



(51) International Patent Classification:

G11B 7/0037 (2006.01) G06F 12/02 (2006.01)
G11B 7/0055 (2006.01) G06F 12/0891 (2016.01)
G11B 7/006 (2006.01)

(21) International Application Number:

PCT/US2017/013212

(22) International Filing Date:

12 January 2017 (12.01.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

14/994,530 13 January 2016 (13.01.2016) US

(71) Applicant: ALIBABA GROUP HOLDING LIMITED

[—/US]; Fourth Floor, One Capital Place, P.O. Box 847,
George Town, Grand Cayman (KY).

(72) Inventor: LI, Shu; Alibaba Group Legal Department, 5/F,

Building 3, No. 969 West Weh Yi Road, Yu Hang District,
Hangzhou, 311121 (CN).

(74) Agent: MURABITO, Anthony, C.; Murabito Hao &
Barnes, LLP, 2 N. Market St., 3rd Floor, San Jose, CA
95113 (US).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN,
KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA,
MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG,
NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS,
RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY,
TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN,
ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ,
TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,
TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,
DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,
LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,

[Continued on next page]

(54) Title: SYSTEM AND METHOD OF WRITE AMPLIFICATION FACTOR MITIGATION AND FLASH LIFESPAN EXTENSION

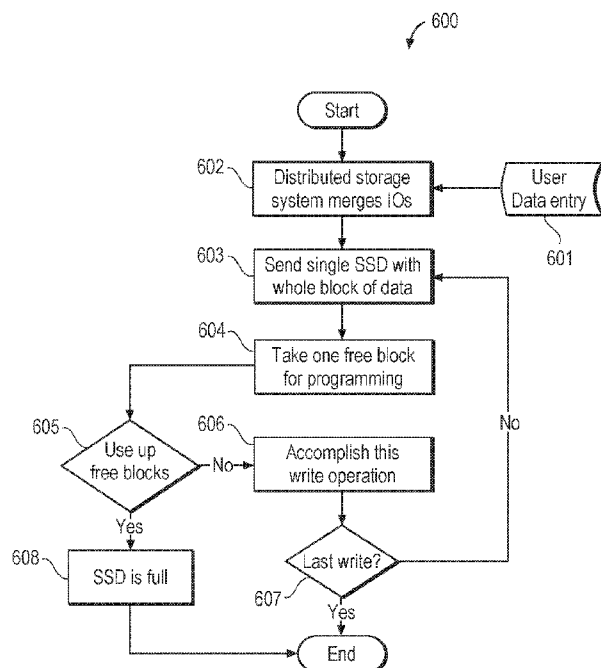


FIG. 6

(57) Abstract: Embodiments of the present invention use a NAND block as the basic write operation unit and ensure that the write operation uses the same basic unit as the erase operation. In this way, the flash product maintains the same level of granularity for read and write operations. The mapping between logical block addressing (LB A) and physical block addressing (PBA) are at the page level. Wear leveling and garbage collection are simplified so the robustness and performance is enhanced. If the data is frequently written, there are no concerns regarding data retention. Embodiments of the present invention evenly distribute hot data using a global optimization perspective based on this observation. When dealing with hot data, the NAND flash's required data retention capability may be adjusted to increase P/E cycles.



SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, —
GW, KM, ML, MR, NE, SN, TD, TG).

*before the expiration of the time limit for amending the
claims and to be republished in the event of receipt of
amendments (Rule 48.2(h))*

Published:

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

SYSTEM AND METHOD OF WRITE AMPLIFICATION FACTOR MITIGATION AND FLASH LIFESPAN EXTENSION

FIELD:

[0001] Embodiments of the present invention generally relate to data storage systems. More specifically, embodiments of the present invention relate to systems and methods for reducing write amplification and extending the lifespan of flash-oriented file systems.

BACKGROUND:

[0002] In general, when any amount of data is written to a NAND flash block, the entire NAND flash block must be erased before new data can be rewritten. In some cases, a flash-based storage device (e.g., a solid state drive (SSD)) is used in a manner similar to traditional hard drives (HDDs), where files are partitioned or merged into logic blocks of 512 bytes, 4k bytes, or more. SSD is typically characterized as high-performance storage device with high throughput, high Input/Output Operations per Second (IOPS), low latency, and low failure rate. An internal index associated with the logic block is referred to as the logical block address (LBA). The blocks are written to specific locations on the storage media, and the address is referred to as the physical block address (PBA). However, some conventional HDD operations, such as defragmentation operations, lead to a degradation of the performance and lifespan of the SSD.

[0003] In a typical case, an SSD receives a write command from a host and stores the associated data on one or more pages of one or more blocks. Initially, all blocks are available to write data and are referred to as free blocks. After data is written to a block, the block may be erased and added to a pool of free blocks. Figure 1 illustrates an exemplary block writing and recycling technique. When a host sends a write request 101 to the SSD, the data is buffered and an available block 102 from a free block pool 103 is selected. Next, the data from the buffer is written to the selected free block. After the data has been written to the selected block, the block is added to a data block pool 104. Some files may be deleted and the corresponding pages are marked as invalid.

[0004] A block 105 with the fewest number of valid pages is selected for garbage collection, and the valid pages in this block are read and written to other blocks. After the block's data is copied and consolidated, the entire block is erased. After the block is erased, the block is considered to have finished one program/erase (P/E) cycle. However, the exemplary block writing and recycling technique depicted in Figure 1 results in write amplification. Copying and rewriting the pages of the selected block takes place internally so these actions are not considered to be writing data from the host. Therefore, from the perspective of the NAND flash side, more data is written than is received from host. This phenomenon is known as write amplification.

[0005] Figure 2 illustrates total P/E cycles for different generations of NAND flash in graph 200. With storage density increasing in subsequent generations, P/E cycles are reduced dramatically. Therefore, each P/E cycle of a NAND flash product becomes more and more important and directly determines the lifespan of the given device. With a fixed P/E cycle budget of 2x nm MLC, for example, each block can be erased 3000 times during the product lifespan. Assuming the average write amplification factor is 3, the maximum amount of data to be written into this NAND product is only 1000 times greater than the capacity of the device. As the total P/E cycles decrease for subsequent generations, the bits required for error correcting code increases. For example, the 3x nm MLC requires an 8-bit ECC and the 2x nm MLC requires a 15-bit ECC.

[0006] All SSDs have a write amplification value that represents the ratio of data written by the host to the SSD compared to the amount of data that is actually written to the SSD. Several factors may increase the write amplification value, including techniques that mitigate read and/or write disturbances and wear-leveling policies that move user data from aged segments into clean segments. Garbage collection policies may further increase write amplification. What is needed is an SSD device that efficiently handles sub-optimal usage (e.g., defragmentation operations), mitigates write amplification, and extends the lifespan of the device, all while keeping device-related costs low.

SUMMARY:

[0007] Methods and systems for managing data storage in flash memory devices are described herein. By using a distributed storage system to merge small files and write data block-by-block, the write amplification factor approaches an ideal value of 1.

[0008] According to one embodiment, a method of distributing data among a plurality of solid state drives to mitigate write amplification is described. The method includes receiving a write request comprising write data, determining a portion of the write data comprising hot data, dividing the hot data into a plurality of stripes, writing each of the plurality of stripes to a different solid state drive such that the hot data is relatively evenly distributed among the plurality of solid state drives.

[0009] According to another embodiment, an apparatus for distributing data among a plurality of solid state drives to mitigate write amplification is disclosed. The apparatus includes a load balancer configured to receive and direct data requests, wherein the data requests comprise write data, a distributed storage system configured to store data, and a plurality of solid state drives coupled to the distributed storage system. The distributed storage system directs the storage of data on the solid state drives such that frequently updated data is relatively evenly distributed amongst the solid state drives.

BRIEF DESCRIPTION OF THE DRAWINGS:

[0010] The accompanying drawings, which are incorporated in and form a part of this specification, illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention:

[0011] Figure 1 is a block diagram illustrating an exemplary block writing and recycling technique for an exemplary SSD.

[0012] Figure 2 is a graph illustrating the total P/E cycles for different generations of NAND flash.

[0013] Figure 3 is a graph illustrating exemplary write amplification factors compared to an overprovisioning percentage for an exemplary SSD according to embodiments of the present invention.

[0014] Figure 4 is a block diagram illustrating an exemplary cloud service system comprising a distributed storage system according to embodiments of the present invention.

[0015] Figure 5 is a block diagram illustrating an exemplary distributed storage system configured as a file blender according to embodiments of the present invention.

[0016] Figure 6 is a flow chart depicting an exemplary sequence of computer implemented steps for performing block-wise programming and erasing for hot data in a multi-SSD storage system according to embodiments of the present invention.

[0017] Figure 7 is a graph illustrating exemplary retention times needed compared to a percentage of maximum cycles according to embodiments of the present invention.

DETAILED DESCRIPTION:

[0018] Reference will now be made in detail to several embodiments. While the subject matter will be described in conjunction with the alternative embodiments, it will be understood that they are not intended to limit the claimed subject matter to these embodiments. On the contrary, the claimed subject matter is intended to cover alternative, modifications, and equivalents, which may be included within the spirit and scope of the claimed subject matter as defined by the appended claims.

[0019] Furthermore, in the following detailed description, numerous specific details are set forth in order to provide a thorough understanding of the claimed subject matter. However, it will be recognized by one skilled in the art that embodiments may be practised without these specific details or with equivalents thereof. In other instances, well-known methods, procedures, components, and circuits have not been described in detail as not to unnecessarily obscure aspects and features of the subject matter.

[0020] Portions of the detailed description that follows are presented and discussed in terms of a method. Although steps and sequencing thereof are disclosed in a figure herein describing the operations of this method, such steps and sequencing are exemplary. Embodiments are well suited to performing various other steps or variations of the steps recited in the flowchart (e.g., Figure 6) of the figures herein, and in a sequence other than that depicted and described herein.

[0021] Some portions of the detailed description are presented in terms of procedures, steps, logic blocks, processing, and other symbolic representations of operations on data bits that can be performed on computer memory. These descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. A procedure, computer-executed step, logic block, process, etc., is here, and generally, conceived to be a self-consistent sequence of steps or instructions leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated in a computer system. It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like.

[0022] It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussions, it is appreciated that throughout, discussions utilizing terms such as “accessing,” “writing,” “including,” “storing,” “transmitting,” “traversing,” “associating,” “identifying” or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission or display devices.

[0023] The following description is presented to enable a person skilled in the art to make and use the embodiments of this invention; it is presented in the context of a particular application and its requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present disclosure. Thus, the present invention is not limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

[0024] Embodiments of the present invention use a NAND block as the basic write operation unit and ensure that the write operation uses the same basic unit as the erase operation. In this way, the flash product maintains the same level of granularity for read and write operations. The mapping between logical block addressing (LBA) and physical block addressing (PBA) are at the page level. Wear leveling and garbage collection are simplified so the robustness and performance is enhanced. If the data is frequently written, there are no concerns regarding data retention. When dealing with hot data, the NAND flash's required data retention capability may be adjusted to increase P/E cycles. By using a distributed storage system to merge small files and write data block-by-block, the write amplification factor approaches an ideal value of 1.

[0025] Flash devices such as SSDs experience increasing write amplification mainly due to small chunks of data updates or deletions. One primary reason for write amplification is the mismatch between the programming unit and the erasing unit. After some pages are updated or deleted, the original pages become invalid. However, all pages in the same block do not become invalid at the same time. Therefore, when an SSD runs out of free blocks, the blocks with the fewest valid pages are chosen for erasing.

[0026] With regard to Figure 3, graph 300 illustrates write amplification factors compared to an overprovisioning percentage for an exemplary SSD according to embodiments of the present invention. Each plot on the graph represents a data group having a different percentage of "hot" data. Hot data is data that is updated frequently. Data that is rarely updated after initially being written is considered "cold" data. Plot 301 represents a data group comprising 10% hot data, plot 302 represents a data group comprising 50% hot data, and plot 303 represents a data group having 100% hot data. In general, graph 300 illustrates that reducing the amount of hot data in a data group effectively mitigates the write amplification factor. Embodiments of the present invention evenly distribute hot data using a global optimization perspective based on this observation.

[0027] With regard to Figure 4, an exemplary cloud service system 400 is depicted according to embodiments of the present invention. Cloud services system 400 may comprise one or more storage systems (e.g., distributed storage system 405) accessed by terminal device 401. Terminal device 401 may comprise a personal computer, tablet, smartphone, wearable device, or any other

device operable to exchange data with a network or storage device. Load balancer 403 receives data requests and orchestrates traffic passing through Firewall 402 so that cloud servers 404a – 404c receive user data and process the data accordingly. Data is stored using the distributed storage system 405 comprising multiple storage servers 406a – 406c. The storage servers 406a – 406c may comprise a single SSD, or a system having an array of SSDs, for example. Storage virtualization may be used so that a user may interact with the cloud services system in the same way as a traditional computer. Behind the virtualization layer, the actual resources are manipulated based on global optimization techniques, where the time difference, region variance, and physical server performance deviations are masked. Consequently, the cloud services system delivers a stable and efficient solution for exchanging large amounts of data.

[0028] According to some embodiments of the present invention, a distributed storage system (e.g., distributed storage system 405) that acts as a file blender to merge incoming data and subsequently spread the data among multiple destinations is used to evenly distribute hot data among the SSDs. The distributed storage system may comprise one or more processors (e.g., CPU 405a) for analysing and directing data to the SSDs, and RAM 405b for storing data. Considering the results depicted in Figure 2 regarding the relationship between hot data and write amplification, one exemplary solution comprises dividing the files with hot data into stripes and distributing the strips of hot data to several (e.g., ten or more) SSDs to avoid a high frequency of hot data access for any particular SSD. This solution utilizes global optimization techniques and makes each SSD's hot data percentage approach the average percentage of hot data. Therefore, local or temporary hot data will be handled efficiently. Because these file are stored in a multiple stripe format across multiple SSDs, the files may be accessed in parallel which leads to improved IOPS and throughput performance.

[0029] With regard to Figure 5, an exemplary distributed storage system 500 configured as a file blender is depicted according to embodiments of the present invention. For example, File 501, File 502, File 503, and File 504 are four files with different sizes and hotness. If it is determined that File 502 is updated frequently, rather than writing B to one SSD, file B may be written to SSDs 506 – 509 as stripes. For example, stripe 502a may be written to SSD 506, stripe 502b to SSD 507, stripe 502c to SSD 508, and stripe 502d to SSD 509. There may be more stripes written to to more SSDs (not pictured). The stripe size or the data amount written from file B may vary between SSDs, and

the stripe size can be adjusted based on individual SSDs capacity and performance. For one physical SSD, only a stripe of hot data from file B is written. Another file, File 503 is also distributed as stripes to SSDs 506 - 509. Stripe 503a is written to SSD 506, stripe 503b is written to SSD 507, stripe 503c is written to SSD 508, and stripe 503d is distributed to SSD 509. The stripe size may vary between different SSDs. Files 502 and 504 may also be distributed to the SSDs as stripes in a similar manner. This technique promotes balanced system wear and reduce the likelihood of single-point failure. Furthermore, reliability is enhanced and operation costs are reduced.

[0030] As mentioned above, the stripe size written to each SSD may vary. Besides global optimization based on the distribution of hot data, embodiments of the present invention use NAND flash block-wise operation to control write amplification. For example, distributed storage system 505 may be used as a data pool to buffer, blend, and merge data to be stored on SSDs 506 - 509. Therefore, utilizing this coherent middle layer, distributed storage system can merge small inputs/outputs (IOs) and increase the block size. As mentioned previously, one root cause of write amplification is the mismatch between the erase operation unit (e.g., block) and the program operation unit (e.g., page). When IOs are merged and written block-by-block, invalid/valid pages are no longer a concern. As a result, an entire block is written or erased at a time, and garbage collection is significantly simplified such that no valid page will be copied from the block to be erased and re-written elsewhere.

[0031] With regard to Figure 6, an exemplary sequence of computer implemented steps 600 for performing block-wise programming and erasing for hot data in a multi-SSD storage system is depicted according to embodiments of the present invention. At step 601, a user data entry is received by a distributed storage system. At step 602, the distributed storage system merges IOs to form a block size of one NAND flash block. At step 603, a whole block of data is sent to a single SSD. At step 604, a flash controller takes one non-empty free block from the free block pool. The entire block is programmed sequentially page-by-page.

[0032] A deletion operation also uses an entire block as the basic unit. Consequently, the garbage collection becomes simpler because if one data block is deleted, the block will be returned to free

block pool for future write operations. At step 605 it is determined if all free blocks have been used. If so, the SSD is determined to be full at step 608 and the process ends. If there are free blocks remaining at step 605, the process continues to step 606 where the data is written to a free block. At step 607 it is determined if there is additional data to be written. If so, the process returns to step 603 and continues. Otherwise, if it is determined that there is no further data to be written at step 607, the process ends.

[0033] With regard to Figure 7, an exemplary graph 700 illustrates retention time needed compared to a percentage of maximum cycles. Graph 700 indicates that by targetting different applications, the retention time can be adjusted to improve the total number of P/E cycles. Based on online data collected from data center, the characteristics of data is extracted to determine the maximum data retention required, then the NAND flash is adjusted accordingly using a new configuration, and the maximal P/E cycles of this NAND flash is increased. At the same time, when the lifespan is unchanged, the increased number of P/E cycles means the write amplification effect is mitigated.

[0034] Embodiments of the present invention are thus described. While the present invention has been described in particular embodiments, it should be appreciated that the present invention should not be construed as limited by such embodiments, but rather construed according to the following claims.

CLAIMS:

What is claimed is:

1. A method of distributing data among a plurality of solid state drives to mitigate write amplification, comprising:
 - receiving a write request comprising write data;
 - determining a portion of the write data comprising hot data;
 - dividing the hot data into a plurality of stripes; and
 - writing each of the plurality of stripes to a different solid state drive.
2. The method of Claim 1, further comprising writing a remainder of the write data, if any, to a free block of one of the plurality of solid state drives.
3. The method of Claim 1, wherein the plurality of stripes are operable to be accessed in parallel.
4. The method of Claim 1, wherein a stripe size of each of the plurality of stripes is adjusted based on a capacity and/or a performance metric of the associated solid state drive.
5. The method of Claim 1, wherein each of the plurality of stripes is merged with another set of stripes before writing and a size of the merged stripes equals a block size.
6. The method of Claim 5, wherein the block size is 4 MB.
7. The method of Claim 5, wherein the block size is 8 MB.
8. An apparatus for distributing data among a plurality of solid state drives to mitigate write amplification, comprising:
 - a load balancer configured to receive and direct data requests, wherein the data requests comprise write data;
 - a distributed storage system coupled to the load balancer configured to store data; and
 - a plurality of solid state drives coupled to the distributed storage system for storage,

wherein the distributed storage system receives data requests from the load balancer and separates any hot data into a plurality of stripes, wherein each of the plurality of stripes is merged with another set of stripes, wherein a size of the merged stripes equals a block size.

9. The apparatus of Claim 8, wherein the block size is 4 MB.
10. The apparatus of Claim 8, wherein the block size is 8 MB.
11. The apparatus of Claim 8, wherein the distributed storage system divides hot data into a plurality of stripes and writes each of the plurality of stripes to a different solid state drives.
12. The apparatus of Claim 11, wherein each of the plurality of stripes is merged with another set of stripes before writing and a size of the merged stripes equals one block.
13. The apparatus of Claim 11, wherein a stripe size of each of the plurality of stripes is adjusted based on a capacity and/or a performance metric of the associated solid state drive.
14. A computer program product tangibly embodied in a computer-readable storage device and comprising instructions that when executed by a processor perform a method for distributing data among a plurality of solid state drives to mitigate write amplification, the method comprising:
 - receiving a write request comprising write data;
 - determining a portion of the write data comprising hot data;
 - dividing the hot data into a plurality of stripes; and
 - writing each of the plurality of stripes to a different solid state drive.
15. The method of Claim 14, further comprising writing a remainder of the write data, if any, to a free block of one of the plurality of solid state drives.
16. The method of Claim 14, wherein the plurality of stripes are operable to be accessed in parallel.

17. The method of Claim 14, wherein a stripe size of each of the plurality of stripes is adjusted based on a capacity and/or a performance metric of the associated solid state drive.
18. The method of Claim 14, wherein each of the plurality of stripes is merged with another set of stripes before writing and a size of the merged stripes equals a block size.
19. The method of Claim 18, wherein the block size is 4 MB.
20. The method of Claim 18, wherein the block size is 8 MB.

1/7

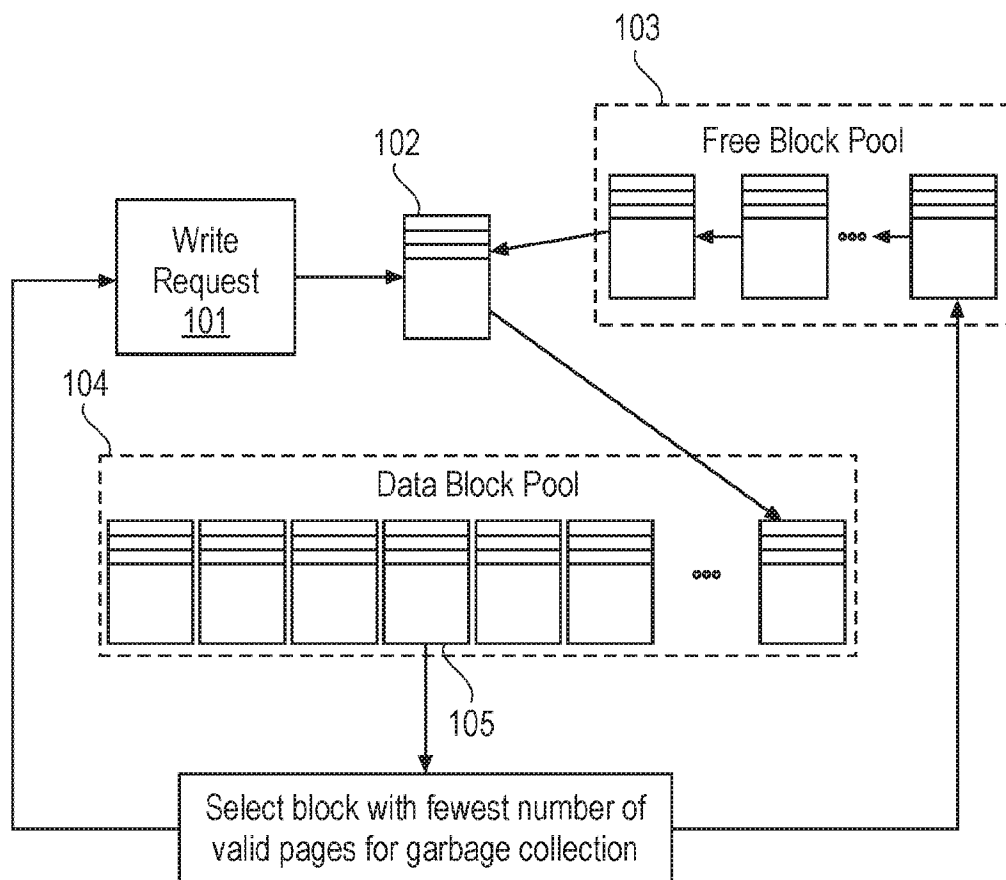


FIG. 1

2/7

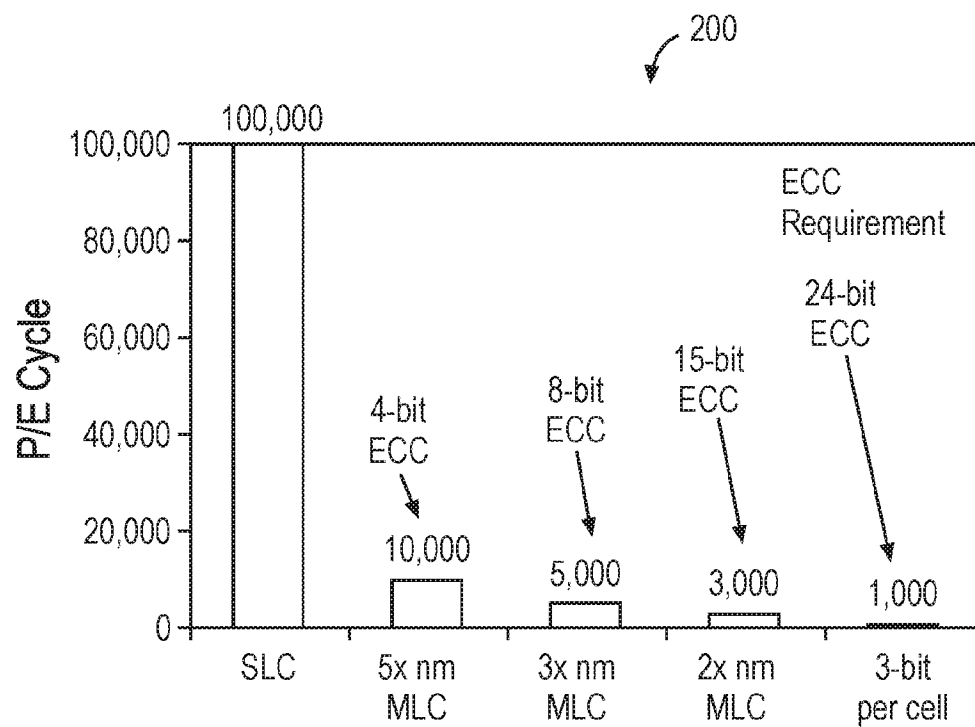


FIG. 2

3/7

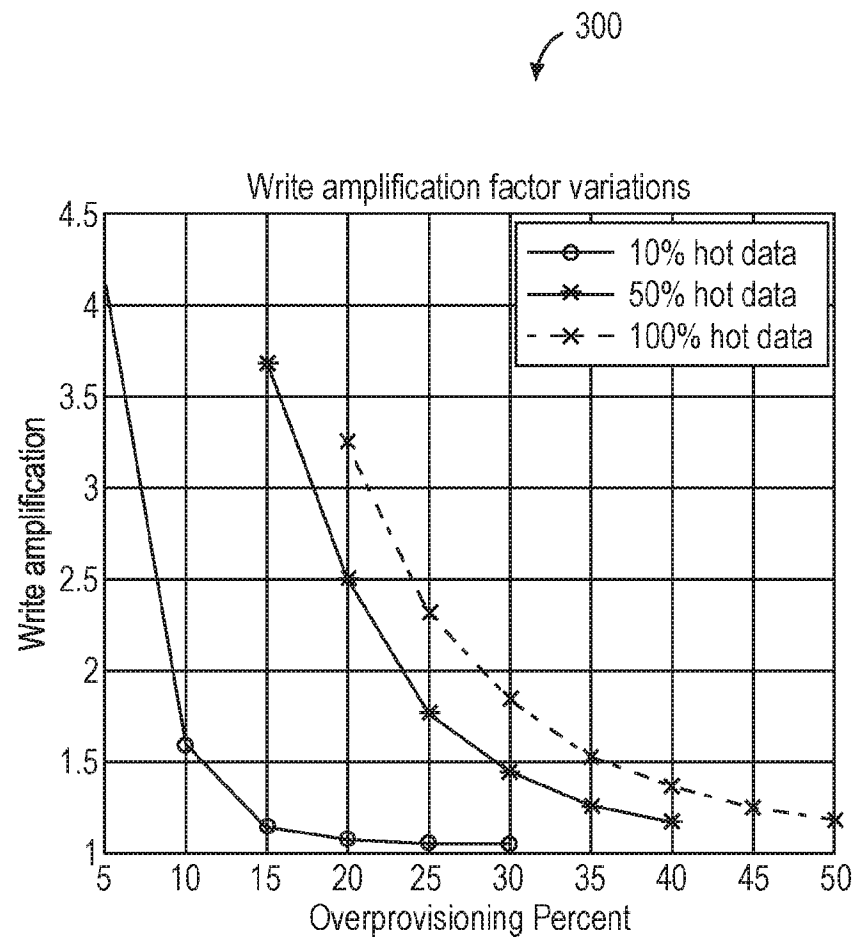


FIG. 3

4/7

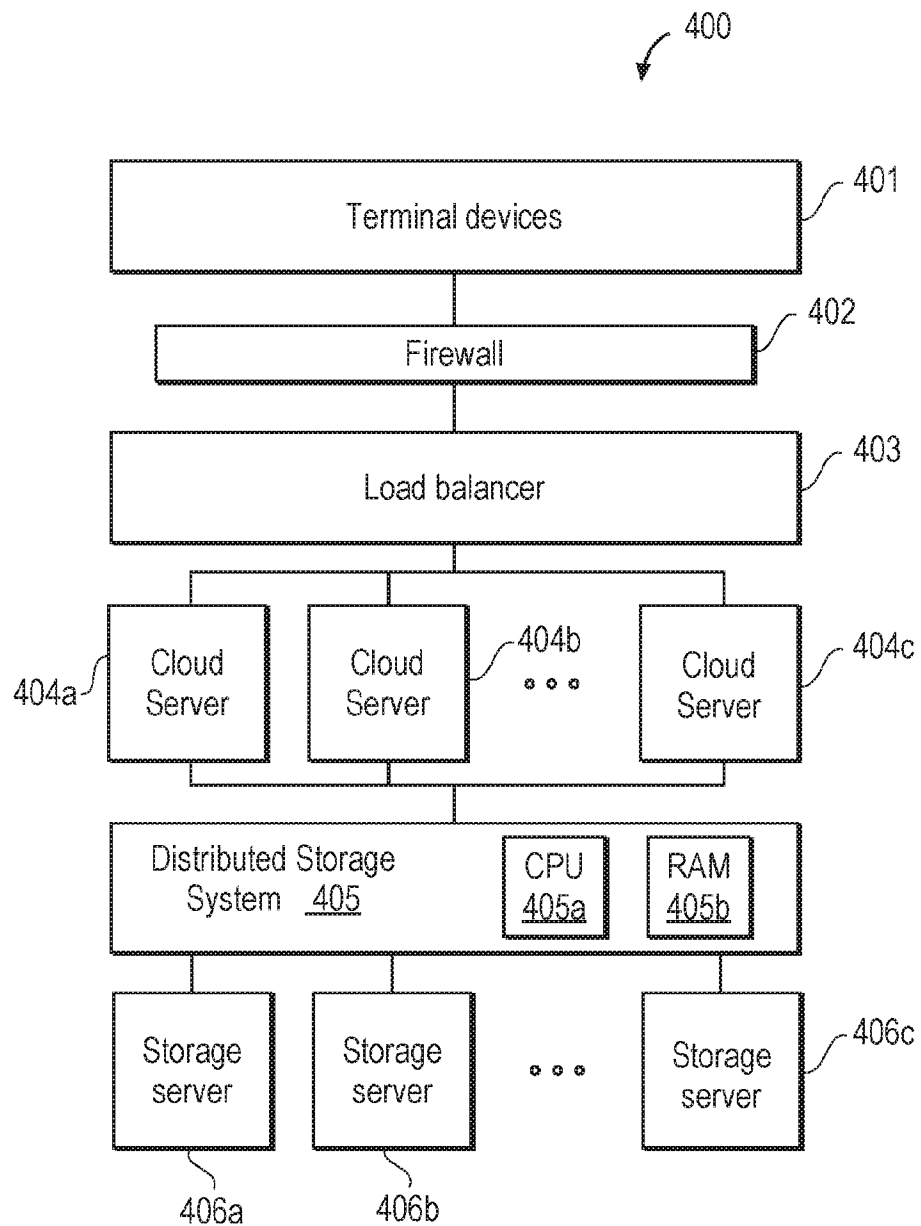


FIG. 4

5/7

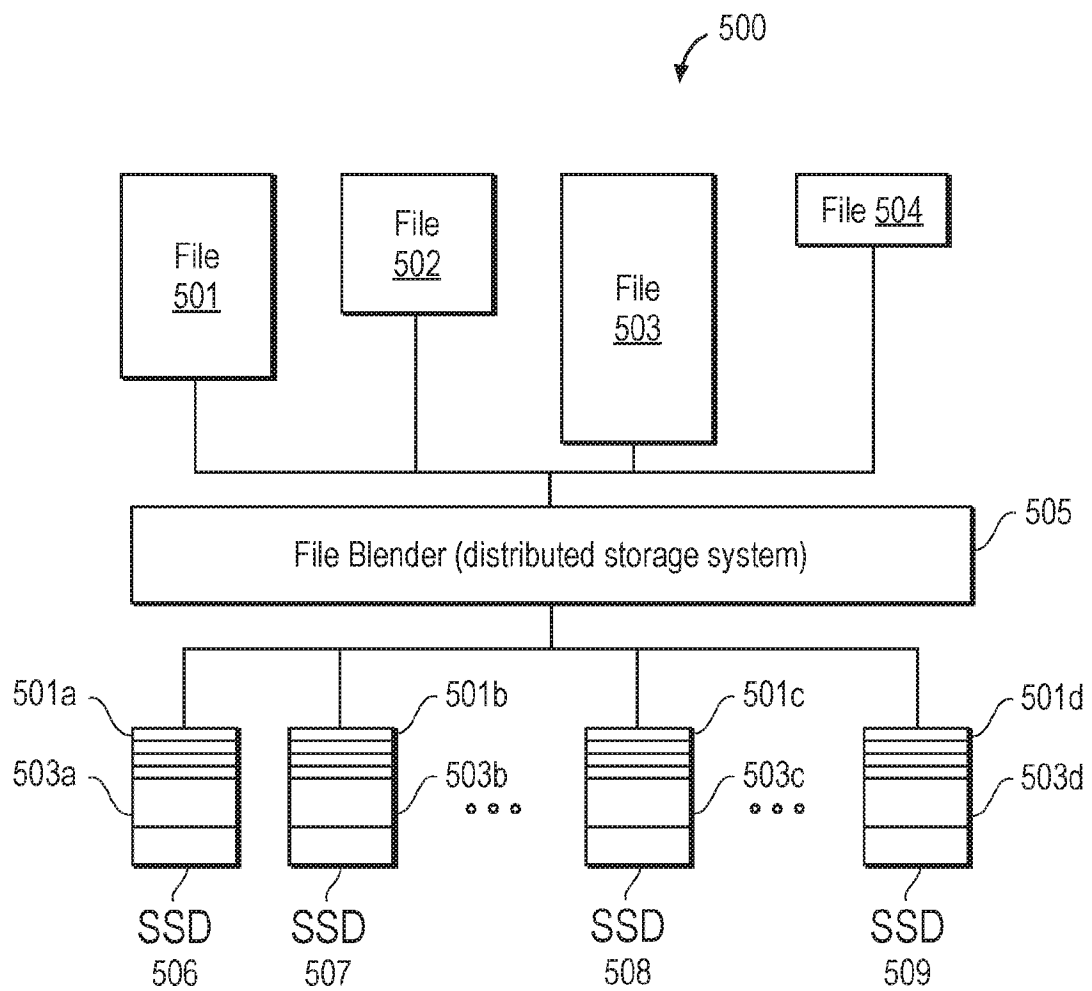


FIG. 5

6/7

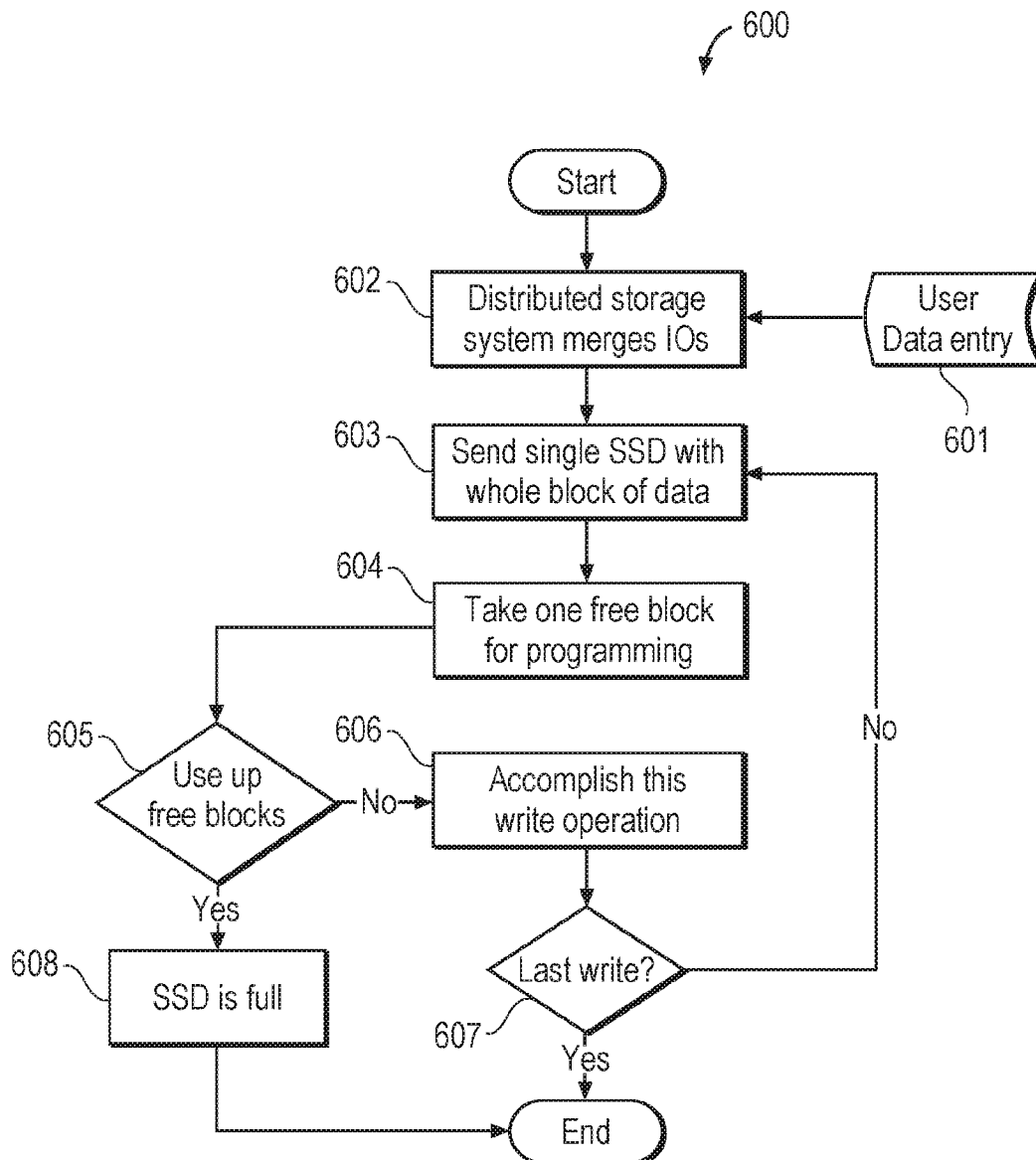


FIG. 6

7/7

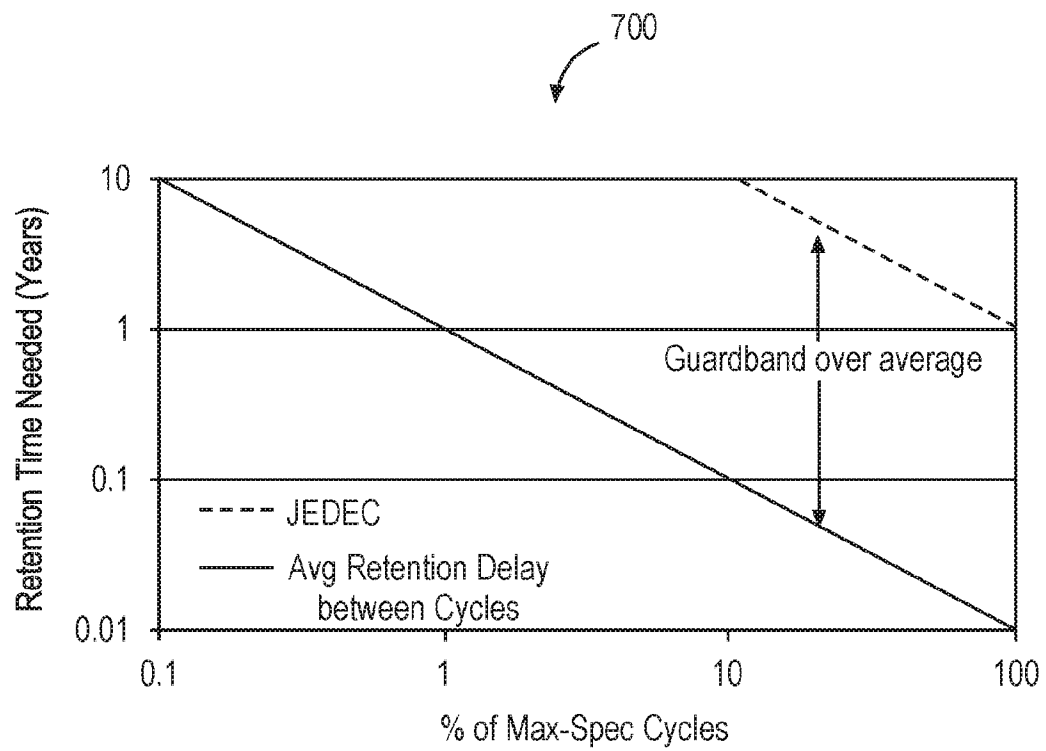


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US17/13212

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G11B 7/0037, 7/0055, 7/006, G06F 12/02, 12/0891 (2017.01)

CPC - G06F 3/0614, 3/06222, 3/0641, 3/0646, 3/0652, 3/0689, 3/0608

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)

See Search History document

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

See Search History document

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 8,363,519 B2 (LUBBERS, C et al.) January 29, 2013; figure 3, abstract, column 2, lines 50-54, lines 36-42, column 3, line 65-column 4, line 5, column 4, lines 46-52, column 5, lines 27-34, column 8, lines 56-65, claims 1, 11, 12	1, 2, 5, 14, 15, 18
--		--
Y	US 5,666,511 A (SUGANUMA, T et al.) September 9, 1997; column 2, lines 34-50	3, 4, 6, 7, 16, 17, 19, 20
Y	US 2015/0277802 A1 (AMAZON TECHNOLOGIES INC.) October 1, 2015; claim 6	3, 16
Y	US 2015/0277802 A1 (AMAZON TECHNOLOGIES INC.) October 1, 2015; claim 6	4, 17
Y	CN 104461914 A (LANGCHAO ELECTRONIC INFORMATION) March 25, 2015; see machine translation; page 6, paragraph 5	6, 7, 19, 20

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

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

"E" earlier 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

07 March 2017 (07.03.2017)

Date of mailing of the international search report

26 MAY 2017

Name and mailing address of the ISA/

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Shane Thomas

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US17/13212

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

-***-Continued within extra sheet-***-

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-7 and 14-20

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US17/13212

-***-Continued from Box No. III - Observations where unity of invention is lacking-***-

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fee must be paid.

Group I: Claims 1-7 and 14-20 are directed towards a method of writing to different solid state drives.

Group II: Claims 8-13 are directed towards an apparatus comprising a load balancer.

The inventions listed as Groups I-II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The special technical features of Group I include at least writing each of the plurality of stripes to a different solid state drive, which are not present in Group II.

The special technical features of Group II include at least a distributed storage system receiving data requests from a load balancer; and wherein each of the plurality of stripes is merged with another set of stripes, wherein a size of the merged stripes equals a block size, which are not present in Group I.

The common technical features shared by Groups I-II are receiving a write request comprising write data; determining a portion of the write data comprising hot data; and dividing the hot data into a plurality of stripes.

However, these common features are previously disclosed by US 8,363,519 B2 to LUBBERS, C et al. (hereinafter "Lubbers"). Lubbers discloses receiving a write request comprising write data (user can access (receiving a request) to write data; column 2, lines 50-54, column 5, lines 27-34); determining a portion of the write data comprising hot data (designate (determining) a selected data set (hot data) to store (write) to the hot data zone; claim 11); and dividing the hot data into a plurality of stripes (selected data set (hot data) characterized as a RAID strip is written to a different (dividing) one of portions of memory in the hot data zone; claim 12).

Since the common technical features are previously disclosed by the Lubbers reference, these common features are not special and so Groups I-II lack unity.