(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7513356号
(P7513356)

(45)発行日　令和6年7月9日(2024.7.9)

(24)登録日　令和6年7月1日(2024.7.1)

(51)国際特許分類

G06F 12/00 (2006.01)	G06F	12/00	550Z
G06F 3/06 (2006.01)	G06F	3/06	301Y
G06F 12/02 (2006.01)	G06F	3/06	306Z
	G06F	12/02	510A
	G06F	12/00	597U

請求項の数　9　（全21頁）

(21)出願番号　特願2021-571287(P2021-571287)
(86)(22)出願日　令和2年5月28日(2020.5.28)
(65)公表番号　特表2022-536056(P2022-536056A)
(43)公表日　令和4年8月12日(2022.8.12)
(86)国際出願番号　PCT/IB2020/055061
(87)国際公開番号　WO2020/250071
(87)国際公開日　令和2年12月17日(2020.12.17)
審査請求日　令和4年10月21日(2022.10.21)
(31)優先権主張番号　16/439,122
(32)優先日　令和1年6月12日(2019.6.12)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
INTERNATIONAL BUSINESS MACHINES CORPORATION
アメリカ合衆国１０５０４ ニューヨーク州 アーモンク ニュー オーチャード ロード
New Orchard Road, Armonk, New York 10504, United States of America
(74)代理人　100112690
弁理士　太佐 種一

最終頁に続く

(54)【発明の名称】　不揮発性メモリにおける摩耗を意識したブロック・モード変換

(57)【特許請求の範囲】

【請求項１】

不揮発性メモリを動作させる方法であって、前記不揮発性メモリは複数の物理ブロックを含み、各物理ブロックはそれぞれの複数のセルを含み、前記複数のセルのそれぞれは個々に複数ビットのデータを記憶することが可能であり、前記方法は、

コントローラが、前記複数の物理ブロックのうちの物理ブロックを、セルあたりより多くのビットを記憶するための第１のモードで動作する物理ブロックを含む第１のプールに割り当て、前記複数の物理ブロックのうちの他の物理ブロックを、セルあたりより少ないビットを記憶するための第２のモードで動作する物理ブロックを含む第２のプールに割り当てることと、

前記コントローラが、前記第１のプールおよび前記第２のプールの間で物理ブロックを転送することと、

を含み、

前記転送することは、

前記第１のプールから前記第２のプールに、より高いビット・エラー・レートを有する物理ブロックを転送することと、

前記第２のプールから前記第１のプールに、より低いビット・エラー・レートを有する物理ブロックを転送することと、

を含む、方法。

【請求項２】

不揮発性メモリを動作させる方法であって、前記不揮発性メモリは複数の物理ブロックを含み、各物理ブロックはそれぞれの複数のセルを含み、前記複数のセルのそれぞれは個々に複数ビットのデータを記憶することが可能であり、前記方法は、
コントローラが、前記複数の物理ブロックのうちの物理ブロックを、セルあたりより多くのビットを記憶するための第１のモードで動作する物理ブロックを含む第１のプールに割り当て、前記複数の物理ブロックのうちの他の物理ブロックを、セルあたりより少ないビットを記憶するための第２のモードで動作する物理ブロックを含む第２のプールに割り当てることと、
前記第１および第２のプールのそれぞれについて、
異なる書き込みヒートの複数の書き込みストリームのそれぞれのデータを記憶するために割り当てられた物理ブロックの数と、前記複数の書き込みストリームのそれぞれにおけるプログラム／消去（Ｐ／Ｅ）サイクル数とを追跡することと、
前記複数の書き込みストリームのそれぞれについて、前記書き込みストリームに割り当てられた物理ブロックの数と、前記書き込みストリームに割り当てられた前記物理ブロックのＰ／Ｅサイクルの総数とに基づいて、それぞれのストリーム摩耗指標を計算することと、
前記コントローラが、前記第１のプールおよび前記第２のプールの間で物理ブロックを転送することと、
を含み、
前記転送することは、（１）前記複数の書き込みストリームの中から選択された書き込みストリームと、（２）前記選択された書き込みストリームの前記ストリーム摩耗指標と、（３）前記選択された書き込みストリームのデータを記憶するために割り当てられた物理ブロックのブロック・ヘルスとに基づいて、転送用に物理ブロックを選択することを含む、
方法。
【請求項３】
前記コントローラが、前記複数の書き込みストリームの全てにおける物理ブロックのブロック・ヘルスを、前記物理ブロックをプログラミングした直後に測定されるビット・エラー・レートから決定することをさらに含む、請求項２に記載の方法。
【請求項４】
前記選択することは、
前記選択された書き込みストリームが前記複数の書き込みストリーム全ての中で最大の摩耗を反映するストリーム摩耗指標を有することに基づいて、前記選択された書き込みストリームに割り当てられた前記物理ブロックの中から最も良好なヘルスの物理ブロックを転送用に選択することと、
前記選択された書き込みストリームが前記複数の書き込みストリーム全ての中で最小の摩耗を反映するストリーム摩耗指標を有することに基づいて、前記選択された書き込みストリームに割り当てられた前記物理ブロックの中から最も不良なヘルスの物理ブロックを転送用に選択することと、
を含む、請求項２に記載の方法。
【請求項５】
前記第１のモードはクアッド・レベル・セル（ＱＬＣ）モードであり、
前記第２のモードはシングル・レベル・セル（ＳＬＣ）モードである、
請求項１または２に記載の方法。
【請求項６】
前記転送することは、前記第１および第２のプールのうちの所与の１つから利用可能なプログラムされていない物理ブロックの数が閾値を満たさないことに基づいて、前記物理ブロックを転送することを含む、請求項１または２に記載の方法。
【請求項７】
不揮発性メモリ用のコントローラを備えるデータ・ストレージ・システムであって、前

記不揮発性メモリは複数の物理ブロックを含み、各物理ブロックはそれぞれの複数のセルを含み、前記複数のセルのそれぞれは個々に複数ビットのデータを記憶することが可能であり、前記コントローラは、
前記複数の物理ブロックのうちの物理ブロックを、セルあたりより多くのビットを記憶するための第１のモードで動作する物理ブロックを含む第１のプールに割り当て、前記複数の物理ブロックのうちの他の物理ブロックを、セルあたりより少ないビットを記憶するための第２のモードで動作する物理ブロックを含む第２のプールに割り当てることと、
前記第１のプールおよび前記第２のプールの間で物理ブロックを転送することと、
を実行するように構成され、
前記転送することは、
前記第１のプールから前記第２のプールに、より高いビット・エラー・レートを有する物理ブロックを転送することと、
前記第２のプールから前記第１のプールに、より低いビット・エラー・レートを有する物理ブロックを転送することと、
を含む、データ・ストレージ・システム。
【請求項８】
不揮発性メモリ用のコントローラを備えるデータ・ストレージ・システムであって、前記不揮発性メモリは複数の物理ブロックを含み、各物理ブロックはそれぞれの複数のセルを含み、前記複数のセルのそれぞれは個々に複数ビットのデータを記憶することが可能であり、前記コントローラは、
前記複数の物理ブロックのうちの物理ブロックを、セルあたりより多くのビットを記憶するための第１のモードで動作する物理ブロックを含む第１のプールに割り当て、前記複数の物理ブロックのうちの他の物理ブロックを、セルあたりより少ないビットを記憶するための第２のモードで動作する物理ブロックを含む第２のプールに割り当てることと、
前記第１および第２のプールのそれぞれについて、
異なる書き込みヒートの複数の書き込みストリームのそれぞれのデータを記憶するために割り当てられた物理ブロックの数と、前記複数の書き込みストリームのそれぞれにおけるプログラム／消去（Ｐ／Ｅ）サイクル数とを追跡することと、
前記複数の書き込みストリームのそれぞれについて、前記書き込みストリームに割り当てられた物理ブロックの数と、前記書き込みストリームに割り当てられた前記物理ブロックのＰ／Ｅサイクルの総数とに基づいて、それぞれのストリーム摩耗指標を計算することと、
前記第１のプールおよび前記第２のプールの間で物理ブロックを転送することと、
を実行するように構成され、
前記転送することは、（１）前記複数の書き込みストリームの中から選択された書き込みストリームと、（２）前記選択された書き込みストリームの前記ストリーム摩耗指標と、（３）前記選択された書き込みストリームのデータを記憶するために割り当てられた物理ブロックのブロック・ヘルスとに基づいて、転送用に物理ブロックを選択することを含む、
データ・ストレージ・システム。
【請求項９】
プロセッサに、請求項１ないし６のいずれか一項に記載の方法を実行させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明はデータ処理およびデータ記憶に関し、より詳細には、ブロックが、異なるセルあたりのビット記憶容量を提供する複数のブロック・モードで動作することが可能な不揮発性メモリ・システムに関する。さらにより詳細には、本発明は、異なるブロック・モード間での不揮発性メモリ・システム内の物理記憶ブロックの摩耗（wear）を意識した変換

に関する。
【背景技術】
【０００２】
ＮＡＮＤフラッシュ・メモリは、メモリ・セルあたり１ビット以上のデータをトランジスタのフローティング・ゲートまたは同様の電荷捕獲構造上の電荷として記憶する、電気的にプログラム可能で消去可能な不揮発性メモリ・テクノロジーである。フローティング・ゲート上の電荷量により、トランジスタの閾値電圧が変化する。適切な読み出し電圧を印加して電流量を測定することにより、メモリ・セルのプログラムされた閾値電圧を決定することができ、ひいては記憶された情報を検出することができる。典型的な実装形態では、ＮＡＮＤフラッシュ・メモリ・アレイは、物理メモリのブロック（「消去ブロック」とも呼ばれる）単位で構造化され、各ブロックは複数の物理ページを含み、各物理ページは多数のメモリ・セルを含む。メモリ・セルにアクセスするために利用されるワード線およびビット線の配置によって、フラッシュ・メモリ・アレイは一般にページ単位でプログラムされるが、ブロック単位で消去される。
【０００３】
フラッシュ・ストレージ・テクノロジーは、フラッシュ・ストレージのギガバイト単価を下げる目的に牽引されて、過去１０年間でビット記憶密度の目覚ましい成長を遂げた。このトレンドの主要な推進力の１つは、フラッシュ・セルあたりに記憶されるビット数が着実に増加していることである。たとえば、フラッシュ・セルあたりに記憶することができるビット数は、１（シングル・レベル・セル（ＳＬＣ））から２（マルチ・レベル・セル（ＭＬＣ））、３（３レベル・セル（ＴＬＣ：Three Level Cell））へ、最近では４（クアッド・レベル・セル（ＱＬＣ））まで着実に増加している。フラッシュ・セルあたりに記憶することができるビット数の増加は、セルあたりのビット数の増加に伴って耐久性およびパフォーマンスが低下するという代償を払っている。
【０００４】
エンタープライズ・クラスのフラッシュ・コントローラは、新世代のＮＡＮＤフラッシュ・メモリの耐久性およびパフォーマンスの制限を改善するために様々な技術を利用する。たとえば、フラッシュ・コントローラは１つまたは複数のフラッシュ管理技術を利用し得、これには、高度な誤り訂正符号の実装、ビット・エラー・レートを低減させるための電圧レベルの調整、書き込みヒート（heat）に基づいてデータを分別することによる内部のガベージ・コレクションのオーバーヘッドの削減、ブロックに基づくブロックの管理（すなわち、ヘルス・ビニング）、ならびにヘルスを意識したウェア・レベリングおよびデータ配置の実行、が含まれる。これらの高度なフラッシュ管理技術によって寿命およびパフォーマンスが向上しているにもかかわらず、ビット密度が増加し続けるので、既存のフラッシュ管理技術のみでは、許容可能なデバイス寿命（たとえば、重い書き込みワークロードの下でのエンタープライズ・クラスのＳＳＤに対する７年以上の保証期間）を保証することができない。たとえば、ＱＬＣ　ＮＡＮＤフラッシュ・テクノロジーでは、ＴＬＣ　ＮＡＮＤフラッシュ・テクノロジーと比較して記憶密度が３３％向上するが、定格耐用年数はＴＬＣ　ＮＡＮＤフラッシュの数分の１（たとえば、ブロックあたりほんの数千プログラム／消去（Ｐ／Ｅ：program/erase）サイクル）である。
【０００５】
デバイスの耐久性を高めるための有望な方法の１つは、ＮＡＮＤフラッシュ・メモリの複数のモードで動作する機能を利用することである。たとえば、一部のＴＬＣ　ＮＡＮＤフラッシュ・メモリは、特定の選択されたブロックがセルあたり１ビットのみを記憶するＳＬＣモードをサポートすることにより、記憶密度を犠牲にして、選択されたブロックのパフォーマンスおよび耐久性を向上させる。ＳＬＣモードで構成されたブロックは、１～２桁多いＰ／Ｅサイクルを維持することができるが、ＴＬＣモードで構成されたブロックの記憶容量の３分の１しか提供することができない。そのようなシステムでは、頻繁に書き込まれるデータをＳＬＣブロックに記憶することができ、頻繁に書き込まれないデータをネイティブＴＬＣモードで動作するブロックに記憶することができる。

【０００６】
複数の動作モードをサポートする既存のフラッシュ・コントローラは階層型アーキテクチャを採用していることが多く、到来したユーザ書き込みは、最初は低い記憶密度の層（ＳＬＣ層など）のブロックに記憶される。低い記憶密度の層のストレージが使い果たされると（または他の何らかの閾値充填レベルに達すると）、フラッシュ・コントローラは、データを低い記憶密度の層から高い記憶密度の層（たとえば、ＴＬＣ層）に移動させることを、好ましくは将来上書きされる可能性が最も低いデータから開始して行うことにより、低い記憶密度の層のスペースを解放するデステージ処理を開始する。次いで、フラッシュ・コントローラは、高い記憶密度の層に存在するデータの管理を従来の方法で行い、たとえば、定期的にガベージ・コレクション用に選択されたブロックから有効なデータを再配置し、ガベージ・コレクションされたブロックを消去し、その後、低い記憶密度の層からデステージされたデータを消去したブロックに再度書き込むことによって行う。一般に、このタイプの従来の階層型ストレージ・アーキテクチャでは、ブロックは低密度のストレージ階層または高密度のストレージ階層のいずれかに静的に割り当てられる。

【発明の概要】

【０００７】
本発明では、異なるビット・記憶密度のブロックを用いた多層メモリ・ストレージ・システムの様々な層への完全に静的なブロックの割り当てでは、所与の層が早期に記憶容量を使い果たしたり、その耐久限度に達したりし得ることを認識している。この問題に対処するために、不揮発性メモリのブロックを動作モード間で選択的に変換することができる。

【０００８】
本発明の一態様では、不揮発性メモリを動作させる方法であって、不揮発性メモリは複数の物理記憶ブロックを含み、各物理記憶ブロックはそれぞれの複数のセルを含み、複数のセルのそれぞれは個々に複数ビットのデータを記憶することが可能であり、方法は、コントローラが、複数の物理ブロックのうちの物理ブロックを、セルあたりより多くのビットを記憶するための第１のモード（たとえば、ＱＬＣモード）で動作する物理ブロックを含む第１のプールに割り当て、複数の物理ブロックのうちの他の物理ブロックを、セルあたりより少ないビットを記憶するための第２のモード（たとえば、ＳＬＣモード）で動作する物理ブロックを含む第２のプールに割り当てることと、コントローラが、第１のプールおよび第２のプールの間で物理ブロックを転送することであって、転送される物理ブロックについて測定されたビット・エラー・レートに少なくとも基づいて行われる、転送することと、を含む、方法を提供する。

【０００９】
少なくとも１つの実施形態では、コントローラは、第１のプールから第２のプールに、より高いビット・エラー・レートを有する物理ブロックを転送し、第２のプールから第１のプールに、より低いビット・エラー・レートを有する物理ブロックを転送する。

【００１０】
少なくとも１つの実施形態では、コントローラは、第１および第２のプールのそれぞれについて、異なる書き込みヒートの複数の書き込みストリームのそれぞれのデータを記憶するために割り当てられた物理ブロックの数と、複数の書き込みストリームのそれぞれにおけるプログラム／消去（Ｐ／Ｅ）サイクル数とを追跡する。追加的に、コントローラは、複数の書き込みストリームのそれぞれについて、書き込みストリームに割り当てられた物理ブロックの数と、書き込みストリームに割り当てられた物理ブロックのＰ／Ｅサイクルの総数とに基づいて、それぞれのストリーム摩耗指標を計算する。次いで、コントローラは、（１）複数の書き込みストリームの中から選択された書き込みストリームと、（２）選択された書き込みストリームのストリーム摩耗指標と、（３）選択された書き込みストリームのデータを記憶するために割り当てられた物理ブロックのブロック・ヘルスとに基づいて、転送用に物理ブロックを選択することができる。複数の書き込みストリームの全てにおける物理ブロックのブロック・ヘルスを、たとえば、物理ブロックをプログラミングした直後に測定されるビット・エラー・レートから決定することができる。

【００１１】
少なくとも１つの実施形態では、選択された書き込みストリームが複数の書き込みストリーム全ての中で最大の摩耗を反映するストリーム摩耗指標を有する場合に、コントローラが、選択された書き込みストリームに割り当てられた物理ブロックの中から最も良好なヘルスの物理ブロックを転送用に選択し、選択された書き込みストリームが複数の書き込みストリーム全ての中で最小の摩耗を反映するストリーム摩耗指標を有する場合に、コントローラが、選択された書き込みストリームに割り当てられた物理ブロックの中から最も不良なヘルスの物理ブロックを転送用に選択するように、コントローラは転送用に物理ブロックを選択する。
【００１２】
少なくとも１つの実施形態では、コントローラは、第１および第２のプールのうちの所与の１つから利用可能なプログラムされていない物理ブロックの数が閾値を満たさないことに基づいて、第１および第２のプールの間で物理ブロックを転送する。
【図面の簡単な説明】
【００１３】
【図１Ａ】一実施形態によるデータ処理環境の高レベルのブロック図である。
【図１Ｂ】図１Ａのデータ・ストレージ・システムのフラッシュ・カードのより詳細なブロック図である。
【図２】一実施形態による例示的なＮＡＮＤフラッシュ・メモリ・モジュールを示す図である。
【図３】一実施形態による、フラッシュ管理機能と、フラッシュ管理に利用されるデータ構造との高レベルのフロー図である。
【図４】物理ブロックのプールが、セルあたりに異なる数のビットを記憶する異なるモードで動作するように構成される例示的な階層型ストレージ・アーキテクチャを示す図である。
【図５】例示的な階層型ストレージ・アーキテクチャについてのライト・アンプリフィケーション対デバイス使用率のグラフである。
【図６】一実施形態による、不揮発性メモリを構成するための例示的な処理の高レベルの論理フローチャートである。
【図７】一実施形態による、ブロック・プール間でブロックを動的に転送するための例示的な処理の高レベルの論理フローチャートである。
【発明を実施するための形態】
【００１４】
図を参照すると、特に図１Ａを参照すると、本明細書でさらに説明する不揮発性メモリ・アレイを有するデータ・ストレージ・システム１２０を含む例示的なデータ処理環境１００の高レベルのブロック図が示されている。図示のように、データ処理環境１００は、命令およびデータを処理する１つまたは複数のプロセッサ１０４を有するプロセッサ・システム１０２などの１つまたは複数のホストを含む。プロセッサ・システム１０２は、プロセッサ１０４によって実行される処理のプログラム・コード、オペランド、または実行結果あるいはその組み合わせを記憶し得るローカル・ストレージ１０６（たとえば、ダイナミック・ランダム・アクセス・メモリ（ＤＲＡＭ）またはディスク）を追加的に含み得る。様々な実施形態では、プロセッサ・システム１０２は、たとえば、モバイル・コンピューティング・デバイス（たとえば、スマートフォンまたはタブレット）、ラップトップまたはデスクトップのパーソナル・コンピュータ・システム、サーバ・コンピュータ・システム（たとえば、インターナショナル・ビジネス・マシーンズ（Ｒ）社から入手できるＰＯＷＥＲ（Ｒ）シリーズのサーバの１つ）、あるいはメインフレーム・コンピュータ・システムとすることができる。プロセッサ・システム１０２はまた、任意のメモリ・キャッシュ、メモリ・コントローラ、ローカル・ストレージ、Ｉ／Ｏバス・ハブなどと組み合わせられた、ＡＲＭ（Ｒ）、ＰＯＷＥＲ（Ｒ）、インテル（Ｒ）ｘ８６、または他の任意のプロセッサなどの様々なプロセッサを使用する組み込みプロセッサ・システムとするこ

ともできる。
【００１５】
各プロセッサ・システム１０２は入力／出力（Ｉ／Ｏ）アダプタ１０８をさらに含み、これはＩ／Ｏチャネル１１０を介してデータ・ストレージ・システム１２０に直接（すなわち、介在デバイスなしで）または間接的に（すなわち、少なくとも１つの中間デバイスを介して）結合される。いくつかの実施形態では、データ・ストレージ・システム１２０は、プロセッサ・システム１０２に統合され得る。様々な実施形態において、Ｉ／Ｏチャネル１１０は、たとえば、ファイバ・チャネル（ＦＣ）、ＦＣオーバー・イーサネット（Ｒ）（ＦＣｏＥ）、インターネット・スモール・コンピュータ・システム・インターフェース（ｉＳＣＳＩ）、インフィニバンド、トランスポート制御プロトコル／インターネット・プロトコル（ＴＣＰ／ＩＰ）、ペリフェラル・コンポーネント・インターコネクト・エクスプレス（ＰＣＩｅ）などを含む、知られているまたは将来開発される通信プロトコルのうちの任意の１つまたは組み合わせを利用し得る。Ｉ／Ｏチャネル１１０を介して伝達されるＩ／Ｏ要求は、プロセッサ・システム１０２がデータ・ストレージ・システム１２０にデータを要求する読み出し要求と、プロセッサ・システム１０２がデータ・ストレージ・システム１２０へのデータの記憶を要求する書き込み要求と含む。
【００１６】
必須ではないが、図示の実施形態では、データ・ストレージ・システム１２０は複数のインターフェース・カード１２２を含み、これを経由して、データ・ストレージ・システム１２０は、Ｉ／Ｏチャネル１１０を介してホストのＩ／Ｏ要求を受信し、それに応答する。各インターフェース・カード１２２は、フォールト・トレランスおよび負荷分散を容易にするために、複数の安価ディスク冗長アレイ（ＲＡＩＤ：Redundant Array of Inexpensive Disks）コントローラ１２４のそれぞれに結合される。転じて、ＲＡＩＤコントローラ１２４のそれぞれは、不揮発性記憶媒体に（たとえば、ＰＣＩｅバスによって）結合され、不揮発性記憶媒体は、図示の例では、ＮＡＮＤフラッシュ・メモリを支える複数のフラッシュ・カード１２６を含む。他の実施形態では、代替または追加あるいはその両方の不揮発性ストレージ・デバイスを利用することができる。
【００１７】
図示の実施形態では、データ・ストレージ・システム１２０の動作は、冗長なシステム管理コントローラ（ＳＭＣ：system management controller）１２３によって管理され、ＳＭＣ１２３はインターフェース・カード１２２およびＲＡＩＤコントローラ１２４に結合される。様々な実施形態では、システム管理コントローラ１２３は、ハードウェア、あるいはファームウェアもしくはソフトウェアまたはその両方を実行するハードウェアを利用して実装することができる。
【００１８】
図１Ｂは、図１Ａのデータ・ストレージ・システム１２０のフラッシュ・カード１２６の例示的な実施形態のより詳細なブロック図を示している。フラッシュ・カード１２６は、フラッシュ・カード１２６とＲＡＩＤコントローラ１２４との間のインターフェースとして機能するゲートウェイ１３０を含む。ゲートウェイ１３０は、汎用プロセッサ（ＧＰＰ：general-purpose processor）１３２に結合され、ＧＰＰ１３２は、ゲートウェイ１３０によって受信した要求に対して前処理を実行し、またはフラッシュ・カード１２６による要求へのサービス提供をスケジューリングし、あるいはその両方を行うように（たとえば、プログラム・コードによって）構成することができる。ＧＰＰ１３２は、ＧＰＰメモリ１３４（たとえば、ダイナミック・ランダム・アクセス・メモリ（ＤＲＡＭ））に結合され、ＧＰＰメモリ１３４は、ＧＰＰ１３２によってその処理の最中に作成、参照、または修正、あるいはその組み合わせが行われるデータ、あるいはフラッシュ・コントローラ１４０のうちの１つまたは複数宛てにゲートウェイ１３０を介して流れるデータをバッファリングできると好都合である。
【００１９】
ゲートウェイ１３０はさらに複数のフラッシュ・コントローラ１４０に結合され、各フ

ラッシュ・コントローラ１４０は、それぞれのＮＡＮＤフラッシュ・メモリ・システム１５０を制御する。フラッシュ・コントローラ１４０は、たとえば、特定用途向け集積回路（ＡＳＩＣ）、および／またはフィールド・プログラマブル・ゲート・アレイ（ＦＰＧＡ）、ならびに／あるいはマイクロプロセッサによって実装することができ、各フラッシュ・コントローラ１４０には、フラッシュ・コントローラ・メモリ１４２（たとえば、ＤＲＡＭ）が関連付けられている。フラッシュ・コントローラ１４０がＦＰＧＡで実装される実施形態では、ＧＰＰ１３２は、データ・ストレージ・システム１２０の起動中にフラッシュ・コントローラ１４０をプログラムして構成し得る。起動後、通常動作時に、フラッシュ・コントローラ１４０は、ＮＡＮＤフラッシュ・メモリ・システム１５０に記憶されたデータを読み出すこと、またはＮＡＮＤフラッシュ・メモリ・システム１５０にデータを記憶すること、あるいはその両方を要求する読み出し要求および書き込み要求をゲートウェイ１３０から受信する。フラッシュ・コントローラ１４０は、たとえば、ＮＡＮＤフラッシュ・メモリ・システム１５０にアクセスして要求されたデータをＮＡＮＤフラッシュ・メモリ・システム１５０に対して読み書きすることによって、またはＮＡＮＤフラッシュ・メモリ・システム１５０に関連付けられたメモリ・キャッシュ（図示せず）にアクセスすることによって、これらの要求にサービス提供する。

【００２０】

フラッシュ・コントローラ１４０は、ＮＡＮＤフラッシュ・メモリ・システム１５０内の特定のメモリ位置へのアクセスを可能にする論理／物理アドレス変換を提供するフラッシュ変換レイヤ（ＦＴＬ：flash translation layer）を実装する。一般に、プロセッサ・システム１０２などのホスト・デバイスからフラッシュ・コントローラ１４０によって受信される要求は、データがアクセスされる（読み出しまたは書き込みされる）論理ブロック・アドレス（ＬＢＡ：logical block address）を含み、書き込み要求の場合は、データ・ストレージ・システム１２０に記憶される書き込みデータも含む。要求は、アクセスするデータの量（またはサイズ）も指定し得る。データ・ストレージ・システム１２０によってサポートされるプロトコルおよび機能に応じて、他の情報も伝達され得る。フラッシュ変換レイヤは、ＲＡＩＤコントローラ１２４から受信したＬＢＡを、ＮＡＮＤフラッシュ・メモリ・システム１５０内の対応する物理位置に割り当てられた物理アドレスに変換する。フラッシュ・コントローラ１４０は、アドレス変換を実行し、またはフラッシュ・コントローラ・メモリ１４２に記憶されると好都合であり得る論理／物理変換テーブル（ＬＰＴ：logical-to-physical translation table）などの論理／物理変換データ構造内に論理アドレスと物理アドレスとの間のマッピングを記憶し、あるいはその両方を行い得る。

【００２１】

ＮＡＮＤフラッシュ・メモリ・システム１５０は、様々な実施形態において多くの形態をとり得る。図１Ｂに示す実施形態では、各ＮＡＮＤフラッシュ・メモリ・システム１５０は、複数の（たとえば、３２個の）個別にアドレス指定可能なＮＡＮＤフラッシュ・メモリ・ストレージ・デバイス１５２を含む。図示の例では、フラッシュ・メモリ・ストレージ・デバイス１５２は、基板に取り付けられたフラッシュ・メモリ・モジュール、たとえば、マルチ・レベル・セル（ＭＬＣ）、３レベル・セル（ＴＬＣ）、またはクアッド・レベル・セル（ＱＬＣ）のＮＡＮＤフラッシュ・メモリ・モジュールの形態をとる。フラッシュ・メモリ・ストレージ・デバイス１５２によって提供される実質的な記憶容量は、たとえば、フラッシュ・コントローラ１４０、またはＧＰＰ１３２、ＲＡＩＤコントローラ１２４、もしくはＳＭＣ１２３などの高レベルのコントローラ、あるいはその両方によるデータ圧縮の実装を通じて増加させることができる。

【００２２】

ここで図２を参照すると、図１ＢのＮＡＮＤフラッシュ・メモリ・ストレージ・デバイス１５２のうちのいずれかを実装するために利用することができる例示的なフラッシュ・メモリ・モジュール２００のブロック図が示されている。フラッシュ・メモリ・モジュール２００は１つまたは複数のメモリ・ダイを含み、各メモリ・ダイは、ＮＡＮＤフラッシ

ュ・メモリ・セルの２次元または３次元アレイから形成される少なくとも１つのメモリ・アレイ２０２を実装する。図２に示すように、メモリ・アレイ２０２内のメモリ・セルは、複数の物理ブロック２０４内に物理的に配置され、各物理ブロックは複数の物理ページ２０６を含む。これらのページはページ・グループ単位で管理することができ、各ページ・グループは、たとえば、共通のワード線に結合された全てのページ、３Ｄ　ＮＡＮＤフラッシュ内の１つまたは複数のレイヤの全てのページ、１つまたは複数のレイヤ内のページのセット、あるいは一般的には同様の特性を有するページで形成することができる。

【００２３】

当業者に知られているように、メモリ・アレイ２０２で利用されるようなＮＡＮＤフラッシュ・メモリは、プログラムされる前に消去されなければならない。さらに、ＮＡＮＤフラッシュ・メモリはその構造によって制約され得（ただし、必ずしもそうではない）、消去できる最小記憶粒度がブロック２０４であり、読み出し要求または書き込み要求によってアクセスできる最小記憶粒度が単一の物理ページ２０６のサイズに固定される。この点に関して、ホスト・デバイスによって提供されるＬＢＡは論理アドレス空間における論理ページに対応し、各論理ページは典型的には４キロバイト（ｋＢ）のサイズを有することを理解されたい。対照的に、物理ページ２０６は典型的にはより大きいサイズ、たとえば、１６ｋＢを有するので、複数の論理ページをホストすることができる。

【００２４】

フラッシュ・メモリ・モジュール２００は、メモリ・アレイ２０２のワード線をアドレス指定することができる行デコーダ２１０と、メモリ・アレイ２０２のビット線をアドレス指定することができる列デコーダ２１２とをさらに含む。また、フラッシュ・メモリ・モジュール２００は、物理ページ２０６のメモリ・セルを並列にプログラムまたは読み出し可能にする読み出し／書き込み回路２１４を含む。フラッシュ・コントローラ２００は制御回路２０５を追加的に含み、制御回路２０５は、メモリ・アレイ２０２内の物理ページ２０６へ行われる読み出しアクセスおよび書き込みアクセス、ブロック２０４の消去を含む、メモリ・アレイ２０２の動作と、メモリ・アレイ２０２に印加される関連する電圧の振幅、持続時間、および極性とのチップ・レベルの制御を提供する。

【００２５】

データ・ストレージ・システム１２０の一例示的実施形態の一般的な物理構造を説明したので、ここでデータ・ストレージ・システム１２０の特定の動作態様を図３を参照して説明し、図３は、一実施形態によるＧＰＰ１３２またはフラッシュ・コントローラ１４０あるいはその両方によって利用されるフラッシュ管理機能およびデータ構造の高レベルのフロー図である。

【００２６】

データ・ストレージ・システム１２０は、一般に、外部デバイス（たとえば、ホスト）がＮＡＮＤフラッシュ・メモリ・システム１５０内の物理メモリ位置を直接アドレス指定すること、またはそれらに直接アクセスすること、あるいはその両方を許可しない。代わりに、データ・ストレージ・システム１２０は一般に、それぞれが連続した論理アドレス空間を有する１つまたは複数の論理ボリュームをホスト・デバイスに提示するように構成され、それによってホスト・デバイスが、論理アドレス空間内の論理ブロック・アドレス（ＬＢＡ）に対してデータを読み書きすることを可能にしつつ、様々なレベルのコントローラ（たとえば、システム管理コントローラ１２３、ＲＡＩＤコントローラ１２４、フラッシュ・コントローラ１４０、およびＧＰＰ１３２）のうちの１つまたは複数が、様々なＬＢＡに関連付けられたデータがＮＡＮＤフラッシュ・メモリ・システム１５０を含む物理メモリ位置において実際に存在する場所を制御することを可能にする。このようにして、ＮＡＮＤフラッシュ・メモリ・システム１５０のパフォーマンスおよび寿命をインテリジェントに管理および最適化することができる。

【００２７】

図示の実施形態では、各フラッシュ・コントローラ１４０は、関連付けられたフラッシュ・コントローラ・メモリ１４２などに記憶することができる論理／物理変換（ＬＰＴ）

テーブル３００などの論理／物理アドレス変換データ構造を使用して、関連付けられたＬＢＡのセットに対して論理／物理アドレス変換を実行する。フラッシュ・コントローラ１４０に供給される論理アドレスは、データ・ストレージ・システム１２０に当初供給された論理アドレスと異なり得ることに留意にすべきであり、その理由は、データ・ストレージ・システム１２０内の様々なコンポーネントが、外部デバイスとフラッシュ・コントローラ１４０との間でアドレス変換動作を実行し得るためである。

【００２８】

ＧＰＰ１３２上で動作しているフラッシュ管理コードは、ＧＰＰメモリ１３４などに記憶され得る即時使用可能（ＲＴＵ：ready-to-use）キュー３０６内で使用準備ができているＮＡＮＤフラッシュ・メモリ・システム１５０の消去済みブロックを追跡する。図示の実施形態では、ＧＰＰ１３２上で動作しているフラッシュ管理コードは、チャネルまたはプレーンあたり（すなわち、データ・バスあたり）１つのＲＴＵキュー３０６を維持し、再利用される各消去済みブロックの識別子が、そのチャネルまたはプレーンに対応するＲＴＵキュー３０６にエンキューされる。追加的に、異なるヘルス・グレード、ヒート・ビン／書き込みストリーム、またはブロック動作モード、すなわち、ＳＬＣもしくはＱＬＣ動作モードに対して別々のＲＴＵキュー３０６が維持され得る。ＧＰＰ１３２上で動作しているフラッシュ管理コードによって実行されるブロック・ストライプ構築機能３２０は、ＲＴＵ３０６キューにエンキューされた消去済みブロックからデータおよび関連するパリティ情報を記憶するための新しいブロック・ストライプを構築する。次いで、新しいブロック・ストライプは、データ配置のためにフラッシュ・コントローラ１４０にキューイングされる。ブロック・ストライプは、異なるチャネルに存在するブロックで形成されることが好ましく、これは、ブロック・ストライプ構築機能３２０が、新しいブロック・ストライプの各ブロックを異なるＲＴＵキュー３０６から引き出すことによって、ブロック・ストライプを構築できると好都合であるということを意味する。一般に、ブロック・ストライプ構築機能３２０は、ほぼ等しいヘルス（すなわち、予想される残存耐用年数）のブロックからストライプを構築しようと試みる。

【００２９】

プロセッサ・システム１０２などのホストから受信した書き込み要求に応答して、フラッシュ・コントローラ１４０のデータ配置機能３１０は、ＬＰＴテーブル３００を参照することにより、書き込み要求に示されたターゲットＬＢＡがＮＡＮＤフラッシュ・メモリ・システム１５０内の物理メモリ・ページに現在マッピングされているか否かを判定し、マッピングされている場合、ターゲットＬＢＡに現在関連付けられている各データページのステータスを変更して、関連付けられたデータが無効になったことを示すようにする。また、データ配置機能３１０は、必要に応じてページ・ストライプを割り当てて、書き込み要求の書き込みデータと、もしあれば、書き込み要求の対象となる既存のページ・ストライプからの更新されないデータ（すなわち、論理ページよりも小さい書き込み要求の場合に、上書きされない、リード・モディファイ・ライトの方法で処理する必要がある、同じ論理アドレスへの前回の書き込みから残っている有効なデータ）とを記憶し、または書き込み要求の書き込みデータと、もしあれば、書き込み要求の対象となる既存のページ・ストライプからの更新されない（すなわち、まだ有効な）データとを、空きスペースが残っている既に割り当て済みのページ・ストライプに記憶し、あるいはその両方を行う。ページ・ストライプは、データを保持するために既に割り当てられているブロック・ストライプから、またはブロック・ストライプ構築機能３２０によって構築される新しいブロック・ストライプから割り当てられ得る。好ましい実施形態では、ページ・ストライプの割り当ては、割り当てに利用可能なブロックのヘルスと、書き込みデータのＬＢＡの「ヒート」（すなわち、推定または測定された書き込みアクセス頻度）とに基づくことができる。次いで、データ配置機能３１０は、書き込みデータと、関連するメタデータ（たとえば、巡回冗長符号（ＣＲＣ）および誤り訂正符号（ＥＣＣ）の値）と、ページ・ストライプのパリティ情報とを割り当てられたページ・ストライプに書き込む。フラッシュ・コントローラ１４０はまた、書き込みデータを記憶するために利用される物理ページを、ホスト

・デバイスによって示されたＬＢＡに関連付けるようにＬＰＴテーブル３００を更新する。その後、フラッシュ・コントローラ１４０は、図３にさらに示すように、ＬＰＴテーブル３００を参照することによって、ホスト読み出し要求にサービス提供するためにデータにアクセスすることができる。

【００３０】

ブロック・ストライプ内の全てのページが書き込まれると、フラッシュ・コントローラ１４０は、ブロック・ストライプを占有済みブロック・キュー３０２の１つに配置し、ＧＰＰ１３２上で動作しているフラッシュ管理コードはこれを利用してガベージ・コレクションを容易にする。上記のように、書き込み処理を通じて、ページが無効化されるので、ＮＡＮＤフラッシュ・メモリ・システム１５０の一部は不使用になる。関連付けられたフラッシュ・コントローラ１４０（またはＧＰＰ１３２、あるいはその両方）は最終的に、ガベージ・コレクタ３１２によって実行されるガベージ・コレクションを通じて、このスペースを回収する必要がある。ガベージ・コレクタ３１２は、たとえば、ブロック・ストライプ内のブロック２０４のヘルス、および消去ブロック２０４内のデータのうちどれくらいが無効であるかを含むいくつかの要因に基づいて、ガベージ・コレクションのために特定のブロック・ストライプを選択する。図示の例では、ガベージ・コレクションがブロック・ストライプ全体に実行され、ＧＰＰ１３２上で動作しているフラッシュ管理コードは、リサイクルする準備ができたブロック・ストライプのログを再配置キュー３０４内にとり、再配置キュー３０４は、関連付けられたフラッシュ・コントローラ・メモリ１４２またはＧＰＰメモリ１３４内に実装できると好都合である。

【００３１】

ＧＰＰ１３２またはフラッシュ・コントローラ１４０によって実行されるフラッシュ管理機能は、再配置キュー３０４にエンキューされたブロック・ストライプに保持されたデータを再配置する再配置機能３１４を追加的に含む。そのようなデータを再配置するために、再配置機能３１４は、データ配置機能３１０に再配置書き込み要求を発行して、古いブロック・ストライプの有効なデータがＮＡＮＤフラッシュ・メモリ・システム１５０内の新しいブロック・ストライプに書き込まれるように要求する。また、再配置機能３１４は、そのデータの論理アドレスおよび物理アドレスの間の現在の関連付けを削除するようにＬＰＴテーブル３００を更新する。残りの有効なデータ全てが古いブロック・ストライプから移動されると、古いブロック・ストライプはブロック・ストライプ解体機能３１６に渡され、ブロック・ストライプ解体機能３１６は、古いブロック・ストライプをその構成要素ブロック２０４に分解して、ブロック２０４の関連付けを解除する。次いで、解体されたブロック・ストライプを以前形成していたブロック２０４のそれぞれは、フラッシュ・コントローラ１４０、または関連するフラッシュ・メモリ・モジュール２００の制御回路２０５、あるいはその両方の指示の下で消去され、各消去済みブロックの対応するプログラム／消去（Ｐ／Ｅ）サイクル・カウント３３４がインクリメントされる。各消去済みブロック２０４のヘルス指標３３２（たとえば、ビット・エラー・レート（ＢＥＲ）指標、訂正不可能なエラー、Ｐ／Ｅサイクル・カウントなど）に基づいて、各消去済みブロック２０４は、ＧＰＰ１３２上で実行されるフラッシュ管理機能のうちのブロック・リタイア機能３１８によってリタイアされ（すなわち、使用中止にされ）、または代替的には、ブロック２０４の識別子を関連付けられたＧＰＰメモリ１３４内の適切な即時使用可能（ＲＴＵ）キュー３０６に配置することによって、再利用のために準備される。

【００３２】

図３にさらに示すように、ＧＰＰ１３２またはフラッシュ・コントローラ１４０あるいはその両方の上で実行されるフラッシュ管理機能は、バックグラウンド・ヘルス・チェッカー３３０を追加的に含む。プロセッサ・システム１０２などのホストの読み出し要求および書き込み要求とは独立して動作するバックグラウンド・ヘルス・チェッカー３３０は、占有済みブロック・キュー３０２に記録されたブロック・ストライプに属するブロックに関する１つまたは複数のブロック・ヘルス指標３３２を継続的に決定する。ブロック・ヘルス指標３３２のうちの１つまたは複数に基づいて、バックグラウンド・ヘルス・チェ

ッカー３３０は、再配置機能３１４による処理のために、ブロック・ストライプを再配置キュー３０４に配置する。バックグラウンド・ヘルス・チェッカーによって監視および記録されることが好ましい主要なブロック・ヘルス指標３３２は、有効なブロックおよび物理ページについて観察されるビット・エラー・レート（ＢＥＲ）指標に関連するものであり、たとえば、各ブロックの最悪ページＢＥＲ、各ブロックの平均ページＢＥＲ、各ブロックの最悪ページＢＥＲおよび平均ページＢＥＲの変化率などを含み得る。可能な最も正確なヘルス推定値を得るために、有効なデータおよび無効なデータの分析からヘルスを決定することにより、ほぼ完全に無効なデータを含むブロックを十分に分析して、ブロック全体のヘルスを決定するようにすることができる。

【００３３】

バックグラウンド・ヘルス・チェッカー３３０に組み込むことができ、または別個に実装することができるＧＰＰ１３２またはフラッシュ・コントローラ１４０あるいはその両方の１つの機能は、メモリ・アレイ２０２のメモリ・セル内に記憶されたデータ・ビットをデコードするために利用される読み出し電圧閾値３３６の定期的な適応（較正）である。これらの読み出し電圧閾値３３６は、任意の所望のレベルの粒度（たとえば、ページごと、ブロック内のページ・グループごと、ブロックごとなど）に個別に定義できるものであるが、１つまたは複数のビット・エラー指標を向上させるように選択されることが好ましい。図３にさらに示すように、ＧＰＰ１３２またはフラッシュ・コントローラ１４０あるいはその両方は、読み出し電圧閾値３３６のインテリジェントな適応を容易にするために、追加の統計量を追跡することが好ましい。これらの追加の統計量は、たとえば、各ブロックが受けたプログラム／消去サイクルの数を特定するＰ／Ｅサイクル・カウント３３４と、所与のメモリ単位（たとえば、ページ、ページ・グループ、またはブロック、あるいはその組み合わせ）がプログラムされた後に読み出された回数を示す読み出しカウント３３８とを含むことができる。また、これらの統計量は、メモリ単位の読み出し電圧閾値を適応させたときからの経過時間を示すメモリ単位ごとのカウンタ、または各ブロックの読み出し電圧閾値を最後に適応させたとき以降にそのブロックが受けたプログラム／消去サイクルの数を示すブロックごとのカウンタ、あるいはその両方などの較正トリガ３４０を含むことができる。

【００３４】

ここで図４を参照すると、不揮発性メモリ・デバイスに実装され得る例示的なストレージ・アーキテクチャが示されている。この例では、ＧＰＰ１３２またはフラッシュ・コントローラ１４０あるいはその両方（本明細書では一般に「コントローラ」と呼ぶ）は、物理ブロック２０４の複数のプールを実装するようにＮＡＮＤフラッシュ・メモリ・システム１５０内のフラッシュ・メモリ・ストレージ・デバイス１５２の一部または全てを構成し、異なるブロック・プール内の物理ブロック２０４は、セルあたりに異なる数のビットを記憶する異なるモードで動作するように構成される。したがって、この例では、メモリ・アレイ２０２の全てが、たとえば、セルあたり４ビットを記憶することが可能なＱＬＣフラッシュ・メモリを含むが、コントローラは、フラッシュ・ストレージ・デバイス１５２のメモリ・アレイ２０２内の物理ブロック２０４の一部を、完全にプログラムされたブロックがセルあたり４ビットを記憶するＱＬＣプール４００に割り当て、メモリ・アレイ２０２の他の物理ブロック２０４を、完全にプログラムされたブロックがセルあたり１ビットのみを記憶するＳＬＣプール４０２に割り当てる。プール・メンバーシップ情報を、たとえば、ＧＰＰメモリ１２４またはフラッシュ・コントローラ・メモリ１４２あるいはその両方に、ブロック・モード・メタデータ３３５として記憶することができる。上記のように、このように複数の異なるモードでブロックを動作させることにより、頻繁に書き込まれるＬＢＡがＳＬＣプール４０２内の物理ブロック２０４にマッピングされる場合に、パフォーマンスまたは耐久性あるいはその両方を向上させることができる。

【００３５】

必須ではないが、いくつかの実施形態では、コントローラは、矢印４０４で示すように、ホスト書き込み要求に含まれるデータが、ＳＬＣプール４０２から選択された物理ブロ

ック２０４に最初に書き込まれる階層構成でＱＬＣプール４００およびＳＬＣプール４０２を編成する。ＳＬＣプール４０２が使用を通じて満杯になり、より多くの容量がＳＬＣプール４０２に必要とされ、または希望され、あるいはその両方になると、コントローラは、矢印４０６で示すように、再配置書き込みによってＳＬＣプール４０２からＱＬＣプール４００にデータをデステージする。ブロックが摩耗を受けたときにブロック・プール４００および４０２の両方で物理ブロック２０４を適切に供給できるようにするために、コントローラは追加的に、記憶容量または耐久性あるいはその両方を提供するために、必要に応じて、または要望に応じて、あるいはその両方に応じて、プール４００および４０２の間でブロック２０４を転送（再割り当て）する。ブロック・プール４００および４０２の間の（データではなく）物理ブロック２０４の転送は、図４に矢印４０８で示している。

【００３６】

本開示は、ブロック・プール４００および４０２の間でブロックを転送する処理は、注意深く設計されなければ、不必要に、（エラー回復活動の増加に起因する）パフォーマンスの低下、さらにはエラー回復が失敗した場合の（たとえば、ビット・エラー・レートが高いブロックがブロック・プール４０２からブロック・プール４００に転送された場合の）データ損失につながり得ることを理解している。一般に、コントローラは、ブロック・プール４００からやや不良な／最も不良なヘルスのブロックを選択してブロック・プール４０２に転送することによって、また、ブロック・プール４０２からやや良好な／最も良好なヘルスのブロックを選択してブロック・プール４００に転送することによって、ブロック・プール４００および４０２の間でブロックを転送する。

【００３７】

本開示は、この一般的なモード変換ポリシーは、一部のシステムには適切であるが、ヒート分別（heat segregation）を利用するシステム、またはデバイス使用率が非常に高いシステムでは準最適であり、これらは摩耗の加速、さらにはデバイスの早期廃棄に直結し得ることを理解している。これらの悪影響の１つの理由は、ＱＬＣプール４００内のブロック２０４の摩耗（たとえば、平均Ｐ／Ｅサイクルによって測定されるもの）が、特定の条件下で、ライト・アンプリフィケーションに部分的に起因して、一般にＳＬＣプール４０２のブロック２０４の摩耗よりもはるかに高い場合があることである。図５は、図４に示す例示的な階層型ストレージ・アーキテクチャについてのライト・アンプリフィケーション対デバイス使用率のグラフである。図示のように、高いデバイス使用率の下では（たとえば、フラッシュ・メモリ・デバイスの９０％以上にデータが書き込まれている場合）、ＱＬＣプール４００のライト・アンプリフィケーションは指数関数的に増加して、ＱＬＣプール４００のブロックがますます急速に摩耗し得る。したがって、以下でより詳細に論じるように、モード変換ポリシーは、ブロック・プール間で転送するブロックを選択する場合に、ブロックの摩耗を考慮に入れることが好ましい。

【００３８】

ここで図６を参照すると、一実施形態による、不揮発性メモリを構成するための例示的な処理の高レベルの論理フローチャートが示されている。図示の処理は、たとえば、データ・ストレージ・システム１２０の起動時に、ハードウェア、ファームウェア、ソフトウェア、またはその組み合わせで、コントローラ（たとえば、ＧＰＰ１３２またはフラッシュ・コントローラ１４０あるいはその両方）によって実行することができる。本明細書に提示した他の論理フローチャートと同様に、動作は厳密な時系列順ではなく論理的な順序で提示しており、いくつかの実施形態では、動作は、提示した順序とは異なる順序で、または同時に実行することができる。

【００３９】

図６の構成処理は、ブロック６００から始まり、次いでブロック６０２に進み、ブロック６０２は、フラッシュ・メモリ・デバイス１５２のコントローラが所望の数のブロック・プールを構成することを示している。図４の例では、コントローラは、ブロック６０２において２つのブロック・プール、すなわち、ＱＬＣブロック・プール４００およびＳＬ

Ｃブロック・プール４０２を実装する。他の実施形態では、より多くのブロック・プール、または別のブロック・モード（たとえば、ＴＬＣもしくはＭＬＣ）で動作するブロック・プール、あるいはその両方を実装することができる。コントローラは追加的に、初期数の物理ブロック２０４を各ブロック・プールに割り当て、各物理ブロック２０４のブロック・プール・メンバーシップに基づいて、各物理ブロック２０４の対応するブロック・モードをブロック・モード・メタデータ３３５に記録する。

【００４０】

処理はブロック６０２からブロック６０４に進み、ブロック６０４は、コントローラが、ブロック６０２において実装された各ブロック・プールが処理される処理ループに入ることを示している。コントローラがブロック６０４において、全てのブロック・プールがまだ処理されていないと判定した場合、処理はブロック６１０に移り、ブロック６１０は、コントローラが、処理すべき次のブロック・プール（たとえば、ＱＬＣプール４００またはＳＬＣプール４０２）を選択することを示している。次いで、ブロック６１２において、コントローラは、選択されたブロック・プールに属するブロック２０４に対して所望の数のヘルス・グレードを構成する。様々な実施形態において、ヘルス・グレードの数は、異なるものにすることができるが、実験的には、２以上８以下のヘルス・グレードの数が効果的であることがわかっている。たとえば、図４に示す一実施形態では、コントローラは、ブロック６１２において４つのヘルス・グレード４１０ａ～４１０ｄまたは４１２ａ～４１２ｄを構成する。本明細書で利用するヘルス・グレードは、具体的には、単なるプログラム／消去（Ｐ／Ｅ）サイクル数以外の少なくとも１つの基準、たとえば、物理メモリの各ブロック（または各ブロック内の物理ページ、あるいはその両方）の１つまたは複数のビット・エラー指標に基づくように定義される。たとえば、一実施形態では、様々なヘルス・グレードが、プログラミング後の少数の読み出しアクセス内で決定される最悪ページＢＥＲのそれぞれの範囲によって定義され、その理由は、このビット・エラー指標がブロックの摩耗の永続的な影響を正確に反映しているためである。測定された最悪ページＢＥＲに基づくと、最も良好なヘルスのブロックは、プログラミング直後の最悪ページＢＥＲが最も低いものであり、最も不良なヘルスのブロックは、プログラミング直後の最悪ページＢＥＲが最も高いものである。

【００４１】

ブロック６１４において、コントローラは追加的に、フラッシュ・メモリ・ストレージ・デバイス１５２によって扱われるＬＢＡに対して所望の数のヒート・ビンを構成する。各ヒート・ビンは、ＬＢＡの書き込みヒート（すなわち、書き込み頻度）のそれぞれの範囲に対応する。様々な実施形態において、ヒート・ビンの数は、異なるものにすることができるが、実験的には、２以上８以下のヒート・ビンの数が効果的であることがわかっている。たとえば、一実施形態では、コントローラはブロック６１４において４つのヒート・ビンを構成し、それぞれ、最も高い頻度で書き込まれるホットＬＢＡ、より少ない頻度で書き込まれるウォームＬＢＡ、さらに少ない頻度で書き込まれるクールＬＢＡ、および最も少ない頻度で書き込まれるコールドＬＢＡに対応する。所与のヒート・ビン内のＬＢＡを対象とするホスト書き込み要求のセットを本明細書では「書き込みストリーム」と呼ぶ。ブロック６１４において構成されるヒート・ビン（ひいては書き込みストリーム）の数は、ブロック６１２において構成されるヘルス・グレードの数と等しくてもよいが、必須ではない。

【００４２】

ブロック６１６において、コントローラは、ブロック６１２において構成されるヒート・ビンとブロック６１４において構成されるヘルス・グレードとの間のマッピングを確立する。１つの特に好ましい実施形態では、ヒートおよびヘルスは直接的に相互に関連付けられ、最もホットなＬＢＡは最も良好なヘルス・グレードにマッピングされ、最もコールドなＬＢＡはより不良なヘルス・グレードにマッピングされる。当然ながら、他の実施形態では、書き込みヒートおよびヘルス・グレードの間の他のマッピングが利用され得る。たとえば、数値の書き込みヒートおよびヘルス・グレードが利用される一実施形態では、

ブロック６１６において確立されるマッピングは、相互にマッピングされたヘルス・グレードおよび書き込みヒートからほぼ一定の結果をもたらすように選択された数学関数によって決定される（たとえば、Ｃ　ｆ（ヒート，ヘルス））。他の実施形態では、書き込みヒートからヘルス・グレードへのマッピングは、全てのヒート・ビンをヘルス・グレードのソートされた優先順位リストにマッピングする。ＲＴＵキュー３０６内のブロックからブロック・ストライプが構築されるときに、所与のヘルス・グレードを有する優先ＲＴＵキュー３０６内にブロックがない場合がある。その際、ソートされた優先順位リストによって、優先ＲＴＵキューに極力近いヘルス・グレードを有する他のＲＴＵキュー３０６からブロックを割り当てることが可能になる。

【００４３】

ブロック６１６に続いて、図６に与えた構成処理は、ブロック６０４に戻る。ブロック６０４において、１つまたは複数の追加のブロック・プールがまだ処理されていないことを示す否定判定がなされた場合、処理は説明済みのブロック６１０～６１６に戻る。しかしながら、ブロック６０２において構成された全てのブロック・プールが処理されたとの判定がブロック６０４においてなされた場合、図６の構成処理はブロック６０６において終了する。

【００４４】

ここで図７を参照すると、一実施形態による、ブロック・プール間でブロックを動的に転送するための例示的な処理の高レベルの論理フローチャートが示されている。図示の処理は、たとえば、データ・ストレージ・システム１２０の動作中に、ハードウェア、ファームウェア、ソフトウェア、またはその組み合わせで、コントローラ（たとえば、ＧＰＰ１３２またはフラッシュ・コントローラ１４０あるいはその両方）によって実行することができる。

【００４５】

図７の処理は７００から始まり、次いでブロック７０２に進み、ブロック７０２は、コントローラが、ブロック・プール４００および４０２のうちの１つのサイズを変更するか否かを判定することを示している。たとえば、コントローラは、１つまたは複数のＲＴＵキュー３０６内の所与のブロック・プールに属するブロックの充填レベルが閾値を満たさないこと、１つまたは複数のＲＴＵキュー３０６内の他のブロック・プールに属するブロックの充填レベル、所与のブロック・プール内のブロックのＰ／Ｅサイクル・カウント、ヘルス・グレードのうちの１つまたは複数が割り当てられた所与のブロック・プール内のブロックの数などを含む１つまたは複数の基準に基づいて、所与のブロック・プールのサイズを変更するか否かを判定し得る。ブロック７０２においてなされる判定は、たとえば、バックグラウンド・ヘルス・チェッカー３３０、ガベージ・コレクタ３１２、ブロック・リタイア機能３１８、またはブロック・ストライプ構築機能３２０あるいはその組み合わせのうちの１つまたは複数によってトリガされ得る。ブロック７０２においてコントローラがブロック・プールのサイズを変更しないと判定したことに応答して、処理はブロック７０２を反復する。ブロック７０２においてブロック・プールのサイズを変更する判定したことに応答して、処理はブロック７０４に進む。

【００４６】

ブロック７０４は、コントローラが、物理ブロック２０４が転送されるブロック・プールを選択することを示している。また、ブロック７０６において、コントローラは、所与のブロック・プールに物理ブロック２０４が追加されるヒート・ビンまたは書き込みストリーム（たとえば、ホット、ウォーム、クール、またはコールド）を選択する。次いで、コントローラは、ブロック７０８において、選択されたヒート・ビン内のＬＢＡに関連付けられたデータを記憶するために現在割り当てられているブロック２０４に対してストリーム摩耗指標を決定する。１つの好ましい実施形態では、コントローラは、各ヒート・ビン／書き込みストリームに割り当てられたブロック２０４を追跡し、それらのブロック２０４が受けたＰ／Ｅサイクルの総数を（たとえば、Ｐ／Ｅサイクル・カウント３３４から）決定する。これらの数値から、コントローラは、選択されたヒート・ビンのストリーム

摩耗指標を、ヒート・ビン内のデータを記憶するブロック２０４の総数を、それらのブロック２０４が受けたＰ／Ｅサイクルの総数で割った値として計算することができる。

【００４７】

次いで、ブロック７０８において決定されたストリーム摩耗指標に基づいて、コントローラは、選択されたブロック・プールに転送する、選択されたヒート・ビンに適切な個別のブロック・ヘルスを有する少なくとも１つのプログラムされていないブロック２０４を他のブロック・プールから選択する（ブロック７１０）。この選択によって、最も良好なヘルスのブロック２０４は、将来受けることになるＰ／Ｅサイクル数が最大となり、最も不良なヘルスのブロック２０４は、将来受けることになるＰ／Ｅサイクル数が最小となるようになる。その結果、全てのタイプのワークロードおよび全てのデバイス使用率のポイントで、デバイスの耐久性が向上する。ブロック７１２において、コントローラは、ブロック７１０において選択されたブロック２０４を選択されたブロック・プールに転送し、ブロック・モード・メタデータ３３５を新しい動作モード（たとえば、ＱＬＣまたはＳＬＣ）で更新し、必要に応じて、ブロックのヘルス・グレードを更新する。その後、図７の処理は、説明済みのブロック７０２に戻る。

【００４８】

説明したように、少なくとも１つの実施形態では、不揮発性メモリは複数の物理記憶ブロックを含み、各物理記憶ブロックはそれぞれの複数のセルを含み、複数のセルのそれぞれは個々に複数ビットのデータを記憶することが可能である。コントローラは、複数の物理ブロックのうちの物理ブロックを、セルあたりより多くのビットを記憶するための第１の（たとえば、ＱＬＣ）モードで動作する物理ブロックを含む第１のプールに割り当て、複数の物理ブロックのうちの他の物理ブロックを、セルあたりより少ないビットを記憶するための第２の（たとえば、ＳＬＣ）モードで動作する物理ブロックを含む第２のプールに割り当てる。コントローラは、第１のプールおよび第２のプールの間で物理ブロックを転送することを、転送される物理ブロックについて測定されたビット・エラー・レートに少なくとも基づいて行う。

【００４９】

本発明は、システム、方法、またはコンピュータ・プログラム製品あるいはその組み合わせであり得る。コンピュータ・プログラム製品は、プロセッサに本発明の態様を実施させるためのコンピュータ可読プログラム命令をその上に有するコンピュータ可読記憶媒体（または複数の媒体）を含み得る。

【００５０】

コンピュータ可読記憶媒体は、命令実行デバイスによる使用のために命令を保持および記憶可能な有形のデバイスとすることができる。コンピュータ可読記憶媒体は、たとえば、限定はしないが、電子ストレージ・デバイス、磁気ストレージ・デバイス、光学ストレージ・デバイス、電磁ストレージ・デバイス、半導体ストレージ・デバイス、またはこれらの任意の適切な組み合わせであり得る。コンピュータ可読記憶媒体のより具体的な例の非網羅的なリストには、ポータブル・コンピュータ・ディスケット、ハード・ディスク、ランダム・アクセス・メモリ（ＲＡＭ）、読み取り専用メモリ（ＲＯＭ）、消去可能プログラム可能読み取り専用メモリ（ＥＰＲＯＭまたはフラッシュ・メモリ）、スタティック・ランダム・アクセス・メモリ（ＳＲＡＭ）、ポータブル・コンパクト・ディスク読み取り専用メモリ（ＣＤ－ＲＯＭ）、デジタル・バーサタイル・ディスク（ＤＶＤ）、メモリー・スティック（Ｒ）、フロッピー（Ｒ）・ディスク、命令が記録されたパンチ・カードまたは溝の隆起構造などの機械的にコード化されたデバイス、およびこれらの任意の適切な組み合わせが含まれる。コンピュータ可読記憶媒体は、本明細書で使用する場合、たとえば、電波または他の自由に伝搬する電磁波、導波管もしくは他の伝送媒体を伝搬する電磁波（たとえば、光ファイバ・ケーブルを通過する光パルス）、または有線で伝送される電気信号などの一過性の信号自体であると解釈されるべきではない。

【００５１】

本明細書に記載のコンピュータ可読プログラム命令は、コンピュータ可読記憶媒体から

それぞれのコンピューティング／処理デバイスに、あるいは、たとえば、インターネット、ローカル・エリア・ネットワーク、ワイド・エリア・ネットワーク、もしくは無線ネットワーク、またはその組み合わせなどのネットワークを介して外部コンピュータまたは外部ストレージ・デバイスにダウンロードすることができる。ネットワークは、銅線伝送ケーブル、光伝送ファイバ、無線伝送、ルータ、ファイアウォール、スイッチ、ゲートウェイ・コンピュータ、またはエッジ・サーバ、あるいはその組み合わせを含み得る。各コンピューティング／処理デバイスのネットワーク・アダプタ・カードまたはネットワーク・インターフェースは、ネットワークからコンピュータ可読プログラム命令を受信し、コンピュータ可読プログラム命令を転送して、それぞれのコンピューティング／処理デバイス内のコンピュータ可読記憶媒体に記憶する。

【００５２】

本発明の動作を実施するためのコンピュータ可読プログラム命令は、アセンブラ命令、命令セット・アーキテクチャ（ＩＳＡ）命令、機械命令、機械依存命令、マイクロコード、ファームウェア命令、状態設定データ、あるいはＳｍａｌｌｔａｌｋ（Ｒ）、Ｃ＋＋などのオブジェクト指向プログラミング言語、および「Ｃ」プログラミング言語または類似のプログラミング言語などの従来の手続き型プログラミング言語を含む、１つまたは複数のプログラミング言語の任意の組み合わせで記述されたソース・コードまたはオブジェクト・コードであり得る。コンピュータ可読プログラム命令は、完全にユーザのコンピュータ上で、部分的にユーザのコンピュータ上で、スタンドアロン・ソフトウェア・パッケージとして、部分的にユーザのコンピュータ上かつ部分的にリモート・コンピュータ上で、あるいは完全にリモート・コンピュータまたはサーバ上で実行され得る。後者のシナリオでは、リモート・コンピュータは、ローカル・エリア・ネットワーク（ＬＡＮ）またはワイド・エリア・ネットワーク（ＷＡＮ）を含む任意のタイプのネットワークを介してユーザのコンピュータに接続され得、または（たとえば、インターネット・サービス・プロバイダを使用してインターネットを介して）外部コンピュータへの接続がなされ得る。一部の実施形態では、たとえば、プログラマブル論理回路、フィールド・プログラマブル・ゲート・アレイ（ＦＰＧＡ）、またはプログラマブル・ロジック・アレイ（ＰＬＡ）を含む電子回路は、本発明の態様を実行するために、コンピュータ可読プログラム命令の状態情報を利用してコンピュータ可読プログラム命令を実行することによって、電子回路を個人向けにし得る。

【００５３】

本発明の態様は、本発明の実施形態による方法、装置（システム）、およびコンピュータ・プログラム製品のフローチャート図またはブロック図あるいはその両方を参照して本明細書で説明している。フローチャート図またはブロック図あるいはその両方の各ブロック、およびフローチャート図またはブロック図あるいはその両方におけるブロックの組み合わせが、コンピュータ可読プログラム命令によって実装できることは理解されよう。

【００５４】

これらのコンピュータ可読プログラム命令を、汎用コンピュータ、専用コンピュータ、または他のプログラム可能データ処理装置のプロセッサに提供して、それらの命令がコンピュータまたは他のプログラム可能データ処理装置のプロセッサを介して実行された場合に、フローチャートまたはブロック図あるいはその両方の１つまたは複数のブロックにおいて指定された機能／行為を実装するための手段が生成されるようなマシンを生成し得る。また、これらのコンピュータ可読プログラム命令を、コンピュータ、プログラム可能データ処理装置、または他のデバイス、あるいはその組み合わせに特定の方法で機能するように指示することが可能なコンピュータ可読記憶媒体に記憶して、命令が記憶されたコンピュータ可読記憶媒体が、フローチャートまたはブロック図あるいはその両方の１つまたは複数のブロックにおいて指定された機能／行為の態様を実装する命令を含む製造品を構成するようにし得る。

【００５５】

また、コンピュータ可読プログラム命令をコンピュータ、他のプログラム可能データ処

理装置、または他のデバイスにロードして、コンピュータ、他のプログラム可能装置、または他のデバイス上で一連の動作ステップを実行させることによって、それらの命令がコンピュータ、他のプログラム可能装置、または他のデバイス上で実行された場合に、フローチャートまたはブロック図あるいはその両方の１つまたは複数のブロックにおいて指定された機能／行為を実装されるようなコンピュータ実装処理を生成し得る。

【００５６】

図中のフローチャートおよびブロック図は、本発明の様々な実施形態によるシステム、方法、およびコンピュータ・プログラム製品の可能な実装形態のアーキテクチャ、機能、および動作を示している。これに関して、フローチャートまたはブロック図の各ブロックは、指定された論理的機能を実装するための１つまたは複数の実行可能命令を含むモジュール、セグメント、または命令の一部を表し得る。一部の代替的実装形態では、ブロックに記載した機能は、図示した順序以外で行われ得る。たとえば、関与する機能に応じて、連続して示した２つのブロックは、実際には実質的に同時に実行され得、またはそれらのブロックは、場合により逆の順序で実行され得る。ブロック図またはフローチャート図あるいはその両方の各ブロック、およびブロック図またはフローチャート図あるいはその両方におけるブロックの組み合わせは、指定された機能もしくは行為を実行するか、または専用ハードウェアおよびコンピュータ命令の組み合わせを実行する専用のハードウェア・ベースのシステムによって実装できることにも気付くであろう。

【００５７】

本発明は、１つまたは複数の好ましい実施形態を参照して説明したように具体的に示してきたが、当業者には、添付の特許請求の範囲に定義した本発明の範囲から逸脱することなく、形態および詳細の様々な変更がなされ得ることは理解されよう。たとえば、特定の機能を指示するフラッシュ・コントローラを含むデータ・ストレージ・システムに関する態様を説明してきたが、本発明は、代替的には、プロセッサがそのような機能を実行することによって、またはそのような機能を実行させることによって、処理することが可能なプログラム・コードを記憶するストレージ・デバイスを含むプログラム製品として実装され得ることを理解されたい。本明細書で使用する場合、「ストレージ・デバイス」は、具体的には、法定製造品（statutory article of manufacture）のみを含み、信号媒体自体、一過性の伝搬信号自体、エネルギー自体を除外するように定義する。

【００５８】

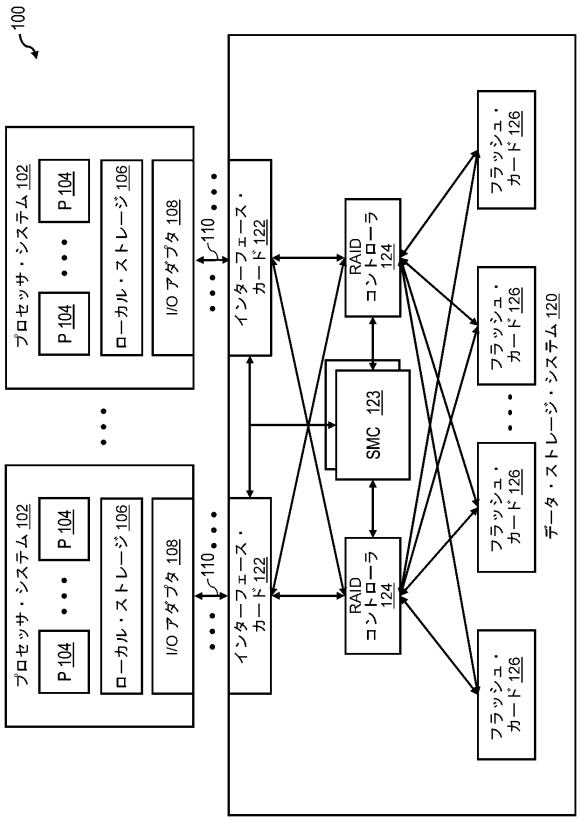
また、ＮＡＮＤフラッシュ・メモリの使用を含む実施形態を説明してきたが、本発明の実施形態は、他の任意のタイプの不揮発性ランダム・アクセス・メモリ（ＮＶＲＡＭ）でも使用できることを理解されたい。

【００５９】

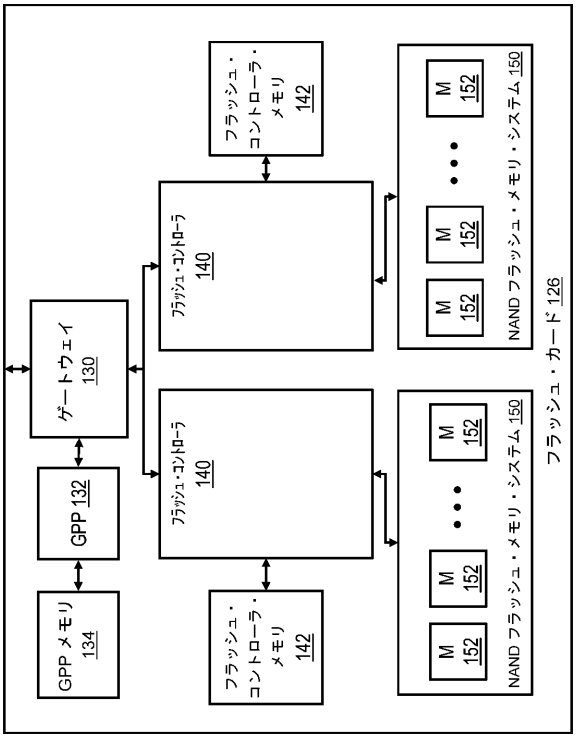
上記で説明した図面および以下の特定の構造および機能についての明細書は、出願人が発明したものの範囲、または添付の特許請求の範囲を制限するために提示していない。むしろ、図面および明細書は、特許保護を求める発明を作成および使用することを当業者に教示するために提供している。当業者は、明確さおよび理解のために、本発明の商用の実施形態の全ての特徴を説明または図示しているわけではないことを理解するであろう。当業者はまた、本発明の態様を組み込んだ実際の商用の実施形態の開発には、商用の実施形態に関する開発者の最終目標を達成するために多数の実装固有の判断が必要となることを理解するであろう。そのような実装固有の判断には、システム関連の、ビジネス関連の、政府関連の、およびその他の制約への準拠が含まれ得るが、おそらくこれらに限定されることはなく、これらの制約は、特定の実装形態、場所、およびその時々によって異なり得る。開発者の努力は絶対的な意味で複雑で時間がかかるものであり得るが、それにもかかわらず、そのような努力は、本開示の利益を得る当業者にとっては日常的な作業であろう。本明細書に開示および教示した発明は、多数の様々な修正および代替形態が可能であることを理解されたい。最後に、「ａ」などであるがこれに限定されない単数形の使用は、項目の数を限定することを意図したものではない。

【図面】

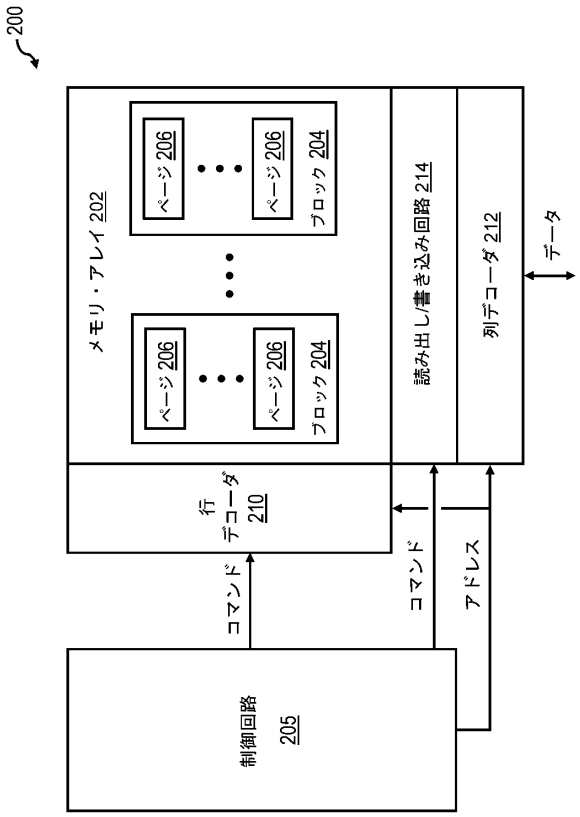
【図１Ａ】



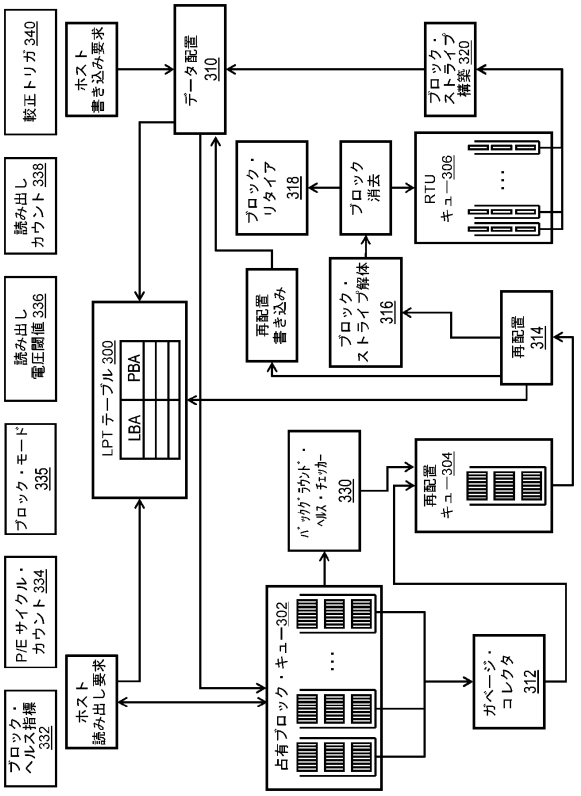
【図１Ｂ】



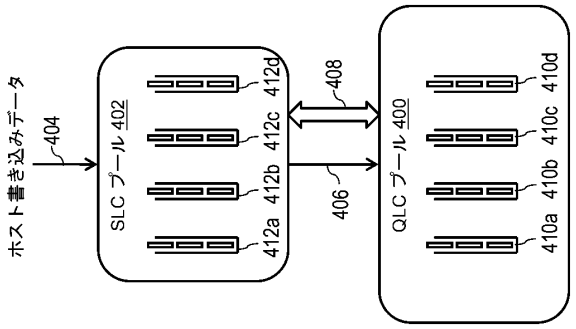
【図２】



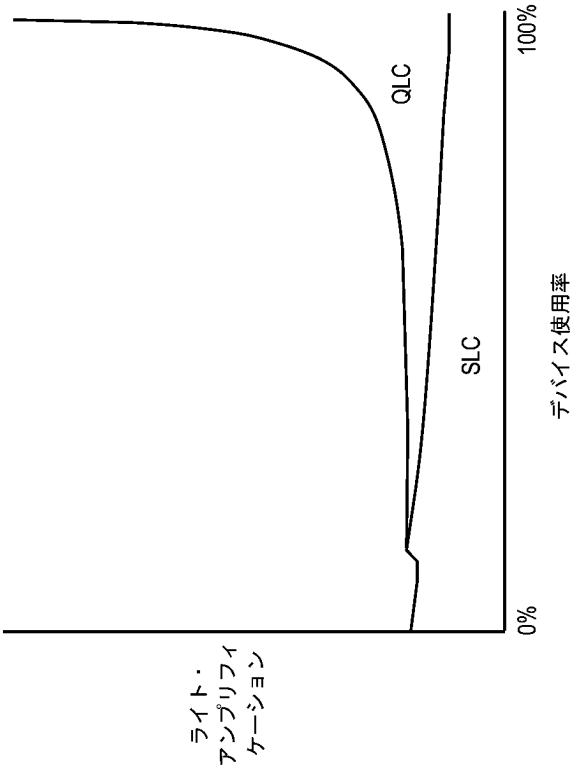
【図３】



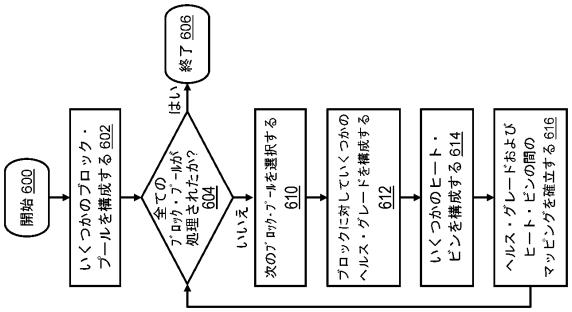
【図４】


ホスト書き込みデータ
404
SLC プール 402
412a
412b
412c
412d
408
406
QLC プール 400
410a
410b
410c
410d

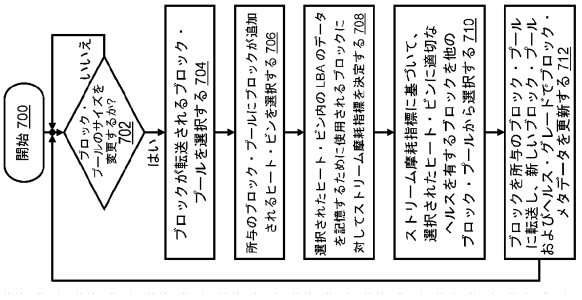
【図５】


100%
0%
QLC
SLC
デバイス使用率
ライト・アンプリフィケーション

【図６】


開始 600
いくつかのブロック・プールを構成する 602
全てのブロック・プールが処理されたか? 604
はい
終了 606
いいえ
次のブロック・プールを選択する 610
ブロックに対していくつかのヘルス・グレードを構成する 612
いくつかのヒート・ビンを構成する 614
ヘルス・グレードおよびヒート・ビンの間のマッピングを確立する 616

【図７】


開始 700
ブロック・プールのサイズを変更するか? 702
いいえ
はい
ブロックが転送されるブロック・プールを選択する 704
所与のブロック・プールにブロックが追加されるヒート・ビンを選択する 706
選択されたヒート・ビン内のLBAのデータを記憶するために使用されるブロックに対してストリーム摩耗指標を決定する 708
ストリーム摩耗指標に基づいて、選択されたヒート・ビンに適切なヘルスを有するブロックを他のブロック・プールから選択する 710
ブロックを所与のブロック・プールに転送し、新しいブロック・プールおよびヘルス・グレードでブロック・メタデータを更新する 712

フロントページの続き

(72)発明者　トミック、ササ
アメリカ合衆国９５１２０－６０９９ カリフォルニア州サンノゼ アルマデン・リサーチ・センター ハリー・ロード６５０
(72)発明者　プレトカ、ローマン、アレクサンダー
スイス８８０３ リュシュリコン ゾイマーシュトラーセ ４
(72)発明者　イオアンノウ、ニコラス
スイス８８０３ リュシュリコン ゾイマーシュトラーセ ４
(72)発明者　パパンドレウ、ニコラス
スイス８８０３ リュシュリコン ゾイマーシュトラーセ ４
(72)発明者　フライ、アーロン、ダニエル
アメリカ合衆国７７０４２－３４７８ テキサス州ヒューストン ウェステイマー・ロード１０７７７
(72)発明者　フィッシャー、ティモシー
アメリカ合衆国７７０４２－３４７８ テキサス州ヒューストン ウェステイマー・ロード１０７７７
(72)発明者　ストイカ、ラドゥ、イオアン
スイス８８０３ リュシュリコン ゾイマーシュトラーセ ４

審査官　松平 英
(56)参考文献　米国特許出願公開第２０１９／０１２１７２５（ＵＳ，Ａ１）
特開２０１３－１１４６７９（ＪＰ，Ａ）
特開２０１５－２１５６７８（ＪＰ，Ａ）
特開２０１８－１２０４３９（ＪＰ，Ａ）
米国特許出願公開第２０１２／０２４００１２（ＵＳ，Ａ１）
米国特許出願公開第２０１３／０１２４７８７（ＵＳ，Ａ１）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ３／０６－３／０８
１１／０８－１１／１０
１２／００－１２／１２８
１２／１６
１３／１６－１３／１８