(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3753259号
(P3753259)
(45)発行日　平成18年3月8日(2006.3.8)　(24)登録日　平成17年12月22日(2005.12.22)

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

請求項の数 9　(全 20 頁)

(21)出願番号　特願平7-330583
(22)出願日　平成7年12月19日(1995.12.19)
(65)公開番号　特開平8-221876
(43)公開日　平成8年8月30日(1996.8.30)
審査請求日　平成14年4月17日(2002.4.17)
(31)優先権主張番号　386,574
(32)優先日　平成7年2月10日(1995.2.10)
(33)優先権主張国　米国(US)

(73)特許権者　398038580
ヒューレット・パッカード・カンパニー
HEWLETT-PACKARD COMPANY
アメリカ合衆国カリフォルニア州パロアルト　ハノーバー・ストリート　3000
(74)代理人　100081721
弁理士　岡田　次生
(72)発明者　マーヴィン・ディー・ネルソン
アメリカ合衆国83704アイダホ州ボイジー、サンフラワー・レーン　9055
(72)発明者　テレサ・エー・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・ラモン・ドライブ　3100

最終頁に続く

(54)【発明の名称】記憶空間を提供する方法およびデータ記憶システム

(57)【特許請求の範囲】
【請求項１】
物理的な記憶空間を規定する複数の記憶ディスクを有する記憶ディスク・アレイにおいて、一つの記憶ディスクが故障したとき、再構築に使用するための記憶空間を提供するための方法であって、前記物理的な記憶空間は、ミラーRAID領域およびパリティRAID領域を含む複数のRAID領域として該物理的な記憶空間を表すRAIDレベルの仮想記憶空間にマップされており、前記方法は、
ディスクの故障が発生する前に、ユーザ・データが上記記憶ディスクのすべてにわたって記憶されるように、ユーザ・データを前記ミラーRAID領域およびパリティRAID領域に格納し、ユーザ・データが増加することに応じて前記ミラーRAID領域から前記パリティRAID領域にデータを動的に移行させ、その際、上記記憶ディスクの一つが故障するときにユーザ・データを再建し冗長性を復元するに十分なホットスペア記憶空間が得られるようにすることと、
を含み、ディスク故障が起るまで前記ホットスペア記憶空間は、ユーザ・データおよび冗長データを格納するために使用する、記憶空間を提供する方法。
【請求項２】
ディスク故障に続いて上記ディスク・アレイにおけるユーザ・データを再建し冗長性を復元するに十分なホットスペア記憶空間が得られないと判断されるとき、このことをユーザに知らせる警告を生成するステップを含む請求項１に記載の方法。
【請求項３】

物理的な記憶空間を規定する、選ばれた容量の複数の記憶ディスクからなるディスク・アレイを含む階層的なデータ記憶システムにおける再構築に使用するための記憶空間を提供する方法であって、

上記物理的な記憶空間を複数のRAID領域として表すRAIDレベルの仮想記憶空間に上記物理的な記憶空間をマップすることを含み、該RAID領域は、データをミラー冗長性に従って格納するミラーRAID領域およびデータをパリティ冗長に従って格納するパリティRAID領域を含んでおり、

ユーザ・データを上記ディスク・アレイの上記記憶ディスクのすべてにわたって分布する態様で、ユーザ・データが増加することに応じて、ミラーRAID領域からパリティRAID領域に動的に移行させて格納し、その際、記憶ディスク故障が生じたとき、ユーザ・データを再建し冗長性を復元するに十分な量のRAIDレベルのホットスペア仮想記憶空間が得られるようにすることと、

前記ホットスペア記憶空間を、上記記憶ディスク故障が生じるまでミラー冗長性に従ってデータを格納するのに使用することと、

を含む記憶空間を提供する方法。

【請求項４】

RAIDレベルの仮想記憶空間をアプリケーション・レベルの仮想記憶空間にマップすることと、

最も大きいサイズの記憶ディスクの容量を含むことなくユーザに割り当てることができるアプリケーション・レベルの仮想記憶空間の容量を計算することと、を含む請求項３に記載の方法。

【請求項５】

十分な量のRAIDレベルの仮想記憶空間が再構築のために利用可能にされるようにするため、上記の計算された容量を越えて付加的なアプリケーション・レベルの仮想記憶空間を割り当てるのを防ぐことを含む、請求項４に記載の方法。

【請求項６】

請求項４に記載の方法において、

上記ディスク・アレイが３つ以上の記憶ディスクを持ち、上記方法はさらに、

最初にユーザ・データをミラーRAID領域に格納することと、

記憶ディスク故障が生じるとき、(a) ミラーまたはパリティRAID領域として割り当てられていない不使用の記憶空間を使って、上記故障した記憶ディスクからユーザ・データを再建し、冗長性を復元すること、または(b)ミラーRAID領域からパリティRAID領域にユーザ・データを移行させて、上記記憶ディスク故障により失われた記憶空間を補う記憶空間を空けること、のいずれかを行うステップと、

を含む上記方法。

【請求項７】

少なくともミラーRAID領域およびパリティRAID領域を提供する複数の記憶ディスクを有するディスク・アレイと、

上記記憶ディスクとの間でユーザ・データの転送をコーディネートするために上記ディスク・アレイに結合されたディスク・アレイ・コントローラと、

メモリ割り当ておよび上記ディスク・アレイ上のデータ記憶を管理するため上記ディスク・アレイ・コントローラに結合されたメモリマネージャーとを含み、該メモリマネージャーは、上記記憶ディスクのすべてにわたってユーザ・データおよび冗長データを分配し、

上記記憶ディスクの一つが故障するとき、ユーザ・データを再建し冗長性を回復するために利用可能にされることができる十分な量のホットスペア記憶空間を維持し、記憶ディスク故障に続く再構築のために該ホットスペア記憶空間が得られるようにしながら、ユーザ・データが増加することに応じて、ユーザ・データを前記ミラーRAID領域から前記パリティRAID領域に動的に移行させるようにしたデータ記憶システム。

【請求項８】

上記記憶ディスクが物理的な記憶空間を規定し、
上記メモリマネージャーが、RAIDレベルの仮想記憶空間を上記記憶ディスクの上記物理的な記憶空間にマップするRAID管理システムを含み、上記RAIDレベルの仮想記憶空間が、物理的な記憶空間を前記ミラーRAID領域およびパリティRAID領域を含む複数のRAID領域として表し、
上記RAID管理システムはさらに、上記RAIDレベルの仮想記憶空間をアプリケーション・レベルの仮想記憶空間にマップし、
上記RAID管理システムは、最も大きいサイズの記憶ディスクの容量を含むことなくユーザに割り当てることができる上記アプリケーション・レベルの仮想記憶空間の容量を計算する、
請求項7に記載のデータ記憶システム。
【請求項９】
上記RAID管理システムは、十分な量の上記RAIDレベルのホットスペア仮想記憶空間が、ユーザ・データを回復し冗長性を復元するために利用可能にされるようにするため、計算された容量を越えて、付加的なアプリケーション・レベルの仮想記憶空間が割り当てられるのを防ぐ、請求項8に記載のデータ記憶システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
この発明は、ディスク・アレイに関し、より具体的には、記憶ディスクが故障する場合、ユーザ・データと冗長情報の再構築のために十分なスペースを提供するためのホット・スペア資源を持つ冗長で階層的なディスク・アレイ・データ記憶システムに関する。この発明は、再構築を容易にするために必要とされるまでユーザ・データを格納するためにホット・スペア資源を使うための方法に関する。
【０００２】
【従来の技術】
ディスク・アレイ・データ記憶システムは、一つの大容量記憶システムを形づくるよう配置され協調するようにされた複数の記憶ディスク・ドライブ装置を持っている。そのような記憶装置のための3つの主要な設計基準がある。これらは、費用、パフォーマンスおよび可用性である。メガバイトのコストが低く、入出力パフォーマンスが高く、データ可用性が高いメモリ装置を生産するすることが望まれる。「可用性」は、その記憶装置に格納したデータをアクセスする能力およびなんらかの故障の場合に継続操作を保証する能力である。典型的には、データ可用性は、データ、またはデータ間の関係が複数のロケーションに格納されいる状況で、冗長性の使用を通して提供される。そのディスク・アレイにおける記憶ディスクが部分的に、あるいは、完全に故障する場合、そのユーザ・データは残りのディスクに格納された冗長データを通して再構成することができる。
【０００３】
冗長データを格納するには2つの一般的な方法がある。第１の方法、すなわち「 ミラー」法によると、データは、複製され、その記憶装置の2つの別々の領域に 格納される。例えば、ディスク・アレイにおいて、同じデータがディスク・アレイの2つの別々のディスクに提供される。ミラー法は、二重記憶技術によって高 いパフォーマンスおよび高いデータ可用性の利点を持っている。しかし、ミラー法は、データを格納する費用を実質的に２倍にするので比較的高価である。
【０００４】
第２の方法、すなわち「パリティ」法では、記憶領域の一部分が、冗長データを格納するために使われるが、冗長な記憶領域のサイズは、オリジナルのデータを格納するのに使われた残りの記憶スペースより小さい。例えば、5つのディス クを持つディスク・アレイにおいて、4つのディスクが、データを格納するのに 使われ、5番目のディスクが冗長データを格納するのに使われる。パリティ法は 、ミラー法より低コストだという利点があるが、ミラー法に比べるとパフォーマンスおよび可用性特性において劣る。

【０００５】
【発明が解決しようとする課題】
データ冗長性は別として、ディスク・アレイ・データ記憶システムのあるものは、故障した記憶ディスクの代用になる付加的な物理的記憶ディスクを保留することによってデータ可用性を改良する。この余分の記憶装置ディスクは、「スペア」と呼ばれる。予備のディスクは、ディスク故障の後、「再構築建」として知られるプロセスで、ディスク・アレイにおいてユーザ・データを再構成し、冗長性を復元するために使われる。ある場合には、この余分な記憶ディスクはディスク・アレイに実際に取り付けられ完全に使用可能であるが、記憶ディスクが故障するまではアイドル状態のままである。これらの運転中の（ライヴ：live）記憶ディスクは、「ホット・スペア」と呼ばれる。
【０００６】
ホット・スペア・ディスクは、ディスク故障の際に有用であるが、すべてのディスクが正常に機能している正常状態ではディスク・アレイになんらのサービスも提供しない。ホット・スペア目的のために物理的な記憶ディスク全体を保留して印を付けることは、付加資源をデータ記憶作用のために十分利用する利益なしでシステムのコストを増やす。
【０００７】
したがって、この発明は、ディスク・アレイのすべての記憶ディスクを十分利用しながらホット・スペア資源を提供することを課題とする。
【０００８】
【課題を解決するための手段】
この発明は、アレイに付けられたすべての記憶ディスクを十分に利用しながら、ホット・スペア資源の余裕を持つディスク・アレイ・データ記憶システムを提供する。ホット・スペアのために物理的ディスク全体を予約して印を付ける代わりに、この発明のシステムおよび方法は、アレイの複数の記憶ディスクにわたって分布した有効記憶スペースから成るホット・スペア・ディスクの論理的容量を保留する。このように、ディスク故障の場合、ユーザ・データを再構成し冗長性を復元するのに、適切な量のスペースを使うことができることを確保しながら、すべての記憶装置ディスクが、ユーザ・データを格納するために使われる。
【０００９】
この発明の一つの面によると、ディスク・アレイ・データ記憶システムは、複数の記憶ディスク、および記憶ディスクの間でユーザ・データの移転をコーディネートするためのディスク・アレイ・コントローラを持つ。RAID管理システムは、RAIDレベルの仮想記憶スペースを記憶ディスクの物理的な記憶スペースにマップするために設けられている。RAIDレベルの仮想記憶スペースは、データをミラー冗長性に従って格納するミラーRAID領域、およびパリティ冗長性に従ってデータを格納するパリティRAID領域を含む複数のRAID領域として物理的な記憶スペースに提示される。RAID管理システムは、またRAIDレベルの仮想記憶スペースを、ユーザに提示されるアプリケーション・レベルの仮想記憶スペースにマップする。
【００１０】
ホット・スペア・オプションがイネーブルされるとき、RAID管理システムは、記憶ディスク故障に続いて冗長性を再構築するために、ホット・スペア・スペースが使用可能になることを保証しながら、ミラー記憶装置にユーザ・データを格納するためにホット・スペアのスペースを使う。スペースは、ミラー記憶装置からパリティ記憶装置へデータを移行することによって使用可能になる。この移行は、ミラー記憶装置の量を減少させながら同時にパリティ記憶装置の量を増やす。この移行プロセスと同時に、データ記憶のためのストライプ・サイズが、故障した記憶ディスクから失われたスペースのために減らされる。
【００１１】
このシステムによると、アレイにおけるすべての記憶ディスクがユーザ・データを格納するため、およびディスク故障の場合にホット・スペアの目的のためで容易に使用可能にできる記憶スペースを維持するために使われる。ホット・スペア記憶スペースのこの使用は

、並列性を増やすことによって、そしてミラー記憶装置のためにより多くの容量を提供することによってかなりパフォーマンスを改善する。

【００１２】

【発明の実施の形態】

図1は、この発明に従ってつくられたデータ記憶システム10を示す。データ記憶システム10は、複数の記憶ディスク12を持つ階層的なディスク・アレイ11、記憶ディスク12との間でのデータの転送をコーディネートするためディスク・アレイ11に結合されたディスク・アレイ・コントローラ14、およびRAID管理システムを含む、ディスク・アレイ・データ記憶システムであるのが好ましい。この発明は、冗長な階層的なディスク・アレイ・システムとしての好ましい実施形態のについて記述される。しかし、この発明の概念は、非階層的なディスク・アレイのようなディスク・アレイ・データ記憶システムの他の型において使用することができることに注意されたい。

【００１３】

説明上、「ディスク」は、それ自身の記憶装置故障を発見する能力を持つ、不揮発性でランダムにアクセス可能な、再書き込み可能な大容量記憶装置である。それは、回転する磁気ディスクおよび光ディスク、ならびにソリッドステートのディスクまたは不揮発性電子記憶素子（PROM、EPROMとEEPROMのような）を含む。「ディスク・アレイ」という用語は、ディスクの集合をいい、それらを一つまたは複数のホスト・コンピュータに接続するために必要なハードウェア、および物理的なディスクの操作を制御するために使われ、ホスト操作環境へそれらを一つまたは複数の仮想ディスクとして提供する管理ソフトウェアを含んでいる。仮想ディスクは、管理ソフトウェアによってディスク・アレイの中で実現される抽象的な構成要素である。

【００１４】

用語「RAID」(Redundant Array of Independent Disks)は、記憶容量の残りの部分に格納されたユーザ・データに関する冗長情報を格納するために、物理的な記憶容量の一部分が使われるディスク・アレイを意味する。そのアレイの構成ディスクの一つが故障するかそれへのアクセス経路が故障した場合、冗長情報がユーザ・データの再生を可能にする。RAIDシステムのより詳細な説明は、The RAID Book: A Source Book for RAID Technology, the RAID Advisory Board, Lino Lakes, Minnesota、1993 年6月9日、に記載されている。

【００１５】

ディスク・アレイ・コントローラ14は、小型コンピュータ・システム・インターフェース（SCSI）のような１つまたは複数のインターフェース・バス13を通してディスク・アレイ11に結合されている。RAID管理システム16は、インターフェース・プロトコル15を通して動作的にディスク・アレイ・コントローラ14に結合されている。また、データ・メモリ・システム10は、I/Oインターフェース・バ ス17を通してホスト・コンピュータ（図示しない）に結合される。RAID管理システム16は、ディスク記憶および信頼性のレベルを制御するため、およびいろいろな信頼性記憶レベルの間でデータを転送するためにデータ・マネージャー手段を提供するよう、別々のコンポーネントとして具体化されることも、ディスク・アレイ・コントローラ14内で、あるいはホスト・コンピュータ内で構成されることもできる。これらの信頼性記憶レベルは、後述するようにミラーまたはパリティ冗長レベルであるのが好ましいが、全く冗長性なしの信頼性記憶レベルを含むことができる。

【００１６】

ディスク・アレイ・コントローラ14は、ディスク・アレイ・コントローラA、14aとディスク・アレイ・コントローラB、14bから成る二重のコントローラとして実施するのが好ましい。二重のコントローラ14aと14bは、連続的なバックアップと冗長性を提供することによって、一つのコントローラが動作不能になるときの信頼性を高める。しかし、この発明は、一つのコントローラその他のアーキテクチャーでも実施することができる。

【００１７】

階層的なディスク・アレイは、その物理的な記憶スペースおよび１つまたは複数の仮想記

憶スペースを含めて、異なる記憶スペースとして特徴づけることができる。記憶装置のこれらのいろいろな見方は、マッピング技法を通して関係づけられる。例えば、そのディスク・アレイの物理的な記憶スペースは、いろいろなデータの信頼度レベルに従って記憶領域を画定する仮想記憶スペースにマップすることができる。仮想記憶スペース内のいくつかの領域が、ミラーまたはRAIDレベル1のような第１の信頼性記憶レベルのために割り当てられ、他の領域がパリ ティまたはRAIDレベル5のような第２の信頼性記憶レベルのために割り当てられ ることができる。いろいろなマッピング技術およびRAIDレベルに関する仮想スペースについて、以下に詳細を説明する。

【００１８】

データ記憶システム10は、異なる記憶スペースを互いにマップするために使われるる仮想マッピング情報を持続的に記憶するためのメモリ・マップ・ストア21を含む。メモリ・マップ・ストアは、ディスク・アレイに対して外部にあり、ディスク・アレイ・コントローラ14に常駐するのが好ましい。異なるビューの間でいろいろなマッピング構成が変わるにつれて、メモリ・マッピング情報がコントローラまたはRAID管理システムによって連続的に、あるいは定期的に更新されることができる。

【００１９】

メモリーマップ・ストア21は、それぞれのコントローラ14aと14bに位置する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 が好ましい。不揮発性RAM（NVRAM）の一例は、バッテリで支援されたRAMである。バッテリ支援RAMは、データ記憶システム10への電力供給が損なわれた場合、ある時間の間メモリの中のデータを維持するために独立のバッテリ源からのエネルギーを使う。一つの好ましい構造は、自己リフレッシュ型のバッテリ支援DRAM（Dynamic RAM）である。

【００２０】

二重のNVRAM 21aおよび21bが、メモリ・マッピング情報の冗長な記憶を行う。仮想マッピング情報が、ミラー冗長技術に従って複製され、両方のNVRAM 21aお よび21bに格納される。このように、NVRAM 21aが、オリジナルのマッピング情報を格納することに専用され、NVRAM 21bが冗長なマッピング情報を格納すること に専用されることができる。代替構造において、ミラーされたメモリーマップ・ストアは、そのデータを二重に記憶するに十分なスペースを持つ単一の不揮発性RAMを使って構成することができる。

【００２１】

図1に示されるように、ディスク・アレイ11は、複数の記憶ディスク・ドライ ブ装置12を持っている。これらの記憶ディスクのサイズの例は、1ないし3ギガバイトである。記憶ディスクは、SCSIバス13とのインターフェイスを提供する機械的ベイに、独立に接続されたり切り離したりされることができる。一つの実施例において、データ記憶システムは、12の機械的ベイを持つよう設計される。4つ のSCSIバスが、これらのベイとディスク・アレイ・コントローラ14とをインターフェイスするのに使われる（すなわち、3つの機械的ベイにつき一つのバス）。 活動ベイがいっぱいにロードされるならば、データ記憶システムは、12ないし36ギガバイトの結合された容量を持つ例になる。ディスク・アレイ・コントローラ14は、どのベイに接続されたかとは関係なく記憶ディスク12を認識する。データ記憶システム10は、ディスク・アレイが運転中である間に、ディスク・アレイの利用可能な機械的ベイに付加的ディスクを「ホットプラグ」（稼働中にプラグ接続する）することができるよう設計されている。

【００２２】

ディスク・アレイ11の中の記憶装置ディスク12は、説明の目的のために、複数のディスク20のミラー・グループ18および複数のディスク24のパリティ・グループ22に配列されているものと観念することができる。ミラー・グループ18は、データを第１の、すなわちミラー冗長レベルに従って格納する、ディスク・アレイの第１の記憶ロケーションすなわちRAID領域を表わす。また、このミラー冗長レベルは、RAIDレベル1と考えられる。RAIDレベル1、あるいはディスク・ミラーリングは、データのあらゆるビットがデータ記憶システム内で複製され記憶される点で１対１の保護を与え、最も高いデータ信頼性を提供する。

ミラー冗長性は、図１において３対のディスク20によって表わされている。オリジナルのデータが、第１組のディスク26に格納されることができ、重複して冗長データが対になった第２組のディスク28に格納される。

【００２３】

図2は、RAIDレベル1に従ったデータの記憶を図示する。垂直な列は、個々のディスク（ディスク0、1、2および3が図示される）を表現する。4つのディスクの ディスク・アレイに含まれる物理的な記憶スペースは、水平な行によって表現されるように、複数のストライプに構成することができる。「ストライプ」は、記憶ディスクを横切って広がり、一つのセグメントがそのアレイの中のそれぞれのディスクと関連する、多数の等しい大きさの記憶スペースのセグメントを含む。すなわち、セグメントは、一つのディスク上にあるストライプの部分である。各ストライプは、記憶ディスクを横切って分配されるあらかじめ決められた量のデータを持つ。ストライプのセグメントのあるものは、オリジナルのデータのために使われ、他のセグメントが冗長データのために使われる。

【００２４】

ミラー冗長性（RAIDレベル1）のこの例において、ディスク0上でストライプ0 のセグメント0に格納されたデータは、複製され、ディスク1上でストライプ0の セグメント0'に格納される。同様に、ディスク2上でストライプ2のセグメント5 に格納されたデータは、ディスク3上でストライプ2のセグメント5'にミラー（複製）される。このように、データの各部分は、複製されてディスク上に格納される。図2の冗長性レイアウトは、説明の目的で用意されている。冗長データは、 図示しているのと同じストライプにきちんと置かれる必要はない。例えば、ディスク0上でストライプ1の中のセグメント2に格納されたデータは、複製されて、 ディスク3上でストライプSのセグメントT'に置かれてもよい。

【００２５】

再び図1を参照すると、ディスク24のパリティ・グループ22は、データがRAID レベル5のような第２の冗長レベルに従って格納される第２の記憶ロケーション すなわちRAID領域を表わす。6つのディスクを用いた説明的なイラストにおいて、オリジナルのデータは、5つのディスク30に格納され、冗長な「パリティ」デ ータが6番目のディスク32に格納される。

【００２６】

図3は、パリティRAID領域のレイアウトをより詳しく示す。図２のミラーRAID 領域レイアウトと同様に、ディスク0、1、2、3の物理的な記憶スペースは、複数の等しい大きさのストライプに構成されることができる。図の例において、データがRAIDレベル5に従って格納され、セグメントに格納された冗長データは、文 字Pによって参照される。冗長なPセグメントは、他のセグメントのパリティをそのストライプに格納する。たとえば、ストライプ0において、ディスク3上の冗長なPセグメントは、ディスク0、1および2のパリティを格納する。各ストライプのためのパリティが、次の式の中の記号によって表わされる排他的ＯＲ機能のようななんらかの機能によって計算される。第１の4つのストライプのためのパリテ ィ（下付き添字数字が対応するストライプを表わす）は、次のようである：

【００２７】

【数１】

P_0 ＝セグメント０ ⊕ セグメント１ ⊕ セグメント２
＝ディスク０ ⊕ ディスク１ ⊕ ディスク２

P_1 ＝セグメント３ ⊕ セグメント４ ⊕ セグメント５
＝ディスク０ ⊕ ディスク１ ⊕ ディスク３

P_2 ＝セグメント６ ⊕ セグメント７ ⊕ セグメント８
＝ディスク０ ⊕ ディスク２ ⊕ ディスク３

P_3 ＝セグメント９ ⊕ セグメント10 ⊕ セグメント11
＝ディスク１ ⊕ ディスク２ ⊕ ディスク３

パリティ冗長性は、ディスクの一つで利用できなくなるデータの再生を可能にする。例えば、セグメント5のデータが利用できなくなったならば、その内容は、セグメント3および4ならびにセグメントPの中のパリティ・データから確かめ ることができる。パリティ記憶は、ミラー記憶より経済的であるが、相対的に信頼性が低く、パフォーマンスが低い。

【００２８】

図1のディスク構成は、概念上の目的で用意されている。実際上、ディスク・ アレイ11は、データをミラーおよびパリティ冗長性に従って格納することができる単に複数のディスク12を持っていてよい。すべてのディスク12によって提供される利用可能な記憶スペースの中で、その記憶スペースの一部がミラー冗長性のために割り当てられ、他の部分がパリティ冗長のために割り当てられる。ディスク12は、個々の区域が複数のセグメントを持つ、複数の等しい大きさに作られた記憶区域（図４において数字35で参照される）を含むよう構成されるのが好ましい。その区域は、記憶スペースの一つの仮想ビューにおいて、RAID領域を形成するようグループ化される。さらに、記憶スペースのもう一つの（ホストで規定される）ビューがユーザまたはホストに提示されるので、RAID領域およびデータ冗長性記憶技術はユーザまたはホストに透明である。これらの機構は、図4を参照 してより詳しく説明する。

【００２９】

データ記憶システム10は、ミラーおよびパリティ記憶法の間でデータの「移行」を管理する。両方の型の冗長性の管理は、RAID管理システム16（図1）で協調 される。RAID管理システム16は、ディスク・アレイの2つの異なる型のRAID領域 を、ミラーRAID領域がパリティRAID領域のためのキャッシュに似た動作をするメモリ・ハイアラーキとして管理する。RAID管理システム16は、規定されたパフォーマンス・プロトコルに従って、ミラーおよびパリティRAID領域の間でデータをシフトし、組織するなど管理する。ミラーおよびパリティRAID領域の間でデータを動かすプロセスは、「移行」と呼ばれる。

【００３０】

データ記憶システム10は、ミラー冗長性が最も高いパフォーマンスと信頼性の余裕を持っているから、よりパフォーマンスに厳格なデータをミラーRAID領域に置こうとする。RAID管理システムによって実行されるパフォーマンス・プロトコルは、2つの好ましい移行方針の一つを含む。「アクセス頻度」として知られる 第１の移行方針によると、階層的なディスク・アレイ上の最も頻繁にアクセスされるデータがミラーRAID領域18に維持される。アクセスされる頻度の小さいデータは、パリティRAID領域22に維持される。「アクセスの最近度」として知られている第２の移行方針によると、最も最近アクセスされたデータがミラーRAID領域18に維持され、アクセス最近度の低いデータがパリティRAID領域22に格納される。他のパフォーマンス・プロトコルを使うこともできる。理想的には、そのようなプロトコルは、特定のコンピュータ・アプリケーションおよびユーザのニーズに基づいて規定される。

【００３１】

さらに、RAID管理システム16は、2つのパラメーターの関数に従って、データ 記憶システムの記憶資源を自動的に「調整する」。この２つのパラメータは、物理的な記憶容量のサイズ、およびデータ記憶システムに格納つつあるユーザ・データの現在量のサイズである。最初は、最も高いパフォーマンスおよび信頼性を持つので、すべてのデータがミラーRAID領域に格納される。より多くのデータがデータ記憶システムに加えられるにつれて、データは、パフォーマンスおよび信頼性を最適化するために、ミラーRAID領域およびパリティRAID領域の間で移行される。データ記憶システムが容量いっぱいに近づくと、冗長性を通しての信頼性を提供しながらユーザのすべての要求に応ずるために、より多くのデータがパリティRAID領域に移行される。それゆえに、この発明のデータ記憶システムは、最大の融通性と順応性を持つ。この発明の記憶システムは、ユーザに特定の記憶体制を選ぶことを要求せず、ユーザの要求に適応することができる。

【００３２】

図4は、データ記憶システム10の利用可能な記憶スペースのメモリ・マッピン グを、マップされた仮想記憶スペースの複数の段として図示する。図の中の垂直に伸びたそれぞれの長方形は、物理的な記憶スペースのビューを表現する。この図の中で、物理的記憶スペース34が２つの仮想記憶装置のビュー40および50で参照される。物理的記憶スペース34は、数字0、1、2、および3で参照される4つの ディスク（図1の中のディスク12のような）によって表現される。そのディスク に関連した4つの長方形は、ディスク1、2および3がほぼ等しい記憶容量を持ち、ディスク0がわずかに小さい記憶容量を持つ物理的な記憶スペースのビュー50を 表わす。そのようなディスクの例としての記憶容量は、1-3ギガバイトである。 記憶スペース34は、領域A0、A1、A2、その他に区分される。個々の領域は、領域A0にストライプ0-Qのような多数のストライプを含む。個々の領域も、多数の区 域35を含む。区域35は、区域が全ディスク・アレイにわたって同じ大きさであるように、あらゆる記憶ディスク上で選ばれた数の均一な大きさのセグメントから成るのが好ましい。区域35の例としてのサイズは、１メガバイトである。

【００３３】

ディスクの記憶スペースは、物理的な記憶スペース34の第１の、中間の、RAIDレベルの仮想ビュー40にマップされる。この第１の仮想ビューは、観念的には、連続してアドレス可能な記憶スペースを表現する第２のアプリケーション・ビューにマップされることができる一組のRAID領域である。その物理的構成およびその記憶スペースのRAIDビューは、アプリケーション・ビューから隠されている。 RAID領域記憶スペース40は、ミラーおよびパリティ記憶スペースを識別する記憶装置のビューである。たとえば、RAID領域42が、M割当てブロック43のミラーRAID領域を表現し、RAID領域44がN割当てブロック45のパリティRAID領域を表現してもよい。割当てブロック43および45は、同じ大きさであるのが好ましい。例としての大きさは、64キロバイトである。これらのRAID領域は、物理的な記憶スペース34上の対応する物理的な領域A0、A1、A2などに関連する。例として、RAID仮想ビューにおける16の64K割当てブロック43または45は、一つの1M区域35の上へ マップすることができる。

【００３４】

ミラーおよびパリティRAID領域は、ディスク・アレイの全記憶スペース34を消費するかもしれず、消費しないかもしれない。したがって、ある種のアプリケーションの間、特定のRAID領域に対応しない、不使用の指定されていない記憶スペースがあることがある。しかし、そのような記憶スペースは、ミラーまたはパリティRAID領域に変換されることができる。また、RAID領域は、ディスク・アレイ上で連続する領域にマップされ、各RAID領域に関連したそれぞれの区域が各記憶ディスクの同じ物理アドレスに位置するものとして示されている点に注意されたい。代わりに、RAID領域は、ディスク・アレイ上の連続しない領域にマップされてもよい。

【００３５】

RAID領域で利用可能な記憶スペースは、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規定され、それらに提示される記憶装置のビューである、第２の、フロント

エンドの、アプリケーション・レベルの仮想ビュー50にマップされる。ユーザまたはホスト・アプリケーション・プログラムから見るとき、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上の利用可能な記憶スペースを示す一つの大きい記憶容量を表わすことができる。仮想記憶スペース50は、個別に0、1、2、...、J-1、J、J+1、...、L－1、L、L+1、...、などとして参照される等しい大きさの記憶仮想ブロック52および53の線形セットのビューを提示する。仮想ブロック52および53は、RAID領域40における割当てブロックと同じサイズであることが好ましい。サイズの例は、64キロバイトである。仮想ブロック記憶スペース50は、RAID領域40で表されるビューの割当てブロックへの参照すなわちポインタ（矢印54によって表現されるように）のテーブルによって 表わされる。したがって、アプリケーション仮想ビュー50における仮想ブロック52および53は、仮想ブロック・テーブルに維持されるポインタを通してRAID仮想ビュー40における割当てブロック43および45と関連する。仮想ブロック・テーブルから参照することができる少くとも2つ型のRAID領域、ミラーおよびパリティ がある。

【００３６】

RAID管理システム16は、物理的な記憶スペースにわたって動的にRAID領域の構成を変えることができる。それぞれの型についてRAID領域の数が、システムで格納されるユーザ・データの量および物理的なディスク記憶スペースのサイズに依存して増やされまたは減少されうる。この結果、RAIDレベルの仮想ビュー40におけるRAID領域のディスクへのマッピング、およびアプリケーション・レベルの仮想ビュー50のRAIDビュー40へのマッピングは、概して変化状態にある。NVRAM 21aおよび21b（図１）におけるメモリ・マップ・ストアは、RAID領域をディスクへマップするためにRAID管理システム16によって使われる現在のマッピング情報を、2つの仮想ビューの間でマップするために使われる情報と同様に維持する。RAID管理システムは、動的にRAIDレベルのマッピングを変えるにつれて、その変更を反映するために、メモリ・マップ・ストアのマッピング情報を更新する。

【００３７】

メモリ・システム10の移行操作を図1および4を参照して説明する。

【００３８】

説明を続ける目的で、アプリケーション・レベルの仮想記憶スペース50の仮想ブロック53は、物理的記憶スペース34の領域A1に格納されたパリティRAID領域44における関連した割当てブロック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ぶことができ、関連した割当てブロック45は、パリティ割当てブロックと呼ぶことができる。同様に、仮想ブロック52は、物理的記憶スペース34の領域A0に格納されたミラーRAID領域42における関連した割当てブロック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、関連した割当てブロック43が「ミラー割当てブロック」と呼ばれる。

【００３９】

一般に、一つのRAID領域から別の領域に移行するために、第１のRAIDレベル型（ミラーまたはレベル1のような）の割当てブロックに関連した仮想ブロックが 選ばれる。次いで、第２のRAIDレベル型（パリティまたはレベル5のような）を 表す不使用の割当てブロックが見いだされる。不使用の割当てブロックを見いだすことができないならば、一つが作成される。データは、選ばれた仮想ブロックと関連する割当てブロックから不使用の割当てブロックに移され、データは冗長レベルを変更される。例えば、ミラー冗長性に従って格納されていたデータは、パリティ冗長性に従って格納され、また、その逆が行われる。最終的なステップとして、RAIDレベルの仮想記憶スペース40へのアプリケーション・レベルの仮想記憶スペース50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。以前に第１のRAIDレベル型の割当てブロックと関連していた、選ばれた仮想ブロックが、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レベル型の割当てブロックを参照する。この処理の間に生じるマッピング変化は、メモリ・マップ・ストア21において更新される。

【００４０】

この発明の好ましい方法と順序に従って、ミラーおよびパリティ記憶領域の間を移行するデータのより詳細な説明を続ける。データをパリティ記憶からミラー記憶に移行するために、次のシーケンスが使われる。

【００４１】

1. RAID管理システムは、ミラーRAID領域42における、不使用のミラー割当てブロック43を見つける。
2. 何も見つけられることができなければ、RAID管理システムは、ミラー割当てブロック（下で説明する）を作り出す。
3. RAID管理システムは、移行すべき仮想ブロックへの新しい記憶要求を中断する。
4. RAID管理システムは、仮想ブロックへのすべての有効なデータ記憶要求が完了されるまで待つ。
5. 仮想ブロック53に関連したパリティ割当てブロック45からのデータが、一時記憶バッファに読み込まれる。
6. データは、次いで、ステップ2において選ばれたミラー割当てブロックに書 き込まれる。
7. 仮想ブロック・テーブルは、ミラー割当て43ブロックにおけるそのデータの新しいロケーションを参照するよう修正される。
8. 中断されていた記憶要求が、再開される。

【００４２】

上記の手順によると、アプリケーション・レベルの仮想ビュー50における仮想ブロック53は、パリティ記憶からミラー記憶に移行した。中間の仮想ビュー40に関し、データは、パリティRAID領域44におけるパリティ割当てブロック45からミラーRAID領域42におけるミラー割当てブロック43に移行されている。物理的な記憶スペースにおいて、データは領域A1から領域A0に動いている。

【００４３】

不使用のミラー割当てブロックを見いだすことができないならば（上のステップ1）、RAID管理システムは、次の3つの技法の好ましいシーケンスをためす。第１に、RAID管理システムは、不使用のRAIDレベルの記憶のシステムしきい値を破ることなくミラーRAID領域に変換されることができる、不使用の（したがって指定されていない）RAID領域を見つけるよう試みる。これは、移行を常に進めることができることを保証するために必要とされる。これが失敗し、そして保留量の不使用のRAIDレベル記憶より多くをシステムが持っているならば、システムは、不使用のパリティ割当てブロックを不使用のRAID領域に集めるためにパリティ記憶装置内でデータを移行させる。この移行によって、上述のようにミラーRAID領域に変換されることができる不使用のRAID領域が生じれば、システムは、それをミラーRAID領域に変換する。さもなければ、システムは、不使用のミラー割当てブロックを見いだすに十分か、または不使用のRAID領域をミラーRAID領域に変換するに十分に、不使用のRAIDレベルの記憶装置をシステムが増やすまで、データをミラー記憶装置からパリティ記憶装置へ移行し、ミラー記憶装置を詰め込み、不使用のRAIDレベルの記憶をパリティに変換する。ミラー割当てブロックがパリティ割当てブロックより大きな物理的な記憶スペースを占めるから、この最後の技法によって、不使用のRAIDレベル記憶装置の量が正味増加する。

【００４４】

不使用のミラー割当てブロックを見つけるためおよび確立するために使われる創設/変換プロトコルは、RAID管理システムがユーザ・データの量および物理的 な記憶スペースのサイズに従ってパリティおよびミラー領域の間で記憶割当てを選択的に調節することを可能にするので、有利である。データ使用および記憶容量が変化するにつれて、RAID管理システムは、ミラー記憶装置に保持されるデータの量を最大にするために上記の3つの技法の１つまたは複数を使う。

【００４５】

RAID管理システムは、スペース作成シーケンスがアイドル時間に不使用RAID領域をつくる

ことによって不使用のミラー割当てブロックを創設するのを、記憶装置要求が待たなければならない状況を避けようとする。しかし、いくらかの状況において、記憶装置要求はスペース作成シーケンスの間、中断されることがある。RAID管理システムは、仮想スペースがRAIDビューより小さくなるような形態で、仮想ブロック記憶スペースを構成する。これは、少くとも1つのRAID領域と等 しい自由スペースが移行その他の目的のためにとっておかれることを保証する。このように、この技法は、常に不使用のミラー割当てブロックを生む。

【００４６】

データをミラー記憶装置からパリティ記憶装置へ移行させるために、以下の順序が使われる：

1. RAID管理システムは、アクセス最近度またはかアクセス頻度のような移行方 針に従ってミラー記憶装置からパリティ記憶装置へ移行させるために仮想ブロックを52から選ぶ。
2. RAID管理システムは、パリティRAID領域44に不使用のパリティ割当てブロック45を見つける。
3. そのようなブロックが見つけられないならば、移行用に保留されたスペース が上述の創設技法に従ってパリティRAID領域に変換される。
4. 移行されるべき仮想ブロックへの新しい記憶装置要求が中断される。
5. RAID管理システムは、仮想ブロックへのすべての有効な記憶装置要求が完了するまで待つ。
6. データは、仮想ブロック52に関連したミラー割当てブロック43から一時記憶バッファに読込まれる。
7. そのデータは、選ばれたパリティ割当てブロック45に書込まれる。
8. 仮想ブロック・テーブルが、パリティ割当てブロック45におけるデータの新しいロケーションを参照するよう修正される。
9. 仮想ブロックへのデータ要求が再開される。

【００４７】

上述の2つの番号を付けたシーケンスは、この発明のメモリ・システムが異な る2つの冗長性レベルの間でデータを移行する方法の例を提供する。

【００４８】

データ冗長は別として、この発明のディスク・アレイ・データ記憶システムは、データ可用性を高めるために「ホット・スペア」能力を提供する。発明の背景において説明したように、先行技術のディスク・アレイは、記憶ディスクが故障する場合に使用する付加的な物理的記憶ディスク、すなわち「スペア」をとっておく。この余分な記憶ディスクは、ディスク・アレイに取り付けられ、完全に動作可能であるが、記憶ディスク故障が起こるまで、働いていないままである。このライブ記憶ディスクは、「ホット・スペア」と呼ばれる。ホット・スペア・ディスクは、ディスク故障の後、「再構築」として知られるプロセスでデータを再構成し冗長性を復元するために使われる。あいにく専用のホット・スペア・ディスクは、すべてのディスクが正しく作用している通常操作の間には有用なサービスを提供せず、ディスク故障の際にだけ有用である。

【００４９】

この発明のディスク・アレイによると、ディスク故障の際にユーザ・データを再構築するに十分な量の記憶スペースが、アレイの多数のそして好ましくはすべての記憶ディスクにわたって分布した形態で維持され、RAID１の記憶装置を増大させるのに使われる。「ホット・スペア」に指定された記憶装置ディスクは一つもない。アレイにおけるすべてのディスクは、再構築のために記憶スペースをつかうことができるような態様でアクセス可能なユーザーデータを格納する。ホット・スペア・スペースの使用は、並列性を増やすことによって、そして、ミラー記憶のためにより多くの容量を提供することによってかなりパフォーマンスを改善する本質において、過去、単にホット・スペア資源のためにだけ保留された付加的な磁気ディスクが、ユーザ・データを格納する通常動作のためにアレイに編入される。

【００５０】
図5は、ディスク故障が起こる前に、記憶ディスク0-5を持つ例としての６ディスクの同質のディスク・アレイを示す。その記憶スペースすなわち記憶空間のすべてが、ミラーまたはパリティ技法に従ってユーザ・データおよび冗長データを格納するために使われる。どのディスクも領域も、ホット・スペア資源のためにだけ保留されることがない。ここで、そのユーザ・データの55%がRAID5のパリティ領域に格納され、45%がRAID1のミラー領域に格納される。たとえすべての記憶空間が普通のデータ記憶のために使われているとしても、ディスク故障の場合、冗長性を復元するためおよびユーザ・データを再建するために、ある量のスペースを使用可能にすることができる。この図において、使用可能にすることができるホット・スペア・スペースが、アレイの記憶ディスクにわたって分布するRAID1の記憶スペースとして現在使われている。

【００５１】
図6は、図5と同じディスクアレイを示すが、記憶ディスク5の故障の後を示している。ディスク5の上に格納されたユーザ・データおよび／または冗長データは、失われる。故障が生じると、ディスク・アレイは、そのシステムにおける冗長性を再構築する再構築プロセスを開始する。スペースは、ミラー記憶からパリティ記憶にデータを移行することにより使用可能にされる。この移行は、ミラー記憶の量を減少させてながら、同時にパリティ記憶の量を増やす。この移行プロセスと並行して、故障した記憶ディスクのために失なわれた一つのセグメントだけシステムのすべてのストライプのサイズが減少する。いまや、ストライプがディスク0-4にわたって広がるだけでディスク5にわたっては延びていないことに注意されたい。このデータ移行およびストライプ・サイズの減少の結果、90%のユーザ・データがRAID5の領域に格納され、10%がRAID1の領域に格納される。冗長性は、この再構築プロセスに続いてディスク・アレイに復元されている。

【００５２】
図7は、ディスク5の故障に続いてデータの移動によってどのように付加的スペースが自由にされるかを図示する。この図は、一般に数字90と92で参照される3つのストライプの２セットを含む記憶ディスク0-5の部分を示す。上の3つのストライプにおけるデータは、ユーザ・データがセグメント1-9に格納され、冗長データがセグメント1'ないし9'に格納されるミラー記憶技法に従って格納される。ディスク5の冗長データ3'、6'および9'がこのディスクの故障のために消失していることに注意されたい。

【００５３】
図7の例において、ストライプ90におけるユーザ・データが低いセットのストライプ92に移行されると仮定する。ここで冗長性記憶技法の相違によって、ユーザ・データ1ないし9をより少ないセグメントに格納することができる。ディスク5の消失にもかかわらず、そのデータの移動によって付加的な3つのセグメント（「不使用」のラベル）が使用可能になった。この３つのセグメントは、ユーザ・データを格納するために使うことができる。

【００５４】
ディスク故障に続く再構築を保証しながらRAID1の記憶装置としてホット・スペア・スペースを使う能力は、メモリ・マネージャーによって管理される。好ましい実施例においいて、メモリ・マネジャーは、ユーザ・データ、冗長データおよびのディスク・アレイ上の不使用スペースの情報を追跡し続けるRAID管理システム（図1）として具体化される。

【００５５】
RAID管理システムは、ユーザに割り当てることができるアプリケーション・レベルの仮想記憶スペースの量（図4におけるスペース50）に人為的な天井を設けることによって、使われた記憶スペースがホット・スペア目的のために、有効にされることを保証する（アプリケーション・レベルの仮想記憶スペースの量が、ホット・スペア・オプションがオフにされれば使用可能であるものから、減らされる点で「人為的」）。この天井を提供するために、RAID管理システムは、ディスク・アレイがその必要な機能（データの移動のような）を実行することができることを保証しながら、ユーザに割り当てることができるアプリケーション・レベルの仮想記憶スペースの容量を定型的に計算する。ユーザがホット・ス

ペア・オプションを選ぶとき、RAID管理システムは、まるでそれがそのアレイにおける最も大きいサイズの記憶ディスクの容量を含まなかったように、アプリケーション・レベルの容量を計算する。最も大きい記憶ディスクの容量は、アレイにおける最も大きいものを含む任意の記憶ディスクが故障する場合に、再構築プロセスが実行されることを保証するのに使われる。すべての記憶ディスクが同じサイズである均質なディスク・アレイにおいては、任意の記憶ディスクの容量を使うことができる。

【００５６】

ひとたび人為的な天井が計算されるならば、RAID管理システムは、計算された容量を越えて付加的なアプリケーション・レベルの仮想記憶スペースが割り当てられるのを防ぐことによってこの限界を強制する。容量分析および強制の結合によって、アプリケーション-レベル仮想記憶スペースのサイズが人為的に減少させられ、これによって、記憶ディスクが故障する場合、確実に十分な量のRAIDレベルの仮想記憶スペースが再構築のための使用可能になる。

【００５７】

同時にホット・スペアを保証しながら、アプリケーション・レベルの仮想記憶スペースの現在の量を、満たすことができなければ、RAID管理システムは、ホット・スペア・スペースが使用可能でないことをユーザに知らせる警告を生成する。ディスク・アレイがすでに一つの記憶ディスクをなくしているとき、この状況が起こるかもしれない。例えば、故障したディスク5をもつ図６のディスク・アレイは、別のディスク損失を受け、しかもユーザ・データの冗長性を維持することはできない。ユーザが最初にホット・スペア・オプションをイネーブルしないで、アプリケーション・レベルの仮想記憶スペースのサイズがセットされた後にこの機構を要求するときに、別の状況が起こりうる。ディスク・アレイは、その時点で、ホット・スペア資源を保証することができないかもしれない。

【００５８】

図8は、正常状態の間、およびディスク故障の場合のディスク・アレイの動作方法を示す。ステップ100において、RAID管理システム16は、ディスク・アレイ12によって提供される物理的な記憶スペースを第１のすなわちRAIDレベルの仮想記憶スペース40（図４参照）にマップする。先に述べたように、RAIDレベルの仮想記憶スペースは、ミラーおよびてパリティRAID領域として物理的な記憶スペースを提示する。ステップ102において、RAID管理システムは、ディスク故障の場合、ディスク・アレイにわたって十分な記憶スペースが使用可能にされることができるようにする。

【００５９】

データが最初にそのシステムに置かれるとき、そのデータは、完全にミラー記憶装置（ステップ104）に格納される。ユーザ・データおよび対応する冗長データは、均一にそのディスクにわたって広がるのがよい。ミラー記憶装置は、パリティ記憶装置と比較してパフォーマンスおよび信頼性がよく、したがって好ましい記憶型である。ユーザ・データの量がそのシステムにおけるそのディスクの物理的容量のおよそ二分の一に届くとき、ユーザ・データが記憶容量のおよそ二分の一を消費し、冗長性データが残りの半分を消費するから、ディスク・アレイは、実効的に満杯である。ユーザ・データの量が、50の%閾値を越えるにつれて、RAID管理システムは、これらのタイプの間でのデータ移動を使って、データをミラーおよびパリティRAID領域両方に格納し始める（ステップ106）。パリティ記憶装置を使うと、同じ合計スペースにミラー記憶よりも多くのユーザ・データを格納することができる。データ記憶システムがいっぱいになり続けると、増加する量のデータを収納するために、比例的により多くのデータがパリティ記憶に移行される。ミラー記憶装置におけるユーザ・データの量は、減少する。

【００６０】

記憶ディスクが、正常動作の間の任意の時点で故障する場合（破線によってステップ108に）、RAID管理システムは、再構築プロセスを始める。ステップ109において、RAID管理システムは、ディスク故障の場合に利用可能にされることができる適切な量のホット・スペア・スペースがあるかどうかを、評価する。言い換えると、故障した記憶ディスクを除

いたシステムで、割り当てた容量のすべてを冗長性とともに格納することができることができるかを評価する。ディスク故障の場合に（すなわち、ステップ109から「NO」分岐）割り当てた容量のすべてを冗長性とともに格納する、適切な量のスペースがなければ、ディスク・アレイ記憶システムは、ホット・スペア資源すなわち冗長性を復元する能力なしで劣化した条件で操作される（ステップ111）。

【００６１】

一方、ホット・スペアのための適切なスペースがあるならば（すなわち、ステップ109からの「YES」分岐）、RAID管理システムは、ユーザ・データを再建して、冗長性を復元する目的のために十分なスペースが利用可能にされているかどうかを、判断する（ステップ110）。ある場合（すなわち、ステップ110からの「YES」分岐）、ディスクアレイは、利用可能な自由スペースを使って、故障した記憶ディスクからユーザ・データを再建し冗長性を復元する。代わりに使用可能にされた十分なスペースがなければ（すなわち、ステップ112から「NO」分岐）、RAID管理システムは、故障した記憶ディスクで失われたスペースを補う利用可能な記憶スペースを空けるために、ミラー記憶装置からパリティ記憶装置にユーザ・データを移行させる。RAID管理システムは、十分な量の使用可能な記憶スペースが空けられ、故障したディスク・スペースの必要性がなくなるまで、ミラーからパリティ記憶装置にデータを移行し続ける（ステップ110および114のループによって表現されるように）。

【００６２】

この発明のディスク・アレイ・システムは、ディスク故障に続いて再構築を保証しながら、すべての記憶ディスクがデータをミラー記憶装置に格納するために等しく使われる点で、先行技術のディスク・アレイより有利である。故障が起こるまで、不使用になる専用記憶ディスクがない。ミラー記憶装置としてのホット・スペア記憶装置の使用は、並列性を増やし冗長な記憶スペースへの書き込みを管理する複雑さを減少させることによって、かなりディスク・アレイ全体のパフォーマンス を改善する。

【００６３】

しかし、この発明は、上述した特定の特長に限られるものではなく、均等論によって解釈される特許請求の範囲で規定する事項を包含する。この発明は、例として次の実施態様を含む。

【００６４】

1.記憶ディスクが故障した場合、記憶ディスク・アレイにおける再構築に使用するための記憶スペースを提供するための方法であって、
複数の記憶ディスクを有するディスク・アレイを準備するステップと、
ユーザ・データをすべての記憶ディスクに格納するステップと、
冗長データを記憶ディスクの中の少なくとも1つに格納するステップと、
記憶ディスクの一つが故障する場合、ユーザ・データを再建し、冗長性を復元するために十分な量の記憶スペースが利用可能にできるようにするステップと、
を含み、ディスク故障が起きるまで上記の量の記憶スペースをユーザ・データおよび冗長データの格納のために使用する、上記記憶スペースを提供する方法。

【００６５】

2. ユーザデータおよび冗長データを格納するために使われる記憶スペースを、そのユーザ・データを再建して、冗長性を復元するためのホット・スペア・スペースに変換するステップを含む上記１の方法。
3.ディスク故障に続いてディスク・アレイにおけるユーザ・データを再建し冗長性を復元するために、十分な量の記憶スペースを利用可能にすることができないことが判断されるとき、ホット・スペア資源が、保証されえないことを知らせる警告を、ユーザに発するステップを含んでいる上記1の方法。

【００６６】

4. 物理的な記憶スペース(34)を規定する選ばれた容量の複数の記憶ディスク（12）のディスク・アレイ（11）を含んでいる階層的なデータ記憶システム（10）おける再構築に使

用する記憶スペースを提供するための方法であって、
物理的な記憶スペース(34)をRAIDレベルの仮想記憶スペース(40)にマップする物理的な記憶スペースを、データをミラー冗長性に従って格納するミラーRAID領域およびデータをパリティ冗長に従って格納するパリティRAID領域を含む複数のRAID領域として提示するステップと、
ユーザ・データがディスク・アレイ（11）のすべての記憶ディスク（12）にわたって分布する態様で少くともミラーまたはパリティRAID領域の1つにユーザ・データを格納するステップと、
記憶ディスク故障の場合、ユーザ・データを再建し冗長性を復元するに十分な量のRAIDレベルの仮想記憶スペース(40)を利用可能にすることができることを確実にするステップと、
を含み、記憶ディスク故障が生じるまでミラー冗長性に従ってデータを格納するのに、ユーザ・データを再建し冗長性を回復するために利用可能にされることができる量の記憶スペースを使う、上記記憶スペースを提供する方法。

【００６７】

5. 次のステップを含む上記4の方法：
RAID-レベル仮想記憶スペース(40)をアプリケーション・レベルの仮想記憶スペース(50)にマップするステップと、
最も大きいサイズ記憶ディスクの容量を含むことなくユーザに割り当てることができるアプリケーション・レベルの仮想記憶スペース(50)の容量を計算するステップ。
6.十分な量のRAIDレベルの仮想記憶スペース(40)が、再構築のために利用可能にされることができることを確保するため、計算された容量を越えて付加的なアプリケーション・レベルの仮想記憶スペースを割り当てるのを防ぐステップを含む上記５の方法。
7.上記4の方法において、ディスク・アレイ（11）が、2つ以上の記憶装置ディスク（12）を持ち、ミラーRAID領域がデータを格納するのに、同じ量のデータを格納するのにパリティRAID領域によって消費されるよりも多くのRAIDレベルの仮想記憶スペース(40)を消費するようになっており、さらに、以下のステップを含む方法：
最初にユーザ・データをミラーRAID領域に格納するステップと、
記憶ディスク故障の場合、以下の2ステップの一つを行うステップ；
(a)故障した記憶ディスクからユーザ・データを再建し、冗長性を復元するための記憶スペースとして、ミラーまたはパリティRAID領域として割り当てられていない不使用の記憶スペースを使う、または
(b)ミラーRAID領域からパリティRAID領域にユーザ・データを移行させ、故障した記憶ディスクで失われたスペースを補うための記憶スペースを空けるためにミラー領域を詰め込む。

【００６８】

8.複数の記憶ディスク（12）を有するディスク・アレイ（11）と；
記憶ディスク(12)との間でユーザ・データの転送をコーディネートするためにディスク・アレイ（11）に結合されたディスク・アレイ・コントローラ（14）と；メモリ割り当ておよびディスク・アレイ（11）上のデータ記憶を管理するためディスク・アレイ・コントローラ(14)に結合されたメモリマネージャー（16）と；を含み、該メモリマネージャーは、すべての記憶ディスク（12）にわたってユーザ・データおよび冗長データを分配し、記憶ディスクの一つが故障する場合、ユーザ・データを再建し冗長性を回復するために利用可能にされることができる十分な量の記憶スペースを維持し、記憶ディスク故障に続く再構築のためにその記憶スペースが利用可能にされることを保証しながら、ユーザ・データおよび冗長性データを格納するためにその量の記憶スペースを使うようにしたデータ記憶システム。

【００６９】

9.記憶ディスク（12）が物理的な記憶空間(34)を規定し、
メモリ・マネージャー（16）が、RAIDレベルの仮想記憶スペース(40)を記憶ディスク（12

）の物理的な記憶スペース(34)にマップするRAID管理システムを含み、RAIDレベルの仮想記憶スペースが、物理的な記憶スペースを複数のRAID領域として提示するRAID領域は、データをミラー冗長に従って格納するミラーRAID領域、およびデータをパリティ冗長に従って格納するパリティRAID領域を含み、
さらに、RAID管理システムは、RAIDレベルの仮想記憶スペース(40)をアプリケーション・レベルの仮想記憶スペース(50)にマップし、最も大きいサイズ記憶ディスクの容量を含むことなくユーザに割り当てることができるアプリケーション・レベルの仮想記憶スペース(50)の容量を計算するようにした上記８のデータ記憶システム。

【００７０】

10． RAID管理システムは、十分な量のRAIDレベルの仮想記憶スペース(40)が、ユーザ・データを回復し冗長性を復元するために利用可能にされることを確実にするために、計算された容量を越えて、付加アプリケーション・レベルの仮想記憶スペース(50)の割り当てを防ぐ上記9のデータ記憶システム。

【００７１】

【発明の効果】

この発明によると、ディスク・アレイのすべての記憶ディスクを十分利用しながらホット・スペア資源が得られる。

【図面の簡単な説明】

【図１】この発明に従う冗長な階層的なディスク・アレイ・データ記憶システムの図表のブロック図である。

【図２】複数の記憶ディスク上の記憶スペースを示す図表で、RAIDレベル1によ るデータ記憶を示している。

【図３】複数の記憶ディスク上の記憶スペースの図表で、RAIDレベル５によるデータ記憶を示す。

【図４】2つの仮想記憶スペースが物理的な記憶スペースの上へマップされてい る、この発明のメモリーマッピング取り決めの図表である。

【図５】複数の記憶ディスク上の記憶スペースの図表であって、記憶ディスクの故障前のホット・スペア資源の使用を示す。

【図６】複数の記憶ディスク上の記憶スペースの図表であって、記憶ディスクの故障後のホット・スペア資源の使用を示す。

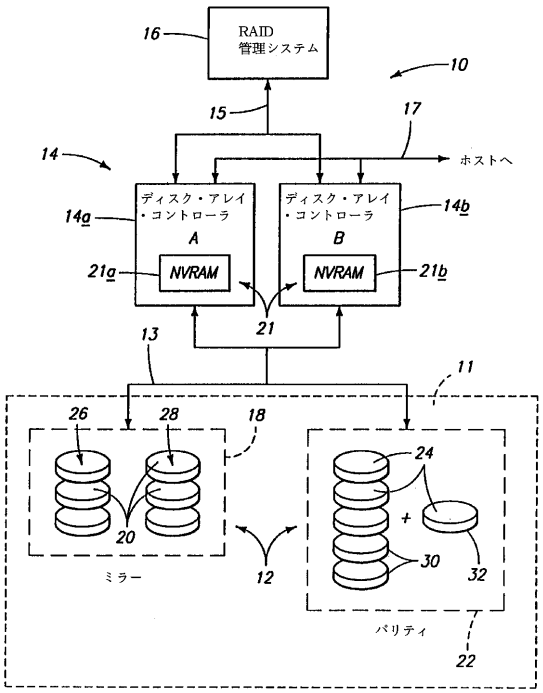
【図７】複数の記憶ディスク上の記憶スペースの図表であって、RAIDレベル1の記憶装置からRAIDレベル5の記憶装置へのデータの移動によってホット・スペアのためにどのようにスペースが使用可能になるかを示す。

【図８】図8は、ホット・スペア・スペースの可用性を保証するようディスク・アレイを操作する方法の流れ図である。

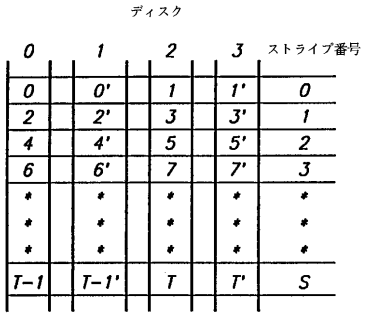
【符号の説明】

１０　データ記憶システム
１１　ディスク・アレイ
１２　記憶ディスク
１３　インターフェイス・バス
１４　ディスク・アレイ・コントローラ
１８　ミラー・グループ
２２　パリティ・グループ

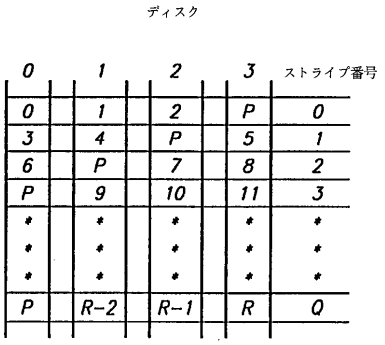
【図１】



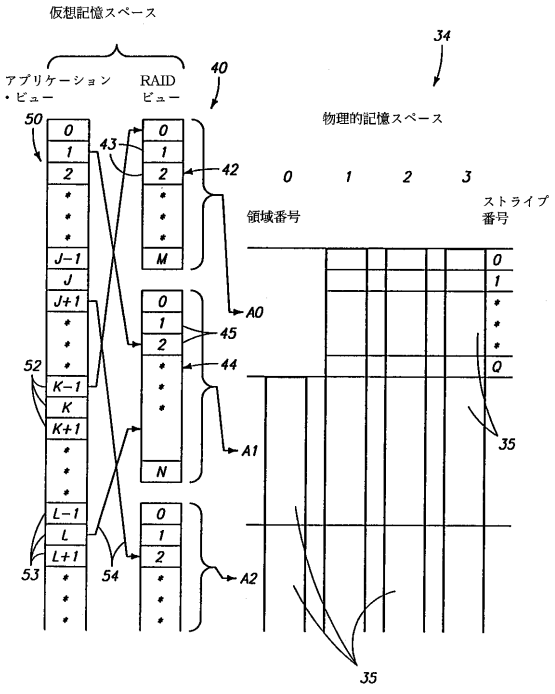
【図２】



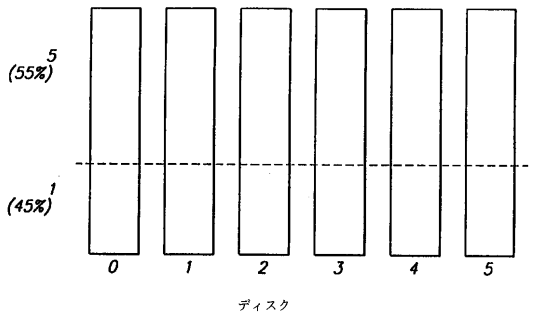
【図３】



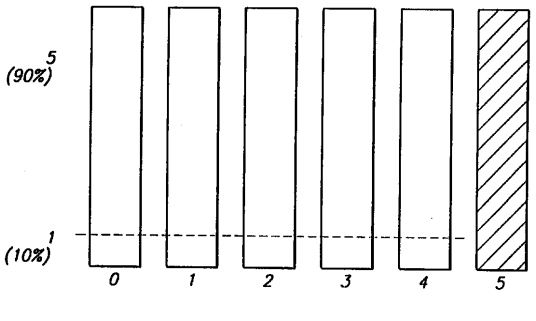
【図４】



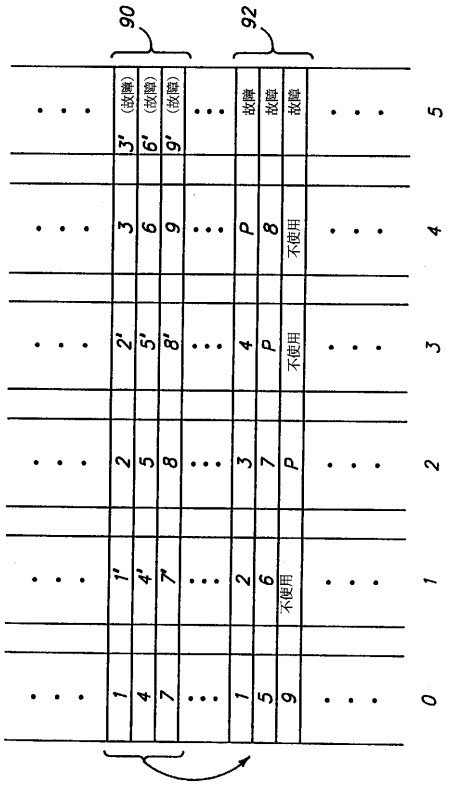
【図５】



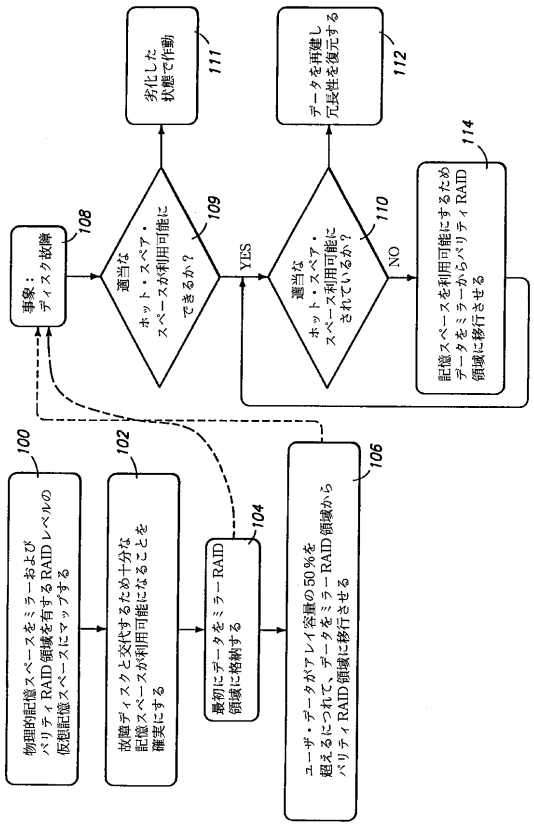
【図６】



【図７】


90
92
0
1
2
3
4
5
故障
不使用
P

【図８】


事象：ディスク故障
108
適当なホット・スペア・スペースが利用可能にできるか？
109
YES
NO
劣化した状態で作動
111
適当なホット・スペア・スペースが利用可能にされているか？
110
データを再建し冗長性を復元する
112
記憶スペースを利用可能にするためデータをミラーからパリティRAID領域に移行させる
114
物理的記憶スペースをミラーおよびパリティRAID領域を有するRAIDレベルの仮想記憶スペースにマップする
100
故障ディスクと交代するため十分な記憶スペースが利用可能になることを確実にする
102
最初にデータをミラーRAID領域に格納する
104
ユーザ・データがアレイ容量の50％を超えるにつれて、データをミラーRAID領域からパリティRAID領域に移行させる
106

フロントページの続き

(72)発明者 ブライアン・エム・ダイアモンド
アメリカ合衆国８３７０４アイダホ州ボイジー、ラムスゲート ９７０４
(72)発明者 マイケル・ビー・ジャコブソン
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 ウェイドゥ・エー・ドルフィン
アメリカ合衆国８３７０４アイダホ州ボイジー、ペトリエ ７１７９
(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、ノース・トゥエンティフォース ３０３０

審査官 木村 貴俊

(56)参考文献 特表平０５－５０５４８１（ＪＰ，Ａ）
Hybrid Redundancy Direct-Access Storage Device Array with Design Options, IBM Technical Disclosure Bulletin, 米国, IBM Corp., １９９４年 2月, Vol. 37 No. 02B, p.141-148

(58)調査した分野(Int.Cl., ＤＢ名)
G06F 3/06- 3/08
G06F 12/00-12/16
G11B 19/00-20/18