(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5594942号
(P5594942)
(45)発行日　平成26年9月24日(2014.9.24)　(24)登録日　平成26年8月15日(2014.8.15)

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

請求項の数 14 (全 15 頁)

(21)出願番号　特願2008-169143(P2008-169143)
(22)出願日　平成20年6月27日(2008.6.27)
(65)公開番号　特開2009-15845(P2009-15845A)
(43)公開日　平成21年1月22日(2009.1.22)
審査請求日　平成23年6月16日(2011.6.16)
(31)優先権主張番号　11/771,980
(32)優先日　平成19年6月29日(2007.6.29)
(33)優先権主張国　米国(US)

前置審査

(73)特許権者　500373758
シーゲイト　テクノロジー　エルエルシー
アメリカ合衆国、９５０１４　カリフォルニア州、クパチーノ、サウス・デ・アンザ・ブールバード、１０２００
(74)代理人　100064746
弁理士　深見　久郎
(74)代理人　100085132
弁理士　森田　俊雄
(74)代理人　100083703
弁理士　仲村　義平
(74)代理人　100096781
弁理士　堀井　豊
(74)代理人　100111246
弁理士　荒川　伸夫

最終頁に続く

(54)【発明の名称】好適なゾーン・スケジューリング

(57)【特許請求の範囲】
【請求項１】
データ・ストレージ・システムであって、
ポリシーエンジンを備え、
前記ポリシー・エンジンは、前記データ・ストレージ・システムに対するネットワーク負荷についての定性データを連続的に収集して前記負荷を動的に特徴付け、該特徴づけにより得られた負荷の特徴に応じて、ライトバック・データに対する転送要求についてのコマンド・キューの内容を記憶媒体上にデータ記憶領域として設けられた複数の物理ゾーンにおいて物理ゾーン単位で選択されたライトバック・データに対する転送要求だけに選択的に限定し、
前記ネットワーク負荷は、該ネットワークを介して前記データ・ストレージ・システムに該ネットワークに結合されるホストから与えられる一連のアクセス・コマンドの少なくとも種別および転送速度により表され、
前記定性データは前記負荷に含まれるアクセス・コマンドが読取りコマンドであるか書込みコマンドであるかを特定するデータであり、前記負荷の特徴は少なくとも当該特定されたアクセス・コマンドの分布を表す、データ・ストレージ・システム。
【請求項２】
前記ポリシー・エンジンに応答して、前記負荷の特徴が、前記負荷が、前記ポリシーエンジンが前記ネットワーク負荷の状態の判定のための基準として用いる予想された定常状態と同じか、またはそれ以下であることを示している場合に、前記複数の物理ゾーンが均

一な割合で選択されてアクセスされる均一な分布にもとづいて前記コマンド・キューの内容を設定して前記複数の物理ゾーンが均一な割合でアクセスされるよう管理し、前記負荷の特徴が、均一でない分布のアクセス・コマンド活動のバーストを示している場合に、物理ゾーンにもとづいて前記コマンド・キューの内容を設定して前記物理ゾーン単位で前記複数の物理ゾーンがアクセスされるように管理するキャッシュ・マネージャを備える、請求項１に記載のデータ・ストレージ・システム。

【請求項３】

前記負荷の特徴が、前記負荷に含まれるアクセス・コマンドのレイテンシに敏感なアクセス・コマンドに対する速度に敏感なアクセス・コマンドの比率で表され、

前記レイテンシに敏感なコマンドは、コマンド処理においてレイテンシの影響が速度による影響よりも大きなコマンドであり、

前記速度に敏感なコマンドはコマンド処理において速度の影響がレイテンシによる影響よりも大きなコマンドである、請求項１または２に記載のデータ・ストレージ・システム。

【請求項４】

前記速度に敏感なコマンドが、ライトバック・キャッシュ・コマンドであり、前記レイテンシに敏感なコマンドが、読取りコマンドおよびライトスルー・キャッシュ・コマンドのうちの少なくとも一方である、請求項３に記載のデータ・ストレージ・システム。

【請求項５】

前記負荷の特徴が、各アクセス・コマンドに関連するサイズで表される、請求項１に記載のデータ・ストレージ・システム。

【請求項６】

各物理ゾーンが、前記記憶媒体の全周にわたる環状部分である、請求項１に記載のデータ・ストレージ・システム。

【請求項７】

独立のドライブ（ＲＡＩＤ）の冗長アレイを定義するために割り当てられた媒体を有するデータ・ストレージのアレイを備え、各物理ゾーンが、２つ以上のデータ・ストレージにわたってストライプ化されて割り当てられたストレージ空間である、請求項１から６のいずれかに記載のデータ・ストレージ・システム。

【請求項８】

前記コマンド・キューがシーク・マネージャに提示され、前記シーク・マネージャは、コマンド・プロファイルを定義するコマンド・キュー内の転送要求を選択的に発行する、請求項１から７のいずれかに記載のデータ・ストレージ・システム。

【請求項９】

前記ポリシー・エンジンが、発行時のコマンドの分布を示すコマンド・プロファイルを決定する際に、シーク・マネージャを管理する規則として負荷の特徴に対応する規則および前記コマンド・キューからの転送要求の処理時に前記システムにおいて実現されるべき目標に対応する規則を決定する、請求項１から８のいずれかに記載のデータ・ストレージ・システム。

【請求項１０】

前記ポリシー・エンジンが、前記負荷に含まれるアクセス・コマンドのレイテンシに敏感なコマンドに対する速度に敏感なコマンドの比率でにおいて、該発行時のコマンドの分布を示すコマンド・プロファイルを前記負荷の特徴に選択的に対応させ、

前記レイテンシに敏感なコマンドは、コマンド処理においてレイテンシの影響が速度による影響よりも大きなコマンドであり、

前記速度に敏感なコマンドはコマンド処理において速度の影響がレイテンシによる影響よりも大きなコマンドである、請求項１から９のいずれかに記載のデータ・ストレージ・システム。

【請求項１１】

前記ポリシー・エンジンが、前記負荷に含まれるアクセス・コマンドのレイテンシに敏

感なコマンドに対する速度に敏感なコマンドの比率において、前記コマンド・プロファイルを前記負荷の特徴とマッチさせ、
前記レイテンシに敏感なコマンドは、コマンド処理においてレイテンシの影響が速度による影響よりも大きなコマンドであり、
前記速度に敏感なコマンドはコマンド処理において速度の影響がレイテンシによる影響よりも大きなコマンドである、請求項９に記載のデータ・ストレージ・システム。
【請求項１２】
前記ポリシー・エンジンが、前記コマンド・プロファイルを前記アクセス・コマンドの所望の最大レイテンシと選択的に対応させる、請求項９に記載のデータ・ストレージ・システム。
【請求項１３】
前記ポリシー・エンジンが、前記コマンド・プロファイルを異なるＬＵＮに割り当てられた優先度に選択的に対応させる、請求項９に記載のデータ・ストレージ・システム。
【請求項１４】
前記ポリシー・エンジンが、有限状態機械である、請求項１に記載のデータ・ストレージ・システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は、概して、データ・ストレージ・システムの分野に関し、特に、分散アレイ・ストレージ・システムにおけるシーク・コマンド・プロファイルを適合可能に管理するための装置および方法に関するが、これに限定されない。
【背景技術】
【０００２】
コンピュータ・ネットワーキングは、工業規格アーキテクチャのデータ転送速度が、インテル社（Ｉｎｔｅｌ　Ｃｏｒｐｏｒａｔｉｏｎ）の８０３８６プロセッサのデータ・アクセス速度に追いつくことができなくなった時に急激に増大し始めた。ローカル・エリア・ネットワーク（ＬＡＮ）は、ネットワーク内のデータ・ストレージ容量を強化することにより、ストレージ・エリア・ネットワーク（ＳＡＮ）に進化した。ユーザは、装置を結合しＳＡＮ内の装置で扱われる関連データにより、直接取付ストレージにより可能となるより１桁上の処理能力、そして扱いやすいコストで有意の利点を実現している。
【０００３】
さらに最近は、データ・ストレージ・サブシステムを制御するためのネットワーク・セントリック・アプローチの方向への動きがある。すなわち、ストレージを強化したのと同じ方法で、サーバから取り出され、ネットワーク自身に送られるストレージの機能を制御するシステムにも同じ動きがある。例えば、ホスト・ベースのソフトウェアは、インテリジェント・スイッチまたは特殊化したネットワーク・ストレージ・サービス・プラットフォームに保守および管理タスクを委託することができる。アプライアンス・ベースの解決方法を使用すれば、ホストで稼働するソフトウェアが必要なくなるし、企業内にノードとして設置されているコンピュータで動作することができる。いずれにせよ、インテリジェント・ネットワーク解決方法は、これらのものをストレージ割当ルーチン、バックアップ・ルーチン、およびホストによらない障害許容スキームとして中央に集めることができる。
【発明の開示】
【発明が解決しようとする課題】
【０００４】
インテリジェンスをホストからネットワークに移動すればこのようないくつかの問題を解決することはできるが、仮想ストレージのプレゼンテーションをホストに変更する際の柔軟性の一般的な不足に関連する固有の問題は解決しない。例えば、データを格納する方法を、通常でないホスト負荷活動のバーストを収容するように適合させる必要がある場合

がある。その各データ・ストレージ容量の自己決定による割当て、管理、および保護、およびグローバルなストレージ要件に適応するように、ネットワークへその容量を仮想ストレージ空間として提示するインテリジェント・データ・ストレージ・サブシステムが求められている。この仮想ストレージ空間は、複数のストレージ・ボリューム内に提供することができる。本発明の目指しているのはこのための解法である。

【課題を解決するための手段】

【０００５】

本発明の実施形態は、概して、バースト負荷条件下でコマンド・キューの内容を選択的に入手することによるダーティ・データのフラッシング性能の最適化に関する。

【０００６】

ある実施形態においては、ポリシー・エンジンが、負荷を動的に特徴付けるために、データ・ストレージ・システムへのネットワーク負荷に関する定性情報を連続的に収集し、キャッシュされたライトバック・データおよびホスト読取りコマンドに対するデータ転送要求から入手するコマンド・キューの内容に負荷の特徴を連続的に相関付け、内容を記憶媒体の複数の予め定義したゾーンのゾーン・ベースで選択されるライトバック・データに対するこれらの転送要求だけに制限するデータ・ストレージ・システムおよび関連する方法が提供される。

【０００７】

特許請求の範囲に記載する本発明を特徴付けるこれらおよび種々の他の機能および利点は、下記の詳細な説明を読み、関連する図面を見れば理解することができるだろう。

【発明を実施するための最良の形態】

【０００８】

図１は、本発明の実施形態を含む例示としてのコンピュータ・システム１００である。１つまたは複数のホスト１０２は、ローカル・エリア・ネットワーク（ＬＡＮ）および／またはワイド・エリア・ネットワーク（ＷＡＮ）１０６により、１つまたは複数のネットワークに取り付けられているサーバ１０４にネットワークで接続している。好適には、ＬＡＮ／ＷＡＮ１０６は、ワールド・ワイド・ウェブを通して通信するために、インターネット・プロトコル・ネットワーキング・インフラストラクチャを使用することが好ましい。ホスト１０２は、多数のインテリジェント記憶素子（ＩＳＥ）１０８のうちの１つまたは複数上に格納しているデータをルーチン的に必要とするサーバ１０４内に常駐しているアプリケーションにアクセスする。それ故、ＳＡＮ１１０は、格納しているデータにアクセスするために、サーバ１０４をＩＳＥ１０８に接続する。ＩＳＥ１０８は、その内部の企業またはデスクトップ・クラスの記憶媒体により、直列ＡＴＡおよびファイバ・チャネルのような種々の選択した通信プロトコルによりデータを格納するために、データ・ストレージ容量１０９を提供する。

【０００９】

図２は、図１のコンピュータ・システム１００の一部の簡単な図面である。３つのホスト・バス・アダプタ（ＨＢＡ）１０３は、ネットワークまたはファブリック１１０を介して１対のＩＳＥ１０８（それぞれＡおよびＢで示す）と相互に作用する。各ＩＳＥ１０８は、好適には、独立ドライブの冗長アレイ（ＲＡＩＤ）として特徴付けられている一組のストレージとして、データ・ストレージ容量１０９上で動作することが好ましい二重化冗長制御装置１１２（Ａ１、Ａ２およびＢ１、Ｂ２で示す）を含む。すなわち、好適には、制御装置１１２およびデータ・ストレージ容量１０９は、種々の制御装置１１２が並列の冗長リンクを使用し、システム１００が格納しているユーザ・データのうちの少なくともいくつかが、少なくとも一組のデータ・ストレージ容量１０９内の冗長フォーマットに格納されるように、障害許容配置を使用することが好ましい。

【００１０】

図３は、本発明の例示としての実施形態により組み立てたＩＳＥ１０８である。シェルフ１１４は、ミッドプレーン１１６と電気的に接続している制御装置１１２に収容する形で係合するための空洞を定める。シェルフ１１４は、キャビネット（図示せず）内に支持

される。１対の複数ドライブ・アセンブリ（ＭＤＡ）１１８は、ミッドプレーン１１６の同じ側面上のシェルフ１１４内に収容される形で係合している。ミッドプレーン１１６の対向側面には、非常電力供給を行うデュアル・バッテリー１２２、デュアル交流電源１２４およびデュアル・インタフェース・モジュール１２６が接続している。好適には、デュアル構成要素は、一方のあるいは両方のＭＤＡ１１８を同時に動作し、それにより構成要素が故障した場合にバックアップ保護を行うように構成することが好ましい。

【００１１】

図４は、それぞれが５つのデータ・ストレージ１２８を支持している上部隔壁１３０および下部隔壁１３２を有するＭＤＡ１１８の拡大分解等角図である。隔壁１３０、１３２は、ミッドプレーン１１６（図３）と係合するコネクタ１３６を有する共通の回路基板１３４と接続するためにデータ・ストレージ１２８を整合する。ラッパー１３８は、電磁妨害シールドを行う。ＭＤＡ１１８のこの例示としての実施形態は、参照により本明細書に組み込むものとする譲受人に譲渡される「複数のディスク・アレイのためのキャリヤ装置および方法（Ｃａｒｒｉｅｒ　Ｄｅｖｉｃｅ　ａｎｄ　Ｍｅｔｈｏｄ　ｆｏｒ　ａ　Ｍｕｌｔｉｐｌｅ　Ｄｉｓｃ　Ａｒｒａｙ）」という名称の米国特許第７，１３３，２９１号の主題である。ＭＤＡ１１８のもう１つの例示としての実施形態は、本発明の譲受人に譲渡される、参照により本明細書に組み込むものとする同じ名称の米国特許第７，１７７，１４５号の主題である。他の等価の実施形態の場合には、ＭＤＡ１１８は、密封されたエンクロージャ内に設置することができる。

【００１２】

図５は、本発明の実施形態と一緒に使用するのに適していて、回転媒体ディスク・ドライブの形をしているデータ・ストレージ１２８の等角図である。動体データ記憶媒体と回転スピンドルを下記の説明のために使用するが、他の等価の実施形態の場合には、固体メモリ素子のような非回転媒体デバイスが使用される。図５の例示としての実施形態の場合には、データ記憶ディスク１３８は、読取り／書込みヘッド（「ヘッド」）１４２にディスク１３８のデータ記憶位置を示すためにモータ１４０により回転する。ヘッド１４２は、ディスク１３８の内部トラックと外部トラックとの間をヘッド１４２が半径方向に移動している間に、ボイス・コイル・モータ（ＶＣＭ）１４６に応じる回転アクチュエータ１４４の遠い方の端部のところに支持されている。ヘッド１４２は、フレックス回路１５０を通して回路基板１４８に電気的に接続している。回路基板１４８は、データ・ストレージ１２８の機能を制御する制御信号を受信し、送信することができる。コネクタ１５２は、回路基板１４８に電気的に接続していて、データ・ストレージ１２８をＭＤＡ１１８の回路基板１３４（図４）と接続することができる。

【００１３】

図６は、制御装置１１２のうちの１つの図面である。制御装置１１２は、１つの集積回路で具体化することもできるし、必要に応じて多数の個々の回路間で分散することもできる。好適には、プログラマブル・コンピュータ・プロセッサであることを特徴とするプロセッサ１５４は、プログラミング・ステップ、および好適には不揮発性メモリ１５６（フラッシュ・メモリまたは類似物など）およびダイナミック・ランダム・アクセス・メモリ（ＤＲＡＭ）１５８内に格納している処理データにより制御を行う。

【００１４】

ファブリック・インタフェース（Ｉ／Ｆ）回路１６０は、ファブリック１１０を介して他の制御装置１１２およびＨＢＡ１０３と通信し、デバイスＩ／Ｆ回路１６２は、ストレージ１２８と通信する。Ｉ／Ｆ回路１６０、１６２および経路制御装置１６４は、キャッシュ１６６を使用するなどして、ＨＢＡ１０３を介してネットワーク・デバイスとＩＳＥ１０８との間でコマンドおよびデータを送るために通信経路を形成する。別々に図示してあるが、経路制御装置１６４およびＩ／Ｆ回路１６０、１６２は一体に形成することができることを理解することができるだろう。

【００１５】

好適には、ホスト処理機能を増大するために、ストレージ１２８への仮想ブロックをフ

ラッシュするようにＲＡＩＤコンテナ・サービス（ＲＣＳ）に要求することにより、キャッシュ・マネージャが、書込みコマンドの特定のサブセットに対してフラッシング活動を作動させるまで、仮想ブロックに対する書込みコマンドはキャッシュ１６６内にライトバック・キャッシュされ、その内部に懸案として保持される。確実に媒体を更新するＲＡＩＤアルゴリズムにより媒体の更新を行う目的で、ＲＣＳは、シーク・マネージャに特定のデータ転送を行うために要求を送るアルゴリズムを実行する。シーク・マネージャは、キャッシュされたライトバック・コマンド、およびもっと優先度の高いホスト読取りコマンドからのデータ転送要求を発行する許可を実際に与えるために、特定のストレージ１２８に対するコマンド・キューを管理する。シーク・マネージャは、実際に転送要求を発行する許可を与える関連するデータ転送を行うためのリソースを割り当てる。

【００１６】

ＩＳＥ１０８のデータ・ストレージ容量は、データをストレージ１２８に格納する場合に、およびデータをストレージ１２８から検索する場合に、参照される論理装置の形に組織される。システム構成情報は、ユーザ・データおよび関連するパリティと、ミラー・データおよび各記憶位置間の関係を定義する。システム構成情報は、さらに、論理ブロック・アドレス（ＬＢＡ）の用語のようなもので、データに割り当てられたストレージ容量のブロックと関連するメモリ記憶位置との間の関係を識別する。システム構成情報は、さらに、論理ブロック・アドレスにマッピングされる仮想ブロック・アドレスを定義することによる仮想化を含むことができる。

【００１７】

制御装置１１２アーキテクチャは、有利にスケーリングすることができる非常に機能的なデータ管理を行い、ストレージ容量の制御を行う。好適には、ストライプ・バッファ・リスト（ＳＢＬ）および他のメタデータ構造を、記憶媒体上のストライプ境界、および記憶処理中ディスク・ストライプと関連するデータを格納するための専用のキャッシュ１６６内の参照データ・バッファと整合することが好ましい。

【００１８】

動作中、キャッシュ１６６は、ＳＡＮ１１０によりＨＢＡ１０３を通して、ユーザ・データおよびＩ／Ｏ転送に関連する他の情報を格納する。要求されなかった不確かなデータを含むストレージ１２８から検索したリードバック・データを、ストレージ１２８宛のアクセス・コマンドのスケジューリングを要求する代わりに、以降の要求したデータがキャッシュ１６６から直接転送されるように、以降の「キャッシュ・ヒット」をあてにして、キャッシュ１６６内に暫くの間保持することができる。同様に、ストレージ１２８に書き込むデータが、キャッシュされるようにライトバック・キャッシュ・ポリシーが使用され、完了肯定応答がＨＢＡ１０３を介して開始ネットワーク・デバイスに返送されるが、ストレージ１２８へのデータの実際の書込みは、後の都合のよい時間にスケジューリングされる。

【００１９】

それ故、通常、制御装置１１２は、各エントリの状態を含むキャッシュ１６６の内容の正確な制御を維持しなければならない。このような制御は、テーブル構造に関連するアドレスを使用するスキップ・リスト配置により実行することが好ましい。スキップ・リストは、キャッシュ１６６の一部内に維持することが好ましいが、必要に応じて他のメモリ・スペースを使用することもできる。

【００２０】

図７は、経路制御装置１６４（図６）内に常駐するキャッシュ・マネージャ１７０、ＲＡＩＤコンテナ・サービス１７２、ポリシー・エンジン１７４、およびシーク・マネージャ１７６を示す機能ブロック図である。シーク・マネージャ１７６は１つしか図示してないが、ストレージ１２８に対して専用のシーク・マネージャ１７６が存在する。そのため、これらのシーク・マネージャは、ポリシー・エンジン１７４からのシーク規則に個々に応答する。

【００２１】

これらの機能ブロックは、ソフトウェアまたはハードウェアで実施することができる。ハードウェアで実施する場合には、ポリシー・エンジン１７４は有限状態機械であるが、これに限定されない。いずれにせよ、ポリシー・エンジン１７４は、経路１７８を介して、Ｉ／Ｏ単位ベースでファブリックＩ／Ｆ１６０経由で受信したアクセス・コマンドについての定性データを連続的に収集する。ポリシー・エンジン１７４は、動的にネットワーク負荷を特徴付け、それに続けてシーク・マネージャ１７６を管理する経路１７９を介してシーク規則をその後で発行する。シーク・マネージャは、経路１８０を介してライトバック・データおよびホスト読取り要求を転送するために、データ転送要求のコマンド・キューに問い合わせ、コマンド・プロファイルを定義するために経路１８２を通してデータ転送要求を発行する許可を選択的に与える。

【００２２】

ポリシー・エンジン１７４は、シーク・マネージャ１７６に対する規則を作成する際に性能の目標１８８に応じることができる。目標１８８は、速度に敏感なコマンドに対するレイテンシに敏感なコマンドの比率（ライトバック・キャッシングに対する書込みコマンドに対する読取りコマンドの比率）でのネットワーク負荷のある要因である所望のコマンド・プロファイルの強化、異なるＬＵＮクラスに割り当てた優先度の強化、所望の読取りコマンドのレイテンシの強化のような量的なものであっても質的なものであってもよいが、これらに限定されない。ポリシー・エンジン１７４は、また、サイズ（帯域幅）のような他のものでホスト負荷を特徴付ける定性データを収集することもできる。

【００２３】

さらに、ポリシー・エンジン１７４は、シーク・マネージャ１７６を管理する規則を作成する際にシステム状態情報１９０に応じることができる。例えば、制限なしで、電源インジケータは、ポリシー・エンジン１７４にＩＳＥ１０８がバッテリーのバックアップ電源に切り替わったことを知らせることができる。この状態の場合、ポリシー・エンジン１７４は、計画した制限付きの電力利用度に関してキャッシュ１６６を積極的にフラッシングするために付随の事態を実施する可能性がある。ポリシー・エンジン１７４は、コマンド・プロファイルをストレージ１２８に調整する場合に、シーク・マネージャ１７６を管理するシーク規則を作成する際に、アクセス・コマンド・データ転送に直接関連しない懸案中のバックグラウンドＩ／Ｏ１９２またはＩ／Ｏの状態に応じることができる。

【００２４】

それ故、ポリシー・エンジン１７４は、キャッシュされたライトバック・コマンドおよびより優先度の高いホスト読取りコマンドから入手したデータ転送のコマンド・キュー内の複数のデータ転送から選択したデータ転送要求を発行する目的で、シーク・マネージャ１７６を管理するシーク規則を定義するために、負荷の特徴、目標１８８、システム状態１９０、およびバックグラウンドＩ／Ｏの任意の組合わせを使用することができる。

【００２５】

例示としての例の場合には、ポリシー・エンジンは、レイテンシに敏感なコマンドに対する速度に敏感なコマンドの比率で、ネットワーク負荷を特徴付ける。この説明のために、ライトバック・キャッシング・スキームを仮定する。それ故、ライトバック・キャッシュ・コマンドは、速度に敏感なコマンドであると見なされる。何故なら、任意の時点でデータ・ストレージ１２８にどの要求をフラッシングするかは大した問題ではないからである。実際には、ダーティ・データとしてキャッシュ１６６内で未決状態である場合に、速度に敏感な要求を上書きすることすらできるからである。問題は、速度に敏感なコマンドを、キャッシュ１６６が飽和状態になることを防止する速度でフラッシングすることである。

【００２６】

一方、１つまたは複数のデータ・ストレージ１２８内に格納しているデータを読み出すためのアクセス・コマンドは、同様に、ネットワーク・アプリケーションが、アクセス・コマンドが満足するまでそれ以上の処理を阻止する恐れがある。アクセス・コマンドを満足させる時間、すなわち、レイテンシ期間は、アプリケーションの性能にとって非常に重

要なものである。そのため、このようなコマンドは、レイテンシに敏感なコマンドと呼ばれる。ある状況の場合には、ホストは、ライトバック・キャッシングを許可しないことを選択することができる。この場合、ライトスルー・キャッシュ・コマンドと呼ばれる書込みコマンドは、同様にレイテンシに敏感なコマンドとして分類される。

【００２７】

定性データを収集する際に、ポリシー・エンジン１７４は、１秒の各間隔のような、しかしこれに限定されない所定の各サンプリング期間中にカウントを照合することが好ましい。書込みコマンドに対する読取りコマンドの比率でデータを収集するために、例えば、フリーランニング・カウンタは、連続的に上記比率を追跡するために１秒刻みで指針を移動させるポインタにより設定することができる。カウンタは、９番目のスロットで現在の１秒の比を計算するため前の８回の１秒サンプル比のような所望の数の前に観察した比率を保持する。１秒刻みの目盛の上では、指針が回転し、指し示した履歴値を減算し、最近のサンプル値を加算し、次に、比率の最近の移動平均を計算するために８で除算を行う。

【００２８】

さらに、例示としての例について説明すると、ポリシー・エンジン１７４が、ホスト負荷バースト活動が発生していることを観察している場合には、ポリシー・エンジンは、キャッシュ１６６内での飽和状態を防止する際に、ネットワーク負荷に関連してコマンド・プロファイルを修正するために、シーク・マネージャ１７６を管理する規則を発行することができる。この規則は、（レイテンシに敏感な）読取りコマンドに対する（速度に敏感な）書込みコマンドの比率でコマンド・プロファイルをネットワーク負荷にマッチさせることができる。この規則は、また飽和状態からできるだけ速く滑らかに回復するために、最大レイテンシおよびＬＵＮクラス優先度のような他の目標を修正することもできるし、または一時的に延期することさえできる。

【００２９】

このような状態の場合、シーク・マネージャ１７６に書込みコマンドを積極的にフラッシングすることが強く求められた場合には、特にストレージ１２８内のディスク１３８（図５）の複数の物理ゾーンの選択した物理ゾーン内に位置するものだけへのコマンド・キュー内の転送要求を入手するために、負荷特徴に関連するキャッシュ・マネージャ１７０の機能を変更することにより、フラッシング性能を有意に改善するという決定が行われる。

【００３０】

図８は、ストライプ・データ記述子（ＳＤＤ）とも呼ばれるデータ構造を使用するノード・ベースでのキャッシュ１６６（図６）を管理するキャッシュ・マネージャ１７０の略図である。各ＳＤＤは、それが関連するデータへの最近および現在のアクセスに関連するデータを保持する。各ＳＤＤは、アクセス履歴、ロックした状態、最後のオフセット、最後のブロック、タイムスタンプ・データ（時刻、ＴＯＤ）、データがどのゾーン（ブック）に属するのかを示す識別子、および使用するＲＡＩＤレベルを含むデータの種々の状態を示す変数を含むことが好ましい。好適には、ＳＤＤに関連するデータのライトバック（「ダーティ」データ）状態は、ダーティ・データ、ダーティ・バッファ、ダーティＬＲＵおよびフラッシングＬＲＵ値と関連して管理することが好ましい。

【００３１】

各ＳＤＤは、対応するＲＡＩＤストライプ１９４（すなわち、特定のパリティ・セットと関連する選択したストレージ１２８上のすべてのデータ）と整合し、特定のストライプ・バッファ・リスト（ＳＢＬ）に適合することが好ましい。各キャッシュ・ノードは、順方向および逆方向にリンクしているリストを使用して、仮想ブロック・アドレス（ＶＢＡ）を介して昇順にリンクしている所与の組の論理ディスクに対する能動ＳＤＤ構造により、いくつかの特定のＳＤＤを参照することが好ましい。

【００３２】

好適には、ＶＢＡ値は、ＲＡＩＤ割当てグリッド・システム（ＲＡＧＳ）と呼ばれるグリッド・システムを使用して、ＲＡＩＤデータ組織と整合される。通常、同一のＲＡＩＤ

ストリップ１９６（例えば、特定のパリティ・セットに貢献するすべてのデータなど）に属するブロックの任意の特定の集合体が、特定のシート上の特定の信頼性のあるストレージ・ユニット（ＲＳＵ）に割り当てられる。ブックはいくつかのシートからできていて、異なるストレージ１２８からのブロックの複数の隣接する組から作られる。実際のシートおよびＶＢＡに基づいて、このブックを、さらに、（冗長性を使用する場合）特定のデバイスまたはデバイスの組を示すゾーンに再分割することができる。

【００３３】

好適には、キャッシュ・マネージャ１７０は、ライトバック・コマンドおよびホスト読取りコマンドをできるだけ効率的に処理するために適合できるように管理することが好ましい。ダーティ・ブロックおよびロックしていないブロックを含む任意のＳＤＤは、ダーティとしてセットし、古さ（例えば、キャッシュ待機フラッシング中にデータが消費した時間など）で分類することが好ましい。特定の古さになると、フラッシングＬＲＵ変数をセットすることができ、フラッシング・リストが更新される。

【００３４】

図９は、コマンド・キュー内の転送要求のアレイ１９８である。これらの要求は、記憶容量およびブック２００が存在するストレージ１２８のブック２００の交点により定義されるセルによりマッピングされる。セルは、Ｘで示すこれらのセルが１つまたは複数の転送要求が懸案中であるこれらの位置を示すように、転送要求により配置される。アレイ１９８は、読取りコマンドを実行するために、アクセス・コマンドが特定の位置にトランスジューサ１４２（図５）を送った場合に使用されるシーク管理ツールであり、ライトバック・データの他の使用可能なブロックは、アクチュエータ１４４（図５）をショート・ストロークする際に現在の位置に関連してサービスを行うために容易に識別することができる。

【００３５】

ダーティ・データに対する転送要求は、シーク・マネージャ１７６に均一な分布が提示されるように、アレイ１９８にランダムに追加される。コマンド・キューに対するこの均一に分布しているベースは、ネットワーク負荷が予想した定常状態である場合に、またはそれ以下である場合に使用することが好ましいＣＳＣＡＮおよびエレベータ・ルーチンのような従来のシーク・スケジューリング・ルーチンにうまく適している。図１０は、例えば、ヘッド１４２が、どのように移動の予め定義した経路に沿って均一に分布している転送要求を発行しているシーク・マネージャ１７６と一緒に、内部の半径２０２から半径方向に外側に向かってディスク１３８を横断し、次に、外側の半径２０４から半径方向に内側に横断するのかを示す。

【００３６】

しかし、本実施形態の場合には、コマンド・キューからのデータのフラッシングの積極性は、アクセス・コマンドのホスト負荷に結びついている。すなわち、比較的大きなホスト負荷がかかっている間に十分積極的にフラッシングしないと、キャッシュ１２６が飽和する恐れがある。それ故、均一でない分布による負荷バースト活動の間にホスト負荷の特徴にコマンド・プロファイルをマッチさせるためにシーク・マネージャ１７４に重きを置くポリシー・エンジン１７４と一緒に、ポリシー・エンジン１７４は、また、それにより自身が均一な分布によるのではなく、物理ゾーンによりコマンド・キューを定義するモードへのキャッシュ・マネージャ１７０の切替も管理する。

【００３７】

図１１は、図９のアレイ１９８を形成しているストレージ１２８のディスク１３８のうちの１つである。これらのストレージ１２８のこのディスク１３８およびすべての他のディスク１３８は、記憶容量のブック２００により定義される環状ゾーン内に区分される。図１２は、各ブック２００内で現在懸案中のライトバック・データに対する転送要求の数の一例である。この実施形態によれば、ポリシー・エンジン１７４は、この場合はブック２００４である、最大数の懸案中のライトバック要求を含むブック２００を選択するためにキャッシュ・マネージャ１７０を選択的に管理し、コマンド・キューの内容をブック２

００４内のこれらのライトバック要求だけに制限する。ある実施形態の場合には、キャッシュ・マネージャ１７０は、すべてのライトバック要求がそのブック２００から発行されるまで、該最多のブック２００に制限することができる。他の実施形態の場合には、キャッシュ・マネージャ１７０は、周期的にライトバック要求の分布を再評価することができ、異なるブック２００がかなり大きい数のライトバック要求を有している場合には、キャッシュ・マネージャは、コマンド・キューの内容を該最多のブック２００内のライトバック要求にシフトする。

【００３８】

図１３は、本発明の実施形態による好適なゾーン・スケジューリングのための方法２０２のステップを示すフローチャートである。この方法２０２は、ライトバック・コマンドおよびホスト読取りコマンドに対する転送要求の均一な分布によるコマンド・キューをローディングするデフォルト・モードでブロック２０４からスタートする。デフォルト・モードは、ポリシー・エンジンが、ネットワーク負荷に関するデータを収集する１秒の間隔のような、しかしこれに限定されない予め定義した間隔中に実施される。最新のデータは、例えば、書込みに対する読取りの比率およびＩ／Ｏ速度で、ホスト負荷を動的に特徴付けるためにブロック２０６で使用される。

【００３９】

ブロック２０８においては、ポリシー・エンジンは、均一でない分布を含むＩ／Ｏコマンドのバーストがネットワーク負荷を監視することによりはっきりわかるか否かを判定する。ブロック２０８における判定が「いいえ」である場合には、制御はブロック２０４に戻り、そのためデフォルト状態が継続する。しかし、ブロック２０８における判定が「はい」である場合には、ブロック２１０において、ポリシー・エンジンは、コマンド・プロファイルの入手の際に連続的に調整する目的で、ホスト負荷の特徴、およびおそらく目標１８８、システム状態１９０およびバックグラウンドＩ／Ｏ１９２を使用して好適なゾーン・スケジューリング規則を呼び出す。例えば、制限無しで、飽和状態でコマンドを読み出すための高い書込みコマンドが発生している場合には、ポリシー・エンジンは、飽和状態から回復するまで、読取りに対する書込みの比率でコマンド・キューの内容をホスト負荷にマッチさせるためにキャッシュ・マネージャを管理することができる。ポリシー・エンジンは、飽和状態からできるだけ速くまた滑らかに回復するために、読取りレイテンシおよびＬＵＮクラス優先度のような他の規則を修正することさえできるし、または一時的に中止することさえできる。好適なゾーン・スケジューリング規則は、負荷データの次のバッチが収集される１秒の間隔のような所定の間隔中に呼び出され、制御は、ブロック２０６に戻る。

【００４０】

通常、この実施形態は、ネットワーク・アクセス・コマンドに応じてデータを転送するためにネットワークに接続するように構成されているストレージ・アレイ、およびコマンド・キュー内の転送要求の分布をアクセス・コマンドのネットワーク負荷の観察した特徴に相関付けることによりコマンド・キューを入手するための手段を予想している。この説明および添付の特許請求の範囲の意味のために、「入手するための手段」という用語は、明らかに、本明細書に記載する構造、および制御装置１１２がネットワーク負荷を特徴付け、特徴によりコマンド・キューの内容を直接調整することができるようにするその等価物を含む。フラッシング・リストを「直接」調整することにより、「入手するための手段」は、明らかに、ネットワーク負荷の特徴およびポリシー・エンジンからの規則に関連して、均一な分布からゾーン分布へのようなそれに応じて内容を切り替えるキャッシュ・マネージャを予測している。この説明および添付の特許請求の範囲の場合、「入手するための手段」は、ライトバック・データおよびコマンド・キューの内容に間接的に影響を与える恐れがある高度に局所化したバースト活動の周期となるＰＯＳ（ｐｏｉｎｔ　ｏｆ　ｓａｌｅ）取引のようなその発生の一致する位置を単に予想していない。

【００４１】

上記説明内で、本発明の種々の実施形態の構造および機能の詳細と一緒に、本発明の種

々の実施形態の多くの特徴および利点を説明してきたが、この詳細な説明は、例示としてのためだけのものであって、添付の特許請求の範囲を説明している用語の広い一般的な意味により示す全範囲に、特に本発明の原理の部材の構造および配置を変えることができることを理解されたい。例えば、本発明の精神および範囲から逸脱することなしに、特定の処理環境により特定の要素を変えることができる。

【００４２】

さらに、本明細書に記載する実施形態は、データ・ストレージ・アレイに関するものであるが、当業者であれば、特許請求の範囲に記載の主題は、それに限定されるものではなく、本発明の精神および範囲から逸脱することなしに、種々の他の処理システムも使用することができることを理解することができるだろう。

【図面の簡単な説明】

【００４３】

【図１】本発明の実施形態を組み込むコンピュータ・システムの図面である。

【図２】図１のコンピュータ・システムの一部の簡単な図面である。

【図３】本発明の実施形態によるインテリジェント記憶素子の分解等角図である。

【図４】図３のインテリジェント記憶素子の複数のドライブ・アレイの分解等角図である。

【図５】図４の複数のドライブ・アレイで使用する例示としてのデータ・ストレージである。

【図６】インテリジェント記憶素子内のアレイ制御装置の機能ブロック図である。

【図７】図６のアレイ制御装置の一部の機能ブロック図である。

【図８】インテリジェント・ストレージ素子内のアレイ制御装置およびデータ・ストレージのアレイの一部を示す略図である。

【図９】ライトバック・データ要求をランダムにマッピングする際にキャッシュ・マネージャが使用するアレイである。

【図１０】ＣＳＣＡＮシーク・スケジューリング技術と一緒に、均一に分布しているライトバック・データ要求を発行するストレージの略図である。

【図１１】データ・ストレージ容量のブックにより定義される環状のゾーンを示すストレージ・ディスクの簡単な平面図である。

【図１２】図１１の各ブック内のライトバック要求の観察したカウントを示す。

【図１３】本発明の実施形態による好適なゾーン・スケジューリングのための方法のステップを示すフローチャートである。

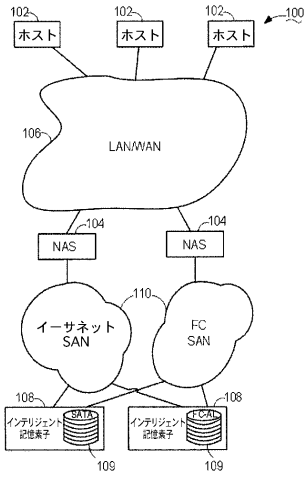
【符号の説明】

【００４４】

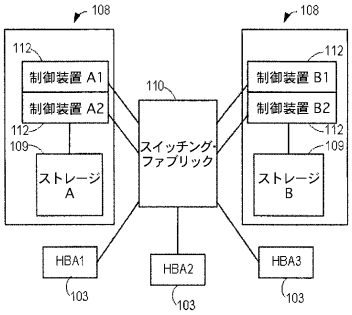
１００　コンピュータ・システム
１０２　ホスト
１０３　ホスト・バス・アダプタ（ＨＢＡ）
１０４　サーバ
１０６　ＬＡＮ／ＷＡＮ
１０８　インテリジェント記憶素子（ＩＳＥ）
１０９　データ・ストレージ容量
１１０　ＳＡＮ
１１２　二重化冗長制御装置
１１４　シェルフ
１１６　ミッドプレーン
１１８　ドライブ・アセンブリ（ＭＤＡ）
１２２　デュアル・バッテリー
１２４　デュアル交流電源
１２６　デュアル・インタフェース・モジュール
１２８　データ・ストレージ

１３０，１３２　隔壁
１３４　回路基板
１３６　コネクタ
１３８　ラッパー
１４０　モータ
１４２　読取り／書込みヘッド
１４４　回転アクチュエータ
１４６　ボイス・コイル・モータ（ＶＣＭ）
１４８　回路基板
１５０　フレックス回路
１５２　コネクタ
１５４　プロセッサ
１５６　不揮発性メモリ
１５８　ダイナミック・ランダム・アクセス・メモリ（ＤＲＡＭ）
１６０，１６２　Ｉ／Ｆ回路
１６４　経路制御装置
１６６　キャッシュ
１７０　キャッシュ・マネージャ
１７２　ＲＡＩＤコンテナ・サービス
１７４　ポリシー・エンジン
１７６　シーク・マネージャ
１７８，１８０，１８２　経路
１８８　目標
１９０　システム状態
１９８　アレイ
２００　ブック

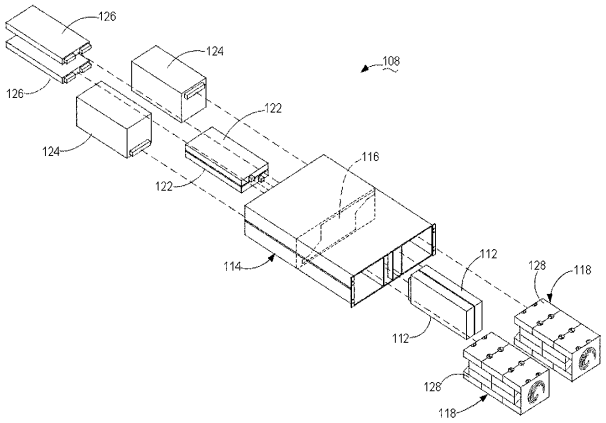
【図１】



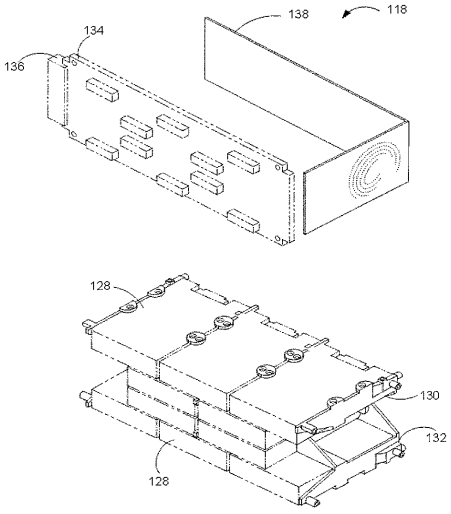
【図２】



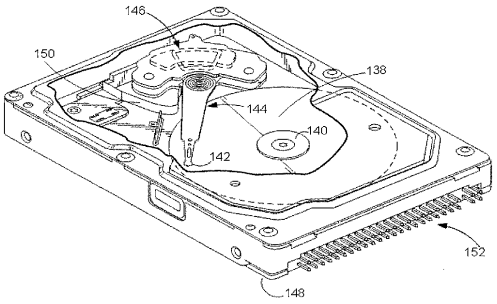
【図３】



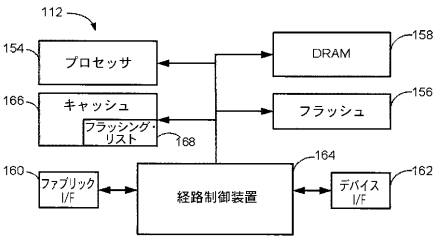
【図４】



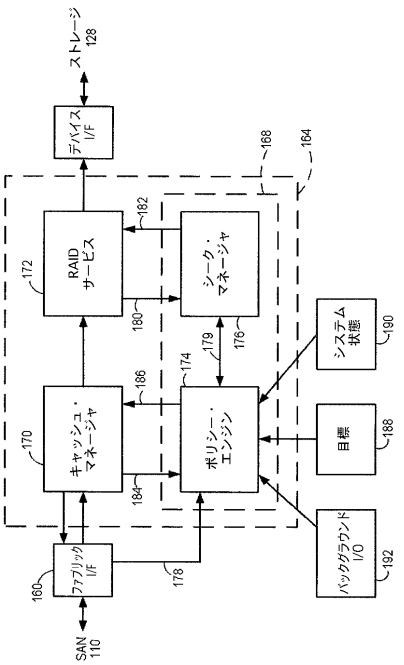
【図５】



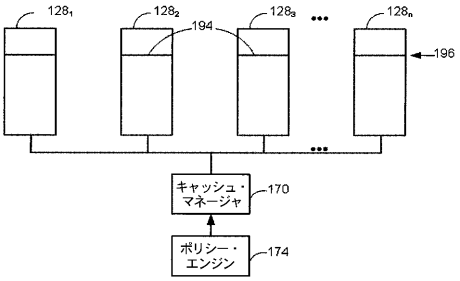
【図６】



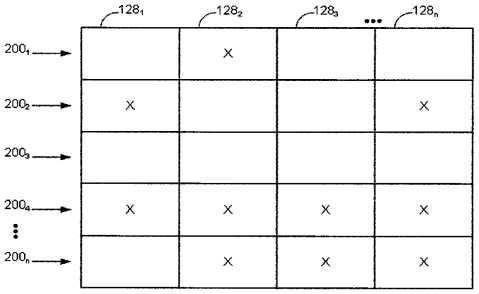
【図７】



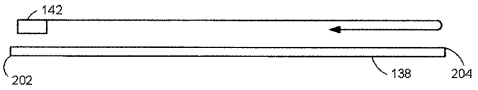
【図８】



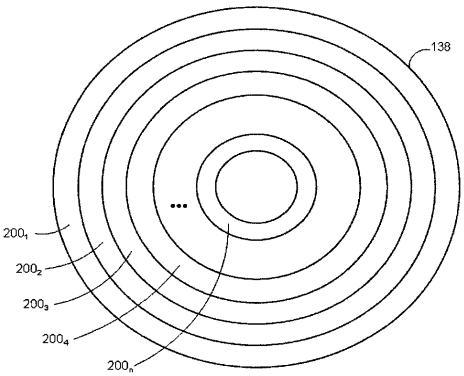
【図９】



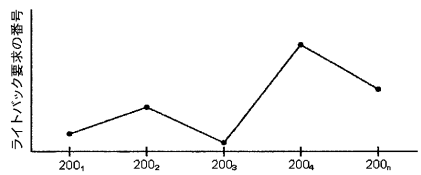
【図１０】



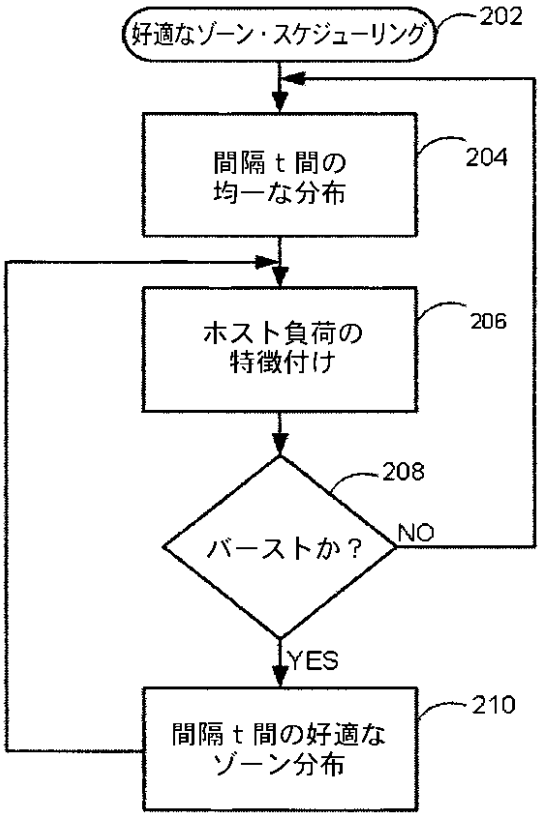
【図１１】



【図１２】



【図１３】



フロントページの続き

(74)代理人 100124523
弁理士 佐々木 眞人
(72)発明者 クラーク エドワード ルッベルス
アメリカ合衆国、コロラド、コロラド スプリングス、 ピニョン バレイ ロード ５３０１
(72)発明者 ロバート マイケル レスター
アメリカ合衆国、コロラド、コロラド スプリングス、 ロショルト ループ １４７１０

審査官 坂東 博司

(56)参考文献 特開平０８－１１５１６９（ＪＰ，Ａ）
特開２００４－２９５８６０（ＪＰ，Ａ）
特開２００６－１２０１１８（ＪＰ，Ａ）
米国特許出願公開第２００２／０１５６９７２（ＵＳ，Ａ１）
米国特許出願公開第２００５／００７６１１５（ＵＳ，Ａ１）
米国特許出願公開第２００６／０２５３６２１（ＵＳ，Ａ１）
特開平０６－１１０７７２（ＪＰ，Ａ）
特開平０５－２５７６１４（ＪＰ，Ａ）
特開２００７－１１５２３３（ＪＰ，Ａ）
Binny S.Gill and Dharmendra S. Modha, Proceedings of the 4th conference on USENIX Conference on File and Storage Technologies, WOW: wise ordering ofr writes - combining spatial and temporal locality in non-volatile caches, USENIX Association, ２００５年１２月１３日, １２９～１４２, ＵＲＬ, http://static.usenix.org/event/fast05/tech/full_papers/gill/gill.pdf#search='WOW Wise Ordering for Writes Combining Spatial and Temporal Locality in NonVolatile Caches'

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ ３／０６