



(51) International Patent Classification:

G06F 12/02 (2006.01)

(21) International Application Number:

PCT/US2017/038317

(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
15/396,063 30 December 2016 (30.12.2016) US

(71) Applicant (for all designated States except): INTEL CORPORATION [US/US]; 2200 Mission College Boulevard, Santa Clara, California 95054 (US).

(72) Inventors: MILLER, Steven C.; 2317 Tuscany Circle, Livermore, California 94550 (US). GRIMSRUD, Knut S.; 48009 SW Morel Lane, Forest Grove, Oregon 97116 (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, 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))

(54) Title: TECHNOLOGIES FOR ENHANCED MEMORY WEAR LEVELING

Data Center 100

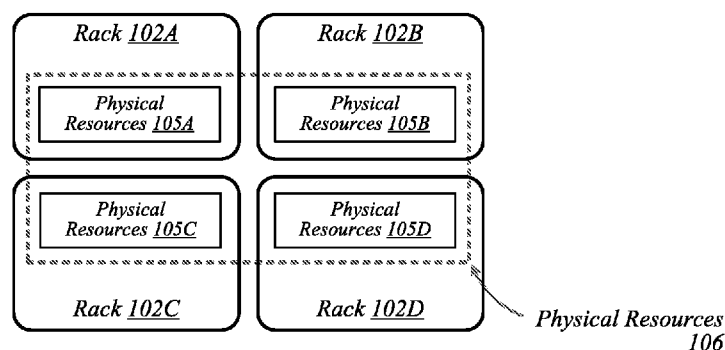


FIG. 1

(57) Abstract: Technologies for enhanced memory wear leveling is disclosed. In the illustrative embodiment, a storage controller on a storage sled performs wear leveling across several storage devices. For example, the storage controller may copy hot data from one storage device that has a high number of erasures to another storage device that has a lower number of erasures in order to make the number of erasures between the devices more even by accumulating further erasures associated with the hot data on the drive that has the lower number of erasures.

TECHNOLOGIES FOR ENHANCED MEMORY WEAR LEVELING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims priority to U.S. Utility Patent Application Serial No. 15/396,063, entitled “TECHNOLOGIES FOR ENHANCED MEMORY WEAR LEVELING,” 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] Solid state drives (SSDs) are data storage devices that rely on memory integrated circuits to store data in a non-volatile or persistent manner. Unlike hard disk drives, solid state drives do not include moving, mechanical parts, such as a movable drive head and/or drive spindle. A typical solid state drive includes a large amount of non-volatile memory, which is oftentimes based on NAND flash memory technology, although NOR flash memory and/or other types of non-volatile memory may be used in some implementations. The majority of data stored on a solid state drive is stored in the non-volatile memory for long-term storage. Flash memory-based solid state drives offer several advantages over traditional magnetic hard drives, but also provide new challenges. In order for data in a NAND flash memory cell to be overwritten, an entire block of data must be erased. Additionally, each block can only be erased a relatively small number of times, such as 1,000-10,000 times, before being rendered unusable.

[0003] In order to address the challenges of flash memory-based solid state drives, techniques such as wear leveling have been developed. A controller in a solid state drive may control where and how data is written and updated in order to spread erasures of the data blocks evenly throughout the drive in order to extend the lifetime of the drive.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] 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.

[0005] 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;

[0006] FIG. 2 is a diagram of an example embodiment of a logical configuration of a rack of the data center of FIG. 1;

[0007] 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;

[0008] 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;

[0009] 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;

[0010] 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;

[0011] FIG. 7 is a diagram of an example embodiment of a sled that may be used with the rack architecture of FIG. 6;

[0012] FIG. 8 is a diagram of an example embodiment of a rack architecture to provide support for sleds featuring expansion capabilities;

[0013] FIG. 9 is a diagram of an example embodiment of a rack implemented according to the rack architecture of FIG. 8;

[0014] FIG. 10 is a diagram of an example embodiment of a sled designed for use in conjunction with the rack of FIG. 9;

[0015] 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;

[0016] FIG. 12 is a simplified block diagram of at least one embodiment of a storage sled of the data center of FIG. 1;

[0017] FIG. 13 is a top perspective view of an example embodiment of a storage sled of FIG. 12;

[0018] FIG. 14 is a bottom perspective view of an example embodiment of a storage sled of FIG. 12;

[0019] FIG. 15 is an environment that may be established by the storage sled of FIG. 13;

[0020] FIG. 16 is at least one embodiment of a flowchart of a method for storing data that may be executed by the storage sled of FIG. 12; and

[0021] FIG. 17 is at least one embodiment of a flowchart of a method for performing wear leveling on data storage that may be executed by the storage sled of FIG. 12.

DETAILED DESCRIPTION OF THE DRAWINGS

[0022] 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.

[0023] 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); (B and C); (A 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); (B and C); (A and C); or (A, B, and C).

[0024] 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 one or more 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).

[0025] 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.

[0026] 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.

[0027] 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.

[0028] 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.

[0029] 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.

[0030] 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 205-4 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.

[0031] 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.

[0032] 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.

[0033] 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.

[0034] 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.

[0035] 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.

[0036] 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.

[0037] 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.

[0038] 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.).

[0039] 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.

[0040] 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.

[0041] 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.

[0042] 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.

[0043] 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.

[0044] 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.

[0045] 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.

[0046] 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.

[0047] 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.

[0048] 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.

[0049] Referring now to FIGS. 12-17, as discussed above, one or more of the sleds 204, 404, 504, 704, 1004 of the data center 100, 300, 400 may be embodied as a storage sled for

storing data. An illustrative storage sled 1200 usable in the data center 100, 300, 400 is shown in FIG. 12. During operation, the storage sled 1200 may receive data for storage on a data storage 1208 local to the storage sled 1200. In the illustrative embodiment, the storage sled 1200 may perform an enhanced memory wear leveling procedure by performing wear leveling across all of the storage devices 1212 that make up the data storage 1208, instead of performing wear leveling across each storage device 1212 individually. For example, the storage sled 1200 may determine that storage device 1212-1 has a higher number of erasures than storage device 1212-2. The storage sled 1200 may then identify hot data that is stored in the storage device 1212-1 and move that hot data to the storage device 1212-2. As a result, future erasures associated with the hot data will be done to the storage device 1212-2 with the fewer number of erasures. Similarly, cold data from storage device 1212-2 may be moved to the storage device 1212-1 so that the data stored on the storage device 1212-1 is associated with a lower frequency of erasures. It should be appreciated that, as used herein, “hot data” refers to data that is updated or overwritten relatively frequently and “cold data” refers to data that is updated or overwritten relatively infrequently. Similarly, as used herein, the “temperature” of a data refers to a relative frequency of how over the data is updated or overwritten (e.g., “hot data” is updated frequently, “warm data” less frequently, etc.).

[0050] Referring specifically now to FIG. 12, an illustrative storage sled 1200 includes a processor 1202, memory 1204, an input/output (I/O) subsystem 1206, the data storage 1208, and a communication circuit 1210. In some embodiments, one or more of the illustrative components of the storage sled 1200 may be incorporated in, or otherwise form a portion of, another component. For example, the memory 1204, or portions thereof, may be incorporated in the processor 1202 in some embodiments.

[0051] The processor 1202 may be embodied as any type of processor capable of performing the functions described herein. For example, the processor 1202 may be embodied as a single or multi-core processor(s), a single or multi-socket processor, a digital signal processor, a graphics processor, a microcontroller, or other processor or processing/controlling circuit. Similarly, the memory 1204 may be embodied as any type of volatile or non-volatile memory or data storage capable of performing the functions described herein. In operation, the memory 1204 may store various data and software used during operation of the storage sled 1200 such as operating systems, applications, programs, libraries, and drivers. The memory 1204 is communicatively coupled to the processor 1202 via the I/O subsystem 1206, which may be embodied as circuitry and/or components to facilitate input/output operations with the processor 1202, the memory 1204, and other components of the storage sled 1200. For

example, the I/O subsystem 1206 may be embodied as, or otherwise include, memory controller hubs, input/output control hubs, firmware devices, communication links (i.e., 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.

[0052] The data storage 1208 may be embodied as any type of device or collection of devices configured for the short-term or long-term storage of data. For example, the data storage 1208 may include any one or more memory devices and circuits, memory cards, solid-state drives, or other data storage devices. In the illustrative embodiment, the data storage 1208 is embodied as several discrete storage devices 1212 (such as storage device 1212-1, storage device 1212-2, storage device 1212-3, etc.). The storage devices 1212 include memory 1214, which may be embodied as any type of storage device, such as a NOR-based or a NAND-based flash storage device. In the illustrative embodiment, the storage devices 1212 are embodied as solid state drives having as memory 1214 NAND-based flash storage located therein. Additionally or alternatively, the memory 1214 may be embodied as any type of data storage capable of storing data in a persistent manner (even if power is interrupted to the non-volatile memory). For example, the memory 1214 may be embodied as any combination of memory devices that use chalcogenide phase change material (e.g., chalcogenide glass), 3-dimensional (3D) cross point memory, 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.

[0053] Each storage device 1212 may include a local controller 1216 to manage the memory 1214 of the corresponding storage device 1212. The local controller 1216 may perform functionality such as wear leveling and garbage collection over the corresponding memory 1214. To do so, the local controller 1216 may store metadata including relevant information such as a number of erasures of each block of the memory 1214 and an indication of the temperature of the data, such as an indication of which data is hot data and which data is cold data. For example, the local controller 1216 may keep track of how frequently data is overwritten or updated. In one illustrative embodiment, the local controller 1216 may determine that data stored in a block is hot data because a relatively large portion of the block has been overwritten in a certain period of time (i.e., a relatively large portion of the block has been marked invalid and moved to a new location with an updated value). Similarly, the illustrative local controller 1216 may determine that data stored in a block is cold data because

a relatively small portion of the block has been overwritten in a certain period of time (i.e., a relatively small portion of the block has been marked invalid and moved to a new location with an updated value). The local controller 1216 may make the information relating to number of erasures and the temperature of the data available to the rest of the storage sled 1200 upon request. In some embodiments, one or more storage devices 1212 may not include a local controller 1216 as part of the storage device 1212, and the corresponding memory 1214 may be managed by the storage sled 1200 (e.g., by the processor 1202 of the sled 1200). In such embodiments, the storage sled 1200 may perform wear leveling and garbage collection over each storage device 1212. Each of the illustrative storage devices 1212 is independently removable from the storage sled 1200. For example, if one storage device 1212 fails, that storage device 1212 may be easily removed and replaced with another storage device 1212. In the illustrative embodiment, the storage devices 1212 are hot swappable (i.e., a storage device 1212 can be removed and replaced without powering down or otherwise interrupting the functioning of the rest of the storage sled 1200).

[0054] Each illustrative storage device 1212 is arranged into several blocks, with each block including several pages, and each page including several cells. Each cell physically stores one or more bits (e.g., 3 bits). Each page may be any appropriate size, such as 512, 1,024, 2,048, 4,096, or 8,192 bits. Similarly, each block may be any appropriate size, such as 16, 32, 64, 128, or 256 pages. The illustrative storage device 1212 can read a single page at a time, write a single page at a time, and erase a single block at a time. In some embodiments, the storage device 1212 may group several blocks together as a single reclaim unit for erasures, such as 2, 4, 6, or 8 blocks. It should be appreciated that, unless explicitly noted otherwise, as used herein, the term “block” may refer to either the smallest storage unit of the storage device 1212 that can be erased, or may refer to the reclaim unit which groups two or more physical blocks together for erasures and wear leveling purposes. However, the illustrative storage device 1212 cannot overwrite a page without first erasing that page (and, therefore, erasing the entire block containing that page). Blocks of the illustrative storage device 1212 can only be erased a limited number of times before failing, such as 1,000, 2,000, 5,000, 10,000, 20,000, 50,000, or 100,000 times.

[0055] The communication circuit 1210 may be embodied as any type of communication circuit, device, or collection thereof, capable of enabling communications between the storage sled 1200 and other devices. To do so, the communication circuit 1210 may be configured to use any one or more communication technology and associated protocols (e.g., Ethernet, Bluetooth®, Wi-Fi®, WiMAX, near field communication (NFC), etc.) to effect

such communication. In the illustrative embodiment, the communication circuit 1210 includes an optical communicator capable of sending and receiving at a high rate, such as a rate of 20, 25, 50, 100, or 200 gigabits per second (Gbps).

[0056] It should be appreciated that, in the illustrative embodiment, the data center (e.g., the data center 100, 300, 400) may include additional sleds, such as accelerator sleds 205-2, memory sleds 205-3, compute sleds 205-4, etc. Each of the various sleds may be configured to be optimized for performing particular tasks, such as compute tasks, memory storage tasks, data storage tasks, etc. For example, a compute sled 205-4 may be configured to be optimized for performing compute tasks, and may include several high-speed processors and large amounts of high-speed memory with little or no data storage, and the storage sled 1200 is configured to be optimized for performing storage tasks, and may include a large amount of data storage 1208 with relatively slow processors 1202 as compared to the processors and data storage of the compute sled 205-4.

[0057] It should be appreciated that the embodiments of the storage sled 1200 described in FIG. 12 are not limiting. For example, in some embodiments, the storage sled 1200 may be embodied as a sled 704 as shown in FIG. 7, a sled 1004 as shown in FIG. 10, or any combination of the sleds 704, 1004, and 1200. Of course, any embodiment of the storage sled 1200 will include the resources necessary (such as the storage devices 1212) to perform the particular task required for a particular embodiment.

[0058] Referring now to FIG. 13, a top perspective view of an illustrative storage sled 1200 is shown. As illustrated, the storage sled 1200 includes a top side 1302. The illustrative storage sled 1200 includes two processors 1202 and a communication circuit 1210 positioned on the top side 1402. The storage sled 1200 further includes a storage cage 1304 positioned at one end of the storage sled 1200 that includes several storage slots 1306 for mounting the physical data storage 1208. In some examples, the illustrative storage sled 1200 shown in FIG. 13 may include sixteen storage devices 1212 (i.e., solid state drives) mounted to storage slots 1306 in the storage cage 1304.

[0059] Referring now to FIG. 14, a bottom perspective view of the illustrative storage sled 1200 is shown. As illustrated, the storage sled 1200 also includes a bottom side 1402. The storage sled 1200 includes memory 1204 positioned within slots 1404 on the bottom side 1402. In some examples, the memory 1204 may include multiple dual in-line memory modules (DIMMs).

[0060] Referring now to FIG. 15, in use, the storage sled 1200 may establish an environment 1500. The illustrative environment 1500 includes a storage controller 1502 and a

communication engine 1504. The various components of the environment 1500 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 1500 may be embodied as circuitry or collection of electrical devices (e.g., a storage controller circuit 1502, a communication engine 1504, etc.). It should be appreciated that, in such embodiments, the storage controller circuit 1502, the communication engine 1504, etc., may form a portion of one or more of the processor 1202, the memory 1204, the I/O subsystem 1206, the data storage 1208, communication circuit 1210, and/or other components of the storage sled 1200. For example, in an illustrative embodiment, the storage controller 1502 is embodied as, or forms a portion of, one or more processors 1202. Additionally, in some embodiments, one or more of the illustrative components may form a portion of another component and/or one or more of the illustrative components may be independent of one another. Further, in some embodiments, one or more of the components of the environment 1500 may be embodied as virtualized hardware components or emulated architecture, which may be established and maintained by the processor 1202 or other components of the storage sled 1200. Additionally, in the illustrative embodiment, the environment 1500 includes storage metadata 1506 which may be embodied as any data which includes metadata for the data stored on the storage devices 1212 as well as metadata relating to the storage devices 1212 themselves. For example, the storage metadata 1506 may include an amount of free space of each storage device 1212, an indication of the temperature of the data stored on each block of each storage device 1212, and information indicative of the number of times each block of each storage device 1212 has been erased.

[0061] The storage controller 1502 is configured to manage any requests for storage or retrieval of data received by the communication circuit 1210. The storage controller 1502 includes a data storer 1508 and a wear leveler 1510. In the illustrative embodiment, the storage controller 1502 may pass any request for retrieval of data or updating or overwriting of data to the appropriate storage device 1212, but, for any new data to be written to the data storage 1208, the storage controller 1502 may determine which storage device 1212 to which the data should be written. The storage controller 1502 may, for example, select the storage device 1212 based on the number of erasures of each storage device 1212 and/or an amount of free space of each storage device 1212.

[0062] The wear leveler 1510 is configured to manage the storage devices 1212 to ensure a desired wear leveling of the storage devices 1212, such as by working to make the number of erasures of each block of the storage device 1212 approximately equal. For

example, the wear leveler 1510 may move hot data from a storage device with a relatively large number of erasures to a storage device with a relatively small number of erasures and move cold data from the storage device with the relatively small number of erasures to the storage device with the relatively large number of erasures. Since the hot data is expected to be associated with more frequent updates and erasures, swapping the hot and cold data would be expected to lead to fewer erasures on the storage device with the relatively large number of erasures and to more erasures on the storage device with the relatively small number of erasures. In the illustrative embodiment, the wear leveler 1510 will only perform wear leveling across storage devices 1212 present on the same storage sled 1200. Additionally or alternatively, the wear leveler 1510 may perform wear leveling across storage devices 1212 present on two or more storage sleds 1200. For example, the wear leveler 1510 may move hot and cold data as described above across different storage sleds 1200 in the data center 100 in order to ensure a desired wear leveling across each storage device 1212 of each storage sled 1200. The wear leveler 1510 may be run periodically, continuously, continually, or when a certain condition is met. For example, the wear leveler 1510 may only be run when the difference in the amount of free storage between two storage devices 1212 reaches a certain threshold or when the difference in the number of erasures between two storage devices reaches a certain threshold.

[0063] The communication engine 1502 is configured to send and receive data using the communication circuit 1210. The communication engine 1502 may use any appropriate protocol to send and receive data.

[0064] Referring now to FIG. 16, in use, the storage sled 1200 may execute a method 1600 for storing data on the storage sled 1200. The method 1600 begins in block 1602, in which the storage sled receives data to be stored. Subsequently, in block 1604, the storage sled 1200 accesses storage metadata including wear leveling information and the amount of free space in each storage device 1212. In block 1606, the storage sled 1200 accesses an indication of a number of erasures for each free block of each storage device 1212.

[0065] In block 1608, the storage sled 1200 selects a storage device 1212 at which to store the data based on the wear leveling information and the amount of free storage on the storage devices 1212. In some embodiments, the storage sled 1200 may select the storage device 1212 having the lowest number of erasures as compared to other storage devices 1212 in block 1610. In block 1612, the storage sled 1200 may store the data in the selected storage device 1212.

[0066] Referring now to FIG. 17, in use, the storage sled 1200 may execute a method 1700 for performing wear-leveling on the storage sled 1200. The method 1700 begins in block 1702, in which the storage sled 1200 determines whether to perform wear leveling. As discussed above, the storage sled 1200 may perform wear leveling periodically, continually, continuously, or when a certain condition is met, such as when the difference in the amount of free storage between two storage devices 1212 reaches a certain threshold or when the difference in the number of erasures between two storage devices reaches a certain threshold.

[0067] In block 1704, if the storage sled 1200 is to perform wear leveling, the method 1700 proceeds to block 1706. Otherwise, if the storage sled 1200 is not to perform wear leveling, the method 1700 loops back to block 1702 in which the storage sled 1200 again determines whether to perform wear leveling. In block 1706, the storage sled 1200 accesses the storage metadata including wear leveling information and the amount of free storage of each storage device 1212. The storage metadata may be stored on the storage sled 1200 separate from the storage devices 1212, or the storage metadata for each storage device 1212 may reside on the corresponding storage device 1212, and the storage sled may access the storage metadata by accessing the storage metadata on each storage device 1212. The storage sled 1200 accesses an indication of a number of erasures for each free block of each storage device 1212 on the storage sled 1200 in block 1708, determines an amount of free space for each storage device 1212 in block 1710, and determines hot and/or cold data for each storage device 1212 in block 1712.

[0068] In block 1714, the storage sled 1200 selects one or more blocks to be moved. For example, the storage sled 1200 may select one or more blocks from a storage device 1212 with a relative low amount of free space. In block 1716, the storage sled 1200 may select hot data from a storage sled 1200 to be moved, such as from a storage sled with a relatively high number of erasures. In block 1718, the storage sled 1200 may select cold data from a storage sled 1200 to be moved, such as from a storage sled with a relatively low number of erasures.

[0069] In block 1720, the storage sled 1200 moves selected data from the corresponding storage device 1212 to a different storage device 1212. The different storage device 1212 may be chosen based on the number of erasures of each storage device 1212, such as by choosing the storage device 1212 with the lowest number of erasures (for hot data) or the highest number of erasures (for cold data). In the illustrative embodiment, the different storage device 1212 chosen is a storage device 1212 on the same storage sled 1200. In some embodiments, the different storage device 1212 chosen may be on a different storage sled 1200. In block 1722, the storage sled 1200 may swap data between the two selected storage devices. For example,

the storage sled 1200 may move cold data from a storage sled 1200 with a relatively low number of erasures to a storage sled 1200 with a relatively high number of erasures and move hot data from the storage sled 1200 with the relatively high number of erasures to the storage sled 1200 with the relatively low number of erasures.

EXAMPLES

[0070] Illustrative examples of the devices, systems, and methods disclosed herein are provided below. An embodiment of the devices, systems, and methods may include any one or more, and any combination of, the examples described below.

[0071] Example 1 includes a storage sled for enhanced wear leveling for flash storage, the storage sled comprising a plurality of storage devices, wherein each storage device of the plurality of storage devices comprises a plurality of blocks; a storage controller to access storage metadata of the plurality of storage devices of the storage sled, wherein the storage metadata comprises, for each storage device of the plurality of storage devices, an indication of a number of erasures of the corresponding storage device and a temperature of data in one or more blocks of one or more storage devices of the plurality of storage devices; select the data in the one or more blocks of the one or more storage devices based on the indication of the number of erasures of the corresponding one or more storage devices and the corresponding temperature of the data; and move the data from the selected one or more storage devices to a different storage device of the plurality of storage devices.

[0072] Example 2 includes the subject matter of Example 1, and wherein to select the data in the one or more blocks of the one or more storage devices comprises to select the one or more storage devices based with the highest number of erasures of the numbers of erasures of the plurality of storage devices.

[0073] Example 3 includes the subject matter of any of Examples 1 and 2, and wherein to select the data in the one or more blocks comprises to select the data based on the temperature of the data indicating a relatively high frequency of writing associated with the data.

[0074] Example 4 includes the subject matter of any of Examples 1-3, and wherein to select the data in the one or more blocks of the one or more storage devices comprises to select the one or more storage devices based with the lowest number of erasures of the numbers of erasures of the plurality of storage devices.

[0075] Example 5 includes the subject matter of any of Examples 1-4, and wherein to select the data in the one or more blocks comprises to select the data based on the temperature of the data indicating a relatively low frequency of writing associated with the data.

[0076] Example 6 includes the subject matter of any of Examples 1-5, and wherein to select the data in the one or more blocks of the one or more storage devices comprises to select a first storage device based on the first storage device having a relatively low number of erasures of the numbers of erasures of the plurality of storage devices; select a second storage device based on the second storage device having a relatively high number of erasures of the numbers of erasures of the plurality of storage devices; select cold data from the first storage device; and select hot data from the second storage device, and wherein to move the data from the selected one or more storage devices to the different storage device of the plurality of storage devices comprises to move the cold data to the second storage device and the hot data to the first storage device.

[0077] Example 7 includes the subject matter of any of Examples 1-6, and wherein to access the storage metadata of the plurality of storage devices comprises to access, for each storage device of the plurality of storage devices, storage device metadata generated by and stored on the corresponding storage device.

[0078] Example 8 includes the subject matter of any of Examples 1-7, and wherein the storage metadata further comprises a temperature of additional data in one or more additional blocks of one or more additional storage devices of the plurality of storage devices, and wherein the storage controller is further to select the additional data in the one or more additional blocks of the one or more additional storage devices based on the indication of the number of erasures of the corresponding one or more additional storage devices and the corresponding temperature of the additional data; and move the additional data from the selected one or more additional storage devices to a storage device on a different storage sled.

[0079] Example 9 includes the subject matter of any of Examples 1-8, and wherein the storage sled comprises a storage cage, wherein the storage cage is configured to establish a plurality of storage slots, and wherein each storage device of the plurality of storage devices is in a storage slot of the plurality of storage slots.

[0080] Example 10 includes the subject matter of any of Examples 1-9, and wherein each storage device of the plurality of storage devices comprises a NAND flash storage device.

[0081] Example 11 includes a method for enhanced wear leveling for flash storage on a storage sled, the method comprising accessing, by the storage sled, storage metadata of a plurality of storage devices of the storage sled, wherein the storage metadata comprises, for

each storage device of the plurality of storage devices, an indication of a number of erasures of the corresponding storage device and a temperature of data in one or more blocks of one or more storage devices of the plurality of storage devices; selecting, by the storage sled, the data in the one or more blocks of the one or more storage devices based on the indication of the number of erasures of the corresponding one or more storage devices and the corresponding temperature of the data; and moving, by the storage sled, the data from the selected one or more storage devices to a different storage device of the plurality of storage devices.

[0082] Example 12 includes the subject matter of Example 11, and wherein selecting the data in the one or more blocks of the one or more storage devices comprises selecting the one or more storage devices based with the highest number of erasures of the numbers of erasures of the plurality of storage devices.

[0083] Example 13 includes the subject matter of any of Examples 11 and 12, and wherein selecting the data in the one or more blocks comprises selecting the data based on the temperature of the data indicating a relatively high frequency of writing associated with the data.

[0084] Example 14 includes the subject matter of any of Examples 11-13, and wherein selecting the data in the one or more blocks of the one or more storage devices comprises selecting the one or more storage devices based with the lowest number of erasures of the numbers of erasures of the plurality of storage devices.

[0085] Example 15 includes the subject matter of any of Examples 11-14, and wherein selecting the data in the one or more blocks comprises selecting the data based on the temperature of the data indicating a relatively low frequency of writing associated with the data.

[0086] Example 16 includes the subject matter of any of Examples 11-15, and wherein selecting the data in the one or more blocks of the one or more storage devices comprises selecting a first storage device based on the first storage device having a relatively low number of erasures of the numbers of erasures of the plurality of storage devices; selecting a second storage device based on the second storage device having a relatively high number of erasures of the numbers of erasures of the plurality of storage devices; selecting cold data from the first storage device; and selecting hot data from the second storage device, and wherein moving the data from the selected one or more storage devices to the different storage device of the plurality of storage devices comprises moving the cold data to the second storage device and the hot data to the first storage device.

[0087] Example 17 includes the subject matter of any of Examples 11-16, and wherein accessing the storage metadata of the plurality of storage devices comprises accessing, for each

storage device of the plurality of storage devices, storage device metadata generated by and stored on the corresponding storage device.

[0088] Example 18 includes the subject matter of any of Examples 11-17, and wherein the storage metadata further comprises a temperature of additional data in one or more additional blocks of one or more additional storage devices of the plurality of storage devices, and further comprising selecting, by the storage sled, the additional data in the one or more additional blocks of the one or more additional storage devices based on the indication of the number of erasures of the corresponding one or more additional storage devices and the corresponding temperature of the additional data; and moving, by the storage sled, the additional data from the selected one or more additional storage devices to a storage device on a different storage sled.

[0089] Example 19 includes the subject matter of any of Examples 11-18, and wherein the storage sled comprises a storage cage, wherein the storage cage is configured to establish a plurality of storage slots, and wherein each storage device of the plurality of storage devices is in a storage slot of the plurality of storage slots.

[0090] Example 20 includes the subject matter of any of Examples 11-19, and wherein each storage device of the plurality of storage devices comprises a NAND flash storage device.

[0091] Example 21 includes one or more computer-readable media comprising a plurality of instructions stored thereon that, when executed, causes a storage sled to perform the methods of any of Examples 11-20.

[0092] Example 22 includes a storage sled for enhanced wear leveling for flash storage, the storage sled comprising means for accessing storage metadata of a plurality of storage devices of the storage sled, wherein the storage metadata comprises, for each storage device of the plurality of storage devices, an indication of a number of erasures of the corresponding storage device and a temperature of data in one or more blocks of one or more storage devices of the plurality of storage devices; means for selecting the data in the one or more blocks of the one or more storage devices based on the indication of the number of erasures of the corresponding one or more storage devices and the corresponding temperature of the data; and means for moving the data from the selected one or more storage devices to a different storage device of the plurality of storage devices.

[0093] Example 23 includes the subject matter of Example 22, and wherein the means for selecting the data in the one or more blocks of the one or more storage devices comprises means for selecting the one or more storage devices based with the highest number of erasures of the numbers of erasures of the plurality of storage devices.

[0094] Example 24 includes the subject matter of any of Examples 22 and 23, and wherein the means for selecting the data in the one or more blocks comprises means for selecting the data based on the temperature of the data indicating a relatively high frequency of writing associated with the data.

[0095] Example 25 includes the subject matter of any of Examples 22-24, and wherein the means for selecting the data in the one or more blocks of the one or more storage devices comprises means for selecting the one or more storage devices based with the lowest number of erasures of the numbers of erasures of the plurality of storage devices.

[0096] Example 26 includes the subject matter of any of Examples 22-25, and wherein the means for selecting the data in the one or more blocks comprises means for selecting the data based on the temperature of the data indicating a relatively low frequency of writing associated with the data.

[0097] Example 27 includes the subject matter of any of Examples 22-26, and wherein the means for selecting the data in the one or more blocks of the one or more storage devices comprises means for selecting a first storage device based on the first storage device having a relatively low number of erasures of the numbers of erasures of the plurality of storage devices; means for selecting a second storage device based on the second storage device having a relatively high number of erasures of the numbers of erasures of the plurality of storage devices; means for selecting cold data from the first storage device; and means for selecting hot data from the second storage device, and wherein the means for moving the data from the selected one or more storage devices to the different storage device of the plurality of storage devices comprises means for moving the cold data to the second storage device and the hot data to the first storage device.

[0098] Example 28 includes the subject matter of any of Examples 22-27, and wherein the means for accessing the storage metadata of the plurality of storage devices comprises means for accessing, for each storage device of the plurality of storage devices, storage device metadata generated by and stored on the corresponding storage device.

[0099] Example 29 includes the subject matter of any of Examples 22-28, and wherein the storage metadata further comprises a temperature of additional data in one or more additional blocks of one or more additional storage devices of the plurality of storage devices, and further comprising means for selecting the additional data in the one or more additional blocks of the one or more additional storage devices based on the indication of the number of erasures of the corresponding one or more additional storage devices and the corresponding

temperature of the additional data; and means for moving the additional data from the selected one or more additional storage devices to a storage device on a different storage sled.

[00100] Example 30 includes the subject matter of any of Examples 22-29, and wherein the storage sled comprises a storage cage, wherein the storage cage is configured to establish a plurality of storage slots, and wherein each storage device of the plurality of storage devices is in a storage slot of the plurality of storage slots.

[00101] Example 31 includes the subject matter of any of Examples 22-30, and wherein each storage device of the plurality of storage devices comprises a NAND flash storage device.

WHAT IS CLAIMED IS:

1. A storage sled for enhanced wear leveling for non-volatile memory, the storage sled comprising:

a plurality of storage devices, wherein each storage device of the plurality of storage devices comprises a plurality of non-volatile memory blocks;

a storage controller to:

access storage metadata of the plurality of storage devices of the storage sled, wherein the storage metadata comprises, for each storage device of the plurality of storage devices, an indication of a number of erasures of the corresponding storage device and a temperature of data in one or more non-volatile memory blocks of one or more storage devices of the plurality of storage devices;

select the data in the one or more non-volatile memory blocks of the one or more storage devices based on the indication of the number of erasures of the corresponding one or more storage devices and the corresponding temperature of the data; and

move the data from the selected one or more storage devices to a different storage device of the plurality of storage devices.

2. The storage sled of claim 1, wherein to select the data in the one or more non-volatile memory blocks of the one or more storage devices comprises to select the one or more storage devices based with the highest number of erasures of the numbers of erasures of the plurality of storage devices.

3. The storage sled of claim 2, wherein to select the data in the one or more non-volatile memory blocks comprises to select the data based on the temperature of the data indicating a relatively high frequency of writing associated with the data.

4. The storage sled of claim 1, wherein to select the data in the one or more non-volatile memory blocks of the one or more storage devices comprises to select the one or more storage devices based with the lowest number of erasures of the numbers of erasures of the plurality of storage devices.

5. The storage sled of claim 4, wherein to select the data in the one or more non-volatile memory blocks comprises to select the data based on the temperature of the data indicating a relatively low frequency of writing associated with the data.

6. The storage sled of claim 1, wherein to select the data in the one or more non-volatile memory blocks of the one or more storage devices comprises to:

select a first storage device based on the first storage device having a relatively low number of erasures of the numbers of erasures of the plurality of storage devices;

select a second storage device based on the second storage device having a relatively high number of erasures of the numbers of erasures of the plurality of storage devices;

select cold data from the first storage device; and

select hot data from the second storage device, and

wherein to move the data from the selected one or more storage devices to the different storage device of the plurality of storage devices comprises to move the cold data to the second storage device and the hot data to the first storage device.

7. The storage sled of claim 1, wherein to access the storage metadata of the plurality of storage devices comprises to access, for each storage device of the plurality of storage devices, storage device metadata generated by and stored on the corresponding storage device.

8. The storage sled of claim 1, wherein the storage metadata further comprises a temperature of additional data in one or more additional non-volatile memory blocks of one or more additional storage devices of the plurality of storage devices, and wherein the storage controller is further to:

select the additional data in the one or more additional non-volatile memory blocks of the one or more additional storage devices based on the indication of the number of erasures of the corresponding one or more additional storage devices and the corresponding temperature of the additional data; and

move the additional data from the selected one or more additional storage devices to a storage device on a different storage sled.

9. The storage sled of claim 1, wherein the storage sled comprises a storage cage, wherein the storage cage is configured to establish a plurality of storage slots, and wherein each storage device of the plurality of storage devices is in a storage slot of the plurality of storage slots.

10. The storage sled of claim 1, wherein each storage device of the plurality of storage devices comprises a NAND flash storage device.

11. A method for enhanced wear leveling for flash storage on a storage sled, the method comprising:

accessing, by the storage sled, storage metadata of a plurality of storage devices of the storage sled, wherein the storage metadata comprises, for each storage device of the plurality of storage devices, an indication of a number of erasures of the corresponding storage device and a temperature of data in one or more non-volatile memory blocks of one or more storage devices of the plurality of storage devices;

selecting, by the storage sled, the data in the one or more non-volatile memory blocks of the one or more storage devices based on the indication of the number of erasures of the corresponding one or more storage devices and the corresponding temperature of the data; and

moving, by the storage sled, the data from the selected one or more storage devices to a different storage device of the plurality of storage devices.

12. The storage sled of claim 11, wherein selecting the data in the one or more non-volatile memory blocks of the one or more storage devices comprises selecting the one or more storage devices based with the highest number of erasures of the numbers of erasures of the plurality of storage devices.

13. The storage sled of claim 12, wherein selecting the data in the one or more non-volatile memory blocks comprises selecting the data based on the temperature of the data indicating a relatively high frequency of writing associated with the data.

14. The storage sled of claim 11, wherein selecting the data in the one or more non-volatile memory blocks of the one or more storage devices comprises selecting the one or more storage devices based with the lowest number of erasures of the numbers of erasures of the plurality of storage devices.

15. The storage sled of claim 14, wherein selecting the data in the one or more non-volatile memory blocks comprises selecting the data based on the temperature of the data indicating a relatively low frequency of writing associated with the data.

16. The storage sled of claim 11, wherein selecting the data in the one or more non-volatile memory blocks of the one or more storage devices comprises:

selecting a first storage device based on the first storage device having a relatively low number of erasures of the numbers of erasures of the plurality of storage devices;

selecting a second storage device based on the second storage device having a relatively high number of erasures of the numbers of erasures of the plurality of storage devices;

selecting cold data from the first storage device; and

selecting hot data from the second storage device, and

wherein moving the data from the selected one or more storage devices to the different storage device of the plurality of storage devices comprises moving the cold data to the second storage device and the hot data to the first storage device.

17. The storage sled of claim 11, wherein accessing the storage metadata of the plurality of storage devices comprises accessing, for each storage device of the plurality of storage devices, storage device metadata generated by and stored on the corresponding storage device.

18. The storage sled of claim 11, wherein the storage metadata further comprises a temperature of additional data in one or more additional non-volatile memory blocks of one or more additional storage devices of the plurality of storage devices, and further comprising:

selecting, by the storage sled, the additional data in the one or more additional non-volatile memory blocks of the one or more additional storage devices based on the indication of the number of erasures of the corresponding one or more additional storage devices and the corresponding temperature of the additional data; and

moving, by the storage sled, the additional data from the selected one or more additional storage devices to a storage device on a different storage sled.

19. The storage sled of claim 11, wherein the storage sled comprises a storage cage, wherein the storage cage is configured to establish a plurality of storage slots, and wherein each storage device of the plurality of storage devices is in a storage slot of the plurality of storage slots.

20. The storage sled of claim 11, wherein each storage device of the plurality of storage devices comprises a NAND flash storage device.

21. One or more computer-readable media comprising a plurality of instructions stored thereon that, when executed, causes a storage sled to perform the methods of any of claims 11-20.

22. A storage sled comprising means to perform the methods of any of claims 11-20.

FIG. 1

Data Center 100

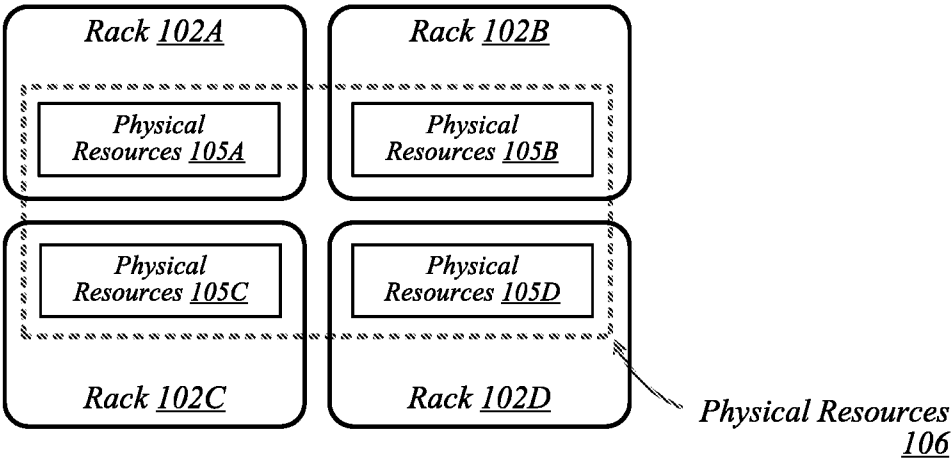


FIG. 2

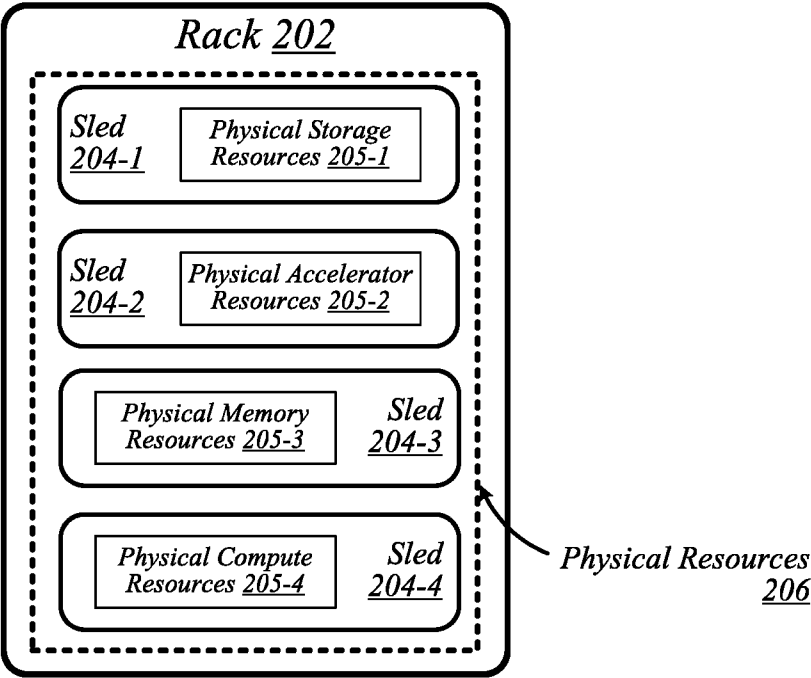


FIG. 3

Data Center 300

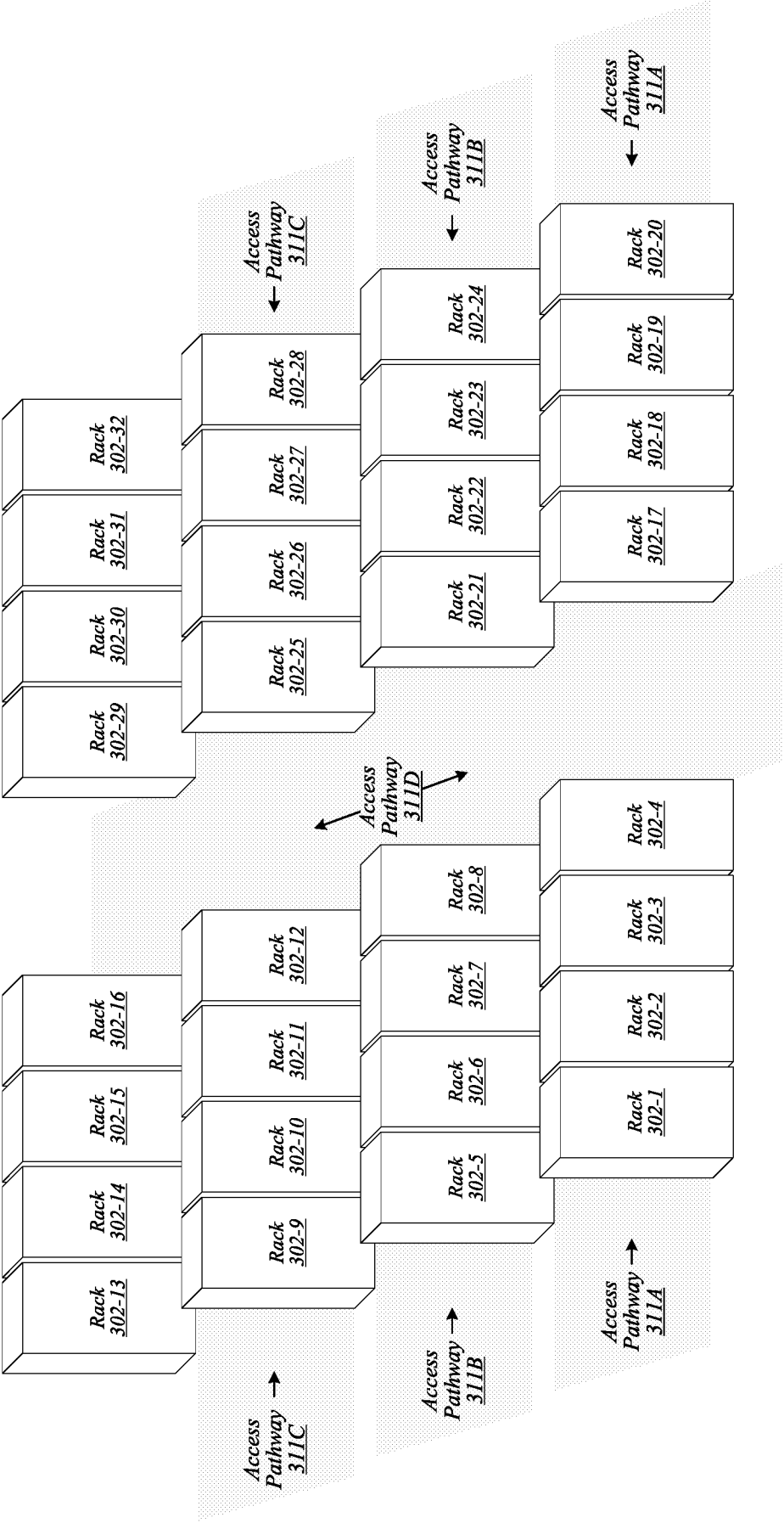


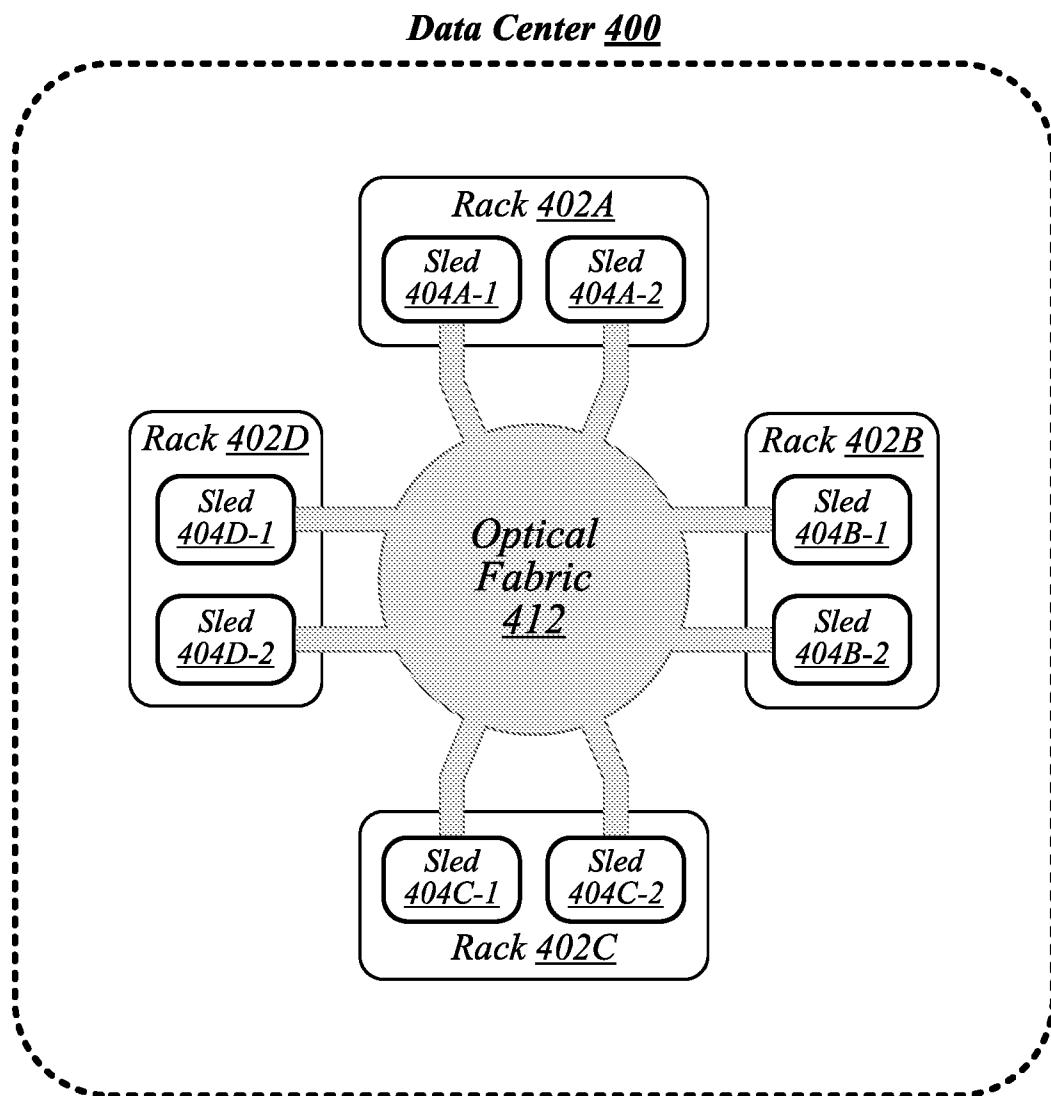
FIG. 4

FIG. 5

500

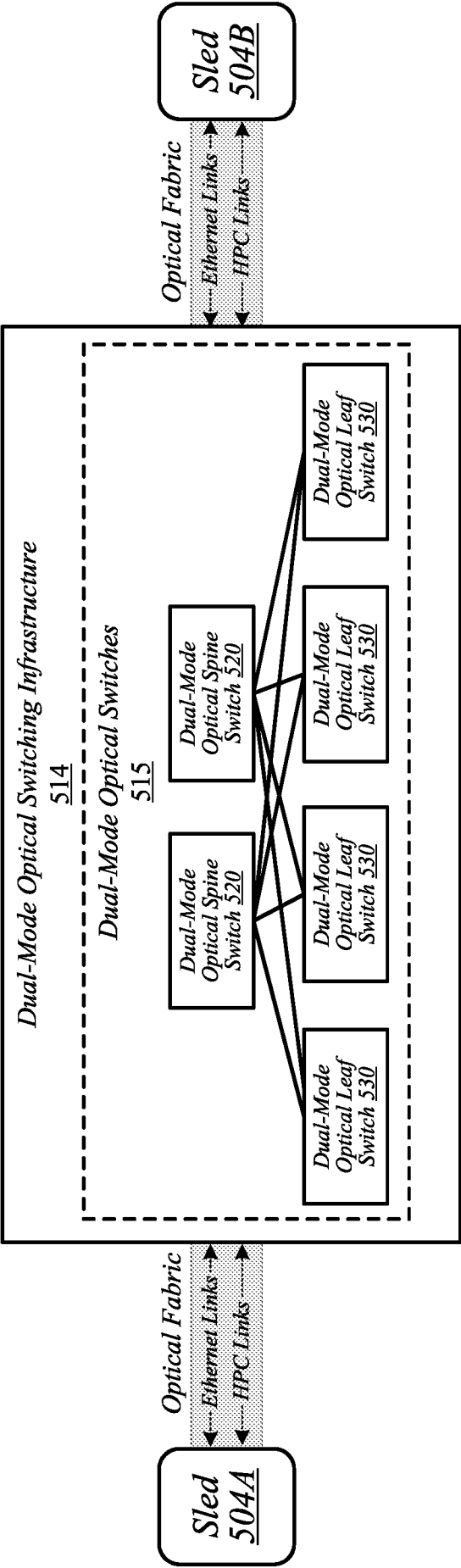


FIG. 6

600

Rack Access Region 601

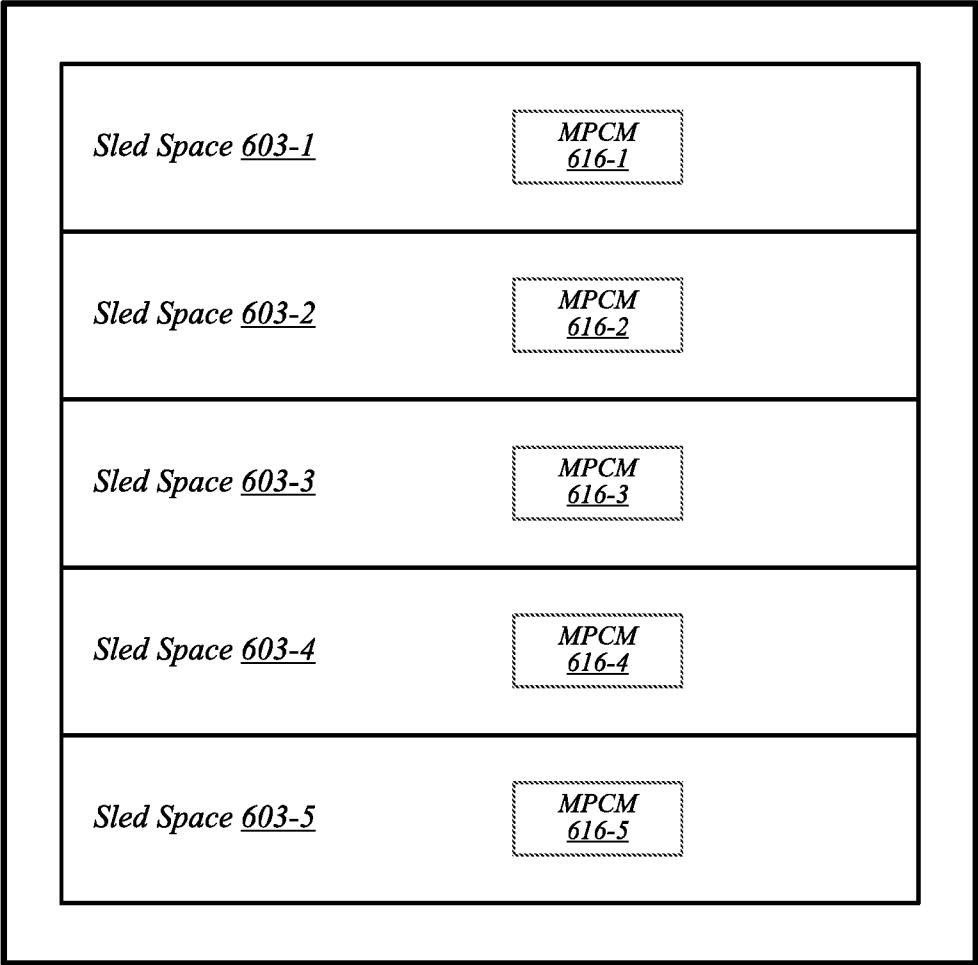


FIG. 7

Sled 704

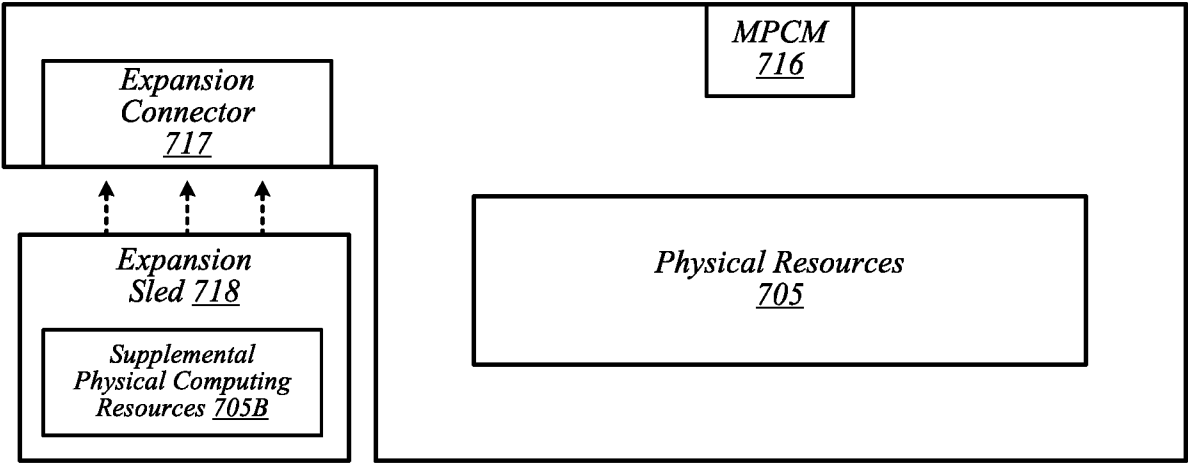


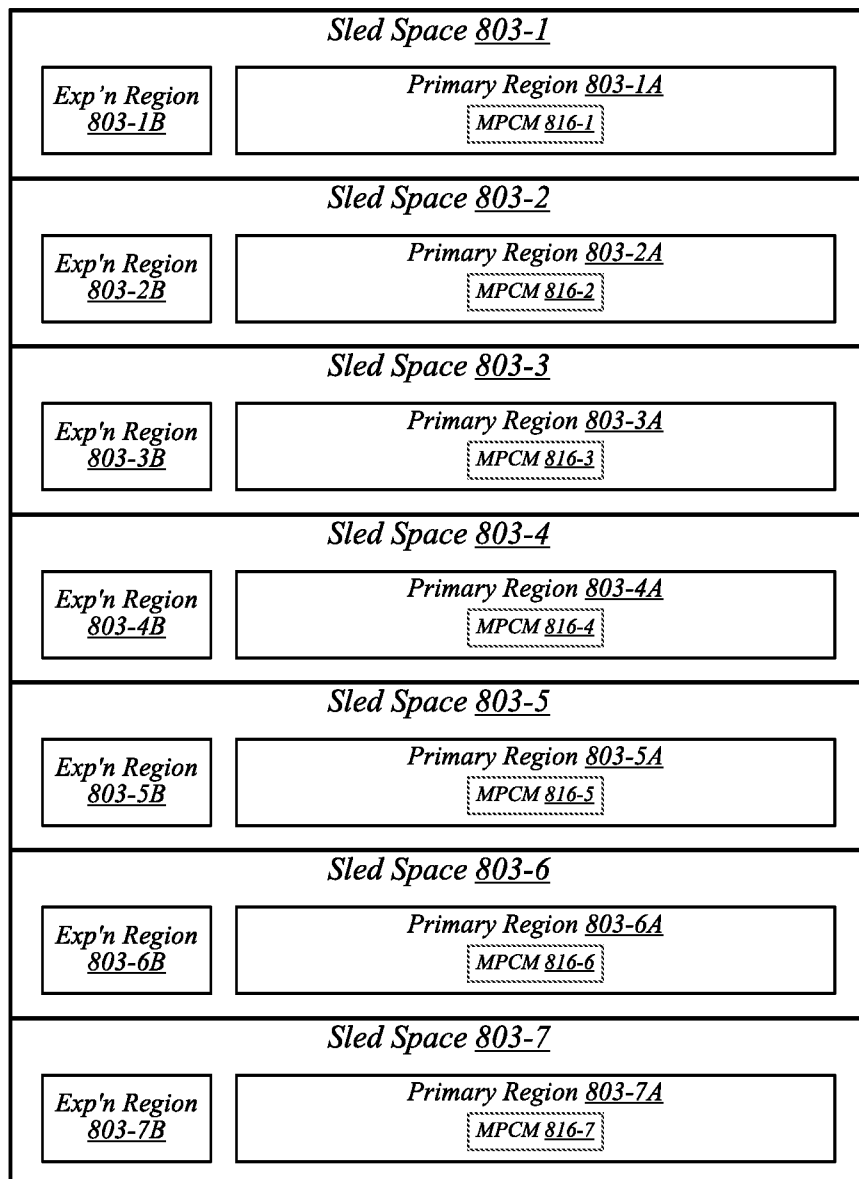
FIG. 8**800**

FIG. 9

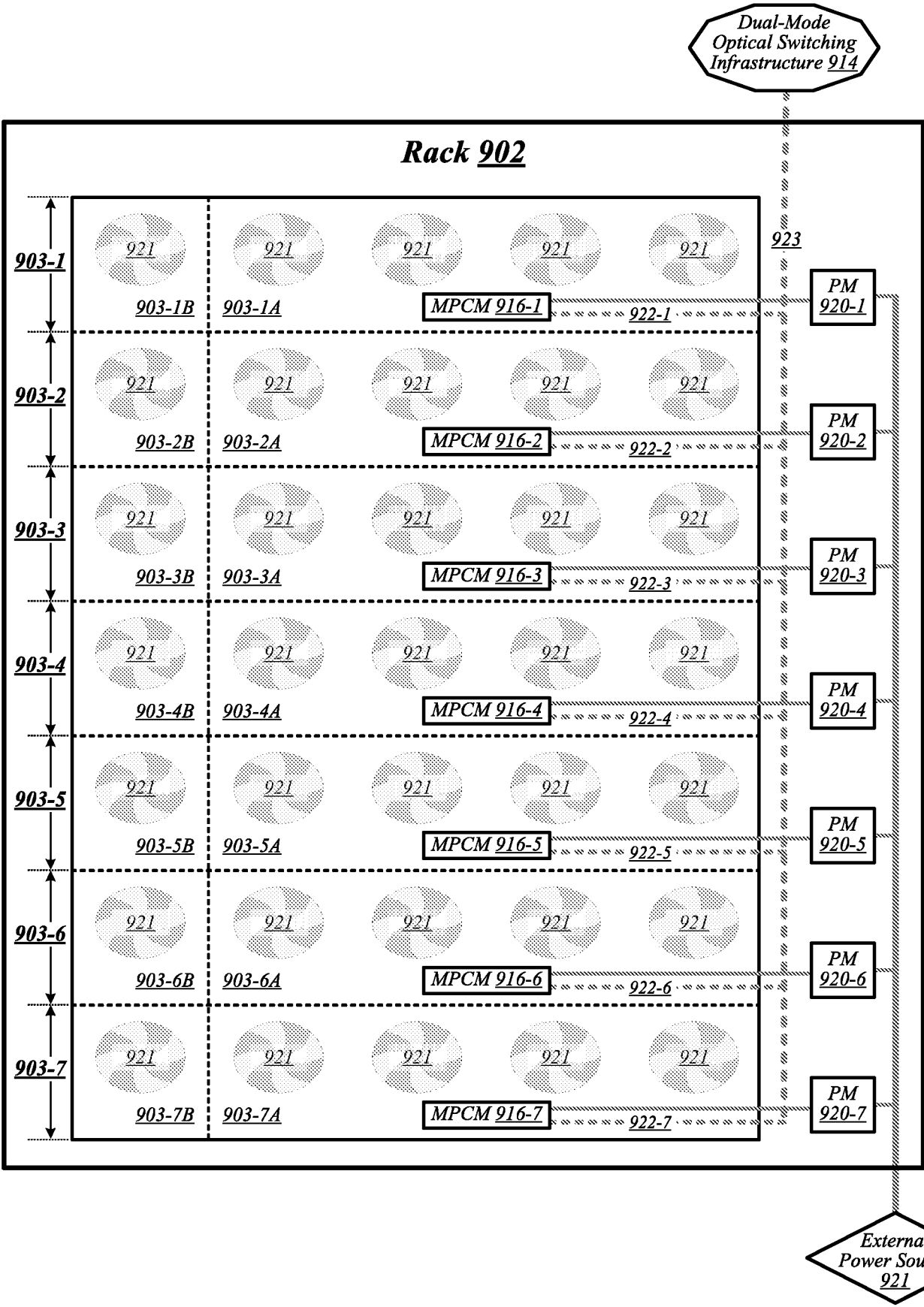


FIG. 10

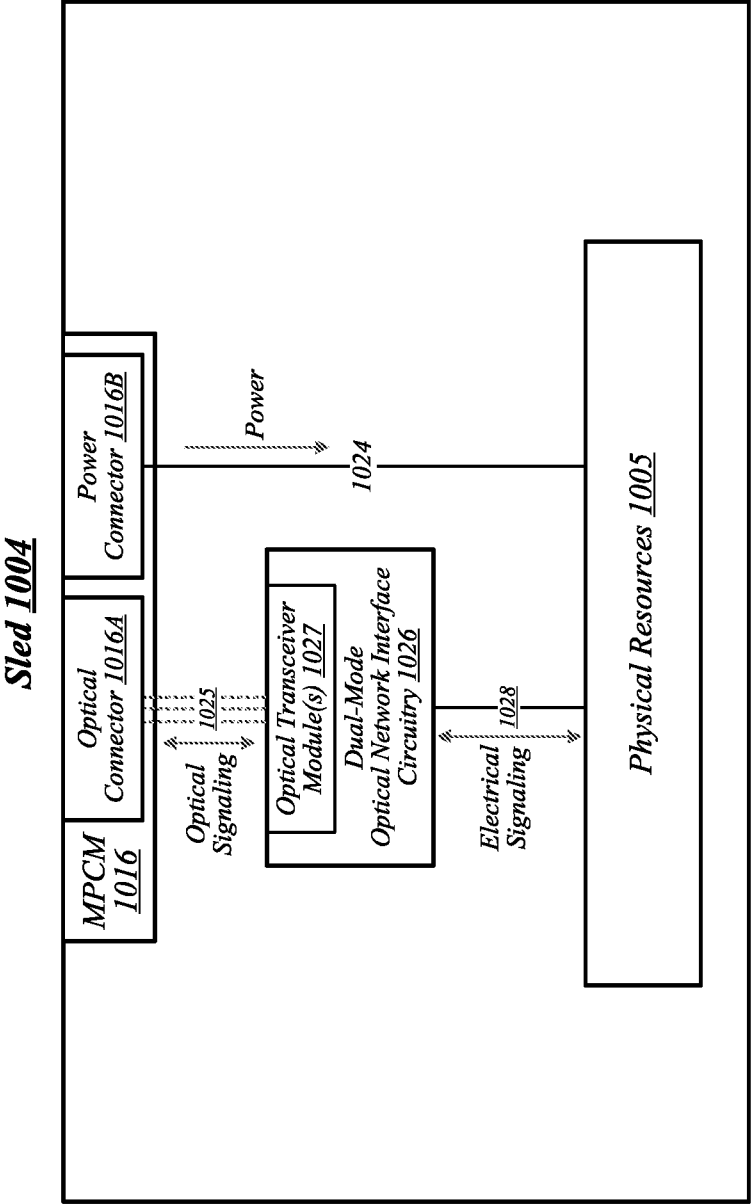
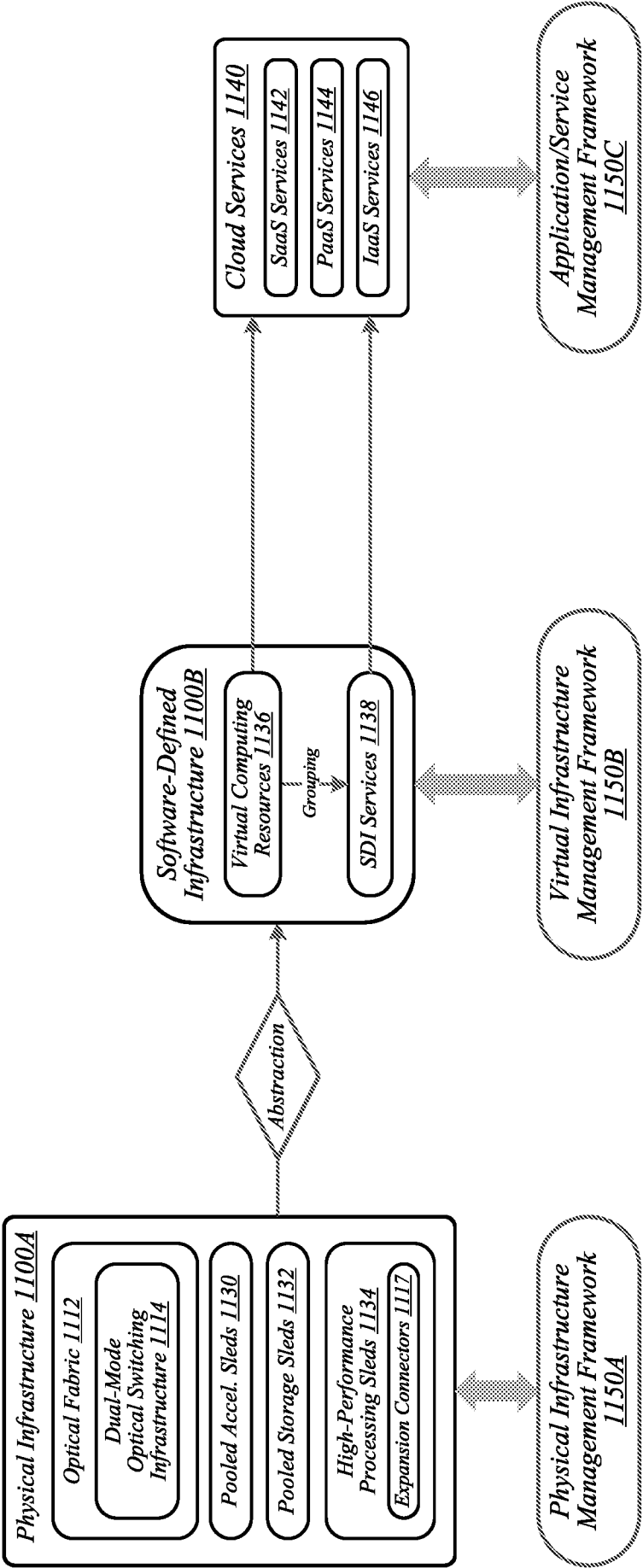


FIG. 11

Data Center 1100



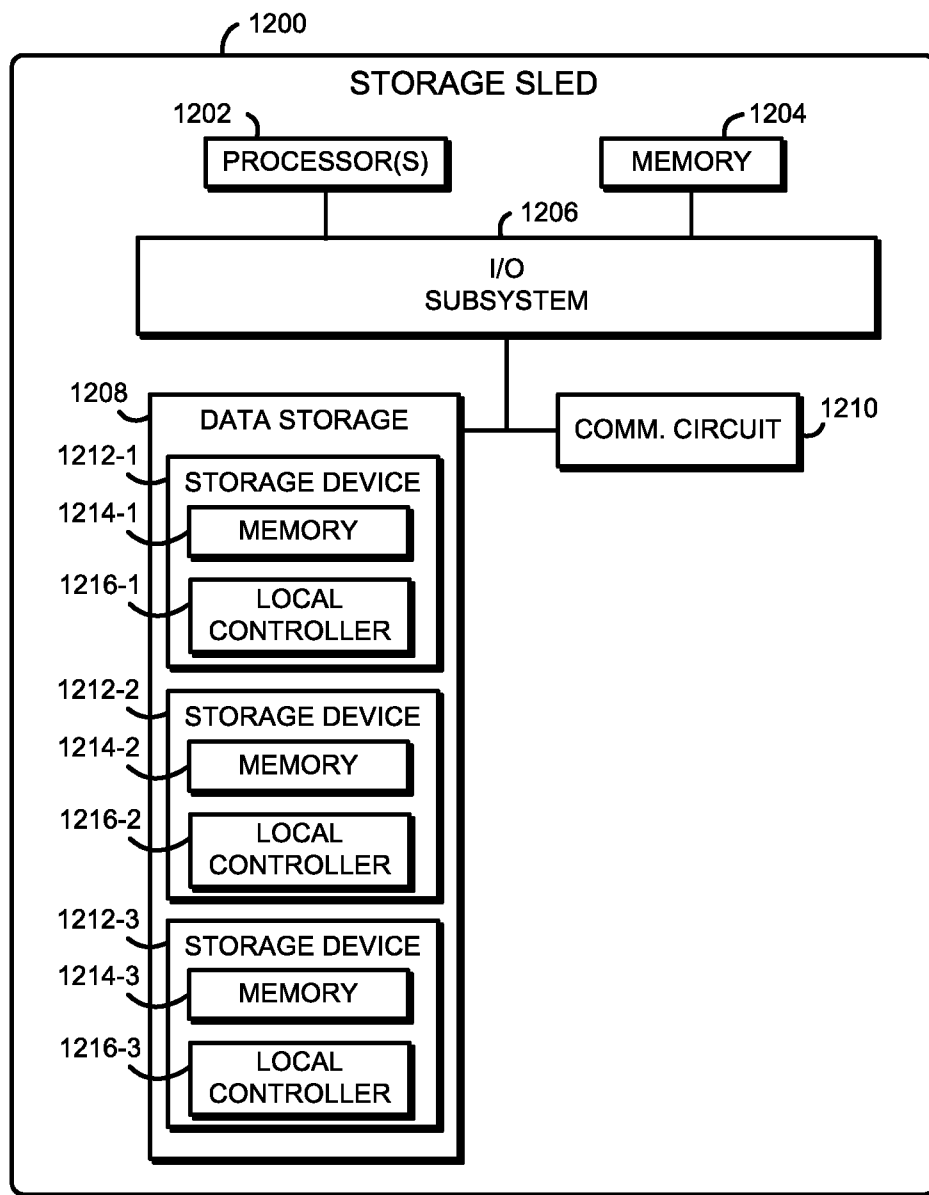


FIG. 12

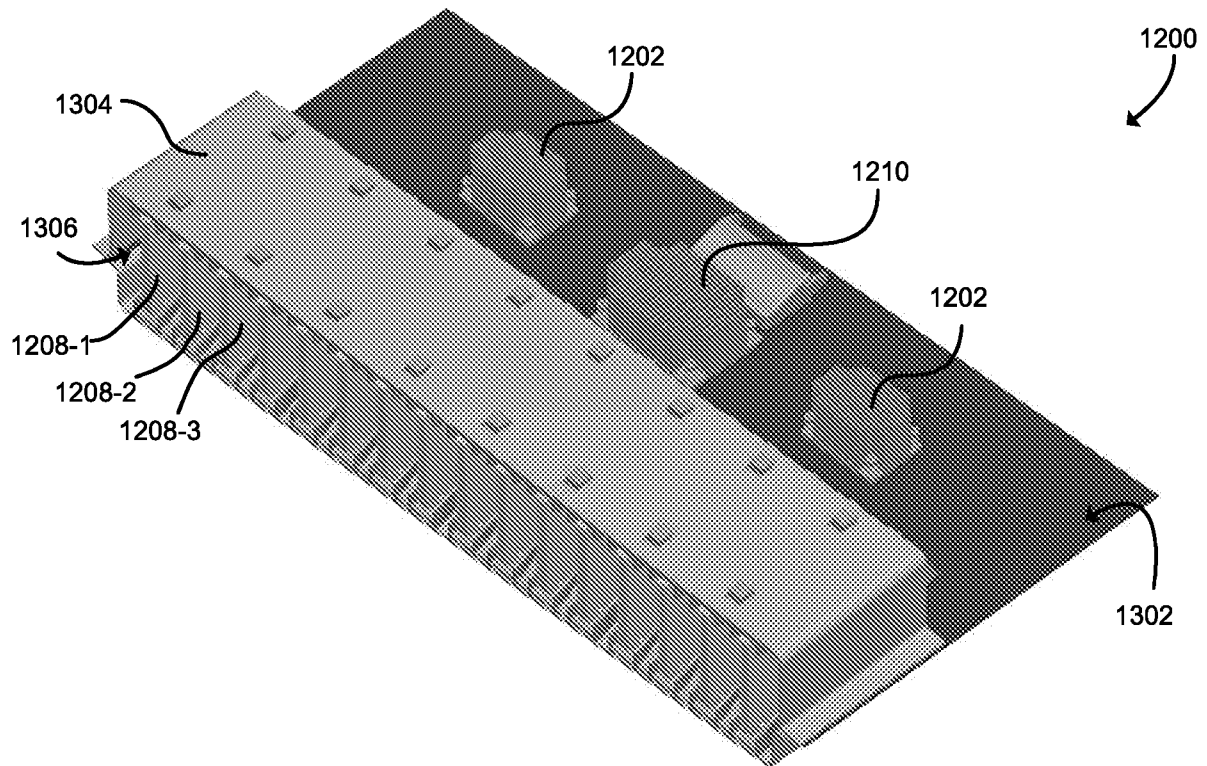


FIG. 13

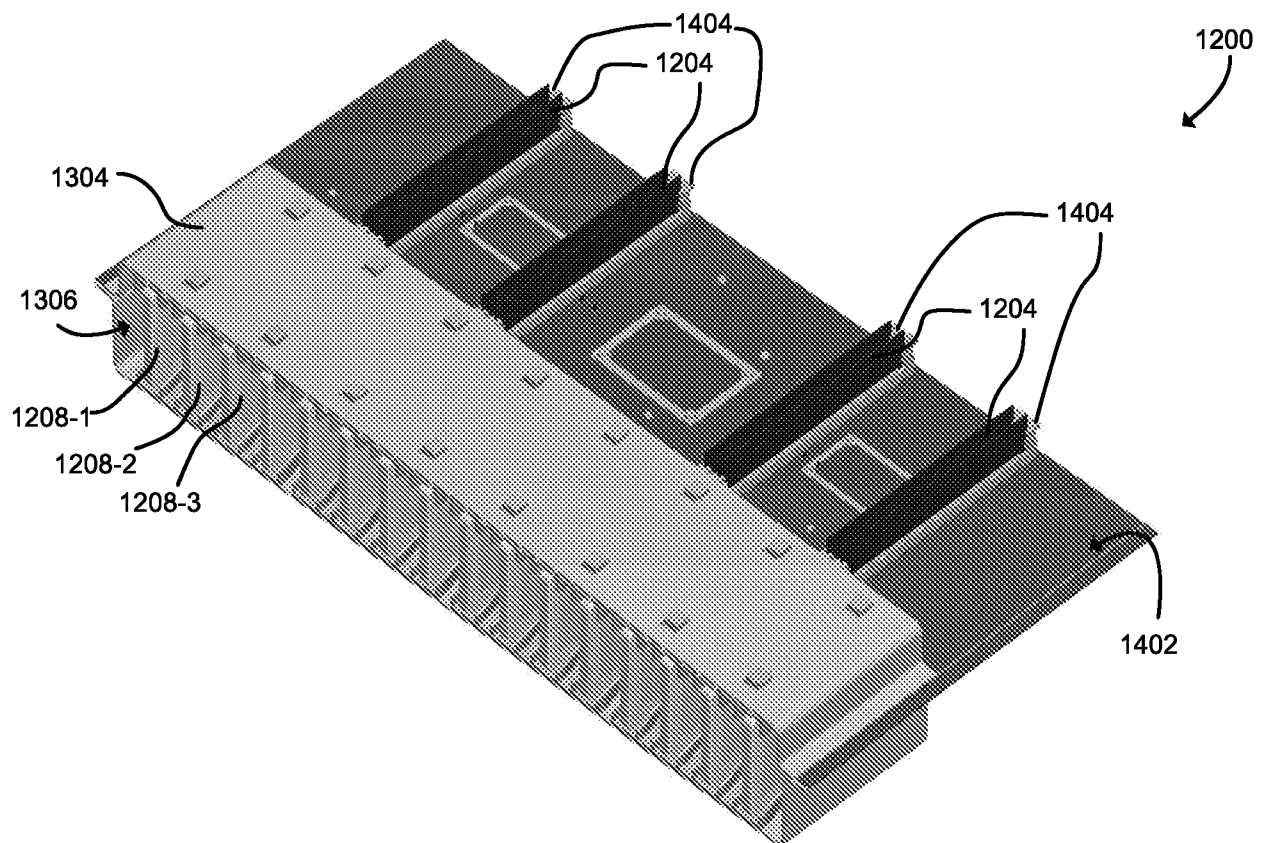


FIG. 14

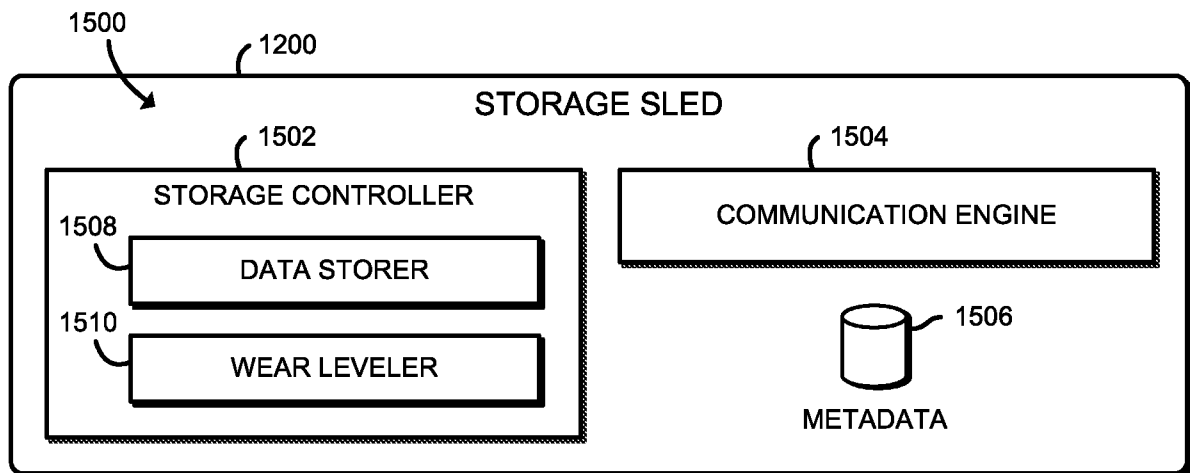


FIG. 15

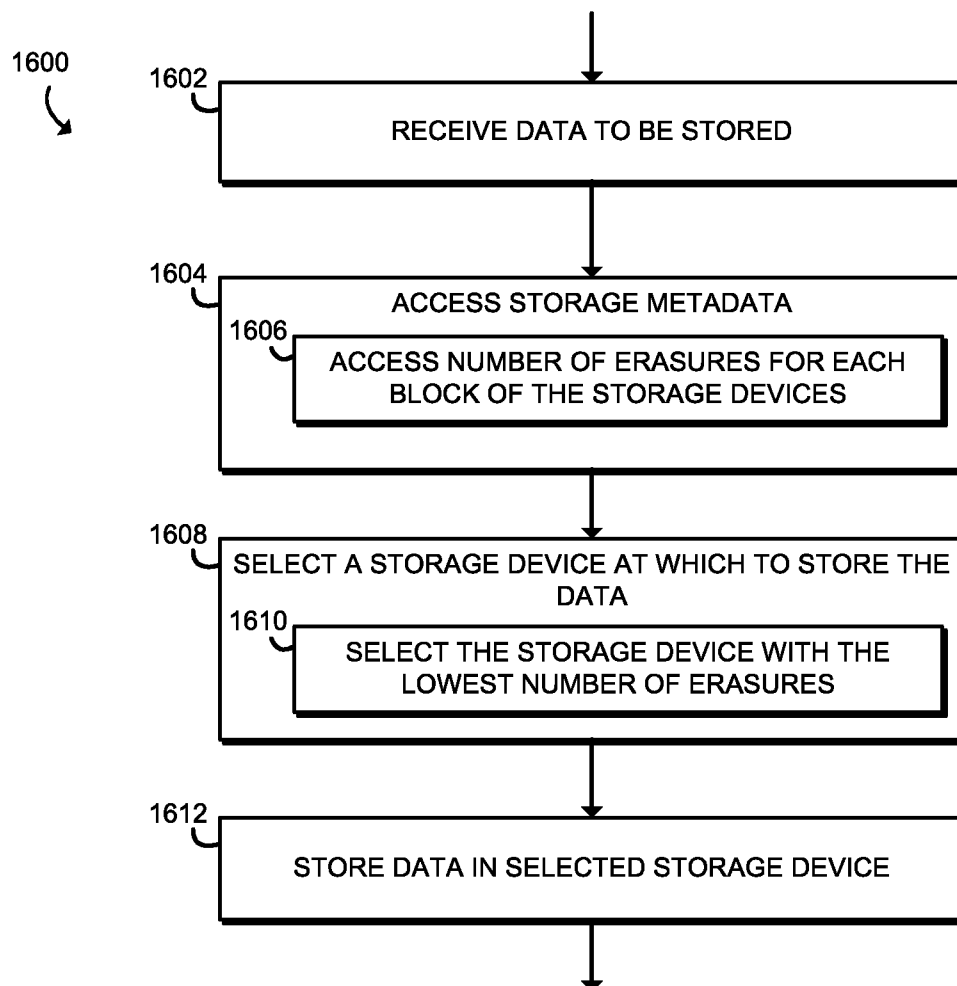


FIG. 16

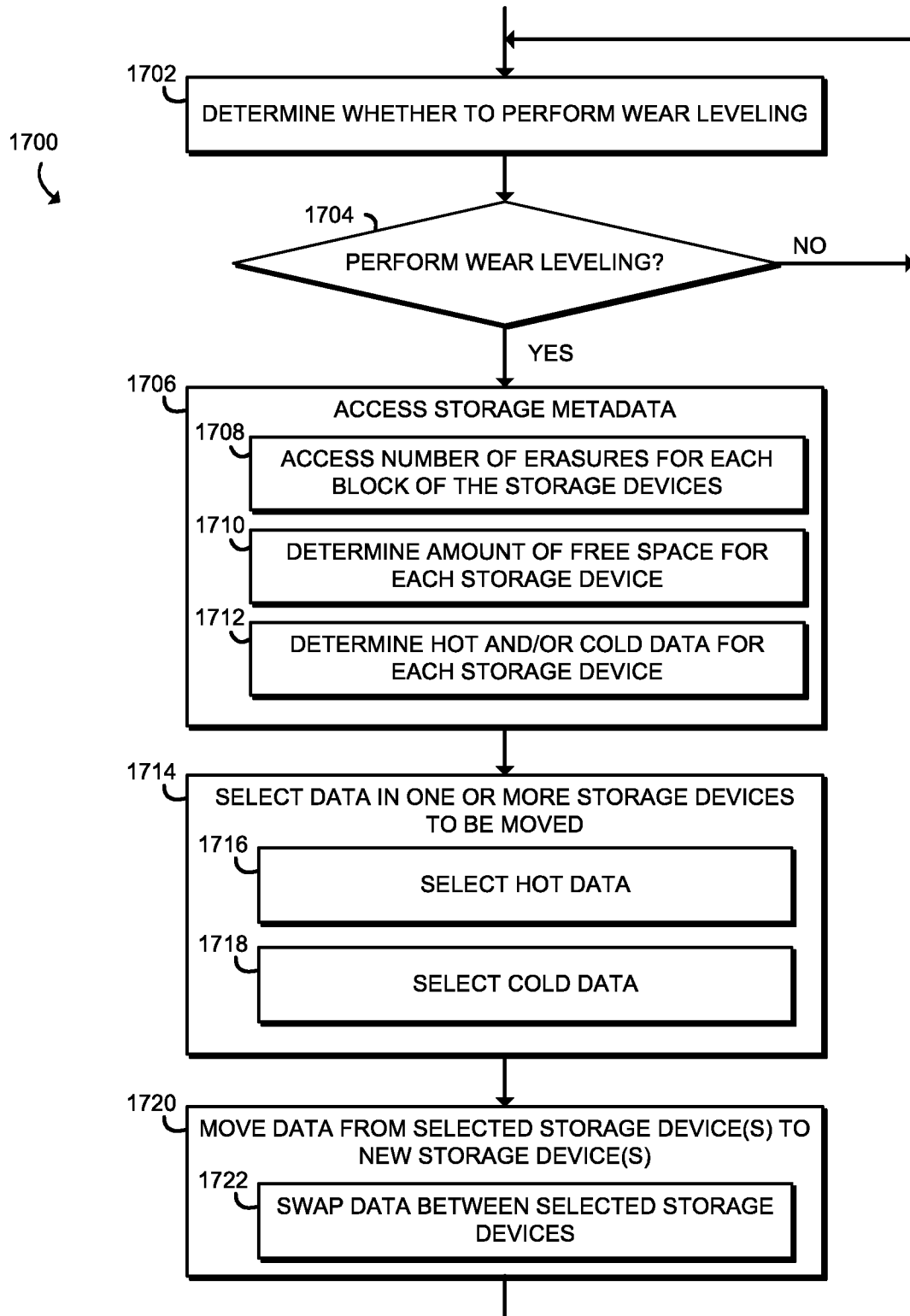


FIG. 17

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2017/038317**A. CLASSIFICATION OF SUBJECT MATTER****G06F 12/02(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 12/02; G06F 12/06; G06F 3/06; G06F 12/00; G11C 16/34

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: wear, leveling, select, erasure, number, temperature, slot, move, different

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2011-0264843 A1 (JONATHAN W. HAINES et al.) 27 October 2011 See paragraphs [0007], [0026], [0037], [0091]-[0092], [0101], [0105]-[0106]; and figure 16.	1-22
Y	US 2016-0062676 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 03 March 2016 See paragraphs [0016], [0021], [0028]; and figure 1.	1-22
A	US 2011-0320688 A1 (YANGSUP LEE) 29 December 2011 See paragraphs [0010], [0057]; and figure 1.	1-22
A	US 2012-0221784 A1 (AMIR BAN) 30 August 2012 See paragraphs [0077], [0147]; and figure 8.	1-22
A	WO 2012-051600 A2 (SON, KYQUANG et al.) 19 April 2012 See paragraphs [0013], [0045]; and figure 1.	1-22



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

16 October 2017 (16.10.2017)

Date of mailing of the international search report

17 October 2017 (17.10.2017)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

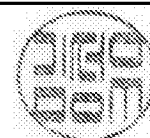
189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

KIM, Seong Woo

Telephone No. +82-42-481-3348



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/038317

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011-0264843 A1	27/10/2011	US 2016-0188226 A1 US 9183134 B2 WO 2011-133646 A1	30/06/2016 10/11/2015 27/10/2011
US 2016-0062676 A1	03/03/2016	US 9292210 B1	22/03/2016
US 2011-0320688 A1	29/12/2011	KR 10-2012-0001405 A	04/01/2012
US 2012-0221784 A1	30/08/2012	US 9021185 B2 WO 2011-061724 A1	28/04/2015 26/05/2011
WO 2012-051600 A2	19/04/2012	TW 201232262 A US 2012-0096217 A1 US 8738846 B2 WO 2012-051600 A3	01/08/2012 19/04/2012 27/05/2014 16/08/2012