

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
6 December 2007 (06.12.2007)

PCT

(10) International Publication Number
WO 2007/140260 A2

- (51) International Patent Classification: **Not classified**
- (21) International Application Number:
PCT/US2007/069669
- (22) International Filing Date: 24 May 2007 (24.05.2007)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
60/808,045 24 May 2006 (24.05.2006) US
- (71) Applicant (for all designated States except US): **COMPELLENT TECHNOLOGIES** [US/US]; 12982 Valley View Road, Eden Prairie, MN 55344 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **KLEMM, Michael, J.** [US/US]; 2301 Rivendell Lane, Minnetonka, MN 55305 (US). **UTTORMARK, Michael, J.** [US/US]; 615 Jewel Lane N, Plymouth, MN 55447 (US).
- (74) Agent: **PADMANABHAN, Devan, V.**; Dorsey & Whitney LLP, Intellectual Property Department, Suite 1500, 50 South Sixth Street, Minneapolis, MN 55402-1498 (US).
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
- Published:**
— without international search report and to be republished upon receipt of that report
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEM AND METHOD FOR RAID MANAGEMENT, REALLOCATION, AND RESTRIPIING

(57) Abstract: The present disclosure relates to systems and methods for RAID Restripping. One method includes selecting an initial RAID device for migration based on at least one score, creating an alternate RAID device, moving data from the initial RAID device to the alternate RAID device, and removing the initial RAID device. The method may be performed automatically by the system or manually. The method may be performed periodically, continuously, after every RAID device migration, upon addition of disk drives, and/or before removal of disk drives, etc. One system includes a RAID subsystem and a disk manager configured to automatically calculate a score for each RAID device, select a RAID device based on the relative scores of the RAID devices, create an alternate RAID device, move data from the selected RAID device to the alternate RAID device, and remove the selected RAID device.



WO 2007/140260 A2

**SYSTEM AND METHOD FOR RAID MANAGEMENT, REALLOCATION,
AND RESTRIPING**

Cross-Reference to Related Application(s)

[001] This application claims priority to U.S. provisional patent application
5 Ser. No. 60/808,045, filed May 24, 2006, which is incorporated herein by reference
in its entirety.

Field of the Invention

[002] The present invention relates generally to disk drive systems and
methods, and more particularly to disk drive systems and methods having a dynamic
10 block architecture RAID Device Management, Reallocation, and Restriping for
optimizing RAID Device layout when changes to RAID parameters or disk
configuration occur.

Background of the Invention

[003] Existing disk drive systems have been designed in such a way that a
15 Virtual Volume is distributed (or mapped) across the physical disks in a manner
which is determined at volume creation time and remains static throughout the
lifetime of the Virtual Volume. That is, the disk drive systems statically allocate
data based on the specific location and size of the virtual volume of data storage
space. Should the Virtual Volume prove inadequate for the desired data storage
20 purposes, the existent systems require the creation of a new Virtual Volume and the
concomitant copying of previously stored data from the old Virtual Volume to the
new in order to change volume characteristics. This procedure is time consuming
and expensive since it requires duplicate physical disk drive space.

[004] These prior art disk drive systems need to know, monitor, and control
25 the exact location and size of the Virtual Volume of data storage space in order to
store data. In addition, the systems often need larger data storage space, whereby
more RAID Devices are added. As a result, emptied data storage space is not used,
and extra data storage devices, e.g. RAID Devices, are acquired in advance for
storing, reading/writing, and/or recovering data in the system. Additional RAID
30 Devices are expensive and not required until extra data storage space is actually
needed.

[005] Therefore, there is a need for improved disk drive systems and methods, and more particularly a need for efficient, dynamic RAID space and time management systems. There is a further need for improved disk drive systems and methods for allowing RAID management, reallocation, and restriping to occur
5 without loss of server or host data access or compromised resiliency.

Brief Summary of the Invention

[006] The present invention, in one embodiment, is a method of RAID Restriping in a disk drive system. The method includes selecting an initial RAID device for migration based on at least one score, creating an alternate RAID device,
10 moving data stored at the initial RAID device to the alternate RAID device; and removing the initial RAID device. The scores may include an initial score, a replacement score, and an overlay score. Furthermore, the method may be performed automatically by the system or manually, such as by a system administrator. The method may be performed periodically, continuously, after every
15 RAID device migration, upon addition of disk drives, and/or before removal of disk drives.

[007] The present invention, in another embodiment, is a disk drive system having a RAID subsystem and a disk manager. The disk manager is configured to automatically calculate a score for each RAID device of the RAID subsystem, select
20 a RAID device from the subsystem based on the relative scores of the RAID devices, create an alternate RAID device, move a portion of the data stored at the selected RAID device to the alternate RAID device, and remove the selected RAID device.

[008] The present invention, in yet another embodiment, is a disk drive
25 system including means for selecting a RAID device for migration based on at least one score calculated for each RAID device, means for creating at least one alternate RAID device, means for moving data stored at the selected RAID device to the at least one alternate RAID device, and means for removing the selected RAID device.

[009] While multiple embodiments are disclosed, still other embodiments
30 of the present invention will become apparent to those skilled in the art from the following detailed description, which shows and describes illustrative embodiments of the invention. As will be realized, the invention is capable of modifications in various obvious aspects, all without departing from the spirit and scope of the

present invention. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not restrictive.

Brief Description of the Drawings

5 [010] While the specification concludes with claims particularly pointing out and distinctly claiming the subject matter that is regarded as forming the present invention, it is believed that the invention will be better understood from the following description taken in conjunction with the accompanying Figures, in which:

[011] FIG. 1A is a disk array having a RAID configuration in accordance with one embodiment of the present invention.

[012] FIG. 1B is the disk array of FIG. 1A having an additional RAID Device.

[013] FIG. 1C is the disk array of FIG. 1B after removing a RAID Device.

15 [014] FIG. 2A is a disk array having a RAID configuration in accordance with another embodiment of the present invention.

[015] FIG. 2B is the disk array of FIG. 2A having an additional RAID Device.

[016] FIG. 2C is the disk array of FIG. 2B after removing a RAID Device and adding another RAID Device.

20 [017] FIG. 2D is the disk array of FIG. 2C after removing yet another RAID device.

[018] FIG. 3A is a disk array having a RAID configuration in accordance with a further embodiment of the present invention.

25 [019] FIG. 3B is the disk array of FIG. 3A illustrating migration of a RAID Device.

[020] FIG. 3C is the disk array of FIG. 3B illustrating further migration of multiple RAID Devices.

[021] FIG. 3D is the disk array of FIG. 3C illustrating yet further migration of multiple RAID Devices.

30 [022] FIG. 3E is the disk array of FIG. 3A in a new RAID configuration.

[023] FIG. 4 is a flow chart of a process of Restriping in accordance with one embodiment of the present invention.

[024] FIG. 5 is a disk array having a RAID configuration with different-sized RAID Devices in accordance with one embodiment of the present invention.

Detailed Description

[025] Various embodiments of the present invention relate generally to disk drive systems and methods, and more particularly to disk drive systems and methods which implement one or more Virtual Volumes spread across one or more RAID Devices, which in turn are constructed upon a set of disk drives. RAID Device Management, Reallocation, and Restripping ("Restripping") provides a system and method for changing the various properties associated with a Virtual Volume such as size, data protection level, relative cost, access speed, etc. This system and method may be initiated by administration action or automatically when changes to the disk configuration occur.

[026] The various embodiments of the present disclosure provide improved disk drive systems having a dynamic block architecture RAID Device Restripping that may optimize RAID Device layout when changes to RAID parameters or disk configuration occur. In one embodiment, the layout of RAID Devices may be primarily rebalanced when disks are added to the system. By rebalancing, virtualization performance may be improved within the system by using the maximum available disk configuration. Restripping also may provide the capability to migrate data away from a group of disks, allowing those disks to be removed from the system without loss of uptime or data protection. Further, Restripping may provide the capability to change RAID parameters giving the user the ability to tune the performance and/or storage capacity even after the data has been written. Restripping additionally may provide an improved disk drive system and method for allowing Restripping to occur without loss of server or host data access or compromised resiliency.

[027] Various embodiments described herein improve on the existent disk drive systems in multiple ways. In one embodiment, the mapping between a Virtual Volume and the physical disk drive space may be mutable on a fine scale. In another embodiment, previously stored data may be migrated automatically in small units, and the appropriate mappings may be updated without the need for an entire duplication of physical resources. In a further embodiment, portions of a Virtual Volume which are already mapped to appropriate resources need not be migrated,

reducing the time needed for reconfiguration of a Volume. In yet another embodiment, the storage system can automatically reconfigure entire groups of Virtual Volumes in parallel. Additionally, the storage system may automatically reconfigure Virtual Volumes when changes to the physical resources occur. Other advantages over prior disk drive systems will be recognized by those skilled in the art and are not limited to those listed.

[028] Furthermore, Restriping and disk categorization may be powerful tools for administrative control of the storage system. Disk drives which, for example, are found to be from a defective manufacturing lot, may be recategorized so that migration away from these disk drives occurs. Similarly, a set of drives may be held in a "reserve" category, and later recategorized to become part of a larger in-use group. Restriping to widen the RAID Devices may gradually incorporate these additional reserve units. It is noted that several benefits may be recognized by the embodiments described herein, and the previous list of examples is not exhaustive and not limiting.

[029] For the purposes of describing the various embodiments herein, a "Volume" may include an externally accessible container for storing computer data. In one embodiment, a container may be presented via the interconnect protocol as a contiguous array of blocks. In a further embodiment, each block may have a fixed size - traditionally 512 bytes. Although, other sized blocks may be used, such as 256, 1,024, etc. bytes. Typically, supported operations performed on data at any given location may include 'write' (store) and 'read' (retrieve). Although, other operations, such as 'verify' may also be supported. The interconnect protocol used to access Volumes may be the same as that used to access disk drives. Thus, in some embodiments, a Volume may appear and function generally identical to that of a disk drive. Volumes traditionally may be implemented as partitions of a disk drive or simple concatenations of disk drives within an array.

[030] A "Virtual Volume," as used herein, may include an externally accessible container for storing data which is constructed from a variety of hardware and software resources and generally may mimic the behavior of a traditional Volume. In particular, a system containing a disk drive array may present multiple Virtual Volumes which utilize non-intersecting portions of the disk array. In this type of system, the storage resources of the individual disk drives may be aggregated in an array, and subsequently partitioned into individual Volumes for use by external

computers. In some embodiments, the external computers may be servers, hosts, etc.

[031] A "RAID Device," as used herein, may include an aggregation of disk partitions which provides concatenation and resiliency to disk drive failure.

5 The RAID algorithms for concatenation and resiliency are well known and include such RAID levels as RAID 0, RAID 1, RAID 0+1, RAID 5, RAID 10, etc. In a given disk array, multiple RAID Devices may reside on any given set of disks. Each of these RAID Devices may employ a different RAID level, have different parameters, such as stripe size, may be spread across the individual disk drives in a different order, may occupy a different subset of the disk drives, etc. A RAID
10 Device may be an internally accessible Virtual Volume. It may provide a contiguous array of data storage locations of a fixed size. The particular RAID parameters determine the mapping between RAID Device addresses and the data storage addresses on the disk drives. In the present disclosure, systems and methods
15 for constructing and modifying externally accessible Virtual Volumes from RAID Devices are described that provide the improved functionality.

[032] **Virtual Volume Construction**

[033] A storage system which utilizes the present disclosure may initially construct a set of RAID Devices having various characteristics on a disk array. The
20 RAID Devices may be logically divided into units referred to herein as "pages," which may be many blocks in size. A typical page size may be 4,096 blocks. Although, in principle any page size from 1 block onwards could be used. However, page sizes generally comprise block numbers in the power of 2. These pages may be managed by Virtual Volume management software. Initially, all the pages from
25 each RAID Device may be marked as free. Pages may be dynamically allocated to Virtual Volumes on an as-needed basis. That is, pages may be allocated when it is determined that a given address is first written. Addresses that are read before being written can be given a default data value. The Virtual Volume management software may maintain the mapping between Virtual Volume addresses and pages
30 within the RAID Devices. It is noted that a given Virtual Volume may be constructed of pages from multiple RAID Devices, which may further have differing properties.

[034] Extending the size of a Virtual Volume constructed in this manner may be accomplished by increasing the range of addresses presented to the server.

The address-to-page mapping may continue with the same allocate-on-write strategy in both the previously available and extended address ranges.

[035] The performance and resiliency properties of a given Virtual Volume may be determined in large part by the aggregate behavior of the pages allocated to that Virtual Volume. The pages inherit their properties from the RAID Device and physical disk drives on which they are constructed. Thus, in one embodiment, page migration between RAID Devices may occur in order to modify properties of a Virtual Volume, other than size. "Migration," as used herein, may include allocating a new page, copying the previously written data from the old page to the new, updating the Virtual Volume mapping, and marking the old page as free. Traditionally, it may not be possible to convert the RAID Device properties (i.e., remap to a new RAID level, stripe size, etc.) and simultaneously leave the data in place.

[036] There are several independent parameters which may be modified to produce different Virtual Volume properties. Several of the scenarios are outlined in detail herein. However, the scenarios described in detail herein are exemplary of various embodiments of the present disclosure and are not limiting. The present disclosure, in some embodiments, may include simultaneous modification of any or all of these parameters.

20 [037] **RAID Parameter Modification**

[038] For purposes of illustration, a disk array 100 containing five disks 102, 104, 106, 108, 110 is shown in Figure 1A. It is recognized that any number of disks may be used in accordance with the various embodiments disclosed herein, and an exemplary five disk system has been randomly chosen for purposes of describing one embodiment. Initially, two RAID Devices, e.g., RAID Devices A 112 and B 114, may be constructed upon a disk array. The remaining space, if any, may be unallocated and unused. Again, it is recognized that any number of RAID Devices may be used in accordance with the various embodiments disclosed herein, and an exemplary two RAID Devices have been randomly chosen for purposes of describing one embodiment. Multiple Virtual Volumes may be constructed from the pages contained in the RAID Devices. If it is desired that the properties of a given Virtual Volume be modified, additional RAID Devices may be constructed in the remaining space and the appropriate pages migrated, as described previously.

[039] Figure 1B depicts an embodiment of a RAID configuration after creating a new RAID Device, e.g., RAID Device C 116, and shows the migration of data from RAID Device A 112. RAID Device C 116 may differ from RAID Device A 112 in RAID level, stripe size, or other RAID parameter, etc. In some
5 embodiments, there may be potential for improved performance by simply relocating to a RAID Device with the same parameters but a different location on the disk drives. For example, the performance of a disk drive may vary from the inside to the outside of the physical platter, and the time for head seeking may be reduced if all data is densely located.

10 [040] When the migration is complete, RAID Device A 112 may be deleted, leaving the example RAID configuration shown in Figure 1C.

[041] The exemplary RAID reconfiguration from that of Figure 1A to that of Figure 1C also demonstrates the ability to move portions of Virtual Volumes. That is, in one embodiment, rather than moving an entire Volume, portions of one or
15 more Virtual Volumes may be migrated. This may be accomplished because a single Virtual Volume may be allocated across a plurality of RAID Devices. Similarly, the example configuration demonstrates the ability to move groups of Virtual Volumes since pages migrated from one RAID Device to another RAID Device may be allocated to a plurality of Virtual Volumes.

20 [042] Adding Disk Drives

[043] Another embodiment having a disk array 200 containing five disks 202, 204, 206, 208, 210 is illustrated in Figure 2A, where two additional disk drives 212, 214 have been added to an existing configuration. It is recognized that any number of disks may be used in accordance with the various embodiments disclosed
25 herein, and an exemplary five disk system has been randomly chosen for purposes of describing one embodiment. Similarly, it is recognized that any number of disks may be added in accordance with the various embodiments disclosed herein, and an exemplary two additional disks have been randomly chosen for purposes of describing one embodiment. In some embodiments, it may be desirable to
30 reconfigure the system and spread the RAID Devices across all seven disks. However, it is recognized that the reconfigured RAID Devices do not need to be spread across all available disks. In an embodiment where the RAID Devices are spread across a plurality of disks, the total throughput of the system can be increased by utilizing more hardware in parallel. Additionally, RAID Device layout

constraints may result in more efficient use of the additional disks. In particular, RAID 5 typically may require a minimum number of independent disks in order to provide resiliency. Commonly encountered RAID 5 implementations may require a minimum of five disks. Thus, it may be desirable to migrate the pages from both
5 RAID Devices A 216 and B 218, for example, to suitable replacements that span all seven disks. A possible sequence for reconfiguration is shown in Figures 2B-D.

[044] In this sequence, the wider RAID Device C 220 may be created and data from RAID Device A 216 may be migrated to RAID Device C 220. RAID Device A 216 may then be deleted, and RAID Device D 222 may be created. RAID
10 Device D 222 may be used to relocate the data previously contained in RAID Device B 218.

[045] In doing so, the only extra space needed on the original disk drives may be that required to create RAID Device C 220. In one embodiment of the example illustration, in the case wherein no other RAID parameter changes, each
15 extent of RAID Device C 220 may be $5/7$ the size of the extent size (i.e., RAID Device C is spread among 5 initial drives + 2 additional drives) used in constructing RAID Device A 216.

[046] It is noted that the process may be entirely reversible and can be used to remove one or more disk drives from a system, such as, for example, if it was
20 desired that disks 212 and 214 be removed from the example configuration of Figure 2D. Similarly, multiple initial RAID Devices may be migrated to a single RAID Device, or a fewer number of RAID Devices (see e.g., Figure 3C). Furthermore, a single initial RAID Device may be migrated to a plurality of new RAID Devices (see e.g., Figure 3B).

25 **[047]** The previous example of one embodiment described with reference to Figures 2A-D demonstrates the ability to migrate data across additional disks when unused space exists on the original disk set. In some embodiments, however, there may be insufficient disk space to migrate, remove, etc. a RAID Device. Nonetheless, it may be possible to migrate data to additional disks. In such a case,
30 disk space may be reallocated to provide the extra space needed to perform the move. If the Replacement Score, described in detail below, of a RAID Device is higher than the initial Score, a permanent RAID Device of equal size may be allocated. No additional decisions may be required. If the Overlay Score, described in detail below, of a RAID Device is higher than the initial Score, temporary space

may be used. This process is detailed for one embodiment having a disk array 300 containing four disks 302, 304, 306, 308 with reference to Figures 3A-E, where a multi-step migration is used. Three disks 310, 312, 314 have been added to the disk array 300. It is recognized that any number of disks may be used in accordance with the various embodiments disclosed herein, and an exemplary four disk system has been randomly chosen for purposes of describing one embodiment. Similarly, it is recognized that any number of disks may be added in accordance with the various embodiments disclosed herein, and an exemplary three additional disks have been randomly chosen for purposes of describing one embodiment.

10 **[048]** The strategy for reconfiguring the system shown in Figure 3A to make use of all the available disk drives in the array may include creating a temporary RAID Device or temporary RAID Devices and migrating the data from RAID Device C 320, for example, away from the original disk drives to temporary RAID Devices D 322 and E 324, for example. Temporary RAID Devices may be used in such cases where the original disk set is at, or near, capacity. In alternate embodiments, the temporary space may not need to be allocated as RAID Devices and may be used in any manner known in the art for suitably holding data. Similarly, although two temporary RAID Devices D 322 and E 324 are illustrated, it is recognized that a fewer or greater number of temporary RAID Devices may be utilized.

20 **[049]** In one embodiment, a data progression process may manage the movement of data between the initial RAID Device and the temporary RAID Device(s), or in other cases, new permanent RAID Device(s). In further embodiments, Restriping may attempt to use the same RAID level, if available. In other embodiments, Restriping may move the data to a different RAID level.

25 **[050]** The size of a temporary RAID Device may depend on the initial RAID Device size and available space within a page pool. The size of the temporary RAID Device may provide sufficient space, such that when the initial RAID Device is deleted, the page pool may continue to operate normally and not allocate more space. The page pool may allocate more space at a configured threshold based on the size of the page pool.

30 **[051]** Once the data has been migrated away from RAID Device C 320, it can be deleted, providing space for a new RAID Device spanning all of the disk drives, e.g., RAID Device X 326. Deleting RAID Device C 320 may return the disk

space RAID Device C 320 consumed to the free space on the disk. At this point, a disk manager may combine adjacent free space allocation into a single larger allocation to reduce fragmentation. Deleting a RAID Device may create free space across a larger number of disks than was previously available. A RAID Device with
5 a higher Score can be created from this free space slice.

[052] After the initial RAID Device C 320 is deleted, Restriping may create a replacement RAID Device X 326, as shown in Figure 3C. In one embodiment, replacement RAID Device X 326 may use as many disks as possible to maximize the benefits of virtualization. Restriping may attempt to allocate a RAID Device
10 larger than the initial RAID Device. In a further embodiment, Restriping may do this using a calculation of the Replacement or Overlay Score divided by the initial Score multiplied by the size of the initial RAID Device. This may create a RAID Device that uses the same amount of disk space per disk as before and may reduce fragmentation of the disk.

[053] By judiciously limiting the size of the initial RAID Devices, e.g., RAID Devices A 316, B 318, and C 320, it may be possible to create RAID Device X 326 such that it can hold all the data from RAID Devices B 318 & E 324, for example, allowing the process to continue until the final configuration is achieved in Figure 3E. That is, RAID Device Y 328 may be created, RAID Devices A 316 and
20 D 322 may be migrated to RAID Device Y 328, and RAID Devices A 316 and D 322 may be deleted.

[054] If a temporary RAID Device or temporary RAID Devices, e.g., RAID Devices D 322 and E 324, were created and marked as temporary, the RAID Devices may be marked for removal, as shown in Figure 3C-E. In an embodiment,
25 as a part of each cycle, the temporary RAID Devices may be removed. As larger replacement RAID Devices are created, the amount of temporary space needed may decline. It is noted again that in some embodiments, a RAID Device migration may not require the allocation of temporary space to migrate or remove the data.

[055] In one embodiment of Restriping, removal of the temporary RAID
30 Devices may use a subset of the steps used for migration or removal of the initial RAID Device, such as the movement of data and deletion of the temporary RAID Devices.

[056] In one embodiment, if the Score of a temporary RAID Device exceeds the Score of the initial RAID Device, the temporary RAID Device may be

considered a permanent RAID Device. That is, it may not be automatically deleted as a part of the process to move a RAID Device. In further embodiments, the temporary RAID Device may be kept only if it has a sufficiently higher Score than the initial RAID Device.

5 **[057]** Restripping may involve a number of further steps to remove an original low-scoring RAID Device and replace it with a new higher-scoring RAID Device. For example, Restripping may account for the possibility that the disks in the system are full, and have no space for another RAID Device. Restripping may trim excess space before attempting to restripe a RAID Device. Trimming excess space
10 may free up additional disk space and increase the success rate of Restripping.

[058] In some embodiments, Restripping may reach a deadlock. For example, the size of the temporary space may consume a portion of the space needed to move the initial RAID Device. If it becomes impossible to remove a RAID Device because all pages cannot be freed, the RAID Device may be marked as
15 failed, and Restripping may move on to the next RAID Device that can or should be migrated.

[059] With reference to Figure 4, a flow chart of one embodiment of a process 400 of Restripping is described. It is recognized that Figure 4 illustrates one embodiment, and various alternative embodiments and processes may be used in
20 accordance with the present disclosure. First, as shown in steps 402 and 404, Restripping may determine whether there is a RAID Device that should or can be migrated, removed, etc. In one embodiment, Restripping may check all of the RAID Devices within a system and select the smallest RAID Device with the lowest relative Score. In other embodiments, Restripping may select other RAID Devices,
25 and Restripping, as described herein, is not limited to selecting the smallest RAID Device and/or the RAID Device with the lowest Score. Generally, however, if movement of the smallest RAID Device fails, a larger RAID Device may likely not succeed either. In an embodiment, the lowest scoring RAID Device may be determined by dividing the Replacement or Overlay Score by the initial Score.
30 Other methods of determining the lowest scoring RAID Device are in accordance with the present disclosure, including using solely the initial Score of the RAID Devices.

[060] In addition to identifying RAID Devices for migration or removal, as shown in Figure 4, RAID Device movement may include a plurality of steps to

optimize the RAID configuration, such as, but not limited to, allocating temporary space (step 406), moving data (step 408), deleting the original RAID Device (step 410), allocating a new RAID Device (step 412), and/or deleting the temporary RAID Device (steps 414 and 416). The foregoing listing of additional steps is exemplary and RAID Device movement need not require each of the listed steps, and in some embodiments, may include further or different steps than those listed. For example, in some embodiments, temporary space may not be used, and therefore, may not be allocated.

[061] In some embodiments, Restriping may limit the movements of RAID Devices. For example, to avoid thrashing the system, Restriping may not need to absolutely maximize the Score of a RAID Device. Restriping may also mark failed RAID Devices so as not to retry them.

[062] Restriping may recognize new disks, create new RAID devices which utilize the additional spaces, and move the data accordingly. After the process is complete, user data and free space may be distributed across the total disk drives, including the initial disks and the additional disks. It is noted that Restriping may replace RAID Devices rather than extend them. It is appreciated that the positioning of free space and user allocations on any given disk may be arbitrary, and the arrangements shown in Figures 1-3, as well as the remaining Figure 5, are for illustration purposes.

[063] Selection of RAID Device for Restriping

[064] In one embodiment, as previously discussed, Restriping may handle:

- Adding Drives - When additional drives are added to the disk drive system, Restriping may identify RAID Devices that use a sub-optimal number of drives. New RAID Devices may be created and the data may be moved. The original RAID Devices may be eliminated.
- Removing Drives – Restriping may detect when disk drives have been marked for removal. RAID Devices which reside on these drives may become candidates for removal, which may be accomplished in a substantially similar manner as for Adding Drives.
- RAID Parameter Changes - RAID Parameters, such as RAID level, number of disks within a stripe, and extent size, may be altered by the user to improve performance. Restriping may compare the desired

parameters against the initial parameters and select nonoptimal RAID Devices for migration and/or removal.

[065] In some embodiments, including embodiments having larger, more complicated systems, it may not be obvious which set of migration operations should be used in order to obtain the desired final configuration or if it is possible to get from the initial configuration to the final desired configuration within the existing resources. In one embodiment, a scoring and optimization technique may be used to select the particular RAID Device for removal and replacement. The scoring function, in an exemplary embodiment, may employ one or more of the following properties:

- RAID Devices which span more disk drives may be preferred.
- RAID Devices which are constructed on a homogeneous set of disk drives may be preferred. Disk drives may be categorized in order to support this function and need not be identical to belong to the same category.
- RAID Devices which match the parameters (RAID level, stripe size, disk region, etc.) of the desired final configuration may be preferred.
- RAID Devices which place redundant data on physically disparate disk drives may be preferred. An example may include disk drives in separate enclosures, on separate communication paths, or having independent power sources.

[066] In another embodiment, Restriping may be divided into three components, such as scoring, examining, and moving. RAID Device scoring may be used to determine the quality of a given RAID Device based on requested parameters and disk space available. In one embodiment, scoring may generate three values. Restriping may provide a Score for an initial RAID Device and the scores of two possible alternative RAID Devices, referred to herein as the Replacement and Overlay Scores. Details of each score for one embodiment are described below:

- Score - The score of the RAID Device in its present state. In one possible embodiment, the Score may be the number of disks used by the RAID Device less fragmentation and parameter issues. See e.g.,

Table 1. In alternative embodiments, a Score may be calculated in any suitable manner.

- 5 • Replacement Score - The maximum score of a RAID Device that could be constructed from existing free space. The Replacement Score may be higher than, lower than, or equal to the Score of the RAID Device.
- 10 • Overlay Score - The maximum score of a RAID Device if the current RAID Device is removed. The Overlay Score may be higher than, lower than, or equal to the Score of the RAID Device. In some

[067] With respect to the Replacement and Overlay Scores, the user accessible blocks for the RAID Device may remain the same as the number of disks
15 changes. The three scores may provide the input parameters to develop a strategy for migrating from lower to higher scoring RAID Devices. In a particular embodiment, if the Replacement Score is higher than the initial Score, a straightforward migration like that described in Figures 1 and 2 may be possible. If the Overlay Score is better than the initial Score, and if sufficient free space is
20 available for a temporary RAID Device or temporary RAID Devices, then a migration strategy outlined in Figure 3 may be possible.

[068] In one embodiment, factors used to determine the Scores may include one or more of the following:

- 25 • Disk Folder - If a RAID Device uses a disk outside of the specified folder, the score of the RAID Device may be lowered. This situation may occur due to administrative action or during sparing, i.e., wherein spare disks may be supplied.
- 30 • Disk Type - If a RAID Device resides on a disk of the wrong type, the score of the RAID Device may be lowered. This situation may occur if a disk fails and a spare of the appropriate type is not available. In such a case, redundancy may be maintained over 'type purity,' and a spare of another disk type may be used.

- Number of Disks Used - In general, wider RAID Devices (e.g., RAID Devices spanning a relatively larger number of disk drives) may be preferred, and the RAID Devices may be given a higher score. In a further embodiment, a maximum width may be considered in order to establish more independent fault domains.
- Number of Disks Available - This may be used to determine the Replacement and Overlay Scores of the RAID Device. In one embodiment, disks, excluding spare disks, may be checked for sufficient space to allocate a replacement RAID Device. In other embodiments, spare disks may be included in the determination of the Scores. If space exists, the Replacement and Overlay Score may be increased.
- Disk Fragmentation – If free disk space exists around the RAID Device, the RAID Device score may be lowered. This may be an indication to Restriping that by migrating the RAID Device, disk fragmentation may be reduced.
- RAID Level - Whether the desired RAID level matches the initial RAID level. This may include the number of disks within a stripe.
- RAID Extent Size - Whether the extent size of the desired RAID Device, i.e., blocks per disk, matches the extent size of the initial RAID Device. In one embodiment, a determination based on RAID extent size may only lower the score of the RAID Device.

[069] Table 1 illustrates an example embodiment of scoring factors that may be used. As illustrated in Table 1, the variables may include Disks In Class, Disks In Folder, RAID Level, RAID Repeat Factor, RAID Extent Size, and RAID Drives in Stripe. Disks In Class, as used in the example scoring factors, may be determined by the equation:

$$(\text{DisksInClass} - 3 * \text{DisksOutOfClass}) * \text{DisksInClassConstant}$$

where DisksInClass may be the number of disks used by the RAID Device that are of the proper class, DisksOutOfClass may be the number of disks used by the RAID Device that are not of the proper class, and DisksInClassConstant may be a multiplicative constant value. Disk classes may include, but are not limited to, 15K

FC, 10K FC, SATA, etc. For example, if a RAID Device was supposed to use 10K FC disks, but included two SATA disks, the value for DisksOutOfClass would be two. Disks In Folder, as used in the example scoring factors, may be determined by the equation:

$$5 \quad (\text{DisksInFolder} - 3 * \text{DisksOutOfFolder}) * \text{DisksInFolderConstant}$$

where DisksInFolder may be the number of disks used by the RAID Device that are in the proper folder of disks, DisksOutOfFolder may be the number of disks used by the RAID Device that are not in the proper folder of disks, and DisksInFolderConstant may be a multiplicative constant value. Disk folders may
10 organize which disks can be used by RAID Devices. Disks may be moved into, and out of, folder objects at any time to change their usage. RAID Level, as used in the example scoring factors, may be zero if the disk is an undesired RAID level. RAID Repeat Factor, RAID Extent Size, and RAID Drives in Stripe may be a computed
15 score of each divided by a factor of two. It is recognized that Table 1 illustrates one embodiment of example scoring factors and one embodiment of how the scoring factors are calculated and used. The example illustrated in Table 1 is for illustration purposes only and is not limiting. Any scoring factors, or group of scoring factors, may be used with the various embodiments disclosed herein. Furthermore, the
20 scoring factors, or group of scoring factors, may be calculated or used in any suitable manner.

Table 1: Example RAID Scoring Factors

Variable	Score
Disk In Class	$(\text{DisksInClass} - 3 * \text{DisksOutOfClass}) * \text{DisksInClassContant}$
Disks In Folder	$(\text{DiskInFolder} - 3 * \text{DisksOutOfFolder}) * \text{DisksInFolderConstant}$
RAID Level	Zero if wrong type
RAID Repeat Factor	Computed Score divided by two
RAID Extent Size	Computed Score divided by two
RAID Drives in Stripe	Computed Score divided by two

[070] In a further embodiment, Restriping may examine the Scores of the RAID Devices to determine which, if any, RAID Devices may be moved.

5 Restriping may move RAID Devices with a score that is lower than either the Replacement or Overlay Scores. That is, in one embodiment, if the Replacement and/or Overlay Score is greater than the initial RAID Device Score the RAID Device may be a candidate to move. In other embodiments, the initial RAID Devices may be selected for migration by any other means, including situations

10 wherein the initial RAID Device Score is higher than the Replacement and Overlay Scores or by manual selection by a user, etc. Restriping may also determine that no RAID Devices should be moved. In a further embodiment, Restriping may pick a single RAID Device from the available RAID Devices to migrate.

[071] If Restriping identifies a RAID Device to move, migration of the

15 RAID Device may occur. In one embodiment, migration may include determining necessary temporary space, movement of data from the RAID Device, cleanup of the initial RAID Device, and elimination of the temporary space. In another embodiment, a dynamic block architecture page pool may use the RAID Devices

and handle the movement of data from lower scoring to higher scoring RAID Devices.

[072] In another embodiment, Restripping may further reevaluate the scores of all RAID Devices after every RAID Device migration since the reallocation of disk space may change the Scores of other RAID Devices. In a further embodiment, the scores of all the RAID Devices may be periodically computed. In some embodiments, Restripping may continually compute the Scores of the RAID Devices. In yet another embodiment, the largest gain in score may be used to select a RAID Device for removal and replacement. A hysteresis mechanism may be used to prevent the process from becoming cyclic.

[073] RAID Device scoring may also handle different-sized disk drives. Figure 5 illustrates an example configuration 500 with different-sized disks 502, 504, 506, 508. Table 2 illustrates an example RAID Device scoring, for the configuration shown in Figure 5, including the scoring information for the RAID Devices 510, 512, 514, 516, 518 based on the configuration 500. Relative numbers are used for simplicity. Although Table 2 illustrates scores relating to RAID Device candidates for migration, Table 2 is not limiting and any scoring combination may result in marking a RAID Device for migration or no migration.

Table 2. RAID Example Scoring

Device	Score	Replacement	Overlay	Restripe
P 510	4	0	4	No, at maximum
Q 512	4	0	4	No, at maximum
R 514	2	2	4	Yes, 50% of maximum
S 516	3	1	4	Yes, 75% of maximum
T 518	2	2	2	No, at maximum no space on smaller disks

[074] From the above description and drawings, it will be understood by those of ordinary skill in the art that the particular embodiments shown and described are for purposes of illustration only and are not intended to limit the scope of the present invention. Those of ordinary skill in the art will recognize that the
5 present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. References to details of particular embodiments are not intended to limit the scope of the invention.

[075] Although the present invention has been described with reference to preferred embodiments, persons skilled in the art will recognize that changes may be
10 made in form and detail without departing from the spirit and scope of the invention.

Claims

We claim:

1. A method of RAID restriping in a disk drive system, comprising the steps of:
 - selecting a RAID device for migration based on at least one score calculated for each of a plurality of RAID devices;
 - creating at least one alternate RAID device;
 - moving data stored at the selected RAID device to the at least one alternate RAID device; and
 - removing the selected RAID device.
2. The method of claim 1, wherein the at least one score comprises an initial score and a replacement score, wherein the initial score relates to the RAID device and the replacement score relates to at least one first hypothetical RAID device located in available disk space.
3. The method of claim 2, wherein the at least one score further comprises an overlay score relating to at least one second hypothetical RAID device located in a combination of available disk space and the disk space taken up by the RAID device.
4. The method of claim 3, wherein the initial score, replacement score, and overlay score are each calculated based on at least one of factors comprising the RAID level, RAID stripe size, RAID extent size, disk category, location on disk, disk enclosure, disk enclosure power supply, and communication path to the disk.
5. The method of claim 4, wherein the factors have varying weights for use in the calculation.
6. The method of claim 5, wherein the initial score, replacement score, and overlay score are each calculated using the same factors.

7. The method of claim 4, wherein selecting a RAID device for migration based on at least one score calculated for each of a plurality of RAID devices comprises selecting the RAID device if at least one of the replacement score or overlay score is better than the initial score.
8. The method of claim 7, wherein the alternate RAID device is based on one of the first and second hypothetical RAID devices.
9. The method of claim 1, wherein the steps of selecting a RAID device for migration, creating at least one alternate RAID device, moving data, and removing the selected RAID device are done automatically without manual intervention.
10. The method of claim 9, wherein the steps are performed without loss of server data access to the disk drive system and compromised resiliency of the data.
11. The method of claim 9, wherein the steps are performed at least one of periodically, continuously, after every RAID device migration, upon addition of disk drives, and before removal of disk drives.
12. The method of claim 1, wherein the steps of selecting a RAID device for migration, creating at least one alternate RAID device, moving data, and removing the selected RAID device are done manually.
13. The method of claim 1, wherein moving data stored at the selected RAID device to the at least one alternate RAID device further comprises creating at least one temporary RAID device.
14. The method of claim 13, further comprising moving data stored at the selected RAID device to the at least one temporary RAID device and then from the temporary RAID device to the at least one alternate RAID device.
15. A disk drive system, comprising:
a RAID subsystem; and

a disk manager having at least one disk storage system controller configured to automatically:

calculate at least one score for each RAID device of a plurality of RAID devices of the RAID subsystem;

select a RAID device from the plurality of RAID devices based on the relative at least one scores of the plurality of RAID devices.

create an alternate RAID device;

move at least a portion of the data stored at the selected RAID device to the alternate RAID device; and

remove the selected RAID device.

16. The disk drive system of claim 15, wherein the at least one score comprises an initial score for the RAID device, a replacement score related to a first hypothetical RAID device, and an overlay score related to a second hypothetical RAID device.

17. The disk drive system of claim 16, wherein the alternate RAID device is based on one of the first and second hypothetical RAID devices.

18. The disk drive system of claim 15, wherein the disk drive system comprises storage space from at least one of a plurality of RAID levels including RAID-0, RAID-1, RAID-5, and RAID-10.

19. The system of claim 18, further comprising RAID levels including RAID-3, RAID-4, RAID-6, and RAID-7.

20. A disk drive system capable of RAID restriping, comprising:
means for selecting a RAID device for migration based on at least one score calculated for each of a plurality of RAID devices;
means for creating at least one alternate RAID device;

means for moving data stored at the selected RAID device to the at least one alternate RAID device; and

means for removing the selected RAID device.

21. A method for reallocating data, comprising:
 - determining a score for a current position of data and a score for an alternate position the data;
 - comparing the score for the current position of the data and the score for the alternate position of the data; and
 - moving the data from the current position to the alternate position when the score for the alternate position of the data is better than the score for the current position of the data.

FIG. 1A

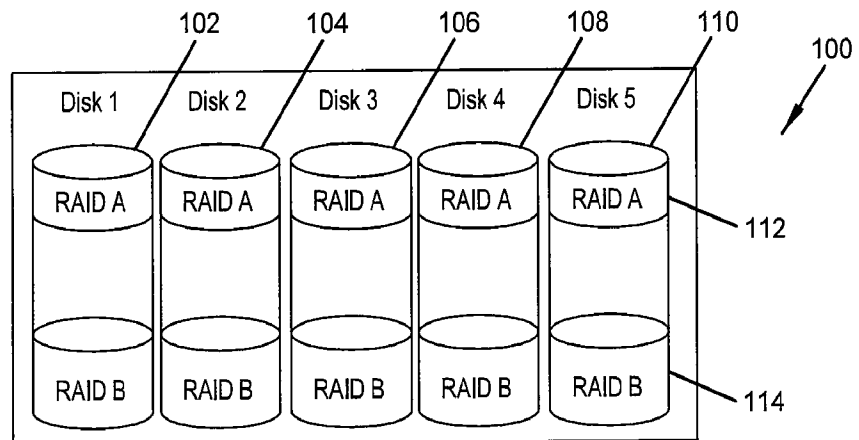


FIG. 1B

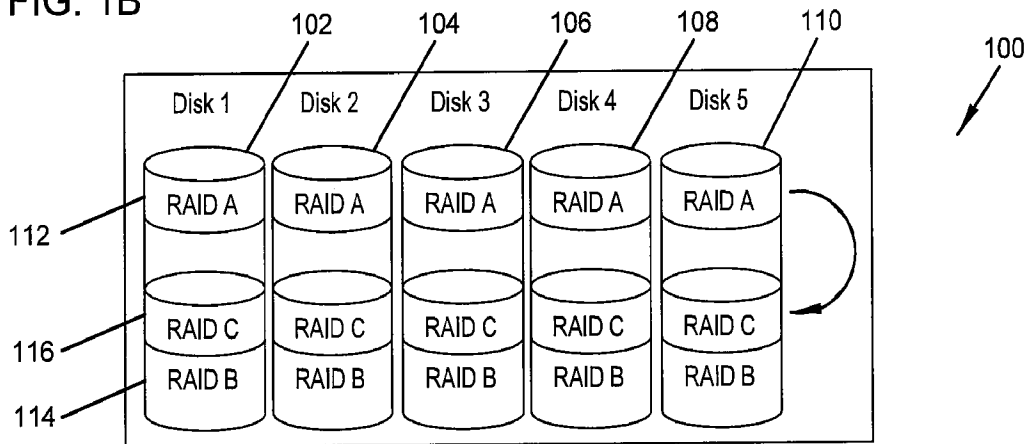


FIG. 1C

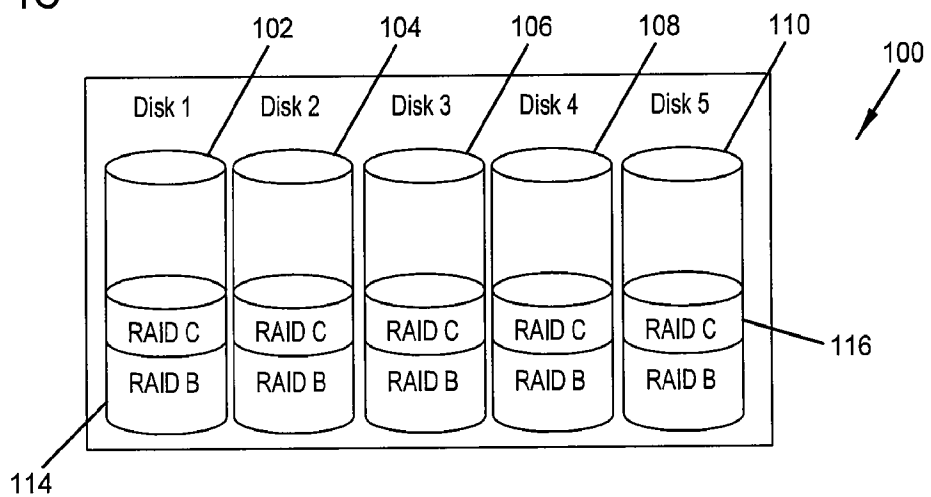


FIG. 2A

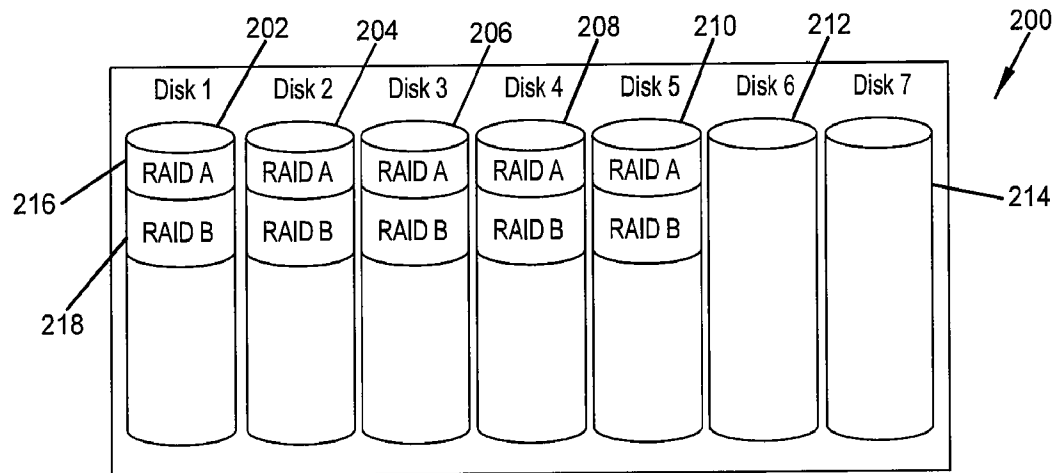


FIG. 2B

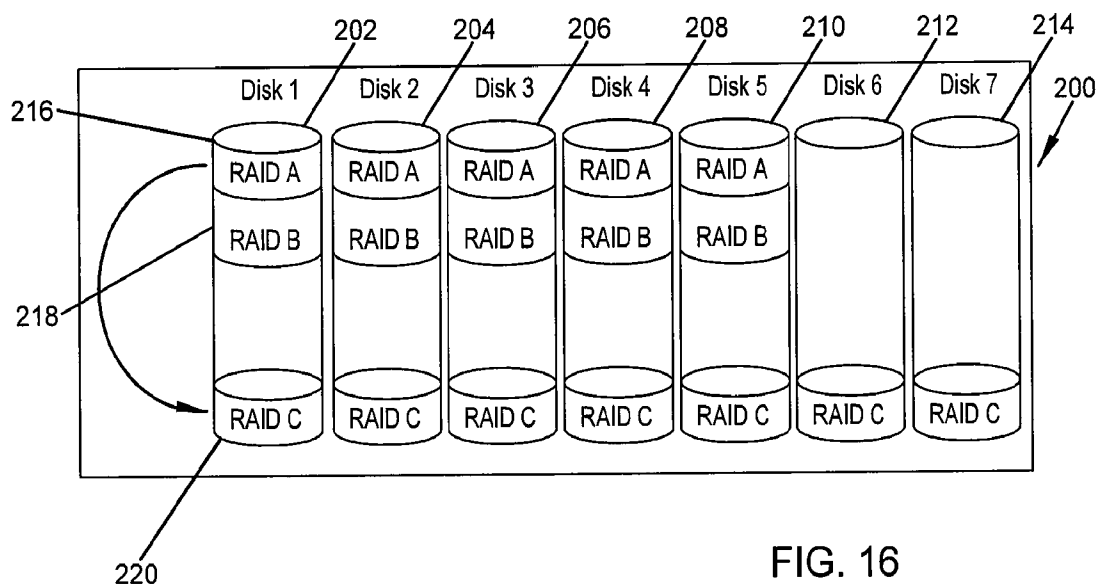


FIG. 16

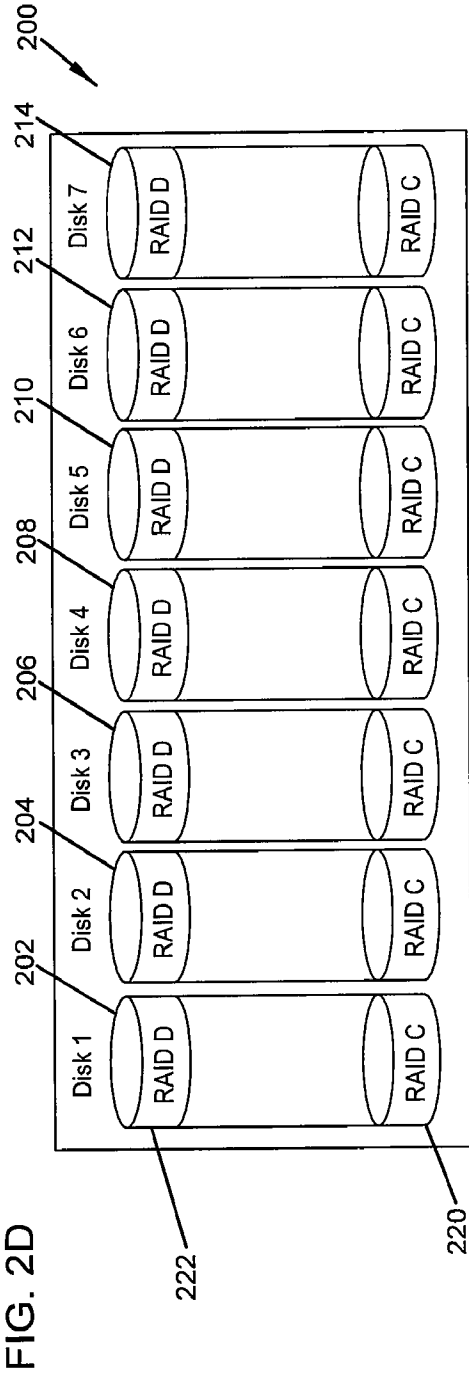
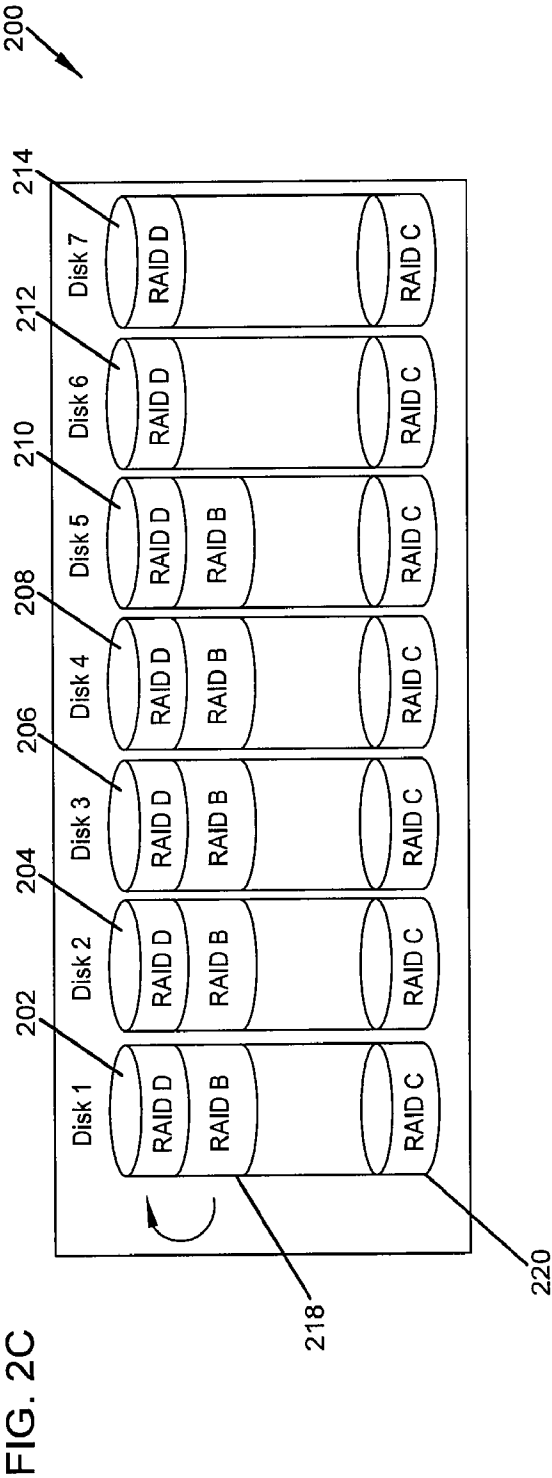


FIG. 3A

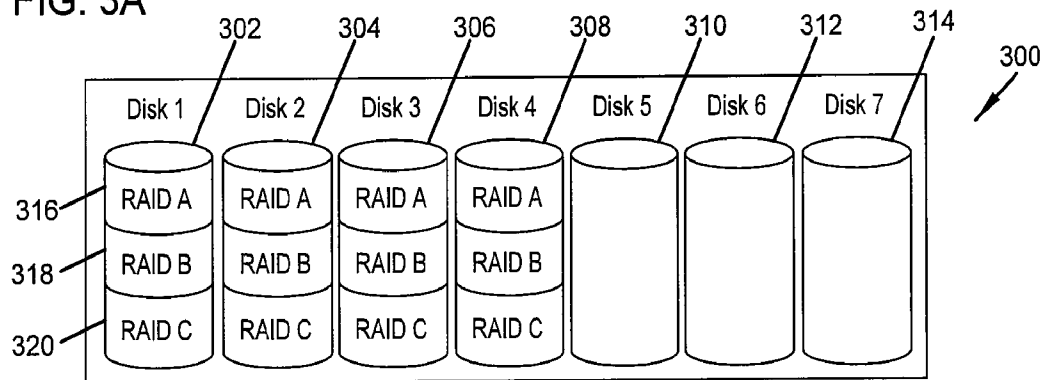


FIG. 3B

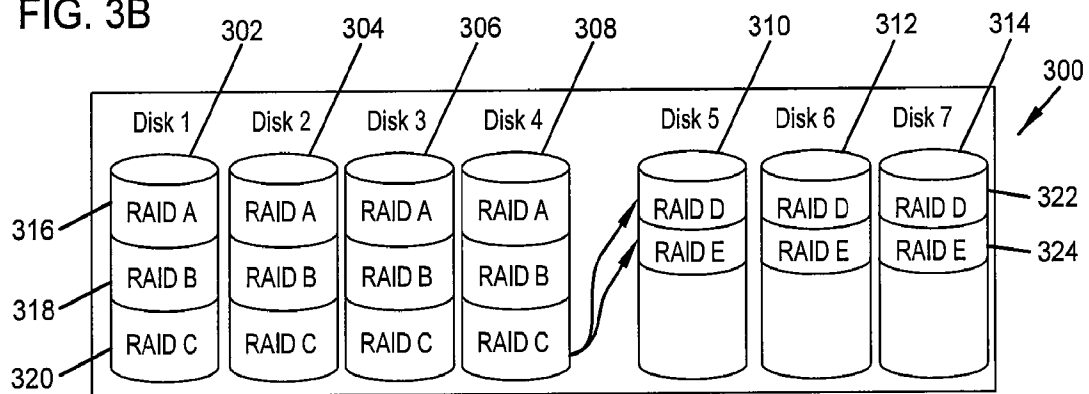


FIG. 3C

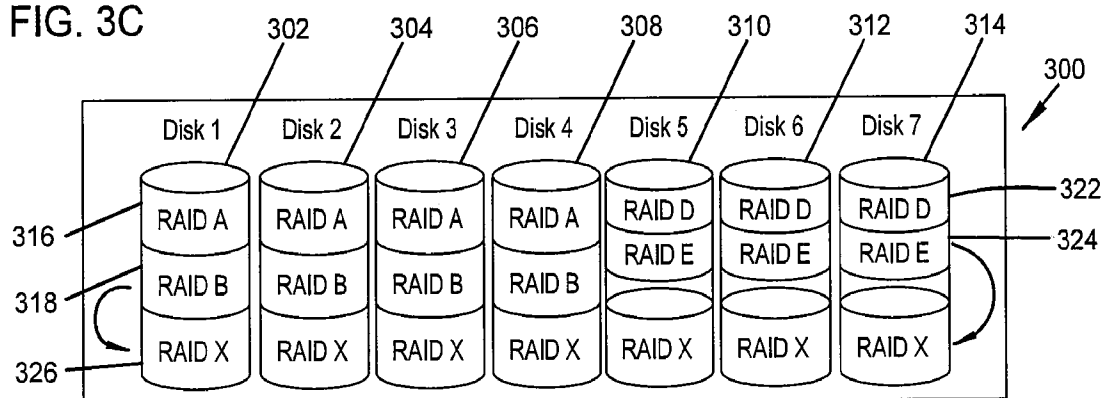


FIG. 3D

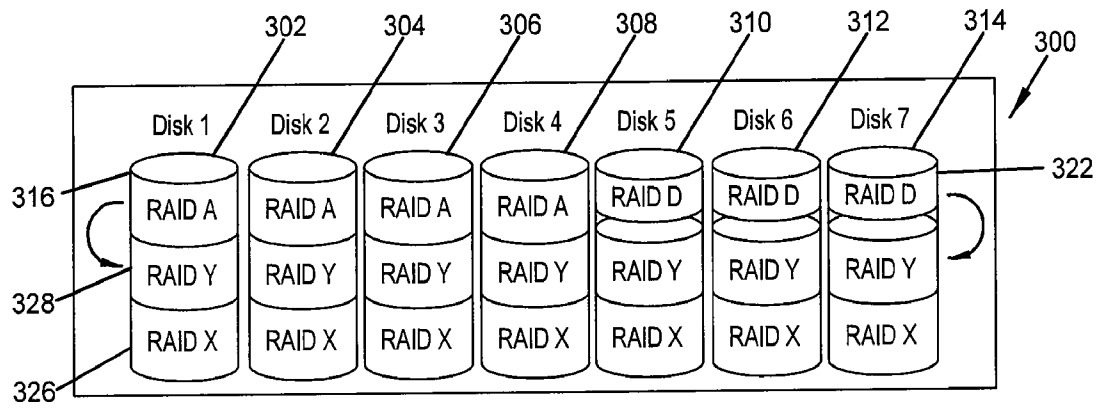


FIG. 3E

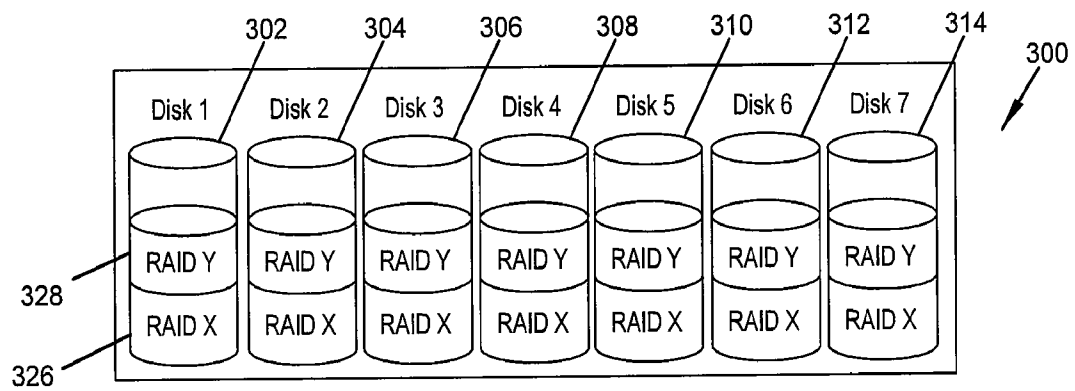


FIG. 4

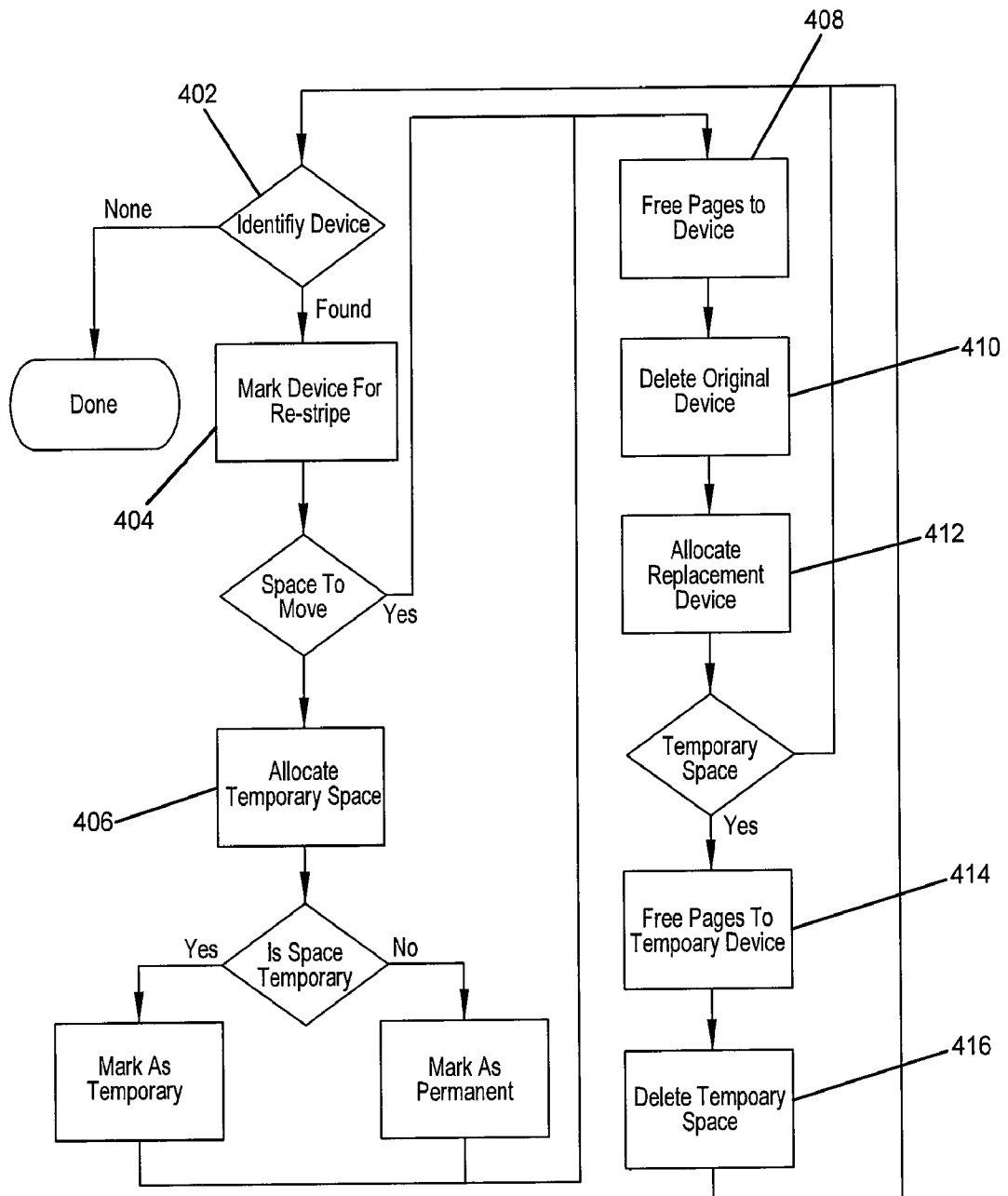


FIG. 5

