(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2016-161970
(P2016-161970A)
(43)公開日　平成28年9月5日(2016.9.5)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 3/06 (2006.01)		G06F	3/06	306Z
		G06F	3/06	540
		G06F	3/06	305C
		G06F	3/06	302Z
		G06F	3/06	301J

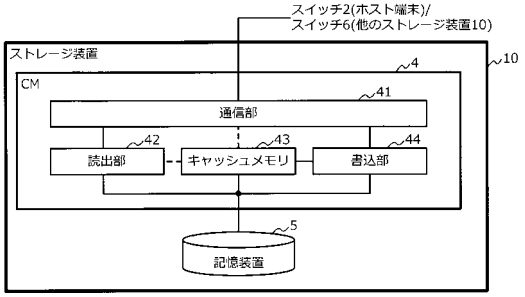
審査請求 未請求　請求項の数 8　ＯＬ　（全 30 頁）

(21)出願番号　特願2015-37096(P2015-37096)
(22)出願日　平成27年2月26日(2015.2.26)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100092978
弁理士　真田　有
(74)代理人　100112678
弁理士　山本　雅久
(72)発明者　塩沢　賢輔
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(54)【発明の名称】ストレージ装置、ストレージシステム、リカバリプログラム、及びリカバリ方法

(57)【要約】
【課題】複数の記憶装置をそなえるストレージシステムにおいて、故障した記憶装置のデータの復旧処理における他の記憶装置からの情報の読み出し性能を向上させる。
【解決手段】１以上の第１記憶装置５と、前記１以上の第１記憶装置５が記憶する複数の第１ブロック情報であって、故障した第２記憶装置５が記憶する複数の第２ブロック情報の復元に用いられる前記複数の第１ブロック情報を、前記１以上の第１記憶装置５から記憶領域のアドレス順に読み出す読出部４２と、前記読出部４２により前記１以上の第１記憶装置５から読み出し済の第１ブロック情報を、前記複数の第２ブロック情報を段階的に復元するために、前記複数の第２ブロック情報の復元先へ出力する出力部４１と、をそなえる。
【選択図】図４



【特許請求の範囲】
【請求項１】
１以上の第１記憶装置と、
前記１以上の第１記憶装置が記憶する複数の第１ブロック情報であって、故障した第２記憶装置が記憶する複数の第２ブロック情報の復元に用いられる前記複数の第１ブロック情報を、前記１以上の第１記憶装置から記憶領域のアドレス順に読み出す読出部と、
前記読出部により前記１以上の第１記憶装置から読み出し済の第１ブロック情報を、前記複数の第２ブロック情報を段階的に復元するために、前記複数の第２ブロック情報の復元先へ出力する出力部と、をそなえることを特徴とする、ストレージ装置。
【請求項２】
前記ストレージ装置とは異なる第１ストレージ装置、又は、前記出力部、から入力された第１ブロック情報に基づき、前記第２ブロック情報を段階的に復元する復元部をさらにそなえることを特徴とする、請求項１記載のストレージ装置。
【請求項３】
前記第１ストレージ装置又は前記出力部から入力される第１ブロック情報を保持する保持部をさらにそなえ、
前記復元部は、所定のタイミングで、前記保持部が保持する１以上の第１ブロック情報に基づき、前記第２ブロック情報を段階的に復元することを特徴とする、請求項２記載のストレージ装置。
【請求項４】
前記第２ブロック情報の復元先は、前記ストレージ装置とは異なる第２ストレージ装置にそなえられた第１記憶装置における空き記憶領域であり、
前記出力部は、前記読出部により前記第１記憶装置から読み出し済の第１ブロック情報を、前記第２ストレージ装置へ送信することを特徴とする、請求項１～３のいずれか１項記載のストレージ装置。
【請求項５】
前記第２ブロック情報の復元先は、前記故障した第２記憶装置の代替となる第３記憶装置における空き記憶領域であり、
前記出力部は、前記読出部により前記第１記憶装置から読み出し済の第１ブロック情報を、前記第３記憶装置へ送信することを特徴とする、請求項１～３のいずれか１項記載のストレージ装置。
【請求項６】
複数の記憶装置と、
前記複数の記憶装置のうちの複数の第１記憶装置の各々が記憶する複数の第１ブロック情報に基づき、前記複数の記憶装置のうちの故障した第２記憶装置が記憶する複数の第２ブロック情報を復元する１以上のストレージ装置と、をそなえ、
前記１以上のストレージ装置は、
複数の第１記憶装置の各々から、前記複数の第１ブロック情報を記憶領域のアドレス順に読み出し、
前記読み出しの処理において前記複数の第１記憶装置の各々から読み出し済の第１ブロック情報に基づき、前記複数の第２ブロック情報を段階的に復元する、
ことを特徴とする、ストレージシステム。
【請求項７】
コンピュータに、
１以上の第１記憶装置が記憶する複数の第１ブロック情報であって、故障した第２記憶装置が記憶する複数の第２ブロック情報の復元に用いられる前記複数の第１ブロック情報を、前記１以上の第１記憶装置から記憶領域のアドレス順に読み出し、
前記読み出しの処理において前記１以上の第１記憶装置から読み出し済の第１ブロック情報を、前記複数の第２ブロック情報を段階的に復元するために、前記複数の第２ブロック情報の復元先へ出力する、

処理を実行させることを特徴とする、リカバリプログラム。
【請求項８】
複数の記憶装置と、前記複数の記憶装置に対する制御を行なう１以上のストレージ装置とをそなえるストレージシステムにおけるリカバリ方法であって、
前記１以上のストレージ装置は、
前記複数の記憶装置のうちの複数の第１記憶装置の各々が記憶する複数の第１ブロック情報に基づき、前記複数の記憶装置のうちの故障した第２記憶装置が記憶する複数の第２ブロック情報を復元し、
前記復元の処理において、
前記１以上のストレージ装置により、
複数の第１記憶装置の各々から、前記複数の第１ブロック情報を記憶領域のアドレス順に読み出し、
前記読み出しの処理において前記複数の第１記憶装置の各々から読み出し済の第１ブロック情報に基づき、前記複数の第２ブロック情報を段階的に復元する、
ことを特徴とする、リカバリ方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ストレージ装置、ストレージシステム、リカバリプログラム、及びリカバリ方法に関する。
【背景技術】
【０００２】
複数のＨＤＤ（Hard Disk Drive）やＳＳＤ（Solid State Drive）等の記憶装置をそなえるストレージ装置では、種々の消失訂正技術により、記憶装置が故障した場合でもデータのリカバリを可能としている。
【０００３】
図２１は、ストレージ装置をそなえるストレージシステム１００におけるディスク故障からの復旧の手法の一例を示す図である。図２１に示すストレージシステム１００は、６つのＨＤＤ－１～ＨＤＤ－６に３つのストライプ－１～ストライプ－３が設定されたディスクグループをそなえ、このディスクグループでは消失訂正符号（Erasure Code）により符号化された情報が各ストライプに格納される。
【０００４】
例えば図２１の上段に示すように、ストライプ－２には、データが４つのＨＤＤ－２～ＨＤＤ－５に分散して格納され（２Ｄ１～２Ｄ４参照）、ストライプ－２のパリティがＨＤＤ－６に格納される（２Ｐ参照）。ここで、ストライプ－２におけるＨＤＤ－１は空きブロックであり、ＨＤＤ－２～ＨＤＤ－６のいずれかの故障時に代替ブロックとして用いられる。また、２Ｄ１とはストライプ－“２”のデータ“Ｄ”のうちの“１”番目のブロックを意味し、２Ｐとはストライプ－“２”のパリティ“Ｐ”のブロックを意味する。なお、以下の説明において、代替ブロックの“空き”の表記を省略する場合がある。
【０００５】
このようなストレージシステム１００においてＨＤＤ－５が故障した場合、図２１の下段に例示するように、ストレージ装置のコントローラモジュール（ＣＭ；Controller Module，図示省略）等の制御装置（以下、ＣＭと表記する）は、ＨＤＤ－５に格納されていたデータを復旧するために再構築（リビルド）を行なう。例えばＣＭは、ストライプ－１について、ＨＤＤ－１～ＨＤＤ－４から１Ｄ１～１Ｄ４のデータを取得し、１Ｄ１～１Ｄ４からパリティ計算を行なって１Ｐを生成し、生成した１ＰをＨＤＤ－６の代替ブロックに書き込む。他のストライプについても同様に、ＣＭは、ストライプ－２について２Ｄ１～２Ｄ３及び２Ｐのデータから２Ｄ４を生成し、生成した２Ｄ４をＨＤＤ－１の代替ブロックに書き込む。また、ＣＭは、ストライプ－３について３Ｄ１、３Ｄ２、３Ｄ４、及び３Ｐのデータから３Ｄ３を生成し、生成した３Ｄ３をＨＤＤ－２の代替ブロックに書き込

む。
【０００６】
関連する技術として、データの各ブロックを消失符号化して、グループ化されたストレージノードに分配し、ストレージノードに障害が発生した場合、未使用のノードに他のノードのデータを使用して故障ノードのデータの再構築を行なう技術が知られている（例えば、特許文献１参照）。
【先行技術文献】
【特許文献】
【０００７】
【特許文献１】特開２０１０－７９８８６号公報
【発明の概要】
【発明が解決しようとする課題】
【０００８】
ストレージシステム１００において、各ＨＤＤ上のブロック配置はＣＭ等による当該ＨＤＤの空き領域の管理に応じて変化する。
【０００９】
従って、図２２の下段に例示するように、ストライプ単位のデータ解放や再割り当てが繰り返されると、各ＨＤＤ内のブロックには、データ又はパリティの情報がストライプ順ではなくランダムな順序で格納されることになる。例えばＨＤＤ－３には、ＨＤＤの記憶領域の先頭から順に、２Ｄ２（ストライプ－２）、１Ｄ３（ストライプ－１）、３Ｄ１（ストライプ－３）のブロックが格納される。
【００１０】
このような図２２の下段に示す状態においてＨＤＤ－５が故障した場合を考える。この場合、各ＨＤＤでは、ストライプ－１～ストライプ－３の構成ブロックの配置がランダムとなっているため、図２３の下段に例示するように、ＣＭによるリビルドの際、ＨＤＤ上のアドレス順とは異なる順序によるＨＤＤへのアクセスが発生する。
【００１１】
例えばＨＤＤ－３では、ＣＭにより、最初にストライプ－１のリカバリのために記憶領域の中央付近のアドレスから１Ｄ３が読み出され、次いでストライプ－２のリカバリのために記憶領域の先頭付近のアドレスから２Ｄ２が読み出される。そして、最後にストライプ－３のリカバリのために記憶領域の末尾（最終アドレス）付近のアドレスから３Ｄ１が読み出される。
【００１２】
また、例えば代替ブロックを持つＨＤＤ－１では、ＣＭにより、最初にストライプ－１のリカバリのために記憶領域の末尾付近のアドレスから１Ｄ１が読み出され、次いでストライプ－２のリカバリのために他のＨＤＤの情報に基づき生成された２Ｄ４が記憶領域の中央付近のアドレスに書き込まれる。そして、最後にストライプ－３のリカバリのために記憶領域の先頭付近のアドレスから３Ｐが読み出される。
【００１３】
このように、ストレージシステム１００においてストライプ単位のデータ解放や再割り当てが繰り返されると、各ＨＤＤでは、リビルドの際にリカバリ対象のストライプの順にアクセスが行なわれるため、このアクセスはランダムアクセスとなる。
【００１４】
図２３の例では、ストレージシステム１００が３つのストライプを管理し、各ＨＤＤに最大３つのブロックが格納されるものとして説明したが、実際にはさらに多くのストライプが管理され、１つのＨＤＤに格納されるブロックも非常に多くなる。例えばＨＤＤが１００ＭＢ／ｓ程度の読み出し性能を持つＳＡＳ（Serial Attached SCSI（Small Computer System Interface））規格に対応した記憶装置であり、ブロックが４ＫＢ程度の小さいサイズである場合、図２３に示す例では各ＨＤＤのアクセスが１０ＭＢ／ｓ程度のランダムアクセスとなってしまう。

【００１５】
以上のように、ストレージ装置を１以上そなえるストレージシステムでは、ＨＤＤに格納されたブロックがストライプ順ではない場合、リビルドの際にストレージ装置においてＨＤＤの読み出しがランダム化され、性能劣化が生じてしまう。
【００１６】
なお、上述した課題は、上述の如く、消失訂正符号により符号化された情報を格納するストレージシステムにおいて生じ得るものである。このようなストレージシステムとしては、ＲＡＩＤ（Redundant Arrays of Inexpensive Disks）５と、ＲＡＩＤ５に対してストライプを追加のディスクにまで拡張するワイドストライプとを組み合わせた構成が挙げられる。また、ＲＡＩＤ５に代えて他のＲＡＩＤ技術（例えばＲＡＩＤ６）或いは複数のＲＡＩＤ技術の組み合わせが採用された構成や、他の消失訂正符号を用いたストレージシステムにおいても、上記課題は同様に生じ得る。
【００１７】
さらに、代替ブロックを持たない、例えば通常のＲＡＩＤ５やＲＡＩＤ６等を採用したストレージシステムにおいても、ＨＤＤに格納されたブロックがストライプ順にならない場合、上記課題は同様に生じ得る。
【００１８】
１つの側面では、本発明は、複数の記憶装置をそなえるストレージシステムにおいて、故障した記憶装置のデータの復旧処理における他の記憶装置からの情報の読み出し性能を向上させることを目的とする。
【課題を解決するための手段】
【００１９】
１つの態様では、本件のストレージ装置は、１以上の第１記憶装置と、前記１以上の第１記憶装置が記憶する複数の第１ブロック情報を、前記１以上の第１記憶装置から記憶領域のアドレス順に読み出す読出部と、をそなえる。前記第１ブロック情報は、故障した第２記憶装置が記憶する複数の第２ブロック情報の復元に用いられる情報である。また、前記ストレージ装置は、前記読出部により前記１以上の第１記憶装置から読み出し済の第１ブロック情報を、前記複数の第２ブロック情報を段階的に復元するために、前記複数の第２ブロック情報の復元先へ出力する出力部をさらにそなえる。
【発明の効果】
【００２０】
１つの側面では、複数の記憶装置をそなえるストレージシステムにおいて、故障した記憶装置のデータの復旧処理における他の記憶装置からの情報の読み出し性能を向上させることができる。
【図面の簡単な説明】
【００２１】
【図１】第１実施形態の一例としてのストレージシステムの構成例を示す図である。
【図２】図１に示すストレージシステムにおけるリビルド処理の一例を説明する図である。
【図３】図１に示すストレージシステムにおけるリビルド処理の一例を説明する図である。
【図４】図１に示すストレージ装置の機能構成例を示す図である。
【図５】図１に示すストレージシステムにおけるリビルド処理の動作例を説明する図である。
【図６】図１に示すストレージシステムにおけるリビルド処理の動作例を説明する図である。
【図７】図１に示すストレージシステムにおけるリビルド処理の動作例を説明する図である。
【図８】図１に示すストレージシステムにおけるリビルド処理の動作例を説明する図である。

【図９】図１に示すストレージシステムにおけるリビルド処理の動作例を説明する図である。
【図１０】図１に示すストレージシステムにおけるリビルド処理の動作例を説明する図である。
【図１１】第１実施形態に係るストレージシステムにおける全体の処理の一例を説明するフローチャートである。
【図１２】第１実施形態に係るストレージシステムにおけるリビルド処理の一例を説明するフローチャートである。
【図１３】第１実施形態に係るストレージシステムにおけるリビルド処理の一例を説明するフローチャートである。
【図１４】第２実施形態に係るストレージシステムにおけるリビルド処理の動作例を説明する図である。
【図１５】第２実施形態に係るストレージシステムの他の適用例におけるリビルド処理の動作を説明する図である。
【図１６】第２実施形態に係るストレージシステムにおける運用ストレージ装置のリビルド処理の一例を説明するフローチャートである。
【図１７】第２実施形態に係るストレージシステムにおける待機ストレージ装置のリビルド処理の一例を説明するフローチャートである。
【図１８】第２実施形態に係るストレージシステムにおける待機ストレージ装置のリビルド処理の他の例を説明するフローチャートである。
【図１９】第１及び第２実施形態に係るストレージ装置のハードウェア構成例を示す図である。
【図２０】第１及び第２実施形態に係るストレージシステムの他の構成例を示す図である。
【図２１】ストレージシステムにおけるディスク故障からの復旧の手法の一例を示す図である。
【図２２】ストレージシステムにおいてストライプ単位のデータ解放や再割り当てが繰り返された状態の一例を示す図である。
【図２３】図２２の下段に示す状態においてディスクが故障した場合のリカバリ処理におけるディスクアクセスの様子を示す図である。
【発明を実施するための形態】
【００２２】
以下、図面を参照して本発明の実施の形態を説明する。ただし、以下に説明する実施形態は、あくまでも例示であり、以下に明示しない種々の変形や技術の適用を排除する意図はない。すなわち、本実施形態を、その趣旨を逸脱しない範囲で種々変形して実施することができる。なお、以下の実施形態で用いる図面において、同一符号を付した部分は、特に断らない限り、同一若しくは同様の部分を表す。
【００２３】
〔１〕第１実施形態
〔１－１〕ストレージシステムの構成例
図１は第１実施形態の一例としてのストレージシステム１の構成例を示す図である。図１に示すように、ストレージシステム１は例示的にスイッチ２及び６、並びに１以上（図１では複数）のストレージ装置１０－１～１０－ｍ（ｍは自然数）をそなえることができる。なお、以下の説明においてストレージ装置１０－１～１０－ｍを区別しない場合には単にストレージ装置１０と表記する。
【００２４】
ストレージシステム１は、ユーザに対してストレージ装置１０の記憶領域を提供するものであり、例えばネットワーク３を介してユーザの使用するユーザ端末（ホスト装置）からストレージ装置１０へのアクセスが可能となっている。ストレージシステム１としては、例えば複数のストレージ装置１０（筐体）をそなえるクラスタ構成のストレージシステ

ムであってもよいし、単一のストレージ装置１０をそなえた構成であってもよい。図１の例では、ストレージシステム１は、ｍ個のストレージ装置１０をそなえ、分散アルゴリズムによりストレージ装置１０が相互に通信可能な分散ストレージシステムである。

【００２５】

スイッチ２は、ストレージ装置１０及びネットワーク３と接続され、ストレージシステム１を使用するユーザのユーザ端末とストレージ装置１０との間の通信（クライアント通信）の切り替え制御等を行なうものである。スイッチ６は、ストレージ装置１０と接続され、ストレージ装置１０間の通信（クラスタ内部通信）の切り替え制御等を行なうものである。

【００２６】

なお、ネットワーク３は、インターネットであってもよいし、ＬＡＮ（Local Area Network）又はＳＡＮ（Storage Area Network）等のイントラネットを形成するネットワークであってもよい。また、図１の例ではストレージシステム１にスイッチ２及び６がそれぞれ１つずつそなえられるものとしたが、これに限定されるものではない。例えば複数のストレージ装置１０が互いに異なる拠点にそなえられる場合には、各拠点にスイッチ２及びスイッチ６を設け、スイッチ２間及びスイッチ６間を、それぞれインターネットやＬＡＮ又はＳＡＮ等のイントラネットを形成するネットワーク等を介して相互に通信可能に接続してもよい。さらに、ストレージシステム１は、スイッチ２及びスイッチ６をクライアント通信及びクラスタ内部通信で共用のスイッチとして、いずれか一方のみそなえてもよい。

【００２７】

ストレージ装置１０は、それぞれ１以上（図１では１つ）のＣＭ４をそなえる。また、ストレージ装置１０－１は記憶装置５－１をそなえ、ストレージ装置１０－ｍは記憶装置５－ｎをそなえる。なお、以下の説明において記憶装置５－１～５－ｎを区別しない場合には単に記憶装置５と表記する。ストレージ装置１０は、図１の例ではそれぞれ１つの記憶装置５をそなえるものとしたが、複数の記憶装置５をそなえてもよい。

【００２８】

ＣＭ４は、スイッチ２を介したユーザ端末からの要求、並びにスイッチ６を介した他のストレージ装置１０（ＣＭ４）からの要求に応じて、記憶装置５の記憶領域に対する種々のアクセス制御を行なうコンピュータ（情報処理装置）の一例である。このアクセス制御には、記憶装置５の故障に伴うリビルド処理が含まれる。例えばＣＭ４は、リビルド処理において、他のストレージ装置１０のＣＭ４とともに、分散アルゴリズムに基づく協調動作を行なうことができる。

【００２９】

ＣＭ４は、例えばＣＰＵ（Central Processing Unit）４ａ、メモリ４ｂ、及びＩＦ（Interface）４ｃ～４ｅをそなえる。

【００３０】

ＣＰＵ４ａは、種々の制御や演算を行なう演算処理装置（プロセッサ）の一例である。ＣＰＵ４ａは、メモリ４ｂ、及びＩＦ４ｃ～４ｅとバスで相互に通信可能に接続され、メモリ４ｂ又は図示しないＲＯＭ（Read Only Memory）等に格納されたプログラムを実行することにより、ＣＭ４における種々の機能を実現することができる。

【００３１】

メモリ４ｂは、種々のデータやプログラムを格納する記憶装置である。第１実施形態に係るメモリ４ｂはさらに、後述するリビルド処理において記憶装置５から読み出した情報及び記憶装置５へ格納する（書き込む）情報を一時的に記憶するキャッシュメモリとして用いられる。なお、メモリ４ｂとしては、例えばＲＡＭ（Random Access Memory）等の揮発性メモリが挙げられる。

【００３２】

ＩＦ４ｃ～４ｅは、それぞれスイッチ２、記憶装置５、スイッチ６との間の接続及び通信の制御等を行なう通信インタフェースである。例えばＩＦ４ｃ及び４ｅはホストアダプ

タであり、ＩＦ４ｄはデバイス（ディスク）アダプタである。これらの通信インタフェース（アダプタ）としては、ＬＡＮ、ＳＡＮ、ＦＣ（Fibre Channel）、インフィニバンド（InfiniBand）等に準拠したアダプタが挙げられる。

【００３３】

記憶装置５は、種々のデータやプログラム等を格納するハードウェアである。記憶装置５としては、例えばＨＤＤ等の磁気ディスク装置や、ＳＳＤ等の半導体ドライブ装置等の各種記憶装置が挙げられる。

【００３４】

〔１－２〕ストレージシステムにおけるリビルド処理について

次に、第１実施形態に係るストレージシステム１におけるリビルド処理について、図２及び図３を参照して簡単に説明する。以下、前提として、記憶装置５がＨＤＤであるものとする。また、ストレージシステム１が６つのＨＤＤ－１～ＨＤＤ－６に３つのストライプ－１～ストライプ－３が設定されたディスクグループをそなえ、このディスクグループでは消失訂正符号により符号化された情報が各ストライプに格納されるものとする（図２３の上段参照）。

【００３５】

例えば図２３の上段に示す状態からＨＤＤ－５が故障した場合に、ストライプ－２の消失したブロック（消失ブロック）である２Ｄ４をＨＤＤ－１に復旧する場合を想定する。ストレージシステム１（ストレージ装置１０のＣＭ４）は、図２の上段に示すように、ストライプ順に関係なく、各ＨＤＤ上で物理アドレス順にシーケンシャルにブロックを読み出す。例えばＨＤＤ－３では、２Ｄ２、１Ｄ３、３Ｄ１の順にブロックが読み出される。なお、ＨＤＤ－２ではＨＤＤ－５の消失ブロック（３Ｄ３）の復旧及び格納前であるため、１Ｄ２、２Ｄ１の順にブロックが読み出される。

【００３６】

そして、ストレージシステム１（ＣＭ４）は、図２の中段に示すように、読み出したブロック（データ（Ｄ）ブロック及びパリティ（Ｐ）ブロック）順に、インクリメンタルに各ストライプを復旧する。以下、データブロック及びパリティブロックを区別しない場合には、これらを情報ブロックと表記する。

【００３７】

例えばＨＤＤ－３の記憶領域の先頭付近のアドレスに格納された２Ｄ２が最初に読み出されると、この２Ｄ２がＨＤＤ－１の新２Ｄ４のブロックに反映される（矢印（１）参照）。次いで、ＨＤＤ－６の記憶領域の中央付近のアドレスに格納された２Ｐが読み出されると、この２ＰがＨＤＤ－１の新２Ｄ４のブロックに反映される（矢印（２）参照）。次に、ＨＤＤ－２の記憶領域の中央付近のアドレスに格納された２Ｄ１が読み出されると、この２Ｄ１がＨＤＤ－１の新２Ｄ４のブロックに反映される（矢印（３）参照）。そして、ＨＤＤ－４の記憶領域の中央付近のアドレスに格納された２Ｄ３が読み出されると、この２Ｄ３がＨＤＤ－１の新２Ｄ４のブロックに反映される（矢印（４）参照）。

【００３８】

なお、図２にはストライプ－２の消失ブロックに着目して新２Ｄ４の復旧について説明したが、ストライプ－１及びストライプ－３の消失ブロックについて新１Ｐ及び新３Ｄ３を復旧する処理も、各ＨＤＤからのシーケンシャルな情報ブロックの読み出しの過程で順次実施される。

【００３９】

以上のように、第１実施形態に係るストレージシステム１は、読み出したブロックを用いて段階的に消失ブロックを復旧することができる。これにより、ストレージシステム１では、記憶装置の故障によるリビルド処理において、ＣＭ４は各記憶装置からシーケンシャルに情報ブロックを読み出すことができるため、正常な記憶装置からの情報の読み出し性能を向上することができる。従って、図２３に示す各記憶装置からの完全なランダムアクセスにより情報を読み出す手法と比較して、最大で１０倍以上のスループットとすることができ、リカバリ性能を大幅に向上させることができる。

【００４０】
ここで、ストレージシステム１が段階的にストライプを復旧することのできる理由を説明する。
【００４１】
ストレージシステム１では、各ストライプの情報ブロックが消失訂正符号により符号化されている。消失訂正符号により符号化された情報ブロックでは、消失ブロックの情報を、或るストライプにおける消失ブロック以外の正常な情報ブロックから算出（復旧）することができる。この算出手法としては、パリティが１つの場合、例えば正常な情報ブロックの排他的論理和（ＸＯＲ）を算出するといった手法が挙げられる。
【００４２】
ＸＯＲ演算は、可換（Commutative）演算である。このため、ストレージシステム１は、各ＨＤＤからシーケンシャルに読み出した情報ブロックを用いて段階的に（インクリメンタルに）各ストライプを復旧することができるのである。以下の説明では、ＸＯＲ演算を示す演算子として“＋”を用い、図面では演算子として“＋”を丸で囲んだ記号を用いる。
【００４３】
例えば図２の下段に示すように、対比例としての図２３の演算では、消失ブロックを復旧するための全ての情報ブロックが揃ってから、新２Ｄ４＝２Ｄ１＋２Ｄ２＋２Ｄ３＋２Ｐの演算が順序通りに行なわれる。一方、第１実施形態に係るストレージシステム１は、図２の矢印（１）及び（２）で読み出した２Ｄ２及び２Ｐの演算を行ない、その演算結果と図２の矢印（３）で読み出した２Ｄ１との演算を行なう。そして、ストレージシステム１は、最後にその演算結果と図２の矢印（４）で読み出した２Ｄ３との演算を行なう。
【００４４】
なお、消失ブロックの情報を算出（復旧）する算出手法としては、上述したＸＯＲ演算に限定されるものではない。例えばストライプごとにパリティが２つ以上含まれる場合には、ＸＯＲ演算を用いたリードソロモン（ＲＳ；Reed-Solomon）符号ベースの演算が行なわれてもよい。ストレージシステム１では、このようなＲＳ符号を消失訂正符号として用いて符号化された情報が各ストライプに格納されている場合でも、段階的に消失ブロック（ストライプ）を復旧することができる。消失ブロックの情報を算出（復旧）する算出手法としては、消失訂正符号がガロア体（有限体）を用いる可換演算の可能な符号であれば、上述したもの以外の種々の手法が用いられてよい。
【００４５】
説明の簡略化のため、以下の説明では、消失ブロックの情報の算出手法としてＸＯＲ演算が用いられるものとする。
【００４６】
ところで、ＣＭ４は、各記憶装置５から情報ブロックを読み出す都度、当該情報ブロックのストライプにおける復旧先の代替ブロック（リカバリ対象ブロック）へＸＯＲ演算等による反映を行なってよい。しかし、当該代替ブロックを有する記憶装置５においてもシーケンシャルリードが行なわれているため、当該代替ブロックを有する記憶装置５では、シーケンシャルリードの最中に代替ブロックへの書き込みアクセスが頻発して読み出し性能が劣化し、リビルド処理の性能低下が生じることがある。
【００４７】
そこで、ＣＭ４（又は記憶装置５）は、図３に示すように、或るストライプの消失ブロックを復旧する場合、当該消失ブロックの復旧に用いる情報ブロックをキャッシュメモリ等に一時的に保持しておくことができる。
【００４８】
例えばＣＭ４は、図３の矢印（１）に示すように、ＨＤＤ－３から２Ｄ２を読み出し、ＨＤＤ－１の新２Ｄ４のための書込キャッシュに反映する。このときストライプ－２の２Ｄ４復旧の進捗は、２Ｄ４として２Ｄ２が設定された状態である。
【００４９】

また、ＣＭ４は、図３の矢印（２）に示すように、ＨＤＤ－６から２Ｐを読み出し、書込キャッシュに反映する。このときストライプ－２の２Ｄ４復旧の進捗は、２Ｄ４として、２Ｄ４に設定された２Ｄ２と２ＰとのＸＯＲ演算結果が設定された状態である。

【００５０】

なお、このタイミングで書込キャッシュの容量が逼迫した場合、ＣＭ４は、書込キャッシュ内の情報ブロック（２Ｄ２＋２Ｐ）をＨＤＤ－１の代替ブロックに書き込む（フラッシュする）。

【００５１】

さらに、ＣＭ４は、図３の矢印（３）に示すように、ＨＤＤ－２から２Ｄ１を読み出し、ＨＤＤ－１の代替ブロック（新２Ｄ４）の情報を書込キャッシュにリロードする。そして、ＣＭ４は、２Ｄ１を書込キャッシュに反映する。このときストライプ－２の２Ｄ４復旧の進捗は、２Ｄ４として、書込キャッシュにリロードされた２Ｄ２＋２Ｐと２Ｄ１とのＸＯＲ演算結果が設定された状態である。

【００５２】

また、ＣＭ４は、図３の矢印（４）に示すように、ＨＤＤ－４から２Ｄ３を読み出し、書込キャッシュに反映する。このときストライプ－２の２Ｄ４復旧の進捗は、２Ｄ４として、２Ｄ４に設定された（２Ｄ２＋２Ｐ）＋２Ｄ１と２Ｄ３とのＸＯＲ演算結果が設定された状態である。

【００５３】

なお、このタイミングで書込キャッシュの容量が逼迫した場合、ＣＭ４は、書込キャッシュ内の情報ブロック（（２Ｄ２＋２Ｐ）＋２Ｄ１）＋２Ｄ３をＨＤＤ－１の代替ブロックに書き込む（フラッシュする）。

【００５４】

以上により書込キャッシュを用いた段階的なリビルド処理が行なわれる。なお、ストライプ－１及びストライプ－３の消失ブロックについて新１Ｐ及び新３Ｄ３を復旧する処理も、ＨＤＤ－２及びＨＤＤ－６に対応する書込キャッシュにおいて、各ＨＤＤからのシーケンシャルな情報ブロックの読み出しの過程で順次実施される。

【００５５】

各記憶装置５についてシーケンシャルに情報ブロックを読み込む場合、できるだけ多くの情報ブロックを保持できるように、大容量のキャッシュメモリが用いられることが好ましい。しかし、ＣＭ４（又は記憶装置５）の各々に大容量のキャッシュメモリを搭載することは、コスト増加の観点から難しい場合がある。

【００５６】

これに対し、上述したストレージシステム１によれば、図３に示すように、書込キャッシュの容量が逼迫した等の場合に、書込キャッシュに保持された情報ブロックをまとめてＸＯＲ演算して代替ブロックに書き込むことができる。そして、ＣＭ４は、容量の空いた書込キャッシュに情報ブロックを蓄積していき、再び書込キャッシュの容量が逼迫した等の場合に、代替ブロックから情報を再読み出しし、書込キャッシュに格納された情報ブロックと再読み出しした情報とをＸＯＲ演算して代替ブロックに書き込むのである。

【００５７】

このように、ストレージシステム１では、図２に示すように各ストライプをインクリメンタルに復旧することができるため、消失ブロックの復旧データである全ての情報ブロックが揃うまで書込キャッシュに情報ブロックを溜め込まなくてよい。これにより、ストレージシステム１は、大容量のキャッシュメモリをそなえなくてもよく、コスト増加を抑制することができる。

【００５８】

〔１－３〕ストレージ装置の構成例

次に、図４を参照してストレージ装置１０の構成例について説明する。ストレージ装置１０（ＣＭ４）は、複数のストレージ装置１０のＣＭ４と協働して、複数の記憶装置５に対する各種制御を行なうことができる。この制御には、ユーザ端末からの書込要求に応じ

て書込データのパリティ演算を行ない、複数の情報ブロックを生成してストライプとして各記憶装置５に分散させる制御が含まれる。また、この制御には、ユーザ端末からの読出要求に応じてストライプから情報ブロックを取得して読出データを構築し、ユーザ端末へ出力する制御も含まれる。

【００５９】

また、例えばＣＭ４は、ユーザ端末からの要求に応じてアクセス対象のストライプに対応する記憶装置５の情報ブロックへのアクセスを行なったり、記憶装置５の故障を検出した場合に他のストレージ装置１０のＣＭ４へ通知を行なってリビルド処理を実行することができる。これらの制御は、既知の種々の手法により行なうことが可能であり、その詳細な説明は省略する。

【００６０】

さらに、第１実施形態に係るストレージ装置１０（ＣＭ４）は、リビルド処理において図２及び図３に示すような動作を実現するため、図４に示すように、例示的に通信部４１、読出部４２、キャッシュメモリ４３、及び書込部４４をそなえることができる。

【００６１】

通信部４１は、他のＣＭ４との間で通信を行なうものであり、例えばリビルド処理に関する制御情報や情報ブロック等の種々の情報を他のＣＭ４との間で送受信する。例えば通信部４１は、図１に示すＣＰＵ４ａ、ＩＦ４ｃ、及びＩＦ４ｅの少なくとも一部の機能により実現することができる。

【００６２】

例えば記憶装置５（第２記憶装置）の故障を検出したＣＭ４は、故障した記憶装置５に格納されていた消失ブロック（第２ブロック情報）に関する情報を通信部４１により他のＣＭ４に通知する。この通知を受信したＣＭ４は、自装置１０がそなえる記憶装置５について、消失ブロックを復旧させる代替ブロックの有無や、消失ブロックと同じストライプの情報ブロックの有無等を判断して、消失ブロックのストライプごとに復旧先の記憶装置５を決定する。例えば記憶装置５に代替ブロックが有り、或る消失ブロックと同じストライプの情報ブロックが無い場合、ＣＭ４は、自装置１０の記憶装置５が当該消失ブロックのストライプの復旧先であることを通信部４１により他のＣＭ４に通知する。

【００６３】

読出部４２は、リビルド処理において、自装置１０がそなえる記憶装置５（第１記憶装置）に格納された情報ブロック（第１ブロック情報）を記憶装置５の物理アドレスの先頭からシーケンシャルに読み出し、読み出した情報ブロックを順次通信部４１に渡す。なお、自装置１０が或るストライプの復旧先（復元先）の記憶装置５をそなえる場合、読出部４２は、シーケンシャルに読み出す過程で当該情報ブロックの読み出しをスキップしてよい。

【００６４】

このように、読出部４２は、１以上の第１記憶装置５が記憶する複数の第１ブロック情報であって、故障した第２記憶装置５が記憶する複数の第２ブロック情報の復元に用いられる複数の第１ブロック情報を、１以上の第１記憶装置５から記憶領域のアドレス順に読み出すものであるといえる。

【００６５】

なお、通信部４１は、読出部４２が読み出した情報ブロックを、当該情報ブロックのストライプの復旧先であるストレージ装置１０（ＣＭ４）へ送信（転送）し、復旧先のＣＭ４は、他のＣＭ４から受信した情報ブロックを書込部４４に出力する。このとき復旧先のＣＭ４は、書込部４４によりキャッシュメモリ４３の使用量を監視し、容量が逼迫した場合、通信部４１により他のＣＭ４に対して容量が逼迫したことを示す通知（或いは情報ブロックの送信を抑止させる通知）を行なうことができる。この通知を受信したＣＭ４（通信部４１）は、読出部４２に対して復旧先のＣＭ４に対応する情報ブロックの読み出しを中止させてもよいし、読出部４２に読み出された当該情報ブロックをキャッシュメモリ４３に一時的に格納（退避）してもよい（図４の破線参照）。

【００６６】
　また、復旧先のＣＭ４は、キャッシュメモリ４３が使用可能になった場合、通信部４１により他のＣＭ４に対して容量が確保できたことを示す通知（或いは情報ブロックの送信を再開させる通知）を行なうことができる。この通知を受信したＣＭ４（通信部４１）は、読出部４２に対して復旧先のＣＭ４に対応する情報ブロックを読み出させてもよいし、キャッシュメモリ４３に格納した当該情報ブロックを復旧先のＣＭ４へ送信してもよい（図４の破線参照）。
【００６７】
　このように、通信部４１は、読出部４２により１以上の第１記憶装置５から読み出し済の第１ブロック情報を、複数の第２ブロック情報を段階的に復元するために、複数の第２ブロック情報の復元先へ出力する出力部の一例であるといえる。
【００６８】
　キャッシュメモリ４３は、図３に示す書込キャッシュの一例であり、例えば図１に示すメモリ４ｂの少なくとも一部の記憶領域を用いることにより実現することができる。キャッシュメモリ４３は、自装置１０が復旧先である場合に、復旧に用いる情報ブロックが格納される記憶領域である。また、上述のように、読出部４２が読み出した情報ブロックを復旧先のＣＭ４へ送信できない場合、当該情報ブロックの退避用の記憶領域として用いられてもよい。
【００６９】
　このように、キャッシュメモリ４３は、他の第１ストレージ装置１０から入力される第１ブロック情報を保持する保持部の一例であるといえる。
【００７０】
　書込部４４は、通信部４１から入力された情報ブロックをキャッシュメモリ４３へ書き込む。また、書込部４４は、例えば定期的に、キャッシュメモリ４３の使用量を監視し、使用量が閾値を超えた（容量が逼迫した）場合、その旨を通信部４１へ通知する。このとき書込部４４は、記憶装置５の代替ブロックに情報ブロックが格納されているか否かを判断する。なお、閾値は、情報ブロックのサイズやキャッシュメモリ４３の容量に応じて予め設定されるものであり、閾値として例えばキャッシュメモリ４３の記憶領域のサイズの８０％～９０％等の値を設定することができる。
【００７１】
　代替ブロックに情報ブロックが格納されていない場合、書込部４４は、キャッシュメモリ４３に格納された複数の情報ブロックについてＸＯＲ演算を行ない、演算結果を記憶装置５の代替ブロック（空き記憶領域）に書き込み、キャッシュメモリ４３をクリアする。一方、代替ブロックに情報ブロックが格納されている場合、書込部４４は、代替ブロックから情報ブロックを読み出し、読み出した情報ブロックと、キャッシュメモリ４３に格納された複数の情報ブロックと、についてＸＯＲ演算を行ない、演算結果を記憶装置５の代替ブロックに書き込み、キャッシュメモリ４３をクリアする。なお、書込部４４は、キャッシュメモリ４３の使用量が閾値以下となった（使用可能になった）場合、その旨を通信部４１へ通知する。
【００７２】
　なお、書込部４４によるキャッシュメモリ４３の容量の監視において、使用量が閾値を超えたか否かを判断する代わりに、キャッシュメモリ４３の残容量を監視し、残容量が閾値以下となったか否かを判断してもよい。或いは、情報ブロックが一定（ブロック単位の）サイズであるため、書込部４４は、キャッシュメモリ４３に格納した情報ブロックの数をカウントし、情報ブロックの数が所定数以上となった場合にキャッシュメモリ４３の容量が逼迫したと判断してもよい。
【００７３】
　また、書込部４４は、例えば受信した（又はキャッシュメモリ４３に格納した）情報ブロックの数をカウントし、カウント値が閾値に達した場合に、当該ストライプの復旧（リカバリ）処理が完了したと判断することができる。なお、カウントする数としては、これ

に限定されるものではなく、代替ブロックに反映した情報ブロックの数であってもよいし、ＸＯＲ演算を行なった回数であってもよい。また、情報ブロックの数をカウントする場合、閾値を「ストライプに含まれる情報ブロックの数」－「当該ストライプに含まれる消失ブロックの数」とすることができる。或いは、ＸＯＲ演算を行なった回数をカウントする場合、閾値を「ストライプに含まれる情報ブロックの数」－「当該ストライプに含まれる消失ブロックの数」－１としてもよい。

【００７４】

このように、書込部４４は、他の第１ストレージ装置１０から入力された第１ブロック情報に基づき、第２ブロック情報を段階的に復元する復元部の一例であるといえる。この復元部の一例としての書込部４４は、入力された第１ブロック情報及び第２ブロック情報の復元先が記憶する情報を用いて、消失訂正符号に基づく演算を行ない、演算結果を第２ブロック情報の復元先へ書き込むのである。また、この復元部の一例としての書込部４４は、所定のタイミングで、キャッシュメモリ４３が保持する１以上の第１ブロック情報に基づき、第２ブロック情報を段階的に復元するのである。

【００７５】

〔１－４〕ストレージシステムにおけるリビルド処理の動作説明

次に、図５～図１０を参照して、ストレージシステム１におけるリビルド処理の動作をストレージ装置１０間の通信に着目して説明する。以下、ストレージシステム１が６台のストレージ装置１０をそなえるものとし、便宜上、これらのストレージ装置１０をノード－１～ノード－６と表記する。また、ノード－５が故障し、ストレージシステム１がノード５のＨＤＤ－５に格納された２Ｄ４、１Ｐ、３Ｄ３の３ブロックのリビルド処理を行なうものとする。

【００７６】

図５の上段に示すように、各ノードのＣＭ４は、ＨＤＤの先頭ブロックを読出部４２により読み出し、読み出した先頭ブロック（情報ブロック）を通信部４１により復旧先の代替ブロック（リカバリ対象ブロック）を有するノードへ転送する。情報ブロックを受信したノードは、キャッシュメモリ４３に格納する。

【００７７】

図５の例では、ノード－１が２Ｄ４（ストライプ－２）のリカバリ対象ブロックを有し、ノード－２が３Ｄ３（ストライプ－３）のリカバリ対象ブロックを有し、ノード－６が１Ｐ（ストライプ－１）のリカバリ対象ブロックを有している。この場合、ノード－２及びノード－４は、それぞれＨＤＤ－２の先頭ブロックの１Ｄ２及びＨＤＤ－４の先頭ブロックの１Ｄ４をノード－６に転送し、ノード－３は、ＨＤＤ－３の先頭ブロックの２Ｄ２をノード－１に転送する。また、ノード－１及びノード－６は、それぞれＨＤＤ－１の先頭ブロックの３Ｐ及びＨＤＤ－６の先頭ブロックの３Ｄ４をノード－２に転送する。

【００７８】

次いで、図５の下段に示すように、ノード－２及びノード－６のＣＭ４は、書込部４４により、キャッシュメモリ４３の容量が枯渇したと判断して、ノード－２及びノード－６のそれぞれのリカバリ対象ブロックについて部分的なリカバリを実施する。このときノード－２及びノード－６のリカバリ対象ブロックに書き込まれる情報である３Ｄ３’及び１Ｐ’は、それぞれ３Ｄ３’＝３Ｐ＋３Ｄ４、１Ｐ’＝１Ｄ２＋１Ｄ４となる。一例として、図６に示すように、ストライプ番号とインデックス番号（パリティを除く）をそれぞれバイナリ化して、３Ｐ＝０００１００、３Ｄ４＝０１１１００、１Ｄ２＝００１０１０、１Ｄ４＝００１１００とした場合、３Ｄ３’及び１Ｐ’はそれぞれ以下の値となる。

【００７９】

３Ｄ３’＝３Ｐ　＋３Ｄ４＝０００１００＋０１１１００＝０１１０００

１Ｐ’　＝１Ｄ２＋１Ｄ４＝００１０１０＋００１１００＝０００１１０

【００８０】

次に、図７の上段に示すように、各ノードのＣＭ４は、ＨＤＤの２ブロック目（２ブロック目がリカバリ対象ブロックであれば３ブロック目）を読出部４２により読み出し、読

み出した情報ブロックを通信部４１によりリカバリ対象ブロックを有するノードへ転送する。情報ブロックを受信したノードは、キャッシュメモリ４３に格納する。

【００８１】

図７の例では、ノード－１及びノード－３は、それぞれＨＤＤ－１の３ブロック目の１Ｄ１及びＨＤＤ－３の２ブロック目の１Ｄ３をノード－６に転送し、ノード－２は、ＨＤＤ－２の３ブロック目の２Ｄ１をノード－１に転送する。また、ノード－４は、ＨＤＤ－４の２ブロック目の３Ｄ２をノード－２に転送する。なお、ノード－６は、ＨＤＤ－６の２ブロック目の２Ｐを転送する前にノード－１のキャッシュメモリ４３の容量が逼迫したため、ノード－１からの通知により転送を抑止（保留）している。

【００８２】

次いで、図７の下段に示すように、ノード－１及びノード－６のＣＭ４は、書込部４４により、キャッシュメモリ４３の容量が枯渇したと判断して、ノード－１及びノード－６のそれぞれのリカバリ対象ブロックについて部分的なリカバリを実施する。このときノード－１及びノード－６のリカバリ対象ブロックに書き込まれる情報である２Ｄ４’及び１Ｐ”は、それぞれ２Ｄ４’＝２Ｄ２＋２Ｄ１、１Ｐ”＝１Ｐ’＋１Ｄ１＋１Ｄ３となる。なお、ノード－６の書込部４４は、リカバリ対象ブロックから１Ｐ’を読み出してから、読み出した１Ｐ’とキャッシュメモリ４３内の１Ｄ１及び１Ｄ３とのＸＯＲ演算を行なう。

【００８３】

一例として、図８に示すように、２Ｄ２＝０１００１０、２Ｄ１＝０１０００１、１Ｄ１＝００１００１、１Ｄ３＝００１０１１とした場合、２Ｄ４’及び１Ｐ”はそれぞれ以下の値となる。

【００８４】

２Ｄ４’＝２Ｄ２＋２Ｄ１　　＝０１００１０＋０１０００１

　　＝０００１１

１Ｐ”　＝１Ｐ’＋１Ｄ１＋１Ｄ３＝０００１１０＋００１００１＋００１０１１

　　＝０００１００

【００８５】

ここで、１Ｐ”については、ストライプ－１における消失ブロック（消失パリティ）以外の情報ブロックのＸＯＲ演算が完了しているため、ストライプ－１のリカバリが完了する。また、図８の下段に示すように、１Ｐ”の値（０００１００）がノード－５のＨＤＤ－５における１Ｐ（０００１００）と一致していることがわかる。

【００８６】

そして、図９の上段に示すように、各ノードのＣＭ４は、ＨＤＤの未読み出しの情報ブロックを読出部４２により読み出し、読み出した情報ブロックを通信部４１によりリカバリ対象ブロックを有するノードへ転送する。情報ブロックを受信したノードは、キャッシュメモリ４３に格納する。

【００８７】

図９の例では、ノード－４及びノード－６は、それぞれＨＤＤ－４の３ブロック目の２Ｄ３及びＨＤＤ－６の２ブロック目の２Ｐをノード－１に転送し、ノード－３は、ＨＤＤ－３の３ブロック目の３Ｄ１をノード－２に転送する。

【００８８】

次いで、図９の下段に示すように、ノード－１及びノード－２のＣＭ４は、書込部４４により、キャッシュメモリ４３の容量が枯渇したと判断して、ノード－１及びノード－２のそれぞれのリカバリ対象ブロックについて部分的なリカバリを実施する。このときノード－１及びノード－２のリカバリ対象ブロックに書き込まれる情報である２Ｄ４”及び３Ｄ３”は、それぞれ２Ｄ４”＝２Ｄ４’＋２Ｐ＋２Ｄ３、３Ｄ３”＝３Ｄ３’＋３Ｄ２＋３Ｄ１となる。なお、ノード－１の書込部４４は、リカバリ対象ブロックから２Ｄ４’を読み出してから、読み出した２Ｄ４’とキャッシュメモリ４３内の２Ｐ及び２Ｄ３とのＸＯＲ演算を行なう。また、ノード－２の書込部４４は、リカバリ対象ブロックから３Ｄ３

’を読み出してから、読み出した３Ｄ３’とキャッシュメモリ４３内の３Ｄ２及び３Ｄ１とのＸＯＲ演算を行なう。

【００８９】

一例として、図１０に示すように、２Ｐ＝０００１００、２Ｄ３＝０１００１１、３Ｄ２＝０１１０１０、３Ｄ１＝０１１００１とした場合、２Ｄ４”及び３Ｄ３”はそれぞれ以下の値となる。

【００９０】

２Ｄ４”＝２Ｄ４’＋２Ｐ　＋２Ｄ３＝０００１１１＋０００１００＋０１００１１
＝０１０１００

３Ｄ３”＝３Ｄ３’＋３Ｄ２＋３Ｄ１＝０１１０００＋０１１０１０＋０１１００１
＝０１１０１１

【００９１】

ここで、２Ｄ４”及び３Ｄ３”のいずれについても、それぞれストライプ－２及びストライプ－３における消失ブロック（消失パリティ）以外の情報ブロックのＸＯＲ演算が完了しているため、ストライプ－２及びストライプ－３のリカバリが完了する。また、図１０の下段に示すように、２Ｄ４”の値（０１０１００）及び３Ｄ３”の値（０１１０１１）が、それぞれノード－５のＨＤＤ－５における２Ｄ４（０１０１００）及び３Ｄ３（０１１０１１）と一致していることがわかる。

【００９２】

このように、第１実施形態においては、故障した記憶装置５の消失ブロック（第２ブロック情報）の復元先は、自装置１０の１以上の第１記憶装置５又は自装置１０とは異なる第２ストレージ装置１０にそなえられた第１記憶装置５における空き記憶領域となる。この場合、通信部４１は、読出部４２により第１記憶装置５から読み出し済の第１ブロック情報を、復元先である第１記憶装置５へ送信するのである。

【００９３】

〔１－５〕ストレージシステムの動作例

次に、上述の如く構成されたストレージシステム１の動作例を、図１１～図１３を参照して説明する。

【００９４】

〔１－５－１〕全体処理の説明

はじめに、図１１を参照して、ストレージシステム１における全体の処理について説明する。

【００９５】

図１１に示すように、ストレージシステム１が正常に運用されている状態において（ステップＳ１、ステップＳ２、及びステップＳ２のＮｏルート）、いずれかのストレージ装置１０が記憶装置５の故障を検出すると（ステップＳ２のＹｅｓルート）、処理がステップＳ３に移行する。ステップＳ３では、各ストレージ装置１０によりリビルドによる障害の復旧が可能か否かが判断される。

【００９６】

リビルドによる障害の復旧が可能である場合（ステップＳ３のＹｅｓルート）、各ストレージ装置１０は、リビルド処理を実施し（ステップＳ４）、処理がステップＳ１に移行する。一方、リビルドによる障害の復旧が不可能である場合（ステップＳ３のＮｏルート）、ストレージシステム１における少なくとも１つのストレージ装置１０がシステムの管理者等へエラー出力を行ない（ステップＳ５）、処理が終了する。

【００９７】

なお、リビルドによる障害の復旧が不可能である場合としては、少なくとも１つのストライプにおいて消失訂正符号の訂正能力を超えた情報ブロックの消失が生じた場合が挙げられる。また、エラー出力の手法としては、システムの管理者が使用する管理者端末へエラーの発生及びエラーの内容を含むメールを送信したり、管理者端末のモニタへエラー出力を行なう等、既知の種々の手法により行なうことが可能である。なお、管理者端末は、

例えばスイッチ２及びネットワーク３、又はスイッチ６を介してストレージ装置１０と相互に通信可能に接続されている。

【００９８】

〔１－５－２〕リビルド処理の説明

次に、図１２及び図１３を参照して、ストレージシステム１におけるリビルド処理（図１１のステップＳ４参照）について説明する。

【００９９】

図１２に示すように、はじめに、ストレージシステム１のストレージ装置１０（ＣＭ４）は、記憶装置５の故障が発生したＣＭ４の通信部４１からの消失ブロックに関する情報の通知に基づき、ＣＭ４間でストライプごとのリカバリ対象ブロックを決定する（ステップＳ１１）。例えば図５の上段に示す構成の場合、各ノードのＣＭ４は、ノード－５のＣＭ４からの通知に基づき、消失ブロックが存在するストライプ－１、ストライプ－２、ストライプ－３について、リカバリ対象ブロックに決定する。この場合、リカバリ対象ブロックは、ノード－１（ストライプ－２）、ノード－２（ストライプ－３）、ノード－６（ストライプ－１）の各記憶装置５の代替ブロックとなる。

【０１００】

次いで、各ＣＭ４は、書込部４４によりキャッシュメモリ４３の初期化を行なう（ステップＳ１２）。なお、ステップＳ１２の処理は、少なくともリカバリ対象ブロックを持つストレージ装置１０のＣＭ４が実施すればよい。

【０１０１】

また、ＣＭ４は、読出部４２により記憶装置５の未読出の情報ブロックを物理アドレスの昇順で１つ読み出す（ステップＳ１３）。なお、自装置１０の記憶装置５にリカバリ対象ブロックが存在する場合、読出部４２により読み出す情報ブロックには、当該リカバリ対象ブロックは含まれない。読出部４２は、読み出す情報ブロックが当該リカバリ対象ブロックであれば、このブロックをスキップして次のアドレスの情報ブロックを読み出す。

【０１０２】

そして、各ＣＭ４は、読み出した情報ブロックが当該情報ブロックに対応するストライプの復旧先（転送先）のＣＭ４でキャッシュメモリ４３に格納可能か否かを判断する（ステップＳ１４）。なお、この判断は、復旧先のＣＭ４から、キャッシュメモリ４３の容量が逼迫したことを示す通知等を受信しているか否かの判断により行なうことができる。

【０１０３】

このような通知を受信しておらず、復旧先のＣＭ４で情報ブロックを格納可能である場合（ステップＳ１４のＹｅｓルート）、処理がステップＳ１６に移行する。一方、このような通知を受信しており、復旧先のＣＭ４で情報ブロックを格納不可能である場合（ステップＳ１４のＮｏルート）、ＣＭ４は、復旧先が情報ブロックを格納可能になるまで待機し（ステップＳ１５）、格納可能になった旨の通知を受けると、処理がステップＳ１６に移行する。なお、ＣＭ４は、この待機において、読み出した情報ブロックをキャッシュメモリ４３に退避しておいてもよい。

【０１０４】

ステップＳ１６では、ＣＭ４が読み出した情報ブロックを通信部４１により復旧先のＣＭ４へ転送する。

【０１０５】

また、ＣＭ４が通信部４１により他のＣＭ４から情報ブロックを受信した場合（ステップＳ１７及びステップＳ１７のＹｅｓルート）、処理が図１３のステップＳ２２に移行する。なお、情報ブロックを受信したということは、自装置１０の記憶装置５にリカバリ対象ブロックが存在することを意味する。

【０１０６】

一方、ＣＭ４が他のＣＭ４から情報ブロックを受信していない場合（ステップＳ１７のＮｏルート）、ＣＭ４は読出部４２により記憶装置５の最終ブロックまで読み出しが完了したか否かを判断する（ステップＳ１８）。最終ブロックまで読み出しが完了していない

場合（ステップＳ１８のＮｏルート）、処理がステップＳ１３に移行し、最後に読み出した情報ブロックの次のアドレスの情報ブロックを読み出す。

【０１０７】

また、ステップＳ１８において、最終ブロックまで読み出しが完了した場合（ステップＳ１８のＹｅｓルート）、自装置１０の記憶装置５が復旧先でなければ（ステップＳ１９及びステップＳ１９のＮｏルート）、リビルド処理が終了する。なお、自装置１０の記憶装置５が復旧先でないとは、自装置１０の記憶装置がリカバリ対象ブロックを持っていない場合である。

【０１０８】

一方、自装置１０の記憶装置５が復旧先であれば（ステップＳ１９のＹｅｓルート）、ＣＭ４は、自装置１０の記憶装置５に対するリカバリ対象ブロックのリカバリが完了しているか否かを判断する（ステップＳ２０）。リカバリが完了している場合（ステップＳ２０のＹｅｓルート）、リカバリ処理が終了する。また、リカバリが完了していない場合（ステップＳ２０のＮｏルート）、ＣＭ４は、通信部４１により他のＣＭ４から情報ブロックを受信するまで待機し（ステップＳ２１）、情報ブロックを受信すると、処理が図１３のステップＳ２２に移行する。

【０１０９】

図１３に示すように、ステップＳ２２では、ＣＭ４の書込部４４が、通信部４１により受信した情報ブロックをキャッシュメモリ４３へ格納する。そして、書込部４４は、キャッシュメモリ４３の容量を監視するための情報として、例えばキャッシュメモリ４３の使用量を示す容量情報を更新する。また、リカバリ対象ブロックのリカバリが完了したか否かを判断するための情報として、例えば受信した（又はキャッシュメモリ４３に格納した）情報ブロック数のカウント値を更新する（ステップＳ２３）。なお、既述のように、キャッシュメモリの使用量についても情報ブロック数のカウント値が用いられてもよく、この場合、書込部４４は、情報ブロック数のカウント値のみを更新すればよい。

【０１１０】

次に、書込部４４は、容量情報が閾値を超えたか否かを判断し（ステップＳ２４）、超えていない場合（ステップＳ２４のＮｏルート）、処理が図１２のステップＳ１８に移行する。

【０１１１】

一方、容量情報が閾値を超えた場合（ステップＳ２４のＹｅｓルート）、書込部４４は、容量情報が閾値を超えたこと（キャッシュ不可の旨）を通信部４１を介して他のＣＭ４に通知する（ステップＳ２５）。そして、書込部４４は、リカバリ対象ブロックに情報が格納済みであるか否かを判断する（ステップＳ２６）。

【０１１２】

リカバリ対象ブロックに情報が格納済みである場合（ステップＳ２６のＹｅｓルート）、書込部４４は、リカバリ対象ブロックの情報を読み出す。そして、書込部４４は、読み出した情報とキャッシュメモリ４３内の情報ブロックとのＸＯＲ演算を実行し（ステップＳ２７）、処理がステップＳ２９に移行する。一方、リカバリ対象ブロックに情報が格納されていない場合（ステップＳ２６のＮｏルート）、書込部４４は、キャッシュメモリ４３内の情報ブロックのＸＯＲ演算を実行し（ステップＳ２８）、処理がステップＳ２９に移行する。

【０１１３】

ステップＳ２９では、書込部４４は、ステップＳ２７又はステップＳ２８におけるＸＯＲ演算結果をリカバリ対象ブロックへ書き込む。なお、キャッシュメモリ４３の内容はリカバリ対象ブロックへ反映されたため、書込部４４はキャッシュメモリ４３をクリアする。また、書込部４４は、キャッシュ可能の旨を通信部４１を介して他のＣＭ４に通知する（ステップＳ３０）。

【０１１４】

そして、書込部４４は、カウント値が閾値に達したか否かを判断し（ステップＳ３１）

、達していない場合（ステップＳ３１のＮｏルート）、つまりリカバリが完了していない場合、処理が図１２のステップＳ１８に移行する。

【０１１５】

一方、カウント値が閾値に達した場合（ステップＳ３１のＹｅｓルート）、書込部４４は、リカバリ対象ブロックのリカバリが完了したと判断し（ステップＳ３２）、処理が図１２のステップＳ１８に移行する。

【０１１６】

このように、リビルド処理では、リカバリ対象ブロックを持たないＣＭ４では読出処理（ステップＳ１３～Ｓ１８）が主に実行され、リカバリ対象ブロックを持つＣＭ４では読出処理と書込処理（ステップＳ１７、ステップＳ２０～Ｓ３２）とが実施される。

【０１１７】

読出処理は、シーケンシャルリードによりスループットを向上させているため、読出処理の最中に書込処理が頻発すると、記憶装置５のアクセス先が変化してシーケンシャルなアクセスが阻害されてしまう。そこで、リカバリ対象ブロックを持つＣＭ４は、受信した情報ブロックをキャッシュメモリ４３に蓄積し、一括して書込処理を行なうことで、書込処理の実行頻度を低下させることができるのである。

【０１１８】

なお、図１２のステップＳ１７における情報ブロックの受信確認の処理は、ステップＳ１３の前（ステップＳ１８のＮｏルートからの合流後）等、任意のタイミングで行なわれてよい。また、図１３のステップＳ２３における情報ブロック数のカウント値の更新処理は、ステップＳ２９の後等に行なわれてもよい。

【０１１９】

〔２〕第２実施形態

上述した第１実施形態に係るストレージシステム１では、リカバリ対象ブロックに運用記憶装置５の代替ブロック（空き領域）が用いられるものとして説明したが、これに限定されるものではない。

【０１２０】

例えばストレージシステム１は、図１４の上段に示すように、運用中のストレージ装置１０（ノード－１～ノード－６）の他に、待機用のストレージ装置１０（ノード－７）をそなえてもよい。このとき、ストレージシステム１は、待機用のストレージ装置１０のホットスワップディスクであるＨＤＤ－７上に、故障した記憶装置５（ＨＤＤ－５）のデータを復元することができる。

【０１２１】

以下、第２実施形態に係るストレージシステム１について説明する。なお、ストレージシステム１及びストレージ装置１０の構成及び機能は、特に言及しない限り、第１実施形態と基本的に同様とすることができる。

【０１２２】

第２実施形態に係るストレージシステム１では、記憶装置５が故障した場合の代替となる記憶装置５（ＨＤＤ－７）を予め用意し、ホットスワップとしてシステムに組み込まれている。例えばノード－５のＨＤＤ－５が故障した場合、図１４の上段に示すように、運用中のノード－１～ノード－４及びノード－６は、それぞれ情報ブロックを待機用のノード－７へ送信する。情報ブロックは複数のノードから順次送られてくるため、情報ブロックを受信したノード－７は、各ストライプについてキャッシュメモリ４３を用いて順次リカバリ対象ブロックへの書込処理（ＸＯＲ演算）を行なう。

【０１２３】

このように、運用中のストレージ装置１０では、消失ブロックのデータを復旧するための読出処理（シーケンシャルリード）を行ない、読み出した情報ブロックを待機用のストレージ装置１０へ送信するだけでよく、書込処理は発生しない。一方、待機用のストレージ装置１０では、情報ブロックを受信し、受信した情報ブロックの記憶装置５への書込処理（ランダムライト）を行なうだけでよい。

【０１２４】
以上のように、第２実施形態に係るストレージシステム１では、復旧用の情報ブロックの読出要求と、復旧先での情報ブロックの書込要求とが別々の記憶装置５に発行されるため、シーケンシャルリードが中断されず、スループットをさらに向上させることができる。
【０１２５】
このため、第２実施形態に係るストレージシステム１では、運用中のストレージ装置１０（ＣＭ４）は、例えば図４に示す機能のうち、少なくとも通信部４１及び読出部４２の機能をそなえていればよい。また、待機用のストレージ装置１０（ＣＭ４）は、例えば図４に示す機能のうち、少なくとも通信部４１、キャッシュメモリ４３、及び書込部４４の機能をそなえていればよい。
【０１２６】
なお、図１４に示す例では、ホットスワップディスクであるＨＤＤ－７が待機用のストレージ装置１０にそなえられるものとして示したが、これに限定されるものではない。例えばＨＤＤ－７は、運用中のストレージ装置１０（ノード－１～ノード－４及びノード－６のいずれか）に追加してそなえられてもよい。この場合、運用中のストレージ装置１０は、図４に示す通信部４１、読出部４２、キャッシュメモリ４３、及び書込部４４の全ての機能をそなえればよい。そして、当該ストレージ装置１０は、例えばＨＤＤ－７以外のＨＤＤから読出部４２により情報ブロックを読み出してキャッシュメモリ４３に格納し、キャッシュメモリ４３内の情報ブロックを書込部４４によりＨＤＤ－７のリカバリ対象ブロックへ反映すればよい。
【０１２７】
なお、図１４の例において、各ストライプの消失ブロックの復旧の進捗は、各記憶装置５での読出順序に関連して、図１４の下段に示すように、先頭の情報ブロック、中央付近の情報ブロック、末尾の情報ブロックの順に、段階的に（この場合３段階で）行なわれる。
【０１２８】
また、第２実施形態に係るストレージシステム１では、上述のように運用中のストレージ装置１０における記憶装置５の代替ブロックは使用されない。換言すれば、第２実施形態に係るストレージシステム１としては、代替ブロックを持たない、例えば通常のＲＡＩＤ５やＲＡＩＤ６等を採用したストレージシステムを用いることもできる。
【０１２９】
例えば図１５の上段に示すように、消失訂正符号としてＲＡＩＤ５を採用したストレージシステム１は、図１４と比較して、ＨＤＤ－１が２Ｄ５を格納し、ＨＤＤ－２が３Ｄ５を格納し、ＨＤＤ－６が１Ｄ５を格納している。このような構成であっても、復旧用の情報ブロックはノード－７へ送信されるため、各ストレージ装置１０は、図１４を参照した説明と同様の処理を行なうことができる。
【０１３０】
ところで、第２実施形態においては、消失ブロックに対応する全てのストライプについて、復旧用の情報ブロックが待機用の記憶装置５（ストレージ装置１０）に送信されることになる。このため、待機用のストレージ装置１０（ＣＭ４）では、大量の書込処理が発生することになる。
【０１３１】
そこで、待機用のストレージ装置１０は、第１実施形態と比較して大容量のキャッシュメモリ４３をそなえることが好ましい。これにより、１つのリカバリ対象ブロックについて或る程度の数の情報ブロックのＸＯＲ演算結果をまとめて反映できるため、１ブロック当たりの書込処理の発生頻度を低減させ、記憶装置５の処理負荷を低減させることができる。また、キャッシュメモリ４３の容量の逼迫により運用中のストレージ装置１０で送信待ちが発生する頻度も低減させることができ、リカバリ性能を向上させることができる。
【０１３２】

又は、待機用のストレージ装置１０は、キャッシュメモリ４３の容量が逼迫するよりも早いタイミングで、キャッシュメモリ４３内の情報ブロックをリカバリ対象ブロックに反映してもよい。このタイミングとしては、例えば所定期間ごとが挙げられる。或いはキャッシュメモリ４３の使用量の閾値を第１実施形態よりも小さい値とすることで、キャッシュメモリ４３内の情報ブロックをパージする頻度を上げてもよい。

【０１３３】

これにより、待機用の記憶装置５における１ブロック当たりの書込処理の発生頻度は増加するものの、キャッシュメモリ４３の容量が逼迫する前にキャッシュメモリ４３内の情報ブロックがリカバリ対象ブロックに反映される。従って、運用中のストレージ装置１０で送信待ちが発生する頻度を低減させる、或いは無くすことができ、リカバリ性能を向上させることができる。

【０１３４】

このように、第２実施形態においては、故障した記憶装置５の消失ブロック（第２ブロック情報）の復元先は、故障した第２記憶装置５の代替となる第３記憶装置５における空き記憶領域となる。この場合、通信部４１は、読出部４２により第１記憶装置５から読み出し済の第１ブロック情報を、第３記憶装置５へ送信するのである。

【０１３５】

次に、図１６～図１８を参照して、第２実施形態に係るストレージシステム１の動作例を説明する。なお、図１６～図１８において、第１実施形態に係る図１２及び図１３に示す符号と同一の符号を付した処理は、図１２及び図１３に示す処理と基本的に同様の処理であるため、重複した説明を省略する。

【０１３６】

第２実施形態に係るストレージシステム１では、運用ストレージ装置１０と待機ストレージ装置１０とで処理が異なる。

【０１３７】

例えば図１６に示すように、運用ストレージ装置１０の処理は、図１２に示すストレージ装置１０の処理から、ステップＳ１２、ステップＳ１７、及びステップＳ１９～Ｓ２１の書込処理に係る処理を省略したものとなる。

【０１３８】

また、図１７に示すように、待機ストレージ装置１０の処理は、図１３に示すストレージ装置１０の処理に対して、以下の処理を追加又は変更したものとなる。

【０１３９】

例えば図１７に示すように、待機ストレージ装置１０のＣＭ４は、リビルド処理において、キャッシュメモリ４３を初期化し（ステップＳ４１）、運用ストレージ装置１０から情報ブロックを受信するまで待機する（ステップＳ４２）。また、ステップＳ２４において容量情報が閾値を超えていない場合（ステップＳ２４のＮｏルート）、及びステップＳ３１においてカウント値が閾値に達していない場合（ステップＳ３１のＮｏルート）、処理がステップＳ４２に移行する。その他の点については、待機ストレージ装置１０の処理は図１３に示すストレージ装置１０の書込処理の動作と基本的に同様である。

【０１４０】

なお、待機ストレージ装置１０は、上述のようにキャッシュメモリ４３の容量が逼迫するよりも早いタイミングでキャッシュメモリ４３内の情報ブロックをリカバリ対象ブロックに反映してもよい。この場合、待機ストレージ装置１０の処理は、図１８に示すように、図１７からステップＳ２４、ステップＳ２５、及びステップＳ３０の処理を省略し、ステップＳ２３の処理を情報ブロック数のカウント値の更新のみを行なうステップＳ５１に置き換えたものとすることができる。

【０１４１】

以上のように、第２実施形態に係るストレージシステム１によっても、第１実施形態と同様の効果を奏することができるほか、ＣＭ４が記憶装置５から情報ブロックをよりシーケンシャルに読み出すことができ、スループットをさらに向上させることができる。

【０１４２】
〔３〕ハードウェア構成例
図１９に例示するように、上述した第１及び第２実施形態に係るストレージ装置１０のＣＭ４は、図１に示すＣＰＵ４ａ及びメモリ４ｂに加えて、記憶部４ｆ、インタフェース部４ｇ、入出力部４ｈ、及び読取部４ｉをそなえることができる。
【０１４３】
記憶部４ｆは、種々のデータやプログラム等を格納するハードウェアである。記憶部４ｆとしては、例えばＨＤＤ等の磁気ディスク装置、ＳＳＤ等の半導体ドライブ装置、フラッシュメモリやＲＯＭ等の不揮発性メモリ等の各種装置が挙げられる。
【０１４４】
例えば記憶部４ｆは、ストレージ装置１０（ＣＭ４）の各種機能の全部もしくは一部を実現するリカバリプログラム４０を格納することができる。ＣＰＵ４ａは、例えば記憶部４ｆに格納されたリカバリプログラム４０をメモリ４ｂ等の記憶装置に展開して実行することにより、故障した記憶装置５の消失ブロックをリカバリ（復旧）することでリビルドを行なう上述したストレージ装置１０（ＣＭ４）の機能を実現することができる。
【０１４５】
インタフェース部４ｇは、有線又は無線による、スイッチ２及び６、並びに記憶装置５等との間の接続及び通信の制御等を行なう通信インタフェースである。なお、図１に示すＩＦ４ｃ～４ｅは、インタフェース部４ｇの一例である。
【０１４６】
入出力部４ｈは、マウス、キーボード、タッチパネル、音声操作のためのマイク等の入力装置（操作部）、並びにディスプレイ、スピーカ、及びプリンタ等の出力装置（出力部、表示部）の少なくとも一方を含むことができる。例えば入力装置は、管理者等による各種操作やデータの入力等の作業に用いられてよく、出力装置は、各種通知等の出力に用いられてよい。
【０１４７】
読取部４ｉは、コンピュータ読取可能な記録媒体４ｊに記録されたデータやプログラムを読み出す装置である。この記録媒体４ｊにはリカバリプログラム４０が格納されてもよい。
【０１４８】
なお、記録媒体４ｊとしては、例えばフレキシブルディスク、ＣＤ、ＤＶＤ、ブルーレイディスク等の光ディスクや、ＵＳＢメモリやＳＤカード等のフラッシュメモリ等の非一時的な記録媒体が挙げられる。なお、ＣＤとしては、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ等が挙げられる。また、ＤＶＤとしては、ＤＶＤ－ＲＯＭ、ＤＶＤ－ＲＡＭ、ＤＶＤ－Ｒ、ＤＶＤ－ＲＷ、ＤＶＤ＋Ｒ、ＤＶＤ＋ＲＷ等が挙げられる。
【０１４９】
上述したストレージ装置１０のハードウェア構成は例示である。従って、ストレージ装置１０内でのハードウェアの増減（例えば任意のブロックの追加や省略）、分割、任意の組み合わせでの統合、バスの追加又は省略等は適宜行なわれてもよい。
【０１５０】
〔４〕その他
以上、本発明の好ましい実施形態について詳述したが、本発明は、かかる特定の実施形態に限定されるものではなく、本発明の趣旨を逸脱しない範囲内において、種々の変形、変更して実施することができる。
【０１５１】
例えば、図４に示すストレージ装置１０（ＣＭ４）の各機能ブロックは、任意の組み合わせで併合してもよく、分割してもよい。
【０１５２】
また、ここまで、ストレージシステム１が複数のストレージ装置１０（筐体）をそなえるクラスタ構成の分散ストレージシステムであるものとして説明したが、これに限定され

るものではない。例えばストレージシステム１は、図２０に示すように、単一のストレージ装置１０内にＪＢＯＤ（Just a Brunck of Disks）５０等の形態で実装されたディスクアレイをそなえてもよい。

【０１５３】

ＪＢＯＤ５０は、複数の記憶装置５を搭載し、複数の記憶装置５に接続されたＩＦ５０ａを介してＣＭ４のＩＦ４ｄと通信を行なうことができる。この場合、ＣＭ４は、読出部４２により複数の記憶装置５からシーケンシャルに読み出した情報ブロックを、書込部４４によりキャッシュメモリ４３に格納すればよい。そして、ＣＭ４は、キャッシュメモリ４３の容量が逼迫した場合に、キャッシュメモリ４３内の情報ブロックを復旧先の記憶装置５のリカバリ対象ブロックに反映すればよい。

【０１５４】

なお、図２０に例示する構成において、ＣＭ４は図４に示す通信部４１、読出部４２、キャッシュメモリ４３、及び書込部４４をそなえることができる。例えば読出部４２は、記憶装置５から情報ブロックをシーケンシャルに読み出して通信部４１に出力する。この情報ブロックは、通信部４１を介して書込部４４によりキャッシュメモリ４３に格納され、リカバリ対象ブロックを有する記憶装置５へ書き込まれる。

【０１５５】

このように、図２０及び図４に例示する構成において、キャッシュメモリ４３は、通信部４１から入力される第１ブロック情報を保持する保持部の一例であるといえる。また、書込部４４は、通信部４１から入力された第１ブロック情報に基づき、第２ブロック情報を段階的に復元する復元部の一例であるといえる。

【０１５６】

なお、ストレージシステム１は、複数のストレージ装置１０（クラスタノード）内にそれぞれＪＢＯＤ５０等により複数の記憶装置５を搭載する、図１及び図２０の双方の形態を持つ構成であってもよい。

【０１５７】

〔５〕付記

以上の第１及び第２実施形態に関し、更に以下の付記を開示する。

【０１５８】

（付記１）

１以上の第１記憶装置と、

前記１以上の第１記憶装置が記憶する複数の第１ブロック情報であって、故障した第２記憶装置が記憶する複数の第２ブロック情報の復元に用いられる前記複数の第１ブロック情報を、前記１以上の第１記憶装置から記憶領域のアドレス順に読み出す読出部と、

前記読出部により前記１以上の第１記憶装置から読み出し済の第１ブロック情報を、前記複数の第２ブロック情報を段階的に復元するために、前記複数の第２ブロック情報の復元先へ出力する出力部と、をそなえることを特徴とする、ストレージ装置。

【０１５９】

（付記２）

前記ストレージ装置とは異なる第１ストレージ装置、又は、前記出力部、から入力された第１ブロック情報に基づき、前記第２ブロック情報を段階的に復元する復元部をさらにそなえることを特徴とする、付記１記載のストレージ装置。

【０１６０】

（付記３）

前記復元部は、前記入力された第１ブロック情報及び前記第２ブロック情報の復元先が記憶する情報を用いて、消失訂正符号に基づく演算を行ない、演算結果を前記第２ブロック情報の復元先へ書き込むことを特徴とする、付記２記載のストレージ装置。

【０１６１】

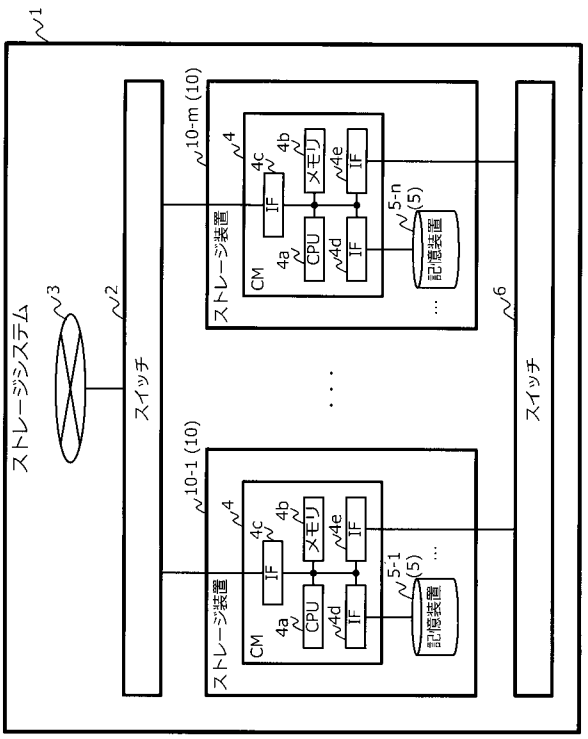
（付記４）

前記第１ストレージ装置又は前記出力部から入力される第１ブロック情報を保持する保

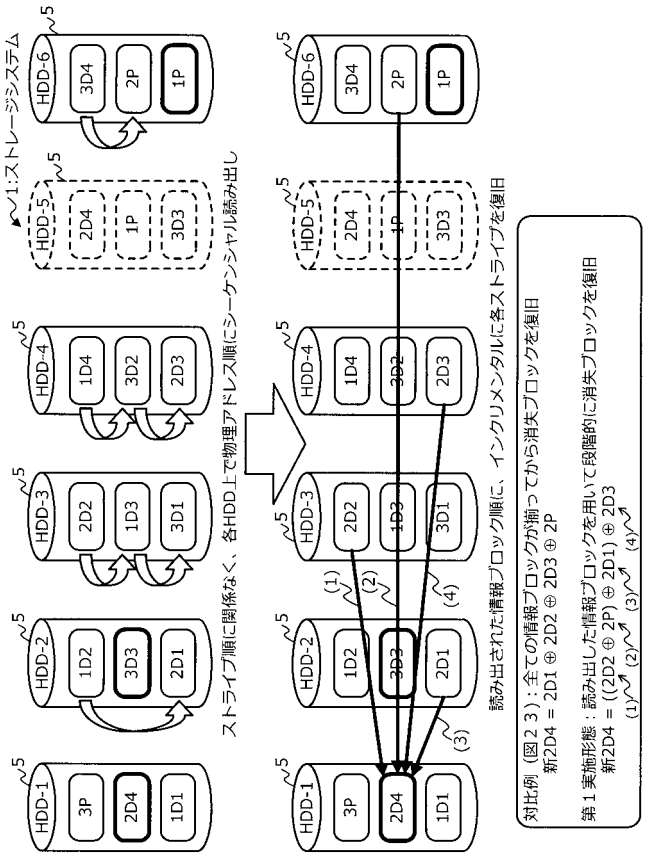
持部をさらにそなえ、
前記復元部は、所定のタイミングで、前記保持部が保持する１以上の第１ブロック情報に基づき、前記第２ブロック情報を段階的に復元することを特徴とする、付記２又は付記３記載のストレージ装置。
【０１６２】
（付記５）
前記第２ブロック情報の復元先は、前記１以上の第１記憶装置、又は、前記ストレージ装置とは異なる第２ストレージ装置にそなえられた第１記憶装置における空き記憶領域であり、
前記出力部は、前記読出部により前記第１記憶装置から読み出し済の第１ブロック情報を、前記復元先である第１記憶装置へ送信することを特徴とする、付記１～４のいずれか１項記載のストレージ装置。
【０１６３】
（付記６）
前記第２ブロック情報の復元先は、前記故障した第２記憶装置の代替となる第３記憶装置における空き記憶領域であり、
前記出力部は、前記読出部により前記第１記憶装置から読み出し済の第１ブロック情報を、前記第３記憶装置へ送信することを特徴とする、付記１～４のいずれか１項記載のストレージ装置。
【０１６４】
（付記７）
複数の記憶装置と、
前記複数の記憶装置のうちの複数の第１記憶装置の各々が記憶する複数の第１ブロック情報に基づき、前記複数の記憶装置のうちの故障した第２記憶装置が記憶する複数の第２ブロック情報を復元する１以上のストレージ装置と、をそなえ、
前記１以上のストレージ装置は、
複数の第１記憶装置の各々から、前記複数の第１ブロック情報を記憶領域のアドレス順に読み出し、
前記読み出しの処理において前記複数の第１記憶装置の各々から読み出し済の第１ブロック情報に基づき、前記複数の第２ブロック情報を段階的に復元する、
ことを特徴とする、ストレージシステム。
【０１６５】
（付記８）
コンピュータに、
１以上の第１記憶装置が記憶する複数の第１ブロック情報であって、故障した第２記憶装置が記憶する複数の第２ブロック情報の復元に用いられる前記複数の第１ブロック情報を、前記１以上の第１記憶装置から記憶領域のアドレス順に読み出し、
前記読み出しの処理において前記１以上の第１記憶装置から読み出し済の第１ブロック情報を、前記複数の第２ブロック情報を段階的に復元するために、前記複数の第２ブロック情報の復元先へ出力する、
処理を実行させることを特徴とする、リカバリプログラム。
【０１６６】
（付記９）
複数の記憶装置と、前記複数の記憶装置に対する制御を行なう１以上のストレージ装置とをそなえるストレージシステムにおけるリカバリ方法であって、
前記１以上のストレージ装置は、
前記複数の記憶装置のうちの複数の第１記憶装置の各々が記憶する複数の第１ブロック情報に基づき、前記複数の記憶装置のうちの故障した第２記憶装置が記憶する複数の第２ブロック情報を復元し、
前記復元の処理において、

前記１以上のストレージ装置により、
複数の第１記憶装置の各々から、前記複数の第１ブロック情報を記憶領域のアドレス順に読み出し、
前記読み出しの処理において前記複数の第１記憶装置の各々から読み出し済の第１ブロック情報に基づき、前記複数の第２ブロック情報を段階的に復元する、
ことを特徴とする、リカバリ方法。
【符号の説明】
【０１６７】
１　ストレージシステム
２，６　スイッチ
３　ネットワーク
４　コントローラモジュール
４ａ　ＣＰＵ
４ｂ　メモリ
４ｃ～４ｅ，５０ａ　インタフェース
４ｆ　記憶部
４ｇ　インタフェース部
４ｈ　入出力部
４ｉ　読取部
４ｊ　記録媒体
５，５－１～５－ｎ　記憶装置
１０，１０－１～１０－ｍ　ストレージ装置
４０　リカバリプログラム
４１　通信部
４２　読出部
４３　キャッシュメモリ
４４　書込部
５０　ＪＢＯＤ

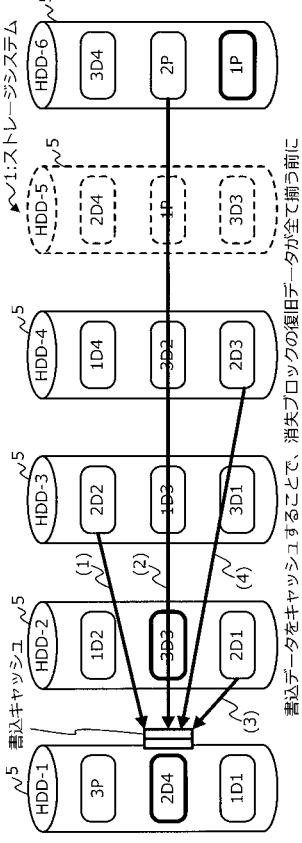
【図１】



【図２】

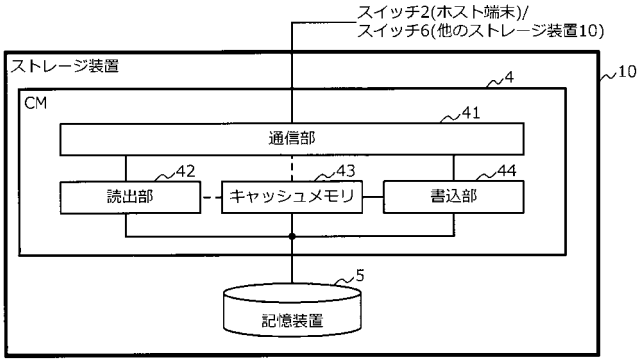


【図３】

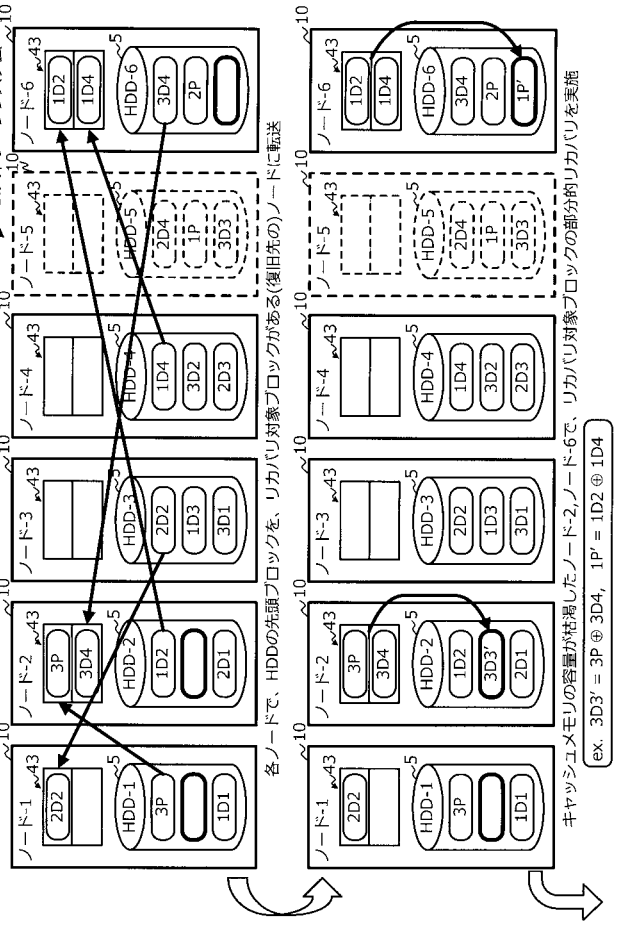


	処理	ストライプ-2の2D4復旧の進捗
(1)	HDD-3から2D2を読み出し、HDD-1の新2D4の書込キャッシュに反映	2D4 = 2D2
(2)	HDD-6から2Pを読み出し、HDD-1の新2D4の書込キャッシュに反映	2D4 = 2D4 ⊕ 2P
	新2D4の書込キャッシュをフラッシュ	
(3)	HDD-2から2D1を読み出し、HDD-1の新2D4を書込キャッシュにリロードし、2D1をキャッシュに反映	2D4 = 2D4 ⊕ 2D1
(4)	HDD-4から2D3を読み出し、HDD-1の新2D4の書込キャッシュに反映	2D4 = 2D4 ⊕ 2D3
	新2D4の書込キャッシュをフラッシュ	

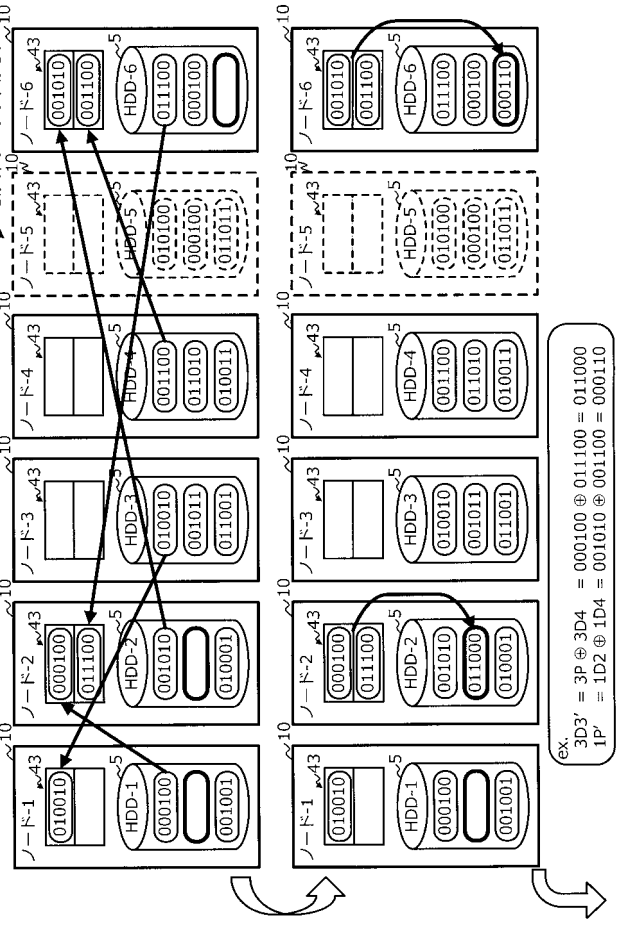
【図４】



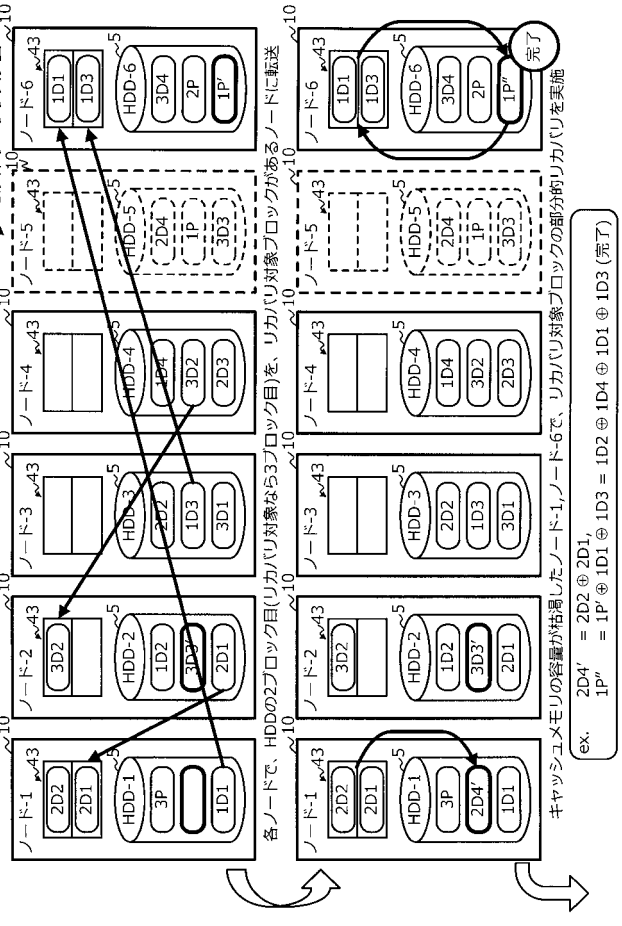
【図５】


1:ストレージシステム
ノード-1
ノード-2
ノード-3
ノード-4
ノード-5
ノード-6
HDD-1
HDD-2
HDD-3
HDD-4
HDD-5
HDD-6
各ノードで、HDDの先頭ブロックを、リカバリ対象ブロックがある(復旧先の)ノードに転送
キャッシュメモリの容量が枯渇したノード-2,ノード-6で、リカバリ対象ブロックの部分的リカバリを実施
ex. 3D3' = 3P ⊕ 3D4, 1P' = 1D2 ⊕ 1D4

【図６】


1:ストレージシステム
ノード-1
ノード-2
ノード-3
ノード-4
ノード-5
ノード-6
HDD-1
HDD-2
HDD-3
HDD-4
HDD-5
HDD-6
ex.
3D3' = 3P ⊕ 3D4 = 000100 ⊕ 011100 = 011000
1P' = 1D2 ⊕ 1D4 = 001010 ⊕ 001100 = 000110

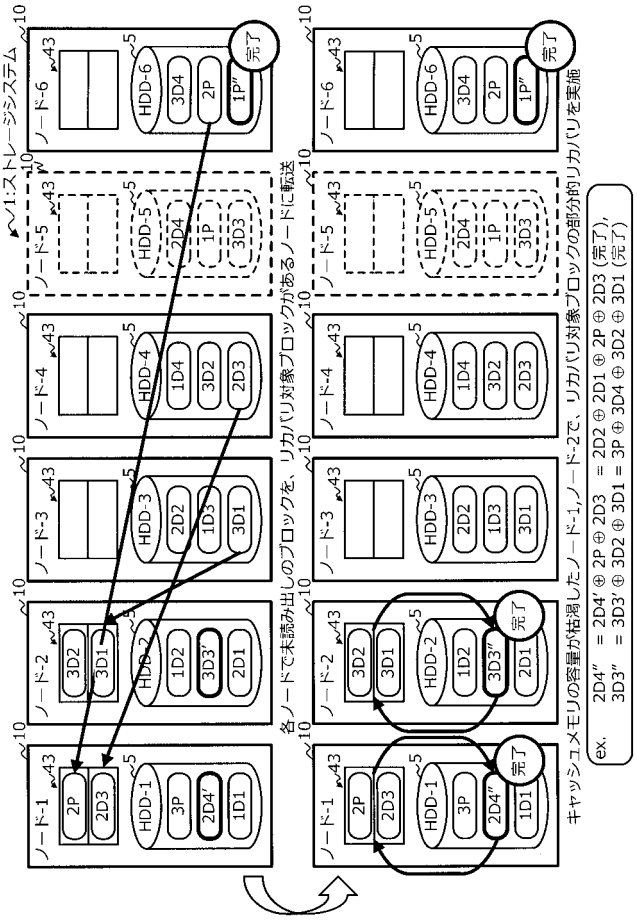
【図７】


1:ストレージシステム
ノード-1
ノード-2
ノード-3
ノード-4
ノード-5
ノード-6
HDD-1
HDD-2
HDD-3
HDD-4
HDD-5
HDD-6
各ノードで、HDDの2ブロック目(リカバリ対象なら3ブロック目)を、リカバリ対象ブロックがあるノードに転送
キャッシュメモリの容量が枯渇したノード-1,ノード-6で、リカバリ対象ブロックの部分的リカバリを実施
完了
ex. 2D4' = 2D2 ⊕ 2D1,
1P'' = 1P' ⊕ 1D1 ⊕ 1D3 = 1D2 ⊕ 1D4 ⊕ 1D1 ⊕ 1D3 (完了)

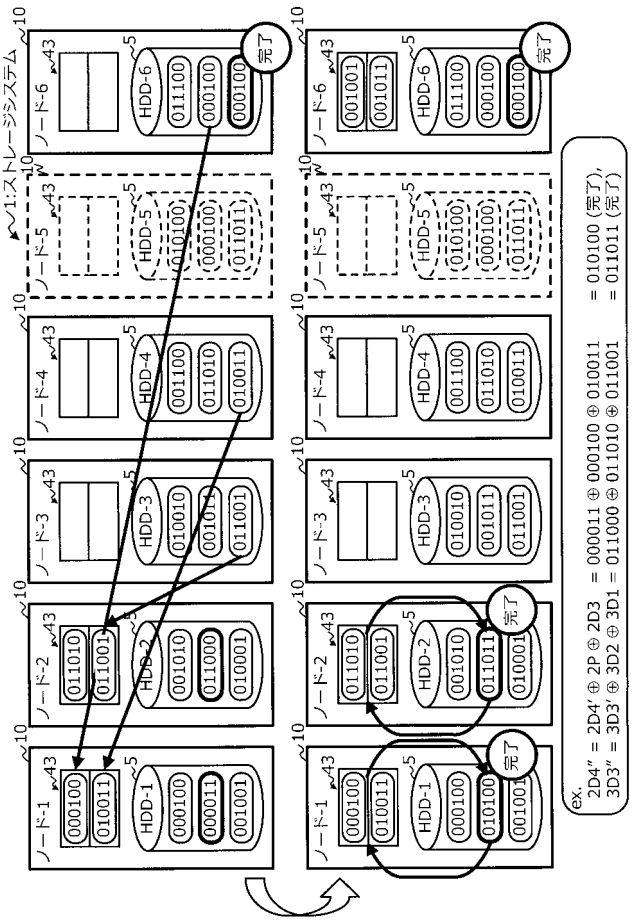
【図８】


1:ストレージシステム
ノード-1
ノード-2
ノード-3
ノード-4
ノード-5
ノード-6
HDD-1
HDD-2
HDD-3
HDD-4
HDD-5
HDD-6
完了
ex.
2D4' = 2D2 ⊕ 2D1 = 010010 ⊕ 010001 = 000011
1P'' = 1P' ⊕ 1D1 ⊕ 1D3 = 000110 ⊕ 001001 ⊕ 001011 = 000100 (完了)

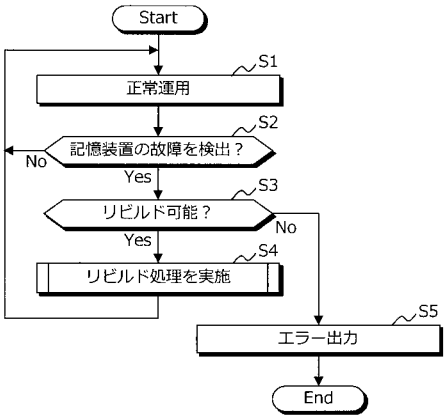
【図９】



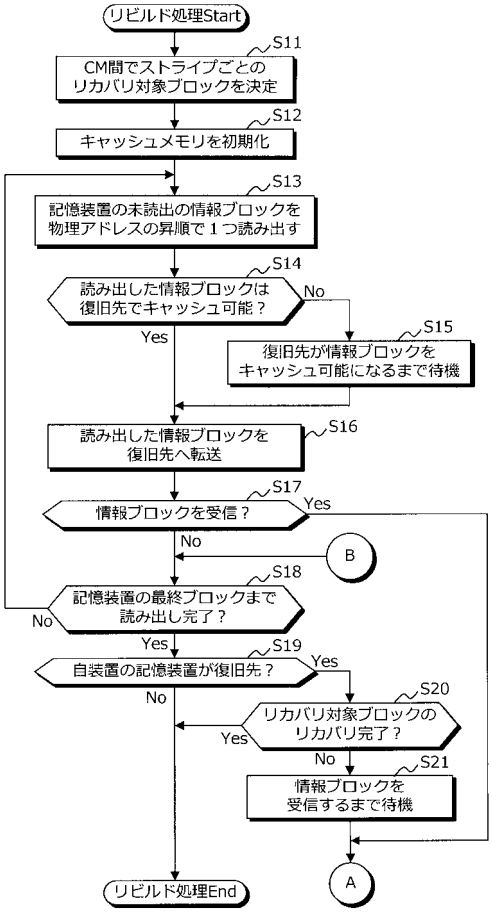
【図１０】



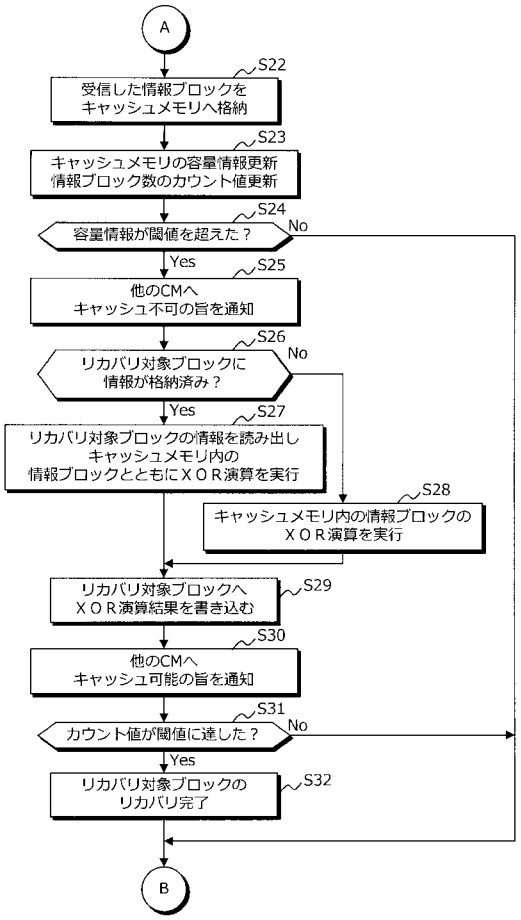
【図１１】



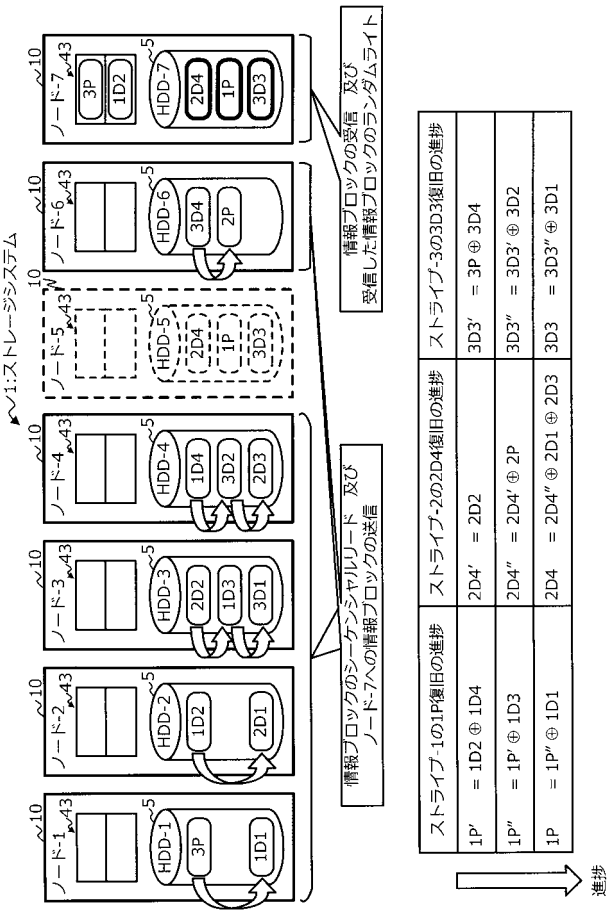
【図１２】



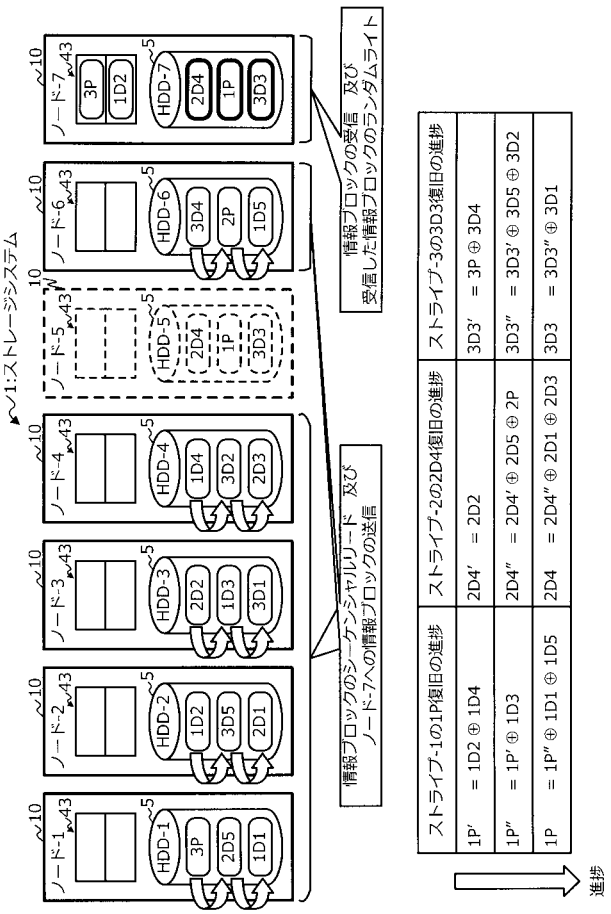
【図１３】



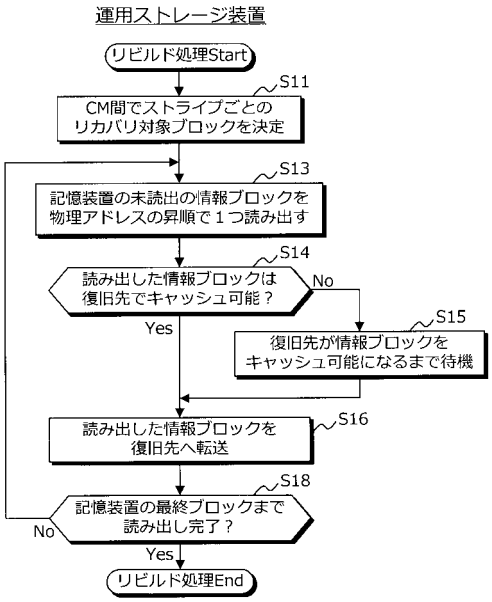
【図１４】



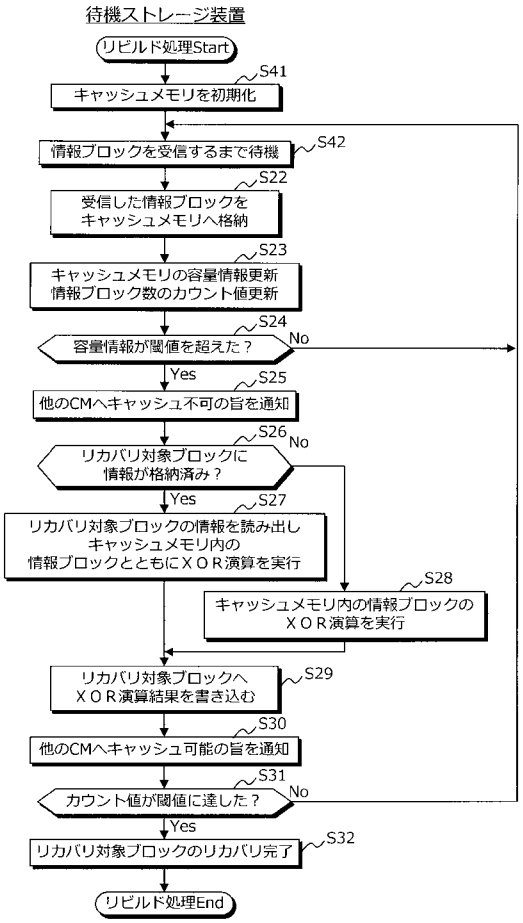
【図１５】



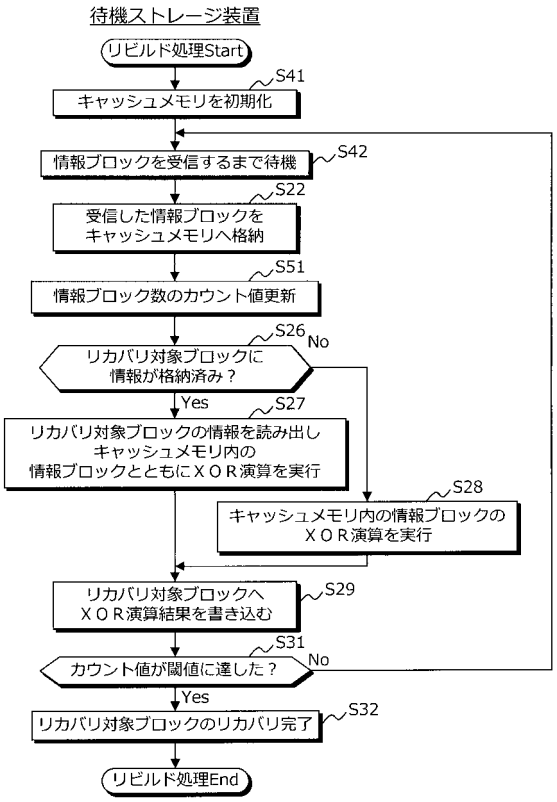
【図１６】



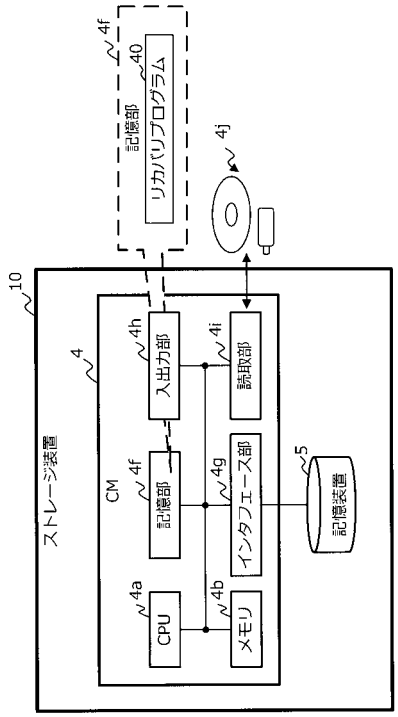
【図１７】



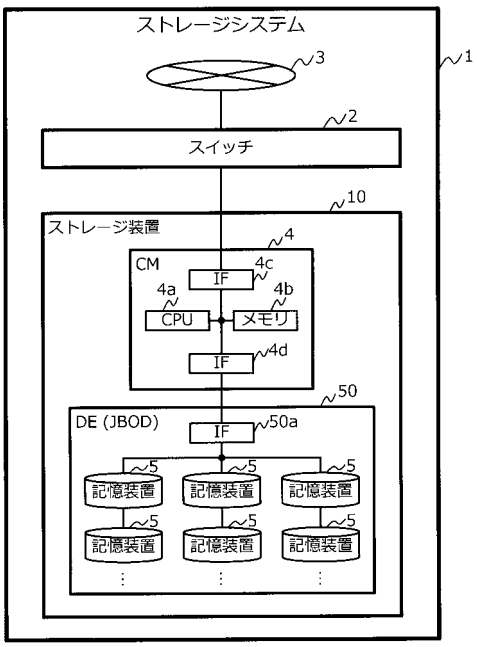
【図１８】



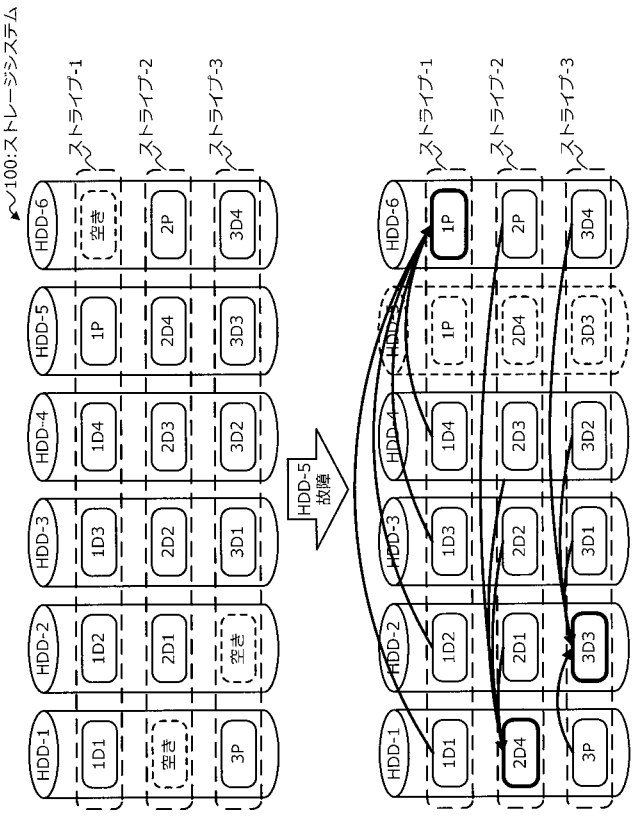
【図１９】



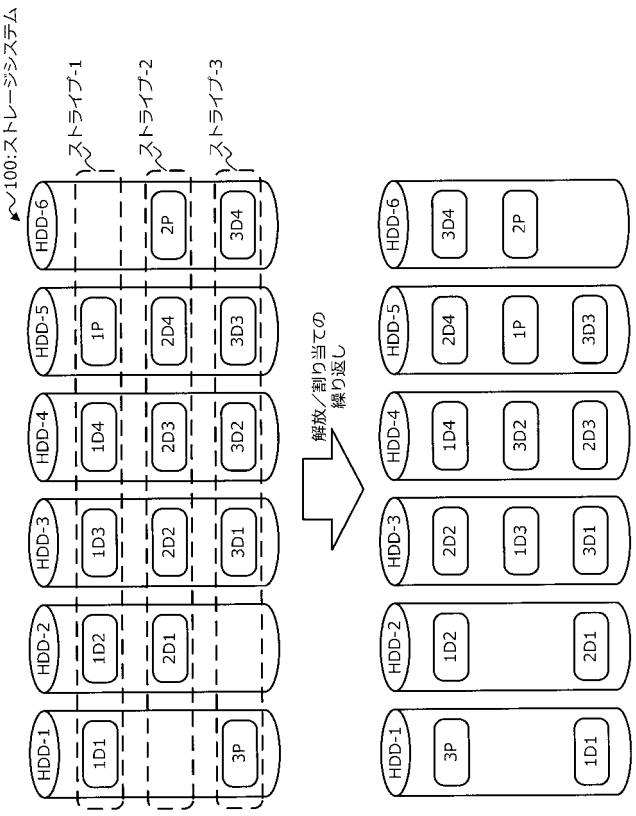
【図２０】



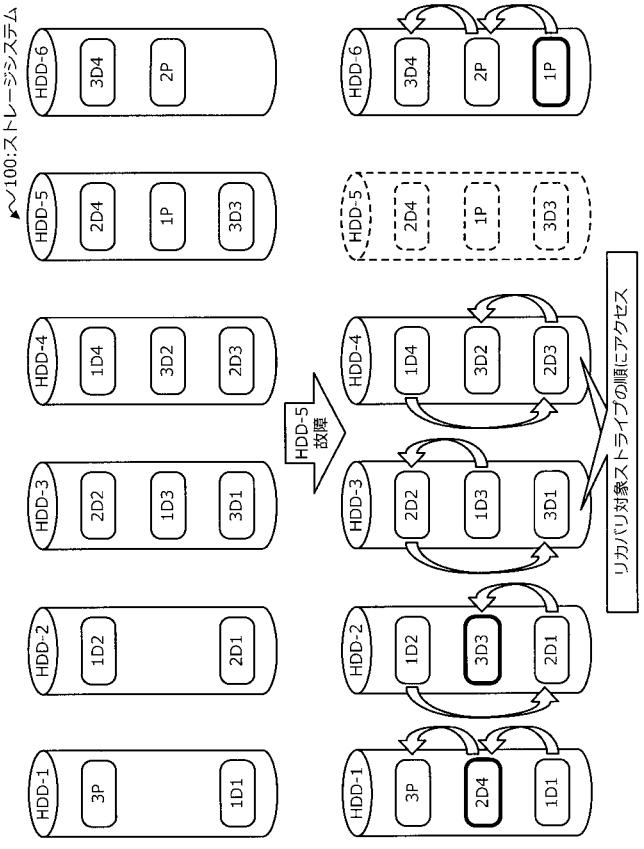
【図２１】


100:ストレージシステム
ストライプ-1
ストライプ-2
ストライプ-3
HDD-1
HDD-2
HDD-3
HDD-4
HDD-5
HDD-6
1D1
1D2
1D3
1D4
1P
2D1
2D2
2D3
2D4
2P
3D1
3D2
3D3
3D4
3P
空き
HDD-5 故障

【図２２】


100:ストレージシステム
ストライプ-1
ストライプ-2
ストライプ-3
HDD-1
HDD-2
HDD-3
HDD-4
HDD-5
HDD-6
1D1
1D2
1D3
1D4
1P
2D1
2D2
2D3
2D4
2P
3D1
3D2
3D3
3D4
3P
解放／割り当ての
繰り返し

【図２３】


100:ストレージシステム
HDD-1
HDD-2
HDD-3
HDD-4
HDD-5
HDD-6
1D1
1D2
1D3
1D4
1P
2D1
2D2
2D3
2D4
2P
3D1
3D2
3D3
3D4
3P
HDD-5 故障
リカバリ対象ストライプの順にアクセス