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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 March 2021 (18.03.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/050875 A1

(51) International Patent Classification:
G06F 11/14 (2006.01) *G06F 11/20* (2006.01)

(21) International Application Number:
PCT/US2020/050408

(22) International Filing Date:
11 September 2020 (11.09.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/900,330 13 September 2019 (13.09.2019) US
16/668,794 30 October 2019 (30.10.2019) US

(71) Applicant: **PURE STORAGE, INC.** [US/US]; 650 Castro Street, Suite 400, Mountain View, California 94041 (US).

(72) Inventors: **GRUNWALD, David**; 650 Castro Street, Suite 260, Mountain View, California 94041 (US). **GILL, Thomas**; 650 Castro Street, Suite 260, Mountain View, California 94041 (US). **COLGROVE, John**; 650 Castro Street, Suite 260, Mountain View, California 94041 (US). **KARR, Ronald**; 650 Castro Street, Suite 260, Mountain View, California 94041 (US). **FAY, Matthew**; 650 Castro Street, Suite 260, Mountain View, California 94041 (US). **PAULSEN, Luke**; 88 Flynn Avenue, Apt. C, Mountain View, California 94043 (US). **WANG, Zong**; 650 Castro Street, Suite 260, Mountain View, California 94041 (US).

(74) Agent: **LENART, Edward** et al.; KENNEDY LENART SPRAGGINS LLP, 301 Congress Avenue, Suite 1350, Austin, Texas 78701 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW,

(54) Title: CONFIGURABLE DATA REPLICATION

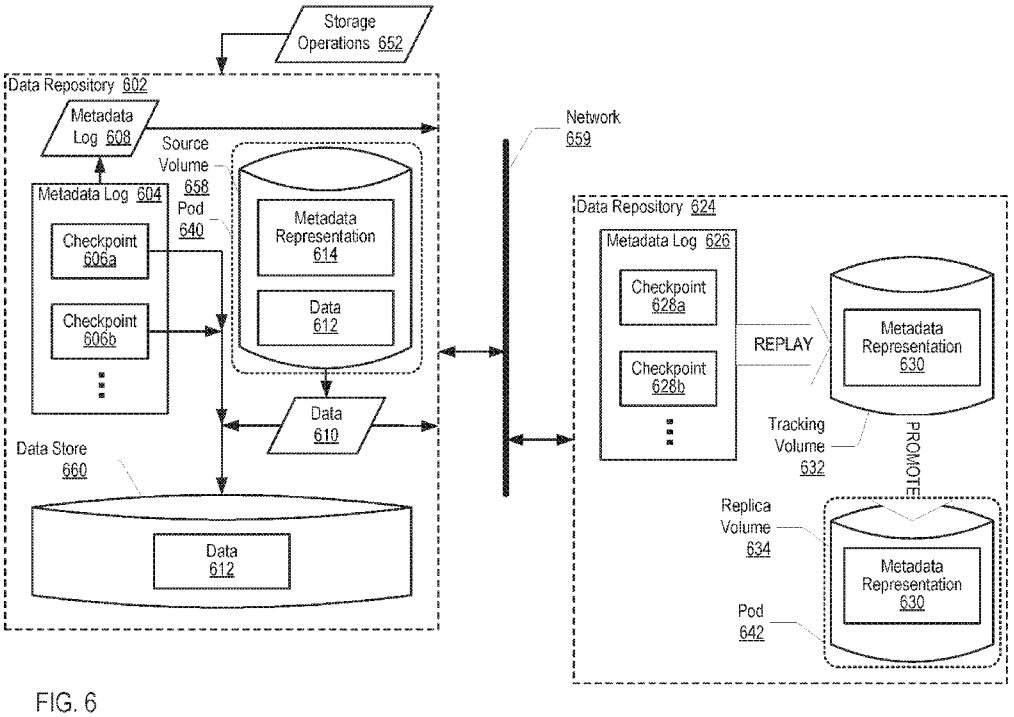


FIG. 6

(57) Abstract: Continuous data protection, including sending, to a target data repository from a source data repository, metadata describing one or more updates to one or more datasets stored within the source data repository; generating, based on the metadata describing the one or more updates to the one or more datasets, an ordered log of metadata describing an ordered application of the one or more updates to the one or more datasets; and generating, on the source data repository and based on the ordered log of metadata, the one or more datasets in accordance with the one or more updates corresponding to a specified point in time.

WO 2021/050875 A1

[Continued on next page]

SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

CONFIGURABLE DATA REPLICATION

BRIEF DESCRIPTION OF DRAWINGS

[0001] Figure 1A illustrates a first example system for data storage in accordance with some implementations.

[0002] Figure 1B illustrates a second example system for data storage in accordance with some implementations.

[0003] Figure 1C illustrates a third example system for data storage in accordance with some implementations.

[0004] Figure 1D illustrates a fourth example system for data storage in accordance with some implementations.

[0005] Figure 2A is a perspective view of a storage cluster with multiple storage nodes and internal storage coupled to each storage node to provide network attached storage, in accordance with some embodiments.

[0006] Figure 2B is a block diagram showing an interconnect switch coupling multiple storage nodes in accordance with some embodiments.

[0007] Figure 2C is a multiple level block diagram, showing contents of a storage node and contents of one of the non-volatile solid state storage units in accordance with some embodiments.

[0008] Figure 2D shows a storage server environment, which uses embodiments of the storage nodes and storage units of some previous figures in accordance with some embodiments.

[0009] Figure 2E is a blade hardware block diagram, showing a control plane, compute and storage planes, and authorities interacting with underlying physical resources, in accordance with some embodiments.

[00010] Figure 2F depicts elasticity software layers in blades of a storage cluster, in accordance with some embodiments.

[00011] Figure 2G depicts authorities and storage resources in blades of a storage cluster, in accordance with some embodiments.

[00012] Figure 3A sets forth a diagram of a storage system that is coupled for data communications with a cloud services provider in accordance with some embodiments of the present disclosure.

[00013] Figure 3B sets forth a diagram of a storage system in accordance with some embodiments of the present disclosure.

[00014] Figure 3C sets forth an example of a cloud-based storage system in accordance with some embodiments of the present disclosure.

[00015] Figure 3D illustrates an exemplary computing device that may be specifically configured to perform one or more of the processes described herein.

[00016] Figure 4A sets forth a block diagram illustrating a plurality of storage systems that support a pod according to some embodiments of the present disclosure.

[00017] Figure 4B sets forth a block diagram illustrating a metadata representation that supports a pod and configurable data replication according to some embodiments of the present disclosure.

[00018] Figure 5 sets forth a block diagram illustrating a plurality of storage systems that support a pod according to some embodiments of the present disclosure.

[00019] Figure 6 sets forth a block diagram illustrating a plurality of storage systems that support configurable data replication according to some embodiments of the present disclosure.

[00020] Figure 7 sets forth a block diagram illustrating a plurality of storage systems that support configurable data replication according to some embodiments of the present disclosure.

[00021] Figure 8 sets forth a flow chart illustrating an additional example method for configurable data replication according to some embodiments of the present disclosure.

[00022] Figure 9 sets forth a flow chart illustrating an example method for continuous data replication according to some embodiments of the present disclosure.

DESCRIPTION OF EMBODIMENTS

[00023] Example methods, apparatus, and products for configurable data replication in accordance with embodiments of the present disclosure are described with reference to the accompanying drawings, beginning with Figure 1A. Figure 1A illustrates an example system for data storage, in accordance with some implementations. System 100 (also referred to as "storage system" herein) includes numerous elements for purposes of illustration rather than limitation. It may be noted that system 100 may include the same, more, or fewer elements configured in the same or different manner in other implementations.

[00024] System 100 includes a number of computing devices 164A-B. Computing devices (also referred to as "client devices" herein) may be embodied, for example, a server in a data center, a workstation, a personal computer, a notebook, or the like. Computing devices 164A-B may be coupled for data communications to one or more storage arrays 102A-B through a storage area network ('SAN') 158 or a local area network ('LAN') 160.

[00025] The SAN 158 may be implemented with a variety of data communications fabrics, devices, and protocols. For example, the fabrics for SAN 158 may include Fibre Channel, Ethernet, Infiniband, Serial Attached Small Computer System Interface ('SAS'), or the like. Data communications protocols for use with SAN 158 may include Advanced Technology Attachment ('ATA'), Fibre Channel Protocol, Small Computer System Interface ('SCSI'), Internet Small Computer System Interface ('iSCSI'), HyperSCSI, Non-Volatile Memory Express ('NVMe') over Fabrics, or the like. It may be noted that SAN 158 is provided for illustration, rather than limitation. Other data communication couplings may be implemented between computing devices 164A-B and storage arrays 102A-B.

[00026] The LAN 160 may also be implemented with a variety of fabrics, devices, and protocols. For example, the fabrics for LAN 160 may include Ethernet (802.3), wireless (802.11), or the like. Data communication protocols for use in LAN 160 may include Transmission Control Protocol ('TCP'), User Datagram Protocol ('UDP'), Internet Protocol ('IP'), HyperText Transfer Protocol ('HTTP'), Wireless Access Protocol ('WAP'), Handheld Device Transport Protocol ('HDTP'), Session Initiation Protocol ('SIP'), Real Time Protocol ('RTP'), or the like.

[00027] Storage arrays 102A-B may provide persistent data storage for the computing devices 164A-B. Storage array 102A may be contained in a chassis (not shown), and storage array 102B may be contained in another chassis (not shown), in implementations. Storage array 102A and 102B may include one or more storage array controllers 110A-D (also referred to as "controller" herein). A storage array controller 110A-D may be embodied as a module of automated computing machinery comprising computer hardware, computer software, or a combination of computer hardware and software. In some implementations, the storage array controllers 110A-D may be configured to carry out various storage tasks. Storage tasks may include writing data received from the computing devices 164A-B to storage array 102A-B, erasing data from storage array 102A-B, retrieving data from storage array 102A-B and providing data to computing devices 164A-B, monitoring and reporting of disk utilization and performance, performing redundancy operations, such as Redundant Array of Independent Drives ('RAID') or RAID-like data redundancy operations, compressing data, encrypting data, and so forth.

[00028] Storage array controller 110A-D may be implemented in a variety of ways, including as a Field Programmable Gate Array ('FPGA'), a Programmable Logic Chip ('PLC'), an Application Specific Integrated Circuit ('ASIC'), System-on-Chip ('SOC'), or any computing device that includes discrete components such as a processing device, central processing unit,

computer memory, or various adapters. Storage array controller 110A-D may include, for example, a data communications adapter configured to support communications via the SAN 158 or LAN 160. In some implementations, storage array controller 110A-D may be independently coupled to the LAN 160. In implementations, storage array controller 110A-D may include an I/O controller or the like that couples the storage array controller 110A-D for data communications, through a midplane (not shown), to a persistent storage resource 170A-B (also referred to as a "storage resource" herein). The persistent storage resource 170A-B main include any number of storage drives 171A-F (also referred to as "storage devices" herein) and any number of non-volatile Random Access Memory ('NVRAM') devices (not shown).

[00029] In some implementations, the NVRAM devices of a persistent storage resource 170A-B may be configured to receive, from the storage array controller 110A-D, data to be stored in the storage drives 171A-F. In some examples, the data may originate from computing devices 164A-B. In some examples, writing data to the NVRAM device may be carried out more quickly than directly writing data to the storage drive 171A-F. In implementations, the storage array controller 110A-D may be configured to utilize the NVRAM devices as a quickly accessible buffer for data destined to be written to the storage drives 171A-F. Latency for write requests using NVRAM devices as a buffer may be improved relative to a system in which a storage array controller 110A-D writes data directly to the storage drives 171A-F. In some implementations, the NVRAM devices may be implemented with computer memory in the form of high bandwidth, low latency RAM. The NVRAM device is referred to as "non-volatile" because the NVRAM device may receive or include a unique power source that maintains the state of the RAM after main power loss to the NVRAM device. Such a power source may be a battery, one or more capacitors, or the like. In response to a power loss, the NVRAM device may be configured to write the contents of the RAM to a persistent storage, such as the storage drives 171A-F.

[00030] In implementations, storage drive 171A-F may refer to any device configured to record data persistently, where "persistently" or "persistent" refers as to a device's ability to maintain recorded data after loss of power. In some implementations, storage drive 171A-F may correspond to non-disk storage media. For example, the storage drive 171A-F may be one or more solid-state drives ('SSDs'), flash memory based storage, any type of solid-state non-volatile memory, or any other type of non-mechanical storage device. In other implementations, storage drive 171A-F may include mechanical or spinning hard disk, such as hard-disk drives ('HDD').

[00031] In some implementations, the storage array controllers 110A-D may be configured for offloading device management responsibilities from storage drive 171A-F in storage array 102A-B. For example, storage array controllers 110A-D may manage control information that may describe the state of one or more memory blocks in the storage drives 171A-F. The control information may indicate, for example, that a particular memory block has failed and should no longer be written to, that a particular memory block contains boot code for a storage array controller 110A-D, the number of program-erase ('P/E') cycles that have been performed on a particular memory block, the age of data stored in a particular memory block, the type of data that is stored in a particular memory block, and so forth. In some implementations, the control information may be stored with an associated memory block as metadata. In other implementations, the control information for the storage drives 171A-F may be stored in one or more particular memory blocks of the storage drives 171A-F that are selected by the storage array controller 110A-D. The selected memory blocks may be tagged with an identifier indicating that the selected memory block contains control information. The identifier may be utilized by the storage array controllers 110A-D in conjunction with storage drives 171A-F to quickly identify the memory blocks that contain control information. For example, the storage controllers 110A-D may issue a command to locate memory blocks that contain control information. It may be noted that control information may be so large that parts of the control information may be stored in multiple locations, that the control information may be stored in multiple locations for purposes of redundancy, for example, or that the control information may otherwise be distributed across multiple memory blocks in the storage drive 171A-F.

[00032] In implementations, storage array controllers 110A-D may offload device management responsibilities from storage drives 171A-F of storage array 102A-B by retrieving, from the storage drives 171A-F, control information describing the state of one or more memory blocks in the storage drives 171A-F. Retrieving the control information from the storage drives 171A-F may be carried out, for example, by the storage array controller 110A-D querying the storage drives 171A-F for the location of control information for a particular storage drive 171A-F. The storage drives 171A-F may be configured to execute instructions that enable the storage drive 171A-F to identify the location of the control information. The instructions may be executed by a controller (not shown) associated with or otherwise located on the storage drive 171A-F and may cause the storage drive 171A-F to scan a portion of each memory block to identify the memory blocks that store control information for the storage drives 171A-F. The storage drives 171A-F may respond by

sending a response message to the storage array controller 110A-D that includes the location of control information for the storage drive 171A-F. Responsive to receiving the response message, storage array controllers 110A-D may issue a request to read data stored at the address associated with the location of control information for the storage drives 171A-F.

[00033] In other implementations, the storage array controllers 110A-D may further offload device management responsibilities from storage drives 171A-F by performing, in response to receiving the control information, a storage drive management operation. A storage drive management operation may include, for example, an operation that is typically performed by the storage drive 171A-F (e.g., the controller (not shown) associated with a particular storage drive 171A-F). A storage drive management operation may include, for example, ensuring that data is not written to failed memory blocks within the storage drive 171A-F, ensuring that data is written to memory blocks within the storage drive 171A-F in such a way that adequate wear leveling is achieved, and so forth.

[00034] In implementations, storage array 102A-B may implement two or more storage array controllers 110A-D. For example, storage array 102A may include storage array controllers 110A and storage array controllers 110B. At a given instance, a single storage array controller 110A-D (e.g., storage array controller 110A) of a storage system 100 may be designated with primary status (also referred to as "primary controller" herein), and other storage array controllers 110A-D (e.g., storage array controller 110A) may be designated with secondary status (also referred to as "secondary controller" herein). The primary controller may have particular rights, such as permission to alter data in persistent storage resource 170A-B (e.g., writing data to persistent storage resource 170A-B). At least some of the rights of the primary controller may supersede the rights of the secondary controller. For instance, the secondary controller may not have permission to alter data in persistent storage resource 170A-B when the primary controller has the right. The status of storage array controllers 110A-D may change. For example, storage array controller 110A may be designated with secondary status, and storage array controller 110B may be designated with primary status.

[00035] In some implementations, a primary controller, such as storage array controller 110A, may serve as the primary controller for one or more storage arrays 102A-B, and a second controller, such as storage array controller 110B, may serve as the secondary controller for the one or more storage arrays 102A-B. For example, storage array controller 110A may be the primary controller for storage array 102A and storage array 102B, and storage array controller 110B may be the secondary controller for storage array 102A and

102B. In some implementations, storage array controllers 110C and 110D (also referred to as "storage processing modules") may neither have primary or secondary status. Storage array controllers 110C and 110D, implemented as storage processing modules, may act as a communication interface between the primary and secondary controllers (e.g., storage array controllers 110A and 110B, respectively) and storage array 102B. For example, storage array controller 110A of storage array 102A may send a write request, via SAN 158, to storage array 102B. The write request may be received by both storage array controllers 110C and 110D of storage array 102B. Storage array controllers 110C and 110D facilitate the communication, e.g., send the write request to the appropriate storage drive 171A-F. It may be noted that in some implementations storage processing modules may be used to increase the number of storage drives controlled by the primary and secondary controllers.

[00036] In implementations, storage array controllers 110A-D are communicatively coupled, via a midplane (not shown), to one or more storage drives 171A-F and to one or more NVRAM devices (not shown) that are included as part of a storage array 102A-B. The storage array controllers 110A-D may be coupled to the midplane via one or more data communication links and the midplane may be coupled to the storage drives 171A-F and the NVRAM devices via one or more data communications links. The data communications links described herein are collectively illustrated by data communications links 108A-D and may include a Peripheral Component Interconnect Express ('PCIe') bus, for example.

[00037] Figure 1B illustrates an example system for data storage, in accordance with some implementations. Storage array controller 101 illustrated in Figure 1B may be similar to the storage array controllers 110A-D described with respect to Figure 1A. In one example, storage array controller 101 may be similar to storage array controller 110A or storage array controller 110B. Storage array controller 101 includes numerous elements for purposes of illustration rather than limitation. It may be noted that storage array controller 101 may include the same, more, or fewer elements configured in the same or different manner in other implementations. It may be noted that elements of Figure 1A may be included below to help illustrate features of storage array controller 101.

[00038] Storage array controller 101 may include one or more processing devices 104 and random access memory ('RAM') 111. Processing device 104 (or controller 101) represents one or more general-purpose processing devices such as a microprocessor, central processing unit, or the like. More particularly, the processing device 104 (or controller 101) may be a complex instruction set computing ('CISC') microprocessor, reduced instruction set computing ('RISC') microprocessor, very long instruction word ('VLIW') microprocessor, or

a processor implementing other instruction sets or processors implementing a combination of instruction sets. The processing device 104 (or controller 101) may also be one or more special-purpose processing devices such as an application specific integrated circuit ('ASIC'), a field programmable gate array ('FPGA'), a digital signal processor ('DSP'), network processor, or the like.

[00039] The processing device 104 may be connected to the RAM 111 via a data communications link 106, which may be embodied as a high speed memory bus such as a Double-Data Rate 4 ('DDR4') bus. Stored in RAM 111 is an operating system 112. In some implementations, instructions 113 are stored in RAM 111. Instructions 113 may include computer program instructions for performing operations in in a direct-mapped flash storage system. In one embodiment, a direct-mapped flash storage system is one that that addresses data blocks within flash drives directly and without an address translation performed by the storage controllers of the flash drives.

[00040] In implementations, storage array controller 101 includes one or more host bus adapters 103A-C that are coupled to the processing device 104 via a data communications link 105A-C. In implementations, host bus adapters 103A-C may be computer hardware that connects a host system (e.g., the storage array controller) to other network and storage arrays. In some examples, host bus adapters 103A-C may be a Fibre Channel adapter that enables the storage array controller 101 to connect to a SAN, an Ethernet adapter that enables the storage array controller 101 to connect to a LAN, or the like. Host bus adapters 103A-C may be coupled to the processing device 104 via a data communications link 105A-C such as, for example, a PCIe bus.

[00041] In implementations, storage array controller 101 may include a host bus adapter 114 that is coupled to an expander 115. The expander 115 may be used to attach a host system to a larger number of storage drives. The expander 115 may, for example, be a SAS expander utilized to enable the host bus adapter 114 to attach to storage drives in an implementation where the host bus adapter 114 is embodied as a SAS controller.

[00042] In implementations, storage array controller 101 may include a switch 116 coupled to the processing device 104 via a data communications link 109. The switch 116 may be a computer hardware device that can create multiple endpoints out of a single endpoint, thereby enabling multiple devices to share a single endpoint. The switch 116 may, for example, be a PCIe switch that is coupled to a PCIe bus (e.g., data communications link 109) and presents multiple PCIe connection points to the midplane.

[00043] In implementations, storage array controller 101 includes a data communications link 107 for coupling the storage array controller 101 to other storage array controllers. In some examples, data communications link 107 may be a QuickPath Interconnect (QPI) interconnect.

[00044] A traditional storage system that uses traditional flash drives may implement a process across the flash drives that are part of the traditional storage system. For example, a higher level process of the storage system may initiate and control a process across the flash drives. However, a flash drive of the traditional storage system may include its own storage controller that also performs the process. Thus, for the traditional storage system, a higher level process (e.g., initiated by the storage system) and a lower level process (e.g., initiated by a storage controller of the storage system) may both be performed.

[00045] To resolve various deficiencies of a traditional storage system, operations may be performed by higher level processes and not by the lower level processes. For example, the flash storage system may include flash drives that do not include storage controllers that provide the process. Thus, the operating system of the flash storage system itself may initiate and control the process. This may be accomplished by a direct-mapped flash storage system that addresses data blocks within the flash drives directly and without an address translation performed by the storage controllers of the flash drives.

[00046] The operating system of the flash storage system may identify and maintain a list of allocation units across multiple flash drives of the flash storage system. The allocation units may be entire erase blocks or multiple erase blocks. The operating system may maintain a map or address range that directly maps addresses to erase blocks of the flash drives of the flash storage system.

[00047] Direct mapping to the erase blocks of the flash drives may be used to rewrite data and erase data. For example, the operations may be performed on one or more allocation units that include a first data and a second data where the first data is to be retained and the second data is no longer being used by the flash storage system. The operating system may initiate the process to write the first data to new locations within other allocation units and erasing the second data and marking the allocation units as being available for use for subsequent data. Thus, the process may only be performed by the higher level operating system of the flash storage system without an additional lower level process being performed by controllers of the flash drives.

[00048] Advantages of the process being performed only by the operating system of the flash storage system include increased reliability of the flash drives of the flash storage system as

unnecessary or redundant write operations are not being performed during the process. One possible point of novelty here is the concept of initiating and controlling the process at the operating system of the flash storage system. In addition, the process can be controlled by the operating system across multiple flash drives. This is contrast to the process being performed by a storage controller of a flash drive.

[00049] A storage system can consist of two storage array controllers that share a set of drives for failover purposes, or it could consist of a single storage array controller that provides a storage service that utilizes multiple drives, or it could consist of a distributed network of storage array controllers each with some number of drives or some amount of Flash storage where the storage array controllers in the network collaborate to provide a complete storage service and collaborate on various aspects of a storage service including storage allocation and garbage collection.

[00050] Figure 1C illustrates a third example system 117 for data storage in accordance with some implementations. System 117 (also referred to as "storage system" herein) includes numerous elements for purposes of illustration rather than limitation. It may be noted that system 117 may include the same, more, or fewer elements configured in the same or different manner in other implementations.

[00051] In one embodiment, system 117 includes a dual Peripheral Component Interconnect ('PCI') flash storage device 118 with separately addressable fast write storage. System 117 may include a storage controller 119. In one embodiment, storage controller 119A-D may be a CPU, ASIC, FPGA, or any other circuitry that may implement control structures necessary according to the present disclosure. In one embodiment, system 117 includes flash memory devices (e.g., including flash memory devices 120a-n), operatively coupled to various channels of the storage device controller 119. Flash memory devices 120a-n, may be presented to the controller 119A-D as an addressable collection of Flash pages, erase blocks, and/or control elements sufficient to allow the storage device controller 119A-D to program and retrieve various aspects of the Flash. In one embodiment, storage device controller 119A-D may perform operations on flash memory devices 120a-n including storing and retrieving data content of pages, arranging and erasing any blocks, tracking statistics related to the use and reuse of Flash memory pages, erase blocks, and cells, tracking and predicting error codes and faults within the Flash memory, controlling voltage levels associated with programming and retrieving contents of Flash cells, etc.

[00052] In one embodiment, system 117 may include RAM 121 to store separately addressable fast-write data. In one embodiment, RAM 121 may be one or more separate

discrete devices. In another embodiment, RAM 121 may be integrated into storage device controller 119A-D or multiple storage device controllers. The RAM 121 may be utilized for other purposes as well, such as temporary program memory for a processing device (e.g., a CPU) in the storage device controller 119.

[00053] In one embodiment, system 117 may include a stored energy device 122, such as a rechargeable battery or a capacitor. Stored energy device 122 may store energy sufficient to power the storage device controller 119, some amount of the RAM (e.g., RAM 121), and some amount of Flash memory (e.g., Flash memory 120a-120n) for sufficient time to write the contents of RAM to Flash memory. In one embodiment, storage device controller 119A-D may write the contents of RAM to Flash Memory if the storage device controller detects loss of external power.

[00054] In one embodiment, system 117 includes two data communications links 123a, 123b. In one embodiment, data communications links 123a, 123b may be PCI interfaces. In another embodiment, data communications links 123a, 123b may be based on other communications standards (e.g., HyperTransport, InfiniBand, etc.). Data communications links 123a, 123b may be based on non-volatile memory express ('NVMe') or NVMe over fabrics ('NVMf') specifications that allow external connection to the storage device controller 119A-D from other components in the storage system 117. It should be noted that data communications links may be interchangeably referred to herein as PCI buses for convenience.

[00055] System 117 may also include an external power source (not shown), which may be provided over one or both data communications links 123a, 123b, or which may be provided separately. An alternative embodiment includes a separate Flash memory (not shown) dedicated for use in storing the content of RAM 121. The storage device controller 119A-D may present a logical device over a PCI bus which may include an addressable fast-write logical device, or a distinct part of the logical address space of the storage device 118, which may be presented as PCI memory or as persistent storage. In one embodiment, operations to store into the device are directed into the RAM 121. On power failure, the storage device controller 119A-D may write stored content associated with the addressable fast-write logical storage to Flash memory (e.g., Flash memory 120a-n) for long-term persistent storage.

[00056] In one embodiment, the logical device may include some presentation of some or all of the content of the Flash memory devices 120a-n, where that presentation allows a storage system including a storage device 118 (e.g., storage system 117) to directly address Flash memory pages and directly reprogram erase blocks from storage system components that are

external to the storage device through the PCI bus. The presentation may also allow one or more of the external components to control and retrieve other aspects of the Flash memory including some or all of: tracking statistics related to use and reuse of Flash memory pages, erase blocks, and cells across all the Flash memory devices; tracking and predicting error codes and faults within and across the Flash memory devices; controlling voltage levels associated with programming and retrieving contents of Flash cells; etc.

[00057] In one embodiment, the stored energy device 122 may be sufficient to ensure completion of in-progress operations to the Flash memory devices 120a-120n stored energy device 122 may power storage device controller 119A-D and associated Flash memory devices (e.g., 120a-n) for those operations, as well as for the storing of fast-write RAM to Flash memory. Stored energy device 122 may be used to store accumulated statistics and other parameters kept and tracked by the Flash memory devices 120a-n and/or the storage device controller 119. Separate capacitors or stored energy devices (such as smaller capacitors near or embedded within the Flash memory devices themselves) may be used for some or all of the operations described herein.

[00058] Various schemes may be used to track and optimize the life span of the stored energy component, such as adjusting voltage levels over time, partially discharging the storage energy device 122 to measure corresponding discharge characteristics, etc. If the available energy decreases over time, the effective available capacity of the addressable fast-write storage may be decreased to ensure that it can be written safely based on the currently available stored energy.

[00059] Figure 1D illustrates a third example system 124 for data storage in accordance with some implementations. In one embodiment, system 124 includes storage controllers 125a, 125b. In one embodiment, storage controllers 125a, 125b are operatively coupled to Dual PCI storage devices 119a, 119b and 119c, 119d, respectively. Storage controllers 125a, 125b may be operatively coupled (e.g., via a storage network 130) to some number of host computers 127a-n.

[00060] In one embodiment, two storage controllers (e.g., 125a and 125b) provide storage services, such as a SCS) block storage array, a file server, an object server, a database or data analytics service, etc. The storage controllers 125a, 125b may provide services through some number of network interfaces (e.g., 126a-d) to host computers 127a-n outside of the storage system 124. Storage controllers 125a, 125b may provide integrated services or an application entirely within the storage system 124, forming a converged storage and compute system. The storage controllers 125a, 125b may utilize the fast write memory within or across storage

devices119a-d to journal in progress operations to ensure the operations are not lost on a power failure, storage controller removal, storage controller or storage system shutdown, or some fault of one or more software or hardware components within the storage system 124.

[00061] In one embodiment, controllers 125a, 125b operate as PCI masters to one or the other PCI buses 128a, 128b. In another embodiment, 128a and 128b may be based on other communications standards (e.g., HyperTransport, InfiniBand, etc.). Other storage system embodiments may operate storage controllers 125a, 125b as multi-masters for both PCI buses 128a, 128b. Alternately, a PCI/NVMe/NVMf switching infrastructure or fabric may connect multiple storage controllers. Some storage system embodiments may allow storage devices to communicate with each other directly rather than communicating only with storage controllers. In one embodiment, a storage device controller 119a may be operable under direction from a storage controller 125a to synthesize and transfer data to be stored into Flash memory devices from data that has been stored in RAM (e.g., RAM 121 of Figure 1C). For example, a recalculated version of RAM content may be transferred after a storage controller has determined that an operation has fully committed across the storage system, or when fast-write memory on the device has reached a certain used capacity, or after a certain amount of time, to ensure improve safety of the data or to release addressable fast-write capacity for reuse. This mechanism may be used, for example, to avoid a second transfer over a bus (e.g., 128a, 128b) from the storage controllers 125a, 125b. In one embodiment, a recalculation may include compressing data, attaching indexing or other metadata, combining multiple data segments together, performing erasure code calculations, etc.

[00062] In one embodiment, under direction from a storage controller 125a, 125b, a storage device controller 119a, 119b may be operable to calculate and transfer data to other storage devices from data stored in RAM (e.g., RAM 121 of Figure 1C) without involvement of the storage controllers 125a, 125b. This operation may be used to mirror data stored in one controller 125a to another controller 125b, or it could be used to offload compression, data aggregation, and/or erasure coding calculations and transfers to storage devices to reduce load on storage controllers or the storage controller interface 129a, 129b to the PCI bus 128a, 128b.

[00063] A storage device controller 119A-D may include mechanisms for implementing high availability primitives for use by other parts of a storage system external to the Dual PCI storage device 118. For example, reservation or exclusion primitives may be provided so that, in a storage system with two storage controllers providing a highly available storage service, one storage controller may prevent the other storage controller from accessing or

continuing to access the storage device. This could be used, for example, in cases where one controller detects that the other controller is not functioning properly or where the interconnect between the two storage controllers may itself not be functioning properly.

[00064] In one embodiment, a storage system for use with Dual PCI direct mapped storage devices with separately addressable fast write storage includes systems that manage erase blocks or groups of erase blocks as allocation units for storing data on behalf of the storage service, or for storing metadata (e.g., indexes, logs, etc.) associated with the storage service, or for proper management of the storage system itself. Flash pages, which may be a few kilobytes in size, may be written as data arrives or as the storage system is to persist data for long intervals of time (e.g., above a defined threshold of time). To commit data more quickly, or to reduce the number of writes to the Flash memory devices, the storage controllers may first write data into the separately addressable fast write storage on one more storage devices.

[00065] In one embodiment, the storage controllers 125a, 125b may initiate the use of erase blocks within and across storage devices (e.g., 118) in accordance with an age and expected remaining lifespan of the storage devices, or based on other statistics. The storage controllers 125a, 125b may initiate garbage collection and data migration data between storage devices in accordance with pages that are no longer needed as well as to manage Flash page and erase block lifespans and to manage overall system performance.

[00066] In one embodiment, the storage system 124 may utilize mirroring and/or erasure coding schemes as part of storing data into addressable fast write storage and/or as part of writing data into allocation units associated with erase blocks. Erasure codes may be used across storage devices, as well as within erase blocks or allocation units, or within and across Flash memory devices on a single storage device, to provide redundancy against single or multiple storage device failures or to protect against internal corruptions of Flash memory pages resulting from Flash memory operations or from degradation of Flash memory cells. Mirroring and erasure coding at various levels may be used to recover from multiple types of failures that occur separately or in combination.

[00067] The embodiments depicted with reference to Figs. 2A-G illustrate a storage cluster that stores user data, such as user data originating from one or more user or client systems or other sources external to the storage cluster. The storage cluster distributes user data across storage nodes housed within a chassis, or across multiple chassis, using erasure coding and redundant copies of metadata. Erasure coding refers to a method of data protection or reconstruction in which data is stored across a set of different locations, such as disks, storage

nodes or geographic locations. Flash memory is one type of solid-state memory that may be integrated with the embodiments, although the embodiments may be extended to other types of solid-state memory or other storage medium, including non- solid state memory. Control of storage locations and workloads are distributed across the storage locations in a clustered peer-to-peer system. Tasks such as mediating communications between the various storage nodes, detecting when a storage node has become unavailable, and balancing I/Os (inputs and outputs) across the various storage nodes, are all handled on a distributed basis. Data is laid out or distributed across multiple storage nodes in data fragments or stripes that support data recovery in some embodiments. Ownership of data can be reassigned within a cluster, independent of input and output patterns. This architecture described in more detail below allows a storage node in the cluster to fail, with the system remaining operational, since the data can be reconstructed from other storage nodes and thus remain available for input and output operations. In various embodiments, a storage node may be referred to as a cluster node, a blade, or a server.

[00068] The storage cluster may be contained within a chassis, i.e., an enclosure housing one or more storage nodes. A mechanism to provide power to each storage node, such as a power distribution bus, and a communication mechanism, such as a communication bus that enables communication between the storage nodes are included within the chassis. The storage cluster can run as an independent system in one location according to some embodiments. In one embodiment, a chassis contains at least two instances of both the power distribution and the communication bus which may be enabled or disabled independently. The internal communication bus may be an Ethernet bus, however, other technologies such as PCIe, InfiniBand, and others, are equally suitable. The chassis provides a port for an external communication bus for enabling communication between multiple chassis, directly or through a switch, and with client systems. The external communication may use a technology such as Ethernet, InfiniBand, Fibre Channel, etc. In some embodiments, the external communication bus uses different communication bus technologies for inter-chassis and client communication. If a switch is deployed within or between chassis, the switch may act as a translation between multiple protocols or technologies. When multiple chassis are connected to define a storage cluster, the storage cluster may be accessed by a client using either proprietary interfaces or standard interfaces such as network file system ('NFS'), common internet file system ('CIFS'), small computer system interface ('SCSI') or hypertext transfer protocol ('HTTP'). Translation from the client protocol may occur at the switch, chassis external communication bus or within each storage node. In some embodiments,

multiple chassis may be coupled or connected to each other through an aggregator switch. A portion and/or all of the coupled or connected chassis may be designated as a storage cluster. As discussed above, each chassis can have multiple blades, each blade has a media access control ('MAC') address, but the storage cluster is presented to an external network as having a single cluster IP address and a single MAC address in some embodiments.

[00069] Each storage node may be one or more storage servers and each storage server is connected to one or more non-volatile solid state memory units, which may be referred to as storage units or storage devices. One embodiment includes a single storage server in each storage node and between one to eight non-volatile solid state memory units, however this one example is not meant to be limiting. The storage server may include a processor, DRAM and interfaces for the internal communication bus and power distribution for each of the power buses. Inside the storage node, the interfaces and storage unit share a communication bus, e.g., PCI Express, in some embodiments. The non-volatile solid state memory units may directly access the internal communication bus interface through a storage node communication bus, or request the storage node to access the bus interface. The non-volatile solid state memory unit contains an embedded CPU, solid state storage controller, and a quantity of solid state mass storage, e.g., between 2-32 terabytes ('TB') in some embodiments. An embedded volatile storage medium, such as DRAM, and an energy reserve apparatus are included in the non-volatile solid state memory unit. In some embodiments, the energy reserve apparatus is a capacitor, super-capacitor, or battery that enables transferring a subset of DRAM contents to a stable storage medium in the case of power loss. In some embodiments, the non-volatile solid state memory unit is constructed with a storage class memory, such as phase change or magnetoresistive random access memory ('MRAM') that substitutes for DRAM and enables a reduced power hold-up apparatus.

[00070] One of many features of the storage nodes and non-volatile solid state storage is the ability to proactively rebuild data in a storage cluster. The storage nodes and non-volatile solid state storage can determine when a storage node or non-volatile solid state storage in the storage cluster is unreachable, independent of whether there is an attempt to read data involving that storage node or non-volatile solid state storage. The storage nodes and non-volatile solid state storage then cooperate to recover and rebuild the data in at least partially new locations. This constitutes a proactive rebuild, in that the system rebuilds data without waiting until the data is needed for a read access initiated from a client system employing the storage cluster. These and further details of the storage memory and operation thereof are discussed below.

[00071] Figure 2A is a perspective view of a storage cluster 161, with multiple storage nodes 150 and internal solid-state memory coupled to each storage node to provide network attached storage or storage area network, in accordance with some embodiments. A network attached storage, storage area network, or a storage cluster, or other storage memory, could include one or more storage clusters 161, each having one or more storage nodes 150, in a flexible and reconfigurable arrangement of both the physical components and the amount of storage memory provided thereby. The storage cluster 161 is designed to fit in a rack, and one or more racks can be set up and populated as desired for the storage memory. The storage cluster 161 has a chassis 138 having multiple slots 142. It should be appreciated that chassis 138 may be referred to as a housing, enclosure, or rack unit. In one embodiment, the chassis 138 has fourteen slots 142, although other numbers of slots are readily devised. For example, some embodiments have four slots, eight slots, sixteen slots, thirty-two slots, or other suitable number of slots. Each slot 142 can accommodate one storage node 150 in some embodiments. Chassis 138 includes flaps 148 that can be utilized to mount the chassis 138 on a rack. Fans 144 provide air circulation for cooling of the storage nodes 150 and components thereof, although other cooling components could be used, or an embodiment could be devised without cooling components. A switch fabric 146 couples storage nodes 150 within chassis 138 together and to a network for communication to the memory. In an embodiment depicted in herein, the slots 142 to the left of the switch fabric 146 and fans 144 are shown occupied by storage nodes 150, while the slots 142 to the right of the switch fabric 146 and fans 144 are empty and available for insertion of storage node 150 for illustrative purposes. This configuration is one example, and one or more storage nodes 150 could occupy the slots 142 in various further arrangements. The storage node arrangements need not be sequential or adjacent in some embodiments. Storage nodes 150 are hot pluggable, meaning that a storage node 150 can be inserted into a slot 142 in the chassis 138, or removed from a slot 142, without stopping or powering down the system. Upon insertion or removal of storage node 150 from slot 142, the system automatically reconfigures in order to recognize and adapt to the change. Reconfiguration, in some embodiments, includes restoring redundancy and/or rebalancing data or load.

[00072] Each storage node 150 can have multiple components. In the embodiment shown here, the storage node 150 includes a printed circuit board 159 populated by a CPU 156, i.e., processor, a memory 154 coupled to the CPU 156, and a non-volatile solid state storage 152 coupled to the CPU 156, although other mountings and/or components could be used in further embodiments. The memory 154 has instructions which are executed by the CPU 156

and/or data operated on by the CPU 156. As further explained below, the non-volatile solid state storage 152 includes flash or, in further embodiments, other types of solid-state memory.

[00073] Referring to Figure 2A, storage cluster 161 is scalable, meaning that storage capacity with non-uniform storage sizes is readily added, as described above. One or more storage nodes 150 can be plugged into or removed from each chassis and the storage cluster self-configures in some embodiments. Plug-in storage nodes 150, whether installed in a chassis as delivered or later added, can have different sizes. For example, in one embodiment a storage node 150 can have any multiple of 4 TB, e.g., 8 TB, 12 TB, 16 TB, 32 TB, etc. In further embodiments, a storage node 150 could have any multiple of other storage amounts or capacities. Storage capacity of each storage node 150 is broadcast, and influences decisions of how to stripe the data. For maximum storage efficiency, an embodiment can self-configure as wide as possible in the stripe, subject to a predetermined requirement of continued operation with loss of up to one, or up to two, non-volatile solid state storage units 152 or storage nodes 150 within the chassis.

[00074] Figure 2B is a block diagram showing a communications interconnect 173 and power distribution bus 172 coupling multiple storage nodes 150. Referring back to Figure 2A, the communications interconnect 173 can be included in or implemented with the switch fabric 146 in some embodiments. Where multiple storage clusters 161 occupy a rack, the communications interconnect 173 can be included in or implemented with a top of rack switch, in some embodiments. As illustrated in Figure 2B, storage cluster 161 is enclosed within a single chassis 138. External port 176 is coupled to storage nodes 150 through communications interconnect 173, while external port 174 is coupled directly to a storage node. External power port 178 is coupled to power distribution bus 172. Storage nodes 150 may include varying amounts and differing capacities of non-volatile solid state storage 152 as described with reference to Figure 2A. In addition, one or more storage nodes 150 may be a compute only storage node as illustrated in Figure 2B. Authorities 168 are implemented on the non-volatile solid state storages 152, for example as lists or other data structures stored in memory. In some embodiments the authorities are stored within the non-volatile solid state storage 152 and supported by software executing on a controller or other processor of the non-volatile solid state storage 152. In a further embodiment, authorities 168 are implemented on the storage nodes 150, for example as lists or other data structures stored in the memory 154 and supported by software executing on the CPU 156 of the storage node 150. Authorities 168 control how and where data is stored in the non-volatile solid state

storages 152 in some embodiments. This control assists in determining which type of erasure coding scheme is applied to the data, and which storage nodes 150 have which portions of the data. Each authority 168 may be assigned to a non-volatile solid state storage 152. Each authority may control a range of inode numbers, segment numbers, or other data identifiers which are assigned to data by a file system, by the storage nodes 150, or by the non-volatile solid state storage 152, in various embodiments.

[00075] Every piece of data, and every piece of metadata, has redundancy in the system in some embodiments. In addition, every piece of data and every piece of metadata has an owner, which may be referred to as an authority. If that authority is unreachable, for example through failure of a storage node, there is a plan of succession for how to find that data or that metadata. In various embodiments, there are redundant copies of authorities 168. Authorities 168 have a relationship to storage nodes 150 and non-volatile solid state storage 152 in some embodiments. Each authority 168, covering a range of data segment numbers or other identifiers of the data, may be assigned to a specific non-volatile solid state storage 152. In some embodiments the authorities 168 for all of such ranges are distributed over the non-volatile solid state storages 152 of a storage cluster. Each storage node 150 has a network port that provides access to the non-volatile solid state storage(s) 152 of that storage node 150. Data can be stored in a segment, which is associated with a segment number and that segment number is an indirection for a configuration of a RAID (redundant array of independent disks) stripe in some embodiments. The assignment and use of the authorities 168 thus establishes an indirection to data. Indirection may be referred to as the ability to reference data indirectly, in this case via an authority 168, in accordance with some embodiments. A segment identifies a set of non-volatile solid state storage 152 and a local identifier into the set of non-volatile solid state storage 152 that may contain data. In some embodiments, the local identifier is an offset into the device and may be reused sequentially by multiple segments. In other embodiments the local identifier is unique for a specific segment and never reused. The offsets in the non-volatile solid state storage 152 are applied to locating data for writing to or reading from the non-volatile solid state storage 152 (in the form of a RAID stripe). Data is striped across multiple units of non-volatile solid state storage 152, which may include or be different from the non-volatile solid state storage 152 having the authority 168 for a particular data segment.

[00076] If there is a change in where a particular segment of data is located, e.g., during a data move or a data reconstruction, the authority 168 for that data segment should be consulted, at that non-volatile solid state storage 152 or storage node 150 having that

authority 168. In order to locate a particular piece of data, embodiments calculate a hash value for a data segment or apply an inode number or a data segment number. The output of this operation points to a non-volatile solid state storage 152 having the authority 168 for that particular piece of data. In some embodiments there are two stages to this operation. The first stage maps an entity identifier (ID), e.g., a segment number, inode number, or directory number to an authority identifier. This mapping may include a calculation such as a hash or a bit mask. The second stage is mapping the authority identifier to a particular non-volatile solid state storage 152, which may be done through an explicit mapping. The operation is repeatable, so that when the calculation is performed, the result of the calculation repeatably and reliably points to a particular non-volatile solid state storage 152 having that authority 168. The operation may include the set of reachable storage nodes as input. If the set of reachable non-volatile solid state storage units changes the optimal set changes. In some embodiments, the persisted value is the current assignment (which is always true) and the calculated value is the target assignment the cluster will attempt to reconfigure towards. This calculation may be used to determine the optimal non-volatile solid state storage 152 for an authority in the presence of a set of non-volatile solid state storage 152 that are reachable and constitute the same cluster. The calculation also determines an ordered set of peer non-volatile solid state storage 152 that will also record the authority to non-volatile solid state storage mapping so that the authority may be determined even if the assigned non-volatile solid state storage is unreachable. A duplicate or substitute authority 168 may be consulted if a specific authority 168 is unavailable in some embodiments.

[00077] With reference to Figure 2A and 2B, two of the many tasks of the CPU 156 on a storage node 150 are to break up write data, and reassemble read data. When the system has determined that data is to be written, the authority 168 for that data is located as above. When the segment ID for data is already determined the request to write is forwarded to the non-volatile solid state storage 152 currently determined to be the host of the authority 168 determined from the segment. The host CPU 156 of the storage node 150, on which the non-volatile solid state storage 152 and corresponding authority 168 reside, then breaks up or shards the data and transmits the data out to various non-volatile solid state storage 152. The transmitted data is written as a data stripe in accordance with an erasure coding scheme. In some embodiments, data is requested to be pulled, and in other embodiments, data is pushed. In reverse, when data is read, the authority 168 for the segment ID containing the data is located as described above. The host CPU 156 of the storage node 150 on which the non-volatile solid state storage 152 and corresponding authority 168 reside requests the data from

the non-volatile solid state storage and corresponding storage nodes pointed to by the authority. In some embodiments the data is read from flash storage as a data stripe. The host CPU 156 of storage node 150 then reassembles the read data, correcting any errors (if present) according to the appropriate erasure coding scheme, and forwards the reassembled data to the network. In further embodiments, some or all of these tasks can be handled in the non-volatile solid state storage 152. In some embodiments, the segment host requests the data be sent to storage node 150 by requesting pages from storage and then sending the data to the storage node making the original request.

[00078] In some systems, for example in UNIX-style file systems, data is handled with an index node or inode, which specifies a data structure that represents an object in a file system. The object could be a file or a directory, for example. Metadata may accompany the object, as attributes such as permission data and a creation timestamp, among other attributes. A segment number could be assigned to all or a portion of such an object in a file system. In other systems, data segments are handled with a segment number assigned elsewhere. For purposes of discussion, the unit of distribution is an entity, and an entity can be a file, a directory or a segment. That is, entities are units of data or metadata stored by a storage system. Entities are grouped into sets called authorities. Each authority has an authority owner, which is a storage node that has the exclusive right to update the entities in the authority. In other words, a storage node contains the authority, and that the authority, in turn, contains entities.

[00079] A segment is a logical container of data in accordance with some embodiments. A segment is an address space between medium address space and physical flash locations, i.e., the data segment number, are in this address space. Segments may also contain meta-data, which enable data redundancy to be restored (rewritten to different flash locations or devices) without the involvement of higher level software. In one embodiment, an internal format of a segment contains client data and medium mappings to determine the position of that data. Each data segment is protected, e.g., from memory and other failures, by breaking the segment into a number of data and parity shards, where applicable. The data and parity shards are distributed, i.e., striped, across non-volatile solid state storage 152 coupled to the host CPUs 156 (See Figures 2E and 2G) in accordance with an erasure coding scheme. Usage of the term segments refers to the container and its place in the address space of segments in some embodiments. Usage of the term stripe refers to the same set of shards as a segment and includes how the shards are distributed along with redundancy or parity information in accordance with some embodiments.

[00080] A series of address-space transformations takes place across an entire storage system. At the top are the directory entries (file names) which link to an inode. Inodes point into medium address space, where data is logically stored. Medium addresses may be mapped through a series of indirect mediums to spread the load of large files, or implement data services like deduplication or snapshots. Medium addresses may be mapped through a series of indirect mediums to spread the load of large files, or implement data services like deduplication or snapshots. Segment addresses are then translated into physical flash locations. Physical flash locations have an address range bounded by the amount of flash in the system in accordance with some embodiments. Medium addresses and segment addresses are logical containers, and in some embodiments use a 128 bit or larger identifier so as to be practically infinite, with a likelihood of reuse calculated as longer than the expected life of the system. Addresses from logical containers are allocated in a hierarchical fashion in some embodiments. Initially, each non-volatile solid state storage unit 152 may be assigned a range of address space. Within this assigned range, the non-volatile solid state storage 152 is able to allocate addresses without synchronization with other non-volatile solid state storage 152.

[00081] Data and metadata is stored by a set of underlying storage layouts that are optimized for varying workload patterns and storage devices. These layouts incorporate multiple redundancy schemes, compression formats and index algorithms. Some of these layouts store information about authorities and authority masters, while others store file metadata and file data. The redundancy schemes include error correction codes that tolerate corrupted bits within a single storage device (such as a NAND flash chip), erasure codes that tolerate the failure of multiple storage nodes, and replication schemes that tolerate data center or regional failures. In some embodiments, low density parity check ('LDPC') code is used within a single storage unit. Reed-Solomon encoding is used within a storage cluster, and mirroring is used within a storage grid in some embodiments. Metadata may be stored using an ordered log structured index (such as a Log Structured Merge Tree), and large data may not be stored in a log structured layout.

[00082] In order to maintain consistency across multiple copies of an entity, the storage nodes agree implicitly on two things through calculations: (1) the authority that contains the entity, and (2) the storage node that contains the authority. The assignment of entities to authorities can be done by pseudo randomly assigning entities to authorities, by splitting entities into ranges based upon an externally produced key, or by placing a single entity into each authority. Examples of pseudorandom schemes are linear hashing and the Replication

Under Scalable Hashing ('RUSH') family of hashes, including Controlled Replication Under Scalable Hashing ('CRUSH'). In some embodiments, pseudo-random assignment is utilized only for assigning authorities to nodes because the set of nodes can change. The set of authorities cannot change so any subjective function may be applied in these embodiments. Some placement schemes automatically place authorities on storage nodes, while other placement schemes rely on an explicit mapping of authorities to storage nodes. In some embodiments, a pseudorandom scheme is utilized to map from each authority to a set of candidate authority owners. A pseudorandom data distribution function related to CRUSH may assign authorities to storage nodes and create a list of where the authorities are assigned. Each storage node has a copy of the pseudorandom data distribution function, and can arrive at the same calculation for distributing, and later finding or locating an authority. Each of the pseudorandom schemes requires the reachable set of storage nodes as input in some embodiments in order to conclude the same target nodes. Once an entity has been placed in an authority, the entity may be stored on physical devices so that no expected failure will lead to unexpected data loss. In some embodiments, rebalancing algorithms attempt to store the copies of all entities within an authority in the same layout and on the same set of machines.

[00083] Examples of expected failures include device failures, stolen machines, datacenter fires, and regional disasters, such as nuclear or geological events. Different failures lead to different levels of acceptable data loss. In some embodiments, a stolen storage node impacts neither the security nor the reliability of the system, while depending on system configuration, a regional event could lead to no loss of data, a few seconds or minutes of lost updates, or even complete data loss.

[00084] In the embodiments, the placement of data for storage redundancy is independent of the placement of authorities for data consistency. In some embodiments, storage nodes that contain authorities do not contain any persistent storage. Instead, the storage nodes are connected to non-volatile solid state storage units that do not contain authorities. The communications interconnect between storage nodes and non-volatile solid state storage units consists of multiple communication technologies and has non-uniform performance and fault tolerance characteristics. In some embodiments, as mentioned above, non-volatile solid state storage units are connected to storage nodes via PCI express, storage nodes are connected together within a single chassis using Ethernet backplane, and chassis are connected together to form a storage cluster. Storage clusters are connected to clients using Ethernet or fiber channel in some embodiments. If multiple storage clusters are configured into a storage grid, the multiple storage clusters are connected using the Internet or other long-distance

networking links, such as a "metro scale" link or private link that does not traverse the internet.

[00085] Authority owners have the exclusive right to modify entities, to migrate entities from one non-volatile solid state storage unit to another non-volatile solid state storage unit, and to add and remove copies of entities. This allows for maintaining the redundancy of the underlying data. When an authority owner fails, is going to be decommissioned, or is overloaded, the authority is transferred to a new storage node. Transient failures make it non-trivial to ensure that all non-faulty machines agree upon the new authority location. The ambiguity that arises due to transient failures can be achieved automatically by a consensus protocol such as Paxos, hot-warm failover schemes, via manual intervention by a remote system administrator, or by a local hardware administrator (such as by physically removing the failed machine from the cluster, or pressing a button on the failed machine). In some embodiments, a consensus protocol is used, and failover is automatic. If too many failures or replication events occur in too short a time period, the system goes into a self-preservation mode and halts replication and data movement activities until an administrator intervenes in accordance with some embodiments.

[00086] As authorities are transferred between storage nodes and authority owners update entities in their authorities, the system transfers messages between the storage nodes and non-volatile solid state storage units. With regard to persistent messages, messages that have different purposes are of different types. Depending on the type of the message, the system maintains different ordering and durability guarantees. As the persistent messages are being processed, the messages are temporarily stored in multiple durable and non-durable storage hardware technologies. In some embodiments, messages are stored in RAM, NVRAM and on NAND flash devices, and a variety of protocols are used in order to make efficient use of each storage medium. Latency-sensitive client requests may be persisted in replicated NVRAM, and then later NAND, while background rebalancing operations are persisted directly to NAND.

[00087] Persistent messages are persistently stored prior to being transmitted. This allows the system to continue to serve client requests despite failures and component replacement. Although many hardware components contain unique identifiers that are visible to system administrators, manufacturer, hardware supply chain and ongoing monitoring quality control infrastructure, applications running on top of the infrastructure address virtualize addresses. These virtualized addresses do not change over the lifetime of the storage system, regardless of component failures and replacements. This allows each component of the storage system

to be replaced over time without reconfiguration or disruptions of client request processing, i.e., the system supports non-disruptive upgrades.

[00088] In some embodiments, the virtualized addresses are stored with sufficient redundancy. A continuous monitoring system correlates hardware and software status and the hardware identifiers. This allows detection and prediction of failures due to faulty components and manufacturing details. The monitoring system also enables the proactive transfer of authorities and entities away from impacted devices before failure occurs by removing the component from the critical path in some embodiments.

[00089] Figure 2C is a multiple level block diagram, showing contents of a storage node 150 and contents of a non-volatile solid state storage 152 of the storage node 150. Data is communicated to and from the storage node 150 by a network interface controller ('NIC') 202 in some embodiments. Each storage node 150 has a CPU 156, and one or more non-volatile solid state storage 152, as discussed above. Moving down one level in Figure 2C, each non-volatile solid state storage 152 has a relatively fast non-volatile solid state memory, such as nonvolatile random access memory ('NVRAM') 204, and flash memory 206. In some embodiments, NVRAM 204 may be a component that does not require program/erase cycles (DRAM, MRAM, PCM), and can be a memory that can support being written vastly more often than the memory is read from. Moving down another level in Figure 2C, the NVRAM 204 is implemented in one embodiment as high speed volatile memory, such as dynamic random access memory (DRAM) 216, backed up by energy reserve 218. Energy reserve 218 provides sufficient electrical power to keep the DRAM 216 powered long enough for contents to be transferred to the flash memory 206 in the event of power failure. In some embodiments, energy reserve 218 is a capacitor, super-capacitor, battery, or other device, that supplies a suitable supply of energy sufficient to enable the transfer of the contents of DRAM 216 to a stable storage medium in the case of power loss. The flash memory 206 is implemented as multiple flash dies 222, which may be referred to as packages of flash dies 222 or an array of flash dies 222. It should be appreciated that the flash dies 222 could be packaged in any number of ways, with a single die per package, multiple dies per package (i.e. multichip packages), in hybrid packages, as bare dies on a printed circuit board or other substrate, as encapsulated dies, etc. In the embodiment shown, the non-volatile solid state storage 152 has a controller 212 or other processor, and an input output (I/O) port 210 coupled to the controller 212. I/O port 210 is coupled to the CPU 156 and/or the network interface controller 202 of the flash storage node 150. Flash input output (I/O) port 220 is coupled to the flash dies 222, and a direct memory access unit (DMA) 214 is coupled to the

controller 212, the DRAM 216 and the flash dies 222. In the embodiment shown, the I/O port 210, controller 212, DMA unit 214 and flash I/O port 220 are implemented on a programmable logic device ('PLD') 208, e.g., a field programmable gate array (FPGA). In this embodiment, each flash die 222 has pages, organized as sixteen kB (kilobyte) pages 224, and a register 226 through which data can be written to or read from the flash die 222. In further embodiments, other types of solid-state memory are used in place of, or in addition to flash memory illustrated within flash die 222.

[00090] Storage clusters 161, in various embodiments as disclosed herein, can be contrasted with storage arrays in general. The storage nodes 150 are part of a collection that creates the storage cluster 161. Each storage node 150 owns a slice of data and computing required to provide the data. Multiple storage nodes 150 cooperate to store and retrieve the data. Storage memory or storage devices, as used in storage arrays in general, are less involved with processing and manipulating the data. Storage memory or storage devices in a storage array receive commands to read, write, or erase data. The storage memory or storage devices in a storage array are not aware of a larger system in which they are embedded, or what the data means. Storage memory or storage devices in storage arrays can include various types of storage memory, such as RAM, solid state drives, hard disk drives, etc. The storage units 152 described herein have multiple interfaces active simultaneously and serving multiple purposes. In some embodiments, some of the functionality of a storage node 150 is shifted into a storage unit 152, transforming the storage unit 152 into a combination of storage unit 152 and storage node 150. Placing computing (relative to storage data) into the storage unit 152 places this computing closer to the data itself. The various system embodiments have a hierarchy of storage node layers with different capabilities. By contrast, in a storage array, a controller owns and knows everything about all of the data that the controller manages in a shelf or storage devices. In a storage cluster 161, as described herein, multiple controllers in multiple storage units 152 and/or storage nodes 150 cooperate in various ways (e.g., for erasure coding, data sharding, metadata communication and redundancy, storage capacity expansion or contraction, data recovery, and so on).

[00091] Figure 2D shows a storage server environment, which uses embodiments of the storage nodes 150 and storage units 152 of Figures 2A-C. In this version, each storage unit 152 has a processor such as controller 212 (see Figure 2C), an FPGA (field programmable gate array), flash memory 206, and NVRAM 204 (which is super-capacitor backed DRAM 216, see Figures 2B and 2C) on a PCIe (peripheral component interconnect express) board in a chassis 138 (see Figure 2A). The storage unit 152 may be implemented as a single board

containing storage, and may be the largest tolerable failure domain inside the chassis. In some embodiments, up to two storage units 152 may fail and the device will continue with no data loss.

[00092] The physical storage is divided into named regions based on application usage in some embodiments. The NVRAM 204 is a contiguous block of reserved memory in the storage unit 152 DRAM 216, and is backed by NAND flash. NVRAM 204 is logically divided into multiple memory regions written for two as spool (e.g., spool_region). Space within the NVRAM 204 spools is managed by each authority 168 independently. Each device provides an amount of storage space to each authority 168. That authority 168 further manages lifetimes and allocations within that space. Examples of a spool include distributed transactions or notions. When the primary power to a storage unit 152 fails, onboard super-capacitors provide a short duration of power hold up. During this holdup interval, the contents of the NVRAM 204 are flushed to flash memory 206. On the next power-on, the contents of the NVRAM 204 are recovered from the flash memory 206.

[00093] As for the storage unit controller, the responsibility of the logical "controller" is distributed across each of the blades containing authorities 168. This distribution of logical control is shown in Figure 2D as a host controller 242, mid-tier controller 244 and storage unit controller(s) 246. Management of the control plane and the storage plane are treated independently, although parts may be physically co-located on the same blade. Each authority 168 effectively serves as an independent controller. Each authority 168 provides its own data and metadata structures, its own background workers, and maintains its own lifecycle.

[00094] Figure 2E is a blade 252 hardware block diagram, showing a control plane 254, compute and storage planes 256, 258, and authorities 168 interacting with underlying physical resources, using embodiments of the storage nodes 150 and storage units 152 of Figs. 2A-C in the storage server environment of Figure 2D. The control plane 254 is partitioned into a number of authorities 168 which can use the compute resources in the compute plane 256 to run on any of the blades 252. The storage plane 258 is partitioned into a set of devices, each of which provides access to flash 206 and NVRAM 204 resources. In one embodiment, the compute plane 256 may perform the operations of a storage array controller, as described herein, on one or more devices of the storage plane 258 (e.g., a storage array).

[00095] In the compute and storage planes 256, 258 of Figure 2E, the authorities 168 interact with the underlying physical resources (i.e., devices). From the point of view of an authority

168, its resources are striped over all of the physical devices. From the point of view of a device, it provides resources to all authorities 168, irrespective of where the authorities happen to run. Each authority 168 has allocated or has been allocated one or more partitions 260 of storage memory in the storage units 152, e.g. partitions 260 in flash memory 206 and NVRAM 204. Each authority 168 uses those allocated partitions 260 that belong to it, for writing or reading user data. Authorities can be associated with differing amounts of physical storage of the system. For example, one authority 168 could have a larger number of partitions 260 or larger sized partitions 260 in one or more storage units 152 than one or more other authorities 168.

[00096] Figure 2F depicts elasticity software layers in blades 252 of a storage cluster, in accordance with some embodiments. In the elasticity structure, elasticity software is symmetric, i.e., each blade's compute module 270 runs the three identical layers of processes depicted in Figure 2F. Storage managers 274 execute read and write requests from other blades 252 for data and metadata stored in local storage unit 152 NVRAM 204 and flash 206. Authorities 168 fulfill client requests by issuing the necessary reads and writes to the blades 252 on whose storage units 152 the corresponding data or metadata resides. Endpoints 272 parse client connection requests received from switch fabric 146 supervisory software, relay the client connection requests to the authorities 168 responsible for fulfillment, and relay the authorities' 168 responses to clients. The symmetric three-layer structure enables the storage system's high degree of concurrency. Elasticity scales out efficiently and reliably in these embodiments. In addition, elasticity implements a unique scale-out technique that balances work evenly across all resources regardless of client access pattern, and maximizes concurrency by eliminating much of the need for inter-blade coordination that typically occurs with conventional distributed locking.

[00097] Still referring to Figure 2F, authorities 168 running in the compute modules 270 of a blade 252 perform the internal operations required to fulfill client requests. One feature of elasticity is that authorities 168 are stateless, i.e., they cache active data and metadata in their own blades' 252 DRAMs for fast access, but the authorities store every update in their NVRAM 204 partitions on three separate blades 252 until the update has been written to flash 206. All the storage system writes to NVRAM 204 are in triplicate to partitions on three separate blades 252 in some embodiments. With triple-mirrored NVRAM 204 and persistent storage protected by parity and Reed-Solomon RAID checksums, the storage system can survive concurrent failure of two blades 252 with no loss of data, metadata, or access to either.

[00098] Because authorities 168 are stateless, they can migrate between blades 252. Each authority 168 has a unique identifier. NVRAM 204 and flash 206 partitions are associated with authorities' 168 identifiers, not with the blades 252 on which they are running in some. Thus, when an authority 168 migrates, the authority 168 continues to manage the same storage partitions from its new location. When a new blade 252 is installed in an embodiment of the storage cluster, the system automatically rebalances load by: partitioning the new blade's 252 storage for use by the system's authorities 168, migrating selected authorities 168 to the new blade 252, starting endpoints 272 on the new blade 252 and including them in the switch fabric's 146 client connection distribution algorithm.

[00099] From their new locations, migrated authorities 168 persist the contents of their NVRAM 204 partitions on flash 206, process read and write requests from other authorities 168, and fulfill the client requests that endpoints 272 direct to them. Similarly, if a blade 252 fails or is removed, the system redistributes its authorities 168 among the system's remaining blades 252. The redistributed authorities 168 continue to perform their original functions from their new locations.

[000100] Figure 2G depicts authorities 168 and storage resources in blades 252 of a storage cluster, in accordance with some embodiments. Each authority 168 is exclusively responsible for a partition of the flash 206 and NVRAM 204 on each blade 252. The authority 168 manages the content and integrity of its partitions independently of other authorities 168. Authorities 168 compress incoming data and preserve it temporarily in their NVRAM 204 partitions, and then consolidate, RAID-protect, and persist the data in segments of the storage in their flash 206 partitions. As the authorities 168 write data to flash 206, storage managers 274 perform the necessary flash translation to optimize write performance and maximize media longevity. In the background, authorities 168 "garbage collect," or reclaim space occupied by data that clients have made obsolete by overwriting the data. It should be appreciated that since authorities' 168 partitions are disjoint, there is no need for distributed locking to execute client and writes or to perform background functions.

[000101] The embodiments described herein may utilize various software, communication and/or networking protocols. In addition, the configuration of the hardware and/or software may be adjusted to accommodate various protocols. For example, the embodiments may utilize Active Directory, which is a database based system that provides authentication, directory, policy, and other services in a WINDOWS™ environment. In these embodiments, LDAP (Lightweight Directory Access Protocol) is one example application protocol for querying and modifying items in directory service providers such as Active Directory. In

some embodiments, a network lock manager ('NLM') is utilized as a facility that works in cooperation with the Network File System ('NFS') to provide a System V style of advisory file and record locking over a network. The Server Message Block ('SMB') protocol, one version of which is also known as Common Internet File System ('CIFS'), may be integrated with the storage systems discussed herein. SMP operates as an application-layer network protocol typically used for providing shared access to files, printers, and serial ports and miscellaneous communications between nodes on a network. SMB also provides an authenticated inter-process communication mechanism. AMAZON™ S3 (Simple Storage Service) is a web service offered by Amazon Web Services, and the systems described herein may interface with Amazon S3 through web services interfaces (REST (representational state transfer), SOAP (simple object access protocol), and BitTorrent). A RESTful API (application programming interface) breaks down a transaction to create a series of small modules. Each module addresses a particular underlying part of the transaction. The control or permissions provided with these embodiments, especially for object data, may include utilization of an access control list ('ACL'). The ACL is a list of permissions attached to an object and the ACL specifies which users or system processes are granted access to objects, as well as what operations are allowed on given objects. The systems may utilize Internet Protocol version 6 ('IPv6'), as well as IPv4, for the communications protocol that provides an identification and location system for computers on networks and routes traffic across the Internet. The routing of packets between networked systems may include Equal-cost multi-path routing ('ECMP'), which is a routing strategy where next-hop packet forwarding to a single destination can occur over multiple "best paths" which tie for top place in routing metric calculations. Multi-path routing can be used in conjunction with most routing protocols, because it is a per-hop decision limited to a single router. The software may support Multi-tenancy, which is an architecture in which a single instance of a software application serves multiple customers. Each customer may be referred to as a tenant. Tenants may be given the ability to customize some parts of the application, but may not customize the application's code, in some embodiments. The embodiments may maintain audit logs. An audit log is a document that records an event in a computing system. In addition to documenting what resources were accessed, audit log entries typically include destination and source addresses, a timestamp, and user login information for compliance with various regulations. The embodiments may support various key management policies, such as encryption key rotation. In addition, the system may support dynamic root passwords or some variation dynamically changing passwords.

[000102] Figure 3A sets forth a diagram of a storage system 306 that is coupled for data communications with a cloud services provider 302 in accordance with some embodiments of the present disclosure. Although depicted in less detail, the storage system 306 depicted in Figure 3A may be similar to the storage systems described above with reference to Figures 1A-1D and Figures 2A-2G. In some embodiments, the storage system 306 depicted in Figure 3A may be embodied as a storage system that includes imbalanced active/active controllers, as a storage system that includes balanced active/active controllers, as a storage system that includes active/active controllers where less than all of each controller's resources are utilized such that each controller has reserve resources that may be used to support failover, as a storage system that includes fully active/active controllers, as a storage system that includes dataset-segregated controllers, as a storage system that includes dual-layer architectures with front-end controllers and back-end integrated storage controllers, as a storage system that includes scale-out clusters of dual-controller arrays, as well as combinations of such embodiments.

[000103] In the example depicted in Figure 3A, the storage system 306 is coupled to the cloud services provider 302 via a data communications link 304. The data communications link 304 may be embodied as a dedicated data communications link, as a data communications pathway that is provided through the use of one or data communications networks such as a wide area network ('WAN') or local area network ('LAN'), or as some other mechanism capable of transporting digital information between the storage system 306 and the cloud services provider 302. Such a data communications link 304 may be fully wired, fully wireless, or some aggregation of wired and wireless data communications pathways. In such an example, digital information may be exchanged between the storage system 306 and the cloud services provider 302 via the data communications link 304 using one or more data communications protocols. For example, digital information may be exchanged between the storage system 306 and the cloud services provider 302 via the data communications link 304 using the handheld device transfer protocol ('HDTP'), hypertext transfer protocol ('HTTP'), internet protocol ('IP'), real-time transfer protocol ('RTP'), transmission control protocol ('TCP'), user datagram protocol ('UDP'), wireless application protocol ('WAP'), or other protocol.

[000104] The cloud services provider 302 depicted in Figure 3A may be embodied, for example, as a system and computing environment that provides services to users of the cloud services provider 302 through the sharing of computing resources via the data communications link 304. The cloud services provider 302 may provide on-demand access

to a shared pool of configurable computing resources such as computer networks, servers, storage, applications and services, and so on. The shared pool of configurable resources may be rapidly provisioned and released to a user of the cloud services provider 302 with minimal management effort. Generally, the user of the cloud services provider 302 is unaware of the exact computing resources utilized by the cloud services provider 302 to provide the services. Although in many cases such a cloud services provider 302 may be accessible via the Internet, readers of skill in the art will recognize that any system that abstracts the use of shared resources to provide services to a user through any data communications link may be considered a cloud services provider 302.

[000105] In the example depicted in Figure 3A, the cloud services provider 302 may be configured to provide a variety of services to the storage system 306 and users of the storage system 306 through the implementation of various service models. For example, the cloud services provider 302 may be configured to provide services to the storage system 306 and users of the storage system 306 through the implementation of an infrastructure as a service ('IaaS') service model where the cloud services provider 302 offers computing infrastructure such as virtual machines and other resources as a service to subscribers. In addition, the cloud services provider 302 may be configured to provide services to the storage system 306 and users of the storage system 306 through the implementation of a platform as a service ('PaaS') service model where the cloud services provider 302 offers a development environment to application developers. Such a development environment may include, for example, an operating system, programming-language execution environment, database, web server, or other components that may be utilized by application developers to develop and run software solutions on a cloud platform. Furthermore, the cloud services provider 302 may be configured to provide services to the storage system 306 and users of the storage system 306 through the implementation of a software as a service ('SaaS') service model where the cloud services provider 302 offers application software, databases, as well as the platforms that are used to run the applications to the storage system 306 and users of the storage system 306, providing the storage system 306 and users of the storage system 306 with on-demand software and eliminating the need to install and run the application on local computers, which may simplify maintenance and support of the application. The cloud services provider 302 may be further configured to provide services to the storage system 306 and users of the storage system 306 through the implementation of an authentication as a service ('AaaS') service model where the cloud services provider 302 offers authentication services that can be used to secure access to applications, data sources, or other resources. The cloud services

provider 302 may also be configured to provide services to the storage system 306 and users of the storage system 306 through the implementation of a storage as a service model where the cloud services provider 302 offers access to its storage infrastructure for use by the storage system 306 and users of the storage system 306. Readers will appreciate that the cloud services provider 302 may be configured to provide additional services to the storage system 306 and users of the storage system 306 through the implementation of additional service models, as the service models described above are included only for explanatory purposes and in no way represent a limitation of the services that may be offered by the cloud services provider 302 or a limitation as to the service models that may be implemented by the cloud services provider 302.

[000106] In the example depicted in Figure 3A, the cloud services provider 302 may be embodied, for example, as a private cloud, as a public cloud, or as a combination of a private cloud and public cloud. In an embodiment in which the cloud services provider 302 is embodied as a private cloud, the cloud services provider 302 may be dedicated to providing services to a single organization rather than providing services to multiple organizations. In an embodiment where the cloud services provider 302 is embodied as a public cloud, the cloud services provider 302 may provide services to multiple organizations. Public cloud and private cloud deployment models may differ and may come with various advantages and disadvantages. For example, because a public cloud deployment involves the sharing of a computing infrastructure across different organization, such a deployment may not be ideal for organizations with security concerns, mission-critical workloads, uptime requirements demands, and so on. While a private cloud deployment can address some of these issues, a private cloud deployment may require on-premises staff to manage the private cloud. In still alternative embodiments, the cloud services provider 302 may be embodied as a mix of a private and public cloud services with a hybrid cloud deployment.

[000107] Although not explicitly depicted in Figure 3A, readers will appreciate that additional hardware components and additional software components may be necessary to facilitate the delivery of cloud services to the storage system 306 and users of the storage system 306. For example, the storage system 306 may be coupled to (or even include) a cloud storage gateway. Such a cloud storage gateway may be embodied, for example, as hardware-based or software-based appliance that is located on premise with the storage system 306. Such a cloud storage gateway may operate as a bridge between local applications that are executing on the storage array 306 and remote, cloud-based storage that is utilized by the storage array 306. Through the use of a cloud storage gateway,

organizations may move primary iSCSI or NAS to the cloud services provider 302, thereby enabling the organization to save space on their on-premises storage systems. Such a cloud storage gateway may be configured to emulate a disk array, a block-based device, a file server, or other storage system that can translate the SCSI commands, file server commands, or other appropriate command into REST-space protocols that facilitate communications with the cloud services provider 302.

[000108] In order to enable the storage system 306 and users of the storage system 306 to make use of the services provided by the cloud services provider 302, a cloud migration process may take place during which data, applications, or other elements from an organization's local systems (or even from another cloud environment) are moved to the cloud services provider 302. In order to successfully migrate data, applications, or other elements to the cloud services provider's 302 environment, middleware such as a cloud migration tool may be utilized to bridge gaps between the cloud services provider's 302 environment and an organization's environment. Such cloud migration tools may also be configured to address potentially high network costs and long transfer times associated with migrating large volumes of data to the cloud services provider 302, as well as addressing security concerns associated with sensitive data to the cloud services provider 302 over data communications networks. In order to further enable the storage system 306 and users of the storage system 306 to make use of the services provided by the cloud services provider 302, a cloud orchestrator may also be used to arrange and coordinate automated tasks in pursuit of creating a consolidated process or workflow. Such a cloud orchestrator may perform tasks such as configuring various components, whether those components are cloud components or on-premises components, as well as managing the interconnections between such components. The cloud orchestrator can simplify the inter-component communication and connections to ensure that links are correctly configured and maintained.

[000109] In the example depicted in Figure 3A, and as described briefly above, the cloud services provider 302 may be configured to provide services to the storage system 306 and users of the storage system 306 through the usage of a SaaS service model where the cloud services provider 302 offers application software, databases, as well as the platforms that are used to run the applications to the storage system 306 and users of the storage system 306, providing the storage system 306 and users of the storage system 306 with on-demand software and eliminating the need to install and run the application on local computers, which may simplify maintenance and support of the application. Such applications may take many forms in accordance with various embodiments of the present disclosure. For example, the

cloud services provider 302 may be configured to provide access to data analytics applications to the storage system 306 and users of the storage system 306. Such data analytics applications may be configured, for example, to receive telemetry data phoned home by the storage system 306. Such telemetry data may describe various operating characteristics of the storage system 306 and may be analyzed, for example, to determine the health of the storage system 306, to identify workloads that are executing on the storage system 306, to predict when the storage system 306 will run out of various resources, to recommend configuration changes, hardware or software upgrades, workflow migrations, or other actions that may improve the operation of the storage system 306.

[000110] The cloud services provider 302 may also be configured to provide access to virtualized computing environments to the storage system 306 and users of the storage system 306. Such virtualized computing environments may be embodied, for example, as a virtual machine or other virtualized computer hardware platforms, virtual storage devices, virtualized computer network resources, and so on. Examples of such virtualized environments can include virtual machines that are created to emulate an actual computer, virtualized desktop environments that separate a logical desktop from a physical machine, virtualized file systems that allow uniform access to different types of concrete file systems, and many others.

[000111] For further explanation, Figure 3B sets forth a diagram of a storage system 306 in accordance with some embodiments of the present disclosure. Although depicted in less detail, the storage system 306 depicted in Figure 3B may be similar to the storage systems described above with reference to Figures 1A-1D and Figures 2A-2G as the storage system may include many of the components described above.

[000112] The storage system 306 depicted in Figure 3B may include storage resources 308, which may be embodied in many forms. For example, in some embodiments the storage resources 308 can include nano-RAM or another form of nonvolatile random access memory that utilizes carbon nanotubes deposited on a substrate. In some embodiments, the storage resources 308 may include 3D crosspoint non-volatile memory in which bit storage is based on a change of bulk resistance, in conjunction with a stackable cross-gridded data access array. In some embodiments, the storage resources 308 may include flash memory, including single-level cell ('SLC') NAND flash, multi-level cell ('MLC') NAND flash, triple-level cell ('TLC') NAND flash, quad-level cell ('QLC') NAND flash, and others. In some embodiments, the storage resources 308 may include non-volatile magnetoresistive random-access memory ('MRAM'), including spin transfer torque ('STT') MRAM, in which data is

stored through the use of magnetic storage elements. In some embodiments, the example storage resources 308 may include non-volatile phase-change memory ('PCM') that may have the ability to hold multiple bits in a single cell as cells can achieve a number of distinct intermediary states. In some embodiments, the storage resources 308 may include quantum memory that allows for the storage and retrieval of photonic quantum information. In some embodiments, the example storage resources 308 may include resistive random-access memory ('ReRAM') in which data is stored by changing the resistance across a dielectric solid-state material. In some embodiments, the storage resources 308 may include storage class memory ('SCM') in which solid-state nonvolatile memory may be manufactured at a high density using some combination of sub-lithographic patterning techniques, multiple bits per cell, multiple layers of devices, and so on. Readers will appreciate that other forms of computer memories and storage devices may be utilized by the storage systems described above, including DRAM, SRAM, EEPROM, universal memory, and many others. The storage resources 308 depicted in Figure 3A may be embodied in a variety of form factors, including but not limited to, dual in-line memory modules ('DIMMs'), non-volatile dual in-line memory modules ('NVDIMMs'), M.2, U.2, and others.

[000113] The storage resources 308 depicted in Figure 3A may include various forms of storage-class memory ('SCM'). SCM may effectively treat fast, non-volatile memory (e.g., NAND flash) as an extension of DRAM such that an entire dataset may be treated as an in-memory dataset that resides entirely in DRAM. SCM may include non-volatile media such as, for example, NAND flash. Such NAND flash may be accessed utilizing NVMe that can use the PCIe bus as its transport, providing for relatively low access latencies compared to older protocols. In fact, the network protocols used for SSDs in all-flash arrays can include NVMe using Ethernet (ROCE, NVME TCP), Fibre Channel (NVMe FC), InfiniBand (iWARP), and others that make it possible to treat fast, non-volatile memory as an extension of DRAM. In view of the fact that DRAM is often byte-addressable and fast, non-volatile memory such as NAND flash is block-addressable, a controller software/hardware stack may be needed to convert the block data to the bytes that are stored in the media. Examples of media and software that may be used as SCM can include, for example, 3D XPoint, Intel Memory Drive Technology, Samsung's Z-SSD, and others.

[000114] The example storage system 306 depicted in Figure 3B may implement a variety of storage architectures. For example, storage systems in accordance with some embodiments of the present disclosure may utilize block storage where data is stored in blocks, and each block essentially acts as an individual hard drive. Storage systems in accordance with some

embodiments of the present disclosure may utilize object storage, where data is managed as objects. Each object may include the data itself, a variable amount of metadata, and a globally unique identifier, where object storage can be implemented at multiple levels (e.g., device level, system level, interface level). Storage systems in accordance with some embodiments of the present disclosure utilize file storage in which data is stored in a hierarchical structure. Such data may be saved in files and folders, and presented to both the system storing it and the system retrieving it in the same format.

[000115] The example storage system 306 depicted in Figure 3B may be embodied as a storage system in which additional storage resources can be added through the use of a scale-up model, additional storage resources can be added through the use of a scale-out model, or through some combination thereof. In a scale-up model, additional storage may be added by adding additional storage devices. In a scale-out model, however, additional storage nodes may be added to a cluster of storage nodes, where such storage nodes can include additional processing resources, additional networking resources, and so on.

[000116] The storage system 306 depicted in Figure 3B also includes communications resources 310 that may be useful in facilitating data communications between components within the storage system 306, as well as data communications between the storage system 306 and computing devices that are outside of the storage system 306. The communications resources 310 may be configured to utilize a variety of different protocols and data communication fabrics to facilitate data communications between components within the storage systems as well as computing devices that are outside of the storage system. For example, the communications resources 310 can include fibre channel ('FC') technologies such as FC fabrics and FC protocols that can transport SCSI commands over FC networks. The communications resources 310 can also include FC over ethernet ('FCoE') technologies through which FC frames are encapsulated and transmitted over Ethernet networks. The communications resources 310 can also include InfiniBand ('IB') technologies in which a switched fabric topology is utilized to facilitate transmissions between channel adapters. The communications resources 310 can also include NVM Express ('NVMe') technologies and NVMe over fabrics ('NVMeoF') technologies through which non-volatile storage media attached via a PCI express ('PCIe') bus may be accessed. The communications resources 310 can also include mechanisms for accessing storage resources 308 within the storage system 306 utilizing serial attached SCSI ('SAS'), serial ATA ('SATA') bus interfaces for connecting storage resources 308 within the storage system 306 to host bus adapters within the storage system 306, internet small computer systems interface ('iSCSI') technologies to provide

block-level access to storage resources 308 within the storage system 306, and other communications resources that that may be useful in facilitating data communications between components within the storage system 306, as well as data communications between the storage system 306 and computing devices that are outside of the storage system 306.

[000117] The storage system 306 depicted in Figure 3B also includes processing resources 312 that may be useful in useful in executing computer program instructions and performing other computational tasks within the storage system 306. The processing resources 312 may include one or more application-specific integrated circuits ('ASICs') that are customized for some particular purpose as well as one or more central processing units ('CPUs'). The processing resources 312 may also include one or more digital signal processors ('DSPs'), one or more field-programmable gate arrays ('FPGAs'), one or more systems on a chip ('SoCs'), or other form of processing resources 312. The storage system 306 may utilize the storage resources 312 to perform a variety of tasks including, but not limited to, supporting the execution of software resources 314 that will be described in greater detail below.

[000118] The storage system 306 depicted in Figure 3B also includes software resources 314 that, when executed by processing resources 312 within the storage system 306, may perform various tasks. The software resources 314 may include, for example, one or more modules of computer program instructions that when executed by processing resources 312 within the storage system 306 are useful in carrying out various data protection techniques to preserve the integrity of data that is stored within the storage systems. Readers will appreciate that such data protection techniques may be carried out, for example, by system software executing on computer hardware within the storage system, by a cloud services provider, or in other ways. Such data protection techniques can include, for example, data archiving techniques that cause data that is no longer actively used to be moved to a separate storage device or separate storage system for long-term retention, data backup techniques through which data stored in the storage system may be copied and stored in a distinct location to avoid data loss in the event of equipment failure or some other form of catastrophe with the storage system, data replication techniques through which data stored in the storage system is replicated to another storage system such that the data may be accessible via multiple storage systems, data snapshotting techniques through which the state of data within the storage system is captured at various points in time, data and database cloning techniques through which duplicate copies of data and databases may be created, and other data protection techniques. Through the use of such data protection techniques, business continuity and

disaster recovery objectives may be met as a failure of the storage system may not result in the loss of data stored in the storage system.

[000119] The software resources 314 may also include software that is useful in implementing software-defined storage ('SDS'). In such an example, the software resources 314 may include one or more modules of computer program instructions that, when executed, are useful in policy-based provisioning and management of data storage that is independent of the underlying hardware. Such software resources 314 may be useful in implementing storage virtualization to separate the storage hardware from the software that manages the storage hardware.

[000120] The software resources 314 may also include software that is useful in facilitating and optimizing I/O operations that are directed to the storage resources 308 in the storage system 306. For example, the software resources 314 may include software modules that perform carry out various data reduction techniques such as, for example, data compression, data deduplication, and others. The software resources 314 may include software modules that intelligently group together I/O operations to facilitate better usage of the underlying storage resource 308, software modules that perform data migration operations to migrate from within a storage system, as well as software modules that perform other functions. Such software resources 314 may be embodied as one or more software containers or in many other ways.

[000121] Readers will appreciate that the presence of such software resources 314 may provide for an improved user experience of the storage system 306, an expansion of functionality supported by the storage system 306, and many other benefits. Consider the specific example of the software resources 314 carrying out data backup techniques through which data stored in the storage system may be copied and stored in a distinct location to avoid data loss in the event of equipment failure or some other form of catastrophe. In such an example, the systems described herein may more reliably (and with less burden placed on the user) perform backup operations relative to interactive backup management systems that require high degrees of user interactivity, offer less robust automation and feature sets, and so on.

[000122] For further explanation, Figure 3C sets forth an example of a cloud-based storage system 318 in accordance with some embodiments of the present disclosure. In the example depicted in Figure 3C, the cloud-based storage system 318 is created entirely in a cloud computing environment 316 such as, for example, Amazon Web Services ('AWS'), Microsoft Azure, Google Cloud Platform, IBM Cloud, Oracle Cloud, and others. The cloud-based

storage system 318 may be used to provide services similar to the services that may be provided by the storage systems described above. For example, the cloud-based storage system 318 may be used to provide block storage services to users of the cloud-based storage system 318, the cloud-based storage system 318 may be used to provide storage services to users of the cloud-based storage system 318 through the use of solid-state storage, and so on.

[000123] The cloud-based storage system 318 depicted in Figure 3C includes two cloud computing instances 320, 322 that each are used to support the execution of a storage controller application 324, 326. The cloud computing instances 320, 322 may be embodied, for example, as instances of cloud computing resources (e.g., virtual machines) that may be provided by the cloud computing environment 316 to support the execution of software applications such as the storage controller application 324, 326. In one embodiment, the cloud computing instances 320, 322 may be embodied as Amazon Elastic Compute Cloud ('EC2') instances. In such an example, an Amazon Machine Image ('AMI') that includes the storage controller application 324, 326 may be booted to create and configure a virtual machine that may execute the storage controller application 324, 326.

[000124] In the example method depicted in Figure 3C, the storage controller application 324, 326 may be embodied as a module of computer program instructions that, when executed, carries out various storage tasks. For example, the storage controller application 324, 326 may be embodied as a module of computer program instructions that, when executed, carries out the same tasks as the controllers 110A, 110B in Figure 1A described above such as writing data received from the users of the cloud-based storage system 318 to the cloud-based storage system 318, erasing data from the cloud-based storage system 318, retrieving data from the cloud-based storage system 318 and providing such data to users of the cloud-based storage system 318, monitoring and reporting of disk utilization and performance, performing redundancy operations, such as RAID or RAID-like data redundancy operations, compressing data, encrypting data, deduplicating data, and so forth. Readers will appreciate that because there are two cloud computing instances 320, 322 that each include the storage controller application 324, 326, in some embodiments one cloud computing instance 320 may operate as the primary controller as described above while the other cloud computing instance 322 may operate as the secondary controller as described above. In such an example, in order to save costs, the cloud computing instance 320 that operates as the primary controller may be deployed on a relatively high-performance and relatively expensive cloud computing instance while the cloud computing instance 322 that operates as the secondary controller may be deployed on a relatively low-performance and

relatively inexpensive cloud computing instance. Readers will appreciate that the storage controller application 324, 326 depicted in Figure 3C may include identical source code that is executed within different cloud computing instances 320, 322.

[000125] Consider an example in which the cloud computing environment 316 is embodied as AWS and the cloud computing instances are embodied as EC2 instances. In such an example, AWS offers many types of EC2 instances. For example, AWS offers a suite of general purpose EC2 instances that include varying levels of memory and processing power. In such an example, the cloud computing instance 320 that operates as the primary controller may be deployed on one of the instance types that has a relatively large amount of memory and processing power while the cloud computing instance 322 that operates as the secondary controller may be deployed on one of the instance types that has a relatively small amount of memory and processing power. In such an example, upon the occurrence of a failover event where the roles of primary and secondary are switched, a double failover may actually be carried out such that: 1) a first failover event where the cloud computing instance 322 that formerly operated as the secondary controller begins to operate as the primary controller, and 2) a third cloud computing instance (not shown) that is of an instance type that has a relatively large amount of memory and processing power is spun up with a copy of the storage controller application, where the third cloud computing instance begins operating as the primary controller while the cloud computing instance 322 that originally operated as the secondary controller begins operating as the secondary controller again. In such an example, the cloud computing instance 320 that formerly operated as the primary controller may be terminated. Readers will appreciate that in alternative embodiments, the cloud computing instance 320 that is operating as the secondary controller after the failover event may continue to operate as the secondary controller and the cloud computing instance 322 that operated as the primary controller after the occurrence of the failover event may be terminated once the primary role has been assumed by the third cloud computing instance (not shown).

[000126] Readers will appreciate that while the embodiments described above relate to embodiments where one cloud computing instance 320 operates as the primary controller and the second cloud computing instance 322 operates as the secondary controller, other embodiments are within the scope of the present disclosure. For example, each cloud computing instance 320, 322 may operate as a primary controller for some portion of the address space supported by the cloud-based storage system 318, each cloud computing instance 320, 322 may operate as a primary controller where the servicing of I/O operations

directed to the cloud-based storage system 318 are divided in some other way, and so on. In fact, in other embodiments where costs savings may be prioritized over performance demands, only a single cloud computing instance may exist that contains the storage controller application. In such an example, a controller failure may take more time to recover from as a new cloud computing instance that includes the storage controller application would need to be spun up rather than having an already created cloud computing instance take on the role of servicing I/O operations that would have otherwise been handled by the failed cloud computing instance.

[000127] The cloud-based storage system 318 depicted in Figure 3C includes cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338. The cloud computing instances 340a, 340b, 340n depicted in Figure 3C may be embodied, for example, as instances of cloud computing resources that may be provided by the cloud computing environment 316 to support the execution of software applications. The cloud computing instances 340a, 340b, 340n of Figure 3C may differ from the cloud computing instances 320, 322 described above as the cloud computing instances 340a, 340b, 340n of Figure 3C have local storage 330, 334, 338 resources whereas the cloud computing instances 320, 322 that support the execution of the storage controller application 324, 326 need not have local storage resources. The cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 may be embodied, for example, as EC2 M5 instances that include one or more SSDs, as EC2 R5 instances that include one or more SSDs, as EC2 I3 instances that include one or more SSDs, and so on. In some embodiments, the local storage 330, 334, 338 must be embodied as solid-state storage (e.g., SSDs) rather than storage that makes use of hard disk drives.

[000128] In the example depicted in Figure 3C, each of the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 can include a software daemon 328, 332, 336 that, when executed by a cloud computing instance 340a, 340b, 340n can present itself to the storage controller applications 324, 326 as if the cloud computing instance 340a, 340b, 340n were a physical storage device (e.g., one or more SSDs). In such an example, the software daemon 328, 332, 336 may include computer program instructions similar to those that would normally be contained on a storage device such that the storage controller applications 324, 326 can send and receive the same commands that a storage controller would send to storage devices. In such a way, the storage controller applications 324, 326 may include code that is identical to (or substantially identical to) the code that would be executed by the controllers in the storage systems described above. In these and similar embodiments, communications

between the storage controller applications 324, 326 and the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 may utilize iSCSI, NVMe over TCP, messaging, a custom protocol, or in some other mechanism.

[000129] In the example depicted in Figure 3C, each of the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 may also be coupled to block-storage 342, 344, 346 that is offered by the cloud computing environment 316. The block-storage 342, 344, 346 that is offered by the cloud computing environment 316 may be embodied, for example, as Amazon Elastic Block Store ('EBS') volumes. For example, a first EBS volume may be coupled to a first cloud computing instance 340a, a second EBS volume may be coupled to a second cloud computing instance 340b, and a third EBS volume may be coupled to a third cloud computing instance 340n. In such an example, the block-storage 342, 344, 346 that is offered by the cloud computing environment 316 may be utilized in a manner that is similar to how the NVRAM devices described above are utilized, as the software daemon 328, 332, 336 (or some other module) that is executing within a particular cloud comping instance 340a, 340b, 340n may, upon receiving a request to write data, initiate a write of the data to its attached EBS volume as well as a write of the data to its local storage 330, 334, 338 resources. In some alternative embodiments, data may only be written to the local storage 330, 334, 338 resources within a particular cloud comping instance 340a, 340b, 340n. In an alternative embodiment, rather than using the block-storage 342, 344, 346 that is offered by the cloud computing environment 316 as NVRAM, actual RAM on each of the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 may be used as NVRAM, thereby decreasing network utilization costs that would be associated with using an EBS volume as the NVRAM.

[000130] In the example depicted in Figure 3C, the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 may be utilized, by cloud computing instances 320, 322 that support the execution of the storage controller application 324, 326 to service I/O operations that are directed to the cloud-based storage system 318. Consider an example in which a first cloud computing instance 320 that is executing the storage controller application 324 is operating as the primary controller. In such an example, the first cloud computing instance 320 that is executing the storage controller application 324 may receive (directly or indirectly via the secondary controller) requests to write data to the cloud-based storage system 318 from users of the cloud-based storage system 318. In such an example, the first cloud computing instance 320 that is executing the storage controller application 324 may perform various tasks such as, for example, deduplicating the data contained in the request,

compressing the data contained in the request, determining where to the write the data contained in the request, and so on, before ultimately sending a request to write a deduplicated, encrypted, or otherwise possibly updated version of the data to one or more of the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338. Either cloud computing instance 320, 322, in some embodiments, may receive a request to read data from the cloud-based storage system 318 and may ultimately send a request to read data to one or more of the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338.

[000131] Readers will appreciate that when a request to write data is received by a particular cloud computing instance 340a, 340b, 340n with local storage 330, 334, 338, the software daemon 328, 332, 336 or some other module of computer program instructions that is executing on the particular cloud computing instance 340a, 340b, 340n may be configured to not only write the data to its own local storage 330, 334, 338 resources and any appropriate block-storage 342, 344, 346 that are offered by the cloud computing environment 316, but the software daemon 328, 332, 336 or some other module of computer program instructions that is executing on the particular cloud computing instance 340a, 340b, 340n may also be configured to write the data to cloud-based object storage 348 that is attached to the particular cloud computing instance 340a, 340b, 340n. The cloud-based object storage 348 that is attached to the particular cloud computing instance 340a, 340b, 340n may be embodied, for example, as Amazon Simple Storage Service ('S3') storage that is accessible by the particular cloud computing instance 340a, 340b, 340n. In other embodiments, the cloud computing instances 320, 322 that each include the storage controller application 324, 326 may initiate the storage of the data in the local storage 330, 334, 338 of the cloud computing instances 340a, 340b, 340n and the cloud-based object storage 348.

[000132] Readers will appreciate that, as described above, the cloud-based storage system 318 may be used to provide block storage services to users of the cloud-based storage system 318. While the local storage 330, 334, 338 resources and the block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n may support block-level access, the cloud-based object storage 348 that is attached to the particular cloud computing instance 340a, 340b, 340n supports only object-based access. In order to address this, the software daemon 328, 332, 336 or some other module of computer program instructions that is executing on the particular cloud computing instance 340a, 340b, 340n may be configured to take blocks of data, package those blocks into objects, and write the

objects to the cloud-based object storage 348 that is attached to the particular cloud computing instance 340a, 340b, 340n.

[000133] Consider an example in which data is written to the local storage 330, 334, 338 resources and the block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n in 1 MB blocks. In such an example, assume that a user of the cloud-based storage system 318 issues a request to write data that, after being compressed and deduplicated by the storage controller application 324, 326 results in the need to write 5 MB of data. In such an example, writing the data to the local storage 330, 334, 338 resources and the block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n is relatively straightforward as 5 blocks that are 1 MB in size are written to the local storage 330, 334, 338 resources and the block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n. In such an example, the software daemon 328, 332, 336 or some other module of computer program instructions that is executing on the particular cloud computing instance 340a, 340b, 340n may be configured to: 1) create a first object that includes the first 1 MB of data and write the first object to the cloud-based object storage 348, 2) create a second object that includes the second 1 MB of data and write the second object to the cloud-based object storage 348, 3) create a third object that includes the third 1 MB of data and write the third object to the cloud-based object storage 348, and so on. As such, in some embodiments, each object that is written to the cloud-based object storage 348 may be identical (or nearly identical) in size. Readers will appreciate that in such an example, metadata that is associated with the data itself may be included in each object (e.g., the first 1 MB of the object is data and the remaining portion is metadata associated with the data).

[000134] Readers will appreciate that the cloud-based object storage 348 may be incorporated into the cloud-based storage system 318 to increase the durability of the cloud-based storage system 318. Continuing with the example described above where the cloud computing instances 340a, 340b, 340n are EC2 instances, readers will understand that EC2 instances are only guaranteed to have a monthly uptime of 99.9% and data stored in the local instance store only persists during the lifetime of the EC2 instance. As such, relying on the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 as the only source of persistent data storage in the cloud-based storage system 318 may result in a relatively unreliable storage system. Likewise, EBS volumes are designed for 99.999% availability. As such, even relying on EBS as the persistent data store in the cloud-based storage system 318 may result in a storage system that is not sufficiently durable. Amazon

S3, however, is designed to provide 99.999999999% durability, meaning that a cloud-based storage system 318 that can incorporate S3 into its pool of storage is substantially more durable than various other options.

[000135] Readers will appreciate that while a cloud-based storage system 318 that can incorporate S3 into its pool of storage is substantially more durable than various other options, utilizing S3 as the primary pool of storage may result in storage system that has relatively slow response times and relatively long I/O latencies. As such, the cloud-based storage system 318 depicted in Figure 3C not only stores data in S3 but the cloud-based storage system 318 also stores data in local storage 330, 334, 338 resources and block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n, such that read operations can be serviced from local storage 330, 334, 338 resources and the block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n, thereby reducing read latency when users of the cloud-based storage system 318 attempt to read data from the cloud-based storage system 318.

[000136] In some embodiments, all data that is stored by the cloud-based storage system 318 may be stored in both: 1) the cloud-based object storage 348, and 2) at least one of the local storage 330, 334, 338 resources or block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n. In such embodiments, the local storage 330, 334, 338 resources and block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n may effectively operate as cache that generally includes all data that is also stored in S3, such that all reads of data may be serviced by the cloud computing instances 340a, 340b, 340n without requiring the cloud computing instances 340a, 340b, 340n to access the cloud-based object storage 348. Readers will appreciate that in other embodiments, however, all data that is stored by the cloud-based storage system 318 may be stored in the cloud-based object storage 348, but less than all data that is stored by the cloud-based storage system 318 may be stored in at least one of the local storage 330, 334, 338 resources or block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n. In such an example, various policies may be utilized to determine which subset of the data that is stored by the cloud-based storage system 318 should reside in both: 1) the cloud-based object storage 348, and 2) at least one of the local storage 330, 334, 338 resources or block-storage 342, 344, 346 resources that are utilized by the cloud computing instances 340a, 340b, 340n.

[000137] As described above, when the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 are embodied as EC2 instances, the cloud computing instances

340a, 340b, 340n with local storage 330, 334, 338 are only guaranteed to have a monthly uptime of 99.9% and data stored in the local instance store only persists during the lifetime of each cloud computing instance 340a, 340b, 340n with local storage 330, 334, 338. As such, one or more modules of computer program instructions that are executing within the cloud-based storage system 318 (e.g., a monitoring module that is executing on its own EC2 instance) may be designed to handle the failure of one or more of the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338. In such an example, the monitoring module may handle the failure of one or more of the cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 by creating one or more new cloud computing instances with local storage, retrieving data that was stored on the failed cloud computing instances 340a, 340b, 340n from the cloud-based object storage 348, and storing the data retrieved from the cloud-based object storage 348 in local storage on the newly created cloud computing instances. Readers will appreciate that many variants of this process may be implemented.

[000138] Consider an example in which all cloud computing instances 340a, 340b, 340n with local storage 330, 334, 338 failed. In such an example, the monitoring module may create new cloud computing instances with local storage, where high-bandwidth instances types are selected that allow for the maximum data transfer rates between the newly created high-bandwidth cloud computing instances with local storage and the cloud-based object storage 348. Readers will appreciate that instances types are selected that allow for the maximum data transfer rates between the new cloud computing instances and the cloud-based object storage 348 such that the new high-bandwidth cloud computing instances can be rehydrated with data from the cloud-based object storage 348 as quickly as possible. Once the new high-bandwidth cloud computing instances are rehydrated with data from the cloud-based object storage 348, less expensive lower-bandwidth cloud computing instances may be created, data may be migrated to the less expensive lower-bandwidth cloud computing instances, and the high-bandwidth cloud computing instances may be terminated.

[000139] Readers will appreciate that in some embodiments, the number of new cloud computing instances that are created may substantially exceed the number of cloud computing instances that are needed to locally store all of the data stored by the cloud-based storage system 318. The number of new cloud computing instances that are created may substantially exceed the number of cloud computing instances that are needed to locally store all of the data stored by the cloud-based storage system 318 in order to more rapidly pull data from the cloud-based object storage 348 and into the new cloud computing instances, as each

new cloud computing instance can (in parallel) retrieve some portion of the data stored by the cloud-based storage system 318. In such embodiments, once the data stored by the cloud-based storage system 318 has been pulled into the newly created cloud computing instances, the data may be consolidated within a subset of the newly created cloud computing instances and those newly created cloud computing instances that are excessive may be terminated.

[000140] Consider an example in which 1000 cloud computing instances are needed in order to locally store all valid data that users of the cloud-based storage system 318 have written to the cloud-based storage system 318. In such an example, assume that all 1,000 cloud computing instances fail. In such an example, the monitoring module may cause 100,000 cloud computing instances to be created, where each cloud computing instance is responsible for retrieving, from the cloud-based object storage 348, distinct 1/100,000th chunks of the valid data that users of the cloud-based storage system 318 have written to the cloud-based storage system 318 and locally storing the distinct chunk of the dataset that it retrieved. In such an example, because each of the 100,000 cloud computing instances can retrieve data from the cloud-based object storage 348 in parallel, the caching layer may be restored 100 times faster as compared to an embodiment where the monitoring module only create 1000 replacement cloud computing instances. In such an example, over time the data that is stored locally in the 100,000 could be consolidated into 1,000 cloud computing instances and the remaining 99,000 cloud computing instances could be terminated.

[000141] Readers will appreciate that various performance aspects of the cloud-based storage system 318 may be monitored (e.g., by a monitoring module that is executing in an EC2 instance) such that the cloud-based storage system 318 can be scaled-up or scaled-out as needed. Consider an example in which the monitoring module monitors the performance of the could-based storage system 318 via communications with one or more of the cloud computing instances 320, 322 that each are used to support the execution of a storage controller application 324, 326, via monitoring communications between cloud computing instances 320, 322, 340a, 340b, 340n, via monitoring communications between cloud computing instances 320, 322, 340a, 340b, 340n and the cloud-based object storage 348, or in some other way. In such an example, assume that the monitoring module determines that the cloud computing instances 320, 322 that are used to support the execution of a storage controller application 324, 326 are undersized and not sufficiently servicing the I/O requests that are issued by users of the cloud-based storage system 318. In such an example, the monitoring module may create a new, more powerful cloud computing instance (e.g., a cloud computing instance of a type that includes more processing power, more memory, etc...) that

includes the storage controller application such that the new, more powerful cloud computing instance can begin operating as the primary controller. Likewise, if the monitoring module determines that the cloud computing instances 320, 322 that are used to support the execution of a storage controller application 324, 326 are oversized and that cost savings could be gained by switching to a smaller, less powerful cloud computing instance, the monitoring module may create a new, less powerful (and less expensive) cloud computing instance that includes the storage controller application such that the new, less powerful cloud computing instance can begin operating as the primary controller.

[000142] Consider, as an additional example of dynamically sizing the cloud-based storage system 318, an example in which the monitoring module determines that the utilization of the local storage that is collectively provided by the cloud computing instances 340a, 340b, 340n has reached a predetermined utilization threshold (e.g., 95%). In such an example, the monitoring module may create additional cloud computing instances with local storage to expand the pool of local storage that is offered by the cloud computing instances. Alternatively, the monitoring module may create one or more new cloud computing instances that have larger amounts of local storage than the already existing cloud computing instances 340a, 340b, 340n, such that data stored in an already existing cloud computing instance 340a, 340b, 340n can be migrated to the one or more new cloud computing instances and the already existing cloud computing instance 340a, 340b, 340n can be terminated, thereby expanding the pool of local storage that is offered by the cloud computing instances. Likewise, if the pool of local storage that is offered by the cloud computing instances is unnecessarily large, data can be consolidated and some cloud computing instances can be terminated.

[000143] Readers will appreciate that the cloud-based storage system 318 may be sized up and down automatically by a monitoring module applying a predetermined set of rules that may be relatively simple of relatively complicated. In fact, the monitoring module may not only take into account the current state of the cloud-based storage system 318, but the monitoring module may also apply predictive policies that are based on, for example, observed behavior (e.g., every night from 10 PM until 6 AM usage of the storage system is relatively light), predetermined fingerprints (e.g., every time a virtual desktop infrastructure adds 100 virtual desktops, the number of IOPS directed to the storage system increase by X), and so on. In such an example, the dynamic scaling of the cloud-based storage system 318 may be based on current performance metrics, predicted workloads, and many other factors, including combinations thereof.

[000144] Readers will further appreciate that because the cloud-based storage system 318 may be dynamically scaled, the cloud-based storage system 318 may even operate in a way that is more dynamic. Consider the example of garbage collection. In a traditional storage system, the amount of storage is fixed. As such, at some point the storage system may be forced to perform garbage collection as the amount of available storage has become so constrained that the storage system is on the verge of running out of storage. In contrast, the cloud-based storage system 318 described here can always 'add' additional storage (e.g., by adding more cloud computing instances with local storage). Because the cloud-based storage system 318 described here can always 'add' additional storage, the cloud-based storage system 318 can make more intelligent decisions regarding when to perform garbage collection. For example, the cloud-based storage system 318 may implement a policy that garbage collection only be performed when the number of IOPS being serviced by the cloud-based storage system 318 falls below a certain level. In some embodiments, other system-level functions (e.g., deduplication, compression) may also be turned off and on in response to system load, given that the size of the cloud-based storage system 318 is not constrained in the same way that traditional storage systems are constrained.

[000145] Readers will appreciate that embodiments of the present disclosure resolve an issue with block-storage services offered by some cloud computing environments as some cloud computing environments only allow for one cloud computing instance to connect to a block-storage volume at a single time. For example, in Amazon AWS, only a single EC2 instance may be connected to an EBS volume. Through the use of EC2 instances with local storage, embodiments of the present disclosure can offer multi-connect capabilities where multiple EC2 instances can connect to another EC2 instance with local storage ('a drive instance'). In such embodiments, the drive instances may include software executing within the drive instance that allows the drive instance to support I/O directed to a particular volume from each connected EC2 instance. As such, some embodiments of the present disclosure may be embodied as multi-connect block storage services that may not include all of the components depicted in Figure 3C.

[000146] In some embodiments, especially in embodiments where the cloud-based object storage 348 resources are embodied as Amazon S3, the cloud-based storage system 318 may include one or more modules (e.g., a module of computer program instructions executing on an EC2 instance) that are configured to ensure that when the local storage of a particular cloud computing instance is rehydrated with data from S3, the appropriate data is actually in S3. This issue arises largely because S3 implements an eventual consistency model where,

when overwriting an existing object, reads of the object will eventually (but not necessarily immediately) become consistent and will eventually (but not necessarily immediately) return the overwritten version of the object. To address this issue, in some embodiments of the present disclosure, objects in S3 are never overwritten. Instead, a traditional 'overwrite' would result in the creation of the new object (that includes the updated version of the data) and the eventual deletion of the old object (that includes the previous version of the data).

[000147] In some embodiments of the present disclosure, as part of an attempt to never (or almost never) overwrite an object, when data is written to S3 the resultant object may be tagged with a sequence number. In some embodiments, these sequence numbers may be persisted elsewhere (e.g., in a database) such that at any point in time, the sequence number associated with the most up-to-date version of some piece of data can be known. In such a way, a determination can be made as to whether S3 has the most recent version of some piece of data by merely reading the sequence number associated with an object - and without actually reading the data from S3. The ability to make this determination may be particularly important when a cloud computing instance with local storage crashes, as it would be undesirable to rehydrate the local storage of a replacement cloud computing instance with out-of-date data. In fact, because the cloud-based storage system 318 does not need to access the data to verify its validity, the data can stay encrypted and access charges can be avoided.

[000148] The storage systems described above may carry out intelligent data backup techniques through which data stored in the storage system may be copied and stored in a distinct location to avoid data loss in the event of equipment failure or some other form of catastrophe. For example, the storage systems described above may be configured to examine each backup to avoid restoring the storage system to an undesirable state. Consider an example in which malware infects the storage system. In such an example, the storage system may include software resources 314 that can scan each backup to identify backups that were captured before the malware infected the storage system and those backups that were captured after the malware infected the storage system. In such an example, the storage system may restore itself from a backup that does not include the malware – or at least not restore the portions of a backup that contained the malware. In such an example, the storage system may include software resources 314 that can scan each backup to identify the presences of malware (or a virus, or some other undesirable), for example, by identifying write operations that were serviced by the storage system and originated from a network subnet that is suspected to have delivered the malware, by identifying write operations that were serviced by the storage system and originated from a user that is suspected to have

delivered the malware, by identifying write operations that were serviced by the storage system and examining the content of the write operation against fingerprints of the malware, and in many other ways.

[000149] Readers will further appreciate that the backups (often in the form of one or more snapshots) may also be utilized to perform rapid recovery of the storage system. Consider an example in which the storage system is infected with ransomware that locks users out of the storage system. In such an example, software resources 314 within the storage system may be configured to detect the presence of ransomware and may be further configured to restore the storage system to a point-in-time, using the retained backups, prior to the point-in-time at which the ransomware infected the storage system. In such an example, the presence of ransomware may be explicitly detected through the use of software tools utilized by the system, through the use of a key (e.g., a USB drive) that is inserted into the storage system, or in a similar way. Likewise, the presence of ransomware may be inferred in response to system activity meeting a predetermined fingerprint such as, for example, no reads or writes coming into the system for a predetermined period of time.

[000150] Readers will appreciate that the various components depicted in Figure 3B may be grouped into one or more optimized computing packages as converged infrastructures. Such converged infrastructures may include pools of computers, storage and networking resources that can be shared by multiple applications and managed in a collective manner using policy-driven processes. Such converged infrastructures may minimize compatibility issues between various components within the storage system 306 while also reducing various costs associated with the establishment and operation of the storage system 306. Such converged infrastructures may be implemented with a converged infrastructure reference architecture, with standalone appliances, with a software driven hyper-converged approach (e.g., hyper-converged infrastructures), or in other ways.

[000151] Readers will appreciate that the storage system 306 depicted in Figure 3B may be useful for supporting various types of software applications. For example, the storage system 306 may be useful in supporting artificial intelligence ('AI') applications, database applications, DevOps projects, electronic design automation tools, event-driven software applications, high performance computing applications, simulation applications, high-speed data capture and analysis applications, machine learning applications, media production applications, media serving applications, picture archiving and communication systems ('PACS') applications, software development applications, virtual reality applications,

augmented reality applications, and many other types of applications by providing storage resources to such applications.

[000152] The storage systems described above may operate to support a wide variety of applications. In view of the fact that the storage systems include compute resources, storage resources, and a wide variety of other resources, the storage systems may be well suited to support applications that are resource intensive such as, for example, AI applications. Such AI applications may enable devices to perceive their environment and take actions that maximize their chance of success at some goal. Examples of such AI applications can include IBM Watson, Microsoft Oxford, Google DeepMind, Baidu Minwa, and others. The storage systems described above may also be well suited to support other types of applications that are resource intensive such as, for example, machine learning applications. Machine learning applications may perform various types of data analysis to automate analytical model building. Using algorithms that iteratively learn from data, machine learning applications can enable computers to learn without being explicitly programmed. One particular area of machine learning is referred to as reinforcement learning, which involves taking suitable actions to maximize reward in a particular situation. Reinforcement learning may be employed to find the best possible behavior or path that a particular software application or machine should take in a specific situation. Reinforcement learning differs from other areas of machine learning (e.g., supervised learning, unsupervised learning) in that correct input/output pairs need not be presented for reinforcement learning and sub-optimal actions need not be explicitly corrected.

[000153] In addition to the resources already described, the storage systems described above may also include graphics processing units ('GPUs'), occasionally referred to as visual processing unit ('VPUs'). Such GPUs may be embodied as specialized electronic circuits that rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. Such GPUs may be included within any of the computing devices that are part of the storage systems described above, including as one of many individually scalable components of a storage system, where other examples of individually scalable components of such storage system can include storage components, memory components, compute components (e.g., CPUs, FPGAs, ASICs), networking components, software components, and others. In addition to GPUs, the storage systems described above may also include neural network processors ('NNPs') for use in various aspects of neural network processing. Such NNPs may be used in place of (or in addition to) GPUs and may be also be independently scalable.

[000154] As described above, the storage systems described herein may be configured to support artificial intelligence applications, machine learning applications, big data analytics applications, and many other types of applications. The rapid growth in these sort of applications is being driven by three technologies: deep learning (DL), GPU processors, and Big Data. Deep learning is a computing model that makes use of massively parallel neural networks inspired by the human brain. Instead of experts handcrafting software, a deep learning model writes its own software by learning from lots of examples. A GPU is a modern processor with thousands of cores, well-suited to run algorithms that loosely represent the parallel nature of the human brain.

[000155] Advances in deep neural networks have ignited a new wave of algorithms and tools for data scientists to tap into their data with artificial intelligence (AI). With improved algorithms, larger data sets, and various frameworks (including open-source software libraries for machine learning across a range of tasks), data scientists are tackling new use cases like autonomous driving vehicles, natural language processing and understanding, computer vision, machine reasoning, strong AI, and many others. Applications of such techniques may include: machine and vehicular object detection, identification and avoidance; visual recognition, classification and tagging; algorithmic financial trading strategy performance management; simultaneous localization and mapping; predictive maintenance of high-value machinery; prevention against cyber security threats, expertise automation; image recognition and classification; question answering; robotics; text analytics (extraction, classification) and text generation and translation; and many others. Applications of AI techniques has materialized in a wide array of products include, for example, Amazon Echo's speech recognition technology that allows users to talk to their machines, Google Translate™ which allows for machine-based language translation, Spotify's Discover Weekly that provides recommendations on new songs and artists that a user may like based on the user's usage and traffic analysis, Quill's text generation offering that takes structured data and turns it into narrative stories, Chatbots that provide real-time, contextually specific answers to questions in a dialog format, and many others. Furthermore, AI may impact a wide variety of industries and sectors. For example, AI solutions may be used in healthcare to take clinical notes, patient files, research data, and other inputs to generate potential treatment options for doctors to explore. Likewise, AI solutions may be used by retailers to personalize consumer recommendations based on a person's digital footprint of behaviors, profile data, or other data.

[000156] Training deep neural networks, however, requires both high quality input data and large amounts of computation. GPUs are massively parallel processors capable of operating on large amounts of data simultaneously. When combined into a multi-GPU cluster, a high throughput pipeline may be required to feed input data from storage to the compute engines. Deep learning is more than just constructing and training models. There also exists an entire data pipeline that must be designed for the scale, iteration, and experimentation necessary for a data science team to succeed.

[000157] Data is the heart of modern AI and deep learning algorithms. Before training can begin, one problem that must be addressed revolves around collecting the labeled data that is crucial for training an accurate AI model. A full scale AI deployment may be required to continuously collect, clean, transform, label, and store large amounts of data. Adding additional high quality data points directly translates to more accurate models and better insights. Data samples may undergo a series of processing steps including, but not limited to: 1) ingesting the data from an external source into the training system and storing the data in raw form, 2) cleaning and transforming the data in a format convenient for training, including linking data samples to the appropriate label, 3) exploring parameters and models, quickly testing with a smaller dataset, and iterating to converge on the most promising models to push into the production cluster, 4) executing training phases to select random batches of input data, including both new and older samples, and feeding those into production GPU servers for computation to update model parameters, and 5) evaluating including using a holdback portion of the data not used in training in order to evaluate model accuracy on the holdout data. This lifecycle may apply for any type of parallelized machine learning, not just neural networks or deep learning. For example, standard machine learning frameworks may rely on CPUs instead of GPUs but the data ingest and training workflows may be the same. Readers will appreciate that a single shared storage data hub creates a coordination point throughout the lifecycle without the need for extra data copies among the ingest, preprocessing, and training stages. Rarely is the ingested data used for only one purpose, and shared storage gives the flexibility to train multiple different models or apply traditional analytics to the data.

[000158] Readers will appreciate that each stage in the AI data pipeline may have varying requirements from the data hub (e.g., the storage system or collection of storage systems). Scale-out storage systems must deliver uncompromising performance for all manner of access types and patterns – from small, metadata-heavy to large files, from random to sequential access patterns, and from low to high concurrency. The storage systems described

above may serve as an ideal AI data hub as the systems may service unstructured workloads. In the first stage, data is ideally ingested and stored on to the same data hub that following stages will use, in order to avoid excess data copying. The next two steps can be done on a standard compute server that optionally includes a GPU, and then in the fourth and last stage, full training production jobs are run on powerful GPU-accelerated servers. Often, there is a production pipeline alongside an experimental pipeline operating on the same dataset. Further, the GPU-accelerated servers can be used independently for different models or joined together to train on one larger model, even spanning multiple systems for distributed training. If the shared storage tier is slow, then data must be copied to local storage for each phase, resulting in wasted time staging data onto different servers. The ideal data hub for the AI training pipeline delivers performance similar to data stored locally on the server node while also having the simplicity and performance to enable all pipeline stages to operate concurrently.

[000159] A data scientist works to improve the usefulness of the trained model through a wide variety of approaches: more data, better data, smarter training, and deeper models. In many cases, there will be teams of data scientists sharing the same datasets and working in parallel to produce new and improved training models. Often, there is a team of data scientists working within these phases concurrently on the same shared datasets. Multiple, concurrent workloads of data processing, experimentation, and full-scale training layer the demands of multiple access patterns on the storage tier. In other words, storage cannot just satisfy large file reads, but must contend with a mix of large and small file reads and writes. Finally, with multiple data scientists exploring datasets and models, it may be critical to store data in its native format to provide flexibility for each user to transform, clean, and use the data in a unique way. The storage systems described above may provide a natural shared storage home for the dataset, with data protection redundancy (e.g., by using RAID6) and the performance necessary to be a common access point for multiple developers and multiple experiments. Using the storage systems described above may avoid the need to carefully copy subsets of the data for local work, saving both engineering and GPU-accelerated servers use time. These copies become a constant and growing tax as the raw data set and desired transformations constantly update and change.

[000160] Readers will appreciate that a fundamental reason why deep learning has seen a surge in success is the continued improvement of models with larger data set sizes. In contrast, classical machine learning algorithms, like logistic regression, stop improving in accuracy at smaller data set sizes. As such, the separation of compute resources and storage

resources may also allow independent scaling of each tier, avoiding many of the complexities inherent in managing both together. As the data set size grows or new data sets are considered, a scale out storage system must be able to expand easily. Similarly, if more concurrent training is required, additional GPUs or other compute resources can be added without concern for their internal storage. Furthermore, the storage systems described above may make building, operating, and growing an AI system easier due to the random read bandwidth provided by the storage systems, the ability to of the storage systems to randomly read small files (50KB) high rates (meaning that no extra effort is required to aggregate individual data points to make larger, storage-friendly files), the ability of the storage systems to scale capacity and performance as either the dataset grows or the throughput requirements grow, the ability of the storage systems to support files or objects, the ability of the storage systems to tune performance for large or small files (i.e., no need for the user to provision filesystems), the ability of the storage systems to support non-disruptive upgrades of hardware and software even during production model training, and for many other reasons.

[000161] Small file performance of the storage tier may be critical as many types of inputs, including text, audio, or images will be natively stored as small files. If the storage tier does not handle small files well, an extra step will be required to pre-process and group samples into larger files. Storage, built on top of spinning disks, that relies on SSD as a caching tier, may fall short of the performance needed. Because training with random input batches results in more accurate models, the entire data set must be accessible with full performance. SSD caches only provide high performance for a small subset of the data and will be ineffective at hiding the latency of spinning drives.

[000162] Although the preceding paragraphs discuss deep learning applications, readers will appreciate that the storage systems described herein may also be part of a distributed deep learning ('DDL') platform to support the execution of DDL algorithms. Distributed deep learning may can be used to significantly accelerate deep learning with distributed computing on GPUs (or other form of accelerator or computer program instruction executor), such that parallelism can be achieved. In addition, the output of training machine learning and deep learning models, such as a fully trained machine learning model, may be used for a variety of purposes and in conjunction with other tools. For example, trained machine learning models may be used in conjunction with tools like Core ML to integrate a broad variety of machine learning model types into an application. In fact, trained models may be run through Core ML converter tools and inserted into a custom application that can be deployed on compatible devices. The storage systems described above may also be paired with other technologies

such as TensorFlow, an open-source software library for dataflow programming across a range of tasks that may be used for machine learning applications such as neural networks, to facilitate the development of such machine learning models, applications, and so on.

[000163] Readers will further appreciate that the systems described above may be deployed in a variety of ways to support the democratization of AI, as AI becomes more available for mass consumption. The democratization of AI may include, for example, the ability to offer AI as a Platform-as-a-Service, the growth of Artificial general intelligence offerings, the proliferation of Autonomous level 4 and Autonomous level 5 vehicles, the availability of autonomous mobile robots, the development of conversational AI platforms, and many others. For example, the systems described above may be deployed in cloud environments, edge environments, or other environments that are useful in supporting the democratization of AI. As part of the democratization of AI, a movement may occur from narrow AI that consists of highly scoped machine learning solutions that target a particular task to artificial general intelligence where the use of machine learning is expanded to handle a broad range of use cases that could essentially perform any intelligent task that a human could perform and could learn dynamically, much like a human.

[000164] The storage systems described above may also be used in a neuromorphic computing environment. Neuromorphic computing is a form of computing that mimics brain cells. To support neuromorphic computing, an architecture of interconnected "neurons" replace traditional computing models with low-powered signals that go directly between neurons for more efficient computation. Neuromorphic computing may make use of very-large-scale integration (VLSI) systems containing electronic analog circuits to mimic neuro-biological architectures present in the nervous system, as well as analog, digital, mixed-mode analog/digital VLSI, and software systems that implement models of neural systems for perception, motor control, or multisensory integration.

[000165] Readers will appreciate that the storage systems described above may be configured to support the storage or use of (among other types of data) blockchains. Such blockchains may be embodied as a continuously growing list of records, called blocks, which are linked and secured using cryptography. Each block in a blockchain may contain a hash pointer as a link to a previous block, a timestamp, transaction data, and so on. Blockchains may be designed to be resistant to modification of the data and can serve as an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way. This makes blockchains potentially suitable for the recording of events, medical records, and other records management activities, such as identity management,

transaction processing, and others. In addition to supporting the storage and use of blockchain technologies, the storage systems described above may also support the storage and use of derivative items such as, for example, open source blockchains and related tools that are part of the IBM™ Hyperledger project, permissioned blockchains in which a certain number of trusted parties are allowed to access the block chain, blockchain products that enable developers to build their own distributed ledger projects, and others. Readers will appreciate that blockchain technologies may impact a wide variety of industries and sectors. For example, blockchain technologies may be used in real estate transactions as blockchain based contracts whose use can eliminate the need for 3rd parties and enable self-executing actions when conditions are met. Likewise, universal health records can be created by aggregating and placing a person's health history onto a blockchain ledger for any healthcare provider, or permissioned health care providers, to access and update.

[000166] Readers will appreciate that the usage of blockchains is not limited to financial transactions, contracts, and the like. In fact, blockchains may be leveraged to enable the decentralized aggregation, ordering, timestamping and archiving of any type of information, including structured data, correspondence, documentation, or other data. Through the usage of blockchains, participants can provably and permanently agree on exactly what data was entered, when and by whom, without relying on a trusted intermediary. For example, SAP's recently launched blockchain platform, which supports MultiChain and Hyperledger Fabric, targets a broad range of supply chain and other non-financial applications.

[000167] One way to use a blockchain for recording data is to embed each piece of data directly inside a transaction. Every blockchain transaction may be digitally signed by one or more parties, replicated to a plurality of nodes, ordered and timestamped by the chain's consensus algorithm, and stored permanently in a tamper-proof way. Any data within the transaction will therefore be stored identically but independently by every node, along with a proof of who wrote it and when. The chain's users are able to retrieve this information at any future time. This type of storage may be referred to as on-chain storage. On-chain storage may not be particularly practical, however, when attempting to store a very large dataset. As such, in accordance with embodiments of the present disclosure, blockchains and the storage systems described herein may be leveraged to support on-chain storage of data as well as off-chain storage of data.

[000168] Off-chain storage of data can be implemented in a variety of ways and can occur when the data itself is not stored within the blockchain. For example, in one embodiment, a hash function may be utilized and the data itself may be fed into the hash function to generate

a hash value. In such an example, the hashes of large pieces of data may be embedded within transactions, instead of the data itself. Each hash may serve as a commitment to its input data, with the data itself being stored outside of the blockchain. Readers will appreciate that any blockchain participant that needs an off-chain piece of data cannot reproduce the data from its hash, but if the data can be retrieved in some other way, then the on-chain hash serves to confirm who created it and when. Just like regular on-chain data, the hash may be embedded inside a digitally signed transaction, which was included in the chain by consensus.

[000169] Readers will appreciate that, in other embodiments, alternatives to blockchains may be used to facilitate the decentralized storage of information. For example, one alternative to a blockchain that may be used is a blockweave. While conventional blockchains store every transaction to achieve validation, a blockweave permits secure decentralization without the usage of the entire chain, thereby enabling low cost on-chain storage of data. Such blockweaves may utilize a consensus mechanism that is based on proof of access (PoA) and proof of work (PoW). While typical PoW systems only depend on the previous block in order to generate each successive block, the PoA algorithm may incorporate data from a randomly chosen previous block. Combined with the blockweave data structure, miners do not need to store all blocks (forming a blockchain), but rather can store any previous blocks forming a weave of blocks (a blockweave). This enables increased levels of scalability, speed and low-cost and reduces the cost of data storage in part because miners need not store all blocks, thereby resulting in a substantial reduction in the amount of electricity that is consumed during the mining process because, as the network expands, electricity consumption decreases because a blockweave demands less and less hashing power for consensus as data is added to the system. Furthermore, blockweaves may be deployed on a decentralized storage network in which incentives are created to encourage rapid data sharing. Such decentralized storage networks may also make use of blockshadowing techniques, where nodes only send a minimal block "shadow" to other nodes that allows peers to reconstruct a full block, instead of transmitting the full block itself.

[000170] The storage systems described above may, either alone or in combination with other computing devices, be used to support in-memory computing applications. In memory computing involves the storage of information in RAM that is distributed across a cluster of computers. In-memory computing helps business customers, including retailers, banks and utilities, to quickly detect patterns, analyze massive data volumes on the fly, and perform their operations quickly. Readers will appreciate that the storage systems described above,

especially those that are configurable with customizable amounts of processing resources, storage resources, and memory resources (e.g., those systems in which blades that contain configurable amounts of each type of resource), may be configured in a way so as to provide an infrastructure that can support in-memory computing. Likewise, the storage systems described above may include component parts (e.g., NVDIMMs, 3D crosspoint storage that provide fast random access memory that is persistent) that can actually provide for an improved in-memory computing environment as compared to in-memory computing environments that rely on RAM distributed across dedicated servers.

[000171] In some embodiments, the storage systems described above may be configured to operate as a hybrid in-memory computing environment that includes a universal interface to all storage media (e.g., RAM, flash storage, 3D crosspoint storage). In such embodiments, users may have no knowledge regarding the details of where their data is stored but they can still use the same full, unified API to address data. In such embodiments, the storage system may (in the background) move data to the fastest layer available – including intelligently placing the data in dependence upon various characteristics of the data or in dependence upon some other heuristic. In such an example, the storage systems may even make use of existing products such as Apache Ignite and GridGain to move data between the various storage layers, or the storage systems may make use of custom software to move data between the various storage layers. The storage systems described herein may implement various optimizations to improve the performance of in-memory computing such as, for example, having computations occur as close to the data as possible.

[000172] Readers will further appreciate that in some embodiments, the storage systems described above may be paired with other resources to support the applications described above. For example, one infrastructure could include primary compute in the form of servers and workstations which specialize in using General-purpose computing on graphics processing units ('GPGPU') to accelerate deep learning applications that are interconnected into a computation engine to train parameters for deep neural networks. Each system may have Ethernet external connectivity, InfiniBand external connectivity, some other form of external connectivity, or some combination thereof. In such an example, the GPUs can be grouped for a single large training or used independently to train multiple models. The infrastructure could also include a storage system such as those described above to provide, for example, a scale-out all-flash file or object store through which data can be accessed via high-performance protocols such as NFS, S3, and so on. The infrastructure can also include, for example, redundant top-of-rack Ethernet switches connected to storage and compute via

ports in MLAG port channels for redundancy. The infrastructure could also include additional compute in the form of whitebox servers, optionally with GPUs, for data ingestion, pre-processing, and model debugging. Readers will appreciate that additional infrastructures are also be possible.

[000173] Readers will appreciate that the systems described above may be better suited for the applications described above relative to other systems that may include, for example, a distributed direct-attached storage (DDAS) solution deployed in server nodes. Such DDAS solutions may be built for handling large, less sequential accesses but may be less able to handle small, random accesses. Readers will further appreciate that the storage systems described above may be utilized to provide a platform for the applications described above that is preferable to the utilization of cloud-based resources as the storage systems may be included in an on-site or in-house infrastructure that is more secure, more locally and internally managed, more robust in feature sets and performance, or otherwise preferable to the utilization of cloud-based resources as part of a platform to support the applications described above. For example, services built on platforms such as IBM's Watson may require a business enterprise to distribute individual user information, such as financial transaction information or identifiable patient records, to other institutions. As such, cloud-based offerings of AI as a service may be less desirable than internally managed and offered AI as a service that is supported by storage systems such as the storage systems described above, for a wide array of technical reasons as well as for various business reasons.

[000174] Readers will appreciate that the storage systems described above, either alone or in coordination with other computing machinery may be configured to support other AI related tools. For example, the storage systems may make use of tools like ONXX or other open neural network exchange formats that make it easier to transfer models written in different AI frameworks. Likewise, the storage systems may be configured to support tools like Amazon's Gluon that allow developers to prototype, build, and train deep learning models. In fact, the storage systems described above may be part of a larger platform, such as IBM™ Cloud Private for Data, that includes integrated data science, data engineering and application building services. Such platforms may seamlessly collect, organize, secure, and analyze data across an enterprise, as well as simplify hybrid data management, unified data governance and integration, data science and business analytics with a single solution.

[000175] Readers will further appreciate that the storage systems described above may also be deployed as an edge solution. Such an edge solution may be in place to optimize cloud computing systems by performing data processing at the edge of the network, near the source

of the data. Edge computing can push applications, data and computing power (i.e., services) away from centralized points to the logical extremes of a network. Through the use of edge solutions such as the storage systems described above, computational tasks may be performed using the compute resources provided by such storage systems, data may be storage using the storage resources of the storage system, and cloud-based services may be accessed through the use of various resources of the storage system (including networking resources). By performing computational tasks on the edge solution, storing data on the edge solution, and generally making use of the edge solution, the consumption of expensive cloud-based resources may be avoided and, in fact, performance improvements may be experienced relative to a heavier reliance on cloud-based resources.

[000176] While many tasks may benefit from the utilization of an edge solution, some particular uses may be especially suited for deployment in such an environment. For example, devices like drones, autonomous cars, robots, and others may require extremely rapid processing — so fast, in fact, that sending data up to a cloud environment and back to receive data processing support may simply be too slow. Likewise, machines like locomotives and gas turbines that generate large amounts of information through the use of a wide array of data-generating sensors may benefit from the rapid data processing capabilities of an edge solution. As an additional example, some IoT devices such as connected video cameras may not be well-suited for the utilization of cloud-based resources as it may be impractical (not only from a privacy perspective, security perspective, or a financial perspective) to send the data to the cloud simply because of the pure volume of data that is involved. As such, many tasks that really on data processing, storage, or communications may be better suited by platforms that include edge solutions such as the storage systems described above.

[000177] Consider a specific example of inventory management in a warehouse, distribution center, or similar location. A large inventory, warehousing, shipping, order-fulfillment, manufacturing or other operation has a large amount of inventory on inventory shelves, and high resolution digital cameras that produce a firehose of large data. All of this data may be taken into an image processing system, which may reduce the amount of data to a firehose of small data. All of the small data may be stored on-premises in storage. The on-premises storage, at the edge of the facility, may be coupled to the cloud, for external reports, real-time control and cloud storage. Inventory management may be performed with the results of the image processing, so that inventory can be tracked on the shelves and restocked, moved, shipped, modified with new products, or discontinued/obsolescent products deleted, etc. The

above scenario is a prime candidate for an embodiment of the configurable processing and storage systems described above. A combination of compute-only blades and offload blades suited for the image processing, perhaps with deep learning on offload-FPGA or offload-custom blade(s) could take in the firehose of large data from all of the digital cameras, and produce the firehose of small data. All of the small data could then be stored by storage nodes, operating with storage units in whichever combination of types of storage blades best handles the data flow. This is an example of storage and function acceleration and integration. Depending on external communication needs with the cloud, and external processing in the cloud, and depending on reliability of network connections and cloud resources, the system could be sized for storage and compute management with bursty workloads and variable conductivity reliability. Also, depending on other inventory management aspects, the system could be configured for scheduling and resource management in a hybrid edge/cloud environment.

[000178] The storage systems described above may alone, or in combination with other computing resources, serves as a network edge platform that combines compute resources, storage resources, networking resources, cloud technologies and network virtualization technologies, and so on. As part of the network, the edge may take on characteristics similar to other network facilities, from the customer premise and backhaul aggregation facilities to Points of Presence (PoPs) and regional data centers. Readers will appreciate that network workloads, such as Virtual Network Functions (VNFs) and others, will reside on the network edge platform. Enabled by a combination of containers and virtual machines, the network edge platform may rely on controllers and schedulers that are no longer geographically co-located with the data processing resources. The functions, as microservices, may split into control planes, user and data planes, or even state machines, allowing for independent optimization and scaling techniques to be applied. Such user and data planes may be enabled through increased accelerators, both those residing in server platforms, such as FPGAs and Smart NICs, and through SDN-enabled merchant silicon and programmable ASICs.

[000179] The storage systems described above may also be optimized for use in big data analytics. Big data analytics may be generally described as the process of examining large and varied data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. Big data analytics applications enable data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by

conventional business intelligence (BI) and analytics programs. As part of that process, semi-structured and unstructured data such as, for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile-phone call-detail records, IoT sensor data, and other data may be converted to a structured form. Big data analytics is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analyses powered by high-performance analytics systems.

[000180] The storage systems described above may also support (including implementing as a system interface) applications that perform tasks in response to human speech. For example, the storage systems may support the execution intelligent personal assistant applications such as, for example, Amazon's Alexa, Apple Siri, Google Voice, Samsung Bixby, Microsoft Cortana, and others. While the examples described in the previous sentence make use of voice as input, the storage systems described above may also support chatbots, talkbots, chatterbots, or artificial conversational entities or other applications that are configured to conduct a conversation via auditory or textual methods. Likewise, the storage system may actually execute such an application to enable a user such as a system administrator to interact with the storage system via speech. Such applications are generally capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic, and other real time information, such as news, although in embodiments in accordance with the present disclosure, such applications may be utilized as interfaces to various system management operations.

[000181] The storage systems described above may also implement AI platforms for delivering on the vision of self-driving storage. Such AI platforms may be configured to deliver global predictive intelligence by collecting and analyzing large amounts of storage system telemetry data points to enable effortless management, analytics and support. In fact, such storage systems may be capable of predicting both capacity and performance, as well as generating intelligent advice on workload deployment, interaction and optimization. Such AI platforms may be configured to scan all incoming storage system telemetry data against a library of issue fingerprints to predict and resolve incidents in real-time, before they impact customer environments, and captures hundreds of variables related to performance that are used to forecast performance load.

[000182] The storage systems described above may support the serialized or simultaneous execution artificial intelligence applications, machine learning applications, data analytics applications, data transformations, and other tasks that collectively may form an AI ladder.

Such an AI ladder may effectively be formed by combining such elements to form a complete data science pipeline, where exist dependencies between elements of the AI ladder. For example, AI may require that some form of machine learning has taken place, machine learning may require that some form of analytics has taken place, analytics may require that some form of data and information architecting has taken place, and so on. As such, each element may be viewed as a rung in an AI ladder that collectively can form a complete and sophisticated AI solution.

[000183] The storage systems described above may also, either alone or in combination with other computing environments, be used to deliver an AI everywhere experience where AI permeates wide and expansive aspects of business and life. For example, AI may play an important role in the delivery of deep learning solutions, deep reinforcement learning solutions, artificial general intelligence solutions, autonomous vehicles, cognitive computing solutions, commercial UAVs or drones, conversational user interfaces, enterprise taxonomies, ontology management solutions, machine learning solutions, smart dust, smart robots, smart workplaces, and many others. The storage systems described above may also, either alone or in combination with other computing environments, be used to deliver a wide range of transparently immersive experiences where technology can introduce transparency between people, businesses, and things. Such transparently immersive experiences may be delivered as augmented reality technologies, connected homes, virtual reality technologies, brain–computer interfaces, human augmentation technologies, nanotube electronics, volumetric displays, 4D printing technologies, or others. The storage systems described above may also, either alone or in combination with other computing environments, be used to support a wide variety of digital platforms. Such digital platforms can include, for example, 5G wireless systems and platforms, digital twin platforms, edge computing platforms, IoT platforms, quantum computing platforms, serverless PaaS, software-defined security, neuromorphic computing platforms, and so on.

[000184] Readers will appreciate that some transparently immersive experiences may involve the use of digital twins of various "things" such as people, places, processes, systems, and so on. Such digital twins and other immersive technologies can alter the way that humans interact with technology, as conversational platforms, augmented reality, virtual reality and mixed reality provide a more natural and immersive interaction with the digital world. In fact, digital twins may be linked with the real-world, perhaps even in real-time, to understand the state of a thing or system, respond to changes, and so on. Because digital twins consolidate massive amounts of information on individual assets and groups of assets (even

possibly providing control of those assets), digital twins may communicate with each other to digital factory models of multiple linked digital twins.

[000185] The storage systems described above may also be part of a multi-cloud environment in which multiple cloud computing and storage services are deployed in a single heterogeneous architecture. In order to facilitate the operation of such a multi-cloud environment, DevOps tools may be deployed to enable orchestration across clouds. Likewise, continuous development and continuous integration tools may be deployed to standardize processes around continuous integration and delivery, new feature rollout and provisioning cloud workloads. By standardizing these processes, a multi-cloud strategy may be implemented that enables the utilization of the best provider for each workload. Furthermore, application monitoring and visibility tools may be deployed to move application workloads around different clouds, identify performance issues, and perform other tasks. In addition, security and compliance tools may be deployed for to ensure compliance with security requirements, government regulations, and so on. Such a multi-cloud environment may also include tools for application delivery and smart workload management to ensure efficient application delivery and help direct workloads across the distributed and heterogeneous infrastructure, as well as tools that ease the deployment and maintenance of packaged and custom applications in the cloud and enable portability amongst clouds. The multi-cloud environment may similarly include tools for data portability.

[000186] The storage systems described above may be used as a part of a platform to enable the use of crypto-anchors that may be used to authenticate a product's origins and contents to ensure that it matches a blockchain record associated with the product. Such crypto-anchors may take many forms including, for example, as edible ink, as a mobile sensor, as a microchip, and others. Similarly, as part of a suite of tools to secure data stored on the storage system, the storage systems described above may implement various encryption technologies and schemes, including lattice cryptography. Lattice cryptography can involve constructions of cryptographic primitives that involve lattices, either in the construction itself or in the security proof. Unlike public-key schemes such as the RSA, Diffie-Hellman or Elliptic-Curve cryptosystems, which are easily attacked by a quantum computer, some lattice-based constructions appear to be resistant to attack by both classical and quantum computers.

[000187] A quantum computer is a device that performs quantum computing. Quantum computing is computing using quantum-mechanical phenomena, such as superposition and entanglement. Quantum computers differ from traditional computers that are based on

transistors, as such traditional computers require that data be encoded into binary digits (bits), each of which is always in one of two definite states (0 or 1). In contrast to traditional computers, quantum computers use quantum bits, which can be in superpositions of states. A quantum computer maintains a sequence of qubits, where a single qubit can represent a one, a zero, or any quantum superposition of those two qubit states. A pair of qubits can be in any quantum superposition of 4 states, and three qubits in any superposition of 8 states. A quantum computer with *n* qubits can generally be in an arbitrary superposition of up to 2^*n* different states simultaneously, whereas a traditional computer can only be in one of these states at any one time. A quantum Turing machine is a theoretical model of such a computer.

[000188] The storage systems described above may also be paired with FPGA-accelerated servers as part of a larger AI or ML infrastructure. Such FPGA-accelerated servers may reside near (e.g., in the same data center) the storage systems described above or even incorporated into an appliance that includes one or more storage systems, one or more FPGA-accelerated servers, networking infrastructure that supports communications between the one or more storage systems and the one or more FPGA-accelerated servers, as well as other hardware and software components. Alternatively, FPGA-accelerated servers may reside within a cloud computing environment that may be used to perform compute-related tasks for AI and ML jobs. Any of the embodiments described above may be used to collectively serve as a FPGA-based AI or ML platform. Readers will appreciate that, in some embodiments of the FPGA-based AI or ML platform, the FPGAs that are contained within the FPGA-accelerated servers may be reconfigured for different types of ML models (e.g., LSTMs, CNNs, GRUs). The ability to reconfigure the FPGAs that are contained within the FPGA-accelerated servers may enable the acceleration of a ML or AI application based on the most optimal numerical precision and memory model being used. Readers will appreciate that by treating the collection of FPGA-accelerated servers as a pool of FPGAs, any CPU in the data center may utilize the pool of FPGAs as a shared hardware microservice, rather than limiting a server to dedicated accelerators plugged into it.

[000189] The FPGA-accelerated servers and the GPU-accelerated servers described above may implement a model of computing where, rather than keeping a small amount of data in a CPU and running a long stream of instructions over it as occurred in more traditional computing models, the machine learning model and parameters are pinned into the high-bandwidth on-chip memory with lots of data streaming though the high-bandwidth on-chip memory. FPGAs may even be more efficient than GPUs for this computing model, as the

FPGAs can be programmed with only the instructions needed to run this kind of computing model.

[000190] The storage systems described above may be configured to provide parallel storage, for example, through the use of a parallel file system such as BeeGFS. Such parallel files systems may include a distributed metadata architecture. For example, the parallel file system may include a plurality of metadata servers across which metadata is distributed, as well as components that include services for clients and storage servers. Through the use of a parallel file system, file contents may be distributed over a plurality of storage servers using striping and metadata may be distributed over a plurality of metadata servers on a directory level, with each server storing a part of the complete file system tree. Readers will appreciate that in some embodiments, the storage servers and metadata servers may run in userspace on top of an existing local file system. Furthermore, dedicated hardware is not required for client services, the metadata servers, or the hardware servers as metadata servers, storage servers, and even the client services may be run on the same machines.

[000191] Readers will appreciate that, in part due to the emergence of many of the technologies discussed above including mobile devices, cloud services, social networks, big data analytics, and so on, an information technology platform may be needed to integrate all of these technologies and drive new business opportunities by quickly delivering revenue-generating products, services, and experiences – rather than merely providing the technology to automate internal business processes. Information technology organizations may need to balance resources and investments needed to keep core legacy systems up and running while also integrating technologies to build an information technology platform that can provide the speed and flexibility in areas such as, for example, exploiting big data, managing unstructured data, and working with cloud applications and services. One possible embodiment of such an information technology platform is a composable infrastructure that includes fluid resource pools, such as many of the systems described above that, can meet the changing needs of applications by allowing for the composition and recomposition of blocks of disaggregated compute, storage, and fabric infrastructure. Such a composable infrastructure can also include a single management interface to eliminate complexity and a unified API to discover, search, inventory, configure, provision, update, and diagnose the composable infrastructure.

[000192] The systems described above can support the execution of a wide array of software applications. Such software applications can be deployed in a variety of ways, including container-based deployment models. Containerized applications may be managed using a

variety of tools. For example, containerized applications may be managed using Docker Swarm, a clustering and scheduling tool for Docker containers that enables IT administrators and developers to establish and manage a cluster of Docker nodes as a single virtual system. Likewise, containerized applications may be managed through the use of Kubernetes, a container-orchestration system for automating deployment, scaling and management of containerized applications. Kubernetes may execute on top of operating systems such as, for example, Red Hat Enterprise Linux, Ubuntu Server, SUSE Linux Enterprise Servers, and others. In such examples, a master node may assign tasks to worker/minion nodes. Kubernetes can include a set of components (e.g., kubelet, kube-proxy, cAdvisor) that manage individual nodes as a well as a set of components (e.g., etcd, API server, Scheduler, Control Manager) that form a control plane. Various controllers (e.g., Replication Controller, DaemonSet Controller) can drive the state of a Kubernetes cluster by managing a set of pods that includes one or more containers that are deployed on a single node. Containerized applications may be used to facilitate a serverless, cloud native computing deployment and management model for software applications. In support of a serverless, cloud native computing deployment and management model for software applications, containers may be used as part of an event handling mechanisms (e.g., AWS Lambdas) such that various events cause a containerized application to be spun up to operate as an event handler.

[000193] The systems described above may be deployed in a variety of ways, including being deployed in ways that support fifth generation ('5G') networks. 5G networks may support substantially faster data communications than previous generations of mobile communications networks and, as a consequence may lead to the disaggregation of data and computing resources as modern massive data centers may become less prominent and may be replaced, for example, by more-local, micro data centers that are close to the mobile-network towers. The systems described above may be included in such local, micro data centers and may be part of or paired to multi-access edge computing ('MEC') systems. Such MEC systems may enable cloud computing capabilities and an IT service environment at the edge of the cellular network. By running applications and performing related processing tasks closer to the cellular customer, network congestion may be reduced and applications may perform better. MEC technology is designed to be implemented at the cellular base stations or other edge nodes, and enables flexible and rapid deployment of new applications and services for customers. MEC may also allow cellular operators to open their radio access network ('RAN') to authorized third-parties, such as application developers and content provider. Furthermore, edge computing and micro data centers may substantially reduce the

cost of smartphones that work with the 5G network because customer may not need devices with such intensive processing power and the expensive requisite components.

[000194] Readers will appreciate that 5G networks may generate more data than previous network generations, especially in view of the fact that the high network bandwidth offered by 5G networks may cause the 5G networks to handle amounts and types of data (e.g., sensor data from self-driving cars, data generated by AR/VR technologies) that weren't as feasible for previous generation networks. In such examples, the scalability offered by the systems described above may be very valuable as the amount of data increases, adoption of emerging technologies increase, and so on.

[000195] For further explanation, Figure 3D illustrates an exemplary computing device 350 that may be specifically configured to perform one or more of the processes described herein. As shown in Figure 3D, computing device 350 may include a communication interface 352, a processor 354, a storage device 356, and an input/output ("I/O") module 358 communicatively connected one to another via a communication infrastructure 360. While an exemplary computing device 350 is shown in Figure 3D, the components illustrated in Figure 3D are not intended to be limiting. Additional or alternative components may be used in other embodiments. Components of computing device 350 shown in Figure 3D will now be described in additional detail.

[000196] Communication interface 352 may be configured to communicate with one or more computing devices. Examples of communication interface 352 include, without limitation, a wired network interface (such as a network interface card), a wireless network interface (such as a wireless network interface card), a modem, an audio/video connection, and any other suitable interface.

[000197] Processor 354 generally represents any type or form of processing unit capable of processing data and/or interpreting, executing, and/or directing execution of one or more of the instructions, processes, and/or operations described herein. Processor 354 may perform operations by executing computer-executable instructions 362 (e.g., an application, software, code, and/or other executable data instance) stored in storage device 356.

[000198] Storage device 356 may include one or more data storage media, devices, or configurations and may employ any type, form, and combination of data storage media and/or device. For example, storage device 356 may include, but is not limited to, any combination of the non-volatile media and/or volatile media described herein. Electronic data, including data described herein, may be temporarily and/or permanently stored in storage device 356. For example, data representative of computer-executable instructions

362 configured to direct processor 354 to perform any of the operations described herein may be stored within storage device 356. In some examples, data may be arranged in one or more databases residing within storage device 356.

[000199] I/O module 358 may include one or more I/O modules configured to receive user input and provide user output. I/O module 358 may include any hardware, firmware, software, or combination thereof supportive of input and output capabilities. For example, I/O module 358 may include hardware and/or software for capturing user input, including, but not limited to, a keyboard or keypad, a touchscreen component (e.g., touchscreen display), a receiver (e.g., an RF or infrared receiver), motion sensors, and/or one or more input buttons.

[000200] I/O module 358 may include one or more devices for presenting output to a user, including, but not limited to, a graphics engine, a display (e.g., a display screen), one or more output drivers (e.g., display drivers), one or more audio speakers, and one or more audio drivers. In certain embodiments, I/O module 358 is configured to provide graphical data to a display for presentation to a user. The graphical data may be representative of one or more graphical user interfaces and/or any other graphical content as may serve a particular implementation. In some examples, any of the systems, computing devices, and/or other components described herein may be implemented by computing device 350.

[000201] For further explanation, Figure 4A sets forth a block diagram illustrating a plurality of storage systems (402, 404, 406) that support a pod according to some embodiments of the present disclosure. Although depicted in less detail, the storage systems (402, 404, 406) depicted in Figure 4A may be similar to the storage systems described above with reference to Figures 1A-1D, Figures 2A-2G, Figures 3A-3B, or any combination thereof. In fact, the storage systems (402, 404, 406) depicted in Figure 4A may include the same, fewer, or additional components as the storage systems described above.

[000202] In the example depicted in Figure 4A, each of the storage systems (402, 404, 406) is depicted as having at least one computer processor (408, 410, 412), computer memory (414, 416, 418), and computer storage (420, 422, 424). Although in some embodiments the computer memory (414, 416, 418) and the computer storage (420, 422, 424) may be part of the same hardware devices, in other embodiments the computer memory (414, 416, 418) and the computer storage (420, 422, 424) may be part of different hardware devices. The distinction between the computer memory (414, 416, 418) and the computer storage (420, 422, 424) in this particular example may be that the computer memory (414, 416, 418) is physically proximate to the computer processors (408, 410, 412) and may store computer

program instructions that are executed by the computer processors (408, 410, 412), while the computer storage (420, 422, 424) is embodied as non-volatile storage for storing user data, metadata describing the user data, and so on. Referring to the example above in Figure 1A, for example, the computer processors (408, 410, 412) and computer memory (414, 416, 418) for a particular storage system (402, 404, 406) may reside within one of more of the controllers (110A-110D) while the attached storage devices (171A-171F) may serve as the computer storage (420, 422, 424) within a particular storage system (402, 404, 406).

[000203] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may attach to one or more pods (430, 432) according to some embodiments of the present disclosure. Each of the pods (430, 432) depicted in Figure 4A can include a dataset (426, 428). For example, a first pod (430) that three storage systems (402, 404, 406) have attached to includes a first dataset (426) while a second pod (432) that two storage systems (404, 406) have attached to includes a second dataset (428). In such an example, when a particular storage system attaches to a pod, the pod's dataset is copied to the particular storage system and then kept up to date as the dataset is modified. Storage systems can be removed from a pod, resulting in the dataset being no longer kept up to date on the removed storage system. In the example depicted in Figure 4A, any storage system which is active for a pod (it is an up-to-date, operating, non-faulted member of a non-faulted pod) can receive and process requests to modify or read the pod's dataset.

[000204] In the example depicted in Figure 4A, each pod (430, 432) may also include a set of managed objects and management operations, as well as a set of access operations to modify or read the dataset (426, 428) that is associated with the particular pod (430, 432). In such an example, the management operations may modify or query managed objects equivalently through any of the storage systems. Likewise, access operations to read or modify the dataset may operate equivalently through any of the storage systems. In such an example, while each storage system stores a separate copy of the dataset as a proper subset of the datasets stored and advertised for use by the storage system, the operations to modify managed objects or the dataset performed and completed through any one storage system are reflected in subsequent management objects to query the pod or subsequent access operations to read the dataset.

[000205] Readers will appreciate that pods may implement more capabilities than just a clustered synchronously replicated dataset. For example, pods can be used to implement tenants, whereby datasets are in some way securely isolated from each other. Pods can also be used to implement virtual arrays or virtual storage systems where each pod is presented as a unique storage entity on a network (e.g., a Storage Area Network, or Internet Protocol

network) with separate addresses. In the case of a multi-storage-system pod implementing a virtual storage system, all physical storage systems associated with the pod may present themselves as in some way the same storage system (e.g., as if the multiple physical storage systems were no different than multiple network ports into a single storage system).

[000206] Readers will appreciate that pods may also be units of administration, representing a collection of volumes, file systems, object/analytic stores, snapshots, and other administrative entities, where making administrative changes (e.g., name changes, property changes, managing exports or permissions for some part of the pod's dataset), on any one storage system is automatically reflected to all active storage systems associated with the pod. In addition, pods could also be units of data collection and data analysis, where performance and capacity metrics are presented in ways that aggregate across all active storage systems for the pod, or that call out data collection and analysis separately for each pod, or perhaps presenting each attached storage system's contribution to the incoming content and performance for each a pod.

[000207] One model for pod membership may be defined as a list of storage systems, and a subset of that list where storage systems are considered to be in-sync for the pod. A storage system may be considered to be in-sync for a pod if it is at least within a recovery of having identical idle content for the last written copy of the dataset associated with the pod. Idle content is the content after any in-progress modifications have completed with no processing of new modifications. Sometimes this is referred to as "crash recoverable" consistency. Recovery of a pod carries out the process of reconciling differences in applying concurrent updates to in-sync storage systems in the pod. Recovery can resolve any inconsistencies between storage systems in the completion of concurrent modifications that had been requested to various members of the pod but that were not signaled to any requestor as having completed successfully. Storage systems that are listed as pod members but that are not listed as in-sync for the pod can be described as "detached" from the pod. Storage systems that are listed as pod members, are in-sync for the pod, and are currently available for actively serving data for the pod are "online" for the pod.

[000208] Each storage system member of a pod may have its own copy of the membership, including which storage systems it last knew were in-sync, and which storage systems it last knew comprised the entire set of pod members. To be online for a pod, a storage system must consider itself to be in-sync for the pod and must be communicating with all other storage systems it considers to be in-sync for the pod. If a storage system can't be certain that it is in-sync and communicating with all other storage systems that are in-sync, then it

must stop processing new incoming requests for the pod (or must complete them with an error or exception) until it can be certain that it is in-sync and communicating with all other storage systems that are in-sync. A first storage system may conclude that a second paired storage system should be detached, which will allow the first storage system to continue since it is now in-sync with all storage systems now in the list. But, the second storage system must be prevented from concluding, alternatively, that the first storage system should be detached and with the second storage system continuing operation. This would result in a "split brain" condition that can lead to irreconcilable datasets, dataset corruption, or application corruption, among other dangers.

[000209] The situation of needing to determine how to proceed when not communicating with paired storage systems can arise while a storage system is running normally and then notices lost communications, while it is currently recovering from some previous fault, while it is rebooting or resuming from a temporary power loss or recovered communication outage, while it is switching operations from one set of storage system controller to another set for whatever reason, or during or after any combination of these or other kinds of events. In fact, any time a storage system that is associated with a pod can't communicate with all known non-detached members, the storage system can either wait briefly until communications can be established, go offline and continue waiting, or it can determine through some means that it is safe to detach the non-communicating storage system without risk of incurring a split brain due to the non-communicating storage system concluding the alternative view, and then continue. If a safe detach can happen quickly enough, the storage system can remain online for the pod with little more than a short delay and with no resulting application outages for applications that can issue requests to the remaining online storage systems.

[000210] One example of this situation is when a storage system may know that it is out-of-date. That can happen, for example, when a first storage system is first added to a pod that is already associated with one or more storage systems, or when a first storage system reconnects to another storage system and finds that the other storage system had already marked the first storage system as detached. In this case, this first storage system will simply wait until it connects to some other set of storage systems that are in-sync for the pod.

[000211] This model demands some degree of consideration for how storage systems are added to or removed from pods or from the in-sync pod members list. Since each storage system will have its own copy of the list, and since two independent storage systems can't update their local copy at exactly the same time, and since the local copy is all that is available on a reboot or in various fault scenarios, care must be taken to ensure that transient

inconsistencies don't cause problems. For example, if one storage systems is in-sync for a pod and a second storage system is added, then if the second storage system is updated to list both storage systems as in-sync first, then if there is a fault and a restart of both storage systems, the second might startup and wait to connect to the first storage system while the first might be unaware that it should or could wait for the second storage system. If the second storage system then responds to an inability to connect with the first storage system by going through a process to detach it, then it might succeed in completing a process that the first storage system is unaware of, resulting in a split brain. As such, it may be necessary to ensure that storage systems won't disagree inappropriately on whether they might opt to go through a detach process if they aren't communicating.

[000212] One way to ensure that storage systems won't disagree inappropriately on whether they might opt to go through a detach process if they aren't communicating is to ensure that when adding a new storage system to the in-sync member list for a pod, the new storage system first stores that it is a detached member (and perhaps that it is being added as an in-sync member). Then, the existing in-sync storage systems can locally store that the new storage system is an in-sync pod member before the new storage system locally stores that same fact. If there is a set of reboots or network outages prior to the new storage system storing its in-sync status, then the original storage systems may detach the new storage system due to non-communication, but the new storage system will wait. A reverse version of this change might be needed for removing a communicating storage system from a pod: first the storage system being removed stores that it is no longer in-sync, then the storage systems that will remain store that the storage system being removed is no longer in-sync, then all storage systems delete the storage system being removed from their pod membership lists. Depending on the implementation, an intermediate persisted detached state may not be necessary. Whether or not care is required in local copies of membership lists may depend on the model storage systems use for monitoring each other or for validating their membership. If a consensus model is used for both, or if an external system (or an external distributed or clustered system) is used to store and validate pod membership, then inconsistencies in locally stored membership lists may not matter.

[000213] When communications fail or one or several storage systems in a pod fail, or when a storage system starts up (or fails over to a secondary controller) and can't communicate with paired storage systems for a pod, and it is time for one or more storage systems to decide to detach one or more paired storage systems, some algorithm or mechanism must be employed to decide that it is safe to do so and to follow through on the detach. One means of

resolving detaches is use a majority (or quorum) model for membership. With three storage systems, as long as two are communicating, they can agree to detach a third storage system that isn't communicating, but that third storage system cannot by itself choose to detach either of the other two. Confusion can arise when storage system communication is inconsistent. For example, storage system A might be communicating with storage system B but not C, while storage system B might be communicating with both A and C. So, A and B could detach C, or B and C could detach A, but more communication between pod members may be needed to figure this out.

[000214] Care needs to be taken in a quorum membership model when adding and removing storage systems. For example, if a fourth storage system is added, then a "majority" of storage systems is at that point three. The transition from three storage systems (with two required for majority) to a pod including a fourth storage system (with three required for majority) may require something similar to the model described previously for carefully adding a storage system to the in-sync list. For example, the fourth storage system might start in an attaching state but not yet attached where it would never instigate a vote over quorum. Once in that state, the original three pod members could each be updated to be aware of the fourth member and the new requirement for a three storage system majority to detach a fourth. Removing a storage system from a pod might similarly move that storage system to a locally stored "detaching" state before updating other pod members. A variant scheme for this is to use a distributed consensus mechanism such as PAXOS or RAFT to implement any membership changes or to process detach requests.

[000215] Another means of managing membership transitions is to use an external system that is outside of the storage systems themselves to handle pod membership. In order to become online for a pod, a storage system must first contact the external pod membership system to verify that it is in-sync for the pod. Any storage system that is online for a pod should then remain in communication with the pod membership system and should wait or go offline if it loses communication. An external pod membership manager could be implemented as a highly available cluster using various cluster tools, such as Oracle RAC, Linux HA, VERITAS Cluster Server, IBM's HACMP, or others. An external pod membership manager could also use distributed configuration tools such as Etcd or Zookeeper, or a reliable distributed database such as Amazon's DynamoDB.

[000216] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may receive a request to read a portion of the dataset (426, 428) and process the request to read the portion of the dataset locally according to some embodiments of the present

disclosure. Readers will appreciate that although requests to modify (e.g., a write operation) the dataset (426, 428) require coordination between the storage systems (402, 404, 406) in a pod, as the dataset (426, 428) should be consistent across all storage systems (402, 404, 406) in a pod, responding to a request to read a portion of the dataset (426, 428) does not require similar coordination between the storage systems (402, 404, 406). As such, a particular storage system that receives a read request may service the read request locally by reading a portion of the dataset (426, 428) that is stored within the storage system's storage devices, with no synchronous communication with other storage systems in the pod. Read requests received by one storage system for a replicated dataset in a replicated cluster are expected to avoid any communication in the vast majority of cases, at least when received by a storage system that is running within a cluster that is also running nominally. Such reads should normally be processed simply by reading from the local copy of a clustered dataset with no further interaction required with other storage systems in the cluster

[000217] Readers will appreciate that the storage systems may take steps to ensure read consistency such that a read request will return the same result regardless of which storage system processes the read request. For example, the resulting clustered dataset content for any set of updates received by any set of storage systems in the cluster should be consistent across the cluster, at least at any time updates are idle (all previous modifying operations have been indicated as complete and no new update requests have been received and processed in any way). More specifically, the instances of a clustered dataset across a set of storage systems can differ only as a result of updates that have not yet completed. This means, for example, that any two write requests which overlap in their volume block range, or any combination of a write request and an overlapping snapshot, compare-and-write, or virtual block range copy, must yield a consistent result on all copies of the dataset. Two operations should not yield a result as if they happened in one order on one storage system and a different order on another storage system in the replicated cluster.

[000218] Furthermore, read requests can be made time order consistent. For example, if one read request is received on a replicated cluster and completed and that read is then followed by another read request to an overlapping address range which is received by the replicated cluster and where one or both reads in any way overlap in time and volume address range with a modification request received by the replicated cluster (whether any of the reads or the modification are received by the same storage system or a different storage system in the replicated cluster), then if the first read reflects the result of the update then the second read should also reflect the results of that update, rather than possibly returning data that preceded

the update. If the first read does not reflect the update, then the second read can either reflect the update or not. This ensures that between two read requests "time" for a data segment cannot roll backward.

[000219] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also detect a disruption in data communications with one or more of the other storage systems and determine whether to the particular storage system should remain in the pod. A disruption in data communications with one or more of the other storage systems may occur for a variety of reasons. For example, a disruption in data communications with one or more of the other storage systems may occur because one of the storage systems has failed, because a network interconnect has failed, or for some other reason. An important aspect of synchronous replicated clustering is ensuring that any fault handling doesn't result in unrecoverable inconsistencies, or any inconsistency in responses. For example, if a network fails between two storage systems, at most one of the storage systems can continue processing newly incoming I/O requests for a pod. And, if one storage system continues processing, the other storage system can't process any new requests to completion, including read requests.

[000220] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also determine whether the particular storage system should remain in the pod in response to detecting a disruption in data communications with one or more of the other storage systems. As mentioned above, to be 'online' as part of a pod, a storage system must consider itself to be in-sync for the pod and must be communicating with all other storage systems it considers to be in-sync for the pod. If a storage system can't be certain that it is in-sync and communicating with all other storage systems that are in-sync, then it may stop processing new incoming requests to access the dataset (426, 428). As such, the storage system may determine whether to the particular storage system should remain online as part of the pod, for example, by determining whether it can communicate with all other storage systems it considers to be in-sync for the pod (e.g., via one or more test messages), by determining whether the all other storage systems it considers to be in-sync for the pod also consider the storage system to be attached to the pod, through a combination of both steps where the particular storage system must confirm that it can communicate with all other storage systems it considers to be in-sync for the pod and that all other storage systems it considers to be in-sync for the pod also consider the storage system to be attached to the pod, or through some other mechanism.

[000221] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also keep the dataset on the particular storage system accessible for management and dataset operations in response to determining that the particular storage system should remain in the pod. The storage system may keep the dataset (426, 428) on the particular storage system accessible for management and dataset operations, for example, by accepting requests to access the version of the dataset (426, 428) that is stored on the storage system and processing such requests, by accepting and processing management operations associated with the dataset (426, 428) that are issued by a host or authorized administrator, by accepting and processing management operations associated with the dataset (426, 428) that are issued by one of the other storage systems, or in some other way.

[000222] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may, however, make the dataset on the particular storage system inaccessible for management and dataset operations in response to determining that the particular storage system should not remain in the pod. The storage system may make the dataset (426, 428) on the particular storage system inaccessible for management and dataset operations, for example, by rejecting requests to access the version of the dataset (426, 428) that is stored on the storage system, by rejecting management operations associated with the dataset (426, 428) that are issued by a host or other authorized administrator, by rejecting management operations associated with the dataset (426, 428) that are issued by one of the other storage systems in the pod, or in some other way.

[000223] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also detect that the disruption in data communications with one or more of the other storage systems has been repaired and make the dataset on the particular storage system accessible for management and dataset operations. The storage system may detect that the disruption in data communications with one or more of the other storage systems has been repaired, for example, by receiving a message from the one or more of the other storage systems. In response to detecting that the disruption in data communications with one or more of the other storage systems has been repaired, the storage system may make the dataset (426, 428) on the particular storage system accessible for management and dataset operations once the previously detached storage system has been resynchronized with the storage systems that remained attached to the pod.

[000224] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also go offline from the pod such that the particular storage system no longer allows management and dataset operations. The depicted storage systems (402, 404, 406) may go

offline from the pod such that the particular storage system no longer allows management and dataset operations for a variety of reasons. For example, the depicted storage systems (402, 404, 406) may also go offline from the pod due to some fault with the storage system itself, because an update or some other maintenance is occurring on the storage system, due to communications faults, or for many other reasons. In such an example, the depicted storage systems (402, 404, 406) may subsequently update the dataset on the particular storage system to include all updates to the dataset since the particular storage system went offline and go back online with the pod such that the particular storage system allows management and dataset operations, as will be described in greater detail in the resynchronization sections included below.

[000225] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also identifying a target storage system for asynchronously receiving the dataset, where the target storage system is not one of the plurality of storage systems across which the dataset is synchronously replicated. Such a target storage system may represent, for example, a backup storage system, as some storage system that makes use of the synchronously replicated dataset, and so on. In fact, synchronous replication can be leveraged to distribute copies of a dataset closer to some rack of servers, for better local read performance. One such case is smaller top-of-rack storage systems symmetrically replicated to larger storage systems that are centrally located in the data center or campus and where those larger storage systems are more carefully managed for reliability or are connected to external networks for asynchronous replication or backup services.

[000226] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also identify a portion of the dataset that is not being asynchronously replicated to the target storage system by any of the other storages systems and asynchronously replicate, to the target storage system, the portion of the dataset that is not being asynchronously replicated to the target storage system by any of the other storages systems, wherein the two or more storage systems collectively replicate the entire dataset to the target storage system. In such a way, the work associated with asynchronously replicating a particular dataset may be split amongst the members of a pod, such that each storage system in a pod is only responsible for asynchronously replicating a subset of a dataset to the target storage system.

[000227] In the example depicted in Figure 4A, the depicted storage systems (402, 404, 406) may also detach from the pod, such that the particular storage system that detaches from the pod is no longer included in the set of storage systems across which the dataset is synchronously replicated. For example, if storage system (404) in Figure 4A detached from

the pod (430) illustrated in Figure 4A, the pod (430) would only include storage systems (402, 406) as the storage systems across which the dataset (426) that is included in the pod (430) would be synchronously replicated across. In such an example, detaching the storage system from the pod could also include removing the dataset from the particular storage system that detached from the pod. Continuing with the example where the storage system (404) in Figure 4A detached from the pod (430) illustrated in Figure 4A, the dataset (426) that is included in the pod (430) could be deleted or otherwise removed from the storage system (404).

[000228] Readers will appreciate that there are a number of unique administrative capabilities enabled by the pod model that can further be supported. Also, the pod model itself introduces some issues that can be addressed by an implementation. For example, when a storage system is offline for a pod, but is otherwise running, such as because an interconnect failed and another storage system for the pod won out in mediation, there may still be a desire or need to access the offline pod's dataset on the offline storage system. One solution may be simply to enable the pod in some detached mode and allow the dataset to be accessed. However, that solution can be dangerous and that solution can cause the pod's metadata and data to be much more difficult to reconcile when the storage systems do regain communication. Furthermore, there could still be a separate path for hosts to access the offline storage system as well as the still online storage systems. In that case, a host might issue I/O to both storage systems even though they are no longer being kept in sync, because the host sees target ports reporting volumes with the same identifiers and the host I/O drivers presume it sees additional paths to the same volume. This can result in fairly damaging data corruption as reads and writes issued to both storage systems are no longer consistent even though the host presumes they are. As a variant of this case, in a clustered application, such as a shared storage clustered database, the clustered application running on one host might be reading or writing to one storage system and the same clustered application running on another host might be reading or writing to the "detached" storage system, yet the two instances of the clustered application are communicating between each other on the presumption that the dataset they each see is entirely consistent for completed writes. Since they aren't consistent, that presumption is violated and the application's dataset (e.g., the database) can quickly end up being corrupted.

[000229] One way to solve both of these problems is to allow for an offline pod, or perhaps a snapshot of an offline pod, to be copied to a new pod with new volumes that have sufficiently new identities that host I/O drivers and clustered applications won't confuse the copied

volumes as being the same as the still online volumes on another storage system. Since each pod maintains a complete copy of the dataset, which is crash consistent but perhaps slightly different from the copy of the pod dataset on another storage system, and since each pod has an independent copy of all data and metadata needed to operate on the pod content, it is a straightforward problem to make a virtual copy of some or all volumes or snapshots in the pod to new volumes in a new pod. In a logical extent graph implementation, for example, all that is needed is to define new volumes in a new pod which reference logical extent graphs from the copied pod associated with the pod's volumes or snapshots, and with the logical extent graphs being marked as copy on write. The new volumes should be treated as new volumes, similarly to how volume snapshots copied to a new volume might be implemented. Volumes may have the same administrative name, though within a new pod namespace. But, they should have different underlying identifiers, and differing logical unit identifiers from the original volumes.

[000230] In some cases it may be possible to use virtual network isolation techniques (for example, by creating a virtual LAN in the case of IP networks or a virtual SAN in the case of fiber channel networks) in such a way that isolation of volumes presented to some interfaces can be assured to be inaccessible from host network interfaces or host SCSI initiator ports that might also see the original volumes. In such cases, it may be safe to provide the copies of volumes with the same SCSI or other storage identifiers as the original volumes. This could be used, for example, in cases where the applications expect to see a particular set of storage identifiers in order to function without an undue burden in reconfiguration.

[000231] Some of the techniques described herein could also be used outside of an active fault context to test readiness for handling faults. Readiness testing (sometimes referred to as "fire drills") is commonly required for disaster recovery configurations, where frequent and repeated testing is considered a necessity to ensure that most or all aspects of a disaster recovery plan are correct and account for any recent changes to applications, datasets, or changes in equipment. Readiness testing should be non-disruptive to current production operations, including replication. In many cases the real operations can't actually be invoked on the active configuration, but a good way to get close is to use storage operations to make copies of production datasets, and then perhaps couple that with the use of virtual networking, to create an isolated environment containing all data that is believed necessary for the important applications that must be brought up successfully in cases of disasters. Making such a copy of a synchronously replicated (or even an asynchronously replicated) dataset available within a site (or collection of sites) that is expected to perform a disaster

recovery readiness test procedure and then starting the important applications on that dataset to ensure that it can startup and function is a great tool, since it helps ensure that no important parts of the application datasets were left out in the disaster recovery plan. If necessary, and practical, this could be coupled with virtual isolated networks coupled perhaps with isolated collection of physical or virtual machines, to get as close as possible to a real world disaster recovery takeover scenario. Virtually copying a pod (or set of pods) to another pod as a point-in-time image of the pod datasets immediately creates an isolated dataset that contains all the copied elements and that can then be operated on essentially identically to the originally pods, as well as allowing isolation to a single site (or a few sites) separately from the original pod. Further, these are fast operations and they can be torn down and repeated easily allowing testing to repeated as often as is desired.

[000232] Some enhancements could be made to get further toward perfect disaster recovery testing. For example, in conjunction with isolated networks, SCSI logical unit identities or other types of identities could be copied into the target pod so that the test servers, virtual machines, and applications see the same identities. Further, the administrative environment of the servers could be configured to respond to requests from a particular virtual set of virtual networks to respond to requests and operations on the original pod name so scripts don't require use of test-variants with alternate "test" versions of object names. A further enhancement can be used in cases where the host-side server infrastructure that will take over in the case of a disaster takeover can be used during a test. This includes cases where a disaster recovery data center is completely stocked with alternative server infrastructure that won't generally be used until directed to do so by a disaster. It also includes cases where that infrastructure might be used for non-critical operations (for example, running analytics on production data, or simply supporting application development or other functions which may be important but can be halted if needed for more critical functions). Specifically, host definitions and configurations and the server infrastructure that will use them can be set up as they will be for an actual disaster recovery takeover event and tested as part of disaster recovery takeover testing, with the tested volumes being connected to these host definitions from the virtual pod copy used to provide a snapshot of the dataset. From the standpoint of the storage systems involved, then, these host definitions and configurations used for testing, and the volume-to-host connection configurations used during testing, can be reused when an actual disaster takeover event is triggered, greatly minimizing the configuration differences between the test configuration and the real configuration that will be used in case of a disaster recovery takeover.

[000233] In some cases it may make sense to move volumes out of a first pod and into a new second pod including just those volumes. The pod membership and high availability and recovery characteristics can then be adjusted separately, and administration of the two resulting pod datasets can then be isolated from each other. An operation that can be done in one direction should also be possible in the other direction. At some point, it may make sense to take two pods and merge them into one so that the volumes in each of the original two pods will now track each other for storage system membership and high availability and recovery characteristics and events. Both operations can be accomplished safely and with reasonably minimal or no disruption to running applications by relying on the characteristics suggested for changing mediation or quorum properties for a pod which were discussed in an earlier section. With mediation, for example, a mediator for a pod can be changed using a sequence consisting of a step where each storage system in a pod is changed to depend on both a first mediator and a second mediator and each is then changed to depend only on the second mediator. If a fault occurs in the middle of the sequence, some storage systems may depend on both the first mediator and the second mediator, but in no case will recovery and fault handling result in some storage systems depending only on the first mediator and other storage systems only depending on the second mediator. Quorum can be handled similarly by temporarily depending on winning against both a first quorum model and a second quorum model in order to proceed to recovery. This may result in a very short time period where availability of the pod in the face of faults depend on additional resources, thus reducing potential availability, but this time period is very short and the reduction in availability is often very little. With mediation, if the change in mediator parameters is nothing more than the change in the key used for mediation and the mediation service used is the same, then the potential reduction in availability is even less, since it now depends only on two calls to the same service versus one call to that service, and rather than separate calls to two separate services.

[000234] Readers will note that changing the quorum model may be quite complex. An additional step may be necessary where storage systems will participate in the second quorum model but won't depend on winning in that second quorum model, which is then followed by the step of also depending on the second quorum model. This may be necessary to account for the fact that if only one system has processed the change to depend on the quorum model, then it will never win quorum since there will never be a majority. With this model in place for changing the high availability parameters (mediation relationship, quorum model, takeover preferences), we can create a safe procedure for these operations to split a pod into

two or to join two pods into one. This may require adding one other capability: linking a second pod to a first pod for high availability such that if two pods include compatible high availability parameters the second pod linked to the first pod can depend on the first pod for determining and instigating detach-related processing and operations, offline and in-sync states, and recovery and resynchronization actions.

[000235] To split a pod into two, which is an operation to move some volumes into a newly created pod, a distributed operation may be formed that can be described as: form a second pod into which we will move a set of volumes which were previously in a first pod, copy the high availability parameters from the first pod into the second pod to ensure they are compatible for linking, and link the second pod to the first pod for high availability. This operation may be encoded as messages and should be implemented by each storage system in the pod in such a way that the storage system ensures that the operation happens completely on that storage system or does not happen at all if processing is interrupted by a fault. Once all in-sync storage systems for the two pods have processed this operation, the storage systems can then process a subsequent operation which changes the second pod so that it is no longer linked to the first pod. As with other changes to high availability characteristics for a pod, this involves first having each in-sync storage system change to rely on both the previous model (that model being that high availability is linked to the first pod) and the new model (that model being its own now independent high availability). In the case of mediation or quorum, this means that storage systems which processed this change will first depend on mediation or quorum being achieved as appropriate for the first pod and will additionally depend on a new separate mediation (for example, a new mediation key) or quorum being achieved for the second pod before the second pod can proceed following a fault that required mediation or testing for quorum. As with the previous description of changing quorum models, an intermediate step may set storage systems to participate in quorum for the second pod before the step where storage systems participate in and depend on quorum for the second pod. Once all in-sync storage systems have processed the change to depend on the new parameters for mediation or quorum for both the first pod and the second pod, the split is complete.

[000236] Joining a second pod into a first pod operates essentially in reverse. First, the second pod must be adjusted to be compatible with the first pod, by having an identical list of storage systems and by having a compatible high availability model. This may involve some set of steps such as those described elsewhere in this paper to add or remove storage systems or to change mediator and quorum models. Depending on implementation, it may be

necessary only to reach an identical list of storage systems. Joining proceeds by processing an operation on each in-sync storage system to link the second pod to the first pod for high availability. Each storage system which processes that operation will then depend on the first pod for high availability and then the second pod for high availability. Once all in-sync storage systems for the second pod have processed that operation, the storage systems will then each process a subsequent operation to eliminate the link between the second pod and the first pod, migrate the volumes from the second pod into the first pod, and delete the second pod. Host or application dataset access can be preserved throughout these operations, as long as the implementation allows proper direction of host or application dataset modification or read operations to the volume by identity and as long as the identity is preserved as appropriate to the storage protocol or storage model (for example, as long as logical unit identifiers for volumes and use of target ports for accessing volumes are preserved in the case of SCSI).

[000237] Migrating a volume between pods may present issues. If the pods have an identical set of in-sync membership storage systems, then it may be straightforward: temporarily suspend operations on the volumes being migrated, switch control over operations on those volumes to controlling software and structures for the new pod, and then resume operations. This allows for a seamless migration with continuous uptime for applications apart from the very brief operation suspension, provided network and ports migrate properly between pods. Depending on the implementation, suspending operations may not even be necessary, or may be so internal to the system that the suspension of operations has no impact. Copying volumes between pods with different in-sync membership sets is more of a problem. If the target pod for the copy has a subset of in-sync members from the source pod, this isn't much of a problem: a member storage system can be dropped safely enough without having to do more work. But, if the target pod adds in-sync member storage systems to the volume over the source pod, then the added storage systems must be synchronized to include the volume's content before they can be used. Until synchronized, this leaves the copied volumes distinctly different from the already synchronized volumes, in that fault handling differs and request handling from the not yet synced member storage systems either won't work or must be forwarded or won't be as fast because reads will have to traverse an interconnect. Also, the internal implementation will have to handle some volumes being in sync and ready for fault handling and others not being in sync.

[000238] There are other problems relating to reliability of the operation in the face of faults. Coordinating a migration of volumes between multi-storage-system pods is a distributed

operation. If pods are the unit of fault handling and recovery, and if mediation or quorum or whatever means are used to avoid split-brain situations, then a switch in volumes from one pod with a particular set of state and configurations and relationships for fault handling, recovery, mediation and quorum to another then storage systems in a pod have to be careful about coordinating changes related to that handling for any volumes. Operations can't be atomically distributed between storage systems, but must be staged in some way. Mediation and quorum models essentially provide pods with the tools for implementing distributed transactional atomicity, but this may not extend to inter-pod operations without adding to the implementation.

[000239] Consider even a simple migration of a volume from a first pod to a second pod even for two pods that share the same first and second storage systems. At some point the storage systems will coordinate to define that the volume is now in the second pod and is no longer in the first pod. If there is no inherent mechanism for transactional atomicity across the storage systems for the two pods, then a naive implementation could leave the volume in the first pod on the first storage system and the second pod on the second storage system at the time of a network fault that results in fault handling to detach storage systems from the two pods. If pods separately determine which storage system succeeds in detaching the other, then the result could be that the same storage system detaches the other storage system for both pods, in which case the result of the volume migration recovery should be consistent, or it could result in a different storage system detaching the other for the two pods. If the first storage system detaches the second storage system for the first pod and the second storage system detaches the first storage system for the second pod, then recovery might result in the volume being recovered to the first pod on the first storage system and into the second pod on the second storage system, with the volume then running and exported to hosts and storage applications on both storage systems. If instead the second storage system detaches the first storage system for the first pod and first storage detaches the second storage system for the second pod, then recovery might result in the volume being discarded from the second pod by the first storage system and the volume being discarded from the first pod by the second storage system, resulting in the volume disappearing entirely. If the pods a volume is being migrated between are on differing sets of storage systems, then things can get even more complicated.

[000240] A solution to these problems may be to use an intermediate pod along with the techniques described previously for splitting and joining pods. This intermediate pod may never be presented as visible managed objects associated with the storage systems. In this

model, volumes to be moved from a first pod to a second pod are first split from the first pod into a new intermediate pod using the split operation described previously. The storage system members for the intermediate pod can then be adjusted to match the membership of storage systems by adding or removing storage systems from the pod as necessary. Subsequently, the intermediate pod can be joined with the second pod.

[000241] For further explanation, Figure 4B sets forth diagrams of metadata representations that may be implemented as a structured collection of metadata objects that, together, may represent a logical volume of storage data, or a portion of a logical volume, in accordance with some embodiments of the present disclosure. Metadata representations 451-50, 451-54, and 451-60 may be stored within a storage system (451-06), and one or more metadata representations may be generated and maintained for each of multiple storage objects, such as volumes, or portions of volumes, stored within a storage system (451-06).

[000242] While other types of structured collections of the metadata objects are possible, in this example, metadata representations may be structured as a directed acyclic graph (DAG) of nodes, where, to maintain efficient access to any given node, the DAG may be structured and balanced according to various methods. For example, a DAG for a metadata representation may be defined as a type of B-tree, and balanced accordingly in response to changes to the structure of the metadata representation, where changes to the metadata representation may occur in response to changes to, or additions to, underlying data represented by the metadata representation. While in this example, there are only two levels for the sake of simplicity, in other examples, metadata representations may span across multiple levels and may include hundreds or thousands of nodes, where each node may include any number of links to other nodes.

[000243] Further, in this example, the leaves of a metadata representation may include pointers to the stored data for a volume, or portion of a volume, where a logical address, or a volume and offset, may be used to identify and navigate through the metadata representation to reach one or more leaf nodes that reference stored data corresponding to the logical address. For example, a volume (451-52) may be represented by a metadata representation (451-50), which includes multiple metadata object nodes (451-52, 451-52A–451-52N), where leaf nodes (451-52A–451-52N) include pointers to respective data objects (451-53A–451-53N, 451-57). Data objects may be any size unit of data within a storage system (451-06). For example, data objects (451-53A–451-53N, 451-57) may each be a logical extent, where logical extents may be some specified size, such as 1MB, 4MB, or some other size.

[000244] In this example, a snapshot (451-56) may be created as a snapshot of a storage object, in this case, a volume (451-52), where at the point in time when the snapshot (451-56) is created, the metadata representation (451-54) for the snapshot (451-56) includes all of the metadata objects for the metadata representation (451-50) for the volume (451-52). Further, in response to creation of the snapshot (451-56), the metadata representation (451-54) may be designated to be read only. However, the volume (451-52) sharing the metadata representation may continue to be modified, and while at the moment the snapshot is created, the metadata representations for the volume (451-52) and the snapshot (451-56) are identical, as modifications are made to data corresponding to the volume (451-52), and in response to the modifications, the metadata representations for the volume (451-52) and the snapshot (451-56) may diverge and become different.

[000245] For example, given a metadata representation (451-50) to represent a volume (451-52) and a metadata representation (451-54) to represent a snapshot (451-56), the storage system (451-06) may receive an I/O operation that writes to data that is ultimately stored within a particular data object (451-53B), where the data object (451-53B) is pointed to by a leaf node pointer (451-52B), and where the leaf node pointer (451-52B) is part of both metadata representations (451-50, 451-54). In response to the write operation, the read only data objects (451-53A–451-53N) referred to by the metadata representation (451-54) remain unchanged, and the pointer (451-52B) may also remain unchanged. However, the metadata representation (451-50), which represents the current volume (451-52), is modified to include a new data object to hold the data written by the write operation, where the modified metadata representation is depicted as the metadata representation (451-60). Further, the write operation may be directed to only a portion of the data object (451-53B), and consequently, the new data object (451-57) may include a copy of previous contents of the data object (451-53B) in addition to the payload for the write operation.

[000246] In this example, as part of processing the write operation, the metadata representation (451-60) for the volume (451-52) is modified to remove an existing metadata object pointer (451-52B) and to include a new metadata object pointer (451-58), where the new metadata object pointer (451-58) is configured to point to a new data object (451-57), where the new data object (451-57) stores the data written by the write operation. Further, the metadata representation (451-60) for the volume (451-52) continues to include all metadata objects included within the previous metadata representation (451-50)—with the exclusion of the metadata object pointer (451-52B) that referenced the target data object,

where the metadata object pointer (451-52B) continues to reference the read only data object (451-53B) that would have been overwritten.

[000247] In this way, using metadata representations, a volume or a portion of a volume may be considered to be snapshotted, or considered to be copied, by creating metadata objects, and without actual duplication of data objects—where the duplication of data objects may be deferred until a write operation is directed at one of the read only data objects referred to by the metadata representations.

[000248] In other words, an advantage of using a metadata representation to represent a volume is that a snapshot or a copy of a volume may be created and be accessible in constant order time, and specifically, in the time it takes to create a metadata object for the snapshot or copy, and to create a reference for the snapshot or copy metadata object to the existing metadata representation for the volume being snapshotted or copied.

[000249] As an example use, a virtualized copy-by-reference may make use of a metadata representation in a manner that is similar to the use of a metadata representation in creating a snapshot of a volume—where a metadata representation for a virtualized copy-by-reference may often correspond to a portion of a metadata representation for an entire volume. An example implementation of virtualized copy-by-reference may be within the context of a virtualized storage system, where multiple block ranges within and between volumes may reference a unified copy of stored data. In such virtualized storage system, the metadata described above may be used to handle the relationship between virtual, or logical, addresses and physical, or real, addresses—in other words, the metadata representation of stored data enables a virtualized storage system that may be considered flash-friendly in that it reduces, or minimizes, wear on flash memory.

[000250] In some examples, logical extents may be combined in various ways, including as simple collections or as logically related address ranges within some larger-scale logical extent that is formed as a set of logical extent references. These larger combinations could also be given logical extent identities of various kinds, and could be further combined into still larger logical extents or collections. A copy-on-write status could apply to various layers, and in various ways depending on the implementation. For example, a copy on write status applied to a logical collection of logical collections of extents might result in a copied collection retaining references to unchanged logical extents and the creation of copied-on-write logical extents (through copying references to any unchanged stored data blocks as needed) when only part of the copy-on-write logical collection is changed.

[000251] Deduplication, volume snapshots, or block range snapshots may be implemented in this model through combinations of referencing stored data blocks, or referencing logical extents, or marking logical extents (or identified collections of logical extents) as copy-on-write.

[000252] Further, with flash storage systems, stored data blocks may be organized and grouped together in various ways as collections are written out into pages that are part of larger erase blocks. Eventual garbage collection of deleted or replaced stored data blocks may involve moving content stored in some number of pages elsewhere so that an entire erase block can be erased and prepared for reuse. This process of selecting physical flash pages, eventually migrating and garbage collecting them, and then erasing flash erase blocks for reuse may or may not be coordinated, driven by, or performed by the aspect of a storage system that is also handling logical extents, deduplication, compression, snapshots, virtual copying, or other storage system functions. A coordinated or driven process for selecting pages, migrating pages, garbage collecting and erasing erase blocks may further take into account various characteristics of the flash memory device cells, pages, and erase blocks such as number of uses, aging predictions, adjustments to voltage levels or numbers of retries needed in the past to recover stored data. They may also take into account analysis and predictions across all flash memory devices within the storage system.

[000253] To continue with this example, where a storage system may be implemented based on directed acyclic graphs comprising logical extents, logical extents can be categorized into two types: leaf logical extents, which reference some amount of stored data in some way, and composite logical extents, which reference other leaf or composite logical extents.

[000254] A leaf extent can reference data in a variety of ways. It can point directly to a single range of stored data (e.g., 64 kilobytes of data), or it can be a collection of references to stored data (e.g., a 1 megabyte "range" of content that maps some number of virtual blocks associated with the range to physically stored blocks). In the latter case, these blocks may be referenced using some identity, and some blocks within the range of the extent may not be mapped to anything. Also, in that latter case, these block references need not be unique, allowing multiple mappings from virtual blocks within some number of logical extents within and across some number of volumes to map to the same physically stored blocks. Instead of stored block references, a logical extent could encode simple patterns: for example, a block which is a string of identical bytes could simply encode that the block is a repeated pattern of identical bytes.

[000255] A composite logical extent can be a logical range of content with some virtual size, which comprises a plurality of maps that each map from a subrange of the composite logical extent logical range of content to an underlying leaf or composite logical extent. Transforming a request related to content for a composite logical extent, then, involves taking the content range for the request within the context of the composite logical extent, determining which underlying leaf or composite logical extents that request maps to, and transforming the request to apply to an appropriate range of content within those underlying leaf or composite logical extents.

[000256] Volumes, or files or other types of storage objects, can be described as composite logical extents. Thus, these presented storage objects can be organized using this extent model.

[000257] Depending on implementation, leaf or composite logical extents could be referenced from a plurality of other composite logical extents, effectively allowing inexpensive duplication of larger collections of content within and across volumes. Thus, logical extents can be arranged essentially within an acyclic graph of references, each ending in leaf logical extents. This can be used to make copies of volumes, to make snapshots of volumes, or as part of supporting virtual range copies within and between volumes as part of EXTENDED COPY or similar types of operations.

[000258] An implementation may provide each logical extent with an identity which can be used to name it. This simplifies referencing, since the references within composite logical extents become lists comprising logical extent identities and a logical subrange corresponding to each such logical extent identity. Within logical extents, each stored data block reference may also be based on some identity used to name it.

[000259] To support these duplicated uses of extents, we can add a further capability: copy-on-write logical extents. When a modifying operation affects a copy-on-write leaf or composite logical extent the logical extent is copied, with the copy being a new reference and possibly having a new identity (depending on implementation). The copy retains all references or identities related to underlying leaf or composite logical extents, but with whatever modifications result from the modifying operation. For example, a WRITE, WRITE SAME, XDWRITEREAD, XPWRITE, or COMPARE AND WRITE request may store new blocks in the storage system (or use deduplication techniques to identify existing stored blocks), resulting in modifying the corresponding leaf logical extents to reference or store identities to a new set of blocks, possibly replacing references and stored identities for a previous set of blocks. Alternately, an UNMAP request may modify a leaf logical extent to

remove one or more block references. In both types of cases, a leaf logical extent is modified. If the leaf logical extent is copy-on-write, then a new leaf logical extent will be created that is formed by copying unaffected block references from the old extent and then replacing or removing block references based on the modifying operation.

[000260] A composite logical extent that was used to locate the leaf logical extent may then be modified to store the new leaf logical extent reference or identity associated with the copied and modified leaf logical extent as a replacement for the previous leaf logical extent. If that composite logical extent is copy-on-write, then a new composite logical extent is created as a new reference or with a new identity, and any unaffected references or identities to its underlying logical extents are copied to that new composite logical extent, with the previous leaf logical extent reference or identity being replaced with the new leaf logical extent reference or identity.

[000261] This process continues further backward from referenced extent to referencing composite extent, based on the search path through the acyclic graph used to process the modifying operation, with all copy-on-write logical extents being copied, modified, and replaced.

[000262] These copied leaf and composite logical extents can then drop the characteristic of being copy on write, so that further modifications do not result in an additional copy. For example, the first time some underlying logical extent within a copy-on-write "parent" composite extent is modified, that underlying logical extent may be copied and modified, with the copy having a new identity which is then written into a copied and replaced instance of the parent composite logical extent. However, a second time some other underlying logical extent is copied and modified and with that other underlying logical extent copy's new identity being written to the parent composite logical extent, the parent can then be modified in place with no further copy and replace necessary on behalf of references to the parent composite logical extent.

[000263] Modifying operations to new regions of a volume or of a composite logical extent for which there is no current leaf logical extent may create a new leaf logical extent to store the results of those modifications. If that new logical extent is to be referenced from an existing copy-on-write composite logical extent, then that existing copy-on-write composite logical extent will be modified to reference the new logical extent, resulting in another copy, modify, and replace sequence of operations similar to the sequence for modifying an existing leaf logical extent.

[000264] If a parent composite logical extent cannot be grown large enough (based on implementation) to cover an address range associated that includes new leaf logical extents to create for a new modifying operation, then the parent composite logical extent may be copied into two or more new composite logical extents which are then referenced from a single "grandparent" composite logical extent which yet again is a new reference or a new identity. If that grandparent logical extent is itself found through another composite logical extent that is copy-on-write, then that another composite logical extent will be copied and modified and replaced in a similar way as described in previous paragraphs. This copy-on-write model can be used as part of implementing snapshots, volume copies, and virtual volume address range copies within a storage system implementation based on these directed acyclic graphs of logical extents. To make a snapshot as a read-only copy of an otherwise writable volume, a graph of logical extents associated with the volume is marked copy-on-write and a reference to the original composite logical extents are retained by the snapshot. Modifying operations to the volume will then make logical extent copies as needed, resulting in the volume storing the results of those modifying operations and the snapshots retaining the original content. Volume copies are similar, except that both the original volume and the copied volume can modify content resulting in their own copied logical extent graphs and subgraphs.

[000265] Virtual volume address range copies can operate either by copying block references within and between leaf logical extents (which does not itself involve using copy-on-write techniques unless changes to block references modifies copy-on-write leaf logical extents). Alternately, virtual volume address range copies can duplicate references to leaf or composite logical extents, which works well for volume address range copies of larger address ranges. Further, this allows graphs to become directed acyclic graphs of references rather than merely reference trees. Copy-on-write techniques associated with duplicated logical extent references can be used to ensure that modifying operations to the source or target of a virtual address range copy will result in the creation of new logical extents to store those modifications without affecting the target or the source that share the same logical extent immediately after the volume address range copy operation.

[000266] Input/output operations for pods may also be implemented based on replicating directed acyclic graphs of logical extents. For example, each storage system within a pod could implement private graphs of logical extents, such that the graphs on one storage system for a pod have no particular relationship to the graphs on any second storage system for the pod. However, there is value in synchronizing the graphs between storage systems in a pod. This can be useful for resynchronization and for coordinating features such as asynchronous

or snapshot based replication to remote storage systems. Further, it may be useful for reducing some overhead for handling the distribution of snapshot and copy related processing. In such a model, keeping the content of a pod in sync across all in-sync storage systems for a pod is essentially the same as keeping graphs of leaf and composite logical extents in sync for all volumes across all in-sync storage systems for the pod, and ensuring that the content of all logical extents is in-sync. To be in sync, matching leaf and composite logical extents should either have the same identity or should have mappable identities. Mapping could involve some set of intermediate mapping tables or could involve some other type of identity translation. In some cases, identities of blocks mapped by leaf logical extents could also be kept in sync.

[000267] In a pod implementation based on a leader and followers, with a single leader for each pod, the leader can be in charge of determining any changes to the logical extent graphs. If a new leaf or composite logical extent is to be created, it can be given an identity. If an existing leaf or composite logical extent is to be copied to form a new logical extent with modifications, the new logical extent can be described as a copy of a previous logical extent with some set of modifications. If an existing logical extent is to be split, the split can be described along with the new resulting identities. If a logical extent is to be referenced as an underlying logical extent from some additional composite logical extent, that reference can be described as a change to the composite logical extent to reference that underlying logical extent.

[000268] Modifying operations in a pod thus comprises distributing descriptions of modifications to logical extent graphs (where new logical extents are created to extend content or where logical extents are copied, modified, and replaced to handle copy-on-write states related to snapshots, volume copies, and volume address range copies) and distributing descriptions and content for modifications to the content of leaf logical extents. An additional benefit that comes from using metadata in the form of directed acyclic graphs, as described above, is that I/O operations that modify stored data in physical storage may be given effect at a user level through the modification of metadata corresponding to the stored data in physical storage—without modifying the stored data in physical storage. In the disclosed embodiments of storage systems, where the physical storage may be a solid state drive, the wear that accompanies modifications to flash memory may be avoided or reduced due to I/O operations being given effect through the modifications of the metadata representing the data targeted by the I/O operations instead of through the reading, erasing, or writing of flash memory. Further, as noted above, in such a virtualized storage system, the

metadata described above may be used to handle the relationship between virtual, or logical, addresses and physical, or real, addresses—in other words, the metadata representation of stored data enables a virtualized storage system that may be considered flash-friendly in that it reduces, or minimizes, wear on flash memory.

[000269] Leader storage systems may perform their own local operations to implement these descriptions in the context of their local copy of the pod dataset and the local storage system's metadata. Further, the in-sync followers perform their own separate local operations to implement these descriptions in the context of their separate local copy of the pod dataset and their separate local storage system's metadata. When both leader and follower operations are complete, the result is compatible graphs of logical extents with compatible leaf logical extent content. These graphs of logical extents then become a type of "common metadata" as described in previous examples. This common metadata can be described as dependencies between modifying operations and required common metadata. Transformations to graphs can be described as separate operations within a set of or more predicates that may describe relationships, such as dependencies, with one or more other operations. In other words, interdependencies between operations may be described as a set of precursors that one operation depends on in some way, where the set of precursors may be considered predicates that must be true for an operation to complete. A fuller description of predicates may be found within Application Reference No. 15/696,418, which is included herein by reference in its entirety. Alternately, each modifying operation that relies on a particular same graph transformation that has not yet been known to complete across the pod can include the parts of any graph transformation that it relies on. Processing an operation description that identifies a "new" leaf or composite logical extent that already exists can avoid creating the new logical extent since that part was already handled in the processing of some earlier operation, and can instead implement only the parts of the operation processing that change the content of leaf or composite logical extents. It is a role of the leader to ensure that transformations are compatible with each other. For example, we can start with two writes come that come in for a pod. A first write replaces a composite logical extent A with a copy of formed as composite logical extent B, replaces a leaf logical extent C with a copy as leaf logical extent D and with modifications to store the content for the second write, and further writes leaf logical extent D into composite logical extent B. Meanwhile, a second write implies the same copy and replacement of composite logical extent A with composite logical extent B but copies and replaces a different leaf logical extent E with a logical extent F which is modified to store the content of the second write, and further writes logical extent

F into logical extent B. In that case, the description for the first write can include the replacement of A with B and C with D and the writing of D into composite logical extent B and the writing of the content of the first write into leaf extend B; and, the description of the second write can include the replacement of A with B and E with F and the writing of F into composite logical extent B, along with the content of the second write which will be written to leaf extent F. A leader or any follower can then separately process the first write or the second write in any order, and the end result is B copying and replacing A, D copying and replacing C, F copying replacing E, and D and F being written into composite logical extent B. A second copy of A to form B can be avoided by recognizing that B already exists. In this way, a leader can ensure that the pod maintains compatible common metadata for a logical extent graph across in-sync storage systems for a pod.

[000270] Given an implementation of storage systems using directed acyclic graphs of logical extents, recovery of pods based on replicated directed acyclic graphs of logical extents may be implemented. Specifically, in this example, recovery in pods may be based on replicated extent graphs then involves recovering consistency of these graphs as well as recovering content of leaf logical extents. In this implementation of recovery, operations may include querying for graph transformations that are not known to have completed on all in-sync storage systems for a pod, as well as all leaf logical extent content modifications that are not known to have completed across all storage systems for the pod. Such querying could be based on operations since some coordinated checkpoint, or could simply be operations not known to have completed where each storage system keeps a list of operations during normal operation that have not yet been signaled as completed. In this example, graph transformations are straightforward: a graph transformation may create new things, copy old things to new things, and copy old things into two or more split new things, or they modify composite extents to modify their references to other extents. Any stored operation description found on any in-sync storage system that creates or replaces any logical extent can be copied and performed on any other storage system that does not yet have that logical extent. Operations that describe modifications to leaf or composite logical extents can apply those modifications to any in-sync storage system that had not yet applied them, as long as the involved leaf or composite logical extents have been recovered properly.

[000271] In another example, as an alternative to using a logical extent graph, storage may be implemented based on a replicated content-addressable store. In a content-addressable store, for each block of data (for example, every 512 bytes, 4096 bytes, 8192 bytes or even 16384 bytes) a unique hash value (sometimes also called a fingerprint) is calculated, based on the

block content, so that a volume or an extent range of a volume can be described as a list of references to blocks that have a particular hash value. In a synchronously replicated storage system implementation based on references to blocks with the same hash value, replication could involve a first storage system receiving blocks, calculating fingerprints for those blocks, identifying block references for those fingerprints, and delivering changes to one or a plurality of additional storage systems as updates to the mapping of volume blocks to referenced blocks. If a block is found to have already been stored by the first storage system, that storage system can use its reference to name the reference in each of the additional storage systems (either because the reference uses the same hash value or because an identifier for the reference is either identical or can be mapped readily). Alternately, if a block is not found by the first storage system, then content of the first storage system may be delivered to other storage systems as part of the operation description along with the hash value or identity associated with that block content. Further, each in-sync storage system's volume descriptions are then updated with the new block references. Recovery in such a store may then include comparing recently updated block references for a volume. If block references differ between different in-sync storage systems for a pod, then one version of each reference can be copied to other storage systems to make them consistent. If the block reference on one system does not exist, then it be copied from some storage system that does store a block for that reference. Virtual copy operations can be supported in such a block or hash reference store by copying the references as part of implementing the virtual copy operation.

[000272] For further explanation, Figure 5 sets forth a flow chart illustrating steps that may be performed by storage systems (402, 404, 406) that support a pod according to some embodiments of the present disclosure. Although depicted in less detail, the storage systems (402, 404, 406), communicating over data communications links (502, 504, 506), and depicted in Figure 5 may be similar to the storage systems described above with reference to Figures 1A-1D, Figures 2A-2G, Figures 3A-3B, Figures 4A and 4B, or any combination thereof. In fact, the storage systems (402, 404, 406) depicted in Figure 5 may include the same, fewer, additional components as the storage systems described above.

[000273] In the example method depicted in Figure 5, a storage system (402) may attach (508) to a pod. The model for pod membership may include a list of storage systems and a subset of that list where storage systems are presumed to be in-sync for the pod. A storage system is in-sync for a pod if it is at least within a recovery of having identical idle content for the last written copy of the dataset associated with the pod. Idle content is the content

after any in-progress modifications have completed with no processing of new modifications. Sometimes this is referred to as "crash recoverable" consistency. Storage systems that are listed as pod members but that are not listed as in-sync for the pod can be described as "detached" from the pod. Storage systems that are listed as pod members, are in-sync for the pod, and are currently available for actively serving data for the pod are "online" for the pod.

[000274] In the example method depicted in Figure 5, the storage system (402) may attach (508) to a pod, for example, by synchronizing its locally stored version of the dataset (426) along with an up-to-date version of the dataset (426) that is stored on other storage systems (404, 406) in the pod that are online, as the term is described above. In such an example, in order for the storage system (402) to attach (508) to the pod, a pod definition stored locally within each of the storage systems (402, 404, 406) in the pod may need to be updated in order for the storage system (402) to attach (508) to the pod. In such an example, each storage system member of a pod may have its own copy of the membership, including which storage systems it last knew were in-sync, and which storage systems it last knew comprised the entire set of pod members.

[000275] In the example method depicted in Figure 5, the storage system (402) may also receive (510) a request to read a portion of the dataset (426) and the storage system (402) may process (512) the request to read the portion of the dataset (426) locally. Readers will appreciate that although requests to modify (e.g., a write operation) the dataset (426) require coordination between the storage systems (402, 404, 406) in a pod, as the dataset (426) should be consistent across all storage systems (402, 404, 406) in a pod, responding to a request to read a portion of the dataset (426) does not require similar coordination between the storage systems (402, 404, 406). As such, a particular storage system (402) that receives a read request may service the read request locally by reading a portion of the dataset (426) that is stored within the storage system's (402) storage devices, with no synchronous communication with other storage systems (404, 406) in the pod. Read requests received by one storage system for a replicated dataset in a replicated cluster are expected to avoid any communication in the vast majority of cases, at least when received by a storage system that is running within a cluster that is also running nominally. Such reads should normally be processed simply by reading from the local copy of a clustered dataset with no further interaction required with other storage systems in the cluster

[000276] Readers will appreciate that the storage systems may take steps to ensure read consistency such that a read request will return the same result regardless of which storage system processes the read request. For example, the resulting clustered dataset content for

any set of updates received by any set of storage systems in the cluster should be consistent across the cluster, at least at any time updates are idle (all previous modifying operations have been indicated as complete and no new update requests have been received and processed in any way). More specifically, the instances of a clustered dataset across a set of storage systems can differ only as a result of updates that have not yet completed. This means, for example, that any two write requests which overlap in their volume block range, or any combination of a write request and an overlapping snapshot, compare-and-write, or virtual block range copy, must yield a consistent result on all copies of the dataset. Two operations cannot yield a result as if they happened in one order on one storage system and a different order on another storage system in the replicated cluster.

[000277] Furthermore, read requests may be time order consistent. For example, if one read request is received on a replicated cluster and completed and that read is then followed by another read request to an overlapping address range which is received by the replicated cluster and where one or both reads in any way overlap in time and volume address range with a modification request received by the replicated cluster (whether any of the reads or the modification are received by the same storage system or a different storage system in the replicated cluster), then if the first read reflects the result of the update then the second read should also reflect the results of that update, rather than possibly returning data that preceded the update. If the first read does not reflect the update, then the second read can either reflect the update or not. This ensures that between two read requests "time" for a data segment cannot roll backward.

[000278] In the example method depicted in Figure 5, the storage system (402) may also detect (514) a disruption in data communications with one or more of the other storage systems (404, 406). A disruption in data communications with one or more of the other storage systems (404, 406) may occur for a variety of reasons. For example, a disruption in data communications with one or more of the other storage systems (404, 406) may occur because one of the storage systems (402, 404, 406) has failed, because a network interconnect has failed, or for some other reason. An important aspect of synchronous replicated clustering is ensuring that any fault handling doesn't result in unrecoverable inconsistencies, or any inconsistency in responses. For example, if a network fails between two storage systems, at most one of the storage systems can continue processing newly incoming I/O requests for a pod. And, if one storage system continues processing, the other storage system can't process any new requests to completion, including read requests.

[000279] In the example method depicted in Figure 5, the storage system (402) may also determine (516) whether to the particular storage system (402) should remain online as part of the pod. As mentioned above, to be 'online' as part of a pod, a storage system must consider itself to be in-sync for the pod and must be communicating with all other storage systems it considers to be in-sync for the pod. If a storage system can't be certain that it is in-sync and communicating with all other storage systems that are in-sync, then it may stop processing new incoming requests to access the dataset (426). As such, the storage system (402) may determine (516) whether to the particular storage system (402) should remain online as part of the pod, for example, by determining whether it can communicate with all other storage systems (404, 406) it considers to be in-sync for the pod (e.g., via one or more test messages), by determining whether the all other storage systems (404, 406) it considers to be in-sync for the pod also consider the storage system (402) to be attached to the pod, through a combination of both steps where the particular storage system (402) must confirm that it can communicate with all other storage systems (404, 406) it considers to be in-sync for the pod and that all other storage systems (404, 406) it considers to be in-sync for the pod also consider the storage system (402) to be attached to the pod, or through some other mechanism.

[000280] In the example method depicted in Figure 5, the storage system (402) may also, responsive to affirmatively (518) determining that the particular storage system (402) should remain online as part of the pod, keep (522) the dataset (426) on the particular storage system (402) accessible for management and dataset operations. The storage system (402) may keep (522) the dataset (426) on the particular storage system (402) accessible for management and dataset operations, for example, by accepting requests to access the version of the dataset (426) that is stored on the storage system (402) and processing such requests, by accepting and processing management operations associated with the dataset (426) that are issued by a host or authorized administrator, by accepting and processing management operations associated with the dataset (426) that are issued by one of the other storage systems (404, 406) in the pod, or in some other way.

[000281] In the example method depicted in Figure 5, the storage system (402) may also, responsive to determining that the particular storage system should not (520) remain online as part of the pod, make (524) the dataset (426) on the particular storage system (402) inaccessible for management and dataset operations. The storage system (402) may make (524) the dataset (426) on the particular storage system (402) inaccessible for management and dataset operations, for example, by rejecting requests to access the version of the dataset

(426) that is stored on the storage system (402), by rejecting management operations associated with the dataset (426) that are issued by a host or other authorized administrator, by rejecting management operations associated with the dataset (426) that are issued by one of the other storage systems (404, 406) in the pod, or in some other way.

[000282] In the example method depicted in Figure 5, the storage system (402) may also detect (526) that the disruption in data communications with one or more of the other storage systems (404, 406) has been repaired. The storage system (402) may detect (526) that the disruption in data communications with one or more of the other storage systems (404, 406) has been repaired, for example, by receiving a message from the one or more of the other storage systems (404, 406). In response to detecting (526) that the disruption in data communications with one or more of the other storage systems (404, 406) has been repaired, the storage system (402) may make (528) the dataset (426) on the particular storage system (402) accessible for management and dataset operations.

[000283] Readers will appreciate that the example depicted in Figure 5 describes an embodiment in which various actions are depicted as occurring within some order, although no ordering is required. Furthermore, other embodiments may exist where the storage system (402) only carries out a subset of the described actions. For example, the storage system (402) may perform the steps of detecting (514) a disruption in data communications with one or more of the other storage systems (404, 406), determining (516) whether to the particular storage system (402) should remain in the pod, keeping (522) the dataset (426) on the particular storage system (402) accessible for management and dataset operations or making (524) the dataset (426) on the particular storage system (402) inaccessible for management and dataset operations without first receiving (510) a request to read a portion of the dataset (426) and processing (512) the request to read the portion of the dataset (426) locally. Furthermore, the storage system (402) may detect (526) that the disruption in data communications with one or more of the other storage systems (404, 406) has been repaired and make (528) the dataset (426) on the particular storage system (402) accessible for management and dataset operations without first receiving (510) a request to read a portion of the dataset (426) and processing (512) the request to read the portion of the dataset (426) locally. In fact, none of the steps described herein are explicitly required in all embodiments as prerequisites for performing other steps described herein.

[000284] For further explanation Figure 6 illustrates a configurable replication system that provides continuous replication with minimal batching and an adjustable recovery point objective. In contrast to the example storage systems described with reference to Figure 5,

which describes use of pods in implementing synchronous replication, in this example, pods are used for asynchronous, or near-synchronous replication.

[000285] However, as described further below, while replication may be asynchronous, efficient use of lightweight journals, also referred to as metadata logs, allows for a short, typical recovery point (the time difference between last update on a source data repository and the clock value of the source data repository associated with the latest consistent dataset available at a target data repository) that can be on the order of a few to 50 or 100 milliseconds, or a short intrinsic or configured recovery point objective (RPO), where the RPO represents a measurement of the maximum tolerable amount of data to lose. For example, in some cases, the RPO may specify that losing data written within the last ten milliseconds may be tolerable, whereas in other cases the RPO may specify that losing data written within few minutes may be tolerable. In some examples, the RPO limit may be more of a function of a typical maximum transfer time. As an illustrative scenario, the earth's moon is a little over one light-second away from the earth, so with sufficient bandwidth to avoid queue delay, an RPO to the moon of 1.2 seconds is possible with a lightweight journal implementation (receiving an acknowledgement from the moon for the primary to confirm the recovery point will take at least another 1.2 seconds).

[000286] In some implementations, the configurable replication system provides for disaster recovery from a failure at a source data repository based on a target data repository being able to provide read and write access with a consistent version of the source data repository in response to the failure of the source data repository. As an example, consider a set of clock values associated with an original dataset that is being updated, where a *source time* represents a clock value for the source dataset, and *includes* all updates which were signaled as completed on the original dataset prior to that time and *excludes* all updates which were received to be processed against the dataset after that time. In this example, any updates which were received to be processed against the dataset at the source time but had not yet been signaled as completed can in general be arbitrarily included or excluded barring any transactional interdependencies.

[000287] Further, a snapshot may represent one such source time for a dataset, and where rolling lightweight checkpoints may represent a sequence of dataset source times. In near-sync replication, checkpoints may be applied as they come in or when they are completely ready to be applied. As a result, in some examples, a tracking dataset always represents some replicated source time clock value which is generally some amount behind the live dataset's source time clock value. In this example, the difference between the replicated dataset source

time clock value and the live dataset source time clock value may be reported as the current available "recovery point" — the distance between the replicated dataset source time clock value and the live dataset source time clock (though propagation delays likely mean that neither source nor target know exactly what this time distance is).

[000288] In some implementations, the lightweight journals may be a basis for implementing continuous data protection — with or without any implementation of data replication. In some examples, continuous data protection provides relatively fine-grained versioning of a dataset for extended periods of time, to allow roll-back or other access to any of those fine-grained versions. For example, these versions can be examined to determine when some update or corruption occurred, allowing a roll-back or other access (such as the formation of a usable snapshot or clone) to the version immediately prior to that update. In some cases, it makes sense to provide access to both the pre-change/pre-corruption dataset as well as the more recent data (or even a set of points-in-time of the dataset before or since the time of the update/corruption) so that other changes can be copied or otherwise reconciled, or for diagnostic purposes.

[000289] Further, continuing with this example, in continuous data protection, checkpoints of a dataset may be replayed up to some limit in order to construct a consistent image. In some cases, such checkpoints may be transformed into a read-only snapshot, or the dataset may also be cloned (or the read-only snapshot may be cloned) to form a read-write volume that may be used for various purposes. In this example, an implementation of continuous data protection may clone a volume to match some point in time, test it to determine whether the volume includes or excludes some data or some corruption, and then if needed re-clone the volume to match some other point in time and test the volume again. In this example, when a point-in-time is determined, that point-in-time may be used as a basis to generate a primary volume or simply copy data out of the volume at that point-in-time.

[000290] Further still, in some implementations, continuous data protection may provide more granular access to these named source time clock values from the source dataset, with granularity limited to the granularity of checkpoints. In some cases, continuous data protection could be either local (the checkpoints are retained on a local storage system and are available for local access), or they can be on a replication target (the checkpoints are retained on a replication target), or both, with each possibly having different retention periods and models for merging checkpoints or converting them to long-duration snapshots.

[000291] In some implementations, a 'pod', as the term is used here and throughout the present application, may be embodied as a management entity that represents a dataset, a set

of managed objects and management operations, a set of access operations to modify or read the dataset, and a plurality of storage systems. Such management operations may modify or query managed objects through a storage system with proper access. Each storage system may store a separate copy of the dataset as a proper subset of the datasets stored and advertised for use by the storage system, where operations to modify managed objects or the dataset performed and completed through any one storage system are reflected in subsequent management objects to query the pod or subsequent access operations to read the dataset.

[000292] In some implementations, a replication relationship is formed as a set of storage systems 602, 624 that replicate some dataset 612 between independent stores, where each storage system 602, 624 may have its own copy and its own separate internal management of relevant data structures for defining storage objects, for mapping objects to physical storage, for deduplication, for defining the mapping of content to snapshots, and so on. In this way, a replication system may use a common management model that is a same, or similar, management model, and use a same, or similar, implementation model and persistent data structures for both synchronous replication and asynchronous replication.

[000293] As illustrated, a source data repository 602 receives storage system operations 652 and may communicate with a target data repository 624 to generate replica data. In this example, the source data repository 602 may be similar to computing device 350 or similar to a storage system 100, 306, 318, as described above with reference to Figures 1A–3D. While exemplary systems are depicted in Figure 6, the components illustrated in Figure 6 are not intended to be exhaustive or limiting.

[000294] As noted above, incoming data storage operations 652 may be received and handled by the source data repository 602, and the data storage operations that update or modify a volume 658, or more generally, modify one or more datasets, may be streamed or transmitted to the target data repository 624 as the data storage operations arrive. In other words, the source data repository 602 may be considered 'active' in that the source data repository 602 accepts write operations and other operations that may modify the stored data, where the target data repository 624 may be considered 'passive' in that the target data repository 624 may accept read operations, but not storage operations that may modify the stored data.

[000295] In this example, the source data repository 602 maintains a metadata log 604, which may be referred to as a journal of modifying data storage operations ordered by checkpoint. In some cases, a journal may equivalently be referred to as a lightweight journal due to the journal including only metadata information, but little or no storage data provided by a user to be stored. In some examples, the metadata log 604 may be generated or updated during a

flush of storage data from NVRAM to a backend bulk storage — where a storage system architecture with NVRAM, and example backend bulk storage, are described above with reference to Figure 2D. In some examples, the metadata, such as checkpoints 604, may be stored in the source data repository 602 as metadata, without being included within a journal, or metadata log structure, where the journal, or metadata log 604 may be constructed on demand, such as in response to one or more checkpoints being ready for transmission to a target data repository 624.

[000296] In some implementations, a checkpoint may also be referred to as an ordered "lightweight checkpoint" of a dataset. A lightweight checkpoint is a data structure that includes metadata describing a set of updates to one or more datasets stored within a data repository. Such a checkpoint is referred to herein as being 'lightweight' because the data structure only includes a reference the actual data associated with a corresponding set of updates (e.g., the data structure only includes a reference to data that was written to a dataset when performing a particular update rather than containing the actual data that was written to the dataset as part of performing the particular update) by holding references to where the actual data associated with a corresponding set of updates is stored. A given set of updates may begin to be staged in NVRAM, or a first tier of a storage system's storage, before the set of updates, or at least a portion of these of updates is flushed to backing storage, or a second tier of the storage system.

[000297] However, in this example, the data referenced by a set of updates within a given checkpoint may survive logical (or address range) overwrites and garbage collection and is not duplicated into a separate metadata journal. Further, lightweight checkpoints may be ordered in that to arrive at a complete and consistent point-in-time image of some point in time of the original dataset, each set of updates described in each lightweight checkpoint between some prior consistent image and the point in time corresponding to a particular lightweight checkpoint should either be applied to form that point-in-time image or the update could be determined to be unnecessary, for example, being due to an overwrite or deletion. In some examples, lightweight checkpoints may be merged, which can be beneficial because merging may release some backing store data that has been overwritten or deleted, for example by having been written in an earlier checkpoint and overwritten in a later one that is merged with the earlier one (in which case the data for the earlier write may no longer be needed), thereby allowing some otherwise held data to be garbage collected.

[000298] Continuing with this example, such lightweight checkpoints are intended to represent very fine-grained consistency point moments in time as consistency points, with

each lightweight checkpoint associated with a set of updates that have been signaled as completed, excluding a set of updates whose processing has not yet started, and potentially including or excluding updates that are concurrent with the moment in time the checkpoint represents. In some examples, formation of a new lightweight checkpoint or a duration, or period, between two checkpoints may be based on time slices, such as every few milliseconds, or operation count slices, such as every 50 to 500 update operations, or based on transfer size or some more complex relationship to update operations, such as counting a few megabytes of modifications or some number of logical extent updates, or they can relate to some explicit operation, such as an operation to explicitly tag or name a particular point-in-time so it can be referenced later such as by a program noticing or being notified when it is received and applied by to replication target, any combination of these and other triggers. Such tags or names could also be searched for within a continuous data protection implementation.

[000299] In some implementations, lightweight checkpoints may differ from snapshots in that they do not affect the durable structure of the storage system beyond whatever side structure is used to store them, apart from the garbage collection or overwrite holds, and lightweight checkpoints may be discarded with minimal effect, other than the release of those garbage collection or overwrite holds. Further, in some cases, lightweight checkpoints may also lack individual administrative handles, perhaps apart from lightweight checkpoints that are explicitly tagged or named. In some example, lightweight checkpoints exist almost exclusively as an ordered list of metadata bundles describing updates, where the ordered list of metadata may be stored in a log-style structure. Further, lightweight checkpoints may be persistent or not persistent, in dependence at least upon an intended use of the lightweight checkpoint. In particular, near-sync replication may have crash or resynchronization recovery mechanisms that may operate independently of lightweight checkpoints and that may then not require persisting of lightweight checkpoint logs, while the target of replication might separately benefit from persisting checkpoints on the target storage system for fault recovery purposes, such as part of making application of lightweight checkpoints atomic.

[000300] In some implementations, if the metadata for a lightweight checkpoint represents logical composite and leaf extents, as described in prior patents, then a lightweight checkpoint may be a set of descriptions for updating these logical composite and leaf extents which are themselves metadata descriptions that reference stored data by content identifier references In some cases, use of content identifiers irrespective of the use of an extent model may also be beneficial in that such use preserves information about duplicates and may be

used as part of a strategy to avoid transfer of content that a target storage system may already be known to store. For further clarification, these prior patents include, U.S. Patent Serial Nos. 16/050,385, 62/598,989, and 15/842,850, which are incorporated herein for all purposes.

[000301] Continuing with this example, the structure of a metadata representation of a dataset may be particularly effective in a Flash storage system because Flash does not allow overwrite in place at the chip level and may generally be driven, at some level, by garbage collection algorithms that can readily account for a wide variety of references that have holds on written data. In some cases, some details may account for the NVRAM aspects which do not have to follow the same write-elsewhere-with-garbage-collection model, but at least the bulk data writes for lightweight checkpoints are not separate writes that require separate storage.

[000302] In some implementations, and as described in other sections of this reference, some applications of lightweight checkpoints may include normal operation of near-sync replication (in contrast to initialization or resynchronization), which may also be referred to as asynchronous replication. In this example, lightweight checkpoints may be transferred over a network link to some target repository that may then apply the lightweight checkpoints to a tracking copy of the original dataset, with lightweight checkpoints (and their referenced data) being held at least until the tracking copy has been updated.

[000303] In some cases, if checkpoints may be received or applied out-of-order, then all intermediate checkpoints may need to be received and applied before the lightweight checkpoint on the source system can be released. Generally, lightweight checkpoints should be applied atomically, such as by using some transaction mechanism. One transaction mechanism is to receive the metadata for a lightweight checkpoint, receive all the data content for a lightweight checkpoint and storing it locally on the target, and then roll forward the tracking copy to incorporate the metadata updates in the lightweight checkpoint with its data references updated to reference the data content stored locally on the target.

[000304] Further, other applications of lightweight checkpoints may include:

- In some examples, a tracking copy may be converted into a snapshot or a clone to provide a stable image at some point in time, thereby allowing use of a point-in-time image for testing purposes or failover purposes;
- In some examples, if a source-to-target interconnect and the target storage repository are not roughly keeping up with the rate that the source storage system itself is receiving data, storing it, and forming and transferring lightweight checkpoints, then

these lightweight checkpoints can start building up. In this scenario, there are several reactions to this that can be used: lightweight checkpoints could be merged to reduce their cost (the source dataset points-in-time associated with named or tagged checkpoints might be preferentially retained); back pressure could be put on the source storage system to reduce the rate at which it receives, processes, or completes updates; a subset of checkpoints could be converted to more durable snapshots; or lightweight checkpoint-based replication could be discarded in favor of replication based on periodic snapshots. In some cases, some number of periodic snapshots might already be kept for resync or connection loss/reconnect purposes so switching to snapshot replication may already be fully ready to go — meaning that lightweight checkpoints since the last snapshot may simply be discarded if replication is not keeping up sufficiently for the lightweight snapshots to be useful (further clarification may be found within U.S. Patent Serial No. 15/842,850, which is incorporated herein for all purposes);

- In some examples, connection loss or other kinds of interruptions to replication may generally be handled by switching to some other scheme, such as snapshot based replication, or by using a resync model similar to what is described for synchronous replication recovery, though without the need to catch all the way up at the very end;
- In some examples, the transfer of data can be initiated by the sender side by simply sending the referenced data to the target storage system along with sending the lightweight checkpoint metadata updates. Further, the transfer of data may instead be initiated by the target storage system: if the lightweight checkpoint metadata lists content identifiers, then the target storage system can reuse references to content it already stores but can then request retrieval of content it does not current store. This can reduce total bandwidth required, though if the network link has to be sized for the update rate, the benefit may be low; and
- In some examples, if the source storage system itself stores content compressed as some kind of compressed blocks, then the compressed blocks may in many cases be transferred directly rather than being uncompressed and then possibly recompressed before being transmitted over the network.

[000305] In some implementations, lightweight checkpoints may be used to implement continuous data protection either on the original storage system — with or without replication being involved — or on a replication target system by storing the lightweight checkpoints on

the target storage system rather than simply applying and then discarding them. In continuous data protection, various point-in-time images of a dataset can be accessed by rolling forward a copy of a dataset to include all lightweight checkpoints up to the lightweight checkpoint corresponding to some source dataset point-in-time of interest.

[000306] For example, if the storage system also implements durable snapshots, then only lightweight checkpoints since the point-in-time of the most immediately prior snapshot may need to be applied. Generally, higher granularity is more interesting for more recent history of a dataset and less granularity is needed farther back, allowing for the possibility of ever more aggressive lightweight checkpoint merging as points-in-time recede, or eventually discarding them in favor of less frequent snapshots.

[000307] Further, if continuous data protection is used to locate a point in time just before where an unwanted change or corruption was introduced, then relatively fine grained lightweight checkpoints (milliseconds to a few seconds to every few minutes) might only need to be kept until plenty of time has elapsed to ensure that corruption will have been noticed and recovery procedures started. After that, 30 minute or hourly or even daily snapshots might be preferable (or such rollbacks may be considered unnecessary). Any specific lightweight checkpoint can be converted into a durable snapshots if snapshots hadn't been created explicitly. If lightweight checkpoints can be named or tagged, continuous data protection could support locating and accessing those named lightweight checkpoints.

[000308] In some implementations, as noted below, under some storage system conditions, or in response to a user-specified configuration, near-synchronous replication may be transitioned to different type of replication, including periodic replication or synchronous replication. Further, in some implementations, a source data storage system may implement synchronous replication that is pod-based among a cluster of storage systems, but where one or more of the source data storage systems also implement lightweight checkpoints for near-synchronous replication with a target storage system that may be initiated in the event of a communication fault with the other storage system in the cluster of storage systems — thereby allowing the source storage system to maintain both synchronous data replication over near distances and to maintain data resiliency over longer distances. Further, in some examples, RPO may be configurable, where the time or operation size of lightweight checkpoints may be configured or adjusted based on, at least, available network bandwidth or supporting flow-control (as discussed above). In some cases, if a set of synchronously replicating storage systems exchange checkpoint information between them as part of their operation, then near-synchronous replication can operate and continue from any of the

storage systems that synchronously replicate the checkpoint information, including continuing after the failure of one such storage system, including parallel transfer of data and metadata from multiple of the synchronously replicating storage systems. Such parallel data transfer could, for example, involve the target of near-synchronous replication requesting data for referenced composite or logical extents or content identifiers from any set or subset of the synchronously replicating storage systems.

[000309] Further, in some implementations, an addition to near-synchronous replication is short-distance synchronous replication of metadata and data updates, combined with longer-distance non-synchronous replication of lightweight checkpoints. In such an example, this may provide what is sometimes called "bunker" replication where a storage system within synchronous replication distance is sized to store enough for in-transit data and metadata but is not sized to store a complete dataset. In this example, if the primary (complete) copy fails but the intermediate "bunker" storage survives, then the further distant non-synchronous target can be caught up by applying the updates that were stored synchronously on the bunker storage. Further, in this example, if both primary and bunker storage fail, then at least the longer-distance storage is consistent and within the longer distance RPO. Continuing with this example, the lightweight checkpoints may be formed and transferred by either the bunker storage system or by the primary storage system, or can be formed and transferred by a combination of the primary storage system and the bunker storage system.

[000310] In some implementations, a metadata log 604 schema may be sorted by (pod, checkpoint), which allows for traversal in a correct order, where a same schema may be used on both a source data repository 602 and a target data repository 624. In this example, a write operation may be encoded in a metadata log 604 by indicating both a physical extent identification along with address information of all writes for a given checkpoint. Further, in some cases, a metadata log 604 may contain operations to modify a metadata representation 614 of the dataset that correspond to system operations, such as copy-on-write (CoW). For example, modifications to a metadata representation 614 may include modifications due to an XCOPY, WSAME, snapshots, CoW, among others. An example of such operation-style metadata may include a sequence of updates to logical and composite extents, with any written content tied to a checkpoint being retained at least until the checkpoint is no longer needed for replication or other purposes. In this case, the metadata log may contain the logical and composite logical extent updates including references to any stored data, with the stored data being a held reference to the content stored in the storage system for its regular use but with any garbage collection or overwrite held off as long as the checkpoint is

retained. Further, in some cases, content overwrites within a checkpoint (including within merged checkpoints if checkpoint merging is supported) may discard the hold on the earlier content replaced by later content described by the checkpoint. In some examples, a metadata log 604 may include metadata representation 614 identifier allocations on a source data repository 602, which allows the target data repository 624 to avoid trying to look up content identifiers that do not exist on the target data repository 624.

[000311] In different embodiments, the lifetime of checkpoint entries 606a, 606b may be configurable to allow for different options for data recovery, including a lifetime extending for an ongoing length of storage services that allows for continuous data protection. In this example, the configurable replication system may provide continuous replication, where as data storage operations that modify a volume or dataset arrive, the storage operations are grouped into checkpoints, and where a given checkpoint may include varying numbers of storage operations. In some examples, a given checkpoint may include metadata for up to 100 storage operations. As noted herein, because a garbage collection process may keep stored data based on references to the stored data location being referenced by either general storage system references within the storage system's general metadata or by a metadata log that includes checkpoints, then the length of the lifetime of the checkpoints corresponds to a length of time for a recovery window for continuous data protection.

[000312] In this example, a checkpoint may be considered a smallest unit of data consistency, where if the metadata log 626 received at the target data repository 624 includes a particular checkpoint, then a replica dataset 634 that is generated by replaying the storage operations in the particular checkpoint will include all storage operations from all checkpoints that were generated prior to the particular checkpoint — and such a policy provides for a crash consistent recovery point for the replica dataset 634. Further, if there is a snapshot that is from a point-in-time earlier than the desired replay point, then only replay checkpoints since that snapshot may be needed during a recovery. In this example, checkpoints may be merged to allow garbage collection of overwritten data, and checkpoints may also be periodically converted to snapshots if that results in a cleaner format or a better or simpler relationship with garbage collection.

[000313] In some implementations, snapshots may be used to coordinate a point in time in the update stream. For example, an application can make some update then issue a snapshot request, and if snapshots are a type of update that is replicated, then when the snapshot appears on the target storage system, that point in time for the application is present. In this example, this could be generalized to some kind of tag, such that a snapshot is not necessarily

needed. Further in this example, if some metadata tag is set on a dataset, or on some component within a dataset, and that tag is handled as a type of update within the log/checkpoint model, then a monitoring program on the target storage system could detect when that point in time of the source dataset has been reached on the target by noticing the appearance of the tag. The storage system could further support a means of notifying programs waiting for such snapshots or named or tagged checkpoints being received and processed on a target storage system. Yet further, when the target storage system has received and processed such snapshots or named or tagged checkpoints, it could send a notification back to the source storage system, which could then, in turn, notify interested programs that the snapshot or named or tagged checkpoint is known to have been received and processed on the target system. Continuing with this example, such a process could be used, for example, by a program running against the source storage system that updates some data, tags a checkpoint, and then takes some action when notified by the source storage system that the tagged checkpoint (and thus the update) is known to have been replicated. For example, a high level task could perform a set of updates which are replicated, and where the action taken is that aspects of the continue only after receiving that notification. In some cases, this in turn allows higher level tasks to be replicated effectively synchronously across long distances even when performing many smaller operations that are not themselves replicated synchronously. For example, a web application might use this to ensure that some requested update to, for example, a user profile is durable across distances before a web page shows the durable change to the user profile.

[000314] While in this example, replication is described in the context of replicating a "volume", in general, the described replication techniques may be applied to any generalized dataset. In other words, in the general case, replication applies to a dataset, which may include one or more volumes, and/or one or more other types of data or collections of data, at a given point in time. In some cases, a dataset may be a dataset specified by a pod, where in a pod the actual set of volumes may change as volumes are added to and removed from the pod, and tracking will reflect that by adding and removing volumes. Further, in some examples, continuous data protection of a pod may result in volumes existing or not existing based on which checkpoint we roll backward to or forward to, and on the volume membership at the pod's source time for that checkpoint.

[000315] Continuing with this example, each incoming write operation may be persisted as described above with reference to Figures 1A–3D, where in addition to the source volume 658 being updated, a reference to the storage location of the data corresponding to the write

operation is added to the metadata log 604. In this way, the metadata log 604 may serve as a buffer that allows recovery after a network outage and support bursts of write operations without impeding the reception or handling of storage operations by the source data repository 602. In this example, as checkpoints 606a, 606b are completed and created within the metadata log 604, the checkpoints 606a, 606b may be replicated to the target data repository 624 by, for example, transmission of one or more messages that include metadata log 608 using a standard communication protocol over one or more networks 659. In this example, independent of the transmission of the metadata log 604, the source data repository 602 may also transmit data 610 corresponding to the checkpoints 606 within the metadata log 604.

[000316] In some implementations, as checkpoints are created within the source data repository 602, a monitoring service may monitor which checkpoints are complete, and determine where the checkpoints may be read. In some examples, a checkpoint may be created as a checkpoint is written into NVRAM, or a first tier of fast data storage. In some cases, the monitoring service may provide an interface for accessing checkpoint data from NVRAM or from a given storage location.

[000317] Continuing with this example, a target data repository 624 may open one or more forwarding streams from the source data repository 602, where on the source data repository 602, each forwarding stream may claim a number of checkpoints from the monitoring service. In this example, a given forwarding stream may fetch metadata log 614 information for one or more checkpoints 606. Similarly, in this example, a given forwarding stream may fetch corresponding storage data for the one or more checkpoints 606. In this way, one or more communication channels may be opened, in some cases in parallel, between the source data repository 602 and the target data repository 624, where the one or more communication channels operate to transfer the metadata log 614 and corresponding data 612 between the source data repository 602 and the target data repository 624.

[000318] In this example, in response to receiving the metadata log 608, the target data repository 624 may persist the checkpoints 628a, 628b into a local metadata log 626. Based on a successful write of the checkpoints 628 into the local metadata log 626, the target data repository 624 may respond to the source data repository 602 with an acknowledgment, where in response to the acknowledgment, the source data repository 602 may — in dependence upon a configuration setting — delete or maintain the checkpoints 606a, 606b.

[000319] In some examples, the target data repository 624 may periodically, or in response to receiving metadata log 626 information from the source data repository 602, replay the

storage operations within the checkpoints 628 to generate and update a tracking volume 632. In some examples, replaying the storage operations may include converting metadata log 626 information into regular formatted metadata for the storage system and converting global content identifiers into local content identifiers; for example, such converting may include mapping content identifiers between the source data repository 602 and the target data repository 624. In this example, a metadata representation 630 may be implemented similarly to metadata representation 614 on the source data repository 602, where the physical location information may be different based on use of physical storage on the target data repository. In some examples, a tracking volume may also be referred to as a "shadow" volume.

[000320] In some implementations, content identifiers may be used to mark written content, including content that the source has already determined was a duplicate of other content that the source knows of (for example, through tracking the source write history, source snapshot history, source virtual copy history, and/or any local duplicate detection). In some examples, the content identifiers may be leveraged when doing recovery, such as after an extended outage, or when converting from near-sync replication to periodic or asynchronous replication.

[000321] In some implementations, delivery of a checkpoint as a set of metadata updates and content identifiers may result in the target storage system noticing which content identifiers the target storage system is already aware of and already stores — the target storage system may then request from the source storage system any content whose content identifiers the target storage system does not already store or is not already aware of. In some cases, except at moon-level distances, checkpoint delivery may still result in sub-second RPOs, and may also reduce data transfer bandwidth if duplicates are common. Further, in this example, until all missing content has been requested and received by the target storage system, the checkpoint may not be considered completed so the checkpoint may not be deleted to allow garbage collection.

[000322] In some examples, the tracking volume 632 is generated in response to a promotion event, where a promotion event may be based on a detected failure, detected impending failure, or detected degradation of responsiveness beyond a compliance policy threshold of the source data repository. In some cases, the promotion event may be automatically generated based on such a detection of a promotion event, and in other cases, the promotion event may be responsive to a user specifying that the replica data on the target data repository 624 be promoted.

[000323] In some implementations, a user may promote a tracking volume 632 in order to use a replica of the source data for different uses, such as for testing — where testing may include modification of the replica data in the tracking volume 632. However, based on a promotion event generating a replica volume 634, any modifications or corruption to the tracking volume that may occur during testing may be undone or reversed by referencing the replica volume 634. In this example, promotion of the tracking volume 632 also includes configuration filtering and/or reconciliation as part of making the tracking volume 632 a new volume available for use by a computational process, a computing device, or a compute node. Further, demotion or deletion of a volume may cause a host to reconfigure a connection to continue to access replica data on the target data repository 624.

[000324] While in some implementations, received metadata log 608 information may be played to generate the tracking volume 632 without storing the metadata log 608, or keeping a stored metadata log 626, the stored metadata log 626 may serve as a basis for providing data consistency guarantees described above with regard to the storage operations in a checkpoint.

[000325] Further, separating the generation of the tracking volume from dependence upon checkpoints as they are received, and instead generating the tracking volume from stored checkpoints supports receiving checkpoints out of order and the option to order the checkpoints prior to building the tracking volume 632. In other words, checkpoints may be transmitted and received out of order, but in general, checkpoint may not be applied out of order, so in some cases applying the checkpoints to a tracking dataset or volume may be delayed until intervening checkpoints are received. This example may be generalized as requiring that all intermediate checkpoints be received before the tracking dataset or volume may be advanced to the time associated a received dataset (irrespective of how checkpoint updates are actually applied).

[000326] Further, in this example, if for some reason, such as a recovery event on the source data repository 602 based on data loss or based on a user or application requesting access to the replica volume or based on a failover request to begin using the replica volume 634 as a primary or user-accessible volume, then the target data repository 624 may promote, or activate, the replica volume 634. In response, the existing checkpoints in the metadata log 626 may be replayed to generate a version of the tracking volume 632 consistent with a most recent checkpoint received, and the tracking volume 632 may be used to create a version of the source volume 658.

[000327] In some examples, in response to a recovery event—such as a source data repository 602 losing a connection with a host computer (not depicted) or applications sending storage operations, performance degradation beyond a threshold value, storage capacity exceeding a threshold value, or a degradation in response times—the target data repository 624 may be promoted to handle all further storage operations from the host computer, and another data repository may be selected. In this example, the replica link from the original source data repository 602 to the target data repository 624 may be reconfigured to flip directions, where the target data repository 624 becomes a new source data repository and another data repository becomes a new target data repository, and where other replica link characteristics stay the same.

[000328] The continuous replication from the source data repository 602 to the target data repository 624 may also be described in terms of pods, where pods and pod characteristics are described above with reference to Figures 4 and 5. As noted above, where Figure 5 describes use of pods in implementing synchronous replication, in this example, pods are used for asynchronous, or near-synchronous replication. In other words, in this example, source volume 658 may be included within a pod 640, and the replica volume 634 may be included within pod 642. In this way, in response to an indication that a user or application intends to use the replica data, and the tracking volume 632 being promoted, the replica pod 642 is updated with the most current contents from the tracking volume 632. While in this example a pod is depicted as include a single volume, in other examples, a pod may generally hold any type and quantity of data, including multiple volumes and/or multiple structured or unstructured datasets.

[000329] Further, in some implementations, as discussed above, there may be a dynamic relationship of volumes to pods, where the dynamic collection of volumes within a pod may be related to a clock value within the update stream on a source storage system. For example, a checkpoint may introduce volumes to a pod, change volume characteristics (name, size, etc.) and may remove volumes. In this example, if there are protection groups or some similar organizational concept within a pod, then these protection groups may also change with those changes being propagated through checkpoints. In this way, a near-sync target storage system may actually take over relatively seamlessly as a periodic replication source with all relationships intact, minus whatever time difference the last processed checkpoint is from the previous active source. In short, in some cases, it is the unified nature of the metadata model between synchronous replication, near synchronous replication (near-sync), and periodic replication (or asynchronous) replication, coupled with the local-to-global-to-

local content identifier and logical and composite extent identifier transformations that provides improvements to various aspects of a storage system and of a storage system replication process.

[000330] As depicted in Figure 6, a data repository 602 stores both data 612 from incoming storage operations 652, and a metadata representation 614 of the data 612. In this example, a metadata representation 614 may be implemented as a structured collection of metadata objects that, together, may represent a logical volume of storage data, or a portion of a logical volume, in accordance with some embodiments of the present disclosure. Metadata representation 614 may be stored within the source data repository 602, and one or more metadata representations may be generated and maintained for each of multiple storage objects, such as volumes, or portions of volumes, stored within the data repository 602.

[000331] In other examples, other types of structured collections of the metadata objects are possible; however, in this example, metadata representations may be structured as a directed acyclic graph (DAG) of nodes, where, to maintain efficient access to any given node, the DAG may be structured and balanced according to various methods. For example, a DAG for a metadata representation may be defined as a type of B-tree, and balanced accordingly in response to changes to the structure of the metadata representation, where changes to the metadata representation may occur in response to changes to, or additions to, underlying data represented by the metadata representation. Generally, metadata representations may span across multiple levels and may include hundreds or thousands of nodes, where each node may include any number of links to other nodes.

[000332] Further, in this example, the leaves of a metadata representation may include pointers to the stored data for a volume, or portion of a volume, where a logical address, or a volume and offset, may be used to identify and navigate through the metadata representation to reach one or more leaf nodes that reference stored data corresponding to the logical address. Data objects may be any size unit of data within the data repository 602. For example, data objects may each be a logical extent, where logical extents may be some specified size, such as 1MB, 4MB, or some other size, such as a system-specified block size.

[000333] In some implementations, as described above with reference to Figures 1A–3D, the data repository 602 may include multiple types of data storage, including NVRAM and Flash storage, where NVRAM may be used as a staging area for incoming storage operations, and where Flash storage may provide long-term, durable storage. In this example, the source volume 658, or portions of the source volume 658, may be stored in NVRAM, and the entire source volume may be stored within Flash memory, or as depicted in Figure 6, data store 660.

[000334] In some implementations, the metadata log 604 is ordered according to checkpoints, and is a journal, or log, describing all changes to stored data, and where checkpoints within the metadata log 604 that have not already been transmitted to the target data repository 624 are transmitted in response to generation or completion of a single checkpoint, or a set of checkpoints, in dependence upon a target RPO. For example, depending on a size of a checkpoint, or a quantity of data-modifying operations described by the checkpoint, more frequent transmission may be made in dependence upon a lower target RPO.

[000335] Further, as described above, checkpoints 606 within the metadata log 604 may include references to stored content such as blocks within data store 660 where that stored content consists of what the storage system would have stored were it not for the replicated checkpoint. In this way, the storage required for the metadata log and checkpoints is reduced considerably versus what would be required for a complete log of all updates that includes both metadata and a duplicate copy of data that was being written to the source storage system. In some examples, a service, or process, or controller, operating on the source data repository 602 may monitor creation of checkpoints, and forward or transmit the checkpoint, or set of checkpoints, to the target data repository 624.

[000336] In some implementations, references within checkpoints, and as a consequence, references within a metadata log, may refer to objects or data stored on the source data repository 602 that have been modified by subsequent storage operations, but where the data stored on the source data repository 602 has not yet been transferred to the target data repository 624. In such a scenario, if a garbage collection process on the source data repository 602 relies only on a reference table maintained by a storage controller managing data within the source data repository 602, then the garbage collection process may delete data or reallocate or otherwise overwrite a storage location that results in data referenced by a metadata log becoming unavailable or no longer valid as a source of content for a replicated checkpoint, thereby compromising the replication. To overcome such a scenario, in some examples, a garbage collection process on the source data repository 602 may reference both a reference table maintained by a storage controller or source data repository 602 process and also a list of references held by lightweight checkpoints, and specifically, a list of references within one or more checkpoints within a metadata log. Over time, checkpoints can be merged together to allow some overwritten content to be released for garbage collection.

[000337] In this way, based at least on both sources of data references — system references and metadata log references — a garbage collection process may preserve data that has not

yet been replicated, but would otherwise be modified or deleted by subsequent storage operations. Such data preservation during garbage collection also holds true for continuous data protection, when checkpoints are retained on a source storage system for some period of time in order to allow for flexible rollback, where the period of time may be configurable to an arbitrary quantity of time. In other words, a garbage collection process may determine that content at a storage location is not needed, and may be reclaimed or garbage collected, based on the content at the storage location not being referenced by any checkpoints in a metadata log or referenced by a storage system reference table.

[000338] In some implementations, as noted above, each checkpoint is exclusive of every other checkpoint, and based on the checkpoints being ordered, the checkpoints may be transmitted in any order to the target data repository 624. In this example, on the target data repository 624, the checkpoints are applied, or replayed, in order to create a consistent version of the data stored on the source data repository 602. In some cases, the data transmitted from the source data repository 602 to the target data repository 624 may be read from data storage within data store 660, for example if the data has been flushed from the NVRAM to Flash, or from the NVRAM, for example if the data continues to be stored in the NVRAM.

[000339] In some implementations, depending on configuration settings with respect to RPO, data may remain on the source data repository 602 for more or less time. In some cases, the longer that data remains on the source data repository 602, the greater the opportunity to perform transformations that may reduce the quantity of data transferred to the target data repository 624. For example, incoming data may be deduplicated, or overwrite previously written data, or may be deleted, among other operations or transformations, which may reduce the quantity of data that is transferred from the source data repository 602 to the target data repository 624.

[000340] In some implementations, the messaging mechanisms may be implemented similarly to the messaging mechanisms described above for synchronous data replication, with reference to Figures 4 and 5.

[000341] For further explanation Figure 7 illustrates a configurable replication system that provides continuous replication with minimal batching and an adjustable recovery point objective. In this example, a management object that specifies a replication policy between a source pod and a replica pod may be referred to as a "replica link".

[000342] A replica link specification may include a specification for a source of data for replication and a target for replica data, including storage data, checkpoints, metadata

representations, or metadata logs (or journals). In some cases, the source of data may be a volume, a structured or unstructured dataset, a bucket, a file within a file system, an entire file system directory, or a combination of sources — where the data sources are stored within a source data repository 602.

[000343] In some cases, there may be one or more replication data targets, where, for example, a source data repository 602 includes multiple pods 640, 704 and multiple, respective replication data targets, illustrated as target data repositories 750, 752. In this example, source pod 640 includes a volume 658, source pod 704 includes a volume 702, replica pod 706 includes replica volume 708, and replica pod 710 includes replica volume 712. Further, as illustrated in Figure 7, there may be one or more replica links 760, 762 that manage replication from the source data repository to one or more target data repositories 750, 752.

[000344] In some implementations, in an example where replication includes the use of snapshots of the source data, a replica link may specify a snapshot policy, which may specify conditions under which a snapshot may be taken. For example, if asynchronous replication, as described above with reference to Figure 6, becomes backed up — such as where the quantity of backed up data and/or metadata pending transfer would result in an RPO that is beyond a threshold RPO value — then a snapshot may be taken. In other examples, the snapshot policy may specify that snapshots are to be taken at a specified schedule, and may specify a length of time for keeping snapshots available.

[000345] Further, in some examples, instead of or in addition to generating snapshots for a source data repository to reduce a backlog of metadata and/or data transmissions to a target data repository, a source data repository may perform one or more transformations or optimizations on the data and/or metadata to be transmitted. For example, if a source data repository determines that data pending transfer is identical to data already transferred, then the source data repository may avoid sending the duplicate data that is pending transfer. As another example, checkpoints within a metadata log may be folded together, where if there is an overwrite between two checkpoints, then the source data repository may avoid sending data that has been overwritten, as reflected by the folded checkpoints.

[000346] Further, a replica link may also specify a replica policy, where the replica policy may include or be exclusively snapshots, specify continuous, but not synchronous replication, or specify synchronous replication. In all cases, a user may be provided with a single user interface, with a single workflow, for a replica link specification allowing for specification of one or more characteristics for data replication.

[000347] In some implementations, a replica link may also specify a compliance policy. For example, a compliance policy may specify that for a particular type of replication policy — for example, continuous, synchronous, asynchronous, snapshot — the replication should adhere to specified parameters. As one example, for a snapshot replication policy, the compliance policy may specify that if a frequency, or schedule, according to which snapshots are taken fails to meet a threshold level of compliance, then a system warning may be generated. Similarly, if data and/or metadata is not being transferred quickly enough to satisfy a specified RPO, or other performance metric, then a system warning or alert may also be generated. Alternately, updates on the source storage system can be slowed down in order to avoid exceeding the RPO.

[000348] However, in other cases, in response to failing to satisfy a threshold level of compliance, other corrective actions may be taken, for example, of a target data repository is a cause of a backup, or has had a drop in performance, or is nearing capacity, then a diagnostic may be initiated to identify correctable issues or an alternate target data repository may be identified for transferring the target replica data to the new target data repository. In some implementations, the replica link may also store attributes of the replication history, such as identifying a point at which a source data repository became frozen or unavailable.

[000349] Generally, a replica link may be used to specify a replication relationship, and depending on whether a pod is active or passive, determines a direction of the replication, where replication occurs in the direction of an active (or activated or promoted) pod to a passive (or deactivated or demoted) pod. In this example, a replication direction may also be changed if all pods connected to the replica link are in communication and reach consensus on a change in replication direction. In this way, a source pod may be protected by creating a replica link to another, deactivated, pod on another data repository, where hosts or host groups may be connected to the deactivated pod on the target data repository to read — nearly synchronous — data from the source pod.

[000350] For further explanation Figure 8 illustrates a configurable data replication system that provides continuous replication with minimal batching and an adjustable recovery point objective. In this example, a target data repository 624 generates replica data based on ordered metadata logs 854 received from a source data repository 602.

[000351] As described above, and in this example, continuous replication is pod-based and provides multiple recovery options. As illustrated in Figure 8, the example method includes: receiving (802), at a source data repository 602, one or more updates 852 to one or more datasets stored within the source data repository 602; generating (804), based on the one or

more updates 852 to the one or more datasets, a metadata 854 describing the one or more updates to the one or more datasets; and generating (806), based on the metadata 854 describing the one or more updates to the one or more datasets, a lightweight checkpoint 856 describing an ordered application of the one or more updates to the one or more datasets.

[000352] Receiving (802), at the source data repository 602, one or more updates 852 to one or more datasets stored within the source data repository 602 may be carried out as described above with regard to Figures 1A–3D, where data may be received at one or more network ports over one or more networks using one or more network communication protocols.

[000353] Generating (804), based on the one or more updates 852 to the one or more datasets, metadata 854 describing the one or more updates to the one or more datasets may be carried out as described above with reference to Figures 6 and 7.

[000354] Generating (806), based on the metadata 854 describing the one or more updates 852 to the one or more datasets, a lightweight checkpoint 856 describing an ordered application of the one or more updates 852 to the one or more datasets may be carried out as described above with reference to Figures 6 and 7, where operations within a given checkpoint within a metadata log for a given dataset, or volume, may be played to generate a replica one or more datasets.

[000355] For further explanation Figure 9 illustrates a continuous data replication system that provides continuous replication for local recovery of lost data. In this example, a source data repository 602 generates ordered metadata logs usable to recover a previous state of a dataset, where the metadata log is also a basis for ensuring that data corresponding to the metadata is not garbage collected.

[000356] As illustrated in Figure 9, the example method includes: sending (902), to a target data repository 624 from a source data repository 602, metadata 952 describing one or more updates to one or more datasets stored within the source data repository 602; generating (904), based on the metadata 952 describing the one or more updates to the one or more datasets, an ordered log of metadata 854 describing an ordered application of the one or more updates to the one or more datasets; and responsive to a recovery event, generate (906), on the source data repository 602 and based on the ordered log of metadata 854, at least a portion of the one or more datasets in accordance with the one or more updates corresponding to a specified point in time.

[000357] Sending (902), to a target data repository 624 from a source data repository 602, metadata 952 describing one or more updates to one or more datasets stored within the source

data repository 602 may be carried out as described above with reference to Figure 8 and sending (802) metadata 952 to a target data repository 624 from a source data repository 602.

[000358] Generating (904), based on the metadata 952 describing the one or more updates to the one or more datasets, an ordered log of metadata 854 describing an ordered application of the one or more updates to the one or more datasets may be carried out as described above with reference to Figure 8 and generating (804) an ordered log of metadata 854.

[000359] Generating (906), responsive to a recovery event 950 and on the source data repository 602 and based on the ordered log of metadata 854, at least a portion of the one or more datasets in accordance with the one or more updates corresponding to a specified point in time may be carried out as described above with reference to Figures 4A–7 describing how references to storage locations for data corresponding to one or more checkpoints within a metadata log may be used to access data at storage locations corresponding to the updates specified in the checkpoints. Further, recovery data for the specified point in time may be generated as described above with reference to Figures 4A–7, where updates may be replayed, in order, to generate recovery data.

[000360] As described above, data corresponding to the updates specified in the checkpoints of the metadata log is protected from garbage collection based on a garbage collection process using both storage system reference tables and also reference information within a metadata log.

[000361] Although embodiments are described above that include receiving, at a source data repository, one or more updates to one or more datasets stored within the source data repository, generating, based on the one or more updates, metadata describing the one or more updates, and generating, based on the metadata describing the one or more updates, a lightweight checkpoint of metadata describing an ordered application of the one or more updates, other embodiments are within the scope of the present disclosure. For example, some embodiments may include receiving one or more updates to one or more datasets stored within the source data repository, generating metadata describing the one or more updates, and generating, based on the metadata describing the one or more updates, a lightweight checkpoint of metadata describing an ordered application of the one or more updates. Other embodiments may include only receiving, at a source data repository, one or more updates to one or more datasets stored within the source data repository and generating, based on the one or more updates, a lightweight checkpoint of metadata describing an ordered application of the one or more updates. Alternatively, some embodiments may include only generating, based on one or more updates, metadata describing the one or more updates and generating,

based on the metadata describing the one or more updates, a lightweight checkpoint of metadata describing an ordered application of the one or more updates.

[000362] Example embodiments are described largely in the context of a fully functional computer system. Readers of skill in the art will recognize, however, that the present disclosure also may be embodied in a computer program product disposed upon computer readable storage media for use with any suitable data processing system. Such computer readable storage media may be any storage medium for machine-readable information, including magnetic media, optical media, or other suitable media. Examples of such media include magnetic disks in hard drives or diskettes, compact disks for optical drives, magnetic tape, and others as will occur to those of skill in the art. Persons skilled in the art will immediately recognize that any computer system having suitable programming means will be capable of executing the steps of the method as embodied in a computer program product. Persons skilled in the art will recognize also that, although some of the example embodiments described in this specification are oriented to software installed and executing on computer hardware, nevertheless, alternative embodiments implemented as firmware or as hardware are well within the scope of the present disclosure.

[000363] Embodiments can include be a system, a method, and/or a computer program product. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present disclosure.

[000364] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals *per se*, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating

through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[000365] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[000366] Computer readable program instructions for carrying out operations of the present disclosure may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Smalltalk, C++ or the like, and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present disclosure.

[000367] Aspects of the present disclosure are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to some embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the

flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[000368] These computer readable program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[000369] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[000370] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

CLAIMS

What is claimed is:

1. A method comprising:

receiving, at a source data repository, one or more updates to one or more datasets stored within the source data repository;

generating, based on the one or more updates, metadata describing the one or more updates; and

generating, based on the metadata describing the one or more updates, a lightweight checkpoint of metadata describing an ordered application of the one or more updates.

2. The method of claim 1, further comprising:

sending, from the source data repository to a target data repository and independent from sending the lightweight checkpoint to the target data repository, data corresponding to the lightweight checkpoint.

3. The method of claim 1 or 2, wherein an ordered log of metadata includes one or more lightweight checkpoints.

4. The method of claims 1, 2, or 3, wherein a quantity of updates described by a given lightweight checkpoint is specified by a configurable setting.

5. The method of claims 4, wherein the configurable setting includes a target recovery point objective.

6. The method of claim 1, 2, 3, 4, or 5, further comprising generating, responsive to a recovery event at least a portion of the one or more datasets in accordance with the one or more updates corresponding to a specified point in time.

7. The method of claim 6 wherein the recovery event is responsive to a detected degradation of responsiveness beyond a compliance policy threshold of the source data repository.

8. A storage system that includes a computer memory and a computer processor, the computer memory including program instructions that, when executed by the computer processor, cause the storage system to carry out the steps of:

receiving, at a target data repository from a source data repository, metadata describing one or more updates to one or more datasets stored within the source data repository;

generating, based on the metadata describing the one or more updates, an ordered log of metadata describing an ordered application of the one or more updates; and

generating, based on the ordered log of metadata, the one or more datasets in accordance with the one or more updates.

9. The storage system of claim 8, wherein the program instructions, when executed by the computer processor, further cause the storage system to carry out the step of:
receiving, at the target data repository from the source data repository and independent from receiving the metadata describing the one or more updates, data corresponding to the one or more updates.

10. The storage system of claim 8 or 9, wherein the ordered log of metadata includes one or more lightweight checkpoints.

11. The storage system of claims 8, 9, or 10, wherein a quantity of updates described by a given lightweight checkpoint is specified by a configurable setting.

12. The storage system of claims 11, wherein the configurable setting includes a target recovery point objective.

13. The storage system of claim 8, 9, 10, 11, or 12, wherein the program instructions, when executed by the computer processor, further cause the storage system to carry out the step of generating, responsive to a recovery event at least a portion of the one or more datasets in accordance with the one or more updates corresponding to a specified point in time.

14. The storage system of claim 13, wherein the recovery event is responsive to a detected failure.

15. The storage system of claim 13, wherein the recovery event is responsive to a detected impending failure.

16. The storage system of claim 13, wherein the recovery event is based on a promotion event to prepare, for testing on the target data repository, a replica of the stored data at the source data repository.

17. A method comprising:
generating, based on the metadata describing the one or more updates to the one or more datasets, an ordered log of metadata describing an ordered application of the one or more updates; and
generating, on the source data repository and based on the ordered log of metadata, at least a portion of the one or more datasets in accordance with the one or more updates corresponding to a specified point in time.

18. The method of claim 17, wherein stored data to generate the one or more datasets is available based on garbage collection on the source data repository using both a

reference table maintained by a storage controller and also a list of references within the ordered log of metadata.

19. The method of claim 17 or 18, wherein a quantity of updates described by a given lightweight checkpoint is specified by a configurable setting.

20. The method of claims 17, 18, or 19 wherein the configurable setting specifies a recovery window of time.

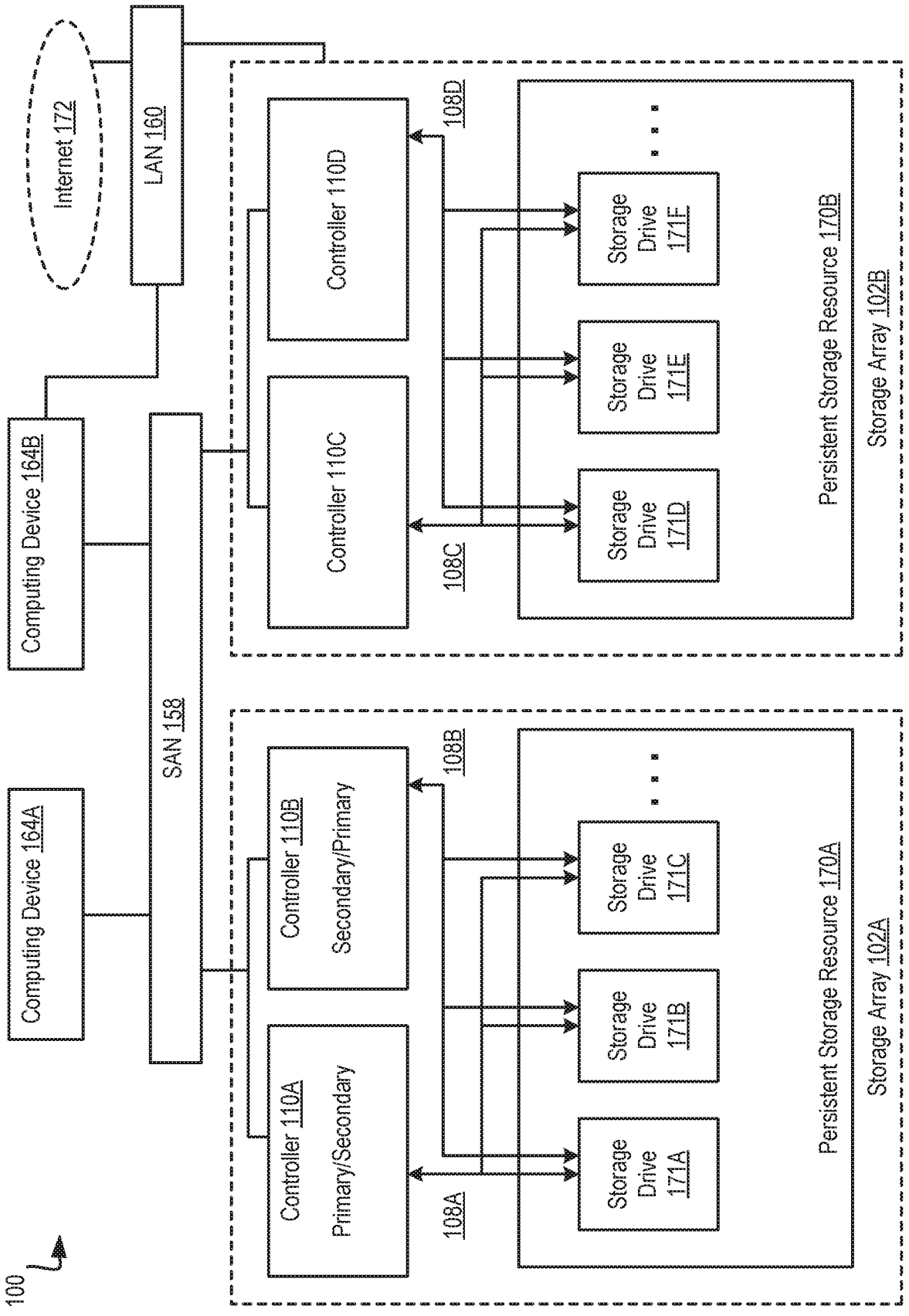

100
Computing Device 164A
Computing Device 164B
SAN 158
LAN 160
Internet 172
Storage Array 102A
Persistent Storage Resource 170A
Controller 110A
Primary/Secondary
Controller 110B
Secondary/Primary
108A
108B
Storage Drive 171A
Storage Drive 171B
Storage Drive 171C
Storage Array 102B
Persistent Storage Resource 170B
Controller 110C
Controller 110D
108C
108D
Storage Drive 171D
Storage Drive 171E
Storage Drive 171F

FIG. 1A

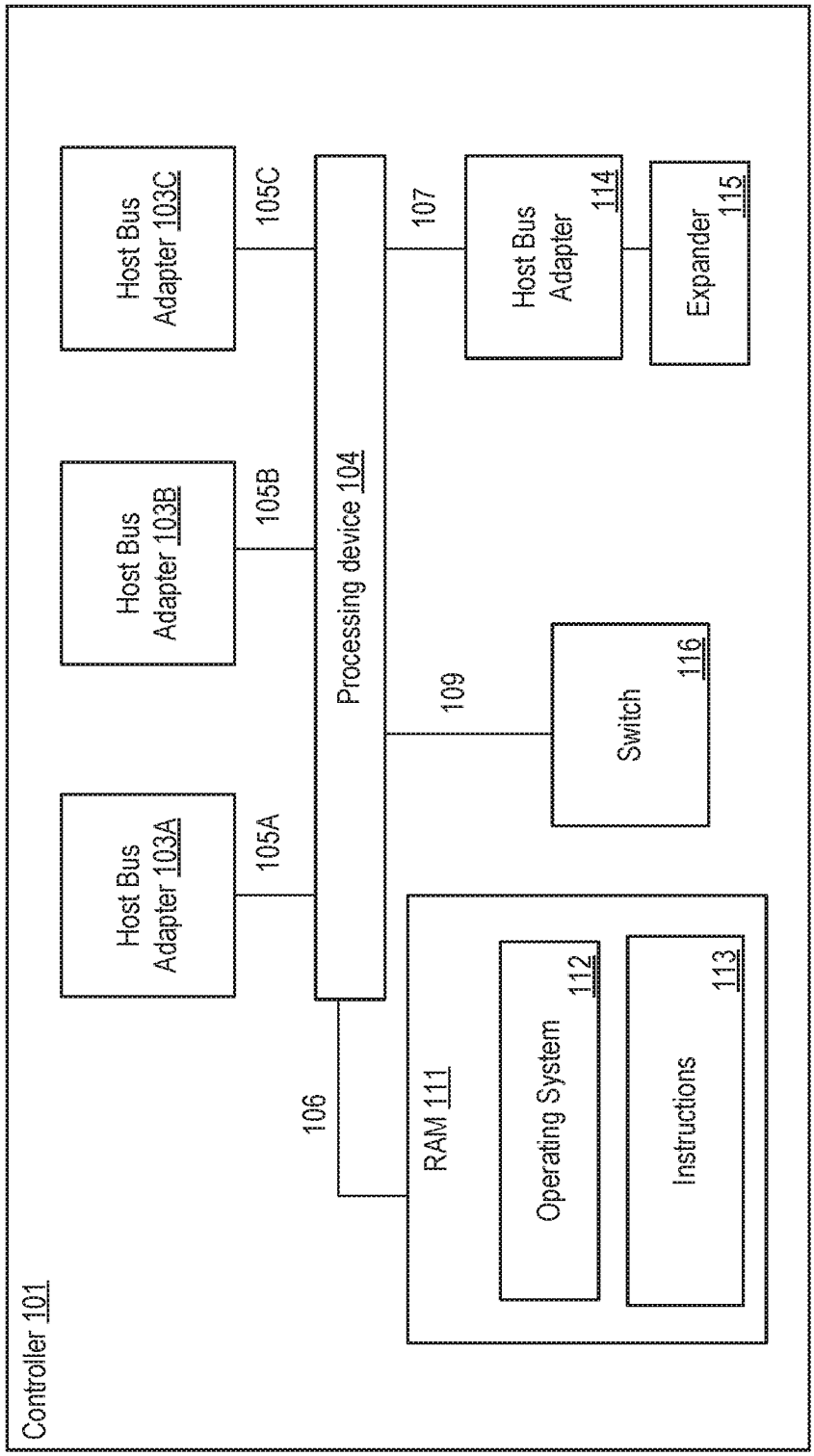

Controller 101
Host Bus Adapter 103A
Host Bus Adapter 103B
Host Bus Adapter 103C
105A
105B
105C
Processing device 104
106
107
109
RAM 111
Operating System 112
Instructions 113
Switch 116
Host Bus Adapter 114
Expander 115

FIG. 1B

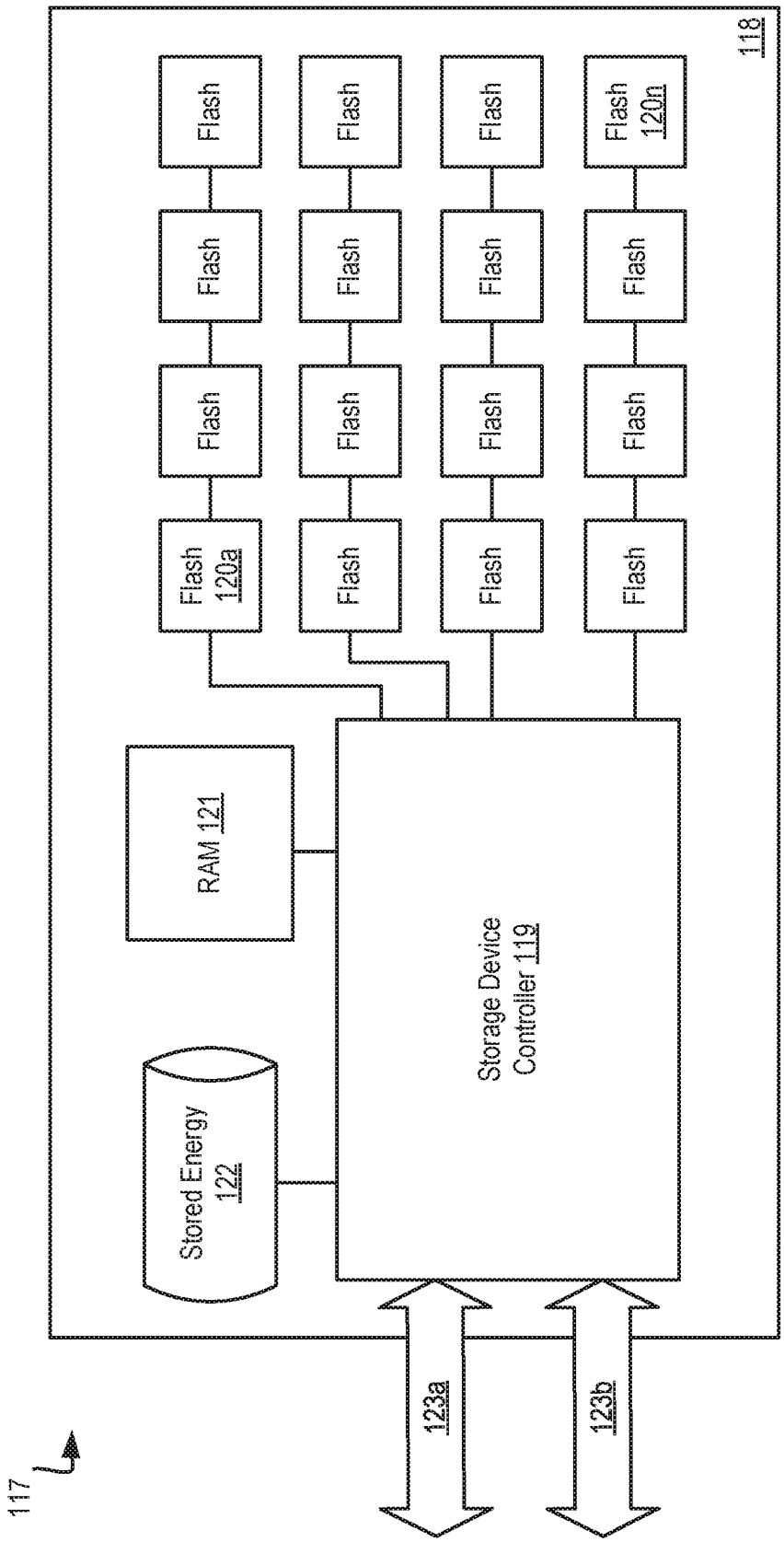

118
Flash
Flash
Flash
Flash 120a
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash 120n
RAM 121
Storage Device Controller 119
Stored Energy 122
123a
123b
117

FIG. 1C

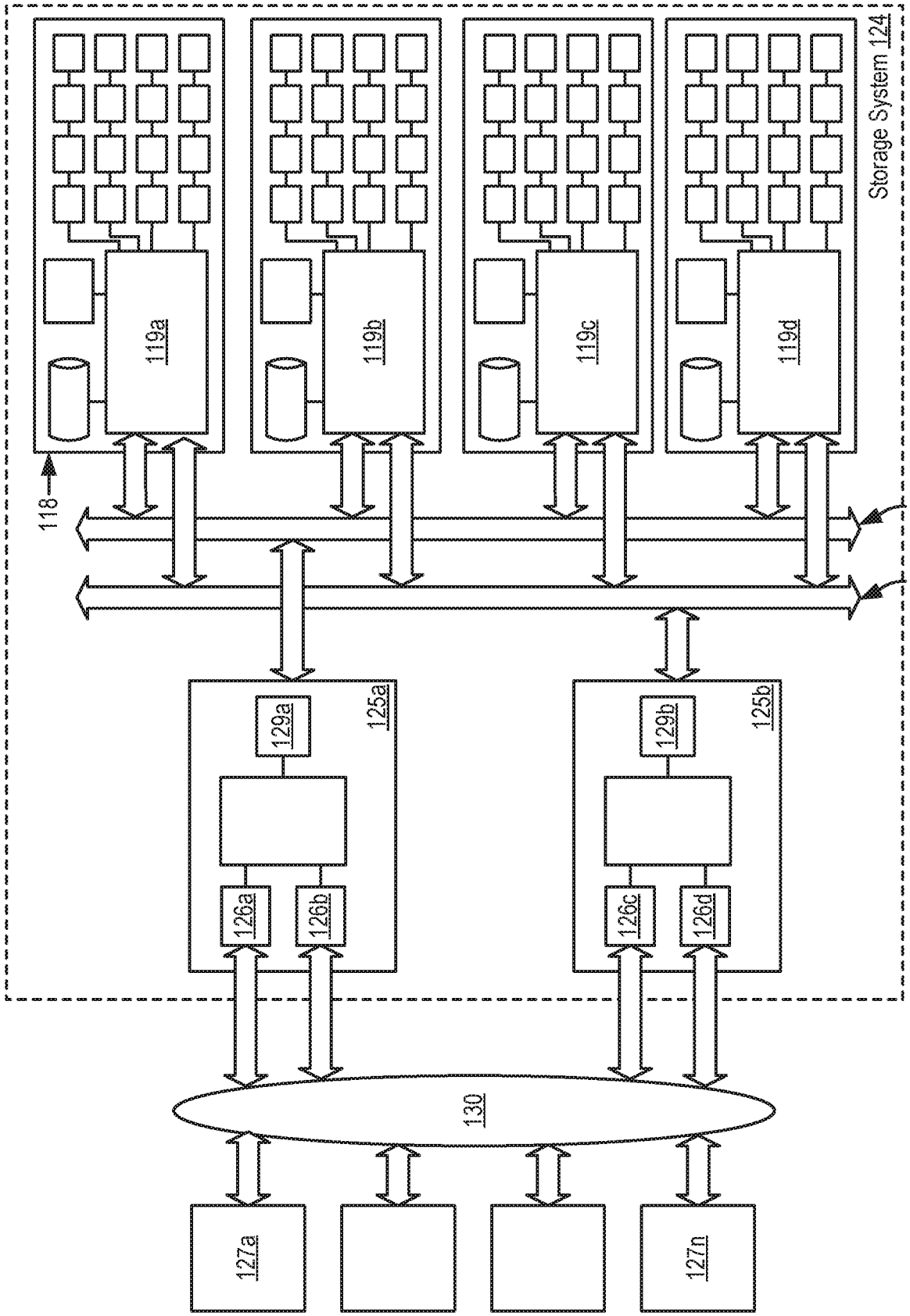

Storage System 124
119a
119b
119c
119d
118
128b
128a
125a
125b
129a
129b
126a
126b
126c
126d
130
127a
127n

FIG. 1D

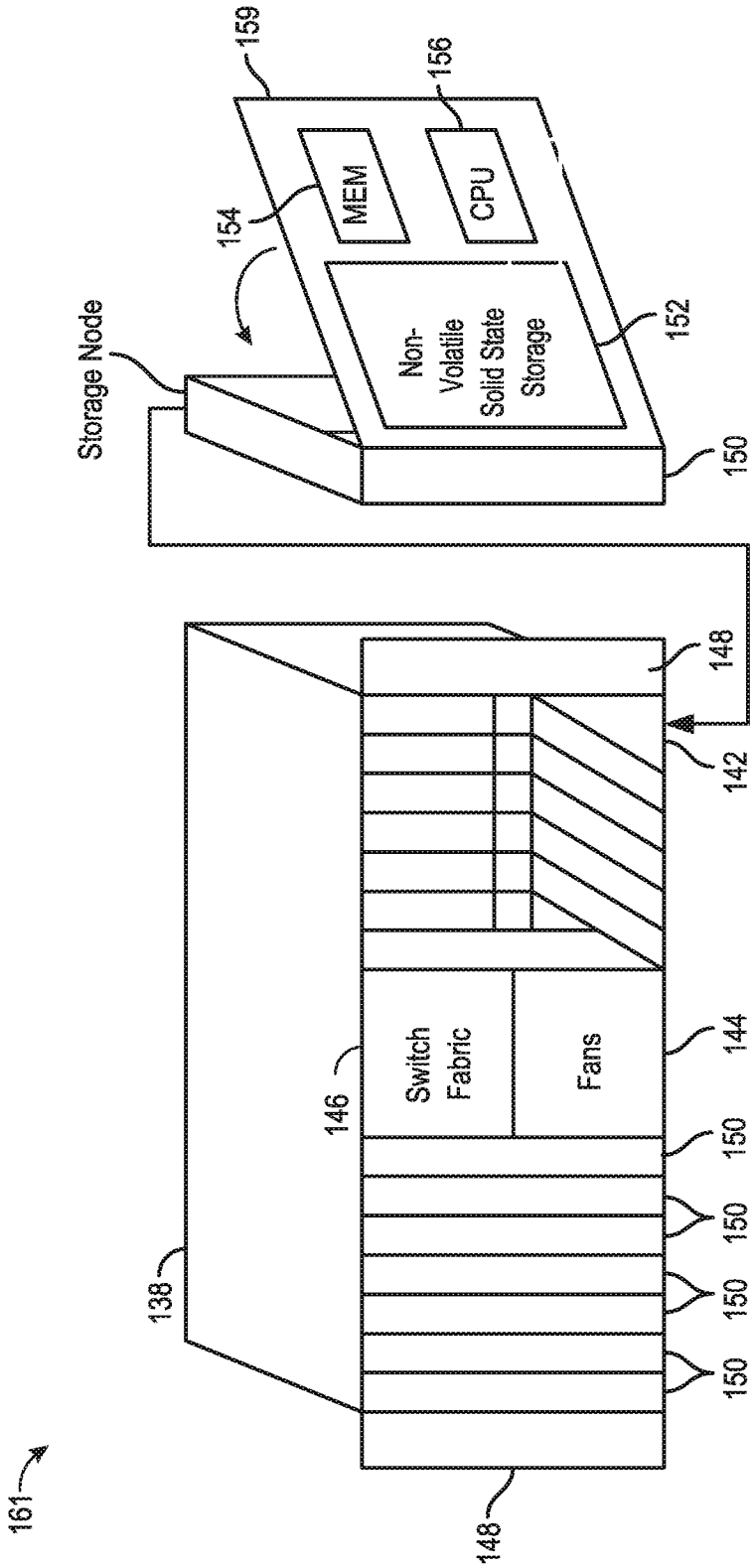

161
138
148
146
Switch Fabric
Fans
144
150
150
150
150
148
142
Storage Node
154
MEM
CPU
156
159
Non-Volatile Solid State Storage
152
150

FIG. 2A

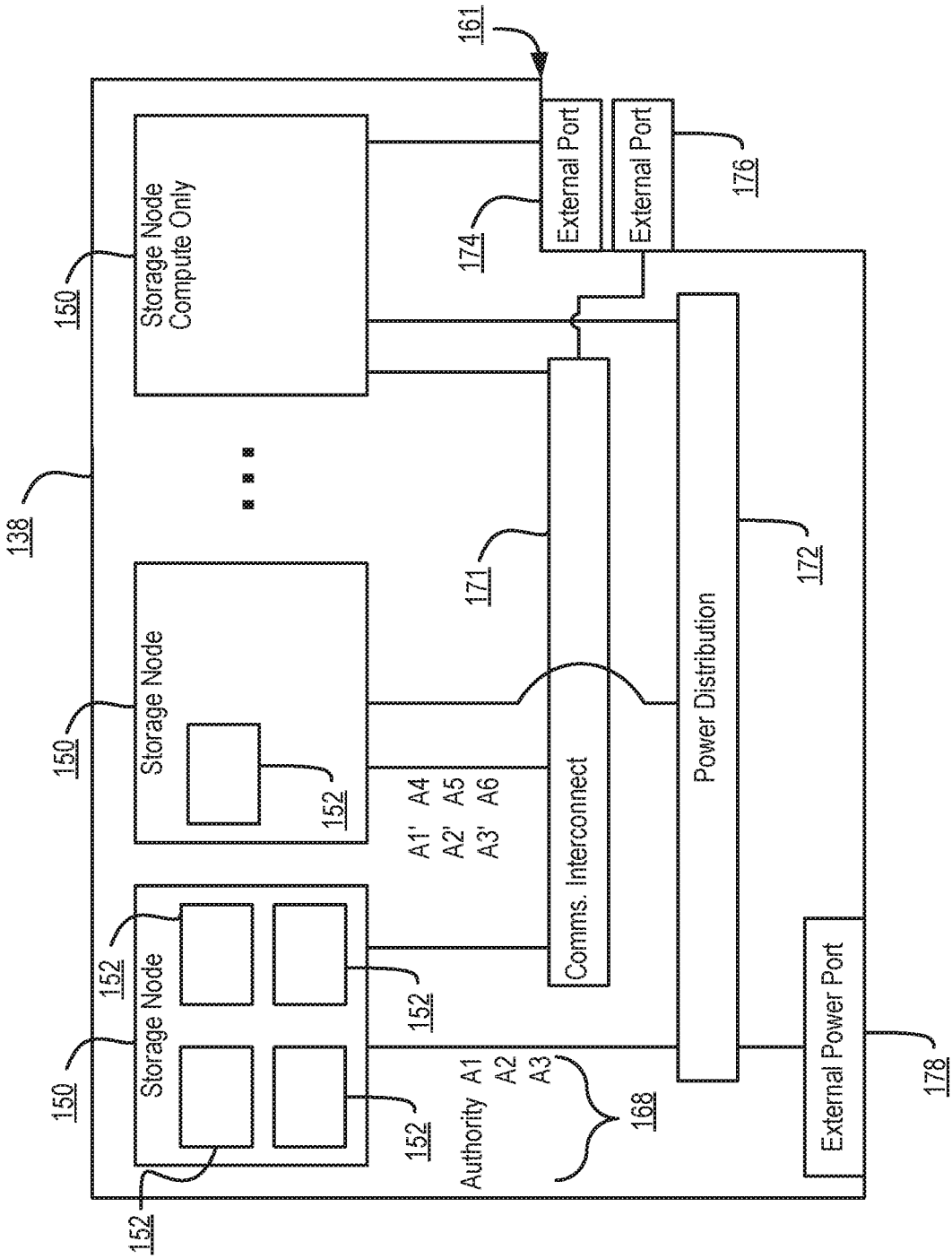

161
External Port
External Port
176
174
150
Storage Node
Compute Only
138
150
Storage Node
152
171
172
A1' A4
A2' A5
A3' A6
Comms. Interconnect
Power Distribution
150
Storage Node
152
152
152
152
Authority A1
A2
A3
168
External Power Port
178

FIG. 2B

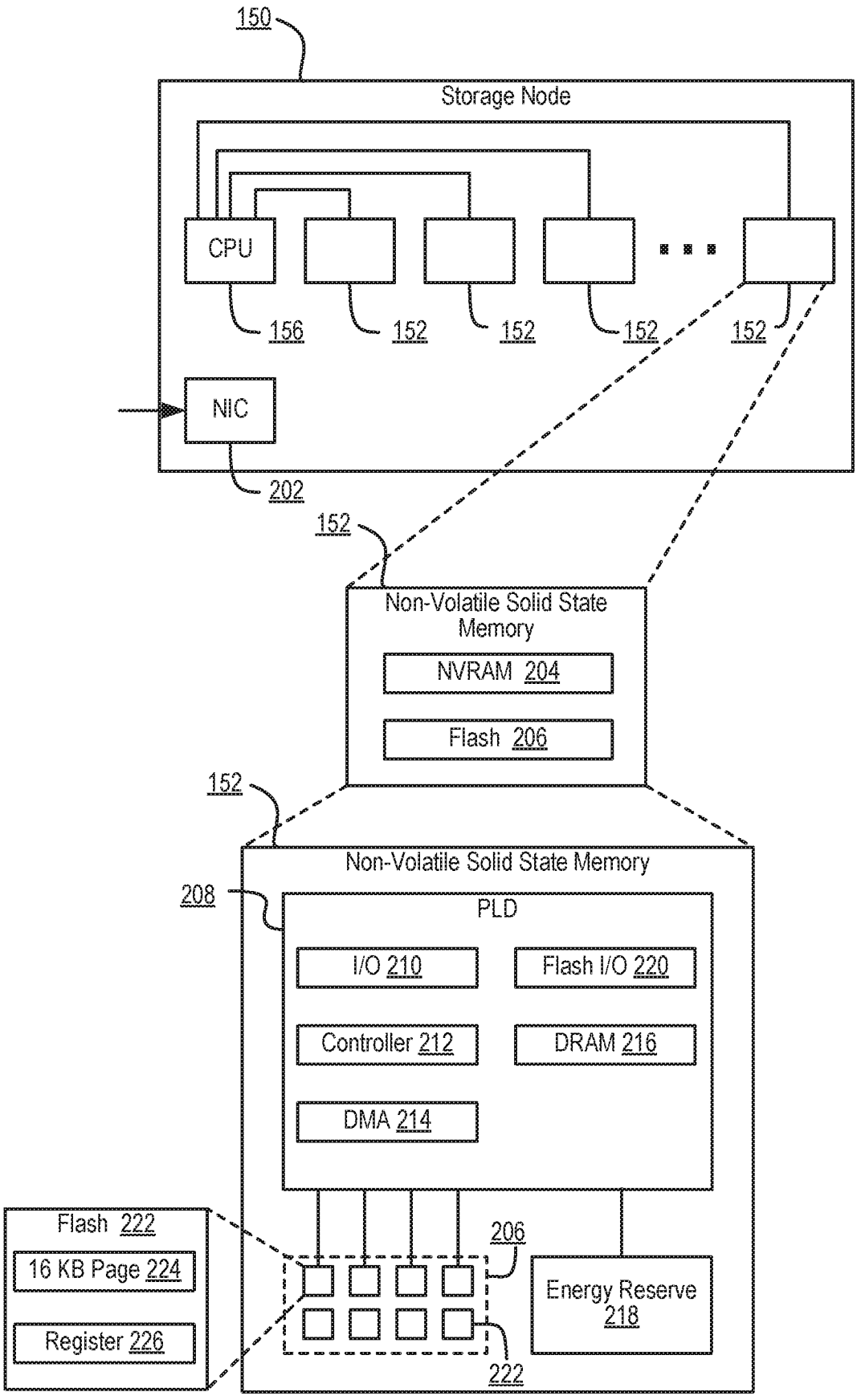

150
Storage Node
CPU
156
152
152
152
152
NIC
202
152
Non-Volatile Solid State Memory
NVRAM 204
Flash 206
152
Non-Volatile Solid State Memory
208
PLD
I/O 210
Flash I/O 220
Controller 212
DRAM 216
DMA 214
Flash 222
16 KB Page 224
Register 226
206
Energy Reserve 218
222

FIG. 2C

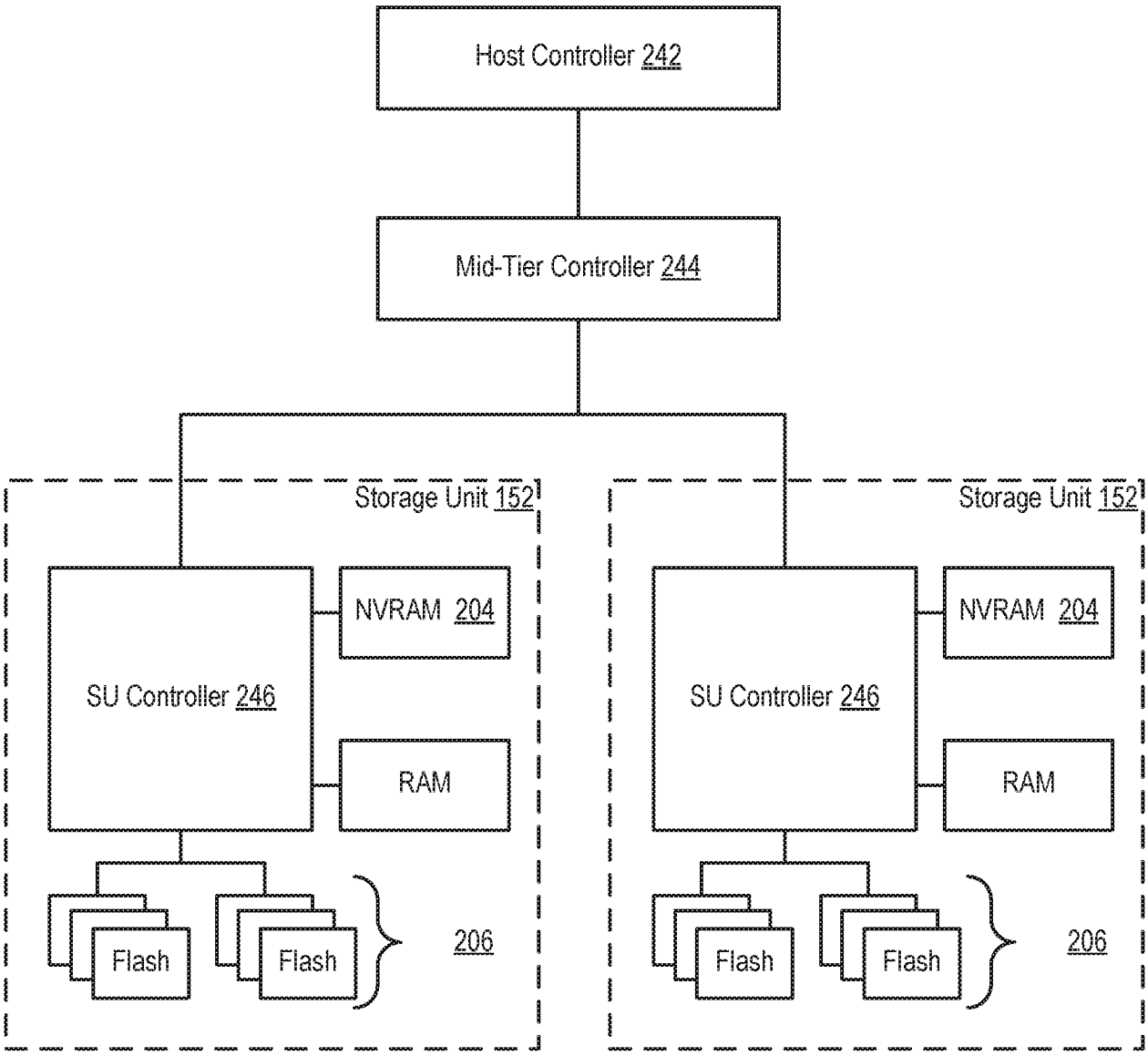

Host Controller 242
Mid-Tier Controller 244
Storage Unit 152
SU Controller 246
NVRAM 204
RAM
Flash
Flash
206
Storage Unit 152
SU Controller 246
NVRAM 204
RAM
Flash
Flash
206

FIG. 2D

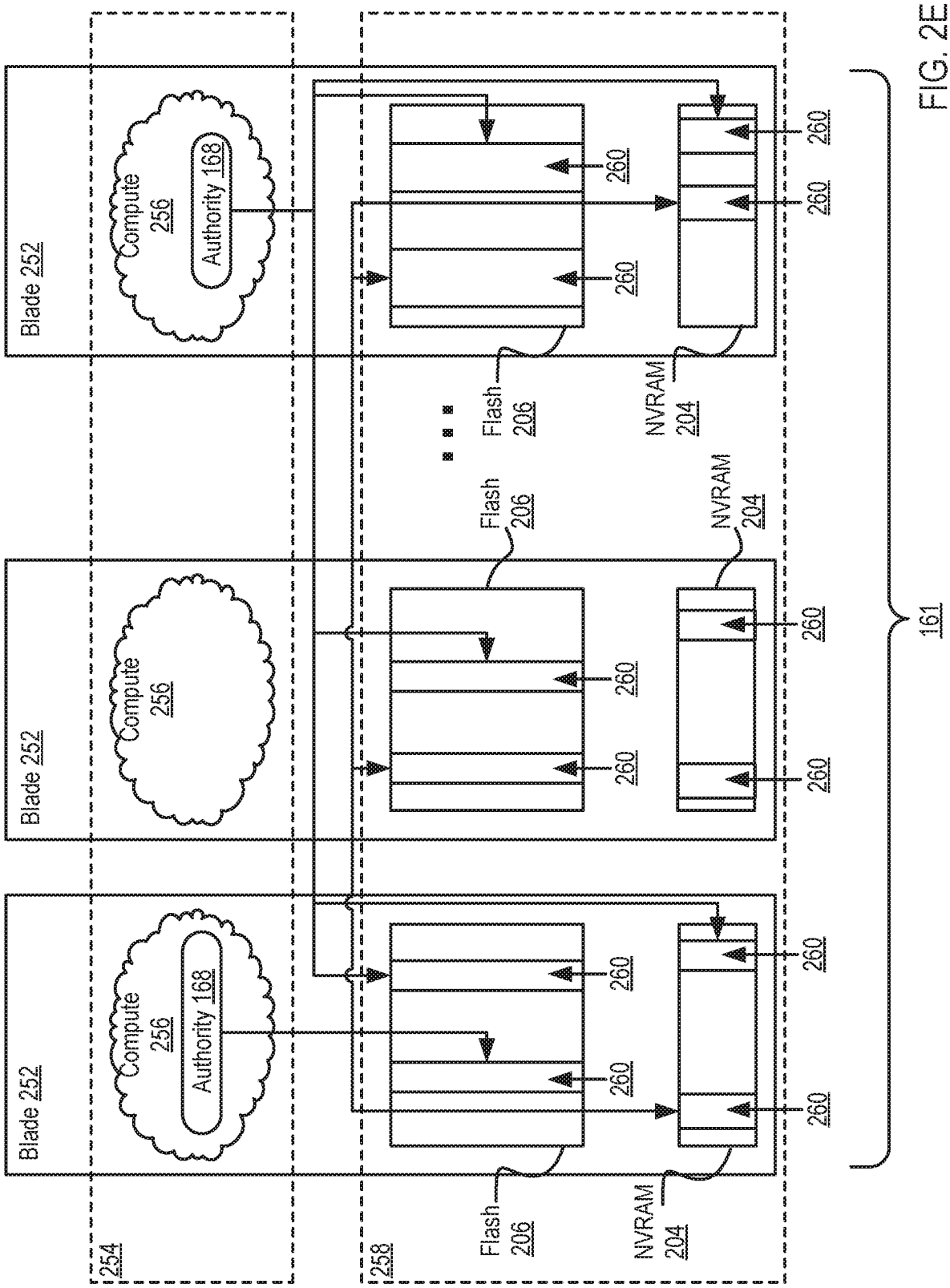

Blade 252
Compute 256
Authority 168
Blade 252
Compute 256
Blade 252
Compute 256
Authority 168
254
258
Flash 206
NVRAM 204
260
161

FIG. 2E

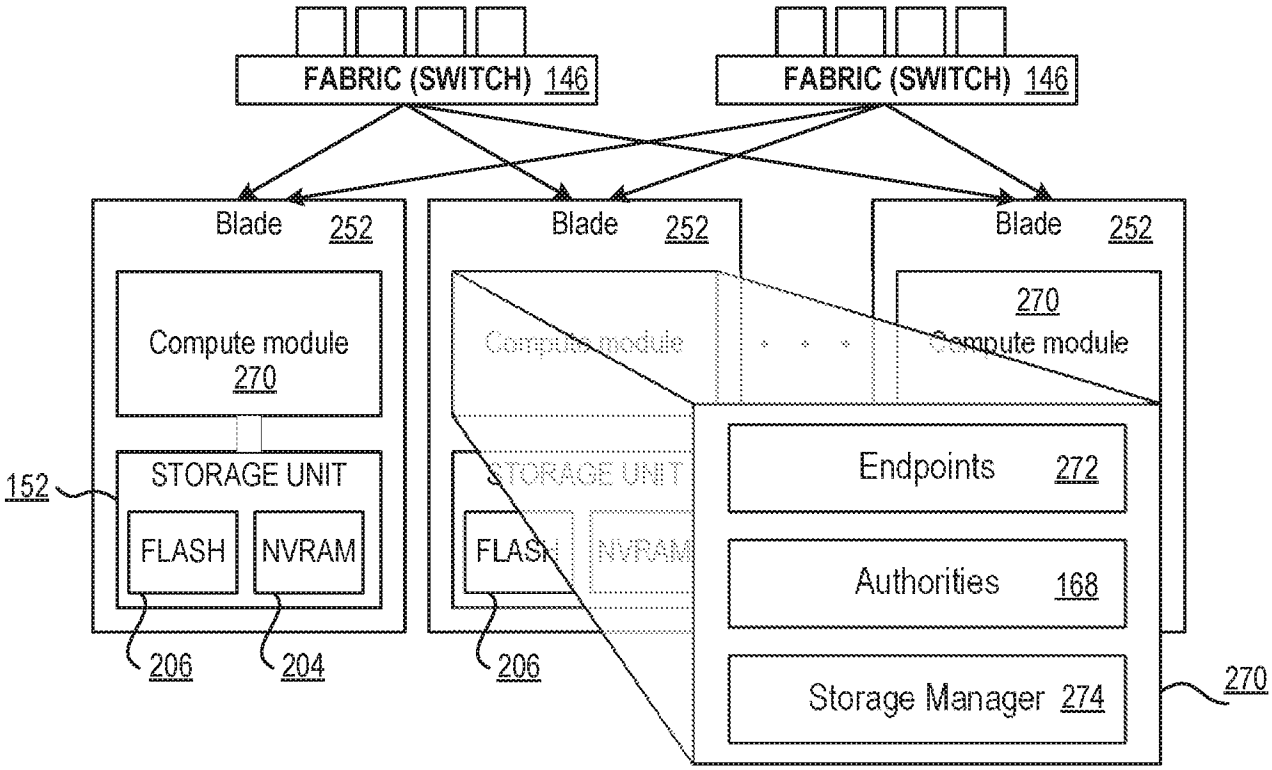

FABRIC (SWITCH) 146
FABRIC (SWITCH) 146
Blade 252
Blade 252
Blade 252
Compute module 270
Compute module
270
Compute module
STORAGE UNIT
FLASH
NVRAM
152
206
204
206
Endpoints 272
Authorities 168
Storage Manager 274
270

FIG. 2F

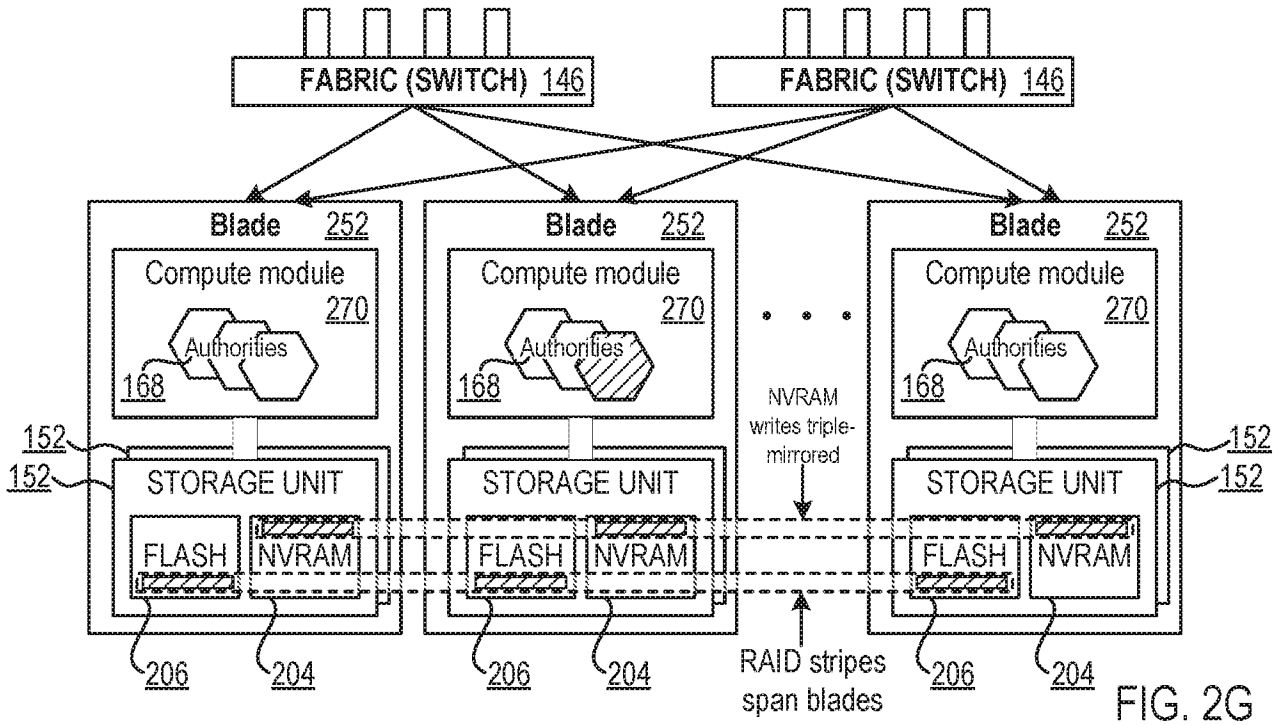

FABRIC (SWITCH) 146
FABRIC (SWITCH) 146
Blade 252
Blade 252
Blade 252
Compute module 270
Authorities 168
152
STORAGE UNIT
FLASH
NVRAM
206
204
NVRAM writes triple-mirrored
RAID stripes span blades

FIG. 2G

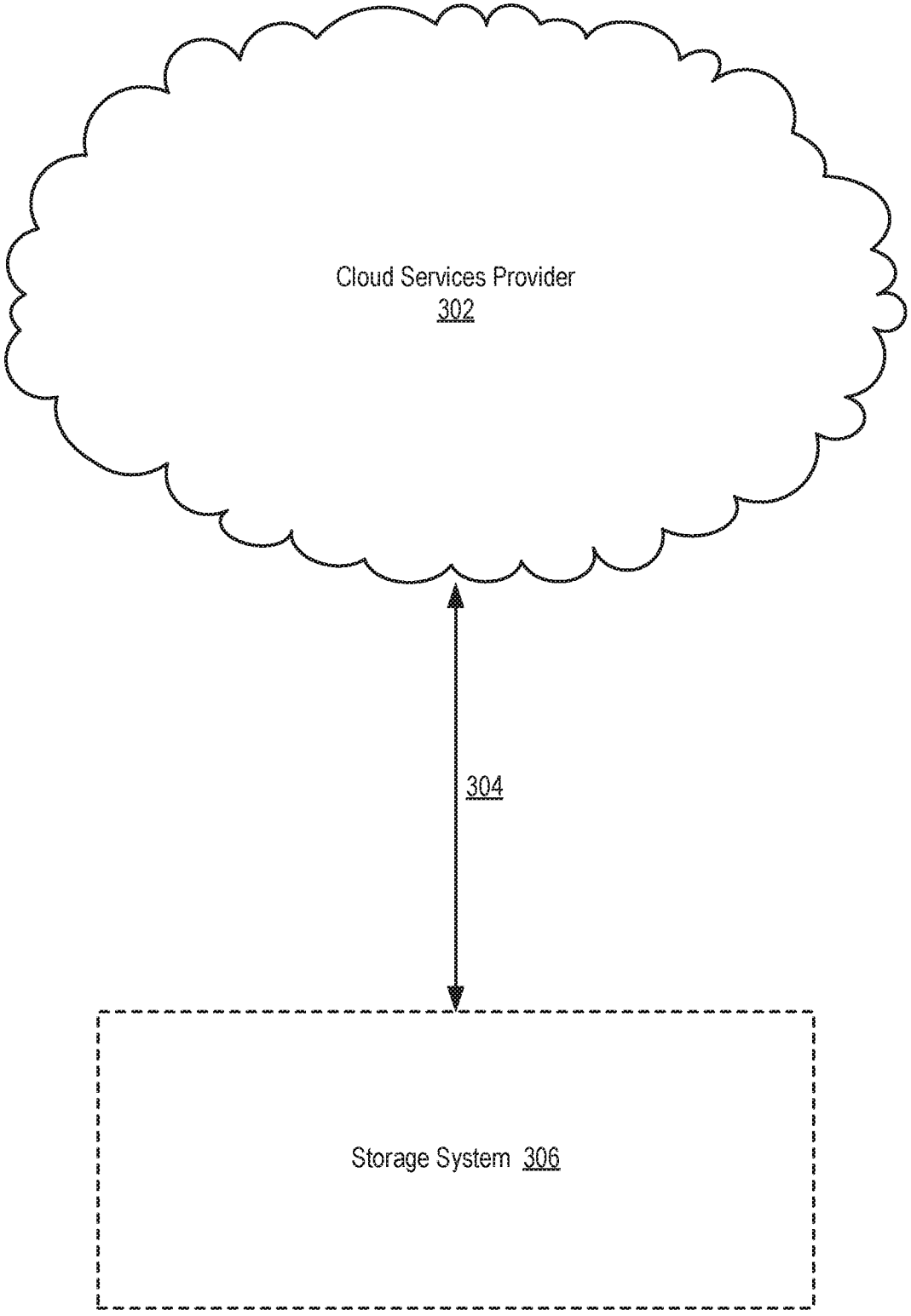


FIG. 3A

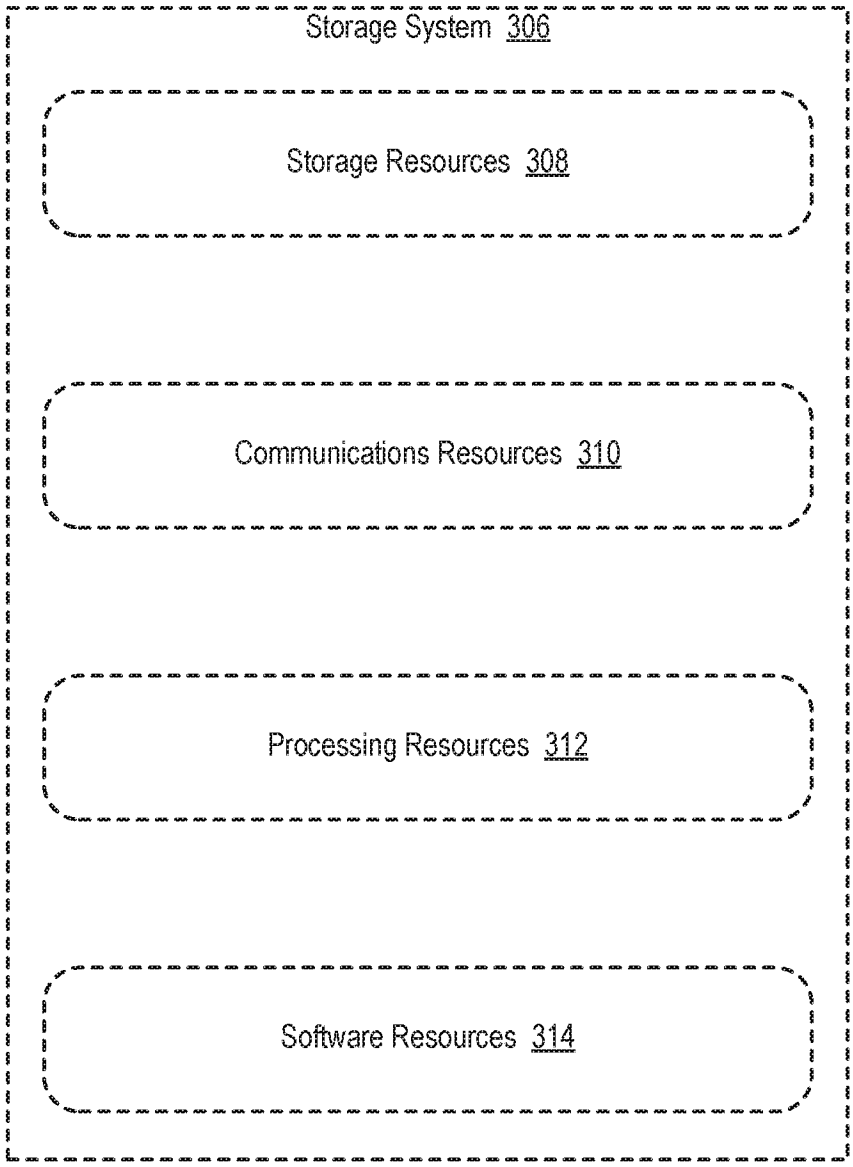

Storage System 306
Storage Resources 308
Communications Resources 310
Processing Resources 312
Software Resources 314

FIG. 3B

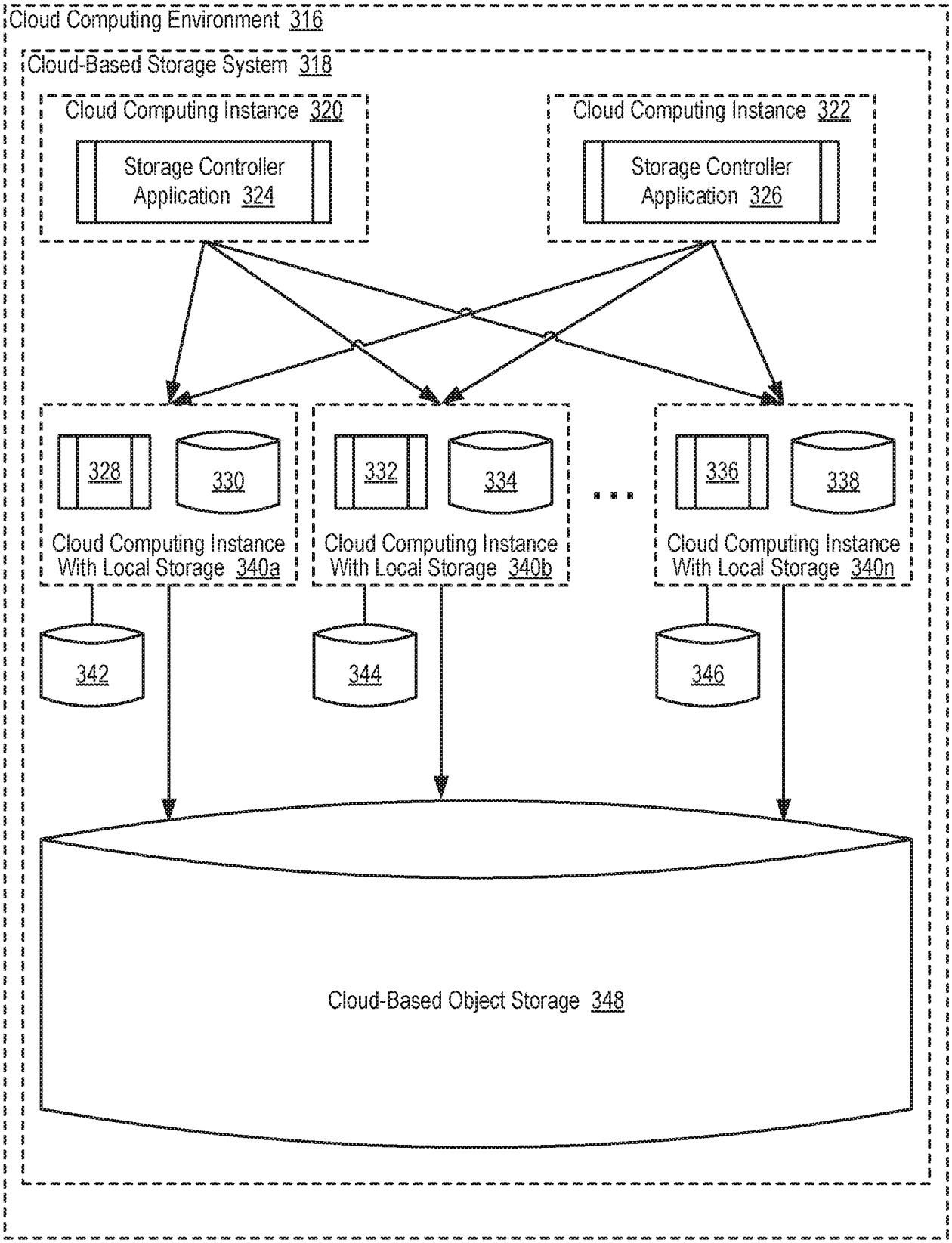

Cloud Computing Environment 316
Cloud-Based Storage System 318
Cloud Computing Instance 320
Storage Controller Application 324
Cloud Computing Instance 322
Storage Controller Application 326
328
330
332
334
336
338
Cloud Computing Instance With Local Storage 340a
Cloud Computing Instance With Local Storage 340b
Cloud Computing Instance With Local Storage 340n
342
344
346
Cloud-Based Object Storage 348

FIG. 3C

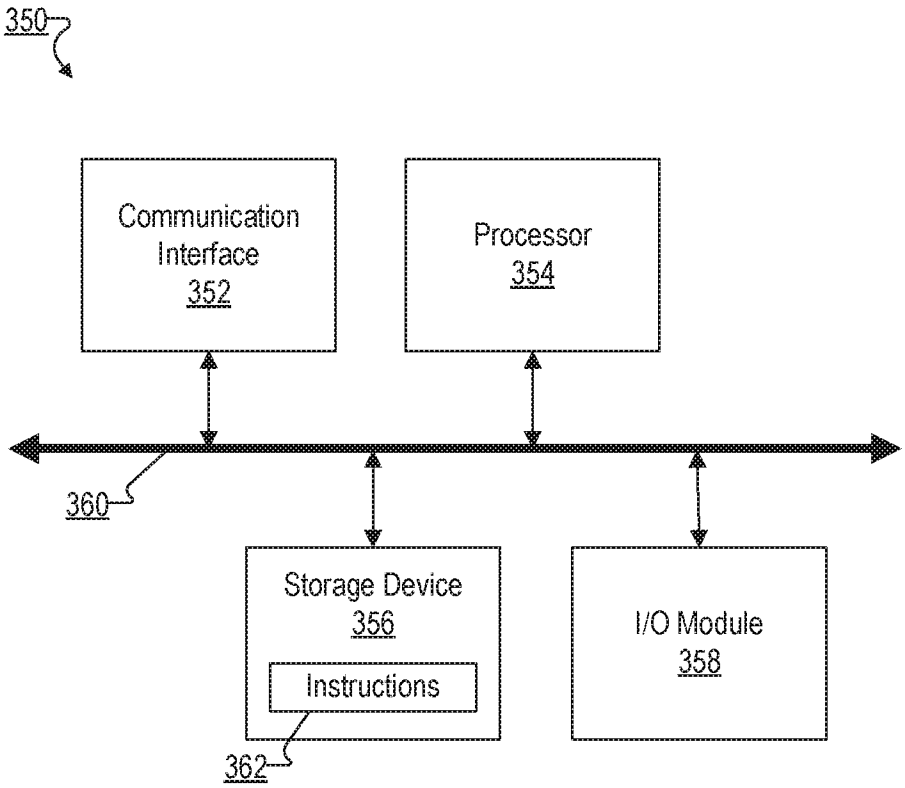

350
Communication Interface 352
Processor 354
360
Storage Device 356
Instructions
362
I/O Module 358

FIG. 3D

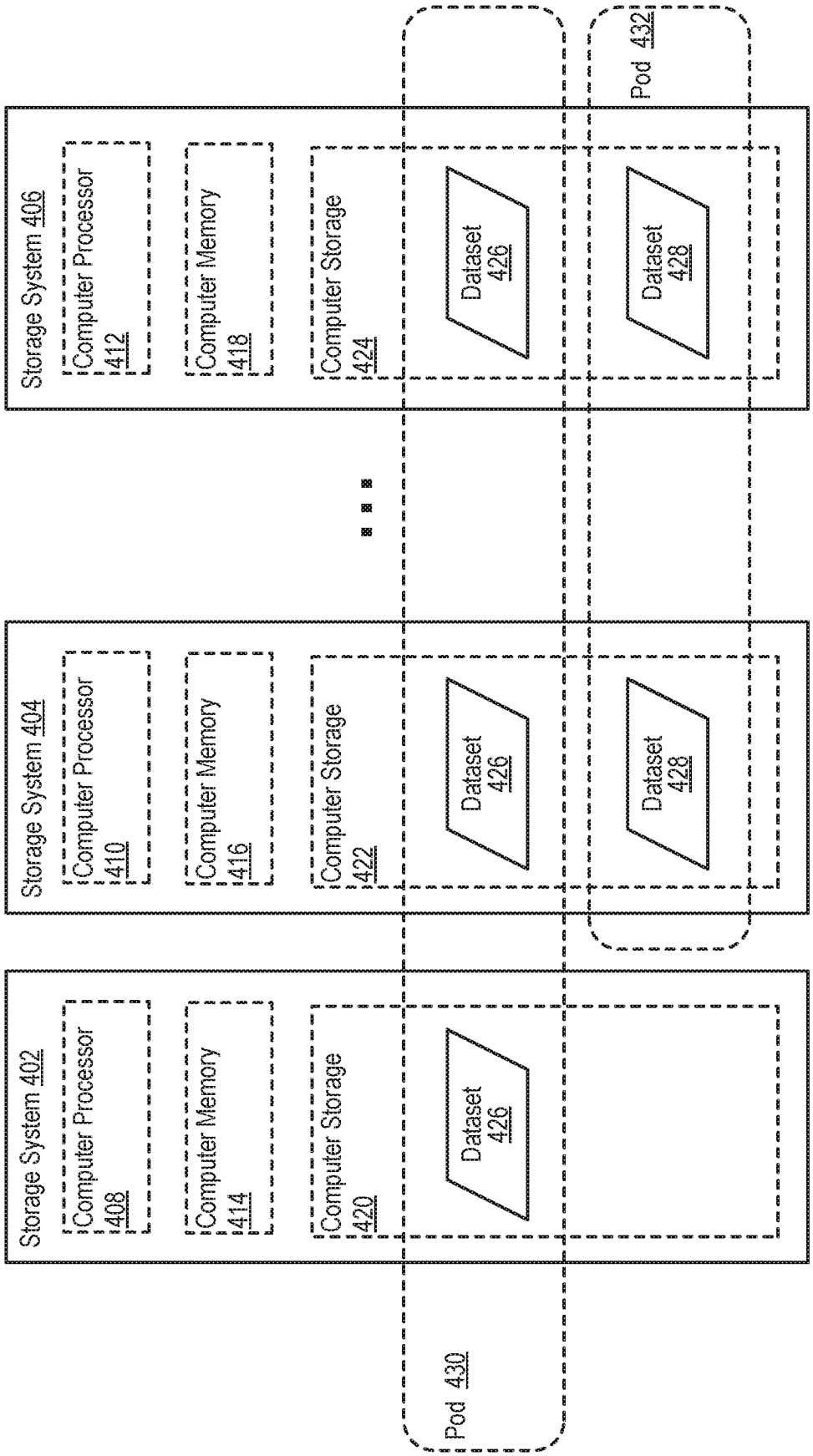

Storage System 402
Computer Processor 408
Computer Memory 414
Computer Storage 420
Dataset 426
Storage System 404
Computer Processor 410
Computer Memory 416
Computer Storage 422
Dataset 426
Storage System 406
Computer Processor 412
Computer Memory 418
Computer Storage 424
Dataset 426
Dataset 428
Dataset 428
Pod 430
Pod 432

FIG. 4A

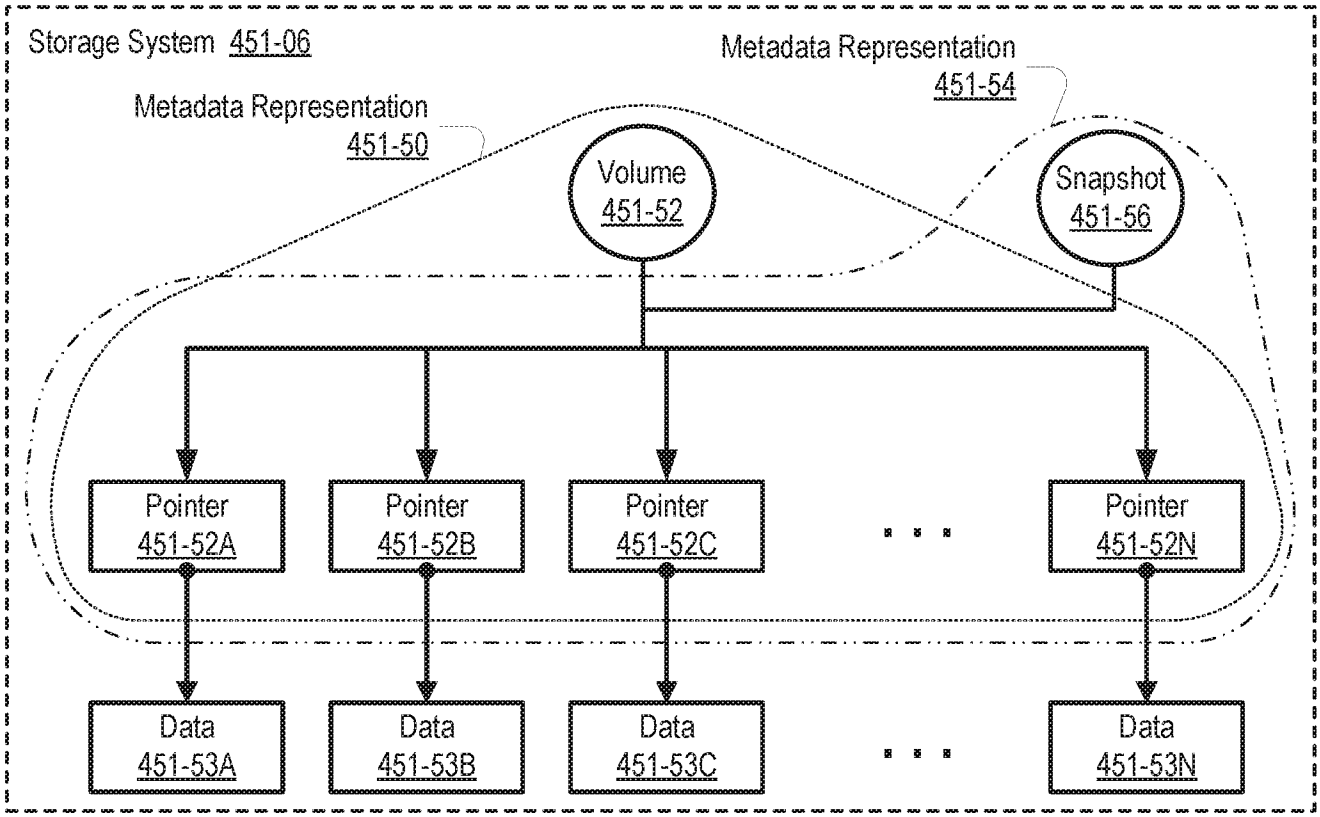
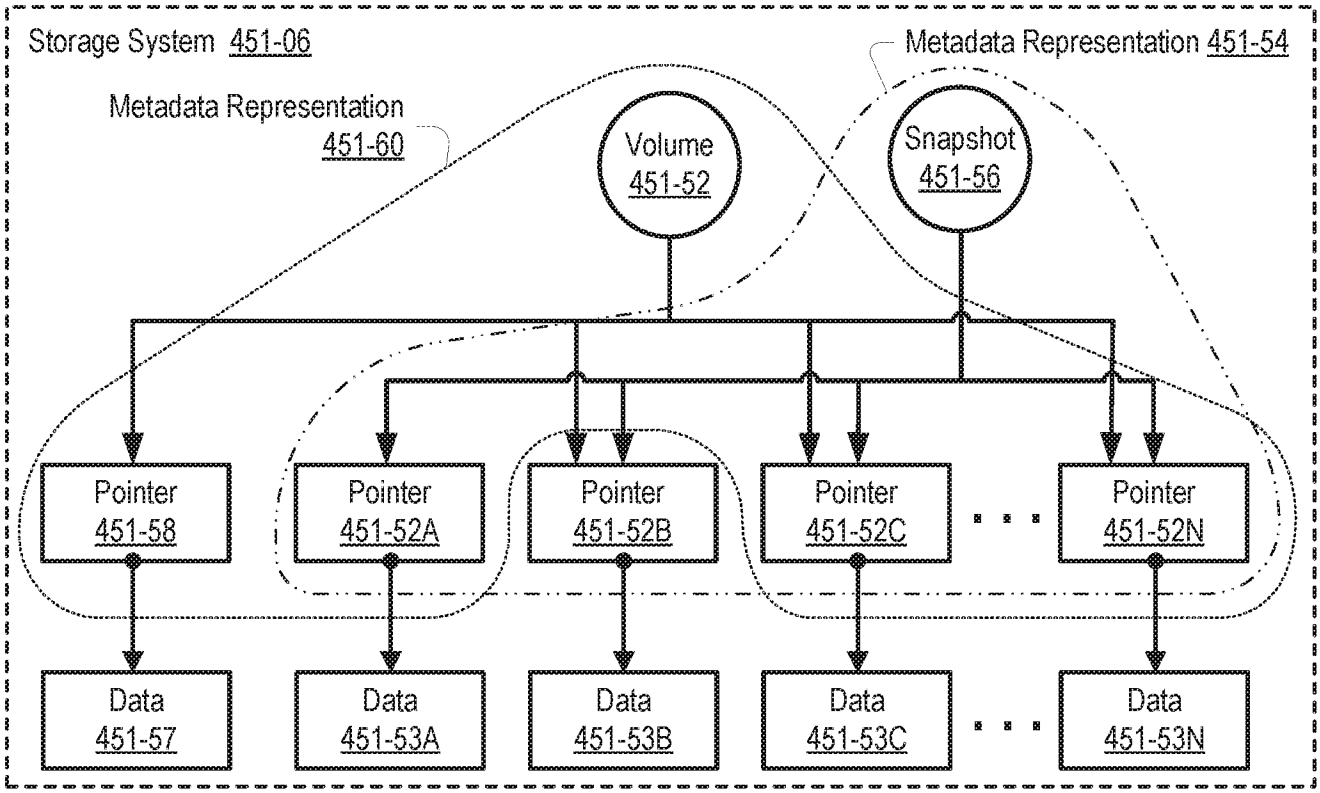


FIG. 4B

Write To Data 451-53B Of Volume 451-52



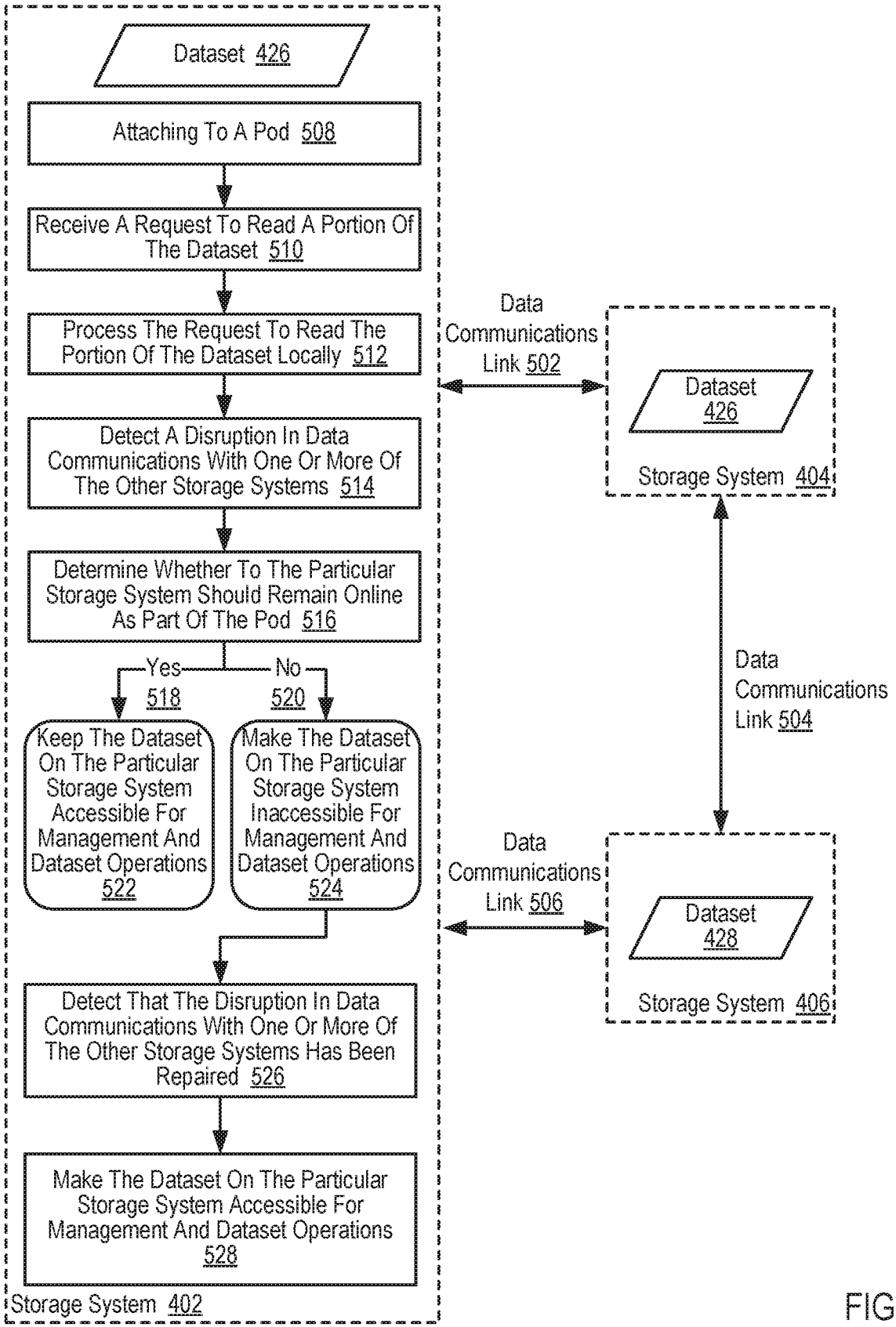

Dataset 426
Attaching To A Pod 508
Receive A Request To Read A Portion Of The Dataset 510
Process The Request To Read The Portion Of The Dataset Locally 512
Detect A Disruption In Data Communications With One Or More Of The Other Storage Systems 514
Determine Whether To The Particular Storage System Should Remain Online As Part Of The Pod 516
Yes 518
No 520
Keep The Dataset On The Particular Storage System Accessible For Management And Dataset Operations 522
Make The Dataset On The Particular Storage System Inaccessible For Management And Dataset Operations 524
Detect That The Disruption In Data Communications With One Or More Of The Other Storage Systems Has Been Repaired 526
Make The Dataset On The Particular Storage System Accessible For Management And Dataset Operations 528
Storage System 402
Data Communications Link 502
Dataset 426
Storage System 404
Data Communications Link 504
Data Communications Link 506
Dataset 428
Storage System 406

FIG. 5

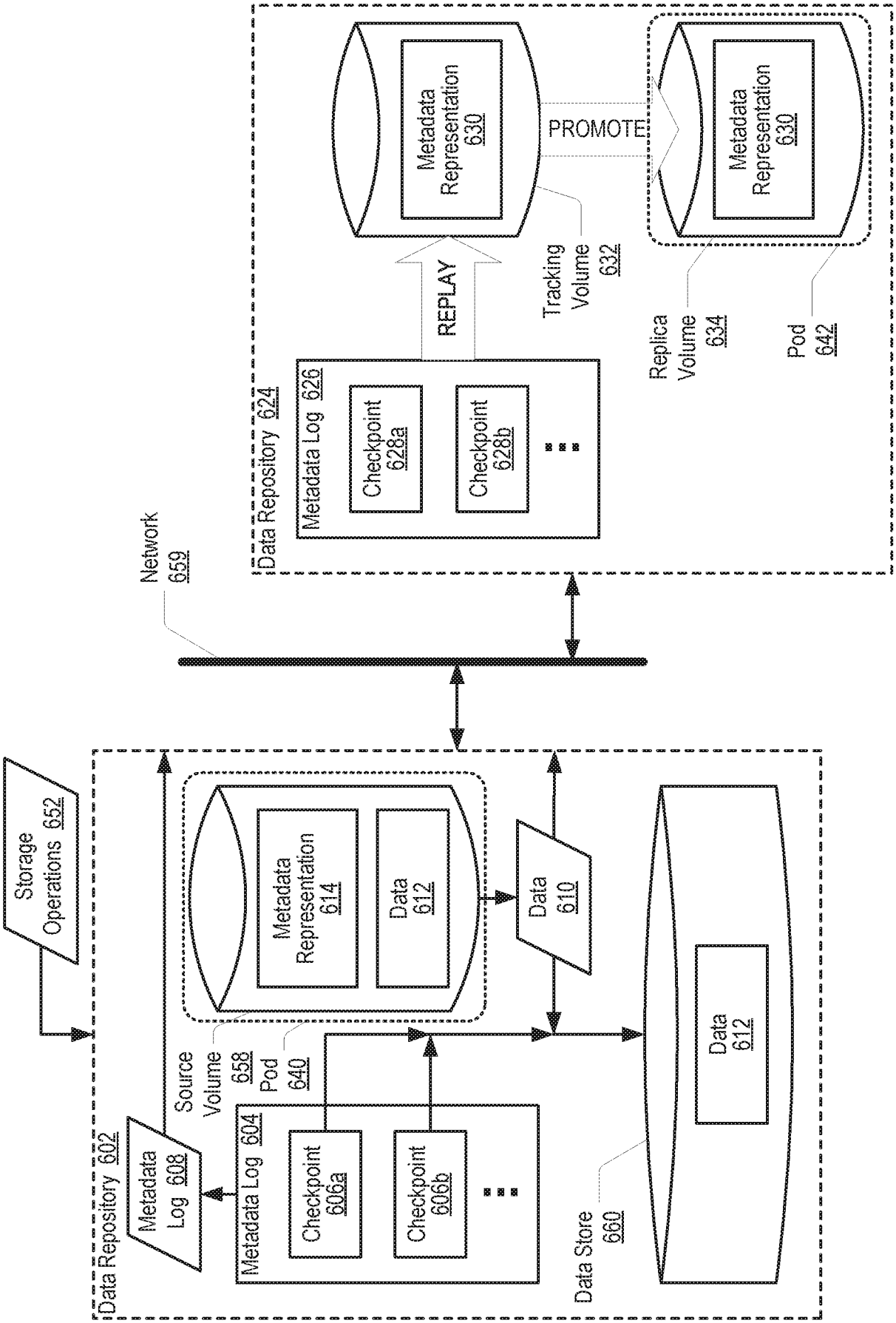

Storage Operations 652
Data Repository 602
Metadata Log 608
Metadata Log 604
Checkpoint 606a
Checkpoint 606b
...
Source Volume 658
Pod 640
Metadata Representation 614
Data 612
Data 610
Data Store 660
Data 612
Network 659
Data Repository 624
Metadata Log 626
Checkpoint 628a
Checkpoint 628b
...
REPLAY
Tracking Volume 632
Metadata Representation 630
PROMOTE
Metadata Representation 630
Replica Volume 634
Pod 642

FIG. 6

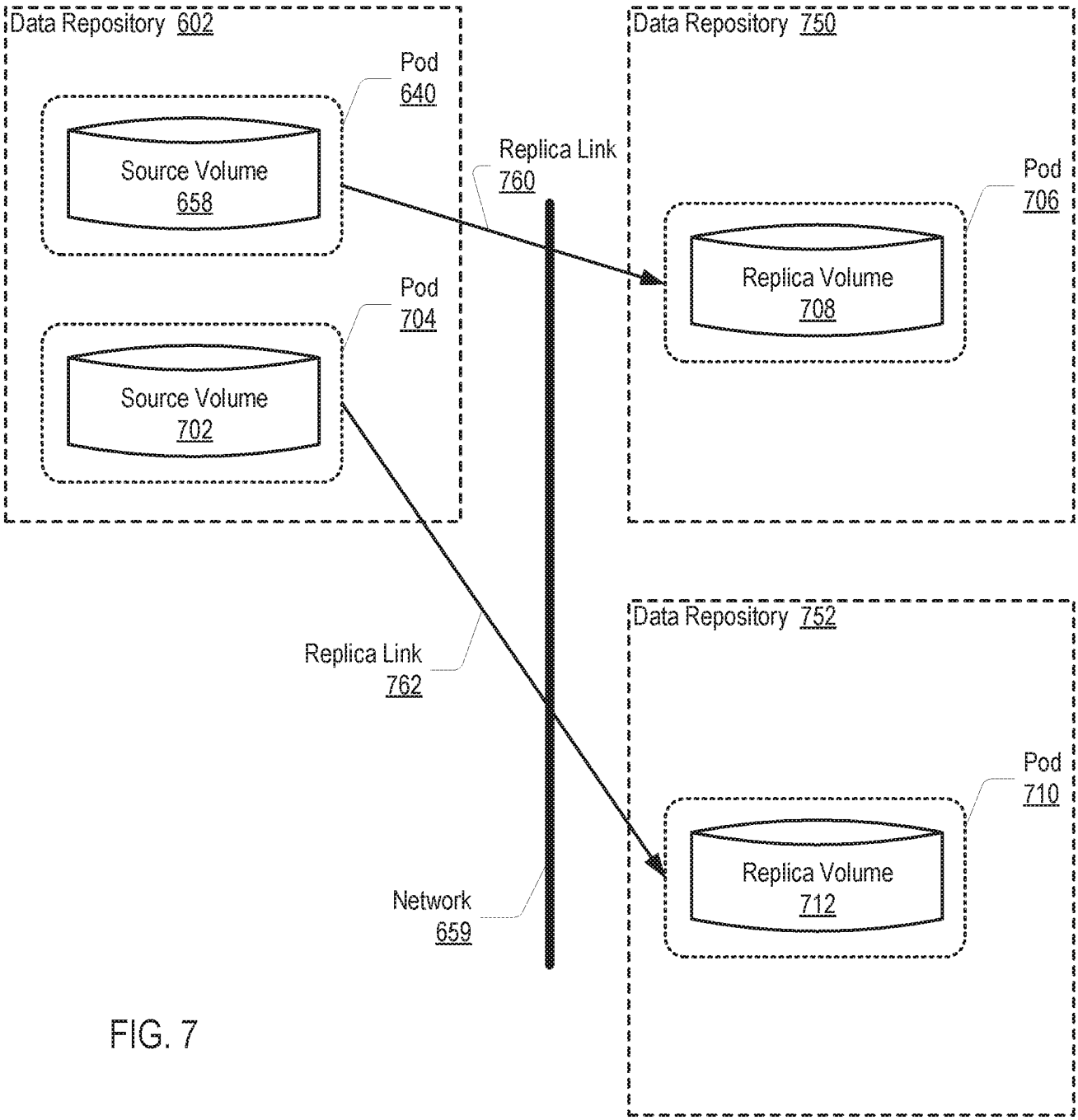

Data Repository 602
Pod 640
Source Volume 658
Pod 704
Source Volume 702
Replica Link 760
Replica Link 762
Network 659
Data Repository 750
Pod 706
Replica Volume 708
Data Repository 752
Pod 710
Replica Volume 712

FIG. 7

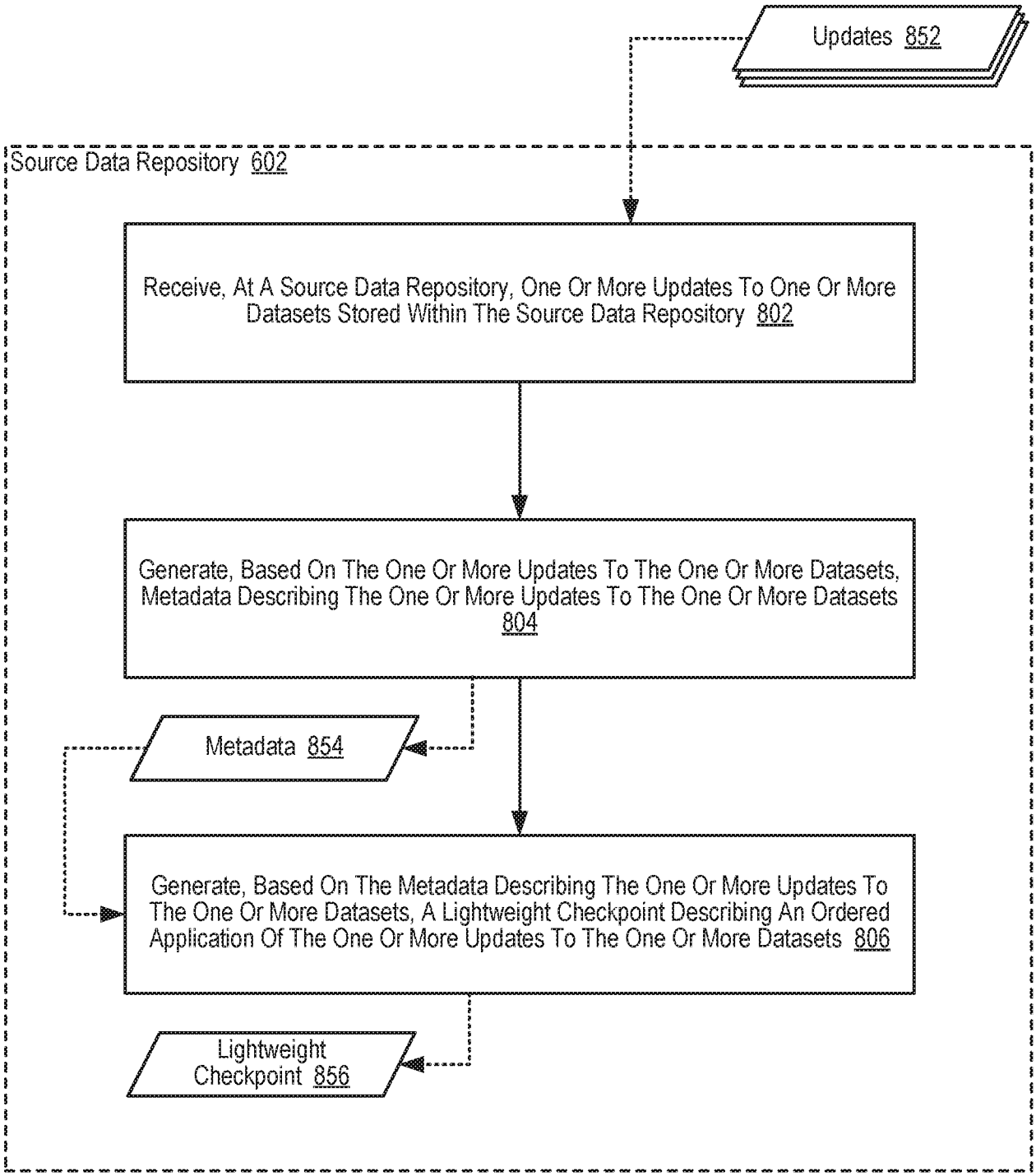

Updates 852
Source Data Repository 602
Receive, At A Source Data Repository, One Or More Updates To One Or More Datasets Stored Within The Source Data Repository 802
Generate, Based On The One Or More Updates To The One Or More Datasets, Metadata Describing The One Or More Updates To The One Or More Datasets 804
Metadata 854
Generate, Based On The Metadata Describing The One Or More Updates To The One Or More Datasets, A Lightweight Checkpoint Describing An Ordered Application Of The One Or More Updates To The One Or More Datasets 806
Lightweight Checkpoint 856

FIG. 8

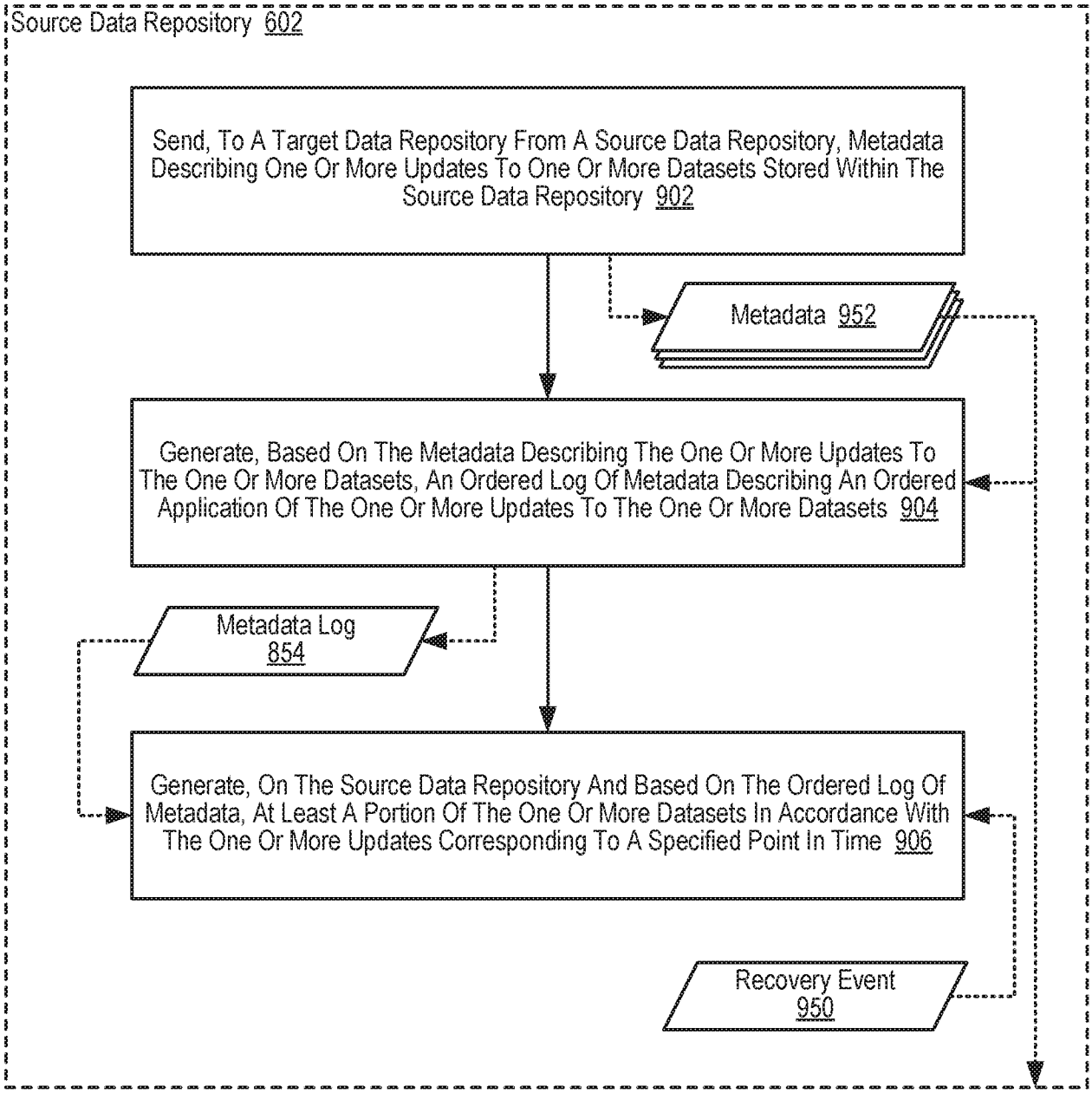

Source Data Repository 602
Send, To A Target Data Repository From A Source Data Repository, Metadata Describing One Or More Updates To One Or More Datasets Stored Within The Source Data Repository 902
Metadata 952
Generate, Based On The Metadata Describing The One Or More Updates To The One Or More Datasets, An Ordered Log Of Metadata Describing An Ordered Application Of The One Or More Updates To The One Or More Datasets 904
Metadata Log 854
Generate, On The Source Data Repository And Based On The Ordered Log Of Metadata, At Least A Portion Of The One Or More Datasets In Accordance With The One Or More Updates Corresponding To A Specified Point In Time 906
Recovery Event 950

FIG. 9

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2020/050408

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F11/14 G06F11/20
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 2016/205187 A1 (SORENSON III JAMES CHRISTOPHER [US] ET AL) 14 July 2016 (2016-07-14) abstract figures 1-2 paragraph [0041] - paragraph [0042] paragraph [0048] paragraph [0061] - paragraph [0081] paragraph [0114] - paragraph [0115] -----	1-3, 6-10, 13-17 4,5,11, 12,18-20
Y	US 2017/031776 A1 (REN JOANNE [US] ET AL) 2 February 2017 (2017-02-02) abstract paragraph [0073] -----	4,5,11, 12,19,20
	-/--	

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 November 2020	23/11/2020

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Lanchès, Philippe

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2020/050408

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 2018/231806 A1 (PURE STORAGE INC [US]) 20 December 2018 (2018-12-20) abstract paragraph [0067] paragraph [00193] paragraph [00204] -----	18

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2020/050408

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2016205187 A1	14-07-2016	CA 2840596 A1	03-01-2013
		CN 103930879 A	16-07-2014
		EP 2727001 A2	07-05-2014
		JP 6073878 B2	01-02-2017
		JP 6139718 B2	31-05-2017
		JP 2014529111 A	30-10-2014
		JP 2016129036 A	14-07-2016
		US 2013007219 A1	03-01-2013
		US 2016205187 A1	14-07-2016
		WO 2013003713 A2	03-01-2013
US 2017031776 A1	02-02-2017	NONE	
WO 2018231806 A1	20-12-2018	US 2018356989 A1	13-12-2018
		WO 2018231806 A1	20-12-2018