PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ^{6} **:** **G06F 3/06, 13/40**	**A1**	**(11) International Publication Number:** **WO 98/00776** **(43) International Publication Date:** 8 January 1998 (08.01.98)

(21) International Application Number: PCT/GB97/01474

(22) International Filing Date: 30 May 1997 (30.05.97)

(30) Priority Data:
08/673,654 28 June 1996 (28.06.96) US

(71) Applicant: SYMBIOS LOGIC INC. [US/US]; 2001 Danfield Court, Fort Collins, CO 80525 (US).

(71) Applicant *(for MG only)*: GILL, David, Alan [GB/GB]; W.P. Thompson & Co., Celcon House, 289-293 High Holborn, London WC1V 7HU (GB).

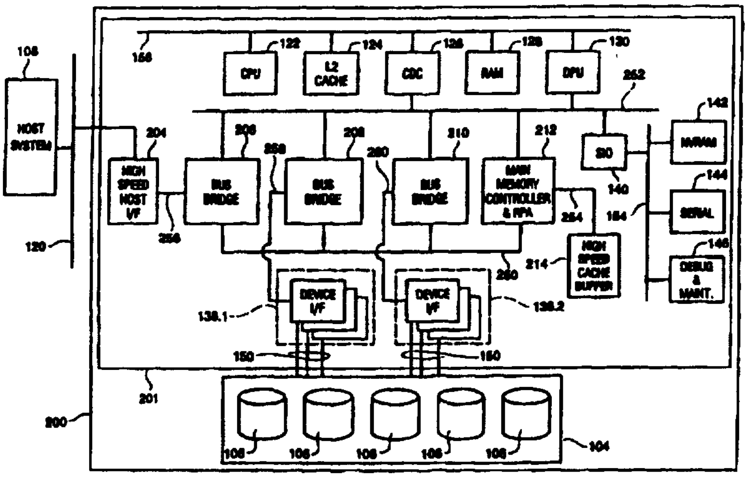
(72) Inventor: WEBER, Bret, S.; 2521 North Tee Time, Wichita, KS 67205 (US).

(74) Agent: GILL, David, Alan; W.P. Thompson & Co., Celcon House, 289-293 High Holborn, London WC1V 7HU (GB).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, 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, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: CACHE MEMORY CONTROLLER IN A RAID INTERFACE



(57) Abstract

The invention provides for a cache memory control architecture within a RAID storage subsystem which simplifies the migration and porting of existing ("legacy") control methods and structures to newer high performance cache memory designs. A centralized high speed cache memory (214) is controlled by a main memory controller circuit (212). One or more bus bridge circuits (206-210) adapt the signals from the bus architecture used by the legacy systems to the high speed cache memory (214). The bus bridge circuits (206-210) each adapt, for example, a PCI bus (256, 258, 260) used for a particular cache access purpose to the signal standards of an intermediate shared memory bus (250). The main memory controller circuit (212) adapts the signals applied to the intermediate shared memory bus (250) to the high speed cache memory bus (254). The hierarchical bus architecture permits older "legacy" control methods and structures to be easily adapted to newer cache memory architectures. In addition, the centralized high speed cache memory (214) and associated legacy system busses serve to distribute the load of cache memory access over simultaneously operable busses. The cache memory architecture of the present invention therefore permits rapid porting and re-usability of older "legacy" control methods and structures while permitting the overall cache memory performance to be scaled up to higher bandwidth demands of modern RAID subsystems.

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

CACHE MEMORY CONTROLLER IN A RAID INTERFACE

The present invention relates to a cache memory controller and in particular, but not exclusively, to a RAID storage subsystem having such a controller.

The present application is related to commonly assigned, co-pending, United States Patent Application serial number 08/357,847, filed December 16, 1994 by Stewart et al., entitled (as amended) **DISK ARRAY STORAGE SYSTEM ARCHITECTURE FOR PARITY OPERATIONS SIMULTANEOUS WITH OTHER DATA OPERATIONS**, which corresponds to EP-A-0,717,357 and which is hereby incorporated by reference.

Modern mass storage subsystems are continuing to provide increasing storage capacities to fulfill user demands from host computer system applications. Due to this critical reliance on large capacity mass storage, demands for enhanced reliability are also high. Various storage device configurations and geometries are commonly applied to meet the demands for higher storage capacity while maintaining or enhancing reliability of the mass storage subsystems.

A popular solution to these mass storage demands for increased capacity and reliability is the use of multiple smaller storage modules configured in geometries that permit redundancy of stored data to assure data integrity in case of various failures. In many such redundant subsystems, recovery from many common failures is automated within the storage subsystem itself due to the use of data redundancy, error codes, and so-called "hot spares" (extra storage modules which may be activated to replace a failed, previously active storage module). These subsystems are typically referred to as redundant arrays of inexpensive (or independent) disks (or more commonly by the acronym RAID). The 1987 publication by David A. Patterson, et al., from University of California at Berkeley entitled *A Case for Redundant Arrays of Inexpensive Disks (RAID),* reviews the fundamental concepts of RAID technology.

There are five "levels" of standard geometries defined in the Patterson publication.

The simplest array, a RAID level 1 system, comprises one or more disks for storing data and an equal number of additional mirror disks for storing copies of the information written to the data disks. The remaining RAID levels, identified as RAID level 2,3,4 and 5 systems, segment the data into portions for storage across several data disks. One of more additional disks are utilized to store error check or parity information. A single unit of storage is spread across the several disk drives and is commonly referred to as a "stripe." The stripe consists of the related data written in each of the disk drive containing data plus the parity (error recovery) information written to the parity disk drive.

RAID storage subsystems typically utilize a control module that shields the user or host system from the details of managing the redundant array. The controller makes the subsystem appear to the host computer as a single, highly reliable, high capacity disk drive. In fact, the RAID controller may distribute the host computer system supplied data across a plurality of the small independent drives with redundancy and error checking information so as to improve subsystem performance and reliability. Frequently RAID subsystems provide large cache memory structures to further improve the performance of the RAID subsystem. The cache memory is associated with the control module such that the storage blocks on the disk array are mapped to blocks in the cache. This mapping is also transparent to the host system. The host system simply requests blocks of data to be read or written and the RAID controller manipulates the disk array and cache memory as required.

Processing of I/O requests generated by a host computer requires significant data bandwidth transferring data to and from the cache memory. For example, data read from the disk array of the RAID storage subsystem into the cache memory may be transferred to a requesting host computer system to satisfy a read I/O request. Other data may be received from a host computer system and transferred to the cache memory to satisfy a write I/O request. Simultaneous with such host initiated transfers of data, background operations performed by the RAID controller of the RAID storage subsystem may post data in the cache memory to the drive array or read data from the disk array to the cache memory. All these exemplary data transfer operations between a host computer and the cache memory and between the disk array and the cache memory require a portion of the total available bandwidth on the bus to the cache memory.

As higher performance host computer connections have developed for connecting RAID storage subsystems to host computer systems, the data transfer bandwidth of the RAID cache memory subsystem has become a performance bottleneck. However, changing the architecture of the cache memory subsystem in the evolutionary design of RAID controllers can create problems in supporting older RAID controller software structures. The investment in older RAID controller software can be significant such that it is highly desirable to port the older control software up to newer RAID controller designs. Old system designs which require support (backward compatibility) in newer controller designs are often referred to as "legacy" systems.

If a RAID subsystem design engineer creates a new cache memory architecture to improve overall RAID performance, older legacy system control software and structures may be abandoned due to fundamental changes in the structure of, operation of, and interface to, the new cache memory subsystem. Redesign of the legacy system control algorithms and structures can add significant costs, complexity, and delay to the development and marketing of new RAID controller systems.

It is evident from the above discussion that an improved architecture for RAID controller cache memory design is required to improve performance of the cache memory subsystem while maintaining compatibility with legacy systems.

The present invention seeks to provide for a memory controller having advantages over known memory controllers.

Summary of the Invention

According to one aspect of the present invention there is provided a cache memory controller comprising main memory controller means, coupled to a central cache memory through a high speed memory bus, for controlling access to said central cache memory, bus bridge means, coupled to said main memory controller means through an intermediate shared memory bus and coupled to a controller specific bus, for accessing said central cache memory by converting signals exchanged on said controller specific bus into signals appropriate for exchange on said intermediate shared memory bus and by converting signals exchanged on said intermediate shared memory bus into signals appropriate for exchange on said high speed memory bus.

According to another aspect of the present invention there is provided a cache memory control architecture within a RAID storage subsystem which simplifies the migration and porting of existing ("legacy") control methods and structures to newer high performance cache memory designs. A centralized high speed cache memory is controlled by a main memory controller circuit. One or more bus bridge circuits adapt the signals from the bus architecture used by the legacy systems to the high speed cache memory. The bus bridge circuits each adapt, for example, a PCI bus used for a particular cache access purpose to the signal standards of an intermediate shared memory bus. The main memory controller circuit adapts the signals applied to the intermediate shared memory bus to the high speed cache memory bus. The hierarchical bus architecture permits older "legacy" control methods and structures to be easily adapted to newer cache memory architectures. In addition, the centralized high speed cache memory and associated legacy system busses serve to distribute the load of cache memory access over simultaneously operable busses. The cache memory architecture of the present invention therefore permits rapid porting and re-usability of older "legacy" control methods and structures while permitting the overall cache memory performance to be scaled up to higher bandwidth demands of modern RAID subsystems.

The present invention solves the above and other problems, and thereby advances the useful arts, by disclosing a RAID controller architecture which provides a high performance cache memory design while simplifying the task of porting (supporting) older "legacy" RAID control software methods. The present invention provides for a hierarchy of lower level interface busses to attach "legacy" RAID control software systems to a centralized high performance cache memory through an intermediate shared memory bus structure. The hierarchical levels of busses are used to distribute the load of the various data exchange and manipulation tasks involving the cache memory. The intermediate shared memory bus is designed to meet the higher performance requirements of modern cached RAID controllers and is shared by the lower level hierarchical busses in order to provide access to the high speed cache memory subsystem.

The present invention includes a main memory controller circuit which manages a wide (preferably 64-bit wide), high speed, cache memory and is adaptable to a wide variety of memory speeds, memory sizes, and memory error detection/correction methods. This main memory controller circuit includes a high performance RAID parity generation and detection portion to enable high speed parity computation functions in parallel with other access to the data stored in the high speed cache memory. Bus bridge circuits of the present invention are used to adapt older "legacy" software control methods and structures to the new high speed cache architecture. The bus bridge adapts signals applied to a bus used by the legacy subsystems to the signal standards of an intermediate shared memory bus. A specific bus bridge circuit is used for each specific type of legacy system cache memory architecture. The intermediate shared memory bus connects one or more bus bridge circuits to the main memory controller circuit. The legacy system control methods and structures may therefore be easily ported and re-used in the higher performance cache environment of the present invention.

The main memory controller circuit is configurable to manage various sizes, geometries, and speeds of memory devices which comprise the high performance cache memory. In a preferred embodiment of the present invention, the high speed cache memory managed by the main memory controller chip is 72-bits wide (64 data bits plus 8 bits for error correcting codes). The main memory controller circuit is also configurable to utilize a variety of memory error detection and correction techniques to permit scaling to newer memory error correction techniques.

The main memory controller circuit permits access to the high performance cache memory through an intermediate shared memory bus for shared connectivity to cache memory bus designs of legacy systems (such as 32 and 64 bit PCI busses). The intermediate shared memory bus provides for high performance shared access to the main memory controller circuit by the bus bridge circuits. Legacy systems are connected to the intermediate shared memory bus by bus bridge circuits of the present invention which adapt the legacy system's PCI bus (for example) to the intermediate shared memory bus of the main memory controller circuit.

Further, the intermediate shared memory bus may utilize a plurality of gigabit-per-second parallel to serial transceiver devices to provide a high speed bus connection with a minimum pin count. For example, eight such transceivers may be utilized to provide a "byte-wide serial" intermediate shared memory bus between the bus bridge circuits and the main memory controller circuit.

The main memory controller circuit and the bus bridge circuits also connect to a 32-bit processor bus independent of the intermediate shared memory bus used for high performance cache memory access. The central processor of the RAID controller connects to the 32-bit processor bus to control the operations of the main memory controller circuit, the bus bridge circuits, and all circuits connected to the legacy bus through the bus bridge circuits. The processor can therefore manage the operation of the bus bridge circuits and the main memory controller circuit without interfering with the operation of the intermediate shared memory bus and associated access to the high performance cache memory. This feature of the present invention enables the main processor to control operations on one bus (the legacy bus and the processor bus) while DMA transfers to the high performance cache memory take place simultaneously on another independent bus (the intermediate shared memory bus).

The circuits of the present invention thereby implement a hierarchical cache memory management structure in which RAID control subsystems, representing a potentially heterogeneous mix of current and legacy control architectures, may share access to a common high performance cache memory subsystem. Each RAID control subsystem connects to the a bus bridge circuit of the present invention which adapts that subsystem's particular cache memory bus architecture (e.g., 32 or 64 bit PCI) to a high performance intermediate shared memory bus architecture of the present invention. The intermediate shared memory bus connects each of the bus bridge circuits to a main memory controller circuit which, in turn, adapts the intermediate shared memory bus to the high performance cache memory. The structure and operation of the high performance cache memory of the present invention is therefore independent of any particular RAID controller architecture. The bus bridge circuits of the present invention provide any translation and adaptation of signals required to mate the cache memory bus of a legacy RAID controller to the high performance cache memory and the main memory controller circuit. This independence of the high performance cache memory from any particular legacy RAID control subsystem architecture permits the high performance cache memory to be updated and improved without impacting the support for older legacy RAID control subsystems. A 32-bit processor bus connects a central processing unit to each of the bus bridge circuits and to the main memory controller circuit to control their operation.

This RAID controller architecture is easily scaled up to higher performance cache memory designs by adapting the main memory controller circuit to the required performance enhancing designs. Older legacy subsystems are maintained through such enhancements by adapting the legacy subsystem's cache memory bus to the main memory controller circuit via a bus bridge circuit.

The present invention can therefore provide a RAID cache memory architecture with improved data bandwidth capacity.

Further the present invention can advantageously provide a RAID cache memory architecture having a centralized high performance cache memory and intermediate high performance bus for the sharing of the centralized cache memory.

According to another advantage, the present invention can provide a RAID cache memory architecture having a main memory controller for the control of a centralized high performance cache memory system, a plurality of bus bridges for adapting other bus architectures to the main memory controller, and an intermediate bus connecting the plurality of bus bridges and the main memory controller.

Preferably, the present invention can provide for RAID cache memory architecture which reduces the complexity of re-using legacy control structures from previous RAID controllers.

The present invention also allows for a RAID cache memory architecture including a centralized high performance cache memory and at least one bus bridge for adapting a slower and/or smaller bus to the centralized high performance cache memory thereby enabling re-use of legacy control structures from previous RAID controllers.

More particularly, the present invention can provide for a RAID cache memory architecture including a centralized high performance cache memory, a plurality of bus bridges for adapting slower and/or smaller busses to the centralized cache memory, and an intermediate bus connecting the plurality of bus bridges and the main memory controller thereby enabling re-use of legacy control structures from previous RAID controllers.

The invention can also provide for a hierarchial memory interface which provides high bandwidth for RAID data transfer while permitting easy scalability to support slower "legacy" RAID controller software.

The invention is described further hereinafter, by way of example only, with reference to the accompanying drawings in which:

Fig. 1 is a block diagram of a typical RAID storage subsystem as known in the art in which each of a plurality of RAID controllers has a unique cache memory architecture;

Fig. 2 is a block diagram of a RAID storage subsystem and embodying centralized high performance cache memory control structures of the present invention;

Fig. 3 is a block diagram providing additional details of the structure of the main memory controller circuit of Fig. 2;

Fig. 4 is a block diagram providing additional details of the structure of the bus bridge circuit of Fig. 2;

Fig. 5 is a block diagram of one embodiment of the intermediate shared memory bus of the present invention as a 64-bit-wide parallel bus; and

Fig. 6 is a block diagram of another embodiment of the intermediate shared memory bus of the present invention as a 128-bit-wide parallel to byte serial bus.

Fig. 1 is a block diagram depicting a typical RAID storage subsystem 100 as is known in the art. RAID storage subsystem 100 is connected by its RAID controller 101 to one or more host systems 108 via bus 120. As is known in the art, bus 120 may commonly be a SCSI bus connection, a Fibre Channel connection, a network connection such as Ethernet or Token Ring, or any of several standard interconnections between host computer systems and storage peripheral devices.

Within RAID storage subsystem 100, a RAID controller 101 processes host computer system generated I/O requests to store and retrieve information from the RAID storage subsystem. RAID controller 101 manages the disk drives of the RAID subsystem to manage the redundancy features of the RAID geometries. RAID controller 101 is connected to RAID array 104 via bus 150. RAID array 104 is comprised of a plurality of storage medium devices such as disk drives 106. As is known in the art, bus 150, may be any of several standard busses utilized in the storage industry to communicate with storage medium devices such as disk drives 106. For example, SCSI, EIDE, IPI, Fibre Channel (FCAL), and other such busses and protocols are commonly used in the storage medium industry to communicate with disk drives. RAID controller 101 utilizes subsets of the disk drives 106 of the RAID array 104 to configure and manage one or more RAID devices on behalf of attached host computer systems.

RAID controller 101 includes CPU 122 in which the RAID control methods are operable. Standard integrated circuit chip sets, well known in the art, are typically utilized to connect CPU 122 to a second level cache (L2 cache 124) and to its main memory (RAM 128) for storage and retrieval of both data and instructions. Cache DRAM controller (CDC

126) and data path unit (DPU 130) exemplify such standard chip sets known (together with system I/O (SIO 140)) as the "Saturn II" chip set manufactured as part number 82420 by Intel Corporation.

CDC 126 and DPU 130 also serve to connect CPU 122 to a standard I/O bus (such as PCI bus 102) for connection with the peripheral I/O devices used by CPU 122. Specifically, SIO 140 connects standard lower speed devices used in initialization of the RAID controller 101 or used in the maintenance or development of the RAID controller 101. NVRAM 142, serial port 144, and debug/maintenance port 146 may be used ,for example, to permanently store the operational programmed instructions for CPU 122 (typically copied to RAM 128 for faster fetch and execution), and to communicate maintenance and debug information to service or design engineers.

Host interface 132, RAID parity assist (RPA) 134, and device interfaces 138 all attach to PCI bus 102 for the exchange of data there between via the common PCI bus 102. Host interface 132 provides a connection between RAID controller 101 and the host connection bus 120 to exchange data between PCI bus 102 and an attached host computer 108. As noted above, connection bus 120 may be a SCSI bus, Fibre Channel, network, or other common connection standards between host computers and peripheral storage devices.

RPA 134 provides connectivity between PCI bus 102 and cache buffer 136 (via bus 152) and provides parity computation assistance for the storage and retrieval of data in cache buffer 136. Host write requests are typically completed by writing the requested data into the cache buffer 136 for later posting to the RAID array 104. Device interfaces 138 connect the disk drives 106 of RAID array 104 to PCI bus 102 for the storage and retrieval of information therefrom.

The RAID parity assist 134 operates on data stored in a local memory (not shown) associated therewith. It calculates parity a burst at a time utilizing an internal 128 byte FIFO (not shown) to store the intermediate results. The intermediate results are not written back to local memory. For maximum performance the control logic for the parity assist engine maintains pointers to each needed data block. The RAID parity assist 134 operates at the full speed of the local memory bus to provide the fastest possible performance for the memory bandwidth.

The parity assist engine contains 4 separate sections that allow additional tasks to be queued up while one task executes. Each task maintains its own control and status register so that task scheduling does not interfere with the currently executing task. Several configuration options are provided to tailor the RPA 134 to the array organization. The engine can be configured as a single engine which works on wide arrays up to 22+1 drives wide, or as a four engine machine which operates with arrays as wide as 4+1 drives. An intermediate mode provides a two engine machine with up to 10+1 drives.

The parity engine includes exclusive-OR logic providing RAID 5 and RAID 3 parity generation/checking as well as move and zero check modes.

One of the most important parts of the RPA architecture is the time independent operation it provides. Blocks of data are not required to be accessed simultaneously for parity calculations. Disk operations which span several drives may be scheduled and executed as soon as possible by each device. Unrelated disk operations may continue even though all drive operations for a single task are not yet complete. This independence improves the performance of the slowest part of the system, the disk drive. It also simplifies the software task of managing concurrent hardware resource requirements.

Further details regarding the structure and operation of RPA 134 are contained in the aforementioned US and European patent applications.

The principle operational purpose of CPU 122 is to exchange data between an attached host computer system 108 and the RAID array 104 through cache buffer 136 in satisfaction of host computer generated I/O requests. As can bee seen from the above discussed structure, all such exchanges of data pass through PCI bus 102. Specifically, host generated I/O requests typically cause the exchange of data between cache buffer 136 and host computer 108 through PCI bus 102 (and through host interface 132 and RPA 134). Likewise, later posting of data by cached RAID controller 101 typically causes the exchange of data between cache buffer 136 and RAID array 104 through PCI bus 102 (and through RPA 134 and device interfaces 138.

As performance requirements have continued to rise, concurrent operations through PCI bus 102 involving the exchange of data in cache buffer 136 have saturated the bandwidth capacity of PCI bus 102. As is known in the art, PCI bus 102 may be replaced or enhanced by higher performance bus architectures. However, fundamental architectural changes in the bus structure often require significant, costly changes in the control methods (the control software) operable within RAID controller 101.

Fig. 2 is a block diagram describing the structure of a RAID storage subsystem 200 in which the present invention is applied to improve the cache memory bandwidth capacity while minimizing the need for costly changes in existing "legacy" control systems of the RAID controller. RAID controller 201 within RAID subsystem 200 is similar in many respects, other than the cache memory architecture, to RAID controller 101 of RAID subsystem 100 of FIG. 1. CPU 122, L2 cache 124, CDC 126, RAM 128, and DPU 130 of FIG. 2 are connected through processor bus 156 as described above with respect to FIG. 1. Other system I/O peripheral devices are connected to the processor bus 252 via System I/O (SIO) 140 and bus 154. NVRAM 142, serial port 144, and debug/maintenance port 146 perform similar functions to that described above with respect to FIG. 1. One of ordinary skill in the art will readily recognize that several equivalent chip sets are available to connect CPU 122, L2 cache 124, and RAM 128 to processor bus 252. DPU 130 and CDC 126 are exemplary of such chip sets presently commercially available from Intel and other integrated circuit manufacturers.

High speed cache buffer 214 comprises a memory array which is both fast and wide to provide adequate bandwidth capacity for aggregated cache buffer accesses in the exchange of information. Main memory controller 212 manages the memory array of high speed cache buffer 214 exchanging data therewith via bus 254. Intermediate shared memory bus 250 connects all other components involved in the exchange of data with high speed cache buffer 214 via main memory controller and RPA 212 and bus 254. All exchange of data between high speed cache buffer 214 and other components of RAID subsystem 200 are managed by the centralized main memory controller 212 and provided thereto via intermediate shared memory bus 250. Main memory controller 212 is therefore isolated from the specific protocols and bus structure of each component within RAID subsystem 200 wishing to exchange data with high speed cache buffer 214 and may therefore be adapted in future evolutionary designs to further enhance bandwidth capacity of the high speed cache buffer 214.

Other components within RAID subsystem 200 requiring the exchange of data with high speed cache buffer 214 through intermediate shared memory bus 250 are connected thereto through a bus bridge. Each bus bridge 206, 208, and 210 adapts the signals applied to their respective, unique, connected bus architecture to the intermediate shared memory bus 250. Specifically, bus bridge 206 adapts signals on bus 256 (e.g., a 32-bit or 64-bit PCI bus) to appropriate signals on intermediate shared memory bus 250 (and vice versa) for purposes of exchanging data between an attached host computer 108, through bus 120 and high speed host interface 204, and the high speed cache buffer 214 under the control of main memory controller 212. In like manner, bus bridges 208 and 210 adapt signals on busses 258 and 260, respectively, for application to intermediate shared memory bus 250 (and vice versa). Bus bridges 208 and 210 thereby provide for the exchange of information between RAID array 104 (through device interfaces 138.1, 138.2) and high speed cache buffer 214 under the control of main memory controller 212.

A first feature of the present invention is that busses 256, 258, and 260 may be implemented for compatibility with older "legacy" control systems operable within RAID controller 201. Legacy systems that depended upon, for example, a 32-bit PCI bus for access to device interfaces 138.1 and 138.2 may be utilized within the RAID controller 201 with the cache architecture of the present invention. Use of such legacy systems reduces the costs and complexity of redesigning new systems compatible with a new cache memory architecture.

A second feature of the present invention is the enhanced aggregate bandwidth for data exchange realized by the plurality of busses hierarchically arranged as described above. Specifically, a high speed cache memory bus architecture is supported by high speed cache buffer 214, bus 254, and main memory controller 212. This high speed cache architecture may be enhanced to utilize the highest performance memory designs available in a particular memory technology. The high speed cache may, for example, make use of DRAM, SRAM, SDRAM, EDRAM, EDO RAM, Burst EDO RAM, or RAMBUS architectures without impacting the portability of legacy control systems in the RAID controller 201. The intermediate shared memory bus 250 provides a high bandwidth shared access path for each of a plurality of legacy system supported busses to exchange data with the high speed cache. Finally, a plurality of legacy system supported busses (e.g., 32 and 64-bit PCI busses) may be intermixed and used, each connected to the intermediate shared memory bus 250 through a bus bridge, to distribute the data exchange bandwidth requirements over a plurality of busses. Each individual bus is therefore less likely to be saturated to its capacity for data exchange.

In particular, data exchange between an attached host computer 108 and the high speed cache buffer 214 may take place concurrently with data exchange between the high speed cache buffer 214 and the disk drives 106 of the RAID array 104. Each of these data exchange operations consumes bandwidth on its own legacy system supported bus (e.g., bus 256 and 258 or 260, respectively). Though the concurrent data exchanges will require use of the intermediate shared memory bus and the main memory controller 212 and bus 254, these components are designed to maximize bandwidth capacity. The cache memory architecture of the present invention is therefore less likely to saturate available data exchange bandwidth.

Main memory controller 212 also may provide RAID parity assist logic (RPA) to provide centralized high speed assistance for the generation and checking of RAID redundancy information. It is common in many RAID storage subsystem to utilize exclusive or (XOR) parity generation and checking for the redundancy information. As used herein, redundancy information includes well known exclusive or parity techniques as well as other encoding methods used to generate and check information used to regenerate erroneous or missing stored data. This feature of the present invention serves to isolate the RAID parity transfers from the individual legacy bus structures as well as the shared memory bus 250. Parity computations are performed internally within RPA circuits of main memory controller 212. This further enhances the distribution of the memory access load from a plurality of legacy systems to a centralized high speed bus structure. The individual legacy busses are therefore less likely to be saturated by such parity accesses. Alternatively, the main memory controller 212 may be directed to allow the device attached to a bus bridge to manage all redundancy information generation and checking.

Processor bus 252 connects CPU 122 to the hierarchical cache memory architecture of the present invention to permit methods operable within CPU 122 to control the operations of the bus bridges 206, 208, and 210, and of the main memory controller 212. This aspect of the present invention enables methods operable within processor 122 to access an individual one of busses (e.g., 256, 258, and 260) without impacting continued high speed data access and processing by the other busses.

The higher bandwidth capacity of the cache memory architecture of the present invention permits the RAID controller 201 to be scaled up for higher performance while minimizing the needs for redesign of older legacy control systems within the RAID controller 201. For example, host interface 204 may be easily scaled up to Fibre Channel host system connections. The higher data bandwidth available in the cache memory architecture of the present invention provides sufficient bandwidth therefor while permitting easier re-use of legacy control systems in the controller. Likewise, device interfaces 138.1 and 138.2 may be scaled up to higher performance disk drive control architectures (e.g., fast SCSI, wide SCSI, SCSI 3, Fibre Channel (FCAL), etc.). The data bandwidth available in the cache architecture of the present invention enables the use of such high performance connections while minimizing the re-design necessary for legacy control structures.

Fig. 3 is a block diagram showing additional details of the main memory controller 212 of Fig. 2. Memory control 308 within main memory controller 212 receives address, data, and control information on intermediate shared memory bus 250. The address, data, and control information together define a requested operation within the attached high speed cache buffer (214 of FIG. 2). Memory interface 310 within memory control 308 manages the signals received from and applied to intermediate shared memory bus 250. RAM control 312 within memory control 308 generates appropriate RAM circuit controls and applies them to bus 254 to perform the requested operation in the attached high speed cache buffer (214 of FIG. 2). Memory control 308 may be configured to appropriately control a wide variety of RAM geometries and technologies including DRAM, SRAM, SDRAM, EDO RAM, etc.

In the preferred embodiment, memory control 308 implements a 64-bit wide memory bus on intermediate shared memory bus 250 and controls a 72-bit wide memory bus on bus 254 (64 data bits plus 8 ECC bits). In the preferred embodiment, both intermediate shared memory bus 250 and bus 254 are clocked at a frequency in excess of 60 megahertz (preferably 66 megahertz). In the preferred embodiment, intermediate shared memory bus 250 and bus 254 can therefore sustain a bandwidth in excess of 500 megabytes per second.

RAID parity assist engine 300 within main memory controller 212 performs all required RAID parity generation and checking as data is stored in and retrieved from the high speed cache buffer (214 of FIG. 2). The parity calculations are performed within the RAID parity assist engine 300 without imposing additional load on the various busses of the RAID controller architecture. In particular, the parity computations are performed with a minimum of access to the cache memory 214 by main memory controller 212. The computed parity data may be retrieved from the RAID parity assist engine 300 for direct storage in the high speed cache buffer 214 or the disk array 104, or may be retrieved by the CPU 122 via processor bus 252 for further computation and processing.

Processor bus interface 316 within main memory controller 212 comprises the connection logic and circuits to connect the main memory controller 212 to the processor bus 252. CPU 122 of FIG. 2 configures and controls the operation of main memory controller 212 via processor bus 252. CPU 122 may access high speed cache memory 214 of FIG. 2 through processor bus 252 and processor bus interface 316 of main memory controller 212. Memory arbitration logic 314 arbitrates access to bus 254 requested through intermediate shared memory bus 250 and through processor bus 252. Such memory arbitration and control functions, and commercially available circuits to perform these functions, are well known to those of ordinary skill in the art.

Fig. 4 is a block diagram providing additional detail of the structure of bus bridges 206, 208, and 210 of Fig. 3. Bus bridges 206, 208, and 210 exchange signals over their respective attached busses 256, 258, and 260, respectively. As noted above, busses 256, 258, and 260 may be, for example, 32 and 64-bit PCI busses to maintain structural compatibility with older "legacy" control systems within RAID controller 201 of Fig. 2. The signals exchanged over busses 256, 258, and 260 are adapted for exchange over intermediate shared memory bus 250. Each bus bridge, 206, 208, or 210, therefore adapts the legacy control system's preferred bus, 256, 258, or 260, respectively, to the intermediate shared memory bus 250 for access to the high speed cache buffer 214 of Fig. 2.

Within each bus bridge 206, 208, or 210, a PCI bus interface 400 manages the exchange of signals on the legacy control system's preferred bus 256, 258, or 260, respectively. PCI bus converter 404 converts the respective PCI bus signals to the signal levels and timing required for application to the intermediate shared memory bus 250. Memory interface 406 applies the converted PCI bus signals to the intermediate shared memory bus 250 as appropriate for control of the shared bus. Memory interface 406 performs all required bus arbitration and negotiation to share the intermediate shared memory bus among the plurality of bus bridges 206, 208, and 210. Conversely, signals received on the intermediate shared memory bus 250 by memory interface 406 are converted to equivalent PCI bus signals by a PCI bus converter 404 within the bus bridge to which the signals are directed. Finally, PCI bus interface 400 within the bus bridge then applies the converted signals to the connected bus 256, 258, or 260 as appropriate.

As is known in the art, selection signals in intermediate shared memory bus 250 are used to arbitrate the use of the bus to exchange converted signals between the intermediate shared memory bus 250 and a selected one of the plurality of bus bridges 206, 208, and 210.

Processor bus 252 is connected to bus bridges 206, 208, and 210 to permit a CPU (122 of FIG. 2) to control and configure the operation of the bus bridges. In addition, CPU 112 exchanges information and configures peripheral devices directly over the connected PCI busses 256, 258, or 260 vi the bus bridges 206, 208, or 210, respectively, without interfering with the exchange of data via intermediate shared memory bus 250. In a manner analogous to that described above, PCI bus converter 404 converts signals between levels and timing required for processor bus 252 as received and generated by processor bus interface 402, and signal levels and timing required for the connected PCI bus 256, 258, or 260 as received and generated by PCI bus interface 400 of the bus bridges.

Bus bridges 206, 208, and 210 convert all signals for exchange between the connected PCI bus 256, 258, or 260 and the intermediate shared memory bus 250 or between the connected PCI bus 256, 258, or 260 and the processor bus 252. This conversion includes all signal level and timing conversions as well as bus width conversions. Specifically, in the preferred embodiment, PCI busses 256, 258, and 260 may be either 32-bit or 64-bit wide PCI busses and may be operated at either 33 or 66 megahertz while the processor bus is a bus appropriate to the selected CPU (for example a 32-bit wide, 33 megahertz PCI bus) and the intermediate shared memory bus is preferably 64-bits wide operating at 66 megahertz. In this preferred embodiment, the shared memory bus 250 is preferably a parallel signal bus providing 64 bit wide data path (plus 8 bits of error correcting codes). In an alternative embodiment, the shared memory bus 250 is implemented as a 128 bit wide data path. To reduce the pin count, the 128 bit wide bus is implemented through use of a plurality of high speed parallel to serial transceivers. For example, eight, sixteen bit wide parallel to serial high speed transceivers may be used in both the bus bridge circuits 206, 208, and 210 and the corresponding memory interface circuits within main memory controller 212 to implement the shared memory bus 250.

Each bus interface, PCI bus interface 400, processor bus interface 402, and memory interface 406, may include FIFO devices as required for speed and bus width matching purposes. The FIFOs permit each bus interface component and the associated bus to operate at peak efficiency for extended bursts though exchanging data with a different bus architecture.

The processor bus interface 402 of the bus bridge circuit essentially implements a bus bridge to translate the signals exchanged between processor bus 252 and the external legacy system bus 256, 258, or 260. The memory interface 406 portion uses FIFO technologies to buffer the exchange of bursts of data between the legacy system bus 256, 258, or 260 and the shared memory bus 250. This buffering enables speed matching between the legacy system bus 256, 258, or 260 and the higher speed shared memory bus 250.

As noted above, the best presently known mode of implementing the intermediate shared memory bus of the present invention is as a 64-data-bit-wide parallel bus clocked at 66 megahertz. Fig. 5 is a block diagram depicting such a structure for intermediate shared memory bus 250 of Fig. 2. Intermediate memory bus 250 of Fig. 5 is connected to the memory interface (FIFO) 406 of bus bridge circuits 206, 208, or 210.

Fig. 6 is a block diagram of an alternative embodiment of the intermediate shared memory bus 250 of the present invention in which eight, high speed, 16-bit-wide parallel to bit serial transceivers 500 are integrated into memory interface 406. This structure provides a 128-bit-wide internal bus structure but an 8-bit-wide (byte) serial interface as applied to shared intermediate memory bus 250. This alternate embodiment of the present invention serves to reduce the external pin count required for interconnect of bus bridge circuits 206, 208, and 210 and main memory controller circuit 212 via intermediate shared memory bus 250. The transceivers 500 are capable of data transfer speeds of at least 1 gigabit per second each to thereby provide bandwidth of 1 gigabyte per second over intermediate shared memory bus 250.

While the invention has been illustrated and described in detail in the drawings and foregoing description, such illustration and description is to be considered as exemplary and not restrictive in character, it being understood that only the preferred embodiment and minor variants thereof have been shown and described and that the invention is not restricted to details of the foregoing embodiments.

CLAIMS

1. A cache memory controller (201) comprising:

a main memory controller (212);

a high speed memory bus (254) coupled to said main memory controller (212) and coupled to a memory bank (214);

an intermediate shared memory bus (250) coupled to said main memory controller (212);

a bus bridge (206-210) coupled to said intermediate shared memory bus (250) and coupled to a controller specific bus (256-260),

and arranged such that signals exchanged on said controller specific bus (256-260) are converted by said bus bridge (206-210) into signals appropriate for exchange on said intermediate shared memory bus (250) and wherein signals exchanged on said intermediate shared memory bus (250) are converted by said main memory controller (212) into signals appropriate for exchange on said high speed memory bus.

2. A controller as claimed in Claim 1, further comprising:

a processor bus (252) coupled to said main memory controller (212) and to said bus bridge (206-210); and

a processor (122) coupled to said processor bus (252) for controlling operation of said main memory controller (212) and for controlling operation of said bus bridge (206-210) and for exchanging information with devices coupled to said controller specific bus (256-260).

3. A controller as claimed in Claim 1 or 2, further comprising:

a plurality of bus bridges (206-210) each coupled to said intermediate shared memory bus (250) and each being coupled to a corresponding independent controller specific bus (256-260), wherein each of said plurality of bus bridges (206-210) is arranged to adapt signals exchanged on its said corresponding independent controller specific bus (256-260) to signals appropriate for exchange on said intermediate shared memory bus (250).

4. A controller as claimed in Claim 1, 2 or 3, and associated with a RAID storage wherein said main memory controller (212) includes:

a redundancy assist to generate redundancy information associated with the storage of data in said RAID storage subsystem and for using said redundancy information in the operation of said RAID storage subsystem to improve data integrity.

5. A cache memory controller (201) comprising:

main memory controller means (212), coupled to a central cache memory (214) through a high speed memory bus (254), for controlling access to said central cache memory (214);

bus bridge means (206-210), coupled to said main memory controller means (212) through an intermediate shared memory bus (250) and coupled to a controller specific bus (256-260), for accessing said central cache memory (214) by converting signals exchanged on said controller specific bus (256-260) into signals appropriate for exchange on said intermediate shared memory bus (250) and by converting signals exchanged on said intermediate shared memory bus (250) into signals appropriate for exchange on said high speed memory bus (254).

6. A controller as claimed in Claim 5, comprising:

processing means (122) coupled to said main memory controller means (212) and to said bus bridge means (206-210) through a processor bus (252), for controlling operation of said main memory controller means (212) and for controlling operation of said bus bridge means (206-210) and for exchanging information with devices coupled to said controller specific bus (256-260).

7. A controller as claimed in Claim 5 or 6, further comprising:

a plurality of bus bridge means (206-210) each coupled to said main memory controller means (212) through said intermediate shared memory bus (250) and each being coupled to a corresponding controller specific bus (256-260), for accessing said central cache memory (214) by converting signals exchanged on said corresponding controller specific bus (256-260) into signals appropriate for exchange on said intermediate shared memory bus (250).

8. A controller as claimed in Claim 5, 6 or 7, and associated with a RAID storage subsystem wherein said main memory controller means (212) includes:

redundancy assist means for generating redundancy information associated with the storage of data in said RAID storage subsystem and for using said redundancy information in the operation of said RAID storage subsystem to improve data integrity.

9. A controller as claimed in Claim 4 or 8, wherein said redundancy assist is controllably enabled and disabled.

10. A controller as claimed in any one of the preceding Claims, wherein said intermediate shared memory bus is capable of exchanging data between said bus bridge (206-210) and said main memory controller (212) at rates in excess of about one gigabytes per second.

11. A controller as claimed in Claim 10, wherein said intermediate shared memory bus (250) comprises a parallel signal bus.

12. A controller as claimed in Claim 11, wherein said intermediate shared memory bus (250) comprises a parallel signal bus having a data width of 64 bits.

13. A controller as claimed in Claim 10, wherein said intermediate shared memory bus (250) comprises:

a multi-bit-wide serial bus;

wherein said bus bridge means (206-210) includes a first plurality of parallel to serial transceivers each having a data transfer rate of at least about one gigabit per second; and

wherein said main memory controller means (212) includes a second plurality of parallel to serial transceivers each having a data transfer rate of at least about one gigabit per second.

14. A controller as claimed in Claim 13, wherein said multi-bit-wide serial bus comprises a 8-bit-wide serial bus and wherein said first plurality of parallel to serial transceivers and said second plurality of parallel to serial transceivers convert 64 parallel signals into signals exchanged over said 8-bit-wide serial bus.

15. A controller as claimed in Claim 13, wherein said multi-bit-wide serial bus comprises a 16-bit-wide serial bus and wherein said first plurality of parallel to serial transceivers and said second plurality of parallel to serial transceivers convert 128 parallel signals into signals exchanged over said 16-bit-wide serial bus.

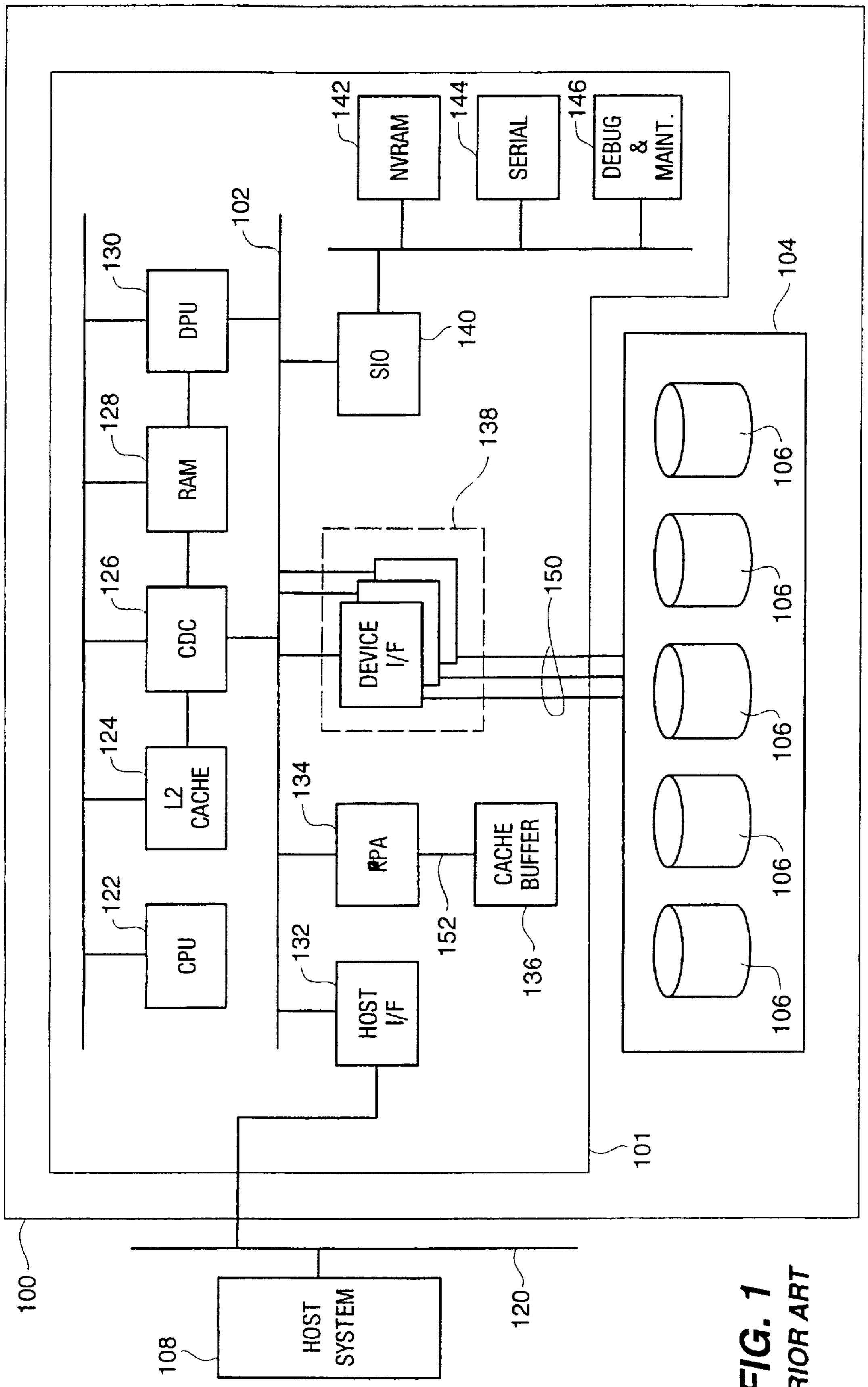


FIG. 1
PRIOR ART

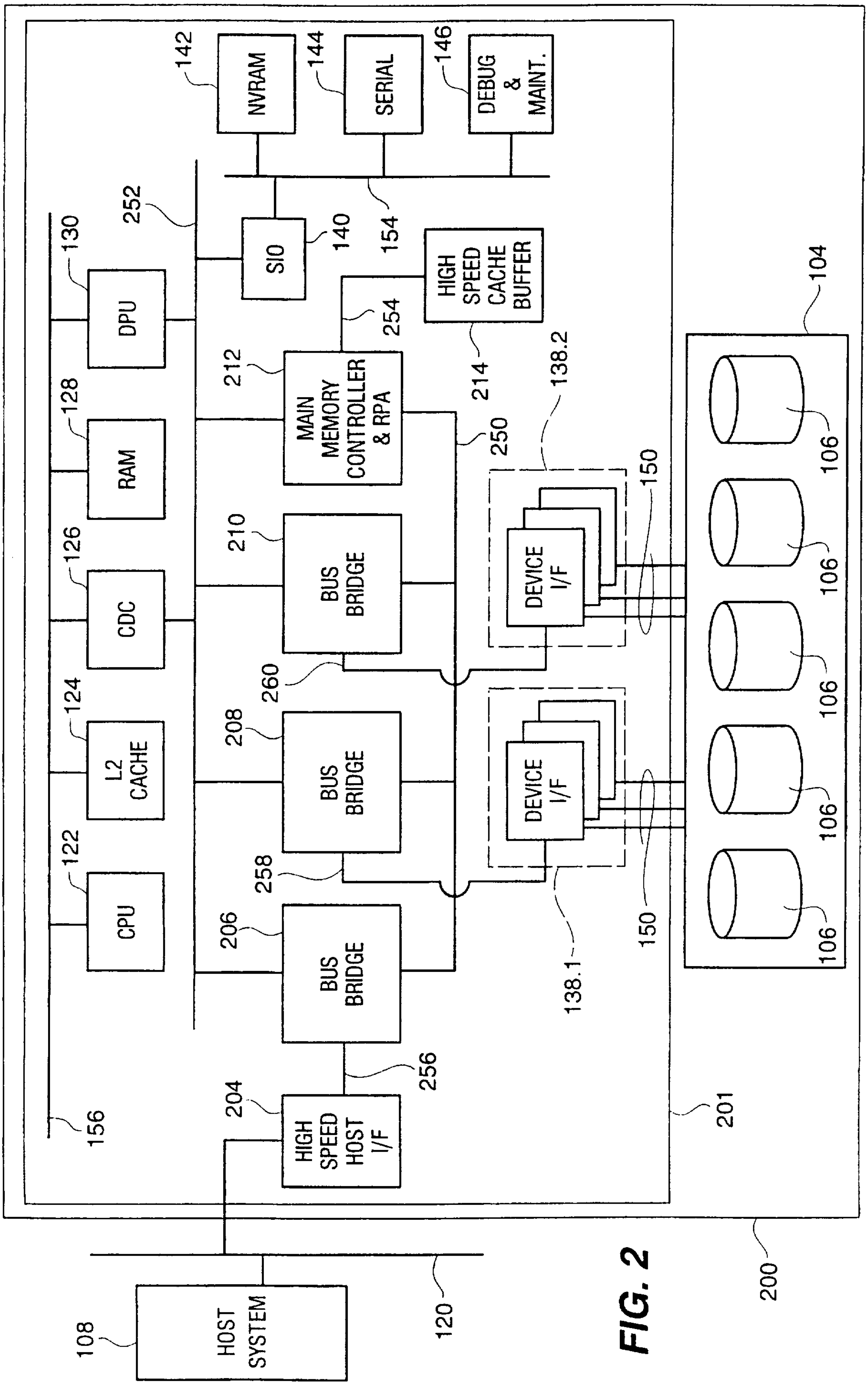

NVRAM
142
SERIAL
144
DEBUG & MAINT.
146
154
SIO
140
252
130
DPU
128
RAM
126
CDC
124
L2 CACHE
122
CPU
156
HIGH SPEED CACHE BUFFER
214
254
MAIN MEMORY CONTROLLER & RPA
212
250
BUS BRIDGE
210
260
BUS BRIDGE
208
258
BUS BRIDGE
206
256
HIGH SPEED HOST I/F
204
201
DEVICE I/F
138.2
DEVICE I/F
138.1
150
150
104
106
106
106
106
106
200
HOST SYSTEM
108
120

FIG. 2

FIG. 3

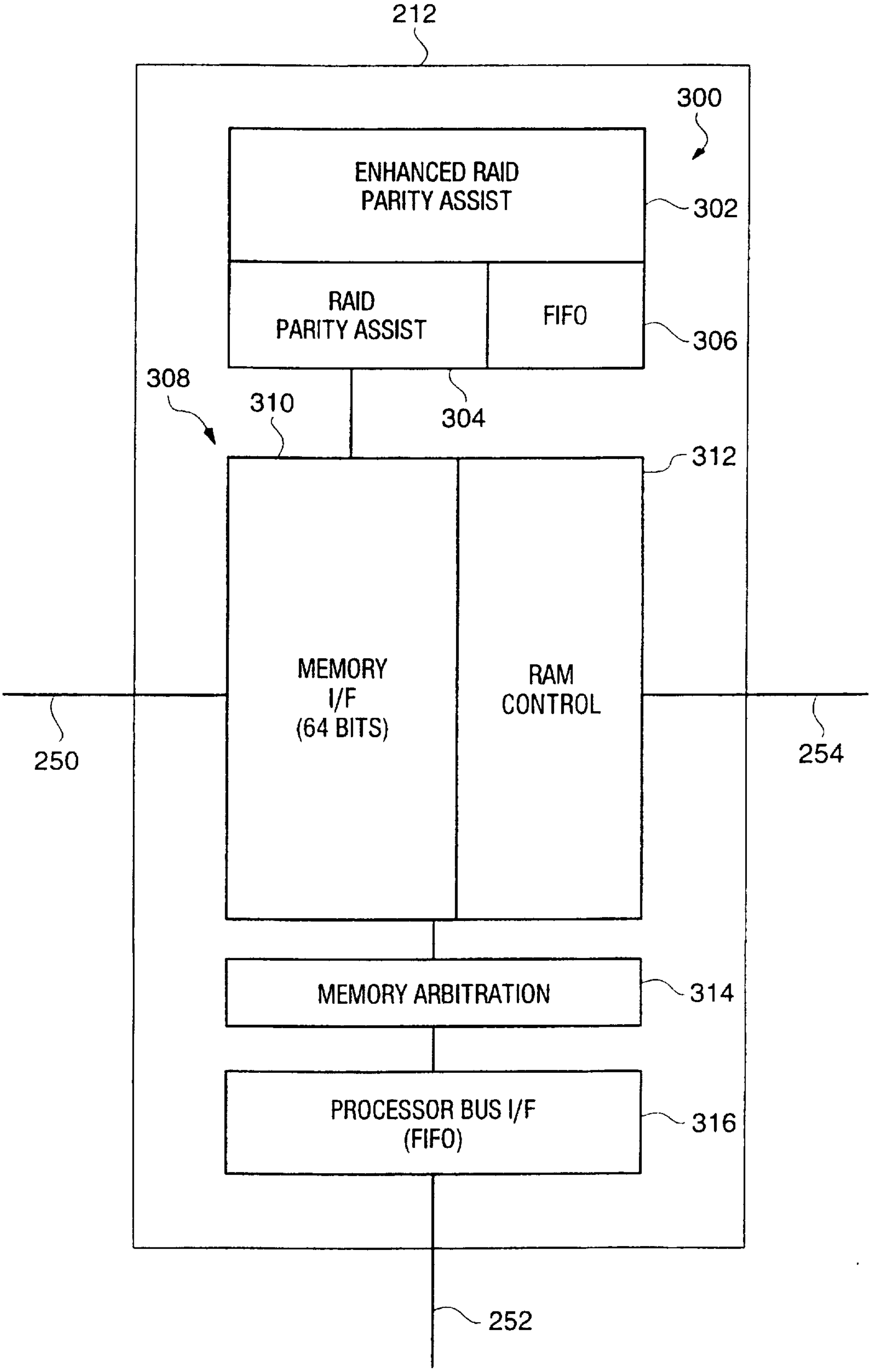

212
300
ENHANCED RAID PARITY ASSIST
302
RAID PARITY ASSIST
FIFO
306
308
310
304
312
MEMORY I/F (64 BITS)
RAM CONTROL
250
254
MEMORY ARBITRATION
314
PROCESSOR BUS I/F (FIFO)
316
252

FIG. 4

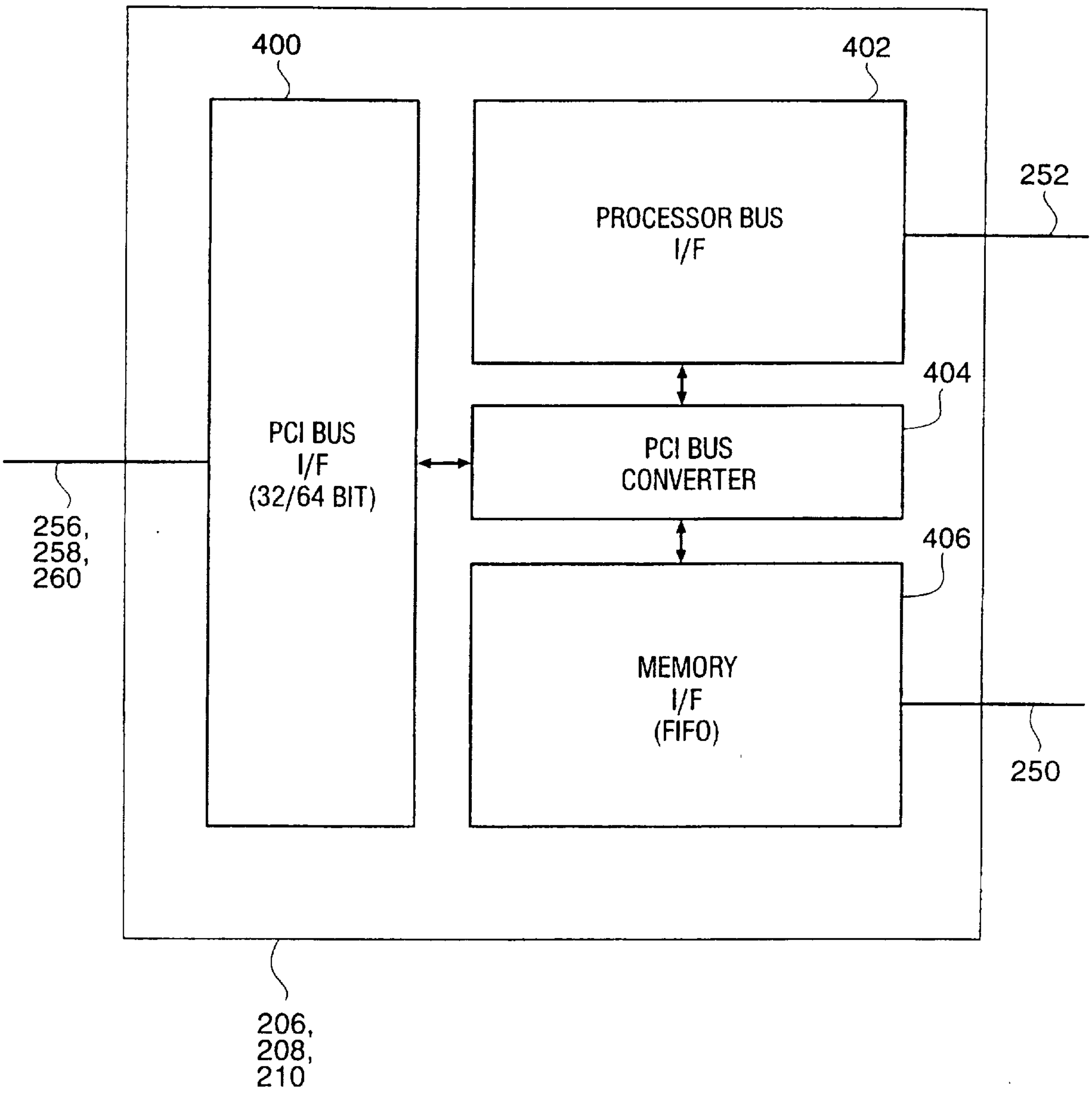

400
402
252
PROCESSOR BUS
I/F
404
PCI BUS
CONVERTER
PCI BUS
I/F
(32/64 BIT)
256,
258,
260
406
MEMORY
I/F
(FIFO)
250
206,
208,
210

FIG. 5

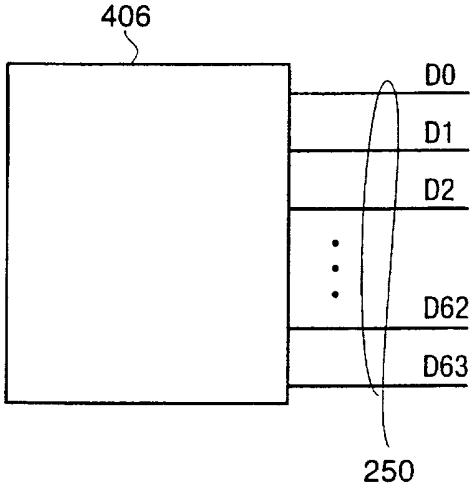
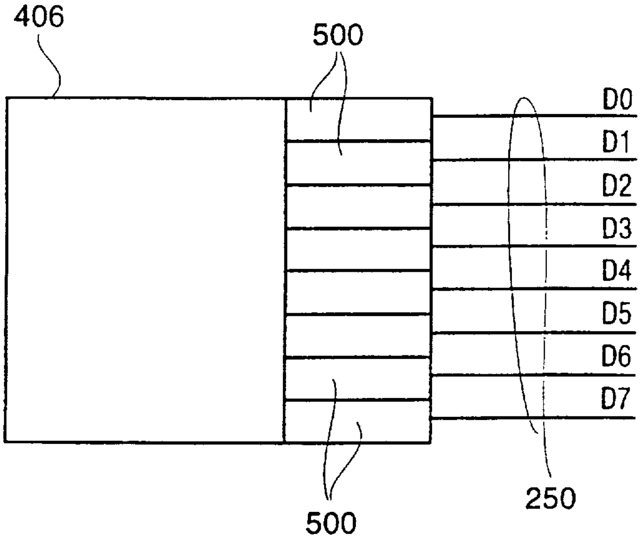


FIG. 6



INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 97/01474

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F3/06 G06F13/40

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)
IPC 6 G06F

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 379 384 A (SOLOMON) 3 January 1995	1,2,5,6, 10-15
Y	see column 1, line 14 - column 5, line 3; figures 1,2	3,7
Y	EP 0 631 241 A (IBM) 28 December 1994 see page 4, line 33 - page 6, line 43; figures 1,7	3,7
A	US 5 353 415 A (WOLFORD ET AL) 4 October 1994 see column 1, line 12 - column 9, line 64; figures 1-3	1,5
A	WO 93 14455 A (DI DATA LTD) 22 July 1993 see page 15, paragraph 2 - page 17, paragraph 4; figure 1	4,8
	-/--	

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

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

"E" earlier 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 when the document is taken alone

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

"&" document member of the same patent family

1

Date of the actual completion of the international search 3 December 1997	Date of mailing of the international search report 11/12/1997
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Gill, S

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 97/01474

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
P,X	EP 0 756 235 A (SYMBIOS LOGIC) 29 January 1997 see page 2, column 1, line 1 - page 5, column 7, line 57; figures 3,4 -----	1,2,4-6, 8-15

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/GB 97/01474

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5379384 A	03-01-95	NONE	
EP 631241 A	28-12-94	US 5542055 A	30-07-96
		BR 9402106 A	13-12-94
		CA 2124618 A	29-11-94
		JP 6348642 A	22-12-94
		KR 9708192 B	21-05-97
US 5353415 A	04-10-94	AU 5403394 A	26-04-94
		WO 9408297 A	14-04-94
WO 9314455 A	22-07-93	AU 3091592 A	03-08-93
		CA 2127380 A	22-07-93
		EP 0620934 A	26-10-94
		JP 8501643 T	20-02-96
		US 5526507 A	11-06-96
EP 756235 A	29-01-97	JP 9114596 A	02-05-97