(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5376624号
(P5376624)
(45)発行日　平成25年12月25日(2013.12.25)　(24)登録日　平成25年10月4日(2013.10.4)

(51)Int.Cl.　　F I
G06F 13/10 (2006.01)　　G06F 13/10 340A
G06F 3/06 (2006.01)　　G06F 3/06 301A
　　G06F 3/06 301F

請求項の数 20　(全 16 頁)

(21)出願番号　特願2008-166092 (P2008-166092)
(22)出願日　平成20年6月25日 (2008.6.25)
(65)公開番号　特開2009-9572 (P2009-9572A)
(43)公開日　平成21年1月15日 (2009.1.15)
審査請求日　平成23年6月7日 (2011.6.7)
(31)優先権主張番号　11/768,850
(32)優先日　平成19年6月26日 (2007.6.26)
(33)優先権主張国　米国 (US)

(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　100109162
弁理士　酒井　將行

最終頁に続く

(54)【発明の名称】ホスト適応シーク技術環境

(57)【特許請求の範囲】
【請求項１】
データ・ストレージ・システムにおけるコントローラであって、前記コントローラは、前記コントローラへのネットワーク負荷についての該ネットワーク負荷の性質を表すデータである定性情報を連続的に収集し該収集された定性情報に従って前記負荷を動的に特徴づけ該特徴づけにより得られる前記負荷の動的特徴を策定するポリシー・エンジン論理を含むホスト適応シーク技術環境（ＨＡＳＴＥ）モジュールを含み、前記ポリシー・エンジン論理は、キャッシュされたライトバック・データを含むフラッシング・リスト内の複数の入出力（Ｉ／Ｏ）要求から選択したＩ／Ｏ要求を発行するシーク・マネージャを支配しているホスト適応シーク技術環境規則であるＨＡＳＴＥ規則を定義する際に、前記負荷の前記動的特徴および予め定義した前記システムにおいて実現すべきシーク性能の目標値を示す性能目標を使用し、前記シーク・マネージャは、前記Ｉ／Ｏ要求の発行により、前記データ・ストレージ・システム内のデータ・ストレージ・デバイスへ前記リスト内のＩ／Ｏ要求をフラッシュする、コントローラ。
【請求項２】
前記動的特徴は、該ネットワーク上のコマンドの転送速度よりも前記データ・ストレージ・システムのレイテンシの影響がより大きいレイテンシに敏感なコマンドに対する前記ネットワーク上のコマンドの転送速度の影響が前記データ・ストレージ・システムのレイテンシの影響よりも大きい速度に敏感なコマンドの比率に関するものであり、前記レイテンシは、アクセスコマンドの指示が満たされるまでに要する期間である、請求項１に記載

のコントローラ。
【請求項３】
前記速度に敏感なコマンドは、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は、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１つである、請求項２に記載のコントローラ。
【請求項４】
前記動的特徴は、各コマンドと関連する帯域幅に関するものである、請求項１に記載のコントローラ。
【請求項５】
前記ＨＡＳＴＥ規則は、該ネットワーク上のコマンドの転送速度よりも前記データ・ストレージ・システムのレイテンシの影響がよりレイテンシに敏感なコマンドに対する前記ネットワーク上のコマンドの転送速度の影響が前記データ・ストレージ・システムのレイテンシの影響よりも大きい速度に敏感なコマンドの比率に関して、前記負荷にコマンドの分布または配列を示すコマンド・プロファイルを選択的にマッチさせ、前記レイテンシは、アクセスコマンドの指示が満たされるまでに要する期間である、請求項１に記載のコントローラ。
【請求項６】
前記ＨＡＳＴＥ規則は、前記コマンドに割り当てられた論理ユニット番号（ＬＵＮ）優先度の強制的な設定に関連する、請求項２に記載のコントローラ。
【請求項７】
前記ＨＡＳＴＥ規則は、前記コマンドの読取りコマンドのレイテンシの強制設定に関連する、請求項２に記載のコントローラ。
【請求項８】
個別データストレージデバイスのアレイに結合されており、前記ポリシー・エンジン論理からの前記ＨＡＳＴＥ規則に個々に応じる前記アレイ内の各データ・ストレージデバイス専用のシーク・マネージャを含む、請求項１に記載のコントローラ。
【請求項９】
ストレージ・システムへのネットワーク負荷のコマンド・ストリームを監視し、該コマンド・ストリーム内のコマンドについての該コマンドの性質を表すデータである定性情報を収集するステップと、
前記収集された定性情報に従って前記負荷を動的に特徴づけ、該特徴づけにより得られる前記負荷の動的特徴を策定するステップと、
前記動的特徴を使用して、前記ネットワークと前記ストレージ・システム間のキャッシュされたライトバック・データの通信を含むフラッシング・リスト内の複数のＩ／Ｏ要求から選択したＩ／Ｏ要求を発行するための、前記動的特徴に関連するホスト適応シーク技術環境規則であるＨＡＳＴＥ規則を生成するステップとを含み、前記Ｉ／Ｏ要求の発行により前記ストレージ・システム内のストレージデバイスへ前記リスト内のＩ／Ｏ要求がフラッシュされる、方法。
【請求項１０】
前記使用するステップは、前記ストレージ・システムに対する予め定義した前記システムにおいて実現すべきシーク性能の目標値を示す性能目標を考慮した規則を使用するステップを含む、請求項９に記載の方法。
【請求項１１】
前記監視するステップは、前記負荷を、該ネットワーク上のコマンドの転送速度よりも前記ストレージ・システムのレイテンシの影響がより大きいレイテンシに敏感なコマンドに対する前記ネットワーク上のコマンドの転送速度の影響が前記データ・ストレージ・システムのレイテンシの影響よりも大きい速度に敏感なコマンドの比率の点から特徴づけるステップを含み、前記レイテンシは、アクセスコマンドの指示が満たされるまでに要する期間である、請求項９に記載の方法。
【請求項１２】

前記監視するステップは、ライトバック・キャッシュ・コマンドが前記速度に敏感なコマンドであり、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が、前記レイテンシに敏感なコマンドであると特徴づけるステップを含む、請求項１１に記載の方法。

【請求項１３】

前記監視するステップは、前記負荷を、各アクセスコマンドに関連する帯域幅の点から特徴づけるステップを含む、請求項９に記載の方法。

【請求項１４】

前記使用するステップは、前記ＨＡＳＴＥ規則を、該ネットワーク上のコマンドの転送速度よりも前記ストレージ・システムのレイテンシの影響がより大きいレイテンシに敏感なコマンドに対する前記ネットワーク上のコマンドの転送速度の影響が前記データ・ストレージ・システムのレイテンシの影響よりも大きい速度に敏感なコマンドの比率に関連づけるステップを含み、前記レイテンシは、アクセスコマンドの指示が満たされるまでに要する期間である、請求項９に記載の方法。

【請求項１５】

前記使用するステップは、該ネットワーク上のコマンドの転送速度よりも前記ストレージ・システムのレイテンシの影響がより大きいレイテンシに敏感なコマンドに対する前記ネットワーク上のコマンドの転送速度の影響が前記ストレージ・システムのレイテンシの影響よりも大きい速度に敏感なコマンドの比率に関して、前記負荷にコマンドの分布または配列を示すコマンド・プロファイルを選択的にマッチさせるステップを含み、前記レイテンシは、アクセスコマンドの指示が満たされるまでに要する期間である、請求項９に記載の方法。

【請求項１６】

前記使用するステップは、前記規則を、前記コマンドの論理ユニット番号（ＬＵＮ）クラス優先度の強制的な設定に関連づけるステップを含む、請求項１１に記載の方法。

【請求項１７】

前記使用するステップは、前記規則を、前記コマンドの読取りコマンドに対するレイテンシの強制設定に関連づけるステップを含む、請求項１１に記載の方法。

【請求項１８】

ストレージ・システムへのネットワーク負荷のコマンド・ストリーム内のコマンドについての該コマンドの性質を表すデータである定性情報を収集するステップと、

前記収集された定性情報を使用して、読取りコマンドに対する書込みコマンドの比率に関して前記負荷を動的に特徴付けるステップと、

前記ネットワークと前記ストレージ・システム間のキャッシュされたライトバック・データの通信を含むフラッシング・リスト内の複数のＩ／Ｏ要求から選択したＩ／Ｏ要求を発行するための、前記動的に特徴付けるステップにより特徴づけられた動的特徴に関連するホスト適応シーク技術環境規則であるＨＡＳＴＥ規則を生成するステップとを含み、前記Ｉ／Ｏ要求の発行により前記ストレージ・システム内のストレージデバイスへ前記リスト内のＩ／Ｏ要求がフラッシュされる、方法。

【請求項１９】

前記使用するステップは、前記負荷の前記動的な特徴付けにより得られる動的特徴が予め定められたしきい値比率未満の場合に第１の規則を生成し、前記負荷の前記動的特徴が前記予め定められたしきい値よりも大きい場合に異なる第２の規則を生成する、請求項１８に記載の方法。

【請求項２０】

前記使用するステップは、前記負荷の前記動的特徴に、複数の選択されたＩ／Ｏ要求をマッチさせるステップを含む、請求項１８に記載の方法。

【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、分散データ・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以上の桁の処理能力、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の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、分散データ・ストレージ・システムのホスト適応シーク技術環境（ＨＡＳＴＥ）に関する。

【０００６】

ある実施形態の場合には、データ・ストレージ・システムおよび関連する方法は、負荷を動的に特徴付けるために、データ・ストレージ・システムにネットワーク負荷についての定性情報を連続的に収集し、コマンド・プロファイルを、特徴に関連するデータ・ストレージ・システムのデータ・ストレージに連続的に相関付けるポリシー・エンジンによりＨＡＳＴＥを実行する。

【０００７】

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。特定の時点でいずれのコマンドを発行するのかの選択は、経路制御装置１６４内に常駐するホスト適応シーク技術環境（ＨＡＳＴＥ：Ｈｏｓｔ Ａｄａｐｔｉｖｅ Ｓｅｅｋ Ｔｅｃｈｎｉｑｕｅ Ｅｎｖｉｒｏｎｍｅｎｔ）モジュール１６８と協働してシーク・マネージャが行う。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】

キャッシュ１６６は、ストライプ・データ記述子（ＳＤＤ）と呼ばれるデータ構造を使用して、制御装置１１２によりノード・ベースで管理される。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、対応するＲＡＩＤストライプ（すなわち、特定のパリティ・セットに関連する選択したストレージ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。

【００２１】

制御装置１１２により管理される各キャッシュ・ノードは、好適には、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）を通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。

【００２２】

好適には、ＶＢＡの値は、ＲＡＩＤ割当てグリッド・システム（ＲＡＧＳ）とも呼ばれるグリッド・システムを使用して、ＲＡＩＤデータ組織と整合される。通常、同じＲＡＩＤストリップ（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できるストレージ・ユニットに割り当てられる。ブックは多数のシートからできていて、異なるストレージ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分割することができる。

【００２３】

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（時刻、ＴＯＤ）、データがいずれのゾーン（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。ＳＤＤに関連するライトバック（「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の値と関連して管理することが好ましい。

【００２４】

制御装置１１２は、システム要件により、多数の異なるレベルのところでのライトバック・データ・プロセスを管理するために同時に動作することが好ましい。第１のレベルは、通常、全ＲＡＩＤストリップが検出された場合に、全ＳＤＤ構造の周期的フラッシングを含む。このことは、ＳＤＤが関連するデータをダーティと識別した場合に、ＲＡＩＤレベル変数に基づいて所与のＳＤＤに対して容易に行うことができる。好適には、このことは、十分な連続している隣接ＳＤＤが、十分ダーティなデータで満たされているか否かを判定するために逆方向のチェックを含む。そうである場合には、これらのＳＤＤ構造は、コマンド・キュー内に入れられ、データのフラッシングを開始するようにとの要求が行われる。

【００２５】

もっと小さな組のデータのフラッシングもＳＤＤをベースとして処理することが好ましい。ダーティ・ブロックおよびロックされていないブロックを含む任意のＳＤＤは、ダーティＬＲＵとしてセットし、古さの程度（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）により区分けすることが好ましい。特定のエージングに達した場合には、フラッシングＬＲＵ変数を設定し、コマンド・キューを更新することが好ましい。

【００２６】

連続しているダーティ・ブロックの特定の範囲がフラッシングに対してスケジューリングされると、制御装置１１２は、最も近い位置を有するＲＡＩＤレベルに基づいて、ダーティ・ブロックの他の範囲、すなわち、シーク時間の点で「近い」ブロック、または同じＲＡＩＤパリティ・ストリップへのアクセスを含むブロックを配置することが好ましい。

【００２７】

この実施形態によれば、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、Ｉ／Ｏコマンドのホスト負荷と結びついている。すなわち、比較的大きな負荷がかかっている時に十分積極的にフラッシングを行わないと、キャッシュ１２６が飽和する恐れがある。逆に、ホストの負荷が比較的低い時にあまりに積極的にフラッシングすると、キャッシュが不足してポテンシャル・キャッシュ・ヒットを満足できないままになる恐れがある。両方のシナリオともＩＳＥ１０８システムの性能に悪影響を及ぼす。

【００２８】

図７は、キャッシュ・マネージャ１７０および経路制御装置１６４（図６）内に常駐するＲＡＩＤコンテナ・サービス１７２を示す機能ブロック図である。この図は、またＨＡＳＴＥモジュール１６８のポリシー・エンジン１７４およびシーク・マネージャ１７６も示す。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ＨＡＳＴＥ規則に個々に応答する。

【００２９】

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内に位置することができる。ハードウェア内に位置する場合には、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、動的にホスト負荷を特徴づけそれに続けてシーク・マネージャ１７６を支配する経路１７９を介してＨＡＳＴＥ規則を発行する。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フラッシングするために、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。ポリシー・エンジン１７４は、経路１８４を通してキャッシュ１６６の状態を引き続き知らされ、経路１８６を通してキャッシュ・マ

ネージャにＨＡＳＴＥ規則を同様に発行することができる。

【００３０】

ポリシー・エンジン１７４は、１つまたは複数のネットワーク要求デバイスからのＩ／Ｏコマンドの現在の速度のようなリアルタイムの負荷についての定量データを収集することができる。ポリシー・エンジン１７４は、負荷を動的に特徴付け、コマンド・プロファイルを特徴付けとの関係においてストレージ１２８に連続的に調整するために、負荷に関する定性データを収集する。例えば、好適には、ポリシー・エンジン１７４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、ホストの負荷を特徴付ける連続データをリアルタイムで収集することが好ましい。

【００３１】

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にいずれの要求をフラッシングするかはたいした問題ではないからである。実際には、速度に敏感な要求は、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さえできるからである。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。

【００３２】

一方、１つまたは複数のストレージ１２８内に格納しているデータを読み出すためのホスト・アクセス・コマンドは、同様に、ホスト・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。

【００３３】

ポリシー・エンジン１７４は、また、関連するデータ・ファイル（帯域幅）のサイズ、アクセス・コマンドを開始するＨＢＡ１０３および／またはネットワーク・デバイス、ストレージ１２８のアクセス履歴、またはブック・アクセス履歴の期間のようなその任意の一部、タイムスタンプ・データ、ＲＡＩＤクラス、およびアクセス・コマンドが送られるＬＵＮクラスのような、しかしこれに限定されない、ホストの負荷を特徴付ける定性データを収集することができる。

【００３４】

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。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。

【００３５】

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、量的（定量的）なものであっても質的（定性的）なものであってもよいが、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）という点でネットワーク負荷のある種の要因である所望のコマンド・プロファイルの強化（強制設定）、異なるＬＵＮクラスへの割り当てられた優先度の強化（強制設定）、所望の読取りコマンド・レイテンシの強化（強制設定）等に限定されな

い。それ故、ポリシー・エンジン１７４は、キャッシュに格納しているライトバック・コマンドおよびもっと高い優先度ホスト読取りコマンドからの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を発行する許可を与える目的で、シーク・マネージャ１７６を支配しているＨＡＳＴＥ規則を定義するために、負荷特性および予め定義した性能目標１８８の両方を使用することができる。

【００３６】

さらに、ポリシー・エンジン１７４は、シーク・マネージャ１７６を支配してい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。この状態において、ポリシー・エンジン１７４は、投影され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不測の事態を実行する可能性が高い。ポリシー・エンジン１７４は、また、ストレージ１２８へのコマンド・プロファイルを調整する時に、シーク・マネージャ１７６を支配しているＨＡＳＴＥ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与しない懸案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。

【００３７】

図８は、本発明の例示としての実施形態によりＨＡＳＴＥを実施するための方法２００のステップを示すフローチャートである。この方法２００は、レイテンシおよびアドレス要因と関連するように、ダーティ・データの均一の分布によりランダムにフラッシングするために、ＨＡＳＴＥ規則を実行するデフォルト・モードでブロック２０２から開始する。デフォルト・モードは、その間に定性ＨＡＳＴＥデータが収集される、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行される。最近のＨＡＳＴＥデータは、書込みに対する読取りの比率で、ホストの負荷を動的に特徴付けるために、ブロック２０４で使用される。

【００３８】

ブロック２０６においては、ポリシー・エンジンは、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分かるか否かを判定する。ブロック２０６における判定が「いいえ」である場合には、制御は、ブロック２０２に戻り、デフォルト状態のままである。しかし、ブロック２０６の判定が「はい」である場合には、ブロック２０８において、ポリシー・エンジンは、ストレージへのコマンド・プロファイル内で引き続き調整を行う目的でＨＡＳＴＥ規則を呼び出すために、ホストの負荷、そして恐らく、目標１８８、システム状態１９０およびバックグラウンドＩ／Ｏ１９２を使用する。例えば、制限無しに、飽和状態で読取りコマンドに対する書込みコマンドの比率が高くなった場合には、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の点でホストの負荷にコマンド・プロファイルを一致させるために、シーク・マネージャを支配することができる。ポリシー・エンジンは、できるだけ迅速に滑らかに飽和状態から回復するために、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、一時的に中止させることもできる。ＨＡＳＴＥ規則は、その間にＨＡＳＴＥデータの次のバッチが収集され、制御がブロック２０４に戻る１秒間隔のような所定の間隔中に呼び出される。

【００３９】

例示としての実施形態の場合には、ＨＡＳＴＥデータは、最近の１秒の間隔中に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の決定、およびその比率の移動平均に対する比較の実施などで、ホストの負荷を質的に特徴付ける。下記の例示としての例のために、すべてのＬＵＮクラスは、すべてのＬＵＮクラスへのシステム・リソースの利用度がバランスのとれたものになるように、同じ優先度を有するものとして処理される。しかし、他の等価の実施形態の場合には、ＬＵＮクラスには、シーク・マネージャ１７６を支配するＨＡＳＴＥ規則に分かれる優先度レベルを割り当てれことができる。また、この例示としての例のために、ＨＡＳＴＥデータは、ＲＡＩＤ－１ストレージ・アレイ

に割り当てられるストレージ１２８の１つはプールに対して入手される。

【００４０】

この例の場合には、ＨＡＳＴＥモードの目標１８８は、各ストレージ１２８が、要求しているネットワーク・デバイスからの到着のその平均速度に比例して書込みコマンドおよび読取りコマンドを混合することである。分析は、適当なＲＡＩＤレベルで分かれる書込みに対する読取りの平均比率を観察することにより開始する。ＲＡＩＤ－１の場合には、例えば、各ホスト書込みコマンドに対して２つの書込みコマンドが発生する。この情報は、（計算のために）１６３８４の仮定した分母値を含む整数の分子であると見なすことができる「速度に敏感な要求の要因」（ＦＲＳＲ）にとり入れられる。例示としての例の場合には、ＦＲＳＲは５４６１である。それ故、「レイテンシに敏感な要求の要因」は下式により計算することができる。

【数１】

$$FLSR = 16384 - 5461$$

$$FLSR = 10923$$

【００４１】

ここで、ストレージ１２８は、２００ＩＯＰＳでアクセス・コマンドの所与の混合物を処理することができるものと仮定する。それ故、この前の８回の１秒サンプル間隔の移動平均は、１６００のアクセス・コマンドに跨る。この移動平均は、観察した１０６７の読取りコマンドおよび５３３の書込みコマンドに基づくものと仮定し、最後の１秒サンプル間隔中には、６６の読取りコマンドおよび３４の書込みコマンドが存在したと仮定する。正規化ＦＲＳＲを計算すると、下式のようになる。

【数２】

$$FRSR = \frac{(533 + 34 = 567) << 14}{(1600 + 100 = 1700)}$$

$$FRSR = 5464$$

【００４２】

ＦＲＳＲの目標は５４６１であり、実際のＦＲＳＲは目標を超える。それ故、ポリシー・エンジン１７４は、現在のコマンド・プロファイルを調整しない。

【００４３】

しかし、上記と同じ状況において、最新の１秒サンプルは、７０の読取りコマンドおよび３０の書込みコマンドを生成したと仮定する。この正規化ＦＲＳＲを計算すると、下式のようになる。

【数３】

$$FRSR = \frac{(533 + 30 = 563) << 14}{(1600 + 100 = 1700)}$$

$$FRSR = 5425$$

【００４４】

目標のＦＲＳＲと実際のＦＲＳＲとを比較する場合のマイナスのデルタは、速度検出（ライトバック）コマンドの数をコマンド・プロファイル内で増大する必要があることを示す。速度検出コマンドに対する適当な修正を決定するために、下記の関係を使用する。

【数４】

$$FRSR = \frac{((RS.IOPS + X) << 14)}{TOTAL.IOPS + X}$$

【００４５】
Ｘについてこの式を解くと、下式のようになる。
【数５】

$$X = \frac{((TOTAL.IOPS * FRSR) - (RS.IOPS << 14))}{FLSR}$$

【００４６】
上記例の場合には、下式のようになる。
【数６】

$$X = \frac{((1700 * 5461) - (563 << 14)}{10923}$$
$$X = 5$$

【００４７】
それ故、シーク・モード１７６は、所望の速度に観察した速度を一致させるために余分の書込みコマンドを実行させる。このことは、例えば、ドライブ自身を、高い優先度（またはレイテンシ検出）コマンドになるように、内部に待ち行列を含む５つの書込みコマンドを促進させることにより行うことができる。
【００４８】
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ＨＡＳＴＥ内のストレージにコマンド・プロファイルを制御するための手段を予想する。この説明および添付の特許請求の範囲の意味のために、「制御するための手段」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、特徴により制御装置・プロファイル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。コマンド・プロファイルを「直接」調整することにより、「制御するための手段」は、制御装置１１２が、特徴に応えてダーティ・データおよび未決のホスト読取り要求からの複数のＩ／Ｏコマンドから選択したＩ／Ｏコマンドの発行を実際に調整することをはっきりと意味する。この説明および添付の特許請求の範囲の意味のために、「制御するための手段」という用語は、それによりキャッシュ・マネージャが、間接的にコマンド・プロファイルに影響を与える恐れがあるフラッシュ・リストを決定する機構の単なる調整を予想していない。
【００４９】
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。
【００５０】
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】

【００５１】

【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】インテリジェント記憶素子内のアレイ制御装置の一部の機能ブロック図である。

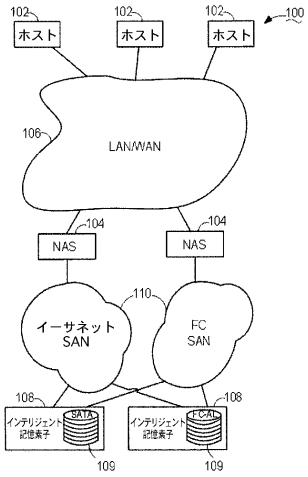
【図８】本発明の実施形態によるＨＡＳＴＥを呼び出すための方法のステップを示すフローチャートである。

【符号の説明】

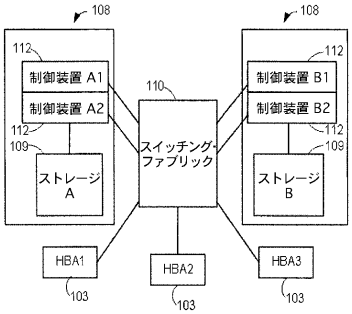
【００５２】

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）
１１０　ＳＡＮ
１０９　データ・ストレージ容量
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
１４８　回路基板
１５０　フレックス回路
１５４　プロセッサ
１６０，１６２　Ｉ／Ｆ回路
１６４　経路制御装置
１６６　キャッシュ
１６８　ホスト適応シーク技術環境（ＨＡＳＴＥ）モジュール

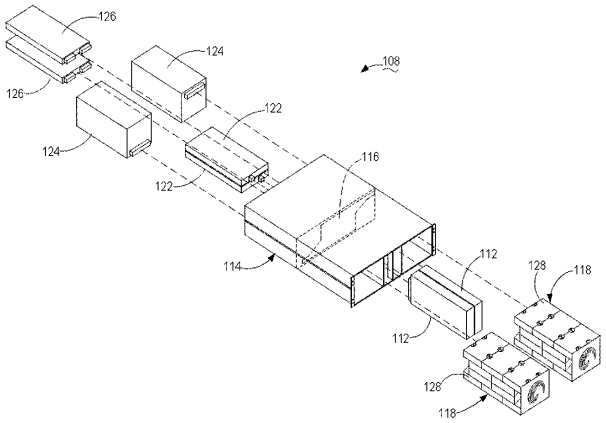
【図１】



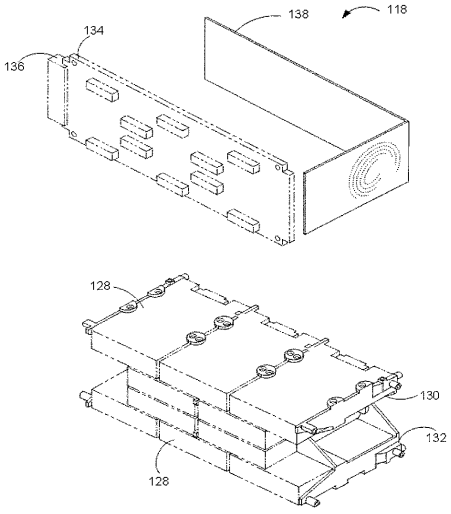
【図２】



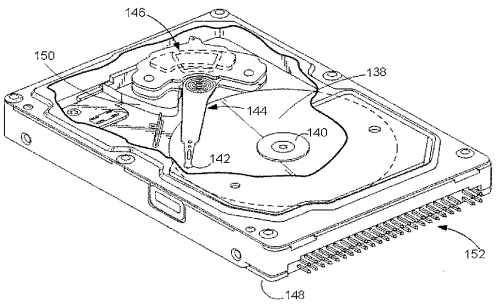
【図３】



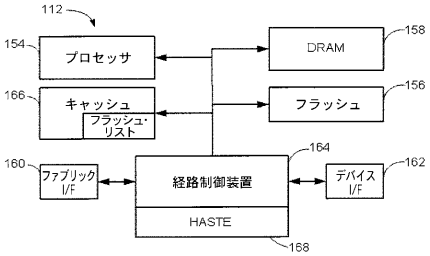
【図４】



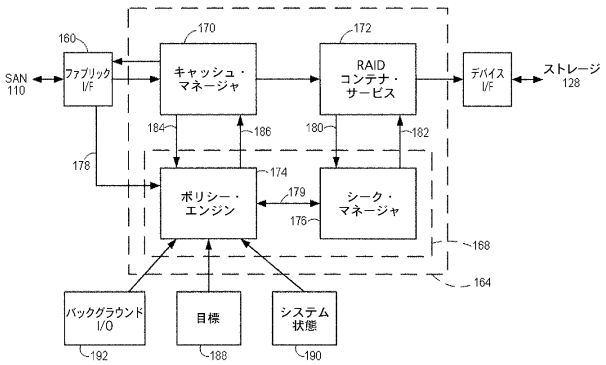
【図５】



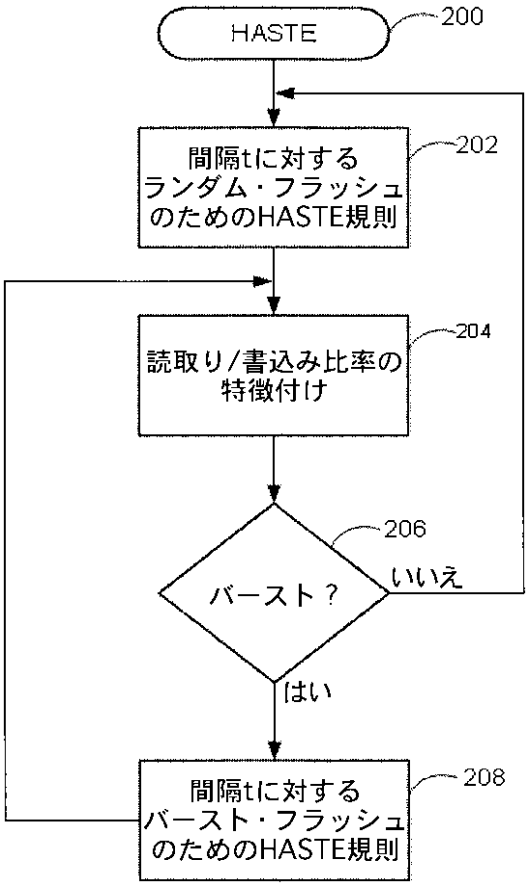
【図６】



【図７】



【図８】


HASTE
200
間隔tに対する
ランダム・フラッシュ
のためのHASTE規則
202
読取り/書込み比率の
特徴付け
204
バースト？
206
いいえ
はい
間隔tに対する
バースト・フラッシュ
のためのHASTE規則
208

フロントページの続き

(74)代理人 100111246
弁理士 荒川 伸夫
(74)代理人 100124523
弁理士 佐々木 眞人
(74)代理人 100066692
弁理士 浅村 皓
(74)代理人 100072040
弁理士 浅村 肇
(74)代理人 100091339
弁理士 清水 邦明
(74)代理人 100094673
弁理士 林 鈺三
(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０６－２４３０４２（ＪＰ, Ａ）
特開平０５－１０８２７４（ＪＰ, Ａ）
特開２００４－２９５８６０（ＪＰ, Ａ）
特表２００６－５２１６４０（ＪＰ, Ａ）
米国特許出願公開第２００２／００１５９７２（ＵＳ, Ａ１）
特表２００２－５４２５６８（ＪＰ, Ａ）
米国特許出願公開第２００４／０１６２９０１（ＵＳ, Ａ１）
米国特許出願公開第２００３／０１４９８３８（ＵＳ, Ａ１）
米国特許出願公開第２００２／０１５６９７２（ＵＳ, Ａ１）
米国特許第０６３３９８１１（ＵＳ, Ｂ１）
米国特許第０６１７００４２（ＵＳ, Ｂ１）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １３／１０
Ｇ０６Ｆ ３／０６