

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization International Bureau



(43) International Publication Date  
14 April 2005 (14.04.2005)

PCT

(10) International Publication Number  
WO 2005/033945 A1

(51) International Patent Classification<sup>7</sup>: G06F 12/00

(21) International Application Number:  
PCT/US2004/032122

(22) International Filing Date:  
30 September 2004 (30.09.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/507,329 30 September 2003 (30.09.2003) US  
10/911,987 5 August 2004 (05.08.2004) US

(71) Applicant (for all designated States except US): SEPARTON, INC. [US/US]; 144 Turnpike Road, Southborough, MA 01778 (US).

(72) Inventor; and

(75) Inventor/Applicant (for US only): SANDORFI, Miklos [US/US]; 14 McKenzie Lane, Foxboro, MA 02035 (US).

(74) Agent: SKRIVANEK, Robert, A., Jr.; Lowrie, Lando & Anastasi, LLP, One Main Street, Cambridge, MA 02142 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

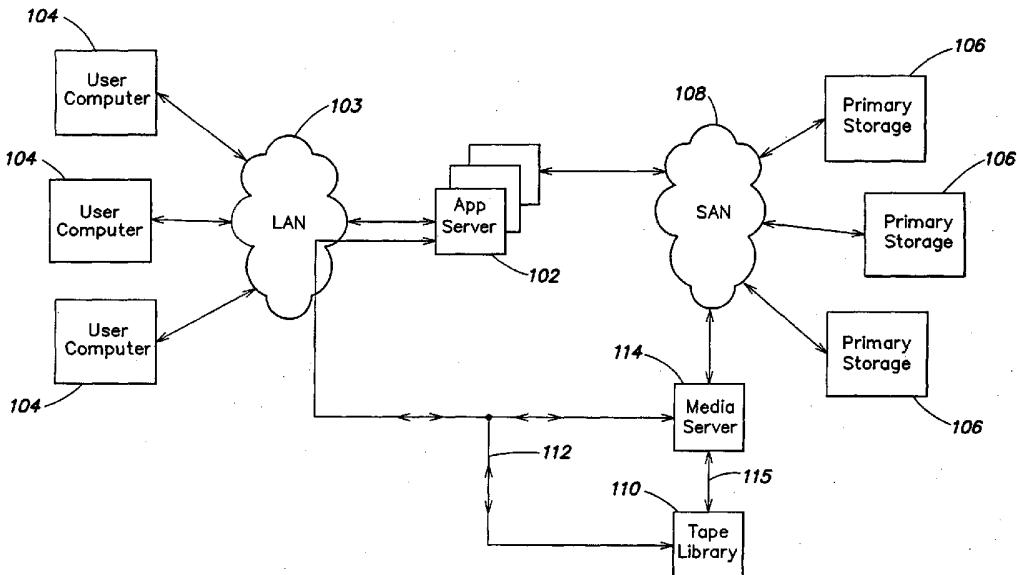
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

[Continued on next page]

(54) Title: EMULATED STORAGE SYSTEM SUPPORTING INSTANT VOLUME RESTORE



WO 2005/033945 A1

(57) Abstract: In a back-up storage system, an apparatus and methods for mounting a data volume corresponding to a back-up data set to a host computer. In one example, a method includes mounting a data volume on a host computer, the data volume comprising at least one data file, the data file corresponding to a most recently backed-up version of the at least one data file stored on a backup storage system, and storing, on the backup storage system, data corresponding to a second version of the at least one data file that is more recent than the most recently backed-up version of the at least one data file stored on the backup storage system while preserving the most recently backed-up version of the at least one data file.



- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

**EMULATED STORAGE SYSTEM SUPPORTING INSTANT VOLUME RESTORE****BACKGROUND**1. **Field of Invention**

5 Aspects of the present invention relate to data storage, and more particularly to apparatus and methods for emulating a tape storage system to provide the equivalent of full back-ups using an existing full back-up and subsequent incremental back-ups and enabling end-users to restore data from such back-ups.

10 2. **Discussion of Related Art**

Many computer systems include one or more host computers and one or more data storage systems that store data used by the host computers. These host computers and storage systems are typically networked together using a network such as a Fibre Channel network, an Ethernet network, or another type of communication network. Fibre Channel is a standard that 15 combines the speed of channel-based transmission schemes and the flexibility of network-based transmission schemes and allows multiple initiators to communicate with multiple targets over a network, where the initiator and the target may be any device coupled to the network. Fibre Channel is typically implemented using a fast transmission media such as optical fiber cables, and is thus a popular choice for storage system networks where large 20 amounts of data are transferred.

An example of a typical networked computing environment including several host computers and back-up storage systems is shown in FIG. 1. One or more application servers 102 are coupled via a local area network (LAN) 103 to a plurality of user computers 104. Both the application servers 102 and the user computers 104 may be considered "host computers." 25 The application servers 102 are coupled to one or more primary storage devices 106 via a storage area network (SAN) 108. The primary storage devices 106 may be, for example, disk arrays such as are available from companies like EMC Corporation, IBM Corporation and others. Alternatively, a bus (not shown) or other network link may provide an interconnect between the application servers and the primary storage system 106. The bus and/or Fibre 30 Channel network connection may operate using a protocol, such as the Small Component System Interconnect (SCSI) protocol, which dictates a format of packets transferred between the host computers (e.g., the application servers 102) and the storage system(s) 106.

- 2 -

It is to be appreciated that the networked computing environment illustrated in FIG. 1 is typical of a large system as may be used by, for example, a large financial institution or large corporation. It is to be understood that many networked computing environments need not include all the elements illustrated in FIG. 1. For example, a smaller networked computing 5 environment may simply include host computers connected directly, or via a LAN, to a storage system. In addition, although FIG. 1 illustrates separate user computers 104, application servers 102 and media servers 114, these functions may be combined into one or more computers.

In addition to primary storage devices 106, many networked computer environments 10 include at least one secondary or back-up storage system 110. The back-up storage system 110 may typically be a tape library, although other large capacity, reliable secondary storage systems may be used. Typically, these secondary storage systems are slower than the primary storage devices, but include some type of removable media (e.g., tapes, magnetic or optical disks) that may be removed and stored off-site.

15 In the illustrated example, the application servers 102 may be able to communicate directly with the back-up storage system 110 via, for example, an Ethernet or other communication link 112. However, such a connection may be relatively slow and may also use up resources, such as processor time or network bandwidth. Therefore, a system such as illustrated may include one or more media servers 114 that may provide a communication link, 20 using for example, Fibre Channel, between the SAN 108 and the back-up storage system 110.

The media servers 114 may run software that includes a back-up/restore application that controls the transfer of data between host computers (such as user computers 104, the media servers 114, and/or the application servers 102), the primary storage devices 106 and the back-up storage system 110. Examples of back-up/restore applications are available from 25 companies such as Veritas, Legato and others. For data protection, data from the various host computers and/or the primary storage devices in a networked computing environment may be periodically backed-up onto the back-up storage system 110 using a back-up/restore application, as is known in the art.

Of course, it is to be appreciated that, as discussed above, many networked computer 30 environments may be smaller and may include fewer components than does the exemplary networked computer environment illustrated in FIG. 1. Therefore, it is also to be appreciated that the media servers 114 may in fact be combined with the application servers 102 in a single host computer, and that the back-up/restore application may be executed on any host computer

that is coupled (either directly or indirectly, such as through a network) to the back-up storage system 110.

One example of a typical back-up storage system is a tape library that includes a number of tape cartridges and at least one tape drive, and a robotic mechanism that controls 5 loading and unloading of the cartridges into the tape drives. The back-up/restore application provides instructions to the robotic mechanism to locate a particular tape cartridge, e.g., tape number 0001, and load the tape cartridge into the tape drive so that data may be written onto the tape. The back-up/restore application also controls the format in which data is written onto the tapes. Typically, the back-up/restore application may use SCSI commands, or other 10 standardized commands, to instruct the robotic mechanism and to control the tape drive(s) to write data onto the tapes and to recover previously written data from the tapes.

Conventional tape library back-up systems suffer from a number of problems including speed, reliability and fixed capacity. Many large companies need to back-up Terabytes of data each week. However, even expensive, high-end tapes can usually only read/write data at 15 speeds of 30-40 Megabytes per second (MB/s), which translates to about 50 Gigabyte per hour (GB/hr). Thus, to back-up one or two Terabytes of data to a tape back-up system may take at least 10 to 20 hours of continuous data transfer time.

In addition, most tape manufacturers will not guarantee that it will be possible to store (or restore) data to/from a tape if the tape is dropped (as may happen relatively frequently in a 20 typical tape library because either a human operator or the robotic mechanism may drop a tape during a move or load operation) or if the tape is exposed to non-ideal environmental conditions, such as extremes in temperature or moisture. Therefore, a great deal of care needs to be taken to store tapes in a controlled environment. Furthermore, the complex machinery of a tape library (including the robotic mechanism) is expensive to maintain and individual tape 25 cartridges are relatively expensive and have limited lifespans.

## SUMMARY OF INVENTION

Embodiments of the present invention provide a back-up storage system that overcomes or alleviates some or all of the problems of conventional tape library systems and 30 that may provide greater flexibility than do conventional tape library systems.

In broad overview, aspects and embodiments of the present invention provide a random-access based storage system that emulates a conventional tape back-up storage system such that a back-up/restore application sees the same view of devices and media as with a

physical tape library. The storage system of the invention uses software and hardware to emulate physical tape media and replace them with one or more random-access disk arrays, translating tape format, linear, sequential data to data that is suitable for storage on disk. In addition, applications implemented in hardware and/or software are provided for recovering 5 the data stored on the back-up storage system.

According to some aspects and embodiments of the present invention, there is provided a mechanism for transforming sequential tape-formatted data into a format suitable for random access I/O. In one embodiment, there is provided a mechanism that includes provisions for mounting the transformed representation of tape-formatted data on a host computer as an NFS 10 (network file system) or CIFS (common Internet file system) mounted volume.

According to other aspects and embodiments of the invention, there is provided a mechanism for diverting writes to the mounted file system to “safe storage” whereby the original data remains unchanged. In one embodiment, a mechanism is provided for tracking real-time changes to the original data so that random access I/O is possible. In another 15 embodiment, there is provided a mechanism for transforming the newly written data back into tape-formatted data suitable for sequential tape-specific I/O.

In one embodiment, a method comprises acts of mounting a data volume on a host computer, the data volume comprising at least one data file, the data file corresponding to a most recently backed-up version of the at least one data file stored on a backup storage system, 20 and storing, on the backup storage system, data corresponding to a second version of the at least one data file that is more recent than the most recently backed-up version of the at least one data file stored on the backup storage system while preserving the most recently backed-up version of the at least one data file. The method may also comprise linking the most recently backed-up version of the at least one data file with the second version of the at least one data 25 file. In one example, the method may also comprise creating a data structure that identifies both the most recently backed-up version of the at least one data file and the second version of the at least one data file. In another example, the second version of the at least one data file may be a modified version of the most recently backed-up version of the at least one data file.

According to another embodiment, a back-up storage system comprises a back-up 30 storage media for storing the back-up data set and a controller including at least one processor configured to execute a set of instructions implementing the method described above.

According to another embodiment, there is provided a computer readable medium having a data structure stored thereon, the data structure comprising a first identifier that

- 5 -

uniquely identifies a system file that corresponds to a back-up data set that includes at least one data file, and at least one second identifier that identifies a respective storage location on a storage medium where a most recent version of each at least one data file of the back-up data set is stored.

5

#### BRIEF DESCRIPTION OF DRAWINGS

The accompanying drawings, are not intended to be drawn to scale. In the drawings, each identical or nearly identical component that is illustrated in various figures is represented by a like numeral. For purposes of clarity, not every component may be labeled in every drawing. In the drawings:

FIG. 1 is a block diagram of one example of a large-scale networked computing environment that includes a back-up storage system;

FIG. 2 is a block diagram of one embodiment of a networked computing environment including a storage system according to aspects of the invention;

FIG. 3 is a block diagram of one embodiment of a storage system according to aspects of the invention;

FIG. 4 is a block diagram illustrating a virtual layout of one embodiment of a storage system according to aspects of the invention;

FIG. 5 is a schematic layout of one example of a system file according to aspects of the invention;

FIG. 6 is one example of a tape directory structure according to aspects of the invention;

FIG. 7 is a diagram depicting one example of a method of creating a synthetic full back-up according to aspects of the invention;

FIG. 8 is a schematic diagram of one example, of a series of back-up data sets including a synthetic full back-up according to aspects of the invention;

FIG. 9 is a diagram of one example of a metadata cache structure;

FIG. 10 is a diagram of one example of a virtual cartridge storing a synthetic full back-up data set;

FIG. 11 is a diagram of another example of a virtual cartridge storing a synthetic full back-up data set;

FIG. 12 is a flow diagram of one embodiment of a method for restoring data from the back-up storage system, according to aspects of the invention;

- 6 -

FIG. 13 is a block diagram of another embodiment of a networked computing environment including a back-up storage system according to aspects of the invention;

FIG. 14 is a diagram of one example of a file descriptor structure according to aspects of the invention;

5 FIG. 15 is a diagram illustrating one example of how file data may be stored in tape format;

FIG. 16 is a diagram illustrating a file descriptor for the file depicted in Fig. 15;

FIG. 17 is a flow diagram of a method of writing data to a mounted data volume in accordance with one embodiment of the present invention;

10 FIG. 18 is an illustration of a newly written file;

FIG. 19 is an illustration of one example of a relationship between an original file, a newly written file, and a resulting modified file, according to aspects of the invention; and

FIG. 20 is a diagram of one example of a file descriptor representing the modified file of FIG. 19.

15

#### DETAILED DESCRIPTION

Various embodiments and aspects thereof will now be described in more detail with reference to the accompanying figures. It is to be appreciated that this invention is not limited in its application to the details of construction and the arrangement of components set forth in 20 the following description or illustrated in the drawings. The invention is capable of other embodiments and of being practiced or of being carried out in various ways. Also, the phraseology and terminology used herein is for the purpose of description and should not be regarded as limiting. The use of "including," "comprising," "having," "containing," "involving," and variations thereof herein, is meant to encompass the items listed thereafter 25 and equivalents thereof as well as additional items.

As used herein, the term "host computer" refers to any computer that has at least one processor, such as a personal computer, a workstation, a mainframe, a networked client, a server, etc. that is capable of communication with other devices, such as a storage system or other host computers. Host computers may include media servers and application servers (as 30 described previously with reference to FIG. 1) as well as user computers (which may be user workstations, PCs, mainframes, etc.). In addition, within this disclosure, the term "networked computer environment" includes any computing environment in which a plurality of host computers are connected to one or more shared storage systems in such a manner that the

storage system(s) can communicate with each of the host computers. Fibre Channel is one example of a communication network that may be used with embodiments of the present invention. However, it is to be appreciated that the networks described herein are not limited to Fibre Channel, and that the various network components may communicate with each other over any network connection, such as Token Ring or Ethernet instead of, or in addition to Fibre Channel, or over combinations of different network connections. Moreover, aspects of the present invention may also be used in bus topologies, such as SCSI or parallel SCSI.

According to various embodiments and aspects of the present invention, there is provided a virtual removable media library back-up storage system that may use one or more disk arrays to emulate a removable media based storage system. Using embodiments of the invention, data may be backed-up onto the disk array(s) using the same back-up/restore application as would have been used to back-up the data onto removable media (such as tapes, magnetic disks, optical disks, etc.), without a user having to make any modifications or adjustments to the existing back-up procedures or having to purchase a new back-up/restore application. In one embodiment, described in detail herein, the removable media that are emulated are tapes, and the back-up storage system of the invention emulates a tape library system including tapes and the robotic mechanism used to handle tapes in a conventional tape library system.

A storage system according to aspects of the invention includes hardware and software that together interface with a host computer (running the back-up/restore application) and a back-up storage media. The storage system may be designed to emulate tapes, or other types of removable storage media, such that the back-up/restore application sees the same view of devices and media as with a physical tape library, and to translate linear, sequential, tape format data into data that is suitable for storage on random-access disks. In this manner, the storage system of the invention may provide enhanced functionality (such as, allowing users to search for individual backed-up user files, as discussed below) without requiring new back-up/restore application software or policies.

Referring to FIG. 2, there is illustrated in block diagram form, one embodiment of a networked computing environment including a back-up storage system 170 according to aspects of the invention. As illustrated, a host computer 120 is coupled to the storage system 170 via a network connection 121. This network connection 121 may be, for example a Fibre Channel connection to allow high-speed transfer of data between the host computer 120 and the storage system 170. It is to be appreciated that the host computer 120 may be, or may

include, one or more application servers 102 (see FIG. 1) and/or media servers 114 (see FIG. 1) and may enable back-up of data from either any of the computers present in the networked computing environment or from a primary storage system 110 (see FIG. 1). In addition, one or more user computers 136 may also be coupled to the storage system 170 via another network 5 connection 138, such as an Ethernet connection. As discussed in detail below, the storage system may enable users of the user computer 136 to view and optionally restore backed-up user files from the storage system.

The storage system includes back-up storage media 126 that may be, for example, one or more disk arrays, as discussed in more detail below. The back-up storage media 126 10 provide the actual storage space for backed-up data from the host computer(s) 120. However, the storage system 170 may also include software and additional hardware that emulates a removable media storage system, such as a tape library, such that, to the back-up/restore application running on the host computer 120, it appears as though data is being backed-up onto conventional removable storage media. Thus, as illustrated in FIG. 2, the storage system 15 170 may include “emulated media” 134 which represent, for example, virtual or emulated removable storage media such as tapes. These “emulated media” 134 are presented to the host computer by the storage system software and/or hardware and appear to the host computer 120 as physical storage media. Further interfacing between the emulated media 134 and the actual back-up storage media 126 may be a storage system controller (not shown) and a switching 20 network 132 that accepts the data from the host computer 120 and stores the data on the back-up storage media 126, as discussed more fully in detail below. In this manner, the storage system “emulates” a conventional tape storage system to the host computer 120.

According to one embodiment, the storage system may include a “logical metadata cache” 242 that stores metadata relating to user data that is backed-up from the host computer 25 120 onto the storage system 170. As used herein, the term “metadata” refers to data that represents information about user data and describes attributes of actual user data. The logical metadata cache 242 represents a searchable collection of data that enables users and/or software applications to randomly locate backed-up user files, compare user files with one another, and otherwise access and manipulate backed-up user files. Two examples of software 30 applications that may use the data stored in the logical metadata cache 242 include a synthetic full back-up application 240 and an end-user restore application 300 that are discussed more fully below.

In brief overview, the synthetic full back-up application 240 is capable of creating a synthetic full back-up data set from one existing full back-up data set and one or more incremental back-up data sets. The synthetic full backup may obviate the need to perform periodic (e.g., weekly) full back-ups, thereby saving considerable time and network resources.

5 Details of the synthetic full back-up application 240 are described further below. The end-user restore application 300, also described more fully in detail below, enables end-users (e.g., operators of the user computers 136) to browse, locate, view and/or restore previously backed-up user files from the storage system 170.

As discussed above, the storage system 170 includes hardware and software that 10 interface with the host computer 120 and the back-up storage media 126. Together, the hardware and software of embodiments of the invention may emulate a conventional tape library back-up system such that, from the point of view of the host computer 120, data appears to be backed-up onto tape, but is in fact backed-up onto another storage medium, such as, for example, a plurality of disk arrays.

15 Referring to FIG. 3, there is illustrated in block diagram form, one embodiment of a storage system 170 according to aspects of the invention. In one example, the hardware of the storage system 170 includes a storage system controller 122 and a switching network 132 that connects the storage system controller 122 to the back-up storage media 126. The storage system controller 122 includes a processor 127 (which may be a single processor or multiple 20 processors) and a memory 129 (such as RAM, ROM, PROM, EEPROM, Flash memory, etc. or combinations thereof) that may run all or some of the storage system software. The memory 129 may also be used to store metadata relating to the data stored on the back-up storage media 126. Software, including programming code that implements embodiments of the present 25 invention, is generally stored on a computer readable and/or writeable nonvolatile recording medium, such as RAM, ROM, optical or magnetic disk or tape, etc., and then copied into memory 129 wherein it may then be executed by the processor 127. Such programming code may be written in any of a plurality of programming languages, for example, Java, Visual Basic, C, C#, or C++, Fortran, Pascal, Eiffel, Basic, COBAL, or combinations thereof, as the present invention is not limited to a particular programming language. Typically, in operation, 30 the processor 127 causes data, such as code that implements embodiments of the present invention, to be read from a nonvolatile recording medium into another form of memory, such as RAM, that allows for faster access to the information by the processor than does the nonvolatile recording medium.

- 10 -

As shown in FIG. 3, the controller 122 also includes a number of port adapters 124a, 124b, 124c that connect the controller 122 to the host computer 120 and to the switching network 132. As illustrated, the host computer 120 is coupled to the storage system via a port adapter 124a, which may be, for example, a Fibre Channel port adapter. Via a storage system controller 122, the host computer 120 backs up data onto the back-up storage media 126 and can recover data from the back-up storage media 126.

In the illustrated example, the switching network 132 may include one or more Fibre Channel switches 128a, 128b. The storage system controller 122 includes a plurality of Fibre Channel port adapters 124b and 124c to couple the storage system controller to the Fibre Channel switches 128a, 128b. Via the Fibre Channel switches 128a, 128b, the storage system controller 122 allows data to be backed-up onto the back-up storage media 126. As illustrated in FIG. 3, the switching network 132 may further include one or more Ethernet switches 130a, 130b that are coupled to the storage system controller 122 via Ethernet port adapters 125a, 125b. In one example, the storage system controller 122 further includes another Ethernet port adapter 125c that may be coupled to, for example, a LAN 103 to enable the storage system 170 to communicate with host computers (e.g., user computers), as discussed below.

In the example illustrated in FIG. 3, the storage system controller 122 is coupled to the back-up storage media 126 via a switching network that includes two Fibre Channel switches and two Ethernet switches. Provision of at least two of each type of switch within the storage system 170 eliminates any single points of failure in the system. In other words, even if one switch (for example, Fibre Channel switch 128a) were to fail, the storage system controller 122 would still be able to communicate with the back-up storage media 126 via another switch. Such an arrangement may be advantageous in terms of reliability and speed. For example, as discussed above, reliability is improved through provision of redundant components and elimination of single points of failure. In addition, in some embodiments, the storage system controller is able to back-up data onto the back-up storage media 126 using some or all of the Fibre Channel switches in parallel, thereby increasing the overall back-up speed. However, it is to be appreciated that there is no requirement that the system comprise two or more of each type of switch, or that the switching network comprise both Fibre Channel and Ethernet switches. Furthermore, in examples wherein the back-up storage media 126 comprises a single disk array, no switches at all may be necessary.

As discussed above, in one embodiment, the back-up storage media 126 may include one or more disk arrays. In one preferred embodiment, the back-up storage media 126 include

a plurality of ATA or SATA disks. Such disks are “off the shelf” products and may be relatively inexpensive compared to conventional storage array products from manufacturers such as EMC, IBM, etc. Moreover, when one factors in the cost of removable media (e.g., tapes) and the fact that such media have a limited lifetime, such disks are comparable in cost to 5 conventional tape-based back-up storage systems. In addition, such disks can read/write data substantially faster than can tapes. For example, over a single Fibre Channel connection, data can be backed-up onto a disk at a speed of at least about 150 MB/s, which translates to about 540 GB/hr, significantly faster (e.g., by an order of magnitude) than tape back-up speeds. In addition, several Fibre Channel connections may be implemented in parallel, thereby 10 increasing the speed even further. In accordance with an embodiment of the present invention, back-up storage media may be organized to implement any one of a number of RAID (Redundant Array of Independent Disks) schemes. For example, in one embodiment, the back-up storage media may be configured as a RAID-5 implementation.

As discussed above, embodiments of the invention emulate a conventional tape library 15 back-up system using disk arrays to replace tape cartridges as the physical back-up storage media, thereby providing a “virtual tape library.” Physical tape cartridges that would be present in a conventional tape library are replaced by what is termed herein as “virtual cartridges.” It is to be appreciated that for the purposes of this disclosure, the term “virtual tape library” refers to an emulated tape library which may be implemented in software and/or 20 physical hardware as, for example, one or more disk array(s). It is further to be appreciated that although this discussion refers primarily to emulated tapes, the storage system may also emulate other storage media, for example, a CD-ROM or DVD-ROM, and that the term “virtual cartridge” refers generally to emulated storage media, for example, an emulated tape or emulated CD. In one embodiment, the virtual cartridge in fact corresponds to one or more 25 hard disks.

Therefore, in one embodiment, a software interface is provided to emulate the tape library such that, to the back-up/restore application, it appears that the data is being backed-up onto tape. However, the actual tape library is replaced by one or more disk arrays such that the data is in fact being backed-up onto these disk array(s). It is to be appreciated that other types 30 of removable media storage systems may be emulated and the invention is not limited to the emulation of tape library storage systems. The following discussion will now explain various aspects, features and operation of the software included in the storage system 170.

It is to be appreciated that although the software may be described as being "included" in the storage system 170, and may be executed by the processor 127 of the storage system controller 122 (see FIG. 3), there is no requirement that all the software be executed on the storage system controller 122. The software programs such as the synthetic full back-up 5 application and the end-user restore application may be executed on the host computers and/or user computers and portions thereof may be distributed across all or some of the storage system controller, the host computer(s), and the user computer(s). Thus, it is to be appreciated that there is no requirement that the storage system controller be a contained physical entity such as a computer. The storage system 170 may communicate with software that is resident 10 on a host computer such as, for example, the media server(s) 114 or application servers 102. In addition, the storage system may contain several software applications that may be run or resident on the same or different host computers. Moreover, it is to be appreciated that the storage system 170 is not limited to a discrete piece of equipment, although in some embodiments, the storage system 170 may be embodied as a discrete piece of equipment. In 15 one example, the storage system 170 may be provided as a self-contained unit that acts as a "plug and play" replacement for conventional tape library back-up systems (i.e., no modification need be made to existing back-up procedures and policies). Such a storage system unit may also be used in a networked computing environment that includes a conventional back-up system to provide redundancy or additional storage capacity.

As discussed above, according to one embodiment, the host computer 120 (which may be, for example, an application server 102 or media server 114, see FIG. 1) may back-up data onto the back-up storage media 126 via the network link (e.g., a Fibre Channel link) 121 that couples the host computer 120 to the storage system 170. It is to be appreciated that although the following discussion will refer primarily to the back-up of data onto the emulated media, 25 the principles apply also to restoring backed-up data from the emulated media. The flow of data between the host computer 120 and the emulated media 134 may be controlled by the back-up/restore application, as discussed above. From the view point of the back-up/restore application, it may appear that the data is actually being backed-up onto a physical version of the emulated media.

Referring to FIG. 4, the storage system software 150 may include one or more logical abstraction layer(s) that represent the emulated media and provide an interface between a back-up/restore application 140 resident on the host computer 120 and the back-up storage media 126. The software 150 accepts tape format data from the back-up/restore application 140 and

translates that data into data suitable for storage on random-access disks (e.g., hard disks, optical disks and the like). In one example, this software 150 is executed on the processor 127 of the storage system controller 122 and may be stored in memory 129 (see FIG. 3).

According to one embodiment, the software 150 may include a layer, referred to herein 5 as the virtual tape library (VTL) layer 142 that may provide a SCSI emulation of tapes, tape drives, and also the robotic mechanisms used to transfer tapes to and from the tape drives. The back-up/restore application 140 may communicate (e.g., back-up or write data to the emulated media) with the VTL 142 using, for example, SCSI commands, represented by arrows 144. Thus, the VTL may form a software interface between the other storage system software and 10 hardware and the back-up/restore application, presenting the emulated storage media 134 (Fig. 2) to a back-up/restore application and allowing the emulated media to appear to the back-up/restore application as conventional removable back-up storage media.

A second software layer referred to herein as the file system layer 146 may provide an 15 interface between the emulated storage media (represented in the VTL) and the physical back-up storage media 126. In one example, the file system, 146 acts as a mini operating system to communicate with the back-up storage media 126 using, for example, SCSI commands, represented by arrows 148, to read and write data to and from the back-up storage media 126.

In one embodiment, the VTL provides generic tape library support and may support 20 any SCSI media changer. Emulated tape devices may include, but are not limited to, an IBM LTO-1 and LTO-2 tape device, a Quantum SuperDLT320 tape device, a Quantum P3000 tape library system, or a StorageTek L180 tape library system. Within the VTL, each virtual cartridge is a file that may grow dynamically as data is stored. This is in contrast to conventional tape cartridges which have a fixed size. One or more virtual cartridges may be stored in a system file as described further below with respect to FIG. 5.

25 FIG. 5 illustrates one example of a data structure within the file system software 146 that illustrates a system file 200 in accordance with an embodiment of the present invention. In this embodiment, the system file 200 includes a header 202 and data 204. The header 202 may include information that identifies each of the virtual cartridges that are stored in that system file. The header may also contain information such as, whether a virtual cartridge is 30 write protected, the dates of creation/modification of the virtual cartridges, etc. In one example, the header 202 includes information uniquely identifying each virtual cartridge and distinguishing each virtual cartridge from other virtual cartridges stored in the storage system. For example, this information may include a name and an identifying number (e.g.,

corresponding to a barcode that would typically be present on a physical tape so that the tape could be identified by the robotic mechanism) of the virtual cartridge. The header 202 may also contain additional information such as a capacity of each of the virtual cartridges, a date of last modification, etc.

5 According to one embodiment of the present invention, the size of the header 202 may be optimized to reflect the type of data being stored (e.g., virtual cartridges representing data back-up from one or more host computer systems) and the number of distinct sets of such data (e.g., virtual cartridges) that the system can track. For example, data that is typically backed-up to a tape storage system is typically characterized by larger data sets representing numerous 10 system and user files. Because the data sets are so large, the number of discrete data files to be tracked may be correspondingly small. Accordingly, in one embodiment, the size of the header 202 may be selected based on a compromise between storing too much data to efficiently keep track of (i.e., the header being too big) and not having space to store a sufficient number of cartridge identifiers (i.e., header being too small). In one exemplary embodiment, the 15 header 202 utilizes the first 32 MB of the system file 200. However it is to be appreciated that the header 202 may have a different size based on system needs and characteristics and that, depending on system needs and capacity, one may select a different size for the header 202.

It is to be appreciated that, from the point of view of the back-up/restore application, the virtual cartridges appear as physical tape cartridges with all the same attributes and 20 features. That is, to the back-up restore application, the virtual cartridges appear as sequentially written tapes. However, in one preferred embodiment, the data stored in the virtual cartridges is not stored in a sequential format on back-up storage media 126. Rather, the data that appears to be written to the virtual cartridges is in fact stored in the storage system's files as randomly accessible, disk-format data. Metadata is used to link the stored 25 data to virtual cartridges so that the back-up/restore application can read and write data in cartridge format.

Thus, in broad overview of one preferred embodiment, user and/or system data (referred to herein as "file data") is received by the storage system 170 from the host computer 120 and is stored on the disk array(s) making up the back-up storage media 126. The software 30 150 (see FIG. 4) and/or hardware of the storage system writes this file data to the back-up storage media 126 in the form of system files, as is described in more detail below. Metadata is extracted from the backed-up file data by the storage system controller to keep track of attributes of the user and/or system files that are backed-up. For example, such metadata for

each file may include the file name, a date of creation or last modification of the file, any encryption information relating to the file, and other information. In addition, metadata may be created by the storage system for each file that links the file to a virtual cartridge. Using such metadata, the software provides to the host computer an emulation of tape cartridges;

5 however the file data is in fact not stored in tape format, but rather in the system files, as discussed below. Storing the data in system files, rather than in sequential cartridge format, may be advantageous in that it allows fast, efficient and random access to individual files without the need to scan through sequential data to find a particular file.

As discussed above, according to one embodiment, file data (i.e., user and/or system data) is stored on the back-up storage media as system files, each system file including a header and data, the data being the actual user and/or system files. The header 202 of each system file 200 includes a tape directory 206 that contains metadata linking the user and/or system files to virtual cartridges. The term "metadata" as used herein refers not to user or system file data, but to data that describes attributes of actual user and/or system data.

10 According to one example, the tape directory may define, down to the byte level, the layout of data on the virtual cartridges. In one embodiment, the tape directory 206 has a table structure, as illustrated in FIG. 6. The table includes a column 220 for the type of information stored (e.g., data, a file marker (FM), etc.), a column 222 for the size of the disk blocks used in bytes, and a column 224 that reflects the number of disk blocks in which the file data is stored. Thus,

15 the tape directory allows the controller to have random (as opposed to sequential) access to any data file stored on back-up storage media 126. For example, referring to FIG. 6, the data file 226 may be quickly located on the virtual tape because the tape directory indicates that the data of file 226 begins one block from the beginning of the system file 200. This one block has no size because it corresponds to a file marker (FM). File markers are not stored in the system

20 file, i.e., file markers correspond to zero data. The tape directory includes file markers because they are used by conventional tapes and the back-up/restore application thus writes file markers along with data files and expects to see file markers when viewing a virtual cartridge. Therefore, file markers are kept track of in the tape directory. However, file markers do not represent any data and are therefore not stored in the data section of the system file. Thus, the

25 data of file 226 begins at the beginning of the data section of the system file, indicated by arrow 205, and is 1024 bytes in length (i.e., one disk block that is 1024 bytes in size). It should be appreciated that other file data may be stored in a block size other than 1024 bytes,

depending on the amount of data, i.e., the size of the data file. For example, larger data files may be stored using larger disk block sizes for efficiency.

In one example, the tape directory may be contained in a “file descriptor” that is associated with each data file backed-up onto the storage system. The file descriptor contains 5 metadata relating the data files 204 stored on the storage system. In one embodiment, the file descriptor may be implemented in accordance with a standardized format, such as the tape archive (tar) format used by most Unix-based computer systems. Each file descriptor may include information such as the name of the corresponding user file, the date the user file was created/modified, the size of the user file, any access restrictions on the user file, etc.

10 Additional information stored in the file descriptor may further include information describing the directory structure from which the data was copied. Thus, the file descriptor may contain searchable metadata about a corresponding data file, as is discussed in more detail below.

From the point of view of the back-up/restore application, any virtual cartridge may contain a plurality of data files and corresponding file descriptors. From the point of view of 15 the storage system software, the data files are stored in system files that may be linked to, for example, a particular back-up job. For example, a back-up executed by one host computer at a particular time may generate one system file that may correspond to one or more virtual cartridges. Virtual cartridges may thus be of any size and may grow dynamically as more user files are stored on the virtual cartridges.

20 Referring again to FIG. 3, as discussed above, the storage system 170 may include a synthetic full back-up software application 240. In one embodiment, the host computer 120 backs-up data onto the emulated media 134, forming one or more virtual cartridges. In some computing environments, a “full back-up,” i.e., a back-up copy of all data stored on the primary storage system in the network (see FIG. 1), may be accomplished periodically (e.g., 25 weekly). This process is typically very lengthy due the large amount of data that is to be copied. Therefore, in many computing environments, additional back-ups, termed incremental back-ups, may be performed between consecutive full back-ups, e.g., daily. An incremental back-up is a process whereby only data that has changed since the last back-up was executed (whether incremental or full) is backed-up. Typically, this changed data is backed-up on a file 30 basis, even though frequently much of the data in the file has not changed. Thus, incremental back-ups are typically much smaller, and therefore much faster to accomplish, than are full back-ups. It is to be appreciated that although many environments typically execute full back-ups once a week and incremental back-ups daily during the week, there is no requirement that

such time frames are used. For example, certain environments may require incremental back-ups several times a day. The principles of the invention apply to any environment using full back-ups (and optionally incremental back-ups), regardless of how often they are executed.

During a full back-up procedure, the host computer may create one or more virtual 5 cartridges containing the backed-up data that comprises a plurality of data files. For clarity, the following discussion will assume that the full back-up generates only one virtual cartridge. However, it is to be appreciated that a full back-up may generate more than one virtual cartridge, and that the principles of the invention apply to any number of virtual cartridges.

According to one embodiment, there is provided a method for creating a synthetic full 10 back-up data set from one existing full back-up data set and one or more incremental back-up data sets. This method may obviate the need to perform periodic (e.g., weekly) full back-ups, thereby saving the user considerable time and network resources. Furthermore, as known to those of skill in the art, restoring data based on a full back-up and one or more incremental back-ups can be a time consuming process because, for example, if the most recent version of a 15 file exists in an incremental back-up, the back-up/restore application will typically restore the file based on the last full back-up and then apply any changes from the incremental back-ups. Providing a synthetic full back-up, therefore, may have an additional advantage of allowing the back-up restore application to more quickly restore data files based on the synthetic full back-up alone, without the need to perform multiple restores from a full back-up and one or more 20 incremental back-ups. It is to be appreciated that the phrase "most recent version" as used herein refers generally to the most recent copy of a data file (i.e., the most recent time that the data file was saved), whether or not the file has a new version number. The term "version" is used generally herein to refer to copies of the same file which may be modified in some way or may have been saved multiple times.

25 Referring to FIG. 7, there is illustrated a schematic representation of a synthetic full back-up procedure. The host computer 120 may execute a full back-up 230 at a first moment in time, for example, on a weekend. The host computer 120 may then execute subsequent incremental back-ups 232a, 232b, 232c, 232d, 232e, for example, on each day during the week. The storage system 170 may then create a synthetic full back-up data set 234, as discussed 30 below.

According to one embodiment, the storage system 170 may include a software application referred to herein as a synthetic full back-up application 240 (see FIG. 3). The synthetic full back-up application 240 may be run on the storage system controller 122 (see

FIG. 2) or may be run on the host computer 120. The synthetic full back-up application includes software commands and interfaces necessary for creating the synthetic full back-up data set 234. In one example, the synthetic full back-up application may perform a logical merge of metadata representations of each of the full back-up data set 230 and the incremental 5 back-up data sets 232 to generate a new virtual cartridge that contains the synthetic full back-up data set 234.

For example, referring to FIG. 8, the existing full back-up data set may include user files F1, F2, F3 and F4. A first incremental back-up data set 232a may include user files F2', a modified version of F2, and F3', a modified version of F3. A second incremental back-up data 10 set 232b may include user files F1', a modified version of F1, and F2'', a further modified version of F2, and a new user file F5. Therefore, the synthetic full back-up data set 234 formed from a logical merge of the full back-up data set 230 and the two incremental data sets 232a and 232b, contains the latest version of each of user files F1, F2, F3, F4 and F5. As seen 15 in FIG. 8, the synthetic full back-up data set therefore contains user files F1', F2'', F3', F4 and F5.

Referring again to FIGS. 3 and 4, the file system software 146 may create a logical metadata cache 242 that stores metadata relating to each user file stored on the emulated media 134. It is to be appreciated that the logical metadata cache is not required to be a physical data 20 cache, but may instead be a searchable collection of data stored on the storage media 126. In another example, the logical metadata cache 242 can be implemented as a database. Where the metadata is stored in a database, conventional database commands (e.g., SQL commands) can be used to perform the logical merge of the full back-up data set and the one or more incremental back-up data sets to create the synthetic full back-up data set.

As discussed above, each data file stored on the emulated media 134 may include a file 25 descriptor that contains metadata relating to the data file, including a location of the file on the back-up storage media 126. In one embodiment, the back-up/restore application running on the host computer 120 stores data in a streaming tape format on the emulated media 134. An example of a data structure 250 representing this tape format is illustrated in FIG. 9. As discussed above, the system file data structure includes headers which may contain information 30 about the data file(s), such as the file descriptor for the data files, the dates of creation and/or modification of the files, security information, the directory structure of the host system from whence the file(s) came, as well as other information linking the files to a virtual cartridge. These headers are associated with the data 254 which is actual user and system files that have

been backed-up (copied) from the host computer, the primary storage system, etc. The system file data structure may also optionally include pads 256 which may appropriately align the next header to a block boundary.

As shown in FIG. 9, in one embodiment, the header data is located in the logical 5 metadata cache 242 to permit rapid searching and random access to the otherwise sequential tape data format. The use of the logical metadata cache, implemented using the file system software 148 on the storage system controller 122, allows translation of the linear, sequential tape data format, stored on the emulated media 134, into the random-access data format stored on physical disks making up the back-up storage media 126. The logical metadata cache 242 10 stores the headers 252 which include the file descriptors for the data files, security information which may be used to control access to the data files, as is discussed in more detail below, and pointers 256 to the actual locations of the data files on the virtual cartridges and the back-up storage media 126. In one embodiment, the logical metadata cache stores data relating to all 15 the data files backed-up in the full back-up data set 230 and each of the incremental data sets 232.

According to one embodiment, the synthetic full back-up application software 240 uses the information stored in the logical metadata cache to create a synthetic full back-up data set. This synthetic full back-up data set is then linked to a synthetic virtual cartridge, created by the synthetic full back-up application 240. To the back-up/restore application, the synthetic full 20 back-up data set appears to be stored on this synthetic virtual cartridge. As discussed above, the synthetic full back-up data set may be created by performing a logical merge of the existing full back-up data set and the incremental back-up data sets. This logical merge may include comparing each of the data files included in each of the existing full back-up data set and the incremental back-up data sets and creating a composite of the latest-modified version of each 25 user file, as discussed above in reference to FIG. 8.

According to one embodiment, the synthetic virtual cartridge 260 includes pointers that point to locations of data files on other virtual cartridges, specifically, the virtual cartridges that contain the existing full back-up data set and the incremental back-up data sets, as shown in FIG. 10. Considering the example given with respect to Fig. 8 above, the synthetic virtual 30 cartridge 260 includes pointers 266 that point (indicated by arrows 268) to the locations in the existing full back-up data set, on virtual cartridge 262, of user file F4 (because the existing full back-up data set contained the latest version of F4) and to the location of, for example, user file F3' in incremental data set 232a on virtual cartridge 264.

The synthetic virtual cartridge also includes a list 270 that contains the identifying numbers (and optionally the names) of all the virtual cartridges that contain data to which the pointers 266 point. This dependent cartridge list 270 may be important for keeping track of where the actual data is and for preventing the dependent virtual cartridges from being erased.

5 In this embodiment, the synthetic full back-up data set does not contain any actual user files, but rather a set of pointers that indicate the locations of the user files on the back-up storage media 126. Therefore, it may be desirable to prevent the actual user files (stored on other virtual cartridges) from being deleted. This may be accomplished in part by keeping a record (dependent cartridge list 270) of the virtual cartridges that contain the data and protecting each

10 of those virtual cartridges from being over-written or deleted. The synthetic virtual cartridge may also include cartridge data 272 such as, the size of the synthetic virtual cartridge, its location on the back-up storage media 126, etc. In addition, the synthetic virtual cartridge may have an identifying number and/or name 274.

According to another embodiment, the synthetic virtual cartridge may include a

15 combination of pointers and actual stored user files. Referring to FIG. 11, in one example, the synthetic virtual cartridge includes pointers 266 that point to locations of data files (the latest versions, as discussed above in reference to FIG. 9) in the existing full back-up data set 230 on virtual cartridge 262. The synthetic virtual cartridge may also include data 278 containing actual data files copied from the incremental data sets 232, as indicated by arrows 280. In this

20 manner, the incremental back-up data sets can be deleted after the synthetic full back-up data set 276 has been created, thereby saving storage space. The synthetic virtual cartridges are relatively small as they contain all or partly pointers rather than copies of all the user files.

It is to be appreciated that synthetic full back-ups may include any combination of pointers and stored file data and are not limited to the examples given above. For example,

25 synthetic full back-ups may include pointers to data files for some files stored on certain incremental and/or full back-ups and may include stored file data copied from other existing full and/or incremental back-ups. Alternatively still, a synthetic full back-up may be created based upon a prior full back-up and any relevant incremental back-ups that does not include any pointers, but rather includes the latest version of actual file data copied from the

30 appropriate full and/or incremental back-ups.

In one embodiment, the synthetic full back-up application software may include a differencing algorithm that enables it to compare the user and system file metadata for each of the existing full back-up data set and the incremental back-up data sets to determine where the

latest version of each of the data files is located. For example, a differencing algorithm could be used to compare the dates of creation and/or modification, the version number (if applicable), etc. between different versions of the same data files in the different back-up sets to select the most recent version of the data file. However, users may often open a user file 5 and save the file (thereby changing its date of modification) without actually changing any of the data inside the file. Therefore, the system may implement a more advanced differencing algorithm that may analyze the data inside the system or user files to determine whether the data has in fact changed. Variations of such differencing algorithms and other types of compare algorithms may be known to those skilled in the art. In addition, as discussed above, 10 where the metadata is stored in a database format, database commands such as SQL commands can also be used to perform the logical merge. The invention may apply any of such algorithms to ensure that the most recent or latest version of each user file may be selected from all compared existing back-up sets so as to properly create the synthetic full back-up data set.

15 As should be appreciated by those skilled in the art, the synthetic full back-up application enables full back-up data sets to be created and made available without requiring the host computer to execute a physical full back-up. Not only does this avoid burdening the host computer with the processor overhead of transferring the data to the back-up storage system, but in embodiments where the synthetic full back-up application is executed on the 20 storage system, it significantly reduces the utilization of network bandwidth. As illustrated in FIG. 7, further synthetic full back-up data sets may be created using a first synthetic full back-up data set 234 and subsequent incremental back-up data sets 236. This may provide a significant time advantage in that files or objects that are not frequently modified may not be frequently copied. Instead, the synthetic full back-up data sets may maintain pointers to these 25 files that have just been copied once.

As discussed above in reference to FIG. 3, the storage system may also include a software application referred to as the end-user restore application 300. Thus, according to another embodiment, there is provided a method for end users to locate and restore back-up data without IT staff intervention and without requiring any changes to existing back-up/restore procedures and/or policies. In a typical back-up storage system, the back-up/restore application running on the host computer 120 is controlled by IT staff and it may be impossible or very difficult for an end-user to access backed-up data without intervention by the IT staff. According to aspects and embodiments of the invention, storage system software is provided

that allows end users to locate and restore their files via, for example, a web-based or other interface with the back-up storage media 126.

It is to be appreciated that, as with the synthetic full back-up application 240, the end-user restore application 300 may be run on the storage system controller 122 (see FIG. 2) or 5 may be run on the host computer 120. The end-user restore application includes software commands and interfaces necessary to allow an authorized user to search the logical metadata cache to locate, an optionally restore, backed-up files from the back-up storage media 126.

According to one embodiment, there is provided software including a user interface 10 that is installed and/or executed on the user computer 136. The user interface may be any type of interface that allows a user to locate files on the back-up storage media. For example, the user interface may be a graphical user interface, may be web-based, or may be a text interface. The user computer is coupled to the storage system 170 via a network connection 138 which may be, for example, an Ethernet connection. Through this network connection 138, an operator of the user computer 136 can access the data stored on the storage system 170.

15 In one example, the end-user restore application 300 includes a user authentication and/or authorization feature. For example, a user may be asked to login via the user interface on the user computer using a username and password. The user computer may communicate the username and password to the storage system (e.g., to the end-user restore application) which may use an appropriate user authentication mechanism to determine whether the user 20 has access to the storage system. Some examples of user authentication mechanisms that may be used include, but are not limited to, a Microsoft Active Directory server, a Unix “yellow pages” server or a Lightweight Directory Access Protocol. The login/user authentication mechanism may communicate with the end-user restore application to exchange the user privileges. For example, some users may be allowed to search only those files that have been 25 created by them or for which they have certain privileges or are identified as an owner. Other users such as, for example, system operators or administrators may be allowed access to all backed-up files, etc.

According to one embodiment, the end-user restore application uses the logical 30 metadata cache to obtain information about all the data files backed-up on the back-up storage media. The end-user restore application presents to the user, via the user interface, a hierarchical directory structure of the user's files sorted by, for example, back-up time/date, username, original user computer directory structure (that may have been obtained when the files were backed-up), or other file characteristics. In one example, the directory structure

presented to the user may vary according to the privileges enabled for that user. The end-user restore application may accept browse requests (i.e., through the user interface, the user may browse the directory structure to locate a desired file) or the user may search for a file by name, date, etc.

5 According to one embodiment, the user may restore backed-up files from the storage system. For example, once the user has located a desired file, as described above, the user may download the file from the storage system via the network connection 138. In one example, this download procedure may be implemented in a manner comparable to any web-based download, as known to those skilled in the art.

10 By allowing end users to access those files for which they have permission to view/download, and by enabling such access through a user interface (e.g., web-based technology), the end-user restore application can enable user to search for and restore their own files without there being any need to alter any back-up policies or procedures.

15 According to another embodiment, there is provided a mechanism and methods whereby users can “mount” a network attached view of the back-up data sets that are stored on the back-up storage media 126. This may allow users to view and access data in the mounted data set(s) as they would data on any other local or network drive coupled to their computer. Thus, for example, a user may restore data availability to an application server (e.g., if the system’s primary storage 106 (see FIG. 1) has suffered a fault) without having to execute a 20 restore process through the media server 114 (see FIG. 1). The restoration of data to the application server using a mounting procedure as described herein may be many orders of magnitude faster than a typical media server facilitated volume restore. It is to be appreciated that the term “mount” as used herein refers to making a data volume or a network component, such as a network drive, available to the operating system of a host computer. The data 25 volume may include, for example, a single data file or system file, a plurality of files, or a directory structure that may include a plurality of files. Common mounting protocols include NFS (network file system) or CIFS (common Internet file system) sharing. These protocols allow a host computer to access resources on another computer over a network connection via an interface that makes it appear as those the remote resources were locally present on the host computer.

30 Referring to FIG. 12, there is illustrated a flow chart representing one embodiment of a method for performing a volume mount according to aspects of the invention. In a first step 290, a user selects a data volume to mount and sends a volume mount request to the back-up

- 24 -

storage system controller 122 (see FIG. 3). Typically, a user may want to restore data from a full back-up data set (rather than an incremental back-up data set) so as to capture a full and accurate representation of the backed-up information. If a current full back-up data set does not exist, (for example, a network manager may execute weekly full back-ups and thus if the 5 user wishes to restore data during the week, a current full back-up may not be available), a synthetic full back-up may be created (as described above) and used to restore selected data.

According to one embodiment, the back-up storage system 170 may include a software application, referred to herein a volume restore application 310 (see FIG. 13), that may control and implement methods for performing a data volume mount and restore procedure. The 10 volume restore application 310, similar to the synthetic full back-up and end-user restore applications, may be executed on the host computers and/or user computers and portions thereof may be distributed across all or some of the storage system controller, the host computer(s), and the user computer(s).

Referring again to FIG. 12, after a volume mount is requested, the volume restore 15 application may query whether a current full back-up data set is available (step 292). If not, the volume restore application may communicate with the synthetic full back-up application 240 (see FIG. 2) to perform the synthetic full back-up process and create a current back-up data set (step 294). The volume restore application may export either a regular full back-up data set or synthetic full back-up data set to perform the requested volume mount as either an 20 NFS or CIFS share. Specifically, the volume restore application queries the logical metadata cache 242 to locate the appropriate metadata that represents the selected full back-up volume identified in step 290.

According to one embodiment, the mount request (step 290) may cause the volume 25 restore application to build one or more filé descriptor structures to facilitate exporting the volume for mounting as an NFS or CIFS share (step 296). Referring to FIG. 14, there is illustrated one embodiment of a file descriptor structure 320 that may be built by the volume restore application, the file descriptor 320 corresponding to a system file in tape format (e.g., system file 332 – see FIG. 15). As discussed above, a file descriptor contains searchable 30 metadata corresponding to data files and system files stored on the storage system. The file descriptor 320 may include a plurality of fields that contain information such as, for example, the filename 322 and file permissions (access control lists) 324 for the data files contained in the volume to be mounted. In addition, the file descriptor contains one or more pointers 326 to the location of source data of a data file (i.e., to identify where on the storage media 126 the

- 25 -

data file is stored), the length 328 of that data file, and a pointer 330 to the next entry (e.g., the next data file) in the linked-list file descriptor structure. If the “next” field is null, e.g., as indicated by reference numeral 331, then it represents that the data file is the last data file known to the system file represented by the file descriptor 320 (e.g., is the last linked-list entry).

5 Each system file that is contained in the data volume to be mounted will be represented by a file descriptor structure such as that illustrated in FIG. 14. Once each system file in the requested volume has a file descriptor 320 built, the file descriptors may be used to locate and export the associated data files in response to an NFS or CIFS request.

As discussed above, in one embodiment, the file descriptor may be implemented in accordance with a standardized format, such as the tape archive (tar) format used by most Unix-based computer systems. An illustration of a typical system file 332 as it would be written in tape format (e.g., in tar format) is illustrated in FIG. 15 as a segment of a tape (e.g., tar) data stream. A corresponding file descriptor 340 for system file 332 is illustrated in FIG. 16. As illustrated in FIG. 15, a file written in tape format includes a header 336 and actual data 338 stored in the system file 332. The data 338 may correspond to one or more data files. In the illustrated example, the system file 332 is 1032 bytes long, however it is to be appreciated that the file may have any length depending on the size of the file and the format in which it is written.

The file descriptor 340 for file 332 is contained in the header 336. As illustrated in FIG. 16 and in accordance with the general example given in FIG. 14, the file descriptor 340 contains a filename 341, security information 344, pointers 342 to the stored data of each data file known to the system file, the length 346 of the corresponding data file, and a “next” entry that identifies the next data file known to the system file, which in the illustrated example is a null 348.

25 Referring again to FIG. 12, once all the file descriptors for the files in the data volume to be mounted have been built, the volume restore application exports the file system based on the built file descriptors to a user-specified mount point as either an NFS or CIFS share (step 298). At this point, the mount is completed (step 299) and the mounted data volume is available for the user to read and/or write data, as discussed below.

30 According to one embodiment, NFS or CIFS read operations (i.e., the user wishes to view data in the mounted data volume) are serviced by searching through the file descriptors 320 for matching file specifications. It is to be appreciated that, according to one embodiment, the user need not actually search the file descriptors himself. Rather, the volume restore

application may include a user interface that presents the data to the user in, for example, a typical directory structure format. The volume restore application may include software that translates user requests for specific files into search commands that access the logical metadata cache and search the file descriptors 320 for the matching system files. Once a file has been 5 located, data transfer to the user computer may be accomplished by following the linked list (i.e., following the pointers stored in the file descriptor to locate the actual data) to build a buffer for the file data which may be sent to the requesting user.

According to another embodiment, a mechanism may be provided for the user to also write new data to the mounted volume. As discussed above, the mounted volume data may 10 appear to the user as an ordinary network drive or other network-stored data. However, in fact the original mounted volume data is actually back-up data that typically needs to be preserved; at least until another back-up data set is created. Therefore, it may be undesirable to allow a user to actually modify the original back-up data. To avoid modifying the back-up data while still allowing the user modify data corresponding to the mounted volume, a mechanism is 15 provided that (transparent to the user) diverts the write to other storage media, as discussed below.

Referring to FIG. 17, there is illustrated a flow chart of one embodiment of a method of processing a write request according to aspects of the invention. In a first step 350, a user 20 requests an NFS or CIFS write operation (typically by selecting a “save” option while editing or viewing a data file). The volume restore application then implements the write request by locating available storage space, writing the data to that space, and updating an appropriate file descriptor to reference the newly written data.

According to one embodiment, the volume restore application queries whether storage 25 space has already been allocated for writing data (step 352) and, if not, allocates storage space (step 354). The storage space may be allocated on the back-up storage media 126 (see FIG. 13). The allocated storage space may be designated specifically to hold only the write data (and optionally, associated metadata).

Referring to FIG. 18, there is illustrated one example of NFS or CIFS write data as 30 stored on the back-up storage media 126. The written data 360 includes, for example, two written portions, W1 362 and W2 364 that correspond to stored data arising as a result of write commands serviced by the volume restore application. For example, W1 and W2 may correspond to modified data files that are contained within the mounted data volume. It is to be appreciated that although the illustration corresponds to two write requests, the principles of

the invention apply to any number of write requests, and the files may be suitably altered to reflect any appropriate number of write requests. The written data 360 also includes a header 366 that includes metadata forming a self-describing relationship between the original data (e.g., file 332) and the newly written data 360. Particularly, the header may include offset 5 information that indicates where the written data portions W1 and W2 logically exist relative to the original data, as described further in relation to FIG. 19.

Referring to FIG. 19, there is illustrated one example of a system file layout after two write requests have been serviced. An original system file 332 is stored on the back-up storage media 126 (see FIG. 13) and presented to the user through the mounting procedure described 10 above. The system file 332 as illustrated in FIG. 19 is in tape format and the data portion 338 may include a plurality of data files (e.g., user files). The data begins at offset zero bytes (point 370) and ends 1032 bytes later at point 372. The written file 360 corresponds to a user's requests to write data to file 332. For example, the user may have modified two data files contained within system file 332, resulting in a written file 360 including W1 and W2. As 15 discussed above, this written file 360 may be stored separately from file 332 on the storage media so as not to alter the original back-up data. A logical modified system file 380 is illustrated and represents the file 332 including the changes (i.e., written files 360) that the user made through the write requests. In other words, in the modified system file 380, W1 and W2 (the user-modified data files) may be used to replace the original data files contained within the 20 data portion of original system file 332 without destroying the backed-up data.

As shown in FIG. 19, the modified system file corresponds to a logical combination (summation) of original system file 332 and written file 360. As shown, the original system file data 338 begins at offset zero as in the original file. At offset 64 (represented by reference numeral 384), the first portion W1 of modified data begins, and ends nine bytes further along 25 at offset 73 represented by reference numeral 386. Thus, W1, a user-modified data file resulting from a user write request, may be used to replace an original data file that was located offset 64 in the original system file 332. The length of W1 is shown to be 9 bytes because W1 exists from offset zero (390) in the written file 360 and ends at offset 9 (392) in the written file 360. The location of the beginning of W1 in the modified file (at offset 64 in the illustrated 30 example) is determined by information stored in the header 366, namely the relative relationship between the written file 360 and the original file 332. The W2 portion is also included in the modified file 380, beginning at offset 1032 (the original end of the file, represented by reference numeral 372) and logically extending the file by 100 bytes. Again,

the length of W2 is determined from information located in the header 366. The new end point of the file is represented by reference numeral 388.

It is to be appreciated that although the modified file 380 is logically created and represents the user-modified version of the original file, the newly written data, represented by 5 file 360 is not actually stored as part of the original file 332. Rather, as discussed above, the newly written data is stored at a specific location on the storage media identified for write data. In this manner, the integrity of the original back-up data is maintained, while allowing the user to apparently write to the mounted volume as they would to an ordinary local or network drive.

The modified file 380 includes a header 382 that includes a file descriptor that 10 represents the modified file. Referring to FIG. 20, there is illustrated an example of such a file descriptor 400. The file descriptor 400 includes a name field 402 that identifies the filename of the modified file 380 and a security field 404 that identifies permission attributes of the modified file 380. The file descriptor 400 also includes a plurality of data fields that include 15 pointers to the original file 332 and pointers to the written file 360 to capture the data stored in each of the original file and the written file. By sequentially following the linked list of pointers given in the file descriptor 400, a representation of the modified file 380 is given.

Referring to FIGS. 19 and 20, one specific example of a file descriptor for a modified file is illustrated and explained. In a first data field 406, there is located a pointer to a location of a first data file, in the modified file 380, which is at offset zero bytes, identified in FIG. 19 20 by reference numeral 408. The following field 410 indicates the length of the data file whose location is specified by pointer 406. In the illustrated example, the length is 64 bytes, as can be seen in FIG. 19 (the data extends between the zero offset point, reference numeral 408, and an offset of 64 bytes, represented by reference numeral 384). The next field 412 indicates that the next data file in modified file 380 is W1, as shown in FIG. 19. Thus, a pointer 414 indicates 25 that the location of the data corresponding to W1 is stored in newly written file 360 at the zero offset point (reference numeral 390 in FIG. 19). The length field 416 indicates that the length of W1 is 9 bytes, which can also be seen by looking at FIG. 19 – W1 extends between offset 64 (reference numeral 384) and offset 73 (reference numeral 386) in the modified file 380. The next field 418 indicates that the next data file in the modified file 380 is a data file from the 30 original system file 332. A pointer in field 420 indicates that the next data file is located at offset 73 in the modified file 380, as shown by reference numeral 386 in FIG. 19. Field 422 indicates that the length of the data file is 959 bytes, which can also be seen by referring to FIG. 19. The next field 424 indicates that the following data file is W2. Again, a pointer in

field 426 identifies the location of W2, namely newly written file 360 at offset 9, which can be seen by referring to FIG. 19. Field 428 indicates that the length of W2 is 100 bytes, and the next field 430 contains a null, indicating the W2 is the last data file in modified file 380, as is shown in FIG. 19. Thus, the file descriptor 400 contains a “roadmap” that identifies the 5 structure of modified file 380 and the location(s) of the data that is contained in modified file 380.

The volume restore application and methods described above represent sequential tape-formatted data in a form suitable for random-access I/O systems such as NFS or CIFS. Linked list file descriptors, such as file descriptor 400, can be used to translate the sequential tape-formatted data into randomly accessible data by recording the locations on the storage media of 10 each data file in a particular tar stream, for example, and also the locations of each data file in the tar stream relative to other data files in the tar stream. In addition, according to one embodiment, the volume restore application may include provisions for representing the changed (i.e., written) data back into a tape (e.g., tar) format so that a back-up/restore 15 application may access the data in a usual manner as described above. According to one embodiment, the instant restore application includes a facility that generates a virtual cartridge which is appropriately formatted with tape headers, pads, data and file markers, in the manner described above in relation to the file system software. In another embodiment, the volume 20 restore application may interface with the file system software to create virtual cartridges as discussed above that contain newly written and modified files.

It should be appreciated that although aspects of the present invention, such as the synthetic full back-up application, the end-user restore application, and the volume restore application are described herein primarily in terms of software, these and other aspects may alternatively be implemented in software, hardware or firmware, or any combination thereof. 25 Thus, for example, embodiments of the present invention may comprise any computer-readable medium (e.g., a computer memory, a floppy disk, a compact disk, a tape, etc.) encoded with a computer program (i.e., a plurality of instructions), which, when executed, at least in part, on a processor of a storage system, performs the functions of the synthetic full back-up application and/or the end-user restore application as described in detail above.

30 In general summary, embodiments and aspects of the invention thus include a storage system and methods that emulate a conventional tape back-up system but may provide enhanced functionality such as being able to create synthetic back-ups and allowing end users to view and restore backed-up files. However, it should be appreciated that various aspects of

- 30 -

the present invention may be used for other than the back-up of computer data. Because the storage system of the present invention may be used to economically store vast amounts of data, and that stored data can be accessed randomly, as opposed to sequentially, and at hard disk access times, embodiments of the present invention may find use outside of traditional 5 back-up storage systems. For example, embodiments of the present invention may be used to store video or audio data representing a wide selection of movies and music and enable video and/or audio on demand.

Having thus described several aspects of at least one embodiment of this invention, it is to be appreciated various alterations, modifications, and improvements will readily occur to 10 those skilled in the art. Such alterations, modifications, and improvements are intended to be part of this disclosure, and are intended to be within the scope of the invention. Accordingly, the foregoing description and drawings are by way of example only.

## CLAIMS

1. A method comprising acts of:
  - mounting a data volume on a host computer, the data volume comprising at least one data file, the data file corresponding to a most recently backed-up version of the at least one data file stored on a backup storage system; and
    - 5 storing, on the backup storage system, data corresponding to a second version of the at least one data file that is more recent than the most recently backed-up version of the at least one data file stored on the backup storage system while preserving the most recently backed-up version of the at least one data file.
- 10 2. The method of claim 1, further comprising an act of:
  - linking the most recently backed-up version of the at least one data file with the second version of the at least one data file.
- 15 3. The method of claim 1, further comprising an act of:
  - creating a data structure that identifies both the most recently backed-up version of the at least one data file and the second version of the at least one data file.
4. The method of claim 3, wherein the second version of the at least one data file is a
  - 20 modified version of the most recently backed-up version of the at least one data file.
5. The method of claim 1, wherein the act of mounting a data volume includes performing one of an NFS mount and a CIFS mount.
- 25 6. The method of claim 1, wherein the act of mounting the data volume includes building a file descriptor containing metadata relating to the most recently backed-up version of the least one data file, the metadata including an identifier that identifies a storage location on the back-up storage media of the most recently backed-up version of the at least one data file.
- 30 7. A back-up storage system comprising
  - a back-up storage media for storing the back-up data set; and
    - a controller including at least one processor configured to execute a set of instructions implementing the method of claim 1.

- 32 -

8. The back-up storage system of claim 7, wherein the back-up data set is a synthetic full back-up data set.

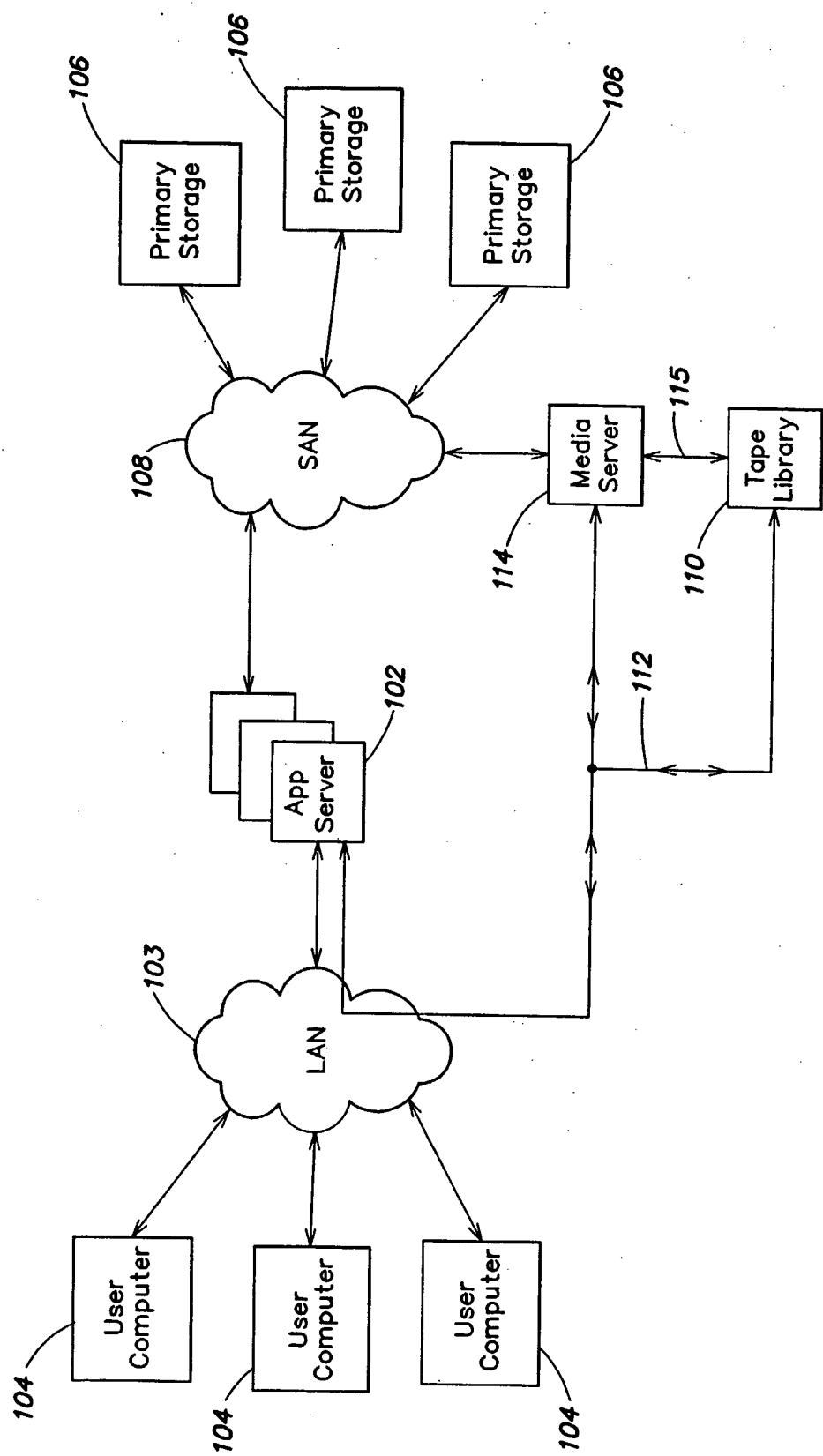
9. A computer readable medium encoded with a plurality of instructions that when 5 executed on at least one processor implement the method of claim 1.

10. The computer readable medium of claim 9, wherein the processor is included in a back-up storage system.

10 11. A computer readable medium having a data structure stored thereon, the data structure comprising:

a first identifier that uniquely identifies a system file that corresponds to a back-up data set that includes at least one data file; and

15 at least one second identifier that identifies a respective storage location on a storage medium where a most recent version of each at least one data file of the back-up data set is stored.



**FIG. 1**  
(Related Art)

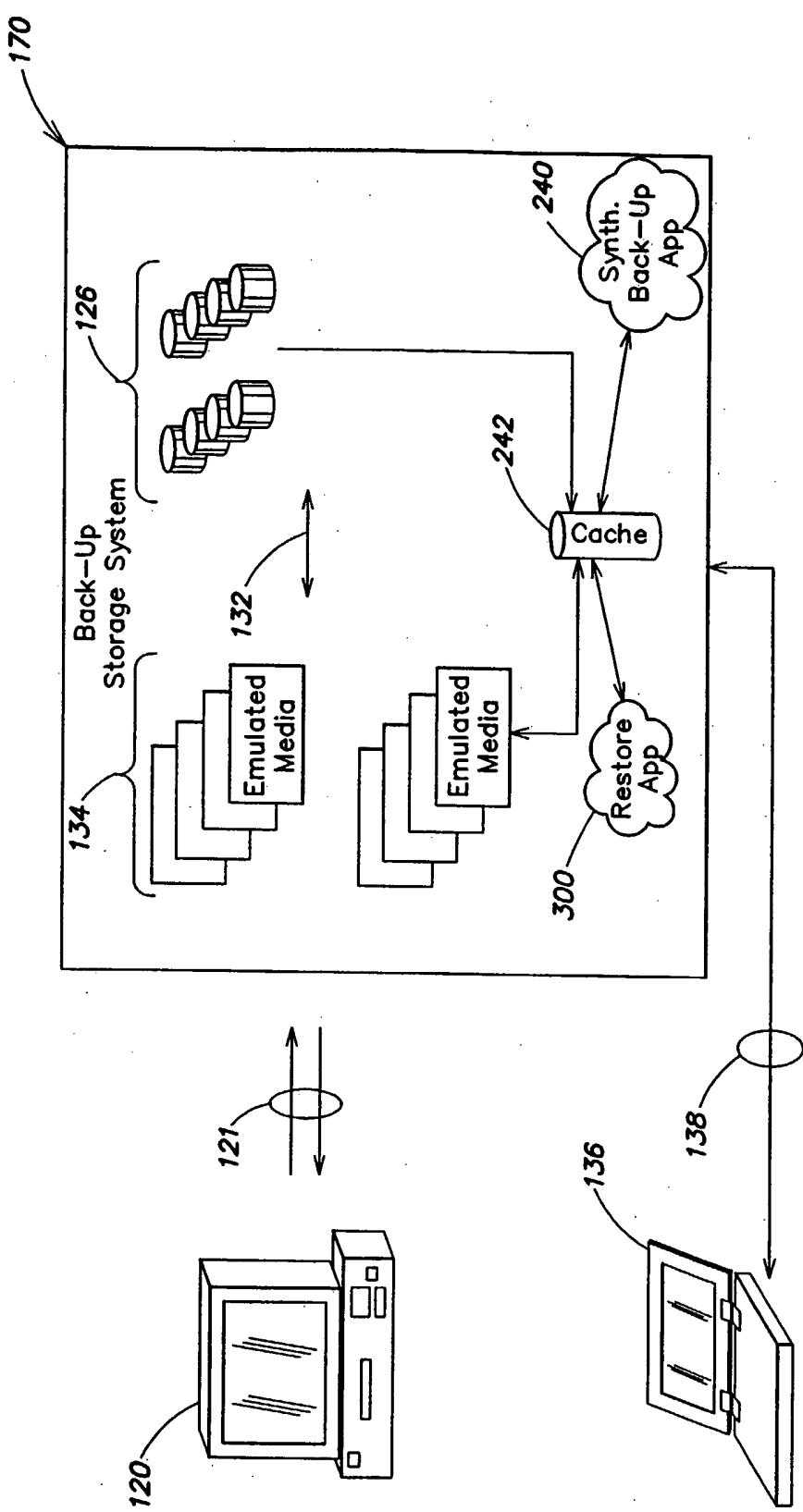


FIG. 2

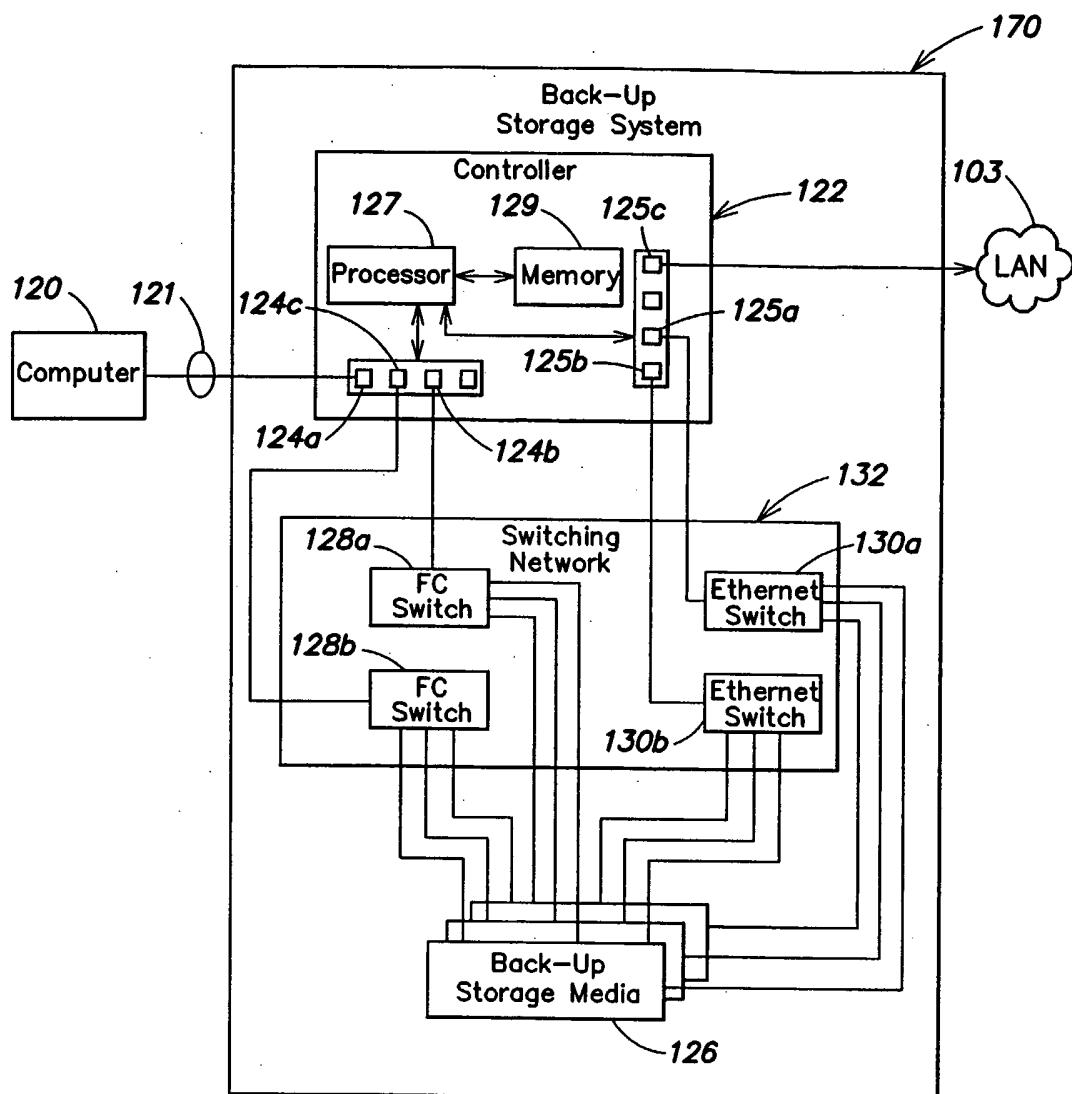
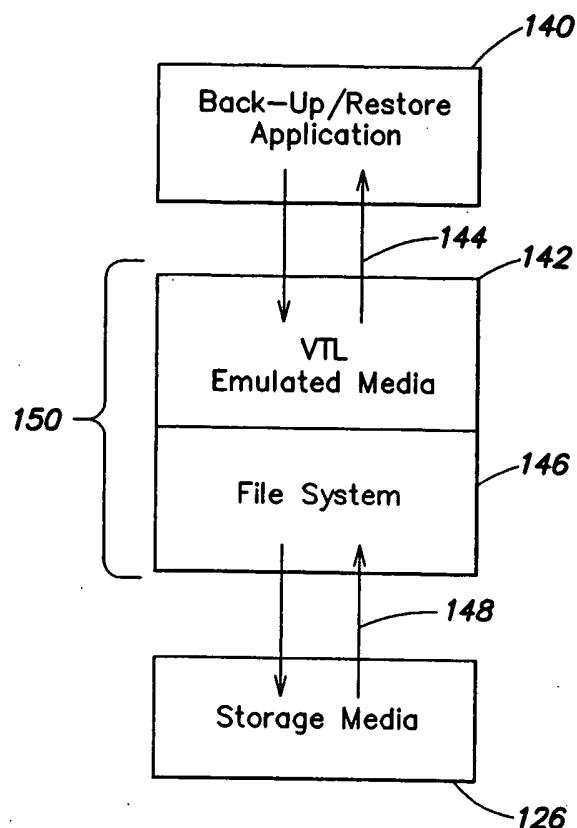
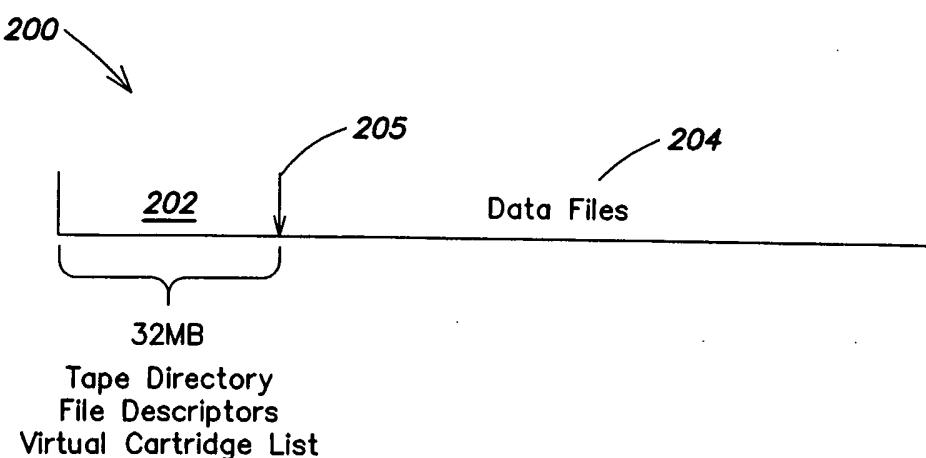
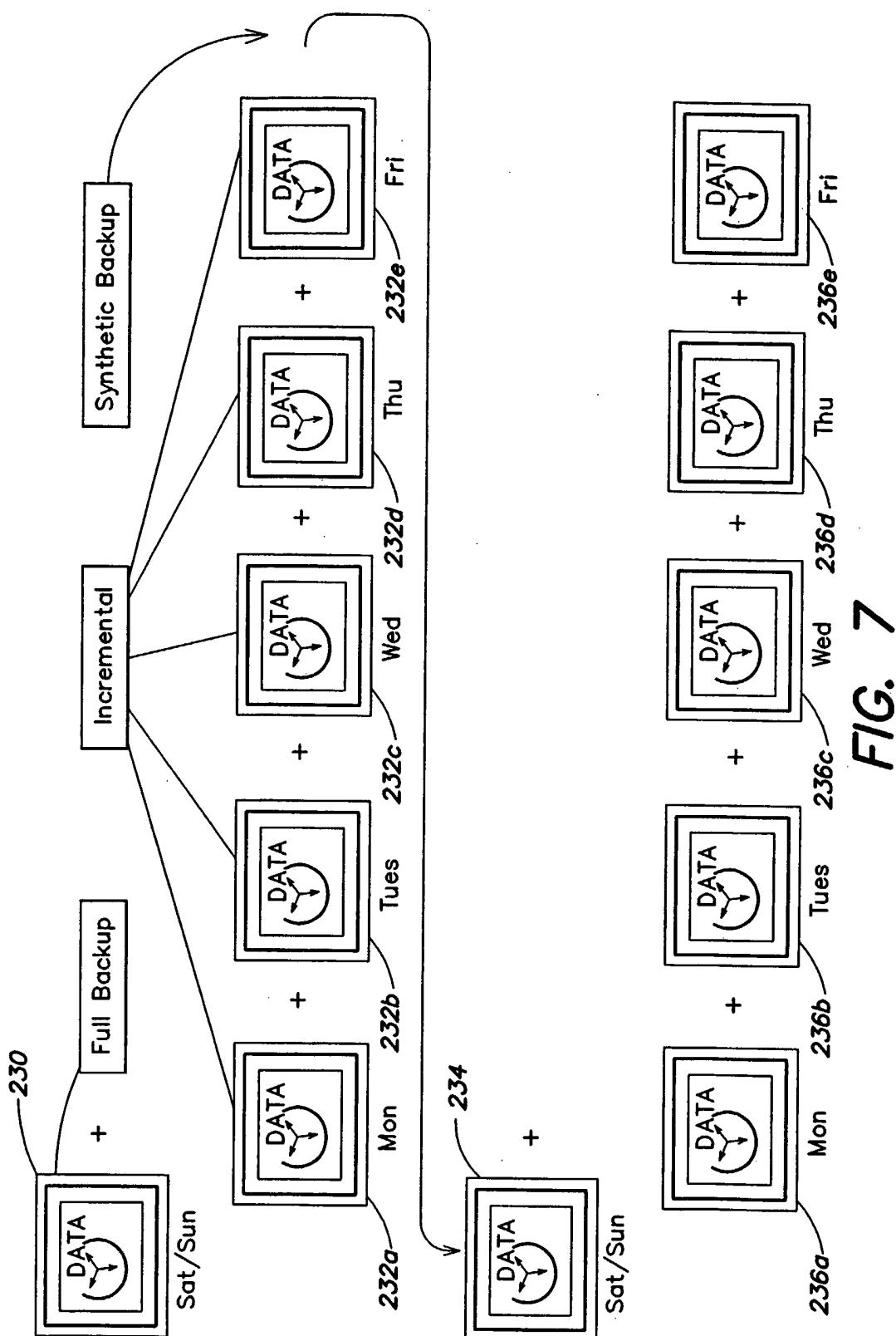


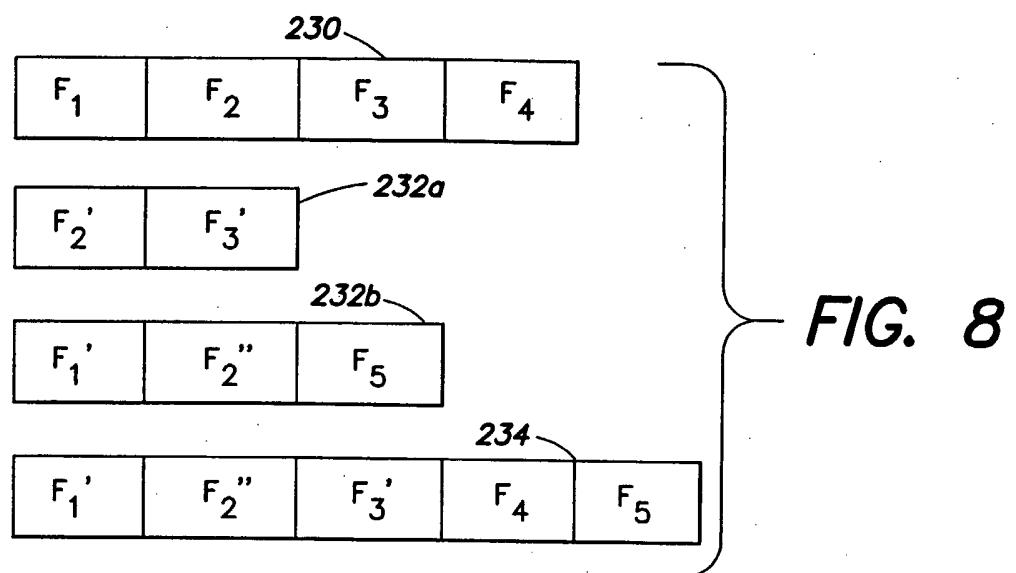
FIG. 3

**FIG. 4****FIG. 5**

Type	Size	Block Count
FM	-	1
Data	1024	4
Data	256k	100,000
FM	-	1
Data	1024	1
•	•	•
•	•	•
•	•	•
•	•	•

FIG. 6





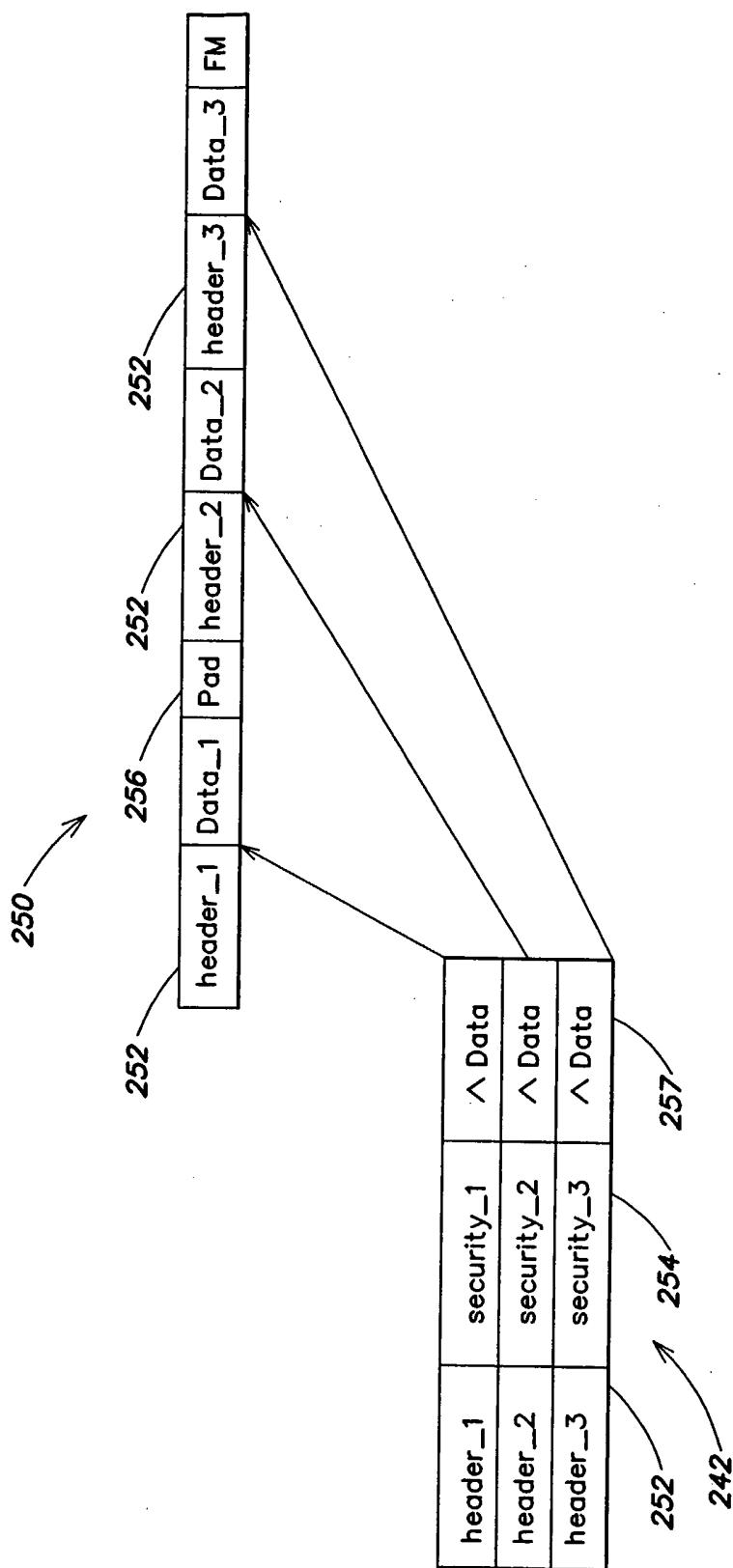


FIG. 9

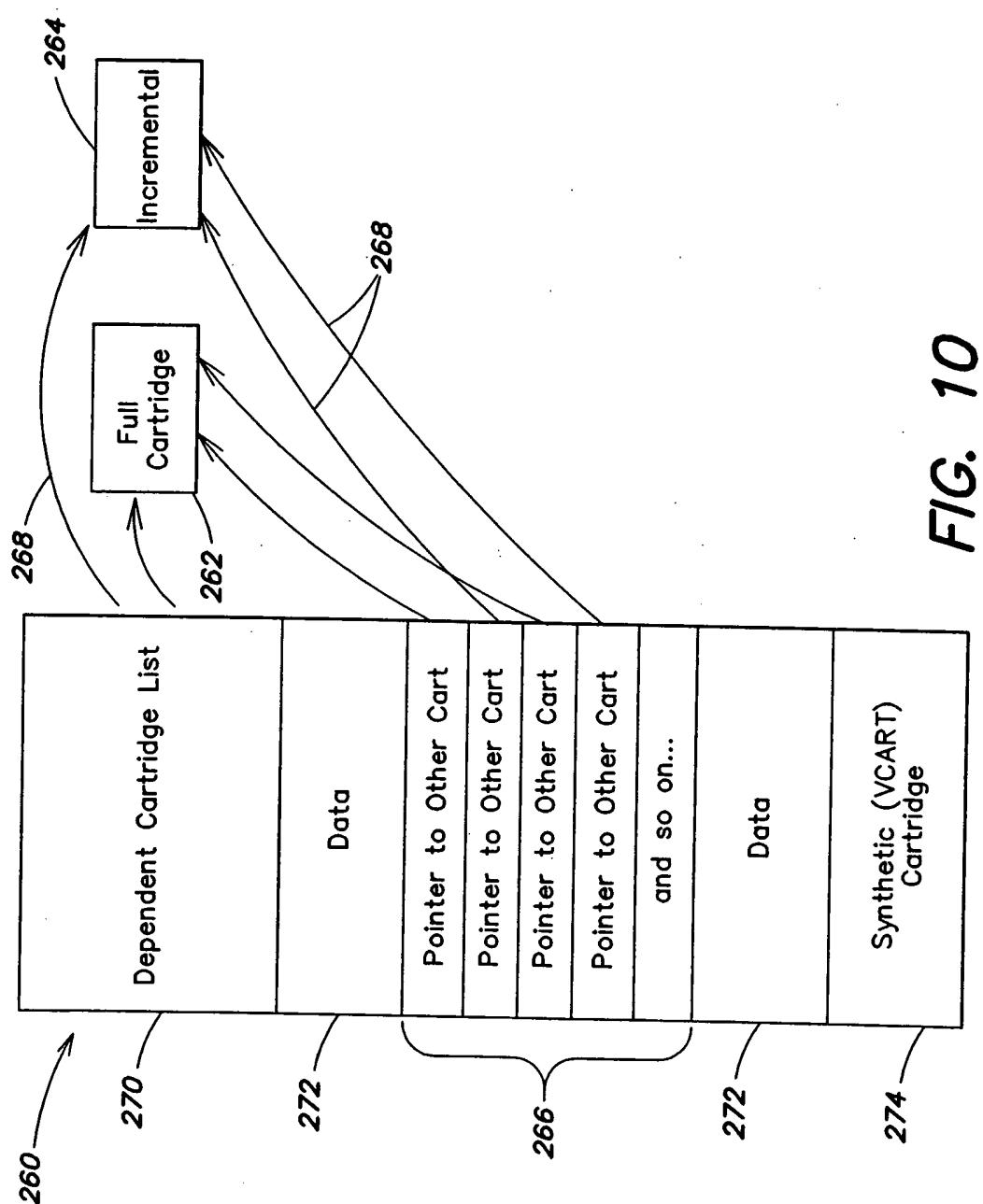
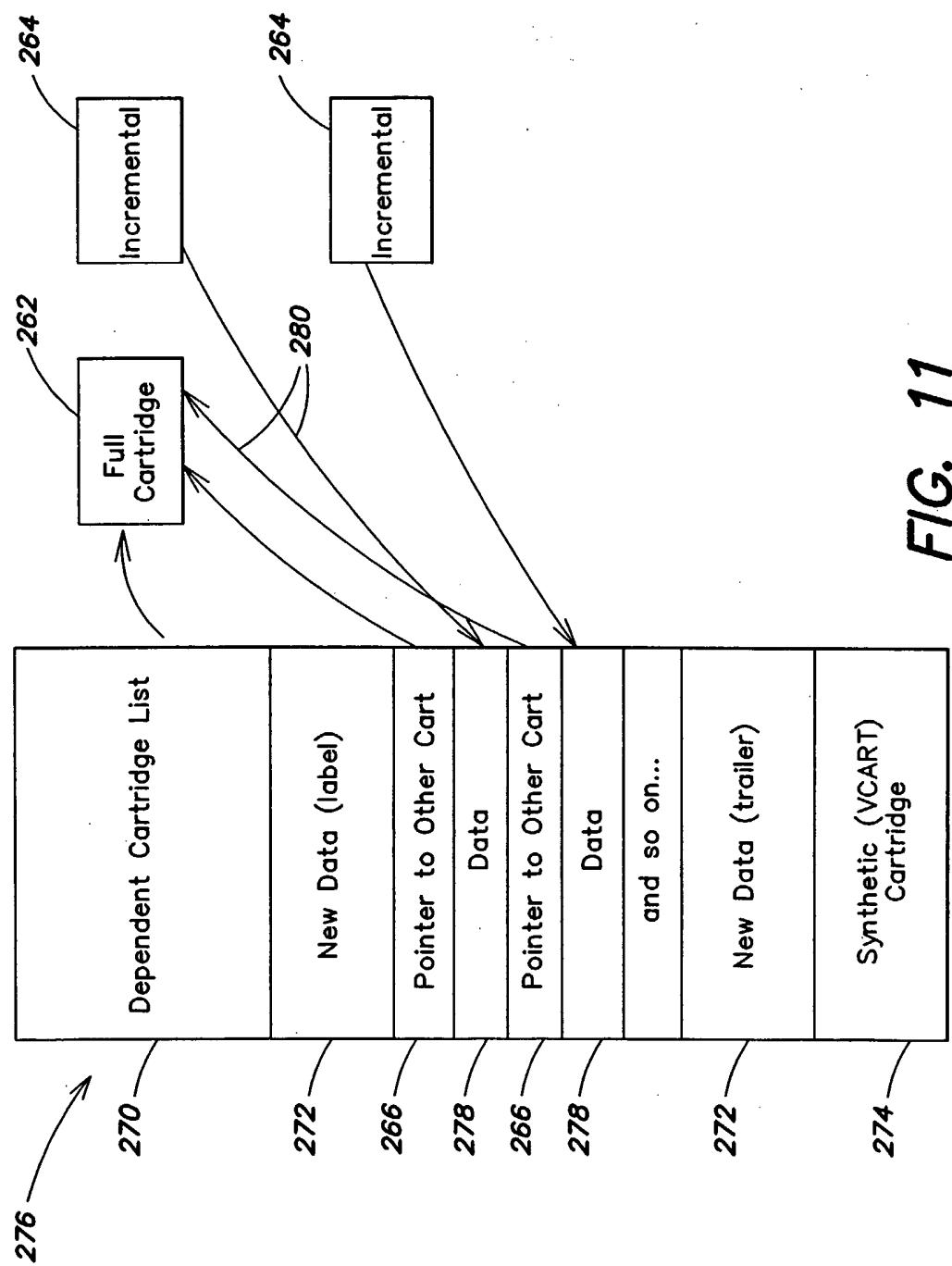
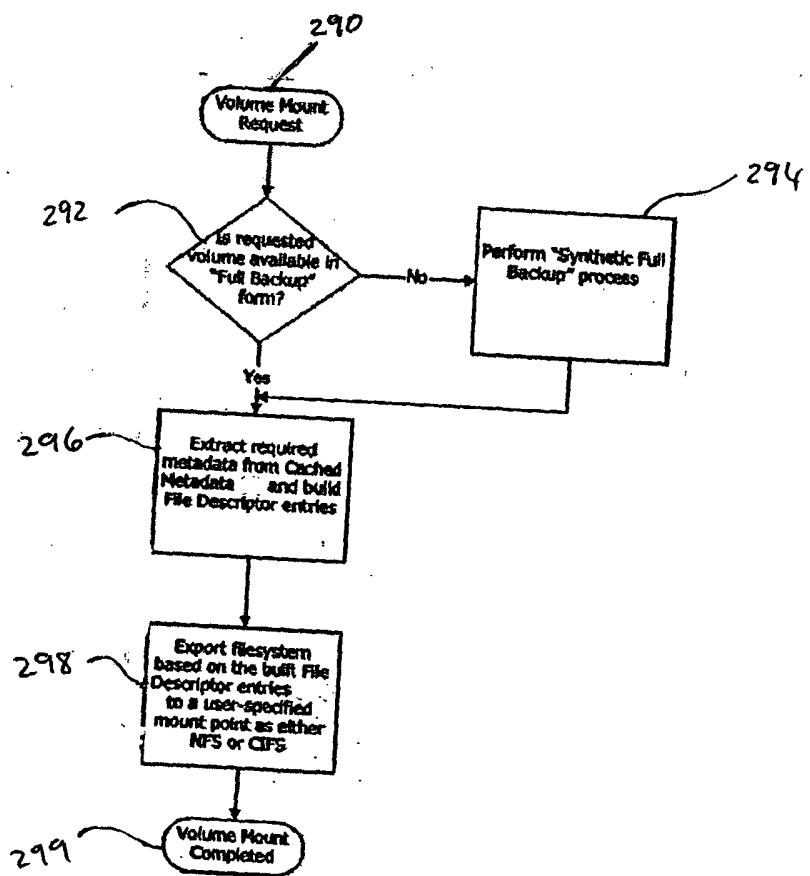
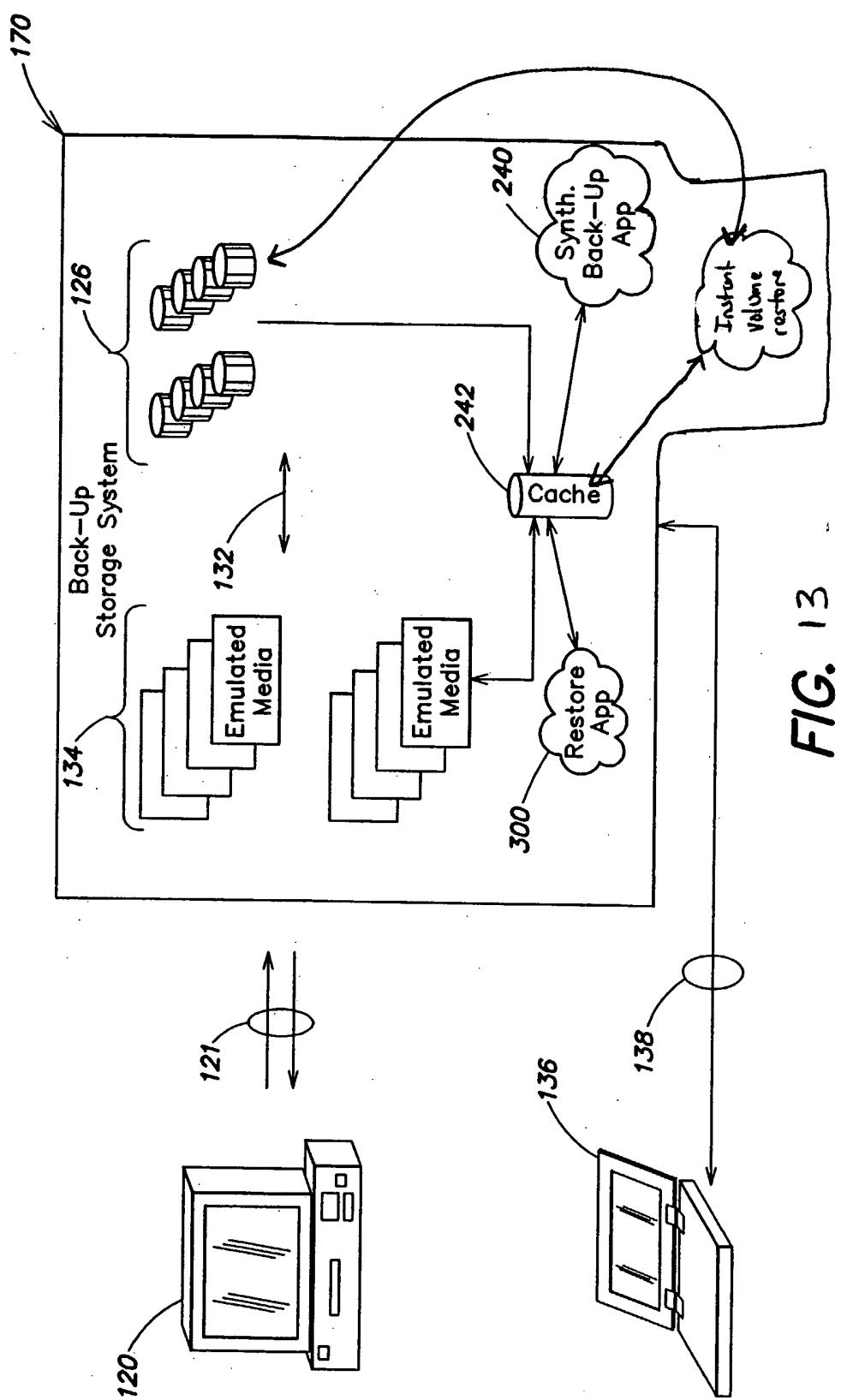


FIG. 10



**Volume Mount Flowchart****FIG. 12**



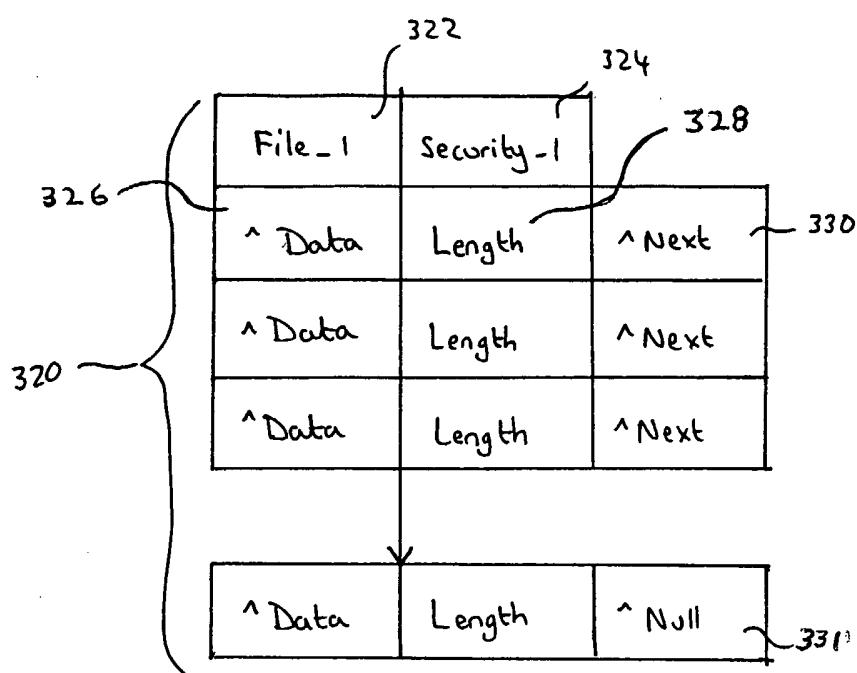


FIG. 14

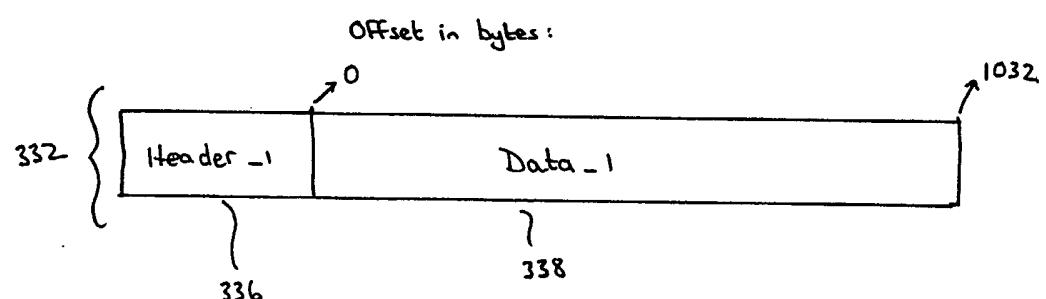


FIG. 15

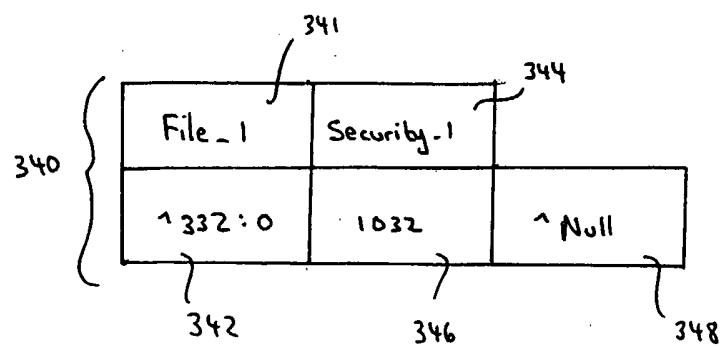
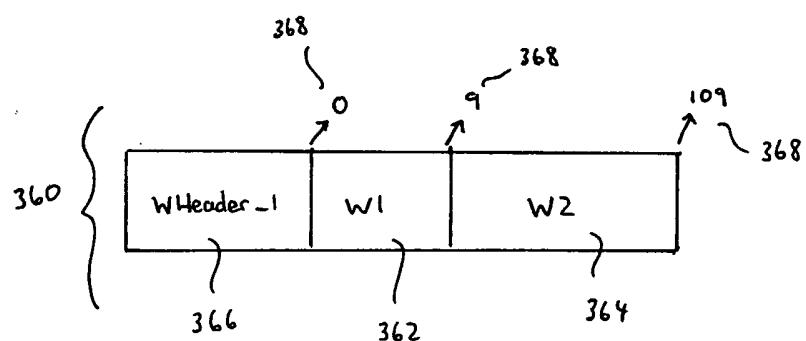
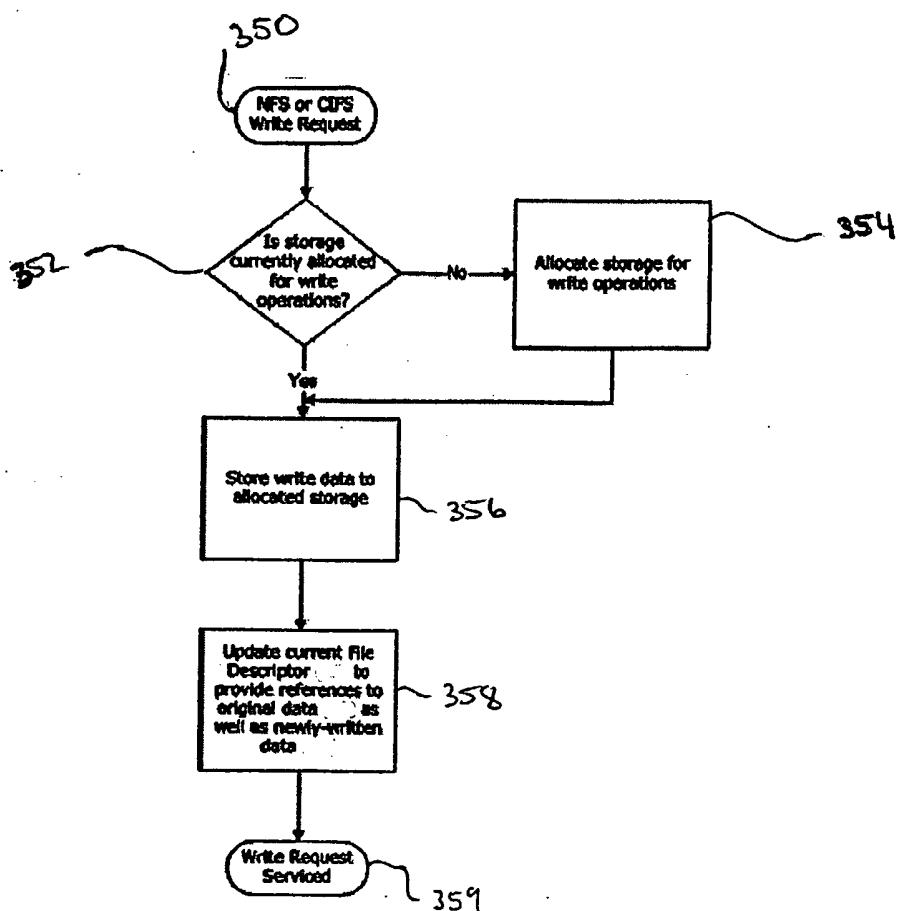


FIG. 16



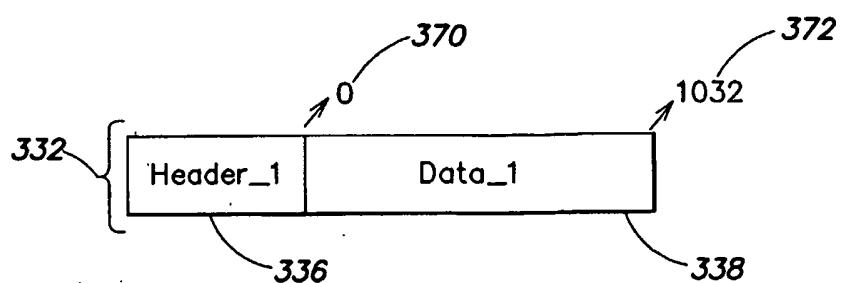
NFS or CIFS Write Data as Stored

FIG. 18

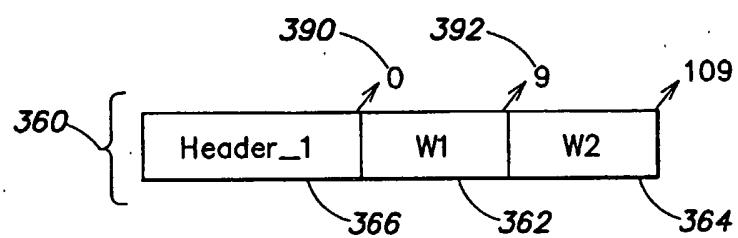


**NFS or CIFS Write Request Flowchart**

FIG.17



(+)



(=)

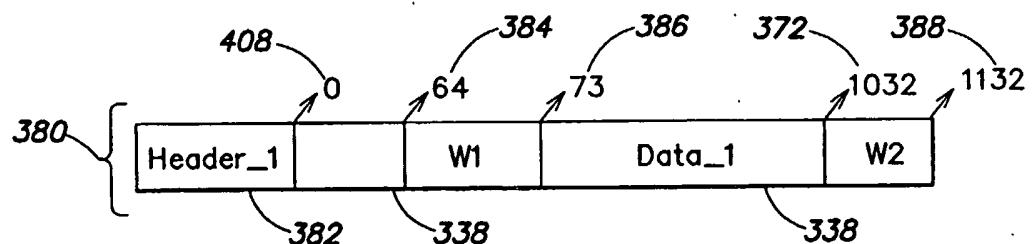
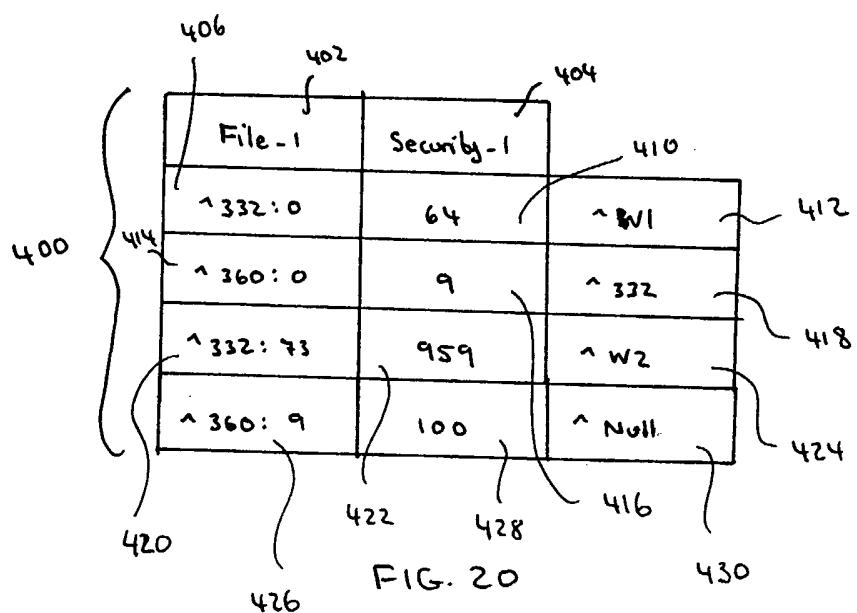


FIG. 19



# INTERNATIONAL SEARCH REPORT

International application No.

PCT/US04/32122

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(7) : G06F 12/00  
 US CL : 711/161,162,165; 707/202,203,204

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 711/161,162,165; 707/202,203,204

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)  
**EAST SEARCH**

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2003/0105912 A1 (NOREN) 05 June 2003 (05.06.2003), see abstract, col. 1, lines 20-27, col. 1, lines 34-55, col. 2, lines 44-62, col. 3, lines 15-45, col. 5, lines 15-19.	1-4, 6-11
---		-----
Y	US 6,230,190 B1 (EDMONDS et al.) 08 May 2001 (08.05.2001), see col. 4, lines 30-40.	5

<input type="checkbox"/> Further documents are listed in the continuation of Box C.	<input type="checkbox"/> See patent family annex.
* Special categories of cited documents:	
"A" document defining the general state of the art which is not considered to be of particular relevance	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E" earlier application or patent published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P" document published prior to the international filing date but later than the priority date claimed	
Date of the actual completion of the international search  30 January 2005 (30.01.2005)	Date of mailing of the international search report  <b>14 FEB 2005</b>
Name and mailing address of the ISA/US  Mail Stop PCT, Attn: ISA/US Commissioner for Patents P.O. Box 1450 Alexandria, Virginia 22313-1450 Facsimile No. (703) 305-3230	Authorized officer <i>Michelle L. Eason</i> <b>DONALD A SPARKS</b> Telephone No. (571) 272-4201