(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4163298号
(P4163298)
(45)発行日　平成20年10月8日(2008.10.8)　(24)登録日　平成20年8月1日(2008.8.1)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 305C
G06F 3/06 304P
G06F 3/06 540

請求項の数 12　(全 18 頁)

(21)出願番号　特願平10-235862
(22)出願日　平成10年8月21日(1998.8.21)
(65)公開番号　特開平11-119919
(43)公開日　平成11年4月30日(1999.4.30)
審査請求日　平成17年8月16日(2005.8.16)
(31)優先権主張番号　08/920-120
(32)優先日　平成9年8月26日(1997.8.26)
(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　100075513
弁理士　後藤　政喜
(74)代理人　100084537
弁理士　松田　嘉夫
(72)発明者　ダグラス・エル・ヴォイグト
アメリカ合衆国　アイダホ，ボイセ，エヌ・24ス　3030

最終頁に続く

(54)【発明の名称】記憶システムへのデータ書き込み方法

(57)【特許請求の範囲】
【請求項１】
コンピュータによって行われる、複数の記憶媒体を有する記憶システムへの書き込み方法であり、メモリに格納されているトランザクションログを、前記複数の記憶媒体に設けられたディスクステージングログ、および、前記ディスクステージングログとは異なるディスクログの２つの領域のうちのいずれか一方の領域に書き込む方法であって、
前記トランザクションログを格納しているメモリがページフルの状態になると、前記ディスクログに前記トランザクションログを書き込むことと、
前記メモリに格納されているトランザクションログの強制引出し要求に応じて、前記ディスクステージングログを構成する記憶媒体のうち、アクセス頻度が最低の記憶媒体（以下、最低使用頻度の記憶媒体）に前記トランザクションログを書き込むことと、
を含むことを特徴とする記憶システムへのデータ書き込み方法。
【請求項２】
前記複数の記憶媒体は、ランダムアクセス記憶媒体であることを特徴とする請求項１に記載の記憶システムへのデータ書き込み方法。
【請求項３】
前記最低使用頻度の記憶媒体を選択することは、入出力動作に基づいて当該の選択が行われることを特徴とする請求項１または２に記載の記憶システムへのデータ書き込み方法。
【請求項４】

前記強制引出し的に書き込む要求は、前記トランザクションログを格納しているメモリがページフルの状態になる前に発生することを特徴とする請求項１、２または３に記載の記憶システムへのデータ書き込み方法。
【請求項５】
前記トランザクションログは、前記選択された最低使用頻度の記憶媒体に非冗長に書き込まれることを特徴とする請求項１に記載の記憶システムへのデータ書き込み方法。
【請求項６】
前記トランザクションログは、前記書き込まれるトランザクションログの順序を示す標識を含むことを特徴とする請求項１に記載の記憶システムへのデータ書き込み方法。
【請求項７】
前記記憶媒体はそれぞれ、前記最低使用頻度の記憶媒体が選択された際の書き込みのためにのみ使用されるように確保された領域を含むことを特徴とする請求項１に記載の記憶システムへのデータ書き込み方法。
【請求項８】
前記確保された領域は、少なくとも２つのサブ領域を含み、
最低使用頻度の記憶媒体が選択される最初の事象の発生時に前記サブ領域の１つに対して書き込みが行なわれ、
最低使用頻度記憶媒体が選択される次の事象の発生時に他方のサブ領域への書き込みが行なわれ、
それによって、同じ最低使用頻度記憶媒体が二度連続して選択された場合であっても、次の連続する書き込みにおいて直前に書き込まれたサブ領域に対して重ね書きが生じないことを特徴とする請求項７に記載の記憶システムへのデータ書き込み方法。
【請求項９】
記憶システムであって、
（ａ）データ記録を保持する第１のメモリと、
（ｂ）前記第１のメモリに接続され、ディスクステージングログ、および、前記ディスクステージングログとは異なるディスクログが設けられる複数の記憶媒体と、
（ｃ）前記第１のメモリの状態を検出する手段と、
（ｄ）前記第１のメモリがページフルの状態であることが検出されると、前記ディスクログに前記データ記録を書き込む手段と、
（ｅ）前記第１のメモリに保持されている前記データ記録の強制引出し要求に応じて、前記ディスクステージングログを構成する記憶媒体のうち、アクセス頻度が最低の記憶媒体（以下、最低使用頻度の記憶媒体）に前記データ記録を書き込む手段と、
を含むことを特徴とする記憶システム。
【請求項１０】
前記第１のメモリがページフルの状態であることが検出された時に、前記ディスクログに前記データ記録を書き込む手段は、前記複数の記憶媒体上の前記データ記録の冗長性を維持することを特徴とする請求項９記載の記憶システム。
【請求項１１】
前記第１のメモリに保持されている前記データ記録の強制引出し要求に応じて、前記最低使用頻度の記憶媒体に前記データ記録を書き込む手段は、前記複数の記憶媒体上において前記データ記録の冗長性を維持しないことを特徴とする請求項９記載の記憶システム。
【請求項１２】
前記最低使用頻度の記憶媒体は、前記データ記録の書き込みのために確保された少なくとも２つのサブ領域を含むことを特徴とする請求項９記載の記憶システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は一般的にはデータ記憶システムに関し、特にディスクアレイ記憶システムのためのトランザクションログ（ transaction log）管理に関する。

【０００２】
【従来の技術】
コンピュータシステムの速度、信頼性および処理能力は絶えず進歩し続けている。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を処理することができる。コンピュータの改良にともなって、大量記憶および入出力（Ｉ／Ｏ）装置の性能に対する要求も高くなる。したがって、進歩し続けるコンピュータシステムに性能上つりあう大量記憶システムを設計することが常に必要とされている。
【０００３】
本発明は特にディスクアレイ型の大量記憶装置に関する。ディスクアレイデータ記憶システムは単一の大量記憶システムを形成するように構成され、統合された多数の記憶ディスクドライブ装置を有する。大量記憶システムには、コスト、性能および利用可能性という３つの主要な設計評価基準がある。メガバイトあたりのコストが低く、入出力性能が高く、データの利用可能性が高いメモリ装置を製作することが最も望ましい。“利用可能性"とは、記憶システムに記憶されたデータにアクセスする能力のことであり、またなんらかの故障があった場合に連続動作を保証する能力のことをいう。通常、データの利用可能性は冗長性を用いて提供され、この場合データあるいはデータ間の関係が複数の場所に記憶される。冗長データの記憶には、“ミラー"法および“パリティ"法の２つの一般的な方法がある。
【０００４】
【発明が解決しようとする課題】
ディスクアレイデータ記憶システムの設計にあたって発生する問題の１つは、システムの誤りあるいは故障の場合における記憶されたデータの正確なマッピング情報の保持の問題に関係するものである。これは前記の冗長データ記憶法のいずれかあるいはその両方を用いるシステムについていえることである。したがって、ディスクアレイマッピング情報の管理において、誤りから回復する目的のためには、最近変更されたマッピング情報がディスク上に確実に記憶されるようにする必要がある場合が多い。このディスク書き込みの必要性は、（ｉ）時間に基づいた頻度状態の更新、（ｉｉ）ログページ・フルが状態、あるいは（ｉｉｉ）特定のホストによる要求等のいくつかの理由で発生する。一般的には、最近のマッピング情報の変更は、ディスクアレイ機能の性能対して最適化されたデータ構造内におけるランダムな場所に蓄積され、さらに他のデータ構造より高速にディスクに書き込む（ポストする（ post ））ことができるログに順次蓄積される。この技術はトランザクション処理技術においては周知である。しかし、ポストの必要性が進行中の他のディスク読み出しあるいは書き込み動作と同時に発生してシステム内に入出力の競合が生じるという問題が発生する場合がある。かかる入出力競合は、特にポストが頻繁に発生する場合にシステム上の重要な性能を阻害することが多い。これはディスクにログを１回ポストするには複数の入出力事象が必然的に発生するためである。たとえば、通常、ログページはまず無効と表示される（すなわち、更新が必要である）。次に、ログページはディスクにコピーされ、その後有効と表示される。最後に、冗長システムでは、冗長ログページがディスクにコピーされる。
【０００５】
以上のことから、また増大し続ける計算速度および管理対象となる膨大な情報量から、ディスクアレイシステム等の性能の改善が常に必要とされている。
【０００６】
したがって、本発明はディスクアレイマッピング情報の管理システムの性能を向上させることを目的とする。また、複数の利用可能な記憶ディスクから選択された任意の最も使用頻度の低いディスクへのログ書き込みを管理および分散してログ入出力と進行中の他の入出力との間におけるディスクアクセスの競合を低減することによって、ディスクログ書き込みのシステム性能を改善し、さらにシステムの誤り又は故障時に記録を確実に回復可能にすることを目的とする。
【０００７】

【課題を解決するための手段】
本発明による記憶システムへの書き込み方法は、コンピュータによって行われる、複数の記憶媒体（１２）を有する記憶システム（１０）への書き込み方法であって、
（ａ）第１のトランザクションログを記憶媒体（１２）に書き込む第１の要求を示す第１の基準を検出することと、
（ｂ）第１の要求に応じて、第１のトランザクションログを、最低使用頻度の記憶媒体であるかないかにかかわらず、記憶媒体（１２）に書き込むことと、
（ｃ）第２のトランザクションログを記憶媒体（１２）に強制引出し的に書き込む要求を含む第２の要求であって、第２のトランザクションログを記憶媒体（１２）に書き込む第２の要求を示す第２の基準を検出することと、
（ｄ）複数の記憶媒体（１２）から最低使用頻度の記憶媒体を選択することと、
（ｅ）第２の要求に応じて、第２のトランザクションログを選択された最低使用頻度の記憶媒体へ書き込むこととを含むことを特徴とする。
また、本発明による記憶システムは、
（ａ）データ記録（１１０）を保持する第１のメモリ（５５）と、
（ｂ）第１のメモリに接続された複数の記憶媒体（１２）と、
（ｃ）第１のメモリ（５５）の状態を検出する手段（１６）と、
（ｄ）第１のメモリにおける第１の検出された状態に応答して第１の記憶管理基準にしたがって記憶媒体（１２）にデータ記録を書き込む手段（１６）と、
（ｅ）第１のメモリにおける第２の検出された状態に応答して第２の記憶管理基準にしたがって記憶媒体（１２）にデータ記録を書き込む手段（１６）と
を含み、
第１の記憶管理基準は、最低使用頻度の記憶媒体であるかないかにかかわらず、記憶媒体（１２）への書き込みを示す基準を含み、第２の記憶管理基準は、複数の記憶媒体における最低使用頻度の記憶媒体への書き込みを示す基準を含み、第１のメモリ（５５）における第１の検出された状態は、第１のメモリの一部がデータ記録で満たされていることを示す状態を含み、第１のメモリ（５５）における第２の検出された状態は、第１のメモリにおいて所定の部分がデータ記録で満たされる前に、第１のメモリ（５５）内のデータ記録を複数の記憶媒体（１２）に強制引出し的に書き込む要求を示す状態を含むことを特徴とする。

【０００８】
一実施形態における本発明の原理によれば、ディスクドライブ等の複数の記憶媒体を有する記憶システムにおいて、第１メモリに記憶されたトランザクションログが、２つの異なるログ領域に選択的にポストされる。詳細には、第１メモリのトランザクションログのページ・フル状態が検出されると、ポストは"ディスクログ"領域に対して行なわれる。ポスト要求がトランザクションログのページ・フル状態が検出される前に発生すると、ポストはただちに"ステージングログ"領域の最も使用頻度の低いディスクに対して行なわれる。
他の原理によれば、"ディスクログ"へのポストは通常の記憶システム管理技術およびデータ冗長性技術を用いて行なわれる。一方、"ステージングログ"領域へのポストは、ステージングログのデータが記憶媒体上で冗長な状態で保持されることのないように通常の記憶システムデータ管理および冗長性技術を用いずに行なわれる。冗長性はトランザクションログがステージングログ領域にコピーされることに加えて第１メモリに残ることによって維持される。

【０００９】
ステージングログ領域は複数の記憶媒体のそれぞれに確保されたスペースを含み、かかる確保されたスペースは各記憶媒体上で論理的に分離された部分に分割される。この構成によって、ステージングログへのポストを確保された部分の間で“トグルする（ toggle ）"ことができる。したがって、２つの連続するステージングログポストによって、どのディスクが最も使用頻度が低いかにかかわりなく確保された領域の同じ部分が重ね書きされ

ることはない。
【００１０】
他の原理によれば、記憶媒体上のログ領域にポストされるデータ記録にシーケンス番号およびディスク群番号が割り当てられる。ディスクログ領域およびステージングログ領域からのデータの回復においては、シーケンス番号およびディスク群番号が参照されて完全なトランザクションログが適正に再構築される。
【００１１】
本発明の他の目的、利点および機能は以下の説明から明らかになるであろう。
【００１２】
【発明の実施の形態】
図１には、本発明の分散書き込みディスクログ法を用いたデータ記憶システム１０のブロック図を示す。図示する例では、データ記憶システム１０は階層的ディスクアレイ１１を含むディスクアレイデータ記憶システムである。本発明は非階層的アレイ（図示せず）にも適用可能である。ディスクアレイ１１はＲＡＩＤ(Redundant Array of Independent Disks)記憶システムを実施するための複数の記憶ディスク１２を含む。データ記憶システム１０はディスクアレイ１１に結合され記憶ディスク１２との間のデータ転送を調整するディスクアレイコントローラ１４を含み、さらにＲＡＩＤ管理システム１６を含む。ＲＡＩＤ管理システム１６は本発明の分散書き込みディスクログ法を実行する手段を含む。
【００１３】
本明細書においては、“ディスク"とは自己の記憶故障を検出することのできる任意の不揮発性のランダムアクセス可能・書き換え可能な大量記憶装置である。ディスクには、回転磁気ディスクおよび光ディスクとソリッドステートディスクの両方あるいは（PROM、EPROMおよびEEPROM等の）不揮発性電子記憶素子を含む。“ディスクアレイ"という用語は、ディスクと、ディスクを１つあるいはそれ以上のホストコンピュータに接続するのに要するハードウエアと、物理的ディスクの動作を制御しそれらをホスト動作環境に１つあるいはそれ以上の仮想ディスクとして提示するのに必要な管理ソフトウエアの集合である。“仮想ディスク"は管理ソフトウエアによってディスクアレイ中に実現される抽象的存在である。
【００１４】
“ＲＡＩＤ"という用語はその物理的記憶容量の一部が記憶容量の残りの部分に記憶されたユーザデータに関する冗長な情報の記憶に用いられるディスクアレイを意味する。この冗長情報によって、このアレイを構成するディスクの１つあるいはこのアレイへのアクセス経路が故障した場合にユーザデータを再生することができる。ＲＡＩＤシステムについては、ミネソタ州Lino LakesのREID Advisory Boardから1993年6月9日に刊行された“The RAID Book: A Source Book for RAID Technology"に詳細に説明されている。ＲＡＩＤシステムを本発明との関係において例示するが、本発明は非ＲＡＩＤシステムにも適用可能であることはいうまでもない。
【００１５】
ディスクアレイコントローラ１４は、small computer system interface（ＳＣＳＩ）等の１つあるいはそれ以上のインターフェースバス１３を介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によってディスクアレイコントローラ１４に操作的に結合されている。ＲＡＩＤ管理システム１６は図示するように別個の要素として（すなわちソフトウエアあるいはファームウエアとして）実施することもでき、あるいはディスクアレイコントローラ１４内あるいはホストコンピュータ内に構成して、ディスクの記憶および信頼性レベルの制御、さまざまな信頼性の記憶装置レベル間でのデータ転送、および本発明の分散書き込みディスクロギングの実施を行なうデータ管理手段を提供することもできる。また、データ記憶システム１０は入出力インターフェースバス１７を介してホストコンピュータ（図示せず）に結合されている。
【００１６】
図示するシステムでは、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およびディスクアレイコントローラ“Ｂ"１４Ｂからなるデュアルコントローラとして実施される。デュアルコントローラ１４Ａおよび１４Ｂは一方のコントローラが動作不能となったとき連続的なバックアップと冗長性を供給することによって信頼性を向上させる。しかし、本発明の方法は単一のコントローラあるいは他のアーキテクチャで実施することができる。実際に、本発明は完全で正確なディスクログの維持がデュアルコントローラ環境におけるよりも重要である単一コントローラアーキテクチャにおいて特に有益である。

【００１７】

階層的ディスクアレイ１１は物理的記憶スペースと１つあるいはそれ以上の仮想記憶スペースを含む異なる記憶スペースとして特徴付けることができる。たとえば、ディスクアレイ１１内の記憶ディスク１２は、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に構成されるものとして概念化することができる。記憶装置のかかる諸相はマッピング技術を用いて関係付けられる。たとえば、ディスクアレイの物理的記憶スペースは記憶領域をさまざまなデータ信頼性レベルに応じて区分した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領域の一部をミラーすなわちＲＡＩＤレベル１の第１の信頼性の記憶レベルに割り当て、他の領域をパリティすなわちＲＡＩＤレベル５の第２の信頼性の記憶レベルに割り当てることができる。かかる領域は同じディスクあるいは別々のディスク上に構成することができ、また任意の組み合わせのディスク上に構成することもできる。

【００１８】

データ記憶システム１０はディスクアレイ１１のマッピングに用いる仮想マッピング情報の永続的な記憶を可能とするメモリマップ記憶域２１を含む。このメモリマップ記憶域はディスクアレイの外部にあり、好適にはディスクアレイコントローラ１４に常駐する。メモリマッピング情報は異なるビュー（ view ）の間でさまざまなマッピング構成が変化するにつれてディスクアレイコントローラ１４あるいはＲＡＩＤ管理システム１６によって連続的あるいは定期的に更新することができる。

【００１９】

好適には、メモリマップ記憶域２１はそれぞれディスクアレイコントローラ“Ａ"１４Ａおよびディスクアレイコントローラ“Ｂ"１４Ｂに設けられた２つの不揮発性ＲＡＭ（ Non-Volatile RAM ）２１Ａおよび２１Ｂとして実施される。これら２つのＮＶＲＡＭ２１Ａおよび２１Ｂはメモリマッピング情報の冗長記憶を可能とする。仮想マッピング情報はミラー冗長性技術によってＮＶＲＡＭ２１ＡおよびＮＶＲＡＭ２１Ｂの両方に複製され記憶される。これによって、ＮＶＲＡＭ２１Ａをオリジナルのマッピング情報の記憶にのみ用い、ＮＶＲＡＭ２１Ｂを冗長マッピング情報の記憶にのみ用いることができる。

【００２０】

図示するように、ディスクアレイ１１は複数の記憶ディスク１２を有する。記憶ディスク１２上の冗長性の管理はＲＡＩＤ管理システム１６によって統御される。ユーザすなわちホストアプリケーションプログラムから見た際、アプリケーションレベルの仮想ビューによって記憶ディスク１２上の利用可能な記憶スペースを示す１つの大きな記憶容量を表わすことができる。ＲＡＩＤ管理システム１６はこの物理的記憶スペース上でのＲＡＩＤ領域の構成を動的に変更することができる。その結果、ＲＡＩＤレベル仮想ビュー内のＲＡＩＤ領域のディスクへのマッピングおよびフロントエンド仮想ビューのＲＡＩＤビューへのマッピングは一般的にはある変化の状態ということになる。ＮＶＲＡＭ２１ＡおよびＮＶＲＡＭ２１Ｂ内のメモリマップ記憶域は、ＲＡＩＤ管理システム１６によるＲＡＩＤ領域のディスクへのマッピングに用いられる現在のマッピング情報および２つの仮想ビューの間でのマッピングに用いられる情報を保持する。ＲＡＩＤ管理システムはＲＡＩＤレベルのマッピングを動的に変更するとき、メモリマップ記憶域のマッピング情報をかかる変更を反映するように更新する。

【００２１】

しかし、ディスクアレイに用いられるＲＡＩＤ機構すなわちデータ記憶機構にかかわりな

く、メモリマップ記憶域２１は一般的にはシステムの使用時全体を通じて常に変化する状態にあることは明らかである。したがって、メモリマップログ記録がメモリに保持され、ＲＡＩＤ管理システム１６によってメモリからディスクに絶えずポストされ、ＮＶＲＡＭ２１の損失時にかかる記録が確実に回復されるようにする。よって、本発明は複数の利用可能な記憶ディスク１２から選択された任意の最も使用頻度の低いディスクへのログ書き込みを管理および分散してログ入出力と進行中の他の入出力との間におけるディスクアクセスの競合を低減することによってディスクログ書き込みのシステム性能を改善するものである。一般的には、これはログの新しい部分を保持するために各記憶ディスク１２上に“ステージングログ"領域を確保することによって行なわれる。そして、トランザクションログメモリのページがいっぱいになる前にポスト要求が発生した場合、ポストは最も使用頻度の低いディスクに確保された“ステージングログ"領域にほとんど即時に実行される。続いて、ログの回復が必要である場合、すべての記憶ディスク１２からの断片がまとめられて単一の完全なイメージが形成される。

【００２２】

図２は、本発明の分散ログ書き込みディスクログ法を示すブロック図である。ＮＶＲＡＭマップ４５はデータ記憶システム１０に用いられるデータが記憶されるディスクアレイコントローラ１４Ａ／１４Ｂ（図１）上における不揮発性のメモリマップ記憶域２１の部分集合を表わす。ディスクマップ５０はディスクアレイ１１に（冗長的に）属するＮＶＲＡＭマップ４５の従来のディスクマップイメージである。ディスクマップ５０へのＮＶＲＡＭマップ４５の定期的記憶（ポスト）によって誤り訂正を行なうために、ＮＶＲＡＭマップ４５の内容の冗長コピーをディスクに維持する手段が提供される。一般的には、ディスクマップ５０へのＮＶＲＡＭマップ４５のポストは、通常のシステム処理および入出力競合状態下で可能である際に（ＲＡＩＤ管理システム１６による制御のもとに）バックグラウンド処理として実行される。よって、ディスクマップ５０へのＮＶＲＡＭマップ４５データのポストは、入出力およびディスクスペースに関する通常のシステム競合の影響を受け、したがってポストが実際にいつ発生するかについては不確定要素がある。

【００２３】

ＲＡＭログイメージ（ＲＬＩ）５５もまた不揮発性メモリ２１の部分集合である。あるいはこれは別個の（好適には不揮発性の）メモリとすることもできる。ＲＬＩ５５はＮＶＲＡＭマップ４５内で発生するインクリメンタルな変化を迅速に記憶／記録するのに用いられる。一実施形態では、ＲＬＩ５５は１６（図ではＮ）のアドレス指定可能な６４Ｋバイトページを含むが、他の構成も可能である。その後ＲＡＩＤ管理システム１６からの要求があった際、ＲＬＩ５５に記憶されたインクリメンタルな変化はディスクログ６０あるいはディスクステージングログ６５にポストされる。

【００２４】

いくつかの要因によって、ＲＡＩＤ管理システム１６にＲＬＩ５５からディスクログ６０あるいはディスクステージングログ６５へのデータのポスト要求を行なう。たとえば、一実施形態では、ＲＬＩが“ページ・フル"状態になった際、ディスクログ６０への“排出（flush）"ポスト要求が発生する。“排出"ポスト要求が発生すると、ＲＬＩ５５からのトランザクションログデータのフル・ページがディスクログ６０に書き込まれる。一方、（ｉ）時間に基づく頻度の要求が発生するか、（ｉｉ）ある特定のホスト要求を受けた場合に、ディスクステージングログ６５への“強制引出し"ポスト要求が発生する。“強制引出し"ポスト要求が発生すると、ＲＬＩ５５から１つあるいはそれ以上のトランザクションログデータのフル・ブロックがディスクステージングログ６５に書き込まれる。この１つあるいはそれ以上のブロックには、前に完全に書き込まれておらず、また未書き込みの１つあるいはそれ以上のトランザクションログ記録を含む（現在のページ内の）ブロックが含まれる。ディスクログ６０に“排出"ポストで書き込みされるページおよびディスクステージングログ６５に“強制引出し" ポストで書き込みされるブロックをここではＲＬＩ５５の“未書き込み"データと称する（ただし、“排出" ポストで書き込みされたページはそれ以前にディスクステージングログ６５に“強制引出し" ポストで書き込みされた

記録を含む場合がある）。いずれの場合にも、（ディスクログ６０あるいはディスクステージングログ６５への）かかるポストによって、（ＲＬＩ５５で捕捉された）ＮＶＲＡＭマップ４５の変化が、ディスクマップ５０が更新されていない際にＮＶＲＡＭ２１に損失が生じた場合においての、誤り回復目的のためにディスクアレイ１１に確実に記憶されることを保証する。

【００２５】

ディスクログ６０はディスクアレイ１１（図１）上に常駐するＲＬＩ５５の従来と同様のディスクイメージである。好適には、ディスクログ６０はＲＬＩ５５と同様に多数のデータページを記憶することができる。図示するように、ディスクログ６０にはデータ記憶用のＮのページが示され、また従来と同様に、連続的あるいは円形にリンクすることができる。ディスクログ６０は（図１の）通常のデータ冗長性機構を用いてディスクアレイ１１上に記憶および管理される。したがって、ディスクログ６０へのＲＬＩ５５における“未書き込み"内容の“排出"ポストは通常の入出力状態で発生し、ディスクアクセスおよびスペースのためのシステム入出力競合の影響を受ける。ディスクログ６０は一般的にはディスクマップ５０より頻繁に更新されるが、ディスクログ６０は（ＲＬＩ５５中で捕捉された）ＮＶＲＡＭマップ４５へのインクリメンタルな変化のみを保持するのに対して、ディスクマップ５０は（最終更新時の）ＮＶＲＡＭマップ４５の完成イメージを保持する。

【００２６】

ディスクステージングログ６５はディスクアレイ１１（図１）のディスクの各部分を表わす確保されたステージング領域７０、７５、８０、８５、９０、９５、１００および１０５を含む。上述したように、一実施形態では、ディスクステージングログ６５は指定された事象の発生時あるいは“ページ・フル"状態以外の時にＲＬＩ５５の内容の記憶に用いられる。しかし、このポスト基準には当業者には明らかなようにシステム設計の変更および／またはユーザによる優先的な指示に合わせて自由度を持たせることができる。いずれの場合にも、（“ページ・フル"状態以外の）ＲＡＩＤ管理システム１６によって要求された所定の事象が発生した場合、ＲＬＩ５５はその“未書き込みの"内容をディスクステージングログ６５のディスク１～Ｍのうち最も使用頻度の低いディスクに“強制引出し"する。最低使用頻度のディスクはディスクアレイ１１のディスク１～Ｍの入出力動作をモニターすることによって検出される。

【００２７】

基本的には、最低使用頻度のディスクへＲＬＩ５５を“強制引出し"することによって、時間の経過とともにディスクアレイ全体にトランザクションログの分散書き込みが実行される。これは、所定の単一のディスクログ６０へのＲＬＩ５５のページ・フル“排出"とは対照的である。パリティ冗長機構を用いる場合、ディスクログ６０は実際には複数のディスクに分散させることができるが基本的には“単一の"すなわち“非分散の"ディスクログである。これは（冗長コピーを考えなければ）１つのディスクドライブ上で１つの基底アドレスのみを用いてログ全体をアドレス指定／アクセスすることができるためである。

【００２８】

最低使用頻度のディスクが選択されるため、“強制引出し"ポストが（進行中の他のシステム呼び出し／書き込み入出力動作との）入出力競合が低減された状態で発生するという利点がある。したがって、ディスクマップ５０あるいはディスクログ６０のポストとは異なり、ディスクステージングログ６５へのこの分散書き込みは一般的にはただちに（あるいは、少なくともより迅速に）完了する。さらに、“強制引出し"ポストは一般的には転送される未書き込みブロックが最小限であるため“排出"ポストより高速である。

【００２９】

ディスクログ６０とは異なり、ディスクステージングログ６５はＲＬＩ５５のインクリメンタルな変化をディスクアレイ１１全体にわたって分散した非冗長的な態様で保持する。これは、ディスクステージングログ６５に発生する書き込みがＲＡＩＤ管理システム１６の通常の冗長性機構から除外されることから非冗長的である。したがって、ディスクステージングログ６５へのポストにおいて発生する入出力ステップはディスクログ６０の場合

に比べて少なくとも１ステップ少ない。トランザクションログはステージングログ領域にコピーされることに加えて第１メモリ（ＲＬＩ５５）にも残るため、“強制引出し"ポスト後にも冗長性が維持される。

【００３０】

一実施形態では、ディスクアレイ１１のディスク１～Mはそれぞれ分散されたログの記憶のために確保された専用のスペース量を有する。たとえば、図示するそれぞれのディスク上には２つの６４Ｋバイトページ７０／７５、８０／８５、９０／９５および１００／１０５が確保される。ディスクステージングログポスト処理中の故障の場合に発生する可能性のある有効データの重ね書き（および損失）を防止するために、それぞれのディスク上に少なくとも２つのページが確保される。すなわち、ＲＬＩ５５はディスクステージングログ６５への書き込み（ポスト／強制引出し）を奇数および偶数ページに交互に実行する（スワッピングすなわちトグル）。たとえば、最初の書き込みでは、ＲＬＩ５５は最低使用頻度のディスクに確保された偶数番号のページ７０、８０、９０あるいは１００にポストを行なう。次に発生する書き込みにおいては、ＲＬＩ５５は最低使用頻度のディスクの奇数番号ページ７５、８５、９５あるいは１０５にポストする。これによって、システムはデータの完全性の別のレベルを保証され、続いて発生するポストにおいて最後にポストされたデータの重ね書きの可能性（すなわち、同じ最低使用頻度ディスクが連続して選択される場合の重ね書き）が防止される。

【００３１】

図３から図６は、ある時間における本発明のディスクステージングログの状態を示すブロック図である。本発明の分散書き込みディスクステージング動作の例をさらに詳細に説明するために、ＲＬＩ５５のログイメージページ５７の一部、およびディスクステージングログ６５の各ページ７０～１０５の一部を示す。すなわち、図３から図６にはそれぞれＲＬＩ５５からの異なるポストに応答してディスクステージングログ６５の状態時間における異なるスナップショットを示す。ＲＬＩ５５のログイメージページ５７およびディスクステージングログ６５の各ページ７０～１０５は破線によって３つの５１２バイトブロック（すなわち部分）Ｂ１、Ｂ２およびＢ３に（論理的に）分割されているものとして示されている。説明を簡略化するために、それぞれの６４Ｋバイトページ中の全ブロックではなく３つのブロックのみを示す。（ＲＬＩ５５の）ログイメージページ５７を本説明および図では“ＬＩ"で示す。さらに、ディスクステージングログ６５中の各ディスクをそれぞれ“Ｄ１～ＤM"で示し、各ディスク中に確保される２つのページをそれぞれ“Ｐ１"あるいは“Ｐ２"で示す。

【００３２】

図３において、論理標識である事象／時刻Ｔ１は（図１のＲＡＩＤ管理システム１６の要求によって）ある特定の事象が発生してＲＬＩ５５におけるログイメージページ５７の未書き込みデータのディスクステージングログ６５への“強制引出し"ポストが開始されるある所定の時点を反映している。事象／時刻Ｔ１はさらにその所定の時点においてＲＬＩのログイメージページ５７にログデータがどれだけ“フル"かを示す位置を同定する。ポストが要求されると、論理標識である事象／時刻Ｔ１によって同定される（ＲＬＩ５５におけるログイメージページ５７の）未書き込みデータの全ブロックがポストされる。５１２バイトデータブロックは（この例では任意のシステム設計条件に対する）最小ポストサイズであるため全てのブロックがポストされる。

【００３３】

したがって、たとえば、事象／時刻Ｔ１の発生時にＲＬＩ５５はログイメージページ５７の（事象／時刻Ｔ１で示される位置）“未書き込みの"内容をディスクステージングログ６５の最低使用頻度のディスク１～Mにおけるページ７０～１０５の１つに（後に詳述する“トグル"状態で）ポスト（ post ）する。すなわち、ログイメージページ５７のブロック“１"（ＬＩＢ１）は“未書き込み"でしかも完全にフルであるためその全体がポストされ、またログイメージのブロック“２"（ＬＩＢ２）も“未書き込み"であるためこれもその全体がポストされる（ただし、ログデータは事象／時刻Ｔ１まではＬＩＢ２の一部し

か満たしていない）。ディスク２が最低使用頻度のディスクとして検出され、ポストがディスクステージングログ６５の偶数ページ番号から開始されるものと（便宜上）仮定すると、ログイメージページ５７はそのブロック内容ＬＩＢ１およびＬＩＢ２をディスクステージングログ６５におけるディスクＤ２のページＰ２に対応するブロックＢ１およびＢ２（すなわち、Ｄ２Ｐ２Ｂ１およびＤ２Ｐ２Ｂ２）にポストする。したがって、ブロックＤ２Ｐ２Ｂ１は（反転ビデオ水平線の形態で示す）すべての有効データを含み、ブロックＤ２Ｐ２Ｂ２は時刻Ｔ１が示す位置までの部分的有効データを含み、ブロックＤ２Ｐ２Ｂ２の残りの部分には（クロスハッチで示す）無効データすなわち“ドント・ケア（ don't care ）"データが含まれる。

【００３４】

図４に示すように、第２の事象／時刻Ｔ２はＲＡＩＤ管理システム１６がＲＬＩ５５によるそのデータのポストを再度要求する時点を同定する。この例では、時刻Ｔ１および時刻Ｔ２の間に記憶されたログイメージページ５７のデータ（すなわち、“未書き込み"データ）をディスクステージングログ６５にポストしなければならない（これは、まだページ・フル状態になっていないためである）。（しかし、事象／時刻Ｔ２以前にログイメージページ５７がすべてトランザクションデータで満たされている場合、ＲＬＩ５５はログイメージページ５７の一部をディスクステージングログ６５にポストするよりむしろその全体をディスクログ６０にポストする。）ディスク１が最低使用頻度のディスクであり、書き込み入出力がフルブロックサイズでのみ発生すると仮定すると、ＬＩＢ２がすべてＤ１Ｐ１Ｂ２に書き込まれる。このとき、奇数ページＰ１（７５）が書き込まれ、前述したページ“トグル"（スワッピング）データ保護技術が実行される。無効データ（すなわち、そのブロックサイズ内にあって指定された時刻Ｔ２より後のデータ）をここでもクロスハッチで示す。

【００３５】

図５にはＲＬＩ５５に対してページ・フル状態になる前にデータのポストが要求される第３の事象／時刻Ｔ３を示す。この例では、時刻Ｔ２と時刻Ｔ３との間に記憶されたログイメージページ５７のデータ（“未書き込み"データ）をポストしなければならない。したがって、この例ではディスク３（Ｄ３）が最低使用頻度のディスクであると仮定すると、ＬＩＢ２のすべてがＤ２Ｐ２Ｂ２にポストされ、ＬＩＢ３のすべてがＤ３Ｐ２Ｂ３にポストされる。この場合も、ページスワッピングを行なうために、“偶数の"ページＰ２（９０）がこのとき書き込まれる。

【００３６】

図６にはＲＬＩ５５に対してページ・フル状態になる前にデータのポストが再度要求される第４の事象／時刻Ｔ４を示す。この例では、時刻Ｔ３と時刻Ｔ４との間に記憶されたログイメージページ５７の“未書き込み"データをポストしなければならない。したがって、この例ではディスク１（Ｄ１）が最低使用頻度のディスクであると仮定すると、ＬＩＢ３のすべてがＤ１Ｐ１Ｂ３に強制引出しされる。

【００３７】

図３～図６からわかるように、最低使用頻度のディスクに対して書き込みが行なわれるだけでなくディスクに対する冗長書き込みが発生しないことからシステムの入出力性能に対する全体的影響が低減される。冗長性はログデータがディスク（ディスクステージングログ６５）上に書き込まれ、しかもＲＬＩ５５にも残ることによって維持される。さらに、一実施形態において、ディスクステージングログ６５への“強制引出し"はＲＬＩ５５がページ・フル状態になる前の事象について発生し、ＲＬＩ５５のページ・フル状態が検出された場合、ＲＬＩ５５からディスクログ６０（図２）への“排出"が発生することに注意しなければならない。

【００３８】

図７には、分散書き込み用のデータ記録１１０の構成を示すブロック図である。ＲＬＩ５５（およびディスクログ６０およびディスクステージングログ６５）内のデータの各５１２バイトブロック（セクタ）はそれぞれ１つあるいはそれ以上のデータ記録１１０を有し

、データ記録１１０はブロックの境界にまたがっている場合もある。説明の目的上、データ記録１１０は簡略化した形態で示す。すなわち、記録に用いられる可能性のあるすべてのフィールドをここでは図示しない。しかし、データ記録１１０は少なくとも記録の長さを同定する長さ標識１１５、ディスクステージングログ６５からデータを回復するための記録の順序付けを同定するためのシーケンス番号１２０、トランザクションログに関係付けられたディスク群を同定するディスク群識別子１２５、記憶される実際のログデータを保持する本体１３０、およびデータ確認用のチェックサム１３５を含む。

【００３９】

シーケンス番号１２０はトランザクションログに新たな記録が追加されるたびに順次インクリメントされる生成番号である。チェックサム１３５はデータ記録１１０全体のチェックサムであり、トランザクションログの回復中に記録の状態を確認するために用いられる。ディスク群識別子１２５はＲＬＩ５５に関係付けられたディスク群の現在のインスタンス（ instance ）の任意の識別子であり、トランザクションログの回復中に“陳腐化した"（すなわち無効な）ディスクステージングログ６５データが使用されていないことを保証するために用いられる。すなわち、回復中に、記録はそのディスク群識別子１２５がディスクの現在のインスタンスに一致する場合に有効と認識され、ディスク群識別子１２５がディスク群の現在のインスタンスに一致しない場合無効と認識される。たとえば、他のディスク群からディスクドライブがスワップされた場合に陳腐化した記録あるいは有効な記録が発生する。この場合、各記録に関係付けられたディスク群識別子によって、トランザクションログ回復処理はその新たなディスク関係付けられたあらゆる陳腐化データを認識し、その使用を避けることができる。つまり、記録のディスク群識別子は現在のディスク群インスタンスに一致しなければならない。

【００４０】

図８は分散書き込みディスクログの処理フローを示すブロック図であり、本発明におけるログトランザクションの管理のために（図１の）ＲＡＩＤ管理システム１６内で実施される処理の相互関係を示す。これらの処理は、好適にはファームウエアで実行される。アプリケーション１５０（たとえば図１のＲＡＩＤ管理システム１６）がＮＶＲＡＭマップ４５（図２）を操作する際、この操作動作を同定するデータ記録１１０がログ管理１５５の制御に追加され、ＲＡＭログイメージ５５（図２）に記録される。記録はいくつかの鍵となる事象の１つが発生するまで絶えず追加される。ＲＬＩ５５中の現在のページがフルになると、ログ管理１５５がデータ管理１６０に制御を渡し、データ管理１６０がディスクドライバ１６５とインターフェースで連結してこのフルページの内容をディスクアレイ１１のディスクログ６０に冗長ポストすることによってそのフルになったページが“排出される"。ＲＬＩ５５中の現在のページがフルではないがログ管理１５５がその事象をディスクアレイ１１のディスクステージング６５への“強制引出し"ポストの要求として検出すると、ログ管理１５５はデータ管理プログラムを迂回してディスクドライバ１６５と直接インターフェースで連結してそのデータをディスクアレイに“強制引出し"する。“強制引出し"ポストの発生後に冗長コピーは書き込まれない。

【００４１】

ここで、すべての図面を参照して、なんらかのメモリ故障あるいはシステム故障のためにログの回復が必要である場合、ディスクログ６０およびディスクステージングログ６５に記憶されたインクリメンタルなログデータの回復のためにいくつかのステップが発生する。まず、ディスクログ６０のすべてのフルページがＲＬＩ５５にコピーされてログデータができるだけ多く再構築される。しかし、ページ５９のような非フルページの場合ディスクステージングログ６５にデータが残っている可能性があり、これもＲＬＩ５５にコピーしなければならない。したがって、ディスク１～Ｍ上のディスクステージングログ６５の全ページ７０～１０５からのログデータの断片をまとめて１つの完全なイメージにしてＲＬＩ５５にコピーしなければならない。

【００４２】

ディスクステージングログ６５からのこの回復に備えて、ＲＬＩ５５を走査してディスク

ログ６０に書き込まれた最後の（最も新しい）記録を示すシーケンス番号１２０を有する記録を発見する。この走査にあたってはログの循環性とシーケンス番号のラッピングの両方を考慮しなければならない。ディスクステージングログ６５内に最後に書き込まれたものに続く次のシーケンス番号（記録）がある場合これを発見しなければならない。したがって、ディスクステージングログ６５を次に走査して（ＲＬＩ５５に次に回復すべき記録を示す）次のシーケンス番号を有する記録を発見する。ディスクステージングログ６５内で次の記録（シーケンス番号）が発見されると、そのディスク群識別子１２５をチェックして、その記録がディスク群の現在のインスタンスに属するものであることを確認する。さらに、その記録のチェックサム１３５を評価してその記録の完全性を判定する。完全である場合、その記録がＲＬＩ５５にコピーされ、トランザクションログ回復処理が続行される。図６に示す例では、ブロックＤ２Ｐ２Ｂ１の最初の記録はこの第１ステップの回復基準を満足する。

【００４３】

続いて、ディスクステージングログ６５を再度走査して、次の連続シーケンス番号を有する次の記録を探す。前に発見された記録の長さ標識１１５から、次の記録が前に発見された記録の長さ標識１１５によって記述されるオフセットで始まることがわかる。したがって、そのオフセットで始まり、適当な連続シーケンス番号１２０を有し、適当なディスク群識別子１２５を有し、チェックサム１３５による有効性分析を満足する次の記録が発見されるまですべてのディスクステージングログ６５が探索される。図６の例では、ブロックＤ２Ｐ２Ｂ１内の第２の記録（視覚的には識別不能）がこの回復基準を満足する。かかる記録が発見されると、その記録がＲＬＩ５５にコピーされる。

【００４４】

一般的には、（ｉ）（次の連続シーケンス番号によって同定される）次の連続する記録を発見し、（ｉｉ）そのディスク群識別子を検証し、（ｉｉｉ）そのチェックサムを検証するこの処理全体が、かかる回復基準のそれぞれを満足するすべての記録が発見されるまでディスクステージングログ６５全体に対して継続的に反復される。さらに例を挙げれば、たとえば図６において、Ｄ２Ｐ２Ｂ１およびＤ２Ｐ２Ｂ２（反転ビデオ水平線で示される）中に同定される有効な記録のそれぞれがまずその順序でＲＬＩ５５に回復される。すると、Ｄ１Ｐ１Ｂ２あるいはＤ３Ｐ２Ｂ２内で発見されるいかなる有効な記録も次の回復ステップを満足する。たとえば、次の有効な記録がまずＤ３Ｐ２Ｂ２の前にＤ１Ｐ１Ｂ２内で発見された場合、Ｄ１Ｐ１Ｂ２内のすべての有効な記録がＲＬＩ５５に（順次一度に１つずつ）コピーされ、Ｄ１Ｐ１Ｂ２内で発見されなかった残りの有効記録は続いてＤ３Ｐ２Ｂ２内で発見され、Ｄ３Ｐ２Ｂ３の有効記録に続く。一方、次の有効な記録が（Ｄ１Ｐ１Ｂ２ではなく）Ｄ３Ｐ２Ｂ２内で発見された場合、これらの有効記録はすべてＲＬＩ５５にコピーされ、次にＤ３Ｐ２Ｂ３の有効記録がＲＬＩ５５にコピーされる。その後にのみＤ１Ｐ１Ｂ３内で発見された最終有効記録が処理され、ＲＬＩ５５にコピーされる。この最終記録はディスクステージングログ６５内の他のいかなる記録にも次の連続シーケンス番号が発見されない際に、ディスクステージングログ６５から回復されることは明らかである。

【００４５】

図示する例では、ディスクログ６０およびディスクステージングログ６５からＲＬＩ５５へのログ回復がこれで完了する。したがって、図８に示すように、ログ管理１５５が（現在はＲＬＩ５５にある）回復された記録をアプリケーション１５０（図１のＲＡＩＤ管理システム１６）の制御に返して、ログ回復が完了し、アプリケーション１５０がＮＶＲＡＭマップ４５をトランザクションログ回復処理を開始させたシステム誤り／故障の前の状態に戻すためにＲＬＩ５５に示されるログ変更の実行に着手可能であることを示す。

【００４６】

ディスクアレイ内の複数のディスクに対する分散書き込み動作を用いてディスクログ書き込み性能を向上させる方法および装置の実施形態を上に説明した。当業者には当該技術分野のさまざまなソフトウエア、ファームウエアおよび／またはハードウエアのうち任意の

ものを用いて容易に実施されることは明らかであろう。さらに、本発明をその具体的実施形態の参照により説明したが、本発明の精神と範囲から逸脱することなく他の代替実施形態および実施方法あるいは変更形態の使用が可能であることは明らかであろう。

【００４７】

以下に本発明の実施の形態を要約する。

【００４８】

１．　複数の記憶媒体（１２）を有する記憶システム（１０）への書き込み方法であって、

（ａ）前記記憶システム（１０）への書き込み要求を示す基準の検出と、

（ｂ）前記複数の記憶媒体（１２）から最低使用頻度の記憶媒体の選択と、

（ｃ）前記選択された最低使用頻度の媒体へのデータの書き込みと

を含むことを特徴とする記憶システムへのデータ書き込み方法。

【００４９】

２．　前記複数の記憶媒体（１２）はランダムアクセス記憶媒体である上記１に記載の記憶システムへのデータ書き込み方法。

【００５０】

３．　前記最低使用頻度の記憶媒体は入出力動作に基づいて選択される上記１または２に記載の記憶システムへのデータ書き込み方法。

【００５１】

４．　前記検出される基準は前記記憶媒体（１２）へのトランザクションログ（５５）の強制引出しポスト要求を含む上記１、２または３に記載の記憶システムへのデータ書き込み方法。

【００５２】

５．　前記強制引出しポスト要求は前記トランザクションログ（５５）の指定された部分（５７）がデータで満たされていることが検出される前に発生する上記４記載の記憶システムへのデータ書き込み方法。

【００５３】

６．　前記データは前記選択された最低使用頻度の記憶媒体に非冗長に書き込まれる請求項１に記載の記憶システムへのデータ書き込み方法。

【００５４】

７．　前記データは前記書き込まれるデータの順序を示す標識（１２０）を含む上記１に記載の記憶システムへのデータ書き込み方法。

【００５５】

８．　前記記憶媒体（１２）はそれぞれ前記最低使用頻度の記憶媒体が選択された際の書き込みのためにのみ使用されるよう確保された領域（６５）を含む上記１記載の記憶システムへのデータ書き込み方法。

【００５６】

９．　前記確保された領域は少なくとも２つのサブ領域（７０、７５、８０、８５、９５、１００、１０５）を含み、最低使用頻度の記憶媒体を選択する第１の事象の発生時に前記サブ領域の１つに対して書き込みが行なわれ、最低使用頻度記憶媒体を選択する次の事象の発生時に他方のサブ領域への書き込みが行なわれ、それによって同じ最低使用頻度記憶媒体が二度連続して選択される場合にも次の連続する書き込みにおいて直前に書き込まれたサブ領域に対して重ね書きが生じないことを特徴とする上記８記載の記憶システムへのデータ書き込み方法。

【００５７】

１０．　記憶システム（１０）であって、

（ａ）データ記録（１１０）を保持する第１メモリ（５５）と、

（ｂ）前記第１メモリに接続された複数の記憶媒体（１２）と、

（ｃ）前記第１メモリ（５５）の状態を検出する手段（１６）と、

（ｄ）前記第１メモリにおける第１の検出された状態に応答して第１の記憶管理基準にし

たがって前記記憶システム（１０）に前記データ記録を書き込む手段（１６）と、
（ｅ）前記第１メモリにおける第２の検出された状態に応答して第２の記憶管理基準にしたがって前記記憶システム（１０）に前記データ記録を書き込む手段（１６）を含み、
前記第２の記憶管理基準は前記複数の記憶媒体における最低使用頻度の記憶媒体への書き込みを含む記憶システム。

【００５８】

１１．　前記第１メモリ（５５）における前記第１の検出された状態は、前記第１メモリの一部が前記データ記録で満たされていることを示す状態を含む上記１０記載の記憶システム。

【００５９】

１２．　前記第１の記憶管理基準は前記複数の記憶媒体（１２）上の前記データ記録の冗長性の維持を含む上記１０記載の記憶システム。

【００６０】

１３．　前記第１メモリにおける前記第２の検出された状態は、前記第１メモリにおいて所定の部分が前記データ記録で満たされる前に前記第１メモリ（５５）内の前記データ記録を前記複数の記憶媒体（１２）に強制引出し的に書き込む要求を示す状態を含む上記１０記載の記憶システム。

【００６１】

１４．　前記第２の記憶管理基準は前記複数の記憶媒体（１２）上において前記データ記録の冗長性を維持しないことを含む上記１０記載の記憶システム。

【００６２】

１５．　前記最低使用頻度の記憶媒体は前記データ記録の書き込みのために確保された少なくとも２つのサブ領域（７０、７５、８０、８５、９０、９５、１００、１０５）を含む上記１０記載の記憶システム。

【００６３】

【発明の効果】

本発明によれば、複数の利用可能な記憶ディスクから選択された任意の最も使用頻度の低いディスクへのログ書き込みを管理および分散してログ入出力と進行中の他の入出力との間におけるディスクアクセスの競合を低減することによって、ディスクログ書き込みのシステム性能を改善でき、且つ、システムの誤り又は故障時に記録を確実に回復することができる。

【００６４】

【図面の簡単な説明】

【図１】本発明の分散書き込みディスクログ法を用いたデータ記憶システムのブロック図である。

【図２】本発明の分散ログ書き込みディスクログ法を示すブロック図である。

【図３】ある時間における本発明のディスクステージングログの状態を示すブロック図である。

【図４】図３の状態以後の、他の時間における本発明のディスクステージングログの状態を示すブロック図である。

【図５】図４の状態以後の、他の時間における本発明のディスクステージングログの状態を示すブロック図である。

【図６】図５の状態以後の、他の時間における本発明のディスクステージングログの状態を示すブロック図である。

【図７】分散書き込み用のデータ記録の構成を示すブロック図である。

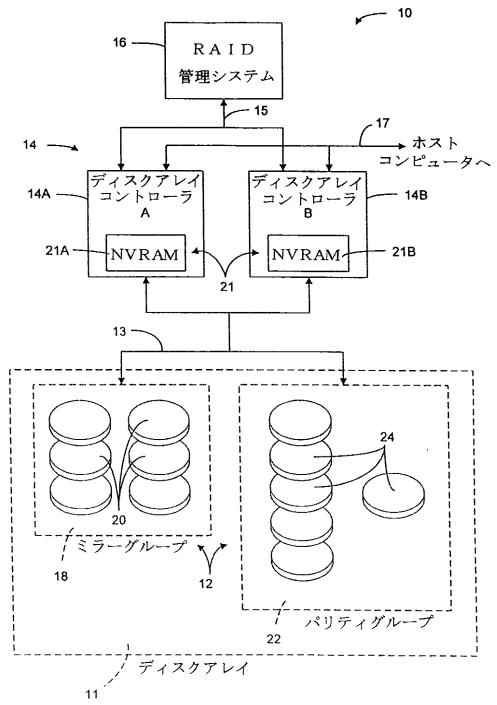
【図８】分散書き込みディスクログの処理フローを示すブロック図である。

【符号の説明】

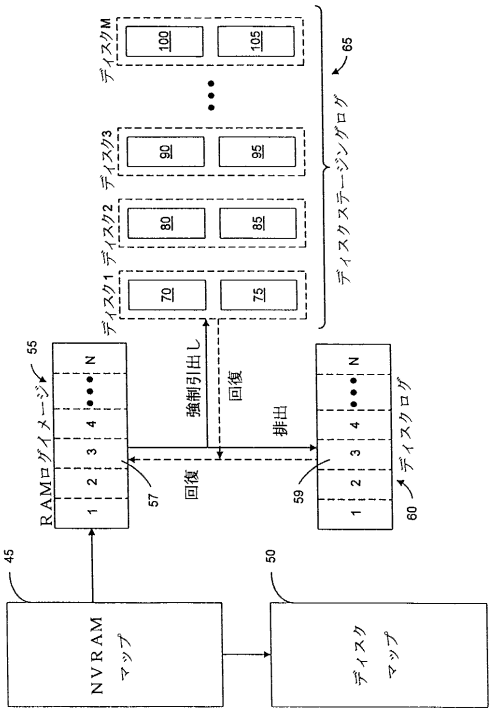
１，２，３，Ｍ　　ディスク
１０　　データ記憶システム
１１　　ディスクアレイ

１２　　記憶ディスク
１４　　ディスクアレイコントローラ
１４Ａ　ディスクアレイコントローラＡ
１４Ｂ　ディスクアレイコントローラＢ
１６　　ＲＡＩＤ管理システム
１８　　ミラーグループ
２１　　メモリマップ記憶域
２１Ａ，２１Ｂ　　ＮＶＲＡＭ
２２　　パリティグループ
４５　　ＮＶＲＡＭマップ
５０　　ディスクマップ
５５　　ＲＡＭログイメージ（ＲＬＩ）
６０　　ディスクログ
６５　　ディスクステージングログ
１１０　　データ記録
１１５　　長さ標識
１２０　　シーケンス番号
１２５　　ディスク群識別子
１３０　　本体
１３５　　チェックサム
１５０　　アプリケーション
１５５　　ログ管理
１６０　　データ管理
１６５　　ディスクドライバ

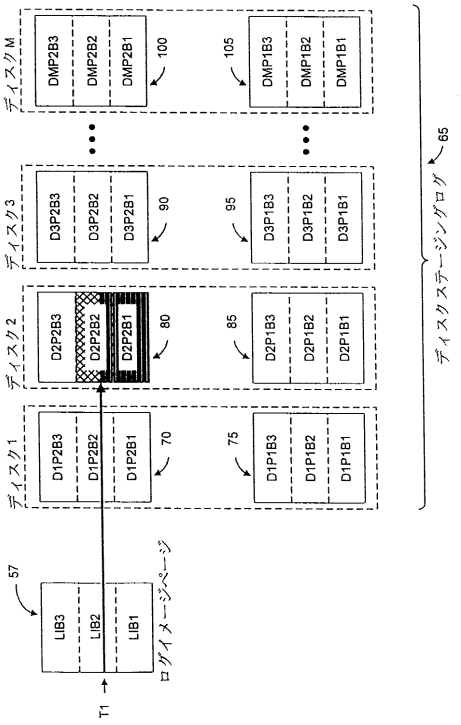
【図１】



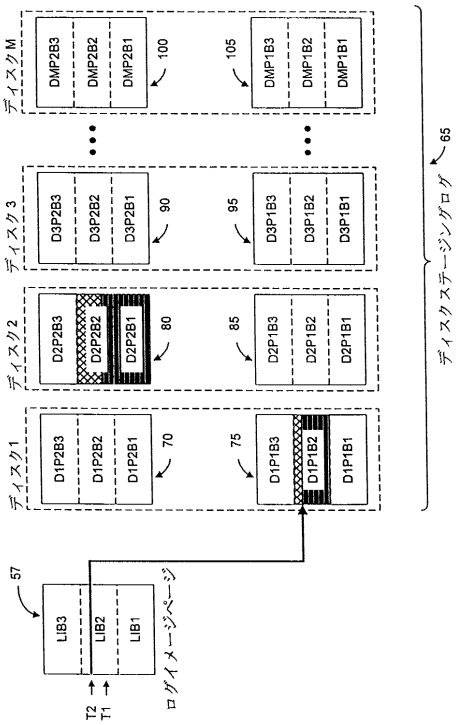
【図２】



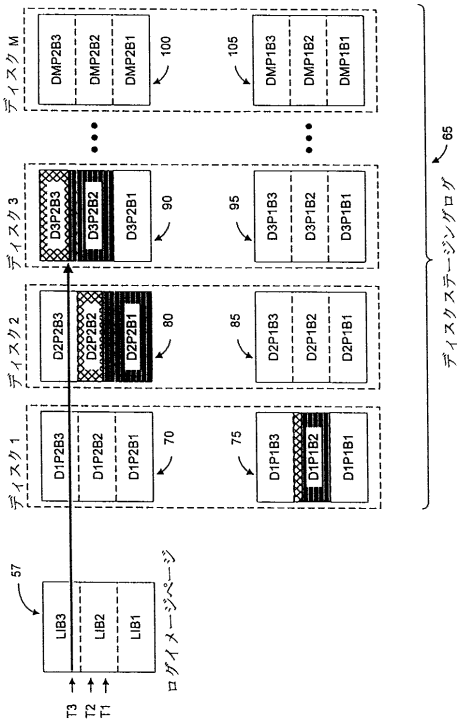
【図３】


ディスクM
ディスク3
ディスク2
ディスク1
DMP2B3
DMP2B2
DMP2B1
DMP1B3
DMP1B2
DMP1B1
D3P2B3
D3P2B2
D3P2B1
D3P1B3
D3P1B2
D3P1B1
D2P2B3
D2P2B2
D2P2B1
D2P1B3
D2P1B2
D2P1B1
D1P2B3
D1P2B2
D1P2B1
D1P1B3
D1P1B2
D1P1B1
100
105
90
95
80
85
70
75
65
ディスクステージングログ
57
LIB3
LIB2
LIB1
ログイメージページ
T1

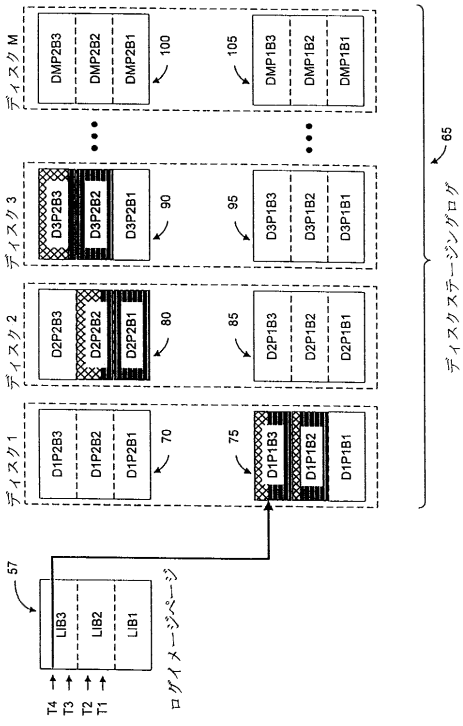
【図４】


ディスクM
ディスク3
ディスク2
ディスク1
DMP2B3
DMP2B2
DMP2B1
DMP1B3
DMP1B2
DMP1B1
D3P2B3
D3P2B2
D3P2B1
D3P1B3
D3P1B2
D3P1B1
D2P2B3
D2P2B2
D2P2B1
D2P1B3
D2P1B2
D2P1B1
D1P2B3
D1P2B2
D1P2B1
D1P1B3
D1P1B2
D1P1B1
100
105
90
95
80
85
70
75
65
ディスクステージングログ
57
LIB3
LIB2
LIB1
ログイメージページ
T2
T1

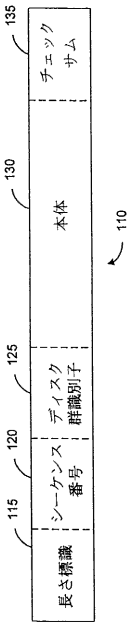
【図５】


ディスクM
ディスク3
ディスク2
ディスク1
DMP2B3
DMP2B2
DMP2B1
DMP1B3
DMP1B2
DMP1B1
D3P2B3
D3P2B2
D3P2B1
D3P1B3
D3P1B2
D3P1B1
D2P2B3
D2P2B2
D2P2B1
D2P1B3
D2P1B2
D2P1B1
D1P2B3
D1P2B2
D1P2B1
D1P1B3
D1P1B2
D1P1B1
100
105
90
95
80
85
70
75
65
ディスクステージングログ
57
LIB3
LIB2
LIB1
ログイメージページ
T3
T2
T1

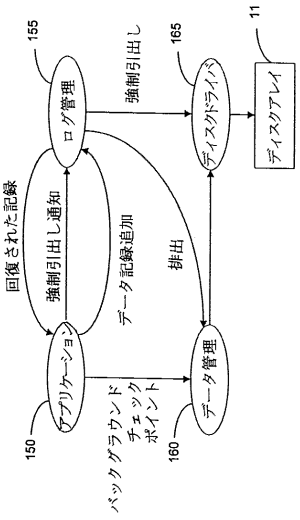
【図６】


ディスクM
ディスク3
ディスク2
ディスク1
DMP2B3
DMP2B2
DMP2B1
DMP1B3
DMP1B2
DMP1B1
D3P2B3
D3P2B2
D3P2B1
D3P1B3
D3P1B2
D3P1B1
D2P2B3
D2P2B2
D2P2B1
D2P1B3
D2P1B2
D2P1B1
D1P2B3
D1P2B2
D1P2B1
D1P1B3
D1P1B2
D1P1B1
100
105
90
95
80
85
70
75
65
ディスクステージングログ
57
LIB3
LIB2
LIB1
ログイメージページ
T4
T3
T2
T1

【図７】


115
長さ標識
120
シーケンス番号
125
ディスク群識別子
130
本体
135
チェックサム
110

【図８】


155
ログ管理
強制引出し
165
ディスクドライバ
11
ディスクアレイ
回復された記録
強制引出し通知
データ記録追加
排出
150
アプリケーション
バックグラウンドチェックポイント
160
データ管理

フロントページの続き

(72)発明者　ドン・エル・ブルケス
アメリカ合衆国　アイダホ，メリディアン，サン・ラモ・ディーアール　３１００
(72)発明者　キルク・エー・ハンソン
アメリカ合衆国　アイダホ，イーグル，ウエスト・ニューフィールド・ディーアール　１１２９

審査官　▲吉▼田　美彦

(56)参考文献　実開平０４－０７３２５１（ＪＰ，Ｕ）
特開平０９－０５４６５８（ＪＰ，Ａ）
特開平０５－０８１０９９（ＪＰ，Ａ）
特開平０６－３２４８１７（ＪＰ，Ａ）
特開平０６－２３１０１２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F　3/06