

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3699165号
(P3699165)

(45) 発行日 平成17年9月28日(2005.9.28)

(24) 登録日 平成17年7月15日(2005.7.15)

(51) Int.Cl.⁷G06F 3/06
G06F 12/16

F I

G06F 3/06 540
G06F 12/16 320L

請求項の数 4 (全 24 頁)

(21) 出願番号	特願平7-207970	(73) 特許権者	398038580
(22) 出願日	平成7年8月15日(1995.8.15)		ヒューレット・パッカード・カンパニー
(65) 公開番号	特開平8-115173		HEWLETT-PACKARD COMPANY
(43) 公開日	平成8年5月7日(1996.5.7)		アメリカ合衆国カリフォルニア州パロアルト
審査請求日	平成14年2月6日(2002.2.6)		ハノーバー・ストリート 3000
(31) 優先権主張番号	319, 385	(74) 代理人	100081721
(32) 優先日	平成6年10月5日(1994.10.5)		弁理士 岡田 次生
(33) 優先権主張国	米国 (US)	(72) 発明者	マイケル・ブレント・ジェイコブソン
			アメリカ合衆国83714アイダホ州ボイジー、オースティン 8884
		(72) 発明者	ダグラス・エル・ヴォイト
			アメリカ合衆国83702アイダホ州ボイジー、エヌ・ツウェンティーフォース 3030
			最終頁に続く

(54) 【発明の名称】 データ記憶装置の記憶容量を拡大する方法

(57) 【特許請求の範囲】

【請求項1】

物理的記憶空間を定義する記憶ディスクを複数持つ階層ディスクアレーを備えたデータ記憶装置の記憶容量を拡大する方法であって、

(a) 前記ディスクアレーのすべてのディスクにわたって延び、前記ディスクアレーの個々の記憶ディスクからの記憶空間の1つまたは複数のセグメントを含み、第1のRAIDレベルにしたがってデータを記憶する第1RAID区域、および第2のRAIDレベルにしたがってデータを記憶する第2RAID区域を含む複数のRAID区域として前記物理的記憶空間を表す第1の仮想記憶空間に、前記物理的記憶空間をマッピングするステップと、

(b) 記憶空間を複数の仮想ブロックとして表す第2の仮想記憶空間に、前記第1の仮想記憶空間をマッピングするステップと、

(c) 前記階層ディスクアレーに少なくとも1つの追加記憶ディスクを加えて、前記データ記憶装置の物理的記憶空間を拡大するステップと、

(d) 前記第2の仮想記憶空間の使用可能な容量を修正して、前記新たに追加された記憶ディスクを加えることによって生じる新しい記憶空間を反映させるステップと、

(e) ステップ(a)でRAID区域にマッピングされていなかった未使用記憶空間を拡大して前記追加記憶ディスク上の空間を含むようにするステップと、

(f) RAID区域を選択するステップと、

(g) 選択したRAID区域から前記階層ディスクアレーの別の場所にデータを移すステ

ップと、

(h) 前記第2の仮想記憶空間から前記第1の仮想記憶空間へのマッピングを更新して、前記階層ディスクアレー内のデータ移動を反映させるステップと、

(i) 前記すべての記憶ディスクと前記新たに追加された記憶ディスクにわたり、前記個々の記憶ディスクおよび前記新たに追加された記憶ディスクからの1つまたは複数のセグメントを含む少なくとも1つの拡大されたRAID区域に、前記選択したRAID区域に対応する物理的記憶空間をマッピングしなおすステップと、

(j) 前記階層ディスクアレーのすべてのRAID区域が前記新たに追加された記憶ディスクのセグメントを含むように拡大されるまで、前記ステップ(f)～(i)を繰り返すステップと、

を有するデータ記憶装置の記憶容量を拡大する方法。

【請求項2】

前記選択したRAID区域内のデータへのアクセスとそのデータ移動との間で起こる衝突を解決するステップをさらに有する請求項1に記載のデータ記憶装置の記憶容量を拡大する方法。

【請求項3】

RAID区域に記憶されているのと同じデータ量を保持するのに十分な大きさがある少なくとも1つの区域を確保するステップをさらに有する請求項1に記載のデータ記憶装置の記憶容量を拡大する方法。

【請求項4】

未使用の記憶空間を探すステップと、

前記未使用の記憶空間を少なくとも1つの拡大されたRAID区域にマッピングするステップと、

をさらに有する請求項1に記載のデータ記憶装置の記憶容量を拡大する方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、階層ディスクアレー・データ記憶装置のようなデータ記憶装置、および、そのような装置にディスクを追加する方法に関する。

【0002】

【従来の技術】

コンピュータ・システムは、速度、信頼性、および処理能力の点で絶え間なく改良が続いている。その結果、コンピュータはより複雑で高度なアプリケーションが処理できる。コンピュータが改良されるのに応じて、大容量記憶装置および入出力装置に対する性能要求が増大している。発展するコンピュータ・システムに性能面で歩調がとれる大容量記憶装置を設計することが常に求められている。

【0003】

本発明は、ディスクアレー型の大容量記憶装置に関する。ディスクアレー・データ記憶装置は、単一の大容量記憶装置を形成するように配列され調整された複数のディスク駆動装置を持つ。大容量記憶装置には、コスト、性能、可用性の3つの主要な設計基準がある。メガバイトあたりの低いコスト、高い入出力性能、および、高いデータ可用性を持つ記憶装置を作ることが望ましい。"可用性"(availability)とは、記憶装置に記憶されているデータをアクセスする能力、および、ある程度の故障が起きても動作が続けられることを保証できる能力のことである。典型的には、データの可用性は、複数の場所に記憶されたデータまたはデータ間の関係を持つ冗長度の使用を通して提供される。

【0004】

冗長データを記憶するには2つの一般的な方法がある。第1の方法すなわち"ミラー"(鏡像: mirror)方法では、データが複製され、記憶装置の2つの別々の区域に記憶される。たとえば、ディスクアレーでは、同一のデータがディスクアレーの2つの別々のディスクの上に提供される。ミラー方法は、二重に記憶する手法により、高性能および高データ可

10

20

30

40

50

用性の利点がある。しかし、ミラー方法は、データを記憶するコストが実質的に2倍になるので、相対的に費用の高いものになる。

【0005】

第2の方法すなわち"パリティ" (奇偶: parity) 方法では、記憶区域の一部を冗長データを記憶するのに使うが、冗長データを記憶する区域は、元のデータを記憶するのに使う記憶空間より小さい。たとえば、5台のディスクを持つディスクアレーでは、4台をデータを記憶するのに使い、5台目をもっぱら冗長データを記憶するのに使う。パリティ方法の利点は、ミラー方法より低費用であることであるが、ミラー方法にくらべると、性能および可用性が低いという特徴がある。

【0006】

ディスクアレー・データ記憶装置の設計で遭遇する1つの問題は、ディスクアレーの記憶容量を増やすためにディスク駆動装置を足す仕事に関係する。従来のディスクアレーでは、次のような長々しい処理が必要である。先ず、いまディスクアレー上に記憶されているデータを (リールからリールのテープ装置のような) 全く別の記憶装置にコピーする。次に、新しい記憶ディスクをディスクアレーに足す。次に、新しい記憶ディスクを組み込むためにディスクアレー全体を再構成する。最後に、データを、別の記憶装置から新しく構成されたディスクアレーにコピーして戻す。この処理の間、ディスクアレーにアクセスすることができず、別のバックアップ装置に一時的に移されたデータは利用することができない。

【0007】

【発明が解決しようとする課題】

本発明の目的は、データの可用性を維持しながら、記憶ディスクを追加することによって、データ記憶装置の記憶容量を拡大する方法を提供することである。

【0008】

【課題を解決するための手段】

本発明の1つの方法では、予定のデータ量を記憶するために、ディスクアレーの物理的な記憶空間を複数のストライプ (stripe) に構成する。ストライプはディスクアレーの複数の記憶ディスクにわたって延び、ディスクアレーの記憶ディスクの各々からの1つまたは複数の同じ大きさのセグメントで構成される。記憶ディスクが足されるとき、データが1つのストライプから、物理的記憶空間のもう1つの部分に移される。次に、ストライプを含む物理的記憶空間が、データを記憶するために拡大されたストライプに再構成される。この拡大されたストライプは、新たに加えられた記憶ディスクを含めた全ての記憶ディスクにわたって延びる。これにより、拡大されたストライプが、新しいデータを受け取る準備が整う。この処理は、全てのストライプが新しいディスクを含むように構成されるまで、ストライプごとに続けられる。ストライプが拡張されている間、データ記憶装置は選択されたストライプ内のどの仮想ブロックも割り当てない。

【0009】

データ記憶装置の記憶容量を拡大するもう1つの方法では、物理的記憶空間のストライプ構成のプリマップ (pre-map) が先ず作られる。1台または複数の記憶ディスクがディスクアレーに加えられると、物理的記憶空間の新しい構成を表すポストマップ (post-map) が作られる。このポストマップが追加された記憶ディスクを考慮に入れる。次に、グループAのセグメントが、個々の記憶ディスクからのセグメントおよび新しく追加された記憶ディスクからのセグメントを持つ、少なくとも1つの拡大されたストライプを含むように定義される。グループBのセグメントは、グループAに入っていないセグメントを含むように定義される。

【0010】

グループAのセグメントの中のデータは、プリマップにしたがった第1のデータ配列からポストマップにしたがった第2のデータ配列に再編成される。この再編成の後、グループAのセグメントは、少なくとも1台のディスクを加えたことで増えた記憶区域によって、少なくとも1つの空のセグメントを持っている。したがって、グループAの空のセグメン

10

20

30

40

50

トを満たすために、データが、グループBの少なくとも1つのセグメントから移動される。この処理は、全てのストライプが新しいディスクを含むように構成されるまで、ステップごとに続けられる。この処理中に起こる、ユーザおよび移行中のデータ移動によるグループAおよびBへのデータアクセスに関する競合は解決され、データは常に使用可能である。

【0011】

本発明は、また、冗長度の異なるレベル、または、RAID (Redundant Array of Independent Disks) レベルにしたがってデータを記憶する階層ディスクアレー・データ記憶装置に使用できるように調整することができる。この方法は、まず、物理的記憶空間を、複数のRAID区域を持つ第1の仮想記憶空間にマッピングすることを含む。各RAID区域は、第1または第2のRAID区域の型(タイプ)として構成できる。第1のRAID区域型は、第1のRAIDレベル(たとえば、RAIDレベル1すなわちミラー冗長性)にしたがってデータを保持し、第2のRAID区域型は、第2のRAIDレベル(たとえば、RAIDレベル5すなわちパリティ冗長性)にしたがってデータを保持する。RAID区域はディスクアレーの複数の記憶ディスクにわたって延び、各記憶ディスクからの同じ大きさの記憶領域を含む。次に、第1の仮想記憶空間は複数の仮想ブロックを持つ第2の仮想記憶空間にマッピングされる。この第2の仮想記憶空間は、ユーザまたはホストによって定義され、ユーザまたはホストに提示される記憶の見え方(ビュー:view)である。

【0012】

この場合、1台または複数の記憶ディスクが階層ディスクアレーに追加されると、選択されたRAID区域に記憶されたデータは、該RAID区域以上の記憶容量を持つ別の区域に移される。第1の仮想記憶空間から第2の仮想記憶空間へのマッピングは、このデータ移動を反映するように更新される。次に、選択されたRAID区域に対応する物理的記憶空間は、新しいディスクを含む全ての記憶ディスクにわたってひろがる拡大されたRAID区域にマッピングしなおされる。この処理は、階層ディスクアレーの全てのRAID区域が新しい記憶ディスク上の領域を含むように拡大されるまで、繰り返される。第2仮想記憶空間の第1仮想記憶空間へのマッピングは、仮想記憶空間内のRAID区域の拡大を反映するように絶えず修正される。

【0013】

新しいディスクを同化するこの容量拡大処理の間、データは使用可能である。データはディスクアレー上に存在しアクセス可能であり続け、別の記憶装置に移されない。移動の結果としてマッピングしなおす処理の間のデータはアクセスできなくなるが、これは束の間のことである。したがって、本発明の方法は、データがユーザに常に使用可能であることを実質的に確実にする。

【0014】

【発明の実施の形態】

図1に、本発明にしたがって構成されたデータ記憶装置10を示す。データ記憶装置10は、複数の記憶ディスク12を持つ階層ディスクアレー11、記憶ディスク12への、および記憶ディスク12からのデータ転送を調整するためにディスクアレー11に結合されたディスクアレー制御装置14、およびRAID管理装置16を含む、ディスクアレー・データ記憶装置であることが望ましい。

【0015】

本明細書の説明の目的のために、"ディスク"とは、不揮発性で、ランダム・アクセスが可能で、自身の記憶故障を検出する能力のある、書き換え可能な大容量記憶装置を指す。それは、回転する磁気および光ディスク、並びに、固体ディスク、または不揮発性の電子的記憶要素、たとえばPROM、EPROM、およびEEPROMを含む。"ディスクアレー"という用語は、複数のディスク、ディスクを1つまたは複数のホスト・コンピュータに接続するのに必要なハードウェア、および物理的ディスクの動作を制御し、物理的ディスクをホストの動作環境に1つまたは複数の仮想ディスクとして提示するのに使われる管

10

20

30

40

50

理ソフトウェアの集まりを指す。"仮想ディスク"とは、管理ソフトウェアによってディスクアレーの中に実現される抽象的実体である。

【0016】

"RAID" (Random Array of Independent Disks) という用語は、記憶容量の残りの部分に記憶されたユーザデータについての冗長情報を記憶するために、その物理的記憶容量の1部が使用されているディスクアレーを意味する。冗長情報は、ディスクアレーを構成するディスクの1つ、または、そのディスクへのアクセス経路が故障したときにユーザデータの再生を可能にする。RAID装置に関する詳しい記述は、「RAID手法の資料集」(The RAIDBook: A Source Book for RAID Technology, published June 6, 1993, by the RAID Advisory Board, Lino Lakes, Minnesota) を参照されたい。

10

【0017】

ディスクアレー制御装置14は、たとえばSCSI (small computer system interface) のような1つまたは複数のインターフェース・バス13を介してディスクアレー11に結合されている。RAID管理装置16は、インターフェース・プロトコル15を介して、ディスクアレー制御装置14を動作させるために結合されている。データ記憶装置10はホスト・コンピュータ(図示せず)にも入出力インターフェース・バス17を介して結合されている。RAID管理装置16は、ディスクの記憶および信頼性レベルを制御し、多様な信頼性記憶レベル間のデータ転送を行うためのデータマネージャ手段を提供するために、別個の装置要素として具体化してもよいし、ディスクアレー制御装置14内またはホスト・コンピュータ内に構成してもよい。これらの信頼性記憶レベルは、以下に述べるように、ミラーまたはパリティ冗長レベルであることが望ましいが、全く冗長性を持たない信頼性記憶レベルを含むこともできる。

20

【0018】

ディスクアレー制御装置14は、ディスクアレー制御装置A 14a、およびディスクアレー制御装置B 14bからなる二重の制御装置として具体化することが望ましい。二重制御装置14aおよび14bは、どちらか1つが動作できなくなった場合にも、継続したバックアップおよび冗長性を提供することによって信頼性を向上させる。しかし、本発明による方法は、1台の制御装置またはその他の構成でも実施することができる。

【0019】

階層ディスクアレー11は、物理的記憶空間および1つまたは複数の仮想記憶空間を含む異なる記憶空間として特徴づけられる。記憶に関するこれらの多様なビューはマッピング手法に関連している。たとえば、ディスクアレーの物理的記憶空間は、多様なデータ信頼性レベルにしたがって記憶区域を描く仮想記憶空間にマッピングすることができる。仮想記憶空間内のある区域を第1信頼性記憶レベル(たとえば、ミラーすなわちRAIDレベル1)のために割り当て、その他の区域を第2信頼性記憶レベル(たとえば、パリティすなわちRAIDレベル5)のために割り当てることができる。RAIDレベルに関しての多様なマッピング手法および仮想空間については以下に詳しく述べる。

30

【0020】

データ記憶装置10は、異なる記憶空間を互いにマッピングするために用いる仮想マッピング情報を持続的に記憶するメモリマップ記憶装置21を含む。メモリマップ記憶装置はディスクアレーの外にあり、ディスクアレー制御装置14内にあることが望ましい。メモリマップ情報は、異なるビューの中で多様なマッピング構成が変わるにしたいが、制御装置またはRAID管理装置によって絶えずまたは定期的に更新される。

40

【0021】

メモリマップ記憶装置21は、それぞれ、制御装置14aおよび14bに置かれる2つの不揮発性RAM (Random Access Memory) 21aおよび21bとして具体化されるのが望ましい。不揮発性RAM (non-volatile RAM: NVRAM) の1つの例は、バッテリー付きのRAMである。バッテリー付きRAMは、データをメモリに維持するために、データ記憶装置10への電源が切られたときに独立したバッテリー電源からのエネルギーを使う。1つの望ましい構成は、自身でリフレッシュできるバッテリー付きDRAM (Dynamic RAM) である

50

。

【0022】

二重のNVRAM 21 aおよび21 bは、メモリマップ情報の冗長記憶を供給する。仮想マッピング情報はミラー冗長方法にしたがって複製され、NVRAM 21 aおよび21 bの両方に記憶される。このようにして、NVRAM 21 aはもっぱら元のマッピング情報を記憶するのに用い、NVRAM 21 bはもっぱら冗長マッピング情報を記憶するのに用いることができる。別の構成として、データを二重に記憶するための十分な空間を備えた1つの不揮発性RAMを用いて、ミラーにしたメモリマップ記憶装置を構成することもできる。

【0023】

図1に示すように、ディスクアレー11は複数のディスク駆動装置12を持つ。これらの記憶ディスクの大きさの例は、1ないし3ギガバイトである。記憶ディスクは、SCSIバス13とのインターフェースを提供する機械的ベイに、それぞれ独立して接続したりはずしたりできる。1つの具体化では、データ記憶装置は、作動する機械的ベイを12個持つように設計される。これらのベイをディスクアレー制御装置14にインターフェースさせるために4つのSCSIバス（すなわち、3つの機械的ベイに対して1つのバス）が使われる。これらのベイがいっぱいになるまでロードされた場合、データ記憶装置は全体で、たとえば、12ないし36ギガバイトの容量を持つ。ディスクアレー制御装置14は、記憶ディスク12がどのベイに差し込まれていても、認識することができる。データ記憶装置10は、ディスクアレーの動作中に、ディスクアレーの使用可能なベイに追加のディスクを差し込む（"ホットプラグ"：hot plug）ことができるように設計されている。新しいディスク装置の追加については、以下に、図5ないし13を参照して詳しく述べる。

【0024】

本発明の説明の目的として、ディスクアレー11の記憶ディスク12は、複数のディスク20を持つミラー・グループ18、および、複数のディスク24を持つパリティ・グループ22に配列されたものとして概念化することができる。ミラー・グループ18は、第1のレベルすなわちミラー冗長レベルにしたがってデータを記憶するディスクアレーの第1の記憶位置すなわちRAID区域を表す。

このミラー冗長レベルは、またRAIDレベル1と考えることができる。RAIDレベル1、すなわちディスクのミラー化は、データの各ビットが複製されてデータ記憶装置内に記憶されるという意味で1対1の保護を提供することにより、最高の信頼性を提供する。ミラー冗長は、図1で、3対のディスク20で表されている。元のデータは第1の組のディスク26に記憶され、複製された冗長データは、対になった第2の組のディスク28に記憶される。

【0025】

図2に、RAIDレベル1にしたがったデータの記憶をより詳しく示す。垂直の列は個々のディスク0、1、2、3を表す。4個のディスクからなるこのディスクアレーに含まれる物理的記憶空間は、水平の行で表される複数のストライプに構成される。1つの"ストライプ"は複数の記憶ディスクにわたって延び、多数の同じ大きさの記憶空間のセグメントで構成され、1つのセグメントはアレーの中の各ディスクに関連している。すなわち、1つのセグメントは1つのディスクの上に存在する1つのストライプの部分である。各ストライプは、記憶ディスク全体に分配されているデータの予定の量を保持する。ストライプのうちのあるセグメントは元のデータのために使われ、その他のセグメントは冗長データのために使われる。

【0026】

ミラー冗長（RAIDレベル1）のこの例では、ディスク0上のストライプ0のセグメント0に記憶されているデータは、複製されて、ディスク1上のストライプ0のセグメント0'に記憶される。同様に、ディスク2のストライプ2のセグメント5に記憶されているデータは、ディスク3のストライプ2のセグメント5'に記憶される。このようにして、各データは複製されディスク上に記憶される。図2の冗長のレイアウトは説明の目的で図

10

20

30

40

50

示したものである。冗長データを同じストライプに、図示したように、きちんと並べる必要はない。たとえば、ディスク0のストライプ1のセグメント2のデータは複製されて、ディスク3のストライプSのセグメントT'に置くことができる。

【0027】

図1を再び参照すると、ディスク24のパリティ・グループ22は、第2の記憶位置すなわちRAID区域を表し、データは第2の冗長レベル、たとえばRAIDレベル5にしたがって記憶される。図に例として示した6個のディスクでは、元のデータは5個のディスク30に記憶され、冗長"パリティ"データは6番目のディスク32に記憶される。

【0028】

図3にパリティRAID区域のレイアウトを詳しく示す。図2のミラーRAID区域のレイアウトと同様に、ディスク0、1、2、3の物理的記憶空間は、複数の同じ大きさのストライプに構成される。例として示したこの図では、データはRAIDレベル5にしたがって記憶され、冗長データはPで表したセグメントに記憶される。冗長Pセグメントは、ストライプの中のその他のセグメントのパリティを記憶する。たとえば、ストライプ0において、ディスク3上の冗長Pセグメントは、ディスク0、1、2のパリティを記憶する。各ストライプのパリティは、たとえば、排他的OR関数のような、ある関数によって計算される。最初の4つのストライプのパリティは次のように表され、下付き数字は対応するストライプを表す。

【0029】

【数1】

$$\begin{aligned} P_0 &= \text{セグメント0} \oplus \text{セグメント1} \oplus \text{セグメント2} \\ &= \text{ディスク0} \oplus \text{ディスク1} \oplus \text{ディスク2} \end{aligned}$$

$$\begin{aligned} P_1 &= \text{セグメント3} \oplus \text{セグメント4} \oplus \text{セグメント5} \\ &= \text{ディスク0} \oplus \text{ディスク1} \oplus \text{ディスク3} \end{aligned}$$

$$\begin{aligned} P_2 &= \text{セグメント6} \oplus \text{セグメント7} \oplus \text{セグメント8} \\ &= \text{ディスク0} \oplus \text{ディスク2} \oplus \text{ディスク3} \end{aligned}$$

$$\begin{aligned} P_3 &= \text{セグメント9} \oplus \text{セグメント10} \oplus \text{セグメント11} \\ &= \text{ディスク1} \oplus \text{ディスク2} \oplus \text{ディスク3} \end{aligned}$$

(記号 $\oplus$ は排他的OR関数を表す)

パリティ冗長は、ディスクの1つで使用不可能になったデータの再生を可能にする。たとえば、セグメント5のデータが使用できなくなったとき、その内容は、セグメント3および4、および、セグメントPのパリティ・データから確かめることができる。パリティ記憶はミラー記憶よりも費用はかからないが、信頼性はより低く性能も低い。

【0030】

図1に示したディスク配列は概念を示す目的のものである。実際には、ディスクアレー11は、ミラーおよびパリティ冗長にしたがってデータを記憶することができる複数のディスク12を単に持つだけである。全てのディスク12によって提供される使用可能な記憶空間のうち、その記憶空間の1部がミラー冗長に割り当てられ、別の1部がパリティ冗長に割り当てられる。ディスク12は複数の同じ大きさの記憶領域（図4の番号35で表すような）を含むように構成され、個々の領域が複数のセグメントを持つようにするのが望ましい。領域は、記憶空間の1つの仮想ビューの中にRAID区域を形成するようにまとめられる。さらに、RAID区域およびデータ冗長記憶手法がユーザまたはホストに透明（トランスペアレント：transparent）になるように、記憶空間の別の（ホストによって定義された）ビューがユーザまたはホストに提示される。これらの特徴については、図4を参照して以下に詳しく述べる。

10

【0031】

データ記憶装置10は、ミラー記憶方法およびパリティ記憶方法の間のデータの"移行"（migration）を管理する。両タイプの冗長性の管理はRAID管理装置16（図1）によって調整される。RAID管理装置16は、ディスクアレーの中の2つの異なった型のRAID区域を、ミラーRAID区域がパリティRAID区域に対してキャッシュとしてはたらくのと同様なメモリ階層として管理する。RAID管理装置16は、定義された性能プロトコルにしたがって、ミラーRAID区域とパリティRAID区域との間のデータをシフトし、組織し、その他の方法で管理する。ミラーRAID区域とパリティRAID区域との間のデータ移動の処理は"移行"と呼ばれる。

20

【0032】

データ記憶装置10は、最高の性能および信頼性がもたらされるように、性能により厳しいデータをミラーRAID区域に置くように努める。RAID管理装置16によって具体化される性能プロトコルは、2つの望ましい移行方針のうちのどちらかを含む。第1の移行方針は"アクセス頻度"と呼ばれ、この方針にしたがって、階層ディスクアレー上の最も頻繁にアクセスされるデータがミラーRAID区域18に維持される。アクセス頻度の低いデータはパリティRAID区域22に維持される。"アクセスの最近性"と呼ばれる第2の移行方針にしたがって、最も最近にアクセスされたデータはミラーRAID区域18に維持され、アクセスの最近性の低いデータはパリティRAID区域22に記憶される。その他の性能プロトコルを使用することもできる。そのようなプロトコルは、具体的なコン

30

【0033】

さらに、RAID管理装置16は、2つのパラメータの関数、すなわち物理的記憶容量の大きさ、およびデータ記憶装置に記憶されているユーザデータの現在量にしたがって、データ記憶装置の記憶資源を自動的に"調整"する。最初は全てのデータがミラーRAID区域に記憶される。これによって最高の性能および信頼性が得られるからである。より多くのデータがデータ記憶装置に加えられるにつれて、性能および信頼性を最適化するために、ミラーRAID区域とパリティRAID区域との間でデータの移行が行われる。データ記憶装置が満杯に近づくにしたがい、冗長性を通して信頼性を提供しながらユーザの要求をすべて満たすようにするために、いっそう多くのデータがパリティRAID区域に移行される。したがって、本発明によるデータ記憶装置は最大の融通性と適応をもたらす。この装置は、特定の記憶方法をユーザが選択する必要はないが、ユーザからのどんな要求をも満たすように適応できる。

40

【0034】

図4に、データ記憶装置10の使用可能な記憶空間のメモリ・マッピングを、マッピングされた仮想記憶空間の複数の層として示す。図の垂直に延びた矩形のそれぞれは、物理的記憶空間のビューを表す。この図において、物理的記憶空間34は2つの仮想記憶ビュー40および50で参照されている。物理的記憶空間34は4個のディスク（図1のディスク12のような）で表され、数字0、1、2、3で参照される。これらのディスクに関連

50

している4つの矩形は物理的記憶空間のビューを表し、ディスク1、2、3は大体同じ記憶容量を持っているが、ディスク0はやや少ない記憶容量を持っている。これらのディスクの記憶容量の例は、1ないし3ギガバイトである。記憶空間34は区域A0、A1、A2等に分割されている。個々の区域は、たとえば区域A0の中のストライプ0～Qのように、多くのストライプを持つ。個々の区域には、また、多くの領域35がある。領域35は、それぞれの領域が全ディスクアレーにわたって同じ大きさになるように、各記憶ディスク上で、均一な大きさの、ある選択された数のセグメントで構成されるのが望ましい。1つの領域35の大きさの例は1メガバイトである。

【0035】

ディスクの記憶空間は、物理的記憶空間34の第1の中間的な、RAIDレベル仮想ビュー40にマッピングされる。この第1仮想ビューは、概念的には、連続的にアドレス可能な記憶空間を表す第2のアプリケーション・ビューにマッピングすることができる1組のRAID区域である。記憶空間の物理的な構成およびRAIDビューはアプリケーション・ビューからは隠れている。

10

【0036】

RAID区域の記憶空間40は、ミラーおよびパリティ記憶空間を区別する記憶のビューである。たとえば、RAID区域42はM個の割り当てブロック43を持つミラーRAID区域を表し、RAID区域44はN個の割り当て区域45を持つパリティRAID区域を表す。割り当てブロック43および45は同じ大きさ（たとえば64kバイト）であることが望ましい。これらのRAID区域は、物理的記憶空間34上の対応する物理的區域A0、A1、A2等に関連する。

20

たとえば、RAID仮想ビューの16個の64k割り当てブロック43または45は、1個の1M（メガ）領域35の上にマッピングすることができる。

【0037】

ミラーおよびパリティRAID区域は、ディスクアレーの記憶空間34の全体を使い切る場合とそうでない場合とがある。したがって、あるアプリケーションの間、特定のRAID区域に対応しない、未使用で未指定の記憶空間がある場合がある。しかし、そのような記憶空間は、ミラーまたはパリティRAID区域のどちらにも変換することができる。また、RAID区域は、ディスクアレー上の連続した区域にマッピングされ、RAID区域に関連する領域のそれぞれは各記憶ディスク上の同じ物理アドレスに置かれるものとして図示してあることに留意されたい。代替方法として、RAID区域はディスクアレー上の不連続な区域にマッピングすることもできる。

30

【0038】

RAID区域で使用可能な記憶空間は、ユーザまたはホスト・アプリケーション・プログラムによって定義され、ユーザまたはホスト・アプリケーション・プログラムに提示される記憶のビューである第2の前面のアプリケーション・レベルの仮想ビュー50にマッピングされる。ユーザまたはホスト・アプリケーション・プログラムから見た場合、アプリケーション・レベルの仮想ビュー50は、記憶ディスク12上で使用可能な記憶空間を示す1つの大きな記憶容量を表す。

仮想記憶空間50は、それぞれが、0、1、2、.....、J-1、J、J+1、.....、L-1、L、L+1、.....、等として参照される同じ大きさの記憶仮想ブロック52および53からなる、線形の1組のビューを提示する。仮想ブロック52および53は、RAID区域40の割り当てブロックと同じ大きさ（たとえば64kバイト）であることが望ましい。仮想ブロック記憶空間50は、RAID区域40で提示されているビューの割り当てブロックを指す参照テーブルまたはポインタ（たとえば図4の矢印54で示す）によって表される。したがって、アプリケーション仮想ビュー50の仮想ブロック52および53は、仮想ブロック・テーブルで維持されているポインタを介して、RAID仮想ビュー40の割り当てブロック43および45に関連づけられている。仮想ブロック・テーブルから参照できるRAID区域の型には少なくとも2つ（すなわちミラーおよびパリティ）の型がある。

40

50

【0039】

RAID管理装置16は、物理的記憶空間全面にわたるRAID区域の構成を動的に変更できる。各型のRAID区域の数は、装置に記憶されつつあるユーザデータの量および物理的なディスク記憶空間の大きさに応じて増減される。その結果、RAIDレベル仮想ビュー40の中のRAID区域のディスク上へのマッピング、および、前面仮想ビュー50のRAIDビュー40へのマッピングは常にその状態が変化している。NVRAM21aおよび21b（図1）のメモリマップ記憶装置は、RAID管理装置16がRAID区域をディスク上にマッピングするために使う最新のマッピング情報、および、2つの仮想ビュー間のマッピングに用いられる情報を維持している。RAID管理装置は、RAIDレベルのマッピングを動的に変更するのに応じて、その変更を反映するためにメモリマップ記憶装置のマッピング情報も更新する。

10

【0040】

図1ないし4を参照して、データ記憶装置10の移行動作について以下に述べる。

【0041】

説明を続ける目的のために、アプリケーション・レベル仮想記憶空間50の仮想ブロック53は、物理的記憶空間34の区域A1に記憶されているパリティRAID区域44の中の関連する割り当てブロック45を参照する。そのような仮想ブロック53は"パリティ仮想ブロック"と呼ばれ、関連する割り当てブロック45は"パリティ割り当てブロック"と呼ばれる。同様に、仮想ブロック52は、物理的記憶空間34の区域A0に記憶されているミラーRAID区域42の中の関連する割り当てブロック43を参照する。そのような仮想ブロック52は、"ミラー仮想ブロック"と呼ばれ、関連する割り当てブロック43は"ミラー割り当てブロック"と呼ばれる。

20

【0042】

一般的に、データを1つのRAID区域から別のRAID区域に移行するためには、第1のRAIDレベル型（たとえば、ミラーすなわちレベル1）の割り当てブロックに関連した仮想ブロックが選択される。次に、第2のRAIDレベル型（たとえば、パリティすなわちレベル5）を表す、未使用の割り当てブロックが探される。未使用の割り当てブロックが見つからなければ、割り当てブロックが作られる。次に、データが、選択された仮想ブロックに以前関連していた割り当てブロックから、未使用の割り当てブロックに転送され、データは冗長レベルの変更を受ける。たとえば、以前ミラー冗長にしたがって記憶されていたデータは、あらたにパリティ冗長にしたがって（またはその逆）記憶される。最後のステップとして、アプリケーション・レベル仮想記憶空間50のRAIDレベル仮想記憶空間40へのマッピング54は、データのシフトを反映するために修正され更新される。第1のRAIDレベル型の割り当てブロックに以前関連していた選択された仮想ブロックは、こんどは、更新されたポインタを介して、移行されたデータを含む第2のRAIDレベル型の割り当てブロックを参照する。この処理中に起こるマッピング変更はメモリマップ記憶装置21の中で更新される。

30

【0043】

以下に、本発明の方法およびシーケンスにしたがった、ミラーおよびパリティ記憶区域の間のデータの移行について説明する。パリティ記憶区域からミラー記憶区域へデータを移行するために、次のシーケンスが使われる。

40

1. RAID管理装置が、ミラーRAID区域42の中に未使用のミラー割り当てブロック43を探す。
2. 見つからなければ、RAID管理装置はミラー割り当てブロックを作る（以下に述べる）。
3. RAID管理装置は、移行される仮想ブロックへの新しい記憶要求を中断する。
4. RAID管理装置は、その仮想ブロックに対する実行中のデータ記憶要求がすべて完了するまで待つ。
5. 仮想ブロック53に関連しているパリティ割り当てブロック45からのデータが臨時のメモリバッファに読み込まれる。

50

6. データが、ステップ2で選ばれたミラー割り当てブロック43に書かれる。
7. 仮想ブロック・テーブルが、ミラー割り当てブロック43にあるデータの新しい位置を参照するように修正される。
8. 中断された記憶要求が再開される。

上述の手続きにより、アプリケーション・レベル仮想ビュー50の仮想ブロック53は、パリティからミラー記憶に移行する。中間的仮想ビュー40に関しては、データは、パリティRAID区域44のパリティ割り当てブロック45からミラーRAID区域42のミラー割り当てブロック43に移行されている。物理的記憶空間では、データは区域A1から区域A0に移動している。

【0044】

未使用のミラー割り当てブロックが見つからなかった場合（上記のステップ1）、RAID管理装置は、3つの手法からなる次のシーケンスを試みる。まず、RAID管理装置は、未使用（したがって未指定）のRAID区域を見つけることを試みる。このRAID区域は、移行が常に進行できることを保証するのに必要な未使用のRAIDレベル記憶のシステムしきい値を犯すことなく、ミラーRAID区域に変換できるものである。これが失敗し、装置が未使用のRAIDレベル記憶の予備量以上を持っている場合、装置は、未使用のパリティ割り当てブロックを未使用のRAID区域に集めるために、パリティ記憶内のデータを移行する。この移行により、上記のようにミラーRAID区域に変換できる未使用のRAID区域が生じる場合、装置は未使用のRAID区域をミラーRAID区域に変換する。そうでない場合、代替方法として、未使用のミラー割り当てブロックを求めるため、または、未使用のRAID区域をミラーRAID区域に変換するのに十分な未使用のRAIDレベル記憶を増やすまで、装置はデータをミラーからパリティ記憶に移行し、ミラー記憶を詰め、未使用のRAIDレベル記憶をパリティに変換する。ミラー割り当てブロックはパリティ割り当てブロックよりも多くの物理的記憶空間を占めるので、この最後の方法は、未使用のRAIDレベル記憶装置の純増につながる。

【0045】

未使用のミラー割り当てブロックを求めて確立するために使われる作成／変換のプロトコルは、ユーザデータの量および物理的記憶空間の大きさに応じて、RAID管理装置がパリティ区域とミラー区域との間のメモリ割り当てを選択的に調節できるので、利点がある。RAID管理装置は、データ使用および記憶容量が変化するようにしたいが、ミラー記憶に保持されるデータ量が最大になるように、上記の1つまたは複数の手法を用いる。

【0046】

RAID管理装置は、空き時間（アイドルタイム）中に未使用のRAID区域を作ることにより、空間作成シーケンスが未使用のミラー割り当てブロックを生じる間、記憶要求が待たなければならない状況を避けるように試みる。しかし、ある状況においては、空間作成シーケンスの間、記憶要求が中断される場合がある。RAID管理装置は、仮想空間がRAIDビューより小さくなるような方法で仮想ブロック記憶空間を構成する。これにより、少なくとも1つのRAID区域に等しい空き空間が、移行またはその他の目的のためにとって置かれることが保証される。このようにして、この手法のシーケンスにより未使用のミラー割り当てブロックが常に生じる。

【0047】

ミラーからパリティ記憶にデータを移行するために、次のシーケンスが使われる。

1. RAID管理装置は、アクセス最近性またはアクセス頻度のような移行方針にしたがってミラーからパリティ記憶に移行するために、52から仮想ブロックを選ぶ。
2. RAID管理装置は、パリティRAID区域44の中に未使用のパリティ割り当てブロック45を探す。
3. そのようなブロックが見つからなかった場合、上述の作成方法にしたがって、移行のために確保しておいた空間がパリティRAID区域に変換される。
4. 移行される仮想ブロックに対する新しい記憶要求が中断される。
5. RAID管理装置は、その仮想ブロックへの実行中の記憶要求がすべて完了するま

10

20

30

40

50

で待つ。

6. データが、仮想ブロック52に関連しているミラー割り当てブロック43から臨時のメモリバッファに読み込まれる。

7. データが、選ばれたパリティ割り当てブロック45に書かれる。

8. 仮想ブロック・テーブルが、パリティ割り当てブロック45の中のデータの新しい位置を参照するために修正される。

9. 仮想ブロックへのデータ要求が再開される。

上述の9ステップからなる2つのシーケンスは、2つの異なる冗長レベルの間のデータの移行を行うために、本発明による記憶装置がどのように動作するか例を提供している。

【0048】

図1に再び戻ると、データ記憶装置10はさらに予備の記憶ディスク33の組31を持っている。データ記憶装置10は、ディスクアレーが動作中に、ディスクアレーの使用可能な機械的ベイに追加ディスクを"ホットプラグ"できるように設計されている。このとき、データ記憶装置の電源は切られない。データ記憶装置の記憶容量を拡大するために、1つまたは複数の予備ディスク33が、SCSIバス13とインターフェースするように、使用可能なベイに単に差し込まれる。制御装置14は新しいディスクを認識し、追加された記憶空間を使用できるRAID区域に同化するために、本発明の処理を開始する。

【0049】

図5ないし13を参照して、以下に、1つまたは複数の記憶ディスクをホットプラグすることによって記憶容量を拡大する方法を述べる。この方法は、データ記憶装置の空き時間にバックグラウンドで実行されることが望ましい。

【0050】

図5および6に、新しい記憶ディスクを組み込むために全RAID区域を一時に1つずつ再構成する、本発明による1つの方法を示す。図5で、数字4で参照される新しい追加ディスクがディスクアレーに加えられる。複数の記憶ディスクを加えることができるが、説明を単純にするために、この例では1つのディスクを加える。新しいディスクは初めは空である。新しいディスクは、ディスクアレーにある既存の記憶ディスクと同じかまたは異なる容量を持つことができる。

【0051】

図6を参照すると、最初のステップ100で、現在構成されているディスクアレー上に記憶空間を確保する。この目的およびその他の目的のために、ディスクアレーには常にいくらかの記憶空間があることが望ましい。たとえば、RAID管理装置16は、図4を参照して上述したように、RAID区域間でデータを移行する目的のために、1つまたは複数のRAID区域に等しい空間を確保する。確保された空間の大きさは、装置が記憶区域間でデータを移動するための空間を持てるのを確実にするために、ユーザデータに使える容量の1部としては計算されない。これは、データ移動または移行の目的以外に、確保された空間の使用を禁止または妨げるしきい値条件を制御装置に設定することによって行うことができる。

【0052】

新しい記憶ディスクを追加する処理の間に使用するために確保した空間は、RAID管理装置16、または制御装置14、またはその他の方法での通常の仕事の1部として管理することができる。確保された空間は、既存の1つのRAID区域以上の容量を持つ。このようにして、確保された空間は、1つのRAID区域に記憶されたデータと少なくとも同じ量のデータを保持するのに十分な大きさを持つ。空間の確保は、第1すなわちRAID仮想ビュー、第2すなわちユーザ仮想ビュー、またはその両者のビューの中で表すことができる。

【0053】

ステップ102で、図5のディスク4のような、1つまたは複数の記憶ディスクがディスクアレーに加えられる。ステップ103で、記憶ディスクの追加によって生じた新しい記憶空間を反映するために、第2の仮想記憶空間50が修正される。したがって、ホストは

10

20

30

40

50

、追加の空間がデータ記憶装置に加えられたことを直ちに認識する。しかし、ホストは、次のRAID区域拡大手法を通してホストの要求を収容する十分な空間が組み込まれるまで、新しいディスク空間への最初のアクセスから遅延を経験する場合がある。

【0054】

ステップ104で、1つの既存のRAID区域が選択される。既存のRAID区域とは、現在データを持っているが、新しいディスク4の空間を組み込むようにまだ構成されていない区域である。ステップ106で、選択されたRAID区域に含まれるデータのすべてが、次の2つの場所、すなわち(1)新しいディスク空間を含むすでに拡大されたRAID区域、または(2)ステップ100で確保された空間を含むディスクアレー上の他の空間、のどちらかに移動される。図5に、選択された区域A1(RAID仮想ビューのRAID区域42に対応する。図4を参照)から区域AVの中に確保された空間へのデータの移動を示す。説明の目的のために、確保された記憶空間は、1つの連続したRAID区域にマッピングされた連続した物理的領域AVに含まれているものとして示されている。しかし、典型的には、確保された記憶空間は、非連続的なかたちでディスクアレー全体に分配され、複数のRAID区域にマッピングされる。

10

【0055】

次に、選択された区域A1に対応する物理的記憶空間は、新しい記憶ディスク4を組み込むためにマッピングしなおされる、即ち再構成される(ステップ108)。このマッピングしなおしによって、複数の記憶ディスク0ないし3、および新しく追加された記憶ディスク4にわたる拡大されたRAID区域が作られる。拡大されたRAID区域は、新しいディスク(すなわち領域35e)を含む、アレーの中の各ディスク(領域35aないし35d)からの同じ大きさの領域(複数の同じ大きさのセグメントで構成される)を含む。既存のRAID区域を拡大されたRAID区域にマッピングしなおす処理(ステップ104ないし108)は、データ記憶装置の空き時間のバックグラウンド処理で行うことができる。

20

【0056】

RAID方法でまだ指定されていない記憶空間は、新しいディスク容量を含むように自動的に成長する。この空間は必要に応じてRAID区域に変換される。

ユーザの記憶要求を容易にするために拡大されたRAID区域がさらに必要な場合、使われていない記憶空間が探され、新しいディスク空間を含む拡大されたRAID区域にマッピングされる。この新しく拡大されたRAID区域は、ステップ106により、既存のRAID区域からデータを受け取るために使うことができることに留意されたい。

30

【0057】

このデータ移動が完了すると、階層ディスクアレー内のデータ移動およびRAID区域の拡大を反映するために、第1および第2の仮想記憶空間に対するメモリマップ記憶装置21のメモリマップ情報が更新される(ステップ114)。より具体的には、拡大処理中に各仮想ブロックの中のデータが移動するたびに、メモリマップ情報が更新される。この処理は、全てのRAID区域が新しいディスク空間を含むように拡大されるまで、ディスクアレー記憶装置を通して一時に1つのRAID区域をインデックスするために続けられる(ステップ116)。本発明のこの方法により、データはユーザに常に使用可能である。

40

【0058】

RAID管理装置16は、この処理中に第2の仮想記憶空間層50で起こるいかなる競合も解決する。具体的には、RAID管理装置16は、仮想ブロックに対するデータアクセス要求と仮想ブロックの移動との間の競合を解決する。競合を解決することにより、データはいつでも使用可能である。RAID管理装置は、多様な手法を通して競合を管理する。手法の1つの例は、選択されたRAID区域に関連する仮想ブロックへのアクセスを一時的に中断することである。"先取りなし"(non-preemption)と呼ばれるもう1つの手法は、競合し合う活動を避けるために、装置の資源を中断なしで使う。資源が1度にもっぱら1つの仕事に使われるので、競合が起こるのが防げる。もう1つの手法は、移動中のデータをアクセスしようとすることから起こる競合活動をログし、データが移動し終わった

50

ときにその競合を解決する方法である。

【0059】

この方法は、RAIDでない環境に一般化することができる。ディスクアレーの物理的記憶空間は、RAID区域の代わりに、図3および4のストライプ0ないしQのような複数のストライプに構成される。ストライプは、ディスクのすべてにわたって延び、ディスクあたり2つ以上のセグメントを持つストライプも可能ではあるが、典型的には、各ディスクから1つのセグメントを含む。ストライプに保持されているデータを一時的に記憶するために、物理的空間の1部を確保してもよいし、しなくてもよい。確保する場合は、その空間部分は、1つのストライプに記憶されるデータ量以上の量を記憶するのに十分な大きさのものである。

10

【0060】

ストライプは拡大のために選択され、そのストライプの中の割り当ては禁止される。選択されたストライプの中のデータはすべて、ディスクアレーの別の記憶場所に移動される。次に、そのストライプを含む物理的記憶空間は、データを記憶するために、新しい記憶ディスクの上のセグメントを含む拡大されたストライプに再構成される。このようにして、拡大されたストライプは、新しく追加された記憶ディスクを含む全ての記憶ディスクにまたがる。次に、拡大されたストライプの中の割り当てが再開される。この処理は、新しいディスクを含むようにすべてのストライプが構成され終わるまで、ストライプごとに続けられる。この処理中、選択されなかった全てのストライプへのデータ可用性は維持される。

20

【0061】

図7ないし13に、再構成の処理中に、ストライプ内でセグメントが再編成され詰められる、本発明によるもう1つの方法を示す。上述した方法と異なり、データは、最初に、選択されたストライプの外の新しい場所ではなく、選択されたストライプ内で移動される。さらに、この方法を使えば、既存の記憶アレー上に空間を確保せずに、オンラインで再構成を実行することができる。

【0062】

図7ないし12に、はじめ4個のディスク0ないし3からなるディスクアレー上の処理を示す。ディスクアレーの物理的記憶空間は、予定のデータ量を記憶するための複数のストライプに構成される（図13のステップ200）。各ストライプは記憶ディスク0ないし3のそれぞれに延び、ディスク0～3それぞれからのセグメントを含む。各ストライプはデータおよび冗長パリティ情報Pを含む。

30

【0063】

ストライプ構成すなわち物理的記憶空間のメモリ・レイアウトおよびそこに記憶されているデータを表すプリマップが、データ記憶装置によって生成され維持される（ステップ202）。プリマップは、ディスクアレー制御装置14（図1）のメモリマップ記憶装置21の中で維持され更新されるテーブルとして具体化することができる。代替方法として、プリマップは、RAID管理装置16に組み込まれた1組の方程式であってもよい。これらの方程式は、ディスクアレーの記憶ディスクの数に基づいてマッピング情報を導き出す。

40

【0064】

次に、新しい追加ディスク4がディスクアレーに加えられる（ステップ204）。新しいディスクははじめは空である。RAID管理装置16は、新しいストライプ構成、すなわち、追加記憶ディスク4を組み込んだ後のディスクアレーの物理的記憶空間のメモリ・レイアウトを表すポストマップを作る（ステップ206）。このポストマップは、同様に、テーブルとして、メモリマップ記憶装置21に保持される。代替方法として、ポストマップは、ディスクアレーに新たに構成されたディスクの新しい数をパラメータとする方程式によって生成することもできる。

【0065】

ステップ208で、グループAのセグメントが定義される。グループAは、新しく追加さ

50

れた記憶ディスクを含む各記憶ディスクからのセグメントを持つ1つの拡大されたストライプからなるのが望ましい。図7で、グループAは、セグメント0、1、2、P、および、ディスク4の新しいセグメントNからなる。次に、グループBのセグメントが定義される（ステップ210）。グループBは、グループAにないセグメントを含む。グループBは、1つの拡大されたストライプを満たすのに十分なセグメントからなることが望ましい。図7ないし12の例では、グループBは、最初は、グループAと同じ大きさで、1つの拡大されたストライプからのセグメントを含んでいる。しかし、処理の後の方で、グループBは、2つの拡大されたストライプからのセグメントを含むために大きくされる。したがって、グループBのセグメントは、物理的に同じストライプにある場合もない場合もある。グループAおよびBが定義された後、RAID管理装置は、拡大処理中に起こり得る競合にそなえて管理を行い始める。

10

【0066】

処理の始めに、グループAのセグメントの中のデータは、プリマップにしたがったセグメントの第1のデータ配列から、ポストマップにしたがったセグメントの第2のデータ配列に再編成する必要がある場合がある（ステップ213）。すなわち、プリマップにしたがって配列されたデータはポストマップに対して評価され、拡大処理の後、データが同じ場所に残るか、または再編成する必要があるかどうか判断される。しかし、何回か処理を繰り返した後、グループAは空のセグメントだけになり、したがって、再編成すべきデータがなくなる。

【0067】

20

したがって、判断ステップ211で、グループAのセグメントがすべて空であるかどうか判断される。そうである場合、再編成ステップは実行されず、データは単にグループBのセグメントからグループAの空のセグメントに移される。

逆に、セグメントが空でない場合、次の判断ステップ212で、ポストマップ構成から見てデータがいま適切な順序にあるかどうか判断される。そうである場合、再編成ステップは再び避けられる。他方、データが適切な順序でないならば、再編成ステップ213が実行される。これらのステップについては、図7ないし12に示した例を参照して、次に詳しく述べる。

【0068】

この例では、プリマップは、図7に示したような、ディスク0～3上のデータ構成を示していると仮定する。さらに、ポストマップは、図11に示したような、新しいディスク4を組み入れた記憶構成を持つと仮定する。

30

【0069】

図7に示したグループAのセグメントのデータは、すでに所望の順序になっていることに留意されたい。再編成は必要ない。代わりに、データは、グループBのセグメント3から、以前パリティ情報を持っていたグループAのセグメントPを満たすために、単に移される。図13の流れ図に沿って説明すると、判断ステップ211は、グループAのセグメントは空でないことを示す"NO"を返し、判断ステップ212は、データが適切な順序になっていることを示す"YES"を返す。したがって、再編成ステップ213はバイパスされ、流れはステップ214に直接進み、データはグループBのセグメントからグループAの空のセグメントを満たすように移される。ステップ216で、グループAのセグメントのデータに対するパリティ情報が再び生成され、新しいディスク4の上の新しいセグメントNに置かれる。次に、グループAおよびBのセグメントに関連する仮想ブロックに関してこの処理中に起こった競合が解決される。

40

【0070】

図8に示すように、グループAは、空になったセグメントV、データセグメント4、パリティ・セグメントP、データ・セグメント5、および新しいセグメントNを、それぞれディスク0ないし4の上に含むように再定義される（ステップ208）。グループBも、セグメント6、P、7、8、Nをそれぞれディスク0ないし4の上に含むように再定義される。グループAのセグメントのデータは適切な順序でなく、したがって、ステップ213

50

で再編成されることに留意されたい。セグメント4のデータが、ディスク1からディスク0の空になったセグメントVに移され、その後、セグメント5のデータが、ディスク3からディスク1のいま空になったセグメントに移される。

【0071】

再編成中に、データが、グループA内のディスク4上の新しいセグメントNにシフトされてもよいことに留意されたい。たとえば、ポストマップがそのように指示した場合、データは、グループAのセグメント4から、新しく追加された記憶ディスク4の未使用の新しいセグメントNに移すことができる。

【0072】

ステップ213でデータがグループA内で再編成された後、データは空のセグメントを満たすようにグループBから移される。ここで使われている"空の"セグメントという意味は、空にされたセグメント（Vで示す）、以前にパリティ情報を記憶していたセグメント（Pで示す）、および、新しいディスクによって提供された新しいセグメント（Nで示す）を意味する。図8で、ディスク0のセグメント6のデータは、ディスク2の空のセグメントPを満たすために移される。また、ディスク2のセグメント7のデータは、ディスク4の新しいセグメントNを満たすために移される。パリティ情報が再び生成され、ディスク3の空にされたセグメントに置かれる。再び、グループAおよびBのセグメントに関連した仮想ブロックに関する競合は解決される。

【0073】

図9は、図13の流れ図を通して行われる3回目の繰返しを示し、グループAおよびBが、それぞれ、第3および第4のストライプをカバーするために再定義される。この繰返しは、グループA内のデータの再編成、および、データの詰めすなわちグループBから提供されたデータによる充填の組合せを再び使う。新しいデータの組に対してパリティ情報が再び生成され、ディスク2のセグメントに記憶される。

【0074】

図10は、4回目の繰返しを示す。グループAが、第4のストライプのセグメントをカバーするために再定義される。グループAはいま全て空のセグメントを持っていることに留意されたい。グループA内には再編成すべきデータがない。

したがって、図13の流れ図のステップ211により、再編成ステップ213はバイパスされ、充填ステップ214に直接進む。したがって、再編成のサブ処理は、始めに実行されるだけで、数回の繰返しの後に急速に止む。実際には、Nを新しいディスクを足す前のアレーの中の記憶ディスクの数とすると、再編成ステップはN-1の繰返しの後に終わる。この例ではNは4に等しく、したがって、再編成サブ処理は3回目の繰返しの後に止まる。

【0075】

グループBは、第5および第6のストライプのセグメントをカバーするために再定義され拡大される。この拡大が起こるのは、グループAの全ての空のセグメントを満たすのに1つのストライプに収まる量のデータよりも多くのデータが使われるからである。

【0076】

図11は、7回目の繰返しを示す。この繰返しにおいて、グループAとBとの間に、未使用のセグメントのギャップの形成が始まる。このギャップは、データがグループBおよびその下のより小さい記憶空間からグループAの上のより大きな拡大された空間に移されるにしたがい、発展する。このギャップは、Nを新しいディスクを足す前のアレーの中の記憶ディスクの数とすると、2N-1の繰返しのときに形成し始める。処理が続くにしたがいギャップは大きくなる。

【0077】

この処理は、グループBがディスクアレー記憶空間の底に到達したときに終わる。その時点で、ギャップはその最大の大きさに成長して、すべて、使用可能な未使用のセグメントを含む。さらに、すべての未使用のセグメントは、ディスクアレーの中の使用されているセグメントの下に置かれる。この処理は、方程式（マッピングのために具体化されている

10

20

30

40

50

場合)に、ディスクアレーのディスクの新しい数をロードすることができ、方程式の使用を容易にするので、利点が多い。

【0078】

図12は、多くの繰返しの後の、処理の一般的なケースを示す。ディスク4上の新しい記憶空間を含めるために、データがグループAの上のセグメントG、G-1、G-2、G-3等に詰められている。他方、グループBの中および下のデータは、新しいディスクを加える前に存在していたままに、セグメントH、H+1、H+2等に残っている。新しいディスク4の上のセグメントは空である。グループAは全て空のセグメントからなる。グループBは、グループAを満たすべきデータからなる。グループAとBとの間のギャップには、多くの空のセグメントEがある。

10

【0079】

処理は、新しいディスク空間を含むようにすべてのストライプが拡大されるまで、ストライプごとに続けられる。その後、新しいディスクを加えた結果の拡大を反映するために、ホストが使用できる最大空間が増やされる。上述の方法は、種々のRAID区域を持つ階層ディスクアレーにも使うことができる。この場合、RAID区域内の領域が選択されたRAID区域内で再編成され、その後、上記の方法にしたがって、選択されたRAID区域の外の領域からのデータで詰められる。そのような領域を全体として扱うこともできるが、図10を参照して述べた方法でRAID区域全体が完了するまで、各RAID区域内のストライプがストライプごとに拡大できるように、複数のストライプに分けて処理することが望ましい。

20

【0080】

本発明による方法は、データの可用性を維持しながら、記憶容量を大きくするために追加の記憶ディスクをホットプラグすることができるので、利点が多い。

新しい空間を区域ごとに再編成することにより、データは、ディスクアレー上に存在しアクセス可能であり続け、従来技術で行われたように別の記憶装置に移す必要がない。

【0081】

本発明には、例として次のような実施態様が含まれる。

(1) 物理的記憶空間を定義する記憶ディスク(12)を複数持つディスクアレー(11)を備えたデータ記憶装置の記憶容量を拡大する方法であって、

(a) 予定のデータ量を記憶させるために、前記ディスクアレー(11)の物理的記憶空間(34)を、それぞれが前記ディスクアレー(11)の複数の記憶ディスク(0~3)にわたって延び、前記ディスクアレーの個々の記憶ディスクからのセグメントを含む複数のストライプに構成するステップと、

30

(b) 前記ディスクアレー(11)に少なくとも1つの追加記憶ディスク(4)を加えるステップと、

(c) ストライプを選択するステップと、

(d) 前記選択されたストライプの少なくとも1つのセグメントから前記ディスクアレー上の別の場所にデータを移すステップと、

(e) データを記憶させるために、前記選択されたストライプを持つ前記物理的記憶空間(34)を、前記複数の記憶ディスク(0~3)および前記追加記憶ディスク(4)にわたる、前記個々の記憶ディスクからのセグメントおよび前記追加記憶ディスクからのセグメントを含む拡大されたストライプに再構成するステップと、

40

を有する方法。

【0082】

(2) 前記方法が、

少なくとも各ストライプに記憶されている量のデータを記憶するのに十分な大きさの、前記ディスクアレー(11)の物理的記憶空間(34)の部分を確認するステップと、

前記選択されたストライプから前記確認された部分に全てのセグメントを移すステップと、

をさらに有する、上記(1)に記載の方法。

50

(3) データを移す前記ステップ (d) が、1つのセグメントから (1) 前記選択されたストライプ内の別のセグメント、または (2) 前記ストライプに含まれる前記追加ディスクからのセグメントのどちらかにデータを移すことによって、前記ストライプ内のセグメントを再編成するステップをさらに有する、上記 (1) に記載の方法。

【0083】

(4) 物理的記憶空間を定義する記憶ディスク (12) を複数持つディスクアレー (11) を備えたデータ記憶装置の記憶容量を拡大する方法であって、

(a) 予定のデータ量を記憶するために、前記ディスクアレー (11) の物理的記憶空間 (34) を、それぞれが前記ディスクアレーの複数の記憶ディスク (0~3) にわたって延び、前記ディスクアレーの個々の記憶ディスクからのセグメントを含む複数のストライプに構成するステップと、

(b) 前記ディスクアレーの前記物理的記憶空間 (34) の構成を示すプリマップを提供するステップと、

(c) 前記ディスクアレー (11) に少なくとも1つの追加記憶ディスク (4) を加えるステップと、

(d) 前記追加記憶ディスクを組み込んだ後の前記ディスクアレーの物理的記憶空間の新しい構成を示すポストマップを作るステップと、

(e) 前記個々の記憶ディスク (0~3) からのセグメントおよび前記追加記憶ディスク (4) からのセグメントを含む、少なくとも1つの拡大されたストライプを含むようにグループAのセグメントを定義するステップと、

(f) グループAにないセグメントを含むためにグループBのセグメントを定義するステップと、

(g) 少なくとも1つの空のセグメントを持つグループAのセグメントの中のデータを、前記プリマップにしたがったセグメントの中の第1のデータ配列から、前記ポストマップにしたがったセグメントの中の第2のデータ配列に再編成するステップと、

(h) グループBの少なくとも1つのセグメントから、グループAの前記少なくとも1つの空のセグメントへデータを移すステップと、
を有する方法。

【0084】

(5) 前記方法が追加ステップとして、(i) 前記ステップ (h) の後、グループAの前記セグメントに含まれるデータに基づいてパリティを再び生成するステップをさらに有する、上記 (4) に記載の方法。

(6) 前記方法が、前記ステップ (e) ~ (h) を繰り返すステップをさらに有し、前記ステップ (e) ~ (h) をある回数繰り返した後で前記再編成するステップ (g) が停止する、上記 (4) に記載の方法。

(7) 前記グループAおよびBのセグメントの中のデータへのアクセスと、そのデータの移動との間の競合を解決するステップをさらに有する、上記 (4) に記載の方法。

【0085】

(8) 物理的記憶空間を定義する記憶ディスク (12) を複数持つ階層ディスクアレー (11) を備えたデータ記憶装置の記憶容量を拡大する方法であって、

(a) 物理的記憶空間 (34) を、前記ディスクアレーの複数の記憶ディスク (0~3) にわたって延び、前記ディスクアレーの個々の記憶ディスクからの記憶空間の1つまたは複数のセグメントを含み、第1のRAIDレベルにしたがってデータを記憶する第1RAID区域、および第2のRAIDレベルにしたがってデータを記憶する第2RAID区域を含む複数のRAID区域として物理的記憶空間を提示する第1の仮想記憶空間 (40) にマッピングするステップと、

(b) 前記第1の仮想記憶空間を、記憶空間を複数の仮想ブロックとして提示する第2の仮想記憶空間 (50) にマッピングするステップと、

(c) 前記データ記憶装置の物理的記憶空間を拡大するために、少なくとも1つの追加記憶ディスク (4) を前記階層ディスクアレーに加えるステップと、

10

20

30

40

50

(d) 前記追加記憶ディスク(4)を加えることによって生じる新しい記憶空間を反映するために前記第2の仮想記憶空間(50)の使用可能な容量を修正するステップと、

(e) R A I D区域を選択するステップと、

(f) 前記選択されたR A I D区域から前記階層ディスクアレー上の別の場所にデータを移すステップと、

(g) 前記階層ディスクアレー内のデータ移動を反映するために、前記第2の仮想記憶空間(50)の前記第1の仮想記憶空間(40)へのマッピングを更新するステップと、

(h) 前記選択されたR A I D区域に対応する物理的記憶空間を、前記複数の記憶ディスクおよび前記追加記憶ディスクにわたる、個々の記憶ディスク(0~3)および追加記憶ディスク(4)からの1つまたは複数のセグメントを含む、少なくとも1つの拡大されたR A I D区域にマッピングしなおすステップと、

(i) 前記階層ディスクアレーのすべてのR A I D区域が追加記憶ディスク上のセグメントを含むように拡大されるまで、前記ステップ(e)~(h)を繰り返すステップと、を有する方法。

【0086】

(9) 前記方法が、R A I D区域に記憶されているのと同じデータ量を保持するのに十分な大きさの少なくとも1つの区域を確保するステップをさらに有し、データを移す前記ステップ(f)が前記選択されたR A I D区域から前記確保された区域にデータを移すステップを有する、上記(8)に記載の方法。

【0087】

(10) 物理的記憶空間を定義する記憶ディスク(12)を複数持つ階層ディスクアレー(11)を備えたデータ記憶装置の記憶容量を拡大する方法であって、

(a) 物理的記憶空間(34)を、前記ディスクアレーの複数の記憶ディスク(0~3)にわたって延び、前記ディスクアレーの個々の記憶ディスクからの記憶空間の1つまたは複数のセグメントを含み、第1のR A I Dレベルにしたがってデータを記憶する第1のR A I D区域、および第2のR A I Dレベルにしたがってデータを記憶する第2のR A I D区域を含む複数のR A I D区域として物理的記憶空間を提示する第1の仮想記憶空間(40)にマッピングするステップと、

(b) 前記第1の仮想記憶空間(40)を、記憶空間を複数の仮想ブロックとして提示する第2の仮想記憶空間(50)にマッピングするステップと、

(c) 前記データ記憶装置の物理的記憶空間(34)を拡大するために、少なくとも1つの追加記憶ディスク(4)を前記階層ディスクアレーに加えるステップと、

(d) 使われていない記憶空間を探すステップと、

(e) 前記使われていない記憶空間を、前記複数の記憶ディスクおよび前記追加記憶ディスクにわたる、個々の記憶ディスクおよび追加記憶ディスクからの1つまたは複数のセグメントを含む、少なくとも1つの拡大されたR A I D区域にマッピングするステップと、

(f) 関連するR A I Dレベルにしたがってデータを現在保持している既存のR A I D区域を探すステップと、

(g) 前記既存のR A I D区域から前記拡大されたR A I D区域へデータを移すステップと、

(h) 前記既存のR A I D区域に対応する物理的記憶空間を、前記複数の記憶ディスクおよび前記追加記憶ディスクにわたる少なくとも1つの拡大されたR A I D区域にマッピングしなおすステップと、

を有する方法。

【0088】

【発明の効果】

本発明の方法により、データの可用性を維持しながら、複数の記憶ディスクを持つディスクアレーに新しく記憶ディスクを追加して、データ記憶装置の記憶容量を拡大することができる。

【図面の簡単な説明】

【図 1】本発明によるディスクアレー・データ記憶装置を示すブロック図。

【図 2】複数の記憶ディスク上の記憶空間、および、RAID レベル 1 によるデータ記憶を示す図。

【図 3】複数の記憶ディスク上の記憶空間、および、RAID レベル 5 によるデータ記憶を示す図。

【図 4】1 つの物理的記憶空間に関して 2 つの仮想記憶空間がマッピングされている、本発明によるメモリマップ配列を示す図。

【図 5】複数の記憶ディスク上の記憶空間を示し、元の RAID 区域が新しいディスク上の空間を含むために拡張できるように、1 つの RAID 区域全体から別の記憶空間へのデータ移動を示す図。

10

【図 6】本発明の 1 つの具体化による、データの可用性を維持しながら、新しいディスクを追加して記憶容量を増やす処理ステップを示す流れ図。

【図 7】複数の記憶ディスク上の記憶空間を示し、新しい追加記憶ディスクの新しいセグメントを含めるために、ストライプ内部のデータを再編成して詰め込み同時にストライプを拡大する、本発明の別の具体化の最初のステップを示す図。

【図 8】図 7 の記憶空間を示し、図 7 に続くステップを示す図。

【図 9】図 7 の記憶空間を示し、図 8 に続くステップを示す図。

【図 10】図 7 の記憶空間を示し、図 9 に続くステップを示す図。

【図 11】図 7 の記憶空間を示し、図 10 に続くステップを示す図。

【図 12】図 7 の記憶空間を示し、図 11 に示したステップの後に続いてとられるステップの一般的なフォーマットを示す図。

20

【図 13】図 7 ないし 12 に示したように、ディスクを追加することによって記憶容量を拡大するステップを示す流れ図。

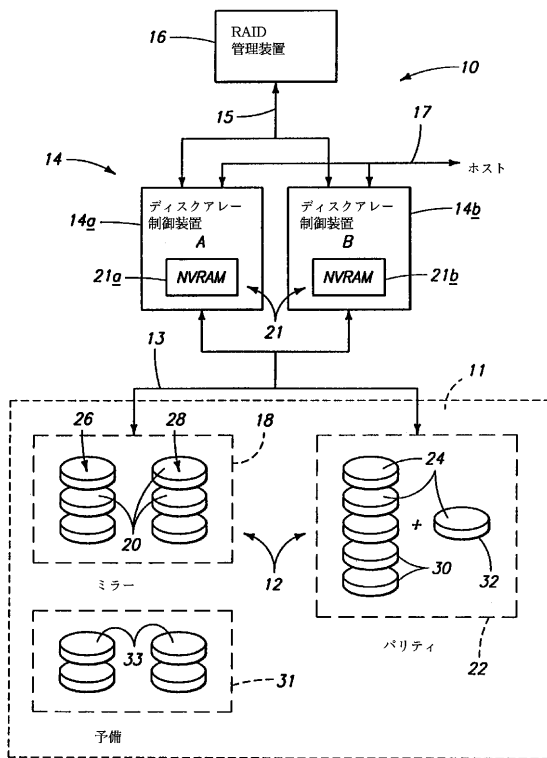
【符号の説明】

1 0	データ記憶装置
1 1	階層ディスクアレー
1 2	記憶ディスクまたはディスク駆動装置
1 3	インターフェース・バス (SCSI バス)
1 4、14a、14b	ディスクアレー制御装置
1 5	インターフェース・プロトコル
1 6	RAID 管理装置
1 7	入出力インターフェース・バス
1 8	ミラー・グループ
2 1	メモリマップ記憶装置
2 1 a、2 1 b	不揮発性 RAM
2 2	パリティ・グループ
2 6	ディスクの第 1 の組
2 8	ディスクの第 2 の組
3 0	5 個のディスク
3 2	6 番目のディスク

30

40

【図 1】



【図 2】

ディスク

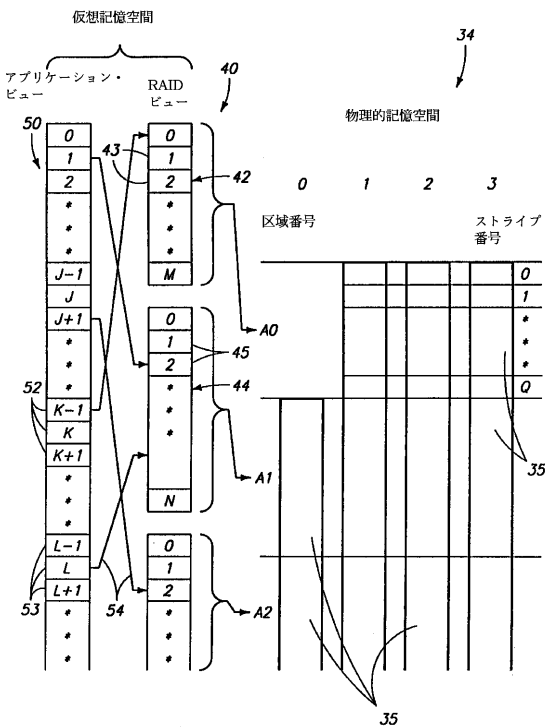
0	1	2	3	ストライプ番号
0	0'	1	1'	0
2	2'	3	3'	1
4	4'	5	5'	2
6	6'	7	7'	3
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
T-1	T-1'	T	T'	S

【図 3】

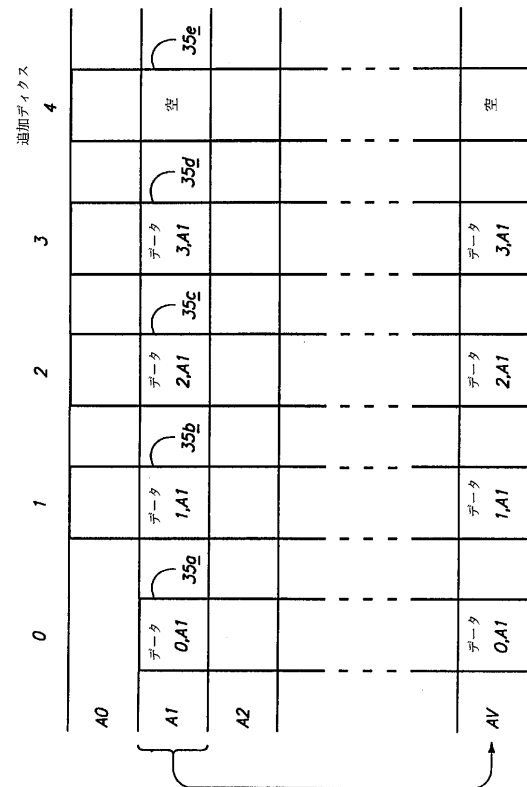
ディスク

0	1	2	3	ストライプ番号
0	1	2	P	0
3	4	P	5	1
6	P	7	8	2
P	9	10	11	3
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
P	R-2	R-1	R	Q

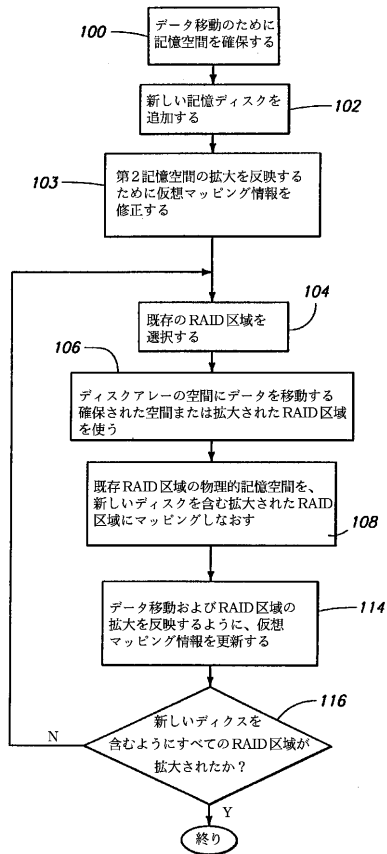
【図 4】



【図 5】



【図 6】



【図 7】



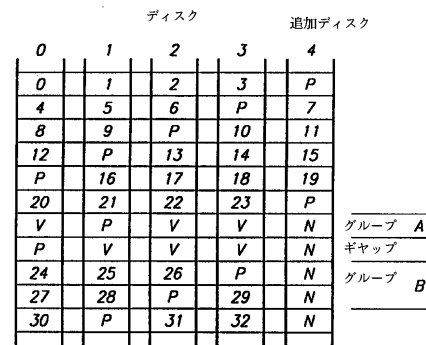
【図 8】



【図 9】



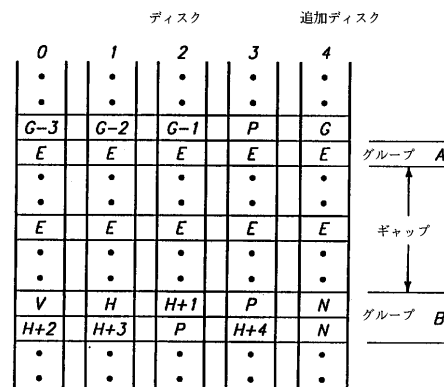
【図 11】



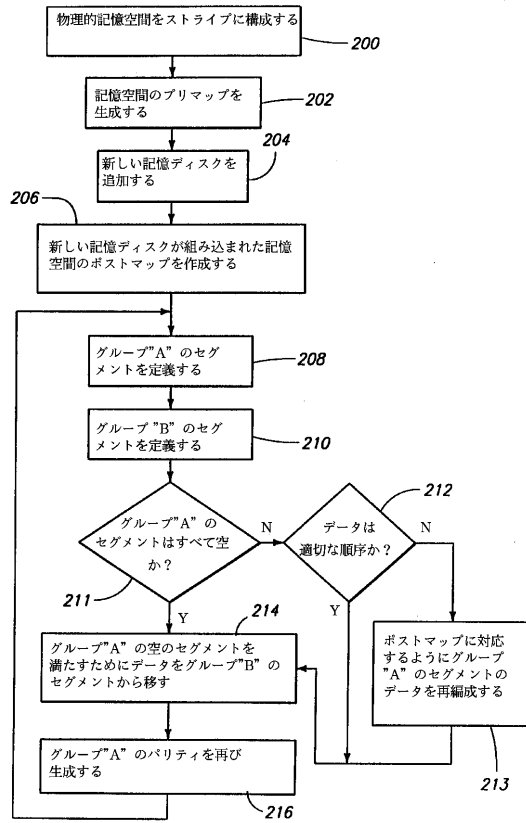
【図 10】



【図 12】



【図 13】



フロントページの続き

- (72)発明者 テレサ・エイ・バークス
アメリカ合衆国 83642 アイダホ州メリディアン、サン・レイモン・ドライブ 3100
- (72)発明者 ブライアン・ダイヤモンド
アメリカ合衆国 83704 アイダホ州ボイジー、ラムズゲイト 9704

審査官 木村 貴俊

- (56)参考文献 特開平07-141121 (JP, A)
特表平05-505481 (JP, A)

- (58)調査した分野(Int.Cl.⁷, DB名)
G06F 3/06-3/08
G06F 12/00-12/16
G11B 20/00-20/18