



(19) **United States**

(12) **Patent Application Publication**

Fallon

(10) **Pub. No.: US 2003/0191876 A1**

(43) **Pub. Date:**

Oct. 9, 2003

(54) **DATA STOREWIDTH ACCELERATOR**

(60) Provisional application No. 60/180,114, filed on Feb. 3, 2000. Provisional application No. 60/333,919, filed on Nov. 28, 2001.

(76) Inventor: **James J. Fallon**, Armonk, NY (US)

Publication Classification

(51) **Int. Cl.⁷** **G06F 13/00**
(52) **U.S. Cl.** **710/68**

Correspondence Address:
Frank V. DeRosa
F. CHAU & ASSOCIATES, LLP
Suite 501
1900 Hempstead Turnpike
East Meadow, NY 11554 (US)

(21) Appl. No.: **10/306,581**

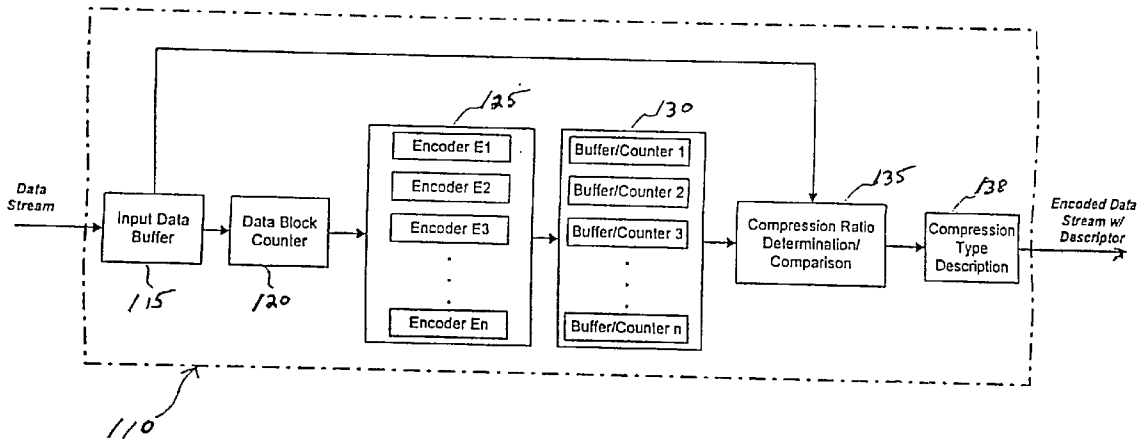
(22) Filed: **Nov. 27, 2002**

Related U.S. Application Data

(63) Continuation-in-part of application No. 09/775,905, filed on Feb. 2, 2001.

(57) **ABSTRACT**

Data storage controllers and data storage devices employing lossless or lossy data compression and decompression to provide accelerated data storage and retrieval bandwidth. In one embodiment of the invention, a composite disk controller provides data storage and retrieval acceleration using multiple caches for data pipelining and increased throughput. In another embodiment of the invention, the disk controller with acceleration is embedded in the storage device and utilized for data storage and retrieval acceleration.



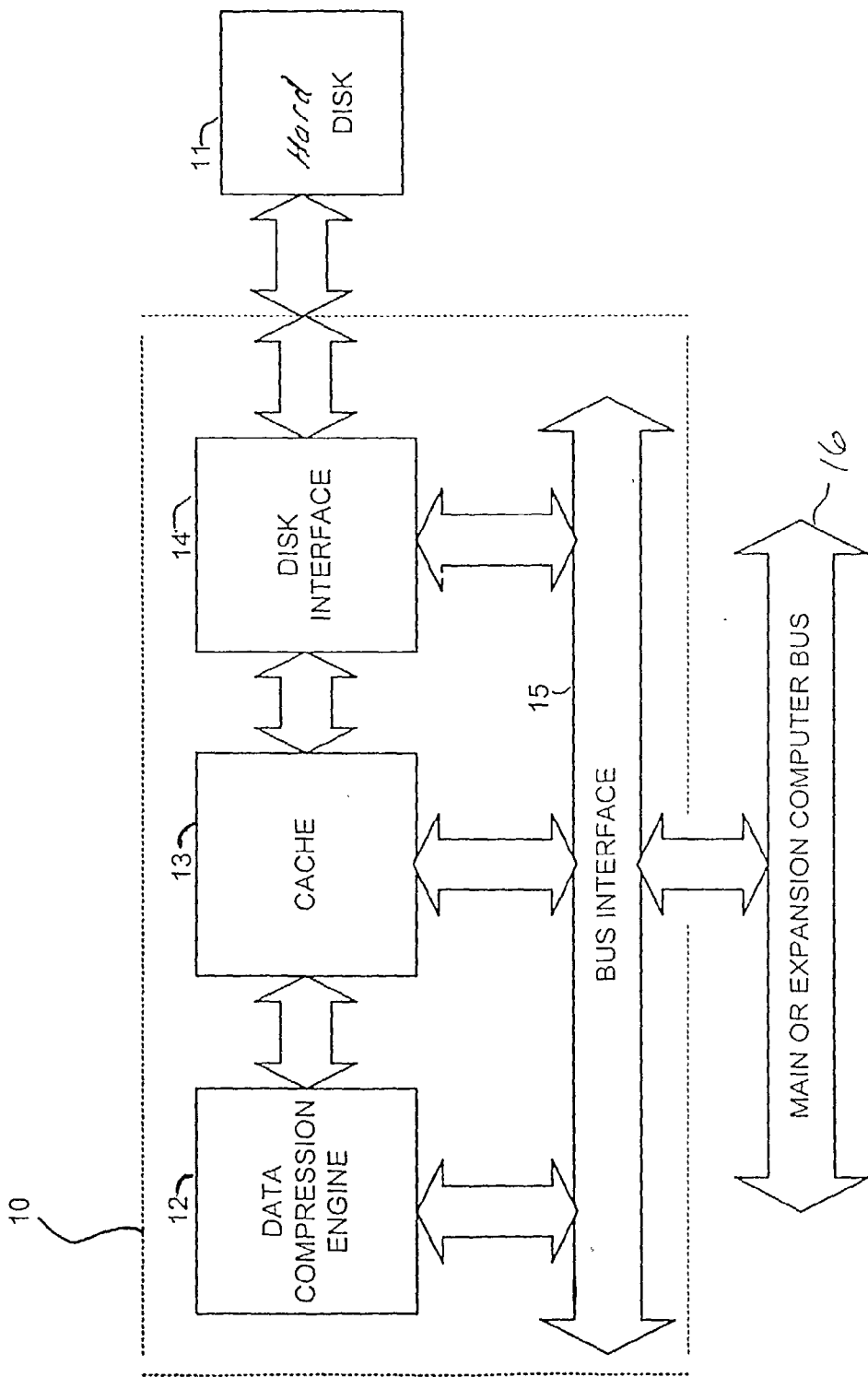


FIGURE 1

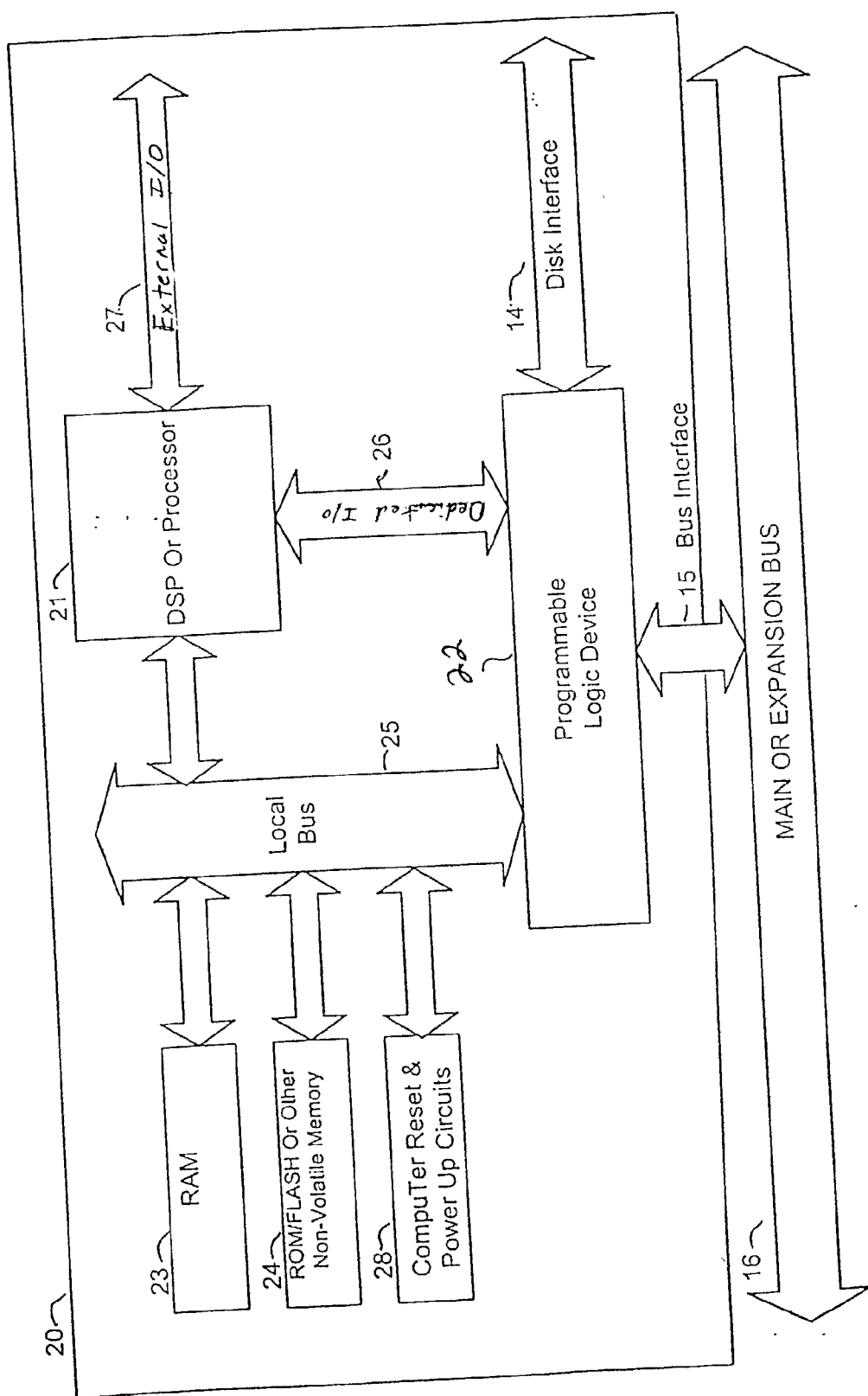


FIGURE 2

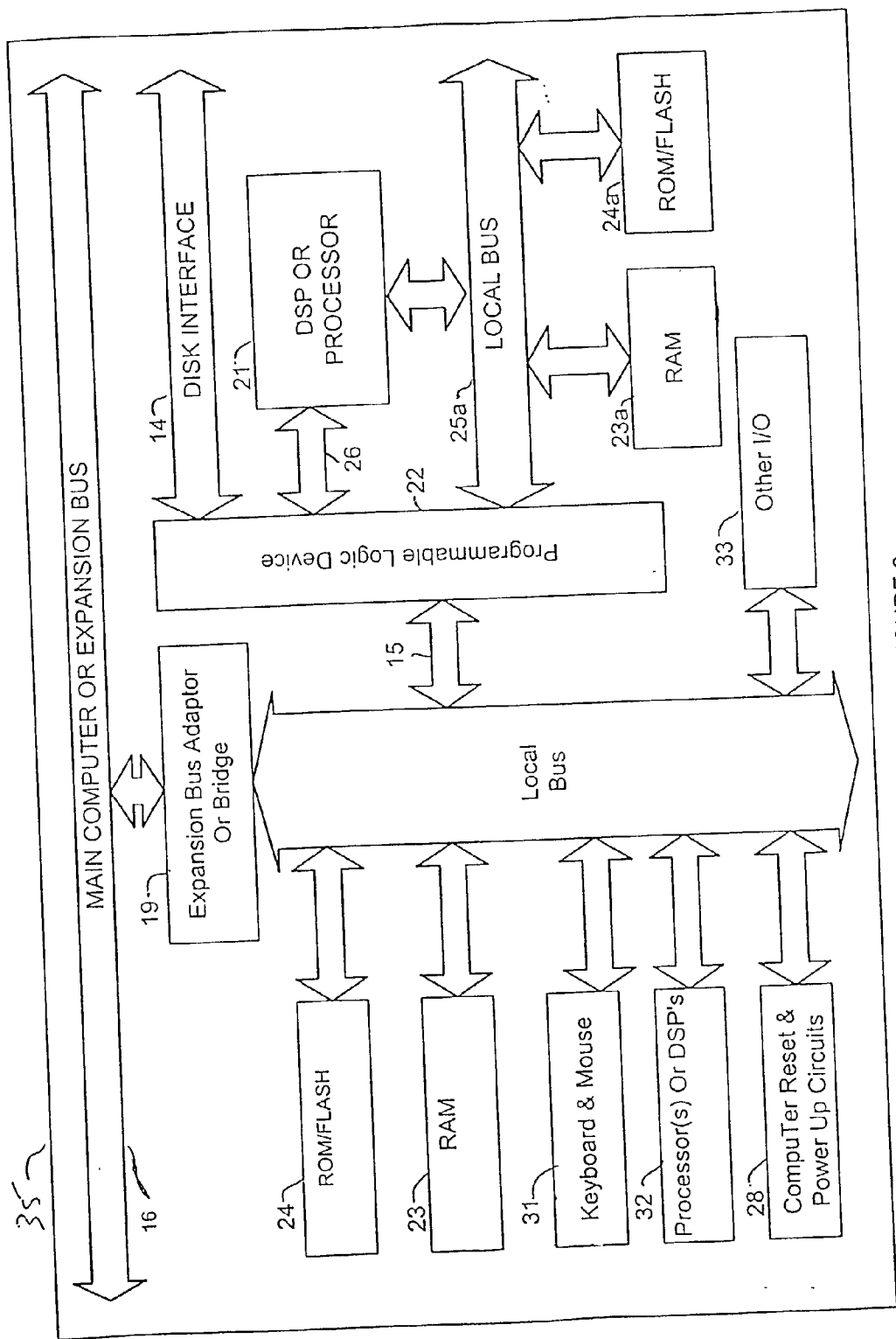


FIGURE 3

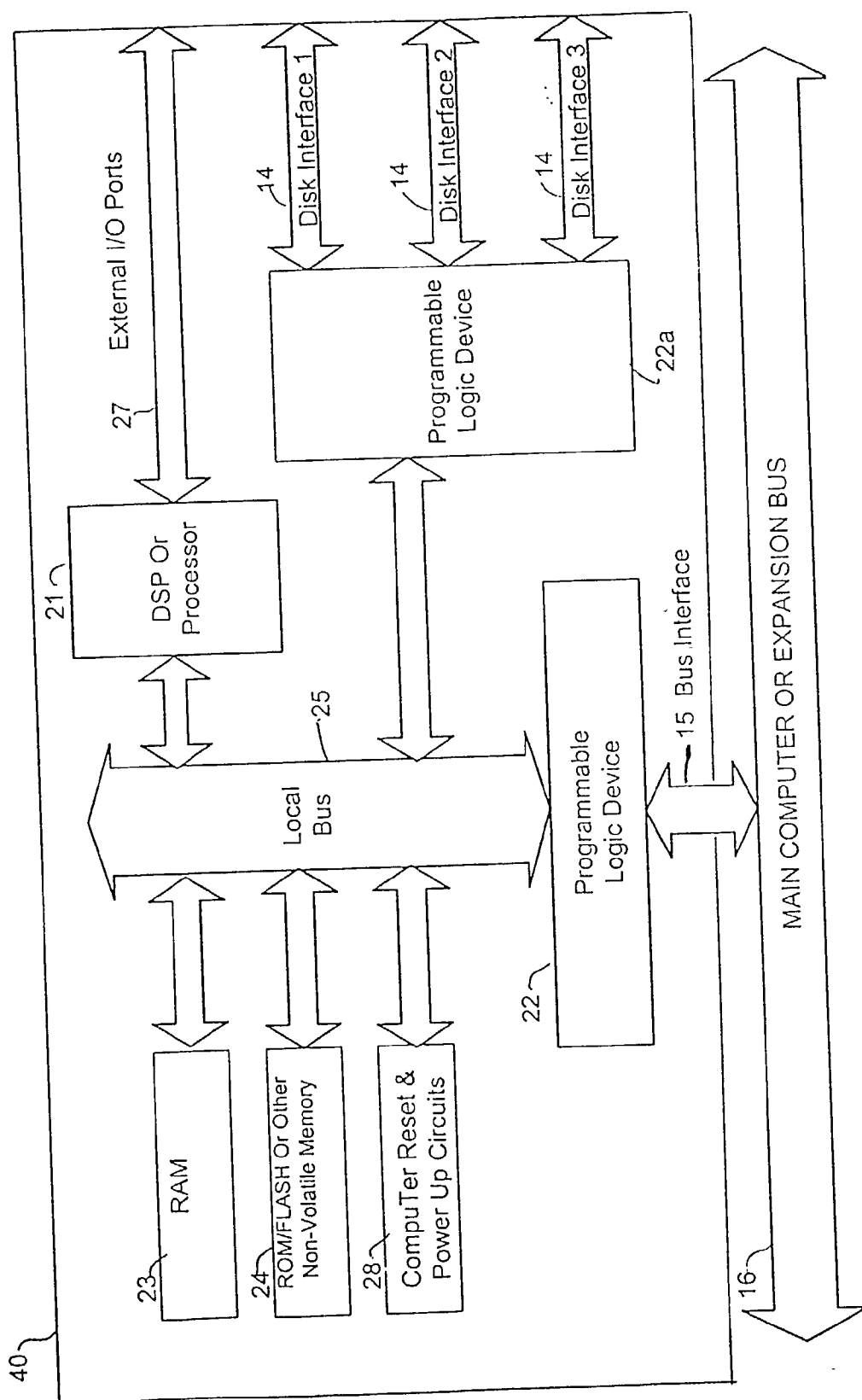


FIGURE 4

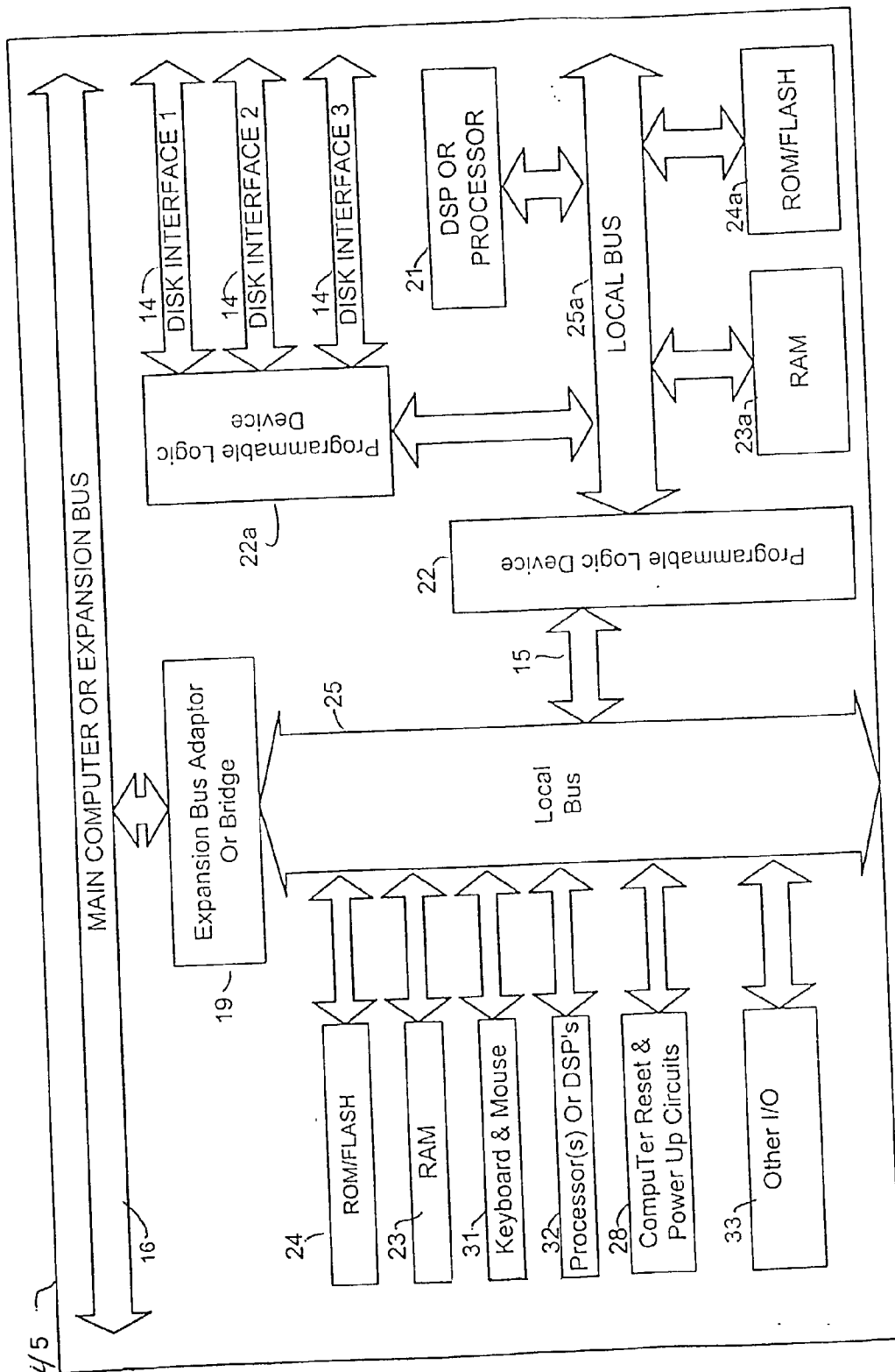


FIGURE 5

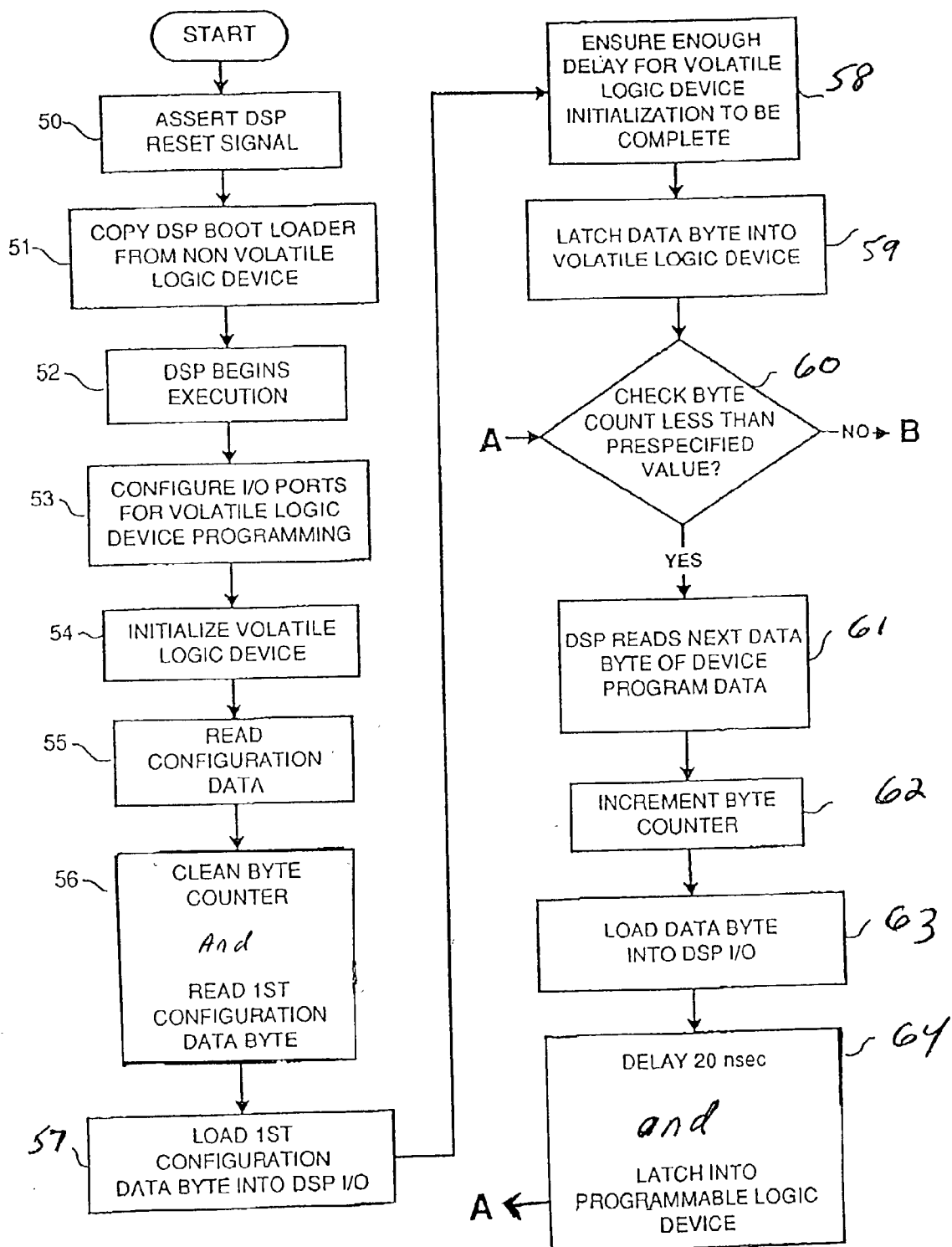


FIG. 6a

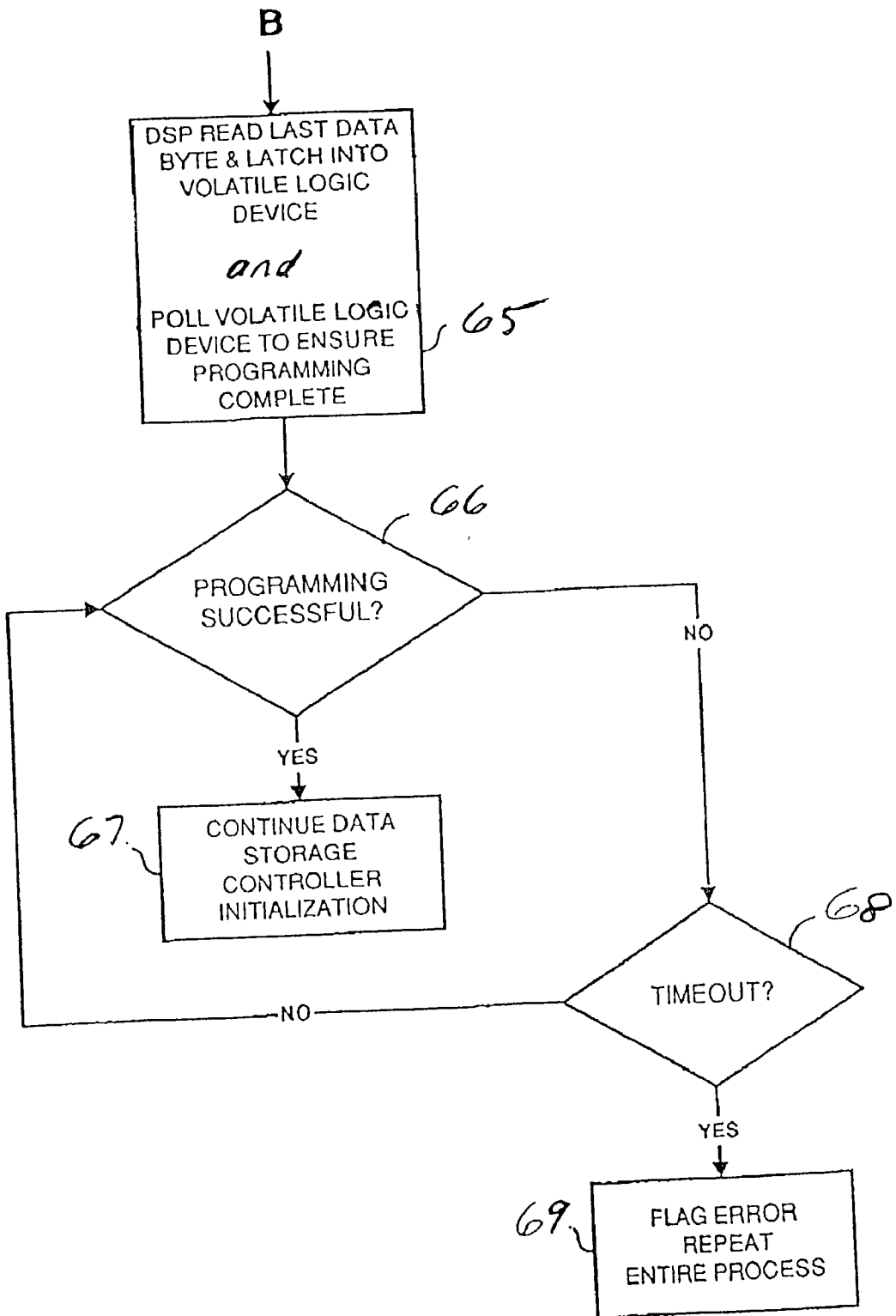


FIG. 6b

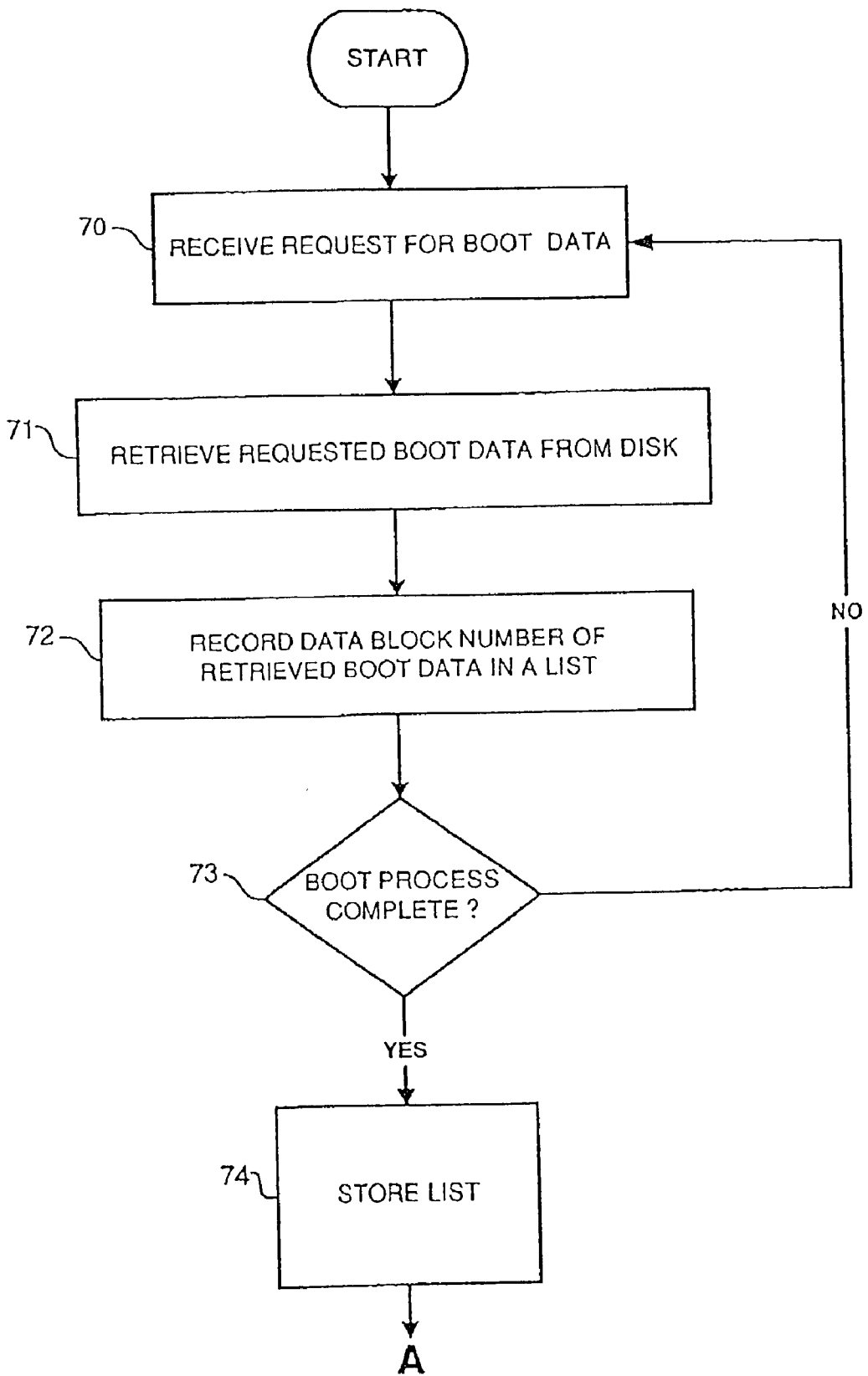


FIG. 7a

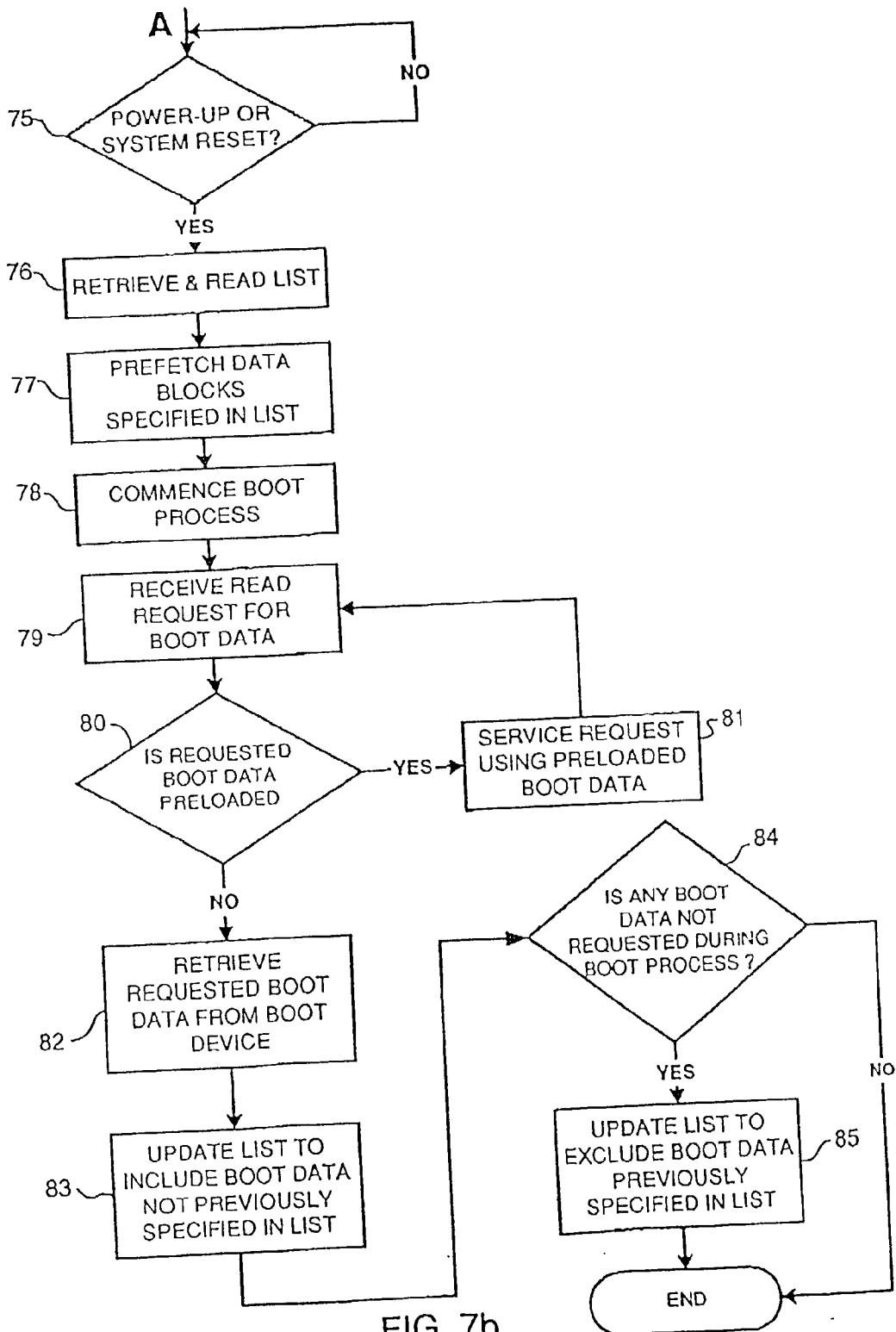


FIG. 7b

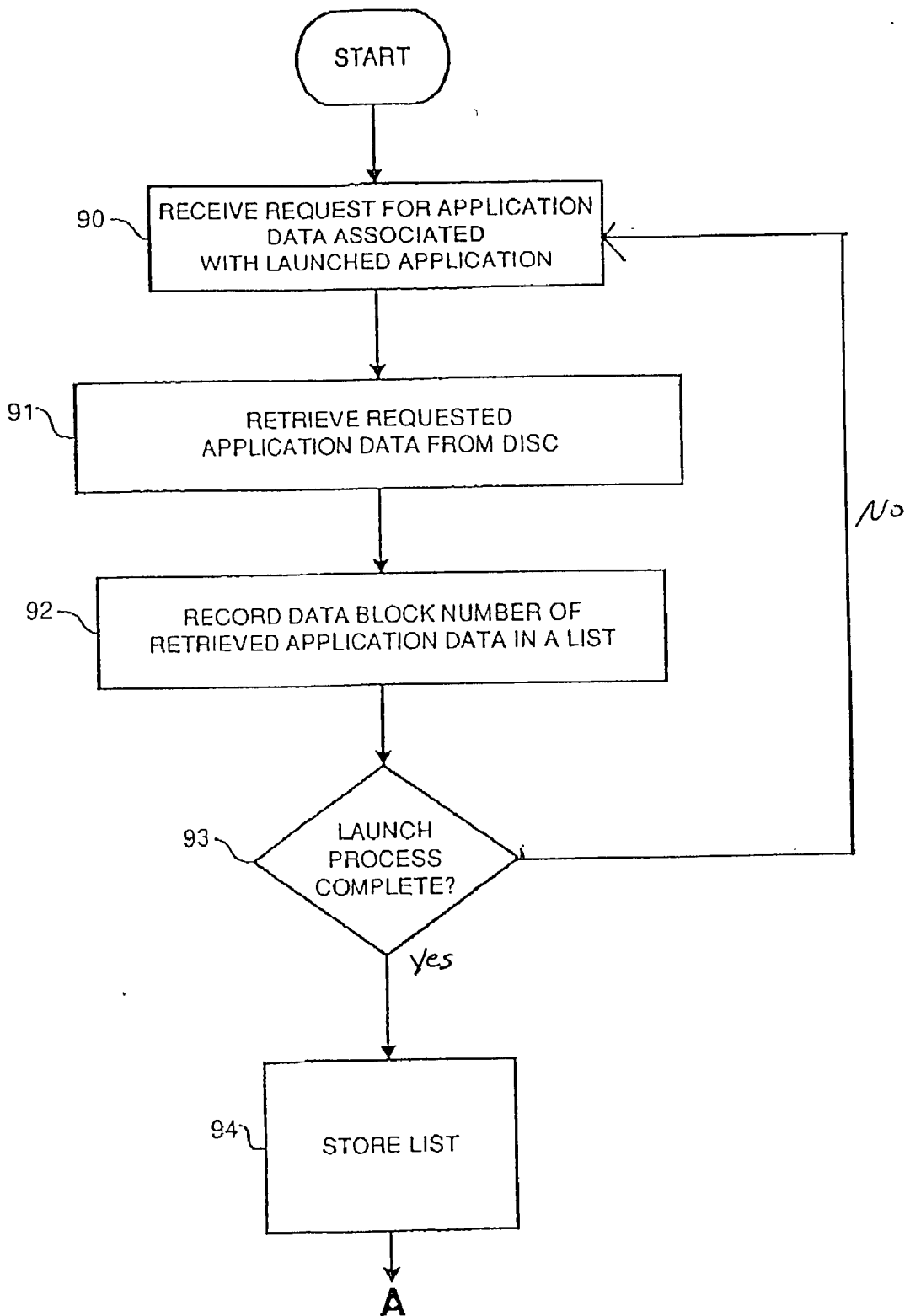
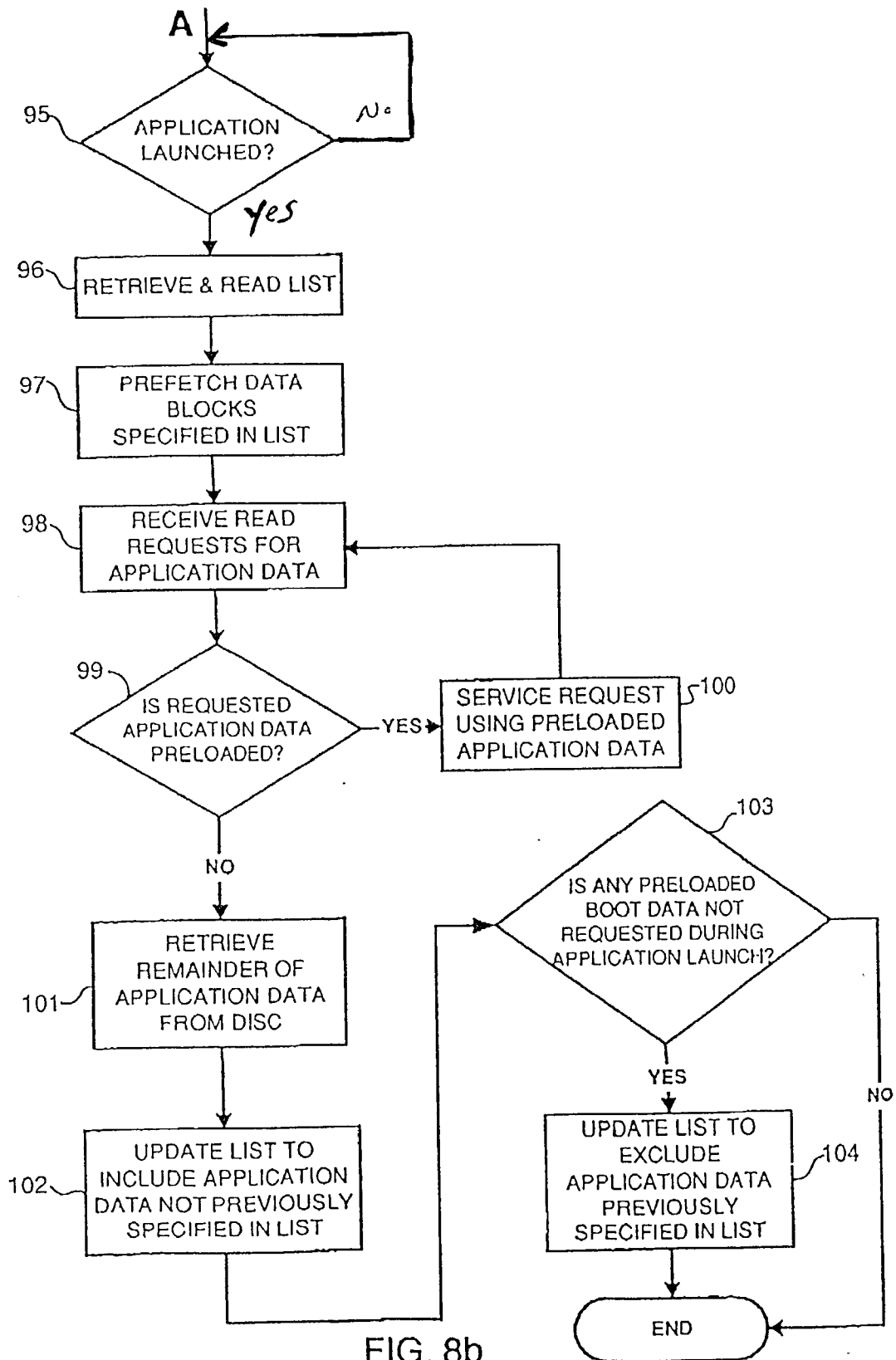


FIG. 8a



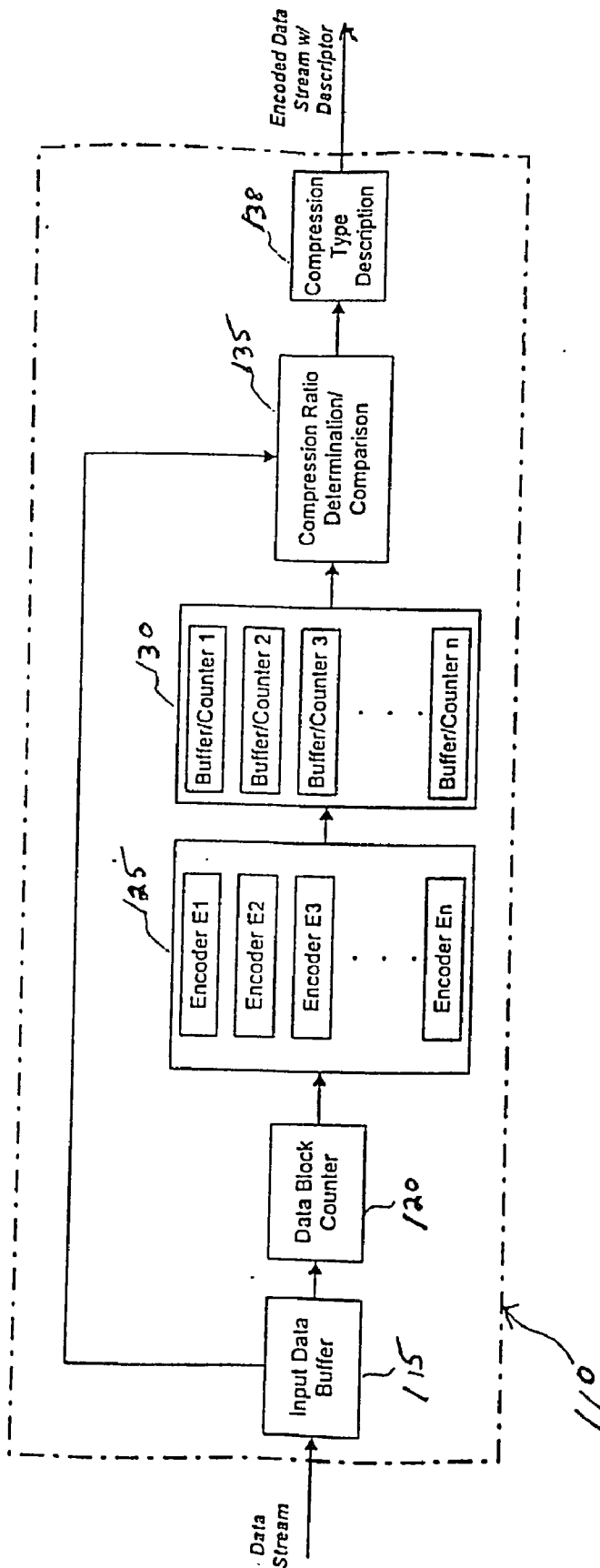


FIGURE 9

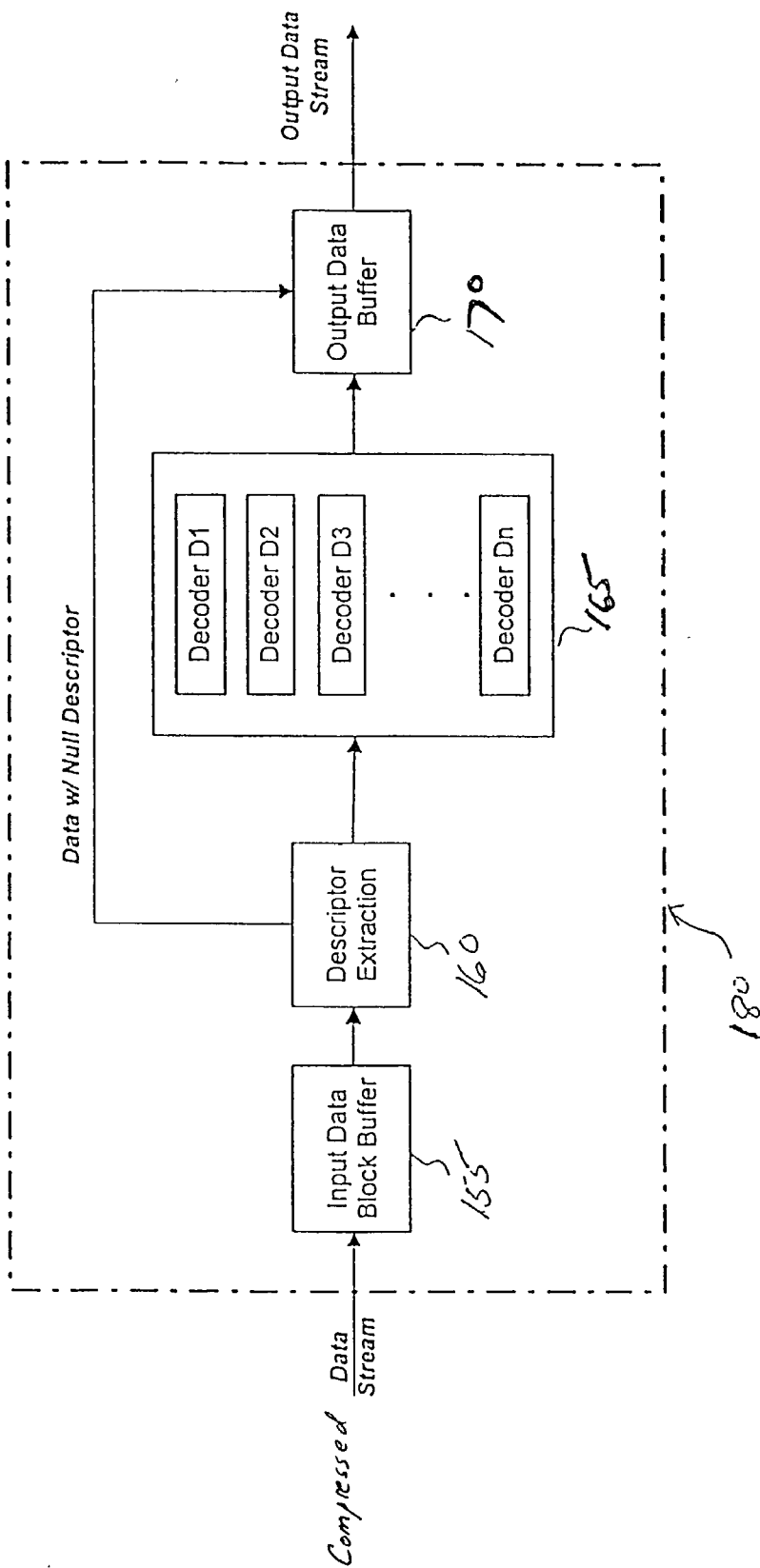


FIG. 10

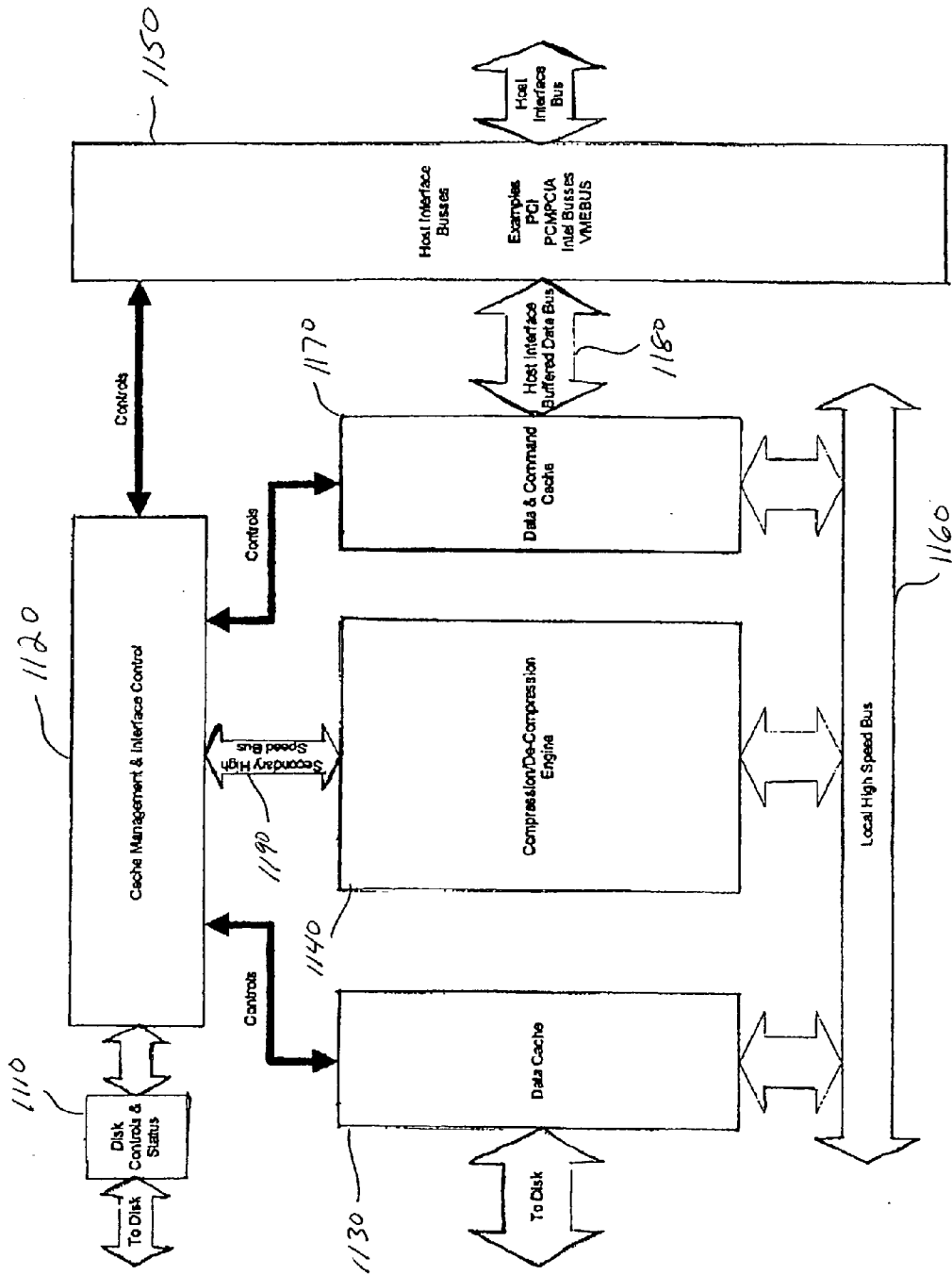


Fig. 11

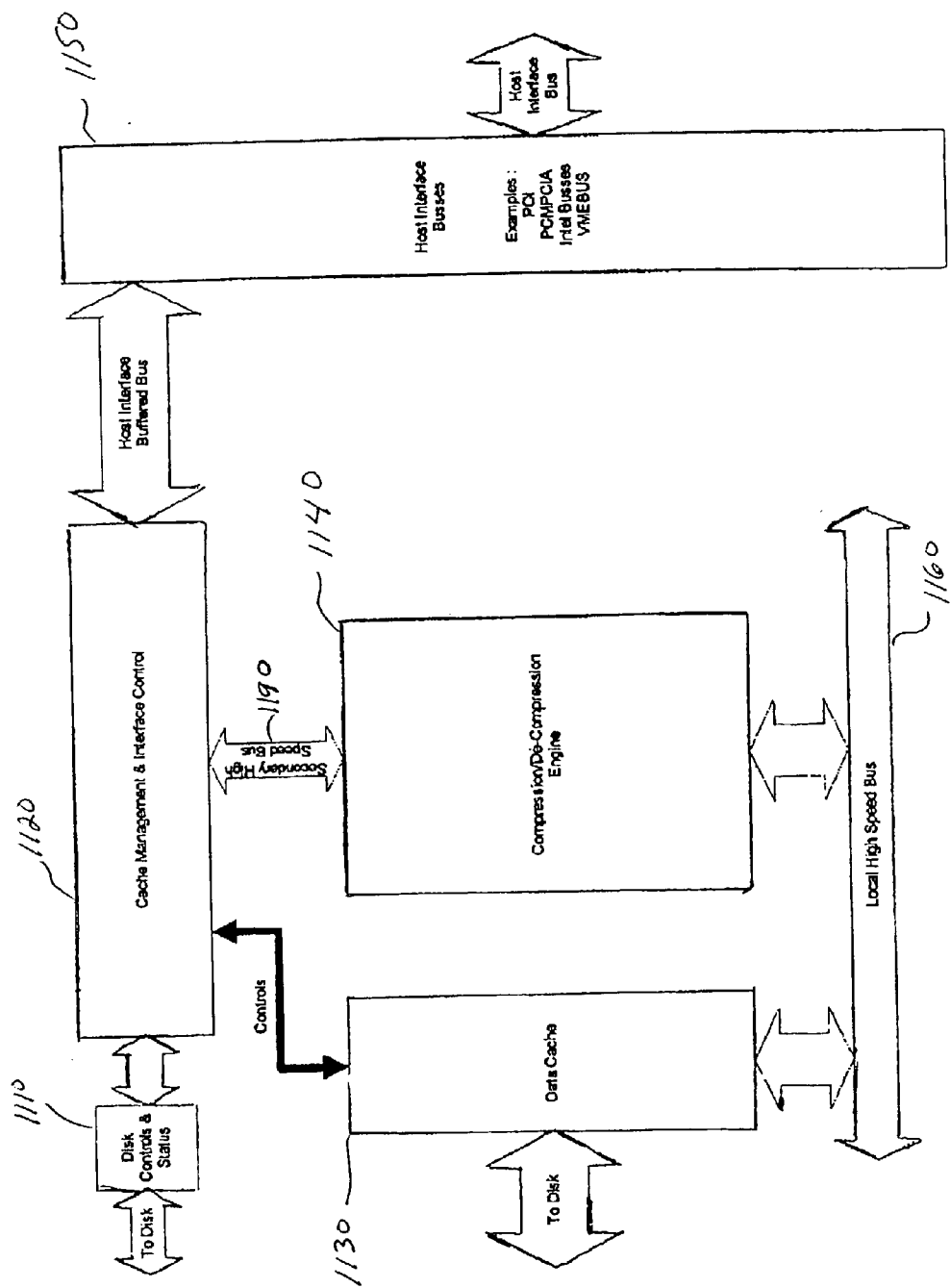


Fig. 12

DATA STOREWIDTH ACCELERATOR

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a Continuation-in-Part of U.S. patent application Ser. No. 09/775,905, filed Feb. 2, 2001, which claims priority to U.S. Provisional Application Serial No. 60/180,114, filed on Feb. 3, 2000, which are fully incorporated herein by reference. In addition, this application claims priority to U.S. Provisional Application Serial No. 60/333,919, filed on Nov. 28, 2001, which is fully incorporated herein by reference.

TECHNICAL FIELD

[0002] The present invention relates generally to systems and methods for data storage and retrieval and, more particularly, to data storage controllers employing lossless and/or lossy data compression and decompression to provide accelerated data storage and retrieval bandwidth.

BACKGROUND

[0003] Modern computers utilize a hierarchy of memory devices. To achieve maximum performance levels, modern processors utilize onboard memory and on board cache to obtain high bandwidth access to both program and data. Limitations in process technologies currently prohibit placing a sufficient quantity of onboard memory for most applications. Thus, in order to offer sufficient memory for the operating system(s), application programs, and user data, computers often use various forms of popular off-processor high speed memory including static random access memory (SRAM), synchronous dynamic random access memory (SDRAM), synchronous burst static ram (SBSRAM). Due to the prohibitive cost of the high-speed random access memory, coupled with their power volatility, a third lower level of the hierarchy exists for non-volatile mass storage devices.

[0004] Furthermore, mass storage devices offer increased capacity and fairly economical data storage. Mass storage devices (such as a "hard disk") typically store the operating system of a computer system, as well as applications and data and rapid access to such data is critical to system performance. The data storage and retrieval bandwidth of mass storage devices, however, is typically much less as compared with the bandwidth of other elements of a computing system. Indeed, over the last decade, although computer processor performance has improved by at least a factor of 50, magnetic disk storage performance has only improved by a factor of 5. Consequently, memory storage devices severely limit the performance of consumer, entertainment, office, workstation, servers, and mainframe computers for all disk and memory intensive operations.

[0005] The explosion in the data storage market will require both an increase in disk densities as well as a reduction in overall size. This latter aspect, ongoing computer miniaturization will not only affect disk architectures but will create pressure to merge current individual functions into more optimized composite implementations.

[0006] The ubiquitous Internet combined with new multimedia applications has put tremendous emphasis on storage volumetric density, storage mass density, storewidth,

and power consumption. Specifically, storage density is limited by the number of bits that are encoded in a mass storage device per unit volume. Similarly mass density is defined as storage bits per unit mass. Storewidth is the data rate at which the data may be accessed. There are various ways of categorizing storewidth in terms, several of the more prevalent metrics include sustained continuous storewidth, burst storewidth, and random access storewidth, all typically measured in megabytes/sec. Power consumption is canonically defined in terms of power consumption per bit and may be specified under a number of operating modes including active (while data is being accessed and transmitted) and standby mode. Hence one fairly obvious limitation within the current art is the need for even more volume, mass, and power efficient data storage.

[0007] Magnetic disk mass storage devices currently employed in a variety of home, business, and scientific computing applications suffer from significant seek-time access delays along with profound read/write data rate limitations. Currently the fastest available disk drives support only a sustained output data rate in the tens of megabytes per second data rate (MB/sec). This is in stark contrast to the modern Personal Computer's Peripheral Component Interconnect (PCI) Bus's low end 32 bit/33 Mhz input/output capability of 264 MB/sec and the PC's internal local bus capability of 800 MB/sec.

[0008] Another problem within the current art is that emergent high performance disk interface standards such as the Small Computer Systems Interface (SCSI-3), Fibre Channel, AT Attachment UltraDMA/66/100, Serial Storage Architecture, and Universal Serial Bus offer only higher data transfer rates through intermediate data buffering in random access memory. These interconnect strategies do not address the fundamental problem that all modern magnetic disk storage devices for the personal computer marketplace are still limited by the same typical physical media restrictions. In practice, faster disk access data rates are only achieved by the high cost solution of simultaneously accessing multiple disk drives with a technique known within the art as data striping and redundant array of independent disks (RAID).

[0009] RAID systems often afford the user the benefit of increased data bandwidth for data storage and retrieval. By simultaneously accessing two or more disk drives, data bandwidth may be increased at a maximum rate that is linear and directly proportional to the number of disks employed. Thus another problem with modern data storage systems utilizing RAID systems is that a linear increase in data bandwidth requires a proportional number of added disk storage devices.

[0010] Another problem with most modern mass storage devices is their inherent unreliability. Many modern mass storage devices utilize rotating assemblies and other types of electromechanical components that possess failure rates one or more orders of magnitude higher than equivalent solid-state devices. RAID systems employ data redundancy distributed across multiple disks to enhance data storage and retrieval reliability. In the simplest case, data may be explicitly repeated on multiple places on a single disk drive, on multiple places on two or more independent disk drives. More complex techniques are also employed that support various trade-offs between data bandwidth and data reliability.

[0011] Standard types of RAID systems currently available include RAID Levels 0, 1, and 5. The configuration selected depends on the goals to be achieved. Specifically data reliability, data validation, data storage/retrieval bandwidth, and cost all play a role in defining the appropriate RAID data storage solution. RAID level 0 entails pure data striping across multiple disk drives. This increases data bandwidth at best linearly with the number of disk drives utilized. Data reliability and validation capability are decreased. A failure of a single drive results in a complete loss of all data. Thus another problem with RAID systems is that low cost improved bandwidth requires a significant decrease in reliability.

[0012] RAID Level 1 utilizes disk mirroring where data is duplicated on an independent disk subsystem. Validation of data amongst the two independent drives is possible if the data is simultaneously accessed on both disks and subsequently compared. This tends to decrease data bandwidth from even that of a single comparable disk drive. In systems that offer hot swap capability, the failed drive is removed and a replacement drive is inserted. The data on the failed drive is then copied in the background while the entire system continues to operate in a performance degraded but fully operational mode. Once the data rebuild is complete, normal operation resumes. Hence, another problem with RAID systems is the high cost of increased reliability and associated decrease in performance.

[0013] RAID Level 5 employs disk data striping and parity error detection to increase both data bandwidth and reliability simultaneously. A minimum of three disk drives is required for this technique. In the event of a single disk drive failure, that drive may be rebuilt from parity and other data encoded on disk remaining disk drives. In systems that offer hot swap capability, the failed drive is removed and a replacement drive is inserted. The data on the failed drive is then rebuilt in the background while the entire system continues to operate in a performance degraded but fully operational mode. Once the data rebuild is complete, normal operation resumes.

[0014] Thus another problem with redundant modern mass storage devices is the degradation of data bandwidth when a storage device fails. Additional problems with bandwidth limitations and reliability similarly occur within the art by all other forms of sequential, pseudo-random, and random access mass storage devices. These and other limitations within the current art are addressed by the present invention.

SUMMARY OF THE INVENTION

[0015] The present invention is generally directed to data storage controllers and data storage devices employing lossless or lossy data compression and decompression to provide accelerated data storage and retrieval bandwidth. In one embodiment of the invention, a composite disk controller provides data storage and retrieval acceleration using multiple caches for data pipelining and increased throughput. In another embodiment of the invention, the disk controller with acceleration is embedded in the storage device and utilized for data storage and retrieval acceleration.

[0016] More particularly, in one aspect of the invention, a controller for controlling storage and retrieval of data to and

from a data storage device comprises a data compression/decompression engine for compressing data stored to the data storage device and for decompressing data retrieved from the data storage device, a first cache, operatively connected to the storage device and the data compression/decompression engine, for temporary storage of (i) compressed data that is read from the data storage device and (ii) compressed data from the data compression/decompression engine that is to be written to the data storage device, a second cache, operatively connected to the data compression/decompression engine and to a host interface bus, for temporary storage of (i) data that is received from a host system over the host interface bus for compression and storage and (ii) data that is to be sent to the host system over the host interface bus, and a cache manager for controlling the first and second data caches under commands received from the data compression/decompression engine.

[0017] In another aspect, the first cache, second cache and the data compression/decompression engine are operatively connected by a first local bus and the data compression/decompression engine is connected to the cache manager over a dedicated bus.

[0018] In another aspect of the invention, a magnetic disk controller for controlling storage and retrieval of data to and from a magnetic disk comprises an embedded data compression/decompression engine for compressing data stored to the magnetic disk and for decompressing data retrieved from the magnetic disk, an embedded first cache, operatively connected to the magnetic disk and the data compression/decompression engine, for temporary storage of (i) compressed data that is read from the magnetic disk and (ii) compressed data from the data compression/decompression engine that is to be written to the magnetic disk, an embedded second cache, operatively connected to the data compression/decompression engine and to a host interface bus, for temporary storage of (i) data that is received from a host system over the host interface bus for compression and storage and (ii) data that is to be sent to the host system over the host interface bus, and an embedded cache manager for controlling the first and second data caches under commands received from the data compression/decompression engine.

[0019] In yet another aspect of the invention, a magnetic disk controller for controlling storage and retrieval of data to and from a magnetic disk comprises an embedded data compression/decompression engine for compressing data stored to the magnetic disk and for decompressing data retrieved from the magnetic disk, an embedded bi-directional cache for temporary storage of (i) compressed data that is read from the magnetic disk and (ii) compressed data from the data compression/decompression engine that is to be written to the magnetic disk, an embedded cache manager for controlling the bi-directional cache, and an embedded virtual file management system for mapping compressed data blocks stored on the disk to corresponding uncompressed data blocks.

[0020] The present invention is realized due to recent improvements in processing speed, inclusive of dedicated analog and digital hardware circuits, central processing units, (and any hybrid combinations thereof), that, coupled with advanced data compression and decompression algorithms are enabling of ultra high bandwidth data compression and decompression methods that enable improved data storage and retrieval bandwidth

[0021] These and other aspects, features and advantages, of the present invention will become apparent from the following detailed description of preferred embodiments that is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] FIG. 1 is a block diagram of a data storage controller according to one embodiment of the present invention.

[0023] FIG. 2 is a block diagram of a data storage controller according to another embodiment of the present invention.

[0024] FIG. 3 is a block diagram of a data storage controller according to another embodiment of the present invention.

[0025] FIG. 4 is a block diagram of a data storage controller according to another embodiment of the present invention.

[0026] FIG. 5 is a block diagram of a data storage controller according to another embodiment of the present invention.

[0027] FIGS. 6a and 6b comprise a flow diagram of a method for initializing a data storage controller according to one aspect of the present invention.

[0028] FIGS. 7a and 7b comprise a flow diagram of a method for providing accelerated loading of an operating system and/or application programs upon system boot, according to one aspect of the present invention.

[0029] FIGS. 8a and 8b comprise a flow diagram of a method for providing accelerated loading of application programs according to one aspect of the present invention.

[0030] FIG. 9 is a diagram of an exemplary data compression system that may be employed in a data storage controller according to the present invention.

[0031] FIG. 10 is a diagram of an exemplary data decompression system that may be employed in a data storage controller according to the present invention.

[0032] FIG. 11 is a block diagram of a data storage controller according to another embodiment of the present invention.

[0033] FIG. 12 is a block diagram of a data storage controller according to another embodiment of the present invention.

[0034] FIG. 13 is a block diagram of a data storage device comprising an embedded data storage accelerator, according to an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0035] In the following description, it is to be understood that system elements having equivalent or similar functionality are designated with the same reference numerals in the Figures. It is to be further understood that the present invention may be implemented in various forms of hardware, software, firmware, or a combination thereof. Preferably, the present invention is implemented on a computer

platform including hardware such as one or more central processing units (CPU) or digital signal processors (DSP), a random access memory (RAM), and input/output (I/O) interface(s). The computer platform may also include an operating system, microinstruction code, and dedicated processing hardware utilizing combinatorial logic or finite state machines. The various processes and functions described herein may be either part of the hardware, microinstruction code or application programs that are executed via the operating system, or any combination thereof.

[0036] It is to be further understood that, because some of the constituent system components described herein are preferably implemented as software modules, the actual system connections shown in the Figures may differ depending upon the manner in that the systems are programmed. It is to be appreciated that special purpose microprocessors, dedicated hardware, or and combination thereof may be employed to implement the present invention. Given the teachings herein, one of ordinary skill in the related art will be able to contemplate these and similar implementations or configurations of the present invention.

[0037] I. System Architectures

[0038] The present invention is directed to data storage controllers that provide increased data storage/retrieval rates that are not otherwise achievable using conventional disk controller systems and protocols to store/retrieve data to/from mass storage devices. The concept of "accelerated" data storage and retrieval was introduced in copending U.S. patent application Ser. No. 09/266,394, filed March 11, 1999, entitled "System and Methods For Accelerated Data Storage and Retrieval" and copending U.S. patent application Ser. No. 09/481,243, filed Jan. 11, 2000, entitled "System and Methods For Accelerated Data Storage and Retrieval," both of which are commonly assigned and incorporated herein by reference. In general, as described in the above-incorporated applications, "accelerated" data storage comprises receiving a digital data stream at a data transmission rate which is greater than the data storage rate of a target storage device, compressing the input stream at a compression rate that increases the effective data storage rate of the target storage device and storing the compressed data in the target storage device. For instance, assume that a mass storage device (such as a hard disk) has a data storage rate of 20 megabytes per second. If a storage controller for the mass storage device is capable of compressing an input data stream with an average compression rate of 3:1, then data can be stored in the mass storage device at a rate of 60 megabytes per second, thereby effectively increasing the storage bandwidth ("storewidth") of the mass storage device by a factor of three. Similarly, accelerated data retrieval comprises retrieving a compressed digital data stream from a target storage device at the rate equal to, e.g., the data access rate of the target storage device and then decompressing the compressed data at a rate that increases the effective data access rate of the target storage device. Advantageously, accelerated data storage/retrieval mitigates the traditional bottleneck associated with, e.g., local and network disk accesses.

[0039] Referring now to FIG. 1, a high-level block diagram illustrates a data storage controller 10 according to one embodiment of the present invention. The data storage controller 10 comprises a data compression engine 12 for

compressing/decompressing data (preferably in real-time or pseudo real-time) stored/retrieved from a hard disk **11** (or any other type of mass storage device) to provide accelerated data storage/retrieval. The DCE **12** preferably employs the data compression/decompression techniques disclosed in U.S. Pat. No. 6,195,024, issued on Feb. 27, 2001, entitled "Content Independent Data Compression Method and System," which is commonly assigned and which is fully incorporated herein by reference. It is to be appreciated that the compression and decompression systems and methods disclosed in U.S. Pat. No. 6,195,024 are suitable for compressing and decompressing data at rates, which provide accelerated data storage and retrieval. A detailed discussion of a preferred "content independent" data compression process will be provided below.

[0040] The data storage controller **10** further comprises a cache **13**, a disk interface (or disk controller) **14** and a bus interface **15**. The storage controller **10** is operatively connected to the hard disk **12** via the disk controller **14** and operatively connected to an expansion bus (or main bus) **16** of a computer system via the bus interface **15**. The disk interface **14** may employ a known disk interface standard such as UltraDMA, SCSI, Serial Storage Architecture, FibreChannel or any other interface that provides suitable disk access data rates. In addition, the storage controller **10** preferably utilizes the American National Standard for Information Systems (ANSI) AT Attachment Interface (ATA/ATAPI-4) to connect the data storage controller **10** to the hard disk **12**. As is known in the art, this standard defines the connectors and cables for the physical interconnects between the data storage controller and the storage devices, along with the electrical and logical characteristics of the interconnecting signals.

[0041] Further, the bus interface **15** may employ a known standard such as the PCI (Peripheral Component Interconnect) bus interface for interfacing with a computer system. The use of industry standard interfaces and protocols is preferable, as it allows the storage controller **10** to be backwards compatible and seamlessly integrated with current systems. However in new designs the present invention may utilize any suitable computer interface or combination thereof.

[0042] It is to be understood that although FIG. 1 illustrates a hard disk **12**, the storage controller **10** may be employed with any form of memory device including all forms of sequential, pseudo-random, and random access storage devices. Storage devices as known within the current art include all forms of random access memory, magnetic and optical tape, magnetic and optical disks, along with various other forms of solid-state mass storage devices. The current invention applies to all forms and manners of memory devices including, but not limited to, storage devices utilizing magnetic, optical, and chemical techniques, or any combination thereof. In addition, the cache **13** may comprise volatile or non-volatile memory, or any combination thereof. Preferably, the cache **13** is implemented in SDRAM (static dynamic random access memory).

[0043] The system of FIG. 1 generally operates as follows. When data is read from disk by the host computer, data flows from the disk **11** through the data storage controller **10** to the host computer. Data is stored in one of several proprietary compression formats on the disk **11** (e.g., "con-

tent independent" data compression). Data blocks are pre-specified in length, comprised of single or multiple sectors, and are typically handled in fractional or whole equivalents of tracks, e.g. $\frac{1}{2}$ track, whole track, multiple tracks, etc. To read disk data, a DMA transfer is setup from the disk interface **14** to the onboard cache memory **13**. The disk interface **14** comprises integral DMA control to allow transfer of data from the disk **11** directly to the onboard cache **13** without intervention by the DCE **12**. It should be noted that the DCE **12** acts as a system level controller and sets-up specific registers within both the disk interface **14** and bus interface **15** to facilitate DMA transfers to and from the cache memory **13**. To initiate a transfer from the disk **11** to the cache **13**, the DMA transfer is setup via specifying the appropriate command (read disk), the source address (disk logical block number), amount of data to be transferred (number of disk logical blocks), and destination address within the onboard cache memory **13**.

[0044] Then, a disk data interrupt signal ("DISKINT#") is cleared (if previously set and not cleared) and the command is initiated by writing to the appropriate address space. Once data has been read from disk **11** and placed into onboard cache memory **13**, the DISKINT# interrupt is asserted notifying the DCE **12** that requested data is now available in the cache memory **13**. Data is then read by the DMA controller within the DCE **12** and placed into local memory for subsequent decompression. The decompressed data is then DMA transferred from the local memory of the DCE **12** back to the cache memory **13**.

[0045] Finally, data is DMA transferred via the bus interface controller **15** from the cache memory **13** to the bus **16**. It is to be understood that in the read mode, the data storage controller acts as a bus master. A bus DMA transfer is then setup via specifying the appropriate command (write to host computer), the source address within the cache memory **13**, the quantity of data words to be transferred (transfers are preferably in 4 byte increments), and the destination address on the host computer. When a bus **16** read or write transaction has completed, the appropriate interrupt signals (respectively referred to as PCIRDINT# and PCIWRINT#) are asserted to the DCE **12**. Either of these interrupts are cleared by a corresponding interrupt service routines through a read or write to the appropriate address of the DCE **12**.

[0046] Similarly, when data is written to the disk **11** from the host computer, data flows from the host computer through the data storage controller **10** and onto disk **11**. Data is normally received from the host computer in uncompressed (raw) format and is compressed by the DCE **12** and stored on the disk **11**. Data blocks from the host are pre-specified in length and are typically handled in blocks that are a fixed multiplier higher than fractional or whole equivalents of tracks, e.g. $\frac{1}{2}$ track, whole track, multiple tracks, etc. This multiplier is preferably derived from the expected average compression ratio that is selected when the disk is formatted with the virtual file management system. To read host computer data, a bus DMA transfer is setup from the host bus **16** to the onboard cache memory **13**. The bus interface controller **15** comprises integral DMA control that allows large block transfers from the host computer directly to the onboard cache **13** without intervention by the DCE **12**. The bus interface controller **15** acts as a host computer Bus Master when executing such transfer. Once

data has been read from the host and placed into onboard cache memory **13**, the data is read by the onboard DMA controller (residing on the DCE **12**) and placed into local memory for subsequent compression. The compressed data is then DMA transferred from the local memory of the DCE **12** back to the cache memory **13**. Finally, data is DMA transferred via the disk controller **14** from the cache **13** to the disk **11**.

[0047] As discussed in greater detail below, upon host computer power-up or external user reset, the data storage controller **10** initializes the onboard interfaces **14, 15** prior to release of the external host bus **16** from reset. The processor of the host computer then requests initial data from the disk **11** to facilitate the computer's boot-up sequence. The host computer requests disk data over the Bus **16** via a command packet issued from the host computer. Command packets are preferably eight words long (in a preferred embodiment, each word comprises 32 bits). Commands are written from the host computer to the data storage controller **10** with the host computer as the Bus Master and the data storage controller **10** as the slave. The data storage controller **10** includes at least one Base Address Register (BAR) for decoding the address of a command queue of the data storage controller **10**. The command queue resides within the cache **13** or within onboard memory of the DCE **12**.

[0048] When a command is received from the host computer, an interrupt (referred to herein as PCICMDINT#) is generated to the DCE processor. The eight-word command is read by the DCE **12** and placed into the command queue. Because the commands occupy a very small amount of memory, the location of the command queue is at the discretion of software and the associated system level performance considerations. Commands may be moved from the bus interface **16** to the command queue by either explicit reads and writes by the DCE processor or, as explained below, by utilizing programmed DMA from an Enhanced DMA Controller (EDMA) residing on the DCE **12**. This second technique may better facilitate system throughput by allowing the EDMA to automatically load commands while the highly pipelined data compression and decompression processing in the DCE is executed fully undisturbed.

[0049] The DCE **12**, disk interface **14** and bus interface **15** commonly share the cache **13**. As explained in detail below, the storage controller **10** preferably provides maximum system bandwidth by allowing simultaneous data transfers between the disk **12** and cache **13**, the DCE **12** and the cache **13**, and the expansion bus **16** and the cache **13**. This is realized by employing an integral DMA (direct memory access) protocol that allows the DCE **12**, disk interface **14** and bus interface **15** to transfer data without interrupting or interfering with other ongoing processes. In particular, as explained in detail below, an integral bandwidth allocation controller (or arbitrator) is preferably employed to allow the DCE **12**, disk controller **14**, and bus interface **15** to access the onboard cache with a bandwidth proportional to the overall bandwidth of the respective interface or processing element. The bandwidth arbitration occurs transparently and does not introduce latency in memory accesses. Bandwidth division is preferably performed with a high degree of granularity to minimize the size of requisite onboard buffers to synchronize data from the disk interface **14** and bus interface **15**.

[0050] It is to be appreciated that the implementation of a storage controller according to the present invention significantly accelerates the performance of a computer system and significantly increases hard disk data storage capacity. For instance, depending on the compression rate, for personal computers running standard Microsoft Windows® based business application software, the storage controller provides: (1) an increase of $n:1$ in disk storage capacity (for example, assuming a compression ration of 3:1, a 20 gigabyte hard drive effectively becomes a 60 gigabyte hard drive) (2) a significant decrease in the computer boot-up time (turn-on and operating system load) and the time for loading application software and (3) User data storage and retrieval is increased by a factor of $n:1$.

[0051] Referring now to FIG. 2, a block diagram illustrates a data storage controller **20** according to another embodiment of the present invention. More specifically, FIG. 2 illustrates a PCB (printed circuit board) implementation of the data storage controller **10** of FIG. 1. The storage controller **20** comprises a DSP (digital signal processor) **21** (or any other micro-processor device) that implements the DCE **12** of FIG. 1. The storage controller **21** further comprises at least one programmable logic device **22** (or volatile logic device). The programmable logic device **22** preferably implements the logic (program code) for instantiating and driving both the disk interface **14** and the bus interface **15** and for providing full DMA capability for the disk and bus interfaces **14, 15**. Further, as explained in detail below, upon host computer power-up and/or assertion of a system-level "reset" (e.g., PCI Bus reset), the DSP **21** initializes and programs the programmable logic device **22** before of the completion of initialization of the host computer. This advantageously allows the data storage controller **20** to be ready to accept and process commands from the host computer (via the bus **16**) and retrieve boot data from the disk (assuming the data storage controller **20** is implemented as the boot device and the hard disk stores the boot data (e.g., operating system, etc.)).

[0052] The data storage controller **20** further comprises a plurality of memory devices including a RAM (random access memory) device **23** and a ROM (read only memory) device **24** (or FLASH memory or other types of non-volatile memory). The RAM device **23** is utilized as on-board cache and is preferably implemented as SDRAM (preferably, 32 megabytes minimum). The ROM device **24** is utilized for non-volatile storage of logic code associated with the DSP **21** and configuration data used by the DSP **21** to program the programmable logic device **22**. The ROM device **24** preferably comprises a one time (erasable) programmable memory (OTP-EPROM) device.

[0053] The DSP **21** is operatively connected to the memory devices **23, 24** and the programmable logic device **22** via a local bus **25**. The DSP **21** is also operatively connected to the programmable logic device **22** via an independent control bus **26**. The programmable logic device **22** provides data flow control between the DSP **21** and the host computer system attached to the bus **16**, as well as data flow control between the DSP **21** and the storage device. A plurality of external I/O ports **27** are included for data transmission and/or loading of one programmable logic devices. Preferably, the disk interface **14** driven by the programmable logic device **22** supports a plurality of hard drives.

[0054] The storage controller **20** further comprises computer reset and power up circuitry **28** (or “boot configuration circuit”) for controlling initialization (either cold or warm boots) of the host computer system and storage controller **20**. A preferred boot configuration circuit and preferred computer initialization systems and protocols are described in U.S. patent application Ser. No. 09/775,897, filed on Feb. 2, 2001, entitled “Systems and Methods for Computer Initialization”, which is commonly assigned and incorporated herein by reference. Preferably, the boot configuration circuit **28** is employed for controlling the initializing and programming the programmable logic device **22** during configuration of the host computer system (i.e., while the CPU of the host is held in reset). The boot configuration circuit **28** ensures that the programmable logic device **22** (and possibly other volatile or partially volatile logic devices) is initialized and programmed before the bus **16** (such as a PCI bus) is fully reset.

[0055] In particular, when power is first applied to the boot configuration circuit **28**, the boot configuration circuit **28** generates a control signal to reset the local system (e.g., storage controller **20**) devices such as a DSP, memory, and I/O interfaces. Once the local system is powered-up and reset, the controlling device (such as the DSP **21**) will then proceed to automatically determine the system environment and configure the local system to work within that environment. By way of example, the DSP **21** of the disk storage controller **20** would sense that the data storage controller **20** is on a PCI computer bus (expansion bus) and has attached to it a hard disk on an IDE interface. The DSP **21** would then load the appropriate PCI and IDE interfaces into the programmable logic device **22** prior to completion of the host system reset. It is to be appreciated that this can be done for all computer busses and boot device interfaces including: PCI, NuBus, ISA, Fiber Channel, SCSI, Ethernet, DSL, ADSL, IDE, DMA, Ultra DMA, and SONET. Once the programmable logic device **22** is configured for its environment, the boot device controller is reset and ready to accept commands over the computer/expansion bus **16**. Details of the boot process using a boot device comprising a programmable logic device will be provided below.

[0056] It is to be understood that the data storage controller **20** may be utilized as a controller for transmitting data (compressed or uncompressed) to and from remote locations over the DSP I/O ports **27** or system bus **16**, for example. Indeed, the I/O ports **27** of the DSP **21** may be used for transmitting data (compressed or uncompressed) that is either retrieved from the disk **11** or received from the host system via the bus **16**, to remote locations for processing and/or storage. Indeed, the I/O ports may be operatively connected to other data storage controllers or to a network communication channels. Likewise, the data storage controller **20** may receive data (compressed or uncompressed) over the I/O ports **27** of the DSP **21** from remote systems that are connected to the I/O ports **27** of the DSP, for local processing by the data storage controller **20**. For instance, a remote system may remotely access the data storage controller (via the I/O ports of the DSP or system bus **16**) to utilize the data compression, in which case the data storage controller would transmit the compressed data back to the system that requested compression.

[0057] The DSP **21** may comprise any suitable commercially available DSP or processor. Preferably, the data stor-

age controller **20** utilizes a DSP from Texas Instruments' 320 series, C62x family, of DSPs (such as TMS320C6211GFN-150), although any other DSP or processor comprising a similar architecture and providing similar functionalities may be employed. The preferred DSP is capable of up to 1.2 billion instructions per second. Additional features of the preferred DSP include a highly parallel eight processor single cycle instruction execution, onboard 4K byte LIP Program Cache, 4K LID Data Cache, and 64K byte Unified L2 Program/Data Cache. The preferred DSP further comprises a 32 bit External Memory Interface (EMIF) that provides for a glueless interface to the RAM **23** and the non-volatile memory **24** (ROM). The DSP further comprises two multi-channel buffered serial ports (McBSPs) and two 32 bit general purpose timers. Preferably, the storage controller disables the I/O capability of these devices and utilizes the I/O ports of the DSP as general purpose I/O for both programming the programmable logic device **22** using a strobed eight bit interface and signaling via a Light Emitting Diode (LED). Ancillary DSP features include a 16 bit Host Port Interface and full JTAG emulation capability for development support. The programmable logic device **22** may comprise any form of volatile or non-volatile memory. Preferably, the programmable logic device **22** comprises a dynamically reprogrammable FPGA (field programmable gate array) such as the commercially available Xilinx Spartan Series XCS40XL-PQ240-5 FPGA. As discussed in detail herein, the FPGA instantiates and drives the disk and bus interfaces **14**, **15**. The non-volatile memory device **24** preferably comprises a 128 Kbyte M27W101-80K one time (erasable) programmable read only memory, although other suitable non-volatile storage devices may be employed. The non-volatile memory device **24** is decoded at a designated memory space in the DSP **21**. The non-volatile memory device **24** stores the logic for the DSP **21** and configuration data for the programmable logic device **22**. More specifically, in a preferred embodiment, the lower 80 Kbytes of the non-volatile memory device **24** are utilized for storing DSP program code, wherein the first 1 k bytes are utilized for the DSP's boot loader. Upon reset of the DSP **21** (via boot configuration circuit **28**), the first 1 K of memory of the non-volatile memory device **24** is copied into an internal RAM of the DSP **21** by e.g., the DSP's Enhanced DMA Controller (EDMA). Although the boot process begins when the CPU of the host system is released from external reset, the transfer of the boot code into the DSP and the DSP's initialization of the programmable logic device actually occurs while the CPU of the host system is held in reset. After completion of the 1K block transfer, the DSP executes the boot loader code and continues thereafter with executing the remainder of the code in non-volatile memory device to program the programmable logic device **22**.

[0058] More specifically, in a preferred embodiment, the upper 48K bytes of the non-volatile memory device **24** are utilized for storing configuration data associated with the programmable logic device **22**. If the data storage controller **20** is employed as the primary boot storage device for the host computer, the logic for instantiating and driving the disk and bus interfaces **14**, **15** should be stored on the data storage controller **20** (although such code may be stored in remotely accessible memory locations) and loaded prior to release of the host system bus **16** from “reset”. For instance, revision 2.2 of the PCI Local Bus specification calls for a typical delay of 100 msec from power-stable before release of PCI

Reset. In practice this delay is currently 200 msec although this varies amongst computer manufacturers. A detailed discussion of the power-on sequencing and boot operation of the data storage controller **20** will be provided below.

[0059] FIG. 3 illustrates another embodiment of a data storage controller **30** wherein the data storage controller **35** is embedded within the motherboard of the host computer system. This architecture provides the same functionality as the system of FIG. 2, and also adds the cost advantage of being embedded on the host motherboard. The system comprises additional RAM and ROM memory devices **23a**, **24a**, operatively connected to the DSP **21** via a local bus **25a**.

[0060] FIG. 4 illustrates another embodiment of a data storage controller. The data storage controller **40** comprises a PCB implementation that is capable of supporting RAID levels 0, 1 and 5. This architecture is similar to those of FIGS. 1 and 2, except that a plurality of programmable logic devices **22**, **22a** are utilized. The programmable logic device **22** is dedicated to controlling the bus interface **15**. The programmable logic device **22a** is dedicated to controlling a plurality of disk interfaces **14**, preferably three interfaces. Each disk interface **14** can connect up to two drives. The DSP in conjunction with the programmable logic device **22a** can operate at RAID level 0, 1 or 5. At RAID level 0, which is disk striping, two interfaces are required. This is also true for RAID level 1, which is disk mirroring. At RAID level 5, all three interfaces are required.

[0061] FIG. 5 illustrates another embodiment of a data storage controller according to the present invention. The data storage controller **45** provides the same functionality as that of FIG. 4, and has the cost advantage of being embedded within the computer system motherboard.

[0062] II. Initializing A Programmable Logic Device

[0063] As discussed above with reference to FIG. 2, for example, the data storage controller **20** preferably employs an onboard Texas Instruments TMS320C6211 Digital Signal Processor (DSP) to program the onboard Xilinx Spartan Series XCS40XL FPGA upon power-up or system level PCI reset. The onboard boot configuration circuit **28** ensures that from system power-up and/or the assertion of a bus reset (e.g., PCI reset), the DSP **21** is allotted a predetermined amount of time (preferably a minimum of 10 msec) to boot the DSP **21** and load the programmable logic device **22**. Because of a potential race condition between either the host computer power-up or assertion of PCI Bus reset and configuration of the programmable logic device **20** (which is used for controlling the boot device and accepting PCI Commands), an "Express Mode" programming mode for configuring the SpartanXL family XCS40XL device is preferably employed. The XCS40XL is factory set to byte-wide Express-Mode programming by setting both the M1/M0 bits of the XCS40XL to 0x0. Further, to accommodate express mode programming of the programmable logic device **22**, the DSP **21** is programmed to utilize its serial ports reconfigured as general purpose I/O. However, after the logic device **22** is programmed, the DSP **21** may then reconfigure its serial ports for use with other devices. Advantageously, using the same DSP ports for multiple purposes affords greater flexibility while minimizing hardware resources and thus reducing product cost.

[0064] The volatile nature of the logic device **22** effectively affords the ability to have an unlimited number of

hardware interfaces. Any number of programs for execution by the programmable logic device **22** can be kept in an accessible memory location (EPROM, hard disk, or other storage device). Each program can contain new disk interfaces, interface modes or subsets thereof. When necessary, the DSP **21** can clear the interface currently residing in the logic device **22** and reprogram it with a new interface. This feature allows the data storage controller **20** to have compatibility with a large number of interfaces while minimizing hardware resources and thus reducing product cost.

[0065] A preferred protocol for programming the programmable logic device can be summarized in the following steps: (1) Clearing the configuration memory; (2) Initialization; (3) Configuration; and (4) Start-Up. When either of three events occur: the host computer is first powered-up or a power failure and subsequent recovery occurs (cold boot), or a front panel computer reset is initiated (warm boot), the host computer asserts RST# (reset) on the PCI Bus. As noted above, the data storage controller **20** preferably comprises a boot configuration circuit **28** that senses initial host computer power turn-on and/or assertion of a PCI Bus Reset ("PCI RST#"). It is important to note that assuming the data storage controller **20** is utilized in the computer boot-up sequence, it should be available exactly 5 clock cycles after the PCI RST# is deasserted, as per PCI Bus Specification Revision 2.2. While exact timings vary from computer to computer, the typical PCI bus reset is asserted for approximately 200 msec from initial power turn-on.

[0066] In general, PCI RST# is asserted as soon as the computer's power exceeds a nominal threshold of about 1 volt (although this varies) and remains asserted for 200 msec thereafter. Power failure detection of the 5 volt or 3.3 volt bus typically resets the entire computer as if it is an initial power-up event (i.e., cold boot). Front panel resets (warm boots) are more troublesome and are derived from a debounced push-button switch input.

[0067] Typical front panel reset times are a minimum of 20 msec, although again the only governing specification limit is 1 msec reset pulse width.

[0068] As discussed in detail below, it may not be necessary to reload the programmable logic device **22** each time the DSP is reset. The boot configuration circuit **20** preferably comprises a state machine output signal that is readable by the DSP **21** to ascertain the type of boot process requested. For example, with a front-panel reset (warm boot), the power remains stable on the PCI Bus, thus the programmable logic device **22** should not require reloading.

[0069] Referring now to FIG. 6, a flow diagram illustrates a method for initializing the programmable logic device **22** according to one aspect of the invention. In the following discussion, it is assumed that the programmable logic device **22** is always reloaded, regardless of the type of boot process. Initially, in FIG. 6a, the DSP **21** is reset by asserting a DSP reset signal (step 50). Preferably, the DSP reset signal is generated by the boot circuit configuration circuit **28** (as described in the above-incorporated U.S. patent application Ser. No. 09/775,897). While the DSP reset signal is asserted (e.g., active low), the DSP is held in reset and is initialized to a prescribed state. Upon deassertion of the DSP Reset signal, the logic code for the DSP (referred to as the "boot loader") is copied from the non-volatile logic device **24** into memory residing in the DSP **21** (step 51). This allows the

DSP to execute the initialization of the programmable logic device 22. In a preferred embodiment, the lower 1K bytes of EPROM memory is copied to the first 1 k bytes of DSP's low memory (0x0000 0000 through 0x0000 03FF). As noted above, the memory mapping of the DSP 21 maps the CE1 memory space located at 0x9000 0000 through 0x9001 FFFF with the OTP EPROM. In a preferred embodiment using the Texas Instrument DSP TMS320c6211GFN-150, this ROM boot process is executed by the EDMA controller of the DSP. It is to be understood, however, that the EDMA controller may be instantiated in the programmable logic device (Xilinx), or shared between the DSP and programmable logic device.

[0070] After the logic is loaded in the DSP 21, the DSP 21 begins execution out of the lower 1K bytes of memory (step 52). In a preferred embodiment, the DSP 21 initializes with at least the functionality to read EPROM Memory (CE1) space. Then, as described above, the DSP preferably configures its serial ports as general purpose I/O (step 53).

[0071] Next, the DSP 21 will initialize the programmable logic device 22 using one or more suitable control signals. (step 54). After initialization, the DSP 21 begins reading the configuration data of the programmable logic device 22 from the non-volatile memory 24 (step 55). This process begins with clearing a Data Byte Counter and then reading the first data byte beginning at a prespecified memory location in the non-volatile memory 24 (step 56). Then, the first output byte is loaded into the DSP's I/O locations with LSB at D0 and MSB at D7 (step 57). Before the first byte is loaded to the logic device 22, a prespecified time delay (e.g., 5 usec) is provided to ensure that the logic device 22 has been initialized (step 58). In particular, this time delay should be of a duration at least equal to the internal setup time of the programmable logic device 22 from completion of initialization. Once this time delay has expired, the first data byte in the I/O bus 26 of the DSP 21 is latched into the programmable logic device 22 (step 59).

[0072] Next, a determination is made as to whether the Data Byte Counter is less than a prespecified value (step 60). If the Data Byte Counter is less than the prespecified value (affirmative determination in step 60), the next successive data byte for the programmable logic device 22 is read from the non-volatile memory 24 (step 61) and the Data Byte Counter is incremented (step 62).

[0073] Next, the read data byte is loaded into the I/O of the DSP (step 63). A time delay of, e.g., 20 nsec is allowed to expire before the data byte is latched to the programmable logic device to ensure that a minimum data set-up time to the programmable logic device 21 is observed (step 64) and the process is repeated (return to step 60). It is to be appreciated that steps 60-64 may be performed while the current data byte is being latched to the programmable logic device. This provides "pipeline" programming of the logic device 22 and minimizes programming duration. When the Data Byte Counter is not less than the prespecified count value negative determination in step 60), as shown in FIG. 6b, the last data byte is read from the non-volatile memory and latched to the programmable logic device 22, and the DSP 21 will then poll a control signal generated by the programmable logic device 22 to ensure that the programming of the logic device 22 is successful (step 65). If programming is complete (affirmative determination in step 66), the process

continues with the remainder of the data storage controller initialization (step 67). Otherwise, a timeout occurs (step 68) and upon expiration of the timeout, an error signal is provided and the programming process is repeated (step 69).

[0074] III. Data Storage and Retrieval Protocols

[0075] A detailed discussion of operational modes of a data storage controller will now be provided with reference to the embodiment of FIG. 2 (although it is to be understood that the following discussion is applicable to all the above-described embodiments). The data storage controller 20 utilizes a plurality of commands to implement the data storage, retrieval, and disk maintenance functions described herein. Each command preferably comprises eight thirty-two bit data words stored and transmitted in little endian format.

[0076] The commands include: Read Disk Data; Write Disk Data; and Copy Disk Data, for example. For example, a preferred format for the "Read Disk Data" command is:

31	16	15	8	7	0
Command Packet Number		Command Type		Command Parameters (00h)	
0000h to FFFFh		00h			
Starting Block Address (Least Significant Word)				04h	
Starting Block Address (Most Significant Word)				08h	
Number of Blocks (Least Significant Word)				0Ch	
Number of Blocks (Most Significant Word)				10h	
Destination Address (Least Significant Word)				14h	
Destination Address (Most Significant Word)				18h	
Checksum			Reserved		1Ch

[0077] The host computer commands the data storage controller 20 over the PCI Bus 16, for example. Upon computer power-up or reset, the host computer issues a PCI Bus Reset with a minimum pulse width of 100 msec (in accordance with PCI Bus Specification Revision 2.2). Upon completion of the PCI Bus reset, the data storage controller 20 is fully initialized and waiting for completion of the PCI configuration cycle. Upon completion of the PCI configuration cycles, the data storage controller will wait in an idle state for the first disk command.

[0078] During operation, the host operating system may issue a command to the data storage controller 20 to store, retrieve, or copy specific logical data blocks. Each command is transmitted over the PCI Bus 16 at the Address assigned to the Base Address Register (BAR) of the data storage controller 20.

[0079] The commands issued by the host system to the data storage controller and the data transmitted to and from the data storage controller are preferably communicated via a 32 bit, 33 MHz, PCI Data Bus. As noted above, the PCI Interface is preferably housed within the onboard Xilinx Spartan XCS40XL-5 40,000 field programmable gate array which instantiates a PCI 32, 32 Bit, 33 MHz PCI Bus Interface (as per PCI Bus Revision 2.2).

[0080] The PCI Bus interface operates in Slave Mode when receiving commands and as a Bus Master when reading or writing data. The source and destination for all data is specified within each command packet. When setting up data transfers, the Enhanced Direct Memory Access (EDMA) Controller of the DSP (or the Xilinx) utilizes two

Control Registers, a 16 Word Data Write to PCI Bus FIFO, a 16 Word Data Read From PCI Bus FIFO, and a PCI Data Interrupt (PCIDATINT). The 32 Bit PCI Address Register holds either the starting Source Address for data storage controller Disk Writes where data is read from the PCI Bus, or the starting Destination Address for data storage controller Disk Reads where data is written to the PCI Bus. The second control register is a PCI Count Register that specifies the direction of the data transfer along with the number of 32 bit Data words to be written to or from the PCI bus.

[0081] Data is written to the PCI Bus from the DSP via a 16 Word PCI Data Write FIFO located within a prespecified address range. Data writes from the DSP to anywhere within the address range place that data word in the next available location within the FIFO. Data is read from the PCI Bus to the DSP via a 16 Word PCI Data Read FIFO located within a prespecified address range and data read by the DSP from anywhere within this address range provides the next data word from the FIFO.

[0082] After completion of the Xilinx initialization by the DSP and subsequent negation of the PCI Bus Reset signal (RST#) by the host computer's PCI Bridge, the data storage controller is ready to accept commands from the host computer via the PCI Bus. When accepting commands it should be noted that the data storage controller is a PCI Target (Slave) Device. Commands are preferably fixed in length at exactly 8 (thirty-two bit) words long. Commands are written from the host computer to the data storage controller via the PCI Bus utilizing the data storage controller's Base Address Register 0 (BAR0). The PCI Bus Reset initially sets the Command FIFO's Counter to zero and also signals the Xilinx's PCI Bus State Controller that the Command FIFO is empty and enable to accept a command.

[0083] Whenever a data write occurs within the valid data range of BAR0, the data word is accepted from PCI Bus and placed in the next available memory position within the Command FIFO. When the last of the 8 thirty-two bit data words is accepted by the PCI Bus (thus completing the command, i.e. last word for the command FIFO to be full), the PCI Bus State Controller is automatically set to Target Abort (within same PCI Transaction) or Disconnect Without Data for all subsequent PCI transactions that try to writes to BAR0. This automatic setting is the responsibility of the Xilinx PCI Data Interface.

[0084] The PCI Command FIFO State Controller then asserts the Command Available Interrupt to the DSP. The DSP services the Command Available Interrupt by reading the command data from a prespecified address range. It should be noted that the command FIFO is read sequentially from any data access that reads data within such address range. It is the responsibility of the DSP to understand that the data is read sequentially from any order of accesses within the data range and should thus be stored accordingly.

[0085] Upon completion of the Command Available Interrupt Service Routine the DSP executes a memory read or write to desired location within the PCI Control Register Space mapped into the DSP's CE3 (Xilinx) memory space. This resets the Command FIFO Counter back to zero. Next, the DSP executes a memory read or write to location in the DSP Memory Space that clears the Command Available Interrupt. Nested interrupts are not possible since the PCI

Bus State Machine is not yet able to accept any Command Data at BAR0. Once the Command Available Interrupt routine has cleared the interrupt and exited, the DSP may then enable the PCI State Machine to accept a new command by reading or writing to PCI Command Enable location within the PCI Command FIFO Control Register Space.

[0086] A preferred architecture has been selected to enable the data storage controller to operate on one command at a time or to accept multiple prioritized commands in future implementations. Specifically, the decoupling of the Command Available Interrupt Service Routine from the PCI State Machine that accepts Commands at BAR0 enables the DSP's "operating system kernel" to accept additional commands at any time by software command. In single command operation, a command is accepted, the Command Available Interrupt Cleared, and the Command executed by the data storage controller in PCI Master Mode prior to the enabling of the PCI State machine to accept new commands.

[0087] In a prioritized multi-command implementation, the "operating system kernel" may elect to immediately accept new commands or defer the acceptance of new commands based upon any software implemented decision criteria. In one embodiment, the O/S code might only allow a pre-specified number of commands to be queued. In another embodiment, commands might only be accepted during processor idle time or when the DSP is not executing time critical (i.e. highly pipelined) compress/decompress routines. In yet another embodiment, various processes are enabled based upon a pre-emptive prioritized based scheduling system.

[0088] As previously stated, the data storage controller retrieves commands from the input command FIFO in 8 thirty-two bit word packets. Prior to command interpretation and execution, a command's checksum value is computed to verify the integrity of the data command and associated parameters. If the checksum fails, the host computer is notified of the command packet that failed utilizing the Command Protocol Error Handler. Once the checksum is verified the command type and associated parameters are utilized as an offset into the command "pointer" table or nay other suitable command/data structure that transfers control to the appropriate command execution routine.

[0089] Commands are executed by the data storage controller with the data storage controller acting as a PCI Master. This is in direct contrast to command acceptance where the data storage controller acts as a PCI Slave. When acting as a PCI Bus Master, the data storage controller reads or writes data to the PCI Bus utilizing a separate PCI Bus Data FIFO (distinct & apart from the Command FIFO). The PCI Data FIFO is 64 (thirty-two bit) words deep and may be utilized for either data reads or data writes from the DSP to the PCI Bus, but not both simultaneously.

[0090] For data to be written from the data storage controller to the Host Computer, the DSP must first write the output data to the PCI Bus Data FIFO. The Data FIFO is commanded to PCI Bus Data Write Mode by writing to a desired location within the Xilinx (CE3) PCI Control Register Space. Upon PCI Bus Reset the default state for the PCI Data FIFO is write mode and the PCI Data FIFO Available Interrupt is cleared. The PCI Data FIFO Available Interrupt should also be software cleared by writing to a prespecified location. Preferably, the first task for the data storage con-

troller is for system boot-up or application code to be downloaded from disk. For reference, PCI Data Read Mode is commanded by writing to location BFF0 0104. The PCI Bus Reset initializes the Data FIFO Pointer to the first data of the 64 data words within the FIFO. However this pointer should always be explicitly initialized by a memory write to location BFF0 0108. This ensures that the first data word written to the FIFO by the DSP performing the data write anywhere in address range B000 0000 to B000 01FF is placed at the beginning of the FIFO. Each subsequent write to any location within this address range then places one thirty-two bit data word into the next available location within the PCI Data FIFO. The FIFO accepts up to 64 thirty-two bit data words although it should be clearly understood that not all data transfers to and from the PCI Bus will consist of a full FIFO. Counting the number of thirty-two bit data words written to the PCI Data FIFO is the responsibility of the DSP Code. It is envisioned that the DSP will, in general, use 64 word DMA data transfers, thus alleviating any additional processor overhead. When the data has been transferred from the DSP to the PCI Data FIFO, the PCI Bus Controller also needs the address of the PCI Target along with the number of data words to be transmitted. In the current data storage controller implementation, the PCI Bus Address is thirty-two bits wide, although future PCI bus implementations may utilize multiword addressing and/or significantly larger (64 bit & up) address widths. The single thirty-two bit address word is written by the DSP to memory location $aaaa+0 \times 10$ in the PCI Control Register Space.

[0091] Finally, the PCI Bus Data Write transaction is initiated by writing the PCI Data FIFO word count to a prespecified memory address. The word count value is always decimal 64 or less ($0 \times 3F$). When the count register is written the value is automatically transferred to the PCI Controller for executing the PCI Bus Master writes.

[0092] When the PCI Bus has completed the transfer of all data words within the PCI Data FIFO the PCI Data FIFO Available Interrupt is set. The DSP PCI Data FIFO Available Interrupt handler will then check to see if additional data is waiting or expected to be written to the PCI Data Bus. If additional data is required the interrupt is cleared and the data transfer process repeats. If no additional data is required to be transferred then the interrupt is cleared and the routine must exit to a system state controller. For example, if the command is complete then master mode must be disabled and then slave mode (command mode) enabled—assuming a single command by command execution data storage controller.

[0093] For data to be read by the data storage controller from the Host Computer, the DSP must command the PCI Bus with the address and quantity of data to be received.

[0094] The PCI Data FIFO is commanded to PCI Bus Data Read Mode by writing to a desired location within the Xilinx (CE3) PCI Control Register Space. Upon PCI Bus Reset the default state for the PCI Data FIFO is Write Mode and the PCI Data FIFO Full Interrupt is cleared. The PCI Data FIFO Full Interrupt should also be cleared via software by writing to such location. The PCI Bus Reset also initializes the PCI Data FIFO Pointer to the first data word of the available 64 data words within the FIFO. However this pointer should always be explicitly initialized by a memory write to prespecified location.

[0095] For data to be read from the PCI Bus by the data storage controller, the Xilinx PCI Bus Controller requires the address of the PCI Target along with the number of data words to be received. In the current data storage controller implementation, the PCI Bus Address is thirty-two bits wide, although future PCI bus implementations may utilize multiword addressing and/or significantly larger (64 bit & up) address widths. The single thirty-two bit address word is written by the DSP to prespecified memory location in the PCI Control Register Space.

[0096] Finally, the PCI Bus Data Read transaction is initiated by writing the PCI Data FIFO word count to prespecified memory address. The word count value is always decimal 64 or less ($0 \times 3F$). When the count register is written the value is automatically transferred to the PCI Controller for executing the PCI Bus Master Read.

[0097] When the PCI Bus has received all the requested data words PCI Data FIFO Full Interrupt is set. The DSP PCI Data FIFO Full Interrupt handler will then check to see if additional data is waiting or expected to be read from the PCI Data Bus. If additional data is required the interrupt is cleared and the data receipt process repeats. If no additional data is required to be transferred, then the interrupt is cleared and the routine exits to a system state controller. For example, if the command is complete then master mode must be disabled and then slave mode (command mode) enabled—assuming a single command by command execution data storage controller.

[0098] It is clearly understood that there are other techniques for handling the PCI Data transfers. The current methodology has been selected to minimize the complexity and resource utilization of the Xilinx Gate Array. It should also be understood that the utilization of asynchronous memory reads and writes to initialize system states and synchronize events at a software level aids in both hardware and system level debug at the expense of increase software overhead. Subsequent embodiments of the gate array may automate resource intensive tasks if system level performance mandates.

[0099] IV. Memory Bandwidth Allocation

[0100] The onboard cache of the data storage controller is shared by the DSP, Disk Interface, and PCI Bus. The best case, maximum bandwidth for the SDRAM memory is 70 megawords per second, or equivalently, 280 megabytes per second. The 32 bit PCI Bus interface has a best case bandwidth of 132 megabytes per second, or equivalently 33 megawords per second. In current practice, this bandwidth is only achieved in short bursts. The granularity of PCI data bursts to/from the data storage controller is governed by the PCI Bus interface data buffer depth of sixteen words (64 bytes). The time division multiplexing nature of the current PCI Data Transfer Buffering methodology cuts the 10 sustained PCI bandwidth down to 66 megabytes/second.

[0101] Data is transferred across the ultraDMA disk interface at a maximum burst rate of 66 megabytes/second. It should be noted that the burst rate is only achieved with disks that contain onboard cache memory. Currently this is becoming more and more popular within the industry. However assuming a disk cache miss, the maximum transfer rates from current disk drives is approximately six megabytes per second. Allotting for technology improvements over time,

the data storage controller has been designed for a maximum sustained disk data rate of 20 megabytes second (5 megawords/second). A design challenge is created by the need for continuous access to the SDRAM memory. Disks are physical devices and it is necessary to continuously read data from disk and place it into memory, otherwise the disk will incur a full rotational latency prior to continuing the read transaction. The maximum SDRAM access latency that can be incurred is the depth of the each of the two disk FIFO s or sixteen data. Assuming the

[0102] FIFO is sixteen words deep the maximum latency time for emptying the other disk FIFO and restoring it to the disk interface is sixteen words at 5 megawords per second or (16x3.2 usec)=1 usec. Each EMIF clock cycle is 14.2857 nsec, thus the maximum latency translates to 224 clock cycles. It should be noted that transfers across the disk interface are 16 bits wide, thus the FPGA is required to translate 32 bit memory transfers to 16 bit disk transfers, and vice-versa.

[0103] The DSP services request for its external bus from two requesters, the Enhanced Direct Memory Access (EDMA) Controller and an external shared memory device controller. The DSP can typically utilize the full 280 megabytes of bus bandwidth on an 8k through 64K byte (2 k word through 16 k word) burst basis. It should be noted that the DSRA does not utilize the SDRAM memory for interim processing storage, and as such only utilizes bandwidth in direct proportion to disk read and write commands.

[0104] For a single read from disk transaction data is transferred from and DMA transfer into SDRAM memory. This data is then DMA transferred by the DSP into onboard DSP memory, processed, and re transferred back to SDRAM in decompressed format (3 words for every one word in). Finally the data is read from SDRAM by the PCI Bus Controller and placed into host computer memory. This equates to eight SDRAM accesses, one write from disk, one read by the DSP, three writes by the DSP and three by the PCI Bus. Disk write transactions similarly require eight SDRAM accesses, three from the PCI, three DSP reads, one DSP write, and one to the disk.

[0105] Neglecting overhead for setting up DMA transfers, arbitration latencies, and memory wait states for setting up SDRAM transactions, the maximum DSRA theoretical SDRAM bandwidth limit for disk reads or writes is 280/8 megabytes second or 35 megabytes second. It should be noted that the best case allocation of SDRAM bandwidth would be dynamic dependent upon the data compression and decompression ratios. Future enhancements to the data storage controller will utilize a programmable timeslice system to allocate SDRAM bandwidth, however this first embodiment will utilize a fixed allocation ratio as follows:

[0106] If all three requesters require SDRAM simultaneously:

PCI Bus Interface	3/8
DSP Accesses	4/8
UltraDMA Disk Interface	1/8

[0107] If only the PCI Bus and DSP require SDRAM:

PCI Bus Interface	4/8
DSP Accesses	4/8

[0108] If only the DSP and Disk require SDRAM:

DSP Accesses	6/8
UltraDMA Disk Interface	2/8

[0109] If only the PCI Bus and Disk require SDRAM:

PCI Bus Interface	6/8
UltraDMA Disk Interface	2/8

[0110] If only one device requires SDRAM it receives the full SDRAM bandwidth. It should be noted that different ratios may be applied based upon the anticipated or actual compression and/or decompression ratios. For example in the case of all three requestors active the following equation applies. Assume that data storage accelerator achieves a compression ratio A:B for example 3:1. The Numerator and denominators of the various allocations are defined as follows:

PCI Bus Interface	A/K
DSP Accesses	(A + B)/K
UltraDMA Disk Interface	B/K

[0111] Where Further define a sum K equal to the sum of the numerators of the PCI Bus interface fraction, the DSP Access fraction, and the UltraDMA Disk Interfaces, i.e. K=2(A+B). Similarly:

[0112] If only the PCI Bus and DSP require SDRAM:

PCI Bus Interface	(A + B)/K
DSP Accesses	(A + B)/K

[0113] If only the DSP and Disk require SDRAM:

DSP Accesses	2A/K
UltraDMA Disk Interface	2B/K

[0114] If only the PCI Bus and Disk require SDRAM:

PCI Bus Interface	2A/K
UltraDMA Disk Interface	2B/K

[0115] It should be noted that the resultant ratios may all be scaled by a constant in order to most effectively utilize the bandwidths of the internal busses and external interfaces. In addition each ratio can be scale by an adjustment factor based upon the time required to complete individual cycles. For example if PCI Bus interface takes 20% longer than all other cycles, the PCI time slice should be adjusted longer accordingly.

[0116] V. Instant Boot Device For Operating System, Application Program and Loading

[0117] Typically, with conventional boot device controllers, after reset, the boot device controller will wait for a command over the computer bus (such as PCI). Since the boot device controller will typically be reset prior to bus reset and before the computer bus starts sending commands, this wait period is unproductive time. The initial bus commands inevitably instruct the boot device controller to retrieve data from the boot device (such as a disk) for the operating system. Since most boot devices are relatively slow compared to the speed of most computer busses, a long delay is seen by the computer user. This is evident in the time it takes for a typical computer to boot.

[0118] It is to be appreciated that a data storage controller (having an architecture as described herein) may employ a technique of data preloading to decrease the computer system boot time. Upon host system power-up or reset, the data storage controller will perform a self-diagnostic and program the programmable logic device (as discussed above) prior to completion of the host system reset (e.g., PCI bus reset) so that the logic device can accept PCI Bus commands after system reset. Further, prior to host system reset, the data storage controller can proceed to pre-load the portions of the computer operating system from the boot device (e.g., hard disk) into the on-board cache memory. The data storage controller preloads the needed sectors of data in the order in which they will be needed. Since the same portions of the operating system must be loaded upon each boot process, it is advantageous for the boot device controller to preload such portions and not wait until it is commanded to load the operating system. Preferably, the data storage controller employs a dedicated 10 channel of the DSP (with or without data compression) to pre-load computer operating systems and applications.

[0119] Once the data is preloaded, when the computer system bus issues its first read commands to the data storage controller seeking operating system data, the data will already be available in the cache memory of the data storage controller. The data storage controller will then be able to instantly start transmitting the data to the system bus. Before transmission to the bus, if the was stored in compressed format on the boot device, the data will be decompressed. The process of preloading required (compressed) portions of the operating system significantly reduces the computer boot process time.

[0120] In addition to preloading operating system data, the data storage controller could also preload other data that the user would likely want to use at startup. An example of this would be a frequently used application such as a word processor and any number of document files.

[0121] There are several techniques that may be employed in accordance with the present invention that would allow

the data storage controller to know what data to preload from the boot device. One technique utilizes a custom utility program that would allow the user to specify what applications/data should be preloaded.

[0122] Another technique (illustrated by the flow diagram of FIGS. 7a and 7b) that may be employed comprises an automatic process that requires no input from the user. With this technique, the data storage controller maintain a list comprising the data associated with the first series of data requests received by the data storage controller by the host system after a power-on/reset. In particular, referring to FIG. 7a, during the computer boot process, the data storage controller will receive requests for the boot data (step 70). In response, the data storage controller will retrieve the requested boot data from the boot device (e.g., hard disk) in the local cache memory (step 71). For each requested data block, the data storage controller will record the requested data block number in a list (step 72). The data storage controller will record the data block number of each data block requested by the host computer during the boot process (repeat steps 70-72). When the boot process is complete (affirmative determination in step 73), the data storage controller will store the data list on the boot device (or other storage device) (step 74).

[0123] Then, upon each subsequent power-on/reset (affirmative result in step 75), the data storage controller would retrieve and read the stored list (step 76) and proceed to preload the boot data specified on the list (i.e., the data associated with the expected data requests) into the onboard cache memory (step 77). It is to be understood that the depending on the resources of the given system (e.g., memory, etc.), the preloading process may be completed prior to commencement of the boot process, or continued after the boot process begins (in which case booting and preloading are performed simultaneously).

[0124] When the boot process begins (step 78) (i.e., the storage controller is initialized and the system bus reset is deasserted), the data storage controller will receive requests for boot data (step 79). If the host computer issues a request for boot data that is pre-loaded in the local memory of the data storage controller (affirmative result in step 80), the request is immediately serviced using the preloaded boot data (step 81). If the host computer issues a request for boot data that is not preloaded in the local memory of the data storage controller (negative determination in step 80), the controller will retrieve the requested data from the boot device, store the data in the local memory, and then deliver the requested boot data to the computer bus (step 82). In addition, the data storage controller would update the boot data list by recording any changes in the actual data requests as compared to the expected data requests already stored in the list (step 83). Then, upon the next boot sequence, the boot device controller would pre-load that data into the local cache memory along with the other boot data previously on the list.

[0125] Further, during the boot process, if no request is made by the host computer for a data block that was pre-loaded into the local memory of the data storage controller (affirmative result in step 84), then the boot data list will be updated by removing the non-requested data block from the list (step 85). Thereafter, upon the next boot sequence, the data storage controller will not pre-load that data into local memory.

[0126] VI. Quick Launch for Operating System, Application Programs and Loading

[0127] It is to be appreciated that the data storage controller (having an architecture as described herein) may employ a technique of data preloading to decrease the time to load application programs (referred to as “quick launch”). Conventionally, when a user launches an application, the file system reads the first few blocks of the file off the disk, and then the portion of the loaded software will request via the file system what additional data it needs from the disk. For example, a user may open a spreadsheet program, and the program may be configured to always load a company spreadsheet each time the program is started. In addition, the company spreadsheet may require data from other spreadsheet files.

[0128] In accordance with the present invention, the data storage controller may be configured to “remember” what data is typically loaded following the launch of the spreadsheet program, for example. The data storage controller may then proceed to preload the company spreadsheet and all the necessary data in the order in which such data is needed. Once this is accomplished, the data storage controller can service read commands using the preloaded data. Before transmission to the bus, if the preloaded data was stored in compressed format, the data will be decompressed. The process of preloading (compressed) program data significantly reduces the time for launching an application.

[0129] Preferably, a custom utility program is employed that would allow the user to specify what applications should be made ready for quick launch.

[0130] FIGS. 8a and 8b comprise a flow diagram of a quick launch method according to one aspect of the present invention. With this technique, the data storage controller maintains a list comprising the data associated with launching an application. In particular, when an application is first launched, the data storage controller will receive requests for the application data (step 90). In response, the data storage controller will retrieve the requested application data from memory (e.g., hard disk) and store it in the local cache memory (step 91). The data storage controller will record the data block number of each data block requested by the host computer during the launch process (step 92). When the launch process is complete (affirmative determination in step 93), the data storage controller will store the data list in a designated memory location (step 94). Then, referring to FIG. 8b, upon each subsequent launch of the application (affirmative result in step 95), the data storage controller would retrieve and read the stored list (step 96) and then proceed to preload the application data specified on the list (i.e., the data associated with the expected data requests) into the onboard cache memory (step 97). During the application launch process, the data storage controller will receive requests for application data (step 98). If the host computer issues a request for application data that is pre-loaded in the local memory of the data storage controller (affirmative result in step 99), the request is immediately serviced using the preloaded data (step 100). If the host computer issues a request for application data that is not preloaded in the local memory of the data storage controller (negative result in step 99), the controller will retrieve the requested data from the hard disk memory, store the data in the local memory, and then deliver the requested application data to the computer

bus (step 101). In addition, the data storage controller would update the application data list by recording any changes in the actual data requests as compared to the expected data requests already stored in the list (step 102).

[0131] Further, during the launch process, if no request is made by the host computer for a data block that was pre-loaded into the local memory of the data storage controller (affirmative result in step 103), then the application data list will be updated by removing the non-requested data block from the list (step 104). Thereafter, upon the next launch sequence for the given application, the data storage controller will not pre-load that data into local memory.

[0132] It is to be understood that the quick boot and quick launch methods described above are preferably implemented by a storage controller according to the present invention and may or may not utilize data compression/decompression by the DSP. However, it is to be understood that the quick boot and quick launch methods may be implemented by a separate device, processor, or system, or implemented in software.

[0133] VII. Content Independent Data Compression

[0134] It is to be understood that any conventional compression/decompression system and method (which comply with the above mentioned constraints) may be employed in the data storage controller for providing accelerated data storage and retrieval in accordance with the present invention. Preferably, the present invention employs the data compression/decompression techniques disclosed in the above-incorporated U.S. Pat. No. 6,195,024.

[0135] Referring to FIG. 9, a detailed block diagram illustrates an exemplary data compression system 110 that may be employed herein. Details of this data compression system are provided in U.S. Pat. No. 6,195,024. In this embodiment, the data compression system 110 accepts data blocks from an input data stream and stores the input data block in an input buffer or cache 115. It is to be understood that the system processes the input data stream in data blocks that may range in size from individual bits through complete files or collections of multiple files. Additionally, the input data block size may be fixed or variable. A counter 120 counts or otherwise enumerates the size of input data block in any convenient units including bits, bytes, words, and double words.

[0136] It should be noted that the input buffer 115 and counter 120 are not required elements of the present invention. The input data buffer 115 may be provided for buffering the input data stream in order to output an uncompressed data stream in the event that, as discussed in further detail below, every encoder fails to achieve a level of compression that exceeds an a priori specified minimum compression ratio threshold.

[0137] Data compression is performed by an encoder module 125 which may comprise a set of encoders E1, E2, E3 . . . En. The encoder set E1, E2, E3 . . . En may include any number “n” (where n may=1) of those lossless encoding techniques currently well known within the art such as run length, Huffman, Lempel-Ziv Dictionary Compression, arithmetic coding, data compaction, and data null suppression. It is to be understood that the encoding techniques are selected based upon their ability to effectively encode different types of input data. It is to be appreciated that a full

complement of encoders are preferably selected to provide a broad coverage of existing and future data types.

[0138] The encoder module **125** successively receives as input each of the buffered input data blocks (or unbuffered input data blocks from the counter module **120**). Data compression is performed by the encoder module **125** wherein each of the encoders **E1 . . . En** processes a given input data block and outputs a corresponding set of encoded data blocks. It is to be appreciated that the system affords a user the option to enable/disable any one or more of the encoders **E1 . . . En** prior to operation. As is understood by those skilled in the art, such feature allows the user to tailor the operation of the data compression system for specific applications. It is to be further appreciated that the encoding process may be performed either in parallel or sequentially. In particular, the encoders **E1** through **En** of encoder module **125** may operate in parallel (i.e., simultaneously processing a given input data block by utilizing task multiplexing on a single central processor, via dedicated hardware, by executing on a plurality of processor or dedicated hardware systems, or any combination thereof). In addition, encoders **E1** through **En** may operate sequentially on a given unbuffered or buffered input data block. This process is intended to eliminate the complexity and additional processing overhead associated with multiplexing concurrent encoding techniques on a single central processor and/or dedicated hardware, set of central processors and/or dedicated hardware, or any achievable combination. It is to be further appreciated that encoders of the identical type may be applied in parallel to enhance encoding speed. For instance, encoder **E1** may comprise two parallel Huffman encoders for parallel processing of an input data block.

[0139] A buffer/counter module **130** is operatively connected to the encoder module **125** for buffering and counting the size of each of the encoded data blocks output from encoder module **125**. Specifically, the buffer/counter **130** comprises a plurality of buffer/counters **BC1, BC2, BC3 . . . BCn**, each operatively associated with a corresponding one of the encoders **E1 . . . En**. A compression ratio module **135**, operatively connected to the output buffer/counter **130**, determines the compression ratio obtained for each of the enabled encoders **E1 . . . En** by taking the ratio of the size of the input data block to the size of the output data block stored in the corresponding buffer/counters **BC1 . . . BCn**. In addition, the compression ratio module **135** compares each compression ratio with an a priori-specified compression ratio threshold limit to determine if at least one of the encoded data blocks output from the enabled encoders **E1 . . . En** achieves a compression that exceeds an a priori-specified threshold. As is understood by those skilled in the art, the threshold limit may be specified as any value inclusive of data expansion, no data compression or expansion, or any arbitrarily desired compression limit. A description module **138**, operatively coupled to the compression ratio module **135**, appends a corresponding compression type descriptor to each encoded data block which is selected for output so as to indicate the type of compression format of the encoded data block. A data compression type descriptor is defined as any recognizable data token or descriptor that indicates which data encoding technique has been applied to the data. It is to be understood that, since encoders of the identical type may be applied in parallel to enhance encoding speed (as discussed above), the data compression type descriptor identifies the corresponding encoding tech-

nique applied to the encoded data block, not necessarily the specific encoder. The encoded data block having the greatest compression ratio along with its corresponding data compression type descriptor is then output for subsequent data processing, storage, or transmittal. If there are no encoded data blocks having a compression ratio that exceeds the compression ratio threshold limit, then the original unencoded input data block is selected for output and a null data compression type descriptor is appended thereto. A null data compression type descriptor is defined as any recognizable data token or descriptor that indicates no data encoding has been applied to the input data block. Accordingly, the unencoded input data block with its corresponding null data compression type descriptor is then output for subsequent data processing, storage, or transmittal.

[0140] Again, it is to be understood that the embodiment of the data compression engine of **FIG. 9** is exemplary of a preferred compression system which may be implemented in the present invention, and that other compression systems and methods known to those skilled in the art may be employed for providing accelerated data storage in accordance with the teachings herein. Indeed, in another embodiment of the compression system disclosed in the above-incorporated U.S. Pat. No. 6,195,024, a timer is included to measure the time elapsed during the encoding process against an a priori-specified time limit. When the time limit expires, only the data output from those encoders (in the encoder module **125**) that have completed the present encoding cycle are compared to determine the encoded data with the highest compression ratio. The time limit ensures that the real-time or pseudo real-time nature of the data encoding is preserved. In addition, the results from each encoder in the encoder module **125** may be buffered to allow additional encoders to be sequentially applied to the output of the previous encoder, yielding a more optimal lossless data compression ratio. Such techniques are discussed in greater detail in the above-incorporated U.S. Pat. No. 6,195,024.

[0141] Referring now to **FIG. 10**, a detailed block diagram illustrates an exemplary decompression system that may be employed herein or accelerated data retrieval as disclosed in the above-incorporated U.S. Pat. No. 6,195,024. In this embodiment, the data compression engine **180** retrieves or otherwise accepts compressed data blocks from one or more data storage devices and inputs the data via a data storage interface. It is to be understood that the system processes the input data stream in data blocks that may range in size from individual bits through complete files or collections of multiple files.

[0142] Additionally, the input data block size may be fixed or variable. The data decompression engine **180** comprises an input buffer **155** that receives as input an uncompressed or compressed data stream comprising one or more data blocks. The data blocks may range in size from individual bits through complete files or collections of multiple files. Additionally, the data block size may be fixed or variable.

[0143] The input data buffer **55** is preferably included (not required) to provide storage of input data for various hardware implementations. A descriptor extraction module **160** receives the buffered (or unbuffered) input data block and then parses, lexically, syntactically, or otherwise analyzes the input data block using methods known by those skilled in the art to extract the data compression type descriptor

associated with the data block. The data compression type descriptor may possess values corresponding to null (no encoding applied), a single applied encoding technique, or multiple encoding techniques applied in a specific or random order (in accordance with the data compression system embodiments and methods discussed above).

[0144] A decoder module 165 includes one or more decoders D1 . . . Dn for decoding the input data block using a decoder, set of decoders, or a sequential set of decoders corresponding to the extracted compression type descriptor. The decoders D1 . . . Dn may include those lossless encoding techniques currently well known within the art, including: run length, Huffman, Lempel-Ziv Dictionary Compression, arithmetic coding, data compaction, and data null suppression. Decoding techniques are selected based upon their ability to effectively decode the various different types of encoded input data generated by the data compression systems described above or originating from any other desired source.

[0145] As with the data compression systems discussed in U.S. application Ser. No. 09/210,491, the decoder module 165 may include multiple decoders of the same type applied in parallel so as to reduce the data decoding time. An output data buffer or cache 170 may be included for buffering the decoded data block output from the decoder module 165. The output buffer 70 then provides data to the output data stream. It is to be appreciated by those skilled in the art that the data compression system 180 may also include an input data counter and output data counter operatively coupled to the input and output, respectively, of the decoder module 165. In this manner, the compressed and corresponding decompressed data block may be counted to ensure that sufficient decompression is obtained for the input data block.

[0146] Again, it is to be understood that the embodiment of the data decompression system 180 of FIG. 10 is exemplary of a preferred decompression system and method which may be implemented in the present invention, and that other data decompression systems and methods known to those skilled in the art may be employed for providing accelerated data retrieval in accordance with the teachings herein.

[0147] VIII. Data Acceleration Controller

[0148] The block diagrams of FIGS. 11 and 12 illustrate disk architectures according to additional embodiments of the present invention. FIG. 11 and FIGS. 12 illustrate composite implementations of disk controller frameworks. Novel aspects of these architectures include, for example, the incorporation of a separate disk controller function and the addition of a data compression/decompression engine.

[0149] The benefits of such architectures are numerous, including, for example, increased reliability due to fewer components, higher throughput performance by eliminating unnecessary interfaces and handling less data due to compression, compression will aid in the thirst for greater disk densities, lower overall system costs due to less components and mechanical interfaces, and the ability to implement transparent pipelining to the host for increased system performance.

[0150] In general, the disk controller architectures in FIGS. 11 and 12 essentially provide an interface directly

from the disk to the host main bus. Any number of busses could be implemented such as PCI, PCMCIA, Pentium, VMEBUS, etc.

[0151] The disk controller framework illustrated in FIG. 11 comprises a disk controls and status module 1110, a cache management and interface control module 1120, a data cache 1130, a compression/decompression engine 1140, host interface busses 1150, a local high-speed bus 1160 and a data and command cache 1170. The data and command cache 1170 is operatively connected to the host interface busses 1150 (via a host interface buffered data bus 80) and to the data compression/decompression engine 1140 via the local bus 1160, to thereby allow data to be transferred to/from the host while the compression engine 1140 works on another buffered data block. The architecture of FIG. 11 maximizes pipelining and produces a maximum throughput.

[0152] FIG. 12 is a block diagram of a disk controller according to another embodiment of the present invention, comprising a disk controls and status module 1110, a cache management and interface control module 1120, a data cache 1130, a compression/decompression engine 1140, host interface busses 1150, and a local high-speed bus 1160. The architecture of FIG. 12 comprises a single buffer between the host and the compression engine 1140 and comprises a more cost effective architecture than that of FIG. 11, but with reduced throughput in certain applications.

[0153] In the frameworks of FIGS. 11 and 12, the data cache 1130 is connected to the data compression/decompression engine 1140 and the actual disk, which enable overlap of disk operations with data storage operations. The local high-speed bus 1160 in both architectures preferably operates primarily in block mode for optimal efficiency. The cache management & interface control module 1120, which communicates with the compression engine 1140 over a dedicated secondary high-speed bus 1190, controls the data flow through the disk system under commands from the compression/decompression engine 1140. This architecture allows data to flow from the disk and the host interface at the same time while the engine 1140 is operating on a separate block of data.

[0154] In a preferred embodiment of the present invention, the data compression/decompression engine 1140 employs the compression/decompression methods described in the above-incorporated U.S. Pat. No. 6,195,024, issued Feb. 27, 2001 to J. Fallon, U.S. Pat. No. 6,309,424, issued on Oct. 30, 2001 to J. Fallon, and/or the techniques described in U.S. patent application Ser. No. 10/016,355, filed on Oct. 29, 2001, which are all commonly assigned and incorporated herein by reference. It should be noted in the following discussion that the use of the term DSP (digital signal processor) may also refer to a plurality of DSP(s), one or a plurality of general purpose processors, finite state machines, micro code, logic, firmware, or software.

[0155] The data compression and decompression functions to accomplish data storewidth acceleration and or storage multiplication may be accomplished in either hardware, software, or any combination thereof.

[0156] In accordance with the present invention, minimizing the cost of the data storage and maximizing retrieval storewidth acceleration and or storage multiplication function is preferably achieved by sharing hardware and or software resources.

[0157] VIII. Data Accelerator Embedded in Disk and Other Data Storage Devices

[0158] FIG. 13 is a block diagram of a magnetic disk controller according to another embodiment of the present invention. It should be noted that although the exemplary embodiment of FIG. 13 describes a magnetic disk, the present invention is applicable to any mass storage device. A magnetic disk implementation according to an embodiment of the present invention comprises a host computer interface 1300 coupled to a compression/decompression engine 1301, a bi-directional cache 1302, a command and status control module 1303, a cache management/virtual file system 1304 and/or full host file system. The system further comprises high performance Giant Magneto Resistive Heads 1305, analog signal processing/data recognition module 1306, a decoding and tracking module 1307, a closed loop positioning module 1308, multistage linear actuators 1309 for positioning, ramp load/unloading of heads, closed loop brushless DC motors 1310 with liquid bearings and associated drivers 1311, and a power conversion and management system 1312. Additional functions (not shown) comprise dynamic head calibrations, static head calibrations, dynamic calibration functions, static calibration functions, and corrections for time, temperature, aging, stress, and radiation, along with error detection and correction functions.

[0159] In one embodiment of the present invention, a DSP is utilized for storewidth acceleration and or storage multiplication, as well as other disk functions or control thereof as mentioned above. For example, in one embodiment, the virtual file management system 1304 and/or cache system 1302 are implemented in the same DSP. In another embodiment, management of a bi-directional cache, separate read/write caches, and or compressed/uncompressed caches, or any combination thereof, is also managed by the DSP. In another embodiment of the present invention, the disk command and control functions are further implemented in the DSP.

[0160] In another embodiment of the present invention, one or more of the following functions may be advantageously implemented in the DSP: data recognition or correlation, decoding and tracking, closed loop servo positioning, multistage linear actuator control for positioning, ramp load/unload management and control, closed loop brushless DC motor control, liquid bearing monitoring (temperature, fluid level), and power conversion and management system.

[0161] Additional functions comprise on the fly dynamic head calibrations with read/write testing, or with pre-stored patterns on the disk, static head calibrations (performed at construction or periodically), dynamic and static calibration functions performed periodically or by command, along with corrections for time, temperature, aging, stress, radiation, along with canonical error detection and correction functions.

[0162] Accordingly, it is to be appreciated that either all the functions or any portion thereof might be implemented in the DSP, with the balance by any available means including dedicated logic, other processors, finite state machines, micro code, firmware, or software.

[0163] In yet another embodiment of the present invention, the virtual file management system 1304 may be

executed in an alternate processor, for example in the host computer, to offload the DSP or other disk processing for other data acceleration, bandwidth multiplication, and aforementioned disk functions. The virtual file management system 1304 is utilized to map uncompressed/unaccelerated blocks to the disk when they are compressed. With the present invention, a compressed data block occupies the same or less space than the original data. As such, the file management system 1304 must be aware of the new data size in order to store data efficiently on the disk. Current storage methods for disks, along with most other mass storage devices, store data in blocks that are either fixed or variable in size. The virtual file management 1304 system allows the appropriate data to be stored or retrieved even when the actual data is different in size than the original uncompressed data. Embedding the virtual file management system 1304 within the mass storage device has the advantage of reducing costs by sharing the processing and virtual file management mapping tasks. Virtual file management and data block mapping may be accomplished via look up tables or mathematical hashing functions. The mapping or hashing function virtualizes the physical disk or mass storage space. Assuming a 3:1 compression ratio, a one terabyte physical disk appears as a three terabyte virtual disk. The virtual file system 1304 allows uncompressed data in the three terabyte virtual space to be stored efficiently in the one terabyte physical space. Mapping data in small blocks has the advantage of being able to store, retrieve, or append with smaller more efficient units of data at the expense of larger look-up tables or more complex hashing functions.

[0164] In another embodiment of the present invention cache management is employed to manage data caching in both the physical space to the virtual space. Data retrieved from the mass storage device is stored compressed in a cache awaiting a cache hit. Typically, when data is retrieved, physically adjacent data or data known to be frequently referenced within the context of the retrieved data is also read from disk. Caches are typically uncompressed and thus require more memory and additional bandwidth for transfer. By caching compressed data on data reads, the cache is capable of storing a significantly larger quantity of data, making the likelihood of cache hits better and thus increasing system performance. Data is then decompressed when called for or in the background when the disk or controller is idle.

[0165] For data writes, the compressed data is stored in the cache awaiting synchronization with the mass storage disk or memory device. Again, the cache is capable of storing a much larger quantity of data (a function of compression ratio) and thus increases overall system performance.

[0166] In other embodiments, any element of the above-mentioned disk functions may be implemented wholly or partially with the disk DSP(s), host processor(s), alternate processor(s), or any combination thereof.

[0167] In yet another embodiment, the virtual file management system and the host operating system file management system are merged, with the results of the compression/decompression process utilized to minimize redundant disk management tasks. This function may be resident within the disk, within the host(s), or any combination thereof. The virtual file management system can consume significant processing resources, as such it may not always

be optimal to embed the virtual file system within the mass storage device. The processing can be resident in the host, thus allowing the disk processor to work on disk management and data acceleration tasks, split between the two, or wholly resident within the storage device.

[0168] In yet another embodiment of the present invention the data acceleration is performed solely with the host. For example in Network Attached Storage Services and other mass storage arrays it may be advantageous to have the data acceleration solely within one or more host CPUs. The CPU(s) can share workload with other functions, such as network interfacing and application processing. The virtual file system can also be allocated to one or more of the same CPUs or it can be performed within the mass storage or disk device.

[0169] In another embodiment of the present invention, the DSP utilized for the data storewidth acceleration or storage multiplication is also utilized to perform the monitoring and actuation functions for a bearing fluid replenishment system as described below. The monitoring and actuation functions for the bearing fluid replenishment system may be performed in a separate processor, DSP, one or a plurality of general purpose processors, finite state machines, micro code, logic, firmware, or software—in addition the control function may be implemented in the host computer.

[0170] A bearing fluid replenishment system according to one embodiment of the present invention is utilized with a bearing fluid reservoir to extend the life of the bearing system. An actuator is utilized to fill the fluid to a constant level. Further the fluid level or pressure within the system is maintained at a constant level. For example, in the case of fluid dynamic bearings, the pressure is preferably maintained a constant level to provide minimum frictional losses while still isolating the rotating parts from contact. In another aspect of the present invention, the pressure and or fluid level is adjusted to optimum as a function of spin rate. In yet another aspect of the present invention, the fluid level or pressure is adjusted to optimum as a function of temperature. In yet another aspect of the present invention, the fluid level and/or pressure is adjusted according to closed loop feedback from the motor current or power consumption to an optimal level. In yet another aspect of the present invention, any combination of the above parameters are utilized in optimal combination to achieve lowest wear, longest life, highest reliability, minimum breakaway friction, minimum frictional losses, lowest power consumption, etc.

[0171] Moving fluid from the reservoir to the bearing chamber and/or back is accomplished by any means including piezo actuators, expandable/compressible diaphragms, mechanical actuators, electromagnetic actuators, gas pressure by heat, or any other means.

[0172] Although illustrative embodiments have been described herein with reference to the accompanying drawings, it is to be understood that the present invention is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention. All such changes and modifications are intended to be included within the scope of the invention as defined by the appended claims.

What is claimed is:

1. A controller for controlling storage and retrieval of data to and from a data storage device, comprising:

a data compression/decompression engine for compressing data stored to the data storage device and for decompressing data retrieved from the data storage device;

a first cache, operatively connected to the storage device and the data compression/decompression engine, for temporary storage of (i) compressed data that is read from the data storage device and (ii) compressed data from the data compression/decompression engine that is to be written to the data storage device;

a second cache, operatively connected to the data compression/decompression engine and to a host interface bus, for temporary storage of (i) data that is received from a host system over the host interface bus for compression and storage and (ii) data that is to be sent to the host system over the host interface bus; and

a cache manager for controlling the first and second data caches under commands received from the data compression/decompression engine.

2. The controller of claim 1, wherein the controller comprises a disk controller and wherein the storage device comprises a hard disk.

3. The controller of claim 1, wherein the first cache, second cache and the data compression/decompression engine are operatively connected by a first local bus.

4. The controller of claim 1, wherein the data compression/decompression engine is connected to the cache manager over a dedicated bus.

5. The controller of claim 1, further comprising a virtual file system.

6. The controller of claim 1, wherein the controller comprises a magnetic disk controller.

7. The controller of claim 1, wherein the data compression/decompression engine, the first cache, and the second cache are implemented on a digital signal processor.

8. The controller of claim 1, wherein the controller provides accelerated data storage.

9. The controller of claim 1, wherein the controller provides accelerated data retrieval.

10. The controller of claim 1, wherein the first cache, second cache and data compression/decompression engine are embedded components of the controller.

11. A magnetic disk controller for controlling storage and retrieval of data to and from a magnetic disk, comprising:

an embedded data compression/decompression engine for compressing data stored to the magnetic disk and for decompressing data retrieved from the magnetic disk;

an embedded first cache, operatively connected to the magnetic disk and the data compression/decompression engine, for temporary storage of (i) compressed data that is read from the magnetic disk and (ii) compressed data from the data compression/decompression engine that is to be written to the magnetic disk;

an embedded second cache, operatively connected to the data compression/decompression engine and to a host interface bus, for temporary storage of (i) data that is received from a host system over the host interface bus

for compression and storage and (ii) data that is to be sent to the host system over the host interface bus; and

an embedded cache manager for controlling the first and second data caches under commands received from the data compression/decompression engine.

12. The magnetic disk controller of claim 11, wherein the first cache, second cache and the data compression/decompression engine are operatively connected by a first local bus.

13. The controller of claim 11, wherein the data compression/decompression engine is connected to the cache manager over a dedicated bus.

14. The controller of claim 11, further comprising an embedded virtual file system.

15. The controller of claim 11, wherein the data compression/decompression engine, the first cache, and the second cache are implemented on a digital signal processor.

16. The controller of claim 1, wherein the controller provides accelerated data storage.

17. The controller of claim 1, wherein the controller provides accelerated data retrieval.

18. A magnetic disk controller for controlling storage and retrieval of data to and from a magnetic disk, comprising:

an embedded data compression/decompression engine for compressing data stored to the magnetic disk and for decompressing data retrieved from the magnetic disk;

an embedded bi-directional cache for temporary storage of (i) compressed data that is read from the magnetic disk and (ii) compressed data from the data compression/decompression engine that is to be written to the magnetic disk;

an embedded cache manager for controlling the bi-directional cache; and

an embedded virtual file management system for mapping compressed data blocks stored on the disk to corresponding uncompressed data blocks.

* * * * *