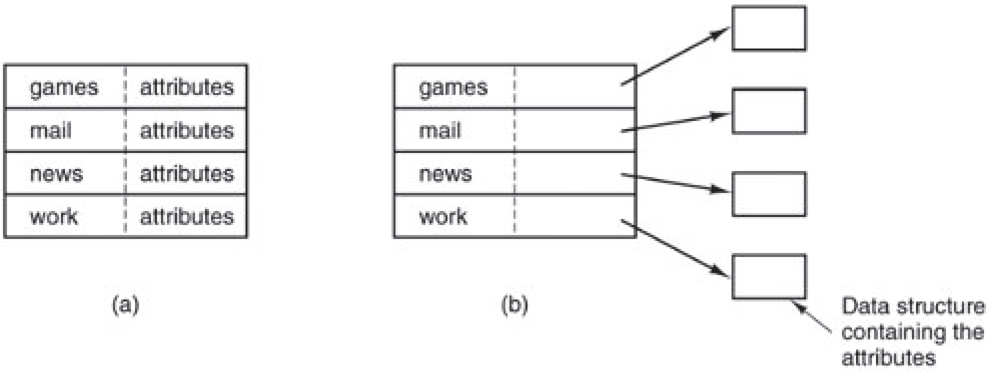
5.2.1. Simple Directories

A directory typically contains a number of entries, one per file. One possibility is shown in Fig. 5-5(a), in which each entry contains the file name, the file attributes, and the disk addresses where the data are stored. Another possibility is shown in Fig. 5-5(b). Here a directory entry holds the file name and a pointer to another data structure where the attributes and disk addresses are found. Both of these systems are commonly used.

Figure 5-5. (a) Attributes in the directory entry. (b) Attributes elsewhere.

(This item is displayed on page 492 in the print version)

[View full size image]



When a file is opened, the operating system searches its directory until it finds the name of the file to be opened. It then extracts the attributes and disk addresses, either directly from the directory entry or from the data structure pointed to, and puts them in a table in main memory. All subsequent references to the file use the information in main memory.

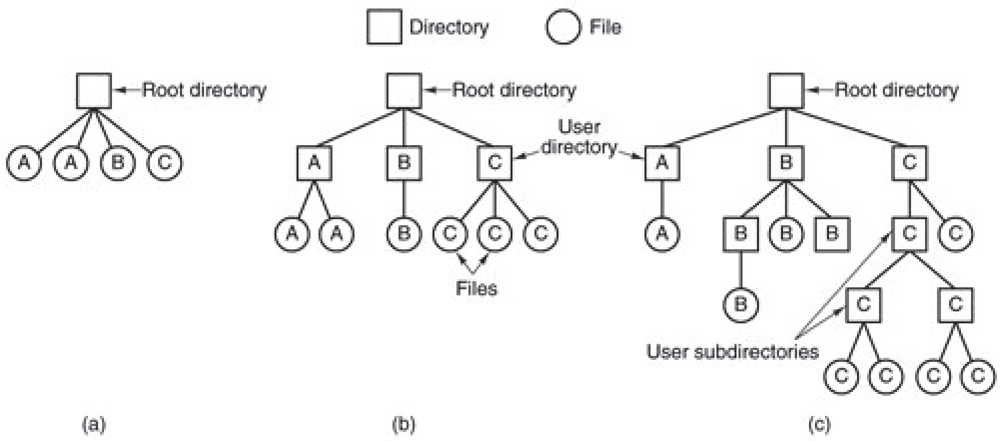
The number of directories varies from system to system. The simplest form of directory system is a single directory containing all files for all users, as illustrated in Fig. 5-6(a). On early personal computers, this single-directory system was common, in part because there was only one user.

Figure 5-6. Three file system designs. (a) Single directory shared by all

users. (b) One directory per user. (c) Arbitrary tree per user. The letters indicate the directory or file's owner.

(This item is displayed on page 493 in the print version)

[View full size image]



[Page 492]

The problem with having only one directory in a system with multiple users is that different users may accidentally use the same names for their files. For example, if user *A* creates a file called *mailbox*, and then later user *B* also creates a file called *mailbox*, *B*'s file will overwrite *A*'s file. Consequently, this scheme is not used on multiuser systems any more, but could be used on a small embedded system, for example, a handheld personal digital assistant or a cellular telephone.

To avoid conflicts caused by different users choosing the same file name for their own files, the next step up is giving each user a private directory. In that way, names chosen by one user do not interfere with names chosen by a different user and there is no problem caused by the same name occurring in two or more directories. This design leads to the system of Fig. 5-6(b). This design could be used, for example, on a multiuser computer or on a simple network of personal computers that shared a common file server over a local area network.

Implicit in this design is that when a user tries to open a file, the operating system knows which user it is in order to know which directory to search. As a consequence, some kind of login procedure is needed, in which the user specifies a login name or identification, something not required with a single-level directory system.

When this system is implemented in its most basic form, users can only access files in their own directories.

5.2.2. Hierarchical Directory Systems

The two-level hierarchy eliminates file name conflicts between users. But another problem is that users with many files may want to group them in smaller subgroups, for instance a professor might want to separate handouts for a class from drafts of chapters of a new textbook. What is needed is a general hierarchy (i.e., a tree of directories). With this approach, each user can have as many directories as are needed so that files can be grouped together in natural ways. This approach is shown in Fig. 5-6(c). Here, the directories *A*, *B*, and *C* contained in the root directory each belong to a different user, two of whom have created subdirectories for projects they are working on.

[Page 493]

The ability to create an arbitrary number of subdirectories provides a powerful structuring tool for users to organize their work. For this reason nearly all modern PC and server file systems are organized this way.

However, as we have pointed out before, history often repeats itself with new technologies. Digital cameras have to record their images somewhere, usually on a flash memory card. The very first digital cameras had a single directory and named the files *DSC0001.JPG, DSC0002.JPG,* etc. However, it did not take very long for camera manufacturers to build file systems with multiple directories, as in Fig. 5-6(b). What difference does it make that none of the camera owners understand how to use multiple directories, and probably could not conceive of any use for this feature even if they did understand it? It is only (embedded) software, after all, and thus costs the camera manufacturer next to nothing to provide. Can digital cameras with full-blown hierarchical file systems, multiple login names, and 255-character file names be far behind?

5.2.3. Path Names

When the file system is organized as a directory tree, some way is needed for specifying file names. Two different methods are commonly used. In the first method, each file is given an **absolute path name** consisting of the path from the root directory to the file. As an example, the path */usr/ast/mailbox* means that the root directory contains a subdirectory *usr/*, which in turn contains a subdirectory *ast/*, which contains the file *mailbox*. Absolute path names always start at the root directory and are unique. In UNIX the components of the path are separated by /. In Windows the separator is \ . Thus the same path name would be written as follows in these two systems:

[Page 494]

```
Windows     \usr\ast\mailbox
UNIX        /usr/ast/mailbox
```

No matter which character is used, if the first character of the path name is the separator, then the path is absolute.

The other kind of name is the **relative path name**. This is used in conjunction with the concept of the **working directory** (also called the **current directory**). A user can designate one directory as the current working directory, in which case all path names not beginning at the root directory are taken relative to the working directory. For example, if the current working directory is */usr/ast,* then the file whose absolute path is */usr/ast/mailbox* can be referenced simply as *mailbox*. In other words, the UNIX command

```
   cp /usr/ast/mailbox /usr/ast/mailbox.bak
```

and the command

```
   cp mailbox mailbox.bak
```

do exactly the same thing if the working directory is */usr/ast/.* The relative form is often more convenient, but it does the same thing as the absolute form.

Some programs need to access a specific file without regard to what the working directory is. In that case, they should always use absolute path names. For example, a spelling checker might need to read */usr/lib/dictionary* to do its work. It should use the full, absolute path name in this case because it does not know what the working directory will be when it is called. The absolute path name will always work, no matter what the working directory is.

Of course, if the spelling checker needs a large number of files from */usr/lib/,* an alternative approach is for it to issue a system call to change its working directory to */usr/lib/,* and then use just *dictionary* as the first parameter to `open`. By explicitly changing the working directory, it knows for sure where it is in the directory tree, so it can then use relative paths.

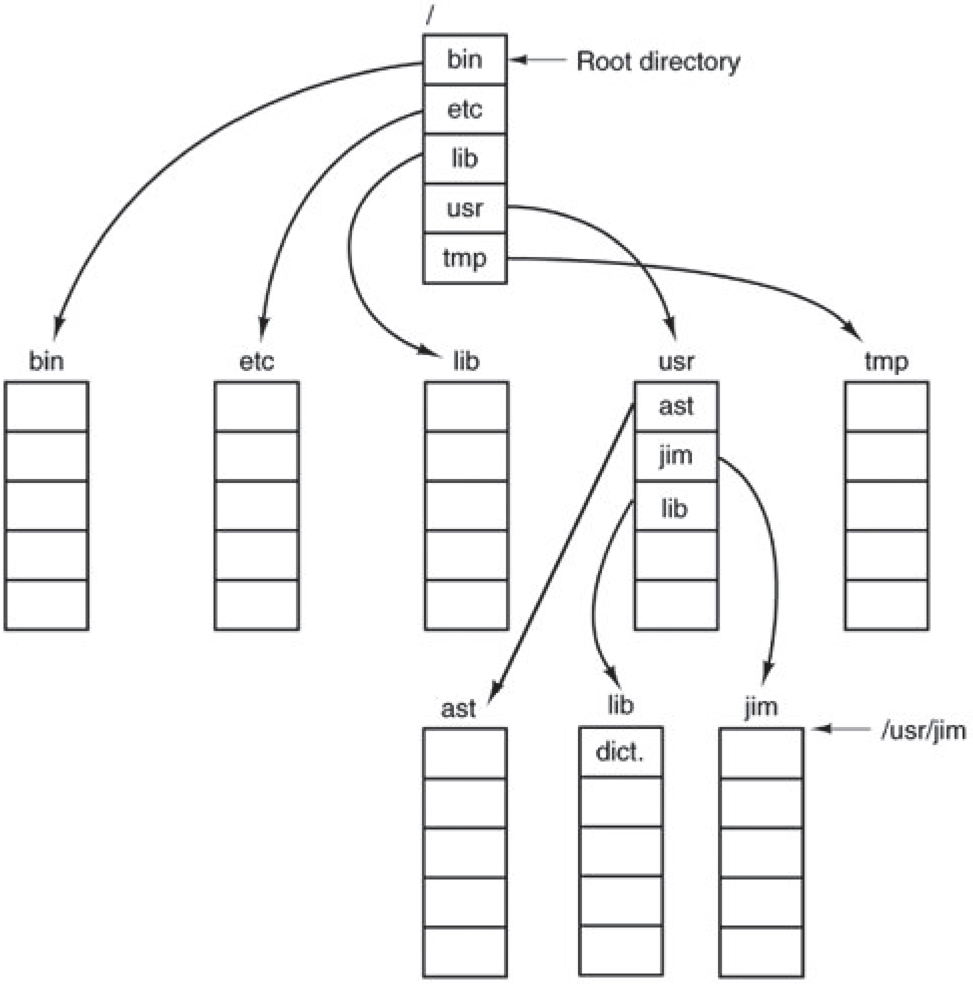
Each process has its own working directory, so when a process changes its working directory and later exits, no other processes are affected and no traces of the change are left behind in the file system. In this way it is always perfectly safe for a process to change its working directory whenever that is convenient. On the other hand, if a *library procedure* changes the working directory and does not change back to where it was when it is finished, the rest of the program may not work since its assumption about where it is may now suddenly be invalid. For this reason, library procedures rarely change the working directory, and when they must, they always change it back again before returning.

Most operating systems that support a hierarchical directory system have two special entries in every directory, "." and "..", generally pronounced "dot" and "dotdot." Dot refers to the current directory; dotdot refers to its parent. To see how these are used, consider the UNIX file tree of Fig. 5-7. A certain process has */usr/ast/* as its working directory. It can use .. to go up the tree. For example, it can copy the file */usr/lib/dictionary* to its own directory using the command

```
cp ../lib/dictionary .
```

Figure 5-7. A UNIX directory tree.

[View full size image]



The first path instructs the system to go upward (to the *usr* directory), then to go down to the directory *lib/* to find the file *dictionary*.

The second argument (dot) names the current directory. When the *cp* command gets a directory name (including dot) as its last argument, it copies all the files there. Of course, a more normal way to do the copy would be to type

```
 cp /usr/lib/dictionary .
```

Here the use of dot saves the user the trouble of typing *dictionary* a second time.

Nevertheless, typing

```
cp /usr/lib/dictionary dictionary
```

also works fine, as does

```
cp /usr/lib/dictionary /usr/ast/dictionary
```

All of these do exactly the same thing.

5.2.4. Directory Operations

The system calls for managing directories exhibit more variation from system to system than system calls for files. To give an impression of what they are and how they work, we will give a sample (taken from UNIX).

1. `Create`. A directory is created. It is empty except for dot and dotdot, which are put there automatically by the system (or in a few cases, by the *mkdir* program).
2. `Delete`. A directory is deleted. Only an empty directory can be deleted. A directory containing only dot and dotdot is considered empty as these cannot usually be deleted.
3. `Opendir`. Directories can be read. For example, to list all the files in a directory, a listing program opens the directory to read out the names of all the files it contains. Before a directory can be read, it must be opened, analogous to opening and reading a file.
4. `Closedir`. When a directory has been read, it should be closed to free up internal table space.
5. `Readdir`. This call returns the next entry in an open directory. Formerly, it was possible to read directories using the usual `read` system call, but that approach has the disadvantage of forcing the programmer to know and deal with the internal structure of directories. In contrast, `readdir` always returns one entry in a standard format, no matter which of the possible directory structures is being used.
6. `Rename`. In many respects, directories are just like files and can be renamed the same way files can be.
7. `Link`. Linking is a technique that allows a file to appear in more than one directory. This system call specifies an existing file and a path name, and creates a link from the existing file to the name specified by the path. In this way, the same file may appear in multiple directories. A link of this kind, which increments the counter in the file's i-node (to keep track of the number of directory entries containing the file), is sometimes called a **hard link**.

8. `Unlink`. A directory entry is removed. If the file being unlinked is only present in one directory (the normal case), it is removed from the file system. If it is present in multiple directories, only the path name specified is removed. The others remain. In UNIX, the system call for deleting files (discussed earlier) is, in fact, `unlink`.

The above list gives the most important calls, but there are a few others as well, for example, for managing the protection information associated with a directory.

PREV NEXT

5.3. File System Implementation

Now it is time to turn from the user's view of the file system to the implementer's view. Users are concerned with how files are named, what operations are allowed on them, what the directory tree looks like, and similar interface issues. Implementers are interested in how files and directories are stored, how disk space is managed, and how to make everything work efficiently and reliably. In the following sections we will examine a number of these areas to see what the issues and trade-offs are.

5.3.1. File System Layout

File systems usually are stored on disks. We looked at basic disk layout in Chap. 2, in the section on bootstrapping MINIX 3. To review this material briefly, most disks can be divided up into partitions, with independent file systems on each partition. Sector 0 of the disk is called the **MBR** (**Master Boot Record**) and is used to boot the computer. The end of the MBR contains the partition table. This table gives the starting and ending addresses of each partition. One of the partitions in the table may be marked as active. When the computer is booted, the BIOS reads in and executes the code in the MBR. The first thing the MBR program does is locate the active partition, read in its first block, called the **boot block**, and execute it. The program in the boot block loads the operating system contained in that partition. For uniformity, every partition starts with a boot block, even if it does not contain a bootable operating system. Besides, it might contain one in the some time in the future, so reserving a boot block is a good idea anyway.

The above description must be true, regardless of the operating system in use, for any hardware platform on which the BIOS is to be able to start more than one operating system. The terminology may differ with different operating systems. For instance the master boot record may sometimes be called the **IPL** (**Initial Program Loader**), **Volume Boot Code**, or simply **masterboot**. Some operating systems do not require a partition to be marked active to be booted, and provide a menu for the user to choose a partition to boot, perhaps with a timeout after which a default choice is automatically taken. Once the BIOS has loaded an MBR or boot sector the actions may vary. For instance, more than one block of a partition may be used to contain the program that loads the operating system. The BIOS can be counted on only to load the first block, but that block may then load additional blocks if the implementer of the operating system writes the boot block that way. An implementer can also supply a custom MBR, but it must work with a standard partition table if multiple operating systems are to be supported.

On PC-compatible systems there can be no more than four **primary partitions** because there is only room for a four-element array of partition descriptors between the master boot record and the end of the first 512-byte sector. Some operating systems allow one entry in the partition table to be an **extended partition** which points to a linked list of **logical partitions**. This makes it possible to have any number of additional partitions. The BIOS cannot start an operating system from a logical partition, so initial startup from a primary partition is required to load code that can manage logical partitions.

An alternative to extended partitions is used by MINIX 3, which allows a partition to contain a

subpartition table. An advantage of this is that the same code that manages a primary partition table can manage a subpartition table, which has the same structure. Potential uses for subpartitions are to have different ones for the root device, swapping, the system binaries, and the users' files. In this way, problems in one subpartition cannot propagate to another one, and a new version of the operating system can be easily installed by replacing the contents of some of the subpartitions but not all.

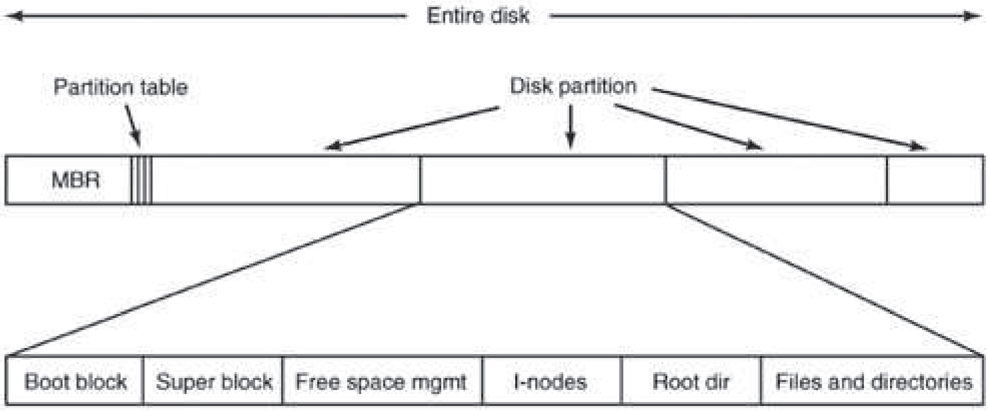
Not all disks are partitioned. Floppy disks usually start with a boot block in the first sector. The BIOS reads the first sector of a disk and looks for a magic number which identifies it as valid executable code, to prevent an attempt to execute the first sector of an unformatted or corrupted disk. A master boot record and a boot block use the same magic number, so the executable code may be either one. Also, what we say here is not limited to electromechanical disk devices. A device such as a camera or personal digital assistant that uses nonvolatile (e.g., flash) memory typically has part of the memory organized to simulate a disk.

Other than starting with a boot block, the layout of a disk partition varies considerably from file system to file system. A UNIX-like file system will contain some of the items shown in Fig. 5-8. The first one is the **superblock**. It contains all the key parameters about the file system and is read into memory when the computer is booted or the file system is first touched.

Figure 5-8. A possible file system layout.

(This item is displayed on page 499 in the print version)

[View full size image]



Next might come information about free blocks in the file system. This might be followed by the i-nodes, an array of data structures, one per file, telling all about the file and where its blocks are located. After that might come the root directory, which contains the top of the file system tree. Finally, the remainder of the disk typically contains all the other directories and files.

5.3.2. Implementing Files

Probably the most important issue in implementing file storage is keeping track of which disk blocks go with which file. Various methods are used in different operating systems. In this section, we will examine a few of them.

Contiguous Allocation

The simplest allocation scheme is to store each file as a contiguous run of disk blocks. Thus on a disk with 1-KB blocks, a 50-KB file would be allocated 50 consecutive blocks. Contiguous disk space allocation has two significant advantages. First, it is simple to implement because keeping track of where a file's blocks are is reduced to remembering two numbers: the disk address of the first block and the number of blocks in the file. Given the number of the first block, the number of any other block can be found by a simple addition.

Second, the read performance is excellent because the entire file can be read from the disk in a single operation. Only one seek is needed (to the first block). After that, no more seeks or rotational delays are needed so data come in at the full bandwidth of the disk. Thus contiguous allocation is simple to implement and has high performance.

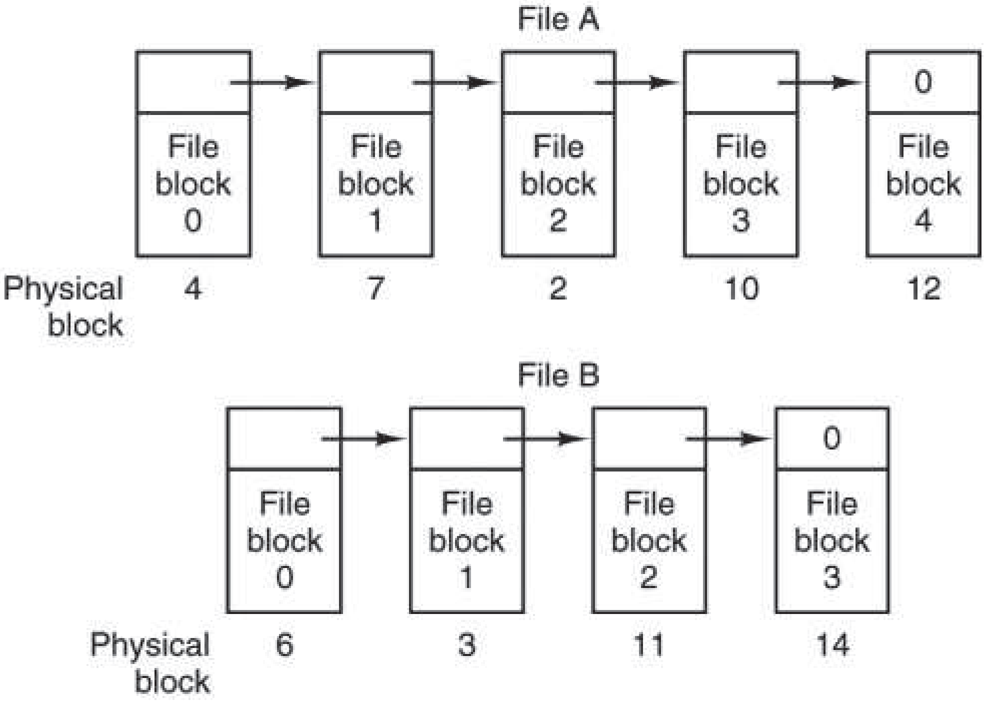
Unfortunately, contiguous allocation also has a major drawback: in time, the disk becomes fragmented, consisting of files and holes. Initially, this fragmentation is not a problem since each new file can be written at the end of disk, following the previous one. However, eventually the disk will fill up and it will become necessary to either compact the disk, which is prohibitively expensive, or to reuse the free space in the holes. Reusing the space requires maintaining a list of holes, which is doable. However, when a new file is to be created, it is necessary to know its final size in order to choose a hole of the correct size to place it in.

As we mentioned in Chap. 1, history may repeat itself in computer science as new generations of technology occur. Contiguous allocation was actually used on magnetic disk file systems years ago due to its simplicity and high performance (user friendliness did not count for much then). Then the idea was dropped due to the nuisance of having to specify final file size at file creation time. But with the advent of CD-ROMs, DVDs, and other write-once optical media, suddenly contiguous files are a good idea again. For such media, contiguous allocation is feasible and, in fact, widely used. Here all the file sizes are known in advance and will never change during subsequent use of the CD-ROM file system. It is thus important to study old systems and ideas that were conceptually clean and simple because they may be applicable to future systems in surprising ways.

Linked List Allocation

The second method for storing files is to keep each one as a linked list of disk blocks, as shown in Fig. 5-9. The first word of each block is used as a pointer to the next one. The rest of the block is for data.

Figure 5-9. Storing a file as a linked list of disk blocks.



Unlike contiguous allocation, every disk block can be used in this method. No space is lost to disk fragmentation (except for internal fragmentation in the last block of each file). Also, it is sufficient for the directory entry to merely store the disk address of the first block. The rest can be found starting there.

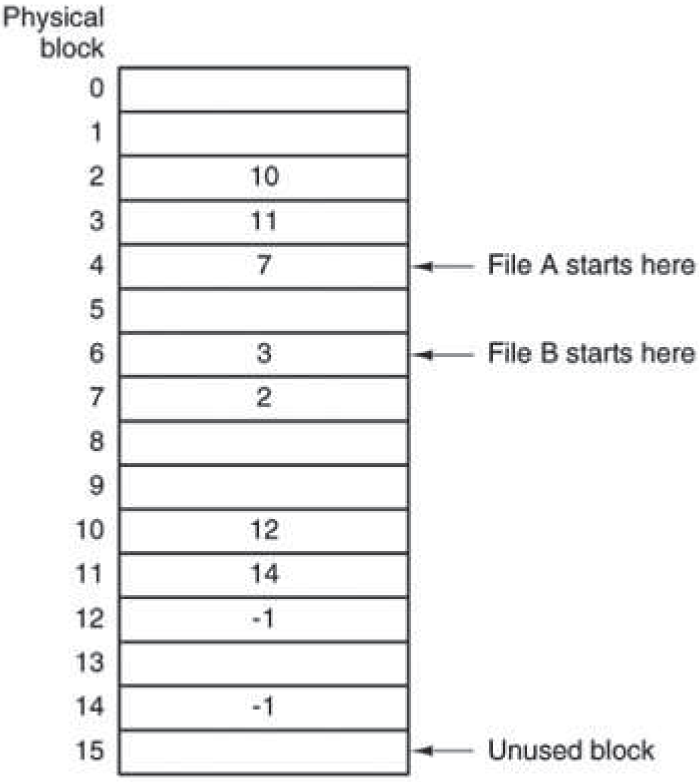
On the other hand, although reading a file sequentially is straightforward, random access is extremely slow. To get to block *n*, the operating system has to start at the beginning and read the *n* 1 blocks prior to it, one at a time. Clearly, doing so many reads will be painfully slow.

Also, the amount of data storage in a block is no longer a power of two because the pointer takes up a few bytes. While not fatal, having a peculiar size is less efficient because many programs read and write in blocks whose size is a power of two. With the first few bytes of each block occupied to a pointer to the next block, reads of the full block size require acquiring and concatenating information from two disk blocks, which generates extra overhead due to the copying.

Linked List Allocation Using a Table in Memory

Both disadvantages of the linked list allocation can be eliminated by taking the pointer word from each disk block and putting it in a table in memory. Figure 5-10 shows what the table looks like for the example of Fig. 5-9. In both figures, we have two files. File *A* uses disk blocks 4, 7, 2, 10, and 12, in that order, and file *B* uses disk blocks 6, 3, 11, and 14, in that order. Using the table of Fig. 5-10, we can start with block 4 and follow the chain all the way to the end. The same can be done starting with block 6. Both chains are terminated with a special marker (e.g., 1) that is not a valid block number. Such a table in main memory is called a **FAT** (**File Allocation Table**).

Figure 5-10. Linked list allocation using a file allocation table in main memory.



Using this organization, the entire block is available for data. Furthermore, random access is much easier. Although the chain must still be followed to find a given offset within the file, the chain is entirely in memory, so it can be followed without making any disk references. Like the previous method, it is sufficient for the directory entry to keep a single integer (the starting block number) and still be able to locate all the blocks, no matter how large the file is.

The primary disadvantage of this method is that the entire table must be in memory all the time. With a 20-GB disk and a 1-KB block size, the table needs 20 million entries, one for each of the 20 million disk blocks. Each entry has to be a minimum of 3 bytes. For speed in lookup, they should be 4 bytes. Thus the table will take up 60 MB or 80 MB of main memory all the time, depending on whether the system is optimized for space or time. Conceivably the table could be put in pageable memory, but it would still occupy a great deal of virtual memory and disk space as well as generating paging traffic. MS-DOS and Windows 98 use only FAT file systems and later versions of Windows also support it.

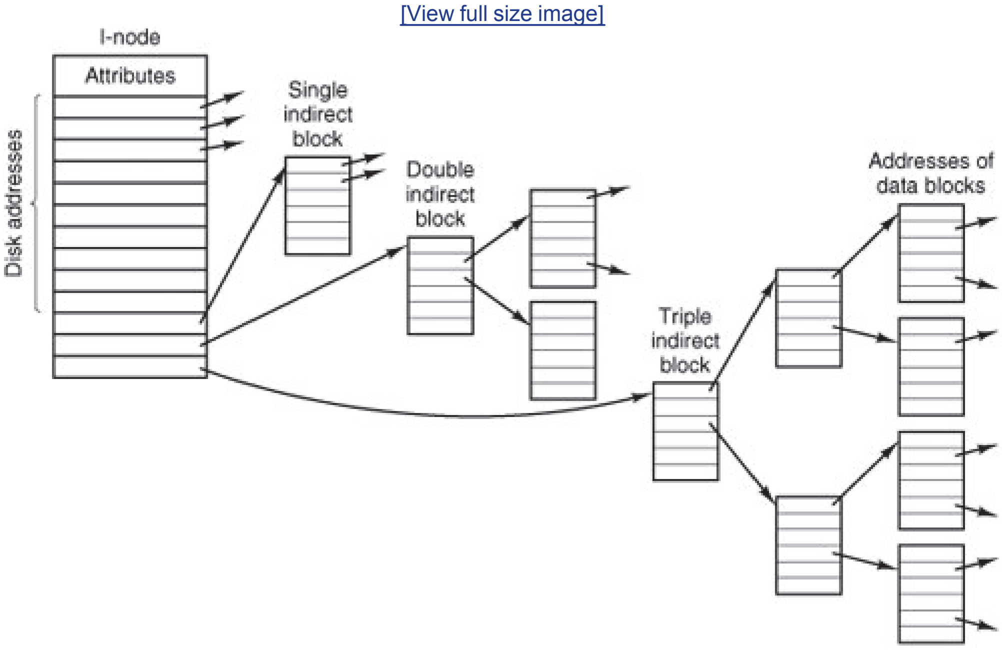
I-Nodes

Our last method for keeping track of which blocks belong to which file is to associate with each file a data structure called an **i-node** (**index-node),** which lists the attributes and disk addresses of the file's blocks. A simple example is depicted in Fig. 5-11. Given the i-node, it is then possible to find all the blocks of the file. The big advantage of this scheme over linked files using an in-memory table is that the i-node need only be in memory when the corresponding file is open. If each i-node occupies *n* bytes and a maximum of *k* files may be open at once, the total memory occupied by the array holding the i-nodes for the open files is only *kn* bytes. Only this much space need be reserved in advance.

Figure 5-11. An i-node with three levels of indirect blocks.

(This item is displayed on page 503 in the print version)

[View full size image]



This array is usually far smaller than the space occupied by the file table described in the previous section. The reason is simple. The table for holding the linked list of all disk blocks is proportional in size to the disk itself. If the disk has *n* blocks, the table needs *n* entries. As disks grow larger, this table grows linearly with them. In contrast, the i-node scheme requires an array in memory whose size is proportional to the maximum number of files that may be open at once. It does not matter if the disk is 1 GB or 10 GB or 100 GB.

One problem with i-nodes is that if each one has room for a fixed number of disk addresses, what happens when a file grows beyond this limit? One solution is to reserve the last disk address not for a data block, but instead for the address of an **indirect block** containing more disk block addresses. This idea can be extended to use **double indirect blocks** and **triple indirect blocks**, as shown in Fig. 5-11.

5.3.3. Implementing Directories

Before a file can be read, it must be opened. When a file is opened, the operating system uses the path name supplied by the user to locate the directory entry. Finding a directory entry means, of course, that the root directory must be located first. The root directory may be in a fixed location relative to the start of a partition. Alternatively, its position may be determined from other information, for instance, in a classic UNIX file system the superblock contains information about the size of the file system data structures that precede the data area. From the superblock the location of the i-nodes can be found. The first i-node will point to the root directory, which is created when a UNIX file system is made. In Windows XP, information in the boot sector (which is really much bigger than one sector) locates the **MFT** (**Master File Table**), which is used to locate other parts of the file system.

[Page 503]

Once the root directory is located a search through the directory tree finds the desired directory entry. The directory entry provides the information needed to find the disk blocks. Depending on the system, this information may be the disk address of the entire file (contiguous allocation), the number of the first block (both linked list schemes), or the number of the i-node. In all cases, the main function of the directory system is to map the ASCII name of the file onto the information needed to locate the data.

A closely related issue is where the attributes should be stored. Every file system maintains file attributes, such as each file's owner and creation time, and they must be stored somewhere. One obvious possibility is to store them directly in the directory entry. In its simplest form, a directory consists of a list of fixed-size entries, one per file, containing a (fixed-length) file name, a structure of the file attributes, and one or more disk addresses (up to some maximum) telling where the disk blocks are, as we saw in Fig. 5-5(a).

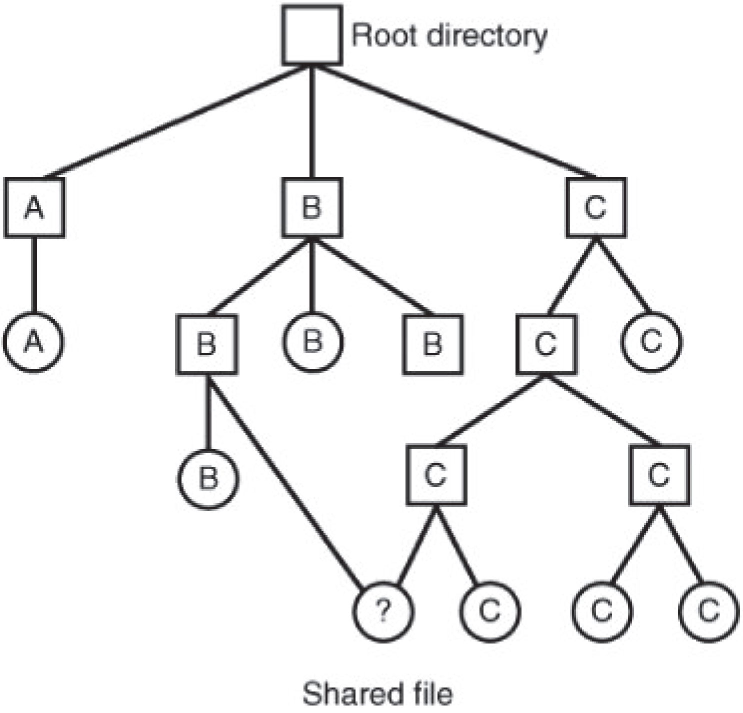
[Page 504]

For systems that use i-nodes, another possibility for storing the attributes is in the i-nodes, rather than in the directory entries, as in Fig. 5-5(b). In this case, the directory entry can be shorter: just a file name and an i-node number.

Shared Files

In Chap. 1 we briefly mentioned **links** between files, which make it easy for several users working together on a project to share files. Figure 5-12 shows the file system of Fig. 5-6(c) again, only with one of *C*'s files now present in one of *B*'s directories as well.

Figure 5-12. File system containing a shared file.



In UNIX the use of i-nodes for storing file attributes makes sharing easy; any number of directory entries can point to a single i-node. The i-node contains a field which is incremented when a new link is added, and which is decremented when a link is deleted. Only when the link count reaches zero are the actual data and the i-node itself deleted.

This kind of link is sometimes called a **hard link**. Sharing files using hard links is not always possible. A major limitation is that directories and i-nodes are data structures of a single file system (partition), so a directory in one file system cannot point to an i-node on another file system. Also, a file can have only one owner and one set of permissions. If the owner of a shared file deletes his own directory entry for that file, another user could be stuck with a file in his directory that he cannot delete if the permissions do not allow it.

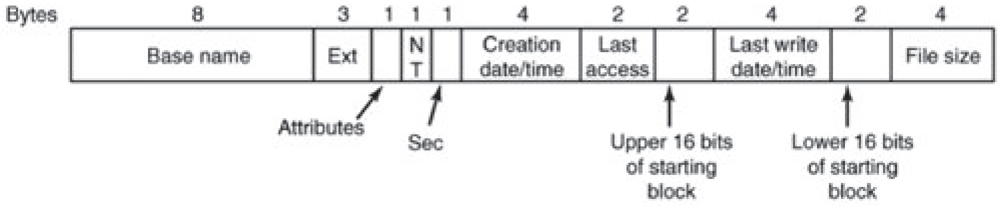
An alternative way to share files is to create a new kind of file whose data is the path to another file. This kind of link will work across mounted file systems. In fact, if a means is provided for path names to include network addresses, such a link can refer to a file on a different computer. This second kind of link is called a **symbolic link** in UNIX-like systems, a **shortcut** in Windows, and an **alias** in Apple's Mac OS. Symbolic links can be used on systems where attributes are stored within directory entries. A little thought should convince you that multiple directory entries containing file attributes would be difficult to synchronize. Any change to a file would have to affect every directory entry for that file. But the extra directory entries for symbolic links do not contain the attributes of the file to which they point. A disadvantage of symbolic links is that when a file is deleted, or even just renamed, a link becomes an orphan.

Directories in Windows 98

The file system of the original release of Windows 95 was identical to the MS-DOS file system, but a second release added support for longer file names and bigger files. We will refer to this as the Windows 98 file system, even though it is found on some Windows 95 systems. Two types of directory entry exist in Windows 98. We will call the first one, shown in Fig. 5-13, a base entry.

Figure 5-13. A Windows 98 base directory entry.

[View full size image]

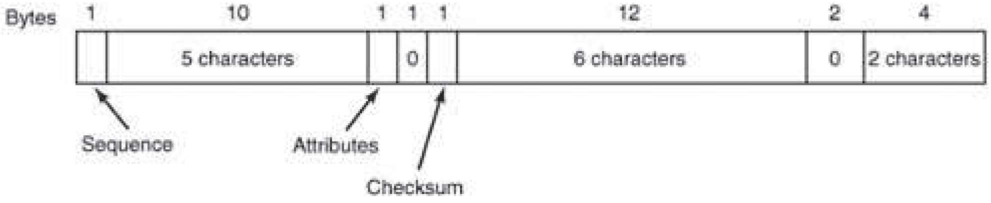


The base directory entry has all the information that was in the directory entries of older Windows versions, and more. The 10 bytes starting with the *NT* field are additions to the older Windows 95 structure, which fortunately (or more likely deliberately, with later improvement in mind) were not previously used. The most important upgrade is the field that increases the number of bits available for pointing to the starting block from 16 to 32. This increases the maximum potential size of the file system from 2^{16} blocks to 2^{32} blocks.

This structure provides only for the old-style 8 + 3 character filenames inherited from MS-DOS (and CP/M). How about long file names? The answer to the problem of providing long file names while retaining compatibility with the older systems was to use additional directory entries. Fig. 5-14 shows an alternative form of directory entry that can contain up to 13 characters of a long file name. For files with long names a shortened form of the name is generated automatically and placed in the *Base name* and *Ext* fields of an Fig. 5-13-style base directory entry. As many entries like that of Fig. 5-14 as are needed to contain the long file name are placed before the base entry, in reverse order. The *Attributes* field of each long name entry contains the value 0x0F, which is an impossible value for older (MS-DOS and Windows 95) files systems, so these entries will be ignored if the directory is read by an older system (on a floppy disk, for instance). A bit in the *Sequence* field tells the system which is the last entry.

Figure 5-14. An entry for (part of) a long file name in Windows 98.

[View full size image]

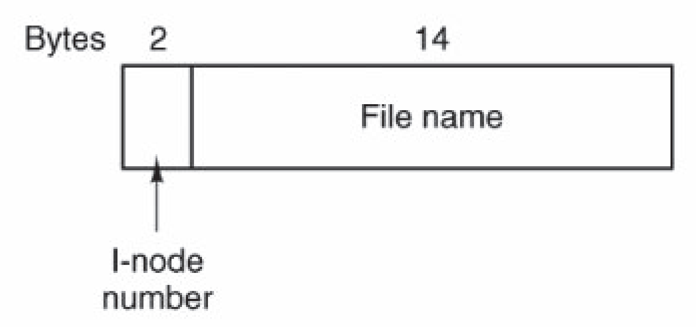


If this seems rather complex, well, it is. Providing backward compatibility so an earlier simpler system can continue to function while providing additional features for a newer system is likely to be messy. A purist might decide not to go to so much trouble. However, a purist would probably not become rich selling new versions of operating systems.

Directories in UNIX

The traditional UNIX directory structure is extremely simple, as shown in Fig. 5-15. Each entry contains just a file name and its i-node number. All the information about the type, size, times, ownership, and disk blocks is contained in the i-node. Some UNIX systems have a different layout, but in all cases, a directory entry ultimately contains only an ASCII string and an i-node number.

Figure 5-15. A Version 7 UNIX directory entry.

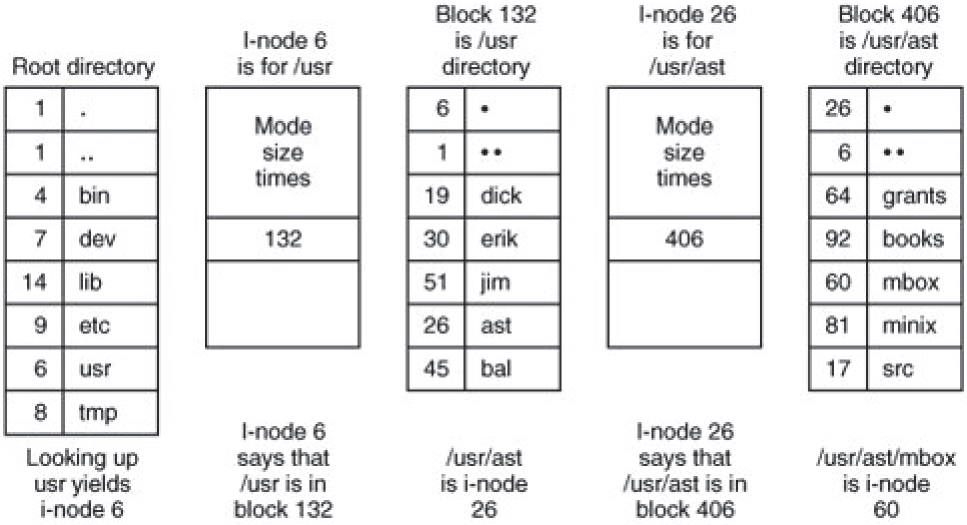


When a file is opened, the file system must take the file name supplied and locate its disk blocks. Let us consider how the path name */usr/ast/mbox* is looked up. We will use UNIX as an example, but the algorithm is basically the same for all hierarchical directory systems. First the system locates the root directory. The i-nodes form a simple array which is located using information in the superblock. The first entry in this array is the i-node of the root directory.

The file system looks up the first component of the path, *usr*, in the root directory to find the i-node number of the file */usr/*. Locating an i-node from its number is straightforward, since each one has a fixed location relative to the first one. From this i-node, the system locates the directory for */usr/* and looks up the next component, *ast*, in it. When it has found the entry for *ast*, it has the i-node for the directory */usr/ast/.* From this i-node it can find the directory itself and look up *mbox*. The i-node for this file is then read into memory and kept there until the file is closed. The lookup process is illustrated in Fig. 5-16.

Figure 5-16. The steps in looking up */usr/ast/mbox.*

[View full size image]


Root directory
1 .
1 ..
4 bin
7 dev
14 lib
9 etc
6 usr
8 tmp
Looking up usr yields i-node 6
I-node 6 is for /usr
Mode size times
132
I-node 6 says that /usr is in block 132
Block 132 is /usr directory
6 •
1 ••
19 dick
30 erik
51 jim
26 ast
45 bal
/usr/ast is i-node 26
I-node 26 is for /usr/ast
Mode size times
406
I-node 26 says that /usr/ast is in block 406
Block 406 is /usr/ast directory
26 •
6 ••
64 grants
92 books
60 mbox
81 minix
17 src
/usr/ast/mbox is i-node 60

Relative path names are looked up the same way as absolute ones, only starting from the working directory instead of starting from the root directory. Every directory has entries for *.* and *..* which are put there when the directory is created. The entry *.* has the i-node number for the current directory, and the entry for *..* has the i-node number for the parent directory. Thus, a procedure looking up *../dick/prog.c* simply looks up *..* in the working directory, finds the i-node number for the parent directory, and searches that directory for *dick*. No special mechanism is needed to handle these names. As far as the directory system is concerned, they are just ordinary ASCII strings, just the same as any other names.

Directories in NTFS

Microsoft's **NTFS** (**New Technology File System**) is the default file system. We do not have space for a detailed description of NTFS, but will just briefly look at some of the problems NTFS deals with and the solutions used.

One problem is long file and path names. NTFS allows long file names (up to 255 characters) and path names (up to 32,767 characters). But since older versions of Windows cannot read NTFS file systems, a complicated backward-compatible directory structure is not needed, and filename fields are variable length. Provision is made to have a second 8 + 3 character name so an older system can access NTFS files over a network.

NTFS provides for multiple character sets by using Unicode for filenames. Unicode uses 16 bits for each character, enough to represent multiple languages with very large symbol sets (e.g., Japanese). But using multiple languages raises problems in addition to representation of different character sets. Even among Latin-derived languages there are subtleties. For instance, in Spanish some combinations of two characters count as single characters when sorting. Words beginning with "ch" or "ll" should appear in sorted lists after words that begin with "cz" or "lz", respectively. The problem of case mapping is more complex. If the default is to make filenames case sensitive, there may still be a need to do case-insensitive searches. For Latin-based languages it is obvious how to do that, at least to native users of these languages. In general, if only one language is in use, users will probably know the rules. However, Unicode allows a mixture of languages: Greek, Russian, and Japanese filenames could all appear in a single directory at an international organization. The NTFS solution is an attribute for each file that defines the case conventions for the language of the filename.

More attributes is the NTFS solution to many problems. In UNIX, a file is a sequence of bytes. In NTFS a file is a collection of attributes, and each attribute is a stream of bytes. The basic NTFS data structure is the **MFT** (**Master File Table**) that provides for 16 attributes, each of which can have a length of up to 1 KB within the MFT. If that is not enough, an attribute within the MFT can be a header that points to an additional file with an extension of the attribute values. This is known as a **nonresident attribute**. The MFT itself is a file, and it has an entry for every file and directory in the file system. Since it can grow very large, when an NTFS file system is created about 12.5% of the space on the partition is reserved for growth of the MFT. Thus it can grow without becoming fragmented, at least until the initial reserved space is used, after which another large chunk of space will be reserved. So if the MFT becomes fragmented it will consists of a small number of very large fragments.

What about data in NTFS? Data is just another attribute. In fact an NTFS file may have more than one data stream. This feature was originally provided to allow Windows servers to serve files to Apple MacIntosh clients. In the original MacIntosh operating system (through Mac OS 9) all files had two data streams, called the resource fork and the data fork. Multiple data streams have other uses, for instance a large graphic image may have a smaller thumbnail image associated with it. A stream can contain up to 2^{64} bytes. At the other extreme, NTFS can handle small files by putting a few hundred bytes in the attribute header. This is called an **immediate file** (Mullender and Tanenbaum, 1984).

We have only touched upon a few ways that NTFS deals with issues not addressed by older and simpler file systems. NTFS also provides features such as a sophisticated protection system, encryption, and data compression. Describing all these features and their implementation would require much more space than we can spare here. For a more throrough look at NTFS see Tanenbaum (2001) or look on the World Wide Web for more information.

5.3.4. Disk Space Management

Files are normally stored on disk, so management of disk space is a major concern to file system designers. Two general strategies are possible for storing an *n* byte file: *n* consecutive bytes of disk space are allocated, or the file is split up into a number of (not necessarily) contiguous blocks. The same trade-off is present in memory management systems between pure segmentation and paging.

As we have seen, storing a file as a contiguous sequence of bytes has the obvious problem that if a file grows, it will probably have to be moved on the disk. The same problem holds for segments in memory, except that moving a segment in memory is a relatively fast operation compared to moving a file from one disk position to another. For this reason, nearly all file systems chop files up into fixed-size blocks that need not be adjacent.

Block Size

Once it has been decided to store files in fixed-size blocks, the question arises of how big the blocks should be. Given the way disks are organized, the sector, the track and the cylinder are obvious candidates for the unit of allocation (although these are all device dependent, which is a minus). In a paging system, the page size is also a major contender. However, having a large allocation unit, such as a cylinder, means that every file, even a 1-byte file, ties up an entire cylinder.

On the other hand, using a small allocation unit means that each file will consist of many blocks. Reading each block normally requires a seek and a rotational delay, so reading a file consisting of many small blocks will be slow.

As an example, consider a disk with 131,072 bytes/track, a rotation time of 8.33 msec, and an average seek time of 10 msec. The time in milliseconds to read a block of *k* bytes is then the sum of the seek, rotational delay, and transfer times:

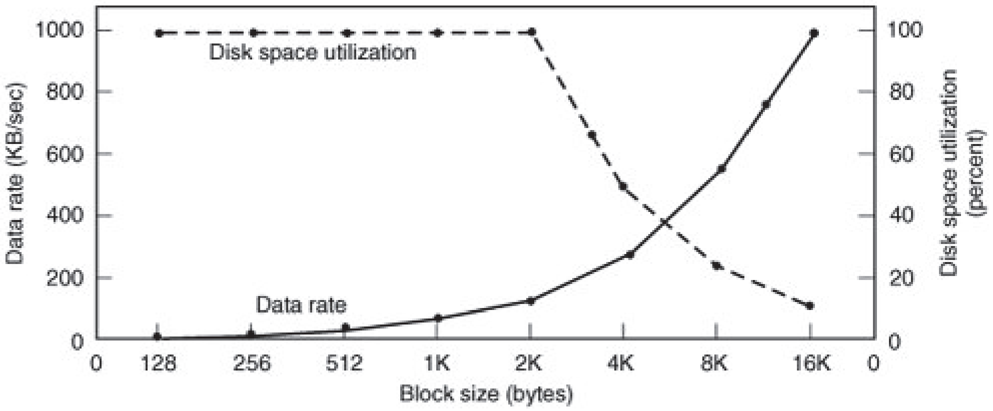
$$10 + 4.165 + (k / 131072) \times 8.33$$

The solid curve of Fig. 5-17 shows the data rate for such a disk as a function of block size.

Figure 5-17. The solid curve (left-hand scale) gives the data rate of a disk. The dashed curve (right-hand scale) gives the disk space efficiency. All files are 2 KB.

(This item is displayed on page 510 in the print version)

[View full size image]



To compute the space efficiency, we need to make an assumption about the mean file size. An early study showed that the mean file size in UNIX environments is about 1 KB (Mullender and Tanenbaum, 1984). A measurement made in 2005 at the department of one of the authors (AST), which has 1000 users and over 1 million UNIX disk files, gives a median size of 2475 bytes, meaning that half the files are smaller than 2475 bytes and half are larger. As an aside, the median is a better metric than the mean because a very small number of files can influence the mean enormously, but not the median. A few 100-MB hardware manuals or a promotional videos or to can greatly skew the mean but have little effect on the median.

In an experiment to see if Windows NT file usage was appreciably different from UNIX file usage, Vogels (1999) made measurements on files at Cornell University. He observed that NT file usage is more complicated than on UNIX. He wrote:

> *When we type a few characters in the* notepad *text editor, saving this to a file will trigger 26 system calls, including 3 failed open attempts, 1 file overwrite and 4 additional open and close sequences.*

Nevertheless, he observed a median size (weighted by usage) of files just read at 1 KB, files just written as 2.3 KB and files read and written as 4.2 KB. Given the fact that Cornell has considerable large-scale scientific computing and the difference in measurement technique (static versus dynamic), the results are reasonably consistent with a median file size of around 2 KB.

For simplicity, let us assume all files are 2 KB, which leads to the dashed curve in Fig. 5-17 for the disk space efficiency.

The two curves can be understood as follows. The access time for a block is completely dominated by the seek time and rotational delay, so given that it is going to cost 14 msec to access a block, the more data that are fetched, the better. Hence the data rate goes up with block size (until the transfers take so long that the transfer time begins to dominate). With small blocks that are powers of two and 2-KB files, no space is wasted in a block. However, with 2-KB files and 4 KB or larger blocks, some disk space is wasted. In reality, few files are a multiple of the disk block size, so some space is always wasted in the last block of a file.

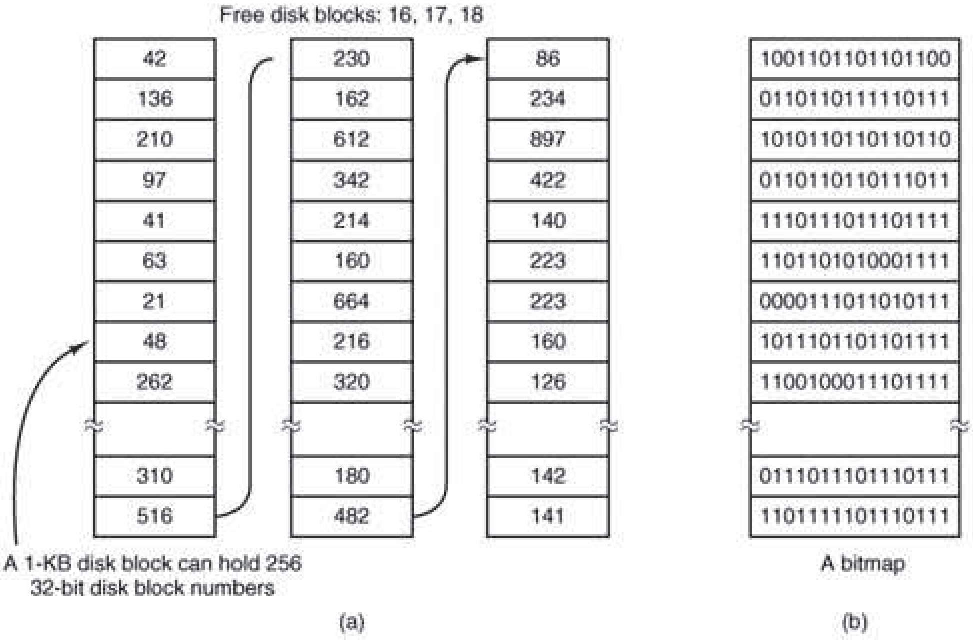
What the curves show, however, is that performance and space utilization are inherently in conflict. Small blocks are bad for performance but good for disk space utilization. A compromise size is needed. For this data, 4 KB might be a good choice, but some operating systems made their choices a long time ago, when the disk parameters and file sizes were different. For UNIX, 1 KB is commonly used. For MS-DOS the block size can be any power of two from 512 bytes to 32 KB, but is determined by the disk size and for reasons unrelated to these arguments (the maximum number of blocks on a disk partition is 2^{16}, which forces large blocks on large disks).

Keeping Track of Free Blocks

Once a block size has been chosen, the next issue is how to keep track of free blocks. Two methods are widely used, as shown in Fig. 5-18. The first one consists of using a linked list of disk blocks, with each block holding as many free disk block numbers as will fit. With a 1-KB block and a 32-bit disk block number, each block on the free list holds the numbers of 255 free blocks. (One slot is needed for the pointer to the next block). A 256-GB disk needs a free list of maximum 1,052,689 blocks to hold all 2^{28} disk block numbers. Often free blocks are used to hold the free list.

Figure 5-18. (a) Storing the free list on a linked list. (b) A bitmap.

[View full size image]



The other free space management technique is the bitmap. A disk with *n* blocks requires a bitmap with *n* bits. Free blocks are represented by 1s in the map, allocated blocks by 0s (or vice versa). A 256-GB disk has 2^{28} 1-KB blocks and thus requires 2^{28} bits for the map, which requires 32,768 blocks. It is not surprising that the bitmap requires less space, since it uses 1 bit per block, versus 32 bits in the linked list model. Only if the disk is nearly full (i.e., has few free blocks) will the linked list scheme require fewer blocks than the bitmap. On the other hand, if there are many blocks free, some of them can be borrowed to hold the free list without any loss of disk capacity.

When the free list method is used, only one block of pointers need be kept in main memory. When a file is created, the needed blocks are taken from the block of pointers. When it runs out, a new block of pointers is read in from the disk. Similarly, when a file is deleted, its blocks are freed and added to the block of pointers in main memory. When this block fills up, it is written to disk.

5.3.5. File System Reliability

Destruction of a file system is often a far greater disaster than destruction of a computer. If a computer is destroyed by fire, lightning surges, or a cup of coffee poured onto the keyboard, it is annoying and will cost money, but generally a replacement can be purchased with a minimum of fuss. Inexpensive personal computers can even be replaced within an hour by just going to the dealer (except at universities, where issuing a purchase order takes three committees, five signatures, and 90 days).

If a computer's file system is irrevocably lost, whether due to hardware, software, or rats gnawing on the backup tapes, restoring all the information will be difficult and time consuming at best, and in many cases will be impossible. For the people whose programs, documents, customer files, tax records, databases, marketing plans, or other data are gone forever, the consequences can be catastrophic. While the file system cannot offer any protection against physical destruction of the equipment and media, it can help protect the information. In this section we will look at some of the issues involved in safeguarding the file system.

Floppy disks are generally perfect when they leave the factory, but they can develop bad blocks during use. It is arguable that this is more likely now than it was in the days when floppy disks were more widely used. Networks and large capacity removable devices such as writeable CDs have led to floppy disks being used infrequently. Cooling fans draw air and airborne dust in through floppy disk drives, and a drive that has not been used for a long time may be so dirty that it ruins the next disk that is inserted. A floppy drive that is used frequently is less likely to damage a disk.

Hard disks frequently have bad blocks right from the start: it is just too expensive to manufacture them completely free of all defects. As we saw in Chap. 3, bad blocks on hard disks are generally handled by the controller by replacing bad sectors with spares provided for that purpose. On these disks, tracks are at least one sector bigger than needed, so that at least one bad spot can be skipped by leaving it in a gap between two consecutive sectors. A few spare sectors are provided on each cylinder so the controller can do automatic sector remapping if it notices that a sector needs more than a certain number of retries to be read or written. Thus the user is usually unaware of bad blocks or their management. Nevertheless, when a modern IDE or SCSI disk fails, it will usually fail horribly, because it has run out of spare sectors. SCSI disks provide a "recovered error" when they remap a block. If the driver notes this and displays a message on the monitor the user will know it is time to buy a new disk when these messages begin to appear frequently.

A simple software solution to the bad block problem exists, suitable for use on older disks. This approach requires the user or file system to carefully construct a file containing all the bad blocks. This technique removes them from the free list, so they will never occur in data files. As long as the bad block file is never read or written, no problems will arise. Care has to be taken during disk backups to avoid reading this file and trying to back it up.

Backups

Most people do not think making backups of their files is worth the time and effortuntil one fine day their disk abruptly dies, at which time most of them undergo a deathbed conversion. Companies, however, (usually) well understand the value of their data and generally do a backup at least once a day, usually to tape. Modern tapes hold tens or sometimes even hundreds of gigabytes and cost pennies per gigabyte. Nevertheless, making backups is not quite as trivial as it sounds, so we will examine some of the related issues below.

Backups to tape are generally made to handle one of two potential problems:

1. Recover from disaster.
2. Recover from stupidity.

The first one covers getting the computer running again after a disk crash, fire, flood, or other natural catastrophe. In practice, these things do not happen very often, which is why many people do not bother with backups. These people also tend not to have fire insurance on their houses for the same reason.

The second reason is that users often accidentally remove files that they later need again. This problem occurs so often that when a file is "removed" in Windows, it is not deleted at all, but just moved to a special directory, the **recycle bin**, so it can be fished out and restored easily later. Backups take this principle further and allow files that were removed days, even weeks ago, to be restored from old backup tapes.

Making a backup takes a long time and occupies a large amount of space, so doing it efficiently and conveniently is important. These considerations raise the following issues. First, should the entire file system be backed up or only part of it? At many installations, the executable (binary) programs are kept in a limited part of the file system tree. It is not necessary to back up these files if they can all be reinstalled from the manufacturers' CD-ROMs. Also, most systems have a directory for temporary files. There is usually no reason to back it up either. In UNIX, all the special files (I/O devices) are kept in a directory */dev/.* Not only is backing up this directory not necessary, it is downright dangerous because the backup program would hang forever if it tried to read each of these to completion. In short, it is usually desirable to back up only specific directories and everything in them rather than the entire file system.

Second, it is wasteful to back up files that have not changed since the last backup, which leads to the idea of **incremental dumps**. The simplest form of incremental dumping is to make a complete dump (backup) periodically, say weekly or monthly, and to make a daily dump of only those files that have been modified since the last full dump. Even better is to dump only those files that have changed since they were last dumped. While this scheme minimizes dumping time, it makes recovery more complicated because first the most recent full dump has to be restored, followed by all the incremental dumps in reverse order, oldest one first. To ease recovery, more sophisticated incremental dumping schemes are often used.

Third, since immense amounts of data are typically dumped, it may be desirable to compress the data before writing them to tape. However, with many compression algorithms, a single bad spot on the backup tape can foil the decompression algorithm and make an entire file or even an entire tape unreadable. Thus the decision to compress the backup stream must be carefully considered.

Fourth, it is difficult to perform a backup on an active file system. If files and directories are being added, deleted, and modified during the dumping process, the resulting dump may be inconsistent. However, since making a dump may take hours, it may be necessary to take the system offline for much of the night to make the backup, something that is not always acceptable. For this reason, algorithms have been devised for making rapid snapshots of the file system state by copying critical data structures, and then requiring future changes to files and directories to copy the blocks instead of updating them in place (Hutchinson et al., 1999). In this way, the file system is effectively frozen at the moment of the snapshot, so it can be backed up at leisure afterward.

Fifth and last, making backups introduces many nontechnical problems into an organization. The best online security system in the world may be useless if the system administrator keeps all the backup tapes in his office and leaves it open and unguarded whenever he walks down the hall to get output from the printer. All a spy has to do is pop in for a second, put one tiny tape in his pocket, and saunter off jauntily. Goodbye security. Also, making a daily backup has little use if the fire that burns down the computers also burns up all the backup tapes. For this reason, backup tapes should be kept off-site, but that introduces more security risks. For a thorough discussion of these and other practical administration issues, see Nemeth et al. (2001). Below we will discuss only the technical issues involved in making file system backups.

Two strategies can be used for dumping a disk to tape: a physical dump or a logical dump. A **physical dump** starts at block 0 of the disk, writes all the disk blocks onto the output tape in order, and stops when it has copied the last one. Such a program is so simple that it can probably be made 100% bug free, something that can probably not be said about any other useful program.

Nevertheless, it is worth making several comments about physical dumping. For one thing, there is no value in backing up unused disk blocks. If the dumping program can get access to the free block data structure, it can avoid dumping unused blocks. However, skipping unused blocks requires writing the number of each block in front of the block (or the equivalent), since it is no longer true that block *k* on the tape was block *k* on the disk.

A second concern is dumping bad blocks. If all bad blocks are remapped by the disk controller and hidden from the operating system as we described in Sec. 5.4.4, physical dumping works fine. On the other hand, if they are visible to the operating system and maintained in one or more "bad block files" or bitmaps, it is absolutely essential that the physical dumping program get access to this information and avoid dumping them to prevent endless disk read errors during the dumping process.

The main advantages of physical dumping are simplicity and great speed (basically, it can run at the speed of the disk). The main disadvantages are the inability to skip selected directories, make incremental dumps, and restore individual files upon request. For these reasons, most installations make logical dumps.

A **logical dump** starts at one or more specified directories and recursively dumps all files and directories found there that have changed since some given base date (e.g., the last backup for an incremental dump or system installation for a full dump). Thus in a logical dump, the dump tape gets a series of carefully identified directories and files, which makes it easy to restore a specific file or directory upon request.

In order to be able to properly restore even a single file correctly, all information needed to recreate the path to that file must be saved to the backup medium. Thus the first step in doing a logical dump is doing an analysis of the directory tree. Obviously, we need to save any file or directory that has been modified. But for proper restoration, all directories, even unmodified ones, that lie on the path to a modified file or directory must be saved. This means saving not just the data (file names and pointers to i-nodes), all the attributes of the directories must be saved, so they can be restored with the original permissions. The directories and their attributes are written to the tape first, and then modified files (with their attributes) are saved. This makes it possible to restore the dumped files and directories to a fresh file system on a different computer. In this way, the dump and restore programs can be used to transport entire file systems between computers.

A second reason for dumping unmodified directories above modified files is to make it possible to incrementally restore a single file (possibly to handle recovery from accidental deletion). Suppose that a full file system dump is done Sunday evening and an incremental dump is done on Monday evening. On Tuesday the directory */usr/jhs/proj/nr3/* is removed, along with all the directories and files under it. On Wednesday morning bright and early, a user wants to restore the file */usr/jhs/proj/nr3/plans/summary* However, is not possible to just restore the file *summary* because there is no place to put it. The directories *nr3/* and *plans/* must be restored first. To get their owners, modes, times, etc., correct, these directories must be present on the dump tape even though they themselves were not modified since the previous full dump.

Restoring a file system from the dump tapes is straightforward. To start with, an empty file system is created on the disk. Then the most recent full dump is restored. Since the directories appear first on the tape, they are all restored first, giving a skeleton of the file system. Then the files themselves are restored. This process is then repeated with the first incremental dump made after the full dump, then the next one, and so on.

Although logical dumping is straightforward, there are a few tricky issues. For one, since the free block list is not a file, it is not dumped and hence it must be reconstructed from scratch after all the dumps have been restored. Doing so is always possible since the set of free blocks is just the complement of the set of blocks contained in all the files combined.

Another issue is links. If a file is linked to two or more directories, it is important that the file is restored only one time and that all the directories that are supposed to point to it do so.

Still another issue is the fact that UNIX files may contain holes. It is legal to open a file, write a few bytes, then seek to a distant file offset and write a few more bytes. The blocks in between are not part of the file and should not be dumped and not be restored. Core dump files often have a large hole between the data segment and the stack. If not handled properly, each restored core file will fill this area with zeros and thus be the same size as the virtual address space (e.g., 2^{32} bytes, or worse yet, 2^{64} bytes).

Finally, special files, named pipes, and the like should never be dumped, no matter in which directory they may occur (they need not be confined to */dev/).* For more information about file system backups, see Chervenak et al. (1998) and Zwicky (1991).

File System Consistency

Another area where reliability is an issue is file system consistency. Many file systems read blocks, modify them, and write them out later. If the system crashes before all the modified blocks have been written out, the file system can be left in an inconsistent state. This problem is especially critical if some of the blocks that have not been written out are i-node blocks, directory blocks, or blocks containing the free list.

To deal with the problem of inconsistent file systems, most computers have a utility program that checks file system consistency. For example, UNIX has *fsck* and Windows has *chkdsk* (or *scandisk* in earlier versions). This utility can be run whenever the system is booted, especially after a crash. The description below tells how *fsck* works. *Chkdsk* is somewhat different because it works on a different file system, but the general principle of using the file system's inherent redundancy to repair it is still valid. All file system checkers verify each file system (disk partition) independently of the other ones.

Two kinds of consistency checks can be made: blocks and files. To check for block consistency, the program builds two tables, each one containing a counter for each block, initially set to 0. The counters in the first table keep track of how many times each block is present in a file; the counters in the second table record how often each block is present in the free list (or the bitmap of free blocks).

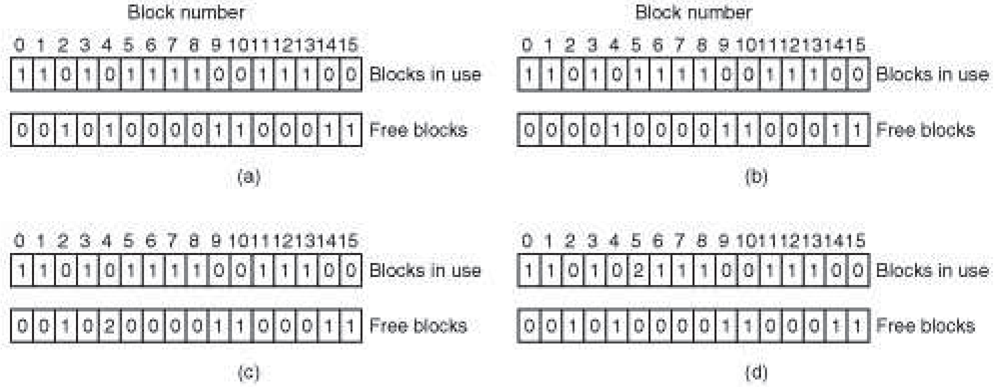
The program then reads all the i-nodes. Starting from an i-node, it is possible to build a list of all the block numbers used in the corresponding file. As each block number is read, its counter in the first table is incremented. The program then examines the free list or bitmap, to find all the blocks that are not in use. Each occurrence of a block in the free list results in its counter in the second table being incremented.

If the file system is consistent, each block will have a 1 either in the first table or in the second table, as illustrated in Fig. 5-19(a). However, as a result of a crash, the tables might look like Fig. 5-19(b), in which block 2 does not occur in either table. It will be reported as being a **missing block**. While missing blocks do no real harm, they do waste space and thus reduce the capacity of the disk. The solution to missing blocks is straightforward: the file system checker just adds them to the free list.

Figure 5-19. File system states. (a) Consistent. (b) Missing block. (c) Duplicate block in free list. (d) Duplicate data block.

(This item is displayed on page 518 in the print version)

[View full size image]



Block number

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Blocks in use	1	1	0	1	0	1	1	1	1	0	0	1	1	1	0	0
Free blocks	0	0	1	0	1	0	0	0	0	1	1	0	0	0	1	1

(a)

Block number

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Blocks in use	1	1	0	1	0	1	1	1	1	0	0	1	1	1	0	0
Free blocks	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1

(b)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Blocks in use	1	1	0	1	0	1	1	1	1	0	0	1	1	1	0	0
Free blocks	0	0	1	0	2	0	0	0	0	1	1	0	0	0	1	1

(c)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Blocks in use	1	1	0	1	0	2	1	1	1	0	0	1	1	1	0	0
Free blocks	0	0	1	0	1	0	0	0	0	1	1	0	0	0	1	1

(d)

Another situation that might occur is that of Fig. 5-19(c). Here we see a block, number 4, that occurs twice in the free list. (Duplicates can occur only if the free list is really a list; with a bitmap it is impossible.) The solution here is also simple: rebuild the free list.

The worst thing that can happen is that the same data block is present in two or more files, as shown in Fig. 5-19(d) with block 5. If either of these files is removed, block 5 will be put on the free list, leading to a situation in which the same block is both in use and free at the same time. If both files are removed, the block will be put onto the free list twice.

The appropriate action for the file system checker to take is to allocate a free block, copy the contents of block 5 into it, and insert the copy into one of the files. In this way, the information content of the files is unchanged (although almost assuredly one is garbled), but the file system structure is at least made consistent. The error should be reported, to allow the user to inspect the damage.

[Page 518]

In addition to checking to see that each block is properly accounted for, the file system checker also checks the directory system. It, too, uses a table of counters, but these are per file, rather than per block. It starts at the root directory and recursively descends the tree, inspecting each directory in the file system. For every file in every directory, it increments a counter for that file's usage count. Remember that due to hard links, a file may appear in two or more directories. Symbolic links do not count and do not cause the counter for the target file to be incremented.

When it is all done, it has a list, indexed by i-node number, telling how many directories contain each file. It then compares these numbers with the link counts stored in the i-nodes themselves. These counts start at 1 when a file is created and are incremented each time a (hard) link is made to the file. In a consistent file system, both counts will agree. However, two kinds of errors can occur: the link count in the i-node can be too high or it can be too low.

If the link count is higher than the number of directory entries, then even if all the files are removed from the directories, the count will still be nonzero and the i-node will not be removed. This error is not serious, but it wastes space on the disk with files that are not in any directory. It should be fixed by setting the link count in the i-node to the correct value.

The other error is potentially catastrophic. If two directory entries are linked to a file, but the i-node says that there is only one, when either directory entry is removed, the i-node count will go to zero. When an i-node count goes to zero, the file system marks it as unused and releases all of its blocks. This action will result in one of the directories now pointing to an unused i-node, whose blocks may soon be assigned to other files. Again, the solution is just to force the link count in the i-node to the actual number of directory entries.

These two operations, checking blocks and checking directories, are often integrated for efficiency reasons (i.e., only one pass over the i-nodes is required). Other checks are also possible. For example, directories have a definite format, with i-node numbers and ASCII names. If an i-node number is larger than the number of i-nodes on the disk, the directory has been damaged.

Furthermore, each i-node has a mode, some of which are legal but strange, such as 0007, which allows the owner and his group no access at all, but allows outsiders to read, write, and execute the file. It might be useful to at least report files that give outsiders more rights than the owner. Directories with more than, say, 1000 entries are also suspicious. Files located in user directories, but which are owned by the superuser and have the SETUID bit on, are potential security problems because such files acquire the powers of the superuser when executed by any user. With a little effort, one can put together a fairly long list of technically legal but still peculiar situations that might be worth reporting.

The previous paragraphs have discussed the problem of protecting the user against crashes. Some file systems also worry about protecting the user against himself. If the user intends to type

```
rm *.o
```

to remove all the files ending with *.o* (compiler generated object files), but accidentally types

```
rm * .o
```

(note the space after the asterisk), *rm* will remove all the files in the current directory and then complain that it cannot find *.o.* In some systems, when a file is removed, all that happens is that a bit is set in the directory or i-node marking the file as removed. No disk blocks are returned to the free list until they are actually needed. Thus, if the user discovers the error immediately, it is possible to run a special utility program that "unremoves" (i.e., restores) the removed files. In Windows, files that are removed are placed in the recycle bin, from which they can later be retrieved if need be. Of course, no storage is reclaimed until they are actually deleted from this directory.

Mechanisms like this are insecure. A secure system would actually overwrite the data blocks with zeros or random bits when a disk is deleted, so another user could not retrieve it. Many users are unaware how long data can live. Confidential or sensitive data can often be recovered from disks that have been discarded (Garfinkel and Shelat, 2003).

5.3.6. File System Performance

Access to disk is much slower than access to memory. Reading a memory word might take 10 nsec. Reading from a hard disk might proceed at 10 MB/sec, which is forty times slower per 32-bit word, and to this must be added 510 msec to seek to the track and then wait for the desired sector to arrive under the read head. If only a single word is needed, the memory access is on the order of a million times as fast as disk access. As a result of this difference in access time, many file systems have been designed with various optimizations to improve performance. In this section we will cover three of them.

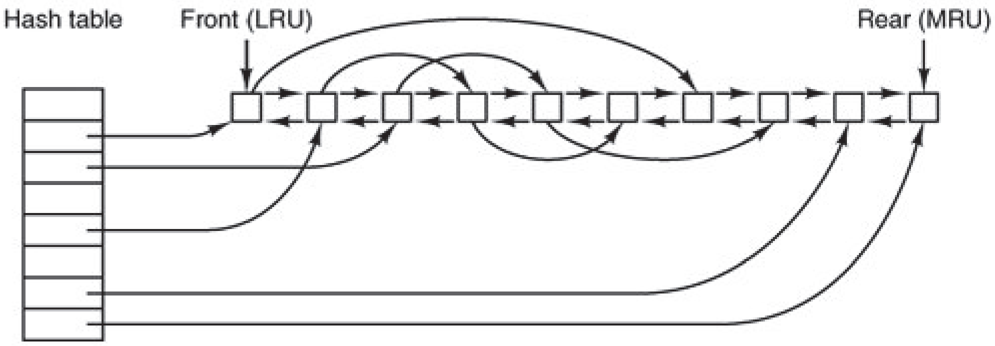
Caching

The most common technique used to reduce disk accesses is the **block cache** or **buffer cache**. (Cache is pronounced "cash" and is derived from the French *cacher*, meaning to hide.) In this context, a cache is a collection of blocks that logically belong on the disk but are being kept in memory for performance reasons.

Various algorithms can be used to manage the cache, but a common one is to check all read requests to see if the needed block is in the cache. If it is, the read request can be satisfied without a disk access. If the block is not in the cache, it is first read into the cache, and then copied to wherever it is needed. Subsequent requests for the same block can be satisfied from the cache.

Operation of the cache is illustrated in Fig. 5-20. Since there are many (often thousands of) blocks in the cache, some way is needed to determine quickly if a given block is present. The usual way is to hash the device and disk address and look up the result in a hash table. All the blocks with the same hash value are chained together on a linked list so the collision chain can be followed.

Figure 5-20. The buffer cache data structures.



When a block has to be loaded into a full cache, some block has to be removed (and rewritten to the disk if it has been modified since being brought in). This situation is very much like paging, and all the usual page replacement algorithms described in Chap. 4, such as FIFO, second chance, and LRU, are applicable. One pleasant difference between paging and caching is that cache references are relatively infrequent, so that it is feasible to keep all the blocks in exact LRU order with linked lists.

In Fig. 5-20, we see that in addition to the collision chains starting at the hash table, there is also a bidirectional list running through all the blocks in the order of usage, with the least recently used block on the front of this list and the most recently used block at the end of this list. When a block is referenced, it can be removed from its position on the bidirectional list and put at the end. In this way, exact LRU order can be maintained.

Unfortunately, there is a catch. Now that we have a situation in which exact LRU is possible, it turns out that LRU is undesirable. The problem has to do with the crashes and file system consistency discussed in the previous section. If a critical block, such as an i-node block, is read into the cache and modified, but not rewritten to the disk, a crash will leave the file system in an inconsistent state. If the i-node block is put at the end of the LRU chain, it may be quite a while before it reaches the front and is rewritten to the disk.

Furthermore, some blocks, such as i-node blocks, are rarely referenced twice within a short interval. These considerations lead to a modified LRU scheme, taking two factors into account:

1. Is the block likely to be needed again soon?
2. Is the block essential to the consistency of the file system?

For both questions, blocks can be divided into categories such as i-node blocks, indirect blocks, directory blocks, full data blocks, and partially full data blocks. Blocks that will probably not be needed again soon go on the front, rather than the rear of the LRU list, so their buffers will be reused quickly. Blocks that might be needed again soon, such as a partly full block that is being written, go on the end of the list, so they will stay around for a long time.

The second question is independent of the first one. If the block is essential to the file system consistency (basically, everything except data blocks), and it has been modified, it should be written to disk immediately, regardless of which end of the LRU list it is put on. By writing critical blocks quickly, we greatly reduce the probability that a crash will wreck the file system. While a user may be unhappy if one of his files is ruined in a crash, he is likely to be far more unhappy if the whole file system is lost.

Even with this measure to keep the file system integrity intact, it is undesirable to keep data blocks in the cache too long before writing them out. Consider the plight of someone who is using a personal computer to write a book. Even if our writer periodically tells the editor to write the file being edited to the disk, there is a good chance that everything will still be in the cache and nothing on the disk. If the system crashes, the file system structure will not be corrupted, but a whole day's work will be lost.

This situation need not happen very often before we have a fairly unhappy user. Systems take two approaches to dealing with it. The UNIX way is to have a system call, `sync`, which forces all the modified blocks out onto the disk immediately. When the system is started up, a program, usually called *update*, is started up in the background to sit in an endless loop issuing `sync` calls, sleeping for 30 sec between calls. As a result, no more than 30 seconds of work is lost due to a system crash, a comforting thought for many people.

The Windows way is to write every modified block to disk as soon as it has been written. Caches in which all modified blocks are written back to the disk immediately are called **write-through caches**. They require more disk I/O than nonwrite-through caches. The difference between these two approaches can be seen when a program writes a 1-KB block full, one character at a time. UNIX will collect all the characters in the cache and write the block out once every 30 seconds, or whenever the block is removed from the cache. Windows will make a disk access for every character written. Of course, most programs do internal buffering, so they normally write not a character, but a line or a larger unit on each `write` system call.

A consequence of this difference in caching strategy is that just removing a (floppy) disk from a UNIX system without doing a `sync` will almost always result in lost data, and frequently in a corrupted file system as well. With Windows, no problem arises. These differing strategies were chosen because UNIX was developed in an environment in which all disks were hard disks and not removable, whereas Windows started out in the floppy disk world. As hard disks became the norm, the UNIX approach, with its better efficiency, became the norm, and is also used now on Windows for hard disks.

Block Read Ahead

A second technique for improving perceived file system performance is to try to get blocks into the cache before they are needed to increase the hit rate. In particular, many files are read sequentially. When the file system is asked to produce block *k* in a file, it does that, but when it is finished, it makes a sneaky check in the cache to see if block *k* + 1 is already there. If it is not, it schedules a read for block *k* + 1 in the hope that when it is needed, it will have already arrived in the cache. At the very least, it will be on the way.

Of course, this read ahead strategy only works for files that are being read sequentially. If a file is being randomly accessed, read ahead does not help. In fact, it hurts by tying up disk bandwidth reading in useless blocks and removing potentially useful blocks from the cache (and possibly tying up more disk bandwidth writing them back to disk if they are dirty). To see whether read ahead is worth doing, the file system can keep track of the access patterns to each open file. For example, a bit associated with each file can keep track of whether the file is in "sequential access mode" or "random access mode." Initially, the file is given the benefit of the doubt and put in sequential access mode. However, whenever a seek is done, the bit is cleared. If sequential reads start happening again, the bit is set once again. In this way, the file system can make a reasonable guess about whether it should read ahead or not. If it gets it wrong once it a while, it is not a disaster, just a little bit of wasted disk bandwidth.

Reducing Disk Arm Motion

Caching and read ahead are not the only ways to increase file system performance. Another important technique is to reduce the amount of disk arm motion by putting blocks that are likely to be accessed in sequence close to each other, preferably in the same cylinder. When an output file is written, the file system has to allocate the blocks one at a time, as they are needed. If the free blocks are recorded in a bitmap, and the whole bitmap is in main memory, it is easy enough to choose a free block as close as possible to the previous block. With a free list, part of which is on disk, it is much harder to allocate blocks close together.

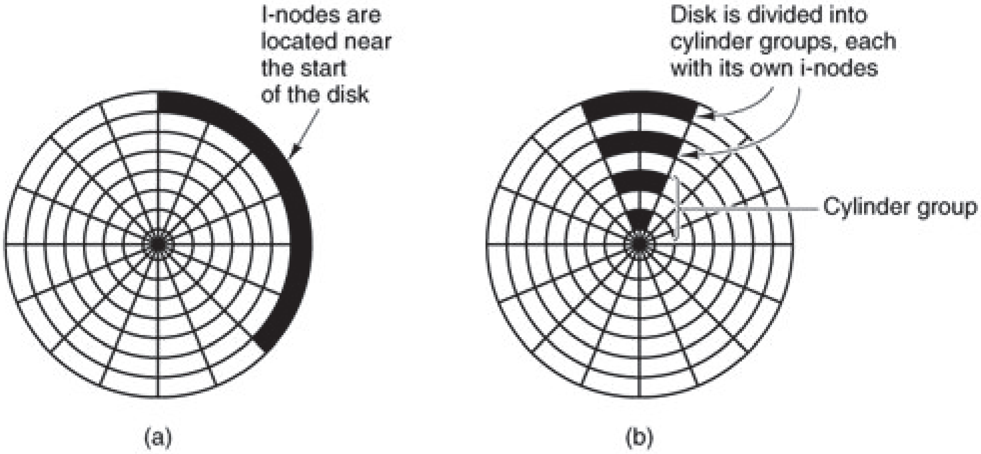
However, even with a free list, some block clustering can be done. The trick is to keep track of disk storage not in blocks, but in groups of consecutive blocks. If sectors consist of 512 bytes, the system could use 1-KB blocks (2 sectors) but allocate disk storage in units of 2 blocks (4 sectors).

This is not the same as having a 2-KB disk blocks, since the cache would still use 1-KB blocks and disk transfers would still be 1 KB but reading a file sequentially on an otherwise idle system would reduce the number of seeks by a factor of two, considerably improving performance. A variation on the same theme is to take account of rotational positioning. When allocating blocks, the system attempts to place consecutive blocks in a file in the same cylinder.

Another performance bottleneck in systems that use i-nodes or anything equivalent to i-nodes is that reading even a short file requires two disk accesses: one for the i-node and one for the block. The usual i-node placement is shown in Fig. 5-21(a). Here all the i-nodes are near the beginning of the disk, so the average distance between an i-node and its blocks will be about half the number of cylinders, requiring long seeks.

Figure 5-21. (a) I-nodes placed at the start of the disk. (b) Disk divided into cylinder groups, each with its own blocks and i-nodes.

[View full size image]


I-nodes are located near the start of the disk
Disk is divided into cylinder groups, each with its own i-nodes
Cylinder group
(a)
(b)

One easy performance improvement is to put the i-nodes in the middle of the disk, rather than at the start, thus reducing the average seek between the i-node and the first block by a factor of two. Another idea, shown in Fig. 5-21(b), is to divide the disk into cylinder groups, each with its own i-nodes, blocks, and free list (McKusick et al., 1984). When creating a new file, any i-node can be chosen, but an attempt is made to find a block in the same cylinder group as the i-node. If none is available, then a block in a nearby cylinder group is used.

5.3.7. Log-Structured File Systems

Changes in technology are putting pressure on current file systems. In particular, CPUs keep getting faster, disks are becoming much bigger and cheaper (but not much faster), and memories are growing exponentially in size. The one parameter that is not improving by leaps and bounds is disk seek time. The combination of these factors means that a performance bottleneck is arising in many file systems. Research done at Berkeley attempted to alleviate this problem by designing a completely new kind of file system, LFS (the **Log-structured File System**). In this section we will briefly describe how LFS works. For a more complete treatment, see Rosenblum and Ousterhout (1991).

The idea that drove the LFS design is that as CPUs get faster and RAM memories get larger, disk caches are also increasing rapidly. Consequently, it is now possible to satisfy a very substantial fraction of all read requests directly from the file system cache, with no disk access needed. It follows from this observation, that in the future, most disk accesses will be writes, so the read-ahead mechanism used in some file systems to fetch blocks before they are needed no longer gains much performance.

To make matters worse, in most file systems, writes are done in very small chunks. Small writes are highly inefficient, since a 50-µsec disk write is often preceded by a 10-msec seek and a 4-msec rotational delay. With these parameters, disk efficiency drops to a fraction of 1 percent.

To see where all the small writes come from, consider creating a new file on a UNIX system. To write this file, the i-node for the directory, the directory block, the i-node for the file, and the file itself must all be written. While these writes can be delayed, doing so exposes the file system to serious consistency problems if a crash occurs before the writes are done. For this reason, the i-node writes are generally done immediately.

From this reasoning, the LFS designers decided to re-implement the UNIX file system in such a way as to achieve the full bandwidth of the disk, even in the face of a workload consisting in large part of small random writes. The basic idea is to structure the entire disk as a log. Periodically, and also when there is a special need for it, all the pending writes being buffered in memory are collected into a single segment and written to the disk as a single contiguous segment at the end of the log. A single segment may thus contain i-nodes, directory blocks, data blocks, and other kinds of blocks all mixed together. At the start of each segment is a segment summary, telling what can be found in the segment. If the average segment can be made to be about 1 MB, almost the full bandwidth of the disk can be utilized.

In this design, i-nodes still exist and have the same structure as in UNIX, but they are now scattered all over the log, instead of being at a fixed position on the disk. Nevertheless, when an i-node is located, locating the blocks is done in the usual way. Of course, finding an i-node is now much harder, since its address cannot simply be calculated from its i-node number, as in UNIX. To make it possible to find i-nodes, an i-node map, indexed by i-node number, is maintained. Entry *i* in this map points to i-node *i* on the disk. The map is kept on disk, but it is also cached, so the most heavily used parts will be in memory most of the time in order to improve performance.

To summarize what we have said so far, all writes are initially buffered in memory, and periodically all the buffered writes are written to the disk in a single segment, at the end of the log. Opening a file now consists of using the map to locate the i-node for the file. Once the i-node has been located, the addresses of the blocks can be found from it. All of the blocks will themselves be in segments, somewhere in the log.

If disks were infinitely large, the above description would be the entire story. However, real disks are finite, so eventually the log will occupy the entire disk, at which time no new segments can be written to the log. Fortunately, many existing segments may have blocks that are no longer needed, for example, if a file is overwritten, its i-node will now point to the new blocks, but the old ones will still be occupying space in previously written segments.

To deal with this problem, LFS has a **cleaner** thread that spends its time scanning the log circularly to compact it. It starts out by reading the summary of the first segment in the log to see which i-nodes and files are there. It then checks the current i-node map to see if the i-nodes are still current and file blocks are still in use. If not, that information is discarded. The i-nodes and blocks that are still in use go into memory to be written out in the next segment. The original segment is then marked as free, so the log can use it for new data. In this manner, the cleaner moves along the log, removing old segments from the back and putting any live data into memory for rewriting in the next segment. Consequently, the disk is a big circular buffer, with the writer thread adding new segments to the front and the cleaner thread removing old ones from the back.

The bookkeeping here is nontrivial, since when a file block is written back to a new segment, the i-node of the file (somewhere in the log) must be located, updated, and put into memory to be written out in the next segment. The i-node map must then be updated to point to the new copy. Nevertheless, it is possible to do the administration, and the performance results show that all this complexity is worthwhile. Measurements given in the papers cited above show that LFS outperforms UNIX by an order of magnitude on small writes, while having a performance that is as good as or better than UNIX for reads and large writes.

PREV NEXT

PREV NEXT

5.4. Security

File systems generally contain information that is highly valuable to their users. Protecting this information against unauthorized usage is therefore a major concern of all file systems. In the following sections we will look at a variety of issues concerned with security and protection. These issues apply equally well to timesharing systems as to networks of personal computers connected to shared servers via local area networks.

5.4.1. The Security Environment

People frequently use the terms "security" and "protection" interchangeably. Nevertheless, it is frequently useful to make a distinction between the general problems involved in making sure that files are not read or modified by unauthorized persons, which include technical, administrative, legal, and political issues on the one hand, and the specific operating system mechanisms used to provide security, on the other. To avoid confusion, we will use the term **security** to refer to the overall problem, and the term **protection mechanisms** to refer to the specific operating system mechanisms used to safeguard information in the computer. The boundary between them is not well defined, however. First we will look at security to see what the nature of the problem is. Later on in the chapter we will look at the protection mechanisms and models available to help achieve security.

Security has many facets. Three of the more important ones are the nature of the threats, the nature of intruders, and accidental data loss. We will now look at these in turn.

Threats

From a security perspective, computer systems have three general goals, with corresponding threats to them, as listed in Fig. 5-22. The first one, **data confidentiality**, is concerned with having secret data remain secret. More specifically, if the owner of some data has decided that these data are only to be made available to certain people and no others, the system should guarantee that release of the data to unauthorized people does not occur. As a bare minimum, the owner should be able to specify who can see what, and the system should enforce these specifications.

Figure 5-22. Security goals and threats.

(This item is displayed on page 527 in the print version)

Goal	Threat
Data confidentiality	Exposure of data
Data integrity	Tampering with data

Goal	Threat
System availability	Denial of service

The second goal, **data integrity**, means that unauthorized users should not be able to modify any data without the owner's permission. Data modification in this context includes not only changing the data, but also removing data and adding false data as well. If a system cannot guarantee that data deposited in it remain unchanged until the owner decides to change them, it is not worth much as an information system. Integrity is usually more important than confidentiality.

The third goal, **system availability**, means that nobody can disturb the system to make it unusable. Such **denial of service** attacks are increasingly common. For example, if a computer is an Internet server, sending a flood of requests to it may cripple it by eating up all of its CPU time just examining and discarding incoming requests. If it takes, say, 100µsec to process an incoming request to read a Web page, then anyone who manages to send 10,000 requests/sec can wipe it out. Reasonable models and technology for dealing with attacks on confidentiality and integrity are available; foiling denial-of-services attacks is much harder.

Another aspect of the security problem is **privacy**: protecting individuals from misuse of information about them. This quickly gets into many legal and moral issues. Should the government compile dossiers on everyone in order to catch *X*-cheaters, where *X* is "welfare" or "tax," depending on your politics? Should the police be able to look up anything on anyone in order to stop organized crime? Do employers and insurance companies have rights? What happens when these rights conflict with individual rights? All of these issues are extremely important but are beyond the scope of this book.

Intruders

Most people are pretty nice and obey the law, so why worry about security? Because there are unfortunately a few people around who are not so nice and want to cause trouble (possibly for their own commercial gain). In the security literature, people who are nosing around places where they have no business being are called **intruders** or sometimes **adversaries**. Intruders act in two different ways. Passive intruders just want to read files they are not authorized to read. Active intruders are more malicious; they want to make unauthorized changes. When designing a system to be secure against intruders, it is important to keep in mind the kind of intruder one is trying to protect against. Some common categories are

1. Casual prying by nontechnical users. Many people have personal computers on their desks that are connected to a shared file server, and human nature being what it is, some of them will read other people's electronic mail and other files if no barriers are placed in the way. Most UNIX systems, for example, have the default that all newly created files are publicly readable.

2. Snooping by insiders. Students, system programmers, operators, and other technical personnel often consider it to be a personal challenge to break the security of the local

computer system. They often are highly skilled and are willing to devote a substantial amount of time to the effort.

3. Determined attempts to make money. Some bank programmers have attempted to steal from the bank they were working for. Schemes have varied from changing the software to truncate rather than round interest, keeping the fraction of a cent for themselves, to siphoning off accounts not used in years, to blackmail ("Pay me or I will destroy all the bank's records.").

4. Commercial or military espionage. Espionage refers to a serious and well-funded attempt by a competitor or a foreign country to steal programs, trade secrets, patentable ideas, technology, circuit designs, business plans, and so forth. Often this attempt will involve wiretapping or even erecting antennas directed at the computer to pick up its electromagnetic radiation.

It should be clear that trying to keep a hostile foreign government from stealing military secrets is quite a different matter from trying to keep students from inserting a funny message-of-the-day into the system. The amount of effort needed for security and protection clearly depends on who the enemy is thought to be.

Malicious Programs

Another category of security pest is malicious programs, sometimes called **malware**. In a sense, a writer of malware is also an intruder, often with high technical skills. The difference between a conventional intruder and malware is that the former refers to a person who is personally trying to break into a system to cause damage whereas the latter is a program written by such a person and then released into the world. Some malware seems to have been written just to cause damage, but some is targeted more specifically. It is becoming a huge problem and a great deal has been written about it (Aycock and Barker, 2005; Cerf, 2005; Ledin, 2005; McHugh and Deek, 2005; Treese, 2004; and Weiss, 2005)

The most well known kind of malware is the **virus**. Basically a virus is a piece of code that can reproduce itself by attaching a copy of itself to another program, analogous to how biological viruses reproduce. The virus can do other things in addition to reproducing itself. For example, it can type a message, display an image on the screen, play music, or something else harmless. Unfortunately, it can also modify, destroy, or steal files (by e-mailing them somewhere).

Another thing a virus can do is to render the computer unusable as long as the virus is running. This is called a **DOS** (**Denial Of Service**) attack. The usual approach is to consume resources wildly, such as the CPU, or filling up the disk with junk. Viruses (and the other forms of malware to be described) can also be used to cause a **DDOS** (**Distributed Denial Of Service**) attack. In this case the virus does not do anything immediately upon infecting a computer. At a predetermined date and time thousands of copies of the virus on computers all over the world start requesting web pages or other network services from their target, for instance the Web site of a political party or a corporation. This can overload the targeted server and the networks that service it.

Malware is frequently created for profit. Much (if not most) unwanted junk e-mail ("spam") is relayed to its final destinations by networks of computers that have been infected by viruses or other forms of malware. A computer infected by such a rogue program becomes a slave, and reports its status to its master, somewhere on the Internet. The master then sends spam to be relayed to all the e-mail addresses that can be gleaned from e-mail address books and other files on the slave. Another kind of malware for profit scheme installs a **key logger** on an infected computer. A key logger records everything typed at the keyboard. It is not too difficult to filter this data and extract information such as username password combinations or credit card numbers and expiration dates. This information is then sent back to a master where it can be used or sold for criminal use.

Related to the virus is the **worm**. Whereas a virus is spread by attaching itself to another program, and is executed when its host program is executed, a worm is a free-standing program. Worms spread by using networks to transmit copies of themselves to other computers. Windows systems always have a *Startup* directory for each user; any program in that folder will be executed when the user logs in. So all the worm has to do is arrange to put itself (or a shortcut to itself) in the *Startup* directory on a remote system. Other ways exist, some much more difficult to detect, to cause a remote computer to execute a program file that has been copied to its file system. The effects of a worm can be the same as those of a virus. Indeed, the distinction between a virus and a worm is not always clear; some malware uses both methods to spread.

Another category of malware is the **Trojan horse**. This is a program that apparently performs a valid functionperhaps it is a game or a supposedly "improved" version of a useful utility. But when the Trojan horse is executed some other function is performed, perhaps launching a worm or virus or performing one of the nasty things that malware does. The effects of a Trojan horse are likely to be subtle and stealthy. Unlike worms and viruses, Trojan horses are voluntarily downloaded by users, and as soon as they are recognized for what they are and the word gets out, a Trojan horse will be deleted from reputable download sites.

Another kind of malware is the **logic bomb**. This device is a piece of code written by one of a company's (currently employed) programmers and secretly inserted into the production operating system. As long as the programmer feeds it its daily password, it does nothing. However, if the programmer is suddenly fired and physically removed from the premises without warning, the next day the logic bomb does not get its password, so it goes off.

Going off might involve clearing the disk, erasing files at random, carefully making hard-to-detect changes to key programs, or encrypting essential files. In the latter case, the company has a tough choice about whether to call the police (which may or may not result in a conviction many months later) or to give in to this blackmail and to rehire the ex-programmer as a "consultant" for an astronomical sum to fix the problem (and hope that he does not plant new logic bombs while doing so).

Yet another form of malware is **spyware**. This is usually obtained by visiting a Web site. In its simplest form spyware may be nothing more than a **cookie**. Cookies are small files exchanged between web browsers and web servers. They have a legitimate purpose. A cookie contains some information that will allow the Web site to identify you. It is like the ticket you get when you leave a bicycle to be repaired. When you return to the shop, your half of the ticket gets matched with your bicycle (and its repair bill). Web connections are not persistent, so, for example, if you indicate an interest in buying this book when visiting an online bookstore, the bookstore asks your browser to accept a cookie. When you have finished browsing and perhaps have selected other books to buy, you click on the page where your order is finalized. At that point the web server asks your browser to return the cookies it has stored from the current session, It can use the information in these to generate the list of items you have said you want to buy.

Normally, cookies used for a purpose like this expire quickly. They are quite useful, and e-commerce depends upon them. But some Web sites use cookies for purposes that are not so benign. For instance, advertisements on Web sites are often furnished by companies other than the information provider. Advertisers pay Web site owners for this privilege. If a cookie is placed when you visit a page with information about, say, bicycle equipment, and you then go to another

Web site that sells clothing, the same advertising company may provide ads on this page, and may collect cookies you obtained elsewhere. Thus you may suddenly find yourself viewing ads for special gloves or jackets especially made for cyclists. Advertisers can collect a lot of information about your interests this way; you may not want to share so much information about yourself.

What is worse, there are various ways a Web site may be able to download executable program code to your computer. Most browsers accept **plug-ins** to add additional function, such as displaying new kinds of files. Users often accept offers for new plugins without knowing much about what the plugin does. Or a user may willingly accept an offer to be provided with a new cursor for the desktop that looks like a dancing kitten. And a bug in a web browser may allow a remote site to install an unwanted program, perhaps after luring the user to a page that has been carefully constructed to take advantage of the vulnerability. Any time a program is accepted from another source, voluntarily or not, there is a risk it could contain code that does you harm.

Accidental Data Loss

In addition to threats caused by malicious intruders, valuable data can be lost by accident. Some of the common causes of accidental data loss are

1. Acts of God: fires, floods, earthquakes, wars, riots, or rats gnawing tapes or floppy disks.
2. Hardware or software errors: CPU malfunctions, unreadable disks or tapes, telecommunication errors, program bugs.
3. Human errors: incorrect data entry, wrong tape or disk mounted, wrong program run, lost disk or tape, or some other mistake.

Most of these can be dealt with by maintaining adequate backups, preferably far away from the original data. While protecting data against accidental loss may seem mundane compared to protecting against clever intruders, in practice, probably more damage is caused by the former than the latter.

5.4.2. Generic Security Attacks

Finding security flaws is not easy. The usual way to test a system's security is to hire a group of experts, known as **tiger teams** or **penetration teams**, to see if they can break in. Hebbard et al. (1980) tried the same thing with graduate students. In the course of the years, these penetration teams have discovered a number of areas in which systems are likely to be weak. Below we have listed some of the more common attacks that are often successful. When designing a system, be sure it can withstand attacks like these.

1. Request memory pages, disk space, or tapes and just read them. Many systems do not erase them before allocating them, and they may be full of interesting information written by the previous owner.

2. Try illegal system calls, or legal system calls with illegal parameters, or even legal system calls with legal but unreasonable parameters. Many systems can easily be confused.

3. Start logging in and then hit DEL, RUBOUT or BREAK halfway through the login sequence. In some systems, the password checking program will be killed and the login considered successful.

4. Try modifying complex operating system structures kept in user space (if any). In some systems (especially on mainframes), to open a file, the program builds a large data structure containing the file name and many other parameters and passes it to the system. As the file is read and written, the system sometimes updates the structure itself. Changing these fields can wreak havoc with the security.

5. Spoof the user by writing a program that types "login:" on the screen and go away. Many users will walk up to the terminal and willingly tell it their login name and password, which the program carefully records for its evil master.

6. Look for manuals that say "Do not do *X*." Try as many variations of *X* as possible.

7. Convince a system programmer to change the system to skip certain vital security checks for any user with your login name. This attack is known as a **trapdoor**.

8. All else failing, the penetrator might find the computer center director's secretary and offer a large bribe. The secretary probably has easy access to all kinds of wonderful information, and is usually poorly paid. Do not underestimate problems caused by personnel.

These and other attacks are discussed by Linde (1975). Many other sources of information on security and testing security can be found, especially on the Web. A recent Windows-oriented work is Johansson and Riley (2005).

5.4.3. Design Principles for Security

Saltzer and Schroeder (1975) have identified several general principles that can be used as a guide to designing secure systems. A brief summary of their ideas (based on experience with MULTICS) is given below.

First, the system design should be public. Assuming that the intruder will not know how the system works serves only to delude the designers.

Second, the default should be no access. Errors in which legitimate access is refused will be reported much faster than errors in which unauthorized access is allowed.

Third, check for current authority. The system should not check for permission, determine that access is permitted, and then squirrel away this information for subsequent use. Many systems check for permission when a file is opened, and not afterward. This means that a user who opens a file, and keeps it open for weeks, will continue to have access, even if the owner has long since changed the file protection.

Fourth, give each process the least privilege possible. If an editor has only the authority to access the file to be edited (specified when the editor is invoked), editors with Trojan horses will not be able to do much damage. This principle implies a fine-grained protection scheme. We will discuss such schemes later in this chapter.

Fifth, the protection mechanism should be simple, uniform, and built into the lowest layers of the system. Trying to retrofit security to an existing insecure system is nearly impossible. Security, like correctness, is not an add-on feature.

Sixth, the scheme chosen must be psychologically acceptable. If users feel that protecting their files is too much work, they just will not do it. Nevertheless, they will complain loudly if something goes wrong. Replies of the form "It is your own fault" will generally not be well received.

5.4.4. User Authentication

Many protection schemes are based on the assumption that the system knows the identity of each user. The problem of identifying users when they log in is called **user authentication**. Most authentication methods are based on identifying something the user knows, something the user has, or something the user is.

Passwords

The most widely used form of authentication is to require the user to type a password. Password protection is easy to understand and easy to implement. In UNIX it works like this: The login program asks the user to type his name and password. The password is immediately encrypted. The login program then reads the password file, which is a series of ASCII lines, one per user, until it finds the line containing the user's login name. If the (encrypted) password contained in this line matches the encrypted password just computed, the login is permitted, otherwise it is refused.

Password authentication is easy to defeat. One frequently reads about groups of high school, or even junior high school students who, with the aid of their trusty home computers, have broken into some top secret system owned by a large corporation or government agency. Virtually all the time the break-in consists of guessing a user name and password combination.

Although more recent studies have been made (e.g., Klein, 1990) the classic work on password security remains the one done by Morris and Thompson (1979) on UNIX systems. They compiled a list of likely passwords: first and last names, street names, city names, words from a moderate-sized dictionary (also words spelled backward), license plate numbers, and short strings of random characters.

They then encrypted each of these using the known password encryption algorithm and checked to see if any of the encrypted passwords matched entries in their list. Over 86 percent of all passwords turned up in their list.

If all passwords consisted of 7 characters chosen at random from the 95 printable ASCII characters, the search space becomes 95^7, which is about 7x 10^{13}. At 1000 encryptions per second, it would take 2000 years to build the list to check the password file against. Furthermore, the list would fill 20 million magnetic tapes. Even requiring passwords to contain at least one lowercase character, one uppercase character, and one special character, and be at least seven characters long would be a major improvement over unrestricted user-chosen passwords.

Even if it is considered politically impossible to require users to pick reasonable passwords, Morris and Thompson have described a technique that renders their own attack (encrypting a large number of passwords in advance) almost useless. Their idea is to associate an *n*-bit random number with each password. The random number is changed whenever the password is changed. The random number is stored in the password file in unencrypted form, so that everyone can read it. Instead of just storing the encrypted password in the password file, the password and the random number are first concatenated and then encrypted together. This encrypted result is stored in the password file.

Now consider the implications for an intruder who wants to build up a list of likely passwords, encrypt them, and save the results in a sorted file, *f*, so that any encrypted password can be looked up easily. If an intruder suspects that *Marilyn* might be a password, it is no longer sufficient just to encrypt *Marilyn* and put the result in *f*. He has to encrypt 2^n strings, such as *Marilyn0000*, *Marilyn0001*, *Marilyn0002*, and so forth and enter all of them in *f*. This technique increases the size of *f* by 2^n. UNIX uses this method with $n= 12$. It is known as **salting** the password file. Some versions of UNIX make the password file itself unreadable but provide a program to look up entries upon request, adding just enough delay to greatly slow down any attacker.

Although this method offers protection against intruders who try to precompute a large list of encrypted passwords, it does little to protect a user *David* whose password is also *David*. One way to encourage people to pick better passwords is to have the computer offer advice. Some computers have a program that generates random easy-to-pronounce nonsense words, such as *fotally*, *garbungy*, or *bipitty* that can be used as passwords (preferably with some upper case and special characters thrown in).

Other computers require users to change their passwords regularly, to limit the damage done if a password leaks out. The most extreme form of this approach is the **one-time password**. When one-time passwords are used, the user gets a book containing a list of passwords. Each login uses the next password in the list. If an intruder ever discovers a password, it will not do him any good, since next time a different password must be used. It is suggested that the user try to avoid losing the password book.

It goes almost without saying that while a password is being typed in, the computer should not display the typed characters, to keep them from prying eyes near the terminal. What is less obvious is that passwords should never be stored in the computer in unencrypted form. Furthermore, not even the computer center management should have unencrypted copies. Keeping unencrypted passwords anywhere is looking for trouble.

A variation on the password idea is to have each new user provide a long list of questions and answers that are then stored in the computer in encrypted form. The questions should be chosen so that the user does not need to write them down. In other words, they should be things no one forgets. Typical questions are:

1. Who is Marjolein's sister?

2. On what street was your elementary school?

3. What did Mrs. Woroboff teach?

At login, the computer asks one of them at random and checks the answer.

Another variation is **challenge-response.** When this is used, the user picks an algorithm when signing up as a user, for example x^2. When the user logs in, the computer types an argument, say 7, in which case the user types 49. The algorithm can be different in the morning and afternoon, on different days of the week, from different terminals, and so on.

Physical Identification

A completely different approach to authorization is to check to see if the user has some item, normally a plastic card with a magnetic stripe on it. The card is inserted into the terminal, which then checks to see whose card it is. This method can be combined with a password, so a user can only log in if he (1) has the card and (2) knows the password. Automated cash-dispensing machines usually work this way.

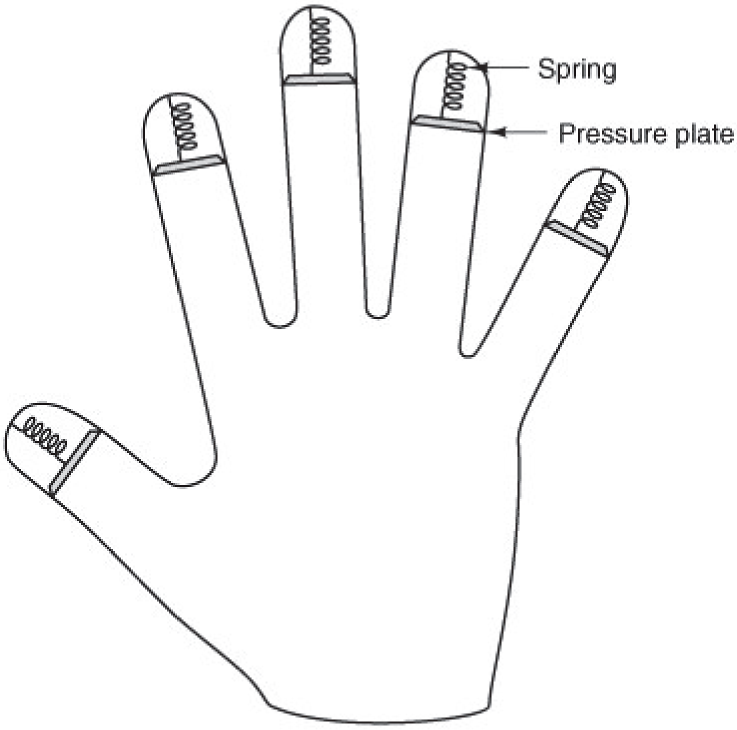
Yet another approach is to measure physical characteristics that are hard to forge. For example, a fingerprint or a voiceprint reader in the terminal could verify the user's identity. (It makes the search go faster if the user tells the computer who he is, rather than making the computer compare the given fingerprint to the entire data base.) Direct visual recognition is not yet feasible but may be one day.

Another technique is signature analysis. The user signs his name with a special pen connected to the terminal, and the computer compares it to a known specimen stored on line. Even better is not to compare the signature, but compare the pen motions made while writing it. A good forger may be able to copy the signature, but will not have a clue as to the exact order in which the strokes were made.

Finger length analysis is surprisingly practical. When this is used, each terminal has a device like the one of Fig. 5-23. The user inserts his hand into it, and the length of each of his fingers is measured and checked against the data base.

Figure 5-23. A device for measuring finger length.

(This item is displayed on page 536 in the print version)



We could go on and on with more examples, but two more will help make an important point. Cats and other animals mark off their territory by urinating around its perimeter. Apparently cats can identify each other this way. Suppose that someone comes up with a tiny device capable of doing an instant urinalysis, thereby providing a foolproof identification. Each terminal could be equipped with one of these devices, along with a discreet sign reading: "For login, please deposit sample here." This might be an absolutely unbreakable system, but it would probably have a fairly serious user acceptance problem.

The same could be said of a system consisting of a thumbtack and a small spectrograph. The user would be requested to jab his thumb against the thumbtack, thus extracting a drop of blood for spectrographic analysis. The point is that any authentication scheme must be psychologically acceptable to the user community. Finger-length measurements probably will not cause any problem, but even something as nonintrusive as storing fingerprints on line may be unacceptable to many people.

Countermeasures

Computer installations that are really serious about securityand few are until the day after an intruder has broken in and done major damageoften take steps to make unauthorized entry much harder. For example, each user could be allowed to log in only from a specific terminal, and only during certain days of the week and hours of the day.

Dial-up telephone lines could be made to work as follows. Anyone can dial up and log in, but after a successful login, the system immediately breaks the connection and calls the user back at an agreed upon number. This measure means than an intruder cannot just try breaking in from any phone line; only the user's (home) phone will do. In any event, with or without call back, the system should take at least 10 seconds to check any password typed in on a dial-up line, and should increase this time after several consecutive unsuccessful login attempts, in order to reduce the rate at which intruders can try. After three failed login attempts, the line should be disconnected for 10 minutes and security personnel notified.

All logins should be recorded. When a user logs in, the system should report the time and terminal of the previous login, so he can detect possible break ins.

The next step up is laying baited traps to catch intruders. A simple scheme is to have one special login name with an easy password (e.g., login name: guest, password: guest). Whenever anyone logs in using this name, the system security specialists are immediately notified. Other traps can be easy-to-find bugs in the operating system and similar things, designed for the purpose of catching intruders in the act. Stoll (1989) has written an entertaining account of the traps he set

PREV NEXT

5.5. Protection Mechanisms

In the previous sections we have looked at many potential problems, some of them technical, some of them not. In the following sections we will concentrate on some of the detailed technical ways that are used in operating systems to protect files and other things. All of these techniques make a clear distinction between policy (whose data are to be protected from whom) and mechanism (how the system enforces the policy). The separation of policy and mechanism is discussed by Sandhu (1993). Our emphasis will be on mechanisms, not policies.

In some systems, protection is enforced by a program called a **reference monitor**. Every time an access to a potentially protected resource is attempted, the system first asks the reference monitor to check its legality. The reference monitor then looks at its policy tables and makes a decision. Below we will describe the environment in which a reference monitor operates.

5.5.1. Protection Domains

A computer system contains many "objects" that need to be protected. These objects can be hardware (e.g., CPUs, memory segments, disk drives, or printers), or they can be software (e.g., processes, files, databases, or semaphores).

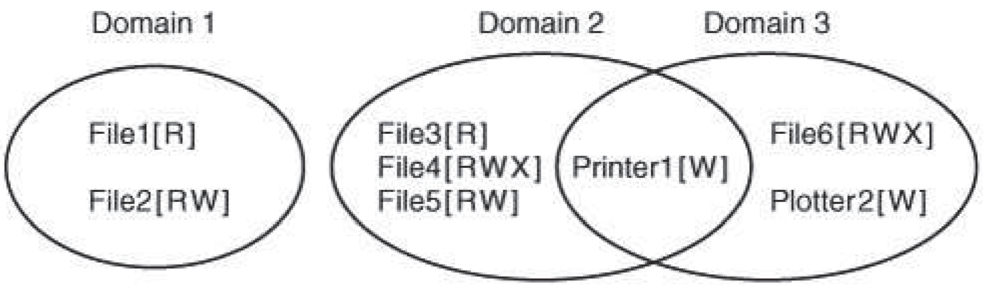
Each object has a unique name by which it is referenced, and a finite set of operations that processes are allowed to carry out on it. The `read` and `write` operations are appropriate to a file; `up` and `down` make sense on a semaphore.

It is obvious that a way is needed to prohibit processes from accessing objects that they are not authorized to access. Furthermore, this mechanism must also make it possible to restrict processes to a subset of the legal operations when that is needed. For example, process *A* may be entitled to read, but not write, file *F*.

In order to discuss different protection mechanisms, it is useful to introduce the concept of a domain. A **domain** is a set of (object, rights) pairs. Each pair specifies an object and some subset of the operations that can be performed on it. A **right** in this context means permission to perform one of the operations. Often a domain corresponds to a single user, telling what the user can do and not do, but a domain can also be more general than just one user.

Figure 5-24 shows three domains, showing the objects in each domain and the rights [Read, Write, eXecute] available on each object. Note that *Printer1* is in two domains at the same time. Although not shown in this example, it is possible for the same object to be in multiple domains, with *different* rights in each one.

Figure 5-24. Three protection domains.



At every instant of time, each process runs in some protection domain. In other words, there is some collection of objects it can access, and for each object it has some set of rights. Processes can also switch from domain to domain during execution. The rules for domain switching are highly system dependent.

To make the idea of a protection domain more concrete, let us look at UNIX. In UNIX, the domain of a process is defined by its UID and GID. Given any (UID, GID) combination, it is possible to make a complete list of all objects (files, including I/O devices represented by special files, etc.) that can be accessed, and whether they can be accessed for reading, writing, or executing. Two processes with the same (UID, GID) combination will have access to exactly the same set of objects. Processes with different (UID, GID) values will have access to a different set of files, although there may be considerable overlap in most cases.

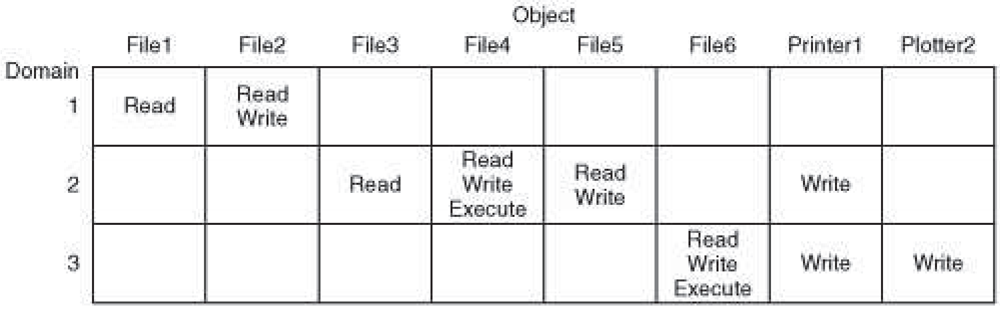
Furthermore, each process in UNIX has two halves: the user part and the kernel part. When the process does a system call, it switches from the user part to the kernel part. The kernel part has access to a different set of objects from the user part. For example, the kernel can access all the pages in physical memory, the entire disk, and all the other protected resources. Thus, a system call causes a domain switch.

When a process does an `exec` on a file with the SETUID or SETGID bit on, it acquires a new effective UID or GID. With a different (UID, GID) combination, it has a different set of files and operations available. Running a program with SETUID or SETGID is also a domain switch, since the rights available change.

An important question is how the system keeps track of which object belongs to which domain. Conceptually, at least, one can envision a large matrix, with the rows being domains and the columns being objects. Each box lists the rights, if any, that the domain contains for the object. The matrix for Fig. 5-24 is shown in Fig. 5-25. Given this matrix and the current domain number, the system can tell if an access to a given object in a particular way from a specified domain is allowed.

Figure 5-25. A protection matrix.

[View full size image]

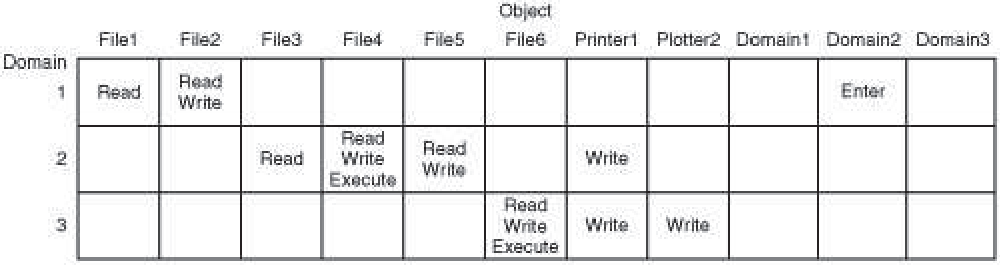


	Object						
Domain	File1	File2	File3	File4	File5	File6	Printer1	Plotter2
1	Read	Read Write						
2			Read	Read Write Execute	Read Write		Write	
3						Read Write Execute	Write	Write

Domain switching itself can be easily included in the matrix model by realizing that a domain is itself an object, with the operation `enter`. Figure 5-26 shows the matrix of Fig. 5-25 again, only now with the three domains as objects themselves. Processes in domain 1 can switch to domain 2, but once there, they cannot go back. This situation models executing a SETUID program in UNIX. No other domain switches are permitted in this example.

Figure 5-26. A protection matrix with domains as objects.

[View full size image]



	Object										
Domain	File1	File2	File3	File4	File5	File6	Printer1	Plotter2	Domain1	Domain2	Domain3
1	Read	Read Write								Enter	
2			Read	Read Write Execute	Read Write		Write				
3						Read Write Execute	Write	Write			

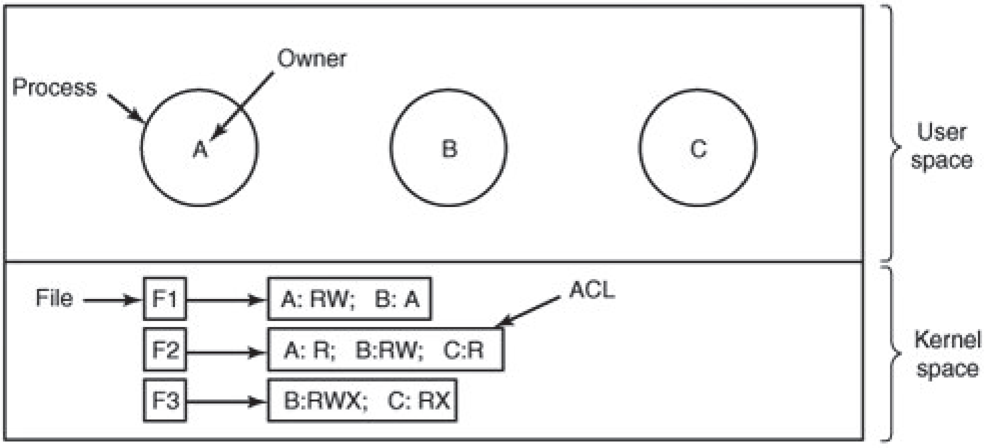
5.5.2. Access Control Lists

In practice, actually storing the matrix of Fig. 5-26 is rarely done because it is large and sparse. Most domains have no access at all to most objects, so storing a very large, mostly empty, matrix is a waste of disk space. Two methods that are practical, however, are storing the matrix by rows or by columns, and then storing only the nonempty elements. The two approaches are surprisingly different. In this section we will look at storing it by column; in the next one we will study storing it by row.

The first technique consists of associating with each object an (ordered) list containing all the domains that may access the object, and how. This list is called the **Access Control List** or **ACL** and is illustrated in Fig. 5-27. Here we see three processes, each belonging to a different domain. *A*, *B*, and *C*, and three files *F1*, *F2*, and *F3*. For simplicity, we will assume that each domain corresponds to exactly one user, in this case, users *A*, *B*, and *C*. Often in the security literature, the users are called **subjects** or **principals**, to contrast them with the things owned, the **objects**, such as files.

Figure 5-27. Use of access control lists to manage file access.

[View full size image]



Each file has an ACL associated with it. File *F1* has two entries in its ACL (separated by a semicolon). The first entry says that any process owned by user *A* may read and write the file. The second entry says that any process owned by user *B* may read the file. All other accesses by these users and all accesses by other users are forbidden. Note that the rights are granted by user, not by process. As far as the protection system goes, any process owned by user *A* can read and write file *F1*. It does not matter if there is one such process or 100 of them. It is the owner, not the process ID, that matters.

File *F2* has three entries in its ACL: *A*, *B*, and *C* can all read the file, and in addition *B* can also write it. No other accesses are allowed. File *F3* is apparently an executable program, since *B* and *C* can both read and execute it. *B* can also write it.

This example illustrates the most basic form of protection with ACLs. More sophisticated systems are often used in practice. To start with, we have only shown three rights so far: read, write, and execute. There may be additional rights as well. Some of these may be generic, that is, apply to all objects, and some may be object specific. Examples of generic rights are `destroy object` and `copy object`. These could hold for any object, no matter what type it is. Object-specific rights might include `append message` for a mailbox object and `sort alphabetically` for a directory `object`.

So far, our ACL entries have been for individual users. Many systems support the concept of a **group** of users. Groups have names and can be included in ACLs. Two variations on the semantics of groups are possible. In some systems, each process has a user ID (UID) and group ID (GID). In such systems, an ACL entry contains entries of the form

```
UID1, GID1: rights1; UID2, GID2: rights2; ...
```

Under these conditions, when a request is made to access an object, a check is made using the caller's UID and GID. If they are present in the ACL, the rights listed are available. If the (UID, GID) combination is not in the list, the access is not permitted.

Using groups this way effectively introduces the concept of a **role**. Consider an installation in which Tana is system administrator, and thus in the group *sysadm*. However, suppose that the company also has some clubs for employees and Tana is a member of the pigeon fanciers club. Club members belong to the group *pigfan* and have access to the company's computers for managing their pigeon database. A portion of the ACL might be as shown in Fig. 5-28.

Figure 5-28. Two access control lists.

File	Access control list
Password	tana, sysadm: RW
Pigeon_data	bill, pigfan: RW; tana, pigfan: RW; ...

If Tana tries to access one of these files, the result depends on which group she is currently logged in as. When she logs in, the system may ask her to choose which of her groups she is currently using, or there might even be different login names and/or passwords to keep them separate. The point of this scheme is to prevent Tana from accessing the password file when she currently has her pigeon fancier's hat on. She can only do that when logged in as the system administrator.

In some cases, a user may have access to certain files independent of which group she is currently logged in as. That case can be handled by introducing **wildcards**, which mean everyone. For example, the entry

```
tana, *: RW
```

for the password file would give Tana access no matter which group she was currently in as.

Yet another possibility is that if a user belongs to any of the groups that have certain access rights, the access is permitted. In this case, a user belonging to multiple groups does not have to specify which group to use at login time. All of them count all of the time. A disadvantage of this approach is that it provides less encapsulation: Tana can edit the password file during a pigeon club meeting.

The use of groups and wildcards introduces the possibility of selectively blocking a specific user from accessing a file. For example, the entry

```
virgil, *: (none); *, *: RW
```

gives the entire world except for Virgil read and write access to the file. This works because the entries are scanned in order, and the first one that applies is taken; subsequent entries are not even examined. A match is found for Virgil on the first entry and the access rights, in this case, (none) are found and applied. The search is terminated at that point. The fact that the rest of the world has access is never even seen.

The other way of dealing with groups is not to have ACL entries consist of (UID, GID) pairs, but to have each entry be a UID or a GID. For example, an entry for the file *pigeon_data* could be

```
debbie: RW; phil: RW; pigfan: RW
```

meaning that Debbie and Phil, and all members of the *pigfan* group have read and write access to the file.

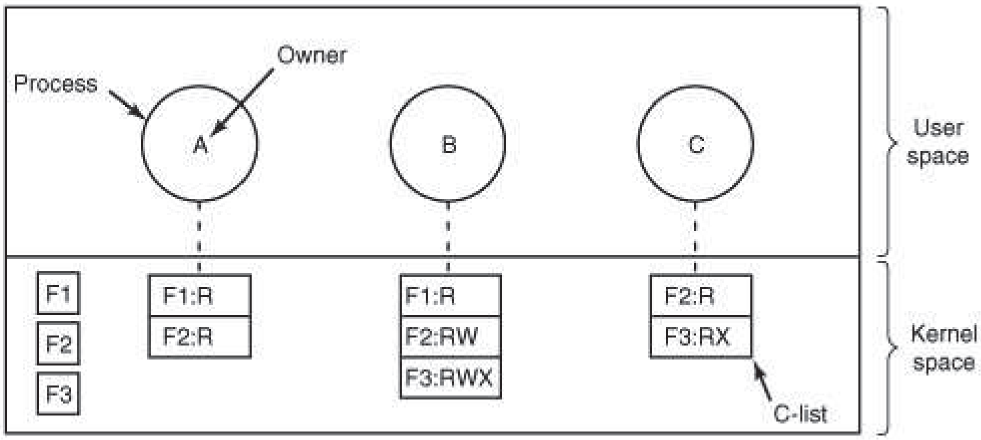
It sometimes occurs that a user or a group has certain permissions with respect to a file that the file owner later wishes to revoke. With access control lists, it is relatively straightforward to revoke a previously granted access. All that has to be done is edit the ACL to make the change. However, if the ACL is checked only when a file is opened, most likely the change will only take effect on future calls to `open`. Any file that is already open will continue to have the rights it had when it was opened, even if the user is no longer authorized to access the file at all.

5.5.3. Capabilities

The other way of slicing up the matrix of Fig. 5-26 is by rows. When this method is used, associated with each process is a list of objects that may be accessed, along with an indication of which operations are permitted on each, in other words, its domain. This list is called a **capability list** or **C-list** and the individual items on it are called **capabilities** (Dennis and Van Horn, 1966; Fabry, 1974). A set of three processes and their capability lists is shown in Fig. 5-29.

Figure 5-29. When capabilities are used, each process has a capability list.

(This item is displayed on page 543 in the print version)



Each capability grants the owner certain rights on a certain object. In Fig. 5-29, the process owned by user *A* can read files *F1* and *F2*, for example. Usually, a capability consists of a file (or more generally, an object) identifier and a bitmap for the various rights. In a UNIX-like system, the file identifier would probably be the i-node number. Capability lists are themselves objects and may be pointed to from other capability lists, thus facilitating sharing of subdomains.

It is fairly obvious that capability lists must be protected from user tampering. Three methods of protecting them are known. The first way requires a **tagged architecture**, a hardware design in which each memory word has an extra (or tag) bit that tells whether the word contains a capability or not. The tag bit is not used by arithmetic, comparison, or similar ordinary instructions, and it can be modified only by programs running in kernel mode (i.e., the operating system). Tagged-architecture machines have been built and can be made to work well (Feustal, 1972). The IBM AS/400 is a popular example.

The second way is to keep the C-list inside the operating system. Capabilities are then referred to by their position in the capability list. A process might say: "Read 1 KB from the file pointed to by capability 2." This form of addressing is similar to using file descriptors in UNIX. Hydra worked this way (Wulf et al., 1974).

The third way is to keep the C-list in user space, but manage the capabilities cryptographically so that users cannot tamper with them. This approach is particularly suited to distributed systems and works as follows. When a client process sends a message to a remote server, for example, a file server, to create an object for it, the server creates the object and generates a long random number, the check field, to go with it. A slot in the server's file table is reserved for the object and the check field is stored there along with the addresses of the disk blocks, etc. In UNIX terms, the check field is stored on the server in the i-node. It is not sent back to the user and never put on the network. The server then generates and returns a capability to the user of the form shown in Fig. 5-30.

Figure 5-30. A cryptographically-protected capability.

Server	Object	Rights	f(Objects,Rights,Check)

The capability returned to the user contains the server's identifier, the object number (the index into the server's tables, essentially, the i-node number), and the rights, stored as a bitmap. For a newly created object, all the rights bits are turned on. The last field consists of the concatenation of the object, rights, and check field run through a cryptographically-secure one-way function, *f*, of the kind we discussed earlier.

When the user wishes to access the object, it sends the capability to the server as part of the request. The server then extracts the object number to index into its tables to find the object. It then computes *f* (*Object*, *Rights*, *Check*) taking the first two parameters from the capability itself and the third one from its own tables. If the result agrees with the fourth field in the capability, the request is honored; otherwise, it is rejected. If a user tries to access someone else's object, he will not be able to fabricate the fourth field correctly since he does not know the check field, and the request will be rejected.

A user can ask the server to produce and return a weaker capability, for example, for read-only access. First the server verifies that the capability is valid. If so, if computes *f* (*Object*, *New_rights*, *Check*) and generates a new capability putting this value in the fourth field. Note that the original *Check* value is used because other outstanding capabilities depend on it.

This new capability is sent back to the requesting process. The user can now give this to a friend by just sending it in a message. If the friend turns on rights bits that should be off, the server will detect this when the capability is used since the *f* value will not correspond to the false rights field. Since the friend does not know the true check field, he cannot fabricate a capability that corresponds to the false rights bits. This scheme was developed for the Amoeba system and used extensively there (Tanenbaum et al., 1990).

In addition to the specific object-dependent rights, such as read and execute, capabilities (both kernel and cryptographically-protected) usually have **generic rights** which are applicable to all objects. Examples of generic rights are

1. Copy capability: create a new capability for the same object.
2. Copy object: create a duplicate object with a new capability.
3. Remove capability: delete an entry from the C-list; object unaffected.
4. Destroy object: permanently remove an object and a capability.

A last remark worth making about capability systems is that revoking access to an object is quite difficult in the kernel-managed version. It is hard for the system to find all the outstanding capabilities for any object to take them back, since they may be stored in C-lists all over the disk. One approach is to have each capability point to an indirect object, rather than to the object itself. By having the indirect object point to the real object, the system can always break that connection, thus invalidating the capabilities. (When a capability to the indirect object is later presented to the system, the user will discover that the indirect object is now pointing to a null object.)

In the Amoeba scheme, revocation is easy. All that needs to be done is change the check field stored with the object. In one blow, all existing capabilities are invalidated. However, neither scheme allows selective revocation, that is, taking back, say, John's permission, but nobody else's. This defect is generally recognized to be a problem with all capability systems.

Another general problem is making sure the owner of a valid capability does not give a copy to 1000 of his best friends. Having the kernel manage capabilities, as in Hydra, solves this problem, but this solution does not work well in a distributed system such as Amoeba.

On the other hand, capabilities solve the problem of sandboxing mobile code very elegantly. When a foreign program is started, it is given a capability list containing only those capabilities that the machine owner wants to give it, such as the ability to write on the screen and the ability to read and write files in one scratch directory just created for it. If the mobile code is put into its own process with only these limited capabilities, it will not be able to access any other system resources and thus be effectively confined to a sandbox without the need to modify its code or run it interpretively. Running code with as few access rights as possible is known as the **principle of least privilege** and is a powerful guideline for producing secure systems.

Briefly summarized, ACLs and capabilities have somewhat complementary properties. Capabilities are very efficient because if a process says "Open the file pointed to by capability 3," no checking is needed. With ACLs, a (potentially long) search of the ACL may be needed. If groups are not supported, then granting everyone read access to a file requires enumerating all users in the ACL. Capabilities also allow a process to be encapsulated easily, whereas ACLs do not. On the other hand, ACLs allow selective revocation of rights, which capabilities do not. Finally, if an object is removed and the capabilities are not or the capabilities are removed and an object is not, problems arise. ACLs do not suffer from this problem.

5.5.4. Covert Channels

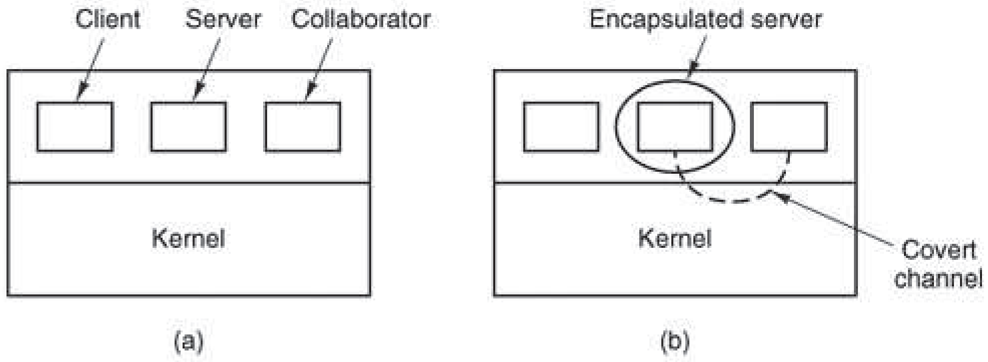
Even with access control lists and capabilities, security leaks can still occur. In this section we discuss how information can still leak out even when it has been rigorously proven that such leakage is mathematically impossible. These ideas are due to Lampson (1973).

Lampson's model was originally formulated in terms of a single timesharing system, but the same ideas can be adapted to LANs and other multiuser environments. In the purest form, it involves three processes on some protected machine. The first process is the client, which wants some work performed by the second one, the server. The client and the server do not entirely trust each other. For example, the server's job is to help clients with filling out their tax forms. The clients are worried that the server will secretly record their financial data, for example, maintaining a secret list of who earns how much, and then selling the list. The server is worried that the clients will try to steal the valuable tax program.

The third process is the collaborator, which is conspiring with the server to indeed steal the client's confidential data. The collaborator and server are typically owned by the same person. These three processes are shown in Fig. 5-31. The object of this exercise is to design a system in which it is impossible for the server process to leak to the collaborator process the information that it has legitimately received from the client process. Lampson called this the **confinement problem**.

Figure 5-31. (a) The client, server, and collaborator processes. (b) The encapsulated server can still leak to the collaborator via covert channels.

[View full size image]



From the system designer's point of view, the goal is to encapsulate or confine the server in such a way that it cannot pass information to the collaborator. Using a protection matrix scheme we can easily guarantee that the server cannot communicate with the collaborator by writing a file to which the collaborator has read access. We can probably also ensure that the server cannot communicate with the collaborator using the system's normal interprocess communication mechanism.

Unfortunately, more subtle communication channels may be available. For example, the server can try to communicate a binary bit stream as follows: To send a 1 bit, it computes as hard as it can for a fixed interval of time. To send a 0 bit, it goes to sleep for the same length of time.

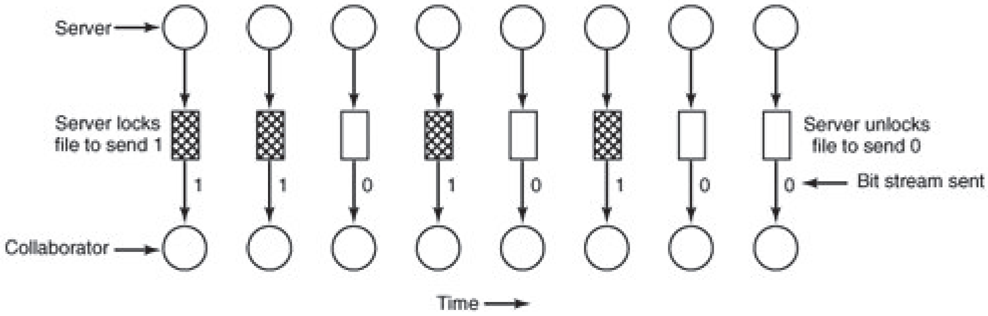
The collaborator can try to detect the bit stream by carefully monitoring its response time. In general, it will get better response when the server is sending a 0 than when the server is sending a 1. This communication channel is known as a **covert channel**, and is illustrated in Fig. 5-31(b).

Of course, the covert channel is a noisy channel, containing a lot of extraneous information, but information can be reliably sent over a noisy channel by using an error-correcting code (e.g., a Hamming code, or even something more sophisticated). The use of an error-correcting code reduces the already low bandwidth of the covert channel even more, but it still may be enough to leak substantial information. It is fairly obvious that no protection model based on a matrix of objects and domains is going to prevent this kind of leakage.

Modulating the CPU usage is not the only covert channel. The paging rate can also be modulated (many page faults for a 1, no page faults for a 0). In fact, almost any way of degrading system performance in a clocked way is a candidate. If the system provides a way of locking files, then the server can lock some file to indicate a 1, and unlock it to indicate a 0. On some systems, it may be possible for a process to detect the status of a lock even on a file that it cannot access. This covert channel is illustrated in Fig. 5-32, with the file locked or unlocked for some fixed time interval known to both the server and collaborator. In this example, the secret bit stream 11010100 is being transmitted.

Figure 5-32. A covert channel using file locking.

[View full size image]



Locking and unlocking a prearranged file, *S* is not an especially noisy channel, but it does require fairly accurate timing unless the bit rate is very low. The reliability and performance can be increased even more using an acknowledgement protocol. This protocol uses two more files, *F1* and *F2*, locked by the server and collaborator, respectively to keep the two processes synchronized. After the server locks or unlocks *S*, it flips the lock status of *F1* to indicate that a bit has been sent. As soon as the collaborator has read out the bit, it flips *F2*'s lock status to tell the server it is ready for another bit and waits until *F1* is flipped again to indicate that another bit is present in *S*. Since timing is no longer involved, this protocol is fully reliable, even in a busy system and can proceed as fast as the two processes can get scheduled. To get higher bandwidth, why not use two files per bit time, or make it a byte-wide channel with eight signaling files, *S0* through *S7*.

Acquiring and releasing dedicated resources (tape drives, plotters, etc.) can also be used for signaling. The server acquires the resource to send a 1 and releases it to send a 0. In UNIX, the server could create a file to indicate a 1 and remove it to indicate a 0; the collaborator could use the `access` system call to see if the file exists. This call works even though the collaborator has no permission to use the file. Unfortunately, many other covert channels exist.

Lampson also mentioned a way of leaking information to the (human) owner of the server process. Presumably the server process will be entitled to tell its owner how much work it did on behalf of the client, so the client can be billed. If the actual computing bill is, say, $100 and the client's income is $53,000 dollars, the server could report the bill as $100.53 to its owner.

Just finding all the covert channels, let alone blocking them, is extremely difficult. In practice, there is little that can be done. Introducing a process that causes page faults at random, or

PREV NEXT

[Page 548 (continued)]

5.6. Overview of the MINIX 3 File System

Like any file system, the MINIX 3 file system must deal with all the issues we have just studied. It must allocate and deallocate space for files, keep track of disk blocks and free space, provide some way to protect files against unauthorized usage, and so on. In the remainder of this chapter we will look closely at MINIX 3 to see how it accomplishes these goals.

In the first part of this chapter, we have repeatedly referred to UNIX rather than to MINIX 3 for the sake of generality, although the external interfaces of the two is virtually identical. Now we will concentrate on the internal design of MINIX 3. For information about the UNIX internals, see Thompson (1978), Bach (1987), Lions (1996), and Vahalia (1996).

The MINIX 3 file system is just a big C program that runs in user space (see Fig. 2-29). To read and write files, user processes send messages to the file system telling what they want done. The file system does the work and then sends back a reply. The file system is, in fact, a network file server that happens to be running on the same machine as the caller.

This design has some important implications. For one thing, the file system can be modified, experimented with, and tested almost completely independently of the rest of MINIX 3. For another, it is very easy to move the file system to any computer that has a C compiler, compile it there, and use it as a free-standing UNIX-like remote file server. The only changes that need to be made are in the area of how messages are sent and received, which differs from system to system.

In the following sections, we will present an overview of many of the key areas of the file system design. Specifically, we will look at messages, the file system layout, the bitmaps, i-nodes, the block cache, directories and paths, file descriptors, file locking, and special files (plus pipes). After studying these topics, we will show a simple example of how the pieces fit together by tracing what happens when a user process executes the `read` system call.

[Page 550]

5.6.1. Messages

The file system accepts 39 types of messages requesting work. All but two are for MINIX 3 system calls. The two exceptions are messages generated by other parts of MINIX 3. Of the system calls, 31 are accepted from user processes. Six system call messages are for system calls which are handled first by the process manager, which then calls the file system to do a part of the work. Two other messages are also handled by the file system. The messages are shown in Fig. 5-33.

Figure 5-33. File system messages. File name parameters are always pointers to the name. The code status as reply value means OK or ERROR.

(This item is displayed on page 549 in the print version)

Messages from users	**Input parameters**	**Reply value**
access	File name, access mode	Status
chdir	Name of new working directory	Status
chmod	File name, new mode	Status
chown	File name, new owner, group	Status
chroot	Name of new root directory	Status
close	File descriptor of file to close	Status
creat	Name of file to be created, mode	File descriptor
dup	File descriptor (for dup2, two fds)	New file descriptor
fcntl	File descriptor, function code, arg	Depends on function
fstat	Name of file, buffer	Status
ioctl	File descriptor, function code, arg	Status
link	Name of file to link to, name of link	Status
lseek	File descriptor, offset, whence	New position
mkdir	File name, mode	Status
mknod	Name of dir or special, mode, address	Status
mount	Special file, where to mount, ro flag	Status
open	Name of file to open, r/w flag	File descriptor
pipe	Pointer to 2 file descriptors (modified)	Status
read	File descriptor, buffer, how many bytes	# Bytes read
rename	File name, file name	Status
rmdir	File name	Status
stat	File name, status buffer	Status
stime	Pointer to current time	Status
sync	(None)	Always OK
time	Pointer to place where current time goes	Status
times	Pointer to buffer for process and child times	Status
umask	Complement of mode mask	Always OK
umount	Name of special file to unmount	Status
unlink	Name of file to unlink	Status
utime	File name, file times	Always OK
write	File descriptor, buffer, how many bytes	# Bytes written

Messages from users	Input parameters	Reply value
Messages from PM	**Input parameters**	**Reply value**
exec	Pid	Status
exit	Pid	Status
fork	Parent pid, child pid	Status
setgid	Pid, real and effective gid	Status
setsid	Pid	Status
setuid	Pid, real and effective uid	Status
Other messages	**Input parameters**	**Reply value**
revive	Process to revive	(No reply)
unpause	Process to check	(See text)

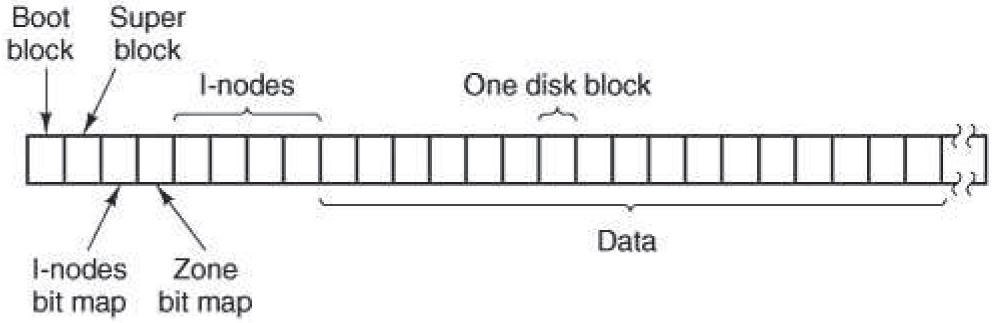
The structure of the file system is basically the same as that of the process manager and all the I/O device drivers. It has a main loop that waits for a message to arrive. When a message arrives, its type is extracted and used as an index into a table containing pointers to the procedures within the file system that handle all the types. Then the appropriate procedure is called, it does its work and returns a status value. The file system then sends a reply back to the caller and goes back to the top of the loop to wait for the next message.

5.6.2. File System Layout

A MINIX 3 file system is a logical, self-contained entity with i-nodes, directories, and data blocks. It can be stored on any block device, such as a floppy disk or a hard disk partition. In all cases, the layout of the file system has the same structure. Figure 5-34 shows this layout for a floppy disk or a small hard disk partition with 64 i-nodes and a 1-KB block size. In this simple example, the zone bitmap is just one 1-KB block, so it can keep track of no more than 8192 1-KB zones (blocks), thus limiting the file system to 8 MB. Even for a floppy disk, only 64 i-nodes puts a severe limit on the number of files, so rather than the four blocks reserved for i-nodes in the figure, more would probably be used. Reserving eight blocks for i-nodes would be more practical but our diagram would not look as nice. For a modern hard disk, both the i-node and zone bitmaps will be much larger than 1 block, of course. The relative size of the various components in Fig. 5-34 may vary from file system to file system, depending on their sizes, how many files are allowed maximum, and so on. But all the components are always present and in the same order.

Figure 5-34. Disk layout for a floppy disk or small hard disk partition, with 64 i-nodes and a 1-KB block size (i.e., two consecutive 512-byte sectors are treated as a single block).

(This item is displayed on page 551 in the print version)

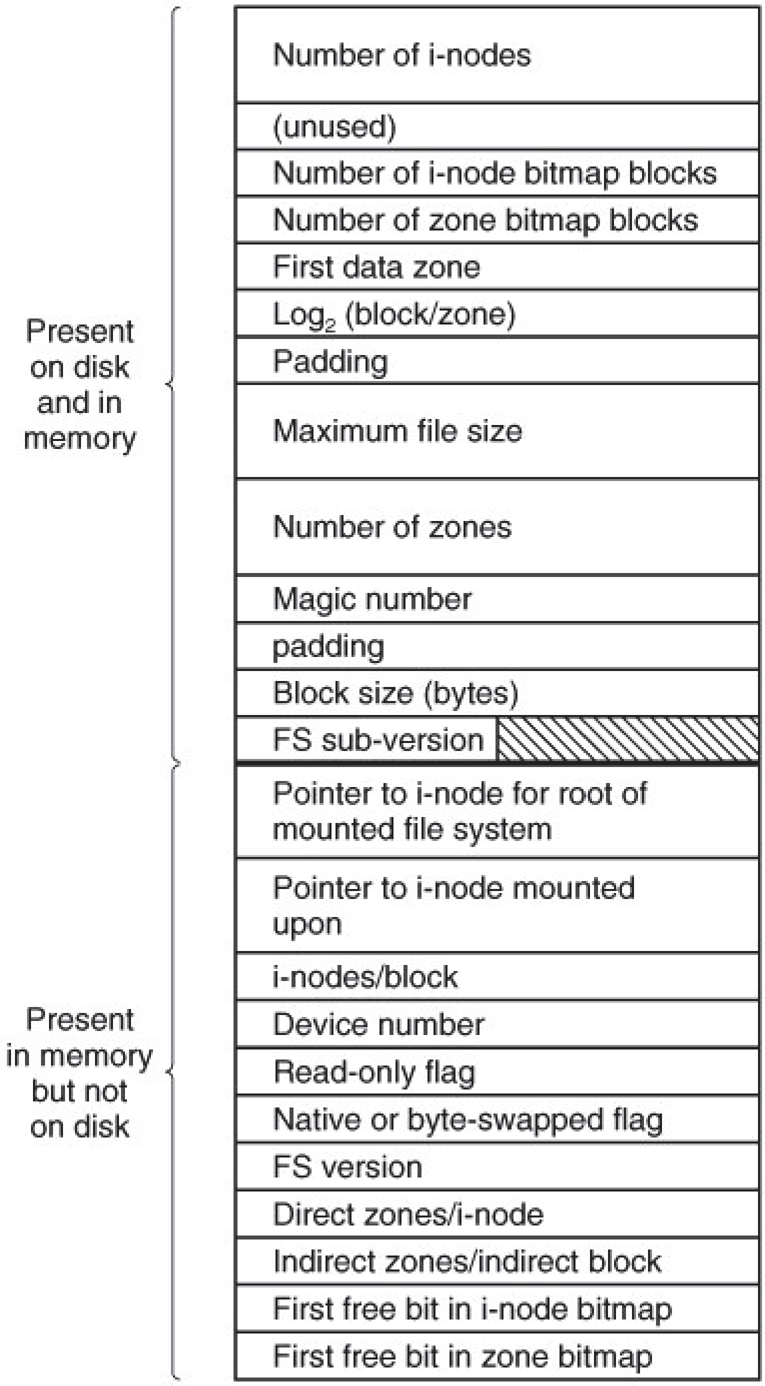


Each file system begins with a **boot block**. This contains executable code. The size of a boot block is always 1024 bytes (two disk sectors), even though MINIX 3 may (and by default does) use a larger block size elsewhere. When the computer is turned on, the hardware reads the boot block from the boot device into memory, jumps to it, and begins executing its code. The boot block code begins the process of loading the operating system itself. Once the system has been booted, the boot block is not used any more. Not every disk drive can be used as a boot device, but to keep the structure uniform, every block device has a block reserved for boot block code. At worst this strategy wastes one block.To prevent the hardware from trying to boot an unbootable device, a **magic number** is placed at a known location in the boot block when and only when the executable code is written to the device. When booting from a device, the hardware (actually, the BIOS code) will refuse to attempt to load from a device lacking the magic number. Doing this prevents inadvertently using garbage as a boot program.

The **superblock** contains information describing the layout of the file system. Like the boot block, the superblock is always 1024 bytes, regardless of the block size used for the rest of the file system. It is illustrated in Fig. 5-35.

Figure 5-35. The MINIX 3 superblock.

(This item is displayed on page 552 in the print version)



Present on disk and in memory

Number of i-nodes

(unused)

Number of i-node bitmap blocks

Number of zone bitmap blocks

First data zone

Log_2 (block/zone)

Padding

Maximum file size

Number of zones

Magic number

padding

Block size (bytes)

FS sub-version

Present in memory but not on disk

Pointer to i-node for root of mounted file system

Pointer to i-node mounted upon

i-nodes/block

Device number

Read-only flag

Native or byte-swapped flag

FS version

Direct zones/i-node

Indirect zones/indirect block

First free bit in i-node bitmap

First free bit in zone bitmap

The main function of the superblock is to tell the file system how big the various pieces of the file system are. Given the block size and the number of i-nodes, it is easy to calculate the size of the i-node bitmap and the number of blocks of inodes. For example, for a 1-KB block, each block of the bitmap has 1024 bytes (8192 bits), and thus can keep track of the status of up to 8192 i-nodes. (Actually the first block can handle only up to 8191 i-nodes, since there is no 0th i-node, but it is given a bit in the bitmap, anyway). For 10,000 i-nodes, two bitmap blocks are needed. Since i-nodes each occupy 64 bytes, a 1-KB block holds up to 16 i-nodes. With 64 i-nodes, four disk blocks are needed to contain them all.

We will explain the difference between zones and blocks in detail later, but for the time being it is sufficient to say that disk storage can be allocated in units (zones) of 1, 2, 4, 8, or in general 2^n blocks. The zone bitmap keeps track of free storage in zones, not blocks. For all standard disks used by MINIX 3 the zone and block sizes are the same (4 KB by default), so to a first approximation a zone is the same as a block on these devices. Until we come to the details of storage allocation later in the chapter, it is adequate to think "block" whenever you see "zone."

Note that the number of blocks per zone is not stored in the superblock, as it is never needed. All that is needed is the base 2 logarithm of the zone to block ratio, which is used as the shift count to convert zones to blocks and vice versa. For example, with 8 blocks per zone, $\log_2 8 = 3$, so to find the zone containing block 128 we shift 128 right 3 bits to get zone 16.

The zone bitmap includes only the data zones (i.e., the blocks used for the bitmaps and i-nodes are not in the map), with the first data zone designated zone 1 in the bitmap. As with the i-node bitmap, bit 0 in the map is unused, so the first block in the zone bitmap can map 8191 zones and subsequent blocks can map 8192 zones each. If you examine the bitmaps on a newly formatted disk, you will find that both the i-node and zone bitmaps have 2 bits set to 1. One is for the nonexistent 0th i-node or zone; the other is for the i-node and zone used by the root directory on the device, which is placed there when the file system is created.

The information in the superblock is redundant because sometimes it is needed in one form and sometimes in another. With 1 KB devoted to the superblock, it makes sense to compute this information in all the forms it is needed, rather than having to recompute it frequently during execution. The zone number of the first data zone on the disk, for example, can be calculated from the block size, zone size, number of i-nodes, and number of zones, but it is faster just to keep it in the superblock. The rest of the superblock is wasted anyhow, so using up another word of it costs nothing.

When MINIX 3 is booted, the superblock for the root device is read into a table in memory. Similarly, as other file systems are mounted, their superblocks are also brought into memory. The superblock table holds a number of fields not present on the disk. These include flags that allow a device to be specified as read-only or as following a byte-order convention opposite to the standard, and fields to speed access by indicating points in the bitmaps below which all bits are marked used. In addition, there is a field describing the device from which the superblock came.

Before a disk can be used as a MINIX 3 file system, it must be given the structure of Fig. 5-34. The utility program *mkfs* has been provided to build file systems. This program can be called either by a command like

```
mkfs /dev/fd1 1440
```

to build an empty 1440 block file system on the floppy disk in drive 1, or it can be given a prototype file listing directories and files to include in the new file system. This command also puts a magic number in the superblock to identify the file system as a valid MINIX file system. The MINIX file system has evolved, and some aspects of the file system (for instance, the size of i-nodes) were different previously. The magic number identifies the version of *mkfs* that created the file system, so differences can be accommodated. Attempts to mount a file system not in MINIX 3 format, such as an MS-DOS diskette, will be rejected by the `mount` system call, which checks the superblock for a valid magic number and other things.

5.6.3. Bitmaps

MINIX 3 keeps tracks of which i-nodes and zones are free by using two bitmaps. When a file is removed, it is then a simple matter to calculate which block of the bitmap contains the bit for the i-node being freed and to find it using the normal cache mechanism. Once the block is found, the bit corresponding to the freed i-node is set to 0. Zones are released from the zone bitmap in the same way.

Logically, when a file is to be created, the file system must search through the bit-map blocks one at a time for the first free i-node. This i-node is then allocated for the new file. In fact, the in-memory copy of the superblock has a field which points to the first free i-node, so no search is necessary until after a node is used, when the pointer must be updated to point to the new next free i-node, which will often turn out to be the next one, or a close one. Similarly, when an i-node is freed, a check is made to see if the free i-node comes before the currently-pointed-to one, and the pointer is updated if necessary. If every i-node slot on the disk is full, the search routine returns a 0, which is why i-node 0 is not used (i.e., so it can be used to indicate the search failed). (When *mkfs* creates a new file system, it zeroes i-node 0 and sets the lowest bit in the bitmap to 1, so the file system will never attempt to allocate it.) Everything that has been said here about the i-node bitmaps also applies to the zone bitmap; logically it is searched for the first free zone when space is needed, but a pointer to the first free zone is maintained to eliminate most of the need for sequential searches through the bitmap.

With this background, we can now explain the difference between zones and blocks. The idea behind zones is to help ensure that disk blocks that belong to the same file are located on the same cylinder, to improve performance when the file is read sequentially. The approach chosen is to make it possible to allocate several blocks at a time. If, for example, the block size is 1 KB and the zone size is 4 KB, the zone bitmap keeps track of zones, not blocks. A 20-MB disk has 5K zones of 4 KB, hence 5K bits in its zone map.

Most of the file system works with blocks. Disk transfers are always a block at a time, and the buffer cache also works with individual blocks. Only a few parts of the system that keep track of physical disk addresses (e.g., the zone bitmap and the i-nodes) know about zones.

Some design decisions had to be made in developing the MINIX 3 file system. In 1985, when MINIX was conceived, disk capacities were small, and it was expected that many users would have only floppy disks. A decision was made to restrict disk addresses to 16 bits in the V1 file system, primarily to be able to store many of them in the indirect blocks. With a 16-bit zone number and a 1-KB zone, only 64-KB zones can be addressed, limiting disks to 64 MB. This was an enormous amount of storage in those days, and it was thought that as disks got larger, it would be easy to switch to 2-KB or 4-KB zones, without changing the block size. The 16-bit zone numbers also made it easy to keep the i-node size to 32 bytes.

As MINIX developed, and larger disks became much more common, it was obvious that changes were desirable. Many files are smaller than 1 KB, so increasing the block size would mean wasting disk bandwidth, reading and writing mostly empty blocks and wasting precious main memory storing them in the buffer cache. The zone size could have been increased, but a larger zone size means more wasted disk space, and it was still desirable to retain efficient operation on small disks. Another reasonable alternative would have been to have different zone sizes on large and small devices.

In the end it was decided to increase the size of disk pointers to 32 bits. This made it possible for the MINIX V2 file system to deal with device sizes up to 4 terabytes with 1-KB blocks and zones and 16 TB with 4-KB blocks and zones (the default value now). However, other factors restrict this size (e.g., with 32-bit pointers, raw devices are limited to 4 GB). Increasing the size of disk pointers required an increase in the size of i-nodes. This is not necessarily a bad thingit means the MINIX V2 (and now, V3) i-node is compatible with standard UNIX i-nodes, with room for three time values, more indirect and double indirect zones, and room for later expansion with triple indirect zones.

Zones also introduce an unexpected problem, best illustrated by a simple example, again with 4-KB zones and 1-KB blocks. Suppose that a file is of length 1-KB, meaning that one zone has been allocated for it. The three blocks between offsets 1024 and 4095 contain garbage (residue from the previous owner), but no structural harm is done to the file system because the file size is clearly marked in the i-node as 1 KB In fact, the blocks containing garbage will not be read into the block cache, since reads are done by blocks, not by zones. Reads beyond the end of a file always return a count of 0 and no data.

Now someone seeks to 32,768 and writes 1 byte. The file size is now set to 32,769. Subsequent seeks to byte 1024 followed by attempts to read the data will now be able to read the previous contents of the block, a major security breach.

The solution is to check for this situation when a write is done beyond the end of a file, and explicitly zero all the not-yet-allocated blocks in the zone that was previously the last one. Although this situation rarely occurs, the code has to deal with it, making the system slightly more complex.

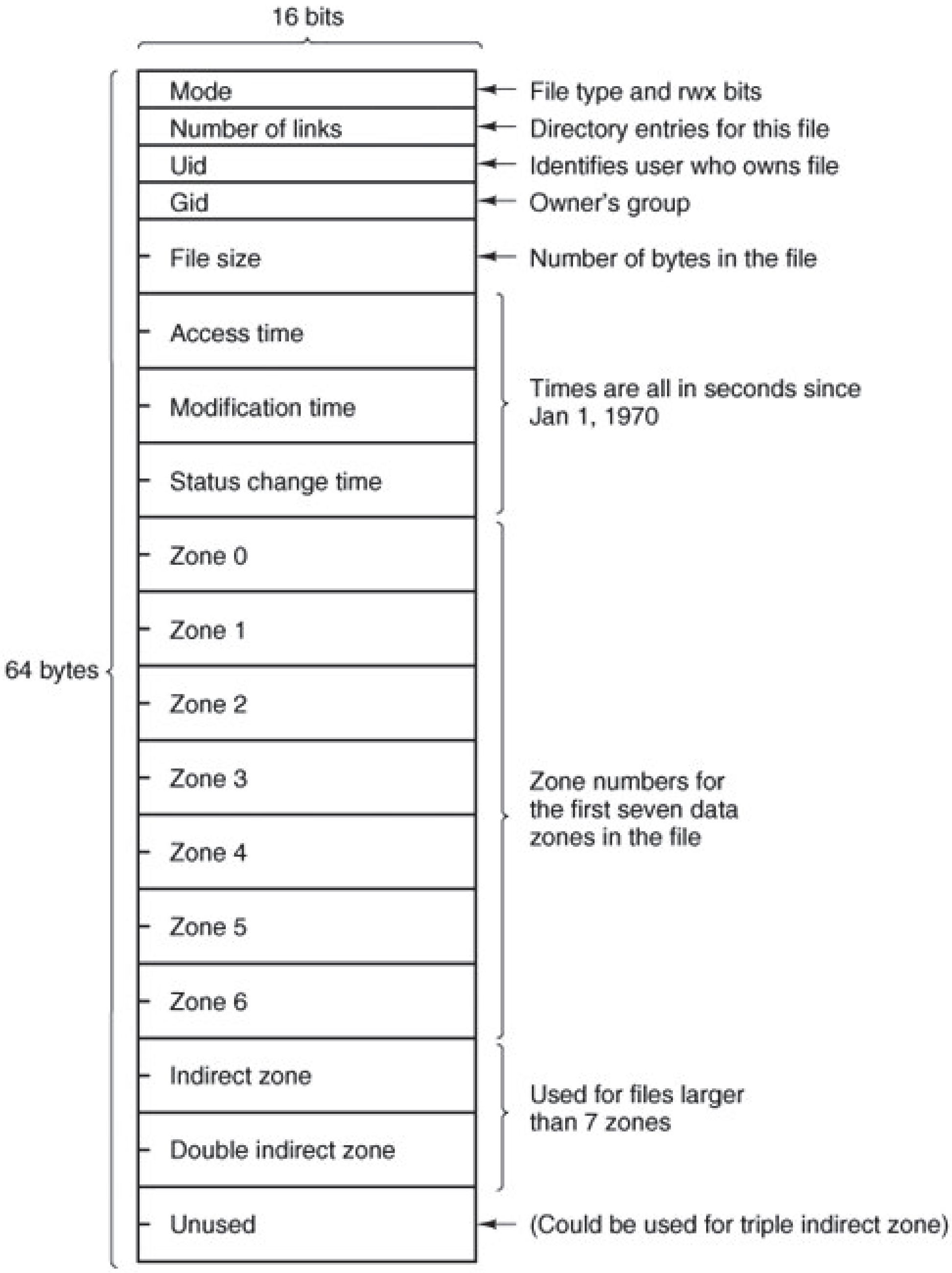
5.6.4. I-Nodes

The layout of the MINIX 3 i-node is given in Fig. 5-36. It is almost the same as a standard UNIX i-node. The disk zone pointers are 32-bit pointers, and there are only 9 pointers, 7 direct and 2 indirect. The MINIX 3 i-nodes occupy 64 bytes, the same as standard UNIX i-nodes, and there is space available for a 10th (triple indirect) pointer, although its use is not supported by the standard version of the FS. The MINIX 3 i-node access, modification time and i-node change times are standard, as in UNIX. The last of these is updated for almost every file operation except a read of the file.

Figure 5-36. The MINIX i-node.

(This item is displayed on page 556 in the print version)

[View full size image]


16 bits
64 bytes
Mode
Number of links
Uid
Gid
File size
Access time
Modification time
Status change time
Zone 0
Zone 1
Zone 2
Zone 3
Zone 4
Zone 5
Zone 6
Indirect zone
Double indirect zone
Unused
File type and rwx bits
Directory entries for this file
Identifies user who owns file
Owner's group
Number of bytes in the file
Times are all in seconds since Jan 1, 1970
Zone numbers for the first seven data zones in the file
Used for files larger than 7 zones
(Could be used for triple indirect zone)

When a file is opened, its i-node is located and brought into the *inode* table in memory, where it remains until the file is closed. The *inode* table has a few additional fields not present on the disk, such as the i-node's device and number, so the file system knows where to rewrite the i-node if it is modified while in memory. It also has a counter per i-node. If the same file is opened more than once, only one copy of the i-node is kept in memory, but the counter is incremented each time the file is opened and decremented each time the file is closed. Only when the counter finally reaches zero is the i-node removed from the table. If it has been modified since being loaded into memory, it is also rewritten to the disk.

The main function of a file's i-node is to tell where the data blocks are. The first seven zone numbers are given right in the i-node itself. For the standard distribution, with zones and blocks both 1 KB, files up to 7 KB do not need indirect blocks. Beyond 7 KB, indirect zones are needed, using the scheme of Fig. 5-10, except that only the single and double indirect blocks are used.

With 1-KB blocks and zones and 32-bit zone numbers, a single indirect block holds 256 entries, representing a quarter megabyte of storage. The double indirect block points to 256 single indirect blocks, giving access to up to 64 megabytes. With 4-KB blocks, the double indirect block leads to 1024 x 1024 blocks, which is over a million 4-KB blocks, making the maximum file zie over 4 GB. In practice the use of 32-bit numbers as file offsets limits the maximum file size to 2^{32} 1 bytes. As a consequence of these numbers, when 4-KB disk blocks are used MINIX 3 has no need for triple indirect blocks; the maximum file size is limited by the pointer size, not the ability to keep track of enough blocks.

[Page 556]

[Page 557]

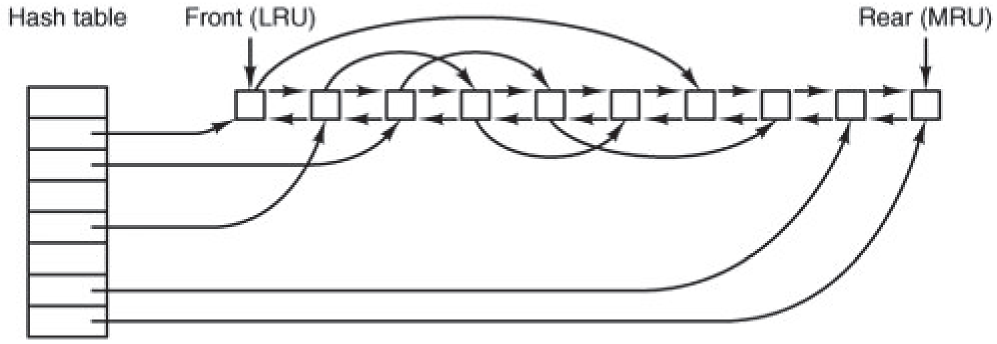
The i-node also holds the mode information, which tells what kind of a file it is (regular, directory, block special, character special, or pipe), and gives the protection and SETUID and SETGID bits. The *link* field in the i-node records how many directory entries point to the i-node, so the file system knows when to release the file's storage. This field should not be confused with the counter (present only in the *inode* table in memory, not on the disk) that tells how many times the file is currently open, typically by different processes.

As a final note on i-nodes, we mention that the structure of Fig. 5-36 may be modified for special purposes. An example used in MINIX 3 is the i-nodes for block and character device special files. These do not need zone pointers, because they don't have to reference data areas on the disk. The major and minor device numbers are stored in the *Zone-0* space in Fig. 5-36. Another way an i-node could be used, although not implemented in MINIX 3, is as an immediate file with a small amount of data stored in the i-node itself.

5.6.5. The Block Cache

MINIX 3 uses a block cache to improve file system performance. The cache is implemented as a fixed array of buffers, each consisting of a header containing pointers, counters, and flags, and a body with room for one disk block. All the buffers that are not in use are chained together in a double-linked list, from most recently used (MRU) to least recently used (LRU) as illustrated in Fig. 5-37.

Figure 5-37. The linked lists used by the block cache.



In addition, to be able to quickly determine if a given block is in the cache or not, a hash table is used. All the buffers containing a block that has hash code *k* are linked together on a single-linked list pointed to by entry *k* in the hash table. The hash function just extracts the low-order *n* bits from the block number, so blocks from different devices appear on the same hash chain. Every buffer is on one of these chains. When the file system is initialized after MINIX 3 is booted, all buffers are unused, of course, and all are in a single chain pointed to by the 0th hash table entry. At that time all the other hash table entries contain a null pointer, but once the system starts, buffers will be removed from the 0th chain and other chains will be built.

When the file system needs to acquire a block, it calls a procedure, *get_block*, which computes the hash code for that block and searches the appropriate list. *Get_block* is called with a device number as well as a block number, and the search compares both numbers with the corresponding fields in the buffer chain. If a buffer containing the block is found, a counter in the buffer header is incremented to show that the block is in use, and a pointer to it is returned. If a block is not found on the hash list, the first buffer on the LRU list can be used; it is guaranteed not to be still in use, and the block it contains may be evicted to free up the buffer.

Once a block has been chosen for eviction from the block cache, another flag in its header is checked to see if the block has been modified since being read in. If so, it is rewritten to the disk. At this point the block needed is read in by sending a message to the disk driver. The file system is suspended until the block arrives, at which time it continues and a pointer to the block is returned to the caller.

When the procedure that requested the block has completed its job, it calls another procedure, *put_block*, to free the block. Normally, a block will be used immediately and then released, but since it is possible that additional requests for a block will be made before it has been released, *put_block* decrements the use counter and puts the buffer back onto the LRU list only when the use counter has gone back to zero. While the counter is nonzero, the block remains in limbo.

One of the parameters to *put_block* tells what class of block (e.g., i-nodes, directory, data) is being freed. Depending on the class, two key decisions are made:

1. Whether to put the block on the front or rear of the LRU list.

2. Whether to write the block (if modified) to disk immediately or not.

Almost all blocks go on the rear of the list in true LRU fashion. The exception is blocks from the RAM disk; since they are already in memory there is little advantage to keeping them in the block cache.

A modified block is not rewritten until either one of two events occurs:

1. It reaches the front of the LRU chain and is evicted.

2. A `sync` system call is executed.

`Sync` does not traverse the LRU chain but instead indexes through the array of buffers in the cache. Even if a buffer has not been released yet, if it has been modified, `sync` will find it and ensure that the copy on disk is updated.

Policies like this invite tinkering. In an older version of MINIX a superblock was modified when a file system was mounted, and was always rewritten immediately to reduce the chance of corrupting the file system in the event of a crash. Superblocks are modified only if the size of a RAM disk must be adjusted at startup time because the RAM disk was created bigger than the RAM image device. However, the superblock is not read or written as a normal block, because it is always 1024 bytes in size, like the boot block, regardless of the block size used for blocks handled by the cache. Another abandoned experiment is that in older versions of MINIX there was a *ROBUST* macro definable in the system configuration file, *include/minix/config.h,* which, if defined, caused the file system to mark i-node, directory, indirect, and bit-map blocks to be written immediately upon release. This was intended to make the file system more robust; the price paid was slower operation. It turned out this was not effective. A power failure occurring when all blocks have not been yet been written is going to cause a headache whether it is an i-node or a data block that is lost.

Note that the header flag indicating that a block has been modified is set by the procedure within the file system that requested and used the block. The procedures *get_block* and *put_block* are concerned just with manipulating the linked lists. They have no idea which file system procedure wants which block or why.

5.6.6. Directories and Paths

Another important subsystem within the file system manages directories and path names. Many system calls, such as `open`, have a file name as a parameter. What is really needed is the i-node for that file, so it is up to the file system to look up the file in the directory tree and locate its i-node.

A MINIX directory is a file that in previous versions contained 16-byte entries, 2 bytes for an i-node number and 14 bytes for the file name. This design limited disk partitions to 64-KB files and file names to 14 characters, the same as V7 UNIX. As disks have grown file names have also grown. In MINIX 3 the V3 file system provides 64 bytes directory entries, with 4 bytes for the i-node number and 60 bytes for the file name. Having up to 4 billion files per disk partition is effectively infinite and any programmer choosing a file name longer than 60 characters should be sent back to programming school.

Note that *paths* such as

/usr/ast/course_material_for_this_year/operating_systems/examination-1.ps

are not limited to 60 charactersjust the individual component names. The use of fixed-length directory entries, in this case, 64 bytes, is an example of a tradeoff involving simplicity, speed, and storage. Other operating systems typically organize directories as a heap, with a fixed header for each file pointing to a name on the heap at the end of the directory. The MINIX 3 scheme is very simple and required practically no code changes from V2. It is also very fast for both looking up names and storing new ones, since no heap management is ever required. The price paid is wasted disk storage, because most files are much shorter than 60 characters.

It is our firm belief that optimizing to save disk storage (and some RAM storage since directories are occasionally in memory) is the wrong choice. Code simplicity and correctness should come first and speed should come second. With modern disks usually exceeding 100 GB, saving a small amount of disk space at the price of more complicated and slower code is generally not a good idea. Unfortunately, many programmers grew up in an era of tiny disks and even tinier RAMs, and were trained from day 1 to resolve all trade-offs between code complexity, speed, and space in favor of minimizing space requirements. This implicit assumption really has to be reexamined in light of current realities.

Now let us see how the path */usr/ast/mbox/* is looked up. The system first looks up *usr* in the root directory, then it looks up *ast* in */usr/,* and finally it looks up *mbox* in */usr/ast/*. The actual lookup proceeds one path component at a time, as illustrated in Fig. 5-16.

The only complication is what happens when a mounted file system is encountered. The usual configuration for MINIX 3 and many other UNIX-like systems is to have a small root file system containing the files needed to start the system and to do basic system maintenance, and to have the majority of the files, including users' directories, on a separate device mounted on /usr. This is a good time to look at how mounting is done. When the user types the command

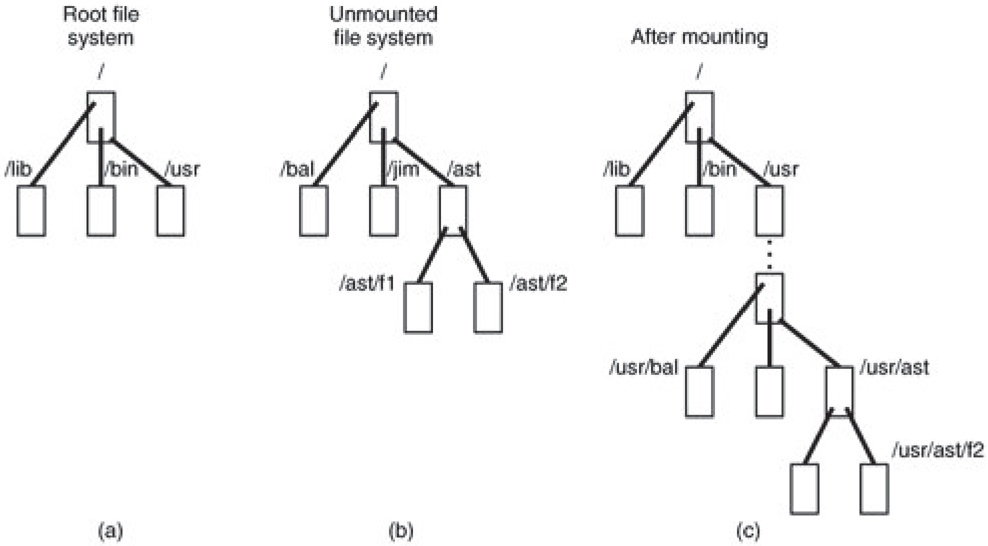
```
mount /dev/c0d1p2 /usr
```

on the terminal, the file system contained on hard disk 1, partition 2 is mounted on top of */usr/* in the root file system. The file systems before and after mounting are shown in Fig. 5-38.

Figure 5-38. (a) Root file system. (b) An unmounted file system. (c) The result of mounting the file system of (b) on */usr/*.

(This item is displayed on page 561 in the print version)

[View full size image]



The key to the whole mount business is a flag set in the memory copy of the i-node of */usr* after a successful mount. This flag indicates that the i-node is mounted on. The `mount` call also loads the super-block for the newly mounted file system into the *super_block* table and sets two pointers in it. Furthermore, it puts the root i-node of the mounted file system in the *inode* table.

In Fig. 5-35 we see that super-blocks in memory contain two fields related to mounted file systems. The first of these, the *i-node-for-root-of-mounted-file-system*, is set to point to the root i-node of the newly mounted file system. The second, the *i-node-mounted-upon*, is set to point to the i-node mounted on, in this case, the i-node for */usr.* These two pointers serve to connect the mounted file system to the root and represent the "glue" that holds the mounted file system to the root [shown as the dots in Fig. 5-38(c)]. This glue is what makes mounted file systems work.

When a path such as */usr/ast/f2* is being looked up, the file system will see a flag in the i-node for */usr/* and realize that it must continue searching at the root inode of the file system mounted on */usr/.* The question is: "How does it find this root i-node?"

The answer is straightforward. The system searches all the superblocks in memory until it finds the one whose *i-node mounted on* field points to */usr/.* This must be the superblock for the file system mounted on */usr/.* Once it has the superblock, it is easy to follow the other pointer to find the root i-node for the mounted file system. Now the file system can continue searching. In this example, it looks for *ast* in the root directory of hard disk partition 2.

5.6.7. File Descriptors

Once a file has been opened, a file descriptor is returned to the user process for use in subsequent `read` and `write` calls. In this section we will look at how file descriptors are managed within the file system.

Like the kernel and the process manager, the file system maintains part of the process table within its address space. Three of its fields are of particular interest. The first two are pointers to the i-nodes for the root directory and the working directory. Path searches, such as that of Fig. 5-16, always begin at one or the other, depending on whether the path is absolute or relative. These pointers are changed by the `chroot` and `chdir` system calls to point to the new root or new working directory, respectively.

The third interesting field in the process table is an array indexed by file descripttor number. It is used to locate the proper file when a file descriptor is presented. At first glance, it might seem sufficient to have the *k*-th entry in this array just point to the i-node for the file belonging to file descriptor *k*. After all, the i-node is fetched into memory when the file is opened and kept there until it is closed, so it is sure to be available.

Unfortunately, this simple plan fails because files can be shared in subtle ways in MINIX 3 (as well as in UNIX). The trouble arises because associated with each file is a 32-bit number that indicates the next byte to be read or written. It is this number, called the **file position**, that is changed by the `lseek` system call. The problem can be stated easily: "Where should the file pointer be stored?"

The first possibility is to put it in the i-node. Unfortunately, if two or more processes have the same file open at the same time, they must all have their own file pointers, since it would hardly do to have an `lseek` by one process affect the next read of a different process. Conclusion: the file position cannot go in the inode.

What about putting it in the process table? Why not have a second array, paralleling the file descriptor array, giving the current position of each file? This idea does not work either, but the reasoning is more subtle. Basically, the trouble comes from the semantics of the `fork` system call. When a process forks, both the parent and the child are required to share a single pointer giving the current position of each open file.

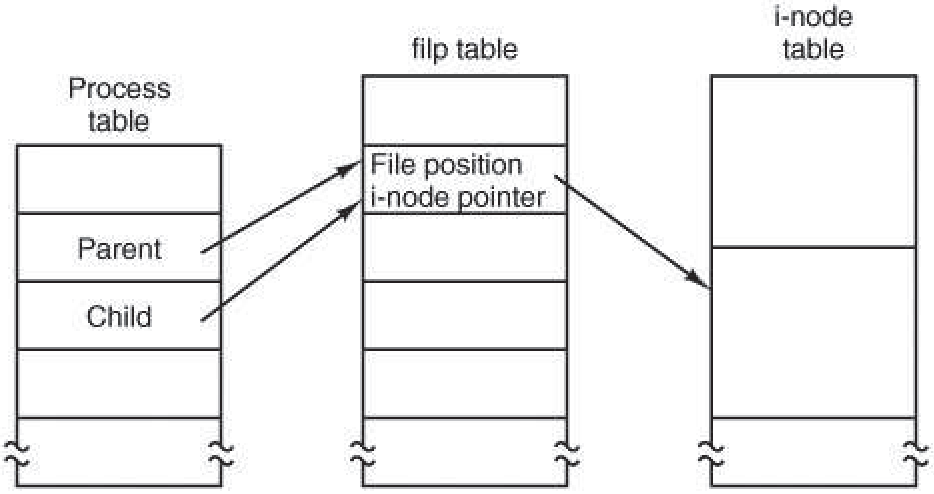
To better understand the problem, consider the case of a shell script whose output has been redirected to a file. When the shell forks off the first program, its file position for standard output is 0. This position is then inherited by the child, which writes, say, 1 KB of output. When the child terminates, the shared file position must now be 1024.

Now the shell reads some more of the shell script and forks off another child. It is essential that the second child inherit a file position of 1024 from the shell, so it will begin writing at the place where the first program left off. If the shell did not share the file position with its children, the second program would overwrite the output from the first one, instead of appending to it.

As a result, it is not possible to put the file position in the process table. It really must be shared. The solution used in UNIX and MINIX 3 is to introduce a new, shared table, *filp*, which contains all the file positions. Its use is illustrated in Fig. 5-39. By having the file position truly shared, the semantics of `fork` can be implemented correctly, and shell scripts work properly.

Figure 5-39. How file positions are shared between a parent and a child.

(This item is displayed on page 563 in the print version)



Although the only thing that the *filp* table really must contain is the shared file position, it is convenient to put the i-node pointer there, too. In this way, all that the file descriptor array in the process table contains is a pointer to a *filp* entry. The *filp* entry also contains the file mode (permission bits), some flags indicating whether the file was opened in a special mode, and a count of the number of processes using it, so the file system can tell when the last process using the entry has terminated, in order to reclaim the slot.

5.6.8. File Locking

Yet another aspect of file system management requires a special table. This is file locking. MINIX 3 supports the POSIX interprocess communication mechanism of **advisory file locking**. This permits any part, or multiple parts, of a file to be marked as locked. The operating system does not enforce locking, but processes are expected to be well behaved and to look for locks on a file before doing anything that would conflict with another process.

The reasons for providing a separate table for locks are similar to the justifications for the *filp* table discussed in the previous section. A single process can have more than one lock active, and different parts of a file may be locked by more than one process (although, of course, the locks cannot overlap), so neither the process table nor the *filp* table is a good place to record locks. Since a file may have more than one lock placed upon it, the i-node is not a good place either.

MINIX 3 uses another table, the *file_lock* table, to record all locks. Each slot in this table has space for a lock type, indicating if the file is locked for reading or writing, the process ID holding the lock, a pointer to the i-node of the locked file, and the offsets of the first and last bytes of the locked region.

5.6.9. Pipes and Special Files

Pipes and special files differ from ordinary files in an important way. When a process tries to read or write a block of data from a disk file, it is almost certain that the operation will complete within a few hundred milliseconds at most. In the worst case, two or three disk accesses might be needed, not more. When reading from a pipe, the situation is different: if the pipe is empty, the reader will have to wait until some other process puts data in the pipe, which might take hours. Similarly, when reading from a terminal, a process will have to wait until somebody types something.

As a consequence, the file system's normal rule of handling a request until it is finished does not work. It is necessary to suspend these requests and restart them later. When a process tries to read or write from a pipe, the file system can check the state of the pipe immediately to see if the operation can be completed. If it can be, it is, but if it cannot be, the file system records the parameters of the system call in the process table, so it can restart the process when the time comes.

Note that the file system need not take any action to have the caller suspended. All it has to do is refrain from sending a reply, leaving the caller blocked waiting for the reply. Thus, after suspending a process, the file system goes back to its main loop to wait for the next system call. As soon as another process modifies the pipe's state so that the suspended process can complete, the file system sets a flag so that next time through the main loop it extracts the suspended process' parameters from the process table and executes the call.

The situation with terminals and other character special files is slightly different. The i-node for each special file contains two numbers, the major device and the minor device. The major device number indicates the device class (e.g., RAM disk, floppy disk, hard disk, terminal). It is used as an index into a file system table that maps it onto the number of the corresponding I/O device driver. In effect, the major device determines which I/O driver to call. The minor device number is passed to the driver as a parameter. It specifies which device is to be used, for example, terminal 2 or drive 1.

In some cases, most notably terminal devices, the minor device number encodes some information about a category of devices handled by a driver. For instance, the primary MINIX 3 console, */dev/console,* is device 4, 0 (major, minor). Virtual consoles are handled by the same part of the driver software. These are devices */dev/ttyc1* (4,1), */dev/ttyc2* (4,2), and so on. Serial line terminals need different low-level software, and these devices, */dev/tty00,* and */dev/tty01* are assigned device numbers 4, 16 and 4, 17. Similarly, network terminals use pseudo-terminal drivers, and these also need different low-level software. In MINIX 3 these devices, *ttyp0*, *ttyp1*, etc., are assigned device numbers such as 4, 128 and 4, 129. These pseudo devices each have an associated device, *ptyp0*, *ptyp1*, etc. The major, minor device number pairs for these are 4,192 and 4,193, etc. These numbers are chosen to make it easy for the device driver to call the low-level functions required for each group of devices. It is not expected that anyone is going to equip a MINIX 3 system with 192 or more terminals.

When a process reads from a special file, the file system extracts the major and minor device numbers from the file's i-node, and uses the major device number as an index into a file system table to map it onto the process number of the corresponding device driver. Once it has identified the driver, the file system sends it a message, including as parameters the minor device, the operation to be performed, the caller's process number and buffer address, and the number of bytes to be transferred. The format is the same as in Fig. 3-15, except that *POSITION* is not used.

If the driver is able to carry out the work immediately (e.g., a line of input has already been typed on the terminal), it copies the data from its own internal buffers to the user and sends the file system a reply message saying that the work is done. The file system then sends a reply message to the user, and the call is finished. Note that the driver does not copy the data to the file system. Data from block devices go through the block cache, but data from character special files do not.

On the other hand, if the driver is not able to carry out the work, it records the message parameters in its internal tables, and immediately sends a reply to the file system saying that the call could not be completed. At this point, the file system is in the same situation as having discovered that someone is trying to read from an empty pipe. It records the fact that the process is suspended and waits for the next message.

When the driver has acquired enough data to complete the call, it transfers them to the buffer of the still-blocked user and then sends the file system a message reporting what it has done. All the file system has to do is send a reply message to the user to unblock it and report the number of bytes transferred.

5.6.10. An Example: The READ System Call

As we shall see shortly, most of the code of the file system is devoted to carrying out system calls. Therefore, it is appropriate that we conclude this overview with a brief sketch of how the most important call, `read`, works.

When a user program executes the statement

```
n = read(fd, buffer, nbytes);
```

to read an ordinary file, the library procedure *read* is called with three parameters. It builds a message containing these parameters, along with the code for `read` as the message type, sends the message to the file system, and blocks waiting for the reply. When the message arrives, the file system uses the message type as an index into its tables to call the procedure that handles reading.

This procedure extracts the file descriptor from the message and uses it to locate the *filp* entry and then the i-node for the file to be read (see Fig. 5-39). The request is then broken up into pieces such that each piece fits within a block. For example, if the current file position is 600 and 1024 bytes have been requested, the request is split into two parts, for 600 to 1023, and for 1024 to 1623 (assuming 1-KB blocks).

For each of these pieces in turn, a check is made to see if the relevant block is in the cache. If the block is not present, the file system picks the least recently used buffer not currently in use and claims it, sending a message to the disk device driver to rewrite it if it is dirty. Then the disk driver is asked to fetch the block to be read.

Once the block is in the cache, the file system sends a message to the system task asking it to copy the data to the appropriate place in the user's buffer (i.e., bytes 600 to 1023 to the start of the buffer, and bytes 1024 to 1623 to offset 424 within the buffer). After the copy has been done, the file system sends a reply message to the user specifying how many bytes have been copied.

When the reply comes back to the user, the library function *read* extracts the reply code and returns it as the function value to the caller.

One extra step is not really part of the `read` call itself. After the file system completes a read and sends a reply, it initiates reading additional blocks, provided that the read is from a block device and certain other conditions are met. Since sequential file reads are common, it is reasonable to expect that the next blocks in a file will be requested in the next read request, and this makes it likely that the desired block will already be in the cache when it is needed. The number of blocks requested depends upon the size of the block cache; as many as 32 additional blocks may be requested. The device driver does not necessarily return this many blocks, and if at least one

PREV NEXT

5.7. Implementation of the MINIX 3 File System

The MINIX 3 file system is relatively large (more than 100 pages of C) but quite straightforward. Requests to carry out system calls come in, are carried out, and replies are sent. In the following sections we will go through it a file at a time, pointing out the highlights. The code itself contains many comments to aid the reader.

In looking at the code for other parts of MINIX 3 we have generally looked at the main loop of a process first and then looked at the routines that handle the different message types. We will organize our approach to the file system differently. First we will go through the major subsystems (cache management, i-node management, etc.). Then we will look at the main loop and the system calls that operate upon files. Next we will look at systems call that operate upon directories, and then, we will discuss the remaining system calls that fall into neither category. Finally we will see how device special files are handled.

5.7.1. Header Files and Global Data Structures

Like the kernel and process manager, various data structures and tables used in the file system are defined in header files. Some of these data structures are placed in system-wide header files in *include/* and its subdirectories. For instance, *include/sys/stat.h* defines the format by which system calls can provide i-node information to other programs and the structure of a directory entry is defined in *include/sys/dir.h.* Both of these files are required by POSIX. The file system is affected by a number of definitions contained in the global configuration file *include/minix/config.h,* such as *NR_BUFS* and *NR_BUF_HASH*, which control the size of the block cache.

File System Headers

The file system's own header files are in the file system source directory *src/fs/.* Many file names will be familiar from studying other parts of the MINIX 3 system. The FS master header file, *fs.h* (line 20900), is quite analogous to *src/kernel/kernel.h* and *src/pm/pm.h.* It includes other header files needed by all the C source files in the file system. As in the other parts of MINIX 3, the file system master header includes the file system's own *const.h, type.h, proto.h,* and *glo.h.* We will look at these next.

Const.h (line 21000) defines some constants, such as table sizes and flags, that are used throughout the file system. MINIX 3 already has a history. Earlier versions of MINIX had different file systems. Although MINIX 3 does not support the old *V1* and *V2* file systems, some definitions have been retained, both for reference and in expectation that someone will add support for these later. Support for older versions is useful not only for accessing files on older MINIX file systems, it may also be useful for exchanging files.

Other operating systems may use older MINIX file systemsfor instance, Linux originally used and still supports MINIX file systems. (It is perhaps somewhat ironic that Linux still supports the original MINIX file system but MINIX 3 does not.) Some utilities are available for MS-DOS and

Windows to access older MINIX directories and files. The superblock of a file system contains a **magic number** to allow the operating system to identify the file system's type; the constants *SUPER_MAGIC*, *SUPER_V2*, and *SUPER_V3* define these numbers for the three versions of the MINIX file system. There are also *_REV*-suffixed versions of these for V1 and V2, in which the bytes of the magic number are reversed. These were used with ports of older MINIX versions to systems with a different byte order (little-endian rather than big-endian) so a removable disk written on a machine with a different byte order could be identified as such. As of the release of MINIX 3.1.0 defining a *SUPER_V3_REV* magic number has not been necessary, but it is likely this definition will be added in the future.

Type.h (line 21100) defines both the old V1 and new V2 i-node structures as they are laid out on the disk. The i-node is one structure that did not change in MINIX 3, so the V2 i-node is used with the V-3 file system. The V2 i-node is twice as big as the old one, which was designed for compactness on systems with no hard drive and 360-KB diskettes. The new version provides space for the three time fields which UNIX systems provide. In the V1 i-node there was only one time field, but a `stat` or `fstat` would "fake it" and return a *stat* structure containing all three fields. There is a minor difficulty in providing support for the two file system versions. This is flagged by the comment on line 21116. Older MINIX 3 software expected the *gid_t* type to be an 8-bit quantity, so *d2_gid* must be declared as type *u16_t*.

Proto.h (line 21200) provides function prototypes in forms acceptable to either old K&R or newer ANSI Standard C compilers. It is a long file, but not of great interest. However, there is one point to note: because there are so many different system calls handled by the file system, and because of the way the file system is organized, the various *do_XXX* functions are scattered through a number of files. *Proto.h* is organized by file and is a handy way to find the file to consult when you want to see the code that handles a particular system call.

Finally, *glo.h* (line 21400) defines global variables. The message buffers for the incoming and reply messages are also here. The now-familiar trick with the *EXTERN* macro is used, so these variables can be accessed by all parts of the file system. As in the other parts of MINIX 3, the storage space will be reserved when *table.c* is compiled.

The file system's part of the process table is contained in *fproc.h* (line 21500). The *fproc* array is declared with the *EXTERN* macro. It holds the mode mask, pointers to the i-nodes for the current root directory and working directory, the file descriptor array, uid, gid, and terminal number for each process. The process id and the process group id are also found here. The process id is duplicated in the part of the process table located in the process manager.

Several fields are used to store the parameters of those system calls that may be suspended part way through, such as reads from an empty pipe. The fields *fp_suspended* and *fp_revived* actually require only single bits, but nearly all compilers generate better code for characters than bit fields. There is also a field for the *FD_CLOEXEC* bits called for by the POSIX standard. These are used to indicate that a file should be closed when an `exec` call is made.

Now we come to files that define other tables maintained by the file system. The first, *buf.h* (line 21600), defines the block cache. The structures here are all declared with *EXTERN*. The array *buf* holds all the buffers, each of which contains a data part, *b*, and a header full of pointers, flags, and counters. The data part is declared as a union of five types (lines 21618 to 21632) because sometimes it is convenient to refer to the block as a character array, sometimes as a directory, etc.

The truly proper way to refer to the data part of buffer 3 as a character array is *buf*[3]. *b.b_ _data* because *buf*[3].*b* refers to the union as a whole, from which the *b_ _data* field is selected. Although this syntax is correct, it is cumbersome, so on line 21649 we define a macro *b_data*, which allows us to write *buf*[3].*b_data* instead. Note that *b_ _data* (the field of the union) contains two underscores, whereas *b_data* (the macro) contains just one, to distinguish them. Macros for other ways of accessing the block are defined on lines 21650 to 21655.

The buffer hash table, *buf_hash*, is defined on line 21657. Each entry points to a list of buffers. Originally all the lists are empty. Macros at the end of *buf.h* define different block types. The *WRITE_IMMED* bit signals that a block must be rewritten to the disk immediately if it is changed, and the *ONE_SHOT* bit is used to indicate a block is unlikely to be needed soon. Neither of these is used currently but they remain available for anyone who has a bright idea about improving performance or reliability by modifying the way blocks in the cache are queued.

Finally, in the last line *HASH_MASK* is defined, based upon the value of *NR_BUF_HASH* configured in *include/minix/config.h. HASH_MASK* is ANDed with a block number to determine which entry in *buf_hash* to use as the starting point in a search for a block buffer.

File.h (line 21700) contains the intermediate table *filp* (declared as *EXTERN*), used to hold the current file position and i-node pointer (see Fig. 5-39). It also tells whether the file was opened for reading, writing, or both, and how many file descriptors are currently pointing to the entry.

The file locking table, *file_lock* (declared as *EXTERN*), is in *lock.h* (line 21800). The size of the array is determined by *NR_LOCKS*, which is defined as 8 in *const.h.* This number should be increased if it is desired to implement a multiuser data base on a MINIX 3 system.

In *inode.h* (line 21900) the i-node table *inode* is declared (using *EXTERN*). It holds i-nodes that are currently in use. As we said earlier, when a file is opened its i-node is read into memory and kept there until the file is closed. The *inode* structure definition provides for information that is kept in memory, but is not written to the disk i-node. Notice that there is only one version, and nothing is version-specific here. When the i-node is read in from the disk, differences between V1 and V2/V3 file systems are handled. The rest of the file system does not need to know about the file system format on the disk, at least until the time comes to write back modified information.

Most of the fields should be self-explanatory at this point. However, *i_seek* deserves some comment. It was mentioned earlier that, as an optimization, when the file system notices that a file is being read sequentially, it tries to read blocks into the cache even before they are asked for. For randomly accessed files there is no read ahead. When an `lseek` call is made, the field *i_seek* is set to inhibit read ahead.

The file *param.h* (line 22000) is analogous to the file of the same name in the process manager. It defines names for message fields containing parameters, so the code can refer to, for example, *m_in.buffer,* instead of *m_in.m1_p1*, which selects one of the fields of the message buffer *m_in*.

In *super.h* (line 22100), we have the declaration of the superblock table. When the system is booted, the superblock for the root device is loaded here. As file systems are mounted, their superblocks go here as well. As with other tables, *super_block* is declared as *EXTERN*.

File System Storage Allocation

The last file we will discuss in this section is not a header. However, just as we did when discussing the process manager, it seems appropriate to discuss *table.c* immediately after reviewing the header files, since they are all included when *table.c* (line 22200) is compiled. Most of the data structures we have mentionedthe block cache, the *filp* table, and so onare defined with the *EXTERN* macro, as are also the file system's global variables and the file system's part of the process table. In the same way we have seen in other parts of the MINIX 3 system, the storage is actually reserved when *table.c* is compiled. This file also contains one major initialized array. *Call_vector* contains the pointer array used in the main loop for determining which procedure handles which system call number. We saw a similar table inside the process manager.

5.7.2. Table Management

Associated with each of the main tablesblocks, i-nodes, superblocks, and so forthis a file that contains procedures that manage the table. These procedures are heavily used by the rest of the file system and form the principal interface between tables and the file system. For this reason, it is appropriate to begin our study of the file system code with them.

Block Management

The block cache is managed by the procedures in the file *cache.c.* This file contains the nine procedures listed in Fig. 5-40. The first one, *get_block* (line 22426), is the standard way the file system gets data blocks. When a file system procedure needs to read a user data block, a directory block, a superblock, or any other kind of block, it calls *get_block*, specifying the device and block number.

Figure 5-40. Procedures used for block management.

(This item is displayed on page 571 in the print version)

Procedure	Function
get_block	Fetch a block for reading or writing
put_block	Return a block previously requested with get_block
alloc_zone	Allocate a new zone (to make a file longer)
free_zone	Release a zone (when a file is removed)
rw_block	Transfer a block between disk and cache
invalidate	Purge all the cache blocks for some device
flushall	Flush all dirty blocks for one device
rw_scattered	Read or write scattered data from or to a device
rm_lru	Remove a block from its LRU chain

When *get_block* is called, it first examines the block cache to see if the requested block is there. If so, it returns a pointer to it. Otherwise, it has to read the block in. The blocks in the cache are linked together on *NR_BUF_HASH* linked lists. *NR_BUF_HASH* is a tunable parameter, along with *NR_BUFS*, the size of the block cache. Both of these are set in *include/minix/config.h.* At the end of this section we will say a few words about optimizing the size of the block cache and the hash table. The *HASH_MASK* is *NR_BUF_HASH* - 1. With 256 hash lists, the mask is 255, so all the blocks on each list have block numbers that end with the same string of 8 bits, that is 00000000, 00000001, ..., or 11111111.

The first step is usually to search a hash chain for a block, although there is a special case, when a hole in a sparse file is being read, where this search is skipped. This is the reason for the test on line 22454. Otherwise, the next two lines set *bp* to point to the start of the list on which the requested block would be, if it were in the cache, applying *HASH_MASK* to the block number. The loop on the next line searches this list to see if the block can be found. If it is found and is not in use, it is removed from the LRU list. If it is already in use, it is not on the LRU list anyway. The pointer to the found block is returned to the caller on line 22463.

If the block is not on the hash list, it is not in the cache, so the least recently used block from the LRU list is taken. The buffer chosen is removed from its hash chain, since it is about to acquire a new block number and hence belongs on a different hash chain. If it is dirty, it is rewritten to the disk on line 22495. Doing this with a call to *flushall* rewrites any other dirty blocks for the same device. This call is is the way most blocks get written. Blocks that are currently in use are never chosen for eviction, since they are not on the LRU chain. Blocks will hardly ever be found to be in use, however; normally a block is released by *put_block* immediately upon being used.

As soon as the buffer is available, all of the fields, including *b_dev*, are updated with the new parameters (lines 22499 to 22504), and the block may be read in from the disk. However, there are two occasions when it may not be necessary to read the block from the disk. *Get_block* is called with a parameter *only_search*. This may indicate that this is a prefetch. During a prefetch an available buffer is found, writing the old contents to the disk if necessary, and a new block number is assigned to the buffer, but the *b_dev* field is set to *NO_DEV* to signal there are as yet no valid data in this block. We will see how this is used when we discuss the *rw_scattered* function. *Only_search* can also be used to signal that the file system needs a block just to rewrite all of it. In this case it is wasteful to first read the old version in. In either of these cases the parameters are updated, but the actual disk read is omitted (lines 22507 to 22513). When the new block has been read in, *get_block* returns to its caller with a pointer to it.

Suppose that the file system needs a directory block temporarily, to look up a file name. It calls *get_block* to acquire the directory block. When it has looked up its file name, it calls *put_block* (line 22520) to return the block to the cache, thus making the buffer available in case it is needed later for a different block.

Put_block takes care of putting the newly returned block on the LRU list, and in some cases, rewriting it to the disk. At line 22544 a decision is made to put it on the front or rear of the LRU list. Blocks on a RAM disk are always put on the front of the queue. The block cache does not really do very much for a RAM disk, since its data are already in memory and accessible without actual I/O. The *ONE_SHOT* flag is tested to see if the block has been marked as one not likely to be needed again soon, and such blocks are put on the front, where they will be reused quickly. However, this is used rarely, if at all. Almost all blocks except those from the RAM disk are put on the rear, in case they are needed again soon.

After the block has been repositioned on the LRU list, another check is made to see if the block should be rewritten to disk immediately. Like the previous test, the test for *WRITE_IMMED* is a vestige of an abandoned experiment; currently no blocks are marked for immediate writing.

As a file grows, from time to time a new zone must be allocated to hold the new data. The procedure *alloc_zone* (line 22580) takes care of allocating new zones. It does this by finding a free zone in the zone bitmap. There is no need to search through the bitmap if this is to be the first zone in a file; the *s_zsearch* field in the superblock, which always points to the first available zone on the device, is consulted. Otherwise an attempt is made to find a zone close to the last existing zone of the current file, in order to keep the zones of a file together. This is done by starting the search of the bitmap at this last zone (line 22603). The mapping between the bit number in the bitmap and the zone number is handled on line 22615, with bit 1 corresponding to the first data zone.

When a file is removed, its zones must be returned to the bitmap. *Free_zone* (line 22621) is responsible for returning these zones. All it does is call *free_bit*, passing the zone map and the bit number as parameters. *Free_bit* is also used to return free i-nodes, but then with the i-node map as the first parameter, of course.

Managing the cache requires reading and writing blocks. To provide a simple disk interface, the procedure *rw_block* (line 22641) has been provided. It reads or writes one block. Analogously, *rw_inode* exists to read and write i-nodes.

The next procedure in the file is *invalidate* (line 22680). It is called when a disk is unmounted, for example, to remove from the cache all the blocks belonging to the file system just unmounted. If this were not done, then when the device were reused (with a different floppy disk), the file system might find the old blocks instead of the new ones.

We mentioned earlier that *flushall* (line 22694), called from *get_block* whenever a dirty block is removed from the LRU list, is the function responsible for writing most data. It is also called by the `sync` system call to flush to disk all dirty buffers belonging to a specific device. `Sync` is activated periodically by the update daemon, and calls *flushall* once for each mounted device. *Flushall* treats the buffer cache as a linear array, so all dirty buffers are found, even ones that are currently in use and are not in the LRU list. All buffers in the cache are scanned, and those that belong to the device to be flushed and that need to be written are added to an array of pointers, *dirty*. This array is declared as *static* to keep it off the stack. It is then passed to *rw_scattered*.

In MINIX 3 scheduling of disk writing has been removed from the disk device drivers and made the sole responsibility of *rw_scattered* (line 22711). This function receives a device identifier, a pointer to an array of pointers to buffers, the size of the array, and a flag indicating whether to read or write. The first thing it does is sort the array it receives on the block numbers, so the actual read or write operation will be performed in an efficient order. It then constructs vectors of contiguous blocks to send to the the device driver with a call to *dev_io*. The driver does not have to do any additional scheduling. It is likely with a modern disk that the drive electronics will further optimize the order of requests, but this is not visible to MINIX 3. *Rw_scattered* is called with the *WRITING* flag only from the *flushall* function described above. In this case the origin of these block numbers is easy to understand. They are buffers which contain data from blocks previously read but now modified. The only call to *rw_scattered* for a read operation is from *rahead* in *read.c.* At this point, we just need to know that before calling *rw_scattered*, *get_block* has been called repeatedly in prefetch mode, thus reserving a group of buffers. These buffers contain block numbers, but no valid device parameter. This is not a problem, since *rw_scattered* is called with a device parameter as one of its arguments.

There is an important difference in the way a device driver may respond to a read (as opposed to a write) request, from *rw_scattered*. A request to write a number of blocks *must* be honored completely, but a request to read a number of blocks may be handled differently by different drivers, depending upon what is most efficient for the particular driver. *Rahead* often calls *rw_scattered* with a request for a list of blocks that may not actually be needed, so the best response is to get as many blocks as can be gotten easily, but not to go wildly seeking all over a device that may have a substantial seek time. For instance, the floppy driver may stop at a track boundary, and many other drivers will read only consecutive blocks. When the read is complete, *rw_scattered* marks the blocks read by filling in the device number field in their block buffers.

The last function in Fig. 5-40 is *rm_lru* (line 22809). This function is used to remove a block from the LRU list. It is used only by *get_block* in this file, so it is declared *PRIVATE* instead of *PUBLIC* to hide it from procedures outside the file.

Before we leave the block cache, let us say a few words about fine-tuning it. *NR_BUF_HASH* must be a power of 2. If it is larger than *NR_BUFS*, the average length of a hash chain will be less than one. If there is enough memory for a large number of buffers, there is space for a large number of hash chains, so the usual choice is to make *NR_BUF_HASH* the next power of 2 greater than *NR_BUFS*. The listing in the text shows settings of 128 blocks and 128 hash lists. The optimal size depends upon how the system is used, since that determines how much must be buffered. The full source code used to compile the standard MINIX 3 binaries that are installed from the CD-ROM that accommpanies this text has settings of 1280 buffers and 2048 hash chains. Empirically it was found that increasing the number of buffers beyond this did not improve performance when recompiling the MINIX 3 system, so apparently this is large enough to hold the binaries for all compiler passes. For some other kind of work a smaller size might be adequate or a larger size might improve performance.

The buffers for the standard MINIX 3 system on the CD-ROM occupy more than 5 MB of RAM. An additional binary, designated *image_small* is provided that was compiled with just 128 buffers in the block cache, and the buffers for this system need only a little more than 0.5 MB. This one can be installed on a system with only 8 MB of RAM. The standard version requires 16 MB of RAM. With some tweaking, it could no doubt be shoehorned into a memory of 4 MB or smaller.

I-Node Management

The block cache is not the only file system table that needs support procedures. The i-node table does, too. Many of the procedures are similar in function to the block management procedures. They are listed in Fig. 5-41.

Figure 5-41. Procedures used for i-node management.

Procedure	Function
get_inode	Fetch an i-node into memory
put_inode	Return an i-node that is no longer needed
alloc_inode	Allocate a new i-node (for a new file)
wipe_inode	Clear some fields in an i-node
free_inode	Release an i-node (when a file is removed)
update_times	Update time fields in an i-node
rw_inode	Transfer an i-node between memory and disk

Procedure	Function
old_icopy	Convert i-node contents to write to V1 disk i-node
new_icopy	Convert data read from V1 file system disk i-node
dup_inode	Indicate that someone else is using an i-node

The procedure *get_inode* (line 22933) is analogous to *get_block*. When any part of the file system needs an i-node, it calls *get_inode* to acquire it. *Get_inode* first searches the *inode* table to see if the i-node is already present. If so, it increments the usage counter and returns a pointer to it. This search is contained on lines 22945 to 22955. If the i-node is not present in memory, the i-node is loaded by calling *rw_inode*.

When the procedure that needed the i-node is finished with it, the i-node is returned by calling the procedure *put_inode* (line 22976), which decrements the usage count *i_count*. If the count is then zero, the file is no longer in use, and the i-node can be removed from the table. If it is dirty, it is rewritten to disk.

If the *i_link* field is zero, no directory entry is pointing to the file, so all its zones can be freed. Note that the usage count going to zero and the number of links going to zero are different events, with different causes and different consequences. If the i-node is for a pipe, all the zones must be released, even though the number of links may not be zero. This happens when a process reading from a pipe releases the pipe. There is no sense in having a pipe for one process.

When a new file is created, an i-node must be allocated by *alloc_inode* (line 23003). MINIX 3 allows mounting of devices in read-only mode, so the superblock is checked to make sure the device is writable. Unlike zones, where an attempt is made to keep the zones of a file close together, any i-node will do. In order to save the time of searching the i-node bitmap, advantage is taken of the field in the superblock where the first unused i-node is recorded.

After the i-node has been acquired, *get_inode* is called to fetch the i-node into the table in memory. Then its fields are initialized, partly in-line (lines 23038 to 23044) and partly using the procedure *wipe_inode* (line 23060). This particular division of labor has been chosen because *wipe_inode* is also needed elsewhere in the file system to clear certain i-node fields (but not all of them).

When a file is removed, its i-node is freed by calling *free_inode* (line 23079). All that happens here is that the corresponding bit in the i-node bitmap is set to 0 and the superblock's record of the first unused i-node is updated.

The next function, *update_times* (line 23099), is called to get the time from the system clock and change the time fields that require updating. *Update_times* is also called by the `stat` and `fstat` system calls, so it is declared *PUBLIC*.

The procedure *rw_inode* (line 23125) is analogous to *rw_block*. Its job is to fetch an i-node from the disk. It does its work by carrying out the following steps:

1. Calculate which block contains the required i-node.
2. Read in the block by calling *get_block*.
3. Extract the i-node and copy it to the *inode* table.
4. Return the block by calling *put_block*.

Rw_inode is a bit more complex than the basic outline given above, so some additional functions are needed. First, because getting the current time requires a kernel call, any need for a change to the time fields in the i-node is only marked by setting bits in the i-node's *i_update* field while the i-node is in memory. If this field is nonzero when an i-node must be written, *update_times* is called.

Second, the history of MINIX adds a complication: in the old *V1* file system the i-nodes on the disk have a different structure from *V2*. Two functions, *old_icopy* (line 23168) and *new_icopy* (line 23214) are provided to take care of the conversions. The first converts between i-node information in memory and the format used by the *V1* filesystem. The second does the same conversion for *V2* and *V3* filesystem disks. Both of these functions are called only from within this file, so they are declared *PRIVATE*. Each function handles conversions in both directions (disk to memory or memory to disk).

Older versions of MINIX were ported to systems which used a different byte order from Intel processors and MINIX 3 is also likely to be ported to such architectures in the future. Every implementation uses the native byte order on its disk; the *sp->native* field in the superblock identifies which order is used. Both *old_icopy* and *new_icopy* call functions *conv2* and *conv4* to swap byte orders, if necessary. Of course, much of what we have just described is not used by MINIX 3, since it does not support the V1 filesystem to the extent that V1 disks can be used. And as of this writing nobody has ported MINIX 3 to a platform that uses a different byte order. But these bits and pieces remain in place for the day when someone decides to make MINIX 3 more versatile.

The procedure *dup_inode* (line 23257) just increments the usage count of the i-node. It is called when an open file is opened again. On the second open, the inode need not be fetched from disk again.

Superblock Management

The file *super.c* contains procedures that manage the superblock and the bitmaps. Six procedures are defined in this file, listed in Fig. 5-42.

Figure 5-42. Procedures used to manage the superblock and bitmaps.

Procedure	Function
alloc_bit	Allocate a bit from the zone or i-node map
free_bit	Free a bit in the zone or i-node map
get_super	Search the superblock table for a device
get_block_size	Find block size to use
mounted	Report whether given i-node is on a mounted (or root) file system
read_super	Read a superblock

When an i-node or zone is needed, *alloc_inode* or *alloc_zone* is called, as we have seen above. Both of these call *alloc_bit* (line 23324) to actually search the relevant bitmap. The search involves three nested loops, as follows:

1. The outer one loops on all the blocks of a bitmap.

2. The middle one loops on all the words of a block.

3. The inner one loops on all the bits of a word.

The middle loop works by seeing if the current word is equal to the one's complement of zero, that is, a complete word full of 1s. If so, it has no free i-nodes or zones, so the next word is tried. When a word with a different value is found, it must have at least one 0 bit in it, so the inner loop is entered to find the free (i.e., 0) bit. If all the blocks have been tried without success, there are no free i-nodes or zones, so the code *NO_BIT* (0) is returned. Searches like this can consume a lot of processor time, but the use of the superblock fields that point to the first unused i-node and zone, passed to *alloc_bit* in *origin*, helps to keep these searches short.

Freeing a bit is simpler than allocating one, because no search is required. *Free_bit* (line 23400) calculates which bitmap block contains the bit to free and sets the proper bit to 0 by calling *get_block*, zeroing the bit in memory and then calling *put_block*.

The next procedure, *get_super* (line 23445), is used to search the superblock table for a specific device. For example, when a file system is to be mounted, it is necessary to check that it is not already mounted. This check can be performed by asking *get_super* to find the file system's device. If it does not find the device, then the file system is not mounted.

In MINIX 3 the file system server is capable of handling file systems with different block sizes, although within a given disk partition only a single block size can be used. The *get_block_size* function (line 23467) is meant to determine the block size of a file system. It searches the superblock table for the given device and returns the block size of the device if it is mounted. Otherwise the minimum block size, *MIN_BLOCK_SIZE* is returned.

The next function, *mounted* (line 23489), is called only when a block device is closed. Normally, all cached data for a device are discarded when it is closed. But, if the device happens to be mounted, this is not desirable. *Mounted* is called with a pointer to the i-node for a device. It just returns *TRUE* if the device is the root device, or if it is a mounted device.

Finally, we have *read_super* (line 23509). This is partially analogous to *rw_block* and *rw_inode*, but it is called only to read. The superblock is not read into the block cache at all, a request is made directly to the device for 1024 bytes starting at an offset of the same amount from the beginning of the device. Writing a superblock is not necessary in the normal operation of the system. *Read_super* checks the version of the file system from which it has just read and performs conversions, if necessary, so the copy of the superblock in memory will have the standard structure even when read from a disk with a different superblock structure or byte order.

Even though it is not currently used in MINIX 3, the method of determining whether a disk was written on a system with a different byte order is clever and worth noting. The magic number of a superblock is written with the native byte order of the system upon which the file system was created, and when a superblock is read a test for reversed-byte-order superblocks is made.

File Descriptor Management

MINIX 3 contains special procedures to manage file descriptors and the *filp* table (see Fig. 5-39). They are contained in the file *filedes.c.* When a file is created or opened, a free file descriptor and a free *filp* slot are needed. The procedure *get_fd* (line 23716) is used to find them. They are not marked as in use, however, because many checks must first be made before it is known for sure that the `creat` or `open` will succeed.

Get_filp (line 23761) is used to see if a file descriptor is in range, and if so, returns its *filp* pointer.

The last procedure in this file is *find_filp* (line 23774). It is needed to find out when a process is writing on a broken pipe (i.e., a pipe not open for reading by any other process). It locates potential readers by a brute force search of the *filp* table. If it cannot find one, the pipe is broken and the write fails.

File Locking

The POSIX record locking functions are shown in Fig. 5-43. A part of a file can be locked for reading and writing, or for writing only, by an `fcntl` call specifying a *F_SETLK* or *F_SETLKW* request. Whether a lock exists over a part of a file can be determined using the *F_GETLK* request.

Figure 5-43. The POSIX advisory record locking operations. These operations are requested by using an FCNTL system call.

Operation	Meaning
F_SETLK	Lock region for both reading and writing
F_SETLKW	Lock region for writing

Operation	Meaning
F_GETLK	Report if region is locked

The file *lock.c* contains only two functions. *Lock_op* (line 23820) is called by the `fcntl` system call with a code for one of the operations shown in Fig. 5-43. It does some error checking to be sure the region specified is valid. When a lock is being set, it must not conflict with an existing lock, and when a lock is being cleared, an existing lock must not be split in two. When any lock is cleared, the other function in this file, *lock_revive* (line 23964), is called. It wakes up all the processes that are blocked waiting for locks.

This strategy is a compromise; it would take extra code to figure out exactly which processes were waiting for a particular lock to be released. Those processes that are still waiting for a locked file will block again when they start. This strategy is based on an assumption that locking will be used infrequently. If a major multiuser data base were to be built upon a MINIX 3 system, it might be desirable to reimplement this.

Lock_revive is also called when a locked file is closed, as might happen, for instance, if a process is killed before it finishes using a locked file.

5.7.3. The Main Program

The main loop of the file system is contained in file *main.c,* (line 24040). After a call to *fs_init* for initialization, the main loop is entered. Structurally, this is very similar to the main loop of the process manager and the I/O device drivers. The call to *get_work* waits for the next request message to arrive (unless a process previously suspended on a pipe or terminal can now be handled). It also sets a global variable, *who*, to the caller's process table slot number and another global variable, *call_nr*, to the number of the system call to be carried out.

Once back in the main loop the variable *fp* is pointed to the caller's process table slot, and the *super_user* flag tells whether the caller is the superuser or not. Notification messages are high priority, and a *SYS_SIG* message is checked for first, to see if the system is shutting down. The second highest priority is a *SYN_ALARM*, which means that a timer set by the file system has expired. A *NOTIFY_MESSAGE* means a device driver is ready for attention, and is dispatched to *dev_status.* Then comes the main attractionthe call to the procedure that carries out the system call. The procedure to call is selected by using *call_nr* as an index into the array of procedure pointers, *call_vecs.*

When control comes back to the main loop, if *dont_reply* has been set, the reply is inhibited (e.g., a process has blocked trying to read from an empty pipe). Otherwise a reply is sent by calling *reply* (line 24087). The final statement in the main loop has been designed to detect that a file is being read sequentially and to load the next block into the cache before it is actually requested, to improve performance.

Two other functions in this file are intimately involved with the file system's main loop. *Get_work* (line 24099) checks to see if any previously blocked procedures have now been revived. If so, these have priority over new messages. When there is no internal work to do the file system calls the kernel to get a message, on line 24124. Skipping ahead a few lines, we find *reply* (line 24159) which is called after a system call has been completed, successfully or otherwise. It sends a reply back to the caller. The process may have been killed by a signal, so the status code returned by the kernel is ignored. In this case there is nothing to be done anyway.

Initialization of the File System

The functions that remain to be discussed in *main.c* are used at system startup. The major player is *fs_init*, which is called by the file system before it enters its main loop during startup of the entire system. In the context of discussing process scheduling in Chapter 2 we showed in Fig. 2-43 the initial queueing of processes as the MINIX 3 system starts up. The file system is scheduled on a queue with lower priority than the process manager, so we can be sure that at startup time the process manager will get a chance to run before the file system. In Chapter 4 we examined the initialization of the process manager. As the PM builds its part of the process table, adding entries for itself and all other processes in the boot image, it sends a message to the file system for each one so the FS can initialize the corresponding entry in the FS part of the file system. Now we can see the other half of this interaction.

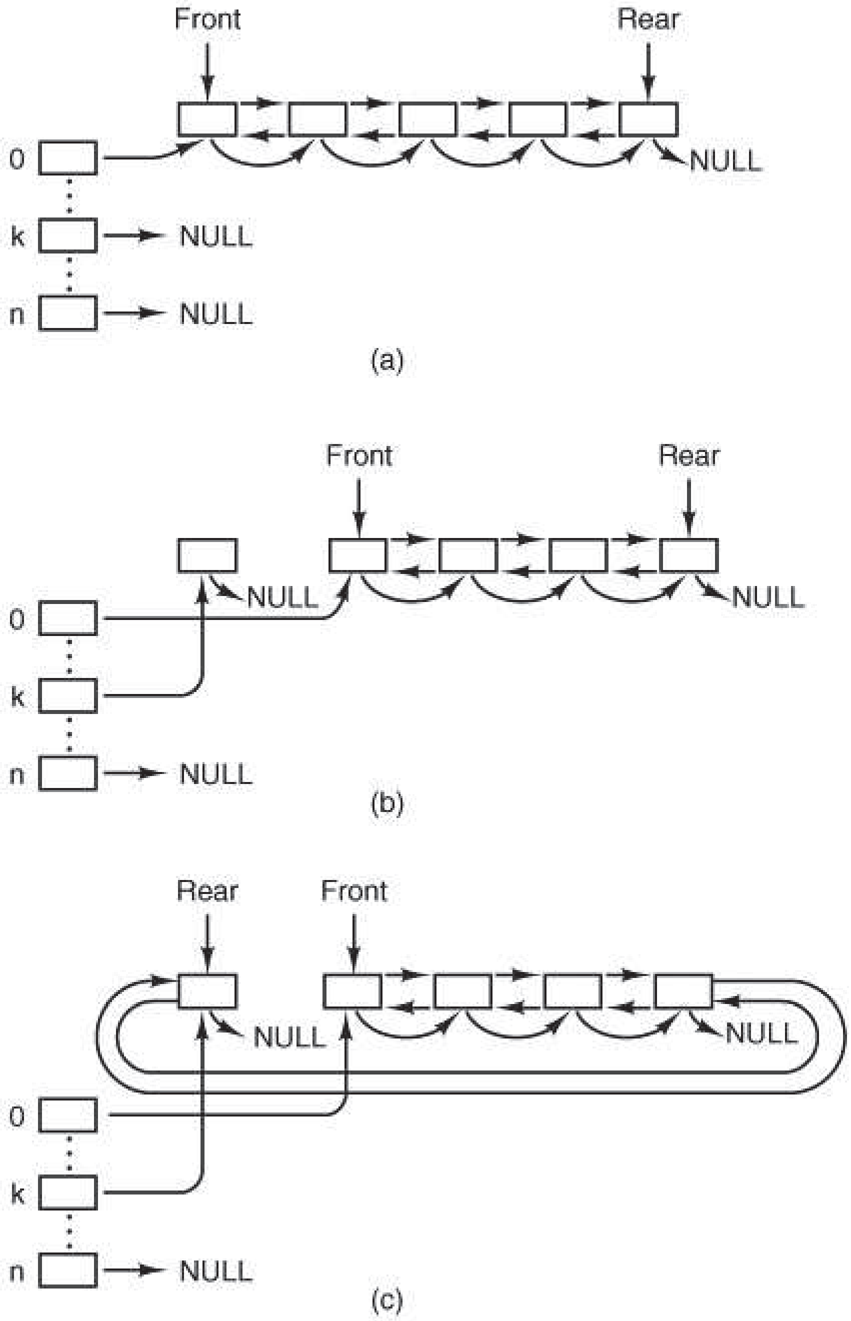
When the file system starts it immediately enters a loop of its own in *fs_init*, on lines 24189 to 24202. The first statement in the loop is a call to `receive`, to get a message sent at line 18235 in the PM's *pm_init* initialization function. Each message contains a process number and a PID. The first is used as an index into the file system's process table and the second is saved in the *fp_pid* field of each selected slot. Following this the real and effective uid and gid for the superuser and a ~0 (all bits set) umask is set up for each selected slot. When a message with the symbolic value *NONE* in the process number field is received the loop terminates and a message is sent back to the process manager to tell it all is OK.

Next, the file system's own initialization is completed. First important constants are tested for valid values. Then several other functions are invoked to initialize the block cache and the device table, to load the RAM disk if necessary, and to load the root device superblock. At this point the root device can be accessed, and another loop is made through the FS part of the process table, so each process loaded from the boot image will recognize the root directory and use the root directory as its working directory (lines 24228 to 24235).

The first function called by *fs_init* after it finshes its interaction with the process manager is *buf_pool*, which begins on line 24132. It builds the linked lists used by the block cache. Figure 5-37 shows the normal state of the block cache, in which all blocks are linked on both the LRU chain and a hash chain. It may be helpful to see how the situation of Fig. 5-37 comes about. Immediately after the cache is initialized by *buf_pool*, all the buffers will be on the LRU chain, and all will be linked into the 0th hash chain, as in Fig. 5-44(a). When a buffer is requested, and while it is in use, we have the situation of Fig. 5-44(b), in which we see that a block has been removed from the LRU chain and is now on a different hash chain.

Figure 5-44. Block cache initialization. (a) Before any buffers have been used. (b) After one block has been requested. (c) After the block has been released.

(This item is displayed on page 581 in the print version)



Normally, blocks are released and returned to the LRU chain immediately. Figure 5-44(c) shows the situation after the block has been returned to the LRU chain. Although it is no longer in use, it can be accessed again to provide the same data, if need be, and so it is retained on the hash chain. After the system has been in operation for awhile, almost all of the blocks can be expected to have been used and to be distributed among the different hash chains at random. Then the LRU chain will look like Fig. 5-37.

The next thing called after *buf_pool* is *build_dmap*, which we will describe later, along with other functions dealing with device files. After that, *load_ram* is called, which uses the next function we will examine, *igetenv* (line 2641). This function retrieves a numeric device identifier from the kernel, using the name of a boot parameter as a key. If you have used the *sysenv* command to to look at the boot parameters on a working MINIX 3 system, you have seen that *sysenv* reports devices numerically, displaying strings like

```
rootdev=912
```

The file system uses numbers like this to identify devices. The number is simply 256 x `major` + `minor`, where `major` and `minor` are the major and minor device numbers. In this example, the major, minor pair is 3, 144, which corresponds to */dev/c0d1p0s0,* a typical place to install MINIX 3 on a system with two disk drives.

Load_ram (line 24260) allocates space for a RAM disk, and loads the root file system on it, if required by the boot parameters. It uses *igetenv* to get the *rootdev*, *ramimagedev*, and *ramsize* parameters set in the boot environment (lines 24278 to 24280). If the boot parameters specify

```
rootdev = ram
```

the root file system is copied from the device named by *ramimagedev* to the RAM disk block by block, starting with the boot block, with no interpretation of the various file system data structures. If the *ramsize* boot parameter is smaller than the size of *ramimagedev*, the RAM disk is made large enough to hold it. If *ramsize* specifies a size larger than the boot device file system the requested size is allocated and the RAM disk file system is adjusted to use the full size specified (lines 24404 to 24420). This is the only time that the file system ever writes a superblock, but, just as with reading a superblock, the block cache is not used and the data is written directly to the device using *dev_io*.

Two items merit note at this point. The first is the code on lines 24291 to 24307 which deals with the case of booting from a CD-ROM. The *cdprobe* function, not discussed in this text, is used. Interested readers are referred to the code in *fs/cdprobe.c,* which can be found on the CD-ROM or the Web site. Second, regardless of the disk block size used by MINIX 3 for ordinary disk access, the boot block is always a 1 KB block and the superblock is loaded from the second 1 KB of the disk device. Anything else would be complicated, since the block size cannot be known until the superblock has been loaded.

Load_ram allocates space for an empty RAM disk if a nonzero *ramsize* is specified without a request to use the RAM disk as the root file system. In this case, since no file system structures are copied, the RAM device cannot be used as a file system until it has been initialized by the *mkfs* command. Alternatively, such a RAM disk can be used for a secondary cache if support for this is compiled into the file system.

The last function in *main.c* is *load_super* (line 24426). It initializes the superblock table and reads in the superblock of the root device.

5.7.4. Operations on Individual Files

In this section we will look at the system calls that operate on individual files one at a time (as opposed to, say, operations on directories). We will start with how files are created, opened, and closed. After that we will examine in some detail the mechanism by which files are read and written. Then that we will look at pipes and how operations on them differ from those on files.

Creating, Opening, and Closing Files

The file *open.c* contains the code for six system calls: `creat`, `open`, `mknod`, `mkdir`, `close`, and `lseek`. We will examine creat and open together, and then look at each of the others.

In older versions of UNIX, the `creat` and `open` calls had distinct purposes. Trying to open a file that did not exist was an error, and a new file had to be created with `creat`, which could also be used to truncate an existing file to zero length. The need for two distinct calls is no longer present in a POSIX system, however. Under POSIX, the `open` call now allows creating a new file or truncating an old file, so the `creat` call now represents a subset of the possible uses of the `open` call and is really only necessary for compatibility with older programs. The procedures that handle `creat` and `open` are *do_creat* (line 24537) and *do_open* (line 24550). (As in the process manager, the convention is used in the file system that system call `xxx` is performed by procedure *do_XXX*.) Opening or creating a file involves three steps:

1. Finding the i-node (allocating and initializing if the file is new).
2. Finding or creating the directory entry.
3. Setting up and returning a file descriptor for the file.

Both the `creat` and the `open` calls do two things: they fetch the name of a file and then they call *common_open* which takes care of tasks common to both calls.

Common_open (line 24573) starts by making sure that free file descriptor and *filp* table slots are available. If the calling function specified creation of a new file (by calling with the *O_CREAT* bit set), *new_node* is called on line 24594. *New_node* returns a pointer to an existing i-node if the directory entry already exists; otherwise it will create both a new directory entry and i-node. If the i-node cannot be created, *new_node* sets the global variable *err_code*. An error code does not always mean an error. If *new_node* finds an existing file, the error code returned will indicate that the file exists, but in this case that error is acceptable (line 24597). If the *O_CREAT* bit is not set, a search is made for the i-node using an alternative method, the *eat_path* function in *path.c,* which we will discuss further on. At this point, the important thing to understand is that if an i-node is not found or successfully created, *common_open* will terminate with an error before line 24606 is reached. Otherwise, execution continues here with assignment of a file descriptor and claiming of a slot in the *filp* table, Following this, if a new file has just been created, lines 24612 to 24680 are skipped.

If the file is not new, then the file system must test to see what kind of a file it is, what its mode is, and so on, to determine whether it can be opened. The call to *forbidden* on line 24614 first makes a general check of the *rwx* bits. If the file is a regular file and *common_open* was called with the *O_TRUNC* bit set, it is truncated to length zero and *forbidden* is called again (line 24620), this time to be sure the file may be written. If the permissions allow, *wipe_inode* and *rw_inode* are called to re-initialize the i-node and write it to the disk. Other file types (directories, special files, and named pipes) are subjected to appropriate tests. In the case of a device, a call is made on line 24640 (using the *dmap* structure) to the appropriate routine to open the device. In the case of a named pipe, a call is made to *pipe_open* (line 24646), and various tests relevant to pipes are made.

The code of *common_open*, as well as many other file system procedures, contains a large amount of code that checks for various errors and illegal combinations. While not glamorous, this code is essential to having an error-free, robust file system. If something is wrong, the file descriptor and *filp* slot previously allocated are deallocated and the i-node is released (lines 24683 to 24689). In this case the value returned by *common_open* will be a negative number, indicating an error. If there are no problems the file descriptor, a positive value, is returned.

This is a good place to discuss in more detail the operation of *new_node* (line 24697), which does the allocation of the i-node and the entering of the path name into the file system for `creat` and `open` calls. It is also used for the `mknod` and `mkdir` calls, yet to be discussed. The statement on line 24711 parses the path name (i.e., looks it up component by component) as far as the final directory; the call to *advance* three lines later tries to see if the final component can be opened.

For example, on the call

```
fd = creat("/usr/ast/foobar", 0755);
```

last_dir tries to load the i-node for */usr/ast/* into the tables and return a pointer to it. If the file does not exist, we will need this i-node shortly in order to add *foobar* to the directory. All the other system calls that add or delete files also use *last_dir* to first open the final directory in the path.

If *new_node* discovers that the file does not exist, it calls *alloc_inode* on line 24717 to allocate and load a new i-node, returning a pointer to it. If no free inodes are left, *new_node* fails and returns *NIL_INODE*.

If an i-node can be allocated, the operation continues at line 24727, filling in some of the fields, writing it back to the disk, and entering the file name in the final directory (on line 24732). Again we see that the file system must constantly check for errors, and upon encountering one, carefully release all the resources, such as i-nodes and blocks that it is holding. If we were prepared to just let MINIX 3 panic when we ran out of, say, i-nodes, rather than undoing all the effects of the current call and returning an error code to the caller, the file system would be appreciably simpler.

As mentioned above, pipes require special treatment. If there is not at least one reader/writer pair for a pipe, *pipe_open* (line 24758) suspends the caller. Otherwise, it calls *release*, which looks through the process table for processes that are blocked on the pipe. If it is successful, the processes are revived.

The `mknod` call is handled by *do_mknod* (line 24785). This procedure is similar to *do_creat*, except that it just creates the i-node and makes a directory entry for it. In fact, most of the work is done by the call to *new_node* on line 24797. If the i-node already exists, an error code will be returned. This is the same error code that was an acceptable result from *new_node* when it was called by *common_open*; in this case, however, the error code is passed back to the caller, which presumably will act accordingly. The case-by-case analysis we saw in *common_open* is not needed here.

The `mkdir` call is handled by the function *do_mkdir* (line 24805). As with the other system calls we have discussed here, *new_node* plays an important part. Directories, unlike files, always have links and are never completely empty because every directory must contain two entries from the time of its creation: the "*.*" and "*..*" entries that refer to the directory itself and to its parent directory. The number of links a file may have is limited, it is *LINK_MAX* (defined in *include/limits.h* as *SHRT_MAX,* 32767 for MINIX 3 on a standard 32-bit Intel system). Since the reference to a parent directory in a child is a link to the parent, the first thing *do_mkdir* does is to see if it is possible to make another link in the parent directory (lines 24819 and 24820). Once this test has been passed, *new_node* is called. If *new_node* succeeds, then the directory entries for *"."* and *".."* are made (lines 24841 and 24842). All of this is straightforward, but there could be failures (for instance, if the disk is full), so to avoid making a mess of things provision is made for undoing the initial stages of the process if it can not be completed.

Closing a file is easier than opening one. The work is done by *do_close* (line 24865). Pipes and special files need some attention, but for regular files, almost all that needs to be done is to decrement the *filp* counter and check to see if it is zero, in which case the i-node is returned with *put_inode*. The final step is to remove any locks and to revive any process that may have been suspended waiting for a lock on the file to be released.

Note that returning an i-node means that its counter in the *inode* table is decremented, so it can be removed from the table eventually. This operation has nothing to do with freeing the i-node (i.e., setting a bit in the bitmap saying that it is available). The i-node is only freed when the file has been removed from all directories.

The final procedure in *open.c* is *do_lseek* (line 24939). When a seek is done, this procedure is called to set the file position to a new value. On line 24968 reading ahead is inhibited; an explicit attempt to seek to a position in a file is incompatible with sequential access.

Reading a File

Once a file has been opened, it can be read or written. Many functions are used during both reading and writing. These are found in the file *read.c*. We will discuss these first and then proceed to the following file, *write.c,* to look at code specifically used for writing. Reading and writing differ in a number of ways, but they have enough similarities that all that is required of *do_read* (line 25030) is to call the common procedure *read_write* with a flag set to *READING*. We will see in the next section that *do_write* is equally simple.

Read_write begins on line 25038. Some special code on lines 25063 to 25066 is used by the process manager to have the file system load entire segments in user space for it. Normal calls are processed starting on line 25068. Some validity checks follow (e.g., reading from a file opened only for writing) and some variables are initialized. Reads from character special files do not go through the block cache, so they are filtered out on line 25122.

The tests on lines 25132 to 25145 apply only to writes and have to do with files that may get bigger than the device can hold, or writes that will create a hole in the file by writing *beyond* the end-of-file. As we discussed in the MINIX 3 overview, the presence of multiple blocks per zone causes problems that must be dealt with explicitly. Pipes are also special and are checked for.

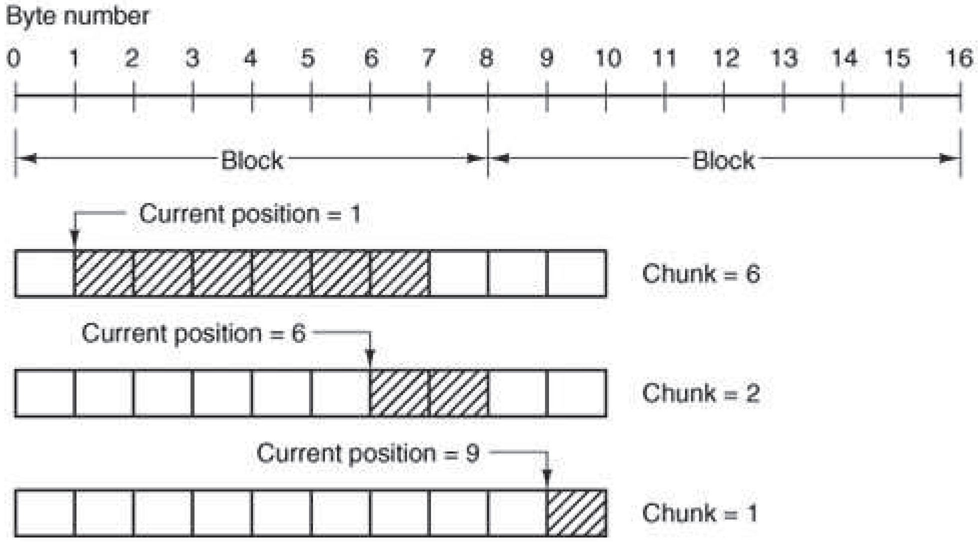
The heart of the read mechanism, at least for ordinary files, is the loop starting on line 25157. This loop breaks the request up into chunks, each of which fits in a single disk block. A chunk begins at the current position and extends until one of the following conditions is met:

1. All the bytes have been read.
2. A block boundary is encountered.
3. The end-of-file is hit.

These rules mean that a chunk never requires two disk blocks to satisfy it. Figure 5-45 shows three examples of how the chunk size is determined, for chunk sizes of 6, 2, and 1 bytes, respectively. The actual calculation is done on lines 25159 to 25169.

Figure 5-45. Three examples of how the first chunk size is determined for a 10-byte file. The block size is 8 bytes, and the number of bytes requested is 6. The chunk is shown shaded.

(This item is displayed on page 587 in the print version)



The actual reading of the chunk is done by *rw_chunk*. When control returns, various counters and pointers are incremented, and the next iteration begins. When the loop terminates, the file position and other variables may be updated (e.g., pipe pointers).

Finally, if read ahead is called for, the i-node to read from and the position to read from are stored in global variables, so that after the reply message is sent to the user, the file system can start getting the next block. In many cases the file system will block, waiting for the next disk block, during which time the user process will be able to work on the data it just received. This arrangement overlaps processing and I/O and can improve performance substantially.

The procedure *rw_chunk* (line 25251) is concerned with taking an i-node and a file position, converting them into a physical disk block number, and requesting the transfer of that block (or a portion of it) to the user space. The mapping of the relative file position to the physical disk address is done by *read_map*, which understands about i-nodes and indirect blocks. For an ordinary file, the variables *b* and *dev* on line 25280 and line 25281 contain the physical block number and device number, respectively. The call to *get_block* on line 25303 is where the cache handler is asked to find the block, reading it in if need be. Calling *rahead* on line 25295 then ensures that the block is read into the cache.

Once we have a pointer to the block, the *sys_vircopy* kernel call on line 25317 takes care of transferring the required portion of it to the user space. The block is then released by *put_block*, so that it can be evicted from the cache later. (After being acquired by *get_block*, it will not be in the LRU queue and it will not be returned there while the counter in the block's header shows that it is in use, so it will be exempt from eviction; *put_block* decrements the counter and returns the block to the LRU queue when the counter reaches zero.) The code on line 25327 indicates whether a write operation filled the block. However, the value passed to *put_block* in *n* does not affect how the block is placed on the queue; all blocks are now placed on the rear of the LRU chain.

Read_map (line 25337) converts a logical file position to the physical block number by inspecting the i-node. For blocks close enough to the beginning of the file that they fall within one of the first seven zones (the ones right in the i-node), a simple calculation is sufficient to determine which zone is needed, and then which block. For blocks further into the file, one or more indirect blocks may have to be read.

Rd_indir (line 25400) is called to read an indirect block. The comments for this function are a bit out of date; code to support the 68000 processor has been removed and the support for the MINIX V1 file system is not used and could also be dropped. However, it is worth noting that if someone wanted to add support for other file system versions or other platforms where data might have a different format on the disk, problems of different data types and byte orders could be relegated to this file. If messy conversions were necessary, doing them here would let the rest of the file system see data in only one form.

Read_ahead (line 25432) converts the logical position to a physical block number, calls *get_block* to make sure the block is in the cache (or bring it in), and then returns the block immediately. It cannot do anything with the block, after all. It just wants to improve the chance that the block is around if it is needed soon,

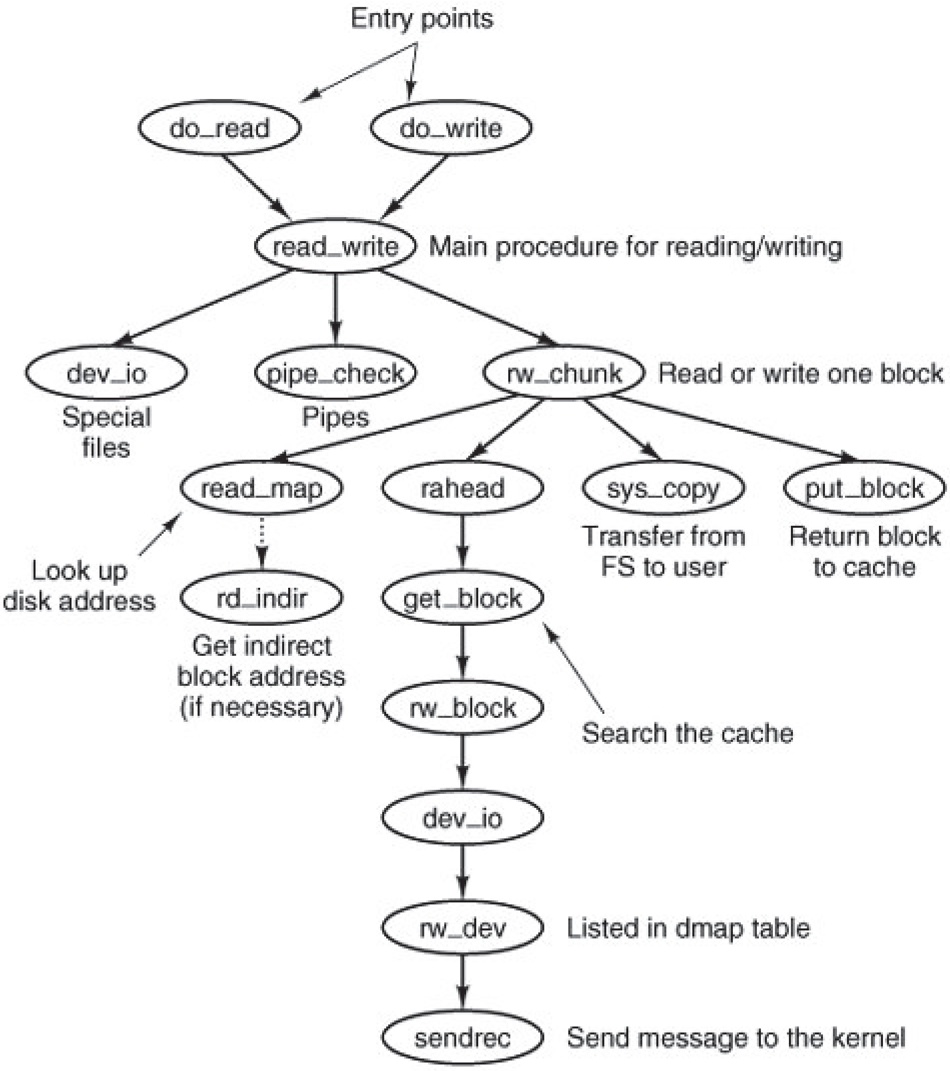
Note that *read_ahead* is called only from the main loop in *main*. It is not called as part of the processing of the `read` system call. It is important to realize that the call to *read_ahead* is performed *after* the reply is sent, so that the user will be able to continue running even if the file system has to wait for a disk block while reading ahead.

Read_ahead by itself is designed to ask for just one more block. It calls the last function in *read.c, rahead*, to actually get the job done. *Rahead* (line 25451) works according to the theory that if a little more is good, a lot more is better. Since disks and other storage devices often take a relatively long time to locate the first block requested but then can relatively quickly read in a number of adjacent blocks, it may be possible to get many more blocks read with little additional effort. A prefetch request is made to *get_block*, which prepares the block cache to receive a number of blocks at once. Then *rw_scattered* is called with a list of blocks. We have previously discussed this; recall that when the device drivers are actually called by *rw_scattered*, each one is free to answer only as much of the request as it can efficiently handle. This all sounds fairly complicated, but the complications make possible a significant speedup of applications which read large amounts of data from the disk.

Figure 5-46 shows the relations between some of the major procedures involved in reading a filein particular, who calls whom.

Figure 5-46. Some of the procedures involved in reading a file.

(This item is displayed on page 589 in the print version)



Writing a File

The code for writing to files is in *write. c*. Writing a file is similar to reading one, and *do_write* (line 25625) just calls *read_write* with the *WRITING* flag.A major difference between reading and writing is that writing requires allocating new disk blocks. *Write_map* (line 25635) is analogous to *read_map*, only instead of looking up physical block numbers in the i-node and its indirect blocks, it enters new ones there (to be precise, it enters zone numbers, not block numbers).

The code of *write_map* is long and detailed because it must deal with several cases. If the zone to be inserted is close to the beginning of the file, it is just inserted into the i-node on (line 25658).

The worst case is when a file exceeds the size that can be handled by a single-indirect block, so a double-indirect block is now required. Next, a single-indirect block must be allocated and its address put into the double-indirect block. As with reading, a separate procedure, *wr_indir*, is called. If the double-indirect block is acquired correctly, but the disk is full so the single-indirect block cannot be allocated, then the double one must be returned to avoid corrupting the bitmap.

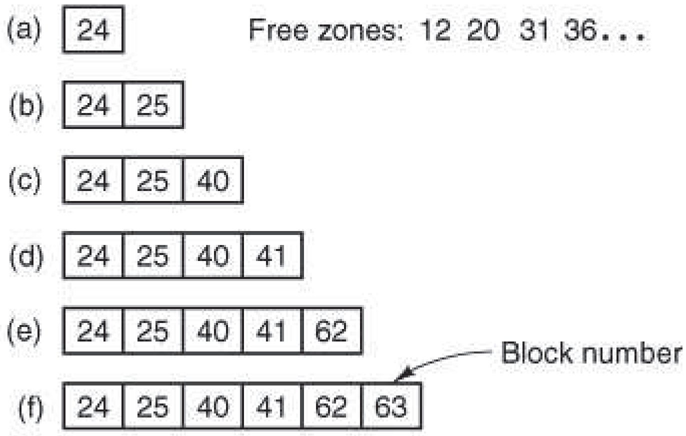
Again, if we could just toss in the sponge and panic at this point, the code would be much simpler. However, from the user's point of view it is much nicer that running out of disk space just returns an error from `write`, rather than crashing the computer with a corrupted file system.

Wr_indir (line 25726) calls the conversion routines, *conv4* to do any necessary data conversion and puts a new zone number into an indirect block. (Again, there is leftover code here to handle the old V1 filesystem, but only the V2 code is currently used.) Keep in mind that the name of this function, like the names of many other functions that involve reading and writing, is not literally true. The actual writing to the disk is handled by the functions that maintain the block cache.

The next procedure in *write.c* is *clear_zone* (line 25747), which takes care of the problem of erasing blocks that are suddenly in the middle of a file. This happens when a seek is done beyond the end of a file, followed by a write of some data. Fortunately, this situation does not occur very often.

New_block (line 25787) is called by *rw_chunk* whenever a new block is needed. Figure 5-47 shows six successive stages of the growth of a sequential file. The block size is 1-KB and the zone size is 2-KB in this example.

Figure 5-47. (a) (f) The successive allocation of 1-KB blocks with a 2-KB zone.



The first time *new_block* is called, it allocates zone 12 (blocks 24 and 25). The next time it uses block 25, which has already been allocated but is not yet in use. On the third call, zone 20 (blocks

40 and 41) is allocated, and so on. *Zero_block* (line 25839) clears a block, erasing its previous contents. This description is considerably longer than the actual code.

Pipes

Pipes are similar to ordinary files in many respects. In this section we will focus on the differences. The code we will discuss is all in *pipe.c.*

First of all, pipes are created differently, by the `pipe` call, rather than the `creat` call. The `pipe` call is handled by *do_pipe* (line 25933). All *do_pipe* really does is allocate an i-node for the pipe and return two file descriptors for it. Pipes are owned by the system, not by the user, and are located on the designated pipe device (configured in *include/minix/config.h),* which could very well be a RAM disk, since pipe data do not have to be preserved permanently.

Reading and writing a pipe is slightly different from reading and writing a file, because a pipe has a finite capacity. An attempt to write to a pipe that is already full will cause the writer to be suspended. Similarly, reading from an empty pipe will suspend the reader. In effect, a pipe has two pointers, the current position (used by readers) and the size (used by writers), to determine where data come from or go to.

The various checks to see if an operation on a pipe is possible are carried out by *pipe_check* (line 25986). In addition to the above tests, which may lead to the caller being suspended, *pipe_check* calls *release* to see if a process previously suspended due to no data or too much data can now be revived. These revivals are done on line 26017 and line 26052, for sleeping writers and readers, respectively. Writing on a broken pipe (no readers) is also detected here.

The act of suspending a process is done by *suspend* (line 26073). All it does is save the parameters of the call in the process table and set the flag *dont_reply* to *TRUE*, to inhibit the file system's reply message.

The procedure *release* (line 26099) is called to check if a process that was suspended on a pipe can now be allowed to continue. If it finds one, it calls *revive* to set a flag so that the main loop will notice it later. This function is not a system call, but is listed in Fig. 5-33(c) because it uses the message-passing mechanism.

The last procedure in *pipe.c* is *do_unpause* (line 26189). When the process manager is trying to signal a process, it must find out if that process is hanging on a pipe or special file (in which case it must be awakened with an *EINTR* error). Since the process manager knows nothing about pipes or special files, it sends a message to the file system to ask. That message is processed by *do_unpause*, which revives the process, if it is blocked. Like *revive*, *do_unpause* has some similarity to a system call, although it is not one.

The last two functions in *pipe.c, select_request_pipe* (line 26247) and *select_match_pipe* (line 26278), support the `select` call, which is not discussed here.

5.7.5. Directories and Paths

We have now finished looking at how files are read and written. Our next task is to see how path names and directories are handled.

Converting a Path to an I-Node

Many system calls (e.g., `open`, `unlink`, and `mount`) have path names (i.e., file names) as a parameter. Most of these calls must fetch the i-node for the named file before they can start working on the call itself. How a path name is converted to an i-node is a subject we will now look at in detail. We already saw the general outline in Fig. 5-16.

The parsing of path names is done in the file *path.c.* The first procedure, *eat_path* (line 26327), accepts a pointer to a path name, parses it, arranges for its i-node to be loaded into memory, and returns a pointer to the i-node. It does its work by calling *last_dir* to get the i-node to the final directory and then calling *advance* to get the final component of the path. If the search fails, for example, because one of the directories along the path does not exist, or exists but is protected against being searched, *NIL_INODE* is returned instead of a pointer to the i-node.

Path names may be absolute or relative and may have arbitrarily many components, separated by slashes. These issues are dealt with by *last_dir*, which begins by examining the first character of the path name to see if it is an absolute path or a relative one (line 26371). For absolute paths, *rip* is set to point to the root i-node; for relative ones, it is set to point to the i-node for the current working directory.

At this point, *last_dir* has the path name and a pointer to the i-node of the directory to look up the first component in. It enters a loop on line 26382 now, parsing the path name, component by component. When it gets to the end, it returns a pointer to the final directory.

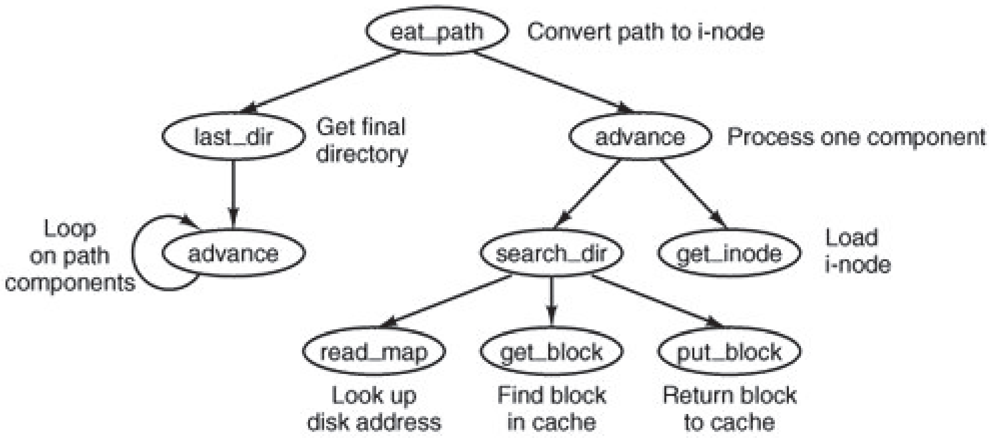
Get_name (line 26413) is a utility procedure that extracts components from strings. More interesting is *advance* (line 26454), which takes as parameters a directory pointer and a string, and looks up the string in the directory. If it finds the string, *advance* returns a pointer to its i-node. The details of transferring across mounted file systems are handled here.

Although *advance* controls the string lookup, the actual comparison of the string against the directory entries is done in *search_dir* (line 26535), which is the only place in the file system where directory files are actually examined. It contains two nested loops, one to loop over the blocks in a directory, and one to loop over the entries in a block. *Search_dir* is also used to enter and delete names from directories. Figure 5-48 shows the relationships between some of the major procedures used in looking up path names.

Figure 5-48. Some of the procedures used in looking up path names.

(This item is displayed on page 593 in the print version)

[View full size image]



Mounting File Systems

Two system calls that affect the file system as a whole are `mount` and `umount`. They allow independent file systems on different minor devices to be "glued" together to form a single, seamless naming tree. Mounting, as we saw in Fig. 5-38, is effectively achieved by reading in the root i-node and superblock of the file system to be mounted and setting two pointers in its superblock. One of them points to the i-node mounted on, and the other points to the root i-node of the mounted file system. These pointers hook the file systems together.

The setting of these pointers is done in the file *mount.c* by *do_mount* on lines 26819 and 26820. The two pages of code that precede setting the pointers are almost entirely concerned with checking for all the errors that can occur while mounting a file system, among them:

1. The special file given is not a block device.
2. The special file is a block device but is already mounted.
3. The file system to be mounted has a rotten magic number.
4. The file system to be mounted is invalid (e.g., no i-nodes).
5. The file to be mounted on does not exist or is a special file.
6. There is no room for the mounted file system's bitmaps.
7. There is no room for the mounted file system's superblock.
8. There is no room for the mounted file system's root i-node.

Perhaps it seems inappropriate to keep harping on this point, but the reality of any practical operating system is that a substantial fraction of the code is devoted to doing minor chores that are not intellectually very exciting but are crucial to making a system usable. If a user attempts to mount the wrong floppy disk by accident, say, once a month, and this leads to a crash and a corrupted file system, the user will perceive the system as being unreliable and blame the designer, not himself.

The famous inventor Thomas Edison once made a remark that is relevant here. He said that "genius" is 1 percent inspiration and 99 percent perspiration. The difference between a good system and a mediocre one is not the brilliance of the former's scheduling algorithm, but its attention to getting all the details right.

Unmounting a file system is easier than mounting onethere are fewer things that can go wrong. *Do_umount* (line 26828) is called to start the job, which is divided into two parts. *Do_umount* itself checks that the call was made by the superuser, converts the name into a device number, and then calls *unmount* (line 26846), which completes the operation. The only real issue is making sure that no process has any open files or working directories on the file system to be removed. This check is straightforward: just scan the whole i-node table to see if any i-nodes in memory belong to the file system to be removed (other than the root i-node). If so, the `umount` call fails.

The last procedure in *mount.c* is *name_to_dev* (line 26893), which takes a special file pathname, gets its i-node, and extracts its major and minor device numbers. These are stored in the i-node itself, in the place where the first zone would normally go. This slot is available because special files do not have zones.

Linking and Unlinking Files

The next file to consider is *link.c,* which deals with linking and unlinking files. The procedure *do_link* (line 27034) is very much like *do_mount* in that nearly all of the code is concerned with error checking. Some of the possible errors that can occur in the call

```
link(file_name, link_name);
```

are listed below:

1. *File_name* does not exist or cannot be accessed.

2. *File_name* already has the maximum number of links.

3. *File_name* is a directory (only superuser can link to it).

4. *Link_name* already exists.

5. *File_name* and *link_name* are on different devices.

If no errors are present, a new directory entry is made with the string *link_name* and the i-node number of *file_name.* In the code, *name1* corresponds to *file_name* and *name2* corresponds to *link_name.* The actual entry is made by *search_dir*, called from *do_ link* on line 27086.

Files and directories are removed by unlinking them. The work of both the `unlink` and `rmdir` system calls is done by *do_unlink* (line 27104). Again, a variety of checks must be made; testing that a file exists and that a directory is not a mount point are done by the common code in *do_unlink*, and then either *remove_dir* or *unlink_file* is called, depending upon the system call being supported. We will discuss these shortly.

The other system call supported in *link.c* is `rename`. UNIX users are familiar with the *mv* shell command which ultimately uses this call; its name reflects another aspect of the call. Not only can it change the name of a file within a directory, it can also effectively move the file from one directory to another, and it can do this atomically, which prevents certain race conditions. The work is done by *do_rename* (line 27162). Many conditions must be tested before this command can be completed. Among these are:

1. The original file must exist (line 27177).
2. The old pathname must not be a directory above the new pathname in the directory tree (lines 27195 to 27212).
3. Neither *.* nor *..* is acceptable as an old or new name (lines 27217 and 27218).
4. Both parent directories must be on the same device (line 27221).
5. Both parent directories must be writable, searchable, and on a writable device (lines 27224 and 27225).
6. Neither the old nor the new name may be a directory with a file system mounted upon it.

Some other conditions must be checked if the new name already exists. Most importantly it must be possible to remove an existing file with the new name.

In the code for *do_rename* there are a few examples of design decisions that were taken to minimize the possibility of certain problems. Renaming a file to a name that already exists could fail on a full disk, even though in the end no additional space is used, if the old file were not removed first, and this is what is done at lines 27260 to 27266. The same logic is used at line 27280, removing the old file name before creating a new name in the same directory, to avoid the possibility that the directory might need to acquire an additional block. However, if the new file and the old file are to be in different directories, that concern is not relevant, and at line 27285 a new file name is created (in a different directory) before the old one is removed, because from a system integrity standpoint a crash that left two filenames pointing to an i-node would be much less serious than a crash that left an i-node not pointed to by any directory entry. The probability of running out of space during a rename operation is low, and that of a system crash even lower, but in these cases it costs nothing more to be prepared for the worst case.

The remaining functions in *link.c* support the ones that we have already discussed. In addition, the first of them, *truncate* (line 27316), is called from several other places in the file system. It steps through an i-node one zone at a time, freeing all the zones it finds, as well as the indirect blocks. *Remove_dir* (line 27375) carries out a number of additional tests to be sure the directory can be removed, and then it in turn calls *unlink_file* (line 27415). If no errors are found, the directory entry is cleared and the link count in the i-node is reduced by one.

5.7.6. Other System Calls

The last group of system calls is a mixed bag of things involving status, directories, protection, time, and other services.

Changing Directories and File Status

The file *stadir.c* contains the code for six system calls: `chdir`, `fchdir`, `chroot`, `stat`, `fstat`, and `fstatfs`. In studying *last_dir* we saw how path searches start out by looking at the first character of the path, to see if it is a slash or not. Depending on the result, a pointer is then set to the working directory or the root directory.

Changing from one working directory (or root directory) to another is just a matter of changing these two pointers within the caller's process table. These changes are made by *do_chdir* (line 27542) and *do_chroot* (line 27580). Both of them do the necessary checking and then call *change* (line 27594), which does some more tests, then calls *change_into* (line 27611) to open the new directory and replace the old one.

Do_fchdir (line 27529) supports `fchdir`, which is an alternate way of effecting the same operation as `chdir`, with the calling argument a file descriptor rather than a path. It tests for a valid descriptor, and if the descriptor is valid it calls *change_into* to do the job.

In *do_chdir* the code on lines 27552 to 27570 is not executed on `chdir` calls made by user processes. It is specifically for calls made by the process manager, to change to a user's directory for the purpose of handling `exec` calls. When a user tries to execute a file, say, *a.out* in his working directory, it is easier for the process manager to change to that directory than to try to figure out where it is.

The two system calls `stat` and `fstat` are basically the same, except for how the file is specified. The former gives a path name, whereas the latter provides the file descriptor of an open file, similar to what we saw for `chdir` and `fchdir`. The top-level procedures, *do_stat* (line 27638) and *do_fstat* (line 27658), both call *stat_inode* to do the work. Before calling *stat_inode*, *do_stat* opens the file to get its i-node. In this way, both *do_stat* and *do_fstat* pass an i-node pointer to *stat_inode*.

All *stat_inode* (line 27673) does is to extract information from the i-node and copy it into a buffer. The buffer must be explicitly copied to user space by a `sys_datacopy` kernel call on lines 27713 and 27714 because it is too large to fit in a message.

Finally, we come to *do_fstatfs* (line 27721). `Fstatfs` is not a POSIX call, although POSIX defines a similar `fstatvfs` call which returns a much bigger data structure. The MINIX 3 `fstatfs` returns only one piece of information, the block size of a file system. The prototype for the call is

```
_PROTOTYPE( int fstatfs, (int fd, struct statfs *st) );
```

The *statfs* structure it uses is simple, and can be displayed on a single line:

```
struct statfs { off_t f_bsize; /* file system block size */ };
```

These definitions are in *include/sys/statfs.h,* which is not listed in Appendix B.

Protection

The MINIX 3 protection mechanism uses the *rwx* bits. Three sets of bits are present for each file: for the owner, for his group, and for others. The bits are set by the `chmod` system call, which is carried out by *do_chmod*, in file *protect.c* (line 27824). After making a series of validity checks, the mode is changed on line 27850.

The `chown` system call is similar to `chmod` in that both of them change an internal i-node field in some file. The implementation is also similar although *do_chown* (line 27862) can be used to change the owner only by the superuser. Ordinary users can use this call to change the group of their own files.

The `umask` system call allows the user to set a mask (stored in the process table), which then masks out bits in subsequent `creat` system calls. The complete implementation would be only one statement, line 27907, except that the call must return the old mask value as its result. This additional burden triples the number of lines of code required (lines 27906 to 27908).

The `access` system call makes it possible for a process to find out if it can access a file in a specified way (e.g., for reading). It is implemented by *do_access* (line 27914), which fetches the file's i-node and calls the internal procedure, *forbidden* (line 27938), to see if the access is forbidden. *Forbidden* checks the uid and gid, as well as the information in the i-node. Depending on what it finds, it selects one of the three *rwx* groups and checks to see if the access is permitted or forbidden.

Read_only (line 27999) is a little internal procedure that tells whether the file system on which its i-node parameter is located is mounted read only or read-write. It is needed to prevent writes on file systems mounted read only.

5.7.7. The I/O Device Interface

As we have mentioned more than once, a design goal was to make MINIX 3a more robust operating system by having all device drivers run as user-space processes without direct access to kernel data structures or kernel code. The primary advantage of this approach is that a faulty device driver will not cause the entire system to crash, but there are some other implications of this approach. One is that device drivers not needed immediately upon startup can be started at any time after startup is complete. This also implies that a device driver can be stopped, restarted, or replaced by a different driver for the same device at any time while the system is running. This flexibility is subject, of course to some restrictionsyou cannot start multiple drivers for the same device. However, if the hard disk driver crashes, it can be restarted from a copy on the RAM disk.

MINIX 3 device drivers are accessed from the file system. In response to user requests for I/O the file system sends messages to the user-space device drivers. The *dmap* table has an entry for every possible major device type. It provides the mapping between the major device number and the corresponding device driver. The next two files we will consider deal with the *dmap* table. The table itself is declared in *dmap.c.* This file also supports initialization of the table and a new system call, `devctl`, which is intended to support starting, stopping, and restarting of device drivers. After that we will look at *device.c* which supports normal runtime operations on devices, such as open, `close`, `read`, `write`, and `ioctl`.

When a device is opened, closed, read, or written, *dmap* provides the name of the procedure to call to handle the operation. All of these procedures are located in the file system's address space. Many of these procedures do nothing, but some call a device driver to request actual I/O. The process number corresponding to each major device is also provided by the table.

Whenever a new major device is added to MINIX 3, a line must be added to this table telling what action, if any, is to be taken when the device is opened, closed, read, or written. As a simple example, if a tape drive is added to MINIX 3, when its special file is opened, the procedure in the table could check to see if the tape drive is already in use.

Dmap.c begins with a macro definition, *DT* (lines 28115 to 28117), which is used to initialize the *dmap* table. This macro makes it easier to add a new device driver when reconfiguring MINIX 3. Elements of the *dmap* table are defined in *include/minix/dmap.h;* each element consists of a pointer to a function to be called on an `open` or `close`, another pointer to a function to be called on a `read` or `write`, a process number (index into process table, not a PID), and a set of flags. The actual table is an array of such elements, declared on line 28132. This table is globally available within the file server. The size of the table is *NR_DEVICES*, which is 32 in the version of MINIX 3 described here, and almost twice as big as needed for the number of devices currently supported. Fortunately, the C language behavior of setting all uninitialized variables to zero will ensure that no spurious information appears in unused slots.

Following the declaration of *dmap* is a *PRIVATE* declaration of *init_dmap*. It is defined by an array of *DT* macros, one for each possible major device. Each of these macros expands to initialize an entry in the global array at compile time. A look at a few of the macros will help with understanding how they are used. *Init_dmap[1]* defines the entry for the memory driver, which is major device 1. The macro looks like this:

```
DT(1, gen_opcl, gen_io, MEM_PROC_NR, 0)
```

The memory driver is always present and is loaded with the system boot image. The "1" as first parameter means that this driver must be present. In this case, a pointer to *gen_opcl* will be entered as the function to call to open or close, and a pointer to *gen_io* will be entered to specify the function to call for reading or writing, *MEM_PROC_NR* tells which slot in the process table the memory driver uses, and "0" means no flags are set. Now look at the next entry, *init_dmap[2].* This is the entry for the floppy disk driver, and it looks like this:

```
DT(0, no_dev, 0, 0, DMAP_MUTABLE)
```

The first "0" indicates this entry is for a driver not required to be in the boot image. The default for the first pointer field specifies a call to *no_dev* on an attempt to open the device. This function returns an *ENODEV*"no such device" error to the caller. The next two zeros are also defaults: since the device cannot be opened there is no need to specify a function to call to do I/O, and a zero in the process table slot is interpreted as no process specified. The meaning of the flag *DMAP_MUTABLE* is that changes to this entry are permitted. (Note that the absence of this flag for the memory driver entry means its entry cannot be changed after initialization.) MINIX 3 can be configured with or without a floppy disk driver in the boot image. If the floppy disk driver is in the boot image and it is specified by a *label=FLOPPY* boot parameter to be the default disk device, this entry will be changed when the file system starts. If the floppy driver is not in the boot image, or if it is in the image but is not specified to be the default disk device, this field will not be changed when FS starts. However, it is still possible for the floppy driver to be activated later. Typically this is done by the */etc/rc* script run when *init* is run.

Do_devctl (line 28157) is the first function executed to service a `devctl` call. The current version is very simple, it recognizes two requests, *DEV_MAP* and *DEV_UNMAP*, and the latter returns a *ENOSYS* error, which means "function not implemented." Obviously, this is a stopgap. In the case of *DEV_MAP* the next function, *map_driver* is called.

It might be helpful to describe how the `devctl` call is used, and plans for its use in the future. A server process, the **reincarnation server** (**RS**) is used in MINIX 3 to support starting user-space servers and drivers after the operating system is up and running. The interface to the reincarnation server is the *service* utility, and examples of its use can be seen in */etc/rc.* An example is

```
service up /sbin/floppy dev /dev/fd0
```

This action results in the reincarnation server making a `devctl` call to start the binary */sbin/floppy* as the device driver for the device special file */dev/fd0.* To do this, RS `execs` the specified binary, but sets a flag that inhibits it from running until it has been transformed into a system process. Once the process is in memory and its slot number in the process table is known, the major device number for the specified device is determined. This information is then included in a message to the file server that requested the `devctl` *DEV_MAP* operation. This is the most important part of the reincarnation server's job from the point of view of initializing the I/O interface. For the sake of completeness we will also mention that to complete initialization of the device driver, RS also makes a `sys_privctl` call to have the system task initialize the driver process's *priv* table entry and allow it to execute. Recall from Chapter 2 that a dedicated *priv* table slot is what makes an otherwise ordinary user-space process into a system process.

The reincarnation server is new, and in the release of MINIX 3 described here it is still rudimentary. Plans for future releases of MINIX 3 include a more powerful reincarnation server that will be able to stop and restart drivers in addition to starting them. It will also be able to monitor drivers and restart them automatically if problems develop. Check the Web site (*www.minix3.org*) and the newsgroup (*comp.os.minix)* for the current status.

Continuing with *dmap.c,* the function *map_driver* begins on line 28178. Its operation is straightforward. If the *DMAP_MUTABLE* flag is set for the entry in the *dmap* table, appropriate values are written into each entry. Three different variants of the function for handling opening and closing of the device are available; one is selected by a *style* parameter passed in the message from RS to the file system (lines 28204 to 28206). Notice that *dmap_flags* is not altered. If the entry was marked *DMAP_MUTABLE* originally it retains this status after the `devctl` call.

The third function in *dmap.c* is *build_map*. This is called by *fs_init* when the file system is first started, before it enters its main loop. The first thing done is to loop over all of the entries in the local *init_dmap* table and copy the expanded macros to the global *dmap* table for each entry that does not have *no_dev* specified as the *dmap_opcl* member. This correctly initializes these entries. Otherwise the default values for an uninitialized driver are set in place in *dmap*. The rest of *build_map* is more interesting. A boot image can be built with multiple disk device drivers. By default *at_wini*, *bios_wini*, and *floppy* drivers are added to the boot image by the *Makefile* in the *src/tools/.* A label is added to each of these, and a *label=* item in the boot parameters determines which one will actually be loaded in the image and activated as the default disk driver. The *env_get_param* calls on line 28248 and line 28250 use library routines that ultimately use the `sys getinfo` kernel call to get the *label* and *controller* boot parameter strings. Finally, *build_map* is called on line 28267 to modify the entry in *dmap* that corresponds to the boot device. The key thing here is setting the process number to *DRVR_PROC_NR*, which happens to be slot 6 in the process table. This slot is magic; the driver in this slot is the default driver.

Now we come to the file *device.c,* which contains the procedures needed for device I/O at run time.

The first one is *dev_open* (line 28334). It is called by other parts of the file system, most often from *common_open* in *main.c* when a `open` operation is determined to be accessing a device special file, but also from *load_ram* and *do_mount*. Its operation is typical of several procedures we will see here. It determines the major device number, verifies that it is valid, and then uses it to set a pointer to an entry in the *dmap* table, and then makes a call to the function pointed to in that entry, at line 28349:

```
r = (*dp->dmap_opcl)(DEV_OPEN, dev, proc, flags)
```

In the case of a disk drive, the function called will be *gen_opcl*, in the case of a terminal device it will be *tty_opcl*. If a *SUSPEND* return code is received there is a serious problem; an open call should not fail this way.

The next call, *dev_close* (line 28357) is simpler. It is not expected that a call will be made to an invalid device, and no harm is done if a close operation fails, so the code is shorter than this text describing it, just one line that will end up calling the same **_opcl* procedure as *dev_open* called when the device was opened.

When the file system receives a notification message from a device driver *dev_status* (line 28366) is called. A notification means an event has occurred, and this function is responsible for finding out what kind of event and initiating appropriate action. The origin of the notification is specified as a process number, so the first step is to search through the *dmap* table to find an entry that corresponds to the notifying process (lines 18371 to 18373). It is possible the notification could have been bogus, so it is not an error if no corresponding entry is found and *dev_status* returns without finding a match. If a match is found, the loop on lines 28378 to 28398 is entered. On each iteration a message is sent to the driver process requesting its status. Three possible reply types are expected. A *DEV_REVIVE* message may be received if the process that originally requested I/O was previously suspended. In this case *revive* (in *pipe.c,* line 26146) is called. A *DEV_IO_READY* message may be received if a `select` call has been made on the device. Finally, a *DEV_NO_STATUS* message may be received, and is, in fact expected, but possibly not until one or both of the first two message types are received. For this reason, the *get_more* variable is used to cause the loop to repeat until the *DEV_NO_STATUS* message is received.

When actual device I/O is needed, *dev_io* (line 28406) is called from *read_write* (line 25124) to handle character special files, and from *rw_block* (line 22661) to handle block special files. It builds a standard message (see Fig. 3-17) and sends it to the specified device driver by calling either *gen_io* or *ctty_io* as specified in the *dp->dmap_driver* field of the *dmap* table. While *dev_io* is waiting for a reply from the driver, the file system waits. It has no internal multiprogramming. Usually, these waits are quite short though (e.g., 50 msec). But it is possible no data will be availablethis is especially likely if the data was requested from a terminal device. In that case the reply message may indicate *SUSPEND*, to temporarily suspend the calling application but let the file system continue.

The procedure *gen_opcl* (line 28455) is called for disk devices, whether floppy disks, hard disks, or memory-based devices. A message is constructed, and, as with reading and writing, the *dmap* table is used to determine whether *gen_io* or *ctty_io* will be used to send the message to the driver process for the device. *Gen_opcl* is also used to close the same devices.

To open a terminal device *tty_opcl* (line 28482) is called. It calls *gen_opcl* after possibly modifying the flags, and if the call made the tty the controlling tty for the active process this is recorded in the process table *fp_tty* entry for that process.

The device */dev/tty* is a fiction which does not correspond to any particular device. This is a magic designation that an interactive user can use to refer to his own terminal, no matter which physical terminal is actually in use. To open or close */dev/tty,* a call is made to *ctty_opcl* (line 28518). It determines whether the *fp_tty* process table entry for the current process has indeed been modified by a previous *ctty_opcl* call to indicate a controlling tty.

The `setsid` system call requires some work by the file system, and this is performed by *do_setsid* (line 28534). It modifies the process table entry for the current process to record that the process is a session leader and has no controlling process.

One system call, `ioctl`, is handled primarily in *device.c.* This call has been put here because it is closely tied to the device driver interface. When an `ioctl` is done, *do_ioctl* (line 28554) is called to build a message and send it to the proper device driver.

To control terminal devices one of the functions declared in *include/termios.h* should be used in programs written to be POSIX compliant. The C library will translate such functions into `ioctl` calls. For devices other than terminals `ioctl` is used for many operations, many of which were described in Chap. 3.

The next function, *gen_io* (line 28575), is the real workhorse of this file. Whether the operation on a device is an `open` or a `close`, a `read` or a `write`, or an `ioctl` this function is called to complete the work. Since */dev/tty* is not a physical device, when a message that refers to it must be sent, the next function, *ctty_io* (line 28652), finds the correct major and minor device and substitutes them into the message before passing the message on. The call is made using the *dmap* entry for the physical device that is actually in use. As MINIX 3 is currently configured a call to *gen_io* will result.

The function *no_dev* (line 28677), is called from slots in the table for which a device does not exist, for example when a network device is referenced on a machine with no network support. It returns an *ENODEV* status. It prevents crashes when nonexistent devices are accessed.

The last function in *device.c* is *clone_opcl* (line 28691). Some devices need special processing upon open. Such a device is "cloned," that is, on a successful open it is replaced by a new device with a new unique minor device number. In MINIX 3 as described here this capability is not used. However, it is used when networking is enabled. A device that needs this will, of course, have an entry in the *dmap* table that specifies *clone_opcl* in the *dmap_opcl* field. This is accomplished by a call from the reincarnation server that specifies *STYLE_CLONE*. When *clone_opcl* opens a device the operation starts in exactly the same way as *gen_opcl*, but on the return a new minor device number may be returned in the *REP_STATUS* field of the reply message. If so, a temporary file is created if it is possible to allocate a new i-node. A visible directory entry is not created. That is not necessary, since the file is already open.

Time

Associated with each file are three 32-bit numbers relating to time. Two of these record the times when the file was last accessed and last modified. The third records when the status of the i-node itself was last changed. This time will change for almost every access to a file except a `read` or `exec`. These times are kept in the i-node. With the `utime` system call, the access and modification times can be set by the owner of the file or the superuser. The procedure *do_utime* (line 28818) in file *time.c* performs the system call by fetching the i-node and storing the time in it. At line 28848 the flags that indicate a time update is required are reset, so the system will not make an expensive and redundant call to *clock_time*.

As we saw in the previous chapter, the real time is determined by adding the time since the system was started (maintained by the clock task) to the real time when startup occurred. The `stime` system call returns the real time. Most of its work is done by the process manager, but the file system also maintains a record of the startup time in a global variable, *boottime*. The process manager sends a message to the file system whenever a `stime` call is made. The file system's *do_stime* (line 28859) updates *boottime* from this message.

5.7.8. Additional System Call Support

There are a number of files that are not listed in Appendix B, but which are required to compile a working system. In this section we will review some files that support additional system calls. In the next section we will mention files and functions that provide more general support for the file system.

The file *misc.c* contains procedures for a few system and kernel calls that do not fit in anywhere else.

Do_getsysinfo is an interface to the `sys_datacopy` kernel call. It is meant to support the information server (IS) for debugging purposes. It allows IS to request a copy of file system data structures so it can display them to the user.

The `dup` system call duplicates a file descriptor. In other words, it creates a new file descriptor that points to the same file as its argument. The call has a variant `dup2`. Both versions of the call are handled by *do_dup* This function is included in MINIX 3 to support old binary programs. Both of these calls are obsolete. The current version of the MINIX 3 C library will invoke the `fcntl` system call when either of these are encountered in a C source file.

`Fcntl`, handled by *do_fcntl* is the preferred way to request operations on an open file. Services are requested using POSIX-defined flags described in Fig. 5-49. The call is invoked with a file descriptor, a request code, and additional arguments as necessary for the particular request. For instance, the equivalent of the old call

Figure 5-49. The POSIX request parameters for the FCNTL system call.

Operation	Meaning
F_DUPFD	Duplicate a file descriptor
F_GETFD	Get the close-on-exec flag

Operation	Meaning
F_SETFD	Set the close-on-exec flag
F_GETFL	Get file status flags
F_SETFL	Set file status flags
F_GETLK	Get lock status of a file
F_SETLK	Set read/write lock on a file
F_SETLKW	Set write lock on a file

```
dup2(fd, fd2);
```

would be

```
fcntl(fd, F_DUPFD, fd2);
```

Several of these requests set or read a flag; the code consists of just a few lines. For instance, the *F_SETFD* request sets a bit that forces closing of a file when its owner process does an `exec`. The *F_GETFD* request is used to determine whether a file must be closed when an `exec` call is made. The *F_SETFL* and *F_GETFL* requests permit setting flags to indicate a particular file is available in nonblocking mode or for append operations.

Do_fcntl handles file locking, also. A call with the *F_GETLK*, *F_SETLK*, or *F_SETLKW* command specified is translated into a call to *lock_op*, discussed in an earlier section.

The next system call is `sync`, which copies all blocks and i-nodes that have been modified since being loaded back to the disk. The call is processed by *do_sync*. It simply searches through all the tables looking for dirty entries. The i-nodes must be processed first, since *rw_inode* leaves its results in the block cache. After all dirty i-nodes are written to the block cache, then all dirty blocks are written to the disk.

The system calls `fork`, `exec`, `exit`, and `set` are really process manager calls, but the results have to be posted here as well. When a process forks, it is essential that the kernel, process manager, and file system all know about it. These "system calls" do not come from user processes, but from the process manager. *Do_fork*, *do_exit*, and *do_set* record the relevant information in the file system's part of the process table. *Do_exec* searches for and closes (using *do_close*) any files marked to be closed-on-exec.

The last function in *misc.c* is not really a system call but is handled like one. *Do_revive* is called when a device driver that was previously unable to complete work that the file system had requested, such as providing input data for a user process, has now completed the work. The file system then revives the process and sends it the reply message.

One system call merits a header file as well as a C source file to support it. *Select.h* and *select.c* provide support for the `select` system call. `Select` is used when a single process has to do deal with multiple I/O streams, as, for instance, a communications or network program. Describing it in detail is beyond the scope of this book.

5.7.9. File System Utilities

The file system contains a few general purpose utility procedures that are used in various places. They are collected together in the file *utility.c*.

Clock_time sends messages to the system task to find out what the current real time is.

Fetch_name is needed because many system calls have a file name as parameter. If the file name is short, it is included in the message from the user to the file system. If it is long, a pointer to the name in user space is put in the message. *Fetch_name* checks for both cases, and either way, gets the name.

Two functions here handle general classes of errors. *No_sys* is the error handler that is called when the file system receives a system call that is not one of its calls. *Panic* prints a message and tells the kernel to throw in the towel when something catastrophic happens. Similar functions can be found in *pm/utility.c* in the process manager's source directory.

The last two functions, *conv2* and *conv4*, exist to help MINIX 3 deal with the problem of differences in byte order between different CPU families. These routines are called when reading from or writing to a disk data structure, such as an i-node or bitmap. The byte order in the system that created the disk is recorded in the superblock. If it is different from the order used by the local processor the order will be swapped. The rest of the file system does not need to know anything about the byte order on the disk.

Finally, there are two other files that provide specialized utility services to the file manager. The file system can ask the system task to set an alarm for it, but if it needs more than one timer it can maintain its own linked list of timers, similar to what we saw for the process manager in the previous chapter. The file *timers.c* provides this support for the file system. Finally, MINIX 3 implements a unique way of using a CD-ROM that hides a simulated MINIX 3 disk with several partitions on a CD-ROM, and allows booting a live MINIX 3 system from the CD-ROM. The MINIX 3 files are not visible to operating systems that support only standard CD-ROM file formats. The file *cdprobe.c* is used at boot time to locate a CD-ROM device and the files on it needed to start MINIX 3.

5.7.10. Other MINIX 3 Components

The process manager discussed in the previous chapter and the file system discussed in this chapter are user-space servers which provide support that would be integrated into a monolithic kernel in an operating system of conventional design. These are not the only server processes in a MINIX 3 system, however. There are other user-space processes that have system privileges and should be considered part of the operating system. We do not have enough space in this book to discuss their internals, but we should at least mention them here.

One we have already mentioned in this chapter. This is the reincarnation server, RS, which can start an ordinary process and turn it into a system process. It is used in the current version of MINIX 3 to launch device drivers that are not part of the system boot image. In future releases it will also be able to stop and restart drivers, and, indeed, to monitor drivers and stop and restart them automatically if they seem to be malfunctioning. The source code for the reincarnation server is in the *src/servers/rs/* directory.

Another server that has been mentioned in passing is the information server, IS. It is used to generate the debugging dumps that can be triggered by pressing the function keys on a PC-style

PREV NEXT

5.8. Summary

When seen from the outside, a file system is a collection of files and directories, plus operations on them. Files can be read and written, directories can be created and destroyed, and files can be moved from directory to directory. Most modern file systems support a hierarchical directory system, in which directories may have subdirectories ad infinitum.

When seen from the inside, a file system looks quite different. The file system-designers have to be concerned with how storage is allocated, and how the system keeps track of which block goes with which file. We have also seen how different systems have different directory structures. File system reliability and performance are also important issues.

Security and protection are of vital concern to both the system users and system designers. We discussed some security flaws in older systems, and generic problems that many systems have. We also looked at authentication, with and without passwords, access control lists, and capabilities, as well as a matrix model for thinking about protection.

Finally, we studied the MINIX 3 file system in detail. It is large but not very complicated. It accepts requests for work from user processes, indexes into a table of procedure pointers, and calls that procedure to carry out the requested system call. Due to its modular structure and position outside the kernel, it can be removed from MINIX 3 and used as a free-standing network file server with only minor modifications.

Internally, MINIX 3 buffers data in a block cache and attempts to read ahead when making sequential access to file. If the cache is made large enough, most program text will be found to be already in memory during operations that repeatedly access a particular set of programs, such as a compilation.

Problems

1. NTFS uses Unicode for naming files. Unicode supports 16-bit characters. Give an advantage of Unicode file naming over ASCII file naming.

2. Some files begin with a magic number. Of what use is this?

3. Fig. 5-4 lists some file attributes. Not listed in this table is parity. Would that be a useful file attribute? If so, how might it be used?

4. Give 5 different path names for the file */etc/passwd.* (*Hint:* think about the directory entries "." and "..".)

5. Systems that support sequential files always have an operation to rewind files. Do systems that support random access files need this too?

6. Some operating systems provide a system call `rename` to give a file a new name. Is there any difference at all between using this call to rename a file, and just copying the file to a new file with the new name, followed by deleting the old one?

7. Consider the directory tree of Fig. 5-7. If */usr/jim/* is the working directory, what is the absolute path name for the file whose relative path name is *../ast/x?*

8. Consider the following proposal. Instead of having a single root for the file system, give each user a personal root. Does that make the system more flexible? Why or why not?

9. The UNIX file system has a call `chroot` that changes the root to a given directory. Does this have any security implications? If so, what are they?

10. The UNIX system has a call to read a directory entry. Since directories are just files, why is it necessary to have a special call? Can users not just read the raw directories themselves?

11. A standard PC can hold only four operating systems at once. Is there any way to increase this limit? What consequences would your proposal have?

12. Contiguous allocation of files leads to disk fragmentation, as mentioned in the text. Is this internal fragmentation or external fragmentation? Make an analogy with something discussed in the previous chapter.

13. Figure 5-10 shows the structure of the original FAT file system used on MS-DOS. Originally this file system had only 4096 blocks, so a table with 4096 (12-bit) entries was enough. If that scheme were to be directly extended to file systems with 2^{32} blocks, how much space would the FAT occupy?

14. An operating system only supports a single directory but allows that directory to have arbitrarily many files with arbitrarily long file names. Can something approximating a hierarchical file system be simulated? How?

15. Free disk space can be kept track of using a free list or a bitmap. Disk addresses require *D* bits. For a disk with *B* blocks, *F* of which are free, state the condition under which the free list uses less space than the bitmap. For *D* having the value 16 bits, express your answer as a percentage of the disk space that must be free.

16. It has been suggested that the first part of each UNIX file be kept in the same disk block as its i-node. What good would this do?

17. The performance of a file system depends upon the cache hit rate (fraction of blocks found in the cache). If it takes 1 msec to satisfy a request from the cache, but 40 msec to satisfy a request if a disk read is needed, give a formula for the mean time required to satisfy a request if the hit rate is *h*. Plot this function for values of *h* from 0 to 1.0.

18. What is the difference between a hard link and a symbolic link? Give an advantage of each one.

19. Name three pitfalls to watch out for when backing up a file system.

20. A disk has 4000 cylinders, each with 8 tracks of 512 blocks. A seek takes 1 msec per cylinder moved. If no attempt is made to put the blocks of a file close to each other, two blocks that are logically consecutive (i.e., follow one another in the file) will require an average seek, which takes 5 msec. If, however, the operating system makes an attempt to cluster related blocks, the mean interblock distance can be reduced to 2 cylinders and the seek time reduced to 100 microsec. How long does it take to read a 100 block file in both cases, if the rotational latency is 10 msec and the transfer time is 20 microsec per block?

21. Would compacting disk storage periodically be of any conceivable value? Explain.

22. What is the difference between a virus and a worm? How do they each reproduce?

23. After getting your degree, you apply for a job as director of a large university computer center that has just put its ancient operating system out to pasture and switched over to UNIX. You get the job. Fifteen minutes after starting work, your assistant bursts into your office screaming: "Some students discovered the algorithm we use for encrypting passwords and posted it on the Internet." What should you do?

24. Two computer science students, Carolyn and Elinor, are having a discussion about i-nodes. Carolyn maintains that memories have gotten so large and so cheap that when a file is opened, it is simpler and faster just to fetch a new copy of the i-node into the i-node table, rather than search the entire table to see if it is already there. Elinor disagrees. Who is right?

25. The Morris-Thompson protection scheme with the *n*-bit random numbers was designed to make it difficult for an intruder to discover a large number of passwords by encrypting common strings in advance. Does the scheme also offer protection against a student user who is trying to guess the superuser password on his machine?

26. A computer science department has a large collection of UNIX machines on its local network. Users on any machine can issue a command of the form

```
machine4 who
```

and have it executed on *machine4*, without having the user log in on the remote machine. This feature is implemented by having the user's kernel send the command and his uid to the remote machine. Is this scheme secure if the kernels are all trustworthy (e.g., large timeshared minicomputers with protection hardware)? What if some of the machines are students' personal computers, with no protection hardware?

27. When a file is removed, its blocks are generally put back on the free list, but they are not erased. Do you think it would be a good idea to have the operating system erase each block before releasing it? Consider both security and performance factors in your answer, and explain the effect of each.

28. Three different protection mechanisms that we have discussed are capabilities, access control lists, and the UNIX *rwx* bits. For each of the following protection problems, tell which of these mechanisms can be used.

(a) Ken wants his files readable by everyone except his office mate.

(b) Mitch and Steve want to share some secret files.

(c) Linda wants some of her files to be public.

For UNIX, assume that groups are categories such as faculty, students, secretaries, etc.

29. Can the Trojan horse attack work in a system protected by capabilities?

30. The size of the *filp* table is currently defined as a constant, *NR_FILPS*, in *fs/const.h.* In order to accommodate more users on a networked system you want to increase *NR_PROCS* in *include/minix/config.h.* How should *NR_FILPS* be defined as a function of *NR_PROCS*?

31. Suppose that a technological breakthrough occurs, and that nonvolatile RAM, which retains its contents reliably following a power failure, becomes available with no price or performance disadvantage over conventional RAM. What aspects of file system design would be affected by this development?

32. Symbolic links are files that point to other files or directories indirectly. Unlike ordinary links such as those currently implemented in MINIX 3, a symbolic link has its own i-node, which points to a data block. The data block contains the path to the file being linked to, and the i-node makes it possible for the link to have different ownership and permissions from the file linked to. A symbolic link and the file or directory to which it points can be located on different devices. Symbolic links are not part of MINIX 3. Implement symbolic links for MINIX 3.

33. Although the current limit to a MINIX 3 file size is determined by the 32-file pointer, in the future, with 64-bit file pointers, files larger than 2^{32} 1 bytes may be allowed, in which case triple indirect blocks may be needed. Modify FS to add triple indirect blocks.

34. Show if setting the (now-unused) ROBUST flag might make the file system more or less robust in the face of a crash. Whether this is the case in the current version of MINIX 3 has not been researched, so it may be either way. Take a good look at what happens when a modified block is evicted from the cache. Take into account that a modified data block may be accompanied by a modified i-node and bitmap.

35. Design a mechanism to add support for a "foreign" file system, so that one could, for instance, mount an MS-DOS file system on a directory in the MINIX 3 file system.

36. Write a pair of programs, in C or as shell scripts, to send and receive a message by a covert channel on a MINIX 3 system. *Hint*: A permission bit can be seen even when a file is otherwise inaccessible, and the *sleep* command or system call is guaranteed to delay for a fixed time, set by its argument. Measure the data rate on an idle system. Then create an artificially heavy load by starting up numerous different background processes and measure the data rate again.

37. Implement immediate files in MINIX 3, that is small files actually stored in the i-node itself, thus saving a disk access to retrieve them.

PREV NEXT

6. Reading List and Bibliography

In the previous five chapters we have touched upon a variety of topics. This chapter is intended as an aid to readers interested in pursuing their study of operating systems further. Section 6.1 is a list of suggested readings. Section 6.2 is an alphabetical bibliography of all books and articles cited in this book.

In addition to the references given below, the *Proceedings of the n-th ACM Symposium on Operating Systems Principles* (ACM) held every other year and the *Proceedings of the n-th International Conference on Distributed Computing Systems* (IEEE) held every year are good places to look for recent papers on operating systems. So is the USENIX *Symposium on Operating Systems Design and Implementation*. Furthermore, *ACM Transactions on Computer Systems* and *Operating Systems Review* are two journals that often have relevant articles.