(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2008-9767
(P2008-9767A)
(43)公開日　平成20年1月17日(2008.1.17)

(51)Int.Cl.	F I	テーマコード（参考）
G06F 3/06 (2006.01)	G06F 3/06 304F	5B065

審査請求 未請求　請求項の数 18　ＯＬ　（全 26 頁）

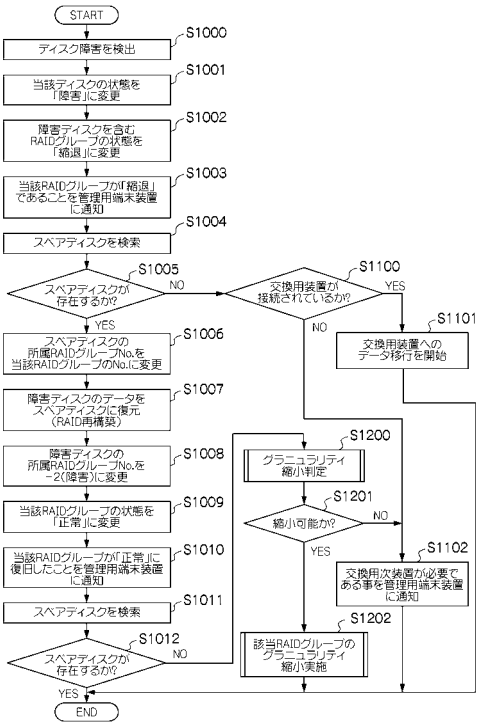
(21)出願番号　特願2006-180256 (P2006-180256)
(22)出願日　平成18年6月29日 (2006.6.29)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100079108
弁理士　稲葉　良幸
(74)代理人　100093861
弁理士　大賀　眞司
(72)発明者　新井　政弘
神奈川県川崎市麻生区王禅寺１０９９番地
株式会社日立製作所システム開発研究所内

Ｆターム(参考)　5B065 BA01 CA30 CC03 EA12 EA23 EA33

(54)【発明の名称】データ処理システム及びその方法並びにストレージ装置

(57)【要約】
【課題】本発明は、多数のスペアディスクを搭載せずとも、ディスク非交換で耐障害性を維持し、小型で長期利用可能なデータ処理システム及びその方法並びにストレージ装置を提供することを目的とする。
【解決手段】上位装置と、前記上位装置に対してデータを読み書きするための第１の記憶領域を提供するストレージ装置とを有するデータ処理システムにおいて、前記ストレージ装置は、複数設けられた前記第１の記憶領域の一部領域に障害が発生した場合に、当該一部領域に格納された前記データを移行するためのスペアの第２の記憶領域が存在しなかったときに、前記複数の第１の記憶領域の他の一部領域を前記第２の記憶領域として動的に確保する確保部とを備えることとした。
【選択図】図１１



【特許請求の範囲】
【請求項１】
上位装置と、前記上位装置に対してデータを読み書きするための第１の記憶領域を提供するストレージ装置とを有するデータ処理システムにおいて、
前記ストレージ装置は、
複数設けられた前記第１の記憶領域の一部領域に障害が発生した場合に、当該一部領域に格納された前記データを移行するためのスペアの第２の記憶領域が存在しなかったときに、前記複数の第１の記憶領域の他の一部領域を前記第２の記憶領域として動的に確保する確保部
を備えることを特徴とするデータ処理システム。
【請求項２】
前記ストレージ装置は、
前記複数の第１の記憶領域の一部領域に格納された前記データを移行するための第２の記憶領域を検索する検索部
を備えることを特徴とする請求項１に記載のデータ処理システム。
【請求項３】
前記複数の第１の記憶領域は、論理ボリュームを構成し、
前記確保部は、前記複数の第１の記憶領域のうち一の記憶領域に予め格納された前記データを、前記複数の第１の記憶領域のうち他の記憶領域に移行して前記論理ボリューム内に前記第２の記憶領域を確保する
ことを特徴とする請求項１に記載のデータ処理システム。
【請求項４】
前記複数の第１の記憶領域は、パリティデータを格納する論理ボリュームを構成し、
前記確保部は、前記パリディデータを削除して前記論理ボリューム内に前記第２の記憶領域を確保する
ことを特徴とする請求項１に記載のデータ処理システム。
【請求項５】
前記複数の第１の記憶領域は、複数の論理ボリュームを構成し、
前記確保部は、前記複数の論理ボリュームから前記データを格納していない未使用の論理ボリュームを前記第２の記憶領域として確保する
ことを特徴とする請求項１に記載のデータ処理システム。
【請求項６】
前記複数の第１の記憶領域は、動的に拡張する仮想ボリュームを構成し、
前記上位装置からアクセスされた前記複数の第１の記憶領域のうち一の記憶領域は、動的に拡張するプールボリュームを構成し、
前記確保部は、前記プールボリュームを縮小して前記第２の記憶領域を確保する
ことを特徴とする請求項１に記載のデータ処理システム。
【請求項７】
上位装置と、前記上位装置に対してデータを読み書きするための第１の記憶領域を提供するストレージ装置とを有するデータ処理システムのデータ処理方法において、
前記ストレージ装置には、複数設けられた前記第１の記憶領域の一部領域に障害が発生した場合に、当該一部領域に格納された前記データを移行するためのスペアの第２の記憶領域が存在しなかったときに、前記複数の第１の記憶領域の他の一部領域を前記第２の記憶領域として動的に確保する確保ステップ
を備えることを特徴とするデータ処理方法。
【請求項８】
前記ストレージ装置には、前記複数の第１の記憶領域の一部領域に格納された前記データを移行するための第２の記憶領域を検索する検索ステップ
を備えることを特徴とする請求項７に記載のデータ処理方法。
【請求項９】

前記複数の第１の記憶領域は、論理ボリュームを構成し、
前記確保ステップは、
前記複数の第１の記憶領域のうち一の記憶領域に予め格納された前記データを、前記複数の第１の記憶領域のうち他の記憶領域に移行して前記論理ボリューム内に前記第２の記憶領域を確保する、
ことを特徴とする請求項７に記載のデータ処理方法。
【請求項１０】
前記複数の第１の記憶領域は、パリティデータを格納する論理ボリュームを構成し、
前記確保部ステップは、
前記パリティデータを削除して前記論理ボリューム内に前記第２の記憶領域を確保する
ことを特徴とする請求項７に記載のデータ処理方法。
【請求項１１】
前記複数の第１の記憶領域は、複数の論理ボリュームを構成し、
前記確保ステップは、
前記複数の論理ボリュームから前記データを格納していない未使用の論理ボリュームを前記第２の記憶領域として確保する
ことを特徴とする請求項７に記載のデータ処理方法。
【請求項１２】
前記複数の第１の記憶領域は、動的に拡張する仮想ボリュームを構成し、
前記上位装置からアクセスされた前記複数の記憶領域のうち一の記憶領域から動的に拡張するプールボリュームを構成し、
前記確保ステップは、
前記プールボリュームを縮小して前記第２の記憶領域を確保する
ことを特徴とする請求項７に記載のデータ処理方法。
【請求項１３】
上位装置と、前記上位装置に対してデータを読み書きするための第１の記憶領域を提供するストレージ装置において、
複数設けられた前記第１の記憶領域の一部領域に障害が発生した場合に、当該一部領域に格納された前記データを移行するためのスペアの第２の記憶領域が存在しなかったときに、前記複数の第１の記憶領域の他の一部領域を前記第２の記憶領域として動的に確保する確保部
を備えることを特徴とするストレージ装置。
【請求項１４】
前記複数の第１の記憶領域の一部領域に格納された前記データを移行するための第２の記憶領域を検索する検索部
を備えることを特徴とする請求項１３に記載のストレージ装置。
【請求項１５】
前記複数の第１の記憶領域は、論理ボリュームを構成し、
前記確保部は、前記複数の第１の記憶領域のうち一の記憶領域に予め格納された前記データを、前記複数の第１の記憶領域のうち他の記憶領域に移行して前記論理ボリューム内に前記第２の記憶領域を確保する
ことを特徴とする請求項１３に記載のストレージ装置。
【請求項１６】
前記複数の第１の記憶領域は、パリティデータを格納する論理ボリュームを構成し、
前記確保部は、前記パリティデータを削除して前記論理ボリューム内に前記第２の記憶領域を確保する
ことを特徴とする請求項１３に記載のストレージ装置。
【請求項１７】
前記複数の第１の記憶領域は、複数の論理ボリュームを構成し、
前記確保部は、前記複数の論理ボリュームから前記データを格納していない未使用の論

理ボリュームを前記第２の記憶領域として確保する
ことを特徴とする請求項１３に記載のストレージ装置。
【請求項１８】
前記複数の第１の記憶領域は、動的に拡張する仮想ボリュームを構成し、
前記上位装置からアクセスされた前記複数の第１の記憶領域のうち一の記憶領域は、動的に拡張するプールボリュームを構成し、
前記確保部は、前記プールボリュームを縮小して前記第２の記憶領域を確保する
ことを特徴とする請求項１３に記載のストレージ装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ＲＡＩＤ（Redundant Array of Inexpensive Disks）を構成するデータ処理システム及びその方法並びにストレージ装置に関し、特にディスクドライブ（以下、単にディスクと呼ぶ）の一部、または、すべてを非交換に搭載するデータ処理システム及びその方法並びにストレージ装置に関する。
【背景技術】
【０００２】
従来、ストレージ装置として、ディスクの障害に起因するデータ損失を防止する方策としてＲＡＩＤ構成を採用したものがある（例えば、非特許文献１参照）。また、更なるディスク障害に備えるため、ＲＡＩＤ回復用にスペアのディスク（以下、スペアディスク）を用意し、迅速な復旧を実現している（例えば、特許文献１参照）。
【非特許文献１】David A.Patterson, Garth Gibson, and Randy H.Katz,"A Case for Redundant Arrays of Inexpensive Disks (ＲＡＩＤ)", University of California Berke.
【特許文献１】特開平７－１４６７６０号公報
【発明の開示】
【発明が解決しようとする課題】
【０００３】
前述のようなストレージ装置では、次のディスクの障害に備えるために、障害が発生したディスクは交換され、新たなディスクがスペアディスクとして補充されることが前提となっている。従い、ディスクを交換できないようなストレージ装置（以下、ディスク非交換型ストレージ装置）では、新たなスペアディスクの補充ができず、耐障害性が低下してしまうという問題が生じる。
【０００４】
一方、非交換を前提に多数のスペアディスクをあらかじめ搭載する方法も考えられるが、この方法では装置コストが高くなる上、装置自体も大きくなってしまうという問題が生じる。
【０００５】
本発明は、多数のスペアディスクを搭載せずとも、ディスク非交換で耐障害性を維持し、小型で長期利用可能なデータ処理システム及びその方法並びにストレージ装置を提供することを目的とする。
【課題を解決するための手段】
【０００６】
かかる課題を解決するため本発明は、上位装置と、前記上位装置に対してデータを読み書きするための第１の記憶領域を提供するストレージ装置とを有するデータ処理システムにおいて、前記ストレージ装置は、複数設けられた前記第１の記憶領域の一部領域に障害が発生した場合に、当該一部領域に格納された前記データを移行するためのスペアの第２の記憶領域が存在しなかったときに、前記複数の第１の記憶領域の他の一部領域を前記第２の記憶領域として動的に確保する確保部とを備えることを特徴とする。
【０００７】

本発明のデータ処理システムによれば、既存のディスクの記憶領域からスペアの記憶領域を動的に確保することができるため、小型で長期利用可能なデータ処理システムの耐障害性を維持することができる。

【０００８】

また、本発明は、上位装置と、前記上位装置に対してデータを読み書きするための第１の記憶領域を提供するストレージ装置とを有するデータ処理システムのデータ処理方法において、前記ストレージ装置には、複数設けられた前記第１の記憶領域の一部領域に障害が発生した場合に、当該一部領域に格納された前記データを移行するためのスペアの第２の記憶領域が存在しなかったときに、前記複数の第１の記憶領域の他の一部領域を前記第２の記憶領域として動的に確保する確保ステップとを備えることを特徴とする。

【０００９】

本発明のデータ処理方法によれば、既存のディスクの記憶領域からスペアの記憶領域を動的に確保することができるため、小型で長期利用可能なデータ処理システムの耐障害性を維持することができる。

【００１０】

さらに、本発明は、上位装置と、前記上位装置に対してデータを読み書きするための第１の記憶領域を提供するストレージ装置において、複数設けられた前記第１の記憶領域の一部領域に障害が発生した場合に、当該一部領域に格納された前記データを移行するためのスペアの第２の記憶領域が存在しなかったときに、前記複数の第１の記憶領域の他の一部領域を前記第２の記憶領域として動的に確保する確保部とを備えることを特徴とする。

【００１１】

本発明のストレージ装置によれば、既存のディスクの記憶領域からスペアの記憶領域を動的に確保することができるため、小型で長期利用可能なストレージ装置の耐障害性を維持することができる。

【発明の効果】

【００１２】

本発明によれば、多数のディスクを搭載せずとも、ディスクを交換することなく、スペアディスク又はスペア領域を確保することができ、小型で長期利用可能なデータ処理システム又はストレージ装置の耐障害性を維持することができる。

【００１３】

また、耐障害性を維持できなくなる可能性が生じた場合（スペアディスク又はスペア領域を確保できなくなった場合）、通知により交換用ストレージ装置の準備の必要性を知ることができる。

【発明を実施するための最良の形態】

【００１４】

以下、本発明の第１から第４の実施の形態を、図面を参照して説明する。

【００１５】

１．第１の実施の形態

図１は本発明の第１から第４の実施の形態におけるストレージ装置の構成を表した図である。ストレージ装置１は、管理用ＬＡＮ５を介して管理用端末装置２と接続されている。また、ストレージ装置１は、ストレージネットワーク４を介して複数のホスト計算機３と接続されている。さらに、ストレージ装置１は交換用の他ストレージ装置との接続インターフェース７を備えている。

【００１６】

ストレージ装置１には、ストレージコントローラ１７１、１７２が備えられており、ストレージネットワークとの間で同時に多くのデータを入出力できるようにしている。

【００１７】

なお、実施の形態によっては、ストレージコントローラは１つでもよいし、３つ以上あってもよい。

【００１８】

また、ストレージ装置１には、複数のディスクＤ０、Ｄ１、Ｄ２・・・ＤＮが備えられている。ストレージコントローラ１７１、１７２は接続インターフェース６を介して、ディスクＤ０～ＤＮと接続されている。これにより、ストレージコントローラ１７１、１７２とディスクＤ０～ＤＮは互いにデータの入出力を行うことができる。ストレージコントローラ１７１，１７２とディスクＤ０～ＤＮとの接続には、ＳＡＴＡ（Serial ATA）やＳＡＳ（Serial Attached SCSI）、ファイバーチャネル（Fibre Channel）等、データ転送に適切な通信路が用いられる。

【００１９】

ストレージコントローラ１７１、１７２は、制御プログラムが動作しており、ディスクＤ０～ＤＮに対するデータの入出力を制御する。また、ストレージコントローラ１７１，１７２は、ディスクＤ０～ＤＮによって構成されるＲＡＩＤの構成を管理する。また、ストレージコントローラ１７１、１７２は、管理用端末装置２と通信を行い、種々のデータをやりとりする。

【００２０】

ディスクＤ０～ＤＮは、ＳＡＴＡ、ＳＡＳまたはファイバーチャネル等で接続可能なディスクである。

【００２１】

管理用端末装置２は、ＣＰＵ、メモリ、記憶装置、インターフェース、入力装置および表示装置が備わるコンピュータ装置である。管理用端末装置２では管理プログラムが動作しており、当該管理プログラムによってストレージ装置１の動作状態を把握し、ストレージ装置１の動作を制御する。なお、管理用端末装置２ではＷｅｂブラウザ等のクライアントプログラムが動作しており、ストレージ装置１から供給される管理プラグラム（Common Gateway InterfaceやJava（登録商標）等）によってストレージ装置１の動作を制御するようにしてもよい。

【００２２】

表示画面２１は管理用端末装置２に備わる表示装置に表示される画面である。

【００２３】

ホスト計算機３は、ＣＰＵ、メモリ、記憶装置およびインターフェースが備わるコンピュータ装置（上位装置）であり、ストレージ装置１から供給されるデータを利用して、データベースサービスやウェブサービス等を利用可能にする。

【００２４】

ストレージネットワーク４は、例えばＳＡＳプロトコル、ファイバーチャネルプロトコルのような、データの転送に適するプロトコルで通信可能なネットワークである。

【００２５】

管理用ＬＡＮ５は、例えば、ＴＣＰ／ＩＰ(Transmission Control Protocol / Internet Protocol)によって、コンピュータ間でデータや制御情報を通信可能であり、例えばイーサネット（登録商標）が用いられる。

【００２６】

図２は、本発明の第１の実施の形態のストレージ装置１の構成を示す概観図である。ストレージコントローラ１７１、１７２、ディスクＤ０～ＤＮ、ファンおよび電源８は、ストレージ装置筐体１に内蔵されており、天板１１で覆われている（図では、内部の構成を表現するために、天板１１の一部を切り取って描画している）。ディスクＤ０～ＤＮは、非交換に設けられている。

【００２７】

また、前面は複数の吸気口を設けたベゼル１２で覆われている。ベゼル１２には、装置の起動・停止に必要な電源ボタン１３、１４、および、障害時に鳴るブザーを止めるブザー停止ボタン１５が備わっている。

【００２８】

なお、図には示さないが、各ディスクのアクセス状況を示す、いくつかのアクセスＬＥＤが前面に備わっていてもよい。

【００２９】
図３は、本発明の第１の実施の形態におけるストレージコントローラ１７１の内部構成を示すブロック図である。なお、ストレージコントローラコントローラ１７２も同様である。
【００３０】
ストレージコントローラ１７１には、ＣＰＵ１９０１、メモリ１９０２、パリティジェネレータおよびデータ転送コントローラ１９０５、データバッファ１９０８、フロントエンド接続インターフェースコントローラ１９０６、バックエンド接続インターフェースコントローラ１９０７、ＬＡＮインターフェースコントローラ１９０９が備えられている。これらは実装に適切なデータ転送路を介して、互いに接続されている。
【００３１】
ＣＰＵ１９０１は、メモリ１９０２に格納される制御プログラム１９０３、管理テーブル１９０４を読み込んでストレージ装置１のさまざまな制御を行う。
【００３２】
パリティジェネレータおよびデータ転送コントローラ１９０５は、ＣＰＵ１９０１の指示に基づいて、メモリ１９０２、フロントエンド接続インターフェースコントローラ１９０６、バックエンド接続インターフェースコントローラ１９０７、データバッファ１９０８との間で互いのデータの転送を行う。また、データバッファ１９０８を用いて、所定のデータに対するパリティ計算を実行する。
【００３３】
フロントエンド接続インターフェース１９０６は、図１にて示した、ホスト計算機３との間でデータの入出力の制御を行い、必要に応じて、データバッファ１９０８にデータを格納したり、取り出したりする。
【００３４】
バックエンド接続インターフェース１９０７は、ディスクＤ０等との間でデータの入出力を行い、必要に応じてデータバッファ１９０８にデータを格納したり、取り出したりする。
【００３５】
データバッファ１９０８は、例えばＤＩＭＭ（Dual Inline Memory Module）等で構成されるメモリであり、バッテリ等によって不揮発にされている。
【００３６】
ＬＡＮインターフェースコントローラ１９０９は、管理用端末装置２とデータや制御情報の入出力を行うインターフェースコントローラである。
【００３７】
なお、第１の実施の形態におけるストレージコントローラ１７２は、上述で説明をしたストレージコントローラ１７１と同様の構成のため、説明を省略する。また、図３に示すストレージコントローラのブロック構成は一例であり、同様の機能を有せば、実装にハードウェア、ソフトウェアを問わない。また、ブロックの構成が異なっていてもよい。
【００３８】
図４は、メモリ１９０２に格納される制御プログラム１９０３及び管理テーブル１９０４を示す説明図である。
【００３９】
制御プログラム１９０３は、ＲＡＩＤ制御プログラム１９１１、ＲＡＩＤグループ設定プログラム１９１２、ＬＵ設定プログラム１９１３および管理・通知プログラム１９１４から構成される。
【００４０】
管理テーブル１９０４は、ディスク情報管理テーブルＴ１、ＲＡＩＤグループ情報管理テーブルＴ２、ＬＵ情報管理テーブルＴ３から構成される。
【００４１】
そうしてＣＰＵ１９０１は、制御プログラム１９０３及び管理テーブル１９０４に基づ

いて、ホスト計算機３と入出力されたデータを処理し、ディスクＤ０等と入出力を行う。また、ＣＰＵ１９０１は、同制御にかかわる障害復旧処理を行う。

【００４２】

ＲＡＩＤグループ設定プログラム１９１２は、管理用端末装置２からの指示やＣＰＵ１９０１の指示に基づき、ＲＡＩＤグループ情報管理テーブルＴ２を用いて、ＲＡＩＤグループの作成、変更または削除を行う。

【００４３】

ＬＵグループ設定プログラム１９１３は、管理用端末装置２からの指示やＣＰＵ１９０１の指示に基づき、ＬＵ情報管理テーブルＴ３を用いて、ＬＵの作成、変更または削除を行う。

【００４４】

管理・通知プログラム１９１４は、管理用端末装置２とデータや制御情報の授受を行う。

【００４５】

ディスク情報管理テーブルＴ１は、ディスクＤ０等の各種情報を記録するテーブルである。ＲＡＩＤグループ情報管理テーブルＴ２は、ディスクＤ０～ＤＮから構成されるＲＡＩＤグループの各種情報を記録するテーブルである。ＬＵ情報管理テーブルＴ３は、ＲＡＩＤ上に構成される論理ディスクであるＬＵ（Logical Unit）の各種情報を記録するテーブルである。

【００４６】

図５は以降の説明での理解を助けるために、ディスク、ＲＡＩＤグループ、ＬＵの関係を模式的に表した図である。ＲＡＩＤグループは複数の物理的なディスク（図１のＤ０～ＤＮ）を用いて任意のレベルのＲＡＩＤ（例えばＲＡＩＤ－５）を構成した際に、そのＲＡＩＤ構成を一単位とするボリュームである。図５の例では、図１のディスクＤ０～Ｄ４をディスク０、１、２、３、４として用いて、ＲＡＩＤが構成されており、そのボリューム（ＲＡＩＤグループ）にＲＡＩＤグループ０のラベルが付されている。

【００４７】

ＬＵは、ＲＡＩＤグループを複数に論理分割したボリュームである。各ＬＵは、ホスト計算機からそれぞれディスクとして見えるので、論理ディスクとも呼ばれる。

【００４８】

図５の例では、ＲＡＩＤグループ０の一部を分割してＬＵ０を構成している。

【００４９】

図６には、データがどのように物理ディスクに格納されているかを示した説明図である。図６は、ＲＡＩＤ－５で構成されたＲＡＩＤグループ上にＬＵ０のデータが格納されている様子を描いている。データはストライプブロックと呼ばれる１以上のセクタから成る構成単位ごとに格納される。ｄ１、ｄ２、・・・、はストライプブロック単位に格納されたデータを示しており、ｐ１、ｐ２、・・・はストライプブロック横一列を単位とするストライプ列ごとにｄ１等から計算生成されるパリティを示している。

【００５０】

ＲＡＩＤ－５は、パリティブロックの位置は各ディスクに分散配置される。そのため、一列目ではディスク４に、二列目ではディスク３に、というふうにずらしながら格納される。

【００５１】

図７は、本発明の第１の実施の形態におけるディスク情報管理テーブルＴ１の説明図である。ディスク情報管理テーブルＴ１には、ディスクＤ０～ＤＮの各種情報が記録される。具体的には、ディスクの番号、ディスクの容量、ディスクの状態（「正常」か「障害」か「未装着」か）、所属するＲＡＩＤグループの番号が記録される。例えば、ディスク番号１のディスクは、容量が５００ＧＢであり、ディスクの状態は「正常」であり、所属するＲＡＩＤグループは０である。

【００５２】

なお、所属するＲＡＩＤグループの番号は、０か正の整数ならば、当該ＲＡＩＤグループの番号を意味し、－１の場合にはスペアディスクを意味し、－２の場合には障害中を意味し、－３の場合には未使用のディスクであることを意味する。

【００５３】

ディスク情報管理テーブルＴ１の内容は、メモリ１９０２上に置かれると共に、管理端末装置２からも確認することができる。

【００５４】

図８は、管理用端末装置２の表示画面２１において、ディスク情報管理テーブルＴ１で管理されるディスク情報を表示した際のディスク管理画面Ｖ１の例である。ユーザーはディスク管理画面Ｖ１を通じて、ディスクの使用状況を随時知ることができる。

【００５５】

図９は、本発明の第１の実施の形態におけるＲＡＩＤグループ情報管理テーブルＴ２の説明図である。ＲＡＩＤグループ情報管理テーブルＴ２には、ディスクＤ０～ＤＮを用いて構成されるＲＡＩＤグループの情報が記録される。具体的には、ＲＡＩＤグループの番号、ＲＡＩＤレベル、ＲＡＩＤに含まれるデータディスク数（パリティ分を除いたディスク数）、ＲＡＩＤの有効容量（実際にデータを書き込める容量）、当該ＲＡＩＤグループの状態（「正常」か「縮退」か「破損」か）、ＬＵへの割当済容量、未割当容量である。なお、ここでいう縮退とは、ＲＡＩＤ回復が可能な障害をいい、具体的にはパリディディスク分台以下の障害をいう。また、破損とは、ＲＡＩＤ回復が不可能な障害をいう。例えば、ＲＡＩＤグループ０は、ＲＡＩＤレベル５（ＲＡＩＤ－５）で構築されており、データディスク数は４であり、有効容量は２０００ＧＢ、状態は「正常」であり、ＬＵへの割当済容量は９００ＧＢ、未割当容量は１１００ＧＢである。

【００５６】

ＲＡＩＤグループ情報管理テーブルＴ１の内容は、メモリ１９０２上に置かれると共に、管理端末装置２からも確認することができる。

【００５７】

図１０は、本発明の第１の実施の形態におけるＬＵ情報管理テーブルＴ３の説明図である。ＬＵ管理テーブルＴ３は各ＲＡＩＤグループ上に構成されるＬＵの情報が記録される。具体的にはＬＵの番号、ＬＵの割当容量、割当元のＲＡＩＤグループ、当該ＬＵの割当開始Logical Brock Address（以下、ＬＢＡという）、終了ＬＢＡである。例えば、ＬＵ番号０のＬＵは、容量が１００ＧＢ、ＲＡＩＤグループ０上に設けられており、開始ＬＢＡは００００００００ｈであり、終了ＬＢＡは０００００Ｆ０００ｈである。ＬＢＡは、先に述べたストライプブロックに一定間隔で割り振られる一意のアドレスである。

【００５８】

ＬＵ情報管理テーブルＴ３は、ディスク情報管理テーブルＴ１と同様に、メモリ１９０２上に置かれると共に、管理端末装置２からも確認することができる。

【００５９】

図１１は、ディスク障害が発生した際に、ＣＰＵ１９０１が制御プログラム１９０３および管理テーブル１９０４に基づいて、障害復旧処理および次スペアディスク確保処理を示すフローチャートである。

【００６０】

すなわち、ＣＰＵ１９０１は、例えばデータの読み書き不可能な場合、電源が切れてしまった場合、データの読み出せない回数がある閾値を越えてしまった場合等のディスク障害を検出すると（Ｓ１０００）、ディスク情報管理テーブルＴ１における当該ディスクの状態を「障害」に変更する（Ｓ１００１）。また、ＣＰＵ１９０１は、ＲＡＩＤグループ情報管理テーブルＴ２において、当該ディスクの所属するＲＡＩＤグループの状態を「縮退」に変更する（Ｓ１００２）。そして、ＣＰＵ１９０１は、管理・通知プログラム１９１４を呼び出し、管理用端末装置２に、当該ディスクの障害および当該ＲＡＩＤグループが「縮退」であることを通知する（Ｓ１００３）。

【００６１】

次に、ＣＰＵ１９０１は、ディスク管理情報テーブルＴ１を用いて、装置内のスペアディスクを検索する（Ｓ１００４）。スペアディスクが存在した場合（Ｓ１００５：ＹＥＳ）、ＣＰＵ１９０１は、ディスク情報管理テーブルＴ１における当該スペアディスクの所属ＲＡＩＤグループの番号－１（スペアディスク）表示を、障害が発生しているＲＡＩＤグループの番号に変更し（Ｓ１００６）、当該スペアディスク上に障害が発生したディスクの内容を復元する。すなわち、ＣＰＵ１９０１は、ＲＡＩＤの再構築を行う（Ｓ１００７）。

【００６２】

再構築が完了すると、ＣＰＵ１９０１は、ディスク情報管理テーブルＴ１を用いて、障害ディスクの所属するＲＡＩＤグループの番号を－２（障害ディスク）に変更する（Ｓ１００８）。また、ＣＰＵ１９０１は、ＲＡＩＤグループ情報管理テーブルＴ２を用いて、当該ＲＡＩＤグループの状態を「縮退」から「正常」に変更する（Ｓ１００９）。

【００６３】

次に、ＣＰＵ１９０１は、管理・通知プログラム１９１４を呼び出し、管理用端末装置２に、当該ＲＡＩＤグループが正常に復旧したことを通知する（Ｓ１０１０）。

【００６４】

続いて、ＣＰＵ１９０１は、既に上述で使用した１台のスペアディスクの他に、残りのスペアディスクを検索するために、２回目のスペアディスク検索を行う（Ｓ１０１１）。スペアディスクが存在した場合は（Ｓ１０１２：ＹＥＳ）、障害復旧処理を終了する。

【００６５】

一方、ステップＳ１００５の１回目のスペアディスク検索において、スペアディスクが存在しない場合には（Ｓ１００５：ＮＯ）、ＣＰＵ１９０１は、外部に交換用の他ストレージ装置が接続されているかを確認する。交換用の他ストレージ装置が接続されている場合には（Ｓ１１００：ＹＥＳ）、ＣＰＵ１９０１は、交換用の他ストレージ装置へのデータ移行を開始する。接続されていない場合には（Ｓ１１００：ＮＯ）、ＣＰＵ１９０１は、管理・通知プログラム１９１４を呼び出し、管理用端末装置２に、交換用次装置が必要であることを通知し（Ｓ１１０２）、障害復旧処理を終了する。

【００６６】

一方、ステップＳ１０１２の２回目のスペアディスク検索において、スペアディスクが存在しない場合には（Ｓ１０１２：ＮＯ）、ＣＰＵ１９０１は、ディスク情報管理テーブルＴ１およびＲＡＩＤグループ情報管理テーブルＴ２を用いてグラニュラリティ縮小可能なＲＡＩＤグループが存在するかを判定する（Ｓ１２００）。

【００６７】

グラニュラリティ縮小可能なＲＡＩＤグループが存在する場合（Ｓ１２０１：ＹＥＳ）、ＣＰＵ１９０１は、最初に見つかったＲＡＩＤグループもしくは、管理端末装置２によって優先度が高く設定されたＲＡＩＤグループのグラニュラリティ縮小を実施し（Ｓ１２０２）、障害復旧処理を終了する。

【００６８】

グラニュラリティ縮小可能なＲＡＩＤグループが存在しない場合（Ｓ１２０１：ＮＯ）、ＣＰＵ１９０１は、管理・通知プログラム１９１４を呼び出し、管理端末装置２に、交換用次装置が必要であることを通知し（Ｓ１１０２）、障害復旧処理を終了する。

【００６９】

図１２は、本発明の第１の実施の形態において、ＣＰＵ１９０１が制御プログラム１９０３および管理テーブル１９０４に基づき、グラニュラリティ縮小判定を示すフローチャートである。グラニュラリティ縮小とは、ＲＡＩＤ構成におけるデータディスクの台数を減少させることを目的に、ストライプブロックおよびパリティブロック内に格納されたデータの移動を行う処理である。第１の実施の形態においては、例えば、ＲＡＩＤ－５にて４Ｄ+1Ｐの構成を３Ｄ+1Ｐに変更するようなケースを指す。以下、フローチャートを用いて説明していく。

【００７０】

すなわち、ＣＰＵ１９０１は、当該ＲＡＩＤグループに存在するストライプ列数ａをａ＝（１台のディスク容量）÷（ストライプブロックサイズ）にて求める（Ｓ２０００）。ここで、１台のディスク容量とは、当該ＲＡＩＤグループを構成するディスク１台の物理容量を指しており、また、ストライプブロックサイズとは、ＣＰＵ１９０１があらかじめ保持する定数である。ＣＰＵ１９０１は、ストライプ列数ａは小数を含む可能性がある値であり、ａの小数点以下を切り上げ、整数とした値を変数Ｘに格納する（Ｓ２００１）。

【００７１】

次に、ＣＰＵ１９０１は、グラニュラリティ縮小後のストライプ列数ｂをｂ＝（現在の列数）×（増加割合）で求める。すなわち、ストライプ列数ｂは、ｂ＝（ＬＵ割当済容量／ストライプブロックサイズ／データディスク数）×｛（データディスク数＋パリティディスク数）／（データディスク数＋パリティディスク数－１）｝で求める（Ｓ２００２）。ここで、パリティディスク数とは、当該ＲＡＩＤグループのＲＡＩＤレベルにより自動的に定まる値であり、例えばＲＡＩＤ－５ならば１である。ストライプ列数ｂは小数を含む可能性がある値であり、ｂの小数点以下を切り捨て、整数とした値をＹに格納する（Ｓ２００３）。

【００７２】

次にＣＰＵ１９０１は、ＸとＹを比較し、Ｙ≧Ｘであった場合には（Ｓ２００４：ＹＥＳ）、グラニュラリティ縮小可能と判定する（Ｓ２００５）。一方、Ｙ＜Ｘであった場合には（Ｓ２００４：ＮＯ）、ＣＰＵ１９０１は、グラニュラリティ縮小不可能と判定する（Ｓ２００７）。

【００７３】

次に、ＣＰＵ１９０１は、全ＲＡＩＤグループを調査したかを判定し（Ｓ２００６）、全て調査を終えている場合は（Ｓ２００６：ＹＥＳ）、終了する。終えていない場合は（Ｓ２００６：ＮＯ）、ＣＰＵ１９０１は、別のＲＡＩＤグループに対して上記判定を繰り返す。

【００７４】

図１３は、本発明の第１の実施の形態において、ＣＰＵ１９０１が制御プログラム１９０３および管理テーブル１９０４に基づいてグラニュラリティ縮小を行う際の手順を示すフローチャートである。

【００７５】

ＣＰＵ１９０１は、当該ＲＡＩＤグループを構成するディスクよりストライプ列１列目のデータブロックをデータバッファ１９０８にリードする（Ｓ３０００）。例えば、図１７に示すように、ＣＰＵ１９０１はデータブロックｄ１～ｄ４までをデータバッファ１９０８にリードする。続いて、ＣＰＵ１９０１は、次のストライプ列のデータブロックをデータバッファ１９０８にリードする（Ｓ３００１）。例えば、図１８に示すように、ＣＰＵ１９０１はデータブロックｄ５～ｄ８までをデータバッファ１９０８にリードする。そして、ＣＰＵ１９０１は、データバッファ１９０８に蓄えられたデータのうち、（データディスク数－１）このデータブロック群を用いて、新パリティｐ´ｘを生成する（Ｓ３００２）。例えば、図１９に示すように、ＣＰＵ１９０１はデータディスク数４－１のデータブロック群ｄ１～ｄ３のうち新パリティｐ´１を生成する。次に、（データディスク数－１）個のデータブロック群と新パリティｐ´ｘで構成される新ストライプ列をディスクにライトする（Ｓ３００３）。

【００７６】

現在リードしてあるストライプ列がＬＵ割当済領域の最終ストライプであった場合には（Ｓ３００４：ＹＥＳ）、ＣＰＵ１９０１は、データバッファ１９０８にある残りのデータブロックから新パリティｐ´ｘを生成し（Ｓ３００５）、ディスクにライトする（Ｓ３００６）。

【００７７】

次に、ＣＰＵ１９０１は、ＲＡＩＤグループ情報管理テーブルＴ２を用いて、当該ＲＡＩＤグループのデータディスク数を１つ減少させる（Ｓ３００７）。また、ＣＰＵ１９０

１は、ディスク情報管理テーブルＴ１を用いて、上記ストライプ列の書き換えで未使用となったディスクの所属ＲＡＩＤグループ番号を－１（スペアディスク）に変更し（Ｓ３００８）、処理を終える。

【００７８】

一方、ステップＳ３００４において、最終ストライプ列でなかった場合には（Ｓ３００４：ＮＯ）、ＣＰＵ１９０１は、ステップＳ３００１に戻り、処理を繰り返す。

【００７９】

図１４～図１６は、前記図１３で説明したグラニュラリティ縮小処理時の、ディスク上のデータブロックｄおよびパリティブロックｐの配置状態を示した図である。

【００８０】

図１４は、グラニュラリティ縮小処理前のデータ配置を示している。データブロックｄおよびパリティブロックｐは、ディスク０～ディスク４上に配置されているのがわかる。

【００８１】

図１５は、ＣＰＵ１９０１がステップＳ３０００～Ｓ３００３を実行することにより、１列目が新ストライプ列に書き換えられた状態を示している。１列目においてデータブロックｄ１、ｄ２、ｄ３はディスク０、１、２に配置され、パリティブロックｐ´１はディスク３に配置されており、そして、ディスク４のブロックは「空き」になっている。

【００８２】

図１６は、グラニュラリティ縮小処理が終了した際のデータブロックｄおよびパリティブロックｐの配置状態を示した図である。データブロックｄ１、ｄ２、ｄ３およびパリティブロックｐ´１は、ディスク０～ディスク３上にのみ配置され、ディスク４は「空き」、すなわち未使用となっていることがわかる。そして、２列目においても、データブロックｄ３、ｄ４、ｄ５およびパリティブロックｐ´２は、ディスク０～ディスク３上にのみ配置され、ディスク４は「空き」となっていることがわかる。３列目以降もディスク４は「空き」が形成されることがわかる。これによりＣＰＵ１９０１が、ステップＳ３００７およびＳ３００８の処理を実行することで、ディスク４を新たなスペアディスクにすることができる。

【００８３】

図１７～図１９は、ステップＳ３０００～Ｓ３００２におけるデータバッファ１９０８を表した図である。

【００８４】

図１７は、ステップＳ３０００にて、ストライプ列一列目のデータブロックｄ１、ｄ２、ｄ３、ｄ４が読み込まれた状態を表している。

【００８５】

図１８は、ステップＳ３００１にて、次のストライプ列のデータブロックｄ５、ｄ６、ｄ７、ｄ８が読み込まれた状態を表している。

【００８６】

図１９は、（データディスク数－１）個のデータブロックｄ１、ｄ２、ｄ３から新パリティｐ´１が生成された状態を表している。

【００８７】

以上のように本実施の形態によるデータ処理システムでは、スペアディスクが存在しない場合に、グラニュラリティ縮小を行うことで既存のディスクからスペアディスクを確保することができる。これにより、ストレージ装置の耐障害性を長期的に維持することができる。また、既存のディスクからスペアディスクを確保するので、ストレージ装置の小型化が実現できる。

【００８８】

一方、耐障害性を維持できなくなる可能性が生じた場合（スペアディスクを確保できなくなった場合）には、通知により交換用ストレージ装置の準備の必要性を知ることができる。

【００８９】

２．第２の実施の形態

第２の実施の形態として、ＲＡＩＤレベルを変更することにより、新たなスペアディスクを確保する方法がある。例えば、ＲＡＩＤ－６（ｎＤ+２Ｐ）をＲＡＩＤ－５（ｎＤ＋１Ｐ）に変更し、この差で生じる１ディスクをスペアディスクとする方法である。本実施の形態について図２０～図２３を用いて説明する。

【００９０】

図２０は、本実施の形態において、ディスク障害が発生した際に、ＣＰＵ１９０１が制御プログラム１９０３および管理テーブル１９０４に基づいて、障害復旧処理および次スペアディスク確保処理を示すフローチャートである。

【００９１】

ステップＳ１０００～Ｓ１０１２、およびステップＳ１１００～Ｓ１１０２については、先に示した第１の実施の形態と同じなので説明を割愛する。

【００９２】

そして、ステップＳ１０１２の２回目のスペアディスク検索においてスペアディスクが存在しない場合（Ｓ１０１２：ＮＯ）、ＣＰＵ１９０１は、ＲＡＩＤレベルを変更可能なＲＡＩＤグループ存在するか、ＲＡＩＤ情報管理テーブルＴ２を用いて検索し、存在した場合には（Ｓ４１００：ＹＥＳ）、該当ＲＡＩＤグループのＲＡＩＤレベル変更を行う（Ｓ４１０１）。なお、該当ＲＡＩＤグループが複数存在する場合には、最初に見つかったＲＡＩＤグループか、管理用端末装置２にて、対象優先度が高く設定されているＲＡＩＤグループに対して処理を行う。

【００９３】

図２１は、本実施の形態におけるＣＰＵ１９０１が制御プログラム１９０３および管理テーブル１９０４に基づいて行うＲＡＩＤレベル変更の手順を示すフローチャートである。

【００９４】

ＣＰＵ１９０１は、ストライプ列１列目のデータブロックｄおよびパリティブロックｐ、ｑをデータバッファ１９０８にリードすると（Ｓ５０００）、パリティブロックｑをデータバッファ１９０８より破棄する。例えば、図２２に示すように、ＣＰＵ１９０１は、ディスク６～ディスク１０に格納されたストライプ列１列目のデータブロックｄ１、ｄ２、ｄ３、パリティブロックｐ１、ｑ１から、図２３に示すように、パリティブロックｑ１を破棄する。次に、ＣＰＵ１９０１は、データブロックｄとパリティブロックｐから構成される新ストライプ列をディスクにライトする（Ｓ５００２）。現在データバッファ１９０８にリードされたストライプ列がＬＵ割当済領域の最終ストライプ列であった場合には（Ｓ５００３：ＹＥＳ）、ＣＰＵ１９０１は、残りのデータブロックｄとパリティブロックｐをディスクにライトする（Ｓ５００４）。

【００９５】

後に、ＣＰＵ１９０１は、ＲＡＩＤグループ情報管理テーブルＴ２を用いて、当該ＲＡＩＤグループのＲＡＩＤレベルの記載を変更する（Ｓ５００５）。次に、ＣＰＵ１９０１は、ディスク情報管理テーブルＴ１を用いて、未使用となったディスクの所属ＲＡＩＤグループ番号を－１（スペアディスク）に変更し（Ｓ５００６）、処理を終える。

【００９６】

一方、ＣＰＵ１９０１は、ステップＳ５００３において、現在データバッファ１９０８にリードされたストライプ列がＬＵ割当済領域の最終ストライプ列でなかった場合には（Ｓ５００３：ＮＯ）、次のストライプ列のデータブロックｄおよびパリティブロックｐ、ｑをデータバッファ１９０８にリードし（Ｓ５００７）、Ｓ５００１に戻って処理を繰り返す。

【００９７】

図２２、図２３は、本実施の形態による、ディスク上のデータブロックｄおよびパリティブロックｐ、ｑの変化を示す図である。

【００９８】

図２２はＲＡＩＤレベル変更を行う前のＲＡＩＤグループの状態である。ＲＡＩＤグループはＲＡＩＤ－６のデータ構成をとっており、データブロックｄと２種のパリティブロックｐ、ｑがディスク６～１０上に分散配置されている。

【００９９】

図２３は、図１９、図２１に示したＲＡＩＤレベル変更の処理により、ＲＡＩＤ－５のデータ構成に変更されたあとのデータ配置を示している。ＲＡＩＤ－６（２パリティ）からＲＡＩＤ－５に変更されたため、パリティブロックｑが破棄され、データブロックｄとパリティブロックｐのみで構成されている。また、データブロックｄおよびパリティブロックｐはディスク６～ディスク９上に配置されており、ディスク１０が「空き」すなわち未使用となっている。従い、ＣＰＵ１９０１が前述のステップＳ５００５、Ｓ５００６を実行することにより、ディスク１０がスペアディスクに設定される。

【００１００】

以上のように本実施の形態によるデータ処理システムでは、スペアディスクが存在しない場合に、ＲＡＩＤのレベル変更を行うことでより安定した性能が実現できるとともに、既存のディスクからスペアディスクを確保することができる。これにより、ストレージ装置の耐障害性を長期的に維持することができる。また、既存のディスクからスペアディスクを確保するので、ストレージ装置の小型化が実現できる。

【００１０１】

一方、耐障害性を維持できなくなる可能性が生じた場合（スペアディスクを確保できなくなった場合）には、通知により交換用ストレージ装置の準備の必要性を知ることができる。

【００１０２】

３．第３の実施の形態

第３の実施の形態として、スペアの領域をディスクとして確保せずＬＵとして確保する方法がある。第１及び第２の実施の形態では物理ディスク上に障害ディスクの内容を復元できるようにしたが、本実施の形態では、ＬＵ、すなわち論理ディスク上に障害ディスクの内容を復元できるようにする。ＬＵとして配されるスペア領域（以下、スペアＬＵ）を確保する方法は、第１及び第２の実施の形態の様に、既存のデータブロックを移動させずとも容易にスペア領域を確保できる点で優れている。一方、障害ディスク上にあった元のＬＵのアドレス（ＬＢＡ）をスペアＬＵ上のアドレスに変換してアクセスするため、性能は、第１及び第２の実施の形態に比べ、低下する恐れがある。以下、図２４～図２７を用いて説明する。

【００１０３】

図２４は、第３の実施の形態におけるメモリ１９０２に格納される制御プログラム１９０３および管理テーブル１９０４を示す説明図である。

【００１０４】

第３の実施の形態における制御プログラム１９０３は、ＲＡＩＤ制御プログラム１９１１、ＲＡＩＤグループ設定プログラム１９１２、ＬＵ設定プログラム１９１３、管理・通知プログラム１９１４から構成される。

【００１０５】

また管理テーブル１９０４は、ディスク情報管理テーブルＴ１、ＲＡＩＤグループ情報管理テーブルＴ２、ＬＵ情報管理テーブルＴ３、スペアＬＵ管理テーブルＴ４、ＬＢＡ－ＬＢＡ変換テーブルＴ５から構成される。

【００１０６】

図２５はスペアＬＵ管理テーブルＴ４の説明図である。スペアＬＵ管理テーブルＴ４は、ＬＵとして配されるスペア領域の情報を記録するテーブルである。具体的には、スペアＬＵ管理テーブルＴ４は、スペアＬＵの番号、容量、割当元ＲＡＩＤグループの番号、開始ＬＢＡ、終了ＬＢＡ、スペアの使用状態（「使用済」か「未使用」か）を記録する。例えば、スペアＬＵ番号０のスペアＬＵは容量が５００ＧＢであり、割当元ＲＡＩＤグループ番号は０であり、開始ＬＢＡは０００００００００ｈであり、終了ＬＢＡは０００４Ｆ０

００ｈであり、使用状態は「未使用」である。

【０１０７】

なお、スペアＬＵ管理テーブルＴ４は、ＬＵ情報管理テーブルＴ３と共にメモリ１９０２におかれる。また、管理用端末装置２を通じて、その情報を確認することができる。

【０１０８】

図２６は、ＬＢＡ‐ＬＢＡ変換テーブルＴ５の説明図である。ＬＢＡ‐ＬＢＡ変換テーブルＴ５は、元のＬＵのアドレス（ＬＢＡ）とスペアＬＵ上のアドレス（ＬＢＡ）との対応を管理するテーブルである。具体的には、ＬＢＡ－ＬＢＡ変換テーブルＴ５は、アクセス元ＬＵの番号、アクセス元のＬＢＡ、スペアＬＵの番号及びアクセス先のＬＢＡから構成され、スペアＬＵのＬＢＡを管理する。例えば、アクセス元ＬＵ１のＬＢＡ００３５６７８０ｈのデータに対するアクセスは、スペアＬＵ０のＬＢＡ１０００４７８０ｈに対するアクセスに変換される。

【０１０９】

図２７は、本実施の形態において、ディスク障害が発生した際に、ＣＰＵ１９０１が制御プログラム１９０３および管理テーブル１９０４に基づいて行う、障害復旧処理およびスペアＬＵの確保処理を示すフローチャートである。

【０１１０】

ステップＳ１０００～Ｓ１０１２、およびステップＳ１１００～Ｓ１１０２については、先に示した実施の形態と同じなので説明を割愛する。

【０１１１】

ステップＳ１００５の１回目のスペアディスク検索においてスペアディスクが存在しない場合（Ｓ１００５：ＮＯ）、ＣＰＵ１９０１はスペアＬＵ管理テーブルＴ４を用いて、未使用のスペアＬＵを検索する（Ｓ６１００）。未使用スペアＬＵが存在した場合（Ｓ６１０１：ＹＥＳ）、ＣＰＵ１９０１は、障害ディスクのデータをスペアＬＵに復元し、ＲＡＩＤ回復を行う（Ｓ６１０２）。

【０１１２】

次にＣＰＵ１９０１は、当該スペアＬＵのＬＢＡ－ＬＢＡ変換テーブルＴ５を更新生成する（Ｓ６１０３）。そしてＣＰＵ１９０１は、スペアＬＵ管理テーブルＴ４を用いて、当該スペアＬＵの使用状態を「使用済」に変更し（Ｓ６１０４）、ステップＳ１００８以降の処理を行う。

【０１１３】

一方、ステップＳ１０１２の２回目のスペアディスク検索において、スペアディスクが存在しない場合（Ｓ１０１２：ＮＯ）、ＣＰＵ１９０１は、ディスク情報管理テーブルＴ１およびＲＡＩＤ情報管理テーブルＴ２を用いて、ディスク１台分の容量以上の未使用容量を持つ、ＲＡＩＤグループを検索する（Ｓ６２００）。このとき、ディスク障害が発生したディスクが所属するＲＡＩＤグループに未使用容量のＬＵが存在していても、基本的には当該未使用容量のＬＵをスペアディスクとして使用せず、他のＲＡＩＤグループのＬＵを検索して使用する。但し、当該未使用容量のＬＵをスペアディスクとして使用する必要がある場合には、ディスク１台分のＬＵに加え、ディスク障害が発生したディスクの所属するＲＡＩＤグループを縮退してもなおも使用できる未使用容量が存在する場合に限り、未使用容量のＬＵをスペアディスクとして割り当てる。

【０１１４】

該当ＲＡＩＤグループが存在した場合（Ｓ６２０１：ＹＥＳ）、ＣＰＵ１９０１は、ＲＡＩＤ情報管理テーブルＴ２を用いて、該当ＲＡＩＤグループのＬＵ割当済容量と未使用容量を変更し、スペアＬＵの容量確保を行う。具体的には、ＣＰＵ１９０１は、ＬＵ割当済容量をディスク１台分の容量だけ増加させ、未使用容量をディスク１台分の容量だけ減少させる（Ｓ６２０２）。例えば、ＣＰＵ１９０１は、スペアＬＵの容量として１００ＧＢの確保が必要な場合には、ＲＡＩＤ情報管理テーブルＴ２のＬＵ割当済容量を１００ＧＢだけ増加させ、未割当容量から１００ＧＢだけ減少させる。

【０１１５】

次に、ＣＰＵ１９０１は、スペアＬＵ管理テーブルＴ４にステップＳ６２０２で確保したＬＵを追加登録し（Ｓ６２０３）、処理を終える。

【０１１６】

以上のように本実施の形態によるデータ処理システムでは、スペアディスクが存在しない場合に、ディスク１台分以上の未使用容量を有するＲＡＩＤグループを検索することで既存のディスクからスペアＬＵ（スペア領域）を確保することができる。これにより、ストレージ装置の耐障害性を長期的に維持することができる。また、既存のディスクからスペア領域を確保するので、ストレージ装置の小型化が実現できる。

【０１１７】

さらに、耐障害性を維持できなくなる可能性が生じた場合（スペアディスクを確保できなくなった場合）、通知により交換用ストレージ装置の準備の必要性を知ることができる。

【０１１８】

４．第４の実施の形態

第４の実施の形態として、容量を動的に割り当てるストレージ装置において適用する形態を説明する。

【０１１９】

容量を動的に割り当てるストレージ装置とは、複数のＲＡＩＤグループの容量をひとつの容量プールとして管理し、実際に書き込まれる部分にだけ動的にエクステントと呼ばれる、小容量同サイズのＬＵを割り当てる制御を行うストレージ装置である。ストレージ装置において、スペア領域は未使用エクステントの集合として形成される。第３の実施の形態において、スペアＬＵを確保するためには、ディスク１台の容量以上の未使用容量が単一のＲＡＩＤグループに必要であった。本実施の形態は、スペア領域をエクステントプール（エクステントの集合体）から確保するため、単一のＲＡＩＤグループである必要がなく、また不連続な領域からも確保できる。以下、図２８～図３５を用いて、本実施の形態について説明する。

【０１２０】

図２８は、第４の実施の形態におけるメモリ１９０２に格納される制御プログラム１９０３および管理テーブル１９０４を示す説明図である。

【０１２１】

第４の実施の形態における制御プログラム１９０３は、ＲＡＩＤ制御プログラム１９１１、ＲＡＩＤグループ設定プログラム１９１２、ＬＵ設定プログラム１９１３および管理・通知プログラム１９１４から構成される。

【０１２２】

また管理テーブル１９０４は、ディスク情報管理テーブルＴ１、ＲＡＩＤグループ情報管理テーブルＴ２、ＬＵ情報管理テーブルＴ３、エクステント管理テーブルＴ６、エクステント数管理テーブルＴ７、仮想ＬＵ用アクセス先エクステント対応テーブルＴ８およびスペアＬＵ用エクステント対応テーブルＴ９から構成される。

【０１２３】

図２９は、本実施の形態の概念を模式的に示した図である。本実施の形態のストレージ装置１は、ＬＵの代わりに仮想的な容量を持った仮想ＬＵをホスト計算機３に対して提供する。ホスト計算機３が仮想ＬＵに書き込みを行うと、ＣＰＵ１９０１は、容量プールから書き込まれたアドレスに対し、エクステントのディスパッチ（割当て）を行う。エクステントとは、ＲＡＩＤグループを分割してできる小容量同サイズのＬＵである。エクステントの集合体は、ＲＡＩＤグループの隔たりなく、１つのエクステントプールとして管理され、必要な分だけディスパッチされる。

【０１２４】

図３０は、エクステント管理テーブルＴ６の説明図である。エクステント管理テーブルＴ６は、エクステントプールの管理実態であり、エクステントの各種情報が記録される。具体的には、エクステント管理テーブルＴ６は、エクステントの番号、割当元のＲＡＩＤ

グループの番号、開始ＬＢＡ、割当状態（「割当済」か「未割当」か）である。例えば、エクステント番号「０」はＲＡＩＤグループ番号「０」上にあり、開始ＬＢＡは０００００００００ｈであり、割当状態は「割当済」である。なお、エクステントのサイズは予め決まっているので終了ＬＢＡの指定アドレス表示は必要ない。

【０１２５】

なお、エクステント管理テーブルＴ６はＲＡＩＤ情報管理テーブルＴ２と共に、メモリ１９０２内に配される。

【０１２６】

図３１は、エクステント数管理テーブルＴ７の説明図である。エクステント数管理テーブルＴ７は、エクステントの仮想ＬＵ領域への割当数、スペア領域への割当数、未使用数、総数を記録している。例えば、図３０の場合、仮想ＬＵ領域への割当数は、200,000,000,000個であり、スペア領域への割当数は0個であり、未使用数は、700,000,000,000個であり、総数は、900,000,000,000である。

【０１２７】

なお、エクステント数管理テーブルＴ７は、エクステント管理テーブルＴ６と同様にメモリ１９０２内に配される。

【０１２８】

図３２は、仮想ＬＵに対する仮想ＬＵ用アクセス先エクステント対応テーブルＴ８である。仮想ＬＵ用アクセス先エクステント対応テーブルＴ８は、仮想ＬＵ上のＬＢＡとディスパッチされたエクステントとの対応関係を記録するテーブルである。具体的には、仮想ＬＵ用アクセス先エクステント対応テーブルＴ８は、仮想ＬＵの番号と仮想ＬＵの開始ＬＢＡ、エクステントの番号で構成される。例えば、仮想ＬＵ０の開始ＬＢＡ００３５６７８０ｈにディスパッチされたエクステント番号は０である。

【０１２９】

仮想ＬＵ用アクセス先エクステント対応テーブルＴ８は、エクステント管理テーブルＴ６と同様に、メモリ１９０２内に配される。

【０１３０】

図３３は、スペアＬＵに対するスペアＬＵ用アクセス先エクステント対応テーブルＴ９である。スペアＬＵ用アクセス先エクステント対応テーブルＴ９は、スペアＬＵの番号、スペアＬＵの開始ＬＢＡ、エクステントの番号で構成される。例えば、スペアＬＵ０の開始ＬＢＡ００３５６７８０ｈにディスパッチされたエクステント番号は１８００である。

【０１３１】

図３４は、エクステントプールを容量表示した容量プールの管理画面である。容量プール管理画面Ｖ２は、管理用端末装置２からアクセスすることができる。

【０１３２】

容量プール管理画面Ｖ２は、仮想ＬＵに割り当てたエクステントと未使用のエクステントを容量表示しており、仮想ＬＵごとの使用率を容量で示している。例えば、仮想ＬＵ０は５００ＧＢ消費している。また未使用エクステントの総容量は１２００ＧＢである。容量プール画面Ｖ２はまた、スペア領域を確保により、未使用エクステント（未使用容量）が減少した際には、その旨を通知するメッセージウィンドウ１６を備えている。

【０１３３】

図３５は、ＣＰＵ１９０１が制御プログラム１９０３および管理テーブル１９０４に基づいて、エクステントを用いてスペアＬＵ用の領域を確保する手順を示すフローチャートである。なお、ステップＳ１０００～Ｓ１０１２、ステップＳ６１００～Ｓ６１０４およびステップＳ１１００～Ｓ１１０２の処理は、上述と同様なので説明は割愛する。

【０１３４】

ＣＰＵ１９０１は、１台分のディスク容量と（未割当数＊エクステントサイズ）とを比較し、未割当エクステントの集合体である未割当エクステント群からスペアＬＵを確保できるか否かを判別する（Ｓ７０００）。ここでエクステントサイズは、ＣＰＵ１９０１においてあらかじめ定められた定数である。

【０１３５】
１台分のディスク容量が（未割当数＊エクステントサイズ）と同じか、それより小さい場合、すなわち、１台分のディスク容量≦（未割当数＊エクステントサイズ）が成り立つ場合には（Ｓ７０００：ＹＥＳ）、ＣＰＵ１９０１は、エクステント数管理テーブルＴ７を用いて、未割当数を（１台分のディスク容量÷エクステントサイズ）分だけ減少させ、スペア領域割当数を（１台分のディスク容量÷エクステントサイズ）分だけ増加させ、エクステントを割当済に変更し（Ｓ７００１）、エクステントカウンタを変更する（Ｓ７００２）。すなわち、ＣＰＵ１９０１は、エクステント管理テーブルＴ６の割当状態を割当済に変更し、エクステント数管理テーブルＴ７の未使用数を減らし、スペア領域割当数を増やす。

【０１３６】
次に、ＣＰＵ１９０１は、スペアＬＵ用アクセス先エクステント対応テーブルＴ９を追加更新する（Ｓ７００３）。ＣＰＵ１９０１は、管理・通知プログラム１９１４を呼び出して、管理用端末装置２に、容量プールが減少した旨を通知し（Ｓ７００４）、容量プール管理画面Ｖ２にこの旨が表示され、この処理が終了する。

【０１３７】
一方、ステップＳ７０００において、１台分のディスク容量に見合う未使用領域が存在しない場合、すなわち、１台分のディスク容量＞（未割当数＊エクステントサイズ）となる場合（Ｓ７０００：ＮＯ）、ＣＰＵ１９０１は、管理・通知プログラム１９１４を呼び出して、管理用端末装置２に、交換用次装置が必要である旨を通知し（Ｓ７００５）、この処理が終了する。

【０１３８】
以上のように本実施の形態によるデータ処理システムでは、スペアディスクが存在しない場合に、エクステントプールから動的にスペア領域を確保することができる。これにより、ストレージ装置の耐障害性を長期的に維持することができる。また、既存のディスクからスペア領域を確保するので、ストレージ装置の小型化が実現できる。

【０１３９】
さらに、耐障害性を維持できなくなる可能性が生じた場合（スペアディスクを確保できなくなった場合）、通知により交換用ストレージ装置の準備の必要性を知ることができる。

【０１４０】
５．他の実施の形態
上述の第１から第４の実施の形態においては、ステップＳ１０００～Ｓ１０１２までを検索部とし、複数設けられた第１の記憶領域をデータブロック及びパリティブロックとして、複数設けられたデータブロック及びパリティブロックの一部領域に障害が発生した場合に当該一部領域に格納されたデータを、第２の記憶領域であるスペアディスク又はスペア領域に移行して動的に記憶領域を確保する確保部を備えた場合について述べてきたが、本発明はこれに限らず、当該一部領域に格納されるべきデータをスペアディスク又はスペア領域に格納するようにしても良い。

【産業上の利用可能性】

【０１４１】
本発明は、１又は複数のストレージ装置や、１又は複数のストレージ装置を有するデータ処理システムに広く適用することができる。

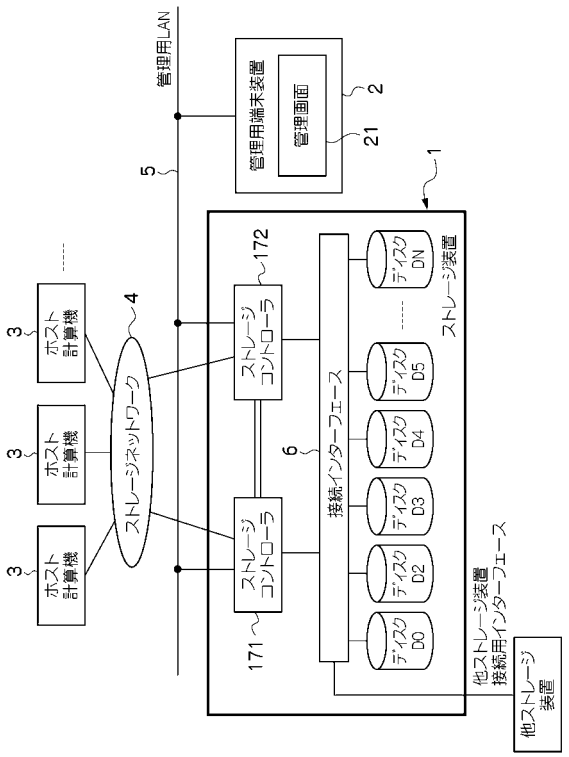
【図面の簡単な説明】

【０１４２】
【図１】本発明におけるストレージ装置の構成を表した図である。
【図２】本発明の第１の実施の形態のストレージ装置の構成を示す概観図である。
【図３】本発明の第１の実施の形態におけるストレージコントローラの内部構成を示すブロック図である。
【図４】本発明の第１の実施の形態におけるメモリに格納される制御プログラムおよび管

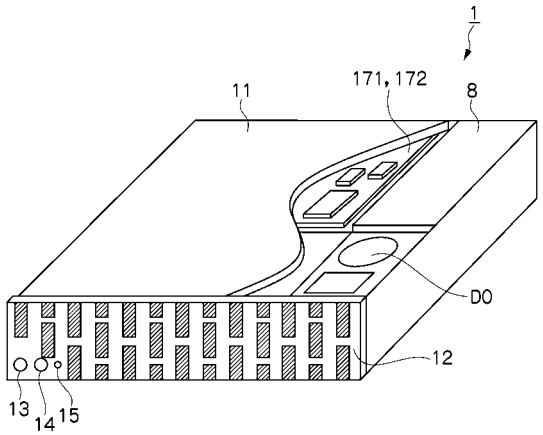
理テーブルを示す説明図である。
【図５】本発明の第１の実施の形態におけるディスク、ＲＡＩＤグループ、ＬＵの関係を模式的に表した図である。
【図６】本発明の第１の実施の形態における物理ディスクの格納状態を示す説明図である。
【図７】本発明の第１の実施の形態におけるディスク情報管理テーブルの説明図である。
【図８】本発明の第１の実施の形態における管理用端末装置の表示画面の画面例である。
【図９】本発明の第１の実施の形態におけるＲＡＩＤグループ情報管理テーブルの説明図である。
【図１０】本発明の第１の実施の形態におけるＬＵ情報管理テーブルの説明図である。
【図１１】本発明の第１の実施の形態におけるディスク障害が発生した際の、障害復旧処理および次スペアディスク確保処理を示すフローチャートである。
【図１２】本発明の第１の実施の形態におけるグラニュラリティ縮小判定を示すフローチャートである。
【図１３】本発明の第１の実施の形態におけるグラニュラリティ縮小処理を行う際の手順を示すフローチャートである。
【図１４】本発明の第１の実施の形態におけるグラニュラリティ縮小処理前のデータ配置を示した図である。
【図１５】本発明の第１の実施の形態におけるグラニュラリティ縮小処理により、１列目が新ストライプ列に書き換えられた状態を示した図である。
【図１６】本発明の第１の実施の形態におけるグラニュラリティ縮小処理が終了した際の、データブロックおよびパリティブロックの配置状態を示した図である。
【図１７】本発明の第１の実施の形態におけるグラニュラリティ縮小処理にて、ストライプ列一列目のデータブロックがデータバッファに読み込まれた状態を表した図である。
【図１８】本発明の第１の実施の形態におけるグラニュラリティ縮小処理にて、次のストライプ列のデータブロックがデータバッファに読み込まれた状態を表した図である。
【図１９】本発明の第１の実施の形態におけるデータバッファ上で新パリティが生成された状態図である。
【図２０】第２の実施の形態において、ディスク障害が発生した際の、障害復旧処理および次スペアディスク確保処理を示すフローチャートである。
【図２１】第２の実施の形態におけるＲＡＩＤレベル変更処理の手順を示すフローチャートである。
【図２２】第２の実施の形態において、ＲＡＩＤレベル変更処理を行う前のＲＡＩＤグループの状態である。
【図２３】第２の実施の形態において、ＲＡＩＤレベル変更処理により、ＲＡＩＤ－５のデータ構成に変更されたあとのデータ配置を示している。
【図２４】第３の実施の形態におけるメモリに格納される制御プログラムおよび管理テーブルを示す説明図である。
【図２５】第３の実施の形態における、スペアＬＵ管理テーブルの説明図である。
【図２６】第３の実施の形態における、ＬＢＡ‐ＬＢＡ変換テーブルの説明図である。
【図２７】第３の実施の形態において、ディスク障害が発生した際の、障害復旧処理およびスペアＬＵの確保処理を示すフローチャートである。
【図２８】第３の実施の形態におけるメモリに格納される制御プログラムおよび管理テーブルを示す説明図である。
【図２９】第４の実施の形態の概念を模式的に示した図である。
【図３０】第４の実施の形態における、エクステント管理テーブルの説明図である。
【図３１】第４の実施の形態における、エクステント数管理テーブルの説明図である。
【図３２】第４の実施の形態における、仮想ＬＵ用アクセス先エクステント対応テーブルの説明図である。
【図３３】第４の実施の形態における、スペアＬＵ用アクセス先エクステント対応テーブ

ルの説明図である。
【図３４】第４の実施の形態における、エクステントプールを容量表示した容量プール管理画面である。
【図３５】第４の実施の形態における、ディスク障害が発生した際の、障害復旧処理およびスペアＬＵの確保処理を示すフローチャートである。
【符号の説明】
【０１４３】
１……ストレージ装置、２……管理用端末装置、３……ホスト計算機、４……ストレージネットワーク、５……管理用ＬＡＮ、６……接続インターフェース、７……交換用ストレージの接続インターフェース、８……ファンおよび電源、１１……天板、１２……ベゼル、１３，１４……電源ボタン、１５……ブザー停止ボタン、１７１，１７２……ストレージコントローラ、１９０１……ＣＰＵ、１９０２……メモリ、１９０５……パリティジェネレータ及びデータ転送コントローラ、１９０６……フロントエンド接続インターフェースコントローラ、１９０７……バックエンド接続インターフェースコントローラ、１９０８……データバッファ、１９０９……ＬＡＮインターフェースコントローラ、Ｖ１……ディスク管理画面、Ｖ２……容量プール管理画面。

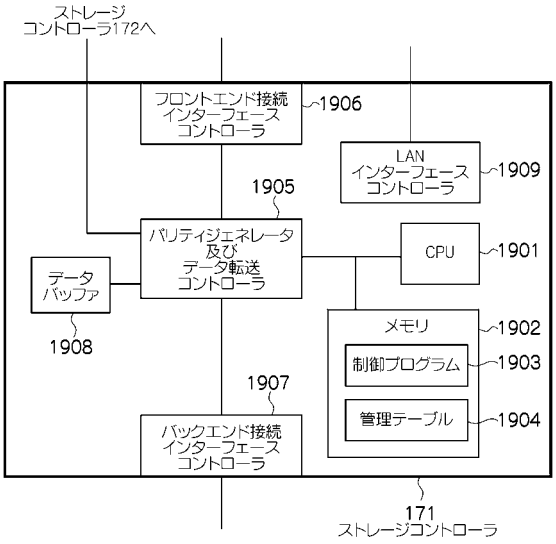
【図１】



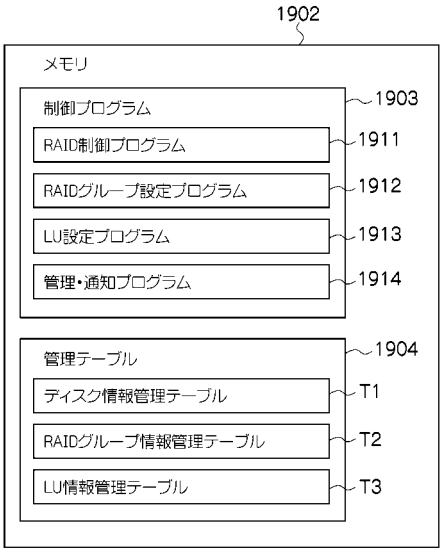
【図２】



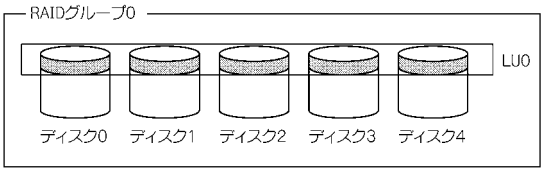
【図３】



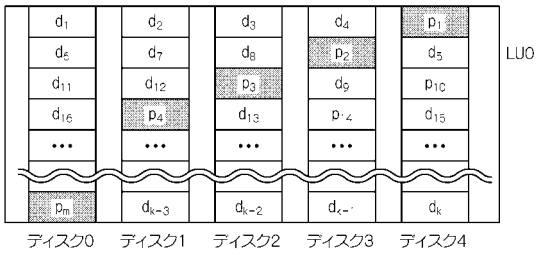
【図４】



【図５】



【図６】

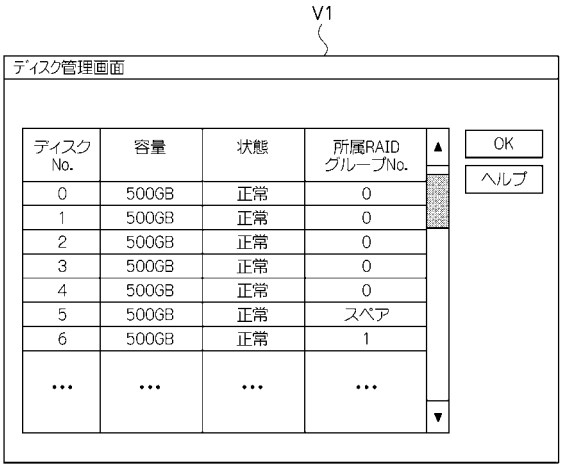


【図７】

T1

ディスクNo.	容量	状態	所属RAIDグループNo.
0	500GB	正常	0
1	500GB	正常	0
2	500GB	正常	0
3	500GB	正常	0
4	500GB	正常	0
5	500GB	正常	-1(スペア)
6	500GB	正常	1
…	…	…	…
N	-	未装着	-

【図８】



【図９】

T2

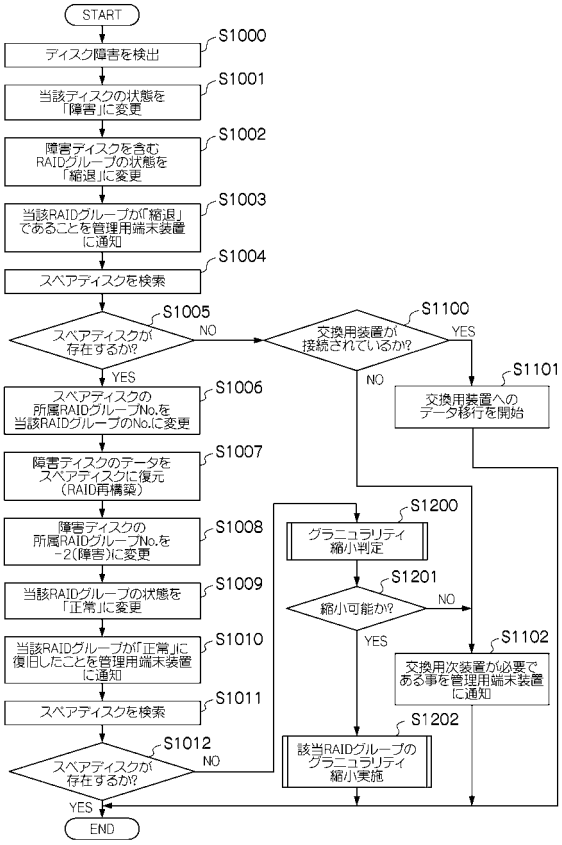
RAIDグループNo.	RAIDレベル	データディスク数	有効容量	状態	LU割当済容量	未割当容量
0	RAID-5	4	2000MB	正常	900GB	1100GB
1	RAID-6	3	1500MB	正常	1000GB	500GB
…	…	…	…	…	…	…

【図１０】

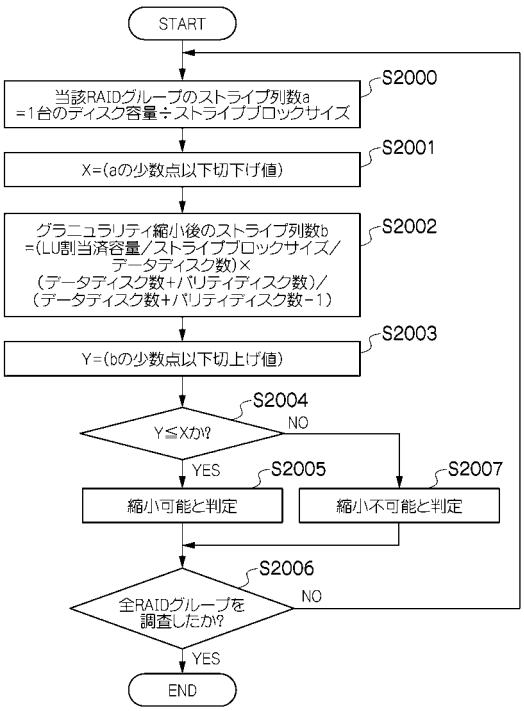
T3

LU No.	容量	RAID グループ No.	開始LBA	終了LBA
0	100GB	0	00000000h	0000F000h
1	300GB	0	00010000h	0003F000h
2	500GB	0	00040000h	0008F000h
…	…	…	…	…

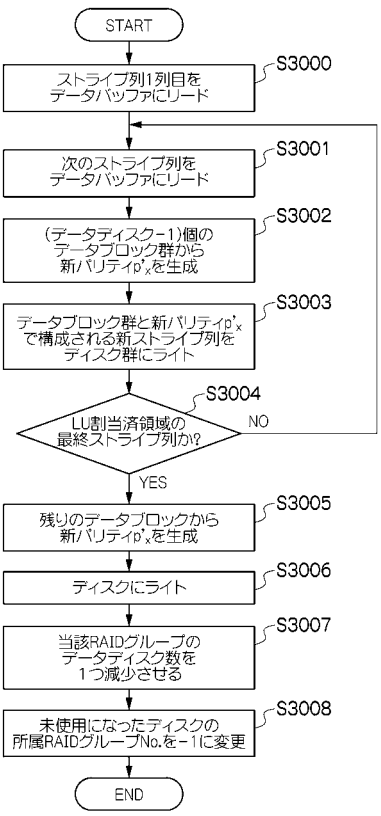
【図１１】



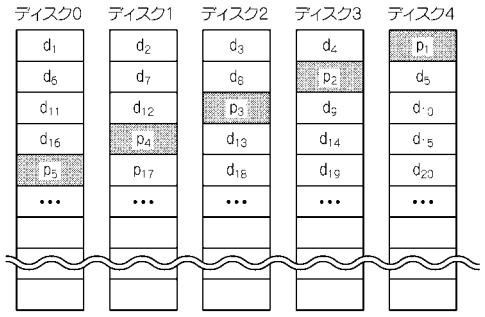
【図１２】



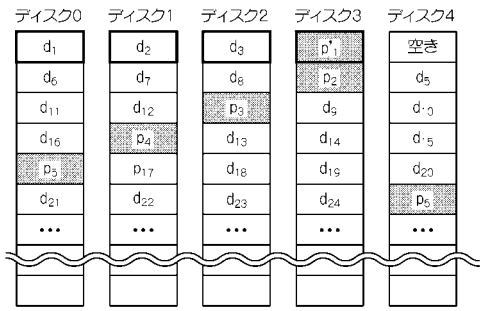
【図１３】



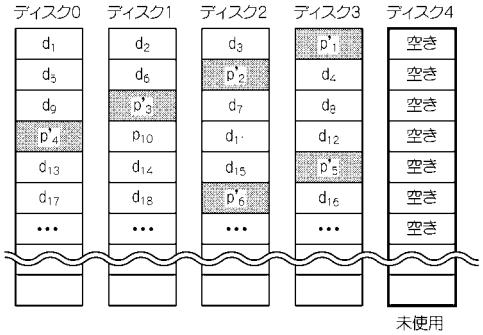
【図１４】



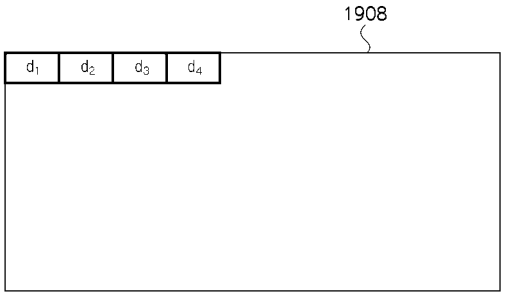
【図１５】



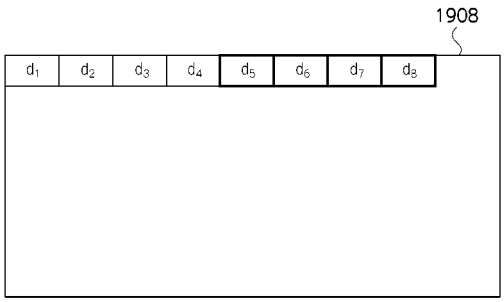
【図１６】



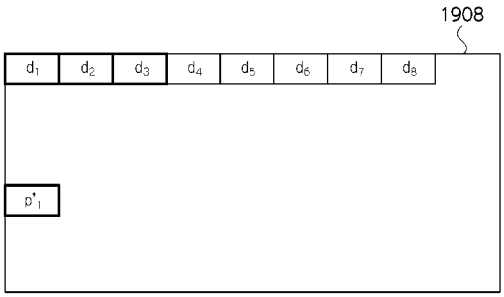
【図１７】



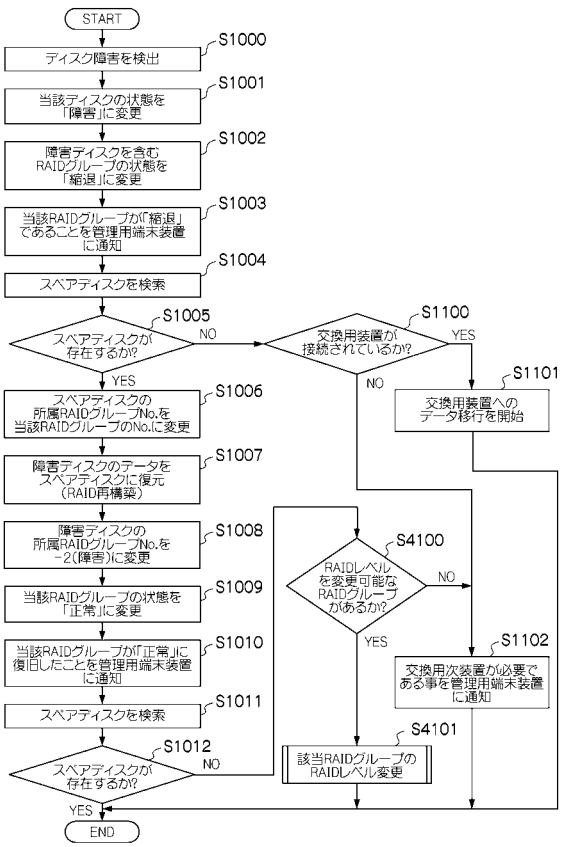
【図１８】



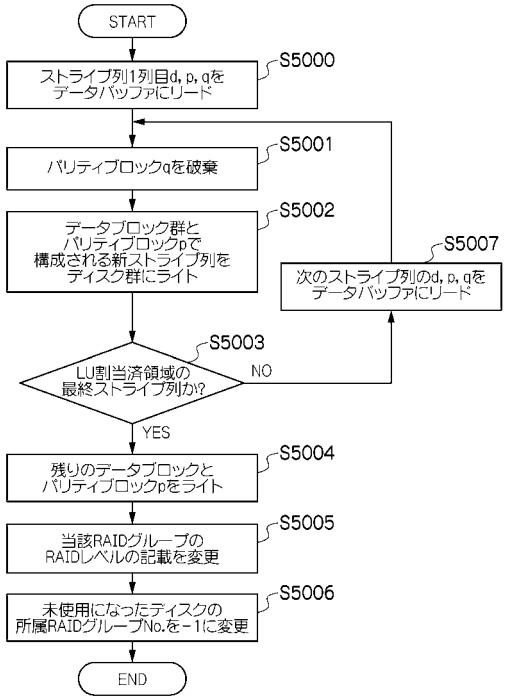
【図１９】



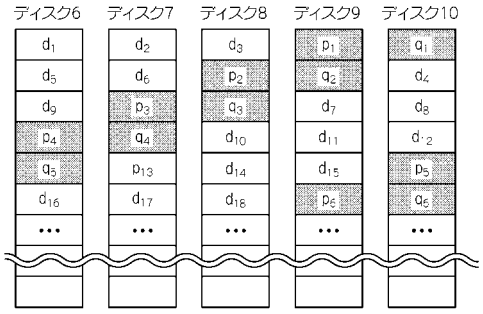
【図２０】



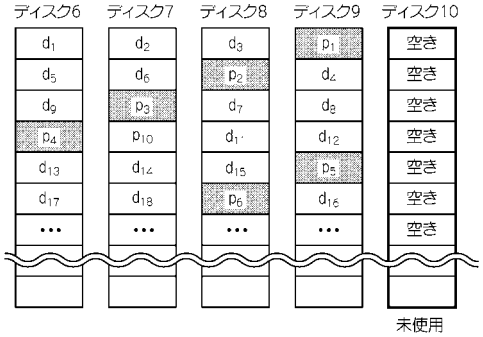
【図２１】



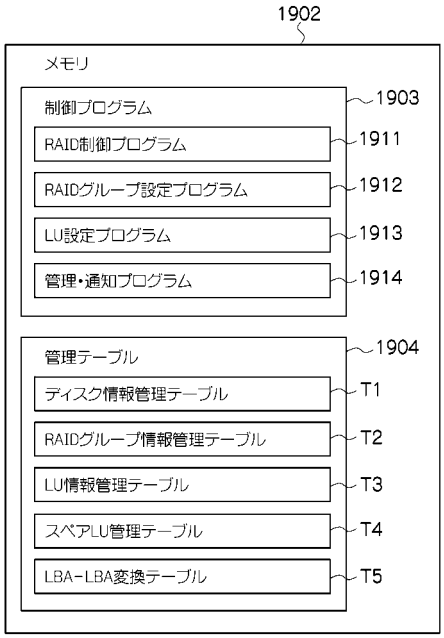
【図２２】



【図２３】



【図２４】



【図２５】

T4

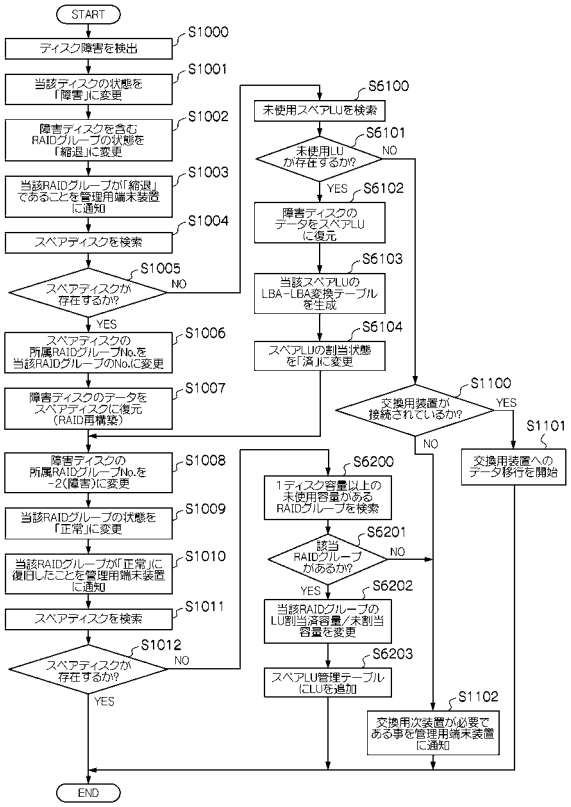
スペア LU No.	容量	RAID グループ No.	開始LBA	終了LBA	使用状態
0	500GB	0	00000000h	0004F000h	未使用
…	…	…	…	…	…

【図２６】

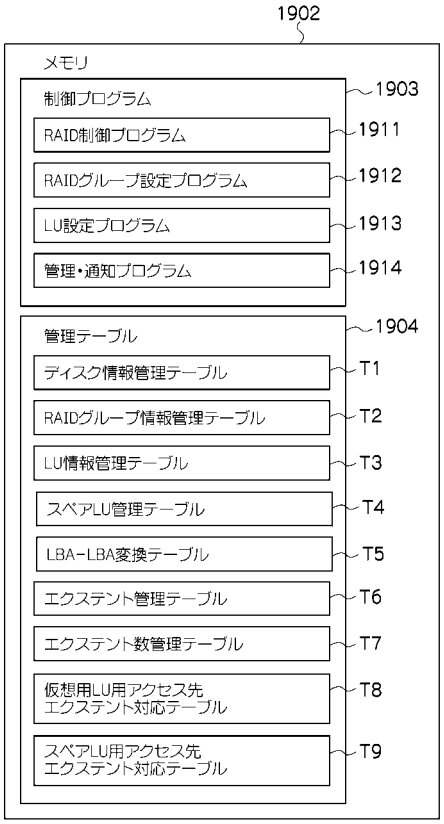
T5

アクセス元 LU No.	アクセス元LBA	スペア LU No.	アクセス先LBA
…	…	…	…
1	00356780h	1	10004780h
1	00356790h	1	10004790h
…	…	…	…

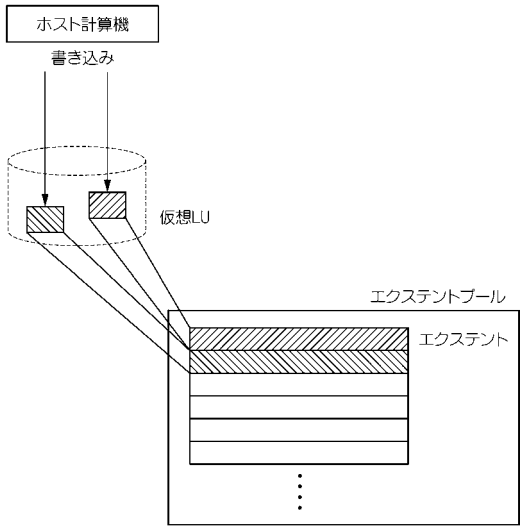
【図２７】



【図２８】



【図２９】



【図３１】

T7

ユーザー領域割当数	200,000,000,000
スペア領域割当数	0
未使用数	700,000,000,000
総数	900,000,000,000

【図３２】

T8

仮想LU No.	LBA	エクステント No.
・・・	・・・	・・・
0	00356780h	0
0	02466790h	1
・・・	・・・	・・・

【図３０】

T6

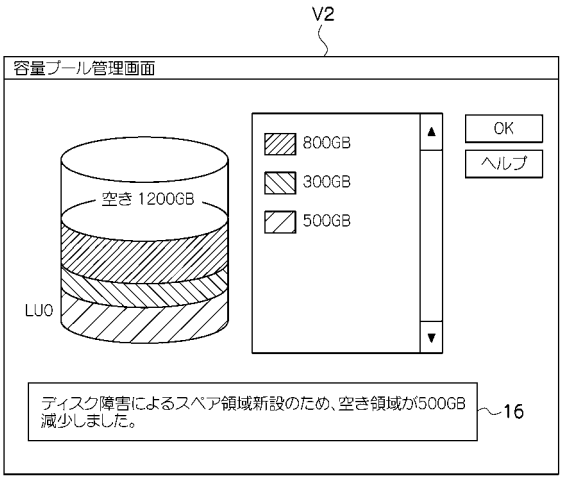
エクステント No.	RAID グループ No.	開始LBA	割当状態
0	0	00000000h	割当済
1	0	00000400h	割当済
2	0	00000800h	未割当
・・・	・・・	・・・	・・・

【図３３】

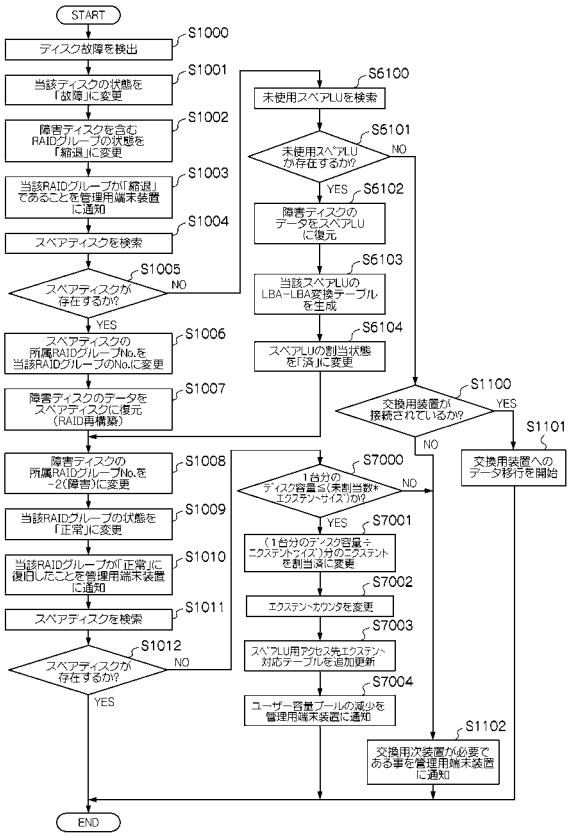
T9

スペアLU No.	開始LBA	エクステント No.
・・・	・・・	・・・
0	00356780h	1800
0	00356790h	8310
・・・	・・・	・・・

【図３４】


V2
容量プール管理画面
空き 1200GB
LU0
800GB
300GB
500GB
OK
ヘルプ
ディスク障害によるスペア領域新設のため、空き領域が500GB減少しました。
16

【図３５】


START
S1000
ディスク故障を検出
S1001
当該ディスクの状態を「故障」に変更
S1002
障害ディスクを含むRAIDグループの状態を「縮退」に変更
S1003
当該RAIDグループが「縮退」であることを管理用端末装置に通知
S1004
スペアディスクを検索
S1005
スペアディスクが存在するか?
NO
YES
S1006
スペアディスクの所属RAIDグループNo.を当該RAIDグループのNo.に変更
S1007
障害ディスクのデータをスペアディスクに復元（RAID再構築）
S1008
障害ディスクの所属RAIDグループNo.を"-2(障害)"に変更
S1009
当該RAIDグループの状態を「正常」に変更
S1010
当該RAIDグループが「正常」に復旧したことを管理用端末装置に通知
S1011
スペアディスクを検索
S1012
スペアディスクが存在するか?
NO
YES
S6100
未使用スペアLUを検索
S6101
未使用スペアLUが存在するか?
NO
YES
S6102
障害ディスクのデータをスペアLUに復元
S6103
当該スペアLUのLBA−LBA変換テーブルを生成
S6104
スペアLUの割当状態を「済」に変更
S1100
交換用装置が接続されているか?
YES
NO
S1101
交換用装置へのデータ移行を開始
S7000
1台分のディスク容量≦(未割当数*エクステントサイズ)か?
NO
YES
S7001
(1台分のディスク容量÷エクステントサイズ)分のエクステントを割当済に変更
S7002
エクステンへカウンタを変更
S7003
スペアLU用アクセス先エクステンへ対応テーブルを追加更新
S7004
ユーザー容量プールの減少を管理用端末装置に通知
S1102
交換用装置が必要である事を管理用端末装置に通知
END