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

(19) **World Intellectual Property Organization**
International Bureau

(43) **International Publication Date**
27 August 2020 (27.08.2020)

WIPO | PCT

(10) **International Publication Number**
WO 2020/168516 A1

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

(21) **International Application Number:**
PCT/CN2019/075713

(22) **International Filing Date:**
21 February 2019 (21.02.2019)

(25) **Filing Language:** English

(26) **Publication Language:** English

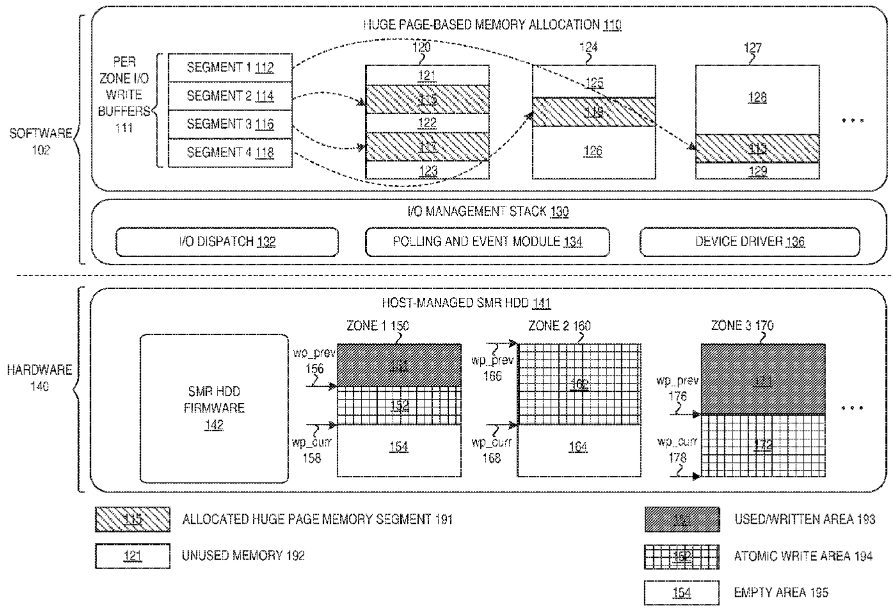
(71) **Applicant: ALIBABA GROUP HOLDING LIMITED** [—/CN]; Fourth Floor, One Capital Place, P.O. Box 847, George Town, Grand Cayman (KY).

(72) **Inventors: HUO, Jianjian**; 16194 Lilac Ln, Los Gatos, California 95032 (US). **CHANG, Cunyin**; Building 1, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, Zhejiang 311121 (CN). **ZHU, Zongpeng**; Building 1, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, Zhejiang 311121 (CN). **LIU, Fei**; 1447 Capuchino Ave, Burlingame, California 94010 (US). **LI, Shu**; 19619 25th Dr. SE, Bothell, Washington 98012 (US).

(74) **Agent: BEIJING TSINGYUANHUI INTELLECTUAL PROPERTY LAW FIRM**; Room 362, 3rd Floor, Suzhou Street No. 1, Haidian District, Beijing 100080 (CN).

(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, JO, 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,

(54) **Title:** METHOD AND SYSTEM FOR FACILITATING FAST ATOMIC WRITE OPERATIONS IN SHINGLED MAGNETIC RECORDING HARD DISK DRIVES



100

FIG. 1

(57) **Abstract:** A method and system for facilitating fast atomic write operations. The system receives a request to write data to a shingled magnetic recording drive (302). The system selects a zone of the drive to which to write the data, wherein the drive stores a previous write pointer indicating a location of an end of a most recent successful write operation in a respective zone (304). The system writes the data at a location of a current write pointer of the selected zone (306). In response to successfully writing the data to the selected zone, the system sets a previous write pointer of the selected zone to a location of an end of the successful write operation (310). In response to not successfully writing the data to the selected zone, the system sets the current write pointer of the selected zone to a location indicated by the previous write pointer of the selected zone (314).

WO 2020/168516 A1

[Continued on next page]

SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

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

METHOD AND SYSTEM FOR FACILITATING FAST ATOMIC WRITE OPERATIONS IN SHINGLED MAGNETIC RECORDING HARD DISK DRIVES

Inventors: Jianjian Huo, Cunyin Chang, Zongpeng Zhu, Fei Liu, and Shu Li

BACKGROUND

Field

[0001] This disclosure is generally related to the field of data storage. More specifically, this disclosure is related to a method and system for facilitating fast atomic write operations in shingled magnetic recording hard disk drives.

Related Art

[0002] The proliferation of the Internet and e-commerce continues to create a vast amount of digital content. Various storage systems have been created to access and store such digital content. Data can be stored in storage systems or storage servers in a data center. The amount of data stored in a data center continues to grow at a rapid pace. To handle this rapid growth, data can be stored in various “tiers” based on the frequency of access. For example, solid state drives (SSDs) are used in the "hot" tier of storage, and generally provide faster access at a higher cost. In contrast, hard disk drives (HDDs) are used in the "warm" and "cold" tiers of storage, and generally provide a lower cost and solid sequential throughput. Some examples of data stored in the warm and cold tiers include user-uploaded photos and videos, enterprise book-keeping data for long term archival purposes, and surveillance recordings. The amount of data stored on these HDDs (for the warm and cold tier of storage) continues to increase at a rapid pace. In some instances, such data must be stored for 50 years or more.

[0003] One solution to lower the cost of storing the increasing amount of warm and cold data on HDDs is a shingled magnetic recording (SMR) HDD. While a conventional magnetic recording (CMR) HDD records data by writing non-overlapping magnetic tracks parallel to each other, an SMR HDD writes new tracks that overlap part of the previously written magnetic track. The tracks partially overlap similar to roof "shingles." A narrower track width can result in a higher track density and an increased storage capacity. Thus, the SMR HDD has an increased areal density and a higher overall per-drive storage capacity, which can result in a decrease in both the storage price per Gigabyte and the total cost of ownership (TCO).

[0004] One common requirement for current storage systems is the preservation of data integrity and consistency in the event of a node or network failure. One method for preserving data integrity and consistency is logging, e.g., journaling. However, because data is first logged to a specific location and then written to a main location in the storage media, the logging method can result in a significant performance degradation, as well as a high write amplification to the storage media. Another method for preserving data integrity and consistency is based on atomic write operations. However, there is currently no support for atomic write operations in hard drives, including SMR HDDs.

SUMMARY

[0005] One embodiment facilitates an atomic write operation. During operation, the system receives a request to write data to a shingled magnetic recording drive. The system selects at least one zone of the drive to which to write the data, wherein the drive stores a previous write pointer indicating a location of an end of a most recent successful write operation in a respective zone. The system writes the data at a location of a current write pointer of the selected zone. In response to successfully writing the data to the selected zone, the system sets a previous write pointer of the selected zone to a location of an end of the successful write operation. In response to not successfully writing the data to the selected zone, the system sets the current write pointer of the selected zone to a location indicated by the previous write pointer of the selected zone.

[0006] In some embodiments, the data is written to the selected zone in a single atomic transaction, thereby facilitating the atomic write operation for successfully writing the data to the shingled magnetic recording drive.

[0007] In some embodiments, the system allocates memory from memory pages for the data to be written, wherein the data is organized into segments and held in a page buffer. A size of a memory page is greater than a first predetermined threshold, and a size of the page buffer is greater than a second predetermined threshold.

[0008] In some embodiments, selecting the at least one zone comprises selecting a plurality of zones to which to write the data. Subsequent to successfully writing one or more segments of the data to the selected zone, the system identifies a next zone of the plurality of zones to which to write remaining segments of the data.

[0009] In some embodiments, in response to successfully writing the data to the selected zone, the system sets the current write pointer of the selected zone to the location of the end of the successful write operation.

[0010] In some embodiments, setting the current write pointer of the selected zone to the location indicated by the previous write pointer of the selected zone is further in response to detecting a failure during a startup or an initialization process.

[0011] In some embodiments, in response to detecting the failure during the startup or initialization process: the system identifies at least one recovery zone on which to perform a recovery process; and the system sets a current write pointer of the identified recovery zone to a location indicated by a previous write pointer of the identified recovery zone.

BRIEF DESCRIPTION OF THE FIGURES

[0012] FIG. 1 illustrates an exemplary environment which facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application.

[0013] FIG. 2 illustrates an exemplary environment which facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application.

[0014] FIG. 3 presents a flowchart illustrating a method for facilitating an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application.

[0015] FIG. 4 presents a flowchart illustrating a method for facilitating an atomic write operation in an SMR HDD, including handling a failure of a write operation, in accordance with an embodiment of the present application.

[0016] FIG. 5 illustrates an exemplary computer system that facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application.

[0017] FIG. 6 illustrates an exemplary apparatus that facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application.

[0018] In the figures, like reference numerals refer to the same figure elements.

DETAILED DESCRIPTION

[0019] The following description is presented to enable any person skilled in the art to make and use the embodiments, and is provided 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 embodiments described herein are not limited to the embodiments shown, but are to be accorded the widest scope consistent with the principles and features disclosed herein.

Overview

[0020] The embodiments described herein at least solve the problem of preserving data integrity and consistency via one or more atomic write operations in a shingled magnetic recording (SMR) hard disk drive (HDD) by tracking both a previous write pointer and a current pointer for each zone of the SMR HDD, and, in addition to, or alternatively, by using a huge page-based memory allocation.

[0021] As described above, the amount of data stored on HDDs (for the warm and cold tier of storage) continues to increase at a rapid pace. In some instances, storage systems must store such data for 50 years or more. One common requirement for current storage systems is the preservation of data integrity and consistency in the event of a node or network failure. One method for preserving data integrity and consistency is logging, e.g., journaling. However, because data is first logged to a specific location and then written to a main location in the storage media, the logging method can result in a significant performance degradation, as well as a high write amplification to the storage media. Another method for preserving data integrity and consistency is based on atomic write operations. However, there is currently no support for atomic write operations in hard drives, including SMR HDDs.

[0022] The embodiments described herein address these challenges by providing a system which facilitates atomic write operations in an SMR HDD. The system includes an integration of hardware and software. On the hardware side, the system can use a retrofitted zone write pointer mechanism. Host-managed (HM) SMR HDDs include specific zones to which data is sequentially written, and the SMR HDD firmware typically maintains a write pointer for each zone. In the embodiments described herein of the retrofitted zone pointer mechanism, for each zone, the SMR HDD firmware can log or track a location of a previous write pointer (e.g., an end of a most recent successful write operation to a given zone) and can update the current write pointer for the zone only after a write operation to the zone has successfully completed. By maintaining the location of both the previous write pointer and the current write pointer, the system can ensure the atomicity of a write operation to a given zone. That is, if a write operation is not successful (i.e., fails to successfully write the entire data to the given zone during the write operation), the unsuccessful write operation does not have any effect. Thus, a partial write cannot occur because the firmware recognizes a zone as a written area based only on the location of the current write pointer, and the location (of the current write pointer) is not updated by the system until after the entire write operation has successfully completed. An exemplary per zone atomic write is described below in relation to FIG. 2.

[0023] In addition to facilitating the per zone atomic write, the embodiments described herein can also facilitate an atomic write across multiple zones, as described below in relation to FIG. 1.

[0024] Furthermore, the SMR HDD firmware can also disable the internal RAM write buffer associated with the zone, zones, and the SMR, to ensure the atomicity of each write operation. However, this may result in decreasing the sequential write performance for each zone, and further result in an increased latency. For example, the HDD may spend more time waiting because there is no RAM buffering for incoming writes, and the operating system typically splits one huge zone write into many small 4 KB aligned write operations. The embodiments described herein address this issue by providing, on the software side, a huge page-based memory allocation, as described below in relation to FIG. 1.

[0025] Thus, the embodiments described herein provide a system which improves the efficiency of a storage system, where the improvements are fundamentally technological. The system provides a technological solution (by maintaining both the previous pointer and the current write pointer and by using a huge page-based memory allocation) to the technological problem of facilitating atomic write operations in SMR HDDs. The system thus provides improvements to the technological field (and related technological applications) of the preservation of data integrity and consistency.

[0026] A "host-managed shingled magnetic recording HDD" ("HM-SMR HDD") adheres to a certain protocol required by the host. In an HM-SMR HDD, the host can manage the shingled nature of the physical media, and writes data sequentially so as not to destroy existing data.

[0027] An "atomic write operation" refers to a write input/output (I/O) request which is performed in a storage media, and can be considered as atomic because the operation either fails or succeeds as a whole. That is, no partial write to the storage media can occur in performing an atomic write operation.

[0028] The term "huge page memory" refers to a memory management method which manages large and fixed-length contiguous blocks of physical memory based on a single entry in a page table or other address mapping structure.

[0029] The term "write amplification" refers to an event associated with flash memory and SMR HDD, where the actual amount of information physically written to the storage media is a multiple of the logical amount intended to be written.

<u>Exemplary Environments Which Facilitate an Atomic Write Operation in an SMR HDD</u>

[0030] The embodiments described herein provide a system which includes both software and hardware components. FIG. 1 illustrates an exemplary environment 100 which facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application. Environment 100 can include software 102 and hardware 140.

[0031] Software 102 can include an I/O management stack 130, which can include, e.g., an I/O dispatch module 132, a polling and event module 134, and a device driver 136. Software 102 can also include a huge page-based memory allocation 110, where memory is allocated to a huge chunk of data in segments corresponding to portions of huge pages. Allocation 110 can include huge pages

120, 124, and 127. Huge page 120 can include unused memory 121, 122, and 123 (as indicated by the white box of unused memory 192 in the index). Similarly, huge page 124 can include unused memory 125 and 126, and huge page 127 can include unused memory 128 and 129. Huge page 120 can include portions which are allocated to corresponding segments of the data to be written, which can be stored in a per-zone I/O write buffer 111. An allocated huge page memory segment can be indicated by a box with left-slanting diagonal lines, as indicated by an allocated huge page memory segment 191 in the index).

[0032] For example: a segment 1 112 of data to be written to the non-volatile memory can be allocated to a portion 113 of huge page 127; a segment 2 114 of the data to be written can be allocated to a portion 115 of huge page 120; a segment 3 116 of the data to be written can be allocated to a portion 117 of huge page 120; and a segment 4 118 of the data to be written can be allocated to a portion 119 of huge page 124. The data of segments 1-4 (112-118) can be stored in a per zone I/O write buffer 111, and subsequently dispatched as I/O to be written to hardware 140.

[0033] As described above, for each zone in an SMR HDD, the SMR HDD firmware can track and store the location of both the current write pointer and the previous write pointer. The SMR HDD firmware can also disable the internal RAM write buffer associated with the zone(s) to ensure the atomicity of each write operation. However, because there is no RAM buffering for incoming writes, this may result in an increased latency due to decreasing the sequential write performance for each zone. The embodiments described herein address this issue by providing, on the software side, huge page-based memory allocation 110. Each per zone I/O write request can consist of multiple segments, e.g., segments 1-4 (112-118), where each segment can correspond to a large chunk allocated from a huge memory page, which is typically in the unit of megabytes (MBs), rather than kilobytes (KBs). A size of a huge memory page can be greater than a first predetermined threshold, and a size of the page buffer can be greater than a second predetermined threshold. These thresholds can be configured by the system or a user. By integrating huge page-based memory allocation 110 with the commonly used scatter/gather mechanism, a host device can issue to the SMR HDD very large single write operations (e.g., up to a typical 256 MB zone size). This can facilitate an improvement in the throughput of atomic write operations in the SMR HDD, as the SMR HDD can spend more time executing the actual write request, rather than waiting for a sequential write operation to be performed (e.g., due to the disabled internal RAM buffer).

[0034] In environment 100, hardware 140 can include a host-managed SMR HDD 141, which can include SMR HDD firmware 142 (e.g., installed in a controller or other component/module). SMR HDD 141 can also include multiple zones to which data is to be written for persistent storage, e.g., a zone 1 150, a zone 2 160, and a zone 3 170. In environment 100, zone areas which are used or include valid and successfully written data are depicted as a shaded box (as indicated by the shaded box of a used/written area 193 in the index). Zone areas in which a current atomic write operation is occurring is depicted by a cross-filled box (as indicated by the cross-filled box of an atomic write area 194 in the index). Zone areas which are empty and available for writing to are depicted by a white box (as indicated by the white box of an empty area 195 in the index). An area of a zone may contain data, but may be considered as empty based on the location of the current write pointer, as described below in relation to FIG. 2.

[0035] Environment 100 depicts how each zone maintains the location of two pointers: a previous write pointer ("wp_prev"), which indicates a location of an end of a most recent successful write operation in a respective zone; and a current write pointer ("wp_curr"), which indicates a location of where a write operation to the respective zone is to begin. The locations of the pair of pointers for each zone in an SMR HDD can be stored in a data structure which is accessible by SMR HDD firmware 142. Various methods may be used to ensure that the pointer locations are appropriately stored and accessible even in the event of a power failure. That is, the data structure may be stored in a volatile memory associated with a host device or with the HDD itself. The system can flush the data structure to a non-volatile memory at periodic intervals. Flushing the data structure can also be based on a predetermined time threshold, the size of the data structure, a capacity of a specific zone or zones associated with storing the data structure, a user configuration, a frequency of access of the data stored in a respective zone, a time-related factor, or any factor. In the embodiments described here, the location of wp_prev and wp_curr can be at the same or different locations, depending on the status and progress of any atomic write operations.

[0036] For example, environment 100 illustrates an atomic write operation of data (e.g., from per zone I/O write buffer 111) to multiple zones of SMR HDD 141. Assume for zone 1 150 that a wp_curr 158 begins at a same location as a wp_prev 156 (as depicted below in relation to FIG. 2), and that the atomic write operation begins at the location indicated by wp_prev 156. The system can update the location of wp_curr 158 only after the data to be written to zone 1 150 has been successfully written to zone 1 150 (i.e., as data 152), or only after all the data has been successfully written to zone 1 150, zone 2 160, and zone 3 170, if the atomic write operation is to be performed across multiple zones. That is, upon determining that the data has been successfully written to a respective zone (or zones, in the case of a write operation across multiple zones), the system can set the location of wp_curr 158 to the end (as shown) of the successful write operation for data 152, and can also set the location of wp_prev 156 to the same end location (not shown).

[0037] Similarly, given a multiple zone write, and assuming for zone 2 160 that a wp_curr 168 begins at a same location as a wp_prev 166, the atomic write operation can continue at the location indicated by wp_prev 166. The system can update the location of wp_curr 168 only after the data to be written to zone 2 160 has been successfully written to zone 2 160 (i.e., as data 162), and only after all the data has been successfully written to zone 1 150, zone 2 160, and zone 3 170 (assuming a multiple zone write). That is, the system can set the location of wp_curr 168 to the end (as shown) of the successful write operation for data 162, and can also set the location of wp_prev 166 to the same end location (not shown).

[0038] Similarly, given a multiple zone write, and assuming for zone 3 170 that a wp_curr 178 begins at a same location as a wp_prev 176, the atomic write operation can continue at the location indicated by wp_prev 176. The system can update the location of wp_curr 178 only after the data to be written to zone 3 170 has been successfully written to zone 3 170 (i.e., as data 172), and only after all the data has been successfully written to zone 1 150, zone 2 160, and zone 3 170 (assuming a multiple zone write). That is, the system can set the location of wp_curr 178 to the end (as shown) of the successful write operation for data 172, and can also set the location of wp_prev 176 to the same end location (not shown).

[0039] If the system detects a failure during an atomic write operation to a single zone (e.g., of data 152 only to zone 1 150), the system can set the location of wp_curr 158 to the location of wp_prev 156, such that the next sequential write to occur in zone 1 150 can begin at the location of the current write pointer, which is set to the location of the previous write pointer.

[0040] If the system detects a failure during an atomic write operation to multiple zones (e.g., of one or more of data 152, 162, and 172 to, respectively, zone 1 150, zone 2 160, and zone 3 170), the system can set the location of the respective current pointers of each zone to the location of the previous pointer of each zone. For example, the system can set the location of wp_curr 158 to the location of wp_prev 156, set the location of wp_curr 168 to the location of wp_prev 166, and set the location of wp_curr 178 to the location of wp_prev 176.

[0041] Thus, by maintaining the locations of the previous write pointer and the current write pointer in each zone of the SMR HDD, and by allocating memory using a huge page-based mechanism, the embodiments described herein can preserve the integrity and consistency of data by facilitating atomic write operations in SMR HDDs.

[0042] FIG. 2 illustrates an exemplary environment 200 which facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application. In environment 200, zone areas which are used or include valid and successfully written data are depicted as a shaded box (as indicated by the shaded box of a used/written area 293 in the index). Zone areas in which a current atomic write operation is occurring is depicted by a cross-filled box (as indicated by the cross-filled box of an atomic write area (in process) 294 in the index). Zone areas which are empty and available for writing to are depicted by a white box (as indicated by the white box of an empty area 295 in the index). An area of a zone may contain data, but may be considered as empty based on the location of the current write pointer (again as indicated by the white box of a considered as empty area 296 in the index).

[0043] Environment 200 can include a zone 1 210 at various times. At a time T1, zone 1 210 can include a used area 202, an empty area 204, a wp_prev 216, and a wp_curr 218. The location of wp_prev 216 and wp_curr 218 can be the same location. The system can perform an atomic write operation (function 252) by writing data beginning at the location of wp_curr 218. That is, at a time T1, zone 1 210 can include data being written to area 206 (indicated as an in-process atomic write area). In some embodiments, the location of the current write pointer may move during the atomic write operation 252, but if the atomic write operation 252 is not successful, the system can set the location of the current write pointer to the location of the previous write pointer. For example, if the atomic write operation is unsuccessful (determination 254), at a time T2, the system can set the location of wp_curr 218 to the location of wp_prev 216. Note that area 206, the area to which the unsuccessful atomic write operation was in process, can be considered as an empty area based on the updated location of the current write pointer.

[0044] On the other hand, if the atomic write operation is successful (determination 256), at a time T3, the system can set the location of the current write pointer to an end of the successful write operation. Note that area 206 is now indicated as a used/written area 206 (by the shaded box). Furthermore, at a time T4, the system can set the location of the previous write pointer to the end of the successful write operation, resulting in similar state as at time T0, where the location of wp_prev 216 and wp_curr 218 are at the same location.

Exemplary Method for Facilitating an Atomic Write Operation in an SMR HDD

[0045] FIG. 3 presents a flowchart 300 illustrating a method for facilitating an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application. During operation, the system receives a request to write data to a shingled magnetic recording (SMR) drive (operation 302). The data can be written to a non-volatile memory of the SMR drive, such as the disk media (e.g., magnetic storage media). The system selects at least one zone of the SMR drive to which to write the data, wherein the SMR drive stores a previous write pointer indicating a location of an end of a most recent successful write operation in a respective zone (operation 304). The system writes the data at a location of a current write pointer of the selected zone (operation 306).

[0046] If the system successfully writes the data to the selected zone (decision 308), the system sets a previous write pointer of the selected zone to a location of an end of the successful write operation (operation 310). The system can also set the current write pointer of the selected zone to the location of the end of the successful write operation (operation 312).

[0047] If the system does not successfully write the data to the selected zone (decision 308), the system sets the current write pointer of the selected zone to a location indicated by the previous write pointer of the selected zone (operation 314).

[0048] If the selected zone is not the last zone to which the data is to be written (decision 320), the operation continues at operation 304. If the selected zone is the last zone to which the data is to be written (decision 320), the system completes the request to write the data to the SMR drive (operation 322), and the operation returns.

Exemplary Method for Handling a Failure of an Atomic Write Operation

[0049] FIG. 4 presents a flowchart 400 illustrating a method for facilitating an atomic write operation in an SMR HDD, including handling a failure of a write operation, in accordance with an embodiment of the present application. During operation, the system detects a failure during startup or during a write operation (operation 402). The system identifies at least one recovery zone on which to perform a recovery process (operation 404). The system sets a current write pointer of the identified recovery zone to a location indicated by a previous write pointer of the identified recovery zone (operation 406).

[0050] If the system determines that there are remaining zones on which to perform recovery (decision 408), the operation continues at operation 402. If the system determines that there are no remaining zones on which to perform recovery (decision 408), the system completes the recovery process (operation 410), and the operation returns.

Exemplary Computer System and Apparatus

[0051] FIG. 5 illustrates an exemplary computer system 500 that facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application. Computer system 500 includes a processor 502, a controller 504, a volatile memory 506, and a storage device 508. Volatile memory 506 can include, e.g., random access memory (RAM), that serves as a managed memory, and can be used to store one or more memory pools. Storage device 508 can include persistent storage which can be managed or accessed via controller 504. Controller 504 can include firmware which can be used to execute the instructions or methods described herein. Furthermore, computer system 500 can be coupled to a display device 510, a keyboard 512, and a pointing device 514. Storage device 508 can store an operating system 516, a content-processing system 518, and data 534.

[0052] Content-processing system 518 can include instructions, which when executed by computer system 500, can cause computer system 500 to perform methods and/or processes described in this disclosure. Specifically, content-processing system 518 can include instructions for receiving and transmitting data packets, including data, I/O data, a write request, data to be written to an SMR HDD, segments of data, and an indicator of a logical address, a physical address, a current write pointer location, and a previous write pointer location (communication module 520).

[0053] Content-processing system 518 can further include instructions for receiving a request to write data to a shingled magnetic recording drive (communication module 520). Content-processing system 518 can include instructions for selecting at least one zone of the drive to which to write the data, wherein the drive stores a previous write pointer indicating a location of an end of a most recent successful write operation in a respective zone (zone-selecting module 522). Content-processing system 518 can include instructions for writing the data at a location of a current write pointer of the selected zone (data-writing module 524). Content-processing system 518 can also include instructions for, in response to successfully writing the data to the selected zone (failure-detecting module 526), setting a previous write pointer of the selected zone to a location of an end of the successful write operation (previous write pointer-managing module 528). Content-processing system 518 can include instructions for, in response to not successfully writing the data to the selected zone (failure-detecting module 526), setting the current write pointer of the selected zone to a location indicated by the previous write pointer of the selected zone (current write pointer-managing module 530).

[0054] Content-processing system 518 can additionally include instructions for allocating memory from memory pages for the data to be written, wherein the data is organized into segments and held in a page buffer (huge page memory-allocating module 532). Content-processing system 518 can include instructions for, in response to detecting the failure during the startup or initialization process (failure-detecting module 526): identifying at least one recovery zone on which to perform a recovery process (zone-selecting module 522); and setting a current write pointer of the identified recovery zone to a location indicated by a previous write pointer of the identified recovery zone (current write pointer-managing module 530).

[0055] Data 534 can include any data that is required as input or that is generated as output by the methods and/or processes described in this disclosure. Specifically, data 534 can store at least: data; an indicator of a location of a previous write pointer and a current write pointer; a segment of data; a logical address; a physical address; a zone; a selected zone; a recovery zone; a plurality of zones; data for an I/O operation; information associated with performing an atomic write operation; allocated memory; data organized into segments; a buffer; a huge page buffer; an incoming per-zone I/O buffer; an indicator of a successful or unsuccessful write operation; and an indicator of a failure detected during a startup or an initialization process.

[0056] FIG. 6 illustrates an exemplary apparatus 600 that facilitates an atomic write operation in an SMR HDD, in accordance with an embodiment of the present application. Apparatus 600 can comprise a plurality of units or apparatuses which may communicate with one another via a wired, wireless, quantum light, or electrical communication channel. Apparatus 600 may be realized using one or more integrated circuits, and may include fewer or more units or apparatuses than those shown in FIG. 6. Further, apparatus 600 may be integrated in a computer system, or realized as a separate device which is capable of communicating with other computer systems and/or devices. Specifically, apparatus 600 can comprise units 602-614 which perform functions or operations similar to modules 520-532 of computer system 500 of FIG. 5, including: a communication unit 602; a zone-selecting unit 604; a data-writing unit 606; a failure-detecting unit 608; a previous write pointer-managing unit 610; a current write pointer-managing unit 612; and a huge page memory-allocating unit 614.

[0057] The data structures and code described in this detailed description are typically stored on a computer-readable storage medium, which may be any device or medium that can store code and/or data for use by a computer system. The computer-readable storage medium includes, but is not limited to, volatile memory, non-volatile memory, magnetic and optical storage devices such as disk drives, magnetic tape, CDs (compact discs), DVDs (digital versatile discs or digital video discs), or other media capable of storing computer-readable media now known or later developed.

[0058] The methods and processes described in the detailed description section can be embodied as code and/or data, which can be stored in a computer-readable storage medium as described above. When a computer system reads and executes the code and/or data stored on the computer-readable storage medium, the computer system performs the methods and processes embodied as data structures and code and stored within the computer-readable storage medium.

[0059] Furthermore, the methods and processes described above can be included in hardware modules. For example, the hardware modules can include, but are not limited to, application-specific integrated circuit (ASIC) chips, field-programmable gate arrays (FPGAs), and other programmable-logic devices now known or later developed. When the hardware modules are activated, the hardware modules perform the methods and processes included within the hardware modules.

[0060] The foregoing embodiments described herein have been presented for purposes of illustration and description only. They are not intended to be exhaustive or to limit the embodiments described herein to the forms disclosed. Accordingly, many modifications and variations will be apparent to practitioners skilled in the art. Additionally, the above disclosure is not intended to limit the embodiments described herein. The scope of the embodiments described herein is defined by the appended claims.

What Is Claimed Is:

1. A computer-implemented method for facilitating an atomic write operation, the method comprising:

receiving a request to write data to a shingled magnetic recording drive;

selecting at least one zone of the drive to which to write the data,

wherein the drive stores a previous write pointer indicating a location of an end of a most recent successful write operation in a respective zone;

writing the data at a location of a current write pointer of the selected zone;

in response to successfully writing the data to the selected zone, setting a previous write pointer of the selected zone to a location of an end of the successful write operation; and

in response to not successfully writing the data to the selected zone, setting the current write pointer of the selected zone to a location indicated by the previous write pointer of the selected zone.

2. The method of claim 1, wherein the data is written to the selected zone in a single atomic transaction, thereby facilitating the atomic write operation for successfully writing the data to the shingled magnetic recording drive.

3. The method of claim 1, further comprising:

allocating memory from memory pages for the data to be written,

wherein the data is organized into segments and held in a page buffer,

wherein a size of a memory page is greater than a first predetermined threshold, and

wherein a size of the page buffer is greater than a second predetermined threshold.

4. The method of claim 1, wherein selecting the at least one zone comprises selecting a plurality of zones to which to write the data, and wherein the method further comprises:

subsequent to successfully writing one or more segments of the data to the selected zone, identifying a next zone of the plurality of zones to which to write remaining segments of the data.

5. The method of claim 1, wherein in response to successfully writing the data to the selected zone, the method further comprises:

setting the current write pointer of the selected zone to the location of the end of the successful write operation.

6. The method of claim 1, wherein setting the current write pointer of the selected zone to the location indicated by the previous write pointer of the selected zone is further in response to detecting a failure during a startup or an initialization process.

7. The method of claim 6, wherein in response to detecting the failure during the startup or initialization process, the method further comprises:

identifying at least one recovery zone on which to perform a recovery process; and

setting a current write pointer of the identified recovery zone to a location indicated by a previous write pointer of the identified recovery zone.

8. A computer system for facilitating data placement, the system comprising:

a processor; and

a memory coupled to the processor and storing instructions, which when executed by the processor cause the processor to perform a method, wherein the computer system is a storage device, the method comprising:

receiving a request to write data to a shingled magnetic recording drive;

selecting at least one zone of the drive to which to write the data,

wherein the drive stores a previous write pointer indicating a location of an end of a most recent successful write operation in a respective zone;

writing the data at a location of a current write pointer of the selected zone;

in response to successfully writing the data to the selected zone, setting a previous write pointer of the selected zone to a location of an end of the successful write operation; and

in response to not successfully writing the data to the selected zone, setting the current write pointer of the selected zone to a location indicated by the previous write pointer of the selected zone.

9. The computer system of claim 8, wherein the data is written to the selected zone in a single atomic transaction, thereby facilitating the atomic write operation for successfully writing the data to the shingled magnetic recording drive.

10. The computer system of claim 8, wherein the method further comprises:

allocating memory from memory pages for the data to be written,

wherein the data is organized into segments and held in a page buffer,

wherein a size of a memory page is greater than a first predetermined threshold, and

wherein a size of the page buffer is greater than a second predetermined threshold.

11. The computer system of claim 8, wherein selecting the at least one zone comprises selecting a plurality of zones to which to write the data, and wherein the method further comprises:

subsequent to successfully writing one or more segments of the data to the selected zone, identifying a next zone of the plurality of zones to which to write remaining segments of the data.

12. The computer system of claim 8, wherein in response to successfully writing the data to the selected zone, the method further comprises:

setting the current write pointer of the selected zone to the location of the end of the successful write operation.

13. The computer system of claim 8, wherein setting the current write pointer of the selected zone to the location indicated by the previous write pointer of the selected zone is further in response to detecting a failure during a startup or an initialization process.

14. The computer system of claim 13, wherein in response to detecting the failure during the startup or initialization process, the method further comprises:

identifying at least one recovery zone on which to perform a recovery process; and

setting a current write pointer of the identified recovery zone to a location indicated by a previous write pointer of the identified recovery zone.

15. A non-transitory computer-readable storage medium storing instructions that when executed by a computer cause the computer to perform a method, the method comprising:

receiving a request to write data to a shingled magnetic recording drive;

selecting at least one zone of the drive to which to write the data,

wherein the drive stores a previous write pointer indicating a location of an end of a most recent successful write operation in a respective zone;

writing the data at a location of a current write pointer of the selected zone;

in response to successfully writing the data to the selected zone, setting a previous write pointer of the selected zone to a location of an end of the successful write operation; and

in response to not successfully writing the data to the selected zone, setting the current write pointer of the selected zone to a location indicated by the previous write pointer of the selected zone.

16. The storage medium of claim 15, wherein the data is written to the selected zone in a single atomic transaction, thereby facilitating the atomic write operation for successfully writing the data to the shingled magnetic recording drive.

17. The storage medium of claim 15, wherein the method further comprises:

allocating memory from memory pages for the data to be written,

wherein the data is organized into segments and held in a page buffer,

wherein a size of a memory page is greater than a first predetermined threshold, and

wherein a size of the page buffer is greater than a second predetermined threshold.

18. The storage medium of claim 15, wherein selecting the at least one zone comprises selecting a plurality of zones to which to write the data, and wherein the method further comprises:

subsequent to successfully writing one or more segments of the data to the selected zone, identifying a next zone of the plurality of zones to which to write remaining segments of the data.

19. The storage medium of claim 15, wherein in response to successfully writing the data to the selected zone, the method further comprises:

setting the current write pointer of the selected zone to the location of the end of the successful write operation.

20. The storage medium of claim 15, wherein setting the current write pointer of the selected zone to the location indicated by the previous write pointer of the selected zone is further in response to detecting a failure during a startup or an initialization process, and

wherein in response to detecting the failure during the startup or initialization process, the method further comprises:

identifying at least one recovery zone on which to perform a recovery process; and

setting a current write pointer of the identified recovery zone to a location indicated by a previous write pointer of the identified recovery zone.

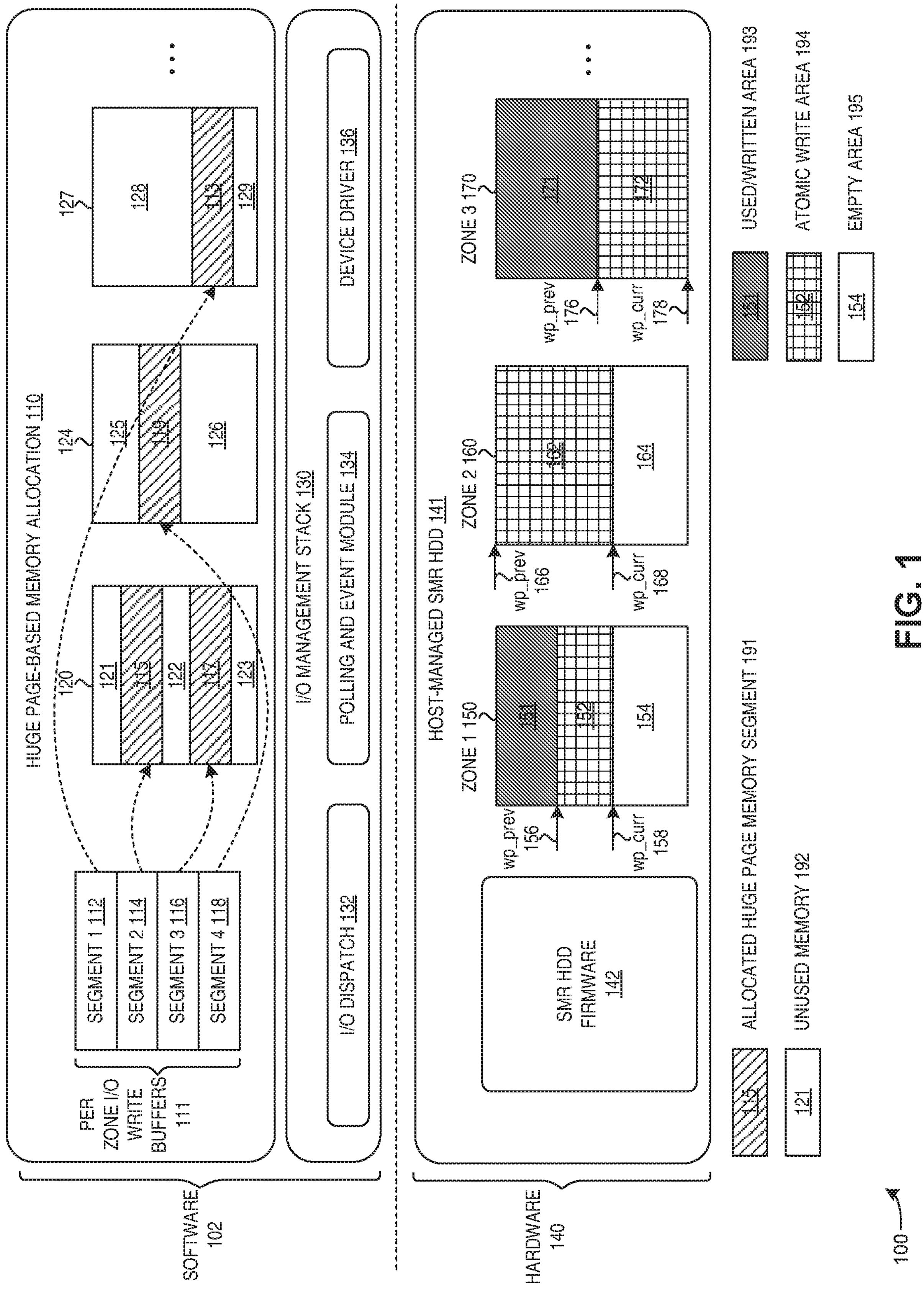

SOFTWARE 102
HUGE PAGE-BASED MEMORY ALLOCATION 110
PER ZONE I/O WRITE BUFFERS 111
SEGMENT 1 112
SEGMENT 2 114
SEGMENT 3 116
SEGMENT 4 118
120
121
115
122
117
123
124
125
119
126
127
128
113
129
I/O MANAGEMENT STACK 130
I/O DISPATCH 132
POLLING AND EVENT MODULE 134
DEVICE DRIVER 136
HARDWARE 140
HOST-MANAGED SMR HDD 141
SMR HDD FIRMWARE 142
ZONE 1 150
wp_prev 156
wp_curr 158
151
152
154
ZONE 2 160
wp_prev 166
wp_curr 168
162
164
ZONE 3 170
wp_prev 176
wp_curr 178
171
172
115 ALLOCATED HUGE PAGE MEMORY SEGMENT 191
121 UNUSED MEMORY 192
151 USED/WRITTEN AREA 193
152 ATOMIC WRITE AREA 194
154 EMPTY AREA 195
100

FIG. 1

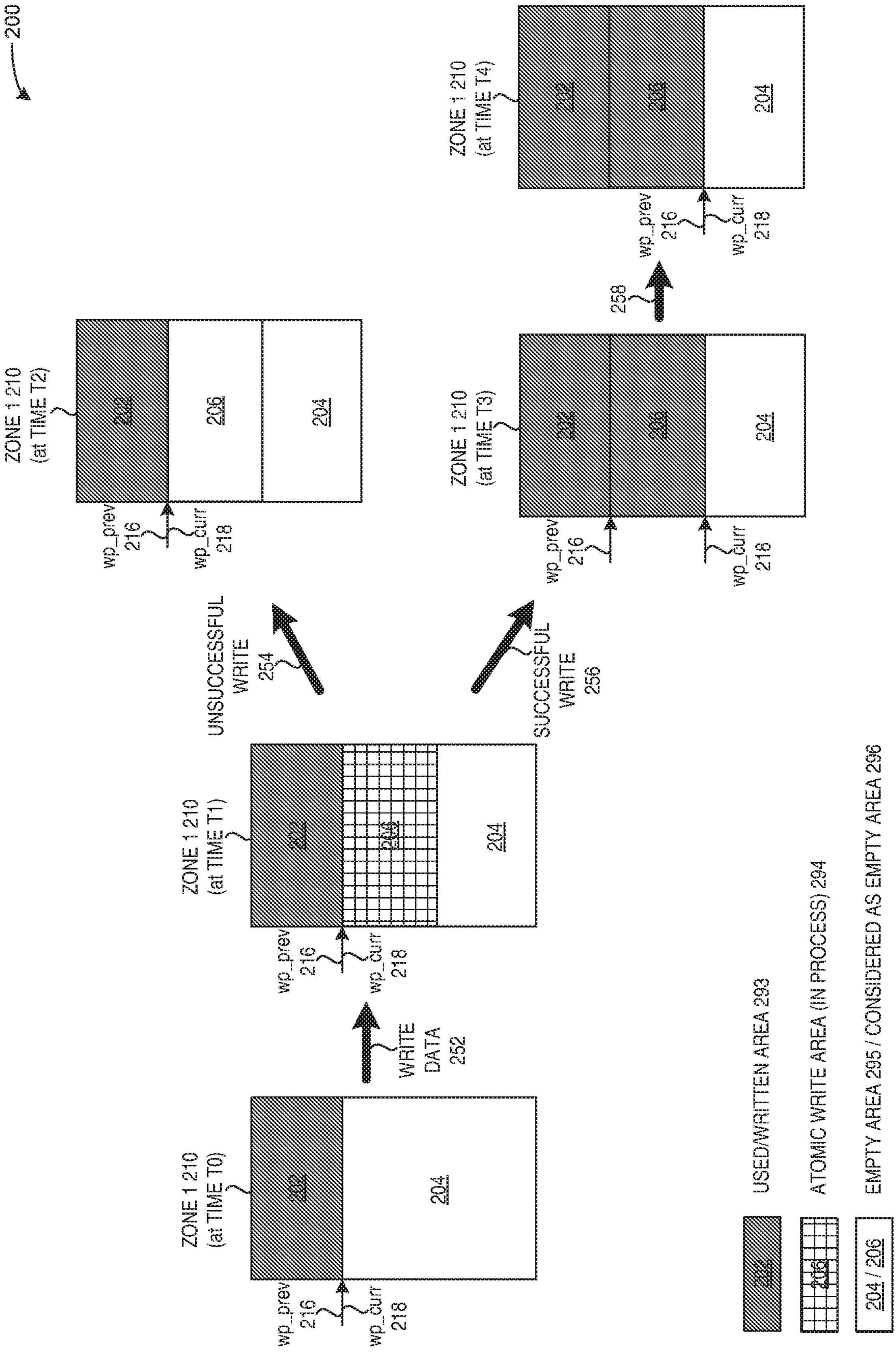

200
ZONE 1 210 (at TIME T0)
202
204
wp_prev 216
wp_curr 218
WRITE DATA 252
ZONE 1 210 (at TIME T1)
202
206
204
wp_prev 216
wp_curr 218
UNSUCCESSFUL WRITE 254
SUCCESSFUL WRITE 256
ZONE 1 210 (at TIME T2)
202
206
204
wp_prev 216
wp_curr 218
ZONE 1 210 (at TIME T3)
202
206
204
wp_prev 216
wp_curr 218
258
ZONE 1 210 (at TIME T4)
202
206
204
wp_prev 216
wp_curr 218
202
USED/WRITTEN AREA 293
206
ATOMIC WRITE AREA (IN PROCESS) 294
204 / 206
EMPTY AREA 295 / CONSIDERED AS EMPTY AREA 296

FIG. 2

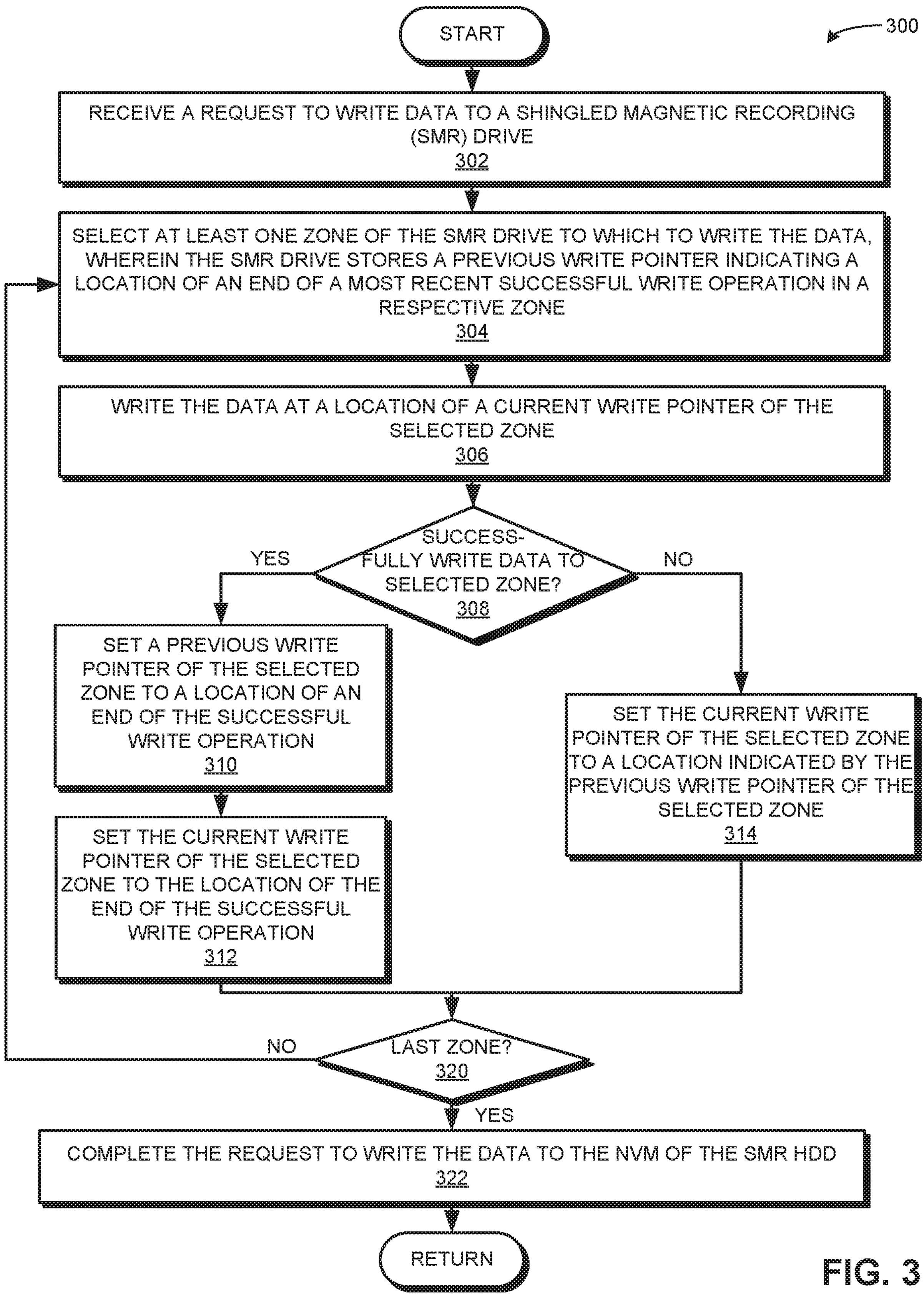

300
START
RECEIVE A REQUEST TO WRITE DATA TO A SHINGLED MAGNETIC RECORDING (SMR) DRIVE
302
SELECT AT LEAST ONE ZONE OF THE SMR DRIVE TO WHICH TO WRITE THE DATA, WHEREIN THE SMR DRIVE STORES A PREVIOUS WRITE POINTER INDICATING A LOCATION OF AN END OF A MOST RECENT SUCCESSFUL WRITE OPERATION IN A RESPECTIVE ZONE
304
WRITE THE DATA AT A LOCATION OF A CURRENT WRITE POINTER OF THE SELECTED ZONE
306
SUCCESS-FULLY WRITE DATA TO SELECTED ZONE?
308
YES
NO
SET A PREVIOUS WRITE POINTER OF THE SELECTED ZONE TO A LOCATION OF AN END OF THE SUCCESSFUL WRITE OPERATION
310
SET THE CURRENT WRITE POINTER OF THE SELECTED ZONE TO THE LOCATION OF THE END OF THE SUCCESSFUL WRITE OPERATION
312
SET THE CURRENT WRITE POINTER OF THE SELECTED ZONE TO A LOCATION INDICATED BY THE PREVIOUS WRITE POINTER OF THE SELECTED ZONE
314
LAST ZONE?
320
NO
YES
COMPLETE THE REQUEST TO WRITE THE DATA TO THE NVM OF THE SMR HDD
322
RETURN

FIG. 3

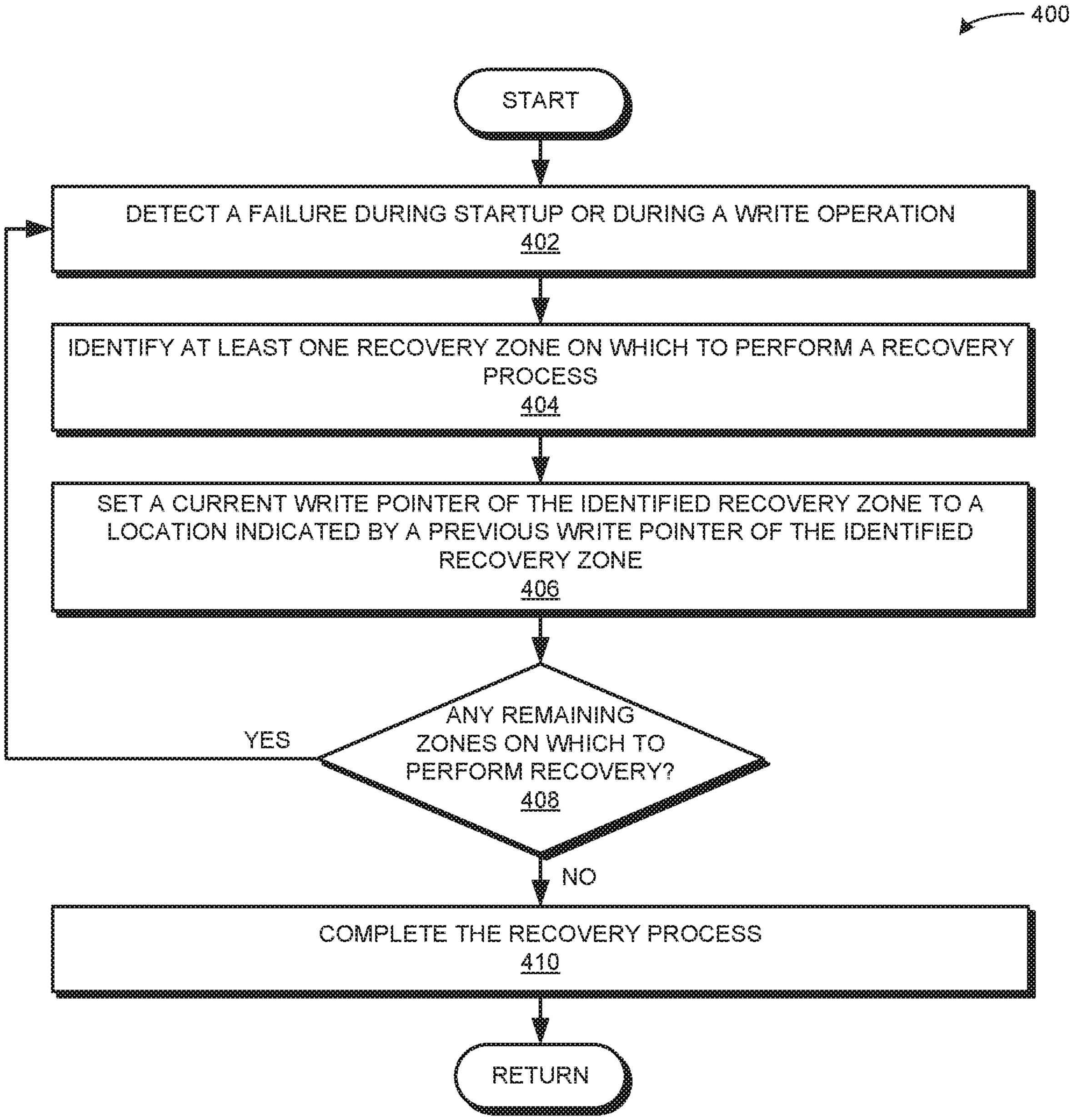

400
START
DETECT A FAILURE DURING STARTUP OR DURING A WRITE OPERATION
402
IDENTIFY AT LEAST ONE RECOVERY ZONE ON WHICH TO PERFORM A RECOVERY PROCESS
404
SET A CURRENT WRITE POINTER OF THE IDENTIFIED RECOVERY ZONE TO A LOCATION INDICATED BY A PREVIOUS WRITE POINTER OF THE IDENTIFIED RECOVERY ZONE
406
ANY REMAINING ZONES ON WHICH TO PERFORM RECOVERY?
408
YES
NO
COMPLETE THE RECOVERY PROCESS
410
RETURN

FIG. 4

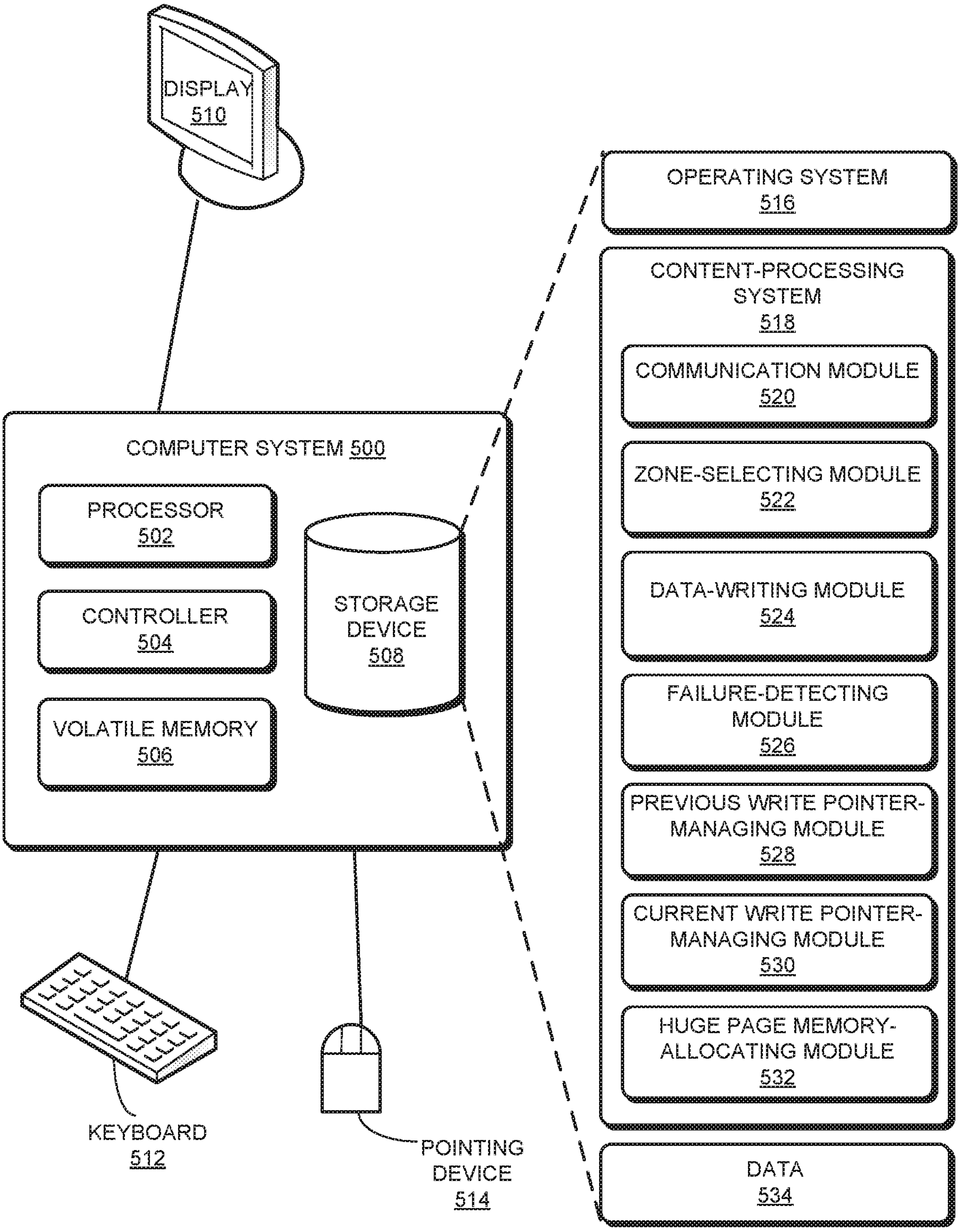

DISPLAY
510
OPERATING SYSTEM
516
CONTENT-PROCESSING
SYSTEM
518
COMMUNICATION MODULE
520
COMPUTER SYSTEM 500
ZONE-SELECTING MODULE
522
PROCESSOR
502
DATA-WRITING MODULE
524
CONTROLLER
504
STORAGE
DEVICE
508
FAILURE-DETECTING
MODULE
526
VOLATILE MEMORY
506
PREVIOUS WRITE POINTER-
MANAGING MODULE
528
CURRENT WRITE POINTER-
MANAGING MODULE
530
HUGE PAGE MEMORY-
ALLOCATING MODULE
532
KEYBOARD
512
POINTING
DEVICE
514
DATA
534

FIG. 5

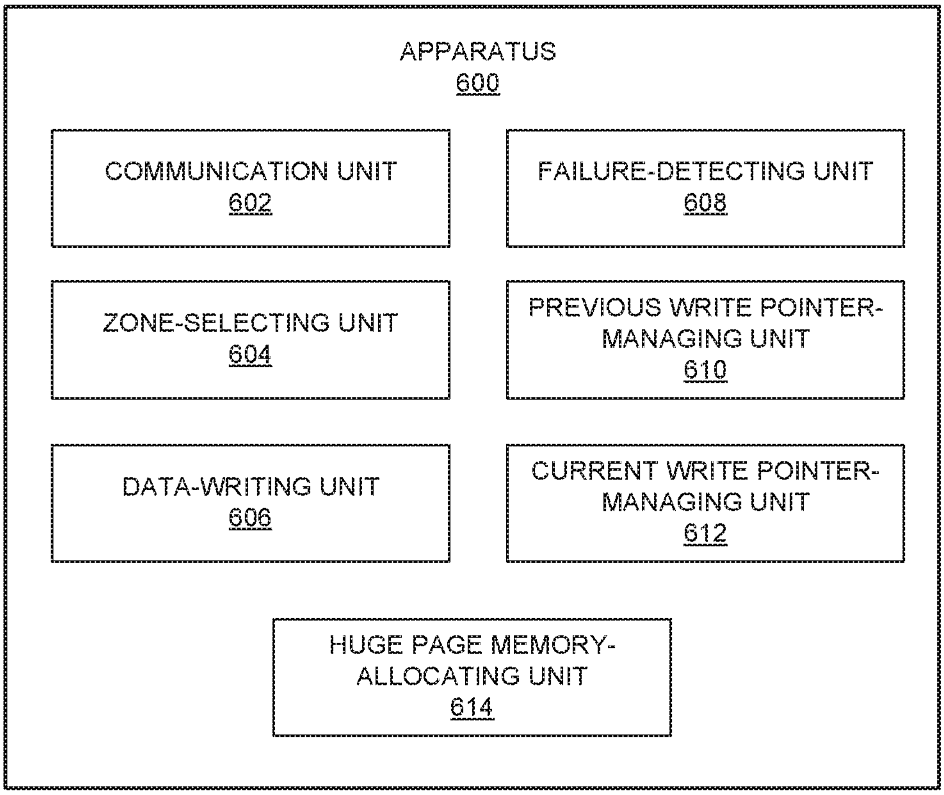

APPARATUS
600
COMMUNICATION UNIT
602
FAILURE-DETECTING UNIT
608
ZONE-SELECTING UNIT
604
PREVIOUS WRITE POINTER-MANAGING UNIT
610
DATA-WRITING UNIT
606
CURRENT WRITE POINTER-MANAGING UNIT
612
HUGE PAGE MEMORY-ALLOCATING UNIT
614

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/075713

A. CLASSIFICATION OF SUBJECT MATTER

G06F 3/06(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

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)

CNPAT, CNKI, WPI, EPODOC: atomic, write, shingled magnetic recording drive, SRM, zone, data, location, pointer, select, previous

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 108228483 A (BEIJING MEMBLAZE TECHNOLOGY CO., LTD.) 29 June 2018 (2018-06-29) description, paragraphs [0016]-[0065]	1-20
Y	CN 107765992 A (HUAWEI TECHNOLOGIES CO., LTD. ET AL.) 06 March 2018 (2018-03-06) description, paragraph [0093]	1-20
A	CN 108108118 A (EPOSTAR ELECTRONICS CO., LTD.) 01 June 2018 (2018-06-01) the whole document	1-20
A	US 2008052444 A1 (WINBOND ELECTRONICS CORPORATION) 28 February 2008 (2008-02-28) the whole document	1-20

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

* Special categories of cited documents:

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

"E" earlier 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
06 November 2019	**20 November 2019**

Name and mailing address of the ISA/CN	Authorized officer
National Intellectual Property Administration, PRC **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088** **China**	**MENG,Zishan**
Facsimile No. **(86-10)62019451**	Telephone No. **86-(10)-53961387**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/075713

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108228483	A	29 June 2018	None			
CN	107765992	A	06 March 2018	None			
CN	108108118	A	01 June 2018	None			
US	2008052444	A1	28 February 2008	TW	I311327	B	21 June 2009
				US	7836242	B2	16 November 2010
				TW	200729232	A	01 August 2007