(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5438827号
(P5438827)

(45)発行日　平成26年3月12日(2014.3.12)　(24)登録日　平成25年12月20日(2013.12.20)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 301J
G06F 3/06 301E
G06F 3/06 301Z
G06F 3/06 540
G06F 3/06 301M

請求項の数 13　(全 40 頁)

(21)出願番号　特願2012-516979 (P2012-516979)
(86)(22)出願日　平成21年10月9日 (2009.10.9)
(65)公表番号　特表2012-531653 (P2012-531653A)
(43)公表日　平成24年12月10日 (2012.12.10)
(86)国際出願番号　PCT/JP2009/005297
(87)国際公開番号　WO2011/042940
(87)国際公開日　平成23年4月14日 (2011.4.14)
審査請求日　平成23年12月27日 (2011.12.27)

(73)特許権者　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　110000279
特許業務法人ウィルフォート国際特許事務所
(72)発明者　川口　裕太郎
神奈川県小田原市中里３２２番２号　株式会社日立製作所　ＲＡＩＤシステム事業部内
(72)発明者　石川　篤
神奈川県小田原市中里３２２番２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

最終頁に続く

(54)【発明の名称】記憶制御装置及び仮想ボリュームの制御方法

(57)【特許請求の範囲】
【請求項１】
仮想的に形成される仮想ボリュームと、
一つまたは複数の記憶装置を含む、複数のＲＡＩＤグループと、
前記各ＲＡＩＤグループにそれぞれストライプ状に設けられる第１実記憶領域であって、それぞれ複数の第２実記憶領域を有する複数の第１実記憶領域を管理するためのプール部と、
上位装置から前記仮想ボリュームに関するライトコマンドが発行された場合、前記各第１実記憶領域のうち所定の第１実記憶領域に含まれる前記各第２実記憶領域の中から所定の第２の実記憶領域を選択し、その所定の第２実記憶領域を前記ライトコマンドに対応する前記仮想ボリューム内の所定領域に対応付ける制御部であって、一つの前記第１実記憶領域に一つの前記仮想ボリュームを対応付ける制御部と、
前記仮想ボリュームに対応付けられている前記各第２実記憶領域の中から移動対象の第２実記憶領域を予め選択し、その移動対象の第２実記憶領域に記憶されているデータの移動先となる移動先第１実記憶領域を、前記各第１実記憶領域のうち前記移動対象の第２実記憶領域が設けられているＲＡＩＤグループ以外の他のＲＡＩＤグループ内の各第１実記憶領域の中から予め選択する移動先決定部と、
予め選択した前記移動対象の第２実記憶領域および前記移動先第１実記憶領域を対応付けて記憶する移動先記憶部と、
前記上位装置が前記移動対象として予め選択した第２実記憶領域に対応するコマンドを

発行した場合は、前記コマンドの処理の中で、前記移動対象の第２実記憶領域に記憶されているデータを、前記移動先記憶部により記憶されている前記移動先第１実記憶領域内に移動させるコマンド処理部と、
を備える記憶制御装置。

【請求項２】
（１）前記移動先記憶部は、
（１－１）前記移動対象の第２実記憶領域を特定する移動対象情報と、
（１－２）前記移動先第１実記憶領域を特定する移動先情報と、
（１－３）前記各ＲＡＩＤグループの負荷に関する負荷情報と、
（１－４）前記各ＲＡＩＤグループの使用容量に関する使用容量情報と、
を記憶しており、
（２）移動先決定部は、
（２－１）所定時刻が到来した場合またはユーザから指示された場合のいずれかの場合に、前記移動先記憶部に記憶されている前記移動先情報を消去させ、さらに、
（２－２）前記各第２実記憶領域のうちヌルデータのみが記憶されている第２実記憶領域を解放して、未使用の第２実記憶領域に変更させ、さらに、
（２－３）前記負荷情報に基づく負荷分散処理であって、相対的に高負荷の第２実記憶領域を相対的に低負荷の第１実記憶領域内に移動させるための、新たな移動先情報を作成して、前記移動先記憶部に記憶させる負荷分散処理と、
（２－４）前記使用容量情報に基づく使用容量平均化処理であって、相対的に使用容量の大きい第１実記憶領域内の第２実記憶領域を相対的に使用容量の小さい第１実記憶領域内に移動させるための、新たな移動先情報を作成して、前記移動先記憶部に記憶させる使用容量平均化処理とを、
それぞれ実行し、
（３）前記コマンド処理部は、
（３－１）前記上位装置から前記移動対象の第２実記憶領域についてのライトコマンドが発行された場合、前記移動対象の第２実記憶領域に記憶されているデータを読み出し、その読み出されたデータと前記ライトコマンドに係るライトデータとをマージして、前記移動先第１実記憶領域内の第２実記憶領域に書き込み、さらに、前記上位装置に前記ライトコマンドの処理が完了した旨を通知し、
（３－２）前記上位装置から前記移動対象の第２実記憶領域についてのリードコマンドが発行された場合、前記移動対象の第２実記憶領域からデータを読み出して前記上位装置に送信し、前記上位装置に前記リードコマンドの処理が完了した旨を通知した後で、前記移動対象の第２実記憶領域から読み出されたデータを前記移動先第１実記憶領域内の第２実記憶領域に書き込む、
請求項１に記載の記憶制御装置。

【請求項３】
前記コマンド処理部は、前記上位装置から前記移動対象の第２実記憶領域についてのライトコマンドが発行された場合、前記移動対象の第２実記憶領域に記憶されているデータを読み出し、その読み出されたデータと前記ライトコマンドに係るライトデータとをマージして、前記移動先第１実記憶領域内の第２実記憶領域に書き込む、
請求項１に記載の記憶制御装置。

【請求項４】
前記コマンド処理部は、前記上位装置から前記移動対象の第２実記憶領域についてのリードコマンドが発行された場合、前記移動対象の第２実記憶領域からデータを読み出して前記上位装置に送信し、前記上位装置に前記リードコマンドの処理が完了した旨を通知した後で、前記移動対象の第２実記憶領域から読み出されたデータを前記移動先第１実記憶

領域内の第２実記憶領域に書き込む、
請求項３に記載の記憶制御装置。

【請求項５】
前記移動先記憶部は、
前記移動対象の第２実記憶領域を特定する移動対象情報と、
前記移動先第１実記憶領域を特定する移動先情報と、
前記各ＲＡＩＤグループの負荷に関する負荷情報と、
を記憶しており、
移動先決定部は、
前記移動先記憶部に記憶されている前記移動先情報を消去させ、
相対的に高負荷の第２実記憶領域を相対的に低負荷の第１実記憶領域内に移動させるための、新たな移動先情報を前記負荷情報に基づいて作成し、前記移動先記憶部に記憶させる、
請求項３に記載の記憶制御装置。

【請求項６】
前記移動先記憶部は、
前記移動対象の第２実記憶領域を特定する移動対象情報と、
前記移動先第１実記憶領域を特定する移動先情報と、
前記各ＲＡＩＤグループの負荷に関する負荷情報と、
前記各ＲＡＩＤグループの使用容量に関する使用容量情報と、
を記憶しており、
移動先決定部は、
前記移動先記憶部に記憶されている前記移動先情報を消去させ、
相対的に高負荷の第２実記憶領域を相対的に低負荷の第１実記憶領域内に移動させるための、新たな移動先情報を前記負荷情報に基づいて作成し、前記移動先記憶部に記憶させ、さらに、
相対的に使用容量の大きい第１実記憶領域内の第２実記憶領域を相対的に使用容量の小さい第１実記憶領域内に移動させるための、新たな移動先情報を前記使用容量情報に基づいて作成し、前記移動先記憶部に記憶させる、
請求項３に記載の記憶制御装置。

【請求項７】
前記移動先決定部は、前記移動先記憶部に記憶されている前記移動先情報を消去させた後で、前記各第２実記憶領域のうちヌルデータのみが記憶されている第２実記憶領域を解放して、未使用の第２実記憶領域に変更させる、
請求項５に記載の記憶制御装置。

【請求項８】
前記移動先決定部は、前記移動対象の第２実記憶領域を前記第１実記憶領域の単位で複数選択する、請求項１に記載の記憶制御装置。

【請求項９】
前記制御部は、
前記仮想ボリュームを生成する場合に、前記仮想ボリューム内の各仮想的記憶領域を、初期データが記憶されている初期化用の第２実記憶領域に対応付け、
前記上位装置から前記仮想ボリュームに関する前記ライトコマンドが発行された場合、前記ライトコマンドに対応する前記仮想的記憶領域の対応付け先を、前記初期化用の第２実記憶領域から、選択された前記所定の第２実記憶領域に切り替える、

請求項１に記載の記憶制御装置。

【請求項１０】
前記制御部は、前回のライト要求に対応して前記仮想ボリュームに対応付けられた前記第２実記憶領域に連続する、未使用の前記第２実記憶領域を、前記所定の第２実記憶領域として前記仮想ボリュームに対応付ける、
請求項９に記載の記憶制御装置。

【請求項１１】
前記仮想ボリュームに記憶されるデータには、所定サイズ毎に保証コードが設定されており、前記保証コードは、前記ＲＡＩＤグループを識別するためのデータと、前記第１実記憶領域を識別するためのデータと、前記第１実記憶領域内における前記第２実記憶領域を識別するためのデータとを含んでいる、
請求項１０に記載の記憶制御装置。

【請求項１２】
ライトコマンドに応じて実記憶領域が割り当てられる仮想ボリュームを制御するための方法であって、
前記仮想ボリュームは、複数の仮想的記憶領域を備えており、
複数のＲＡＩＤグループを管理するためのプール部を作成し、
前記各ＲＡＩＤグループは、複数の記憶装置を跨るようにしてストライプ状に形成される第１実記憶領域であって、前記仮想的記憶領域に対応する第２実記憶領域をそれぞれ複数ずつ有する第１実記憶領域を複数備えており、
前記上位装置から前記仮想ボリュームに関するライトコマンド要求が発行された場合、一つの前記第１実記憶領域に複数の前記仮想ボリュームが対応付けられないようにして、前記ライトコマンドに対応する前記仮想的記憶領域に、前記各第１実記憶領域のうち所定の第１実記憶領域に含まれる所定の第２実記憶領域を対応付け、
前記仮想的記憶領域に対応付けられる前記所定の第２実記憶領域に、前記上位装置から受領したライトデータを記憶させ、
前記仮想的記憶領域に対応付けられている前記各第２実記憶領域の中から、相対的に高負荷の第２実記憶領域を移動対象の第２実記憶領域として予め選択し、
前記各第１実記憶領域のうち前記移動対象の第２実記憶領域が設けられているＲＡＩＤグループ以外の他のＲＡＩＤグループ内の各第１実記憶領域であって、相対的に低負荷の第１実記憶領域を、前記移動対象の第２実記憶領域に記憶されているデータの移動先となる移動先第１実記憶領域として予め選択し、
予め選択した前記移動対象の第２実記憶領域および前記移動先第１実記憶領域を対応付けて移動先記憶部に記憶させ、
前記上位装置が前記移動対象として予め選択した第２実記憶領域に対応するライトコマンドを発行した場合は、前記移動対象の第２実記憶領域に記憶されているデータを読み出し、その読み出されたデータと前記ライトコマンドに係るライトデータとをマージして、前記移動先第１実記憶領域内の第２実記憶領域に書き込む、
仮想ボリュームの制御方法。

【請求項１３】
前記上位装置から前記移動対象の第２実記憶領域についてのリードコマンドが発行された場合には、前記移動対象の第２実記憶領域からデータを読み出して前記上位装置に送信し、前記上位装置に前記リードコマンドの処理が完了した旨を通知した後で、前記移動対象の第２実記憶領域から読み出されたデータを前記移動先第１実記憶領域内の第２実記憶領域に書き込む、
請求項１２に記載の仮想ボリュームの制御方法。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、記憶制御装置及び仮想ボリュームの制御方法に関する。
【背景技術】
【０００２】
企業等のユーザは、記憶制御装置を用いてデータを管理する。記憶制御装置は、ＲＡＩＤ（Redundant Array of Independent Disks）に基づく記憶領域上に、論理ボリュームを形成する。その論理ボリュームは、ホストコンピュータ（以下、ホスト）に提供される。
【０００３】
ユーザの使用するデータ量は日々増大するため、現状に合わせて設定されたボリュームサイズでは、いずれ容量が不足する。これに対し、データ量の増加を見越して、ボリュームサイズを現在必要なサイズよりも過大に設定すると、不要不急のディスクドライブが多くなり、コストが増加する。
【０００４】
そこで、仮想的な論理ボリュームを用意し、実際の使用に応じて、仮想的な論理ボリュームに実記憶領域を割り当てる技術が提案されている（特許文献１、特許文献２）。なお、仮想的な論理ボリュームに割り当てられた実記憶領域が特定の実ボリュームに偏在するのを防止するために、データを再配置させる技術も知られている（特許文献３）。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】米国特許第６８２３４４２号明細書
【特許文献２】特開２００７－３１０８６１号公報
【特許文献３】特開２００８－２３４１５８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
前記文献（ＵＳ６，８２３，４４２Ｂ１）には、ストレージサーバシステムが、仮想ボリューム上のブロックアドレスに関するライト要求を受信した場合に、そのブロックアドレスに対応する仮想ボリュームページアドレスに対して、論理的なデータページを割り当てる。そして、その論理的なデータページにデータが書き込まれる。
【０００７】
前記文献には、複数の物理ディスク上のエリアから得られるチャンクレット（chunklet）という概念のエリアに基づいて、特定のＲＡＩＤレベルを有する論理ディスクを構成するための管理方法が記載されている。
【０００８】
しかしながら、その管理方法は、物理ディスクドライブ単位にＲＡＩＤグループを構成する、記憶制御装置の物理エリア管理方法と全く異なる。従って、前記文献に記載の管理方法を、物理ディスクドライブ単位にＲＡＩＤグループを構成する記憶制御装置に、そのまま適用することはできない。
【０００９】
仮に、前記文献に記載の技術を前記管理方法に適用する場合は、通常の論理ボリュームと仮想的な論理ボリュームとの両方を提供可能な記憶制御装置において、通常の論理ボリュームと仮想的な論理ボリュームとで、それぞれ物理エリアの管理方法が異なってしまい、記憶制御装置の構造が複雑化するという問題を生じる。ここで、通常の論理ボリュームとは、ボリューム生成時に、そのボリュームサイズと同容量の物理エリア（物理的記憶領域）が予め割り当てられる論理ボリュームを意味する。仮想的な論理ボリュームとは、ボリュームサイズが仮想化された論理ボリュームであって、ライト要求に応じて、物理エリアが割り当てられる論理ボリュームを意味する。

【００１０】
つまり、前記文献に記載の技術を、物理ディスクドライブ単位でＲＡＩＤグループを構成する記憶制御装置にもしも適用したとすると、異なる複数の管理方法で物理エリアを管理しなければならず、構成が複雑化し、開発コストも増大する。
【００１１】
さらに、前記文献では、ライト要求の受領時に、仮想ボリュームページアドレスに対応するテーブルページが割り当てられていない場合、ストレージサーバシステムは、まず最初にテーブルページを割当て、その次に、論理的なデータページを割り当てる。従って、前記文献に記載の技術では、テーブルページを割り当てた後で、データページを割り当てる必要があり、ライト処理の性能が低下するという問題がある。
【００１２】
さらに、前記文献では、上述のような割当て処理を行うため、データページ専用のプールとテーブルページ専用のプールとをそれぞれ別々に設ける必要があり、システム構造が複雑化する。
【００１３】
そこで、本発明の目的は、ＲＡＩＤグループの物理的構成を考慮して仮想ボリュームに効率的に記憶領域を対応付けることのできる記憶制御装置及び仮想ボリュームの制御方法を提供することにある。本発明の他の目的は、各ＲＡＩＤグループを均等に使用して仮想ボリュームを構成することのできる記憶制御装置及び仮想ボリュームの制御方法を提供することにある。本発明のさらに別の目的は、仮想ボリュームに効率的に記憶領域を対応付けることができ、かつ、仮想ボリュームの応答性能を向上できるようにした記憶制御装置及び仮想ボリュームの制御方法を提供することにある。本発明の更なる目的は、後述する実施形態の記載から明らかになるであろう。
【課題を解決するための手段】
【００１４】
上記課題を解決すべく、本発明の第１観点に従う記憶制御装置は、仮想的に形成される仮想ボリュームと、一つまたは複数の記憶装置を含む、複数のＲＡＩＤグループと、各ＲＡＩＤグループにそれぞれストライプ状に設けられる第１実記憶領域であって、それぞれ複数の第２実記憶領域を有する複数の第１実記憶領域を管理するためのプール部と、上位装置から仮想ボリュームに関するライトコマンドが発行された場合、各第１実記憶領域のうち所定の第１実記憶領域に含まれる各第２実記憶領域の中から所定の第２の実記憶領域を選択し、その所定の第２実記憶領域をライトコマンドに対応する仮想ボリューム内の所定領域に対応付ける制御部であって、一つの第１実記憶領域に一つの仮想ボリュームを対応付ける制御部と、仮想ボリュームに対応付けられている各第２実記憶領域の中から移動対象の第２実記憶領域を選択し、その移動対象の第２実記憶領域に記憶されているデータの移動先となる移動先第１実記憶領域を、各第１実記憶領域のうち移動対象の第２実記憶領域が設けられているＲＡＩＤグループ以外の他のＲＡＩＤグループ内の他の各第１実記憶領域の中から選択する移動先決定部と、移動対象の第２実記憶領域と、移動先第１実記憶領域とを対応付けて記憶する移動先記憶部と、上位装置が移動対象の第２実記憶領域に対応するコマンドを発行した場合は、コマンドの処理の中で、移動対象の第２実記憶領域に記憶されているデータを、移動先記憶部により記憶されている移動先第１実記憶領域内に移動させるコマンド処理部と、を備える。
【００１５】
第２観点では、第１観点において、移動先記憶部は、移動対象の第２実記憶領域を特定する移動対象情報と、移動先第１実記憶領域を特定する移動先情報と、各ＲＡＩＤグループの負荷に関する負荷情報と、各ＲＡＩＤグループの使用容量に関する使用容量情報と、を記憶しており、移動先決定部は、所定時刻が到来した場合またはユーザから指示された場合のいずれかの場合に、移動先記憶部に記憶されている移動先情報を消去させ、さらに、各第２実記憶領域のうちヌルデータのみが記憶されている第２実記憶領域を解放して、未使用の第２実記憶領域に変更させ、さらに、負荷情報に基づく負荷分散処理であって、

相対的に高負荷の第２実記憶領域を相対的に低負荷の第１実記憶領域内に移動させるための、新たな移動先情報を作成して、移動先記憶部に記憶させる負荷分散処理と、使用容量情報に基づく使用容量平均化処理であって、相対的に使用容量の大きい第１実記憶領域内の第２実記憶領域を相対的に使用容量の小さい第１実記憶領域内に移動させるための、新たな移動先情報を作成して、移動先記憶部に記憶させる使用容量平均化処理とを、それぞれ実行し、コマンド処理部は、上位装置から移動対象の第２実記憶領域についてのライトコマンドが発行された場合、移動対象の第２実記憶領域に記憶されているデータを読み出し、その読み出されたデータとライトコマンドに係るライトデータとをマージして、移動先第１実記憶領域内の第２実記憶領域に書き込み、さらに、上位装置にライトコマンドの処理が完了した旨を通知し、上位装置から移動対象の第２実記憶領域についてのリードコマンドが発行された場合、移動対象の第２実記憶領域からデータを読み出して上位装置に送信し、上位装置にリードコマンドの処理が完了した旨を通知した後で、移動対象の第２実記憶領域から読み出されたデータを移動先第１実記憶領域内の第２実記憶領域に書き込むようになっている。

【００１６】

第３観点では、第１観点において、コマンド処理部は、上位装置から移動対象の第２実記憶領域についてのライトコマンドが発行された場合、移動対象の第２実記憶領域に記憶されているデータを読み出し、その読み出されたデータとライトコマンドに係るライトデータとをマージして、移動先第１実記憶領域内の第２実記憶領域に書き込む。

【００１７】

第４観点では、コマンド処理部は、上位装置から移動対象の第２実記憶領域についてのリードコマンドが発行された場合、移動対象の第２実記憶領域からデータを読み出して上位装置に送信し、上位装置にリードコマンドの処理が完了した旨を通知した後で、移動対象の第２実記憶領域から読み出されたデータを移動先第１実記憶領域内の第２実記憶領域に書き込む。

【００１８】

第５観点では、第３観点において、移動先記憶部は、移動対象の第２実記憶領域を特定する移動対象情報と、移動先第１実記憶領域を特定する移動先情報と、各ＲＡＩＤグループの負荷に関する負荷情報と、を記憶しており、移動先決定部は、移動先記憶部に記憶されている移動先情報を消去させ、相対的に高負荷の第２実記憶領域を相対的に低負荷の第１実記憶領域内に移動させるための、新たな移動先情報を負荷情報に基づいて作成し、移動先記憶部に記憶させる。

【００１９】

第６観点では、第３観点において、移動先記憶部は、移動対象の第２実記憶領域を特定する移動対象情報と、移動先第１実記憶領域を特定する移動先情報と、各ＲＡＩＤグループの負荷に関する負荷情報と、各ＲＡＩＤグループの使用容量に関する使用容量情報と、を記憶しており、移動先決定部は、移動先記憶部に記憶されている移動先情報を消去させ、相対的に高負荷の第２実記憶領域を相対的に低負荷の第１実記憶領域内に移動させるための、新たな移動先情報を負荷情報に基づいて作成し、移動先記憶部に記憶させ、さらに、相対的に使用容量の大きい第１実記憶領域内の第２実記憶領域を相対的に使用容量の小さい第１実記憶領域内に移動させるための、新たな移動先情報を使用容量情報に基づいて作成し、移動先記憶部に記憶させる。

【００２０】

第７観点では、第５観点において、移動先決定部は、移動先記憶部に記憶されている移動先情報を消去させた後で、各第２実記憶領域のうちヌルデータのみが記憶されている第２実記憶領域を解放して、未使用の第２実記憶領域に変更させる。

【００２１】

第８観点では、移動先決定部は、移動対象の第２実記憶領域を第１実記憶領域の単位で複数選択する。

【００２２】

第９観点では、制御部は、仮想ボリュームを生成する場合に、仮想ボリューム内の各仮想的記憶領域を、初期データが記憶されている初期化用の第２実記憶領域に対応付け、上位装置から仮想ボリュームに関するライトコマンドが発行された場合、ライトコマンドに対応する仮想的記憶領域の対応付け先を、初期化用の第２実記憶領域から、選択された所定の第２実記憶領域に切り替える。

【００２３】

第１０観点では、制御部は、前回のライト要求に対応して仮想ボリュームに対応付けられた第２実記憶領域に連続する、未使用の第２実記憶領域を、所定の第２実記憶領域として仮想ボリュームに対応付ける。

【００２４】

第１１観点では、仮想ボリュームに記憶されるデータには、所定サイズ毎に保証コードが設定されており、保証コードは、ＲＡＩＤグループを識別するためのデータと、第１実記憶領域を識別するためのデータと、第１実記憶領域内における第２実記憶領域を識別するためのデータとを含んでいる。

【００２５】

第１２観点でに従う仮想ボリュームの制御方法は、ライトコマンドに応じて実記憶領域が割り当てられる仮想ボリュームを制御するための方法であって、仮想ボリュームは、複数の仮想的記憶領域を備えており、複数のＲＡＩＤグループを管理するためのプール部を作成し、各ＲＡＩＤグループは、複数の記憶装置を跨るようにしてストライプ状に形成される第１実記憶領域であって、仮想的記憶領域に対応する第２実記憶領域をそれぞれ複数ずつ有する第１実記憶領域を複数備えており、上位装置から仮想ボリュームに関するライトコマンド要求が発行された場合、一つの第１実記憶領域に複数の仮想ボリュームが対応付けられないようにして、ライトコマンドに対応する仮想的記憶領域に、各第１実記憶領域のうち所定の第１実記憶領域に含まれる所定の第２実記憶領域を対応付け、仮想的記憶領域に対応付けられる所定の第２実記憶領域に、上位装置から受領したライトデータを記憶させ、仮想的記憶領域に対応付けられている各第２実記憶領域の中から、相対的に高負荷の第２実記憶領域を移動対象の第２実記憶領域として選択し、各第１実記憶領域のうち移動対象の第２実記憶領域が設けられているＲＡＩＤグループ以外の他のＲＡＩＤグループ内の他の各第１実記憶領域であって、相対的に低負荷の第１実記憶領域を、移動対象の第２実記憶領域に記憶されているデータの移動先となる移動先第１実記憶領域として選択し、移動対象の第２実記憶領域と、移動先第１実記憶領域とを対応付けて移動先記憶部に記憶させ、上位装置が移動対象の第２実記憶領域に対応するライトコマンドを発行した場合は、移動対象の第２実記憶領域に記憶されているデータを読み出し、その読み出されたデータとライトコマンドに係るライトデータとをマージして、移動先第１実記憶領域内の第２実記憶領域に書き込む。

【００２６】

第１３観点では、第１２観点において、上位装置から移動対象の第２実記憶領域についてのリードコマンドが発行された場合には、移動対象の第２実記憶領域からデータを読み出して上位装置に送信し、上位装置にリードコマンドの処理が完了した旨を通知した後で、移動対象の第２実記憶領域から読み出されたデータを移動先第１実記憶領域内の第２実記憶領域に書き込む。

【００２７】

第１４観点に従う記憶制御装置は、仮想的に形成される仮想ボリュームと、一つまたは複数の記憶装置を含む、複数のＲＡＩＤグループと、各ＲＡＩＤグループにそれぞれストライプ状に設けられる第１実記憶領域であって、それぞれ複数の第２実記憶領域を有する複数の第１実記憶領域を管理するためのプール部と、上位装置から仮想ボリュームに関するライトコマンドが発行された場合、各第１実記憶領域のうち所定の第１実記憶領域に含まれる各第２実記憶領域の中から所定の第２の実記憶領域を選択し、その所定の第２実記憶領域をライトコマンドに対応する仮想ボリューム内の所定領域に対応付ける制御部であって、一つの第１実記憶領域に一つの仮想ボリュームを対応付ける制御部と、仮想ボリュ

ームに対応付けられている各第１実記憶領域の中から移動対象の第１実記憶領域を選択し、その移動対象の第１実記憶領域に記憶されているデータの移動先となる移動先ＲＡＩＤグループを、移動対象の第１実記憶領域が設けられているＲＡＩＤグループ以外の他のＲＡＩＤグループの中から選択する移動先決定部と、移動対象の第１実記憶領域と、移動先ＲＡＩＤグループとを対応付けて記憶する移動先記憶部と、上位装置が移動対象の第１実記憶領域に対応するコマンドを発行した場合は、コマンドの処理の中で、移動対象の第１実記憶領域に記憶されているデータを、移動先記憶部により記憶されている移動先ＲＡＩＤグループ内に移動させるコマンド処理部と、を備える。

【００２８】

本発明の構成の少なくとも一部は、コンピュータプログラムとして構成できる。このコンピュータプログラムは、記録媒体に固定して配布したり、通信ネットワークを介して配信することができる。さらに、前記観点の組合せ以外の他の組合せも本発明の範囲に含まれる。

【図面の簡単な説明】

【００２９】

【図１】図１は、本発明の実施形態の全体概念を示す説明図である。

【図２】図２は、記憶制御装置を含むシステムの全体構成を示す説明図である。

【図３】図３は、記憶制御装置のブロック図である。

【図４】図４は、プログラム及びテーブルを示す図である。

【図５】図５は、仮想ボリュームとチャンク及びページの関係を示す図である。

【図６】図６は、チャンク及びページを管理するテーブル群を示す図である。

【図７】図７は、仮想ボリュームの生成時におけるテーブル群の接続の様子を示す説明図である。

【図８】図８は、ライトデータを書き込む場合に、仮想ボリュームに割り当てるページを初期設定用ページから所定のページに切り替える様子を示す図である。

【図９】図９は、移動先管理テーブルを示す図である。

【図１０】図１０は、記憶制御装置の全体動作の流れを示す図である。

【図１１】図１１は、プール作成処理を示すフローチャートである。

【図１２】図１２は、データに付加される保証コードの構成を示す図である。

【図１３】図１３は、チャンクの状態遷移を示す図である。

【図１４】図１４は、チャンクを管理するためのキューを示す図である。

【図１５】図１５は、フォーマット処理を示すフローチャートである。

【図１６】図１６は、フォーマット処理の一部を示すフローチャートである。

【図１７】図１７は、仮想ボリューム生成処理を示すフローチャートである。

【図１８】図１８は、ライト処理を示すフローチャートである。

【図１９】図１９は、図１８に続くフローチャートである。

【図２０】図２０は、図１９に続くフローチャートである。

【図２１】図２１は、チャンクを仮想ボリュームに割り当てる処理を示すフローチャートである。

【図２２】図２２は、ページを仮想ボリュームに割り当てる処理を示すフローチャートである。

【図２３】図２３は、ページ状態を変更する処理を示すフローチャートである。

【図２４】図２４は、リード処理を示すフローチャートである。

【図２５】図２５は、図２４に続くフローチャートである。

【図２６】図２６は、図２５に続くフローチャートである。

【図２７】図２７は、移動先を決定する処理のフローチャートである。

【図２８】図２８は、負荷分散処理を示すフローチャートである。

【図２９】図２９は、使用容量を平均化させる処理のフローチャートである。

【図３０】図３０は、本発明の効果の一つを示す図である。

【図３１】図３１は、第２実施例に係り、移動先を決定する処理を示すフローチャートで

ある。
【図３２】図３２は、０データのみを記憶するページを解放する処理を示すフローチャートである。
【発明を実施するための形態】
【００３０】
以下、図面に基づいて、本発明の実施の形態を説明する。最初に、本発明の概要を説明し、次に、実施例について説明する。本発明は、後述のように、仮想ボリューム５への実記憶領域の割当てを、チャンク７単位で行う。チャンク７は、複数のページ８から構成される。一つのチャンク７には、一つの仮想ボリューム５が対応付けられる。つまり、一つのチャンク７が異なる複数の仮想ボリューム５に対応付けられることはない。このため、チャンク７の記憶領域を効率的に使用することができる。
【００３１】
仮想ボリューム５の生成時に、各仮想的記憶領域５Ａと初期設定用のページ８とが予め対応付けられる。ホスト２から仮想ボリューム５へのライト要求が発行されると、チャンク７内のページ８が順番に使用されて、ライト要求に関わる仮想的記憶領域５Ａに割り当てられる。その割り当てられたページ８にライトデータが書き込まれる。データの書込み時には、ライト要求に係る仮想的記憶領域５Ａの接続先が、初期設定用のページ８から、チャンク７内の所定のページ８に切り替えられる。所定のページ８とは、前回のライト処理時に使用されたページに連続するページである。つまり、ライトデータを書き込む際には、仮想的記憶領域５Ａに割り当てられるページを、初期設定用のページ８から所定のページ８に切り替えるだけで済むため、仮想ボリューム５の応答性能を向上できる。
【００３２】
さらに、チャンク７内に空きページ８が無くなると、新たなチャンク７が選択されて仮想ボリューム５に割り当てられる。新たなチャンク７は、別のＲＡＩＤグループ６ｂ内のチャンク群から選択される。これにより、各ＲＡＩＤグループ６ａ，６ｂ間で負荷を分散させることができる。
【００３３】
さらに、本発明では、負荷等に基づいて移動対象のページ８またはチャンク７を選択して、その移動先を決定する。そして、移動対象のページ８またはチャンク７にホスト２がアクセスした場合、ホスト２から発行されるコマンド（ライトコマンドまたはリードコマンド）を処理しながら、データを移動させる。
【００３４】
図１は、本発明の実施形態の概要を示す説明図である。図１に関する以下の記載は、本発明の理解及び実施に必要な程度で本発明の概要を示しており、本発明の範囲は図１に示す構成に限定されない。
【００３５】
図１に示すシステムは、例えば、記憶制御装置１と、「上位装置」としてのホスト２とを備える。ホスト２は、例えば、サーバコンピュータまたはメインフレームコンピュータのようなコンピュータ装置として構成される。ホスト２がホストコンピュータの場合、例えば、ＦＩＣＯＮ（Fibre Connection：登録商標）、ＥＳＣＯＮ（Enterprise System Connection：登録商標）、ＡＣＯＮＡＲＣ（Advanced Connection Architecture：登録商標）、ＦＩＢＡＲＣ（Fibre Connection
Architecture：登録商標）等の通信プロトコルに従って、データ通信が行われる。ホスト２がサーバコンピュータ等の場合、例えば、ＦＣＰ（Fibre Channel Protocol）またはｉＳＣＳＩ（internet Small Computer System Interface）等の通信プロトコルに従って、データ通信が行われる。
【００３６】
記憶制御装置１は、ホスト２に通信ネットワークを介して接続される。記憶制御装置１は、例えば、コントローラ３と、記憶装置４と、仮想ボリューム５（１），５（２）とを備える。特に区別する必要がない場合、仮想ボリューム５と呼ぶ。各ＲＡＩＤグループ６

ａ，６ｂは、それぞれ複数ずつの記憶装置４から構成される。特に区別しない場合、ＲＡＩＤグループ６と呼ぶ。

【００３７】

記憶装置４としては、例えば、ハードディスクデバイス、半導体メモリデバイス、光ディスクデバイス、光磁気ディスクデバイス、磁気テープデバイス、フレキシブルディスクデバイス等のデータを読み書き可能な種々のデバイスを利用可能である。

【００３８】

記憶装置４としてハードディスクデバイスを用いる場合、例えば、ＦＣ（Fibre Channel）ディスク、ＳＣＳＩ（Small Computer System Interface）ディスク、ＳＡＴＡディスク、ＡＴＡ（AT Attachment）ディスク、ＳＡＳ（Serial Attached SCSI）ディスク等を用いることができる。また、例えば、フラッシュメモリ、ＦｅＲＡＭ（Ferroelectric Random Access Memory）、ＭＲＡＭ（Magnetoresistive Random Access Memory）、相変化メモリ（Ovonic Unified Memory）、ＲＲＡＭ（Resistance RAM）」等の種々の記憶装置４を用いることもできる。さらに、例えば、フラッシュメモリデバイスとハードディスクドライブのように、種類の異なる記憶装置４を混在させる構成でもよい。

【００３９】

各ＲＡＩＤグループ６ａ，６ｂの有する物理的記憶領域は、ストライプ状の複数のチャンク７に区切られる。各チャンク７は、連続する複数のページ８から構成される。チャンク７は「第１実記憶領域」に該当し、ページ８は「第２実記憶領域」に該当する。理解のために、一方のＲＡＩＤグループ６ａに属する第１チャンク７に符号”ａ１”を与え、第１チャンク７（ａ１）に属する各ページに連番を添える。他方のＲＡＩＤグループ６ｂについても同様である。従って、例えば、”ａ２－３”は、ＲＡＩＤグループ６ａの第２チャンク内の３番目のページであることを意味し、”ｂ１－１”は、ＲＡＩＤグループ６ｂの第１チャンク内の１番目のページであることを意味する。

【００４０】

仮想ボリューム５は、複数の仮想的記憶領域５Ａから構成される。仮想的記憶領域５Ａとページ８のサイズは同一である。一つの例では、１枚のページ８のサイズはＳＺ１バイト（例えば、３２ＭＢ）、一つのチャンク７のサイズはＳＺ２バイト（例えば、１ＧＢ）、仮想ボリューム５のサイズはＳＺ３バイト（例えば、１０ＧＢ）である。この場合、一つの仮想ボリューム５はＮ１個（例えば、１０個）のチャンク７から構成され、一つのチャンク７はＮ２枚（例えば、３２枚）のページ８から構成される。上記括弧内の数値は、理解のための一例に過ぎず、本発明の範囲は上記数値に限定されない。上述のページサイズ、チャンクサイズ、仮想ボリュームサイズ等は、可変に設定できる。

【００４１】

コントローラ３は、記憶制御装置１の動作を制御する。例えば、コントローラ３は、ユーザからの指示に基づいて、ＲＡＩＤグループ６ａ，６ｂ及び仮想ボリューム５を生成させる。また、コントローラ３は、ホスト２から発行されるコマンド（リードコマンド、ライトコマンド）に応じて処理を実行し、その処理結果をホスト２に送信する。

【００４２】

さらに、コントローラ３は、ライトコマンドを受領した場合、ライトコマンドにより指定される仮想的記憶領域５Ａに、ページ８が割り当てられているか否かを判断する。指定される仮想的記憶領域５Ａにページ８が割り当てられていない場合、コントローラ３は、チャンク７内の所定のページ８を、指定される仮想的記憶領域５Ａに割り当てる。コントローラ３は、割り当てられた所定ページ８にライトデータを書き込む。

【００４３】

コントローラ３は、一つのチャンク７に一つの仮想ボリューム５だけが対応付けられるように、ページ割当てを制御する。一つのチャンク７に含まれる各ページ８は、一つの仮想ボリューム５にのみ割り当てられる。一つのチャンク７内にそれぞれ異なる複数の仮想ボリューム５に割り当てられるページ８が混在することはない。一つのチャンク７内では、論理アドレスの値を問わずに、連続するページ８が使用される。

【００４４】
一方の仮想ボリューム５（１）を例に挙げて説明すると、最初のライトコマンドについて、チャンク７（ａ１）内の先頭ページ８（ａ１－１）が使用され、次のライトコマンドについては、その先頭ページ８（ａ１－１）に続く次のページ８（ａ１－２）が使用され、さらに別のライトコマンドについては、次のページ８（ａ１－３）が使用される。そして、最後のページ８（ａ１－４）を使用した後で、さらにライトコマンドを受領した場合、新たなチャンク７（ｂ１）が仮想ボリューム５（１）に割り当てられる。
【００４５】
他方の仮想ボリューム５（２）には、ＲＡＩＤグループ６ａの第２チャンク７内の先頭ページ８（ａ２－１）が割り当てられている。もしも、仮想ボリューム５（２）を対象とする新たなライトコマンドが発行された場合、次のページ８が仮想ボリューム５（２）に割り当てられる。そのページ８には、”ａ２－２”の符号が添えられるはずであるが、図１では省略する。
【００４６】
このように、仮想ボリューム５には、ホスト２に見せかけているボリュームサイズよりも小さいサイズの実記憶領域（ページ７、チャンク８）が割り当てられた状態で、ホスト２に提供される。そして、ホスト２からのライトコマンドに応じて、必要量の実記憶領域が動的に割り当てられる。
【００４７】
さらに、コントローラ３は、以下に述べるように、各ＲＡＩＤグループ６ａ，６ｂ間でデータを移動させる。コントローラ３は、例えば、データへのアクセス頻度（負荷）等に基づいて、移動対象のデータを選択する。コントローラ３は、移動対象のデータを移動させる移動先として、低負荷のＲＡＩＤグループ６を選択する。または、コントローラ３は、低負荷のＲＡＩＤグループ内のチャンクを移動先として選択することもできる。
【００４８】
上述のように、コントローラ３は、移動対象のデータ及びその移動先を決定し（Ｓ１）、記憶させる（Ｓ２）。ホスト２が移動対象のデータにアクセスすると、コントローラ３は、移動対象データを読み出してコマンド処理を行うとともに、その移動対象データを移動先の記憶領域に移動させる（Ｓ３）。
【００４９】
このように構成される本実施形態では、複数のページ８を有するチャンク７の単位で、仮想ボリューム５に実記憶領域（物理的記憶領域）を割り当て、かつ、一つのチャンク７を一つの仮想ボリューム５にのみ割り当てる。従って、後述のように、チャンク７内の記憶領域を有効に使用できる。また、通常の論理ボリュームと同様にして、仮想ボリューム５を管理でき、制御構造を簡素化できる。
【００５０】
本実施形態では、複数のＲＡＩＤグループ６ａ，６ｂを均等に使用するべく、各ＲＡＩＤグループ６ａ，６ｂからそれぞれチャンク７（ａ１），７（ｂ１）を選択して、仮想ボリューム５（１）に割り当てる。これにより、各ＲＡＩＤグループ６ａ，６ｂの負荷を均一化することができる。
【００５１】
本実施形態では、例えば、ＲＡＩＤグループ６ａ，６Ｂ間の負荷を分散させべくデータ移動計画を予め作成しておき、ホスト２が移動対象のデータにアクセスした場合に、ホスト２の要求する処理を実行しながら、移動対象のデータを予め設定された移動先に移動させる。従って、ホスト２からのアクセスを契機として、通常のコマンド処理の中でデータ移動を行うことができる。リード処理の場合は、読み出したデータを移動先に書き込む作業が発生するが、移動先は低負荷の記憶領域の中から選択されるため、記憶制御装置の応答性能に与える影響は少ない。
【００５２】
さらに、本実施形態では、格納先の移動が予定されたものの、ホスト２にアクセスされ

なかったデータは、移動契機が発生しないので、移動されることはない。従って、実記憶領域へのアクセスを低減し、ホスト２により使用されるデータのみを効率的に移動させることができる。

【００５３】

さらに、本実施形態では、ホスト２によるアクセスを移動契機とするため、移動先を決定するタイミングと、実際の移動タイミングとの間に時間遅れが生じる。しかし、ホスト２により頻繁にアクセスされるデータは、直ちに低負荷の記憶領域に移動されるため、タイミングのずれによる影響は小さい。以下、本実施形態を詳細に説明する。

【実施例１】

【００５４】

図２は、本実施例に係る記憶制御装置１０を含む情報処理システムの全体構成を示す説明図である。この情報処理システムは、例えば、少なくとも一つの記憶制御装置１０と、一つまたは複数のホスト２０と、少なくとも一つの管理サーバ７０とを含んで構成することができる。

【００５５】

先に図１で述べた実施形態との対応関係を説明する。記憶制御装置１０は記憶制御装置１に、ホスト２０はホスト２に、コントローラ３０はコントローラ３に、記憶装置４０は記憶装置４に、仮想ボリューム５０Ｖは仮想ボリューム５に、ＲＡＩＤグループ９０はＲＡＩＤグループ６ａ，６ｂに、それぞれ対応する。図１で述べた説明と重複する説明は、できるだけ省略する。

【００５６】

ホスト２０と記憶装置１０とは、第１通信ネットワーク８０を介して接続される。第１通信ネットワーク８０は、例えば、ＦＣ－ＳＡＮ（Fibre Channel-Storage Area Network）やＩＰ－ＳＡＮ（Internet Protocol_SAN）のように構成される。

【００５７】

管理サーバ７０は、記憶制御装置１０の設定を変更等するための装置である。管理サーバ７０は、例えば、ＬＡＮ（Local Area Network）のような第２通信ネットワーク８１を介して、記憶制御装置１０に接続されている。なお、ホスト２０にストレージ管理機能を設けて、ホスト２０側から記憶制御装置１０の設定変更等を行うように構成することもできる。

【００５８】

記憶制御装置１０の詳細は後述するが、記憶制御装置１０は、仮想ボリューム５０Ｖと通常ボリューム５０Ｎとを備える。なお、図中では、論理ボリュームを”ＬＵ”と表記している。ＬＵとは、Logical Unitの略である。

【００５９】

仮想ボリューム５０Ｖは、図１で述べたように、仮想的に生成される論理ボリュームであって、ホスト２０からのライトコマンドに応じて記憶領域が割り当てられるボリュームである。つまり、仮想ボリューム５０Ｖは、ホスト２０に提供されるボリュームサイズと、実際に有する記憶領域のサイズとが一致しない。通常ボリューム５０Ｎは、ＲＡＩＤグループ９０の有する記憶領域に基づいて生成されるボリュームである。

【００６０】

プール部６０は、複数のＲＡＩＤグループ９０の有する記憶領域を管理する。プール部６０で管理されている記憶領域は、チャンク９１（図５参照）単位で仮想ボリューム５０Ｖに割り当てられる。

【００６１】

コントローラ３０は、各ボリューム５０Ｖ，５０Ｎの論理アドレスをＲＡＩＤグループ９０の物理アドレスに変換等して、データを記憶装置４０に書き込んだり、あるいは、記憶装置４０から読み出したデータの物理アドレスを論理アドレスに変換等して、データをホスト２０に送信する。

【００６２】

さらに、後述のように、コントローラ３０は、データ移動計画を作成し、ホスト２０からのアクセスを契機としてデータ移動計画を実行する。

【００６３】

図３は、コントローラ３０の構成を示すブロック図である。記憶制御装置１０には、複数の増設筐体４３を接続することができる。増設筐体４３は、複数の記憶装置４０を収容している。ユーザは、必要に応じて、増設筐体４３を接続することにより、システムの総記憶容量を増大させることができる。

【００６４】

記憶制御装置１０は、複数のコントローラ３０（＃０），３０（＃１）を備える。いずれか一方のコントローラ３０が障害等によって停止した場合でも、他方のコントローラ３０により動作を継続することができる。以下、特に区別する必要が無い場合、コントローラ３０と称する。

【００６５】

コントローラ３０は、例えば、第１通信回路３１０（図中、ＦＥＩ／Ｆ）と、第２通信回路３２０（図中、ＳＡＳ）と、データ転送制御回路３３０（図中、ＤＣＴＬ）と、キャッシュメモリ３４０（図中、ＣＭ）と、ローカルメモリ３５０（図中、ＬＭ）と、マイクロプロセッサ３６０（図中、ＭＰＵ）と、メモリコントローラ３７０（図中、ＭＣ）と、エキスパンダ３８０（図中、ＥＸＰ）とを備える。

【００６６】

第１通信回路３１０は、ホスト２０と通信を行うための制御回路である。第２通信回路３２０は、各記憶装置４０と通信を行うための制御回路である。データ転送制御回路３３０は、記憶制御装置１０内のデータの流れを制御するための回路である。各データ転送制御回路３３０は、互いに接続されている。キャッシュメモリ３４０は、例えば、ホスト２０から受領したライトデータと、記憶装置４０から読み出されるデータとを記憶する。さらに、キャッシュメモリ３４０には、記憶制御装置１０の構成または動作を管理するための管理用データ等が記憶される場合もある。

【００６７】

ローカルメモリ３５０は、例えば、マイクロプロセッサ３６０により使用される各種データを記憶する。マイクロプロセッサ３６０は、記憶装置４０またはローカルメモリ３５０からコンピュータプログラムを読み込んで実行することにより、後述のように、記憶制御装置１０の動作を制御する。

【００６８】

メモリコントローラ３７０は、マイクロプロセッサ３６０をローカルメモリ３５０及びデータ転送制御回路３３０に接続させるための制御回路である。エキスパンダ３８０は、第２通信回路３２０の通信ポートを拡張するための回路である。なお、図３に示す構成は一例であって、他の構成の記憶制御装置にも本発明を適用できる。

【００６９】

図４は、記憶制御装置１０の有するプログラム及びテーブルの例を示す。例えば、メモリ３６０にはプログラム１１０～１３０が記憶され、キャッシュメモリ３４０にはテーブル１４０～１６０が記憶される。または、特定の記憶装置４０内に、プログラム及びテーブルを格納し、必要に応じてメモリにロードして使用する構成でもよい。

【００７０】

ＲＡＩＤ制御プログラム１１０は、ＲＡＩＤを制御するためのプログラムである。ＲＡＩＤ制御プログラム１１０は、ドライブ管理テーブル１４０等を用いて、ＲＡＩＤ構成を管理する。

【００７１】

コマンド処理プログラム１２０は、ホスト２０からのコマンドを処理するためのプログラムである。コマンドとしては、例えば、ライトコマンド、リードコマンド、その他のコマンドを挙げることができる。その他のコマンドとしては、例えば、ボリュームの作成、ボリュームの削除、ボリュームコピーの指示、仕様等を問い合わせるためのコマンドを挙

げることができる。コマンド処理プログラム１２０は、ライトコマンド及びリードコマンドを処理する場合、その処理対象のデータに「データの移動先」が関連づけられているか否かを確認する。もしも、処理対象データにデータの移動先が関連づけられている場合、コマンド処理プログラム１２０は、コマンド処理の中で、処理対象データを予め設定されている移動先に移動させる。

【００７２】

移動先決定プログラム１３０は、仮想ボリューム５０Ｖに記憶されているデータのうち、第１条件を満たすデータを移動対象データとして選択し、第２条件を満たす移動先に移動させる。

【００７３】

例えば、第１条件は、予め設定されている負荷の値を超えるＲＡＩＤグループ９０のうち、最も高い負荷値を有すること、である。例えば、第２条件は、予め設定される他の負荷値を下回るＲＡＩＤグループ９０のうち、最も低い負荷値を有すること、である。

【００７４】

他の第１条件は、使用容量が予め設定される使用容量の値よりも大きいＲＡＩＤグループ９０のうち最も使用容量が大きいこと、である。他の第２条件は、予め設定される他の使用容量値を下回るＲＡＩＤグループ９０のうち、最も小さい使用容量を有すること、である。使用容量は、百分率形式で表すこともできる。

【００７５】

ホストアクセスを契機としてデータを移動させると、移動元ＲＡＩＤグループ９０は、その負荷及び使用容量がともに減少する。移動先ＲＡＩＤグループ９０は、その負荷及び使用容量がともに増加する。

【００７６】

ドライブ管理テーブル１４０は、各記憶装置４０を管理するための情報である。ドライブ管理テーブル１４０は、例えば、各記憶装置４０の種類、容量、所属するＲＡＩＤグループの番号等を管理する。

【００７７】

ページ管理テーブル１５０は、各ページ９２及び各チャンク９１等を管理する。ページ管理テーブル１５０の詳細は、図６～図８で後述する。

【００７８】

移動先管理テーブル１６０は、移動対象データの移動先に関する情報と、移動先を決定するために使用される情報とを管理する。移動先管理テーブル１６０の詳細は、図９で後述する。

【００７９】

図５は、仮想ボリューム５０Ｖとチャンク９１との関係を示す説明図である。仮想ボリューム５０Ｖは、複数の仮想的記憶領域５００を有する。チャンク９１は、複数のページ９２を有する。図５では、便宜上、各ページ９２を、ストライプ状のチャンク９１を横方向に区切って形成するかのように示すが、実際には、各ページ９２は、ストライプ列に沿うようにして形成される。

【００８０】

図５では、最初に、第１のチャンク９１（＃０）が仮想ボリューム５０Ｖに割り当てられたとする。仮想ボリューム５０Ｖへのライトコマンド（ライト要求）が受領される度に、第１チャンク９１（＃０）内のページ９２が順番に選択されて、ライトコマンドに対応する仮想的記憶領域５００に対応付けられる。ライトデータは、仮想的記憶領域５００に対応付けられたページ９２に書き込まれる。つまり、そのページ９２を構成する各記憶装置４０の各記憶領域にライトデータが書き込まれる。

【００８１】

第１チャンク９１（＃０）の最初のページ９２（１－０）が使用された後、次のページ９２（１－１）が使用され、さらに次のページ９２（１－２）が使用される。そして、第１チャンク９１（＃０）の最終ページ９２（１－４）まで使用されたとする。これにより

、第１チャンク９１（＃０）の全ページ９２が使用されたことになる。
【００８２】
新たなライトコマンドがホスト２０から発行されると、第２のチャンク９１（＃１）が他のＲＡＩＤグループ９０内から選択され、第２チャンク９１（＃１）の先頭ページ９２（２－０）が使用される。以下、第１チャンク９１（＃０）で述べたと同様に、第２チャンク９１（＃１）内の各ページ９２が順番に使用される。
【００８３】
このように、仮想ボリューム５０Ｖには、複数のチャンク９１（＃０），９１（＃１）が対応付けられる。それらのチャンク９１（＃０），９１（＃１）は、それぞれ別々のＲＡＩＤグループ９０の中から選択される。つまり、仮想ボリューム５０Ｖには、複数のＲＡＩＤグループ９０の中から選択される複数のチャンク９１が対応付けられる。ホスト２０による仮想ボリューム５０Ｖへのデータ書込みに応じて、対応付けられたチャンク９１内のページ９２が順番に使用される。
【００８４】
なお、ホスト２０が仮想ボリューム５０Ｖ内のデータ消去を要求する場合、消去対象のデータの論理アドレスに対応するページ９２は開放され、未使用ページに戻る。未使用ページには、別のライトデータが書き込まれる。
【００８５】
ホスト２０が仮想ボリューム５０Ｖを使用すればするほど、多くのページ９２が仮想ボリューム５０Ｖに次第に割り当てられていく。これにより、仮想ボリューム５０Ｖが実際に有する物理的記憶領域の量と、仮想ボリューム５０Ｖがホスト２０に見せかけている容量との差は縮まっていく。
【００８６】
プール部６０内に、新たなＲＡＩＤグループ９０が追加された場合を検討する。図示は省略するが、その新たなＲＡＩＤグループの番号を＃２とする。チャンク９１（＃１）を使い切った後であれば、新たなＲＡＩＤグループ９０（＃２）からチャンク９１（＃２）が選択されるかも知れない。
【００８７】
しかし、せっかく新たなＲＡＩＤグループ９０（＃２）が追加されても、前のチャンク９１（＃１）を使い終わるまで、新たなＲＡＩＤグループ９０（＃２）を使用できないならば、それは無駄である。
【００８８】
そこで、本実施例では、各ＲＡＩＤグループ９０（＃０）～９０（＃２）の負荷または／及び使用容量を均等にするためのデータ移動計画を作成し、ホストアクセスを契機として、データを移動させる。これにより、プール部６０に新たに追加されるＲＡＩＤグループ９０（＃２）を比較的速やかに使い始めることができる。
【００８９】
図６は、仮想ボリューム５０Ｖを管理するためのページ管理テーブル１５０を示す説明図である。図６に示すように、ページ管理テーブル１５０は、複数のテーブル１５１～１５５等を含んで構成される。ページ管理テーブル１５０は、キャッシュメモリ３４０上に設けられ、その後、所定の記憶装置４０内の所定の管理領域に格納される。また、ページ管理テーブル１５０は、ローカルメモリ３５０にもコピーされて、マイクロプロセッサ３６０により使用される。
【００９０】
プールインデックス１５１は、例えば、プール部６０で管理されているＲＡＩＤグループ９０の識別番号等の、プール部６０の構成（状態及び属性を含む）に関する情報を管理するための情報である。
【００９１】
仮想ボリュームインデックス１５２は、仮想ボリューム５０Ｖの構成に関する情報を管理するための情報である。仮想ボリュームインデックス１５２は、例えば、仮想ボリュー

ム５０Ｖに割り当てられているチャンク９１の識別番号、及び、リンクされる仮想アドレスインデックス１５３の情報等を管理する。

【００９２】

仮想アドレスインデックス１５３は、仮想アドレスブロック１５４へのポインタを管理するための情報である。仮想アドレスブロック１５４は、ページアドレス情報１５５へのポインタを管理するための情報である。

【００９３】

例えば、仮想ボリューム５０Ｖのボリュームサイズを１０ＧＢとすると、仮想アドレスインデックス１５３は、仮想ボリューム５０Ｖの仮想アドレス領域を４ＧＢずつの領域に分けて管理する（最初の２つの領域は４ＧＢ、最後の一つの領域は２ＧＢである。）。仮想アドレスブロック１５４は、それぞれ４ＧＢの範囲をカバー可能である。このように、本実施例では、仮想ボリューム５０Ｖの有する仮想アドレス領域を、仮想アドレスインデックス１５３と仮想アドレスブロック１５４との２段階に分けて階層管理する。これにより、範囲を絞り込んで検索することができ、該当ページ９２に速やかにアクセスすることができる。なお、上記の数値（４ＧＢ、１０ＧＢ、２ＧＢ）は、説明のための一例に過ぎず、本発明はそれらの数値に限定されない。前記各数値は可変に設定できる。

【００９４】

ページアドレス情報１５５は、仮想ボリューム５０Ｖを構成する各仮想的記憶領域５００（つまり、仮想ページ５００）の構成情報を管理するための情報である。ページアドレス情報１５５には、例えば、仮想ページに対応付けられる物理ページ９２を示す物理アドレス及びページ状態が含まれる。

【００９５】

プール内ＲＡＩＤグループインデックス１１０は、プール部６０で管理されている各ＲＡＩＤグループ９０の構成情報を管理するための情報である。プール内ＲＡＩＤグループインデックス１１０は、例えば、プール部６０内の各ＲＡＩＤグループ９０が有する各チャンク９１の情報等を管理する。また、プール内ＲＡＩＤグループインデックス１１０は、未割当てチャンクキューの先頭及び末尾をそれぞれ示すためのポインタを含む。

【００９６】

チャンクインデックス１２０は、各仮想ボリューム５０Ｖにどこまでチャンク９１が割り当てられているかを管理するためのポインタを含む。つまり、チャンクインデックス１２０は、各仮想ボリューム５０Ｖに割り当てられているチャンク数などを管理する。

【００９７】

図７は、図６に示すページ管理テーブル１５０が初期化された状態を示す。仮想ボリューム５０Ｖの生成時に、ページ管理テーブル１５０は、図７に示すように初期化される。仮想ボリューム５０Ｖに含まれる各仮想的記憶領域５００（仮想ページ５００）は、特定のチャンク９１内の特定のページ９２にマッピングされる。

【００９８】

例えば、図７に示すように、ＲＡＩＤグループ９０内の先頭チャンク９１の最終ページ９２が、初期化用の特定ページとして使用される。先頭チャンク９１には、上述のテーブル群などの管理情報が記憶される。先頭チャンク９１内の先頭ページ９２から所定数のページ９２までは、管理情報の退避領域として使用される。設定可能な仮想ボリュームの数等によって異なるが、管理情報の合計サイズはチャンクサイズ（例えば、１ＧＢ）未満となる。従って、少なくとも、先頭チャンク９１の最終ページ９２に管理情報が格納されることはない。つまり、先頭チャンク９１の最終ページ９２は、管理情報の格納先として使用されることがない。

【００９９】

そこで、先頭チャンク９１の最終ページである、初期化用のページ９２に、予めゼロデータのみを記憶させる。そして、ページ管理テーブル１５０を初期化する場合には、仮想ボリューム５０Ｖ内の全ての仮想的記憶領域５００を、初期化用のページ９２に対応付けておく。

【０１００】
これにより、仮想ボリューム５０Ｖの定義時（仮想ボリューム５０Ｖの生成時）に、テーブル群のマッピングに異常が生じていないことを予め確認することができる。さらに、図８に太線で示すように、仮想ボリューム５０Ｖにライトコマンドが発行された場合には、そのライトコマンドで指定される論理アドレスに対応する仮想的記憶領域５００を、初期設定用のページ９２からライトデータを書き込むべき所定のページに、接続し直すだけでよい。従って、仮想的記憶領域５００の対応付け先を切り替えるだけで、ライトデータを格納させることができ、仮想ボリューム５０Ｖの応答性能を高めることができる。
【０１０１】
図９は、移動先管理テーブル１６０の一例を示す。移動先管理テーブル１６０も、前記ページ管理テーブル１５０と同様に、例えば、キャッシュメモリ等に記憶させることができる。
【０１０２】
移動先管理テーブル１６０は、例えば、チャンクインデックス１６１と、チャンク情報管理テーブル１６２とを対応付けることにより構成される。チャンク情報管理テーブル１６２は、各チャンク毎に管理される管理情報として、移動先情報と、移動先決定用情報とを記憶する。
【０１０３】
移動先情報とは、移動対象のデータを格納するための格納先を示す情報である。移動先情報としては、例えば、移動先として決定されたＲＡＩＤグループ９０の番号等が用いられる。なお、ＲＡＩＤグループ番号に限らず、ＲＡＩＤグループ番号とチャンク番号の組合せ等を用いてもよい。
【０１０４】
移動先決定用情報とは、移動対象のデータの移動先を決定するために使用される情報である。移動先決定用情報としては、例えば、そのデータへのアクセス頻度等を用いることができる。後述のように、ＲＡＩＤグループ９０の使用容量（使用率）を移動先決定用情報として用いることもできる。
【０１０５】
さらに、移動先を決定するために用いられる負荷情報としては、アクセス頻度に限らず、例えば、アクセスサイズ（ライトデータのサイズ等）、アクセスパターン（シーケンシャルアクセスかランダムアクセスか）等を用いてもよい。
【０１０６】
チャンクインデックス１６１は、仮想ボリューム５０Ｖと、チャンク情報管理テーブル１６２の所定エントリとを対応付ける。所定エントリとは、チャンクインデックス１６１の示す仮想ボリューム５０Ｖに対応付けられているチャンク９０の管理情報を記憶しているエントリである。
【０１０７】
図１０は、記憶制御装置１０の全体動作を理解するためのフローチャートである。このフローチャートには、ユーザの手順も含まれている。まず最初に、ユーザは、管理サーバ７０を介して記憶制御装置１０に所定の指示を与えることにより、プール部６０にＲＡＩＤグループ９０を生成させ、さらに、そのＲＡＩＤグループ９０をフォーマットさせ、未割当てチャンクキュー等を作成させる（Ｓ１０）。ここで、プール部６０内で管理される、仮想ボリューム５０Ｖ用のＲＡＩＤグループ９０と、通常ボリューム５０Ｎ用のＲＡＩＤグループ９０とは、連番で管理される。
【０１０８】
続いて、ユーザは、管理サーバ７０から記憶制御装置１０に別の所定の指示を与えることにより、仮想ボリューム５０Ｖを作成させる（Ｓ１１）。上述の通り、仮想ボリューム５０Ｖの生成時に、各仮想的記憶領域５００はそれぞれ初期設定用のページ９２に対応付けられる。ここで、仮想ボリューム５０Ｖと通常ボリューム５０Ｎとは、連番で管理される。これにより、仮想ボリューム５０Ｖと通常ボリューム５０Ｎとを共通の管理方式で管

理することができ、記憶制御装置１０内に仮想ボリューム５０Ｖと通常ボリューム５０Ｎとを混在させることができる。

【０１０９】

続いて、ユーザは、ホスト２０と仮想ボリューム５０Ｖとを接続させる（Ｓ１２）。ユーザは、ホスト２０に繋がるＬＵＮ（Logical Unit Number）に、仮想ボリューム５０Ｖを接続させ、さらに、ＷＷＮ（Logical Unit Number）の登録等を行わせる。

【０１１０】

ホスト２０は、仮想ボリューム５０Ｖを認識し（Ｓ１３）、仮想ボリューム５０Ｖに向けてライトコマンド等のコマンドを発行する。記憶制御装置１０は、ホスト２０からのコマンドに応じた処理を行い、その結果をホスト２０に送信する（Ｓ１４）。

【０１１１】

記憶制御装置１０は、所定の周期により、または、管理サーバ７０からのユーザ指示により、仮想ボリューム５０Ｖに対応付けられている各ＲＡＩＤグループ９０内のデータについて、移動先を決定して記憶する（Ｓ１４）。

【０１１２】

記憶制御装置１０は、移動対象として選択されたデータにホスト２０がアクセスすると、そのホストアクセスに関するコマンドを処理しながら、予定されたデータ移動を実施する（Ｓ１５）。以下、各処理の詳細を図を改めて説明する。

【０１１３】

図１１は、プール作成処理を示すフローチャートである。以下に述べる各フローチャートは、各処理の概要を示す。いわゆる当業者であれば、図示されたステップの入れ替え、変更、削除、あるいは新たなステップの追加を行うことができるであろう。

【０１１４】

図１１に示す処理は、プール作成要求が与えられると開始される。以下、動作の主体をコントローラ３０とする。コントローラ３０は、作成対象のプール部６０のプールインデックス１５１について、そのプール状態を更新し（Ｓ２０）、ＲＡＩＤグループ９０を作成する（Ｓ２１）。コントローラ３０は、プール内ＲＡＩＤグループインデックス１１０の状態を”処理中”に変化させる（Ｓ２２）。

【０１１５】

コントローラ３０は、管理情報を退避させるための領域をＲＡＩＤグループ９０内に設定し（Ｓ２３）、さらに、チャンクインデックス１２０を作成する（Ｓ２４）。プール部６０内の全チャンク９１について、以下のステップＳ２６～Ｓ２８がそれぞれ実行される（Ｓ２５）。

【０１１６】

コントローラ３０は、対象チャンクに対応するページアドレス情報１５５を初期化し（Ｓ２６）、対象チャンクの状態を”フォーマット待ち”に変更する（Ｓ２７）。コントローラ３０は、管理情報を退避させるための退避要求キューに、管理情報の対比要求をエンキューする（Ｓ２８）。

【０１１７】

各チャンク９１についてＳ２６～Ｓ２８を実行した後、コントローラ３０は、プール内ＲＡＩＤグループインデックス１１０の状態を”有効”に変更する（Ｓ２９）。そして、コントローラ３０は、キャッシュメモリ３４０にヌルデータをステージングさせて（Ｓ３０）、本処理を終了する。ライトデータの書き込まれていない仮想的記憶領域５００からのデータ読み出しが要求された場合に、物理的記憶領域であるページ９２にアクセスすることなく、ホスト２０にヌルデータを返すためである。

【０１１８】

図１２は、データ及び保証コードについて説明する図である。図１２（ａ）に示すように、本実施例では、例えば、５１２バイトのデータＤ１０毎に８バイトの保証コードＤ１１を付加して、記憶装置４０に記憶させる。保証コードＤ１１は、論理アドレスを検証するための部分と、ビットエラーを検証するための部分とを含むことができる。以下では、

論理アドレスを検証するための部分に着目して説明する。

【０１１９】

図１２（ｂ）は、通常ボリューム５０Ｎに記憶されるデータに付加される、保証コードの構成を示す。通常ボリューム５０Ｎに関する保証コードのうち論理アドレスの検証に使用される部分は、４ビットの予約領域Ｄ１１０と、１２ビットのＬＵＮ領域Ｄ１１１と、１６ビットのＬＢＡ領域Ｄ１１２とを含んでいる。ＬＢＡとは、Logical Block Addressの略である。ＬＵＮ領域Ｄ１１１には、通常ボリューム５０Ｎに対応付けられるＬＵＮが格納される。ＬＢＡ領域Ｄ１１２には、データＤ１０の論理アドレスが格納される。

【０１２０】

図１２（ｃ）は、仮想ボリューム５０Ｖに記憶されるデータに付加される、保証コードの構成を示す。仮想ボリューム５０Ｖに関する保証コードのうち論理アドレスの検証に使用される部分は、４ビットの予約領域Ｄ１１０と、８ビットのＲＡＩＤグループ識別領域Ｄ１１３と、４ビットのチャンク識別領域Ｄ１１４と、１６ビットのチャンク内ＬＢＡオフセット領域Ｄ１１５とを含んでいる。ＲＡＩＤグループ識別領域Ｄ１１３には、ＲＡＩＤグループ９０を識別するための情報が格納される。チャンク識別領域Ｄ１１４には、チャンク９１を識別するための情報のうち下位の４ビットが格納される。チャンク内ＬＢＡオフセット領域Ｄ１１５には、データＤ１０が格納されているチャンク９１内において、そのチャンク９１の先頭論理アドレスからのオフセット値が格納される。

【０１２１】

図１３，図１４に基づいて、チャンク９１を管理するためのキューを説明する。図１３は、キューの使用方法を示す説明図である。プール部６０が作成されると、プール部６０で管理される各ＲＡＩＤグループ９０の有する各チャンク９１は、フォーマット待ちチャンクキューＱ１０に登録されて管理される。フォーマットが開始されると、フォーマット中のチャンク９１は、処理完了待ちチャンクキューＱ２０に移される。そして、フォーマットが完了すると、フォーマット済チャンク９１は、未割当てチャンクキューＱ３０に移される。

【０１２２】

図１４は、各キューＱ１０，Ｑ２０，Ｑ３０を模式的に示す説明図である。フォーマット待ちチャンクキューＱ１０は、各ＲＡＩＤグループ９０毎に用意される。そして、フォーマット処理を開始する場合、フォーマット待ちチャンクキューＱ１０内の各ＲＡＩＤグループ９０から所定の順番でチャンク９１が取り出されて、処理完了待ちキューＱ２０に接続される。フォーマットが完了したチャンク９１は、上述の通り、未割当てチャンクキューＱ３０に接続される。処理完了待ちチャンクキューＱ２０に繋がれた順番で、フォーマットは完了するため、処理完了待ちキューＱ２０内の各チャンク９１の順番と、未割当てチャンクキューＱ３０内の各チャンク９１の順番とは、通常の場合、一致する。

【０１２３】

図１５，図１６に基づいてチャンク９１をフォーマットする処理を説明する。図１５は、フォーマット処理の全体を示し、図１６は、フォーマット処理の一部を示す。

【０１２４】

コントローラ３０は、フォーマット待ちチャンクキューＱ１０を確認することにより、フォーマット待ちのチャンク９１が有るか否かを判定する（Ｓ５０）。フォーマット待ちチャンクキューＱ１０にチャンク９１が登録されている場合（S50:YES）、コントローラ３０は、プール部６０内の各ＲＡＩＤグループ９０毎に、ラウンドロビン方式でチャンクを選択し、以下のステップＳ５１～Ｓ５４を実行する。

【０１２５】

図１４に示す例で選択方法を説明すると、第１ＲＡＩＤグループ９０（＃１）から一つのチャンク９１（１－１）を選択した後、第２ＲＡＩＤグループ９０（＃２）から別の一つのチャンク９１（２－１）を選択し、さらに、第３ＲＡＩＤグループ９０（＃３）からさらに別の一つのチャンク９１（３－１）を選択する。選択されたチャンク９１（１－１）、９１（２－１），９１（３－１）について、後述するＳ５２～Ｓ５４をそれぞれ実行

する。チャンク９１（１－２），９１（２－２），９１（３－２）のセットについても、Ｓ５２～Ｓ５４を実行する。さらに、チャンク９１（１－３），９１（２－３），９１（３－３）のセットについても、Ｓ５２～Ｓ５４を実行する。以下同様である。

【０１２６】

コントローラ３０は、フォーマット待ちチャンクキューＱ１０から、対象ＲＡＩＤグループ９０のチャンク９１を一つ選択し、フォーマット待ちチャンクキューＱ１０からデキューさせる（Ｓ５２）。コントローラ３０は、選択されたチャンク９１を処理完了待ちチャンクキューＱ２０にエンキューし（Ｓ５３）、選択されたチャンク９１についてのフォーマットジョブを実行する（Ｓ５４）。フォーマットジョブの詳細については、図１６と共に後述する。

【０１２７】

フォーマットジョブが完了すると、コントローラ３０は、フォーマット済のチャンク９１を、処理完了待ちチャンクキューＱ２０からデキューし（Ｓ５５）、そのフォーマット済のチャンク９１を未割当てチャンクキューＱ３０にエンキューする（Ｓ５６）。

【０１２８】

図１６は、図１４中にＳ５４で示されるフォーマットジョブの詳細を示すフローチャートである。コントローラ３０は、処理対象のチャンク９１の識別番号を取得し（Ｓ６０）、対象チャンク９１をフォーマットする範囲を決定する（Ｓ６１）。そして、コントローラ３０は、対象チャンク９１についてのエクステントロックを取得する（Ｓ６２）。これにより、対象チャンク９１が別のプログラムによって使用されるのを防止する。

【０１２９】

コントローラ３０は、対象チャンク９１について所定サイズ毎に、後述するＳ６４～Ｓ７０を実行する（Ｓ６３）。つまり、コントローラ３０は、対象チャンク９１を、所定サイズの単位領域毎にフォーマットするようになっている。

【０１３０】

コントローラ３０は、データ用のキャッシュセグメントを確保し（Ｓ６４）、続いて、パリティ用のキャッシュセグメントを確保する（Ｓ６５）。キャッシュセグメントとは、キャッシュメモリ３４０の記憶領域を管理する単位である。

【０１３１】

コントローラ３０は、論理アドレスを算出し（Ｓ６６）、ゼロデータの生成を要求し（Ｓ６７）、さらに、パリティの生成を要求する（Ｓ６８）。ゼロデータは、Ｓ６４で確保されたキャッシュセグメントを用いて生成される。パリティは、Ｓ６５で確保されたキャッシュセグメントを用いて生成される。コントローラ３０は、データ用に確保されたキャッシュセグメントを開放させ（Ｓ６９）、さらに、パリティ用に確保されたキャッシュセグメントも開放させる（Ｓ７０）。

【０１３２】

コントローラ３０は、対象チャンク９１のエクステントロックを開放し（Ｓ７１）、対象チャンク９１についてのフォーマット完了を確認してから（S72:YES）、図１５の処理に戻る。

【０１３３】

図１７は、仮想ボリューム５０Ｖを作成する処理を示すフローチャートである。コントローラ３０は、プール部６０の状態が正常であるか否かを判定する（Ｓ８０）。プール部６０に障害が発生している場合（S80:NO）、仮想ボリューム５０Ｖを生成することはできないため、コントローラ３０は、エラー処理を実行する（Ｓ８１）。エラー処理では、例えば、プール部６０に異常が生じているために仮想ボリューム５０Ｖを生成できない旨を、ユーザに通知する。

【０１３４】

プール部６０が正常な場合（S80:YES）コントローラ３０は、仮想ボリュームインデックス１５２の状態を”無効”から”処理中”に変更させる（Ｓ８２）。予め用意されている仮想ボリュームの識別番号には、仮想ボリュームの状態の初期値として”無効”が予め

設定されている。仮想ボリューム５０Ｖの生成中では、その状態は”無効”から”処理中”に変化する。仮想ボリューム５０Ｖの生成が完了すると、その状態は”処理中”から”有効”に変化する。

【０１３５】

コントローラ３０は、状態が”処理中”に変更された仮想ボリューム５０Ｖについて、仮想アドレスインデックス１５３を作成し（Ｓ８３）、さらに、その仮想アドレスインデックス１５３に対応付けられる仮想アドレスブロック１５４を作成する（Ｓ８４）。さらに、コントローラ３０は、仮想アドレスブロック１５４に対応付けられるページアドレス情報１５５を作成する（Ｓ８５）。

【０１３６】

コントローラ３０は、プール内ＲＡＩＤグループインデックス１１０を更新し（Ｓ８６）、仮想ボリュームインデックス１５２の状態を”処理中”から”有効”に変更する（Ｓ８７）。コントローラ３０は、仮想ボリューム５０Ｖを正常に作成することができたか否かを確認し（Ｓ８８）、仮想ボリューム５０Ｖを正常に作成できた場合には（S88:YES）、本処理を終了させる。仮想ボリュームを正常に作成できなかった場合（S88:NO）、エラー処理が行われる（Ｓ８９）。エラー処理では、例えば、仮想ボリューム５０Ｖを正常に作成できなかった旨をユーザに通知する。

【０１３７】

なお、便宜上、仮想ボリューム５０Ｖを正常に作成できたか否かを最後に判定するかのように説明したが、実際には、各テーブル１０１－１０４の作成時にそれぞれ正常に作成できたか否かが判定される。そして、正常に作成できなかった場合には、エラー処理が行われる。

【０１３８】

図１７に示す処理を行うことにより、図５に示すテーブル群が作成され、仮想ボリューム５０Ｖ内の各仮想的記憶領域５００は、初期設定用のページ９２にそれぞれ対応付けられる。従って、仮想ボリューム５０Ｖを正常に作成できた時点で、各仮想的記憶領域５００と実ページ９２との対応付けを正常に行えることが確認される。

【０１３９】

図１６は、ライトコマンドの処理を示すフローチャートである。コントローラ３０は、ホスト２０からコマンドを受領すると、そのコマンドが仮想ボリューム５０Ｖを対象とするライトコマンドであるか否かを判定する（Ｓ１００）。

【０１４０】

そのコマンドが、通常ボリューム５０Ｎを対象とするライトコマンドである場合（S100:NO）、コントローラ３０は、通常のライト処理を実行する（Ｓ１０１）。通常のライト処理では、例えば、ホスト２０から受領したライトデータをキャッシュメモリ３４０に書き込み、キャッシュメモリ３４０へのライトデータ書込みが完了した時に、ホスト２０に処理完了を通知する。その後、適切なタイミングを見計らって、キャッシュメモリ３４０に記憶されたライトデータを、記憶装置４０に書き込む。

【０１４１】

仮想ボリューム５０Ｖを対象とするライトコマンドを受領した場合（S100:YES）、コントローラ３０は、移動先管理テーブル１６０を参照することにより、書き込み対象の仮想ページ５００に割り当てられている実ページ９２に移動先が設定されているか否かを判定する（Ｓ１０２）。

【０１４２】

ライトデータの書き込まれる実ページ９２に移動先が設定されている場合（S102:YES）、コントローラ３０は、移動先のＲＡＩＤグループ９０に利用可能なチャンク９１が有るか否かを判定する（Ｓ１０３）。利用可能なチャンク９１とは、既に作成されており、仮想ボリューム５０Ｖに対応付けられているチャンクである。このように、既に作成されているチャンク９１を利用して、ページを移動させることにより、割り当てられるチャンクの数を少なくできる。つまり、記憶容量を効率的に使用できる。

【０１４３】
利用可能なチャンク９１が無い場合（S103:NO）、コントローラ３０は、初期状態になっている新規チャンク９１が有るか否かを判定する（Ｓ１０４）。つまり、コントローラ３０は、直ちに使用できる新規チャンク９１が、移動先ＲＡＩＤグループ９０内に用意されているか否かを判定する。
【０１４４】
使用可能な新規チャンク９１が移動先ＲＡＩＤグループ９０内に無い場合（S104:NO）、コントローラ３０は、エラー処理を行う（Ｓ１０７）。エラー処理では、例えば、「記憶容量不足です。」、「ＲＡＩＤグループを作成してください。」等のエラーメッセージを、管理サーバ７０を介してユーザに通知する。
【０１４５】
移動先ＲＡＩＤグループ９０内に利用可能なチャンク９１が存在する場合（S103:YES）、コントローラ３０は、Ｓ１０４をスキップしてＳ１０５に移る。
【０１４６】
コントローラ３０は、書き込み対象の実ページ９２に記憶されるデータを、移動先ＲＡＩＤグループ９０内に設けられる新ページ９２に移動させるべく、チャンク割当て変更処理（Ｓ１０５）と、ページ割当て変更処理（Ｓ１０６）とを実行する。チャンク割当て変更処理及びページ割当て変更処理については、後述する。その後、コントローラ３０は、図１９のＳ１１４に移る。
【０１４７】
つまり、Ｓ１０２～Ｓ１０６は、移動先の設定されている実ページ９２についてライトコマンドが発行された場合に、移動先チャンク９１内の移動先ページ９２を、ライトコマンドの対象でる仮想ページ５００に対応付ける。換言すれば、書き込み対象の仮想ページ５００の割当て先を、移動元ページ９２から移動先ページ９２に切り替える。
【０１４８】
図１９は、図１８に続くフローチャートである。書き込み対象の実ページ９２に移動先ＲＡＩＤグループ９０が設定されていな場合（S102:NO）、コントローラ３０は、ライトコマンドで指定された仮想ボリューム５０Ｖについて、現在使用中のチャンク９１が有るか否かを判定する（Ｓ１０８）。使用中のチャンク９１が有る場合（S108:YES）、後述のＳ１１２に移行する。使用中のチャンク９１が無い場合（S108:NO）、コントローラ３０は、仮想ボリューム５０Ｖに新たに割り当てるべきチャンク９１が初期状態になっているか否かを判定する（Ｓ１０９）。
【０１４９】
通常の場合、図１５に示すフォーマット処理によって、新規チャンクは初期状態に設定されている。それにもかかわらず、新規チャンク９１が初期状態になっていない場合（S109:NO）、エラー処理が行われる（Ｓ１１０）。エラー処理では、例えば、初期状態のチャンク９１が存在しない旨を、管理サーバ７０を介してユーザに通知する。
【０１５０】
新規チャンク９１が初期状態になっている場合（S109:YES）、コントローラ３０は、チャンク割当て変更処理を行う（Ｓ１１１）。チャンク割当て変更処理の詳細は、図２１で述べるが、先に簡単に説明すると、コントローラ３０は、未割当てチャンクキューＱ３０から一つのチャンク９１を選択して仮想ボリューム５０Ｖに対応付け、そのチャンク９１の状態を”割当済み（使用中）”に変更等する。
【０１５１】
コントローラ３０は、チャンク９１内の各ページ９２のうち、使用しようとするページ９２が初期状態になっているか否かを判定する（Ｓ１１２）。使用しようとするページ９２が初期状態の場合（S112:YES）、ページ割当て変更処理が行われる（Ｓ１１３）。
【０１５２】
ページ割当て変更処理の詳細は、図２２で後述する。簡単に説明すると、ページ割当て変更処理では、仮想ボリューム５０Ｖに割り当てられるページ９２の状態を”割当済み（

使用中）”に変更し、仮想アドレスブロック１５４及びページアドレス情報１５５を更新させる。Ｓ１１３の後、Ｓ１１４に移る。

【０１５３】

使用しようとするページ９２が初期状態ではない場合（S112:NO）、つまり、使用しようとするページ９２が初期設定用のページではない場合、Ｓ１１３はスキップされて、Ｓ１１４に移る。

【０１５４】

コントローラ３０は、ライトデータを記憶させるためのキャッシュセグメントを確保し（Ｓ１１４）、さらに、ライトデータを転送するためのＤＭＡ（Direct Memory Access）転送リストを作成する（Ｓ１１５）。そして、コントローラ３０は、保証コードのアドレス部分（ＬＡ）を算出する（Ｓ１１６）。

【０１５５】

図２０を参照する。コントローラ３０は、ホスト２０から受領したライトデータをキャッシュメモリ３４０にＤＭＡ転送させる（Ｓ１１７）。キャッシュメモリ３４０にライトデータを記憶させた後で、コントローラ３０は、ライトコマンドの処理が完了した旨をホスト２０に通知する（Ｓ１１８）。

【０１５６】

キャッシュメモリ３４０へのライトデータ書込み完了後に、ホスト２０に処理完了を通知する方式を非同期方式と呼ぶ。これに対し、記憶装置４０にライトデータが書き込まれるのを待ってから、ホスト２０に処理完了を通知する方式を同期方式と呼ぶ。非同期方式または同期方式のいずれを用いても良い。

【０１５７】

コントローラ３０は、ライトデータに設定されている論理アドレスを、記憶装置４０に記憶させるための物理アドレスに変換する（Ｓ１１９）。コントローラ３０は、キャッシュセグメントを確保する（Ｓ１２０）。コントローラ３０は、パリティを生成するために必要な旧データを記憶装置４０から読み出して（Ｓ１２１）、Ｓ１２０で確保したキャッシュセグメントに格納させる。

【０１５８】

コントローラ３０は、ホスト２０から受領したライトデータと記憶装置４０から読み出した旧データとに基づいて、新たなパリティを算出する（Ｓ１２２）。コントローラ３０は、キャッシュメモリ３４０に記憶されたライトデータを、記憶装置４０（図２０中、ディスクと表示）に転送して記憶させる（Ｓ１２３）。

【０１５９】

コントローラ３０は、キャッシュメモリ３４０に記憶されているライトデータの状態を”ダーティ”から”クリーン”に変更する（Ｓ１２４）。”ダーティ”状態とは、キャッシュメモリ３４０だけに記憶されている状態を示す。”クリーン状態”とは、記憶装置４０に書き込まれた状態を示す。最後に、コントローラ３０は、Ｓ１１４及びＳ１２０で確保されたキャッシュセグメントを開放し、本処理を終了する（Ｓ１２５）。

【０１６０】

上述の通り、書き込み対象の実ページ９２（移動元ページ）に移動先が設定されている場合、Ｓ１０２～Ｓ１０６によって、その移動元ページ９２に対応付けられていた仮想ページ５００は、移動先ページ９２に新たに対応付けられる。従って、Ｓ１１９～Ｓ１２５では、移動先ページ９２にライトデータが書き込まれる。これにより、ライトコマンドを処理する中で、移動対象データを移動元ページ９２から移動先ページ９２に移動させることができる。

【０１６１】

図２１は、図１８のＳ１０５及び図１９のＳ１１１に示されるチャンク割当て変更処理の詳細を示すフローチャートである。コントローラ３０は、処理対象チャンク９１の状態を所定の状態に変更させる（Ｓ１３０）。

【０１６２】

例えば、新規チャンク９１を仮想ボリューム５０Ｖに割り当てる場合、新規チャンク９１の状態は”未割当て（未使用）”から”割当済み（使用中）”に変更される。また、例えば、仮想ボリューム５０Ｖに割り当てられているチャンク９１を開放する場合、そのチャンク９１の状態は”割当済み（使用中）”から"フォーマット待ち"に変更される。

【０１６３】

コントローラ３０は、チャンク９１の状態を変更した後で、そのチャンク９１に対応する仮想アドレスインデックス１５３を更新させる（Ｓ１３１）。さらに、コントローラ３０は、チャンクインデックス１２０を更新させ（Ｓ１３２）、続いて、プール内ＲＡＩＤグループインデックス１１０を更新させる（Ｓ１３３）。

【０１６４】

図２２は、図１８中のＳ１０６及び図１９のＳ１１３に示されるページ割当て変更処理の詳細を示すフローチャートである。コントローラ３０は、処理対象ページ９２の状態を所定の状態に変更させる（Ｓ１４０）。例えば、新規ページ９２を仮想ボリューム５０Ｖ内の仮想的記憶領域５００に割り当てる場合、その新規ページ９２の状態は”未割当て（未使用）”から”割当済み（使用中）”に変更される。また、例えば、仮想ボリューム５０Ｖに割当済みのページ９２を開放する場合、その割当済みページ９２の状態は”割当済み（使用中）”から”フォーマット待ち”に変更される。仮想ボリューム５０Ｖに新たなページ９２を割り当てる方法の詳細については、図２３で後述する。

【０１６５】

ページ状態を変更した後、コントローラ３０は、処理対象ページ９２に対応する仮想アドレスブロック１５４を更新させ（Ｓ１４１）、さらに、処理対象ページ９２に対応するページアドレス情報１５５を更新させる（Ｓ１４２）。そして、コントローラ３０は、管理情報（図６に示すテーブル群）を、管理情報退避領域（図１１のＳ２３参照）に退避させる（Ｓ１２３）。

【０１６６】

図２３は、図２２のＳ１４０で示される処理の一部を示すフローチャートである。上述の通り、ホスト２０からのライトコマンドに応じて、仮想ボリューム５０Ｖにチャンク単位で実記憶領域が割り当てられ、その割り当てられたチャンク９１内の実ページ９２が順番に使用されていく。割り当てられたチャンク９１を使い切ると、新たなチャンク９１が仮想ボリューム５０Ｖに割り当てられる。以下、図２３を参照しながら、ページ９２の使用方法を説明する。

【０１６７】

コントローラ３０は、各ページ９２の状態を管理するためのページ状態管理テーブル１５８を備える。そのテーブル１５８は、例えば、チャンク識別番号の欄１５８０と、ページ識別番号の欄１５８１と、ページの状態を示す欄１５８２とを備える。

【０１６８】

ページ状態欄１５８２には、例えば、”使用中（あるいは割当済み）”、”未使用（あるいは未割当て）”、”開放（あるいはフォーマット待ち）”等のような予め用意されている状態のうちのいずれか一つの状態が設定される。なお、説明の便宜上、ページ状態管理テーブル１５８は、ページアドレス情報１５５と別体の情報であるかのように示すが、実際には、ページアドレス情報１５５だけで各ページの状態を管理できる。

【０１６９】

コントローラ３０は、テーブル１５８を参照し、現在使用中のチャンク９１内に未使用ページ９２が有る場合、その未使用ページ９２を使用する（Ｓ１４００）。現在使用中のチャンク９１内に未使用ページ９２が無い場合、コントローラ３０は、テーブル１５８を参照し、現在使用中のチャンク９１内に開放ページ９２が有る場合、その開放されたページ９２を使用する（Ｓ１４０１）。

【０１７０】

現在使用中のチャンク９１内に未使用ページ９２も開放ページ９２もいずれも存在しない場合、コントローラ３０は、テーブル１５８を参照し、使用済チャンク９１内の開放ペ

ージを使用する（Ｓ１４０２）。つまり、コントローラ３０は、対象の仮想ボリューム５０Ｖについて既に使用されたチャンク９１から、開放されたページ９２を検出して再使用する。

【０１７１】

なお、使用済チャンク９１内に開放ページ９２が無い場合、コントローラ３０は、図１９で述べたように、未使用のチャンク９１を仮想ボリューム５０Ｖに対応付け、そのチャンク９１の先頭ページ９２を仮想ボリューム５０Ｖに割り当てる。

【０１７２】

図２４は、リードコマンドの処理を示すフローチャートである。コントローラ３０は、ホスト２０から受領したコマンドが、仮想ボリューム５０Ｖを対象とするリードコマンドであるか否かを判別する（Ｓ１５０）。

【０１７３】

ホスト２０から受領したコマンドが通常ボリューム５０Ｎへのリードコマンドである場合（S150:NO）、コントローラ３０は、通常のリード処理を実行する（Ｓ１５１）。例えば、コントローラ３０は、ホスト２０から要求されたデータがキャッシュメモリ３４０に記憶されているか否かを判定する。要求されたデータがキャッシュメモリ３４０上に存在する場合、コントローラ３０は、キャッシュメモリ３４０からデータを読み出してホスト２０に送信する。ホスト２０の要求するデータがキャッシュメモリ３４０上に存在しない場合、コントローラ３０は、記憶装置４０からデータを読み出してキャッシュメモリ３４０に記憶させ、そのデータをホスト２０に送信する。

【０１７４】

ホスト２０から発行されたコマンドが、仮想ボリューム５０Ｖからデータを読み出すためのリードコマンドである場合（S150:YES）、コントローラ３０は、リード対象の仮想ボリューム５０Ｖについてエクステントロックを取得する（Ｓ１５２）。

【０１７５】

リードコマンドは、データの読出し先の論理アドレスを指定する。コントローラ３０は、指定された論理アドレスに対応する仮想的記憶領域５００を検出し、図６に示すテーブル群を参照して、その仮想的記憶領域５００に割り当てられているページ９２の状態を取得する。コントローラ３０は、リード対象のページ９２の状態が”初期状態”であるか否かを判定する（Ｓ１５３）。

【０１７６】

リード対象のページ９２が初期状態の場合（S153:YES）、コントローラ３０は、ホスト２０に送信すべきヌルデータがキャッシュメモリ３４０に記憶されているか否かを判定する（Ｓ１６０）。図１１のＳ３０で述べたように、プール部６０の作成時に、キャッシュメモリ３４０の所定のキャッシュセグメントにヌルデータが予め記憶される。従って、通常の場合、Ｓ１６０では”ＹＥＳ”と判定されて、後述のＳ１６１に移る。キャッシュメモリ３４０にヌルデータが記憶されていない場合（S160:NO）、後述のＳ１５５に移り、所定チャンクの所定ページ（例えば、初期設定用ページ）から、ヌルデータが読み出されて、ホスト２０に送信される（Ｓ１５５～Ｓ１５９，Ｓ１７１～Ｓ１７５）。

【０１７７】

Ｓ１５３に戻る。リード対象のページ９２が初期状態では無い場合（S153:NO）、つまり、リード対象ページにライトデータが書き込まれている場合、コントローラ３０は、リード対象データに関するパリティを算出する（Ｓ１５４）。そして、コントローラ３０は、キャッシュセグメントを確保し（Ｓ１５５）、第２通信回路３２０に向けてリード要求を発行する（Ｓ１５６）。コントローラ３０は、論理アドレスを物理アドレスに変換し（Ｓ１５７）、保証コードのアドレス部分（ＬＡ）を算出する（Ｓ１５８）。コントローラ３０は、第２通信回路３２０を介して、記憶装置４０からキャッシュメモリ３４０にリード対象データを転送させる（Ｓ１５９）。

【０１７８】

図２５に移る。コントローラ３０は、キャッシュメモリ３４０から第１通信回路３１０

にＤＭＡ転送させるための、ＤＭＡ転送リストを設定する（Ｓ１６１）。続いて、コントローラ３０は、キャッシュメモリ３４０上に記憶されているデータを、第１通信回路３１０を介してホスト２０に送信させる（Ｓ１６２）。そして、コントローラ３０は、リードコマンドに係る仮想ページ５００に対応付けられている実ページ９２に、移動先ＲＡＩＤグループ９０が設定されているか否かを判定する（Ｓ１６３）。リード対象の実ページ９２に移動先ＲＡＩＤグループ９０が設定されていない場合（S163:NO）、図２６に示すフローチャートに移る。

【０１７９】

リード対象の実ページ９２に移動先ＲＡＩＤグループ９０が設定されている場合（S163:YES）、コントローラ３０は、移動先ＲＡＩＤグループ９０に利用可能なチャンク９１が有るか否かを判定する（Ｓ１６４）。利用可能なチャンク９１が無い場合（S164:NO）、移動先ＲＡＩＤグループ９０に初期状態の新規チャンク９１が有るか否かを判定する（Ｓ１６５３）。

【０１８０】

移動先ＲＡＩＤグループ９０内に初期状態の新規チャンク９１が無い場合（S165:NO）、コントローラ３０は、図２６に示すフローチャートに移る。移動対象のデータを移動先ＲＡＩＤグループ９０に移動させることができないためである。

【０１８１】

一方、移動先ＲＡＩＤグループ９０内に利用可能なチャンク９１が有る場合（S164:YES）、または、移動先ＲＡＩＤグループ９０が初期状態の新規チャンク９１を有する場合（S165:YES）、のいずれかの場合には、コントローラ３０は、リードコマンドの処理が完了した旨をホスト２０に通知する（Ｓ１６６）。

【０１８２】

さらに、コントローラ３０は、読み出したデータを移動先ＲＡＩＤグループ９０内に移動させるべく、以下のステップを実行する。コントローラ３０は、図２１で述べたチャンク割当て処理（Ｓ１６７）と、図２２で述べたページ割当て変更処理（Ｓ１６８）とを実行する。これにより、リード対象の仮想ページ５００は、移動元ページ９２から、移動先ＲＡＩＤグループ９０内の新たなページ９２に対応付けられる。

【０１８３】

コントローラ３０は、キャッシュメモリ３４０上のデータを、記憶装置４０に転送させて書き込ませる（Ｓ１６９）。コントローラ３０は、キャッシュセグメントを開放し（Ｓ１７０）、さらに、エクステントロックを開放する（Ｓ１７１）。

【０１８４】

一方、リード対象の実ページ９２に移動先ＲＡＩＤグループ９０が設定されていない場合（S163:NO）、図２６に移る。

【０１８５】

コントローラ３０は、Ｓ１５５で確保したキャッシュセグメントを開放し（Ｓ１７２）、エクステントロックも開放する（Ｓ１７３）。最後に、コントローラ３０は、リードコマンドの処理が完了した旨をホスト２０に通知し（Ｓ１７４）、本処理を終了する。

【０１８６】

このように、コントローラ３０は、リードコマンドを処理する中で、移動対象のデータを予め設定される移動先に移動させる。そこで、移動先を決定する方法を説明する。

【０１８７】

図２７は、移動先を決定するための処理を示すフローチャートである。データを再配置させるための処理と呼び変えてもよい。本処理は、例えば、予め設定される所定の周期毎に実行される。または、ユーザからの指示に応じて本処理を実行することもできる。

【０１８８】

コントローラ３０は、移動先管理テーブル１６０に設定されている移動先ＲＡＩＤグループ９０に関する情報をクリアさせる（Ｓ１８０）。つまり、新たなデータ移動計画を作成する前に、前回作成されたデータ移動計画をリセットする。以下、各プール部６０毎に

（Ｓ１８１）、負荷を分散させる処理（Ｓ１８２）と、使用容量を平均化させる処理（Ｓ１８３）とが行われる。

【０１８９】

図２８は、図２７のＳ１８２に示される負荷分散処理のフローチャートである。コントローラ３０は、予め設定される条件に従って、移動元ＲＡＩＤグループ９０と、移動先ＲＡＩＤグループ９０とを選択する（Ｓ１９０）。

【０１９０】

例えば、コントローラ３０は、使用容量が５０％以下であるＲＡＩＤグループの中から、負荷が最も低いＲＡＩＤグループを移動先ＲＡＩＤグループとして選択する。さらに、コントローラ３０は、ＲＡＩＤグループの中から、負荷が最も高いＲＡＩＤグループを移動元ＲＡＩＤグループとする。

【０１９１】

もしも、上記条件を満たすＲＡＩＤグループが存在せず、移動元ＲＡＩＤグループまたは移動先ＲＡＩＤグループのいずれか一つでも決定することができなかった場合、コントローラ３０は、図２８に示すループを抜ける。

【０１９２】

コントローラ３０は、移動元チャンクを決定する（Ｓ１９１）。今までの説明では、ページ単位でデータを移動させる場合を述べたが、チャンク単位でデータを移動させることもできる。

【０１９３】

コントローラ３０は、移動元ＲＡＩＤグループに所属する各チャンクの中で負荷が高いチャンクから順番に移動元チャンクとして決定する。移動元チャンクに指定されるチャンクの負荷の合計は、ＲＡＩＤグループ間の負荷差の５０％以内とする。

【０１９４】

データの移動によって、移動元ＲＡＩＤグループの負荷が低下し、移動先ＲＡＩＤグループの負荷は増大する。ＲＡＩＤグループ間の負荷差を考慮して移動元チャンクを選定することにより、データ移動の結果、移動元ＲＡＩＤグループの負荷が移動先ＲＡＩＤグループの負荷よりも小さくなるのを防止できる。

【０１９５】

コントローラ３０は、上記方法で選ばれた移動元ＲＡＩＤグループ、移動先グループ及び移動元チャンクを、移動先管理テーブル１６０に記憶させる（Ｓ１９２）。コントローラ３０は、Ｓ１９０で選択された移動元ＲＡＩＤグループ及び移動先ＲＡＩＤグループを、処理対象から除外する（Ｓ１９３）。

【０１９６】

コントローラ３０は、プール部６０内のＲＡＩＤグループ数が１以下になるか、または、上記条件を満たすＲＡＩＤグループが見つからなくなるまで、本処理を繰り返す。

【０１９７】

図２９は、図２７のＳ１８３に示される負荷分散処理のフローチャートである。コントローラ３０は、予め設定される条件に従って、移動元ＲＡＩＤグループ９０と、移動先ＲＡＩＤグループ９０とを選択する（Ｓ２００）。

【０１９８】

例えば、コントローラ３０は、使用容量が５０％以下のＲＡＩＤグループの中から、使用容量が最も小さいＲＡＩＤグループを移動先ＲＡＩＤグループとして選択する。さらに、コントローラ３０は、ＲＡＩＤグループの中で使用容量が最も大きいＲＡＩＤグループを移動元ＲＡＩＤグループとして選択する。

【０１９９】

もしも、上記条件を満たすＲＡＩＤグループが存在せず、移動元ＲＡＩＤグループまたは移動先ＲＡＩＤグループのいずれか一つでも決定することができなかった場合、コントローラ３０は、図２９に示すループを抜ける。

【０２００】

コントローラ３０は、移動元チャンクを決定する（Ｓ１９１）。例えば、コントローラ３０は、移動元ＲＡＩＤグループに所属するチャンクのうち未だ移動先が決定されておらず、かつ、負荷が低いチャンクから順番に、移動元チャンクに決定する。負荷が０のチャンクは選択対象から外される。

【０２０１】

コントローラ３０は、移動元チャンクを決定する（Ｓ２０１）。移動元チャンクとして選択されるチャンクの数は、ＲＡＩＤグループ間の未使用チャンク数差の５０％以内に保たれる。

【０２０２】

さらに、移動先ＲＡＩＤグループの負荷が、移動元ＲＡＩＤグループの負荷よりも高い場合、移動元チャンクとして選定されるチャンクの数は、移動先ＲＡＩＤグループにおいて移動先が指定されているチャンクの数以下に保持される。

【０２０３】

このように構成される本実施例では、ホストアクセスが発生する前に、仮想ボリューム５０Ｖに割り当てられている実ページ９２のデータを移動させるためのデータ移動計画を作成して保存する。本実施例では、移動対象データについてホストアクセスが発生すると、そのホストアクセスに係わるコマンド処理の中で、移動対象のデータを移動させることができる。つまり、データマイグレーション用の特別なプログラムを、コマンド処理のためのプログラムと別に実行するのではなく、コマンドを処理するための一連の流れの中でデータ移動を行う。

【０２０４】

従って、本実施例では、比較的簡易な構成で、比較的効率的にデータを移動させることができ、プール部６０内の実記憶領域を略均等に使用することができ、特定のＲＡＩＤグループにアクセスが偏ったりするのを抑制できる。これにより、記憶制御装置１０の応答性能の低下を防止できる。

【０２０５】

上述のデータ移動に関する本実施例の効果は、本実施例の基本的構成と結合することにより発揮される。本実施例の基本的構成とは、仮想ボリューム５０Ｖにチャンク単位で実記憶領域を割当て、かつ、一つのチャンク９１は一つの仮想ボリューム５０Ｖに専属させる構成である。

【０２０６】

本実施例では、図３０に示すように、物理的記憶領域であるＲＡＩＤグループ９０を効率的に使用することができる。図３０は、本発明の効果を模式的に示す説明図である。図３０（ａ）は、本発明を適用しない場合を示し、図３０（ｂ）は、本発明を適用した場合を示す。

【０２０７】

通常ボリューム５０Ｎの場合を先に説明する。通常ボリューム５０Ｎの場合は、ＲＡＩＤグループ９０内の連続した記憶領域を使用することができる。従って、一つのストライプ列に、複数の通常ボリューム５０Ｎに関するデータが混在することは無い。

【０２０８】

仮想ボリューム５０Ｖの場合、必要に応じて実記憶領域が割り当てられ、データは離散的に管理される。もしも、仮想ボリュームにページ単位で実記憶領域を割り当てる場合は、同一のストライプ列に複数の仮想ボリュームに関するデータが混在しないように制御する必要がある。何故なら、一つのストライプ列に複数のボリュームが混在すると、パリティ生成等の処理が複雑化し、データ入出力時のオーバーヘッドが増大して、記憶制御装置の性能が低下するためである。

【０２０９】

そこで、図３０（ａ）に示すように、各ページの先頭をストライプ列の先頭に一致させるようにして、データを記憶する方法が考えられる。この方法によれば、横一列のストライプ内に、異なるボリュームのデータが混在するという事態は生じない。

【０２１０】
しかし、本実施例の記憶制御装置１０は、ＲＡＩＤグループ９０を構成する記憶装置４０の台数を自由に設定することができるため、ページサイズとストライプサイズとは必ずしも一致しない。ページサイズとストライプサイズとが不一致の場合において、ページの先頭をストライプ列の先頭に一致させるようにしてデータを配置すると、図３０（ａ）に空白領域として示すように無駄な領域が発生する。従って、図３０（ａ）に示す方法では、ＲＡＩＤグループ９０の有する実記憶領域を有効に利用することができず、その利用効率が低いという問題がある。
【０２１１】
そこで、本発明では、図３０（ｂ）に示すように、実記憶領域をページ単位で仮想ボリュームに割り当てるのではなく、複数ページを有するチャンク９１単位で実記憶領域を仮想ボリュームに割り当てる。そして、上述のように、仮想ボリューム５０Ｖに関するライトコマンドを受領するたびに、チャンク９１内のページ９２を連続的に使用する。チャンク９１内の各ページ９２は、同一の仮想ボリューム５０Ｖに対応付けられる。異なる仮想ボリュームに関するページ９２が同一のチャンク９１内に混在することはない。このように、本実施例によれば、ＲＡＩＤグループ９０の有する実記憶領域を効率的に使用することができる。
【０２１２】
本実施例では、上述の通り、仮想ボリューム５０Ｖの識別番号と通常ボリューム５０Ｎの識別番号とを特に区別することなく、各ボリューム５０Ｖ，５０Ｎを連続番号で管理している。また、本実施例では、仮想ボリューム５０Ｖのために使用されるＲＡＩＤグループ９０と、通常ボリューム５０Ｎが設けられるＲＡＩＤグループ９０とを特に区別せずに、連続番号で管理している。従って、本実施例の記憶制御装置１０は、仮想ボリューム５０Ｖと通常ボリューム５０Ｎとを比較的簡易な制御構造で共通に管理し、両方のボリューム５０Ｖ，５０Ｎを混在させることができる。
【０２１３】
本実施例では、複数のＲＡＩＤグループ９０からチャンク９１を順番に選択して、仮想ボリューム５０Ｖに割り当てる。従って、プール部６０内の各ＲＡＩＤグループ９０の負荷を均等にすることができる。
【０２１４】
本実施例では、図７，図１７で述べたように、仮想ボリューム５０Ｖを作成するときに各テーブル１５１－１５５間を関連づけ、仮想ボリューム５０Ｖ内の全ての仮想的記憶領域５００を、初期設定用のページ９２に割り当てる。
【０２１５】
従って、本実施例では、仮想ボリューム５０Ｖを正常に作成できた時点で、各仮想的記憶領域５００と実ページ９２との対応付けが正常に行われることを確認できる。つまり、本実施例では、ライトコマンドを受領する前に、仮想ボリューム５０Ｖへのチャンク９１及び初期設定用ページ９２の仮の割り当てが完了している。これにより、仮想ボリューム５０Ｖが正常に動作するか否かを、ライトコマンド受領前に事前に確認することができ、信頼性及び使い勝手が向上する。
【０２１６】
さらに、本実施例では、ライトコマンドを受領した場合に、ライトコマンドで指定される論理アドレスに対応する仮想的記憶領域５００の対応付け先を、仮割当てされた初期設定用ページ９２から、所定チャンク９１内の所定ページ９２に切り替えるだけで済む。これにより、比較的速やかにライトコマンドを処理することができ、記憶制御装置１０の応答性能を高めることができる。
【実施例２】
【０２１７】
図３１，図３２に基づいて第２実施例を説明する。本実施例は、第１実施例の変形例に該当する。従って、第１実施例との相違を中心に説明する。本実施例は、データ移動計画

を作成する場合に、ヌルデータのみが記憶されている実ページ９２を仮想ボリューム５０Ｖから開放し、未使用ページに戻す。

【０２１８】

図３１は、本実施例による移動先決定処理を示すフローチャートである。本フローチャートは、図２７に示すフローチャートに比べて、ゼロデータ（ヌルデータ）を削除するための処理（Ｓ２１０）をさらに備える。移動先決定処理では、最初に、ゼロデータのみを記憶する実ページ９２を未使用ページに戻し、その後で、負荷分散処理及び使用容量平均化処理を実行する。

【０２１９】

図３２は、ゼロデータ削除処理のフローチャートである。コントローラ３０は、各ＲＡＩＤグループ毎に、以下の処理を行う（Ｓ２１０１）。コントローラ３０は、処理対象のＲＡＩＤグループ９０の各ページ９２のうち、ゼロデータのみが記憶されているページ９２を検出する（Ｓ２１０２）。

【０２２０】

コントローラ３０は、ゼロデータのみを記憶するページ９２に対応する仮想ページ５００を、図７で述べた初期化用の特定ページに対応付け（Ｓ２１０３）、ページ管理テーブル１６０を更新させる（Ｓ２１０４）。

【０２２１】

上述の構成を有する本実施例も第１実施例と同様の効果を奏する。さらに、本実施例では、データ移動計画を作成する場合に、ヌルデータのみが記憶されている実ページ９２を開放して、未使用の実ページ９２に戻すため、プール部６０内の実記憶領域を有効に利用できる。つまり、本実施例では、データ移動計画を作成するたびに、プール部６０内のページ９２のうち無駄に使用されているページ９２を開放できる。

【０２２２】

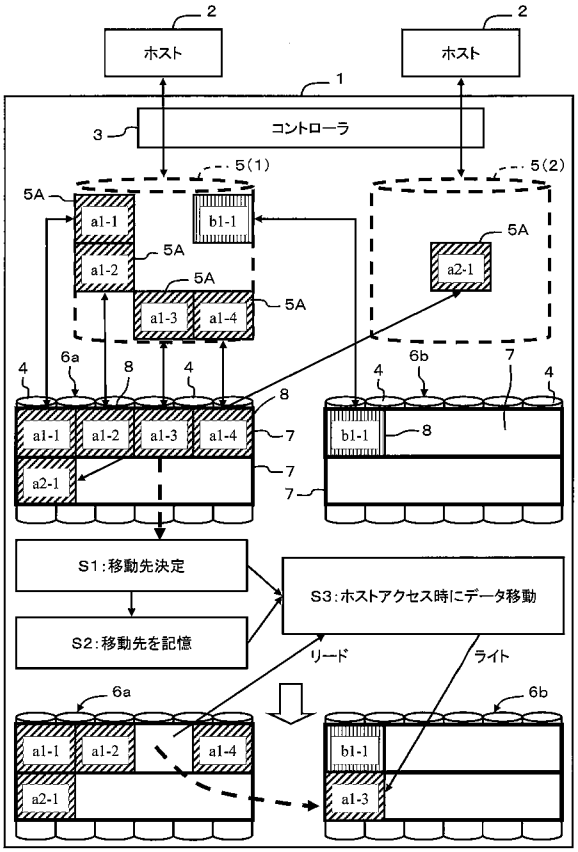
なお、本発明は、上述した実施形態に限定されない。当業者であれば、例えば、上記各実施例を適宜組み合わせる等のように、本発明の範囲内で、種々の追加や変更等を行うことができる。

【符号の説明】

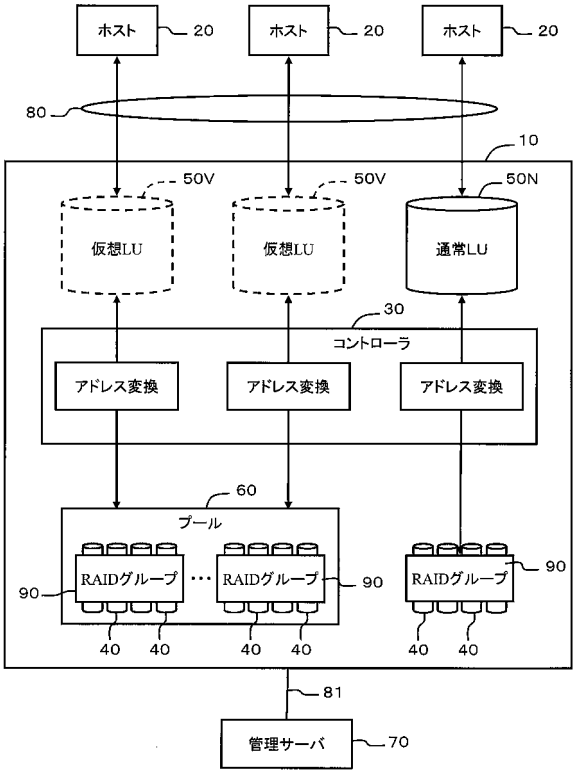
【０２２３】

１：記憶制御装置、２：ホスト、３：コントローラ、４：記憶装置、５（１），５（２）：仮想ボリューム、５Ａ：仮想的記憶領域、６ａ，６ｂ：ＲＡＩＤグループ、７：チャンク、８：ページ、１０記憶制御装置、２０：ホスト、３０：コントローラ、４０：記憶装置、５０Ｖ：仮想ボリューム、６０：プール部、７０：管理サーバ、９０：ＲＡＩＤグループ、９１：チャンク、９２：ページ、５００：仮想ページ。

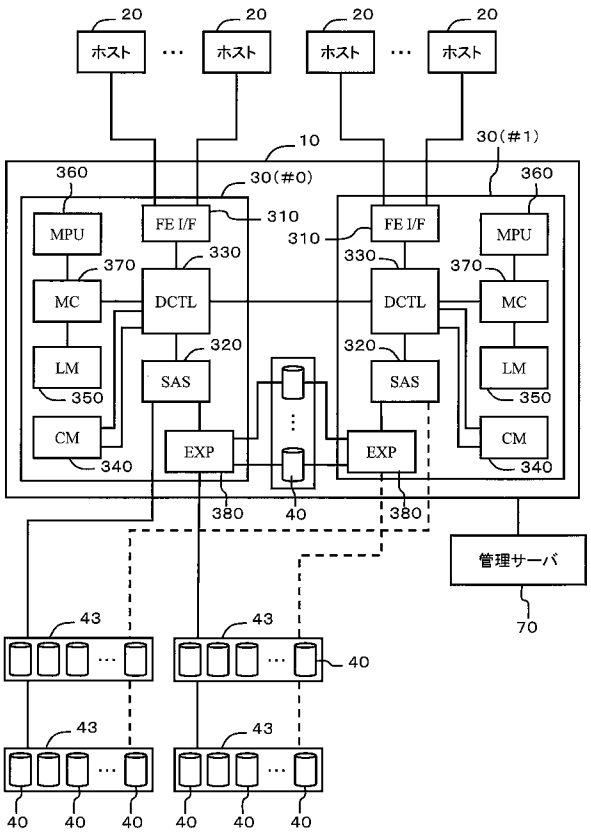
【図１】


ホスト
ホスト
コントローラ
5(1)
5(2)
a1-1
a1-2
a1-3
a1-4
b1-1
a2-1
S1:移動先決定
S2:移動先を記憶
S3:ホストアクセス時にデータ移動
リード
ライト

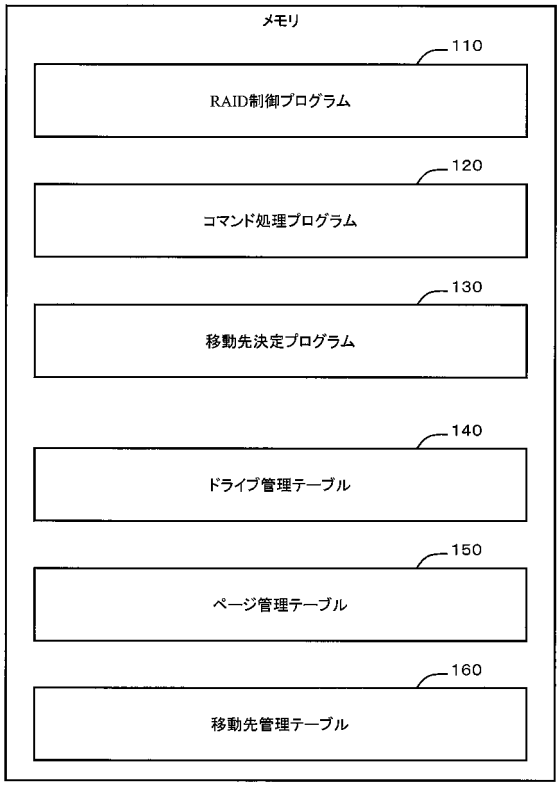
【図２】


ホスト
ホスト
ホスト
80
10
50V
50V
50N
仮想LU
仮想LU
通常LU
30
コントローラ
アドレス変換
アドレス変換
アドレス変換
60
プール
RAIDグループ
RAIDグループ
RAIDグループ
90
40
81
管理サーバ
70

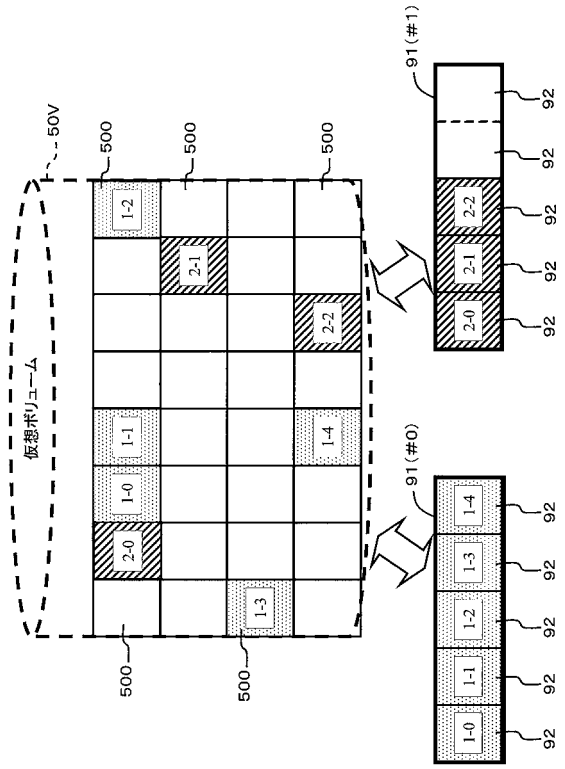
【図３】


ホスト
10
30(#0)
30(#1)
MPU
FE I/F
DCTL
MC
LM
SAS
CM
EXP
310
320
330
340
350
360
370
380
40
管理サーバ
70
43

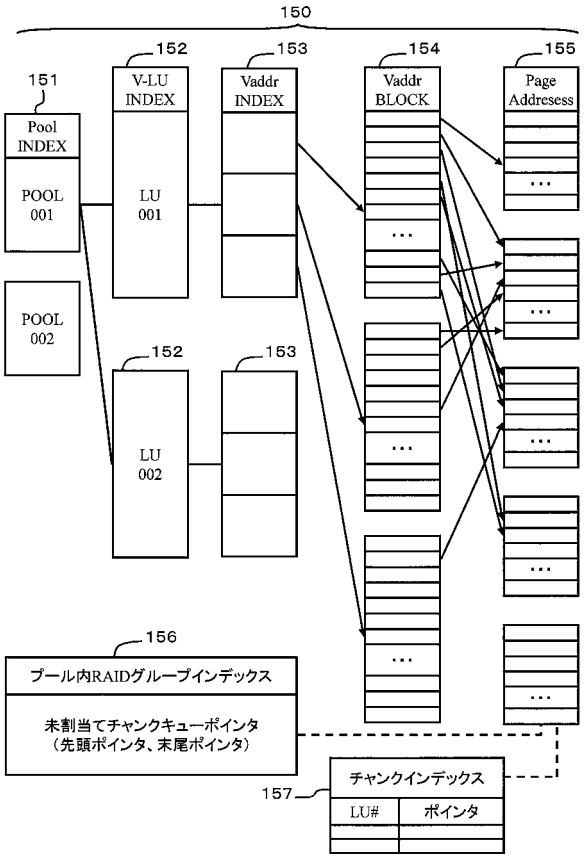
【図４】


メモリ
RAID制御プログラム
110
コマンド処理プログラム
120
移動先決定プログラム
130
ドライブ管理テーブル
140
ページ管理テーブル
150
移動先管理テーブル
160

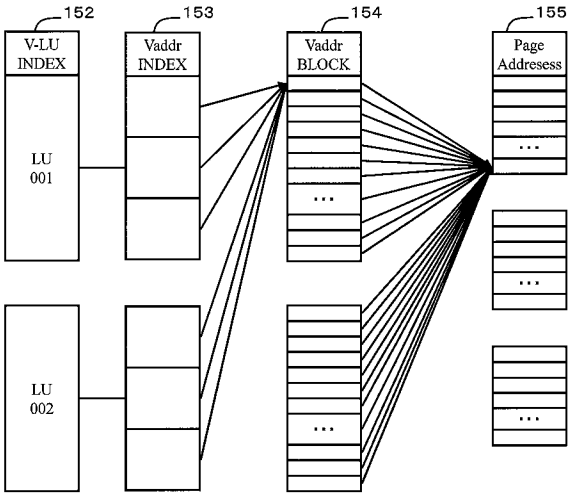
【図５】


50V
500
500
500
仮想ボリューム
1-2
2-1
2-2
1-1
1-4
1-0
2-0
1-3
91(#1)
91(#0)
92
2-2
2-1
2-0
1-4
1-3
1-2
1-1
1-0

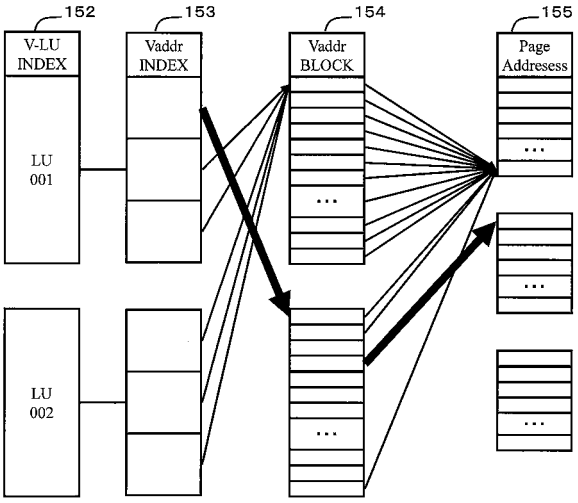
【図６】


150
151
152
153
154
155
Pool INDEX
V-LU INDEX
Vaddr INDEX
Vaddr BLOCK
Page Addresess
POOL 001
POOL 002
LU 001
LU 002
156
プール内RAIDグループインデックス
未割当てチャンクキューポインタ
（先頭ポインタ、末尾ポインタ）
157
チャンクインデックス
LU#
ポインタ

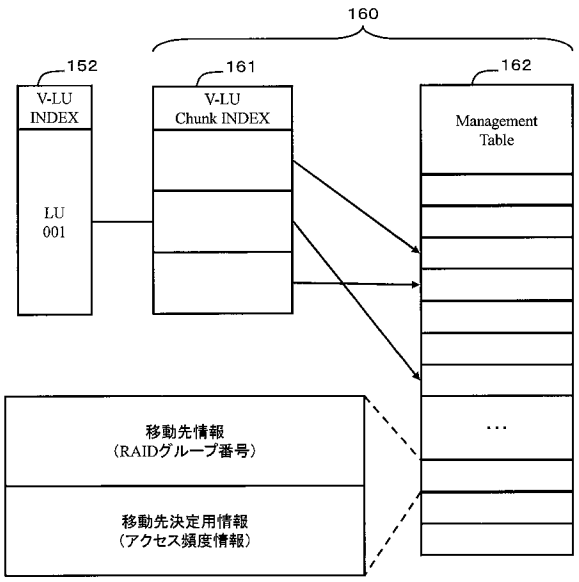
【図７】


152
153
154
155
V-LU INDEX
Vaddr INDEX
Vaddr BLOCK
Page Addresess
LU 001
LU 002

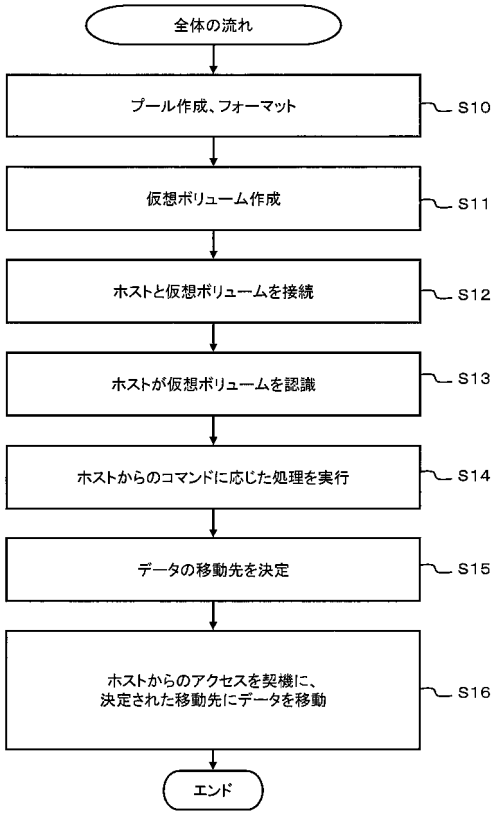
【図８】


152
153
154
155
V-LU INDEX
Vaddr INDEX
Vaddr BLOCK
Page Addresess
LU 001
LU 002

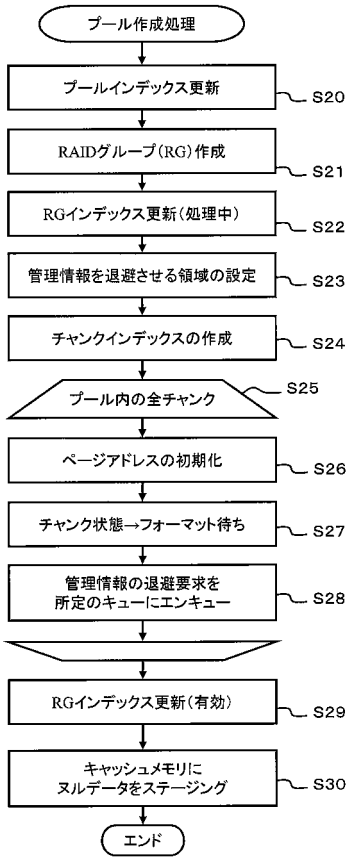
【図９】



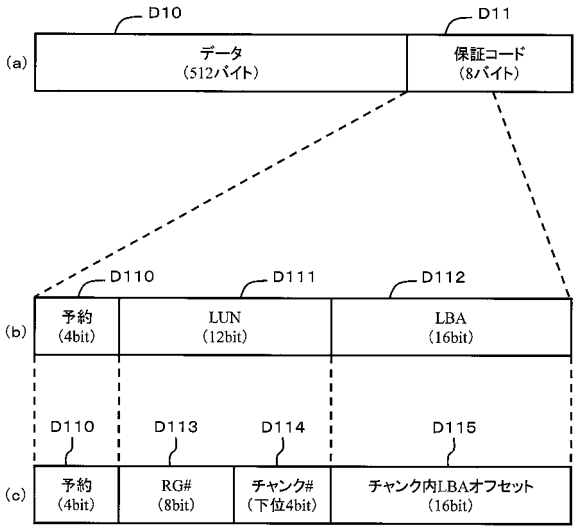
【図１０】



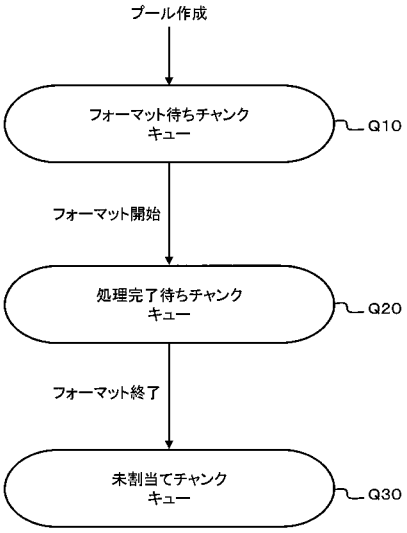
【図１１】



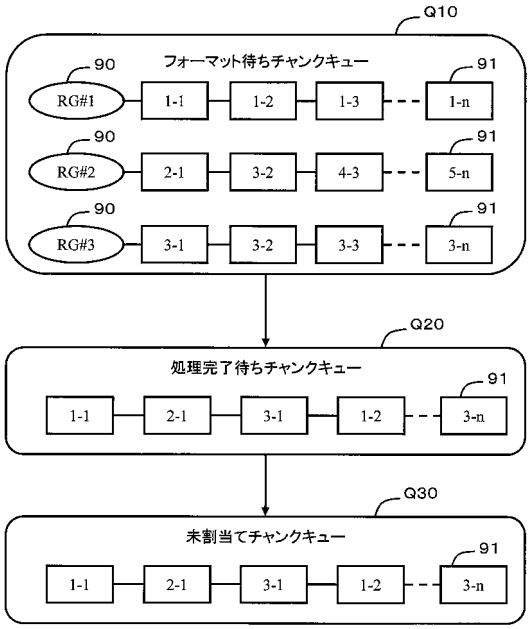
【図１２】



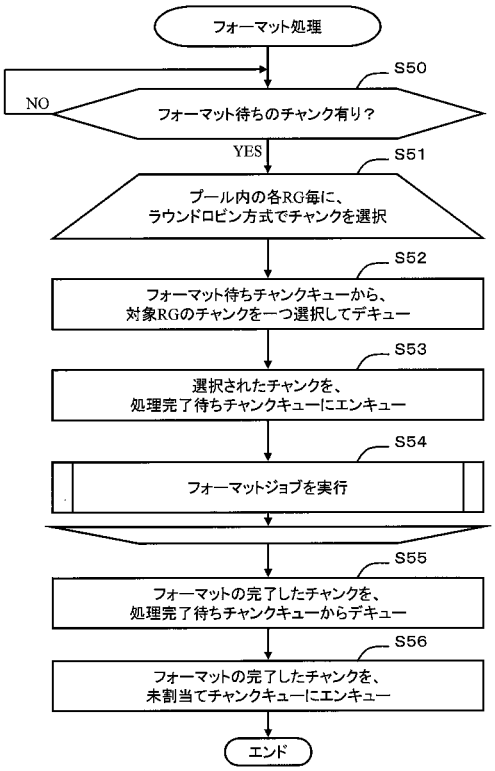
【図１３】



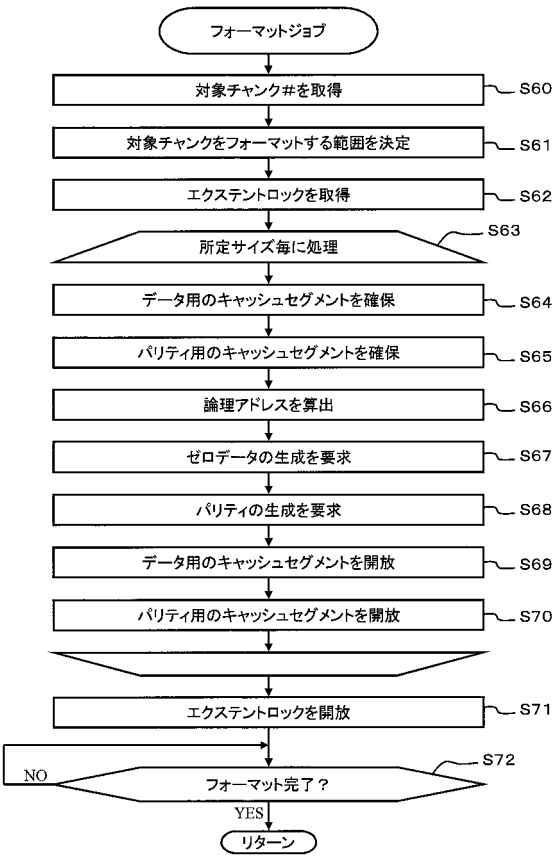
【図１４】



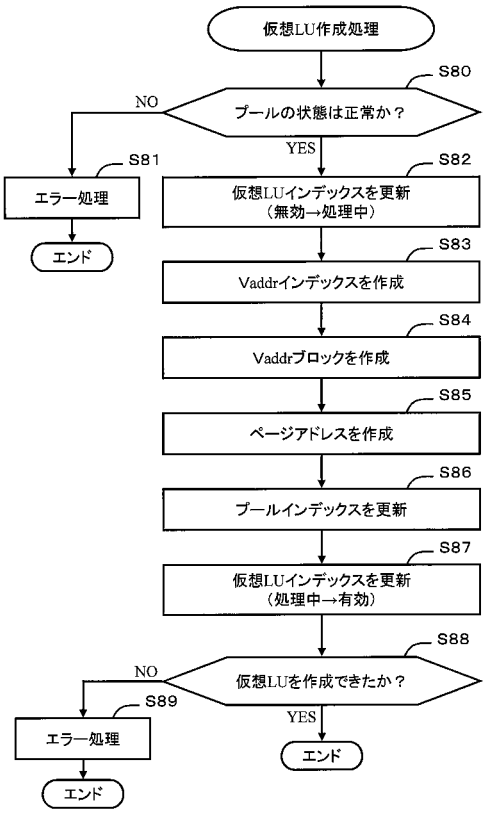
【図１５】



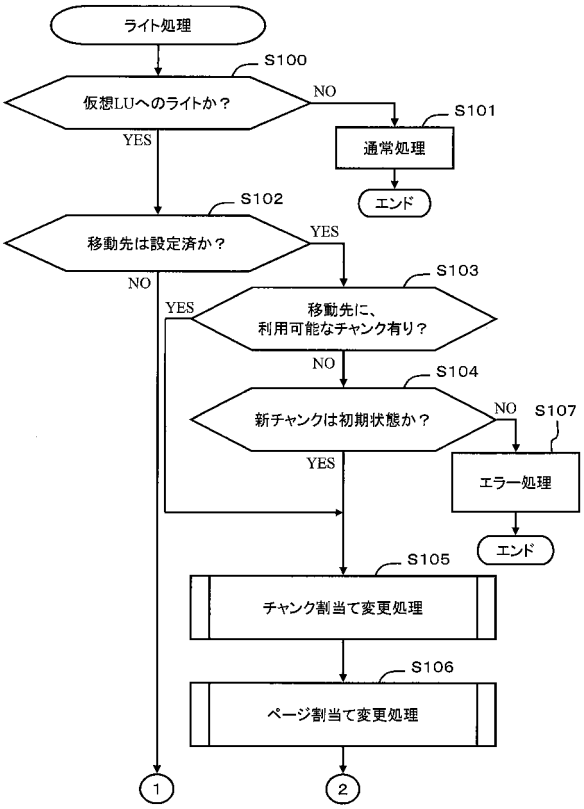
【図１６】



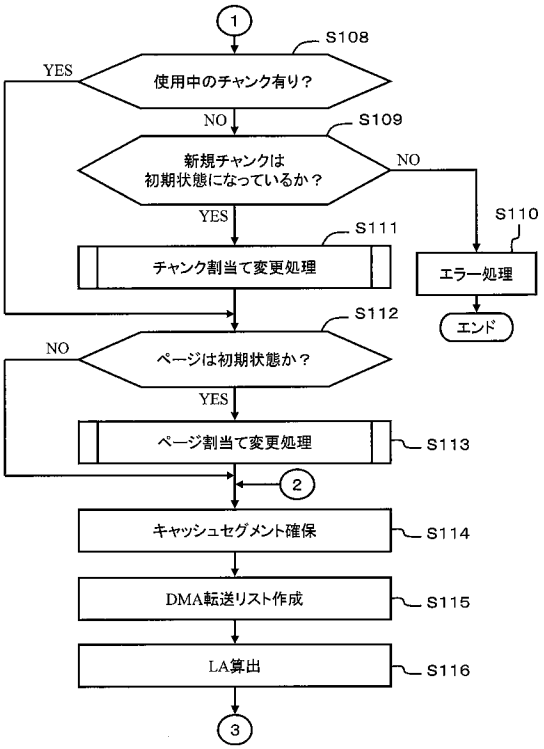
【図１７】



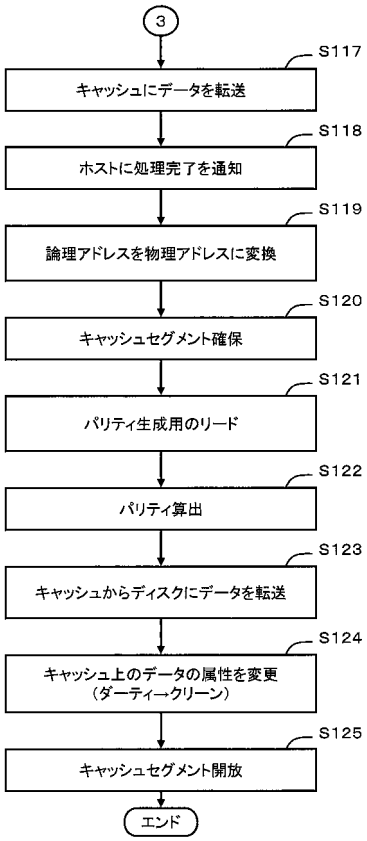
【図１８】



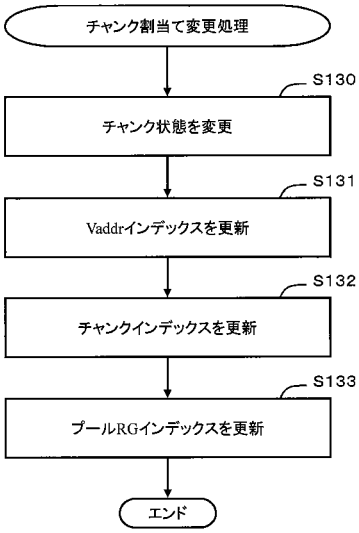
【図１９】



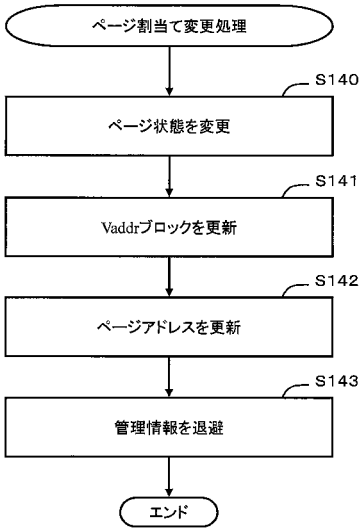
【図２０】



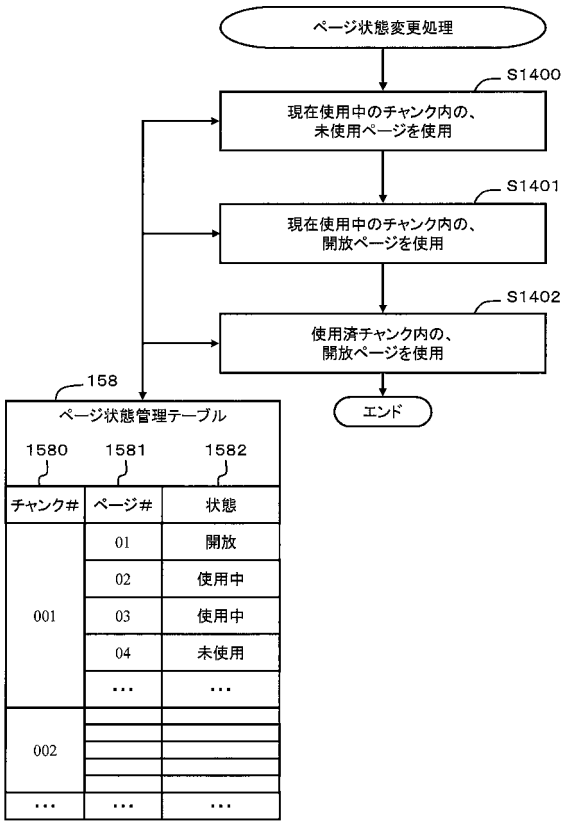
【図２１】



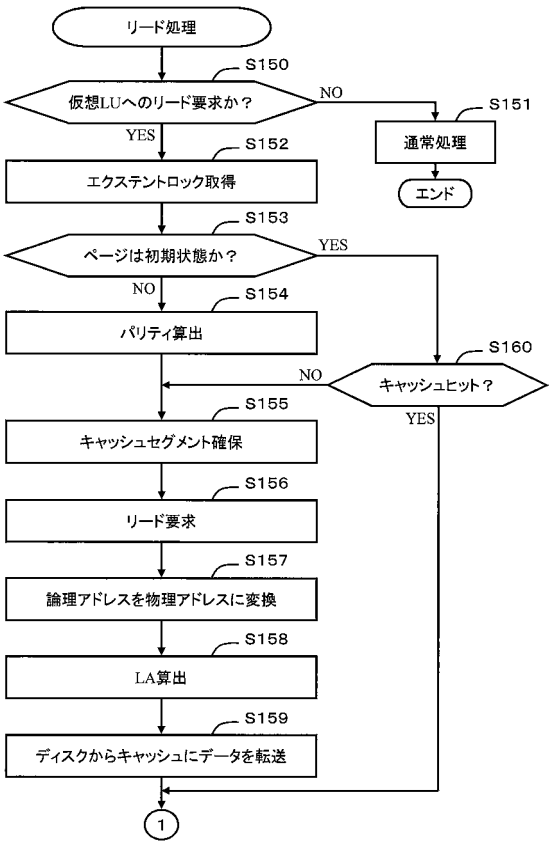
【図２２】



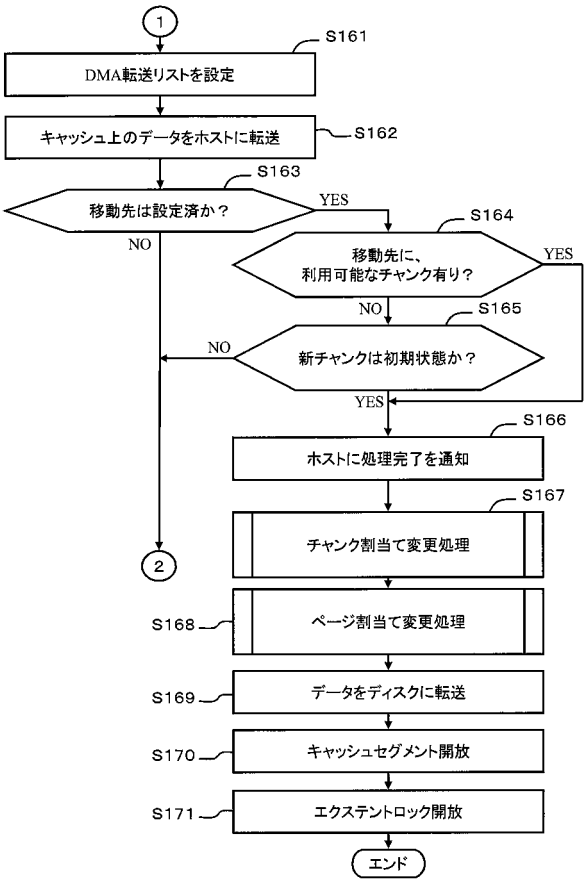
【図２３】



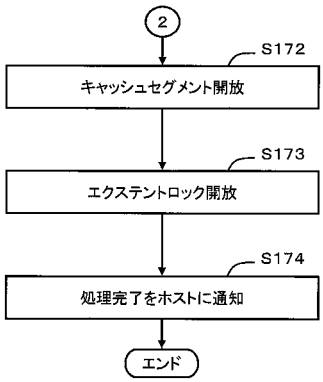
【図２４】



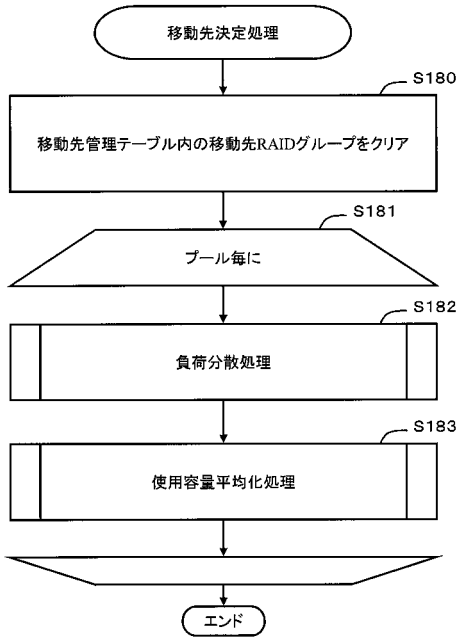
【図２５】



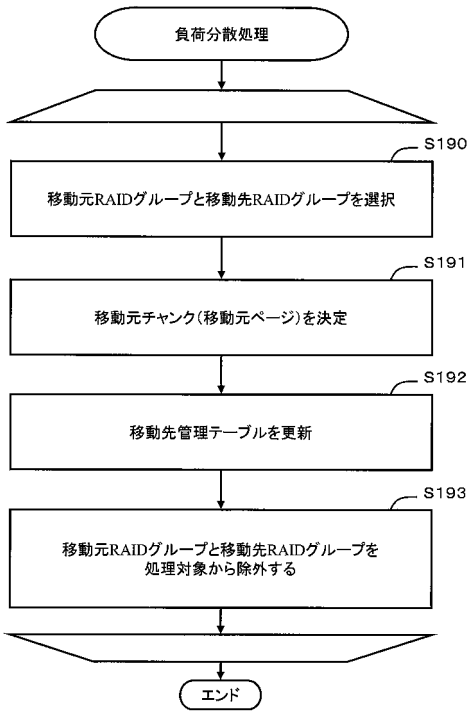
【図２６】



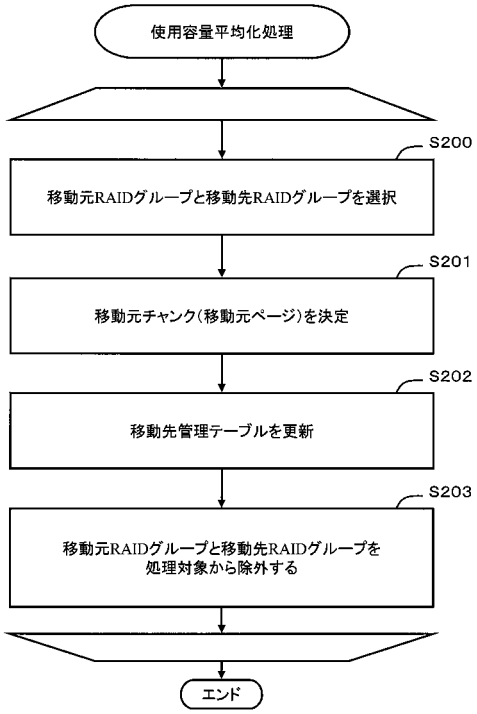
【図２７】



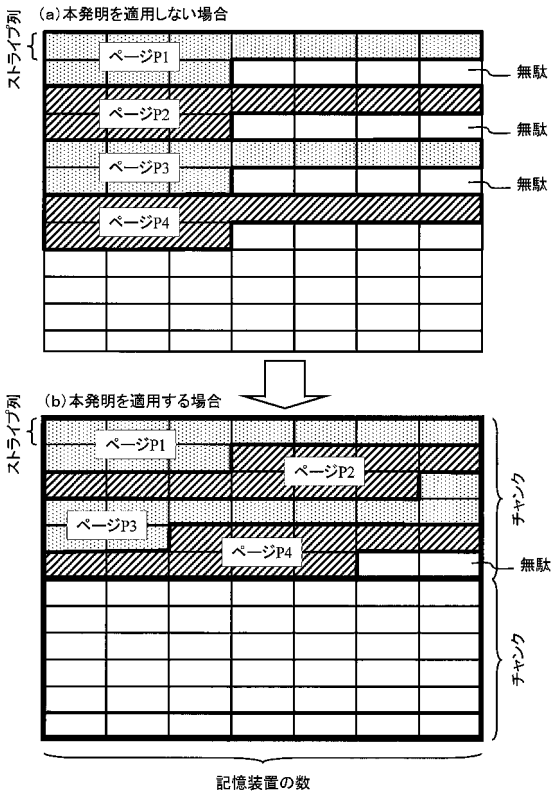
【図２８】



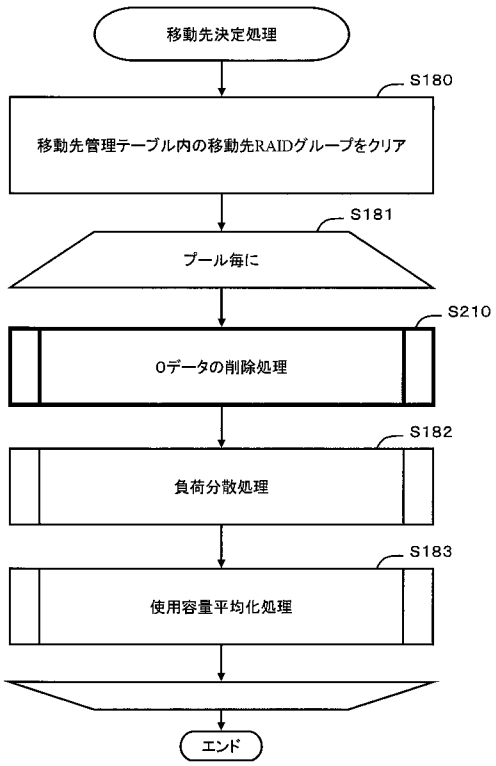
【図２９】



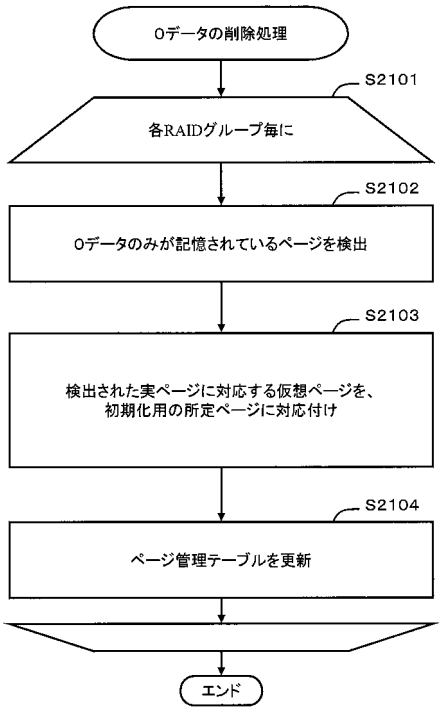
【図３０】



【図３１】



【図３２】



フロントページの続き

(72)発明者　内海　勝広
神奈川県小田原市中里３２２番２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

審査官　木村　貴俊

(56)参考文献　特開２００８－０５９３５３（ＪＰ，Ａ）
特開２００８－０９０７４１（ＪＰ，Ａ）
特開２００８－０４６９８６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　３／０６－　３／０８
Ｇ０６Ｆ　１２／００－１２／１６
Ｇ０６Ｆ　１３／００－１３／４２