



(51) International Patent Classification:

G06F 12/02 (2006.01) G06F 11/30 (2006.01)

(21) International Application Number:

PCT/RU2017/000273

(22) International Filing Date:

27 April 2017 (27.04.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: EMC IP HOLDING COMPANY LLC
[US/US]; 176 South Street, Hopkinton, MA 01748 (US).

(72) Inventor: DALMATOV, Nikolay Alexandrovich;
Slavy, 13, 108, St.Petersburg, 192239 (RU).

(74) Agent: NILOVA, Maria Innokentievna; PATENTICA,
Box 1125, St.Petersburg, 190000 (RU).

(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, 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).

Published:

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

(54) Title: RE-PLACING DATA WITHIN A MAPPED-RAID ENVIRONMENT

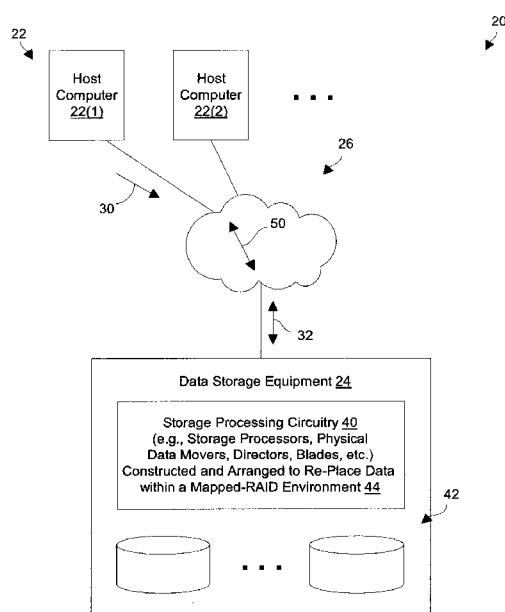


FIG. 1

(57) Abstract: A technique balances data storage activity within a mapped-RAID environment. The technique involves selecting, by processing circuitry, a source slice of storage from multiple slices of storage of the mapped-RAID environment, the source slice containing particular data to be relocated. The technique further involves selecting, by the processing circuitry, a destination slice of storage from the multiple slices of storage of the mapped-RAID environment. The technique further involves relocating, by the processing circuitry, the particular data from the source slice to the destination slice to balance data storage activity within the mapped-RAID environment. The mapped-RAID environment includes multiple storage devices. Each storage device provides multiple non-overlapping device extents. Each slice of the multiple slices of storage of the mapped-RAID environment is formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment.



Re-Placing Data within a Mapped-RAID Environment

BACKGROUND

A conventional data storage system presents a mapped logical unit of storage (LUN) for block-level access. The mapped LUN includes slices of storage that are formed from physical drives configured as a redundant array of independent disk (RAID) group. For example, in accordance with RAID Level 5, each block-level data stripe (multiple data blocks and a parity block) extends across all of the physical drives of the RAID group (one block being placed on each physical drive), and these data stripes form slices of storage of the LUN.

During operation, the data storage system may move slices of data between RAID groups of different storage tiers. For example, the data storage system may promote a frequently accessed slice of data to a first RAID group in a solid state memory storage tier (i.e., a fast tier) from a second RAID group in a magnetic memory storage tier (i.e., a slow tier) to reduce access latency for that slice of data. Additionally, the data storage system may demote a slice of data that is infrequently accessed from first RAID group in the solid state memory storage tier to the second RAID group in the magnetic memory storage tier to free up storage space within the solid state memory storage tier.

SUMMARY

Unfortunately, there are deficiencies to the above-described conventional data storage system in which each block-level data stripe extends across all of the physical drives of the RAID group. For example, in response to failure of a physical drive, such a data storage system requires excessive access to a new physical drive during data restoration. Such access results in relatively long data restoration time thus leaving the data storage system vulnerable to data loss if second physical drive fails before data restoration completes.

However, suppose that the data is stored in a mapped-RAID environment instead of the above-described conventional data storage system in which each block-level data stripe extends across all of the physical drives of the RAID group. Although data restoration may complete faster in the mapped-RAID environment since the restored data may be written across more than one existing storage device rather than a single new storage device, the mapped-RAID

environment may provide uneven input/output (I/O) loading and/or wearing of storage devices. That is, the mapped-RAID environment may form imbalances (e.g., hot spots) since each block-level data stripe does not extend across all of the physical drives of the RAID group. If there is no compensation provided for these imbalances, latency to overloaded storage devices that form hot spots may be sub-optimal and these overloaded storage devices may fail prematurely.

In contrast to the above-described conventional data storage system in which each block-level data stripe extends across all of the physical drives of the RAID group, improved techniques are directed to re-placing (or relocating) data within a mapped-RAID environment to provide improved I/O load and wear balancing among storage devices. In particular, such techniques identify imbalances (e.g., heavily accessed slices of data) and move the imbalances among the storage devices of the mapped-RAID environment (e.g., redistribute the heavily accessed slices of data). Such operation statistically balances the storage devices in terms of I/O performance and wearing thus reducing overall latency and reducing failures due to premature wear.

One embodiment is directed to a method of balancing data storage activity within a mapped-RAID environment. The method includes selecting, by processing circuitry, a source slice of storage from multiple slices of storage of the mapped-RAID environment, the source slice containing particular data to be relocated. The method further includes selecting, by the processing circuitry, a destination slice of storage from the multiple slices of storage of the mapped-RAID environment. The method further includes relocating, by the processing circuitry, the particular data from the source slice to the destination slice to balance data storage activity within the mapped-RAID environment. The mapped-RAID environment includes multiple storage devices. Each storage device provides multiple non-overlapping device extents. Each slice of the multiple slices of storage of the mapped-RAID environment is formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment.

In some arrangements, the method further includes, prior to selecting the source slice and the destination slice, performing a set of input/output (I/O) operations that accesses the particular data while the particular data resides in the source slice. In these arrangements, the method further includes updating, in response to performing the set of I/O operations, a set of statistics that tracks data storage activity of the mapped-RAID environment.

In some arrangements, selecting the source slice of storage from the multiple slices of

storage of the mapped-RAID environment includes identifying the source slice among other slices of the mapped-RAID environment based on the set of statistics that tracks data storage activity of the mapped-RAID environment indicating that the source slice has experienced higher data storage activity than the other slices of the mapped-RAID environment.

5 In some arrangements, selecting the destination slice of storage from the multiple slices of storage of the mapped-RAID environment includes identifying the destination slice among the other slices of the mapped-RAID environment based on the set of statistics that tracks data storage activity of the mapped-RAID environment indicating that the destination slice has experienced less data storage activity than the other slices of the mapped-RAID environment.

10 In some arrangements, the method further includes, while performing the set of I/O operations that accesses the particular data while the particular data resides in the source slice, maintaining an initial set of address mappings that map a portion of a logical unit of storage (LUN), which is addressable by a set applications, to the source slice.

15 In some arrangements, the method further includes, upon relocating the particular data from the source slice to the destination slice, maintaining a subsequent set of address mappings that map the portion of the LUN to the destination slice.

20 In some arrangements, the method further includes, after relocating the particular data from the source slice to the destination slice, performing another set of I/O operations that accesses the particular data while the particular data resides in the destination slice. In these arrangements, the method further includes updating the set of statistics that tracks data storage activity of the mapped-RAID environment in response to performing the other set of I/O operations.

25 Another embodiment is directed to data storage equipment which includes a communications interface, memory, and control circuitry coupled to the communications interface and the memory. The memory includes a mapped-RAID environment having multiple storage devices, each storage device providing multiple non-overlapping device extents. The memory stores instructions which, when carried out by the control circuitry, cause the control circuitry to:

- 30 (A) select a source slice of storage from multiple slices of storage of the mapped-RAID environment, the source slice containing particular data to be relocated,

- (B) select a destination slice of storage from the multiple slices of storage of the mapped-RAID environment, and
- (C) relocate the particular data from the source slice to the destination slice to balance data storage activity within the mapped-RAID environment, each slice of the multiple slices of storage of the mapped-RAID environment being formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment.

Yet another embodiment is directed to a computer program product having a non-transitory computer readable medium which stores a set of instructions to balance data storage activity within a mapped-RAID environment. The set of instructions, when carried out by computerized circuitry, causing the computerized circuitry to perform a method of:

- (A) selecting a source slice of storage from multiple slices of storage of the mapped-RAID environment, the source slice containing particular data to be relocated;
- (B) selecting a destination slice of storage from the multiple slices of storage of the mapped-RAID environment; and
- (C) relocating the particular data from the source slice to the destination slice to balance data storage activity within the mapped-RAID environment, the mapped-RAID environment including multiple storage devices, each storage device providing multiple non-overlapping device extents, each slice of the multiple slices of storage of the mapped-RAID environment being formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment.

It should be understood that, in the cloud context, at least some of electronic circuitry is formed by remote computer resources distributed over a network. Such an electronic environment is capable of providing certain advantages such as high availability and data protection, transparent operation and enhanced security, big data analysis, etc.

Other embodiments are directed to electronic systems and apparatus, processing circuits, computer program products, and so on. Some embodiments are directed to various methods,

electronic components and circuitry which are involved in re-placing data within a mapped-RAID environment.

BRIEF DESCRIPTION OF THE DRAWINGS

5 The foregoing and other objects, features and advantages will be apparent from the following description of particular embodiments of the present disclosure, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating the principles of various embodiments of the present disclosure.

10 Fig. 1 is a block diagram of a data storage setting which re-replaces data within a mapped-RAID environment.

 Fig. 2 is a block diagram of data storage equipment of the data storage setting of Fig. 1.

 Fig. 3 is a block diagram illustrating particular storage device details of the mapped-RAID environment.

15 Fig. 4 is a block diagram illustrating particular storage RAID group and tier details of the mapped-RAID environment.

 Fig. 5 is a flowchart of a procedure which is performed by the data storage equipment of Fig. 2.

20 DETAILED DESCRIPTION

 An improved technique is directed to relocating (or moving) data within a mapped-RAID environment to provide improved input/output (I/O) performance and wear balancing among storage devices. Along these lines, such a technique identifies imbalances (e.g., heavily accessed slices of data) and moves that data among the storage devices (e.g., redistributes the heavily
25 accessed slices of data to less frequently accessed locations). Such operation statistically balances the storage devices in terms of I/O loading and wearing thus reducing overall latency and reducing failures due to premature wear.

 Fig. 1 is a block diagram of a data storage setting (or environment) 20 which relocates (or re-replaces) data within a mapped-RAID environment. The data storage setting 20 includes host
30 computers 22(1), 22(2), ... (collectively, host computers 22), data storage equipment 24, and a communications medium 26.

Each host computer 22 is constructed and arranged to perform useful work. For example, a host computer 22 may operate as a web server, a file server, an email server, an enterprise server, combinations thereof, and so on, which provides I/O requests 30 (e.g., small computer system interface or SCSI commands) to the data storage equipment 24. Such I/O requests 30
5 direct the data storage equipment 24 to store and retrieve host data 32 on behalf of the host computers 22.

The data storage equipment 24 includes storage processing circuitry 40 and an array of storage devices 42. The storage processing circuitry 40 may include one or more physical storage processors or engines, data movers, director boards, blades, I/O modules, storage drive
10 controllers, switches, other hardware, combinations thereof, and so on. The storage processing circuitry 40 is constructed and arranged to provide a variety of data storage system services. For example, the storage processing circuitry 40 processes the I/O requests 30 from the host computers 22 by robustly and reliably storing host data 32 (and associated metadata) in a mapped-RAID environment 44 provided by the array of storage devices 42 and retrieving the host
15 data 32 from the mapped-RAID environment 44. Additionally, the storage processing circuitry 40 moves data between different slices of the mapped-RAID environment 44 to provide even I/O loading and wear. Accordingly, I/O latency and failures due to premature wearing are reduced.

The communications medium 26 is constructed and arranged to connect the various components of the data storage setting 20 together to enable these components to communicate
20 via exchanging electronic signals 50 (e.g., see the double arrow 50). At least a portion of the communications medium 26 is illustrated as a cloud to indicate that the communications medium 26 is capable of having a variety of different topologies including backbone, hub-and-spoke, loop, irregular, combinations thereof, and so on. Along these lines, the communications medium 26 may include copper-based data communications devices and cabling, fiber optic communications
25 devices and cabling, wireless communication devices, combinations thereof, etc. Furthermore, the communications medium 26 is capable of supporting SAN-based communications, LAN-based communications, cellular communications, WAN-based communications, distributed infrastructure communications, other topologies, combinations thereof, etc.

During operation, the storage processing circuitry 40 of the data storage equipment 24
30 writes host data 32 to the array of storage devices 42 and reads host data 32 from the array of storage devices 42 in response to the I/O requests 30. Routinely, the storage processing circuitry

40 re-replaces data (e.g., host data 32, metadata, etc.) among slices of the mapped-RAID environment 44 to more evenly distribute loading and wear. As a result, there is lower I/O latency and fewer failures due to premature wearing. Further details will now be provided with reference to Fig. 2.

5 Fig. 2 shows particular details of electronic circuitry 60 which forms the storage processing circuitry 40 of the data storage equipment 24 (also see Fig. 1) in accordance with certain embodiments. As shown, the electronic circuitry 60 includes a communications interface 70, memory 72, and processing circuitry 74, and other circuitry 76.

10 The communications interface 70 is constructed and arranged to connect the data storage equipment 24 to the communications medium 26 (also see Fig. 1) to enable electronic communications with other devices of the data storage setting 20 (e.g., the host computers 22). Such communications may be SAN-based, IP-based, cellular-based, cable-based, fiber-optic based, wireless, cloud-based, combinations thereof, and so on. Accordingly, the communications interface 70 enables the data storage equipment 24 to robustly and reliably communicate with
15 other external apparatus.

 The memory 72 is intended to represent both volatile storage (e.g., DRAM, SRAM, etc.) and non-volatile storage (e.g., flash memory, magnetic memory, etc.). In some arrangements, the memory 72 utilizes a portion of the array of storage devices 42 (Fig. 1). The memory 72 stores a variety of software constructs 80 including operating system code 82, data storage code 84, data
20 relocation code and data 86, as well as other code and data 88.

 The operating system code 82 refers to an operating system which is involved in controlling various electronic resources of the data storage equipment 24. For example, the operating system manages/allocates processor cycles, memory space, application installations, access control and privileges, and so on.

25 The data storage code 84 refers to the instructions that are involved in routine data storage operations (e.g., write and read operations). The data storage code 84 may include operating parameters, drivers, other metadata, etc. that form one or more portions of the I/O stack, mapping tables, and so on. In some arrangements, the data storage code 84 is tightly integrated with the operating system code 82 (e.g., a kernel).

30 The data relocation code and data 86 refers to specialized instructions that identify source slices containing data for relocation, and then relocate that data. In some arrangements, the data

relocation code and data 86 includes a database which stores a set of statistics that tracks data storage activity of the mapped-RAID environment 44. Such statistics are then used to identify candidate source and destination slices for re-placement.

5 The other code and data 88 refers to various other instructions, parameters, data, etc. For example, the data storage equipment 24 may be equipped with other tools and utilities, user-level applications, and so on.

10 The processing circuitry 74 is constructed and arranged to operate in accordance with the various software constructs 80 stored in the memory 72. As will be explained in further detail shortly, the processing circuitry 74 executes code of the various software constructs 80 to form specialized circuitry that concurrently carries out data storage operations as well as perform data redistribution operations.

15 Such processing circuitry 74 may be implemented in a variety of ways including via one or more processors (or cores) running specialized software, application specific ICs (ASICs), field programmable gate arrays (FPGAs) and associated programs, discrete components, analog circuits, other hardware circuitry, combinations thereof, and so on. In the context of one or more processors executing software, a computer program product 90 is capable of delivering all or portions of the software constructs 80 to the electronic circuitry 60. In particular, the computer program product 90 has a non-transitory (or non-volatile) computer readable medium which stores a set of instructions which controls one or more operations of the electronic circuitry 60.
20 Examples of suitable computer readable storage media include tangible articles of manufacture and apparatus which store instructions in a non-volatile manner such as CD-ROM, DVD, flash memory, disk memory, tape memory, and the like.

25 The other circuitry 76 represents additional hardware. For example, the other circuitry 76 may include a storage drive (or array) interface to communicate with the storage drives 42 (e.g., SATA, SAS, PCIe, etc.). As another example, the other circuitry 76 may include a user terminal or service processor that enables a human user to control and/or configure the data storage equipment 24, and so on. Further details will now be provided with reference to Figs. 3 and 4.

30 Figs 3 and 4 show particular storage details of the mapped-RAID environment 44. Fig. 3 shows particular slice details regarding device extents, RAID extents, stripes, slices, and LUNs. Fig. 4 shows particular slice relocation details.

As shown in Fig. 3, the data storage equipment 24 divides each storage device 42 into multiple device extents 100(1), 100(2), 100(3), ... (collectively, device extents 100). An example size for each device extent 100 is 1 Gigabyte (GB) although other sizes are suitable for use as well.

5 The data storage equipment 24 then combines device extents 100 from different storage devices 42 to form RAID extents 110. Each RAID extent 110 includes storage for data and parity such as RAID 4+1 which includes four device extents 100 that hold data and one device extent 100 that holds parity. For example, as shown in Fig. 3, the RAID extent 110(1) is formed by a device extent DE1 from the storage device 42(5), a device extent DE2 from the storage device
10 42(2), a device extent DE3 from the storage device 42(7), a device extent DE4 from the storage device 42(8), a device extent DE5 from storage the device 42(1). Generally, each storage device 42 that contributes storage to a RAID extent 110 should be of the same type (e.g., all solid state drives of the same capacity and speed, all magnetic drives of the same capacity and speed, etc.).

15 The data storage equipment 24 combines multiple RAID extents 110(1), 110(2), ... to form a RAID group 120. By way of example, if the size of each device extent 100 is 1GB, the size of each RAID extent 110 is 5GB in a RAID 4+1 implementation. Of course, other RAID implementations and other RAID extent sizes are suitable for use in the mapped-RAID environment 44.

20 It should be understood that, in the RAID 4+1 example, a RAID group 120 that includes two RAID extents 110 distributes data across a maximum of ten (10) storage devices 42 (i.e., where there are no storage devices 42 in common between the two RAID extents 110). Likewise, in the RAID 4+1 example, a RAID group 120 that includes three RAID extents 110 distributes data across a maximum of fifteen (15) storage devices 42, and so on.

25 The data storage equipment 24 divides each RAID extent 110 (see the RAID extent 110(1) in Fig. 3) into multiple storage stripes 130(1), 130(2), 130(3), 130(4), ... (collectively, storage stripes 130). In one RAID 4+1 example arrangement, each storage stripe 130 is capable of holding 320 Kilobytes (KB) of data and parity (e.g., four segments 132 that hold data and one segment 132 that holds parity). In such an example arrangement, the size of each stripe segment 132 is 64KB.

30 The data storage equipment 24 then combines multiple storage stripes 130 together to form slices 140(1), ... 140(N-1), 140(N), 140(N+1), ... (collectively, slices 140) of a logical unit

of storage (LUN) 150. The arrow 142 pointing to a portion of the slice 140(N) delineated by hash marks is intended to illustrate the relationship between a storage stripe 130 and a slice 140. In some arrangements, each slice 140 is 256MB in size, although other sizes (e.g., 128MB, 512MB, 1GB, etc.) are suitable for use as well.

5 It should be understood that such a LUN 150 is addressable in that data can be read using offsets (e.g., logical block addresses which are mapped to underlying storage). In some arrangements, such a LUN 150 is thickly provisioned in that actual physical storage from the storage devices 42 supports the entire LUN address space. In other arrangements, such a LUN 150 is thinly provisioned in that physical storage from the storage devices 42 is allocated only upon storage of actual data.

10 Fig. 4 shows details of how a data of a slice 140 can be relocated to another slice 140. By way of example, the storage devices 42 form multiple storage tiers 160(1), 160(2), 160(3) (collectively, tiers 160). The storage tier 160(1) may include fast storage devices 42 (e.g., solid state devices), the storage tier 160(2) may include moderate speed storage devices 42 (e.g., SAS devices), and the storage tier 160(1) may include slow storage devices 42 (e.g., near-line SAS devices). Other tier arrangements are suitable for the data storage equipment 24 as well such as storage that includes only a single tier 150, only two tiers 150, four tiers 150, and so on.

15 Additionally, each storage tier 150 includes at least one RAID group 120 (also see Fig. 3). For example, the tier 150(3) includes two distinct RAID groups 120(1) and 120(2). It should be understood that other numbers of RAID groups are suitable for each tier 150 (e.g., one, three, etc.). Moreover, the different tiers 150 may have different numbers of RAID groups 120.

20 As shown in Fig. 4, the data storage equipment 24 maintains a database 170 that contains a set of statistics that tracks data storage activity of the mapped-RAID environment 44 (also see the data relocation code and data 86 in Fig. 2). Such a database 170 includes storage I/O loading information such as how frequently each storage device 42 is accessed per unit of time such as per day, per hour, etc. Additionally, the database 170 includes wear data such as how many times each storage device 42 has been accessed overall. Furthermore, the database 170 includes heat (or temperature) data indicating how often each data slice 140 of the LUN 150 is accessed per unit of time such as per day, per hour, etc. Other metrics and data are suitable for storing in the database 170 as well such as indications whether each data slice of the LUN 150 is data or metadata, moving averages, predefined thresholds, which device extents 100 form which RAID

extents 110, which RAID extents 110 for which RAID groups 120, which storage devices 42 belong to which RAID groups 120, and so on.

In some embodiments, the storage processing circuitry 40 is configured and operative to read wear data from each storage device 42 periodically. Along these lines, the storage
5 processing circuitry 40 obtains the current program-erase (P/E) cycle number and the maximum number of P/E cycles. From this information, the storage processing circuitry 40 is able to ascertain the percentage of wear of each storage device 42 (e.g., the current P/E divided by the maximum number of P/E cycles as a percentage). Additionally, the storage processing circuitry 40 is able to estimate how many cycles are consumed between updates (e.g., the number of P/E
10 cycles consumed since the last time the storage processing circuitry read the wear data, and so on). Other metrics can be obtained, derived, and/or estimated as well.

During operation, the storage processing circuitry 40 selects data within a source slice 140(S) to re-place within the mapped-RAID environment 44 in order to improve loading and wear characteristics. In particular, the storage processing circuitry 40 identifies a source slice
15 140(S) containing hot data among the other slices 140 based on the database 170 containing the set of statistics that tracks data storage activity (e.g., the most frequently accessed slice 140, a slice 140 having a number of writes that exceeds a predefined write threshold, a slice 140 having a number of accesses that exceeds a predefined access threshold, etc.). In some arrangements, the database 170 includes moving averages, and the storage processing circuitry 40 compares
20 particular slice metrics to one or more of the moving averages in order to identify a candidate source slice 140(S) that stands out from a relative perspective among the other slices 140. Accordingly, the results of such evaluation by the storage processing circuitry 40 may be viewed essentially as predictions as to when source slice activity justifies relocating the slice data.

In some embodiments, the storage processing circuitry selects a destination slice 140(D)
25 within the mapped-RAID environment 44 to be an available slice 140 that is different from the current source slice 140(S). As a result, I/O and wear are balanced statistically. That is, from time to time, the source slice 140(S) moves from one place to another and the corresponding I/O moves with it.

In alternative embodiments, the storage processing circuitry 40 selects a destination slice
30 140(D) within the mapped-RAID environment 44. Similarly, the storage processing circuitry 40 identifies a destination slice 140(D) among other slices 140 based on the set statistics that tracks

data storage activity (e.g., the least frequently accessed slice 140, a slice 140 having a number of writes that is below a predefined write threshold, a slice 140 having a number of accesses that is below a predefined access threshold, etc.). Again, in some arrangements, the storage processing circuitry 40 compares particular slice metrics to one or more moving averages in order to identify, as the destination slice 140(D), a slice 140 that stands out from a relative perspective among the other slices 140.

Once the source slice 140(S) and the destination slice 140(D) have been chosen, the storage processing circuitry 40 relocates data (the shaded portions) from the source slice 140(S) to the destination slice 140(D) (see the arrow 180 in Fig. 4). In particular, the slice data is redistributed among the storage devices 42 based on the set of statistics within the database 170. Such operation statistically rebalances the storage devices 42 in terms of I/O loading and wearing thus reducing overall latency and reducing failures due to premature wear.

Upon relocating the particular data from the source slice 140(S) to the destination slice 140(D), the storage processing circuitry 40 updates a set of address mappings that maps the LUN 150 to actual storage. Accordingly, subsequent I/O operations to access the slice data are no longer directed to the source slice 140(S) but are now directed to the destination slice 140(S).

It should be understood that while host data 32 is moved from a source slice 140(S) to a destination slice 140(D), the host data 32 may be accessed by the set of host computers 22. In particular, the data may be read, modified, written, etc. Accordingly, all changes are eventually reflected in the host data 32 that resides in the destination slice 140(D). Moreover, the storage processing circuitry 40 continues to update the set of statistics within the database 170 so that the database 170 continues to track the data storage activity of the mapped-RAID environment 44.

Furthermore, the storage processing circuitry 40 may impose policies on data relocation. Along these lines, the storage processing circuitry 40 may restrict metadata (e.g., data used by the data storage equipment 24 during operation that is transparent to the host computers 22) to a particular storage tier 160 (e.g., the fastest storage tier 160). Other policies are suitable for use as well.

It should be further understood that such moving of slice data from a source slice 140(S) to a destination slice 140(D) may occur within the same RAID group 120 within the same storage tier 150. Such re-placement of the slice data may be warranted for frequently accessed slice data

that resides in a relatively lightly used RAID group 120 or when the data storage system 24 (or the particular storage tier 160) includes only one RAID group 120.

Alternatively, the storage processing circuitry 40 may move slice data from a source slice 140 in one RAID group 120 to a destination slice 140 in a different RAID group 120 within the same storage tier 150. Such re-placement of the slice data may be warranted for frequently accessed slice data that resides a heavily used RAID group 120 in an appropriate storage tier 160 or when the data storage system 24 includes only one storage tier 160.

As yet another alternative, the storage processing circuitry 40 may move slice data from a slice 140 in one RAID group 120 to a slice 140 in a different RAID group 120 within different storage tiers 150. Such re-placement of the slice data may be warranted for frequently accessed slice data that resides a heavily used RAID group 120 in an inappropriate storage tier 160 (e.g., when the slice data should be promoted or demoted due to a change in access frequency, etc., also see the darkened arrows extending between tiers 160 in Fig. 4).

It should be understood that such moving of data slices may occur in the background and be transparent to the host computers 22 (Fig. 1). Along these lines, such data re-placement may be performed while the data storage equipment 24 concurrently performs I/O operations in response to I/O requests 30 from the host computers 22. For example, such re-placement may take place during idle cycles, period of low data storage activity, during other operations such as data promotion/demotion, combinations thereof, etc. Further details will now be provided with reference to Fig. 5.

Fig. 5 is a flowchart of a procedure 200 which is performed by the data storage equipment 24. Such a procedure 200 balances data storage activity within a mapped-RAID environment 44.

At 202, the data storage equipment 24 selects a source slice of storage from multiple slices of storage of the mapped-RAID environment. The source slice contains particular data to be relocated. Such selection may be based on an evaluation of historical data storage activity (e.g., also see the database 170 in Fig. 4).

At 204, the data storage equipment 24 select a destination slice of storage from the multiple slices of storage of the mapped-RAID environment. Again, such selection may be based on an evaluation of historical data storage activity.

At 206, data storage equipment 24 relocates the particular data from the source slice to the destination slice to balance data storage activity within the mapped-RAID environment. The mapped-RAID environment includes multiple storage devices. Each storage device provides multiple non-overlapping device extents. Each slice of the multiple slices of storage of the mapped-RAID environment is formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment. It should be understood that such operation may take place while host computers 22 access host data within the mapped-RAID environment (i.e., I/O operations are performed concurrently).

As described above, improved techniques are directed to re-placing (or relocating) data within a mapped-RAID environment 44 to provide improved I/O load and wear balancing among storage devices 42. In particular, such techniques identify imbalances (e.g., heavily accessed slices of data) and move the imbalances among the storage devices 42 of the mapped-RAID environment 44 (e.g., redistribute the heavily accessed slices of data). Such operation statistically balances the storage devices 42 in terms of I/O performance and wearing thus reducing overall latency and reducing failures due to premature wear.

One should appreciate that the above-described techniques do not merely read/write data. Rather, the disclosed techniques involve improvements to the technology of data storage. For example, such techniques improve data access speeds by balancing I/O activity across multiple storage devices 42. Additionally, such techniques reduce the chance of encountering an unexpected storage device failure due to earlier than anticipated wear on a particular storage device 42, and so on.

While various embodiments of the present disclosure have been particularly shown and described, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the present disclosure as defined by the appended claims.

For example, it should be understood that various components of the data storage setting such as the host computers 22 are capable of being implemented in or “moved to” the cloud, i.e., to remote computer resources distributed over a network. Here, the various computer resources may be distributed tightly (e.g., a server farm in a single facility) or over relatively large distances (e.g., over a campus, in different cities, coast to coast, etc.). In these situations, the

network connecting the resources is capable of having a variety of different topologies including backbone, hub-and-spoke, loop, irregular, combinations thereof, and so on. Additionally, the network may include copper-based data communications devices and cabling, fiber optic devices and cabling, wireless devices, combinations thereof, etc. Furthermore, the network is capable of supporting LAN-based communications, SAN-based communications, combinations thereof, and so on.

The individual features of the various embodiments, examples, and implementations disclosed within this document can be combined in any desired manner that makes technological sense. Furthermore, the individual features are hereby combined in this manner to form all possible combinations, permutations and variants except to the extent that such combinations, permutations and/or variants have been explicitly excluded or are impractical. Support for such combinations, permutations and variants is considered to exist within this document. Such modifications and enhancements are intended to belong to various embodiments of the disclosure.

CLAIMS

What is claimed is:

1. A method of balancing data storage activity within a mapped-RAID (redundant array of independent disk) environment, the method comprising:
 - selecting, by processing circuitry, a source slice of storage from multiple slices of storage of the mapped-RAID environment, the source slice containing particular data to be relocated;
 - selecting, by the processing circuitry, a destination slice of storage from the multiple slices of storage of the mapped-RAID environment; and
 - relocating, by the processing circuitry, the particular data from the source slice to the destination slice to balance data storage activity within the mapped-RAID environment, the mapped-RAID environment including multiple storage devices, each storage device providing multiple non-overlapping device extents, each slice of the multiple slices of storage of the mapped-RAID environment being formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment.
2. A method as in claim 1, further comprising:
 - prior to selecting the source slice and the destination slice, performing a set of input/output (I/O) operations that accesses the particular data while the particular data resides in the source slice and updating, in response to performing the set of I/O operations, a set of statistics that tracks data storage activity of the mapped-RAID environment.
3. A method as in claim 2 wherein selecting the source slice of storage from the multiple slices of storage of the mapped-RAID environment includes:
 - identifying the source slice among other slices of the mapped-RAID environment based on the set of statistics that tracks data storage activity of the mapped-RAID

environment indicating that the source slice has experienced higher data storage activity than the other slices of the mapped-RAID environment.

4. A method as in claim 3 wherein selecting the destination slice of storage from the multiple slices of storage of the mapped-RAID environment includes:
 - identifying the destination slice among the other slices of the mapped-RAID environment based on the set of statistics that tracks data storage activity of the mapped-RAID environment indicating that the destination slice has experienced less data storage activity than the other slices of the mapped-RAID environment.
5. A method as in claim 4, further comprising:
 - while performing the set of I/O operations that accesses the particular data while the particular data resides in the source slice, maintaining an initial set of address mappings that map a portion of a logical unit of storage (LUN), which is addressable by a set applications, to the source slice.
6. A method as in claim 5, further comprising:
 - upon relocating the particular data from the source slice to the destination slice, maintaining a subsequent set of address mappings that map the portion of the LUN to the destination slice.
7. A method as in claim 6, further comprising:
 - after relocating the particular data from the source slice to the destination slice, performing another set of I/O operations that accesses the particular data while the particular data resides in the destination slice, and updating the set of statistics that tracks data storage activity of the mapped-RAID environment in response to performing the other set of I/O operations.
8. A method as in claim 7 wherein performing the set of I/O operations that accesses the particular data while the particular data resides in the source slice includes addressing, by

the set of applications, the portion of the LUN using block addresses; and

wherein performing the other set of I/O operations that accesses the particular data while the particular data resides in the destination slice includes addressing, by the set of applications, the portion of the LUN using the same block addresses that the set of applications used to address the particular data while the particular data resided in the source slice.

9. A method as in claim 8, further comprising:

while identifying the source slice, identifying the destination slice and relocating the particular data from the source slice to the destination slice, performing other I/O operations that access data within the LUN.

10. A method as in claim 9 wherein performing other I/O operations that access the data within the LUN includes:

accessing the particular data on behalf of a set of host computers while the particular data is being relocated from the source slice to the destination slice.

11. A method as in claim 4 wherein identifying the source slice among other slices of the mapped-RAID environment based on the set of statistics includes:

generating a first prediction that the source slice will encounter a future access behavior that exceeds a predefined high access threshold, and

choosing the source slice among the other slices in response to the first prediction; and

wherein identifying the destination slice among the other slices of the mapped-RAID environment based on the set of statistics includes:

generating a second prediction that the destination slice will encounter a future access behavior that is below a predefined low access threshold, and

choosing the destination slice among the other slices in response to the second prediction.

12. A method as in claim 4 wherein identifying the source slice among other slices of the mapped-RAID environment based on the set of statistics includes:
- choosing the source slice among the other slices in response to a total number of write operations to the source slice exceeding a first predefined write cycle threshold; and
- wherein identifying the destination slice among the other slices of the mapped-RAID environment based on the set of statistics includes:
- choosing the destination slice among the other slices in response to the total number of write operations to the destination slice being below a second predefined write cycle threshold.
13. A method as in claim 4 wherein identifying the source slice among other slices of the mapped-RAID environment based on the set of statistics includes:
- generating a first prediction that the source slice will encounter a future access behavior that exceeds a predefined high access threshold, and
 - choosing the source slice among the other slices in response to the first prediction and the total number of write operations to the source slice exceeding a first predefined write cycle threshold; and
- wherein identifying the destination slice among the other slices of the mapped-RAID environment based on the set of statistics includes:
- generating a second prediction that the destination slice will encounter a future access behavior that is below a predefined low access threshold, and
 - choosing the destination slice among the other slices in response to the second prediction and the total number of write operations to the destination slice being below a second predefined write cycle threshold.
14. A method as in claim 1 wherein the source slice resides within a first RAID extent that is formed by a first group of device extents, each device extent of the first group residing on a different storage device of the mapped-RAID environment; wherein the destination slice resides within a second RAID extent that is formed by a

second group of device extents, each device extent of the second group residing on a different storage device of the mapped-RAID environment; and
wherein the first group of device extents reside on a first group of storage devices, the second group of device extents reside on a second group of storage devices, and there is at least one storage device of the second group of storage devices that does not belong to the first group of storage devices.

15. A method as in claim 14 wherein none of the storage devices of the second group of storage devices belongs to the first group of storage devices.
16. A method as in claim 14 wherein the mapped-RAID environment includes multiple RAID groups; and
wherein the first RAID extent and the second RAID extent reside within the same RAID group.
17. A method as in claim 14 wherein the mapped-RAID environment includes multiple RAID groups;
wherein the first RAID extent resides within a first RAID group which is formed by a first group of RAID extents;
wherein the second RAID extent resides within a second RAID group which is formed by a second group of RAID extents; and
wherein none of the RAID extents of the second group of RAID extents belongs to the first group of RAID extents.
18. A method as in claim 14 wherein the mapped-RAID environment includes multiple storage tiers, each storage tier being formed by multiple storage devices and providing a different access speed behavior;
wherein the first RAID extent resides within a first storage tier;
wherein the second RAID extent resides within a second storage tier that is different from the first storage tier.

19. Data storage equipment, comprising:
- a communications interface;
 - memory including a mapped-RAID (redundant array of independent disk) environment having multiple storage devices, each storage device providing multiple non-overlapping device extents; and
 - control circuitry coupled to the communications interface and the memory, the memory storing instructions which, when carried out by the control circuitry, cause the control circuitry to:
 - select a source slice of storage from multiple slices of storage of the mapped-RAID environment, the source slice containing particular data to be relocated,
 - select a destination slice of storage from the multiple slices of storage of the mapped-RAID environment, and
 - relocate the particular data from the source slice to the destination slice to balance data storage activity within the mapped-RAID environment, each slice of the multiple slices of storage of the mapped-RAID environment being formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment.
20. A computer program product having a non-transitory computer readable medium which stores a set of instructions to balance data storage activity within a mapped-RAID (redundant array of independent disk) environment; the set of instructions, when carried out by computerized circuitry, causing the computerized circuitry to perform a method of:
- selecting a source slice of storage from multiple slices of storage of the mapped-RAID environment, the source slice containing particular data to be relocated;
 - selecting a destination slice of storage from the multiple slices of storage of the mapped-RAID environment; and
 - relocating the particular data from the source slice to the destination slice to

balance data storage activity within the mapped-RAID environment, the mapped-RAID environment including multiple storage devices, each storage device providing multiple non-overlapping device extents, each slice of the multiple slices of storage of the mapped-RAID environment being formed of storage stripes extending across device extents provided by a group of storage devices that includes less than all of the storage devices of the mapped-RAID environment.

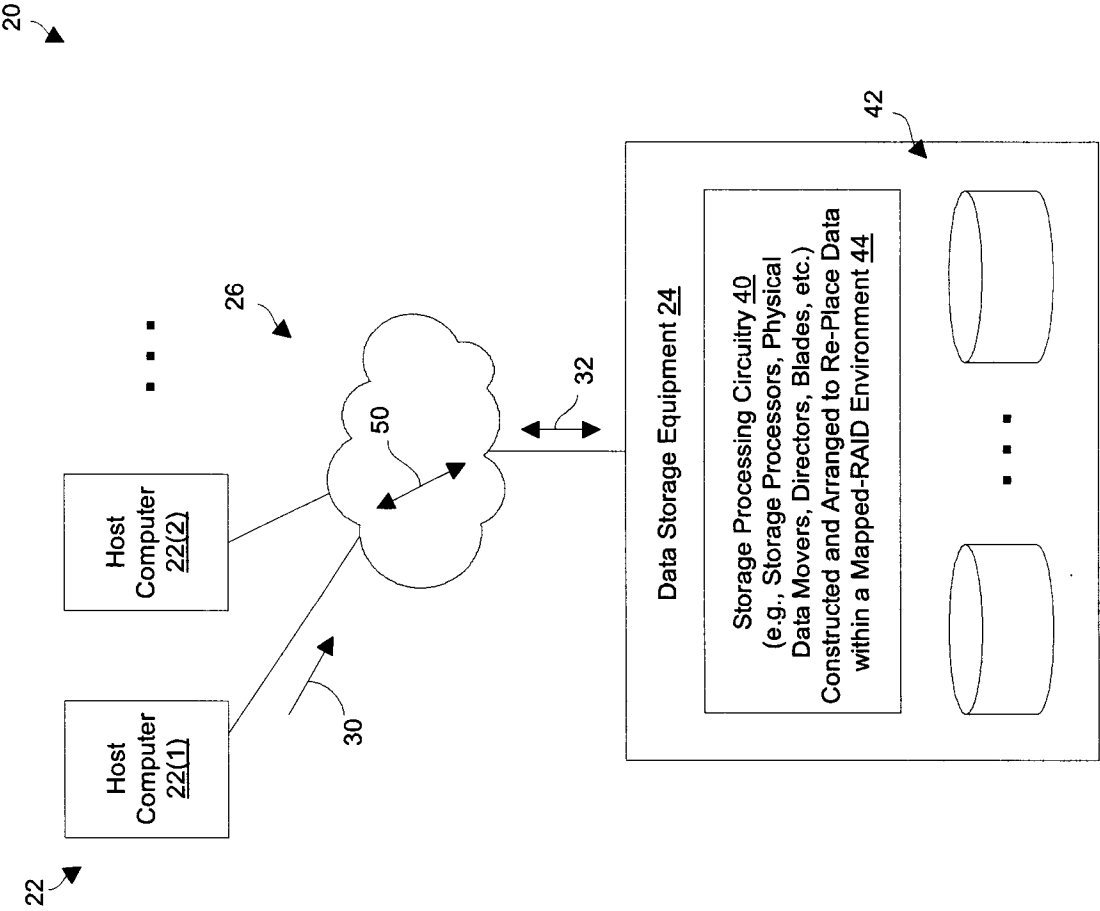


FIG. 1

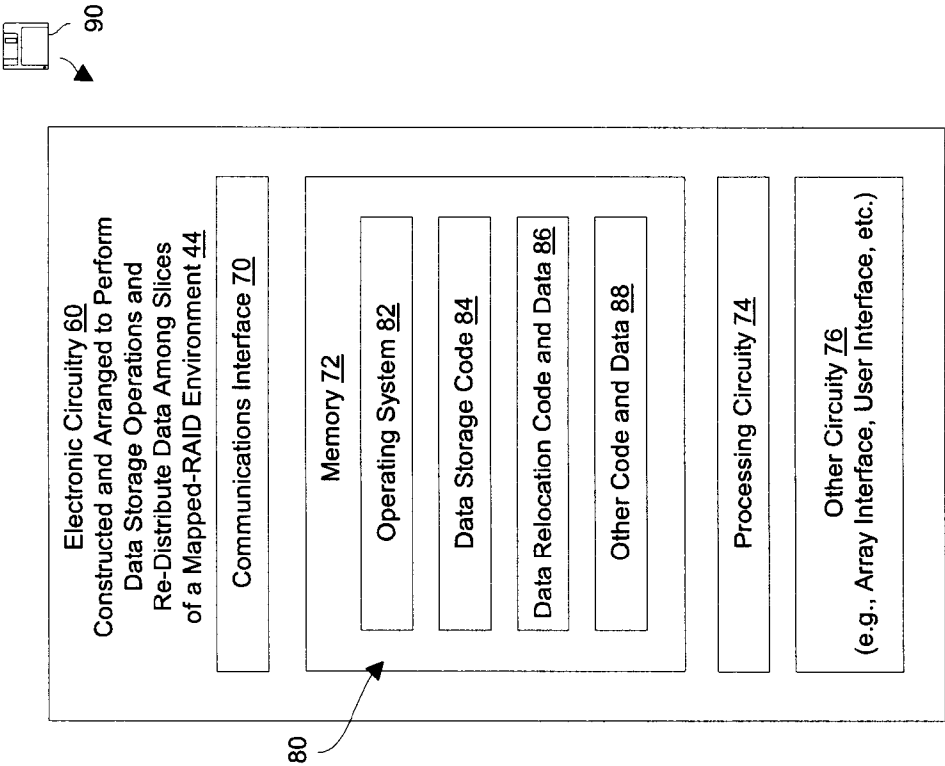


FIG. 2

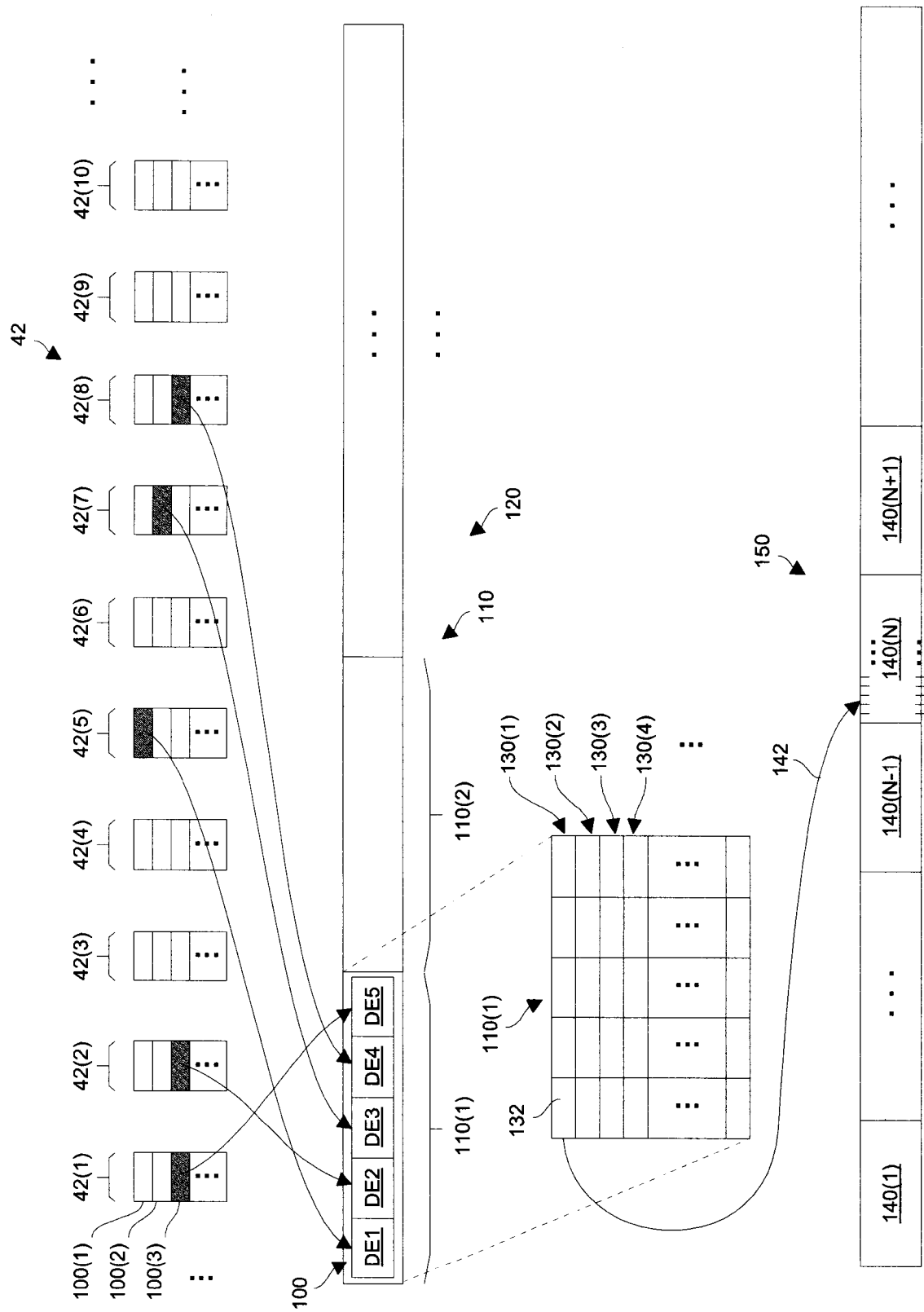
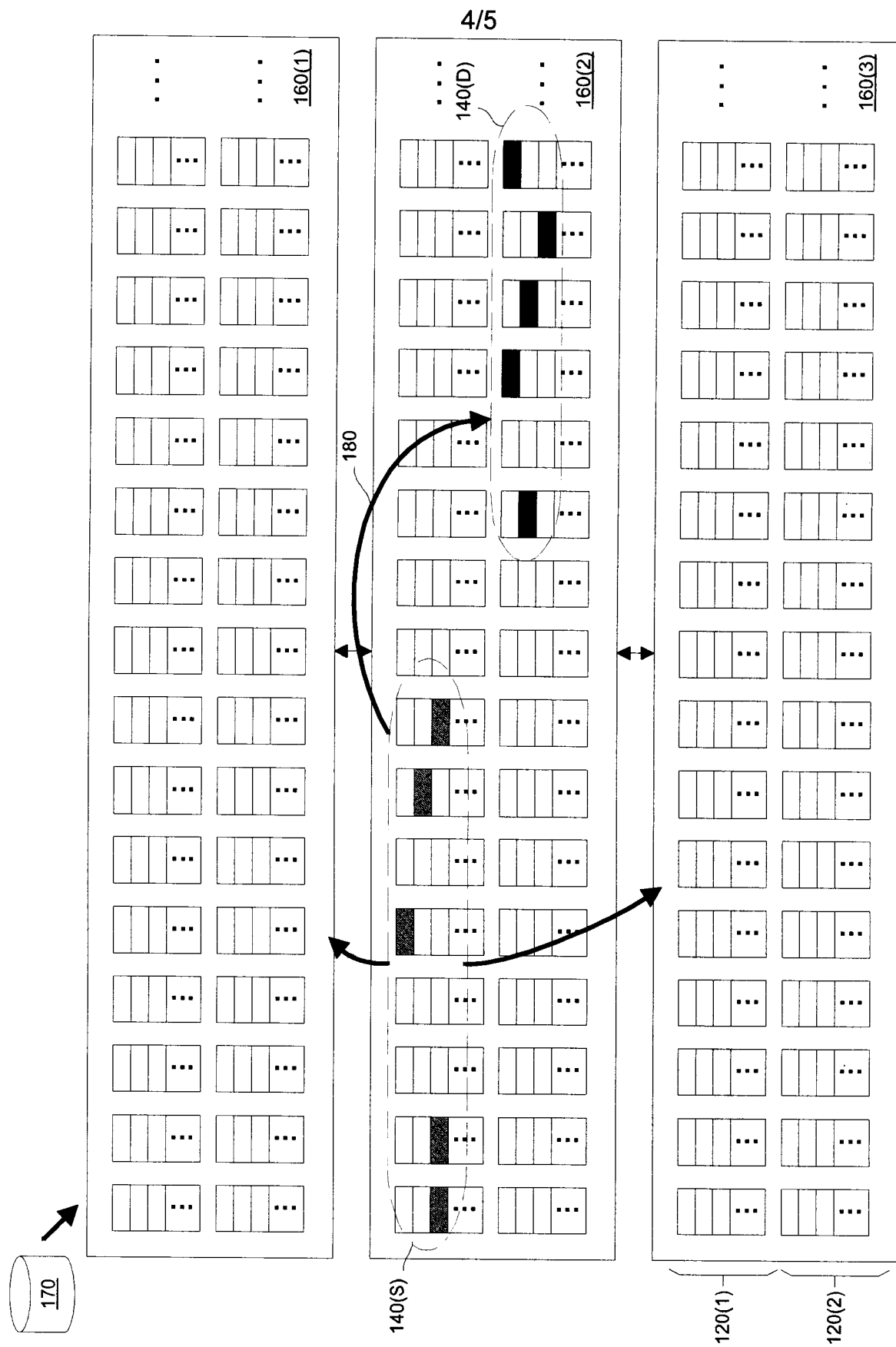


FIG. 3



5/5

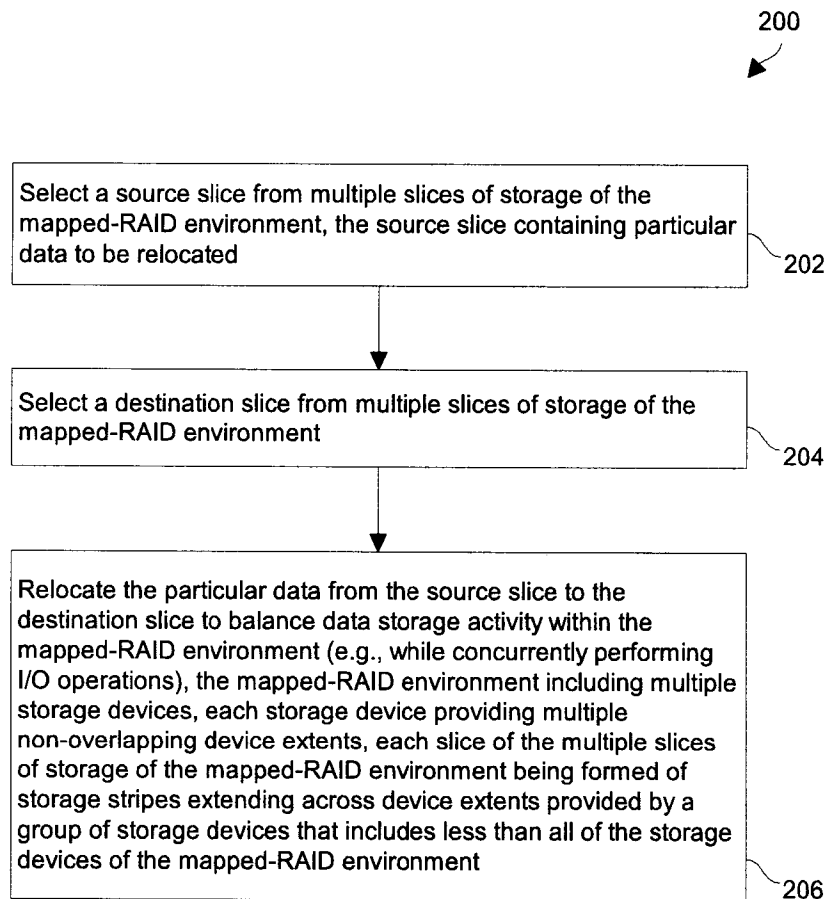


FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/RU 2017/000273

A. CLASSIFICATION OF SUBJECT MATTER <i>G06F 12/02 (2006.01)</i> <i>G06F 11/30 (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 3/00, 3/06, 11/00, 11/30, 12/00, 12/02, 13/00, 13/14, 13/20 Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) PatSearch (RUPTO internal), USPTO, PAJ, Esp@cenet, DWPI, EAPATIS, PATENTSCOPE, Information Retrieval System of FIPS		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2008/0109601 A1 (MICHAEL J. KLEMM et al.) 08.05.2008, abstract, [0002], [0025], [0026], [0031], [0045], [0049], [0051], [0061], [0072], [0092], fig. 1A, 2A, 2B, 3A, 3B, 4	1, 14-20
Y		2
A		3-13
Y	US 2003/0065883 A1 (EITAN BACHMAT) 03.04.2003, abstract, [0011]-[0013]	2
A	US 5345584 A (LACLEAD ENTERPRISES) 06.09.1994	1-20
A	US 2016/0364149 A1 (DELL INTERNATIONAL L.L.C.) 15.12.2016	1-20
☐ 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 document 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 19 December 2017 (19.12.2017)	Date of mailing of the international search report 25 January 2018 (25.01.2018)	
Name and mailing address of the ISA/RU: Federal Institute of Industrial Property, Berezhkovskaya nab., 30-1, Moscow, G-59, GSP-3, Russia, 125993 Facsimile No: (8-495) 531-63-18, (8-499) 243-33-37	Authorized officer T. Kiseleva Telephone No. (499) 240-25-91	