



- (51) **International Patent Classification:**
H04L 12/879 (2013.01) *H04Q 11/00* (2006.01)
H04L 12/861 (2013.01)
- (21) **International Application Number:**
PCT/US2017/038297
- (22) **International Filing Date:**
20 June 2017 (20.06.2017)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
62/365,969 22 July 2016 (22.07.2016) US
62/376,859 18 August 2016 (18.08.2016) US
62/427,268 29 November 2016 (29.11.2016) US
15/395,765 30 December 2016 (30.12.2016) US
- (71) **Applicant:** INTEL CORPORATION [US/US]; 2200 Mission College Boulevard, Santa Clara, California 95054 (US).
- (72) **Inventor:** MILLER, Steven C.; 2317 Tuscany Circle, Livermore, California 94550 (US).
- (74) **Agent:** KELLETT, Glen M.; Barnes & Thornburg LLP, 11 South Meridian Street, Indianapolis, Indiana 46204 (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, 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, SA,

(54) **Title:** TECHNOLOGIES FOR ACCELERATING DATA WRITES

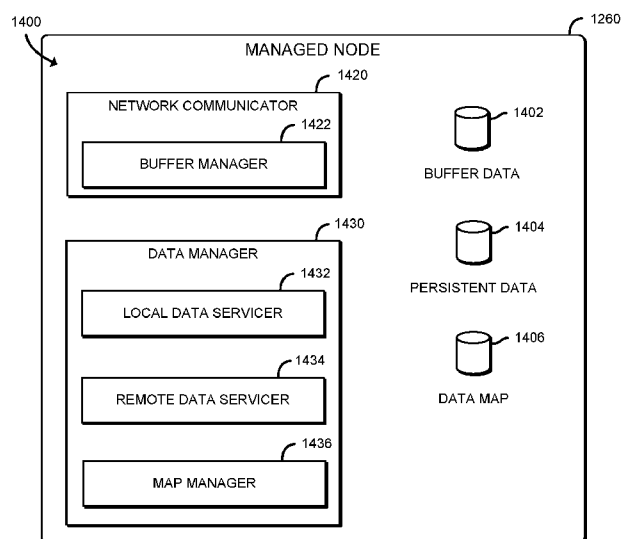


FIG. 14

(57) **Abstract:** Technologies for accelerating data writes include a managed node that includes a network interface controller that includes a power loss protected buffer and non-volatile memory. The managed node is to receive, through the network interface controller, a write request from a remote device. The write request includes a data block. The managed node is also to write the data block to the power loss protected buffer of the network interface controller, and send, in response to receipt of the data block and prior to a write of the data block to the non-volatile memory, an acknowledgement to the remote device. The acknowledgement is indicative of a successful write of the data block to the non-volatile memory. The managed node is also to write, after the acknowledgement has been sent, the data block from the power loss protected buffer to the non-volatile memory. Other embodiments are also described and claimed.

SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

TECHNOLOGIES FOR ACCELERATING DATA WRITES**CROSS-REFERENCE TO RELATED APPLICATIONS**

[0001] The present application claims priority to U.S. Utility Patent Application Serial No. 15/395,765, entitled “TECHNOLOGIES FOR ACCELERATING DATA WRITES,” which was filed on December 30, 2016 and which claims priority to U.S. Provisional Patent Application No. 62/365,969, filed July 22, 2016; U.S. Provisional Patent Application No. 62/376,859, filed August 18, 2016; and U.S. Provisional Patent Application No. 62/427,268, filed November 29, 2016.

BACKGROUND

[0002] In a typical cloud-based computing environment (e.g., a data center), compute nodes execute workloads (e.g., applications, processes, services, etc.) on behalf of customers. As the workloads are executed, requests may be issued to write and/or read blocks of data to one or more data storage devices (e.g., non-volatile memory) which may be remotely located from the one or more processors executing the corresponding workloads. For a typical write request, the data block may be sent through a network of the data center to one or more remotely located data storage devices. A processor local to each data storage device coordinates writing the received data block to the data storage device and, afterwards, sends an acknowledgement message back through the network indicating that the data block was successfully written to the data storage device. Once the workload receives acknowledgement that the data block has been safely written, the workload proceeds with other operations. However, the process of writing the data block to each data storage device consumes time and, as such, can adversely affect the quality of service of the workload.

BRIEF DESCRIPTION OF THE DRAWINGS

[0003] The concepts described herein are illustrated by way of example and not by way of limitation in the accompanying figures. For simplicity and clarity of illustration, elements illustrated in the figures are not necessarily drawn to scale. Where considered appropriate, reference labels have been repeated among the figures to indicate corresponding or analogous elements.

[0004] FIG. 1 is a diagram of a conceptual overview of a data center in which one or more techniques described herein may be implemented according to various embodiments;

[0005] FIG. 2 is a diagram of an example embodiment of a logical configuration of a rack of the data center of FIG. 1;

[0006] FIG. 3 is a diagram of an example embodiment of another data center in which one or more techniques described herein may be implemented according to various embodiments;

[0007] FIG. 4 is a diagram of another example embodiment of a data center in which one or more techniques described herein may be implemented according to various embodiments;

[0008] FIG. 5 is a diagram of a connectivity scheme representative of link-layer connectivity that may be established among various sleds of the data centers of FIGS. 1, 3, and 4;

[0009] FIG. 6 is a diagram of a rack architecture that may be representative of an architecture of any particular one of the racks depicted in FIGS. 1-4 according to some embodiments;

[0010] FIG. 7 is a diagram of an example embodiment of a sled that may be used with the rack architecture of FIG. 6;

[0011] FIG. 8 is a diagram of an example embodiment of a rack architecture to provide support for sleds featuring expansion capabilities;

[0012] FIG. 9 is a diagram of an example embodiment of a rack implemented according to the rack architecture of FIG. 8;

[0013] FIG. 10 is a diagram of an example embodiment of a sled designed for use in conjunction with the rack of FIG. 9;

[0014] FIG. 11 is a diagram of an example embodiment of a data center in which one or more techniques described herein may be implemented according to various embodiments;

[0015] FIG. 12 is a simplified block diagram of at least one embodiment of a system for accelerating data writes among a set of managed nodes;

[0016] FIG. 13 is a simplified block diagram of at least one embodiment of a managed node of the system of FIG. 12;

[0017] FIG. 14 is a simplified block diagram of at least one embodiment of an environment that may be established by a managed node of FIGS. 12 and 13; and

[0018] FIGS. 15-17 are a simplified flow diagram of at least one embodiment of a method for managing write acceleration that may be performed by a managed node of FIGS. 12-14.

DETAILED DESCRIPTION OF THE DRAWINGS

[0019] While the concepts of the present disclosure are susceptible to various modifications and alternative forms, specific embodiments thereof have been shown by way of

example in the drawings and will be described herein in detail. It should be understood, however, that there is no intent to limit the concepts of the present disclosure to the particular forms disclosed, but on the contrary, the intention is to cover all modifications, equivalents, and alternatives consistent with the present disclosure and the appended claims.

[0020] References in the specification to “one embodiment,” “an embodiment,” “an illustrative embodiment,” etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may or may not necessarily include that particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to effect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described. Additionally, it should be appreciated that items included in a list in the form of “at least one A, B, and C” can mean (A); (B); (C); (A and B); (A and C); (B and C); or (A, B, and C). Similarly, items listed in the form of “at least one of A, B, or C” can mean (A); (B); (C); (A and B); (A and C); (B and C); or (A, B, and C).

[0021] The disclosed embodiments may be implemented, in some cases, in hardware, firmware, software, or any combination thereof. The disclosed embodiments may also be implemented as instructions carried by or stored on a transitory or non-transitory machine-readable (e.g., computer-readable) storage medium, which may be read and executed by one or more processors. A machine-readable storage medium may be embodied as any storage device, mechanism, or other physical structure for storing or transmitting information in a form readable by a machine (e.g., a volatile or non-volatile memory, a media disc, or other media device).

[0022] In the drawings, some structural or method features may be shown in specific arrangements and/or orderings. However, it should be appreciated that such specific arrangements and/or orderings may not be required. Rather, in some embodiments, such features may be arranged in a different manner and/or order than shown in the illustrative figures. Additionally, the inclusion of a structural or method feature in a particular figure is not meant to imply that such feature is required in all embodiments and, in some embodiments, may not be included or may be combined with other features.

[0023] FIG. 1 illustrates a conceptual overview of a data center 100 that may generally be representative of a data center or other type of computing network in/for which one or more techniques described herein may be implemented according to various embodiments. As shown in FIG. 1, data center 100 may generally contain a plurality of racks, each of which may house

computing equipment comprising a respective set of physical resources. In the particular non-limiting example depicted in FIG. 1, data center 100 contains four racks 102A to 102D, which house computing equipment comprising respective sets of physical resources 105A to 105D. According to this example, a collective set of physical resources 106 of data center 100 includes the various sets of physical resources 105A to 105D that are distributed among racks 102A to 102D. Physical resources 106 may include resources of multiple types, such as – for example – processors, co-processors, accelerators, field-programmable gate arrays (FPGAs), memory, and storage. The embodiments are not limited to these examples.

[0024] The illustrative data center 100 differs from typical data centers in many ways. For example, in the illustrative embodiment, the circuit boards (“sleds”) on which components such as CPUs, memory, and other components are placed are designed for increased thermal performance. In particular, in the illustrative embodiment, the sleds are shallower than typical boards. In other words, the sleds are shorter from the front to the back, where cooling fans are located. This decreases the length of the path that air must travel across the components on the board. Further, the components on the sled are spaced further apart than in typical circuit boards, and the components are arranged to reduce or eliminate shadowing (i.e., one component in the air flow path of another component). In the illustrative embodiment, processing components such as the processors are located on a top side of a sled while near memory, such as dual in-line memory modules (DIMMs), are located on a bottom side of the sled. As a result of the enhanced airflow provided by this design, the components may operate at higher frequencies and power levels than in typical systems, thereby increasing performance. Furthermore, the sleds are configured to blindly mate with power and data communication cables in each rack 102A, 102B, 102C, 102D, enhancing their ability to be quickly removed, upgraded, reinstalled, and/or replaced. Similarly, individual components located on the sleds, such as processors, accelerators, memory, and data storage drives, are configured to be easily upgraded due to their increased spacing from each other. In the illustrative embodiment, the components additionally include hardware attestation features to prove their authenticity.

[0025] Furthermore, in the illustrative embodiment, the data center 100 utilizes a single network architecture (“fabric”) that supports multiple other network architectures including Ethernet and Omni-Path. The sleds, in the illustrative embodiment, are coupled to switches via optical fibers, which provide higher bandwidth and lower latency than typical twisted pair cabling (e.g., Category 5, Category 5e, Category 6, etc.). Due to the high bandwidth, low latency interconnections and network architecture, the data center 100 may, in use, pool resources, such as memory, accelerators (e.g., graphics accelerators, FPGAs, application specific integrated circuits (ASICs), etc.), and data storage drives that are physically

disaggregated, and provide them to compute resources (e.g., processors) on an as needed basis, enabling the compute resources to access the pooled resources as if they were local. The illustrative data center 100 additionally receives usage information for the various resources, predicts resource usage for different types of workloads based on past resource usage, and dynamically reallocates the resources based on this information.

[0026] The racks 102A, 102B, 102C, 102D of the data center 100 may include physical design features that facilitate the automation of a variety of types of maintenance tasks. For example, data center 100 may be implemented using racks that are designed to be robotically-accessed, and to accept and house robotically-manipulatable resource sleds. Furthermore, in the illustrative embodiment, the racks 102A, 102B, 102C, 102D include integrated power sources that receive a greater voltage than is typical for power sources. The increased voltage enables the power sources to provide additional power to the components on each sled, enabling the components to operate at higher than typical frequencies.

[0027] FIG. 2 illustrates an exemplary logical configuration of a rack 202 of the data center 100. As shown in FIG. 2, rack 202 may generally house a plurality of sleds, each of which may comprise a respective set of physical resources. In the particular non-limiting example depicted in FIG. 2, rack 202 houses sleds 204-1 to 204-4 comprising respective sets of physical resources 205-1 to 205-4, each of which constitutes a portion of the collective set of physical resources 206 comprised in rack 202. With respect to FIG. 1, if rack 202 is representative of – for example – rack 102A, then physical resources 206 may correspond to the physical resources 105A comprised in rack 102A. In the context of this example, physical resources 105A may thus be made up of the respective sets of physical resources, including physical storage resources 205-1, physical accelerator resources 205-2, physical memory resources 205-3, and physical compute resources 205-5 comprised in the sleds 204-1 to 204-4 of rack 202. The embodiments are not limited to this example. Each sled may contain a pool of each of the various types of physical resources (e.g., compute, memory, accelerator, storage). By having robotically accessible and robotically manipulatable sleds comprising disaggregated resources, each type of resource can be upgraded independently of each other and at their own optimized refresh rate.

[0028] FIG. 3 illustrates an example of a data center 300 that may generally be representative of one in/for which one or more techniques described herein may be implemented according to various embodiments. In the particular non-limiting example depicted in FIG. 3, data center 300 comprises racks 302-1 to 302-32. In various embodiments, the racks of data center 300 may be arranged in such fashion as to define and/or accommodate various access pathways. For example, as shown in FIG. 3, the racks of data center 300 may be

arranged in such fashion as to define and/or accommodate access pathways 311A, 311B, 311C, and 311D. In some embodiments, the presence of such access pathways may generally enable automated maintenance equipment, such as robotic maintenance equipment, to physically access the computing equipment housed in the various racks of data center 300 and perform automated maintenance tasks (e.g., replace a failed sled, upgrade a sled). In various embodiments, the dimensions of access pathways 311A, 311B, 311C, and 311D, the dimensions of racks 302-1 to 302-32, and/or one or more other aspects of the physical layout of data center 300 may be selected to facilitate such automated operations. The embodiments are not limited in this context.

[0029] FIG. 4 illustrates an example of a data center 400 that may generally be representative of one in/for which one or more techniques described herein may be implemented according to various embodiments. As shown in FIG. 4, data center 400 may feature an optical fabric 412. Optical fabric 412 may generally comprise a combination of optical signaling media (such as optical cabling) and optical switching infrastructure via which any particular sled in data center 400 can send signals to (and receive signals from) each of the other sleds in data center 400. The signaling connectivity that optical fabric 412 provides to any given sled may include connectivity both to other sleds in a same rack and sleds in other racks. In the particular non-limiting example depicted in FIG. 4, data center 400 includes four racks 402A to 402D. Racks 402A to 402D house respective pairs of sleds 404A-1 and 404A-2, 404B-1 and 404B-2, 404C-1 and 404C-2, and 404D-1 and 404D-2. Thus, in this example, data center 400 comprises a total of eight sleds. Via optical fabric 412, each such sled may possess signaling connectivity with each of the seven other sleds in data center 400. For example, via optical fabric 412, sled 404A-1 in rack 402A may possess signaling connectivity with sled 404A-2 in rack 402A, as well as the six other sleds 404B-1, 404B-2, 404C-1, 404C-2, 404D-1, and 404D-2 that are distributed among the other racks 402B, 402C, and 402D of data center 400. The embodiments are not limited to this example.

[0030] FIG. 5 illustrates an overview of a connectivity scheme 500 that may generally be representative of link-layer connectivity that may be established in some embodiments among the various sleds of a data center, such as any of example data centers 100, 300, and 400 of FIGS. 1, 3, and 4. Connectivity scheme 500 may be implemented using an optical fabric that features a dual-mode optical switching infrastructure 514. Dual-mode optical switching infrastructure 514 may generally comprise a switching infrastructure that is capable of receiving communications according to multiple link-layer protocols via a same unified set of optical signaling media, and properly switching such communications. In various embodiments, dual-mode optical switching infrastructure 514 may be implemented using one or more dual-mode

optical switches 515. In various embodiments, dual-mode optical switches 515 may generally comprise high-radix switches. In some embodiments, dual-mode optical switches 515 may comprise multi-ply switches, such as four-ply switches. In various embodiments, dual-mode optical switches 515 may feature integrated silicon photonics that enable them to switch communications with significantly reduced latency in comparison to conventional switching devices. In some embodiments, dual-mode optical switches 515 may constitute leaf switches 530 in a leaf-spine architecture additionally including one or more dual-mode optical spine switches 520.

[0031] In various embodiments, dual-mode optical switches may be capable of receiving both Ethernet protocol communications carrying Internet Protocol (IP packets) and communications according to a second, high-performance computing (HPC) link-layer protocol (e.g., Intel's Omni-Path Architecture's, Infiniband) via optical signaling media of an optical fabric. As reflected in FIG. 5, with respect to any particular pair of sleds 504A and 504B possessing optical signaling connectivity to the optical fabric, connectivity scheme 500 may thus provide support for link-layer connectivity via both Ethernet links and HPC links. Thus, both Ethernet and HPC communications can be supported by a single high-bandwidth, low-latency switch fabric. The embodiments are not limited to this example.

[0032] FIG. 6 illustrates a general overview of a rack architecture 600 that may be representative of an architecture of any particular one of the racks depicted in FIGS. 1 to 4 according to some embodiments. As reflected in FIG. 6, rack architecture 600 may generally feature a plurality of sled spaces into which sleds may be inserted, each of which may be robotically-accessible via a rack access region 601. In the particular non-limiting example depicted in FIG. 6, rack architecture 600 features five sled spaces 603-1 to 603-5. Sled spaces 603-1 to 603-5 feature respective multi-purpose connector modules (MPCMs) 616-1 to 616-5.

[0033] FIG. 7 illustrates an example of a sled 704 that may be representative of a sled of such a type. As shown in FIG. 7, sled 704 may comprise a set of physical resources 705, as well as an MPCM 716 designed to couple with a counterpart MPCM when sled 704 is inserted into a sled space such as any of sled spaces 603-1 to 603-5 of FIG. 6. Sled 704 may also feature an expansion connector 717. Expansion connector 717 may generally comprise a socket, slot, or other type of connection element that is capable of accepting one or more types of expansion modules, such as an expansion sled 718. By coupling with a counterpart connector on expansion sled 718, expansion connector 717 may provide physical resources 705 with access to supplemental computing resources 705B residing on expansion sled 718. The embodiments are not limited in this context.

[0034] FIG. 8 illustrates an example of a rack architecture 800 that may be representative of a rack architecture that may be implemented in order to provide support for sleds featuring expansion capabilities, such as sled 704 of FIG. 7. In the particular non-limiting example depicted in FIG. 8, rack architecture 800 includes seven sled spaces 803-1 to 803-7, which feature respective MPCMs 816-1 to 816-7. Sled spaces 803-1 to 803-7 include respective primary regions 803-1A to 803-7A and respective expansion regions 803-1B to 803-7B. With respect to each such sled space, when the corresponding MPCM is coupled with a counterpart MPCM of an inserted sled, the primary region may generally constitute a region of the sled space that physically accommodates the inserted sled. The expansion region may generally constitute a region of the sled space that can physically accommodate an expansion module, such as expansion sled 718 of FIG. 7, in the event that the inserted sled is configured with such a module.

[0035] FIG. 9 illustrates an example of a rack 902 that may be representative of a rack implemented according to rack architecture 800 of FIG. 8 according to some embodiments. In the particular non-limiting example depicted in FIG. 9, rack 902 features seven sled spaces 903-1 to 903-7, which include respective primary regions 903-1A to 903-7A and respective expansion regions 903-1B to 903-7B. In various embodiments, temperature control in rack 902 may be implemented using an air cooling system. For example, as reflected in FIG. 9, rack 902 may feature a plurality of fans 919 that are generally arranged to provide air cooling within the various sled spaces 903-1 to 903-7. In some embodiments, the height of the sled space is greater than the conventional “1U” server height. In such embodiments, fans 919 may generally comprise relatively slow, large diameter cooling fans as compared to fans used in conventional rack configurations. Running larger diameter cooling fans at lower speeds may increase fan lifetime relative to smaller diameter cooling fans running at higher speeds while still providing the same amount of cooling. The sleds are physically shallower than conventional rack dimensions. Further, components are arranged on each sled to reduce thermal shadowing (i.e., not arranged serially in the direction of air flow). As a result, the wider, shallower sleds allow for an increase in device performance because the devices can be operated at a higher thermal envelope (e.g., 250W) due to improved cooling (i.e., no thermal shadowing, more space between devices, more room for larger heat sinks, etc.).

[0036] MPCMs 916-1 to 916-7 may be configured to provide inserted sleds with access to power sourced by respective power modules 920-1 to 920-7, each of which may draw power from an external power source 921. In various embodiments, external power source 921 may deliver alternating current (AC) power to rack 902, and power modules 920-1 to 920-7 may be configured to convert such AC power to direct current (DC) power to be sourced to inserted

sleds. In some embodiments, for example, power modules 920-1 to 920-7 may be configured to convert 277-volt AC power into 12-volt DC power for provision to inserted sleds via respective MPCMs 916-1 to 916-7. The embodiments are not limited to this example.

[0037] MPCMs 916-1 to 916-7 may also be arranged to provide inserted sleds with optical signaling connectivity to a dual-mode optical switching infrastructure 914, which may be the same as – or similar to – dual-mode optical switching infrastructure 514 of FIG. 5. In various embodiments, optical connectors contained in MPCMs 916-1 to 916-7 may be designed to couple with counterpart optical connectors contained in MPCMs of inserted sleds to provide such sleds with optical signaling connectivity to dual-mode optical switching infrastructure 914 via respective lengths of optical cabling 922-1 to 922-7. In some embodiments, each such length of optical cabling may extend from its corresponding MPCM to an optical interconnect loom 923 that is external to the sled spaces of rack 902. In various embodiments, optical interconnect loom 923 may be arranged to pass through a support post or other type of load-bearing element of rack 902. The embodiments are not limited in this context. Because inserted sleds connect to an optical switching infrastructure via MPCMs, the resources typically spent in manually configuring the rack cabling to accommodate a newly inserted sled can be saved.

[0038] FIG. 10 illustrates an example of a sled 1004 that may be representative of a sled designed for use in conjunction with rack 902 of FIG. 9 according to some embodiments. Sled 1004 may feature an MPCM 1016 that comprises an optical connector 1016A and a power connector 1016B, and that is designed to couple with a counterpart MPCM of a sled space in conjunction with insertion of MPCM 1016 into that sled space. Coupling MPCM 1016 with such a counterpart MPCM may cause power connector 1016 to couple with a power connector comprised in the counterpart MPCM. This may generally enable physical resources 1005 of sled 1004 to source power from an external source, via power connector 1016 and power transmission media 1024 that conductively couples power connector 1016 to physical resources 1005.

[0039] Sled 1004 may also include dual-mode optical network interface circuitry 1026. Dual-mode optical network interface circuitry 1026 may generally comprise circuitry that is capable of communicating over optical signaling media according to each of multiple link-layer protocols supported by dual-mode optical switching infrastructure 914 of FIG. 9. In some embodiments, dual-mode optical network interface circuitry 1026 may be capable both of Ethernet protocol communications and of communications according to a second, high-performance protocol. In various embodiments, dual-mode optical network interface circuitry 1026 may include one or more optical transceiver modules 1027, each of which may be capable

of transmitting and receiving optical signals over each of one or more optical channels. The embodiments are not limited in this context.

[0040] Coupling MPCM 1016 with a counterpart MPCM of a sled space in a given rack may cause optical connector 1016A to couple with an optical connector comprised in the counterpart MPCM. This may generally establish optical connectivity between optical cabling of the sled and dual-mode optical network interface circuitry 1026, via each of a set of optical channels 1025. Dual-mode optical network interface circuitry 1026 may communicate with the physical resources 1005 of sled 1004 via electrical signaling media 1028. In addition to the dimensions of the sleds and arrangement of components on the sleds to provide improved cooling and enable operation at a relatively higher thermal envelope (e.g., 250W), as described above with reference to FIG. 9, in some embodiments, a sled may include one or more additional features to facilitate air cooling, such as a heat pipe and/or heat sinks arranged to dissipate heat generated by physical resources 1005. It is worthy of note that although the example sled 1004 depicted in FIG. 10 does not feature an expansion connector, any given sled that features the design elements of sled 1004 may also feature an expansion connector according to some embodiments. The embodiments are not limited in this context.

[0041] FIG. 11 illustrates an example of a data center 1100 that may generally be representative of one in/for which one or more techniques described herein may be implemented according to various embodiments. As reflected in FIG. 11, a physical infrastructure management framework 1150A may be implemented to facilitate management of a physical infrastructure 1100A of data center 1100. In various embodiments, one function of physical infrastructure management framework 1150A may be to manage automated maintenance functions within data center 1100, such as the use of robotic maintenance equipment to service computing equipment within physical infrastructure 1100A. In some embodiments, physical infrastructure 1100A may feature an advanced telemetry system that performs telemetry reporting that is sufficiently robust to support remote automated management of physical infrastructure 1100A. In various embodiments, telemetry information provided by such an advanced telemetry system may support features such as failure prediction/prevention capabilities and capacity planning capabilities. In some embodiments, physical infrastructure management framework 1150A may also be configured to manage authentication of physical infrastructure components using hardware attestation techniques. For example, robots may verify the authenticity of components before installation by analyzing information collected from a radio frequency identification (RFID) tag associated with each component to be installed. The embodiments are not limited in this context.

[0042] As shown in FIG. 11, the physical infrastructure 1100A of data center 1100 may comprise an optical fabric 1112, which may include a dual-mode optical switching infrastructure 1114. Optical fabric 1112 and dual-mode optical switching infrastructure 1114 may be the same as – or similar to – optical fabric 412 of FIG. 4 and dual-mode optical switching infrastructure 514 of FIG. 5, respectively, and may provide high-bandwidth, low-latency, multi-protocol connectivity among sleds of data center 1100. As discussed above, with reference to FIG. 1, in various embodiments, the availability of such connectivity may make it feasible to disaggregate and dynamically pool resources such as accelerators, memory, and storage. In some embodiments, for example, one or more pooled accelerator sleds 1130 may be included among the physical infrastructure 1100A of data center 1100, each of which may comprise a pool of accelerator resources – such as co-processors and/or FPGAs, for example – that is globally accessible to other sleds via optical fabric 1112 and dual-mode optical switching infrastructure 1114.

[0043] In another example, in various embodiments, one or more pooled storage sleds 1132 may be included among the physical infrastructure 1100A of data center 1100, each of which may comprise a pool of storage resources that is available globally accessible to other sleds via optical fabric 1112 and dual-mode optical switching infrastructure 1114. In some embodiments, such pooled storage sleds 1132 may comprise pools of solid-state storage devices such as solid-state drives (SSDs). In various embodiments, one or more high-performance processing sleds 1134 may be included among the physical infrastructure 1100A of data center 1100. In some embodiments, high-performance processing sleds 1134 may comprise pools of high-performance processors, as well as cooling features that enhance air cooling to yield a higher thermal envelope of up to 250W or more. In various embodiments, any given high-performance processing sled 1134 may feature an expansion connector 1117 that can accept a far memory expansion sled, such that the far memory that is locally available to that high-performance processing sled 1134 is disaggregated from the processors and near memory comprised on that sled. In some embodiments, such a high-performance processing sled 1134 may be configured with far memory using an expansion sled that comprises low-latency SSD storage. The optical infrastructure allows for compute resources on one sled to utilize remote accelerator/FPGA, memory, and/or SSD resources that are disaggregated on a sled located on the same rack or any other rack in the data center. The remote resources can be located one switch jump away or two-switch jumps away in the spine-leaf network architecture described above with reference to FIG. 5. The embodiments are not limited in this context.

[0044] In various embodiments, one or more layers of abstraction may be applied to the physical resources of physical infrastructure 1100A in order to define a virtual infrastructure,

such as a software-defined infrastructure 1100B. In some embodiments, virtual computing resources 1136 of software-defined infrastructure 1100B may be allocated to support the provision of cloud services 1140. In various embodiments, particular sets of virtual computing resources 1136 may be grouped for provision to cloud services 1140 in the form of SDI services 1138. Examples of cloud services 1140 may include – without limitation – software as a service (SaaS) services 1142, platform as a service (PaaS) services 1144, and infrastructure as a service (IaaS) services 1146.

[0045] In some embodiments, management of software-defined infrastructure 1100B may be conducted using a virtual infrastructure management framework 1150B. In various embodiments, virtual infrastructure management framework 1150B may be designed to implement workload fingerprinting techniques and/or machine-learning techniques in conjunction with managing allocation of virtual computing resources 1136 and/or SDI services 1138 to cloud services 1140. In some embodiments, virtual infrastructure management framework 1150B may use/consult telemetry data in conjunction with performing such resource allocation. In various embodiments, an application/service management framework 1150C may be implemented in order to provide quality of service (QoS) management capabilities for cloud services 1140. The embodiments are not limited in this context.

[0046] As shown in FIG. 12, an illustrative system 1210 for accelerating data writes among a set of managed nodes 1260 includes an orchestrator server 1240 in communication with the set of managed nodes 1260. Each managed node 1260 may be embodied as an assembly of resources (e.g., physical resources 206), such as compute resources (e.g., physical compute resources 205-4), storage resources (e.g., physical storage resources 205-1), accelerator resources (e.g., physical accelerator resources 205-2), or other resources (e.g., physical memory resources 205-3) from the same or different sleds (e.g., the sleds 204-1, 204-2, 204-3, 204-4, etc.) or racks (e.g., one or more of racks 302-1 through 302-32). Each managed node 1260 may be established, defined, or “spun up” by the orchestrator server 1240 at the time a workload is to be assigned to the managed node 1260 or at any other time, and may exist regardless of whether any workloads are presently assigned to the managed node 1260. The system 1210 may be implemented in accordance with the data centers 100, 300, 400, 1100 described above with reference to FIGS. 1, 3, 4, and 11. In the illustrative embodiment, the set of managed nodes 1260 includes managed nodes 1250, 1252, and 1254. While three managed nodes 1260 are shown in the set, it should be understood that in other embodiments, the set may include a different number of managed nodes 1260 (e.g., tens of thousands). The system 1210 may be located in a data center and provide storage and compute services (e.g., cloud services) to a client device 1220 that is in communication with the system 1210 through a network 1230.

The orchestrator server 1240 may support a cloud operating environment, such as OpenStack, and assign workloads to the managed nodes 1260 for execution.

[0047] The managed nodes 1260 may execute the workloads, such as in virtual machines or containers, on behalf of a user of the client device 1220. As the workloads are executed, requests may be issued to write data, referred to herein as data blocks, and/or read data blocks. Given that the physical data storage resources 205-1 may be located on a different sled 204-1 than the compute resources (e.g., one or more processors) 205-4 used by a managed node 1260 to execute a workload, the requests may be transmitted through the network 1230 from the compute resources 205-4 to one or more physical storage resources 205-1. The storage resources 205-1 of the managed node 1260 may receive, with a network interface controller, a request that includes a data block to be written, write the data block to a power loss protected buffer (e.g., volatile memory) of the network interface controller, and send an acknowledgement message back through the network 1230 indicative of successful storage of the data block to non-volatile memory, and subsequently write the data block to the non-volatile memory. In some embodiments, a master storage sled of physical storage resources may write the data block to the power loss protected buffer, and to dynamic random access memory (DRAM), and coordinate replication of the data block in one or more “follower” storage sleds. Further, in some embodiments, in storing the data block, one or more of the storage sleds may encrypt and shard (e.g., partition) the data block. By sending the acknowledgement when the data block is in the power loss protected buffer of the network interface controller, rather than after the data block has actually been written to the relatively slower non-volatile memory of the one or more physical storage resources 205-1, the managed node 1260 reduces the amount of time that the workload waits for the acknowledgement and improves the overall quality of service (e.g., latency, throughput, etc.) provided by the workload.

[0048] Referring now to FIG. 13, the managed node 1260 may be embodied as any type of compute device capable of performing the functions described herein, including executing a workload, writing data blocks, and reading data blocks. For example, the managed node 1260 may be embodied as a computer, a distributed computing system, one or more sleds (e.g., the sleds 204-1, 204-2, 204-3, 204-4, etc.), a server (e.g., stand-alone, rack-mounted, blade, etc.), a multiprocessor system, a network appliance (e.g., physical or virtual), a desktop computer, a workstation, a laptop computer, a notebook computer, a processor-based system, or a network appliance. As shown in FIG. 13, the illustrative managed node 1260 includes a central processing unit (CPU) 1302, a main memory 1304, an input/output (I/O) subsystem 1306, communication circuitry 1308, and one or more data storage devices 1314. Of course, in other embodiments, the managed node 1260 may include other or additional components, such as

those commonly found in a computer (e.g., display, peripheral devices, etc.). Additionally, in some embodiments, one or more of the illustrative components may be incorporated in, or otherwise form a portion of, another component. For example, in some embodiments, the main memory 1304, or portions thereof, may be incorporated in the CPU 1302.

[0049] The CPU 1302 may be embodied as any type of processor capable of performing the functions described herein. The CPU 1302 may be embodied as a single or multi-core processor(s), a microcontroller, or other processor or processing/controlling circuit. In some embodiments, the CPU 1302 may be embodied as, include, or be coupled to a field programmable gate array (FPGA), an application specific integrated circuit (ASIC), reconfigurable hardware or hardware circuitry, or other specialized hardware to facilitate performance of the functions described herein. As discussed above, the managed node 1260 may include resources distributed across multiple sleds and in such embodiments, the CPU 1302 may include portions thereof located on the same sled or different sled. Similarly, the main memory 1304 may be embodied as any type of volatile (e.g., dynamic random access memory (DRAM), etc.) or non-volatile memory or data storage capable of performing the functions described herein. In some embodiments, all or a portion of the main memory 1304 may be integrated into the CPU 1302. In operation, the main memory 1304 may store various software and data used during operation, such as data blocks and a map of locations of data blocks among different data storage devices of the managed node 1260 and/or other managed nodes 1260, operating systems, applications, programs, libraries, and drivers. As discussed above, the managed node 1260 may include resources distributed across multiple sleds and in such embodiments, the main memory 1304 may include portions thereof located on the same sled or different sled.

[0050] The I/O subsystem 1306 may be embodied as circuitry and/or components to facilitate input/output operations with the CPU 1302, the main memory 1304, and other components of the managed node 1260. For example, the I/O subsystem 1306 may be embodied as, or otherwise include, memory controller hubs, input/output control hubs, integrated sensor hubs, firmware devices, communication links (e.g., point-to-point links, bus links, wires, cables, light guides, printed circuit board traces, etc.), and/or other components and subsystems to facilitate the input/output operations. In some embodiments, the I/O subsystem 1306 may form a portion of a system-on-a-chip (SoC) and be incorporated, along with one or more of the CPU 1302, the main memory 1304, and other components of the managed node 1260, on a single integrated circuit chip.

[0051] The communication circuitry 1308 may be embodied as any communication circuit, device, or collection thereof, capable of enabling communications over the network

1230 between the managed node 1260 and another compute device (e.g., the orchestrator server 1240 and/or one or more other managed nodes 1260). The communication circuitry 1308 may be configured to use any one or more communication technology (e.g., wired or wireless communications) and associated protocols (e.g., Ethernet, Bluetooth®, Wi-Fi®, WiMAX, etc.) to effect such communication.

[0052] The illustrative communication circuitry 1308 includes a network interface controller (NIC) 1310, which may also be referred to as a host fabric interface (HFI). The NIC 1310 may be embodied as one or more add-in-boards, daughtercards, network interface cards, controller chips, chipsets, or other devices that may be used by the managed node 1260 to connect with another compute device (e.g., the orchestrator server 1240 and/or physical resources of one or more managed nodes 1260). In some embodiments, the NIC 1310 may be embodied as part of a system-on-a-chip (SoC) that includes one or more processors, or included on a multichip package that also contains one or more processors. In some embodiments, the NIC 1310 may include a processor (not shown) local to the NIC 1310. In such embodiments, the local processor of the NIC 1310 may be capable of performing one or more of the functions of the CPU 1302 described herein. Additionally, the NIC 1310 includes a power loss protected buffer 1312 which may be embodied as any volatile local memory device that, when a power loss imminent condition is detected, may write any data present in the power loss protected buffer to non-volatile memory (e.g., to one or more of the data storage devices 1314). The power loss protected buffer 1312 may include an independent power supply, such as capacitors or batteries that allow the power loss protected buffer 1312 to operate for a period of time even after power to the managed node 1260 has been interrupted. As discussed above, the managed node may include resources distributed across multiple sleds and in such embodiments, the communication circuitry 1308 may include portions thereof located on the same sled or different sled. In the illustrative embodiment, the NIC 1310 in every sled having physical storage resources 205-1 (e.g., data storage devices 1314) includes the power loss protected buffer 1312.

[0053] The one or more illustrative data storage devices 1314, may be embodied as any type of devices configured for short-term or long-term storage of data such as, for example, solid-state drives (SSDs), hard disk drives, memory cards, and/or other memory devices and circuits. Each data storage device 1314 may include a system partition that stores data and firmware code for the data storage device 1314. Each data storage device 1314 may also include an operating system partition that stores data files and executables for an operating system. In the illustrative embodiment, each data storage device 1314 includes non-volatile memory. Non-volatile memory may be embodied as any type of data storage capable of storing

data in a persistent manner (even if power is interrupted to the non-volatile memory). For example, the non-volatile memory may be embodied as Flash memory (e.g., NAND memory). In other examples, the non-volatile memory may be embodied as any combination of memory devices that use chalcogenide phase change material (e.g., chalcogenide glass), or other types of byte-addressable, write-in-place non-volatile memory, ferroelectric transistor random-access memory (FeTRAM), nanowire-based non-volatile memory, phase change memory (PCM), memory that incorporates memristor technology, magnetoresistive random-access memory (MRAM) or Spin Transfer Torque (STT)-MRAM.

[0054] Additionally, the managed node 1260 may include a display 1316. The display 1316 may be embodied as, or otherwise use, any suitable display technology including, for example, a liquid crystal display (LCD), a light emitting diode (LED) display, a cathode ray tube (CRT) display, a plasma display, and/or other display usable in a compute device. The display 1316 may include a touchscreen sensor that uses any suitable touchscreen input technology to detect the user's tactile selection of information displayed on the display including, but not limited to, resistive touchscreen sensors, capacitive touchscreen sensors, surface acoustic wave (SAW) touchscreen sensors, infrared touchscreen sensors, optical imaging touchscreen sensors, acoustic touchscreen sensors, and/or other type of touchscreen sensors.

[0055] Additionally or alternatively, the managed node 1260 may include one or more peripheral devices 1318. Such peripheral devices 1318 may include any type of peripheral device commonly found in a compute device such as speakers, a mouse, a keyboard, and/or other input/output devices, interface devices, and/or other peripheral devices.

[0056] The client device 1220 and the orchestrator server 1240 may have components similar to those described in FIG. 13, with the exception that the power loss protected buffer 1312 may be absent in the client device 1220 and/or the orchestrator server 1240. The description of those components of the managed node 1260 is equally applicable to the description of components of the client device 1220 and the orchestrator server 1240 and is not repeated herein for clarity of the description. Further, it should be appreciated that any of the client device 1220 and the orchestrator server 1240 may include other components, sub-components, and devices commonly found in a computing device, which are not discussed above in reference to the managed node 1260 and not discussed herein for clarity of the description.

[0057] As described above, the client device 1220, the orchestrator server 1240 and the managed nodes 1260 are illustratively in communication via the network 1230, which may be embodied as any type of wired or wireless communication network, including global networks

(e.g., the Internet), local area networks (LANs) or wide area networks (WANs), cellular networks (e.g., Global System for Mobile Communications (GSM), 3G, Long Term Evolution (LTE), Worldwide Interoperability for Microwave Access (WiMAX), etc.), digital subscriber line (DSL) networks, cable networks (e.g., coaxial networks, fiber networks, etc.), or any combination thereof.

[0058] Referring now to FIG. 14, in the illustrative embodiment, the managed node 1260 may establish an environment 1400 during operation. The illustrative environment 1400 includes a network communicator 1420 and a data manager 1430. Each of the components of the environment 1400 may be embodied as hardware, firmware, software, or a combination thereof. As such, in some embodiments, one or more of the components of the environment 1400 may be embodied as circuitry or a collection of electrical devices (e.g., network communicator circuitry 1420, data manager circuitry 1430, etc.). It should be appreciated that, in such embodiments, one or more of the network communicator circuitry 1420 or the data manager circuitry 1430 may form a portion of one or more of the CPU 1302, the main memory 1304, the I/O subsystem 1306, the communication circuitry 1308, and/or other components of the managed node 1260. In the illustrative embodiment, the environment 1400 includes buffer data 1402 which may be embodied as any data (e.g., data blocks) present in the power loss protected buffer 1312 of the NIC 1310. The environment 1400, in the illustrative embodiment, also includes persistent data 1404 which may be embodied as any data that has been written to non-volatile memory of the managed node 1260, such as data blocks that have been written from the power loss protected buffer 1312 (e.g., the buffer data 1402) to one or more of the data storage devices 1314. Additionally, in the illustrative embodiment, the environment 1400 includes a data map 1406 which may be embodied as any data indicative of locations where data blocks have been stored in the data storage devices 1314 (i.e., non-volatile memory) of the managed node 1260 and/or in one or more other managed nodes 1260. In the illustrative embodiment, each data block is identified by a key (e.g., a unique identifier, such as an alphanumeric code), such that the key and the corresponding data block form a key-value pair.

[0059] In the illustrative environment 1400, the network communicator 1420, which may be embodied as hardware, firmware, software, virtualized hardware, emulated architecture, and/or a combination thereof as discussed above, is configured to facilitate inbound and outbound network communications (e.g., network traffic, network packets, network flows, etc.) to and from the managed node 1260, respectively and to accelerate data writes requested through the network 1230. To do so, the network communicator 1420 is configured to receive and process data packets from one system or computing device (e.g., the orchestrator server 1240, a managed node 1260, etc.) and to prepare and send data packets to another computing

device or system (e.g., another managed node 1260). Accordingly, in some embodiments, at least a portion of the functionality of the network communicator 1420 may be performed by the communication circuitry 1308, and, in the illustrative embodiment, by the NIC 1310. In the illustrative embodiment, the network communicator 1420 includes a buffer manager 1422, which, in the illustrative embodiment, is configured to store a received data block from a write request in the power loss protected buffer 1312, send an early acknowledgement message through the network 1230 in response to the write request, indicating that the data block has been written to non-volatile memory, and coordinate writing the data block to the non-volatile memory (e.g., from the buffer data 1402 to the persistent data 1404).

[0060] The data manager 1430, which may be embodied as hardware, firmware, software, virtualized hardware, emulated architecture, and/or a combination thereof as discussed above, is configured to manage writing and reading of data to resources of the managed node 1260 (e.g., the data storage devices 1314) and/or to and from other managed nodes 1260. To do so, in the illustrative embodiment, the data manager 1430 includes a local data servicer 1432, a remote data servicer 1434, and a map manager 1436. The local data servicer 1432, in the illustrative embodiment, is configured to write data blocks and associated keys to the one or more data storage devices 1314 of the managed node 1260 and/or read data blocks from the one or more data storage devices 1314 of the managed node 1260. The remote data servicer 1434, in the illustrative embodiment, is configured to write data blocks and/or read data blocks to and from the data storage devices 1314 of one or more other managed nodes 1260 by issuing corresponding requests and receiving corresponding responses through the network 1230. The map manager 1436, in the illustrative embodiment, is configured to track where data blocks are stored among the data storage devices 1314 of the managed node 1260 and/or other managed nodes 1260. In doing so, the map manager 1436 may store keys in association with location identifiers, such as unique identifiers of data storage devices 1314 in which the data blocks are stored, and/or logical block addresses of the data blocks.

[0061] It should be appreciated that each of the local data servicer 1432, the remote data servicer 1434, and the map manager 1436 may be separately embodied as hardware, firmware, software, virtualized hardware, emulated architecture, and/or a combination thereof and may be distributed across multiple sleds. For example, the local data servicer 1432 may be embodied as a hardware component, while the remote data servicer 1434 and the map manager 1436 are embodied as virtualized hardware components or as some other combination of hardware, firmware, software, virtualized hardware, emulated architecture, and/or a combination thereof.

[0062] Referring now to FIG. 15, in use, the managed node 1260 may execute a method 1500 for managing write acceleration to improve the quality of service of workloads. The

method 1500 begins with block 1502, in which the managed node 1260 determines whether to manage write acceleration. In the illustrative embodiment, the managed node 1260 determines to manage write acceleration if the managed node 1260 is powered on and has access to (e.g., locally and/or through the network 1230) the one or more data storage devices 1314 which, in the illustrative embodiment, are located on different sleds. The managed node 1260 may, in other embodiments, determine whether to manage write acceleration based on other factors. Regardless, in response to a determination to manage write acceleration, in the illustrative embodiment, the method 1500 advances to block 1504 in which the managed node 1260 may receive, at the network interface controller 1310, a request from a remote device (e.g., a compute resource 205-3 on a different sled) to write a data block. In doing so, the managed node 1260 may receive a key associated with the data block to be written, as indicated in block 1506. In block 1508, the managed node determines whether a write request has been received. In response to a determination that a write request has been received, the method 1500 advances to block 1510 in which a storage sled of the managed node 1260 writes the data block to the power loss protected buffer 1312 of the network interface controller 1310. In doing so, the storage sled may additionally write the data block to DRAM (e.g., main memory 1304), as indicated in block 1512. The storage sled may do so in embodiments in which the storage sled is a “master” storage sled that is to coordinate storage of the data block among itself and other “follower” storage sleds. As indicated in block 1514, the storage sled of the managed node 1260 may additionally shard (e.g., partition) and encode the data block with one or more error correction codes. In the remainder of this description, any discussion of a storage sled performing an operation on a data block should be interpreted to mean performing an operation on a data block or a shard. Further, as indicated in block 1516, the storage sled of the managed node 1260 may forward the data block and/or shards (e.g., if the storage sled performed block 1514) to one or more follower storage sleds for storage thereon. As indicated in block 1518, the storage sled may then receive acknowledgements from the follower storage sleds.

[0063] Subsequently, in block 1520, the storage sled of the managed node 1260 sends an acknowledgement from the network interface controller 1310 of the storage sled to the remote device (e.g., the sled with compute resources) that originally sent the write request. Subsequently, or if the storage sled determined at block 1522 that a write request was not received, the storage sleds of the managed node 1260 determine whether to write one or more data blocks from their power loss protected buffer 1312 to their non-volatile memory (e.g., one or more of the data storage devices 1314), as indicated in block 1522. The storage sleds may write the data block immediately, or may write the data block to non-volatile memory based on one or more conditions, as described herein. For example, a storage sled of the managed node

1260 may determine whether the number of data blocks present in its power loss protected buffer 1312 satisfies a predefined threshold (e.g., is at least four data blocks), as indicated in block 1524. Additionally or alternatively, the storage sled of the managed node 1260 may determine whether the amount of unused (e.g., free) memory in the power loss protected buffer 1312 satisfies a predefined threshold (e.g., the amount of free memory is less than the size of a data block), as indicated in block 1526. A storage sled of the managed node 1260 may, additionally or alternatively, determine whether a predefined time interval has elapsed, as indicated in block 1528. For example, a storage sled of the managed node 1260 may be configured to write any previously unwritten data blocks from the power loss protected buffer 1312 to the non-volatile memory every second, or at any other time interval. As indicated in block 1530, in determining whether to write the data blocks from the power loss protected buffer 1312 to the non-volatile memory, a storage sled of the managed node 1260 may determine whether a power loss imminent condition is present (e.g., power to the NIC 1310 has been interrupted and the NIC 1310 is operating on reserve energy stored in a capacitor or other energy storage device). Subsequently, the method 1500 advances to block 1532 of FIG. 16, in which the storage sleds of the managed node 1260 perform additional operations based on whether each storage sled of the managed node 1260 has determined to write data blocks from the power loss protected buffer 1312 to the non-volatile memory (e.g., the one or more data storage devices 1314).

[0064] Referring now to FIG. 16, if a storage sled of the managed node 1260 determines to write one or more data blocks from the power loss protected buffer 1312, the method 1500 advances to block 1534 in which the storage sled of the managed node 1260 writes the one or more data blocks to the non-volatile memory (e.g., the one or more data storage devices 1314). In doing so, as indicated in block 1536, the storage managed node 1260 may write the one or more data blocks in association with keys received with them, as described with reference to block 1506 of FIG. 15. By writing the data blocks in association with their corresponding keys, the managed node 1260 forms a key-pair in the non-volatile memory for each written data block. In block 1538, the managed node 1260 may update the data map 1406 to indicate the locations (e.g., logical block addresses) within the one or more data storage devices 1314 of the storage sleds and/or network addresses (e.g., media access control addresses) of data storage devices 1314, where each data block was stored, along with the corresponding keys. As indicated in block 1540, each follower storage sled of the managed node 1260 may send a completion indicator to the master storage sled, indicating that the data block has been written to non-volatile memory. As indicated in block 1542, the master storage sled may receive the completion indicators through the network 1230. Further, as indicated in block 1544, the

storage sled may free up the power loss protected buffer 1312 of its own NIC 1310 and, if the storage sled is a master storage sled, send a notification to the follower storage sleds to free up their corresponding power loss protected buffers 1312.

[0065] Subsequently, or if the storage sleds of the managed node 1260 determined in block 1532 not to write any data blocks, the managed node 1260 may receive a request to read a data block, as indicated in block 1546. In doing so, the managed node 1260 may receive a request from another managed node 1260 (e.g., a requestor managed node 1260), as indicated in block 1548. Alternatively, the managed node 1260 may receive the request from the present managed node 1260 executing the workload, as indicated in block 1550. In receiving the request, the managed node 1260 may receive the key associated with the data block to be read, as indicated in block 1552. In block 1554, the managed node 1260 determines whether a read request has been received. If not, the method 1500 returns to block 1502 of FIG. 15, in which the managed node 1260 determines whether to continue managing write acceleration. Otherwise, the method 1500 advances to block 1556, in which the managed node 1260 determines whether the requested data block is stored in the non-volatile memory of the present managed node 1260 (e.g., in the one or more data storage devices 1314 of one or more storage sleds). In doing so, as indicated in block 1558, the managed node 1260 may search the data map 1406 for the key. In block 1560, the managed node 1260 determines whether the data block is available in the non-volatile memory (e.g., whether the key was found in the data map 1406). If not, the method 1500 advances to block 1562 of FIG. 17 in which the managed node 1260 determines whether the data block is present in the power loss protected buffer 1312 of a storage sled or the DRAM (e.g., memory 1304) of a storage sled (e.g., the master storage sled). Otherwise, the method advances to block 1570 of FIG. 17, in which the managed node 1260 reads the data block from the location where the data block was found in block 1556 of FIG. 16.

[0066] Referring now to FIG. 17, in determining whether the data block is present in the power loss protected buffer 1312 or the DRAM 1304, the managed node 1260 may search the power loss protected buffer 1312 and the DRAM 1304 for the key, as indicated in block 1564. Subsequently, in block 1566, the managed node 1260 determines whether the data block is available in the power loss protected buffer 1312 or the DRAM 1304 (e.g., whether the key was found in the power loss protected buffer 1312). If not, the method 1500 advances to block 1568 in which the managed node 1260 returns an error message in response to the request, indicating that the data block was not found. Subsequently, the method 1500 loops back to block 1502 of FIG. 15 in which the managed node 1260 determines whether to continue managing write acceleration.

[0067] If the managed node 1260 instead determined that the data block was found in the power loss protected buffer 1312 or the DRAM 1304, or as stated above, if the managed node 1260 determined that the data block is present in the non-volatile memory, the managed node 1260, in block 1570, reads the data block from the corresponding location where the data block was determined to exist. In doing so, the managed node 1260 may read the data block from the non-volatile memory (e.g., the one or more data storage devices 1314), as indicated in block 1572. Alternatively, the managed node 1260 may read the data block from the power loss protected buffer 1312 or DRAM 1304, as indicated in block 1574. Subsequently, in block 1576, the managed node 1260 returns the read data block in response to the read request. After the managed node 1260 has returned the data block in response to the read request, the method 1500 loops back to block 1502 of FIG. 15, in which the managed node 1260 determines whether to continue managing write acceleration.

EXAMPLES

[0068] Illustrative examples of the technologies disclosed herein are provided below. An embodiment of the technologies may include any one or more, and any combination of, the examples described below.

[0069] Example 1 includes a managed node to accelerate data writes, the managed node comprising a network interface controller that includes a power loss protected buffer; non-volatile memory; and a network communicator to receive, through the network interface controller, a write request from a remote device, wherein the write request includes a data block, write the data block to the power loss protected buffer of the network interface controller, and send, in response to receipt of the data block and prior to a write of the data block to the non-volatile memory, an acknowledgement to the remote device, wherein the acknowledgement is indicative of a successful write of the data block to the non-volatile memory; and a data manager to write, after the acknowledgement has been sent, the data block from the power loss protected buffer to the non-volatile memory.

[0070] Example 2 includes the subject matter of Example 1, and wherein the power loss protected memory comprises volatile memory coupled to a temporary power source.

[0071] Example 3 includes the subject matter of any of Examples 1 and 2, and wherein the network communicator is further to determine whether a power loss imminent condition is present; and wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that a power loss imminent condition is present, the data block to the non-volatile memory.

[0072] Example 4 includes the subject matter of any of Examples 1-3, and wherein the network communicator is further to determine whether a predefined time interval has elapsed; and wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that the predefined time interval has elapsed, the data block to the non-volatile memory.

[0073] Example 5 includes the subject matter of any of Examples 1-4, and wherein the network communicator is further to determine whether the number of data blocks present in the power loss protected buffer satisfies a predefined threshold; and wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that the number of data blocks present in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

[0074] Example 6 includes the subject matter of any of Examples 1-5, and wherein the network communicator is further to determine whether an amount of unused memory in the power loss protected buffer satisfies a predefined threshold; and wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that the amount of unused memory in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

[0075] Example 7 includes the subject matter of any of Examples 1-6, and wherein the network communicator is further to forward the data block to a follower storage sled for storage.

[0076] Example 8 includes the subject matter of any of Examples 1-7, and wherein to write the data block to the non-volatile memory comprises to write the data block in association with a key, wherein the key uniquely identifies the data block in the non-volatile storage.

[0077] Example 9 includes the subject matter of any of Examples 1-8, and wherein to receive the data block comprises to receive the data block with an associated key in the power loss protected buffer; and to write the data block to the non-volatile memory further comprises to store a location of the data block in a data map in association with the key

[0078] Example 10 includes the subject matter of any of Examples 1-9, and wherein the data manager is further to determine whether a request to read the data block has been received; determine, in response to the request to read the data block, a location of the data block; and read the data block from the determined location.

[0079] Example 11 includes the subject matter of any of Examples 1-10, and wherein to receive the request to read the data block comprises to receive a request that includes a key associated with the data block; to determine the location of the data block comprises to search a

data map that is indicative of keys and corresponding locations in the non-volatile memory; and to read the data block comprises to read the data block from the non-volatile memory.

[0080] Example 12 includes the subject matter of any of Examples 1-11, and wherein to determine the location of the data block comprises to determine whether the data block is present in the power loss protected buffer; and read, in response to a determination that the data block is present in the power loss protected buffer, the data block from the power loss protected buffer.

[0081] Example 13 includes the subject matter of any of Examples 1-12, and wherein the data manager is further to shard and encrypt the data block for storage on multiple storage sleds.

[0082] Example 14 includes the subject matter of any of Examples 1-13, and wherein the network communicator is further to receive an indication from one or more follower storage sleds that the data block has been written to non-volatile memory; and send, in response to receipt of the one or more indications, a notification to the follower storage sleds to free the data block from a power loss protected buffer of each follower storage sled.

[0083] Example 15 includes a method for accelerating data writes, the method comprising receiving, by a managed node through a network interface controller of the managed node, a write request from a remote device, wherein the write request includes a data block; writing, by the managed node, the data block to a power loss protected buffer of the network interface controller; sending, by the managed node, in response to receipt of the data block and prior to a write of the data block to a non-volatile memory of the managed node, an acknowledgement to the remote device, wherein the acknowledgement is indicative of a successful write of the data block to the non-volatile memory; and writing, by the managed node after the acknowledgement has been sent, the data block from the power loss protected buffer to the non-volatile memory.

[0084] Example 16 includes the subject matter of Example 15, and wherein writing the data block to a power loss protected buffer comprises writing the data block to volatile memory of the network interface controller, wherein the volatile memory is coupled to a temporary power source.

[0085] Example 17 includes the subject matter of any of Examples 15 and 16, and further including determining, by the managed node, whether a power loss imminent condition is present; and wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that a power loss imminent condition is present, the data block to the non-volatile memory.

[0086] Example 18 includes the subject matter of any of Examples 15-17, and further including determining, by the managed node, whether a predefined time interval has elapsed; and wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that the predefined time interval has elapsed, the data block to the non-volatile memory.

[0087] Example 19 includes the subject matter of any of Examples 15-18, and further including determining, by the managed node, whether the number of data blocks present in the power loss protected buffer satisfies a predefined threshold; and wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that the number of data blocks present in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

[0088] Example 20 includes the subject matter of any of Examples 15-19, and further including determining, by the managed node, whether an amount of unused memory in the power loss protected buffer satisfies a predefined threshold; and wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that the amount of unused memory in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

[0089] Example 21 includes the subject matter of any of Examples 15-20, and further including forwarding the data block to one or more follower storage sleds for storage.

[0090] Example 22 includes the subject matter of any of Examples 15-21, and wherein writing the data block to the non-volatile memory comprises writing the data block in association with a key, wherein the key uniquely identifies the data block in the non-volatile storage.

[0091] Example 23 includes the subject matter of any of Examples 15-22, and wherein receiving the data block comprises receiving the data block with an associated key in the power loss protected buffer; and writing the data block to the non-volatile memory further comprises storing a location of the data block in a data map in association with the key.

[0092] Example 24 includes the subject matter of any of Examples 15-23, and further including determining, by the managed node, whether a request to read the data block has been received; determining, by the managed node and in response to the request to read the data block, a location of the data block; and reading, by the managed node, the data block from the determined location.

[0093] Example 25 includes the subject matter of any of Examples 15-24, and wherein receiving the request to read the data block comprises receiving a request that includes a key associated with the data block; determining the location of the data block comprises searching a

data map that is indicative of keys and corresponding locations in the non-volatile memory; and reading the data block comprises reading the data block from the non-volatile memory.

[0094] Example 26 includes the subject matter of any of Examples 15-25, and wherein determining the location of the data block comprises determining whether the data block is present in the power loss protected buffer; and reading, in response to a determination that the data block is present in the power loss protected buffer, the data block from the power loss protected buffer.

[0095] Example 27 includes the subject matter of any of Examples 15-26, and wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing the data block to flash memory.

[0096] Example 28 includes the subject matter of any of Examples 15-27, and wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing the data block to one or more solid state drives.

[0097] Example 29 includes one or more computer-readable storage media comprising a plurality of instructions that, when executed by a managed node, cause the managed node to perform the method of any of Examples 15-28.

[0098] Example 30 includes a managed node comprising means for receiving, through a network interface controller of the managed node, a write request from a remote device, wherein the write request includes a data block; means for writing the data block to a power loss protected buffer of the network interface controller; means for sending, in response to receipt of the data block and prior to a write of the data block to a non-volatile memory of the managed node, an acknowledgement to the remote device, wherein the acknowledgement is indicative of a successful write of the data block to the non-volatile memory; and means for writing, after the acknowledgement has been sent, the data block from the power loss protected buffer to the non-volatile memory.

[0099] Example 31 includes the subject matter of Example 30, and wherein the means for writing the data block to a power loss protected buffer comprises means for writing the data block to volatile memory of the network interface controller, wherein the volatile memory is coupled to a temporary power source.

[00100] Example 32 includes the subject matter of any of Examples 30 and 31, and further including means for determining whether a power loss imminent condition is present; and wherein the means for writing the data block from the power loss protected buffer to the non-volatile memory comprises means for writing, in response to a determination that a power loss imminent condition is present, the data block to the non-volatile memory.

[00101] Example 33 includes the subject matter of any of Examples 30-32, and further including means for determining whether a predefined time interval has elapsed; and wherein the means for writing the data block from the power loss protected buffer to the non-volatile memory comprises means for writing, in response to a determination that the predefined time interval has elapsed, the data block to the non-volatile memory.

[00102] Example 34 includes the subject matter of any of Examples 30-33, and further including means for determining whether the number of data blocks present in the power loss protected buffer satisfies a predefined threshold; and wherein the means for writing the data block from the power loss protected buffer to the non-volatile memory comprises means for writing, in response to a determination that the number of data blocks present in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

[00103] Example 35 includes the subject matter of any of Examples 30-34, and further including means for determining whether an amount of unused memory in the power loss protected buffer satisfies a predefined threshold; and wherein the means for writing the data block from the power loss protected buffer to the non-volatile memory comprises means for writing, in response to a determination that the amount of unused memory in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

[00104] Example 36 includes the subject matter of any of Examples 30-35, and further including means for forwarding the data block to one or more follower storage sleds for storage.

[00105] Example 37 includes the subject matter of any of Examples 30-36, and wherein the means for writing the data block to the non-volatile memory comprises means for writing the data block in association with a key, wherein the key uniquely identifies the data block in the non-volatile storage.

[00106] Example 38 includes the subject matter of any of Examples 30-37, and wherein the means for receiving the data block comprises means for receiving the data block with an associated key in the power loss protected buffer; and the means for writing the data block to the non-volatile memory further comprises means for storing a location of the data block in a data map in association with the key.

[00107] Example 39 includes the subject matter of any of Examples 30-38, and further including means for determining whether a request to read the data block has been received; means for determining, in response to the request to read the data block, a location of the data block; and means for reading the data block from the determined location.

[00108] Example 40 includes the subject matter of any of Examples 30-39, and wherein the means for receiving the request to read the data block comprises means for receiving a request that includes a key associated with the data block; the means for determining the

location of the data block comprises means for searching a data map that is indicative of keys and corresponding locations in the non-volatile memory; and the means for reading the data block comprises means for reading the data block from the non-volatile memory.

[00109] Example 41 includes the subject matter of any of Examples 30-40, and wherein the means for determining the location of the data block comprises means for determining whether the data block is present in the power loss protected buffer; and means for reading, in response to a determination that the data block is present in the power loss protected buffer, the data block from the power loss protected buffer.

[00110] Example 42 includes the subject matter of any of Examples 30-41, and wherein the means for writing the data block from the power loss protected buffer to the non-volatile memory comprises means for writing the data block to flash memory.

[00111] Example 43 includes the subject matter of any of Examples 30-42, and wherein the means for writing the data block from the power loss protected buffer to the non-volatile memory comprises means for writing the data block to one or more solid state drives.

WHAT IS CLAIMED IS:

1. A managed node to accelerate data writes, the managed node comprising:
a network interface controller that includes a power loss protected buffer;
non-volatile memory; and
a network communicator to receive, through the network interface controller, a write request from a remote device, wherein the write request includes a data block, write the data block to the power loss protected buffer of the network interface controller, and send, in response to receipt of the data block and prior to a write of the data block to the non-volatile memory, an acknowledgement to the remote device, wherein the acknowledgement is indicative of a successful write of the data block to the non-volatile memory; and
a data manager to write, after the acknowledgement has been sent, the data block from the power loss protected buffer to the non-volatile memory.
2. The managed node of claim 1, wherein the power loss protected memory comprises volatile memory coupled to a temporary power source.
3. The managed node of claim 1, wherein the network communicator is further to determine whether a power loss imminent condition is present; and
wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that a power loss imminent condition is present, the data block to the non-volatile memory.
4. The managed node of claim 1, wherein the network communicator is further to determine whether a predefined time interval has elapsed; and
wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that the predefined time interval has elapsed, the data block to the non-volatile memory.
5. The managed node of claim 1, wherein the network communicator is further to determine whether the number of data blocks present in the power loss protected buffer satisfies a predefined threshold; and
wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that the number of data blocks present in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

6. The managed node of claim 1, wherein the network communicator is further to determine whether an amount of unused memory in the power loss protected buffer satisfies a predefined threshold; and

wherein to write the data block from the power loss protected buffer to the non-volatile memory comprises to write, in response to a determination that the amount of unused memory in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

7. The managed node of claim 1, wherein the network communicator is further to forward the data block to a follower storage sled for storage.

8. The managed node of claim 1, wherein to write the data block to the non-volatile memory comprises to write the data block in association with a key, wherein the key uniquely identifies the data block in the non-volatile storage.

9. The managed node of claim 1, wherein:
to receive the data block comprises to receive the data block with an associated key in the power loss protected buffer; and
to write the data block to the non-volatile memory further comprises to store a location of the data block in a data map in association with the key

10. The managed node of claim 1, wherein the data manager is further to:
determine whether a request to read the data block has been received;
determine, in response to the request to read the data block, a location of the data block; and
read the data block from the determined location.

11. The managed node of claim 10, wherein:
to receive the request to read the data block comprises to receive a request that includes a key associated with the data block;
to determine the location of the data block comprises to search a data map that is indicative of keys and corresponding locations in the non-volatile memory; and
to read the data block comprises to read the data block from the non-volatile memory.

12. The managed node of claim 10, wherein to determine the location of the data block comprises to:

determine whether the data block is present in the power loss protected buffer;
and

read, in response to a determination that the data block is present in the power loss protected buffer, the data block from the power loss protected buffer.

13. A method for accelerating data writes, the method comprising:

receiving, by a managed node through a network interface controller of the managed node, a write request from a remote device, wherein the write request includes a data block;

writing, by the managed node, the data block to a power loss protected buffer of the network interface controller;

sending, by the managed node, in response to receipt of the data block and prior to a write of the data block to a non-volatile memory of the managed node, an acknowledgement to the remote device, wherein the acknowledgement is indicative of a successful write of the data block to the non-volatile memory; and

writing, by the managed node after the acknowledgement has been sent, the data block from the power loss protected buffer to the non-volatile memory.

14. The method of claim 13, wherein writing the data block to a power loss protected buffer comprises writing the data block to volatile memory of the network interface controller, wherein the volatile memory is coupled to a temporary power source.

15. The method of claim 13, further comprising:

determining, by the managed node, whether a power loss imminent condition is present; and

wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that a power loss imminent condition is present, the data block to the non-volatile memory.

16. The method of claim 13, further comprising determining, by the managed node, whether a predefined time interval has elapsed; and

wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that the predefined time interval has elapsed, the data block to the non-volatile memory.

17. The method of claim 13, further comprising determining, by the managed node, whether the number of data blocks present in the power loss protected buffer satisfies a predefined threshold; and

wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that the number of data blocks present in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

18. The method of claim 13, further comprising determining, by the managed node, whether an amount of unused memory in the power loss protected buffer satisfies a predefined threshold; and

wherein writing the data block from the power loss protected buffer to the non-volatile memory comprises writing, in response to a determination that the amount of unused memory in the power loss protected buffer satisfies the predefined threshold, the data block to the non-volatile memory.

19. The method of claim 13, further comprising forwarding the data block to one or more follower storage sleds for storage.

20. The method of claim 13, wherein writing the data block to the non-volatile memory comprises writing the data block in association with a key, wherein the key uniquely identifies the data block in the non-volatile storage.

21. The method of claim 13, wherein:
receiving the data block comprises receiving the data block with an associated key in the power loss protected buffer; and
writing the data block to the non-volatile memory further comprises storing a location of the data block in a data map in association with the key.

22. The method of claim 13, further comprising:
determining, by the managed node, whether a request to read the data block has been received;
determining, by the managed node and in response to the request to read the data block, a location of the data block; and
reading, by the managed node, the data block from the determined location.
23. The method of claim 22, wherein:
receiving the request to read the data block comprises receiving a request that includes a key associated with the data block;
determining the location of the data block comprises searching a data map that is indicative of keys and corresponding locations in the non-volatile memory; and
reading the data block comprises reading the data block from the non-volatile memory.
24. One or more computer-readable storage media comprising a plurality of instructions that, when executed by a managed node, cause the managed node to perform the method of any of claims 13-23.
25. A managed node comprising means for performing the method of any of claims 13-23.

Data Center 100

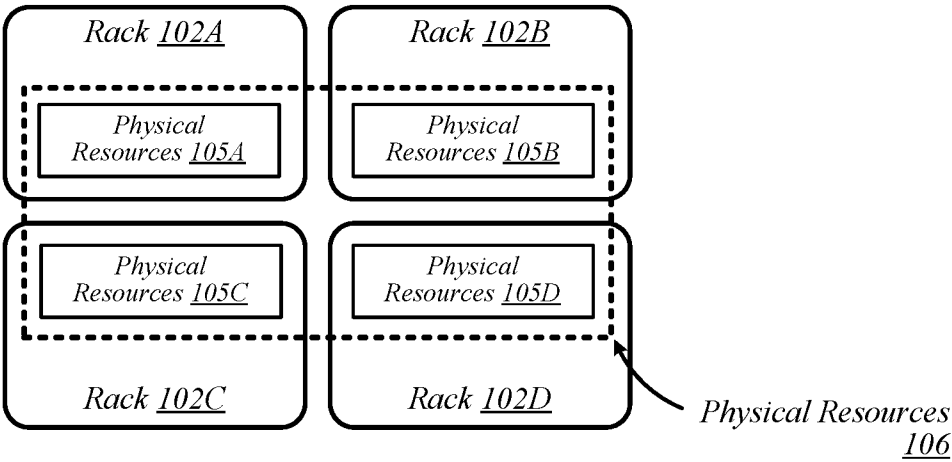


FIG. 1

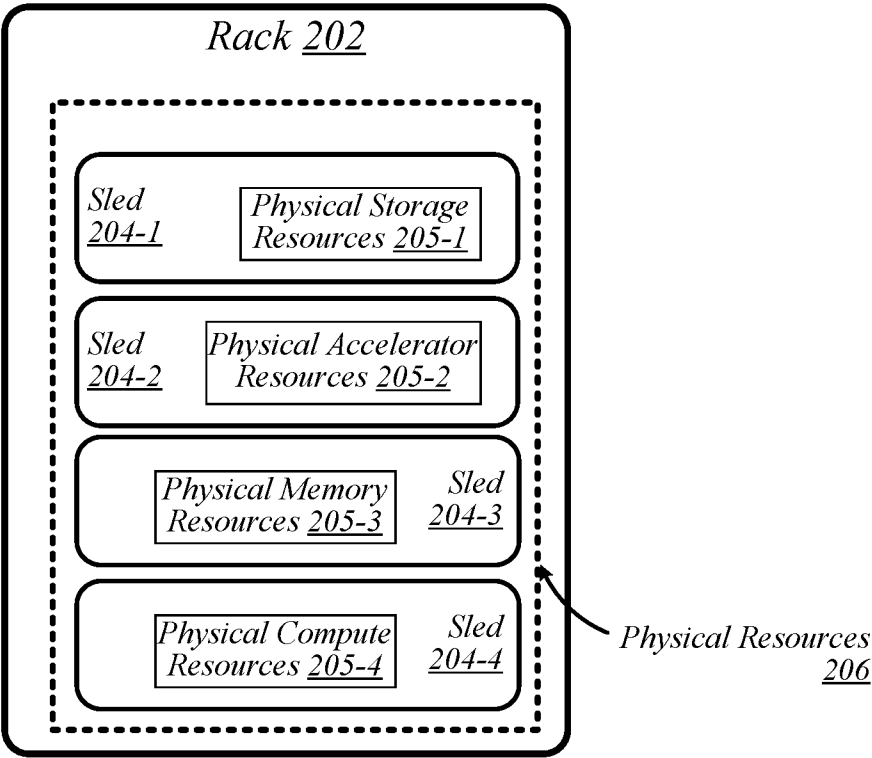


FIG. 2

Data Center 300

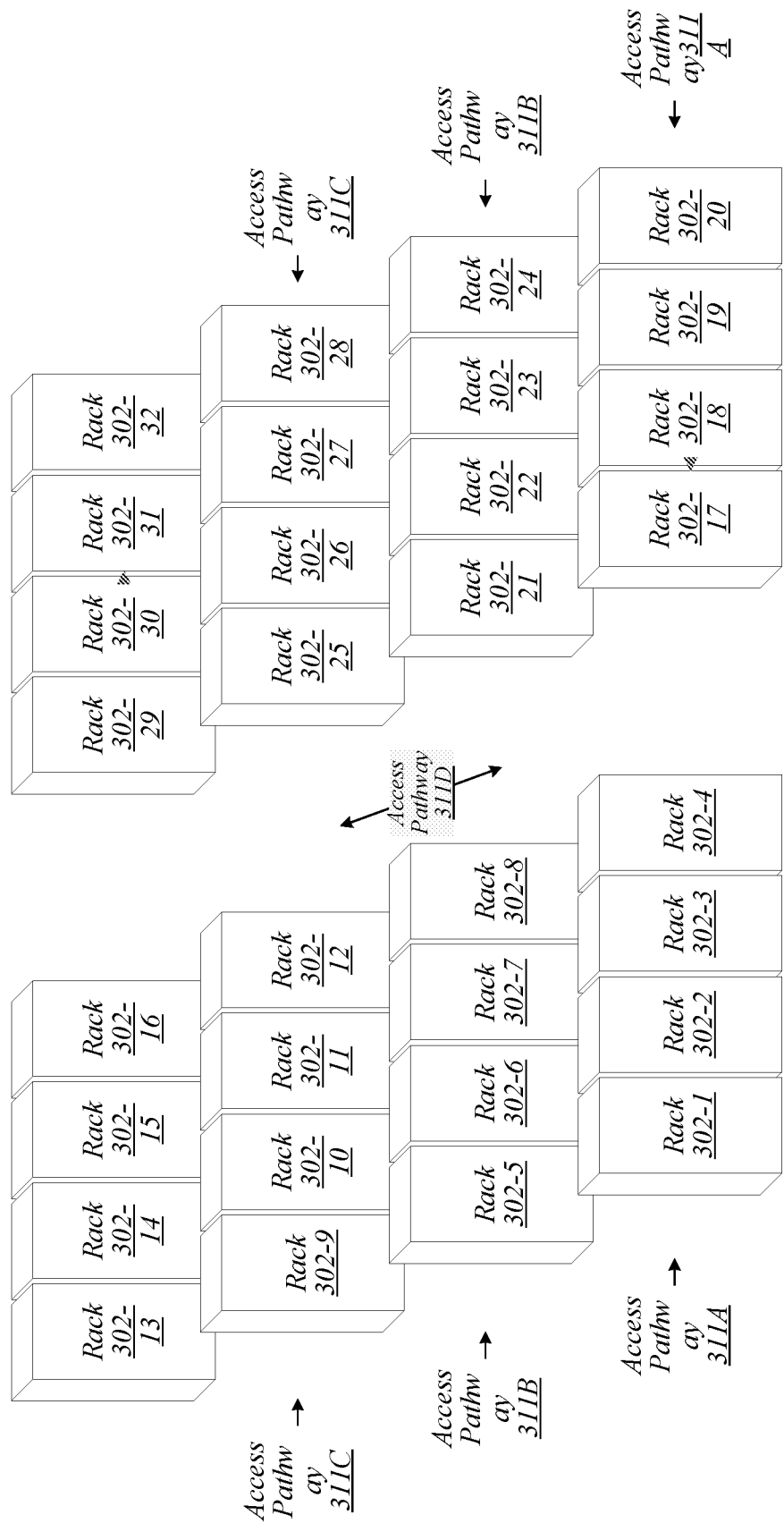


FIG. 3

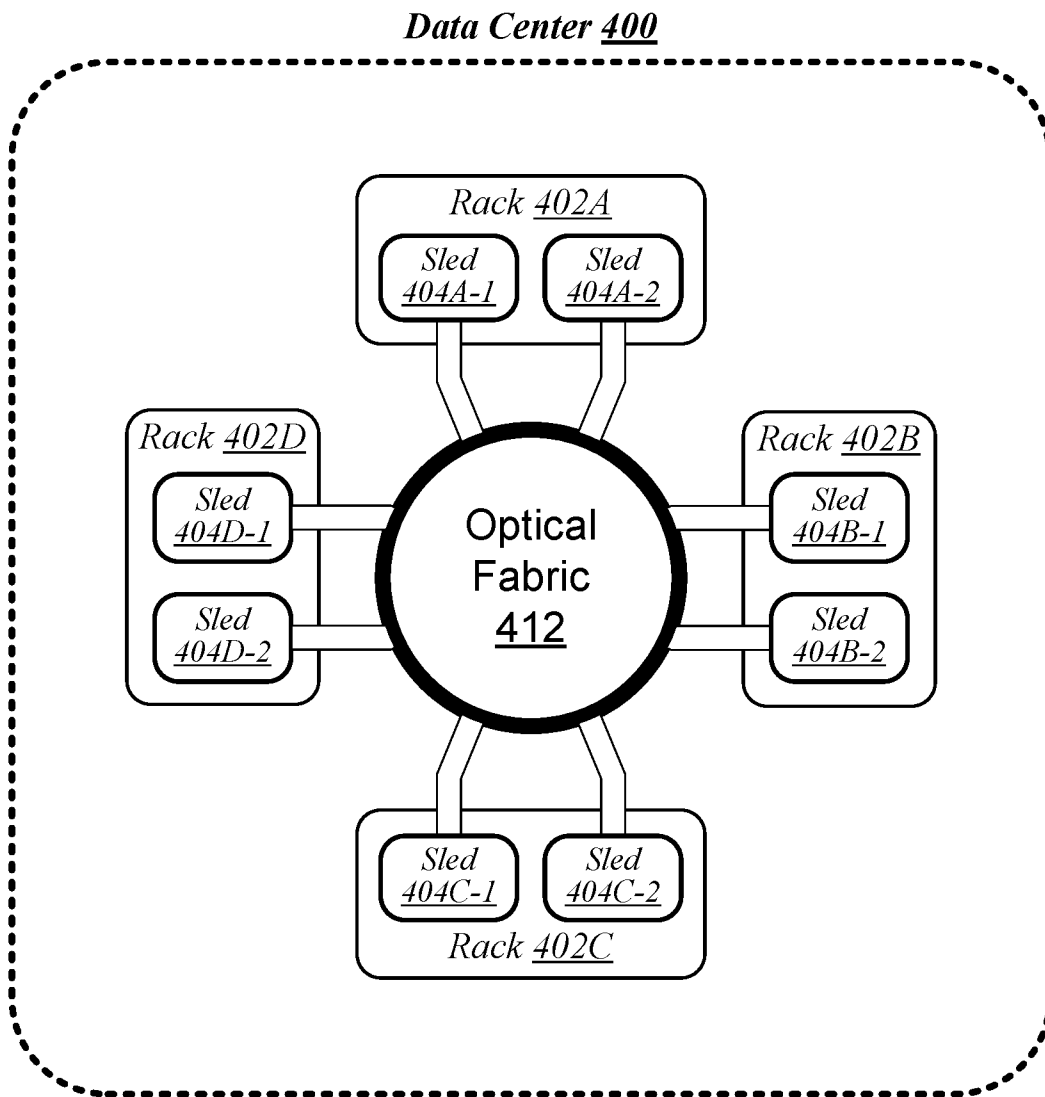


FIG. 4

500

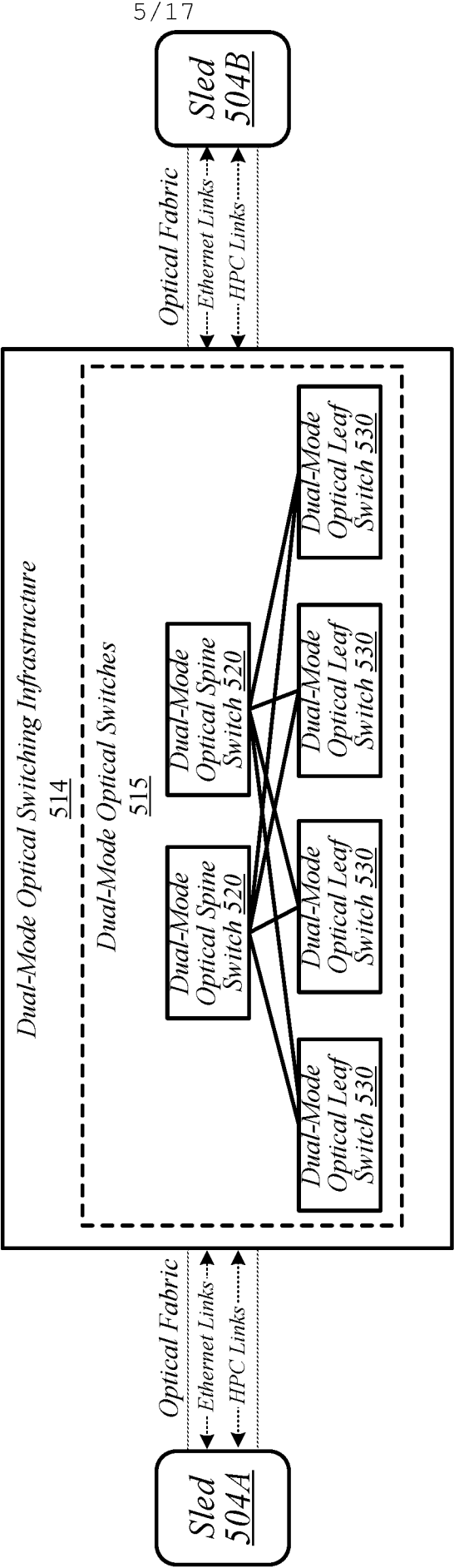


FIG. 5

600

Rack Access Region 601

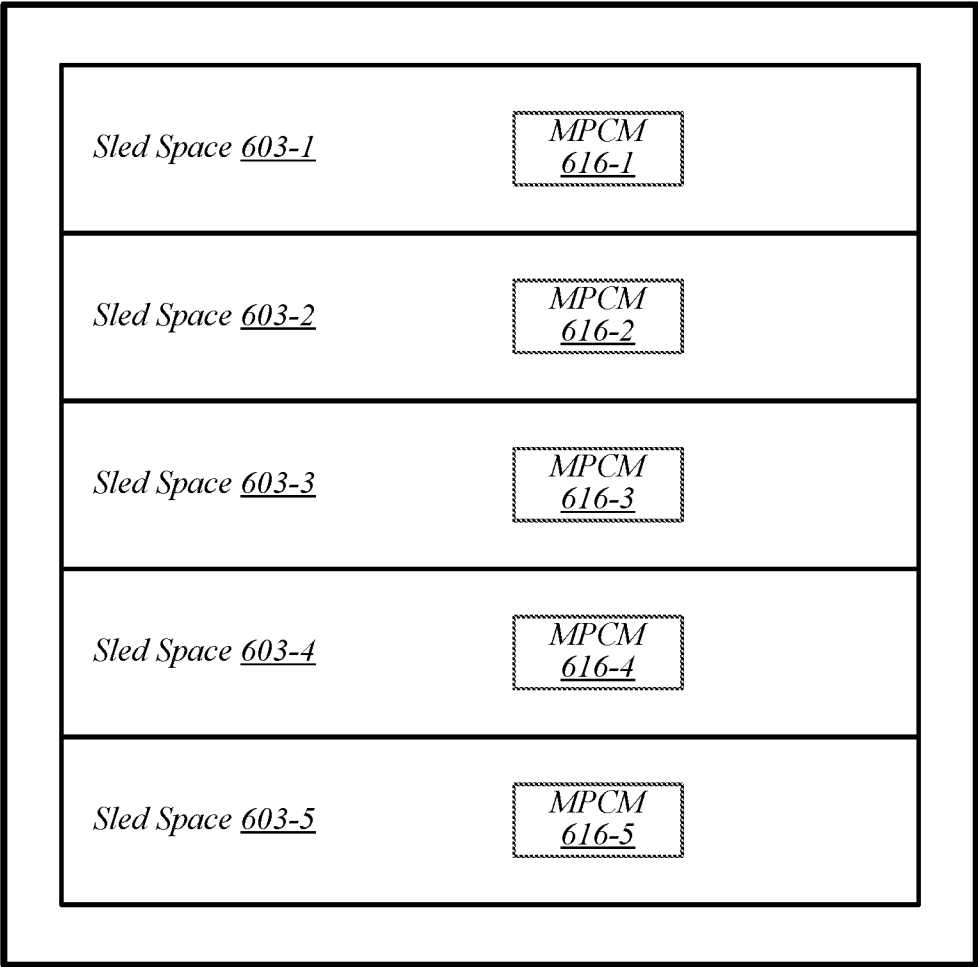


FIG. 6

Sled 704

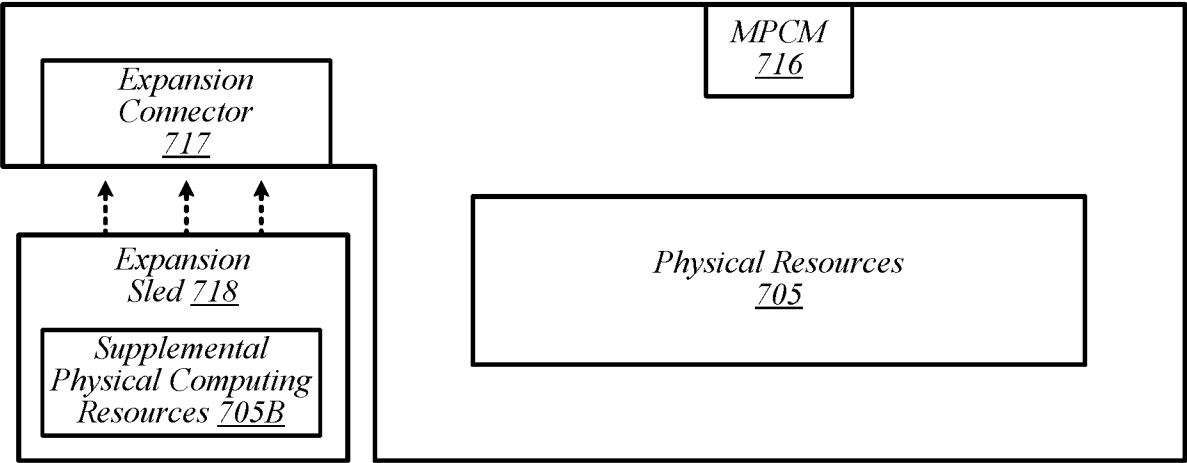
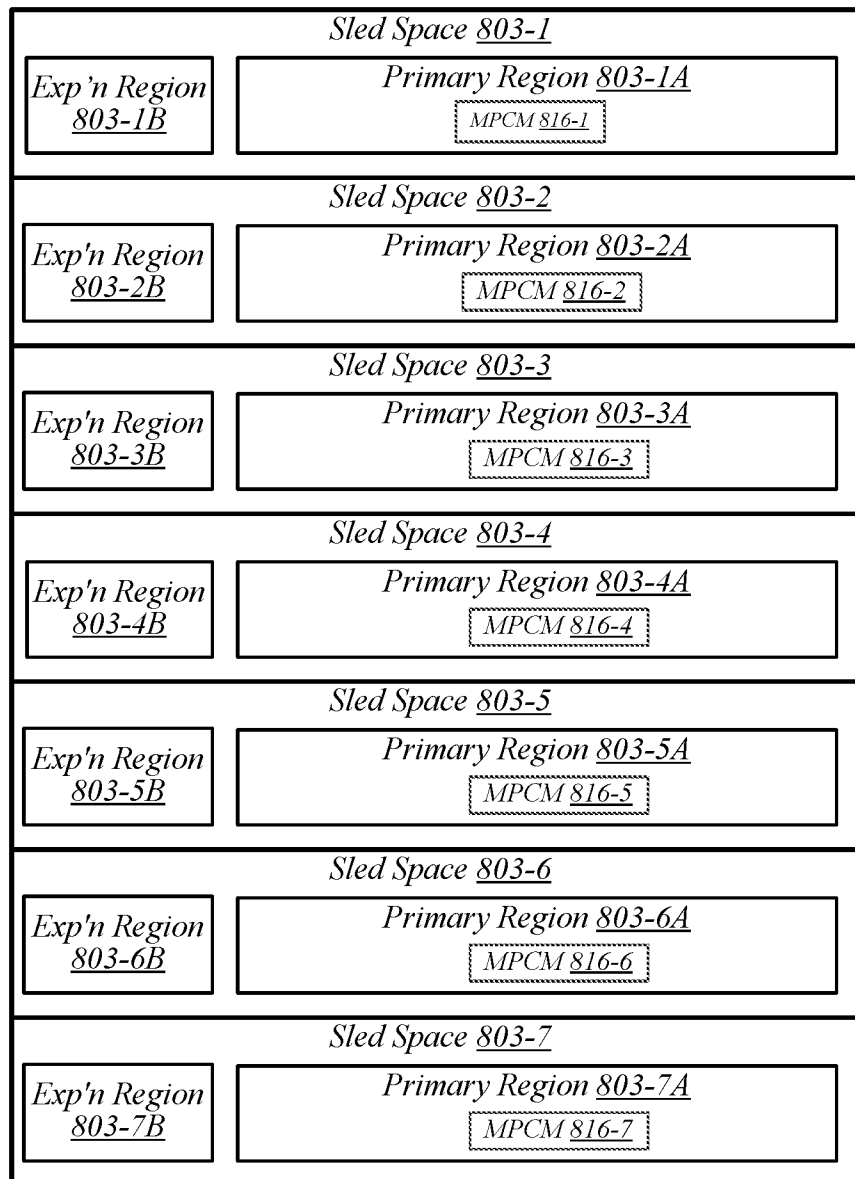


FIG. 7

800**FIG. 8**

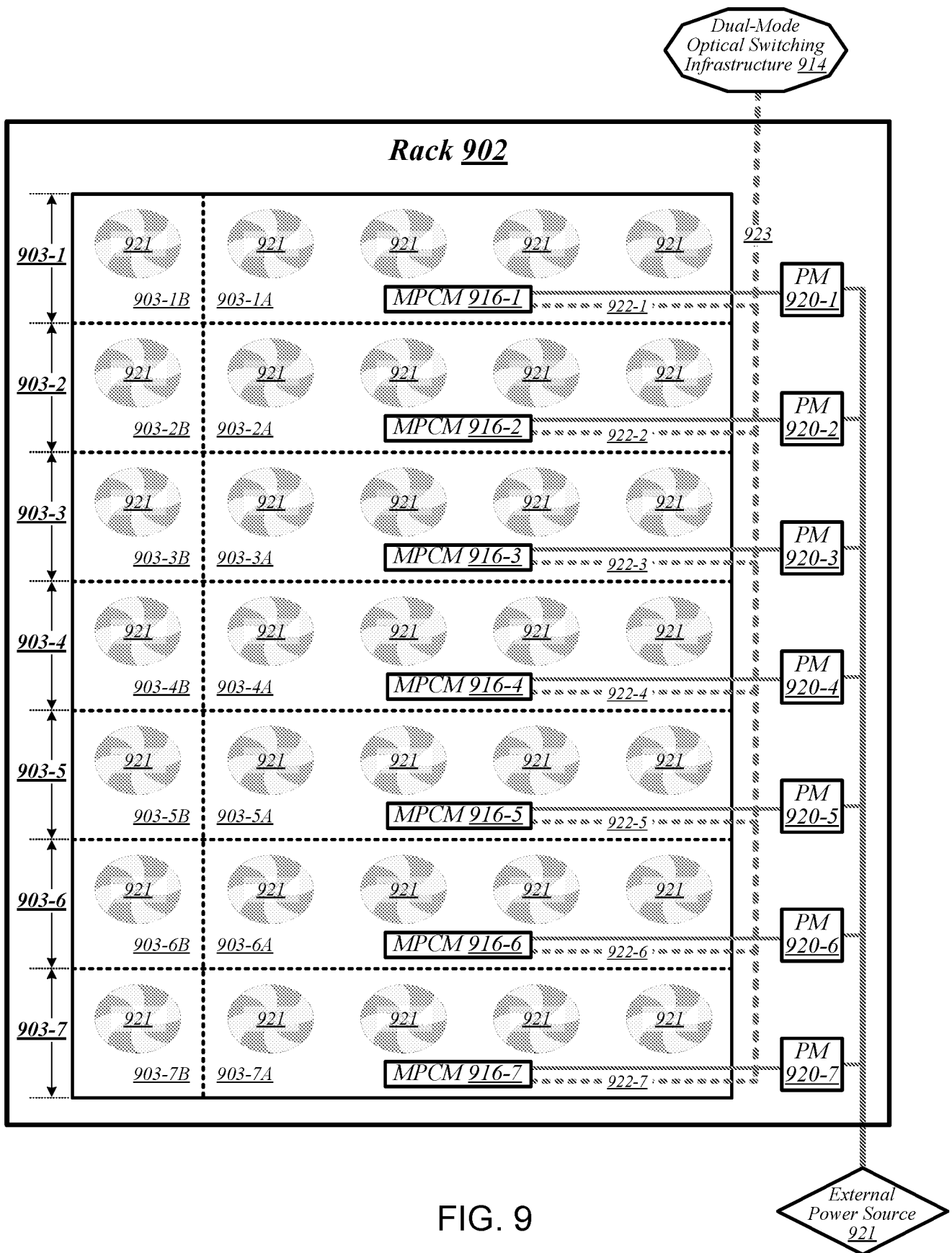


FIG. 9

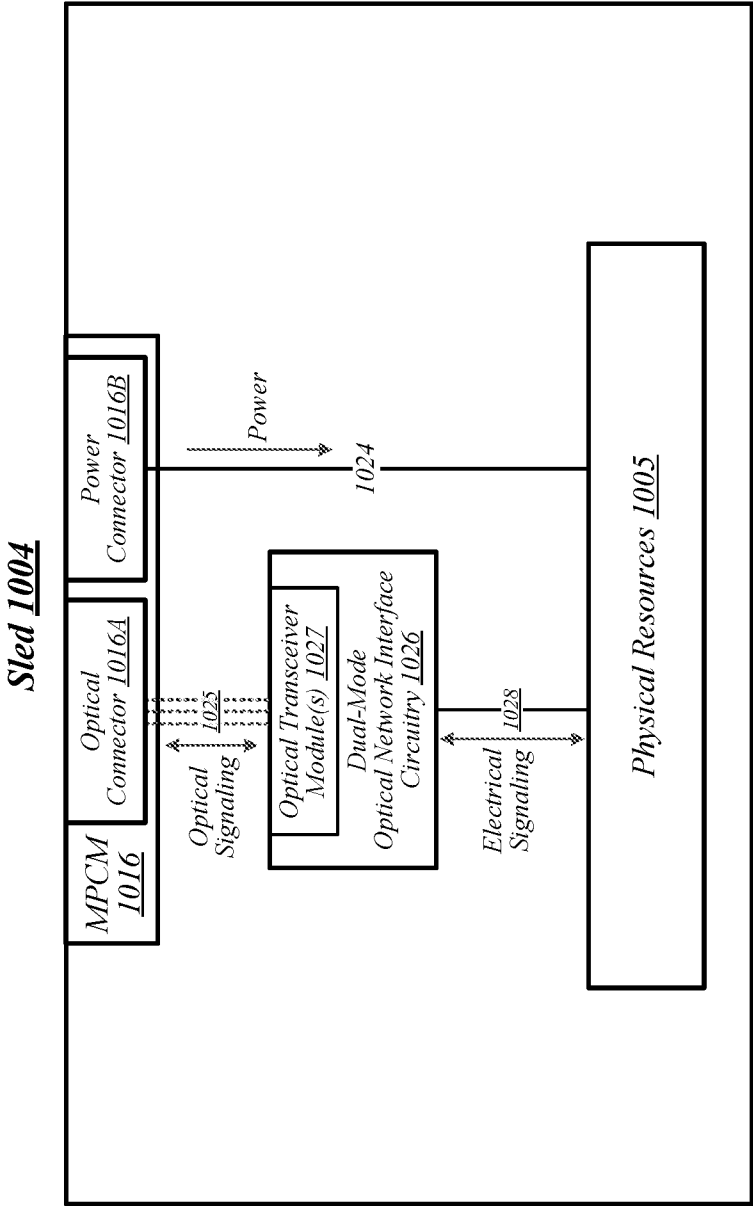


FIG. 10

Data Center 1100

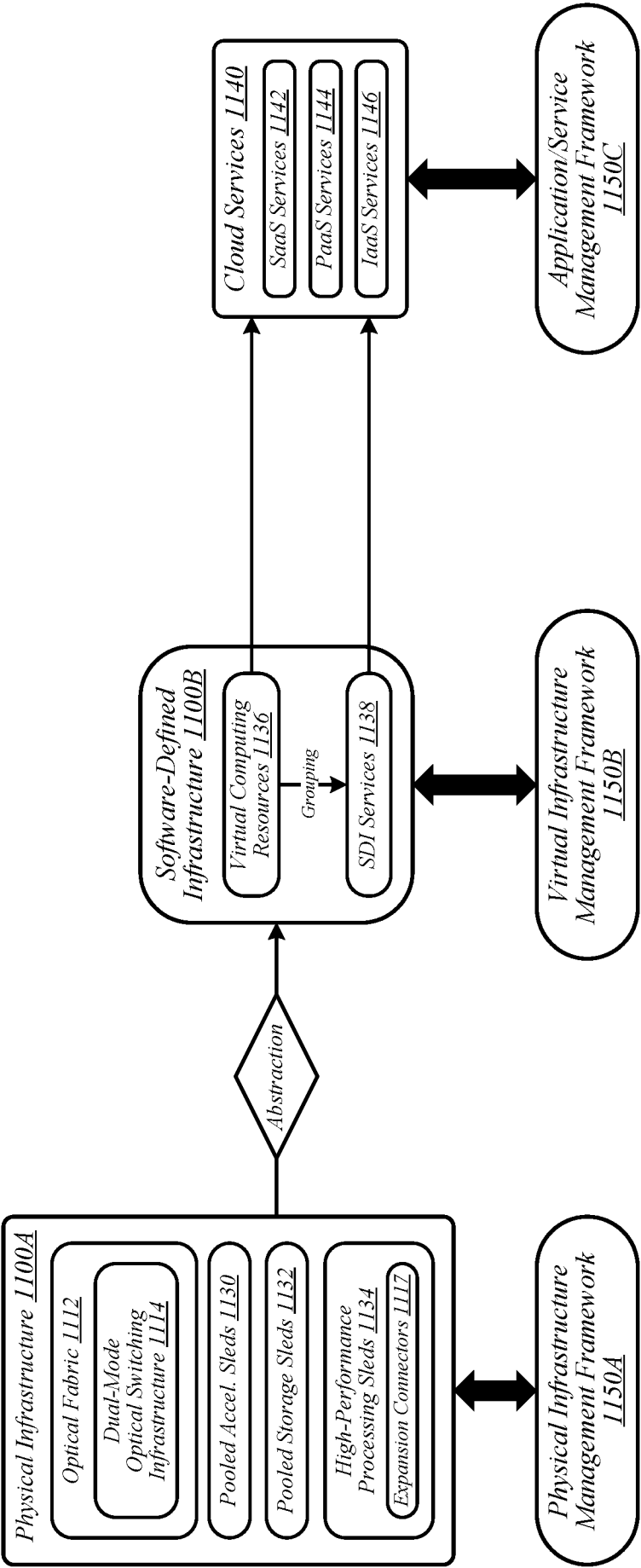


FIG. 11

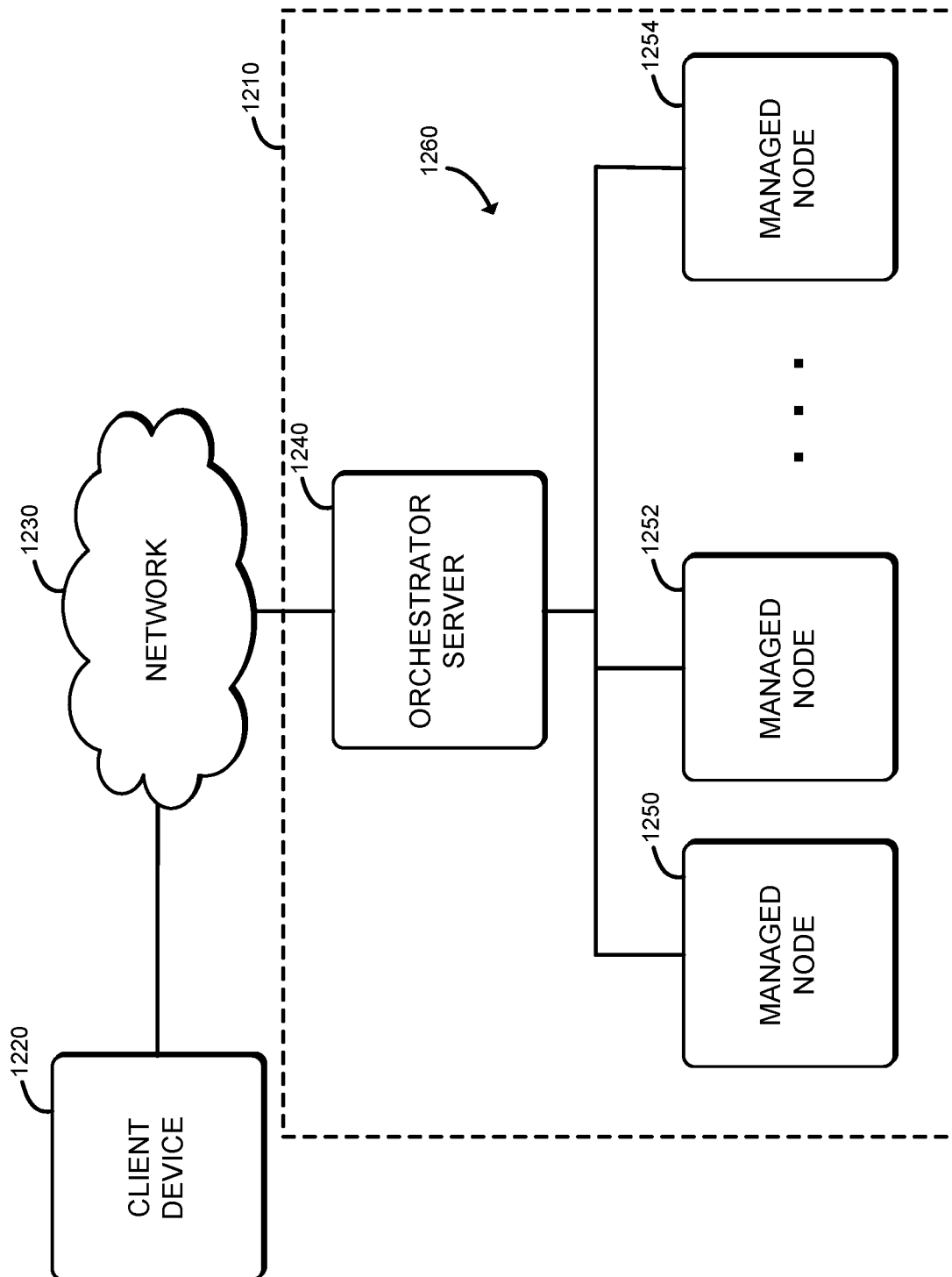


FIG. 12

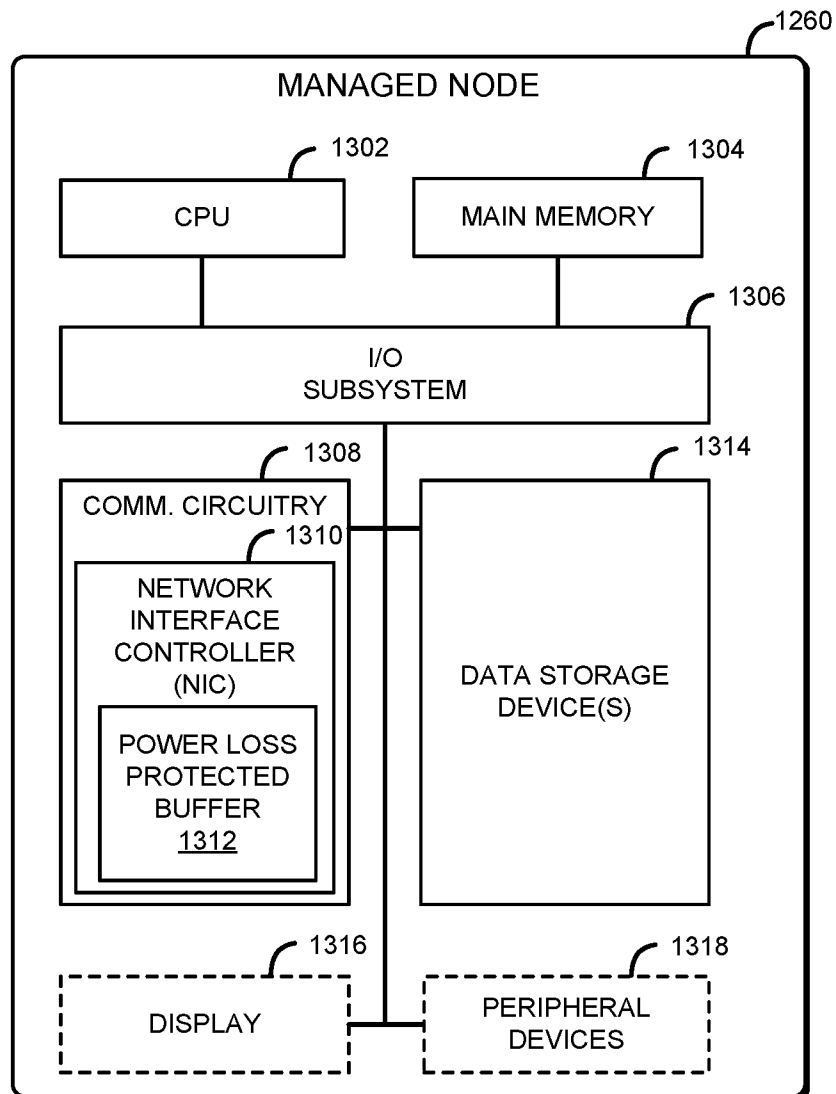


FIG. 13

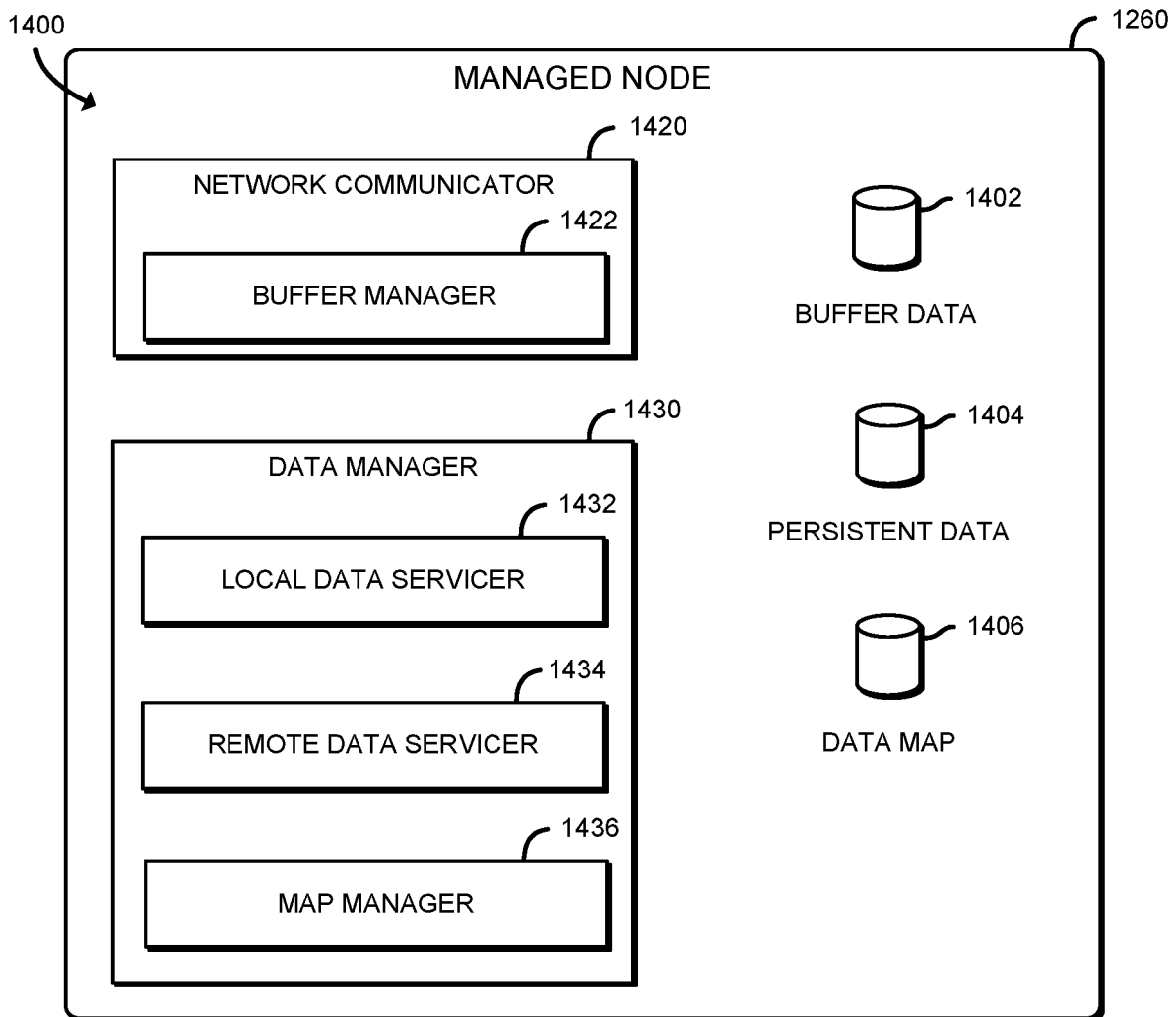


FIG. 14

15/17

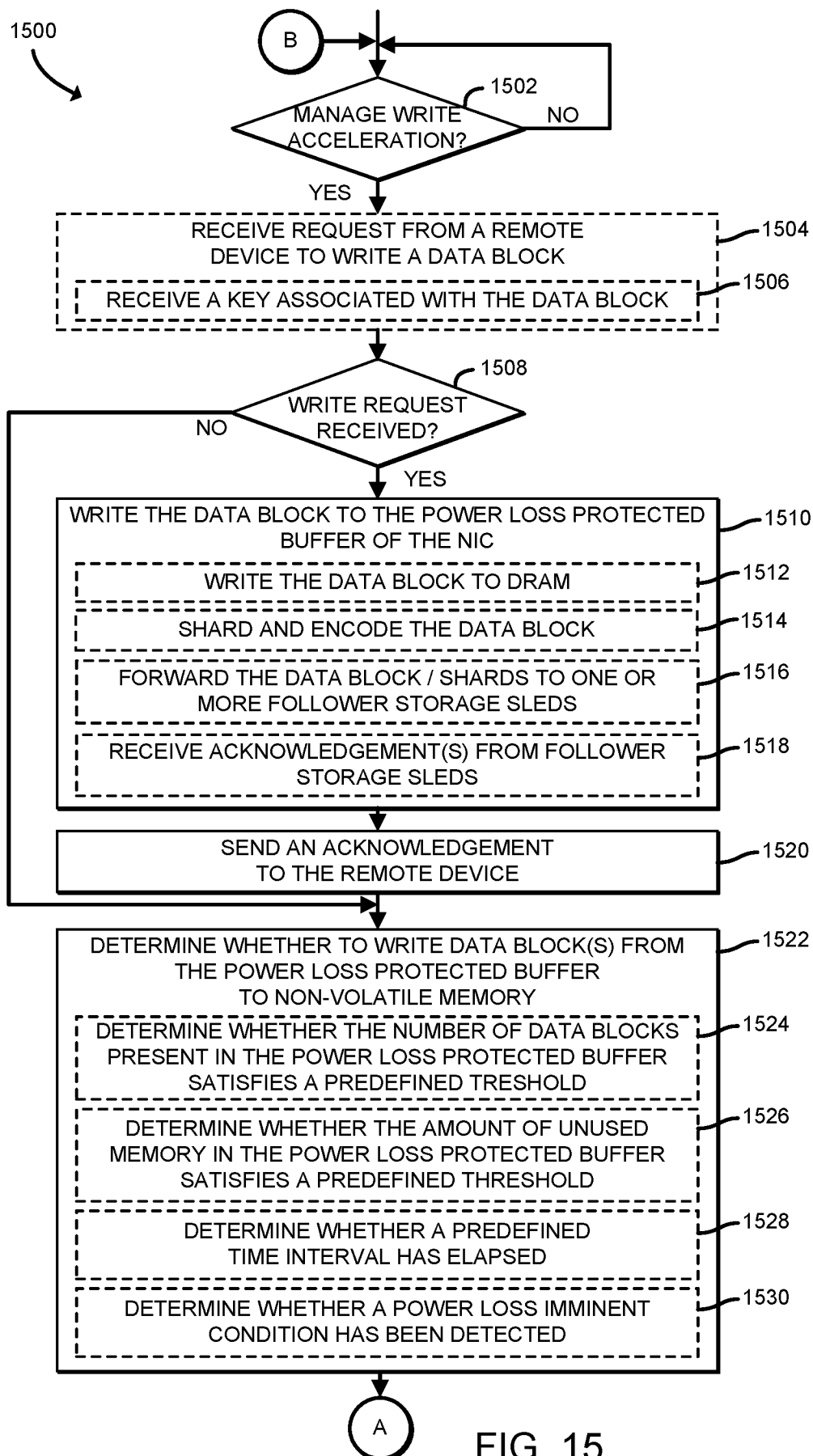


FIG. 15

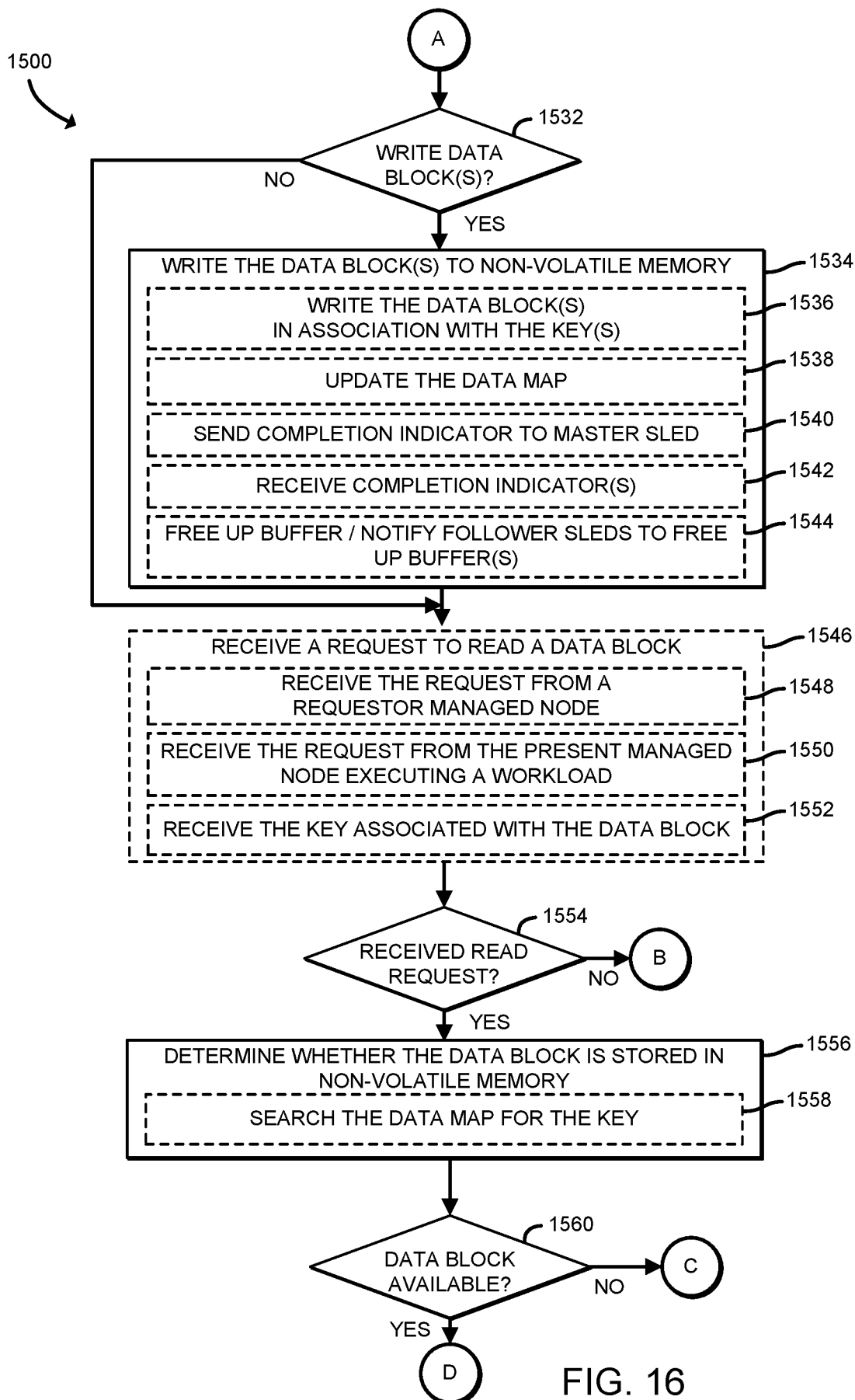


FIG. 16

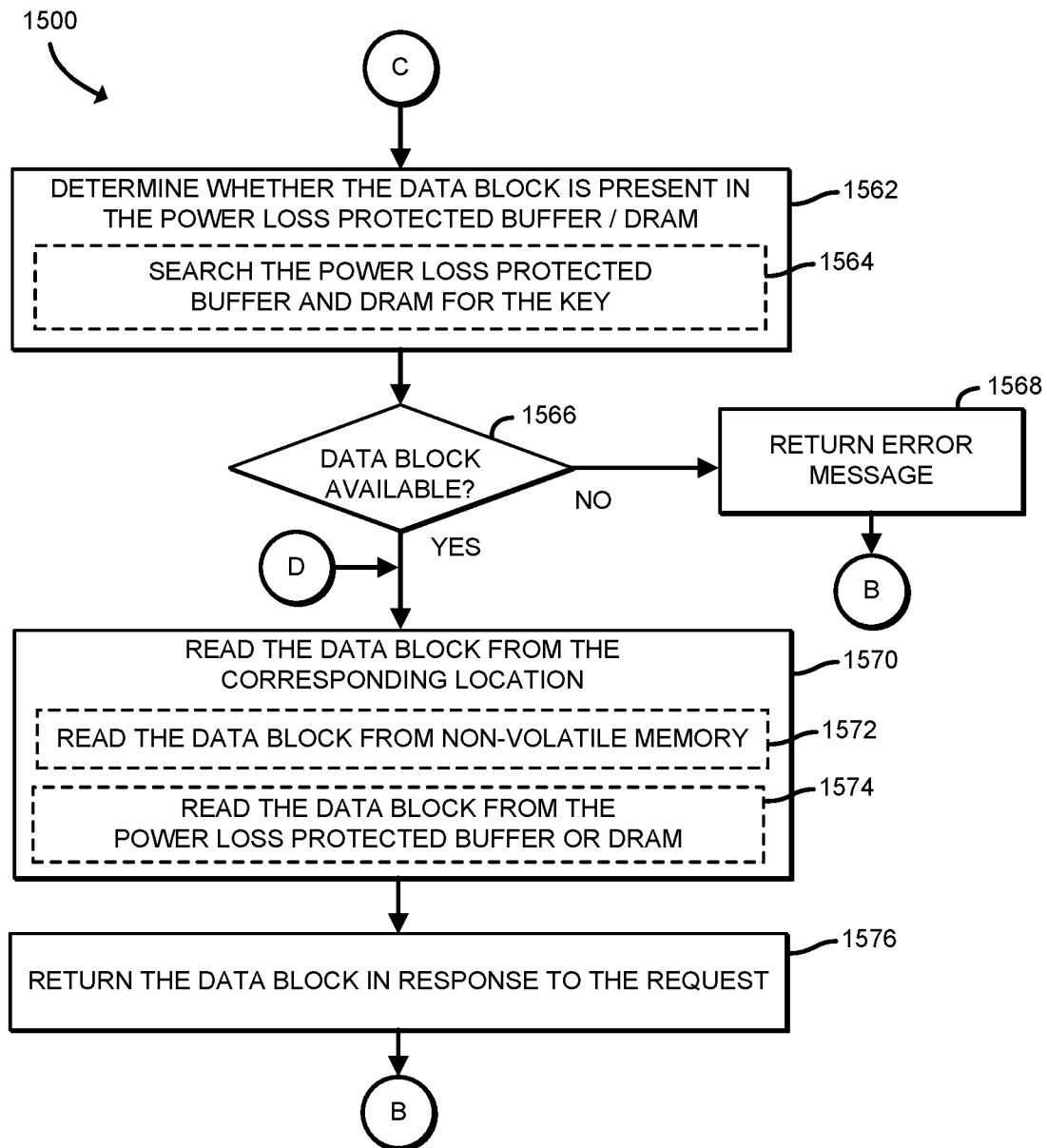


FIG. 17

A. CLASSIFICATION OF SUBJECT MATTER**H04L 12/879(2013.01)i, H04L 12/861(2013.01)i, H04Q 11/00(2006.01)i**

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)

H04L 12/879; G06F 3/06; G11C 11/56; G06F 17/30; G11C 5/14; G11C 16/34; G06F 12/00; G11C 16/06; H04L 12/861; H04Q 11/00

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

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

eKOMPASS(KIPO internal) & Keywords: data write, date read, managed node, protect power loss, non-volatile memory, buffer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015-0155050 A1 (SEAGATE TECHNOLOGY LLC) 04 June 2015 See paragraphs [0003], [0012], [0015]–[0016], [0025], [0043], [0056], [0068]; and claim 1.	1, 3–7, 13, 15–19 , 24–25
Y		8–12, 20–23
A		2, 14
Y	US 8082236 B2 (ARSENIY AHARONOV et al.) 20 December 2011 See column 4, line 51; and claim 8.	8–12, 20–23
A	US 2016-0139832 A1 (KABUSHIKI KAISHA TOSHIBA) 19 May 2016 See paragraphs [0036]–[0054]; and figure 2.	1–25
A	US 9361044 B2 (LYNDON S. CHIU et al.) 07 June 2016 See paragraphs [0011]–[0013]; and figure 1.	1–25
A	US 2011-0096601 A1 (LEE M. GAVENS et al.) 28 April 2011 See paragraphs [0072]–[0075]; and figure 1.	1–25



Further documents are listed in the continuation of Box C.



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

18 September 2017 (18.09.2017)

Date of mailing of the international search report

20 September 2017 (20.09.2017)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

Lee, Eun Kyu

Telephone No. +82-42-481-3580



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/038297

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015-0155050 A1	04/06/2015	None	
US 8082236 B2	20/12/2011	IL 187042 A IL 187042 D0 US 2009-0112823 A1 US 2012-0185442 A1 US 8606764 B2 WO 2009-057099 A1	09/02/2008 09/02/2008 30/04/2009 19/07/2012 10/12/2013 07/05/2009
US 2016-0139832 A1	19/05/2016	JP 2013-061799 A US 2013-0067146 A1 US 2014-0379972 A1 US 8862809 B2 US 9298240 B2	04/04/2013 14/03/2013 25/12/2014 14/10/2014 29/03/2016
US 9361044 B2	07/06/2016	CN 102722339 A CN 102722339 B US 2012-0254503 A1 US 2016-0357670 A1	10/10/2012 26/04/2017 04/10/2012 08/12/2016
US 2011-0096601 A1	28/04/2011	CN 102667943 A CN 102667943 B CN 102667944 A CN 102667944 B CN 102667945 A CN 102667945 B EP 2494552 A1 EP 2494553 A1 EP 2494553 B1 EP 2494554 A1 EP 2494554 B1 JP 05443612 B2 JP 05596791 B2 JP 2013-509651 A JP 2013-509667 A JP 2013-509668 A KR 10-1621584 B1 KR 10-1727349 B1 KR 10-2012-0087948 A KR 10-2012-0093295 A TW 201140604 A US 2011-0099418 A1 US 2011-0099460 A1 US 2012-0272120 A1 US 8214700 B2 US 8386861 B2 US 8423866 B2 US 8634240 B2 WO 2011-056478 A1	12/09/2012 18/05/2016 12/09/2012 13/01/2016 12/09/2012 13/01/2016 05/09/2012 05/09/2012 27/07/2016 05/09/2012 23/12/2015 19/03/2014 24/09/2014 14/03/2013 14/03/2013 14/03/2013 16/05/2016 14/04/2017 07/08/2012 22/08/2012 16/11/2011 28/04/2011 28/04/2011 25/10/2012 03/07/2012 26/02/2013 16/04/2013 21/01/2014 12/05/2011

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/038297Patent document
cited in search reportPublication
datePatent family
member(s)Publication
date

WO 2011-056482 A1

12/05/2011

WO 2011-056483 A1

12/05/2011