

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4754852号  
(P4754852)

(45) 発行日 平成23年8月24日(2011.8.24)

(24) 登録日 平成23年6月3日(2011.6.3)

(51) Int.Cl. F 1  
G 0 6 F 3/06 (2006.01) G 0 6 F 3/06 3 0 4 B

請求項の数 13 (全 28 頁)

(21) 出願番号	特願2005-73669 (P2005-73669)	(73) 特許権者	000005223 富士通株式会社
(22) 出願日	平成17年3月15日 (2005.3.15)		神奈川県川崎市中原区上小田中4丁目1番1号
(65) 公開番号	特開2006-259894 (P2006-259894A)	(74) 代理人	100074099 弁理士 大菅 義之
(43) 公開日	平成18年9月28日 (2006.9.28)	(74) 代理人	100133570 弁理士 ▲徳▼永 民雄
審査請求日	平成18年4月25日 (2006.4.25)	(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
審判番号	不服2009-17251 (P2009-17251/J1)	(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
審判請求日	平成21年9月15日 (2009.9.15)		

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) 【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の記憶装置の復元済み領域の情報を前記第1の予備記憶装置内に保持したままで、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して前記第1の予備記憶装置内で情報が保持された領域以外の領域に書き込む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

10

20

する請求項 1 または 2 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 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 または 2 記載のストレージ制御装置。

【請求項 5】

前記第 1 の記憶装置内での復元処理の現在位置を示す復元先頭位置と前記第 2 の記憶装置内での復元処理を停止すべき位置を示す復元停止位置の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、前記第 2 の記憶装置が故障したときに、前記復元先頭位置に基づいて前記復元停止位置を設定することを特徴とする請求項 4 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 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 または 2 記載のストレージ制御装置。

【請求項 7】

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 または 2 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 または 2 記載のストレージ制御装置。

【請求項 9】

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 8 記載のストレージ制御装置。

【請求項 10】

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 9 記載のストレージ制御装置。

10

20

30

40

50

## 【請求項 1 1】

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 の記憶装置の復元済み領域の情報を前記第 1 の予備記憶装置内に保持したままで、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。

10

## 【請求項 1 2】

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 の記憶装置の復元済み領域の情報を前記第 1 の予備記憶装置内に保持したままで、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。

20

## 【請求項 1 3】

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 の記憶装置の復元済み領域の情報を前記第 1 の予備記憶装置内に保持したままで、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。

30

## 【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。

40

## 【背景技術】

## 【0002】

RAID は、複数のハードディスクを組み合わせ、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、RAID 0 ~ RAID 6 の 7 つのレベルに分類されている。このうち、RAID 3 ~ RAID 6 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)。

50

## 【 0 0 0 3 】

R A I D 6 は 2 つのディスクの故障に対応した R A I D レベルである。R A I D 6 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。

## 【 0 0 0 4 】

例えば、図 2 3 に示すように、ディスク 1 0 ~ 1 4 の 5 つのディスクからなる R A I D 装置において、ディスク 1 0 が故障してデータ D 0 が失われた場合、予備ディスクであるホットスペア 1 5 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 1 1 ~ 1 3 に格納されたデータ D 1、D 2、およびパリティ P からデータ D 0 が復元される。

10

## 【 0 0 0 5 】

これに対して、図 2 4 に示すように、ディスク 1 0 および 1 1 が故障してデータ D 0 および D 1 が失われた場合、ホットスペア 1 5 および 1 6 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 1 2 ~ 1 4 に格納されたデータ D 2 およびパリティ P、Q からデータ D 0 および D 1 が復元される。

## 【 0 0 0 6 】

一般に、R A I D 6 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。

20

## 【 0 0 0 7 】

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 2 5 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 ( H S ) 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、H S # 0 および H S # 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 0 3 - 2 4 0 1 2 3 号公報

30

## 【 発明の開示 】

## 【 発明が解決しようとする課題 】

## 【 0 0 0 8 】

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 2 5 に示した 2 ディスク故障の再構築処理では、H S # 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。

## 【 0 0 0 9 】

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、H S # 0 および H S # 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。

40

## 【 0 0 1 0 】

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 R A I D 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 】

## 【 0 0 1 1 】

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 1 0 1 は、再構築手段 1 1 1 および 1 1 2 を備え、複数の記憶装置 1 0 2 - 1 ~ 1 0 2 - 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

50

。

## 【 0 0 1 2 】

再構築手段 1 1 1 は、記憶装置 1 0 2 - 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 1 0 2 - 1 の情報を復元し、予備記憶装置 1 0 3 - 1 に書き込む。再構築手段 1 1 2 は、記憶装置 1 0 2 - 1 の情報を復元している間に記憶装置 1 0 2 - 2 が故障したとき、記憶装置 1 0 2 - 1 および 1 0 2 - 2 以外の記憶装置に格納された情報を用いて記憶装置 1 0 2 - 1 の未復元領域の情報と記憶装置 1 0 2 - 2 の情報を復元し、予備記憶装置 1 0 3 - 1 の対応する領域と予備記憶装置 1 0 3 - 2 にそれぞれ書き込む。

## 【 0 0 1 3 】

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 1 0 2 - 1 の情報を復元している間に記憶装置 1 0 2 - 2 が故障したとき、記憶装置 1 0 2 - 1 の復元済み領域の情報はそのまま予備記憶装置 1 0 3 - 1 内に保存され、記憶装置 1 0 2 - 1 の未復元領域と記憶装置 1 0 2 - 2 の全領域を対象として復元処理が行われる。

## 【 0 0 1 4 】

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 1 0 2 - 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 1 0 2 - 1 の復元済み領域に対応する記憶装置 1 0 2 - 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。

## 【 0 0 1 5 】

ストレージ制御装置 1 0 1 は、例えば、後述する図 2 のコントローラ 2 1 1、図 2 1 のホストバスアダプタ 1 9 1 1、または図 2 2 のホスト装置 2 0 0 1 に対応する。

## 【 発明の効果 】

## 【 0 0 1 6 】

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 】

## 【 0 0 1 7 】

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 2 0 1 および R A I D 装置 2 0 2 からなり、R A I D 装置 2 0 2 は、コントローラ 2 1 1 および D i s k # 0 ~ D i s k # 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。

## 【 0 0 1 8 】

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 2 0 1 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 2 1 1 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。

## 【 0 0 1 9 】

コントローラ 2 1 1 は、プロセッサ 2 2 1、メモリ 2 2 2、およびキャッシュメモリ 2 2 3 を備え、D i s k # 0 ~ D i s k # 3 の故障時の再構築処理を行う。プロセッサ 2 2 1 は、メモリ 2 2 2 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 2 2 3 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。

## 【 0 0 2 0 】

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

10

20

30

40

50

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 2 1 1 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。

10

##### 【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。

20

##### 【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 2 5 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 2 5 の再構築処理よりも早くディスク # 0 の復元が完了する。

30

#### 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 2 1 1 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。

40

##### 【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 #

50

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。

10

#### 【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。

20

#### 【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。

30

#### 【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。

40

#### 【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】

50

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 2 1 1 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。

10

##### 【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。

20

##### 【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 2 1 1 は、キャッシュメモリ 2 2 3 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。

30

##### 【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 2 1 1 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 2 1 1 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。

40

##### 【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 2 1 1 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、これを越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 2 1 1 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

50

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 2 1 1 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。

【 0 0 4 3 】

例えば、図 8 に示すように、最初にディスク # 0 が故障し、H S # 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。

【 0 0 4 4 】

コントローラ 2 1 1 は、H S # 0 および H S # 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。H S # 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。

【 0 0 4 5 】

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。

【 0 0 4 6 】

次に、図 9 から図 1 9 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 2 1 1 は、契機となったディスクを担当 M a i n に指定し、必要に応じて他の故障ディスクを担当 S u b として追加する。再構築処理では、図 9 に示すように、担当 M a i n と担当 S u b 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。

【 0 0 4 7 】

また、コントローラ 2 1 1 は、R A I D 装置 2 0 2 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 2 2 3 に保持する。具体的には、図 1 0 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。

【 0 0 4 8 】

故障ディスクの復元先頭位置は、そのディスクを担当 M a i n 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。

【 0 0 4 9 】

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。

【 0 0 5 0 】

図 1 1 は、再構築処理のフローチャートである。コントローラ 2 1 1 は、まず、担当 M a i n 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 1 1 0 1）、復元ルーチンを実行して復元処理を行う（ステップ 1 1 0 2）。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。

10

20

30

40

50

## 【 0 0 5 1 】

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。

## 【 0 0 5 2 】

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。

10

## 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。

## 【 0 0 5 3 】

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）。

20

## 【 0 0 5 4 】

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）。

## 【 0 0 5 5 】

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。

30

## 【 0 0 5 6 】

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）。

## 【 0 0 5 7 】

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）。

40

## 【 0 0 5 8 】

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）。

## 【 0 0 5 9 】

50

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)。

10

【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)。

20

【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)。

30

【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)。

40

## 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】

50

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づいて選択される。

【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)。

10

【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。

20

【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)。

30

【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】

40

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

50

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）。

10

【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）。

20

【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。

30

【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）。

40

【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

50

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

10

20

30

40

50

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。

10

【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）。

20

【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。

30

【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）。

40

【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

50

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

10

20

30

40

50

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

10

20

30

40

50

in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。

10

【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。

20

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、再構築処理が実行される。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。

30

【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。

40

【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

50

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

10

20

30

40

50

、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리카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。

10

【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。

20

【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。

30

【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。

40

【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。

50

- 【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。

10

## 【符号の説明】

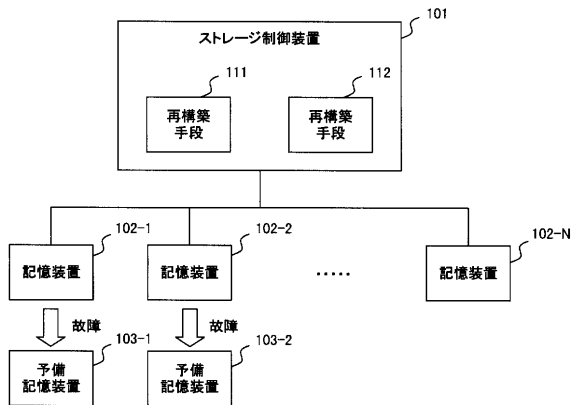
## 【0138】

- 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

20

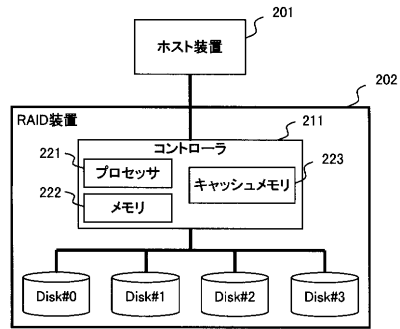
【図1】

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



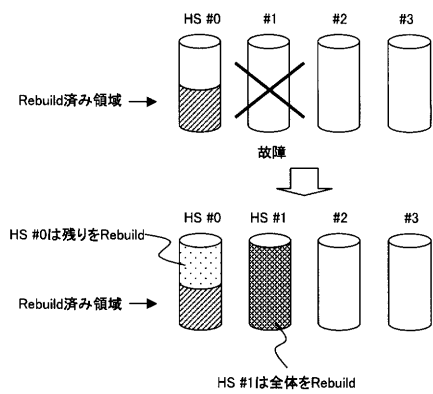
【図3】

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

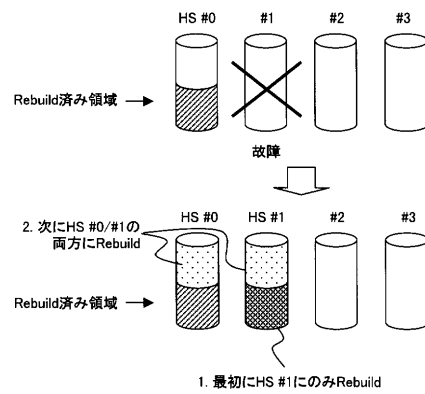
【図4】

改善案1を示す図

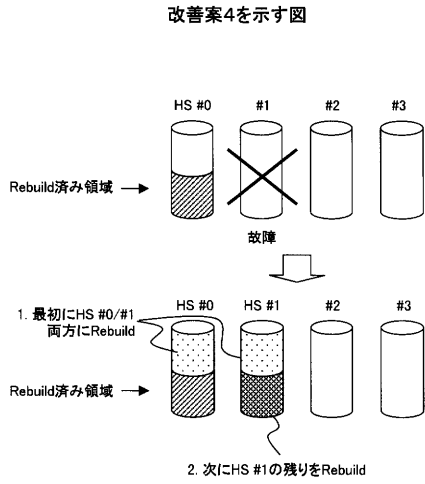


【図5】

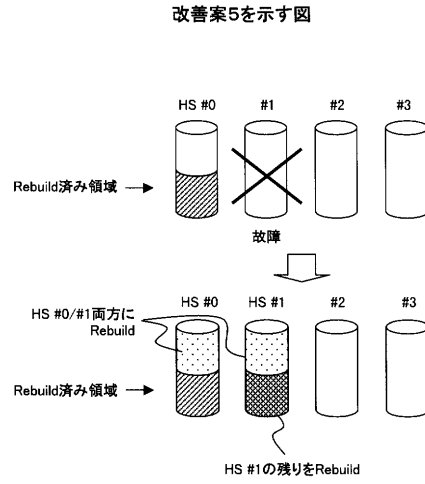
改善案2を示す図



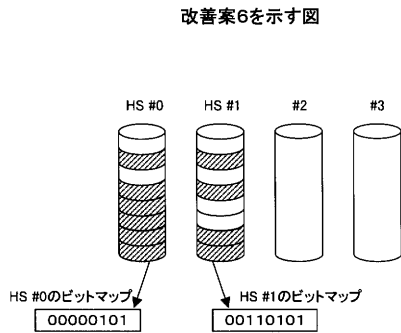
【図6】



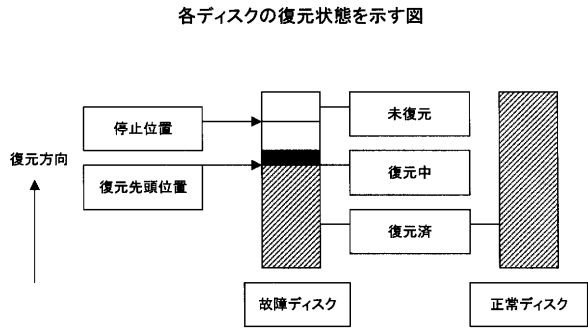
【図7】



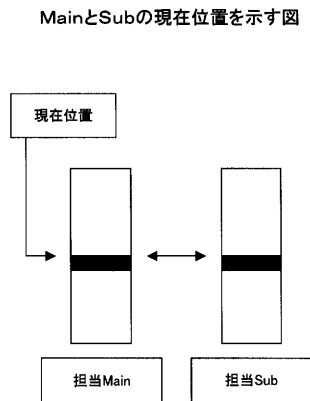
【図8】



【図10】

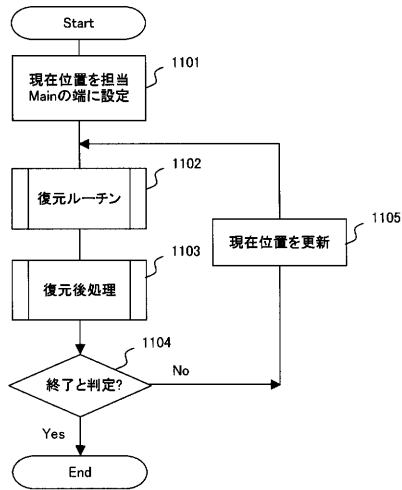


【図9】



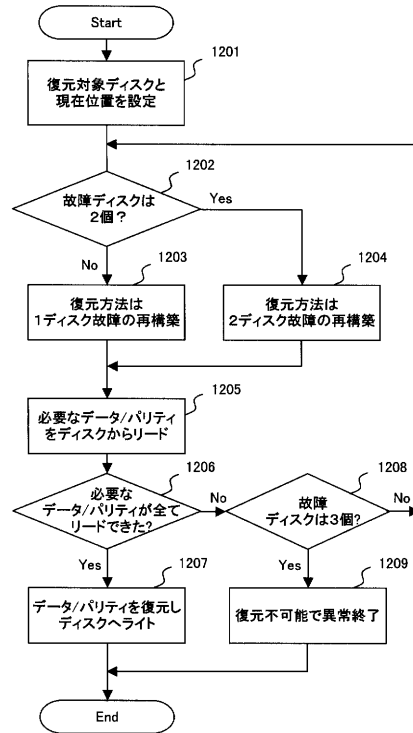
【図11】

再構築処理のフローチャート



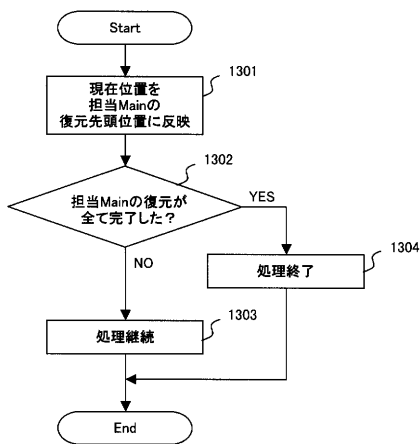
【図12】

改善案1の復元ルーチンのフローチャート



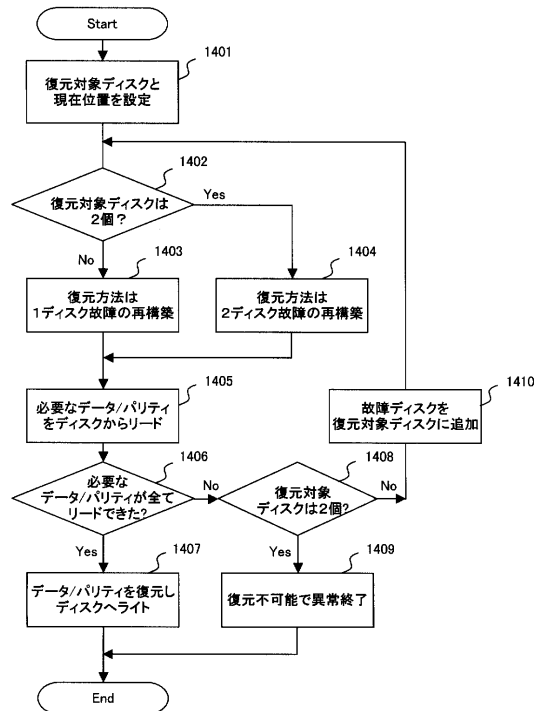
【図13】

改善案1の復元後処理のフローチャート



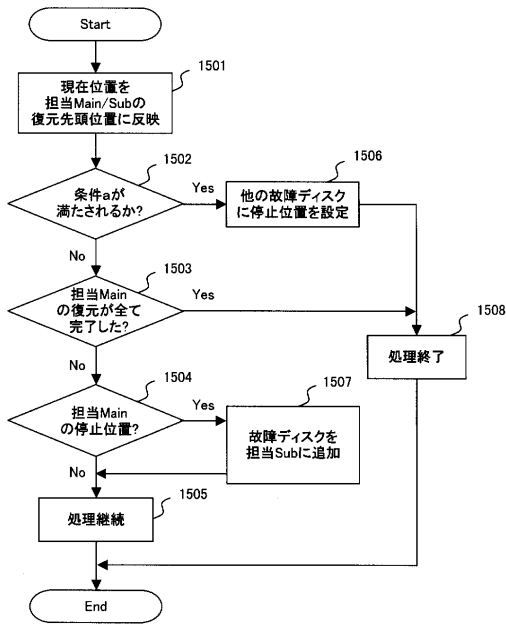
【図14】

改善案2~5の復元ルーチンのフローチャート



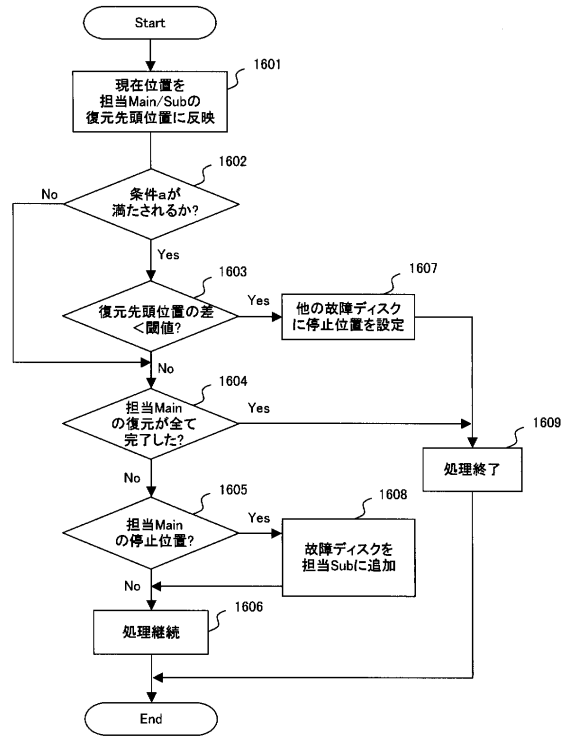
【図15】

改善案2の復元後処理のフローチャート



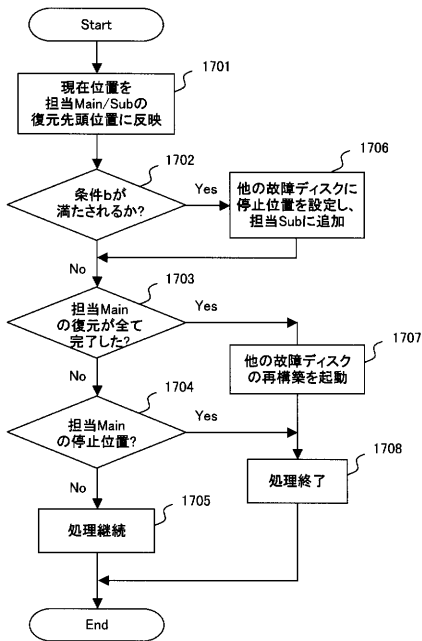
【図16】

改善案3の復元後処理のフローチャート



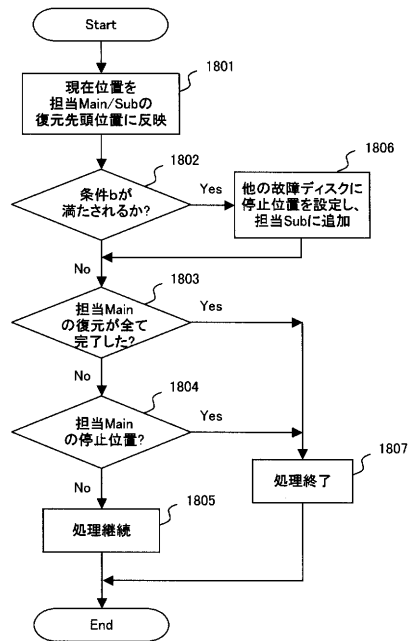
【図17】

改善案4の復元後処理のフローチャート



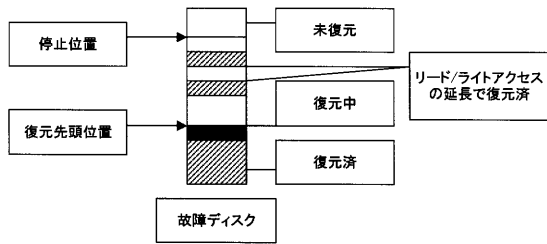
【図18】

改善案5の復元後処理のフローチャート



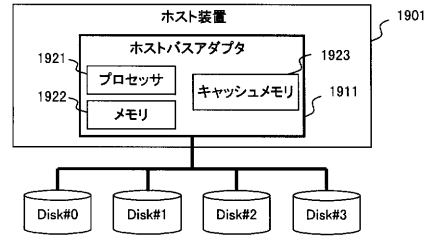
【図19】

改善案6の復元状態を示す図



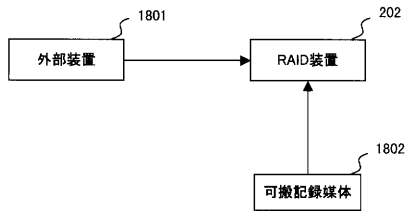
【図21】

第2のストレージシステムの構成図



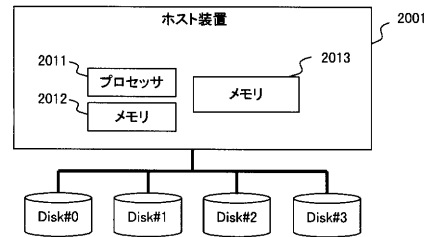
【図20】

プログラムおよびデータの提供方法を示す図



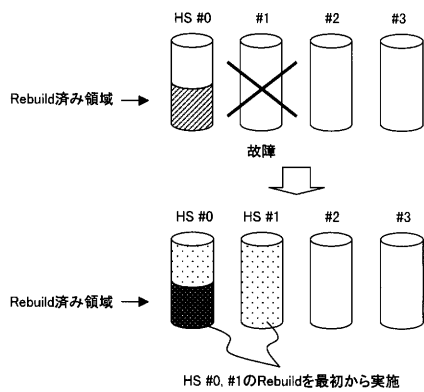
【図22】

第3のストレージシステムの構成図



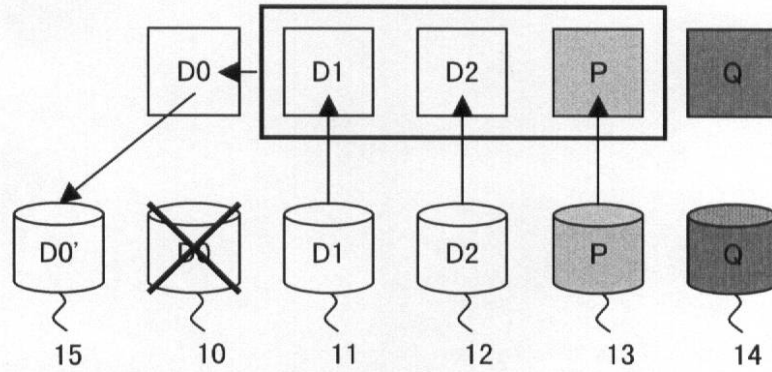
【図25】

2ディスク故障の再構築処理を示す図



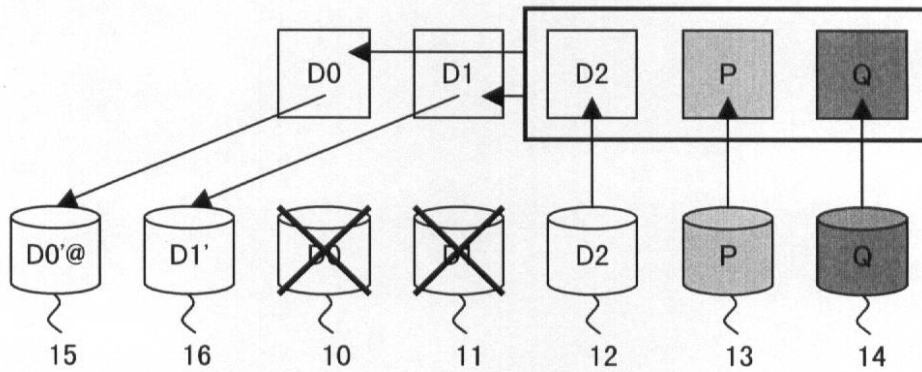
【図23】

1ディスク故障のデータ復元を示す図



【図24】

2ディスク故障のデータ復元を示す図



## 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

## 合議体

- 審判長 水野 恵雄  
審判官 中野 裕二  
審判官 稲葉 和生

- (56)参考文献 特開平7-28710(JP,A)  
特開2004-30577(JP,A)  
特開2003-233469(JP,A)  
特開平9-146850(JP,A)  
特開2001-290746(JP,A)  
特開2001-147785(JP,A)  
特開2000-259359(JP,A)  
特開2003-233468(JP,A)  
特開2003-85019(JP,A)  
特開2002-123372(JP,A)

## (58)調査した分野(Int.Cl., DB名)

G06F 3/06-3/08