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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 August 2013 (15.08.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/118170 A1

(51) International Patent Classification:
*G06F 3/06* (2006.01) *G06F 12/02* (2006.01)

(21) International Application Number:
PCT/JP2012/000843

(22) International Filing Date:
8 February 2012 (08.02.2012)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **HITACHI, LTD.** [JP/JP]; 6-6, Marunouchi 1-chome, Chiyoda-ku, Tokyo, 1008280 (JP).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KOSEKI, Hideyuki** [JP/JP]; c/o HITACHI, LTD., Yokohama Research Laboratory, 292, Yoshida-cho, Totsuka-ku, Yokohama-shi, Kanagawa, 2440817 (JP). **OGAWA, Junji** [JP/JP]; c/o HITACHI, LTD., Yokohama Research Laboratory, 292, Yoshida-cho, Totsuka-ku, Yokohama-shi, Kanagawa, 2440817 (JP).

(74) Agent: **WILLFORT INTERNATIONAL**; Kanda-Ogawamachi Tosei Bldg. II 7F, 3, Kanda-Ogawamachi 3-chome, Chiyoda-ku, Tokyo, 1010052 (JP).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published**:

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

(54) **Title**: STORAGE APPARATUS WITH A PLURALITY OF NONVOLATILE SEMICONDUCTOR STORAGE UNITS AND CONTROL METHOD THEREOF TO PLACE HOT DATA IN STORAGE UNITS WITH HIGHER RESIDUAL LIFE AND COLD DATA IN STORAGE UNITS WITH LOWER RESIDUAL LIFE

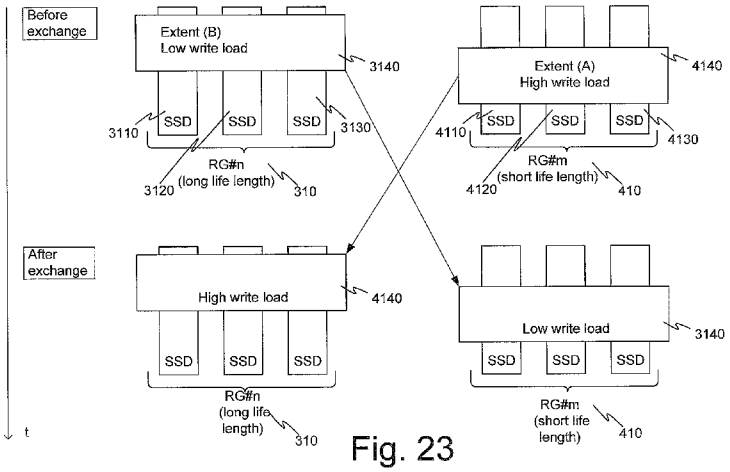


Fig. 23

(57) **Abstract**: A storage apparatus is provided with a plurality of nonvolatile semiconductor storage media and a storage controller that is a controller that is coupled to the plurality of semiconductor storage media. The storage controller identifies a first semiconductor storage unit that is at least one semiconductor storage media and a second semiconductor storage unit that is at least one semiconductor storage media and that is provided with a remaining length of life shorter than that of the first semiconductor storage unit based on the remaining life length information that has been acquired. The storage controller moreover identifies a first logical storage region for the first semiconductor storage unit and a second logical storage region that is provided with a write load higher than that of the first logical storage region for the second semiconductor storage unit based on the statistics information that indicates the statistics that is related to a write for every logical storage region. The storage controller reads data from the first logical storage region and the second logical storage region, and writes data that has been read from the first logical storage region to the second logical storage region and/or writes data that has been read from the second logical storage region to the first logical storage region.

WO 2013/118170 A1

# Description

## Title of Invention: STORAGE APPARATUS WITH A PLURALITY OF NONVOLATILE SEMICONDUCTOR STORAGE UNITS AND CONTROL METHOD THEREOF TO PLACE HOT DATA IN STORAGE UNITS WITH HIGHER RESIDUAL LIFE AND COLD DATA IN STORAGE UNITS WITH LOWER RESIDUAL LIFE

### Technical Field

[0001] The present invention relates to a storage control of a storage apparatus that is provided with a plurality of nonvolatile semiconductor storage media.

### Background Art

[0002] A storage apparatus is provided with a physical storage medium that is configured to store data and a controller that is configured to control a physical storage medium in general. The controller provides a data store space (a logical volume in general) to a computer (such as a host) that is coupled to a storage apparatus.

[0003] A storage apparatus enables an I/O processing to be executed at a high speed and can show a high fault tolerance to a failure of a physical storage medium by using a plurality of physical storage media with a RAID (Redundant Array of Independent (or Inexpensive) Disks) configuration.

[0004] A storage apparatus is equipped with an HDD (Hard Disk Drive) as a physical storage medium in general. In recent years however, a physical storage medium that is provided with a flash memory (hereafter referred to as an FM) such as an SSD (Solid State Drive) has attracted attention as a new physical storage medium as substitute for an HDD.

[0005] The SSD is provided with a merit of an extremely high speed of an I/O processing as compared with an HDD. However, there is an upper limit to the frequency of write of data for the SSD and a length of life of an SSD is shorter than that of an HDD disadvantageously. The disadvantages of the SSD will be described in the following.

[0006] In the case in which data of a flash memory (a flash memory of a NAND type typically) is tried to be rewritten, data cannot be over written on a physical region that has stored the data. In order to rewrite data to the data on the physical region, after an erasing processing is executed to data on the physical region in a unit of a block that is an erasing unit of a flash memory (hereafter referred to as a block erasing), it is necessary to write data on the physical region in which the block erasing has been executed.

[0007] However, the number of times of a block erasing (hereafter referred to as a frequency of erasing) for each block is limited because of a physical restriction of a flash memory. In the case in which a frequency of erasing of a block exceeds the limit, data cannot be stored into the block. In other words, a length of life of an SSD is when a

frequency of erasing of all blocks that configure the SSD exceeds the upper limit.

[0008] A length of life of an SSD is lengthened by using a method called a wear leveling (hereafter referred to as a WL) for a general SSD. This is a technique for leveling a frequency of erasing between blocks and for suppressing only a specific block from being degraded by controlling a store location of data in such a manner that data that is updated with a low frequency is stored into a block that is provided with a more frequency of erasing and data that is updated with a high frequency is stored into a block that is provided with a less frequency of erasing.

[0009] In the case in which an SSD is adopted as a physical storage medium of a storage apparatus, a plurality of SSDs is mounted on the storage apparatus in general. In other words, even in the case in which only a specific block can be suppressed from being degraded, an imbalance occurs to loads between SSDs and a load is concentrated solely on a specific SSD in some cases. Patent Literature 1 discloses a method for implementing a long life of the entire of a storage apparatus by applying the WL to SSDs and by leveling an erasing frequency between a plurality of SSDs. A WL that is executed between physical storage media such as an SSD is referred to as an "inter-device WL" in the following.

## Citation List

## Patent Literature

[0010] PTL 1: WO/2011/010344

## Summary of Invention

## Technical Problem

[0011] A controller of a storage apparatus (hereafter referred to as a storage controller) decides data of a movement target for an inter-device WL. Consequently, it is necessary that the storage controller comprehends the internal information of an SSD. The internal information is information such as a data write amount to a flash memory and a remaining erasing frequency. In the case in which a granularity of the information is finer (that is, the information is more detailed), an execution accuracy of an inter-device WL is improved. This means that the storage controller can comprehend an SSD, data of the SSD, and an SSD to which the data is to be moved. For instance, Patent Literature 1 discloses a technique in which the storage controller can comprehend the internal information of the SSD in a physical block unit in detail to implement an inter-device WL. For Patent Literature 1, the storage controller controls the information of an erasing frequency of a plurality of blocks in each SSD.

[0012] However in this case, since the internal information of each SSD is detailed, an amount of the internal information of each SSD is large, and a memory of a large capacity is required for the storage controller. A huge amount of internal information

from each SSD is stored into the memory. Consequently, in the case in which the storage controller levels an erasing frequency, it is necessary that the storage controller refers to a huge amount of internal information that has been stored into the memory. Therefore, a load of the storage controller is larger.

[0013] Such a problem may occur for a storage apparatus that is provided with nonvolatile semiconductor storage media in which an erasing frequency is limited other than the SSD.

## Solution of Problem

[0014] A storage apparatus is provided with a plurality of nonvolatile semiconductor storage media and a storage controller that is a controller that is coupled to the plurality of semiconductor storage media.

[0015] Each of the semiconductor storage media unit is configured by at least one non-volatile semiconductor storage media and is a basis of a logical storage region. The storage controller writes data based on data of a write target to a semiconductor storage unit that is a basis of a logical storage region of a write destination of a plurality of logical storage regions. The storage controller acquires the internal information from each of the semiconductor storage media on a regular basis or on an irregular basis for instance, and stores the internal information that has been acquired for every semi-conductor storage medium.

[0016] The storage controller stores the statistics information that indicates the statistics that is related to a write for every logical storage region, and stores the remaining life length information that is the information that is related to a remaining length of life of each of the semiconductor storage media. The remaining life length information can be a numerical value that indicates a remaining length of life itself, can be an attribute that has an effect on a remaining length of life (for instance, a storage media type such as a single level cell and a multi-level cell), and can be a numerical value that is used for predicting (calculating) a remaining length of life. The storage apparatus can be provided with a physical storage medium of other type (such as a hard disk drive) in addition to a plurality of semiconductor storage media.

[0017] The storage controller identifies a first semiconductor storage unit and a second semiconductor storage unit that is provided with a remaining length of life shorter than that of the first semiconductor storage unit based on the remaining life length in-formation that has been acquired.

[0018] The storage controller moreover identifies a first logical storage region for the first semiconductor storage unit and a second logical storage region that is provided with a write load higher than that of the first logical storage region for the second semi-conductor storage unit based on the statistics information that indicates the statistics that is related to a write for every logical storage region.

[0019] The storage controller reads data from the first logical storage region and the second logical storage region, and writes data that has been read from the first logical storage region to the second logical storage region and/or writes data that has been read from the second logical storage region to the first logical storage region.

[0020] The semiconductor storage unit can be one SSD or can be a RAID group that is configured by at least two SSDs for instance.

[0021] The logical storage region can be a logical address range based on one SSD or can be a logical address range that is over at least two SSDs that configure a RAID group for instance.

## Advantageous Effects of Invention

[0022] An increase in a load of the storage controller can be reduced, and a leveling of an erasing frequency between nonvolatile semiconductor storage media can be executed with a high degree of accuracy.

## Brief Description of Drawings

[0023] [fig.1]Fig. 1 is a view showing a configuration example of a storage system 10000 that includes a storage apparatus 10 in accordance with a first embodiment.

[fig.2]Fig. 2 is a schematic illustrative drawing showing the capacity virtualization technique.

[fig.3]Fig. 3 is a view showing a configuration example of a cache memory 103.

[fig.4]Fig. 4 is a view showing a configuration example of a Disk management TBL 13100.

[fig.5]Fig. 5 is a view showing a configuration example of an RG management TBL 13200.

[fig.6]Fig. 6 is a view showing a configuration example of a Pool management TBL 13300.

[fig.7]Fig. 7 is a view showing a configuration example of an extent management TBL 13400.

[fig.8]Fig. 8 is a view showing a configuration example of a virtual volume management TBL 13500.

[fig.9]Fig. 9 is a view showing a configuration example of a statistics information management TBL 13600.

[fig.10]Fig. 10 is a view showing a configuration example of an FM WR amount prediction TBL 13700 in accordance with a first embodiment.

[fig.11]Fig. 11 is a view showing a configuration example of an SSD 700.

[fig.12]Fig. 12 is a view showing a configuration example of a flash memory 780.

[fig.13]Fig. 13 is a view showing a configuration example of an address space of an SSD.

[fig.14]Fig. 14 is a view showing a configuration example of a cache memory 716.

[fig.15]Fig. 15 is a view showing a configuration example of a logical physical conversion TBL 23100.

[fig.16]Fig. 16 is a view showing a configuration example of a statistics information management TBL 23200 in an SSD.

[fig.17]Fig. 17 is a view showing a configuration example of the SSD internal information 25000.

[fig.18]Fig. 18 is a sequence drawing showing an example of a flow of a processing from a transmission of a write request from a host computer 30 to a completion of the processing of the write request.

[fig.19]Fig. 19 is a sequence drawing showing an example of a flow of a processing from a transmission of a read request from a host computer 30 to a completion of the processing of the read request.

[fig.20]Fig. 20 is a sequence drawing showing an example of a flow of a WL control processing.

[fig.21]Fig. 21 is a schematic illustrative drawing showing S301 (a life length prediction of an SSD) of Fig. 20.

[fig.22]Fig. 22 is a sequence drawing showing an example of a flow of an FM WR amount prediction in accordance with a first embodiment (S303 of Fig. 20).

[fig.23]Fig. 23 is a schematic illustrative drawing showing an execution pattern A of an inter-device WL.

[fig.24]Fig. 24 is a sequence drawing showing an example of an execution pattern A of an inter-device WL.

[fig.25]Fig. 25 is a schematic illustrative drawing showing an execution pattern B of an inter-device WL.

[fig.26]Fig. 26 is a sequence drawing showing an example of an execution pattern B of an inter-device WL.

[fig.27]Fig. 27 is a schematic illustrative drawing showing an execution pattern C of an inter-device WL.

[fig.28]Fig. 28 is a schematic illustrative drawing showing an execution pattern D of an inter-device WL.

[fig.29]Fig. 29 is a sequence drawing showing an example of the execution patterns C and D of an inter-device WL.

[fig.30]Fig. 30 is a schematic illustrative drawing showing an execution pattern E of an inter-device WL.

[fig.31]Fig. 31 is a sequence drawing showing an example of an execution pattern E of an inter-device WL.

[fig.32]Fig. 32 is a sequence drawing showing an example of an execution pattern F of

an inter-device WL.

[fig.33]Fig. 33 is a view showing a configuration example of an FM WR amount prediction TBL 13700 in accordance with a second embodiment.

[fig.34]Fig. 34 is a view showing a configuration example of a WA information storage table 13800 in accordance with a second embodiment.

[fig.35]Fig. 35 is a sequence drawing showing an example of a flow of an FM WR amount prediction in accordance with a second embodiment (S303 of Fig. 20).

[fig.36]Fig. 36 is a view showing a configuration example of an FM WR amount prediction TBL 13700 in accordance with a second embodiment.

[fig.37]Fig. 37 is a view showing a summary of an embodiment.

[fig.38]Fig. 38 is a view showing an example of a relationship between an inter-device WL and a virtual volume.

## Description of Embodiments

[0024] Some embodiments of the present invention will be described below in detail with reference to the drawings.

[0025] In the following descriptions, while a wide variety of information will be described in the expression of "xxx table" in some cases, a wide variety of information can be represented by a data structure other than a table. In order to indicate that a wide variety of information is not depended on a data structure, the expression of "xxx table" can also be referred to as "xxx information".

[0026] In the following descriptions, while a number is adopted as the identification information of an element (such as an extent), the identification information of other types (such as a name and an identifier) can also be adopted.

[0027] In the following descriptions, the processing will be described while a "program" is handled as a subject in some cases. In the case in which the program is executed by a processor (for instance, a CPU (Central Processing Unit)) that is included in a controller (a storage controller and an SSD controller), the processor executes the predetermined processing by using a storage resource (such as a memory) and/or a communication interface apparatus (such as a communication port) as it decides proper. Consequently, a subject of a processing can also be a controller or a processor. Moreover, the processor can include a hardware circuit that executes a part or a whole of a processing. A computer program can be installed from a program source. The program source can be a program distribution server or a storage medium that can be read by a computer for instance.

[0028] In the following descriptions, a physical storage medium is represented as "Disk" as a matter of practical convenience in some cases. However, this notation does not always mean that a physical storage medium is a storage medium in a disk shape. In

the following description, Disk indicates an SSD in many cases.

[0029] In the following descriptions, a unit of a period or a time is not restricted. For instance, a unit of a period or a time can be represented as any one or a combination of at least two of a year, a month, a day, an hour, a minute, and a second.

[0030] In the following descriptions, the nonvolatile semiconductor storage medium that is included in an SSD is a flash memory (FM). For the flash memory, an erasing is executed in a unit of a block. The flash memory is a flash memory in which a read/ write is executed in a unit of a page, a flash memory of a NAND type in a quintessential way. However, the flash memory can also be a flash memory of other type as substitute for a NAND type (for instance, a NOR type). Moreover, a non-volatile semiconductor storage medium of other type such as a phase change memory can also be adopted as substitute for a flash memory.

[0031] In the first place, the summary of the present embodiment will be described in the following.

[0032] Fig. 37 is a view showing a summary of an embodiment.

[0033] A storage apparatus 10 is provided with a storage controller 100 and a plurality of physical storage media 11. Each of the physical storage media 11 is a basis of at least two logical storage regions. The logical storage region can be an extent described later or can be a whole or a part of a logical volume (LU: Logical Unit) that is provided to an upper level apparatus (such as a host computer) that is coupled to the storage apparatus 10. The plurality of physical storage media 11 can configure a RAID group. The storage controller 100 executes a leveling of an erasing frequency between physical storage media 11.

[0034] In the first place, the storage controller 100 acquires the internal information that is related to each of physical storage media from the physical storage media 11. The internal information is the information that is an index of a length of life and a consumption status of the physical storage media 11.

[0035] In the case in which the physical storage media 11 is an SSD (storage media that includes a flash memory) for instance, the internal information is the information that includes a total erasing frequency of a plurality of blocks and a real write amount that is a total amount of data that has been written to a flash memory. The storage controller 100 judges a remaining length of life of an SSD based on the internal information. In other words, the storage controller 100 identifies an SSD 11 that is provided with a short length of life and that is a movement source of data and an SSD 11 that is provided with a long length of life and that is the movement source of the data. The storage controller 100 can reduce an overhead caused by an information transfer between the storage controller 100 and the SSD 11 by acquiring the internal information in a unit of an SSD. The storage controller 100 can acquire the internal in-

formation for every logical address range in an SSD and for every physical range (for instance, in a unit of a DIMM).

[0036] In the next place, the storage controller 100 identifies data to be moved from the SSD 11 that is provided with a short length of life. A range of a storage region (a certain address range) in moving data is called an extent for an inter-device WL. The extent can be an address range (a logical region) based on one SSD as shown by a reference symbol 60 and can be an address range (a logical region) that is disposed over a plurality of SSDs as shown by a reference symbol 70. A part of the extent can be a target of the inter-device WL, and an aggregate of a plurality of SSDs can be a target of the inter-device WL. An address range (a logical region) based on one SSD can also be called an extent, and an address range (a logical region) that is disposed over a plurality of SSDs can also be called an extent group.

[0037] A load for a data movement is larger in a small unit such as a block unit and a page unit of a flash memory. Consequently, by executing the inter-device WL in a unit called an extent that is larger than a block and a page, a load of a data movement can be prevented from being increased.

[0038] The storage controller 100 measures a write data amount to an extent for every extent. In the case in which a write data amount to an SSD is large for instance, a rewriting of a block occurs on a number of occasions and an erasing frequency is increased in accordance with that. In the case in which an erasing frequency of a block is increased, the erasing frequency of a block reaches the upper limit and a read/write to the block is not possible. Moreover, an SSD in which such a block is increased cannot be used. Consequently, the storage controller 100 finds an extent in which a write data amount is large and moves data that is included in the extent from an SSD that is provided with a short length of life to an SSD that is provided with a long length of life. In the case in which an extent is disposed over a plurality of SSDs, data in an extent in which a write data amount is large is also moved between SSD groups (a plurality of SSDs).

[0039] By this configuration, an erasing frequency can be leveled between SSDs without increasing a load of the storage controller 100.

[0040] Moreover, a degradation rate of an SSD that is provided with a short length of life can be suppressed by exchanging data of an extent in which a write data amount is large for an SSD that is provided with a short length of life and data of an extent in which a write data amount is small for an SSD that is provided with a long length of life.

[0041] The present embodiment will be described in detail in the following.

## Embodiment 1

[0042] Fig. 1 is a view showing a configuration example of a storage system 10000 that

includes a storage apparatus 10 in accordance with a first embodiment.

[0043] A storage system 10000 is provided with a storage apparatus 10 and a host computer 30.

[0044] The host computer 30 is an example of an upper level apparatus that utilizes the storage apparatus 10. The host computer 30 is an application server for instance. The host computer 30 and the storage apparatus 10 communicate with each other via a SAN (Storage Area Network) 20. As the SAN 20, a fiber channel, an SCSI (Small Computer System Interface), an iSCSI (internet Small Computer System Interface), a USB (Universal Serial Bus), an IEEE 1394 bus, and a SAS (Serial Attached SCSI) can be used for instance. As substitute for the SAN 20, a communication network of other type (such as a LAN (Local Area Network)) can also be adopted. In the figure, there is one host computer 30 and one storage apparatus 10. However, there can be a plurality of host computers 30 and/or a plurality of storage apparatuses 10.

[0045] The host computer 30 issues a control command or the like to the storage apparatus 10 by executing the control software (not shown) that issues a control command or the like to the storage apparatus 10. In the case in which the storage apparatus 10 executes the control command, a modification of a RAID level of a RAID group which the storage apparatus 10 is provided with can be executed. The RAID group is a physical storage medium group that is configured by a plurality of SSDs (or HDDs) and that stores data in accordance with a predetermined RAID level. As a computer that issues a control command to the storage apparatus 10, a computer other than the host computer 30 can also be used.

[0046] The storage apparatus 10 is provided with a storage controller 100 and a Disk Box 110 that is coupled to the storage controller 100.

[0047] The storage controller 100 controls an operation of the storage apparatus 10. The storage controller 100 is provided with a communication interface device, a memory, and a control device that is coupled to the communication interface device and the memory. As a communication interface device, there is a Host I/F 101 that is a communication interface device of a front end and a Disk I/F 107 that is a communication interface device of a back end. As a memory, there is a cache memory 103. As a control device, there is a processor (such as a CPU (Central Processing Unit)) 104. The Host I/F 101, the cache memory 103, the processor 104, and the Disk I/F 107 are coupled to an internal network 102 by a dedicated connection bus such as a PCI (Peripheral Component Interconnect) and can be communicated with each other via the internal network 102. The cache memory 103 is coupled to an internal network 102 by a dedicated connection bus such as a DDR3 (Double Data Rate3).

[0048] The Host I/F 101 is an interface by which the storage apparatus 10 is coupled to the SAN 20.

[0049] The internal network 102 is a network for coupling devices that exist in the storage apparatus 10 to each other. The internal network 102 includes a switch. As substitute for the internal network 102, the ASICs (Application Specific Integrated Circuit) that is provided with a switch function, a DMA transfer, and an assist function such as a RAID operation can also be used.

[0050] The processor 104 controls the entire of the storage apparatus 10. A plurality of processors 104 exist. In this case, the plurality of processors 104 can control the storage apparatus 10 in consort with each other or while paying a share.

[0051] The cache memory 103 is a region for storing a computer program and data that are required for controlling the storage apparatus 10 by the processor 104.

[0052] The Disk I/F 107 is an interface for coupling the storage controller 100 and the Disk Box 110 to each other.

[0053] The Disk Box 110 is provided with a plurality of Disks of different types (for instance, nonvolatile physical storage media such as an HDD 111 and an SSD 700). The RAID group is configured by Disks of the same type. A logical volume as a storage space of user data is provided from each RAID group. In the figure, the HDD 111 and the SSD 700 are shown as the Disk that configures the Disk Box 110. However, the HDD 111 can also be omitted.

[0054] Fig. 11 is a view showing a configuration example of the SSD 700.

[0055] The SSD 700 is provided with a plurality of flash memories and an SSD controller 710 that is coupled to the flash memories. The SSD controller 710 controls an operation of the SSD 700.

[0056] The SSD controller 710 is provided with a communication interface device, a memory, and a control device that is coupled to the communication interface device and the memory. As the communication interface device, there can be mentioned for instance a Disk I/F 711 that is a communication interface device of a front end and a Flash I/F 717 that is a communication interface device of a back end. As the memory, there can be mentioned for instance a cache memory 716. As the control device, there can be mentioned for instance a processor 713. The Disk I/F 711, the processor 713, the cache memory 716, and the Flash I/F 717 are coupled to each other via an internal network 712.

[0057] The Disk I/F 711 is coupled to the Disk I/F 107 of the storage controller 100 via a dedicated communication bus. The internal network 712 is a network that is configured to couple devices to each other. The internal network 712 can include a switch and can be replaced by the ASICs that are provided with a switch function. The processor 713 controls the entire of the SSD 700. The cache memory 716 is a region that is configured to store a computer program and data that are required for controlling the SSD 700 by the processor 713. The Flash I/F 717 is an interface that is configured to

couple the SSD controller 710 and a flash memory 780 to each other.

[0058] In the present embodiment, an SSD is a storage media that is provided with a plurality of flash memories and a controller that controls the plurality of flash memories, and the external shape of the SSD is not limited to a form factor.

[0059] Fig. 12 is a view showing a configuration example of the flash memory 780.

[0060] The flash memory 780 is provided with a plurality of blocks 782 and a flash memory controller 781 that is coupled to the blocks. The flash memory controller (hereafter referred to as a flash controller) 781 controls the flash memory 780. Each of blocks 782 is configured by a plurality of pages 783. A read of data from the flash memory 780 and a write of data to the flash memory 780 are executed in a unit of a page. A data erasing is executed in a unit of a block.

[0061] The flash memory 780 is a flash memory of a NAND type for instance. Data cannot be overwritten to the flash memory of a NAND type. Consequently, in the case in which new data is tried to be written to a page to which data has been written, data is erased and data is written to a page in which data has been erased.

[0062] The SSD controller 710 executes a leveling of an erasing frequency to a plurality of blocks 782 in the SSD 700. The SSD controller 710 stores data that is provided with a small update frequency into a block that is provided with a large erasing frequency and stores data that is provided with a large update frequency into a block that is provided with a small erasing frequency to level an erasing frequency to a block in the SSD 700.

[0063] Moreover, the SSD controller 710 can also execute a leveling of an erasing frequency by selecting a free block that is provided with a small erasing frequency and storing data into the block. In this case, a plurality of blocks can be divided into a plurality of groups depending on an erasing frequency, and a block can be selected from a group that is provided with a small erasing frequency.

[0064] A long length of life of an SSD can be achieved by leveling an erasing frequency of a plurality of blocks in an SSD.

[0065] Fig. 13 is a view showing a configuration example of an address space of an SSD.

[0066] A logical address space 900 and a physical address space 800 exist for the SSD 700. The logical address space 900 is a unique address space that is provided to the storage controller 100 by the SSD 700. On the other hand, the physical address space 800 is an address space in which actual data is stored. The SSD controller 710 dynamically modifies a mapping of an address range (region) that configures the logical address space 900 and an address range (region) that configures the physical address space 800, thereby implementing a WL or the like.

[0067] A correspondence between the logical address space 900 and the physical address space 800 is managed using a logical physical conversion table 23100 described later by the processor 713. In the present embodiment, the storage controller 100 does not

manage a logical physical conversion table in an SSD in a direct manner. Consequently, in the case in which the processor 713 does not notify the storage controller 100 of the SSD internal information, the storage controller 100 cannot comprehend a correspondence relationship between the logical address space 900 and the physical address space 800.

[0068] In the present embodiment, the SSD 700 manages a chunk that is an aggregate of blocks. In Fig. 13, a chunk 810 is formed by blocks 811 to 814, a chunk 820 is formed by blocks 821 to 824, and a chunk 830 is formed by blocks 831 to 834. Moreover, the chunk 810 provides a logical address (LBA) space #a 901, the chunk 820 provides a logical address (LBA) space #b 902, and the chunk 830 provides a logical address (LBA) space #c 903.

[0069] In the next place, the capacity virtualization technique will be described in the following.

[0070] The capacity virtualization technique (for instance, Thin Provisioning) is a technique for providing a virtual capacity larger than a physical capacity which the storage apparatus 10 is provided with to the side of the host computer 30. The host computer 30 accesses a virtual logical volume (a virtual volume). For the capacity virtualization technique, in the case in which the storage apparatus 10 receives a write request, a physical storage region is allocated to a virtual region (a virtual extent) of the write destination of the data of the write request. In the present embodiment, a unit of a physical storage region that is allocated for the capacity virtualization technique is called an extent unit. Moreover, a size of the extent can be rich in diversity in the range from several MB to several GB.

[0071] The capacity virtualization technique will be described in detail with reference to Fig. 2 in the following. The RAID group (RG) is configured by the Disks (such as SSDs) of the same type. A Pool Volume 500 is configured based on at least one RG.

[0072] The extent is a storage region that is obtained by dividing the Pool Volume 500, that is, a part of the Pool Volume 500. In the figure, the Pool Volume 500 is configured based on three RGs of an RG 200, an RG 300, and an RG 400. The RG 200 will be described in the following.

[0073] The RG 200 is configured by a Disk1 210, a Disk2 220, and a Disk3 230. Moreover, the RG 200 constructs a RAID 5, and a Data (D in the figure) and Parity (P in the figure) are stored into a stripe line based on Disks (210 to 230). Here, a "stripe line" is a storage region that is configured by the same address region of a plurality of Disks that configure the same RG. That is, the stripe line is arranged on a plurality of Disks that configure the same RG. The set of the Data and Parity is stored into the stripe line. In the figure, a D 211, D 221, and P 231 are stored into one stripe line 250 for instance. In the case in which at least two same Data or Parity of the same stripe line exist in the

same Disk by a data movement, the redundancy of the RAID is degraded. Consequently, the storage controller 100 controls a stored destination of data in such a manner that at least two Data or Parity that have been stored into one stripe line do not exist in the same Disk as a result of an execution of an inter-device WL.

[0074] A storage region based on the RG is divided into a plurality of extents. Moreover, a storage region based on the RG is made of a logical address space 900 that is provided by a plurality of SSDs.

[0075] In the next place, a configuration of an extent will be described in the following. For instance, each of extents is configured by at least one stripe line.

[0076] In Fig. 2, an extent 240, an extent 241, and an extent 242 are formed from the RG 200. An extent 340, an extent 341, and an extent 342 are formed from a storage region of the RG 300. An extent 440, an extent 441, and an extent 442 are formed from a storage region of the RG 400. The extent 240 is configured by two stripe lines, that is, a stripe line in which Data 214, Parity 224, and Data 234 have been stored and a stripe line in which Parity 215, Data 225, and Data 235 have been stored. Similarly, an extent 301 and an extent 401 are also formed from the RG 300 and the RG 400.

[0077] The extent is a storage region of which a size is larger than that of a block. A size of the extent is N times (N is an integer number that is equal to or larger than 2) of that of a block for instance. Chunk described later is an aggregate of a plurality of extents.

[0078] A virtual volume 600 is a virtual logical volume that is configured to store user data by the host computer 30. A capacity that is defined as a capacity of the virtual volume 600 can be a storage capacity larger than a total capacity of the storage media that is included in the storage apparatus 10. The virtual volume 600 is configured by virtual extents of arbitrary number of 601 to 607. For instance, although one extent is corresponded to one virtual extent shown in Fig. 2, a plurality of extents can also be corresponded to one virtual extent. The virtual volume 600 is provided with a virtual address (a logical address that configures a virtual volume), and the virtual address is divided in a predetermined range to configure the virtual extent.

[0079] The virtual extents 601 to 604 that are shown by solid lines are virtual extents in which an extent has been allocated from the RGs 200, 300, and 400. In other words, an extent 301 has been allocated to a virtual extent 601, an extent 242 has been allocated to a virtual extent 602, an extent 402 has been allocated to a virtual extent 603, and an extent 240 has been allocated to a virtual extent 604.

[0080] The virtual extents 605 to 607 that are shown by dotted lines are not virtual extents to which an address that has been specified as a write destination of data belongs. Consequently, an extent has not been allocated to the virtual extents 605 to 607.

[0081] Fig. 38 is a view showing an example of a relationship between an inter-device WL and a virtual volume.

[0082] In the case in which data is moved between SSDs, the host computer 30 must modify an access destination of data to be an SSD of a movement destination. However, in the case in which an address is converted by using the virtual volume 600, the host computer 30 can access data of a movement destination without modifying an access destination. In other words, an association with a virtual address of the virtual extent 610 is changed from a logical address of the extent 70 of a movement source to a logical address of the extent 80 of a movement destination by the storage controller 100. The host computer 30 can execute an inter-device WL by accessing the virtual volume 600 without modifying an address of an access destination.

[0083] The case in which a virtual volume is used for an address conversion is an example, and an address conversion can also be executed without using a virtual volume.

[0084] Fig. 3 is a view showing a configuration example of a cache memory 103 that is included in the storage controller 100.

[0085] The cache memory 103 is provided with a Program Area 12000, a Table Area 13000, and a Data Cache Area 14000. The Program Area 12000 and the Table Area 13000 are regions in which a program for controlling the storage apparatus 10 and a wide variety of tables are stored. The Data Cache Area 14000 is a region that is used for storing user data on a temporary basis.

[0086] The Program Area 12000 stores a Write I/O Program 12100, a Read I/O Program 12200, a life length management Program 12300, an SSD information acquisition Program 12400, and an inter-SSD WL Program 12500.

[0087] The Write I/O Program 12100 is a program for processing a write request from the host computer 30. The Read I/O Program 12200 is a program for processing a read request from the host computer 30. The life length management Program 12300 is a program for managing a life length of a wide variety of Disks such as an SSD 700 by the processor 104. The SSD information acquisition Program 12400 is a program for acquiring the internal information of the SSD 700. The inter-SSD WL Program 12500 is a program for executing the inter-device WL.

[0088] The Table Area 13000 is provided with a Disk management Table (hereafter Table is referred to as TBL) 13100, an RG management TBL 13200, a Pool management TBL 13300, an extent management TBL 13400, a virtual volume management TBL 13500, a statistics information management TBL 13600, an FM (Flash Memory) write amount prediction TBL 13700, and a WA information storage TBL 13800.

[0089] The Disk management TBL 13100 is a table for storing the information related to Disk that has been stored into a Disk Box 110. The RG management TBL 13200 is a table for storing the information related to the RAID group. The Pool management TBL 13300 is a table for storing the information of a Pool Volume. The extent management TBL 13400 is a table for storing the information related to an extent. The

virtual volume management TBL 13500 is a table for storing the information related to a virtual volume. The statistics information management TBL 13600 is a table for storing a wide variety of information related to a performance of the storage apparatus 10. The FM write amount prediction TBL 13700 is a table that is used for predicting a data write amount in an SSD. The WA information storage TBL 13800 is a table for storing a predicted value of a rate of a write data amount that is increased by a processing in an SSD based on a write I/O pattern to the SSD.

[0090] The user data 14100 that conforms to a write request and a read request is stored into the Data Cache Area 14000 on a temporary basis. The user data 14100 is data that is used by the host computer 30.

[0091] Fig. 4 is a view showing a configuration example of a Disk management TBL 13100.

[0092] The information that is managed by the Disk management TBL 13100 is used for judging a length of life of each Disk in large part. The Disk management TBL 13100 is provided with a Disk# 13101, a Disk Type 13102, an information update date 13103, a remaining guarantee period 13104, a total Write amount 13105, a remaining erasing frequency 13106, and a remaining life length 13107 for every Disk.

[0093] The Disk# 13101 is an identifier of the Disk and a unique number. The Disk Type 13102 indicates a type of the Disk and indicates an SSD (SLC), an SSD (MLC), and an HDD for instance. Here, there are two types of an SLC (Single Level Cell) type and an MLC (Multi Level Cell) type depending on a type of a NAND flash memory to be used. The SLC is a flash memory of a high speed, a long life length, and a small capacity, and enables a block erasing of hundreds of thousands of order to tens of thousands of order. On the other hand, the MLC is a flash memory of a low speed, a short life length, and a large capacity, and enables a block erasing of tens of thousands of order to thousands of order.

[0094] The information update date 13103 indicates a latest date when the information related to the Disk was updated (for instance, a date when the life length information was updated). Here, the information update date 13103 is represented by the number of days. The remaining guarantee period 13104 indicates a remaining period of a guarantee period of the Disk decided by a vender.

[0095] The guarantee period of the Disk is a period in which a normal operation of the Disk is guaranteed (for instance, 5 years). A flash memory is deteriorated by an increase in an erasing frequency due to a write of data. In the case in which a length of life of a flash memory is reached, a read/write of data is not possible, or a data retention characteristic is extremely degraded. Consequently, as a period in which the Disk can be normally used, a threshold value is specified based on an erasing frequency and a total write amount in advance. In the case in which a usage period of the Disk exceeds the guarantee period, the Disk is exchanged.

[0096] The total Write amount 13105 is an integrated value of a write amount that has occurred in an SSD (an amount of data that has been written to a flash memory in an SSD), and is updated based on the SSD internal information. The remaining erasing frequency 13106 indicates a value that conforms to the number of times of a block erasing that can be executed for an SSD. Here, the remaining erasing frequency 13106 is represented in percentage terms and is updated based on the SSD internal information. In the case in which the numerical value becomes zero, it is indicated that the number of times of a erasing of all blocks in a flash memory reaches the upper limit.

[0097] The information that is registered as the total Write amount 13105 and the remaining erasing frequency 13106 is included in the information that is acquired from the SSD. The total Write amount 13105 and the remaining erasing frequency 13106 are updated based on the information that has been acquired from each SSD by the storage controller 100. The information is acquired by using a general-purpose interface such as an S.M.A.R.T. (Self-Monitoring Analysis and Reporting Technology).

[0098] The remaining life length 13107 indicates the remaining period of time until the remaining erasing frequency 13106 of the SSD reaches zero. The remaining life length 13107 is a numerical value that is predicted by the processor 104 (the life length management Program 12300). Here, the remaining life length 13107 is represented by the number of days.

[0099] That the remaining life length 13107 is below the remaining guarantee period 13104 means that a load is concentrated to the specific SSD due to a dispersion of an I/O for instance, the SSD is provided with a load higher than a load that has been assumed by a vender, and the SSD reaches the length of life before the expiration date for use that is guaranteed by a vender. An inter-device WL is executed between SSDs, thereby avoiding a concentration of a write to a specific SSD and lengthening a length of life of the SSD.

[0100] Moreover, the remaining life length 13107 can be calculated based on the information update date 13103, the total Write amount 13105, and the remaining erasing frequency 13106 or the like. A calculation method of the remaining life length will be described later.

[0101] The Disk management TBL 13100 stores the information related to a life length management of the Disk mainly. Consequently, it is not necessary that the information of the total Write amount 13105 to the remaining life length 13107 is an acquisition target of the processor 104 for the Disk that is not provided with a restriction of the number of times of rewrite (for instance, an HDD).

[0102] Fig. 5 is a view showing a configuration example of an RG management TBL 13200.

[0103] The RG management TBL 13200 is provided with an RG# 13201, a Disk Type

13202, a RAID Level 13203, a RAID configuration 13204, and a Disk# 13205 for every RG.

[0104] The RG# 13201 is an identifier of an RG and a unique number. The Disk Type 13202 indicates a type of the Disk that configures an RG. The RG is configured by the Disk of the same type.

[0105] The RAID Level 13203 indicates a RAID level of an RG and is a variety of values such as RAID 1 + 0, 1, 3, 4, 5, and 6 for instance. The RAID configuration 13204 indicates the number of data Disks that configures the RAID (Disks that store data) and the number of parity Disks (Disks that store parity). The Disk# 13205 indicates the number of the Disk that configures an RG and includes an effective numerical value that is equivalent to a numerical value of the RAID configuration 13204.

[0106] Fig. 6 is a view showing a configuration example of a Pool management TBL 13300.

[0107] The Pool management TBL 13300 is provided with a Pool# 13301, an RG# 13302, an RG remaining capacity 13303, and a Pool remaining capacity 13304 for every Pool.

[0108] The Pool# 13301 is an identifier of a Pool and a unique number. The RG# 13302 indicates an RG number of all RGs that configure the Pool. The RG remaining capacity 13303 indicates a remaining capacity for every RG. The Pool remaining capacity 13304 indicates a remaining capacity for the Pool and is equivalent to a sum total value of the RG remaining capacity 13303 of an RG that configures the Pool. Moreover, the RG remaining capacity 13303 and the Pool remaining capacity 13304 are reduced in the case in which a write of data occurs in an unallocated region of an extent of a virtual volume, and are updated to be the reduced value by the processor 104 (Write I/O Program 12100).

[0109] Fig. 7 is a view showing a configuration example of an extent management TBL 13400.

[0110] The extent management TBL 13400 is provided with an extent# 13401, an RG# 13402, a Size 13403, a Disk Type 13404, a Disk# 13405, a Stripe# 13406, a Start-LBA 13407, a Last-LBA 13408, an allocation flag 13409 for every extent.

[0111] The extent# 13401 is an identifier of an extent and a unique number. The RG# 13402 indicates a number of an RG that is a basis of an extent. The Size 13403 indicates a capacity of an extent. Here, a unit of the Size 13403 is Byte.

[0112] The Disk Type 13404 indicates a type of the Disk that is included in an RG that is a basis of an extent.

[0113] The Disk# 13405, the Stripe# 13406, the Start-LBA 13407, and the Last-LBA 13408 indicate a Disk on which the extent is prepared based, a stripe line that configures the Disk on which the extent is prepared based, and a physical space from the number of LBA to the number of LBA of a stripe line by which the extent is prepared.

[0114] The allocation flag 13409 indicates whether or not the extent has been allocated to a

virtual volume. Here, a flag 13409 of an allocated extent is "done" and a flag 13409 of an unallocated extent is "not done".

[0115] Fig. 8 is a view showing a configuration example of a virtual volume management TBL 13500.

[0116] The virtual volume management TBL 13500 is provided with a virtual volume # 13501, a virtual capacity 13502, a real used capacity 13503, a virtual extent # 13504, and an allocating extent # 13505 for every virtual volume.

[0117] The virtual volume # 13501 is an identifier of a virtual volume and a unique number. The virtual capacity 13502 is a virtual capacity of a virtual volume. The virtual capacity is provided to the host computer 30. The real used capacity 13503 is a sum total value of a capacity of an extent that has been allocated to a virtual volume as a practical matter.

[0118] The virtual extent # 13504 is an identifier of a virtual extent that is included in a virtual volume. It is indicated that virtual extents #0 to #n are included in a virtual volume #0.

[0119] The allocating extent# 13505 is an identifier (a number) of an extent that has been allocated to a virtual volume. The storage controller 100 manages an allocating status of an extent. In other words, an extent #0 is allocated to a virtual extent #0, and an extent #100 is allocated to a virtual extent #1. However, the allocating extent # 13505 of a virtual extent to which an extent has been unallocated is "-". In other words, an extent that is a physical storage region is not allocated to a virtual extent #n.

[0120] Fig. 9 is a view showing a configuration example of a statistics information management TBL 13600.

[0121] The statistics information management TBL 13600 is the information that is related to an access status such as a read/write to each extent. The storage controller 100 monitors a read/write status to each extent and configures the statistics information management TBL 13600.

[0122] The statistics information management TBL 13600 is provided with an Disk# 13601, an extent# 13602, a WR (an abbreviation of Write) IOPS (Input Output Per Second) 13603, an RD (an abbreviation of Read) IOPS 13604, an average WR I/O Size 13605, a WR I/O Pattern 13606, a WR rate 13607, an WR amount 13608, and an RD amount 13609 for every Disk. At least one of the information 13601 to 13609 is referred to as statistics information in some cases in the following.

[0123] The Disk# 13601 is an identifier of the Disk and a unique number. The extent# 13602 is an identifier of an extent based on the Disk and a unique number.

[0124] The WR IOPS 13603 and the RD IOPS 13604 indicate an occurrence frequency of a WR (write) I/O and an RD (read) I/O that have occurred in an address range that is corresponded to an extent in an address range of the Disk. Here, IOPS is an ab-

breviation of Input/Output Per Second.

[0125] The average WR I/O Size 13605 indicates an average size of data that is associated with a WR I/O request from the host computer 30.

[0126] The WR I/O Pattern 13606 indicates whether a WR I/O is a random pattern (RND) or a sequential pattern (SEQ). The WR rate 13607 indicates a rate of a write I/O to all I/O. The WR amount 13608 and the RD amount 13609 indicate a total amount of WR data that has occurred in an extent (data in which an address range that is corresponded to an extent in an address range of the Disk is a write destination) and a total amount of RD data that has occurred in an extent (data in which an address range that is corresponded to an extent in an address range of the Disk is a read source), respectively.

[0127] In the case in which an extent is disposed over a plurality of Disks, a value of an IOPS of the extent can be calculated as a total sum or an average value of values that are managed for every disk.

[0128] Fig. 9 shows the case in which the statistics information for every Disk is managed. However, the statistics information can also be managed for every RG. In the case in which the statistics information is acquired for every RG, the statistics information can be acquired without distinguishing the Disk for an extent that is disposed over a plurality of Disks.

[0129] In the present embodiment, the statistics information is monitored in a unit of an extent. Consequently, a monitoring load can be reduced as compared with the case in which the statistics information is monitored in a block unit or a page unit of a flash memory.

[0130] The storage controller 100 monitors an access status for an extent that is being allocated based on the virtual volume management TBL 13500. The statistics information is an accumulated value from an allocation of an extent to a virtual volume and a trend of a read/write. The statistics information can also be an accumulated value in a unit time.

[0131] Depending on a timing of a judgment of necessity of an inter-device WL, the statistics information of an extent of a target of a judgment of necessity can be reset. For instance, the storage controller 100 can reset only the statistics information that is related to an extent that has been moved by an inter-device WL.

[0132] Fig. 10 is a view showing a configuration example of an FM WR amount prediction TBL 13700 in the present embodiment.

[0133] The FM WR amount prediction TBL 13700 is configured based on the statistics information management TBL 13600. The FM WR amount prediction TBL 13700 is provided with a Disk# 13701, an extent# 13702, a WR amount 13706, and an FM WR predicted amount 13708 for every Disk.

[0134] The Disk# 13701 is an identifier of the Disk and a unique number. The extent#

13702 is an identifier of an extent and a unique number. In the case in which an extent is disposed over a plurality of SSDs that configure an RG, the extent described here means a part of an extent to be precise.

[0135] The WR amount 13706 indicates a total amount of WR data in which an address range that is corresponded to an extent in an address range of the Disk is a write destination.

[0136] The FM WR predicted amount 13708 indicates a predicted amount of data that is written to a flash memory as a practical matter.

[0137] In the present embodiment, the WR amount 13706 for an extent (or an extent part) is considered as a real WR amount for a region that is corresponded to the extent for the Disk (SSD). For instance, in the case in which a write data amount to an extent is large, a rewrite of data to a block for an SSD that is a basis of the extent occurs on a number of occasions, and an erasing frequency of a block is increased. This is because a length of life of an SSD that is provided with the block is shortened in the case in which an erasing frequency of a block is increased. Consequently, a highly accurate inter-device WL can be executed by deciding an extent of a movement target based on a write data amount to the extent in the case in which an inter-device WL is executed.

[0138] Fig. 14 is a view showing a configuration example of a cache memory 716 of the SSD 700.

[0139] The cache memory 716 is provided with a Program Area 22000, a Table Area 23000, and a Data Cache Area 24000.

[0140] The Program Area 22000 and the Table Area 23000 are regions in which a program for controlling the SSD 700 and a wide variety of tables are stored. The Data Cache Area 24000 is a region that is used to store user data on a temporary basis.

[0141] The Program Area 22000 is provided with a Write I/O Program 22100, a Read I/O Program 22200, an in-SSD WL Program 22300, a free capacity generation Program 22400, and an SSD internal information communication Program 22500.

[0142] The Write I/O Program 22100 is a program for processing a write request from the storage controller 100. The Read I/O Program 22200 is a program for processing a read request from the storage controller 100. The in-SSD WL Program 22300 is a program for executing a WL in the SSD. The free capacity generation Program 22400 is a program for executing a reclamation processing that is executed for avoiding a depletion of a free space of the SSD 700.

[0143] The SSD internal information communication Program 22500 is a program for creating the internal information of an SSD in accordance with a request of the storage controller 100 and notifying the storage controller 100 of the SSD internal information.

[0144] The Table Area 23000 is provided with a logical physical conversion TBL 23100 and a statistics information management TBL 23200.

[0145] The logical physical conversion TBL 23100 is a table for managing a correspondence between a logical address space and a physical address space of the SSD 700. The statistics information management TBL 23200 is a table for managing the statistics information of the SSD.

[0146] The Data Cache Area 24000 is provided with a region that is used to store the user data 24100. Here, the user data 24100 is data that is written to a block and data that has been read from a block.

[0147] Fig. 15 is a view showing a configuration example of the logical physical conversion TBL 23100.

[0148] The logical physical conversion TBL 23100 is provided with a Start-LBA 23100, a Chunk# 23102, a Block# 23103, and a Page# 23104.

[0149] The Start-LBA 23100 indicates a starting position of an LBA that is provided to the storage controller 100. Since an I/O unit of a flash memory is a page, a numerical value of the Start-LBA 23100 is managed by a multiple number of a page size. The Chunk# 23102, the Block# 23103, and the Page# 23104 indicate the information of a physical address space that is corresponded to the Start-LBA 23100. The logical physical conversion TBL 23100 is provided with the information related to a chunk, a block, and a page that are corresponded to each Start-LBA 23100.

[0150] Fig. 16 is a view showing a configuration example of the statistics information management TBL 23200 in the SSD.

[0151] The information that is included in the statistics information management TBL 23200 is notified to the storage controller 100. The statistics information management TBL 23200 is provided with the information 23210 of a unit of an SSD, and the information 23220 of a unit of a Chunk.

[0152] The statistics information management TBL 23200 can be provided with at least one of the information 23210 of a unit of an SSD and the information 23220 of a unit of a Chunk (an aggregate of a plurality of extents). In the present embodiment, the statistics information management TBL 23200 is provided with the information 23210 of a unit of an SSD.

[0153] The information 23210 of a unit of an SSD is provided with an SSD# 23211, a WR amount 23212, and a remaining erasing frequency 23213.

[0154] The SSD# 23211 indicates an identifier of an SSD. The WR amount 23212 indicates a total amount of data that been written to an SSD. The remaining erasing frequency 23213 indicates a remaining frequency of erasing until a length of life for an SSD. The remaining erasing frequency 23213 is a value (for instance, a total sum) based on a remaining erasing frequency of a block that is included in an SSD.

[0155] The information 23220 of a unit of a Chunk is provided with a Chunk# 23221, a WR amount 23222, and a remaining erasing frequency 23223.

[0156] The Chunk# 23221 indicates an identifier of a Chunk. The WR amount 23222 indicates a total amount of data that been written to a Chunk. The remaining erasing frequency 23223 indicates a remaining frequency of erasing for a Chunk. The remaining erasing frequency 23223 is a value (for instance, a total sum) based on a remaining erasing frequency of a block that configures the Chunk.

[0157] As the information that is transmitted from an SSD, any one of the information 23210 of a unit of an SSD and the information 23220 of a unit of a Chunk.

[0158] A numerical value of the present table 23200 is updated (for instance, can be added) in the case in which a write of data or the erasing of data occurs. The present table 23200 can store not only the information related to a write but also the information related to a read.

[0159] Fig. 17 is a view showing a configuration example of the SSD internal information 25000.

[0160] The SSD internal information 25000 is the information that is transmitted from an SSD. The SSD internal information 25000 is provided with the address information 25100 and the statistics information 25200 for instance.

[0161] The address information 25100 is the information that is created based on the logical physical conversion TBL 23100. The logical physical conversion TBL 23100 is configured to notify the storage controller 100 of a correspondence status between a logical address and a physical address for instance. The logical physical conversion TBL 23100 is the information that indicates a Chunk that is corresponded to an LBA to be more precise for instance.

[0162] In the case of the present embodiment, since a processing is executed based on the information 23210 in a unit of an SSD, the SSD address information 25100 is not notified from the SSD controller 710 to the storage controller 100. In the embodiment 3 described later, the address information 25100 is notified from the SSD controller 710 to the storage controller 100.

[0163] The statistics information 25200 is the information that is created based on the statistics information management TBL 23200. The statistics information 25200 (see Fig. 16) is the information that is configured to make the storage controller 100 to create the information related to a life length of each SSD for instance. To be more precise, the statistics information 25200 includes a data amount that has been written to an SSD and a remaining number of the erasing enable number of times for instance.

[0164] An amount of the information that is included in the SSD internal information 25000 is varied depending on a notification granularity in a unit of an SSD and a unit of a Chunk (see the statistics information management TBL 23200). An amount of the information in the case of a unit of an SSD is smaller than that in the case of a unit of a Chunk. Consequently, an overhead of a communication is smaller in the case of a unit

of an SSD. On the other hand, an SSD notifies the storage controller of the information in a unit of a Chunk, it is necessary to transfer the information for a plurality of Chunks in the SSD. However, the storage controller can comprehend the internal information in more detail and can execute the inter-device WL.

[0165] As described above in the present embodiment, a notification unit is a unit of an SSD. In the present embodiment, the address information 25100 is not included and the statistics information 25200 is provided with only the information 23210 in a unit of an SSD of the statistics information management TBL 23200.

[0166] The information that is notified from the SSD controller 710 to the storage controller as the SSD internal information 25000 can include the information related to a life length of each SSD that has been created by the SSD controller 710 in addition to the information configured to make the storage controller 100 to create the information related to a life length of each SSD.

[0167] The SSD can also manage the information in a plurality of granularity such as a unit of an SSD and a unit of a Chunk, and can modify the granularity of the information to be notified of.

[0168] Fig. 18 is a sequence drawing showing an example of a flow of a processing from a transmission of a write request from the host computer 30 to a completion of the processing of the write request.

[0169] The host computer 30 transmits write data and a write request to the storage controller 100 (S100).

[0170] The storage controller 100 receives the write data and the write request and judges whether or not an extent has been allocated to a write destination range (an address range for a virtual volume) that is indicated by the write request based on the virtual volume management table 13500 (S101). The storage controller 100 can allocate an extent to a write destination range that is indicated by the write request while taking the opportunity of receiving the write request. In the case in which an extent has been unallocated as a result of the judgment (S101: No), the storage controller 100 finds an unallocated extent, allocates a virtual extent of the extent to the write destination range (S102), and stores the received data to the cache memory 103 (S103). The storage controller 100 can allocate an extent in the case in which the storage controller 100 transmits data from the cache memory 103 to the SSD.

[0171] On the other hand, in the case in which an extent has already been allocated as a result of the judgment of the S101 (S101: Yes), the storage controller 100 proceeds to the S103.

[0172] In the next place, the storage controller 100 transmits the data that has been stored into the cache memory 103 and the write request to the SSD controller 710 of an SSD 700 that is a stored destination of data (S104).

[0173] The SSD controller 710 receives the data and the write request from the storage controller 100, decides a flash memory that is a stored destination of the received data based on the logical physical conversion TBL 23100, and stores the data into the decided flash memory (S105).

[0174] After that, the SSD controller 710 updates the statistics information management TBL 23200 related to the received data (S106). The SSD controller 710 transmits a completion response of a transfer as a response of the write request that has been received in the S105 to the storage controller 100 (S107).

[0175] The storage controller 100 receives the completion response of a transfer from the SSD controller 710, and updates the statistics information management TBL 23200 (S108).

[0176] After that, the storage controller 100 transmits a completion response of a transfer as a response of the write request that has been received in the S101 to the host computer 30 (S109).

[0177] The host computer 30 receives the completion response of a transfer from the storage controller 100, and terminates a sequence of processing (S110).

[0178] The storage controller 100 can transmit the completion response of a transfer to the host computer 30 at the time point when data is stored into the cache memory 103, and then transmit data from the cache memory 103 to an SSD at an arbitrary timing. The present processing is called a post line processing, and is known as one means for improving a write processing performance of the storage controller 100.

[0179] In Fig. 18, an operation of the storage controller 100 is an operation that has been executed by an execution of an I/O Program 12100, and an operation of the SSD controller 710 is an operation that has been executed by an execution of a Write I/O Program 22100.

[0180] Fig. 19 is a sequence drawing showing an example of a flow of a processing from a transmission of a read request from the host computer 30 to a completion of the processing of the read request.

[0181] The host computer 30 transmits write data and a read request to the storage controller 100 (S200).

[0182] The storage controller 100 receives the read request of data, identifies an SSD that is a basis of an extent that has been allocated to a read source range (an address range of a virtual volume) that conforms to the read request, and transmits the read request of data (also referred to as a staging request) to the SSD controller 710 of the identified SSD (S201).

[0183] The SSD controller 710 receives the read request of data, identifies a physical address range that is corresponded to a logical address range that conforms to the read request based on the logical physical conversion TBL 23100, reads data from the

physical address range (at least one page), and transmits the read data to the storage controller 100 (S202). At this time, the SSD controller 710 updates the statistics information management TBL 23200 (S203).

[0184] The storage controller 100 receives data from the SSD controller 710 as a response of the read request that has been transmitted in the S201, stores the received date into the cache memory 103 (S204), and updates the statistics information management TBL 13600 (S205). After that, the storage controller 100 transmits data that has been stored in the S204 to the host computer 30 (S206).

[0185] The host computer 30 receives data from the storage controller 100 as a response of the read request that has been transmitted in the S200, and terminates a sequence of processing (S207).

[0186] In Fig. 19, an operation of the storage controller 100 is an operation that has been executed by an execution of a Read I/O Program 12200, and an operation of the SSD controller 710 is an operation that has been executed by an execution of a Read I/O Program 22200.

[0187] Fig. 20 is a sequence drawing showing an example of a flow of an inter-device WL control processing.

[0188] The storage controller 100 requests a notification of the internal information 25000 (see Fig. 17) to the SSD controller 710, and receives the internal information of the SSD 700 from the SSD controller 710 in response to the request (S300).

[0189] The storage controller 100 updates the Disk management TBL 13100 and calculates a remaining length of life of each SSD based on the internal information (S301). A method for calculating a remaining length of life will be described later with reference to Fig. 21.

[0190] In the next place, the storage controller 100 judges whether or not the inter-device WL is required (S302). This can be judged by checking whether or not there is an SSD that reaches a length of life prior to a guarantee period (a remaining life length 13107 is shorter than a remaining guarantee period 13104) or whether or not there is a dispersion of a write amount between SSDs (a total write amount 13105 is not equalized) for instance. In the case in which an extent is disposed over a plurality of SSDs and there is at least one SSD that reaches a length of life prior to a guarantee period among the plurality of SSDs, the storage controller 100 can also execute an inter-device WL.

[0191] In the case in which the inter-device WL is required as a result of the judgment of the S302 (S302: Yes), the storage controller 100 proceeds to the S303 and predicts a write amount of a flash memory for every extent based on the statistics information (13701 to 13706) of the storage controller 100 and the internal information (23211 to 23213) of the SSD 700 (S303: the details will be described with reference to Fig. 22).

[0192] In the next place, the storage controller 100 executes the inter-device WL based on the predicted result of the S303 (S304: the details will be described with reference to Figs. 22 to 32). After that, the storage controller 100 updates the information related to an extent that has stored data that has been moved in the inter-device WL (for instance, the information that has been stored into a table) (S305). Subsequently, the storage controller 100 terminates the present processing.

[0193] In the case in which the inter-device WL is not required as a result of the judgment of the S302 (S302: No), the storage controller 100 terminates the present processing.

[0194] The present processing can also be executed at any timing. The present processing can also be executed at the same interval of time (for instance, every other day). Moreover, the present processing can be executed in conjunction with an I/O request of the host computer 30. Moreover, the present processing can be executed at the timing when a specific command is received. Moreover, the present processing can be executed in the case in which a user instructs an execution opportunity of the present processing from the control software of the host computer 30.

[0195] In Fig. 20, an operation of the storage controller 100 is an operation that has been executed by executing the inter-SSD WL Program 12500.

[0196] Fig. 21 is a schematic illustrative drawing showing S301 (a life length prediction of an SSD) of Fig. 20.

[0197] A method for predicting a reaching time to a length of life from the rate of decline of a remaining erasing frequency will be described in the following. A length of life of an SSD is a period in which the SSD can be used as a practical matter. In the figure, a horizontal axis indicates a time and a vertical axis indicates a remaining erasing frequency. In the case in which a write is concentrated to an SSD, a life length is shortened. Consequently, it is necessary that an inter-device WL is executed (more specifically, data that has been stored into an extent is exchanged between SSDs) in such a manner that other SSD that is provided with a long life length is used in a positive manner. In the present processing, an SSD that is provided with a short life length (a life length is equal to or less than a threshold value) is detected.

[0198] In Fig. 21, t (previous time) 30000 is a point of time when a life length prediction was executed at a previous time and a point of time that is indicated by a value that has been stored into the information update date 13103. EZ (previous time) 30003 is a remaining erasing frequency at a point of time of a life length prediction of a previous time and a number of times that is indicated by a value that has been stored into the remaining erasing frequency 13106.

[0199] In the next place, t (present time) 30001 is a point of time when a life length prediction was executed at a present time, and EZ (present time) 30002 is a remaining erasing frequency of a present time and a value that can be acquired from the SSD

internal information that is issued this time. The following expression can be calculated by using above information:

Inclination (I) = (EZ (present time) - EZ (previous time)) divided by (t (present time) - t (previous time))

By using the above expression, an inclination (I) 3004 can be calculated. The larger the inclination (I) 3004 is, the higher the rate of decline of a remaining erasing frequency is.

[0200] The following expression can be calculated by using the inclination (I) 30004 that has been calculated and EZ (present time) 30001 that is a remaining erasing frequency of a present time:

t (life length reach) = - (EZ (present time) divided by inclination (I))

By using the above expression, a time when a remaining erasing frequency becomes zero, that is, t (life length reach) 30005 that is a life length reaching time can be calculated. A remaining life length period is calculated from a remaining erasing frequency and the rate of decline and is stored into the Disk management TBL 13100.

[0201] It is important that the storage media and the storage apparatus are used for a predetermined period (a guarantee period of 5 years for instance) and data is guaranteed. Consequently, in the case in which an inter-device WL is executed in which a guarantee period is one index, a plurality of storage media (such as SSDs) can be used for a guarantee period or longer.

[0202] In the case in which the all Write total amount of data that can be written until an SSD reaches a length of life is known, a vertical axis can be substituted with a remaining erasing frequency and a remaining Write amount (a value that is obtained by subtracting a total Write amount 13105 from the all Write total amount) can also be used.

[0203] Fig. 22 is a sequence drawing showing an example of a flow of S303 (a prediction of a write amount of a flash memory) of Fig. 20.

[0204] The processor 104 acquires a WR amount 13706 to an extent that is the statistics information that is required for the FM WR amount prediction TBL 13700 based on the information of the statistics information management TBL 13600 (S3031).

[0205] In the next place, the processor 104 makes the WR amount 13706 that has been acquired to be an FM WR predicted amount to a flash memory in an SSD to which an extent belongs (S3032).

[0206] In the next place, a summary of the S304 (an inter-device WL) of Fig. 20 will be described.

[0207] As an inter-device WL, there are some patterns, for instance, the following five execution patterns (patterns A to E):

Pattern A: data is moved between RGs in a unit of an extent (Fig. 23);

Pattern B: data is moved between RGs in a unit of an extent, and a storage location of a plurality of data elements (data (and parity)) based on data in the extent is optimized for the RG of a movement destination (Fig. 25);

Pattern C: data is moved between RGs in a specific unit of data in an extent (Fig. 27);

Pattern D: data is moved in an RG in a specific unit of data in an extent (Fig. 28);

Pattern E: data is moved between different devices in accordance with a performance characteristic of a device (Fig. 30); and

Pattern F: data is moved in accordance with a life length characteristic of a device (Fig. 32).

[0208] For the patterns A to C, the condition is that at least two RGs exist. Moreover, for the pattern D, there is at least one RG. Moreover, for the pattern E and the pattern F, the condition is that at least two RGs that are provided with different Disk types exist.

[0209] In the present embodiment, a write load is an FM WR predicted amount. In other words, a high (low) write load means that an FM WR predicted amount is large (small).

[0210] Fig. 23 is a schematic illustrative drawing showing an execution pattern A of an inter-device WL.

[0211] For the pattern A, data in an extent is moved between different RGs. An RG #n 310 is configured by an SSD 3110, an SSD 3120, and an SSD 3130, and there is an extent #3140 based on an RG #n 310. Data based on data 3111 in the SSD 3110, data 3121 in the SSD 3120, and data 3131 in the SSD 3130 has been stored into the extent #3140.

[0212] An RG #m 410 is configured by an SSD 4110, an SSD 4120, and an SSD 4130, and there is an extent #4140 based on an RG #m 410. Data based on data 4111 in the SSD 4110, data 4121 in the SSD 4120, and data 4131 in the SSD 4130 has been stored into the extent #4140.

[0213] The RG #m 410 is a RAID group in which a length of life is short. For instance, the RG #m 410 includes an SSD in which a remaining length of life is shorter than a threshold value. In other words, a RAID group in which a length of life is short includes an SSD in which a length of life is shorter than a guarantee period. An extent (A) in which an FM WR predicted amount is largest is a target of an inter-device WL among a plurality of extents that are included in the RAID group.

[0214] The RG #n 310 is a RAID group in which a length of life is long. For instance, an SSD that is included in the RAID group is provided with a remaining length of life that is larger than a threshold value. An extent (B) in which an FM WR predicted amount is largest is a target of an inter-device WL among a plurality of extents that are included in the RAID group.

[0215] Even in the case in which a length of life of any RAID group is in a guarantee period, in the case in which a divergence of a length of life between RAID groups (a di-

vergence of a length of life of an SSD that configures each RAID group) is equal to or larger than a predetermined value, the storage controller 100 can execute an inter-device WL.

[0216] As described above, in the case in which there is a divergence between a remaining length of life of the RG #n 310 and a remaining length of life of the RG #m 410, an inter-device WL is executed.

[0217] More specifically, data that has been stored into the extent # 3140 and data that has been stored into the extent # 4140 are exchanged with each other for instance. By this exchange, data that is provided with a high write load is stored into the RG #n 310 that is provided with a long remaining length of life, and data that is provided with a low write load is stored into the RG #m 410 that is provided with a short remaining length of life.

[0218] Here, the "data exchange" means that the storage controller 100 executes the following processing for instance:

(*) the storage controller 100 identifies a range of a logical address of the extent # 3140 in which data has been stored and a range of a logical address of the extent # 4140 in which data has been stored by referring to the extent management TBL 13400.

(*) the storage controller 100 issues a read request to an SSD that includes the identified logical address range (more specifically, issues a read request to an SSD that is a basis of the extent # 3140 and an SSD that is a basis of the extent # 4140), reads the data from the SSD, and stores the read data into the cache memory 103 on a temporary basis. At this time, the SSD that has received the read request identifies a physical page in which data has been stored, reads the data, and transmits the data to the storage controller based on a logical address that is included in the read request and the logical physical conversion TBL 23100. The data that has been read from the SSD based on the extent # 3140 (first data) and the data that has been read from the SSD based on the extent # 4140 (second data) are stored into the cache memory 103 on a temporary basis.

(*) the storage controller 100 identifies an SSD that includes a logical address of the extent, and issues a write request of the data that has been stored into the cache memory 103 on a temporary basis to the SSD. More specifically, a write request of the first data is issued to an SSD based on the extent # 4140 and a write request of the second data is issued to an SSD based on the extent # 3140. As a result, the first data from the extent # 3140 is stored into the extent # 4140, and the second data from the extent # 4140 is stored into the extent # 3140. At this time, the SSD that has received the write request identifies a free physical page from the logical physical conversion TBL 23100, writes the received data to the identified free physical page, and updates the logical physical conversion TBL 23100.

[0219] In the following, the "data exchange" means a processing in which data that has been stored into each of two extents is identified by the extent management TBL 13400, the identified data is stored into the cache memory 103 on a temporary basis, and the stored data is stored into the other extent that is different from an extent of a read source as described above.

[0220] In the case in which data in the extent # 3140 that is provided with a small write data amount is stored into the RG #m 410 that is provided with a short remaining length of life, a write data amount to an SSD that configures the RG #m 410 that is provided with a short remaining length of life is reduced. Consequently, a reduction of a remaining length of life can be suppressed. On the other hand, in the case in which data in the extent # 4140 that is provided with a large write data amount is stored into the RG #n 310 that is provided with a long remaining length of life, a write data amount of the RG #n 310 that is provided with a short remaining length of life is increased. Consequently, a divergence of a remaining length of life of the RG #n 310 and a remaining length of life of the RG #m 410 can be reduced by the data exchange. That is, an erasing frequency between devices can be leveled.

[0221] Since a data movement occurs in a unit of an extent in which a stripe line is maintained in the pattern A, a reduction of a redundancy of the RAID due to switching does not occur.

[0222] More specifically, the "data exchange" is equivalent to a modification of an allocation order of the allocating extent # 13505 of the virtual volume management TBL 13500 and a modification of a value of the Disk # 13405, the Stripe # 13406, the Start-LBA 13407, and the Last-LBA 13408 of the extent management TBL 13400 for instance.

[0223] In the present processing, a data movement is executed between SSDs in order to reduce a divergence of a remaining length of life of the RG #n 310 and a remaining length of life of the RG #m 410 by exchanging data that has been stored into the extent # 3140 and data that has been stored into the extent # 4140 as described above. However, data that has been stored into the extent # 4140 (a short life length and a high write load) can also be moved to an extent in which data has not been stored.

[0224] Fig. 24 is a sequence drawing showing an example of an execution pattern A of an inter-device WL.

[0225] In the sequence drawing of the following description, the storage controller 100 transmits an exchange instruction of data that has been stored into an extent to the SSD controller 710, and the SSD controller 710 that has received the exchange instruction of data transmits data related to the exchange instruction of data to the storage controller 100.

[0226] The processor 104 selects an RG that is provided with a short length of life and that

is a data movement source of an inter-device WL based on the Disk management TBL 13100 and the RG management TBL 13200. Here, the RG that is provided with a short length of life is an RG that is provided with at least one (or at least two) SSDs in which a remaining life length period is equal to or less than the predetermined threshold value (a short length of life among a plurality of SSDs that are included in an RG. For instance, the threshold value is determined based on a guarantee period. That is, an RG that includes an SSD that reaches a length of life before a guarantee period is an RG that is provided with a short length of life. In the case in which an SSD reaches a length of life before a guarantee period, it is thought that a write is concentrated to the SSD. Consequently, data that is provided with a high write load is moved from such an SSD, thereby enabling a long length of life of an SSD.

[0227] In the case in which there are RGs that are provided with a short length of life, any RG can be a target of an inter-device WL. In this case, a length of life of each RG is a length of life of an SSD that is provided with a shortest length of life among SSDs that are included in each RG, and an inter-device WL can be executed in order from an RG that is provided with a shorter length of life. Moreover, an inter-device WL can be executed in order from an RG in which the number of SSDs that are provided with a length of life shorter than a guarantee period is larger.

[0228] The processor 104 refers to the FM WR amount prediction TBL 13700, and selects an extent (A1) that is provided with a large FM WR predicted amount in RGs that are provided with a short length of life (S30410).

[0229] An extent that is provided with a large FM WR predicted amount is an extent that is provided with a largest FM WR predicted amount among a plurality of extents. The extents that are provided with a FM WR predicted amount that is equal to or larger than a threshold value can be grouped, and one extent can be selected from the group. In this case, it is not necessary that an extent that is provided with a largest FM WR predicted amount is searched, thereby shortening a processing time.

[0230] In the next place, the processor 104 selects an RG that is provided with a long length of life based on the Disk management TBL 13100 and the RG management TBL 13200. Here, the RG that is provided with a long length of life is an RG that is not provided with an SSD that is provided with a short length of life among a plurality of SSDs that are included in an RG. In the case in which there is not such an RG, an RG in which the number of SSDs that are provided with a short length of life is less can be an RG that is provided with a longer length of life.

[0231] Moreover, the processor 104 refers to the FM WR amount prediction TBL 13700, and selects an extent (A1) that is provided with a small FM WR predicted amount in RGs that are provided with a long length of life (S30411).

[0232] An extent that is provided with a small FM WR predicted amount is an extent that is

provided with a smallest FM WR predicted amount among a plurality of extents. The extents that are provided with a FM WR predicted amount that is equal to or less than a threshold value can be grouped, and one extent can be selected from the group.

[0233] The processor 104 then judges whether or not an FM WR predicted amount of the extent (A1) is larger than an FM WR predicted amount of the extent (B1) (S30412).

[0234] In the case in which an FM WR predicted amount of the extent (B1) is larger than an FM WR predicted amount of the extent (A1) and the extent (B1) is moved to an RG that is provided with a short remaining length of life, a write data amount to the RG is more increased, and an erasing frequency is not leveled between devices. Consequently, by this judgment, an erasing frequency can be leveled between devices in an appropriate manner without executing an unnecessary data movement.

[0235] In the case in which the result of the judgment is positive (S30412: Yes), the processor 104 exchanges data in the extent (A1) and data in the extent (B1) with each other (S30413) and terminates the present processing. On the other hand, in the case in which the result of the judgment is negative (S30412: No), the processor 104 stops the data exchange (S30414) and terminates the present processing.

[0236] Fig. 25 is a schematic illustrative drawing showing an execution pattern B of an inter-device WL.

[0237] For the pattern B, a data movement (a data exchange) is executed in a unit of an extent between RGs, and an optimization of a data store location in an extent is also executed. Even in the case in which data is moved between RGs, it is not always true that an optimization of a data store location in an extent is executed.

[0238] An RG #n 320 is configured by an SSD 3210, an SSD 3220, and an SSD 3230, and there is an extent # 3240 based on an RG #n 320. Data that is based on data in the SSD 3210, data in the SSD 3220, and data in the SSD 3230 has been stored into the extent # 3240.

[0239] An RG #m 420 is configured by an SSD 4210, an SSD 4220, and an SSD 4230, and there is an extent # 4240 based on an RG #m 420. Data that is based on data in the SSD 4210, data in the SSD 4220, and data in the SSD 4230 has been stored into the extent # 4240.

[0240] The extent # 3240 of the RG #n 320 is provided with the regions (3211 and 3221) that are provided with a high FM WR predicted amount and the region (3231) that is provided with a middle FM WR predicted amount. In other words, the extent # 3240 is provided with regions (extent parts) that are provided with different FM WR predicted amounts. In the case in which a logical address range based on one SSD is called an "extent", it can also be said that an extent group # 3240 is provided with extents that are provided with different FM WR predicted amounts. Moreover, the RG #n 320 is a RAID group that is provided with a short length of life.

[0241] On the other hand, the extent # 4240 of the RG #m 420 has stored data that is provided with a low FM WR predicted amount (4211, 4221, and 4231). However, a length of life of an SSD that configures the RG #m 420 is dispersed. More specifically, an SSD 4210 and an SSD 4230 are provided with a long length of life (a remaining erasing frequency is high) for instance, and an SSD 4220 is provided with a middle remaining length of life.

[0242] In other words, there is a divergence between a remaining length of life of the RG #n 320 and a remaining length of life of the RG #m 420.

In order to reduce a divergence between a remaining length of life of the RG #n 320 and a remaining length of life of the RG #m 420, a WL in which the extents # 3240 and # 4240 are targets of an inter-device WL is executed between SSDs. More specifically, the following will be executed:

(*) Data that has been stored into the extent # 3240 and data that has been stored into the extent # 4240 are exchanged with each other. More specifically, data that is provided with a large FM WR predicted amount is stored into the RG #n 320 that is provided with a long length of life, and data that is provided with a small FM WR predicted amount is stored into the RG #m 420 that is provided with a short length of life for instance. By this data exchange, a divergence between a remaining length of life of the RG #n 320 and a remaining length of life of the RG #m 420 can be reduced.

(*) In order to solve a dispersion of a remaining length of life between SSDs that configure the RG #m 420, the stored destinations of at least two data of a plurality of data based on data that has been stored into the extent # 3240 are exchanged for the extent # 3240. More specifically, the data 3231 that is provided with a middle FM WR predicted amount is stored into the SSD 4220 that is provided with a middle remaining length of life, and the data 3221 that is provided with a large FM WR predicted amount is stored into the SSD 4230 that is provided with a long length of life for instance. That is, the stored locations of the data 3231 and 3221 are exchanged with each other for the extent # 3240. By this configuration, it is expected that a dispersion of a remaining length of life between SSDs can be solved for the RG #m 420 after an inter-device WL is executed.

[0243] Fig. 26 is a sequence drawing showing an example of an execution pattern B of an inter-device WL.

[0244] The processor 104 selects an RG that is provided with a short length of life based on the Disk management TBL 13100 and the RG management TBL 13200. Moreover, the processor 104 refers to the FM WR amount prediction TBL 13700, and selects an extent (A2) that is provided with a large FM WR predicted amount in RGs that are provided with a short length of life (S30420).

[0245] In the next place, the processor 104 selects an RG that is provided with a long length

of life based on the Disk management TBL 13100 and the RG management TBL 13200. Moreover, the processor 104 refers to the FM WR amount prediction TBL 13700, and selects an extent (B2) that is provided with a small FM WR predicted amount in RGs that are provided with a long length of life (S30420).

[0246] The processor 104 then judges whether or not an FM WR amount of the extent (A2) is larger than an FM WR amount of the extent (B2) (S30422).

[0247] In the case in which the result of the judgment is positive (S30422: Yes), the processor 104 judges whether or not an optimization of a data store location in an extent is possible (S30424). The optimization of a data store location in an extent is to store the data into an SSD that is provided with a remaining length of life that is suitable for a degree of an FM WR predicted amount of the data, more specifically, to store the data that is provided with a large FM WR predicted amount into an SSD that is provided with a long remaining length of life and to store the data that is provided with a small FM WR predicted amount into an SSD that is provided with a short remaining length of life.

[0248] In the case in which the result of the judgment is positive (S30424: Yes), the processor 104 exchanges data in the extent (A2) and data in the extent (B2) with each other, executes a switching of a data location (an optimization of a data store location) in an extent (S30425), and terminates the present processing.

[0249] In the case in which the result of the judgment of the S30424 is negative (S30424: No), the processor 104 exchanges data in the extent (A2) and data in the extent (B2) with each other (S30426) and terminates the present processing without executing an optimization of a data store location.

[0250] In the case in which the result of the judgment of the S30422 is negative (S30422: No), the processor 104 stops the data movement (S30423) and terminates the present processing.

[0251] Fig. 27 is a schematic illustrative drawing showing an execution pattern C of an inter-device WL.

[0252] For the pattern C, a data movement between RGs is executed in a unit of an extent part. There are an extent (A) 3340 based on the RG #n 330 and an extent (B) 4340 based on the RG #m. An SSD 3330 belongs to the RG #n 330 and is provided with a short length of life. However, the SSD 3330 is based on a region (an extent part of the extent (A) 3340) that is provided with a large FM WR predicted amount and data has been stored into the region 3331. On the other hand, an SSD 4310 belongs to the RG #n 430 and is provided with a long length of life. However, the SSD 4310 is based on a region (an extent part of the extent (B) 4340) that is provided with a small FM WR predicted amount and data has been stored into the region 4331.

[0253] In order to level a remaining length of life of the SSD 3330 and the SSD 4310, a data

movement (a data exchange) is executed. More specifically, the data is exchanged between the region 3331 that is provided with a large FM WR predicted amount and the region 4311 that is provided with a small FM WR predicted amount.

[0254] However, there is a possibility that a stripe line cannot be maintained due to the data movement. In the case in which data or parity that belongs to the same stripe line is stored into the same SSD, a redundancy of the RAID is deteriorated.

[0255] To avoid the above problem, the processor 104 can refers to the extent management TBL 13400 in a movement of data, and can suppress the movement of data in the case in which a redundancy of the RAID is deteriorated

[0256] Before a data movement, one extent (A) 3340 (in other words, three extent parts of the same extent (A) 3340) is allocated to one virtual extent for instance. After the data movement, two extents (A) and (B) (in other words, an extent part of the extent (A) 3340 and an extent part of the extent (B) 4340) are corresponded to the virtual extent for instance. More specifically, after the data movement, an allocated destination to which the extent (A) 3340 has been allocated is modified from a part of the extent (A) 3340 (a data movement source) to a part of the extent (B) 4340 (a data movement destination). Similarly, an allocated destination of a part of a virtual extent to which the extent (B) 4340 has been allocated is modified from a part of the extent (B) 4340 (a data movement source) to a part of the extent (A) 3340 (a data movement destination).

[0257] Fig. 28 is a schematic illustrative drawing showing an execution pattern D of an inter-device WL.

[0258] For the pattern D, a data movement is executed in a unit of an extent part in an RG.

[0259] The pattern D that is different from the pattern C is not a movement between RGs but a data movement in the same RG. The basic concept of the pattern D is equal to that of the pattern C, and is to level a write load of the SSD 3410, the SSD 3420, and the SSD 3430 that configure the RG #n 340.

[0260] For instance, the RG #n 340 is configured by the SSD 3410 that is provided with a long remaining length of life, the SSD 3420 that is provided with a middle remaining length of life, and the SSD 3430 that is provided with a short remaining length of life. The extents # 3440 and # 3450 are based on the RG #n 340.

[0261] The data that has been stored into the SSD 3430 that is provided with a short length of life and that has been stored into the region 3413 that is provided with a large FM WR predicted amount is moved to the SSD 3410 that is provided with a long length of life. Moreover, the data that has been stored into the SSD 3410 that is provided with a long length of life and that has been stored into the region 3411 that is provided with a small FM WR predicted amount is moved to the SSD 3430 that is provided with a short length of life.

[0262] Similarly to the pattern C, a redundancy of the RAID is at risk of being deteriorated

due to a movement of data. In that case, the processor 104 suppresses the movement of data.

[0263] For the execution patterns C and D of an inter-device WL, data that configures the same stripe line is not stored into the same extent. Consequently, the processor 104 can store a correspondence relationship between an extent part of a movement source of the data and an extent part of a movement destination into the cache memory 103. In the case in which the processor 104 reads a data group that configures the same stripe line and that is dispersed to a plurality of extents, a movement source can be a plurality of extents based on the correspondence relationship that has been stored. Alternatively, in the case in which data that configures the same stripe line is dispersed to a plurality of extents, the processor 104 can modify the correspondence relationship between a part of a logical address range of an extent and a physical address range of an SSD. By this configuration, some extent can based on an SSD of a part of one RAID group and an SSD of a part of another RAID group.

[0264] Fig. 29 is a sequence drawing showing an example of the execution patterns C and D of an inter-device WL.

[0265] The processor 104 selects an RG that is formed based on the Disk that is provided with a short length of life based on the Disk management TBL 13100 and the RG management TBL 13200. Moreover, the processor 104 refers to the FM WR amount prediction TBL 13700 and selects an extent (A3) that is provided with a large FM WR predicted amount in Disks that are provided with a short length of life (S30430).

[0266] In the next place, the processor 104 selects a Disk that is provided with a long length of life based on the Disk management TBL 13100 and the FM WR amount prediction TBL 13700. Here, the processor 104 selects a Disk that is provided with a long length of life among Disks that configure an RG that is different from an RG that is provided with a Disk that is provided with a short length of life in the pattern C. Moreover, the processor 104 selects a Disk that is provided with a long length of life among Disks that configure an RG that is equal to an RG that is provided with a Disk that is provided with a short length of life in the pattern D.

[0267] In the next place, the processor 104 selects the data (B3) that is provided with a small FM WR amount among Disks that are provided with a long length of life (S30431). The processor 104 then judges whether or not an FM WR amount of the data (A3) is larger than an FM WR amount of the data (B3) (S30432).

[0268] In the case in which the result of the judgment is positive (S30432: Yes), the processor 104 refers to the extent management TBL 13400 and judges whether or not a redundancy of the RAID is deteriorated due to a data movement (S30433).

[0269] In the case in which the result of the judgment is negative (S30433: No), the processor 104 executes a data movement for exchanging a store location of the data

(A) and the data (B) with each other (S30433) and terminates the present processing.

[0270] On the other hand, in the case in which the result of the judgment is positive (S30433: Yes), the processor 104 stops a movement of data (S30435) and terminates the present processing.

[0271] Moreover, in the case in which the result of the judgment is negative (S30432: No), the processor 104 stops a movement of data (S30435) and terminates the present processing.

[0272] Fig. 30 is a schematic illustrative drawing showing an execution pattern E of an inter-device WL.

[0273] For the pattern E, a data movement is executed in accordance with a characteristic of a device that affects a remaining length of life of the storage media.

[0274] The RG #n 350 is configured by an SSD of an SLC type. The RG #m 450 is configured by an SSD of an MLC type. The RG #o 550 is configured by an HDD. In other words, the RG #n 350 is characterized by a high speed and a long life length. The RG #m 450 is characterized by a high speed and a short life length. The RG #o 550 is characterized by a low speed and no restriction of rewriting in a substantial way.

[0275] For the pattern E, a device that is most suitable for a stored destination of data is selected by using a difference in a characteristic for every device.

[0276] More specifically, the data that has been stored into an extent that is provided with a small FM WR predicted amount among is stored into an extent that is provided with a small FM WR predicted amount of the SSD of an SLC type, the data that is provided with a high read load is stored into an extent that is provided with a small FM WR predicted amount of the SSD of an MLC type, and the data that is provided with a low read load is moved to an HDD.

[0277] As described above, in the case in which a data movement is executed between devices that are provided with different characteristic, each of data can be stored into a device of a type that is most suitable for the I/O characteristic.

[0278] To simplify the descriptions, Fig. 30 shows an example of the pattern A (a data movement between RGs). However, the pattern E can also be applied to both of the pattern B (a data movement between RGs and a data movement between SSDs that are basis of the same extent) and the pattern C (a data movement between RGs in a specific data unit).

[0279] Fig. 31 is a sequence drawing showing an example of an execution pattern E of an inter-device WL.

[0280] The processor 104 selects a specific extent that is a processing target based on the Disk management TBL 13100, the RG management TBL 13200, the extent management TBL 13400, and the FM WR amount prediction TBL 13700 (S30440). In the description of Fig. 31 in the following, an extent that has been selected in the

S30440 is referred to as a "target extent".

[0281] In the next place, the processor 104 judges whether or not a target extent includes a region that is provided with a large FM WR predicted amount based on the Disk management TBL 13100, the RG management TBL 13200, the extent management TBL 13400, and the FM WR amount prediction TBL 13700 (S30441). In the case in which the result of the judgment is positive (S30441: Yes), the processor 104 moves data in the target extent to an extent based on an SSD of an SLC type (S30443) and terminates the present processing.

[0282] On the other hand, in the case in which the result of the judgment of the S30441 is negative (S30441: No), the processor 104 judges whether or not a read load of a target extent is high (S30442).

[0283] In the case in which the result of the judgment is positive (S30442: Yes), the processor 104 moves data in the target extent to an extent based on an SSD of an MLC type (S30444) and terminates the present processing.

[0284] On the other hand, in the case in which the result of the judgment of the S30442 is negative (S30442: No), the processor 104 moves data in the target extent to an extent based on an HDD (S30445) and terminates the present processing.

[0285] Fig. 32 is a sequence drawing showing an example of an execution pattern F of an inter-device WL.

[0286] For the pattern F, the data is moved in accordance with a life length of a device. The processor 104 selects a specific extent that is a processing target based on the Disk management TBL 13100, the RG management TBL 13200, the extent management TBL 13400, and the FM WR amount prediction TBL 13700 (S30450).

[0287] An extent that is selected here is an extent that is provided with a large FM WR predicted amount. In the description of Fig. 32 in the following, an extent that has been selected in the S30450 is referred to as a "target extent".

[0288] In the next place, the processor 104 judges whether or not data in the target extent can be stored into an extent based on an SSD of an MLC type from a point of view of a length of life (S30451). More specifically, the processor 104 refers to the Disk management TBL 13100 and checks whether or not a remaining life length period 13107 of an SSD is shorter than a remaining guarantee period 13104 of the SSD for each of the SSD of an MLC type that configures the same RG. This is because in the case in which a remaining life length period 13107 of an SSD is shorter than a remaining guarantee period 13104 of the SSD for each of the SSD of an MLC type that configures the same RG, a high write load that exceeds an acceptable amount occurs for the SSD of an MLC type, and a problem of a length of life cannot be solved even in the case in which a movement destination of data is an extent based on the SSD of an MLC type.

[0289] In the case in which a remaining life length period 13107 of all SSDs of an MLC type is shorter than a remaining guarantee period 13104 of the SSD (S30451: No), the processor 104 tries to move data to an extent based on the SSD of an SLC type that is provided with a length of life longer than that of the SSD of an MLC type (S30452). This is possible in the case in which the processor 104 refers to the Disk management TBL 13100 and judges whether or not a remaining life length period 13107 of an SSD is shorter than a remaining guarantee period 13104 of the SSD for the SSD of an SLC type that configures the same RG.

[0290] In the case in which the SSD of an SLC type is not suitable as a movement destination (S30452: No), the processor 104 decides that an HDD that is not provided with an upper limit of a rewriting is a movement destination of data (S30455).

[0291] In the case in which the SSD of an MLC type has the capacity to a length of life as a result of the judgment of the step S30451 (S30451: Yes), the processor 104 decides that the SSD of an MLC type is a movement destination of data (S30453).

[0292] In the case in which the SSD of an SLC type has the capacity to a length of life as a result of the judgment of the step S30452 (S30452: Yes), the processor 104 decides that the SSD of an SLC type is a movement destination of data (S30454).

## Embodiment 2

[0293] An embodiment 2 will be described in the next place. The present embodiment includes many of common parts with the embodiment 1. Consequently, in the present embodiment, a part that is different from the embodiment 1 will be described mainly. In the case in which a WL target is selected in the embodiment 2, a predicted WA (see Fig. 36) is used in addition to the FM WR predicted amount.

[0294] Since a unique processing of the SSD such as a WL and a reclamation processing occurs in the SSD 700 in general, there is a characteristic in which an amount of data that is written to a flash memory as a practical matter is larger than an amount of data that has been received from the storage controller 100 by the SSD 700. This is called a WA (Write Amplification). An increase in a write data amount due to a processing in an SSD depends on an access pattern and a size of write data or the like.

[0295] In the case in which a write data amount is increased by a unique processing in the SSD, an erasing frequency is also increased in accordance with the step. In the present embodiment consequently, an accuracy of an inter-device WL is improved by predicting a WA.

[0296] The WA will be described in detail in the following in the first place.

[0297] The WA is a rate that is obtained by dividing the following (b) by (a) ((b) / (a)):

(a) an amount of data that has been received from the storage controller 100 by the SSD; and

(b) an amount of data that is written to a flash memory as a practical matter.

[0298] For instance, in the case in which the SSD 700 receives WR data from the storage controller 100 in the state in which no data has been written to the SSD 700, the data is written to a free page without any change, whereby the predicted WA have a high probability of being "1.0". Moreover for instance, in the case in which the effective data has been written to a page of the SSD 700 and a reclamation processing is required without a free page, a page that is a movement destination of the effective data and a page in which data from the storage controller 100 is to be written are required, whereby the predicted WA exceeds "1.0".

[0299] However, it is not always true that the WA 13707 is small even in the case in which the WR amount 13706 is small. This is because a unique processing of the SSD such as a WL and a reclamation processing occurs in the SSD 700.

[0300] The reclamation processing is a processing in which a free block is generated by collecting pages that store the effective data to write to another block, by generating a block that is provided with only the ineffective data, and by erasing data in the block in the case in which a free block is started to be depleted for instance.

[0301] Moreover in general, the WA has a characteristic in which a random I/O (also referred to as a random access) is larger than a sequential I/O (also referred to as a sequential access).

[0302] The sequential I/O is an I/O to a continuous LBA space in general. Consequently, there is a high possibility that new data is written to all pages that configure one block. Therefore, all data that exist in the block are not ineffective data in some case and a free page can be formed only by executing an erasing processing to the block. Accordingly, since there is a low necessity of moving data for the sequential I/O, a page that is a movement destination of data is not consumed and there is a high possibility that the WA is "1" or a numerical value that is close to "1".

[0303] The random I/O is an I/O to a discontinuous LBA in general and a plurality of blocks is I/O destinations in some cases for instance. In this case, there is a high possibility that much effective data is included in one block. Consequently, a data amount that is moved in a reclamation processing is larger as compared with a sequential write. Therefore, there is a high possibility that the WA is a numerical value that is larger than "1".

[0304] Moreover in general, the WA is larger for a small size I/O as compared with a large size I/O. Here, a small size I/O means that a size of data that is associated with an I/O command is small. A large size I/O means that a size of data that is associated with an I/O command is large. In the case in which data of 512 B is transmitted to an SSD as a write target for instance and the minimum write unit (page) is 8192 B, one page of 8192 B is consumed and data of 512 B is stored. This is equal to that a write of data of 16 times as compared with a size of data to an SSD is executed to an internal flash

memory. This depends on an I/O size and a page size.

[0305] Moreover, a value of a WA also depends on an existence or non-existence of a compression function of an SSD, an existence or non-existence of a duplication exclusion function of an SSD, and a type of a compression algorithm of a compression function. This is because data in which a compression effect or a duplication exclusion effect is high is provided with a small write data size.

[0306] Moreover, a value of a WA also depends on a cache hit rate in an SSD. This is because a write to a flash memory does not occur in the case in which write data is updated on a cache in an SSD.

[0307] Moreover, a value of a WA also depends on a data storage rate of an SSD (a rate of the total amount of user data to a capacity of an SSD).

[0308] In the case in which user data of 50 GB has been stored into an SSD that is provided with a physical capacity of 100 GB for instance, the SSD can utilize a remaining region of 50 GB as a free page. Even in the case in which all of the user data of 50 GB is updated to new data, the new data can be written to a remaining free page of 50 GB. At this time, since the updated data is invalid data all, a free page can be formed by only executing an erasing processing in the reclamation processing. Consequently, in the case in which a data storage rate of an SSD is small (more specifically, a half or less), an efficiency of the reclamation is improved and there is a high possibility that the WA comes close to "1".

[0309] As described above, a WA is affected by a wide variety of factors. In the present embodiment, the storage controller predicts a WA in consideration of these (see Fig. 34).

[0310] Fig. 33 is a view showing a configuration example of an FM WR amount prediction TBL 13700 in accordance with the present embodiment.

[0311] The FM WR amount prediction TBL 13700 includes the information that is required for predicting a data amount that is written in an SSD as a practical matter.

[0312] The FM WR amount prediction TBL 13700 is configured based on the statistics information management TBL 13600. The FM WR amount prediction TBL 13700 is provided with a Disk# 13701, an extent# 13702, an average WR I/O Size 13703, a WR I/O Pattern 13704, a WR rate 13705, a WR amount 13706, a predicted WA (Write Amplification) 13707, and an FM WR predicted amount 13708 for every Disk.

[0313] The Disk# 13701 is an identifier of the Disk and a unique number. The extent# 13702 is an identifier of an extent and a unique number.

[0314] The average WR I/O Size 13703 is an average size of data in which an address range that is corresponded to an extent in an address range of the Disk is a WR I/O destination. The WR I/O Pattern 13704 indicates whether a pattern of a WR I/O is a random pattern or a sequential pattern. The WR rate 13705 indicates a rate of a WR command of an I/O to an address range that is corresponded to an extent in an address

range of the Disk. The WR amount 13706 indicates a total amount of WR data in which an address range that is corresponded to an extent in an address range of the Disk is a write destination.

[0315] The predicted WA 13707 is a numerical value that predicts a multiple number of an increase of WR data from the storage controller 100 in the SSD 700. The predicted WA is based on a WA information storage TBL 13800 described later.

[0316] The FM WR predicted amount 13708 indicates a predicted amount of data that is written to a flash memory as a practical matter. The value is a numerical value that is obtained based on the WR amount 13706 and the predicted WA 13707 (a product of the WR amount 13706 and the predicted WA 13707).

[0317] The present table 13700 is a table for comprehending a data amount in a unit of an extent in the SSD 700 for a write data amount that has been transmitted from the storage controller 100 to the SSD 700.

[0318] The Disk# 13701 to the WR amount 13706 are updated based on the statistics information management TBL 13600.

[0319] Fig. 34 is a view showing a configuration example of a WA information storage table 13800 in accordance with a second embodiment.

[0320] The WA information storage table 13800 is provided with a WR I/O pattern 13801, an average WR I/O size 13802, and a predicted WA 13803. It is thought that the WA is also affected by factors other than a WR I/O pattern and an average WR I/O size. However, the present embodiment is based on a concept that an influence of a WR I/O pattern and an average WR I/O size to the WA is large. As substitute for or in addition to an average WR I/O size, the maximum (or minimum) WR I/O size in a unit time can also be adopted. Moreover, the predicted WA 13803 can also be determined based on not only the WR I/O pattern 13801 and the average WR I/O size 13802 but also any items (a WR IOPS 13603, an RD IOPS 13604, a WR rate 13607, a WR amount 13608, and an RD amount 13609) of the statistics information management TBL 13600. Moreover, the information that is included in the statistics information (see Fig. 9) is the information that can be easily acquired by monitoring an I/O by the storage controller 100 (the statistics information may be installed in advance as a function of the storage controller 100). Consequently, an WA can be predicted without executing a communication overhead between the storage controller and an SSD.

[0321] The WR I/O pattern 13801 indicates an I/O pattern of a write. The average WR I/O size 13802 indicates an average value of a size of write data. The predicted WA 13803 indicates a value of a WA that is predicted to the I/O pattern and the average I/O size.

[0322] In the present embodiment, in the case in which an I/O pattern is sequential, a predicted WA is 1.0 regardless of an average I/O size. In the case in which an I/O pattern is random, a predicted WA is larger as an I/O size is smaller.

[0323] In the case in which an SSD is provided with a compression function or a duplication exclusion function, the compression function and the duplication exclusion function can be added to the WA information storage table 13800. The data of a specific pattern (such as all zero data) is provided with a high compression effect, and a write amount in an SSD is small. Consequently, in the case in which the storage controller 100 has a high tendency to write a data pattern in which a compression effect or a duplication exclusion effect is high to an SSD, the predicted WA is made small.

[0324] Moreover, in the case in which the storage controller 100 acquires a value of a cache hit rate in an SSD from the SSD and a cache hit rate is high, the predicted WA can be made small.

[0325] Moreover, in the case in which the storage controller 100 acquires a data storage rate in an SSD from the SSD and a data storage rate is small, the predicted WA can be made small.

[0326] As described above, in the case in which the storage controller 100 acquires the internal information from the SSD and predicts a WA, an accuracy of the predicted WA can be more improved.

[0327] Since a basic flow of an inter-device WL is equal to that of the embodiment 1, the descriptions are omitted. The processing of the S303 shown in Fig. 20 is different as described in the following.

[0328] Fig. 35 is a sequence drawing showing an example of a flow of S303 (a prediction of a write amount of a flash memory) of Fig. 20.

[0329] The processor 104 acquires a WR amount to an extent from the statistics information management TBL 13600 (S3033).

[0330] The processor 104 acquires an average WR I/O Size 13703, a WR I/O Pattern 13704, and a WR amount 13706 that are the statistics information that is required for the FM WR amount prediction TBL 13700 based on the information of the statistics information management TBL 13600 (S3034).

[0331] In the next place, the processor 104 acquires a value of a predicted WR13803 from the WA information storage TBL 13800 based on the information that has been acquired in the S3034 (S3031).

[0332] In the next place, the processor 104 calculates the FM WR predicted amount 13708 based on the predicted WA 13803 and the WR amount 13706 to an extent. Here, the FM WR predicted amount 13708 is calculated by a product of the predicted WA 13803 and the WR amount 13706 to an extent.

[0333] In the present embodiment, the FM WR predicted amount 13708 is obtained not only by the WR amount 13706 but also based on the WR amount 13706 to an extent and the predicted WA 13707. Consequently, an inter-device WL with a higher degree of precision can be executed, thereby lengthening a life length of an SSD.

## Embodiment 3

[0334] An embodiment 3 will be described in the next place. The present embodiment includes many of common parts with the embodiments 1 and 2. Consequently, in the present embodiment, a part that is different from the embodiments 1 and 2 will be described mainly.

[0335] In the embodiment 3, a granularity of the information that is acquired by the storage controller 100 is expanded in a Chunk unit. In this case, the statistics information 25200 shown in Fig. 17 can include the information 23210 in an SSD unit and the information 23220 in a Chunk (an aggregate of a plurality of extents) unit. Moreover, the address information 25100 (see Fig. 17) includes the information in a unit of an SSD and the logical physical conversion information in a Chunk unit.

[0336] Consequently, the processor 104 can comprehend a Chunk to which an extent belongs. In other words, the information that is acquired by the storage controller 100 for the present embodiment is more detailed as compared with that of the embodiments 1 and 2. Consequently, a predicted accuracy of an FM WR amount is improved, thereby improving an accuracy of an inter-device WL. A specific point that has been modified from the embodiment 1 will be described with reference to Fig. 36.

[0337] Fig. 36 is a view showing a configuration example of an FM WR amount prediction TBL 13700 in accordance with the present embodiment.

[0338] A difference from the embodiment 1 (Fig. 10) is that the FM WR amount prediction TBL 13700 includes the information of the Chunk# 13710 and the Chunk WR amount 13711. The Chunk# 13710 indicates the number of a Chunk to which an extent belongs. The Chunk WR amount 13711 indicates the total value of an FM WR amount that is corresponded to a Chunk (an amount of data that has been written to a flash memory region that is corresponded to a Chunk). In the case in which the information of the Chunk WR amount 13711 in which a notification granularity is higher than that of the information in a unit of an SSD is referred to, the present embodiment enables the FM WR predicted amount 13708 to be calculated with a higher degree of accuracy as compared with the embodiment 1.

[0339] By the above method, even in the case in which it is difficult that the internal information of an SSD is comprehended in detail, an inter-device WL can be implemented with a high degree of accuracy in the case in which the storage controller 100 predicts an internal status of an SSD (for instance, the FM WR predicted amount 13708) based on the statistics information 25200.

[0340] For an execution of an inter-device WL, as the information for making a decision for predicting a data write amount in the SSD by the storage controller 100, it is also possible to use a write I/O frequency and/or a data write amount for every extent.

[0341] In the embodiment 1 to the embodiment 3, the storage controller 100 executes a

processing for converting a remaining erasing frequency to the number of remaining dates. However, an SSD can also directly notify of the number of remaining dates. This can be implemented in the case in which an SSD is provided with a function for converting a remaining erasing frequency to the number of remaining dates as shown in Fig. 21 for instance.

[0342] In the embodiment 1 to the embodiment 3, an erasing frequency or an FM WR predicted amount is used for calculating a remaining length of life. However, a flash memory is provided with a characteristic in which as a time from a write (programming) of data to an erasing is shorter, the flash memory is deteriorated more easily. Consequently, in the case in which a remaining length of life is calculated, not only an erasing frequency and an FM WR predicted amount but also an index "degree of deterioration" in consideration of a time from a write to an erasing can also be used.

[0343] As a degree of deterioration, an accumulation of values (points) that are corresponded to a time elapsed from a previous write for every erasing of a block can be used for instance. More specifically, a degree of deterioration can be managed by adding points that are corresponded to an elapsed time for every erasing as 40 points in the case in which data is erased within 5 minutes from a time when data was written and 35 points in the case in which data is erased within 10 minutes from a time when data was written for instance.

[0344] In the case in which the maximum value (upper limit) of a degree of deterioration is determined in advance, a length of life can be predicted by a rate of increase in a degree of deterioration. By this step, a length of life can be predicted with a higher degree of accuracy as compared with the case in which only an FM WR predicted amount and an erasing frequency are used, thereby executing an inter-device WL with a high degree of accuracy.

[0345] In the embodiment 1 to the embodiment 3, a processing for executing an inter-device WL in a unit of an extent was described. However, an inter-device WL can also be executed in a unit of an LU (Logical unit). In this case, the storage controller 100 acquires the statistics information in a unit of an LU. For other processing, a RAID group of a movement source (a RAID group that is provided with a short length of life) and a RAID group of a movement destination (a RAID group that is provided with a long length of life) are selected based on a length of life of a RAID group similarly to the above embodiments, and data that has been stored into an LU is moved based on a predicted write amount (a write load) of the RAID group of a movement source and the RAID group of a movement destination that have been selected.

## Reference Signs List

[0346] 10: Storage apparatus

# Claims

[Claim 1] A storage apparatus, comprising:
a plurality of nonvolatile semiconductor storage unit that is provided with a memory controller; and
a storage controller that is a controller that is coupled to the plurality of semiconductor storage unit,
wherein each of the semiconductor storage unit is configured by at least one nonvolatile semiconductor storage media and is a basis of a logical storage region,
the storage controller writes data to a semiconductor storage unit that is a basis of a logical storage region of a write destination of a plurality of logical storage regions,
the storage controller acquires the internal information from each of the semiconductor storage media on a regular basis or on an irregular basis for instance and stores the internal information that has been acquired for every semiconductor storage medium,
the storage controller stores the statistics information that indicates the statistics that is related to a write for every logical storage region and stores the remaining life length information that is the information that is related to a remaining length of life of each of the semiconductor storage media for every semiconductor storage medium,
(A) the storage controller identifies a first semiconductor storage unit and a second semiconductor storage unit that is provided with a remaining length of life shorter than that of the first semiconductor storage unit based on the remaining life length information that has been acquired,
(B) the storage controller moreover identifies a first logical storage region for the first semiconductor storage unit and a second logical storage region that is provided with a write load higher than that of the first logical storage region for the second semiconductor storage unit based on the statistics information that indicates the statistics that is related to a write for every logical storage region, and
(C) the storage controller reads data from the first logical storage region group and the second logical storage region, writes data that has been read from the first logical storage region to the second logical storage region, and/or writes data that has been read from the second logical storage region to the first logical storage region.

[Claim 2] A storage apparatus according to claim 1, wherein:
the internal information for every semiconductor storage medium includes a numerical value that is related to a remaining length of life of the semiconductor storage medium as the remaining life length information and is the information in a unit larger than the minimum unit of a storage region that is included in the semiconductor storage medium, and
the storage controller predicts a remaining length of life of the semiconductor storage medium based on the numerical value in the internal information that has been acquired at a first point of time and the numerical value in the internal information that has been acquired at a second point of time before the first point of time for each of the semiconductor storage media.

[Claim 3] A storage apparatus according to claim 2, wherein:
a write load for a logical storage region based on the semiconductor storage media is a load that conforms to a write amount that is a total amount of data that has been transmitted to the logical storage region as a write destination.

[Claim 4] A storage apparatus according to claim 3, wherein:
a write load for the logical storage region is a load that is based on the write amount and a predicted write increase-decrease rate that is obtained based on the statistics information by the storage controller, and
the predicted write increase-decrease rate for the logical storage region is obtained based on the statistics that is related to a write to the logical storage region.

[Claim 5] A storage apparatus according to claim 4, wherein:
the statistics that is related to a write to the logical storage region are an average size of data that conforms to a write to the logical storage region and/or whether a write destination of a write to the logical storage region is sequential or random.

[Claim 6] A storage apparatus according to claim 2, wherein:
a plurality of RAID groups is formed by the plurality of semiconductor storage media,
the semiconductor storage unit is a RAID group,
at least two logical storage region groups are formed for every RAID group,
the logical storage region group is an aggregate of at least two logical

storage regions that are corresponded to at least two semiconductor storage media that configure the RAID group,
in the (A), a first RAID group and a second RAID group that is provided with a remaining length of life that is shorter than that of the first RAID group are identified based on a remaining length of life of each semiconductor storage medium,
the first RAID group and the second RAID group are the first semiconductor storage unit and the second semiconductor storage unit, and
in the (B), the first logical storage region group is a logical storage region group based on the first RAID group and the second logical storage region group is a logical storage region group based on the second RAID group.

[Claim 7] A storage apparatus according to claim 6, wherein:
the storage controller executes the following (G) in the case in which the condition of the following (F) is satisfied:
(F) for at least one of the first RAID group and the second RAID group, at least two semiconductor storage media based on the logical storage region group to which data has been written in the (C) varies in a remaining length of life and at least two logical storage regions that configure the logical storage region group varies in a write load; and
(G) data is exchanged between a first logical storage region and a second logical storage region for the same logical storage region group based on a write load of the at least two logical storage regions and a remaining length of life of at least two semiconductor storage media based on the logical storage region to which data has been written in the (C),
the first logical storage region is based on the first semiconductor storage medium of at least two semiconductor storage media,
the second logical storage region is based on the second semiconductor storage medium of at least two semiconductor storage media,
the second semiconductor storage medium is a semiconductor storage medium that is provided with a remaining length of life that is shorter than that of the first semiconductor storage medium, and
the second logical storage region is a logical storage region that is provided with a write load that is higher than that of the first logical storage region.

[Claim 8] A storage apparatus according to claim 7, wherein:
each logical storage region is configured by a plurality of stripe lines,

and

in the case in which the (G) is executed and at least two data that are stored into the same stripe line are stored into the same semiconductor storage medium, the (G) is not executed even if the condition of the (F) is satisfied.

[Claim 9] A storage apparatus according to claim 1, wherein:

a plurality of RAID groups is formed by the plurality of semiconductor storage media,

at least two logical storage region groups are formed for every RAID group,

in the (A), a first semiconductor storage medium in a first RAID group and a second semiconductor storage medium in a second RAID group that is provided with a remaining length of life that is shorter than that of the first RAID group are identified based on a remaining length of life of each semiconductor storage medium,

the first semiconductor storage medium and the second semiconductor storage medium are the first semiconductor storage unit and the second semiconductor storage unit, and

in the (B), the first logical storage region is a logical storage region based on the first semiconductor storage medium and the second logical storage region is a logical storage region based on the second semiconductor storage medium.

[Claim 10] A storage apparatus according to claim 9, wherein:

the storage controller provides a virtual volume that is configured by a plurality of virtual regions, allocates a logical storage region group to a virtual region of a write destination, and writes data of a write target to the logical storage region group, and

in the (C), the second logical storage region is allocated to a first virtual region to which a logical storage region group based on the first RAID group has been allocated as substitute for the first storage region of the logical storage region group, and/or the first logical storage region is allocated to a second virtual region to which a logical storage region group based on the second RAID group has been allocated as substitute for the second storage region of the logical storage region group.

[Claim 11] A storage apparatus according to claim 2, wherein:

a RAID groups is formed by the plurality of semiconductor storage media,

at least two logical storage regions are formed based on the RAID

group,
in the (A), a first semiconductor storage medium and a second semiconductor storage medium that are included in a first RAID group are identified based on a remaining length of life of each semiconductor storage medium,
the first semiconductor storage medium and the second semiconductor storage medium are the first semiconductor storage unit and the second semiconductor storage unit, and
in the (B), the first logical storage region is a logical storage region based on the first semiconductor storage medium and the second logical storage region is a logical storage region based on the second semiconductor storage medium.

[Claim 12] A storage apparatus according to claim 11, wherein:
each logical storage region group is configured by a plurality of stripe lines, and
in the case in which at least two data that are stored into the same stripe line are stored into the same semiconductor storage medium, the (C) is not executed.

[Claim 13] A storage apparatus according to claim 1, wherein:
the logical storage region of a write destination of data that has been read in the (C) is decided based on a read load of the logical storage region in addition to the write load for the logical storage region, and a read load that conforms to an amount of data that is read in a unit time.

[Claim 14] A storage apparatus according to claim 2, wherein:
the storage controller stores the information that indicates a period of guarantee of the semiconductor storage media, and
the first semiconductor storage unit and the second semiconductor storage unit
the characteristic is a remaining length of life and a period of guarantee are decided based on a result of a comparison of a remaining length of life of each semiconductor storage unit and a period of guarantee of each semiconductor storage unit.

[Claim 15] A storage apparatus according to claim 2, wherein:
the numerical value is a numerical value that is related to at least one of a remaining frequency of erasing and a real write amount, and
the real write amount of a semiconductor storage unit is a total amount of data that has been written in the semiconductor storage unit as a

practical matter.

[Claim 16] A storage apparatus according to claim 2, wherein:
a unit of the predicted remaining length of life is a day.

[Claim 17] A storage apparatus according to claim 2, wherein:
(A) the storage controller identifies a semiconductor storage unit in which a remaining length of life is equal to or less than a threshold value based on the remaining life length information that has been acquired,
(B) the storage controller identifies a first logical storage region for the identified semiconductor storage unit and a second logical storage region that is provided with a write load higher than that of the first logical storage region based on the statistics information that is related to a write for every logical storage region, and
(C) the storage controller reads data from the first logical storage region group and the second logical storage region, writes data that has been read from the first logical storage region to the second logical storage region, and/or writes data that has been read from the second logical storage region to the first logical storage region.

[Claim 18] A storage control method comprising the steps of:
identifying a first semiconductor storage unit that is a semiconductor storage medium and a second semiconductor storage unit that is at least one semiconductor storage medium and that is provided with a remaining length of life shorter than that of the first semiconductor storage unit based on the remaining life length information that is the information that is related to a remaining length of life of each non-volatile semiconductor storage medium,
identifying a first logical storage region for the first semiconductor storage unit and a second logical storage region that is provided with a write load higher than that of the first logical storage region for the second semiconductor storage medium based on the statistics information that indicates the statistics that is related to a write of a plurality of semiconductor storage regions based on a plurality of semiconductor storage units, and
reading data from the first logical storage region group and the second logical storage region, writing data that has been read from the first logical storage region to the second logical storage region, and/or writing data that has been read from the second logical storage region to the first logical storage region.

[Fig. 1]

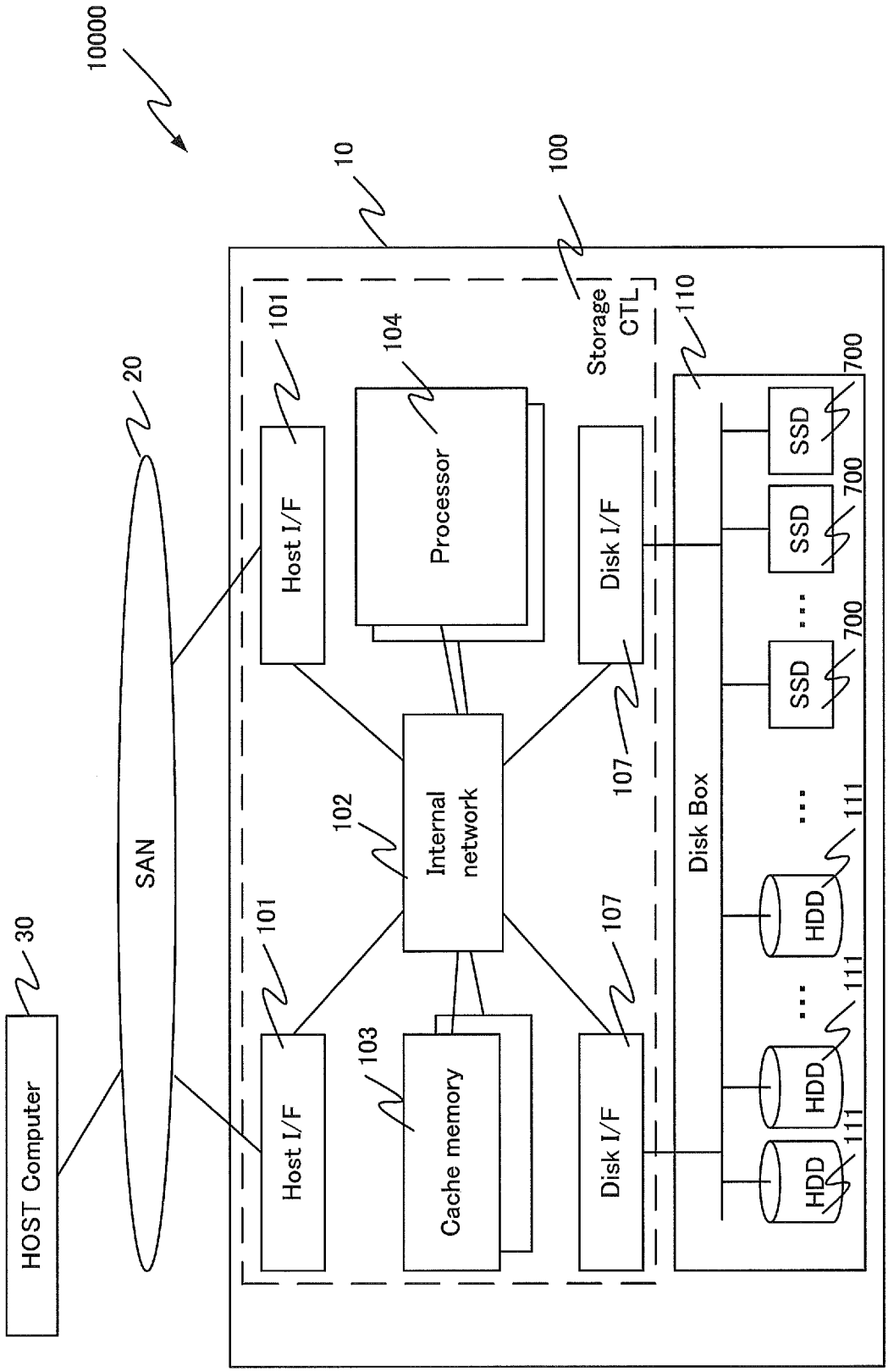
10000
10
100
20
30
SAN
HOST Computer
Storage CTL
Host I/F
Processor
Disk I/F
Internal network
Cache memory
Disk Box
SSD
HDD
101
102
103
104
107
110
111
700

Fig. 1

[Fig. 2]

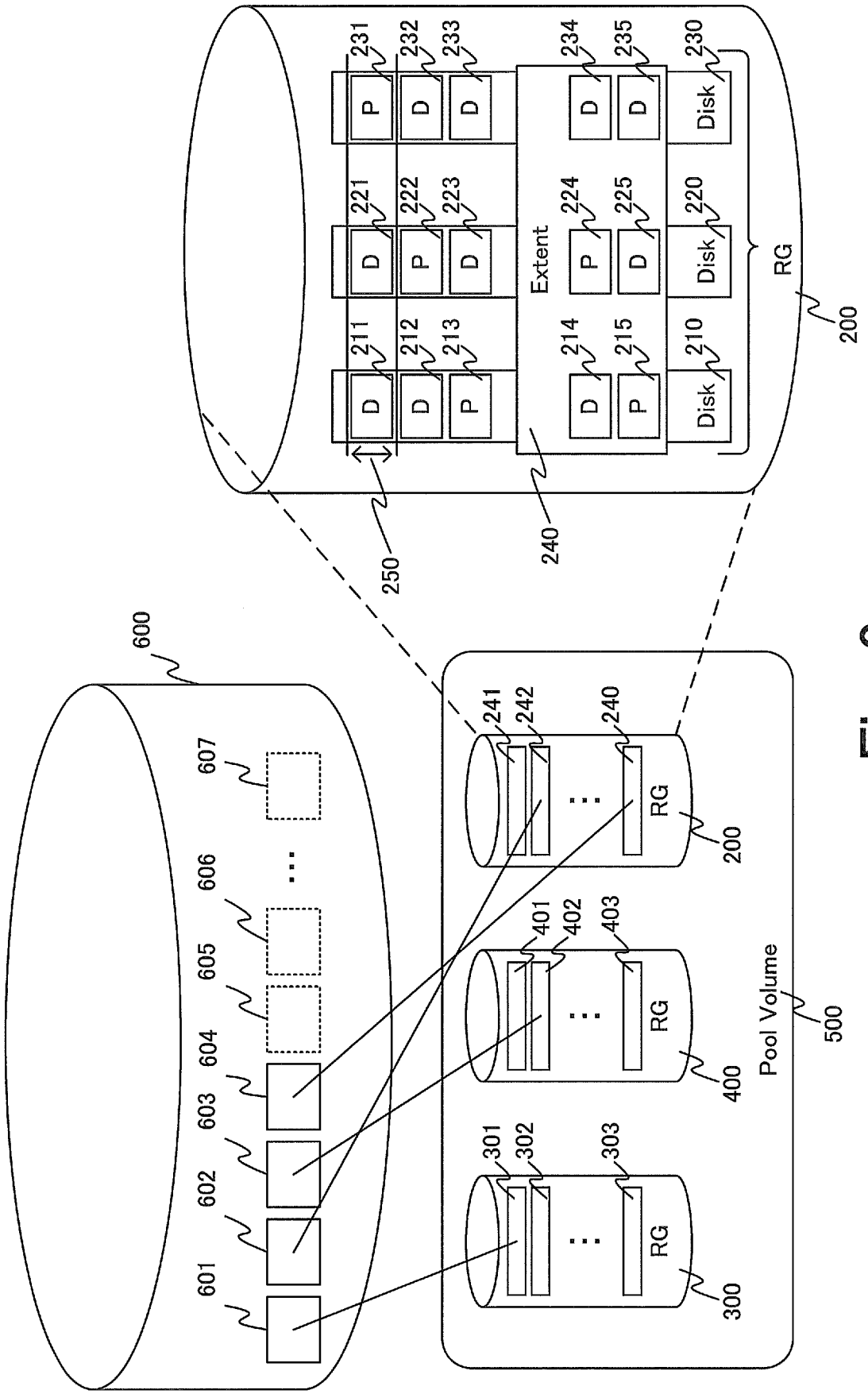
600
601
602
603
604
605
606
607
Pool Volume
500
RG
200
300
400
Extent
240
250
231
232
233
234
235
230
221
222
223
224
225
220
211
212
213
214
215
210
P
D
Disk

Fig. 2

[Fig. 3]

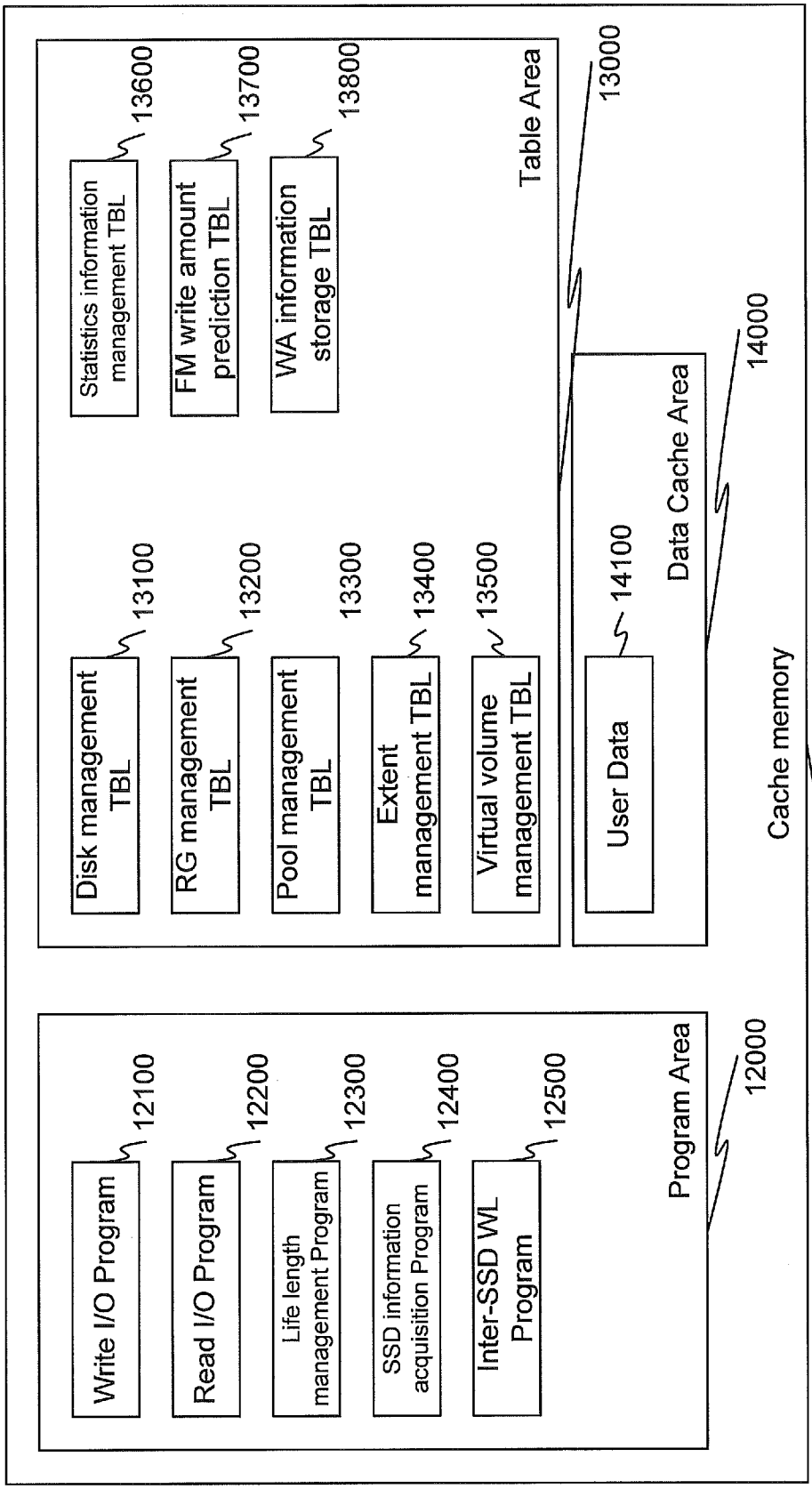
Cache memory
103
Program Area
12000
Write I/O Program
12100
Read I/O Program
12200
Life length management Program
12300
SSD information acquisition Program
12400
Inter-SSD WL Program
12500
Table Area
13000
Disk management TBL
13100
RG management TBL
13200
Pool management TBL
13300
Extent management TBL
13400
Virtual volume management TBL
13500
Statistics information management TBL
13600
FM write amount prediction TBL
13700
WA information storage TBL
13800
Data Cache Area
14000
User Data
14100

Fig. 3

[Fig. 4]

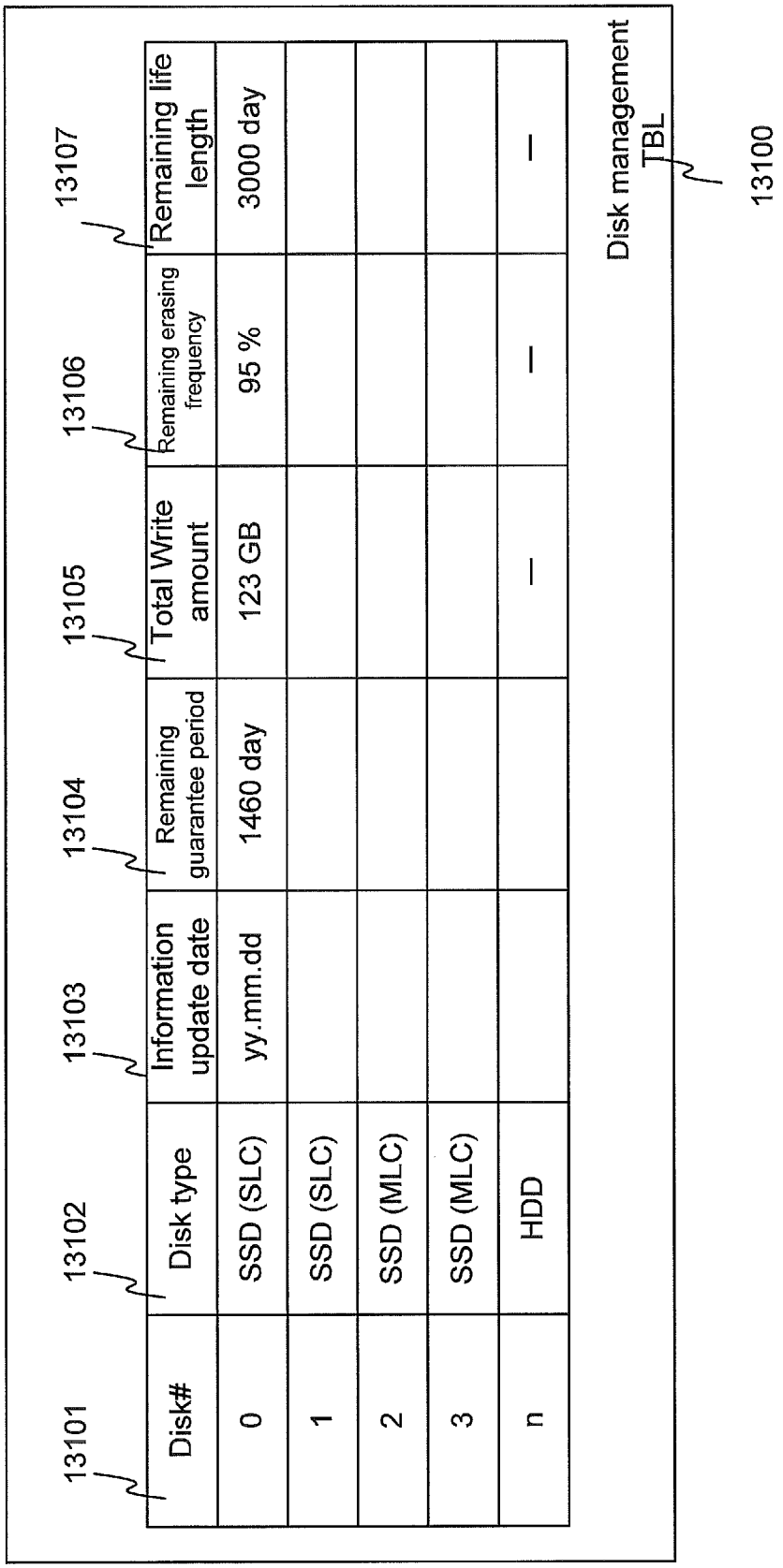
13101
13102
13103
13104
13105
13106
13107
Disk#
Disk type
Information update date
Remaining guarantee period
Total Write amount
Remaining erasing frequency
Remaining life length
0
SSD (SLC)
yy.mm.dd
1460 day
123 GB
95 %
3000 day
1
SSD (SLC)
2
SSD (MLC)
3
SSD (MLC)
n
HDD
—
—
—
Disk management TBL
13100

Fig. 4

[Fig. 5]

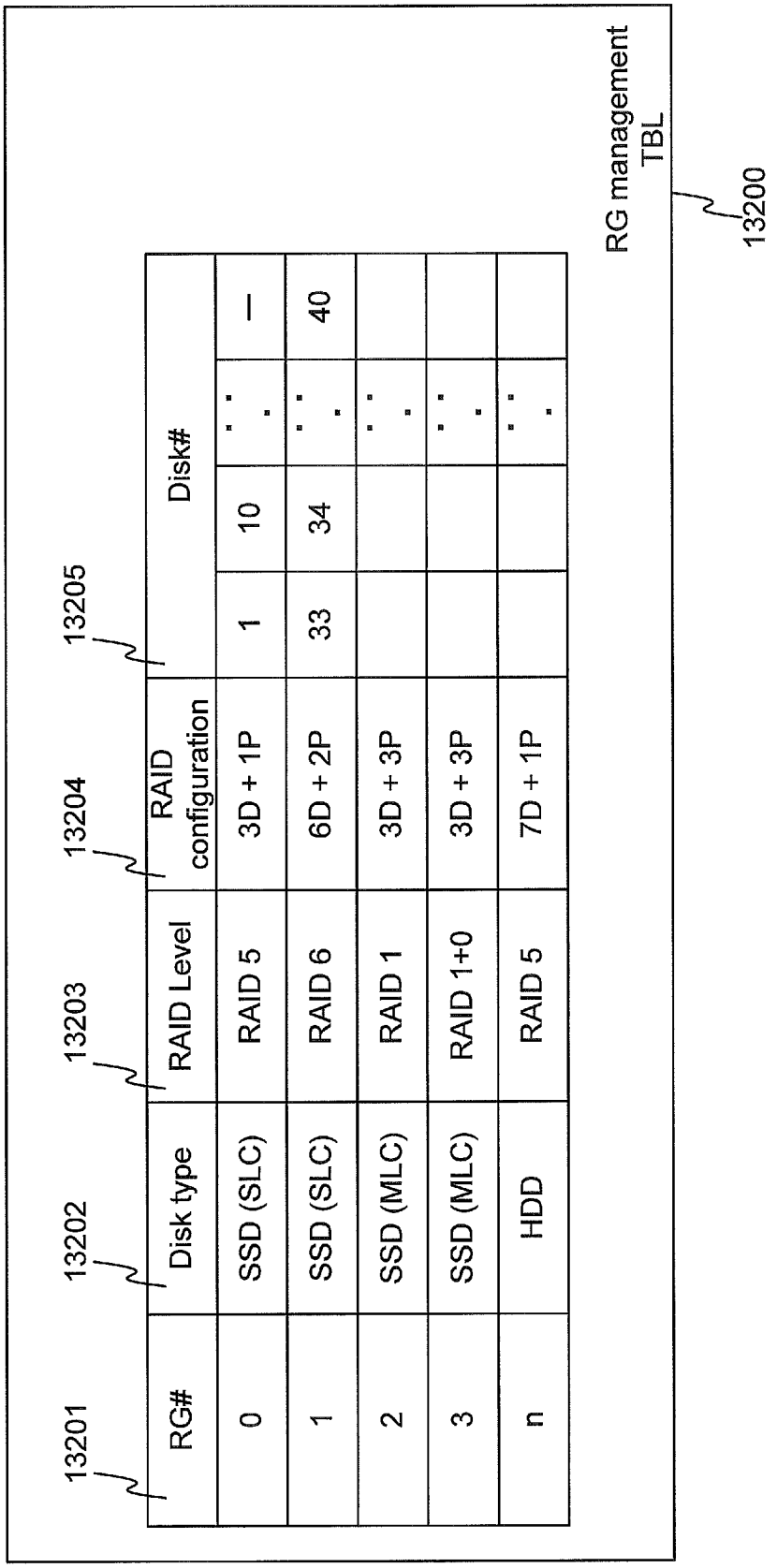
RG management TBL
13200
13201
13202
13203
13204
13205
RG#
Disk type
RAID Level
RAID configuration
Disk#
1
10
—
0
SSD (SLC)
RAID 5
3D + 1P
33
34
40
1
SSD (SLC)
RAID 6
6D + 2P
2
SSD (MLC)
RAID 1
3D + 3P
3
SSD (MLC)
RAID 1+0
3D + 3P
n
HDD
RAID 5
7D + 1P

Fig. 5

[Fig. 6]

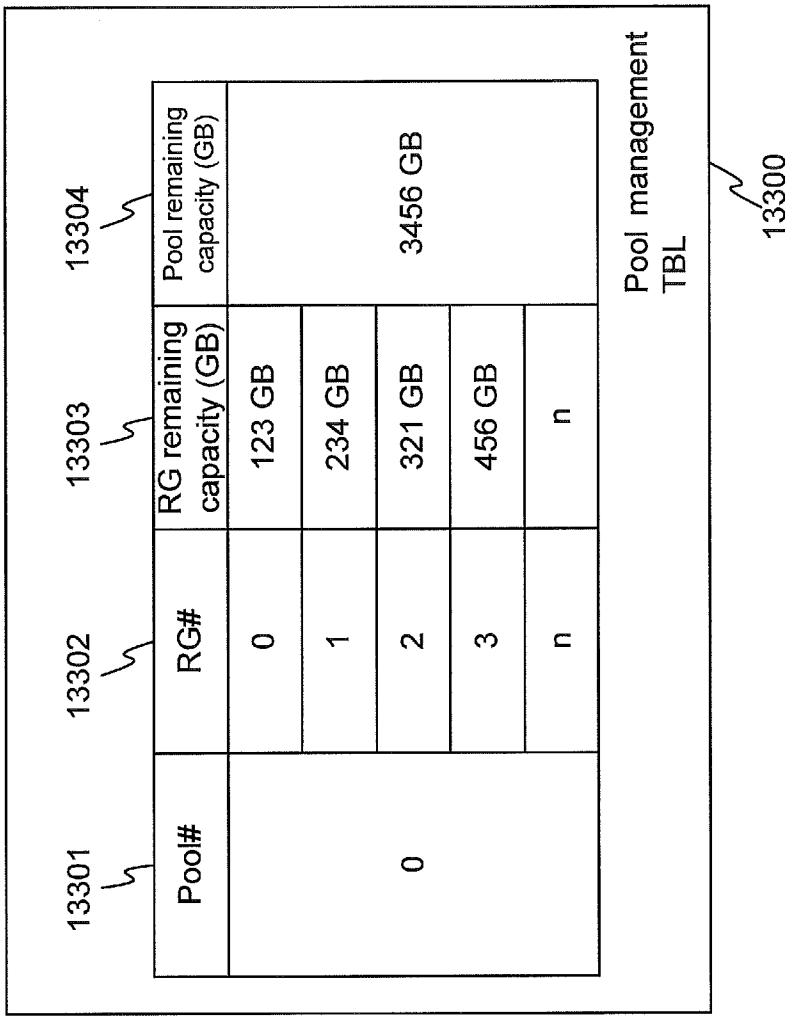
13301
13302
13303
13304
Pool#
RG#
RG remaining capacity (GB)
Pool remaining capacity (GB)
0
0
123 GB
3456 GB
1
234 GB
2
321 GB
3
456 GB
n
n
Pool management TBL
13300

Fig. 6

[Fig. 7]

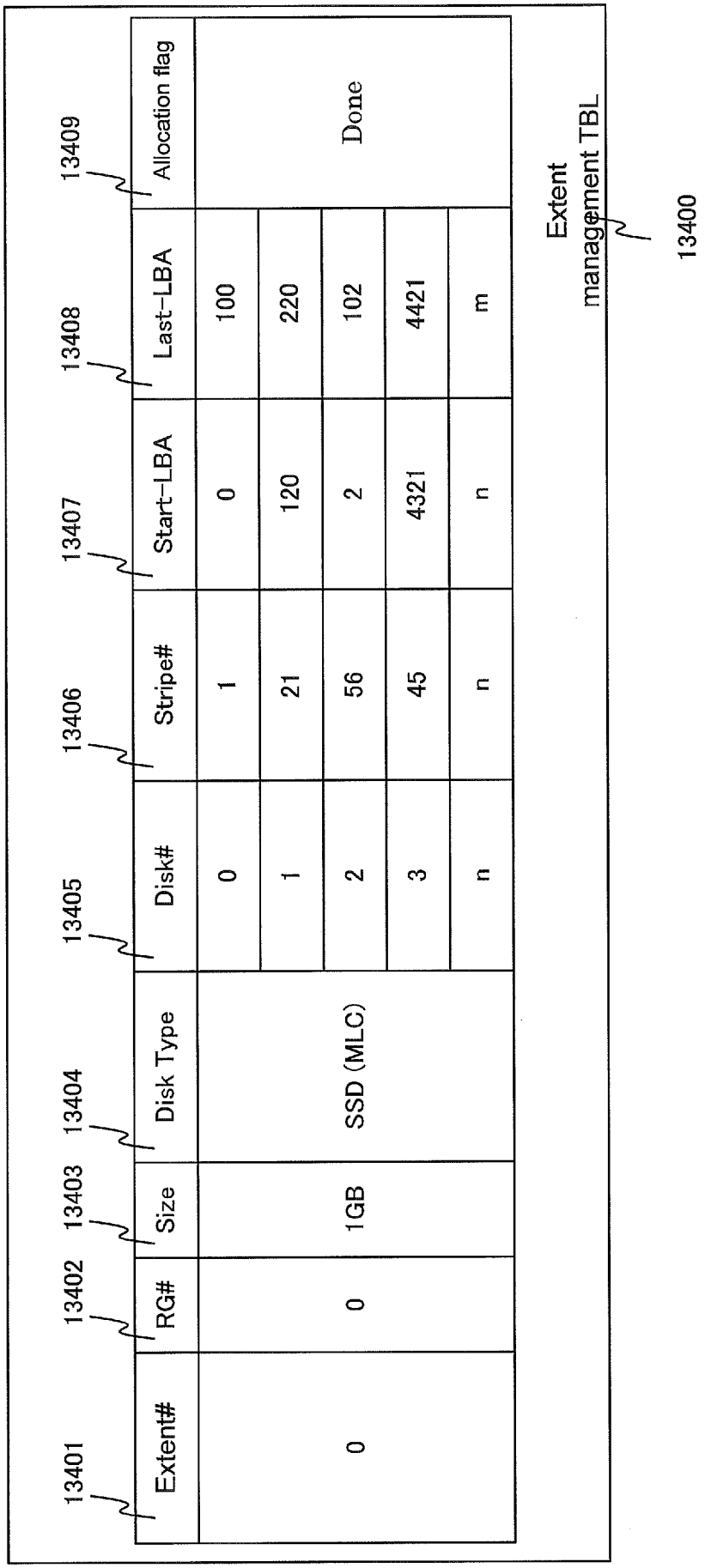
Extent management TBL
13400
13401
Extent#
13402
RG#
13403
Size
13404
Disk Type
13405
Disk#
13406
Stripe#
13407
Start-LBA
13408
Last-LBA
13409
Allocation flag
0
0
1GB
SSD (MLC)
0
1
2
3
n
1
21
56
45
n
0
120
2
4321
n
100
220
102
4421
m
Done

Fig. 7

[Fig. 8]

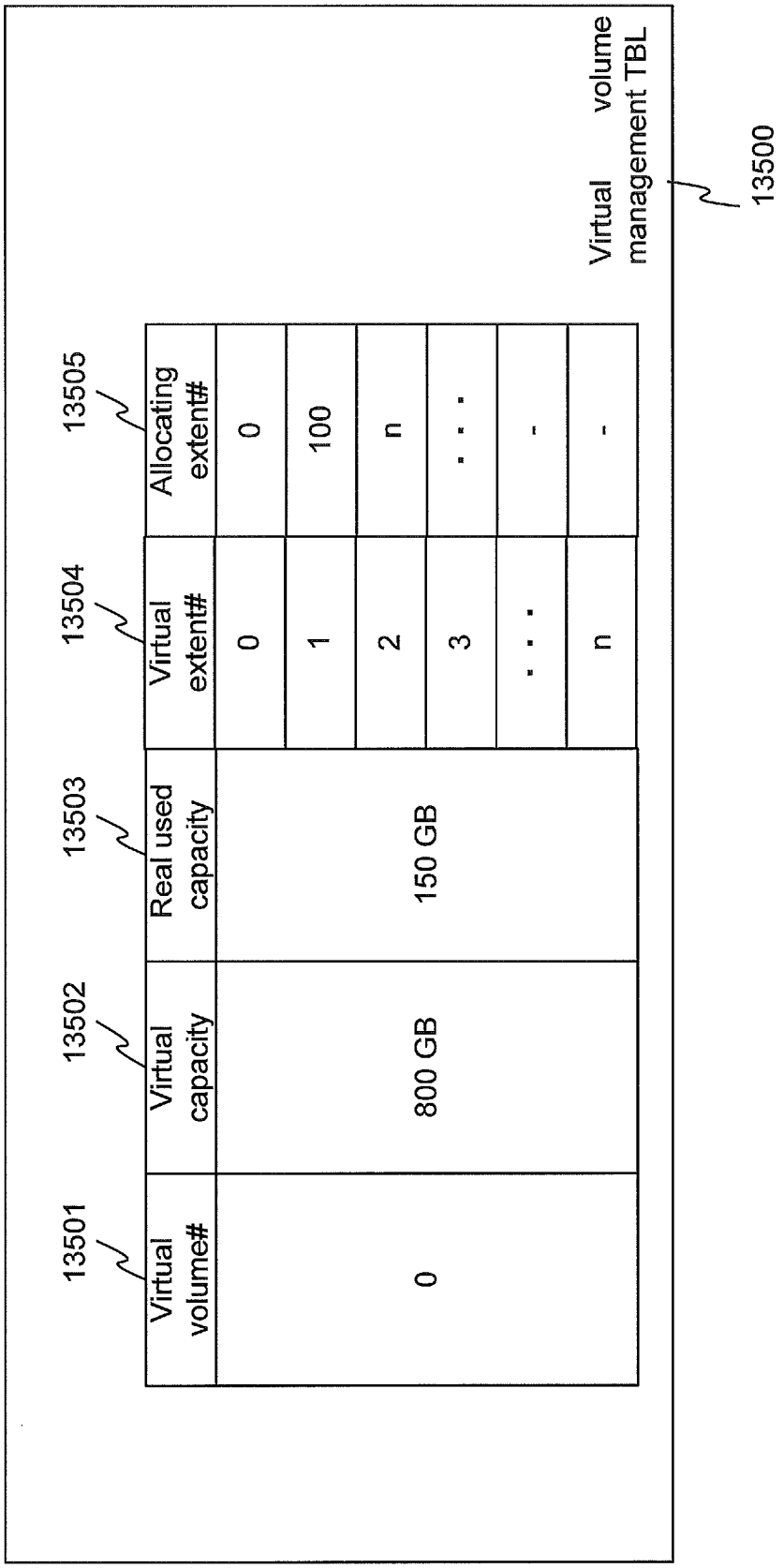
Virtual volume management TBL
13500
13501
Virtual volume#
0
13502
Virtual capacity
800 GB
13503
Real used capacity
150 GB
13504
Virtual extent#
0
1
2
3
n
13505
Allocating extent#
0
100
n
-
-

Fig. 8

[Fig. 9]

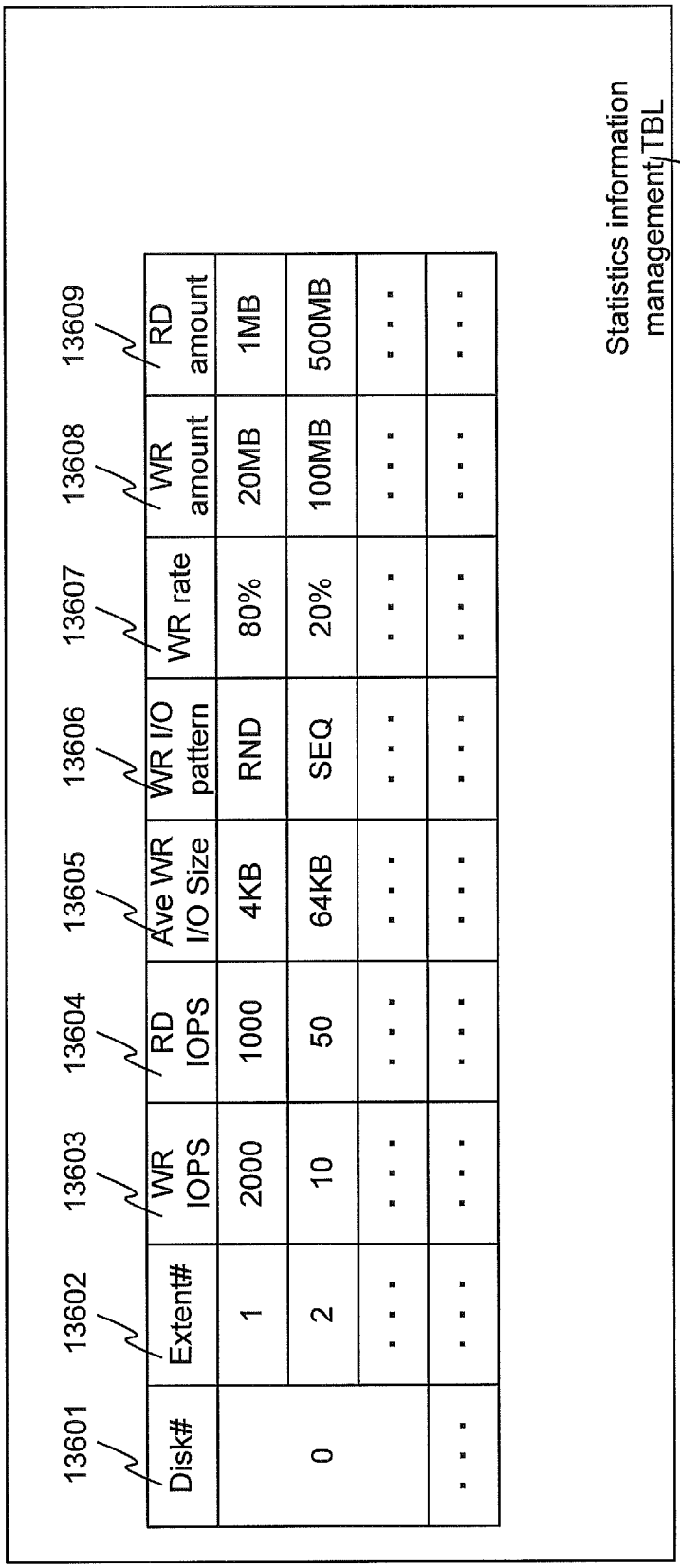
13601
13602
13603
13604
13605
13606
13607
13608
13609
Disk#
Extent#
WR IOPS
RD IOPS
Ave WR I/O Size
WR I/O pattern
WR rate
WR amount
RD amount
0
1
2000
1000
4KB
RND
80%
20MB
1MB
2
10
50
64KB
SEQ
20%
100MB
500MB
Statistics information management TBL
13600

Fig. 9

[Fig. 10]

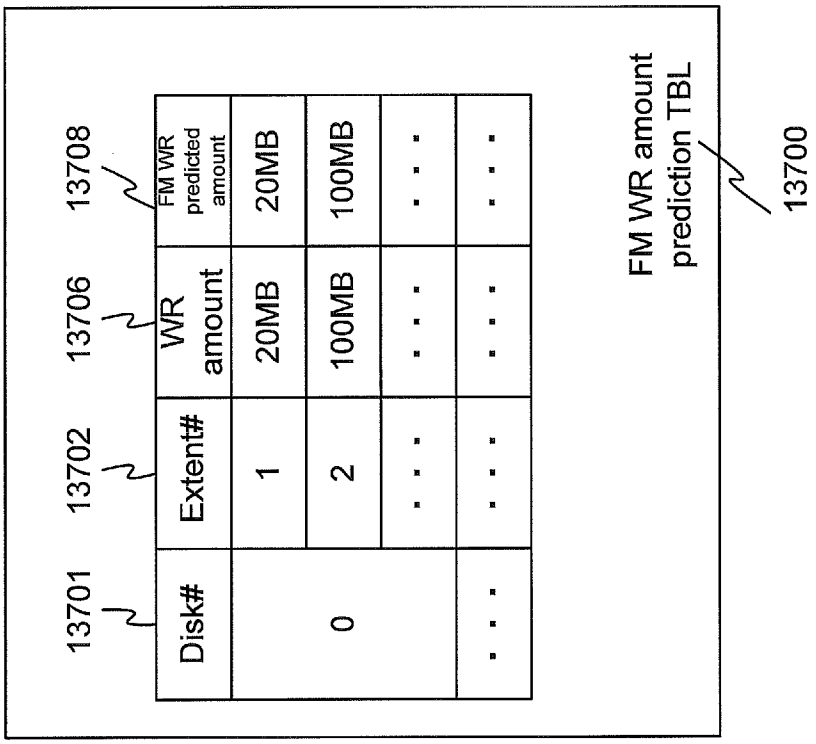
13701
13702
13706
13708
Disk#
Extent#
WR amount
FM WR predicted amount
0
1
20MB
20MB
2
100MB
100MB
FM WR amount prediction TBL
13700

Fig. 10

[Fig. 11]

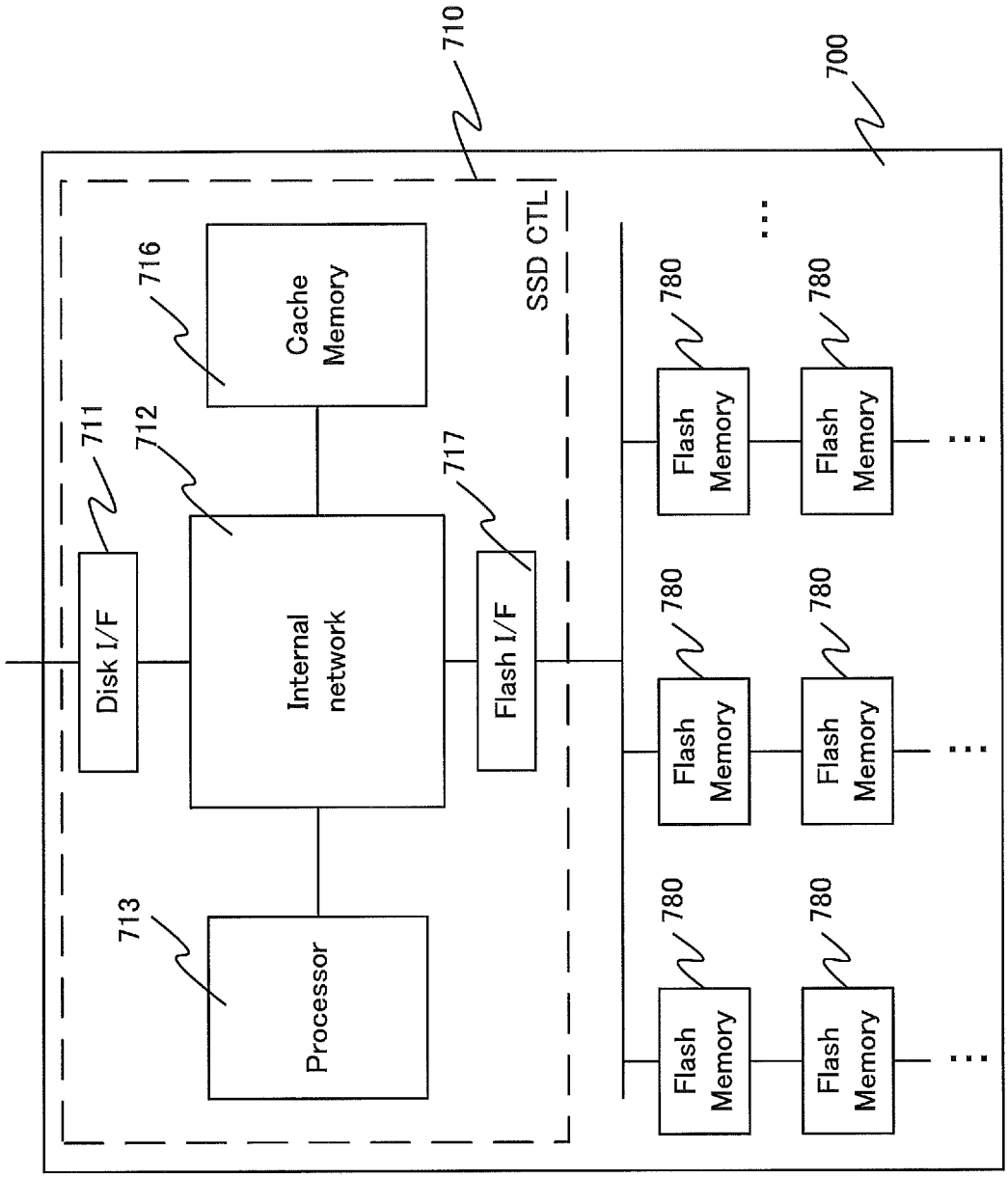
710
700
SSD CTL
716
Cache Memory
711
712
717
Disk I/F
Internal network
Flash I/F
713
Processor
780
Flash Memory

Fig. 11

[Fig. 12]

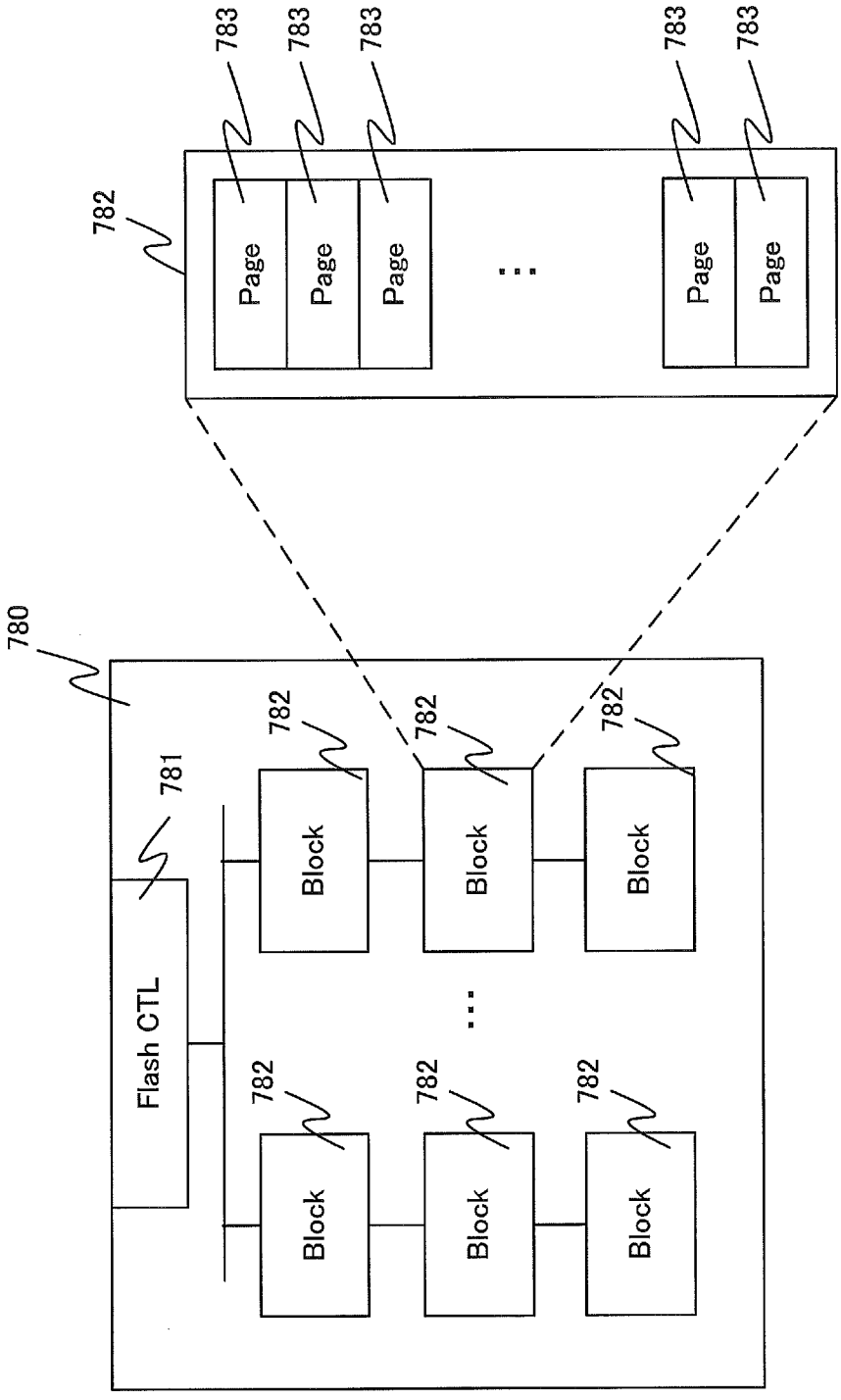


Fig. 12

[Fig. 13]

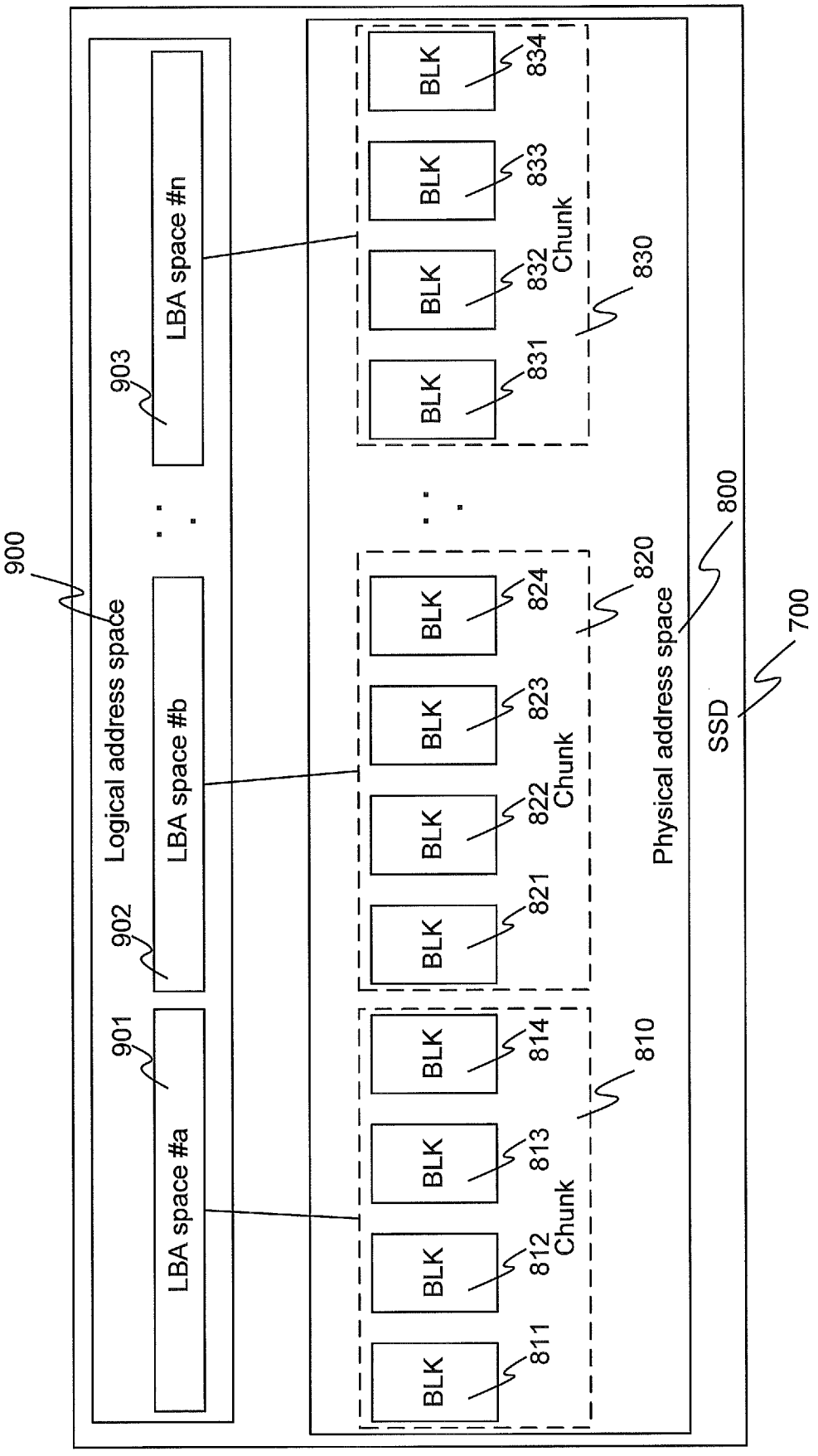
900
Logical address space
901
LBA space #a
902
LBA space #b
903
LBA space #n
SSD
700
800
Physical address space
810
820
830
Chunk
BLK
811
812
813
814
821
822
823
824
831
832
833
834

Fig. 13

[Fig. 14]

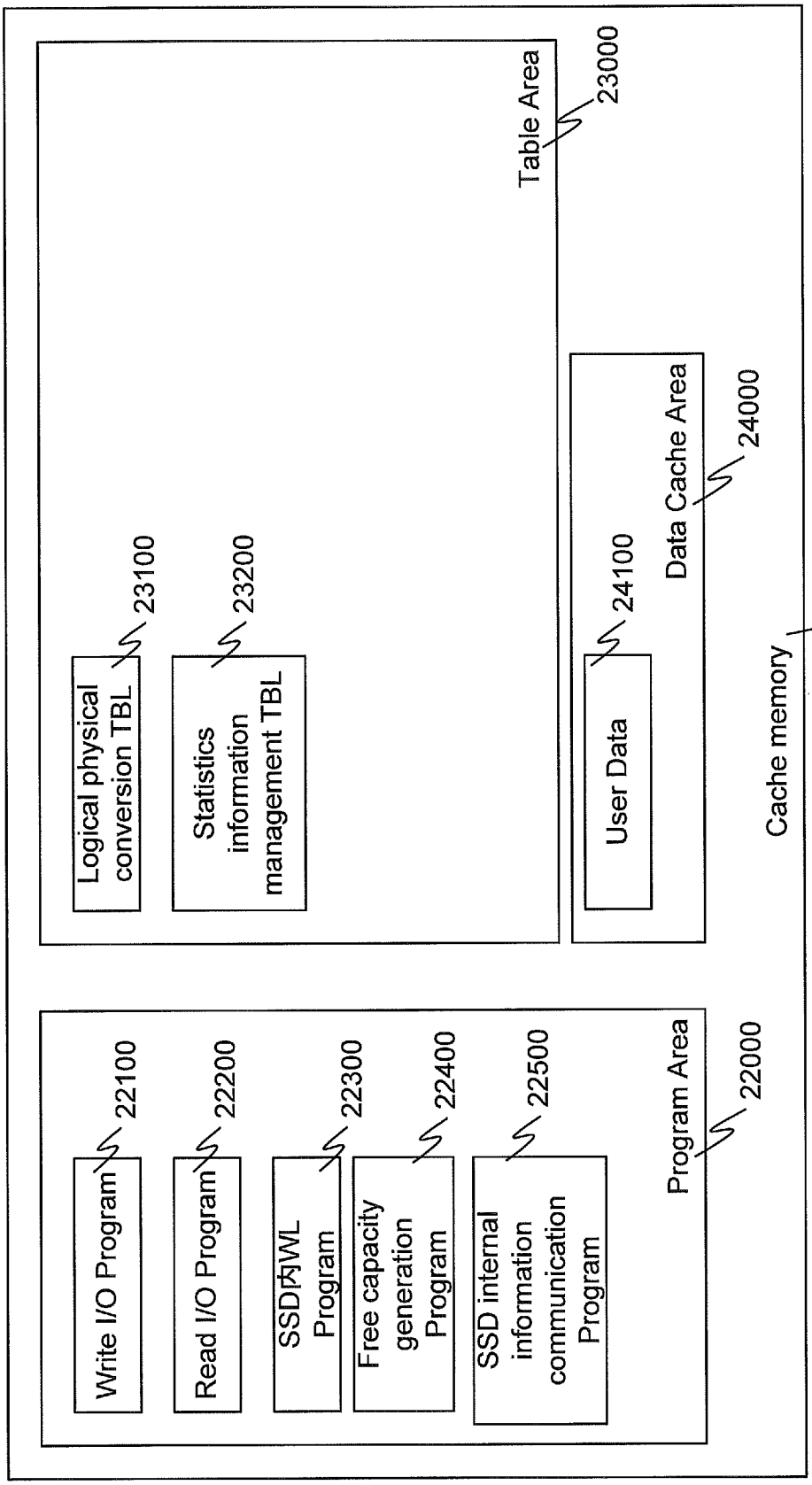
Cache memory
716
Program Area
22000
Write I/O Program
22100
Read I/O Program
22200
SSD内WL Program
22300
Free capacity generation Program
22400
SSD internal information communication Program
22500
Data Cache Area
24000
User Data
24100
Table Area
23000
Logical physical conversion TBL
23100
Statistics information management TBL
23200

Fig. 14

[Fig. 15]

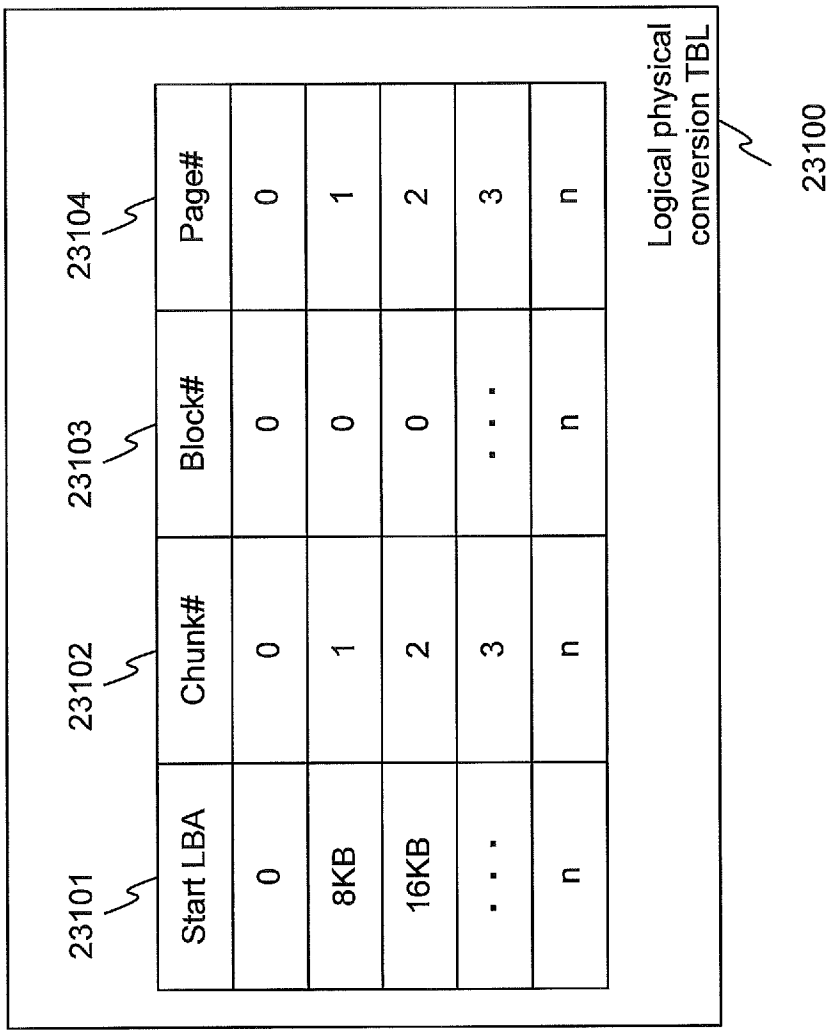
Logical physical conversion TBL
23100
23101
23102
23103
23104
Start LBA
Chunk#
Block#
Page#
0
0
0
0
8KB
1
0
1
16KB
2
0
2
. . . .
3
. . . .
3
n
n
n
n

Fig. 15

[Fig. 16]

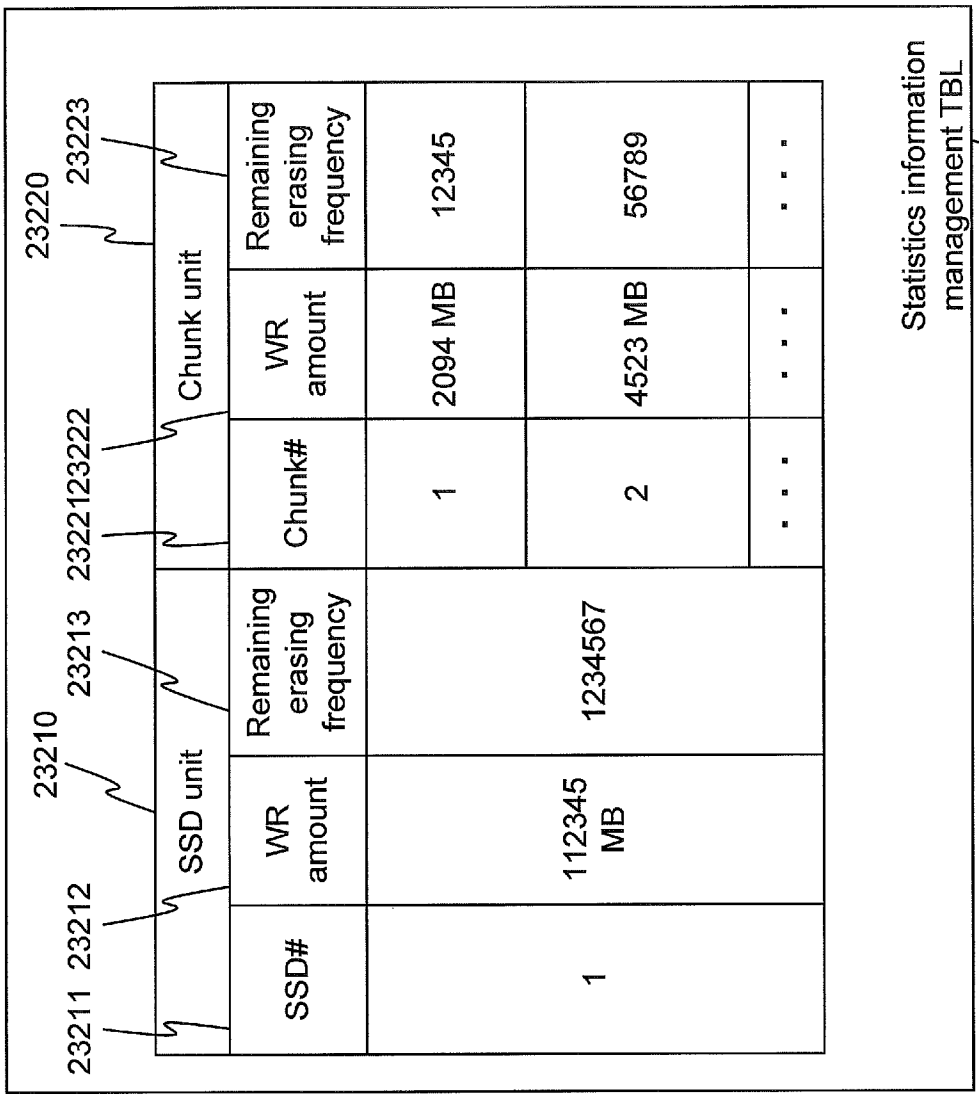
23210
23211
23212
SSD unit
SSD#
WR amount
Remaining erasing frequency
1
112345 MB
1234567
23213
23220
23221
23222
23223
Chunk unit
Chunk#
WR amount
Remaining erasing frequency
1
2094 MB
12345
2
4523 MB
56789
Statistics information management TBL
23200

Fig. 16

[Fig. 17]

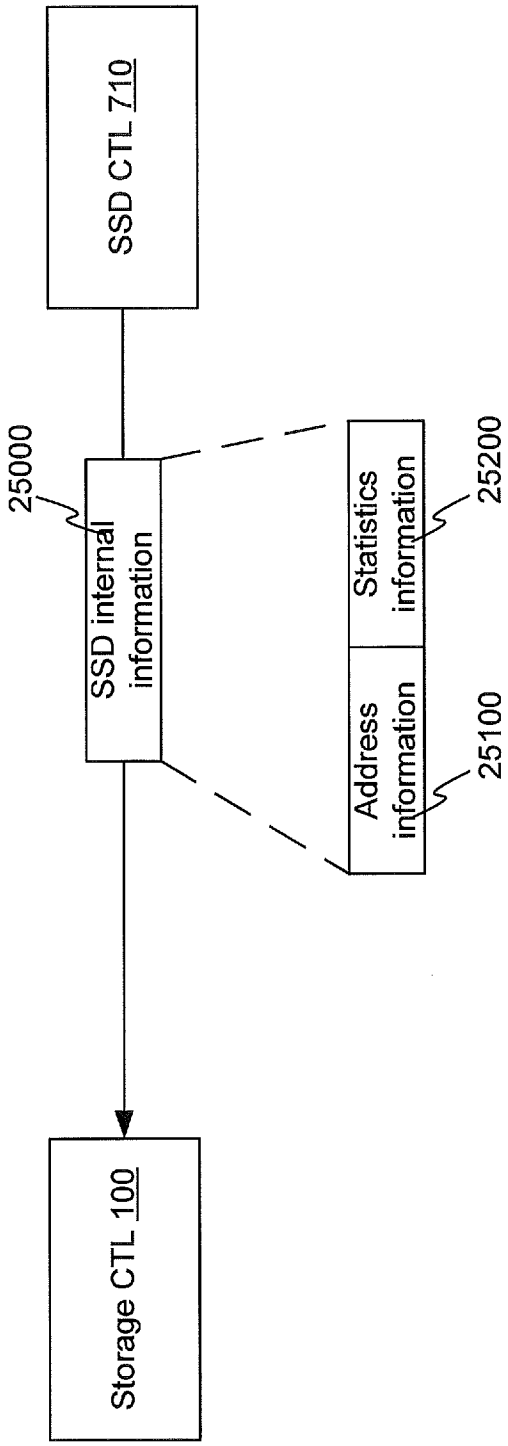
SSD CTL 710
25000
SSD internal information
Statistics information
25200
Address information
25100
Storage CTL 100

Fig. 17

[Fig. 18]

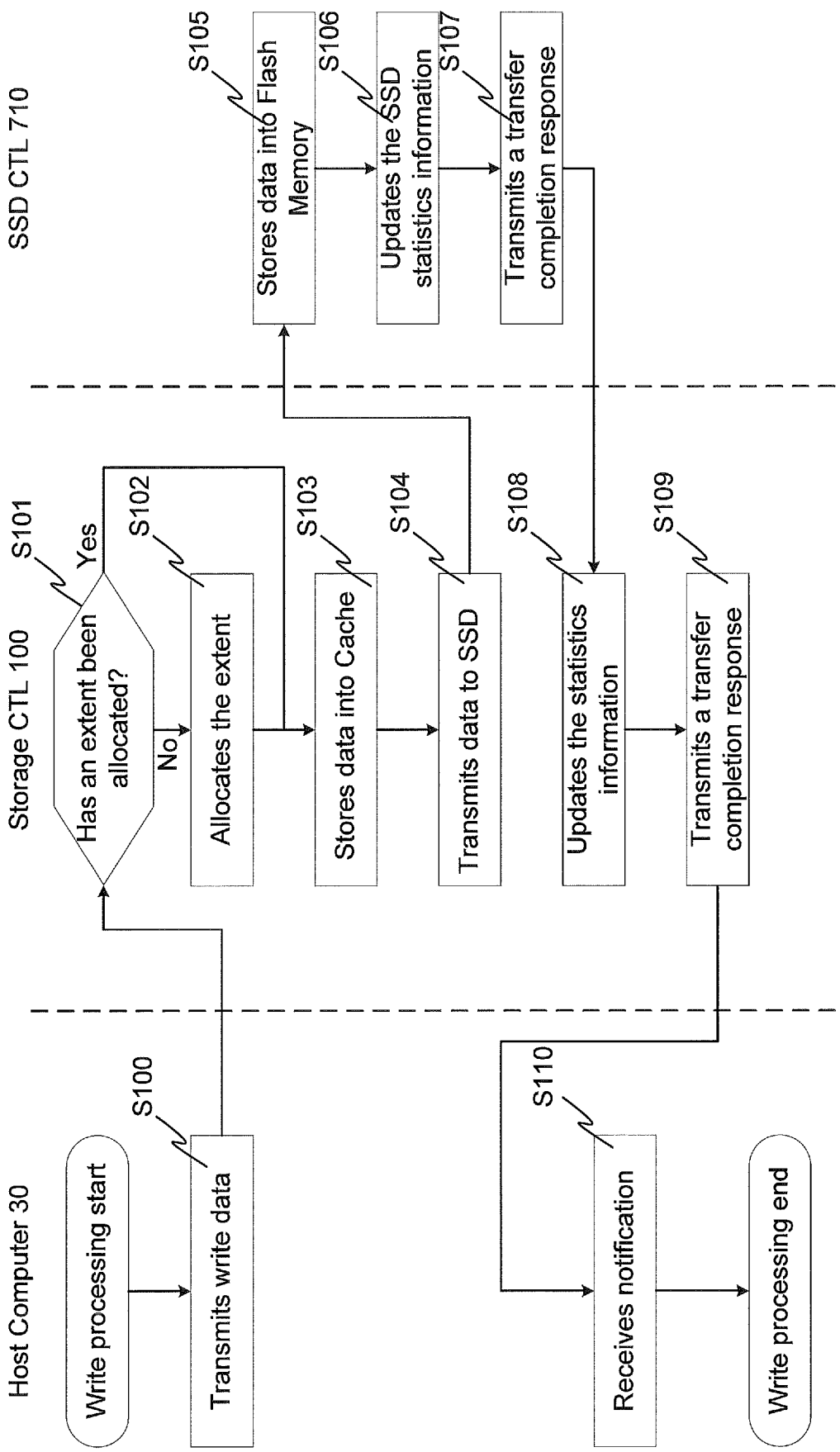
Host Computer 30
Write processing start
S100
Transmits write data
Storage CTL 100
S101
Has an extent been allocated?
Yes
No
S102
Allocates the extent
S103
Stores data into Cache
S104
Transmits data to SSD
SSD CTL 710
S105
Stores data into Flash Memory
S106
Updates the SSD statistics information
S107
Transmits a transfer completion response
S108
Updates the statistics information
S109
Transmits a transfer completion response
S110
Receives notification
Write processing end

Fig. 18

[Fig. 19]

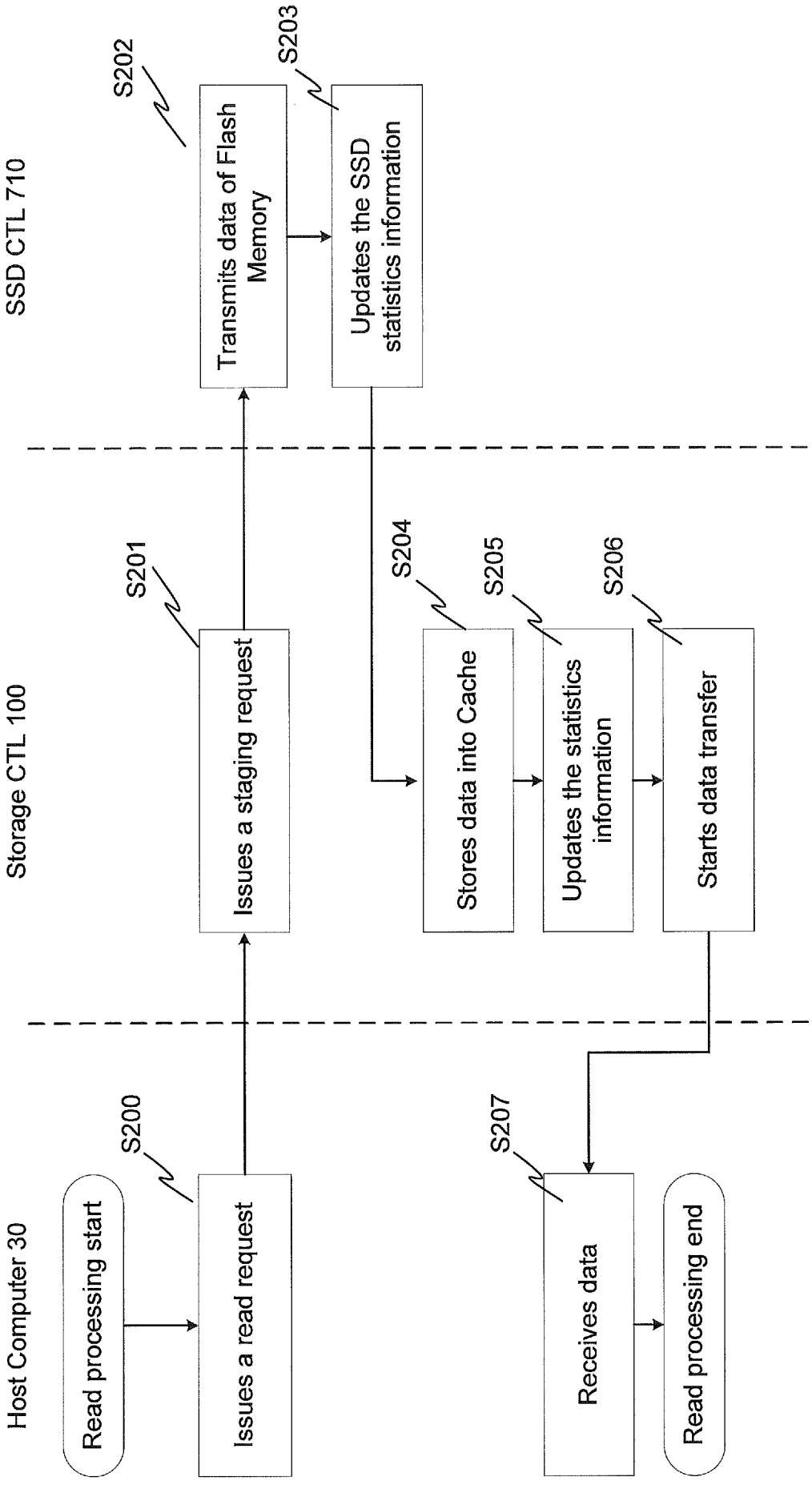
Host Computer 30
Storage CTL 100
SSD CTL 710
Read processing start
S200
Issues a read request
S201
Issues a staging request
S202
Transmits data of Flash Memory
S203
Updates the SSD statistics information
S204
Stores data into Cache
S205
Updates the statistics information
S206
Starts data transfer
S207
Receives data
Read processing end

Fig. 19

[Fig. 20]

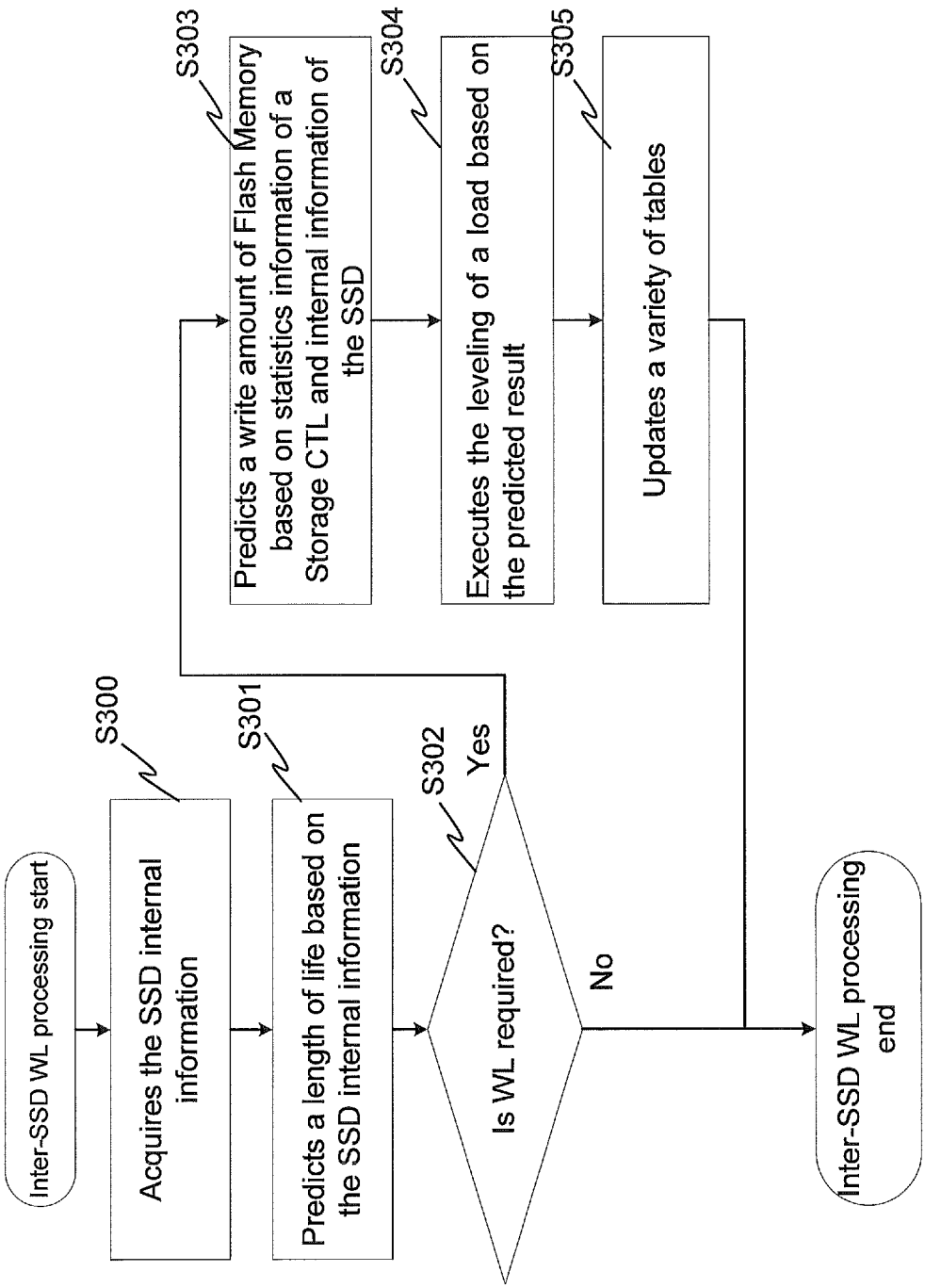
Inter-SSD WL processing start
S300
Acquires the SSD internal information
S301
Predicts a length of life based on the SSD internal information
S302
Is WL required?
Yes
No
S303
Predicts a write amount of Flash Memory based on statistics information of a Storage CTL and internal information of the SSD
S304
Executes the leveling of a load based on the predicted result
S305
Updates a variety of tables
Inter-SSD WL processing end

Fig. 20

[Fig. 21]

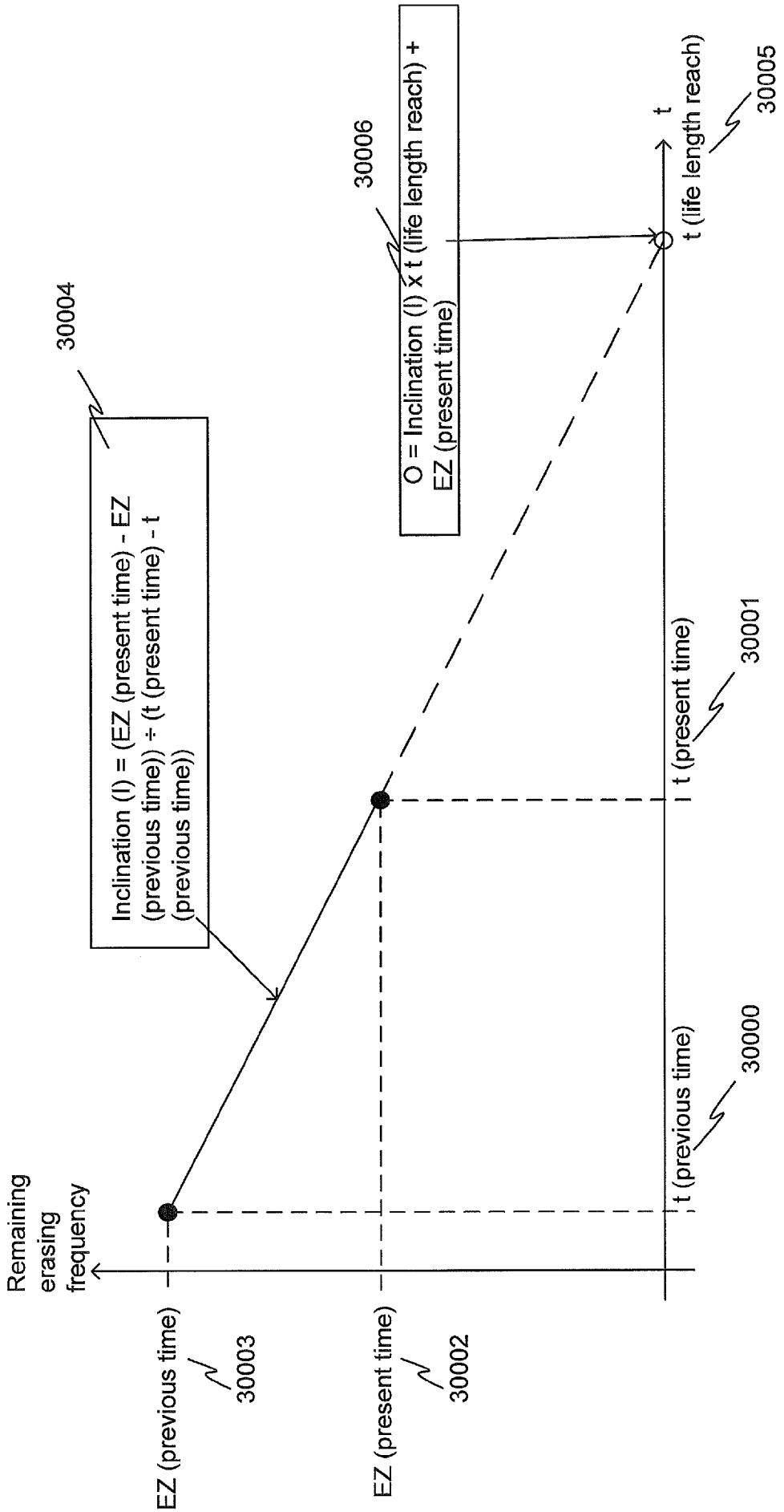
Remaining erasing frequency
EZ (previous time)
30003
EZ (present time)
30002
t (previous time)
30000
t (present time)
30001
t (life length reach)
30005
t
30004
Inclination (I) = (EZ (present time) - EZ (previous time)) ÷ (t (present time) - t (previous time))
30006
O = Inclination (I) x t (life length reach) + EZ (present time)

Fig. 21

[Fig. 22]

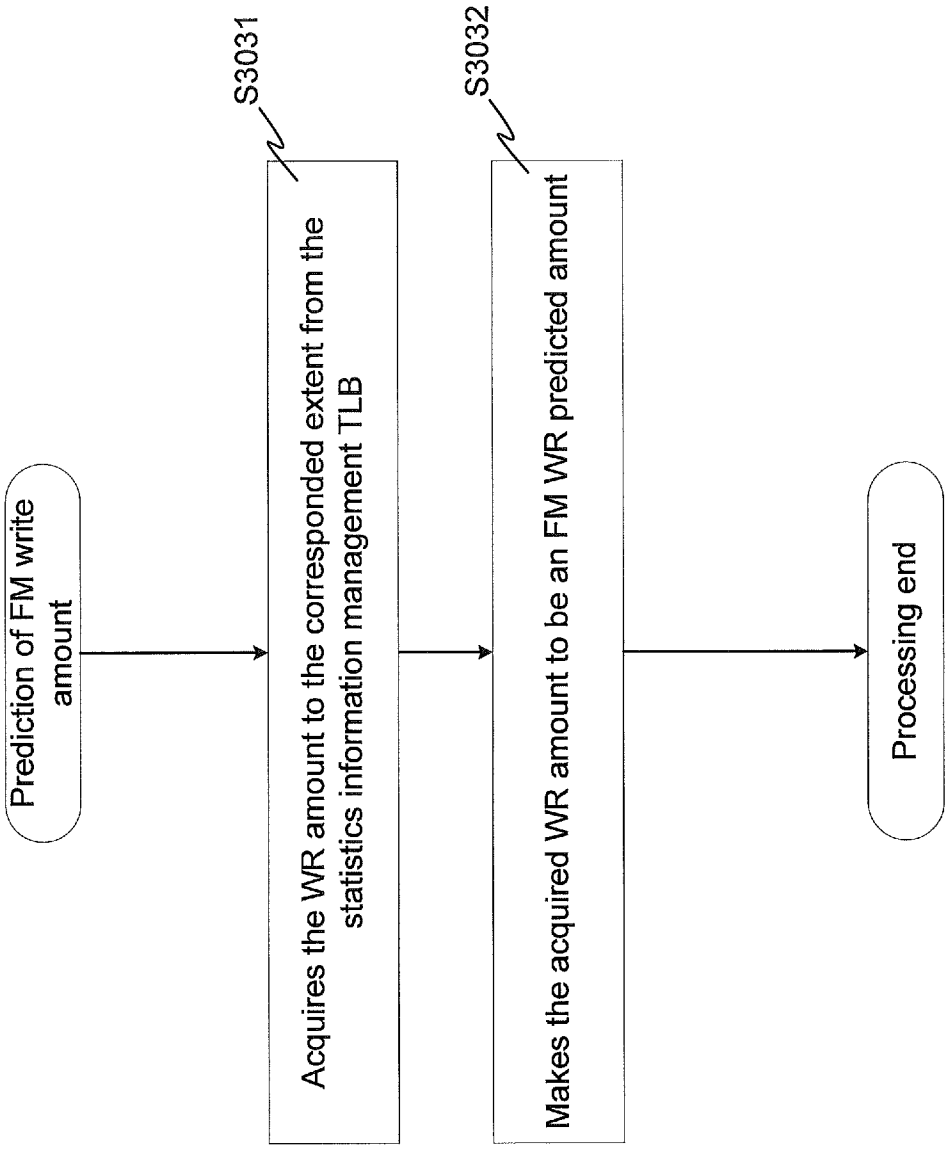
Prediction of FM write amount
S3031
Acquires the WR amount to the corresponded extent from the statistics information management TLB
S3032
Makes the acquired WR amount to be an FM WR predicted amount
Processing end

Fig. 22

[Fig. 23]

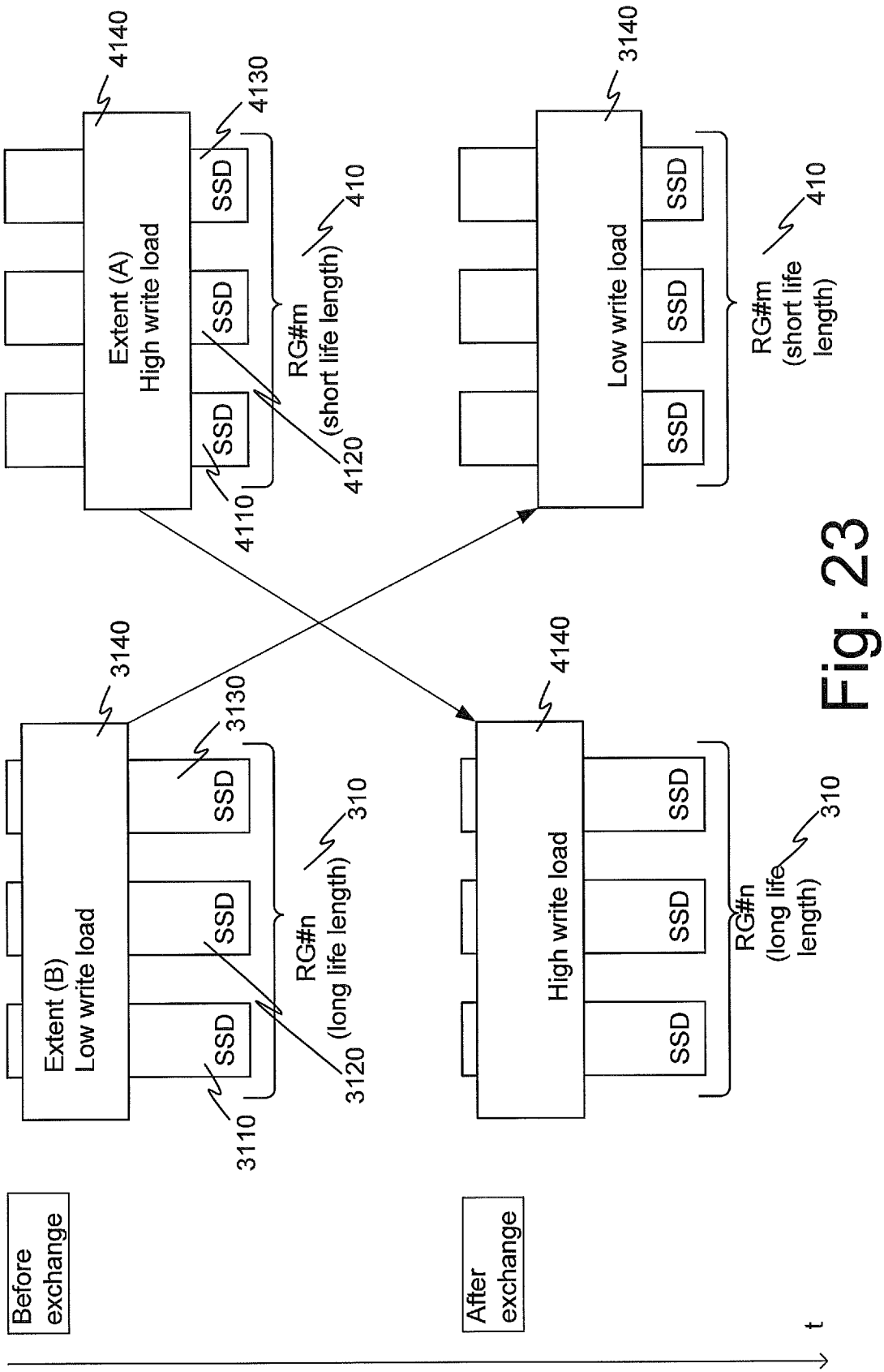
Before exchange
After exchange
t
Extent (A)
High write load
Extent (B)
Low write load
Low write load
High write load
SSD
RG#m
(short life length)
RG#n
(long life length)
4140
4130
4120
4110
410
3140
3130
3120
3110
310

Fig. 23

[Fig. 24]

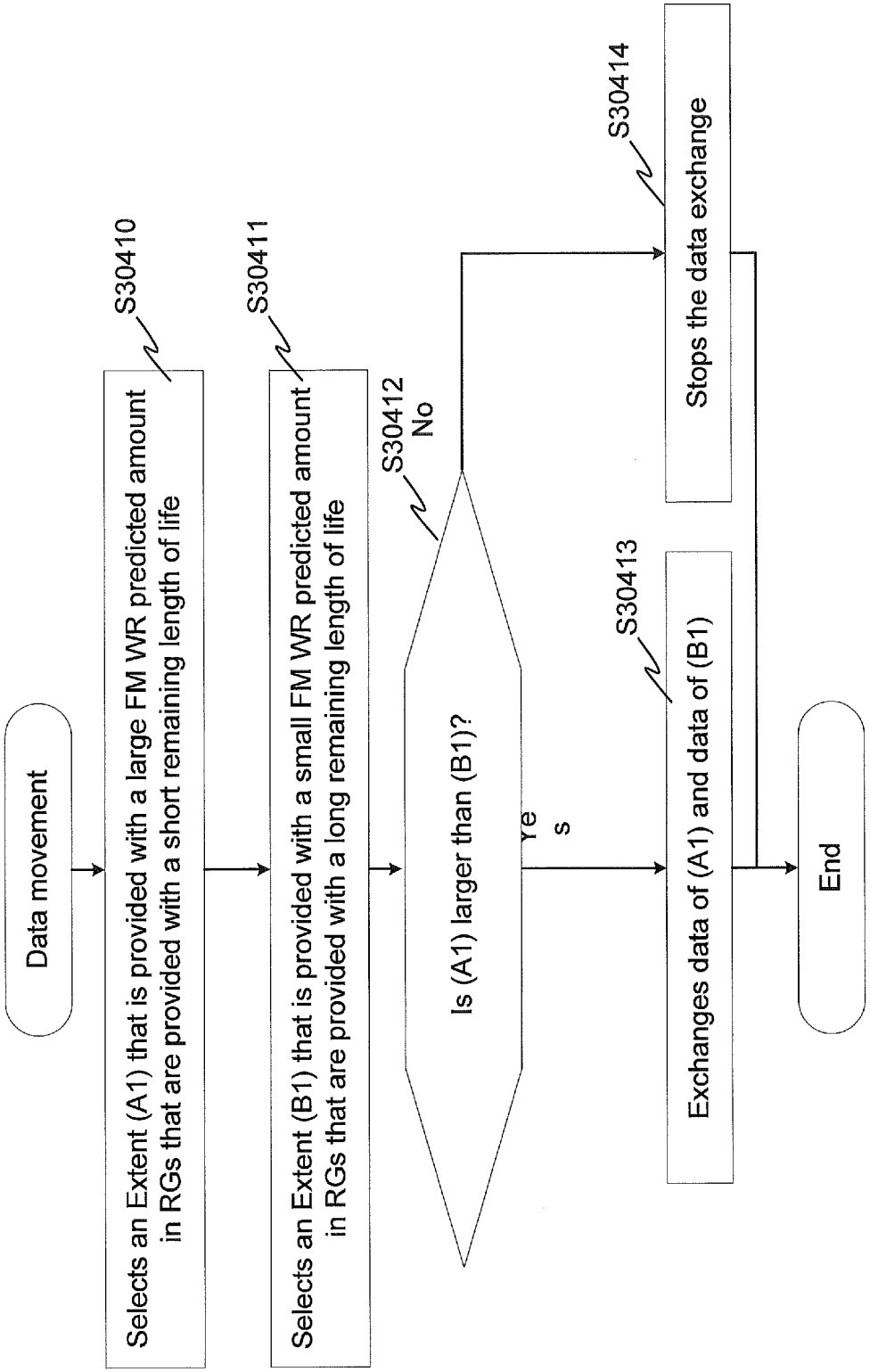
Data movement
S30410
Selects an Extent (A1) that is provided with a large FM WR predicted amount in RGs that are provided with a short remaining length of life
S30411
Selects an Extent (B1) that is provided with a small FM WR predicted amount in RGs that are provided with a long remaining length of life
S30412
Is (A1) larger than (B1)?
Yes
No
S30413
Exchanges data of (A1) and data of (B1)
S30414
Stops the data exchange
End

Fig. 24

[Fig. 25]

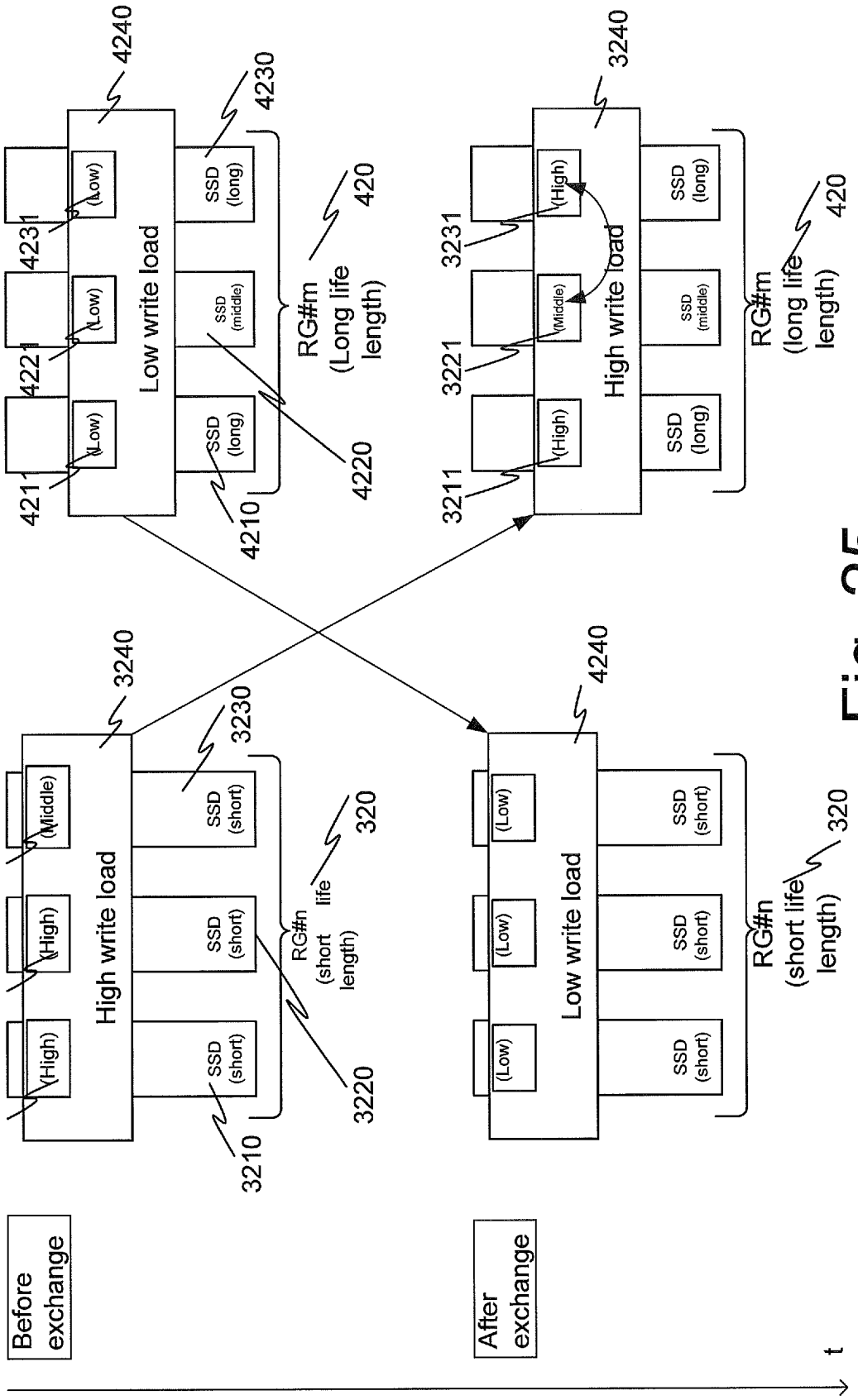
Before exchange
After exchange
t
High write load
Low write load
(High)
(Middle)
(Low)
SSD (short)
SSD (middle)
SSD (long)
RG#n (short life length)
RG#m (Long life length)
RG#m (long life length)
3211
3221
3231
3210
3220
3230
3240
320
420
4210
4220
4230
4240

Fig. 25

[Fig. 26]

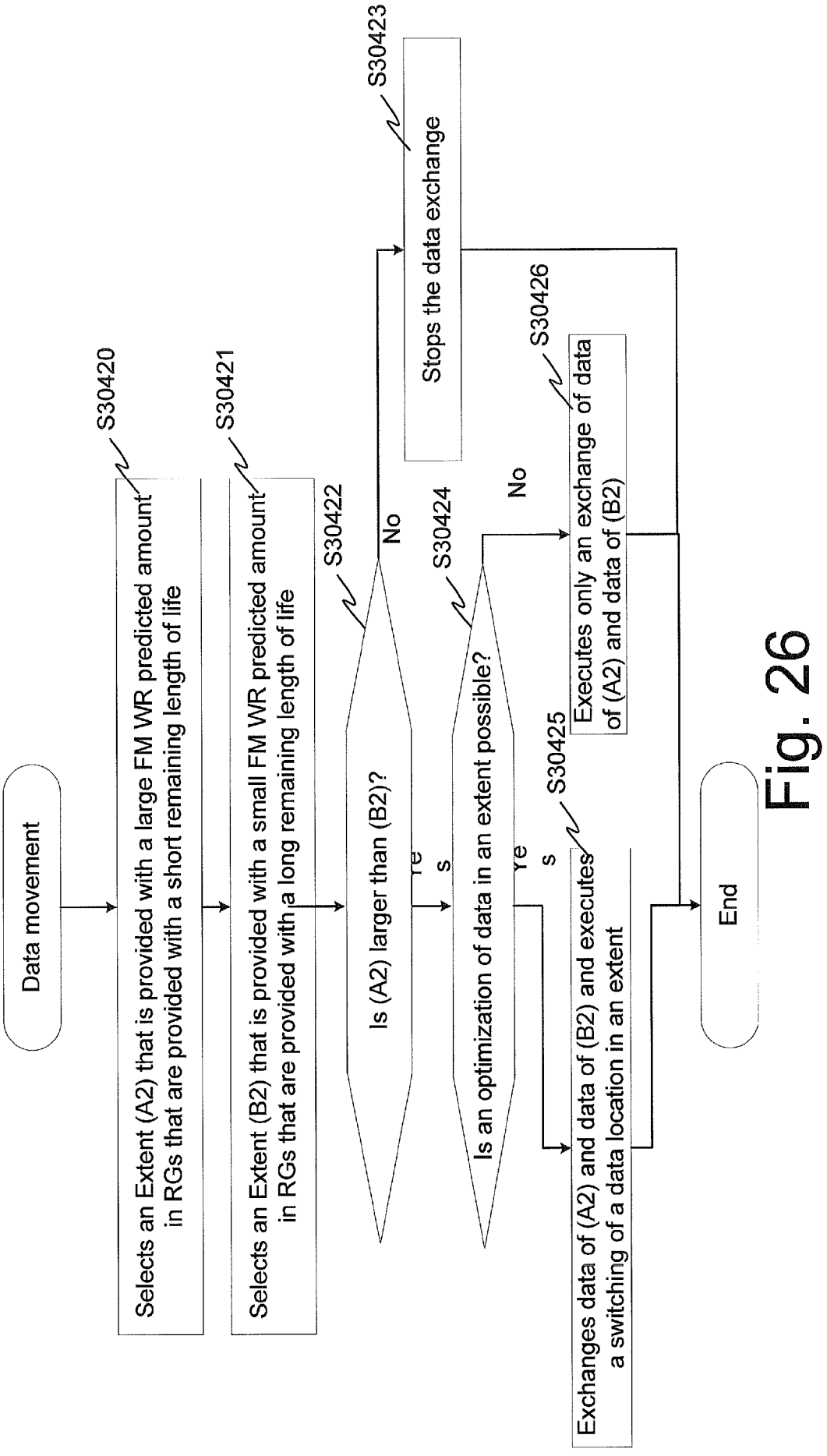
Data movement
S30420
Selects an Extent (A2) that is provided with a large FM WR predicted amount in RGs that are provided with a short remaining length of life
S30421
Selects an Extent (B2) that is provided with a small FM WR predicted amount in RGs that are provided with a long remaining length of life
S30422
Is (A2) larger than (B2)?
Yes
No
S30423
Stops the data exchange
S30424
Is an optimization of data in an extent possible?
Yes
No
S30425
Exchanges data of (A2) and data of (B2) and executes a switching of a data location in an extent
S30426
Executes only an exchange of data of (A2) and data of (B2)
End

Fig. 26

[Fig. 27]

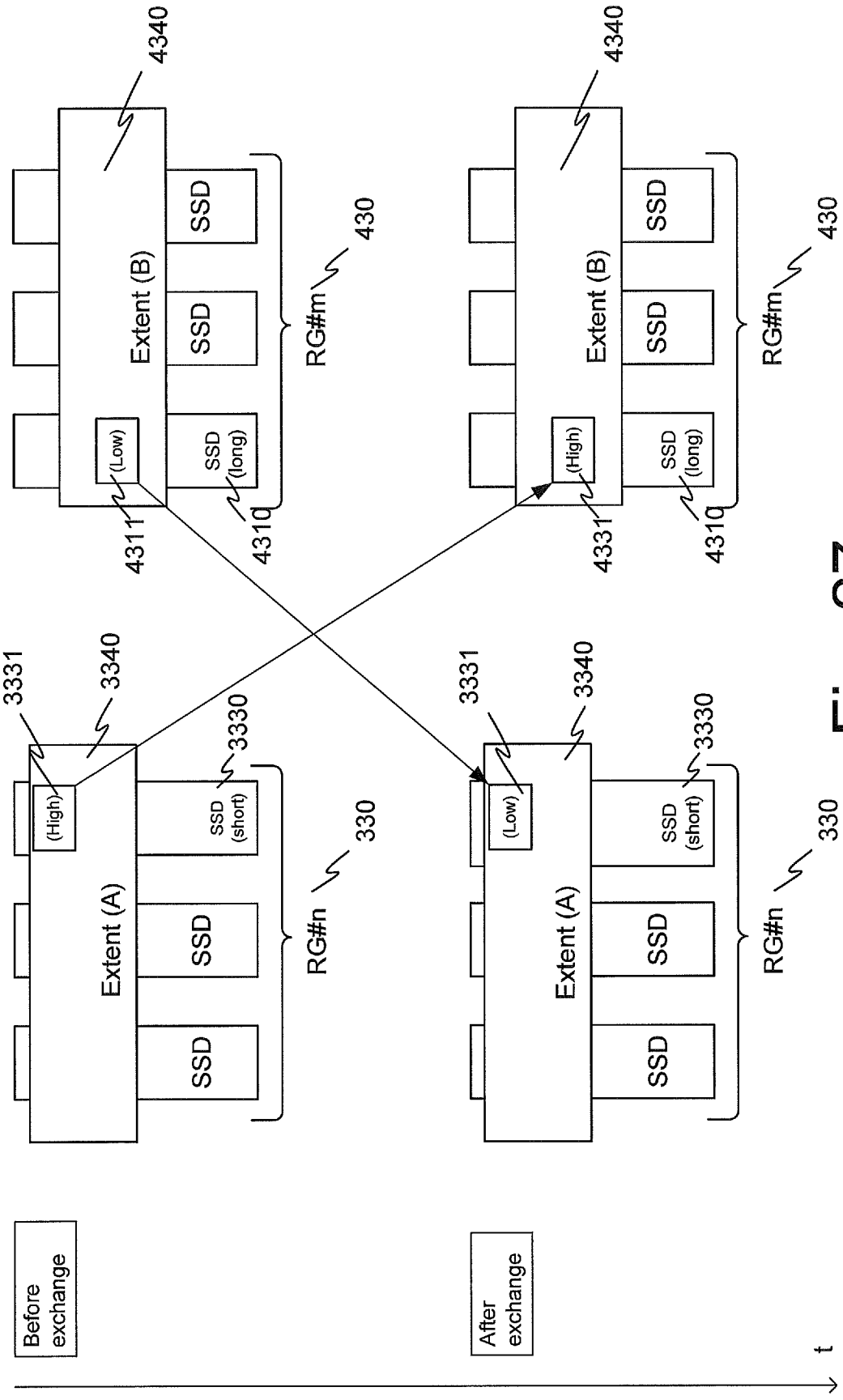
Before exchange
After exchange
t
RG#n
330
RG#m
430
Extent (A)
Extent (B)
SSD
SSD (short)
SSD (long)
(High)
(Low)
3331
3340
3330
4340
4311
4310
4331

Fig. 27

[Fig. 28]

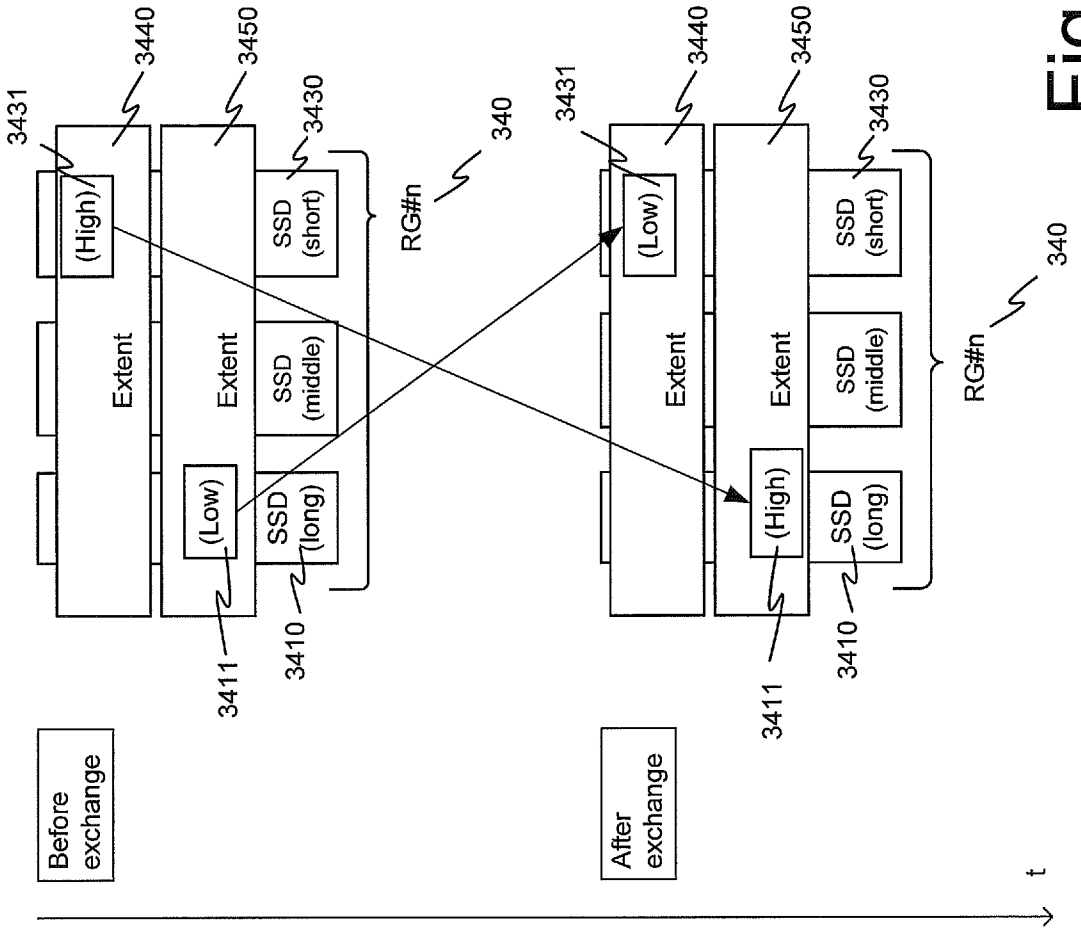
Before exchange
After exchange
t
RG#n
340
3410
3411
3430
3431
3440
3450
Extent
(High)
(Low)
SSD (short)
SSD (middle)
SSD (long)

Fig. 28

[Fig. 29]

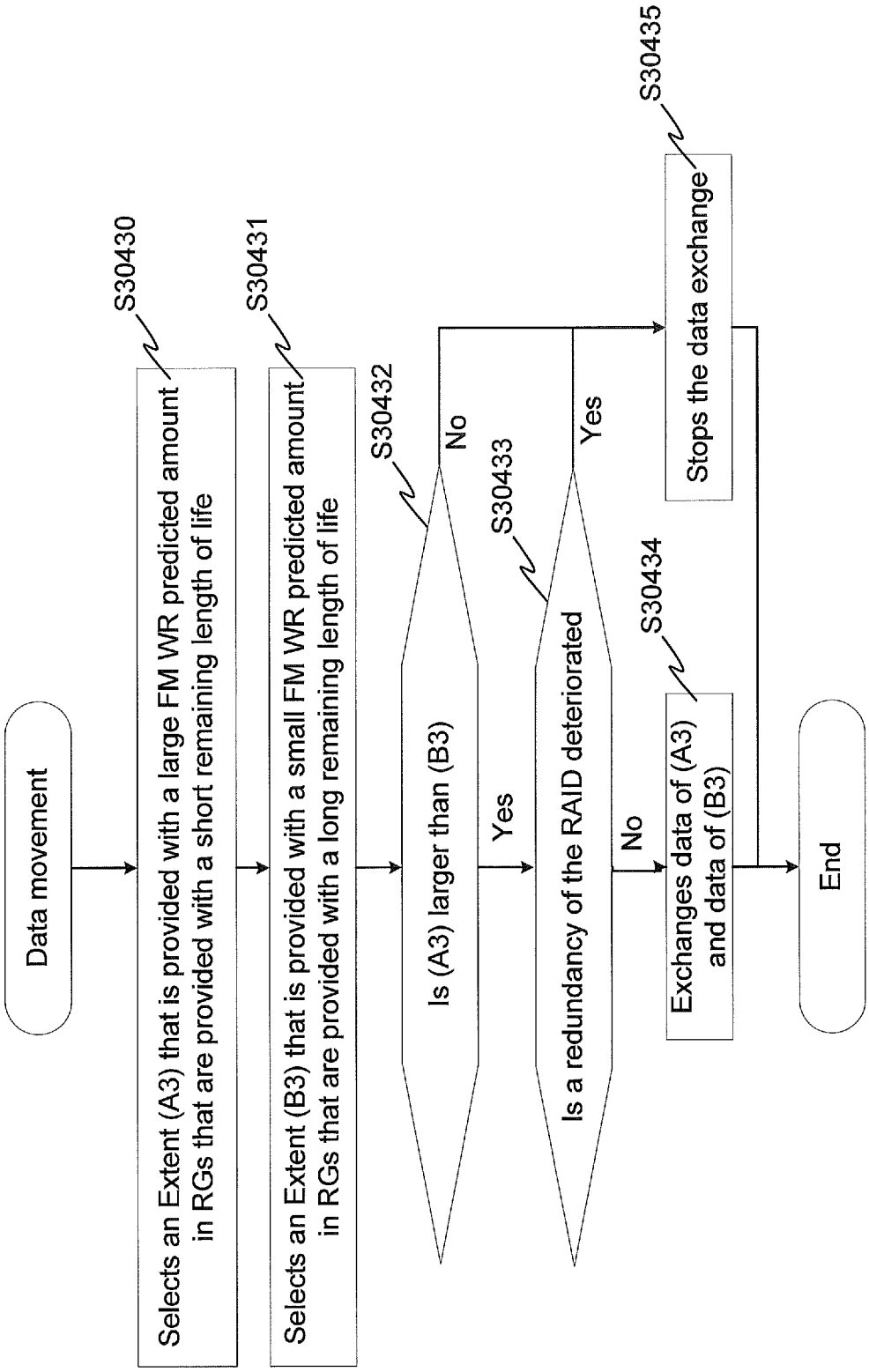
Data movement
S30430
Selects an Extent (A3) that is provided with a large FM WR predicted amount in RGs that are provided with a short remaining length of life
S30431
Selects an Extent (B3) that is provided with a small FM WR predicted amount in RGs that are provided with a long remaining length of life
S30432
Is (A3) larger than (B3)
No
Yes
S30433
Is a redundancy of the RAID deteriorated
Yes
No
S30434
Exchanges data of (A3) and data of (B3)
S30435
Stops the data exchange
End

Fig. 29

[Fig. 30]

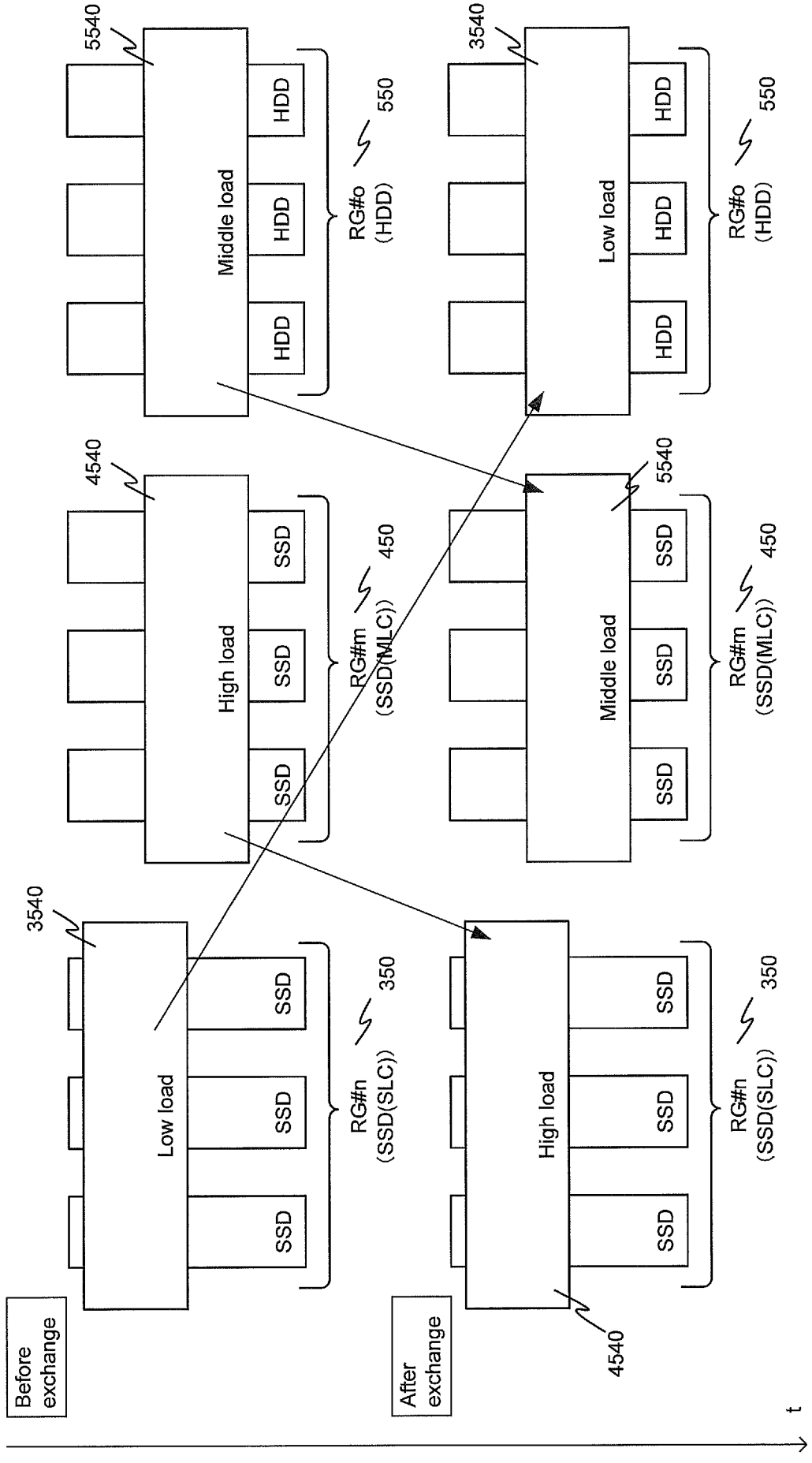
Before exchange
After exchange
t
Low load
High load
Middle load
Low load
High load
Middle load
SSD
HDD
RG#n (SSD(SLC))
RG#m (SSD(MLC))
RG#o (HDD)
350
450
550
3540
4540
5540

Fig. 30

[Fig. 31]

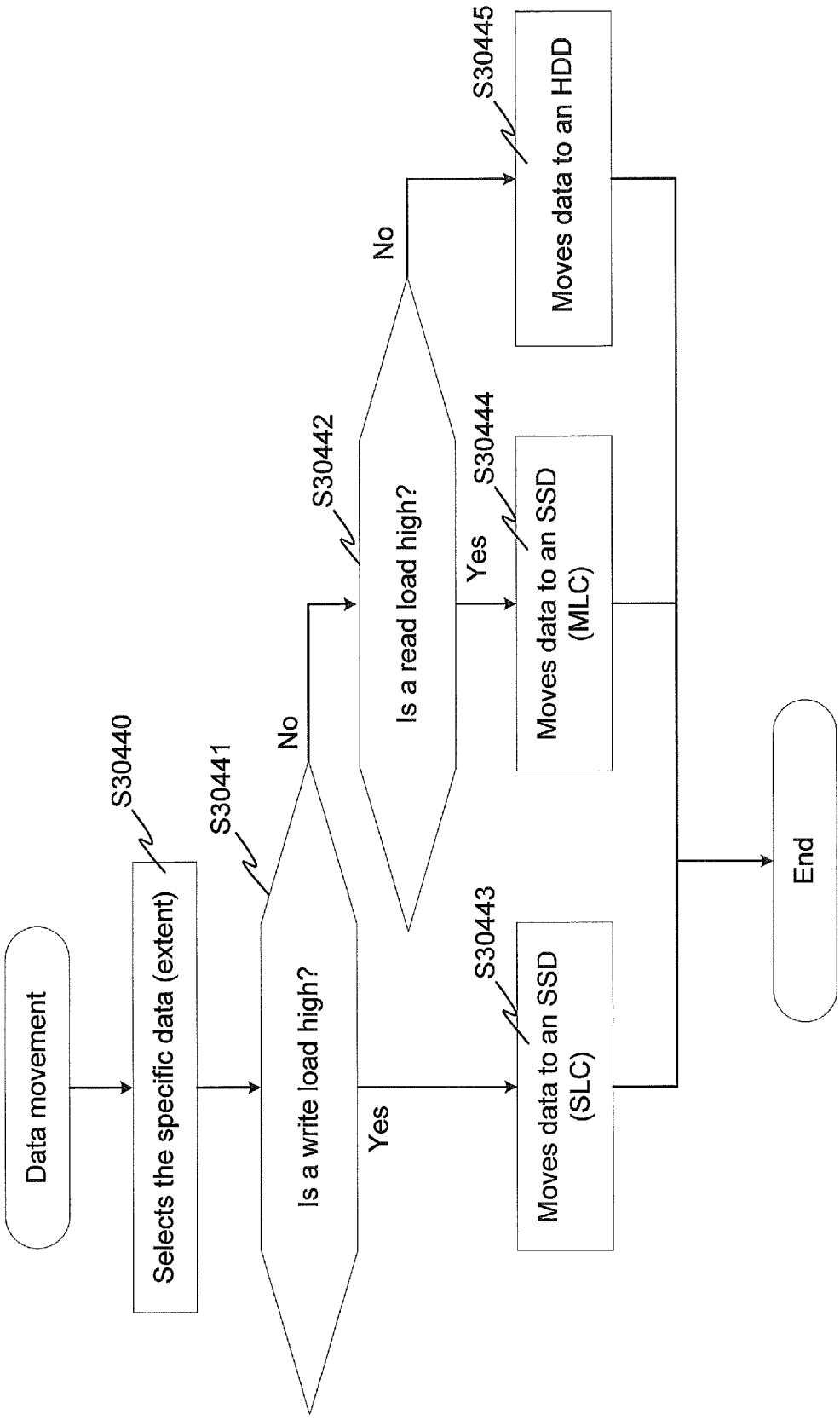
Data movement
S30440
Selects the specific data (extent)
S30441
Is a write load high?
Yes
S30443
Moves data to an SSD (SLC)
No
S30442
Is a read load high?
Yes
S30444
Moves data to an SSD (MLC)
No
S30445
Moves data to an HDD
End

Fig. 31

[Fig. 32]

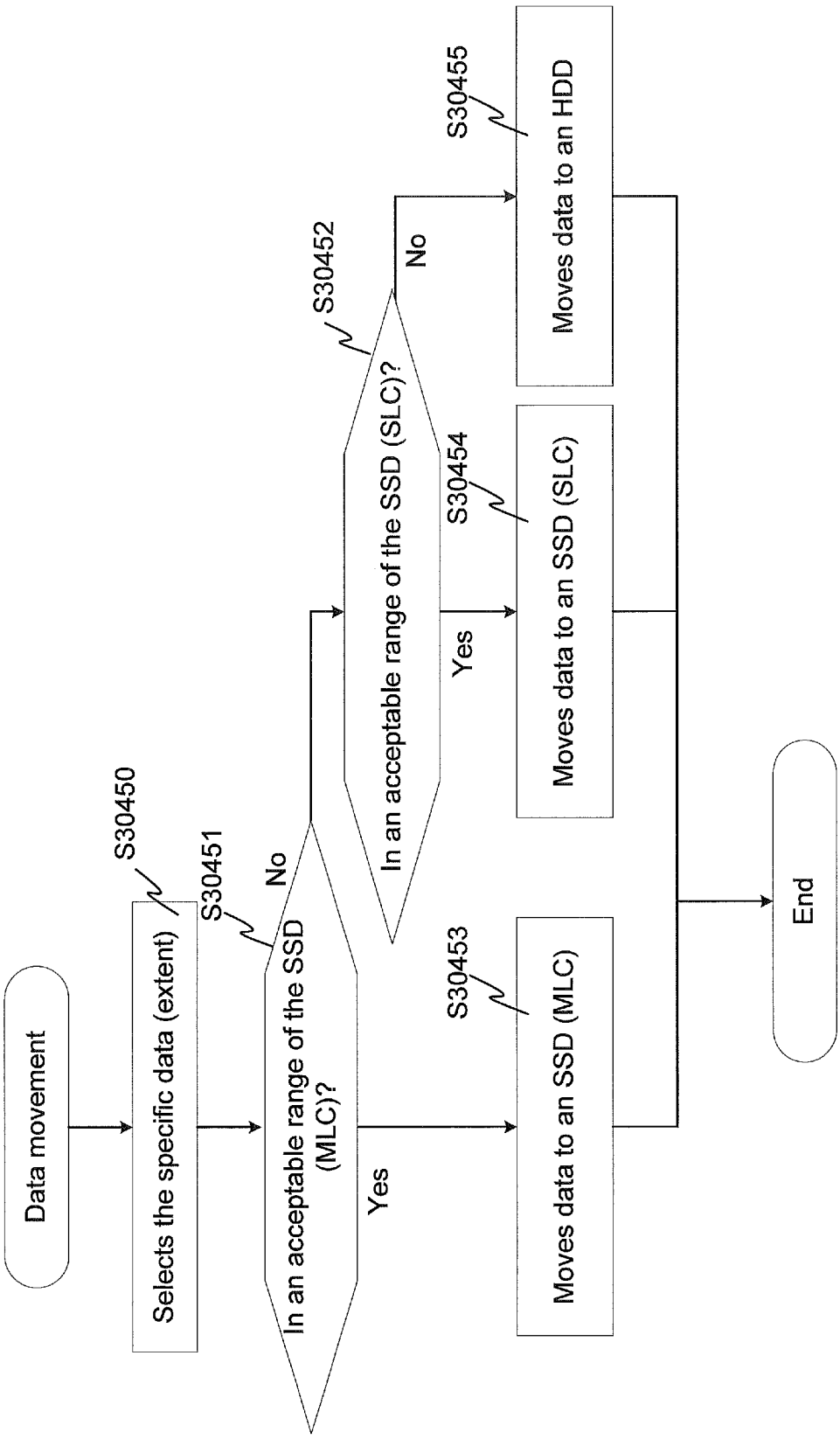
Data movement
Selects the specific data (extent)
S30450
In an acceptable range of the SSD (MLC)?
S30451
No
Yes
In an acceptable range of the SSD (SLC)?
S30452
No
Yes
Moves data to an SSD (MLC)
S30453
Moves data to an SSD (SLC)
S30454
Moves data to an HDD
S30455
End

Fig. 32

[Fig. 33]

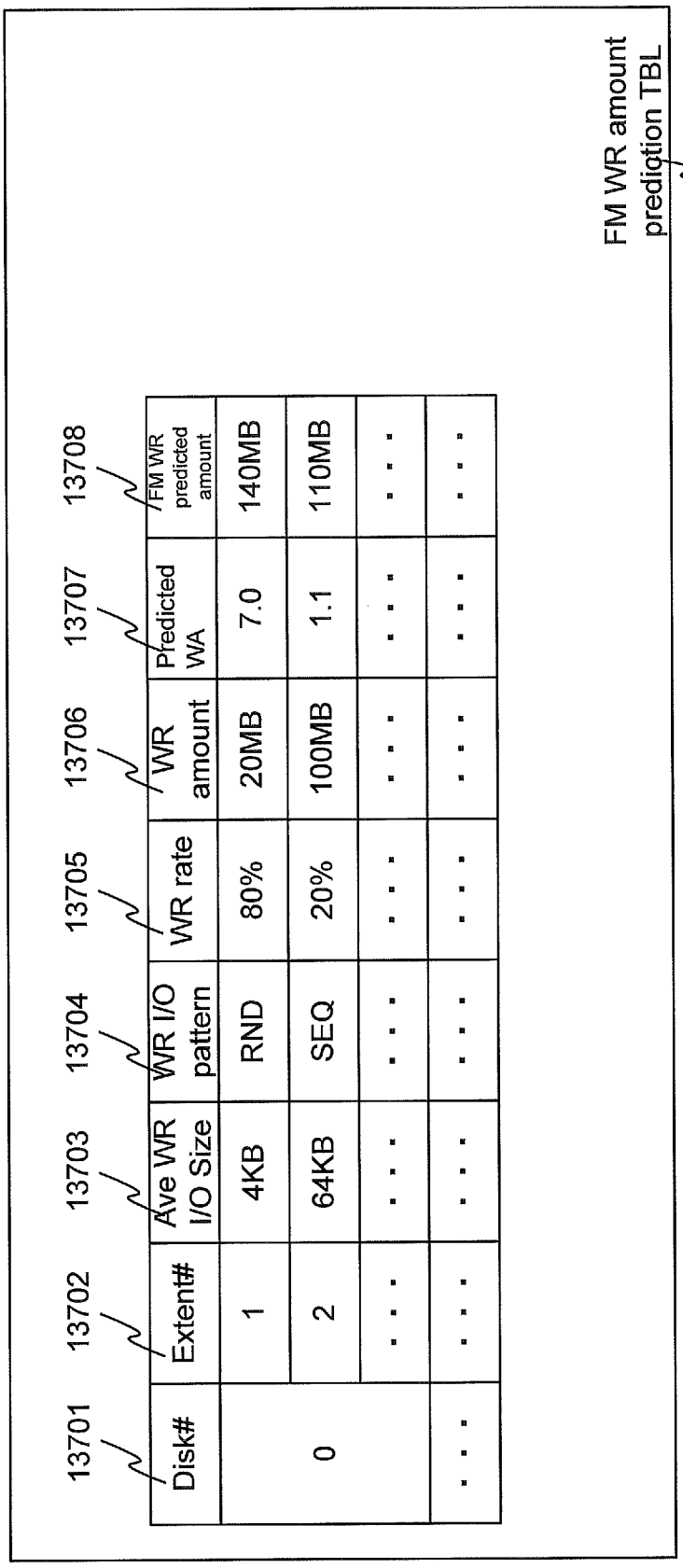
FM WR amount prediction TBL
13700
13701
13702
13703
13704
13705
13706
13707
13708
Disk#
Extent#
Ave WR I/O Size
WR I/O pattern
WR rate
WR amount
Predicted WA
FM WR predicted amount
0
1
4KB
RND
80%
20MB
7.0
140MB
2
64KB
SEQ
20%
100MB
1.1
110MB

Fig. 33

[Fig. 34]

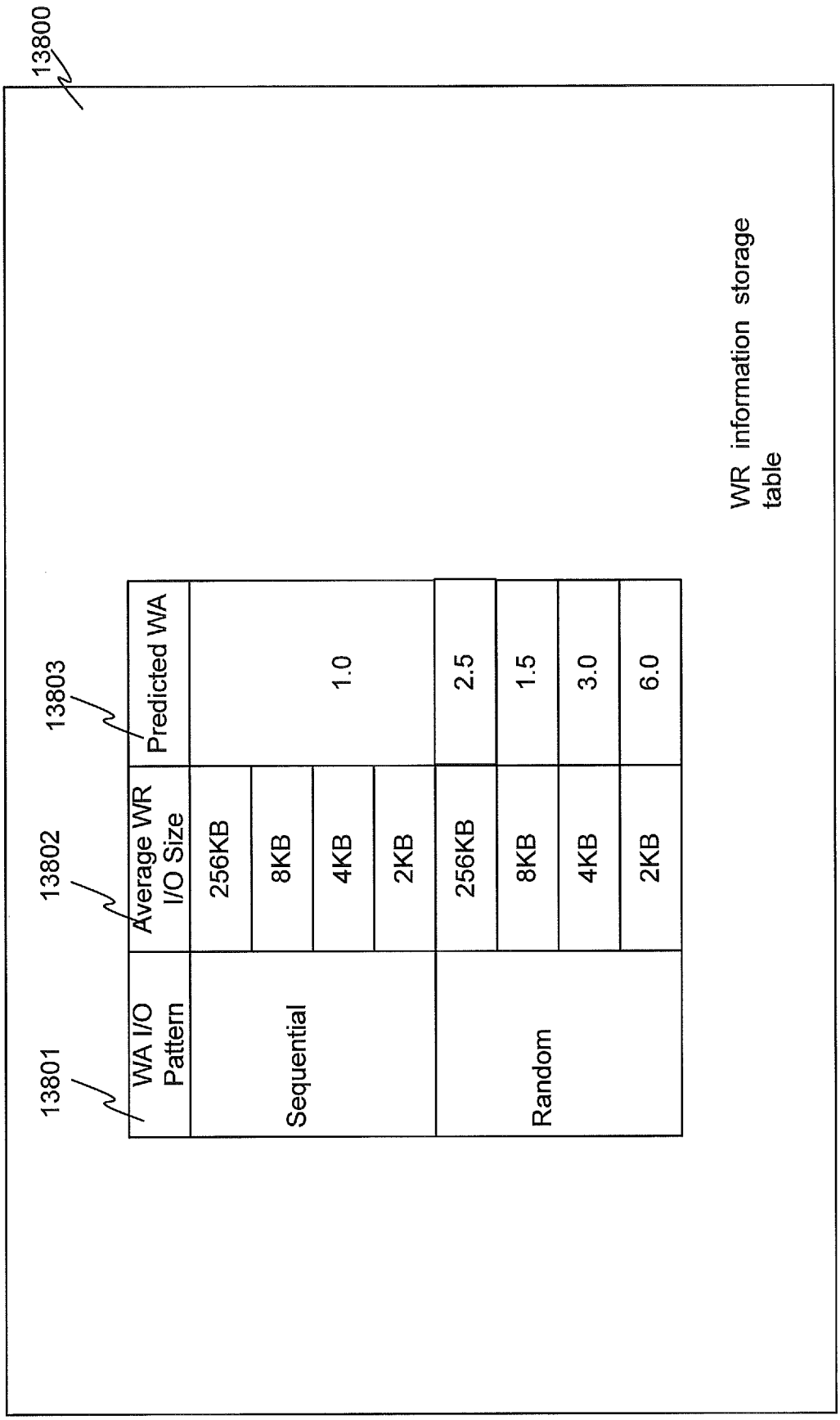
13800
13801
13802
13803
WA I/O Pattern
Average WR I/O Size
Predicted WA
Sequential
256KB
8KB
4KB
2KB
1.0
Random
256KB
8KB
4KB
2KB
2.5
1.5
3.0
6.0
WR information storage table

Fig. 34

[Fig. 35]

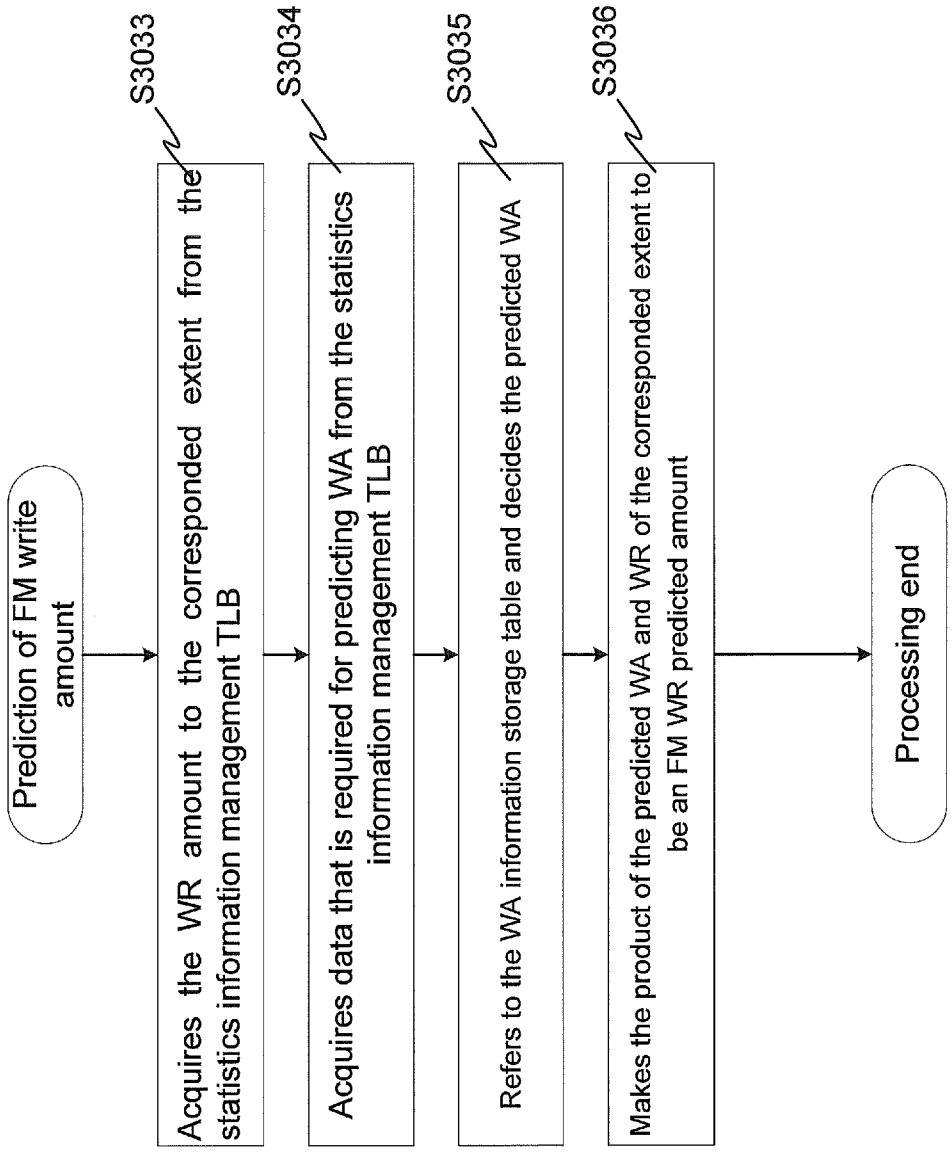
Prediction of FM write amount
S3033
Acquires the WR amount to the corresponded extent from the statistics information management TLB
S3034
Acquires data that is required for predicting WA from the statistics information management TLB
S3035
Refers to the WA information storage table and decides the predicted WA
S3036
Makes the product of the predicted WA and WR of the corresponded extent to be an FM WR predicted amount
Processing end

Fig. 35

[Fig. 36]

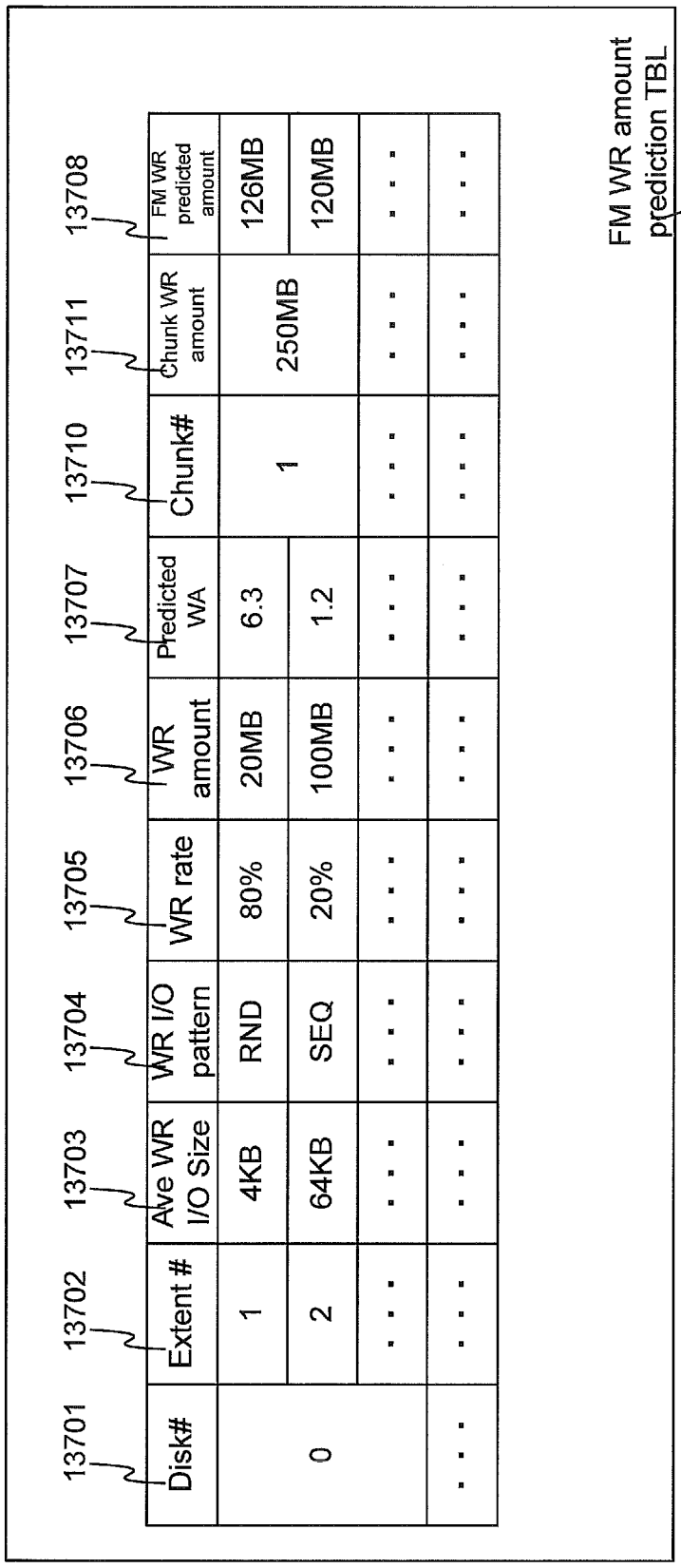
FM WR amount prediction TBL
13700
13701 Disk#
13702 Extent #
13703 Ave WR I/O Size
13704 WR I/O pattern
13705 WR rate
13706 WR amount
13707 Predicted WA
13710 Chunk#
13711 Chunk WR amount
13708 FM WR predicted amount
0
1
4KB
RND
80%
20MB
6.3
1
250MB
126MB
2
64KB
SEQ
20%
100MB
1.2
120MB

Fig. 36

[Fig. 37]

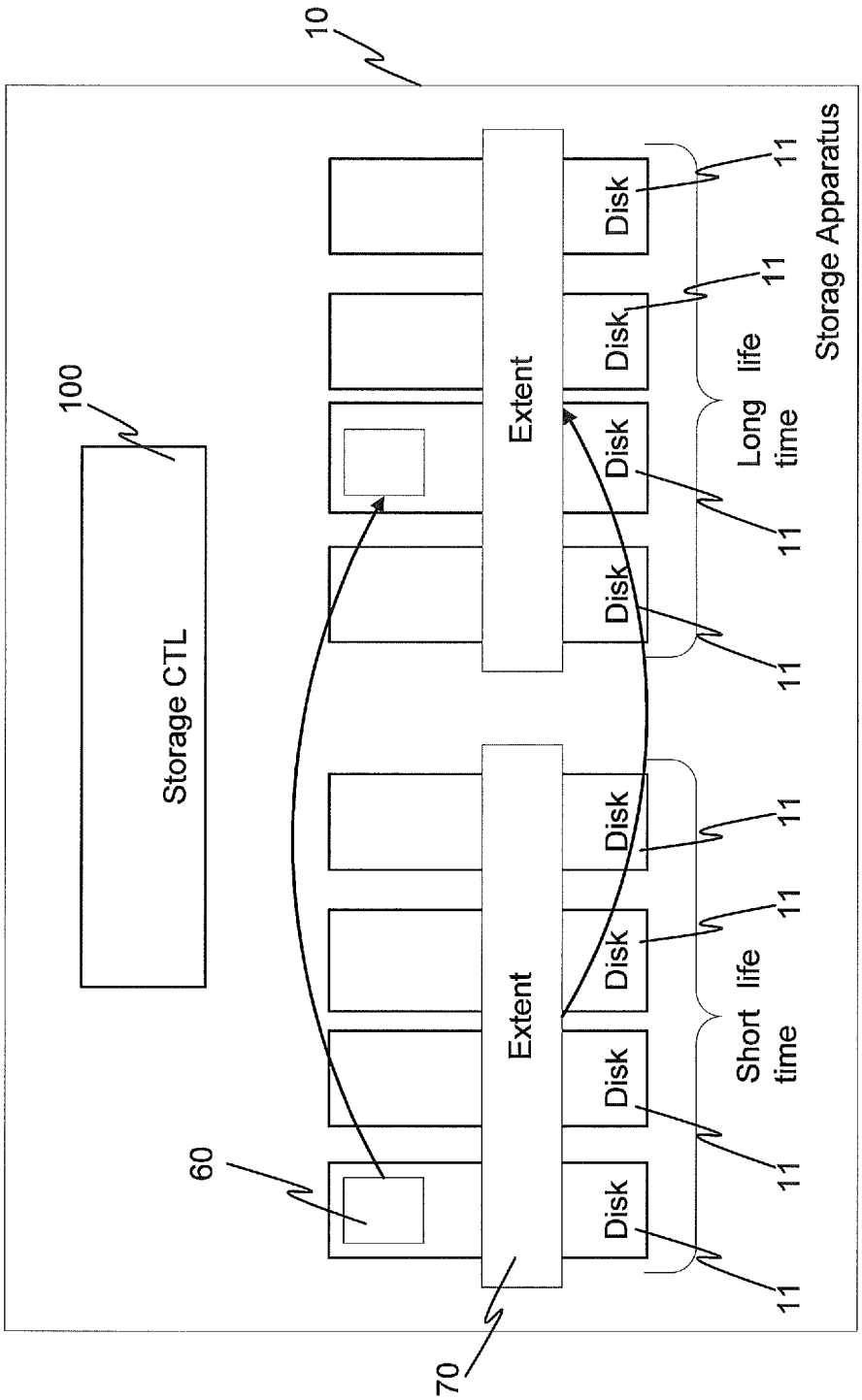
10
Storage Apparatus
100
Storage CTL
60
70
Extent
Extent
Disk
11
Long life time
Short life time

Fig. 37

[Fig. 38]

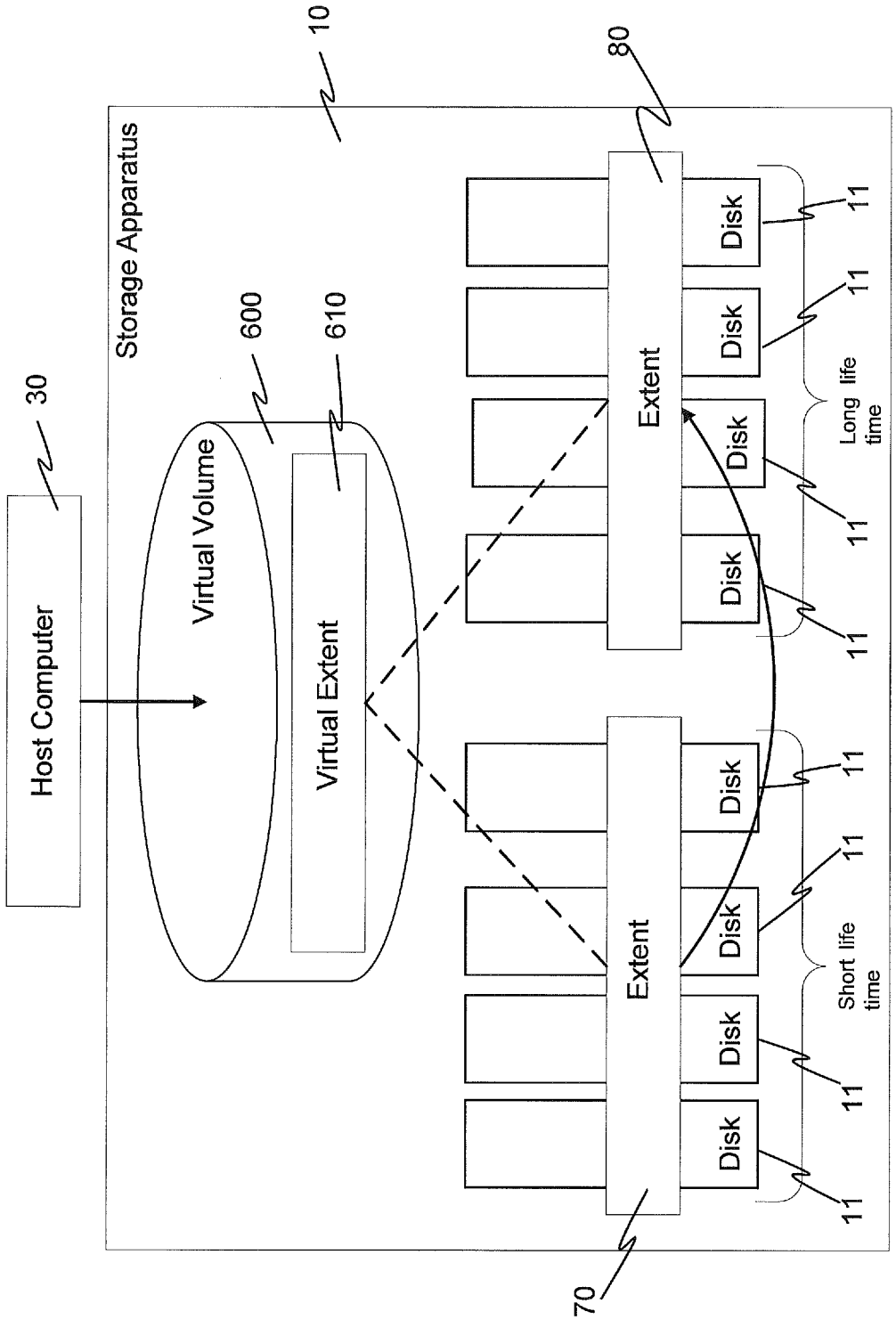
10
80
Storage Apparatus
600
610
30
Host Computer
Virtual Volume
Virtual Extent
Extent
Disk
11
Long life time
Short life time
70

Fig. 38

# INTERNATIONAL SEARCH REPORT

International application No
PCT/JP2012/000843

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F3/06 G06F12/02
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | EP 1 876 519 A2 (HITACHI LTD [JP])<br>9 January 2008 (2008-01-09)<br>abstract; figures 2,8,9,13,17,18 | 1-18 |
| X | GB 2 479 235 A (INTEL CORP [US])<br>5 October 2011 (2011-10-05)<br>claims 2,3,5,10,12,20; figures 3A,3B,3C | 1-18 |
| X | US 5 737 742 A (ACHIWA KYOSUKE [JP] ET AL)<br>7 April 1998 (1998-04-07)<br>abstract; figures 3,4,7,10A,10B,11,13,17, | 1-18 |
| X | EP 1 840 722 A2 (HITACHI LTD [JP])<br>3 October 2007 (2007-10-03)<br>figures 1,6,9,15-19 | 1-18 |

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 17 July 2012 | 25/07/2012 |

| Name and mailing address of the ISA/ | Authorized officer |
|---|---|
| European Patent Office, P.B. 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel. (+31-70) 340-2040,<br>Fax: (+31-70) 340-3016 | Manfrin, Max |

1

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/JP2012/000843

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| EP 1876519 A2 | 09-01-2008 | EP 1876519 A2 | 09-01-2008 |
| | | JP 2008015769 A | 24-01-2008 |
| | | US 2008010398 A1 | 10-01-2008 |
| GB 2479235 A | 05-10-2011 | GB 2479235 A | 05-10-2011 |
| | | KR 20110110720 A | 07-10-2011 |
| | | US 2011246705 A1 | 06-10-2011 |
| US 5737742 A | 07-04-1998 | JP 3507132 B2 | 15-03-2004 |
| | | JP 8016482 A | 19-01-1996 |
| | | US 5737742 A | 07-04-1998 |
| | | US 5930193 A | 27-07-1999 |
| EP 1840722 A2 | 03-10-2007 | CN 101046771 A | 03-10-2007 |
| | | CN 101645041 A | 10-02-2010 |
| | | EP 1840722 A2 | 03-10-2007 |
| | | EP 2365428 A1 | 14-09-2011 |
| | | JP 4863749 B2 | 25-01-2012 |
| | | JP 2007265265 A | 11-10-2007 |
| | | US 2007233931 A1 | 04-10-2007 |
| | | US 2008276038 A1 | 06-11-2008 |
| | | US 2010205359 A1 | 12-08-2010 |
| | | US 2011231600 A1 | 22-09-2011 |