



(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/US2017/038301

(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
15/395,572 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, Liv-
ermore, 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,
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,

(54) Title: TECHNOLOGIES FOR DISTRIBUTING DATA TO IMPROVE DATA THROUGHPUT RATES

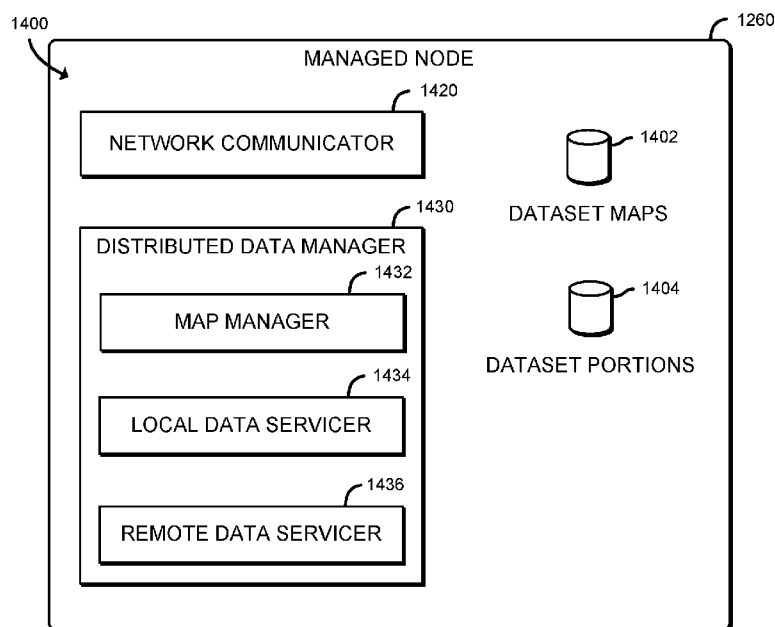


FIG. 14

(57) Abstract: Technologies for managing distributed data to improve data throughput rates include a managed node to distribute a dataset over multiple data storage devices coupled to a network. Each data storage device has a peak data throughput rate. The managed node is further to request a corresponding portion of the dataset from each data storage device, receive the requested portions of the dataset at a combined data throughput rate that is greater than the peak data throughput rate of any of the data storage devices, and combine the received portions of the dataset to reconstruct the dataset. Other embodiments are also described and claimed.



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 DISTRIBUTING DATA TO IMPROVE DATA THROUGHPUT RATES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims priority to U.S. Utility Patent Application Serial No. 15/395,572, entitled “TECHNOLOGIES FOR DISTRIBUTING DATA TO IMPROVE DATA THROUGHPUT RATES,” 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), data may be written to and retrieved from data storage devices as workloads (e.g., applications, processes, services, etc.) are executed on behalf of customers. The data storage devices typically have a peak data throughput rate at which they can write and/or retrieve data. As such, in a system in which the peak data throughput rate of a data storage device is less than the data throughput rate of the data communication bus that couples the data storage device to a compute device requesting the access to the data, the peak data throughput rate of the data storage device becomes a bottleneck and may reduce the performance of any workloads executed by the compute device. To address such bottlenecks, administrators of data centers may purchase more expensive data storage devices that provide greater data throughput rates. As a result, the cost of the data center increases.

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 managing the distribution of data among a set of managed nodes to improve data access throughput;

[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. 15-16 are a simplified flow diagram of at least one embodiment of a method for managing distributed data to increase data access throughput 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 managing the distribution of data among a set of managed nodes 1260 to improve data access throughput 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 the managed nodes 1250 may execute one or more applications or processes (i.e., workloads), such as in virtual machines or containers, on behalf of a user of the client device 1220.

[0047] As discussed in more detail herein, the managed nodes 1260 may write data to and read data from multiple data storage devices (e.g., physical storage resources 205-1 located in one or more of the managed nodes 1260). In doing so, the managed nodes 1260 may partition a dataset to be written into multiple portions and write each portion to a different data storage device (e.g., different SSDs). Each data storage device may have a data throughput rate that is less than the throughput rate of the communication bus (e.g., the optical fabric 412 described with reference to FIG. 4) connecting a physical compute resource 205-4 (e.g., a processor executing a workload) to the data storage devices (e.g., the physical storage resources 205-1). As such, by writing and/or reading different portions of the dataset with multiple data storage devices, a dataset may be written and read at a faster rate than would be possible using any one data storage device.

[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, partitioning a dataset into multiple portions and writing the portions to different data storage devices, reading multiple portions of the dataset from multiple data storage devices, combining the portions to reconstruct the dataset, and applying one or more error correction schemes to portions of the dataset to identify and correct errors. 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 1312. 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 portions of datasets, a map of the locations (e.g., data storage devices 1312 in various managed nodes 1260 and keys associated with the portions) where portions of datasets are stored, 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 one or more other 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 local processor (not shown) and/or a local memory (not shown) that are both 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 or alternatively, in such embodiments, the local memory of the NIC 1310 may be integrated into one or more components of the managed node 1260 at the board level, socket level, chip level, and/or other levels. As discussed above, the managed node 1260 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.

[0053] The one or more illustrative data storage devices 1312, 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 1312 may include a system partition that stores data and firmware code for the data storage device 1312. Each data storage device 1312 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 1312 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, in the illustrative embodiment, the non-volatile memory is embodied as Flash memory (e.g., NAND memory). In other embodiments, 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 1314. The display 1314 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 1314 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 1316. Such peripheral devices 1316 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. 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 distributed 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, distributed data manager circuitry 1430, etc.). It should be appreciated that, in such embodiments, one or more of the network communicator circuitry 1420 or the distributed 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 one or more dataset maps 1402 which may be embodied as any data indicative of the locations (e.g., data storage devices 1312 on one or more of the managed nodes 1260) where the portions of each dataset are stored. In the illustrative embodiment, the dataset maps 1402 additionally include a key associated with each portion, to be used in a request to access the associated value (e.g., the corresponding dataset portion 1404) of a key-value pair on a data storage device 1312. Additionally, the dataset maps 1402 include locations of redundant copies of portions of each dataset, to be requested if a data storage device 1312 or managed node 1260 on which the data storage device 1212 is physically located is inoperative (e.g., has lost network connectivity or has otherwise become unavailable to provide a portion of the dataset). Additionally, in the illustrative embodiment, the environment 1400 includes dataset portions 1404 which may be embodied as any data representing a subset of a dataset stored on behalf of a workload executed by the present managed node 1260 or another managed node 1260 in the set. As described above, the dataset portions 1404 may be associated with unique keys (e.g., alphanumeric codes, etc.) to be used to identify the portion 1404. Additionally, one or more of the dataset portions 1404 may be a redundant copy of another dataset portion 1404

stored on another data storage device 1312, and may be encoded using an error correction scheme (e.g., a low density parity check (LDPC) scheme, a Reed-Solomon scheme, etc.).

[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 orchestrator server 1240, respectively. 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.

[0060] The distributed 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 data access (e.g., writing data and/or reading data) to and from data storage devices 1312 local to the managed node 1260 or available in one or more other managed nodes 1260 to obtain a higher data throughput rate than would be available if the data was written to and read from a single data storage device 1312. To do so, in the illustrative embodiment, the distributed data manager 1430 includes a map manager 1432, a local data servicer 1434, and a remote data servicer 1436. The map manager 1432, in the illustrative embodiment, is configured to track where portions 1404 of datasets are stored among the data storage devices 1312 of the set of managed nodes 1260, partition datasets used by workloads executed by the present managed node 1260 into the portions 1404, including redundant portions for error correction schemes, associate unique keys (e.g., generated by the map manager 1432 based on a hash of the portion 1404 combined with an address such as a media access control address of the managed node 1260 to store the portion 1404 and a unique address (e.g., media access control address) of the present managed node 1260, and/or based on any other suitable method for uniquely identifying the portion 1404) with the portions 1404, track the availability of the data storage devices 1312 and the associated managed nodes 1260 to determine where to write and read dataset portions 1404, and recombine read dataset portions 1404 into the original datasets.

[0061] The local data servicer 1434, in the illustrative embodiment, is configured to write dataset portions 1404 in association with assigned keys (e.g., determined by the map

manager 1432) to one or more data storage devices 1312 local to the managed node 1260, read requested dataset portions 1404 (e.g., dataset portions 1404 identified by their corresponding keys) from the local data storage devices 1312, and apply any error correction algorithms in the processes of writing or reading the dataset portions 1404. The remote data servicer 1436, in the illustrative embodiment, is configured to issue requests to other managed nodes 1260 (e.g., managed nodes 1260 determined by the map manager 1432) to write dataset portions 1404 in association with keys provided by the map manager 1432 and issue requests to read dataset portions 1404 from the other managed nodes 1260 using keys provided by the map manager 1432.

[0062] It should be appreciated that each of the map manager 1432, the local data servicer 1434, and the remote data servicer 1436 may be separately embodied as hardware, firmware, software, virtualized hardware, emulated architecture, and/or a combination thereof and/or may be embodied as distributed services across multiple managed nodes 1260. For example, the map manager 1432 may be embodied as a hardware component, while the local data servicer 1434 and the remote data servicer 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.

[0063] Referring now to FIG. 15, in use, the managed node 1260 may execute a method 1500 for managing distributed data across multiple data storage devices 1312 to improve the data throughput rate for writing and/or reading data, as compared to writing to and/or reading from a single data storage device 1312. The method 1500 begins with block 1502, in which the managed node 1260 determines whether to manage distributed data. In the illustrative embodiment, the managed node 1260 determines to manage distributed data if the managed node 1260 is powered on and has access to (e.g., locally and/or through the fabric 412) multiple data storage devices 1312. In other embodiments, the managed node 1260 may determine whether to manage distributed data based on other factors. Regardless, in response to a determination to manage distributed data, in the illustrative embodiment, the method 1500 advances to block 1504 in which the managed node 1260 may receive a request from a workload executed by the managed node 1260 to write a dataset. In block 1506, the managed node 1260 determines whether a write request (e.g., a request to write a dataset) has been received. If the managed node 1260 has not received a write request, the method 1500 advances to block 1530 of FIG. 16, in which the managed node 1260 may receive a read request from a workload. Otherwise, if the managed node 1260 has received a write request, the

method 1500 advances to block 1508 in which the managed node 1260 distributes the dataset to be written across multiple data storage devices 1312.

[0064] In distributing the dataset, in the illustrative embodiment, the managed node 1260 partitions the dataset into multiple portions 1404 (e.g., subsets), as indicated in block 1510. For example, the managed node 1260 may divide the size of the dataset by a number of portions 1404 to be written, such that each portion 1404 is of equal size. In other embodiments, the managed node 1260 may partition the dataset into unequally sized portions 1404. As indicated in block 1512, the managed node 1260 may generate redundant portions 1404 using an error correction scheme. The redundant portions 1404 may be copies of other portions 1404 or may be complementary portions 1404 suitable for use in reconstructing a dataset when one or more of the portions 1404 cannot be recovered (e.g., the result of an XOR operation on one or more of the other portions 1404).

[0065] In block 1514, in the illustrative embodiment, the managed node 1260 determines an assignment of the portions 1404 to data storage devices 1312 in the present managed node 1260 and/or other managed nodes 1260. The managed node 1260, in the illustrative embodiment, may determine to distribute the portions 1404 across multiple managed nodes 1260. By doing so, if any one managed node 1260 becomes unavailable, a relatively large percentage of the portions 1404 may still be obtained from the other managed nodes 1260. Further, as indicated in block 1516, the managed node 1260 may assign the redundant portions 1404 to managed nodes 1260 that are different from the managed nodes 1260 that are to store the original portions 1404 (e.g., the portions 1404 that the redundant portions 1404 would be used to recreate), so that both the original and redundant version of a portion 1404 do not become lost if the corresponding managed node 1260 becomes inoperative. In block 1518, in the illustrative embodiment, the managed node 1260 associates a key with each portion 1404. As described above, the key uniquely identifies each portion 1404 and may be generated by executing a hash function on the portion 1404 and combining the hash with target location information such as by appending a unique address (e.g., media access control address) of the managed node 1260 to store the data and a unique address of the present managed node 1260, or based on any other method for uniquely identifying the portion 1404. In block 1520, in the illustrative embodiment, the managed node 1260 may generate and store a map of the portions 1404, the corresponding keys, and the data storage devices 1312 that are to store the portions 1404 (e.g., a dataset map 1402). When a portion 1404 is to be stored on a remote managed node 1260, the present managed node 1260 may not have information regarding the specific data storage devices 1312 present in the remote managed node 1260. Accordingly, in such

embodiments, the present managed node 1260 stores an identifier (e.g., the media access control address or other unique identifier) of the remote managed node 1260 where the one or more portions 1404 are to be stored, rather than identifiers of specific data storage devices 1312 within the remote managed node 1260.

[0066] In block 1522, the managed node 1260 writes the portions 1404 to the multiple data storage devices 1312, such as based on the determination of the assignment of the portions 1404 from block 1514. In doing so, the managed node 1260 may write one or more portions 1404 to data storage devices 1312 local to the present managed node 1260, as indicated in block 1524. In doing so, in the illustrative embodiment, the managed node 1260 stores the corresponding portions 1404 in one or more of the local data storage devices 1312 with their corresponding keys (e.g., in a table of the keys and corresponding logical block addresses where the portions 1404 are written). Additionally, as indicated in block 1526, the managed node 1260 may write one or more portions 1404 to remote data storage devices 1312 of other managed nodes 1260, such as by issuing requests to those managed nodes 1260 with the portions 1404 to write and the keys to be associated with the portions 1404. As indicated in block 1528, by concurrently writing the various portions 1404 to different data storage devices 1312, the managed node 1260, in effect, writes the dataset at a combined rate that is greater than the peak data throughput rate of any one of the data storage devices 1312. Subsequently, the method 1500 advances to block 1530 of FIG. 16, in which the managed node 1260 may receive a request from a workload executed by the managed node 1260 to read a dataset.

[0067] Referring now to FIG. 16, as described above, the managed node 1260 may receive a request from a workload to read a dataset (e.g., the previously written dataset or another dataset that was written at a different time). In block 1532, the managed node 1260 determines whether a read request was received. If not, the method 1500 advances to block 1502 of FIG. 15, in which the managed node 1260 determines whether to continue managing distributed data. Otherwise, the method 1500 advances to block 1534 in which the managed node 1260 reads the portions 1404 of the dataset from the multiple data storage devices 1312 on which the portions 1404 are stored. In doing so, as indicated in block 1536, the managed node 1260 determines the data storage devices 1312 where the portions 1404 of the dataset are stored. In the illustrative embodiment, in determining the data storage device 1312, the managed node 1260 accesses the map (e.g., the dataset map 1402) of the portions 1404, the corresponding keys, and the corresponding data storage devices 1312 where the portions 1404 are stored, as indicated in block 1538. As described above, with reference to block 1520, in some embodiments, the dataset map 1402 may include an identifier (e.g., media access control

address) of a remote managed node 1260 where a particular portion 1404 is stored, rather than the specific data storage device 1312 within that remote managed node 1260. In block 1540, the managed node 1260 may read from one or more local data storage devices 1312. In doing so, as indicated in block 1542, the managed node 1260 reads a portion 1404 stored in association with a key (e.g., a key from the dataset map 1402) from a local data storage device 1312. As indicated in block 1544, the managed node 1260 may read from remote data storage devices 1312, such as by issuing read requests to the remote managed nodes 1260 having one or more data storage devices 1312 in which corresponding portions 1404 are stored. In reading from the remote data storage devices 1312, in the illustrative embodiment, the managed node 1260 reads a portion 1404 stored in association with a key (e.g., by issuing a request to the remote managed node 1260 to read the portion 1404 associated with the corresponding key), as indicated in block 1546. The managed node 1260 may apply an error correction scheme (e.g., a low density parity check scheme, a Reed-Solomon scheme, etc.) to correct errors in any portions 1404 read from the local data storage devices 1312 and/or read by remote managed nodes 1260 as the portions 1404 are read, and/or may apply an error correction scheme later, when combining the portions 1404 as described herein.

[0068] As indicated in block 1548, in reading the portions 1404, the managed node 1260 may identify one or more inoperative data storage devices 1312 (e.g., local data storage devices 1312 that are inoperative and/or one or more managed nodes 1260 that have become unresponsive or have reported an inoperative status for one or more data storage devices 1312 local to them) and read the corresponding redundant portions 1404 from other data storage devices 1312. As indicated in block 1550, in reading the portions 1404, the managed node 1260 effectively reads the dataset requested by the workload at a combined rate that is greater than the peak data throughput rate of any one of the data storage devices 1312 one which a portion 1404 of the dataset is stored. After reading the portions 1404 of the dataset, the managed node 1260, in block 1552, combines the read portions 1404 to reconstruct the dataset requested by the workload. In doing so, the managed node 1260 may apply an error correction scheme (e.g., a low density parity check, a Reed-Solomon scheme, etc.) to correct any data corruption present in the read portions 1404. Afterwards, the method 1500 loops back to block 1502 of FIG. 15 in which the managed node 1260 determines whether to continue managing distributed data.

EXAMPLES

[0069] 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.

[0070] Example 1 includes a managed node to manage distributed data, the managed node comprising a distributed data manager to distribute a dataset over multiple data storage devices coupled to a network, wherein each data storage device has a peak data throughput rate; and a network communicator to request a corresponding portion of the dataset from each data storage device and receive the requested portions of the dataset at a combined data throughput rate that is greater than the peak data throughput rate of any one of the data storage devices; wherein the distributed data manager is further to combine the received portions of the dataset to reconstruct the dataset.

[0071] Example 2 includes the subject matter of Example 1, and wherein to request the corresponding portion of the dataset from each data storage device comprises to receive a request from a workload for the dataset; determine, in response to the request from the workload, the corresponding data storage device on which each portion is stored; and request the corresponding portion after determining the corresponding data storage devices.

[0072] Example 3 includes the subject matter of any of Examples 1 and 2, and wherein to distribute the dataset over multiple data storage devices comprises to distribute the dataset in response to a request from a workload to store the dataset.

[0073] Example 4 includes the subject matter of any of Examples 1-3, and wherein to distribute the dataset comprises to write the portions on data storage devices that are physically located on different managed nodes.

[0074] Example 5 includes the subject matter of any of Examples 1-4, and wherein to distribute the dataset comprises to write the portions on solid state drives.

[0075] Example 6 includes the subject matter of any of Examples 1-5, and wherein the distributed data manager is further to associate each portion with a key and wherein to request the corresponding portion comprises to request the portion stored in association with each key.

[0076] Example 7 includes the subject matter of any of Examples 1-6, and wherein the distributed data manager is further to store a map indicative of locations of the portions of the dataset among the data storage devices.

[0077] Example 8 includes the subject matter of any of Examples 1-7, and wherein to request the corresponding portion comprises to access the map to determine the data storage device on which each corresponding portion is stored.

[0078] Example 9 includes the subject matter of any of Examples 1-8, and wherein to distribute the dataset comprises to write at least one redundant portion of the data set to at least one of the data storage devices.

[0079] Example 10 includes the subject matter of any of Examples 1-9, and wherein to request a corresponding portion comprises to determine whether a data storage device on which one of the portions is stored is inoperative; determine, in response to a determination that the data storage device is inoperative, an alternative data storage device on which a redundant version of the portion is stored; and request the redundant version of the portion from the alternative data storage device.

[0080] Example 11 includes the subject matter of any of Examples 1-10, and wherein to combine the received portions comprises to apply an error correction scheme to the received portions to correct corrupted data.

[0081] Example 12 includes the subject matter of any of Examples 1-11, and wherein to distribute the dataset over multiple data storage devices comprises to apply an error correction scheme to generate one or more redundant versions of one or more of the portions; and write the redundant versions to data storage devices in managed nodes that are separate from original versions of the corresponding portions.

[0082] Example 13 includes the subject matter of any of Examples 1-12, and wherein to distribute the dataset comprises to write the portions of the dataset to the data storage devices at a data throughput rate that is greater than the peak data throughput rate of any of the data storage devices.

[0083] Example 14 includes a method for managing distributed data, the method comprising distributing, by a managed node, a dataset over multiple data storage devices coupled to a network, wherein each data storage device has a peak data throughput rate; requesting, by the managed node, a corresponding portion of the dataset from each data storage device; receiving, by the managed node, the requested portions of the dataset at a combined data throughput rate that is greater than the peak data throughput rate of any one of the data storage devices; and combining, by the managed node, the received portions of the dataset to reconstruct the dataset.

[0084] Example 15 includes the subject matter of Example 14, and wherein requesting the corresponding portion of the dataset from each data storage device comprises receiving a request from a workload for the dataset; determining, in response to the request from the workload, the corresponding data storage device on which each portion is stored; and requesting the corresponding portion after determining the corresponding data storage devices.

[0085] Example 16 includes the subject matter of any of Examples 14 and 15, and wherein distributing the dataset over multiple data storage devices comprises distributing the dataset in response to a request from a workload to store the dataset.

[0086] Example 17 includes the subject matter of any of Examples 14-16, and wherein distributing the dataset comprises writing the portions on data storage devices that are physically located on different managed nodes.

[0087] Example 18 includes the subject matter of any of Examples 14-17, and wherein distributing the dataset comprises writing the portions on solid state drives.

[0088] Example 19 includes the subject matter of any of Examples 14-18, and further including associating, by the managed node, each portion with a key and wherein requesting the corresponding portion comprises requesting the portion stored in association with each key.

[0089] Example 20 includes the subject matter of any of Examples 14-19, and further including storing, by the managed node, a map indicative of locations of the portions of the dataset among the data storage devices.

[0090] Example 21 includes the subject matter of any of Examples 14-20, and wherein requesting the corresponding portion comprises accessing the map to determine the data storage device on which each corresponding portion is stored.

[0091] Example 22 includes the subject matter of any of Examples 14-21, and wherein distributing the dataset comprises writing at least one redundant portion of the data set to at least one of the data storage devices.

[0092] Example 23 includes the subject matter of any of Examples 14-22, and wherein requesting a corresponding portion comprises determining whether a data storage device on which one of the portions is stored is inoperative; determining, in response to a determination that the data storage device is inoperative, an alternative data storage device on which a redundant version of the portion is stored; and requesting the redundant version of the portion from the alternative data storage device.

[0093] Example 24 includes the subject matter of any of Examples 14-23, and wherein combining the received portions comprises applying an error correction scheme to the received portions to correct corrupted data.

[0094] Example 25 includes the subject matter of any of Examples 14-24, and wherein distributing the dataset over multiple data storage devices comprises applying an error correction scheme to generate one or more redundant versions of one or more of the portions; and writing the redundant versions to data storage devices in managed nodes that are separate from original versions of the corresponding portions.

[0095] Example 26 includes the subject matter of any of Examples 14-25, and wherein distributing the dataset comprises writing the portions of the dataset to the data storage devices at a data throughput rate that is greater than the peak data throughput rate of any of the data storage devices.

[0096] Example 27 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 14-26.

[0097] Example 28 includes a managed node comprising means for distributing a dataset over multiple data storage devices coupled to a network, wherein each data storage device has a peak data throughput rate; means for requesting a corresponding portion of the dataset from each data storage device; means for receiving the requested portions of the dataset at a combined data throughput rate that is greater than the peak data throughput rate of any one of the data storage devices; and means for combining the received portions of the dataset to reconstruct the dataset.

[0098] Example 29 includes the subject matter of Example 28, and wherein the means for requesting the corresponding portion of the dataset from each data storage device comprises means for receiving a request from a workload for the dataset; means for determining, in response to the request from the workload, the corresponding data storage device on which each portion is stored; and means for requesting the corresponding portion after determining the corresponding data storage devices.

[0099] Example 30 includes the subject matter of any of Examples 28 and 29, and wherein the means for distributing the dataset over multiple data storage devices comprises means for distributing the dataset in response to a request from a workload to store the dataset.

[00100] Example 31 includes the subject matter of any of Examples 28-30, and wherein the means for distributing the dataset comprises means for writing the portions on data storage devices that are physically located on different managed nodes.

[00101] Example 32 includes the subject matter of any of Examples 28-31, and wherein the means for distributing the dataset comprises means for writing the portions on solid state drives.

[00102] Example 33 includes the subject matter of any of Examples 28-32, and further including means for associating each portion with a key and wherein the means for requesting the corresponding portion comprises means for requesting the portion stored in association with each key.

[00103] Example 34 includes the subject matter of any of Examples 28-33, and further including means for storing a map indicative of locations of the portions of the dataset among the data storage devices.

[00104] Example 35 includes the subject matter of any of Examples 28-34, and wherein the means for requesting the corresponding portion comprises means for accessing the map to determine the data storage device on which each corresponding portion is stored.

[00105] Example 36 includes the subject matter of any of Examples 28-35, and wherein the means for distributing the dataset comprises means for writing at least one redundant portion of the data set to at least one of the data storage devices.

[00106] Example 37 includes the subject matter of any of Examples 28-36, and wherein the means for requesting a corresponding portion comprises means for determining whether a data storage device on which one of the portions is stored is inoperative; means for determining, in response to a determination that the data storage device is inoperative, an alternative data storage device on which a redundant version of the portion is stored; and means for requesting the redundant version of the portion from the alternative data storage device.

[00107] Example 38 includes the subject matter of any of Examples 28-37, and wherein the means for combining the received portions comprises means for applying an error correction scheme to the received portions to correct corrupted data.

[00108] Example 39 includes the subject matter of any of Examples 28-38, and wherein the means for distributing the dataset over multiple data storage devices comprises means for applying an error correction scheme to generate one or more redundant versions of one or more of the portions; and means for writing the redundant versions to data storage devices in managed nodes that are separate from original versions of the corresponding portions.

[00109] Example 40 includes the subject matter of any of Examples 28-39, and wherein the means for distributing the dataset comprises means for writing the portions of the dataset to the data storage devices at a data throughput rate that is greater than the peak data throughput rate of any of the data storage devices.

WHAT IS CLAIMED IS:

1. A managed node to manage distributed data, the managed node comprising:
 - a distributed data manager to distribute a dataset over multiple data storage devices coupled to a network, wherein each data storage device has a peak data throughput rate; and
 - a network communicator to request a corresponding portion of the dataset from each data storage device and receive the requested portions of the dataset at a combined data throughput rate that is greater than the peak data throughput rate of any one of the data storage devices;wherein the distributed data manager is further to combine the received portions of the dataset to reconstruct the dataset.
2. The managed node of claim 1, wherein to request the corresponding portion of the dataset from each data storage device comprises to:
 - receive a request from a workload for the dataset;
 - determine, in response to the request from the workload, the corresponding data storage device on which each portion is stored; and
 - request the corresponding portion after determining the corresponding data storage devices.
3. The managed node of claim 1, wherein to distribute the dataset over multiple data storage devices comprises to distribute the dataset in response to a request from a workload to store the dataset.
4. The managed node of claim 1, wherein to distribute the dataset comprises to write the portions on data storage devices that are physically located on different managed nodes.
5. The managed node of claim 1, wherein to distribute the dataset comprises to write the portions on solid state drives.
6. The managed node of claim 1, wherein the distributed data manager is further to associate each portion with a key and wherein to request the corresponding portion comprises to request the portion stored in association with each key.

7. The managed node of claim 1, wherein the distributed data manager is further to store a map indicative of locations of the portions of the dataset among the data storage devices.

8. The managed node of claim 7, wherein to request the corresponding portion comprises to access the map to determine the data storage device on which each corresponding portion is stored.

9. The managed node of claim 1, wherein to distribute the dataset comprises to write at least one redundant portion of the data set to at least one of the data storage devices.

10. The managed node of claim 1, wherein to request a corresponding portion comprises to:

determine whether a data storage device on which one of the portions is stored is inoperative;

determine, in response to a determination that the data storage device is inoperative, an alternative data storage device on which a redundant version of the portion is stored; and

request the redundant version of the portion from the alternative data storage device.

11. The managed node of claim 1, wherein to combine the received portions comprises to apply an error correction scheme to the received portions to correct corrupted data.

12. The managed node of claim 1, wherein to distribute the dataset over multiple data storage devices comprises to:

apply an error correction scheme to generate one or more redundant versions of one or more of the portions; and

write the redundant versions to data storage devices in managed nodes that are separate from original versions of the corresponding portions.

13. A method for managing distributed data, the method comprising:
distributing, by a managed node, a dataset over multiple data storage devices coupled to a network, wherein each data storage device has a peak data throughput rate;

requesting, by the managed node, a corresponding portion of the dataset from each data storage device;

receiving, by the managed node, the requested portions of the dataset at a combined data throughput rate that is greater than the peak data throughput rate of any one of the data storage devices; and

combining, by the managed node, the received portions of the dataset to reconstruct the dataset.

14. The method of claim 13, wherein requesting the corresponding portion of the dataset from each data storage device comprises:

receiving a request from a workload for the dataset;

determining, in response to the request from the workload, the corresponding data storage device on which each portion is stored; and

requesting the corresponding portion after determining the corresponding data storage devices.

15. The method of claim 13, wherein distributing the dataset over multiple data storage devices comprises distributing the dataset in response to a request from a workload to store the dataset.

16. The method of claim 13, wherein distributing the dataset comprises writing the portions on data storage devices that are physically located on different managed nodes.

17. The method of claim 13, wherein distributing the dataset comprises writing the portions on solid state drives.

18. The method of claim 13, further comprising associating, by the managed node, each portion with a key and wherein requesting the corresponding portion comprises requesting the portion stored in association with each key.

19. The method of claim 13, further comprising storing, by the managed node, a map indicative of locations of the portions of the dataset among the data storage devices.

20. The method of claim 19, wherein requesting the corresponding portion comprises accessing the map to determine the data storage device on which each corresponding portion is stored.

21. The method of claim 13, wherein distributing the dataset comprises writing at least one redundant portion of the data set to at least one of the data storage devices.

22. The method of claim 13, wherein requesting a corresponding portion comprises:
determining whether a data storage device on which one of the portions is stored is inoperative;

determining, in response to a determination that the data storage device is inoperative, an alternative data storage device on which a redundant version of the portion is stored; and

requesting the redundant version of the portion from the alternative data storage device.

23. The method of claim 13, wherein combining the received portions comprises applying an error correction scheme to the received portions to correct corrupted data.

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

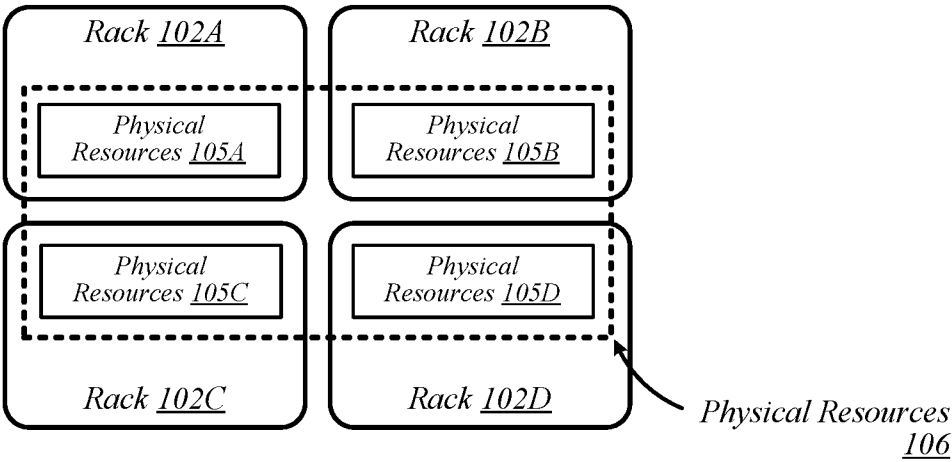


FIG. 1

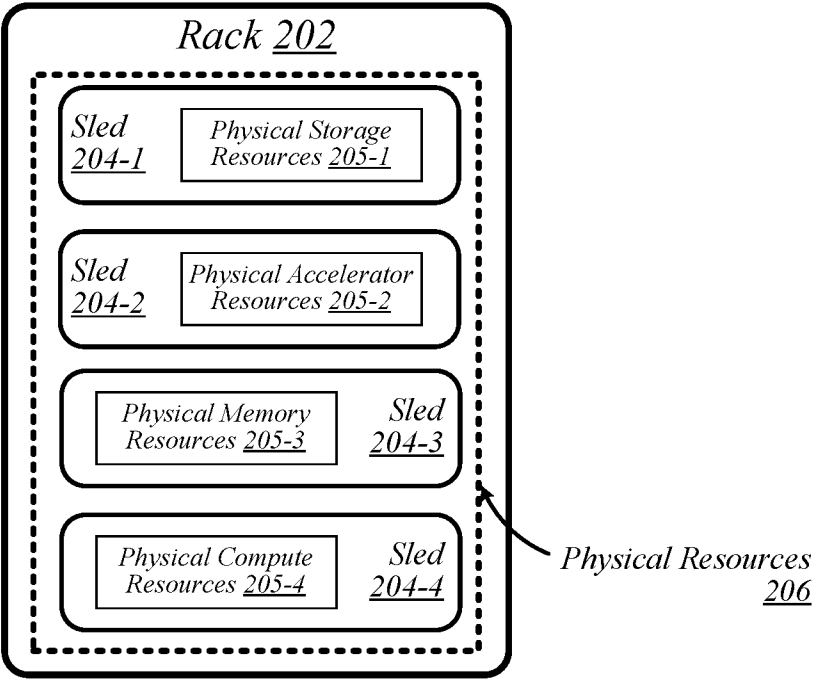


FIG. 2

Data Center 300

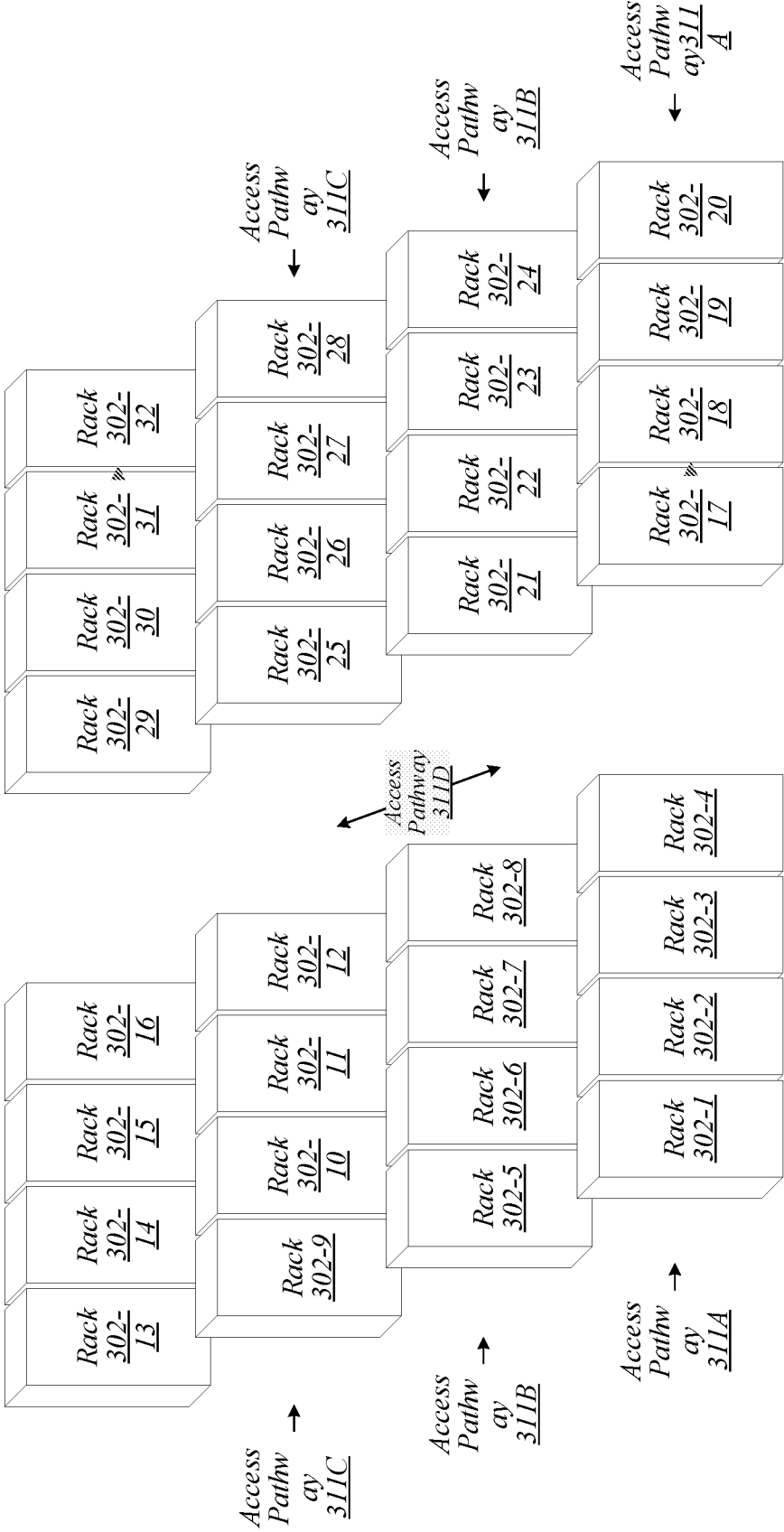


FIG. 3

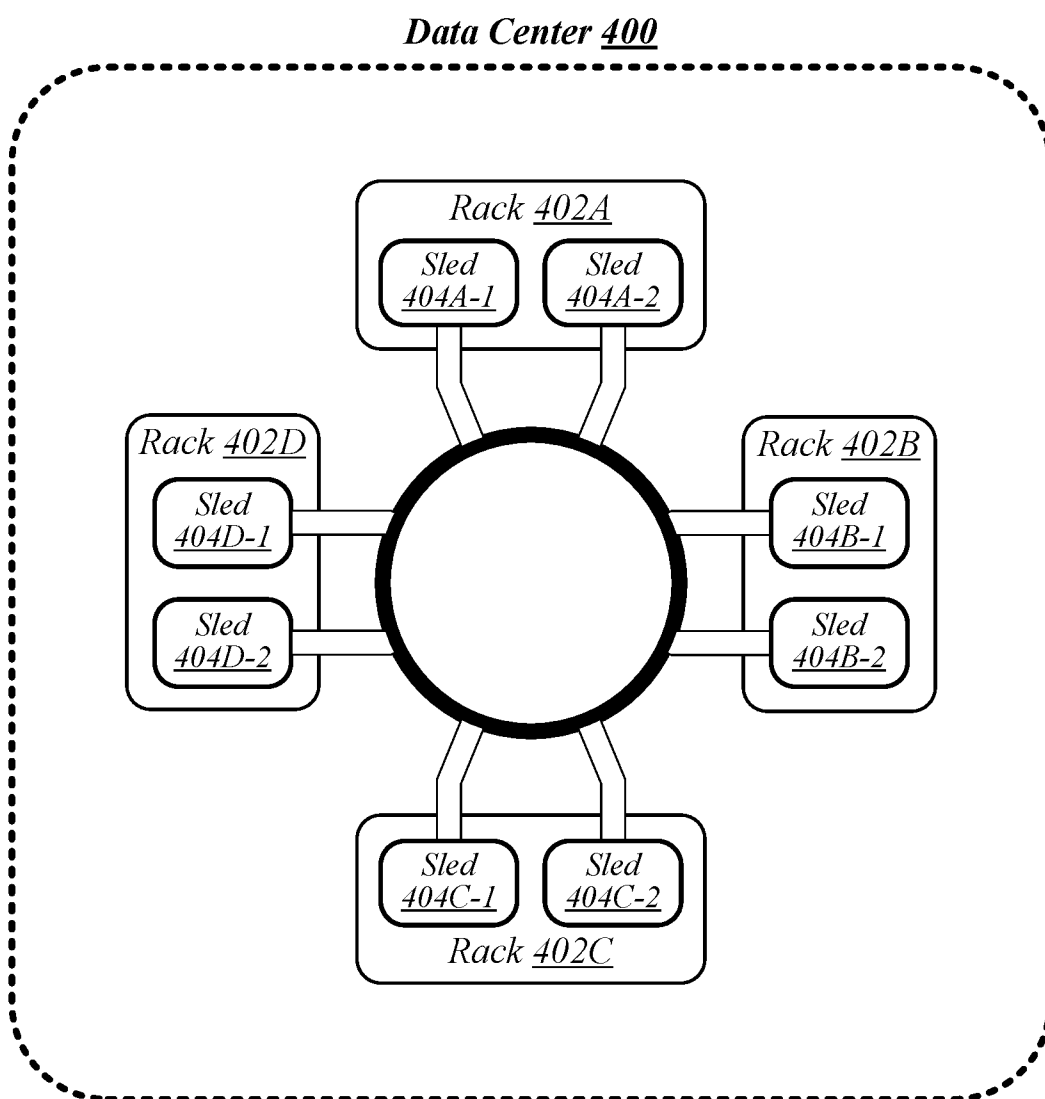


FIG. 4

500

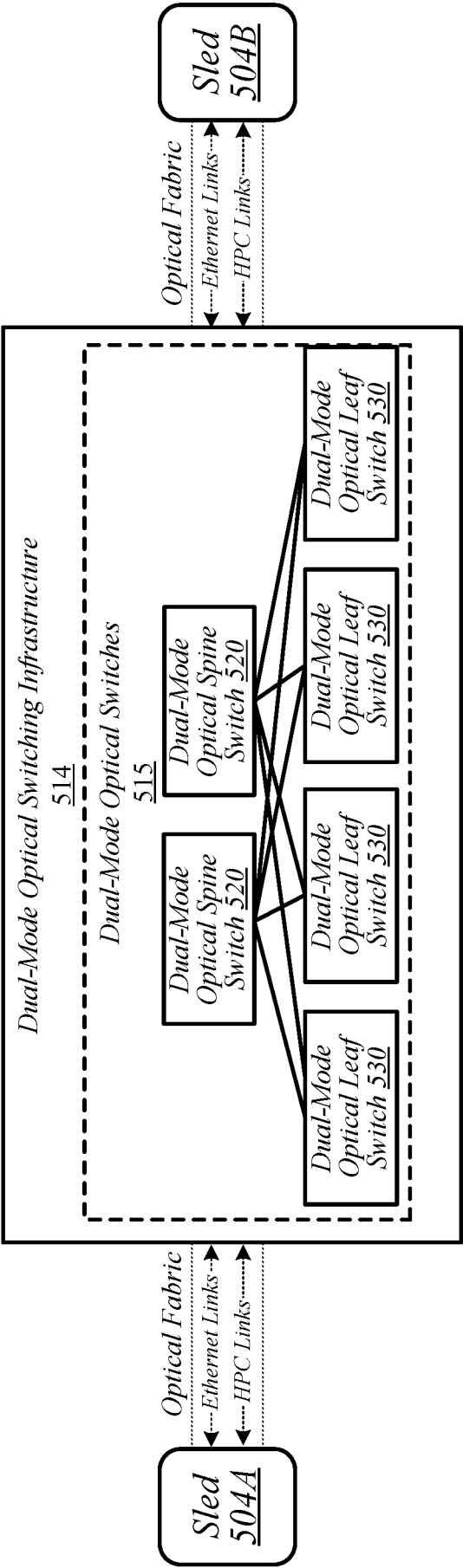


FIG. 5

600

Rack Access Region 601

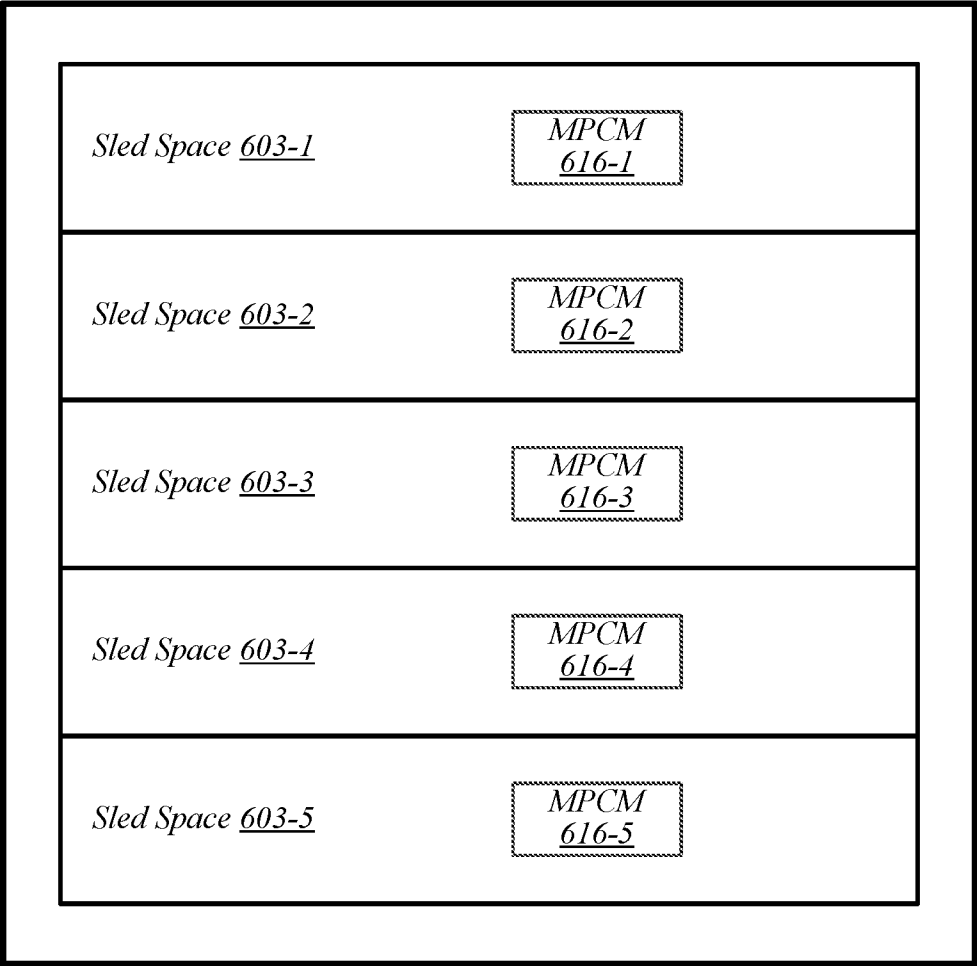


FIG. 6

Sled 704

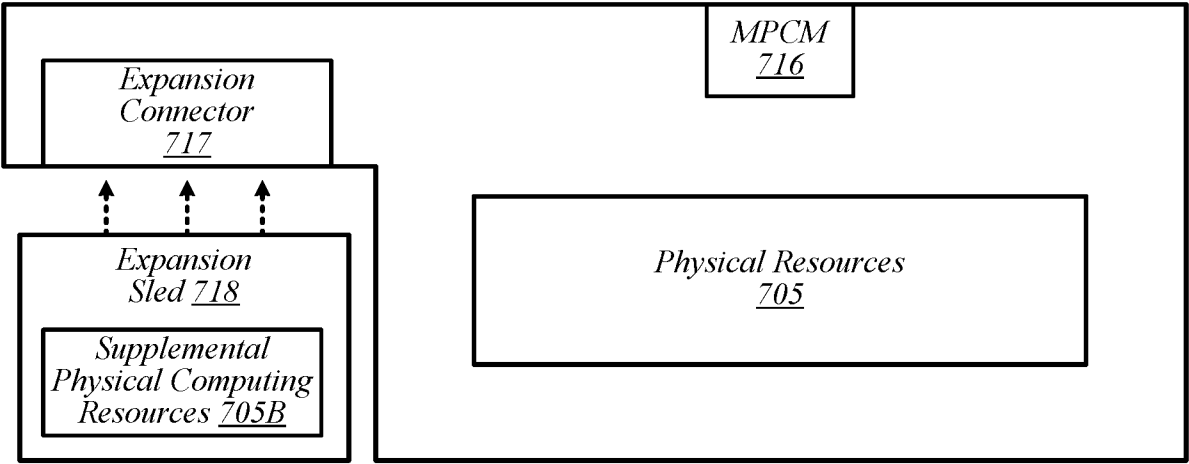


FIG. 7

800

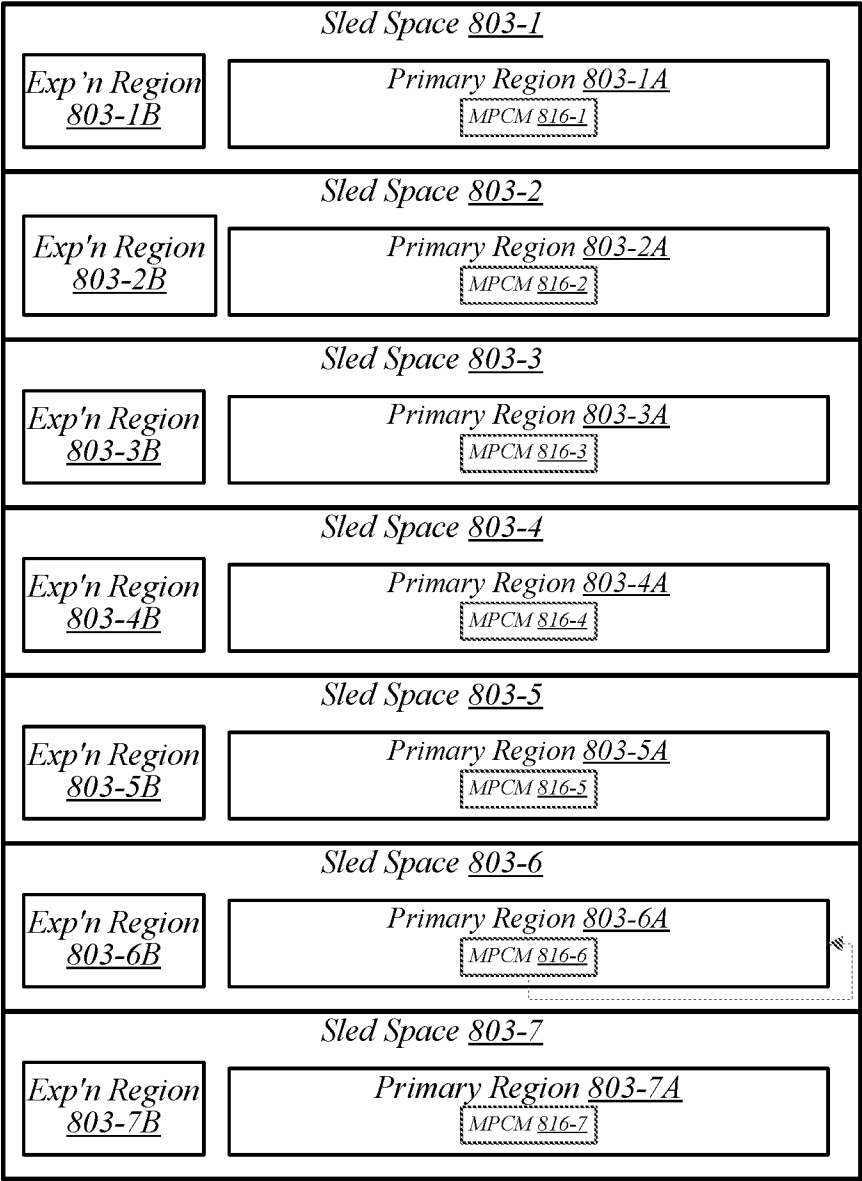


FIG. 8

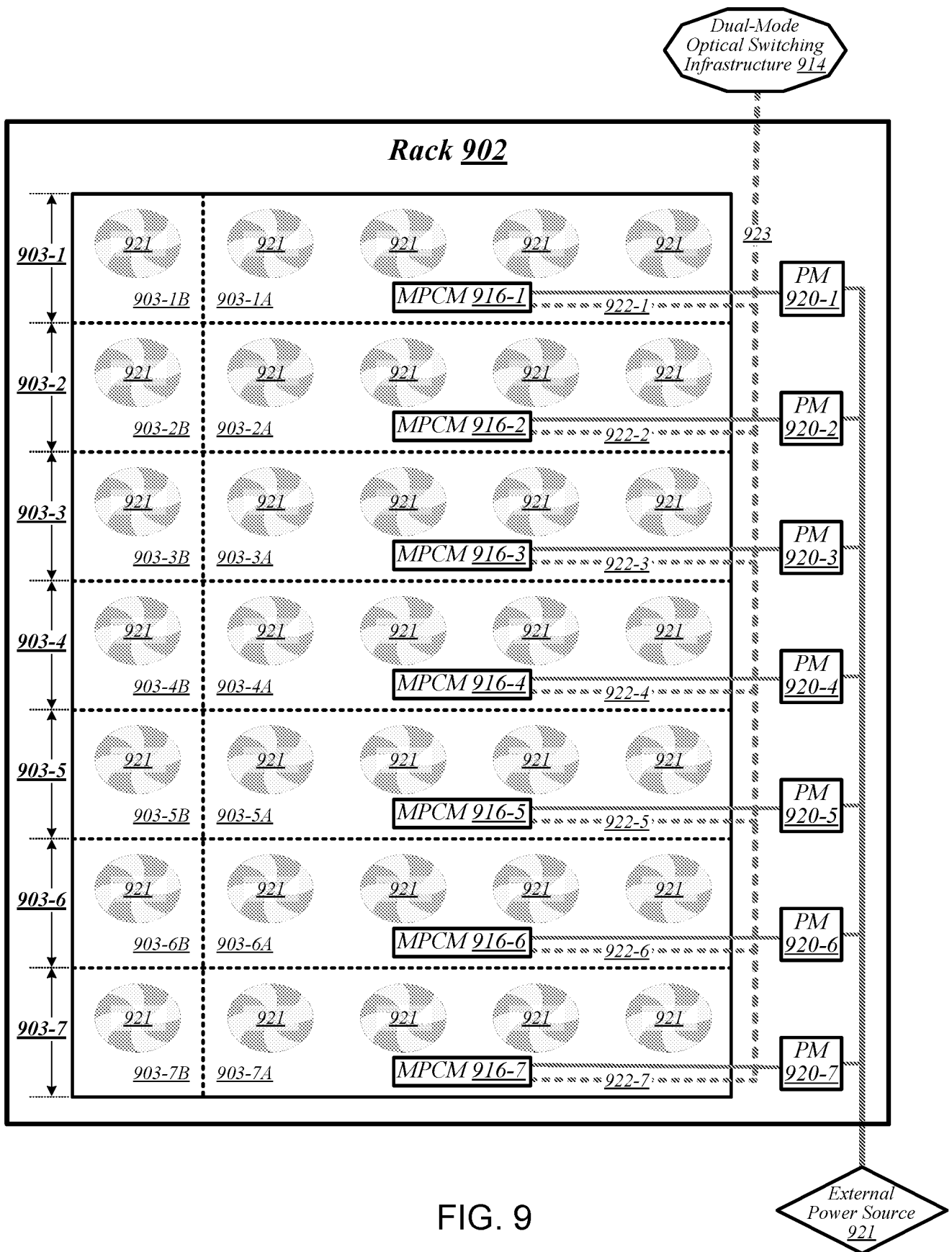


FIG. 9

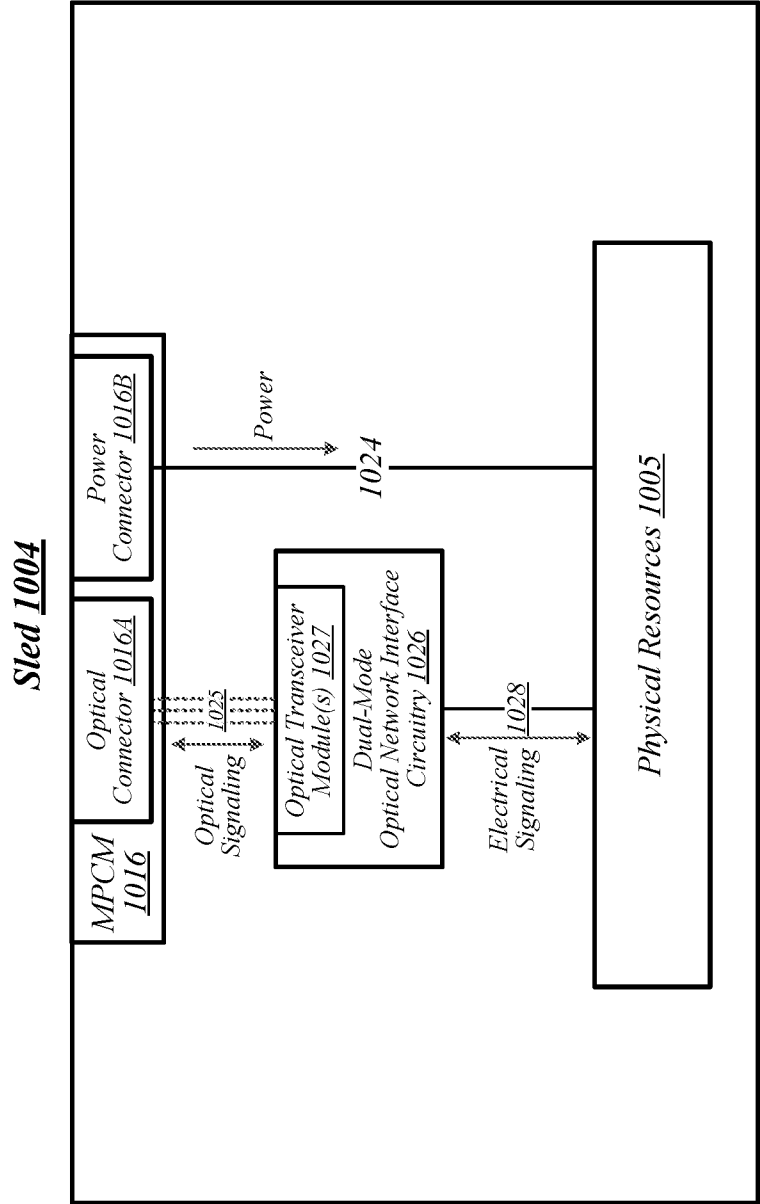


FIG. 10

Data Center 1100

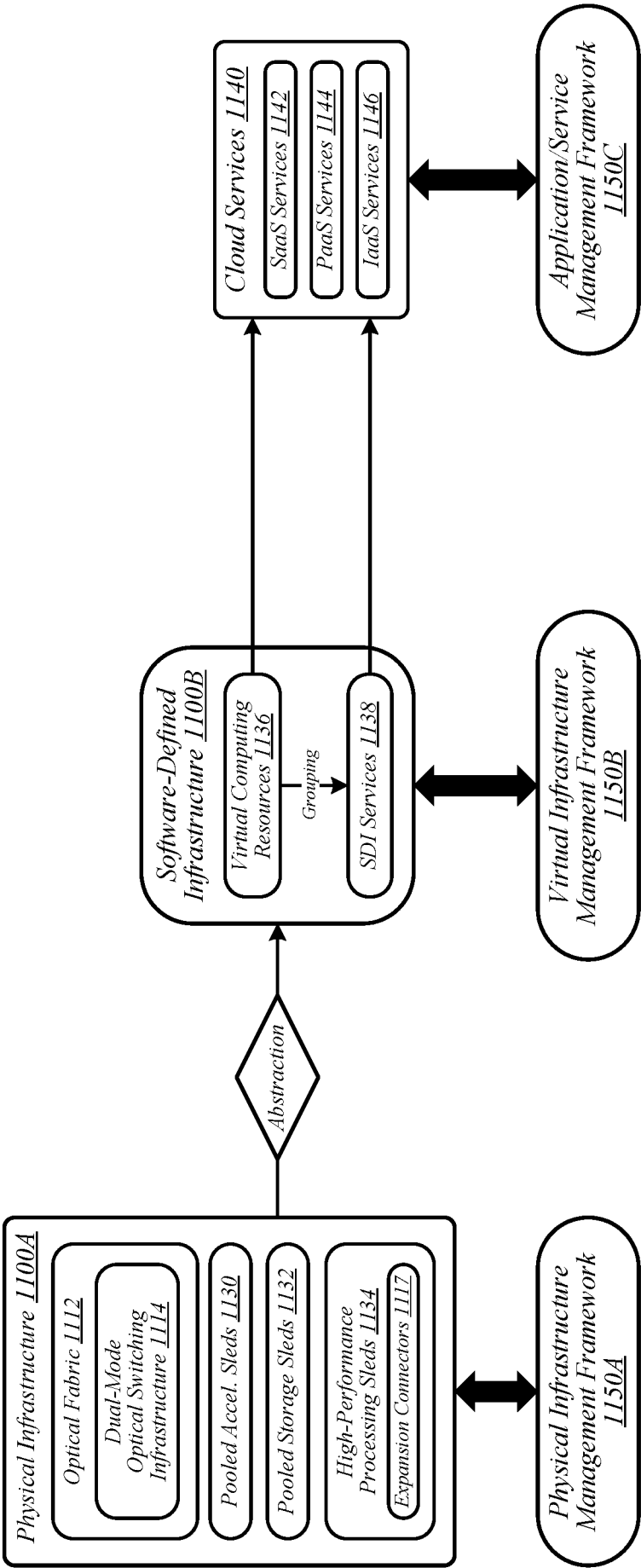


FIG. 11

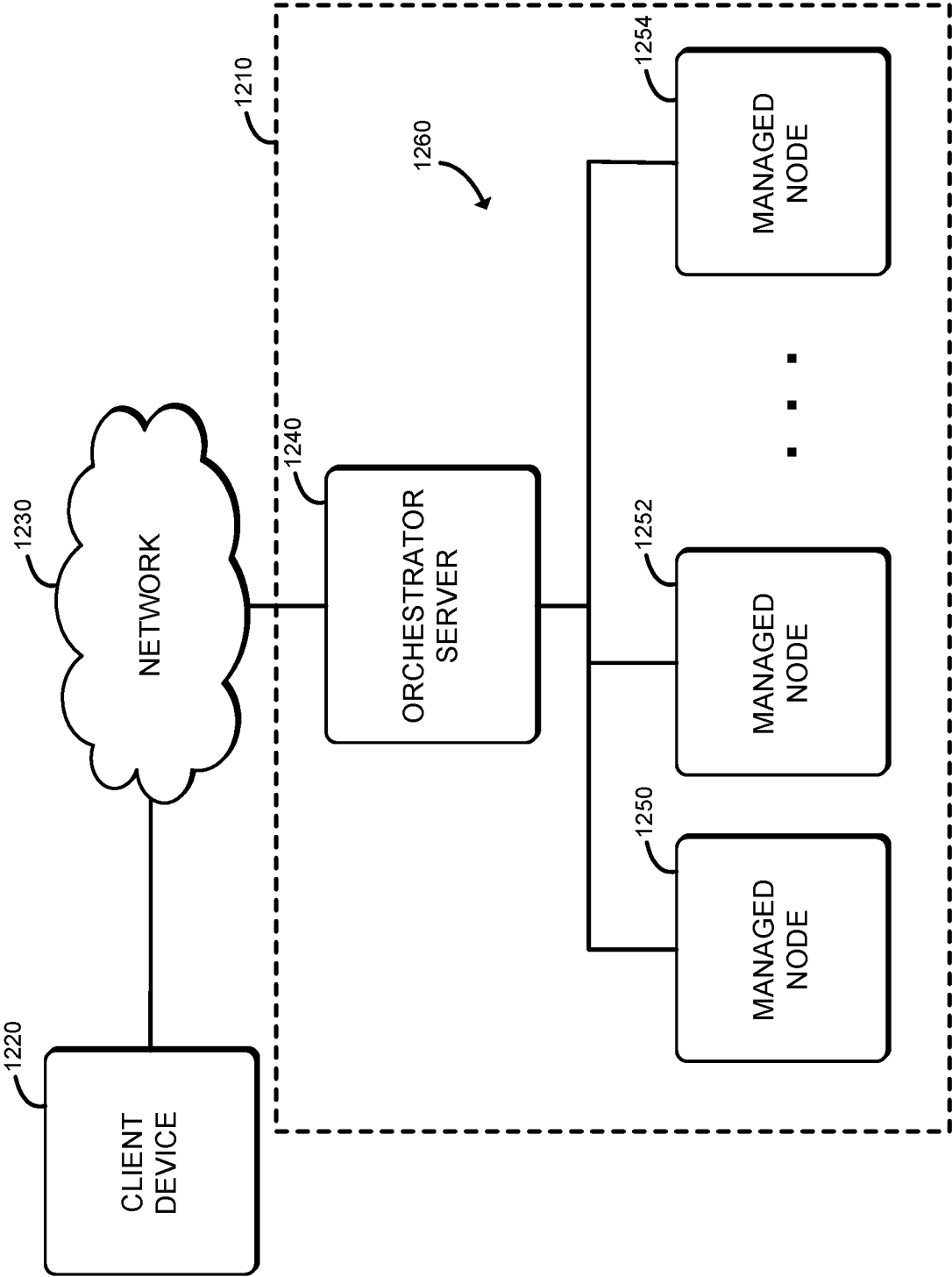


FIG. 12

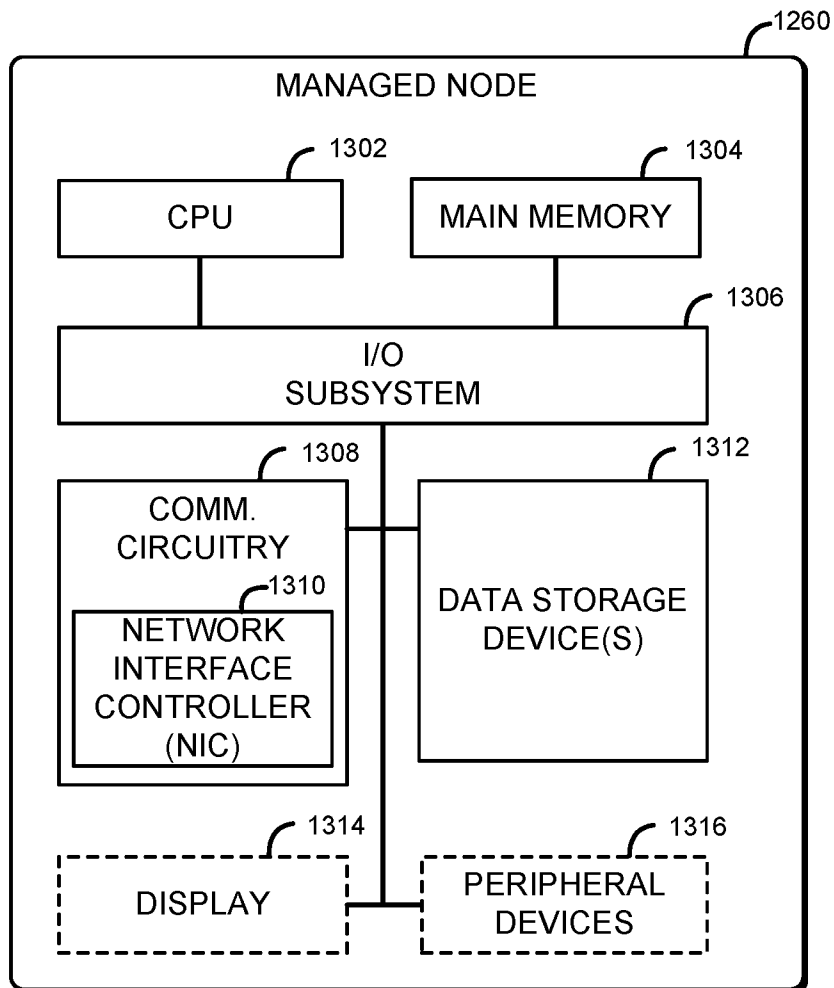


FIG. 13

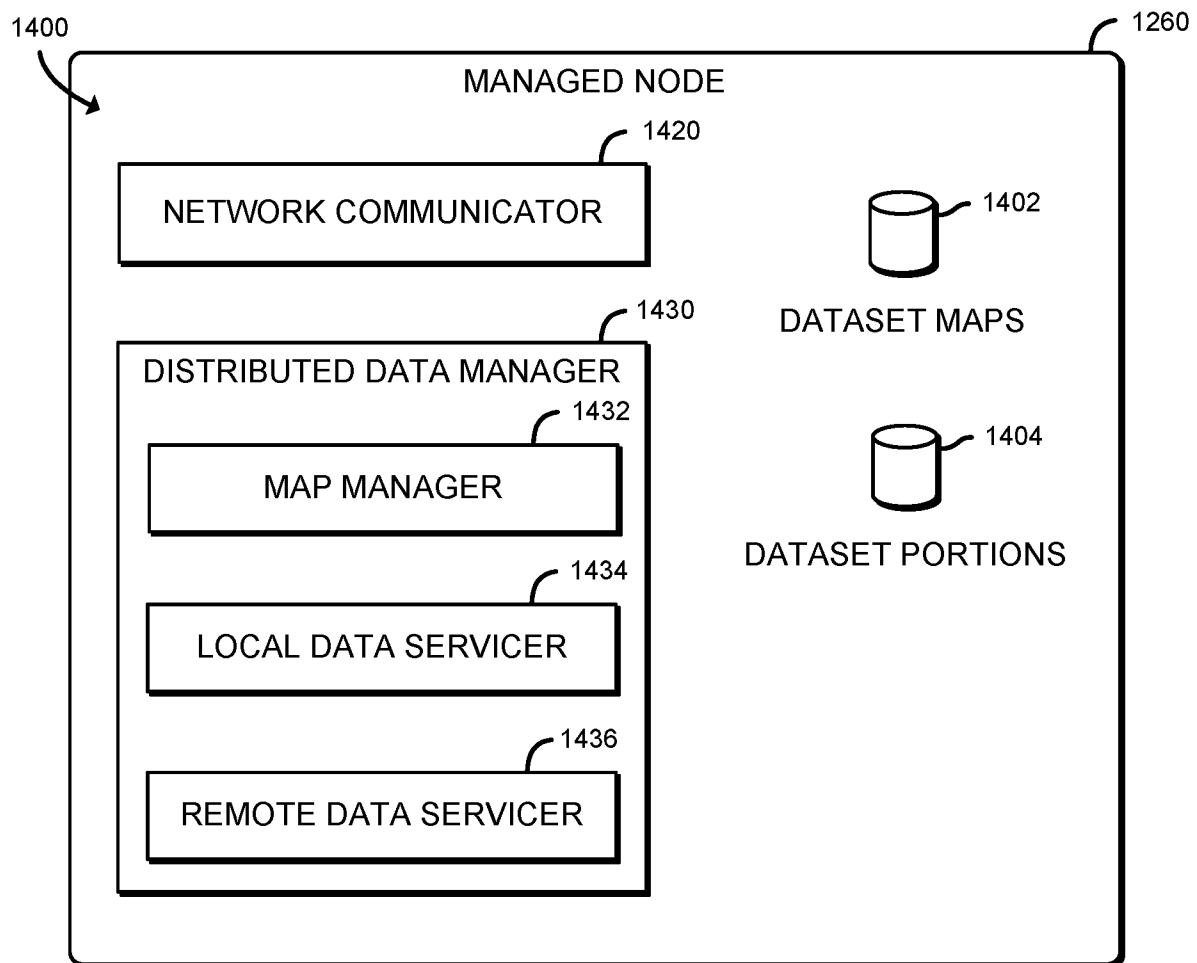


FIG. 14

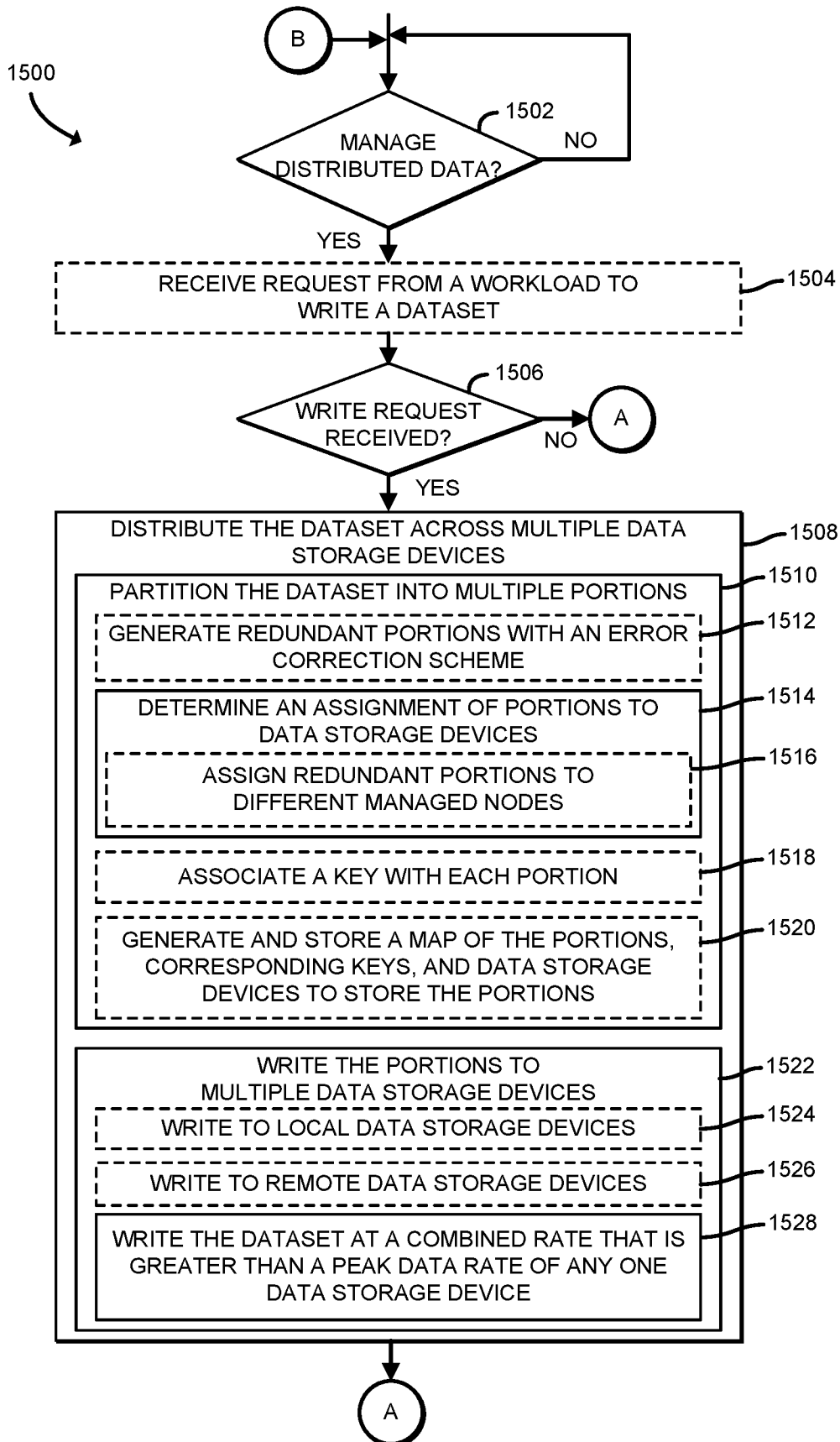
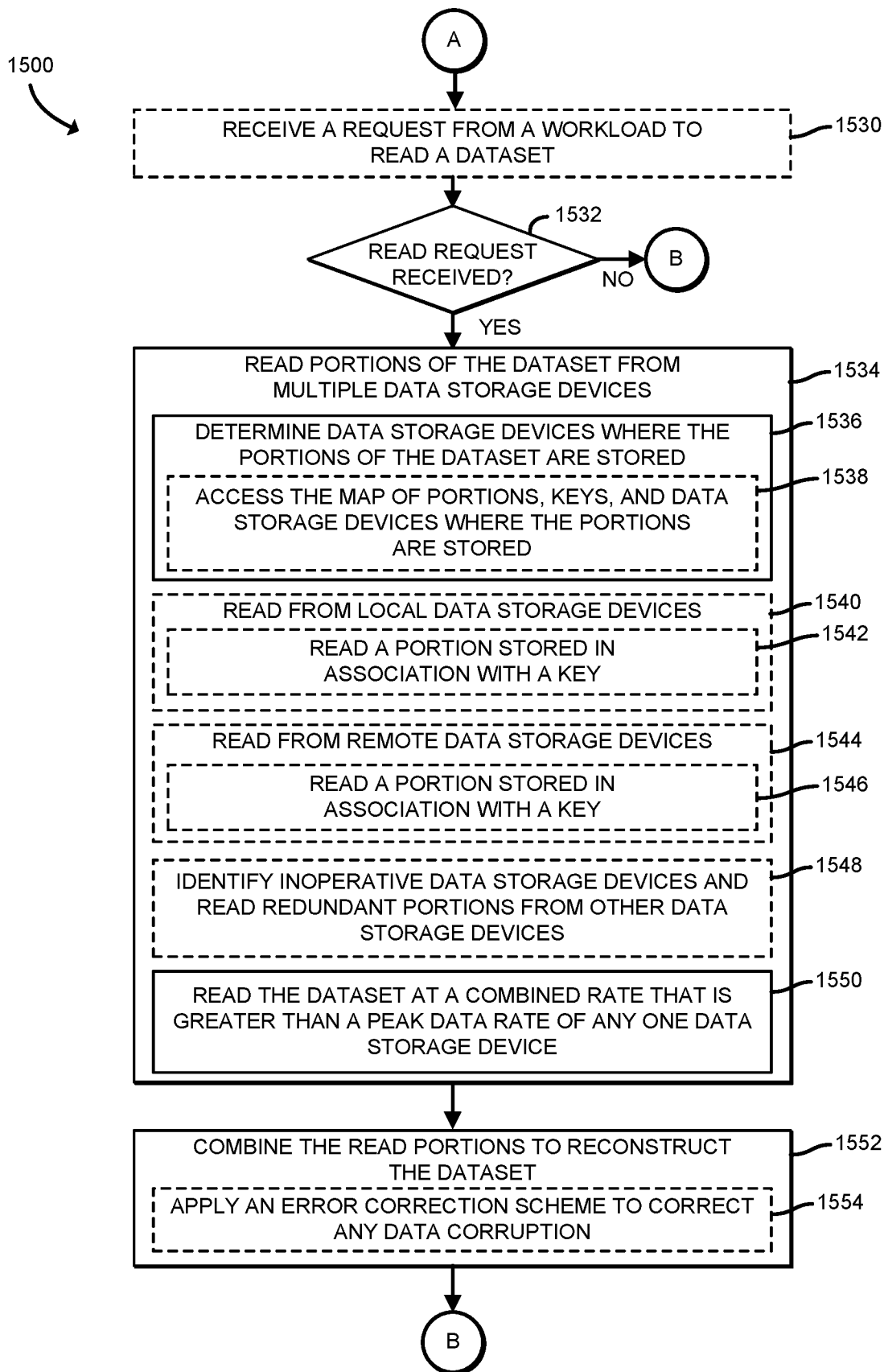


FIG. 15

16/16



A. CLASSIFICATION OF SUBJECT MATTER**G06F 17/30(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)

G06F 17/30; G06F 12/02; G06F 11/14; G06F 13/16; G06F 12/14; G06F 12/00; G06F 13/28; G06F 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: storage, data, segment, throughput, redundant, bottleneck, reassemble, combine, and similar terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2014-0156903 A1 (EMPIRE TECHNOLOGY DEVELOPMENT LLC) 05 June 2014 See paragraphs [0025]-[0028] and [0031]-[0038]; and figures 1-2.	1-25
Y	US 2010-0161926 A1 (HONG LI et al.) 24 June 2010 See paragraphs [0013] and [0022]-[0023]; claim 1; and figure 1.	1-25
Y	US 8386841 B1 (DILIP RENADE) 26 February 2013 See column 3, line 42 - column 4, line 26; claim 1; and figures 2-3.	9-10, 12, 21-22
A	US 2012-0011301 A1 (RYAN J. GOSS et al.) 12 January 2012 See paragraphs [0148]-[0150] and figure 15.	1-25
A	US 2013-0339784 A1 (CRAIG A. BICKELMAN et al.) 19 December 2013 See paragraph [0029]; claim 1; and figure 3.	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

25 October 2017 (25.10.2017)

Date of mailing of the international search report

26 October 2017 (26.10.2017)

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, Dal Kyong

Telephone No. +82-42-481-8440



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/038301

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2014-0156903 A1	05/06/2014	US 9501437 B2 WO 2014-077823 A2 WO 2014-077823 A3	22/11/2016 22/05/2014 11/06/2015
US 2010-0161926 A1	24/06/2010	None	
US 8386841 B1	26/02/2013	None	
US 2012-0011301 A1	12/01/2012	None	
US 2013-0339784 A1	19/12/2013	None	