

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
26 October 2006 (26.10.2006)

PCT

(10) International Publication Number  
**WO 2006/113334 A2**

(51) International Patent Classification:

G06F 11/00 (2006.01)

(21) International Application Number:

PCT/US2006/013835

(22) International Filing Date: 13 April 2006 (13.04.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

11/107,551 15 April 2005 (15.04.2005) US

(71) Applicant (for all designated States except US): **MICROSOFT CORPORATION** [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(72) Inventors: **NICHOLSON, Clark D.**; One Microsoft Way, Redmond, Washington 98052-6399 (US). **FORTIN, Michael R.**; One Microsoft Way, Redmond, Washington 98052-6399 (US). **WILEY, Shaun B.**; One Microsoft Way, Redmond, Washington 98052-6399 (US). **ERGAN, Cenk**; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

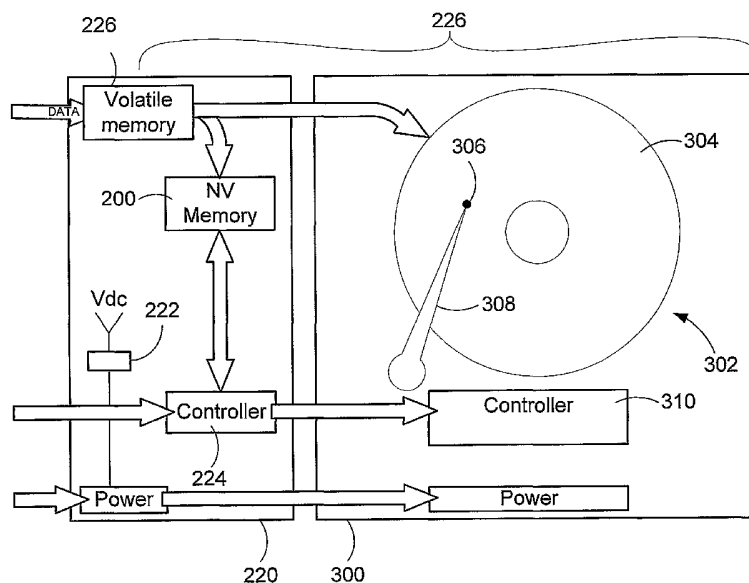
- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))
- as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))

Published:

- without international search report and to be republished upon receipt of that report

[Continued on next page]

(54) Title: IN-LINE NON VOLATILE MEMORY DISK READ CACHE AND WRITE BUFFER



(57) Abstract: A method and apparatus to improve the read/write performance of a hard drive is presented. A device having solid state, non-volatile (NV) memory is added in-line to the conventional hard drive and acts as a read/write cache. Data specified by the operating system is stored in the NV memory. The operating system provides a list of data to be put in NV memory. The data includes data to be pinned in NV memory and data that is dynamic. Pinned data persists in NV memory until the operating system commands it to be flushed. Dynamic data can be flushed by the hard drive controller. Data sent by an application for storage is temporarily stored in NV memory in data blocks until the operating system commits it to the disk.



---

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

## **IN-LINE NON VOLATILE MEMORY DISK READ CACHE AND WRITE BUFFER**

### **FIELD OF THE INVENTION**

**[0001]** This invention relates generally to computer systems and, more particularly, relates to improving the performance of hard disk drives.

### **BACKGROUND OF THE INVENTION**

**[0002]** Certain persistent storage devices are used on computer systems in order to provide persistent memory for the functioning of the computer system. For example, a magnetic hard disk drive assembly is well known in the art. However, such persistent memory has some drawbacks. It is time consuming to prepare the magnetic hard drive for reading or writing data. In order to do so, in most cases, it is necessary to physically initiate spinning of the disk, and to position a read/write head over the proper location on the disk.

**[0003]** However, the computer system's microprocessor executes instructions only from fast randomly addressable memory such as DRAM or some other type of volatile electronic memory. In order to initialize this addressable memory to boot the operating system, a small amount of non-volatile boot-up memory is typically provided in an EPROM (electronically programmable read-only memory) or similar device. When the computer is booted, a boot program stored in the non-volatile memory is used. The boot program contains code for reading an initial portion of the operating system on the hard disk that is

responsible for loading and initializing the remaining portions of the operating system.

[0004] The transfer of the operating system, executables, and data from the hard drive is slow, particularly when the operating system is a large operating system, a file is large, etc. This is prevalent not only in booting the computer system, but also in restoring the computer system from hibernate mode. A significant factor in the transfer time is due to the disk drive spin up speed. A relatively small disk spinning at a relatively slow RPM requires 5 to 6 seconds to spin up and be usable. Larger disks such as multi-platter devices and those spinning at faster RPMs require 10 to 12 seconds or more to spin up. On server systems having hundreds or thousands of disk drives that spin up in a largely sequential manner, the disk spin up time can be over ninety percent of the total time to boot the server system. The hard drive consumes a significant share of the power used by a computer system, contributes substantially to thermal load, and causes the PC to be noisy due to the fan noise needed to cool the system.

[0005] Methods have been developed for increasing the speed in booting the computer system and restoring the computer system from hibernation. One example is disclosed in to Feigenbaum et al U.S. Pat. No. 5,307,497. The '497 patent teaches that a portion of an operating system can be stored in read-only memory (ROM) for fast access of the portion during power-on or reset. However, the portion stored in ROM can not be changed. In other words, the method disclosed in the '497 patent does not adapt to changing user preferences regarding operating system settings and applications used, to restoring the computer system from hibernation, or to updated or new versions of an operating system.

[0006] Another example developed by industry is storing the state of the computer memory to a backup memory that is separate from the disk drive and stores the entire

computer memory. This backup memory is large because it stores the entire computer memory.

### BRIEF SUMMARY OF THE INVENTION

[0007] The present invention improves the read/write performance of a hard drive and reduces the number of times the hard drive disk is accessed. The operating system provides a list of data to be put in solid state, non-volatile ("NV") memory in a device in-line to the hard drive and motherboard and attached to the hard drive, thereby allowing standard mechanical drives to be used with standard mother boards. The data includes data to be pinned in NV memory and data that is dynamic. Pinned data persists in NV memory until the operating system commands it to be flushed. Dynamic data is data that can be prioritized by the operating system but are otherwise controllable by the hard drive controller.

[0008] The NV memory is provided in a unit that installs on the IDE connector or equivalent on a disk drive. The unit can be strapped onto the disk drive so that it is relatively difficult to remove the unit from the disk drive. The strapping reduces the possibility that the drive would be removed from the computer system after a power failure and be corrupted. In one embodiment, the device includes fast volatile memory to buffer bursts of data written to the disk and an energy storage device, such as a small battery, that provides sufficient power to keep the device powered long enough to move data from volatile memory to the NV memory to the disk drive platters in case of sudden power failure. During operation, the unit's controller caches commands from the operating system and buffers writes so that the mechanical disk can be spun down, thereby reducing latency,

power consumption, and wear and tear on the drive.

[0009] The NV memory is used during system operation to temporarily buffer data written to the drive. When available resources on the NV memory falls below a threshold, the data stored in the NV memory is transferred to the disk drive. This reduces the number of times the disk drive is spun up when the system is operating.

[0010] An interface for enabling an operating system to control a NV memory cache in the unit is also provided. The interface comprises commands for the operating system, the unit, and the hard drive to communicate. The interface includes a command to determine capabilities of the hard drive; a command to pin a list of blocks in the NV memory; a command to pre-cache a list of blocks in the NV memory, a command to flush commands from the NV memory to a hard drive disk, a command to write a set of blocks that comprise a write transaction to NV memory and a command to commit the write transaction to the hard drive disk.

[0011] Additional features and advantages of the invention will be made apparent from the following detailed description of illustrative embodiments which proceeds with reference to the accompanying figures.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[0012] While the appended claims set forth the features of the present invention with particularity, the invention, together with its objects and advantages, may be best

understood from the following detailed description taken in conjunction with the accompanying drawings of which:

[0013] FIG. 1 is a block diagram generally illustrating an exemplary computer system on which the present invention resides;

[0014] FIG. 2 is a block diagram of the exemplary computer system of FIG. 1 generally illustrating locations where the present invention may be implemented;

[0015] FIG. 3 is a block diagram illustrating an in-line NV memory unit embodying an embodiment of the invention connected to a standard disk drive;

[0016] FIGS. 4a and 4b are isometric views illustrating the in-line NV memory unit of FIG. 3 connected to a standard disk drive;

[0017] FIG. 5 is a flow chart illustrating a first portion of a process for storing static and dynamic configuration data during power-down or entering into a hibernate mode in order to decrease a boot time and awaken time of a computer system;

[0018] FIG. 6 is a flow chart illustrating a remaining portion of the process for storing static and dynamic configuration data during power-down or entering into a hibernate mode in order to decrease a boot time and awaken time of a computer system;

[0019] FIG. 7 is a flow chart illustrating a process for decreasing a boot time of a computer system and an awaken time from hibernation of a computer system;

[0020] FIG. 8 is a block diagram illustrating a storage network stack of an operating system interacting with the present invention;

[0021] FIG. 9 is a block diagram illustrating a storage network stack controlling the NVM memory in accordance with the present invention;

[0022] FIG. 10 is a block diagram illustrating an alternate storage network stack controlling the NVM memory in accordance with the present invention; and

[0023] FIG. 11 is a flow chart illustrating a process for storing data in NV memory in accordance with the teachings of the present invention.

#### DETAILED DESCRIPTION OF THE INVENTION

[0024] The present invention reduces power consumption of conventional hard disk drives and improves the read/write performance of hard disk drives, decreases the boot time and the hibernate awaken time of a computer system. Static and dynamic configuration data are stored in solid state, nonvolatile memory technology such as flash memory. Static configuration data is the operating system and application program instructions and data that do not change each time the operating system boots or an application launches except after upgrades to the operating system have been installed or applications have been installed. Dynamic configuration data is program instructions and data that changes with each boot of the operating system or each launch of an application and may change at any time during execution. In the description that follows, the term "NV memory" will be used to refer to solid state non-volatile memory. NV memory means any type of memory that



retains its data when power is removed and that can be erased and reprogrammed with new data. The NV memory may include battery-backed up memory. The NV memory is located inline with the disk drive interface between the motherboard and the hard disk drive.

[0025] Future disks for mobile systems may draw from battery power to keep its memory alive without actually drawing power to keep the disk spinning. The NV memory may store file data, file metadata, as well as known disk sectors such as the boot sector and partition tables. During power down or entry into a hibernate mode, the computer system determines the static and dynamic configuration data to be stored in NV memory based on a history of prior uses. Data is also stored in the NV memory during system operation to reduce the number of times the disk drive is spun up. For mobile PCs and the like, reducing the number of times the disk drive is spun up improves user experience because power is reduced, thereby allowing smaller and lighter mobile PCs and the like that last longer on a given battery capacity. It also reduces the need to run noisy, high speed fans that are loud and annoying to the user. When the computer system is powered up or awakened from hibernation, the configuration data in NV memory is used to initialize the dynamic system memory without requiring the disk drive to spin up. In cases where *additional files are required from the disk*, the disk drive can be spun up simultaneously to the initialization from NV memory so that the disk spin up does not slow down the initialization process.

[0026] Turning to the drawings, wherein like reference numerals refer to like elements, the invention is illustrated as being implemented in a suitable computing environment. Although not required, the invention will be described in the general context of computer-

executable instructions, such as program modules, being executed by a personal computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Moreover, those skilled in the art will appreciate that the invention may be practiced with other computer system configurations, including hand-held devices, multi-processor systems, microprocessor based or programmable consumer electronics, network PCs, minicomputers, mainframe computers, and the like. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

[0027] FIG. 1 illustrates an example of a suitable computing system environment 100 on which the invention may be implemented. The computing system environment 100 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the invention. Neither should the computing environment 100 be interpreted as having any dependency or requirement relating to any one or combination of components illustrated in the exemplary operating environment 100.

[0028] The invention is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well known computing systems, environments, and/or configurations that may be suitable for use with the invention include, but are not limited to, personal computers, server computers, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, set top

boxes, programmable consumer electronics, network PCs, minicomputers, mainframe computers, distributed computing environments that include any of the above systems or devices, and the like.

[0029] The invention may be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote computer storage media including memory storage devices.

[0030] With reference to FIG. 1, an exemplary system for implementing the invention includes a general purpose computing device in the form of a computer 110. Components of computer 110 may include, but are not limited to, a processing unit 120, a system memory 130, and a system bus 121 that couples various system components including the system memory to the processing unit 120. The system bus 121 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Associate (VESA) local bus, and Peripheral Component Interconnect (PCI) bus also known as Mezzanine bus.

[0031] Computer 110 typically includes a variety of computer readable media.

Computer readable media can be any available media that can be accessed by computer 110 and includes both volatile and nonvolatile media, removable and non-removable media. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media. Computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by computer 110. Communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of the any of the above should also be included within the scope of computer readable media.

[0032] The system memory 130 includes computer storage media in the form of volatile and/or nonvolatile memory such as read only memory (ROM) 131 and random access memory (RAM) 132. A basic input/output system 133 (BIOS), containing the basic

routines that help to transfer information between elements within computer 110, such as during start-up, is typically stored in ROM 131. RAM 132 typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit 120. By way of example, and not limitation, FIG. 1 illustrates operating system 134, application programs 135, other program modules 136, and program data 137.

[0033] The computer 110 may also include other removable/non-removable, volatile/nonvolatile computer storage media. By way of example only, FIG. 1 illustrates a hard disk drive 141 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 151 that reads from or writes to a removable, nonvolatile magnetic disk 152, and an optical disk drive 155 that reads from or writes to a removable, nonvolatile optical disk 156 such as a CD ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media that can be used in the exemplary operating environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The hard disk drive 141 is typically connected to the system bus 121 through a non-removable memory interface such as interface 140, and magnetic disk drive 151 and optical disk drive 155 are typically connected to the system bus 121 by a removable memory interface, such as interface 150.

[0034] The drives and their associated computer storage media discussed above and illustrated in FIG. 1, provide storage of computer readable instructions, data structures, program modules and other data for the computer 110. In FIG. 1, for example, hard disk drive 141 is illustrated as storing operating system 144, application programs 145, other program modules 146, and program data 147. Note that these components can either be the

same as or different from operating system 134, application programs 135, other program modules 136, and program data 137. Operating system 144, application programs 145, other program modules 146, and program data 147 are given different numbers hereto illustrate that, at a minimum, they are different copies. A user may enter commands and information into the computer 110 through input devices such as a keyboard 162 and pointing device 161, commonly referred to as a mouse, trackball or touch pad. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 120 through a user input interface 160 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). A monitor 191 or other type of display device is also connected to the system bus 121 via an interface, such as a video interface 190. In addition to the monitor, computers may also include other peripheral output devices such as speakers 197 and printer 196, which may be connected through an output peripheral interface 195.

[0035] The computer 110 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 180. The remote computer 180 may be another personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the personal computer 110, although only a memory storage device 181 has been illustrated in FIG. 1. The logical connections depicted in FIG. 1 include a local area network (LAN) 171 and a wide area network (WAN) 173, but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

[0036] When used in a LAN networking environment, the personal computer 110 is connected to the LAN 171 through a network interface or adapter 170. When used in a WAN networking environment, the computer 110 typically includes a modem 172 or other means for establishing communications over the WAN 173, such as the Internet. The modem 172, which may be internal or external, may be connected to the system bus 121 via the user input interface 160, or other appropriate mechanism. In a networked environment, program modules depicted relative to the personal computer 110, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, FIG. 1 illustrates remote application programs 185 as residing on memory device 181. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

[0037] In the description that follows, the invention will be described with reference to acts and symbolic representations of operations that are performed by one or more computers, unless indicated otherwise. As such, it will be understood that such acts and operations, which are at times referred to as being computer-executed, include the manipulation by the processing unit of the computer of electrical signals representing data in a structured form. This manipulation transforms the data or maintains it at locations in the memory system of the computer, which reconfigures or otherwise alters the operation of the computer in a manner well understood by those skilled in the art. The data structures where data is maintained are physical locations of the memory that have particular properties defined by the format of the data. However, while the invention is being described in the foregoing context, it is not meant to be limiting as those of skill in the art will appreciate that various of the acts and operation described hereinafter may also be implemented in hardware.

[0038] Turning now to FIG. 2, the NV memory 200 may be located in the computer 110 as a separate component 220 of the computer that is located inline with the disk drive interface between the motherboard and the hard disk drive. The NV memory 200 as a separate component 220 provides the capability to improve performance on legacy systems that do not have NV memory 200 without having to install a hard drive having NV memory.

[0039] Turning now to FIGs. 3 and 4a-4b, the persistent storage device 300 (i.e., hard disk drive 141) has a disk assembly 302 that includes a non-volatile memory storage medium such as one or more disks 304, a read/write head 306 for reading and writing to the disk 304, and a spindle/armature assembly 308 for moving the read/write head 306 to locations on the disk 304. The controller 310 commands the disk assembly to read data from the storage medium or write data to the storage medium. The NV memory 200 is provided in a device 220 that installs on the IDE connector 320 or equivalent (and the power connection 322) on a disk drive. The ATA (Advanced Technology Attachment) cable 324 connects to the device 220 and the motherboard 111. Power is provided to the device 220 via connector 326. While FIGs. 4a and 4b illustrate the device 220 installing on the IDE connector 320 of a standard desk-top hard drive with power provided via a separate connector, those skilled in the art will recognize that the device 220 is not limited to such connectors and that the device 220 can be installed on other connectors or connector such as, for example and not limitation, the hard drive connector of a mobile device such as a lap-top hard disk drive. The device 220 can be strapped or permanently mounted onto the disk drive so that it is difficult or nearly impossible to remove the device from the disk drive. The strapping/permanent mounting reduces the possibility that the



drive 300 would be removed from the computer system after a power failure and be corrupted.

[0040] In one embodiment, an energy storage device 222 is provided, such as a small battery, that provides sufficient power to maintain power to volatile memory 226 and non-volatile memory powered long enough to move data from the volatile memory to the NV memory in the event of sudden power failure.

[0041] During operation, the device's controller 224 caches or intercepts commands from the operating system and buffers writes to the NVM memory 200 instead of the disk 304 so that the mechanical disk 304 can be spun down, thereby reducing latency, power consumption, and wear and tear on the drive. The controller 224 also interprets the commands from the operating system and passes the commands on to the controller 310.

[0042] The computer 110 can detect the presence of NV memory by observing access times of 500 microseconds or less for large regions of the disk 304 instead of the typical times of approximately 10 milliseconds. In an alternate embodiment, the NV memory 200 is exposed to the operating system so that the operating system and other components can query the size of the NV memory 200, control the state of the NV memory 200, write data to the NV memory 200 using a transactional procedure, and as stated above, use the NV memory 200 as a write buffer. One embodiment of APIs that may be used is in commonly assigned U.S. Patent Application No. 10/837,986 filed May 3, 2004 and titled "Non-Volatile Memory Cache Performance Improvement" and is hereby incorporated in its entirety by reference.

[0043] Operation of the present invention will now be described with reference to FIGs. 5 to 7. It should be noted that the steps described below can be performed sequentially, in parallel, or a combination of sequential and parallel. The spin-up time of the disk 304 is determined (step 500). For relatively small disks spinning at relatively slow speeds, the spin-up time is typically 3 to 6 seconds. For larger disks, multi-platter disks, and disks spinning at faster speeds, the spin-up time approaches 10 to 12 seconds. In one embodiment, the spin-up time is inferred by determining the time it takes to load a page into system memory from the disk drive that is not spinning (i.e., not being accessed). The spin-up time is approximately equal to the load time because the load time is dominated by the spin-up time.

[0044] When the computer system is to be powered down (step 502), the static configuration data and dynamic configuration data to be stored in NV memory is determined (step 504). The static configuration data and dynamic configuration data is selected so that the time of initializing the static configuration data and dynamic configuration data is approximately equal to the spin-up time of the disk drive. The static and dynamic configuration data is configuration data that does not need data from the disk drive to be initialized. It may include configuration data for keyboard devices, pointing devices, the USB subsystem, the video subsystem, and the like. The size of the NV memory required is equal to the size of the static configuration data and dynamic configuration data. In many operating systems, the size of the NV memory needed ranges from 16 MB to 64 MB. 128 MB of NV memory may be needed to reduce the cost of CPU compression in some operating systems.

[0045] The static configuration data is system configuration data that is used for all

users. For example, in the Windows™ operating system, the static configuration data includes boot strapping code to load the static and dynamic configuration data stored on NV memory and the configuration data stored on the disk drive. Other static configuration data in the Windows™ operating system includes files necessary to initialize the kernel such as registry initialization (smss.exe), video initialization (csrss.exe), the shell, services that run from dynamic link libraries (svchost.exe), security services (lsass.exe) such as IPsec and network login. For a computer system that is powering down, the static configuration data that is stored in NV memory includes the boot strapping code, kernel initialization files, the video initialization files, the shell, and services (e.g., dynamic link libraries) that are needed while the disk drive is spinning up.

[0046] The dynamic configuration data is system configuration data that is on a per-user basis. For example, dynamic configuration data includes desktop settings (e.g., background image, icon location, etc.), security credentials, internet favorite files, fonts, start-up programs, keyboard settings, personal data files such as scheduled tasks, financial information, and the like. For a computer system that is powering down, the dynamic configuration data that is stored in NV memory includes the desktop settings, keyboard settings, and security credentials

[0047] The selected static configuration data and dynamic configuration data is stored in the NV memory 200 (step 506). The remaining configuration data is stored on the disk drive (step 508). In one embodiment, the amount of NV memory is larger than the minimum size required. Power down of the system is decreased by storing data that is being sent to the disk for storage that the operating system can not control such as an application writing to the disk to the NV memory 200 for storage (step 510). This reduces

the number of writes to the disk, which results in a faster power down time. The computer system is then powered down.

[0048] If the system is going into hibernation (step 514), similar steps occur. Static configuration data and dynamic configuration data to be stored in NV memory is determined (step 516) and stored in memory (step 518). This data is the data that is necessary to "repaint" (i.e., restore) the graphical user interface quickly. The static configuration data includes kernel initialization files, the video initialization files, the shell and services that run from dynamic link libraries that were in use prior to hibernation. The dynamic configuration data includes desktop settings, keyboard settings, fonts and the personal data files that were being used prior to hibernation.

[0049] The remaining configuration data is stored on the disk drive (step 520). Writes received from applications are stored in available NV memory first and stored on the disk drive if the NV memory is not available (step 522). The computer system then enters hibernate mode (step 524).

[0050] In one embodiment, the NV memory is used during system operation. If an application or the operating system is sending data to the disk for storage (step 526), the data is stored in the NV memory 200 (step 528). When the system is being powered down or entering hibernation, the data stored in NV memory 200 is transferred to the disk drive. The inline device may receive new commands that are used to control hybrid disks that include a NV memory cache. These new commands will be translated by the inline device controller 224 to legacy commands that the hard disk drive controller 310 can interpret (step 530). If the NV memory available space falls below a threshold (step 532), the data

in the NV memory is transferred to the disk drive (step 534). One reason for transferring data to the disk drive is to make room for configuration data that is required to resume operation from hibernate or to boot the system. Storing the data in NV memory decreases the number of times the disk drive is spun-up during operation. This improves a user's experience with computing devices, and pocket sized devices in particular, because power consumption is reduced, which subsequently reduces the requirement for a noisy high speed fan. Additionally, the noise of the disk drive spinning and writing data to disk is not heard as often.

[0051] FIG. 7 illustrates the steps taken to boot the computer system or restore the computer system from hibernate mode. When the computer system boots from the disk drive or wakes from the hibernate mode, the computer system initializes the system DRAM by loading it with the static configuration data and dynamic configuration data stored in the NV memory 200 (step 700). If all required initialization data is not stored in NV Memory and additional information must be loaded from the disk, configuration data is first loaded from NV memory and during that time, the disk drive is spinning up (step 702). This effectively eliminates the spin-up time of the disk drive from the boot time and the wake time. Once the disk drive has spun up, the remaining configuration data that is stored on the disk drive is loaded to initialize the system DRAM (step 704). The experience the user sees is that when the user hits the power button, she should see a responsive interactive logon screen by the time of or shortly after the time that the monitor 191 or other type of display device turns on instead of having to wait for the disk drive to spin up. During the time the user clicks on her user icon and types in her password, the services needed to display the desktop are initialized. A responsive and usable desktop is shown to the user after the disk spins up and the data in the NV is verified to be consistent with the data on

the disk drive. In one embodiment, all of the cached data is not compared. However, if there are mismatches within the partition or file system metadata, the computer system needs to bug-check and enter a service / help mode such that a slow boot can be attempted at next boot. A help button or switch can be used to bypass the NV memory and turn on the computer system in safe mode for recovery and backward compatibility purposes.

[0052] In an alternate embodiment, the NV memory 200 is exposed to the operating system so that the operating system and other components can query the size of the NV memory 200, control the state of the NV memory 200, write data to the NV memory 200 using a transactional procedure, and use the NV memory 200 as a write buffer. This embodiment may be implemented in a variety of ways. The NV memory may be located in the device 220 as illustrated in FIG. 3. If data is cached in NV memory and is also on a device strapped to the disk, the processor operating system uses the data cached in NV memory 200 instead of data from the disk.

[0053] FIG. 8 shows the storage software stack of an operating system interacting with hardware in one embodiment of the implementation where the inline device 220 makes the hard disk drive look like a hybrid drive, containing spinning magnetic media and solid state non-volatile memory and responding to a new set of commands defined for hybrid disks. A predictive volatile read cache control module 802 predicts the data that will be required in the coming interval of time and reads that data from the hard disk drive 141 into volatile system memory 130, thus minimizing reads from the disk. This allows the disk platter to be spun down for long periods of time when combined with the write buffering capability of the inline NV cache device 220.

[0054] Turning now to FIG. 9 an alternative embodiment of the invention where the operating system and drivers control the non-volatile memory is illustrated. The storage driver stack 800 of an operating system is shown. In the description that follows, the hard drive 300 and inline NV memory device 220 are shown as in FIG. 3. The operating system 134, application programs 135, other program modules 136 send data (e.g., documents, spreadsheets, files, graphics and the like) to the storage driver stack 800. The predictive cache control module 802 controls what is in the volatile memory 132 and hard drive 300 and NV memory 200. The file system filter driver 804 receives data from the predictive cache control module 802. The principal function of file system filter driver 804 is to gracefully handle errors such as when the NV memory cache 200 is full. The file system filter driver 804 passes the data through the file system (i.e., NTFS) 806, which the operating system 134 uses for storing and retrieving files from a hard disk in the hard drive 300.

[0055] The write cache filter driver 808 receives data from the file system 806 and the predictive cache control module 802. The write cache filter driver 808 sends data to be buffered in the NV memory 200 to the NVM driver 224 (e.g., controller 224). The NVM driver 224 detects when the NV memory 200 no longer has capacity (i.e., it is full) and sends a message to file system filter driver 804 to indicate the NV memory 200 is full. In response, the file system filter driver 804 informs the predictive cache control module 802 to stop sending data to the NV memory 200 until the NV memory 200 has storage capacity. Additionally, the predictive cache control module 802 may provide the NVM driver 224 a list of data block to flush from the NV memory 200 to the hard drive 300.

[0056] The write cache filter driver 200 synchronizes the NV memory 200 and hard

drive 300. If data is stored in the NV memory 200 and the hard drive 300, the write cache filter driver 808 reads from the NV memory 200 when it needs to synchronize the write cache in NV memory 200 with the hard drive 300. Data is transactionally written to the NV memory 200 and from the NV memory 200 to hard drive 300. This allows the system 100 to seamlessly continue working. When a write is completed, a transaction completion message is sent to the write cache driver 808. Transaction writing allows a write to be retried if the write fails or gets interrupted (i.e., no completion message is received). If the system reboots due to a loss of power or any other reason during a transaction between the NV memory 200 to the hard drive 300, the write cache filter driver 808 synchronizes the hard drive 300 with the data stored in the NV memory 200 after the system is running again. For a system where the hard drive has NV memory, the controller in the hard drive synchronizes the data.

[0057] The storage driver 810 receives data to be sent to the hard drive 300. The storage driver 810 sends data in blocks. It may be iSCSI or equivalent. iSCSI enables block storage transfer. It is a protocol that encapsulates SCSI commands in TCP/IP enabling high performance storage area networks (SANs) over standard networking technology. The storage driver 810 converts the data into data blocks (SCSI-3 data if iSCSI is used) and sends the data blocks to the hard drive 300.

[0058] As previously indicated, the predictive cache control module 802 determines what is stored in the NV memory 200. In general, the predictive cache control module 802 looks at a user's behavior over an extended period of time and determines what programs and data the user is accessing frequently and stores the data, which includes static and dynamic configuration data in the NV memory 200. Data as used herein includes personal



data files such as scheduled tasks, financial information, data files, and the like. The module 802 starts with defaults until the behavior of a user can be predicted. The static and dynamic configuration data is based on a history of prior use and comprises data that is frequently required by the user. The static configuration data is data that does not change periodically. The dynamic configuration data is data that changes periodically and is on a per-user basis. The configuration data may include keyboard devices, pointing devices, the USB subsystem, the video subsystem, output peripheral drivers (e.g., printer driver) the operating system, application programs, data files and the like.

[0059] The NVM driver 812 and storage driver 810 are combined into a single module 1000 (see FIG. 10) in the in the case where the NV memory 200 is attached to the hard drive 226, consisting of the device 220 and hard drive 300. The hard drive driver 1000 communicates with the controller 224 and hard drive controller 310 to determine the disk's capabilities and the disk's state, to specify which blocks to place in NV memory and the priority of those blocks, which blocks to preserve indefinitely in NV memory, when to flush the NV memory 200 to disk 304, and when to commit a block of write operations to the disk 304. This improves performance for the user by having blocks in NV memory 200 that can be quickly loaded into system memory. For example, the operating system 134 could request that the hard drive 226 places in NV memory 200 the files that the operating system 134 needs to boot the computer 110. The next time the computer system is powered up, after BIOS POST (power on self test) finishes, the operating system 134 can boot from NV memory 200 without waiting for the disk 304 to spin up. The NV memory cache 200 may contain data that the user has identified via an interface to keep instantly available, such as the user's Contacts list in Microsoft Outlook®. The user could power on his or her computer, retrieve contact information, and power off again in the time it would take to

boot a computer that doesn't have NV memory used as a storage medium.

[0060] The NV memory cache 200 consists of an array of cached memory blocks (e.g., an array of cached disk blocks, pages, sectors, files, and the like) with each block capable of being pinned. In one embodiment, each block has a tag and a pin bit associated with it to indicate whether the block is pinned. Non-pinned blocks contains data that can be prioritized by the operating system 134 but are otherwise in control of the disk controller 310. The pinned blocks are blocks in the array that are controlled by the operating system 134 and cannot be moved out of NV memory by the controller 310 unless the operating system 134 gives permission to move a pinned block out of NV memory. Data in a pinned block persists in NV memory until the operating system unpins it. The memory may be pinned in such a way that it persists between boots of the computer system. The data in pinned blocks typically includes static configuration data and dynamic configuration data.

[0061] The storage device 226 (the combination of hard disk 300 and device 220) reports its NVM capabilities and the current state of the spinning media to the operating system 134. The capabilities and current state includes information such as size of the NV memory cache 200, features supported by the NV memory (e.g., whether or not it can "pin" blocks), and the state of the disk 304 (e.g., stopped, spinning, spinning-up, or spinning-down).

[0062] The operating system 134 communicates with the hard disk controller and device controller 224 to provide information and commands to the controller 310. The information includes a list of blocks to pin in the NV memory 200 such that they persist in the cache between boots, a list of blocks to "pre-cache" in the NV memory 200 while the

user is working, and the list contains the priority of pre-caching the blocks. The operating system predicts the next set of blocks that will be write-accessed, based on usage scenarios and provides a list of the next set of blocks to the NV memory 200 for pre-caching. The priority of pre-caching the blocks includes whether the pre-caching should be done immediately, including spinning up the disk 304 if necessary, or can be done the next time the disk 304 is spinning.

[0063] The information also includes a list of blocks to flush from the NV memory 200 to disk 304, an instruction to flush all of the NV memory 200 to disk 304, an instruction to flush the pinned blocks in NV memory 200 to disk 304, an instruction to flush the dynamic (i.e., non-pinned) blocks in NV memory 200 to disk 304, a designation of a set of blocks or segments that represent a write transaction (these blocks are written to a temporary location in NV memory and not stored permanently until a commit command is received from the operating system 134), and an instruction to commit the last write transaction to permanent storage in NV memory 200.

[0064] Turning now to FIG. 11, during operation, the operating system 134 uses commands to determine the size of the NV memory cache 200, features supported by the NV memory and the state of the disk (step 1100). The operating system 134 changes operation modes based on the size of NV memory. If the size is sufficient to boot from the NV memory without having to use the disk 304, the operating system commands the hard drive to move the required files to boot to the NV memory prior to shutting down the system or before entering hibernation (step 1102). In addition to boot files, the operating system 134 determines via the predictive cache control module the data the user typically uses at start-up and commands the hard drive to move those files to NV memory if there is

sufficient available space in the NV memory prior to shutting down or entering hibernation.

If the device 220 has the capability to pin blocks into NV memory, the operating system 134 determines what should be pinned in NV memory and sends the hard drive the list of blocks that should be pinned (step 1104). The operating system 134 may periodically send this list to the controller 310 as conditions change. For example, if the user hasn't used a particular program recently that was heavily used in the past, the program is moved off the list and the program is no longer pinned in memory.

[0065] The predictive cache module 802 predicts what data the user will need based on usage scenarios (step 1106). A list of data is sent to the hard drive 300 (step 1108). These data are cached in NV memory in accordance with the priority set by the operating system 134 based upon what the user is doing. For example, if a user has started a particular application, such as Microsoft Word®, the list provided to the controller 310 contains a list of files that will be needed soon.

[0066] The available space of the NV memory 200 is monitored. When the available space falls below a threshold, the NV memory is flushed (step 1110). The operating system 134 may instruct the controller 224 to flush all data in NV memory 200 to disk 304, flush pinned blocks to disk 304, flush dynamic blocks to disk 304, or a combination of the above. The operating system 134 also sends flush commands when usage changes. The flushing can be done in the background at a time when the disk 304 is already spinning to handle a read that isn't in the NV memory 200 or the disk 304 can be spun up to do it.

[0067] When an application sends data to be stored in the hard drive 300, the operating system command the controller 224 to store the data in data blocks in a temporary location

in NV memory 300 (step 1112). These data blocks represent a write transaction. The storage of data in NV memory 200 is done to reduce the number of times the hard drive is spun up, which results in reduced power consumption. Additionally, the predictive cache module 800 may indicate that this data may be used soon based on usage scenarios. The data blocks are stored in NV memory until the operating system 134 sends an instruction to commit the write transaction to permanent storage in the disk 304. The temporary storage of data allows the operating system to operate in block storage mode and manage the storage of data blocks in accordance with operating system needs. Data to be sent to the operating system that has not been cached in NV memory is sent to the operating system directly from the disk drive.

[0068] It can be seen that a method and apparatus to reduce power consumption in a hard drive has been provided by reducing the number of times the disk in the hard drive spins. Additionally, the pre-caching of data in NV memory enhances a user's experience. The data that is stored in NV memory appears almost instantly to the user when compared to loading data from a rotating disk.

[0069] In view of the many possible embodiments to which the principles of this invention may be applied, it should be recognized that the embodiment described herein with respect to the drawing figures is meant to be illustrative only and should not be taken as limiting the scope of invention. For example, those of skill in the art will recognize that the elements of the illustrated embodiment shown in software may be implemented in hardware and vice versa or that the illustrated embodiment can be modified in arrangement and detail without departing from the spirit of the invention. Additionally, the steps described above can be performed sequentially, in parallel, in a different order, or a

combination of sequential and parallel. Therefore, the invention as described herein contemplates all such embodiments as may come within the scope of the following claims and equivalents thereof.

## WE CLAIM:

1. An in-line persistent storage device comprising:  
a housing adapted to be connected to a disk drive connector of a disk drive;  
an array of solid state, non-volatile ("NV") memory blocks mounted within the housing and in communication with the disk drive and a controller, the controller in communication with the disk drive, the controller for controlling the NV memory and adapted to store at least one data block received from a computing device in at least one memory block in NV memory and flush the at least one block of data in response to receiving a flush command, the array of solid state NV memory blocks.
2. The in-line persistent storage device of claim 1 wherein the controller is within the housing.
3. The in-line persistent storage device of claim 1 further comprising a volatile memory buffer faster than the NV memory and an energy storage device.
4. The in-line persistent storage device of claim 3 wherein the energy storage device is sized such that there is sufficient power to keep the volatile buffer and the NV memory buffer powered long enough to move data from at least one of the volatile memory and the array of NV memory to the disk drive in the event of sudden power failure.
5. The in-line persistent storage device of claim 3 wherein the energy storage device is a battery or a capacitor.

6. The in-line persistent storage device of claim 1 further comprising a pair of connectors, one of the pair of connectors connecting to the disk drive connector and the other of the pair of connectors connecting to a cable connected to a motherboard.

7. The in-line persistent storage device of claim 1 wherein the housing is integrally attached to the disk drive.

8. The in-line persistent storage device of claim 1 further comprising a pair of power connectors, one of the pair of power connectors connecting to a disk drive power connector and the other of the pair of connectors connecting to a power cable.

9. The in-line persistent storage device of claim 1 wherein the controller intercepts commands from an operating system and buffers writes to the NVM memory.

10. The in-line persistent storage device of claim 1 wherein the controller is further adapted to load at least one block of data in NV memory from the disk drive in response to receiving a list of data to be pre-cached

11. The in-line persistent storage device of claim 1 wherein the controller is further adapted to pin a block of data in the NV memory.

12. The in-line persistent storage device of claim 1 wherein the controller is further adapted to immediately command the disk drive to spin up to load the at least one block of data in NV memory.



13. The in-line persistent storage device of claim 1 wherein each memory block is one of a pinned data block and a non-pinned data block and the controller is further adapted to flush pinned data blocks in response to receiving a command to flush pinned data blocks.

14. The in-line persistent storage device of claim 1 wherein each memory block is one of a pinned data block and a non-pinned data block and the controller is further adapted to flush non-pinned data blocks in response to receiving a command to flush a dynamic cache.

15. The in-line persistent storage device of claim 1 wherein each memory block is one of a pinned data block and a non-pinned data block and the controller is further adapted to flush pinned data blocks and non-pinned data blocks in response to receiving a command to flush all data blocks.

16. The in-line persistent storage device of claim 1 wherein the controller is further adapted to provide capabilities to the computer system in response to receiving a request to provide capabilities.

17. The in-line persistent storage device of claim 16 wherein the controller provides a size of the NV memory, features supported by the NV memory, and the state of the disk drive when the controller provides capabilities to the computer system,

18. The in-line persistent storage device of claim 1 wherein the controller is

further adapted to:

determine when the NV memory available space is below a threshold; and  
flush at least a portion of the NV memory to the disk drive if the NV memory available space is below the threshold.

19. The in-line persistent storage device of claim 1 wherein the controller is further adapted to intercept commands directed to the disk drive, interpret the commands, and pass the commands on to the disk drive.

20. The inline persistent storage device of claim 1 wherein the controller is further adapted to receive commands from the operating system, translates commands, and passes them on to the HDD.

21. The inline storage device of claim 1 wherein the inline storage device has straps or adhesive associated with it that allow it to be fixed to the disk drive.

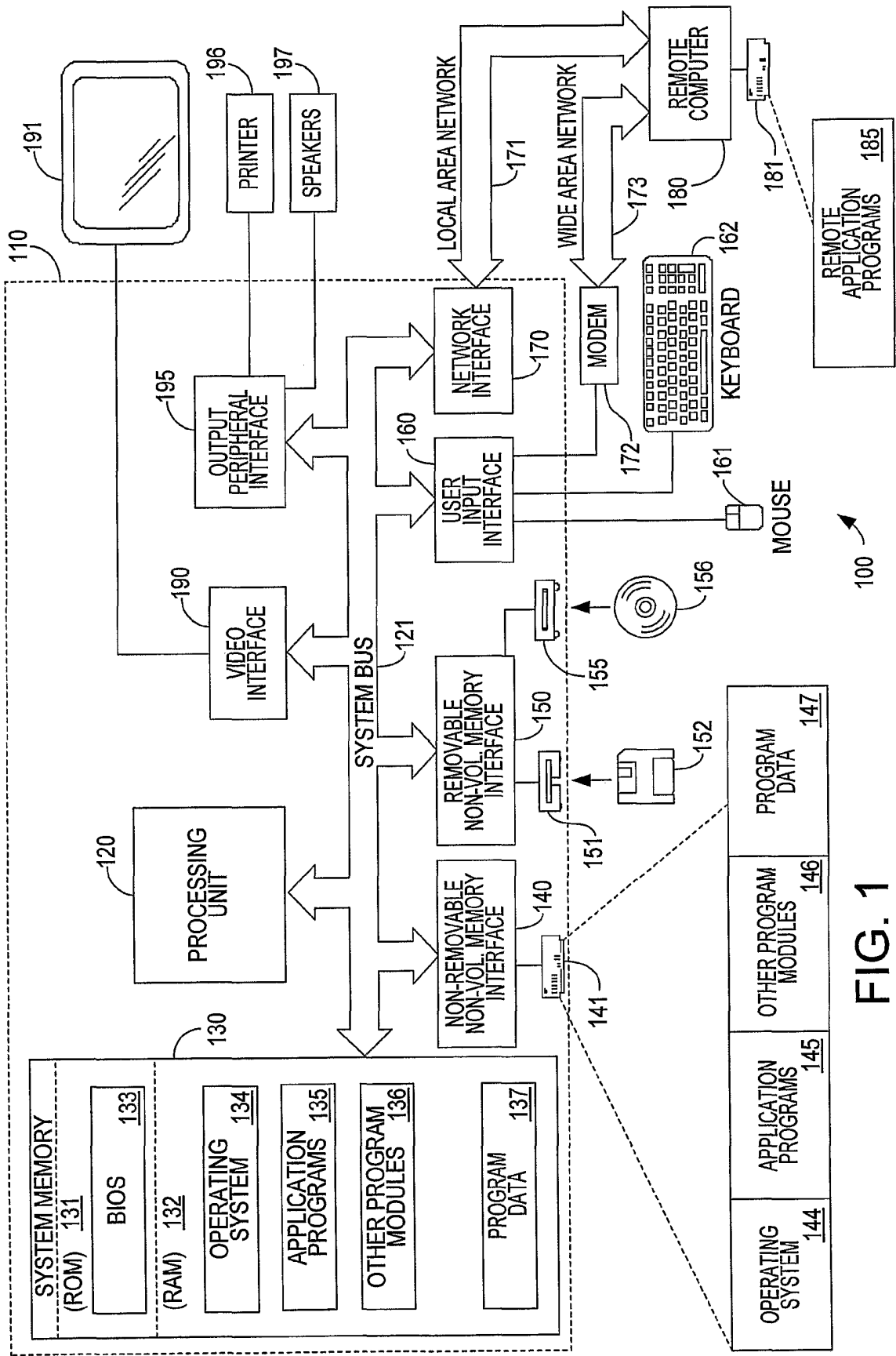


FIG. 1

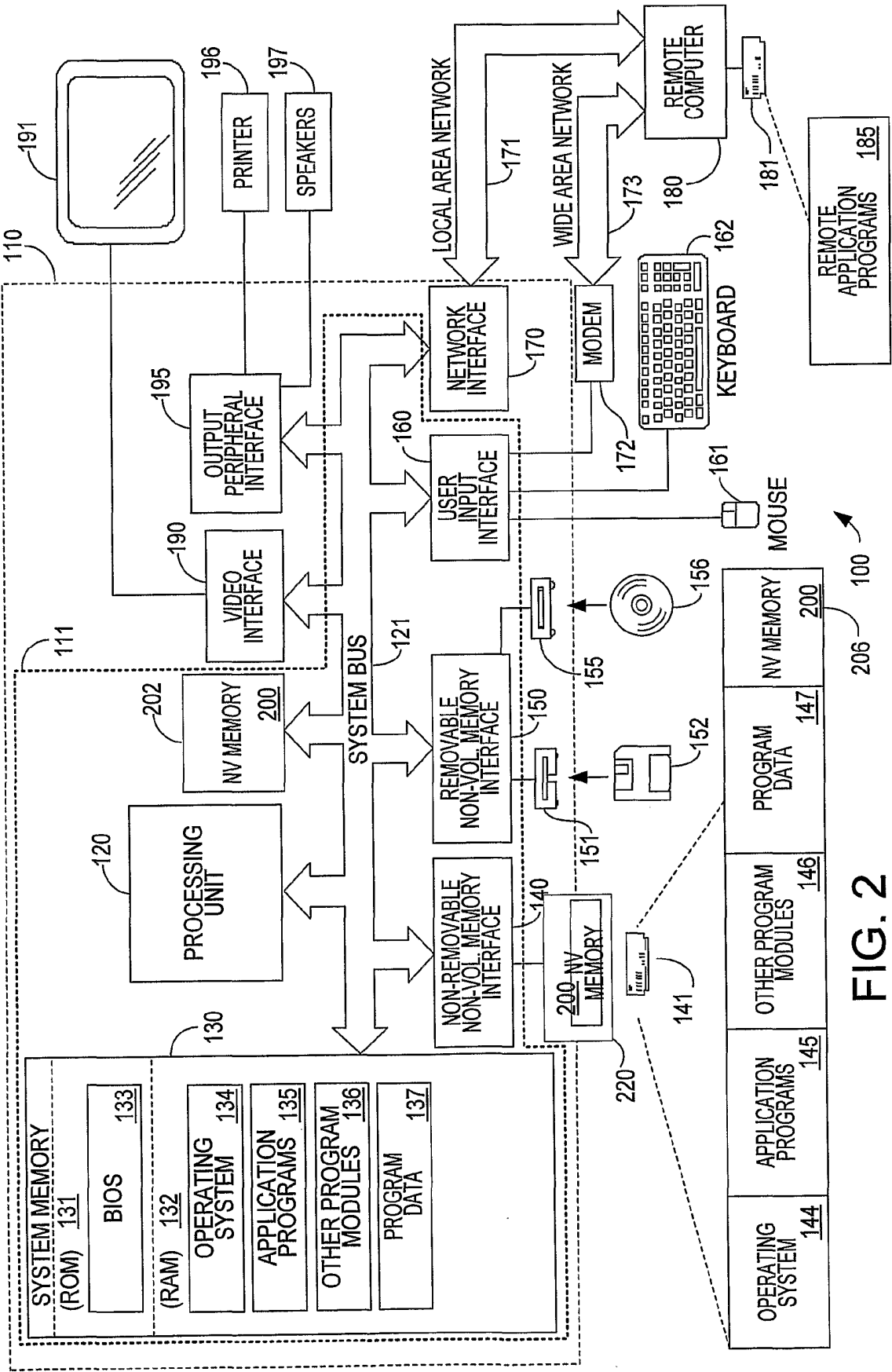
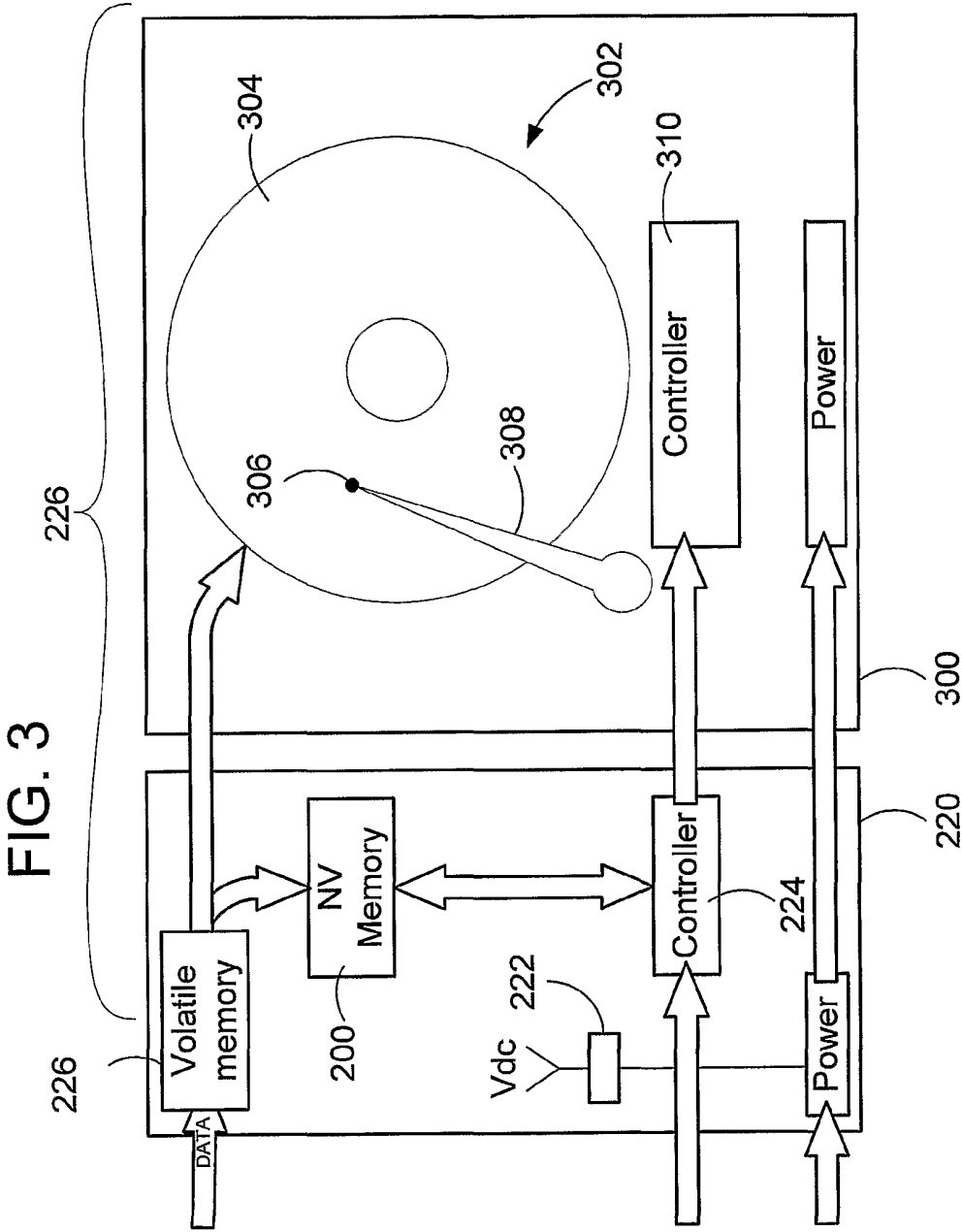
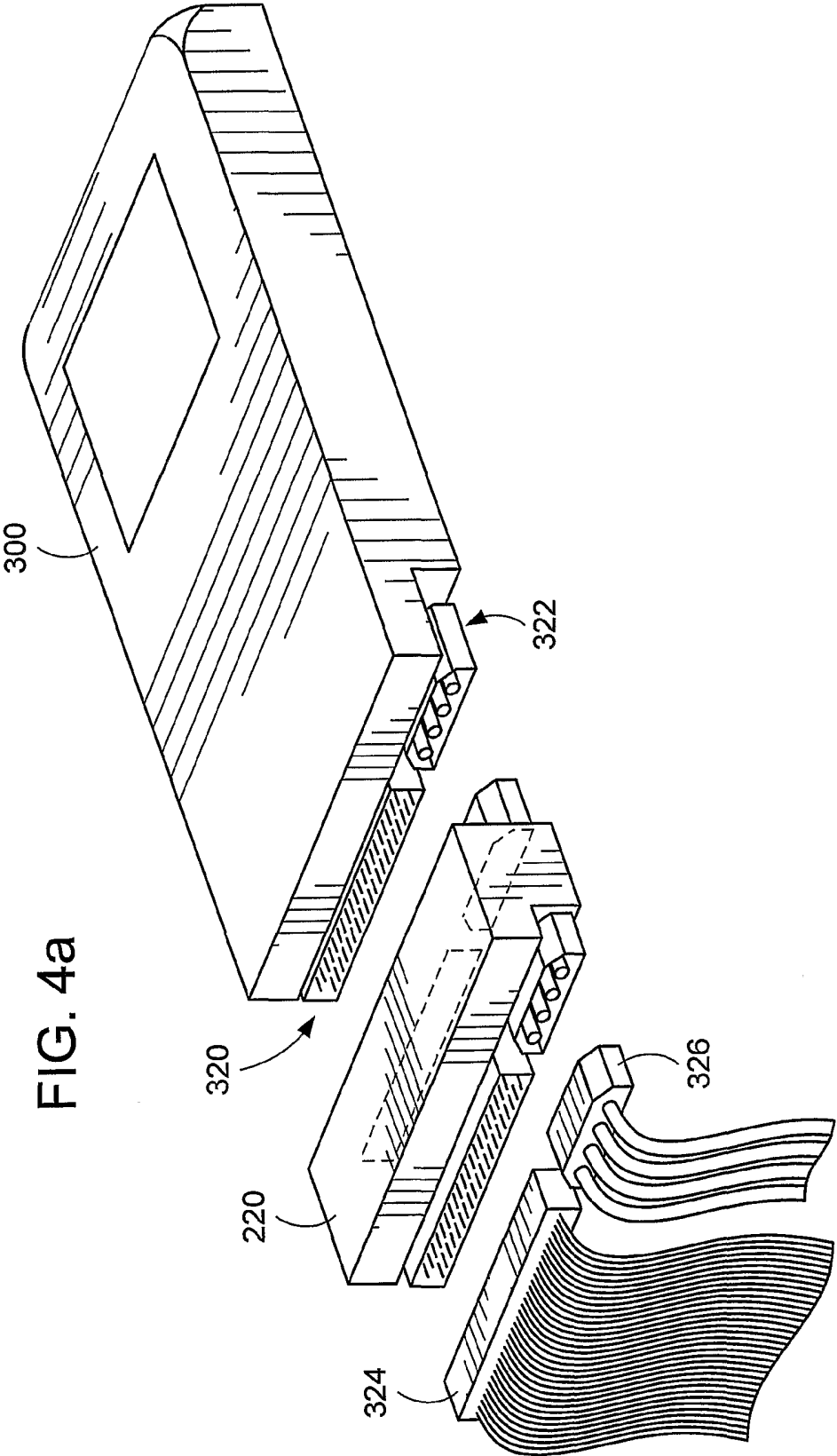


FIG. 2





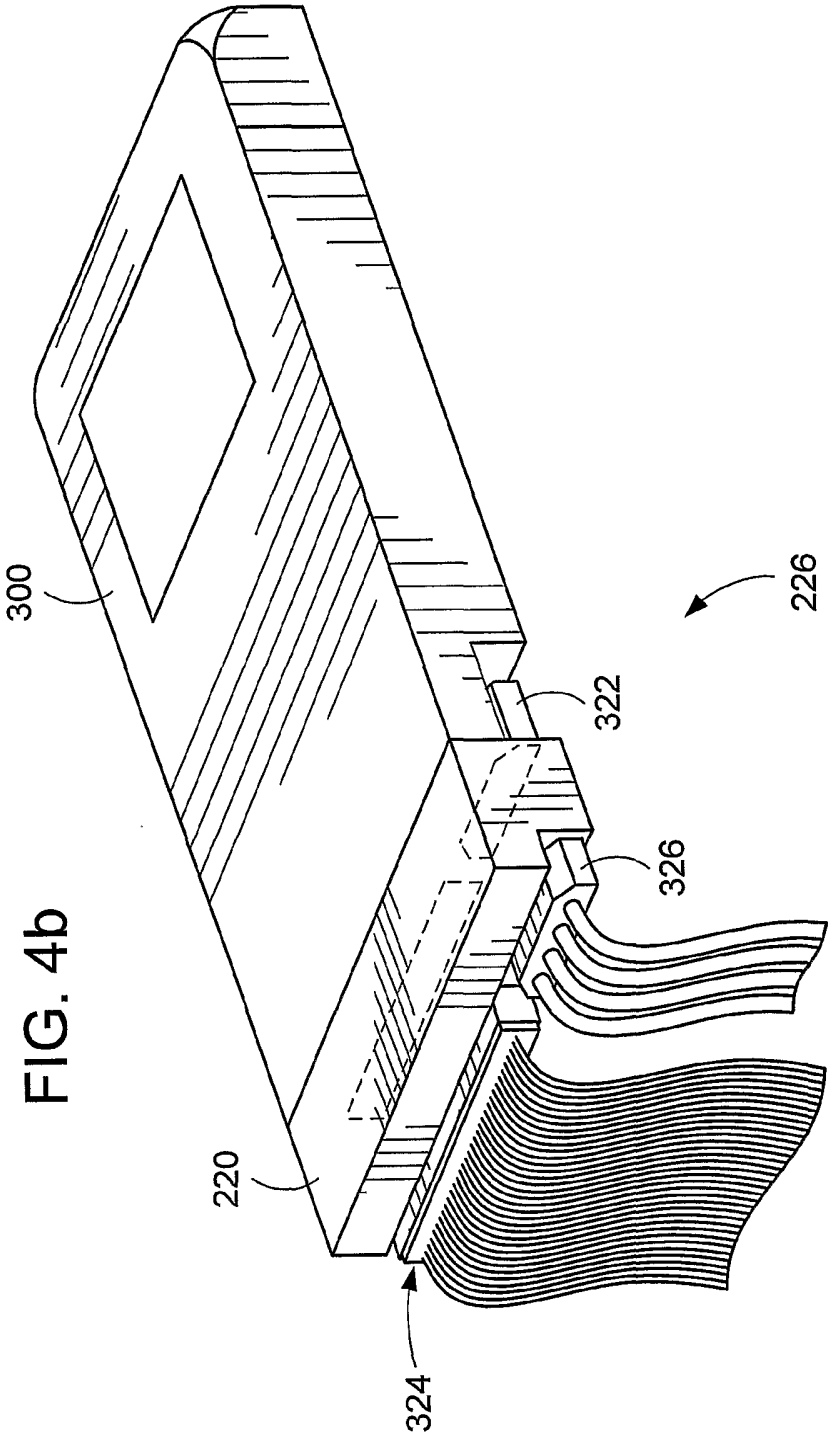


FIG. 5

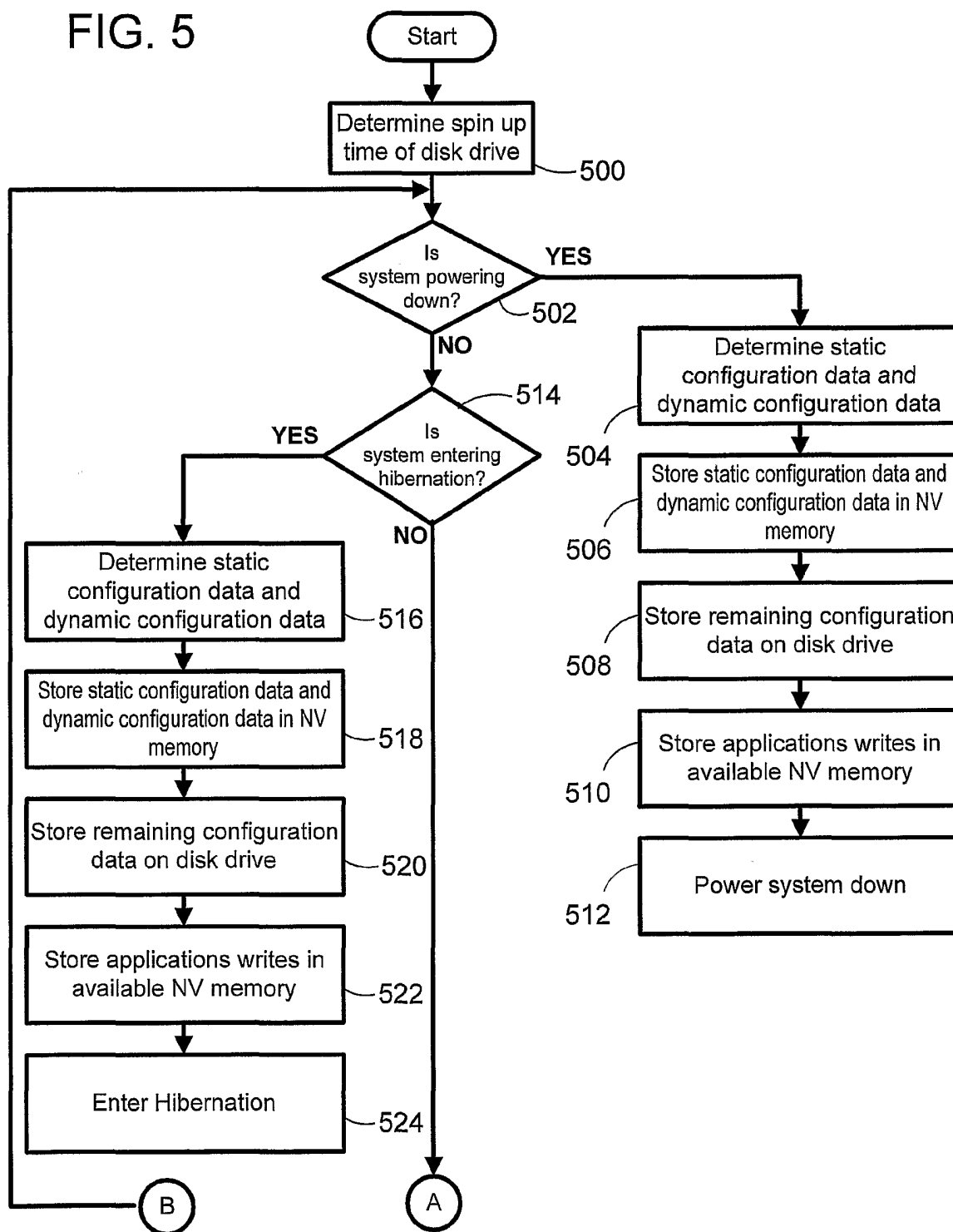




FIG. 6

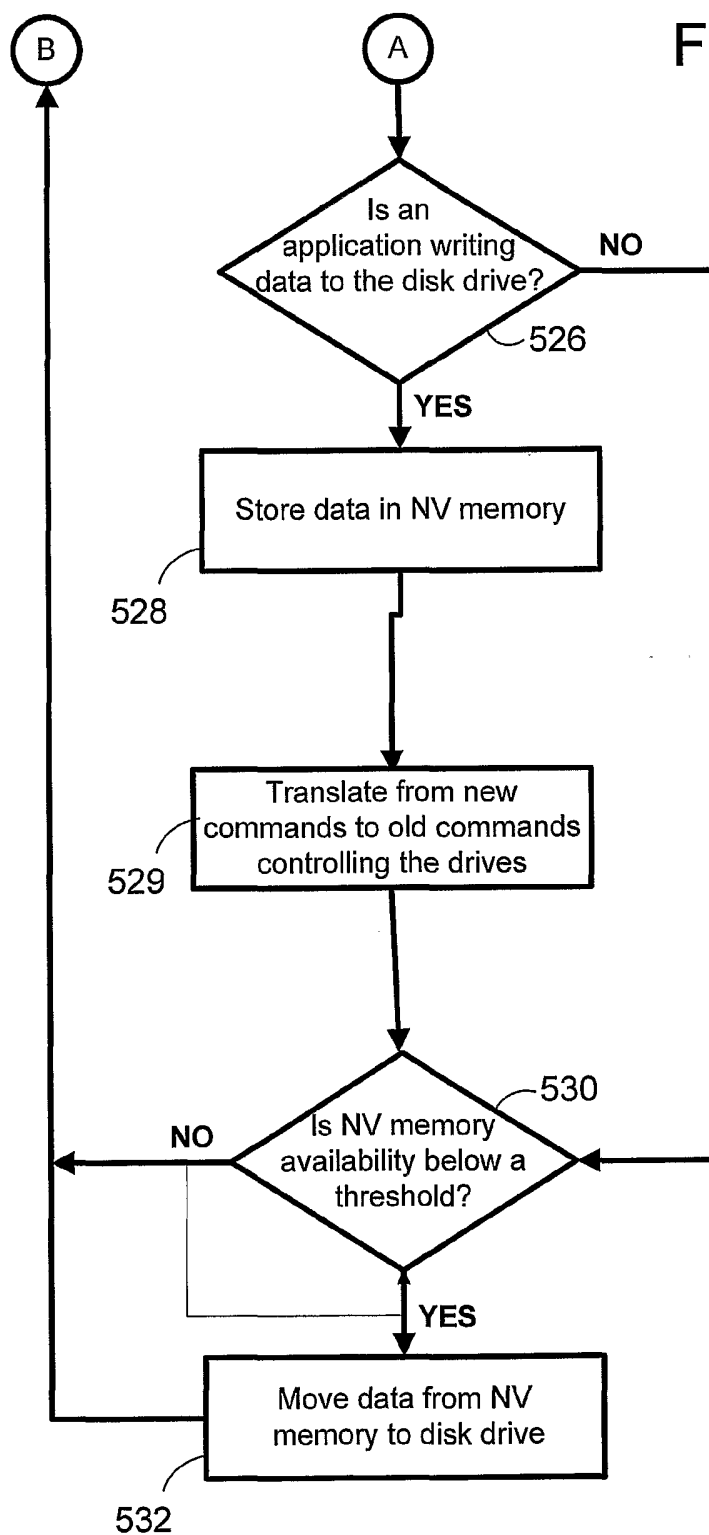


FIG. 7

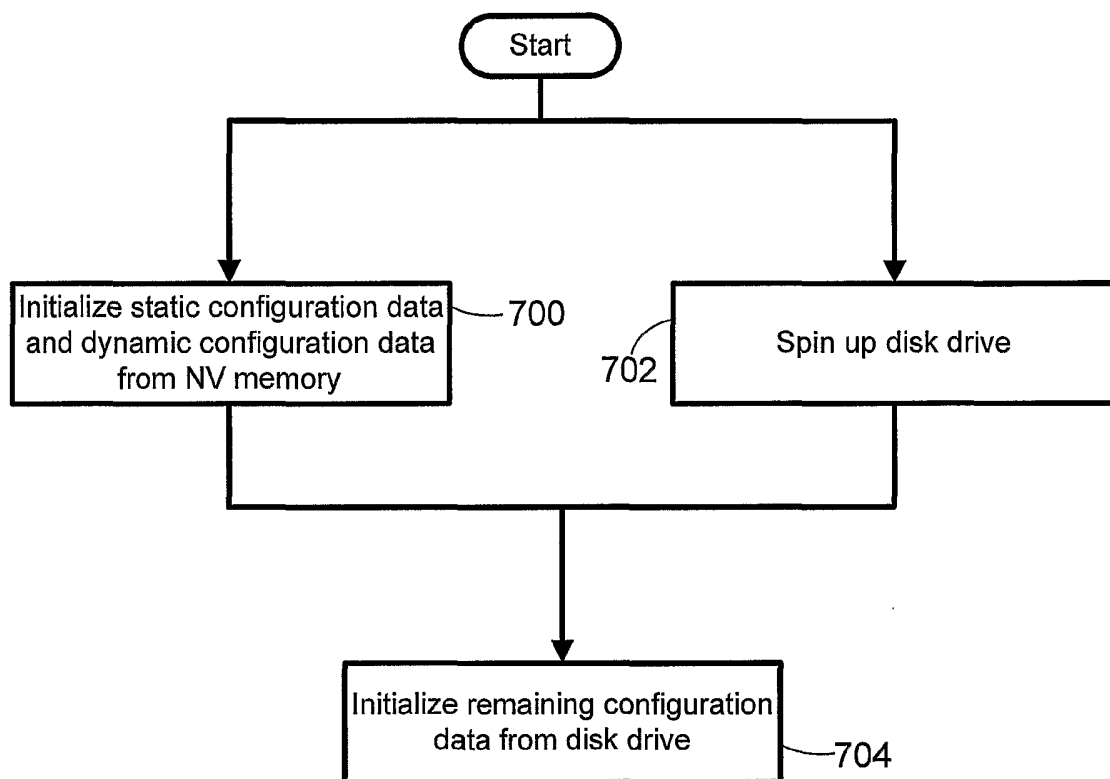
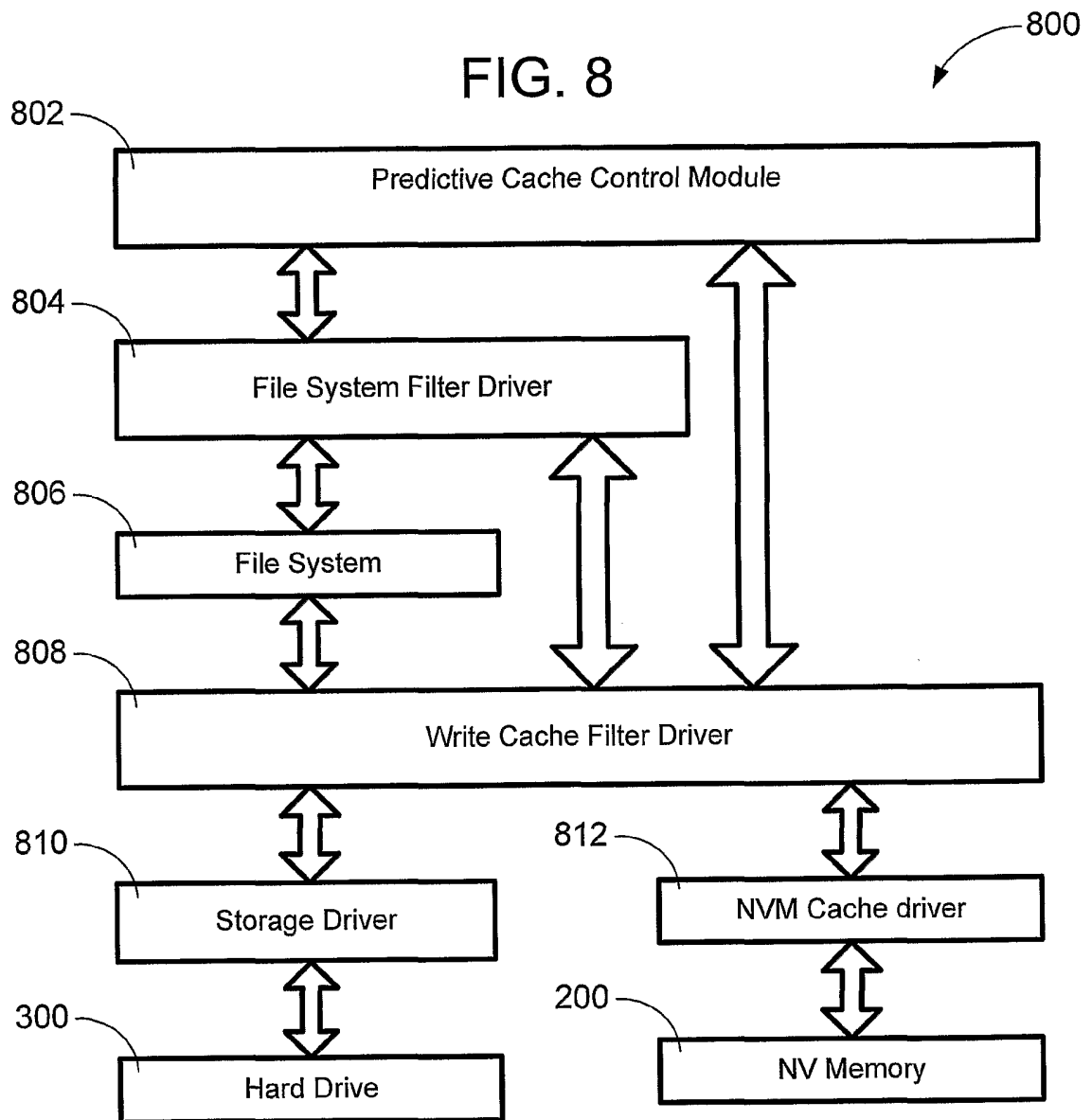


FIG. 8



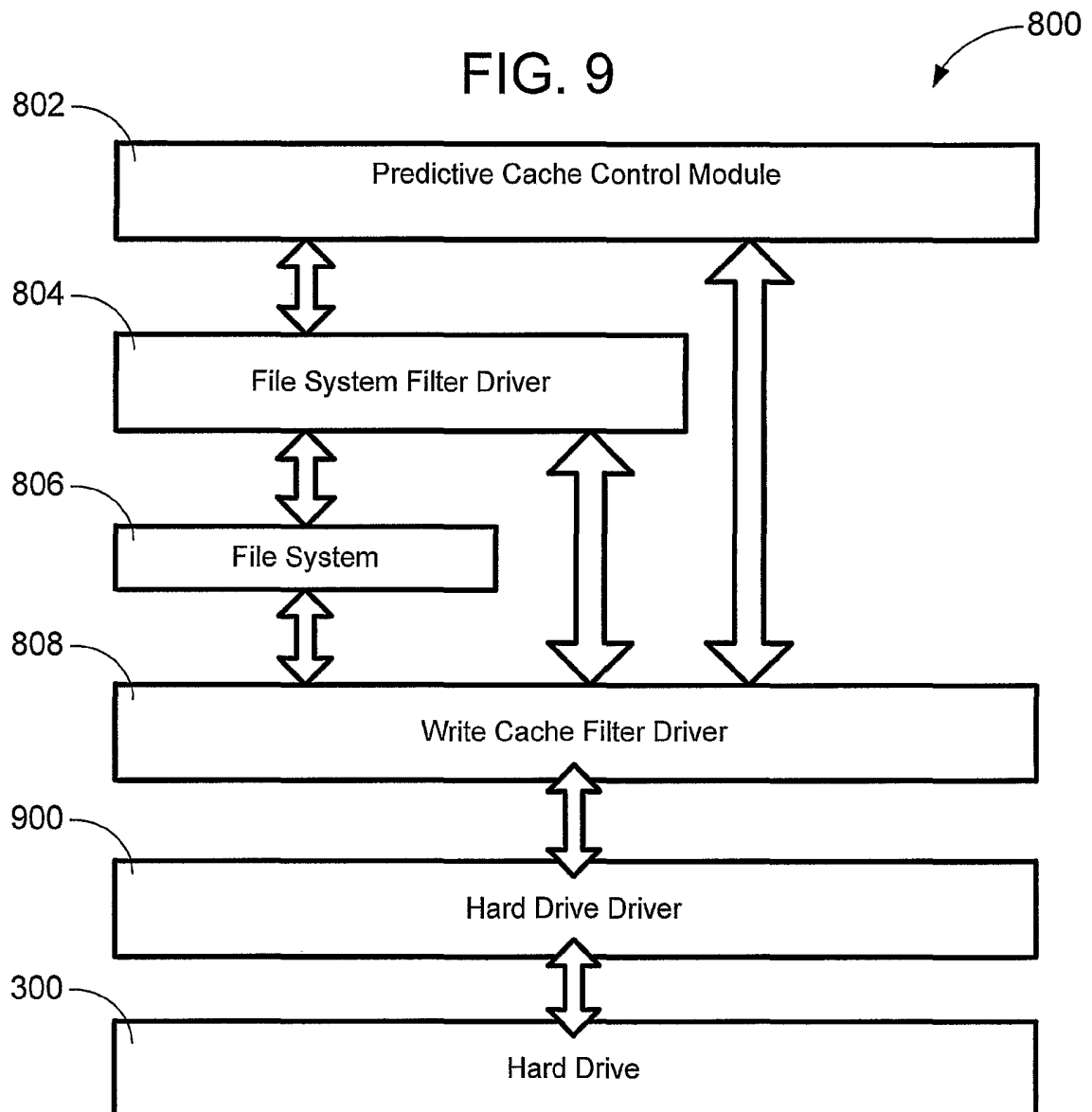


FIG. 10

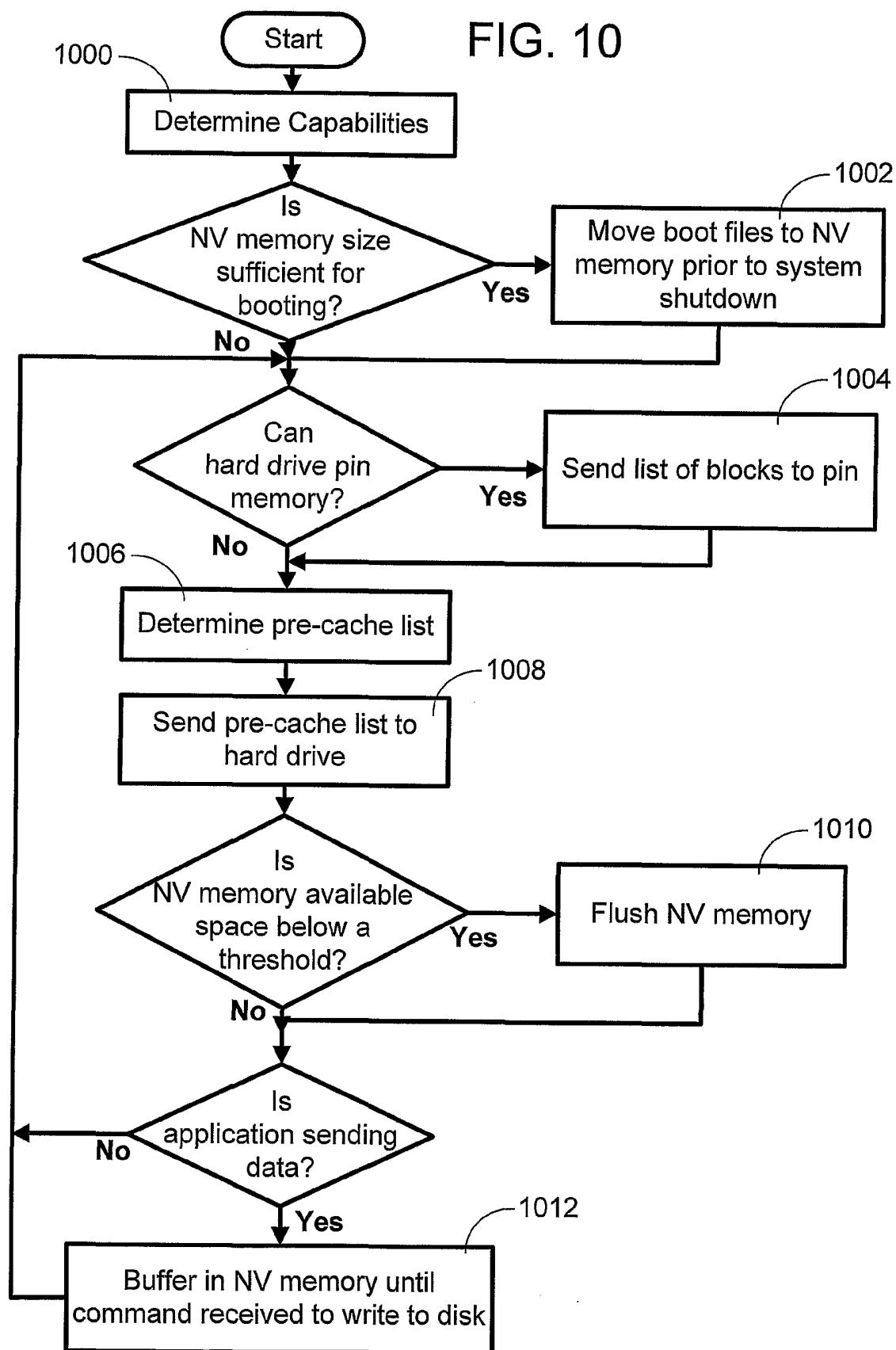


FIG. 11

