Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2532998 C 2011/03/08

(11)(21) 2 532 998

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2004/07/07
(87) Date publication PCT/PCT Publication Date: 2005/01/20
(45) Date de délivrance/Issue Date: 2011/03/08
(85) Entrée phase nationale/National Entry: 2006/01/16
(86) N° demande PCT/PCT Application No.: EP 2004/051383
(87) N° publication PCT/PCT Publication No.: 2005/006198
(30) Priorité/Priority: 2003/07/14 (US10/619,649)

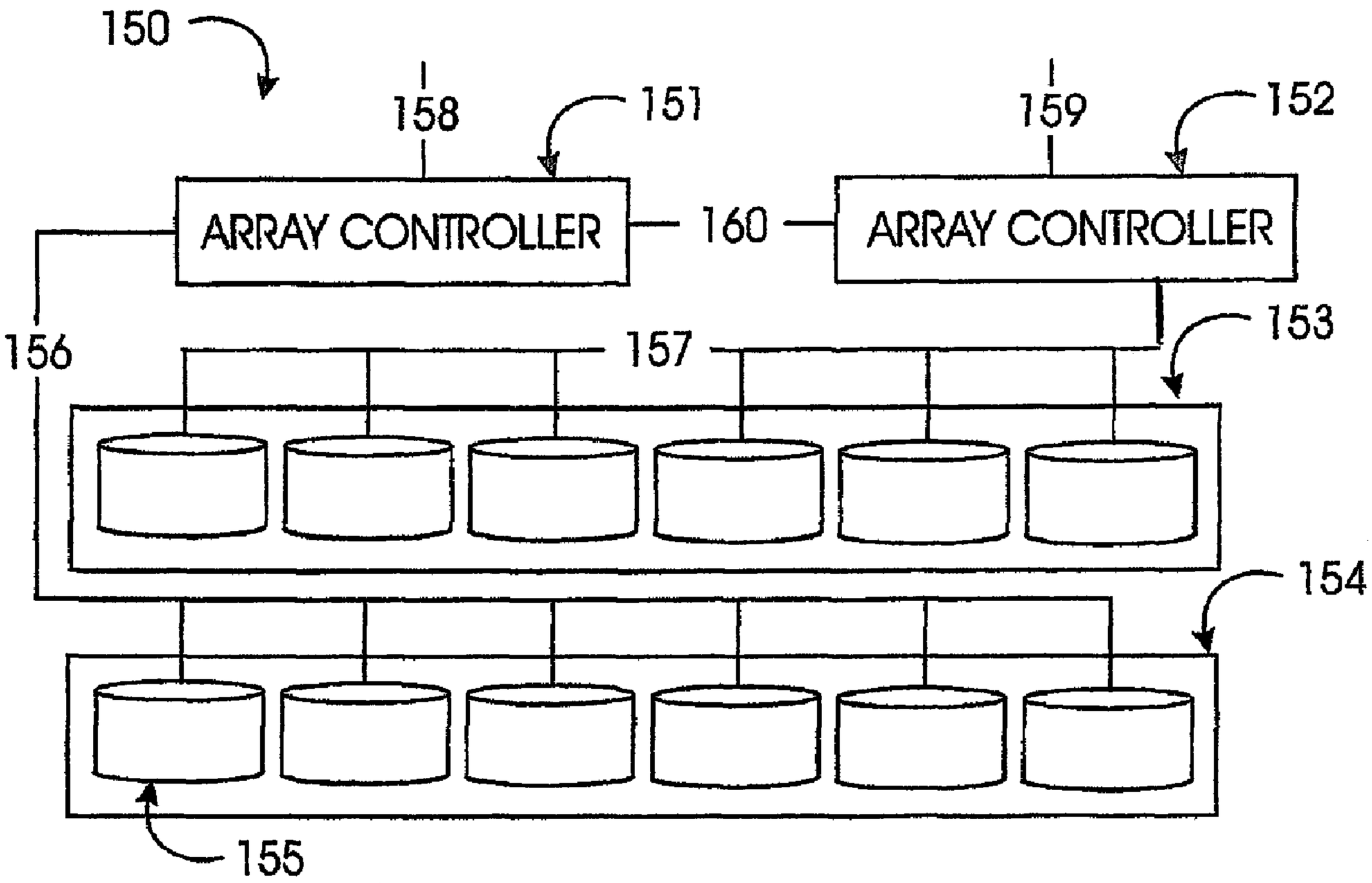
(51) Cl.Int./Int.Cl. *G06F 12/16* (2006.01)

(72) Inventeurs/Inventors:
HETZLER, STEVEN ROBERT, US;
SMITH, DANIEL FELIX, US

(73) Propriétaire/Owner:
INTERNATIONAL BUSINESS MACHINES CORPORATION, US

(74) Agent: WANG, PETER

(54) Titre : REDONDANCE DANS UN SYSTEME DE MEMOIRE MATRICIELLE
(54) Title: REDUNDANCY IN ARRAY STORAGE SYSTEM



(57) Abrégé/Abstract:
Error tolerance is increased for a storage system having a plurality of arrays by making local redundancy in a selected array globally available throughout the storage system. To achieve the increased error tolerance, a donor array is selected from the plurality of arrays when the difference between a minimum distance of the donor array and a minimum distance of a recipient array is greater or equal to 2. A donor storage unit is selected in the donor array and recipient information is then rebuilt from the recipient array on the selected storage unit. The selected storage unit is indicated to the donor array as having been donated before the lost

CA 2532998 C 2011/03/08

(11)(21) 2 532 998

(13) C

(57) Abrégé(suite)/Abstract(continued):
information is rebuilt on the selected storage unit. Preferably, the minimum Hamming distance of the recipient array is d ≥ 2 before the donor array is selected from the plurality of arrays.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 January 2005 (20.01.2005)

PCT

(10) International Publication Number
WO 2005/006198 A2

(51) International Patent Classification[7]: G06F 12/16

(21) International Application Number:
PCT/EP2004/051383

(22) International Filing Date: 7 July 2004 (07.07.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/619,649 14 July 2003 (14.07.2003) US

(71) Applicant *(for all designated States except US)*: **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, New York 10504 (US).

(71) Applicant *(for MG only)*: **IBM UNITED KINGDOM LIMITED** [GB/GB]; PO Box 41, North Harbour, Portsmouth Hampshire PO6 3AU (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **HETZLER, Steven, Robert** [US/US]; 246 Eleanor Avenue, Los Altos, California 94022 (US). **SMITH, Daniel, Felix** [GB/US]; 7048 Via Blanca, San Jose, California 95139 (US).

(74) Agent: **LITHERLAND, David, Peter**; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Winchester, Hampshire SO21 2JN (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, 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, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, 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, IT, LU, MC, 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.

WO 2005/006198 A2

(54) Title: REDUNDANCY IN ARRAY STORAGE SYSTEM

(57) Abstract: Error tolerance is increased for a storage system having a plurality of arrays by making local redundancy in a selected array globally available throughout the storage system. To achieve the increased error tolerance, a donor array is selected from the plurality of arrays when the difference between a minimum distance of the donor array and a minimum distance of a recipient array is greater or equal to 2. A donor storage unit is selected in the donor array and recipient information is then rebuilt from the recipient array on the selected storage unit. The selected storage unit is indicated to the donor array as having been donated before the lost information is rebuilt on the selected storage unit. Preferably, the minimum Hamming distance of the recipient array is $d \geq 2$ before the donor array is selected from the plurality of arrays.

Description

REDUNDANCY IN ARRAY STORAGE SYSTEM

Technical Field

[001] The present invention relates to storage systems. In particular, the present invention relates to a method for configuring a storage system comprising a plurality of arrays of storage units and thereby increasing the number of storage-unit failures that the storage system can tolerate without loss of data stored in the system.

Background Art

[002] The following definitions are used herein and are offered for purposes of illustration and not limitation:

[003] An "element" is a block of data on a storage unit.

[004] A "base array" is a set of elements that comprise an array unit for an ECC.

[005] An "array" is a set of storage units that holds one or more base arrays.

[006] A "stripe" is a base array within an array.

[007] n is the number of data units in the base array.

[008] r is the number of redundant units in the base array.

[009] m is the number of storage units in the array.

[010] d is the minimum Hamming distance of the array.

[011] D is the minimum Hamming distance of the storage system.

[012] Large storage systems typically comprise multiple separate arrays of storage units. Each array is conventionally protected against a certain number of storage-unit failures (also called erasures) by an Erasure (or Error) Correcting Code (ECC) in, for example, a mirroring configuration or a RAID 5 (Redundant Array of Independent Disks Level 5) configuration. ECC codes provide redundant storage units that are local to each array, and increase reliability for a storage system by handling unit failures that are localized to a subset of the arrays.

[013] Storage capacity of Hard Disk Drive (HDD) -based storage systems is increasing faster than improvements in component reliability. Consequently, minimum Hamming distance d = 2 schemes, such as RAID 5 and mirroring techniques, no longer provide adequate protection for sufficient reliability at the system level. Alternative designs, such as RAID 6 (dual parity) at distance d = 3, double mirroring at distance d = 3, and RAID 51 at distance d = 4, have been proposed to address deficiencies in system reliability. It is common practice in storage systems to provide spare units to decrease the system repair time and increase the maintenance interval. Adding spares, however, increases the cost of the system and decreases the storage efficiency.

[014] Other approaches for improving system reliability include use of higher order parity

codes. For example, J.S. Plank, "A Tutorial on Reed-Solomon Coding for Fault-Tolerance in RAID-like Systems," Software – Practice & Experience, 27(9), September 1997, pp. 995discloses an example of a Reed-Solomon code.

[015] Additionally, E.J. Schwabe et al., "Evaluating Approximately Balanced Parity-Declustering Layouts in Disk Arrays," ACM 0-89791-813-4/96/05 1996, disclose data layouts for efficient positioning of redundant information for improved performance.

[016] P. Chen et al., "RAID: High-Performance, Reliable Secondary Storage," ACM Computing Surveys, Vol. 26, June 1994, pp. 145provide an overview of RAID. M. Holland et al., "Parity Declustering for Continuous Operation In Redundant Disk Arrays," Proceedings of the 5th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-V), pp. 23October 1992, disclose declustered parity for RAID systems. G.A. Alvarez et al., "Tolerating Multiple Failures in RAID Architectures," ACM 0-89791-901-7/97/0006 1997 describe the properties and construction of a general multiple parity array using 8-bit finite fields.

[017] U.S. Patent No. 5,579,475 to M.M. Blaum et al., entitled "Method and Means for Encoding and Rebuilding the Data Contents of Up to Two Unavailable DASDs in a DASD Array Using Simple Non-Recursive Diagonal and Row Parity," discloses the operation of a distance d = 3 array. N.K. Ouchi, "Two-Level DASD Failure Recover Method," IBM Technical Disclosure Bulletin, Vol. 36:03, March 1993, discloses operations required for reconstructing data from a distance d = 3 array having failures.

[018] Nevertheless, some array designs, such as product codes (including RAID 51), have vulnerabilities to certain patterns of storage unit failures. These arrays behave somewhat as if they possess local redundancy.

Disclosure of Invention

[019] The present invention provides a technique that improves the reliability of a storage system by making local redundancy in an array to be globally available throughout a system of arrays. Additionally, the present invention provides a technique that mitigates the failure pattern sensitivity of a storage system. Further still, the present invention provides a technique that allows maintenance of the storage system to be deferred for considerably longer than with a conventional storage system.

[020] The advantages of embodiments of the present invention are provided by a method for increasing failure tolerance of a storage system having a plurality of arrays such that each array has a plurality of storage units. The arrays of the storage system include redundancy based on an erasure or error correcting code, such as a parity code, a Winograd code, a symmetric code, a Reed-Solomon code, an EVENODD code or a derivative of an EVENODD code. The failure tolerance of a storage system is given by the minimum Hamming distance D of the system. The minimum distance of the system

is, accordingly, the minimum of all the minimum Hamming distances of the respective stripes, that is, $D = \min(d_i)$. A donor array is selected from the plurality of arrays when the difference between a minimum distance of the donor array and a minimum distance of a recipient array is greater or equal to 2. A donor storage unit is selected in the donor array based on a minimal performance impact on the donor array. A recipient storage unit is selected from the recipient array. At least a portion of lost information is then rebuilt from the recipient array onto the selected storage unit in the donor array. The recipient information is selected based on an improved performance of the recipient array. The selected storage unit is indicated to the donor array as having been donated before the lost information is rebuilt on the selected storage unit. Preferably, the minimum Hamming distance of the recipient array is $d \geq 2$ before the donor array is selected from the plurality of arrays. When a spare storage unit becomes available, the spare storage unit is assigned to a selected array in a conventional manner.

[021] When a storage unit in the donor array fails during the step of rebuilding at least a portion of recipient information from the recipient array on the selected storage unit, the step rebuilding is terminated and a second donor array is selected. At least a portion of lost recipient information from the recipient array is rebuilt on the selected storage unit in the second donor array. The selection of the second donor array proceeds by re-evaluating the conditions described previously.

[022] The present invention also provides a method of increasing the failure tolerance of an array of storage units that is vulnerable to selected patterns of failures. According to the invention, a recipient storage unit is selected from the array of storage units subsequent to a storage unit failure. A donor storage unit is selected from the array of storage units such that a failure tolerance of the array is increased following a rebuild operation. Lastly, at least a portion of lost recipient information is rebuilt onto the donor storage unit. When a spare unit becomes available, the spare unit is assigned to the array in a conventional manner.

[023] Further still, the present invention provides a data storage system having a plurality of arrays and a system array controller. Each array has a plurality of storage units and includes redundancy based on an erasure or error correcting code, such as a parity code, a Winograd code, a symmetric code, a Reed-Solomon code, an EVENODD code or a derivative of an EVENODD code. The system array controller is coupled to each array and detects a failure of a storage unit in a first array of the plurality of arrays. The system controller then selects a storage unit in a second array of the plurality of arrays when a difference between a minimum distance of the second array and a minimum distance of the first array is greater or equal to 2. At least a portion of lost information from the first array is rebuilt onto the selected storage unit of the second array. Each storage unit can be an HDD, a volatile Random Access Memory device, a non-volatile

Random Access Memory device, an optical storage device, or a tape storage device.

[024] The present invention also provides a data storage system having an array of a plurality of storage units and an array controller. The array includes redundancy based on an erasure or error correcting code, such as a parity code, a Winograd code a symmetric code, a Reed-Solomon code, an EVENODD code or a derivative of an EVENODD code. The array is also vulnerable to selected patterns of failures and/or a non-uniform failure probability. The array controller is coupled to the array and detects a failure of a first storage unit in the array. The array controller then selects a second storage unit in the array such that a failure tolerance of the array is increased following a rebuild operation, and rebuilds at least a portion of information from the first storage unit onto the second storage unit. The second storage unit is selected based on a failure pattern of the array and/or based on a predetermined target pattern. The minimum Hamming distance of the array is $d \geq 2$ before the array controller selects the second storage unit, and is increased upon completion of rebuilding at least a portion of information from the first storage unit onto the second storage unit. Each storage unit can be an HDD, a volatile Random Access Memory device, a non-volatile Random Access Memory device, an optical storage device or a tape storage device.

Brief Description of the Drawings

[025] The present invention is illustrated by way of example and not by limitation in the accompanying figures in which like reference numerals indicate similar elements and in which:

[026] Figure 1a shows a typical configuration of a storage system with a plurality of arrays connected to a common storage controller;

[027] Figure 1b shows a typical configuration of a storage system with a plurality of arrays connected to separate storage controllers;

[028] Figure 2 shows an exemplary set of two arrays for illustrating the present invention;

[029] Figure 3 shows the arrays of Figure 2 following failure of two drives in one of the arrays for illustrating the present invention;

[030] Figure 4 shows the arrays of Figure 2 following an APX operation according to an embodiment of the present invention;

[031] Figure 5 shows an exemplary array that has sensitivity to patterns of storage unit failures;

[032] Figure 6 shows an exemplary pattern of six storage unit failures of the array of Figure 5 that leads to data loss; and

[033] Figure 7 shows an exemplary target pattern of nine storage unit failures of the array of Figure 5 that has distance $d = 3$.

Mode for the Invention

[034] The present application is related to US Patent No.7,533,325 issued on May 12, 2009, entitled "Anamorphic Codes", US Patent No.7,379,974 issued on May 27, 2008, entitled "Multi-path Data Retrieval from Redundant Array," and US Patent No.7,254,754 issued on August 7, 2007 entitled "RAID 3 + 3". The present application is also related to co-pending and co-assigned US Patent No.7.350,126 issued on March 25, 2008.

[035] The present invention dramatically improves the reliability of a storage system and allows maintenance of the storage system to be deferred for considerably longer than can be with a comparable storage system without parity exchange. Thus, the present invention provides a significant reliability improvement over the degree of reliability provided by RAID systems. In contrast to RAID systems, which treat each array of a multi-array storage system as an individual entity, the present invention globally couples the individual arrays of a multi-array storage system, thereby allowing the redundancy of one array to be utilized by another array. Such a process is referred to herein as an autonomic parity exchange (APX) or as an APX operation.

[036] According to an embodiment of the present invention, an APX operation allows local redundancy in an array to be globally available throughout a system of arrays, thereby increasing system reliability as the number of storage units increases. APX also reduces or eliminates the need for spare storage units.

[037] To illustrate the benefits of the present invention using a specific example, consider a 48-unit storage system comprising eight arrays of six storage units and having no spare storage units. [Further, the exemplary storage system uses a symmetric code having distance d = 4. A symmetric code has an equal number of data units and redundant units, and the ability to recover from the loss of any combination of half the units. With APX, the distance of the storage system remains at D = 2 with up to nine I storage-unit failures. Assuming that the APX operations can complete with fewer than two failures during the operations, the storage systems reaches D - 1 with ten failures. In contrast, a RAID 6 system using three arrays of sixteen storage units reaches D = 1 with as few as two storage-unit failures. In further contrast, a RAID 51 system using three arrays of sixteen storage units reaches D = 1 with as few as three storage-unit failures.

[038] Moreover, a storage system utilizing APX gracefully degrades as failures accumulate, thereby permitting maintenance of the system to be deferred with an ac companying significant cost savings. Accordingly, the annual warranty costs for a storage system utilizing APX will be significantly less than the annual warranty cost for a comparable storage system without APX. For a conventional system, service is typically requested when the spare- storage-unit pool is exhausted. When APX is used,

service could be requested after up to nine unit failures for the exemplary system. Thus, when APX is used, the system can maintain a given distance over a longer interval compared to a system without APX.

[039] APX allows arrays within a set of arrays to exchange redundancy, thereby overcoming exposure to failures that are concentrated on a single array of the set of arrays. For example, if a first array has a minimum Hamming distance that is less than the minimum Hamming distance of a second array by 2 or more, the second array can donate a storage unit to the first array. Afterward, the failure tolerance of the first array will be increased and the failure tolerance of the second array will be reduced, but to a level that is not less than the first array. Accordingly, the minimum distance of the system will be increased, thereby increasing the failure tolerance of the system.

[040] Figure 1a shows an exemplary storage system, indicated generally as 100, comprising two storage arrays 102 and 103 that are connected to a common array controller 101. Storage arrays 102 and 103 comprise multiple storage units 104 and communicate with array controller 101 over interface 105. Array controller 101 communicates to other controllers and host systems over interface 106. Such a configuration allows an array controller to communicate with multiple storage arrays.

[041] Figure 1b shows an exemplary storage system, indicated generally as 150, comprising two storage arrays 153 and 154, that are respectively connected to different array controllers 152 and 151. Storage array 153 communicates with array controller 152 over interface 157, and storage array 154 communicates with array controller 151 over interface 156. Array controllers 151 and 152 respectively communicate with other array controllers and storage systems over interfaces 158 and 159. Also shown in Figure 1b is a communication connection 160 that allows array controllers 151 and 152 to communicate with each other.

[042] The array controllers shown in figures 1a and 1b may be designed as hardware or software controllers. The term controller will be used herein generally to refer to any of the configurations described above.

[043] Figure 2 shows an exemplary set of two arrays 201 and 202 for illustrating the present invention. Array 201 includes storage units 1Aand array 202 includes storage units 2A Storage units A, B and C of each array are data storage units and storage units D, E and F of each array are redundant storage units with an MDS code. Accordingly, both arrays 201 and 202 have a minimum Hamming distance d = 4. The configuration shown in Figure 2 is referred to as a symmetric code in which the number of data storage units equals the number of redundant storage units. Redundancy is calculated so that any three unit failures can be corrected by the symmetric code. Erasure or Error correcting codes (ECCs), such as parity codes, Winograd codes, symmetric codes, Reed-Solomon codes, EVENODD codes and derivatives of EVENODD codes, can be

used for generating the redundancy. As used herein, the term "distance" means the minimum Hamming distance.

[044] Figure 3 shows the arrays of Figure 2 following failure of storage units 1C and 1E in array 201. After storage units 1C and 1E fail, array 201 has distance d = 2, while array 202 still has distance d = 4. Array 201 can tolerate only one further failure without the possibility of a data loss event. Array 202 can still tolerate three failures without the possibility of a data loss event. The overall system distance is that of array 201, i.e., distance D = 2. The overall failure tolerance of the system can be increased by rebalancing the redundancy using APX, because the distance of the system is increased from D = 2 to D = 3.

[045] Rebalancing the redundancy is achieved by donating a storage unit contained within array 202 (referred to as the donor array), and then providing the donated storage unit to array 201 (referred to as the recipient array) as if the donated storage unit were a spare unit. The donor array must be made aware that the donated storage unit is no longer part of the donor array to prevent the donor array from reading or writing data on the donated unit. It may be beneficial to assign one of the failed storage units from the recipient array to the donor array so that both arrays maintain a constant number of storage units and maintain knowledge of the failed units. No information can be written to the donated unit by the donor array. The system can select which storage unit to donate. The primary criterion for selecting a donor unit is based on selecting a donor unit that has the least impact on the donor array reliability. A secondary criterion is based on selecting the storage unit having the least impact on performance, such as the unit with the most expensive redundancy calculation. The system can select which data from the failed units to rebuild onto the donated unit. The primary criterion for selecting the information to be rebuilt is based on the information set that provides the greatest increase in reliability. A secondary criterion is to select the information set that provides the best array performance following the rebuild operation. In the example of Figure 3 above, recipient array 201 will have the best performance by rebuilding the information set of unit 1C because unit 1C is a data unit. Similarly, donor array 202 will have the least performance impact by donating a unit storing redundant information, such as unit 2F. In both cases, after the APX operation, read commands could thus be processed without the need to reconstruct the data from the redundant units of the storage arrays 201 and 202.

[046] Donating a storage unit from a donor array to a recipient array requires that the storage system be able to assign storage units from one array to another array. When the donor and recipient arrays are connected to a common array controller 101, as shown in Figure 1a, then controller 101 can perform this operation internally. When the donor and recipient arrays are connected to separate controllers 151 and 152, as

shown in Figure lb, then controllers 151 and 152 exchange information. For example, the controllers could expose the individual drives over communication connection 160, such as in the manner of a Just a Bunch of Disks (JBOD) array configuration. Alternatively, the controllers could exchange information regarding the data to be read i and written from the locations on the storage units involved.

[047] It is possible to achieve the donation by artificially indicating that the donated unit has failed in the donor array. It is, however, beneficial to perform the donation in a piece-wise manner. Segments of data that have been written with recipient array data during the rebuild belong to the recipient. Until the rebuild has completed, the donor array may be permitted to write to the donated unit to keep data in the remaining segments up to date. In the situation in which a storage unit in the donor array fails during the parity exchange, it may be preferred to terminate the donation operation, and initiate a new donation operation with a different donor array. The donor array can then rebuild onto the previously donated segments to increase the distance of the array. For example, if the failure occurs during the initial portion of the donation operation, this method of reversing the donation and initiating a new donation should decrease the duration that the system is operating at reduced distance.

[048] Once a spare storage unit becomes available, such as through maintenance, it can be assigned to replace any of the failed units. Information is rebuilt onto Me spare in a well-known mater. Assigning one of the failed units of the recipient array to the donor array can facilitate this operation because it indicates to which array a failed unit belongs.

[049] Figure 4 shows the arrays of Figure 2 following fan APX operation according to the present invention. Rebuilt data in Figure 4 is underlined. While there are still two failed storage units in the system, there is only one failed storage unit in each array. For the system configuration of Figure 4, each array now has distance d = 3 and can tolerate two further failures without possible loss of data. The overall distance of the system is now D = 3.

[050] Using APX, a storage system can tolerate more failures than would otherwise be the case. In the example of Figures 2-4, the first point of system failure would be at four unit failures without utilizing APX. In contrast, when APX is used, the first point of system failure is six unit failures. The improvement has not been achieved by requiring additional redundancy or sparing, but by adjusting global assignment of redundancy within the storage system to meet observed storage-unit-failure conditions. The improvement provided by APX increases with increasing number of arrays in a system.

[051] In a storage system having equal distance arrays, the donor arrays must be at least distance d = 3 so that an APX operation can be performed. That is, a donor array must have distance at least 2 greater than the distance of the recipient array. Generally, an

APX operation is preferably performed when a recipient array is at distance 2 or greater, thereby protecting against hard errors or a further storage unit failure during an APX operation.

[052] Many conventional systems, such as RAID 5, use distributed parity (also called "de-clustered parity") for spreading access patterns. In such systems, each storage unit has portions assigned to each unit type (e.g., data I, data 2, redundancy I, etc.). APX can also be applied to systems using distributed parity. In such a case, selection of a donor storage unit is less critical because the redundancy is spread across all the units. The system can select to simply rebuild any one of the recipient's failed storage units.

[053] When there are multiple choices for donor arrays, the selection criteria for a donor array can be based on considerations such as utilization, age of devices, and previous error history.

[054] The illustrative example shown in Figures 2-4 performs parity exchange with arrays in which the number of redundant units is the same as the number of data units. When the number of storage units in the array is greater than the number of storage units in the base array, an APX operation can be performed in combination with a dodging operation, such as disclosed by US Patent No.7,533,325.

[055] Some array designs are sensitive to patterns of unit failures. In such arrays, both the donor unit and recipient unit may come from the same array. A donor unit can be selected on the basis of the array configuration. Figure 5 shows a product code al ray 500 comprising 30 storage units, in which rows 501-505 form RAID 6 arrays (d = 3) and columns 511-516 form RAID 5 arrays (d = 2) In array 500, D = 6 with 14 redundant drives. Only certain positional arrangements of six failures, however, will cause the array to fail, as illustrated in Figure 6, in which storage units lJ, lK, lM, 3J, 3K and 3M have failed. The pattern of failures can be recognized as three failures in a first row matching three failures in a second row. This failure tolerance of the system can be increased by using APX. The system would choose the donor unit and the recipient unit in such a manner, thereby avoiding these patterns and returning the array back to a D = 3 condition.

[056] For the example of Figure 5, the system can maintain D = 3 with nine failures when they are in an arrangement such as shown in Figure 7. Here, the array has nine failed units, lN, 2N, 3N, 4N, 5H, 5J, 5K, 5L and 5M. The system can choose a target pattern that has a high failure tolerance. As failures occur, donor units are selected from the target pattern and assigned to failed recipients that are not part of the target pattern. For i the example of Figure 5, an APX operation can be performed several times such that the array can tolerate at least ten failures before reaching D = 2. The effective distance

has thereby been increased from D = 6 to D = 12 by utilizing APX. APX allows the maintenance to be deferred until ten of the 30 units have failed, with the array remaining at least D = 2. Without APX, maintaining D = 2 can require maintenance with as few as four unit failures.

[057] The terms "failure" and "pattern" as used herein refer to the erasure of information from the logical position in the array, not the physical position. Such non-MDS arrays can be formed from product codes as illustrated, low density parity codes, non-uniform graph codes, or any codes that have particular pattern vulnerabilities.

[058] APX can be used beyond simply increasing the minimum distance of a storage system. Many other factors may be included in determining whether to perform APX and to choose donors and recipients. For example, the individual failure probabilities of components when they are non-uniform, the combinations of failures that lead to data loss, and the effects on system performance may all be considered. In such cases, the minimum distance of the system could remain unchanged following APX.

[059] APX can be used with other array types having minimum distance $d \geq 3$. Additionally, a smaller array size allows APX to be used more efficiently, and allows large systems consisting of small arrays to achieve high failure tolerance. When a storage system has a spare pool, it is best to perform rebuilds onto the spare pool before performing an APX operation.

[060] APX can also be performed on a subset of the data on a storage unit. For example, in some configurations the rebuild time may be decreased. Consider the example of Figure 4. Instead of rebuilding the contents of unit 1C onto unit 2F, it may be beneficial to rebuild half of unit 1C onto half of unit 2F, and the other half of unit 1E onto half of unit 2E. The net result would be both arrays at D = 3, but the rebuild time may be reduced because the same amount of data is being rebuilt, but two donor drives are being used. Other combinations are clearly possible as well.

[061] While the present invention has been described in terms of storage arrays formed from HDD storage units, the present invention is applicable to storage systems formed from arrays of other memory devices, such as Random Access Memory (RAM) storage devices (both volatile and non-volatile), optical storage devices, and tape storage devices. Additionally, it is suitable to virtualized storage systems, such as arrays built out of network -attached storage. It is further applicable to any redundant system in which there is some state information that associates a redundant component to particular subset of components, and that state information may be transferred using a donation operation.

[062] While the invention has been described with respect to specific examples including presently preferred modes of carrying out the invention, those skilled in the art will appreciate that there are numerous variations and permutations of the above described

systems and techniques that fall within the spirit and scope of the invention as set forth in the appended claims.

What is claimed is:

1. A method for increasing failure tolerance of a storage system having a plurality of arrays, each array having a plurality of storage units, the method comprising steps of:

selecting a recipient array from the plurality of arrays;

selecting a donor array from the plurality of arrays when a difference between a minimum Hamming distance of the donor array and a minimum Hamming distance of the recipient array is greater or equal to 2;

selecting a donor storage unit in the donor array; and

rebuilding at least a portion of lost recipient information from the recipient array on the selected storage unit in the donor array.

2. The method according to claim 1, wherein the minimum Hamming distance of the recipient array is greater than or equal 2 before the step of selecting the donor array from the plurality of arrays.

3. The method according to claim 1, farther comprising a step of indicating to the donor array that the selected storage has been donated before the step of rebuilding the lost information on the selected storage unit.

4. The method according to claim 1, wherein the storage units are hard disk drives.

5. The method according to claim 1, wherein the storage units are RAM storage devices.

6. The method according to claim 1, further including a step of selecting a recipient storage unit from the recipient array.

7. The method according to claim 6, wherein at least a portion of the lost contents of the recipient storage unit are rebuilt onto the donor storage unit.

8. The method according to claim 1, wherein the arrays of the storage system include redundancy based on an erasure or error correcting code.

9. The method according to claim 8, wherein the erasure or error correcting code is a parity code.

10. The method according to claim 8, wherein the erasure or error correcting code is a Winograd code.

11. The method according to claim 8, wherein the erasure or error correcting code is a symmetric code.

12. The method according to claim 8, wherein the erasure or error correcting code is a Reed-Solomon code.

13. The method according to claim 8, wherein the erasure or error correcting code is an EVENODD code.

14. The method according to claim 8, wherein the erasure or error correcting code is a derivative of an EVENODD code.

15. The method according to claim 8, wherein the arrays of the storage system includes redundancy based on a product of a plurality of erasure or error correcting codes.

16. The method according to claim 15, wherein at least one of the erasure or error correcting codes is a parity code.

17. The method according to claim 15, wherein at least one of the erasure or error correcting codes is a Winograd code.

18. The method according to claim 15, wherein at least one of the erasure or error correcting code is a symmetric code.

19. The method according to claim 15, wherein at least one of the erasure or error correcting code is a Reed-Solomon code.

20. The method according to claim 15, wherein at least one of the erasure or error correcting code is an EVENODD code.

21. The method according to claim 15, wherein at least one of the erasure or error correcting code is a derivative of an EVENODD code.

22. The method according to claim 1, wherein when a storage unit in the donor array fails during the step of rebuilding at least a portion of recipient information from the recipient array on the selected storage unit, the method further comprising steps of:

terminating the step of rebuilding at least a portion of recipient information from the recipient array on the selected storage unit;

selecting a second donor array from the plurality of arrays when a difference between a minimum Hamming distance of the second donor array and a minimum Hamming distance of the recipient array is greater or equal to 2;

selecting a donor storage unit in the second donor array; and

rebuilding at least a portion of lost recipient information from the recipient array on the selected storage unit in the second donor array.

23. The method according to claim 1, wherein when a spare storage unit becomes available, the method further comprising a step of assigning the spare storage unit to a selected array.

24. A data storage system, comprising:

a plurality of arrays, each array having a plurality of storage units;

and a system array controller coupled to each array, the system array controller

detecting a failure of a storage unit in a first array of the plurality of arrays,

selecting a storage unit in a second array of the plurality of arrays when a difference between a minimum Hamming distance of the second array and a minimum Hamming distance of the first array is greater or equal to 2, and

rebuilding at least a portion of information from the first array onto the selected storage unit of the second array.

25. The data storage system according to claim 24, wherein when a spare unit becomes available, the spare unit is assigned to the second array.

26. The data storage system according to claim 24, wherein at least one array has a non-uniform failure probability.

27. The data storage system according to claim 24, wherein the system array controller includes a plurality of array controllers, an array controller being coupled to at least one array of the plurality of arrays, each respective array controller detecting a failure of a storage unit in each array associated with the array controller, a first array controller of the plurality of array controllers selecting a storage unit in an array associated with the first array controller when a difference between a minimum Hamming distance of the array of the selected storage unit and a minimum Hamming distance of an array associated with a second array controller of the plurality of array controllers is greater or equal to 2, and the first and second array controllers rebuilding at least a portion of lost information from the array associated with the second array controller onto the selected storage unit in the array associated with the first array controller.

28. The data storage system according to claim 27, wherein when a spare unit becomes available, the spare unit is assigned to the array of the selected storage unit.

29. The data storage system according to claim 27, wherein at least one array has a non-uniform failure probability.

30. The data storage system according to claim 24, wherein the arrays of the data storage system include redundancy based on an erasure or error correcting code.

31. The data storage system according to claim 30, wherein the erasure or error correcting code is a parity code.

32. The data storage system according to claim 30, wherein the erasure or error correcting code is a Winograd code.

33. The data storage system according to claim 30, wherein the erasure or error correcting code is a symmetric code.

34. The data storage system according to claim 30, wherein the erasure or error correcting code is a Reed-Solomon code.

35. The data storage system according to claim 30, wherein the erasure or error correcting code is an EVENODD code.

36. The data storage according to claim 30, wherein the erasure or error correcting code is a derivative of an EVENODD code.

37. The data storage system according to claim 24, wherein the arrays of the storage system includes redundancy based on a product of a plurality of erasure or error correcting codes.

38. The data storage system according to claim 37, wherein at least one of the erasure or error correcting codes is a parity code.

39. The data storage system according to claim 37, wherein at least one of the erasure or error correcting codes is a Winograd code.

40. The data storage system according to claim 37, wherein at least one of the erasure or error correcting codes is a symmetric code.

41. The data storage system according to claim 37, wherein at least one of the erasure or error correcting codes is a Reed-Solomon code.

42. The data storage system according to claim 37, wherein at least one of the erasure or error correcting codes is an EVENODD code.

43. The data storage system according to claim 37, wherein at least one of the erasure or error correcting codes is a derivative of an EVENODD code.

44. The data storage system according to claim 24, wherein each storage unit is one of a Hard Disk Drive, a volatile Random Access Memory device, a non-volatile Random Access Memory device, an optical storage device, and a tape storage device.

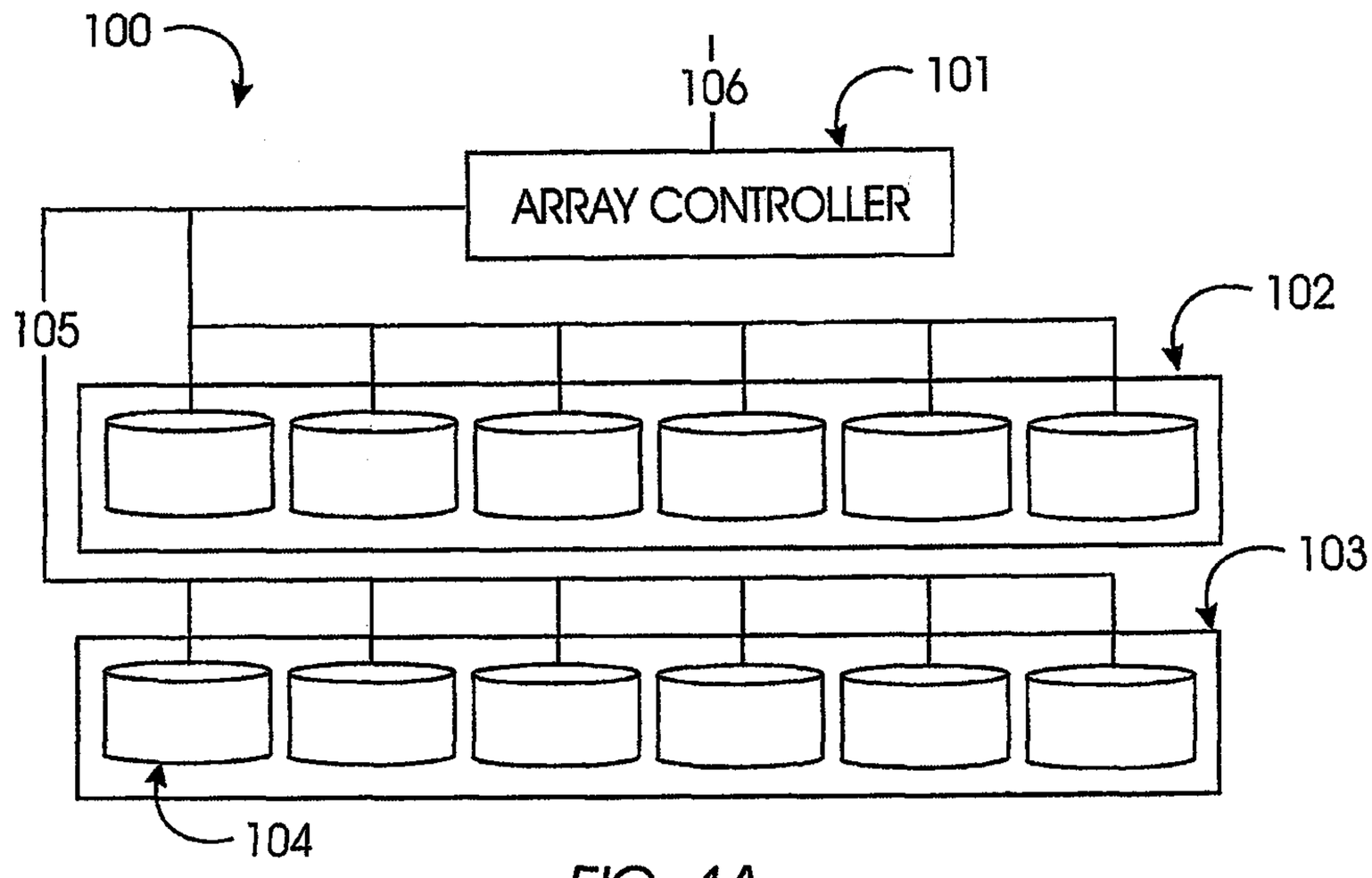

100
106
101
ARRAY CONTROLLER
105
102
103
104

FIG. 1A

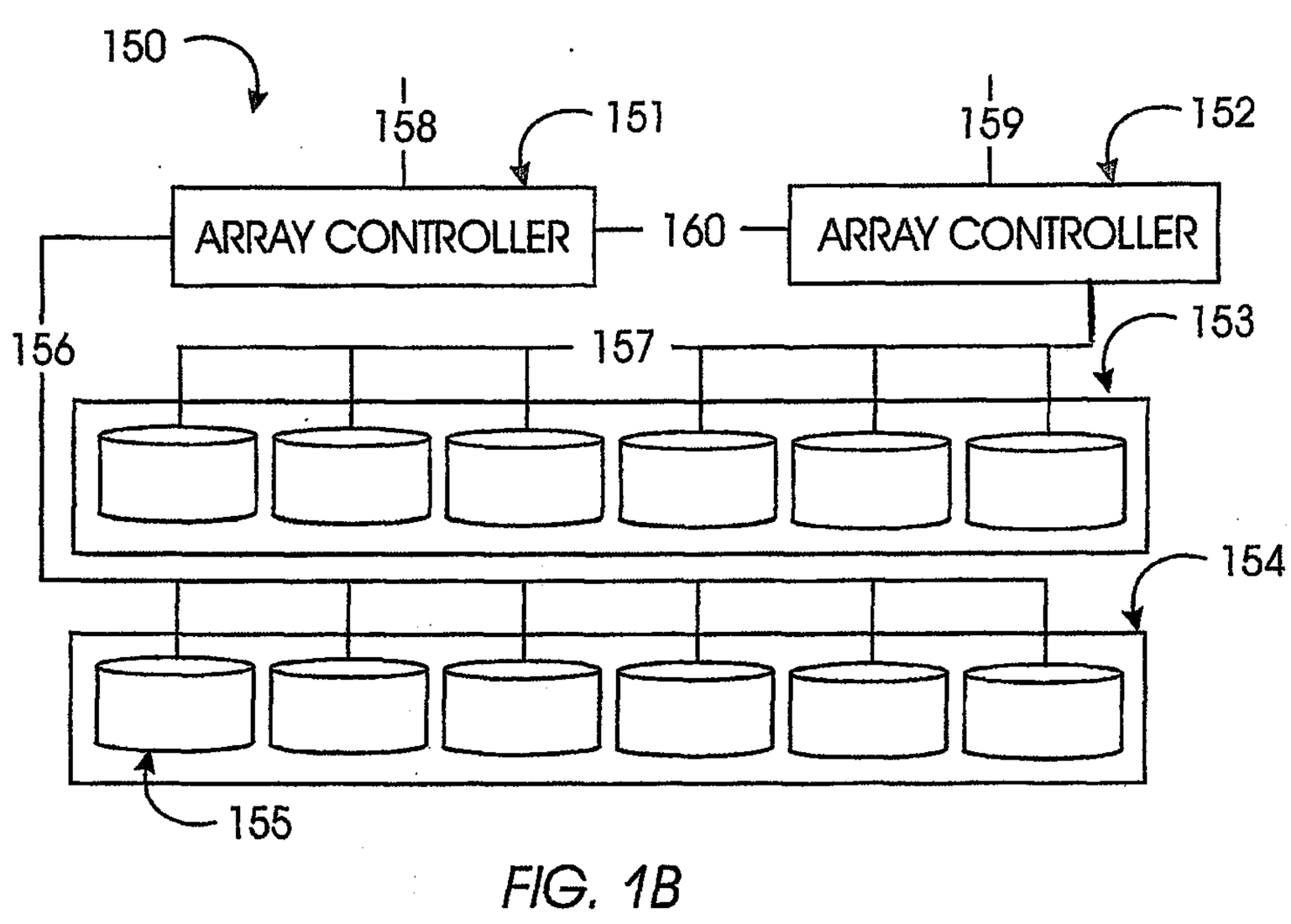

150
158
151
159
152
ARRAY CONTROLLER
160
ARRAY CONTROLLER
156
157
153
154
155

FIG. 1B

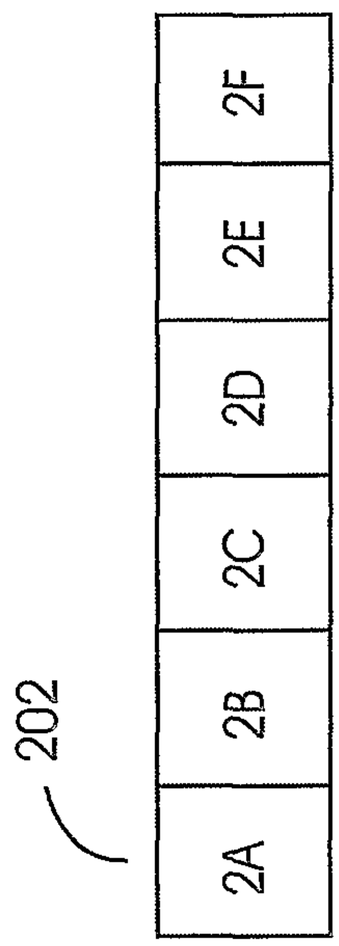


FIG. 2

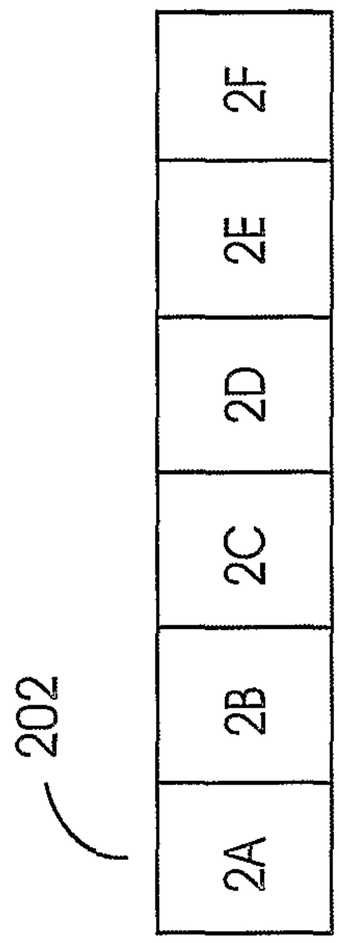


FIG. 3

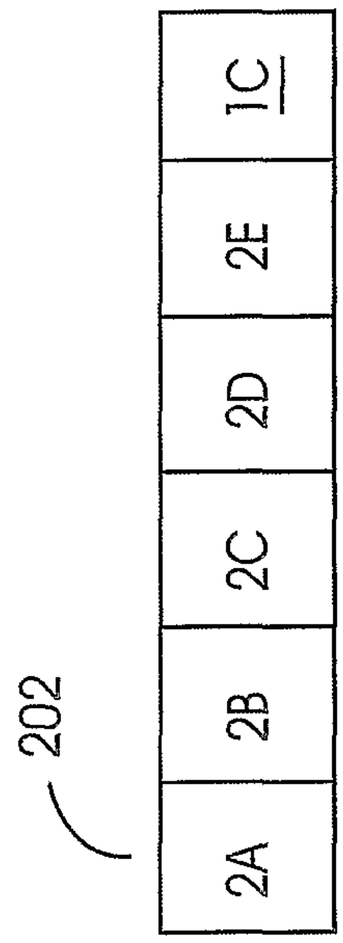
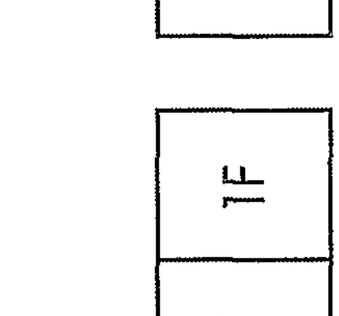
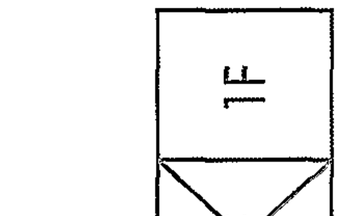
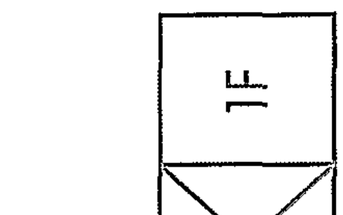


FIG. 4







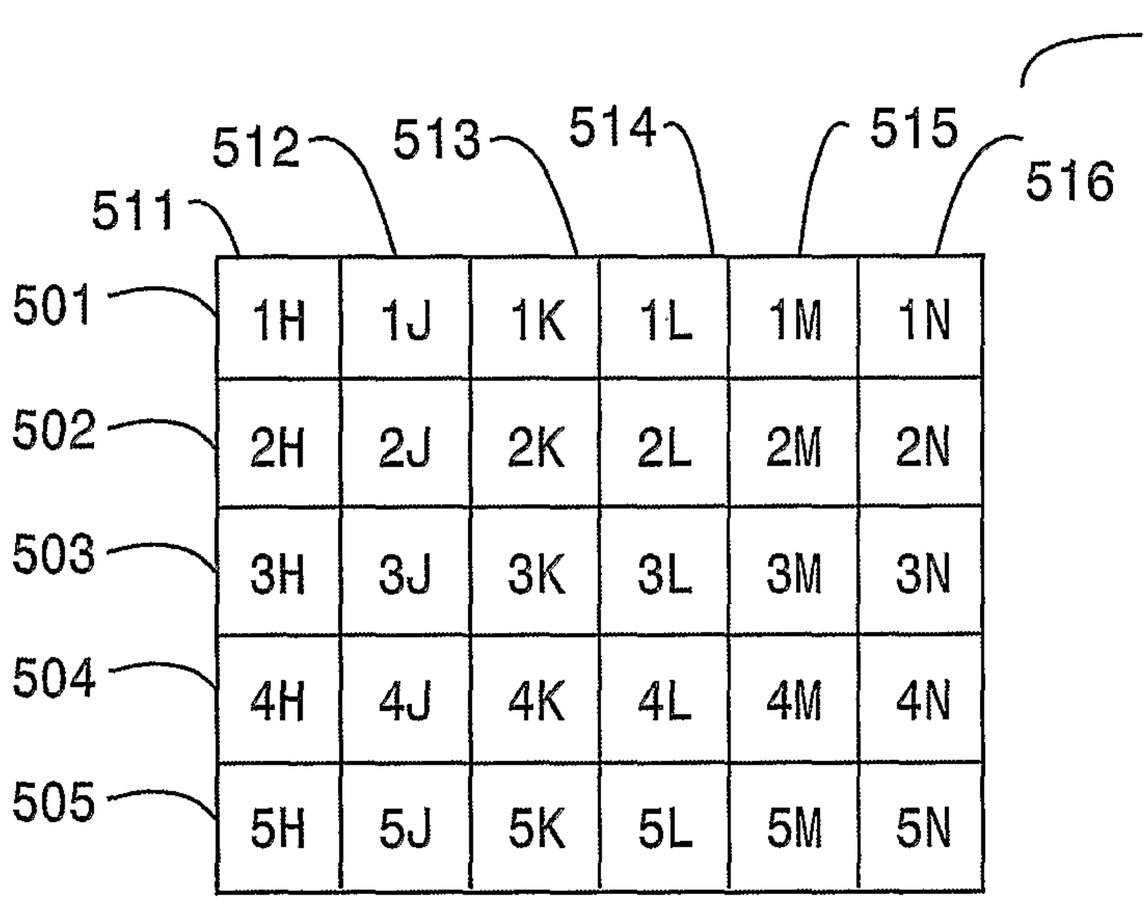


FIG. 5

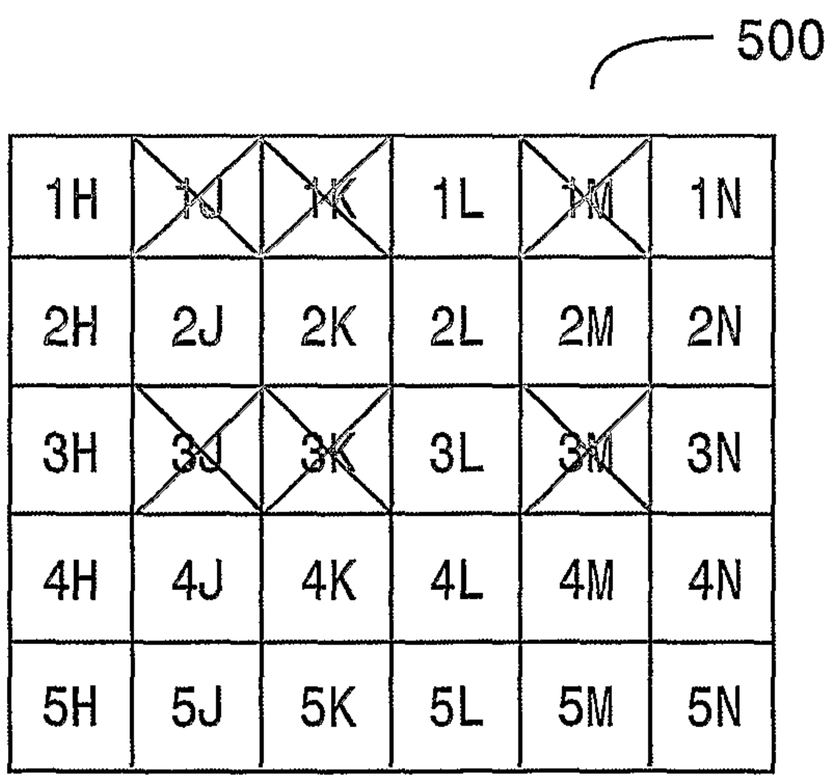


FIG. 6

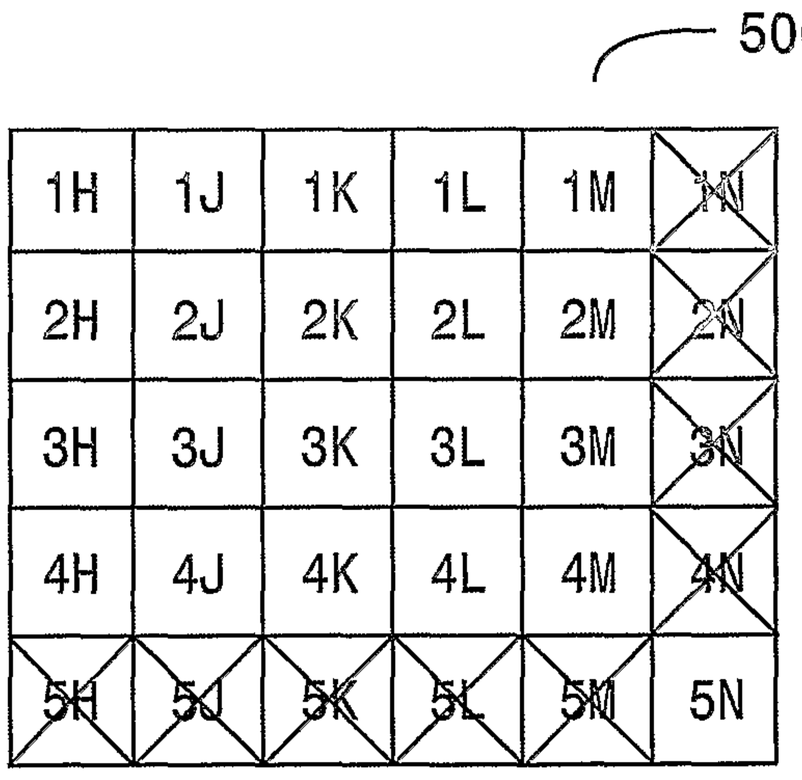
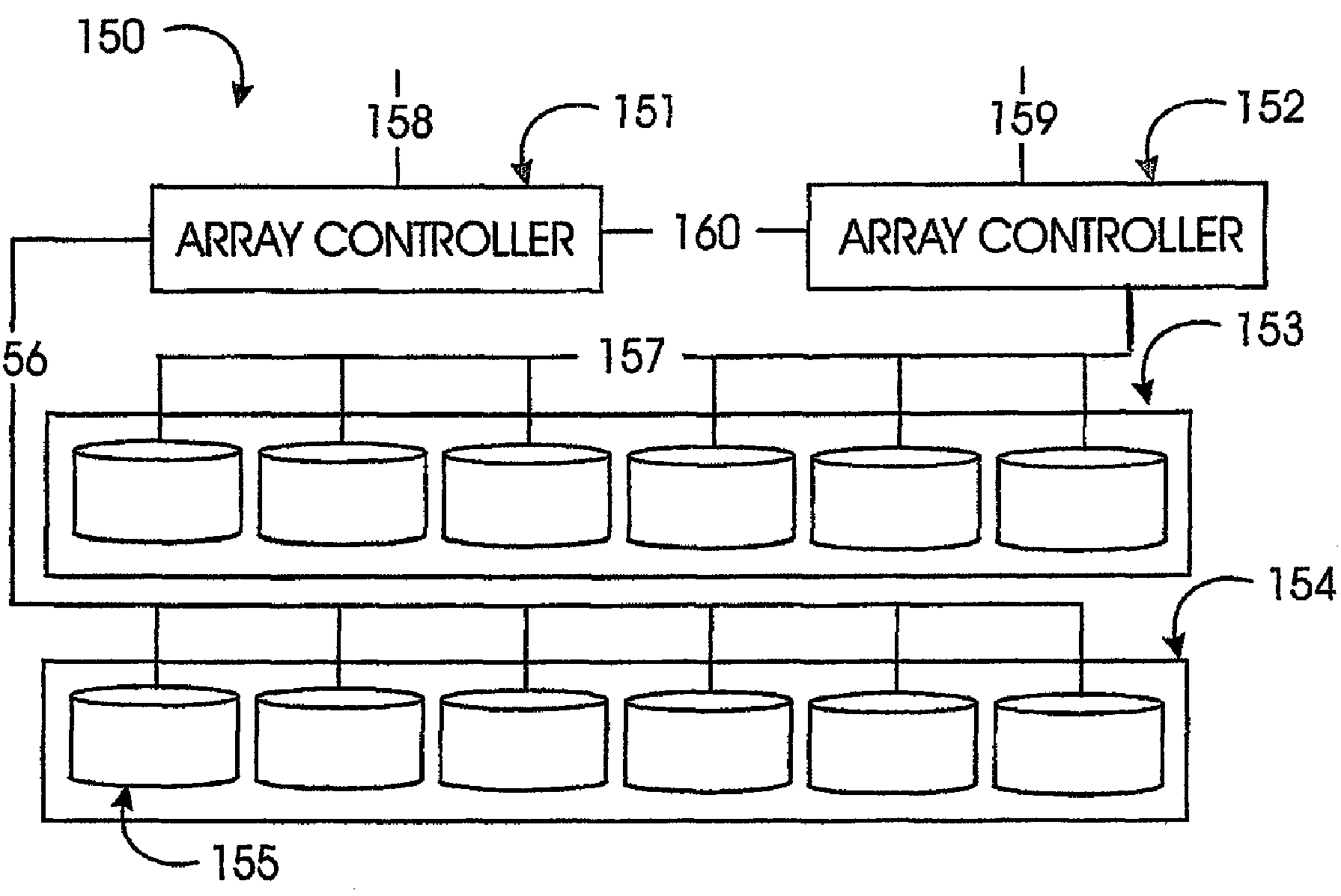


FIG. 7


150
158
151
159
152
ARRAY CONTROLLER
160
ARRAY CONTROLLER
156
157
153
154
155