

**495**

USPTO PATENT FULL-TEXT AND IMAGE DATABASE

<a href="#">Home</a>	<a href="#">Quick</a>	<a href="#">Advanced</a>	<a href="#">Pat Num</a>	<a href="#">Help</a>
<a href="#">Hit List</a>	<a href="#">Previous</a>	<a href="#">Next</a>	<a href="#">Bottom</a>	
<a href="#">View Cart</a>		<a href="#">Add to Cart</a>		
<a href="#">Images</a>				

( 21 of 27 )

United States Patent  
Johnson , et al.

5,175,852  
December 29, 1992

Distributed file access structure lock

**Abstract**

A distributed file management system (DFS) with a plurality of nodes and a plurality of files is disclosed. The DFS uses the *UNIX* operating system tree structure employing inodes (data structures containing the administrative information of each file) to manage the local files and surrogate inodes (s.sub.-- inode) to manage access to files existing on another node. In addition, the DFS uses a file access structure lock (fas.sub.-- lock) to manage multiple requests to a single file. The primary reason for the addition of the fas.sub.-- lock for each file is to avoid the problem of deadlocks. The inodes and s.sub.-- inodes use the fas.sub.-- lock to synchronize their accesses to a file and avoid a deadlock situation where both s.sub.-- inode and inode await the use of a file that is locked by the other.

Inventors: **Johnson; Donavon W.** (Georgetown, TX); **Shaheen-Gouda; Amal A.** (Austin, TX);  
**Smith; Todd A.** (Austin, TX)

Assignee: **International Business Machines Corporation** (Armonk, NY)

Appl. No.: **418750**

Filed: **October 4, 1989**

**Current U.S. Class:**

**707/8; 707/9**

**Intern'l Class:**

**G06F 015/16; G06F 013/14**

**Field of Search:**

**364/200 MS File, 900 MS File**

**References Cited [Referenced By]****U.S. Patent Documents**

<u>4104718</u>	Aug., 1978	Poublan et al.	364/200.
<u>4224664</u>	Sep., 1980	Trinchieri	364/200.
<u>4414624</u>	Nov., 1983	Summer, Jr. et al.	364/200.
<u>4527237</u>	Jul., 1985	Frieder et al.	364/200.
<u>4819159</u>	Apr., 1989	Shipley et al.	364/200.

### Other References

Rifkin et al., "RFS Architectual Overview", pp. 1-12.

*Primary Examiner:* Clark; David L.

*Assistant Examiner:* Von Buhr; Maria N.

*Attorney, Agent or Firm:* Whitham & Marhoefer

---

### Parent Case Text

---

This is a continuation of application Ser. No. 014,900 filed Feb. 13, 1987, now abandoned.

---

### Claims

---

Having thus described our invention, what we claim as new and desire to secure by Letters Patent is set forth in the following claims:

1. A distributed system having data in a file residing at a server data processing system, said file data being accessible by a plurality of processes in at least one client data processing system and said server data processing system, said server data processing system and said at least one client data processing system connected by a communication means, said distributed system comprising:

a first lock means, in said server data processing system, for serializing access to the data in the file by processes at the server data processing system;

a second lock means, in said client data processing system, for serializing access to data in a cache in the client data processing system corresponding to said data in said file by processes at the client data processing system;

a third lock means in said server data processing system for serializing access to a file access structure list containing descriptions of locks granted by said first lock means at said server data processing system; and

means for using said third lock means, to lock said file access structure list in said server data processing system, instead of using said first lock means to lock said data in said file, by an operation which is capable of causing a remote procedure call to be generated between the server data processing system and the client data processing system, said remote procedure call requiring access to the data in the cache at the client data processing system and to the data in the file at the server data processing system, thereby avoiding a lock simultaneously existing on both said first lock means and said second lock means during a remote procedure call.

2. The distributed system of claim 1 wherein said third lock means serializes use of said first lock means when one of said plurality of processes closes said file.

3. The distributed system of claim 1 wherein said third lock means serializes use of said first lock means

when one of said plurality of processes opens said file.

4. The distributed system of claim 1 wherein said third lock means serializes use of said first lock means when a list of processes having current access to said file is interrogated.

5. The distributed system of claim 1 wherein said third lock means serializes use of said first lock means during a write to said file when said file is open for write access in more than one of a plurality of data processing systems.

6. The distributed system of claim 1 wherein said first lock means for serializing access to said data in said file at said server data processing system is locked by a process during an operation to said file in said server data processing system.

7. The distributed system of claim 1 wherein said first lock means for serializing access to said data in said file at said server data processing system is locked by a process during a read operation to said data in said file at said server data processing system.

8. The distributed system of claim 1 wherein said first lock means for serializing access to said data in said file at said server data processing system is locked by a process during a write operation to said data in said file at said server data processing system.

9. The distributed system of claim 1 wherein said first lock means for serializing access to said data in said file at said server data processing system is unlocked by a process before said remote procedure call is sent to said client data processing system from said server data processing system if the remote procedure call requires a lock on said second lock means for serializing access to said data in said cache in said client data processing system.

10. The distributed system of claim 1 wherein said first lock means for serializing access to said data in said file is unlocked by a process at said server data processing system when said server data processing system originates said remote procedure call to the client data processing system, thereby allowing the server data processing system to accept read and write requests after said remote procedure call is sent.

11. The distributed system of claim 1 wherein said second lock means for serializing access to said data in said cache at said client data processing system is locked by a process during an operation at said client data processing system requiring access to said data in said cache in said client data processing system.

12. The distributed system of claim 1 wherein said second lock means for serializing access to said data in said cache at said client data processing system is unlocked by a process if an operation at said client data processing system requires a lock on said third lock means.

13. The distributed system of claim 1 wherein said second lock means for serializing access to said data in said cache at said client data processing system is locked by a process during an operation at said client data processing system requiring access to said data in said cache until said remote procedure call is sent from client data processing system to said server data processing system.

14. The distributed system of claim 1 wherein said second lock means for serializing access to said data in said cache is unlocked at one of a plurality of client data processing systems by a process when said remote procedure call is sent to said server data processing system from said one of said plurality of client data processing systems.

15. A method, in a data processing system, of preventing a deadlock between a first lock that serializes access to data in a file at a server data processing system and a second lock that serializes file access within a client process in a client data processing system, said method comprising:

locking a third lock for serializing access to a list of data corresponding to at least said first lock, representing client data processing systems having current access to said file;

locking said first lock for serializing access to said data in said file at said server data processing system; and

unlocking said first lock at said server data processing system before sending, by said server data processing system, a remote procedure call to said client data processing system to lock said second lock, by an operation executing at said client data processing system requiring access to the file.

16. A method, in a data processing system, of preventing a deadlock between a first lock that serializes access to data in a file at a server data processing system and a second lock that serializes file access in a cache in a client data processing system, said method comprising:

locking said second lock at said client data processing system by an operation executing at said client data processing system and accessing data in said cache corresponding to said file;

unlocking said second lock by an operation at said client data processing system before a remote procedure call request is sent from said client data processing system to said server data processing system;

locking a third lock in said server data processing system, by said remote procedure call request, for serializing access to a list of client data processing systems having current access to said file;

locking said first lock by an operation executing at said server data processing system and requiring access to said file at said server data processing system; and

unlocking said first lock by said server data processing system while maintaining said third lock before sending said remote procedure call to said client data processing system if said remote procedure call requires a lock on said second lock, thereby allowing said server data processing system to accept read and write operations requiring a lock on said first lock after said remote procedure call is sent.

17. A method, in a data processing system, of preventing a deadlock between a first lock that serializes access to data in a file at a server data processing system and a second lock that serializes access to data, corresponding to said file, in a cache in a client data processing system, said method comprising:

locking said first lock by an operation executing at said server data processing system and requiring access to said file at said server data processing system;

locking said second lock by an operation executing at the client data processing system requiring access to said data in said cache;

unlocking said second lock by an operation at said client data processing system if said operation generates a remote procedure call from said client data processing system to said server data processing system; and

a locking a third lock for serializing access to a list of files representing client accesses to said files in

said server data processing system by a second operation executing in said server data processing system in response to said remote procedure call received by said server data processing system from said client data processing system, thereby avoiding locking of both said first lock and said second lock during said remote procedure call.

18. A method of preventing deadlocks in a distributed data processing system of the type having at least one server data processing system having at least one file physically residing at the server data processing system and having a first data structure representing said file at said server data processing system, and at least one client data processing system having access to data from said file by a communications link between said server data processing system and said client data processing system, said client data processing system having access to said data from said file from a cache at the client data processing system, said client data processing system further having a second data structure representing said cached data at said client data processing system, said method of preventing deadlocks in a distributed processing system comprising:

locking said first data structure during execution of an operation from a process, at said server data processing system, requiring access to at least a portion of said file at said server data processing system;

locking said second data structure during execution of an operation from a process, at said client data processing system, requiring access to at least a portion of said file in the cache at said client data processing system; and

unlocking said second data structure before controlling access to at least a portion of said file residing at said server data processing system by locking a third data structure, in said server data processing system, instead of locking said first data structure by said at least one client data processing system, thereby maintaining the control of the locking of said first data structure by said server data processing system and eliminating said first data structure as a critical locking resource between said client data processing system and said server data processing system.

---

### *Description*

---

#### CROSS REFERENCE TO RELATED APPLICATIONS

This application is related in subject matter to the following applications filed concurrently herewith and assigned to a common assignee:

Application Ser. No. 07/014,899 filed by A. Chang, G. H. Neuman, A. A. Shaheen-Gouda, and T. A. Smith for A System And Method For Using Cached Data At A Local Node After Re-opening A File At A Remote Node In A Distributed Networking Environment; now U.S. Pat. No. 4,897,781, issued Jan. 30, 1990.

Application Ser. No. 07/014,884 filed by D. W. Johnson, L. W. Henson, A. A. Shaheen-Gouda, and T. A. Smith for Negotiating Communication Conventions Between Nodes in a Network; now abandoned.

Application Ser. No. 07/014,897 filed by D. W. Johnson, G. H. Neuman, C. H. Sauer, A. A. Shaheen-Gouda, and T. A. Smith for A System And Method For Accessing Remote Files In A Distributed Networking Environment; now U.S. Pat. No. 4,887,204, issued Dec. 12, 1989.

Application Ser. No. 07/014,891 filed by L. W. Henson, A. A. Shaheen-Gouda, and T. A. Smith for Distributed File and Record Locking; now abandoned.

Application Ser. No. 07/014,892 filed by D. W. Johnson, L. K. Loucks, C. H. Sauer, and T. A. Smith for Single System Image; Uniquely Defining an Environment For Each User In a Data Processing System; now abandoned.

Application Ser. No. 07/014,888 filed by D. W. Johnson, L. K. Loucks, A. A. Shaheen-Gouda for Interprocess Communication Queue Location Transparency now U.S. Pat. No. 5,133,053, issued Jul. 21, 1992.

Application Ser. No. 07/014,889 filed by D. W. Johnson, A. A. Shaheen-Gouda, and T. A. Smith for Directory Cache Management In a Distributed Data Processing System.

The disclosures of the foregoing co-pending applications are incorporated herein by reference.

## DESCRIPTION

### Field of the Invention

This invention generally relates to improvements in operating systems for a distributed data processing system and, more particularly, to an operating system for a multi-processor system interconnected by a local area network (LAN) or a wide area network (WAN). IBM's System Network Architecture (SNA) may be used to construct the LAN or WAN. The operating system according to the invention permits the accessing of files by processors in the system, no matter where those files are located in the system. The preferred embodiment of the invention is disclosed in terms of a preferred embodiment which is implemented in a version of the UNIX<sup>sup.1</sup> operating system; however, the invention could be implemented in other and different operating systems.

<sup>sup.1</sup> Developed and licensed by AT&T. UNIX is a registered trademark of AT&T in the U.S.A. and other countries.

## BACKGROUND OF THE INVENTION

Virtual machine operating systems are known in the prior art which make a single real machine appear to be several machines. These machines can be very similar to the real machine on which they are run or they can be very different. While many virtual machine operating systems have been developed, perhaps the most widely used is VM/370 which runs on the IBM System/370. The VM/370 operating system creates the illusion that each of several users operating from terminals has a complete System/370 with varying amounts of disk and memory capacity.

The physical disk devices are managed by the VM/370 operating system. The physical volumes residing on disk are divided into virtual volumes of various sizes and assigned and accessed by users carrying out a process called mounting. Mounting defines and attaches physical volumes to a VM/370 operating system and defines the virtual characteristics of the volumes such as size, security and ownership.

Moreover, under VM/370 a user can access and use any of the other operating systems running under VM/370 either locally on the same processor or remotely on another processor. A user in Austin can use a function of VM/370 called "passthru" to access another VM/370 or MVS/370 operating system on the same processor or, for example, a processor connected into the same SNA network and located in Paris, France. Once the user has employed this function, the files attached to the other operating system are

available for processing by the user.

There are some significant drawbacks to this approach. First, when the user employs the "passthru" function to access another operating system either locally or remotely, the files and operating environment that were previously being used are no longer available until the new session has been terminated. The only way to process files from the other session is to send the files to the other operating system and effectively make duplicate copies on both disks. Second, the user must have a separate "logon" on all the systems that are to be accessed. This provides the security necessary to protect the integrity of the system, but it also creates a tremendous burden on the user. For further background, the reader is referred to the text book by Harvey M. Deitel entitled *An Introduction to Operating Systems*, published by Addison-Wesley (1984), and in particular to Chapter 22 entitled "VM: A Virtual Machine Operating System". A more in depth discussion may be had by referring to the text book by Harold Lorin and Harvey M. Deitel entitled *Operating Systems*, published by Addison-Wesley (1981), and in particular to Chapter 16 entitled "Virtual Machines".

The invention to be described hereinafter was implemented in a version of the UNIX operating system but may be used in other operating systems having characteristics similar to the UNIX operating system. The UNIX operating system was developed by Bell Telephone Laboratories, Inc., for use on a Digital Equipment Corporation (DEC) minicomputer but has become a popular operating system for a wide range of minicomputers and, more recently, microcomputers. One reason for this popularity is that the UNIX operating system is written in the C programming language, also developed at Bell Telephone Laboratories, rather than in assembly language so that it is not processor specific. Thus, compilers written for various machines to give them C capability make it possible to transport the UNIX operating system from one machine to another. Therefore, application programs written for the UNIX operating system environment are also portable from one machine to another. For more information on the UNIX operating system, the reader is referred to UNIX.TM. System, User's Manual, System V, published by Western Electric Co., January 1983. A good overview of the UNIX operating system is provided by Brian W. Kernighan and Rob Pike in their book entitled *The Unix Programming Environment*, published by Prentice-Hall (1984). A more detailed description of the design of the UNIX operating system is to be found in a book by Maurice J. Bach, *Design of the Unix Operating System*, published by Prentice-Hall (1986).

AT&T Bell Labs has licensed a number of parties to use the UNIX operating system, and there are now several versions available. The most current version from AT&T is version 5.2. Another version known as the Berkeley version of the UNIX operating system was developed by the University of California at Berkeley. Microsoft, the publisher of the popular MS-DOS and PC-DOS operating systems for personal computers, has a version known under their trademark as XENIX. With the announcement of the IBM RT<sup>sup.2</sup> PC (RISC (reduced instruction set computer) Technology Personal Computer) in 1985, IBM Corp. released a new operating system called AIX<sup>sup.3</sup> (Advanced Interactive Executive) which is compatible at the application interface level with AT&T's UNIX operating system, version 5.2, and includes extensions to the UNIX operating system, version 5.2. For more description of the AIX operating system, the reader is referred to AIX Operating System Technical Reference, published by IBM Corp., First Edition (Nov. 1985).

.sup.2 RT and RT PC are trademarks of IBM Corporation.

.sup.3 AIX is a trademark of IBM Corporation.

The invention is specifically concerned with distributed data processing systems characterized by a plurality of processors interconnected in a network. As actually implemented, the invention runs on a plurality of IBM RT PCs interconnected by IBM's Systems Network Architecture (SNA), and more



specifically SNA LU 6.2 Advanced Program to Program Communication (APPC). SNA uses as its link level Ethernet.sup.4, a local area network (LAN) developed by Xerox Corp., or SDLC (Synchronous Data Link Control). A simplified description of local area networks including the Ethernet local area network may be found in a book by Larry E. Jordan and Bruce Churchill entitled Communications and Networking for the IBM PC, published by Robert J. Brady (a Prentice-Hall company) (1983). A more definitive description of communications systems for computers, particularly of SNA and SDLC, is to be found in a book by R. J. Cypser entitled Communications Architecture for Distributed Systems, published by Addison-Wesley (1978). It will, however, be understood that the invention may be implemented using other and different computers than the IBM RT PC interconnected by other networks than the Ethernet local area network or IBM's SNA.

.sup.4 Ethernet is a trademark of Xerox Corporation.

As mentioned, the invention to be described hereinafter is directed to a distributed data processing system in a communication network. In this environment, each processor at a node in the network potentially may access all the files in the network no matter at which nodes the files may reside. As shown in FIG. 1, a distributed network environment 1 may consist of two or more nodes A, B and C connected through a communication link or network 3. The network 3 can be a local area network (LAN) as mentioned or a wide area network (WAN), the latter comprising a switched or leased teleprocessing (TP) connection to other nodes or to a SNA network of systems. At any of the nodes A, B or C there may be a processing system 10A, 10B or 10C, such as the aforementioned IBM RT PC. Each of these systems 10A, 10B and 10C may be a single user system or a multi-user system with the ability to use the network 3 to access files located at a remote node in the network. For example, the processing system 10A at local node A is able to access the files 5B and 5C at the remote nodes B and C.

The problems encountered in accessing remote nodes can be better understood by first examining how a standalone system accesses files. In a standalone system, such as 10 shown in FIG. 2, a local buffer 12 in the operating system 11 is used to buffer the data transferred between the permanent storage 2, such as a hard file or a disk in a personal computer, and the user address space 14. The local buffer 12 in the operating system 11 is also referred to as a local cache or kernel buffer. For more information on the UNIX operating system kernel, see the aforementioned books by Kernighan et al. and Bach. The local cache can be best understood in terms of a memory resident disk. The data retains the physical characteristics that it had on disk; however, the information now resides in a medium that lends itself to faster data transfer rates very close to the rates achieved in main system memory.

In the standalone system, the kernel buffer 12 is identified by blocks 15 which are designated as device number and logical block number within the device. When a read system call 16 is issued, it is issued with a file descriptor of the file 5, and a byte range within the file 5 as shown in step 101 in FIG. 3. The operating system 11 takes this information and converts it to device number and logical block numbers of the device in step 102. Then the operating system 11 reads the cache 12 according to the device number and logical block numbers in step 103.

Any data read from the disk 2 is kept in the cache block 15 until the cache block 15 is needed. Consequently, any successive read requests from an application program 4 that is running on the processing system 10 for the same data previously read from the disk is accessed from the cache 12 and not the disk 2. Reading from the cache is less time consuming than accessing the disk; therefore, by reading from the cache, performance of the application 4 is improved. Obviously, if the data which is to be accessed is not in the cache, then a disk access must be made, but this requirement occurs infrequently.

Similarly, data written from the application 4 is not saved immediately on the disk 2 but is written to the

cache 12. This again saves time, improving the performance of the application 4. Modified data blocks in the cache 12 are saved on the disk 2 periodically under the control of the operating system 11.

Use of a cache in a standalone system that utilizes the AIX operating system, which is the environment in which the invention was implemented, improves the overall performance of the system disk and minimizes access time by eliminating the need for successive read and write disk operations.

In the distributed networking environment shown in FIG. 1, there are two ways the processing system 10C in local node C could read the file 5A from node A. In one way, the processing system 10C could copy the whole file 5A and then read it as if it were a local file 5C residing at node C. Reading the file in this way creates a problem if another processing system 10B at node B, for example, modifies the file 5A after the file 5A has been copied at node C. The processing system 10C would not have access to the latest modifications to the file 5A.

Another way for processing system 10C to access a file 5A at node A is to read one block at a time as the processing system at node C requires it. A problem with this method is that every read has to go across the network communications link 3 to the node A where the file resides. Sending the data for every successive read is time consuming.

Accessing files across a network presents two competing problems as illustrated above. One problem involves the time required to transmit data across the network for successive reads and writes. On the other hand, if the file data is stored in the node to reduce network traffic, the file integrity may be lost. For example, if one of the several nodes is also writing to the file, the other nodes accessing the file may not be accessing the latest updated file that has just been written. As such, the file integrity is lost, and a node may be accessing incorrect and outdated files. Within this document, the term "server" will be used to indicate the processing system where the file is permanently stored, and the term client will be used to mean any other processing system having processes accessing the file. The invention to be described hereinafter is part of an operating system which provides a solution to the problem of managing distributed information.

Other approaches to supporting a distributed data processing system in the UNIX operating system environment are known. For example, Sun Microsystems has released a Network File System (NFS) and Bell Laboratories has developed a Remote File System (RFS). The Sun Microsystems NFS has been described in a series of publications including S. R. Kleiman, "Vnodes: An Architecture for Multiple File System Types in Sun UNIX", Conference Proceedings, USENIX 1986 Summer Technical Conference and Exhibition, pp. 238 to 247; Russel Sandberg et al., "Design and Implementation of the Sun Network Filesystem", Conference Proceedings, Usenix 1985, pp. 119 to 130; Dan Walsh et al., "Overview of the Sun Network File System", pp. 117 to 124; JoMei Chang, "Status Monitor Provides Network Locking Service for NFS"; JoMei Chang, "SunNet", pp. 71 to 75; and Bradley Taylor, "Secure Networking in the Sun Environment", pp. 28. The AT&T RFS has also been described in a series of publications including Andrew P. Rifkin et al., "RFS Architectural Overview", USENIX Conference Proceedings, Atlanta, Ga. (June 1986), pp. 1 to 12; Richard Hamilton et al., "An Administrator's View of Remote File Sharing", pp. 1 to 9; Tom Houghton et al., "File Systems Switch", pp. 1 to 2; and David J. Olander et al., "A Framework for Networking in System V", pp. 1 to 8.

One feature of the distributed services system in which the subject invention is implemented which distinguishes it from the Sun Microsystems NFS, for example, is that Sun's approach was to design what is essentially a stateless machine. More specifically, the server in a distributed system may be designed to be stateless. This means that the server does not store any information about client nodes, including such information as which client nodes have a server file open, whether client processes have a file open in read.sub.-- only or read.sub.-- write modes, or whether a client has locks placed on byte ranges of the

file. Such an implementation simplifies the design of the server because the server does not have to deal with error recovery situations which may arise when a client fails or goes off-line without properly informing the server that it is releasing its claim on server resources.

An entirely different approach was taken in the design of the distributed services system in which the present invention is implemented. More specifically, the distributed services system may be characterized as a "statefull implementation". A "statefull" server, such as that described here, does keep information about who is using its files and how the files are being used. This requires that the server have some way to detect the loss of contact with a client so that accumulated state information about that client can be discarded. The cache management strategies described here, however, cannot be implemented unless the server keeps such state information. The management of the cache is affected, as described below, by the number of client nodes which have issued requests to open a server file and the read/write modes of those opens.

### SUMMARY OF THE INVENTION

It is therefore a general object of this invention to provide a distributed services system for an operating system which supports a multi-processor data processing system interconnected in a communications network that provides user transparency as to file location in the network and as to performance.

It is another, more specific object of the invention to provide a technique for providing a distributed file management system (DFS) with a file access control structure lock (fas.sub.-- lock) for preventing the problem of deadlocks.

According to the invention, these objects are accomplished by creating a fas.sub.-- lock for each file accessed from a remote system. The fas.sub.-- lock is used to lock instead of locking the file's inode. This makes it possible for the DFS to regulate accesses to files and avoid the problem of a deadlock occurring.

### BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, aspects and advantages of the invention will be better understood from the following detailed description of the preferred embodiment of the invention with reference to the accompanying drawings, in which:

FIG. 1 is a block diagram showing a typical distributed data processing system in which the subject invention is designed to operate;

FIG. 2 is a block diagram illustrating a typical standalone processor system;

FIG. 3 is a flowchart showing the steps performed by an operating system when a read system call is made by an application running on a processor;

FIG. 4 is a block diagram of the data structure illustrating the scenario for following a path to a file operation at a local node as performed by the operating system which supports the subject invention;

FIGS. 5 and 6 are block diagrams of the data structures illustrating the before and after conditions of the scenario for a mount file operation at a local node as performed by the operating system;

FIG. 7 is a block diagram, similar to FIG. 1, showing a distributed data processing system according to the invention;

FIG. 8 is a block diagram of the data structure for the distributed file system shown in FIG. 7;

FIGS. 9A to 9F are block diagrams of component parts of the data structure shown in FIG. 8;

FIGS. 10, 11 and 12 are block diagrams of the data structures illustrating the scenarios for a mount file operation and following a path to a file at a local and remote node in a distributed system as performed by the operating system;

FIG. 13 is a block diagram showing in more detail a portion of the distributed data processing system shown in FIG. 7;

FIG. 14 is a state diagram illustrating the various synchronization modes employed by the operating system which supports the present invention;

FIG. 15 is a block diagram, similar to FIG. 13, which illustrates the synchronous mode operations;

FIG. 16 is a state diagram, similar to the state diagram of FIG. 14, which shows an example of the synchronization modes of the distributed file system; and

FIG. 17 is a diagram showing the control flow of accesses to a file by two client nodes.

FIG. 18 is a diagram showing a deadlock when two operations are currently executing.

FIG. 19 is a diagram showing the execution steps of an open request from a client node.

#### DESCRIPTION OF THE PREFERRED EMBODIMENT

The following disclosure describes solutions to problems which are encountered when creating a distributed file system in which the logic that manages a machine's files is altered to allow files that physically reside in several different machines to

appear to be part of the local machine's file system. The implementation described is an extension of the file system of the AIX operating system. Reference should be made to the above-referenced Technical Reference for more information on this operating system. Specific knowledge of the following AIX file system concepts is assumed: tree structured file systems; directories; and file system organization, including inodes.

In a UNIX operating system, an individual disk (or diskette or partition of a disk) contains a file system. The essential aspects of the file system that are relevant to this discussion are listed below:

- a) each file on an individual file system is uniquely identified by its inode number;
- b) directories are files, and thus a directory can be uniquely identified by its inode number;
- c) a directory contains an array of entries of the following form:

name--inode number, where the inode number may be that of an individual file or that of another directory; and

- d) by convention, the inode number of the file system's root directory is inode number 2.

Following the path "/dir1/dir2/file" within a device's file system thus involves the following steps:

1. Read the file identified by inode number 2 (the device's root directory).
2. Search the directory for an entry with name=dir1.
3. Read the file identified by the inode number associated with dir1 (this is the next directory in the path).
4. Search the directory for an entry with name=dir2.
5. Read the file identified by the inode number associated with dir2 (this is the next directory in the path).
6. Search the directory for an entry with name=file.
7. The inode number associated with file in this directory is the inode number of the file identified by the path "/dir1/dir2/file".

The file trees which reside on individual file systems are the building blocks from which a node's aggregate file tree is built. A particular device (e.g., hard file partition) is designated as the device which contains a node's root file system. The file tree which resides on another device can be added to the node's file tree by performing a mount operation. The two principal parameters to the mount operation are (1) the name of the device which holds the file to be mounted and (2) the path to the directory upon which the device's file tree is to be mounted. This directory must already be part of the node's file tree; i.e., it must be a directory in the root file system, or it must be a directory in a file system which has already been added (via a mount operation) to the node's file tree.

After the mount has been accomplished, paths which would ordinarily flow through the "mounted over" directory instead flow through the root inode of the mounted file system. A mount operation proceeds as follows:

1. Follow the path to the mount point and get the inode number and device number of the directory which is to be covered by the mounted device.
2. Create a data structure which contains essentially the following:
  - a) the device name and inode number of the covered directory; and
  - b) the device name of the mounted device.

The path following in the node's aggregate file tree consists of (a) following the path in a device file tree until encountering an inode which has been mounted over (or, of course, the end of the path); (b) once a mount point is encountered, using the mount data structure to determine which device is next in the path; and (c) begin following the path at inode 2 (the root inode) in the device indicated in the mount structure.

The mount data structures are volatile; they are not recorded on disk. The list of desired mounts must be re-issued each time the machine is powered up as part of the Initial Program Load (IPL). The preceding discussion describes how traditional UNIX operating systems use mounts of entire file systems to create

file trees and how paths are followed in such a file tree. Such an implementation is restricted to mounting the entire file system which resides on a device. The invention described herein is based on an enhancement, embodying the concept of a virtual file system, which allows (1) mounting a portion of the file system which resides on a device by allowing the mounting of directories in addition to allowing mounting of devices, (2) mounting either remote or local directories over directories which are already part of the file tree, and (3) mounting of files (remote or local) over files which are already part of the file tree.

In the virtual file system, the operations which are performed on a particular device file system are clearly separated from those operations which deal with constructing and using the node's aggregate file tree. A node's virtual file system allows access to both local and remote files.

The management of local files is a simpler problem than management of remote files. For this reason, the discussion of the virtual file system is broken into two parts. The first part describes only local operations. This part provides a base from which to discuss remote operations. The same data structures and operations are used for both remote and local operations. The discussion on local operations describes those aspects of the data and procedures which are relevant to standalone operations. The discussion on remote operations adds information pertinent to remote operations without, however, reiterating what was discussed in the local operations section.

FIG. 4 shows the relationship that exists among the data structures of the virtual file system. Every mount operation creates a new virtual file system (vfs) data structure. The essential elements in this structure are (a) a pointer to the root vnode (virtual node) of this virtual file system (e.g., the arrow from block 21 to block 23), and (b) a pointer to the vnode which was mounted over when this virtual file system was created (e.g., the arrow from block 25 to block 24).

Whenever an inode needs to be represented in the file system independent portion of the system, it is represented by a vnode. The essential elements in this structure are the following:

- a) a pointer to the vfs which contains the vnode (e.g., the arrow from block 22 to block 21);
- b) a pointer to the vfs which is mounted over this vnode (e.g., the arrow from block 24 to block 25; but note however that not all vnodes are the mount point for a virtual file system, i.e., a null pointer indicates that this vnode is not a mount point);
- c) a pointer to either a surrogate inode or a real inode (e.g., the arrow from block 26 to block 32); and
- d) a pointer to a node table entry (this is a non-null only when the file is a remote file).

The AIX operating system, in common with other UNIX operating systems, keeps a memory resident table which contains information about each inode that is being used by the system. For instance, when a file is opened, its inode is read from the disk and a subset of this inode information, together with some additional information, is stored in the inode table. The essential elements of an inode table entry are (a) a pointer to the head of a file access structure list and (b) information from the disk inode, the details of which are not relevant here.

The file access structure records information about which nodes have the file open, and about the modes (read only or read.sub.-- write) of these opens. There is a separate file access structure for each node which has the file open. This state information enables the server to know how each client is using the server file.

The file system supports a set of operations which may be performed on it. A process interacts with a file system by performing a file system operation as follows:

1. The user calls one of the operations providing (perhaps) some input parameters.
2. The file system logic performs the operation, which may alter the internal data state of the file
3. The file system logic returns to the calling user, perhaps returning some return parameters. The operations which can be performed on a file system are referred to as "vn.sub.-- operations" or "vn.sub.-- ops". There are several vn.sub.-- ops, but the ones which are important to this discussion are described below:

#### VN.sub.-- LOOKUP

In the vn.sub.-- lookup operation, the essential iterative step in following a path in a file system is to locate the name of a path component in a directory file and use the associated inode number to locate the next directory in the chain. The pseudo code for the vn.sub.-- lookup operation is listed below:

---

```

function lookup
input: directory vnode pointer,
name to be looked up in directory
output: vnode pointer to named file/dir.
convert directory vnode pointer
to an inode pointer;
use private data pointer of vnode
lock directory's inode;
if( we don't have search permission in
directory )
unlock directory inode;
return error;
search directory for name;
if( found )
create file handle for name;
use inode found in directory entry;
get pointer to vnode for file handle;
unlock directory inode;
return pointer to vnode;
else -- not found
unlock directory inode;
return error;

```

---

#### VN.sub.-- OPEN

The function vn.sub.-- open creates a file access structure (or modifies an existing one) to record what open modes (READ/WRITE or READ.sub.-- ONLY) to open a file. The pseudo code for the vn.sub.-- open operation is listed below:

---

```

function vn.sub.-- open
inputs: vnode pointer for file to be opened
open flags (e.g., read-only,

```

```

        read/write)
create mode -- file mode bits if
        creating
output: return code indicating success or
        failure
get pointer to file's inode from vnode;
lock inode;
if( not permitted access )
unlock inode;
return( error );
get the file access structure for this
client;
if there is no file access structure
allocate one
if( couldn't allocate file access
structure )
unlock inode;
return( error );
update file access structure read-only,
read/write, and text counts;
if( truncate mode is set )
truncate file:
unlock the inode;

```

---

## LOOKUPPN

The lookuppn operation is the function which follows paths. Its input is a path (e.g., "/dir1/dir2/file"), and its return is a pointer to the vnode which represents the file. Lookuppn calls vn.sub.-- lookup to read one directory, then it checks to see if the vnode returned by vn.sub.-- lookup has been mounted over. If the vnode is not mounted over, then lookuppn calls vn.sub.-- lookup in the same file system. If the vnode has been mounted over, then lookuppn follows the pointer from the mounted over vnode (e.g., block 24 in FIG. 4) to the vfs of the mounted file system (e.g., block 25 in FIG. 4). From the vfs, it follows the pointer to the root vnode (e.g., block 26 in FIG. 4) and issues a new vn.sub.-- lookup giving as input the vfs's root vnode and the name which constitutes the next element in the path. The pseudo code for the lookuppn function is listed below:

---

```

function lookuppn
input: pathname
output: pointer to vnode for named file
if( first character of path is '/' )
current vnode for search is user's root
directory vnode;
else
current vnode for search is user's
current directory vnode;
repeat
if( next component of path is "." )
while( current vnode is root of a
virtual file system )
current vnode becomes the vnode that
the virtual file system is mounted
over;
if( there is not mounted over vnode )

```



```

        return( error ); -- ".." past root
        of file system
    use vn.sub.-- lookup to look up path component
    in current vnode;
    if( vn.sub.-- lookup found component );
    current vnode becomes the vnode
        returned by vn.sub.-- lookup;
    while( current vnode is mounted over )
        follow current vnode's pointer to vfs
        structure that represents the
        mounted virtual file system;
    current vnode becomes root vnode of
        the mounted vfs;
    else -- vn.sub.-- lookup couldn't file component
    return( error ); -- search failed
    until( there are no additional path
    components );
    return( current vnode );

```

---

The operation will be illustrated by describing the scenarios of following a path to a file and mounting a directory. First, in following a path to a file, suppose an application process issues a system call (e.g., open) for file "/u/dept54/status". This request is accomplished by the operating system in the following manner with reference to FIG. 4 (operations which are basically unchanged from the UNIX operating system are not explained here in any detail). The following assumptions are made: First, the vfs represented by block 21 is the root virtual file system. Second, the file "/u" is represented by vnode block 24 and inode block 31. Third, a previous mount operation has mounted a device onto the directory "/u". This mount created the vfs represented by block 25. Fourth, all of the directories and files involved are on the same device. Fifth, the following directory entries exist in the indicated directories:

DIRECTORY			
INODE NUMBER	NAME	INODE NUMBER	
2	"u"	15	
45	"dept54"	71	
71	"status"	12	

---

The code which implements the system call calls `lookuppn` to follow the path. `Lookuppn` starts at the root vnode (block 23) of the root virtual file system (block 21) and calls `vn.sub.-- lookup` to look up the name "u" in the directory file represented by this vnode. `Vn.sub.-- lookup` finds in the directory that the name "u" is associated with inode 15 in block 31. `Vn.sub.-- lookup` must return a pointer to a vnode associated with inode 15. To do this it first brings inode 15 into the inode table. Then it checks to see if there is already a vnode in the parent vfs (the input vnode (block 23) has a pointer to the parent vfs) for this vnode. In this case there is. `Vn.sub.-- lookup` then finds the vnode (block 24) in the root vfs (block 21) and returns a pointer to the vnode. `Lookuppn` discovers that the returned vnode is mounted over in the parent vfs. It follows the "mounted over" pointer from the vnode (block 24) to the mounted vfs (block 25). `Lookuppn` follows the "root vnode" pointer to the root vnode (block 26) of the new vfs (block 25). `Lookuppn` now calls `vn.sub.-- lookup` again, this time inputting a pointer to the root vnode (block 26) and name "dept54". As before, `vn.sub.-- lookup` reads the directory, finds the inode

associated with the name, finds or creates a vnode for this inode in the parent vfs (block 25) and returns a pointer to this vnode. Lookuppn calls vn.sub.-- lookup once more inputting the vnode for the just found directory and the name "status". Vn.sub.-- lookup reads the directory, finds the inode associated with the name (block 34), finds or creates a vnode (block 28) for this inode in the parent vfs (block 25) and returns a pointer to this vnode. The code which implements the system call now performs the requested operation on the file.

Suppose now that an application process issues a "mount" system call to mount the directory "/u/gorp" over the directory "/u/foo". The following scenario explains how this request is accomplished by the operating system (again, operations which are basically unchanged from UNIX operating system are not explained in any detail).

This scenario refers to FIG. 5, which represents initial conditions, and FIG. 6, which represents the final conditions, with the following assumptions: First, the vfs represented by block 41 is the root virtual file system. Second, all of the directories and files involved are on the same device. Third, the following directory entries exist in the indicated directories:

DIRECTORY		
INODE NUMBER	NAME	INODE NUMBER
2	"u"	15
2	"etc"	83
15	"gorp"	92
83	"foo"	75
75	"file1"	89

The code which implements the mount system call performs the following operations. Lookuppn is called to follow the path to the directory which is to be mounted over--"/etc/foo". At the completion of this operation, the root vfs (block 41) contains a vnode for "/etc/foo" (block 44) which has a pointer to the root vfs (block 41) and pointer to an inode table entry (block 45) for inode 75. Lookuppn is called to follow a path to the directory which is to be mounted --"/etc/gorp". At the completion of this operation, the root vfs (block 41) contains a vnode for "/etc/gorp" (block 49) which has a pointer to the root vfs (block 41) and a pointer to an inode table entry (block 48) for inode 92. Now the mount logic creates the new virtual file system by first creating a new vfs (block 46) and then creating a root vnode for this vfs (block 47) with a pointer back to its parent vfs (block 46) and a pointer to the root inode (inode 92, block 48) of the vfs. A "mounted over" pointer is installed in the covered vnode (block 44) in the root vfs (block 41), and a pointer to the vnode upon which it is mounted (block 44) is installed in the new vfs.

The foregoing illustrates the data structure for standalone operation. Reference is now made to FIG. 7 which shows a distributed system similar to that shown in FIG. 1 in which the operating system which supports the present invention has been implemented. In the following description, the term "server" is used to indicate the node where a file is permanently stored, and the term "client" is used to mean any other node having processes accessing the file. It is to be understood, however, that the term "server" does not mean a dedicated server as that term is used in some local area network systems. The distributed services system in which the invention is implemented is a truly distributed system supporting a wide variety of applications running at different nodes in the system which may access files located anywhere in the system.

The data structure for the distributed system shown in FIG. 7 is illustrated in FIG. 8, and the component parts of that data structure are shown in FIGS. 9A to 9F. With reference to FIG. 8, a client node may have access to files which reside in a remote server node. Such a client gains access to a server's files by mounting one of the server's directories. In the client node, the data structures created by a remote mount operation compare to those created by mounting a local entity in the following ways: Just as in the local case, a remote mount creates a vfs in the client node (e.g., block 54). Just as in the local case, use of a file in a virtual file system which contains remote files creates a vnode structure in the client node (e.g., block 57). Just as in the local case, the vnode structure has a pointer to an inode table entry (e.g., block 63). The inode table entry, however, does not contain the inode information from the remote file. Instead, the inode table entry contains a surrogate inode. This surrogate inode stands for, or represents, the remote inode.

In the server node, some data structures are constructed to allow the server to record state information about how remote nodes are using its files. More specifically, each server has a "dummy vfs" (e.g., block 71) to provide a vfs to hold files open by remote clients. A server has only one dummy vfs, and all files open by all remote clients reside in this vfs. The dummy vfs is not a part of the server's file tree. For each file which is open by a remote node, there is a vnode (e.g., block 74) in the server's dummy vfs. Each file which is open by a

remote node has an inode table entry in the server's inode table (e.g., block 85). This inode table entry is the same as that which exists because a local process at the server has a file open. For example, block 84, which is in the table because of a remote open, has the same structure as block 88, which is in the table because of an operation at the server.

When a client and server communicate about a server file, they need a way to identify the file. This is done with a file handle. When a client request causes the server to reply with a designation of a particular file (e.g., a remote lookup request), the file is identified by a file handle. When a client request carries a designation of a particular file (e.g., a remote open request), the file is identified by a file handle. The file handle contains the following fields: device number, inode number, and inode generation number.

The need for a file handle is illustrated by the following scenario. Suppose a client makes a request of a server and gets a file handle in reply. The client stores and remembers the file handle. Some activity at the server causes the file to be deleted and the inode slot reused for another file. The client makes a request of the server using the stored file handle. The server receives the file handle and performs the operation on the new file. This would be an unacceptable operation.

This flaw is prevented by use of the inode generation number. The inode generation number is stored on disk as a field in the inode. When the server deletes a file, it increments the inode generation number. If a request arrives at a server, the file handle is broken apart, the device number and inode number are used to locate the inode, and then the file handle inode generation number is compared to the inode's inode generation number. If they are different, then the request is rejected.

When a client wants to open a file which resides on a remote server, it uses a network transport mechanism to establish a connection with the server. Subsequent transactions regarding this file (e.g., read, write, etc.) flow on this connection. Each node contains a node table. A node uses entries in its node table (e.g., block 70) to record information about existing connections to remote nodes.

There are a limited number of operations that one node in the network can request another node to perform on its behalf. These operations are called dfs.sub.-- ops. When a node makes a request of

another node, the following operations occur: First, the requesting node sends a message which specifies which dfs.sub.-- operation is being requested and carries the parameters appropriate to that request. Next, the receiving node receives the request and performs the specified operation. Finally, the receiving node sends a message which carries the reply parameters appropriate for the dfs.sub.-- operation.

There is a close correlation between the vn.sub.-- ops that are issued within a local node to a file system and the dfs.sub.-- ops that are issued over the network. A typical operation on a remote file is as follows: First, a local kernel issues a vn.sub.-- op, not knowing whether the file being operated on is remote or local. Second, since the file resides in a remote node, the file system implementation code sends the corresponding dfs.sub.-- op to the node which holds the file. Note that if the file had been a local file, the operation would have been performed, the return parameters would have been returned, and the task would have been complete. Third, the node which holds the file receives the dfs.sub.-- operation request and requests its local file system to perform the corresponding vn.sub.-- operation. The return parameters from this vn.sub.-- op are used to construct the return parameters for the dfs.sub.-- op. Fourth, the requesting node receives the dfs.sub.-- op reply from the server node and uses the dfs.sub.-- op return parameters to construct the return parameters to the original vn.sub.-- operation request.

The operation will be illustrated by describing the scenarios of mounting a remote directory over a local directory and following a path to a file. In the first scenario, suppose that an application process in a client node issues a "mount" system call to mount a server node's directory "/u/gorp" over the local client directory "/etc/foo". The following scenario explains how this request is accomplished. This scenario refers to FIG. 10, which represents the initial conditions, and to FIG. 11, which represents the final condition, with the following assumptions: The vfs represented by block 51 is the root virtual file system of the server's file tree, and all the server directories and files involved are on the same device. The following entries exist in the indicated directories:

Server Node			
DIRECTORY			
INODE NUMBER	NAME	INODE NUMBER	
2	"u"	15	
15	"gorp"	92	
92	"file2"	67	

  

Client Node			
DIRECTORY			
INODE NUMBER	NAME	INODE NUMBER	
2	"etc"	83	
83	"foo"	75	

The code which implements the mount system calls lookupn to follow the path to the directory which is to be mounted over--"/etc/foo". At the completion of this operation, the root vfs (block 51) contains a vnode for "/etc/foo" (block 53) which has a pointer to the root vfs (block 51) and a pointer to an inode table entry (block 61) for inode 75. Since the directory being mounted resides in a remote node, a dfs.sub.-- mount request is issued to the server node, passing the path "/u/gorp" as the path to the object to be mounted. Upon receiving the dfs.sub.-- mount request, the server node calls lookupn to follow the path to the directory which is to be mounted--"/u/gorp". At the completion of this lookup operation, the server's root vfs (block 71) contains a vnode for "/u/gorp" which has a pointer to the root vfs and pointer

to an inode table entry for inode 92. The server uses the information in the inode (device 0, inode 92) to construct a file handle for the file "/u/gorp". The server returns this file handle in the reply to the dfs.sub.-- mount request and then releases the vnode and inode. Finally, the client receives the file handle in the reply to the dfs.sub.-- mount request and performs the operations necessary to create the new virtual file system as follows:

- a) Create a new vfs (block 54).
- b) Create a root vnode for this vfs (block 55) with a pointer back to its parent vfs (block 54) and a pointer to the root inode of the vfs (block 62). Since the root inode of this vfs is a remote directory, the inode pointed to from the root vnode is a surrogate inode. This surrogate inode contains the file handle returned by the server in response to the client's dfs.sub.-- mount request.
- c) Install a "mounted over" pointer in the covered vnode (block 53) in the root vfs (block 51).
- d) Install in the new vfs (block 54) a pointer to the vnode upon which it is mounted (block 53).

Suppose now that after executing the remote mount described above (mount server /u/gorp over client etc/foo) a client process issues a system call to operate on the file "/etc/foo/file2". The block numbers in the following scenario refer to FIG. 11, which represents initial conditions, and FIG. 12, which represents the system state after the open operation. First, the code which implements the system call calls lookuppn to follow the path. Lookuppn starts at the root vnode (block 52) of the root virtual file system (block 51) and calls vn.sub.-- lookup to look up the name "u" in the directory file represented by this vnode. Vn.sub.-- lookup finds in the directory that the name "u" is associated with inode 15. Vn.sub.-- lookup constructs a vnode and inode in the root virtual file system for inode 15 and returns to lookuppn a pointer to this vnode. Lookuppn calls vn.sub.-- lookup again, this time to look up the name "foo" in the directory identified by inode 15. Vn.sub.-- lookup reads the indicated directory and discovers that the name "foo" is associated with inode 75 in block 61. There already exists in the root vfs (block 51) a vnode (block 53) for this inode (block 61), so vn.sub.-- lookup returns a pointer to this vnode. Lookuppn discovers that the vnode is mounted over (the "mounted over" pointer in block 53 points to block 54). Lookuppn thus follows the "mounted over" pointer to the next vfs (block 54) and follows its root vnode pointer to the root vnode (block 55) of the virtual file system. Lookuppn now calls vn.sub.-- lookup for the next element ("file2") of the path giving vn.sub.-- lookup a pointer to block 55 and the name "file2". The directory to be searched resides in a remote node and is identified by the file handle stored in the client surrogate inode (block 62). Vn.sub.-- lookup issues a dfs.sub.-- lookup to the server which holds the file, sending the file handle which identifies the directory and the name ("file2") which is to be looked up. When the server receives the dfs.sub.-- lookup, it uses the file handle to identify the directory to be read and issues a vn.sub.-- lookup to search for the name "file2" in this directory. Vn.sub.-- lookup reads the directory and discovers that the inode number associated with the name "file2" is 67. Vn.sub.-- lookup constructs a vnode and inode in the dummy virtual file system for inode 67 and returns to lookuppn a pointer to this vnode. Dfs.sub.-- lookup uses the information in data structures returned by vn.sub.-- lookup to construct a file handle for the file identified by inode 67. It returns this file handle to the client, as the reply to the dfs.sub.-- lookup request, and releases the vnode and inode. In the client, a vnode (block 58) and surrogate inode (block 63) are created for the found file. Since "file2" is the last piece of the path, lookuppn returns to its caller a pointer to the found vnode (block 58). The code which implements the system call now performs the requested operation on the file.

In the distributed services system in which the invention is implemented as shown in FIG. 7, a local cache 12A, 12B and 12C exists at every node A, B and C. If file 5 permanently resides at node A on disk 2A, use of the cache 12A by local processes 13A executing at the server node A is the same as that

in a stand alone system as discussed above. However, remote processes 13B and 13C executing at nodes B and C, respectively, access file 5 through a two step caching scheme using a server cache and a client cache as shown in FIG. 3. The server node gets blocks of file 5 from disk 2A and stores it in the server cache 12A. Client node B goes out over the network 3 and gets blocks of file 5 from the server cache 12A. Client node B stores the blocks of file 5 as it existed in the server cache 12A into the client cache 12B. When the user address space 14B of client node B seeks data from any block of file 5, the client cache 12B is accessed instead of going across the network 3 for each access. Using the client cache 12B to access a remote file 5 can significantly improve the performance since it can save network traffic and overhead.

The system and method of this invention manages the use of the client cache 12B and server cache 12A in a distributed environment to achieve high performance while preserving the file access semantics at the application program level. This allows existing programs which run on a standalone system to run on a distributed system without any modification. The file access semantics preserves a file's integrity as it is being opened by different processes that issue read and write system calls to access and modify the file. The file access semantics require that only one I/O operation is allowed on any byte range at a time, and once an I/O operation starts, it cannot be pre-empted by any other I/O operation to the same byte range of the file.

An example of this is given by referring to FIG. 13. If process 131 issues a write system call to a byte range N1-N2 in file 5, the write system call can only be executed when the entire byte range N1-N2 is available for access by process 131, and no read operation involving the byte range N1-N2 is being executed. During the execution of the write system call, all other operations involving the byte range N1-N2 in file 5 are suspended until the write is completed. The write is not completed until the bytes are written to the local cache 12A. When a write request is complete, the written data in the cache 12A is visible to any subsequent read operation by any of the other processes 131 to 13N.

Another requirement of file access semantics is that when a file byte range such as N1-N2, which can be a record or a set of related records accessed by the same I/O operation, is visible to a read process, the file byte range N1-N2 must always have a consistent set of data reflecting the last update to this range. This range is never available for access while a write operation is being executed. In this way, the next read issued by a process will read the data just written and not the old outdated data.

In a distributed networking environment of this invention as shown in FIG. 7, the execution of read and write system calls from different application programs 4A and 4B and processes 131 to 13N and 231 to 23N are synchronized such that the file access semantics as previously discussed are preserved. The system and method of this invention guarantees synchronization by utilizing various cache synchronization (sync) modes. For a specific file 5, the I/O calls are synchronized by either the client B or the server A depending on the location of the processes 131 to 13N or 231 to 23N which have the file 5 open for access, and the sync mode.

The three synchronization modes are shown in FIG. 14 and are described with reference to FIG. 7. The first mode 104 is referred to as ASYNC s.sub.-- mode, or asynchronous mode. The file 5 operates in this mode 104 if file 5 is open for read/write access by processes 13C executing at only one client remote node C, as shown in block 107 of FIG. 14. In this mode 104, all of the control is in the client node C. Both the server cache 12A and client cache 12C are used for these read/write operations. A read or write operation requires access to the server cache 12A only if it cannot be satisfied from the client cache 12C. Modified blocks at the client 12C are written to the server 12A by the periodic sync operation, when the file 5 is closed by all processes 13C in the client node C, or when more room is needed in the client cache for other data. Additionally, modified blocks are written to the server when the file changes from ASYNC s.sub.-- mode to FULLSYNC s.sub.-- mode.

A second mode 105 is READONLY s.sub.-- mode. The READONLY s.sub.-- mode 105 is used for files 5 that are open for read only access from processes 13C in only one node C, or from processes 13B and 13C in more than one node B and C as shown in block 108 in FIG. 14. In this mode 105, the server cache 12A and the client caches 12B and/or 12C are used. The read request is issued for a block or more at a time. Every other read request from the same client, either B or C, to the specific block does not go to the server 12. Instead, it is read from the respective client cache, either B or C. In other words, a read operation does not require access to the server 12A if it can be satisfied from the client cache 12C or 12B. In summary, the file 5 operates in mode 105 if the file 5 is open for read only access by any of the processes 13A, 13B or 13C in any of the nodes A, B or C.

A third mode 106 is FULLSYNC s.sub.-- mode. The FULLSYNC s.sub.-- mode 106 is used for files 5 open in more than one node A, B and at least one node has the file 5 open for write access. In the FULLSYNC s.sub.-- mode 106, the client cache 12C or 12B is bypassed, and only the server cache 12A is used. All read and write operations are executed at the server 12A.

In a distributed environment 1 as shown in FIG. 7, most files 5 will more frequently be open for read only by processes 13A, 13B and 13C at several nodes A, B and C in the READONLY s.sub.-- mode 105 shown in FIG. 14 or open for update at only one node in the ASYNC s.sub.-- mode 104. It will be less frequent that there will be an open for read and write access by processes executing at more than one node in the FULLSYNC s.sub.-- mode 106. In both the READONLY s.sub.-- mode 42 and the ASYNC s.sub.-- mode 104, the use of a client cache 12B, shown in FIG. 13, significantly reduces the remote read/write response time of accessing file 5 and improves overall system performance.

As shown in FIG. 15, in the FULLSYNC s.sub.-- mode, the client cache is not used. The client node B accesses the file 5 from the server A over the network 3 for each read and write operation. Although the read/write response time increases in this mode, the file access semantics are preserved since a client does not retain a file 5 in local cache that has not been updated along with the corresponding file residing at the server.

Utilizing the three modes to manage the use of the client cache optimizes overall system performance by combining both an overall average increase in read/write response speed with file integrity. Using a client cache in some situations decreases the read/write response time; while not using a client cache in other situations preserves the file system semantics.

A file's sync mode is not only dependent on which nodes have the file open and whether the file is open for read or write, but also on whether the device where the file resides is open in raw access mode. Raw access for a device means that a block of data LBN1 shown in FIG. 13 within a device 2A is accessed. In this way, the reads and writes of the device 2A read and write to a block LBN1 of device 2A. It is not relevant to which file the block belongs. The device 2A can be opened for raw access from a process 131 to 13N at the server node A. It can not be opened for raw access from a remote node B or C.

In FIG. 13, the cache 12A is managed as blocks LBN1 of device 2A, similar to a standalone system as described above with reference to FIG. 2. The server A looks at the server cache 12A as a logical block LBN1 within a device 2A. The client B has no knowledge of where the file 5 resides on the device 2A. All the client B knows is that it accesses a file 5 on block number N1 on device 2A. The client cache 12B handles the data as logical blocks N1 of file 5. In the server cache 12A, the data is handled as logical blocks LBN1 of devices 2A. In handling the data this way, the server can guarantee that if data is written to the device as a raw device, and if there is another read of a block of the file that happens to be the same block that was written to the device, then the read would see the newly written data. This preserves the file system semantics.

If the file is being accessed in a client node B, and the file is in ASYNC or READONLY mode, as shown in FIG. 13, the client operating system 11B does not convert the file descriptor and byte range within the file in the system call READ (file descriptor, N1) 16 to the device number and the logical block number in the device. The client does convert the file descriptor and byte range to a file handle, node identifier, and logical block number within the file. In the client cache 12B, there are blocks 17 that are designated by file handle, node identifier, and logical block number within the file. When a read 16 is issued from a client application 4B, the request for the read goes to the operating system 11B with the file descriptor and the byte range within the file. The operating system then looks in the client cache 12B. If the file handle, node identifier, and logical block number within the file is there, the cache 12B is read; on the other hand, if it is not there, the read is sent to the server. The server then takes the file handle and the logical block number within the file and converts it to a device number and logical block in the device. This conversion is necessary since the server cache 12A is managed by device number and block number within the device as it is in a standalone system. After the read is sent to the server, it is handled the same as if the read was coming from its own application in a standalone system as described with reference to FIG. 2.

A closed file does not have a synchronization mode. However, once a file is first opened by a process, the file's sync mode is initialized according to the following as illustrated in FIG. 16. The sync mode for a file is initialized to ASYNC 104 if the device where the file resides is closed 112; i.e., it is not open as a special device and the file is open for write access at one remote node 113. The sync mode for a file is READONLY 105 if the device where the file resides is closed, and the file is open for read only access in one or more nodes 114, or both the device and the file are open for read only access 115. The sync mode for a file is initialized to FULLSYNC 106 if the device where the file resides is open as a block special device for read/write access 116, or the file is open in more than one node and at least one of the opens is for writing. A block special device means that there is a raw access to the device.

Once a file is initialized to a mode, if the conditions change, the file mode may change. Transitions from one mode to another, as shown by lines 118 to 123 in FIG. 16, may occur under the following conditions. If a file is presently in ASYNC mode 104, and the number of nodes where the file is open becomes two or more, 124, then the sync mode changes to FULLSYNC 106 as shown via line 119. Also, if there is an open of the block special device D where the file resides, 125, the sync mode will change from ASYNC 104 to FULLSYNC 106. In a close operation for the file, if the close operation is not the last close of the file, and the file is still open for write, there is no mode change. However, if the close operation is the last close of the file for write access such that all the remaining opens are for read access, 83, then the new mode becomes READONLY 105 as shown via line 121. If the close operation is the last close of the file, then there is no sync mode.

If a file is presently in READONLY s.sub.-- mode 105 and there is a file open operation, there will not be a mode change if the open is for read. However, if the open is for write, then the new sync mode is ASYNC 104 if all the opens are in one client node, 127, as shown via line 120. Otherwise the sync mode is FULLSYNC. Furthermore, if the device where the file resides is open for read/write access, 130, the new sync mode for the file is FULLSYNC mode 106. For a close operation, if the close is the last close of the file, there is no sync mode for the file. If the file is still open at one or more nodes after a close operation, there is no change to the sync mode.

If a file is presently in FULLSYNC mode 106 and there is another open for the file, or the device where the file resides is opened, there is no sync mode change. If after a close operation of the file, there remains an open for read/write access at the remote node, and the block special device where the file resides is not open, the sync mode is changed to ASYNC s.sub.-- mode 104, as shown by block 141 via line 118. The sync mode is changed from FULLSYNC 106 to READONLY 105 if the block special



device where the file resides is not open, and the file is open for read only access at one or more nodes as shown by block 142 on line 122, or if the block special device where the file resides is open for read only access and the file is open for read only access as shown in block 143 on line 122.

All open and close operations for files and devices are resolved at the server node. The server determines the sync mode of an open file when executing any operation that may change the mode. The server also performs the change of the synchronization modes. As the server gets new opens or closes for the file, a change in synchronization modes for the file may be triggered. If the required sync mode is not the current one, the server sends a "change sync mode" remote procedure call to all the clients with the file open. After a file is opened for the first time, the client that opened the file is informed of the mode of the file. If the mode is either ASYNC or READONLY, the client can start using the client cache for reads, and also for writes if the mode is ASYNC, as shown in FIG. 13. The client does not have to read or write over the communications link to the server. If the mode is FULLSYNC as shown in FIG. 15, the client cache is not used, and the client must send the read or write over the communications link 3 to the server.

The server A, in FIG. 15, always sets the mode 151 of the file 5. The mode of the file is the same at every node that has the file open. The server A also knows which nodes have the file open, and whether the opens are for reads or writes. The server A does not have to know which processes 131 to 13N, 231 to 23N within a node have a file open. The server keeps all the above information in a file access structure list 150, as shown in FIG. 15. Each element of the file access structure list 150 contains a node which has the file open 152, the number of opens for read 153 in the node, and the number of opens for write 154 in the node.

### UNIX File Locking

In a UNIX operating system, processes can lock byte ranges within files so that other processes can not have access to the ranges. Locks apply to a byte range of a file. A lock over the entire extent of a file locks the file and could be called a file lock. A lock over an arbitrary byte range is sometimes called a record lock, but for purposes of this disclosure, we will refer to record and file locks simply as locks.

Two types of locks are supported in this system, write locks and read locks. Write locks are exclusive locks: if a range of a file is write locked, no other locks can exist on that range. The other type of locks, read locks, are shared locks: any number of overlapping read locks can apply to a segment of a file. Notice that an existing read lock does not block other read locks, but it does block other write locks. An existing write lock blocks all other locks for a given range. Write locks can only be applied to file descriptors that have been opened with write access.

Files are in either in enforcement-mode or not in enforcement-mode. Locks on a file not in enforcement-mode are called advisory locks. An advisory lock does not provide absolute protection for a file or record; however, it does prevent a process from reading or writing the locked file or record. Advisory locks only affect the results of calls to lockf(2) or fcntl(2). They must be used by processes that are cooperating through the use of lockf(2) or fcntl(2) to query the status of locks on the shared files that they are accessing. The advantage of advisory locks is that they do not have to be interrogated by the operating system kernel during reading or writing operations. An enforced lock, like an advisory lock affects subsequent calls to lockf(2) and fcntl(2). In addition, each read(2), write(2), open(2), creat(2), fclear(2), fruncate(2), and shmat(2) will have to insure that no read or write locked portion of the file is being changed and that no write locked portion of the file is being accessed.

Three different UNIX operating system commands of the fcntl(2) system call are related to locking:

F.sub.-- GETLK Find the first existing lock that would prevent the lock described by fcntl(2)'s argument from being granted to the caller.

F.sub.13 SETLK Grant the lock described by fcntl(2)'s argument to the caller. If the lock cannot be granted because an existing lock interferes with the request, return the description of this existing lock.

F.sub.-- SETLKW Grant the lock described by fcntl(2)'s argument to the caller. If the lock cannot be granted because an existing lock interferes with the request, check for deadlock and if no deadlock will be caused, have the caller wait. Each time an interfering lock is cleared, the kernel will again attempt to establish the requested lock by searching for any interfering locks. A process could wait forever. Although deadlocks that involve only file locks on a single node will be detected, deadlocks due to multiple file locks on multiple nodes can occur. A process could never deadlock, but it could live-lock due to interfering locks for various files being alternately set.

### Lock Tables

Locks are associated with open files, so it is natural to keep lock information with the information about an open file, in the file's inode structure. The inode structure has a fixed size in the UNIX operating system so the lock information had to be stored in another structure with the inode only containing an address where the lock structure resides. The locks are kept in a linked list as a set of entries from the kernel data structure called the lock table.

The UNIX operating system did not provide support for a distributed file system. In a distributed system, there may be several nodes using the same file. The lock table for a file is always located in a single node. If the file is in ASYNCH mode, the lock table (along with the active node information) is kept at the single node where the file is open. When the file is in READONLY or FULLSYNCH mode, the lock table is kept at the server. There are two important implications of this architecture. First, processes may have to use remote procedure calls (RPC)s to set or test locks. These RPCs will run on the file's server. Second, when the synch mode of a file changes, the file's lock table may have to be moved from the client to the server or vice-versa. The entries of an inode's lock table correspond to locks over segments of the inode's file. To represent a lock, a lock set entry must contain information identifying the range of bytes locked, the type of lock (read or write), the owner of the lock.

Some searching of the kernel data structures is eliminated by storing a flag in the lock entry indicating whether or not a process is waiting on the entry's lock. This flag is another piece of information that is contained in the actual implementation of the lock set entries. Finally, the lock set entries will probably contain a pointer field used to link entries belonging to the same file together. We will study the implementation details by taking a detailed look at the following operations for manipulating the elements of a lock set:

- 1) A way of iterating through the locks in the sets of locks associated with a particular inode.
- 2) A function that adds a lock to the set of locks for an inode.
- 3) A function that removes a lock from the set of locks for an inode.

An entry in a lock table represents a lock with a minimum of the byte range locked and the owner information for the lock. Operations that return the following attributes of locks will be needed:

- 1) The range of the segment locked
- 2) The type of the lock (read, write or temporary)
- 3) The lock owner

A problem occurs when an owner is associated with a lock. The owner of a lock is identified by a pointer to the owner's proc table. The proc table is a collection of management information concerning each active process in the system. The data structure source code for the proc table is provided below with information describing each field's purpose:

---

```

STRUCT  PROC      {
STRUCT  SEG.sub.-- LIST
                Segment Ids
                P.sub.-- SEGS;
INT      P.sub.-- FLAG;
                Flag for testing
CHAR     P.sub.-- STAT;
                Status of process
CHAR     P.sub.-- PRI;
                Priority, Neg. High
CHAR     P.sub.-- TIME;
                Resident time Sched.
CHAR     P.sub.-- CPU;
                CPU Usage
CHAR     P.sub.-- NICE;
USHORT   P.sub.-- UID;
                Real User Id.
USHORT   P.sub.-- SUID;
                Effective User Id.
SHORT    P.sub.-- PGRP;
                Process Group Leader
SHORT    P.sub.-- PID;
                Unique Process Id.
SHORT    P.sub.-- PPID;
                Process Id. of Parent
LONG     P.sub.-- ADDR;
                Process Address
SHORT    P.sub.-- SIZE;
                Size of the image
STRUCT   USER *P.sub.-- UPTR;
                Location of
                ;      Unblock for this
                ;      process
CHAR     P.sub.-- CURSIG;
INT      P.sub.-- SIG;
                Signals Pending
INT      P.sub.-- SIGMASK;
                Signal Mask
INT      P.sub.-- SIGIGNORE;
                Signals being
                ;      ignored
INT      P.sub.-- SIGCATCH;
                Signals to Catch
INT      P.sub.-- SIGACTION;
                1=Bell style,
                ;      0=4.2 style
Union    {
                CADDR.sub.-- T P.sub.-- CAD;
                INT P.sub.-- INT;
                } P.sub.-- UNW;
#define P.sub.-- W.sub.-- CHAN
                P.sub.-- UNW.P.sub.-- CAD;

```

```

;          Pointer to the
;          Lock Table that a
;          process is waiting
;          on.
#define P.sub.-- ARG
P.sub.-- UNW.P.sub.-- INT;
STRUCT TEXT *P.sub.-- TEXTP;
;          Pointer to text
;          structure
STRUCT PROC *P.sub.-- LINK ;
;          Linked list of
;          Running Processes
INT P.sub.-- CLKTIM
;          Alarm Clock Signal
INT P.sub.-- SMBEG
;          Beg. of Shared Mem.
INT P.sub.-- SMEND
;          End of Shared Mem.
CADDR.sub.-- T
P.sub.-- FREQ
;          Null or Regset
LONG P.sub.-- PFLT
;          # of Page Faults
INT P.sub.-- EPID
;          Locking id process
INT P.sub.-- SYSID
;          NID of the remote
;          process }; End of
the structure

```

---

The proc table is a useful way to identify the owner of a lock element because deadlock detection needs to access the proc table for each lock's owner. In a distributed system, the owner might not be a local process. When an owner is a remote process, the owner information needs to include the node id for the node that the owner resides on. The information that is required is:

- 1) An operation that returns the address of the owner's proc table.
- 2) An operation that compares two owners and returns TRUE if the owners are the same process and FALSE otherwise.
- 3) An operation that returns the node id of the owner. These operations use field references within structures to perform the indicated operations.

#### Waiting For A Lock

In standalone operating system a process that tries to establish a lock may have to wait for an existing lock to clear first. Before waiting (going to sleep) the process must check the lock sets of all the inodes of the system to insure that no deadlock will occur if it does wait. A waiting process has its proc table use the W.sub.-- CHAN field of the proc table to point to the lock table entry that it is waiting on.

In a DFS, waiting is not as easy. There are two ways to wait on a blocking lock: 1) directly on a local lock set entry and 2) indirectly on a server lock set entry. Direct waiting is identical to the standalone

waiting described above. It is used anytime that a process must wait for a lock that occurs locally in a lock table. It is important to remember that a lock table for a file resides in only one node. If the calling process is not in the same node then the lock table is in the server. A process that attempts to lock a region of a file that has its lock table located in a remote node (server) waits on the lock indirectly. This indirect waiting is done by a RPC that invokes a transaction program in the server and waits on the lock. In a standalone UNIX operating system, a process never enters the SLEEP state if a lock can be granted or if waiting could create a deadlock. In a distributed system, a process performing a lock request that does not reside in the same node as the lock table always enters the SLEEP state, at least briefly. It must do so while it is waiting for the RPC to return. In order to avoid unnecessary network communication, a process that is waiting for the lock RPC will not know if it is waiting because the RPC transaction program is in turn waiting for a blocking lock or if it is waiting because the transaction program has not finished running in the server where no blocking lock was found. In a distributed system, deadlocks can span several nodes; therefore, to allow a process to wait only if there are no deadlocks is not practical using a similar architecture to the standalone environment due to the overhead and checking that must be done. Our invention is a method for providing such capabilities without the associated overhead and complexity associated with a distributed version of the standalone deadlock prevention.

### Deadlocks

A deadlock exists when there are a chain of processes executing and they are vying for resources that are already busy. For example, if process one is waiting for a resource that process two controls, process two is waiting for a resource that process three controls and so on until the last process is waiting for a resource that process one has control of. A standalone UNIX operating system prevents deadlocks involving file and record locking from occurring. In the past the UNIX operating system has not supported distributed systems and has not been able to prevent such deadlocks from occurring. The circular chain of processes forming a deadlock that we want to prevent are linked by two kinds of links: 1) processes pointing to locks through the process W.sub.-- CHAN fields of the process proc table and 2) locks pointing to processes through the owner fields of the locks. By following these links, a search for circularity can be performed to determine if a deadlock will occur. Because an UNIX operating system process can only wait on one thing (a single w.sub.-- chan) and a lock is owned by only one process, the searching of the chains is not hard. The only complication is the possibility of indirect waiting. Indirect waiting occurs when a process is waiting for a response from a RPC that has been sent to another node. To handle indirect waiting, applications are required to set a timer before executing a RPC to ensure an awakening if a deadlock does occur.

### Distributed File Support Lock Control

The file access calls in a UNIX operating system environment make use of a data structure called FLOCK. The structure of FLOCK is presented below:

---

```

STRUCT FLOCK {
SHORT          L.sub.-- TYPE;
SHORT          L.sub.-- WHENCE;
LONG           L.sub.-- START;
LONG           L.sub.-- LEN;
SHORT          L.sub.-- PID;
SHORT          L.sub.-- NID;
};

```

---

Another important structure for implementing distributed file support (DFS) is the file access structure. The source code for the file access structure is provided below with some descriptive information to show the detailed logic:

---

```

STRUCT      FILE.sub.-- ACCESS
{ /*
File Access Structure Pointer
*/
STRUCT      FILE.sub.-- ACCESS *FA.sub.-- NEXT; ;
/*
File Access Structure Flag
*/
SHORT      FA.sub.-- FLAG; /*
File Access Structure Total Users
*/
SHORT      FA.sub.-- COUNT; /*
File Access Structure Read/Only Count
*/
SHORT      FA.sub.-- OROCNT; /*
File Access Structure Read/Write Count
*/
SHORT      FA.sub.-- ORWCNT; /*
File Access Structure Executing Processes
*/
SHORT      FA.sub.-- TXTCNT; /*
File Access Structure Node Structure Ptr.
*/
STRUCT      NODE *FA.sub.-- NID; /*
File Access Structure Node ID
*/
INT         FA.sub.-- NID; /*
File Access Structure S.sub.-- INODE Pointer
*/
STRUCT      INODE *FA.sub.-- SIP; };

```

---

The DFS provides a lock control subroutine to coordinate the locking of files in a distributed environment. The subroutine's interface source code is provided below to present the detailed logic.

---

```

STRUCT FLOCK {
SHORT      L.sub.-- TYPE;
SHORT      L.sub.-- WHENCE;
LONG       L.sub.-- START;
LONG       L.sub.-- LEN;
SHORT      L.sub.-- PID;
SHORT      L.sub.-- NID;
};
DFS.sub.-- LOCK.sub.-- CONTROL(FH, LOCK.sub.-- INFO, CMD,
FAS.sub.-- FLAG)
FILE.sub.-- HANDLE.sub.-- T FH;
STRUCT FLOCK *LOCK.sub.-- INFO;
SHORT CMD, FAS.sub.-- FLAG;
RETURN(ERRNO, MODE, START, LENGTH, NID, PID)

```

---

The DFS.sub.-- LOCK.sub.-- CONTROL subroutine has parameters FH (the handle of the file), LOCK.sub.-- INFO (a pointer (address) of the FLOCK structure, CMD (process to be carried out) and FAS.sub.-- FLAG (file access structure flag). The file handle is necessary to uniquely identify the file that the command is to be run against. The LOCK.sub.-- INFO is necessary to identify the FLOCK structure and the system information on the file contained within. The CMD is set to have the remote process wait until any blocking locks are removed and then process the specified CMD. The CMD field can be set to carry out an unlock, test, set or test&set command. The l.sub.-- len and l.sub.-- start fields in the FLOCK structure are used to specify the lock range for the command to apply to. The FAS.sub.-- FLAG is set to tell the server to check the S.sub.-- MODE in the file access structure lock to see if it is in FULL.sub.-- SYNC S.sub.-- MODE. If the file is in FULL.sub.-- SYNC S.sub.-- MODE then the server proceeds with the command; however, if it is not, then an error is returned. If the file access structure flag is not set, then the server does not acquire and check the file access structure lock before proceeding with execution. This check allows an inquiry before processing and avoids the possibility of deadlocks.

#### File Locking Subroutine

The LOCKF subroutine is the principal interface used by applications to lock or unlock files for write exclusive privileges. The interface source code of the subroutine is presented below to demonstrate its usage:

---

```
#include <SYS/LOCKF.H>
int lockf(fildes, function, size);
int fildes, function;
long size;
```

---

Locks may be set or released by UNIX operating system FCNTL(2) in a similar manner to LOCKF. The external interface between applications and the LOCKF and FCNTL subroutines is unchanged between UNIX operating system and this invention; however, the internals have been modified to support a distributed environment.

#### Internal Details

If a remote file is to be locked, the UNIX operating system LOCKF and FCNTL system calls are intercepted and an RPC DFS.sub.-- LOCK.sub.-- CONTROL is executed. The server node receives the remote process call and carries out the lock request. The request could entail locking a single record, a set of records or the whole file. The server then tells the client to awaken by sending a signal while the client surrogate inode is waiting for a reply from the DFS.sub.-- LOCK.sub.-- CONTROL RPC. The client confirms the reception of the lock and sends an acknowledgement to the remote server. The server updates the lock table after receiving the acknowledgement from the client surrogate inode. If the server does not confirm the reception of DFS.sub.-- LOCK.sub.-- CONTROL's acknowledgement, then DFS.sub.-- LOCK.sub.-- CONTROL removes the lock from the lock table.

#### File Access Structure Lock

The file access structure lock fas.sub.-- lock is used to synchronize the use of the inodes and surrogate inodes (s.sub.-- inode) for open files in a distributed file system (DFS). The synchronization is performed to avoid a deadlock situation which can occur if an inode and the s.sub.-- inodes are locked.

In a standalone AIX operating system, execution of system calls that require access to a file F are serialized by locking the inode for F during the entire execution time of any system call for that file. In DFS, if file F is open at a remote node C, a s.sub.-- inode is created at node C to represent file F. Thus, two resources are involved: the inode for the specific file at the server node where the file resides, and the s.sub.-- inode at the client node where the file is open. To serialize system calls executing at client C, the s.sub.-- inode for file F is locked during the execution time for each call. If an access to the server is required to read a block of data not available in the client cache, the inode for file F is also locked.

Locking the inode for file F in the server and the s.sub.-- inode for file F in the client for the entire execution time of each system call can lead to a deadlock situation if the order in which the two resources are acquired is not always carried out in the same order. Typically, the s.sub.-- inode is locked first and then the server is accessed via a remote procedure call (RPC) and the inode is locked. However, there are some exceptions to the above order. Under certain conditions, the server may lock the inode and then send a RPC to the client which requires the locking of the s.sub.-- inode.

A deadlock can occur in any one of the following situations in which two operations are currently executing 01 and 02, as shown in FIG. 18, where 01 is a read operation, and 02 is an open operation:

- a) 01 is executing at a client node. 01 locks the s.sub.-- inode and tries to lock the inode in the server for a read operation.
- b) 02 is executing in the server. 02 locks the inode and initiates a RPC to the client node to open a file. The execution of the RPC request in the client node waits on the s.sub.-- inode to lock it.

Because both operations are executing and require the same two resources and each have acquired one and are waiting on the other locked resource, a deadlock situation is present. In examining the cause, note that the deadlock occurs during the execution of the RPC from the server to the client. The inode on the server is locked first and an attempt is made to lock the s.sub.-- inode. This is the reverse of most cases where the s.sub.-- inode is locked first and then sends a RPC to lock the inode.

To prevent the above problem from occurring, the server could unlock the inode before originating a RPC to lock the s.sub.-- inode to the client. Unlocking the inode during the execution of the open operation will solve the above problem; however, it complicates the sync mode change management for open files since more than one open and/or close operation can occur at the server in parallel. It may also introduce another problem as shown in FIG. 17. In FIG. 17 file F at 10 is open in ASYNC mode by only one process in client node C-1 at 20. Two operations are in progress: a close operation from C-1 at 20 and an open operation at 60 for the same file F at label 10 from another client C-2 at label 40. The close operation from C-1 at label 20 will lock the s.sub.-- inode (which has a use count of 1) and sends a "dfs.sub.-- close" RPC at label 50 to the server at label 30. The open operation from C-2 at label 40 sends a "dfs.sub.-- open" RPC at label 70 to the server at label 30. This RPC arrives at the server and executes before the "dfs.sub.-- close" RPC at label 50 from C-1 at label 20. The s.sub.-- mode for file F is ASYNC, so the server unlocks the inode and sends a "dfs.sub.-- chng.sub.-- sync.sub.-- mode" RPC at label 80 to C-1 at label 20 requesting that file F at label 10 be changed to FULLSYNC s.sub.-- mode. This RPC will arrive at C-1 at label 20 and wait for the s.sub.-- inode to be unlocked. Next, the "dfs.sub.-- close" RPC at label 50 arrives at the server. Since the inode for file F at label 10 is not locked at the server, the close operation executes at the server and a "dfs.sub.-- close.sub.-- ack" RPC at label



90 is sent to C-1 at label 20. When the "dfs.sub.-- close.sub.-- ack" RPC at label 90 arrives at C-1 at label 20, the use count on the s.sub.-- inode is decremented and since the use count's value is zero, the s.sub.-- inode is released at label 100. This leaves no s.sub.-- inode for the sync mode change to be applied to in C-1 in label 20.

A solution to this problem is to have the sync mode change procedure increment the use count of the s.sub.-- inode before waiting on it. However, this approach raises more management headaches for the file management system, since at this time file F is not open at C-1 and its s.sub.-- mode is not FULLSYNC. A better approach is to introduce a new lock, the file access structure lock (fas.sub.-- lock) to serialize operations which access or change the access list to a file. Use of the fas.sub.-- lock will eliminate the inode from being a critical resource. The two critical resources will be the s.sub.-- inode in the client node and the fas.sub.-- lock at the server. To prevent a deadlock, any operation executing at a client node which requires holding the fas.sub.-- lock should unlock the s.sub.-- inode before a RPC is sent to the server.

Operations that may generate an RPC from the server to clients must acquire the fas.sub.-- lock before it starts executing in the server. Examples of situations in a UNIX operating system and/or an AIX operating system environment are: Remote Procedure Calls (RPC):

* DFS.sub.-- OPEN	* DFS.sub.-- CREATE
* DFS.sub.-- CLOSE	* DFS.sub.-- GET.sub.-- ATTR
* DFS.sub.-- SET.sub.-- ATTR	
	* DFS.sub.-- LOOKUP
* DFS.sub.-- CHNG.sub.-- SYNC.sub.-- MODE	
	System Calls From Server
Processes:	
* OPEN	* CLOSE
* CREAT	* STAT
* FULLSTAT	* CHMOD
* EXIT	

The above UNIX operating system and AIX operating system operations correspond to the following vn.sub.-- ops:

* vn.sub.-- open	* vn.sub.-- create
* vn.sub.-- close	* vn.sub.-- getattr
* vn.sub.-- setattr	* vn.sub.-- lookup

An example of a vn.sub.-- ops execution is discussed below and shown in FIG. 19. The operation (an open) is executed at a client node and locks the s.sub.-- inode as usual if any local processing is necessary. If one of the above listed RPCs (dfs.sub.-- open) is sent to the server, then the s.sub.-- inode is unlocked before the RPC is sent. In the server, the RPC request will lock the fas.sub.-- lock or wait on it if it is busy, then it will lock the inode for a file F. If it is a local server operation, the executing process will acquire the fas.sub.-- lock and then lock the inode. If a DFS.sub.-- CHNG.sub.-- SYNC.sub.-- MODE or a DFS.sub.-- GET.sub.-- ATTR RPC is sent from a server to a client, then the

inode is unlocked before sending the RPC. Thus, the server can accept read and write operations after the RPC is sent. When the response messages from all the clients are received, the server locks the inode to finish any remaining local processing. If the operation was initiated at a client node, an acknowledgement is sent to that client. The inode is then unlocked and the fas.sub.-- lock is released.

The fas.sub.-- lock provides the means to synchronize the inode use and avoid deadlock operations. The fas.sub.-- lock synchronizes the access to the file access structure list containing information about the nodes that have the file open. The fas.sub.-- lock is locked when a process opens or closes a file, or when the file access structure list is being interrogated. The fas.sub.-- lock is also locked during a write operation when the file is in FULLSYNC mode.

The inode lock synchronizes access to the data in the file at the server. The inode lock is locked during a read or write operation to the file in the server. The inode lock is unlocked before the server sends a remote procedure call to the client, if the remote procedure call requires a lock on the s.sub.-- inode lock.

The s.sub.-- inode lock synchronizes the file access within a client process. The s.sub.-- inode is locked if an operation at the client is accessing the file. The s.sub.-- inode is unlocked if the operation at the client locks the fas.sub.-- lock. The s.sub.-- inode is also unlocked before a remote procedure call is sent from the client to the server.

If both the fas.sub.-- lock and the inode lock are locked, the inode lock is unlocked before the server originates a remote procedure call to the client. This allows the fas.sub.-- lock and s.sub.-- inode lock to be locked. The locks are unlocked in the reverse sequence that they became locked.

While the invention has been described in terms of a preferred embodiment in a specific operating system environment, those skilled in the art will recognize that the invention can be practiced, with modification, in other and different operating systems within the spirit and scope of the appended claims.

\* \* \* \* \*

