PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :		(11) International Publication Number:
G06F 11/10	A1	WO 92/15057
		(43) International Publication Date: 3 September 1992 (03.09.92)

(21) International Application Number: PCT/US92/01257

(22) International Filing Date: 18 February 1992 (18.02.92)

(30) Priority data:
658,317 20 February 1991 (20.02.91) US

(71) Applicant: MICROPOLIS CORPORATION [US/US]; 21211 Nordhoff, Chatsworth, CA 91311 (US).

(72) Inventor: ANDERSON, Michael, H. ; 4341 N. Ashtree, Moorpark, CA 93011 (US).

(74) Agents: ROSE, Alan, C. et al.; Poms, Smith, Lande & Rose, 2121 Avenue of the Stars, Suite 1400, Los Angeles, CA 90067 (US).

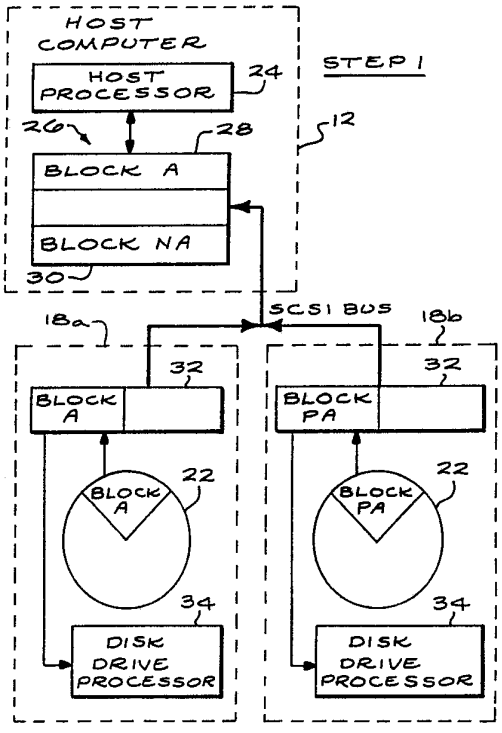
(81) Designated States: AT (European patent), BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), MC (European patent), NL (European patent), SE (European patent).

Published
With international search report.

(54) Title: PARITY CALCULATION IN AN EFFICIENT ARRAY OF MASS STORAGE DEVICES

(57) Abstract

An efficient redundant array of mass storage devices includes a plurality of hard disk drives (22), a controller (20) or processor associated with each hard disk drive for calculating partial parity data and parity data, a host computer (12) and at least one bus (16) for communications between the host computer (12) and the plurality of hard disk drives (22). The controller (20) of a drive calculates the partial parity, which is the Exclusive Or function of the old data and the new data which is to be stored into the drive (22). New data is written to the location in the disk drive from which the old data was obtained. The partial parity data is transferred to the controller (20) of another drive which contains the old parity data for the location to which the new data was written, and that controller calculates the new parity which is the Exclusive Or of the partial parity and the old parity. The new parity is written to the location on that disk drive which formerly held the old parity.



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.

AT	Austria	**FI**	Finland	**ML**	Mali
AU	Australia	**FR**	France	**MN**	Mongolia
BB	Barbados	**GA**	Gabon	**MR**	Mauritania
BE	Belgium	**GB**	United Kingdom	**MW**	Malawi
BF	Burkina Faso	**GN**	Guinea	**NL**	Netherlands
BG	Bulgaria	**GR**	Greece	**NO**	Norway
BJ	Benin	**HU**	Hungary	**PL**	Poland
BR	Brazil	**IE**	Ireland	**RO**	Romania
CA	Canada	**IT**	Italy	**RU**	Russian Federation
CF	Central African Republic	**JP**	Japan	**SD**	Sudan
CG	Congo	**KP**	Democratic People's Republic of Korea	**SE**	Sweden
CH	Switzerland			**SN**	Senegal
CI	Côte d'Ivoire	**KR**	Republic of Korea	**SU**	Soviet Union
CM	Cameroon	**LI**	Liechtenstein	**TD**	Chad
CS	Czechoslovakia	**LK**	Sri Lanka	**TG**	Togo
DE	Germany	**LU**	Luxembourg	**US**	United States of America
DK	Denmark	**MC**	Monaco		
ES	Spain	**MG**	Madagascar		

"PARITY CALCULATION IN AN EFFICIENT ARRAY OF MASS STORAGE DEVICES"

FIELD OF THE INVENTION

This invention relates in general to data storage devices for computers, and more particularly to mass storage memory arrays.

BACKGROUND OF THE INVENTION

In the recent history of computers, advances in the performance and speed of central processing units (CPU's) have far outraced the advances in the performance of hard disk drives, although hard disk drives have also made great advances in the recent past. However, the far greater increases in the performance of CPU's has begun to cause input/output (I/O) bottlenecks when the CPU accesses a disk drive because the increases in disk drive performance have not caught up to the improved performance of CPU's.

Another problem is that although top quality disk drives offer a mean time between failure of about 150,000 hours, in systems using multiple disk drives failures will occur. One approach to addressing the failure problem is called mirroring. In the mirroring technique, the host computer writes data to two disk drives simultaneously. If one disk drive fails, a copy of all the data is immediately available on the other drive. Mirroring protects against a failure in one drive but it requires a user to purchase twice as much storage to hold the data and programs. Mirroring also does not address the I/O bottle-neck problem.

Another approach which addresses the problem of a failure in a multiple disk drive system is disclosed in U.S. Patent No. 4,870,643 to Bultman, et al., which issued on September 26, 1989. The Bultman patent discloses a system with five standard 5¼" Winchester disk drives with

successive bytes of digital information routed to four of the drives. The fifth drive contains parity information. Control circuitry is provided so that any one of the five standard drives may be unplugged and replaced without interruption of the operation of the storage system. The Bultman computer configuration uses less drives for storing the same amount of information as the mirroring technique discussed above but does not address the I/O bottleneck problem.

A recently proposed computer configuration for partially alleviating these problems was set forth in an article titled "Strength (and Safety) in Numbers" in the December 1990 issue of Byte Magazine written by Michael H. Anderson. That computer configuration is called Redundant Arrays of Inexpensive Disks, and is referenced by its acronym "RAID".

A RAID system is a group of intelligent disk drives under the control of a single device driver or host computer. The proposed RAID system offers significantly higher performance than a single disk drive. Data can be striped or dispensed among several drives so that several of the drives are accessed in parallel to read one block of data which was striped across the several drives. This provides for quicker access than retrieving the block from a single drive.

In a RAID system, check bytes are stored, also preferably in an interleaved pattern across all of the drives. The check (or parity) byte is the sum of the data stored on the other drives in the same position. Therefore, if one drive fails the data which was stored on that drive can be quickly recreated by a calculation involving a check byte and data on the other non-failed drives for the same position. All of the above calculations are performed by the host computer.

When the host is instructed to write a block of new data which may be designated "NA" to a disk drive, the

host computer first reads the old data "A" from the position to which the new data "NA" will be stored and the corresponding check bytes or parity bytes "PA" as shown in Fig. 2. Incidentally, the old parity designated "PA" is the parity involving not only the data "A", but also data from other disk drives. The host then calculates the new check bytes NPA by subtracting out the old data "A" and summing in the new data "PA". The check bytes are then rewritten over the old check bytes and the new data "NA" is written to the position where the old data "A" previously resided.

Unfortunately, though the above described RAID technique offers some improvement in performance, the host processor is forced to do a large number of calculations and multiple transfers of data over the bus which connects the host and the disk drives each time a block of data is to be written to a disk drive. The large number of calculations and the bus transfers reduces the performance of the host computer and the overall system.

As compared with the foregoing prior art arrangements, the principal objects of the present invention are to reduce the involvement of the host computer in the data storage process and to minimize the number of data transfers along the bus which interconnects the disk drives and the host computer.

SUMMARY OF THE INVENTION

In accordance with the present invention, an efficient, redundant array of mass storage devices which includes a plurality of hard disk drives, a controller or processor associated with each hard disk drive, a host computer and at least one bus for communications between the host computer and the plurality of hard disk drives includes partial parity calculating means for calculating the Exclusive Or function of old data and new data, and

parity calculating means for calculating the Exclusive Or of partial parity and old parity.

Another aspect of the present invention encompasses a method for writing data to and calculating new parity data for a plurality of peripheral storage devices and a system which includes a host computer connected by a bus to a plurality of disk drives, each with its own controller. First, the system calculates the partial parity which is the Exclusive Or function of the old data and the new data which is to be stored in a location which currently holds the old data. In a preferred embodiment, the partial parity is calculated at the controller for the storage device to which the new data is written.

Next, the new data is written to the location in the peripheral storage device from which the old data was obtained. The partial parity data is then written to the controller of a peripheral storage device which contains the old parity data for the location to which the new data was written. The controller which receives the partial parity calculates the new parity which is the Exclusive Or function of the partial parity and the old parity. Finally, the new parity data is written to the location in the peripheral storage device which formerly held the old parity data.

Other objects and advantages of the present invention will be apparent to those skilled in the art from the accompanying drawings and detailed descriptions.

DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a computer system;

Fig. 2 is a pictorial representation of a mass storage system known in the prior art;

Fig. 3 through 6 are block diagram representations of the present invention during different phases of the write operation;

Fig. 7 is a representation of the arrangement of data and parity information across four disk drives.

DETAILED DESCRIPTION

Turning now to the drawings, Fig. 1 shows a preferred embodiment of the present invention. The host computer 12, preferably a micro-computer based on the Extended Industry Standard Architecture (EISA) is shown with a plurality of interface adaptor cards 14 for coupling the host computer to a plurality of buses 16, preferably a Small Computer System Interface Bus or SCSI bus. The SCSI buses interconnect the host computer and a plurality of peripheral storage devices 18.

Each peripheral storage device 18 preferably includes a dedicated spindle processor or controller 20 for each peripheral storage device, and a rotary disk drive 22 or other similar mass storage device. In one preferred embodiment, the rotary disk drives 22 are 5¼" Winchester hard disk drives. The peripheral storage devices 18 could be implemented using 1580 series or 1370 series disk drives manufactured by Micropolis of Chatsworth, California.

Fig. 2 shows a prior art RAID system as was discussed in the BACKGROUND OF THE INVENTION section of this specification.

Referring now to Fig. 3, the host computer 12 is shown with more detail including the host computer's host (central) processor 24 and its dynamic memory 26. The dynamic memory 26 is shown containing a block of data A and an identical size block of new data NA. Data block A was retrieved from the buffer memory 32 from peripheral storage device 18a. Data A had originated in the disk drive 22. Data A was then transferred to buffer memory 32 and then to the host computer's dynamic memory 26.

As can also be seen in Fig. 3, within peripheral storage device 18b the parity data (check bytes) PA has

been written from disk drive 22 to buffer memory 32. The parity data PA provides a parity check for all the data stored in all of the disk drives in the system in the same location as data block A. The parity data is discussed in more detail below.

As can be seen in Fig. 4, the host processor 24 then generates partial parity data (PPA) which is the Exclusive Or of data A and data NA. Data NA is then written to the buffer storage of peripheral memory device 18a and from there data NA is written to the same location formerly occupied by the old data A in disk drive 22 within peripheral memory device 18a.

The partial parity PPA is written to buffer memory 32 of peripheral memory device 18b. Disk drive processor 34 within peripheral memory device 18b then calculates the new parity data NPA which is the Exclusive Or of PPA and PA. The new parity data NPA is then written to the same location from which PA was retrieved in disk drive 22 within peripheral memory device 18b.

The above described configuration approach allows the disk drive processor of the peripheral storage device containing the old priority data PA to calculate the new priority data NPA by performing the Exclusive Or of PPA and PA. This saves the host computer from having to perform that function. Also, the SCSI bus has one less data transfer along it. The data PA is never transferred along the bus. Therefore less bus time is also utilized which frees up the SCSI bus for data retrieval.

Another embodiment of the present invention is shown in Figs. 5 and 6. Fig. 5 shows the dynamic memory 26 of the host computer 12 containing the new data NA to be written to one of the peripheral storage devices. Peripheral storage device 18a is shown with the old data A in its buffer memory 32. The new data NA is also shown as having been written to the buffer memory 32 of

peripheral 18a. The buffer memory 32 of peripheral device 18b contains the old parity data PA.

The disk drive processor 34 of peripheral 18a calculates the partial parity data PPA which is the Exclusive Or of A and NA. Fig. 6 shows PPA stored n buffer memory 32 of peripheral 18a after the calculation. PPA is then written to the buffer memory 32 of peripheral 18b. NA is also written to the location which held A in the disk drive 22. The disk drive processor 35 of peripheral 18b then calculates NPA (the new priority data) which is the Exclusive Or of PPA and PA. NA is then written to the location on disk drive 22 which previously contained PA.

In this embodiment, two disk drive processors calculated NPA and PPA. This saves the host computer from having to perform both of those calculations. Also, as in the previously described embodiment, the SCSI bus has one less data transfer along it as compared with the prior art systems. Therefore, less bus time is utilized, which frees up the SCSI bus for other functions.

The parity data allows the system to recreate the data stored on one disk drive from the data stored on the other drives. The parity data is the Exclusive Or of the data stored in the same location on each drive.

In a three drive system with the parity data for drives 1 and 2 stored on drive 3, the parity data would operate as follows. If location A on drive 1 contained the value 7 and location A on drive 2 contained the value 3 then the parity data stored in location A on drive 3 would be the Exclusive Or of 7 and 3.

In a four bit system 7 is represented as 0111 and three is represented as 0011. The Exclusive Or of 0111 and 0011 is 0100. If drive 1 fails, the data that was stored in location A of drive 1 is the Exclusive Or of the parity data for location A, 0100 (which is stored in location A on drive 3) and the data stored in location A

on drive 2, 0011. The Exclusive Or of 0100 and 0011 is 0111 or 7. This approach works for any number of drives.

Fig. 7 is a pictorial representation of the interleaving of data and parity information across four peripheral storage devices. Pie shaped sections 1, 2 and 3 represent identical locations on three different drives with P(1-2-3) representing the parity data which is stored in the same position as Sections 1, 2 and 3 on the fourth drive. Similarly, the parity data for locations 5, 6 and 4 is shown stored on the third drive in sequence.

Interleaving of the data can greatly increase the access speed of the overall peripheral memory storage system. When a data block is accessed, each drive can simultaneously access its portion of the requested data and transmit it in parallel to the host.

In conclusion, it is to be understood that the foregoing detailed description and accompanying drawings relate to illustrative implementations of the invention. The invention is not limited to these illustrative implementations. Thus, by way of example and not of limitation, a system using five, six or seven drives could be used. Also, storage units other than Winchester type disk drives, including optical storage devices, could be used as a peripheral storage device. Accordingly, the present invention is not limited to the system as described in detail herein and as shown in the accompanying drawings.

CLAIMS

What is Claimed is:

1. A method for writing data to and calculating new parity data for an efficient, redundant array of mass storage devices in a system which includes a host computer connected via a bus to a plurality of disk drives, each with its own controller, comprising the steps of:

calculating the partial parity which is the Exclusive Or function of the old data and the new data which is to be stored in the location which holds the old data;

writing the new data to the location in the peripheral storage device from which the old data was obtained;

transferring the partial parity to the controller of a peripheral storage device which contains the old parity data for the location to which the new data was written;

calculating the new parity which is the Exclusive Or function of the partial parity and the old parity at the controller which receives the partial parity; and

transferring the new parity to the location in the peripheral storage device which formerly held the old parity.

2. The method as described in Claim 1 wherein the partial parity is calculated by the a first controller which is the controller of the storage device to which the new data is being written.

3. The method as described in Claim 1 wherein the host computer calculates the partial parity.

4. The method as described in Claim 2 further including the steps of:

the first controller receiving the new data from the host computer; and

the first controller obtaining the old data from its associated disk drive.

5. The method as described in Claim 4 wherein said first controller performs the step of transferring the partial parity to the controller of a peripheral storage device which contains the old parity data for the location to which the new data was written.

6. An efficient, redundant array of mass storage devices which includes a plurality of hard disk drives, a host computer and at least one bus for communications between the host computer and the plurality of hard disk drives comprising:

partial parity calculating means for calculating the Exclusive Or function of old data and new data;

a plurality of controller means, one associated with each disk drive, for calculating the new parity which is the Exclusive Or function of partial parity and old parity; and

means for communicating the partial parity from said partial parity calculating means to said controller means.

7. An efficient, redundant array of mass storage devices as defined in claim 4 wherein said system further includes a plurality of partial parity calculating means, associated with each disk drive.

8. An efficient, redundant array of mass storage devices as defined in Claim 6, wherein said hard disk drives are 5¼" disk drives.

9. An efficient, redundant array of mass storage devices as defined in Claim 6, wherein said bus is a small computer system interface bus.

10. An efficient, redundant array of mass storage devices as defined in claim 7 further including means for storing the new parity on the same disk drive and in the same location as the old parity previously resided.

11. An efficient, redundant array of mass storage devices as defined in claim 10 further including means for storing the new parity in the disk drive associated with the controller means which has calculated the new parity.

12. A computer system including an efficient, redundant array of mass storage devices comprising:

a host computer;

a plurality of hard disk drives;

a least one bus interconnecting said hard disk drives and said host computer; and

a plurality of controllers, each associated with one hard disk drive, including means for calculating parity data by calculating the Exclusive Or function of partial parity data and old parity data.

13. A computer system as defined in Claim 10 wherein said controller further includes means for calculating partial parity data by calculating the Exclusive Or function of old data and new data.

14. A computer system as defined in Claim 13 wherein said controller further includes means for receiving new data from said host computer and retrieving old data from the disk drive associated with said controller.

15. A computer system as defined in Claim 14 wherein said controller further includes means for transferring said partial parity data to a second controller.

16. A computer system as defined in Claim 14 wherein said controller further includes means for storing the new parity data on the disk drive associated with that controller and in the same location as the old parity data previously resided.

17. A computer system as defined in Claim 14 wherein said controller further includes means for storing the new data in the same location as the old data previously resided on the disk drive associated with that controller.

18. A computer system as defined in Claim 12 wherein said hard disk drives ar 5¼" Winchester drives.

19. A computer system as defined in Claim 16 wherein said at least one bus is a Small Computer System Interface bus.

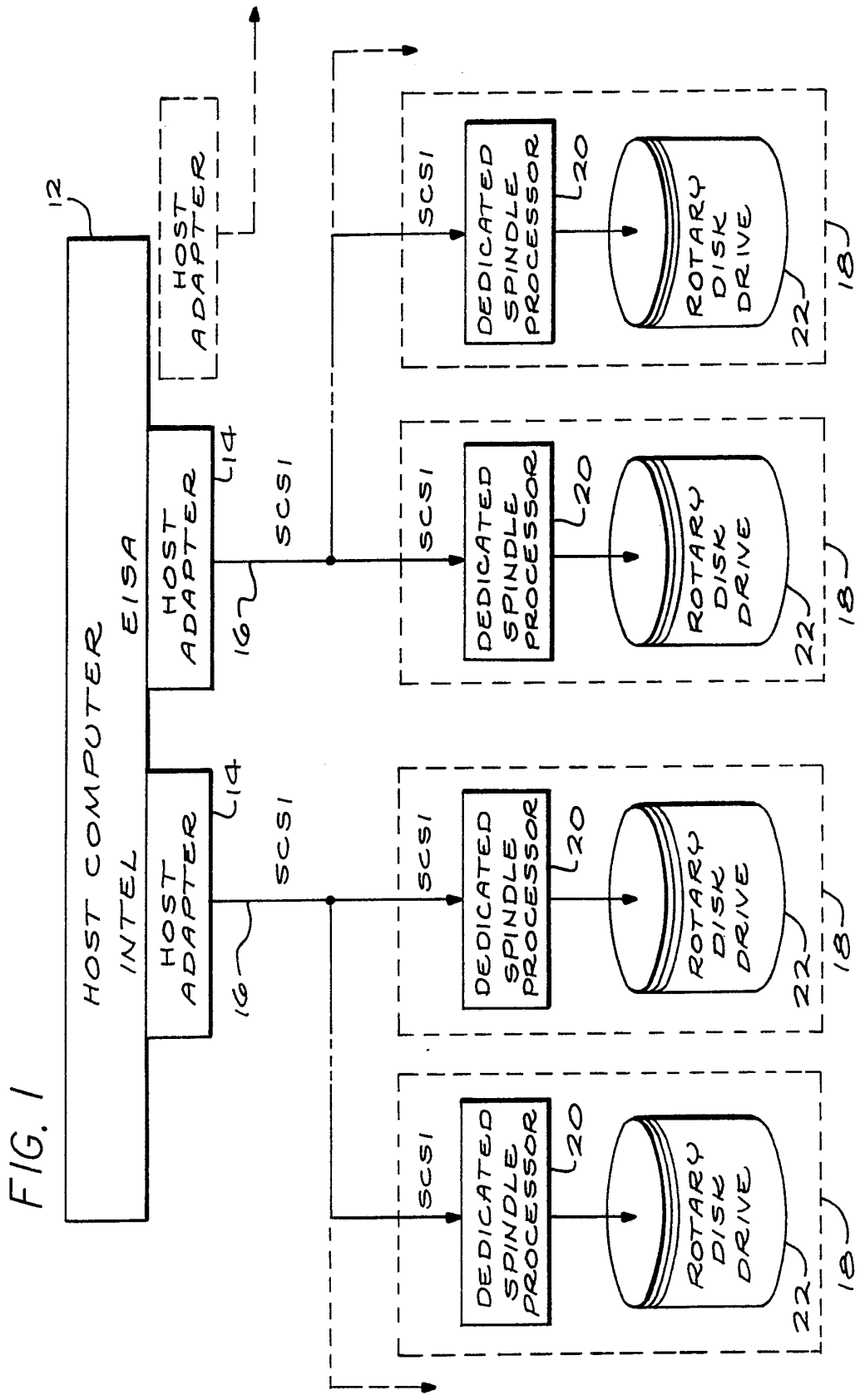

HOST COMPUTER
INTEL
EISA
12
HOST ADAPTER
HOST ADAPTER
14
16
SCSI
SCSI
DEDICATED SPINDLE PROCESSOR
20
ROTARY DISK DRIVE
22
18

FIG. 1

FIG. 2
PRIOR ART

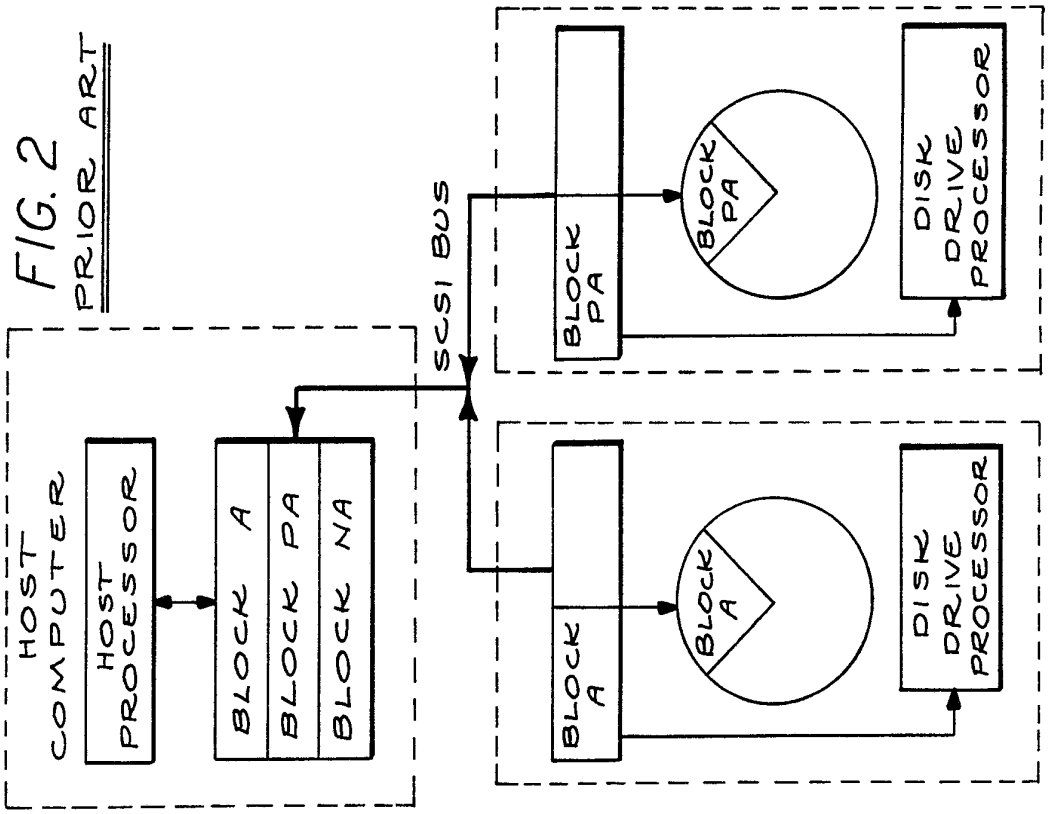


FIG. 3
STEP 1

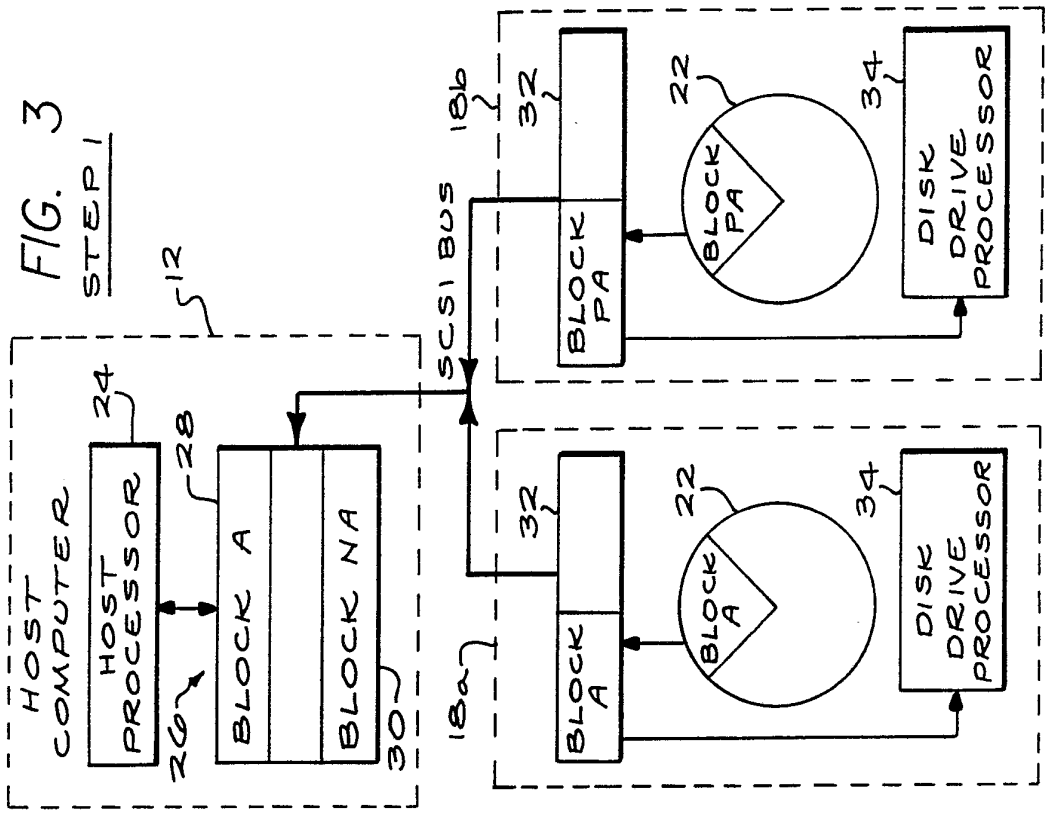


FIG. 5

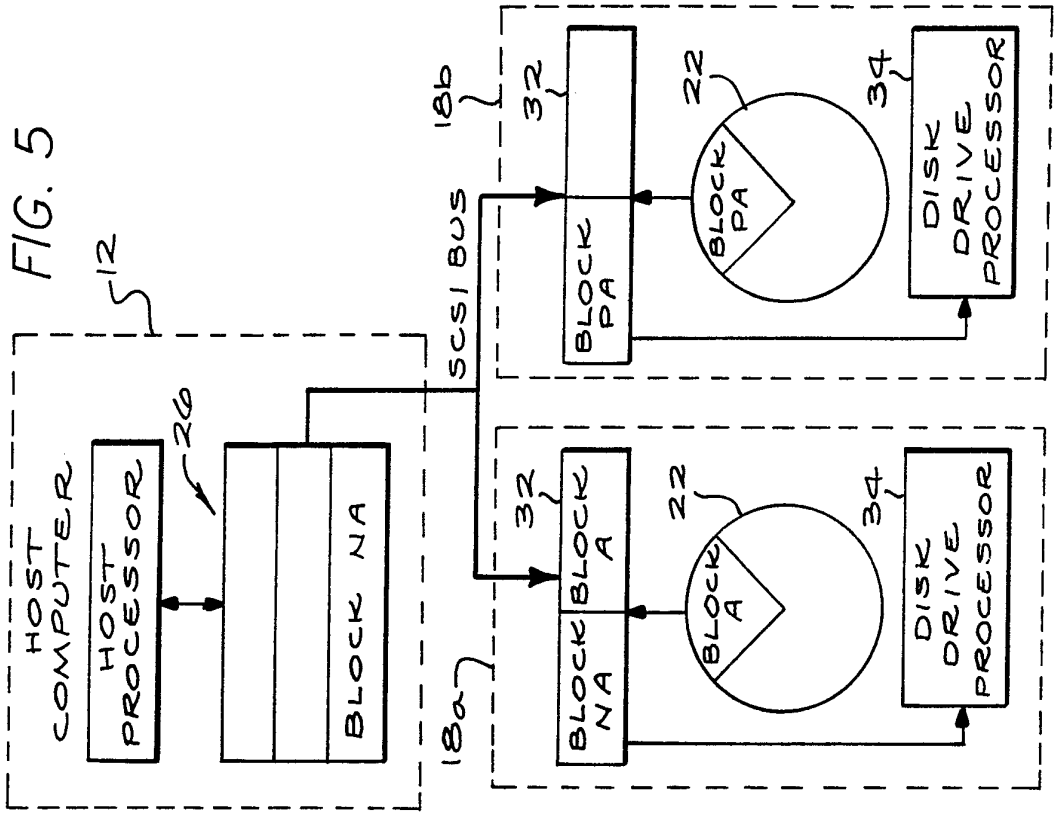
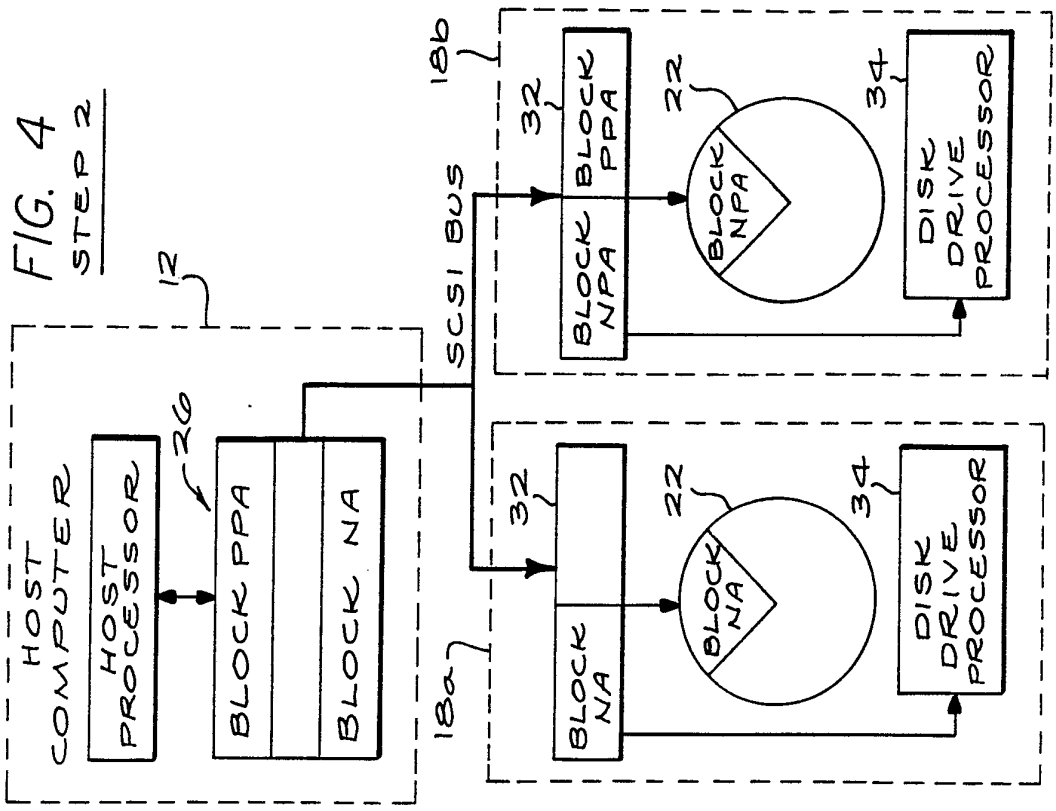


FIG. 4 STEP 2



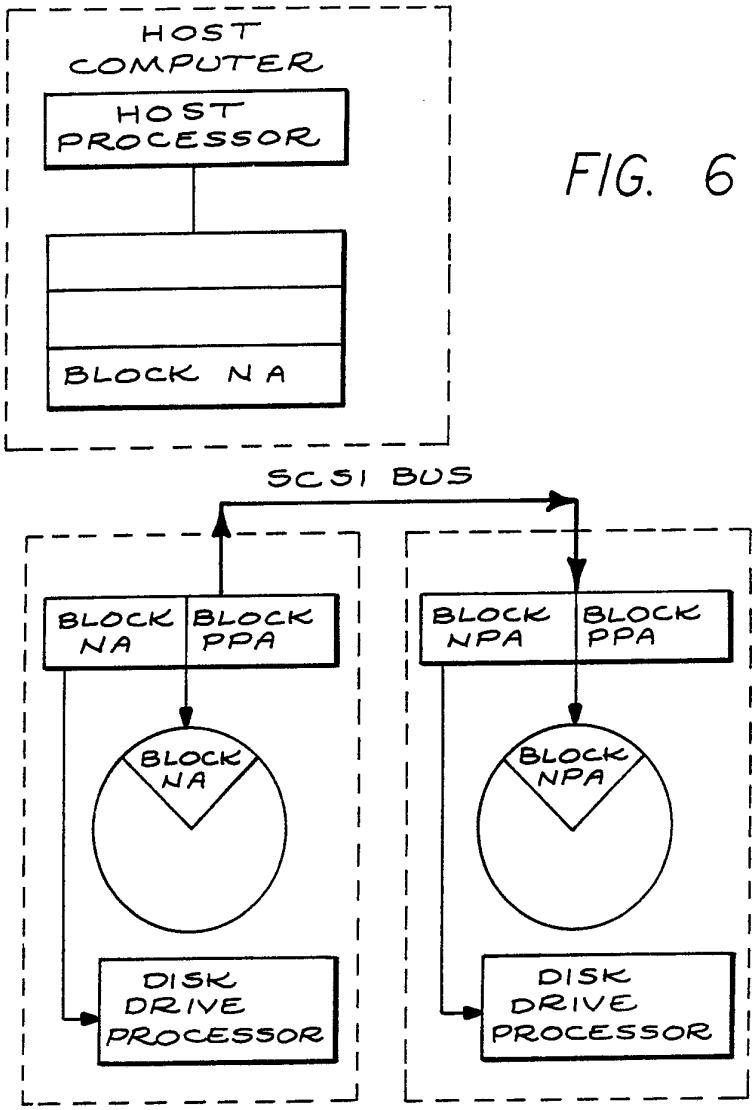
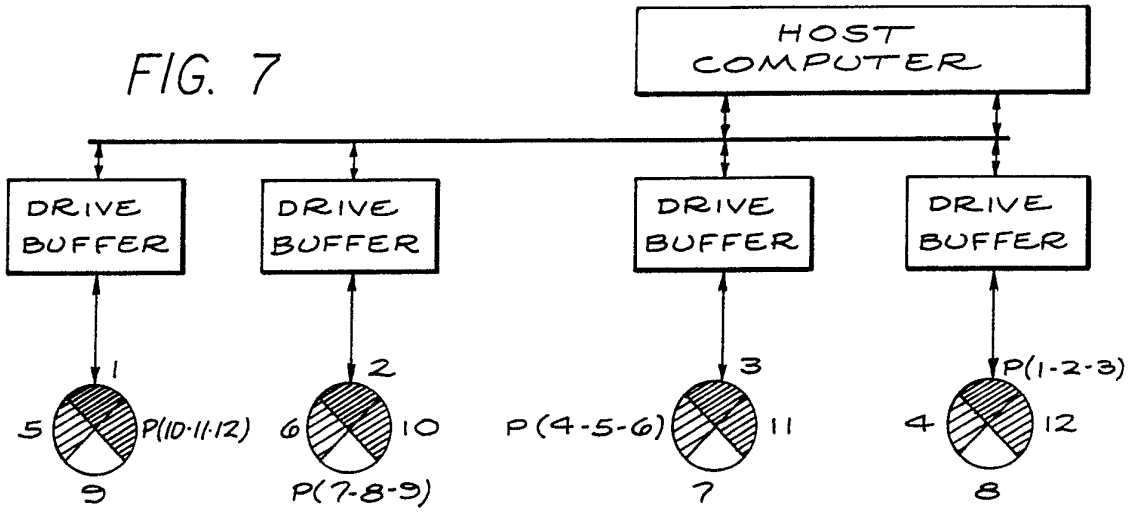

HOST COMPUTER
HOST PROCESSOR
BLOCK NA
SCSI BUS
BLOCK NA
BLOCK PPA
BLOCK NPA
BLOCK PPA
BLOCK NA
BLOCK NPA
DISK DRIVE PROCESSOR
DISK DRIVE PROCESSOR

FIG. 6

FIG. 7


HOST COMPUTER
DRIVE BUFFER
DRIVE BUFFER
DRIVE BUFFER
DRIVE BUFFER
1
2
3
P(1-2-3)
5
P(10-11-12)
9
6
10
P(7-8-9)
P(4-5-6)
11
7
4
12
8

INTERNATIONAL SEARCH REPORT

International Application No. PCT/US92/ 01257

I. CLASSIFICATION OF SUBJECT MATTER

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC (5): G06F 11/10
U.S.Cl.: 371/51.1

II. FIELDS SEARCHED

Minimum Documentation Searched

Classification System	Classification Symbols
U.S.	371/10.1,49.1,51.1 360/53; 365/200

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched

III. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
A	US, A, 4,016,409 (KIM) 05 April 1977 See entire document.	1-19
A	US, A, 3,789,204 (BARLOW) 29 January 1974 See entire document.	1-19
A	US, A, 3,887,901 (MOORE, III) 03 June 1975 See entire document.	1-19
A	US, A, 4,433,388 (OOSTERBAAN) 21 February 1984 See entire document.	1-19
A,P	US, A, 5,088,081 (FARR) 11 February 1992 See entire document.	1-19

* Special categories of cited documents:

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

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
06 May 1992	12 JUN 1992
International Searching Authority	**Signature of Authorized Officer**
ISA/US	Charles Atkinson

Form PCT/ISA/210 (second sheet) (Rev.11-87)