



US006073209A

United States Patent [19]
Bergsten

[11] Patent Number: 6,073,209
[45] Date of Patent: Jun. 6, 2000

[54] DATA STORAGE CONTROLLER
PROVIDING MULTIPLE HOSTS WITH
ACCESS TO MULTIPLE STORAGE
SUBSYSTEMS

5,794,013 8/1998 McBrearty 395/500.48
5,926,833 7/1999 Rasulian et al. 711/147

FOREIGN PATENT DOCUMENTS

PCT/US93/
05853 6/1993 WIPO .
WO 94/00816 1/1994 WIPO .
PCT/US94/
01447 2/1994 WIPO .
WO 94/19743 9/1994 WIPO .

OTHER PUBLICATIONS

Brochure, "CacheXchange: General Description, Legacy Series 690/9090," Andor International, Ltd., 1991.

Primary Examiner—Hiep T. Nguyen
Attorney, Agent, or Firm—Blakely, Sokoloff, Taylor & Zafman LLP

[75] Inventor: James R. Bergsten, Saratoga, Calif.
[73] Assignee: Ark Research Corporation, San Jose, Calif.

[21] Appl. No.: 08/828,888
[22] Filed: Mar. 31, 1997

[51] Int. Cl.⁷ G06F 12/02
[52] U.S. Cl. 711/114; 711/148; 711/153;
711/162; 707/204; 714/204
[58] Field of Search 707/204; 711/162,
711/148, 114, 153; 714/6

[56] References Cited

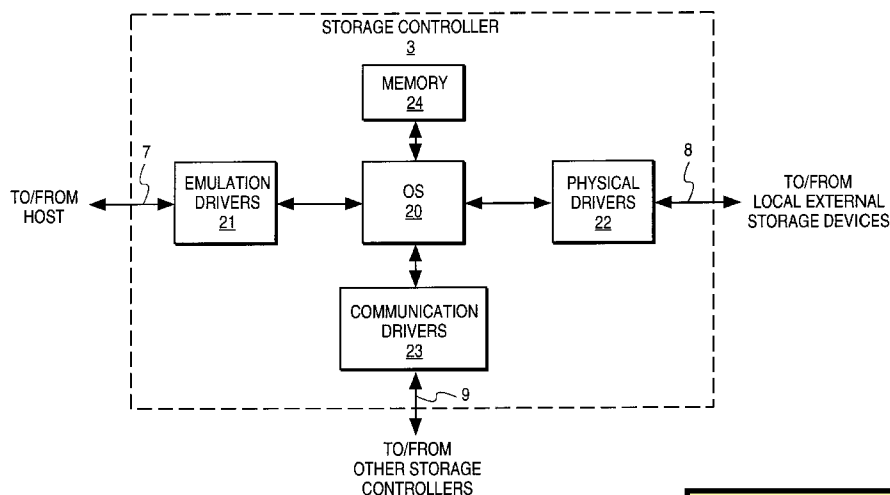
U.S. PATENT DOCUMENTS

3,889,237 6/1975 Alferness et al. 711/148
4,718,006 1/1988 Nishida 711/148
5,155,845 10/1992 Beal et al. 395/575
5,408,629 4/1995 Tsuchiya et al. 711/151
5,455,932 10/1995 Major et al. 711/162
5,459,842 10/1995 Begun et al. .
5,459,857 10/1995 Ludlam et al. 395/182.04
5,481,276 1/1996 Dickey et al. .
5,504,882 4/1996 Chai et al. 395/182.03
5,530,845 6/1996 Hiatt et al. .
5,588,132 12/1996 Cardoza 711/148
5,615,364 3/1997 Marks 707/202
5,673,381 9/1997 Huai et al. 395/180
5,673,382 9/1997 Cannon et al. 395/182.04
5,675,802 10/1997 Allen et al. 395/703
5,680,580 10/1997 Beardsley et al. 711/162
5,692,147 11/1997 Larson et al. .
5,699,510 12/1997 Petersen et al. 385/185.07
5,708,769 1/1998 Stallmo 714/6
5,729,735 3/1998 Mayering 707/10
5,740,433 4/1998 Carr et al. 707/202
5,742,792 4/1998 Yanai et al. 711/162
5,765,200 6/1998 McIlvain et al. 711/170
5,768,623 6/1998 Judd et al. 395/857

[57] ABSTRACT

A computer network comprises a number of storage controllers, each coupled to one of a plurality of storage arrays, each storage array including at least one mass storage device. Each storage controller may be coupled to at least one host processing system and to at least one other storage controller to control access of the host processing systems to the mass storage devices. Multiple copies of data are maintained in storage arrays that are geographically remote to each other, such that any copy can be accessed by any host. Each storage controllers includes an interface with a host that emulates a mass storage device and an interface with a local storage array that emulates a host. The interfaces to the host and local storage arrays are independent of the type of host or devices in the local storage array. Two or more hosts may be dissimilar to each other, and two or more storage arrays may include dissimilar mass storage devices. Hosts access stored data using virtual addressing. During a data access, the storage controller connected to the accessing host maps a virtual address provided by the host to a real physical location in any of the storage arrays, such that the actual location of the data is transparent to the host. The storage controllers provide automatic back-up and error correction as well as write protection of back-up copies.

40 Claims, 18 Drawing Sheets



MICROSOFT CORP.
EXHIBIT 1004

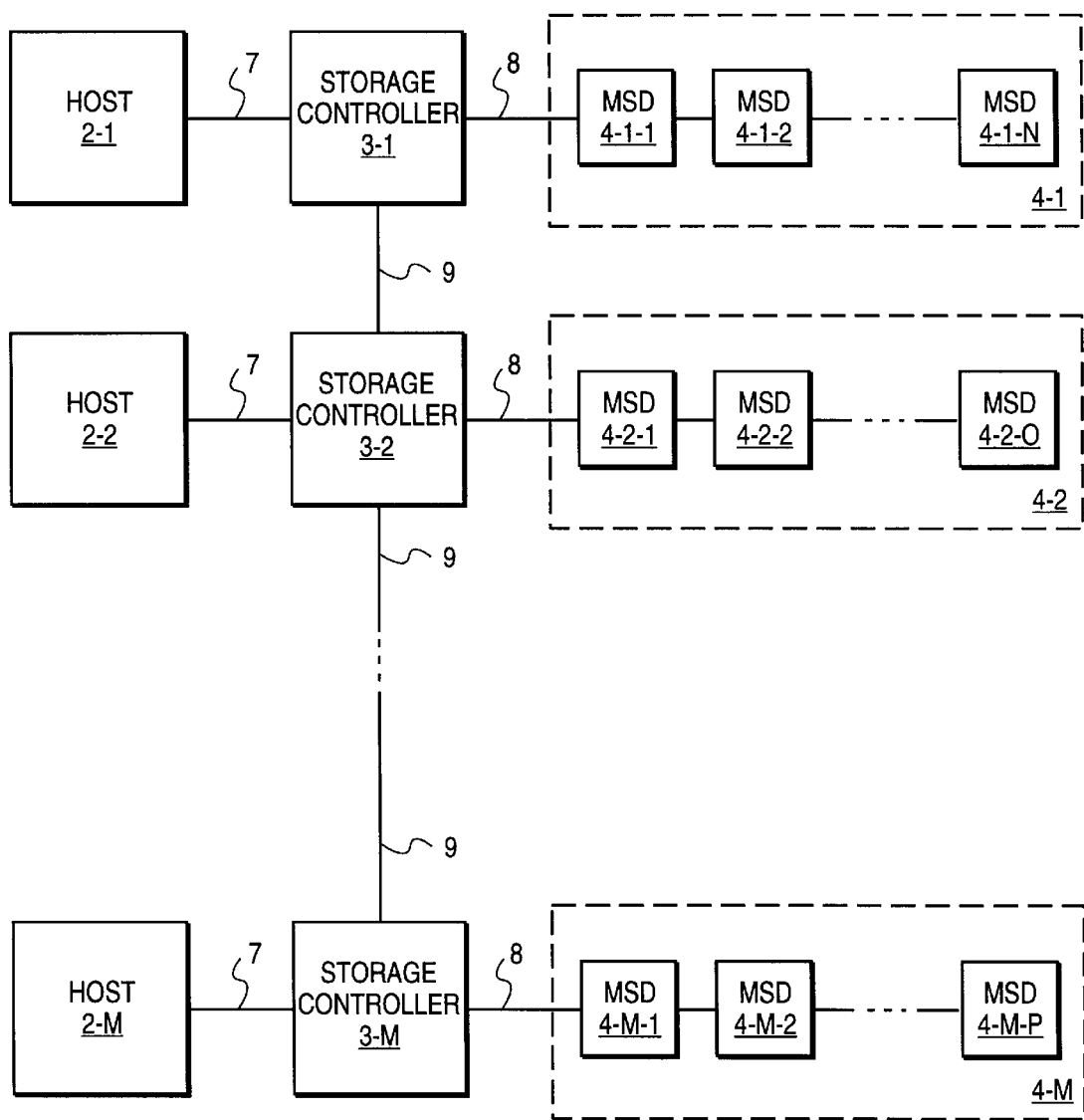


FIG. 1

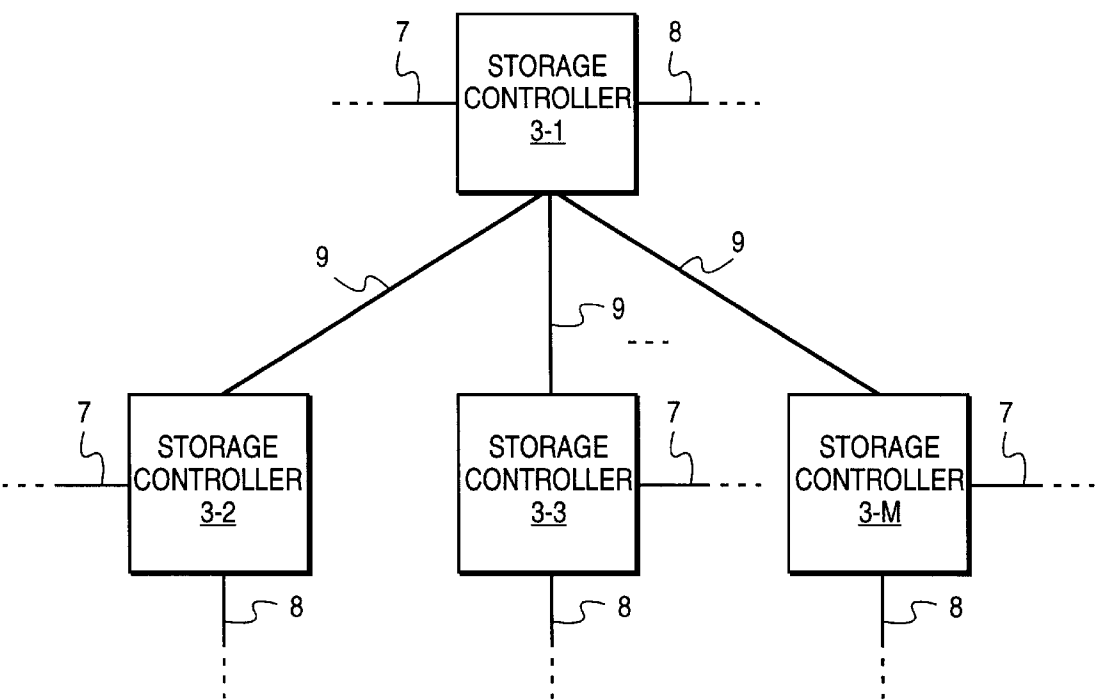


FIG. 2

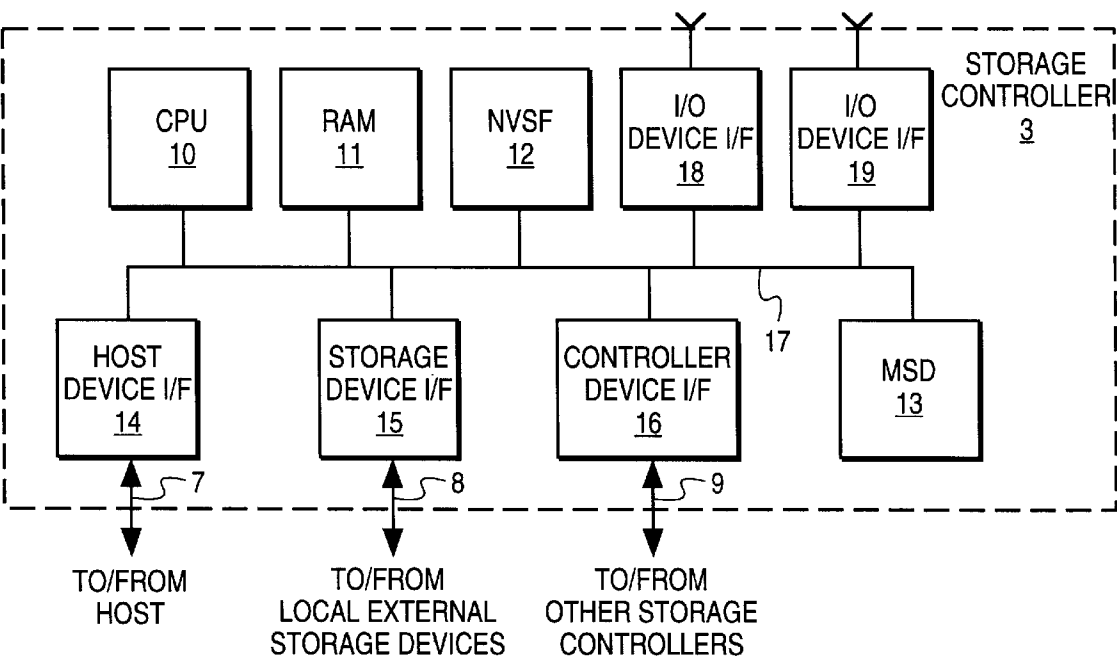


FIG. 3

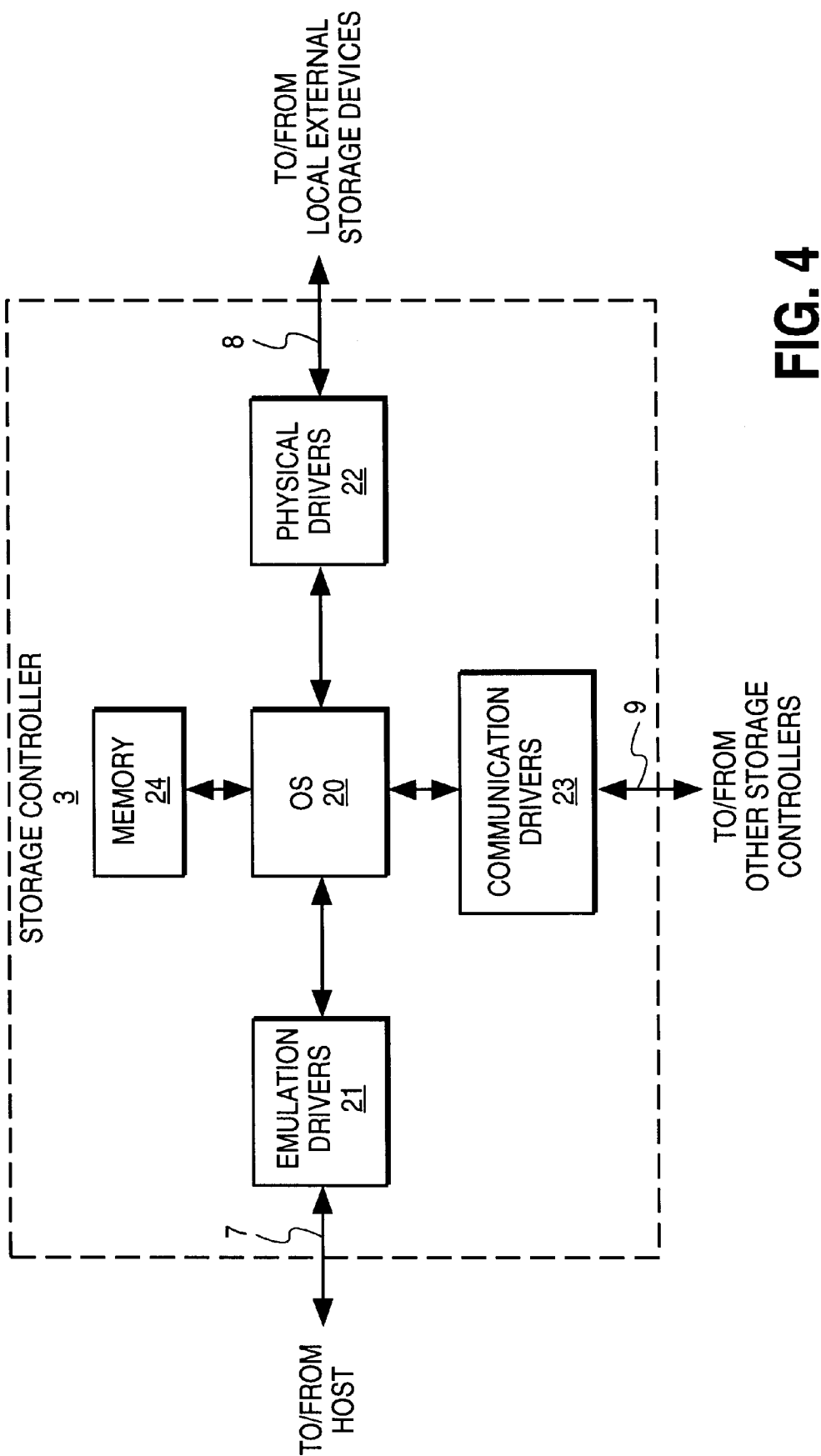


FIG. 4

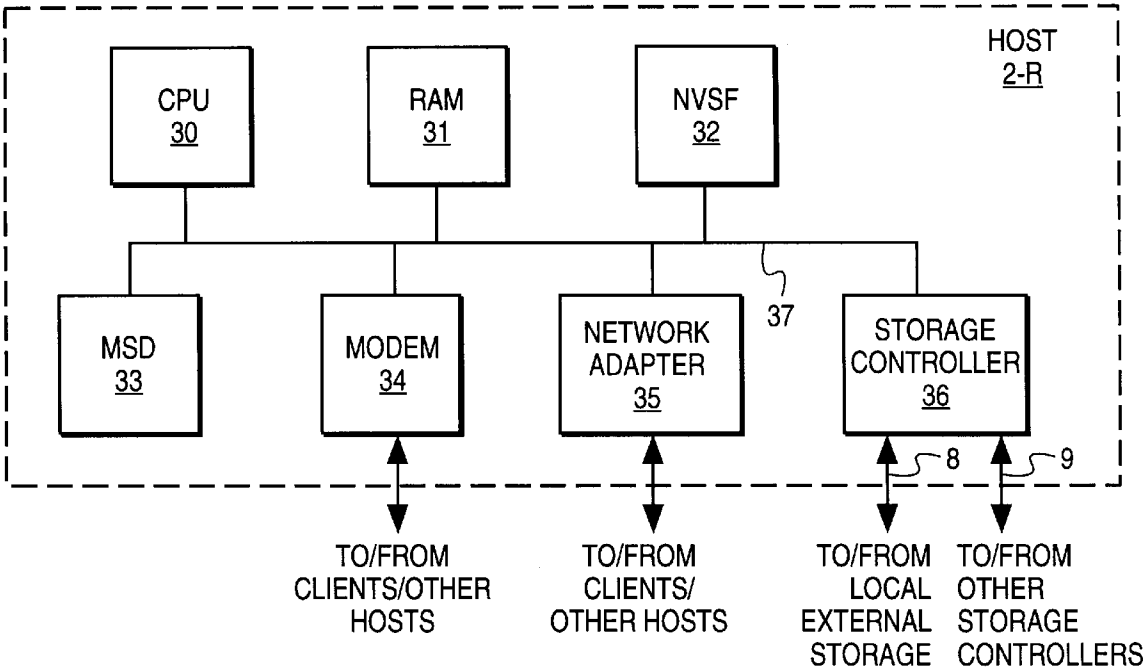


FIG. 5

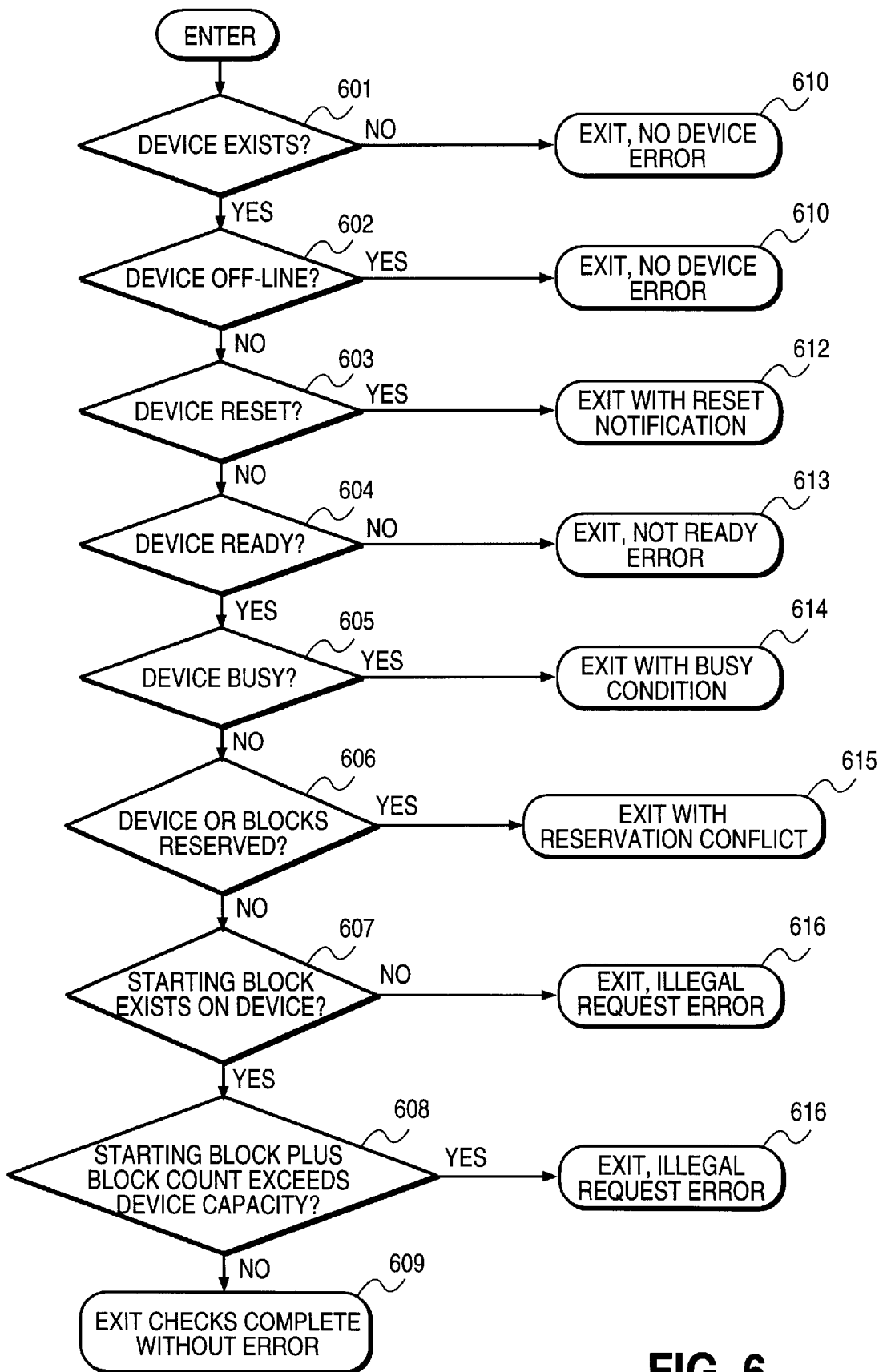
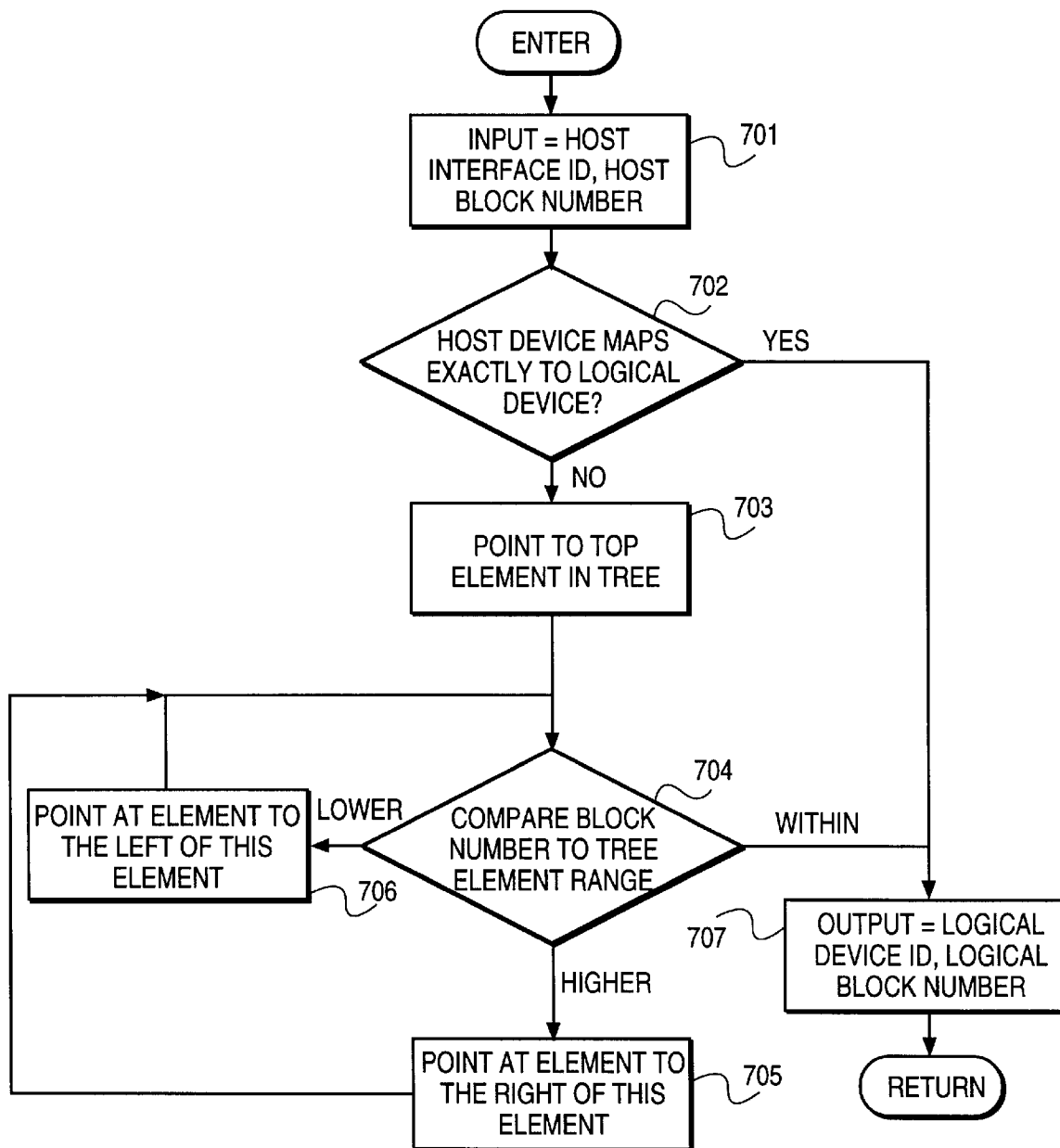


FIG. 6

**FIG. 7**

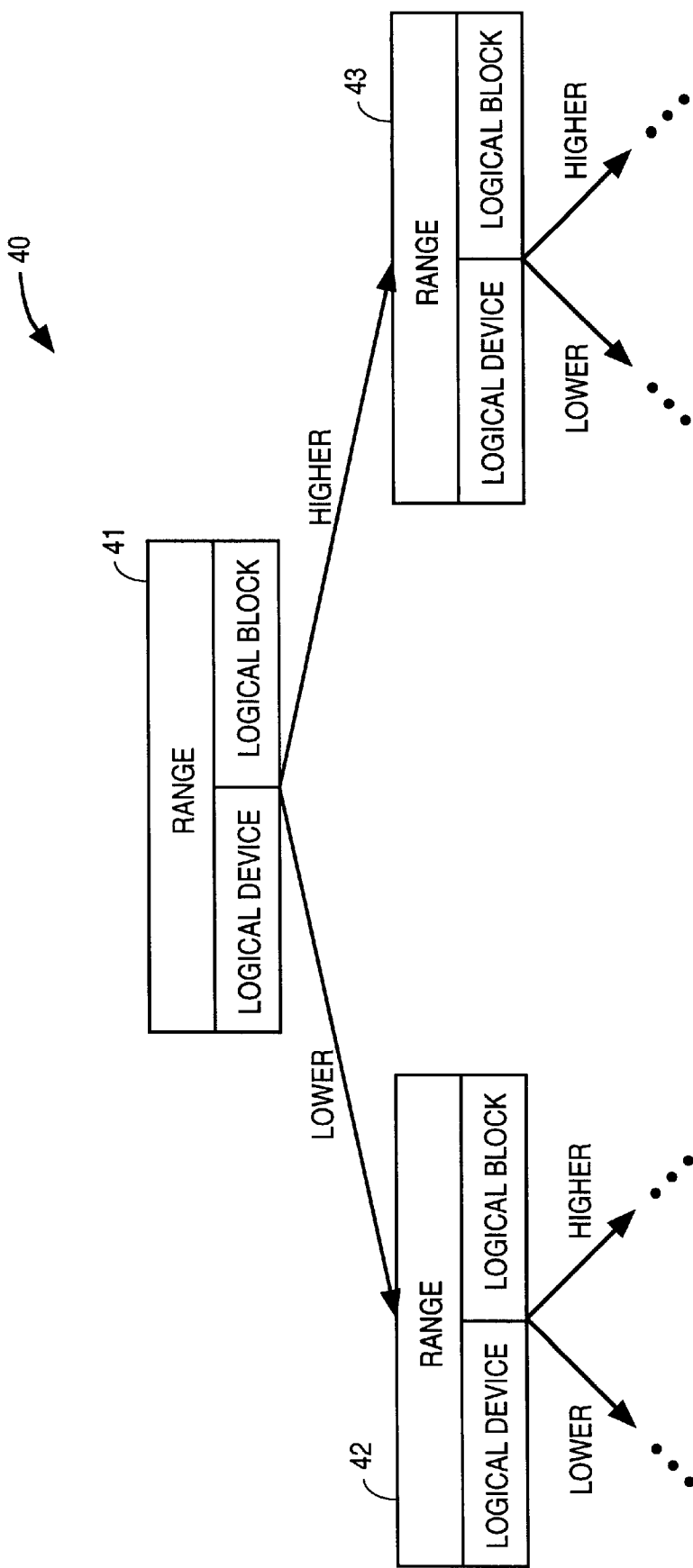
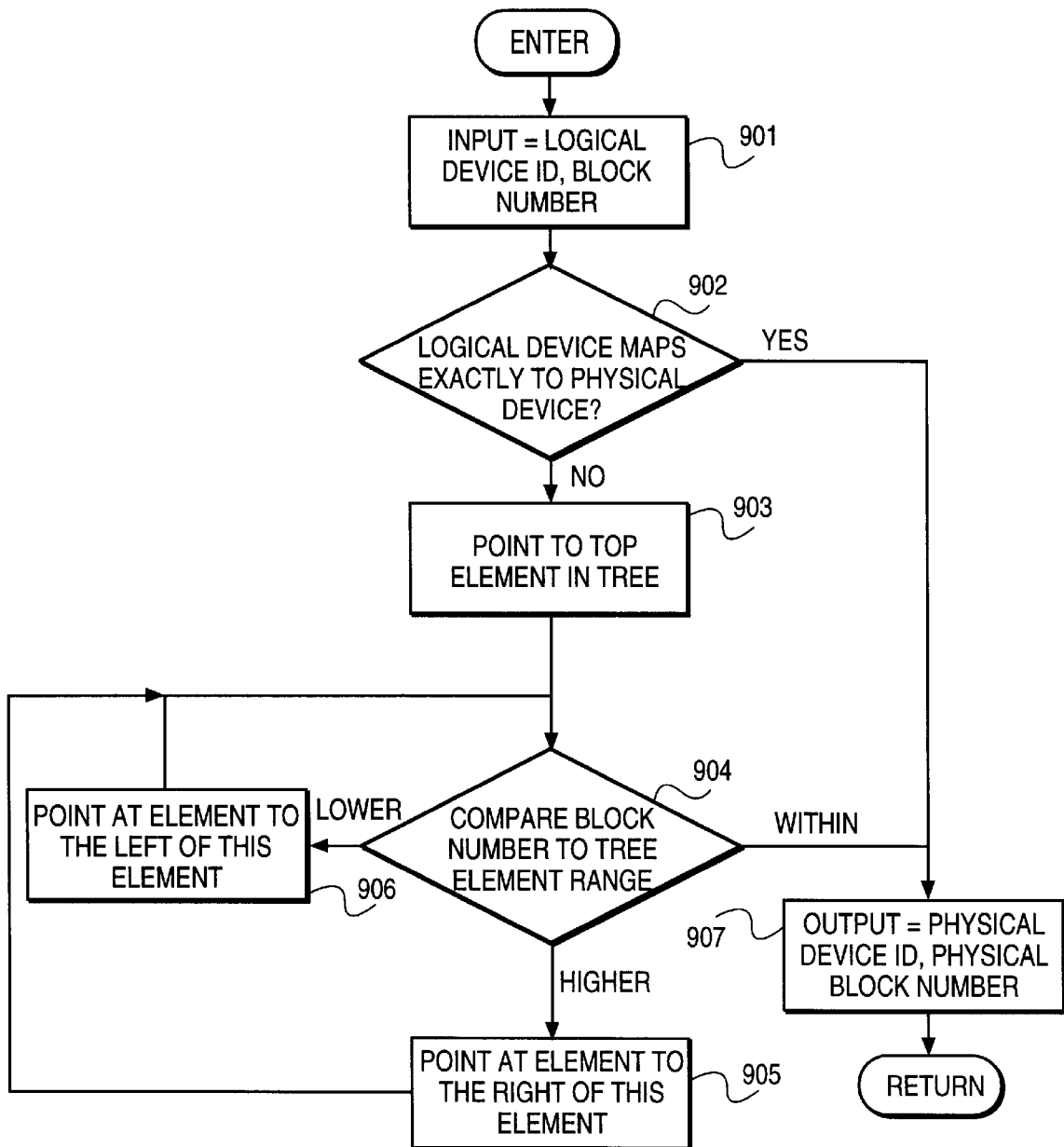


FIG. 8

**FIG. 9**

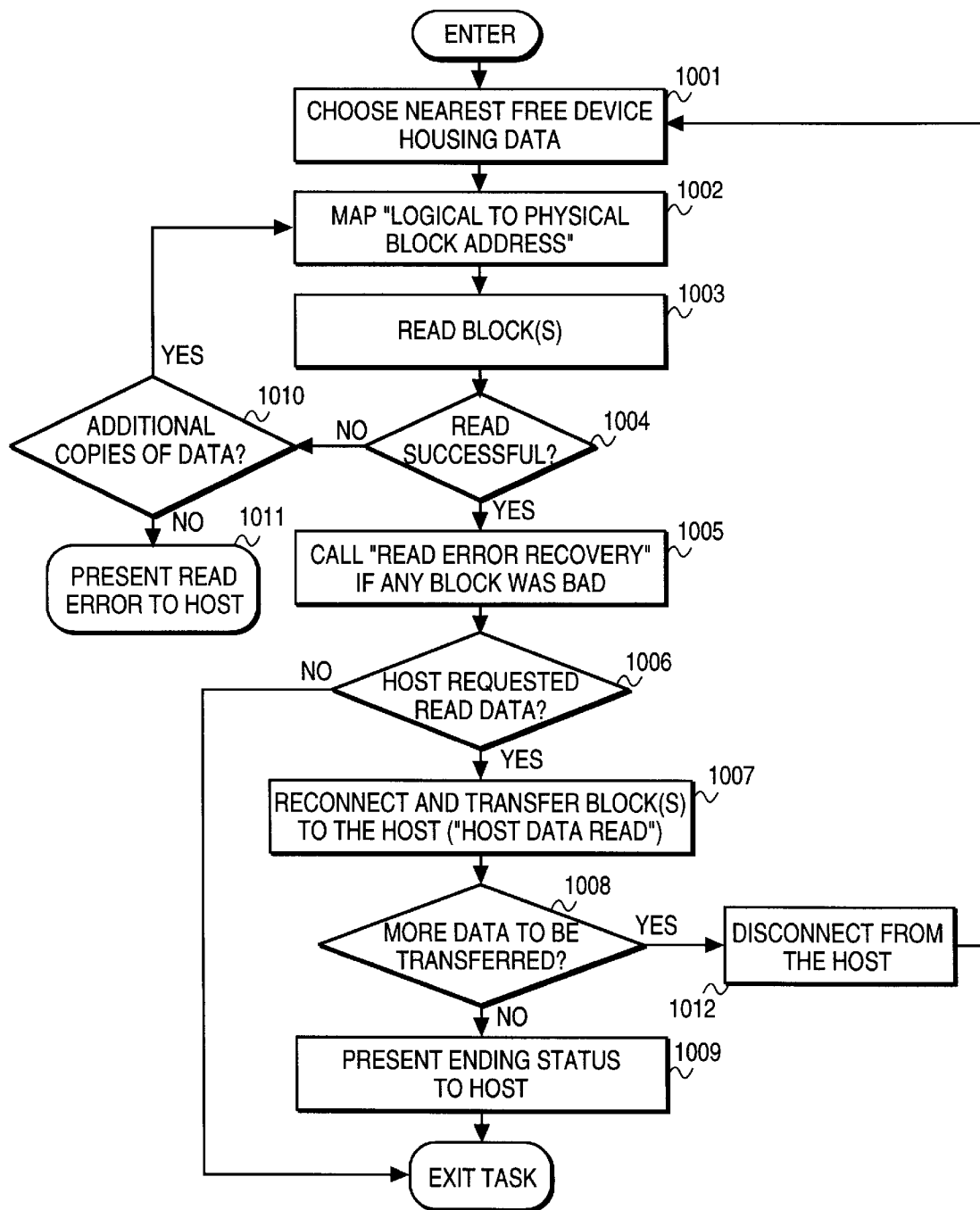


FIG. 10

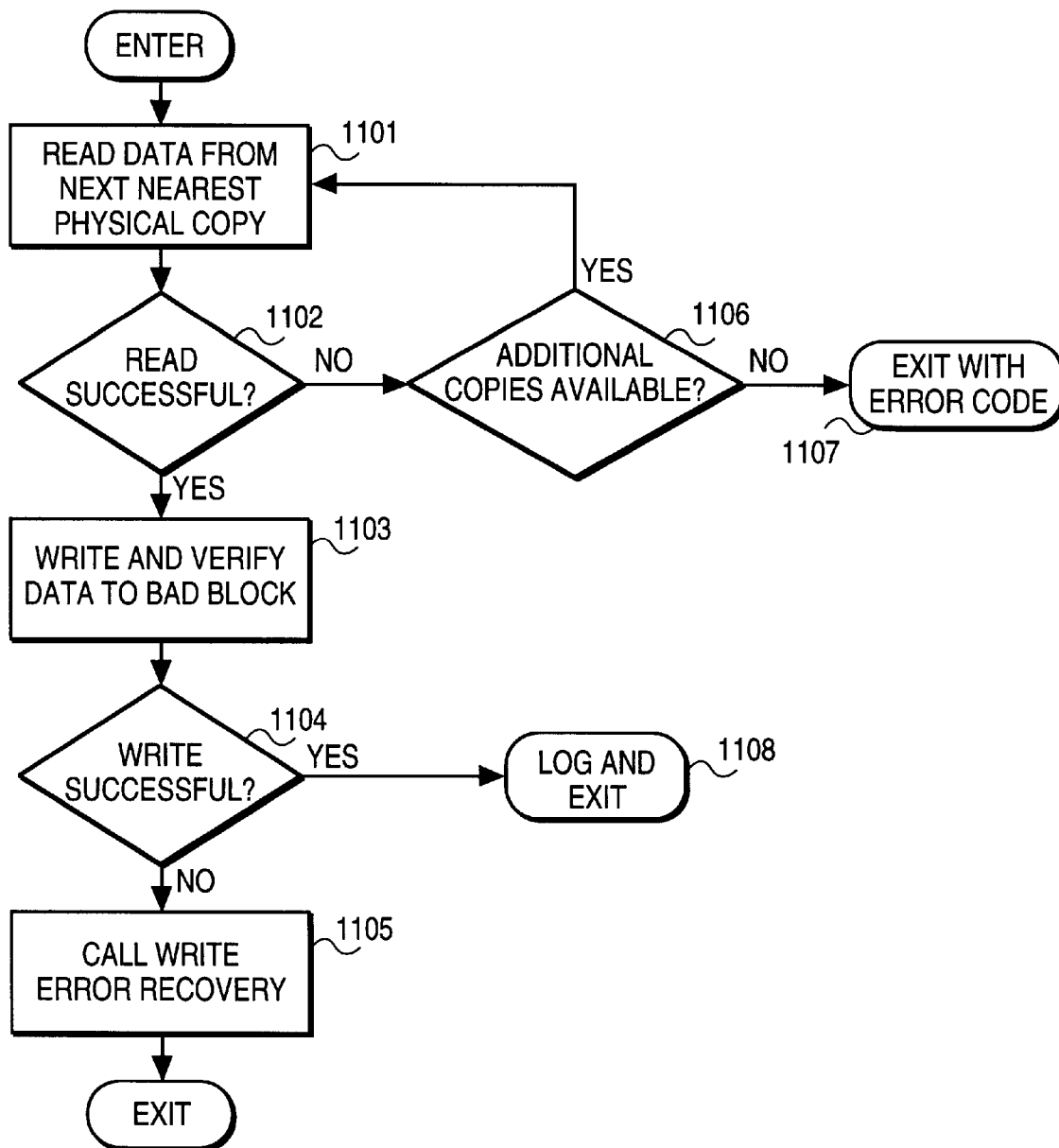


FIG. 11

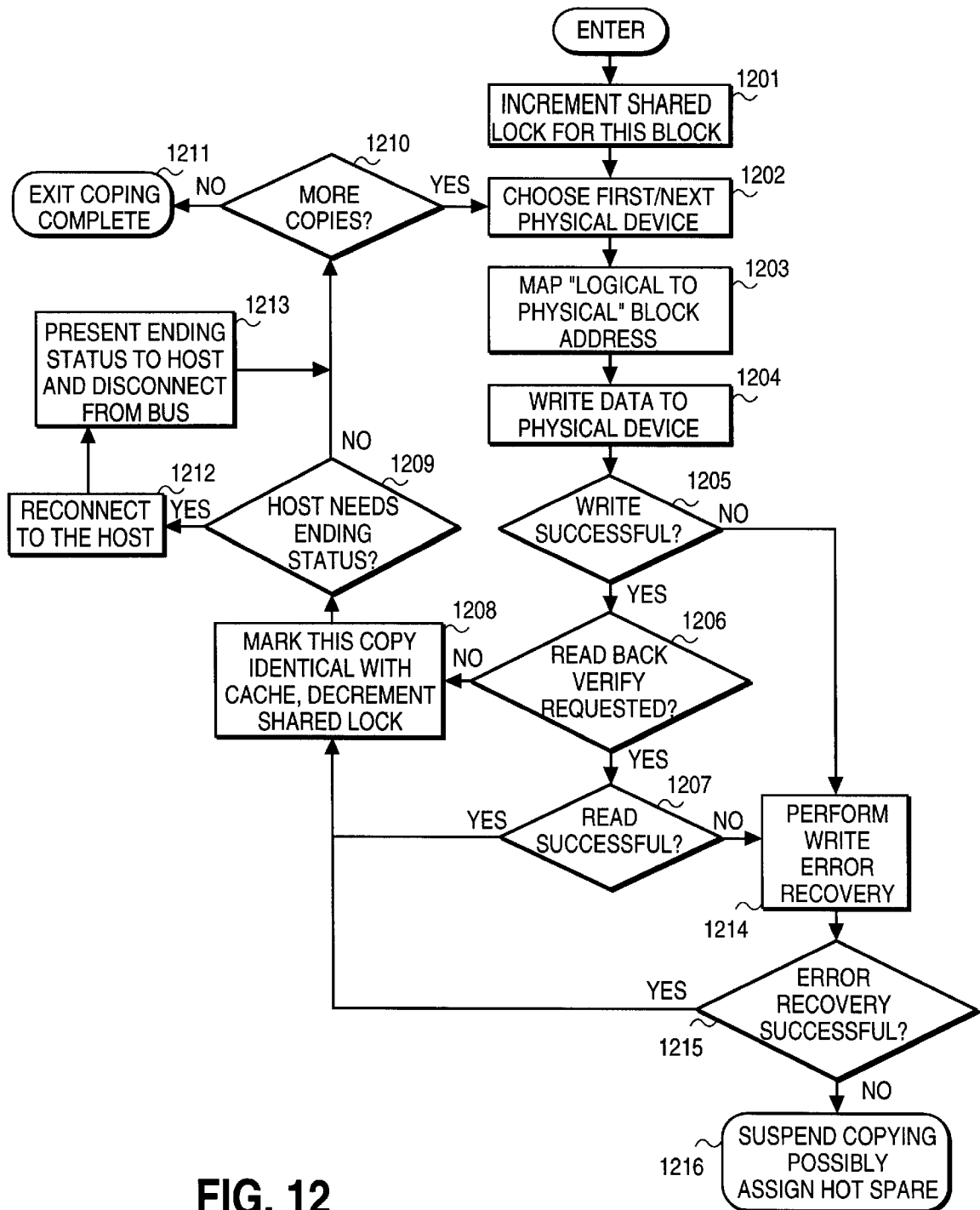
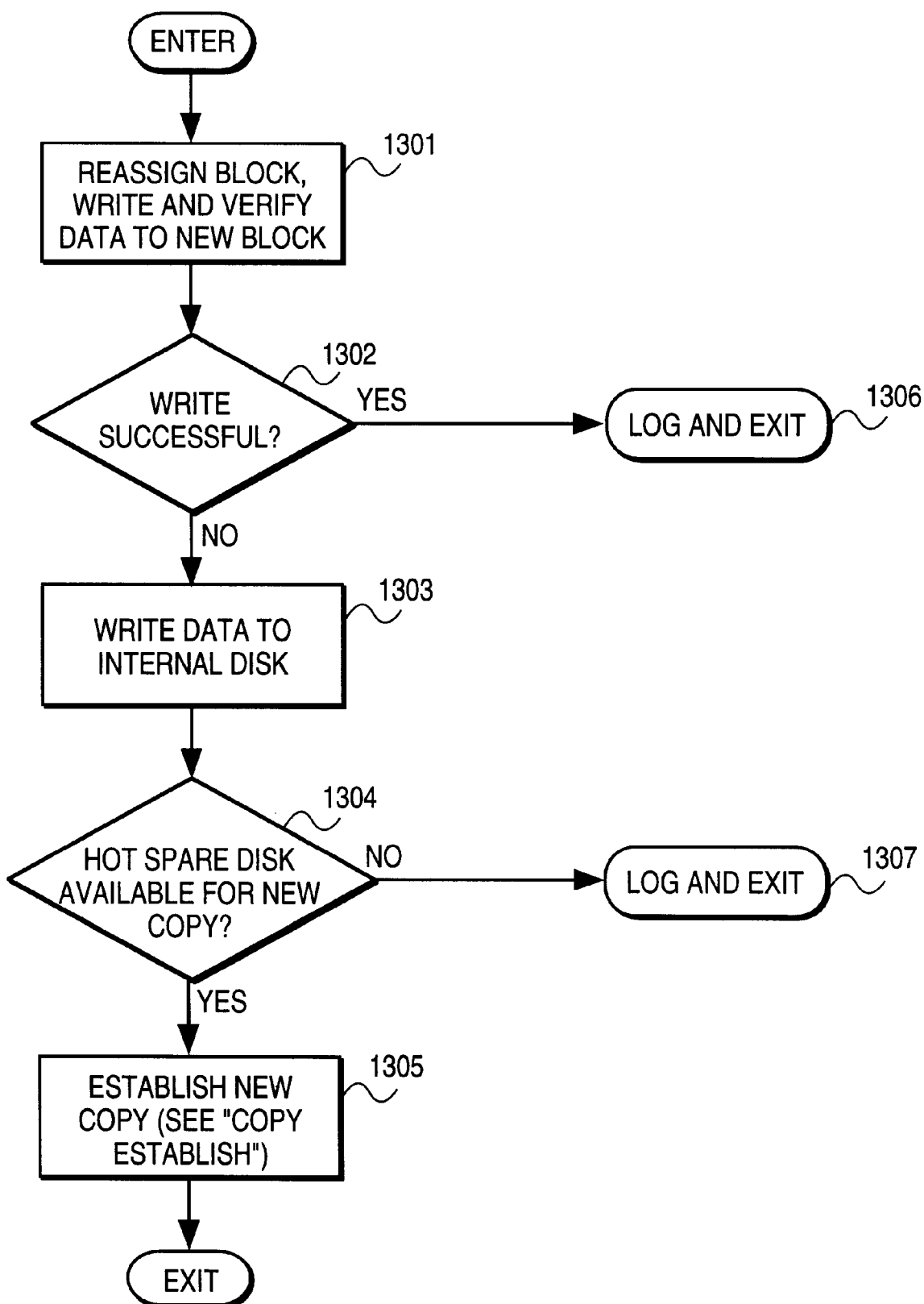


FIG. 12

**FIG. 13**

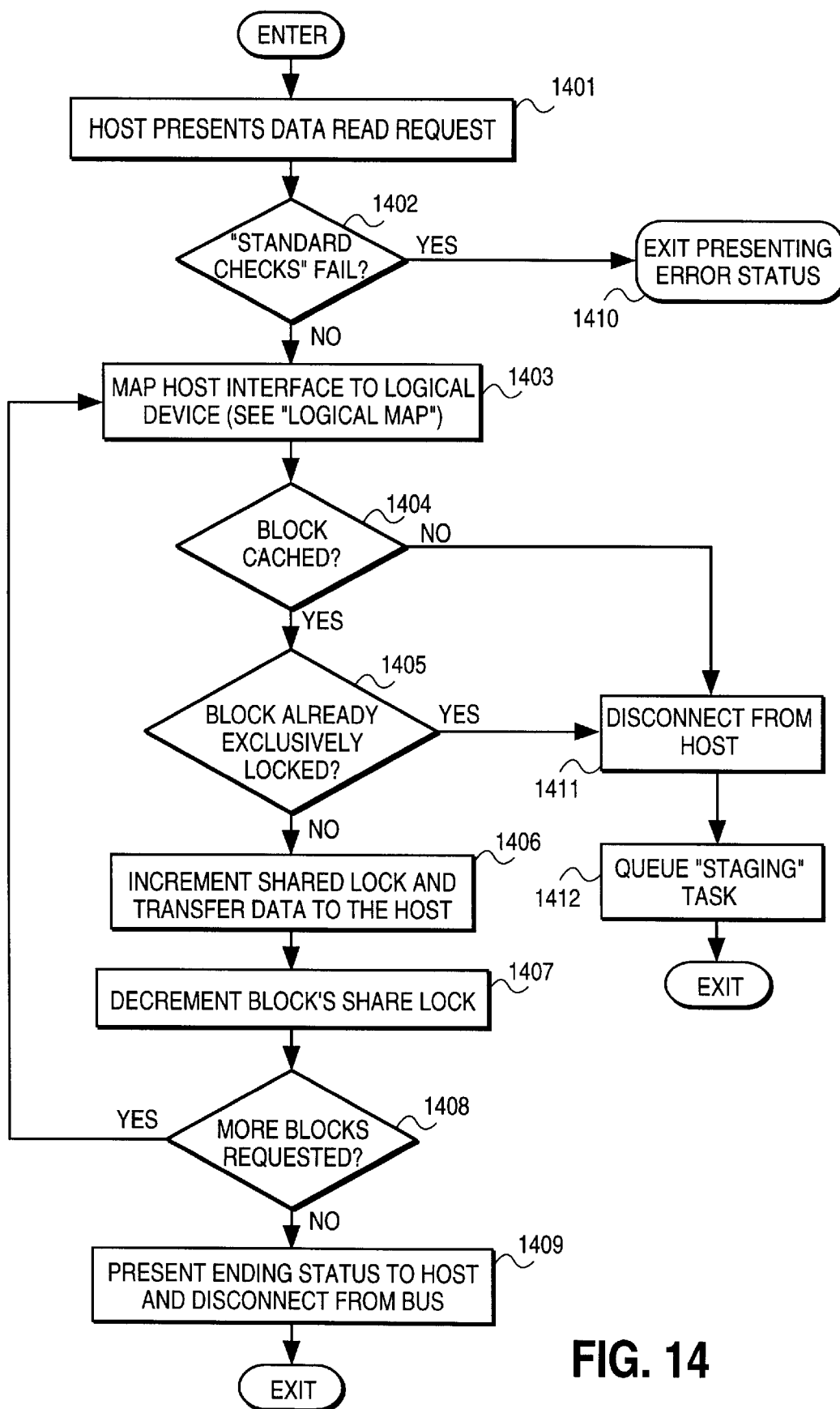
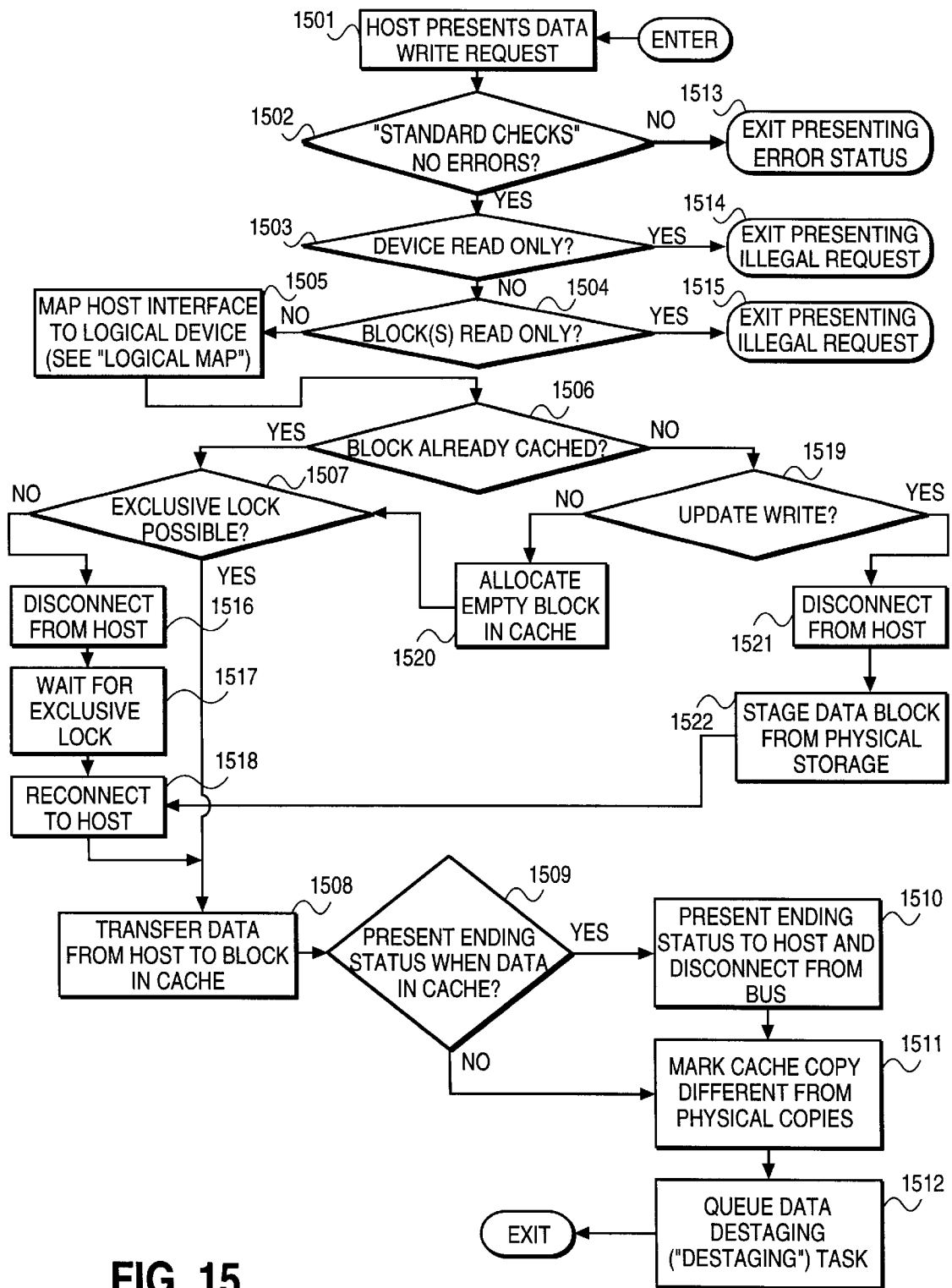


FIG. 14



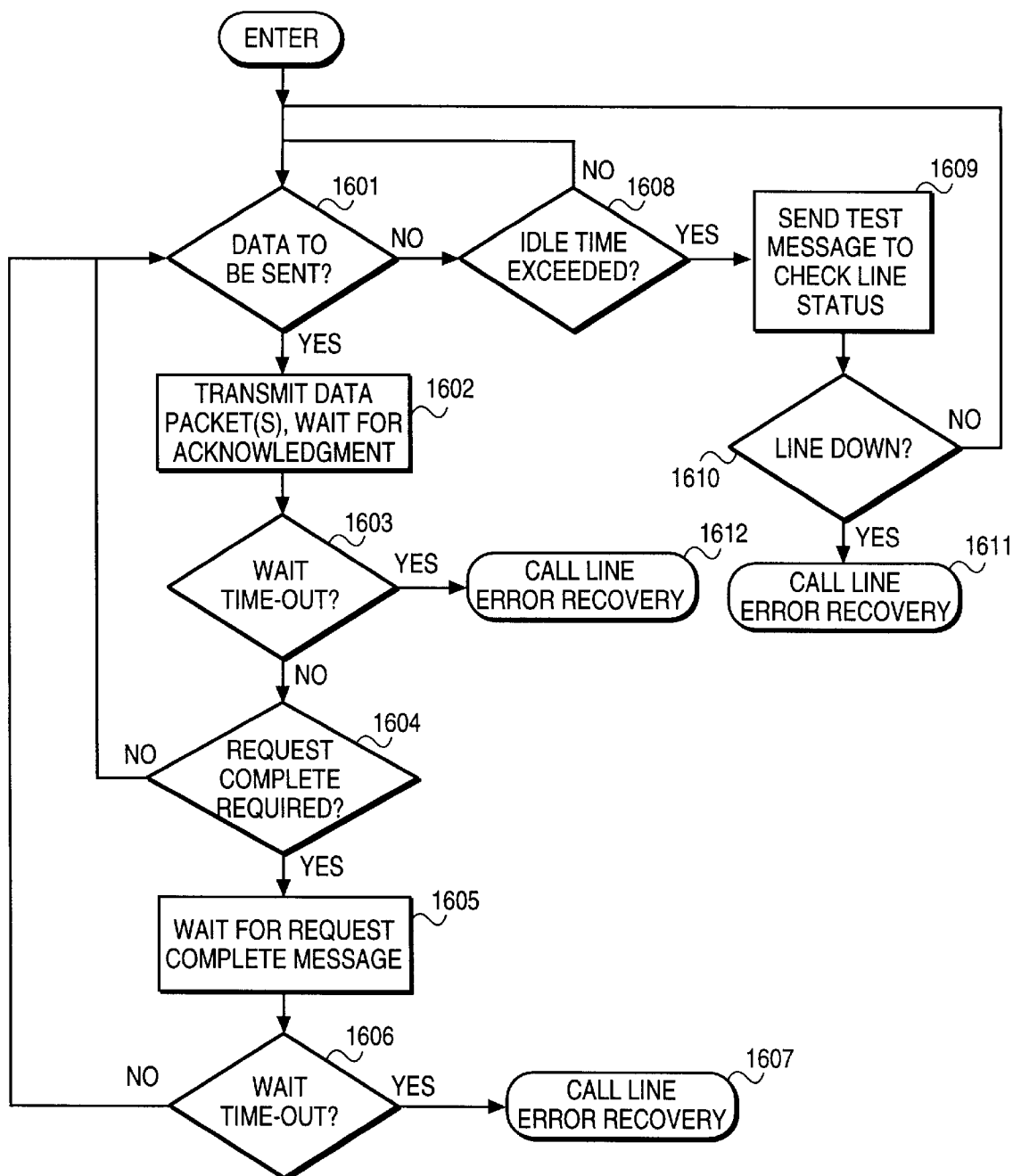
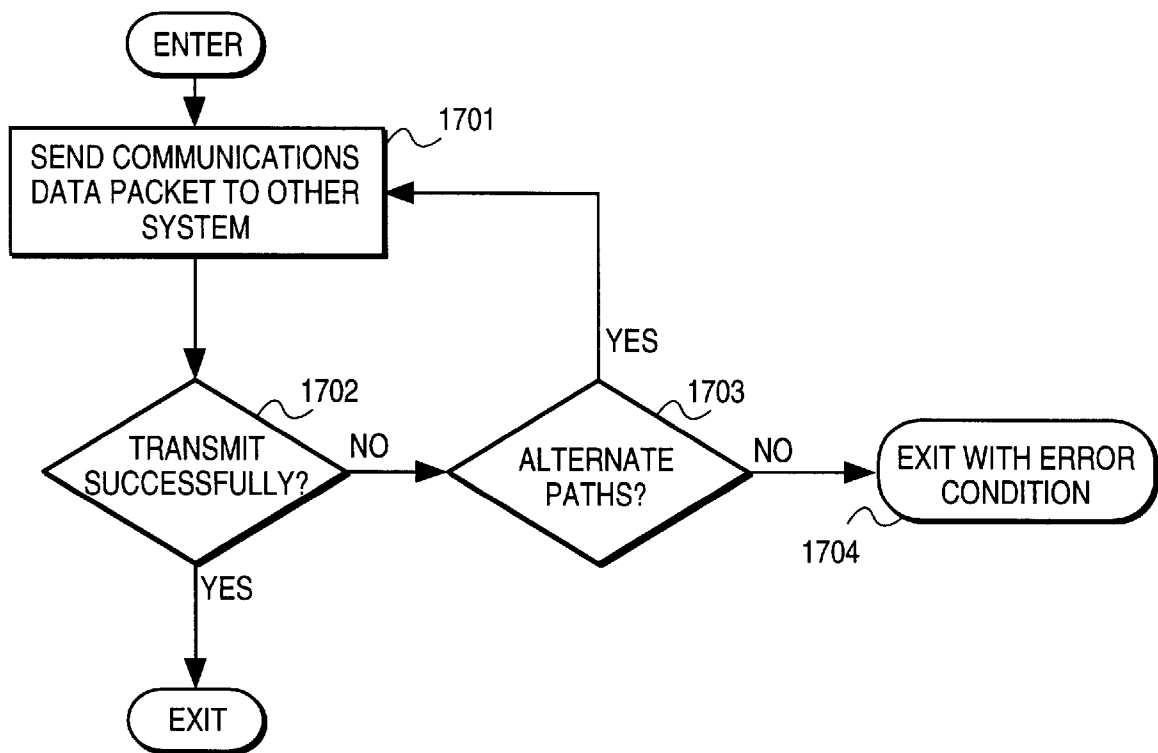
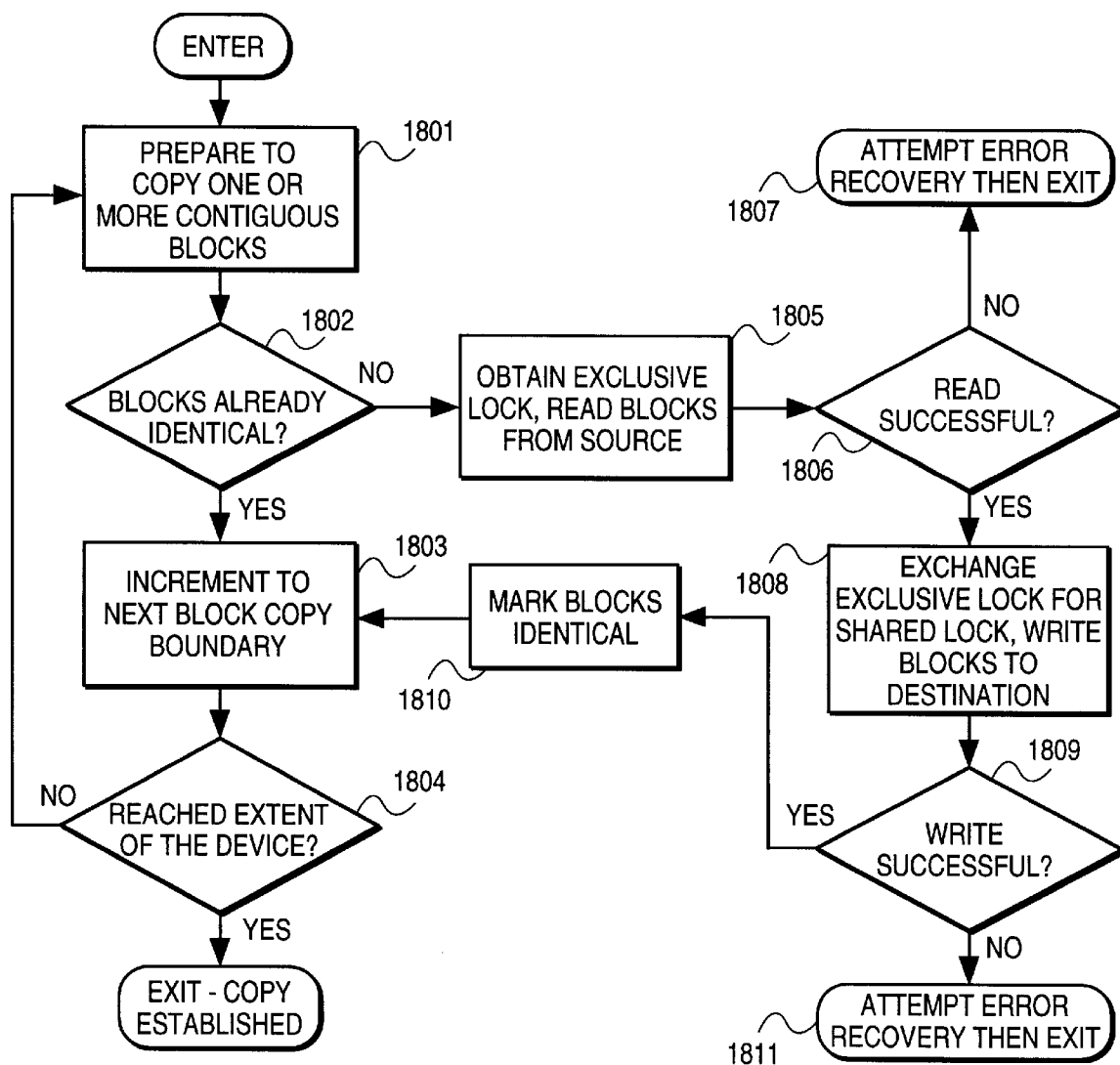
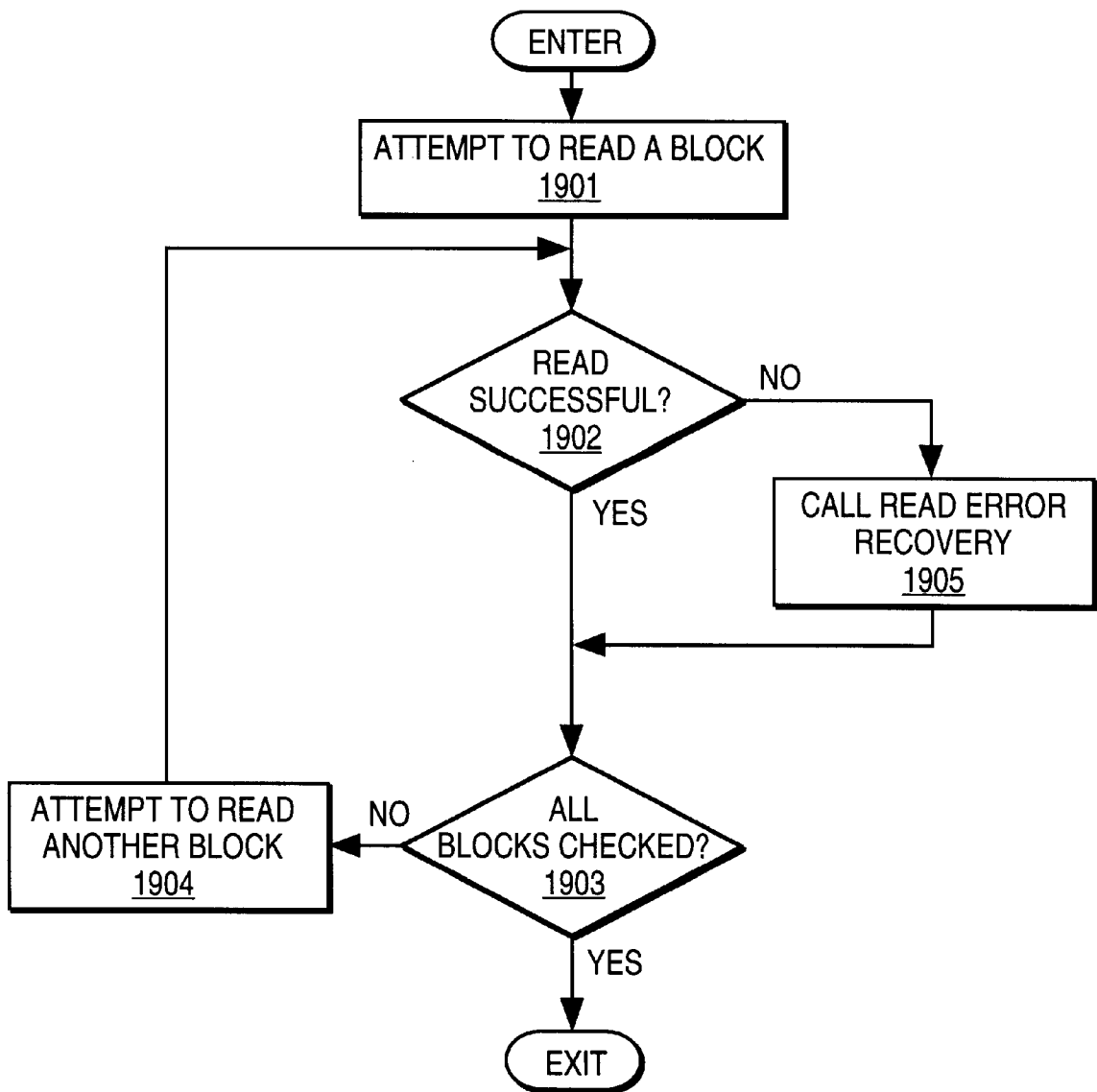


FIG. 16

**FIG. 17**

**FIG. 18**

**FIG. 19**

DATA STORAGE CONTROLLER PROVIDING MULTIPLE HOSTS WITH ACCESS TO MULTIPLE STORAGE SUBSYSTEMS

FIELD OF THE INVENTION

The present invention pertains to the field of computer systems. More particularly, the present invention relates to devices which control data flow between one or more host processing systems and one or more data storage subsystems.

BACKGROUND OF THE INVENTION

Computer systems sometimes handle valuable or irreplaceable data. Data maintained by a computer system may be of vital importance, for example, in business applications such as airline reservations, bank account management, electronic funds transfers, shipping and receiving, and inventory control. Consequently, there is a need to ensure that valuable data is adequately protected against loss or damage.

It is common to store large volumes of data on non-volatile mass storage devices, such as magnetic or optical disks. Occasionally, however, a mass storage device will fail, resulting in the loss of stored data. Consequently, it is a common practice to store a back-up copy of data that is considered valuable on a separate, back-up storage device. For practical reasons, however, it is often necessary to locate the back-up storage device in the same geographic vicinity as the primary storage device. As a result, both the primary data and the back-up copy may be lost or destroyed due to theft, vandalism, or natural disasters, such as fires, floods, and earthquakes. Hence, there is a need for the capability to store multiple back-up copies of data in geographically separate locations, while still permitting quick and easy access by a host computer to any back-up copy.

In particular, it is desirable to provide a storage controller which provides the above-mentioned functionality. It is further desirable that such a storage controller allow multiple host computers at different locations to access any copy of stored data. It is further desirable that such a storage controller automatically create and manage multiple back-up copies while the host computers are "on line" in a manner that is transparent to the host computers and their users. It is further desirable that such a storage controller automatically detect and correct errors in stored data and automatically replace faulty copies. It is further desirable that such a storage controller not be dependent upon any particular hardware or software configuration of any host computer or mass storage device which it services.

SUMMARY OF THE INVENTION

The present invention includes a storage controller that includes a data interface with a host processing system and a data interface with at least one storage array. Each storage array includes at least one mass storage device. The host interface is independent of a device type of the host processing system, and the storage interface is independent of the device type of the mass storage devices. The data storage controller further includes control circuitry coupled to the host interface and the storage interface. The control circuitry is configured to enable the host processing system to access each of the storage arrays.

Other features of the present invention will be apparent from the accompanying drawings and from the detailed description which follows.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate, similar elements and in which:

FIG. 1 illustrates a computing system in which a number of storage controllers provide multiple host computers with access to multiple storage arrays.

FIG. 2 illustrates a computing system according to an embodiment in which one storage controller is connected to a number of other storage controllers.

FIG. 3 is a block diagram of a storage controller.

FIG. 4 is a block diagram illustrating functional modules of a storage controller.

FIG. 5 is a block diagram illustrating a host computer system including a storage controller according to the present invention.

FIG. 6 is a flow diagram illustrating a check routine performed in response to an input/output request.

FIG. 7 is a flow diagram illustrating a routine for mapping a host address to a logical address.

FIG. 8 illustrates a tree hierarchy used by a storage controller to map a host address to a logical address.

FIG. 9 is a flow diagram illustrating a routine for mapping a logical address to a physical address.

FIG. 10 is a flow diagram illustrating a routine performed by a storage controller for staging data from a mass storage device.

FIG. 11 is a flow diagram illustrating a read error recovery routine.

FIG. 12 is a flow diagram illustrating a routine performed by a storage controller for destaging data to a mass storage device.

FIG. 13 is a flow diagram illustrating a write error recovery routine.

FIG. 14 is a flow diagram illustrating a routine performed by a storage controller for reading data from a mass storage device in response to a request from a host.

FIG. 15 is a flow diagram illustrating a routine performed by a storage controller for writing data to a mass storage device in response to a request from a host.

FIG. 16 is a flow diagram illustrating a routine performed by a storage controller for transmitting data to a remote storage controller.

FIG. 17 is a flow diagram illustrating a remote communications error recovery routine.

FIG. 18 is a flow diagram illustrating a routine for establishing copies.

FIG. 19 is a flow diagram illustrating a routine performed by a storage controller for performing automatic detection and correction of data errors.

DETAILED DESCRIPTION

A storage controller capable of providing multiple host computers system with access to multiple storage arrays is described. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be evident, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to facilitate description of the present invention.

As will be described below in detail, the present invention includes a storage controller which allows multiple host computer systems at different locations to access any of multiple copies of stored data. The storage controller automatically creates and manages multiple back-up copies while the host computer systems are "on line" in a manner that is both non-disruptive of, and transparent to, the host computer systems and their users. Further, the storage controller automatically detects and corrects errors in stored data and automatically replaces faulty copies. Moreover, the storage controller is not dependent upon any particular hardware or software configuration of the host computer system which it services or the mass storage devices which it accesses. The storage controller emulates a local storage array for the host computer system which it services and emulates a host computer system for the local storage array which it accesses. Host computer systems access stored data using virtual device addresses, which are mapped to real device addresses by the storage controller.

FIG. 1 illustrates a computing system in which a number of storage controllers of the present invention provide a number of host computer systems with access to a number of storage arrays. Specifically, the computing system includes M storage controllers, 3-1 through 3-M; M host computers, 2-1 through 2-M, which are coupled to storage controllers 3-1 through 3-M, respectively; and M storage arrays, 4-1 through 4-M, which are coupled to the storage controllers 3-1 through 3-M, respectively. Each of the storage arrays includes a number of mass storage devices (MSDs) coupled to a storage controller in a daisy chain configuration. Specifically, storage array 4-1 includes N MSDs, 4-1-1 through 4-1-N; storage array 4-2 includes O MSDs, 4-2-1 through 4-2-O; and, storage array 4-M includes P MSDs, 4-M-1 through 4-M-P.

Each of the storage controllers is coupled to another storage controller via a communication link 9. A portion of a communication link 9 between two geographically-separated storage controllers may be provided by a local area network (LAN). For example, in one embodiment, a given communication link 9 may be implemented partly on a Fast Ethernet; other portions of the link 9 can be implemented as an ATM (Asynchronous Transfer Mode) link, a T1 or T3 link, an FDDI (Fiber Distributed Data Interface) link, or any other suitable type of link.

Note that any of the data communication paths 7, 8, and 9 may actually consist of two or more redundant, physical paths. Therefore, a failure of any single physical connection does not affect the ability to access any stored data.

Each of the host computer systems 2 may be any conventional computer system. For example, a host computer system may be a personal computer (PC), a mini-computer, or a mainframe. In addition, any of the host computer systems may function as a server for one or more client computer systems (not shown).

Each MSD may include non-volatile facility for storing large volumes of data, such as a magnetic disk or tape, an optical storage device such as CD-ROM (Compact Disk-ROM), CD-R (CD-recordable), DVD (Digital Versatile Disk), a magneto-optical (MO) device, or the like. The MSDs within the computing system need not be of the same device type. That is, the MSDs in any given storage array may use a different type of storage medium from those in any other storage array.

Each storage array may be located geographically distant from the other storage arrays. Multiple copies are generally maintained on different, geographically-separated storage

arrays. Hence, the loss of one or more MSDs in a given storage array will not result in the complete loss of data. With respect to a given (local) storage controller, any or all of the other (remote) storage controllers, host computer systems, and storage arrays may therefore be located at distant locations to the local storage controller.

Storage controllers 3-1 through 3-M function cooperatively to provide any of host computer systems 2-1 through 2-M with access to any of storage arrays 4-2 through 4-M. Each one of storage controllers 3-1 through 3-M directly services one local host computer system and one local storage array in one embodiment. For example, in the embodiment shown in FIG. 1, storage controller 3-1 services and is directly coupled to its local host computer system 2-1 via a data communication path 7. Storage controller 3-1 also is directly coupled to its local data storage array 4-1 via another communication path 8.

In one embodiment, each storage controller communicates with its local host computer system and its local storage array using standard SCSI (Small Computer Systems Interface) protocols. Consequently, operation of a storage controller of the present invention in the manner described herein is not dependent upon the particular hardware or software configuration of any host computer or storage array, as long as those devices are SCSI-compatible. Note that in other embodiments, however, the data communication paths 7 and 8 may conform to other protocols and standards, such as serial SCSI, Fiber Channel, or ESCON. Thus, because data paths 7 and 8 are conventional interfaces, a storage controller of the present invention can be used concurrently with host computers and MSDs having different configurations. For example, one host computer in the system may be a mainframe computer system while another host computer is a PC. Similarly, one storage array in the system may consist of conventional magnetic hard disk drives while another storage array consists of CD-ROM drives.

The storage controllers in the computing system cooperate to allow any of the host computer systems to access data located in any of the storage arrays. For example, host computer system 2-1 may access data located on any MSD in storage array 4-1 by communicating with storage controller 3-1. In addition, host computer system 2-1 may access data located in remote storage array 4-M by communicating with remote storage controller 3-M via storage controllers 3-1, 3-2, etc. As will be explained below, data is accessed using virtual addressing, such that a host computer system has no knowledge of which physical storage device is being accessed.

The storage controllers operate in peer-to-peer relationships (as opposed to master-slave relationships) with each other when responding to remote access requests. Any storage controller can access data in a remote storage array serviced by any other storage controller. Communication between individual storage controllers takes place on communication links 9 using a common protocol implemented by all storage controllers in the system. Note that various different protocols might be used for this purpose. A description of the details of such a protocol is not necessary for a full understanding of the present invention and is therefore not included herein.

The remote access capabilities of the storage controllers can be used advantageously in a variety of ways. The storage controllers can be used to enable sharing and centralized control of storage media. For example, the MSDs in any of the storage arrays may include CD-ROM devices, so that

CD-ROMs in one a storage arrays may be shared by all host computers. Use of the devices can be recorded by one of the storage controllers. Access to CD-ROM data can therefore be monitored and controlled by a user, such as a system administrator, using the storage controller. The present invention also permits low-volume production of CD-Rs without specialized equipment. For example, a limited quantity of CD-R disks can be produced by writing to a single, logical CD-R drive corresponding to multiple physical CD-R devices within one or more storage arrays.

The present invention is not limited to the specific configuration shown in FIG. 1. For example, the system configuration might alternatively include only a single host computer which is coupled to multiple geographically-separated storage arrays using multiple storage controllers. In another embodiment, a storage controller may have a direct interface with more than one host or may have redundant interfaces with the same host. Each storage controller might also be coupled to more than one other storage controller, as illustrated in FIG. 2, in which storage controller 3-1 is connected to a number of other storage controllers, 3-2 through 3-M.

The storage controllers provide a number of functions and services to the host computer systems. In particular, the storage controllers automatically provide multiple back-up copies of stored data in a manner that is transparent to, and non-disruptive of, all of the host computer systems. For example, when a given data file is saved by a local host computer to its local storage array, the local storage controller causes one or more back-up copies of the data file to be stored in one or more of the remote storage arrays, without any intervention from or interruption to the local host computer. The process of backing-up data can be initiated by any storage controller in the system. This allows centralized back-up of any or all of the host computers. The number of back-up copies to be created and the locations to which the back-up copies are to be stored can be configured in advance by a system administrator. For example, these parameters can be programmed into a storage controller by connecting a keyboard and a display to ports on the storage controller. Alternatively, the storage controller can be configured by using one of the host computers to transmit configuration commands and data to the storage controller. Further, a remote storage controller can be configured via a local storage controller. The storage controller of the present invention also performs automatic block error detection and reassignment, as will be described below in detail.

The remote data access, data mirroring, and path redundancy provided by the present invention allow recovery from many possible failure modes, such as failure of communication medium, failure a host computer, or failure of a storage device.

In general, when the system of FIG. 1 is first set up, or when MSDs are added to the system, all data from a "primary" MSD are automatically copied to one or more newly-added MSDs. Such copying is performed while the host computers are operating, although the copying process is transparent to the host computers 2-1 through 2-M. Copying can be throttled to reduce the impact on system performance. Once all copies are established (brought into a consistent state), only changed data are subsequently written to the secondary MSDs in response to changes dictated by a host computer.

Each of the storage controllers also provides virtualized data access and emulation, as mentioned above. A local storage controller will emulate its local storage array from

the viewpoint of its local host computer system; similarly, the local storage controller will emulate its local host computer system from the viewpoint of its local storage array. Such emulation is implemented, in part, by using a common communication interface for data communication paths 7 and 8, such as SCSI. Again, in other embodiments, the data communication paths 7 and 8 may conform to other protocols and standards, such as serial SCSI, Fiber Channel, or ESCON.

A local host computer accesses data by transmitting a (virtual) host address to its local storage controller. The host address is then mapped to a real address representing a location on one or more physical MSDs, as will be described below in detail. The mapping is completely transparent to all of the host computers. To improve performance, the storage controller may distribute a single copy among multiple MSDs, which may be located in different storage arrays. Hence, in the above described mapping process, a single host address may map to multiple physical addresses, which may be distributed among multiple MSDs, and such MSDs may further be located in different storage arrays.

FIG. 3 illustrates the hardware architecture of a storage controller of the present invention according to one embodiment. The storage controller 3 includes a central processing unit (CPU) 10, random-access memory (RAM) 11, a non-volatile storage facility (NVSF) 12, an MSD 13, each of which is coupled to a bus 17. Bus 17 may represent multiple physical or logical buses, which may be interconnected by various adapters and/or controllers. NVSF 12 may be, or may include, a programmable non-volatile storage device, such as flash memory or electrically erasable programmable read-only memory (EEPROM). MSD 13 may be any conventional device that is suitable for non-volatile storage of large volumes of data, such as any of those discussed above. Also coupled to the bus 17 are a host device interface 14, a storage device interface 15, a controller device interface 16, and input/output (I/O) device interfaces 18 and 19. I/O device interfaces 18 and 19 are also coupled to separate, external connectors on the storage controller 3.

The host device interface 14 connects communication path 7 to the bus 17 in order to connect a local host computer to the storage controller 3. The storage device interface 15 connects communication path 8 to the bus 17 in order to connect a local storage array to storage controller 3. In one embodiment, host device interface 14 and storage device interface 15 each comprise a SCSI adapter, such that communication between storage controller 3 and the local host computer and storage array is performed using SCSI protocols. As is well-known, the SCSI standard allows multiple peripheral devices to be connected to a host device in a daisy chain configuration. From the viewpoint of the local storage array, the storage controller 3 emulates the host device. From the viewpoint of the host device, the storage controller 3 emulates the local storage array.

The controller device interface 16 connects communication path 9 to the bus 17 in order to connect a remote storage controller to storage controller 3. Controller device interface 16 may be an Ethernet, ATM, T1, T3, or FDDI adapter, or any other suitable device, depending upon the nature of the communication link 9.

I/O device interfaces 18 and 19 may be used to connect a keyboard and a monitor to the bus 17. I/O interface 18 and 19 may therefore be used by a systems administrator to perform various functions, such as initially configuring the storage controller 3, inputting commands and control information to the storage controller 3, or obtaining status information.

mation from the storage controller 3. Further, these interfaces 18 and 19 can be used to remotely perform these same functions on a remote storage controller via (local) storage controller 3 and communication link 9.

In one embodiment, the present invention is carried out in the storage controller by the CPU 10 executing sequences of instructions that are contained in a memory. More specifically, execution of the sequences of instructions contained in the memory causes the CPU 10 to perform steps according to the present invention which will be described below. For example, instructions may be loaded into RAM 11 for execution by the CPU 10 from a persistent store, such as the NVSF 12 or MSD 13. In other embodiments, hardwired circuitry may be used in place of, or in combination with, software instructions to implement the present invention. Thus, the present invention is not limited to any specific combination of hardware circuitry and software, nor to any particular source for the instructions executed by a computer system.

FIG. 4 shows the functional modules within a storage controller 3 of the present invention. Each of these functional modules may be embodied in hardware, software, or a combination of hardware and software. In one embodiment, the storage controller 3 includes an operating system (OS) 20 coupled to emulation drivers 21, physical drivers 22, communication drivers 23, and local memory 24. Memory 24 may include any combination of RAM 11, NVSF 12, or MSD 13 (see FIG. 3).

The OS 20 controls the overall operation of the storage controller, including emulation, data flow, caching of data in the local memory 24, control of data mirroring, and error recovery. Emulation drivers 21 are controlled by the OS 20 to provide communication of data between the storage controller 3 and the local host computer 2 via communication path 7. In particular, the emulation drivers 21 receive read and write requests from the host computer 2 and convert the read and write requests into a format recognized by the OS 20. The emulation drivers 21 further function to transfer any data received from the local storage array or a remote storage array to the local host computer. The communication drivers 23 are controlled by the OS 20 to provide communication of data between the storage controller 3 and remote storage controllers via communication path 9. The physical drivers 22 are controlled by the OS 20 to transfer data between the storage controller 3 and its local storage array. Such transfer includes the transformation of commands and control data from the format used by the storage controller 3 to the format recognized by the local storage array (e.g., SCSI).

In an alternative embodiment, a storage controller 3 may be an internal component of a host computer, rather than a separate, external unit. FIG. 5 illustrates a host computer 2-R according to one such embodiment. The host computer 2-R includes a CPU 30, a RAM 31, an NVSF 32, an MSD 33, a modem 34, a network adapter 35, and a storage controller 36, each of which is coupled to a bus 37. The local bus 37 may be any conventional local bus, such as a PCI (Peripheral Component Interconnect) bus. In this embodiment, the storage controller 36 may be constructed on an add-in card inserted in the host computer 2-R. The storage controller 36 provides all of the functionality of the storage controllers described above and has essentially the same architecture. However, in contrast with the storage controllers described above, the storage controller is directly connected to the local bus 37 of the host processing system. This embodiment therefore eliminates the need for the emulation drivers 21 illustrated in FIG. 4. Communication paths 8 and 9 to the

local storage array and the remote storage controllers, respectively, are provided through one or more external connectors on the host computer 2-R.

The modem 34 and network adapter 35 each allow the host computer 2-R to communicate other computer systems, such as clients or other hosts.

The various functions of the storage controller of the present invention will now be described in greater detail. The storage controller is configured to provide capability to gracefully recover from numerous types of failure modes, including block errors, device failures, and failure of communication paths, in a manner that is transparent to all hosts. For example, if an error occurs during an I/O operation, and the error is determined to result from a communication media failure (e.g., failure of one of the communication paths 7, 8 or 9), the storage controller automatically performs the operation using an alternate communication path, if available. If the error is determined to be a device failure (i.e., failure of an MSD), then the storage controller automatically attempts to perform the write operation to an alternate storage device. If the error is a block error, the system automatically attempts to refresh the block or, if necessary, to reassign the data to an alternate block.

In response to each I/O request (read or write) from a host, the storage controller performs a series of standard checks before carrying out the operation. FIG. 6 illustrates a routine for performing these standard checks. It is first determined, in step 601, whether the storage device identified by the host exists. As noted above, this device will be a virtual device. The method of identifying the physical MSD to be accessed is discussed below. If the device does exist, then the routine proceeds to step 602. Otherwise, the routine exits in step 610 with a "no device" error being provided to the host. If the device exists, then in step 602, it is determined whether the device to be accessed is off-line. If not, the routine proceeds to step 603. If the device is off-line, the routine exits in step 610 with a "no device" error being provided to the host. If the device is not off-line, then in step 603, it is determined whether the device has been reset. If so, the routine exits in step 612 with a reset notification being provided to the host. If the device has not been reset, then the routine proceeds to step 604, in which it is determined whether the device is ready. If the device is not ready, then in step 613 the routine exits with a "not ready" error being provided to the host. If the device is ready, then in step 605 it is determined whether the device is busy. If so, the routine exits in step 614 with the setting of a "busy" condition flag. If the device is not busy, then in step 606 it is determined whether the device or particular blocks to be accessed are reserved. If so, the routine exits in step 615 with a "reservation conflict" error being provided to the host. If the device or blocks to be accessed are not reserved, then in step 607 it is determined whether the starting block of the file to be accessed exists on the device. If not, the routine exits in step 616 with an "illegal request" error being provided to the host. If starting block exists on the device, then it is determined in step 608 whether the starting block -plus- block count exceeds the device capacity. If so, the routine exits in step 616 with an "illegal request" error being provided to the host. Otherwise, the routine exits in step 609 with a "checks complete without error" condition being provided to the host.

As mentioned above, a storage controller according to the present invention uses virtual-to-real device mapping to provide transparency of I/O operations to the host computers, as will now be described. A single host (virtual) address may map to a single physical address, or, to improve performance, the storage controller may map a single host

address to multiple physical addresses, which may be distributed among multiple MSDs located in different storage arrays. The storage controller 3 maps a host address to one or more physical addresses using a two-part process that is transparent to all hosts. Specifically, a host address received from a host is first mapped to a logical address, and the logical address is then mapped to a physical (real) address in one or more MSDs.

FIG. 7 illustrates the process of mapping a virtual address to a logical address. Because a given storage controller may have an interface with more than one host or multiple paths to the same host, each host address includes a host interface identifier (ID) that identifies the particular host or redundant path with which the access request is associated. Hence, in step 701, the storage controller receives a host interface ID and a host (memory) block number. In step 702, the storage controller determines whether the host device ID and block number map exactly to a logical device. This determination can be made by checking the status of a single bit representing whether or not the mapping is exact.

The storage controller maintains and uses a tree structure such as that illustrated in FIG. 8 to map the host interface ID and block number to a logical device. The tree structure 40 of FIG. 8 comprises a number of elements 41, 42 and 43. Each element includes a field designating a range of memory blocks as well as pointers to both a logical device and a logical block for that range. In addition, each element includes a pointer to another element corresponding to a different memory range. For example, the range of memory blocks represented by element 42 may be a lower range than that represented by element 41. Accordingly, element 41 would include a pointer to element 42 which is designated as a pointer to lower memory ranges. Similarly, element 43 may correspond to a range of memory blocks that is higher than the range of element 41. Accordingly, element 41 includes a pointer to element 43 which is identified as a pointer to higher memory ranges. Each element in the tree structure 40 similarly includes pointers to elements representing lower and higher memory ranges (except for elements in the bottom-most level of the tree).

Hence, referring again to FIG. 7, if it is determined in step 702 that the mapping between the host device and the logical device is exact, then in step 707 the output of the process is simply a logical device ID and a logical block number. If there is not an exact mapping, however, then in step 703 the process accesses the top element (e.g., element 41 in FIG. 8) in the virtual-to-logical mapping tree structure. In step 704, the host block number is compared to the element range in the top element in the tree. If the host block number falls below that range, then the process points to the element representing the next lower range (e.g., element 42 in FIG. 8) in step 706; step 704 is then repeated. If, in step 704, the host block number falls higher than the range of the top element in the tree, then in step 705 the process points to the element in the tree representing the next higher range (e.g., element 43 in FIG. 8); the process then repeats from step 704. If the block number is determined to fall within the range of the top (or current) element in the tree in step 704, then the process proceeds to step 707 by outputting the logical device and logical block number in the top (or current) element.

FIG. 9 illustrates the second part of the address mapping process, in which a logical address is mapped to a physical address. In step 901, the logical device ID and logical block number determined in the host-to-logical mapping process (FIG. 7) is input. In step 902, the storage controller determines whether the logical device maps exactly to the physi-

cal device. As in the above-described routine, this determination can be made based upon the status of a single bit. If the mapping is exact, then the process proceeds to step 907, in which the appropriate physical device ID and physical block number are output. Otherwise, the process proceeds to step 903, in which the routine points to the top element of a logical-to-physical mapping tree. The logical-to-physical mapping tree is similar in structure to, and is accessed in essentially the same manner as, the host-to-logical mapping tree of FIG. 8. In step 904, the logical block number is compared to the element range in the top element in the tree. If the logical block number falls below that range, then the process points to the element representing the next lower range in step 906; step 904 is then repeated. If the logical block number falls higher than the range of the top element in the tree in step 904, then in step 905 the process points to the element in the tree representing the next higher range; the process then repeats from step 904. If the block number is determined to fall within the range of the top (or current) element in the tree in step 904, then the process proceeds to step 907 by outputting the correct physical device and physical block number in the top (or current) element.

At various times, the storage controller must read data from one or more MSDs into its internal memory ("the cache") or transfer data from the cache to one or more MSDs; these procedures are known as "staging" and "destaging", respectively. Staging is generally performed, for example, when executing a host read request. Destaging is generally performed when executing a host write request.

FIG. 10 illustrates a routine for staging data, which can be performed by a storage controller of the present invention. In step 1001, the storage controller selects the "nearest" free device housing the data of interest. In this context, the term "nearest" refers to the mass storage device associated with the smallest overall access time or earliest availability, rather than referring to physical distance. (Note, however, that physical distance may be a factor in the total access time.) In a given situation, the nearest storage device may be the local memory within the local storage controller, an MSD within the local storage array, or an MSD within a remote storage array. After selecting the nearest free device, in step 1002 the logical block address of the data is mapped to a physical block address in the manner described above. In step 1003, the block or blocks are read from the selected device using the physical block address. If the read is successful (step 1004), then in step 1005 a read error recovery routine is called if any block was bad. The read error recovery routine is discussed below in connection with FIG. 11. If the read is not successful, and no additional copies of the data are available (step 1010), then a read error message is presented to the host and the routine exits in step 1011. If the read was unsuccessful and additional copies are available, then the routine repeats from step 1002 by mapping the block of the next available copy to a physical block. After a successful read (step 1004), and after calling the read error recovery routine if necessary (step 1005), then in step 1006 it is determined whether the staged data was requested by the host; if not, the routine exits. If so, the storage controller reconnects to the host (i.e., acquires exclusive use of communication link 8, which may be a SCSI bus) and transfers the block or blocks to the host in step 1007. If there is more data to be transferred (step 1008), then the storage controller disconnects from the host in step 1012, and the routine then repeats from step 1001. The storage controller disconnects from the host at this time because it is desirable not to monopolize the link 8 with the host while accessing the additional data. If there is no more data to be transferred,

11

then an ending status message is presented to the host in step 1009, and the routine then exits.

FIG. 11 illustrates the read error recovery routine. In response to a read error, in step 1101 the data of interest is read from the next-nearest physical copy. Again, “nearest” refers to overall access time or earliest availability, rather than physical distance. If the read is successful (step 1102), then an attempt is made in step 1103 to write the data to the bad block and to verify the write operation (i.e., to refresh the data block). If the write is successful (step 1104), then that fact is logged in step 1108, and the routine exits. If the attempt to rewrite the data is unsuccessful, then a write error recovery routine is called in step 1105, after which the routine exits. The write error recovery routine is discussed below in connection with FIG. 13. If the read operation (step 1101) is unsuccessful, and if no additional copies are available (step 1106), the routine exits in step 1107 with an error code presented to the host. If additional copies are available after an unsuccessful read, then the routine repeats from step 1101 using the next-nearest physical copy.

FIG. 12 illustrates a routine for destaging data. It is appropriate at this point to consider certain signals that are used in conjunction with destaging. Associated with each block of data are the two single-bit binary quantities, Exclusive Lock and Shared Lock. Only a single device is permitted access to a given block of data for purposes of performing a write operation; this constraint is implemented using Exclusive Lock. A write on a block is performed by obtaining an Exclusive Lock on that block. For purposes of writing to a block, Exclusive Lock for that block can be set for only if neither Shared Lock nor Exclusive Lock is already set. When a write operation on a block is initiated, Exclusive Lock is set for that block, so that no other device may access that block during the write operation. In contrast, multiple devices are permitted to read from a given block of data at the same time. A read on a block is performed by obtaining a Shared Lock on that block. A device may obtain a Shared Lock on a data block, regardless of the current status of Shared Lock, provided Exclusive Lock for that block is not already set. When a read is initiated on a given block, Shared Lock for that block is set, so that no device may write to that block during the read operation.

The storage controller can optionally be configured to delay and “merge” multiple destaging tasks to avoid multiple updates of the same physical blocks. In other words, instead of executing each destaging task, the storage controller may simply accumulate and log several destaging tasks in internal memory without executing them, and then subsequently perform a single “merged” destaging task to implement the aggregate effect of the accumulated destaging tasks.

Referring now to FIG. 12, when destaging is to be performed with respect to a given block, then Shared Lock for that block is set in step 1201. In step 1202, the first (or next) physical device is selected. In step 1203, the logical block address is mapped to a physical block address in the manner described above. In step 1204 the data is written to the physical device identified in step 1203. If the write is successful (step 1205), and if no read back verification was requested (step 1206), then in step 1208 the copy is marked as identical with that stored in the cache, and Shared Lock is cleared. It is desirable to mark a copy as identical with that stored in the cache in order to keep an accurate map of consistency between copies. If the write is not successful, then in step 1214 a call is made to the write error recovery routine. If error recovery is successful (step 1215), then the routine proceeds to step 1208 (described above). Otherwise,

12

copying is suspended and a “hot spare” is assigned (if available) in step 1216, and the routine exits. Note that strict adherence to the ordering of steps shown in FIG. 12 is not required in order to perform destaging according to the present invention; hence, multiple copies of data can be written simultaneously during destaging.

Following step 1208, it is next determined whether a host requires an ending status message (step 1209). If not, and if there are no additional copies stored in physical storage (step 1210), then the routine exits in step 1211. If no ending status is required by a host (step 1209), and there are additional copies (step 1210), then the routine repeats from step 1202 with selection of the next physical device. If there are no additional copies, the routine exits as in step 1211. If ending status is required by a host (step 1209), then in step 1212 the storage controller reconnects to the host (reacquires possession of the bus 8). Ending status is then presented to the host and the storage controller disconnects from the bus in step 1213. The routine then proceeds from step 1210.

FIG. 13 illustrates the write error recovery routine. In response to a write error, in step 1301 the faulty block is reassigned and written to a new block; this step is accompanied by verification of the write operation. If the write is successful (step 1302), then in step 1306 the new location of the block is logged and the routine exits. If the write was not successful, then the data is written to the internal MSD (i.e., MSD 13 in FIG. 3) in step 1303. Note that the internal MSD of the storage controller is generally not accessible to a host. After writing to the internal MSD, it is determined in step 1304 whether a hot spare MSD is available for the new copy. If no hot spare is available, then in step 1307 the location of the data is logged and the routine exits. If a hot spare is available, then in step 1305 an “establish” routine is performed to establish a new copy of the data on the hot spare. The establish routine is discussed below in connection with FIG. 18. The routine exits after establishing the new copy.

FIG. 14 illustrates a routine for performing a standard read operation in response to a request by a host. In step 1401, a storage controller receives a data read request from a host. Next, in step 1402 the standard checks routine (see FIG. 6) is performed. If any of the standards checks fail, then the routine exits in step 1410, presenting an error status message to the host. If none of the standard checks fail, then in step 1403 the storage controller identifies the appropriate logical device based on the host interface and virtual block number received from the host in step 1403. This is done in the manner discussed above in connection with FIG. 8. Next, if the requested block is cached (stored in the storage controller’s internal memory 24), then it is determined in step 1405 whether Exclusive Lock is set for the requested block. If Exclusive Lock is set or if the block was not cached, then in step 1411 the storage controller disconnects from the host, and in step 1412 a staging task is queued. The routine exists after queuing a staging task. If Exclusive Lock was not set in step 1405, then in step 1406 Shared Lock is set and the data is transferred to the host from the cache. After transferring the data to the host, Shared Lock is cleared in step 1407. Then, it is determined in step 1408 whether additional blocks have been requested. If so, the routine repeats from step 1403. If no additional blocks have been requested, then in step 1409 an ending status message is presented to the host, and the storage controller disconnects from the bus 8. The routine then exits.

As noted above, the storage controller may maintain multiple copies of a given set of data on different MSDs (which fact may be transparent to the hosts). Consequently, to improve performance, the storage controller can option-

ally be configured to concurrently read different subsets (i.e., blocks) of the set of data from different copies that are stored on different devices. This technique, which may be referred to as “spreading” reads, may reduce the overall time required to read the complete set of data in some cases.

FIG. 15 illustrates a routine for performing a standard write operation in response to a request from a host. In step 1501, the storage controller receives a write request from a host. Next, in step 1502 the standard checks (FIG. 6) are performed. If any of the standard checks fail, then the routine exits in step 1513, presenting an error status message to the host. Assuming none of the standard checks fail, then in steps 1503 and 1504, respectively, it is determined whether the device or blocks to be written to are designated as read-only. If the device or any of the blocks are read-only, then the routine exits, presenting an “illegal request” message to the host (steps 1514 and 1515). If neither the device nor any block is read-only, then in step 1505 the storage controller maps the host interface to the appropriate logical device using the technique described above (FIG. 7). After identifying the appropriate logical device, it is determined in step 1506 whether the block is already cached in the storage controller’s internal memory 24. If the block is already cached, then in step 1507 it is determined whether it possible to obtain Exclusive Lock on the requested block. As noted above, in order to obtain Exclusive Lock, neither Shared Lock nor Exclusive Lock for that block can be set. If Exclusive Lock can be obtained, then in step 1508 the data is transferred from the host to the appropriate block in the internal cache 24. If Exclusive Lock is not possible, then in step 1516 the storage controller disconnects from the host and waits for Exclusive Lock in step 1517. Once Exclusive Lock is obtained, then in step 1518 the storage controller reconnects to the host and, in step 1508, transfers the data from the host to the internal cache 24, as described above.

Referring again to step 1506, if the block to be written was not already cached, then in step 1519 it is determined whether the write operation is an update write (i.e., the writing of a partial block). With SCSI protocol, partial blocks cannot be written; only complete blocks can be written. However, other protocols which might be used in conjunction with the present invention might permit an update write. Accordingly, if the requested write operation is not an update write, then in step 1520 an empty block is allocated in the internal cache and the routine proceeds to step 1507 by obtaining Exclusive Lock, if possible. If the requested write operation is an update write, then in step 1521 the storage controller disconnects from the host. After disconnecting from the host, the storage controller stages the block from physical storage in step 1522 and then reconnects to the host in step 1518. The data is then transferred from the host to the appropriate block in the internal cache 24 in step 1508, as described above. Note that the process of disconnecting from the host while staging a block to be updated (steps 1521 and 1522) may be performed by an asynchronous task; that is,

After a block is transferred from the host to the internal cache 24 in step 1508, then if an ending status message is required (step 1509), the status message is presented to the host and the storage controller disconnects from the bus 8 in step 1510. After disconnecting from the bus in step 1510, or if no ending status message was required, then in step 1511 the cache copy is marked as different from the physical copies. Next, in step 1512 a destaging task is queued, and the routine exits.

As in the case of a read operation, if multiple copies are maintained, the storage controller can optionally be config-

ured to concurrently write different portions of a data set to different copies stored on different devices (i.e., “spread” the writes) in order to improve performance. Also, the storage controller can be configured to “simulate” a write to an MSD without actually writing to a physical device for certain cases. This procedure may be desirable for purposes of writing temporary data, such as data generated only for a specific job. In such cases, the storage controller simply writes the data to its internal memory 24 and then, if required, presents ending status to the host.

As mentioned above, a local storage controller of the present invention can communicate with one or more remote storage controllers for purposes of accessing a remote storage array. FIG. 16 illustrates a routine which can be performed by a local storage controller for transferring data to a remote storage controller. If there is data to be sent (step 1601) to a remote storage controller, then in step 1602 the local storage controller transmits the data onto the link 9 in packet form and waits for an acknowledgment from the remote storage controller. If a wait time-out occurs (step 1603) before an acknowledgment is received, then in step 1612 a line error recovery routine is called. The line error recovery routine is discussed below in connection with FIG. 17. If no wait time-out occurs before the acknowledgment is received, then if a request complete message is required by the host (step 1604) the storage controller waits in step 1605 for a “request complete” message from the remote storage controller. If no “request complete” message is received before a wait time-out occurs (step 1606), then the storage controller calls the line error recovery routine in step 1607. If no “request complete” message is required, or if no time-out occurs before the “request complete” message arrives, the routine repeats from step 1601.

FIG. 17 illustrates the line error recovery routine. In step 1701, a communications data packet is sent to the remote storage controller. If the transmission is successful (step 1702), then if alternate paths to that storage controller exist (step 1703), the routine repeats from step 1701 using an alternate path. Otherwise, the routine exits in step 1704 with the notification of an error condition. If the transmission was successful in step 1702, then the routine exits.

FIG. 18 illustrates the establish routine for establishing (ensuring consistency between) copies. Assume now that at least one copy of a given file has already been created from an “original” (which may be a previously-created copy). In step 1801, the storage controller identifies one or more contiguous blocks of data in the file. In step 1802, it is determined whether the corresponding blocks are identical between the original and the copy. If so, then in step 1803 one or more other contiguous blocks of data are examined, starting from the next block copy boundary. Next, in step 1804, if the extent of the storage device has not been reached, then the routine repeats from step 1801. If the extent of the device has been reached, then the copy has been established and the routine exits. If the blocks are determined not to be identical in step 1802, then in step 1805 the storage controller obtains (sets) Exclusive Lock for each of those blocks and reads the blocks from the source file. If the read is successful (step 1806), then Exclusive Lock is exchanged for Shared Lock and the blocks are written to the destination device in step 1808. If the read was not successful, then in step 1807 the read error recovery routine is called. After exchanging Exclusive Lock for Shared Lock and writing to the destination (step 1808), in step 1809 it is determined whether the write was successful. If not, the write error recovery routine is called in step 1811. If the write was successful, then in step 1810 the corresponding

blocks of the two different copies are marked as identical and the routine proceeds from step **1803** (described above). Note that during the establish routine double buffering may occur; that is, one or more subsequent reads may overlap in time with previous writes in order to improve performance. 5

A storage controller of the present invention also implements automatic block error detection and reassignment. When the storage controller is idle, it automatically examines data stored in any of the MSDs for errors and corrects detected errors, when possible. FIG. **19** illustrates a routine performed by the storage controller for implementing these functions. In step **1901**, the storage controller attempts to read a block of data from a given location. Note that this given location may be within the local storage array or it may be within one of the remote storage arrays. This attempted read (step **1901**) operation is not performed in response to any request from a host computer; it is performed on the storage controller's own initiative in response to the system becoming idle. In step **1902**, it is determined whether the block was successfully read. If the block was successfully read, then in step **1903** it is determined whether all blocks have been examined; if so, the routine ends. If all blocks have not been examined, then the storage controller attempts to read another block of data in step **1904**, and the process repeats from step **1902**. If the block was not read successfully, then in step **1905** the storage controller calls the read error recovery routine (FIG. **11**). 10

The storage controller of the present invention can also be used to migrate data from one storage array to another. The storage controller can be configured to perform data migration automatically in a manner that is transparent to, and non-disruptive of, any host computer. Further, in contrast with prior systems, which maintain only a single back-up copy of given file, the storage controller of the present invention allows two or more back-up copies of a file to be created and maintained in addition to a primary copy. Any copy is then accessible to any host computer connected to the system. 15

A storage controller of the present invention further allows data blocks to be write protected, so that a block cannot be modified from any host computer. Write protection may be desirable for purposes such as virus protection or implementation of security firewalls. Write protection can be achieved by configuring the storage controller appropriately at set-up time or by inputting a write protect command to the storage controller from a host computer. 20

Thus, a storage controller capable of providing multiple host computers system with access to multiple storage arrays has been described. Although the present invention has been described with reference to specific exemplary embodiments, it will be evident that various modifications and changes may be made to these embodiments without departing from the broader spirit and scope of the invention as set forth in the claims. Accordingly, the specification and drawings are to be regarded in an illustrative rather than a restrictive sense. 25

What is claimed is:

1. A storage controller, comprising:

a host interface providing a data interface with at least one host processing system independent of a device type of each said at least one host processing system, the storage controller physically separate from each said at least one host processing system; 30

a storage interface providing a data interface with at least one storage array, the storage controller physically separate from each said at least one storage array, each 65

said at least one storage array including at least one mass storage device, the storage interface further providing the data interface with said at least one storage array independent of the device type(s) of the mass storage devices in each said array; and

control circuitry coupled to the host interface and the data interface, the control circuitry configured to enable said at least one host processing system to access each said at least one storage array, such that the presence of the storage controller is transparent to each said at least one host processing system and each said at least one storage array.

2. A storage controller according to claim **1**, wherein each said at least one host processing system has a hardware configuration and a software configuration, wherein the storage controller is configured to enable said at least one host processing system to access said at least one storage array regardless of the hardware configuration or the software configuration(s) of said at least one host processing system.

3. A storage controller according to claim **1**, wherein the storage controller is configured to enable said at least one host processing system to access a mass storage device having any of a plurality of different device types.

4. A storage controller according to claim **1**, wherein the storage controller is configured to:

receive from one of said at least one host processing system a request specifying a virtual storage location of data; and

determine a physical storage location of the data based on the request. 30

5. A storage controller according to claim **4**, wherein the virtual storage location corresponds to a plurality of physical storage locations distributed among a plurality of mass storage devices.

6. A storage controller according to claim **1**, wherein: the storage controller is coupled to a local storage array; the storage controller is further coupled to a remote storage controller;

the remote storage controller is coupled to a remote host processing system; and

the storage controller is further configured to provide the remote host processing system with access to the local storage array. 35

7. A storage controller according to claim **6**, wherein the remote storage controller is coupled to a remote storage array, wherein the storage controller is further configured to provide the local host processing system with access to data stored in the remote storage array. 40

8. A storage controller according to claim **1**, wherein each of a plurality of copies of a set of data is stored on a different one of the mass storage devices, each copy of the set of data having a plurality of subsets, and wherein the storage controller is configured to access a different subset of the set of data in each of the plurality of copies to access the complete set of data. 45

9. A storage controller according to claim **1**, wherein a first one of the storage arrays may comprise mass storage devices having a different storage medium than a second one of the storage arrays.

10. A storage controller according to claim **1**, wherein: said at least one host processing systems comprises at least three host processing systems; and said at least one storage array comprises at least three storage arrays. 50

11. A storage controller according to claim **1**, wherein the control circuitry is configured to communicate with a plurality of other storage controllers. 55

12. A data storage controller for providing a plurality of host processing systems with access to a plurality of storage arrays, each of the storage arrays including a plurality of mass storage devices, the data storage controller comprising:

- a processor; and
- a memory coupled to the processor, the memory having instructions stored therein which, when executed by the processor, cause the data storage controller to:
 - provide a data interface with a local host processing system independent of a device type of the local host processing system, the storage controller physically separate from the local host processing system;
 - emulate a local storage array to the local host processing system, such that the presence of the storage controller is functionally transparent to the local host processing system;
 - provide a data interface with the local storage array independent of the type of mass storage devices in the local storage array, the storage controller physically separate from the local storage array;
 - emulate the host processing system to the mass storage devices of the local storage array, such that the presence of the storage controller is transparent to the local storage array;
 - provide an interface to a remote data storage controller, the remote data storage controller for coupling to a remote storage array; and
 - control communication of data such that any of the plurality of host processing systems can access data in any of the plurality of storage arrays.

13. A data storage controller according to claim 12, wherein the instructions further comprise instructions which, when executed by the processor, cause the data storage controller to receive from one of the host processing systems a request specifying a virtual storage location of a data block and to determine a real storage location of the data block based on the virtual storage location.

14. A data storage controller according to claim 12, wherein the instructions further comprise instructions which, when executed by the processor, allow the data storage controller to provide any of the host processing systems with access to any of the storage arrays using virtual addresses, such that actual locations of stored data in the storage arrays are transparent to the host processing systems.

15. A data storage controller according to claim 14, wherein one of the virtual addresses may be mapped to a plurality of real storage locations distributed among a plurality of the mass storage devices.

16. A data storage controller according to claim 12, wherein the instructions further comprise instructions which, when executed by the processor, cause the data storage controller to provide a remote host processing system of the plurality of host processing systems with access to the local storage array, the remote host processing system coupled to the remote data storage controller.

17. A data storage controller according to claim 16, wherein the remote storage controller is coupled to a remote storage array, and wherein the instructions further comprise instructions which, when executed by the processor, cause the data storage controller to provide the local host processing system with access to data stored in the remote storage array.

18. A data storage controller according to claim 12, wherein the data storage controller is coupled to the host processing system via a SCSI interface.

19. A data storage controller according to claim 18, wherein the data storage controller is coupled to the local storage array via a SCSI interface.

20. A data storage controller according to claim 12, further comprising means for connecting a mass storage device having any of a plurality of different device types to the host processing system.

21. A data storage controller according to claim 20, wherein a first one of the storage arrays may comprise mass storage devices having a different storage medium than a second one of the storage arrays.

22. A data storage controller according to claim 12, wherein the local host processing system has a hardware configuration and a software configuration, the data storage controller further comprising means for providing the data interface to the host processing system, such that the local host processing system can access the local storage array independent of both the hardware configuration and the software configuration of the local host processing system.

23. A data storage controller according to claim 12, wherein:

the plurality of processing systems includes at least three host processing systems; and

the plurality of storage arrays includes at least three storage arrays.

24. A data storage controller according to claim 12, wherein the instructions which cause the data storage controller to control communication of data comprise instructions which, when executed by the processor, cause the data storage controller to communicate with a plurality of other data storage controllers.

25. In a storage controller coupled to a plurality of host processing systems and a plurality of storage arrays, a method of providing the plurality of host processing systems with access to the plurality of storage arrays, the plurality of host processing systems including a local host processing system, the plurality of storage arrays including a local storage array, each of the storage arrays including a plurality of mass storage devices, the method comprising:

providing a data interface with the local host processing system, the host processing system physically separate from the storage controller;

emulating the local storage array, such that the presence of the storage controller is transparent to the local host processing system;

providing a data interface with the local storage array, the local storage array physically separate from the storage controller, wherein the local storage array may include any of a plurality of different storage mediums;

emulating the local host processing system, such that the presence of the storage controller is transparent to the local storage array;

providing an interface to a remote data storage controller coupled to a remote storage array of the plurality of storage arrays; and

controlling communication of data such that any of the plurality of host processing systems can access data in any of the plurality of storage arrays.

26. A method according to claim 25, wherein the local storage array includes mass storage devices of a first storage medium, and the remote storage array includes mass storage devices of a second storage medium different from the first storage medium.

27. A method according to claim 25, further comprising the step of providing the host processing system with access to the storage arrays using virtual addressing, wherein physical locations of data stored in the storage arrays are transparent to the host processing systems.

28. A method according to claim 25, wherein the plurality of processing systems comprises at least three host processing systems.

19

29. A method according to claim 28, wherein the plurality of storage arrays comprises at least three storage arrays.

30. A method according to claim 25, wherein the step of controlling communication of data comprises the step of communicating with a plurality of other storage controllers to cause a plurality of copies of a set of data to be stored distributed among the plurality of storage arrays.

31. A system comprising:

- a plurality of host processing systems;
- a plurality of storage arrays, each of the storage arrays including at least one mass storage device; and
- a plurality of storage controllers, each of the storage controllers coupled to one of the host processing systems and physically separate from the host processing systems, each of the storage controllers coupled to one of the storage arrays and physically separate from the storage arrays, each of the storage controllers further coupled to another one of the storage controllers, each of the storage controllers having a first interface with one of the host processing systems, the first interface for emulating a mass storage device such that the presence of the storage controller is transparent to the host processing systems, each of the storage controllers further having a second interface with a local storage array, the second interface for emulating a host processing system such that the presence of the storage controller is transparent to storage arrays, the plurality of storage controllers cooperating to permit any of the host processing systems to access data stored in any of the storage arrays.

32. A system according to claim 31, wherein for each of the storage controllers, the second interface is capable of interfacing with mass storage devices having any of a plurality of different device types.

33. A system according to claim 32, wherein two or more of the storage arrays may use dissimilar storage media.

34. A system according to claim 31, wherein the plurality of processing systems comprises at least three host processing systems.

20

35. A systems comprising:

- a plurality of host processing systems;
- a plurality of storage arrays, each of the storage arrays including at least one mass storage device; and
- a plurality of storage controllers, each of the storage controllers coupled to one of the host processing systems and physically separate from said one of the host processing systems, coupled to one of the storage arrays and physically separate from said one of the storage arrays, and coupled to another one of the storage controllers, wherein the storage controllers cooperate to permit any of the host processing systems to access data stored in any of the storage arrays, each of the storage controllers including means for supporting any one of a plurality of types of mass storage media in a corresponding one of the storage arrays, wherein the presence of each of the storage controllers is transparent to each of the host processing systems and each of the storage arrays.

36. A system according to claim 35, wherein each of the storage controllers further comprises means for supporting any one of a plurality of configurations of a corresponding host processing system.

37. A system according to claim 36, each of the storage controllers comprising an interface with a host processing system for emulating a mass storage device and an interface with a local storage array for emulating a host processing system.

38. A system according to claim 37, each of the storage devices further comprising translation means for representing the storage arrays as at least one virtual device to the host processing systems.

39. A system according to claim 35, wherein the storage controllers function as peers with respect to each other.

40. A system according to claim 35, wherein the plurality of processing systems comprises at least three host processing systems.

* * * * *