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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 March 2016 (10.03.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/036713 A1

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

(21) International Application Number:
PCT/US2015/047898

(22) International Filing Date:
1 September 2015 (01.09.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/044,905 2 September 2014 (02.09.2014) US
14/621,263 12 February 2015 (12.02.2015) US

(71) Applicant: **SANDISK TECHNOLOGIES INC.** [US/US]; Two Legacy Town Center, 6900 North Dallas Parkway, Plano, TX 75024 (US).

(72) Inventors: **SAMUELS, Allen**; Two Legacy Town Center, 6900 North Dallas Parkway, Plano, TX 75024 (US). **KRUGER, Warren, Fritz**; Two Legacy Town Center, 6900 North Dallas Parkway, Plano, TX 75024 (US). **TRUONG, Linh, Tien**; Two Legacy Town Center, 6900 North Dallas Parkway, Plano, TX 75024 (US).

(74) Agents: **WILLIAMS, Gary, S.** et al.; Morgan Lewis & Bockius LLP, 2 Palo Alto Square, 3000 El Camino Real, Suite 700, Palo Alto, CA 94306 (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, 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, KN, KP, KR, 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, 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))*

(54) Title: NOTIFICATION OF TRIGGER CONDITION TO REDUCE DECLARED CAPACITY OF A STORAGE DEVICE IN A MULTI-STORAGE-DEVICE STORAGE SYSTEM

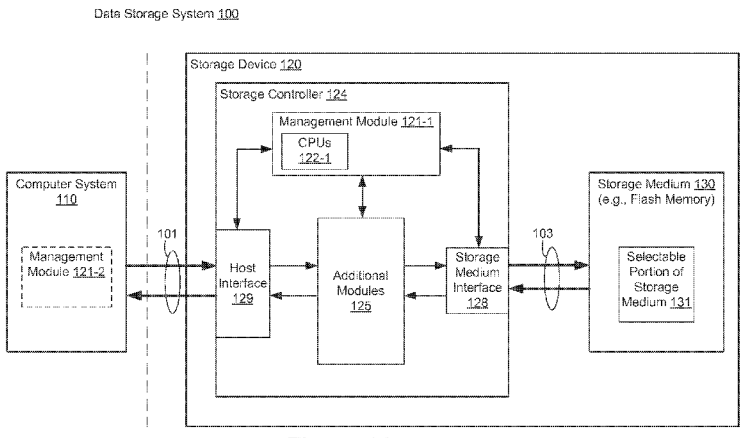


Figure 1A

(57) Abstract: Systems, methods and/or devices are used to enable notification of a trigger condition to reduce declared capacity of a storage device in a multi-storage-device storage system. In one aspect, the method includes: (1) obtaining, for each storage device of a plurality of storage devices, one or more metrics of the storage device, the storage device including non¬ volatile memory, (2) detecting a trigger condition for reducing declared capacity of the non¬ volatile memory of a respective storage device of the plurality of storage devices, the trigger condition detected in accordance with the one or more metrics of two or more of the storage devices, and (3) notifying a host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device.

WO 2016/036713 A1

Notification of Trigger Condition to Reduce Declared Capacity of a Storage Device in a Multi-Storage-Device Storage System

TECHNICAL FIELD

[0001] The disclosed embodiments relate generally to memory systems, and in particular, to notification of a trigger condition to reduce declared capacity of a storage device (e.g., comprising one or more flash memory devices) in a multi-storage-device storage system.

BACKGROUND

[0002] Semiconductor memory devices, including flash memory, typically utilize memory cells to store data as an electrical value, such as an electrical charge or voltage. A flash memory cell, for example, includes a single transistor with a floating gate that is used to store a charge representative of a data value. Flash memory is a non-volatile data storage device that can be electrically erased and reprogrammed. More generally, non-volatile memory (e.g., flash memory, as well as other types of non-volatile memory implemented using any of a variety of technologies) retains stored information even when not powered, as opposed to volatile memory, which requires power to maintain the stored information. Increases in storage density have been facilitated in various ways, including increasing the density of memory cells on a chip enabled by manufacturing developments, and transitioning from single-level flash memory cells to multi-level flash memory cells, so that two or more bits can be stored by each flash memory cell.

[0003] Repeated erasure and reprogramming of flash memory cells causes degradation of the charge storage capability (wear). Eventually, the charge storage capability degrades to the point where it becomes impossible or infeasible to recover the original data (e.g., an unrecoverable codeword is read from the flash memory device, the computational resources required to recover a codeword exceed a predefined threshold, or a count of program-erase (P/E) cycles for the flash memory device exceeds a threshold value) and the device is considered worn out. Wear-leveling technology is often used to distribute the wear across the multiple portions of a flash memory device. In a typical system, once the wear limit of a portion of a flash memory device is reached, the entire flash memory device is considered to have failed.

SUMMARY

[0004] Various embodiments of systems, methods and devices within the scope of the appended claims each have several aspects, no single one of which is solely responsible for the attributes described herein. Without limiting the scope of the appended claims, after considering this disclosure, and particularly after considering the section entitled "Detailed Description" one will understand how the aspects of various embodiments are used to enable notification of a trigger condition to reduce declared capacity of a storage device in a multi-storage-device storage system. In one aspect, a host to which a respective storage device of a plurality of storage devices of a storage system is operatively coupled is notified of a trigger condition for reducing declared capacity of non-volatile memory of the respective storage device.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] So that the present disclosure can be understood in greater detail, a more particular description may be had by reference to the features of various embodiments, some of which are illustrated in the appended drawings. The appended drawings, however, merely illustrate pertinent features of the present disclosure and are therefore not to be considered limiting, for the description may admit to other effective features.

[0006] Figure 1A is a block diagram illustrating an implementation of a data storage system, in accordance with some embodiments.

[0007] Figure 1B is a block diagram illustrating an implementation of a data storage system, in accordance with some embodiments.

[0008] Figure 1C is a block diagram illustrating an implementation of a data storage system, in accordance with some embodiments.

[0009] Figure 2A-1 is a block diagram illustrating an implementation of a management module, in accordance with some embodiments.

[0010] Figure 2A-2 is a block diagram illustrating an implementation of a management module, in accordance with some embodiments.

[0011] Figure 2B-1 is a block diagram illustrating an implementation of a system management module, in accordance with some embodiments.

[0012] Figure 2B-2 is a block diagram illustrating an implementation of a system management module, in accordance with some embodiments.

[0013] Figure 2C-1 is a block diagram illustrating an implementation of a cluster management module, in accordance with some embodiments.

[0014] Figure 2C-2 is a block diagram illustrating an implementation of a cluster management module, in accordance with some embodiments.

[0015] Figure 2D is a block diagram illustrating an implementation of an amelioration module included in Figures 2A-1 and 2A-2, in accordance with some embodiments.

[0016] Figure 3 is a block diagram of a logical address space, and more specifically a logical block address (LBA) space, in accordance with some embodiments.

[0017] Figure 4 is a block diagram of a mapping table and physical address space, in accordance with some embodiments.

[0018] Figure 5A is a prophetic diagram of voltage distributions that may be found in a single-level flash memory cell (SLC) over time, in accordance with some embodiments.

[0019] Figure 5B is a prophetic diagram of voltage distributions that may be found in a multi-level flash memory cell (MLC) over time, in accordance with some embodiments.

[0020] Figure 6 illustrates a flowchart representation of a method of managing a storage system, in accordance with some embodiments.

[0021] Figures 7A-7D illustrate a flowchart representation of a method of managing a storage system, in accordance with some embodiments.

[0022] Figures 8A-8D illustrate a flowchart representation of a method of managing a storage system, in accordance with some embodiments.

[0023] Figures 9A-9D illustrate a flowchart representation of a method of managing a storage system, in accordance with some embodiments.

[0024] Figures 10A-10C illustrate a flowchart representation of a method of managing a storage system, in accordance with some embodiments.

[0025] Figures 11A-11C illustrate a flowchart representation of a method of managing a storage system, in accordance with some embodiments.

[0026] In accordance with common practice the various features illustrated in the drawings may not be drawn to scale. Accordingly, the dimensions of the various features may be arbitrarily expanded or reduced for clarity. In addition, some of the drawings may

not depict all of the components of a given system, method or device. Finally, like reference numerals may be used to denote like features throughout the specification and figures.

DETAILED DESCRIPTION

[0027] When a multi-level flash cell has reached its wear limit it typically still has charge retention capability sufficient to store a reduced number of charge levels. Often it is the case that a substantial number of erasure and reprogramming cycles can be performed on a wear-limited multi-level flash cell with full recovery of the stored data, provided that a reduced number of charge levels is used and expected. For example, a flash memory device operating in 3 bits per cell mode (TLC) typically can perform between 500 and 1500 erasure and reprogramming cycles before being considered worn out. However, at that point in time it will typically still have sufficient charge storage capability to operate in the single bit per cell mode (SLC) for an additional 10,000 to 20,000 erasure and reprogramming cycles before the SLC wear limit is encountered. Thus the lifetime of the flash memory device may be extended provided that it can be allowed to store less data. Currently it is difficult for a storage system to utilize this extended capability because storage system mechanisms for managing and working with a storage device whose capacity decreases with usage, by decreasing over-provisioning, are inadequate. Consequently what is desired are mechanisms for managing and working with such a storage device, including mechanisms to inform the surrounding system of its impending (or imminent) reduction in capacity so that the system may adjust its operation accordingly. Potentially, memory devices with other forms of non-volatile memory may benefit from the same or similar mechanisms as those described in this document.

[0028] The various embodiments described herein include systems, methods and/or devices used to enable triggering a process to reduce declared capacity of a storage device. Some embodiments include systems, methods and/or devices to detect a trigger condition in accordance with one or more metrics of a storage device and enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of non-volatile memory of the storage device.

[0029] (A1) More specifically, some embodiments include a method of managing a storage system. In some embodiments, the method includes, at a storage device of the storage system: (1) generating one or more metrics of the storage device, the storage device including non-volatile memory, (2) detecting a trigger condition in accordance with the one or more metrics of the storage device, and (3) enabling an amelioration process associated

with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device.

[0030] (A1-1) In some embodiments of the method of A1, the method further includes: (1) prior to detecting the trigger condition, detecting a first wear condition of the non-volatile memory of the storage device, wherein a total storage capacity of the non-volatile memory of the storage device includes declared capacity and over-provisioning, and (2) in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the storage device without reducing declared capacity of the non-volatile memory of the storage device.

[0031] (A1-2) In some embodiments of the method of A1-1, detecting the trigger condition includes detecting a second wear condition distinct from the first wear condition.

[0032] (A2) In some embodiments of the method of A1 or A1-1 or A1-2, enabling the amelioration process associated with the detected trigger condition includes notifying a host to which the storage device is operatively coupled of the trigger condition.

[0033] (A3) In some embodiments of the method of A1 or A1-1 or A1-2, enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a query from a host to which the storage device is operatively coupled, and (2) in response to receiving the query, reporting the trigger condition.

[0034] (A4) In some embodiments of the method of A1 or A1-1 or A1-2, enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition.

[0035] (A5) In some embodiments of the method of A1 or A1-1 or A1-2, enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition.

[0036] (A6) In some embodiments of the method of any of A2 to A5, the host includes a client on behalf of which data is stored in the storage system.

[0037] (A7) In some embodiments of the method of any of A2 to A5, the host includes a storage system controller of the storage system.

[0038] (A8) In some embodiments of the method of any of A2 to A5, the host includes a cluster controller of the storage system.

[0039] (A9) In some embodiments of the method of any of A1 to A8, enabling the amelioration process associated with the detected trigger condition includes scheduling the amelioration process to be performed on the storage device.

[0040] (A10) In some embodiments of the method of any of A1 to A9, enabling the amelioration process associated with the detected trigger condition includes determining one or more parameters for the amelioration process.

[0041] (A11) In some embodiments of the method of A10, enabling the amelioration process associated with the detected trigger condition further includes reporting at least a subset of the one or more parameters for the amelioration process.

[0042] (A12) In some embodiments of the method of any of A1 to A11, generating one or more metrics of the storage device includes generating at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the storage device.

[0043] (A13) In some embodiments of the method of any of A1 to A12, the one or more metrics of the storage device include one or more status metrics corresponding to the storage device's ability to retain data.

[0044] (A14) In some embodiments of the method of any of A1 to A13, the one or more metrics of the storage device include one or more performance metrics corresponding to performance of the storage device.

[0045] (A15) In some embodiments of the method of any of A1 to A14, the one or more metrics of the storage device include one or more wear metrics corresponding to wear on the storage device.

[0046] (A16) In some embodiments of the method of any of A1 to A15, the one or more metrics of the storage device include one or more time metrics.

[0047] (A17) In some embodiments of the method of any of A1 to A16, the one or more metrics of the storage device include values of the one or more metrics from more than one time.

[0048] (A18) In some embodiments of the method of any of A1 to A17, the method further includes, after enabling the amelioration process, (1) re-evaluating the trigger

condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, aborting the amelioration process to reduce declared capacity of the non-volatile memory of the storage device.

[0049] (A19) In some embodiments of the method of any of A1 to A18, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device includes a process to reduce utilization of the non-volatile memory of the storage device.

[0050] (A20) In some embodiments of the method of any of A1 to A19, the storage device comprises one or more flash memory devices.

[0051] (A21) In another aspect, a storage device includes (1) non-volatile memory (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) one or more processors, and (3) controller memory (e.g., non-volatile memory or volatile memory in or coupled to the controller) storing one or more programs, which when executed by the one or more processors cause the storage device to perform or control performance of any of the methods A1 to A20 described herein.

[0052] (A23) In yet another aspect, any of the methods A1 to A20 described above are performed by a storage device including means for performing any of the methods described herein.

[0053] (A25) In yet another aspect, a storage system includes (1) a storage medium (e.g., comprising one or more non-volatile storage devices, such as flash memory devices) (2) one or more processors, and (3) memory (e.g., non-volatile memory or volatile memory in the storage system) storing one or more programs, which when executed by the one or more processors cause the storage system to perform or control performance of any of the methods A1 to A20 described herein.

[0054] (A26) In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a storage device, the one or more programs including instructions for performing any of the methods described herein.

[0055] The various embodiments described herein include systems, methods and/or devices used to enable triggering, at a host, a process to reduce declared capacity of a storage device. Some embodiments include systems, methods and/or devices to detect a trigger

condition in accordance with one or more metrics of a storage device and enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of non-volatile memory of the storage device.

[0056] (B1) More specifically, some embodiments include a method of managing a storage system. In some embodiments, the method includes, at a host to which a storage device of the storage system is operatively coupled: (1) obtaining one or more metrics of the storage device, the storage device including non-volatile memory, (2) detecting a trigger condition in accordance with the one or more metrics of the storage device, and (3) enabling an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device.

[0057] (B1-1) In some embodiments of the method of B1, the method further includes: (1) prior to detecting the trigger condition, detecting a first wear condition of the non-volatile memory of the storage device, wherein a total storage capacity of the non-volatile memory of the storage device includes declared capacity and over-provisioning, and (2) in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the storage device without reducing declared capacity of the non-volatile memory of the storage device.

[0058] (B1-2) In some embodiments of the method of B1-1, detecting the trigger condition includes detecting a second wear condition distinct from the first wear condition.

[0059] (B2) In some embodiments of the method of B1 or B1-1 or B1-2, the host includes a client on behalf of which data is stored in the storage system.

[0060] (B3) In some embodiments of the method of B1 or B1-1 or B1-2, the host includes a storage system controller of the storage system.

[0061] (B4) In some embodiments of the method of B1 or B1-1 or B1-2, the host includes a cluster controller of the storage system.

[0062] (B5) In some embodiments of the method of any of B1 to B4, enabling the amelioration process associated with the detected trigger condition includes scheduling the amelioration process to be performed on the storage device.

[0063] (B6) In some embodiments of the method of any of B1 to B5, enabling the amelioration process associated with the detected trigger condition includes determining one or more parameters for the amelioration process.

[0064] (B7) In some embodiments of the method of B6, enabling the amelioration process associated with the detected trigger condition further includes conveying at least a subset of the one or more parameters for the amelioration process to the storage device.

[0065] (B8) In some embodiments of the method of any of B1 to B7, obtaining one or more metrics of the storage device includes obtaining at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the storage device.

[0066] (B9) In some embodiments of the method of any of B1 to B8, the one or more metrics of the storage device include one or more status metrics corresponding to the storage device's ability to retain data.

[0067] (B10) In some embodiments of the method of any of B1 to B9, the one or more metrics of the storage device include one or more performance metrics corresponding to performance of the storage device.

[0068] (B11) In some embodiments of the method of any of B1 to B10, the one or more metrics of the storage device include one or more wear metrics corresponding to wear on the storage device.

[0069] (B12) In some embodiments of the method of any of B1 to B11, the one or more metrics of the storage device include one or more time metrics.

[0070] (B13) In some embodiments of the method of any of B1 to B12, the one or more metrics of the storage device include values of the one or more metrics from more than one time.

[0071] (B14) In some embodiments of the method of any of B1 to B13, the method further includes, after enabling the amelioration process, (1) re-evaluating the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, aborting the amelioration process to reduce declared capacity of the non-volatile memory of the storage device.

[0072] (B15) In some embodiments of the method of any of B1 to B14, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device includes a process to reduce utilization of the non-volatile memory of the storage device.

[0073] (B16) In some embodiments of the method of any of B1 to B15, the storage device comprises one or more flash memory devices.

[0074] (B17) In another aspect, a storage system includes (1) one or more storage devices (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) a host to which the one or more storage devices are operatively coupled, (3) one or more processors, and (4) controller memory storing one or more programs, which when executed by the one or more processors cause the host to perform or control performance of any of the methods B1 to B16 described herein.

[0075] (B19) In yet another aspect, any of the methods B1 to B16 described above are performed by a host system, coupled to one or more storage devices, the host system including means for performing any of the methods described herein.

[0076] (B21) In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a storage system, the one or more programs including instructions for performing any of the methods described herein.

[0077] (B22) In yet another aspect, a host system includes (1) an interface for operatively coupling to a storage system, (2) one or more processors, and (3) controller memory storing one or more programs, which when executed by the one or more processors cause the host system to perform or control performance of any of the methods B1 to B16 described herein.

[0078] In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a host system, the one or more programs including instructions for performing any of the methods described herein.

[0079] The various embodiments described herein include systems, methods and/or devices used to enable triggering a process to reduce declared capacity of a storage device in a multi-storage-device storage system. Some embodiments include systems, methods and/or devices to detect a trigger condition in accordance with one or more metrics of one or more storage devices of a plurality of storage devices of a storage system and enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of non-volatile memory of the respective storage device.

[0080] (C1) More specifically, some embodiments include a method of managing a storage system. In some embodiments, the method includes: (1) obtaining, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory, (2) detecting a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of one or more storage devices of the plurality of storage devices, and (3) enabling an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device.

[0081] (C1-1) In some embodiments of the method of C1, the method further includes: (1) prior to detecting the trigger condition, detecting a first wear condition of the non-volatile memory of the respective storage device, wherein a total storage capacity of the non-volatile memory of the respective storage device includes declared capacity and over-provisioning, and (2) in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the respective storage device without reducing declared capacity of the non-volatile memory of the respective storage device.

[0082] (C1-2) In some embodiments of the method of C1-1, detecting the trigger condition includes detecting a second wear condition distinct from the first wear condition.

[0083] (C2) In some embodiments of the method of C1 or C1-1 or C1-2, enabling the amelioration process associated with the detected trigger condition includes notifying a host to which the respective storage device is operatively coupled of the trigger condition.

[0084] (C3) In some embodiments of the method of C1 or C1-1 or C1-2, enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a query from a host to which the respective storage device is operatively coupled, and (2) in response to receiving the query, reporting the trigger condition.

[0085] (C4) In some embodiments of the method of C1 or C1-1 or C1-2, enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the respective storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition.

[0086] (C5) In some embodiments of the method of C1 or C1-1 or C1-2, enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the respective storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition.

[0087] (C6) In some embodiments of the method of any of C2 to C5, the host includes a client on behalf of which data is stored in the storage system.

[0088] (C7) In some embodiments of the method of any of C2 to C5, the host includes a storage system controller of the storage system.

[0089] (C8) In some embodiments of the method of any of C2 to C5, the host includes a cluster controller of the storage system.

[0090] (C9) In some embodiments of the method of any of C1 to C8, enabling the amelioration process associated with the detected trigger condition includes scheduling the amelioration process to be performed on the respective storage device.

[0091] (C10) In some embodiments of the method of any of C1 to C9, enabling the amelioration process associated with the detected trigger condition includes determining one or more parameters for the amelioration process.

[0092] (C11) In some embodiments of the method of C10, enabling the amelioration process associated with the detected trigger condition further includes reporting at least a subset of the one or more parameters for the amelioration process.

[0093] (C12) In some embodiments of the method of any of C1 to C11, obtaining one or more metrics of the respective storage device includes obtaining at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the respective storage device.

[0094] (C13) In some embodiments of the method of any of C1 to C12, the one or more metrics of the respective storage device include one or more status metrics corresponding to the respective storage device's ability to retain data.

[0095] (C14) In some embodiments of the method of any of C1 to C13, the one or more metrics of the respective storage device include one or more performance metrics corresponding to performance of the respective storage device.

[0096] (C15) In some embodiments of the method of any of C1 to C14, the one or more metrics of the respective storage device include one or more wear metrics corresponding to wear on the respective storage device.

[0097] (C16) In some embodiments of the method of any of C1 to C15, the one or more metrics of the respective storage device include one or more time metrics.

[0098] (C17) In some embodiments of the method of any of C1 to C16, the one or more metrics of the respective storage device include values of the one or more metrics from more than one time.

[0099] (C18) In some embodiments of the method of any of C1 to C17, the method further includes, after enabling the amelioration process: (1) re-evaluating the trigger condition in accordance with the one or more metrics of the respective storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, aborting the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device.

[00100] (C19) In some embodiments of the method of any of C1 to C18, the obtaining, the enabling, or both the obtaining and the enabling are performed by one or more subsystems of the storage system distinct from the plurality of storage devices.

[00101] (C20) In some embodiments of the method of any of C1 to C19, the respective storage device comprises one or more flash memory devices.

[00102] (C21) In another aspect, a storage system includes (1) non-volatile memory, (2) one or more processors, and (3) controller memory storing one or more programs, which when executed by the one or more processors cause the storage system to perform or control performance of any of the methods C1 to C20 described herein.

[00103] (C23) In yet another aspect, any of the methods C1 to C20 described above are performed by a storage system including means for performing any of the methods described herein.

[00104] (C25) In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a storage system, the one or more programs including instructions for performing any of the methods described herein.

[00105] (C26) In yet another aspect, a storage system includes (1) a plurality of storage devices, (2) one or more subsystems having one or more processors, and (3) memory storing one or more programs, which when executed by the one or more processors cause the one or more subsystems to perform or control performance of any of the methods C1 to C20 described herein.

[00106] In yet another aspect, a host system includes (1) an interface for operatively coupling to a storage system, (2) one or more processors, and (3) controller memory storing one or more programs, which when executed by the one or more processors cause the host system to perform or control performance of any of the methods C1 to C20 described herein.

[00107] In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a host system, the one or more programs including instructions for performing any of the methods described herein.

[00108] The various embodiments described herein include systems, methods and/or devices used to enable notification of a trigger condition to reduce declared capacity of a storage device. Some embodiments include systems, methods and/or devices to notify a host to which a storage device is operatively coupled of a trigger condition for reducing declared capacity of non-volatile memory of the storage device.

[00109] (D1) More specifically, some embodiments include a method of managing a storage system. In some embodiments, the method includes, at a storage device of the storage system, the storage device including non-volatile memory: (1) detecting a trigger condition for reducing declared capacity of the non-volatile memory of the storage device, and (2) notifying a host to which the storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the storage device. In some embodiments, the amelioration process is performed, at least in part, by an apparatus other than the storage device (e.g., performed at least in part by the host, or by a storage system controller or by a cluster controller of a data storage system that includes at least one storage device distinct from the storage device).

[00110] (D1-1) In some embodiments of the method of D1, the method further includes: (1) prior to detecting the trigger condition, detecting a first wear condition of the

non-volatile memory of the storage device, wherein a total storage capacity of the non-volatile memory of the storage device includes declared capacity and over-provisioning, and (2) in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the storage device without reducing declared capacity of the non-volatile memory of the storage device.

[00111] (D1-2) In some embodiments of the method of D1-1, detecting the trigger condition includes detecting a second wear condition distinct from the first wear condition.

[00112] (D2) In some embodiments of the method of D1 or D1-1 or D1-2, the host includes a client on behalf of which data is stored in the storage system.

[00113] (D3) In some embodiments of the method of D1 or D1-1 or D1-2, the host includes a storage system controller of the storage system.

[00114] (D4) In some embodiments of the method of D1 or D1-1 or D1-2, the host includes a cluster controller of the storage system.

[00115] (D5) In some embodiments of the method of any of D1 to D4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes notifying the host with an unsolicited communication.

[00116] (D6) In some embodiments of the method of any of D1 to D4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes: (1) receiving a query from the host, and (2) in response to receiving the query, reporting the trigger condition.

[00117] (D7) In some embodiments of the method of any of D1 to D4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes: (1) receiving a command from the host, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition.

[00118] (D8) In some embodiments of the method of any of D1 to D4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes: (1) receiving a command from the host, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition.

[00119] (D9) In some embodiments of the method of any of D1 to D8, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device further includes notifying the host that the storage device is in read-only mode.

[00120] (D10) In some embodiments of the method of any of D1 to D9, the trigger condition is detected by the storage device in accordance with one or more metrics of the storage device.

[00121] (D11) In some embodiments of the method of D10, the method further includes, after notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device: (1) re-evaluating the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, notifying the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device.

[00122] (D12) In some embodiments of the method of any of D1 to D11, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device includes a process to reduce utilization of the non-volatile memory of the storage device.

[00123] (D13) In some embodiments of the method of any of D1 to D12, the storage device comprises one or more flash memory devices.

[00124] (D14) In another aspect, a storage device includes (1) non-volatile memory (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) one or more processors, and (3) controller memory (e.g., non-volatile memory or volatile memory in or coupled to the controller) storing one or more programs, which when executed by the one or more processors cause the storage device to perform or control performance of any of the methods D1 to D13 described herein.

[00125] (D16) In yet another aspect, any of the methods D1 to D13 described above are performed by a storage device including means for performing any of the methods described herein.

[00126] (D18) In yet another aspect, a storage system includes (1) a storage medium (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) one or more processors, and (3) memory (e.g., non-volatile memory or volatile memory in the storage system) storing one or more programs, which when executed by the one or more

processors cause the storage system to perform or control performance of any of the methods D1 to D13 described herein.

[00127] (D19) In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a storage device, the one or more programs including instructions for performing any of the methods described herein.

[00128] The various embodiments described herein include systems, methods and/or devices used to enable notification of a trigger condition to reduce declared capacity of a storage device in a multi-storage-device storage system. Some embodiments include systems, methods and/or devices to notify a host to which a respective storage device of a plurality of storage devices of a storage system is operatively coupled of a trigger condition for reducing declared capacity of non-volatile memory of the respective storage device.

[00129] (E1) More specifically, some embodiments include a method of managing a storage system. In some embodiments, the method includes: (1) obtaining, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory, (2) detecting a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of two or more of the storage devices of the plurality of storage devices in the storage system, and (3) notifying a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device. In some embodiments, or in some circumstances, the notification of the trigger condition causes performance of the amelioration process to be enabled.

[00130] (E1-1) In some embodiments of the method of E1, the method further includes: (1) prior to detecting the trigger condition, detecting a first wear condition of the non-volatile memory of the respective storage device, wherein a total storage capacity of the non-volatile memory of the respective storage device includes declared capacity and over-provisioning; and (2) in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the respective storage

device without reducing declared capacity of the non-volatile memory of the respective storage device.

[00131] (E1-2) In some embodiments of the method of E1-1, detecting the trigger condition includes detecting a second wear condition distinct from the first wear condition.

[00132] (E2) In some embodiments of the method of E1 or E1-1 or E1-2, the host includes a client on behalf of which data is stored in the storage system.

[00133] (E3) In some embodiments of the method of E1 or E1-1 or E1-2, the host includes a storage system controller of the storage system.

[00134] (E4) In some embodiments of the method of E1 or E1-1 or E1-2, the host includes a cluster controller of the storage system.

[00135] (E5) In some embodiments of the method of any of E1 to E4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying the host with an unsolicited communication.

[00136] (E6) In some embodiments of the method of any of E1 to E4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes: (1) receiving a query from the host, and (2) in response to receiving the query, reporting the trigger condition.

[00137] (E7) In some embodiments of the method of any of E1 to E4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes: (1) receiving a command from the host, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition.

[00138] (E8) In some embodiments of the method of any of E1 to E4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes: (1) receiving a command from the host, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition.

[00139] (E9) In some embodiments of the method of any of E1 to E4, notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying the host that the respective storage device is in read-only mode.

[00140] (E10) In some embodiments of the method of any of E1 to E9, the method further includes, after notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device: (1) re-evaluating the trigger condition in accordance with the one or more metrics of the two or more storage devices of the plurality of storage devices in the storage system, and (2) in accordance with a determination that the trigger condition is no longer valid, notifying the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device.

[00141] (E11) In some embodiments of the method of any of E1 to E10, the obtaining, the notifying, or both the obtaining and the notifying are performed by one or more subsystems of the storage system distinct from the plurality of storage devices.

[00142] (E12) In some embodiments of the method of any of E1 to E11, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device includes a process to reduce utilization of the non-volatile memory of the respective storage device.

[00143] (E13) In some embodiments of the method of any of E1 to E12, the respective storage device comprises one or more flash memory devices.

[00144] (E14) In another aspect, a storage system includes (1) non-volatile memory (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) one or more processors, and (3) controller memory (e.g., non-volatile memory or volatile memory in or coupled to a controller of the storage system) storing one or more programs, which when executed by the one or more processors cause the storage system to perform or control performance of any of the methods E1 to E13 described herein.

[00145] (E16) In yet another aspect, any of the methods E1 to E13 described above are performed by a storage system including means for performing any of the methods described herein.

[00146] (E18) In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a storage system, the one or more programs including instructions for performing any of the methods described herein.

[00147] (E19) In yet another aspect, a storage system includes (1) a plurality of storage devices, (2) one or more subsystems having one or more processors, and (3) memory (e.g.,

non-volatile memory or volatile memory in the storage system) storing one or more programs, which when executed by the one or more processors cause the one or more subsystems to perform or control performance of any of the methods E1 to E13 described herein.

[00148] In yet another aspect, a host system includes (1) an interface for operatively coupling to a storage system, (2) one or more processors, and (3) controller memory storing one or more programs, which when executed by the one or more processors cause the host system to perform or control performance of any of the methods E1 to E13 described herein.

[00149] In yet another aspect, some embodiments include a non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a host system, the one or more programs including instructions for performing any of the methods described herein.

[00150] Numerous details are described herein in order to provide a thorough understanding of the example embodiments illustrated in the accompanying drawings. However, some embodiments may be practiced without many of the specific details, and the scope of the claims is only limited by those features and aspects specifically recited in the claims. Furthermore, well-known methods, components, and circuits have not been described in exhaustive detail so as not to unnecessarily obscure pertinent aspects of the embodiments described herein.

[00151] Data storage systems, including those described below, use a variety of techniques to avoid data loss caused by a variety of failure mechanisms, including storage media failure, communication failures, and failures at the system and subsystem level. A common feature of these mechanisms is the use of data redundancy to protect data, to compensate for actual and potential data errors (e.g., media errors, lost data, transmission errors, inaccessible data, etc.). One class of redundancy mechanisms is known as error correction codes (ECCs). Numerous types of error correction codes are well known (e.g., BCH, LDPC, Reed-Solomon, etc.), as are numerous schemes for storing them with or in conjunction with the data that is being protected. Another class of redundancy mechanisms is erasure codes (e.g., pyramid, fountain, partial MDS, locally repairable, simple regenerating, etc.)

[00152] Another type or level of redundancy mechanism is typically called RAID (redundant array of independent disks), even when the storage media are not "disks" in the

traditional sense. There are multiple forms of RAID, or RAID schemes, providing different levels of data protection (e.g., RAID-1, RAID-5, RAID-6, RAID-10, etc.). Typically, in systems that use RAID, "stripes" of data stored in multiple distinct storage locations are treated as a set, and stored with sufficient redundant data that any data in a stripe that would have been lost, in a partial or complete failure of any one of the storage locations, is recovered using the other data in the stripe, possibly including the redundant data.

[00153] A third type of redundancy mechanism is replication of data to multiple storage locations, typically in distinct failure domains. Systems implementing this type of redundancy mechanism typically store three or more replicas of each data set or data item. Typically either each replica is in a distinct failure domain from the other replicas, or at least one replica is in a distinct failure domain from the other replicas.

[00154] The embodiments described below work in conjunction with the data redundancy mechanisms described above (used alone or in combination). Some of the data storage systems described below have an architecture or configuration designed to implement a particular redundancy mechanism. Furthermore, some of the embodiments described below may utilize more than one of the redundancy mechanisms described above, either alone or in combination. Furthermore, some of the embodiments are able to store data encoded with different redundancy mechanisms simultaneously. Furthermore, even within a single mechanism, the selection of parameters (i.e., codeword size relative to data size) may vary dynamically. Hence, altering the redundancy mechanism directly affects the amount of data stored and in turn the utilization.

[00155] Figure 1A is a block diagram illustrating data storage system 100, in accordance with some embodiments. While some example features are illustrated, various other features have not been illustrated for the sake of brevity and so as not to obscure pertinent aspects of the example embodiments disclosed herein. To that end, as a non-limiting example, data storage system 100 includes a storage device 120, which includes a storage controller 124 and a storage medium 130, and is used in conjunction with or includes a computer system 110. In some embodiments, storage medium 130 is a single flash memory device while in other embodiments storage medium 130 includes a plurality of flash memory devices. In some embodiments, storage medium 130 is NAND-type flash memory or NOR-type flash memory. In some embodiments, storage medium 130 includes one or more three-dimensional (3D) memory devices, as further defined herein. Further, in some embodiments storage controller 124 is a solid-state drive (SSD) controller. However, other types of storage

media may be included in accordance with aspects of a wide variety of embodiments (e.g., PCRAM, ReRAM, STT-RAM, etc.). In some embodiments, a flash memory device includes one or more flash memory die, one or more flash memory packages, one or more flash memory channels or the like. In some embodiments, data storage system 100 can contain one or more storage device 120s.

[00156] Computer system 110 is coupled to storage controller 124 through data connections 101. However, in some embodiments computer system 110 includes storage controller 124, or a portion of storage controller 124, as a component and/or a subsystem. For example, in some embodiments, some or all of the functionality of storage controller 124 is implemented by software executed on computer system 110. Computer system 110 may be any suitable computer device, such as a computer, a laptop computer, a tablet device, a netbook, an internet kiosk, a personal digital assistant, a mobile phone, a smart phone, a gaming device, a computer server, or any other computing device. Computer system 110 is sometimes called a host, host system, client, or client system. In some embodiments, computer system 110 is a server system, such as a server system in a data center. In some embodiments, computer system 110 includes one or more processors, one or more types of memory, a display and/or other user interface components such as a keyboard, a touch screen display, a mouse, a track-pad, a digital camera and/or any number of supplemental devices to add functionality. In some embodiments, computer system 110 does not have a display and other user interface components.

[00157] Storage medium 130 is coupled to storage controller 124 through connections 103. Connections 103 are sometimes called data connections, but typically convey commands in addition to data, and optionally convey metadata, error correction information and/or other information in addition to data values to be stored in storage medium 130 and data values read from storage medium 130. In some embodiments, however, storage controller 124 and storage medium 130 are included in the same device (i.e., an integral device) as components thereof. Furthermore, in some embodiments, storage controller 124 and storage medium 130 are embedded in a host device (e.g., computer system 110), such as a mobile device, tablet, other computer or computer controlled device, and the methods described herein are performed, at least in part, by the embedded memory controller. Storage medium 130 may include any number (i.e., one or more) of memory devices including, without limitation, non-volatile semiconductor memory devices, such as flash memory device(s). For example, flash memory device(s) can be configured for enterprise storage

suitable for applications such as cloud computing, for database applications, primary and/or secondary storage, or for caching data stored (or to be stored) in secondary storage, such as hard disk drives. Additionally and/or alternatively, flash memory device(s) can also be configured for relatively smaller-scale applications such as personal flash drives or hard-disk replacements for personal, laptop, and tablet computers. In some embodiments, storage medium 130 includes one or more three-dimensional (3D) memory devices, as further defined herein.

[00158] Storage medium 130 is divided into a number of addressable and individually selectable blocks, such as selectable portion 131. In some embodiments, the individually selectable blocks are the minimum size erasable units in a flash memory device. In other words, each block contains the minimum number of memory cells that can be erased simultaneously. Each block is usually further divided into a plurality of pages and/or word lines, where each page or word line is typically an instance of the smallest individually accessible (readable) portion in a block. In some embodiments (e.g., using some types of flash memory), the smallest individually accessible unit of a data set, however, is a sector, which is a subunit of a page. That is, a block includes a plurality of pages, each page contains a plurality of sectors, and each sector is the minimum unit of data for reading data from the flash memory device.

[00159] As noted above, while data storage densities of non-volatile semiconductor memory devices are generally increasing, a drawback of increasing storage density is that the stored data is more prone to being stored and/or read erroneously. In some embodiments, error control coding can be utilized to limit the number of uncorrectable errors that are introduced by electrical fluctuations, defects in the storage medium, operating conditions, device history, write-read circuitry, etc., or a combination of these and various other factors.

[00160] In some embodiments, storage controller 124 includes a management module 121-1, a host interface 129, a storage medium I/O interface 128, and additional module(s) 125. Storage controller 124 may include various additional features that have not been illustrated for the sake of brevity and so as not to obscure pertinent features of the example embodiments disclosed herein, and a different arrangement of features may be possible. Host interface 129 provides an interface to computer system 110 through data connections 101. Similarly, storage medium I/O 128 provides an interface to storage medium 130 though connections 103. In some embodiments, storage medium I/O 128 includes read and write

circuitry, including circuitry capable of providing reading signals to storage medium 130 (e.g., reading threshold voltages for NAND-type flash memory).

[00161] In some embodiments, management module 121-1 includes one or more processing units (CPUs, also sometimes called processors) 122-1 configured to execute instructions in one or more programs (e.g., in management module 121-1). In some embodiments, the one or more CPUs 122-1 are shared by one or more components within, and in some cases, beyond the function of storage controller 124. Management module 121-1 is coupled to host interface 129, additional module(s) 125 and storage medium I/O 128 in order to coordinate the operation of these components. In some embodiments, one or more modules of management module 121-1 are implemented in management module 121-2 of computer system 110. In some embodiments, one or more processors of computer system 110 (not shown) are configured to execute instructions in one or more programs (e.g., in management module 121-2). Management module 121-2 is coupled to storage device 120 in order to manage the operation of storage device 120.

[00162] Additional module(s) 125 are coupled to storage medium I/O 128, host interface 129, and management module 121-1. As an example, additional module(s) 125 may include an error control module to limit the number of uncorrectable errors inadvertently introduced into data during writes to memory or reads from memory. In some embodiments, additional module(s) 125 are executed in software by the one or more CPUs 122-1 of management module 121-1, and, in other embodiments, additional module(s) 125 are implemented in whole or in part using special purpose circuitry (e.g., to perform encoding and decoding functions). In some embodiments, additional module(s) 125 are implemented in whole or in part by software executed on computer system 110.

[00163] In some embodiments, an error control module, included in additional module(s) 125, includes an encoder and a decoder. In some embodiments, the encoder encodes data by applying an error control code to produce a codeword, which is subsequently stored in storage medium 130. When the encoded data (e.g., one or more codewords) is read from storage medium 130, the decoder applies a decoding process to the encoded data to recover the data, and to correct errors in the recovered data within the error correcting capability of the error control code. Those skilled in the art will appreciate that various error control codes have different error detection and correction capacities, and that particular codes are selected for various applications for reasons beyond the scope of this disclosure. As such, an exhaustive review of the various types of error control codes is not provided

herein. Moreover, those skilled in the art will appreciate that each type or family of error control codes may have encoding and decoding algorithms that are particular to the type or family of error control codes. On the other hand, some algorithms may be utilized at least to some extent in the decoding of a number of different types or families of error control codes. As such, for the sake of brevity, an exhaustive description of the various types of encoding and decoding algorithms generally available and known to those skilled in the art is not provided herein.

[00164] In some embodiments, during a write operation, host interface 129 receives data to be stored in storage medium 130 from computer system 110. The data received by host interface 129 is made available to an encoder (e.g., in additional module(s) 125), which encodes the data to produce one or more codewords. The one or more codewords are made available to storage medium I/O 128, which transfers the one or more codewords to storage medium 130 in a manner dependent on the type of storage medium being utilized.

[00165] In some embodiments, a read operation is initiated when computer system (host) 110 sends one or more host read commands (e.g., via data connections 101, or alternatively a separate control line or bus) to storage controller 124 requesting data from storage medium 130. Storage controller 124 sends one or more read access commands to storage medium 130, via storage medium I/O 128, to obtain raw read data in accordance with memory locations (addresses) specified by the one or more host read commands. Storage medium I/O 128 provides the raw read data (e.g., comprising one or more codewords) to a decoder (e.g., in additional module(s) 125). If the decoding is successful, the decoded data is provided to host interface 129, where the decoded data is made available to computer system 110. In some embodiments, if the decoding is not successful, storage controller 124 may resort to a number of remedial actions or provide an indication of an irresolvable error condition.

[00166] As explained above, a storage medium (e.g., storage medium 130) is divided into a number of addressable and individually selectable blocks and each block is optionally (but typically) further divided into a plurality of pages and/or word lines and/or sectors. While erasure of a storage medium is performed on a block basis, in many embodiments, reading and programming of the storage medium is performed on a smaller subunit of a block (e.g., on a page basis, word line basis, or sector basis). In some embodiments, the smaller subunit of a block consists of multiple memory cells (e.g., single-level cells or multi-level cells). In some embodiments, programming is performed on an entire page. In some

embodiments, a multi-level cell (MLC) NAND flash typically has four possible states per cell, yielding two bits of information per cell. Further, in some embodiments, a MLC NAND has two page types: (1) a lower page (sometimes called fast page), and (2) an upper page (sometimes called slow page). In some embodiments, a triple-level cell (TLC) NAND flash has eight possible states per cell, yielding three bits of information per cell. Although the description herein uses TLC, MLC, and SLC as examples, those skilled in the art will appreciate that the embodiments described herein may be extended to memory cells that have more than eight possible states per cell, yielding more than three bits of information per cell.

[00167] The encoding format of the storage media (i.e., TLC, MLC, or SLC and/or a chose data redundancy mechanism) is a choice made when data is actually written to the storage media. Often in this specification there is described an event, condition, or process that is said to set the encoding format, alter the encoding format of the storage media, etc. It should be recognized that the actual process may involve multiple steps, e.g., erasure of the previous contents of the storage media followed by the data being written using the new encoding format and that these operations may be separated in time from the initiating event, condition or procedure.

[00168] As an example, if data is written to a storage medium in pages, but the storage medium is erased in blocks, pages in the storage medium may contain invalid (e.g., stale) data, but those pages cannot be overwritten until the whole block containing those pages is erased. In order to write to the pages with invalid data, the pages (if any) with valid data in that block are read and re-written to a new block and the old block is erased (or put on a queue for erasing). This process is called garbage collection. After garbage collection, the new block contains the pages with valid data and may have free pages that are available for new data to be written, and the old block can be erased so as to be available for new data to be written. Since flash memory can only be programmed and erased a limited number of times, the efficiency of the algorithm used to pick the next block(s) to re-write and erase has a significant impact on the lifetime and reliability of flash-based storage systems.

[00169] Figure 1B is a block diagram illustrating data storage system 140, in accordance with some embodiments. While some example features are illustrated, various other features have not been illustrated for the sake of brevity and so as not to obscure pertinent aspects of the example embodiments disclosed herein. To that end, as a non-limiting example, data storage system 140 (sometimes called a scale-up storage system, a single node storage system, etc.) includes a plurality of storage devices 160 (e.g., storage

devices 160-1 to 160-m) and a storage system controller 150, and is used in conjunction with a computer system 142. In some embodiments, storage devices 160 include management modules 161 (e.g., storage device 160-1 includes management module 161-1 and storage device 160-m includes management module 161-m). Some of the features described above with respect to storage device 120 (Figure 1A) and management module 121-1 (Figure 1A) are applicable to storage devices 160 and management modules 161, respectively, and for sake of brevity and simplicity, the details are not repeated here.

[00170] Computer system 142 is coupled to storage system controller 150 through connections 141. However, in some embodiments computer system 142 includes a part of or the entire storage system controller 150 as a component and/or a subsystem. For example, in some embodiments, some or all of the functionality of storage system controller 150 is implemented by software executed on computer system 142. Computer system 142 may be any suitable computer device, such as a computer, a laptop computer, a tablet device, a netbook, an internet kiosk, a personal digital assistant, a mobile phone, a smart phone, a gaming device, a computer server, or any other computing device. In some embodiments, computer system 142 is a server system, such as a server system in a data center. Computer system 142 is sometimes called a host, host system, client, or client system. In some embodiments, computer system 142 includes one or more processors, one or more types of memory, a display and/or other user interface components such as a keyboard, a touch screen display, a mouse, a track-pad, a digital camera and/or any number of supplemental devices to add functionality. In some embodiments, computer system 142 does not have a display and other user interface components.

[00171] In some embodiments, storage system controller 150 includes a system management module 151-1, and additional module(s) 155. Storage system controller 150 may include various additional features that have not been illustrated for the sake of brevity and so as not to obscure pertinent features of the example embodiments disclosed herein, and a different arrangement of features may be possible. For example, in some embodiments, storage system controller 150 additionally includes an interface for each of the storage devices 160 coupled to storage system controller 150. Storage devices 160 are coupled to storage system controller 150 through connections 143 (e.g., storage device 160-1 is coupled to storage system controller 150 through connections 143-1 and storage device 160-m is coupled to storage system controller 150 through connections 143-m). In some embodiments, connections 143-1 through 143-m are implemented as a communication media over which

commands and data are communicated using a protocol such as SCSI, SATA, Infiniband, Ethernet, Token Ring, or the like.

[00172] In some embodiments, system management module 151-1 includes one or more processing units (CPUs, also sometimes called processors) 152-1 configured to execute instructions in one or more programs (e.g., in system management module 151-1). In some embodiments, the one or more CPUs 152-1 are shared by one or more components within, and in some cases, beyond the function of storage system controller 150. System management module 151-1 is coupled to additional module(s) 155 in order to coordinate the operation of these components. In some embodiments, one or more modules of system management module 151-1 are implemented in system management module 151-2 of computer system 142 (sometimes called a host, host system, client, or client system). In some embodiments, one or more processors (sometimes called CPUs or processing units) of computer system 142 (not shown) are configured to execute instructions in one or more programs (e.g., in system management module 151-2). System management module 151-2 is coupled to storage system controller 150 in order to manage the operation of storage system controller 150.

[00173] Additional module(s) 155 are coupled to system management module 151-1. In some embodiments, additional module(s) 155 are executed in software by the one or more CPUs 152-1 of system management module 151-1, and, in other embodiments, additional module(s) 155 are implemented in whole or in part using special purpose circuitry. In some embodiments, additional module(s) 155 are implemented in whole or in part by software executed on computer system 142.

[00174] During a write operation, storage system controller 150 receives data to be stored in storage devices 160 from computer system 142 (sometimes called a host, host system, client, or client system). In some embodiments, storage system controller 150 maps a virtual logical address from computer system 142 to an address, which determines or identifies the one or more of storage devices 160 to which to write the data.

[00175] A read operation is initiated when computer system 142 sends one or more host read commands to storage system controller 150 requesting data from storage devices 160. In some embodiments, storage system controller 150 maps a virtual logical address from computer system 142 to an address, which determines or identifies the one or more of storage devices 160 from which to read the requested data.

[00176] Figure 1C is a block diagram illustrating an implementation of data storage system 170, in accordance with some embodiments. While some example features are illustrated, various other features have not been illustrated for the sake of brevity and so as not to obscure pertinent aspects of the example embodiments disclosed herein. To that end, as a non-limiting example, data storage system 170 (sometimes called a scale-out storage system, a multiple node storage system or a storage cluster system) includes a plurality of storage subsystems 192 (e.g., storage subsystems 192-1 to 192-s) and a cluster controller 180, and is used in conjunction with a computer system 172. In some embodiments, storage subsystems 192 include storage system controllers 190 and storage devices 194 (e.g., storage subsystem 192-1 includes storage system controller 190-1 and storage devices 194-1 through 194-n). Some of the features described above with respect to data storage system 140 (Figure 1B) are applicable to storage subsystems 192, and for sake of brevity, the details are not repeated here. In some embodiments there may be a plurality of cluster controllers 180 that may communicate with each other to coordinate their activities.

[00177] Computer system 172 is coupled to cluster controller 180 through connections 171. However, in some embodiments computer system 172 includes cluster controller 180 as a component and/or a subsystem. For example, in some embodiments, some or all of the functionality of cluster controller 180 is implemented by software executed on computer system 172. Computer system 172 may be any suitable computer device, such as a computer, a laptop computer, a tablet device, a netbook, an internet kiosk, a personal digital assistant, a mobile phone, a smart phone, a gaming device, a computer server, or any other computing device. In some embodiments, computer system 172 is a server system, such as a server system in a data center. Computer system 172 is sometimes called a host, host system, client, or client system. In some embodiments, computer system 172 includes one or more processors, one or more types of memory, a display and/or other user interface components such as a keyboard, a touch screen display, a mouse, a track-pad, a digital camera and/or any number of supplemental devices to add functionality. In some embodiments, computer system 172 does not have a display and other user interface components.

[00178] In some embodiments, cluster controller 180 includes a cluster management module 181-1, and additional module(s) 185. Cluster controller 180 may include various additional features that have not been illustrated for the sake of brevity and so as not to obscure pertinent features of the example embodiments disclosed herein, and a different arrangement of features may be possible. For example, in some embodiments, cluster

controller 180 additionally includes an interface for each of the storage subsystems 192 coupled to cluster controller 180. Storage subsystems 192 are coupled to cluster controller 180 through connections 173 (e.g., storage subsystem 192-1 is coupled to cluster controller 180 through connections 173-1 and storage subsystem 192-s is coupled to cluster controller 180 through connections 173-s). In some embodiments, connections 173 may be implemented as a shared communication network, e.g., Token Ring, Ethernet, Infiniband, etc.

[00179] In some embodiments, cluster management module 181-1 includes one or more processing units (CPUs, also sometimes called processors) 182-1 configured to execute instructions in one or more programs (e.g., in cluster management module 181-1). In some embodiments, the one or more CPUs 182-1 are shared by one or more components within, and in some cases, beyond the function of cluster controller 180. Cluster management module 181-1 is coupled to additional module(s) 185 in order to coordinate the operation of these components. In some embodiments, one or more modules of cluster management module 181-1 are implemented in cluster management module 181-2 of computer system 172 (sometimes called a host, host system, client, or client system). In some embodiments, one or more processors (sometimes called CPUs or processing units) of computer system 172 (not shown) are configured to execute instructions in one or more programs (e.g., in cluster management module 181-2). Cluster management module 181-2 is coupled to cluster controller 180 in order to manage the operation of cluster controller 180.

[00180] Additional module(s) 185 are coupled to cluster management module 181-1. In some embodiments, additional module(s) 185 are executed in software by the one or more CPUs 182-1 of cluster management module 181-1, and, in other embodiments, additional module(s) 185 are implemented in whole or in part using special purpose circuitry. In some embodiments, additional module(s) 185 are implemented in whole or in part by software executed on computer system 172.

[00181] In some embodiments, during a write operation, cluster controller 180 receives data to be stored in storage subsystems 192 from computer system 172 (sometimes called a host, host system, client, or client system). In some embodiments, cluster controller 180 maps a virtual logical address from computer system 172 to an address format understandable by storage subsystems 192 and to identify a storage subsystem of storage subsystems 192 to which to write the data. In some embodiments, cluster controller 180 may convert the data to be stored into a plurality of sets of data, each set of data is stored on one storage subsystem of storage subsystems 192. In one embodiment, the conversion process may be as simple as a

partitioning of the data to be stored. In another embodiment, the conversion process may redundantly encode the data to be stored so as to provide enhanced data integrity and access in the face of failures of one or more storage subsystems of storage subsystems 192 or communication thereto.

[00182] In some embodiments, a read operation is initiated when computer system 172 sends one or more host read commands to cluster controller 180 requesting data from storage subsystems 192. In some embodiments, cluster controller 180 maps a virtual logical address from computer system 172 to an address format understandable by storage subsystems 192, to determine or identify the storage subsystem of storage subsystems 192 from which to read the requested data. In some embodiments, more than one storage subsystem of storage subsystems 192 may have data read in order to satisfy the read operation, e.g. for data reconstruction.

[00183] As used herein, the term "host" or "host system" may be construed to mean (1) a computer system (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, or computer system 172, Figure 1C) on behalf of which data is stored in a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, or data storage system 170, Figure 1C), (2) a storage system controller (e.g., storage system controller 150, Figure 1B) of a storage system (e.g., data storage system 140, Figure 1B), (3) a cluster controller (e.g., cluster controller 180, Figure 1C) of a storage system (e.g., data storage system 170, Figure 1C), and/or (4) any computing entity (e.g., a computer, a process running on a computer, a mobile phone, an internet kiosk, a tablet computer, a laptop computer, a desktop computer, a server computer, etc.) that is operatively coupled either directly or indirectly to a storage system, depending on the context. For example, in some circumstances, with respect to data storage system 140 (Figure 1B), the term "host" may refer to computer system 142 or storage system controller 150, depending on the context. As another example, in some circumstances, with respect to data storage system 170 (Figure 1C), the term "host" may refer to computer system 172 or cluster controller 180, depending on the context. Further, in some contexts, the host is or includes a client or client system, on behalf of which data is stored in a storage system.

[00184] Figure 2A-1 is a block diagram illustrating a management module 121-1, in accordance with some embodiments, as shown in Figure 1A. Management module 121-1 typically includes one or more processing units (sometimes called CPUs or processors) 122-1 for executing modules, programs and/or instructions stored in memory 206-1 and thereby

performing processing operations, memory 206-1 (sometimes called controller memory), and one or more communication buses 208-1 for interconnecting these components. The one or more communication buses 208-1 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. Management module 121-1 is coupled to host interface 129, additional module(s) 125, and storage medium I/O 128 by the one or more communication buses 208-1. Memory 206-1 includes high-speed random access memory, such as DRAM, SRAM, DDR RAM or other random access solid state memory devices, and may include non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. Memory 206-1 optionally includes one or more storage devices remotely located from the CPU(s) 122-1. Memory 206-1, or alternatively the non-volatile memory device(s) within memory 206-1, comprises a non-transitory computer readable storage medium. In some embodiments, memory 206-1, or the non-transitory computer readable storage medium of memory 206-1 stores the following programs, modules, and data structures, or a subset or superset thereof:

- translation table 212-1 that is used for mapping logical addresses to physical addresses (e.g., in some embodiments, translation table 212-1 includes mapping table 402, Figure 4);
- data read module 214-1 that is used for reading data from one or more codewords, pages or blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- data write module 216-1 that is used for writing data to one or more codewords, pages or blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- data erase module 218-1 that is used for erasing data from one or more blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- garbage collection module 220-1 that is used for garbage collection for one or more blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- metrics module 222-1 that is used for generating and/or obtaining one or more metrics of a storage device (e.g., storage device 120, Figure 1A);
- trigger detection module 224-1 that is used for detecting a trigger condition (e.g., in accordance with one or more metrics of a storage device);
- enabling module 226-1 that is used for enabling an amelioration process associated with a trigger condition (e.g., detected by trigger detection module 224-1);

- notification module 228-1 that is used for notifying a host to which a storage device is operatively coupled of a trigger condition (e.g., detected by trigger detection module 224-1) and/or of an absence of the trigger condition;
- amelioration module 230-1 that is used for performing an amelioration process to reduce declared capacity of non-volatile memory of a storage device (e.g., storage device 120, Figure 1A), optionally including:
    - detection module 231-1 that is used for detecting an amelioration trigger for reducing declared capacity of the non-volatile memory of the storage device;
    - utilization module 232-1 that is used for reducing utilization of the non-volatile memory of the storage device; and
    - capacity module 234-1 that is used for reducing declared capacity of the non-volatile memory of the storage device.

[00185] Each of the above identified elements may be stored in one or more of the previously mentioned memory devices, and corresponds to a set of instructions for performing a function described above. The above identified modules or programs (i.e., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various embodiments. In some embodiments, memory 206-1 may store a subset of the modules and data structures identified above. Furthermore, memory 206-1 may store additional modules and data structures not described above. In some embodiments, the programs, modules, and data structures stored in memory 206-1, or the non-transitory computer readable storage medium of memory 206-1, provide instructions for implementing some of the methods described below. In some embodiments, some or all of these modules may be implemented with specialized hardware circuits that subsume part or all of the module functionality.

[00186] Although Figure 2A-1 shows management module 121-1 in accordance with some embodiments, Figure 2A-1 is intended more as a functional description of the various features which may be present in management module 121-1 than as a structural schematic of the embodiments described herein. In practice, and as recognized by those of ordinary skill in the art, the programs, modules, and data structures shown separately could be combined and some programs, modules, and data structures could be separated.

[00187] Figure 2A-2 is a block diagram illustrating a management module 121-2 of computer system 110 (Figure 1A), in accordance with some embodiments. Management module 121-2 typically includes one or more processing units (sometimes called CPUs or processors) 122-2 for executing modules, programs and/or instructions stored in memory 206-2 and thereby performing processing operations, memory 206-2, and one or more communication buses 208-2 for interconnecting these components. The one or more communication buses 208-2 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. Management module 121-2 is coupled to storage device 120 by the one or more communication buses 208-2. Memory 206-2 (sometimes called host memory) includes high-speed random access memory, such as DRAM, SRAM, DDR RAM or other random access solid state memory devices, and may include non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. Memory 206-2 optionally includes one or more storage devices remotely located from the CPU(s) 122-2. Memory 206-2, or alternatively the non-volatile memory device(s) within memory 206-2, comprises a non-transitory computer readable storage medium. In some embodiments, memory 206-2, or the non-transitory computer readable storage medium of memory 206-2 stores the following programs, modules, and data structures, or a subset or superset thereof:

- translation table 212-2 that is used for mapping logical addresses to physical addresses (e.g., in some embodiments, translation table 212-2 includes mapping table 402, Figure 4);
- data read module 214-2 that is used for reading data from one or more codewords, pages or blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- data write module 216-2 that is used for writing data to one or more codewords, pages or blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- data erase module 218-2 that is used for erasing data from one or more blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- garbage collection module 220-2 that is used for garbage collection for one or more blocks in a storage medium (e.g., storage medium 130, Figure 1A);
- metrics module 222-2 that is used for generating and/or obtaining one or more metrics of a storage device (e.g., storage device 120, Figure 1A);

- trigger detection module 224-2 that is used for detecting a trigger condition (e.g., in accordance with one or more metrics of a storage device);
- enabling module 226-2 that is used for enabling an amelioration process associated with a trigger condition (e.g., detected by trigger detection module 224-2);
- notification module 228-2 that is used for notifying an application, module or process, of the host (i.e., computer system 110, Figure 1A), of a trigger condition (e.g., detected by trigger detection module 224-2) and/or of an absence of the trigger condition;
- amelioration module 230-2 that is used for performing an amelioration process to reduce declared capacity of non-volatile memory of a storage device (e.g., storage device 120, Figure 1A), optionally including:
    - detection module 231-2 that is used for detecting an amelioration trigger for reducing declared capacity of the non-volatile memory of the storage device;
    - utilization module 232-2 that is used for reducing utilization of the non-volatile memory of the storage device; and
    - capacity module 234-2 that is used for reducing declared capacity of the non-volatile memory of the storage device.

[00188] Each of the above identified elements may be stored in one or more of the previously mentioned memory devices, and corresponds to a set of instructions for performing a function described above. The above identified modules or programs (i.e., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various embodiments. In some embodiments, memory 206-2 may store a subset of the modules and data structures identified above. Furthermore, memory 206-2 may store additional modules and data structures not described above. In some embodiments, the programs, modules, and data structures stored in memory 206-2, or the non-transitory computer readable storage medium of memory 206-2, provide instructions for implementing some of the methods described below. In some embodiments, some or all of these modules may be implemented with specialized hardware circuits that subsume part or all of the module functionality.

[00189] Although Figure 2A-2 shows management module 121-2 in accordance with some embodiments, Figure 2A-2 is intended more as a functional description of the various

features which may be present in management module 121-2 than as a structural schematic of the embodiments described herein. In practice, and as recognized by those of ordinary skill in the art, the programs, modules, and data structures shown separately could be combined and some programs, modules, and data structures could be separated.

[00190] Figure 2B-1 is a block diagram illustrating a system management module 151-1, in accordance with some embodiments, e.g., embodiments in which the system management module is in a storage system controller, as shown in Figure 1B. System management module 151-1 typically includes one or more processing units (sometimes called CPUs or processors) 152-1 for executing modules, programs and/or instructions stored in memory 246-1 and thereby performing processing operations, memory 246-1 (sometimes called storage system controller memory or controller memory), and one or more communication buses 248-1 for interconnecting these components. The one or more communication buses 248-1 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. System management module 151-1 is coupled to additional module(s) 155 by the one or more communication buses 248-1. Memory 246-1 includes high-speed random access memory, such as DRAM, SRAM, DDR RAM or other random access solid state memory devices, and may include non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. Memory 246-1 optionally includes one or more storage devices remotely located from the CPU(s) 152-1. Memory 246-1, or alternatively the non-volatile memory device(s) within memory 246-1, comprises a non-transitory computer readable storage medium. In some embodiments, memory 246-1, or the non-transitory computer readable storage medium of memory 246-1 stores the following programs, modules, and data structures, or a subset or superset thereof:

- system mapping module 250-1 that is used for mapping virtual logical addresses (e.g., used by computer system 142, Figure 1B) to intermediate addresses (e.g., which are mapped by storage devices 160 to physical addresses, Figure 1B);
- metrics module 252-1 that is used for generating and/or obtaining one or more metrics of a storage device (e.g., any of storage devices 160-1 to 160-m, Figure 1B);
- trigger detection module 254-1 that is used for detecting a trigger condition (e.g., in accordance with one or more metrics of a storage device);

- enabling module 256-1 that is used for enabling an amelioration process associated with a trigger condition (e.g., detected by trigger detection module 254-1);
- notification module 258-1 that is used for notifying a host to which a storage device is operatively coupled of a trigger condition (e.g., detected by trigger detection module 254-1) and/or of an absence of the trigger condition;
- amelioration module 260-1 that is used for performing an amelioration process to reduce declared capacity of non-volatile memory of a storage device (e.g., storage device 160, Figure 1B), optionally including:
    - detection module 261-1 that is used for detecting an amelioration trigger for reducing declared capacity of the non-volatile memory of the storage device;
    - utilization module 262-1 that is used for reducing utilization of the non-volatile memory of the storage device; and
    - capacity module 264-1 that is used for reducing declared capacity of the non-volatile memory of the storage device; and
- optionally, data redundancy module 266-1 that is used for redundantly encoding data (e.g., to implement a particular RAID (redundant array of independent disks) level); and
- optionally, communication module 268-1 that is used for facilitating communications with other devices, for example via a storage area network (SAN).

[00191] Each of the above identified elements may be stored in one or more of the previously mentioned memory devices, and corresponds to a set of instructions for performing a function described above. The above identified modules or programs (i.e., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various embodiments. In some embodiments, memory 246-1 may store a subset of the modules and data structures identified above. Furthermore, memory 246-1 may store additional modules and data structures not described above. In some embodiments, the programs, modules, and data structures stored in memory 246-1, or the non-transitory computer readable storage medium of memory 246-1, provide instructions for implementing some of the methods described below. In some embodiments, some or all of these modules may be implemented with specialized hardware circuits that subsume part or all of the module functionality.

[00192] Although Figure 2B-1 shows system management module 151-1 in accordance with some embodiments, Figure 2B-1 is intended more as a functional description of the various features which may be present in system management module 151-1 than as a structural schematic of the embodiments described herein. In practice, and as recognized by those of ordinary skill in the art, the programs, modules, and data structures shown separately could be combined and some programs, modules, and data structures could be separated.

[00193] Figure 2B-2 is a block diagram illustrating a system management module 151-2, in accordance with some embodiments, e.g., embodiments in which the system management module is located in the host, as shown in Figure 1B; in some such embodiments the storage system is called a host-managed storage system. System management module 151-2 typically includes one or more processing units (sometimes called CPUs or processors) 152-2 for executing modules, programs and/or instructions stored in memory 246-2 and thereby performing processing operations, memory 246-2 (sometimes called host memory), and one or more communication buses 248-2 for interconnecting these components. The one or more communication buses 248-2 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. System management module 151-2 is coupled to storage system controller 150 by the one or more communication buses 248-2. Memory 246-2 includes high-speed random access memory, such as DRAM, SRAM, DDR RAM or other random access solid state memory devices, and may include non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. Memory 246-2 optionally includes one or more storage devices remotely located from the CPU(s) 152-2. Memory 246-2, or alternatively the non-volatile memory device(s) within memory 246-2, comprises a non-transitory computer readable storage medium. In some embodiments, memory 246-2, or the non-transitory computer readable storage medium of memory 246-2 stores the following programs, modules, and data structures, or a subset or superset thereof:

- system mapping module 250-2 that is used for mapping virtual logical addresses (e.g., used by computer system 142, Figure 1B) to intermediate addresses (e.g., which are mapped by storage devices 160 to physical addresses, Figure 1B);
- metrics module 252-2 that is used for generating and/or obtaining one or more metrics of a storage device (e.g., any of storage devices 160-1 to 160-m, Figure 1B);

- trigger detection module 254-2 that is used for detecting a trigger condition (e.g., in accordance with one or more metrics of a storage device);
- enabling module 256-2 that is used for enabling an amelioration process associated with a trigger condition (e.g., detected by trigger detection module 254-2);
- notification module 258-2 that is used for notifying a host to which a storage device is operatively coupled of a trigger condition (e.g., detected by trigger detection module 254-2) and/or of an absence of the trigger condition;
- amelioration module 260-2 that is used for performing an amelioration process to reduce declared capacity of non-volatile memory of a storage device (e.g., storage device 160, Figure 1B), optionally including:
    - detection module 261-2 that is used for detecting an amelioration trigger for reducing declared capacity of the non-volatile memory of the storage device;
    - utilization module 262-2 that is used for reducing utilization of the non-volatile memory of the storage device; and
    - capacity module 264-2 that is used for reducing declared capacity of the non-volatile memory of the storage device;
- optionally, data redundancy module 266-2 that is used for redundantly encoding data (e.g., to implement a particular RAID (redundant array of independent disks) level) ; and
- optionally, communication module 268-2 that is used for facilitating communications with other devices, for example, via a storage area network (SAN).

[00194] Each of the above identified elements may be stored in one or more of the previously mentioned memory devices, and corresponds to a set of instructions for performing a function described above. The above identified modules or programs (i.e., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various embodiments. In some embodiments, memory 246-2 may store a subset of the modules and data structures identified above. Furthermore, memory 246-2 may store additional modules and data structures not described above. In some embodiments, the programs, modules, and data structures stored in memory 246-2, or the non-transitory computer readable storage medium of memory 246-2, provide instructions for implementing

some of the methods described below. In some embodiments, some or all of these modules may be implemented with specialized hardware circuits that subsume part or all of the module functionality.

[00195] Although Figure 2B-2 shows system management module 151-2 in accordance with some embodiments, Figure 2B-2 is intended more as a functional description of the various features which may be present in system management module 151-2 than as a structural schematic of the embodiments described herein. In practice, and as recognized by those of ordinary skill in the art, the programs, modules, and data structures shown separately could be combined and some programs, modules, and data structures could be separated.

[00196] Figure 2C-1 is a block diagram illustrating a cluster management module 181-1, in accordance with some embodiments, as shown in Figure 1C. Cluster management module 181-1 typically includes one or more processing units (sometimes called CPUs or processors) 182-1 for executing modules, programs and/or instructions stored in memory 276-1 and thereby performing processing operations, memory 276-1, and one or more communication buses 278-1 for interconnecting these components. The one or more communication buses 278-1 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. Cluster management module 181-1 is coupled to additional module(s) 185 by the one or more communication buses 278-1. Memory 276-1 includes high-speed random access memory, such as DRAM, SRAM, DDR RAM or other random access solid state memory devices, and may include non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. Memory 276-1 optionally includes one or more storage devices remotely located from the CPU(s) 182-1. Memory 276-1, or alternatively the non-volatile memory device(s) within memory 276-1, comprises a non-transitory computer readable storage medium. In some embodiments, memory 276-1, or the non-transitory computer readable storage medium of memory 276-1 stores the following programs, modules, and data structures, or a subset or superset thereof:

- cluster mapping module 280-1 that is used for mapping virtual logical addresses (e.g., used by computer system 172, Figure 1C) to intermediate addresses (e.g., which are mapped by storage subsystems 192 to physical addresses, Figure 1C);

- metrics module 282-1 that is used for generating and/or obtaining one or more metrics of a storage device (e.g., any of the storage devices 194-1 to 194-n or 194-j to 194-k, Figure 1C);
- trigger detection module 284-1 that is used for detecting a trigger condition (e.g., in accordance with one or more metrics of a storage device);
- enabling module 286-1 that is used for enabling an amelioration process associated with a trigger condition (e.g., detected by trigger detection module 284-1);
- notification module 288-1 that is used for notifying a host to which a storage device is operatively coupled of a trigger condition (e.g., detected by trigger detection module 284-1) and/or of an absence of the trigger condition;
- amelioration module 290-1 that is used for performing an amelioration process to reduce declared capacity of non-volatile memory of a storage device (e.g., storage device 194, Figure 1C), optionally including:
    - detection module 291-1 that is used for detecting an amelioration trigger for reducing declared capacity of the non-volatile memory of the storage device;
    - utilization module 292-1 that is used for reducing utilization of the non-volatile memory of the storage device; and
    - capacity module 294-1 that is used for reducing declared capacity of the non-volatile memory of the storage device;
- optionally, data redundancy module 296-1 that is used for redundantly encoding data; and
- optionally, communication module 298-1 that is used for facilitating communications with other devices, for example, via a storage area network (SAN).

[00197] Each of the above identified elements may be stored in one or more of the previously mentioned memory devices, and corresponds to a set of instructions for performing a function described above. The above identified modules or programs (i.e., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various embodiments. In some embodiments, memory 276-1 may store a subset of the modules and data structures identified above. Furthermore, memory 276-1 may store additional modules and data structures not described above. In some embodiments, the

programs, modules, and data structures stored in memory 276-1, or the non-transitory computer readable storage medium of memory 276-1, provide instructions for implementing some of the methods described below. In some embodiments, some or all of these modules may be implemented with specialized hardware circuits that subsume part or all of the module functionality.

[00198] Although Figure 2C-1 shows cluster management module 181-1 in accordance with some embodiments, Figure 2C-1 is intended more as a functional description of the various features which may be present in cluster management module 181-1 than as a structural schematic of the embodiments described herein. In practice, and as recognized by those of ordinary skill in the art, the programs, modules, and data structures shown separately could be combined and some programs, modules, and data structures could be separated.

[00199] Figure 2C-2 is a block diagram illustrating a cluster management module 181-2, in accordance with some embodiments, e.g., embodiments in which the cluster management module is located, at least in part, in the host, as shown in Figure 1C; in some such embodiments the storage system uses host-based cluster management. Cluster management module 181-2 typically includes one or more processing units (sometimes called CPUs or processors) 182-2 for executing modules, programs and/or instructions stored in memory 276-2 and thereby performing processing operations, memory 276-2 (sometimes called host memory), and one or more communication buses 278-2 for interconnecting these components. The one or more communication buses 278-2 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. Cluster management module 181-2 is coupled to cluster controller 180 by the one or more communication buses 278-2. Memory 276-2 includes high-speed random access memory, such as DRAM, SRAM, DDR RAM or other random access solid state memory devices, and may include non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. Memory 276-2 optionally includes one or more storage devices remotely located from the CPU(s) 182-2. Memory 276-2, or alternatively the non-volatile memory device(s) within memory 276-2, comprises a non-transitory computer readable storage medium. In some embodiments, memory 276-2, or the non-transitory computer readable storage medium of memory 276-2 stores the following programs, modules, and data structures, or a subset or superset thereof:

- cluster mapping module 280-2 that is used for mapping virtual logical addresses (e.g., used by computer system 172, Figure 1C) to intermediate addresses (e.g., which are mapped by storage subsystems 192 to physical addresses, Figure 1C);
- metrics module 282-2 that is used for generating and/or obtaining one or more metrics of a storage device (e.g., any of the storage devices 194-1 to 194-n or 194-j to 194-k, Figure 1C);
- trigger detection module 284-2 that is used for detecting a trigger condition (e.g., in accordance with one or more metrics of a storage device);
- enabling module 286-2 that is used for enabling an amelioration process associated with a trigger condition (e.g., detected by trigger detection module 284-2);
- notification module 288-2 that is used for notifying a host to which a storage device is operatively coupled of a trigger condition (e.g., detected by trigger detection module 284-2) and/or of an absence of the trigger condition;
- amelioration module 290-2 that is used for performing an amelioration process to reduce declared capacity of non-volatile memory of a storage device (e.g., storage device 194, Figure 1C), optionally including:
    - detection module 291-2 that is used for detecting an amelioration trigger for reducing declared capacity of the non-volatile memory of the storage device;
    - utilization module 292-2 that is used for reducing utilization of the non-volatile memory of the storage device; and
    - capacity module 294-2 that is used for reducing declared capacity of the non-volatile memory of the storage device;
- optionally, data redundancy module 296-2 that is used for redundantly encoding data; and
- optionally, communication module 298-2 that is used for facilitating communications with other devices, for example, via a storage area network (SAN).

[00200] Each of the above identified elements may be stored in one or more of the previously mentioned memory devices, and corresponds to a set of instructions for performing a function described above. The above identified modules or programs (i.e., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-

arranged in various embodiments. In some embodiments, memory 276-2 may store a subset of the modules and data structures identified above. Furthermore, memory 276-2 may store additional modules and data structures not described above. In some embodiments, the programs, modules, and data structures stored in memory 276-2, or the non-transitory computer readable storage medium of memory 276-2, provide instructions for implementing some of the methods described below. In some embodiments, some or all of these modules may be implemented with specialized hardware circuits that subsume part or all of the module functionality.

[00201] Although Figure 2C-2 shows cluster management module 181-2 in accordance with some embodiments, Figure 2C-2 is intended more as a functional description of the various features which may be present in cluster management module 181-2 than as a structural schematic of the embodiments described herein. In practice, and as recognized by those of ordinary skill in the art, the programs, modules, and data structures shown separately could be combined and some programs, modules, and data structures could be separated.

[00202] Figure 2D is a block diagram illustrating amelioration module 230 included in management module 121-1 of Figure 2A-1 and/or management module 121-2 of Figure 2A-2, in accordance with some embodiments. As described above with respect to Figures 2A-1 and 2A-2, in some embodiments, amelioration module 230 includes utilization module 232 and capacity module 234. In some embodiments, utilization module 232 includes the following programs and/or modules, or a subset or superset thereof:

- trimming module 235 that is used for trimming from a storage device at least a portion of previously-written data that is no longer used by a host (e.g., not live data 332, Figure 3);
- deleting module 236 that is used for deleting from a storage device discardable data that is used by a host; and
- moving module 237 that is used for moving a portion of data that is used by a host from a storage device to another storage device.

[00203] In some embodiments, capacity module 234 includes the following programs and/or modules, or a subset or superset thereof:

- LBA reduction module 238 (sometimes called a logical address reduction module) that is used for reducing a range of logical addresses, reducing a count of logical addresses, and/or making specific logical addresses unavailable to a host; and

- advertising module 239 that is used for advertising a reduced declared capacity of non-volatile memory of a storage device or storage subsystem.

[00204] In some embodiments, the amelioration process includes a utilization reduction process (e.g., performed by utilization module 232) and a declared capacity reduction process (e.g., performed by capacity module 234). In some embodiments, the amelioration process has a target reduced declared capacity to be achieved by the amelioration process, and utilizes the target reduced declared capacity to determine a target amount of utilization reduction to be achieved by the amelioration process. In some circumstances, such as when the amelioration process has a target reduced declared capacity to be achieved by the amelioration process, and the amount of a storage device utilized is less than the target reduced declared capacity, the target amount of utilization reduction is zero. In such circumstances, performance of the utilization reduction process, or one or more portions of the utilization reduction process, is unneeded and therefore skipped or forgone. Furthermore, in some embodiments, the amelioration process (e.g., periodically, semi-continuously, irregularly, initially and/or finally) recomputes or re-evaluates a number of parameters, such as the target reduced declared capacity and/or the target amount of utilization reduction, as those parameters may change in value due to the amelioration process and/or normal storage operations (e.g., read, write, erase and trim or unmap operations). In some circumstances, in accordance with the recomputed or re-evaluated parameters, the utilization reduction is re-prioritized, re-scheduled, or aborted. Although Figure 2D uses the example of amelioration module 230 included in Figures 2A-1 and 2A-2, the description of Figure 2D similarly applies to other amelioration modules (e.g., amelioration module 260-1 of Figure 2B-1, amelioration module 260-2 of Figure 2B-2, amelioration module 290-1 of Figure 2C-1, and/or amelioration module 290-2 of Figure 2C-2), and for sake of brevity, the details are not repeated here.

[00205] The trim operation indicates that specific portions of the LBA space (320, Figure 3) are converted to unallocated LBA space (340), thereby reducing utilization. The trim operation typically includes invalidating one or more entries of a mapping table (e.g., mapping table 402, Figure 4) used to translate logical addresses in a logical address space to physical addresses in a physical address space of a storage device. As a result of the trim operation, any data that was previously stored using the specific portion of the LBA space is no longer available to the host (e.g., is discarded). The physical pages used to previously store this data may be reused for other purposes. The reuse may be done coincident in time

with the trim operation or at a future time (e.g., during garbage collection). As discussed elsewhere, the reused physical pages may be used with a different encoding format or different redundancy mechanism than the encoding format or different redundancy mechanism used by those same physical pages before the reuse. The trim operation is sometimes also referred to as an unmap operation. The trim operation, as used herein, is not necessarily identical to the trim operation of the SATA protocol. The unmap operation, as used herein, is not necessarily identical to the unmap operation of the SCSI protocol.

[00206] Figure 3 is a block diagram of a logical block address (LBA) space 320 (sometimes called logical address (LA) space), in accordance with some embodiments. In some embodiments, a logical address is the address at which an item (e.g., a file or other data) resides from the perspective of a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, and/or computer system 172, Figure 1C). In some embodiments, a logical address (e.g., in LBA space 320) differs from a physical address (e.g., in physical space 318) due to the operation of a mapping function or address translator (e.g., a function or module that includes translation table 212-1, Figure 2A-1, or mapping table 402, Figure 4). In some embodiments, a logical block address (LBA) is mapped to a physical flash address (e.g., a physical page number (PPN), including a bank, block, and page), as described further with respect to Figure 4.

[00207] In some embodiments, a logical address space includes allocated logical address space (e.g., allocated LBA space 342) and unallocated logical address space (e.g., unallocated LBA space 340). In some embodiments, unallocated logical address space is logical address space at which no data is stored. In some embodiments, unallocated logical address space includes logical address space that has never been written to and/or has been discarded (previously written data may be discarded through a trim or unmap operation, and is sometimes called trimmed logical address space). For example, in Figure 3, unallocated LBA space 340 includes trimmed LBA space 330. In some embodiments, allocated logical address space is logical address space that was previously-written by a host, the previously-written data including data that is no longer used by a host (e.g., not live data 332) and data that is still in use by the host (e.g., live data 334). In some embodiments, not live data is data in a portion of the logical address space that is marked as free, available or unused in the metadata of a file system. Optionally, a file system may choose to convert not live address space into unallocated address space through a trim or unmap operation.

[00208] In Figure 3, allocated LBA space 342 represents an amount of allocated space, and unallocated LBA space 340 represents an amount of unallocated space. However, neither allocated LBA space 342 nor unallocated LBA space 340 is necessarily a single contiguous region of LBA space 320. Similarly, live data 334 and not live data 332 in Figure 3 represent amounts (e.g., counts of LBAs) of live data and not live data, respectively. However, neither live data 334 nor not live data 332 is necessarily a single contiguous region of LBA space 320 or allocated LBA space 342, nor do the positions of live data 334 and not live data 332 illustrated in Figure 3 have any correlation to the logical or physical address values of the live data and not live data. Typically, live data 334 and/or not live data 332 will be present in multiple regions of LBA space 320, and are thus non-contiguous. Optionally, however, a remapping or coalescing process, which can also be called defragmentation, can be performed to consolidate some or all live data 334 into a contiguous region of LBA space 320.

[00209] Allocated logical address space (342) is space that is utilized. The utilization reduction modules and processes discussed herein are modules, applications and processes whose purpose is to reduce the size of the allocated logical address space, and thus reduce utilization of non-volatile memory in a storage device or data storage system. Typically, reducing the size of the allocated logical address space requires reducing the amount of live data 334 and/or not live data 332 stored by a storage device, or storage system, thereby converting a portion of the allocated logical address space into unallocated logical address space. In some embodiments, portions of not live data 332 are trimmed, and thereby converted into unallocated logical address space through the use of trim or unmap operations.

[00210] In some embodiments, a logical address may be outside of LBA Space (320) and is therefore unavailable. A previously available logical address can be made unavailable by reducing the size of the LBA space (320) such that that address is no longer within LBA space (320) and hence becomes unavailable (e.g. it is an undefined operation or erroneous operation to request a normal storage operation to a logical address that is outside of LBA space (320)). As noted above, LBA Space (320) can be reduced by a command to the storage device, or a host can limit its usage of logical addresses to a reduced range of logical addresses therefore effectively reducing LBA space (320).

[00211] In some embodiments, the total number of allocated logical addresses (342) is limited. In such embodiments, specific logical addresses are considered to be unavailable if usage of them would cause the system to exceed the limited total number. For example, if

the total number of allocated logical addresses is limited to five and the currently allocated addresses are 1, 3, 19, 45 and 273838 then any specific logical address other than these five (e.g., 6, 15, 137, etc.) would be considered unavailable.

[00212] Figure 4 is a block diagram of a mapping table 402 and physical address space 410, in accordance with some embodiments. In some embodiments, mapping table 402 is used to translate a logical block address (LBA) from the perspective of a host (e.g., computer system 110, Figure 1A) to a physical address in a physical address space (e.g., physical address space 410) of non-volatile memory in a storage device (e.g., storage device 120, Figure 1A). In some embodiments, an LBA is the address of the smallest unit of stored data that is addressable by a host (e.g., 512 B or 4096 B). In some embodiments, LBAs are a sequence of integers organized in numerical order in the logical address space. In some embodiments, LBAs are integers chosen from a logical address space but need not be contiguous. For example, in implementations that utilize a sparse logical address space, the amount of addressable space is governed by a limit on the number of logical addresses that can be allocated, but those logical addresses are distributed over a larger address space than the maximum number of logical addresses that can be allocated (e.g., to a host or a set of hosts or clients).

[00213] In some embodiments, mapping table 402 is stored in memory associated with the storage device (e.g., in memory 206-1, as part of translation table 212-1, Figure 2A-1). In some embodiments, a physical address is a physical page number (PPN), including a bank number, a block number, and a page number. In the example shown in Figure 4, LBA 0 is mapped to bank 1 (e.g., Bank 420-1), block 3 (e.g., Block 421-3), page 37 (pages not shown in Figure 4) of physical address space 410. Figure 4 shows that physical address space 410 includes a plurality of non-volatile memory blocks 421, 422 423, 424. As described above, and as shown in the representation of block 424-p, each non-volatile memory block in the physical address space of a storage device typically includes a plurality of pages 426, where each page is typically an instance of the smallest individually accessible (e.g., readable or writable) portion of a block. Although Figure 4 illustrates one example of a logical address to physical address mapping, in other embodiments, different mappings may be used. For example, in some embodiments, each of the logical address entries corresponds to multiple (e.g., eight) logical addresses (e.g., 8 LBAs per logical address entry). In some embodiments mapping table 402 need not contain contiguous LBA addresses and may be organized in any

manner to facilitate lookup operations, e.g., hash table, binary tree, content addressable memory, and others.

[00214] As discussed below with reference to Figure 5A, a single-level flash memory cell (SLC) stores one bit ("0" or "1"). Thus, the storage density of a SLC memory device is one bit of information per memory cell. A multi-level flash memory cell (MLC), however, can store two or more bits of information per cell by using different ranges within the total voltage range of the memory cell to represent a multi-bit bit-tuple. In turn, the storage density of a MLC memory device is multiple-bits per cell (e.g., two bits per memory cell).

[00215] Flash memory devices utilize memory cells to store data as electrical values, such as electrical charges or voltages. Each flash memory cell typically includes a single transistor with a floating gate that is used to store a charge, which modifies the threshold voltage of the transistor (i.e., the voltage needed to turn the transistor on). The magnitude of the charge, and the corresponding threshold voltage, is used to represent one or more data values. In some embodiments, during a read operation, a reading threshold voltage is applied to the control gate of the transistor and the resulting sensed current or voltage is mapped to a data value.

[00216] The terms "cell voltage" and "memory cell voltage," in the context of flash memory cells, typically means the threshold voltage of the memory cell, which is the minimum voltage that needs to be applied to the gate of the memory cell's transistor in order for the transistor to conduct current. Similarly, reading threshold voltages (sometimes also called reading signals, reading voltages, and/or read thresholds) applied to a flash memory cells are gate voltages applied to the gates of the flash memory cells to determine whether the memory cells conduct current at that gate voltage. In some embodiments, when a flash memory cell's transistor conducts current at a given reading threshold voltage, indicating that the cell voltage is less than the reading threshold voltage, the raw data value for that read operation is a "1," and otherwise the raw data value is a "0."

[00217] Figure 5A is a simplified, prophetic diagram of voltage distributions 300a found in a single-level flash memory cell (SLC) over time, in accordance with some embodiments. The voltage distributions 300a shown in Figure 5A have been simplified for illustrative purposes. In this example, the SLC's voltage range extends approximately from a voltage, V_{SS}, at a source terminal of an NMOS transistor to a voltage, V_{DD}, at a drain terminal of the NMOS transistor. As such, voltage distributions 300a extend between V_{SS} and V_{DD}.

[00218] Sequential voltage ranges 301 and 302 between source voltage V_{SS} and drain voltage V_{DD} are used to represent corresponding bit values "1" and "0," respectively. Each voltage range 301, 302 has a respective center voltage V_1 301b, V_0 302b. As described below, in many circumstances the memory cell current sensed in response to an applied reading threshold voltages is indicative of a memory cell voltage different from the respective center voltage V_1 301b or V_0 302b corresponding to the respective bit value written into the memory cell. Errors in cell voltage, and/or the cell voltage sensed when reading the memory cell, can occur during write operations, read operations, or due to "drift" of the cell voltage between the time data is written to the memory cell and the time a read operation is performed to read the data stored in the memory cell. For ease of discussion, these effects are collectively described as "cell voltage drift." Each voltage range 301, 302 also has a respective voltage distribution 301a, 302a that may occur as a result of any number of a combination of error-inducing factors, examples of which are identified above.

[00219] In some implementations, a reading threshold voltage V_R is applied between adjacent center voltages (e.g., applied proximate to the halfway region between adjacent center voltages V_1 301b and V_0 302b). Optionally, in some implementations, the reading threshold voltage is located between voltage ranges 301 and 302. In some implementations, reading threshold voltage V_R is applied in the region proximate to where the voltage distributions 301a and 302a overlap, which is not necessarily proximate to the halfway region between adjacent center voltages V_1 301b and V_0 302b.

[00220] In order to increase storage density in flash memory, flash memory has developed from single-level (SLC) cell flash memory to multi-level cell (MLC) flash memory so that two or more bits can be stored by each memory cell. As discussed below with reference to Figure 5B, a MLC flash memory device is used to store multiple bits by using voltage ranges within the total voltage range of the memory cell to represent different bit-tuples. A MLC flash memory device is typically more error-prone than a SLC flash memory device created using the same manufacturing process because the effective voltage difference between the voltages used to store different data values is smaller for a MLC flash memory device. Moreover, due to any number of a combination of factors, such as electrical fluctuations, defects in the storage medium, operating conditions, device history, and/or write-read circuitry, a typical error includes a stored voltage level in a particular MLC being in a voltage range that is adjacent to the voltage range that would otherwise be representative of the correct storage of a particular bit-tuple. As discussed in greater detail below with

reference to Figure 5B, the impact of such errors can be reduced by gray-coding the data, such that adjacent voltage ranges represent single-bit changes between bit-tuples.

[00221] Figure 5B is a simplified, prophetic diagram of voltage distributions 300b found in a multi-level flash memory cell (MLC) over time, in accordance with some embodiments. The voltage distributions 300b shown in Figure 5B have been simplified for illustrative purposes. The cell voltage of a MLC approximately extends from a voltage, V_{SS}, at the source terminal of a NMOS transistor to a voltage, V_{DD}, at the drain terminal. As such, voltage distributions 300b extend between V_{SS} and V_{DD}.

[00222] Sequential voltage ranges 311, 312, 313, 314 between the source voltage V_{SS} and drain voltages V_{DD} are used to represent corresponding bit-tuples "11," "01," "00," "10," respectively. Each voltage range 311, 312, 313, 314 has a respective center voltage 311b, 312b, 313b, 314b. Each voltage range 311, 312, 313, 314 also has a respective voltage distribution 311a, 312a, 313a, 314a that may occur as a result of any number of a combination of factors, such as electrical fluctuations, defects in the storage medium, operating conditions, device history (e.g., number of program-erase (P/E) cycles), and/or imperfect performance or design of write-read circuitry.

[00223] Ideally, during a write operation, the charge on the floating gate of the MLC would be set such that the resultant cell voltage is at the center of one of the ranges 311, 312, 313, 314 in order to write the corresponding bit-tuple to the MLC. Specifically, the resultant cell voltage would be set to one of V_{11} 311b, V_{01} 312b, V_{00} 313b and V_{10} 314b in order to write a corresponding one of the bit-tuples "11," "01," "00" and "10." In reality, due to the factors mentioned above, the initial cell voltage may differ from the center voltage for the data written to the MLC.

[00224] Reading threshold voltages V_{RA}, V_{RB} and V_{RC} are positioned between adjacent center voltages (e.g., positioned at or near the halfway point between adjacent center voltages) and, thus, define threshold voltages between the voltage ranges 311, 312, 313, 314. During a read operation, one of the reading threshold voltages V_{RA}, V_{RB} and V_{RC} is applied to determine the cell voltage using a comparison process. However, due to the various factors discussed above, the actual cell voltage, and/or the cell voltage received when reading the MLC, may be different from the respective center voltage V_{11} 311b, V_{01} 312b, V_{00} 313b or V_{10} 314b corresponding to the data value written into the cell. For example, the actual cell voltage may be in an altogether different voltage range, strongly indicating that the MLC is storing a different bit-tuple than was written to the MLC. More commonly, the actual cell

voltage may be close to one of the read comparison voltages, making it difficult to determine with certainty which of two adjacent bit-tuples is stored by the MLC.

[00225] Errors in cell voltage, and/or the cell voltage received when reading the MLC, can occur during write operations, read operations, or due to "drift" of the cell voltage between the time data is written to the MLC and the time a read operation is performed to read the data stored in the MLC. For ease of discussion, sometimes errors in cell voltage, and/or the cell voltage received when reading the MLC, are collectively called "cell voltage drift."

[00226] One way to reduce the impact of a cell voltage drifting from one voltage range to an adjacent voltage range is to gray-code the bit-tuples. Gray-coding the bit-tuples includes constraining the assignment of bit-tuples such that a respective bit-tuple of a particular voltage range is different from a respective bit-tuple of an adjacent voltage range by only one bit. For example, as shown in Figure 5B, the corresponding bit-tuples for adjacent ranges 301 and 302 are respectively "11" and "01," the corresponding bit-tuples for adjacent ranges 302 and 303 are respectively "01" and "00," and the corresponding bit-tuples for adjacent ranges 303 and 304 are respectively "00" and "10." Using gray-coding, if the cell voltage drifts close to a read comparison voltage level, the error is typically limited to a single bit within the 2-bit bit-tuple.

[00227] Although the description of Figure 5B uses an example in which $q = 2$ (i.e., 2 bits per cell in a MLC flash memory), those skilled in the art will appreciate that the embodiments described herein may be extended to memory cells that have more than four possible states per cell, yielding more than two bits of information per cell. For example, in some embodiments, a triple-level memory cell (TLC) has eight possible states per cell, yielding three bits of information per cell. As another example, in some embodiments, a quad-level memory cell (QLC) has 16 possible states per cell, yielding four bits of information per cell. As another example, in some embodiments, a cell might store only 6 states, yielding approximately 2.5 bits of information per cell, meaning that two cells together would provide 36 possible states, more than sufficient to store 5 bits of information per pair of cells.

[00228] Figure 6 illustrates a flowchart representation of a method 600 of managing a storage system, in accordance with some embodiments. At least in some embodiments, method 600 is performed by a storage device (e.g., storage device 120, Figure 1A) or one or more components of the storage device (e.g., storage controller 124 and/or storage medium

130, Figure 1A), wherein the storage device is operatively coupled with a host system (e.g., computer system 110, Figure 1A). In some embodiments, method 600 is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a device, such as the one or more processing units (CPUs) 122-1 of management module 121-1, shown in Figures 1A and 2A-1. In some embodiments, method 600 is performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, and/or data storage system 170, Figure 1C) or one or more components of the storage system (e.g., computer system 110 and/or storage device 120, Figure 1A, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C). In some embodiments, some of the operations of method 600 are performed at a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, and/or computer system 172, Figure 1C) and information is transmitted to a storage device (e.g., storage device 120, Figure 1A) and/or one or more subsystems of a storage system (e.g., storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C). In some embodiments, method 600 is governed, at least in part, by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a subsystem of a storage system, such as the one or more processing units (CPUs) 152-1 of system management module 151-1, shown in Figures 1B and 2B-1 or the one or more processing units (CPUs) 182-1 of cluster management module 181-1, shown in Figures 1C and 2C-1. In some embodiments, method 600 is governed, at least in part, by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a host (processors of host not shown in Figures 1A, 1B, and 1C). For ease of explanation, the following describes method 600 as performed by a storage device (e.g., storage device 120, Figure 1A). However, those skilled in the art will appreciate that in other embodiments, one or more of the operations described in method 600 are performed by a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, and/or computer system 172, Figure 1C) and/or one or more subsystems of a storage system distinct from the storage device (e.g., storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C).

[00229] A storage device (e.g., storage device 120, Figure 1A), triggers (602) a condition for reducing declared capacity of non-volatile memory of the storage device in accordance with one or more metrics of the storage device (e.g., including one or more status metrics corresponding to the storage device's ability to retain data, one or more performance metrics corresponding to performance of the storage device, one or more wear metrics

corresponding to wear on the storage device, and/or one or more time metrics). Metrics of the storage device include metrics (e.g., wear metrics such as P/E cycle counts, write operation counts and the like) of the non-volatile storage media (e.g., storage medium 130, Figure 1A) of the storage device, but are not necessarily limited to such metrics. For example, some metrics (e.g., some performance metrics, such as latency metrics, metrics that measure how long it takes or how many operations are required to complete a write or erase operation, etc.) of the storage device reflect both storage media performance as well as controller and/or other storage device component performance.

[00230] In some embodiments, the metrics of the storage device used to determine the trigger condition include a write amplification metric of the storage device. Another metric of the storage device that is used, in some embodiments, to determine the trigger condition is an over-provisioning metric (e.g., quantity or percentage of total storage capacity that is in excess of the declared capacity of the storage device, and/or quantity or percentage of total storage capacity that is in excess of the declared capacity of the storage device after a projected conversion of a number of memory blocks (or other portions of the storage device) from a current encoding format (e.g., TLC, MLC and/or data redundancy mechanism) to a lower storage density encoding format (e.g., MLC, SLC and/or data redundancy mechanism). For example, in some embodiments, a trigger condition is determined if a projected over-provisioning metric, corresponding to a number of blocks (or other portions) of the storage device removed from service (e.g., that have been or will be removed from service) due to wear or due to failure of those blocks (or other portions) to meet a predefined quality of service metric, falls below a predefined threshold (e.g., a non-zero threshold such as 2 percent or 5 percent or the like), or falls below a threshold determined in accordance with a measured or projected write amplification of the storage device.

[00231] Write amplification is a phenomenon where the actual amount of physical data written to a storage medium (e.g., storage medium 130 in storage device 120) is a multiple of the logical amount of data written by a host (e.g., computer system 110, sometimes called a host) to the storage medium. As discussed above, when a block of storage medium must be erased before it can be re-written, the garbage collection process to perform these operations results in re-writing data one or more times. This multiplying effect increases the number of writes required over the life of a storage medium, which shortens the time it can reliably operate. The formula to calculate the write amplification of a storage system is given by equation:

$$\frac{amount\ of\ data\ written\ to\ a\ storage\ medium}{amount\ of\ data\ written\ by\ a\ host}$$

[00232] One of the goals of any flash memory based data storage system architecture is to reduce write amplification as much as possible so that available endurance is used to meet storage medium reliability and warranty specifications. Higher system endurance also results in lower cost as the storage system may need less over-provisioning. By reducing write amplification, the endurance of the storage medium is increased and the overall cost of the storage system is decreased. Generally, garbage collection is performed on erase blocks with the fewest number of valid pages for best performance and best write amplification.

[00233] In some embodiments, the trigger condition is detected in accordance with a non-linear and/or linear combination of the one or more metrics. For example, in some embodiments, the trigger condition is detected by comparing a wear metric such as P/E cycle counts to a previously determined value, e.g., a threshold value. In some embodiments, the trigger condition can also be asserted by other means, e.g., by a human operator or scheduled by a human operator. For example, it may be desirable to initiate the amelioration process because of the expected availability or unavailability of resources. For example, it may be desirable to initiate the amelioration process because performance characteristics of the storage device (including reliability) are altered.

[00234] In some embodiments, the trigger condition is detected in accordance with historical knowledge of the one or more metrics. For example, historical knowledge can be a running average of one or more metrics. In another example, historical knowledge can be used to determine (e.g., compute) one or more projected values of one or more metrics at a particular time in the future (e.g., an hour, day, week, or month in the future), and the trigger condition can be detected in accordance with the one or more projected values. The latter methodology can be particularly useful for avoiding events that result in loss of data (e.g., due to wear out), or more generally for avoiding events that significantly impact on the quality of service provided by a storage system, and for enabling a storage system to undertake ameliorative measures prior to there being an urgent need to do so. For example, in some embodiments, the trigger condition is detected by comparing a historical wear metric such as P/E cycle counts to a previously determined value to anticipate wear out of a portion of the storage media. Similarly, in some embodiments, the trigger condition is detected by comparing a historical metric, such as the bit error rate (BER), or the rate of change of the metric, BER (of the storage media, or a portion of the storage media), or a projected value (e.g., a projected BER rate at a particular time in the future, as determined based on a current

or historical BER and a rate of change of the BER), against a previously determined value to anticipate performance degradation due to increased computation requirements of error correction.

[00235] In a storage system with a plurality of storage devices the trigger condition may be dependent on metrics obtained from a plurality of the storage devices. The amelioration process may operate on more than one storage device at a time, either sequentially or in parallel. For example, a storage system may have a fixed maximum rate of capacity reduction independent of how many storage devices are currently being operated on in parallel by the amelioration process (e.g., maximum rate of data movement between the storage devices while reducing utilization). The trigger condition should include considering, separately and in combination, the metrics of the plurality of storage devices when determining the targeted capacity reduction and, due to the fixed maximum rate, the scheduling of the amelioration process.

[00236] The storage device notifies (604) a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device. In some embodiments, upon receipt of the notification of the trigger condition the host sends an amelioration trigger to initiate the amelioration process (606).

[00237] The storage device or a host detects the amelioration trigger and, in accordance with the detected amelioration trigger, performs an amelioration process (606) to reduce declared capacity of the non-volatile memory of the storage device. In some embodiments, the amelioration process includes a process to reduce utilization (608), a process to reduce declared capacity (610), and/or a process to advertise (612) a reduced declared capacity. As described above with respect to Figure 2D, in some embodiments, the amelioration process (606) includes a utilization reduction process (608) (e.g., performed by utilization module 232, Figures 2A-1 and 2A-2, utilization module 262, Figures 2B-1 and 2B-2, or utilization module 292, Figures 2C-1 and 2C-2) and/or a declared capacity reduction process (610) (e.g., performed by capacity module 234, Figures 2A-1 and 2A-2, capacity module 264, Figures 2B-1 and 2B-2, or capacity module 294, Figures 2C-1 and 2C-2). In some circumstances, such as when the amount of the storage device utilized by the host is less than the reduced declared capacity, performance of the process to reduce utilization

(608), or one or more portions of the process to reduce utilization, is skipped or forgone. Although Figure 6 shows operations 608, 610, and 612 as sequential processes of the amelioration process (606), in some embodiments, these processes may be overlapped, non-sequential, and/or performed in a different order.

[00238] In some embodiments, prior to the operations described above in Figure 6 (e.g., operations 602, 604, and/or 606), method 600 includes reducing (601) over-provisioning of the non-volatile memory of the storage device. As described above, over-provisioning refers to a quantity or percentage of total storage capacity that is in excess of the declared capacity of the storage device. In some embodiments, over-provisioning refers to the difference between the physical capacity of the storage device (e.g., the physical capacity less capacity set aside for management data structures and metadata) for storing user data (e.g., data stored in the storage system on behalf of a host or host system), and the logical capacity presented as available for a host or user. For example, in some embodiments, if a non-volatile memory of a storage device has 12 GB of total storage capacity (e.g., total storage capacity for storing user data) and 10 GB of declared capacity, then the non-volatile memory of the storage device has 2 GB of over-provisioning. Unlike declared capacity, which is the storage capacity available to a host, the extra capacity of over-provisioning is not visible to the host as available storage. Instead, over-provisioning is used to increase endurance of a storage device (e.g., by distributing the total number of writes and erases across a larger population of blocks and/or pages over time), improve performance (e.g., by providing additional buffer space for managing P/E cycles and improving the probability that a write operation will have immediate access to a pre-erased block), and reduce write amplification.

[00239] In some embodiments, reducing (601) over-provisioning includes: (1) detecting a first wear condition of non-volatile memory of a storage device of a storage system, wherein a total storage capacity of the non-volatile memory of the storage device includes declared capacity and over-provisioning, and (2) in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the storage device without reducing declared capacity of the non-volatile memory of the storage device. In some embodiments, performing a remedial action that reduces over-provisioning includes marking one or more blocks of the non-volatile memory as unusable. In some embodiments, performing a remedial action that reduces over-provisioning includes converting one or more MLC blocks to SLC, or more generally,

changing the physical encoding format of one or more blocks of the non-volatile memory. In some embodiments, reducing over-provisioning is performed by an over-provisioning module of management module 121, system management module 151, or cluster management module 181 (e.g., in memory 206 of Figures 2A-1 and 2A-2, in memory 246 of Figures 2B-1 and 2B-2, or in memory 276 of Figures 2C-1 and 2C-2, respectively, but not explicitly shown). Furthermore, in some circumstances or in some embodiments, over-provisioning reducing operation 601 is performed multiple times prior to the first time operation 602 is performed. For example, over-provisioning reducing operation 601 may be repeated each of multiple times that a predefined wear condition is detected, until over-provisioning falls to or below a predefined minimum level.

[00240] In some embodiments, the first wear condition is detected in accordance with one or more metrics of the storage device (e.g., including one or more status metrics corresponding to the storage device's ability to retain data, one or more performance metrics corresponding to performance of the storage device, one or more wear metrics corresponding to wear on the storage device, and/or one or more time metrics), as described above with respect to operation 602. In some embodiments, the first wear condition is detected in accordance with a determination that the one or more metrics of the storage device satisfy a first criterion and over-provisioning of the non-volatile memory of the storage device is greater than a predefined threshold (e.g., 2 percent of the declared capacity, at least 100 blocks, or 40 blocks + n% of declared capacity, etc.).

[00241] In some embodiments, detecting the trigger condition (as described above with respect to operation 602) comprises detecting a second wear condition distinct from the first wear condition. For example, in some embodiments, the trigger condition (or the second wear condition) is detected in accordance with a determination that the one or more metrics of the storage device satisfy a second criterion (e.g., the first criterion used for the first wear condition or another criterion) and over-provisioning of the non-volatile memory of the storage device is less than or equal to (e.g., not greater than) a predefined threshold (e.g., 2 percent of the declared capacity, at least 100 blocks, or 40 blocks + n% of declared capacity, etc.).

[00242] Figures 7A-7D illustrate a flowchart representation of a method 700 of managing a storage system, in accordance with some embodiments. At least in some embodiments, method 700 is performed by a storage device (e.g., storage device 120, Figure 1A, storage device 160, Figure 1B, or storage device 194, Figure 1C) or one or more

components of the storage device (e.g., storage controller 124, Figure 1A), wherein the storage device is operatively coupled with a host system (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). In some embodiments, method 700 is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a device, such as the one or more processing units (CPUs) 122-1 of management module 121-1, shown in Figures 1A and 2A-1. In some embodiments, method 700 is performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, and/or data storage system 170, Figure 1C) or one or more components of the storage system (e.g., storage device 120, Figure 1A, storage device 160, Figure 1B, or storage device 194, Figure 1C). In some embodiments, some of the operations of method 700 are performed at a storage device (e.g., storage device 120, Figure 1A, storage device 160, Figure 1B, or storage device 194, Figure 1C) and information is transmitted to a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). For ease of explanation, the following describes method 700 as performed by a storage device (e.g., storage device 120, Figure 1A) of a storage system (e.g., data storage system 100, Figure 1A). However, those skilled in the art will appreciate that in other embodiments, one or more of the operations described in method 700 are performed by a storage device of another storage system (e.g., storage device 160 of data storage system 140, Figure 1B, or storage device 194 of data storage system 170, Figure 1C).

[00243] At a storage device of a storage system (702), the storage device (e.g., storage device 120, Figure 1A) generates (704) one or more metrics of the storage device, the storage device including non-volatile memory. In some embodiments, a metrics module (e.g., metrics module 222-1, Figure 2A-1) is used to generate one or more metrics of the storage device, the storage device including non-volatile memory, as described above with respect to Figure 2A-1.

[00244] In some embodiments, the storage device comprises (712) one or more flash memory devices. In some embodiments, the storage device comprises a storage medium (e.g., storage medium 130, Figure 1A), and the storage medium comprises one or more non-volatile storage devices, such as flash memory devices. In some embodiments, the storage medium (e.g., storage medium 130, Figure 1A) is a single flash memory device, while in

other embodiments the storage medium includes a plurality of flash memory devices. For example, in some embodiments, the storage medium includes dozens or hundreds of flash memory devices, organized in parallel memory channels, such as 16, 32 or 64 flash memory devices per memory channel, and 8, 16 or 32 parallel memory channels. In some embodiments, the non-volatile storage medium (e.g., storage medium 130, Figure 1A) includes NAND-type flash memory or NOR-type flash memory. In other embodiments, the storage medium comprises one or more other types of non-volatile storage devices.

[00245] In some embodiments, generating (714) one or more metrics of the storage device includes generating at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the storage device. In some embodiments, at least one metric is generated for each block of a plurality of blocks of the storage device. In some embodiments, at least one metric is generated for each page of a plurality of pages of the storage device. In some embodiments, at least one metric is generated for each region of a plurality of regions of the storage device. In some embodiments, some metrics are generated on a block basis, some metrics are generated on a page basis, some metrics are generated on a region basis, and/or some metrics are generated on a storage device basis.

[00246] In some embodiments, the one or more metrics of the storage device include (716) one or more status metrics corresponding to the storage device's ability (e.g., ability of the storage device's storage medium or storage media) to retain data. In some embodiments, storage controller 124 (Figure 1A) or a component thereof (e.g., metrics module 222-1, Figure 2A-1) generates and/or maintains one or more status metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the storage device. In some embodiments, the one or more status metrics indicate a respective memory portion's ability to retain data. In some embodiments, the one or more status metrics associated with a respective memory portion, of a plurality of memory portions of the storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more status metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a bytes written field indicating a number of bytes of data written to pages in the respective memory portion, (b) a program-erase (P/E) cycle field indicating a current count of the number of P/E cycles performed on the respective memory portion, (c) a bit error rate (BER) field indicating a number of errors detected in a codeword read from pages of the respective memory portion, and (d) other usage information indicating the health, performance, and/or endurance of the

respective memory portion, as it relates to the respective memory portion's ability to retain data. In some embodiments, the one or more status metrics indicate the storage device's ability, as a whole, to retain data. For example, as the storage device ages, the one or more status metrics reflect the storage device's diminished ability to retain data (e.g., data read from the storage device typically have more errors as the storage device ages).

[00247] In some embodiments, the one or more metrics of the storage device include (718) one or more performance metrics corresponding to performance of the storage device. In some embodiments, storage controller 124 (Figure 1A) or a component thereof (e.g., metrics module 222-1, Figure 2A-1) generates and/or maintains one or more performance metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the storage device. In some embodiments, the one or more performance metrics correspond to performance of a respective memory portion of the plurality of memory portions. In some embodiments, the one or more performance metrics associated with a respective memory portion, of a plurality of memory portions of the storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more performance metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a measure of latency, and (b) transaction time. In some embodiments, the one or more performance metrics correspond to performance of the storage device as a whole. For example, in some embodiments, the one or more performance metrics include a measure of latency for the storage device and/or transaction time for the storage device.

[00248] In some embodiments, the one or more metrics of the storage device include (720) one or more wear metrics corresponding to wear on the storage device. In some embodiments, storage controller 124 (Figure 1A) or a component thereof (e.g., metrics module 222-1, Figure 2A-1) generates and/or maintains one or more wear metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the storage device. In some embodiments, the one or more wear metrics correspond to wear on a respective memory portion of the plurality of memory portions. In some embodiments, the one or more wear metrics associated with a respective memory portion, of a plurality of memory portions of the storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more wear metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a count of cumulative writes to the respective memory portion, (b) a

count of cumulative reads from the respective memory portion, (c) a count of P/E cycles performed on the respective memory portion, and (d) a BER for the respective memory portion. In some embodiments, the one or more wear metrics correspond to wear on the storage device as a whole. For example, in some embodiments, the one or more wear metrics include a count of cumulative writes to the storage device (e.g., to the storage device's storage medium or storage media), a count of cumulative reads from the storage device (e.g., from the storage device's storage medium or storage media), a count of P/E cycles performed on the storage device and/or a BER for the storage device.

[00249] In some embodiments, the one or more metrics of the storage device include (722) one or more time metrics. In some embodiments, storage controller 124 (Figure 1A) or a component thereof (e.g., metrics module 222-1, Figure 2A-1) generates and/or maintains one or more time metrics. In some embodiments, the one or more time metrics include a wall-clock time.

[00250] In some embodiments, the one or more metrics of the storage device include (724) values of the one or more metrics from more than one time. For example, in some embodiments, the one or more metrics of the storage device include a count of cumulative writes to the storage device (e.g., to the storage device's storage medium or storage media) at a first time and a count of cumulative writes to the storage device (e.g., to the storage device's storage medium or storage media) at a second time. In some embodiments, values of the one or more metrics from more than one time include historical knowledge of the one or more metrics. For example, in some embodiments, the one or more metrics from more than one time include a running average of the one or more metrics. In another example, historical knowledge can be used to determine (e.g., compute) one or more projected values of one or more metrics at a particular time in the future (e.g., an hour, day, week, or month in the future). In some embodiments, historical knowledge of the one or more metrics of the storage device is used to detect a trigger condition, as described below with respect to operation 706.

[00251] At a storage device of a storage system (702), the storage device (e.g., storage device 120, Figure 1A) detects (706) a trigger condition in accordance with the one or more metrics of the storage device. In some embodiments, the trigger condition is detected in accordance with a non-linear and/or linear combination of the one or more metrics. In some embodiments, the trigger condition is detected in accordance with historical knowledge of the one or more metrics, as described above with respect to operation 602 of Figure 6. In some

embodiments, a trigger detection module (e.g., trigger detection module 224-1, Figure 2A-1) is used to detect a trigger condition in accordance with the one or more metrics of the storage device, as described above with respect to Figure 2A-1. Furthermore, in some embodiments, prior to detecting the trigger condition (706), the storage device detects a wear condition and reduces over-provisioning of the non-volatile memory of the storage device, without reducing declared capacity of the non-volatile memory of the storage device, as described above with respect to operation 601 of Figure 6.

[00252] At a storage device of a storage system (702), the storage device (e.g., storage device 120, Figure 1A) enables (708) an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device. In some embodiments, the amelioration process includes a process to reduce utilization by a host, a process to reduce declared capacity of the non-volatile memory of the storage device, and/or a process to advertise a reduced declared capacity. In some embodiments, the amelioration process includes altering an encoding format (e.g., from TLC to SLC and/or changing the redundancy mechanism) of at least a portion of the non-volatile memory of the storage device. In some embodiments, altering the encoding format of at least a portion of the non-volatile memory of the storage device includes setting the encoding format of an entirety of the non-volatile memory of the storage device to a low-density physical encoding format, for example SLC. For example, prior to the amelioration process the storage device includes some blocks (e.g., 98%) encoded as TLC and other blocks (e.g., 2%) encoded as SLC, and after the amelioration process all blocks are encoded using the lower-density physical encoding format, SLC. The latter example may correspond to a storage device that initially stores all client data using TLC and all storage device metadata using SLC. The amelioration process converts all of the client data from TLC to SLC without changing the encoding format of the storage device metadata (SLC). In some embodiments, an enabling module (e.g., enabling module 226-1, Figure 2A-1) is used to enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device, as described above with respect to Figure 2A-1.

[00253] In some embodiments, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device includes a process to reduce utilization of the non-volatile memory of the storage device, for example as described above with respect to operation 608 of Figure 6.

[00254] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (726) notifying a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the storage device is operatively coupled of the trigger condition. In some embodiments, notifying the host of the trigger condition includes notifying the host with an unsolicited communication. For example, in some embodiments, the unsolicited communication includes an interrupt communication. As another example, in some embodiments, the unsolicited communication includes a remote direct memory access (RDMA). As yet another example, in some embodiments, the unsolicited communication includes a TCP connection request or a TCP data transmission. In some embodiments, the unsolicited communication includes any other form of unsolicited communication.

[00255] In some embodiments, the host includes (728) a client on behalf of which data is stored in the storage system (e.g., data storage system 100, Figure 1A; data storage system 140, Figure 1B; data storage system 170, Figure 1C). In some embodiments, the client is or includes an entity on behalf of which data is stored in the storage system. For example, in some embodiments, the host is (1) computer system 110 (Figure 1A) or a client process, module or application executed by computer system 110, (2) computer system 142 (Figure 1B) or a client process, module or application executed by computer system 142, and/or (3) computer system 172 (Figure 1C) or a client process, module or application executed by computer system 172.

[00256] In some embodiments, the host includes (730) a storage system controller of the storage system (e.g., data storage system 140, Figure 1B). In some embodiments, the storage system controller controls and/or coordinates operations among one or more storage devices. For example, in some embodiments, the host is storage system controller 150 (Figure 1B). In some of these embodiments, the data storage system (e.g., data storage system 140, Figure 1B) is called a scale-up system.

[00257] In some embodiments, the host includes (732) a cluster controller of the storage system (e.g., data storage system 170, Figure 1C). In some embodiments, the cluster controller controls and/or coordinates operations among one or more data storage subsystems, as shown for example in Figure 1C, where each of the data storage subsystems may be implemented as a data storage system having one or more storage devices (e.g., data storage system 140, Figure 1B). For example, in some embodiments, the host is cluster controller

180 (Figure 1C). In some of these embodiments, the data storage system (e.g., data storage system 170, Figure 1C) is called a scale-out system or sometimes known as a clustered storage system.

[00258] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (734): (1) receiving a query from a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the storage device is operatively coupled, and (2) in response to receiving the query, reporting the trigger condition. For example, in some embodiments, the host polls for the trigger condition and the storage device receives the query from the host and in response to receiving the query, reports the trigger condition.

[00259] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (736): (1) receiving a command from a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the storage device and/or a write request to the storage device. In some embodiments, the command includes a request for temperature of the storage device. In some embodiments, the command includes a request for some other status of the storage device. In some embodiments, the notification of the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the storage device, and the storage device (1) receives the read request from the host, and (2) in response to receiving the read request, the storage device sends data corresponding to the read request and a notification of the trigger condition.

[00260] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (738): (1) receiving a command from a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to

obtain information with respect to the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the storage device and/or a write request to the storage device. In some embodiments, the command includes a request for temperature of the storage device. In some embodiments, the command includes a request for some other status of the storage device. In some embodiments, the notification that prompts the host to obtain information with respect to the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the storage device, and the storage device (1) receives the read request from the host, and (2) in response to receiving the read request, the storage device sends data corresponding to the read request and a notification (e.g., by setting a notification bit) that prompts the host to obtain information with respect to the trigger condition. In some embodiments, the mechanism used for returning such a notification when responding to a command from the host is a SCSI deferred error or deferred error response code.

[00261] Although a few examples of notification are described here, those skilled in the art will appreciate that the embodiments described herein may be extended to other notification methods (e.g., read-only mode notification, as described below with respect to operation 1028 of Figure 10C).

[00262] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (740) scheduling the amelioration process to be performed on the storage device. For example, in some embodiments, the trigger condition feeds back to the storage device and the storage device enables the amelioration process by scheduling the amelioration process to be performed on the storage device. In some embodiments, however, the amelioration process is performed, at least in part, by an apparatus other than the storage device (e.g., performed at least in part by the host, or by a storage system controller or by a cluster controller of a data storage system that includes at least one storage device distinct from the storage device).

[00263] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (742) determining one or more parameters for the amelioration process. In some embodiments, the one or more parameters for the amelioration process include a level of urgency for the amelioration process, a target reduced declared capacity of the non-volatile memory of the storage device, and/or a target amount of reduction in utilization of the non-volatile memory of the storage device, or any combination

or subset thereof. For example, in some embodiments or in some circumstances, the one or more parameters for the amelioration process include a parameter indicating that the urgency level is high (e.g., the amelioration process needs to begin within the next hour) and a parameter indicating that at least 1 GB of storage capacity needs to be reduced in the storage device.

[00264] In some embodiments, enabling the amelioration process associated with the detected trigger condition further includes (744) reporting at least a subset of the one or more parameters for the amelioration process. For example, in some embodiments, enabling the amelioration process associated with the detected trigger condition further includes reporting a target reduction in storage capacity of the non-volatile memory in the storage device.

[00265] In some embodiments, after enabling the amelioration process (710): the storage device (1) re-evaluates the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, aborts the amelioration process to reduce declared capacity of the non-volatile memory of the storage device. For example, in some circumstances, the one or more metrics of the storage device may change such that the trigger condition is no longer valid (e.g., the amelioration process is no longer needed). For example, during the operation of the amelioration process, normal storage operations will continue to be performed (e.g., read, write, delete, trim, etc.). Normal storage operations include operations like trim that explicitly reduce the storage device utilization, possibly enough to merit aborting the amelioration process. Other storage activity such as garbage collection may also reduce utilization, possibly enough to merit aborting the amelioration process. In some embodiments, the trigger condition is (e.g., periodically, semi-continuously, irregularly, initially, finally, etc.) re-evaluated in accordance with the one or more metrics of the storage device, as the one or more metrics may change in value due to the amelioration process and/or normal storage operations (e.g., read, write, erase and trim or unmap operations). In some embodiments, a trigger detection module (e.g., trigger detection module 224-1, Figure 2A-1) and/or an enabling module (e.g., enabling module 226-1, Figure 2A-1) are used to, after enabling the amelioration process, (1) re-evaluate the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, abort the amelioration process to reduce declared capacity of the non-volatile memory of the storage device, as described above with respect to Figure 2A-1.

[00266] In some embodiments, any operations of method 700 described above are performed by a storage device, the storage device including (1) non-volatile memory (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) one or more processors, and (3) controller memory (e.g., non-volatile memory or volatile memory in or coupled to the controller) storing one or more programs, which when executed by the one or more processors cause the storage device to perform or control performance of any of the methods described herein.

[00267] In some embodiments, any operations of method 700 described above are performed by a storage device including means for performing any of the methods described herein.

[00268] In some embodiments, any operations of method 700 described above are performed by a storage system comprising (1) a storage medium (e.g., comprising one or more non-volatile storage devices, such as flash memory devices) (2) one or more processors, and (3) memory (e.g., non-volatile memory or volatile memory in the storage system) storing one or more programs, which when executed by the one or more processors cause the storage system to perform or control performance of any of the methods described herein.

[00269] Figures 8A-8D illustrate a flowchart representation of a method 800 of managing a storage system, in accordance with some embodiments. In some embodiments, method 800 is performed by a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C) or one or more components of the host (e.g., management module 121-2, Figure 1A, system management module 151, Figure 1B, or cluster management module 181, Figure 1C). In some embodiments, method 800 is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a host, such as the one or more processing units (CPUs) 152-1 of system management module 151-1, shown in Figures 1B and 2B-1, or the one or more processing units (CPUs) 182-1 of cluster management module 181-1, shown in Figures 1C and 2C-1. In some embodiments, method 800 is performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, or data storage system 170, Figure 1C) or one or more components of the storage system (e.g., computer system 110, Figure 1A, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). In some embodiments, some of the operations of method 800 are performed at a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, or computer

system 172, Figure 1C) and information is transmitted to a storage device (e.g., storage device 120, Figure 1A) and/or one or more subsystems of a storage system (e.g., storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). In some embodiments, method 800 is governed, at least in part, by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a host (processors of host not shown in Figures 1A, 1B, and 1C).

[00270] In some embodiments, the host includes (812) a client on behalf of which data is stored in the storage system (e.g., data storage system 100, Figure 1A; data storage system 140, Figure 1B; data storage system 170, Figure 1C). In some embodiments, the client is or includes an entity on behalf of which data is stored in the storage system. For example, in some embodiments, the host is computer system 110 (Figure 1A) or a client process or application executed by computer system 110, or is a computer system 142 (Figure 1B) or a client process or application executed by computer system 142, or computer system 172 (Figure 1C) or a client process or application executed by computer system 172.

[00271] In some embodiments, the host includes (814) a storage system controller of the storage system (e.g., data storage system 140, Figure 1B). In some embodiments, the storage system controller controls and/or coordinates operations among one or more storage devices. For example, in some embodiments, the host is storage system controller 150 (Figure 1B). In some of these embodiments, the data storage system (e.g., data storage system 140, Figure 1B) is called a scale-up system.

[00272] In some embodiments, the host includes (816) a cluster controller of the storage system (e.g., data storage system 170, Figure 1C). In some embodiments, the cluster controller controls and/or coordinates operations among one or more data storage subsystems, as shown for example in Figure 1C, where each of the data storage subsystems may be implemented as a data storage system having one or more storage devices (e.g., data storage system 140, Figure 1B). For example, in some embodiments, the host is cluster controller 180 (Figure 1C). In some of these embodiments, the data storage system (e.g., data storage system 170, Figure 1C) is called a scale-out system or a clustered storage system.

[00273] At a host to which a storage device of the storage system is operatively coupled (802), the host obtains (804) one or more metrics of the storage device (e.g., storage device 120, Figure 1A, a particular storage device 160, Figure 1B, or a particular storage device 194, Figure 1C), the storage device including non-volatile memory. In some embodiments, the storage device generates and/or maintains the one or more metrics of the

storage device, and the host obtains the one or more metrics from the storage device. In some embodiments, one or more subsystems of a storage system distinct from the storage device (e.g., storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C) generate and/or maintain the one or more metrics of the storage device, and the host obtains the one or more metrics from the one or more subsystems. In some embodiments, the host generates and/or maintains the one or more metrics of the storage device. In some embodiments, a metrics module (e.g., metrics module 222-2, Figure 2A-2) is used to obtain one or more metrics of the storage device, the storage device including non-volatile memory, as described above with respect to Figure 2A-2.

[00274] In some embodiments, the storage device comprises (818) one or more flash memory devices. In some embodiments, the storage device comprises a storage medium (e.g., storage medium 130, Figure 1A), and the storage medium comprises one or more non-volatile storage devices, such as flash memory devices. In some embodiments, the storage medium (e.g., storage medium 130, Figure 1A) is a single flash memory device, while in other embodiments the storage medium includes a plurality of flash memory devices. For example, in some embodiments, the storage medium includes dozens or hundreds of flash memory devices, organized in parallel memory channels, such as 16, 32 or 64 flash memory devices per memory channel, and 8, 16 or 32 parallel memory channels. In some embodiments, the non-volatile storage medium (e.g., storage medium 130, Figure 1A) includes NAND-type flash memory or NOR-type flash memory. In other embodiments, the storage medium comprises one or more other types of non-volatile storage devices.

[00275] In some embodiments, obtaining (804) one or more metrics of the storage device includes obtaining (820) at least one metric, of the one or more metrics, for each memory portion (e.g., block, page, region, etc.) of a plurality of memory portions of the storage device. In some embodiments, at least one metric is generated for each block of a plurality of blocks of the storage device. In some embodiments, at least one metric is generated for each page of a plurality of pages of the storage device. In some embodiments, at least one metric is generated for each region of a plurality of regions of the storage device. In some embodiments, some metrics are generated on a block basis, some metrics are generated on a page basis, some metrics are generated on a region basis, and/or some metrics are generated on a storage device basis.

[00276] In some embodiments, the one or more metrics of the storage device include (822) one or more status metrics corresponding to the storage device's ability to retain data.

In some embodiments, the host or a component thereof (e.g., metrics module 222-2, Figure 2A-2) generates and/or maintains one or more status metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the storage device. In some embodiments, the one or more status metrics indicate a respective memory portion's ability to retain data. In some embodiments, the one or more status metrics associated with a respective memory portion, of a plurality of memory portions of the storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more status metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a bytes written field indicating a number of bytes of data written to pages in the respective memory portion, (b) a program-erase (P/E) cycle field indicating a current count of the number of P/E cycles performed on the respective memory portion, (c) a bit error rate (BER) field indicating a number of errors included in a codeword read from pages of the respective memory portion, and (d) other usage information indicating the health, performance, and/or endurance of the respective memory portion, as it relates to the respective memory portion's ability to retain data. In some embodiments, the one or more status metrics indicate the storage device's ability, as a whole, to retain data. For example, as the storage device ages, the one or more status metrics reflect the storage device's diminished ability to retain data (e.g., data read from the storage device includes more errors as the storage device ages).

[00277] In some embodiments, the one or more metrics of the storage device include (824) one or more performance metrics corresponding to performance of the storage device. In some embodiments, the host or a component thereof (e.g., metrics module 222-2, Figure 2A-2) generates and/or maintains one or more performance metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the storage device. In some embodiments, the one or more performance metrics correspond to performance of a respective memory portion of the plurality of memory portions. In some embodiments, the one or more performance metrics associated with a respective memory portion, of a plurality of memory portions of the storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more performance metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a measure of latency, and (b) transaction time. In some embodiments, the one or more performance metrics correspond to performance of the storage device as a whole. For example, in some embodiments, the one

or more performance metrics include a measure of latency for the storage device and/or transaction time for the storage device.

[00278] In some embodiments, the one or more metrics of the storage device include (826) one or more wear metrics corresponding to wear on the storage device. In some embodiments, the host or a component thereof (e.g., metrics module 222-2, Figure 2A-2) generates and/or maintains one or more wear metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the storage device. In some embodiments, the one or more wear metrics correspond to wear on a respective memory portion of the plurality of memory portions. In some embodiments, the one or more wear metrics associated with a respective memory portion, of a plurality of memory portions of the storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more wear metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a count of cumulative writes to the respective memory portion, (b) a count of cumulative reads from the respective memory portion, (c) a count of P/E cycles performed on the respective memory portion, and (d) a BER for the respective memory portion. In some embodiments, the one or more wear metrics correspond to wear on the storage device as a whole. For example, in some embodiments, the one or more wear metrics include a count of cumulative writes to the storage device, a count of cumulative reads from the storage device, a count of P/E cycles performed on the storage device and/or a BER for the storage device.

[00279] In some embodiments, the one or more metrics of the storage device include (828) one or more time metrics. In some embodiments, the host or a component thereof (e.g., metrics module 222-2, Figure 2A-2) generates and/or maintains one or more time metrics. In some embodiments, the one or more time metrics include a wall-clock time.

[00280] In some embodiments, the one or more metrics of the storage device include (830) values of the one or more metrics from more than one time. For example, in some embodiments, the one or more metrics of the storage device include a count of cumulative writes to the storage device at a first time and a count of cumulative writes to the storage device at a second time. In some embodiments, values of the one or more metrics from more than one time include historical knowledge of the one or more metrics. For example, in some embodiments, the one or more metrics from more than one time include a running average of the one or more metrics. In another example, historical knowledge can be used to determine (e.g., compute) one or more projected values of one or more metrics at a particular time in the

future (e.g., an hour, day, week, or month in the future). In some embodiments, historical knowledge of the one or more metrics of the storage device is used to detect a trigger condition, as described below with respect to operation 806.

[00281] At a host to which a storage device of the storage system is operatively coupled (802), the host detects (806) a trigger condition in accordance with the one or more metrics of the storage device. In some embodiments, the trigger condition is detected in accordance with a non-linear and/or linear combination of the one or more metrics. In some embodiments, the trigger condition is detected in accordance with historical knowledge of the one or more metrics. In some embodiments, a trigger detection module (e.g., trigger detection module 224-2, Figure 2A-2) is used to detect a trigger condition in accordance with the one or more metrics of the storage device, as described above with respect to Figure 2A-2. Furthermore, in some embodiments, prior to detecting the trigger condition (806), the host detects a wear condition and reduces over-provisioning of the non-volatile memory of the storage device, without reducing declared capacity of the non-volatile memory of the storage device, as described above with respect to operation 601 of Figure 6.

[00282] At a host to which a storage device of the storage system is operatively coupled (802), the host enables (808) an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device. In some embodiments, the amelioration process includes a process to reduce utilization by a host, a process to reduce declared capacity of the non-volatile memory of the storage device, and/or a process to advertise a reduced declared capacity. In some embodiments, the amelioration process includes altering an encoding format (e.g., from TLC to SLC and/or changing the redundancy mechanism) of at least a portion of the non-volatile memory of the storage device. In some embodiments, altering the encoding format of at least a portion of the non-volatile memory of the storage device includes setting the encoding format of an entirety of the non-volatile memory of the storage device to a low-density physical encoding format, for example SLC. For example, prior to the amelioration process the storage device includes some blocks (e.g., 98%) encoded as TLC and other blocks (e.g., 2%) encoded as SLC, and after the amelioration process all blocks are encoded using the lower-density physical encoding format, SLC. The latter example may correspond to a storage device that initially stores all client data using TLC and all storage device metadata using SLC. The amelioration process converts all of the client data from TLC to SLC without changing the encoding format of the storage device metadata (SLC). In

some embodiments, an enabling module (e.g., enabling module 226-2, Figure 2A-2) is used to enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device, as described above with respect to Figure 2A-2.

[00283] In some embodiments, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device includes a process to reduce utilization of the non-volatile memory of the storage device, for example as described above with respect to operation 608 of Figure 6.

[00284] In some embodiments, enabling (808) the amelioration process associated with the detected trigger condition includes scheduling (832) the amelioration process to be performed on the storage device. For example, in some embodiments, the trigger condition feeds back to the host and the host enables the amelioration process by scheduling the amelioration process to be performed on the storage device.

[00285] In some embodiments, enabling (808) the amelioration process associated with the detected trigger condition includes determining (834) one or more parameters for the amelioration process. In some embodiments, the one or more parameters for the amelioration process include a level of urgency for the amelioration process, a target reduced declared capacity of the non-volatile memory of the storage device, and/or a target amount of reduction in utilization of the non-volatile memory of the storage device, or any combination or subset thereof. For example, in some embodiments, the one or more parameters for the amelioration process include a parameter indicating that the urgency level is high (e.g., the amelioration process needs to begin within the next hour) and a parameter indicating that at least 1 GB of storage capacity needs to be reduced in the storage device.

[00286] In some embodiments, enabling (808) the amelioration process associated with the detected trigger condition further includes conveying (836) at least a subset of the one or more parameters for the amelioration process to the storage device. For example, in some embodiments, enabling the amelioration process associated with the detected trigger condition further includes conveying to the storage device a target amount of storage capacity of the non-volatile memory that needs to be reduced in the storage device. As another example, in some embodiments, enabling the amelioration process associated with the detected trigger condition further includes conveying to the storage device which portions of the storage medium (e.g., storage medium 130, Figure 1A) to reformat by altering an encoding format.

[00287] In some embodiments, after enabling the amelioration process (810): the host re-evaluates the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, aborts the amelioration process to reduce declared capacity of the non-volatile memory of the storage device. For example, in some circumstances, the one or more metrics of the storage device may change such that the trigger condition is no longer valid (e.g., the amelioration process is no longer needed). For example, during the operation of an amelioration process (e.g., operation 606, Figure 6), normal storage operations will continue to be performed (e.g., read, write, delete, trim, etc.). Normal storage operations include operations like trim that explicitly reduce the storage device utilization, possibly enough to merit aborting the amelioration process. Other storage activity such as garbage collection may also reduce utilization, possibly enough to merit aborting the amelioration process. In some embodiments, the trigger condition is (e.g., periodically, semi-continuously, irregularly, initially, finally, etc.) re-evaluated in accordance with the one or more metrics of the storage device, as the one or more metrics may change in value due to the amelioration process and/or normal storage operations (e.g., read, write, erase and trim or unmap operations). In some embodiments, a trigger detection module (e.g., trigger detection module 224-2, Figure 2A-2) and/or an enabling module (e.g., enabling module 226-2, Figure 2A-2) are used to, after enabling the amelioration process, (1) re-evaluate the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, abort the amelioration process to reduce declared capacity of the non-volatile memory of the storage device, as described above with respect to Figure 2A-2.

[00288] In some embodiments, any of the methods described above are performed by a storage system, the storage system including (1) one or more storage devices (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) a host to which the one or more storage devices are operatively coupled, (3) one or more processors, and (4) controller memory storing one or more programs, which when executed by the one or more processors cause the host to perform or control performance of any of the methods described herein.

[00289] In some embodiments, any of the methods described above are performed by a host system, coupled to one or more storage devices, the host system including means for performing any of the methods described herein.

[00290] In some embodiments, any operations of method 800 described above are performed by a host system comprising (1) an interface for operatively coupling to a storage system, (2) one or more processors, and (3) controller memory storing one or more programs, which when executed by the one or more processors cause the host system to perform or control performance of any of the methods described herein.

[00291] Figures 9A-9D illustrate a flowchart representation of a method 900 of managing a storage system, in accordance with some embodiments. At least in some embodiments, method 900 is performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, or data storage system 170, Figure 1C) or one or more components of the storage system (e.g., computer system 110, Figure 1A, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). In some embodiments, method 900 is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a storage system, such as the one or more processing units (CPUs) 152-1 of system management module 151-1, shown in Figures 1B and 2B-1, the one or more processing units (CPUs) 182-1 of cluster management module 181-1, shown in Figures 1C and 2C-1, or one or more processors of an included host (e.g., computer system 110, Figure 1A), shown in Figure 2A-2. For ease of explanation, the following describes method 900 as performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, or data storage system 170, Figure 1C). However, those skilled in the art will appreciate that in other embodiments, one or more of the operations described in method 900 are performed by one or more subsystems of the storage system distinct from the storage device (e.g., storage system controller 150, Figure 1B or cluster controller 180, Figure 1C).

[00292] A storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, or data storage system 170, Figure 1C) obtains, (902) for each storage device (e.g., storage device 120, Figure 1A, or any of storage devices 160-1 to 160-m of Figure 1B, or any of storage devices 194-1 to 194-n or 194-j to 194-k of Figure 1C) of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory. Although Figure 1A only shows one storage device 120, in some embodiments, data storage system 100 of Figure 1A includes a plurality of storage devices, of which storage device 120 is one example. In some embodiments, the storage device generates and/or maintains the one or more metrics of the storage device, and the storage system obtains the one or more metrics from the storage

device. In some embodiments, one or more subsystems of a storage system distinct from the storage device (e.g., storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C) generate and/or maintain the one or more metrics of the storage device, and the storage system obtains the one or more metrics from the one or more subsystems. In some embodiments, a metrics module (e.g., metrics module 222, Figures 2A-1 and 2A-2; metrics module 252-1, Figure 2B-1; or metrics module 282-1, Figure 2C-1) is used to obtain, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1.

[00293] In some embodiments, obtaining (902) one or more metrics of a respective storage device (e.g., storage device 120 of Figure 1A, a respective storage device 160 of Figure 1B, or a respective storage device 194 of Figure 1C) of the plurality of storage devices of the storage system includes obtaining (912) at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the respective storage device. In some embodiments, at least one metric is generated for each block of a plurality of blocks of the respective storage device. In some embodiments, at least one metric is generated for each page of a plurality of pages of the respective storage device. In some embodiments, at least one metric is generated for each region of a plurality of regions of the respective storage device. In some embodiments, some metrics are generated on a block basis, some metrics are generated on a page basis, some metrics are generated on a region basis, and/or some metrics are generated on a storage device basis.

[00294] In some embodiments, the one or more metrics of the respective storage device include (914) one or more status metrics corresponding to the respective storage device's ability to retain data. In some embodiments, the storage system or a component thereof (e.g., metrics module 222-1, Figure 2A-1, of the respective storage device; metrics module 252-1, Figure 2B-1, of storage system controller 150; or metrics module 282-1, Figure 2C-1, of cluster controller 180) generates and/or maintains one or more status metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the respective storage device. In some embodiments, the one or more status metrics indicate a respective memory portion's ability to retain data. In some embodiments, the one or more status metrics associated with a respective memory portion, of a plurality of memory portions of the respective storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more

status metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a bytes written field indicating a number of bytes of data written to pages in the respective memory portion, (b) a program-erase (P/E) cycle field indicating a current count of the number of P/E cycles performed on the respective memory portion, (c) a bit error rate (BER) field indicating a number of errors included in a codeword read from pages of the respective memory portion, and (d) other usage information indicating the health, performance, and/or endurance of the respective memory portion, as it relates to the respective memory portion's ability to retain data. In some embodiments, the one or more status metrics indicate the respective storage device's ability, as a whole, to retain data. For example, as the respective storage device ages, the one or more status metrics reflect the respective storage device's diminished ability to retain data (e.g., data read from the respective storage device includes more errors as the respective storage device ages).

[00295] In some embodiments, the one or more metrics of the respective storage device include (916) one or more performance metrics corresponding to performance of the respective storage device. In some embodiments, the storage system or a component thereof (e.g., metrics module 222-1, Figure 2A-1, of the respective storage device; metrics module 252-1, Figure 2B-1, of storage system controller 150; or metrics module 282-1, Figure 2C-1, of cluster controller 180) generates and/or maintains one or more performance metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the respective storage device. In some embodiments, the one or more performance metrics correspond to performance of a respective memory portion of the plurality of memory portions. In some embodiments, the one or more performance metrics associated with a respective memory portion, of a plurality of memory portions of the respective storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more performance metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a measure of latency, and (b) transaction time. In some embodiments, the one or more performance metrics correspond to performance of the respective storage device as a whole. For example, in some embodiments, the one or more performance metrics include a measure of latency for the respective storage device and/or transaction time for the respective storage device.

[00296] In some embodiments, the one or more metrics of the respective storage device include (918) one or more wear metrics corresponding to wear on the respective

storage device. In some embodiments, the storage system or a component thereof (e.g., metrics module 222-1, Figure 2A-1, of the respective storage device; metrics module 252-1, Figure 2B-1, of storage system controller 150; or metrics module 282-1, Figure 2C-1, of cluster controller 180) generates and/or maintains one or more wear metrics for each memory portion of a plurality of memory portions (e.g., in storage medium 130, Figure 1A) of the respective storage device. In some embodiments, the one or more wear metrics correspond to wear on a respective memory portion of the plurality of memory portions. In some embodiments, the one or more wear metrics associated with a respective memory portion, of a plurality of memory portions of the respective storage device, are stored in a characterization vector corresponding to the respective memory portion. In some embodiments, the one or more wear metrics stored in the characterization vector for the respective memory portion include a subset or superset of: (a) a count of cumulative writes to the respective memory portion, (b) a count of cumulative reads from the respective memory portion, (c) a count of P/E cycles performed on the respective memory portion, and (d) a BER for the respective memory portion. In some embodiments, the one or more wear metrics correspond to wear on the respective storage device as a whole. For example, in some embodiments, the one or more wear metrics include a count of cumulative writes to the respective storage device, a count of cumulative reads from the respective storage device, a count of P/E cycles performed on the respective storage device and/or a BER for the respective storage device.

[00297] In some embodiments, the one or more metrics of the respective storage device include (920) one or more time metrics. In some embodiments, the storage system or a component thereof (e.g., metrics module 222-1, Figure 2A-1, of the respective storage device; metrics module 252-1, Figure 2B-1, of storage system controller 150; or metrics module 282-1, Figure 2C-1, of cluster controller 180) generates and/or maintains one or more time metrics. In some embodiments, the one or more time metrics include a wall-clock time.

[00298] In some embodiments, the one or more metrics of the respective storage device include (922) values of the one or more metrics from more than one time. For example, in some embodiments, the one or more metrics of the respective storage device include a count of cumulative writes to the respective storage device at a first time and a count of cumulative writes to the respective storage device at a second time. In some embodiments, values of the one or more metrics from more than one time include historical knowledge of the one or more metrics. For example, in some embodiments, the one or more

metrics from more than one time include a running average of the one or more metrics. In another example, historical knowledge can be used to determine (e.g., compute) one or more projected values of one or more metrics at a particular time in the future (e.g., an hour, day, week, or month in the future). In some embodiments, historical knowledge of the one or more metrics of the respective storage device is used to detect a trigger condition, as described below with respect to operation 904.

[00299] The storage system detects (904) a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of one or more storage devices of the plurality of storage devices. Alternatively, in some embodiments, the trigger condition detected in accordance with the one or more metrics of one or more storage devices of the plurality of storage devices is a trigger condition for reducing declared capacity of the non-volatile memory of a subset (e.g., one or more, including all) of the one or more storage devices, and the associated amelioration process described below with respect to operation 908 is an amelioration process to reduce declared capacity of the non-volatile memory of a subset (e.g., one or more, including all) of the one or more storage devices. In some embodiments, the trigger condition is detected in accordance with a non-linear and/or linear combination of the one or more metrics. In some embodiments, the trigger condition is detected in accordance with historical knowledge of the one or more metrics. In some embodiments, a trigger detection module (e.g., trigger detection module 224, Figures 2A-1 and 2A-2; trigger detection module 254-1, Figure 2B-1; or trigger detection module 284-1, Figure 2C-1) is used to detect a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of one or more storage devices of the plurality of storage devices, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1. Examples of trigger conditions are discussed above with reference to Figure 6. Furthermore, in some embodiments, prior to detecting the trigger condition (904), the storage system detects a wear condition and reduces over-provisioning of the non-volatile memory of the respective storage device, without reducing declared capacity of the non-volatile memory of the respective storage device, as described above with respect to operation 601 of Figure 6.

[00300] In some embodiments, one or more metrics from the plurality of storage devices of the storage system are combined to detect the trigger condition. For example, it

may be advantageous to examine the rate of change of historical wear metrics such as P/E cycle counts in the plurality of storage devices in order to allow the amelioration process sufficient time to complete before any one device reaches a wear limit.

[00301] In some embodiments, the respective storage device comprises (906) one or more flash memory devices. In some embodiments, the respective storage device comprises a storage medium (e.g., storage medium 130, Figure 1A), and the storage medium comprises one or more non-volatile storage devices, such as flash memory devices. In some embodiments, the storage medium (e.g., storage medium 130, Figure 1A) is a single flash memory device, while in other embodiments the storage medium includes a plurality of flash memory devices. For example, in some embodiments, the storage medium includes dozens or hundreds of flash memory devices, organized in parallel memory channels, such as 16, 32 or 64 flash memory devices per memory channel, and 8, 16 or 32 parallel memory channels. In some embodiments, the non-volatile storage medium (e.g., storage medium 130, Figure 1A) includes NAND-type flash memory or NOR-type flash memory. In other embodiments, the storage medium comprises one or more other types of non-volatile storage devices.

[00302] The storage system enables (908) an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device. In some embodiments, the amelioration process includes altering an encoding format (e.g., from TLC to SLC and/or changing the redundancy mechanism) of at least a portion of the non-volatile memory of the respective storage device. In some embodiments, altering the encoding format of at least a portion of the non-volatile memory of the storage device includes setting the encoding format of an entirety of the non-volatile memory of the storage device to a low-density physical encoding format, for example SLC. For example, prior to the amelioration process the storage device includes some blocks (e.g., 98%) encoded as TLC and other blocks (e.g., 2%) encoded as SLC, and after the amelioration process all blocks are encoded using the lower-density physical encoding format, SLC. The latter example may correspond to a storage device that initially stores all client data using TLC and all storage device metadata using SLC. The amelioration process converts all of the client data from TLC to SLC without changing the encoding format of the storage device metadata (SLC).

[00303] In some embodiments, the amelioration process includes a process to reduce utilization by a host, a process to reduce declared capacity of the non-volatile memory of the respective storage device, and/or a process to advertise a reduced declared capacity. In some

embodiments, an enabling module (e.g., enabling module 226, Figures 2A-1 and 2A-2; enabling module 256-1, Figure 2B-1; or enabling module 286-1, Figure 2C-1) is used to enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1.

[00304] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (924) notifying a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the respective storage device is operatively coupled of the trigger condition. In some embodiments, notifying the host of the trigger condition includes notifying the host with an unsolicited communication. For example, in some embodiments, the unsolicited communication includes an interrupt communication.

[00305] In some embodiments, the host includes (926) a client on behalf of which data is stored in the storage system (e.g., data storage system 100, Figure 1A; data storage system 140, Figure 1B; data storage system 170, Figure 1C). In some embodiments, the client is or includes an entity on behalf of which data is stored in the storage system. For example, in some embodiments, the host is (1) computer system 110 (Figure 1A) or a client process, module or application executed by computer system 110, (2) computer system 142 (Figure 1B) or a client process or application executed by computer system 142, and/or (3) computer system 172 (Figure 1C) or (4) a client process, module or application executed by computer system 172.

[00306] In some embodiments, the host includes (928) a storage system controller of the storage system (e.g., data storage system 140, Figure 1B). In some embodiments, the storage system controller controls and/or coordinates operations among one or more storage devices. For example, in some embodiments, the host is storage system controller 150 (Figure 1B). In some of these embodiments, the data storage system (e.g., data storage system 140, Figure 1B) is called a scale-up system.

[00307] In some embodiments, the host includes (930) a cluster controller of the storage system (e.g., data storage system 170, Figure 1C). In some embodiments, the cluster controller controls and/or coordinates operations among one or more data storage subsystems, as shown for example in Figure 1C, where each of the data storage subsystems may be implemented as a data storage system having one or more storage devices (e.g., data storage

system 140, Figure 1B). For example, in some embodiments, the host is cluster controller 180 (Figure 1C). In some of these embodiments, the data storage system (e.g., data storage system 170, Figure 1C) is called a scale-out system.

[00308] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (932): (1) receiving a query from a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the respective storage device is operatively coupled, and (2) in response to receiving the query, reporting the trigger condition. For example, in some embodiments, computer system 142 (Figure 1B) polls for the trigger condition and storage system controller 150 (Figure 1B) receives the query from computer system 142 and in response to receiving the query, reports the trigger condition. As another example, in some embodiments, computer system 172 (Figure 1C) polls for the trigger condition and cluster controller 150 (Figure 1C) receives the query from computer system 172 and in response to receiving the query, reports the trigger condition. As yet another example, in some embodiments, the host polls for the trigger condition and the respective storage device receives the query from the host and in response to receiving the query, reports the trigger condition.

[00309] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (934): (1) receiving a command from a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the respective storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the respective storage device and/or a write request to the respective storage device. In some embodiments, the command includes a request for temperature of the respective storage device. In some embodiments, the command includes a request for some other status of the respective storage device. In some embodiments, the notification of the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the respective storage device, and the respective storage device (1) receives the read request from the host, and (2) in response to receiving the read request, the respective storage device sends data corresponding to the read request and a notification of

the trigger condition. As another example, in some embodiments, the host (e.g., computer system 142, Figure 1B; or computer system 172, Figure 1C) issues a read request for data, and the storage system or one or more components of the storage system (e.g., storage system controller 150, Figure 1B; or cluster controller 180, Figure 1C) (1) receives the read request from the host, and (2) in response to receiving the read request, the storage system or one or more components of the storage system sends data corresponding to the read request and a notification of the trigger condition.

[00310] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (936): (1) receiving a command from a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) to which the respective storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the respective storage device and/or a write request to the respective storage device. In some embodiments, the command includes a request for temperature of the respective storage device. In some embodiments, the command includes a request for some other status of the respective storage device. In some embodiments, the notification that prompts the host to obtain information with respect to the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the respective storage device, and the respective storage device (1) receives the read request from the host, and (2) in response to receiving the read request, the respective storage device sends data corresponding to the read request and a notification (e.g., by setting a notification bit) that prompts the host to obtain information with respect to the trigger condition. As another example, in some embodiments, the host (e.g., computer system 142, Figure 1B; or computer system 172, Figure 1C) issues a read request for data, and the storage system or one or more components of the storage system (e.g., storage system controller 150, Figure 1B; or cluster controller 180, Figure 1C) (1) receives the read request from the host, and (2) in response to receiving the read request, the respective storage device sends data corresponding to the read request and a notification (e.g., by setting a notification bit) that prompts the host to obtain information with respect to the trigger condition. In some embodiments, the mechanism used for returning such a

notification when responding to a command from the host is a SCSI deferred error or deferred error response code.

[00311] Although a few examples of notification are described here, those skilled in the art will appreciate that the embodiments described herein may be extended to other notification methods, such as notification methods described herein with reference to method 1000.

[00312] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (938) scheduling the amelioration process to be performed on the respective storage device. For example, in some embodiments, the trigger condition feeds back to storage system controller 150 (Figure 1B) and storage system controller 150 enables the amelioration process by scheduling the amelioration process to be performed on the respective storage device. As another example, in some embodiments, the trigger condition feeds back to cluster controller 180 (Figure 1C) and cluster controller 180 enables the amelioration process by scheduling the amelioration process to be performed on the respective storage device.

[00313] In some embodiments, enabling the amelioration process associated with the detected trigger condition includes (940) determining one or more parameters for the amelioration process. In some embodiments, the one or more parameters for the amelioration process include a level of urgency for the amelioration process and a targeted amount of storage capacity of the non-volatile memory that needs to be reduced in the respective storage device. For example, in some embodiments, the one or more parameters for the amelioration process include a parameter indicating that the urgency level is high (e.g., the amelioration process needs to begin within the next hour) and a parameter indicating that 1 GB of storage capacity needs to be reduced in the respective storage device.

[00314] In some embodiments, enabling the amelioration process associated with the detected trigger condition further includes (942) reporting at least a subset of the one or more parameters for the amelioration process. For example, in some embodiments, enabling the amelioration process associated with the detected trigger condition further includes reporting a targeted amount of storage capacity of the non-volatile memory that needs to be reduced in the respective storage device.

[00315] In some embodiments, the obtaining, the enabling, or both the obtaining and the enabling are performed (944) by one or more subsystems of the storage system distinct

from the plurality of storage devices. For example, in some of these embodiments, the obtaining, the enabling, or both the obtaining and the enabling are performed by a storage system controller (e.g., storage system controller 150, Figure 1B) of the storage system (e.g., data storage system 140, Figure 1B). As another example, in some of these embodiments, the obtaining, the enabling, or both the obtaining and the enabling are performed by a cluster controller (e.g., cluster controller 180, Figure 1C) of the storage system (e.g., data storage system 170, Figure 1C). In some embodiments, method 900, or at least the performing operation 902 and/or enabling operation 908 of method 900, is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a device, such as the one or more processing units (CPUs) 152-1 of system management module 151-1, shown in Figures 1B and 2B-1 or the one or more processing units (CPUs) 182-1 of cluster management module 181-1, shown in Figures 1C and 2C-1.

[00316] In some embodiments, after enabling the amelioration process (910): the storage system re-evaluates the trigger condition in accordance with the one or more metrics of the one or more storage devices of the plurality of storage devices, and (2) in accordance with a determination that the trigger condition is no longer valid, aborts the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device. For example, in some embodiments, the one or more metrics may change such that the trigger condition is no longer present (e.g., the amelioration process is no longer needed). For example, normal storage operations (read, write, trim, etc.) performed coincident in time with the amelioration process may cause the metrics to change so that the trigger condition is no longer present. In some embodiments, a trigger detection module (e.g., trigger detection module 224, Figures 2A-1 and 2A-2; trigger detection module 254-1, Figure 2B-1; or trigger detection module 284-1, Figure 2C-1) and/or an enabling module (e.g., enabling module 226, Figures 2A-1 and 2A-2; enabling module 256-1, Figure 2B-1; or enabling module 286-1, Figure 2C-1) are used to, after enabling the amelioration process, (1) re-evaluate the trigger condition in accordance with the one or more metrics of the one or more storage devices of the plurality of storage devices, and (2) in accordance with a determination that the trigger condition is no longer valid, abort the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1.

[00317] In some embodiments, any of the method described above are performed by a storage system, the storage system including (1) non-volatile memory, (2) one or more processors, and (3) controller memory storing one or more programs, which when executed by the one or more processors cause the storage system to perform or control performance of any of the methods described herein.

[00318] In some embodiments, any of the methods described above are performed by a storage system including means for performing any of the methods described herein.

[00319] In some embodiments, any of the methods described above are performed by a storage system, the storage system including (1) a plurality of storage devices, (2) one or more subsystems having one or more processors, and (3) memory storing one or more programs, which when executed by the one or more processors cause the one or more subsystems to perform or control performance of any of the methods described herein.

[00320] Figures 10A-10C illustrate a flowchart representation of a method 1000 of managing a storage system, in accordance with some embodiments. At least in some embodiments, method 1000 is performed by a storage device (e.g., storage device 120, Figure 1A, storage device 160, Figure 1B, or storage device 194, Figure 1C) or one or more components of the storage device (e.g., storage controller 124, Figure 1A), wherein the storage device is operatively coupled with a host system (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). In some embodiments, method 1000 is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a device, such as the one or more processing units (CPUs) 122-1 of management module 121-1, shown in Figures 1A and 2A-1. In some embodiments, method 1000 is performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, and/or data storage system 170, Figure 1C) or one or more components of the storage system (e.g., storage device 120, Figure 1A, storage device 160, Figure 1B, or storage device 194, Figure 1C). In some embodiments, some of the operations of method 1000 are performed at a storage device (e.g., storage device 120, Figure 1A, storage device 160, Figure 1B, or storage device 194, Figure 1C) and information is transmitted to a host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). For ease of explanation, the following describes method 1000 as performed by a storage device (e.g., storage device 120, Figure 1A)

of a storage system (e.g., data storage system 100, Figure 1A). However, those skilled in the art will appreciate that in other embodiments, one or more of the operations described in method 1000 are performed by a storage device of another storage system (e.g., storage device 160 of data storage system 140, Figure 1B, or storage device 194 of data storage system 170, Figure 1C).

[00321] At a storage device of the storage system, the storage device including non-volatile memory (1002), the storage device (e.g., storage device 120, Figure 1A) detects (1004) a trigger condition for reducing declared capacity of the non-volatile memory of the storage device. In some embodiments, the trigger condition is detected in accordance with a non-linear and/or linear combination of one or more metrics of the storage device. In some embodiments, the trigger condition is detected in accordance with historical knowledge of the one or more metrics, as described above with respect to operation 602 of Figure 6. In some embodiments, the trigger condition is detected when the storage device transitions from normal operation to read-only mode. In some embodiments, a trigger detection module (e.g., trigger detection module 224-1, Figure 2A-1) is used to detect a trigger condition for reducing declared capacity of the non-volatile memory of the storage device, as described above with respect to Figure 2A-1. Furthermore, in some embodiments, prior to detecting the trigger condition (1004), the storage device detects a wear condition and reduces over-provisioning of the non-volatile memory of the storage device, without reducing declared capacity of the non-volatile memory of the storage device, as described above with respect to operation 601 of Figure 6.

[00322] In some embodiments, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device includes a process to reduce utilization of the non-volatile memory of the storage device, for example as described above with respect to operation 608 of Figure 6.

[00323] In some embodiments, the storage device comprises (1006) one or more flash memory devices. In some embodiments, the storage device comprises a storage medium (e.g., storage medium 130, Figure 1A), and the storage medium comprises one or more non-volatile storage devices, such as flash memory devices. In some embodiments, the storage medium (e.g., storage medium 130, Figure 1A) is a single flash memory device, while in other embodiments the storage medium includes a plurality of flash memory devices. For example, in some embodiments, the storage medium includes dozens or hundreds of flash memory devices, organized in parallel memory channels, such as 16, 32 or 64 flash memory

devices per memory channel, and 8, 16 or 32 parallel memory channels. In some embodiments, the non-volatile storage medium (e.g., storage medium 130, Figure 1A) includes NAND-type flash memory or NOR-type flash memory. In other embodiments, the storage medium comprises one or more other types of non-volatile storage devices.

[00324] In some embodiments, the trigger condition is (1008) detected by the storage device in accordance with one or more metrics of the storage device. In some embodiments, the one or more metrics of the storage device include one or more status metrics corresponding to the storage device's ability to retain data, one or more performance metrics corresponding to performance of the storage device, one or more wear metrics corresponding to wear on the storage device, and/or one or more time metrics, as described above with respect to operation 602 of Figure 6. Additional example of metrics of a storage device are described herein with respect to method 900. Examples of trigger conditions are discussed above with reference to Figure 6.

[00325] At a storage device of the storage system, the storage device including non-volatile memory (1002), the storage device (e.g., storage device 120, Figure 1A) notifies (1010) a host to which the storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the storage device. In some embodiments, a notification module (e.g., notification module 228-1, Figure 2A-1) is used to notify a host to which the storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device, as described above with respect to Figure 2A-1.

[00326] In some embodiments, the host includes (1014) a client on behalf of which data is stored in the storage system (e.g., data storage system 100, Figure 1A; data storage system 140, Figure 1B; data storage system 170, Figure 1C). In some embodiments, the client is or includes an entity on behalf of which data is stored in the storage system. For example, in some embodiments, the host is (1) computer system 110 (Figure 1A) or a client process, module or application executed by computer system 110, (2) computer system 142 (Figure 1B) or a client process, module or application executed by computer system 142, and/or (3) computer system 172 (Figure 1C) or a client process, module or application executed by computer system 172.

[00327] In some embodiments, the host includes (1016) a storage system controller of the storage system (e.g., data storage system 140, Figure 1B). In some embodiments, the

storage system controller controls and/or coordinates operations among one or more storage devices. For example, in some embodiments, the host is storage system controller 150 (Figure 1B). In some of these embodiments, the data storage system (e.g., data storage system 140, Figure 1B) is called a scale-up system.

[00328] In some embodiments, the host includes (1018) a cluster controller of the storage system (e.g., data storage system 170, Figure 1C). In some embodiments, the cluster controller controls and/or coordinates operations among one or more data storage subsystems, as shown for example in Figure 1C, where each of the data storage subsystems may be implemented as a data storage system having one or more storage devices (e.g., data storage system 140, Figure 1B). For example, in some embodiments, the host is cluster controller 180 (Figure 1C). In some of these embodiments, the data storage system (e.g., data storage system 170, Figure 1C) is called a scale-out system, sometimes known as a clustered storage system.

[00329] In some embodiments, notifying (1010) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes notifying (1020) the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) with an unsolicited communication. For example, in some embodiments, the unsolicited communication includes an interrupt communication. As another example, in some embodiments, the unsolicited communication includes a remote direct memory access (RDMA). As yet another example, in some embodiments, the unsolicited communication includes a TCP connection request. In some embodiments, the unsolicited communication includes any other form of unsolicited communication.

[00330] In some embodiments, notifying (1010) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes (1022): (1) receiving a query from the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C), and (2) in response to receiving the query, reporting the trigger condition. For example, in some embodiments, the host polls for the trigger condition and the storage device receives the query from the host and in response to receiving the query, reports the trigger condition.

[00331] In some embodiments, notifying (1010) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes (1024):

(1) receiving a command from the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C), and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the storage device and/or a write request to the storage device. In some embodiments, the command includes a request for temperature of the storage device. In some embodiments, the command includes a request for some other status of the storage device. In some embodiments, the notification of the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the storage device, and the storage device (1) receives the read request from the host, and (2) in response to receiving the read request, the storage device sends data corresponding to the read request and a notification of the trigger condition.

[00332] In some embodiments, notifying (1010) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes (1026): (1) receiving a command from the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C), and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the storage device and/or a write request to the storage device. In some embodiments, the command includes a request for temperature of the storage device. In some embodiments, the command includes a request for some other status of the storage device. In some embodiments, the notification that prompts the host to obtain information with respect to the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the storage device, and the storage device (1) receives the read request from the host, and (2) in response to receiving the read request, the storage device sends data corresponding to the read request and a notification (e.g., by setting a notification bit) that prompts the host to obtain information with respect to the trigger condition. In some embodiments, the mechanism used for returning a notification when responding to a command from the host is a SCSI deferred error or deferred error response code.

[00333] In some embodiments, notifying (1010) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device further includes notifying (1028) the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) that the storage device is in read-only mode. In some embodiments, notifying the host that the storage device is in read-only mode includes notifying the host that the storage device is not accepting any more write commands. In some embodiments, notifying the host that the storage device is in read-only mode includes sending one or more rejections of write commands.

[00334] In some embodiments, after notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device (1012), the storage device (1) re-evaluates the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, notifies the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device. For example, in some circumstances, the one or more metrics of the storage device may change such that the trigger condition is no longer valid (e.g., the amelioration process is no longer needed). For example, during the operation of an amelioration process (e.g., operation 606, Figure 6), normal storage operations will continue to be performed (e.g., read, write, delete, trim, etc.). Normal storage operations include operations like trim that explicitly reduce the storage device utilization, possibly enough to merit aborting the amelioration process. Other storage activity such as garbage collection may also reduce utilization, possibly enough to merit aborting the amelioration process.

[00335] In some embodiments, the trigger condition (e.g., periodically, semi-continuously, initially, finally, occasionally or irregularly) is recomputed or re-evaluated in accordance with the one or more metrics of the storage device, as the one or more metrics may change in value due to the amelioration process and/or normal storage operations (e.g., read, write, erase and trim or unmap operations). In some embodiments, a trigger detection module (e.g., trigger detection module 224-1, Figure 2A-1) and/or a notification module (e.g., notification module 228-1, Figure 2A-1) are used to, after notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device: (1) re-evaluate the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer

valid, notify the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device, as described above with respect to Figure 2A-1.

[00336] In some embodiments, any operations of method 1000 described above are performed by a storage device, the storage device including (1) non-volatile memory (e.g., comprising one or more non-volatile storage devices, such as flash memory devices) (2) one or more processors, and (3) controller memory (e.g., non-volatile memory or volatile memory in or coupled to the controller) storing one or more programs, which when executed by the one or more processors cause the storage device to perform or control performance of any of the methods described herein.

[00337] In some embodiments, any operations of method 1000 described above are performed by a storage device including means for performing any of the methods described herein.

[00338] In some embodiments, any operations of method 1000 described above are performed by a storage system comprising (1) a storage medium (e.g., comprising one or more non-volatile storage devices, such as flash memory devices) (2) one or more processors, and (3) memory (e.g., non-volatile memory or volatile memory in the storage system) storing one or more programs, which when executed by the one or more processors cause the storage system to perform or control performance of any of the methods described herein.

[00339] Figures 11A-11C illustrate a flowchart representation of a method 1100 of managing a storage system, in accordance with some embodiments. At least in some embodiments, method 1100 is performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, or data storage system 170, Figure 1C) or one or more components of the storage system (e.g., computer system 110, Figure 1A, storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C). In some embodiments, method 1100 is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a storage system, such as the one or more processing units (CPUs) 152-1 of system management module 151-1, shown in Figures 1B and 2B-1, the one or more processing units (CPUs) 182-1 of cluster management module 181-1, shown in Figures 1C and 2C-1, or one or more processors of an included host (e.g., computer system 110, Figure 1A), shown in Figure 2A-2. For ease of explanation, the following describes method 1100 as performed by a storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure

1B, or data storage system 170, Figure 1C). However, those skilled in the art will appreciate that in other embodiments, one or more of the operations described in method 1100 are performed by one or more subsystems of the storage system distinct from the storage device (e.g., storage system controller 150, Figure 1B or cluster controller 180, Figure 1C).

[00340] A storage system (e.g., data storage system 100, Figure 1A, data storage system 140, Figure 1B, or data storage system 170, Figure 1C) obtains (1102), for each storage device (e.g., storage device 120, Figure 1A, or any of storage devices 160-1 to 160-m of Figure 1B, or any of storage devices 194-1 to 194-n or 194-j to 194-k of Figure 1C) of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory. Although Figure 1A only shows one storage device 120, in some embodiments, data storage system 100 of Figure 1A includes a plurality of storage devices, of which storage device 120 is one example. In some embodiments, the storage device generates and/or maintains the one or more metrics of the storage device, and the storage system obtains the one or more metrics from the storage device. In some embodiments, one or more subsystems of a storage system distinct from the storage device (e.g., storage system controller 150, Figure 1B, or cluster controller 180, Figure 1C) generate and/or maintain the one or more metrics of the storage device, and the storage system obtains the one or more metrics from the one or more subsystems. In some embodiments, a metrics module (e.g., metrics module 222, Figures 2A-1 and 2A-2; metrics module 252-1, Figure 2B-1; or metrics module 282-1, Figure 2C-1) is used to obtain, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1.

[00341] The storage system detects (1104) a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of two or more of the storage devices of the plurality of storage devices in the storage system. In some embodiments, the one or more metrics of the respective storage device include one or more status metrics corresponding to the respective storage device's ability to retain data, one or more performance metrics corresponding to performance of the respective storage device, one or more wear metrics corresponding to wear on the respective storage device, and/or one or more time metrics, as described above with respect to operation 602 of Figure 6. In some embodiments, the trigger condition is detected in accordance with a

non-linear and/or linear combination of the one or more metrics of the respective storage device. In some embodiments, the trigger condition is detected in accordance with historical knowledge of the one or more metrics of the respective storage device, as described above with respect to operation 602 of Figure 6. In some embodiments, the trigger condition is detected when the respective storage device transitions from normal operation to read-only mode. In some embodiments, one or more metrics from the plurality of storage devices of the storage system are combined to detect the trigger condition. For example, it may be advantageous to examine the rate of change of historical wear metrics such as P/E cycle counts in the plurality of storage devices in order to allow the amelioration process sufficient time to complete before any one device reaches a wear limit. In some embodiments, a trigger detection module (e.g., trigger detection module 224, Figures 2A-1 and 2A-2; trigger detection module 254-1, Figure 2B-1; or trigger detection module 284-1, Figure 2C-1) is used to detect a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of the respective storage device, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1. Furthermore, in some embodiments, prior to detecting the trigger condition (1104), the storage system detects a wear condition and reduces over-provisioning of the non-volatile memory of the respective storage device, without reducing declared capacity of the non-volatile memory of the respective storage device, as described above with respect to operation 601 of Figure 6.

[00342] In some embodiments, the respective storage device comprises (1106) one or more flash memory devices. In some embodiments, the respective storage device comprises a storage medium (e.g., storage medium 130, Figure 1A), and the storage medium comprises one or more non-volatile storage devices, such as flash memory devices. In some embodiments, the storage medium (e.g., storage medium 130, Figure 1A) is a single flash memory device, while in other embodiments the storage medium includes a plurality of flash memory devices. For example, in some embodiments, the storage medium includes dozens or hundreds of flash memory devices, organized in parallel memory channels, such as 16, 32 or 64 flash memory devices per memory channel, and 8, 16 or 32 parallel memory channels. In some embodiments, the non-volatile storage medium (e.g., storage medium 130, Figure 1A) includes NAND-type flash memory or NOR-type flash memory. In other embodiments, the storage medium comprises one or more other types of non-volatile storage devices.

[00343] The storage system notifies (1108) a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device. In some embodiments, a notification module (e.g., notification module 228, Figures 2A-1 and 2A-2, notification module 258-1, Figure 2B-1, or notification module 288-1, Figure 2C-1) is used to notify a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition associated with an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1. In some embodiments, or in some circumstances, the notification of the trigger condition causes performance of the amelioration process to be enabled. Furthermore, in some embodiments, or in some circumstances, the amelioration process reduces declared capacity of two or more of the storage devices in the storage system.

[00344] In some embodiments, the host includes (1112) a client on behalf of which data is stored in the storage system (e.g., data storage system 100, Figure 1A; data storage system 140, Figure 1B; data storage system 170, Figure 1C). In some embodiments, the client is or includes an entity on behalf of which data is stored in the storage system. For example, in some embodiments, the host is (1) computer system 110 (Figure 1A) or a client process, module or application executed by computer system 110, (2) computer system 142 (Figure 1B) or a client process, module or application executed by computer system 142, and/or (3) computer system 172 (Figure 1C) or a client process, module or application executed by computer system 172.

[00345] In some embodiments, the host includes (1114) a storage system controller of the storage system (e.g., data storage system 140, Figure 1B). In some embodiments, the storage system controller controls and/or coordinates operations among one or more storage devices. For example, in some embodiments, the host is storage system controller 150 (Figure 1B). In some of these embodiments, the data storage system (e.g., data storage system 140, Figure 1B) is called a scale-up system.

[00346] In some embodiments, the host includes (1116) a cluster controller of the storage system (e.g., data storage system 170, Figure 1C). In some embodiments, the cluster controller controls and/or coordinates operations among one or more data storage subsystems,

as shown for example in Figure 1C, where each of the data storage subsystems may be implemented as a data storage system having one or more storage devices (e.g., data storage system 140, Figure 1B). For example, in some embodiments, the host is cluster controller 180 (Figure 1C). In some of these embodiments, the data storage system (e.g., data storage system 170, Figure 1C) is called a scale-out system or a clustered storage system.

[00347] In some embodiments, notifying (1108) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying (1118) the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) with an unsolicited communication. For example, in some embodiments, the unsolicited communication includes an interrupt communication. As another example, in some embodiments, the unsolicited communication includes a remote direct memory access (RDMA). As yet another example, in some embodiments, the unsolicited communication includes a TCP connection request. In some embodiments, the unsolicited communication includes any other form of unsolicited communication.

[00348] In some embodiments, notifying (1108) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes (1120): (1) receiving a query from the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C), and (2) in response to receiving the query, reporting the trigger condition. For example, in some embodiments, the host polls for the trigger condition and the storage system receives the query from the host and in response to receiving the query, reports the trigger condition.

[00349] In some embodiments, notifying (1108) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes (1122): (1) receiving a command from the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C), and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the respective storage device and/or a write request to the respective storage device. In some embodiments, the command includes a request for temperature of the respective storage device. In some embodiments,

the command includes a request for some other status of the respective storage device. In some embodiments, the notification of the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the respective storage device, and the storage system (1) receives the read request from the host, and (2) in response to receiving the read request, the storage system sends data corresponding to the read request and a notification of the trigger condition.

[00350] In some embodiments, notifying (1108) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes (1124): (1) receiving a command from the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C), and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition. In some embodiments, the command includes an I/O (input/output) request. In some embodiments, the I/O request includes a read request from the respective storage device and/or a write request to the respective storage device. In some embodiments, the command includes a request for temperature of the respective storage device. In some embodiments, the command includes a request for some other status of the respective storage device. In some embodiments, the notification that prompts the host to obtain information with respect to the trigger condition is piggy-backed on a response to the command from the host. For example, in some embodiments, the host issues a read request for data from the respective storage device, and the storage system (1) receives the read request from the host, and (2) in response to receiving the read request, the storage system sends data corresponding to the read request and a notification (e.g., by setting a notification bit) that prompts the host to obtain information with respect to the trigger condition. In some embodiments, the mechanism used for returning a notification when responding to a command from the host is a SCSI deferred error or deferred error response code.

[00351] In some embodiments, notifying (1108) the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying (1126) the host (e.g., computer system 110, Figure 1A, computer system 142, Figure 1B, computer system 172, Figure 1C, storage system controller 150, Figure 1B, and/or cluster controller 180, Figure 1C) that the respective storage device is in read-only

mode. In some embodiments, notifying the host that the respective storage device is in read-only mode includes notifying the host that the respective storage device is not accepting any more write commands. In some embodiments, notifying the host that the respective storage device is in read-only mode includes sending one or more rejections of write commands.

[00352] In some embodiments, the obtaining, the notifying, or both the obtaining and the notifying are performed (1128) by one or more subsystems of the storage system distinct from the plurality of storage devices. For example, in some of these embodiments, the obtaining, the notifying, or both the obtaining and the notifying are performed by a storage system controller (e.g., storage system controller 150, Figure 1B) of the storage system (e.g., data storage system 140, Figure 1B). As another example, in some of these embodiments, the obtaining, the notifying, or both the obtaining and the notifying are performed by a cluster controller (e.g., cluster controller 180, Figure 1C) of the storage system (e.g., data storage system 170, Figure 1C). In some embodiments, method 1100, or at least the obtaining operation 1102 and/or the notifying operation 1108 of method 1100, is governed by instructions that are stored in a non-transitory computer readable storage medium and that are executed by one or more processors of a device, such as the one or more processing units (CPUs) 152-1 of system management module 151-1, shown in Figures 1B and 2B-1 or the one or more processing units (CPUs) 182-1 of cluster management module 181-1, shown in Figures 1C and 2C-1.

[00353] In some embodiments, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device includes (1130) a process to reduce utilization of the non-volatile memory of the respective storage device, for example as described above with respect to operation 608 of Figure 6.

[00354] In some embodiments, after notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device (1110), the storage system (1) re-evaluates the trigger condition in accordance with the one or more metrics of the two or more storage devices of the plurality of storage devices in the storage system, and (2) in accordance with a determination that the trigger condition is no longer valid, notifies the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device. For example, in some circumstances, the one or more metrics of the respective storage device may change such that the trigger condition is no longer valid (e.g., the amelioration process is no longer needed). For example, during the operation of an amelioration process (e.g., operation 606, Figure 6),

normal storage operations will continue to be performed (e.g., read, write, delete, trim, etc.). Normal storage operations include operations like trim that explicitly reduce the respective storage device utilization, possibly enough to merit aborting the amelioration process. Other storage activity such as garbage collection may also reduce utilization, possibly enough to merit aborting the amelioration process.

[00355] In some embodiments, the trigger condition (e.g., periodically, semi-continuously, initially, finally, occasionally or irregularly) is recomputed or re-evaluated in accordance with the one or more metrics of the respective storage device, as the one or more metrics may change in value due to the amelioration process and/or normal storage operations (e.g., read, write, erase and trim or unmap operations). In some embodiments, a trigger detection module (e.g., trigger detection module 224, Figures 2A-1 and 2A-2; trigger detection module 254-1, Figure 2B-1; or trigger detection module 284-1, Figure 2C-1) and/or a notification module (e.g., notification module 228, Figures 2A-1 and 2A-2, notification module 258-1, Figure 2B-1, or notification module 288-1, Figure 2C-1) are used to, after notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device: (1) re-evaluate the trigger condition in accordance with the one or more metrics of the respective storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, notify the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, as described above with respect to Figures 2A-1, 2A-2, 2B-1, and 2C-1.

[00356] In some embodiments, any operations of method 1100 described above are performed by a storage system, the storage system including (1) non-volatile memory (e.g., comprising one or more non-volatile storage devices, such as flash memory devices), (2) one or more processors, and (3) controller memory (e.g., non-volatile memory or volatile memory in or coupled to a controller of the storage system) storing one or more programs, which when executed by the one or more processors cause the storage system to perform or control performance of any of the methods described herein.

[00357] In some embodiments, any operations of method 1100 described above are performed by a storage system including means for performing any of the methods described herein.

[00358] In some embodiments, any operations of method 1100 described above are performed by a storage system including (1) a plurality of storage devices, (2) one or more subsystems having one or more processors, and (3) memory (e.g., non-volatile memory or

volatile memory in the storage system) storing one or more programs, which when executed by the one or more processors cause the one or more subsystems to perform or control performance of any of the methods described herein.

[00359] Semiconductor memory devices include volatile memory devices, such as dynamic random access memory ("DRAM") or static random access memory ("SRAM") devices, non-volatile memory devices, such as resistive random access memory ("ReRAM"), electrically erasable programmable read only memory ("EEPROM"), flash memory (which can also be considered a subset of EEPROM), ferroelectric random access memory ("FRAM"), and magnetoresistive random access memory ("MRAM"), and other semiconductor elements capable of storing information. Each type of memory device may have different configurations. For example, flash memory devices may be configured in a NAND or a NOR configuration.

[00360] The semiconductor memory elements located within and/or over a substrate may be arranged in two or three dimensions, such as a two dimensional memory structure or a three dimensional memory structure.

[00361] The term "three-dimensional memory device" (or 3D memory device) is herein defined to mean a memory device having multiple memory layers or multiple levels (e.g., sometimes called multiple memory device levels) of memory elements, including any of the following: a memory device having a monolithic or non-monolithic 3D memory array; or two or more 2D and/or 3D memory devices, packaged together to form a stacked-chip memory device.

[00362] One of skill in the art will recognize that this invention is not limited to the two dimensional and three dimensional structures described but cover all relevant memory structures within the spirit and scope of the invention as described herein and as understood by one of skill in the art.

[00363] It will be understood that, although the terms "first," "second," etc. may be used herein to describe various elements, these elements should not be limited by these terms. These terms are only used to distinguish one element from another. For example, a first storage device could be termed a second storage device, and, similarly, a second storage device could be termed a first storage device, without changing the meaning of the description, so long as all occurrences of the "first storage device" are renamed consistently and all occurrences of the "second storage device" are renamed consistently. The first

storage device and the second storage device are both storage devices, but they are not the same storage device.

[00364] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the claims. As used in the description of the embodiments and the appended claims, the singular forms "a," "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will also be understood that the term "and/or" as used herein refers to and encompasses any and all possible combinations of one or more of the associated listed items. It will be further understood that the terms "comprises" and/or "comprising," when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[00365] As used herein, the term "if" may be construed to mean "when" or "upon" or "in response to determining" or "in accordance with a determination" or "in response to detecting," that a stated condition precedent is true, depending on the context. Similarly, the phrase "if it is determined [that a stated condition precedent is true]" or "if [a stated condition precedent is true]" or "when [a stated condition precedent is true]" may be construed to mean "upon determining" or "in response to determining" or "in accordance with a determination" or "upon detecting" or "in response to detecting" that the stated condition precedent is true, depending on the context.

[00366] The foregoing description, for purpose of explanation, has been described with reference to specific embodiments. However, the illustrative discussions above are not intended to be exhaustive or to limit the claims to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The embodiments were chosen and described in order to best explain principles of operation and practical applications, to thereby enable others skilled in the art.

What is claimed is:

1. A method of managing a storage system, the method comprising:

obtaining, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory;

detecting a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of two or more of the storage devices of the plurality of storage devices in the storage system; and

notifying a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device.

2. The method of claim 1, further comprising:

prior to detecting the trigger condition, detecting a first wear condition of the non-volatile memory of the respective storage device, wherein a total storage capacity of the non-volatile memory of the respective storage device includes declared capacity and over-provisioning; and

in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the respective storage device without reducing declared capacity of the non-volatile memory of the respective storage device.

3. The method of claim 2, wherein detecting the trigger condition comprises detecting a second wear condition distinct from the first wear condition.

4. The method of any of claims 1-2, wherein the host includes a client on behalf of which data is stored in the storage system.

5. The method of any of claims 1-2, wherein the host includes a storage system controller of the storage system.

6. The method of any of claims 1-2, wherein the host includes a cluster controller of the storage system.

7. The method of any of claims 1-6, wherein notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying the host with an unsolicited communication.

8. The method of any of claims 1-6, wherein notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes:

receiving a query from the host; and

in response to receiving the query, reporting the trigger condition.

9. The method of any of claims 1-6, wherein notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes:

receiving a command from the host; and

in response to receiving the command, sending a response to the command and a notification of the trigger condition.

10. The method of any of claims 1-6, wherein notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes:

receiving a command from the host; and

in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition.

11. The method of any of claims 1-6, wherein notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying the host that the respective storage device is in read-only mode.

12. The method of any of claims 1-9, further comprising:

after notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device:

re-evaluating the trigger condition in accordance with the one or more metrics of the two or more storage devices of the plurality of storage devices in the storage system; and

in accordance with a determination that the trigger condition is no longer valid, notifying the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device.

13. The method of any of claims 1-12, wherein the obtaining, the notifying, or both the obtaining and the notifying are performed by one or more subsystems of the storage system distinct from the plurality of storage devices.

14. The method of any of claims 1-13, wherein the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device includes a process to reduce utilization of the non-volatile memory of the respective storage device.

15. The method of any of claims 1-14, wherein the respective storage device comprises one or more flash memory devices.

16. A storage system, comprising:

non-volatile memory;

one or more processors; and

controller memory storing one or more programs, which when executed by the one or more processors cause the storage system to perform operations comprising:

obtaining, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory;

detecting a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of the respective storage device; and

notifying a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition associated with an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device.

17. The storage system of claim 16, wherein the one or more programs further include instructions that, when executed by the one or more processors, cause the storage system to perform operations comprising:

prior to detecting the trigger condition, detecting a first wear condition of the non-volatile memory of the respective storage device, wherein a total storage capacity of the non-volatile memory of the respective storage device includes declared capacity and over-provisioning; and

in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the respective storage device without reducing declared capacity of the non-volatile memory of the respective storage device.

18. The storage system of any of claims 16-17, wherein the one or more programs further include instructions that, when executed by the one or more processors, cause the storage system to perform the method of any of claims 3-15.

19. A non-transitory computer readable storage medium, storing one or more programs configured for execution by one or more processors of a storage system, the one or more programs including instructions for:

obtaining, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory;

detecting a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of two or more of the storage devices of the plurality of storage devices in the storage system; and

notifying a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device.

20. The non-transitory computer readable storage medium of claim 19, wherein the one or more programs further include instructions for:

prior to detecting the trigger condition, detecting a first wear condition of the non-volatile memory of the respective storage device, wherein a total storage capacity of the non-volatile memory of the respective storage device includes declared capacity and over-provisioning; and

in response to detecting the first wear condition, performing a remedial action that reduces over-provisioning of the non-volatile memory of the respective storage device without reducing declared capacity of the non-volatile memory of the respective storage device.

21. The non-transitory computer readable storage medium of any of claims 19-20, wherein the one or more programs further include instructions for performing the method of any of claims 3-15.

22. A storage system, comprising:

means for obtaining, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory;

means for detecting a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of the respective storage device; and

means for notifying a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition associated with an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device.

23. The storage system of claim 22, further comprising:

means for detecting, prior to detecting the trigger condition, a first wear condition of the non-volatile memory of the respective storage device, wherein a total storage capacity of the non-volatile memory of the respective storage device includes declared capacity and over-provisioning; and

means for performing, in response to detecting the first wear condition, a remedial action that reduces over-provisioning of the non-volatile memory of the respective storage device without reducing declared capacity of the non-volatile memory of the respective storage device.

24. The storage system of any of claims 22-23, further comprising means for performing the method of any of claims 3-15.

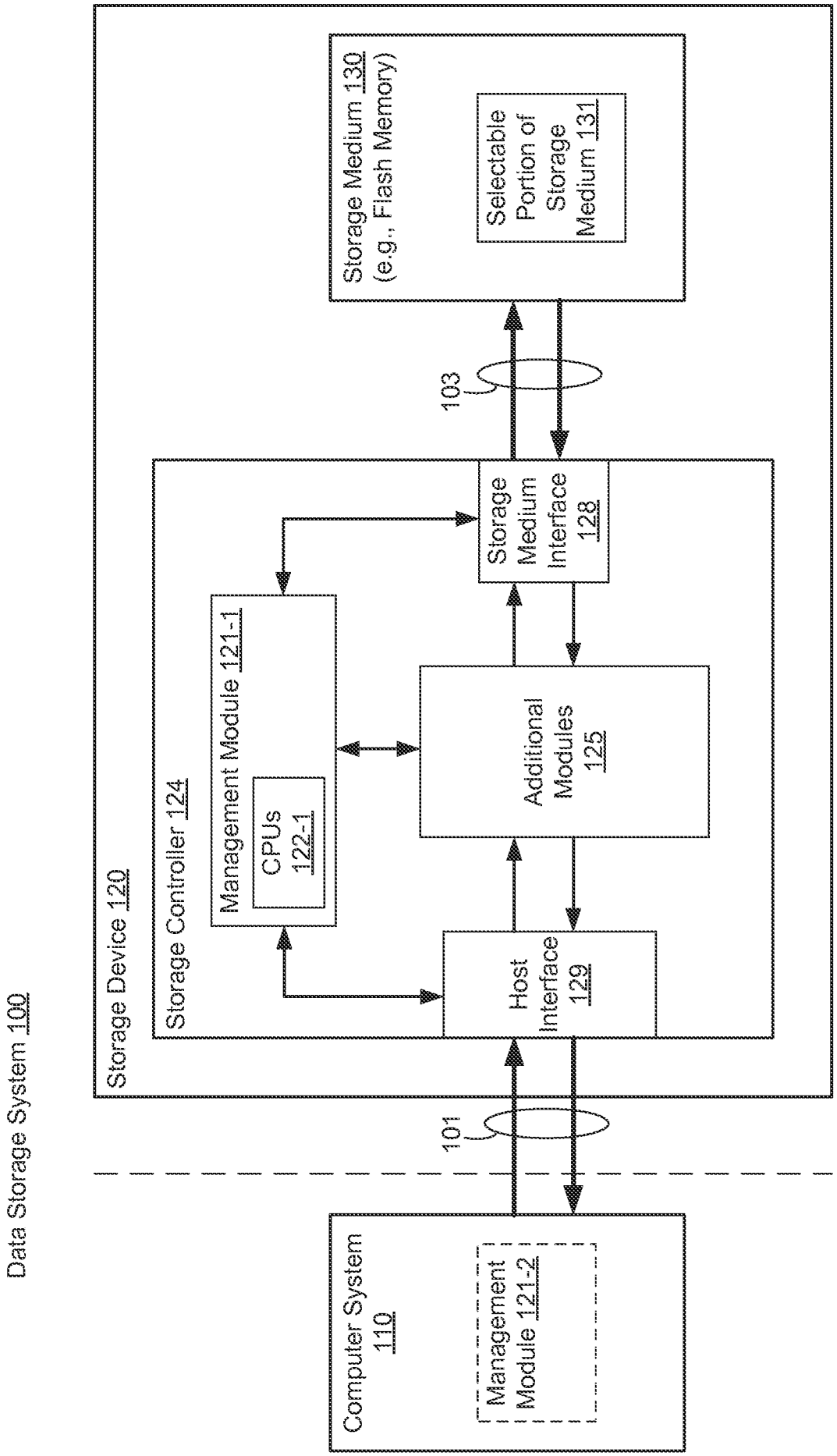

Data Storage System 100
Storage Device 120
Storage Controller 124
Management Module 121-1
CPUs 122-1
Storage Medium Interface 128
Additional Modules 125
Host Interface 129
Storage Medium 130 (e.g., Flash Memory)
Selectable Portion of Storage Medium 131
103
101
Computer System 110
Management Module 121-2

Figure 1A

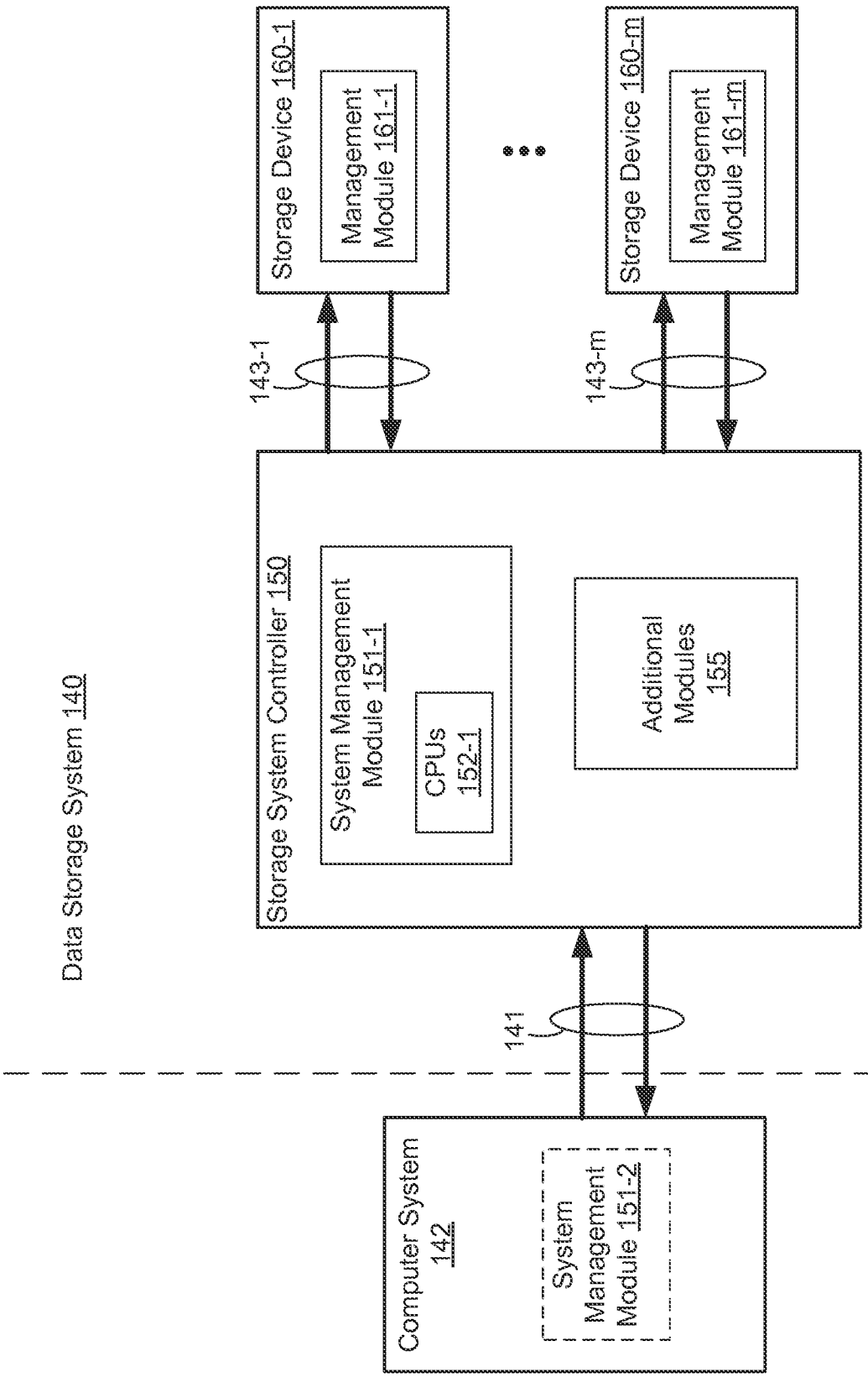

Data Storage System 140
Storage System Controller 150
System Management Module 151-1
CPUs 152-1
Additional Modules 155
Storage Device 160-1
Management Module 161-1
Storage Device 160-m
Management Module 161-m
143-1
143-m
141
Computer System 142
System Management Module 151-2

Figure 1B

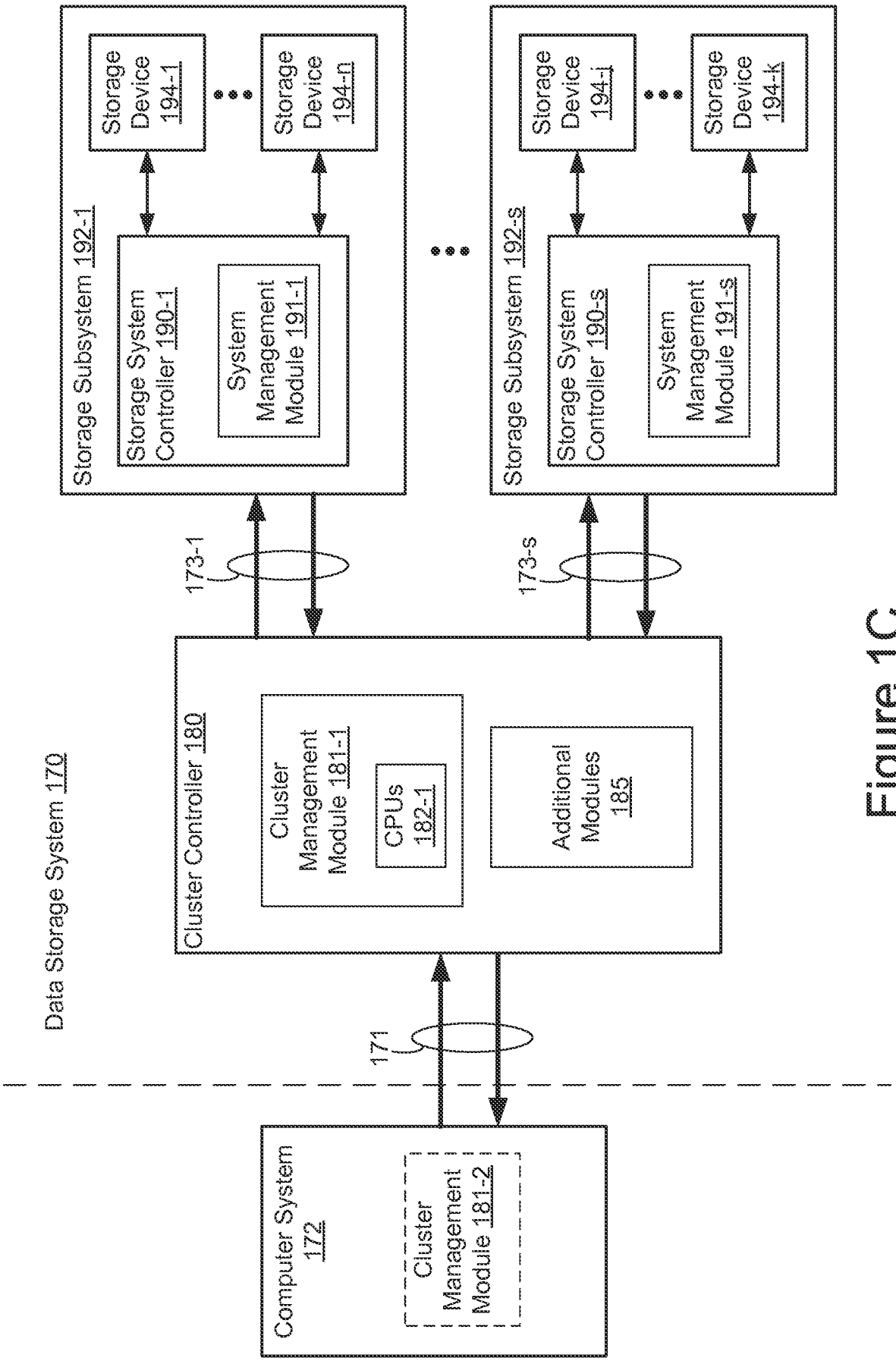

Data Storage System 170
Computer System 172
Cluster Management Module 181-2
171
Cluster Controller 180
Cluster Management Module 181-1
CPUs 182-1
Additional Modules 185
173-1
173-s
Storage Subsystem 192-1
Storage System Controller 190-1
System Management Module 191-1
Storage Device 194-1
Storage Device 194-n
Storage Subsystem 192-s
Storage System Controller 190-s
System Management Module 191-s
Storage Device 194-i
Storage Device 194-k

Figure 1C

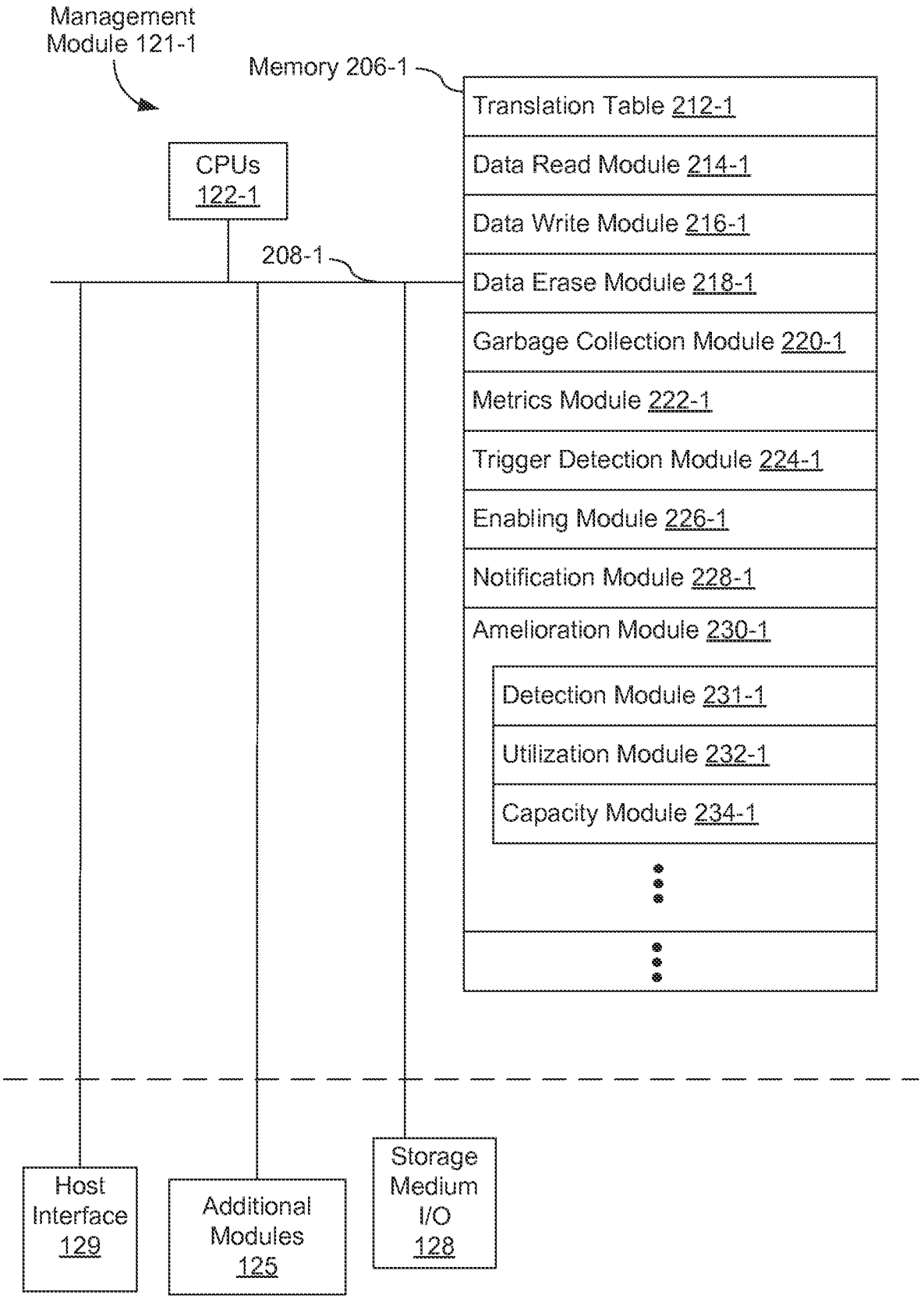

Management Module 121-1
Memory 206-1
Translation Table 212-1
CPUs 122-1
Data Read Module 214-1
Data Write Module 216-1
208-1
Data Erase Module 218-1
Garbage Collection Module 220-1
Metrics Module 222-1
Trigger Detection Module 224-1
Enabling Module 226-1
Notification Module 228-1
Amelioration Module 230-1
Detection Module 231-1
Utilization Module 232-1
Capacity Module 234-1
Host Interface 129
Additional Modules 125
Storage Medium I/O 128

Figure 2A-1

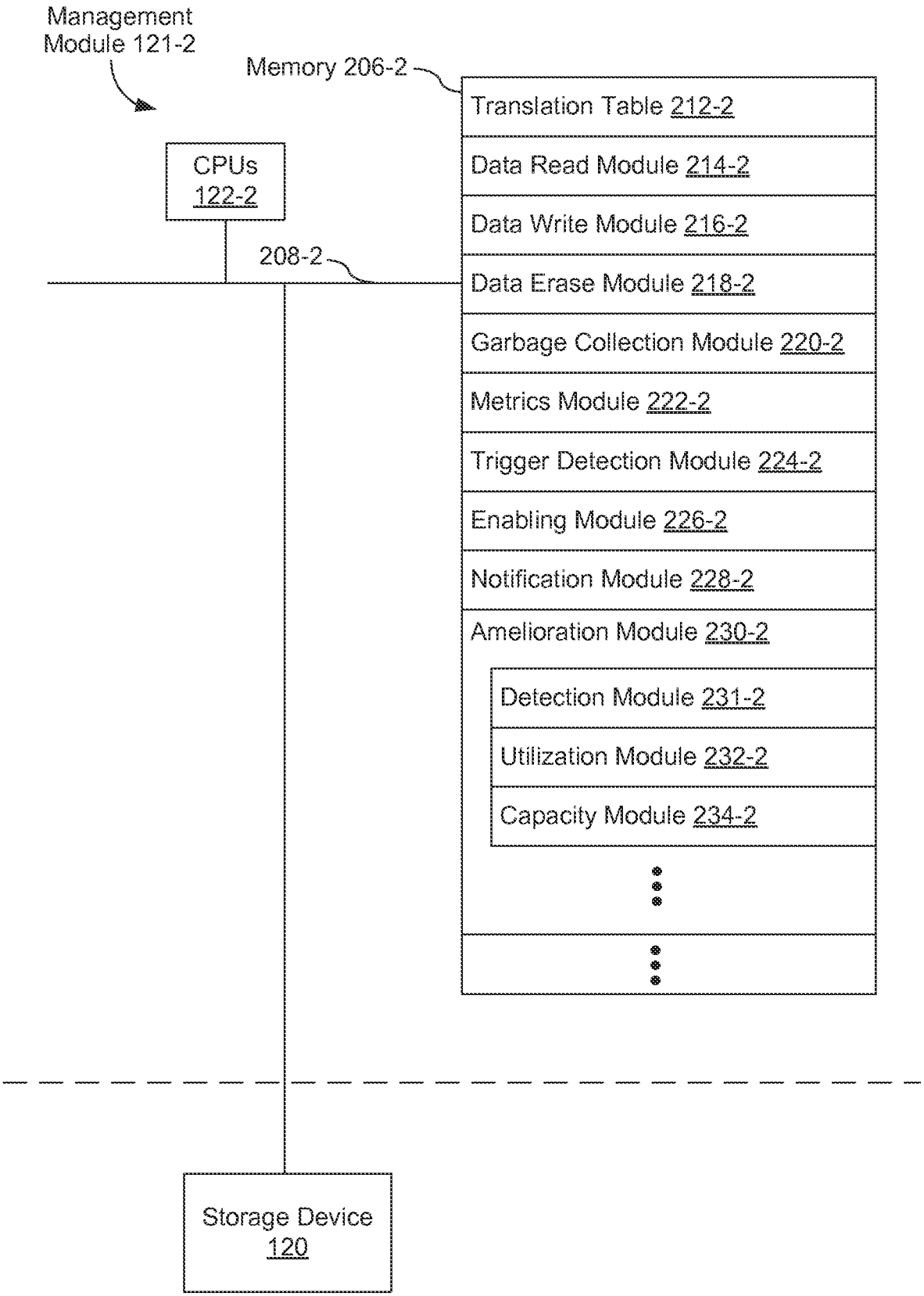

Management Module 121-2
Memory 206-2
Translation Table 212-2
CPUs 122-2
Data Read Module 214-2
Data Write Module 216-2
208-2
Data Erase Module 218-2
Garbage Collection Module 220-2
Metrics Module 222-2
Trigger Detection Module 224-2
Enabling Module 226-2
Notification Module 228-2
Amelioration Module 230-2
Detection Module 231-2
Utilization Module 232-2
Capacity Module 234-2
Storage Device 120

Figure 2A-2

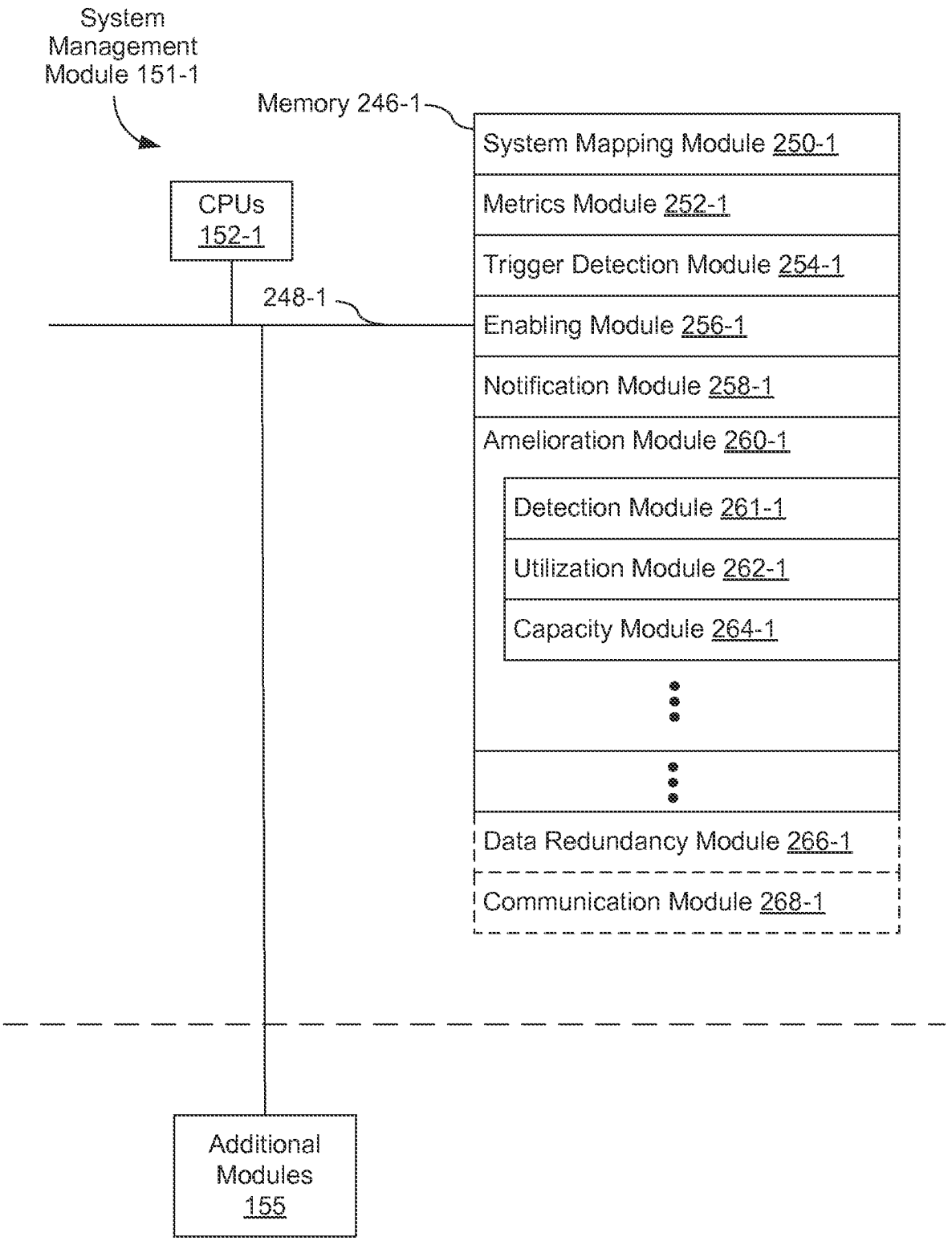

System Management Module 151-1
Memory 246-1
System Mapping Module 250-1
Metrics Module 252-1
Trigger Detection Module 254-1
Enabling Module 256-1
Notification Module 258-1
Amelioration Module 260-1
Detection Module 261-1
Utilization Module 262-1
Capacity Module 264-1
Data Redundancy Module 266-1
Communication Module 268-1
CPUs 152-1
248-1
Additional Modules 155

Figure 2B-1

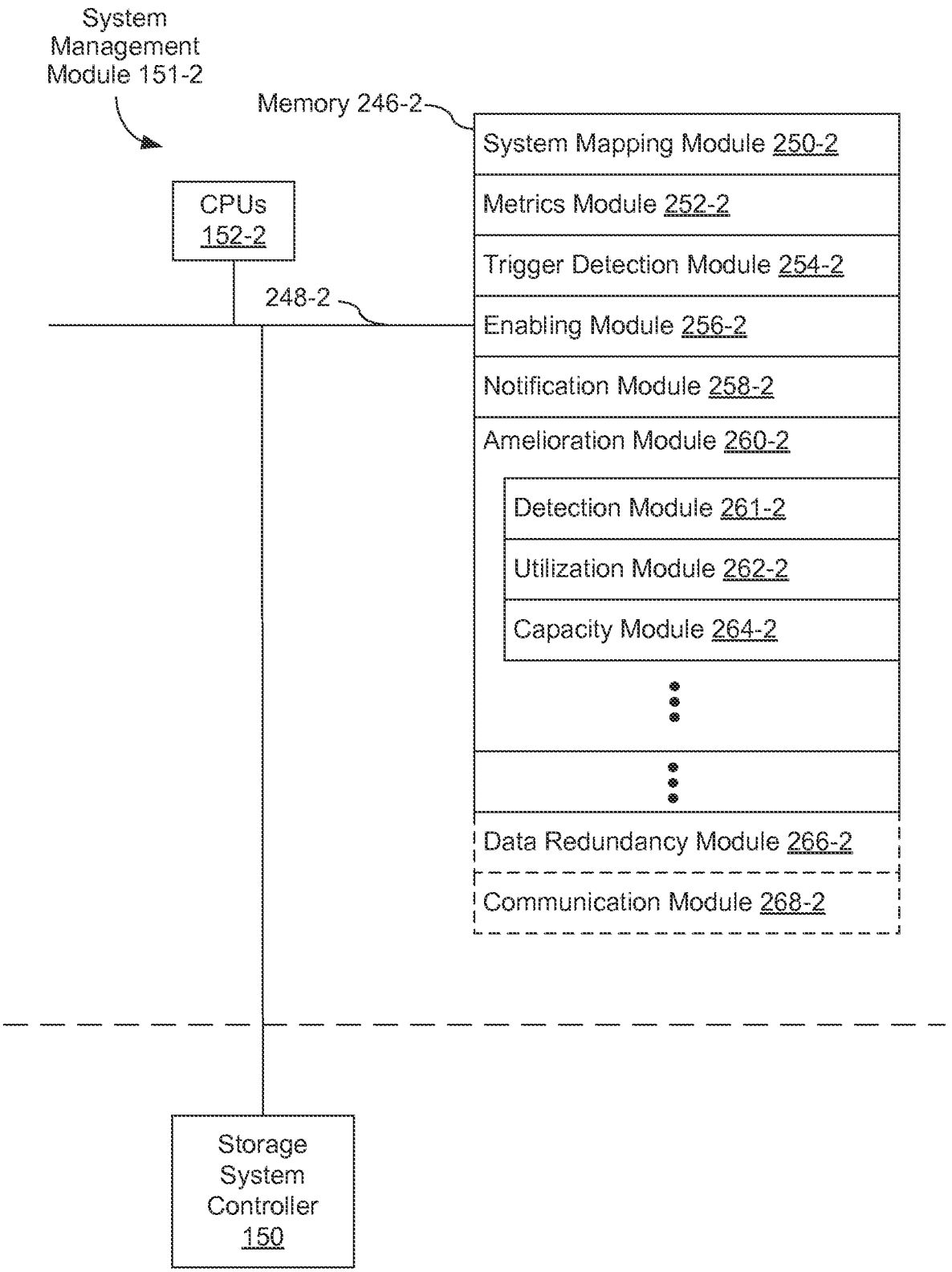

System Management Module 151-2
Memory 246-2
System Mapping Module 250-2
Metrics Module 252-2
Trigger Detection Module 254-2
Enabling Module 256-2
Notification Module 258-2
Amelioration Module 260-2
Detection Module 261-2
Utilization Module 262-2
Capacity Module 264-2
Data Redundancy Module 266-2
Communication Module 268-2
CPUs 152-2
248-2
Storage System Controller 150

Figure 2B-2

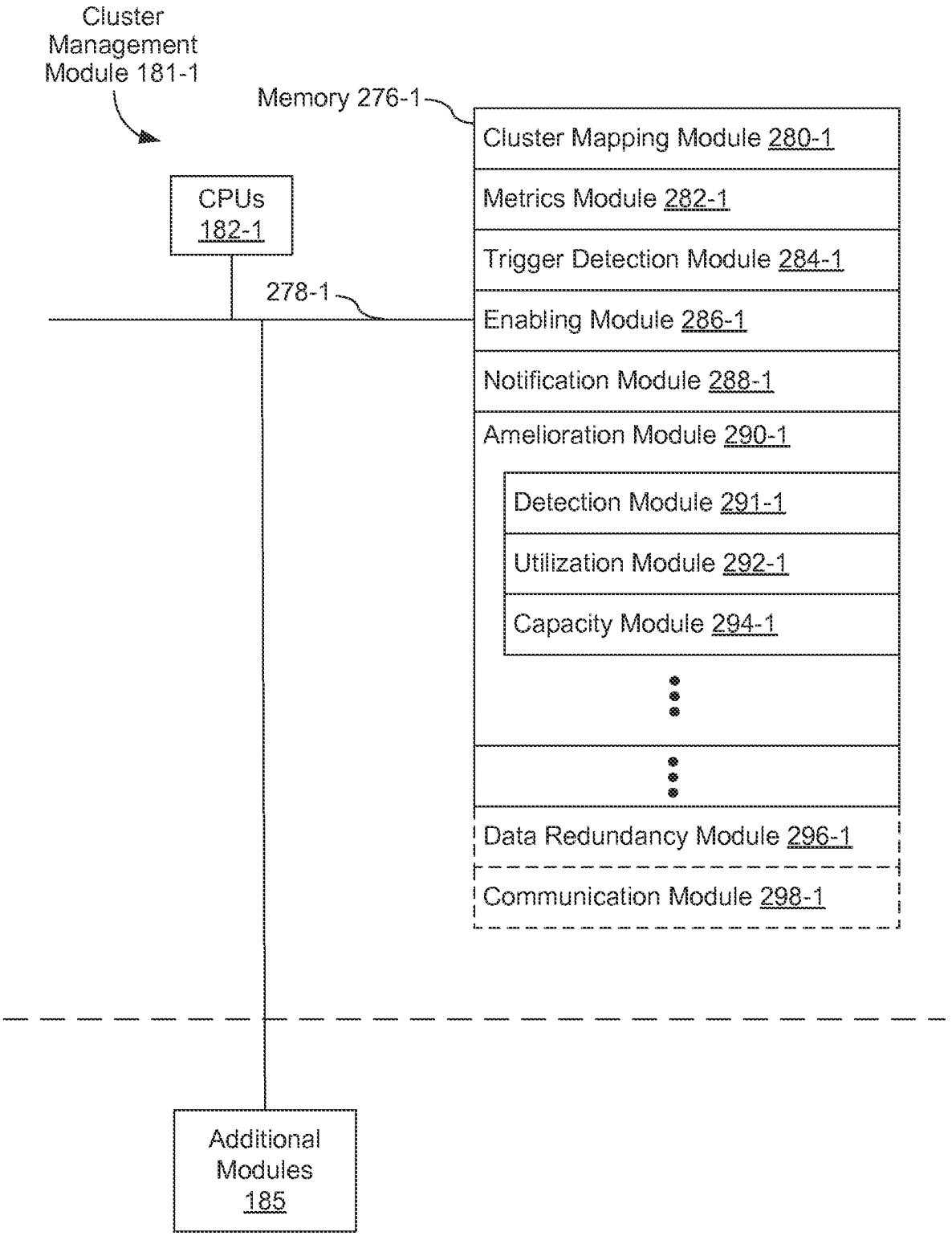

Cluster Management Module 181-1
Memory 276-1
Cluster Mapping Module 280-1
Metrics Module 282-1
Trigger Detection Module 284-1
CPUs 182-1
278-1
Enabling Module 286-1
Notification Module 288-1
Amelioration Module 290-1
Detection Module 291-1
Utilization Module 292-1
Capacity Module 294-1
Data Redundancy Module 296-1
Communication Module 298-1
Additional Modules 185

Figure 2C-1

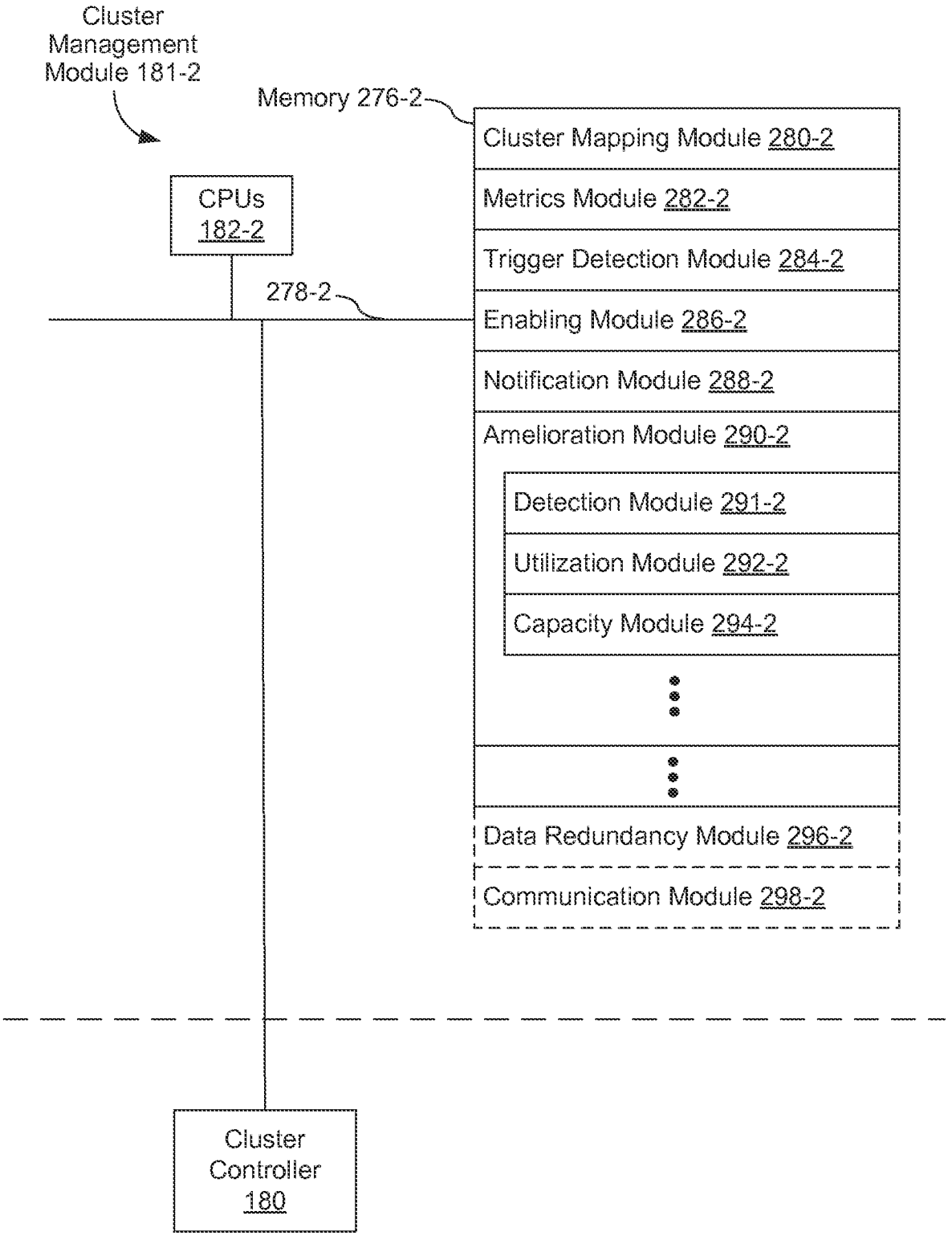

Cluster Management Module 181-2
Memory 276-2
Cluster Mapping Module 280-2
Metrics Module 282-2
Trigger Detection Module 284-2
CPUs 182-2
278-2
Enabling Module 286-2
Notification Module 288-2
Amelioration Module 290-2
Detection Module 291-2
Utilization Module 292-2
Capacity Module 294-2
Data Redundancy Module 296-2
Communication Module 298-2
Cluster Controller 180

Figure 2C-2

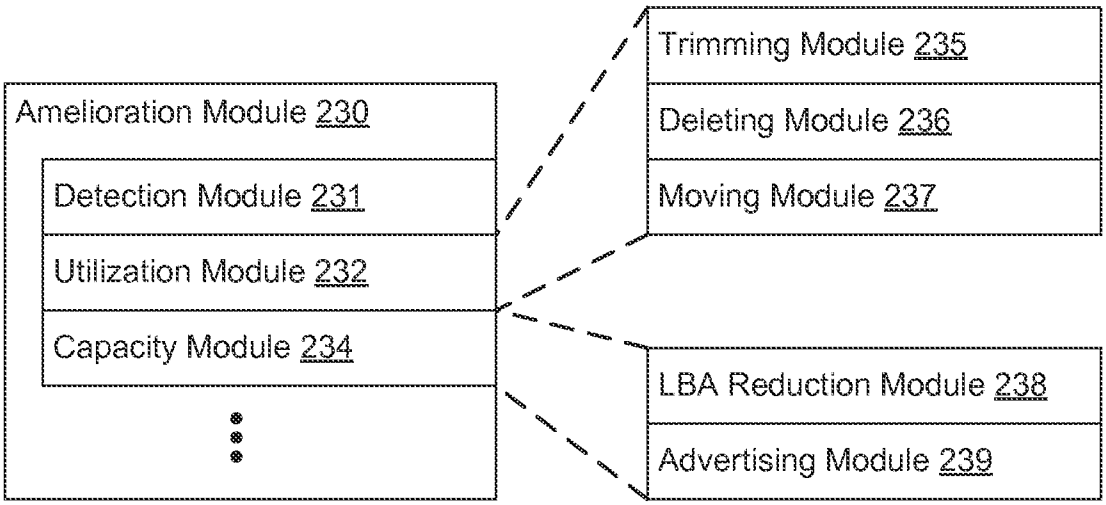

Amelioration Module 230
Detection Module 231
Utilization Module 232
Capacity Module 234
Trimming Module 235
Deleting Module 236
Moving Module 237
LBA Reduction Module 238
Advertising Module 239

Figure 2D

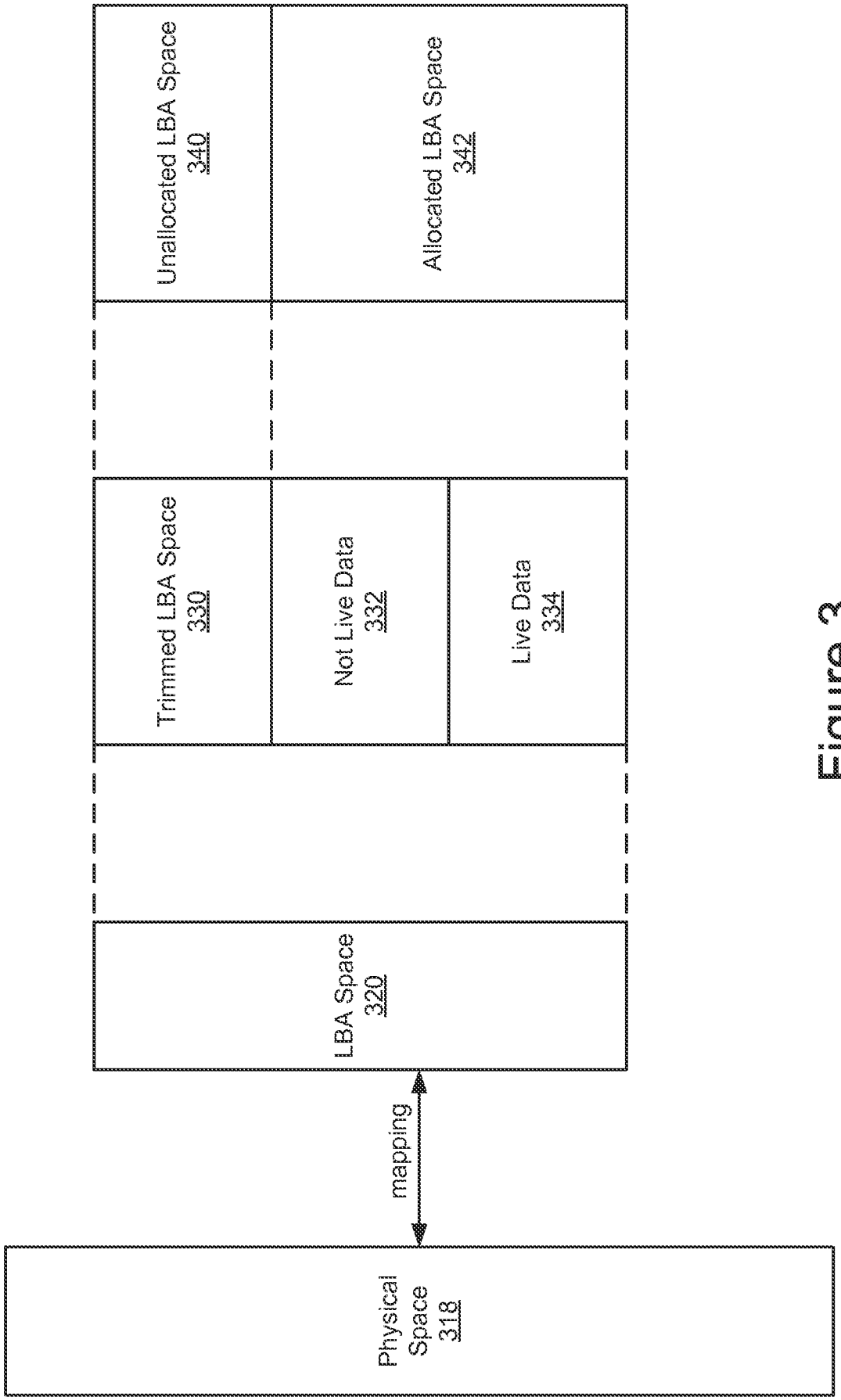

Physical Space 318
mapping
LBA Space 320
Trimmed LBA Space 330
Not Live Data 332
Live Data 334
Unallocated LBA Space 340
Allocated LBA Space 342

Figure 3

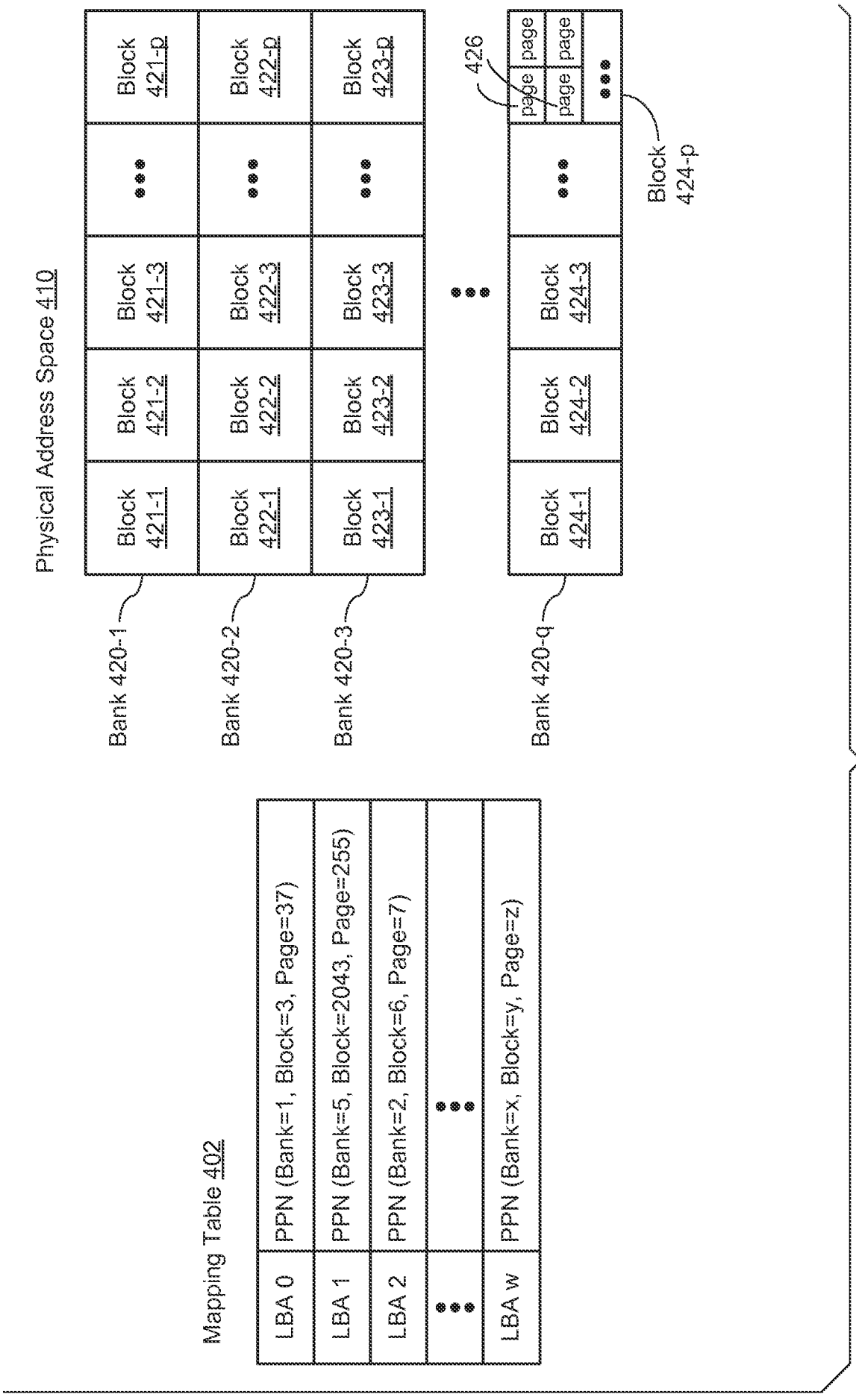

Physical Address Space 410
Bank 420-1
Bank 420-2
Bank 420-3
Bank 420-q
Block 421-1
Block 421-2
Block 421-3
Block 421-p
Block 422-1
Block 422-2
Block 422-3
Block 422-p
Block 423-1
Block 423-2
Block 423-3
Block 423-p
Block 424-1
Block 424-2
Block 424-3
Block 424-p
426
page
Mapping Table 402
LBA 0
PPN (Bank=1, Block=3, Page=37)
LBA 1
PPN (Bank=5, Block=2043, Page=255)
LBA 2
PPN (Bank=2, Block=6, Page=7)
LBA w
PPN (Bank=x, Block=y, Page=z)

Figure 4

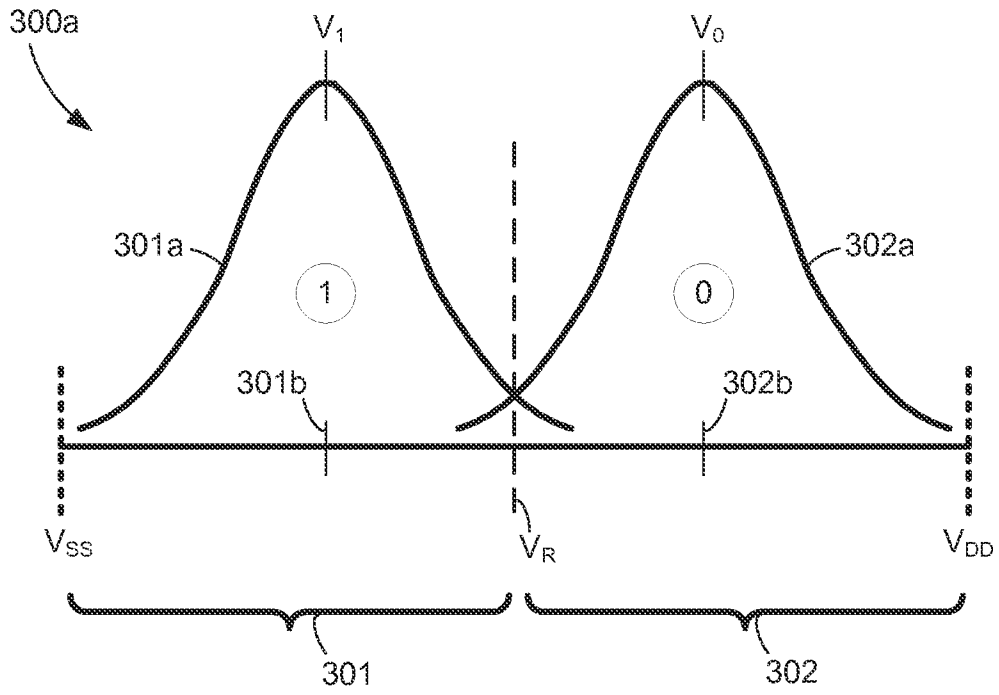


Figure 5A

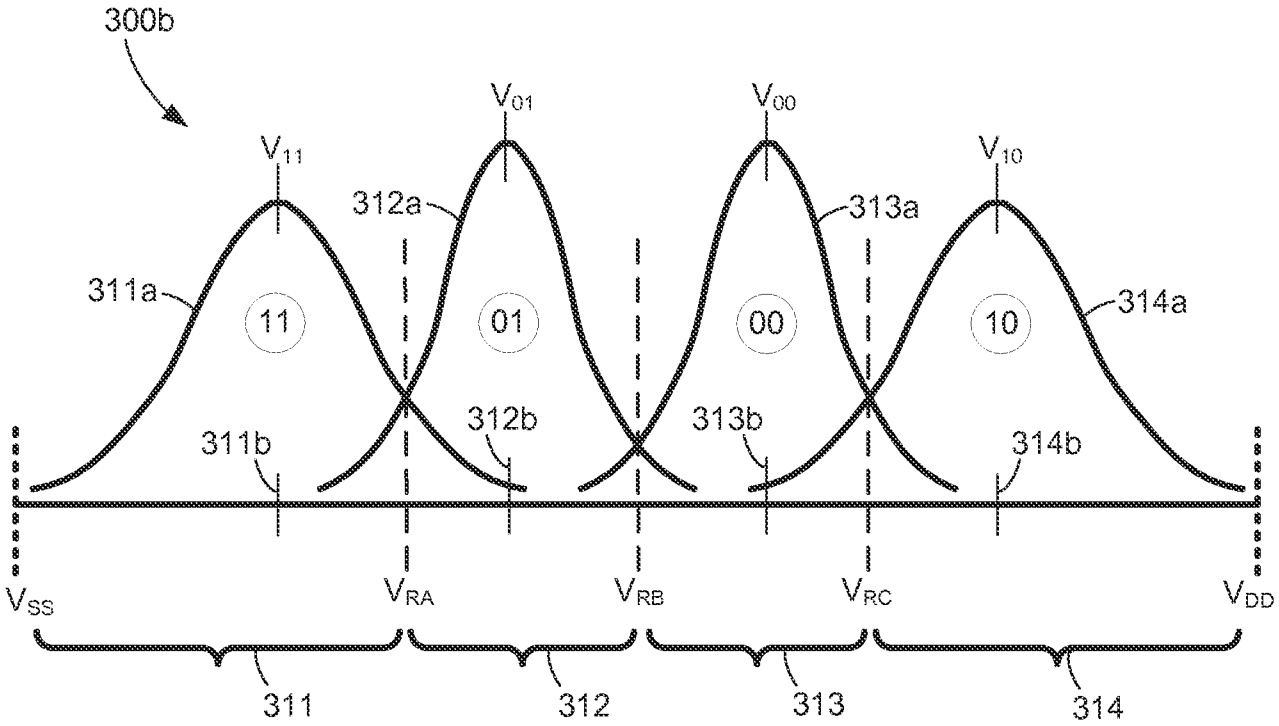


Figure 5B

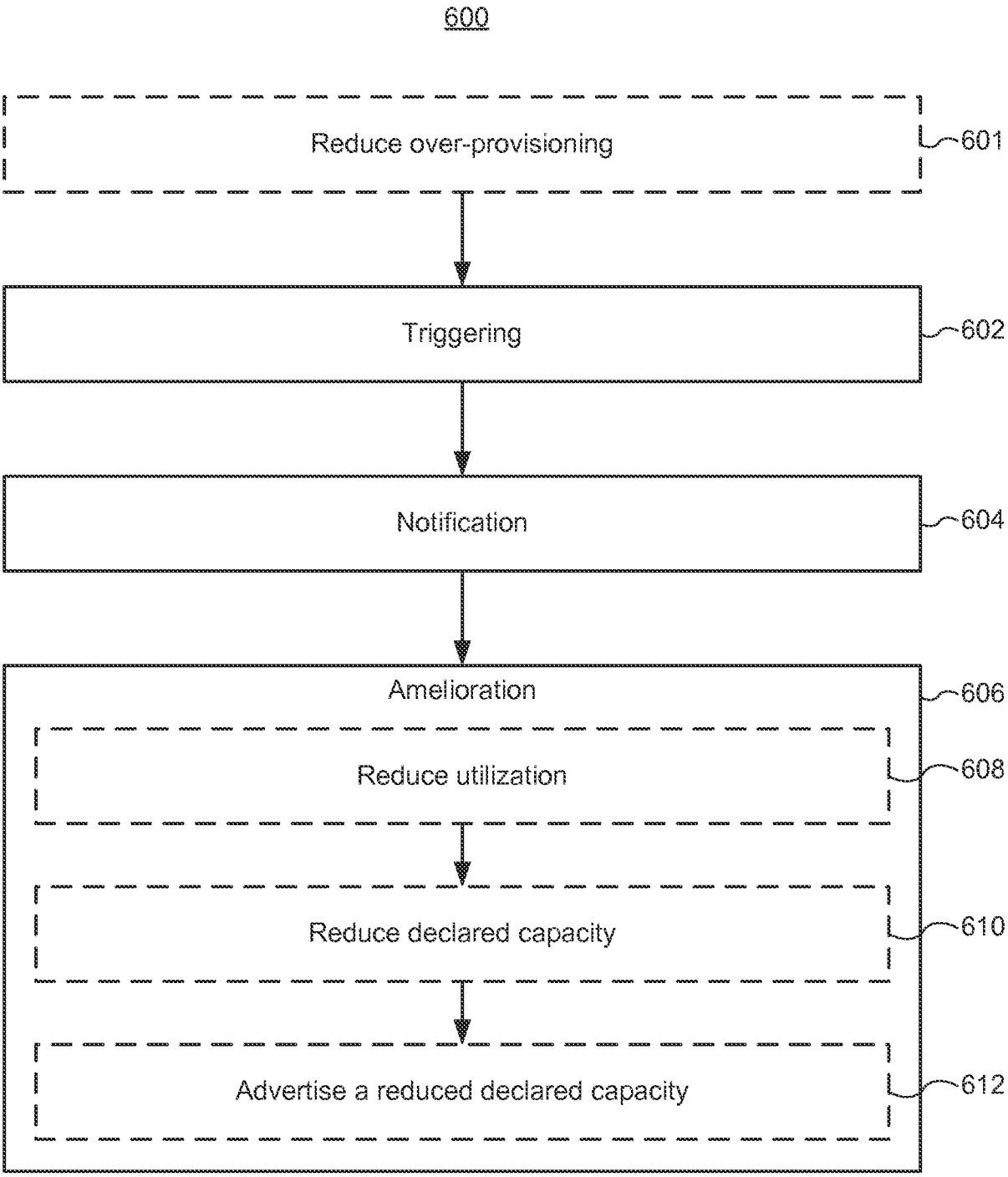

600
Reduce over-provisioning
601
Triggering
602
Notification
604
Amelioration
606
Reduce utilization
608
Reduce declared capacity
610
Advertise a reduced declared capacity
612

Figure 6

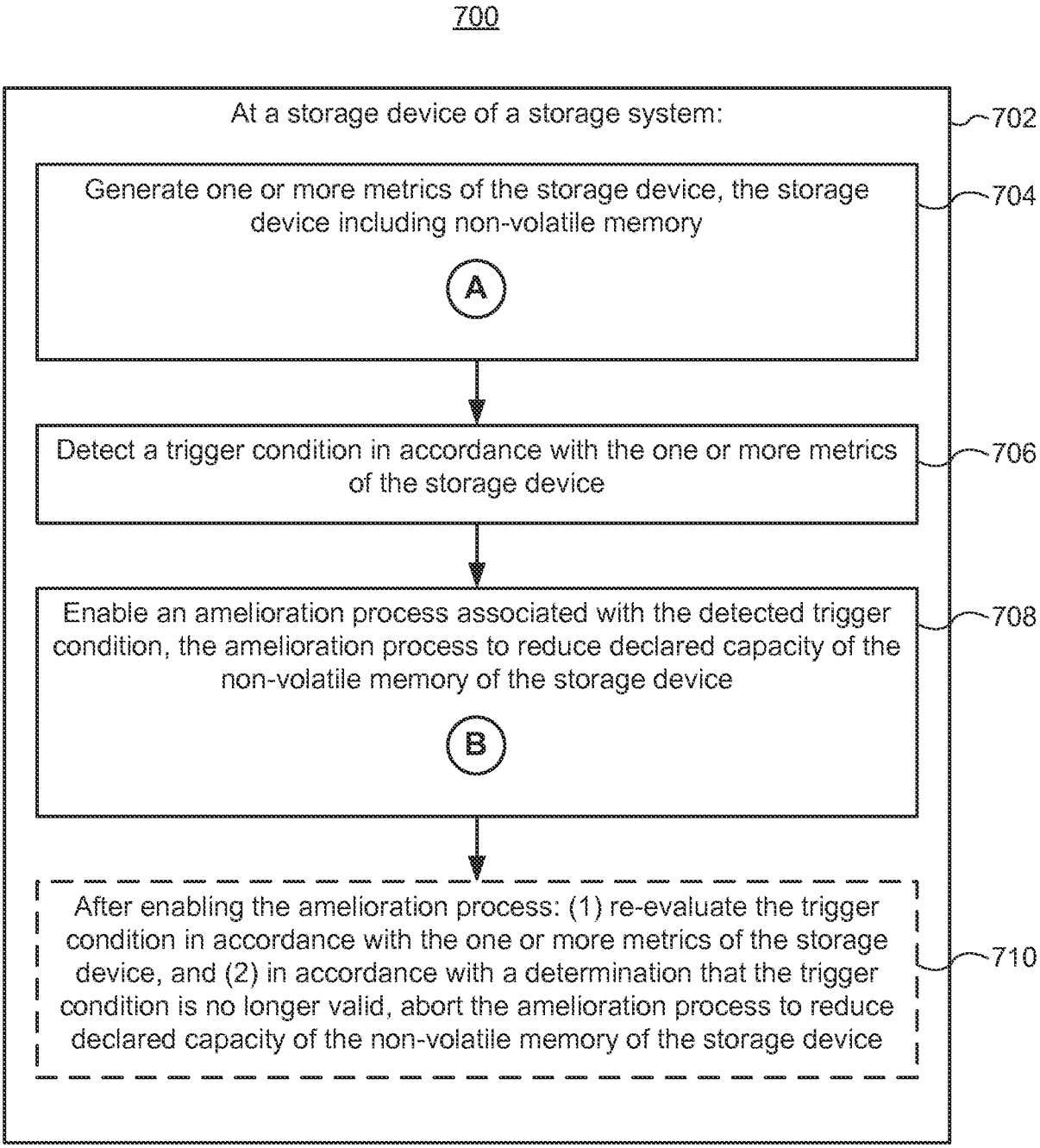

700
At a storage device of a storage system:
702
Generate one or more metrics of the storage device, the storage device including non-volatile memory
704
A
Detect a trigger condition in accordance with the one or more metrics of the storage device
706
Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device
708
B
After enabling the amelioration process: (1) re-evaluate the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, abort the amelioration process to reduce declared capacity of the non-volatile memory of the storage device
710

Figure 7A

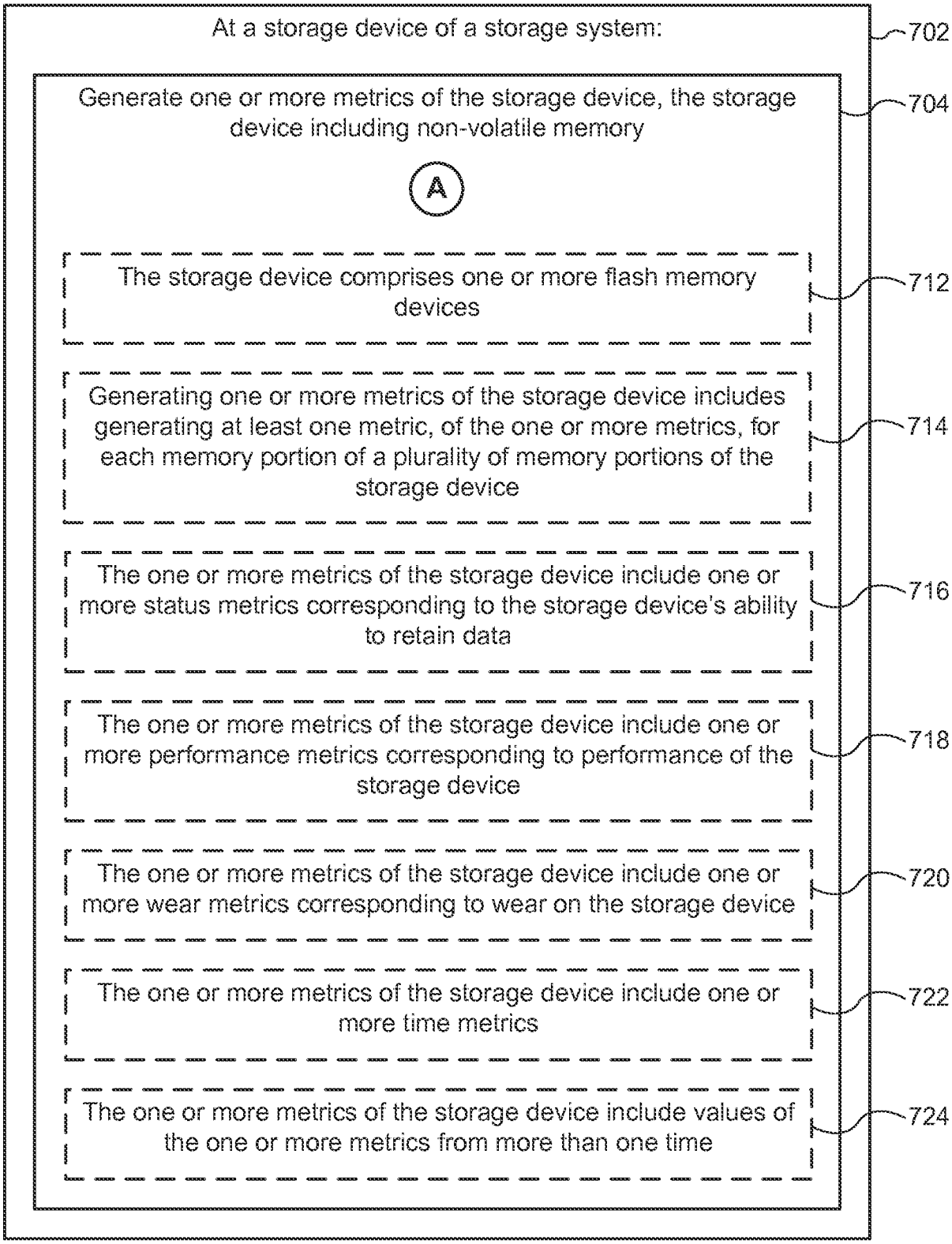

At a storage device of a storage system:
702
Generate one or more metrics of the storage device, the storage device including non-volatile memory
704
A
The storage device comprises one or more flash memory devices
712
Generating one or more metrics of the storage device includes generating at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the storage device
714
The one or more metrics of the storage device include one or more status metrics corresponding to the storage device's ability to retain data
716
The one or more metrics of the storage device include one or more performance metrics corresponding to performance of the storage device
718
The one or more metrics of the storage device include one or more wear metrics corresponding to wear on the storage device
720
The one or more metrics of the storage device include one or more time metrics
722
The one or more metrics of the storage device include values of the one or more metrics from more than one time
724

Figure 7B

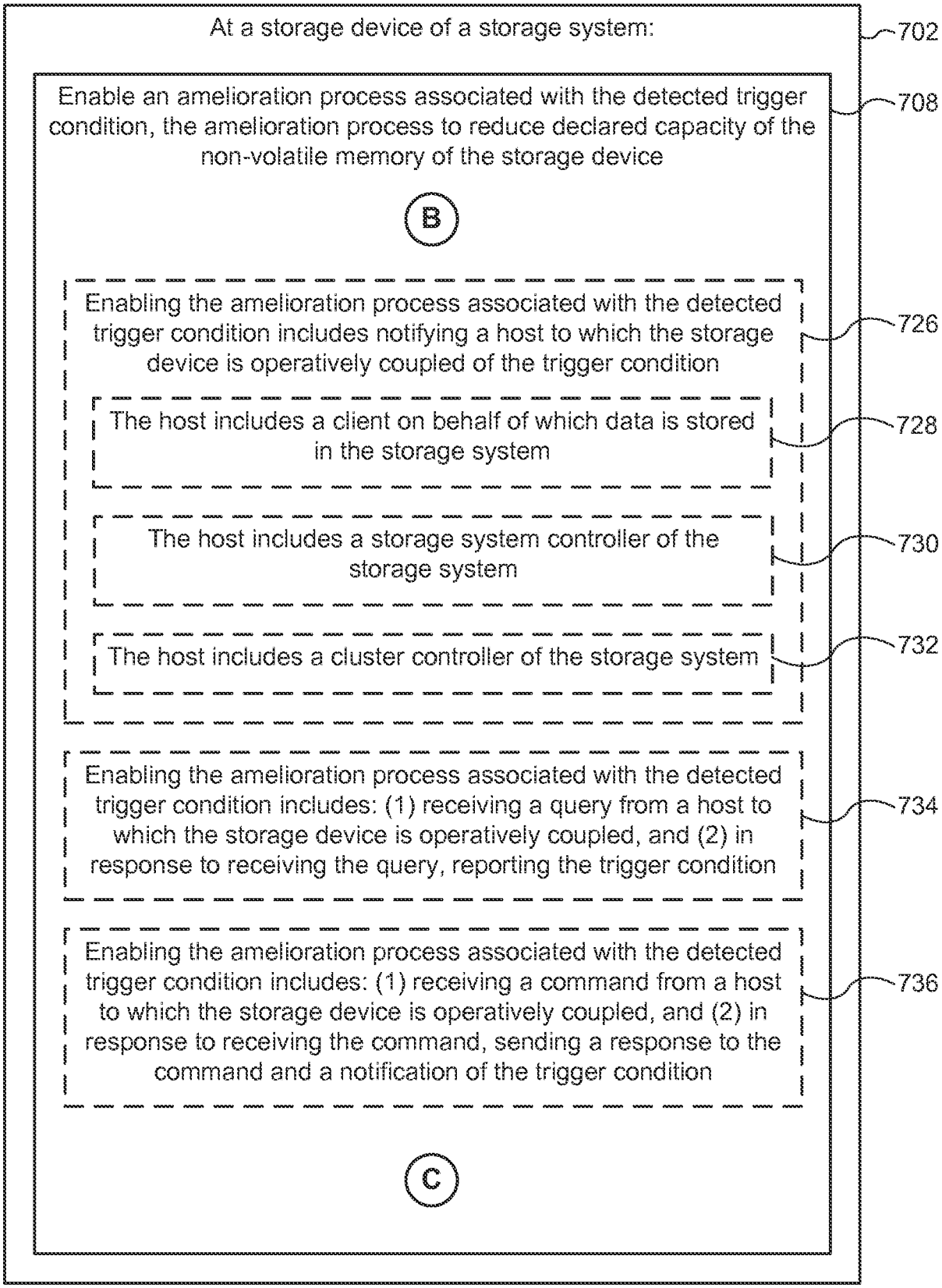

At a storage device of a storage system:
702
Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device
708
B
Enabling the amelioration process associated with the detected trigger condition includes notifying a host to which the storage device is operatively coupled of the trigger condition
726
The host includes a client on behalf of which data is stored in the storage system
728
The host includes a storage system controller of the storage system
730
The host includes a cluster controller of the storage system
732
Enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a query from a host to which the storage device is operatively coupled, and (2) in response to receiving the query, reporting the trigger condition
734
Enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition
736
C

Figure 7C

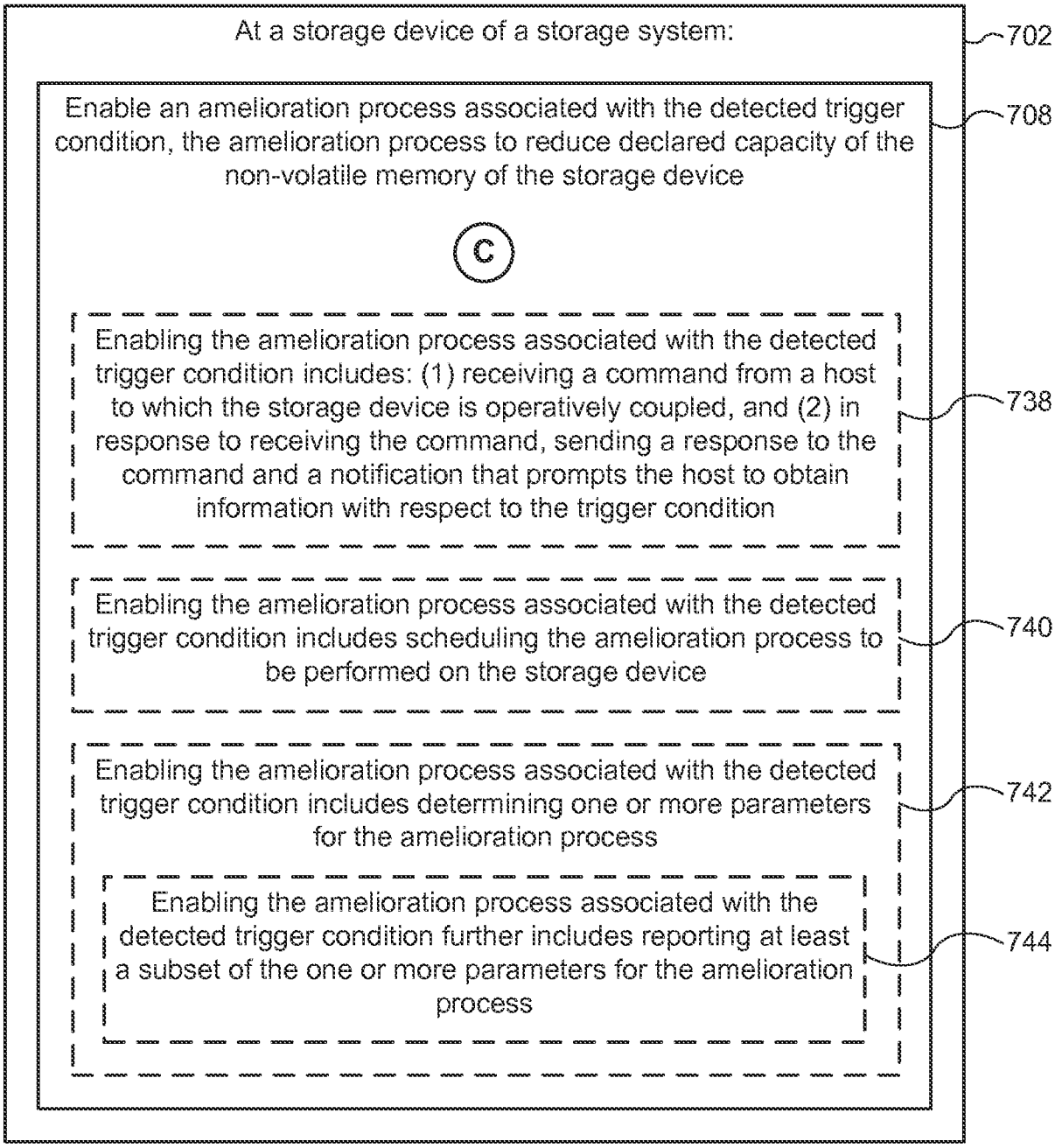

At a storage device of a storage system:
702
Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device
708
C
Enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition
738
Enabling the amelioration process associated with the detected trigger condition includes scheduling the amelioration process to be performed on the storage device
740
Enabling the amelioration process associated with the detected trigger condition includes determining one or more parameters for the amelioration process
742
Enabling the amelioration process associated with the detected trigger condition further includes reporting at least a subset of the one or more parameters for the amelioration process
744

Figure 7D

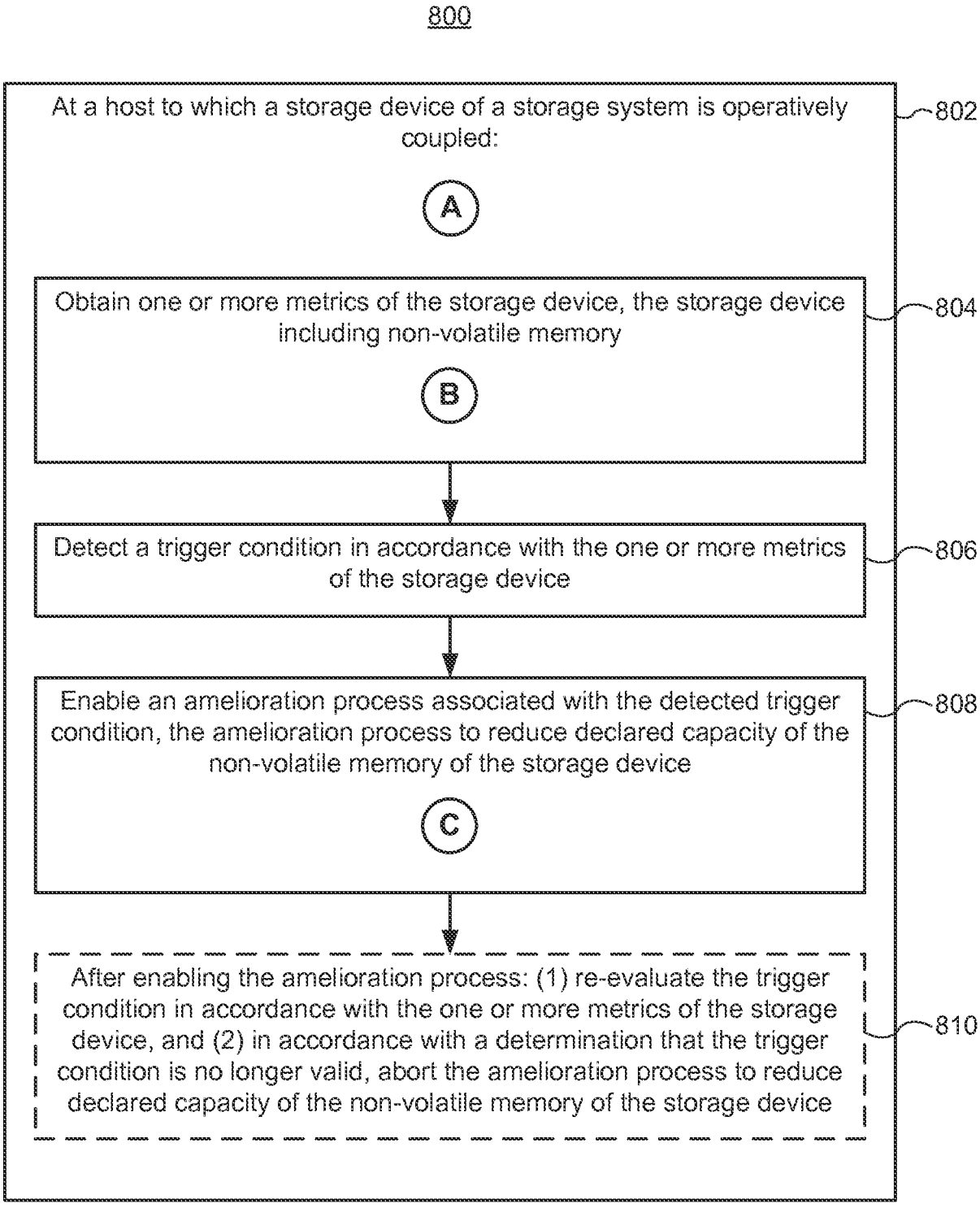

800
At a host to which a storage device of a storage system is operatively coupled:
802
A
Obtain one or more metrics of the storage device, the storage device including non-volatile memory
804
B
Detect a trigger condition in accordance with the one or more metrics of the storage device
806
Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device
808
C
After enabling the amelioration process: (1) re-evaluate the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, abort the amelioration process to reduce declared capacity of the non-volatile memory of the storage device
810

Figure 8A

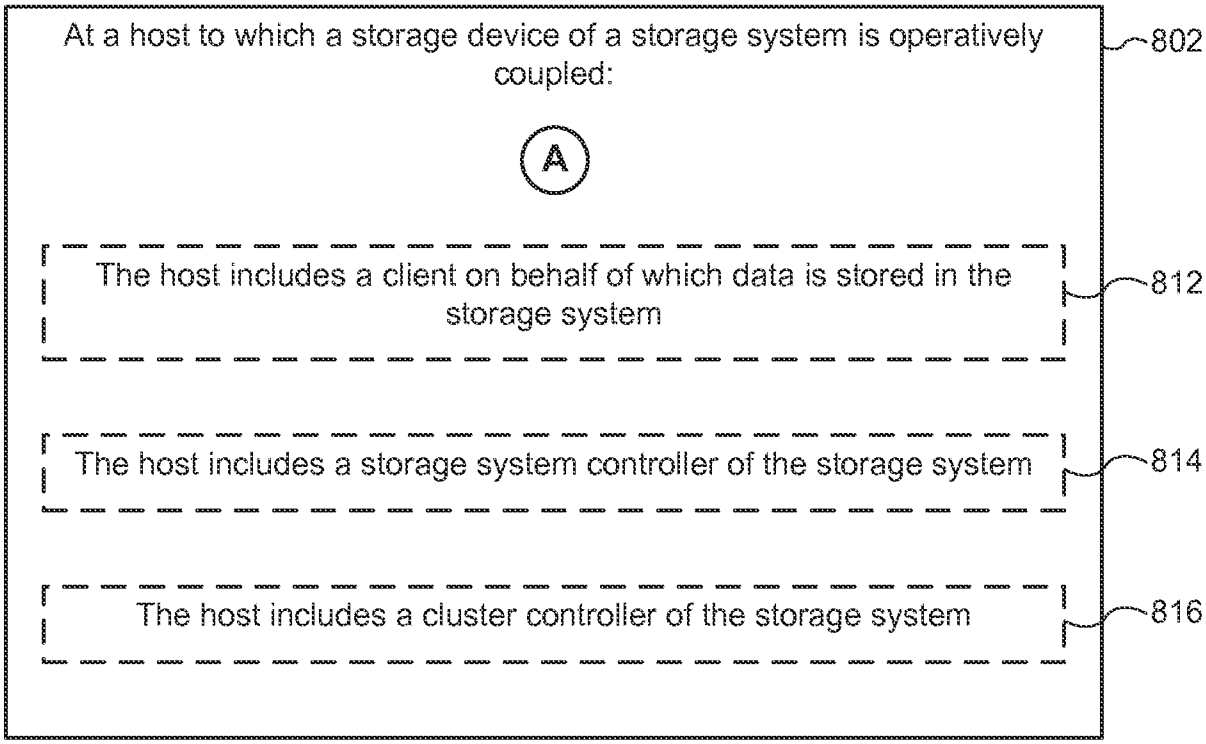

At a host to which a storage device of a storage system is operatively coupled:
802
A
The host includes a client on behalf of which data is stored in the storage system
812
The host includes a storage system controller of the storage system
814
The host includes a cluster controller of the storage system
816

Figure 8B

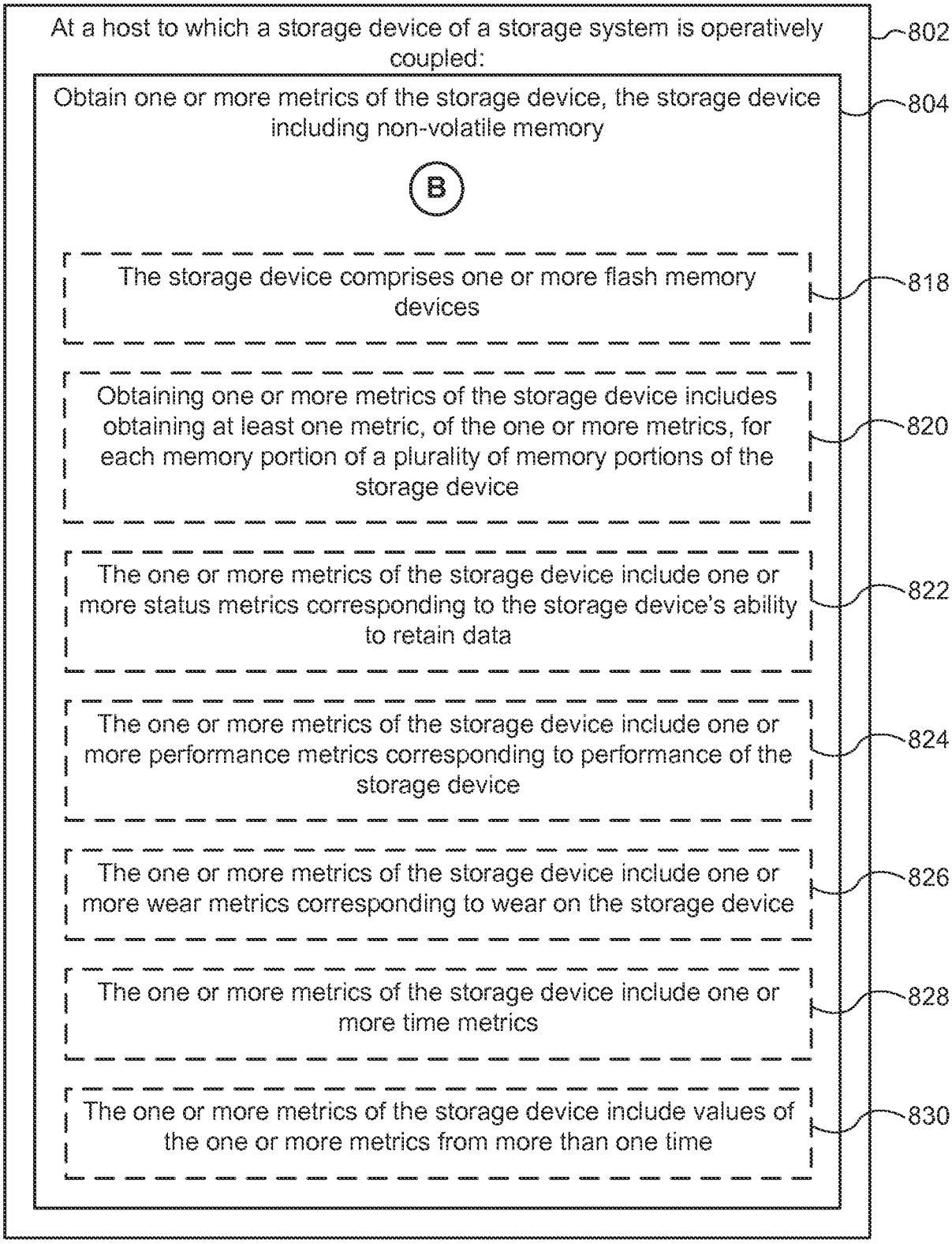

At a host to which a storage device of a storage system is operatively coupled:
802
Obtain one or more metrics of the storage device, the storage device including non-volatile memory
804
B
The storage device comprises one or more flash memory devices
818
Obtaining one or more metrics of the storage device includes obtaining at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the storage device
820
The one or more metrics of the storage device include one or more status metrics corresponding to the storage device's ability to retain data
822
The one or more metrics of the storage device include one or more performance metrics corresponding to performance of the storage device
824
The one or more metrics of the storage device include one or more wear metrics corresponding to wear on the storage device
826
The one or more metrics of the storage device include one or more time metrics
828
The one or more metrics of the storage device include values of the one or more metrics from more than one time
830

Figure 8C

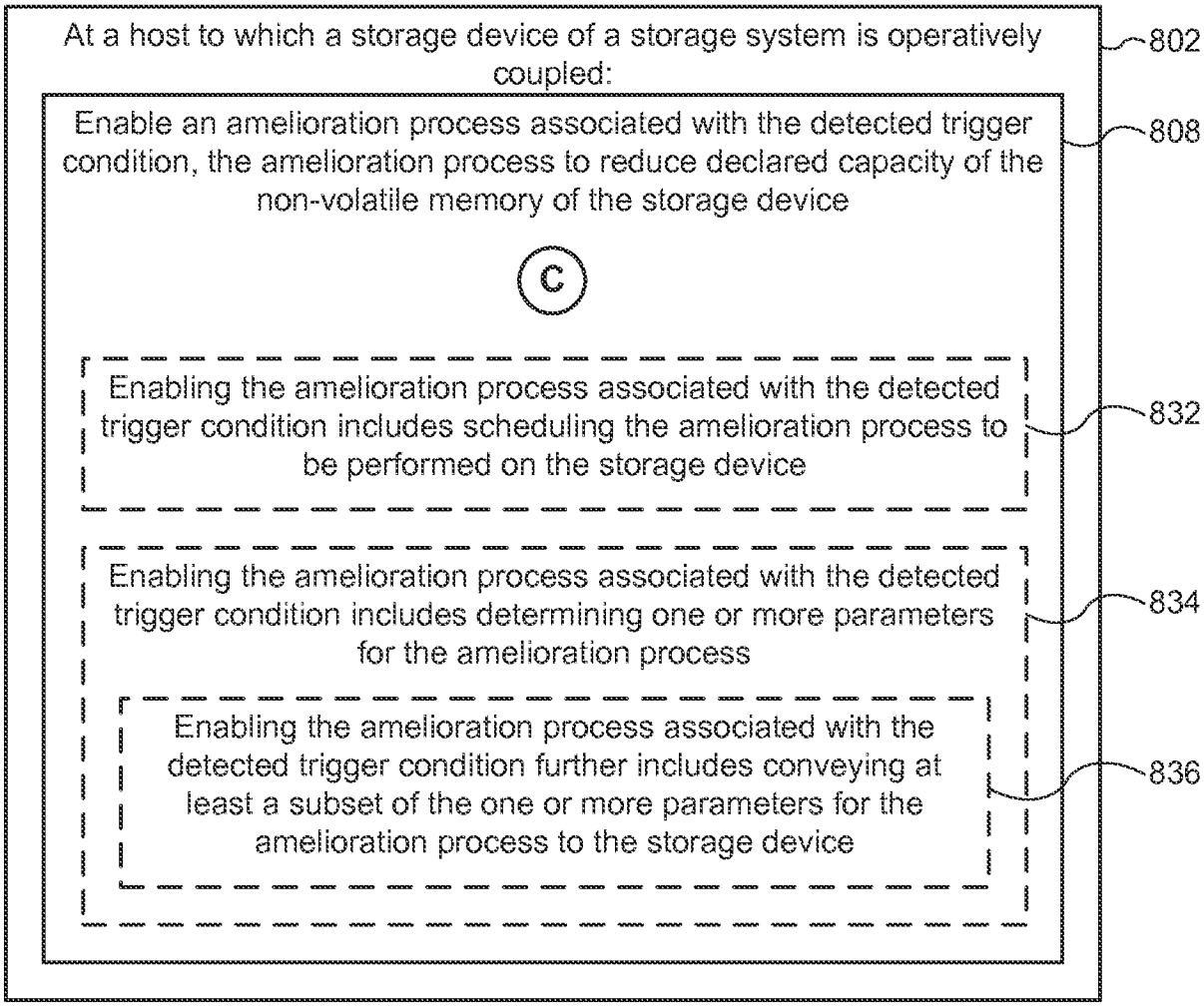

At a host to which a storage device of a storage system is operatively coupled:
802
Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the storage device
808
C
Enabling the amelioration process associated with the detected trigger condition includes scheduling the amelioration process to be performed on the storage device
832
Enabling the amelioration process associated with the detected trigger condition includes determining one or more parameters for the amelioration process
834
Enabling the amelioration process associated with the detected trigger condition further includes conveying at least a subset of the one or more parameters for the amelioration process to the storage device
836

Figure 8D

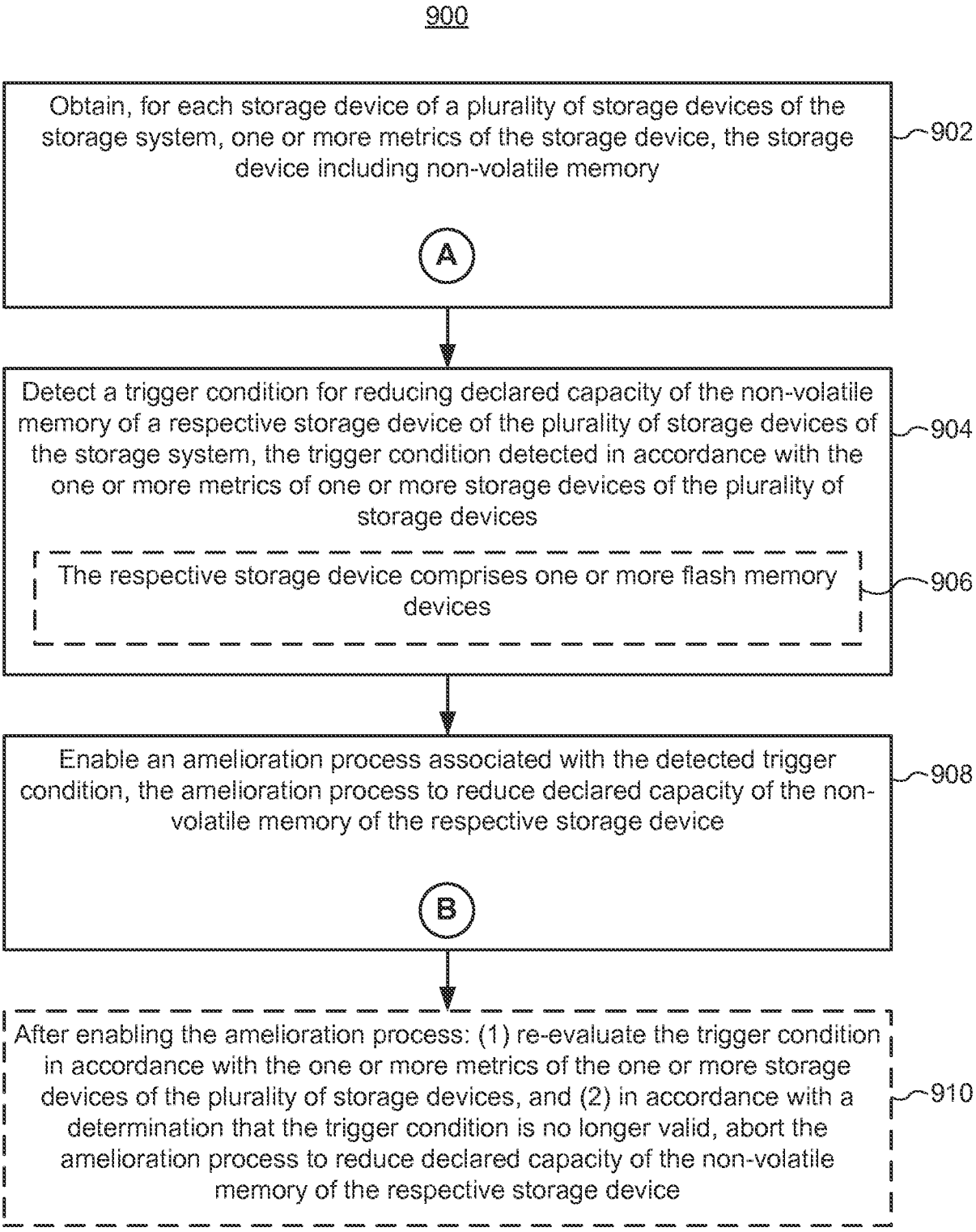

900
Obtain, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory
902
A
Detect a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of one or more storage devices of the plurality of storage devices
904
The respective storage device comprises one or more flash memory devices
906
Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device
908
B
After enabling the amelioration process: (1) re-evaluate the trigger condition in accordance with the one or more metrics of the one or more storage devices of the plurality of storage devices, and (2) in accordance with a determination that the trigger condition is no longer valid, abort the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device
910

Figure 9A

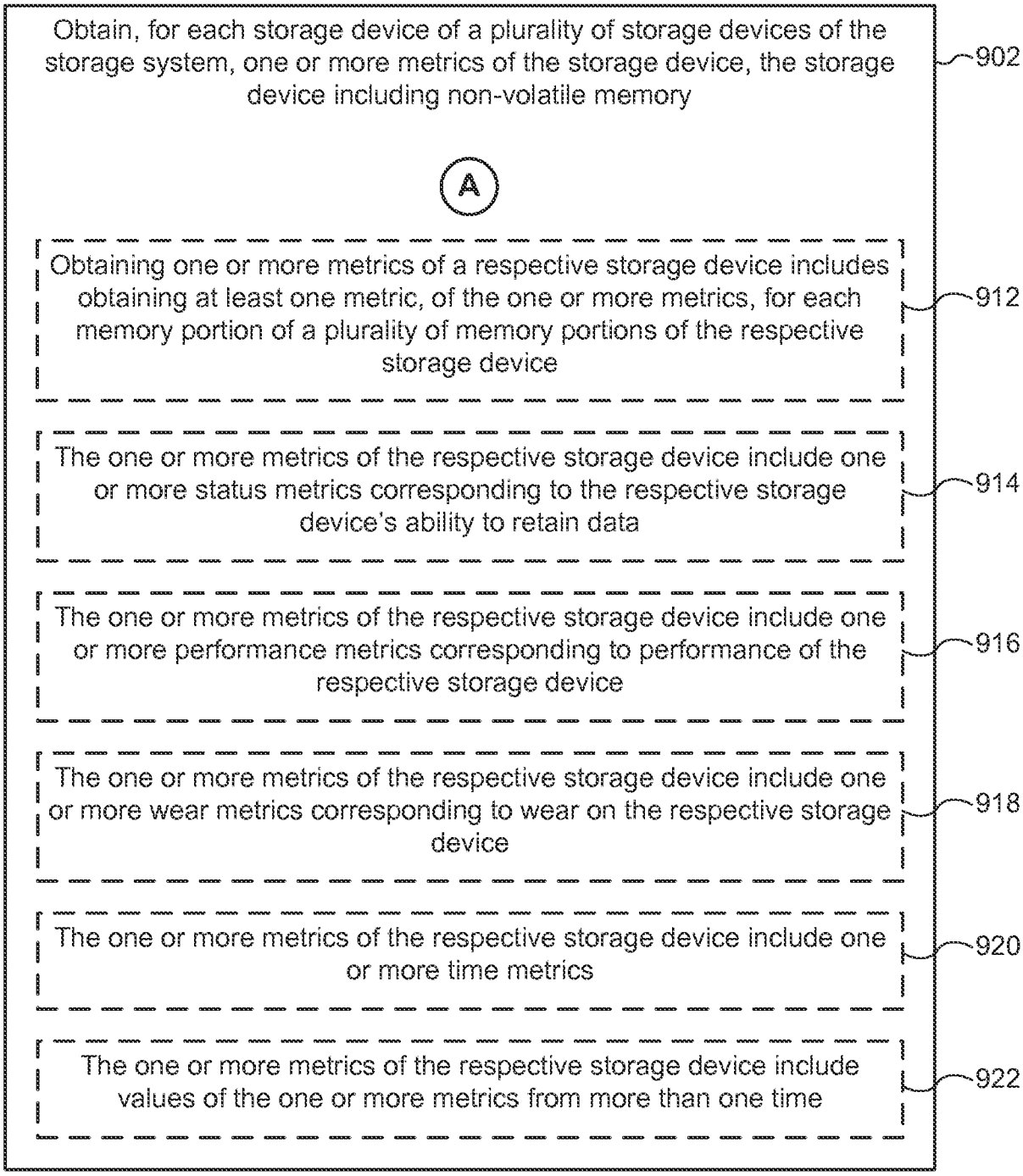

Obtain, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory
902
A
Obtaining one or more metrics of a respective storage device includes obtaining at least one metric, of the one or more metrics, for each memory portion of a plurality of memory portions of the respective storage device
912
The one or more metrics of the respective storage device include one or more status metrics corresponding to the respective storage device's ability to retain data
914
The one or more metrics of the respective storage device include one or more performance metrics corresponding to performance of the respective storage device
916
The one or more metrics of the respective storage device include one or more wear metrics corresponding to wear on the respective storage device
918
The one or more metrics of the respective storage device include one or more time metrics
920
The one or more metrics of the respective storage device include values of the one or more metrics from more than one time
922

Figure 9B

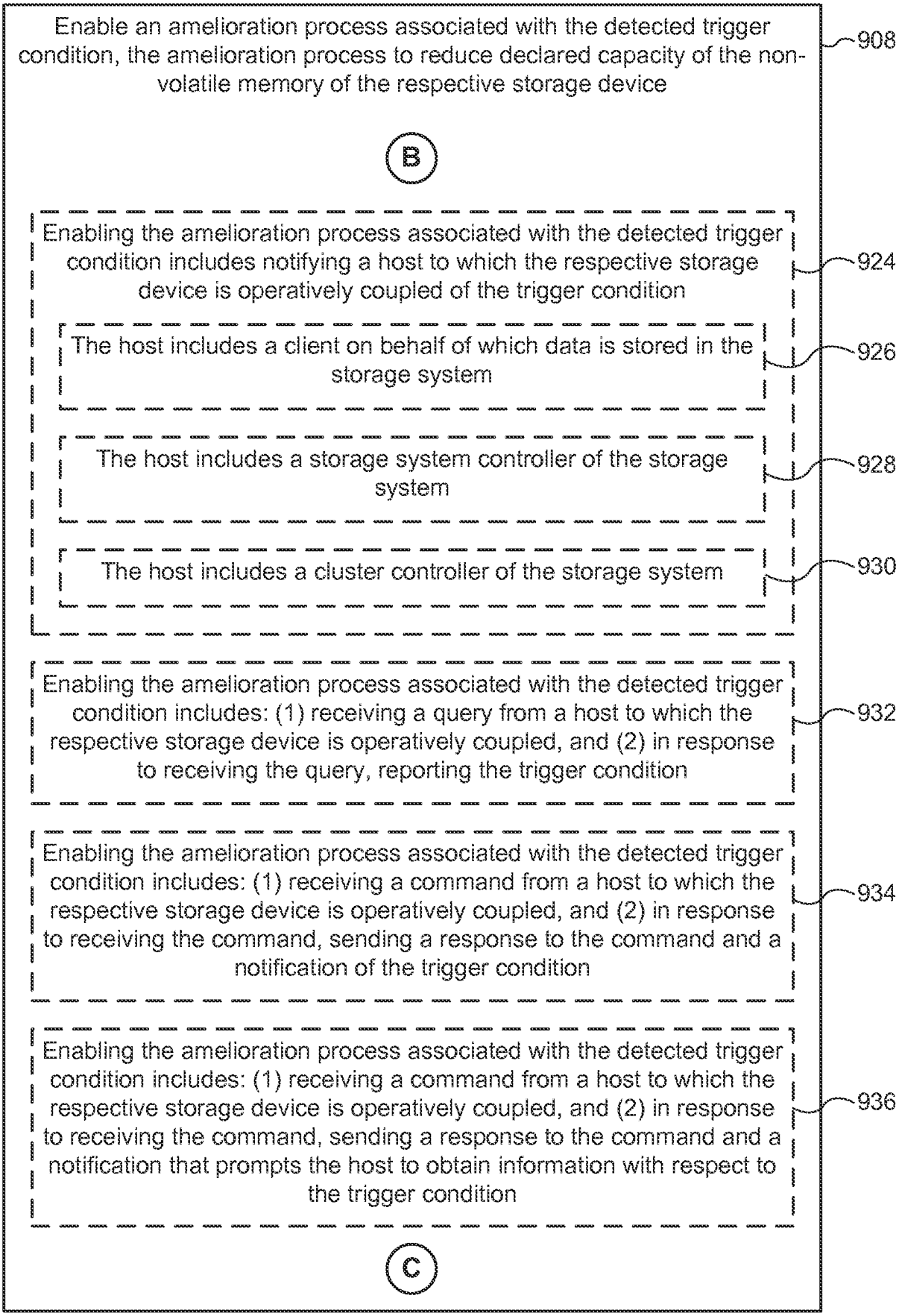

Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device
908
B
Enabling the amelioration process associated with the detected trigger condition includes notifying a host to which the respective storage device is operatively coupled of the trigger condition
924
The host includes a client on behalf of which data is stored in the storage system
926
The host includes a storage system controller of the storage system
928
The host includes a cluster controller of the storage system
930
Enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a query from a host to which the respective storage device is operatively coupled, and (2) in response to receiving the query, reporting the trigger condition
932
Enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the respective storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition
934
Enabling the amelioration process associated with the detected trigger condition includes: (1) receiving a command from a host to which the respective storage device is operatively coupled, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition
936
C

Figure 9C

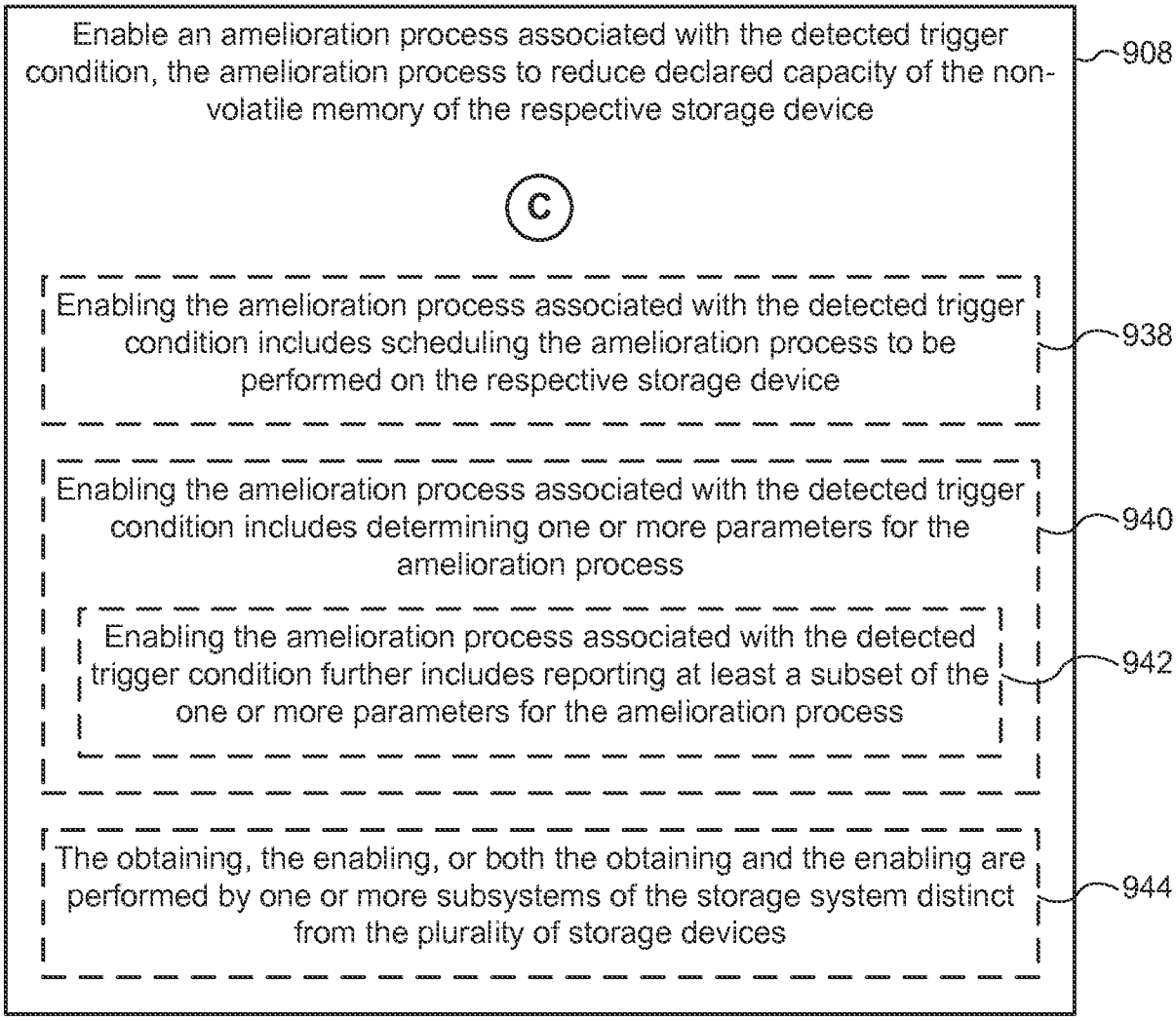

Enable an amelioration process associated with the detected trigger condition, the amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device
908
C
Enabling the amelioration process associated with the detected trigger condition includes scheduling the amelioration process to be performed on the respective storage device
938
Enabling the amelioration process associated with the detected trigger condition includes determining one or more parameters for the amelioration process
940
Enabling the amelioration process associated with the detected trigger condition further includes reporting at least a subset of the one or more parameters for the amelioration process
942
The obtaining, the enabling, or both the obtaining and the enabling are performed by one or more subsystems of the storage system distinct from the plurality of storage devices
944

Figure 9D

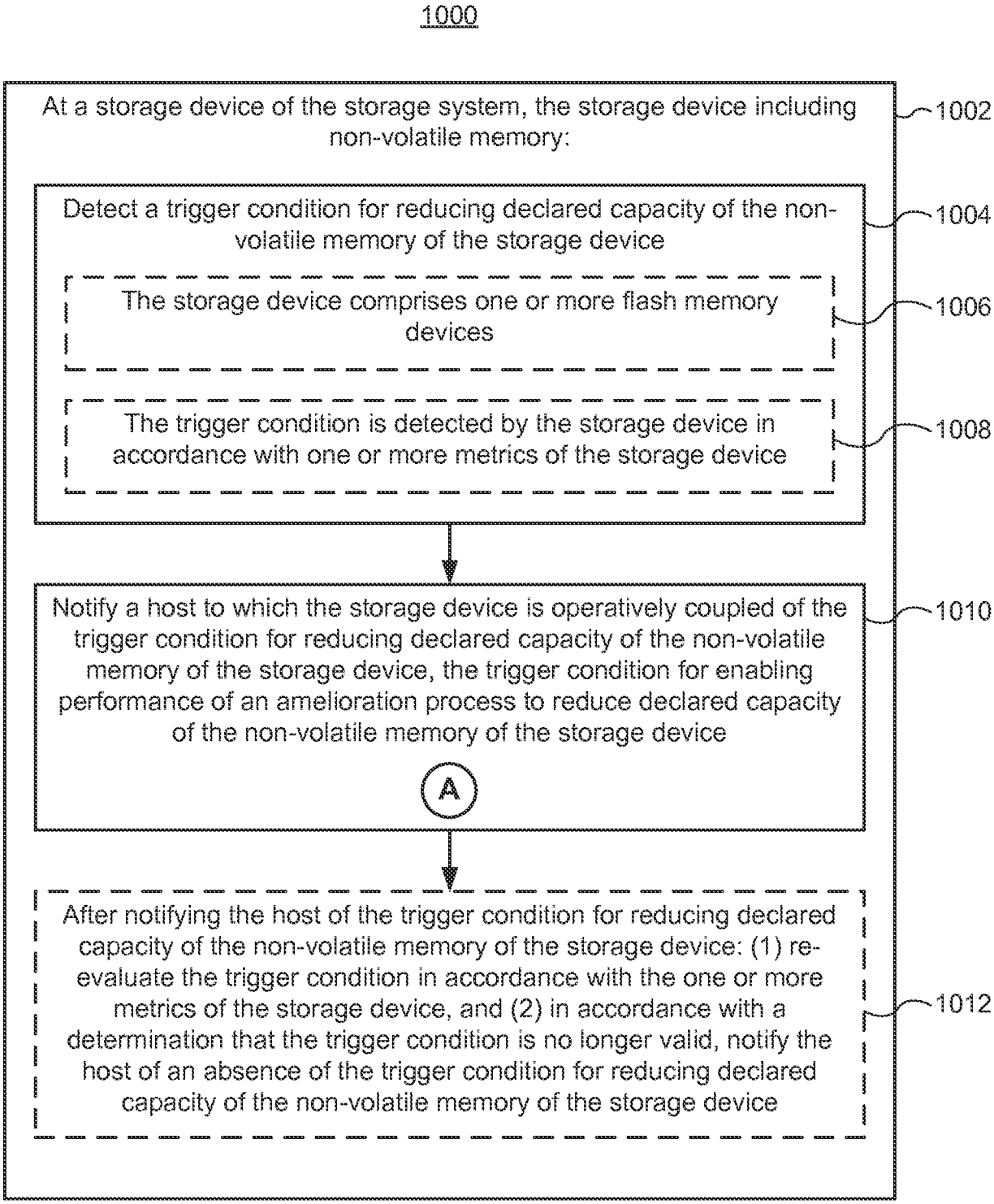

1000
At a storage device of the storage system, the storage device including non-volatile memory:
1002
Detect a trigger condition for reducing declared capacity of the non-volatile memory of the storage device
1004
The storage device comprises one or more flash memory devices
1006
The trigger condition is detected by the storage device in accordance with one or more metrics of the storage device
1008
Notify a host to which the storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the storage device
1010
A
After notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device: (1) re-evaluate the trigger condition in accordance with the one or more metrics of the storage device, and (2) in accordance with a determination that the trigger condition is no longer valid, notify the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device
1012

Figure 10A

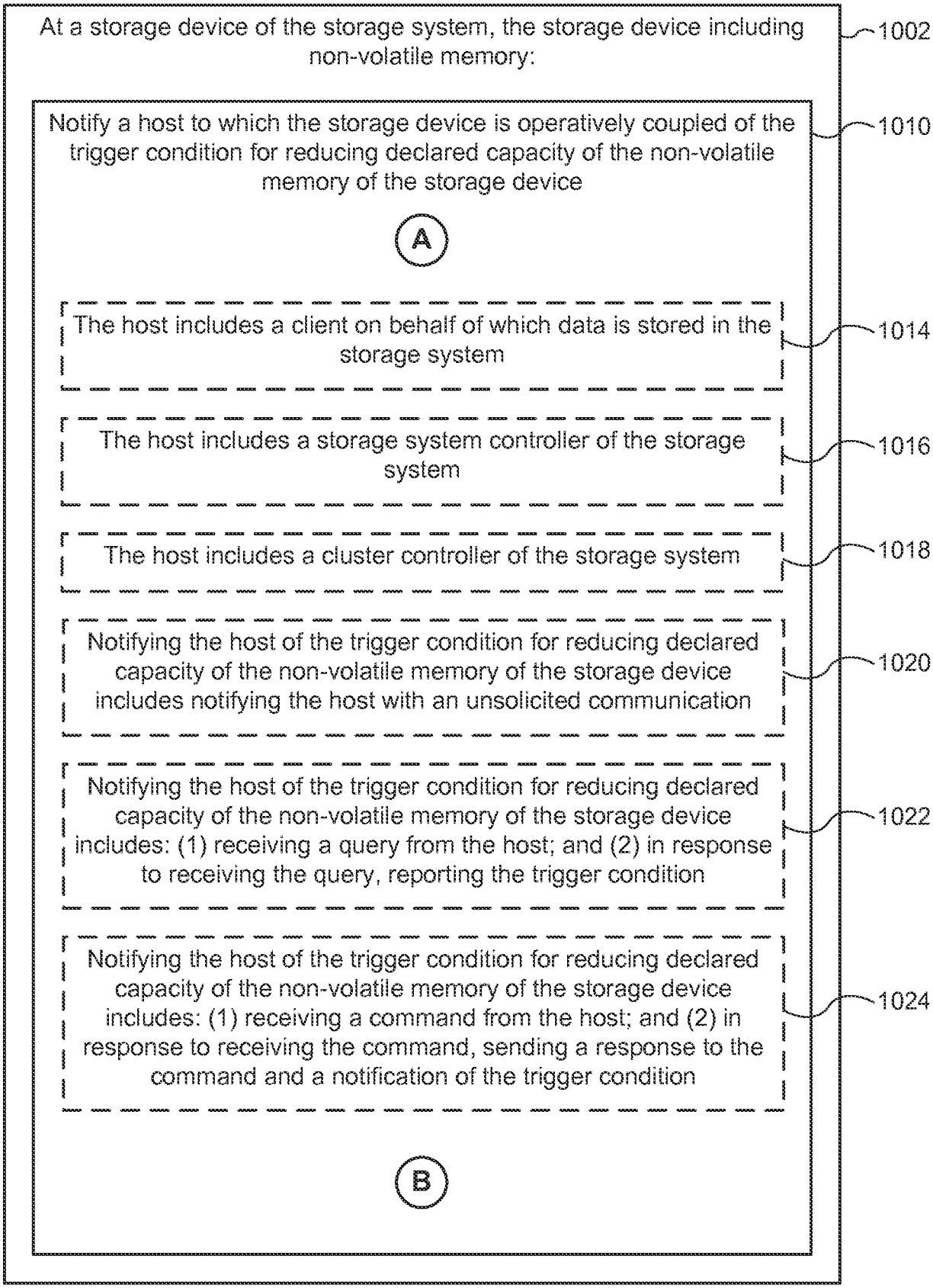

At a storage device of the storage system, the storage device including non-volatile memory:
1002
Notify a host to which the storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device
1010
A
The host includes a client on behalf of which data is stored in the storage system
1014
The host includes a storage system controller of the storage system
1016
The host includes a cluster controller of the storage system
1018
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes notifying the host with an unsolicited communication
1020
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes: (1) receiving a query from the host; and (2) in response to receiving the query, reporting the trigger condition
1022
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes: (1) receiving a command from the host; and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition
1024
B

Figure 10B

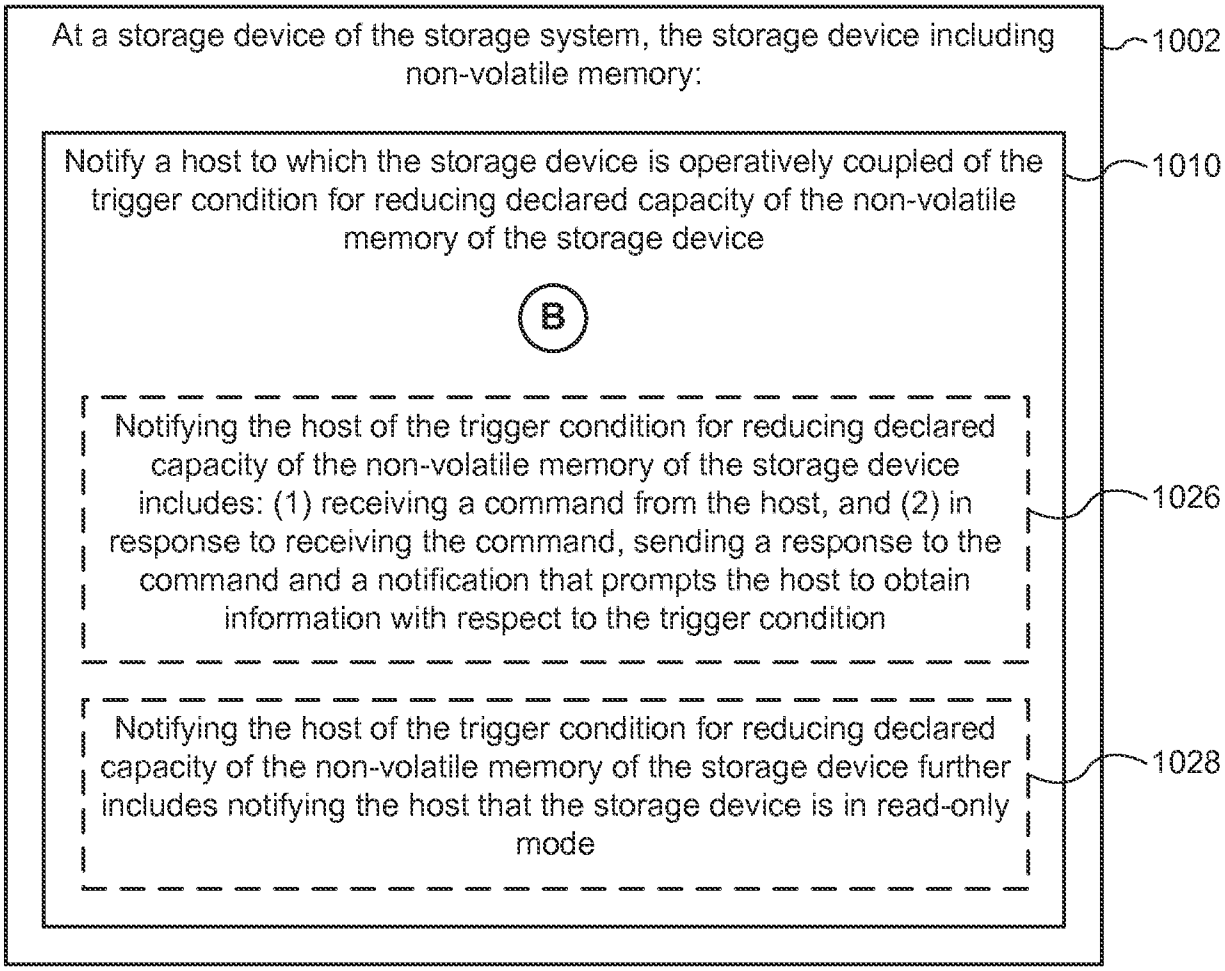

At a storage device of the storage system, the storage device including non-volatile memory:
1002
Notify a host to which the storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device
1010
B
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device includes: (1) receiving a command from the host, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition
1026
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the storage device further includes notifying the host that the storage device is in read-only mode
1028

Figure 10C

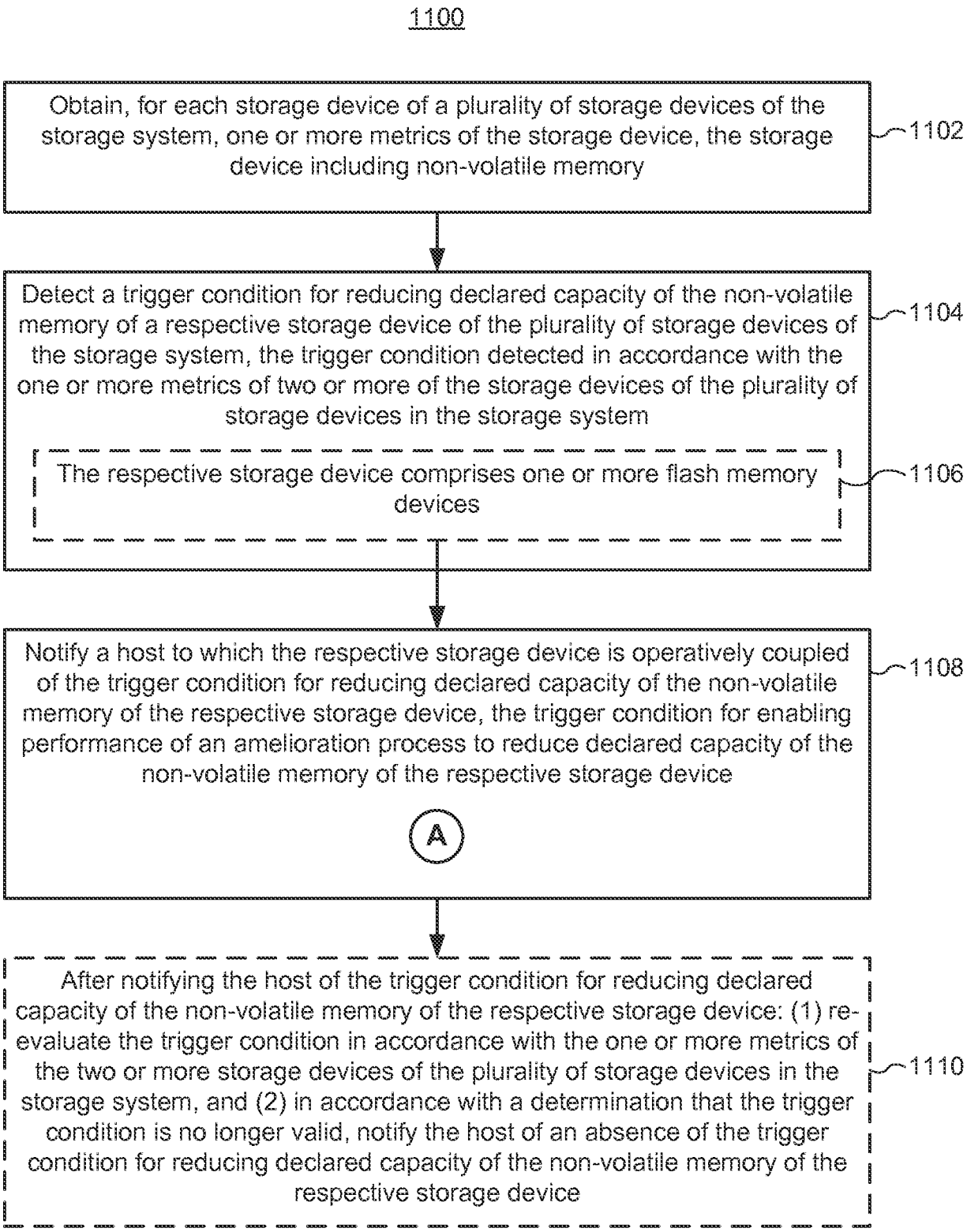

1100
Obtain, for each storage device of a plurality of storage devices of the storage system, one or more metrics of the storage device, the storage device including non-volatile memory
1102
Detect a trigger condition for reducing declared capacity of the non-volatile memory of a respective storage device of the plurality of storage devices of the storage system, the trigger condition detected in accordance with the one or more metrics of two or more of the storage devices of the plurality of storage devices in the storage system
1104
The respective storage device comprises one or more flash memory devices
1106
Notify a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device
1108
A
After notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device: (1) re-evaluate the trigger condition in accordance with the one or more metrics of the two or more storage devices of the plurality of storage devices in the storage system, and (2) in accordance with a determination that the trigger condition is no longer valid, notify the host of an absence of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device
1110

Figure 11A

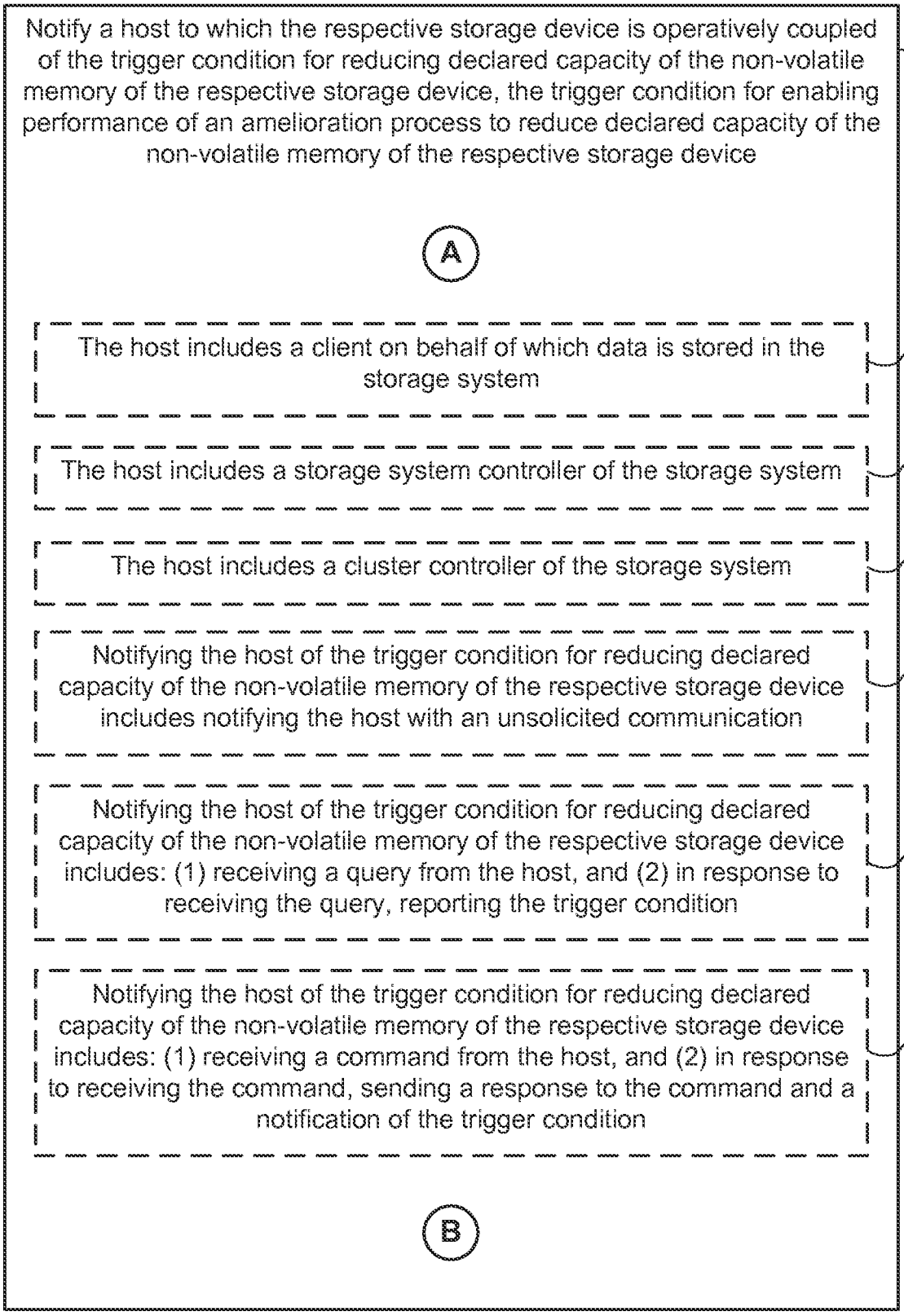

Notify a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device
1108
A
The host includes a client on behalf of which data is stored in the storage system
1112
The host includes a storage system controller of the storage system
1114
The host includes a cluster controller of the storage system
1116
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying the host with an unsolicited communication
1118
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes: (1) receiving a query from the host, and (2) in response to receiving the query, reporting the trigger condition
1120
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes: (1) receiving a command from the host, and (2) in response to receiving the command, sending a response to the command and a notification of the trigger condition
1122
B

Figure 11B

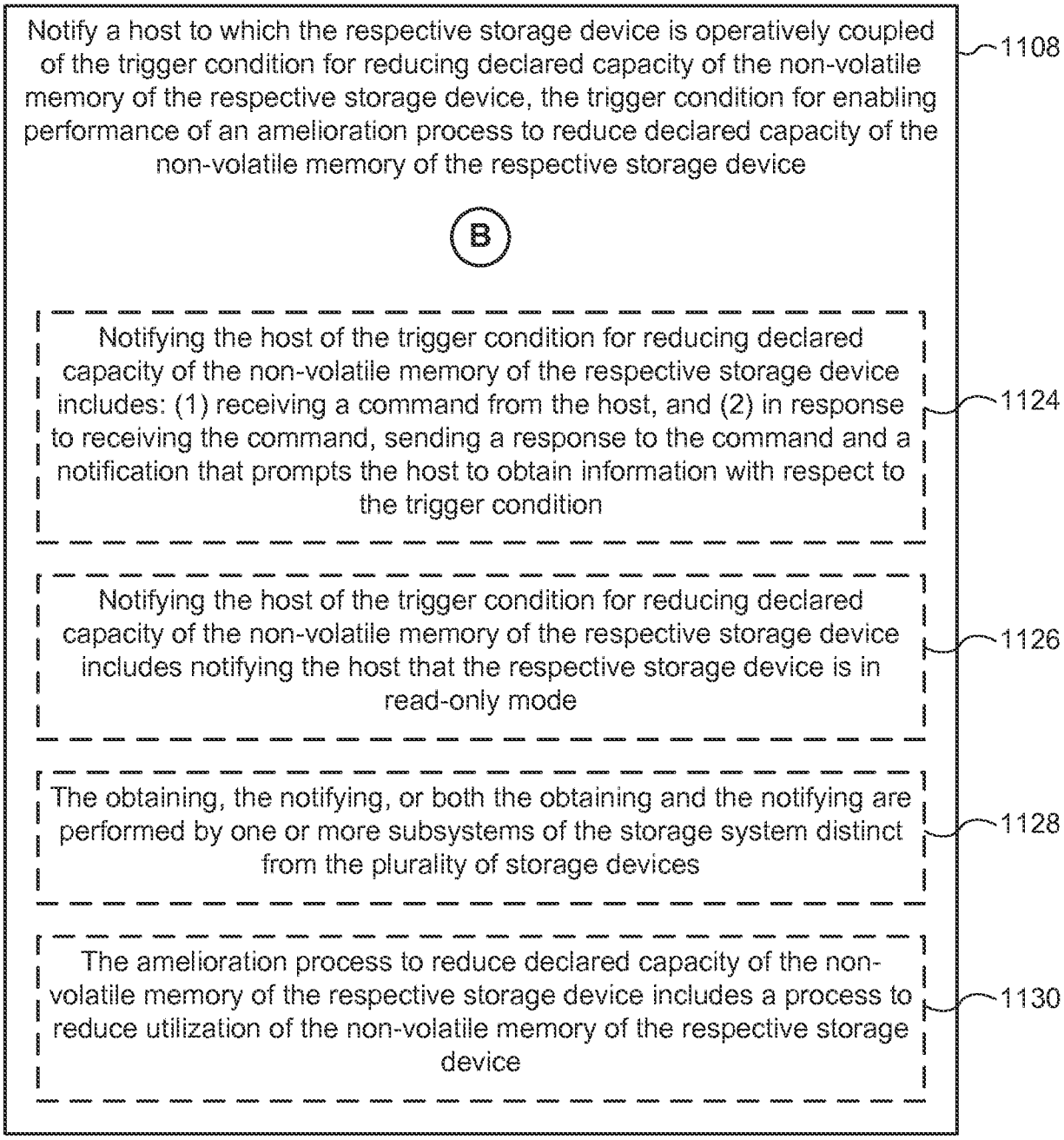

Notify a host to which the respective storage device is operatively coupled of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device, the trigger condition for enabling performance of an amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device
1108
B
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes: (1) receiving a command from the host, and (2) in response to receiving the command, sending a response to the command and a notification that prompts the host to obtain information with respect to the trigger condition
1124
Notifying the host of the trigger condition for reducing declared capacity of the non-volatile memory of the respective storage device includes notifying the host that the respective storage device is in read-only mode
1126
The obtaining, the notifying, or both the obtaining and the notifying are performed by one or more subsystems of the storage system distinct from the plurality of storage devices
1128
The amelioration process to reduce declared capacity of the non-volatile memory of the respective storage device includes a process to reduce utilization of the non-volatile memory of the respective storage device
1130

Figure 11C

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/047898

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

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

B. FIELDS SEARCHED

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2010/008140 A1 (LEE CHANG-HYUN [KR]) 14 January 2010 (2010-01-14) figures 9, 7 paragraph [0048] - paragraph [0062] -----	1-24

☐ 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
25 September 2015	08/10/2015

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

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2015/047898

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2010008140 A1	14-01-2010	CN 101625894 A	13-01-2010
		KR 20100007191 A	22-01-2010
		US 2010008140 A1	14-01-2010