

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
9 November 2006 (09.11.2006)

PCT

(10) International Publication Number
WO 2006/118593 A2

(51) International Patent Classification:
G06F 3/06 (2006.01)

(US). **WYCKOFF HYER, Robert, Jr.**; 305 Covington Court, Mars, PA 16046 (US).

(21) International Application Number:
PCT/US2005/031220

(74) Agents: **DREGER, Duane, H.** et al.; Cesari and McKenna, LLP, 88 Black Falcon Avenue, Boston, MA 02210 (US).

(22) International Filing Date:
1 September 2005 (01.09.2005)

(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, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(25) Filing Language: English

(26) Publication Language: English

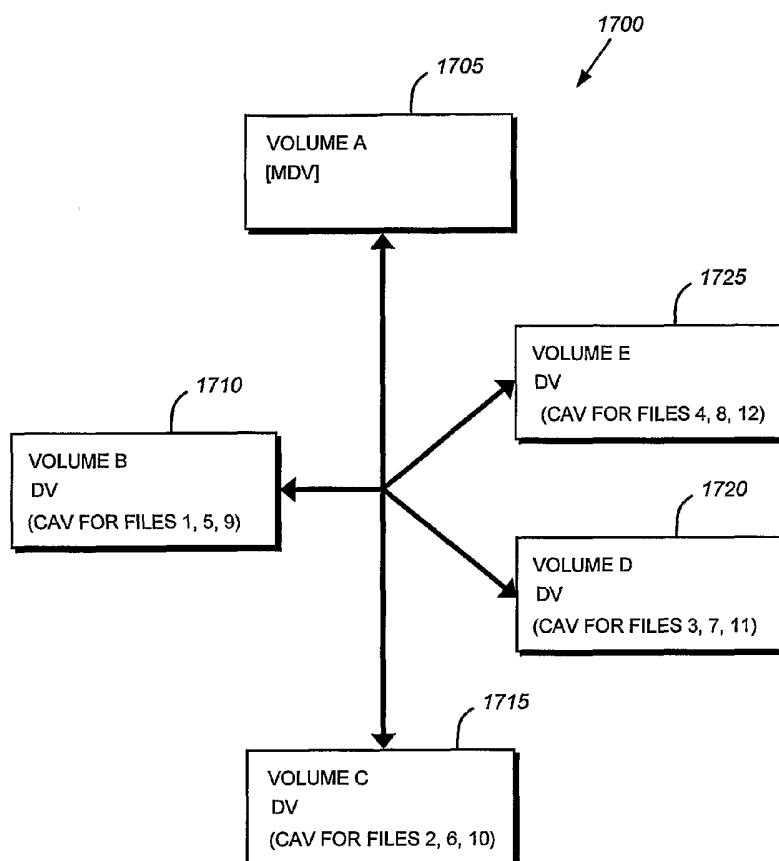
(30) Priority Data:
11/119,277 29 April 2005 (29.04.2005) US

(71) Applicant: **NETWORK APPLIANCE, INC.** [US/US];
495 East Java Drove, Sunnyvale, CA 94089 (US).

(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,

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR MULTI-TIERED META-DATA CACHING AND DISTRIBUTION IN A CLUSTERED COMPUTER ENVIRONMENT



(57) Abstract: A system and method caches and distributes meta-data for one or more data containers stored on a plurality of volumes configured as a striped volume set (SVS) and served by a plurality of nodes interconnected as a cluster. The SVS comprises one meta 5data volume (MDV) configured to store a canonical copy of certain meta-data, including access control lists and directories, associated with all data containers stored on the SVS, and one or more data volumes (DV) configured to store, at least, data content of those containers. In addition, for each data container stored on the SVS, one volume is designated a container attribute volume (CAV) and, as such, is configured to store ("cache") a io canonical copy of certain, rapidly-changing attribute meta-data, including time stamps and container length, associated with that container.



FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, 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).

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.

Published:

— *without international search report and to be republished
upon receipt of that report*

SYSTEM AND METHOD FOR MULTI-TIERED META-DATA CACHING AND DISTRIBUTION IN A CLUSTERED COMPUTER ENVIRONMENT

FIELD OF THE INVENTION

5 The present invention relates to clustered computer environments and, more particularly, to caching and distributing meta-data in a cluster.

BACKGROUND OF THE INVENTION

A storage system typically comprises one or more storage devices into which information may be entered, and from which information may be obtained, as desired. The storage system includes a storage operating system that functionally organizes the system by, *inter alia*, invoking storage operations in support of a storage service implemented by the system. The storage system may be implemented in accordance with a variety of storage architectures including, but not limited to, a network-attached storage environment, a storage area network and a disk assembly directly attached to a client or host computer. The storage devices are typically disk drives organized as a disk array, wherein the term "disk" commonly describes a self-contained rotating magnetic media storage device. The term disk in this context is synonymous with hard disk drive (HDD) or direct access storage device (DASD).

The storage operating system of the storage system may implement a high-level module, such as a file system, to logically organize the information stored on volumes as a hierarchical structure of data containers, such as files and logical units. For example, each "on-disk" file may be implemented as set of data structures, i.e., disk blocks, configured to store information, such as the actual data for the file. These data blocks are organized within a volume block number (vbn) space that is maintained by the file system. The file system may also assign each data block in the file a corresponding "file offset" or file block number (fbn). The file system typically assigns sequences of fbns on a per-file basis, whereas vbns are assigned over a larger volume address space. The file

system organizes the data blocks within the vbn space as a "logical volume"; each logical volume may be, although is not necessarily, associated with its own file system.

A known type of file system is a write-anywhere file system that does not overwrite data on disks. If a data block is retrieved (read) from disk into a memory of the storage system and "dirtied" (i.e., updated or modified) with new data, the data block is thereafter stored (written) to a new location on disk to optimize write performance. A write-anywhere file system may initially assume an optimal layout such that the data is substantially contiguously arranged on disks. The optimal disk layout results in efficient access operations, particularly for sequential read operations, directed to the disks. An example of a write-anywhere file system that is configured to operate on a storage system is the Write Anywhere File Layout (WAFL™) file system available from Network Appliance, Inc., Sunnyvale, California.

The storage system may be further configured to operate according to a client/server model of information delivery to thereby allow many clients to access data containers stored on the system. In this model, the client may comprise an application, such as a database application, executing on a computer that "connects" to the storage system over a computer network, such as a point-to-point link, shared local area network (LAN), wide area network (WAN), or virtual private network (VPN) implemented over a public network such as the Internet. Each client may request the services of the storage system by issuing file-based and block-based protocol messages (in the form of packets) to the system over the network.

A plurality of storage systems may be interconnected to provide a storage system environment configured to service many clients. Each storage system may be configured to service one or more volumes, wherein each volume stores one or more data containers. Yet often a large number of data access requests issued by the clients may be directed to a small number of data containers serviced by a particular storage system of the environment. A solution to such a problem is to distribute the volumes serviced by the particular storage system among all of the storage systems of the environment. This, in turn, distributes the data access requests, along with the processing resources needed to service such requests, among all of the storage systems, thereby reducing the individual process-

ing load on each storage system. However, a noted disadvantage arises when only a single data container, such as a file, is heavily accessed by clients of the storage system environment. As a result, the storage system attempting to service the requests directed to that data container may exceed its processing resources and become overburdened, with a concomitant degradation of speed and performance.

One technique for overcoming the disadvantages of having a single data container that is heavily utilized is to stripe the data container across a plurality of volumes configured as a striped volume set (SVS), where each volume is serviced by a different storage system, thereby distributing the load for the single data container among a plurality of storage systems. A technique for data container striping is described in the above-referenced U.S. Patent Application Serial No 11/119,278, entitled STORAGE SYSTEM ARCHITECTURE FOR STRIPING DATA CONTAINER CONTENT ACROSS VOLUMES OF A CLUSTER.

In one data container striping arrangement, meta-data (e.g., time stamps, file length and access control lists, (ACL) associated with a data container (e.g., a file) may also be striped across (or stored on) each volume of the SVS. However, a disadvantage of this arrangement involves the latency incurred when acquiring and updating such meta-data needed to service a data access request directed to the file. For example, changes to the meta-data resulting from servicing of the request may need to be propagated to every volume. This substantially increases the number of accesses to the storage systems serving the volumes, thereby adversely impacting the processing capabilities available for servicing other client data access requests.

An alternative arrangement may be to designate a single volume of the SVS as a meta-data volume (MDV) configured to store a canonical (i.e., definitive) copy of the meta-data associated with all files stored on the SVS. Here, each storage system hosting a SVS volume and servicing a data access request directed to a file is required to contact the MDV in order to ascertain and update the status of meta-data associated with the file. This arrangement places a substantial load on the storage system serving the MDV with a concomitant decrease in system performance. Moreover, depending on the load of the SVS, the meta-data requests to/from the MDV may become a bottleneck, causing certain

storage systems to stall (wait) until their meta-data requests have been processed before servicing client data access requests.

SUMMARY OF THE INVENTION

The present invention overcomes the disadvantages of the prior art by providing a system and method for caching and distributing meta-data for one or more data containers stored on a plurality of volumes configured as a striped volume set (SVS) and served by a plurality of nodes interconnected as a cluster. To that end, the SVS comprises one meta-data volume (MDV) configured to store a canonical copy of certain meta-data, including access control lists and directories, associated with all data containers stored on the SVS, and one or more data volumes (DV) configured to store, at least, data content of those containers. In addition, for each data container stored on the SVS, one volume is designated a container attribute volume (CAV) and, as such, is configured to store a canonical copy of certain, rapidly-changing attribute meta-data, including time stamps and container length, associated with that container.

According to an aspect of the invention, the meta-data are organized into various categories (e.g., MD1-MD3) along functional boundaries and are resident on various volumes to optimize data access (e.g., read and write) paths through the SVS. These categories include (i) MD1 meta-data that changes on every data access (read/write) request served by the SVS, (ii) MD2 meta-data that may be retrieved (but not changed) on every request and (iii) MD3 meta-data that is unused for the read/write requests. Since it changes on every read/write request served by the DVs of the SVS, the MD1 meta-data is canonically resident on the CAV and generally cached on the DVs. Likewise, since it may be retrieved, but does not change, on every request served by the DVs, the MD2 meta-data is canonically resident on the MDV and generally cached on all DVs of the SVS, including the volume designated as CAV. Finally, since it is unused for a read/write request, the MD3 meta-data is canonically, and solely, resident on the MDV.

According to another aspect of the invention, the CAV cooperates with the MDV and DVs of the SVS to provide a multi-tier caching and distribution architecture that off-loads meta-data access requests that would otherwise be directed to the MDV. Here, the

CAV is utilized as a first tier caching and distribution point for storing and distributing most meta-data, e.g., MD1 and MD2, for use by the DVs of the SVS. In the case of the MD1 meta-data, the multi-tier caching and distribution architecture enables the DVs to constantly access the CAV for meta-data changes that occur on every read/write request.

5 The MD1 access path is relatively “short” and, notably, offloads all such frequent requests from the MDV. In the case of MD2 meta-data, the multi-tier caching and distribution architecture enables the DV to access the CAV for meta-data retrieved on every read/write request. In general, the MD2 access path terminates at the CAV for current meta-data; however, since it does not store a canonical copy of the MD2 meta-data, the

10 CAV may have to acquire a current version of the meta-data from the MDV if its copy is invalid.

Advantageously, the present invention facilitates (i) propagating of meta-data throughout the nodes of the cluster for storage on various volumes (“caches”) of the SVS and (ii) updating of those caches in an efficient manner. That is, the novel multi-tier

15 caching and distribution architecture reduces the latency incurred when acquiring and updating such meta-data needed to service a data access request directed to a data container to thereby provide a high-performance cluster.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and further advantages of invention may be better understood by referring to the following description in conjunction with the accompanying drawings in

20 which like reference numerals indicate identical or functionally similar elements:

Fig. 1 is a schematic block diagram of a plurality of nodes interconnected as a cluster in accordance with an embodiment of the present invention;

Fig. 2 is a schematic block diagram of a node in accordance with an embodiment

25 of the present invention;

Fig. 3 is a schematic block diagram of a storage operating system that may be advantageously used with the present invention;

Fig. 4 is a schematic block diagram illustrating the format of a cluster fabric (CF) message in accordance with an embodiment of with the present invention;

Fig. 5 is a schematic block diagram illustrating the format of a data container handle in accordance with an embodiment of the present invention;

Fig. 6 is a schematic block diagram of an exemplary inode in accordance with an embodiment of the present invention;

5 Fig. 7 is a schematic block diagram of an exemplary buffer tree in accordance with an embodiment of the present invention;

Fig. 8 is a schematic block diagram of an illustrative embodiment of a buffer tree of a file that may be advantageously used with the present invention;

10 Fig. 9 is a schematic block diagram of an exemplary aggregate in accordance with an embodiment of the present invention;

Fig. 10 is a schematic block diagram of an exemplary on-disk layout of the aggregate in accordance with an embodiment of the present invention;

Fig. 11 is a schematic block diagram illustrating a collection of management processes in accordance with an embodiment of the present invention;

15 Fig. 12 is a schematic block diagram of a volume location database (VLDB) volume entry in accordance with an embodiment of the present invention;

Fig. 13 is a schematic block diagram of a VLDB aggregate entry in accordance with an embodiment of the present invention;

20 Fig. 14 is a schematic block diagram of a striped volume set (SVS) in accordance with an embodiment of the present invention;

Fig. 15 is a schematic block diagram of a VLDB SVS entry in accordance with an embodiment of the present invention;

25 Fig. 16 is a schematic block diagram illustrating the periodic sparseness of file content stored on volumes of a SVS in accordance with an embodiment of the present invention;

Fig. 17 is a schematic block diagram of an exemplary striped volume set in accordance with an embodiment of the present invention;

Fig. 18 is a schematic block diagram of an exemplary striped volume set in accordance with an embodiment of the present invention

30 Fig. 19 is a flowchart of the steps of a procedure for modifying/creating attributes in accordance with an embodiment of the present invention; and

Fig. 20 is a flowchart of a procedure for processing a data access operation in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF AN ILLUSTRATIVE EMBODIMENT

5 A. Cluster Environment

Fig. 1 is a schematic block diagram of a plurality of nodes 200 interconnected as a cluster 100 and configured to provide storage service relating to the organization of information on storage devices. The nodes 200 comprise various functional components that cooperate to provide a distributed storage system architecture of the cluster 100. To
10 that end, each node 200 is generally organized as a network element (N-blade 310) and a disk element (D-blade 350). The N-blade 310 includes functionality that enables the node 200 to connect to clients 180 over a computer network 140, while each D-blade 350 connects to one or more storage devices, such as disks 130 of a disk array 120. The nodes 200 are interconnected by a cluster switching fabric 150 which, in the illustrative
15 embodiment, may be embodied as a Gigabit Ethernet switch. An exemplary distributed file system architecture is generally described in U.S. Patent Application Publication No. US 2002/0116593 titled METHOD AND SYSTEM FOR RESPONDING TO FILE SYSTEM REQUESTS, by M. Kazar et al. published August 22, 2002. It should be noted that while there is shown an equal number of N and D-blades in the illustrative cluster
20 100, there may be differing numbers of N and/or D-blades in accordance with various embodiments of the present invention. For example, there may be a plurality of N-blades and/or D-blades interconnected in a cluster configuration 100 that does not reflect a one-to-one correspondence between the N and D-blades. As such, the description of a node 200 comprising one N-blade and one D-blade should be taken as illustrative only.

25 The clients 180 may be general-purpose computers configured to interact with the node 200 in accordance with a client/server model of information delivery. That is, each client may request the services of the node, and the node may return the results of the services requested by the client, by exchanging packets over the network 140. The client may issue packets including file-based access protocols, such as the Common Internet

File System (CIFS) protocol or Network File System (NFS) protocol, over the Transmission Control Protocol/Internet Protocol (TCP/IP) when accessing information in the form of files and directories. Alternatively, the client may issue packets including block-based access protocols, such as the Small Computer Systems Interface (SCSI) protocol encapsulated over TCP (iSCSI) and SCSI encapsulated over Fibre Channel (FCP), when ac-
5 cessing information in the form of blocks.

B. Storage System Node

Fig. 2 is a schematic block diagram of a node 200 that is illustratively embodied as a storage system comprising a plurality of processors 222a,b, a memory 224, a net-
10 work adapter 225, a cluster access adapter 226, a storage adapter 228 and local storage 230 interconnected by a system bus 223. The local storage 230 comprises one or more storage devices, such as disks, utilized by the node to locally store configuration information (e.g., in configuration table 235) provided by one or more management processes that execute as user mode applications 1100 (see Fig. 11). The cluster access adapter 226
15 comprises a plurality of ports adapted to couple the node 200 to other nodes of the cluster 100. In the illustrative embodiment, Ethernet is used as the clustering protocol and interconnect media, although it will be apparent to those skilled in the art that other types of protocols and interconnects may be utilized within the cluster architecture described herein. In alternate embodiments where the N-blades and D-blades are implemented on
20 separate storage systems or computers, the cluster access adapter 226 is utilized by the N/D-blade for communicating with other N/D-blades in the cluster 100.

Each node 200 is illustratively embodied as a dual processor storage system executing a storage operating system 300 that preferably implements a high-level module, such as a file system, to logically organize the information as a hierarchical structure of
25 named directories, files and special types of files called virtual disks (hereinafter generally "blocks") on the disks. However, it will be apparent to those of ordinary skill in the art that the node 200 may alternatively comprise a single or more than two processor system. Illustratively, one processor 222a executes the functions of the N-blade 310 on the node, while the other processor 222b executes the functions of the D-blade 350.

The memory 224 illustratively comprises storage locations that are addressable by the processors and adapters for storing software program code and data structures associated with the present invention. The processor and adapters may, in turn, comprise processing elements and/or logic circuitry configured to execute the software code and manipulate the data structures. The storage operating system 300, portions of which is typically resident in memory and executed by the processing elements, functionally organizes the node 200 by, *inter alia*, invoking storage operations in support of the storage service implemented by the node. It will be apparent to those skilled in the art that other processing and memory means, including various computer readable media, may be used for storing and executing program instructions pertaining to the invention described herein.

The network adapter 225 comprises a plurality of ports adapted to couple the node 200 to one or more clients 180 over point-to-point links, wide area networks, virtual private networks implemented over a public network (Internet) or a shared local area network. The network adapter 225 thus may comprise the mechanical, electrical and signaling circuitry needed to connect the node to the network. Illustratively, the computer network 140 may be embodied as an Ethernet network or a Fibre Channel (FC) network. Each client 180 may communicate with the node over network 140 by exchanging discrete frames or packets of data according to pre-defined protocols, such as TCP/IP.

The storage adapter 228 cooperates with the storage operating system 300 executing on the node 200 to access information requested by the clients. The information may be stored on any type of attached array of writable storage device media such as video tape, optical, DVD, magnetic tape, bubble memory, electronic random access memory, micro-electro mechanical and any other similar media adapted to store information, including data and parity information. However, as illustratively described herein, the information is preferably stored on the disks 130 of array 120. The storage adapter comprises a plurality of ports having input/output (I/O) interface circuitry that couples to the disks over an I/O interconnect arrangement, such as a conventional high-performance, FC link topology.

Storage of information on each array 120 is preferably implemented as one or more storage "volumes" that comprise a collection of physical storage disks 130 cooper-

ating to define an overall logical arrangement of volume block number (vbn) space on the volume(s). Each logical volume is generally, although not necessarily, associated with its own file system. The disks within a logical volume/file system are typically organized as one or more groups, wherein each group may be operated as a Redundant Array of Independent (or Inexpensive) Disks (RAID). Most RAID implementations, such as a RAID-4 level implementation, enhance the reliability/integrity of data storage through the redundant writing of data "stripes" across a given number of physical disks in the RAID group, and the appropriate storing of parity information with respect to the striped data. An illustrative example of a RAID implementation is a RAID-4 level implementation, although it should be understood that other types and levels of RAID implementations may be used in accordance with the inventive principles described herein.

C. Storage Operating System

To facilitate access to the disks 130, the storage operating system 300 implements a write-anywhere file system that cooperates with one or more virtualization modules to "virtualize" the storage space provided by disks 130. The file system logically organizes the information as a hierarchical structure of named directories and files on the disks. Each "on-disk" file may be implemented as set of disk blocks configured to store information, such as data, whereas the directory may be implemented as a specially formatted file in which names and links to other files and directories are stored. The virtualization module(s) allow the file system to further logically organize information as a hierarchical structure of blocks on the disks that are exported as named logical unit numbers (luns).

In the illustrative embodiment, the storage operating system is preferably the NetApp® Data ONTAP™ operating system available from Network Appliance, Inc., Sunnyvale, California that implements a Write Anywhere File Layout (WAFL™) file system. However, it is expressly contemplated that any appropriate storage operating system may be enhanced for use in accordance with the inventive principles described herein. As such, where the term "WAFL" is employed, it should be taken broadly to refer to any storage operating system that is otherwise adaptable to the teachings of this invention.

Fig. 3 is a schematic block diagram of the storage operating system 300 that may be advantageously used with the present invention. The storage operating system com-

prises a series of software layers organized to form an integrated network protocol stack or, more generally, a multi-protocol engine 325 that provides data paths for clients to access information stored on the node using block and file access protocols. The multi-protocol engine includes a media access layer 312 of network drivers (e.g., gigabit Ethernet drivers) that interfaces to network protocol layers, such as the IP layer 314 and its supporting transport mechanisms, the TCP layer 316 and the User Datagram Protocol (UDP) layer 315. A file system protocol layer provides multi-protocol file access and, to that end, includes support for the Direct Access File System (DAFS) protocol 318, the NFS protocol 320, the CIFS protocol 322 and the Hypertext Transfer Protocol (HTTP) protocol 324. A VI layer 326 implements the VI architecture to provide direct access transport (DAT) capabilities, such as RDMA, as required by the DAFS protocol 318. An iSCSI driver layer 328 provides block protocol access over the TCP/IP network protocol layers, while a FC driver layer 330 receives and transmits block access requests and responses to and from the node. The FC and iSCSI drivers provide FC-specific and iSCSI-specific access control to the blocks and, thus, manage exports of luns to either iSCSI or FCP or, alternatively, to both iSCSI and FCP when accessing the blocks on the node 200.

In addition, the storage operating system includes a series of software layers organized to form a storage server 365 that provides data paths for accessing information stored on the disks 130 of the node 200. To that end, the storage server 365 includes a file system module 360 in cooperating relation with a volume striping module (VSM) 370, a RAID system module 380 and a disk driver system module 390. The RAID system 380 manages the storage and retrieval of information to and from the volumes/disks in accordance with I/O operations, while the disk driver system 390 implements a disk access protocol such as, e.g., the SCSI protocol. The VSM 370 illustratively implements a striped volume set (SVS) of the present invention. As described further herein, the VSM cooperates with the file system 360 to enable storage server 365 to service a volume of the SVS. In particular, the VSM 370 implements a Locate() function 375 to compute the location of data container content in the SVS volume to thereby ensure consistency of such content served by the cluster.

The file system 360 implements a virtualization system of the storage operating system 300 through the interaction with one or more virtualization modules illustratively

embodied as, e.g., a virtual disk (vdisk) module (not shown) and a SCSI target module 335. The vdisk module enables access by administrative interfaces, such as a user interface of a management framework 1110 (see Fig. 11), in response to a user (system administrator) issuing commands to the node 200. The SCSI target module 335 is generally
5 disposed between the FC and iSCSI drivers 328, 330 and the file system 360 to provide a translation layer of the virtualization system between the block (lun) space and the file system space, where luns are represented as blocks.

The file system 360 is illustratively a message-based system that provides logical volume management capabilities for use in access to the information stored on the storage devices, such as disks. That is, in addition to providing file system semantics, the file
10 system 360 provides functions normally associated with a volume manager. These functions include (i) aggregation of the disks, (ii) aggregation of storage bandwidth of the disks, and (iii) reliability guarantees, such as mirroring and/or parity (RAID). The file system 360 illustratively implements the WAFL file system (hereinafter generally the
15 “write-anywhere file system”) having an on-disk format representation that is block-based using, e.g., 4 kilobyte (kB) blocks and using index nodes (“inodes”) to identify files and file attributes (such as creation time, access permissions, size and block location). The file system uses files to store meta-data describing the layout of its file system; these meta-data files include, among others, an inode file. A file handle, i.e., an identifier
20 that includes an inode number, is used to retrieve an inode from disk.

Broadly stated, all inodes of the write-anywhere file system are organized into the inode file. A file system (fs) info block specifies the layout of information in the file system and includes an inode of a file that includes all other inodes of the file system. Each logical volume (file system) has an fsinfo block that is preferably stored at a fixed location within, e.g., a RAID group. The inode of the inode file may directly reference (point
25 to) data blocks of the inode file or may reference indirect blocks of the inode file that, in turn, reference data blocks of the inode file. Within each data block of the inode file are embedded inodes, each of which may reference indirect blocks that, in turn, reference data blocks of a file.

Operationally, a request from the client 180 is forwarded as a packet over the computer network 140 and onto the node 200 where it is received at the network adapter 225. A network driver (of layer 312 or layer 330) processes the packet and, if appropriate, passes it on to a network protocol and file access layer for additional processing prior to forwarding to the write-anywhere file system 360. Here, the file system generates operations to load (retrieve) the requested data from disk 130 if it is not resident "in core", i.e., in memory 224. If the information is not in memory, the file system 360 indexes into the inode file using the inode number to access an appropriate entry and retrieve a logical vbn. The file system then passes a message structure including the logical vbn to the RAID system 380; the logical vbn is mapped to a disk identifier and disk block number (disk,dbn) and sent to an appropriate driver (e.g., SCSI) of the disk driver system 390. The disk driver accesses the dbn from the specified disk 130 and loads the requested data block(s) in memory for processing by the node. Upon completion of the request, the node (and operating system) returns a reply to the client 180 over the network 140.

It should be noted that the software "path" through the storage operating system layers described above needed to perform data storage access for the client request received at the node may alternatively be implemented in hardware. That is, in an alternate embodiment of the invention, a storage access request data path may be implemented as logic circuitry embodied within a field programmable gate array (FPGA) or an application specific integrated circuit (ASIC). This type of hardware implementation increases the performance of the storage service provided by node 200 in response to a request issued by client 180. Moreover, in another alternate embodiment of the invention, the processing elements of adapters 225, 228 may be configured to offload some or all of the packet processing and storage access operations, respectively, from processor 222, to thereby increase the performance of the storage service provided by the node. It is expressly contemplated that the various processes, architectures and procedures described herein can be implemented in hardware, firmware or software.

As used herein, the term "storage operating system" generally refers to the computer-executable code operable on a computer to perform a storage function that manages data access and may, in the case of a node 200, implement data access semantics of a general purpose operating system. The storage operating system can also be imple-

mented as a microkernel, an application program operating over a general-purpose operating system, such as UNIX® or Windows NT®, or as a general-purpose operating system with configurable functionality, which is configured for storage applications as described herein.

5 In addition, it will be understood to those skilled in the art that the invention described herein may apply to any type of special-purpose (e.g., file server, filer or storage serving appliance) or general-purpose computer, including a standalone computer or portion thereof, embodied as or including a storage system. Moreover, the teachings of this invention can be adapted to a variety of storage system architectures including, but not
10 limited to, a network-attached storage environment, a storage area network and disk assembly directly-attached to a client or host computer. The term “storage system” should therefore be taken broadly to include such arrangements in addition to any subsystems configured to perform a storage function and associated with other equipment or systems. It should be noted that while this description is written in terms of a write any where file
15 system, the teachings of the present invention may be utilized with any suitable file system, including a write in place file system.

D. CF Protocol

In the illustrative embodiment, the storage server 365 is embodied as D-blade 350 of the storage operating system 300 to service one or more volumes of array 120. In addition,
20 the multi-protocol engine 325 is embodied as N-blade 310 to (i) perform protocol termination with respect to a client issuing incoming data access request packets over the network 140, as well as (ii) redirect those data access requests to any storage server 365 of the cluster 100. Moreover, the N-blade 310 and D-blade 350 cooperate to provide a highly-scalable, distributed storage system architecture of the cluster 100. To that end,
25 each blade includes a cluster fabric (CF) interface module 340a,b adapted to implement intra-cluster communication among the blades, including D-blade-to-D-blade communication for data container striping operations described herein.

The protocol layers, e.g., the NFS/CIFS layers and the iSCSI/FC layers, of the N-blade 310 function as protocol servers that translate file-based and block based data access requests from clients into CF protocol messages used for communication with the D-
30

blade 350. That is, the N-blade servers convert the incoming data access requests into file system primitive operations (commands) that are embedded within CF messages by the CF interface module 340 for transmission to the D-blades 350 of the cluster 100. Notably, the CF interface modules 340 cooperate to provide a single file system image
5 across all D-blades 350 in the cluster 100. Thus, any network port of an N-blade that receives a client request can access any data container within the single file system image located on any D-blade 350 of the cluster.

Further to the illustrative embodiment, the N-blade 310 and D-blade 350 are implemented as separately-scheduled processes of storage operating system 300; however,
10 in an alternate embodiment, the blades may be implemented as pieces of code within a single operating system process. Communication between an N-blade and D-blade is thus illustratively effected through the use of message passing between the blades although, in the case of remote communication between an N-blade and D-blade of different nodes, such message passing occurs over the cluster switching fabric 150. A known
15 message-passing mechanism provided by the storage operating system to transfer information between blades (processes) is the Inter Process Communication (IPC) mechanism. The protocol used with the IPC mechanism is illustratively a generic file and/or block-based "agnostic" CF protocol that comprises a collection of methods/functions constituting a CF application programming interface (API). Examples of such an agnostic protocol are the SpinFS and SpinNP protocols available from Network Appliance, Inc. The
20 SpinFS protocol is described in the above-referenced U.S. Patent Application Publication No. US 2002/0116593.

The CF interface module 340 implements the CF protocol for communicating file system commands among the blades of cluster 100. Communication is illustratively ef-
25 fected by the D-blade exposing the CF API to which an N-blade (or another D-blade) issues calls. To that end, the CF interface module 340 is organized as a CF encoder and CF decoder. The CF encoder of, e.g., CF interface 340a on N-blade 310 encapsulates a CF message as (i) a local procedure call (LPC) when communicating a file system command to a D-blade 350 residing on the same node 200 or (ii) a remote procedure call (RPC)
30 when communicating the command to a D-blade residing on a remote node of the cluster

100. In either case, the CF decoder of CF interface 340b on D-blade 350 de-encapsulates the CF message and processes the file system command.

Fig. 4 is a schematic block diagram illustrating the format of a CF message 400 in accordance with an embodiment of with the present invention. The CF message 400 is illustratively used for RPC communication over the switching fabric 150 between remote blades of the cluster 100; however, it should be understood that the term "CF message" may be used generally to refer to LPC and RPC communication between blades of the cluster. The CF message 400 includes a media access layer 402, an IP layer 404, a UDP layer 406, a reliable connection (RC) layer 408 and a CF protocol layer 410. As noted, the CF protocol is a generic file system protocol that conveys file system commands related to operations contained within client requests to access data containers stored on the cluster 100; the CF protocol layer 410 is that portion of message 400 that carries the file system commands. Illustratively, the CF protocol is datagram based and, as such, involves transmission of messages or "envelopes" in a reliable manner from a source (e.g., an N-blade 310) to a destination (e.g., a D-blade 350). The RC layer 408 implements a reliable transport protocol that is adapted to process such envelopes in accordance with a connectionless protocol, such as UDP 406.

A data container, e.g., a file, is accessed in the file system using a data container handle. Fig. 5 is a schematic block diagram illustrating the format of a data container handle 500 including a SVS ID field 502, an inode number field 504, a unique-ifier field 506, a striped flag field 508 and a striping epoch number field 510. The SVS ID field 502 contains a global identifier (within the cluster 100) of the SVS within which the data container resides. The inode number field 504 contains an inode number of an inode (within an inode file) pertaining to the data container. The unique-ifier field 506 contains a monotonically increasing number that uniquely identifies the data container handle 500. The unique-ifier is particularly useful in the case where an inode number has been deleted, reused and reassigned to a new data container. The unique-ifier distinguishes that reused inode number in a particular data container from a potentially previous use of those fields. The striped flag field 508 is illustratively a Boolean value that identifies whether the data container is striped or not. The striping epoch number field 510 indi-

ates the appropriate striping technique for use with this data container for embodiments where the SVS utilizes differing striping techniques for different data containers.

E. File System Organization

In the illustrative embodiment, a data container is represented in the write-
5 anywhere file system as an inode data structure adapted for storage on the disks 130. Fig. 6 is a schematic block diagram of an inode 600, which preferably includes a meta-data section 605 and a data section 660. The information stored in the meta-data section 605 of each inode 600 describes the data container (e.g., a file) and, as such, includes the type (e.g., regular, directory, vdisk) 610 of file, its size 615, time stamps (e.g., access and/or
10 modification time) 620 and ownership, i.e., user identifier (UID 625) and group ID (GID 630), of the file. The meta-data section 605 also includes a generation number 631, and a meta-data invalidation flag field 650. As described further herein, meta-data invalidation flag field 634 is used to indicate whether meta-data in the inode is usable or whether it should be re-acquired from the MDV. The contents of the data section 660 of each inode
15 may be interpreted differently depending upon the type of file (inode) defined within the type field 610. For example, the data section 660 of a directory inode contains meta-data controlled by the file system, whereas the data section of a regular inode contains file system data. In this latter case, the data section 660 includes a representation of the data associated with the file.

20 Specifically, the data section 660 of a regular on-disk inode may include file system data or pointers, the latter referencing 4kB data blocks on disk used to store the file system data. Each pointer is preferably a logical vbn to facilitate efficiency among the file system and the RAID system 380 when accessing the data on disks. Given the restricted size (e.g., 128 bytes) of the inode, file system data having a size that is less than
25 or equal to 64 bytes is represented, in its entirety, within the data section of that inode. However, if the length of the contents of the data container exceeds 64 bytes but less than or equal to 64kB, then the data section of the inode (e.g., a first level inode) comprises up to 16 pointers, each of which references a 4kB block of data on the disk.

Moreover, if the size of the data is greater than 64kB but less than or equal to 64
30 megabytes (MB), then each pointer in the data section 660 of the inode (e.g., a second

level inode) references an indirect block (e.g., a first level L1 block) that contains 1024 pointers, each of which references a 4kB data block on disk. For file system data having a size greater than 64MB, each pointer in the data section 660 of the inode (e.g., a third level L3 inode) references a double-indirect block (e.g., a second level L2 block) that
5 contains 1024 pointers, each referencing an indirect (e.g., a first level L1) block. The indirect block, in turn, contains 1024 pointers, each of which references a 4kB data block on disk. When accessing a file, each block of the file may be loaded from disk 130 into the memory 224.

When an on-disk inode (or block) is loaded from disk 130 into memory 224, its
10 corresponding in-core structure embeds the on-disk structure. For example, the dotted line surrounding the inode 600 indicates the in-core representation of the on-disk inode structure. The in-core structure is a block of memory that stores the on-disk structure plus additional information needed to manage data in the memory (but not on disk). The additional information may include, e.g., a “dirty” bit 670. After data in the inode (or
15 block) is updated/modified as instructed by, e.g., a write operation, the modified data is marked “dirty” using the dirty bit 670 so that the inode (block) can be subsequently “flushed” (stored) to disk. The in-core and on-disk format structures of the WAFL file system, including the inodes and inode file, are disclosed and described in the previously incorporated U.S. Patent No. 5,819,292 titled METHOD FOR MAINTAINING
20 CONSISTENT STATES OF A FILE SYSTEM AND FOR CREATING USER-ACCESSIBLE READ-ONLY COPIES OF A FILE SYSTEM by David Hitz et al., issued on October 6, 1998.

Fig. 7 is a schematic block diagram of an embodiment of a buffer tree of a file that may be advantageously used with the present invention. The buffer tree is an internal
25 representation of blocks for a file (e.g., file 700) loaded into the memory 224 and maintained by the write-anywhere file system 360. A root (top-level) inode 702, such as an embedded inode, references indirect (e.g., level 1) blocks 704. Note that there may be additional levels of indirect blocks (e.g., level 2, level 3) depending upon the size of the file. The indirect blocks (and inode) contain pointers 705 that ultimately reference data
30 blocks 706 used to store the actual data of the file. That is, the data of file 700 are contained in data blocks and the locations of these blocks are stored in the indirect blocks of

the file. Each level 1 indirect block 704 may contain pointers to as many as 1024 data blocks. According to the “write anywhere” nature of the file system, these blocks may be located anywhere on the disks 130.

A file system layout is provided that apportions an underlying physical volume
5 into one or more virtual volumes (or flexible volume) of a storage system, such as node 200. An example of such a file system layout is described in U.S. Patent Application Serial No. 10/836,817 titled EXTENSION OF WRITE ANYWHERE FILE SYSTEM LAYOUT, by John K. Edwards et al. and assigned to Network Appliance, Inc. The underlying physical volume is an aggregate comprising one or more groups of disks, such
10 as RAID groups, of the node. The aggregate has its own physical volume block number (pvbn) space and maintains meta-data, such as block allocation structures, within that pvbn space. Each flexible volume has its own virtual volume block number (vvbn) space and maintains meta-data, such as block allocation structures, within that vvbn space. Each flexible volume is a file system that is associated with a container file; the container
15 file is a file in the aggregate that contains all blocks used by the flexible volume. Moreover, each flexible volume comprises data blocks and indirect blocks that contain block pointers that point at either other indirect blocks or data blocks.

In one embodiment, pvbns are used as block pointers within buffer trees of files (such as file 700) stored in a flexible volume. This “hybrid” flexible volume embodiment
20 involves the insertion of only the pvbn in the parent indirect block (e.g., inode or indirect block). On a read path of a logical volume, a “logical” volume (vol) info block has one or more pointers that reference one or more fsinfo blocks, each of which, in turn, points to an inode file and its corresponding inode buffer tree. The read path on a flexible volume is generally the same, following pvbns (instead of vvbn) to find appropriate locations of blocks; in this context, the read path (and corresponding read performance) of a
25 flexible volume is substantially similar to that of a physical volume. Translation from pvbn-to-disk,dbn occurs at the file system/RAID system boundary of the storage operating system 300.

In an illustrative dual vbn hybrid flexible volume embodiment, both a pvbn and
30 its corresponding vvbn are inserted in the parent indirect blocks in the buffer tree of a

file. That is, the pvbn and vvbn are stored as a pair for each block pointer in most buffer tree structures that have pointers to other blocks, e.g., level 1(L1) indirect blocks, inode file level 0 (L0) blocks. Fig. 8 is a schematic block diagram of an illustrative embodiment of a buffer tree of a file 800 that may be advantageously used with the present invention. A root (top-level) inode 802, such as an embedded inode, references indirect (e.g., level 1) blocks 804. Note that there may be additional levels of indirect blocks (e.g., level 2, level 3) depending upon the size of the file. The indirect blocks (and inode) contain pvbn/vvbn pointer pair structures 808 that ultimately reference data blocks 806 used to store the actual data of the file.

The pvbns reference locations on disks of the aggregate, whereas the vvbn reference locations within files of the flexible volume. The use of pvbns as block pointers 808 in the indirect blocks 804 provides efficiencies in the read paths, while the use of vvbn block pointers provides efficient access to required meta-data. That is, when freeing a block of a file, the parent indirect block in the file contains readily available vvbn block pointers, which avoids the latency associated with accessing an owner map to perform pvbn-to-vvbn translations; yet, on the read path, the pvbn is available.

Fig. 9 is a schematic block diagram of an embodiment of an aggregate 900 that may be advantageously used with the present invention. Luns (blocks) 902, directories 904, qtrees 906 and files 908 may be contained within flexible volumes 910, such as dual vbn flexible volumes, that, in turn, are contained within the aggregate 900. The aggregate 900 is illustratively layered on top of the RAID system, which is represented by at least one RAID plex 950 (depending upon whether the storage configuration is mirrored), wherein each plex 950 comprises at least one RAID group 960. Each RAID group further comprises a plurality of disks 930, e.g., one or more data (D) disks and at least one (P) parity disk.

Whereas the aggregate 900 is analogous to a physical volume of a conventional storage system, a flexible volume is analogous to a file within that physical volume. That is, the aggregate 900 may include one or more files, wherein each file contains a flexible volume 910 and wherein the sum of the storage space consumed by the flexible volumes is physically smaller than (or equal to) the size of the overall physical volume. The ag-

gregate utilizes a *physical* pvbn space that defines a storage space of blocks provided by the disks of the physical volume, while each embedded flexible volume (within a file) utilizes a *logical* vvbn space to organize those blocks, e.g., as files. Each vvbn space is an independent set of numbers that corresponds to locations within the file, which loca-
 5 tions are then translated to dbns on disks. Since the flexible volume 910 is also a logical volume, it has its own block allocation structures (e.g., active, space and summary maps) in its vvbn space.

A container file is a file in the aggregate that contains all blocks used by a flexible volume. The container file is an internal (to the aggregate) feature that supports a flexible
 10 volume; illustratively, there is one container file per flexible volume. Similar to a pure logical volume in a file approach, the container file is a hidden file (not accessible to a user) in the aggregate that holds every block in use by the flexible volume. The aggregate includes an illustrative hidden meta-data root directory that contains subdirectories of flexible volumes:

15 *WAFL/fsid/filesystem file, storage label file*

Specifically, a physical file system (WAFL) directory includes a subdirectory for each flexible volume in the aggregate, with the name of subdirectory being a file system identifier (fsid) of the flexible volume. Each fsid subdirectory (flexible volume) contains at least two files, a filesystem file and a storage label file. The storage label file is illus-
 20 tratively a 4kB file that contains meta-data similar to that stored in a conventional raid label. In other words, the storage label file is the analog of a raid label and, as such, contains information about the state of the flexible volume such as, e.g., the name of the flexible volume, a universal unique identifier (uuid) and fsid of the flexible volume, whether it is online, being created or being destroyed, etc.

25 Fig. 10 is a schematic block diagram of an on-disk representation of an aggregate 1000. The storage operating system 300, e.g., the RAID system 380, assembles a physical volume of pvbns to create the aggregate 1000, with pvbns 1 and 2 comprising a “physical” volinfo block 1002 for the aggregate. The volinfo block 1002 contains block pointers to fsinfo blocks 1004, each of which may represent a snapshot of the aggregate.
 30 Each fsinfo block 1004 includes a block pointer to an inode file 1006 that contains inodes

of a plurality of files, including an owner map 1010, an active map 1012, a summary map 1014 and a space map 1016, as well as other special meta-data files. The inode file 1006 further includes a root directory 1020 and a “hidden” meta-data root directory 1030, the latter of which includes a namespace having files related to a flexible volume in which
5 users cannot “see” the files. The hidden meta-data root directory includes the *WAFL/fsid/* directory structure that contains filesystem file 1040 and storage label file 1090. Note that root directory 1020 in the aggregate is empty; all files related to the aggregate are organized within the hidden meta-data root directory 1030.

In addition to being embodied as a container file having level 1 blocks organized
10 as a container map, the filesystem file 1040 includes block pointers that reference various file systems embodied as flexible volumes 1050. The aggregate 1000 maintains these flexible volumes 1050 at special reserved inode numbers. Each flexible volume 1050 also has special reserved inode numbers within its flexible volume space that are used for, among other things, the block allocation bitmap structures. As noted, the block allo-
15 cation bitmap structures, e.g., active map 1062, summary map 1064 and space map 1066, are located in each flexible volume.

Specifically, each flexible volume 1050 has the same inode file structure/content as the aggregate, with the exception that there is no owner map and no
WAFL/fsid/filesystem file, storage label file directory structure in a hidden meta-data root
20 directory 1080. To that end, each flexible volume 1050 has a volinfo block 1052 that points to one or more fsinfo blocks 1054, each of which may represent a snapshot, along with the active file system of the flexible volume. Each fsinfo block, in turn, points to an inode file 1060 that, as noted, has the same inode structure/content as the aggregate with the exceptions noted above. Each flexible volume 1050 has its own inode file 1060 and
25 distinct inode space with corresponding inode numbers, as well as its own root (fsid) directory 1070 and subdirectories of files that can be exported separately from other flexible volumes.

The storage label file 1090 contained within the hidden meta-data root directory 1030 of the aggregate is a small file that functions as an analog to a conventional raid la-
30 bel. A raid label includes physical information about the storage system, such as the vol-

ume name; that information is loaded into the storage label file 1090. Illustratively, the storage label file 1090 includes the name 1092 of the associated flexible volume 1050, the online/offline status 1094 of the flexible volume, and other identity and state information 1096 of the associated flexible volume (whether it is in the process of being created
5 or destroyed).

F. VLDB

Fig. 11 is a schematic block diagram illustrating a collection of management processes that execute as user mode applications 1100 on the storage operating system 300 to provide management of configuration information (i.e. management data) for the nodes
10 of the cluster. To that end, the management processes include a management framework process 1110 and a volume location database (VLDB) process 1130, each utilizing a data replication service (RDB 1150) linked as a library. The management framework 1110 provides a user to an administrator 1170 interface via a command line interface (CLI) and/or a web-based graphical user interface (GUI). The management framework is illus-
15 tratively based on a conventional common interface model (CIM) object manager that provides the entity to which users/system administrators interact with a node 200 in order to manage the cluster 100.

The VLDB 1130 is a database process that tracks the locations of various storage components (e.g., SVSs, flexible volumes, aggregates, etc.) within the cluster 100 to
20 thereby facilitate routing of requests throughout the cluster. In the illustrative embodiment, the N-blade 310 of each node accesses a configuration table 235 that maps the SVS ID 502 of a data container handle 500 to a D-blade 350 that "owns" (services) the data container within the cluster. The VLDB includes a plurality of entries which, in turn, provide the contents of entries in the configuration table 235; among other things, these
25 VLDB entries keep track of the locations of the flexible volumes (hereinafter generally "volumes 910") and aggregates 900 within the cluster. Examples of such VLDB entries include a VLDB volume entry 1200 and a VLDB aggregate entry 1300.

Fig. 12 is a schematic block diagram of an exemplary VLDB volume entry 1200. The entry 1200 includes a volume ID field 1205, an aggregate ID field 1210 and, in alter-
30 nate embodiments, additional fields 1215. The volume ID field 1205 contains an ID that

identifies a volume 910 used in a volume location process. The aggregate ID field 1210 identifies the aggregate 900 containing the volume identified by the volume ID field 1205. Likewise, Fig. 13 is a schematic block diagram of an exemplary VLDB aggregate entry 1300. The entry 1300 includes an aggregate ID field 1305, a D-blade ID field 1310 and, in alternate embodiments, additional fields 1315. The aggregate ID field 1305 contains an ID of a particular aggregate 900 in the cluster 100. The D-blade ID field 1310 contains an ID of the D-blade hosting the particular aggregate identified by the aggregate ID field 1305.

The VLDB illustratively implements a RPC interface, e.g., a Sun RPC interface, which allows the N-blade 310 to query the VLDB 1130. When encountering contents of a data container handle 500 that are not stored in its configuration table, the N-blade sends an RPC to the VLDB process. In response, the VLDB 1130 returns to the N-blade the appropriate mapping information, including an ID of the D-blade that owns the data container. The N-blade caches the information in its configuration table 235 and uses the D-blade ID to forward the incoming request to the appropriate data container. All functions and interactions between the N-blade 310 and D-blade 350 are coordinated on a cluster-wide basis through the collection of management processes and the RDB library user mode applications 1100.

To that end, the management processes have interfaces to (are closely coupled to) RDB 1150. The RDB comprises a library that provides a persistent object store (storing of objects) for the management data processed by the management processes. Notably, the RDB 1150 replicates and synchronizes the management data object store access across all nodes 200 of the cluster 100 to thereby ensure that the RDB database image is identical on all of the nodes 200. At system startup, each node 200 records the status/state of its interfaces and IP addresses (those IP addresses it "owns") into the RDB database.

G. Storage System Architecture

The present invention is related to a storage system architecture illustratively comprising two or more volumes 910 distributed across a plurality of nodes 200 of cluster 100. The volumes are organized as a SVS and configured to store content of data con-

ainers, such as files and luns, served by the cluster in response to multi-protocol data access requests issued by clients 180. Notably, the content of each data container is apportioned among the volumes of the SVS to thereby improve the efficiency of storage service provided by the cluster. To facilitate a description and understanding of the present invention, data containers are hereinafter referred to generally as "files".

The SVS comprises a meta-data volume (MDV) and one or more data volumes (DV). The MDV is configured to store a canonical, or definitive, copy of certain meta-data, including access control lists (ACLs) and directories, associated with all files stored on the SVS, whereas each DV is configured to store, at least, data content of those files.

For each file stored on the SVS, one volume is designated the container attribute volume (CAV) and, to that end, is configured to store ("cache") certain, rapidly-changing attribute meta-data, including time stamps and file length, associated with that file to thereby offload access requests that would otherwise be directed to the MDV. In the illustrative embodiment described herein, determination of the CAV for a file is based on a simple rule: designate the volume holding the first stripe of content (data) for the file as the CAV for the file. Not only is this simple rule convenient, but it also provides an optimization for small files. That is, a CAV may be able to perform certain operations without having to communicate with other volumes of the SVS if the file is small enough to fit within the specified stripe width. Ideally, the first stripes of data for files are distributed among the DVs of the SVS to thereby facilitate even distribution of CAV designations among the volumes of the SVS. In alternate embodiments, data for files is striped across the MDV and the DVs.

Fig. 14 is a schematic block diagram of the inode files of an SVS 1400 in accordance with an embodiment of the present invention. The SVS 1400 illustratively comprises three volumes, namely MDV 1405 and two DVs 1410, 1415. It should be noted that in alternate embodiments additional and/or differing numbers of volumes may be utilized in accordance with the present invention. Illustratively, the MDV 1405 stores a plurality of inodes, including a root directory (RD) inode 1420, a directory (DIR) inode 1430, file (F) inodes 1425, 1435, 1445 and an ACL inode 1440. Each of these inodes illustratively includes meta-data (M) associated with the inode. In the illustrative em-

bodiment, each inode on the MDV 1405 does not include data (D); however, in alternate embodiments, the MDV may include user data.

In contrast, each DV 1410, 1415 stores only file (F) inodes 1425, 1435, 1445 and ACL inode 1440. According to the inventive architecture, a DV does not store directories or other device inodes/constructs, such as symbolic links; however, each DV does store F inodes, and may store cached copies of ACL inodes, that are arranged in the same locations as their respective inodes in the MDV 1405. A particular DV may not store a copy of an inode until an I/O request for the data container associated with the inode is received by the D-Blade serving a particular DV. Moreover, the contents of the files denoted by these F inodes are periodically sparse according to SVS striping rules, as described further herein. In addition, since one volume is designated the CAV for each file stored on the SVS 1400, DV 1415 is designated the CAV for the file represented by inode 1425 and DV 1410 is the CAV for the files identified by inodes 1435, 1445. Accordingly, these CAVs cache certain, rapidly-changing attribute meta-data (M) associated with those files such as, e.g., file size 615, as well as access and/or modification time stamps 620.

The SVS is associated with a set of striping rules that define a stripe algorithm, a stripe width and an ordered list of volumes within the SVS. The striping rules for each SVS are illustratively stored as an entry of VLDB 1130 and accessed by SVS ID. Fig. 15 is a schematic block diagram of an exemplary VLDB SVS entry 1500 in accordance with an embodiment of the present invention. The VLDB entry 1500 includes a SVS ID field 1505 and one or more sets of striping rules 1530. In alternate embodiments additional fields 1535 may be included. The SVS ID field 1505 contains the ID of a SVS which, in operation, is specified in data container handle 500.

Each set of striping rules 1530 illustratively includes a stripe width field 1510, a stripe algorithm ID field 1515, an ordered list of volumes field 1520 and, in alternate embodiments, additional fields 1525. The striping rules 1530 contain information for identifying the organization of a SVS. For example, the stripe algorithm ID field 1515 identifies a striping algorithm used with the SVS. In the illustrative embodiment, multiple striping algorithms could be used with a SVS; accordingly, stripe algorithm ID is needed

to identify which particular algorithm is utilized. Each striping algorithm, in turn, specifies the manner in which file content is apportioned as stripes across the plurality of volumes of the SVS. The stripe width field 1510 specifies the size/width of each stripe. The ordered list of volumes field 1520 contains the IDs of the volumes comprising the SVS.

5 Moreover, the ordered list of volumes may specify the function and implementation of the various volumes and striping rules of the SVS. For example, the first volume in the ordered list may denote the MDV of the SVS, whereas the ordering of volumes in the list may denote the manner of implementing a particular striping algorithm, e.g., round-robin.

A Locate() function 375 is provided that enables the VSM 370 and other modules
10 (such as those of N-blade 310) to locate a D-blade 350 and its associated volume of a SVS 1400 in order to service an access request to a file. The Locate() function takes as arguments, at least (i) a SVS ID 1505, (ii) an offset within the file, (iii) the inode number for the file and (iv) a set of striping rules 1530, and returns the volume 910 on which that offset begins within the SVS 1400. For example, assume a data access request directed
15 to a file is issued by a client 180 and received at the N-blade 310 of a node 200, where it is parsed through the multi-protocol engine 325 to the appropriate protocol server of N-blade 310.

To determine the location of a D-blade 350 to which to transmit a CF message 400, the N-blade 310 may first retrieve a SVS entry 1500 to acquire the striping rules
20 1530 (and list of volumes 1520) associated with the SVS. The N-blade 310 then executes the Locate() function 375 to identify the appropriate volume to which to direct an operation. Thereafter, the N-Blade may retrieve the appropriate VLDB volume entry 1200 to identify the aggregate containing the volume and the appropriate VLDB aggregate entry 1300 to ultimately identify the appropriate D-blade 350. The protocol server of N-blade
25 310 then transmits the CF message 400 to the D-blade 350.

Fig. 16 is a schematic block diagram illustrating the periodic sparseness of file content stored on volumes A 1605, B 1610 and C 1615 of SVS 1600. As noted, file content is periodically sparse according to the SVS striping rules, which specify a striping algorithm (as indicated by stripe algorithm ID field 1515) and a size/width of each stripe
30 (as indicated by stripe width field 1510). Note that, in the illustrative embodiment, a

stripe width is selected to ensure that each stripe may accommodate the actual data (e.g., stored in data blocks 806) referenced by an indirect block (e.g., level 1 block 804) of a file.

In accordance with an illustrative round robin striping algorithm, volume A 1605
5 contains a stripe of file content or data (D) 1620 followed, in sequence, by two stripes of
sparseness (S) 1622, 1624, another stripe of data (D) 1626 and two stripes of sparseness
(S) 1628, 1630. Volume B 1610, on the other hand, contains a stripe of sparseness (S)
1632 followed, in sequence, by a stripe of data (D) 1634, two stripes of sparseness (S)
1636, 1638, another stripe of data (D) 1640 and a stripe of sparseness (S) 1642. Volume
10 C 1615 continues the round robin striping pattern and, to that end, contains two stripes of
sparseness (S) 1644, 1646 followed, in sequence, by a stripe of data (D) 1648, two stripes
of sparseness (S) 1650, 1652 and another stripe of data (D) 1654.

H. Meta-data Caching and Distribution

The present invention is directed to a system and method for caching and distrib-
15 uting meta-data for one or more data containers stored on a plurality of volumes config-
ured as a SVS. As noted, the SVS comprises one MDV configured to store a canonical
copy of certain meta-data, including access control lists and directories, associated with
all data containers stored on the SVS, and one or more DVs configured to store, at least,
data content of those containers. In addition, for each data container stored on the SVS,
20 one volume is designated the CAV and, as such, is configured to store ("cache") a ca-
nonical copy of certain, rapidly-changing attribute meta-data, including time stamps and
container length, associated with that container.

Fig. 17 is a schematic block diagram of an exemplary five volume SVS environ-
ment 1700 in accordance with an embodiment the present invention. The SVS 1700
25 comprises five volumes, namely volume A 1705, volume B 1710, volume C 1715, vol-
ume D 1720 and volume E 1725. It should be noted that five volumes are shown for il-
lustrative purposes only and that the teachings of the present invention may be utilized
with SVSs having any number of volumes. In the illustrative environment 1700, volume
A 1705 is designated the MDV, with the other four volumes functioning as DVs associ-
30 ated with the SVS.

Twelve data containers, e.g., files (files 1-12), are illustratively stored on the volumes of the SVS, wherein each volume serves as the CAV for any file whose first stripe is stored therein. Notably, the CAV is a role that the MDV or DV serves for a particular file to store (and serve) rapidly-changing attribute meta-data for the file. Thus, for example, volume B 1710 serves as the CAV for files 1, 5, 9. Similarly, volume C 1715 serves as the CAV for files 2, 6 and 10, volume D 1720 serves as the CAV for files 3, 7 and 11 and volume E serves as the CAV for files 4, 8 and 12. Volume A 1705, which serves as the MDV for the SVS does not, in the illustrative embodiment, serve as the CAV for any files. In alternate embodiments, the MDV may serve as the CAV for files. By distributing the role of the CAV among the SVS volumes, each volume serves as a CAV for an approximately equal number of files.

According to an aspect of the invention, the meta-data associated with the files stored on the SVS are organized into various categories (e.g., MD1-MD3) along functional boundaries and are resident on various volumes to optimize data access (e.g., read and write) paths through the SVS. These categories include (i) MD1 meta-data that changes on every data access (read/write) request served by the SVS, (ii) MD2 meta-data that may be retrieved (but not changed) on every request and (iii) MD3 meta-data that is unused for the read/write requests. Since it changes on every read/write request served by the DVs of the SVS, the MD1 meta-data is canonically resident on the CAV and generally cached on the DVs. Likewise, since it may be retrieved, but does not change, on every request served by the DVs, the MD2 meta-data is canonically resident on the MDV and generally cached on all DVs of the SVS, including the volume designated as CAV. Finally, since it is unused for a read/write request, the MD3 meta-data is canonically, and solely, resident on the MDV. Table 1 illustrates the various categories and residences of meta-data according to the present invention.

Type	When Utilized	Where Stored	Examples of Meta-Data
MD1	Changes on every Read/Write	Canonical copy on CAV Cached at each DV and MDV	File length Access Time
MD2	Read for Read/Write Operations	Canonical copy at MDV Cached at all DVs	Access Control Lists Lock State Ownership
MD3	Not Used for Read/Write Operations	Canonical copy on MDV	Link Count

Table 1

According to another aspect of the invention, the CAV cooperates with the MDV and DVs of the SVS to provide a multi-tier caching and distribution architecture that off-loads meta-data access requests that would otherwise be directed to the MDV. Fig. 18 is a schematic block diagram of an exemplary multi-tier meta-data caching and distribution hierarchical environment 1800 in accordance with an embodiment of the present invention. As noted, MDV 1805 stores a canonical copy of MD2 and MD3 meta-data for all of the files stored on the SVS. Here, the CAV is utilized as a first tier caching and distribution point for storing and distributing most meta-data, e.g., MD1 and MD2, for use by the DVs 1815.

As noted above, every volume within the SVS serves as a CAV for a portion of the files stored on the SVS. Illustratively, the volume storing the first stripe of the file is deemed to be the CAV. Thus, different volumes of the environment 1800 may serve as the CAV 1810 depending on the particular files. Moreover, the MDV may serve as the CAV for certain files. In such a case, the roles of the CAV and MDV are merged due to the fact that the volume contains a canonical copy of all (MD1-MD3) meta-data. For those files stored on a SVS volume having a merged CAV/MDV role, no first tier caching and distribution point is utilized in the hierarchical environment 1800.

In the illustrative embodiment, the multi-tier caching and distribution architecture facilitates propagating of meta-data throughout the nodes of the cluster for storage on various volumes ("caches") of the SVS and updating of those caches in an efficient manner that ensures cache consistency in the cluster. As for the latter, a "pull" model is illustratively employed for cache updates. That is, the meta-data stored on a particular data volume is deemed to be valid unless it has been explicitly invalidated by, for example, assertion of the meta-data invalidation flag 634. In response to assertion of the invalidation flag, the volume (e.g., CAV or DV) queries its "higher-level" volume of the hierarchy (e.g., MDV or CAV, respectively) to obtain (pull) a valid copy of the meta-data in order to update its cache.

In the case of the MD1 meta-data, the multi-tier caching and distribution architecture enables the DVs to constantly access the CAV for meta-data changes that occur on every read/write request. The MD1 access path of the caching and distribution hierarchy 1800 is relatively "short" and, notably, offloads all such frequent requests from the MDV. In the case of MD2 meta-data, the multi-tier caching and distribution architecture enables the DV to access the CAV for meta-data retrieved on every read/write request. In general, the MD2 access path of the hierarchy terminates at the CAV for current MD2 meta-data; however, since it does not store a canonical copy of that meta-data, the CAV may have to acquire a current version of the MD2 meta-data from the MDV if its copy is invalid.

It should be noted that meta-data cached locally at each volume may be deemed to be valid unless certain conditions have been met. As noted, one of these conditions is that the meta-data has not been explicitly invalidated, by, for example, assertion of the meta-data invalidation flag 634 of the inode stored on each data volume. Another condition is that the data volumes may have a set of rules to ensure cache consistency in addition to the requirements of checking for explicit invalidation. One such technique to ensure cache consistency is described in the United States Patent Application Serial Number 112056-0210, entitled SYSTEM AND METHOD FOR GUARANTEEING CACHE CONSISTENCY IN A CLUSTER, by Richard Jernigan et al, the contents of which are hereby incorporated by reference.

In the illustrative embodiment, all commands for a file stored on the SVS that modify and/or create an attribute of the file are directed to the MDV. In response, the MDV instructs the CAV to invalidate its cached copy of the meta-data by, e.g., asserting the meta-data invalidation flag in the inode of the file. The MDV then modifies its copy of the meta-data before responding to the command. Subsequently, when attempting to access its cached meta-data, the CAV detects that the meta-data is invalid and queries the MDV server for an updated copy of the meta-data.

Fig. 19 is a flow chart detailing the steps of a procedure 1900 for processing a modification or creation command directed to a file attribute in accordance with an embodiment of the present invention. The procedure 1900 begins in step 1905 and continues to step 1910 where the VSM serving the MDV receives a request to modify/create an attribute of a file. In the illustrative embodiment, all received requests for modifying/creating attributes are directed to the VSM of the MDV for processing and, to that end, each N-Blade is programmed to forward any such requests to the MDV. In response to receiving the request, the MDV sends a request to the CAV to invalidate its cached meta-data in step 1915. Upon receiving the request, the VSM of the CAV invalidates its meta-data by, for example, marking the meta-data invalidation flag 634 within the inode of the data volume as invalid. This causes the CAV to query the MDV to update its cached value before using the meta-data in any operation. Additionally, the VSM of the CAV transmits a request to all DVs to invalidate their cached meta-data. In response, each VSM of a DV invalidates its meta-data by, for example, asserting the meta-data invalidation flag 634. The CAV then acknowledges completion of the invalidation in step 1925. Upon receiving the acknowledgment, the MDV modifies the canonical copy of the meta-data per the received request in step 1930. The MDV then responds to the N-Blade in step 1935 before the procedure completes in 1940.

Fig. 20 is a flow chart detailing the steps of a procedure 2000 for processing a data access operation in accordance with the novel multi-tier meta-data caching and distribution architecture of the present invention. The procedure 2000 begins in step 2005 and continues to step 2010 where a VSM of a data volume receives a data access operation directed to a file from, e.g., an N-Blade during normal cluster operation. In step 2015, the VSM first determines whether its local cached copy of meta-data (e.g., stored

on the data volume) needed to process the operation is invalid. The VSM may render this determination by, for example, examining the meta-data invalidation flag 634 of the inode of the data volume to which the data access operation is directed. Note that, in alternate embodiments of the present invention, this determination may be rendered by determining whether a predetermined amount of time has elapsed since the data volume queried the CAV for the file. Such additional checks are described in the above-incorporated United States Patent Application entitled SYSTEM AND METHOD FOR GUARANTEEING CACHE CONSISTENCY IN A CLUSTER.

If the local cached copy of the meta-data is not invalid, the procedure branches to step 2045 where the VSM of the data volume processes the data access operation before completing in step 2050. However, if the cached meta-data is invalid on the data volume (step 2015), the procedure continues to step 2020 where the VSM on the data volume queries the VSM of the CAV for a current, updated copy of that meta-data. This VSM-to-VSM query may be accomplished by, for example, forwarding a CF command message from the D-Blade hosting the data volume to the D-Blade hosting the CAV. It should be noted that, since the role of the CAV is on a per file basis, the load of such queries is distributed among the various volumes and D-blades of the cluster servicing those volumes.

In step 2025, the VSM of the CAV determines whether its meta-data is invalid by, e.g., examining the meta-data invalidation flag 634 status in the inode of the appropriate file. Note that the meta-data may be invalid on the CAV for a number of reasons, one of which being that the meta-data, such as an access control list, has not yet been created. In addition, the CAV's meta-data may have been explicitly invalidated by the MDV while processing a modify/create command directed to an attribute. If the meta-data on the CAV is invalid, the procedure continues to step 2030 where the VSM of the CAV queries the VSM of the MDV for the current, updated version of the meta-data. Upon receiving and storing a current copy of the meta-data, the VSM of the CAV marks its cache as updated and valid by, e.g., de-asserting the meta-data invalidation flag 634.

In step 2035, the VSM of the CAV distributes the current copy of the meta-data to the VSM of the data volume which, in turn, updates its local cached copy of the meta-

data in step 2040. In Step 2045, the VSM of the data volume then utilizes the meta-data when processing the data access operation and the procedure completes in step 2050. As can be appreciated from procedure 2000, the use of the CAV as a first tier caching and distribution point reduces the processing load on the MDV. Moreover, since each volume serves as the CAV for a portion of the files (data containers) stored on the SVS, the processing load is distributed among those SVS volumes. Accordingly, the novel multi-tier caching and distribution architecture reduces the processing load on the MDV to thereby obviate a processing bottleneck in the SVS.

In sum, the present invention advantageously facilitates (i) propagating of meta-data throughout the nodes of the cluster for storage on various volumes ("caches") of the SVS and (ii) updating of those caches in an efficient manner. That is, the novel multi-tier caching and distribution architecture reduces the latency incurred when acquiring and updating such meta-data needed to service a data access request directed to a data container to thereby provide a high-performance cluster.

The foregoing description has been directed to particular embodiments of this invention. It will be apparent, however, that other variations and modifications may be made to the described embodiments, with the attainment of some or all of their advantages. Specifically, it should be noted that the principles of the present invention may be implemented in non-distributed file systems. Furthermore, while this description has been written in terms of N and D-blades, the teachings of the present invention are equally suitable to systems where the functionality of the N and D-blades are implemented in a single system. Alternately, the functions of the N and D-blades may be distributed among any number of separate systems, wherein each system performs one or more of the functions. Additionally, the procedures, processes and/or modules described herein may be implemented in hardware, software, embodied as a computer-readable medium having program instructions, firmware, or a combination thereof. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.

What is claimed is:

CLAIMS

- 1 1. A method for distributing meta-data relating to a data container in a striped vol-
2 ume set, the method comprising the steps of:
3 providing a meta-data volume storing a canonical copy of the meta-data relating
4 to the data container; and
5 designating one of one or more data volumes as a container attribute volume for
6 the data container, the container attribute volume storing a cached copy of the meta-data
7 relating to the data container and wherein the container attribute volume serves as a first
8 tier caching and distribution point within the striped volume set.
- 1 2. The method of claim 1 further comprising the steps of:
2 receiving a request to modify/create the meta-data by a volume striping module
3 serving the meta-data volume;
4 invalidating the cached copy of the meta-data stored on the container attribute
5 volume; and
6 processing the received request to modify/create the canonical copy of the meta-
7 data.
- 1 3. The method of claim 2 wherein the step of invalidating the cached copy of the
2 meta-data comprises the step of setting a meta-data invalidation flag to identify that the
3 cached copy of the meta-data is invalid.
- 1 4. The method of claim 2 wherein the step of invalidating the cached copy of the
2 meta-data comprises the step of transmitting a request to the one or more data volumes to
3 invalidate a cached copy of the meta-data on each of the one or more data volumes.
- 1 5. The method of claim 1 wherein the container attribute volume is selected as the
2 data volume storing a first stripe of data of the data container.
- 1 6. The method of claim 1 wherein the meta-data comprises an access control list for

2 the data container.

1 7. The method of claim 1 wherein the data container comprises a file.

1 8. The method of claim 1 wherein the data container comprises a logical unit num-
2 ber.

1 9. The method of claim 1 further comprising the steps of:
2 receiving, by a volume striping module serving one of the one or more data vol-
3 umes, a data access operation;
4 determining whether a copy of the meta-data stored on the data volume is valid;
5 in response to determining that the copy of the meta-data is not valid, requesting a
6 volume striping module serving the container attribute volume to send a copy of its
7 cached meta-data to the volume striping module serving the data volume;
8 determining, by the volume striping module serving the container attribute vol-
9 ume, whether its cached copy of the meta-data is valid; and
10 in response to determining that the container attribute volume's cached copy of
11 the meta-data is not valid, updating the cached copy of the meta-data by querying the
12 meta-data volume.

1 10. The method of claim 1 wherein the meta-data comprises meta-data that is not
2 modified on every data access request.

1 11. The method of claim 10 wherein the meta-data comprises an access control list.

1 12. A system for distributing meta-data relating to a data container in a striped set,
2 the system comprising:
3 a meta-data volume storing a canonical copy of the meta-data;
4 one or more data volumes organized to store stripes of data of the data container,
5 one of the one or more data volumes being designated as a container attribute volume for
6 the data container; and

7 wherein the container attribute volume stores a cached copy of the meta-data and
8 includes a meta-data invalidation flag identifying a status of the cached copy of the meta-
9 data.

1 13. The system of claim 12 wherein the container attribute volume is selected as the
2 data volume storing a first stripe of data of the data container.

1 14. The system of claim 12 wherein the meta-data comprises an access control list for
2 the data container.

1 15. The system of claim 12 wherein the data container comprises a file.

1 16. The system of claim 12 wherein the data container comprises a logical unit num-
2 ber.

1 17. The system of claim 12 wherein the meta-data comprises meta-data that is not
2 modified on every data access request.

1 18. The system of claim 17 wherein the meta-data comprises an access control list.
2

1 19. A system for distributing meta-data relating to a data container in a striped vol-
2 ume set, the system comprising:

3 a plurality of computers, each of the plurality of computers comprising a volume
4 striping module, each volume striping module adapted to service one or more of a plural-
5 ity of volumes logically organized into the volume striping set;

6 wherein one of the plurality of volumes is designated as a meta-data volume stor-
7 ing a canonical copy of meta-data relating to the data container; and

8 wherein one of the plurality of volumes is designated as a container attribute vol-
9 ume for the data container, the container attribute volume storing a cached copy of the
10 meta-data relating to the data container and having a meta-data invalidation flag identify-

11 ing a status of the cached copy of the meta-data relating to the data container.

1 20. The system of claim 19 wherein the container attribute volume is selected as the
2 data volume storing a first stripe of data of the data container.

1 21. The system of claim 19 wherein the meta-data comprises an access control list for
2 the data container.

1 22. The system of claim 19 wherein the data container comprises a file.

1 23. The system of claim 19 wherein the data container comprises a logical unit num-
2 ber.

1 24. The system of claim 19 wherein the meta-data comprises meta-data that is not
2 modified on every data access request.

1 25. The system of claim 19 wherein the meta-data comprises an access control list.
2
3

1 26. A computer readable medium for distributing meta-data relating to a data con-
2 tainer in a striped volume set having a meta-data volume storing a canonical copy of the
3 meta-data relating to the data container and wherein one of one or more data volumes is
4 designated as a container attribute volume for the data container and storing a cached
5 copy of the meta-data relating to the data container, the computer readable medium in-
6 cluding program instructions for performing the steps of:

7 receiving a request to modify/create the meta-data relating to the data container by
8 a volume striping module serving the meta-data volume;

9 invalidating the cached copy of the meta-data relating to the data container; and

10 processing the received request to modify/create the canonical copy of the meta-
11 data relating to the data container.

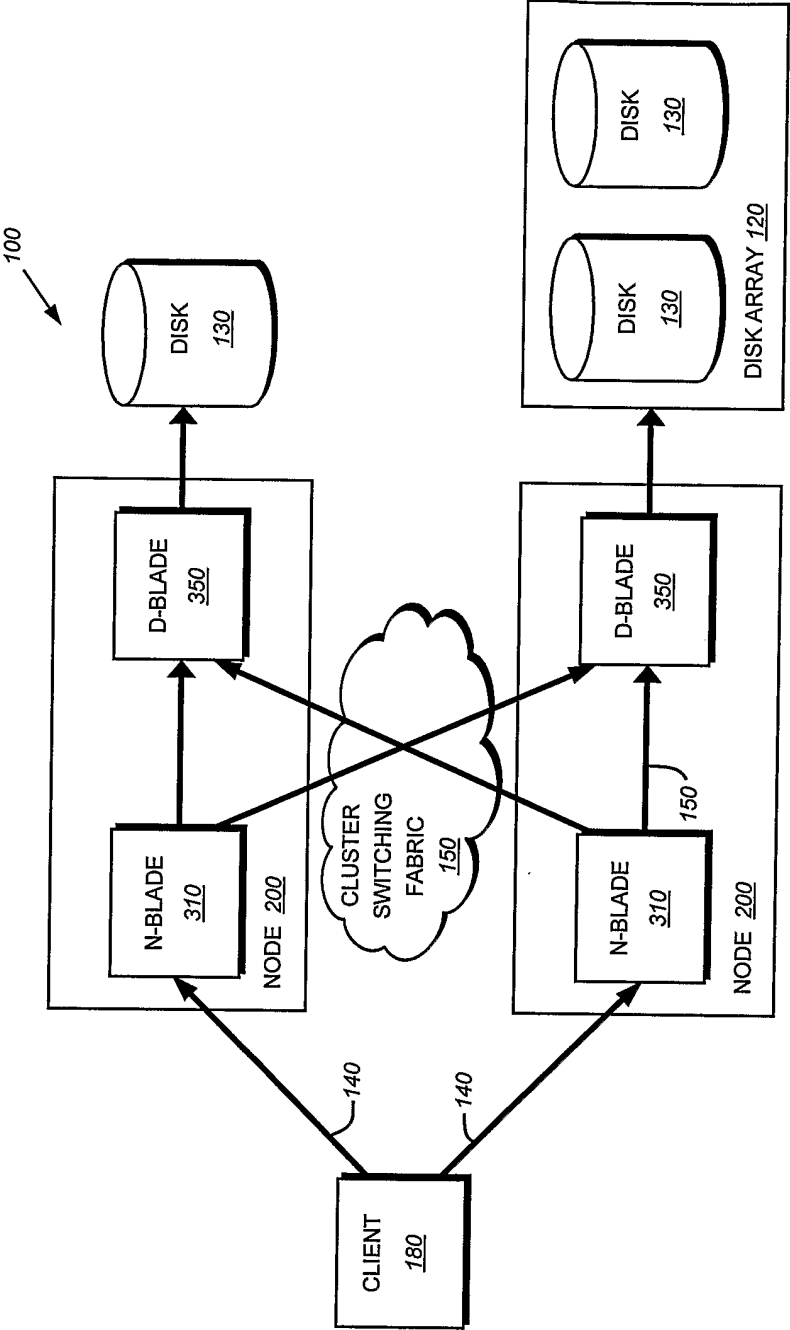


FIG. 1

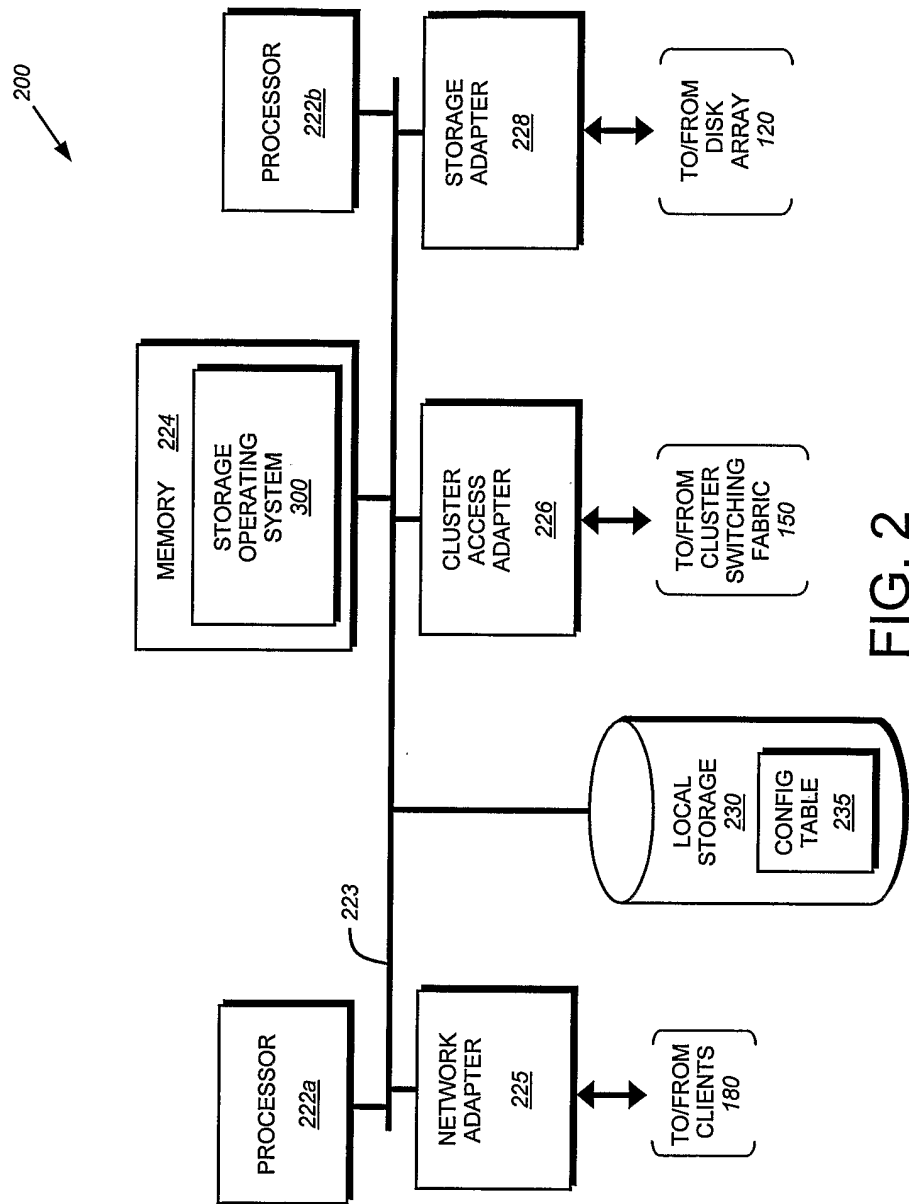


FIG. 2

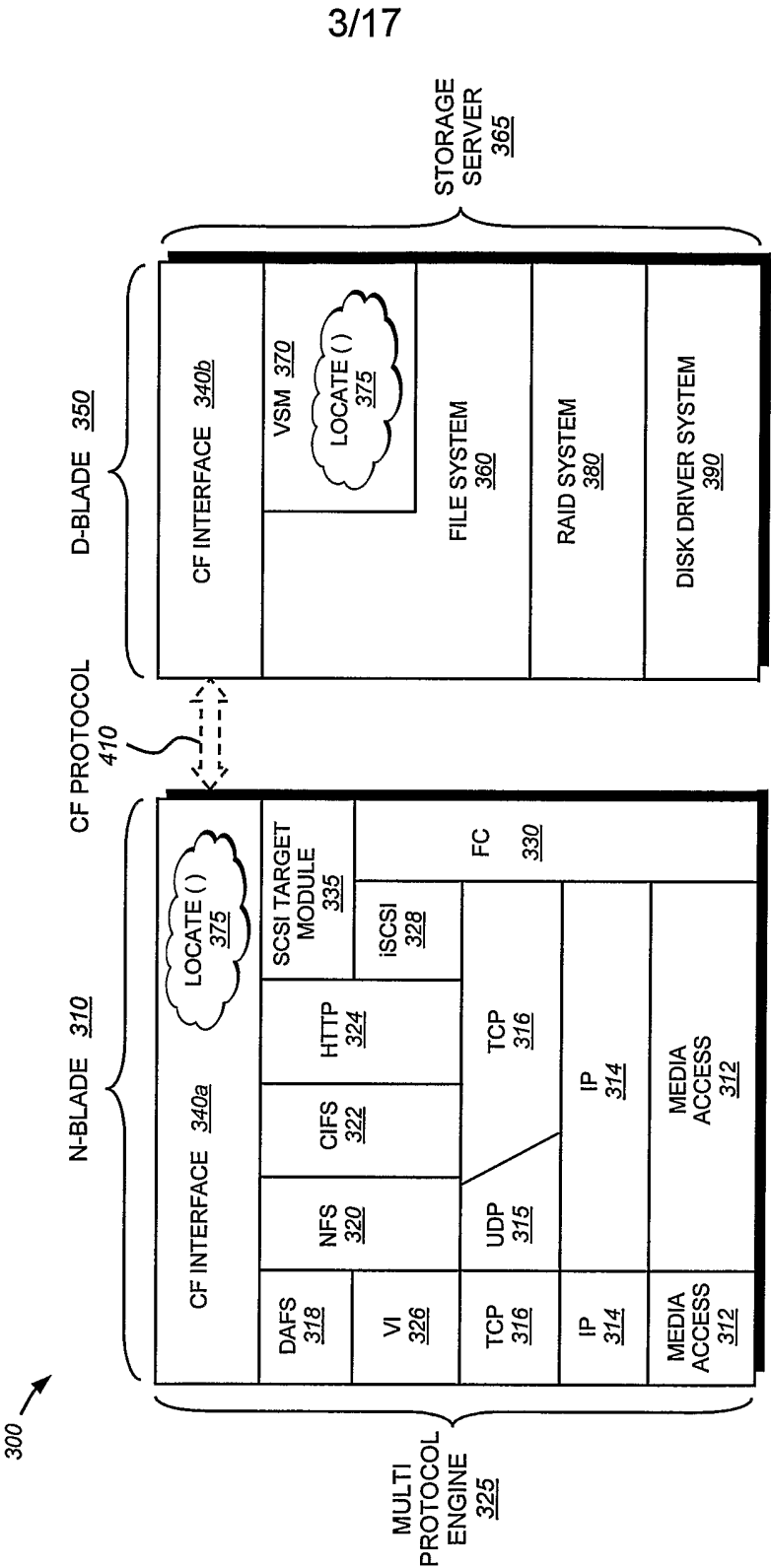


FIG. 3

4/17

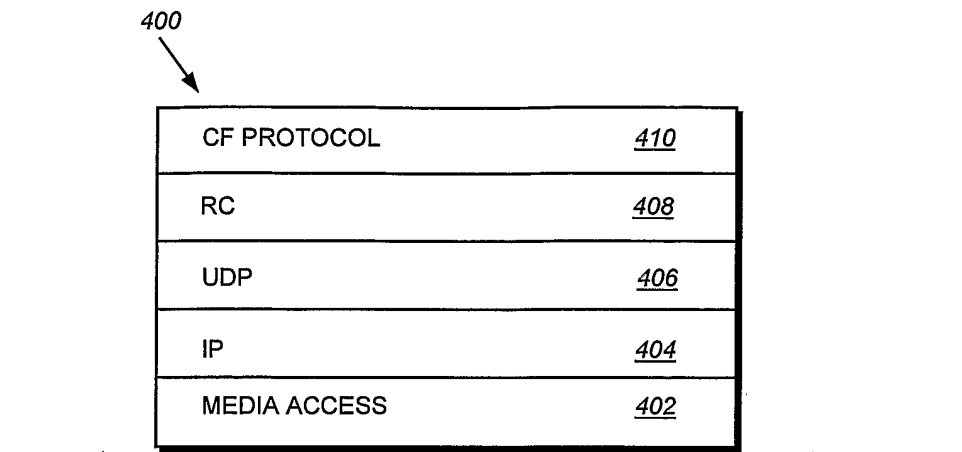


FIG. 4

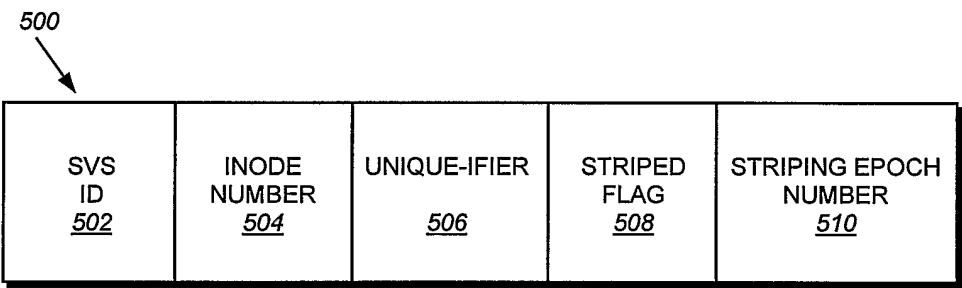


FIG. 5

5/17

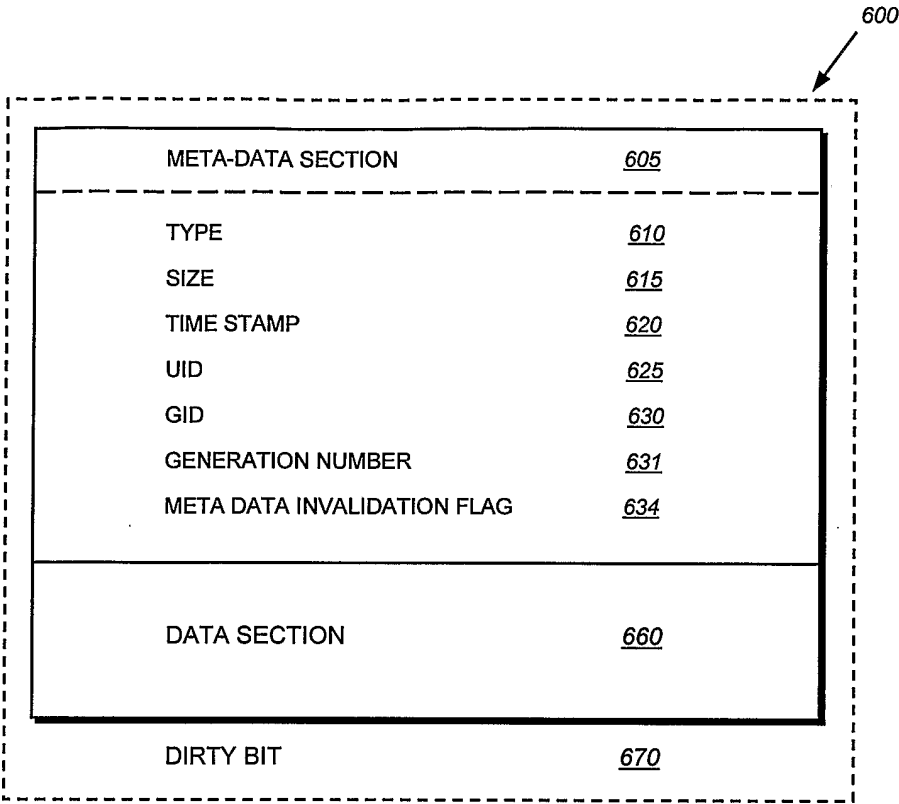


FIG. 6

6/17

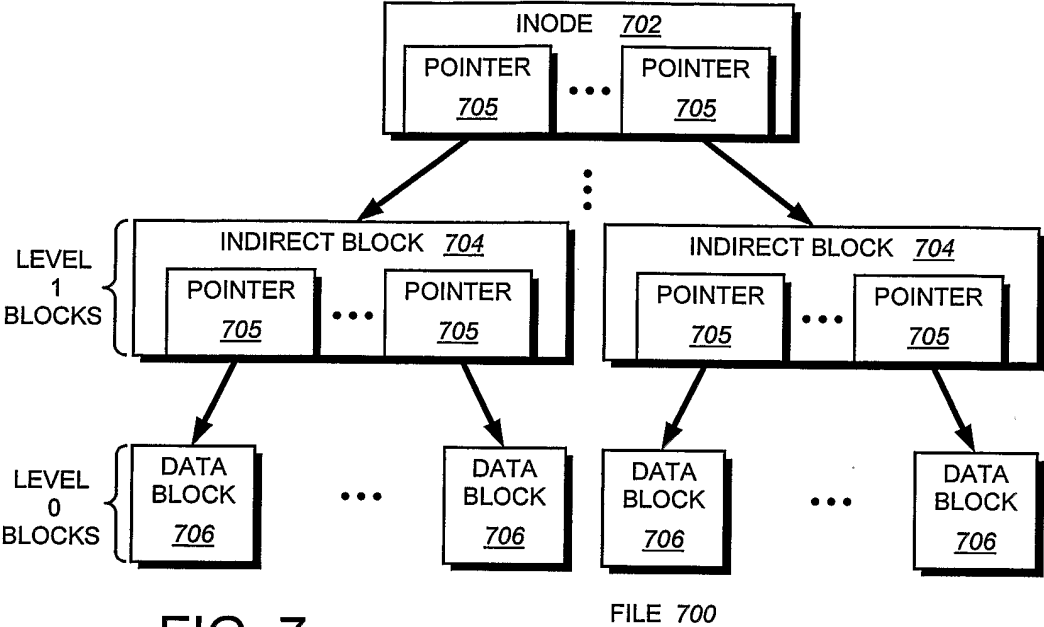


FIG. 7

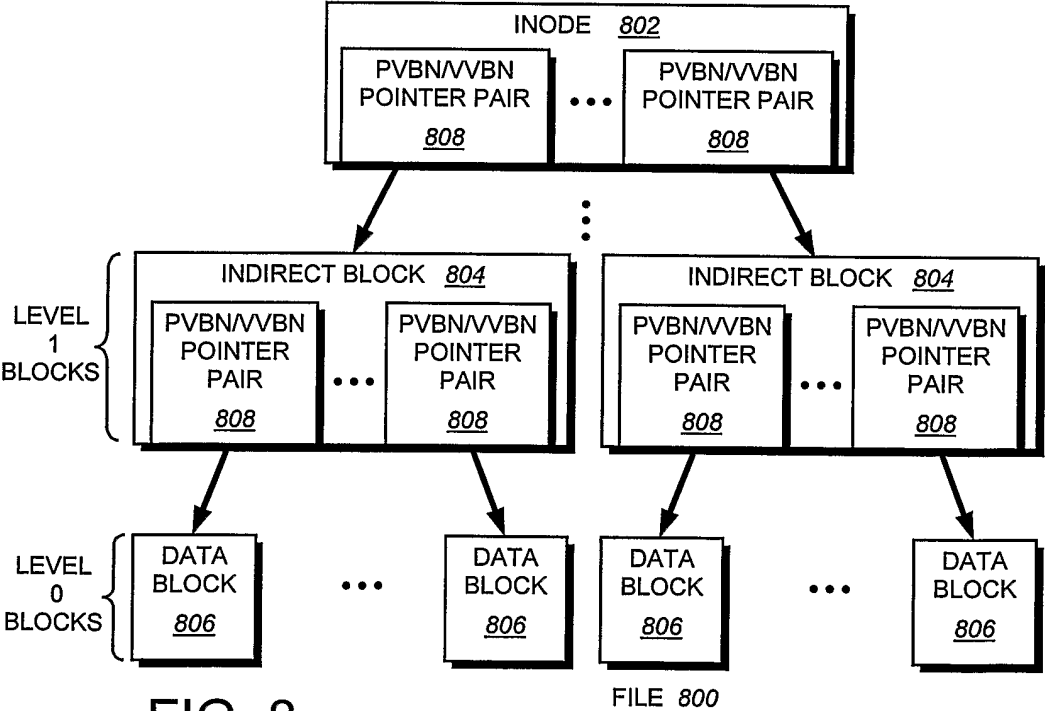


FIG. 8

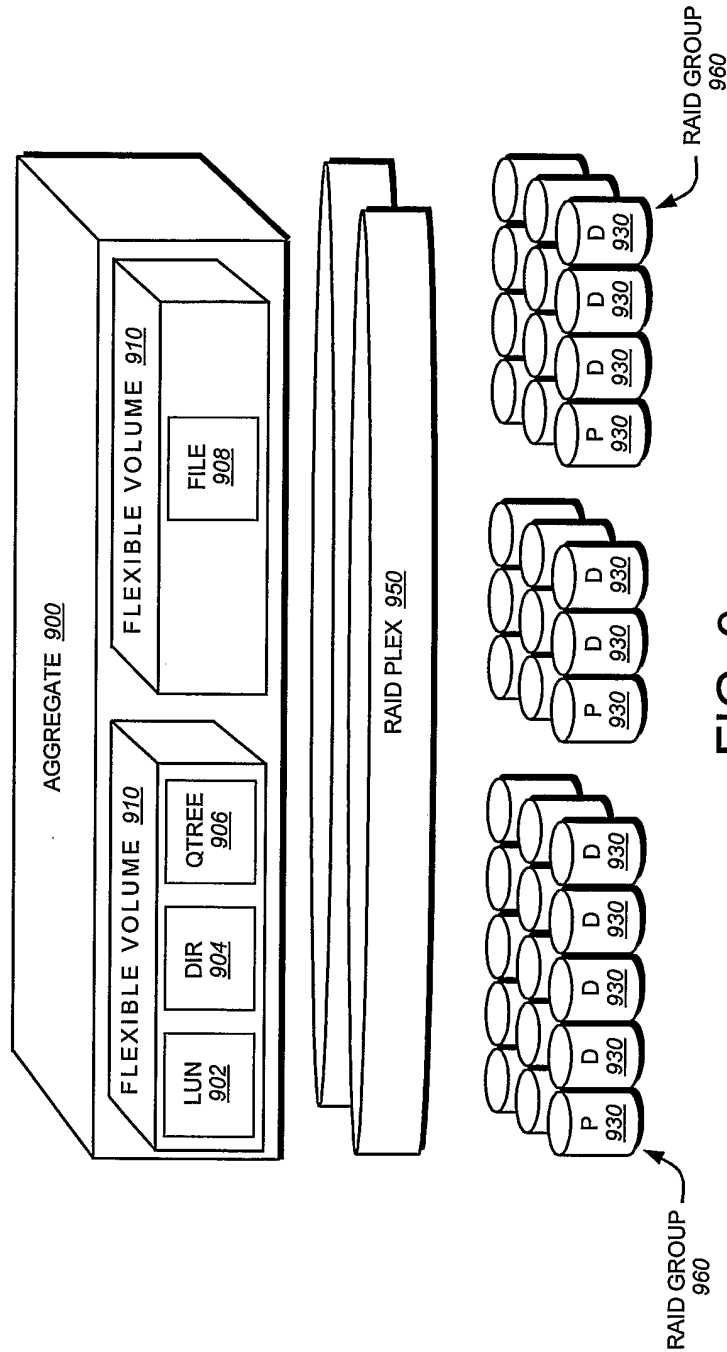
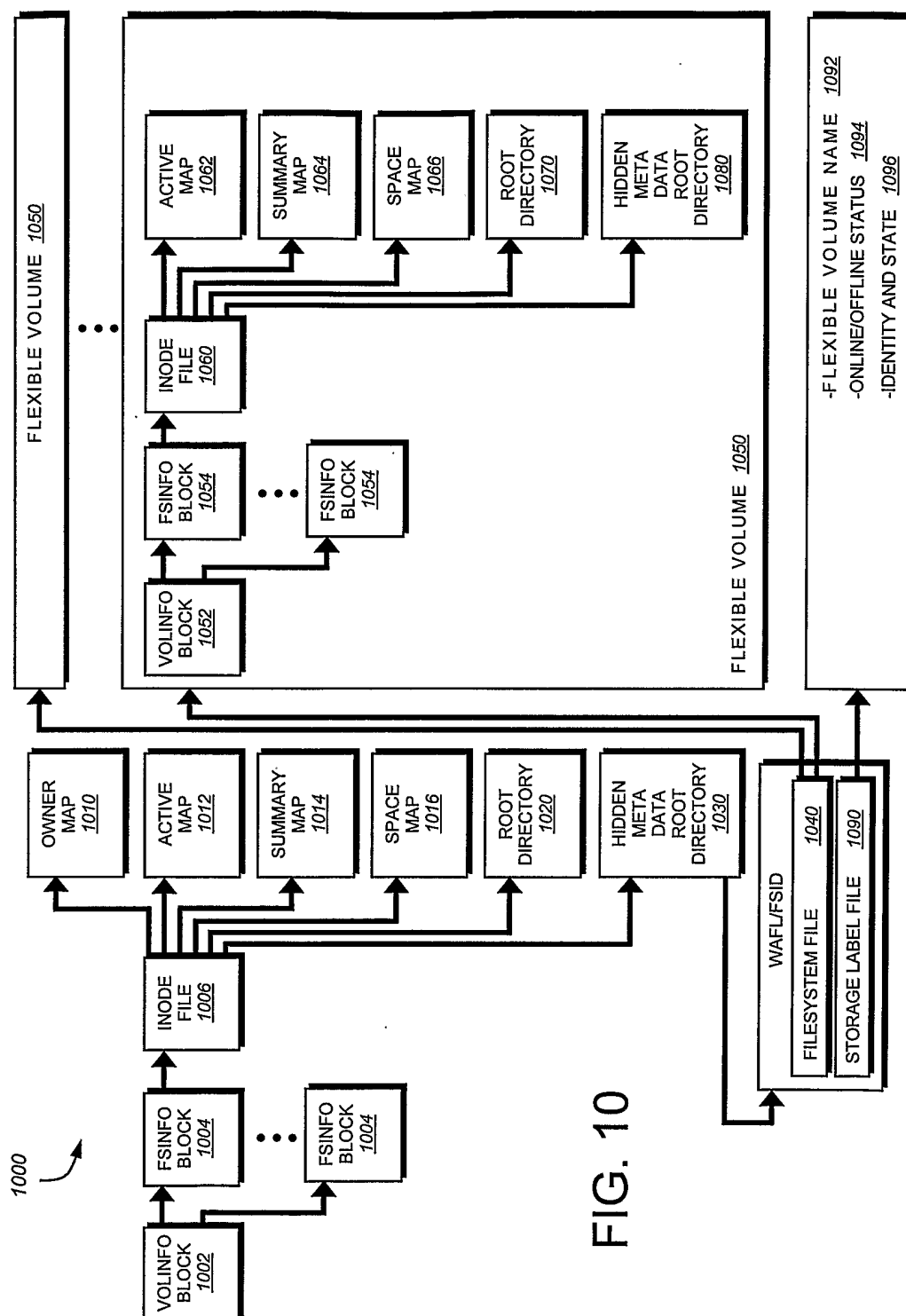


FIG. 9



9/17

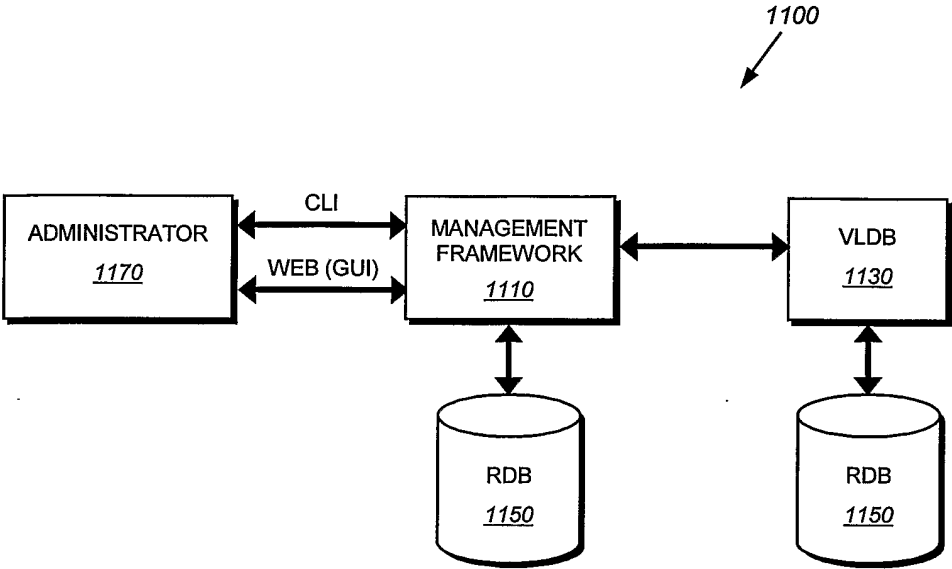


FIG. 11

10/17

1200

VOLUME ID	<u>1205</u>
AGGREGATE ID	<u>1210</u>
⋮	<u>1215</u>

FIG. 12

1300

AGGREGATE ID	<u>1305</u>
D-BLADE ID	<u>1310</u>
⋮	<u>1315</u>

FIG. 13

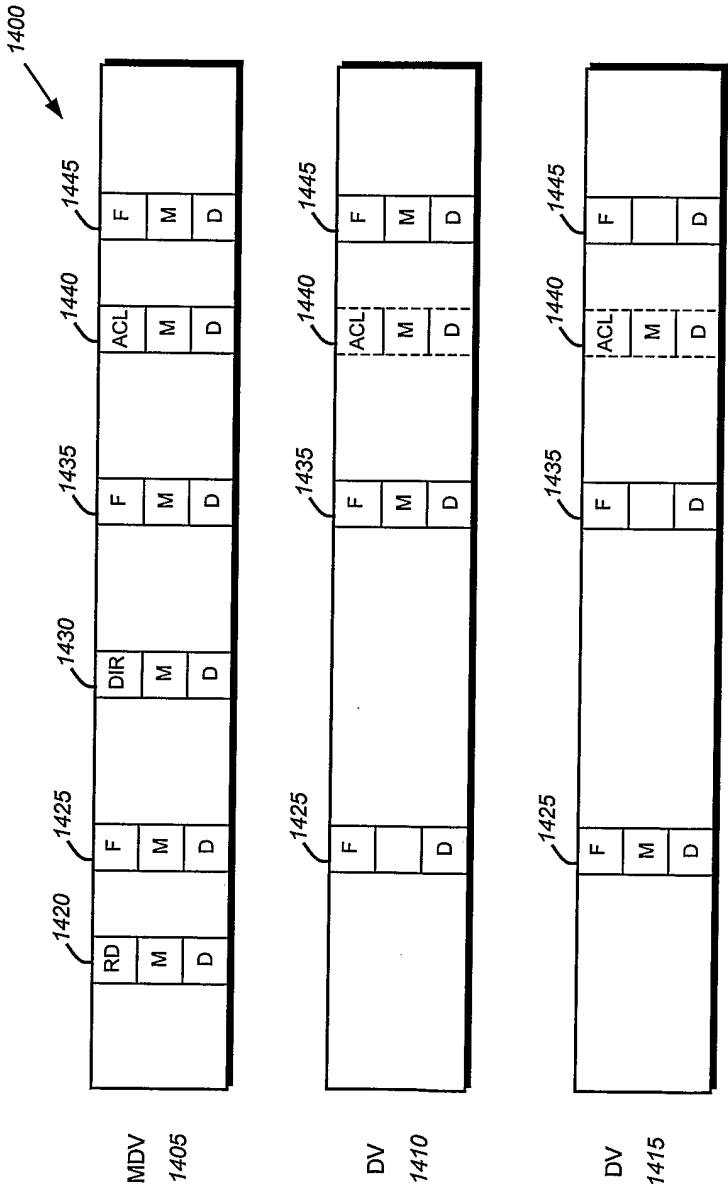


FIG. 14

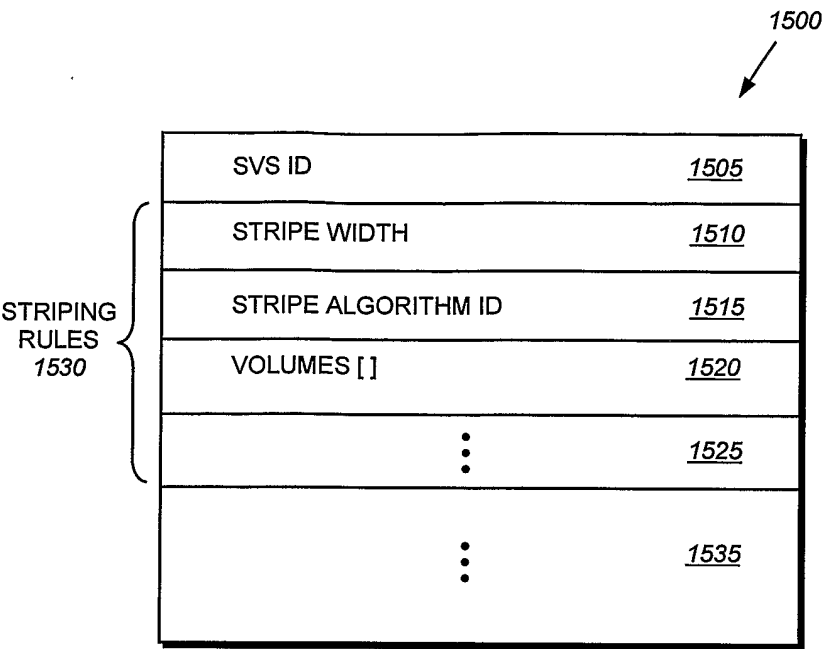


FIG. 15

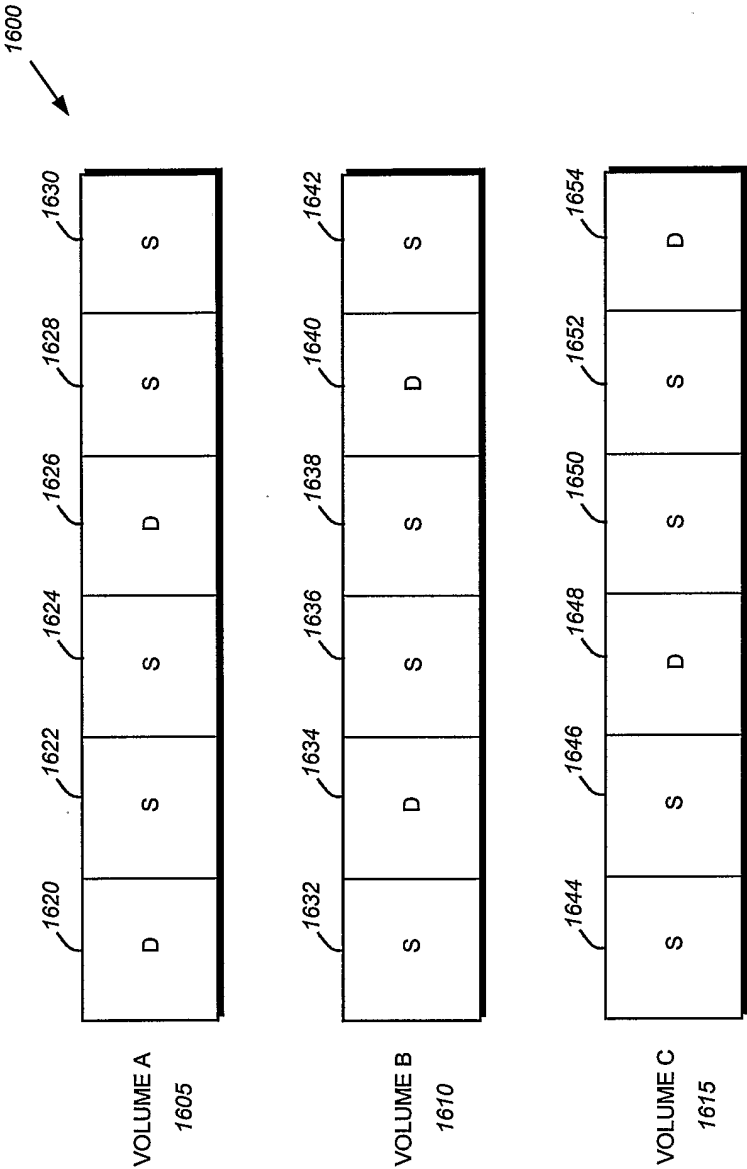


FIG. 16

14/17

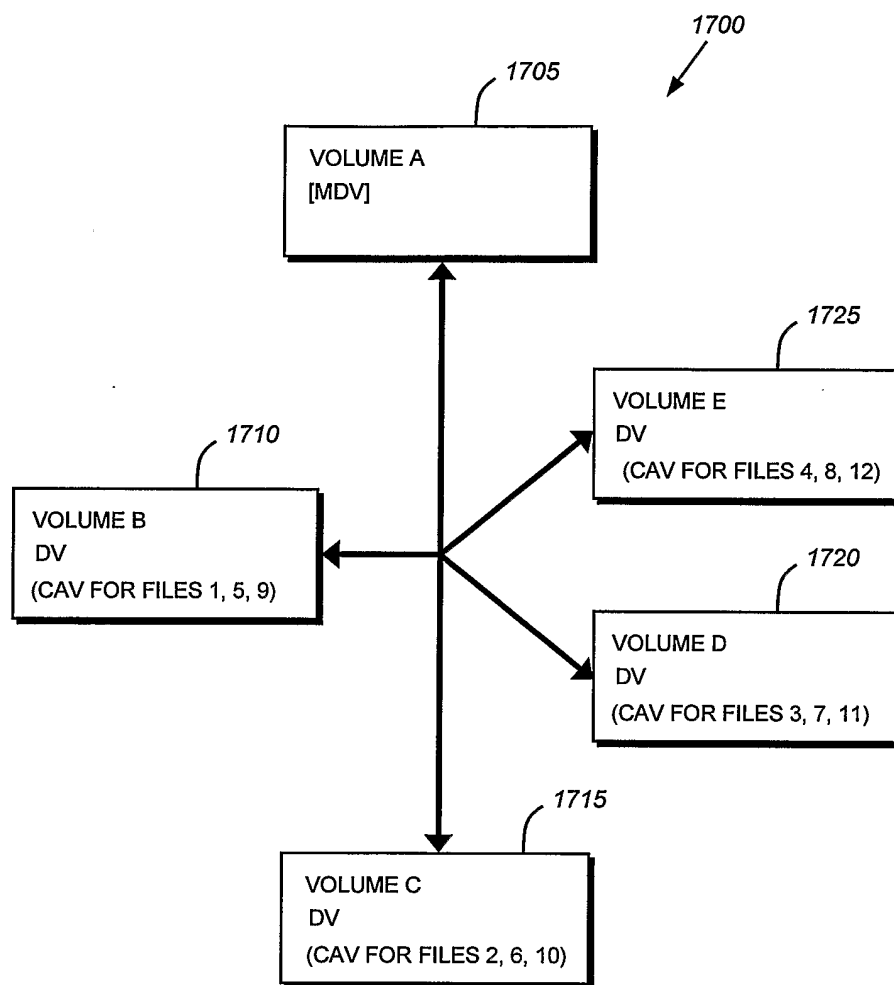


FIG. 17

15/17

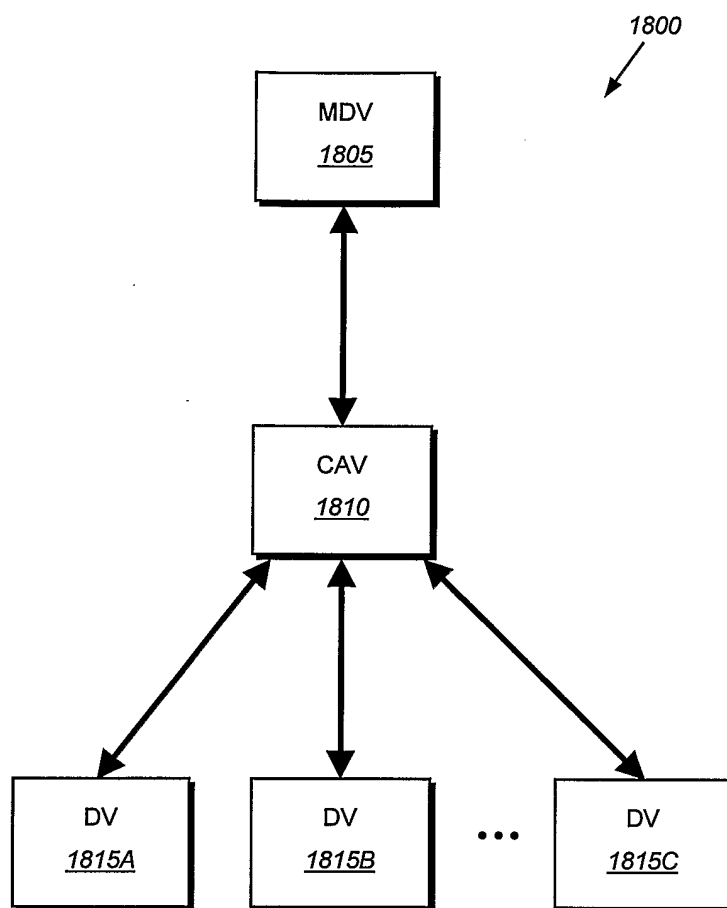


FIG. 18

16/17

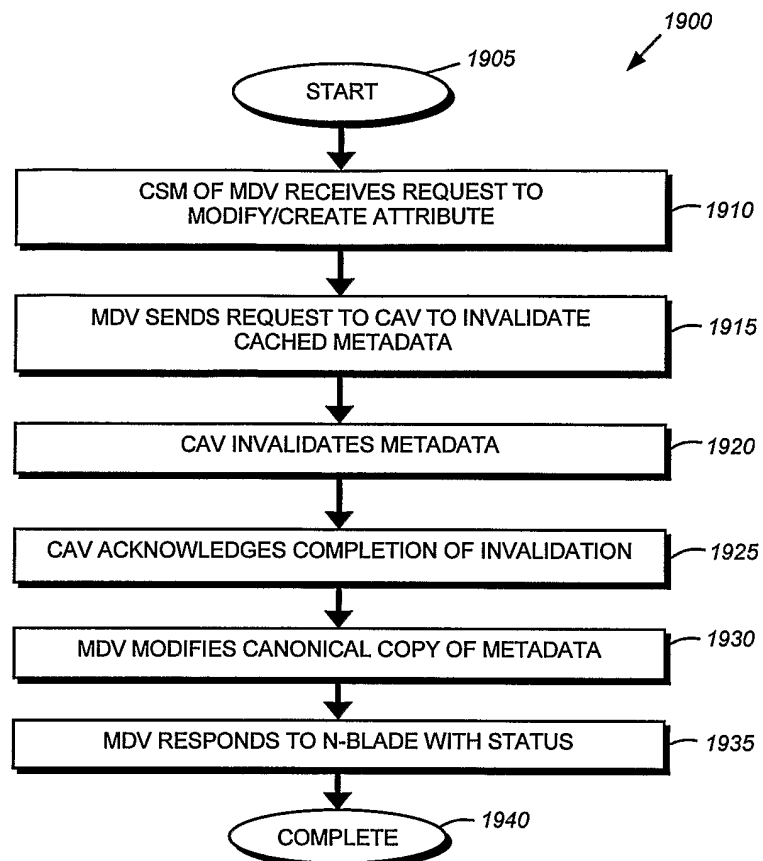


FIG. 19

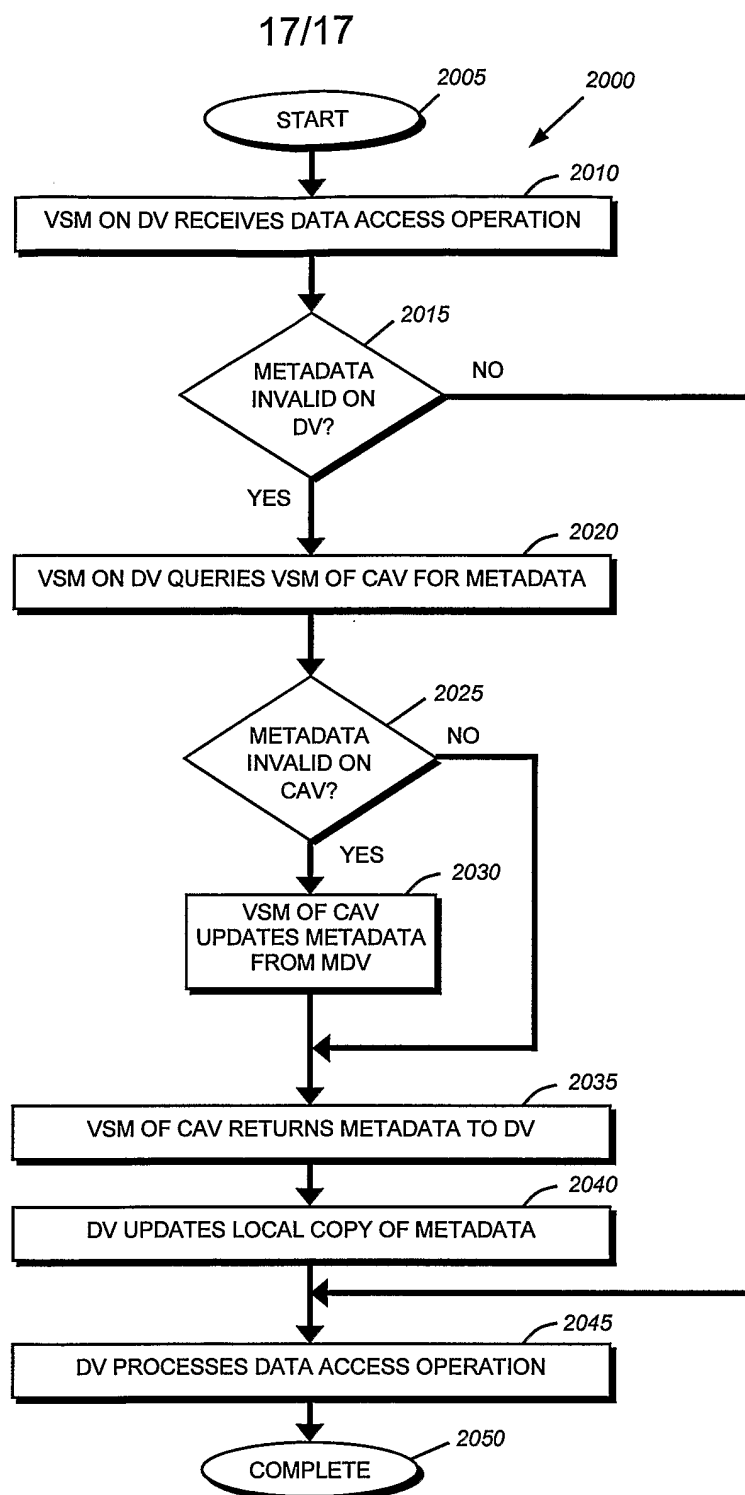


FIG. 20