(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2008-16024
(P2008-16024A)
(43)公開日　平成20年1月24日(2008.1.24)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 3/06	(2006.01)	G06F 3/06	302A	5B005
G06F 12/08	(2006.01)	G06F 3/06	540	5B065
		G06F 12/08	501F	
		G06F 12/08	557	
		G06F 12/08	543B	

審査請求 未請求　請求項の数 19　ＯＬ　外国語出願　（全 15 頁）

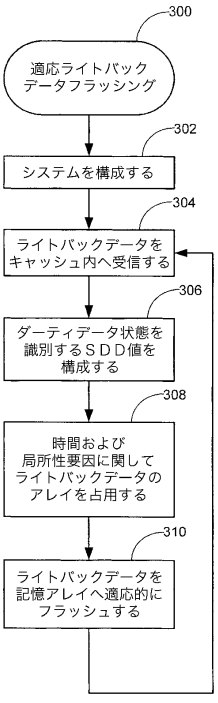
(21)出願番号　特願2007-172478 (P2007-172478)
(22)出願日　平成19年6月29日 (2007.6.29)
(31)優先権主張番号　11/480,128
(32)優先日　平成18年6月30日 (2006.6.30)
(33)優先権主張国　米国 (US)

(71)出願人　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，カリフォルニア，スコッツ　バレイ，ピー．オー．ボックス　６６３６０，　ディスク　ドライブ　９２０
(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　鉌三

最終頁に続く

(54)【発明の名称】キャッシュされたデータのダイナミック適応フラッシング

(57)【要約】
【課題】キャッシュされたライトバックデータを記憶アレイにフラッシングする方法および装置を得る。
【解決手段】ライトバックデータのセットが記憶アレイの異なる場所にわたってデータの実質的に均一な分布を維持する目的でキャッシュメモリ内に整列して蓄積される。その後、データの整列されたセットは実質的にライトバックデータの付加セットがホストによりキャッシュメモリへ提供されるレートでキャッシュメモリから記憶アレイへ転送される。好ましくは、各ライトバックデータは複数の隣接データブロックを含み、好ましくは記憶アレイに関してデータの選択された近接範囲内の別個のアクセスコマンドの操作と共に記憶装置に書き込まれる（フラッシュされる）。好ましくは、アレイ内のライトバックデータの各セットに対してストライプデータ記述子（ＳＤＤ）が維持される。
【選択図】図８



【特許請求の範囲】
【請求項１】
記憶アレイの場所に関して記憶アレイ内のライトバックデータのセットのアレイを形成するステップと、実質的にライトバックデータの付加セットがホストによりキャッシュメモリへ提供されるレートでライトバックデータの前記セットをキャッシュメモリから記憶アレイへ転送するステップと、を含む方法。
【請求項２】
請求項１に記載の方法であって、ライトバックデータの各セットは複数の隣接データブロックを含む方法。
【請求項３】
請求項１に記載の方法であって、転送ステップは記憶アレイのトランスデューサを記録媒体に隣接する選択された場所へ移動させて選択されたアクセスコマンドをサービスするステップと、選択されたアクセスコマンドの選択された場所とライトバックデータの選択されたセットに対する関連場所との近接に関してライトバックデータの選択されたセットをキャッシュメモリから媒体へ転送するステップと、を含む方法。
【請求項４】
請求項３に記載の方法であって、前記近接はトランスデューサの選択されたシーク長を含む方法。
【請求項５】
請求項１に記載の方法であって、さらに、アレイ内のライトバックデータの各セットに対してストライプデータ記述子（ＳＤＤ）を与えるステップを含み、ＳＤＤはライトバックデータを記憶アレイへフラッシングするための準備完了として識別する値を含む方法。
【請求項６】
請求項１に記載の方法であって、キャッシュメモリはライトバックデータのｍセットを格納し、蓄積ステップは記憶アレイへのフラッシングに備えてアレイの異なる非隣接場所に関連する前記ライトバックデータのｎセットを配列するステップを含み、ｎはｍよりも小さい方法。
【請求項７】
請求項１に記載の方法であって、さらに、ライトバックデータの前記付加セットがキャッシュメモリへ移されるレートを監視するステップを含み、それに関連して転送ステップを実施して前記キャッシュメモリ内に前記ライトバックデータの実質的に均一な分布を維持する方法。
【請求項８】
記憶アレイの関連する場所へ転送するのに利用できるライトバックデータのセットを格納するキャッシュメモリと、ライトバックデータの前記セットをキャッシュメモリ内に配列して実質的に均一な分布を記憶アレイを横切って提供し、かつ実質的にライトバックデータの付加セットがホストによりキャッシュメモリへ提供されるレートでライトバックデータの前記セットをキャッシュメモリから記憶アレイへ転送するように構成されたプロセッサと、を含む装置。
【請求項９】
請求項８に記載の装置であって、プロセッサはライトバックデータのセットを選択的にフラッシングリストへ移すキャッシュ・マネージャを含む装置。
【請求項１０】
請求項８に記載の装置であって、記憶アレイは記録媒体に隣接する選択された場所へ移されて選択されたアクセスコマンドをサービスするように構成されたトランスデューサを含み、プロセッサは選択されたアクセスコマンドの選択された場所とライトバックデータの選択されたセットに対する関連場所との間の近接に関してライトバックデータの選択されたセットをキャッシュメモリから媒体へ転送する装置。
【請求項１１】
請求項１０に記載の装置であって、前記近接はトランスデューサの選択されたシーク長

を含む装置。
【請求項１２】
請求項８に記載の装置であって、プロセッサはアレイ内のライトバックデータの各セットに対するストライプデータ記述子（ＳＤＤ）を発生し、ＳＤＤはライトバックデータを記憶アレイへのフラッシング準備完了として識別する値を含む装置。
【請求項１３】
請求項８に記載の装置であって、キャッシュメモリはライトバックデータのｍセットを格納し、プロセッサは前記ライトバックデータのｎセットを記憶アレイフラッシングするために配置し、ｎはｍよりも小さい装置。
【請求項１４】
請求項８に記載の装置であって、プロセッサはさらにライトバックデータの前記付加セットがキャッシュメモリへ移されるレートを監視し、それに関して前記データを転送してキャッシュメモリ内に前記ライトバックデータの実質的に均一な分布を維持する装置。
【請求項１５】
記憶アレイの関連する場所へ転送するのに利用できるライトバックデータの複数のセットを配列するキャッシュメモリと、実質的にライトバックデータの付加セットがホストによりキャッシュメモリへ提供されるレートでライトバックデータの前記セットを前記アレイから記憶アレイへ転送する第１の手段と、を含む装置。
【請求項１６】
請求項１５に記載の装置であって、第１の手段はキャッシュ・マネージャを含む装置。
【請求項１７】
請求項１５に記載の装置であって、記憶アレイは記録媒体に隣接する選択された場所へ移されて選択されたアクセスコマンドをサービスするように構成されたトランスデューサを含み、第１の手段は選択されたアクセスコマンドの選択された場所とライトバックデータの選択されたセットに対する関連場所との間の近接に関してライトバックデータの選択されたセットをキャッシュメモリから媒体へ転送する装置。
【請求項１８】
請求項１７に記載の装置であって、前記近接はトランスデューサの選択されたシーク長を含む装置。
【請求項１９】
請求項１５に記載の装置であって、第１の手段はさらにライトバックデータの前記付加セットがキャッシュメモリへ移されるレートを監視し、前記データを転送してプール内にライトバックデータの実質的に均一な分布を維持する装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は一般的にデータ記憶システムの分野に関し、制約はしないが、特に、記憶アレイへのキャッシュされたデータのダイナミック適応フラッシング方法および装置に向けられている。
【背景技術】
【０００２】
高速かつ効率的方法でデータにアクセスするために記憶装置が使用される。ある種の記憶装置は、媒体表面上に規定されたトラックにデータを書込みその後データを読み出す１つ以上のデータトランスデューサと共に、回転可能な記録媒体を使用する。
【０００３】
マルチデバイス・アレイ（ＭＤＡ）は多数の記録装置を利用して統合されたメモリスペースを形成する。ＭＤＡに対する１つの広く利用されるフォーマットはＲＡＩＤ（ｒｅｄｕｎｄａｎｔ　ａｒｒａｙ　ｏｆ　ｉｎｄｅｐｅｎｄｅｎｔ　ｄｉｓｃｓ）構成を利用し、入力データはアレイ内の多数の記憶装置にわたって格納される。ＲＡＩＤレベルに応じて、ミラーリング、ストライピングおよびパリティコード発生を含むさまざまな技術を利

用して格納されたデータの完全性を高めることができる。
【発明の開示】
【発明が解決しようとする課題】
【０００４】
記憶容量および性能のレベルは絶えず高まることが要求され続けているため、このようなアレイ内の記憶装置の動作が管理される方法の改善が継続的に必要とされている。本発明の好ましい実施例は一般的にこれらおよびその他の改善に向けられる。
【課題を解決するための手段】
【０００５】
本発明の好ましい実施例は一般的に記憶アレイへキャッシュされたライトバックデータをフラッシングする装置および方法に向けられている。
【０００６】
好ましい実施例では、記憶アレイの異なる場所にわたってデータの実質的に均一な分布を維持する目的で、ライトバックデータのセットがキャッシュメモリ内に整列して蓄積される。その後、データの整列されたセットは実質的にライトバックデータの付加セットがホストによりキャッシュメモリへ提供されるレートでキャッシュメモリから記憶アレイへ転送される。
【０００７】
ライトバックデータの各セットは好ましくは複数の隣接データブロックを含み、好ましくは記憶アレイに関してデータの選択された近接範囲内の別個のアクセスコマンドの操作と共に記憶装置に書き込まれる（フラッシュされる）。好ましくは、アレイ内のライトバックデータの各セットに対してストライプデータ記述子（ＳＤＤ）が維持される。
【０００８】
このようにして、キャッシュされたライトバックデータのフラッシングは全体ホストＩ／Ｏアクセスレートに著しい変動を生じることはない。
【０００９】
下記の詳細な説明を読みかつ添付図を検討すれば、特許請求された本発明を特徴づけるこれらおよびさまざまな他の特徴および利点が明白となる。
【実施例】
【００１０】
図１はユーザデータを格納かつ検索するように構成された典型的な記憶装置１００を示す。好ましくは、装置１００はハードディスクドライブとして特徴づけられるが、所望により、他の装置構成も容易に利用することができる。
【００１１】
ベースデッキ１０２がトップカバー（図示せず）と嵌合して密閉筐体を形成する。スピンドルモータ１０４が筐体内に搭載されて、好ましくは、磁気記録ディスクとして特徴づけられる媒体１０６を制御可能に回転させる。
【００１２】
制御可能に移動可能なアクチュエータ１０８はボイスコイルモータ（ＶＣＭ）１１２へ電流を印加することにより、媒体表面上に規定されたトラックに隣接するリード／ライト・トランスデューサ１１０のアレイを移動させる。フレックス回路アセンブリ１１４はアクチュエータ１０８および外部搭載印刷回路板（ＰＣＢ）１１６上の装置制御エレクトロニクス間の電気通信パスを提供する。
【００１３】
図２はｎ個の記憶装置（ＳＤ）１００を有利に内蔵して統合記憶アレイ１２２を形成する典型的なネットワークシステム１２０を一般的に示している。冗長コントローラ１２４，１２６が好ましくは記憶アレイ１２２およびサーバ１２８間でデータを転送するように動作する。サーバ１２８はローカルエリアネットワーク（ＬＡＮ）、インターネット、等のファブリック１３０に接続されている。
【００１４】

遠隔ユーザは、それぞれ、パーソナルコンピュータ（ＰＣ）１３２，１３４，１３６を介してファブリック１３０にアクセスする。このようにして、選択されたユーザは記憶スペース１２２にアクセスして、所望により、データを書込みまたは検索することができる。

【００１５】

好ましくは、装置１００およびコントローラ１２４，１２６はマルチデバイス・アレイ（ＭＤＡ）１３８に内蔵される。好ましくは、ＭＤＡは１つ以上の選択されたＲＡＩＤ（ｒｅｄｕｎｄａｎｔ ａｒｒａｙ ｏｆ ｉｎｄｅｐｅｎｄｅｎｔ ｄｉｓｃｓ）構成を使用して装置１００に亘ってデータを格納する。図２には１つだけのＭＤＡと３人の遠隔ユーザしか例示されていないが、これは単なる説明が目的であって制約的意味合いは無く、所望により、ネットワークシステム１２０は任意数およびタイプのＭＤＡ、サーバ、クライアントおよびホスト装置、ファブリック構成およびプロトコル、等を利用することができる。図３は図２のネットワーク内で有用なアレイコントローラ構成１４０を示している。

【００１６】

図３は中間バス１４６（「Ｅバス」と呼ばれる）により接続された２つのインテリジェント記憶プロセッサ（ＩＳＰ）１４２，１４４を説明する。好ましくは、各ＩＳＰ１４２，１４４は共通コントローラボード上の別個の集積回路パッケージ内に配置される。好ましくは、各ＩＳＰ１４２，１４４はファイバチャネル・サーバリンク１４８，１５０を介してアップストリーム・アプリケーション・サーバと通信し、かつファイバチャネル・記憶リンク１５２，１５４を介して記憶装置１００と通信する。

【００１７】

ポリシープロセッサ１５６，１５８はコントローラ１４０に対するリアルタイム・オペレーティングシステム（ＲＴＯＳ）を実行し、ＰＣＩバス１６０，１６２を介して各ＩＳＰ１４２，１４４と通信する。ポリシープロセッサ１５６，１５８は、さらに、カスタマイズされた論理を実行して定められた記憶アプリケーションに対してＩＳＰ１４２，１４４と共に精巧な処理タスクを実施する。所望により、ＩＳＰ１４２，１４４およびポリシープロセッサ１５６，１５８は動作中に必要とされるメモリモジュール１６４，１６６にアクセスする。

【００１８】

図４は図３の選択されたＩＳＰに対する好ましい構造を提供する。纏めて１６８に示すいくつかのファンクション・コントローラはホスト交換、直接メモリアクセス（ＤＭＡ）、排他的ｏｒ（ＸＯＲ）、コマンドルーティング、メタデータ制御、およびディスク交換等のいくつかのコントローラ動作に対するファンクション・コントローラ・コア（ＦＣＣ）として機能する。好ましくは、各ＦＣＣはメモリ交換その他のスケジューリングタスクを容易にする非常に柔軟な機能セットおよびインターフェイスを含んでいる。

【００１９】

一般的に１７０に示すいくつかのリストマネージャは、キャッシュテーブル管理、メタデータ保守、およびバッファ管理等のコントローラ動作中にさまざまなデータおよびメモリ管理タスクに使用される。好ましくは、リストマネージャ１７０はメモリ上の単純ではあるが明確に定義された操作を実施してＦＣＣ１６８により指令されるタスクを遂行する。好ましくは、各リストマネージャはＦＣＣによるメモリアクセスに対するメッセージプロセッサとして作動して、好ましくは、規定されたプロトコルに従って受信メッセージにより規定される操作を実行する。

【００２０】

リストマネージャ１７０は、それぞれ、交換メモリブロック１７２、キャッシュテーブル・ブロック１７４、バッファメモリ・ブロック１７６およびＳＲＡＭ１７８を含むいくつかのメモリモジュールと通信しかつ制御する。ファンクション・コントローラ１６８およびリストマネージャ１７０は、それぞれ、クロスポイントスイッチ（ＣＰＳ）モジュール１８０を介して通信する。このようにして、コントローラ１６８の選択されたファンク

ション・コアはＣＰＳ１８０を介した対応するリストマネージャ１７０への通信パスウェイを確立して状態を通信し、メモリモジュールをアクセスし、あるいは所望のＩＳＰ動作を呼び出すことができる。

【００２１】

同様に、選択されたリストマネージャ１７０はＣＰＳ１８０を介してファンクション・コントローラ１６８へ応答を通信し返すことができる。図示されてはいないが、好ましくは、図４の各エレメント間にその間のデータ転送を調整するために別個のデータバス接続が確立される。所望により、他の構成を容易に利用できることがお判りであろう。

【００２２】

ＰＣＩインターフェイス（Ｉ／Ｆ）モジュール１８２はポリシープロセッサ１５６およびＩＳＰ１４２間でトランザクションを確立して指示する。Ｅ－ＢＵＳＩ／Ｆモジュール１８４はＦＣＣおよび各ＩＳＰ１４２，１４４のリストマネージャ間のＥ－ＢＵＳ１４６を介した通信を容易にする。所望により、ポリシープロセッサ１５６，１５８はＥ－ＢＵＳ１４６を介したシステムの他の部分との通信を開始し受信することもできる。

【００２３】

図３および４のコントローラ・アーキテクチュアはアレイに対するスケーラブルで、非常に機能的なデータ管理および制御を有利に提供する。好ましくは、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は記録媒体上のストライプ境界および記憶トランザクション中にディスクストライプに関連付けられるデータを専用に格納するキャッシュ内の基準データバッファに揃えられる。

【００２４】

処理効率をさらに高めるために、好ましくは、コントローラ・アーキテクチュアは新しいライトバックデータ・キャッシング方法論を利用する。一般的に、これはメモリ内の記憶装置１００に書き込まれるデータのキャッシング、および後でこのようなライトバックデータの記憶装置１００への転送をスケジューリングする（フラッシング）ことを含んでいる。

【００２５】

一般的に、データの時間および局所性の両方を考慮する２次元方法を使用してライトバックデータの隣接ブロックのセットがキャッシュメモリ内に配列される。キャッシュされたライトバックデータの実質的に均一な分布が維持されて、他のアクセス操作と共にデータを書き込むための最適機会を提供する。

【００２６】

好ましくは、データの隣接ブロックのセットは付加ライトバックデータセットがホストによりキャッシュメモリへ提供されるレートと実質的に一致するレートでキャッシュメモリから記憶アレイへ書き込まれる。このようにして、観察されたホストＩ／Ｏ転送レートの大きな変動は実質的に解消される。

【００２７】

図５に示すように、好ましくは、キャッシュされたデータはストライプ・データ記述子（ＳＤＤ）１９２と呼ばれるデータ構造を使用してキャッシュ・マネージャ（ＣＭ）１９０によりノードベースで管理される。各ＳＤＤは関連するデータへの最近および現在のアクセスに関するデータを保持する。各ＳＤＤは好ましくは対応するＲＡＩＤストライプ１９４（すなわち、特定のパリティセットに関連する選択された装置１００上の全データ）に揃えられ、特定のＳＢＬ１９６に従う。

【００２８】

ＣＭ１９０により管理される各キャッシュノードは、好ましくは、ある特定のＳＤＤを参照し、論理ディスクの定められたセット（装置１００のサブセット）に対するアクティブＳＤＤは、好ましくは、標準フォワードおよびバックワード・リンクリストを使用して仮想ブロックアドレス（ＶＢＡ）を介して昇ベキの順でリンクされる。

【００２９】

好ましくは、ＶＢＡ値はしばしばＲＡＩＤ Ａｌｌｏｃａｔｉｏｎ ＧｒｉｄＳｙｓｔ

ｅｍ（ＲＡＧＳ）と呼ばれるグリッドシステムを使用してＲＡＩＤデータ編成と揃えられる。一般的に、同じＲＡＩＤストリップ１９８に属するブロックの任意特定の集り（たとえば、特定のパリティセットに寄与する全データ）が特定のシート上の特定の信頼できる記憶ユニット（ＲＳＵ）に割当てられる。

【００３０】

数枚のシートからなるブックが異なる装置１００からのブロックの多数の隣接セットから構成される。実際のシートおよびＶＢＡに基づいて、ブックは特定の装置または装置セットを示す（冗長度が利用される場合）ゾーンへ細分割される。

【００３１】

好ましくは、各ＳＤＤはアクセス歴、ロック状態、最終オフセット、最終ブロック、タイムスタンプデータ（一日の時間、ＴＯＤ）、データがどのゾーン（ブック）に属するかの識別子、および利用されるＲＡＩＤレベルを含むデータのさまざまな状態を示す変数を含んでいる。好ましくは、ＳＤＤに関連するデータのライトバック（「ダーティ」データ）状態がダーティデータ、ダーティバッファ、ダーティＬＲＵおよびフラッシングＬＲＵ値に関して管理される。

【００３２】

好ましくは、ＣＭ１９０は同時に作動していくつかの異なるレベルのライトバックデータ・プロセスをシステム要求条件に応じて管理する。第１のレベルは一般的にフルＲＡＩＤストリップ１９８が検出される時のフルＳＤＤ構造の周期的フラッシングを含んでいる。これはＳＤＤが関連するデータをダーティとして識別する時にＲＡＩＤレベル変数に基づいて定められたＳＤＤ１９２に対して容易に実施することができる。好ましくは、これは十分連続的な隣接ＳＤＤ構造がダーティデータで十分満たされているかを決定するバックワード検査を含んでいる。そうであれば、ＳＤＤ構造はフラッシングリスト（１９９に示す）上に置かれ、データのフラッシングを開始する要求がなされる。フラッシングリスト状態はＳＤＤ１９２のフラッシングＬＲＵ値を使用して設定することができる。

【００３３】

データの小さなセットのフラッシングは好ましくはＳＤＤベースで処理される。ダーティブロックを有しロックブロックの無い任意のＳＤＤは好ましくはダーティＬＲＵとして設定され経過期間（ａｇｅ）（たとえば、データがキャッシュ待機フラッシングにおいて過ごした時間）によりソートされる。特定の経過期間に達すると、好ましくはフラッシングＬＲＵ変数が設定されフラッシングリスト１９９が更新される。

【００３４】

好ましくは、フラッシングリストからのデータのフラッシングの攻撃性は実質的に付加ダーティデータがキャッシュ内に入るレートでダーティデータを押出すように適応的に調節される。特定範囲の連続的ダーティブロックをフラッシングする予定の時は、ＣＭ１９０は好ましくは近似局所性を有するＲＡＩＤレベルに基づいて他の範囲のダーティブロック、すなわち、シーク時間に関して「すぐ近い」または同じＲＡＩＤパリティストリップ１９９へのアクセスを伴うブロックを捜し出す。

【００３５】

コントローラアーキテクチュアがこれを実施する好ましい方法を図６について見ることができ、それはキャッシュされたライトバックデータのアレイ２００を表している。アレイ２００はＣＭ１９０またはコントローラの他の処理ブロックにより維持される。

【００３６】

アレイ２００内の各セル２０２は一般的に記憶装置１００内の選択された局所性に対応し、ブック、シートおよび／またはアレイ内のゾーンとして編成することができる。装置内の境界は、たとえば、各コラムが異なる記憶装置１００を表しコラム内の各セルはその装置を横切る異なる半径方向バンドを表すように選択することができる。

【００３７】

セルはフラッシングリスト１９９にフラグを付されている隣接ライトバックデータのセットで「占用されている（ｐｏｐｕｌａｔｅｄ）」。特に、各占用ブロック２０４（図６

に「Ｘ」で示す）は記憶装置１００内の異なる論理的または物理的場所に対応する変動するサイズのデータブロックの１つの（またはいくつかの）グループを表す。このようにして、キャッシュ内のライトバックデータのセットは記憶アレイ１２２へ転送されることを予測してプールすることができる。

【００３８】

アレイ２００はさまざまなデータ装置１００にわたってデータのライトバックをスケジューリングする有用なフォーマットを提供する。好ましい実施例では、記憶装置１００内の選択された場所にわたってアクセスする（たとえば、リードコマンドを実施する）ように特定のアクセスコマンドがスケジュールされると、アレイ２００はアクセスコマンドと共に効率的にサービスされるライトバックデータの利用可能なブロックを識別するように参照される。

【００３９】

アレイ２００を横切って占用セル２０４の実質的に均一な分布を維持し、着信ダーティデータのレートをキャッシュと整合させるためにダーティセットが選択的にアレイ２００に加えられる。あるシステム負荷の元で、ＣＭ１９０は比較的多数のフラッシング操作をロードアップして、論理的セットまたは物理的媒体を横切ってＩＤからＯＤへ進行するライト等の、ショートシークのクラスタを生成するように構成することができる。

【００４０】

関連する実施例では、図６のアレイ２００は個別のＲＡＩＤストライプ（対応するＳＤＤ１９２を有する図５の１９４等）を表すようにセル２０２を配置するように構成することができ、アレイ２００のコラムは前記したＲＡＧＳグリッドのコラムに対応することができる。このような場合、定められたロー内の特定の占用セル２０４のフラッシングを使用して同じロー内の他の占用セルにより使用されるパリティ（グリッド内には示されていない）への参照を示すことができる。

【００４１】

このようなフラッシング操作の同時スケジューリングにより、特に、ＲＡＩＤ－５およびＲＡＩＤ－６環境において性能改善を導くことができ、それはそのロー内の全占用セルに対する４つの（ＲＡＩＤ－６の場合は６つの）Ｉ／Ｏアクセスの内の２つが同じパリティＲＡＩＤストライプ１９４にアクセスするためである。

【００４２】

もう１つの好ましい実施例では、アレイ２００は各コラム（またはロー）が別個の記憶装置１００を表し、各セル２０２は一般的に装置の媒体１０６の半径方向幅にわたって異なる隣接ゾーン、すなわち領域、に対応するように配置される。

【００４３】

図７の「W」表記は一般的にこれらのさまざまな場所内のデータの未決ライトバックセットに対応し、したがって各セル２０２内の装置にわたって分布されるライト機会を表す。各セル２０２内のW表記の空間的場所は一般的に関連する領域内のそのデータの論理的または物理的場所を表す。W表記は媒体１０６へ書き込むためにフラッシングリスト１９９から利用可能な全ライトバックデータセットを必ずしも表すものではない。

【００４４】

リード優先環境では、リードコマンドが優先権を有し、したがって一般的にライトコマンドのサービスよりも優先的にサービスされる。しかしながら、ライト優勢環境では、一般的にリードに較べて比較的多数のライトコマンドがある。1つのこのようなリードコマンドは図７において「Ｒ」表記で識別され、一般的に関連するデータが検索される媒体１０６の領域に対応する。

【００４５】

この実施例では、キャッシュ・マネージャ１９０は好ましくは関連するリードコマンドを実施してＲ表記のデータを検索するよう装置１００に指令する。このリード操作の終りに、キャッシュ・マネージャ１９０は、さらに、好ましくは一般的にリードコマンドの近くの（たとえば、同じセル２０２内の）１つ以上ライトバック操作を装置１００に実施さ

せるように進行する。

【００４６】

図７は「丸Ｗ」表記を使用してこのようにサービスされる２つのこのようなライトバックデータセットを識別する、すなわち、２つの丸Ｗ表記ライトバックフラッシュが関連するリードコマンドの終りに生じる（Ｒ表記）。リードコマンド近くの（たとえば、セル２０２内の）全ライトバックデータがフラッシュされる必要はなく、望ましくないことさえあることに注目願いたい。しかしながら、近くのデータの少なくともいくつかはフラッシュされ、装置１００のトランスデューサ１１０は一般的にこの近くにあるため、これらのライトバックデータ・フラッシュ操作は低減されたシークレーテンシ（ｓｅｅｋ　ｌａｔｅｎｃｉｅｓ）で実施することができる。

【００４７】

選択されたライトバックデータがフラッシュされると、キャッシュ・マネージャ１９０は、フラッシュリストから得られる、この同じ領域への付加ライトバックデータ・セットでアレイ２００を「埋め戻す」ように進行する。このようにして、新しいライトコマンドは媒体１０６のさまざまな半径方向幅にわたってライトバックデータ機会の均一な分布を実質的に維持するようにアレイ２００に対して計られる（ｍｅｔｅｒｅｄ）。

【００４８】

ディスクアクセスを要する未決リードコマンドが無い限り、キャッシュ・マネージャ１９０は一般的に前と同様にライトバックデータをフラッシュするように動作する。しかしながら、各新しいリードコマンドが発せられると、リードコマンドに優先権が与えられ１つ以上の付加ライトバックセットがリードコマンドの一般的に近くからフラッシュされる。

【００４９】

好ましくは、これは次のリードコマンドがどこへ向けられても効率的方法でフラッシュすることができる１つ以上のライトバックデータセットがその近くにある環境を提供する。好ましい実施例では、キャッシュ・マネージャ１９０は、任意の定められた時間に未決の３０を超えるＷおよび２を超えるＲが無いように（待ち行列コマンドに対する合計３２の「スロット」に対して）、各装置１００に対するアレイ２００内の「Ｒ」に対する「Ｗ」の選択された比率を維持するように動作する。しかしながら、他の比率も容易に使用することができる。比率はキャッシュ・マネージャ１９０により経験されるリード／ライト・コマンドミックス内のバースト変化に関して時間と共に調節することもできる。

【００５０】

図８はライトバックデータ・フラッシング・ルーチン３００を説明し、一般的に本発明の好ましい実施例に従って実施されるステップを表す。

【００５１】

システムはステップ３０２において初期構成される。好ましくは、これはフラッシングリスト１９９に対するさまざまな境界および記憶装置１００の物理的領域をカバーする対応するアレイ２００の初期識別を含む。所望により、記憶スペースの適切なサブセットに対する異なるアレイおよびリストを維持することができ、あるいは単一の統合リスト／アレイを維持することができる。

【００５２】

正規のシステム操作が次に開始され、ステップ３０４に示すように、これはキャッシュメモリへのライトバック（ダーティ）データの周期的提供を含む。このようなライトバックデータは主として図１のＰＣ１３２，１３４，１３６等のホストからのデータライト操作から生じるものと考えられ、その場合、コントローラ１２４は好ましくはライトバックデータを選択されたキャッシュ場所（図４の１７６等）に格納し、ライト完了信号を開始装置へ返す。しかしながら、ライトバックデータはシステム状態データ、選択されたメモリバックアップ、メタデータ、等の内部発生ライトとすることもできる。

【００５３】

ステップ３０６に示すように、ＳＤＤ１９２は好ましくは関連するライトバックデータ

に対して更新される。ダーティデータおよびダーティバッファ値はデータをダーティとして識別するように初期設定することができる。その後、データセットはフルストリップ１９８とのデータの関係、経過期間、および付加データのキャッシュ内への進入レートを含む前記したいくつかの要因に関してフラッシングリスト１９９へ移される。アレイ２００は前記ブロックのフラッシングリスト１９９への移動に応答してフラッシングに利用できる隣接データブロックのセットを識別するようにステップ３０８において対応的に占用される。

【００５４】

ステップ３１０において、ライトバックデータの選択されたセットは記憶装置１００へフラッシュされる。好ましくは、これは他の近接アクセス操作と共に生じるが、装置１００にわたってより大きい逐次フラッシング操作をスケジュールすることもできる。前記したように、ＣＭ１９０または他のプロセスは好ましくはキャッシュされたライトバックデータがさらにキャッシュメモリへ導入されるレートに関して利用できるライトバックデータブロックの実質的に均一な分布を維持するように動作する。

【００５５】

ここで検討されたさまざまな好ましい実施例は従来技術に優る利点を提供する。開示された方法論は時間と局所性の両方がフラッシング・アルゴリズムの要因として記憶装置１００へのデータの効率的フラッシングを提供する点においてダイナミックである。さらに、方法論は付加ダーティデータがキャッシュメモリへ導入されるレートに実質的に揃えるのに適応的である。好ましい実施例では、キャッシュメモリ内の付加ダーティデータのセットはアレイ２００へかつアレイ２００から記憶ディスクへ選択的に計られて実質的に平坦な負荷量を維持する。このようにして、ホストＩ／Ｏ内の著しい変動が回避される。

【００５６】

ここに提示された好ましい実施例は複数のディスクドライブ記憶装置を利用するマルチデバイスアレイに向けられているが、それは単なる説明の目的であって制約的意味合いはない。むしろ、特許請求される発明は任意数のさまざまな環境で利用して効率的なデータ処理を促進させることができる。

【００５７】

本発明のさまざまな実施例の非常に多くの特徴および利点を、本発明のさまざまな実施例の構造および機能の詳細と共に、前記した明細書に記載してきたが、この詳細な説明は説明用にすぎず詳細、特に、部品の構造および配置に関して本発明の原理内で添付特許請求の範囲が表現される用語の広範な一般的意味により示される限界まで変更を行うことができる。たとえば、特定のエレメントは特定の応用に応じて本発明の精神および範囲を逸脱することなく変動することができる。

【図面の簡単な説明】

【００５８】

【図１】本発明の好ましい実施例に従って構成かつ作動される記憶装置を一般的に示す図である。

【図２】図１に示すようないくつかの記憶装置を利用するネットワークシステムの機能的ブロック図である。

【図３】図２のコントローラの好ましいアーキテクチュアの一般的表現を示す図である。

【図４】図３の選択されたインテリジェント記憶プロセッサの機能的ブロック図である。

【図５】好ましい実施例に従って記憶アレイにデータをフラッシュするように動作するキャッシュ・マネージャを一般的に示す図である。

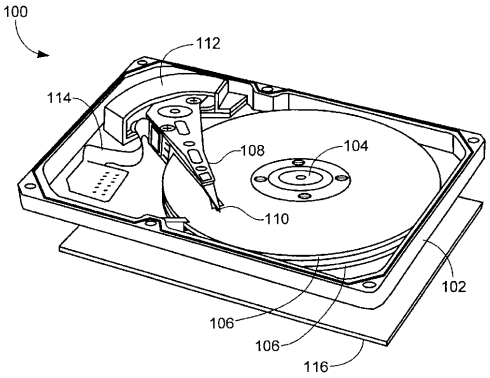
【図６】好ましい実施例に従って記憶アレイのいくつかの異なる場所にわたってライトバックデータ機会の分布を提供する図５のキャッシュ・マネージャにより維持されるライトバックデータセットのアレイを表す図である。

【図７】もう１つの好ましい実施例に従った図６のアレイの一部を示す図である。

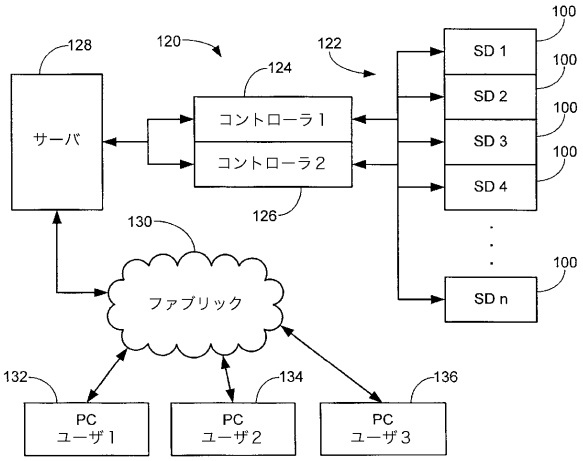
【図８】好ましい実施例に従って実施されるステップを示すライトバックデータ・フラッシング・ルーチンに対するフロー図である。

【符号の説明】
【００５９】
１００　記憶装置
１０２　ベースデッキ
１０４　スピンドルモータ
１０６　回転媒体
１０８　アクチュエータ
１１０　リード／ライト・トランスデューサ
１１２　ボイスコイルモータ
１１４　フレックス回路アセンブリ
１１６　印刷回路板
１２０　ネットワーク・システム
１２２　記録アレイ
１２４，１２６　冗長コントローラ
１２８　サーバ
１３０　ファブリック
１３２，１３４，１３６　パーソナルコンピュータ
１４０　アレイコントローラ構成
１４２，１４４　インテリジェント記憶プロセッサ
１４６　中間バス
１４８，１５０　ファイバチャネル・サーバリンク
１５２，１５４　ファイバチャネル記憶リンク
１５６，１５８　ポリシー・プロセッサ
１６０，１６２　ＰＣＩバス
１６４，１６６　アクセスメモリ・モジュール
１６８　ファンクション・コントローラ
１７０　リストマネージャ
１７２　交換メモリブロック
１７４　キャッシュテーブル・ブロック
１７６　バッファメモリ・ブロック
１７８　ＳＲＡＭ
１８０　クロスポイントスイッチ・モジュール
１８２　ＰＣＩインターフェイス・モジュール
１８４　Ｅ－ＢＵＳ　ＩＦモジュール
１９０　キャッシュ・マネージャ
１９２　ストライプ・データ記述子
１９４，１９８　ＲＡＩＤストライプ
１９６　ＳＢＬ
１９９　フラッシングリスト
２００　アレイ
２０２，２０４　セル

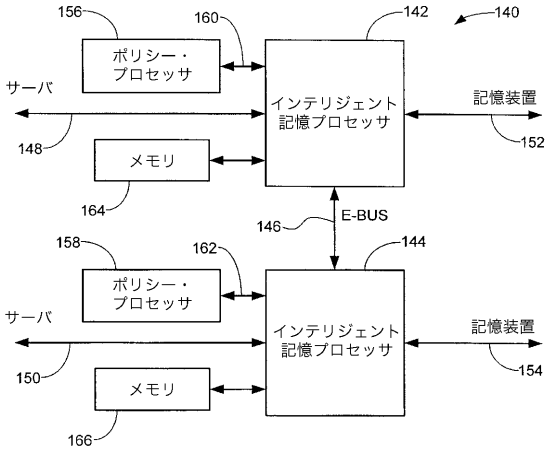
【図１】



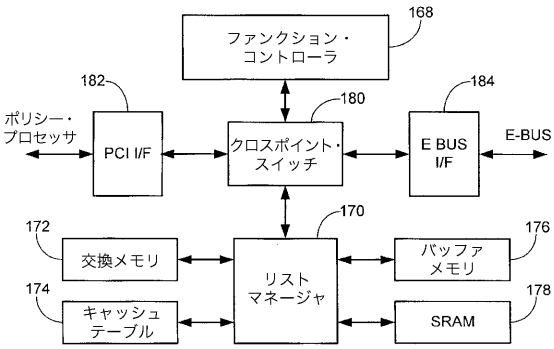
【図２】



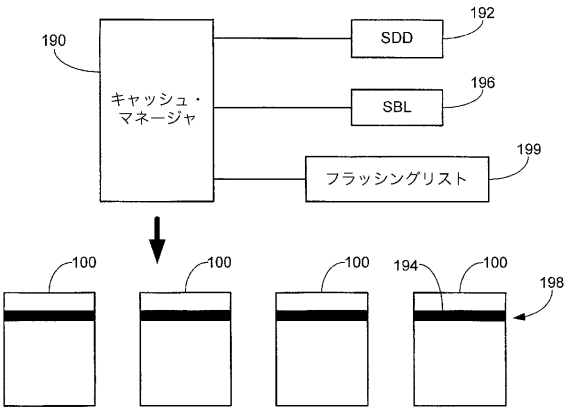
【図３】



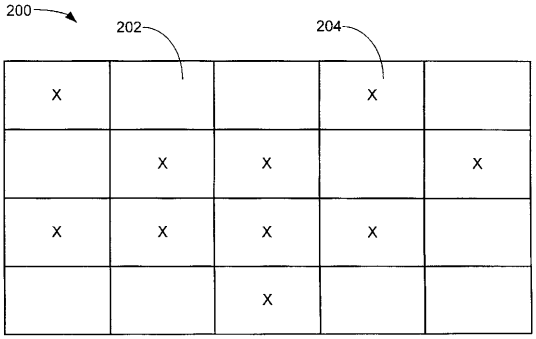
【図４】



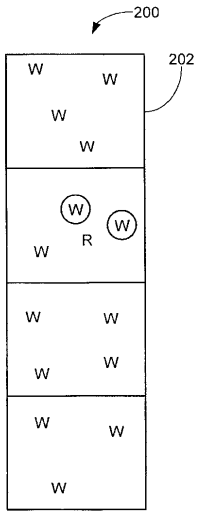
【図５】



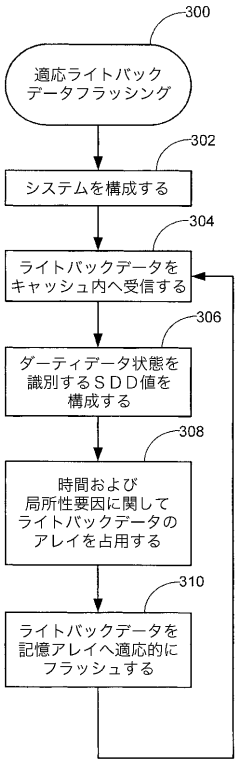
【図６】



【図７】



【図８】


300
適応ライトバック
データフラッシング
302
システムを構成する
304
ライトバックデータを
キャッシュ内へ受信する
306
ダーティデータ状態を
識別するＳＤＤ値を
構成する
308
時間および
局所性要因に関して
ライトバックデータの
アレイを占用する
310
ライトバックデータを
記憶アレイへ適応的に
フラッシュする

フロントページの続き

(72)発明者　クラーク　イー．レッベルス
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　バレー　ロード　５３０１
(72)発明者　マイケル　ディー．ウォーカー
アメリカ合衆国、コロラド、コロラドスプリングス、　イースト　ウィラメッド　２１９
(72)発明者　デーヴィッド　ピー．デセンゾ
アメリカ合衆国、コロラド、プエブロ、　ミッドナイト　アベニュー　３１９

Ｆターム(参考) 5B005 JJ12　MM11　NN02　VV01
5B065 CA30　CH01

【外国語明細書】
2008016024000001.pdf