(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2006-259894
(P2006-259894A)
(43)公開日　平成18年9月28日(2006.9.28)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	304B	5B001
G06F 11/08	(2006.01)	G06F	11/08	310B	5B018
G06F 12/16	(2006.01)	G06F	12/16	320L	5B065

審査請求 有　請求項の数 11　ＯＬ　（全 28 頁）

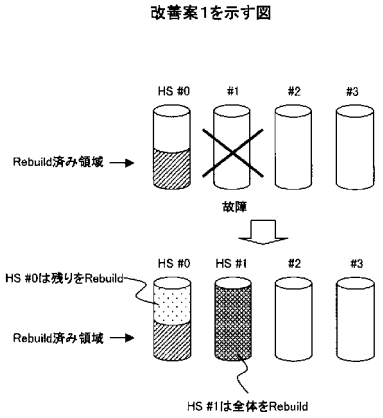
(21)出願番号　特願2005-73669(P2005-73669)
(22)出願日　平成17年3月15日(2005.3.15)

(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　100074099
弁理士　大菅　義之
(74)代理人　100067987
弁理士　久木元　彰
(72)発明者　望月　信哉
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
(72)発明者　伊藤　実希夫
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

最終頁に続く

(54)【発明の名称】ストレージ制御装置および方法

(57)【要約】
【課題】　パリティを用いてデータの冗長化を行うストレージ装置において、データ／パリティを格納する２つの記憶装置が故障したときの再構築処理を改善して、効率よく冗長性を回復する。
【解決手段】　故障ディスク＃０のデータ／パリティの復元中にディスク＃１が故障したとき、既に復元されて予備ディスクＨＳ＃０に書き込まれているデータ／パリティを破棄せずに、他の正常ディスクに格納されたデータ／パリティを用いて、ディスク＃０の未復元領域とディスク＃１のデータ／パリティを復元し、ＨＳ＃０の対応する領域とＨＳ＃１にそれぞれ書き込む。
【選択図】図４



【特許請求の範囲】
【請求項１】
複数の記憶装置にデータおよびパリティを分散して格納する制御を行うことで、データの冗長化を実現するストレージ制御装置であって、
前記複数の記憶装置のうち第１の記憶装置が故障したとき、該第１の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の情報を復元し、第１の予備記憶装置に書き込む第１の再構築手段と、
前記第１の記憶装置の情報を復元している間に第２の記憶装置が故障したとき、該第１および第２の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の未復元領域の情報と第２の記憶装置の情報を復元し、前記第１の予備記憶装置の対応する領域と第２の予備記憶装置にそれぞれ書き込む第２の再構築手段と
を備えることを特徴とするストレージ制御装置。
【請求項２】
前記第２の再構築手段は、前記第１の記憶装置の未復元領域の情報を復元する処理と前記第２の記憶装置の情報を復元する処理をそれぞれ独立に並行して実行することを特徴とする請求項１記載のストレージ制御装置。
【請求項３】
前記第２の再構築手段は、前記第１および第２の記憶装置以外の記憶装置に格納された、前記第１の記憶装置の復元済み領域に対応する領域の情報を用いて、該第２の記憶装置の対応する領域の情報を復元した後、該第１および第２の記憶装置以外の記憶装置に格納された、前記第１の記憶装置の未復元領域に対応する領域の情報を用いて、該第１の記憶装置の未復元領域および該第２の記憶装置の対応する領域の情報を復元することを特徴とする請求項１記載のストレージ制御装置。
【請求項４】
前記第２の再構築手段は、前記第１の記憶装置の復元進捗位置と前記第２の記憶装置の復元進捗位置の差を閾値と比較し、該復元進捗位置の差が該閾値以上であれば、前記第１の記憶装置の未復元領域の情報を復元する処理と該第２の記憶装置の情報を復元する処理をそれぞれ独立に並行して実行し、該復元進捗位置の差が該閾値未満であれば、前記第１および第２の記憶装置以外の記憶装置に格納された、前記第１の記憶装置の復元済み領域に対応する領域の情報を用いて、該第２の記憶装置の対応する領域の情報を復元した後、該第１および第２の記憶装置以外の記憶装置に格納された、前記第１の記憶装置の未復元領域に対応する領域の情報を用いて、該第１の記憶装置の未復元領域および該第２の記憶装置の対応する領域の情報を復元することを特徴とする請求項１記載のストレージ制御装置。
【請求項５】
前記第２の再構築手段は、前記第１および第２の記憶装置以外の記憶装置に格納された、前記第１の記憶装置の未復元領域に対応する領域の情報を用いて、該第１の記憶装置の未復元領域および該第２の記憶装置の対応する領域の情報を復元した後、該第１および第２の記憶装置以外の記憶装置に格納された、該第１の記憶装置の復元済み領域に対応する領域の情報を用いて、該第２の記憶装置の対応する領域の情報を復元することを特徴とする請求項１記載のストレージ制御装置。
【請求項６】
前記第２の再構築手段は、前記第１および第２の記憶装置以外の記憶装置に格納された、前記第１の記憶装置の復元済み領域に対応する領域の情報を用いて、該第２の記憶装置の対応する領域の情報を復元する処理と、該第１および第２の記憶装置以外の記憶装置に格納された、前記第１の記憶装置の未復元領域に対応する領域の情報を用いて、該第１の記憶装置の未復元領域および該第２の記憶装置の対応する領域の情報を復元する処理を、並行して実行することを特徴とする請求項１記載のストレージ制御装置。
【請求項７】
前記第１および第２の記憶装置の所定領域毎に復元済みか否かを表すビットマップ情報

を格納する格納手段をさらに備え、前記第２の再構築手段は、該ビットマップ情報を参照しながら復元済み以外の領域の情報を復元することを特徴とする請求項１乃至６記載のストレージ制御装置。
【請求項８】
前記第２の再構築手段は、前記第１および第２の記憶装置に対するアクセス要求が発生したとき、アクセス対象の情報を復元し、前記ビットマップ情報の該アクセス対象の情報に対応する位置に復元済みと記録することを特徴とする請求項７記載のストレージ制御装置。
【請求項９】
データの冗長化を実現するために、データおよびパリティを分散して格納する複数の記憶装置と、
前記複数の記憶装置のうち第１の記憶装置が故障したとき、該第１の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の情報を復元し、第１の予備記憶装置に書き込む第１の再構築手段と、
前記第１の記憶装置の情報を復元している間に第２の記憶装置が故障したとき、該第１および第２の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の未復元領域の情報と第２の記憶装置の情報を復元し、前記第１の予備記憶装置の対応する領域と第２の予備記憶装置にそれぞれ書き込む第２の再構築手段と
を備えることを特徴とするストレージ装置。
【請求項１０】
複数の記憶装置にデータおよびパリティを分散して格納する制御を行うことで、データの冗長化を実現するプロセッサのためのプログラムであって、
前記複数の記憶装置のうち第１の記憶装置が故障したとき、該第１の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の情報を復元し、第１の予備記憶装置に書き込み、
前記第１の記憶装置の情報を復元している間に第２の記憶装置が故障したとき、該第１および第２の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の未復元領域の情報と第２の記憶装置の情報を復元し、前記第１の予備記憶装置の対応する領域と第２の予備記憶装置にそれぞれ書き込む
処理を前記プロセッサに実行させることを特徴とするプログラム。
【請求項１１】
複数の記憶装置にデータおよびパリティを分散して格納することで、データの冗長化を実現するストレージ制御方法であって、
前記複数の記憶装置のうち第１の記憶装置が故障したとき、該第１の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の情報を復元し、第１の予備記憶装置に書き込み、
前記第１の記憶装置の情報を復元している間に第２の記憶装置が故障したとき、該第１および第２の記憶装置以外の記憶装置に格納された情報を用いて該第１の記憶装置の未復元領域の情報と第２の記憶装置の情報を復元し、前記第１の予備記憶装置の対応する領域と第２の予備記憶装置にそれぞれ書き込む
ことを特徴とするストレージ制御方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ＲＡＩＤ（Redundant Array of Inexpensive Disks）のような複数の記憶装置にデータおよびパリティを分散して格納し、記憶装置の故障時にデータおよびパリティの再構築（Rebuild ）処理を行うストレージ制御装置および方法に関する。
【背景技術】
【０００２】
ＲＡＩＤは、複数のハードディスクを組み合わせて、冗長化された１台のハードディス

クとして管理する技術であり、ディスクへのデータ配置やデータの冗長化の方法により、ＲＡＩＤ０～ＲＡＩＤ６の７つのレベルに分類されている。このうち、ＲＡＩＤ３～ＲＡＩＤ６では、データから生成されたパリティをデータとは別に格納することで、冗長化を実現している。ディスク故障時には、このパリティを用いて故障したディスクのデータを復元する再構築処理が行われる（例えば、特許文献１参照）。

【０００３】

ＲＡＩＤ６は２つのディスクの故障に対応したＲＡＩＤレベルである。ＲＡＩＤ６では、異なる２種類のパリティＰとＱがそれぞれが異なるディスクに分散して格納され、１つのディスクが故障したとき（１ディスク故障）の再構築処理と２つのディスクが故障したとき（２ディスク故障）の再構築処理では、復元方法が異なる。

【０００４】

例えば、図２３に示すように、ディスク１０～１４の５つのディスクからなるＲＡＩＤ装置において、ディスク１０が故障してデータＤ０が失われた場合、予備ディスクであるホットスペア１５を用いて１ディスク故障の再構築処理が行われる。そして、他のディスク１１～１３に格納されたデータＤ１、Ｄ２、およびパリティＰからデータＤ０が復元される。

【０００５】

これに対して、図２４に示すように、ディスク１０および１１が故障してデータＤ０およびＤ１が失われた場合、ホットスペア１５および１６を用いて２ディスク故障の再構築処理が行われる。そして、他のディスク１２～１４に格納されたデータＤ２およびパリティＰ、ＱからデータＤ０およびＤ１が復元される。

【０００６】

一般に、ＲＡＩＤ６ではデータおよびパリティを担当するディスクがストライプ毎に異なるため、故障ディスクに格納された情報およびその復元に必要となる情報の種類もストライプ毎に異なってくる。そこで、以下の説明では、各ディスクに格納された情報をデータ／パリティと記すことにする。

【０００７】

１ディスク故障から２ディスク故障になった場合、再構築処理も１ディスク故障の処理から２ディスク故障の処理へ切り替えて行う必要がある。例えば、図２５に示すように、最初に故障したディスク＃０をホットスペア（ＨＳ）と入れ替えて１ディスク故障の再構築処理を実行している間に、２番目のディスク＃１が故障した場合、１ディスク故障の再構築処理ではデータ／パリティの復元ができなくなる。そこで、１ディスク故障の再構築処理を中止し、ＨＳ＃０およびＨＳ＃１を対象に２ディスク故障の再構築処理を開始することが想定される。

【特許文献１】特開平０３－２４０１２３号公報

【発明の開示】

【発明が解決しようとする課題】

【０００８】

上述した２ディスク故障の再構築方法には、次のような問題がある。

図２５に示した２ディスク故障の再構築処理では、ＨＳ＃０の再構築済み領域へ既に格納された復元データ／パリティが破棄され、再び最初から再構築が実施されることになるため、復元データ／パリティが有効活用されない。

【０００９】

また、１ディスク故障の再構築処理に比べて、より大きな処理コストを要する２ディスク故障の再構築処理を、ＨＳ＃０およびＨＳ＃１の全領域に対して実施するため、冗長性の回復までに多くの時間がかかる。

【００１０】

本発明の課題は、パリティを用いてデータの冗長化を行うＲＡＩＤのようなストレージ装置において、データ／パリティを格納する２つの記憶装置が故障したときの再構築処理を改善して、効率よく冗長性を回復することである。

【課題を解決するための手段】
【００１１】
図１は、本発明のストレージ制御装置の原理図である。図１のストレージ制御装置１０１は、再構築手段１１１および１１２を備え、複数の記憶装置１０２－１～１０２－Ｎにデータおよびパリティを分散して格納する制御を行うことで、データの冗長化を実現する。
【００１２】
再構築手段１１１は、記憶装置１０２－１が故障したとき、それ以外の記憶装置に格納された情報を用いて記憶装置１０２－１の情報を復元し、予備記憶装置１０３－１に書き込む。再構築手段１１２は、記憶装置１０２－１の情報を復元している間に記憶装置１０２－２が故障したとき、記憶装置１０２－１および１０２－２以外の記憶装置に格納された情報を用いて記憶装置１０２－１の未復元領域の情報と記憶装置１０２－２の情報を復元し、予備記憶装置１０３－１の対応する領域と予備記憶装置１０３－２にそれぞれ書き込む。
【００１３】
各記憶装置には、データまたはパリティが情報として格納されている。故障した記憶装置の情報は、その時点で故障していない正常な記憶装置に格納された情報を用いて復元され、対応する予備記憶装置に書き込まれる。記憶装置１０２－１の情報を復元している間に記憶装置１０２－２が故障したとき、記憶装置１０２－１の復元済み領域の情報はそのまま予備記憶装置１０３－１内に保存され、記憶装置１０２－１の未復元領域と記憶装置１０２－２の全領域を対象として復元処理が行われる。
【００１４】
このような再構築制御によれば、１ディスク故障から２ディスク故障になったとき、既に復元されている情報が消去されずに有効利用される。また、記憶装置１０２－１の復元済み領域は復元対象から除外されるため、再構築処理に要する時間が削減される。さらに、記憶装置１０２－１の復元済み領域に対応する記憶装置１０２－２の領域については、処理コストの小さな１ディスク故障の再構築処理を適用できるため、さらなる効率化が可能である。
【００１５】
ストレージ制御装置１０１は、例えば、後述する図２のコントローラ２１１、図２１のホストバスアダプタ１９１１、または図２２のホスト装置２００１に対応する。
【発明の効果】
【００１６】
本発明によれば、複数の記憶装置にデータおよびパリティを分散して格納するストレージ装置において、２つの記憶装置が故障したときの再構築処理が改善され、データの冗長性が効率よく回復する。
【発明を実施するための最良の形態】
【００１７】
以下、図面を参照しながら、本発明を実施するための最良の形態を詳細に説明する。
図２は、実施形態のストレージシステムの構成例を示している。図２のストレージシステムは、ホスト装置２０１およびＲＡＩＤ装置２０２からなり、ＲＡＩＤ装置２０２は、コントローラ２１１およびＤｉｓｋ＃０～Ｄｉｓｋ＃３の４つのディスクを備えるストレージ装置に対応する。
【００１８】
各ディスクは、１つ以上の磁気ディスク装置からなり、ホスト装置２０１は、各ディスクを１つの記憶装置とみなしてデータのリード／ライトを行う。ただし、コントローラ２１１に接続されるディスクの数は４つに限られるわけではなく、一般にはより多くのディスクが接続される。
【００１９】
コントローラ２１１は、プロセッサ２２１、メモリ２２２、およびキャッシュメモリ２

２３を備え、Ｄｉｓｋ＃０～Ｄｉｓｋ＃３の故障時の再構築処理を行う。プロセッサ２２１は、メモリ２２２に格納されたプログラムを実行することにより、キャッシュメモリ２２３をデータバッファとして利用しながら再構築処理を行う。

【００２０】

本実施形態においては、２ディスク故障の再構築処理を改善するために、図３に示す改善案１～６を採用する。まず、図４から図８までを参照しながら、各改善案の概要について説明する。

１．個別実行案（改善案１）

コントローラ２１１は、１番目のディスクの再構築と２番目のディスクの再構築を独立に行う。特に、両方のディスクの故障部分に対応する二重故障部分の再構築処理では、２つのディスクのデータ／パリティをデータバッファ上で復元しながら、一方のディスクの復元データ／パリティのみをディスクにライトし、もう一方のディスクの復元データ／パリティはライトせずに破棄する。

【００２１】

例えば、図４に示すように、最初にディスク＃０が故障し、ＨＳ＃０を用いてディスク＃０のデータ／パリティを復元している間にディスク＃１が故障した場合について説明する。この場合、ディスク＃１が故障するまでの間は、正常なディスク＃１～＃３のデータ／パリティを用いて１ディスク故障の再構築処理によりディスク＃０のデータ／パリティが復元される。

【００２２】

ディスク＃１が故障すると、ＨＳ＃０については、再構築済み領域のデータ／パリティを破棄せずにそのまま保持し、残りの部分のデータ／パリティのみを２ディスク故障の再構築処理により復元する。この処理では、正常なディスク＃２～＃３のデータ／パリティを用いてＨＳ＃０の残りの部分のデータ／パリティが復元される。このとき、ディスク＃１のデータ／パリティも同時に生成されるが、ディスク＃１の再構築処理は別途独立に行われるため、生成されたデータ／パリティは破棄される。

【００２３】

ＨＳ＃１については、ＨＳ＃０の再構築処理と並行して、正常なディスク＃２～＃３のデータ／パリティを用いて２ディスク故障の再構築処理により全体のデータ／パリティを復元する。このとき、２ディスク故障の再構築処理により同時に生成されるディスク＃０のデータ／パリティは破棄される。

【００２４】

このような再構築処理によれば、ディスク＃１が故障したときに、ディスク＃０の再構築処理が、復元対象のディスクのみが故障している１ディスク故障の再構築処理から、復元対象以外にも故障ディスクが存在する２ディスク故障の再構築処理に切り替えられ、復元済み領域のデータ／パリティはそのまま保持される。したがって、図２５に示したように復元済み領域のデータ／パリティを再度復元し直す必要がなく、図２５の再構築処理よりも早くディスク＃０の復元が完了する。

２．同じ進捗位置になるまで待つ案（改善案２）

コントローラ２１１は、２番目のディスクの故障時に、１番目のディスクの再構築処理を一旦停止し、２番目のディスクのみを１番目と同じ進捗位置になるまで再構築する。そして、同じ進捗位置になったら、それ以降は両方同時に再構築する。

【００２５】

例えば、図５に示すように、最初にディスク＃０が故障し、ＨＳ＃０を用いてディスク＃０のデータ／パリティを復元している間にディスク＃１が故障した場合について説明する。ディスク＃１が故障するまでの動作は、図４の場合と同様である。

【００２６】

ディスク＃１が故障すると、最初に、ＨＳ＃０の再構築済み領域に対応するＨＳ＃１のデータ／パリティのみを、ＨＳ＃０の再構築済み領域のデータ／パリティと正常なディスク＃２～＃３のデータ／パリティを用いて１ディスク故障の再構築処理により復元する。

そして、その部分のデータ／パリティが復元されたら、次に、ＨＳ＃０およびＨＳ＃１の残りの部分のデータ／パリティを、正常なディスク＃２～＃３のデータ／パリティを用いて２ディスク故障の再構築処理により同時に復元する。

【００２７】

このような再構築処理によれば、改善案１と同様の利点に加えて、復元中のディスク＃０の復元済み領域のデータ／パリティを他のディスク＃１の復元に有効利用することができ、ディスク＃１の再構築処理が効率化される。

３．混在案（改善案３）

改善案１と改善案２の混在案である。コントローラ２１１は、処理中に２つのディスクＨＳ＃０およびＨＳ＃１の再構築進捗位置をモニタしながら、復元方法を選択する。具体的には、進捗位置の差をチェックし、その差が所定の閾値以上であれば改善案１を適用して、ＨＳ＃０の再構築を優先させ、差が閾値未満であれば改善案２を適用して、互いの進捗位置が同じとなるまでＨＳ＃１の再構築を行う。進捗位置の差のチェックは、所定単位のデータ／パリティを復元する度に毎回行うか、あるいは一定範囲のデータ／パリティを復元する度に行う。

【００２８】

改善案２を適用した場合、ＨＳ＃０およびＨＳ＃１の進捗位置の差は、冗長性が失われている二重故障部分の復元開始を待ち合わせる時間に対応し、進捗位置の差が大きいほど冗長性の回復は遅れることになる。そこで、待ち合わせ時間が一定時間以上となる場合は、それを避けるために改善案１を適用して、ＨＳ＃０およびＨＳ＃１の再構築処理を並行して行う。

【００２９】

ただし、改善案１とは異なり、ＨＳ＃１の二重故障部分以外のデータ／パリティは、ＨＳ＃０の再構築済み領域のデータ／パリティと正常なディスク＃２～＃３のデータ／パリティを用いて１ディスク故障の再構築処理により復元する。一般に、１ディスク故障の再構築処理に比べて、２ディスク故障の再構築処理の方が多くの計算量が必要となる。このため、先行しているＨＳ＃０の２ディスク故障の再構築処理より、ＨＳ＃１の１ディスク故障の再構築処理の方が早く進行し、時間と共に進捗位置の差が縮まってくる可能性が高い。

【００３０】

このような再構築処理によれば、改善案２と同様の利点に加えて、待ち合わせによる冗長性回復の遅れを回避することができる。また、並行処理によりプロセッサ資源等が有効利用されることも期待できる。

４．二重故障部分を先に復元する案（改善案４）

コントローラ２１１は、２番目のディスクの故障時に、１番目のディスクの現在の進捗位置を復元完了位置（後述する停止位置）として保存し、その位置から両方同時の再構築を行う。そして、両方同時の再構築が最後まで終了したら、２番目のディスクの未済領域復旧のため、２番目のディスクの最初から復元完了位置までの再構築を行う。

【００３１】

例えば、図６に示すように、最初にディスク＃０が故障し、ＨＳ＃０を用いてディスク＃０のデータ／パリティを復元している間にディスク＃１が故障した場合について説明する。ディスク＃１が故障するまでの動作は、図４の場合と同様である。

【００３２】

ディスク＃１が故障すると、最初に、ＨＳ＃０の再構築未済領域に対応するＨＳ＃０およびＨＳ＃１のデータ／パリティを、正常なディスク＃２～＃３のデータ／パリティを用いて２ディスク故障の再構築処理により同時に復元する。そして、その部分のデータ／パリティが復元されたら、次に、ＨＳ＃１の残りの部分のデータ／パリティを、ＨＳ＃０の再構築済み領域のデータ／パリティと正常なディスク＃２～＃３のデータ／パリティを用いて１ディスク故障の再構築処理により復元する。

【００３３】

再構築処理中にＲＡＩＤ６の通常のリード／ライトアクセスがあったときは、アクセス対象データの再構築済み／未済をチェックするか、あるいは、リード要求に対してデータを再構築しながら返すというように、通常のリード／ライトアクセスをすべて縮退動作にする。

【００３４】

このような再構築処理によれば、改善案２と同様の利点に加えて、二重故障部分のデータ／パリティを先に復元することで、ＲＡＩＤグループとしての冗長性を短時間で回復することが可能になる。

５．二重故障復元と復元済みＨＳを用いた復元を並行して実施する案（改善案５）

コントローラ２１１は、改善案４において後で実施していた２番目のディスクの最初から復元完了位置までの再構築処理を、二重故障部分の再構築処理を待ち合わせずに並行して実施する。

【００３５】

例えば、図７に示すように、最初にディスク＃０が故障し、ＨＳ＃０を用いてディスク＃０のデータ／パリティを復元している間にディスク＃１が故障した場合について説明する。ディスク＃１が故障するまでの動作は、図４の場合と同様である。

【００３６】

ディスク＃１が故障すると、ＨＳ＃０およびＨＳ＃１の二重故障部分のデータ／パリティを、正常なディスク＃２～＃３のデータ／パリティを用いて２ディスク故障の再構築処理により復元する処理と、ＨＳ＃１の残りの部分のデータ／パリティを、ＨＳ＃０の再構築済み領域のデータ／パリティと正常なディスク＃２～＃３のデータ／パリティを用いて１ディスク故障の再構築処理により復元する処理とを、並行して実施する。

【００３７】

再構築処理中に通常のリード／ライトアクセスがあったときは、改善案４の場合と同様に、アクセス対象データの再構築済み／未済をチェックするか、あるいは縮退動作を行う。

【００３８】

このような再構築処理によれば、ＲＡＩＤグループとしての冗長性の回復は改善案４より遅くなるが、冗長性が失われた二重故障部分と冗長性が残っている他の部分の復元処理が並行して行われるため、全体の復元に要する時間が短縮される。

６．ランダム案（改善案６）

コントローラ２１１は、キャッシュメモリ２２３上のビットマップを用いて、ディスクの所定領域毎に再構築動作を行う。再構築順序としては、上記改善案１～５のいずれかを採用する。基本的には、データ／パリティの保全の観点から改善案４の順序を採用することが望ましいが、他の改善案の順序でも動作可能である。

【００３９】

また、通常のリード／ライトアクセスの延長で再構築動作が行われたときも、ビットマップには再構築済みと記録する。このため、リード／ライトアクセスとは無関係のシーケンシャルな再構築と、リード／ライトアクセスの延長で行われるワンポイント再構築の２種類の処理を併用して、再構築処理が行われる。ワンポイント再構築では、リード要求に対してデータを復元した場合、コントローラ２１１は復元データをディスクの対応位置にライトする。

【００４０】

コントローラ２１１は、再構築対象のディスク毎に、例えば、１ストライプを１ビットに対応付けたビットマップを用意して、進捗状況を管理する。もし、ビットマップが失われた場合は、最初から再構築を開始する。１つの論理ブロックを１ビットに対応付けたビットマップを用いてもよい。

【００４１】

また、最初にビットマップ用のメモリ領域が獲得できなかった場合、コントローラ２１１は以下のいずれかの動作を行う。

・上述の別の改善案で再構築を実行する。
・ビットマップのサイズに上限を設定しておき、これを超えたビットマップを必要とする場合、再構築は資源の獲得を待ってから行う。

【００４２】

コントローラ２１１は、再構築動作を並行して行うことが可能であり、電源オフ／オンを考慮して、ビットマップのバックアップ／リストアの機能も有する。コントローラ２１１が冗長化（二重化）されている場合、ビットマップもコントローラ間で二重化することが基本であるが、二重化しなくてもデータロストに繋がることはない。上述したように、ビットマップが失われた場合は再構築が再始動される。

【００４３】

例えば、図８に示すように、最初にディスク＃０が故障し、ＨＳ＃０を用いてディスク＃０のデータ／パリティを復元している間にディスク＃１が故障した場合について説明する。

【００４４】

コントローラ２１１は、ＨＳ＃０およびＨＳ＃１のそれぞれに対してビットマップを生成し、各ストライプを１ビットのデータで管理する。再構築未済のストライプについては対応するビットに“１”が記録され、再構築済みのストライプについては“０”が記録される。ＨＳ＃０の再構築済み領域の全ストライプについては、ディスク＃１が故障した時点で“０”が記録されており、残りのストライプについては再構築が行われたときに“０”が記録される。

【００４５】

このような再構築処理によれば、改善案１～５と同様の利点に加えて、ワンポイント再構築により復元されたデータ／パリティをディスクに書き戻して復元済み領域として扱うことで、処理が効率化される。

【００４６】

次に、図９から図１９までを参照しながら、上述した各改善案の詳細について説明する。

各改善案の再構築処理は、ディスクの故障を契機として起動されるか、または他の再構築処理により起動される。コントローラ２１１は、契機となったディスクを担当Ｍａｉｎに指定し、必要に応じて他の故障ディスクを担当Ｓｕｂとして追加する。再構築処理では、図９に示すように、担当Ｍａｉｎと担当Ｓｕｂで共通となる現在位置について、復元処理が行われる。ここで、現在位置とは、現在、復元処理が実施されている場所を表す。

【００４７】

また、コントローラ２１１は、ＲＡＩＤ装置２０２を構成する全ディスクの情報を、すべての改善案に共通の制御情報としてキャッシュメモリ２２３に保持する。具体的には、図１０に示すように、各ディスクについて、復元状態（復元済み、復元中、および未復元）、復元先頭位置、および停止位置（必要であれば）の情報が保持される。ただし、正常ディスクについては、すべての領域が復元済みに設定される。

【００４８】

故障ディスクの復元先頭位置は、そのディスクを担当Ｍａｉｎとして実施されている復元処理の現在位置と一致し、復元処理が未実施の場合はそのディスクの端（図９の下端）に設定される。復元処理は、復元先頭位置から上へ向かう方向に進行する。停止位置は、復元処理を停止すべき場所を表す。

【００４９】

復元状態は、ストリップ、論理ブロック等の単位領域毎に管理され、現在位置、復元先頭位置、および停止位置等の位置情報は、その単位領域のアドレスまたはその単位領域が属するストライプの識別子を用いて管理される。

【００５０】

図１１は、再構築処理のフローチャートである。コントローラ２１１は、まず、担当Ｍａｉｎの端（図９の下端）をそのディスクの現在位置として設定し（ステップ１１０１）

、復元ルーチンを実行して復元処理を行う（ステップ１１０２）。この復元処理では、復元対象ディスクの復元対象となるデータ／パリティが他のディスクのデータ／パリティを用いて生成され、対応するホットスペアにライトされる。１回の復元処理では、論理ブロック、ストライプ等の所定単位のデータ／パリティが復元されるが、一般的には、ストライプが所定単位として用いられる。

【００５１】

次に、再構築制御のための復元後処理を行う（ステップ１１０３）。この復元後処理では、復元先頭位置の設定や復元処理を終了するか否かの判定等が行われる。その後、復元後処理において復元処理終了と判定されたか否かをチェックする（ステップ１１０４）。復元処理終了と判定されなければ、現在位置を１ストライプだけ進めて（ステップ１１０５）、ステップ１１０２以降の処理を繰り返し、復元処理終了と判定されれば、再構築処理を終了する。

【００５２】

ステップ１１０２の復元ルーチンおよびステップ１１０３の復元後処理は、改善案毎に異なるため、以下、改善案１～６について順番に説明する。

１．改善案１

改善案１では、図１１の再構築処理は各ディスクの故障を契機として起動され、契機となったディスクが担当Ｍａｉｎに設定されるが、他の故障ディスクが担当Ｓｕｂとして追加されることはない。したがって、２番目のディスクが故障した後は、２つの再構築処理が並行して実行され、各再構築処理における復元対象ディスクは担当Ｍａｉｎのみとなる。

【００５３】

図１２は、改善案１の復元ルーチンのフローチャートである。コントローラ２１１は、まず、担当Ｍａｉｎを復元対象ディスクに設定し、その現在位置を設定して（ステップ１２０１）、故障ディスクの個数をチェックする（ステップ１２０２）。そして、故障ディスクが１個であれば、復元方法を１ディスク故障の再構築に決定する（ステップ１２０３）。

【００５４】

次に、現在位置のストライプに属するデータ／パリティのうち、１ディスク故障の再構築に必要なものを正常ディスクからリードし（ステップ１２０５）、それらがすべてリードできたか否かをチェックする（ステップ１２０６）。データ／パリティがすべてリードできた場合は、それらを用いて同じストライプに属する復元対象ディスクのデータ／パリティを復元し、対応するホットスペアにライトする（ステップ１２０７）。

【００５５】

ステップ１２０６においてリードエラーが発生した場合は、リード対象のディスクが故障したものと判断する。そこで、故障ディスクの個数をチェックし（ステップ１２０８）、それが２個であれば、ステップ１２０２以降の処理を行う。

【００５６】

そして、復元方法を２ディスク故障の再構築に変更し（ステップ１２０４）、２ディスク故障の再構築に必要なデータ／パリティを正常ディスクからリードする（ステップ１２０５）。データ／パリティがすべてリードできた場合は、それらを用いて復元対象ディスクのデータ／パリティを復元し、ホットスペアにライトする（ステップ１２０７）。

【００５７】

故障ディスクが２個のときにさらにリードエラーが発生した場合は、故障ディスクが３個となるため（ステップ１２０８）、復元不可能と判断し、エラー処理を行う（ステップ１２０９）。

【００５８】

図１３は、改善案１の復元後処理のフローチャートである。コントローラ２１１は、まず、復元ルーチンで用いた現在位置を担当Ｍａｉｎの復元先頭位置に設定し（ステップ１３０１）、担当Ｍａｉｎのすべての領域の復元が完了したか否かをチェックする（ステッ

プ１３０２）。ここでは、復元ルーチンで用いた現在位置が担当Ｍａｉｎの端（図９の上端）に達していれば、すべての領域の復元が完了したものと判定される。すべての領域の復元が完了していなければ、復元処理継続と判定し（ステップ１３０３）、すべての領域の復元が完了すれば、復元処理終了と判定する（ステップ１３０４）。

【００５９】

例えば、図４に示したように、最初にディスク＃０が故障すると、ディスク＃０を担当Ｍａｉｎとして再構築処理が起動される。このとき、故障ディスクは１個であるから（図１２のステップ１２０２）、復元方法は１ディスク故障の再構築となり（ステップ１２０３）、正常なディスク＃１～＃３のデータ／パリティのうち、ディスク＃０のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１２０５）。そして、リードされたデータ／パリティを用いてディスク＃０のデータ／パリティが復元され、ＨＳ＃０にライトされる（ステップ１２０７）。

【００６０】

ディスク＃０の現在位置は、その復元先頭位置として設定され（図１３のステップ１３０１）、復元処理継続と判定される（ステップ１３０３）。復元先頭位置は、ディスク＃０に対するリード／ライトアクセス等の他の処理から参照される。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行される（図１１のステップ１１０５）。

【００６１】

次に、ディスク＃１が故障すると、故障ディスクは２個となるから（ステップ１２０２）、復元方法は２ディスク故障の再構築となり（ステップ１２０４）、正常なディスク＃２～＃３のデータ／パリティのうち、ディスク＃０および＃１のデータ／パリティを２ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１２０５）。そして、リードされたデータ／パリティを用いてディスク＃０および＃１のデータ／パリティが復元され、そのうちディスク＃０のデータ／パリティがＨＳ＃０にライトされる（ステップ１２０７）。

【００６２】

復元後処理については、ディスク＃１の故障前と同様である。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃０のすべての領域の復元が完了すれば（ステップ１３０４）、ディスク＃０の再構築処理を終了する（ステップ１１０４）。

【００６３】

さらに、ディスク＃１の故障時には、ディスク＃１を担当Ｍａｉｎとしてもう１つの再構築処理が起動される。このとき、故障ディスクは２個であるから（ステップ１２０２）、復元方法は２ディスク故障の再構築処理となり（ステップ１２０４）、正常なディスク＃２～＃３のデータ／パリティのうち、ディスク＃０および＃１のデータ／パリティを２ディスク故障の再構築処理により復元するために必要なものがリードされる（ステップ１２０５）。

【００６４】

そして、リードされたデータ／パリティを用いてディスク＃０および＃１のデータ／パリティが復元され、そのうちディスク＃１のデータ／パリティがＨＳ＃１にライトされる（ステップ１２０７）。

【００６５】

復元後処理については、ディスク＃０の場合と同様である。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃１のすべての領域の復元が完了すれば（ステップ１３０４）、ディスク＃１の再構築処理を終了する（ステップ１１０４）。

２．改善案２

改善案２では、改善案１と同様に、図１１の再構築処理は各ディスクの故障を契機として起動され、契機となったディスクが担当Ｍａｉｎに設定される。２番目のディスクが故

障すると、１番目の故障ディスクを担当Ｍａｉｎとする再構築処理が中断され、２番目の故障ディスクを担当Ｍａｉｎとする再構築処理が開始される。そして、２番目の故障ディスクの現在位置が１番目と同じ進捗位置に達すると、１番目の故障ディスクが担当Ｓｕｂとして追加される。

【００６６】

図１４は、改善案２の復元ルーチンのフローチャートである。この場合、図１２の復元ルーチンとは異なり、復元対象ディスクは担当Ｍａｉｎと担当Ｓｕｂで表され、２個まで設定できる。また、復元方法は、故障ディスクの個数ではなく、復元対象ディスクの個数に基いて選択される。

【００６７】

コントローラ２１１は、まず、担当Ｍａｉｎ／担当Ｓｕｂを復元対象ディスクに設定し、担当Ｍａｉｎの現在位置を設定する（ステップ１４０１）。このとき、担当Ｓｕｂが設定されていなければ、担当Ｍａｉｎのみが復元対象ディスクに設定される。

【００６８】

次に、復元対象ディスクの個数をチェックし（ステップ１４０２）、復元対象ディスクが１個であれば、復元方法を１ディスク故障の再構築に決定する（ステップ１４０３）。そして、現在位置のストライプに属するデータ／パリティのうち、１ディスク故障の再構築に必要なものを正常ディスクからリードし（ステップ１４０５）、それらがすべてリードできたか否かをチェックする（ステップ１４０６）。データ／パリティがすべてリードできた場合は、それらを用いて同じストライプに属する復元対象ディスクのデータ／パリティを復元し、対応するホットスペアにライトする（ステップ１４０７）。

【００６９】

ステップ１４０６においてリードエラーが発生した場合は、リード対象のディスクが故障したものと判断する。そこで、復元対象ディスクの個数をチェックし（ステップ１４０８）、それが１個であれば、故障ディスクを復元対象ディスクに追加して（ステップ１４１０）、ステップ１４０２以降の処理を行う。

【００７０】

そして、復元方法を２ディスク故障の再構築に変更し（ステップ１４０４）、２ディスク故障の再構築に必要なデータ／パリティを正常ディスクからリードする（ステップ１４０５）。データ／パリティがすべてリードできた場合は、それらを用いて２個の復元対象ディスクのデータ／パリティを復元し、それぞれ対応するホットスペアにライトする（ステップ１４０７）。

【００７１】

復元対象ディスクが２個のときにさらにリードエラーが発生した場合は、故障ディスクが３個となるため（ステップ１４０８）、復元不可能と判断し、エラー処理を行う（ステップ１４０９）。

【００７２】

図１５は、改善案２の復元後処理のフローチャートである。コントローラ２１１は、まず、復元ルーチン終了時の担当Ｍａｉｎの現在位置を担当Ｍａｉｎ／Ｓｕｂの復元先頭位置に設定し（ステップ１５０１）、以下の条件ａが満たされるか否かをチェックする（ステップ１５０２）。

条件ａ：担当Ｍａｉｎ以外に他の故障ディスクが存在し、担当Ｍａｉｎおよびその故障ディスクのいずれにも停止位置が設定されておらず、かつ、その故障ディスクの復元先頭位置が担当Ｍａｉｎのそれより後ろ（下方）である。

【００７３】

他の故障ディスクの復元先頭位置が担当Ｍａｉｎより後ろの場合、担当Ｍａｉｎより復元処理が遅れていることを意味する。条件ａが満たされれば、担当Ｍａｉｎの復元先頭位置を他の故障ディスクの停止位置に設定し（ステップ１５０６）、担当Ｍａｉｎの再構築処理を中断するために復元処理終了と判定する（ステップ１５０８）。

【００７４】

条件ａが満たされなければ、次に、担当Ｍａｉｎのすべての領域の復元が完了したか否かをチェックする（ステップ１５０３）。すべての領域の復元が完了していれば、復元処理終了と判定する（ステップ１５０８）。

【００７５】

すべての領域の復元が完了していなければ、次に、担当Ｍａｉｎに停止位置が設定されており、かつ、担当Ｍａｉｎの現在位置がその停止位置であるか否かをチェックする（ステップ１５０４）。現在位置が停止位置であれば、他の故障ディスクを担当Ｓｕｂに追加し（ステップ１５０７）、復元処理継続と判定する（ステップ１５０５）。

【００７６】

現在位置が停止位置でない場合、および、停止位置が設定されていない場合は、そのまま復元処理継続と判定する（ステップ１５０５）。

例えば、図５に示したように、最初にディスク＃０が故障すると、ディスク＃０を担当Ｍａｉｎとして再構築処理が起動される。このとき、復元対象ディスクは１個であるから（図１４のステップ１４０１）、復元方法は１ディスク故障の再構築となり（ステップ１４０３）、正常なディスク＃１～＃３のデータ／パリティのうち、ディスク＃０のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１４０５）。そして、リードされたデータ／パリティを用いてディスク＃０のデータ／パリティが復元され、ＨＳ＃０にライトされる（ステップ１４０７）。

【００７７】

ディスク＃０の現在位置は、その復元先頭位置として設定され（図１５のステップ１５０１）、他の故障ディスクはないので（ステップ１５０２）、復元処理継続と判定される（ステップ１５０５）。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行される（図１１のステップ１１０５）。

【００７８】

次に、ディスク＃１が故障すると、担当Ｍａｉｎであるディスク＃０の復元先頭位置はその現在位置に一致しており（ステップ１５０１）、他の故障ディスクであるディスク＃１の復元先頭位置はその下端に一致しているため、条件ａが満たされる（ステップ１５０２）。そこで、ディスク＃０の復元先頭位置がディスク＃１の停止位置に設定され（ステップ１５０６）、復元処理終了と判定される（ステップ１５０８）。これにより、ディスク＃０を担当Ｍａｉｎとする再構築処理が中断される（ステップ１１０４）。

【００７９】

このとき、ディスク＃１を担当Ｍａｉｎとして別の再構築処理が起動される。復元対象ディスクは１個であるから（ステップ１４０１）、復元方法は１ディスク故障の再構築となり（ステップ１４０３）、ディスク＃０、＃２、および＃３のデータ／パリティのうち、ディスク＃１のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１４０５）。ただし、ディスク＃０については、ＨＳ＃０にライトされた復元済みのデータ／パリティがリードされる。

【００８０】

そして、リードされたデータ／パリティを用いてディスク＃１のデータ／パリティが復元され、ＨＳ＃１にライトされる（ステップ１４０７）。

ディスク＃１の現在位置は、その復元先頭位置として設定され（ステップ１５０１）、ディスク＃１には既に停止位置が設定されているため、条件ａは満たされない（ステップ１５０２）。また、ディスク＃１の現在位置はその停止位置に到達していないので（ステップ１５０４）、復元処理継続と判定される（ステップ１５０５）。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行される（ステップ１１０５）。

【００８１】

そして、ディスク＃１の現在位置がその停止位置に到達すると（ステップ１５０４）、処理が中断していたディスク＃０が担当Ｓｕｂとして追加され（ステップ１５０７）、復元処理継続と判定される（ステップ１５０５）。これにより、現在位置が更新される（ステップ１１０５）。

【００８２】
これにより、復元対象ディスクは２個となるから（ステップ１４０１）、復元方法は２ディスク故障の再構築処理となり（ステップ１４０４）、正常なディスク＃２～＃３のデータ／パリティのうち、ディスク＃０および＃１のデータ／パリティを２ディスク故障の再構築処理により復元するために必要なものがリードされる（ステップ１４０５）。そして、リードされたデータ／パリティを用いてディスク＃０および＃１のデータ／パリティが復元され、それぞれＨＳ＃０およびＨＳ＃１にライトされる（ステップ１４０７）。
【００８３】
ディスク＃１の現在位置は、ディスク＃０および＃１の復元先頭位置として設定され（ステップ１５０１）、ディスク＃１の現在位置はその停止位置を越えているため（ステップ１５０４）、復元処理継続と判定される（ステップ１５０５）。
【００８４】
このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃１のすべての領域の復元が完了すれば（ステップ１５０８）、ディスク＃１を担当Ｍａｉｎとする再構築処理を終了する（ステップ１１０４）。この時点で、担当Ｓｕｂであるディスク＃０の現在位置もその上端に達しているため、ディスク＃０の復元も同時に完了する。
３．改善案３
改善案３では、改善案２と同様に、図１１の再構築処理は各ディスクの故障を契機として起動され、契機となったディスクが担当Ｍａｉｎに設定される。２番目のディスクが故障すると、２つの故障ディスクの進捗位置の差に応じて改善案１または２が選択される。
【００８５】
進捗位置の差が閾値以上であれば改善案１が選択され、２つの再構築処理が並行して実行される。ただし、改善案１とは異なり、２番目の故障ディスクのデータ／パリティは、１番目の故障ディスクの復元済みのデータ／パリティと正常なディスクのデータ／パリティを用いて１ディスク故障の再構築処理により復元される。
【００８６】
そして、進捗位置の差が閾値未満になると改善案２が選択され、１番目の故障ディスクを担当Ｍａｉｎとする再構築処理が中断され、２番目の故障ディスクを担当Ｍａｉｎとする再構築処理が開始される。そして、２番目の故障ディスクの現在位置が１番目と同じ進捗位置に達すると、１番目の故障ディスクが担当Ｓｕｂとして追加される。
【００８７】
改善案３の復元ルーチンのフローチャートは改善案２と同様であり、復元後処理のフローチャートは図１６のようになる。図１６の復元後処理は、図１５の復元後処理にステップ１６０３の判定を追加した構成を有する。
【００８８】
ステップ１６０３において、コントローラ２１１は、他の故障ディスクと担当Ｍａｉｎの復元先頭位置の差を閾値と比較する。そして、復元先頭位置の差が閾値未満であれば、担当Ｍａｉｎの復元先頭位置を他の故障ディスクの停止位置に設定して（ステップ１６０７）、復元処理終了と判定する（ステップ１６０９）。また、復元先頭位置の差が閾値以上であれば、ステップ１６０４以降の処理を行う。
【００８９】
したがって、改善案２で説明した条件ａが満たされ、かつ、他の故障ディスクと担当Ｍａｉｎの復元先頭位置の差が閾値未満であれば、他の故障ディスクに停止位置が設定され（ステップ１６０７）、それ以外の場合は停止位置は設定されない。
【００９０】
例えば、図５に示したように、最初にディスク＃０が故障すると、ディスク＃０を担当Ｍａｉｎとして再構築処理が起動され、ディスク＃１が故障するまで改善案２と同様の処理が行われる。
【００９１】

次に、ディスク＃１が故障すると、条件ａが満たされるため（図１６のステップ１６０２）、ディスク＃０および＃１の復元先頭位置の差が閾値と比較される（ステップ１６０３）。ここで、ディスク＃０が故障してから十分に時間が経過している場合は、ディスク＃０の復元処理がかなり進行しており、復元先頭位置の差は閾値を超えていると考えられる。この場合、ステップ１６０４以降の処理が行われ、ディスク＃０には停止位置が設定されていないので（ステップ１６０５）、復元処理継続と判定される（ステップ１６０６）。

【００９２】

このとき、復元対象ディスクは１個であるから（図１４のステップ１４０１）、復元方法は１ディスク故障の再構築となり（ステップ１４０３）、ディスク＃１～＃３のデータ／パリティのうち、ディスク＃０のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１４０５）。

【００９３】

ここで、ディスク＃１のデータ／パリティがリードされた場合、リードエラーが発生し、ディスク＃１が復元対象ディスクに追加されて（ステップ１４１０）、復元対象ディスクは２個となる。

【００９４】

このため、復元方法は２ディスク故障の再構築処理に変更され（ステップ１４０４）、正常なディスク＃２～＃３のデータ／パリティのうち、ディスク＃０および＃１のデータ／パリティを２ディスク故障の再構築処理により復元するために必要なものがリードされる（ステップ１４０５）。そして、リードされたデータ／パリティを用いてディスク＃０および＃１のデータ／パリティが復元され、そのうちディスク＃０のデータ／パリティがＨＳ＃０にライトされる（ステップ１４０７）。

【００９５】

ディスク＃０の現在位置は、その復元先頭位置として設定され（ステップ１６０１）、条件ａが満たされるが（ステップ１６０２）、復元先頭位置の差はまだ閾値より大きい（ステップ１６０３）。また、ディスク＃０には停止位置が設定されていないので（ステップ１６０５）、復元処理継続と判定される（ステップ１６０６）。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行される（図１１のステップ１１０５）。

【００９６】

さらに、ディスク＃１の故障時には、ディスク＃１を担当Ｍａｉｎとしてもう１つの再構築処理が起動される。復元対象ディスクは１個であるから（ステップ１４０１）、復元方法は１ディスク故障の再構築となり（ステップ１４０３）、ディスク＃０、＃２、および＃３のデータ／パリティのうち、ディスク＃１のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１４０５）。ただし、ディスク＃０については、ＨＳ＃０にライトされた復元済みのデータ／パリティがリードされる。

【００９７】

そして、リードされたデータ／パリティを用いてディスク＃１のデータ／パリティが復元され、ＨＳ＃１にライトされる（ステップ１４０７）。

ディスク＃１の現在位置は、その復元先頭位置として設定され（ステップ１６０１）、他の故障ディスクであるディスク＃０の復元先頭位置は担当Ｍａｉｎであるディスク＃１の復元先頭位置より前にあるため、条件ａは満たされない（ステップ１６０２）。また、ディスク＃１には停止位置が設定されていないので（ステップ１６０５）、復元処理継続と判定される（ステップ１６０６）。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行される（ステップ１１０５）。

【００９８】

こうして、ディスク＃０を担当Ｍａｉｎとする２ディスク故障の再構築処理と、ディスク＃１を担当Ｍａｉｎとする１ディスク故障の再構築処理とが並行して実行され、ディス

ク＃１の復元先頭位置が徐々にディスク＃０の復元先頭位置に接近してくる。

【００９９】

そして、ディスク＃０を担当Ｍａｉｎとする再構築処理において、ディスク＃０および＃１の復元先頭位置の差が閾値未満になると（ステップ１６０３）、ディスク＃０の復元先頭位置がディスク＃１の停止位置に設定され（ステップ１６０７）、復元処理終了と判定される（ステップ１６０９）。これにより、ディスク＃０を担当Ｍａｉｎとする再構築処理が中断される（ステップ１１０４）。

【０１００】

その後、ディスク＃１を担当Ｍａｉｎとする再構築処理のみが継続されるが、ディスク＃１には既に停止位置が設定されているため、やはり条件ａは満たされない（ステップ１６０２）。また、ディスク＃１の現在位置はその停止位置に到達していないので（ステップ１６０５）、復元処理継続と判定される（ステップ１６０６）。

【０１０１】

そして、ディスク＃１の現在位置がその停止位置に到達すれば（ステップ１６０５）、処理が中断していたディスク＃０が担当Ｓｕｂとして追加され（ステップ１６０８）、復元処理継続と判定される（ステップ１６０６）。これにより、現在位置が更新される（ステップ１１０５）。

【０１０２】

これにより、復元対象ディスクは２個となるから（ステップ１４０１）、復元方法は２ディスク故障の再構築処理となり（ステップ１４０４）、正常なディスク＃２～＃３のデータ／パリティのうち、ディスク＃０および＃１のデータ／パリティを２ディスク故障の再構築処理により復元するために必要なものがリードされる（ステップ１４０５）。そして、リードされたデータ／パリティを用いてディスク＃０および＃１のデータ／パリティが復元され、それぞれＨＳ＃０およびＨＳ＃１にライトされる（ステップ１４０７）。

【０１０３】

ディスク＃１の現在位置は、ディスク＃０および＃１の復元先頭位置として設定され（ステップ１６０１）、ディスク＃１の現在位置はその停止位置を越えているため（ステップ１６０５）、復元処理継続と判定される（ステップ１６０６）。

【０１０４】

このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃１のすべての領域の復元が完了すれば（ステップ１６０９）、ディスク＃１を担当Ｍａｉｎとする再構築処理を終了する（ステップ１１０４）。この時点で、担当Ｓｕｂであるディスク＃０の現在位置もその上端に達しているため、ディスク＃０の復元も同時に完了する。

４．改善案４

改善案４では、図１１の再構築処理はディスクの故障を契機として起動されるか、あるいは他の再構築処理から起動され、契機となったディスクが担当Ｍａｉｎに設定される。前者の契機では、ＲＡＩＤグループあたり１つの再構築処理のみが起動される。したがって、２番目のディスクが故障しても、既にＲＡＩＤグループとして再構築処理が起動済みであれば、新たな再構築処理は起動されない。

【０１０５】

２番目のディスクが故障すると、１番目の故障ディスクの現在位置が２番目の故障ディスクの停止位置として設定され、２番目の故障ディスクが担当Ｓｕｂとして追加されて、再構築処理が続行される。そして、１番目の故障ディスクの復元が完了すると、２番目の故障ディスクを担当Ｍａｉｎとして、その下端から停止位置までの再構築処理が行われる。改善案４の復元ルーチンのフローチャートは改善案２と同様である。

【０１０６】

図１７は、改善案４の復元後処理のフローチャートである。コントローラ２１１は、まず、復元ルーチン終了時の担当Ｍａｉｎの現在位置を担当Ｍａｉｎ／Ｓｕｂの復元先頭位置に設定し（ステップ１７０１）、以下の条件ｂが満たされるか否かをチェックする（ス

テップ１７０２）。
条件ｂ：担当Ｍａｉｎ以外に他の故障ディスクが存在し、担当Ｍａｉｎおよびその故障ディスクのいずれにも停止位置が設定されていない。
【０１０７】
条件ｂが満たされれば、担当Ｍａｉｎの復元先頭位置を他の故障ディスクの停止位置に設定し、その故障ディスクを担当Ｓｕｂに追加する（ステップ１７０６）。そして、担当Ｍａｉｎのすべての領域の復元が完了したか否かをチェックする（ステップ１７０３）。条件ｂが満たされなければ、そのままステップ１７０３の処理を行う。
【０１０８】
すべての領域の復元が完了していれば、他の故障ディスクがあるか否かをチェックし、そのような故障ディスクがあれば、それを担当Ｍａｉｎとする別の再構築処理を起動する（ステップ１７０７）。そして、復元処理終了と判定する（ステップ１７０８）。他の故障ディスクがなければ、別の再構築処理を起動することなく復元処理終了と判定する（ステップ１７０８）。
【０１０９】
すべての領域の復元が完了していなければ、次に、担当Ｍａｉｎの現在位置がその停止位置である否かをチェックする（ステップ１７０４）。現在位置が停止位置であれば、復元処理終了と判定する（ステップ１７０８）。
【０１１０】
現在位置が停止位置でない場合、および、停止位置が設定されていない場合は、そのまま復元処理継続と判定する（ステップ１７０５）。
例えば、図６に示したように、最初にディスク＃０が故障すると、ディスク＃０を担当Ｍａｉｎとして再構築処理が起動される。このとき、復元対象ディスクは１個であるから（図１４のステップ１４０１）、復元方法は１ディスク故障の再構築となり（ステップ１４０３）、正常なディスク＃１～＃３のデータ／パリティのうち、ディスク＃０のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１４０５）。そして、リードされたデータ／パリティを用いてディスク＃０のデータ／パリティが復元され、ＨＳ＃０にライトされる（ステップ１４０７）。
【０１１１】
ディスク＃０の現在位置は、その復元先頭位置として設定され（図１７のステップ１７０１）、他の故障ディスクはないので（ステップ１７０２）、復元処理継続と判定される（ステップ１７０５）。このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行される（図１１のステップ１１０５）。
【０１１２】
次に、ディスク＃１が故障すると、条件ｂが満たされ（ステップ１７０２）、ディスク＃０の復元先頭位置がディスク＃１の停止位置に設定され、ディスク＃１が担当Ｓｕｂとして追加される（ステップ１７０６）。しかし、ディスク＃０には停止位置が設定されていないので（ステップ１７０４）、復元処理継続と判定される（ステップ１７０５）。
【０１１３】
このとき、復元対象ディスクは２個となるから（ステップ１４０１）、復元方法は２ディスク故障の再構築処理となり（ステップ１４０４）、正常なディスク＃２～＃３のデータ／パリティのうち、ディスク＃０および＃１のデータ／パリティを２ディスク故障の再構築処理により復元するために必要なものがリードされる（ステップ１４０５）。そして、リードされたデータ／パリティを用いてディスク＃０および＃１のデータ／パリティが復元され、それぞれＨＳ＃０およびＨＳ＃１にライトされる（ステップ１４０７）。
【０１１４】
ディスク＃０の現在位置は、ディスク＃０および＃１の復元先頭位置として設定され（ステップ１７０１）、ディスク＃１には既に停止位置が設定されているため、条件ｂは満たされない（ステップ１７０２）。また、ディスク＃０には停止位置が設定されていないので（ステップ１７０４）、復元処理継続と判定される（ステップ１７０５）。

【０１１５】
このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃０のすべての領域の復元が完了すれば（ステップ１７０３）、ディスク＃１を担当Ｍａｉｎとする別の再構築処理が起動され（ステップ１７０７）、復元処理終了と判定される（ステップ１７０８）。これにより、ディスク＃０を担当Ｍａｉｎとする再構築処理が終了する（ステップ１１０４）。この時点で、担当Ｓｕｂであるディスク＃１の現在位置は、その上端に達している。

【０１１６】
次に、ディスク＃１を担当Ｍａｉｎとする再構築処理において、ディスク＃１の下端が現在位置として設定される（ステップ１１０１）。このとき、復元対象ディスクは１個であるから（ステップ１４０１）、復元方法は１ディスク故障の再構築となり（ステップ１４０３）、ディスク＃０、＃２、および＃３のデータ／パリティのうち、ディスク＃１のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１４０５）。ただし、ディスク＃０については、ＨＳ＃０にライトされた復元済みのデータ／パリティがリードされる。

【０１１７】
そして、リードされたデータ／パリティを用いてディスク＃１のデータ／パリティが復元され、ＨＳ＃１にライトされる（ステップ１４０７）。
ディスク＃１の現在位置は、その復元先頭位置として設定され（ステップ１７０１）、ディスク＃１には既に停止位置が設定されているため、条件ｂは満たされない（ステップ１７０２）。また、ディスク＃１の現在位置はその停止位置に到達していないので（ステップ１７０４）、復元処理継続と判定される（ステップ１７０５）。

【０１１８】
このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃１の現在位置が停止位置に到達する。このとき、ディスク＃１の現在位置はその上端に達していないため、すべての領域の復元が完了していないと判定される（ステップ１７０３）。しかし、現在位置が停止位置に一致するので（ステップ１７０４）、復元処理終了と判定される（ステップ１７０８）。これにより、ディスク＃１を担当Ｍａｉｎとする再構築処理が終了し（ステップ１１０４）、ディスク＃１の復元が完了する。

５．改善案５
改善案５では、改善案１と同様に、図１１の再構築処理は各ディスクの故障を契機として起動され、契機となったディスクが担当Ｍａｉｎに設定される。

【０１１９】
２番目のディスクが故障すると、改善案４と同様に、１番目の故障ディスクの現在位置が２番目の故障ディスクの停止位置として設定され、２番目の故障ディスクが担当Ｓｕｂとして追加されて、再構築処理が続行される。それと同時に、２番目の故障ディスクを担当Ｍａｉｎとする再構築処理が起動され、１番目の故障ディスクを担当Ｍａｉｎとする再構築処理と並行して実行される。

【０１２０】
改善案５の復元ルーチンのフローチャートは改善案２と同様であり、復元後処理のフローチャートは図１８のようになる。図１８の復元後処理は、図１７の復元後処理からステップ１７０７の処理を除いた構成を有する。

【０１２１】
例えば、図７に示したように、最初にディスク＃０が故障すると、ディスク＃０を担当Ｍａｉｎとして再構築処理が起動され、ディスク＃１が故障するまで改善案４と同様の処理が行われる。

【０１２２】
次に、ディスク＃１が故障すると、条件ｂが満たされ（ステップ１８０２）、ディスク＃０の復元先頭位置がディスク＃１の停止位置に設定され、ディスク＃１が担当Ｓｕｂと

して追加される（ステップ１８０６）。しかし、ディスク＃０には停止位置が設定されていないので（ステップ１８０４）、復元処理継続と判定される（ステップ１８０５）。

【０１２３】

このとき、復元対象ディスクは２個となるから（ステップ１４０１）、復元方法は２ディスク故障の再構築処理となり（ステップ１４０４）、正常なディスク＃２～＃３のデータ／パリティのうち、ディスク＃０および＃１のデータ／パリティを２ディスク故障の再構築処理により復元するために必要なものがリードされる（ステップ１４０５）。そして、リードされたデータ／パリティを用いてディスク＃０および＃１のデータ／パリティが復元され、それぞれＨＳ＃０およびＨＳ＃１にライトされる（ステップ１４０７）。

【０１２４】

ディスク＃０の現在位置は、ディスク＃０および＃１の復元先頭位置として設定され（ステップ１８０１）、ディスク＃１には既に停止位置が設定されているため、条件ｂは満たされない（ステップ１８０２）。また、ディスク＃０には停止位置が設定されていないので（ステップ１８０４）、復元処理継続と判定される（ステップ１８０５）。

【０１２５】

このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃０のすべての領域の復元が完了すれば（ステップ１８０３）、復元処理終了と判定される（ステップ１８０７）。これにより、ディスク＃０を担当Ｍａｉｎとする再構築処理が終了する（ステップ１１０４）。この時点で、担当Ｓｕｂであるディスク＃１の現在位置は、その上端に達している。

【０１２６】

さらに、ディスク＃１の故障時には、ディスク＃１を担当Ｍａｉｎとしてもう１つの再構築処理が起動され、ディスク＃１の下端が現在位置として設定される（ステップ１１０１）。このとき、復元対象ディスクは１個であるから（ステップ１４０１）、復元方法は１ディスク故障の再構築となり（ステップ１４０３）、ディスク＃０、＃２、および＃３のデータ／パリティのうち、ディスク＃１のデータ／パリティを１ディスク故障の再構築により復元するために必要なものがリードされる（ステップ１４０５）。ただし、ディスク＃０については、ＨＳ＃０にライトされた復元済みのデータ／パリティがリードされる。

【０１２７】

そして、リードされたデータ／パリティを用いてディスク＃１のデータ／パリティが復元され、ＨＳ＃１にライトされる（ステップ１４０７）。

ディスク＃１の現在位置は、その復元先頭位置として設定され（ステップ１８０１）、ディスク＃１には既に停止位置が設定されているため、条件ｂは満たされない（ステップ１８０２）。また、ディスク＃１の現在位置はその停止位置に到達していないので（ステップ１８０４）、復元処理継続と判定される（ステップ１８０５）。

【０１２８】

このような復元ルーチンおよび復元後処理が１ストライプ毎に繰り返し実行され（ステップ１１０５）、ディスク＃１の現在位置が停止位置に到達する。このとき、ディスク＃１の現在位置はその上端に達していないため、すべての領域の復元が完了していないと判定される（ステップ１８０３）。しかし、現在位置が停止位置に一致するので（ステップ１８０４）、復元処理終了と判定される（ステップ１８０７）。これにより、ディスク＃１を担当Ｍａｉｎとする再構築処理が終了する（ステップ１１０４）。

【０１２９】

ディスク＃０を担当Ｍａｉｎとする２ディスク故障の再構築処理と、ディスク＃１を担当Ｍａｉｎとする１ディスク故障の再構築処理は並行して実行され、両方の再構築処理が終了した時点でディスク＃１の復元が完了する。

６．改善案６

改善案１～５の再構築処理において、各ディスクのストリップや論理ブロックのような所定領域毎にその復元状態を表すビットマップを、制御情報として追加する。ディスク全

体に対する進捗の制御は、改善案１～５により行われる。

【０１３０】

コントローラ２１１は、再構築処理または復元ルーチンの実行時に、ビットマップ中の復元位置に対応するビット情報を参照する。そして、図１９に示すように、リード／ライトアクセスの延長等により既に復元済みであれば、復元処理をスキップする。これにより、余計な復元処理のコストを削減することができる。

【０１３１】

図２０は、コントローラ２１１のプロセッサ２２１が処理に用いるプログラムおよびデータの提供方法を示している。情報処理装置等の外部装置１８０１や可搬記録媒体１８０２に格納されたプログラムおよびデータは、ＲＡＩＤ装置２０２のメモリ２２２にロードされる。

【０１３２】

外部装置１８０１は、そのプログラムおよびデータを搬送する搬送信号を生成し、通信ネットワーク上の任意の伝送媒体を介してＲＡＩＤ装置２０２に送信する。可搬記録媒体１８０２は、メモリカード、フレキシブルディスク、光ディスク、光磁気ディスク等の任意のコンピュータ読み取り可能な記録媒体である。プロセッサ２２１は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【０１３３】

図２１および２２は、ストレージシステムの別の構成例を示している。図２１は、ホスト装置に実装されたホストバスアダプタが再構築処理を行う例を示しており、図２２は、ホスト装置に実装されたソフトウェアが再構築処理を行う例を示している。いずれの構成においても、必要なプログラムおよびデータは、ＲＡＩＤ装置２０２の場合と同様にして提供される。

【０１３４】

図２１のストレージシステムは、ホスト装置１９０１およびＤｉｓｋ＃０～Ｄｉｓｋ＃３からなり、ホスト装置１９０１は、ホストバスアダプタ１９１１を備える。ホストバスアダプタ１９１１は、プロセッサ１９２１、メモリ１９２２、およびキャッシュメモリ１９２３を備え、Ｄｉｓｋ＃０～Ｄｉｓｋ＃３の故障時の再構築処理を行う。このとき、プロセッサ１９２１は、メモリ１９２２に格納されたプログラムを実行することにより、上述した再構築処理を行う。

【０１３５】

図２２のストレージシステムは、ホスト装置２００１およびＤｉｓｋ＃０～Ｄｉｓｋ＃３からなる。ホスト装置２００１は、プロセッサ２０１１およびメモリ２０１２、２０１３を備え、Ｄｉｓｋ＃０～Ｄｉｓｋ＃３の故障時の再構築処理を行う。このとき、プロセッサ２０１１は、メモリ２０１２に格納されたプログラムを実行することにより、メモリ２０１３上で上述した再構築処理を行う。

【０１３６】

なお、以上の実施形態では、ディスク装置として磁気ディスク装置が用いられているが、本発明は、光ディスク装置、光磁気ディスク装置等の他のディスク装置や、テープ装置のような他の記憶装置を用いたストレージシステムに対しても、適用可能である。

【図面の簡単な説明】

【０１３７】

【図１】本発明のストレージ制御装置の原理図である。

【図２】第１のストレージシステムの構成図である。

【図３】再構築処理の改善案を示す図である。

【図４】改善案１を示す図である。

【図５】改善案２を示す図である。

【図６】改善案４を示す図である。

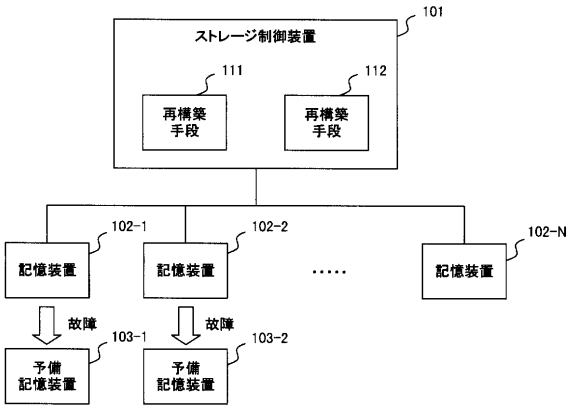
【図７】改善案５を示す図である。

【図８】改善案６を示す図である。

【図９】ＭａｉｎとＳｕｂの現在位置を示す図である。
【図１０】各ディスクの復元状態を示す図である。
【図１１】再構築処理のフローチャートである。
【図１２】改善案１の復元ルーチンのフローチャートである。
【図１３】改善案１の復元後処理のフローチャートである。
【図１４】改善案２～５の復元ルーチンのフローチャートである。
【図１５】改善案２の復元後処理のフローチャートである。
【図１６】改善案３の復元後処理のフローチャートである。
【図１７】改善案４の復元後処理のフローチャートである。
【図１８】改善案５の復元後処理のフローチャートである。
【図１９】改善案６の復元状態を示す図である。
【図２０】プログラムおよびデータの提供方法を示す図である。
【図２１】第２のストレージシステムの構成図である。
【図２２】第３のストレージシステムの構成図である。
【図２３】１ディスク故障のデータ復元を示す図である。
【図２４】２ディスク故障のデータ復元を示す図である。
【図２５】２ディスク故障の再構築処理を示す図である。
【符号の説明】
【０１３８】
１０１　ストレージ制御装置
１０２－１、１０２－２、１０２－Ｎ　記憶装置
１０３－１、１０３－２　予備記憶装置
１１１、１１２　再構築手段
２０１、１９０１、２００１　ホスト装置
２０２　ＲＡＩＤ装置
２１１　コントローラ
２１１、１９２１、２０１１　プロセッサ
２２２、１９２２、２０１２、２０１３　メモリ
２２３、１９２３　キャッシュメモリ
１８０１　外部装置
１８０２　可搬記録媒体
１９１１　ホストバスアダプタ
Ｄｉｓｋ０、Ｄｉｓｋ１、Ｄｉｓｋ２、Ｄｉｓｋ３　ディスク

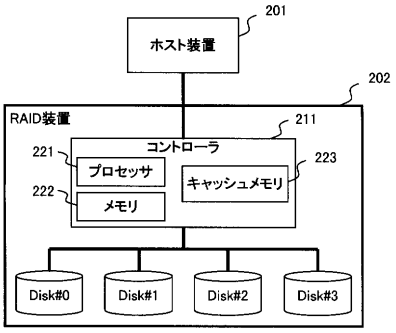
【図１】

本発明のストレージ制御装置の原理図



【図２】

第1のストレージシステムの構成図



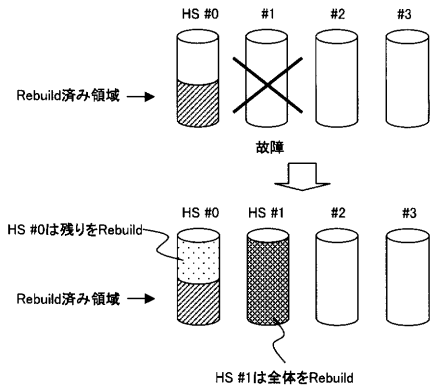
【図３】

再構築処理の改善案を示す図

	方式	進捗管理方法	工数・難易度
1	個別実行案	現状通り	容易
2	同じ進捗位置になるまで待つ案	現状通り	容易
3	混在案	現状通り	中度
4	二重故障部分を先に復元する案	現状に加えて、復元完了位置を管理	通常処理に影響あり
5	二重故障復元と、復元済みHSを用いた復元を並行して実施する案	現状に加えて、復元完了位置を管理	
6	ランダム案	ビットマップ管理	工数がかかる 通常処理に影響あり

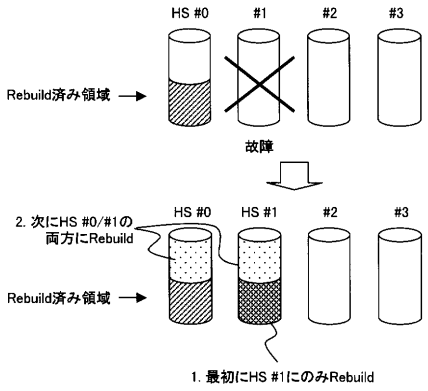
【図４】

改善案1を示す図



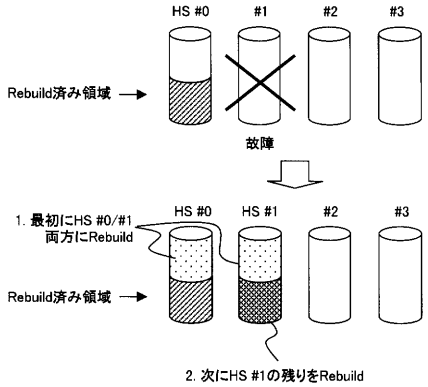
【図５】

改善案2を示す図



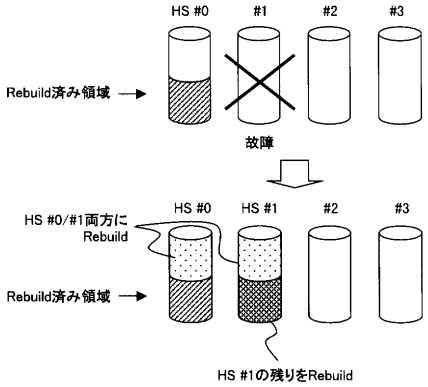
【図６】

改善案4を示す図



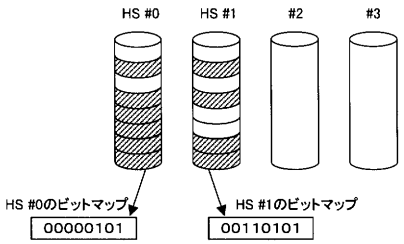
【図７】

改善案5を示す図



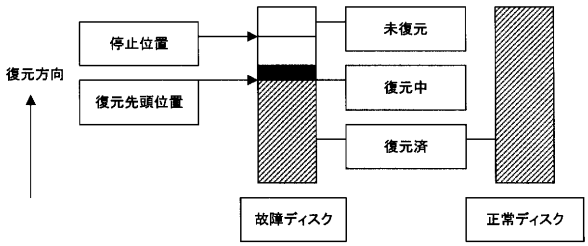
【図８】

改善案6を示す図



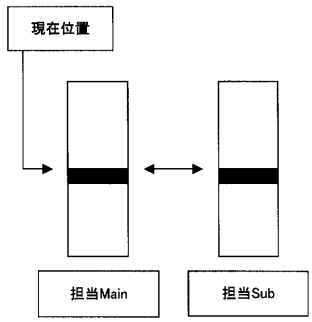
【図１０】

各ディスクの復元状態を示す図



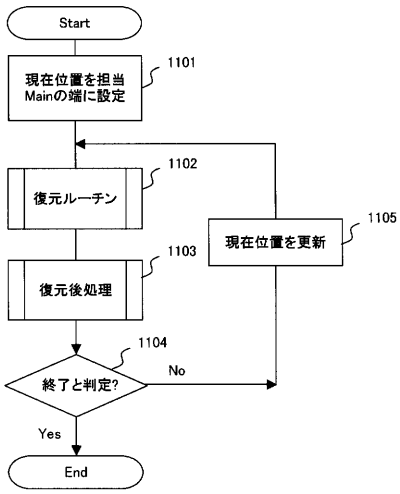
【図９】

MainとSubの現在位置を示す図



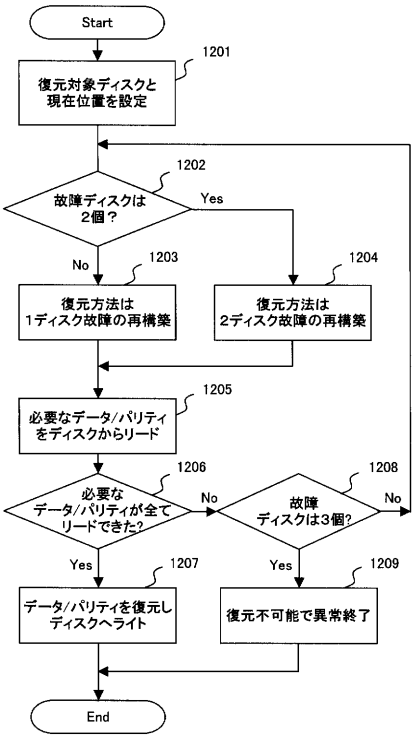
【図１１】

再構築処理のフローチャート



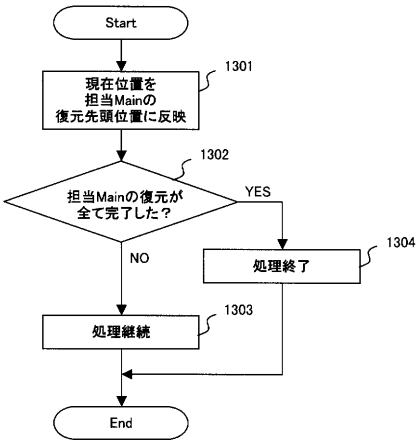
【図１２】

改善案1の復元ルーチンのフローチャート



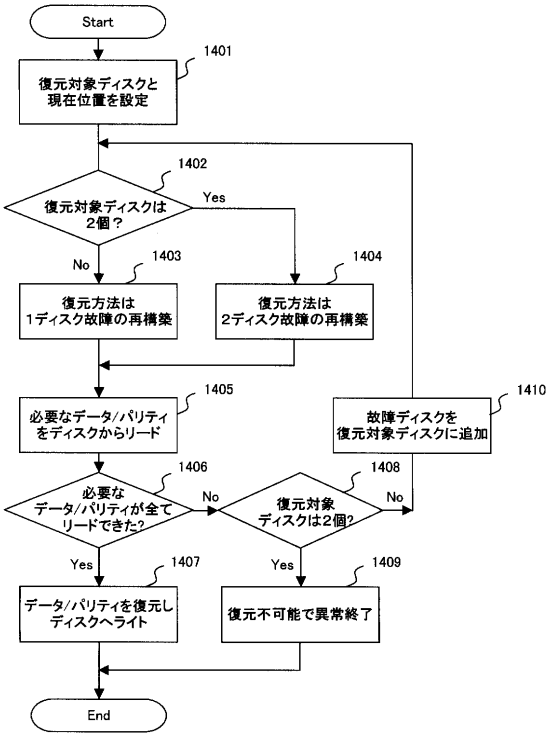
【図１３】

改善案1の復元後処理のフローチャート



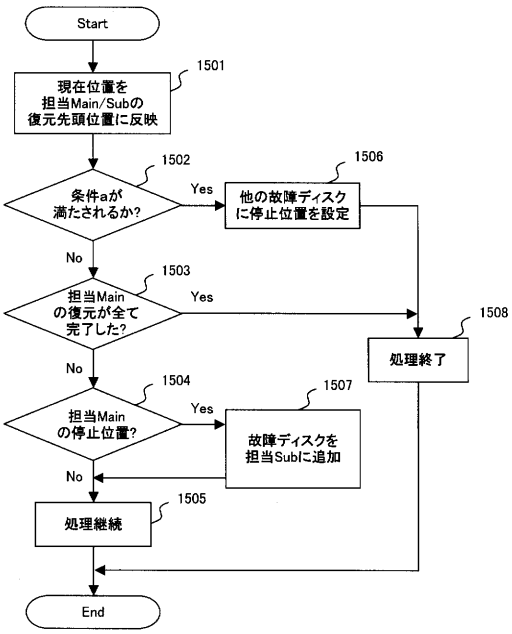
【図１４】

改善案2～5の復元ルーチンのフローチャート



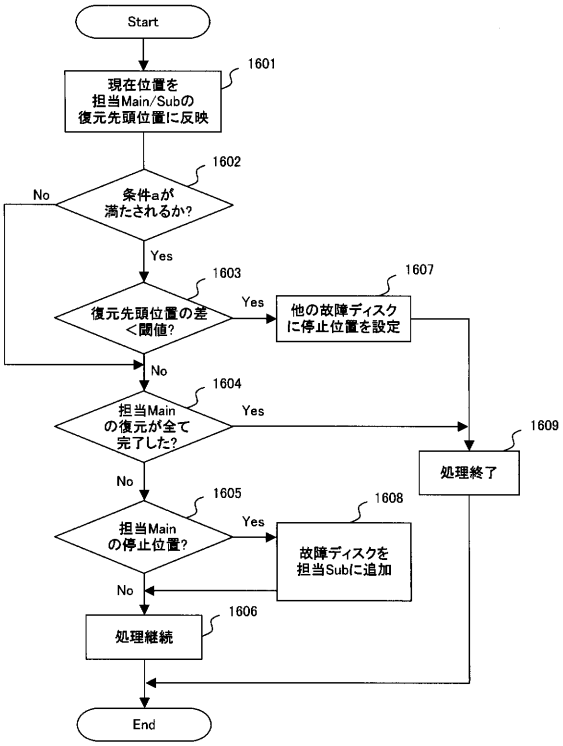
【図１５】

改善案2の復元後処理のフローチャート



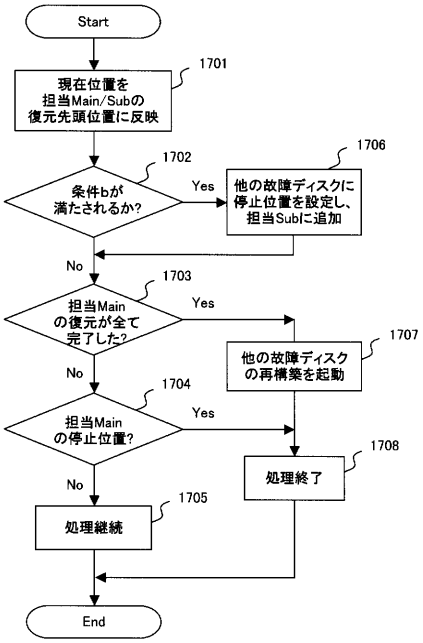
【図１６】

改善案3の復元後処理のフローチャート



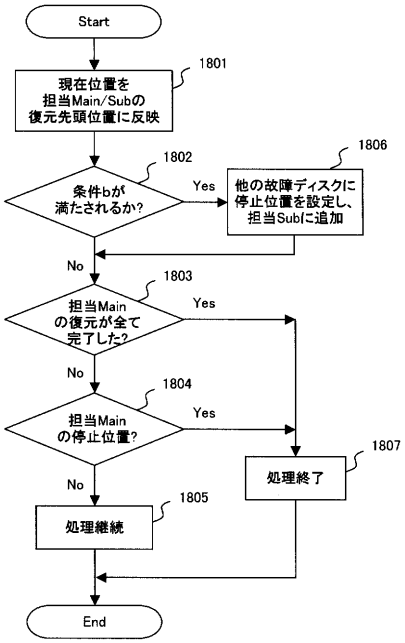
【図１７】

改善案4の復元後処理のフローチャート



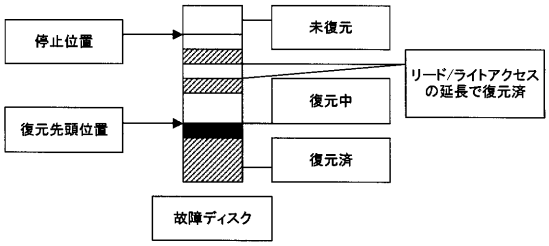
【図１８】

改善案5の復元後処理のフローチャート



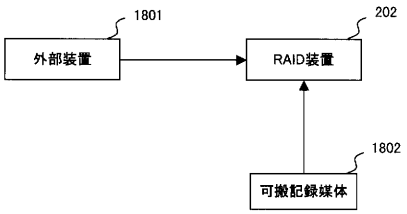
【図１９】

改善案6の復元状態を示す図



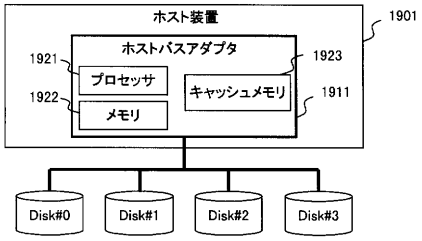
【図２０】

プログラムおよびデータの提供方法を示す図



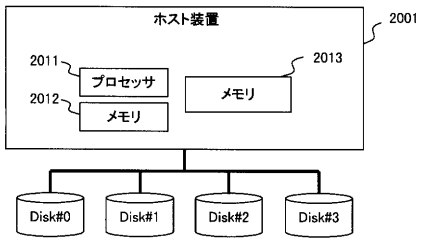
【図２１】

第2のストレージシステムの構成図



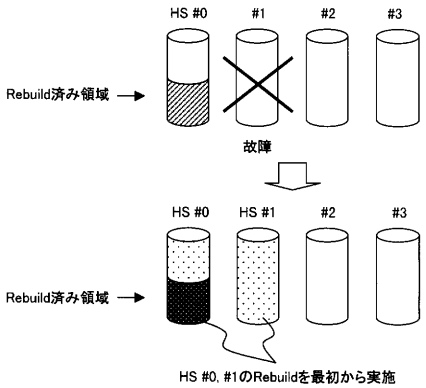
【図２２】

第3のストレージシステムの構成図

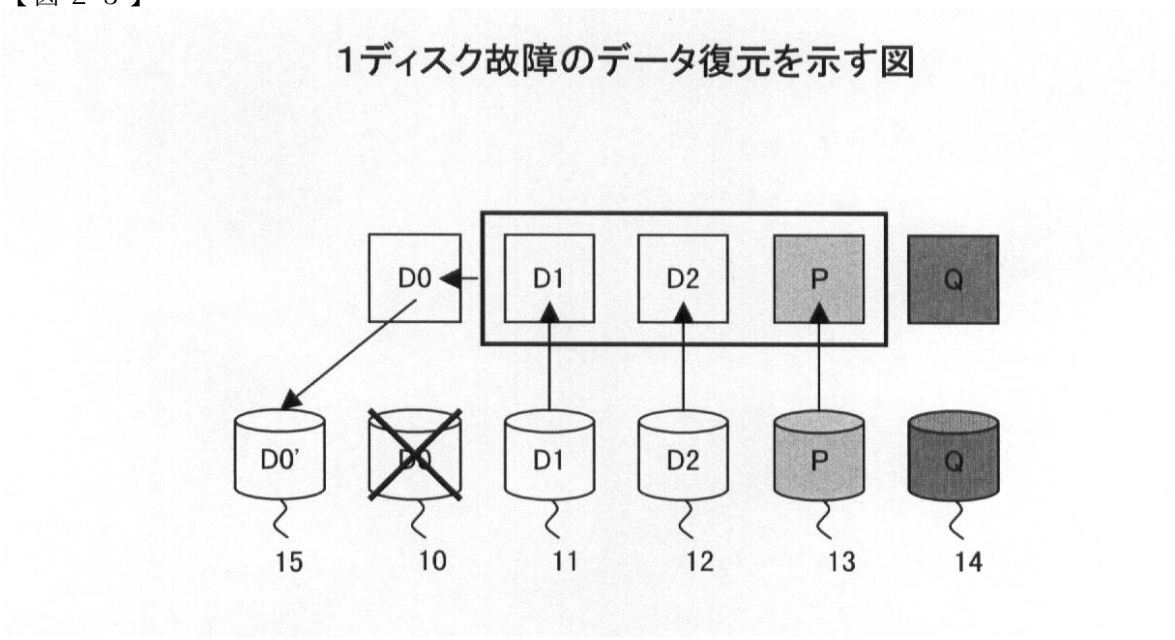


【図２５】

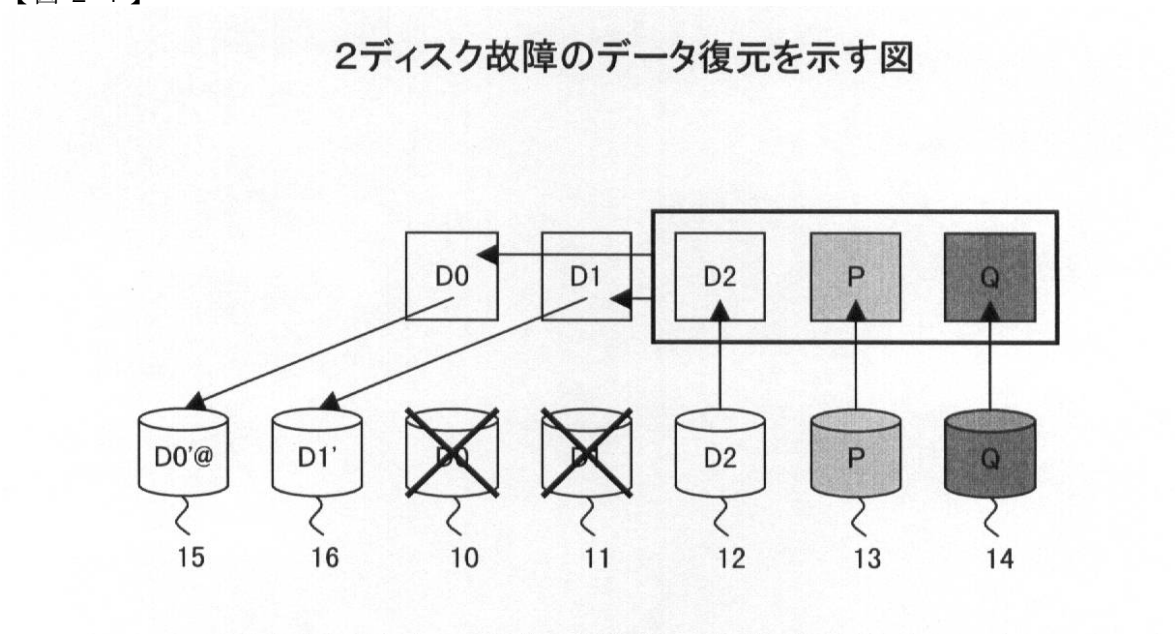
2ディスク故障の再構築処理を示す図



【図２３】


1ディスク故障のデータ復元を示す図
D0
D1
D2
P
Q
D0'
D0
D1
D2
P
Q
15
10
11
12
13
14

【図２４】


2ディスク故障のデータ復元を示す図
D0
D1
D2
P
Q
D0'@
D1'
D0
D1
D2
P
Q
15
16
10
11
12
13
14

フロントページの続き

(72)発明者　大黒谷　秀治郎
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
(72)発明者　池内　和彦
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
(72)発明者　▲高▼橋　秀夫
神奈川県川崎市中原区上小田中４丁目１番１号　株式会社富士通コンピュータテクノロジーズ内
(72)発明者　紺田　與志仁
神奈川県川崎市中原区上小田中４丁目１番１号　株式会社富士通コンピュータテクノロジーズ内
(72)発明者　佐藤　靖丈
神奈川県川崎市中原区上小田中４丁目１番１号　株式会社富士通コンピュータテクノロジーズ内
(72)発明者　越智　弘昭
神奈川県川崎市中原区上小田中４丁目１番１号　株式会社富士通コンピュータテクノロジーズ内
(72)発明者　牧野　司
神奈川県川崎市中原区上小田中４丁目１番１号　株式会社富士通コンピュータテクノロジーズ内
(72)発明者　久保田　典秀
神奈川県川崎市中原区上小田中４丁目１番１号　株式会社富士通コンピュータテクノロジーズ内

Ｆターム(参考) 5B001 AB01　AB02　AD04
5B018 HA04　HA14　MA12
5B065 BA01　CA30　EA03　EA24