(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4703945号
(P4703945)
(45)発行日　平成23年6月15日(2011.6.15)　(24)登録日　平成23年3月18日(2011.3.18)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 305C
G06F 3/06 540

請求項の数 24　(全 19 頁)

(21)出願番号　特願2003-62749(P2003-62749)
(22)出願日　平成15年3月10日(2003.3.10)
(65)公開番号　特開2004-30577(P2004-30577A)
(43)公開日　平成16年1月29日(2004.1.29)
審査請求日　平成18年3月2日(2006.3.2)
審判番号　不服2008-15087(P2008-15087/J1)
審判請求日　平成20年6月16日(2008.6.16)
(31)優先権主張番号　10/094086
(32)優先日　平成14年3月8日(2002.3.8)
(33)優先権主張国　米国(US)

(73)特許権者　303039534
ネットアップ，インコーポレイテッド
アメリカ合衆国　カリフォルニア　９４０８９，　サニーヴェール，　イースト　ジャバ　ドライブ　４９５
(74)代理人　100087642
弁理士　古谷　聡
(74)代理人　100076680
弁理士　溝部　孝彦
(74)代理人　100121061
弁理士　西山　清春
(72)発明者　スティーブン・アール・クレイマン
アメリカ合衆国カリフォルニア州９４０２２，ロスアルトス，エル・モンテ・コート・１５７

最終頁に続く

(54)【発明の名称】ストレージアレイにおける複数の記憶装置故障を訂正する方法

(57)【特許請求の範囲】

【請求項１】

ストレージアレイにおける複数の記憶装置故障を訂正するためのシステムであって、

連結された複数のサブアレイを有するストレージアレイであって、各サブアレイが、データ記憶装置の集合と、局所パリティ記憶装置とを含み、２重故障保護符号化方法を使用して、恰も各サブアレイが単独で存在するかのように、各サブアレイに対して同様に対角パリティ集合が割り当てられ、該ストレージアレイが、各サブアレイにおける同等の対角パリティ集合から計算された対角パリティを保持する大域パリティ記憶装置を更に含み、各サブアレイの前記局所パリティ記憶装置と共に前記大域パリティ記憶装置を使用してサブアレイ内の２重故障を訂正するように構成され、前記局所パリティ記憶装置が、特定のサブアレイ内の全てのデータ記憶装置に関する行パリティを記憶する記憶装置であり、前記大域パリティ記憶装置が、各サブアレイの対角パリティ集合に沿って対角パリティブロックを計算し、前記サブアレイの対応する対角パリティ集合に沿って計算された対角パリティブロックを排他的論理和演算により１つにまとめることにより計算される対角パリティを記憶する記憶装置である、ストレージアレイを含むシステム。

【請求項２】

前記局所パリティ記憶装置は、前記サブアレイの行パリティ集合内の単一装置の故障の訂正に使用される単一装置誤り訂正方法を使用して符号化された値を記憶するように構成される、請求項１に記載のシステム。

【請求項３】

前記行パリティ集合はブロックの行である、請求項２に記載のシステム。
【請求項４】
２重装置故障に対する保護を提供する前記２重故障保護符号化方法は、前記単一装置誤り訂正方法に依存しないものである、請求項２又は請求項３に記載のシステム。
【請求項５】
前記２重故障保護符号化方法は行－対角符号化である、請求項１～４のうちのいずれか一項に記載のシステム。
【請求項６】
前記単一装置誤り訂正方法は行パリティである、請求項２～５のうちのいずれか一項に記載のシステム。
【請求項７】
各サブアレイは集中パリティ装置アレイとして編成される、請求項１～６のうちのいずれか一項に記載のシステム。
【請求項８】
各サブアレイの局所パリティ記憶装置、及びデータ記憶装置の集合を、局所パリティブロックの位置が記憶装置ごとに異なる分散パリティ装置アレイに置き換えた、請求項１～６のうちのいずれか一項に記載のシステム。
【請求項９】
前記記憶装置は、磁気テープ、光学式ＤＶＤ、バブルメモリ、電子的ランダムアクセスメモリ又は磁気ディスク装置である、請求項１～８のうちのいずれか一項に記載のシステム。
【請求項１０】
ストレージアレイにおける２重故障の訂正のためのデータを符号化するためにコンピュータで実施される方法であって、
連結された複数のサブアレイとしてストレージアレイを構成するステップであって、各サブアレイが、データ記憶装置の集合と局所パリティ記憶装置とを含み、前記ストレージアレイが、対角パリティを保持するための大域パリティ記憶装置を更に含み、前記局所パリティ記憶装置に、特定のサブアレイ内の全てのデータ記憶装置に関する行パリティを記憶し、前記大域パリティ記憶装置に、対角パリティを記憶する、ストレージアレイを構成するステップと、
２重故障保護符号化方法を使用して、恰も各サブアレイが単独で存在するかのように、各サブアレイに対して同様に対角パリティ集合を割り当てるステップと、
各サブアレイの対角パリティ集合に沿って対角パリティブロックを計算し、前記サブアレイの対応する対角パリティ集合の計算された対角パリティブロックを排他的論理和演算により１つにまとめることにより前記対角パリティを計算するステップと、
からなる方法。
【請求項１１】
単一装置誤り訂正方法を使用して符号化されたパリティ値を各サブアレイの局所パリティ記憶装置に記憶するステップを更に含む、請求項１０に記載のコンピュータで実施される方法。
【請求項１２】
第２の装置故障に対する保護を提供する前記２重故障保護符号化方法は、前記単一装置誤り訂正方法に依存しないものである、請求項１０に記載のコンピュータで実施される方法。
【請求項１３】
前記構成するステップは、各サブアレイを集中パリティ装置アレイとして編成することからなる、請求項１０に記載のコンピュータで実施される方法。
【請求項１４】
前記構成するステップは、各サブアレイの局所パリティ記憶装置、及びデータ記憶装置の集合を、局所パリティブロックの位置が記憶装置ごとに異なる分散パリティ装置アレイ

に置き換えることからなる、請求項１０に記載のコンピュータで実施される方法。
【請求項１５】
前記記憶装置は、ビデオテープ、磁気テープ、光学式ＤＶＤ、バブルメモリ、電子的ランダムアクセスメモリ、又は、磁気ディスク装置である、請求項１０に記載のコンピュータで実施される方法。
【請求項１６】
ストレージアレイにおける２重故障を訂正するための装置であって、
連結された複数のサブアレイとしてストレージアレイを構成する手段であって、各サブアレイが、データ記憶装置の集合と局所パリティ記憶装置とを含み、前記ストレージアレイが、対角パリティを保持するための大域パリティ記憶装置を更に含み、前記局所パリティ記憶装置に、特定のサブアレイ内の全てのデータ記憶装置に関する行パリティを記憶し、前記大域パリティ記憶装置に、対角パリティを記憶する、ストレージアレイを構成する手段と、
２重故障保護符号化方法を使用して、恰も各サブアレイが単独で存在するかのように、各サブアレイに対して同様に対角パリティ集合を割り当てる手段と、
各サブアレイの対角パリティ集合に沿って対角パリティブロックを計算し、前記サブアレイの対応する対角パリティ集合の計算された対角パリティブロックを排他的論理和演算により１つにまとめることにより前記対角パリティを計算する手段と、
各サブアレイに関連する前記局所パリティ記憶装置及び前記ストレージアレイに関連する前記大域パリティ記憶装置に対するパリティ復号演算を使用して、前記アレイ内の記憶装置故障を訂正する手段と、
からなる装置。
【請求項１７】
前記構成する手段は、各サブアレイを集中パリティ装置アレイとして編成する手段からなる、請求項１６の装置。
【請求項１８】
前記構成する手段は、各サブアレイの局所パリティ記憶装置、及びデータ記憶装置の集合を、局所パリティブロックの位置が記憶装置ごとに異なる分散パリティ装置アレイに置き換える手段からなる、請求項１６の装置。
【請求項１９】
前記記憶装置は、ビデオテープ、磁気テープ、光学式ＤＶＤ、バブルメモリ、電子的ランダムアクセスメモリ、又は、磁気ディスク装置である、請求項１６に記載の装置。
【請求項２０】
前記パリティ符号化演算及び復号化演算は、フィールド・プログラマブル・ゲートアレイ又は特定用途向け集積回路等の専用ハードウェアで実施される、請求項１６に記載の装置。
【請求項２１】
ストレージアレイにおける２重故障を訂正するための実行可能プログラム命令を含むコンピュータ読取可能媒体であって、該実行可能プログラム命令が、
連結された複数のサブアレイとしてストレージアレイを構成するプログラム命令であって、各サブアレイが、データ記憶装置の集合と局所パリティ記憶装置とを含み、前記ストレージアレイが、対角パリティを保持するための大域パリティ記憶装置を更に含み、前記局所パリティ記憶装置に、特定のサブアレイ内の全てのデータ記憶装置に関する行パリティを記憶し、前記大域パリティ記憶装置に、対角パリティを記憶する、ストレージアレイを構成するプログラム命令と、
２重故障保護符号化方法を使用して、恰も各サブアレイが単独で存在するかのように、各サブアレイに対して同様に対角パリティ集合を割り当てるためのプログラム命令と、
各サブアレイの対角パリティ集合に沿って対角パリティブロックを計算し、前記サブアレイの対応する対角パリティ集合の計算された対角パリティブロックを排他的論理和演算により１つにまとめることにより前記対角パリティを計算するためのプログラム命令と

、
各サブアレイに関連する前記局所パリティ記憶装置及び前記ストレージアレイに関連する前記大域パリティ記憶装置を使用して、前記アレイ内の記憶装置故障訂正するためのプログラム命令と、
からなる、コンピュータ読取可能媒体。
【請求項２２】
前記計算するためのプログラム命令は、各サブアレイの対角パリティ集合の対角パリティブロックを計算するためのプログラム命令を含む、請求項２１に記載のコンピュータ読取可能媒体。
【請求項２３】
前記計算するためのプログラム命令は、前記サブアレイの対応する対角パリティ集合の計算された対角パリティブロックを論理的に結合し、対角パリティとして前記大域パリティ記憶装置に記憶するためのプログラム命令を更に含む、請求項２２に記載のコンピュータ読取可能媒体。
【請求項２４】
前記計算するためのプログラム命令は、前記大域パリティ記憶装置に記憶された対角パリティから他のサブアレイの前記結合された対角パリティブロックを減算することによって、任意のサブアレイ前記計算された対角パリティブロックを復元するためのプログラム命令を含む、請求項２３に記載のコンピュータ読取可能媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明はストレージシステムのアレイに関し、詳しくは、スレージアレイ内の任意の１台の故障した記憶装置、または、任意の２台の故障した記憶装置の組み合わせを効率よく復元するための技術に関するものである。
【０００２】
【従来の技術】
ストレージシステムは通常１以上の記憶装置を含み、要求に応じてそれらの記憶装置にデータを入力したりそれらの記憶装置からデータを取得したりすることができる。ストレージシステムは、限定はしないが、ネットワークに取り付けられたストレージ環境、ストレージエリアネットワーク、及び、クライアントまたはホストコンピュータに直接取り付けられたディスクアセンブリを含む、様々なストレージアーキテクチャに従って実施される。記憶装置は典型的にはディスクドライブであり、ここで「ディスク」という用語は一般に独立型の回転式磁気媒体記憶装置を意味している。この文脈での用語「ディスク」は、ハードディスクドライブ（ＨＤＤ）やダイレクトアクセス記憶装置（ＤＡＳＤ）と同義である。
【０００３】
ストレージシステム内のディスクは通常、１以上のグループに編成され、各グループがＲＡＩＤ（Redundant Array of Independent(or Inexpensive) Disks）として運用されている。大半のＲＡＩＤ実施形態は、ＲＡＩＤグループ内の所定数の物理ディスクに「ストライプ状」にまたがるデータの冗長書き込み、及び、そのストライプ状になったデータに関する冗長情報の適切な記憶により、データ記憶の信頼性／完全性を向上させている。この冗長情報によって、記憶装置が故障したときのデータ損失の復旧が可能になる。
【０００４】
ディスクアレイを運用する場合、ディスクが故障し得ることを考慮している。高性能ストレージシステムの目標は、ＭＴＴＤＬ（Mean time to data loss）を可能な限り長くすることであり、システムの推定サービス寿命よりも長くすることが好ましい。１以上のディスクが故障した場合、データが失われる可能性があり、その装置からデータを復旧させることは不可能になる。データの損失を回避する一般的な手段としては、ミラーリング、バックアップ、パリティ保護が挙げられる。ミラーリングは、ディスク等のストレージリソ

ースの消費という観点からは、高価な解決方法である。バックアップは、バックアップが作成された後に変更されたデータを保護することができない。パリティ手段は、わずか１台のディスクドライブをシステムに追加するだけでデータの冗長符号化を提供し、単一削除（１台のディスクの損失）を許容するので、一般的である。

【０００５】

パリティ保護は、ディスク等の記憶装置上のデータの損失を防止するためにコンピュータシステムで用いられる。パリティ値は、異なるデータを有する多数の同様のディスクにわたって、あるワードサイズ（通常１ビット）のデータを足し合わせる（通常モジュロ２で）ことにより計算される。すなわち、パリティは、各々のディスク上の対応する位置にあるビットから構成される１ビット幅のベクトルについて計算される。１ビット幅のベクトルについて計算される場合、パリティは、合計として計算される場合と、その補数として計算される場合とがあり、これらはそれぞれ偶数パリティ、奇数パリティと呼ばれる。１ビットベクトルに対する加算および減算は、いずれも排他的論理和（ＸＯＲ）演算と同じである。そして、複数のディスクのうちのいずれか１台の損失、あるいは、複数のディスクのうちのいずれか１台の任意の部分のデータの損失から、データが保護される。パリティを記憶しているディスクが失われた場合、パリティはデータから再生成することができる。データディスクのうちの１つが失われた場合、そのデータは、残ったディスクの内容を加え合わせてその結果を記憶されているパリティから減算することにより再生成することができる。

【０００６】

通常、ディスクはパリティグループに分割され、パリティグループの各々が１以上のデータディスクと１つのパリティディスクとを含む。パリティ集合は複数のデータブロックと１つのパリティブロックとを含むブロックの集合であり、ここでパリティブロックはそれらのデータブロックすべてのＸＯＲをとったものである。パリティグループは、１以上のパリティ集合を選択する元になるディスクの集合である。ディスク空間はストライプに分割され、各ストライプが各ディスクの中から１ブロックを保持している。あるストライプのブロックは通常、パリティグループ内の各ディスク上の同じ位置に存在する。ストライプ内では、１ブロックを除くすべてのブロックがデータを保持するブロック（「データブロック」）であり、１ブロックはそれらのデータ全てのＸＯＲをとることによって計算されたパリティを保持するブロック（「パリティブロック」）である。これらのパリティブロックをすべて１つのディスク上に記憶し、すべてのパリティ情報（パリティ情報のみ）を保持する１つのディスクを設けた場合、ＲＡＩＤ－４実施形態になる。それらのパリティブロックを各ストライプの異なるディスク内に保持する場合、たいていは循環パターンが用いられ、実施形態はＲＡＩＤ－５になる。用語「ＲＡＩＤ」及びその実施形態は広く知られており、１９９８年６月、データの管理に対する国際会議（ＳＩＧＭＯＤ）の議事録で、Ｄ．Ａ．Ｐａｔｔｅｒｓｏｎ、Ｇ．Ａ．Ｇｉｂｓｏｎ、およびＲ．Ｈ．Ｋａｔｚによる「A Case for Redundant Arrays of Inexpensive Disks(RAID)」に開示されている。

【０００７】

本明細書で用いられる場合、「符号化」という用語はデータブロックのうちの所定の部分集合にわたる冗長性値の計算を意味しているのに対して、「復号化」という用語はデータブロックの部分集合及び冗長性値を用いた冗長性計算の際の同じプロセスによるデータブロックまたはパリティブロックの復元を意味している。パリティグループ内で１台のディスクが故障した場合、そのディスクの内容は、残ったデータブロックのすべての内容を加え合わせ、その結果をパリティブロックから減算することにより、予備ディスク上に復号（復元）することができる。１ビットフィールドに対する２の補数加算及び減算はいずれもＸＯＲ演算と等しいので、この復元は、すべての生き残ったデータブロック及びパリティブロックのＸＯＲをとることから構成される。同様に、パリティディスクが失われた場合も、パリティディスクは生き残ったデータから同じ方法で再計算することができる。

【０００８】

データビットのＸＯＲ合計をパリティビット値として直接記憶するのが一般的である。こ

の方法は、通常「偶数パリティ」と呼ばれる。あるいは、データビットのＸＯＲ合計の補数をパリティビット値として記憶する場合もあり、この方法は「奇数パリティ」と呼ばれる。本明細書で開示する本発明について、偶数パリティを用いるか奇数パリティを用いるかは指定していない。しかしながら、そのような区別が問題となる場合には、本明細書で開示するアルゴリズムは偶数パリティを用いるものとして説明している。また、当業者であれば、本発明の教示に従って奇数パリティも使用できることは、明らかであろう。

【０００９】

パリティ手段は、一般にパリティグループ内の１つの故障に対する保護を提供するものである。これらの手段は、故障が異なるパリティグループ内で発生する限りは、複数のディスク故障に対する保護を提供することも可能である。しかしながら、パリティグループ内で２つのディスクが同時に故障した場合、復元不能なデータの損失をこうむる。パリティグループ内で２台のディスクが同時の故障することはかなり一般的に起こりうるものであり、特にその原因は、ディスクの磨耗、及び、ディスクの動作に関する環境要因である。この文脈で、パリティグループ内での２台のディスクの同時の故障は「２重故障」と呼ばれる。

【００１０】

２重故障は、典型的には、１台のディスク故障と、その最初の故障からの復旧を試みている間に生じた他のディスクの故障の結果として発生する。復旧時間または復元時間は、ストレージシステムのアクティビティのレベルに依存している。すなわち、故障したディスクを復元している間、ストレージシステムは「オンライン」のままであり、データへのアクセス（読み出し及び／又は書き込み）の要求（クライアントまたはユーザからの）に対してサービスを提供することが可能である。ストレージシステムが要求に対するサービスの提供に忙しい場合、復元の所要時間は増加することになる。失われたデータを復元するためには生き残ったディスクのすべてを読み出す必要があるので、この復元処理時間はストレージシステムのディスクの数やサイズが増加するのに応じても増加する。さらに、２重故障の確率は、パリティグループ内のディスク数の２乗に比例する。しかしながら、パリティグループを小さくすることは、各パリティグループが冗長データを扱うために１ディスク全体を必要とするので、費用がかかる。

【００１１】

ディスクの他の故障形態は、ディスク内の１ブロックまたは１セクタが読み出せなくなるというメディア読み込みエラーである。ストレージアレイにパリティが保持されていれば、その読み出し不能なデータを復元することができる。しかしながら、あるディスクがすでに故障しているときに、アレイ内の他のディスクに対して読み出しエラーが起きると、データが失われる。これが２重故障の第２の形態である。２重故障の第３の形態は、同ストライプ内での２つのメディア読み込みエラーであり、めったに起こらないが起こる可能性はある。

【００１２】

従って、２重故障に対して耐性のある技術を提供することが望まれている。この技術は、より大きなパリティグループを有するより大きなディスクシステムの構成を可能にする一方、一台のディスク故障に長い時間（例えば数時間）を要した後の復元であっても、システムが２重故障に耐えることを保証する。このような技術により、ストレージシステムに対する特定の設計上の制限を緩和することができる。例えば、ストレージシステムに比較的低コストのディスクを用いても、高いＭＴＴＤＬを維持できるようになる。低コストのディスクは高コストのディスクに比べて一般に寿命が短く、寿命までに故障する確率も高い。従って、ストレージシステムがパリティグループ内の２重ディスク故障に耐えることができれば、このようなディスクの使用がもっと可能になる。

【００１３】

既知の２重故障訂正パリティ手段は、失われた（故障した）ディスクの順次復元を可能にするＥＶＥＮＯＤＤ　ＸＯＲベースの技術である。ＥＶＥＮＯＤＤパリティには、ちょうど２台のディスク分の冗長データが必要であり、これが最適である。このパリティ技術に

よると、すべてのディスクブロックは２つのパリティ集合に属しており、一方はすべてのデータディスクにわたって計算される一般的なＲＡＩＤ４スタイルのＸＯＲであり、他方は対角方向に隣り合ったディスクブロックの集合にわたって計算される。大まかに説明すると、ディスクを同サイズのブロックに分割し、ディスクにわたってストライプを形成する。各ストライプ内において、対角方向に隣接したディスクブロックの集合によって形成されるパリティを保持するために指定したディスクを対角パリティディスクと呼び、それを保持するパリティを対角パリティと呼ぶ。このブロックの集合を行パリティ集合、または「行」と呼ぶ。行のブロックのうち１ブロックがその行のパリティを保持するために選択され、残りのブロックがデータを保持する。各ストライプ内には、１台を除いて対角パリティ以外のすべてのディスクの各々から１ブロックが選択され、選択されたブロックのうちの２つが同じ行に属しないという制限をさらに加える。これを対角パリティ集合または「対角」と呼ぶ。

【００１４】

ＥＶＥＮＯＤＤ技術における対角パリティ集合は、１つを除きすべてのデータディスクからブロックを含む。ｎ台のデータディスクの場合、１ストライプ内にはｎ－１行のブロックが存在する。各ブロックが１つの対角上にあるので、長さｎ－１ブロックのｎ個の対角が存在する。注意して欲しいのは、ＥＶＥＮＯＤＤ手段はｎが素数である場合にしか機能しない点である。ＥＶＥＮＯＤＤ技術については、１９９５年ブラウム他による「A variant of EVENODD is Failure in RAID Architecture」と題したIEEE Transactions on Computers Vol.44 No.2の論文に開示されている。１９９６年１１月２６日に発行された「Method and Means for Encoding and Rebuilding the Data Contents of up to Two Unavailable DASDs in a DASD Array using Simple Non-Recursive Diagonal and Row Parity」と題したブラウム他による米国特許第５，５７９，４７５号には、様々なＥＶＥＮＯＤＤが開示されている。上記の論文および特許はここで参照することにより完全に説明したものとして取り込まれる。

【００１５】

ＥＶＥＮＯＤＤ技術は、ｐを素数として全部でｐ＋２台のディスクを使用し、そのうちのｐ台のディスクがデータを保持し、残りの２台のディスクがパリティ情報を保持する。一方のパリティディスクは行パリティブロックを保持する。行パリティは、各データディスクの同じ位置にあるすべてのデータブロックのＸＯＲとして計算される。他方のパリティディスクは対角パリティブロックを保持する。対角パリティは、複数のデータディスク上に対角パターンに配置されたｐ－１個のデータブロックから構成される。これらのブロックは、ｐ－１行のストライプにグループ化される。これは、データブロックの行パリティ集合への割り当てには影響を与えない。しかしながら、対角は、対角内のすべてのブロックが同じブロックのストライプに入るようなパターンに構成される。これは、ほとんどの対角は、ディスクからディスクへと進む際に、ストライプ内で「循環」することを意味している。

【００１６】

具体的には、ｎ×（ｎ－１）のデータブロックのアレイの場合、対角がアレイの端部で「循環する」ならば、長さｎ－１の対角がちょうどｎ個存在する。ＥＶＥＮＯＤＤパリティ配置の復元で重要なのは、各対角パリティ集合がデータディスクのうちの１台からは情報を保持していないことである。しかしながら、対角のパリティブロックを記憶するためのブロックが存在する以外に、もう１つだけ対角が存在する。すなわち、ＥＶＥＮＯＤＤパリティ配置では、独立したパリティブロックを持たない対角パリティ集合になる。この余分な「抜けている」パリティブロックを収容するため、ＥＶＥＮＯＤＤ配置は、ある特別な対角のパリティ結果と、その他の対角の各々のパリティブロックとのＸＯＲをとる。

【００１７】

図１は、従来のＥＶＥＮＯＤＤパリティ配置に従って構成された従来技術のディスクアレイ１００を示す略ブロック図である。各データブロックＤａｂはパリティ集合ａ及びｂに属しており、各パリティ集合のパリティブロックをＰａで表記している。ある特別な対角

（Ｘ）については、対応するパリティ集合が存在しないことに注意して欲しい。ここにＥＶＯＮＯＤＤの特徴が現れている。２つの故障からの復元を可能にするためには、各ディスクは少なくとも１つの対角パリティ集合に貢献してはならない。ｎ×（ｎ－１）のデータブロックのアレイを用いた場合、対角パリティ集合はｎ－１個のデータブロック要素を有する。上記のように、このような配置では、すべての対角についてパリティブロックを記憶するための位置を持つのではない。そのため、余分な（抜けている）対角パリティブロックのパリティ（Ｘ）は、その対角パリティを他の対角パリティブロックの各々のパリティとＸＯＲをとることにより記録される。具体的には、この抜けている対角パリティ集合のパリティは対角パリティブロックＰ４～Ｐ７の各々とＸＯＲをとられ、それらのブロックがＰ４Ｘ～Ｐ７Ｘで表記されている。

【００１８】

２台のデータディスクの故障から復元するためには、まず、全てのパリティブロックのＸＯＲをとることにより、パリティブロックを持たない対角のパリティを再計算する。例えば、全ての行パリティの合計は全てのデータブロックの合計である。全ての対角パリティの合計は、全てのデータブロックの合計から抜けている対角パリティブロックの合計を引いたものである。従って、すべてのパリティブロックのＸＯＲは、すべてのブロックの合計（行パリティ合計）から抜けている対角を除いた全てのブロックの合計を引いたものに等しく、これがちょうど抜けている対角のパリティになる。実際には、各対角パリティブロックについて１つ、抜けている対角パリティのｎ－１個の複製がその結果に加算される。ｎが素数であるからｎ－１は偶数であり、あるブロックを自分自身と偶数回ＸＯＲをとった結果はゼロブロックになる。従って、付加的な抜けているパリティが各々に加算された対角パリティブロックの和は、その付加的な対角パリティが無くても、対角パリティブロックの和に等しい。

【００１９】

次に、対角パリティブロックの各々からその抜けている対角パリティを減算する。２台のデータディスクが故障した後には、１ブロックしか失われていない対角パリティ集合が少なくとも２つ存在する。それらのパリティ集合の各々の中から失われたブロックは、一方の集合がパリティブロックを持たない対角であれば、復元することが可能である。これらのブロックが復元されると、２つの行パリティ集合について１要素を除いたすべての要素が利用可能になる。これにより、それらの行の失われた要素の復元が可能になる。他方の対角上でこの復元を行なうと、それらの対角上で１つだけ失われたブロックを復元するのに十分な情報が提供される。行パリティ及び対角パリティを交互に利用したこの復元パターンは、すべての失われたブロックを復元し終えるまで継続される。

【００２０】

ｎが素数なので、復元の際には、すべての対角に出くわすまで、即ちすべての失われたデータブロックが復元されるまで、循環が形成されることがない。ｎが素数でなければ状況は異なる。両方のパリティディスクが失われた場合は、データからパリティの単純な復元を実施することが可能である。データディスクと対角パリティディスクが失われた場合は、行パリティを用いてデータディスクの単純なＲＡＩＤ－４スタイルの復元を実施した後、対角パリティディスクの復元を実施する。データディスクと行パリティディスクが失われた場合は、１つの対角パリティを計算することができる。すべての対角が同じパリティを有するので、各対角上の失われたブロックを順次計算してゆくことができる。

【００２１】

各データブロックがある対角パリティ集合の要素になっているので、２台のデータディスクが失われた場合（２重故障）でも、１要素しか失わないパリティ集合が２つ存在する。各ディスクには、そのディスク上に現れない２つのパリティ集合がある。従って、２重故障の場合、復元可能な２つのパリティ集合が存在する。ＥＶＥＮＯＤＤは、両方のパリティディスクの故障からも、１台のデータディスクと１台のパリティディスクとの任意の組み合わせの故障からも復元を可能にする。また、この技術は任意の単一のディスク故障からの復元も可能にする。

【００２２】
ＥＶＥＮＯＤＤは必要なディスク数に関しては最適であるが、この技術のディスク効率は復元性能のコストで達成される。ＥＶＥＮＯＤＤはディスクアレイ全体を１つの単位として扱う。アレイ内の何らかのディスクが故障した場合、システムはアレイ内のすべてのディスクにアクセスして失われたブロックを復元しなければならない。ｎデータブロックのアレイ内で１台のディスクが故障した場合、ｎ－１台の残ったディスクすべてに加えて行パリティディスクも読み出すことにより、それらのアクセスのうちの１／ｎが満足されるにすぎない。他のディスクに対するアクセスは一回の読み出し処理で満足するので、１つの読み出し当たりの平均アクセス数は、２－１／ｎである。これは、ｎが大きい場合、復元中に、２の要素によってディスクアレイの性能が低下することを意味している。また、故障から復旧させるためにしなければならないシステムの仕事量（及び、システムが拘束される場合にはその復元時間）も、ディスクアレイの大きさに比例する。２ｎ台のディスクを有するシステムは、ｎ台のディスクを有するシステムの２倍の時間を復旧に要する。
【００２３】
発明の概要
本発明は、ストレージアレイ内の複数の記憶装置故障を効率的に訂正するための技術を含む。このストレージアレイは複数の連結されたサブアレイを含み、各サブアレイは、データ記憶装置の集合と、ブロックからなる行（すなわち行パリティ集合）内の単一装置の故障を訂正するために用いる単一装置誤り訂正方法を用いて符号化されたパリティを記憶するための１つの局所パリティ記憶装置とを含む。２重故障保護符号化方法を使用して、恰も各サブアレイが唯一単独で存在するかのようにして、各サブアレイに対して同様に対角パリティ集合を割り当てる。このアレイは、サブアレイの各々における同等の対角パリティ集合をまとめて論理加算することにより計算された対角パリティを保持するための１台の大域パリティ記憶装置をさらに含む。
【００２４】
本発明によると、対角パリティブロックは各サブアレイの対角パリティ集合に沿って計算される。次いで、サブアレイの対応する対角パリティ集合の計算された対角パリティブロックを、例えばＸＯＲ演算を用いて論理的に結合し、対角パリティとして対角パリティ記憶装置上に記憶する。その後、任意のサブアレイの計算された対角パリティブロックの内容は、大域パリティ記憶装置上に記憶された対角パリティから他のサブアレイの対角パリティブロックを結合したものを減算することにより、復元することができる。従って、局所パリティ記憶装置と共に大域パリティ記憶装置を用いることにより、１つのサブアレイ内のいかなる２重故障も訂正することができる。
【００２５】
注意すべき点は、本発明で用いる２重故障保護符号化方法は単一装置誤り訂正方法に依存しないことである。さらに、単一装置故障からの復旧に用いる方法は、行志向であって各サブアレイにおけるブロックの行が独立している限り、すなわち復旧がブロックの他の行からの情報に依存することがない限り、何も制限がない。この独立性が維持されるならば、これらの行の大きさは対角パリティを計算するために用いる行の大きさに関連した大きさである必要はない。
【００２６】
有利なことに、本発明では、アレイのサブアレイ内の２台の記憶装置の同時故障からの復旧が可能であるように構成されたアレイにおける単一故障の有効な復元も可能である。異なるサブアレイにおける何らかのデータブロックの故障時には、本発明は、例えば局所行パリティ等の単一装置故障復旧方法を用いてデータブロックの復旧を可能にする。サブアレイ内の任意の２ブロックの故障時には、本発明は、局所行パリティと大域対角パリティとの組み合わせを用いて復旧を容易にする。つまり、２重故障を有するサブアレイが１つだけである限り、大域パリティ記憶装置の内容から他のサブアレイの対角パリティ寄与を取り除くことができるので、データを復元することが可能である。さらに、本発明の技術は、故障の無い動作中にアレイに記憶されるパリティを計算するための計算負荷も低減す

る。本技術は更に、パリティ計算のオーバヘッドも低減し、従来の方法に比べて少ない計算しか必要としない。

【００２７】

本発明の上記の利点及び更なる利点は添付の図面と共に下記の説明を参照することによりさらによく理解することができ、図面中の似たような符号は同じ要素または機能的に類似した要素を示している。

【００２８】

図２は、本発明に有利に用いられるストレージシステム２２０を含む環境２００を示す略ブロック図である。本明細書で説明する本発明の技術は、ストレージシステム２２０として実現された、あるいは、ストレージシステム２２０を含む、スタンドアロンのコンピュータやその一部を含むいかなる種類の特別な目的（例えばファイルサーバやファイラー）のコンピュータまたは汎用コンピュータにも適用することができる。さらに、本発明の教示は、これらに限定はしないが、ネットワークに取り付けられたストレージ環境、ストレージエリアネットワーク、及びクライアントまたはホストコンピュータに直接取り付けられたディスクアセンブリを含む、様々なストレージシステムアーキテクチャに適合させることができる。従って、「ストレージシステム」という用語は、ストレージ機能を実施するように構成され、他の装置またはシステムに関連付けられた任意のサブシステムに加えて、それらの構成も含むように広く解釈しなければならない。

【００２９】

例示的実施形態において、ストレージシステム２２０は、システムバス２２５によって相互接続されたプロセッサ２２２、メモリ２２４及びストレージアダプタ２２８を含む。メモリ２２４は、本発明に関するソフトウェアプログラムコード及びデータ構造を記憶するためにプロセッサ及びアダプタによってアドレス指定可能な記憶場所を含む。そして、プロセッサ及びアダプタは、そのソフトウェアコードを実行し、データ構造を操作するように構成された処理要素及び／又は論理回路を含む。ストレージオペレーティングシステム６００は通常、その一部がメモリ内に存在し、処理要素によって実行され、とりわけ、ストレージシステムによって実行される記憶処理を呼び出すことにより、システム２００の機能を構成する。当業者であれば、本明細書で説明する本発明の技術に関するプログラム命令の記憶及び実行のために、様々なコンピュータ読取可能媒体を含む他の処理手段及びメモリ手段を用いることもできることは、明らかであろう。

【００３０】

ストレージアダプタ２２８は、システム２２０上で実行されているストレージオペレーティングシステム６００と協働して、ユーザ（またはクライアント）から要求された情報にアクセスする。その情報は、データ及びパリティ情報含めて情報を記憶するように構成された、ビデオテープ、光学式ＤＶＤ、磁気テープ、バブルメモリ、電子的ランダムアクセスメモリ、マイクロ電気機械、及び、何らかの他の媒体などの任意の種類の書込可能記憶要素媒体の取り付けられたストレージアレイ上に記憶される。しかしながら、本明細書で例示的に説明するように、この情報は、アレイ４００のディスク２３０（ＨＤＤ及び／又はＤＡＳＤ）等の記憶装置上に記憶される。ストレージアダプタは、従来の高性能ファイバーチャネルシリアル接続トポロジ等のＩ／Ｏ相互接続構成を介してそれらのディスクに接続された入出力（Ｉ／Ｏ）インタフェースを含む。

【００３１】

アレイ３００への情報の記憶は、ディスク空間の全体的な論理的配置を定義する、物理的記憶ディスク２３０の集まりを含む１以上のストレージ「ボリューム」として実施されることが好ましい。各ボリュームは通常、必須ではないが、そのボリューム自体のファイルシステムに関連している。ボリューム／ファイルシステム内のディスクは通常、１以上のグループに編成され、各グループがＲＡＩＤ（Redundant Array of Independent(or Inexpensive) Disks）として動作している。ほとんどのＲＡＩＤ実施形態は、ＲＡＩＤグループ内の所定数の物理ディスクに「ストライプ状」にまたがるデータの冗長書き込みと、そのストライプ化されたデータに関するパリティ情報の適当な記憶とによって、データ記憶

の信頼性／完全性を向上させている。

【００３２】

本発明は、複数の連結されたサブアレイを有するストレージアレイ内の複数の記憶装置故障を効率的に訂正するための技術を含む。本発明の技術は、ストレージオペレーティングシステム６００のディスクストレージ層（図６に符号６２４で示す）で実施し、２重故障保護符号化方法を使用して、恰も各サブアレイが唯一単独で存在するかのようにして、各サブアレイに対して同様に対角パリティ集合を割り当てる。ストレージアレイの各サブアレイには、データ記憶装置（ディスク）の集合と、ブロックからなる行（例えば行パリティ集合）内の単一ディスクの故障を訂正するために用いる単一装置誤り訂正方法を用いて符号化されたパリティ値を記憶する局所パリティディスクとが含まれる。このアレイは、対角パリティを保持する一台の大域パリティディスクをさらに含む。

【００３３】

図３は、複数の連結サブアレイ３１０として編成されたストレージアレイ３００の略ブロック図であり、各サブアレイはデータディスクの集合（D_1、D_2）とパリティディスク（P_{R1}、P_{R2}）を含む。例として、各サブアレイ３１０は例えばＲＡＩＤ－４スタイルの集中パリティになるように構成され、ディスクアレイ［A_0、A_2、...、A_n］が所定数（例えば７台）のデータディスク３２０と１台の行パリティディスク３３０とを含む。各サブアレイの要素は、Ｃｋ（ｋ＝０...ｎ）で表すものとする。アレイ内の２台のディスクの同時故障からの復旧を可能にするため、アレイの各サブアレイ（及び行パリティディスク）について対角パリティを設けるのではなく、アレイ全体について１つの対角パリティを設ける。従って、アレイは、１つの対角パリティグループについて対角パリティを保持する大域パリティディスクP_D３５０をさらに含み、そのパリティはサブアレイ３１０の各々における同等の対角パリティ集合をディスクストレージ層でまとめて論理的に加算することにより計算される。サブアレイ内の２重故障は、アレイ全体に関するこのただ１つの大域対角パリティディスク３５０を用いて訂正することができる。従って、この新規の技術は、アレイ内の２台の記憶装置（ディスク）の同時故障からの有効な復旧を可能にするために必要なディスク数を削減するものである。

【００３４】

本発明によると、各サブアレイの対角パリティ集合に沿って対角パリティブロックが計算される。次いで、サブアレイの対応する対角パリティ集合の計算された対角パリティブロックは、例えば排他的論理和（ＸＯＲ）演算により論理的に結合され、１つの大域パリティディスク３５０上に対角パリティとして記憶される。その後、大域パリティディスク上に記憶された対角パリティから他のサブアレイの結合された対角パリティブロックを減算することにより、任意のサブアレイの計算された対角パリティブロックの内容を復元することができる。従って１つのサブアレイだけが２重故障の影響を受け、他のサブアレイは実質的に関係ない場合、局所パリティディスクと共に大域パリティディスクを用いることにより、サブアレイ内のいかなる２重故障も訂正することができる。

【００３５】

注意すべき点は、本発明で用いる２重故障保護符号化方法は単一装置誤り訂正方法に依存しないことである。さらに、単一ディスク故障からの復旧に用いる方法は、行志向であって各サブアレイにおけるブロックの行が独立したものである限り、すなわち復旧がブロックの他の行からの情報に依存することがない限り、何も制限がない（即ち、必ずしも「行パリティ」である必要はない）。この独立性が維持されるならば、これらの行の大きさは対角パリティを計算するために用いる行の大きさに関連した大きさである必要はない。

【００３６】

例示的実施形態において、抜けているディスクがゼロであるものと仮定することにより、各サブアレイ３１０は、適当な素数ｐに切り上げられた最大サブアレイと同じディスク数で構成されているかのように扱われる。各サブアレイはｐ－１行のブロックをさらに有する。ｍが任意の正の整数であるものとすると、この新規の装置故障訂正技術は、（ｍ＊ｐ＋１）×（ｐ－１）アレイのブロックを扱えることが好ましい。さらにサブアレイの連結

は「行－対角」２重故障保護符号化に基くのが好ましいが、従来のＥＶＥＮＯＤＤ（ＥＯ）符号化等の他の２重故障保護符号化方法を本発明に用いることもできる。

【００３７】

行－対角（Ｒ－Ｄ）符号化は、ディスクアレイにおける行パリティ及び対角パリティを用いた２重故障訂正復旧を提供するパリティ技術である。アレイ内の２台のディスクを完全にパリティ専用にし、残りのディスクがデータを保持する。任意の１台または任意の２台の同時ディスク故障の後でも、データを失うことなく、アレイの内容を完全に復元することができる。本発明と共に有利に用いられるＲ－Ｄパリティ技術の一例は、「Row-Diagonal Parity Technique for Enabling Efficient Recovery from Double Failures in a Storage Array.」と題した同時継続の共通所有者の米国特許出願第１０／０３５，６０７号明細書に開示されている。

【００３８】

図４は、Ｒ－Ｄパリティ符号化技術に従って編成されたディスクアレイ４００を示す略ブロック図である。ｎがアレイ内のディスク数、ｐを素数としてｎ＝ｐ＋１であるものと仮定する。最初のｎ－２台のディスク（Ｄ０～３）がデータを保持し、ディスクｎ－１（ＲＰ）が単一装置訂正アルゴリズムを用いて符号化された値、例えばデータディスクＤ０～Ｄ３についての行パリティを保持し、ディスクｎ（ＤＰ）が対角パリティを保持している。ディスクをブロックに分割し、ブロックをストライプに編成し、各ストライプがｎ－２（即ちｐ－１）行のブロックになるようにする。対角パリティディスクは、アレイの対角パリティ集合（「対角」）に沿って計算されたパリティ情報を記憶する。ストライプ内のブロックは、ｐ個の対角に編成され、各々の対角がデータディスク及び行パリティディスクの中からｐ－１ブロックを保持し、１つを除いてすべての対角が対角パリティディスク上にパリティブロックを有する。また、１ストライプ当たりｎ－１個の対角が存在する。

【００３９】

データブロック及び行パリティブロックには、各ブロックが１対角パリティ集合に属し、各行内で各ブロックが異なる対角パリティ集合に属するように、番号が付けられる。$D_{a,b}$及び$P_{a,b}$という記述は、データ（Ｄ）ブロック及びパリティ（Ｐ）ブロックの特定の行（ａ）及び対角（ｂ）パリティ計算への寄与をそれぞれ表している。すなわち、$D_{a,b}$という記述は、それらのブロックが行パリティａ及び対角パリティｂの計算に使用される行または対角に属していることを意味しており、$P_{a,b}$は、行パリティ集合ａについてのパリティを記憶するとともに、対角パリティ集合ｂにも寄与することを意味している。例えば、「^」がＸＯＲ演算を表すものとすると、$P_{0,8}=D_{0,4}\wedge D_{0,5}\wedge D_{0,6}\wedge D_{0,7}$となる。この記述には、特定の対角についての対角パリティの計算のために使用される行パリティブロックも含まれ、例えば$P_4=D_{0,4}\wedge D_{3,4}\wedge D_{2,4}\wedge P_{1,4}$である。対角パリティディスク上に記憶された対角パリティブロックの各々は、アレイ内の他のディスク（行パリティディスクを含む）のうちの一台を除いて全てのディスクからの寄与を表していることに注意して欲しい。例えば、対角パリティブロックP_4は、Ｄ０（$D_{0,4}$）、Ｄ２（$D_{3,4}$）、Ｄ３（$_{2,4}$）及びＲＰ（$P_{1,4}$）からの寄与を有するが、Ｄ１からの寄与は有しない。また、対角８（P_8）についての対角パリティは、計算もされず、対角パリティディスクＤＰに記憶もされないことに注意して欲しい。

【００４０】

具体的には、ディスクＤＰ上の対角パリティブロックは、それらのＸＯＲ計算を行なう際に行パリティブロックも含める。言い換えると、ディスクＤＰ上に記憶された対角パリティは、データディスクの内容だけでなく、行パリティディスクの内容にも従って計算される。さらに、対角パリティディスクは、１つを除いてストライプの対角の各々についてパリティブロックを保持する。アレイ４００に示すように対角パリティブロックを符号化することにより、このシステムは、対角パリティ（P_8）が無くても、任意の２台のディスク故障から復旧させることができる。これは、対角パリティディスクＤＰ上に記憶された対角パリティの計算に、行パリティブロックが要素として含まれた結果である。

【００４１】

Ｒ－Ｄパリティ技術の復旧（復元プロセス）態様は、故障によりサブアレイ内の２台のデータディスク（または１台のデータディスクと行パリティディスク）が同時に失われた場合に呼び出される。故障した２台のデータディスク（または１台のデータディスクと行パリティディスク）には任意の組み合わせがあるため、失われたデータを復元するのに行パリティを直ちに用いることはできないが、対角パリディだけは用いることができる。与えられたアレイの構造および編成（即ちストライプ長とストライプ深さが等しくない）の場合、各対角は、ディスクのうちの１台からはブロックを含まない（抜けている）。そのため、２台のデータディスクが失われた場合でも、１要素しか失わない対角が２つ存在する。即ち、２台の失われたディスクの各々について、そのディスクと交わらない対角が１つ存在するので、その対角中のブロックはそのディスクの故障により失われることがない。対角パリティブロックは１つの対角を除くすべての対角について対角パリティディスク上に記憶されているので、抜けているブロックのうちの少なくとも一方、通常２つの復元が、対角パリティを用いて開始される。

【００４２】

失われたブロックのうちの一方を復元した後、行パリティを用いてその行にある他方の失われたブロックを復元することにより、行の復元を計算することができる。他方のブロックを復元する場合、そのブロックがパリティの記憶された対角に属しているか否かについて判定を行なう。そのブロックがパリティのある対角に属している場合、その対角上にある他のディスクから、対角パリティを用いてその対角上にある他方の失われたブロックを復元することができる。すなわち、抜けている対角を除くすべての対角について、その対角上の１ブロックが復元できれば、他方のブロックも復元することが可能である。その後、その行パリティ集合のうちの他方の失われたブロックを復元する。しかしながら、ブロックがパリティの無い対角（即ち、抜けている対角）に属していた場合、すべてのブロックを復元し終わったか否かについて判定を行なう。まだ復元し終っていない場合、対角パリティに基づく第１の復元と、続く行パリティに基づく復元とからなるパターンを繰り返し、抜けている対角パリティ集合の計算に用いられる最後のデータブロックに達するまでそれを継続する。すべてのブロックを復元し終えると、復元処理は終了する。

【００４３】

図５は、Ｒ－Ｄ符号化に基くサブアレイ３１０の連結を有するストレージアレイ３００に適用されるような、新規の複数故障訂正技術を含むステップのシーケンスを示すフロー図である。このシーケンスはステップ５００で開始され、ステップ５０２へ進み、ここで行パリティディスク３３０を含むすべてのサブアレイＡ[０－ｎ]を連結し、Ｃｋ全体のデータディスクと行パリティディスク３３０の総数が素数であるようにする。ステップ５０４で、対角パリティディスク３５０を追加してアレイ３００を形成する。ステップ５０６で、各サブアレイの対角パリティをＲ－Ｄパリティ技術に従って計算し、ＸＯＲ演算を用いて各サブアレイの同等のパリティ計算結果を結合し、それらを対角パリティディスク上に記憶することにより、対角パリティディスク３５０の内容を符号化する。

【００４４】

ステップ５０８でアレイが故障する。その故障が単一ディスク故障である場合（ステップ５１０）、ステップ５１２で、その故障がサブアレイ内のディスクであるか否かについて判定を行なう。サブアレイ内のディスクであった場合、ステップ５１４で、そのサブアレイに関する局所行パリティを用いてその故障したデータディスクまたは行パリティディスクを復元する。そしてシーケンスはステップ５３２で終了する。単一故障がサブアレイのディスクでなかった場合、故障した大域対角パリティディスクをアレイ全体のすべてのサブアレイのすべてのディスク（データディスク及び行パリティディスク）を用いて復元する。この理由は、対角パリティ集合（即ち、対角）がディスクのアレイ全体にわたって分布しているからである。特に、故障した大域対角パリティディスク３５０上に記憶された対角パリティは、例えばＸＯＲ演算を用いてサブアレイ３１０における同等の対角パリティ集合を論理的に結合することによりステップ５１６で復元される。そしてシーケンスはステップ５３２で終了する。

【００４５】
故障が単一ディスク故障でない場合、ステップ５１８で、そのアレイ故障がサブアレイ内の２重故障であるか否かについて判定を行なう。サブアレイ内の２重故障でなかった場合、ステップ５２０で、故障のうちの１つに対角パリティディスクが含まれるか否かについて判定を行なう。対角パリティディスクが含まれない場合、各ディスク故障は異なるサブアレイで発生したデータディスク故障または行パリティディスク故障であるから、ステップ５２２で、各サブアレイの故障したディスクを局所行パリティを用いて復元する。そしてシーケンスはステップ５３２で終了する。

【００４６】
故障のうちの１つに大域対角パリティディスクが含まれる場合、ステップ５２４で、他の故障したディスクに行パリティディスクが含まれるか否かについて判定を行なう。行パリティディスクが含まれる場合、まず故障した行パリティディスクをそのサブアレイのデータディスクから復元し、次いでそれらのサブアレイにおける同等の対角パリティ集合から対角パリティディスクを復元することにより、行対角パリティディスクと対角パリティディスクの故障を復元する（ステップ５２６）。そしてシーケンスはステップ５３２で終了する。他の故障したディスクに行パリティディスクが含まれない場合、まずそのサブアレイに関する局所行パリティからデータディスクを復元し、次いでそれらのサブアレイにおける同等の対角パリティ集合から対角パリティディスクを復元することにより、データディスクと対角パリティディスクの故障を復元する（ステップ５２８）。そしてシーケンスはステップ５３２で終了する。

【００４７】
ステップ５３０で、サブアレイ内の２台のディスク故障（２重故障）をＲ－Ｄ復元プロセスを用いて大域的に復元する。この場合、ディスク内で発生する２つの故障は同じ行パリティによって保護されているので、復元のためには対角パリティが必要である。本発明によると、２重故障を有しているサブアレイが１つだけである限り、対角パリティから他のサブアレイの寄与を取り除くことができるので、データを復元することができる。具体的には、対角パリティディスクの内容から２重故障ではないサブアレイの対角パリティを減算し、次いでＲ－Ｄパリティ技術を利用して故障したサブアレイのデータ及び／又は行パリティを復元する。対角パリティディスクに対する条件はＲ－Ｄパリティ技術に関して説明したものと同じであるので、対角パリティディスクを用いて故障したサブアレイ内の少なくとも１つのデータブロックが復元される。そのブロックを復元した後、サブアレイ内の行パリティを用いて他の故障したディスクにおける対応するブロックを復元する。Ｒ－Ｄ復元プロセスに従ってこの処理を継続する。そしてシーケンスはステップ５３２で終了する。

【００４８】
本技術とＲ－Ｄ技術との差は、アレイ内の任意数のディスクを仮想的に行パリティディスクにできることであることに注目して欲しい。行パリティディスクはアレイ内のサブアレイを実質的に定義している。局所行パリティに基く復元には、サブアレイのデータディスク（即ち、行パリティ集合）しか必要ない。そのため本発明の訂正技術によると、サブアレイ内の２台のディスクの同時故障からの復旧が可能であるように構成されたアレイ３００において、より効率的（容易）な単一故障の復旧が可能になる。

【００４９】
さらに本発明は、既存のデータディスク及びパリティディスクに１台のパリティディスクを追加することにより、アレイ内の２重故障に対する保護を提供することができる。そしてＲ－Ｄパリティ復元アルゴリズムを用いることができる。

【００５０】
また、本明細書で説明する技術は、いかなる１つのサブアレイにおいても３以上の故障が存在せず２つの故障をもつサブアレイが２以上存在しない場合、及び、何らかのサブアレイに２つの故障が存在する場合であっても対角パリティディスクが故障していない場合には、アレイ３００内の３以上の故障を訂正することも可能であるということに注目して欲

しい。例えば、３つのサブアレイが存在し、各々のサブアレイが１以上のデータディスクと１つの行パリティディスクとを含むものと仮定する。本発明では、アレイ全体内の総数４台のディスク故障について、各サブアレイ内の単一（データまたは行パリティ）ディスク故障、及び、アレイ内のどこかで生じたもう１つのディスク故障からの復旧が可能である。１つのサブアレイ内の２台のディスク故障の場合、復元は、１つの要素しか失っていない対角パリティ集合を探すことから開始される。つまり、復元は、故障したディスクのうちの一方に現れない対角パリティ集合の対角パリティのうちの抜けているブロックから開始される。次いで、行パリティ集合内の他の抜けているブロックの復元が可能になり、抜けている対角パリティ集合の計算に用いられる最後のデータブロックに達するまで、行－対角復元手順を継続する。

【００５１】

有利なことに、本発明によると、アレイのサブアレイ内の２台の記憶装置の同時故障からの復旧が可能であるように構成されたアレイにおける単一故障の効率的な復旧も可能になる。異なるサブアレイにおける何らかのデータブロックの故障時には、本発明は、例えば局所行パリティなどの単一装置故障復旧方法を用いてデータブロックの復旧を可能にする。サブアレイ内の任意の２ブロックの故障時には、本発明は、局所行パリティと大域対角パリティとの組み合わせを用いて復旧を容易にする。つまり、２重故障を有するサブアレイが１つだけである限り、大域対角パリティ記憶装置の内容から他のサブアレイの対角パリティ寄与を取り除くことができるので、データを復旧させることができる。

【００５２】

図６は、本発明に有利に用いることができるストレージオペレーティングシステム６００を示す略ブロック図である。例示的実施形態では、このストレージオペレーティングシステムは、カリフォルニア州サニーベールにあるネットワークアプライアンス社から入手可能なＮｅｔＡｐｐ　Ｄａｔａ　ＯＮＴＡＰ（登録商標）オペレーティングシステムが好ましく、Ｗｒｉｔｅ　Ａｎｙｗｈｅｒｅ　Ｆｉｌｅ　Ｌａｙｏｕｔ（ＷＡＦＬ（登録商標））ファイルシステムを実施する。本明細書で用いる場合、「ストレージオペレーティングシステム」という用語は、一般にストレージシステムにおいてストレージ機能を実施することができるコンピュータ実行可能コードを指しており、例えばファイルシステムセマンティックを実施したり、データアクセスを管理したりする。その意味で、ＯＮＴＡＰソフトウェアは、マイクロカーネルとして実施されるそのようなストレージオペレーティングシステムの一例であり、ＷＡＦＬファイルシステムセマンティック及びデータアクセスの管理を実施するためのＷＡＦＬ層を含む。また、このストレージオペレーティングシステムは、ＵＮＩＸ（登録商標）やＷｉｎｄｏｗｓＮＴ（登録商標）等の汎用オペレーティングシステム上で動作するアプリケーションプログラムとして実施することもできるし、本明細書で説明するストレージアプリケーションのために構成されたストレージ機能または設定可能な機能を有する汎用オペレーティングシステムとして実施することもできる。

【００５３】

ストレージオペレーティングシステムは、ネットワークドライバ（例えばイーサネット（登録商標）ドライバ）のメディアアクセス層６１０を含む一連のソフトウェア層からなる。このネットワークオペレーティングシステムは、インターネットプロトコル（ＩＰ）層６１２、並びに、ＩＰ層がサポートするトランスポート手段であるＴＣＰ（Transport Control Protcol）層６１４及びＵＤＰ(User Datagram Protocol）層６１６等のネットワークプロトコル層をさらに含む。ファイルシステムプロトコル層は、マルチプロトコルアクセスを提供するため、ＣＩＦＳ（Common Internet File System）プロトコル６１８、ＮＦＳ（Network File System）プロトコル６２０、及び、ＨＴＴＰ（Hypertext Transfer Protocol）プロトコル６２２をサポートしている。さらに、オペレーティングシステム６００は、ＲＡＩＤプロトコル等のディスクストレージプロトコルを実施するディスクストレージ層６２４と、ＳＣＳＩ（Small Computer Systems Interface）プロトコル等のディスクアクセスプロトコルを実施するディスクドライバ層６２６とを含む。ディスクソフトウェア層とネットワークプロトコル層／ファイルシステムプロトコル層とを橋渡しするの

は、好ましくはＷＡＦＬファイルシステムを実施するＷＡＦＬ層６８０である。
【００５４】
ストレージシステムで受信したユーザ要求に対して、データストレージアクセスを実施するのに必要な上記のストレージオペレーティングシステム層を通るソフトウェア「パス」は、代替的にハードウェアで実施することもできる。すなわち、本発明の代替の実施形態では、このストレージアクセス要求データパス６５０は、ＦＰＧＡ（Field Programmable Gate Array）やＡＳＩＣ（Application Specific Integrated Cirquit）の中に実現された論理回路として実施することもできる。この種のハードウェアによる実施形態は、ユーザ要求に応答してシステム２２０によって提供されるサービスの能力を向上させることができる。さらに、本発明のさらに他の実施形態として、アダプタ２２８の処理要素は、プロセッサ２２２からストレージアクセス処理の負荷の一部または全部を低減させるように構成し、ストレージシステムによって提供されるサービスの性能を向上させることもできる。
【００５５】
本明細書で説明する様々な処理、アーキテクチャが、ハードウェアでもファームウェアでも、あるいはソフトウェアでも実施できることは、明らかである。例えば、本発明の一般的な実施形態は、マイクロプロセッサが組み込まれた汎用または専用のコンピュータ上で動作するソフトウェアコードを含む場合がある。しかしながら、本発明は、ＦＰＧＡ、ＡＳＩＣ、または、何らかの他のハードウェアあるいはソフトウェア実施形態で実施することも全く問題なく可能であり、場合によっては好ましい。当業者であれば、本明細書で説明する本発明のアルゴリズムは様々な技術的手段を用いて実施できることが分かるであろう。
【００５６】
本明細書で説明した例示的実施形態は、各サブアレイの局所パリティブロックがすべて同じディスク上に記憶される、集中パリティ配置に関するものである。本発明のさらに別の実施形態として、本発明の技術は、異なる行の集合のサブアレイ内で局所パリティブロックの位置がディスクごとにシフトされている、分散パリティ配置（例えばＲＡＩＤ－５）等の他のサブアレイ構成と共に用いることも可能である。しかしながら、本発明のスケーリング態様（即ち、将来的に既存のデータブロック及びパリティブロックを再編成することなくディスクをアレイに追加する能力）は、対角パリティ集合がゼロ値のブロックを有する「仮想」（不在）ディスクの存在を考慮しているため、集中パリティ技術にのみ適用される。分散パリティ配置を用いたこの種のスケーリングは極めて難しく、循環したパリティがこのような仮想ディスクに当たってしまう場合がある。
【００５７】
さらに、本発明は２個の記憶装置からｐ個の記憶装置までの範囲の大きさのサブアレイ上で動作する。つまり、本発明は、２個からｐ個の装置からなるサブアレイをｐ－１行で反復することにより、任意のサブアレイ内の２重故障保護を提供するとともに、ストレージアレイ全体における２重故障保護を提供する。ストレージアレイ全体についての大域対角パリティ装置から他のサブアレイの計算された対角パリティを取り除くことにより、任意の１つのサブアレイについての「サブアレイ」対角パリティ装置の内容を復元できるということが、その証拠である（１つの大域対角パリティ装置はサブアレイの同等のサブアレイ対角パリティ装置の追加であることに注意して欲しい）。本発明は、２重故障保護符号化方法を適用可能であるための制約に合致するストライプのブロック化と複数の装置が各サブアレイ内（対角パリティ装置以外）に必要であり、本明細書ではそれらをＲ－Ｄ（またはＥＯ）符号化されたアレイとして説明している。
【００５８】
ここまでストレージアレイ内の複数の記憶装置故障を効率的に訂正するための例示的実施形態を図示及び説明してきたが、本発明の思想と範囲の中で様々な改造及び修正が可能であるものと考えられる。例えば、代替の実施形態において、本発明は、フォワード・エラー・コレクション技術として通信の分野で用いることもでき、例えば待ち時間の長いリン

ク（衛星など）を介したデータのマルチキャスト配送などを可能にする。この実施形態では、データを、パケットや電子通信媒体（ネットワーク）を介した伝送に適したデータ単位等の記憶要素に分割して、ｐ番目のパケット毎に直前のｐ－１個のパケットの行パリティＸＯＲを保持するようにする。当業者であれば、本発明の原理に従ってパケットの他の編成及び構成を用いることも可能であることが分かるである。行パリティパケットは各サブグループ（集合）内の最大データパケットと少なくとも同じ大きさをもつ必要があり、対角パリティパケットは任意のサブグループ内の最大データパケットと少なくとも同じ大きさをもつ必要があることに注意して欲しい。また、パケットの任意のサブグループ内のパケット数と少なくとも同じ大きさをもつ最小の素数をｐとしたとき、対角パリティパケットの大きさはｐ－１ビットである。ｐの集合から１パケットが欠落した場合、そのパケットは行パリティから復元することができる。ｐの一集合から２パケットが欠落した場合、対角パリティを用いて復旧を行なうことができる。

【００５９】

上記の説明は、本発明の特定の実施形態に関して行なっている。しかしながら、説明した実施形態に対して他の様々な変更及び修正を行い、本発明の利点のうちのいくらかまたは全部を得ることも可能であることは明らかである。従って、付記した請求の範囲の目的は、かかる変更及び修正が本発明の真の思想及び範囲に入るようにすることである。

【図面の簡単な説明】

【図１】従来のＥＶＥＮＯＤＤパリティ配置に従って構成された従来技術のディスクアレイを示す略ブロック図である。

【図２】本発明に有利に用いられる、ストレージシステムを含む環境を示す略ブロック図である。

【図３】本発明に有利に用いられる、複数の連結されたサブアレイを含むストレージアレイを示す略ブロック図である。

【図４】行－対角（Ｒ－Ｄ）パリティ符号化技術に従って編成されたディスクアレイを示す略ブロック図である。

【図５】本発明に従ってＲ－Ｄ符号化に基くサブアレイの連結に適用される新規の装置故障訂正技術を含むステップのシーケンスを示すフロー図である。

【図６】本発明に有利に用いられるストレージオペレーティングシステムを示す略ブロック図である。

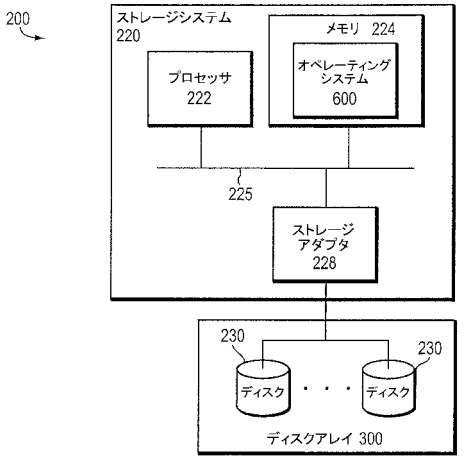
【図１】

100

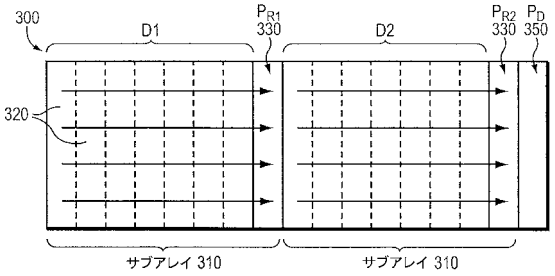
データディスク0	データディスク1	データディスク2	データディスク3	データディスク4	行パリティディスク	対角パリティディスク
D04	D05	D06	D07	D0X	P0	P4X
D15	D16	D17	D1X	D14	P1	P5X
D26	D27	D2X	D24	D25	P2	P6X
D37	D3X	D34	D35	D36	P3	P7X

（従来技術）

【図２】



【図３】



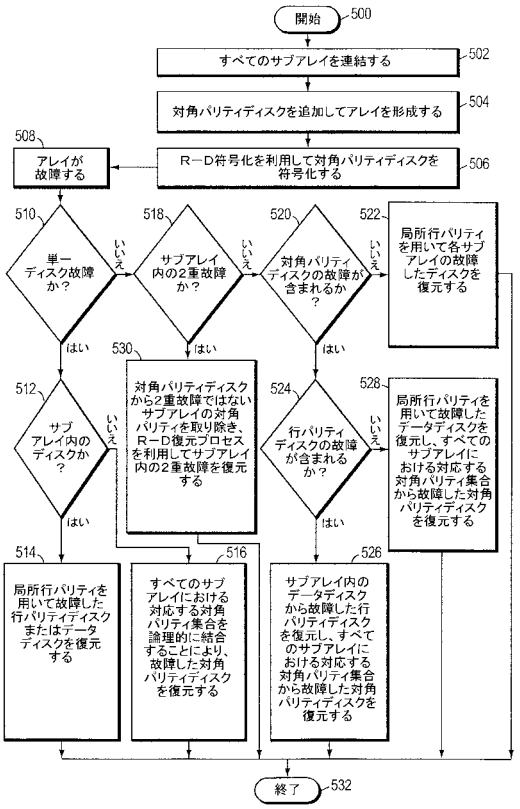
【図４】

400

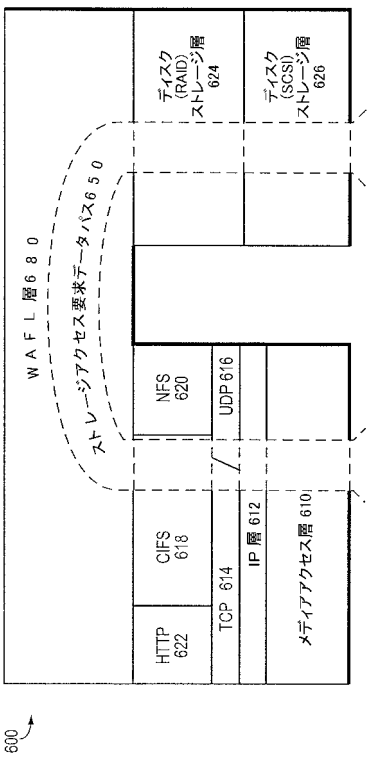
	D0	D1	D2	D3	RP	DP
N-2	$D_{0,4}$	$D_{0,5}$	$D_{0,6}$	$D_{0,7}$	$P_{0,8}$	P_4
	$D_{1,5}$	$D_{1,6}$	$D_{1,7}$	$D_{1,8}$	$P_{1,4}$	P_5
	$D_{2,6}$	$D_{2,7}$	$D_{2,8}$	$D_{2,4}$	$P_{2,5}$	P_6
	$D_{3,7}$	$D_{3,8}$	$D_{3,4}$	$D_{3,5}$	$P_{3,6}$	P_7

N-1

【図５】



【図６】



フロントページの続き

(72)発明者　ロバート・エム・イングリッシュ
アメリカ合衆国カリフォルニア州９４０２５，メンロパーク，イースト・クリーク・プレイス・４
(72)発明者　ピーター・エフ・コルベット
アメリカ合衆国マサチューセッツ州０２４２０，レキシントン，サマー・ストリート・３３

合議体
審判長　和田　志郎
審判官　清水　稔
審判官　丸山　高政

(56)参考文献　特開平７－２００１８７（ＪＰ，Ａ）
国際公開第９９／５９１５７（WO，Ａ１）
特表２００２－５１５６２０（ＪＰ，Ａ）
特開２００１－３２５７７３（ＪＰ，Ａ）
特表平１１－５０５６８５（ＪＰ，Ａ）
特開平７－２８７１０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F　3/06