960700-7 IPOS

ORIGINAL

This Sheet Has Been Replaced

REPLACED SHEETS SG
☑ Description : paged ....1.... to paged ..19.....
☑ Claim(s) : paged ..20.... to paged ..23.....
☑ Drawing(s) : paged ....1/2.... to paged ....2/2....
☑ Abstract : paged ..24.... to paged .....
DATE OF RECEIPT .....

A DATA RECOVERY CHANNEL IN A FAULT TOLERANT DISK DRIVE ARRAY AND A METHOD OF CORRECTING ERRORS THEREIN

This invention relates to data recovery channels in fault tolerant disk drive arrays and methods of correcting errors therein.

A peripheral data storage system is a type of secondary memory accessible by a data processing system, such as a computer. Memory, in a computer, is where programs and work files are stored as digital data. Computer memory can include either, and commonly includes both, moving type memory and non-moving type memory. Non-moving memory is typically directly addressed by the computer's central processing unit. Moving memory, such as magnetic disk drives and magnetic tape, is not directly addressed, and is commonly referred to as secondary memory or peripheral memory.

Moving memory typically has much greater data storage capacity than directly addressed memory and has much longer access times. Moving memory is also typically not volatile. That is, it survives turning the computer off. Non-moving type memory is typically faster and more expensive per unit of memory than moving type memory, and has less capacity. Moving type memories are generally used for long term storage of large programs and substantial bodies of information, such as a data base file, which are not in constant use by the computer, or which are too bulky to provide short term direct access memory capacity for.

The storage media of the moving type memory are physically alterable objects. That is to say, they can be magnetised, grooved, pitted or altered in some detectable fashion to record information. Preferably

This Sheet Has Been Replaced



ORIGINAL

the storage media is at the same time physically resilient, portable, cheap, of large capacity, and resistant to accidental alteration. An example of an analogous medium is a phonograph record where a wavy spiral groove represents an analog information signal. The various species of storage media used in moving type memory for computers include magnetic tape, floppy disks, compact disk-ROM, write-once, read-many optical disks, and, most recently, erasable magneto-optic disks. Each of these storage media exhibit detectable physical changes to the media representing binary data. To read, and where applicable to erase and write data to the media, mechanical apparatus is provided which can be directed to the proper location on the physical media and carry out the desired function.

A magnetic disk drive includes a transducer, a magnetic media disk and associated electronic circuitry to drive or monitor the transducer and transfer the data between the physical medium and the computer to which the drive is connected. The conversion of data from electronic signal to physical feature for storage, extended retention of the data as physical features, and the conversion of physical feature back to electronic signal offers numerous opportunities for error to be introduced to the data record. Transducer and disk present to one another surfaces that are in essentially constant motion. While the inter-action between transducer and media is magnetic, the inter-action between their surfaces is a mechanical one, affected over time by factors such as friction, wear, media flaking and collisions between the transducer and the media. All of these occurrences can be sources of error. In addition, both individual transducers and magnetic media surfaces are subject to failures, such as

This Sheet Has Been Replaced

ORIGINAL

opening of a transducer electro-magnet winding or imperfections in the magnetic media surface. Mechanical failure can affect the magnetic inter-action of the transducer and the media, and consequently can affect the ability of data recovery circuitry to read data records from the media surface. The faults described above, and others too numerous to discuss in any detail here, can result in loss of data records ranging from one bit to all of the data of the record stemming from failure of the entire drive. The possibility of error in read back, an error or loss of data on the disk, has led to the incorporation of error detection and correction techniques in disk drives.

Data in digital processing systems is typically stored to a disk drive by sectors. Such sectors will include certain redundant information to be used for checking the accuracy of the record and possible correction of the sector upon read back. An example of such redundant information is "Error Checking and Correcting Code", sometimes referred to as "Error Correction Code" or, "ECC".

An ECC for a sector or record will literally comprise data bits supplementing the regular data bits of the record. Where an ECC is used each record conforms to specific rules of construction which permit use of the supplemental bits to detect and, under certain circumstances, correct errors in the record. In actual use, error syndromes are generated from data and the ECC upon read back of the sector. Non-zero error syndromes indicates that error is present. Where error is "random", which for the purpose of this specification means error within the capacity of the ECC to correct, the error syndromes can be used to correct the sector. However, operation on the error syndromes to correct

This Sheet Has Been Replaced

ORIGINAL

error is relatively time consuming in terms of performance of the disk drive in a computer. Analysis of the error syndromes to determine whether error is correctable, that is whether error is random or massive, can be done in a relatively short period of time.

A second example of redundant data is parity data. Parity data is generated for a logically associated group of binary bits, typically by a modulo addition operation on the group to generate a check digit. An example of a check digit in a binary system is where the digits are logically "ORed", generating a "1" if the number of "1's" in the group is odd or a "0" otherwise. Parity, absent information indicating location of an error, is usable only to identify the existence of error. However, given independent identification of location of the error in a group, the use of parity permits rapid correction of the error. Only one bit of error per parity bit can be corrected.

A computer can use more than one disk drive for the storage of data. Alternatively, a plurality of disk drives can be organised to operate together and thereby appear to the computer as one peripheral storage unit. The term disk drive array is used in this specification to indicate a group of disk drives operating in a parallel, synchronous fashion allowing transfer of data bits of record in parallel to the individual drives of the array and appearing to the computer as a single, data storage peripheral device. An interface operating between the disk drive array and the data processing system transfers data to and from storage in parallel to increase data transmission band width.

Such parallel, synchronised disk drives have characteristics offering opportunity for improvement in redundancy for stored data. Simplistically, data on

This Sheet Has Been Replaced



ORIGINAL

one disk drive can be mirrored on another. In US-A-4,775,978 there is disclosed a data error correction system for a mass data storage system comprising an array of synchronised disk data storage units. The data storage system receives data blocks from a host data processing system for storage. A data block divider stripes the data among a plurality of the disk data storage units as columns. A data word is divided among the plurality of disk drives one bit to each drive. Data stored to the same address on each disk are supplemental by a column of parity data stored to yet another data storage unit. Each data column, assigned to a particular disk drive, is supplemented with an error correction code. Upon read back, two levels of data error correction are provided. Data columns are read and stored to buffers. Error syndromes are generated upon read back and used to correct random error. When a data storage unit fails, or where error correction code for a column is inadequate to correct errors in a column, the column (i.e. data storage unit) where the error occurs is identified and the parity data in combination with the data of the other columns is used to re-construct data in the missing column. The system provides a highly fault tolerant data storage system.

The present invention seeks to provide an improved data recovery channel applicable to a data error correction system such as that disclosed in US-A-4,775,978.

According to one aspect of the present invention there is provided a data recovery channel in a fault tolerant disk drive array, the disk drive array including a plurality of disk drives and an input interface for receiving data records from a data processing system, the input interface organising each

This Sheet Has Been Replaced

ORIGINAL

data record received into data sectors for storage on the disk drives, identifying a plurality of sectors as a related group and generating a parity sector for the group, where each data and parity sector includes an appended error correction code, the group of data and parity sectors being stored by striping across the available disk drives, the data recovery channel being characterised by comprising: a disk drive controller associated with each disk drive including means for recovering sectors from the disk drive, means for determining error syndromes for each sector as recovered, and means for detecting a disk drive fault; a data buffer for retaining the data and parity recovered with each sector; an error status buffer for retaining indication of non-zero error syndromes generated upon recovery of a sector or occurrence of a disk fault associated with attempted recovery of a sector; an error syndrome buffer for retaining any non-zero error syndromes generated during recovery of data and parity sectors; syndrome processing means connected to the error syndrome buffer for determining the correctability of the data of a sector, and for correcting the data in the data buffers, if possible; parity correction means for receiving associated groups of data and parity information from the data buffers and correcting the data from up to one sector; and an array controller for operating upon the error status buffer and indications of correctability of data from the syndrome processing means to direct the correction method for each sector of a group.

Preferably the array controller further comprises: means for monitoring the error status buffer; means for transferring the data and parity for a group of sectors to the parity correction means when the number of errors

This Sheet Has Been Replaced

ORIGINAL

indicated for a group of sectors is or is reduced to one or less; and means for re-setting the status indications for a sector in the error status buffer where analysis of the error syndromes for the sector shows that the data from the sector can be corrected. Each status buffer may have a memory location for each sector recovered of a data record.

The array controller may further comprise means for initiating a disk read re-try when more than one disk fault is indicated for a group, and means for indicating a read failure when a disk read re-try fails.

The array controller may further comprise means for initiating a disk read re-try when the number of sectors having a fault cannot be reduced to one in number; and means for indicating a read failure when a disk read re-try fails.

The data recovery channel may be in combination with a disk drive array, the disk drive array comprises synchronised, fault independent disk drive units.

According to another aspect of the present invention there is provided a method of correcting errors appearing in data recovered from a plurality of disk drives in an array, wherein the data is organised on the disk by sectors and groups of sectors form related rows, each row having a parity data sector and each sector including an error correction code, the method being characterised by comprising the steps of: (a) generating error syndromes for each sector upon recovery of data from the disk drives; (b) storing non-zero error syndromes for use in correction of data; (c) generating and storing error status indications for each sector recovered, the indications including a no error condition, a non-zero error syndrome condition and a drive fault condition; (d) initiating a

This Sheet Has Been Replaced

ORIGINAL

correction protocol including the steps of:
(1) checking the error status for each sector for a row and continuing to the next row of sectors if no errors are present, (2) if error is indicated for one sector, correcting the sector using parity correction and returning to step (1) for the next row of sectors, (3) if error is indicated for more than one sector of the row, using the error syndromes to locate sectors correctable with the error syndromes and correcting the data for those sectors until one sector having error is left, and then returning to step (2), (4) where the number of errors in a row cannot be reduced to one, initiating a re-try read of a data storage unit from which a sector having not correctable error was recovered and returning to step (1), if successful, and (5) indicating a read failure and aborting the data recovery attempt if step (4) is not successful.

The method may include the additional step after step (d) (4) of initiating a recovery of the disk drive and returning to step (1) if successful.

The invention is illustrated, merely by way of example, in the accompanying drawings, in which:-

Figure 1 is a block diagram of a data processing system;

Figure 2 is a block diagram of a data recovery channel according to the present invention; and

Figure 3 illustrates organisation of data buffers of the data recovery channel of Figure 2.

Figure 1 shows a data processing system 10 with which a data recovery channel according to the present invention can be advantageously used. The description of the architecture of the data processing system 10 is intended only to give an environment for explanation of the present invention and is not intended as a

This Sheet Has Been Replaced



ORIGINAL

description of a particular computer architecture with which the present invention is used. The data processing system 10 is entirely conventional and includes a central processing unit 12, a direct access memory 14 and a data storage peripheral 16. The central processing unit 12 stores data and programming steps to, and recovers data and programming steps over a data bus 18 from the memory 14 and the data storage peripheral 16. The data storage peripheral 16 can transfer data and programming steps over the data bus 18 directly to the memory 14 or to the central processing unit 12.

The central processing unit 12 exercises control over the memory 14 and the data storage peripheral 16 over a control bus 20 and an address bus 22, thereby directing the timing of the transfer of data and locations to which the data is delivered or from which it is called. The data storage peripheral 16 may be a tape drive, a disk drive, or some other form of indirectly addressed, mass data storage structure. Communication between the data storage peripheral 16 and the rest of the data processing system 10 is through an input/output interface 24. Where the data storage peripheral 16 is in an array of disk drives, a data recovery channel according to the present invention can be advantageously incorporated in the interface 24.

Figure 2 illustrates the data storage peripheral 16, the interface 24 and a data recovery channel 26 according to the present invention. The data storage peripheral 16 has n+1 disk drives 28-0 to 28-n. Each disk drive is under the direct control of one of n+1 drive controllers 30-0 to 30-n. The drive controllers 30-0 to 30n, among other functions, provide a passage for data between the disk drives 28-0 to 28-n and data

This Sheet Has Been Replaced

ORIGINAL

buffers 32-0 to 32-n, respectively.

Data records on the disk drives 28-0 to 28-n are organised by sectors. A sector includes preamble information, data which is derived from the record applied to the data storage peripheral 16 for storage and an error detection and correction code generated from the data. The generation of such sectors and the operation of writing the sectors to the disk drives 28-0 to 28-n forms no part of the present invention. The sectors stored on the disk drives 28-0 to 28-(n-1) contain the data received over the data bus 18 for storage and are referred to here as data sectors. The sectors stored on the disk drive 28-n contain parity data related to the data stored in the first n drives and are referred to here as parity sectors. Parity sectors are otherwise like data sectors in that they include error correction codes.

Read back of data records from the disk drives 28-0 to 28-n is done simultaneously and in parallel by the drive controllers 30-0 to 30-n upon receipt of a read command from an array controller 34 in the interface 24. Groups of sectors, designated here as rows, are inter-related. A row includes a plurality of data sectors, one from each disk, and a parity sector. The parity data in the parity sector is generated from logically associated bits between the data sectors. For example, the first parity bit in the parity sector may be parity for the first data bit from each data sector, the second parity bit in the parity sector is calculated over the second data bits of the data sectors and so on. The sectors of a row are stored to the same logical addresses in the data buffers 32-0 to 32-n. During execution of a read back of data records, the drive controllers 30-0 to 30-n operate on the recovered data

This Sheet Has Been Replaced



ORIGIN.

and ECC to generate error syndromes for each sector. The error syndromes, if non-zero, are loaded into known locations in an error syndrome buffer 36 for possible later use. A non-zero error syndrome set for a sector indicates presence of a code error in the sector, which may or may not be correctable using the error syndromes. The drive controllers 30-0 to 30-n also monitor the disk drives 28-0 to 28-n for indications of drive failure or fault.

A data record includes a plurality of rows of sectors. During read back of a record from the disk drives 28-0 to 28-n by the drive controllers 30-0 to 30-n, the drive controllers generate error status indications for each sector read and store the error status indications in status buffers 38-0 to 38-n. An error status indication of 0,0 indicates a sector free of detected error; an error status indication of 0,1 indicates a sector for which non-zero error syndromes were generated upon read back; an error status record of 1,0 indicates occurrence of a disk fault. An error status record of 1,1 is preferably not used. In summary, the first digit of the record is used to indicate a disk fault and the second digit is used to indicate a code fault in a sector.

Upon completion of reading of the sectors from the disk drives 28-0 to 28-n, the data from each data and parity sector will be stored to a pre-determined location in the data buffers 32-0 to 32-n, an error status indication for each data and parity sector will be stored to a logically associated address in the status buffers 38-0 to 38-n, and the non-zero error syndromes for each sector having code errors will be stored in the syndrome buffer 36.

The array controller 34 is a programmed micro-

This Sheet Has Been Replaced



ORIGINAL

computer and controls the error correction processing in the data recovery channel 26 by monitoring the status buffers 38-0 to 38-n and communicating with syndrome processing circuitry 42 over a bus 43. The data recovery channel 26 includes parity correction circuitry 40, the error syndrome buffer 36, the error syndrome processing circuitry 42 and the data buffers 32-0 to 32-n. The error syndrome processing circuitry 42 may be and preferably is combined with the array controller 34 by programming the array controller to perform the functions of the error syndrome processing circuitry. Illustration of the array controller 34 and the syndrome processing circuitry 42 as separate components is an aid to understanding and a feasible hardware configuration.

During a read operation, the array controller 34 monitors the error status records stored to the status buffers 38-0 to 38-n. A read operation is allowed to continue as long as no more than one drive fault error occurs in any given row of the recovered record, i.e. the read operation continues as long as the error status 1,0 does not repeat at the same logical address across the status buffers 38-0 to 38-n. Where two drive faults are indicated for a row, the read is aborted and extra-ordinary recovery techniques may be employed, such as a re-read attempt or an attempted recovery of the disk drive. If all such techniques fail the read operation is aborted and a drive array failure condition is indicated.

Absent indication of simultaneous drive faults for two or more of the disk drives 28-0 to 28-n, the read operation continues until completion of recovery of an entire record. The array controller 34 then directs error correction for the recovered record, row by row. As previously noted, both operation on parity data and

This Sheet Has Been Replaced

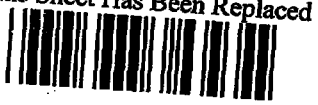
ORIGINAL

on the error syndromes is available. The array controller 34 interrogates the error status words for a row and determines the number of sectors for which drive fault and ECC error is present. Where error of either type is indicated as occurring in none or only one sector for the row, all of the sectors of the row are transferred from the data buffers 32-0 to 32-n to the parity correction circuitry 40 for correction, if required, through a sector divider 44 for reconstruction of the original words of the record and transmission out to the data bus 18. Parity correction provides superior speed over an error syndrome operation for correction of a sector, therefore it is not even determined whether a sector having error is correctable using the error syndromes. The array controller 34 indicates to the parity correction circuitry 40 the sector having error. Correction involves the regeneration of the data contained in the sector having the error from the remaining data and parity information.

Where positive error status indications exist for two or more sectors of a row, utilisation of the error syndromes for correction of one or more sectors is attempted. The array controller 34 first locates those sectors having code errors, indicated in the status buffers 38-0 to 38-n by 0,1 records for each sector. The array controller 34 then directs the syndrome processing circuitry 42 to analyse the error syndromes for a given sector, typically beginning with the error syndromes for the first sector for which an error status of 0,1 exists. By "first" will typically be meant the sector associated with the lowest numbered data buffer (e.g. data buffer 32-1 before data buffer 32-3).

The first operation performed by the syndrome

This Sheet Has Been Replaced



ORIGINAL

processing circuitry 42 is a determination whether the error syndromes for a sector can be used to correct the data. If the error is correctable the error syndrome processing circuitry 42 corrects the data in the appropriate data buffer 32-0 to 32-n. The array controller 34 re-sets the status buffer of the associated error status word to indicate no error in the sector. The array controller 34 will continue correction of sectors using error syndromes, if possible, until only one defective sector remains for a row. When only one error containing sector remains, regardless of whether the error can be corrected using error syndromes or not, the entire row is transferred to the parity correction circuitry 40 for parity correction of the sector containing error. The corrected row is then transmitted to the sector divider 44 for re-construction of the original words of the data record and on to the data bus 18. The array controller 34 then advances the correction algorithm to the next row. Correction continues until all sectors of the file are corrected and transmitted on to the data bus 18.

Should two or more sectors of a row have massive errors, i.e. error exceeding the capacity of the error syndromes, correction by the array controller 34 is not attempted. Instead, recovery is aborted and the data from at least one of the affected disk drives 28-0 to 28-n is re-read. Numerous conventional data recovery methods may be attempted, before the recovery attempt is aborted.

The original data words of a record are re-assembled before transmission on the data bus 18. The sector divider 44 is a bi-directional element which receives data words over the data bus 18 and routes each subsequent bit or byte of the word to a succeeding disk

This Sheet Has Been Replaced



ORIGINAL

drive, wrapping around to the first disk drive after assignment of n bits (or bytes). The sector divider 44 also operates to re-assemble the original words when a record is transmitted out of the parity correction circuitry 40.

Figure 3 illustrates the storage of an exemplary data record among the memory units comprising the data buffers 32-0 to 32-n after recovery of the record from the disk drives 28-0 to 28-n. The exemplary record includes 4 x n data sectors, stored four (numbered 0 to 3 for each buffer) to a data buffer, and four parity sectors (numbered 0 to 3) stored in a data buffer 32-0 to 32-n. Like numbered data and parity sectors belong to a row, e.g. sectors 1 of the data bufers 32-0 to 32-n are a row, sectors 3 of the data buffers 32-0 to 32-n are another row. One data bit from each data sector in a row relates to a particular parity bit from the parity sector for the row. In a typical arrangement only the original data of the sector having bits numbered 0 to 255 are in a data buffer after down loading a sector from its disk drive.

The status buffers 38-0 to 28-n consist of storage locations corresponding one for one with each location for sectors to be stored in data buffers 32-0 to 32-n. As recovery of data and parity sectors is attempted from the disk drives 28-0 to 28-n and the data or parity transferred to locations in the data buffers 32-0 to 32-n, error status indication words for the sectors are generated and stored in the corresponding memory locations of the status buffers 38-0 to 38-n. The first bit location of the error status indication words are the indicators of drive fault, the second bit location of the words are indication of non-zero error syndromes. Initiation of data recovery and transfer in the data

This Sheet Has Been Replaced



ORIGINAL

buffers results in all locations of the error status buffers 38-0 to 38-n being re-set. An error status of 0,0, or re-set value, is indication of no error in the corresponding data or parity sector.

Sector 3 stored in the data buffer 32-0 illustrates indication of non-zero error syndromes for a sector. The status buffer 38-0 has four error status locations 38-0(0) to 38-0(3). The error condition of sector 3 of data buffer 32-0 is reflected by the second bit of location 38-0(3) which is set at 1. Similarly, sector 1 of buffer 32-n shows a drive fault error. Location 28-n(1) shows a first bit location as 1 indicating the error condition of the sector.

The present data recovery channel corrects the data of a sector recovered from a disk drive, wherein sectors include error correction codes, groups of sectors are related by the existence of parity data for the group. Two methods of correction for sectors are used, one being parity, the other method using error syndromes. Correction employs the following priority if no more than one sector is lost per group due to drive failure:

(1) Correction of a sector by parity within a parity related group;

(2) Correction of sectors by error correction syndromes until the number of sectors evidencing error in a parity related group is reduced to one, then correcting that sector with parity;

(3) If two or more sectors evidence drive fault and a non-correctable error condition in a parity related group, re-trying the read for at least one disk drive.

This Sheet Has Been Replaced



ORIGINAL

(4) If the re-try read fails, attempting a "recover" of at least one disk drive (if the disk controller admits the command); and

(5) Signalling the data storage peripheral "FAIL" condition if all the above fail.

As will be appreciated, the data recovery channel preferentially operates on parity data for error correction over utilisation of error syndromes for error correction, though both remain available. In summary, absent indication of errors from more than one sector for a row of sectors, correction of error in the defective sector is made by use of parity information. Where error is indicated for more than one sector in a row of sectors, correction using the error syndromes is attempted, sector by sector, until the number of sectors in the row having error is reduced to one. Parity is used to correct the remaining defective sector.

Error syndromes are generated for each sector upon recovery of the sector from a disk drive. A non-zero error syndrome set for a sector indicates the possible presence of error and results in generation of a signal indicating such occurrence. A failure to recover a sector results in generation of a signal indicating a drive fault. A status buffer saves indications of fault from a failing drive, the occurrence of non-zero error syndromes or an error free condition for each recovered sector. An error syndrome buffer retains the error syndromes if non-zero.

An array controller monitors the status buffer and continues the read operation as long as no more than one drive fault is indicated with respect to a row of data

This Sheet Has Been Replaced

ORIGINAL

sectors. Upon completion of reading of all the rows of a file, the data recovery channel has access to the sector status information, recovered data sectors and to the error syndromes. Error correction of the sectors, re-construction of the original data words and transmission out of the original record is then attempted.

The array controller interrogates the sector status information and either executes, or directs execution, of error correction routines to restore the original data. Utilisation of the error syndromes for data re-construction is deferred in favour of parity re-construction if errors occur in only one sector for each row of sectors.

Error status for each sector is retained in an error status buffer. Where the status buffer indicates no errors for a row, the original data words are re-constructed and the data transmitted to the utilising computer. Where error of either type is indicated for one sector of a row, parity is used to re-construct the affected sector. Upon indication of error for more than one sector of a row, the array controller attempts restoration of the defective sectors of the row utilising error syndrome data. The array controller or syndrome processing circuitry which may be a program executed by the array controller, operates on the error syndromes for a first defective sector from the row to determine if the sector is correctable. If the sector is correctable, correction is carried out, the restored data is stored to the data buffer and the error status buffer for the sector is re-set. Once the number of sectors in a row with error of either type is reduced to one, parity correction is done and the entire row is transferred out to the utilising computer. Where the

This Sheet Has Been Replaced

ORIGINAL

number of sectors in a row with error cannot be reduced below two using error syndromes, the read operation is interrupted and the more extensive data recovery methods, directed at the affected disk drives are attempted, e.g. a re-read of disk drives.

This Sheet Has Been Replaced

ORIGINAL

CLAIMS

1. A data recovery channel in a fault tolerant disk drive array (16), the disk drive array including a plurality of disk drives (28-0 to 28-n) and an input interface (24) for receiving data records from a data processing system (10), the input interface (24) organising each data record received into data sectors for storage on the disk drives, identifying a plurality of sectors as a related group and generating a parity sector for the group, where each data and parity sector includes an appended error correction code, the group of data and parity sectors being stored by striping across the available disk drives, the data recovery channel being characterised by comprising: a disk drive controller (30-0 to 30-n) associated with each disk drive (28-0 to 28-n) including means for recovering sectors from the disk drive, means for determining error syndromes for each sector as recovered, and means for detecting a disk drive fault; a data buffer (32-0 to 32-n) for retaining the data and parity recovered with each sector; an error status buffer (38-0 to 38-n) for retaining indication of non-zero error syndromes generated upon recovery of a sector or occurrence of a disk fault associated with attempted recovery of a sector; an error syndrome buffer (36) for retaining any non-zero error syndromes generated during recovery of data and parity sectors; syndrome processing means (42) connected to the error syndrome buffer for determining the correctability of the data of a sector, and for correcting the data in the data buffers, if possible; parity correction means (40) for receiving associated groups of data and parity information from the data buffers and correcting the data from up to one sector;

* C L M S *

This Sheet Has Been Replaced

ORIGINAL

and an array controller (34) for operating upon the error status buffer and indications of correctability of data from the syndrome processing means to direct the correction method for each sector of a group.

2. A data recovery channel as claimed in claim 1, characterised in that the array controller (34) further comprises: means for monitoring the error status buffer (38-0 to 38-n); means for transferring the data and parity for a group of sectors to the parity correction means (40) when the number of errors indicated for a group of sectors is or is reduced to one or less; and means for re-setting the status indications for a sector in the error status buffer (38-0 to 38-n) where analysis of the error syndromes for the sector shows that the data from the sector can be corrected.

3. A data recovery channel as claimed in claim 2 characterised in that each status buffer (38-0 to 38-n) has a memory location for each sector recovered of a data record.

4. A data recovery channel as claimed in claim 3 characterised in that the array controller 34 further comprises: means for initiating a disk read re-try when more than one disk fault is indicated for a group, and means for indicating a read failure when a disk read re-try fails.

5. A data recovery channel as claimed in claim 3 or 4 characterised in that the array controller further comprises: means for initiating a disk read re-try when the number of sectors having a fault cannot be reduced to one in number; and means for indicating a read failure when a disk read re-try fails.

6. A data recovery channel as claimed in any preceding claim in combination with a disk drive array characterised in that the disk drive array comprises

* C L M S *

This Sheet Has Been Replaced

ORIGINAL

synchronised, fault independent disk drive units.

7. A method of correcting errors appearing in data recovered from a plurality of disk drives (28-0 to 28-n) in an array, wherein the data is organised on the disk by sectors and groups of sectors form related rows, each row having a parity data sector and each sector including an error correction code, the method being characterised by comprising the steps of: (a) generating error syndromes for each sector upon recovery of data from the disk drives; (b) storing non-zero error syndromes for use in correction of data; (c) generating and storing error status indications for each sector recovered, the indications including a no error condition, a non-zero error syndrome condition and a drive fault condition; (d) initiating a correction protocol including the steps of: (1) checking the error status for each sector for a row and continuing to the next row of sectors if no errors are present, (2) if error is indicated for one sector, correcting the sector using parity correction and returning to step (1) for the next row of sectors, (3) if error is indicated for more than one sector of the row, using the error syndromes to locate sectors correctable with the error syndromes and correcting the data for those sectors until one sector having error is left, and then returning to step (2), (4) where the number of errors in a row cannot be reduced to one, initiating a re-try read of a data storage unit from which a sector having not correctable error was recovered and returning to step (1), if successful, and (5) indicating a read failure and aborting the data recovery attempt if step (4) is not successful.

* C L M S *

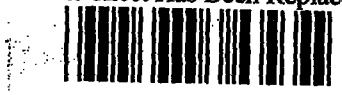
This Sheet Has Been Replaced

ORIGINAL

8. A method as claimed in claim 7 characterised by including an additional step after step (d) (4) of initiating a recovery of the disk drive and returning to step (1) if successful.

* C L M S *

This Sheet Has Been Replaced



ORIGINAL

ABSTRACT

A DATA RECOVERY CHANNEL IN A FAULT TOLERANT DISK DRIVE ARRAY AND A METHOD OF CORRECTING ERRORS THEREIN

A disk drive array includes a plurality of disk drives (28-0 to 28-n) in an input interface (24) for receiving data records from a data processing system (10). The input interface (24) organises each data record received in the data sectors for storage on the disk drives, identifies a plurality of sectors as a related group and generates a parity sector for the group, where each data and parity sector includes an appended error correction code. The groups of data and parity sectors are stored by striping across the available disk drives. A data recovery channel according to the present invention comprises: a disk drive controller (30-0 to 30-n) associated with each disk drive (28-0 to 28-n); a data buffer (32-0 to 32-n) retaining the data and parity recovered with each sector; an error status buffer (38-0 to 38-n) retaining indication of non-zero error syndromes generated upon recovery of a sector or occurrence of a disk fault associated with attempted recovery of a sector; an error syndrome buffer (36) for retaining any non-zero error syndromes generated during recovery of data and parity sectors; syndrome processing circuitry (42) connected to the error syndrome buffer for determining the correctability of the data of a sector, and for correcting the data in the data buffers, if possible; parity correction circuitry (40) receiving associated groups of data and parity information from the data buffers and correcting the data from up to one sector;

This Sheet Has Been Replaced



ORIGINAL

and an array controller (34) operating upon the error status buffer and indications of correctability of data from the syndrome processing circuitry to direct the correction method for each sector of a group.

[Figure 2]

This Sheet Has Been Replaced

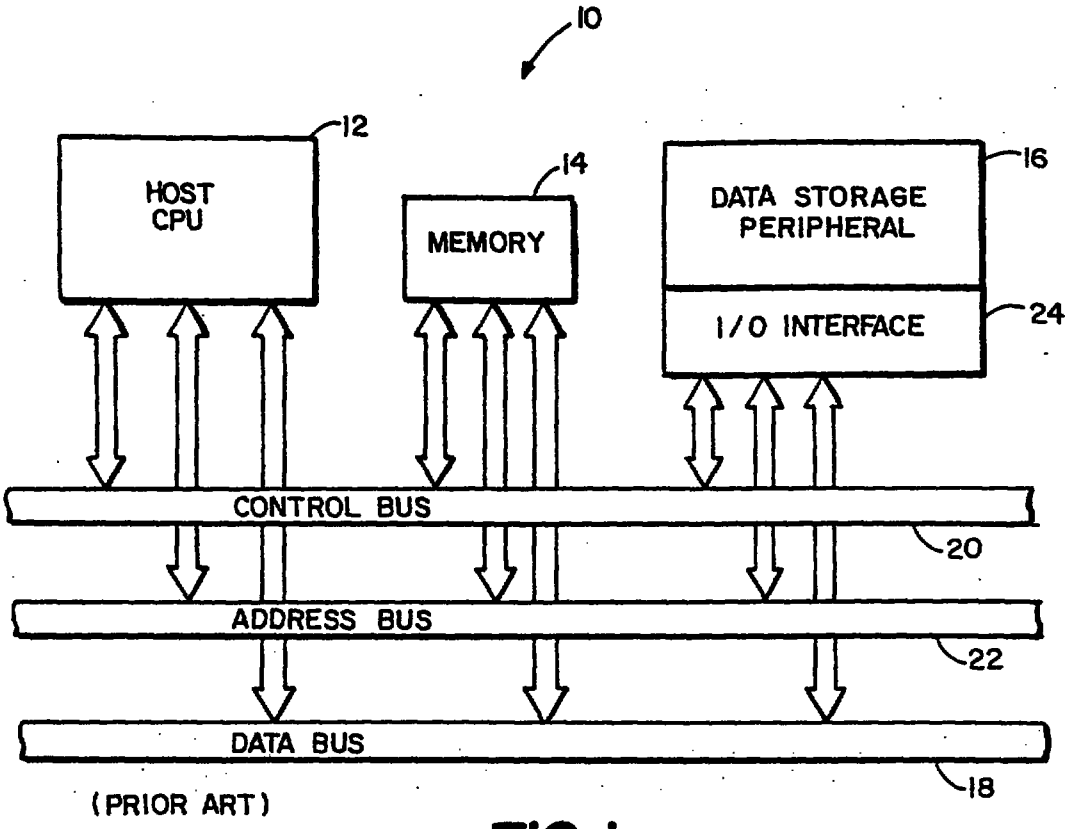


FIG. 1

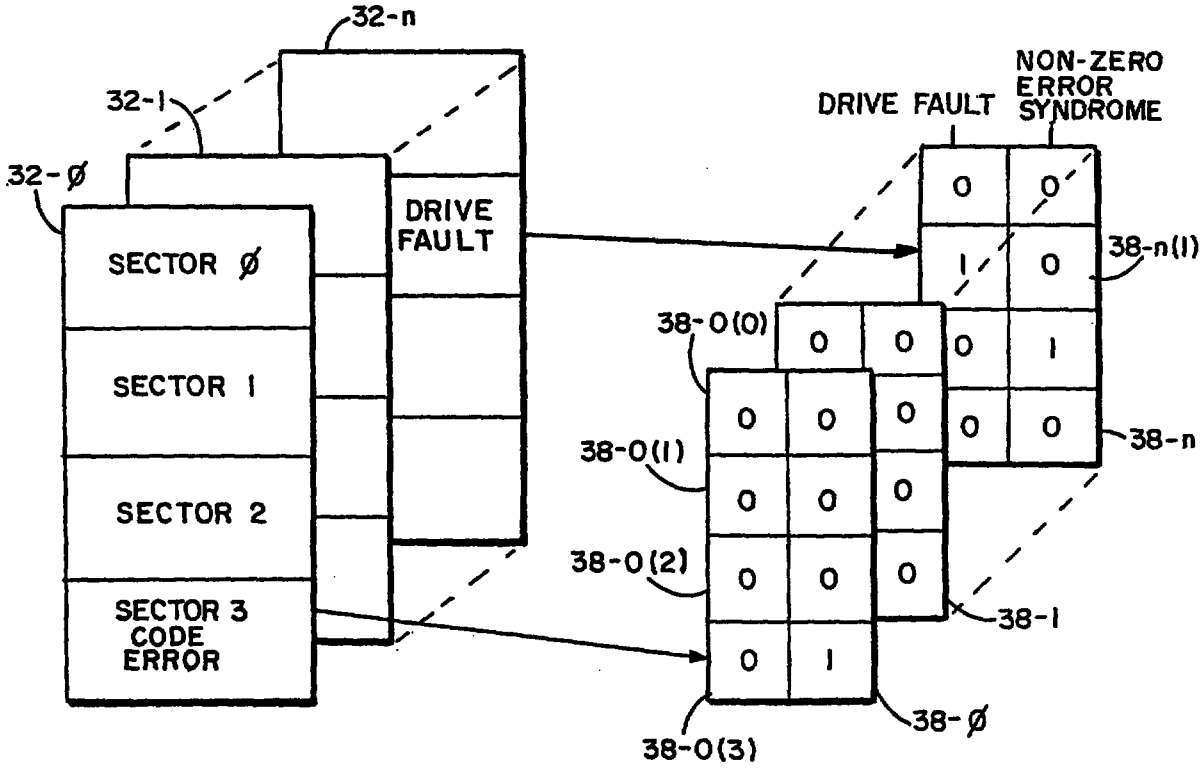
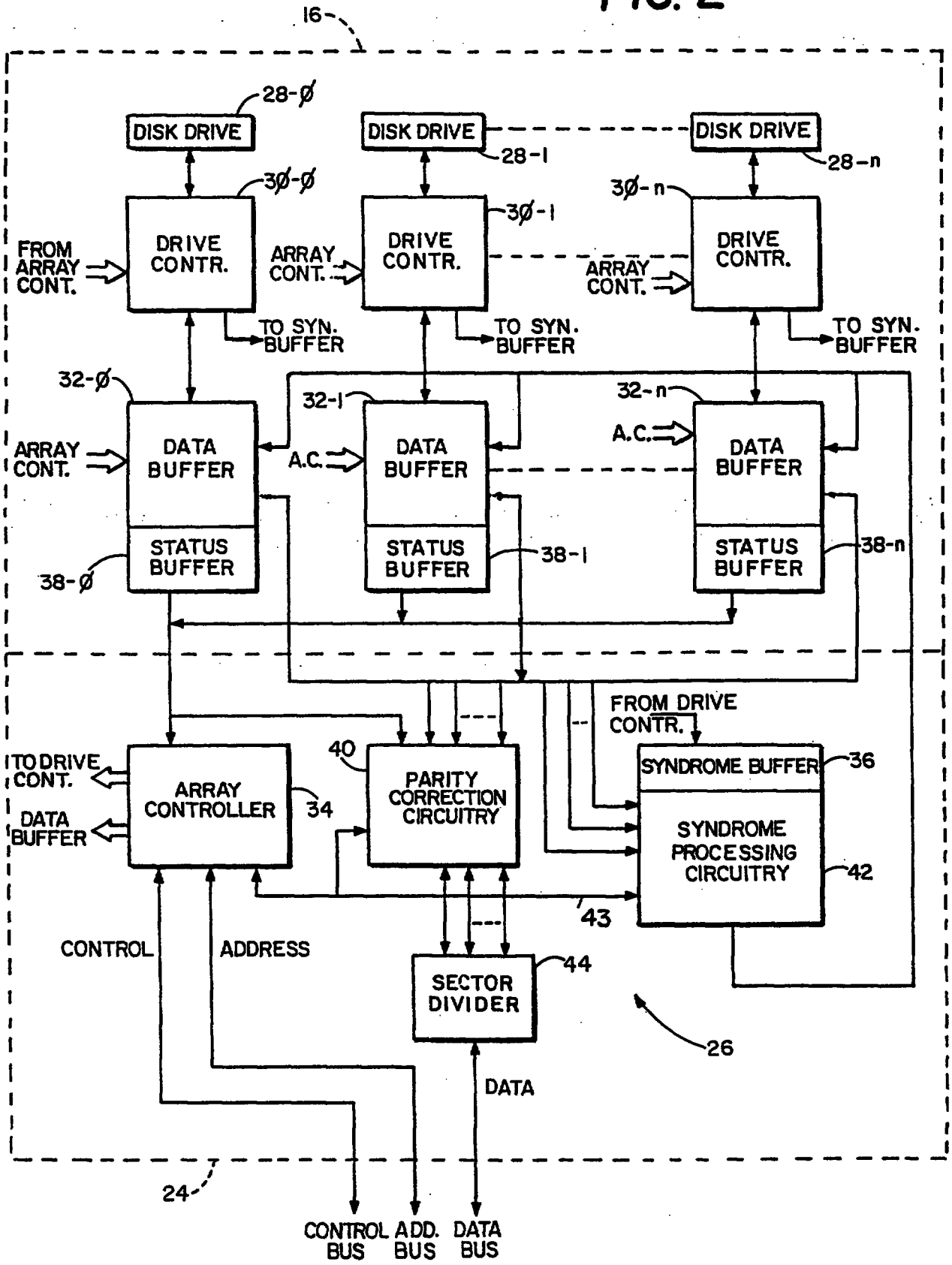


FIG. 3

This Sheet Has Been Replaced

FIG. 2



9603700-7 IBDS

REPLACEMENT SHEETS SG
☑ Description : paged ..1.. to paged ..19..
☑ Claim(s) : paged ..20.. to paged ..23..
☑ Drawing(s) : paged ..1/3.. to paged ..3/3..
☑ Abstract : paged ..24.. to paged ..-..
DATE OF RECEIPT ..18/11/96..

REPLACED SHEETS SG
☑ Description : paged ..1.. to paged ..19..
☑ Claim(s) : paged ..20.. to paged ..23..
☑ Drawing(s) : paged ..1/3.. to paged ..3/3..
☑ Abstract : paged ..24.. to paged ..-..
DATE OF RECEIPT ..

A DATA RECOVERY CHANNEL IN A FAULT TOLERANT DISK DRIVE ARRAY AND A METHOD OF CORRECTING ERRORS THEREIN

This invention relates to data recovery channels in fault tolerant disk drive arrays and methods of correcting errors therein.

A peripheral data storage system is a type of secondary memory accessible by a data processing system, such as a computer. Memory, in a computer, is where programs and work files are stored as digital data. Computer memory can include either, and commonly includes both, moving type memory and non-moving type memory. Non-moving memory is typically directly addressed by the computer's central processing unit. Moving memory, such as magnetic disk drives and magnetic tape, is not directly addressed, and is commonly referred to as secondary memory or peripheral memory.

Moving memory typically has much greater data storage capacity than directly addressed memory and has much longer access times. Moving memory is also typically not volatile. That is, it survives turning the computer off. Non-moving type memory is typically faster and more expensive per unit of memory than moving type memory, and has less capacity. Moving type memories are generally used for long term storage of large programs and substantial bodies of information, such as a data base file, which are not in constant use by the computer, or which are too bulky to provide short term direct access memory capacity for.

The storage media of the moving type memory are physically alterable objects. That is to say, they can be magnetised, grooved, pitted or altered in some detectable fashion to record information. Preferably

This Sheet Has Been Replaced

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

the storage media is at the same time physically resilient, portable, cheap, of large capacity, and resistant to accidental alteration. An example of an analogous medium is a phonograph record where a wavy spiral groove represents an analog information signal. The various species of storage media used in moving type memory for computers include magnetic tape, floppy disks, compact disk-ROM, write-once, read-many optical disks, and, most recently, erasable magneto-optic disks. Each of these storage media exhibit detectable physical changes to the media representing binary data. To read, and where applicable to erase and write data to the media, mechanical apparatus is provided which can be directed to the proper location on the physical media and carry out the desired function.

A magnetic disk drive includes a transducer, a magnetic media disk and associated electronic circuitry to drive or monitor the transducer and transfer the data between the physical medium and the computer to which the drive is connected. The conversion of data from electronic signal to physical feature for storage, extended retention of the data as physical features, and the conversion of physical feature back to electronic signal offers numerous opportunities for error to be introduced to the data record. Transducer and disk present to one another surfaces that are in essentially constant motion. While the inter-action between transducer and media is magnetic, the inter-action between their surfaces is a mechanical one, affected over time by factors such as friction, wear, media flaking and collisions between the transducer and the media. All of these occurrences can be sources of error. In addition, both individual transducers and magnetic media surfaces are subject to failures, such as

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

opening of a transducer electro-magnet winding or imperfections in the magnetic media surface. Mechanical failure can affect the magnetic inter-action of the transducer and the media, and consequently can affect the ability of data recovery circuitry to read data records from the media surface. The faults described above, and others too numerous to discuss in any detail here, can result in loss of data records ranging from one bit to all of the data of the record stemming from failure of the entire drive. The possibility of error in read back, an error or loss of data on the disk, has led to the incorporation of error detection and correction techniques in disk drives.

Data in digital processing systems is typically stored to a disk drive by sectors. Such sectors will include certain redundant information to be used for checking the accuracy of the record and possible correction of the sector upon read back. An example of such redundant information is "Error Checking and Correcting Code", sometimes referred to as "Error Correction Code" or, "ECC".

An ECC for a sector or record will literally comprise data bits supplementing the regular data bits of the record. Where an ECC is used each record conforms to specific rules of construction which permit use of the supplemental bits to detect and, under certain circumstances, correct errors in the record. In actual use, error syndromes are generated from data and the ECC upon read back of the sector. Non-zero error syndromes indicates that error is present. Where error is "random", which for the purpose of this specification means error within the capacity of the ECC to correct, the error syndromes can be used to correct the sector. However, operation on the error syndromes to correct

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

error is relatively time consuming in terms of performance of the disk drive in a computer. Analysis of the error syndromes to determine whether error is correctable, that is whether error is random or massive, can be done in a relatively short period of time.

A second example of redundant data is parity data. Parity data is generated for a logically associated group of binary bits, typically by a modulo addition operation on the group to generate a check digit. An example of a check digit in a binary system is where the digits are logically "ORed", generating a "1" if the number of "1's" in the group is odd or a "0" otherwise. Parity, absent information indicating location of an error, is usable only to identify the existence of error. However, given independent identification of location of the error in a group, the use of parity permits rapid correction of the error. Only one bit of error per parity bit can be corrected.

A computer can use more than one disk drive for the storage of data. Alternatively, a plurality of disk drives can be organised to operate together and thereby appear to the computer as one peripheral storage unit. The term disk drive array is used in this specification to indicate a group of disk drives operating in a parallel, synchronous fashion allowing transfer of data bits of record in parallel to the individual drives of the array and appearing to the computer as a single, data storage peripheral device. An interface operating between the disk drive array and the data processing system transfers data to and from storage in parallel to increase data transmission band width.

Such parallel, synchronised disk drives have characteristics offering opportunity for improvement in redundancy for stored data. Simplistically, data on

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

one disk drive can be mirrored on another. In US-A-4,775,978 there is disclosed a data error correction system for a mass data storage system comprising an array of synchronised disk data storage units. The data storage system receives data blocks from a host data processing system for storage. A data block divider stripes the data among a plurality of the disk data storage units as columns. A data word is divided among the plurality of disk drives one bit to each drive. Data stored to the same address on each disk are supplemental by a column of parity data stored to yet another data storage unit. Each data column, assigned to a particular disk drive, is supplemented with an error correction code. Upon read back, two levels of data error correction are provided. Data columns are read and stored to buffers. Error syndromes are generated upon read back and used to correct random error. When a data storage unit fails, or where error correction code for a column is inadequate to correct errors in a column, the column (i.e. data storage unit) where the error occurs is identified and the parity data in combination with the data of the other columns is used to re-construct data in the missing column. The system provides a highly fault tolerant data storage system.

The present invention seeks to provide an improved data recovery channel applicable to a data error correction system such as that disclosed in US-A-4,775,978.

According to one aspect of the present invention there is provided a data recovery channel in a fault tolerant disk drive array, the disk drive array including a plurality of disk drives and an input interface for receiving data records from a data processing system, the input interface organising each

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

data record received into data sectors for storage on the disk drives, identifying a plurality of sectors as a related group and generating a parity sector for the group, where each data and parity sector includes an appended error correction code, the group of data and parity sectors being stored by striping across the available disk drives, the data recovery channel being characterised by comprising: a disk drive controller associated with each disk drive including means for recovering sectors from the disk drive, means for determining error syndromes for each sector as recovered, and means for detecting a disk drive fault; a data buffer for retaining the data and parity recovered with each sector; an error status buffer for retaining indication of non-zero error syndromes generated upon recovery of a sector or occurrence of a disk fault associated with attempted recovery of a sector; an error syndrome buffer for retaining any non-zero error syndromes generated during recovery of data and parity sectors; syndrome processing means connected to the error syndrome buffer for determining the correctability of the data of a sector, and for correcting the data in the data buffers, if possible; parity correction means for receiving associated groups of data and parity information from the data buffers and correcting the data from up to one sector; and an array controller for operating upon the error status buffer and indications of correctability of data from the syndrome processing means to direct the correction method for each sector of a group.

Preferably the array controller further comprises: means for monitoring the error status buffer; means for transferring the data and parity for a group of sectors to the parity correction means when the number of errors

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

indicated for a group of sectors is or is reduced to one or less; and means for re-setting the status indications for a sector in the error status buffer where analysis of the error syndromes for the sector shows that the data from the sector can be corrected. Each status buffer may have a memory location for each sector recovered of a data record.

The array controller may further comprise means for initiating a disk read re-try when more than one disk fault is indicated for a group, and means for indicating a read failure when a disk read re-try fails.

The array controller may further comprise means for initiating a disk read re-try when the number of sectors having a fault cannot be reduced to one in number; and means for indicating a read failure when a disk read re-try fails.

The data recovery channel may be in combination with a disk drive array, the disk drive array comprises synchronised, fault independent disk drive units.

According to another aspect of the present invention there is provided a method of correcting errors appearing in data recovered from a plurality of disk drives in an array, wherein the data is organised on the disk by sectors and groups of sectors form related rows, each row having a parity data sector and each sector including an error correction code, the method being characterised by comprising the steps of: (a) generating error syndromes for each sector upon recovery of data from the disk drives; (b) storing non-zero error syndromes for use in correction of data; (c) generating and storing error status indications for each sector recovered, the indications including a no error condition, a non-zero error syndrome condition and a drive fault condition; (d) initiating a

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

correction protocol including the steps of:
(1) checking the error status for each sector for a row and continuing to the next row of sectors if no errors are present, (2) if error is indicated for one sector, correcting the sector using parity correction and returning to step (1) for the next row of sectors, (3) if error is indicated for more than one sector of the row, using the error syndromes to locate sectors correctable with the error syndromes and correcting the data for those sectors until one sector having error is left, and then returning to step (2), (4) where the number of errors in a row cannot be reduced to one, initiating a re-try read of a data storage unit from which a sector having not correctable error was recovered and returning to step (1), if successful, and (5) indicating a read failure and aborting the data recovery attempt if step (4) is not successful.

The method may include the additional step after step (d) (4) of initiating a recovery of the disk drive and returning to step (1) if successful.

The invention is illustrated, merely by way of example, in the accompanying drawings, in which:-

Figure 1 is a block diagram of a data processing system;

Figure 2 is a block diagram of a data recovery channel according to the present invention; and

Figure 3 illustrates organisation of data buffers of the data recovery channel of Figure 2.

Figure 1 shows a data processing system 10 with which a data recovery channel according to the present invention can be advantageously used. The description of the architecture of the data processing system 10 is intended only to give an environment for explanation of the present invention and is not intended as a

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

description of a particular computer architecture with which the present invention is used. The data processing system 10 is entirely conventional and includes a central processing unit 12, a direct access memory 14 and a data storage peripheral 16. The central processing unit 12 stores data and programming steps to, and recovers data and programming steps over a data bus 18 from the memory 14 and the data storage peripheral 16. The data storage peripheral 16 can transfer data and programming steps over the data bus 18 directly to the memory 14 or to the central processing unit 12.

The central processing unit 12 exercises control over the memory 14 and the data storage peripheral 16 over a control bus 20 and an address bus 22, thereby directing the timing of the transfer of data and locations to which the data is delivered or from which it is called. The data storage peripheral 16 may be a tape drive, a disk drive, or some other form of indirectly addressed, mass data storage structure. Communication between the data storage peripheral 16 and the rest of the data processing system 10 is through an input/output interface 24. Where the data storage peripheral 16 is in an array of disk drives, a data recovery channel according to the present invention can be advantageously incorporated in the interface 24.

Figure 2 illustrates the data storage peripheral 16, the interface 24 and a data recovery channel 26 according to the present invention. The data storage peripheral 16 has n+1 disk drives 28-0 to 28-n. Each disk drive is under the direct control of one of n+1 drive controllers 30-0 to 30-n. The drive controllers 30-0 to 30n, among other functions, provide a passage for data between the disk drives 28-0 to 28-n and data

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

buffers 32-0 to 32-n, respectively.

Data records on the disk drives 28-0 to 28-n are organised by sectors. A sector includes preamble information, data which is derived from the record applied to the data storage peripheral 16 for storage and an error detection and correction code generated from the data. The generation of such sectors and the operation of writing the sectors to the disk drives 28-0 to 28-n forms no part of the present invention. The sectors stored on the disk drives 28-0 to 28-(n-1) contain the data received over the data bus 18 for storage and are referred to here as data sectors. The sectors stored on the disk drive 28-n contain parity data related to the data stored in the first n drives and are referred to here as parity sectors. Parity sectors are otherwise like data sectors in that they include error correction codes.

Read back of data records from the disk drives 28-0 to 28-n is done simultaneously and in parallel by the drive controllers 30-0 to 30-n upon receipt of a read command from an array controller 34 in the interface 24. Groups of sectors, designated here as rows, are inter-related. A row includes a plurality of data sectors, one from each disk, and a parity sector. The parity data in the parity sector is generated from logically associated bits between the data sectors. For example, the first parity bit in the parity sector may be parity for the first data bit from each data sector, the second parity bit in the parity sector is calculated over the second data bits of the data sectors and so on. The sectors of a row are stored to the same logical addresses in the data buffers 32-0 to 32-n. During execution of a read back of data records, the drive controllers 30-0 to 30-n operate on the recovered data

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced



and ECC to generate error syndromes for each sector. The error syndromes, if non-zero, are loaded into known locations in an error syndrome buffer 36 for possible later use. A non-zero error syndrome set for a sector indicates presence of a code error in the sector, which may or may not be correctable using the error syndromes. The drive controllers 30-0 to 30-n also monitor the disk drives 28-0 to 28-n for indications of drive failure or fault.

A data record includes a plurality of rows of sectors. During read back of a record from the disk drives 28-0 to 28-n by the drive controllers 30-0 to 30-n, the drive controllers generate error status indications for each sector read and store the error status indications in status buffers 38-0 to 38-n. An error status indication of 0,0 indicates a sector free of detected error; an error status indication of 0,1 indicates a sector for which non-zero error syndromes were generated upon read back; an error status record of 1,0 indicates occurrence of a disk fault. An error status record of 1,1 is preferably not used. In summary, the first digit of the record is used to indicate a disk fault and the second digit is used to indicate a code fault in a sector.

Upon completion of reading of the sectors from the disk drives 28-0 to 28-n, the data from each data and parity sector will be stored to a pre-determined location in the data buffers 32-0 to 32-n, an error status indication for each data and parity sector will be stored to a logically associated address in the status buffers 38-0 to 38-n, and the non-zero error syndromes for each sector having code errors will be stored in the syndrome buffer 36.

The array controller 34 is a programmed micro-

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced



computer and controls the error correction processing in the data recovery channel 26 by monitoring the status buffers 38-0 to 38-n and communicating with syndrome processing circuitry 42 over a bus 43. The data recovery channel 26 includes parity correction circuitry 40, the error syndrome buffer 36, the error syndrome processing circuitry 42 and the data buffers 32-0 to 32-n. The error syndrome processing circuitry 42 may be and preferably is combined with the array controller 34 by programming the array controller to perform the functions of the error syndrome processing circuitry. Illustration of the array controller 34 and the syndrome processing circuitry 42 as separate components is an aid to understanding and a feasible hardware configuration.

During a read operation, the array controller 34 monitors the error status records stored to the status buffers 38-0 to 38-n. A read operation is allowed to continue as long as no more than one drive fault error occurs in any given row of the recovered record, i.e. the read operation continues as long as the error status 1,0 does not repeat at the same logical address across the status buffers 38-0 to 38-n. Where two drive faults are indicated for a row, the read is aborted and extra-ordinary recovery techniques may be employed, such as a re-read attempt or an attempted recovery of the disk drive. If all such techniques fail the read operation is aborted and a drive array failure condition is indicated.

Absent indication of simultaneous drive faults for two or more of the disk drives 28-0 to 28-n, the read operation continues until completion of recovery of an entire record. The array controller 34 then directs error correction for the recovered record, row by row. As previously noted, both operation on parity data and

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

on the error syndromes is available. The array controller 34 interrogates the error status words for a row and determines the number of sectors for which drive fault and ECC error is present. Where error of either type is indicated as occurring in none or only one sector for the row, all of the sectors of the row are transferred from the data buffers 32-0 to 32-n to the parity correction circuitry 40 for correction, if required, through a sector divider 44 for re-construction of the original words of the record and transmission out to the data bus 18. Parity correction provides superior speed over an error syndrome operation for correction of a sector, therefore it is not even determined whether a sector having error is correctable using the error syndromes. The array controller 34 indicates to the parity correction circuitry 40 the sector having error. Correction involves the re-generation of the data contained in the sector having the error from the remaining data and parity infor-mation.

Where positive error status indications exist for two or more sectors of a row, utilisation of the error syndromes for correction of one or more sectors is attempted. The array controller 34 first locates those sectors having code errors, indicated in the status buffers 38-0 to 38-n by 0,1 records for each sector. The array controller 34 then directs the syndrome processing circuitry 42 to analyse the error syndromes for a given sector, typically beginning with the error syndromes for the first sector for which an error status of 0,1 exists. By "first" will typically be meant the sector associated with the lowest numbered data buffer (e.g. data buffer 32-1 before data buffer 32-3).

The first operation performed by the syndrome

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced



processing circuitry 42 is a determination whether the error syndromes for a sector can be used to correct the data. If the error is correctable the error syndrome processing circuitry 42 corrects the data in the appropriate data buffer 32-0 to 32-n. The array controller 34 re-sets the status buffer of the associated error status word to indicate no error in the sector. The array controller 34 will continue correction of sectors using error syndromes, if possible, until only one defective sector remains for a row. When only one error containing sector remains, regardless of whether the error can be corrected using error syndromes or not, the entire row is transferred to the parity correction circuitry 40 for parity correction of the sector containing error. The corrected row is then transmitted to the sector divider 44 for re-construction of the original words of the data record and on to the data bus 18. The array controller 34 then advances the correction algorithm to the next row. Correction continues until all sectors of the file are corrected and transmitted on to the data bus 18.

Should two or more sectors of a row have massive errors, i.e. error exceeding the capacity of the error syndromes, correction by the array controller 34 is not attempted. Instead, recovery is aborted and the data from at least one of the affected disk drives 28-0 to 28-n is re-read. Numerous conventional data recovery methods may be attempted, before the recovery attempt is aborted.

The original data words of a record are re-assembled before transmission on the data bus 18. The sector divider 44 is a bi-directional element which receives data words over the data bus 18 and routes each subsequent bit or byte of the word to a succeeding disk

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

drive, wrapping around to the first disk drive after assignment of n bits (or bytes). The sector divider 44 also operates to re-assemble the original words when a record is transmitted out of the parity correction circuitry 40.

Figure 3 illustrates the storage of an exemplary data record among the memory units comprising the data buffers 32-0 to 32-n after recovery of the record from the disk drives 28-0 to 28-n. The exemplary record includes 4 x n data sectors, stored four (numbered 0 to 3 for each buffer) to a data buffer, and four parity sectors (numbered 0 to 3) stored in a data buffer 32-0 to 32-n. Like numbered data and parity sectors belong to a row, e.g. sectors 1 of the data bufers 32-0 to 32-n are a row, sectors 3 of the data buffers 32-0 to 32-n are another row. One data bit from each data sector in a row relates to a particular parity bit from the parity sector for the row. In a typical arrangement only the original data of the sector having bits numbered 0 to 255 are in a data buffer after down loading a sector from its disk drive.

The status buffers 38-0 to 28-n consist of storage locations corresponding one for one with each location for sectors to be stored in data buffers 32-0 to 32-n. As recovery of data and parity sectors is attempted from the disk drives 28-0 to 28-n and the data or parity transferred to locations in the data buffers 32-0 to 32-n, error status indication words for the sectors are generated and stored in the corresponding memory locations of the status buffers 38-0 to 38-n. The first bit location of the error status indication words are the indicators of drive fault, the second bit location of the words are indication of non-zero error syndromes. Initiation of data recovery and transfer in the data

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

buffers results in all locations of the error status buffers 38-0 to 38-n being re-set. An error status of 0,0, or re-set value, is indication of no error in the corresponding data or parity sector.

Sector 3 stored in the data buffer 32-0 illustrates indication of non-zero error syndromes for a sector. The status buffer 38-0 has four error status locations 38-0(0) to 38-0(3). The error condition of sector 3 of data buffer 32-0 is reflected by the second bit of location 38-0(3) which is set at 1. Similarly, sector 1 of buffer 32-n shows a drive fault error. Location 28-n(1) shows a first bit location as 1 indicating the error condition of the sector.

The present data recovery channel corrects the data of a sector recovered from a disk drive, wherein sectors include error correction codes, groups of sectors are related by the existence of parity data for the group. Two methods of correction for sectors are used, one being parity, the other method using error syndromes. Correction employs the following priority if no more than one sector is lost per group due to drive failure:

(1) Correction of a sector by parity within a parity related group;

(2) Correction of sectors by error correction syndromes until the number of sectors evidencing error in a parity related group is reduced to one, then correcting that sector with parity;

(3) If two or more sectors evidence drive fault and a non-correctable error condition in a parity related group, re-trying the read for at least one disk drive.

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

(4) If the re-try read fails, attempting a "recover" of at least one disk drive (if the disk controller admits the command); .and

(5) Signalling the data storage peripheral "FAIL" condition if all the above fail.

As will be appreciated, the data recovery channel preferentially operates on parity data for error correction over utilisation of error syndromes for error correction, though both remain available. In summary, absent indication of errors from more than one sector for a row of sectors, correction of error in the defective sector is made by use of parity information. Where error is indicated for more than one sector in a row of sectors, correction using the error syndromes is attempted, sector by sector, until the number of sectors in the row having error is reduced to one. Parity is used to correct the remaining defective sector.

Error syndromes are generated for each sector upon recovery of the sector from a disk drive. A non-zero error syndrome set for a sector indicates the possible presence of error and results in generation of a signal indicating such occurrence. A failure to recover a sector results in generation of a signal indicating a drive fault. A status buffer saves indications of fault from a failing drive, the occurrence of non-zero error syndromes or an error free condition for each recovered sector. An error syndrome buffer retains the error syndromes if non-zero.

An array controller monitors the status buffer and continues the read operation as long as no more than one drive fault is indicated with respect to a row of data

This is the replacement sheet
Date: 18 NOV 1996

This Sheet Has Been Replaced

sectors. Upon completion of reading of all the rows of a file, the data recovery channel has access to the sector status information, recovered data sectors and to the error syndromes. Error correction of the sectors, re-construction of the original data words and transmission out of the original record is then attempted.

The array controller interrogates the sector status information and either executes, or directs execution, of error correction routines to restore the original data. Utilisation of the error syndromes for data re-construction is deferred in favour of parity re-construction if errors occur in only one sector for each row of sectors.

Error status for each sector is retained in an error status buffer. Where the status buffer indicates no errors for a row, the original data words are re-constructed and the data transmitted to the utilising computer. Where error of either type is indicated for one sector of a row, parity is used to re-construct the affected sector. Upon indication of error for more than one sector of a row, the array controller attempts restoration of the defective sectors of the row utilising error syndrome data. The array controller or syndrome processing circuitry which may be a program executed by the array controller, operates on the error syndromes for a first defective sector from the row to determine if the sector is correctable. If the sector is correctable, correction is carried out, the restored data is stored to the data buffer and the error status buffer for the sector is re-set. Once the number of sectors in a row with error of either type is reduced to one, parity correction is done and the entire row is transferred out to the utilising computer. Where the

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced



number of sectors in a row with error cannot be reduced below two using error syndromes, the read operation is interrupted and the more-extensive data recovery methods, directed at the affected disk drives are attempted, e.g. a re-read of disk drives.

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

C L A I M S

1. A data recovery channel in a fault tolerant disk drive array (16), the disk drive array including a plurality of disk drives (28-0 to 28-n) and an input interface (24) for receiving data records from a data processing system (10), the input interface (24) organising each data record received into data sectors for storage on the disk drives, identifying a plurality of sectors as a related group and generating a parity sector for the group, where each data and parity sector includes an appended error correction code, the group of data and parity sectors being stored by striping across the available disk drives, the data recovery channel being characterised by comprising: a disk drive controller (30-0 to 30-n) associated with each disk drive (28-0 to 28-n) including means for recovering sectors from the disk drive, means for determining error syndromes for each sector as recovered, and means for detecting a disk drive fault; a data buffer (32-0 to 32-n) for retaining the data and parity recovered with each sector; an error status buffer (38-0 to 38-n) for retaining indication of non-zero error syndromes generated upon recovery of a sector or occurrence of a disk fault associated with attempted recovery of a sector; an error syndrome buffer (36) for retaining any non-zero error syndromes generated during recovery of data and parity sectors; syndrome processing means (42) connected to the error syndrome buffer for determining the correctability of the data of a sector, and for correcting the data in the data buffers, if possible; parity correction means (40) for receiving associated groups of data and parity information from the data buffers and correcting the data from up to one sector;

* C L M S *

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

and an array controller (34) for operating upon the error status buffer and indications of correctability of data from the syndrome processing means to direct the correction method for each sector of a group.

2. A data recovery channel as claimed in claim 1, characterised in that the array controller (34) further comprises: means for monitoring the error status buffer (38-0 to 38-n); means for transferring the data and parity for a group of sectors to the parity correction means (40) when the number of errors indicated for a group of sectors is or is reduced to one or less; and means for re-setting the status indications for a sector in the error status buffer (38-0 to 38-n) where analysis of the error syndromes for the sector shows that the data from the sector can be corrected.

3. A data recovery channel as claimed in claim 2 characterised in that each status buffer (38-0 to 38-n) has a memory location for each sector recovered of a data record.

4. A data recovery channel as claimed in claim 3 characterised in that the array controller 34 further comprises: means for initiating a disk read re-try when more than one disk fault is indicated for a group, and means for indicating a read failure when a disk read re-try fails.

5. A data recovery channel as claimed in claim 3 or 4 characterised in that the array controller further comprises: means for initiating a disk read re-try when the number of sectors having a fault cannot be reduced to one in number; and means for indicating a read failure when a disk read re-try fails.

6. A data recovery channel as claimed in any preceding claim in combination with a disk drive array characterised in that the disk drive array comprises

* C L M S *

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

synchronised, fault independent disk drive units.

7. A method of correcting errors appearing in data recovered from a plurality of disk drives (28-0 to 28-n) in an array, wherein the data is organised on the disk by sectors and groups of sectors form related rows, each row having a parity data sector and each sector including an error correction code, the method being characterised by comprising the steps of: (a) generating error syndromes for each sector upon recovery of data from the disk drives; (b) storing non-zero error syndromes for use in correction of data; (c) generating and storing error status indications for each sector recovered, the indications including a no error condition, a non-zero error syndrome condition and a drive fault condition; (d) initiating a correction protocol including the steps of: (1) checking the error status for each sector for a row and continuing to the next row of sectors if no errors are present, (2) if error is indicated for one sector, correcting the sector using parity correction and returning to step (1) for the next row of sectors, (3) if error is indicated for more than one sector of the row, using the error syndromes to locate sectors correctable with the error syndromes and correcting the data for those sectors until one sector having error is left, and then returning to step (2), (4) where the number of errors in a row cannot be reduced to one, initiating a re-try read of a data storage unit from which a sector having not correctable error was recovered and returning to step (1), if successful, and (5) indicating a read failure and aborting the data recovery attempt if step (4) is not successful.

* C L M S *

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

8. A method as claimed in claim 7 characterised by including an additional step after step (d) (4) of initiating a recovery of the disk drive and returning to step (1) if successful.

* C L M S *

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced



ABSTRACT

A Data Recovery Channel in a Fault Tolerant Disk Drive Array and a Method of Correcting Errors Therein

An input interface for an array of a plurality of associated disc drives, which stores data in related sectors and apportions a parity sector for related data sectors, where each data and parity sector include an appended error correction code. A data recovery channel comprises a controller (30-0 to 30-n) for each disc drive (28-0 to 28-n); a data buffer (32-0 to 32-n) retaining data and parity for a sector; an error status buffer (38-0 to 38-n) retaining indications of non-zero error syndromes or disk fault associated with data recovery; an error syndrome buffer (36) retaining any non zero error syndromes generated during recovery of data and parity sectors; syndrome processing circuitry (42) for determining correctability of and correcting data; parity correction circuitry (40) for correcting parity data in a sector; and an array controller (34) operating upon the error status buffer and indications of correctability of data to direct the correction method for each related sector.

Figure 2

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

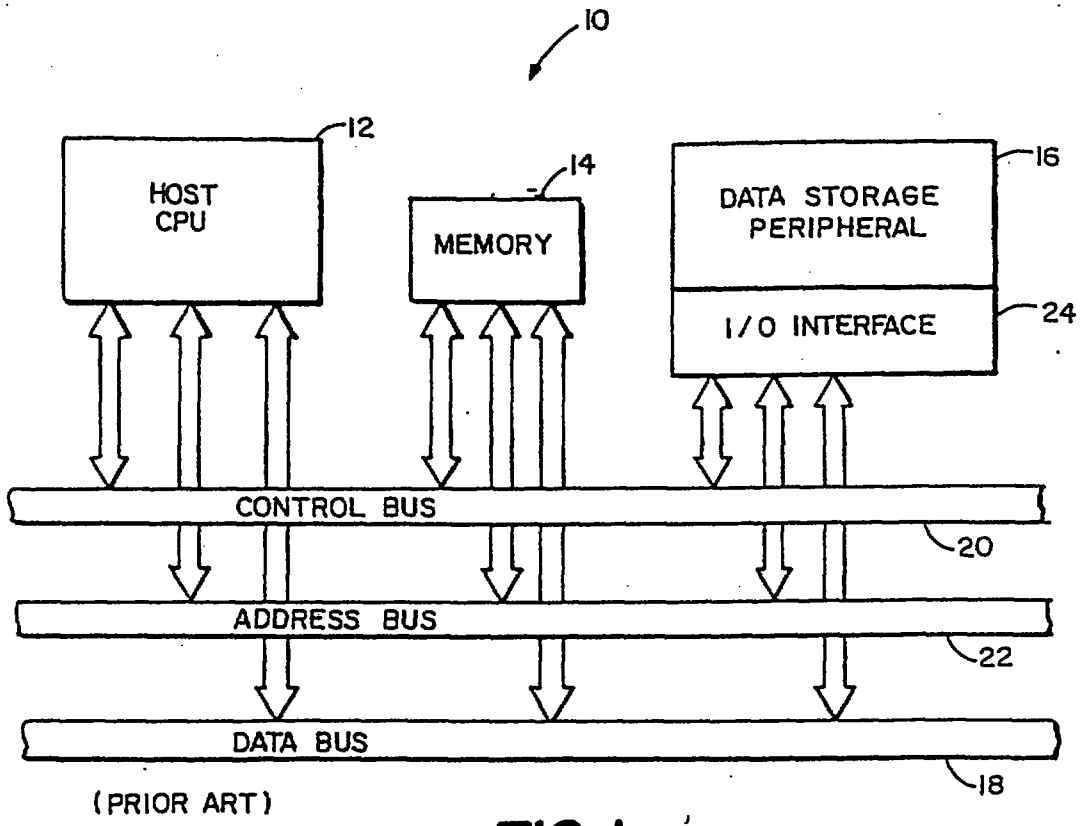


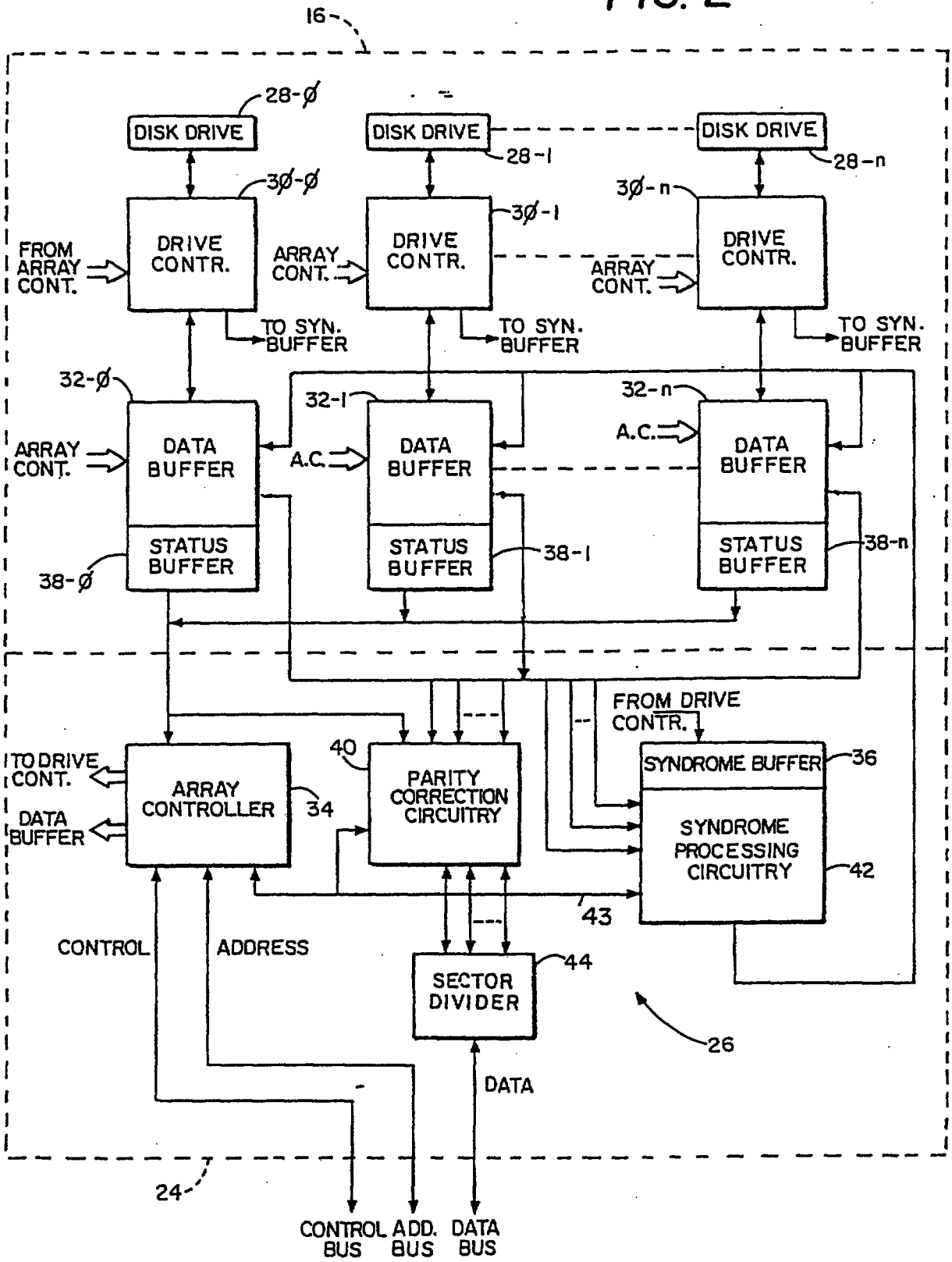
FIG. 1

This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

FIG. 2



This is the replacement sheet

Date: 18 NOV 1996

This Sheet Has Been Replaced

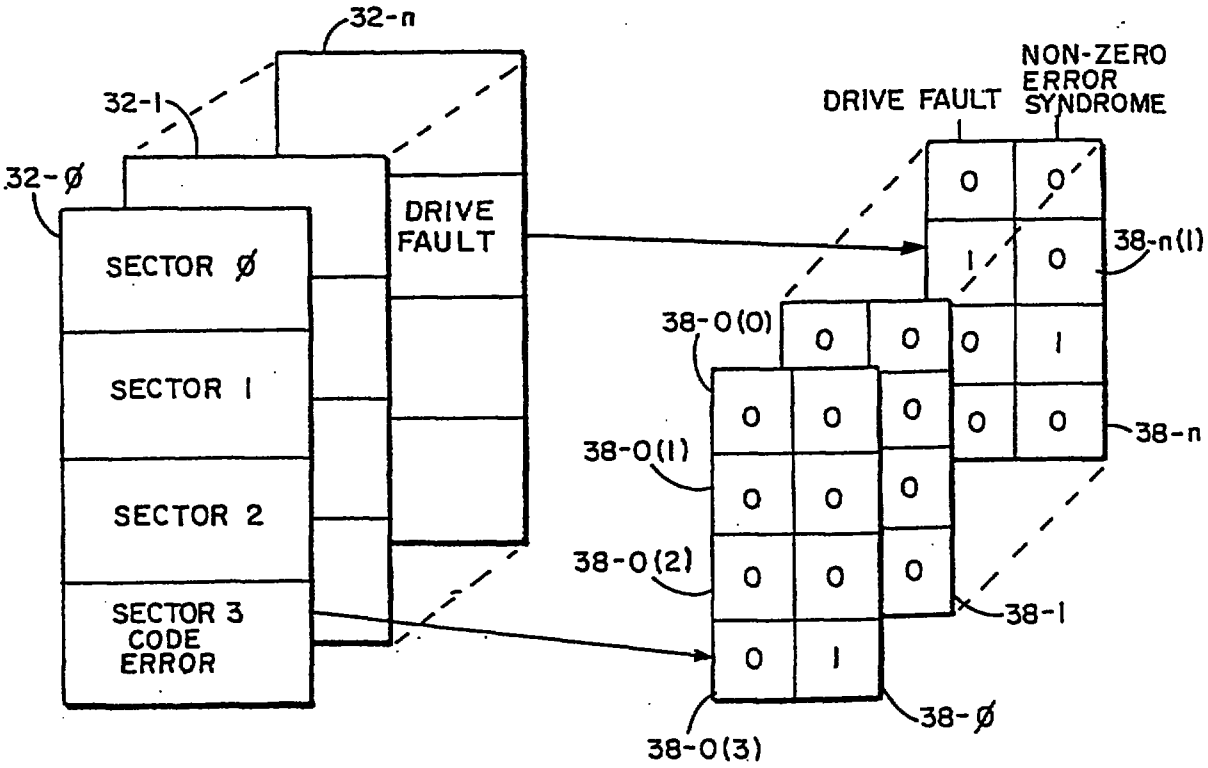


FIG. 3

This is the replacement sheet
Date: 18 NOV 1996

9603700-7 IP03

REPLACEMENT SHEETS SG
☑ Description : paged 1 to paged 20
☑ Claim(s) : paged 21 to paged 23
☑ Drawing(s) : paged 1/3 to paged 3/3
☑ Abstract : paged 24 to paged ---

DATE OF RECEIPT 7/4/97

A DATA RECOVERY CHANNEL IN A FAULT TOLERANT DISK DRIVE ARRAY AND A METHOD OF CORRECTING ERRORS THEREIN

This invention relates to data recovery channels in fault tolerant disk drive arrays and methods of correcting errors therein.

A peripheral data storage system is a type of secondary memory accessible by a data processing system, such as a computer. Memory, in a computer, is where programs and work files are stored as digital data. Computer memory can include either, and commonly includes both, moving type memory and non-moving type memory. Non-moving memory is typically directly addressed by the computer's central processing unit. Moving memory, such as magnetic disk drives and magnetic tape, is not directly addressed, and is commonly referred to as secondary memory or peripheral memory.

Moving memory typically has much greater data storage capacity than directly addressed memory and has much longer access times. Moving memory is also typically not volatile. That is, it survives turning the computer off. Non-moving type memory is typically faster and more expensive per unit of memory than moving type memory, and has less capacity. Moving type memories are generally used for long term storage of large programs and substantial bodies of information, such as a data base file, which are not in constant use by the computer, or which are too bulky to provide short term direct access memory capacity for.

The storage media of the moving type memory are physically alterable objects. That is to say, they can be magnetised, grooved, pitted or altered in some detectable fashion to record information. Preferably

This is the replacement sheet

Date: 07 APR 1999

the storage media is at the same time physically resilient, portable, cheap, of large capacity, and resistant to accidental alteration. An example of an analogous medium is a phonograph record where a wavy spiral groove represents an analog information signal. The various species of storage media used in moving type memory for computers include magnetic tape, floppy disks, compact disk-ROM, write-once, read-many optical disks, and, most recently, erasable magneto-optic disks. Each of these storage media exhibit detectable physical changes to the media representing binary data. To read, and where applicable to erase and write data to the media, mechanical apparatus is provided which can be directed to the proper location on the physical media and carry out the desired function.

A magnetic disk drive includes a transducer, a magnetic media disk and associated electronic circuitry to drive or monitor the transducer and transfer the data between the physical medium and the computer to which the drive is connected. The conversion of data from electronic signal to physical feature for storage, extended retention of the data as physical features, and the conversion of physical feature back to electronic signal offers numerous opportunities for error to be introduced to the data record. Transducer and disk present to one another surfaces that are in essentially constant motion. While the inter-action between transducer and media is magnetic, the inter-action between their surfaces is a mechanical one, affected over time by factors such as friction, wear, media flaking and collisions between the transducer and the media. All of these occurrences can be sources of error. In addition, both individual transducers and magnetic media surfaces are subject to failures, such as

This is the replacement sheet

Date: 07 APR 1999

opening of a transducer electro-magnet winding or imperfections in the magnetic media surface. Mechanical failure can affect the magnetic inter-action of the transducer and the media, and consequently can affect the ability of data recovery circuitry to read data records from the media surface. The faults described above, and others too numerous to discuss in any detail here, can result in loss of data records ranging from one bit to all of the data of the record stemming from failure of the entire drive. The possibility of error in read back, an error or loss of data on the disk, has led to the incorporation of error detection and correction techniques in disk drives.

Data in digital processing systems is typically stored to a disk drive by sectors. Such sectors will include certain redundant information to be used for checking the accuracy of the record and possible correction of the sector upon read back. An example of such redundant information is "Error Checking and Correcting Code", sometimes referred to as "Error Correction Code" or, "ECC".

An ECC for a sector or record will literally comprise data bits supplementing the regular data bits of the record. Where an ECC is used each record conforms to specific rules of construction which permit use of the supplemental bits to detect and, under certain circumstances, correct errors in the record. In actual use, error syndromes are generated from data and the ECC upon read back of the sector. Non-zero error syndromes indicates that error is present. Where error is "random", which for the purpose of this specification means error within the capacity of the ECC to correct, the error syndromes can be used to correct the sector. However, operation on the error syndromes to correct

This is the replacement sheet

Date: 07 APR 1999

error is relatively time consuming in terms of performance of the disk drive in a computer. Analysis of the error syndromes to determine whether error is correctable, that is whether error is random or massive, can be done in a relatively short period of time.

A second example of redundant data is parity data. Parity data is generated for a logically associated group of binary bits, typically by a modulo addition operation on the group to generate a check digit. An example of a check digit in a binary system is where the digits are logically "ORed", generating a "1" if the number of "1's" in the group is odd or a "0" otherwise. Parity, absent information indicating location of an error, is usable only to identify the existence of error. However, given independent identification of location of the error in a group, the use of parity permits rapid correction of the error. Only one bit of error per parity bit can be corrected.

A computer can use more than one disk drive for the storage of data. Alternatively, a plurality of disk drives can be organised to operate together and thereby appear to the computer as one peripheral storage unit. The term disk drive array is used in this specification to indicate a group of disk drives operating in a parallel, synchronous fashion allowing transfer of data bits of record in parallel to the individual drives of the array and appearing to the computer as a single, data storage peripheral device. An interface operating between the disk drive array and the data processing system transfers data to and from storage in parallel to increase data transmission band width.

Such parallel, synchronised disk drives have characteristics offering opportunity for improvement in redundancy for stored data. Simplistically, data on

This is the replacement sheet

Date: 07 APR 1999

one disk drive can be mirrored on another. In US-A-4,775,978 corresponding to EP-A-0 274817 there is disclosed a data error correction system for a mass data storage system comprising an array of synchronised disk data storage units. The data storage system receives data blocks from a host data processing system for storage. A data block divider stripes the data among a plurality of the disk data storage units as columns. A data word is divided among the plurality of disk drives one bit to each drive. Data stored to the same address on each disk are supplemental by a column of parity data stored to yet another data storage unit. Each data column, assigned to a particular disk drive, is supplemented with an error correction code. Upon read back, two levels of data error correction are provided. Data columns are read and stored to buffers. Error syndromes are generated upon read back and used to correct random error. When a data storage unit fails, or where error correction code for a column is inadequate to correct errors in a column, the column (i.e. data storage unit) where the error occurs is identified and the parity data in combination with the data of the other columns is used to re-construct data in the missing column. The system provides a highly fault tolerant data storage system.

The present invention seeks to provide an improved data recovery channel applicable to a data error correction system such as that disclosed in US-A-4,775,978.

According to one aspect of the present invention there is provided a data recovery channel in a fault tolerant disk drive array, the disk drive array including a plurality of disk drives and an input interface for receiving data records from a data processing system, the input interface organising each

This is the replacement sheet

Date: 07 APR 1999

data record received into data sectors for storage on the disk drives, identifying a plurality of sectors as a related group and generating a parity sector for the group, where each data and parity sector includes an appended error correction code, the group of data and parity sectors being stored by striping across the available disk drives, the data recovery channel being characterised by comprising: a disk drive controller associated with each disk drive including means for recovering sectors from the disk drive, means for determining error syndromes for each sector as recovered, and means for detecting a disk drive fault; a data buffer for retaining the data and parity recovered with each sector; an error status buffer for retaining indication of non-zero error syndromes generated upon recovery of a sector or occurrence of a disk fault associated with attempted recovery of a sector; an error syndrome buffer for retaining any non-zero error syndromes generated during recovery of data and parity sectors; syndrome processing means connected to the error syndrome buffer for determining the correctability of the data of a sector, and for correcting the data in the data buffers, if possible; parity correction means for receiving associated groups of data and parity information from the data buffers and correcting the data from up to one sector; and an array controller for operating upon the error status buffer and indications of correctability of data from the syndrome processing means to direct for each sector of a group the correction method claimed in the independent method claim.

Preferably the array controller further comprises: means for monitoring the error status buffer; means for transferring the data and parity for a group of sectors to the parity correction means when the number of errors

This is the replacement sheet

Date: 07 APR 1999

indicated for a group of sectors is or is reduced to one or less; and means for re-setting the status indications for a sector in the error status buffer where analysis of the error syndromes for the sector shows that the data from the sector can be corrected. Each status buffer may have a memory location for each sector recovered of a data record.

The array controller may further comprise means for initiating a disk read re-try when more than one disk fault is indicated for a group, and means for indicating a read failure when a disk read re-try fails.

The array controller may further comprise means for initiating a disk read re-try when the number of sectors having a fault cannot be reduced to one in number; and means for indicating a read failure when a disk read re-try fails.

The data recovery channel may be in combination with a disk drive array, the disk drive array comprises synchronised, fault independent disk drive units.

According to another aspect of the present invention there is provided a method of correcting errors appearing in data recovered from a plurality of disk drives in an array, wherein the data is organised on the disk by sectors and groups of sectors form related rows, each row having a parity data sector and each sector including an error correction code, the method being characterised by comprising the steps of:

(a) a disk drive controller associated with each disk drive generating error syndromes for each sector upon recovery of data from the disk drives; (b) retaining the data and parity recovered with each sector in a data buffer and storing non-zero error syndromes for use in correction of data in an error syndrome buffer; (c) generating and storing in an error status buffer error status indications for each sector recovered, the indications including a no error condition, a non-zero error syndrome condition and a drive fault condition; (d) an array controller and syndrome processing means initiating a

This is the replacement sheet

Date: 07 APR 1999

correction protocol including the steps of:
(1) checking the error status for each sector for a row and continuing to the next row of sectors if no errors are present, (2) if error is indicated for one sector, correcting the sector using parity correction means and returning to step (1) for the next row of sectors, (3) if error is indicated for more than one sector of the row, using the error syndromes to locate sectors correctable with the error syndromes and correcting the data for those sectors until one sector having error is left, and then returning to step (2), (4) where the number of errors in a row cannot be reduced to one, initiating a re-try read of a data storage unit from which a sector having not correctable error was recovered and returning to step (1), if successful, and (5) indicating a read failure and aborting the data recovery attempt if step (4) is not successful.

The method may include the additional step after step (d) (4) of initiating a recovery of the disk drive and returning to step (1) if successful.

The invention is illustrated, merely by way of example, in the accompanying drawings, in which:-

Figure 1 is a block diagram of a data processing system;

Figure 2 is a block diagram of a data recovery channel according to the present invention; and

Figure 3 illustrates organisation of data buffers of the data recovery channel of Figure 2.

Figure 1 shows a data processing system 10 with which a data recovery channel according to the present invention can be advantageously used. The description of the architecture of the data processing system 10 is intended only to give an environment for explanation of the present invention and is not intended as a

This is the replacement sheet

Date: 07 APR 1999

description of a particular computer architecture with which the present invention is used. The data processing system 10 is entirely conventional and includes a central processing unit 12, a direct access memory 14 and a data storage peripheral 16. The central processing unit 12 stores data and programming steps to, and recovers data and programming steps over a data bus 18 from the memory 14 and the data storage peripheral 16. The data storage peripheral 16 can transfer data and programming steps over the data bus 18 directly to the memory 14 or to the central processing unit 12.

The central processing unit 12 exercises control over the memory 14 and the data storage peripheral 16 over a control bus 20 and an address bus 22, thereby directing the timing of the transfer of data and locations to which the data is delivered or from which it is called. The data storage peripheral 16 may be a tape drive, a disk drive, or some other form of indirectly addressed, mass data storage structure. Communication between the data storage peripheral 16 and the rest of the data processing system 10 is through an input/output interface 24. Where the data storage peripheral 16 is in an array of disk drives, a data recovery channel according to the present invention can be advantageously incorporated in the interface 24.

Figure 2 illustrates the data storage peripheral 16, the interface 24 and a data recovery channel 26 according to the present invention. The data storage peripheral 16 has n+1 disk drives 28-0 to 28-n. Each disk drive is under the direct control of one of n+1 drive controllers 30-0 to 30-n. The drive controllers 30-0 to 30n, among other functions, provide a passage for data between the disk drives 28-0 to 28-n and data

This is the replacement sheet

Date: 07 APR 1999

buffers 32-0 to 32-n, respectively.

Data records on the disk drives 28-0 to 28-n are organised by sectors. A sector includes preamble information, data which is derived from the record applied to the data storage peripheral 16 for storage and an error detection and correction code generated from the data. The generation of such sectors and the operation of writing the sectors to the disk drives 28-0 to 28-n forms no part of the present invention. The sectors stored on the disk drives 28-0 to 28-(n-1) contain the data received over the data bus 18 for storage and are referred to here as data sectors. The sectors stored on the disk drive 28-n contain parity data related to the data stored in the first n drives and are referred to here as parity sectors. Parity sectors are otherwise like data sectors in that they include error correction codes.

Read back of data records from the disk drives 28-0 to 28-n is done simultaneously and in parallel by the drive controllers 30-0 to 30-n upon receipt of a read command from an array controller 34 in the interface 24. Groups of sectors, designated here as rows, are inter-related. A row includes a plurality of data sectors, one from each disk, and a parity sector. The parity data in the parity sector is generated from logically associated bits between the data sectors. For example, the first parity bit in the parity sector may be parity for the first data bit from each data sector, the second parity bit in the parity sector is calculated over the second data bits of the data sectors and so on. The sectors of a row are stored to the same logical addresses in the data buffers 32-0 to 32-n. During execution of a read back of data records, the drive controllers 30-0 to 30-n operate on the recovered data

This is the replacement sheet

Date: 07 APR 1999

and ECC to generate error syndromes for each sector. The error syndromes, if non-zero, are loaded into known locations in an error syndrome buffer 36 for possible later use. A non-zero error syndrome set for a sector indicates presence of a code error in the sector, which may or may not be correctable using the error syndromes. The drive controllers 30-0 to 30-n also monitor the disk drives 28-0 to 28-n for indications of drive failure or fault.

A data record includes a plurality of rows of sectors. During read back of a record from the disk drives 28-0 to 28-n by the drive controllers 30-0 to 30-n, the drive controllers generate error status indications for each sector read and store the error status indications in status buffers 38-0 to 38-n. An error status indication of 0,0 indicates a sector free of detected error; an error status indication of 0,1 indicates a sector for which non-zero error syndromes were generated upon read back; an error status record of 1,0 indicates occurrence of a disk fault. An error status record of 1,1 is preferably not used. In summary, the first digit of the record is used to indicate a disk fault and the second digit is used to indicate a code fault in a sector.

Upon completion of reading of the sectors from the disk drives 28-0 to 28-n, the data from each data and parity sector will be stored to a pre-determined location in the data buffers 32-0 to 32-n, an error status indication for each data and parity sector will be stored to a logically associated address in the status buffers 38-0 to 38-n, and the non-zero error syndromes for each sector having code errors will be stored in the syndrome buffer 36.

The array controller 34 is a programmed micro-

This is the replacement sheet

Date: 07 APR 1999

computer and controls the error correction processing in the data recovery channel 26 by monitoring the status buffers 38-0 to 38-n and communicating with syndrome processing circuitry 42 over a bus 43. The data recovery channel 26 includes parity correction circuitry 40, the error syndrome buffer 36, the error syndrome processing circuitry 42 and the data buffers 32-0 to 32-n. The error syndrome processing circuitry 42 may be and preferably is combined with the array controller 34 by programming the array controller to perform the functions of the error syndrome processing circuitry. Illustration of the array controller 34 and the syndrome processing circuitry 42 as separate components is an aid to understanding and a feasible hardware configuration.

During a read operation, the array controller 34 monitors the error status records stored to the status buffers 38-0 to 38-n. A read operation is allowed to continue as long as no more than one drive fault error occurs in any given row of the recovered record, i.e. the read operation continues as long as the error status 1,0 does not repeat at the same logical address across the status buffers 38-0 to 38-n. Where two drive faults are indicated for a row, the read is aborted and extra-ordinary recovery techniques may be employed, such as a re-read attempt or an attempted recovery of the disk drive. If all such techniques fail the read operation is aborted and a drive array failure condition is indicated.

Absent indication of simultaneous drive faults for two or more of the disk drives 28-0 to 28-n, the read operation continues until completion of recovery of an entire record. The array controller 34 then directs error correction for the recovered record, row by row. As previously noted, both operation on parity data and

This is the replacement sheet

Date: 07 APR 1995

on the error syndromes is available. The array controller 34 interrogates the error status words for a row and determines the number of sectors for which drive fault and ECC error is present. Where error of either type is indicated as occurring in none or only one sector for the row, all of the sectors of the row are transferred from the data buffers 32-0 to 32-n to the parity correction circuitry 40 for correction, if required, through a sector divider 44 for reconstruction of the original words of the record and transmission out to the data bus 18. Parity correction provides superior speed over an error syndrome operation for correction of a sector, therefore it is not even determined whether a sector having error is correctable using the error syndromes. The array controller 34 indicates to the parity correction circuitry 40 the sector having error. Correction involves the regeneration of the data contained in the sector having the error from the remaining data and parity information.

Where positive error status indications exist for two or more sectors of a row, utilisation of the error syndromes for correction of one or more sectors is attempted. The array controller 34 first locates those sectors having code errors, indicated in the status buffers 38-0 to 38-n by 0,1 records for each sector. The array controller 34 then directs the syndrome processing circuitry 42 to analyse the error syndromes for a given sector, typically beginning with the error syndromes for the first sector for which an error status of 0,1 exists. By "first" will typically be meant the sector associated with the lowest numbered data buffer (e.g. data buffer 32-1 before data buffer 32-3).

The first operation performed by the syndrome

This is the replacement sheet

Date: 07 APR 1999

processing circuitry 42 is a determination whether the error syndromes for a sector can be used to correct the data. If the error is correctable the error syndrome processing circuitry 42 corrects the data in the appropriate data buffer 32-0 to 32-n. The array controller 34 re-sets the status buffer of the associated error status word to indicate no error in the sector. The array controller 34 will continue correction of sectors using error syndromes, if possible, until only one defective sector remains for a row. When only one error containing sector remains, regardless of whether the error can be corrected using error syndromes or not, the entire row is transferred to the parity correction circuitry 40 for parity correction of the sector containing error. The corrected row is then transmitted to the sector divider 44 for re-construction of the original words of the data record and on to the data bus 18. The array controller 34 then advances the correction algorithm to the next row. Correction continues until all sectors of the file are corrected and transmitted on to the data bus 18.

Should two or more sectors of a row have massive errors, i.e. error exceeding the capacity of the error syndromes, correction by the array controller 34 is not attempted. Instead, recovery is aborted and the data from at least one of the affected disk drives 28-0 to 28-n is re-read. Numerous conventional data recovery methods may be attempted, before the recovery attempt is aborted.

The original data words of a record are re-assembled before transmission on the data bus 18. The sector divider 44 is a bi-directional element which receives data words over the data bus 18 and routes each subsequent bit or byte of the word to a succeeding disk

This is the replacement sheet

Date: 07 APR 1999

drive, wrapping around to the first disk drive after assignment of n bits (or bytes). The sector divider 44 also operates to re-assemble the original words when a record is transmitted out of the parity correction circuitry 40.

Figure 3 illustrates the storage of an exemplary data record among the memory units comprising the data buffers 32-0 to 32-n after recovery of the record from the disk drives 28-0 to 28-n. The exemplary record includes 4 x n data sectors, stored four (numbered 0 to 3 for each buffer) to a data buffer, and four parity sectors (numbered 0 to 3) stored in a data buffer 32-0 to 32-n. Like numbered data and parity sectors belong to a row, e.g. sectors 1 of the data bufers 32-0 to 32-n are a row, sectors 3 of the data buffers 32-0 to 32-n are another row. One data bit from each data sector in a row relates to a particular parity bit from the parity sector for the row. In a typical arrangement only the original data of the sector having bits numbered 0 to 255 are in a data buffer after down loading a sector from its disk drive.

The status buffers 38-0 to 38-n consist of storage locations corresponding one for one with each location for sectors to be stored in data buffers 32-0 to 32-n. As recovery of data and parity sectors is attempted from the disk drives 28-0 to 28-n and the data or parity transferred to locations in the data buffers 32-0 to 32-n, error status indication words for the sectors are generated and stored in the corresponding memory locations of the status buffers 38-0 to 38-n. The first bit location of the error status indication words are the indicators of drive fault, the second bit location of the words are indication of non-zero error syndromes. Initiation of data recovery and transfer in the data

This is the replacement sheet

Date: 07 APR 1999

buffers results in all locations of the error status buffers 38-0 to 38-n being re-set. An error status of 0,0, or re-set value, is indication of no error in the corresponding data or parity sector.

Sector 3 stored in the data buffer 32-0 illustrates indication of non-zero error syndromes for a sector. The status buffer 38-0 has four error status locations 38-0(0) to 38-0(3). The error condition of sector 3 of data buffer 32-0 is reflected by the second bit of location 38-0(3) which is set at 1. Similarly, sector 1 of buffer 32-n shows a drive fault error. Location 28-n(1) shows a first bit location as 1 indicating the error condition of the sector.

The present data recovery channel corrects the data of a sector recovered from a disk drive, wherein sectors include error correction codes, groups of sectors are related by the existence of parity data for the group. Two methods of correction for sectors are used, one being parity, the other method using error syndromes. Correction employs the following priority if no more than one sector is lost per group due to drive failure:

(1) Correction of a sector by parity within a parity related group;

(2) Correction of sectors by error correction syndromes until the number of sectors evidencing error in a parity related group is reduced to one, then correcting that sector with parity;

(3) If two or more sectors evidence drive fault and a non-correctable error condition in a parity related group, re-trying the read for at least one disk drive.

This is the replacement sheet

Date: 07 APR 1999

(4) If the re-try read fails, attempting a "recover" of at least one disk drive (if the disk controller admits the command); and

(5) Signalling the data storage peripheral "FAIL" condition if all the above fail.

As will be appreciated, the data recovery channel preferentially operates on parity data for error correction over utilisation of error syndromes for error correction, though both remain available. In summary, absent indication of errors from more than one sector for a row of sectors, correction of error in the defective sector is made by use of parity information. Where error is indicated for more than one sector in a row of sectors, correction using the error syndromes is attempted, sector by sector, until the number of sectors in the row having error is reduced to one. Parity is used to correct the remaining defective sector.

Error syndromes are generated for each sector upon recovery of the sector from a disk drive. A non-zero error syndrome set for a sector indicates the possible presence of error and results in generation of a signal indicating such occurrence. A failure to recover a sector results in generation of a signal indicating a drive fault. A status buffer saves indications of fault from a failing drive, the occurrence of non-zero error syndromes or an error free condition for each recovered sector. An error syndrome buffer retains the error syndromes if non-zero.

An array controller monitors the status buffer and continues the read operation as long as no more than one drive fault is indicated with respect to a row of data

This is the replacement sheet

Date: 07 APR 1999

sectors. Upon completion of reading of all the rows of a file, the data recovery channel has access to the sector status information, recovered data sectors and to the error syndromes. Error correction of the sectors, re-construction of the original data words and transmission out of the original record is then attempted.

The array controller interrogates the sector status information and either executes, or directs execution, of error correction routines to restore the original data. Utilisation of the error syndromes for data re-construction is deferred in favour of parity re-construction if errors occur in only one sector for each row of sectors.

Error status for each sector is retained in an error status buffer. Where the status buffer indicates no errors for a row, the original data words are re-constructed and the data transmitted to the utilising computer. Where error of either type is indicated for one sector of a row, parity is used to re-construct the affected sector. Upon indication of error for more than one sector of a row, the array controller attempts restoration of the defective sectors of the row utilising error syndrome data. The array controller or syndrome processing circuitry which may be a program executed by the array controller, operates on the error syndromes for a first defective sector from the row to determine if the sector is correctable. If the sector is correctable, correction is carried out, the restored data is stored to the data buffer and the error status buffer for the sector is re-set. Once the number of sectors in a row with error of either type is reduced to one, parity correction is done and the entire row is transferred out to the utilising computer. Where the

This is the replacement sheet

Date: 07 APR 1999

number of sectors in a row with error cannot be reduced below two using error syndromes, the read operation is interrupted and the more extensive data recovery methods, directed at the affected disk drives are attempted, e.g. a re-read of disk drives.

This is the replacement sheet

Date: 07 APR 1999



CLAIMS

1. A method of correcting errors appearing in data recovered from a plurality of disk drives (28-0 to 28-n) in an array, wherein the data is organised on the disk by sectors and groups of sectors form related rows, each row having a parity data sector and each sector including an error correction code, the method being characterised by comprising the steps of: (a) a disk drive controller (30-0 to 30-n)associated with each disk drive (28-0 to 28-n) generating error syndromes for each sector upon recovery of data from the disk drives; (b) retaining the data parity recovered with each sector in a data buffer (32-0 to 32-n) and storing non-zero error syndromes for use in correction of data in an error syndrome buffer (36); (c) generating and storing error in an error status buffer (38-0 to 38-n) status indications for each sector recovered, the indications including a no error condition, a non-zero error syndrome condition and a drive fault condition; (d) an array controller (34) and syndrome processing means (42) initiating a correction protocol including the steps of : (1) checking the error status for each sector for a row and continuing to the next row of sectors if no errors are present, (2) if error is indicated for one sector, correcting the sector using parity correction means (40) and returning to step (1) for the next row of sectors, (3) if error is indicated for more than one sector of the row, using the error syndromes to locate sectors correctable with the error syndromes and correcting the data for those sectors until one sector having error is left, and then returning to step (2), (4) where the number of errors in a row cannot be reduced to one, initiating a re-try read of a data storage unit from which a sector having not correctable error was recovered and returning to step (1), if successful, and (5) indicating a read failure and aborting the data recovery attempt if step (4) is not successful.

This is the replacement sheet

Date: 07 APR 1999

* C L M S *

2. A method as claimed in claim 1 characterised by including an additional step after step (d) (4) of initiating a recovery of the disk drive and returning to step (1) if successful.

3. A data recovery channel in a fault tolerant disk drive array (16), the disk drive array including a plurality of disk drives (28-0 to 28-n) and an input interface (24) for receiving data records from a data processing system (10), the input interface (24) organising each data record received into data sectors for storage on the disk drives, identifying a plurality of sectors as a related group and generating a parity sector for the group, where each data and parity sector includes an appended error correction code, the group of data and parity sectors being stored by striping across the available disk drives, the data recovery channel being characterised by comprising: a disk drive controller (30-0 to 30-n) associated with each disk drive (28-0 to 28-n) including means for recovering sectors from the disk drive, means for determining error syndromes for each sector as recovered, and means for detecting a disk drive fault; a data buffer (32-0 to 32-n) for retaining the data and parity recovered with each sector; an error status buffer (38-0 to 38-n) for retaining indication of non-zero error syndromes generated upon recovery of a sector or occurrence of a disk fault associated with attempted recovery of a sector; an error syndrome buffer (36) for retaining any non-zero error syndromes generated during recovery of data and parity sectors; syndrome processing means (42)

This is the replacement sheet

Date: 07 APR 1999

* C L M S *

connected to the error syndrome buffer for determining the correctability of the data of a sector, and for correcting the data in the data buffers, if possible; parity correction means (40) for receiving associated groups of data and parity information from the data buffers and correcting the data from up to one sector; and an array controller (34) for operating upon the error status buffer and indications of correctability of data from the syndrome processing means to direct the correction method of claim 1 for each sector of a group.

4. A data recovery channel as claimed in claim 3, characterised in that the array controller (34) further comprises: means for monitoring the error status buffer (38-0 to 38-n); means for transferring the data and parity for a group of sectors to the parity correction means (40) when the number of errors indicated for a group of sectors is or is reduced to one or less; and means for re-setting the status indications for a sector in the error status buffer (38-0 to 38-n) where analysis of the error syndromes for the sector shows that the data from the sector can be corrected.

5. A data recovery channel as claimed in claim 4 characterised in that each status buffer (38-0 to 38-n) has a memory location for each sector recovered of a data record.

6. A data recovery channel as claimed in claim 5 characterised in that the array controller 34 further comprises: means for initiating a disk read re-try when more than one disk fault is indicated for a group, and means for indicating a read failure when a disk read re-try fails.

This is the replacement sheet

Date: 07 APR 1999

* C L M S *

7. A data recovery channel as claimed in claim 5 or 6 characterised in that the array controller further comprises: means for initiating a disk read re-try when the number of sectors having a fault cannot be reduced to one in number; and means for indicating a read failure when a disk read re-try fails.

8. A data recovery channel as claimed in any of claims 3 to 7 in combination with a disk drive array characterised in that the disk drive array comprises synchronised, fault independent disk drive units.

This is the replacement sheet

Date: 07 APR 1999

ABSTRACT

A Data Recovery Channel in a Fault Tolerant Disk Drive Array and a Method of Correcting Errors Therein

An input interface for an array of a plurality of associated disc drives, which stores data in related sectors and apportions a parity sector for related data sectors, where each data and parity sector include an appended error correction code. A data recovery channel comprises a controller (30-0 to 30-n) for each disc drive (28-0 to 28-n); a data buffer (32-0 to 32-n) retaining data and parity for a sector; an error status buffer (38-0 to 38-n) retaining indications of non-zero error syndromes or disk fault associated with data recovery; an error syndrome buffer (36) retaining any non zero error syndromes generated during recovery of data and parity sectors; syndrome processing circuitry (42) for determining correctability of and correcting data; parity correction circuitry (40) for correcting parity data in a sector; and an array controller (34) operating upon the error status buffer and indications of correctability of data to direct the correction method for each related sector.

[Figure 2]

This is the replacement sheet

Date: 07 APR 1999

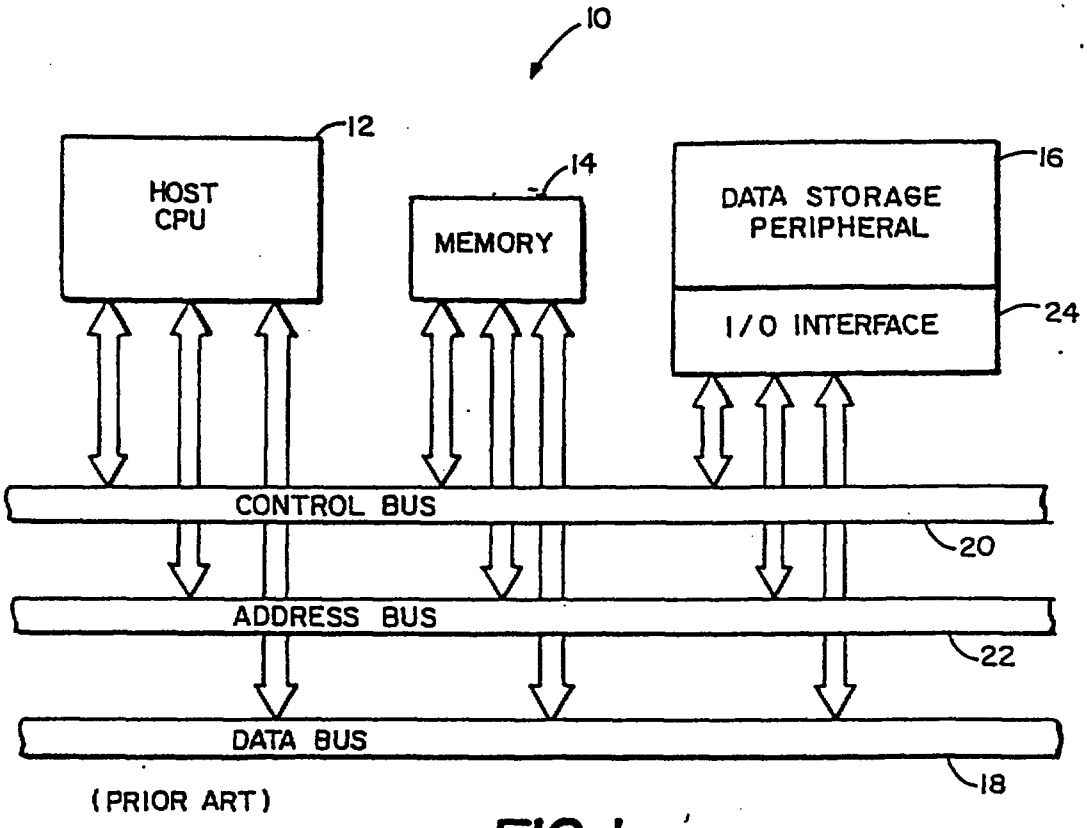
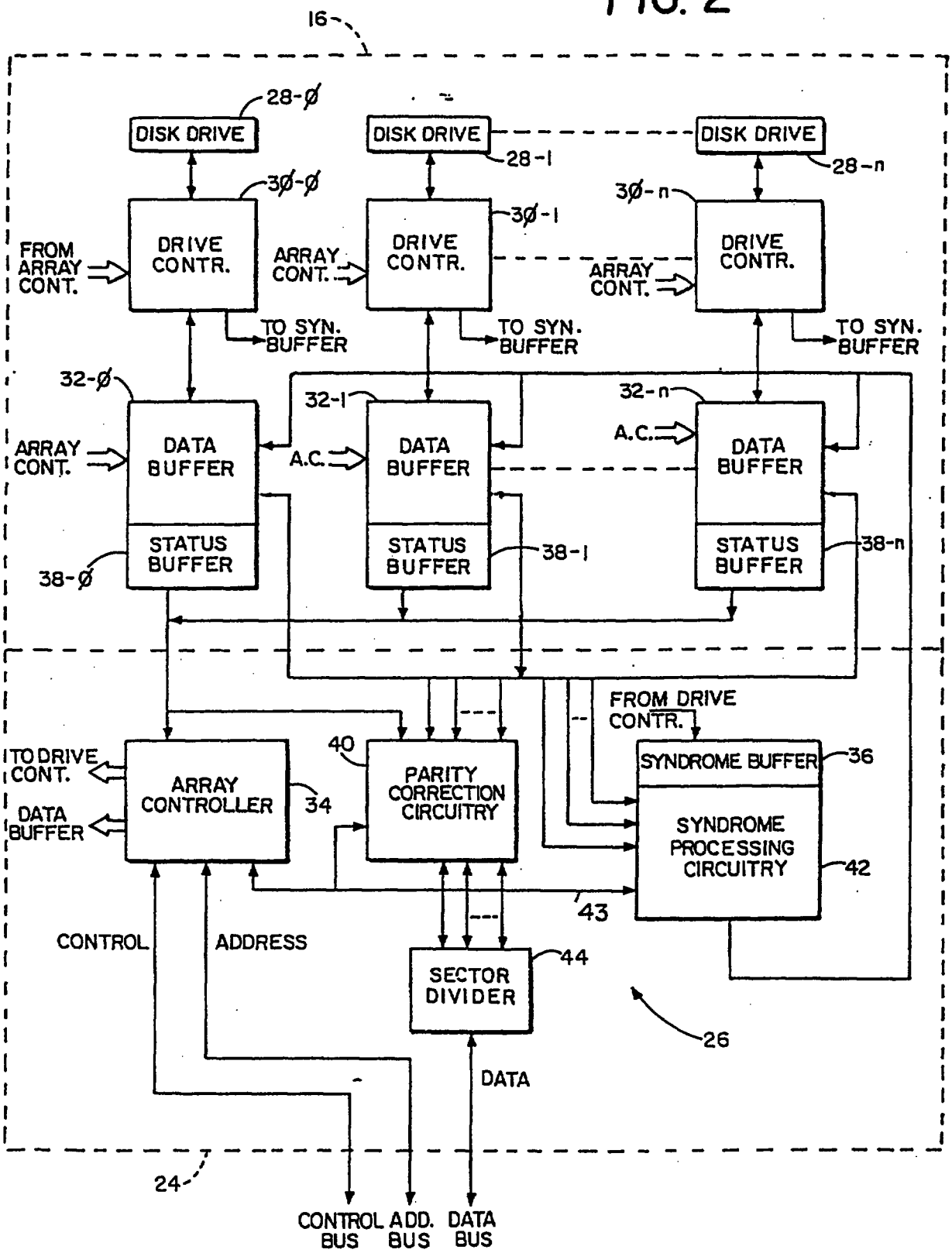


FIG. 1

This is the replacement sheet

Date: 07 APR 1999

FIG. 2



This is the replacement sheet

Date: 07 APR 1999

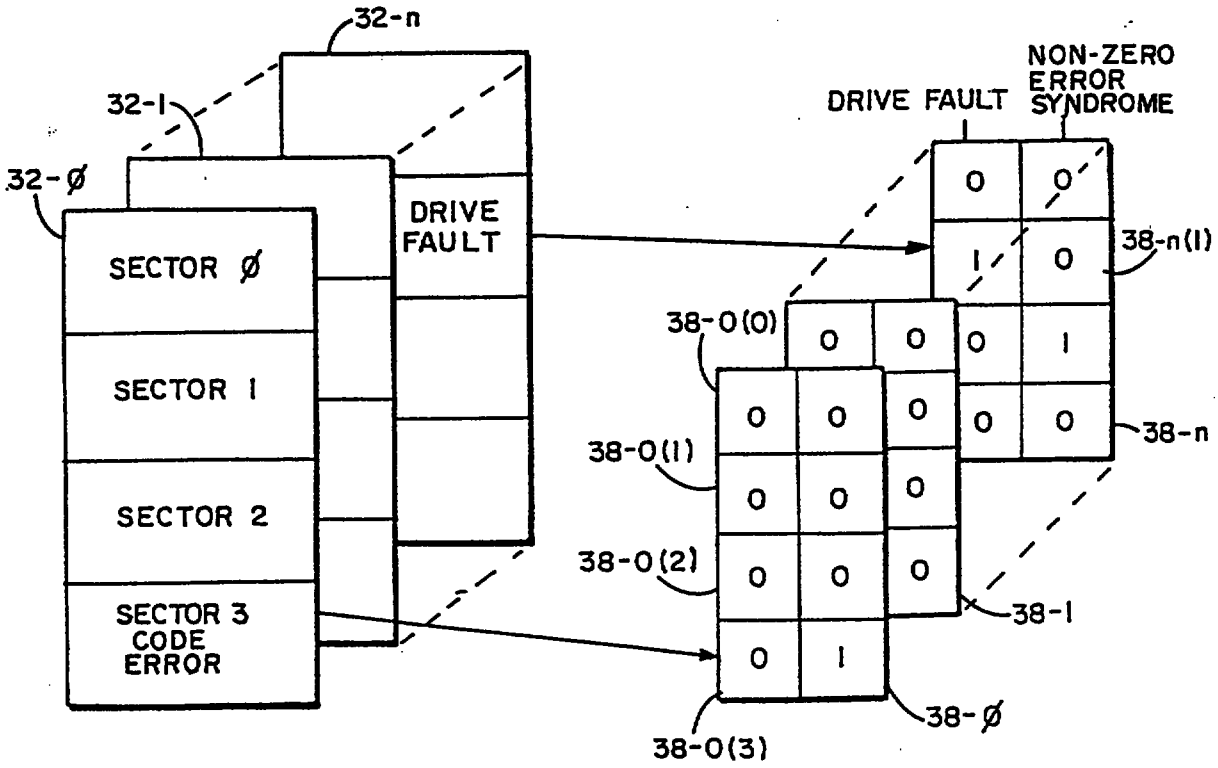


FIG. 3

This is the replacement sheet

Date: 07 APR 1999