(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5638744号
(P5638744)

(45)発行日　平成26年12月10日(2014.12.10)　(24)登録日　平成26年10月31日(2014.10.31)

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

請求項の数 14　(全 14 頁)

(21)出願番号　特願2008-166109 (P2008-166109)
(22)出願日　平成20年6月25日 (2008.6.25)
(65)公開番号　特開2009-9573 (P2009-9573A)
(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22日 (2011.6.22)
審判番号　不服2013-17342 (P2013-17342/J1)
審判請求日　平成25年9月9日 (2013.9.9)
(31)優先権主張番号　11/768,849
(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　100111246
弁理士　荒川　伸夫

最終頁に続く

(54)【発明の名称】コマンド・キュー・ローディング

(57)【特許請求の範囲】
【請求項１】
データ・ストレージ・システムにおける制御装置であって、前記制御装置は、
ネットワーク・デバイスから前記制御装置へのコマンドを含む負荷の動的特徴を作成するために、および前記動的特徴を、前記コマンドに含まれるデータ転送要求のコマンド・キューの深さに連続的に相関させるように対応のコマンドキューの深さを選択するために、前記ネットワーク・デバイスから前記制御装置への前記負荷に関する定性データを連続的に前記負荷に含まれるコマンドから収集するポリシー・エンジンを含む経路制御装置を備え、前記定性データは、前記負荷内のコマンドを速度に敏感なコマンドとレイテンシに敏感なコマンドとを識別して分類し、前記動的特徴は、前記収集された定性データに従って前記負荷を動的に特徴づけた結果得られたものであり、
前記速度に敏感なコマンドは、前記ネットワークデバイスからの該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コマンドであり、前記レイテンシに敏感なコマンドは、前記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コマンドである、制御装置。
【請求項２】
前記経路制御装置は、さらに、前記ポリシーエンジンに応答して、前記負荷が、デフォルトモードとして期待される定常状態と同じか、またはそれ以下であることを前記動的特徴が示している場合に、コマンド・キューの深さとして第１のコマンド・キューの深さを

設定し、前記動的特徴がコマンド活動のバーストを示している場合に、前記第１のコマンド・キューの深さより深い第２のコマンド・キューの深さを設定するキャッシュ・マネージャを備える、請求項１に記載の制御装置。

【請求項３】

前記第２のコマンド・キューの深さが、前記第１のコマンド・キューの深さより少なくとも約１桁深い、請求項２に記載の制御装置。

【請求項４】

前記選択したコマンド・キューの深さが、前記選択したコマンド・キューの深さを有する前記コマンド・キューからデータ転送コマンドを選択的に発行してコマンドの分布または配列を示すコマンド・プロファイルを規定するシーク・マネージャに提示される、請求項３に記載の制御装置。

【請求項５】

前記ポリシー・エンジンが、前記動的特徴に関連して、また前記コマンド・プロファイルを定義する際に、前記シーク・マネージャが現在実現する目標について前記シーク・マネージャのシーク動作を規定する規則を設定する、請求項４に記載の制御装置。

【請求項６】

前記ポリシー・エンジンの規則が、前記レイテンシに敏感なコマンドに対する前記速度に敏感なコマンドの数の比率の点において前記コマンド・プロファイルを前記動的特徴に相関させる、請求項５に記載の制御装置。

【請求項７】

前記ポリシー・エンジンの規則が、前記レイテンシに敏感なコマンドの数に対する前記速度に敏感なコマンドの数の比率の点において前記コマンド・プロファイルを前記動的特徴とマッチさせる、請求項６に記載の制御装置。

【請求項８】

前記ポリシー・エンジンの規則が、前記コマンド・プロファイルを前記コマンドの所望の最大レイテンシと相関させる、請求項５に記載の制御装置。

【請求項９】

前記ポリシー・エンジンの規則が、前記コマンド・プロファイルを異なるＬＵＮのグループに割り当てられた優先度と相関させる、請求項５に記載の制御装置。

【請求項１０】

前記動的特徴が、前記レイテンシに敏感なコマンドに対する前記速度に敏感なコマンドの比率である、請求項３から９のいずれかに記載の制御装置。

【請求項１１】

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１０に記載の制御装置。

【請求項１２】

前記負荷の動的特徴が、各コマンドに関連するファイル・サイズである、請求項１１に記載の制御装置。

【請求項１３】

前記ポリシー・エンジンが、有限状態機械を備える、請求項１２に記載の制御装置。

【請求項１４】

制御装置によって行われる方法であって、

ネットワーク・デバイスからストレージ・システムへのコマンド・ストリーム負荷を監視して、前記コマンド・ストリーム内のコマンドについての定性データを前記コマンドから収集して、前記定性データに従って前記負荷を動的に特徴づけ、該特徴づけにより前記負荷の動的特徴を作成するステップを含み、前記定性データは、前記コマンドを前記ネットワークデバイスからの該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と前記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とを識別して分類し、
前記動的特徴を使用して、前記コマンドに含まれるデータ転送要求の選択すべきコマンド・キューの深さを、前記動的特徴が前記コマンドに含まれるデータ転送要求の対応のコマンド・キューの深さに連続的に相関させるように規定するステップをさらに含む、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、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、バースト負荷の状況下でシーク・コマンド・キューを選択的にローディング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、負荷の特徴をライトバック・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コマンド・キューの深さに連続的に相関付けるために、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】
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・ストレージ素子（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、データをドライブ１２８に格納する場合に、およびデータをドライブ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（ＲＳＵ）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、ホスト読取り要求のレイテンシに悪影響を与える恐れがある。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、ＲＡＩＤコンテナ・サービス１７２、ポリシー・エンジン１７４、および経路制御装置１６４（図６）内に常駐するシーク・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。

【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で実装することができる。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コマンド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。

【００３０】

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、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、Ｉ／Ｏコマンドに関連するファイル・サイズのような、しかしこれに限定されない他のものでホストの負荷を特徴付ける定性データを収集することもできる。

【００３１】

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施する。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。

【００３２】

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。

【００３３】

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。

【００３４】

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原因となり得る。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。この場合には、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。

【００３５】

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ル間隔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、自立式カウンタは、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、現在の１秒比率を照合するために９番目のスロットを含む、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、最新のサンプル値を加算し、次に、比率の最新の移動平均を計算するために８で除算を行う。

【００３６】

例示としての例について引き続き説明すると、ポリシー・エンジン１７４が、負荷バーストが発生していることを観察した場合には、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ために、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、読取りコマンドに対する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、一時的に延期することさえできる。

【００３７】

このような状態の場合、シーク・マネージャ１７６にキャッシュ・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、転送要求のコマンド・キューの深さを増大することにより、フラッシング性能を有意に改善することができることが分かっている。すなわち、この実施形態は、また、シーク・マネージャ１７６に提示されたコマンド・キューを「ロードアップ」するキャッシュ・マネージャ１７０を管理する規則を定義するために、負荷の特徴を使用するポリシー・エンジン１７４を予想している。

【００３８】

図８および図９は、それぞれ本発明の実施形態による、定常状態の負荷の状態およびＩ／Ｏバースト状態中のアレイ制御装置１１２の関連部分を示す略図である。１９４および１９６で示すキャッシュ１６６内のライトバック・キャッシュされたデータのリストは、通常、両方の状態の場合、数千ノード程度である。ドライブ・キュー１９８は、比較的短く、両方の状態の場合一定である。しかし、キャッシュ・マネージャ１７０は、定常状態の場合、第１のコマンド・キューの深さ２００を定義するために、ポリシー・エンジン１７４からの規則に応じ、バースト状態において第１のコマンド・キューの深さ２００より有意に深い第２のコマンド・キューの深さ２０２を定義することに留意されたい。通常、コマンド・キューの深さ２００は、約２０～４０転送要求の範囲内である。飽和状態でバースト状態が発生している場合には、調整したコマンド・キューの深さ２０２は、約２００～５００の要求の範囲内にあるか、または少なくとも定常状態の状態より大きいことが好ましい。

【００３９】

図１０は、本発明の実施形態によるコマンド・キュー・ローディングのための方法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、２０～４０程度の転送要求のようなデフォルト・モード・コマンド・キューの深さで、ブロック２０から開始する。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、書込みに対する読取りの比率およびＩ／Ｏ速度等で、ホスト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。

【００４０】

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わ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、そのためデフォルト状態が継続する。しかし、ブロック２０６における判定が「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、デフォルト深さより約１レベル深いような有意に深い深さにコマンド・キューのローディングを呼び出す目的で、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を使用する。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・プロファイルをホストにマッチさせるために、シーク・マネージャを管理することができる。それを支持して、ポリシー・エンジンは、この実施形態によりコマンド・キューの深さを有意に増大する。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、一時的に中止することさえできる。ローディングしたコマンド・キュー規則は、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る。次に、制御はブロック２０４に戻る。

【００４１】

通常、この実施形態は、ネットワーク・アクセス・コマンドに応じて転送データへネットワークを接続するように構成されているストレージ・アレイ、およびデータ転送要求のサイズをアクセス・コマンドのネットワーク負荷の観察した特徴と相関付けることにより、コマンド・キューをローディング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、「ローディングのための手段」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。

【００４２】

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。

【００４３】

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】

【００４４】

【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・ストレージ素子の分解等角図である。

【図４】図３のインテリジェント・ストレージ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。

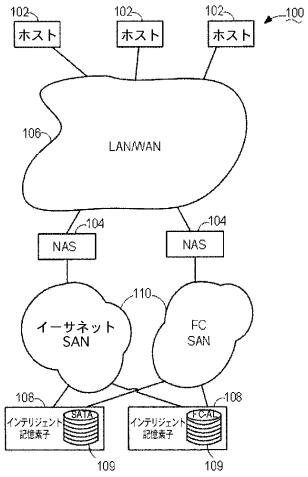
【図６】図３のインテリジェント・ストレージ素子内のアレイ制御装置の機能ブロック図である。

【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。

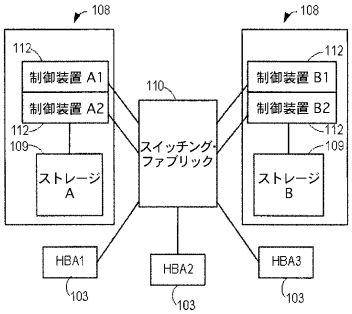
【図８】予想した定常状態のネットワーク負荷状態中のコマンド・キューのサイズを示す略図である。

【図９】図８類似の図面であるが、Ｉ／Ｏバースト状態中にシーク・マネージャに提示されるローディングしたコマンド・キューを示す。
【図１０】本発明の実施形態によるコマンド・キュー・ローディ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】
【００４５】
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・ストレージ素子（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

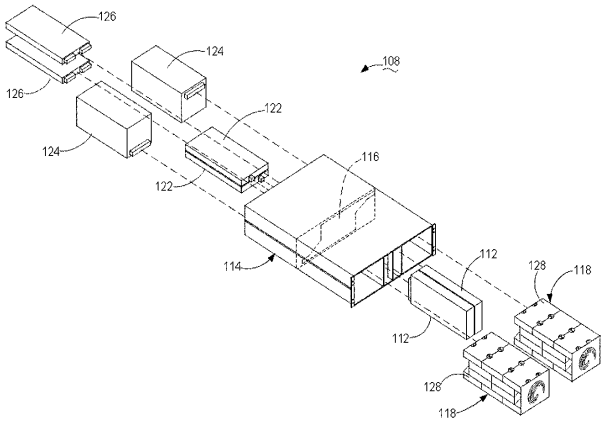
【図１】



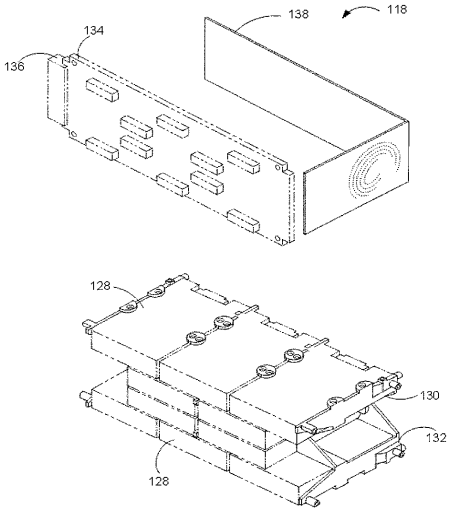
【図２】



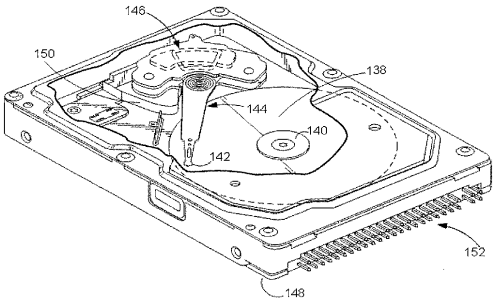
【図３】



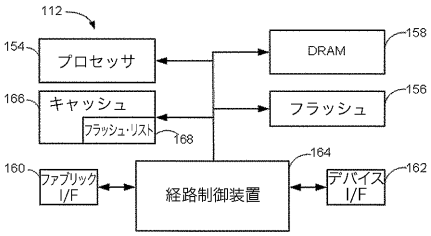
【図４】



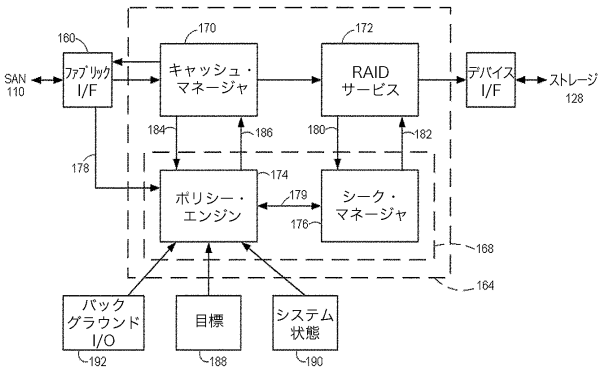
【図５】



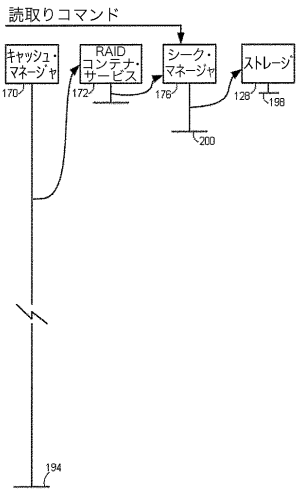
【図６】



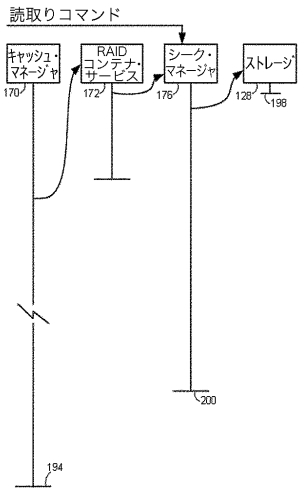
【図７】



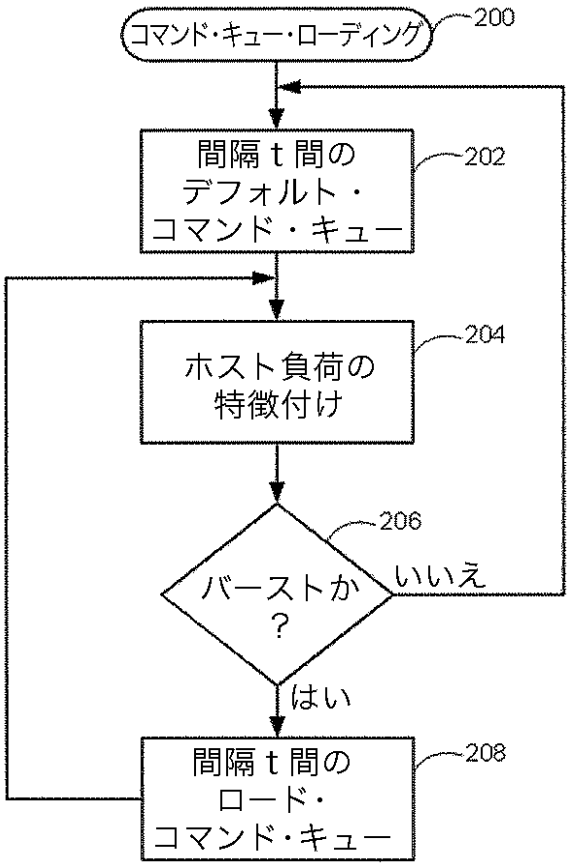
【図８】


読取りコマンド
キャッシュ・マネージャ
170
RAID コンテナ・サービス
172
シーク・マネージャ
176
ストレージ
128
198
200
194

【図９】


読取りコマンド
キャッシュ・マネージャ
170
RAID コンテナ・サービス
172
シーク・マネージャ
176
ストレージ
128
198
200
194

【図１０】


コマンド・キュー・ローディング
200
間隔ｔ間の
デフォルト・
コマンド・キュー
202
ホスト負荷の
特徴付け
204
バーストか
？
206
いいえ
はい
間隔ｔ間の
ロード・
コマンド・キュー
208

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

合議体
審判長 乾 雅浩
審判官 千葉 輝久
審判官 和田 志郎

(56)参考文献 米国特許第５４２６７３６（ＵＳ，Ａ）
特表２００６－５２１６４０（ＪＰ，Ａ）
特開２００４－２９５８６０（ＪＰ，Ａ）
特開平６－２４３０４２（ＪＰ，Ａ）
特開２００６－１８６８９（ＪＰ，Ａ）
特開２００７－１１５２３３（ＪＰ，Ａ）
特開２００４－１７１４１１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F3/06-3/08