



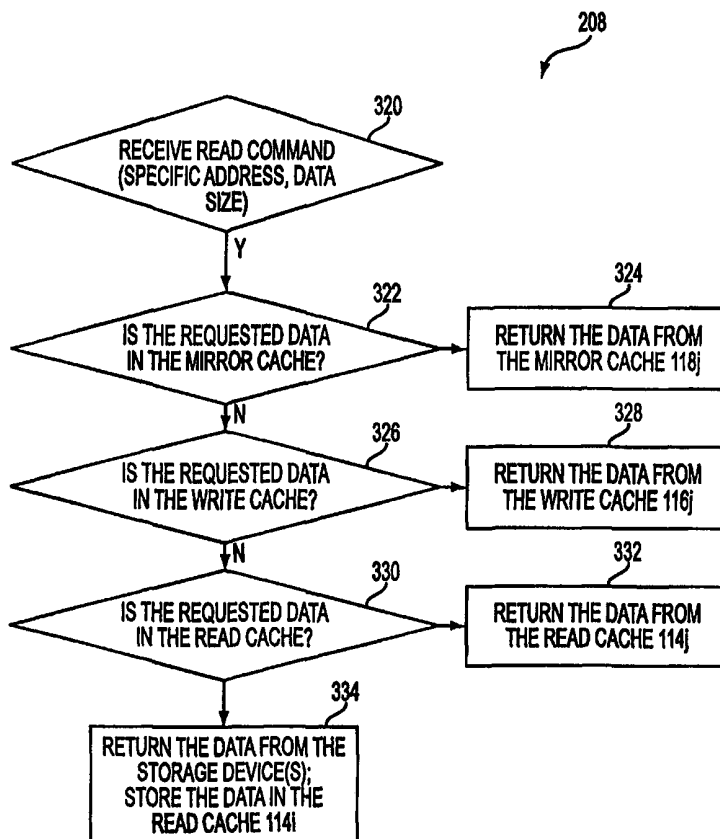
INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : G06F 12/08		A1	(11) International Publication Number: WO 00/43888
			(43) International Publication Date: 27 July 2000 (27.07.00)
(21) International Application Number: PCT/US00/01921 (22) International Filing Date: 25 January 2000 (25.01.00) (30) Priority Data: 09/236,504 25 January 1999 (25.01.99) US (71) Applicant: MYLEX CORPORATION [US/US]; 34551 Ardenwood Boulevard, Fremont, CA 94555-3607 (US). (72) Inventor: HUBIS, Walter, A.; 2022 Centennial Drive, Louisville, CO 80027 (US). (74) Agents: ANANIAN, R., Michael et al.; Flehr Hohbach Test Albritton & Herbert LLP, Suite 3400, 4 Embarcadero Center, San Francisco, CA 94111-4187 (US).		(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). Published <i>With international search report.</i>	

(54) Title: FULL CACHE COHERENCY ACROSS MULTIPLE RAID CONTROLLERS

(57) Abstract

A method for providing cache coherency in a RAID system (100) in which multiple RAID controllers (104) provide read/write access to shared storage devices (108) for multiple host computers (102). Each controller includes read (114), write (116) and write mirror (118) caches and the controllers and the shared storage devices are coupled to one another via common backend buses (110). Whenever a controller receives a write command (302) from a host the controller writes the data to the shared devices, its write cache and the write mirror caches of the other controllers. Whenever a controller receives a read command (320) from a host the controller attempts to return the requested data from its write mirror cache, write cache and read cache and the storage devices, in that order.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

FULL CACHE COHERENCY ACROSS MULTIPLE RAID CONTROLLERS

The present invention relates generally to methods for caching disk reads and writes in a RAID system and, particularly, to methods for maintaining the coherency of multiple caches used for caching disk reads and disk writes in a RAID system.

5

BACKGROUND OF THE INVENTION

FIG. 1 shows a block diagram of a typical multiple-controller RAID system 100 (RAID is an acronym for "Redundant Array of Independent Disks"). Each host computer 102 is connected to a respective RAID controller 104 through either a Fibre Channel or SCSI bus 106 via a host bus adapter (HBA). Each RAID controller 104 coordinates reading and writing requests from a respective host 102 directed to a shared set of storage devices 108 to which the RAID controllers 104 are connected via a backend Fibre Channel or SCSI disk bus 110. The controllers 104 use the same storage devices 108 so that each host computer 102 can access the same data. FIG. 1 shows only two controllers; however, the illustrated architecture is extendable to systems of N controllers (where N is an integer greater than 2). The controllers 104 have cache memories 112 in which they temporarily store the data most recently read and written by the host 102. The operation of these cache memories 112 is now described with reference to FIG. 2.

FIG. 2 shows a block diagram of the caches 112, which include a read cache 114, a write cache 116 and a write mirror cache 118. A controller 104i (where "i" represents any integer) places write data 103 (FIG. 1) from the host 102 into its write cache 116i and data 105 (FIG. 1) read from the controller 104 by the host 102 into its read cache 114i. Each write mirror cache 118i duplicates the contents 107j of another controller's write cache 116j. The write mirror cache 118i is written to by a controller 104j around the time it initiates a write operation. The write mirror caches 118 allow a duplicate copy of the write data 107 to be stored in a second controller so that a failure of either controller 104 will not result in the loss of data.

Data 107 for the write mirror caches 118 is transferred between the controllers through the backend SCSI or Fibre Channel disk buses (busses) 110. The data in a mirrored cache 118 is used only if a controller 104 involved in a write fails, in which case the mirrored data is transferred to the disks 108 for storage.

5

The problem with this method is that the caches may not be synchronized, which can cause the hosts to receive inconsistent data following read operations. For example, if the host controller 104-1 performs a write to a disk device 108 and the second host system 102-1 attempts to read the same data, a copy of which is already in the read
10 cache 114-2 of the second controller 104-2, the second host would receive state data as the read caches are not updated across controllers. Further, copying all read data across the controllers would severely compromise performance. This problem will become increasingly important as clustering environments increase in popularity.

15

SUMMARY OF THE INVENTION

In summary, the present invention is a method to insure cache coherency across multiple RAID controllers. This invention synchronizes both the write and read caches,
20 whereas prior art implementations only synchronize the write cache data.

In particular, the present invention synchronizes the read caches without needing to copy the read cache data between controllers. As a result, the present invention provides full cache coherency without using excessive bandwidth for inter-controller
25 data transfers.

One embodiment of the present invention for use with multiple RAID controllers having associated read, write and mirror caches, where the controllers are connected via one or more backend channels to each other and a set of storage devices, includes the
30 following steps:

1. In response to a command to write data to a specific address, a first controller initiates the write operation and copies the particular data to the mirror caches of one or more other controllers through the one or more backend channels.
- 5 2. The first controller then confirms that the write operation is complete.
3. In response to a command to read data at the specific address, a second controller that is one of the other controllers:
 - a. checks for the data in its mirror cache and, if the data is there, returns that data to the host computer;
 - 10 b. checks for the data in its write cache, and, if the data is there, returns that data to the host;
 - c. checks for the data in its read cache and, if the data is there, returns that data to the host; and
 - d. if the data is not in any of the caches, gets the data from the
15 storage devices, places the data into its read cache and returns that data to the host.

In one embodiment, the backend channels can be any combination of SCSI and/or Fibre Channel buses. In another embodiment, the storage devices are disks. In none of the
20 embodiments is it required that a controller has a one-to-one relationship with its associated read, write and write mirror caches or with a host for which it executes read and write commands.

The present invention also encompasses RAID systems incorporating RAID controllers
25 programmed to implement the preceding method, as well as computer programs and computer program products recorded on tangible recording media which incorporate the inventive method.

BRIEF DESCRIPTION OF THE DRAWINGS

Additional objects and features of the invention will be more readily apparent from the following detailed description and appended claims when taken in conjunction with the
5 drawings, in which:

FIG. 1 shows a block diagram of a typical multiple-controller RAID system 100 (RAID is an acronym for “Redundant Array of Independent Disks”);

10 FIG. 2 shows a block diagram of the caches 112, which include a read cache 114, a write cache 116 and a write mirror cache 118;

FIG. 3 shows a block diagram of a RAID controller in which the present invention is implemented;

15

FIG. 4 shows a flow diagram of a RAID controller write method implemented in accordance with the present invention;

FIG. 5 shows a flow diagram of a RAID controller read method implemented in
20 accordance with the present invention; and

FIG. 6 shows a sequence diagram illustrating actions performed and messages exchanged by sending and receiving RAID controllers in one embodiment of a cache mirroring system.

25

DESCRIPTION OF THE PREFERRED EMBODIMENT

FIG. 3 shows a block diagram of a computer system 100 in which the present invention
30 is implemented. The computer system 100 includes many of the same components illustrated and described with reference to FIG. 1. In particular, the computer system

100 includes at least two host computers 102i, 102j, each coupled to a RAID (Redundant Arrays of Independent Disks) controller 104i, 104j. The RAID controllers 104 provide the hosts 102 with read/write access to the shared storage devices 108, which are coupled to the controllers 104 via one or more backend disk buses (referred
5 to herein after as the backend bus) 110. In different embodiments the backend bus 110 can be a Fibre Channel bus, a SCSI bus, or another type of high-speed bus (e.g., Ethernet, IPI, HIPPI, Fire-Wire or IDE). In different embodiments the shared storage devices 108 are magnetic disk drives, magnetic tape drives, optical disks, or another type of non-volatile storage (e.g., RAMDISK drives).

10

Each RAID controller 104 includes a processor 200 and a memory 202. The memory 202 can be any combination of a fast memory, such as a semiconductor random access memory (RAM), a fast non-volatile memory, such as a read only memory (ROM) or an erasable read only memory EPROM, and a slow magnetic memory, such as a hard
15 disk. The memory 202 includes a read cache 114, a write cache 116 and a write mirror cache 118 (referred to hereinafter as a "mirror cache"). The memory 202 also includes controller routines 204, which are programs that are executed by the processor 200 and determine, among other things, the operation of the controller 104 in response to read and write commands 126, 128 issued by the host computer 102. In one embodiment
20 a read command 126 designates the address 132 and size 134 of data 140 to be read from the storage devices 108 and a write command 128 includes data 138 to be written to a particular address 136 of the storage devices 108.

The controller routines 204 include, but are not limited to, read and write programs 208,
25 206. In one embodiment the controller routines 204 are firmware, meaning that they are stored in non-volatile memory and therefore are available for execution as soon as the controller 104 is powered-up. The controller routines 204 can also be implemented as conventional software that is stored on a hard disk (not shown), or other non-volatile memory, and is loaded into RAM or other fast memory for execution by the processor
30 200 under control of an operating system (not shown). The read and write programs

208, 206, which embody principles of the present invention, are now described with reference to FIGS. 4 and 5, respectively.

FIG. 4 shows a flow chart of selected steps performed by a controller 104i under control
5 of the write program 206. In the conventional manner, in response to a host write command 128 (step 302) the controller 104i writes the designated write data 138 to the shared storage devices 108 and to its write cache 116i (304). In accordance with the present invention, the controller 104i also copies the designated data 138 to the mirror caches 118j of each of the other controllers 104j (306). The controller 104i
10 accomplishes this “mirror copy” operation by broadcasting the write data 138 over the backend bus 110 using a special mirror cache write command. Finally, the controller 104i confirms the write operation’s completion (308). If it cannot confirm completion, the controller 104i re-attempts the write operation using the write data 138 previously stored in the write cache 116i. One embodiment of the mirror copy operation 306 is
15 now described with reference to FIG. 6.

FIG. 6 shows a sequence diagram of one embodiment of the mirror copy operation 306 that can be performed by a sending controller 104i and one or more receiving controllers 104j. This diagram shows the sequence of actions performed and messages
20 exchanged by the controllers 104 in which the special mirror cache write command is implemented using the “Vendor-Unique” command format defined by the SCSI protocol specification. In this embodiment, the sending controller 104i bundles the data 402 to be cached with cache meta-data 404 (information about the address 404a and size 404b of the cache data) and imbeds this data into the data-phase of a Vendor-
25 Unique command 406 (6.1). Vendor-Unique commands are known in the art and is generally a command that allows the unique characteristics of the controller (for example, those characteristics not defined in the SCSI or Fibre channel specifications and therefore possibly not available via standard SCSI or Fibre Channel commands or protocols) to be determined and set, as well as allowing other special operations to the
30 controlled device. (Additional background information about the Vendor-Unique command may be found in the SCSI-III protocol specification, which is incorporated

herein by reference: SCSI-3 Primary Commands (SPC) {Date: 1997/03/28, Rev. 11a, Status: Published, Project 0995-D} X3.301:1997). The sender 104i then initiates a transfer of the cache and meta data 402, 404 to the receiving controller 104j using the Vendor-Unique command 406 (6.2) . The receiver 104j, which is configured to
5 recognize the Vendor-Unique command 406 and to have a-priori knowledge of the structure of the data 402, 404, receives the data (6.3) and transfers that data into the correct position in the receiver's write mirror cache 118 (6.4). The receiver 104j then acknowledges completion of the command 406 through the usual SCSI mechanisms (6.5). Sending the command 406 to any device other than an appropriately configured
10 controller 104 will result in an error condition for that device. In this methodology, the command 406 is sent to a single receiver as the SCSI protocol does not define a "broadcast" method. In the situation of multiple receivers, the command is sent independently to all receivers. This can occur simultaneously since multiple back-end (disk-side) channels are connected to the controllers, allowing a command to be sent
15 to each receiver on a different back end channel.

As a result of this write method, following execution of a write command 128 a copy of the new write data 138 is resident in the mirror caches 118j of all controllers 104j that did not perform the write command 128. The mirrored data can be used by
20 subsequent read operations initiated by the controllers 104j, ensuring that a read command 126 issued for the new data 138 returns the newest version 140 of that data, which is not the case with the prior art methods. The read method of the present invention, which makes this possible, is now described with reference to FIG. 5.

25 FIG. 5 shows a flow chart of selected steps performed by a controller 104j under control of the read program 208. In accordance with the present invention, a controller 104j carries out a host read command 126 in such a way as to ensure that it returns the current version of the requested read data to the host 102j. In particular, in response to the read command 126 (320), the controller 104j first looks in its mirror cache 118j for
30 the designated read data 140 (i.e., the data at address 132 of size 134) (322). If the read data 140 is in the mirror cache 118j (322-Y), the controller returns that data to the host

102j (324). If the read data 140 is not in the mirror cache (322-N), the controller 104j checks its write cache 116i (326). If the read data 140 is in its write cache 116j (326-Y), the controller 104j returns that data to the host 102j (328). If the read data 140 is not in the mirror cache (322-N), the controller checks its read cache 114i (330). If the
5 read data 140 is in its read cache 114j (330-Y), the controller 104j returns that data to the host 102j (332). If the read data 140 is not in the mirror cache (330-N), the controller returns the designated read data from the storage devices 108 (334) and stores the same data in its read cache 116i for subsequent use (334). (Note: generally, the controller 104 writes any data returned to the host to its read cache 114).

10

By checking the write mirror cache first in response to a read command, this embodiment ensures that a controller 104 returns to a host 102 the current version of the requested read data, even if previous versions of the requested data are already resident in the controller's write and/or read caches. Similarly, by requiring the
15 controller 104 to return the requested data preferentially from its write cache 116 instead of its read cache 114 in the event the requested data is not in the mirror cache 118, this embodiment ensures that the controller 104 returns the most recent version of data it has updated. Finally, by providing for the controller 104 to supply the requested data from its read cache 114 when the other two options fail, the described embodiment
20 ensures that data already read by the controller 104 is returned to the host 102 with minimum delay.

Thus, the present invention maintains cache coherency in a RAID system including multiple hosts and RAID controllers. In one embodiment, shown in FIG. 3, this
25 advantage is provided without requiring additional, high bandwidth data transfers between controllers 104. This is possible because the controllers 104 not involved in a write operation simply receive the write data 138 as it is being written to the shared storage devices 108 via the backend bus.

30 In summary, in one embodiment for use in a RAID system having multiple RAID controllers and a set of storage devices, the host read and write processing includes:

(1) in response to a write command 128 to write first data 138 to the storage devices 108, a first controller 104i writes the first data 138 to the storage devices 108 and copies the first data to mirror caches 118j associated with one or more other controllers 104j; and

5 (2) in response to a read command 126 to read second data 134 from the storage devices 108, a second controller 104 checks for the second data 134 in an associated one of the mirror caches and, if the data 134 is in the associated mirror cache, returns the second data to the host computer 102 that issued the read command 126.

10 In another alternate embodiment the first controller 104i copies the first data 138 to the associated mirror caches 118j by broadcasting the first data 138 to the associated mirror caches over a backend bus 110 to which the controllers 104 and the storage devices 108 are coupled. In a related embodiment the broadcasting step is implemented so that it adds no more than minimal overhead to the step of writing the first data 138 to the
15 storage devices 108. Yet another related embodiment provides this minimal overhead by performing the broadcasting and writing steps simultaneously.

In embodiments where the RAID controllers 104 have associated read and write caches 114, 116, the host read and write processing includes the following steps in addition to
20 the two outlined above:

(3) in response to the write command 128 to write first data 138 to the storage devices 108, the first controller 104i also writes the first data to its associated write cache 116i;

(4) in response to the read command 126 to read second data 134 from the
25 storage devices 108, the second controller 104:

(a) checks for the second data 134 in the associated write cache 116, and, if the data is there, returns the second data 134 to the host 102;

(b) checks for the second data 134 in the read cache 114 and, if the data is there, returns the second data 134 to the host 102; and

(c) if the second data 134 is not in the associated caches 116, 118, retrieves the second data from the storage devices 108 and returns the second data 134 to the host computer 102 that issued the read command.

5 We now review aspects of the invention by way of highlighting selected particular embodiments of the invention. In a first aspect the invention provides a synchronization method for use in an RAID system having multiple RAID controllers and a set of storage devices, the method including the steps: in response to a write command to write first data to the storage devices, a first one of the controllers writes
10 the first data to the storage devices and copies the first data to mirror caches associated with one or more other controllers; and in response to a read command to read second data from the storage devices, a second controller checks for the data in an associated one of the mirror caches and, if the data is in the associated mirror cache, returns the second data to a host computer that issued the read command.

15

In a second aspect, this synchronization method further provides that the RAID controllers have associated read and write caches, the method further comprises in response to the write command to write first data to the storage devices, the first controller also writes the first data to its associated write cache; in response to the read
20 command to read second data from the storage devices, the second controller: checks for the second data in the associated write cache, and, if the data is there, returns the second data to the host; checks for the second data in the read cache and, if the data is there, returns the second data to the host; and if the second data is not in the associated caches, retrieves the second data from the storage devices and returns the second data
25 to the host computer that issued the read command.

In a third aspect, the synchronization method provides that the copying of the first data to the associated mirror caches comprises broadcasting the first data to the associated mirror caches over a backend bus to which the controllers and the storage devices are
30 coupled. In a fourth aspect, the synchronization method additionally provides that the broadcasting of the first data adds no more than minimal overhead to the writing of the

first data to the storage devices. In a fifth aspect, the broadcasting and the writing of the first data are performed simultaneously.

In a sixth aspect, the backend bus comprises any combination of: one or more Fibre Channel buses; or one or more SCSI buses. In a seventh aspect, the storage devices
5 comprise magnetic disks.

In an eighth embodiment, the synchronization method provides that the copying of the first data to the associated mirror caches comprises transmitting, using a SCSI Vendor Unique command, the first data to the associated mirror caches over a backend bus to
10 which the controllers and the storage devices are coupled.

In a ninth aspect, the invention provides a cache system for use in a RAID system including a plurality of RAID controllers providing access to a set of storage devices, comprising: a plurality of mirror caches accessible to the controllers; a first RAID being
15 controller configured, when it receives a write command to write data to a specific address, to copy the data to the mirror caches of one or more different RAID controllers in addition to writing the data to the storage devices, any of the RAID controllers being configured, after receiving a read command to read the data at the specific address, to attempt first to retrieve the data from an associated one of the mirror caches.

20

In a tenth aspect, the cache system is further defined such that the plurality of controllers and the set of storage devices are connected via a backend bus. In an eleventh aspect, the cache system is further defined such that the backend bus comprises any combination of: one or more Fibre Channel buses; or one or more SCSI buses.

25

In a twelfth aspect the cache system further comprises: a plurality of write caches accessible to the controllers; a plurality of read caches accessible to the controllers; such that: in response to a read command to read data at the specific address, the second controller: checks for the data in its mirror cache and, if the data is there, returns the
30 data to a host computer that issued the read command; checks for the data in its write cache, and, if the data is there, returns the data to the host; checks for the data in its read

cache and, if the data is there, returns the data to the host; and if the data is not in any of the caches, retrieves the data from the storage devices and returns the data to the host.

In a thirteenth aspect, the cache system is further defined to such that the first RAID
5 controller is configured to copy the first data to the mirror caches using a SCSI Vendor Unique command transmitted over a SCSI bus to which the controllers and the storage devices are coupled. In a fourteenth aspect, the cache system is further defined such that the SCSI bus comprises a backend bus.

10 In a fifteenth aspect, the invention further provides a synchronization method for use in a data storage system having at least first and second controllers and at least one storage device, the method including: in response to a write command to write first data to the storage device, the first controller writes the first data to the storage device and copies the first data to a mirror cache associated with the second controller; and in
15 response to a read command to read second data from the storage device, the second controller checks for the data in an associated one of the mirror caches and, if the data is in the associated mirror cache, returns the second data to a host computer that issued the read command.

20 In a sixteenth aspect, the synchronization method further provides that the controllers have associated read and write caches, the method further comprising: in response to the write command to write first data to the storage device, the first controller also writes the first data to its associated write cache; in response to the read command to read second data from the storage device, the second controller: checks for the second
25 data in the associated write cache, and, if the data is there, returns the second data to the host; checks for the second data in the read cache and, if the data is there, returns the second data to the host; and if the second data is not in the associated caches, retrieves the second data from the storage device and returns the second data to the host computer that issued the read command.

In a seventeenth aspect, the synchronization method provides that the copying of the first data to the associated mirror caches comprises broadcasting the first data to the associated mirror caches over a bus to which the controllers and the storage devices are coupled. In an eighteenth aspect, this synchronization method additionally provides
5 that the broadcasting of the first data adds substantially no overhead to the writing of the first data to the storage devices. In a nineteenth aspect, the synchronization method further provides that the broadcasting and the writing of the first data are performed substantially simultaneously. In a twentieth aspect, the synchronization method utilizes a bus comprises any combination of: one or more Fibre Channel buses; or one or more
10 SCSI buses.

In a twenty-first aspect, the synchronization method is utilized in conjunction with magnetic disc drive storage devices wherein the storage devices comprise magnetic disk drives.

15

In a twenty-second embodiment, the synchronization method is further defined such that the copying of the first data to the associated mirror caches comprises transmitting, using a SCSI Vendor Unique command, the first data to the associated mirror caches over a bus to which the controllers and the storage devices are coupled. In a twenty-
20 third aspect, this synchronization method is further defined such that the bus is a backend bus coupling the controllers and the storage devices.

In a twenty-fourth aspect, the invention also provides a computer program as well as a computer program product for use in conjunction with a computer system, the
25 computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism, comprising: a program module that directs at least one of a plurality of controllers connected to a host computer, and one or more disk storage devices grouped into a data storage system, to function in a specified manner, the program module including
30 instructions for: a first one of the controllers writes the first data to the storage devices and copies the first data to mirror caches associated with one or more other controllers

in response to a write command to write first data to the storage devices; and a second controller checks for the data in an associated one of the mirror caches and, if the data is in the associated mirror cache, returns the second data to a host computer that issued the read command in response to a read command to read second data from the storage
5 devices.

In a twenty-fifth aspect, the computer program and computer program product are further defined such that the computer program mechanism and computer mechanism further include a program module including instructions: the first controller also writes
10 the first data to its associated write cache in response to the write command to write first data to the storage devices; the second controller, in response to the read command to read second data from the storage devices: checks for the second data in the associated write cache, and, if the data is there, returns the second data to the host; checks for the second data in the read cache and, if the data is there, returns the second
15 data to the host; and if the second data is not in the associated caches, retrieves the second data from the storage devices and returns the second data to the host computer that issued the read command.

While the present invention has been described with reference to a few specific
20 embodiments, the description is illustrative of the invention and is not to be construed as limiting the invention. Various modifications may occur to those skilled in the art without departing from the true spirit and scope of the invention as defined by the appended claims.

25 For example, in none of the embodiments is it required that the controllers 104 have a one-to-one relationship with a set of associated read, write and write mirror caches 114, 116, 118 or with a host 102 for which it executes read and write commands. Additionally, it is not required that each controller 104 has a full complement of associated read, write and write mirror caches 114, 116, 118. Instead, all that is
30 required by the present invention is that each controller 104 have an associated write mirror cache or other quickly accessed memory location into which other controllers

copy host write data for subsequent, speedy retrieval by the former controller 104 in response to a host read command.

WHAT IS CLAIMED IS:

1. A synchronization method for use in an RAID system having multiple RAID controllers and a set of storage devices, the method including:
 - 5 in response to a write command to write first data to the storage devices, a first one of the controllers writes the first data to the storage devices and copies the first data to mirror caches associated with one or more other controllers; and
 - in response to a read command to read second data from the storage devices, a second controller checks for the data in an associated one of the mirror caches and, if
 - 10 the data is in the associated mirror cache, returns the second data to a host computer that issued the read command.
2. The synchronization method of claim 1, wherein the RAID controllers have associated read and write caches, the method further comprising:
 - 15 in response to the write command to write first data to the storage devices, the first controller also writes the first data to its associated write cache;
 - in response to the read command to read second data from the storage devices, the second controller:
 - checks for the second data in the associated write cache, and, if the data
 - 20 is there, returns the second data to the host;
 - checks for the second data in the read cache and, if the data is there, returns the second data to the host; and
 - if the second data is not in the associated caches, retrieves the second data from the storage devices and returns the second data to the host computer
 - 25 that issued the read command.
3. The synchronization method of claim 1, wherein the copying of the first data to the associated mirror caches comprises broadcasting the first data to the associated mirror caches over a backend bus to which the controllers and the storage devices are
- 30 coupled.

4. The synchronization method of claim 3, wherein the broadcasting of the first data adds no more than minimal overhead to the writing of the first data to the storage devices.
- 5 5. The synchronization method of claim 4, wherein the broadcasting and the writing of the first data are performed simultaneously.
6. The synchronization method of claim 3, wherein the backend bus comprises any combination of:
- 10 one or more Fibre Channel buses; or
one or more SCSI buses.
7. The synchronization method of claim 1, wherein the storage devices comprise magnetic disks.
- 15 8. The synchronization method of claim 1, wherein the copying of the first data to the associated mirror caches comprises transmitting, using a SCSI Vendor Unique command, the first data to the associated mirror caches over a backend bus to which the controllers and the storage devices are coupled.
- 20 9. A cache system for use in a RAID system including a plurality of RAID controllers providing access to a set of storage devices, comprising:
a plurality of mirror caches accessible to the controllers;
a first RAID being controller configured, when it receives a write command to
25 write data to a specific address, to copy the data to the mirror caches of one or more different RAID controllers in addition to writing the data to the storage devices,
any of the RAID controllers being configured, after receiving a read command to read the data at the specific address, to attempt first to retrieve the data from an associated one of the mirror caches.

10. The cache system of claim 9, wherein the plurality of controllers and the set of storage devices are connected via a backend bus.
11. The cache system of claim 10, wherein the backend bus comprises any
5 combination of:
one or more Fibre Channel buses; or
one or more SCSI buses.
12. The cache system of claim 9, further comprising:
10 a plurality of write caches accessible to the controllers;
a plurality of read caches accessible to the controllers; such that:
in response to a read command to read data at the specific address, the second
controller:
checks for the data in its mirror cache and, if the data is there, returns the
15 data to a host computer that issued the read command;
checks for the data in its write cache, and, if the data is there, returns the
data to the host;
checks for the data in its read cache and, if the data is there, returns the
data to the host; and
20 if the data is not in any of the caches, retrieves the data from the storage
devices and returns the data to the host.
13. The cache system of claim 9, wherein the first RAID controller is configured
to copy the first data to the mirror caches using a SCSI Vendor Unique command
25 transmitted over a SCSI bus to which the controllers and the storage devices are
coupled.
14. The cache system of claim 13, wherein the SCSI bus comprises a backend bus.

15. A synchronization method for use in a data storage system having at least first and second controllers and at least one storage device, the method including:

in response to a write command to write first data to the storage device, the first controller writes the first data to the storage device and copies the first data to a mirror
5 cache associated with the second controller; and

in response to a read command to read second data from the storage device, the second controller checks for the data in an associated one of the mirror caches and, if the data is in the associated mirror cache, returns the second data to a host computer that issued the read command.

10

16. The synchronization method of claim 15, wherein the controllers have associated read and write caches, the method further comprising:

in response to the write command to write first data to the storage device, the first controller also writes the first data to its associated write cache;

15 in response to the read command to read second data from the storage device, the second controller:

checks for the second data in the associated write cache, and, if the data is there, returns the second data to the host;

20 checks for the second data in the read cache and, if the data is there, returns the second data to the host; and

if the second data is not in the associated caches, retrieves the second data from the storage device and returns the second data to the host computer that issued the read command.

25 17. The synchronization method of claim 15, wherein the copying of the first data to the associated mirror caches comprises broadcasting the first data to the associated mirror caches over a bus to which the controllers and the storage devices are coupled.

18. The synchronization method of claim 17, wherein the broadcasting of the first
30 data adds substantially no overhead to the writing of the first data to the storage devices.

19. The synchronization method of claim 18, wherein the broadcasting and the writing of the first data are performed substantially simultaneously.
20. The synchronization method of claim 18, wherein the bus comprises any
5 combination of: one or more Fibre Channel buses; or one or more SCSI buses.
21. The synchronization method of claim 15, wherein the storage devices comprise magnetic disk drives.
- 10 22. The synchronization method of claim 15, wherein the copying of the first data to the associated mirror caches comprises transmitting, using a SCSI Vendor Unique command, the first data to the associated mirror caches over a bus to which the controllers and the storage devices are coupled.
- 15 23. The synchronization method of claim 22, wherein the bus is a backend bus coupling the controllers and the storage devices.
24. A computer program for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a
20 computer program mechanism embedded therein, the computer program mechanism, comprising:
- a program module that directs at least one of a plurality of controllers connected to a host computer, and one or more disk storage devices grouped into a data storage system, to function in a specified manner, the program module including instructions
25 for:
- a first one of the controllers writes the first data to the storage devices and copies the first data to mirror caches associated with one or more other controllers in response to a write command to write first data to the storage devices; and
- a second controller checks for the data in an associated one of the mirror caches
30 and, if the data is in the associated mirror cache, returns the second data to a host

computer that issued the read command in response to a read command to read second data from the storage devices.

25. A computer program as in claim 24, the computer program mechanism and
5 computer mechanism further including a program module including instructions

the first controller also writes the first data to its associated write cache in response to the write command to write first data to the storage devices;

the second controller, in response to the read command to read second data from the storage devices:

10 checks for the second data in the associated write cache, and, if the data is there, returns the second data to the host;

checks for the second data in the read cache and, if the data is there, returns the second data to the host; and

15 if the second data is not in the associated caches, retrieves the second data from the storage devices and returns the second data to the host computer that issued the read command.

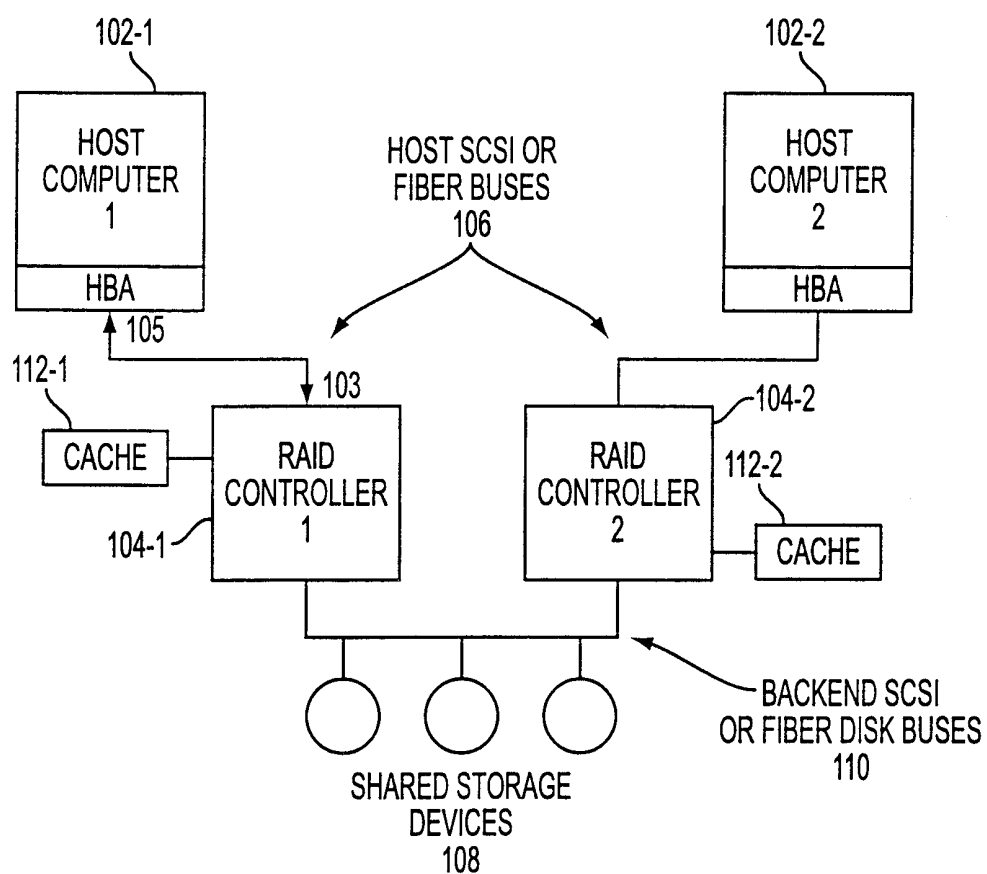


FIG. 1
(PRIOR ART)

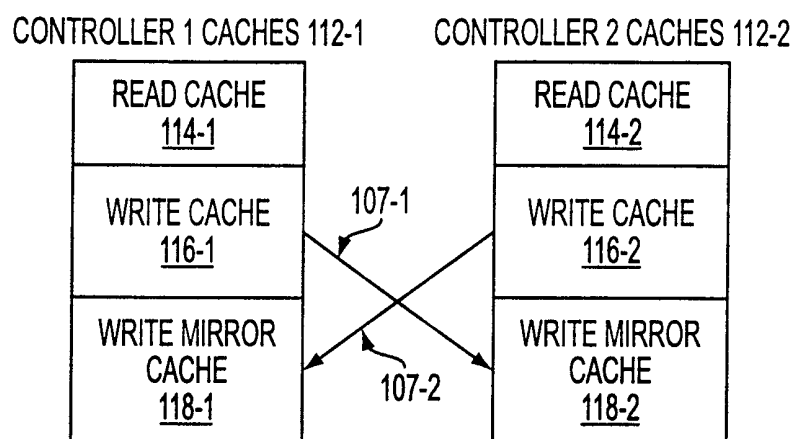


FIG. 2
(PRIOR ART)

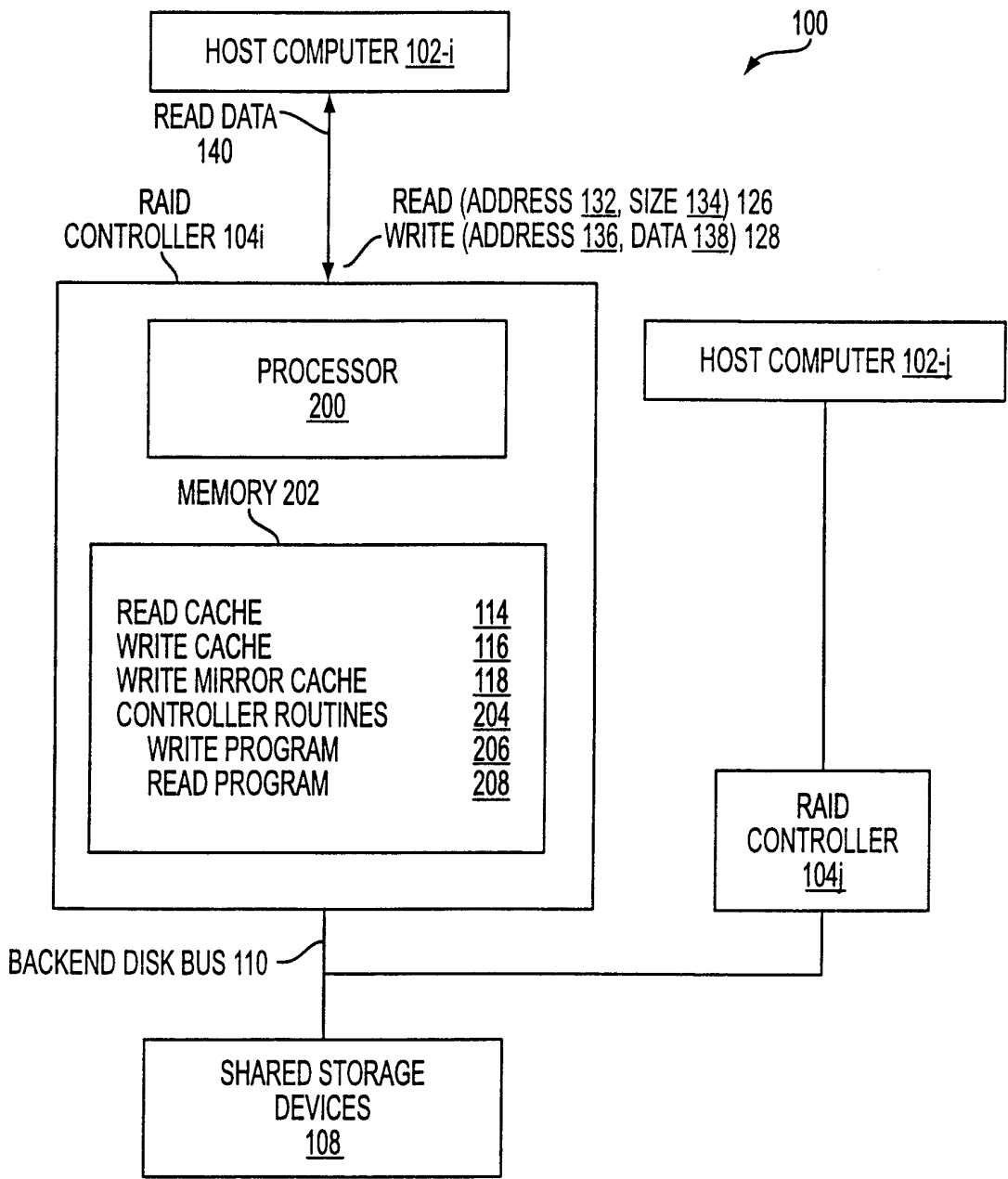


FIG. 3

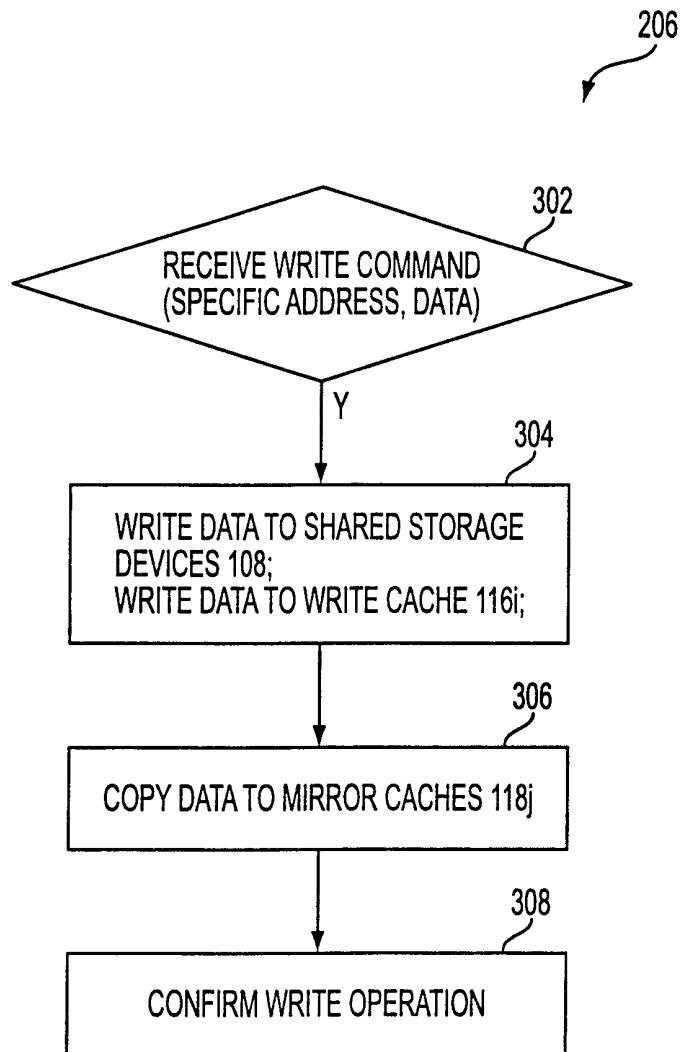


FIG. 4

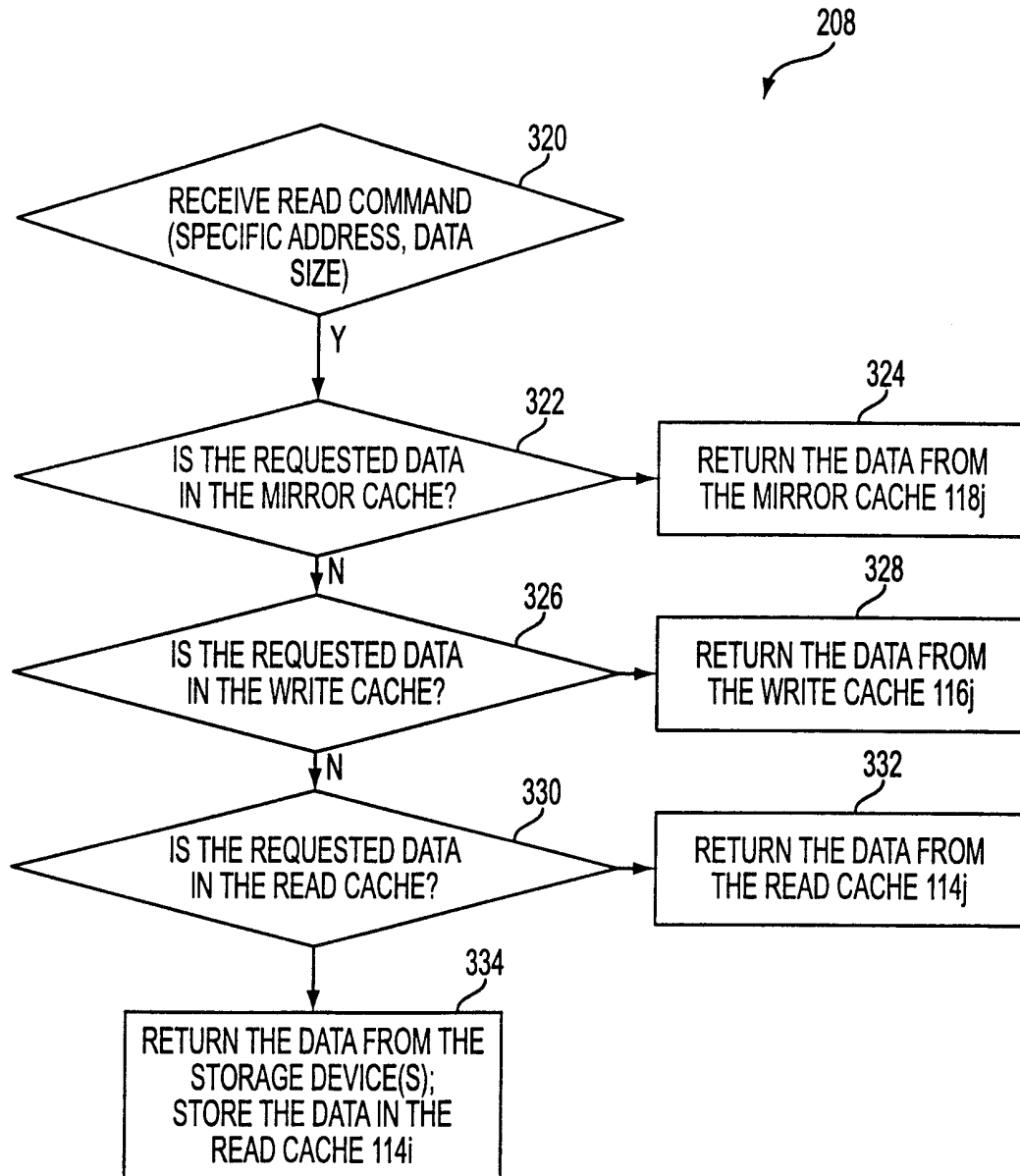
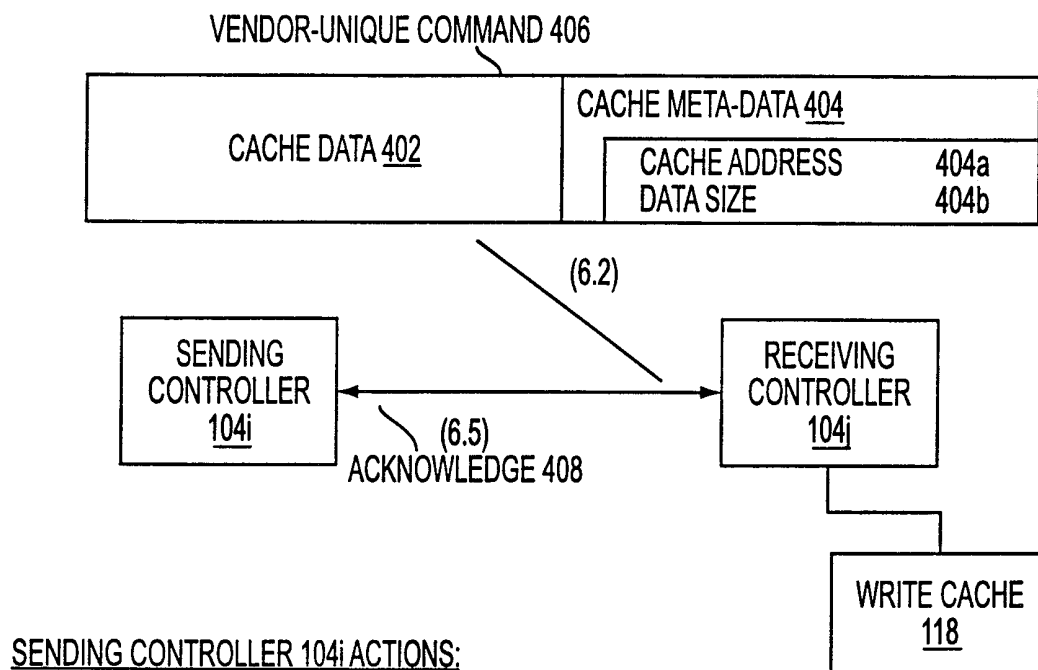


FIG. 5

6/6

SENDING CONTROLLER 104i ACTIONS:

(6.1) PREPARE THE VENDOR-UNIQUE COMMAND 406 WITH BUNDLED CACHE DATA AND CACHE META-DATA.
 (6.2) INITIATE TRANSFER OF VENDOR-UNIQUE COMMAND.

RECEIVING CONTROLLER 104j ACTIONS:

(6.3) RECEIVE THE VENDOR-UNIQUE COMMAND.
 (6.4) TRANSFER CACHE DATA AND CACHE META-DATA TO CORRECT POSITION IN WRITE MIRROR CACHE 118.
 (6.5) RETURN ACKNOWLEDGE 408

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US00/01921

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 12/08

US CL : 711/113; 714/6

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)

U.S. : 711/113, 114, 120, 124, 141, 142, 143; 714/5, 6, 7

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)

http://iel.ihs.com/

search terms: RAID, read cache, write cache, mirror

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	US 5,974,506 A (SICOLA et al) 26 October 1999, col. 3 line 58 to col. 4 line 20.	1-25
Y	US 5,761,705 A (DEKONING et al) 02 June 1998, col. 5 line 66 to col. 6 line 21.	1-25
Y	US 5,636,355 A (RAMAKRISHNAN et al) 03 June 1997, col. 8 lines 40-43.	1-25
A	VARMA, A et al., Destage Algorithms for Disk Arrays with Non-Volatile Caches, 1995 ACM.	1-25

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

* Special categories of cited documents:	*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
A document defining the general state of the art which is not considered to be of particular relevance	*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
B earlier document published on or after the international filing date	*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
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)	*A* document member of the same patent family
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	

Date of the actual completion of the international search 27 MARCH 2000	Date of mailing of the international search report 18 APR 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer GARY J. PORTKA Telephone No. (703) 305-3900 <i>Joni Hill</i>