

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号
特許第5048760号
(P5048760)

(45) 発行日 平成24年10月17日 (2012. 10. 17)

(24) 登録日 平成24年7月27日 (2012. 7. 27)

(51) Int. Cl.

F I

G O 6 F 3 / 0 6 (2 0 0 6 . 0 1)

G O 6 F 3 / 0 6 3 O 5 C

請求項の数 17 (全 19 頁)

(21) 出願番号	特願2009-512308 (P2009-512308)	(73) 特許権者	506051681
(86) (22) 出願日	平成19年5月24日 (2007. 5. 24)		コンベレント・テクノロジーズ
(65) 公表番号	特表2009-538482 (P2009-538482A)		アメリカ合衆国ミネソタ州55344, エ
(43) 公表日	平成21年11月5日 (2009. 11. 5)		デン・プレイリー, スメタナ・レイン 7
(86) 国際出願番号	PCT/US2007/069669		625
(87) 国際公開番号	W02007/140260	(74) 代理人	100140109
(87) 国際公開日	平成19年12月6日 (2007. 12. 6)		弁理士 小野 新次郎
審査請求日	平成22年5月21日 (2010. 5. 21)	(74) 代理人	100089705
(31) 優先権主張番号	60/808, 045		弁理士 社本 一夫
(32) 優先日	平成18年5月24日 (2006. 5. 24)	(74) 代理人	100075270
(33) 優先権主張国	米国 (US)		弁理士 小林 泰
		(74) 代理人	100080137
			弁理士 千葉 昭男
		(74) 代理人	100096013
			弁理士 富田 博行

最終頁に続く

(54) 【発明の名称】 R A I D管理、再割振り、およびリストライピングのためのシステムおよび方法

(57) 【特許請求の範囲】

【請求項 1】

ディスク・ドライブ・システムにおける R A I D のリストライピングの方法であって、
複数の R A I D デバイスのそれぞれについて計算された初期スコアと少なくとも 1 つの
第 2 のスコアとの比較結果に基づいて、前記複数の R A I D デバイスから、移行のための
R A I D デバイスを選択するステップであって、

前記初期スコアは、現在の状態における前記 R A I D デバイスに関連するものであ
り、スコアを求めるための 1 以上のファクタに基づいて計算されるものであり、

前記第 2 のスコアは、使用可能なディスク空間に位置する少なくとも 1 つの仮想 R
A I D デバイスに関連するものであり、スコアを求めるための 1 以上のファクタに基づい
て計算されるものである、

選択するステップと、

前記少なくとも 1 つの仮想 R A I D デバイスに基づいて、少なくとも 1 つの代替 R A I
D デバイスを作成するステップと、

前記少なくとも 1 つの代替 R A I D デバイスへ、選択した前記 R A I D デバイスに格納
されたデータを移動するステップと、

選択した前記 R A I D デバイスを取り除くステップと
を備える方法。

【請求項 2】

請求項 1 に記載の方法であって、少なくとも 1 つの前記第 2 のスコアが、置換スコアを

含み、前記置換スコアは、存在する使用可能なディスク空間に位置する少なくとも1つの仮想RAIDデバイスと関連する、方法。

【請求項3】

請求項1に記載の方法であって、少なくとも1つの前記第2のスコアが、存在する使用可能なディスク空間と、前記RAIDデバイスにより占有されるディスク空間の少なくとも一部との組合せの中に位置した少なくとも1つの仮想RAIDデバイスと関連するオーバーレイ・スコアを含む、方法。

【請求項4】

請求項1に記載の方法であって、スコアを求めるための1以上の前記ファクタが、RAIDレベル、RAIDストライプ・サイズ、RAID範囲サイズ、ディスク・カテゴリ、ディスク上の位置、ディスク筐体、ディスク筐体電源、およびディスクへの通信経路のうちの少なくとも1つを含む、方法。

10

【請求項5】

請求項4に記載の方法であって、前記ファクタが、前記計算で使用される可変の重みを有する、方法。

【請求項6】

請求項1に記載の方法であって、前記初期スコアおよび少なくとも1つの前記第2のスコアのそれぞれが、スコアを求めるための同じファクタを使用して計算される、方法。

【請求項7】

請求項4に記載の方法であって、複数のRAIDデバイスのそれぞれについて計算された初期スコアと少なくとも1つの第2のスコアとの比較結果に基づいて、移行のためのRAIDデバイスを選択する前記ステップが、少なくとも1つの前記第2のスコアが前記初期スコアよりも良好である場合に前記RAIDデバイスを選択するステップを含む、方法。

20

【請求項8】

請求項1に記載の方法であって、移行のためのRAIDデバイスを選択する前記ステップと、少なくとも1つの代替RAIDデバイスを作成する前記ステップと、データを移動する前記ステップと、選択した前記RAIDデバイスを取り除く前記ステップとが、手動介入なしに自動的に行われる、方法。

【請求項9】

請求項8に記載の方法であって、前記ステップが、前記ディスク・ドライブ・システムへのサーバ・データ・アクセスを損失することなく、また、前記データの弾性を損なうことなく、実施される、方法。

30

【請求項10】

請求項8に記載の方法であって、前記ステップが、周期的、継続的、RAIDデバイス移行後ごと、ディスク・ドライブの追加時、およびディスク・ドライブの取外し前、のうちの少なくとも1つで実施される、方法。

【請求項11】

請求項1に記載の方法であって、前記少なくとも1つの代替RAIDデバイスへ、選択した前記RAIDデバイスに格納されたデータを移動する前記ステップが、少なくとも1つの一時的RAIDデバイスを作成するステップを更に含む、方法。

40

【請求項12】

請求項11に記載の方法であって、選択した前記RAIDデバイスに格納されたデータを前記少なくとも1つの一時的RAIDデバイスへ移動し、次いで、前記一時的RAIDデバイスから前記少なくとも1つの代替RAIDデバイスへ移動するステップを更に含む方法。

【請求項13】

ディスク・ドライブ・システムであって、

RAIDサブシステムと、

少なくとも1つのディスク記憶システム・コントローラを有するディスク・マネージャ

50

であって、自動的に、

複数の R A I D デバイスについて計算された初期スコアと少なくとも 1 つの第 2 のスコアとの比較結果に基づいて、前記複数の R A I D デバイスから R A I D デバイスを選択するように構成され、前記初期スコアは、現在の状態における前記 R A I D デバイスに関連するものであり、スコアを求めるための 1 以上のファクタに基づいて計算されるものであり、かつ、前記第 2 のスコアは、使用可能なディスク空間に位置する少なくとも 1 つの仮想 R A I D デバイスに関連するものであり、スコアを求めるための 1 以上のファクタに基づいて計算されるものであり、かつ、

前記少なくとも 1 つの仮想 R A I D デバイスに基づいて、少なくとも 1 つの代替 R A I D デバイスを作成し、

選択した前記 R A I D デバイスに格納された前記データの少なくとも一部を前記代替 R A I D デバイスへ移動し、

選択した前記 R A I D デバイスを取り除く
ように構成されたディスク・マネージャと
を備えるディスク・ドライブ・システム。

【請求項 14】

請求項 13 に記載のディスク・ドライブ・システムであって、少なくとも 1 つ前記第 2 のスコアが、存在する使用可能なディスク空間と、前記 R A I D デバイスにより占有されるディスク空間の少なくとも一部との組合せの中に位置した少なくとも 1 つの仮想 R A I D デバイスに関連するオーバーレイ・スコアを含む、ディスク・ドライブ・システム。

【請求項 15】

請求項 14 に記載のディスク・ドライブ・システムであって、前記少なくとも 1 つの代替 R A I D デバイスが、前記少なくとも 1 つの仮想 R A I D デバイスに基づく、ディスク・ドライブ・システム。

【請求項 16】

請求項 13 に記載のディスク・ドライブ・システムであって、R A I D-0、R A I D-1、R A I D-5、および R A I D-10 を含む複数の R A I D レベルのうちの少なくとも 1 つからの記憶空間を含む、ディスク・ドライブ・システム。

【請求項 17】

請求項 16 に記載のシステムであって、R A I D-3、R A I D-4、R A I D-6、および R A I D-7 を含む R A I D レベルを更に含む、システム。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願の相互参照：本願は、2006 年 5 月 24 日出願の米国仮特許出願第 60/808045 号に対する優先権を主張するものであり、これは参照により全体が本明細書に組み込まれる。

【0002】

本発明は、一般にはディスク・ドライブのシステムおよび方法に関し、より詳細には、R A I D パラメータまたはディスク構成に対する変更が行われたときに R A I D デバイスのレイアウトを最適化する動的ブロック・アーキテクチャ R A I D デバイスの管理、再配分、および再ストライピングを有するディスク・ドライブのシステムおよび方法に関する。

【背景技術】

【0003】

既存のディスク・ドライブ・システムは、仮想ボリューム（仮想ボリューム）が物理ディスクにわたって分散する（またはマッピングされる）ように設計されているが、それは、ボリューム作成時に決定され、仮想ボリュームの存続時間全体にわたって静的にとどまる様式とされる。即ち、ディスク・ドライブ・システムは、データ記憶空間の仮想ボリュームの特定の位置およびサイズに基づいて、データを静的に割り振る。仮想ボリュームが

10

20

30

40

50

所望のデータ記憶目的に対して不適當であることが判明した場合、既存のシステムは、ボリュームの特性を変更するために、新しい仮想ボリュームの作成と、それに付随しての、古い仮想ボリュームから新しい仮想ボリュームへの、以前に格納したデータのコピーとを必要とする。この手続きは、重複する物理ディスク・ドライブ空間を必要とするので、時間と費用とがかかる。

【0004】

こうした従来技術のディスク・ドライブ・システムは、データを格納するために、データ記憶空間の仮想ボリュームの厳密な位置およびサイズを認識し、監視し、制御することを必要とする。更に、システムはしばしば、より多くのデータ記憶空間を必要とし、それによりより多くのRAID Device (RAIDデバイス)が追加される。その結果、空にされたデータ記憶空間は使用されず、余分のデータ記憶装置、例えばRAIDデバイスが、システムでデータを格納、読取り／書込み、および／または復元するためにあらかじめ取得される。追加のRAIDデバイスは、費用がかかり、余分のデータ記憶空間が実際に必要となるまでは不要である。

10

【発明の概要】

【発明が解決しようとする課題】

【0005】

従って、改良型のディスク・ドライブのシステムおよび方法が求められており、より具体的には、効率的な動的なRAID空間および時間の管理のシステムが求められている。更に、サーバやホストのデータ・アクセスを損失することなく、また、弾性(resiliency)を損なうことなく、RAIDの管理、再割振り(reallocation)、および再ストライピング(リストライピング、restriping)を行うことを可能にする改良されたディスク・ドライブのシステムおよび方法が求められている。

20

【課題を解決するための手段】

【0006】

一実施形態では、本発明は、ディスク・ドライブ・システムでのRAIDリストライピングの方法である。この方法は、少なくとも1つのスコアに基づいて移行(migration)のための初期RAIDデバイスを選択すること、代替のRAIDデバイスを作成すること、初期RAIDデバイスに格納されたデータを代替RAIDデバイスへ移動すること、および初期RAIDデバイスを取り除くことを含む。スコアは、初期スコア、置換スコア、およびオーバーレイ・スコアを含むことができる。更に、この方法は、システムにより自動的に実施することができ、またはシステム管理者などにより手動で実施することができる。この方法は、周期的に、継続的に、RAIDデバイスの移行後ごとに、ディスク・ドライブの追加時に、および／またはディスク・ドライブを取り除く前などに、実施することができる。

30

【0007】

別の実施形態では、本発明は、RAIDサブシステムおよびディスク・マネージャを有するディスク・ドライブ・システムである。ディスク・マネージャは、RAIDサブシステムの各RAIDデバイスについてのスコアを自動的に計算し、RAIDデバイスの相対的スコアに基づいてサブシステムからRAIDデバイスを選択し、代替RAIDデバイスを作成し、選択したRAIDデバイスに格納されたデータの一部を代替RAIDデバイスへ移動し、選択したRAIDデバイスを取り除くように、構成される。

40

【0008】

更に別の実施形態では、本発明はディスク・ドライブ・システムであり、このディスク・ドライブ・システムは、各RAIDデバイスについて計算された少なくとも1つのスコアに基づいて移行のためのRAIDデバイスを選択する手段と、少なくとも1つの代替RAIDデバイスを作成する手段と、選択したRAIDデバイスに格納されたデータを少なくとも1つの代替RAIDデバイスへ移動する手段と、選択したRAIDデバイスを取り除く手段とを含む。

【0009】

50

複数の実施形態が開示されるが、本発明の例示的な実施形態を示し説明する以下の詳細な説明から、本発明の更に別の実施形態が当業者には明らかとなるであろう。理解するであろうが、本発明は、本発明の精神および範囲から逸脱することなく、様々な自明な態様における変更が可能である。従って、図面および詳細な説明は、例示的な性質のものとなされるべきであり、限定的なものとなされるべきではない。

【0010】

本明細書は、本発明を形成するとみなされる主題を具体的に指摘して明確に特許請求する特許請求の範囲で完結するが、添付の図と共に行われる以下の説明より、本発明がより良く理解され则认为られる。

【図面の簡単な説明】

10

【0011】

【図1A】図1Aは、本発明の一実施形態によるRAID構成を有するディスク・アレイである。

【図1B】図1Bは、追加のRAIDデバイスを有する図1Aのディスク・アレイである。

【図1C】図1Cは、RAIDデバイスを取り除いた後の図1Bのディスク・アレイである。

【図2A】図2Aは、本発明の別の実施形態によるRAID構成を有するディスク・アレイである。

【図2B】図2Bは、追加のRAIDデバイスを有する図2Aのディスク・アレイである。

20

【図2C】図2Cは、RAIDデバイスを取り除いて別のRAIDデバイスを追加した後の図2Bのディスク・アレイである。

【図2D】図2Dは、更に別のRAIDデバイスを取り除いた後の図2Cのディスク・アレイである。

【図3A】図3Aは、本発明の別の実施形態によるRAID構成を有するディスク・アレイである。

【図3B】図3Bは、RAIDデバイスの移行を示している図3Aのディスク・アレイである。

【図3C】図3Cは、複数のRAIDデバイスの別の移行を示している図3Bのディスク・アレイである。

30

【図3D】図3Dは、複数のRAIDデバイスの更に別の移行を示している図3Cのディスク・アレイである。

【図3E】図3Eは、新しいRAID構成の図3Aのディスク・アレイである。

【図4】図4は、本発明の一実施形態によるリストライピングのプロセスの流れ図である。

【図5】図5は、本発明の一実施形態による異なるサイズのRAIDデバイスを有するRAID構成を有するディスク・アレイである。

【発明を実施するための形態】

【0012】

40

本発明の様々な実施形態は、一般にはディスク・ドライブのシステムおよび方法に関し、より詳細には、1組のディスク・ドライブ上に構築される1または複数のRAIDデバイスにわたって分散する1または複数の仮想ボリュームを実装するディスク・ドライブのシステムおよび方法に関する。RAIDデバイスの管理、リアロケーション（Reallocation、再配分）、およびリストライピング（「Restriping」）は、サイズ、データ保護レベル、相対的成本、アクセス速度などの仮想ボリュームに関連する様々な特性を変更するシステムおよび方法を提供する。このシステムおよび方法は、ディスク構成（コンフィギュレーション）に対する変更が行われたときに、管理処置により又は自動的に開始することができる。

【0013】

50

本開示の様々な実施形態は、RAIDパラメータまたはディスクのコンフィギュレーションに対する変更が行われたときにRAIDデバイス・レイアウトを最適化する動的ブロック・アーキテクチャRAIDデバイス・リストラIPPINGを有する改良型のディスク・ドライブ・システムを提供する。一実施形態では、ディスクがシステムに追加されたときに、まずRAIDデバイスのレイアウトをリバランス（バランスの取り直し）することができる。リバランスすることにより、最大使用可能ディスク構成を使用することによりシステム内の仮想化性能を改善することができる。リストラIPPINGはまた、ディスクのグループからデータを移行する機能も提供することができ、アップタイム（使用可能時間）やデータ保護の損失なしに、こうしたディスクをシステムから取り除くことが可能となる。更に、リストラIPPINGは、RAIDパラメータを変更する機能を提供することができ、データが書き込まれた後であっても性能および／または記憶容量を調整する能力をユーザに与える。リストラIPPINGは更に、サーバやホストのデータ・アクセスを損なうことなく、また、弾性を損なうことなく、リストラIPPINGを行うことを可能にする改良型のディスク・ドライブのシステムおよび方法を提供することができる。

10

【0014】

本明細書に記載の様々な実施形態は、複数の方式で既存のディスク・ドライブ・システムを改善する。一実施形態では、仮想ボリュームと物理ディスク・ドライブ空間との間のマッピングが、細密なスケールで変更可能である。別の実施形態では、以前に格納されたデータを小単位で自動的に移行することができ、物理的資源を完全に複製する必要なしに適切なマッピングを更新することができる。別の実施形態では、適切な資源へ既にマッピングされている仮想ボリュームの部分を移行する必要はなく、ボリューム（Volume）の再構成（再コンフィギュレーション）に必要な時間が削減される。更に別の実施形態では、記憶システムが、仮想ボリュームのグループ全体を自動的に並列的に再構成する。更に、記憶システムは、物理的構成に対する変更が行われたときに仮想ボリュームを自動的に再構成することができる。従来のディスク・ドライブ・システムに勝る他の利点を当業者は理解するであろうし、そうした利点は、列挙されたものに限定されない。

20

【0015】

更に、リストラIPPINGおよびディスクの分類は、記憶システムの管理制御にとって強力なツールとなることができる。例えば、欠陥のある製造ロットからのものであることが判明しているディスク・ドライブが再分類されて、こうしたディスク・ドライブからの移行が行われるようにできる。同様に、1組のドライブを「予備（リザーブ）」カテゴリに保持し、後に、より大きい使用中グループの一部となるように再分類することができる。RAIDデバイスを広げるためのリストラIPPINGは、こうした追加の予備ユニットを徐々に組み込むことができる。本明細書に記載の実施形態により幾つかの利点を理解することができること、および先の例のリストは包括的なものではなく、限定を行うためのものではないことに、留意されたい。

30

【0016】

本明細書で様々な実施形態の説明では、「Volume（ボリューム）」は、コンピュータ・データを格納する外部アクセス可能なコンテナ（入れもの）を含むことができる。一実施形態では、コンテナを、相互接続プロトコルを介して、ブロックの連続するアレイとして提示することができる。別の実施形態では、各ブロックは、固定サイズ、伝統的には512バイトを有することができる。しかし、256バイト、1024バイトなどのような他のサイズのブロックを使用することもできる。典型的に、任意の所与の位置でデータに対して実施される、サポートされる操作は、「書込み」（格納）および「読取り」（検索）を含むことができる。しかし、「検証」などの他の操作もサポートすることができる。ボリュームにアクセスするために使用される相互接続プロトコルは、ディスク・ドライブにアクセスするために使用されるものと同じでよい。従って、ある実施形態では、ボリュームは、ディスク・ドライブのそれと概して同等に見え、機能することができる。伝統的に、ボリュームを、ディスク・ドライブのパーティション、またはアレイ内のディスク・ドライブの単純な連結として、実装することができる。

40

50

【0017】

本明細書で使用される「仮想ボリューム (Virtual Volume)」は、様々なハードウェア資源およびソフトウェア資源から構築されるデータを格納する外部アクセス可能コンテナを含むことができ、一般には従来型のボリュームの挙動を模倣する。具体的には、ディスク・ドライブ・アレイを含むシステムは、ディスク・アレイの非交差部分（交差していない部分）を使用する複数の仮想ボリュームを提示する。このタイプのシステムでは、個々のディスク・ドライブの記憶資源は、アレイで統合することができ、その後で、外部コンピュータで使用するために個々のボリュームに分割することができる。ある実施形態では、外部コンピュータはサーバ、ホストなどである。

【0018】

本明細書で使用される「RAID Device (RAIDデバイス)」は、ディスク・パーティションの集合を含むことができ、連結と、ディスク・ドライブ障害に対する弾性 (resiliency) とを提供する。連結および弾性のためのRAIDアルゴリズムは周知であり、RAID0、RAID1、RAID0+1、RAID5、RAID10などのRAIDレベルを含む。所与のディスク・アレイでは、複数のRAIDデバイスが、任意の所与のディスク組に常駐することができる。こうしたRAIDデバイスのそれぞれは、異なるRAIDレベルを使用しうるものであり、異なるパラメータ、例えばストライプ・サイズなど、を有し得るものであり、異なる順序で個々のディスク・ドライブにわたって分散し得るものであり、ディスク・ドライブの異なるサブセットを占有し得るものなどである。RAIDデバイスは、内部アクセス可能な仮想ボリュームであり得る。それは、固定サイズのデータ記憶位置の連続するアレイを提供することができる。特定のRAIDパラメータが、RAIDデバイスアドレスと、ディスク・ドライブのデータ記憶アドレスとの間のマッピングを決定する。本開示では、改良された機能を提供する、RAIDデバイスから外部アクセス可能仮想ボリュームを構築するためおよび変更するためのシステムおよび方法が説明される。

【0019】

仮想ボリューム構築

本開示を使用する記憶システムは、最初に、ディスク・アレイ上に様々な特性を有する1組のRAIDデバイスを構築することができる。RAIDデバイスは、本明細書で「ページ」と呼ぶ単位に論理的に分割することができ、ページのサイズは多数のブロックであり得る。典型的なページ・サイズは4096ブロックでよい。しかし、原理上は、1ブロック以上の任意のページ・サイズを使用することができる。しかしながら、ページ・サイズは一般に、2のべき乗のブロック数を含む。こうしたページを仮想ボリューム管理ソフトウェアにより管理することができる。最初に、各RAIDデバイスからのすべてのページをフリーとマークすることができる。必要に応じて、ページを仮想ボリュームに動的に割り振ることができる。即ち、所与のアドレスが最初に書き込まれると判定されたときにページを割り振ることができる。書き込まれる前に読み取られるアドレスには、デフォルト・データ値を与えることができる。仮想ボリューム管理ソフトウェアは、仮想ボリュームアドレスと、RAIDデバイス内のページとの間のマッピングを維持することができる。所与の仮想ボリュームを複数のRAIDデバイスからのページで構成することができ、これは異なる特性を更に有し得ることに留意されたい。

【0020】

このようにして構築された仮想ボリュームのサイズを拡張することは、サーバへ提示されるアドレスの範囲を増大させることにより実施することができる。アドレスからページへのマッピングは、以前に使用可能であったアドレス範囲と拡張されたアドレス範囲との両方で、同じallocate-on-write (アロケート・オン・ライト) 方式で継続することができる。

【0021】

所与の仮想ボリュームの性能および弾性の特性は、大部分は、その仮想ボリュームに割り振られたページの集団的挙動により決定することができる。ページは、その特性を、そ

れらが構築されるRAIDデバイスおよび物理ディスク・ドライブから継承する。従って、一実施形態では、RAIDデバイス間でのページの移行は、サイズ以外の仮想ボリュームの特性を変更するために行う。本明細書で使用される「移行(Migration)」は、新しいページを割り振ること、以前に書き込まれたデータを旧ページから新ページへコピーすること、仮想ボリューム・マッピングを更新すること、および旧ページをフリーとマークすることを含むことができる。伝統的には、RAIDデバイスの特性を変換して(即ち、新しいRAIDレベル、ストライプ・サイズなどへ再マップする)それと同時にデータを同じ場所に残すことは、できない。

【0022】

異なる仮想ボリューム特性を生成するように変更することのできる幾つかの独立パラメータが存在する。シナリオのうちの幾つかが本明細書で述べられる。しかしながら、本明細書で詳細に説明されるシナリオは本開示の様々な実施形態の例であり、限定を意図するものではない。ある実施形態では、本開示は、こうしたパラメータの何れかまたは全ての同時変更を含むことができる。

【0023】

RAIDパラメータ変更

例示のために、5個のディスク102、104、106、108、110を含むディスク・アレイ100が図1Aに示されている。本明細書で開示される様々な実施形態に従って任意の数のディスクを使用することができ、例示的な5個のディスク・システムは、一実施形態を説明するために無作為に選ばれたものであることを理解されたい。最初に、2つのRAIDデバイス、例えばRAIDデバイスA 112およびRAIDデバイスB 114をディスク・アレイ上に構築することができる。残りの空間が存在するならば、それは未割り振りおよび未使用でよい。この場合も、本明細書で開示される様々な実施形態に従って任意の数のRAIDデバイスを使用することができ、例示的な2個のRAIDデバイスは、一実施形態を説明するために無作為に選ばれたものであることを理解されたい。複数の仮想ボリュームを、RAIDデバイスに含まれるページから構築することができる。所与の仮想ボリュームの特性を変更することが望まれる場合、残りの空間内に追加のRAIDデバイスを構築することができ、前述のように、適切なページを移行することができる。

【0024】

図1Bは、新しいRAIDデバイス、例えばRAIDデバイスC 116を作成した後のRAID構成の一実施形態を示し、RAIDデバイスA 112からのデータの移行を示す。RAIDデバイスC 116は、RAIDレベル、ストライプ・サイズ、または他のRAIDパラメータなどの点で、RAIDデバイスA 112とは異なり得る。ある実施形態では、同じパラメータを有するがディスク・ドライブ上の位置が異なるRAIDデバイスへ単に再配置することにより、性能が向上する可能性がある。例えば、ディスク・ドライブの性能は、物理プラッタの内側と外側とでは異なるものであり得、また、すべてのデータが密に配置される場合にはヘッド・シーキング(head seeking)のための時間が削減され得る。

【0025】

移行が完了したとき、RAIDデバイスA 112を取り除くことができ、図1Cに示される例示的RAID構成が残される。

【0026】

図1Aのものから図1Cのものへの例示的RAIDの再構成は、仮想ボリュームの幾つかの部分の移動できることも示す。即ち、一実施形態では、ボリューム全体を移動するのではなく、1または複数の仮想ボリュームの幾つかの部分の移行することができる。このことを実施することができるのは、単一の仮想ボリュームを複数のRAIDデバイスにわたって割り振ることができるからである。同様に、この例示的構成は、仮想ボリュームの幾つかのグループを移動できることも示す。なぜなら、あるRAIDデバイスから別のRAIDデバイスへ移行されたページを複数の仮想ボリュームに割り振ることができるから

である。

【0027】

ディスク・ドライブの追加

5個のディスク202、204、206、208、210を含むディスク・アレイ200を有する別の実施形態が図2Aに示されており、図2Aでは、2つの追加のディスク・ドライブ212、214が既存の構成に追加されている。本明細書で開示される様々な実施形態に従って任意の数のディスクを使用することができ、例示的な5個のディスク・システムは、一実施形態を説明するために無作為に選ばれたものであることを理解されたい。同様に、本明細書で開示される様々な実施形態に従って任意の数のディスクを追加することができ、例示的な2個の追加ディスクは、一実施形態を説明するために無作為に選ばれたものであることを理解されたい。ある実施形態では、システムを再構成し、RAIDデバイスをすべての7個のディスクにわたって分散させることが望ましいことがある。しかしながら、再構成されたRAIDデバイスを使用可能なすべてのディスクにわたって分散させる必要はないことを理解されたい。RAIDデバイスが複数のディスクにわたって分散する一実施形態では、システムの合計スループットは、より多くのハードウェアを並列に使用することにより向上させることができる。更に、RAIDデバイスのレイアウトの制約の結果として、追加のディスクの使用がより効率的となり得る。具体的には、RAID5は典型的に、弾性を実現するために、必要とする独立ディスクの数が最小限であり得る。一般に直面するRAID5の実装は、必要とするディスクが最小5個である。従って、RAIDデバイスA 216とRAIDデバイスB 218との両方から、例えばすべての7個のディスクにまたがる適切な置換手段へページを移行することが望ましいことがある。再構成のための可能なシーケンスが図2B～図2Dに示されている。

10

20

【0028】

このシーケンスでは、より広範囲のRAIDデバイスC 220を作成することができ、RAIDデバイスA 216からのデータをRAIDデバイスC 220へ移行することができる。次いで、RAIDデバイスA 216を削除することができ、RAIDデバイスD 222を作成することができる。RAIDデバイスD 222を使用して、以前にRAIDデバイスB 218に含まれていたデータを再配置することができる。

【0029】

それを行う際に、元のディスク・ドライブ上での必要な余分の空間は、RAIDデバイスC 220を作成するために必要な空間だけである。例示的な一実施形態では、他のRAIDパラメータが変化しない場合には、RAIDデバイスC 220の各範囲は、RAIDデバイスA 216を構築する際に使用される範囲サイズの5/7のサイズでよい（即ち、RAIDデバイスCは、5個の初期ドライブ+2個の追加のドライブ、の間で分散する）。

30

【0030】

プロセスは完全に可逆にすることができ、また、システムから1または複数のディスク・ドライブを取り除くのに使用することができ、例えばディスク212および214を図2Dの例示的構成から取り除くことが望まれた場合などに使用することができることに、留意されたい。同様に、複数の初期RAIDデバイスを単一のRAIDデバイスへ、またはより少数のRAIDデバイスへ、移行することができる（例えば図3Cを参照）。更に、単一の初期RAIDデバイスを複数の新しいRAIDデバイスへ移行することができる（例えば図3Bを参照）。

40

【0031】

図2A～図2Dを参照しながら説明した先の一実施形態の例は、元のディスク・セット上に未使用空間が存在するときに、追加の各ディスクにわたってデータを移行できることを示す。しかしながら、ある実施形態では、RAIDデバイスを移行すること、取り除くことなどのためのディスク空間が不十分であり得る。それでも、データを追加のディスクへ移行することが可能であり得る。そのような場合、ディスク空間を再割振りして、移動を実施するために必要な余分の空間を設けることができる。以下でより詳細に説明するR

50

A I DデバイスのR e p l a c e m e n t S c o r e (置換スコア)が初期S c o r e (スコア)よりも高い場合、等しいサイズの永続的R A I Dデバイスを割り振ることができる。追加の判定は不要である。以下で詳細に説明するR A I DデバイスのO v e r l a y S c o r e (オーバーレイ・スコア)が初期スコアよりも高い場合、一時的空間を使用する。このプロセスは、4個のディスク302、304、306、308を含むディスク・アレイ300を有する一実施形態について、図3A～図3Eを参照しながら詳述され、そこでは多段移行が使用される。3個のディスク310、312、314がディスク・アレイ300に追加されている。本明細書で開示される様々な実施形態に従って任意の数のディスクを使用することができ、例示的な4個のディスク・システムは、一実施形態を説明するために無作為に選ばれたものであることを理解されたい。同様に、本明細書で開示される様々な実施形態に従って任意の数のディスクを追加することができ、例示的な3個の追加ディスクは、一実施形態を説明するために無作為に選ばれたものであることを理解されたい。

10

【0032】

図3Aに示されるシステムを再構成して、アレイ内のすべての使用可能なディスク・ドライブを使用する方策は、一時的R A I Dデバイス(1または複数)を作成すること、および、例えば、R A I DデバイスC 320からのデータを、元のディスク・ドライブから、例えば、一時的R A I DデバイスD 322およびE 324へ移行することを含む。一時的R A I Dデバイスは、元のディスク・セットが容量限界または容量限界近くにある場合に、使用する。代替の実施形態では、一時的空間を、R A I Dデバイスとして割り振る必要はなく、また、データを適切に保持するための技術分野で周知の方式で 사용할ことができる。同様に、2つの一時的R A I DデバイスD 322およびE 324が示されているが、より少数または多数の一時的R A I Dデバイスを使用できることを理解されたい。

20

【0033】

一実施形態では、データ・プログレッション・プロセス(data progression process)が、初期R A I Dデバイスと、一時的R A I Dデバイス(1または複数)との間、または他のケースでは、新しい永続的R A I Dデバイス(1または複数)との間のデータの移動を管理することができる。別の実施形態では、リストラIPPINGは、使用可能であれば同じR A I Dレベルを使用することを試みる。別の他の実施形態では、リストラIPPINGは、データを異なるR A I Dレベルへ移動する。

30

【0034】

一時的R A I Dデバイスのサイズは、初期R A I Dデバイスサイズと、ページ・プール(page pool)内の使用可能な空間とに依存する。一時的R A I Dデバイスのサイズは十分な空間を提供することができ、初期R A I Dデバイスが削除されたときにページ・プールが引き続き正常に動作することができ、より多くの空間を割り振ることがない。ページ・プールは、ページ・プールのサイズに基づいて、構成されたしきい値で、より多くの空間を割り振ることができる。

【0035】

データがR A I DデバイスC 320から移行された後、R A I DデバイスC 320を削除することができ、それにより、すべてのディスク・ドライブにまたがる新しいR A I Dデバイス、例えばR A I DデバイスX 326のための空間が提供される。R A I DデバイスC 320を削除することは、R A I DデバイスC 320が消費したディスク空間をディスクの自由空間へ戻す。この時点で、ディスク・マネージャは、隣接する自由空間割振りを単一のより大きな割振りとして組み合わせて、断片化を低減することができる。R A I Dデバイスを削除することは、以前に使用可能であったよりも多くの自由空間を多数のディスクにわたって作成することができる。より高いS c o r eを有するR A I Dデバイスをこの自由空間のスライスから作成することができる。

40

【0036】

初期R A I DデバイスC 320が削除された後、図3Cに示されるように、リストラ

50

イピングが置換RAIDデバイスX 326を作成する。一実施形態では、置換RAIDデバイスX 326は、可能な限り多くのディスクを使用して、仮想化の利点を最大にすることができる。リストライピングは、初期RAIDデバイスよりも大きいRAIDデバイスを割り振ることを試みる。別の実施形態では、リストライピングは、これを、置換スコアまたはオーバーレイ・スコアを初期スコアで割り、初期RAIDデバイスのサイズを掛けた計算値を使用して、行うことができる。これは、ディスク当たりのディスク空間の使用量が以前と同じであるRAIDデバイスを作成することができ、ディスクの断片化を低減することができる。

【0037】

初期RAIDデバイス、例えばRAIDデバイスA 316、B 318、およびC 320のサイズを慎重に制限することにより、RAIDデバイスB 318およびE 324からのすべてのデータを保持することができるようRAIDデバイスX 326を作成することが可能となり、図3Eで最終的構成が達成されるまでプロセスが継続することが可能となる。即ち、RAIDデバイスY 328を作成することができ、RAIDデバイスA 316およびD 322をRAIDデバイスY 328へ移行することができ、RAIDデバイスA 316およびD 322を削除することができる。

【0038】

一時的RAIDデバイス（1または複数）、例えばRAIDデバイスD 322およびE 324が作成されて一時的とマークされた場合、RAIDデバイスは、図3C～Eに示されるように、取外しのためにマークされ得る。一実施形態では、各サイクルの一部として、一時的RAIDデバイスを取り除くことができる。より大きい置換RAIDデバイスが作成されると、必要とされる一時的空間の量を低減する。ある実施形態では、RAIDデバイスの移行が、データを移行または除去するための一時的空間の割り振りを必要としないことがあることに、留意されたい。

【0039】

リストライピングの一実施形態では、一時的RAIDデバイスの取外しは、データの移動や一時的RAIDデバイスの削除などのような、初期RAIDデバイスの移行または取外しのために使用されるステップのサブセットを使用する。

【0040】

一実施形態では、一時的RAIDデバイスのスコアが初期RAIDデバイスのスコアを超過した場合、一時的RAIDデバイスを永続的RAIDデバイスとみなすことができる。即ち、これは、RAIDデバイスを移動するプロセスの一部として自動的に削除されない。別の実施形態では、一時的RAIDデバイスは、一時的RAIDデバイスが初期RAIDデバイスよりも十分高いスコアを有する場合にのみ、保持される。

【0041】

リストライピングは、元の低スコアのRAIDデバイスを取り外し、それを新しい高スコアのRAIDデバイスと置き換えるための幾つかの別のステップを含むことができる。例えば、リストライピングは、システム内のディスクが一杯であり別のRAIDデバイスのための空間を有さないという可能性に対するものである。リストライピングは、RAIDデバイスをリストライピングする試みの前に過剰な空間を削除（トリミング）する。過剰な空間をトリミングすることは、追加のディスク空間を解放することができ、リストライピングの成功率を向上させる。

【0042】

ある実施形態では、リストライピングがデッドロックに達する可能性がある。例えば、一時的空間のサイズが、初期RAIDデバイスを移動するために必要な空間の一部を消費する。すべてのページが解放できないものであるためにRAIDデバイスを取り除くことが不可能となった場合、RAIDデバイスを故障とマークすることができ、リストライピングは、移動することができるか又は移動すべきである次のRAIDデバイスへ移る。

【0043】

図4を参照しながら、リストライピングのプロセス400の一実施形態のフロー・チャ

10

20

30

40

50

ートが説明される。図4は一実施形態を示すものであり、本開示に従って様々な代替実施形態およびプロセスを使用できることを理解されたい。まず、ステップ402および404に示されるように、リストラIPPINGは、移行、取外しなどをすべきRAIDデバイス、または移行、取外しなどをすることのできるRAIDデバイスがあるかどうかを判定する。一実施形態では、リストラIPPINGは、システム内のRAIDデバイスのすべてをチェックし、最低の相対的スコアを有する最小のRAIDデバイスを選択する。他の実施形態では、リストラIPPINGは他のRAIDデバイスを選択することができ、本明細書に記載のリストラIPPINGは、最小のRAIDデバイスおよび／または最低のスコアを有するRAIDデバイスを選択することに限定されない。しかしながら、一般に、最小のRAIDデバイスの移動が失敗した場合、より大きいRAIDデバイスでも成功しない可能性が高い。一実施形態では、最低スコアのRAIDデバイスは、置換スコアまたはオーバーレイ・スコアを初期スコアで割ることにより、求めることができる。最低スコアのRAIDデバイスを求める他の方法は本開示に従い、それには、RAIDデバイスの初期スコアだけを使用することを含む。

10

【0044】

移行または取外しのためのRAIDデバイスを識別することに加えて、図4に示されるように、RAIDデバイス移動は、限定はしないが、一時的空間を割り振ること（ステップ406）、データを移動すること（ステップ408）、元のRAIDデバイスを削除すること（ステップ410）、新しいRAIDデバイスを割り振ること（ステップ412）、および／または一時的RAIDデバイスを削除すること（ステップ414および416）などのような、RAID構成を最適化するための複数のステップを含むことができる。上記の列挙した追加のステップの例示的なものであり、RAIDデバイスの移動は、列挙された各ステップを必要とするわけではなく、また、ある実施形態では、列挙されたステップとは別のステップまたは異なるステップを含むことができる。例えば、ある実施形態では、一時的空間を使用できず、従って、一時的空間を割り振ることができない。

20

【0045】

ある実施形態では、リストラIPPINGは、RAIDデバイスの移動を制限することができる。例えば、システムをスラッシングさせることを回避するために、リストラIPPINGは、RAIDデバイスのスコアを絶対的に最大にする必要はない。リストラIPPINGはまた、故障したRAIDデバイスを再試行しないように、それをマークすることができる。

30

【0046】

リストラIPPINGは、新しいディスクを認識し、追加の空間を使用する新しいRAIDデバイスを作成し、それに応じてデータを移動することができる。プロセスが完了した後、ユーザ・データおよび自由空間を、初期ディスクおよび追加のディスクを含むディスク・ドライブ全体にわたって分散させることができる。リストラIPPINGはRAIDデバイスを拡張するのではなく、RAIDデバイスを置換することに、留意されたい。任意の所与のディスク上の自由空間およびユーザ割振りの配置は任意でよく、図1～図3ならびに残りの図5に示される配置は例示のためのものであることを理解されたい。

【0047】

リストラIPPINGのためのRAIDデバイスの選択

40

一実施形態では、先に論じたように、リストラIPPINGは以下のものを処理することができる。

・ドライブの追加： 追加のドライブがディスク・ドライブ・システムに追加されたとき、リストラIPPINGは、最適以下の数のドライブを使用するRAIDデバイスを識別することができる。新しいRAIDデバイスを作成することができ、データを移動することができる。元のRAIDデバイスは除去することができる。

・ドライブの取外し： リストラIPPINGは、ディスク・ドライブが取外しのためにマークされている場合、それを検出することができる。こうしたドライブ上に常駐するRAIDデバイスは取外しの候補となることができ、取外しは、ドライブの追加に関するものとはほぼ同様に実施することができる。

50

・ R A I D パラメータの変更： 性能を改善するために、R A I D レベル、ストライプ内のディスク数、範囲サイズなどの R A I D パラメータをユーザが変更することができる。リストライピングは、所望されるパラメータと初期パラメータとを比較し、最適ではない R A I D デバイスを移行および／または取外しのために選択することができる。

【0048】

より大規模の、より複雑なシステムを有する実施形態を含むある実施形態では、所望される最終的構成を得るために何れの移行操作を使用すべきかや、また、既存の資源内で初期構成から最終的な所望される構成にすることが可能かどうか、明らかでない可能性がある。一実施形態では、スコアリングおよび最適化の技法を使用して、取外しおよび置換のための特定の R A I D デバイスを選択することができる。例示的な実施形態では、スコアリング機能が、以下の特性のうちの 1 または複数のものを使用する。

【0049】

- ・ より多くのディスク・ドライブにまたがる R A I D デバイスが好ましい。
- ・ 均質な 1 組のディスク・ドライブ上に構築される R A I D デバイスが好ましい。ディスク・ドライブは、この機能をサポートするために分類することができ、同じカテゴリに属するように同一である必要はない。
- ・ 所望される最終的構成のパラメータ（R A I D レベル、ストライプ・サイズ、ディスク領域など）と合致する R A I D デバイスが好ましい。
- ・ 物理的に異なるディスク・ドライブに冗長データを配置する R A I D デバイスが好ましい。一例は、別々の筐体内のディスク・ドライブ、別々の通信経路上のディスク・ドライブ、または独立した電源を有するディスク・ドライブを含むことができる。

【0050】

別の実施形態では、リストライピングを、スコアリング、検査、移動などのような 3 つの構成要素に分割することができる。R A I D デバイスのスコアリングを使用して、要求されたパラメータおよび使用可能なディスク空間に基づいて所与の R A I D デバイスの品質を求めることができる。一実施形態では、スコアリングは 3 つの値を生成することができる。リストライピングは、初期 R A I D デバイスに関するスコアと、本明細書で置換スコアおよびオーバーレイ・スコアと呼ぶ 2 つの可能な代替 R A I D デバイスのスコアとを提供する。一実施形態に関する各スコアの詳細が以下で説明される。

【0051】

- ・ スコア： R A I D デバイスの現状態でのスコア。可能な一実施形態では、スコアは、R A I D デバイスで使用されるディスク数から断片化およびパラメータの問題を引いたものであり得る。例えば表 1 を参照されたい。代替の実施形態では、任意の適切な方式でスコアを計算することができる。
- ・ 置換スコア： 存在する自由空間から構築することのできる R A I D デバイスの最大スコア。置換スコアは、R A I D デバイスのスコアよりも高い場合も、低い場合も、等しい場合もある。
- ・ オーバーレイ・スコア： 現 R A I D デバイスが取り外される場合の R A I D デバイスの最大スコア。オーバーレイ・スコアは、R A I D デバイスのスコアよりも高い場合も、低い場合も、等しい場合もある。ある実施形態では、ディスク・フル状態などにおいて、現 R A I D デバイスにより既に割り振られている空間を使用してより良好な R A I D デバイスを作成できるかどうかを判定するために、オーバーレイ・スコアが望ましい。

【0052】

置換スコアおよびオーバーレイ・スコアに関して、R A I D デバイスに関するユーザ・アクセス可能ブロックは、ディスク変更の数と同じにとどまる。3 つのスコアは、低スコアの R A I D デバイスから高スコアの R A I D デバイスへ移行する方策を開発するための入力パラメータを提供することができる。特定の実施形態では、置換スコアが初期スコアよりも高い場合、図 1 および図 2 に記載されているような直接的な移行が可能である。オーバーレイ・スコアが初期スコアよりも良好であり、かつ十分な自由空間が一時的 R A I D デバイス（1 または複数）に対して使用可能である場合、図 3 で略述される移行方策が

可能である。

【0053】

一実施形態では、スコアを求めるのに使用されるファクタは、以下のうちの1または複数のものを含むことができる。

【0054】

・ディスク・フォルダ： R A I Dデバイスが指定のフォルダの外部のディスクを使用する場合、R A I Dデバイスのスコアが低くされる。この状況は、管理処置のために生じ得るものであり、また、スペアリングの間、即ち、スペア（予備）のディスクが供給されるときに生じ得る。

・ディスク・タイプ： R A I Dデバイスが誤ったタイプのディスクに常駐する場合、R A I Dデバイスのスコアが低くされる。この状況は、ディスクが故障し、適切なタイプの予備が使用可能ではない場合に生じ得る。そのような場合、冗長性は「タイプ純度（type purity）」にわたって維持されることができ、別のディスク・タイプの予備が使用されることができる。

・使用されるディスク数： 一般には、より広範囲のR A I Dデバイス（例えば、比較的多数のディスク・ドライブにまたがるR A I Dデバイス）が好ましく、そのR A I Dデバイスにはより高いスコアを与えられる。別の実施形態では、より独立性の高い障害領域を確立するために、最大幅が考慮され得る。

・使用可能なディスク数： これは、R A I Dデバイスの置換スコアおよびオーバーレイ・スコアを求めるために使用される。一実施形態では、予備ディスクを除くディスクが、置換R A I Dデバイスを割り振るための十分な空間があるかをチェックされる。他の実施形態では、予備ディスクが、スコアの決定において含まれる。空間が存在する場合、置換スコアおよびオーバーレイ・スコアが増加される。

・ディスク断片化： R A I Dデバイスの周りに空きディスク空間が存在する場合、R A I Dデバイススコアは低くされる。これは、R A I Dデバイスを移行することによりディスク断片化が低減され得ることを、リストラIPPINGに対して示すものである。

・R A I Dレベル： 所望されるR A I Dレベルが初期R A I Dレベルに合致するかどうか。これは、ストライプ内のディスクの数を含むことができる。

・R A I D範囲サイズ（RAID Extent Size）： 所望されるR A I Dデバイスの範囲サイズ、即ち、ディスク当たりのブロックが、初期R A I Dデバイスの範囲サイズと合致するかどうか。一実施形態では、R A I D範囲サイズに基づく決定は、R A I Dデバイスのスコアを低くすることができるだけである。

【0055】

表1は、使用することのできるスコアリング要素の例示的な実施形態を示す。表1に示されるように、変数は、D i s k s I n C l a s s（クラス内ディスク）、D i s k s I n F o l d e r（フォルダ内ディスク）、R A I D L e v e l（R A I Dレベル）、R A I D R e p e a t F a c t o r（R A I D反復ファクタ）、R A I D E x t e n t S i z e（R A I D範囲サイズ）、およびR A I D D r i v e s i n S t r i p e（ストライプ内R A I Dドライブ）を含むことができる。例示的なスコアリング・ファクタにおいて使用される「クラス内ディスク」は、以下の式で求めることができる。

【0056】

$$(D i s k s I n C l a s s - 3 \times D i s k s O u t O f C l a s s) \times D i s k s I n C l a s s C o n s t a n t$$

（（クラス内ディスク－3×クラス外ディスク）×クラス内ディスク定数）

【0057】

この式では、D i s k s I n C l a s sは、適切なクラスにあるR A I Dデバイスで 사용되는ディスクの数であり、D i s k s O u t O f C l a s sは、適切なクラスではないR A I Dデバイスで 사용되는ディスクの数であり、D i s k s I n C l a s s C o n s t a n tは乗算定数値である。ディスク・クラスは、限定はしないが、15 K F C、

10

20

30

40

50

10 K FC、SATAなどを含むことができる。例えば、RAIDデバイスが、10 K FCディスクを使用するようにサポートされるが、2つのSATAディスクを含んでいた場合、DisksOutOfClassに関する値は2となる。例示的なスコアリング・ファクタで使用されるDisksInFolderは、以下の式で求めることができる。

【0058】

$$(DisksInFolder - 3 \times DisksOutOfFolder) \times DisksInFolderConstant$$

((フォルダ内ディスク - 3 × フォルダ外ディスク) × フォルダ内ディスク定数)

【0059】

上記の式では、DisksInFolderは、ディスクの適切なフォルダ内にあるRAIDデバイスで使用されるディスクの数であり、DisksOutOfFolderは、ディスクの適切なフォルダ内にはないRAIDデバイスで使用されるディスクの数であり、DisksInFolderConstantは乗算定数値である。ディスク・フォルダは、RAIDデバイスで何れのディスクを使用することができるかを編成することができる。任意の時間にディスクをフォルダ・オブジェクト内に移動したり、フォルダ・オブジェクトの外に移動して、その使用法を変更することができる。例示的なスコアリング・ファクタで使用されるRAID Level (RAIDレベル)は、ディスクが望ましくないRAIDレベルである場合に0である。RAID Repeat Factor (RAID反復ファクタ)、RAID Extent Size (RAID範囲サイズ)、およびRAID Drives in Stripe (ストライプ内RAIDドライブ)は、それぞれの計算スコアを2で割ったものである。表1は、例示的なスコアリング・ファクタの一実施形態と、スコアリング・ファクタがどのように計算および使用されるかについての一実施形態とを示すことが、認識される。表1に示される例は、例示のために過ぎず、限定のためのものではない。任意のスコアリング・ファクタ、またはスコアリング・ファクタのグループを、本明細書で開示される様々な実施形態と共に使用することができる。更に、スコアリング・ファクタまたはスコアリング・ファクタのグループを任意の適切な方式で計算したり使用したりすることができる。

【0060】

【表1】

表1：例示的RAIDスコアリング・ファクタ

変数	スコア
クラス内ディスク	$(DisksInClass - 3 \times DisksOutOfClass) \times DisksInClassConstant$
フォルダ内ディスク	$(DiskInFolder - 3 \times DisksOutOfFolder) \times DisksInFolderConstant$
RAIDレベル	誤ったタイプの場合は0
RAID反復ファクタ	(計算されたスコア) ÷ 2
RAID範囲サイズ	(計算されたスコア) ÷ 2
ストライプ内RAIDドライブ	(計算されたスコア) ÷ 2

【0061】

別の実施形態では、リストライピングは、RAIDデバイスのスコアを検討して、RAIDデバイスが存在するならばどのRAIDデバイスを移動するかを判定する。リストライピングは、置換スコアまたはオーバーレイ・スコアのどちらかよりも低いスコアを有するRAIDデバイスを移動することができる。即ち、一実施形態では、置換スコアおよび/またはオーバーレイ・スコアが初期RAIDデバイススコアよりも大きい場合、そのRAIDデバイスは移動の候補とすることができる。他の実施形態では、初期RAIDデバイススコアが置換スコアおよびオーバーレイ・スコアよりも高い状況を含めての、任意の他の手段により、またはユーザによる手動選択などにより、初期RAIDデバイスを移行

の対象として選択することができる。リストライピングはまた、RAIDデバイスを移動すべきでないと判定することもできる。別の実施形態では、リストライピングは、使用可能なRAIDデバイスから単一のRAIDデバイスを選んで移行することができる。

【0062】

リストライピングが、移動するRAIDデバイスを識別した場合、そのRAIDデバイスの移行が行われる。一実施形態では、移行は、必要な一時的空間を決定すること、RAIDデバイスからデータを移動すること、初期RAIDデバイスのクリーンアップを行うこと、および一時的空間を除くことを含むことができる。別の実施形態では、動的ブロック・アーキテクチャ・ページ・プールが、RAIDデバイスを使用し、低いスコアのRAIDデバイスから高いスコアのRAIDデバイスへのデータの移動を処理することができる。

10

【0063】

別の実施形態では、リストライピングは更に、RAIDデバイス移行後毎にすべてのRAIDデバイスのスコアを再評価する。なぜなら、ディスク空間の再割振りが他のRAIDデバイスのスコアを変更する可能性があるからである。別の実施形態では、すべてのRAIDデバイスのスコアを周期的に計算することができる。ある実施形態では、リストライピングは、RAIDデバイスのスコアを継続的に計算することができる。更に別の実施形態では、スコアにおける最大の増加を使用して、取外しおよび置換のためのRAIDデバイスを選択することができる。ヒステリシス機構を使用して、プロセスが巡回的(cyclic)となるのを防ぐことができる。

20

【0064】

RAIDデバイスのスコアリングは、異なるサイズのディスク・ドライブに対処することもできる。図5は、異なるサイズのディスク502、504、506、508を有する例示的な構成500を示す。表2は、構成500に基づくRAIDデバイス510、512、514、516、518に関するスコアリング情報を含む、図5に示される構成に関する例示的なRAIDデバイスのスコアリングを示す。話を簡単にするために、相対的数字が使用される。表2は、移行のためのRAIDデバイス候補に関するスコアを示すが、表2は限定のためのものではなく、任意のスコアリングの組合せの結果として、RAIDデバイスに対して移行または移行なしとマークすることができる。

【0065】

30

【表2】

デバイス	スコア	置換	オーバーレイ	再ストライプ
P510	4	0	4	いいえ、最大
Q512	4	0	4	いいえ、最大
R514	2	2	4	はい、最大の50%
S516	3	1	4	はい、最大の75%
T518	2	2	2	いいえ、最小、より小さい ディスクに空間はない

【0066】

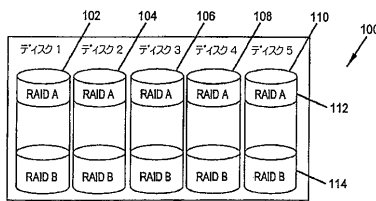
40

上記の説明および図面により示し説明した特定の実施形態は例示のためのものに過ぎず、本発明の範囲を限定することを意図するものではないことを理解されよう。本発明の精神または不可欠な特徴から逸脱することなく、本発明を他の特定の形態で実施できることを当業者は理解されよう。特定の実施形態の詳細に対する参照は、本発明の範囲を限定することを意図するものではない。

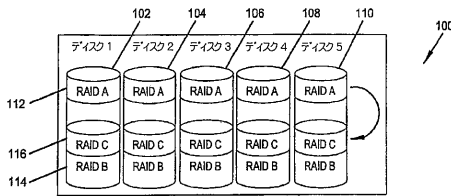
【0067】

本発明を好ましい実施形態を参照しながら説明したが、本発明の精神および範囲から逸脱することなく、形態および細部の変更を行えることを当業者は理解されよう。

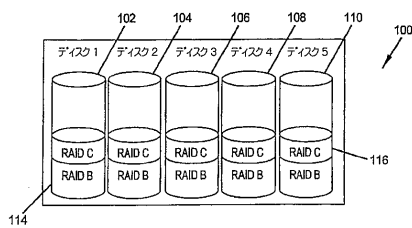
【図 1 A】



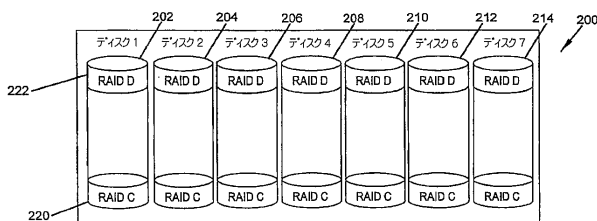
【図 1 B】



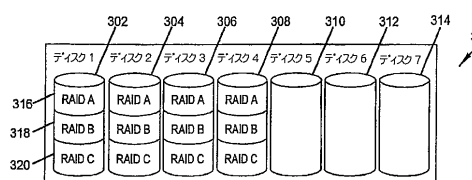
【図 1 C】



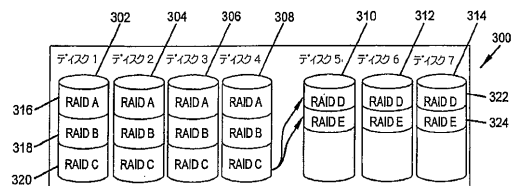
【図 2 D】



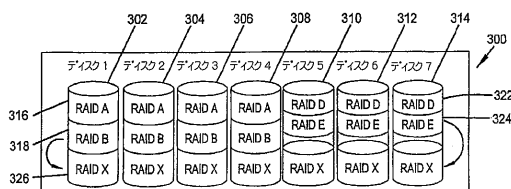
【図 3 A】



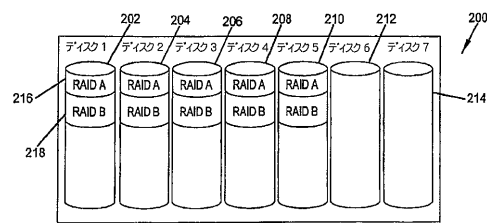
【図 3 B】



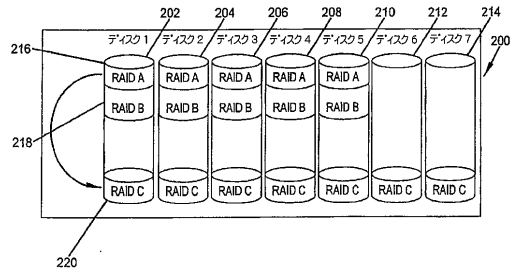
【図 3 C】



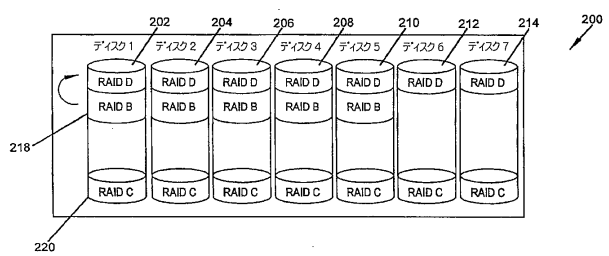
【図 2 A】



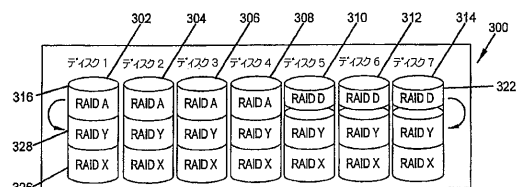
【図 2 B】



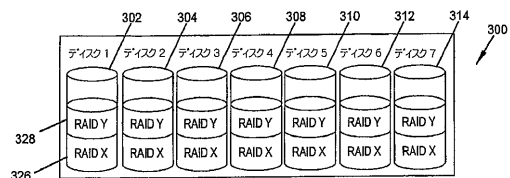
【図 2 C】



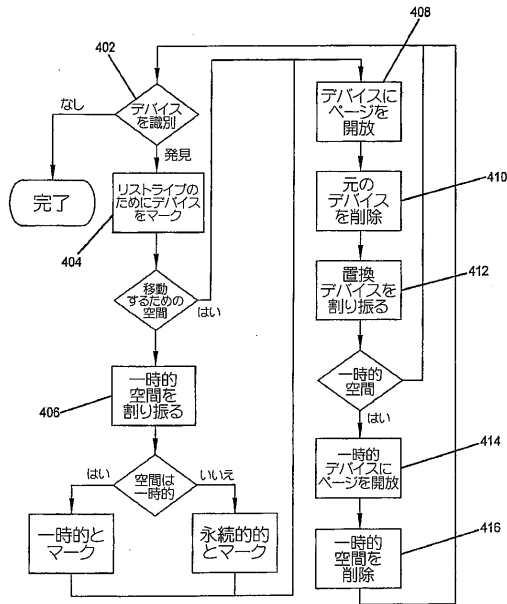
【図 3 D】



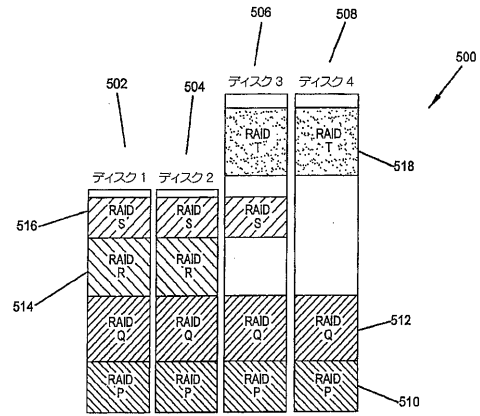
【図 3 E】



【図 4】



【図 5】



フロントページの続き

(74)代理人 100096068

弁理士 大塚 住江

(72)発明者 クレム, マイケル・ジェイ

アメリカ合衆国ミネソタ州55305, ミネトンカ, リヴェンデル・レイン 2301

(72)発明者 ウッターマーク, マイケル・ジェイ

アメリカ合衆国ミネソタ州55447, プリマス, ジュエル・レイン・ノース 615

審査官 坂東 博司

(56)参考文献 特開平08-115173 (JP, A)

特開2005-149274 (JP, A)

特開2005-332123 (JP, A)

特開2002-182859 (JP, A)

米国特許出願公開第2002/0103969 (US, A1)

欧州特許出願公開第00706113 (EP, A1)

米国特許出願公開第2005/0182992 (US, A1)

米国特許出願公開第2006/0161752 (US, A1)

米国特許第06052759 (US, A)

(58)調査した分野(Int.Cl., DB名)

G06F 3/06