(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2008-226149
(P2008-226149A)
(43)公開日　平成20年9月25日(2008.9.25)

| (51)Int.Cl. | | F I | | | テーマコード（参考） |
|---|---|---|---|---|---|
| G06F 3/06 | (2006.01) | G06F | 3/06 | 302A | 5B005 |
| G06F 12/16 | (2006.01) | G06F | 12/16 | 310J | 5B018 |
| G06F 12/00 | (2006.01) | G06F | 12/16 | 310Q | 5B065 |
| G06F 12/08 | (2006.01) | G06F | 12/16 | 310R | 5B082 |
| | | G06F | 12/00 | 542L | |

審査請求 未請求　請求項の数 17　ＯＬ　（全 30 頁）　最終頁に続く

(21)出願番号　特願2007-67142(P2007-67142)
(22)出願日　平成19年3月15日(2007.3.15)

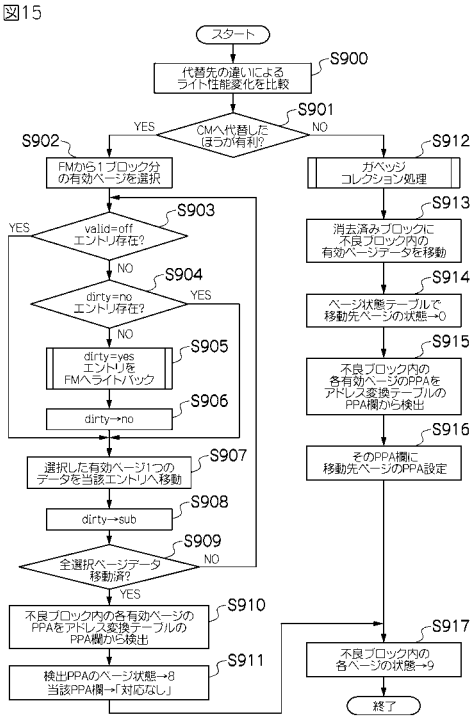
(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100079108
弁理士　稲葉　良幸
(74)代理人　100093861
弁理士　大賀　眞司
(72)発明者　水島　永雅
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
(72)発明者　中村　崇仁
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内

最終頁に続く

(54)【発明の名称】ストレージシステム及びストレージシステムのライト性能低下防止方法

(57)【要約】
【解決課題】ストレージシステムのライト性能の低下を抑制する。
【解決手段】フラッシュメモリと、キャッシュメモリと、フラッシュメモリのデータの読み出し、書き込み及び消去と、キャッシュメモリのデータの読み出し及び書き込みを制御し、フラッシュメモリ内に不良なブロックが発生したことを検出するコントローラと、データのライト処理を要求するコマンドを発行するホスト計算機とを含むストレージシステムにおいて、コントローラは、フラッシュメモリ内に不良ブロックが発生したことを検出すると、フラッシュメモリに格納された所定のデータをキャッシュメモリに移動し、その移動したデータを更新するためのコマンドをホスト計算機から受信しても、そのコマンドに基づくデータをフラッシュメモリへ書き込むことを禁止する。
【選択図】図１５

図15



【特許請求の範囲】
【請求項１】
ページ単位でデータを書き込み、複数の前記ページから構成されるブロックを単位として前記データを消去するとともに複数の前記ブロックを有し、前記データの更新に前記ページを含む前記ブロックの消去を必要とするフラッシュメモリと、
前記フラッシュメモリに書き込むべきデータを前記フラッシュメモリよりも高速に書き込むとともに一時的に記憶するキャッシュメモリと、
前記フラッシュメモリのデータの読み出し、書き込み及び消去と、前記キャッシュメモリのデータの読み出し及び書き込みを制御し、前記フラッシュメモリ内に不良な前記ブロックが発生したことを検出するコントローラと、
前記データのライト処理を要求するコマンドを発行するホスト計算機とを備え、
前記コントローラは、
前記フラッシュメモリ内に前記不良ブロックが発生したことを検出したときに、前記フラッシュメモリに格納された所定のデータを前記キャッシュメモリに移動し、その移動したデータを更新するためのコマンドを前記ホスト計算機から受信しても、そのコマンドに基づくデータを前記フラッシュメモリへ書き込むことを禁止する、
ことを特徴とするストレージシステム。
【請求項２】
請求項１記載のストレージシステムであって、
前記移動したデータを更新するためのコマンドに基づいて、前記キャッシュメモリ内に移動したデータを更新する、
ことを特徴とするストレージシステム。
【請求項３】
請求項１に記載のストレージシステムであって、
前記キャッシュメモリに記憶する前記データを管理するキャッシュ管理テーブルを備え、
前記キャッシュ管理テーブルは、前記データを前記フラッシュメモリへ書き込むことを禁止する禁止情報を保持する、
ことを特徴とするストレージシステム。
【請求項４】
請求項１に記載のストレージシステムであって、
前記フラッシュメモリに記憶するデータの論理ページアドレスと物理ページアドレスとの対応関係を管理するアドレス変換テーブルを備え、
前記アドレス変換テーブルは、前記データの論理ページアドレスに対応する物理ページアドレスが存在しないことを表すアドレス不存在情報を保持する、
ことを特徴とする前記ストレージシステム。
【請求項５】
請求項１に記載のストレージシステムであって、
前記コントローラは、
前記ホスト計算機のコマンドに基づいてデータを前記キャッシュメモリに書き込むときに前記データの論理ページアドレスと同じ論理ページアドレスの前記データが前記キャッシュメモリに存在する確率を示すヒット率を計算し、
前記不良ブロックが発生したことを検出したときに、その検出時までに計算した前記ヒット率に基づいて、前記不良ブロックの代替先として前記フラッシュメモリ内のその他のブロックを適用したときの前記ストレージシステムの第１のライト性能と、前記フラッシュメモリに格納された前記データの一部を前記キャッシュメモリに移動したときの前記ストレージシステムの第２のライト性能とを計算し、その計算結果に基づいて、前記第２のライト性能が前記第１のライト性能に勝ると判定したならば、前記フラッシュメモリに格納された前記データを前記キャッシュメモリに移動する、
ことを特徴とするストレージシステム。

【請求項６】
請求項１に記載のストレージシステムであって、
前記キャッシュメモリは不揮発性のランダムアクセスメモリである、
ことを特徴とするストレージシステム。
【請求項７】
請求項６に記載のストレージシステムであって、
前記不揮発性のランダムアクセスメモリは、相変化ＲＡＭである、
ことを特徴とするストレージシステム。
【請求項８】
請求項１に記載のストレージシステムであって、
前記キャッシュメモリに移動した前記データを、前記キャッシュメモリが冗長的に保持する、
ことを特徴とするストレージシステム。
【請求項９】
請求項８に記載のストレージシステムであって、
前記キャッシュメモリとは、異なるキャッシュメモリを備え、
前記冗長的な保持は、前記キャッシュメモリへ移動した前記データを、前記異なるキャッシュメモリへコピーして行なう、
ことを特徴とするストレージシステム。
【請求項１０】
請求項８に記載のストレージシステムであって、
前記キャッシュメモリとは、異なる複数のキャッシュメモリを備え、
前記冗長的な保持は、前記キャッシュメモリへ移動した前記データを、前記異なる複数のキャッシュメモリとＲＡＩＤグループを組んで行なう、
ことを特徴とするストレージシステム。
【請求項１１】
ページ単位でデータを書き込み、複数の前記ページから構成されるブロックを単位として前記データを消去するとともに複数の前記ブロックを有し、前記データの更新に前記ページを含む前記ブロックの消去を必要とするフラッシュメモリと、
前記フラッシュメモリに書き込むべきデータを前記フラッシュメモリよりも高速に書き込むとともに一時的に記憶するキャッシュメモリと、
前記フラッシュメモリのデータの読み出し、書き込み及び消去と、前記キャッシュメモリのデータの読み出し及び書き込みを制御し、前記フラッシュメモリ内に不良な前記ブロックが発生したことを検出するコントローラと、
前記データのライト処理を要求するコマンドを発行するホスト計算機とを含むストレージシステムのライト性能低下防止方法において、
前記コントローラは、
前記フラッシュメモリ内に前記不良ブロックが発生したことを検出し、
前記フラッシュメモリに格納された所定のデータを前記キャッシュメモリに移動し、
その移動したデータを更新するためのコマンドを前記ホスト計算機から受信しても、そのコマンドに基づくデータを前記フラッシュメモリへ書き込むことを禁止する、
ことを特徴とするストレージシステムのライト性能低下防止方法。
【請求項１２】
請求項１１に記載のストレージシステムのライト性能低下防止方法であって、
前記キャッシュメモリに記憶する前記データを管理するキャッシュ管理テーブルにより保持される禁止情報に基づいて前記データを前記フラッシュメモリへ書き込むことを禁止する、
ことを特徴とする方法。
【請求項１３】
請求項１１記載のストレージシステムのライト性能低下防止方法であって、

前記コントローラは、
前記ホスト計算機のコマンドに基づいてデータを前記キャッシュメモリに書き込むときに前記データの論理ページアドレスと同じ論理ページアドレスの前記データが前記キャッシュメモリに存在する確率を示すヒット率を計算し、
前記不良ブロックが発生したことを検出したときに、その検出時までに計算した前記ヒット率に基づいて、前記不良ブロックの代替先として前記フラッシュメモリ内のその他のブロックを適用したときの前記ストレージシステムの第１のライト性能と、前記フラッシュメモリに格納された前記データの一部を前記キャッシュメモリに移動したときの前記ストレージシステムの第２のライト性能とを計算し、その計算結果に基づいて、前記第２のライト性能が前記第１のライト性能に勝ると判定したならば、前記フラッシュメモリに格納された前記データを前記キャッシュメモリに移動する、
ことを特徴とする方法。
【請求項１４】
請求項１１記載のストレージシステムのライト性能低下防止方法であって、
前記キャッシュメモリは不揮発性のランダムアクセスメモリである、
ことを特徴とする方法。
【請求項１５】
請求項１１記載のストレージシステムのライト性能低下防止方法であって、
前記キャッシュメモリに移動した前記データを、前記キャッシュメモリが冗長的に保持する、
ことを特徴とする方法。
【請求項１６】
請求項１５に記載のストレージシステムのライト性能低下防止方法であって、
前記冗長的な保持は、前記キャッシュメモリへ移動した前記データを、前記キャッシュメモリとは、異なるキャッシュメモリへコピーして行なう、
ことを特徴とする方法。
【請求項１７】
請求項１５に記載のストレージシステムのライト性能低下防止方法であって、
異なる複数のキャッシュメモリを備え、
前記冗長的な保持は、前記キャッシュメモリへ移動した前記データを、前記キャッシュメモリとは異なる複数のキャッシュメモリとＲＡＩＤグループを組んで行なう、
ことを特徴とする方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、電気的に書き換え可能な不揮発性メモリを使用したストレージシステム及びストレージシステムのライト性能低下防止方法に関し、特に不揮発性メモリとしてフラッシュメモリを用い、フラッシュメモリよりも高速なランダムアクセスメモリをキャッシュメモリとして搭載したストレージシステムにおいて、不良化したフラッシュメモリブロックの代替にキャッシュメモリを使用するものに適用しても好適なものである。
【背景技術】
【０００２】
フラッシュメモリは直接的に上書きのできない不揮発性メモリであり、書き換えを実施するためには、複数個の書き込み単位（ページ）から構成される消去単位（ブロック）を消去し、それらのページを未書き込み状態にする必要がある。
【０００３】
したがって、フラッシュメモリを主たる記憶媒体とする一般的なストレージシステムでは、ホスト装置が格納データを更新したい場合、フラッシュメモリの一部にあらかじめ確保した予備領域に新データを書いて有効データとする一方で旧データを無効データとし、後に旧データを含むブロックを消去して新たな予備領域とする、といった制御を行う。な

お、消去するブロック内にその他の有効データが残っている場合は、消去の前にそれらを別のブロックの未書き込みのページに退避しなければならない。

【０００４】

一般に、フラッシュメモリ上の予備領域は他のブロックが不良化したときの代替領域としても利用される。フラッシュメモリのブロックは１０万回程度の書き換え寿命しか保証されてないため、ホスト装置がストレージシステムの格納データの更新を繰り返していると日常的に不良ブロックが発生し、その数は徐々に増加する。そして、これが予備領域を満たすほどの数になると、上記の格納データ更新制御が困難になり、そのストレージシステムは書き換え不能となる。

【０００５】

特許文献１には、フラッシュメモリを記憶媒体とするデータファイルストレージシステムにおいて、キャッシュメモリを使用した以下の手続きによって、ストレージシステムの書き換え寿命を延長する方法が記載されている。（１）ホスト装置からのフラッシュメモリ向けのデータファイルを、フラッシュメモリよりかなり多くのライト・消去のサイクルに耐えうるキャッシュメモリに一時保管する。（２）ホスト装置からのライト要求に応じて、新データファイルをフラッシュメモリの代わりにキャッシュメモリに書く。（３）データファイルの識別子および各データファイルが最後にキャッシュメモリに書かれてからの時間を、タグメモリに格納する。（４）キャッシュメモリに新データファイル用の追加空間が必要になったら、タグメモリの参照によって最後の書込から最長の時間がたったデータファイルを優先してキャッシュメモリからフラッシュメモリに移動する。上記（１）～（４）の処理を行うことで、フラッシュメモリへの実際のライト回数や関連ストレスを大きく減らしている。

【特許文献１】米国特許第５９３６９７１号公報

【発明の開示】

【発明が解決しようとする課題】

【０００６】

フラッシュメモリを記憶媒体とするストレージシステムにおいて、フラッシュメモリの不良ブロックの数の増加に伴って、上記の格納データ更新制御に使用する予備領域のサイズは徐々に減少する。正味の格納データ量に対して更新制御用の予備領域サイズの割合が小さいほど、ストレージシステムの格納データ更新時の作業効率（ホスト装置の書き換え単位あたりのフラッシュメモリのライト回数）は低下する。これは、上記の格納データ更新制御において、消去ブロック内に残された他の有効データを退避するときの平均作業量が多くなることが原因である。

【０００７】

その結果、不良ブロックの数が増加するにつれて、ストレージシステムのライト性能は徐々に低下するという問題があった。

【０００８】

本発明は、以上の点を考慮してなれたもので、ストレージシステムのライト性能の低下を抑制することができるストレージシステム及びストレージシステムのライト性能低下防止方法を提案しようとするものである。

【課題を解決するための手段】

【０００９】

本発明は、ページ単位でデータを書き込み、複数の前記ページから構成されるブロックを単位として前記データを消去するとともに複数の前記ブロックを有し、前記データの更新に前記ページを含む前記ブロックの消去を必要とするフラッシュメモリと、前記フラッシュメモリに書き込むべきデータを前記フラッシュメモリよりも高速に書き込むとともに一時的に記憶するキャッシュメモリと、前記フラッシュメモリのデータの読み出し、書き込み及び消去と、前記キャッシュメモリのデータの読み出し及び書き込みを制御し、前記フラッシュメモリ内に不良な前記ブロックが発生したことを検出するコントローラと、前記データのライト処理を要求するコマンドを発行するホスト計算機とを備え、前記コント

ローラは、前記フラッシュメモリ内に前記不良ブロックが発生したことを検出したときに、前記フラッシュメモリに格納された所定のデータを前記キャッシュメモリに移動し、その移動したデータを更新するためのコマンドを前記ホスト計算機から受信しても、そのコマンドに基づくデータを前記フラッシュメモリへ書き込むことを禁止するストレージシステムである。

【００１０】

すなわち、ページ単位でデータを書き込み、複数のページから構成されるブロックを単位としてデータを消去するとともに複数のブロックを有し、データの更新にページを含むブロックの消去を必要とするフラッシュメモリと、フラッシュメモリに書き込むべきデータをフラッシュメモリよりも高速に書き込むとともに一時的に記憶するキャッシュメモリと、フラッシュメモリのデータの読み出し、書き込み及び消去と、キャッシュメモリのデータの読み出し及び書き込みを制御し、フラッシュメモリ内に不良なブロックが発生したことを検出するコントローラと、データのライト処理を要求するコマンドを発行するホスト計算機とを含むストレージシステムのライト性能低下防止方法において、コントローラは、フラッシュメモリ内に不良ブロックが発生したことを検出し、前記キャッシュメモリに格納された所定のデータをキャッシュメモリに移動し、その移動したデータを更新するためのコマンドをホスト計算機から受信しても、そのコマンドに基づくデータをフラッシュメモリへ書き込むことを禁止することにより、フラッシュメモリの不良ブロック数の増加に伴って格納データ更新時の作業効率が低下するのを抑制することができるので、キャッシュメモリのヒット率が低い状況下において、従来よりもストレージシステムのライト性能の低下を抑制することができる。

【００１１】

また、上記ストレージシステムにおけるキャッシュメモリを不揮発性のランダムアクセスメモリ、例えば、相変化ＲＡＭ（Random Access Memory）とすることにより、ラッシュメモリからキャッシュメモリに代替されたデータを補助電源なしに保持することができるため、ストレージシステムの電源消費電力を削減し、データを突然の電源遮断などの障害による消失から保護するという効果を奏する。

【発明の効果】

【００１２】

本発明によれば、ストレージシステムのライト性能の低下を抑制するという効果を奏する。

【発明を実施するための最良の形態】

【００１３】

以下、本発明の実施形態について図面を参照しながら説明する。

【００１４】

（１）ストレージシステムの構成

【００１５】

図１は、本発明を適用したストレージシステム１０のハードウェア構成を簡単に示す図である。

【００１６】

ストレージシステム１０は、ストレージコントローラ１２０及びフラッシュメモリ・モジュール（ＦＭＭ）１５１～１５４、１６１～１６４、１７１～１７４、１８１～１８４を備える。ストレージコントローラ１２０は、チャネルアダプタ１２１、１２２、キャッシュメモリ１２３、１２４、ストレージアダプタ１２５、１２６、共有メモリ１２９及び相互接続網１２７、１２８を備える。

【００１７】

なお、ストレージコントローラ１２０は、チャネルアダプタ１２１、１２２、キャッシュメモリ１２３、１２４及びストレージアダプタ１２５、１２６、共有メモリ１２９を有する場合を例示しているが、それらの個数は限定されるものではない

【００１８】

相互接続網１２７及び１２８は、例えば、スイッチ等であり、ストレージコントローラ１２０を構成する各装置を相互に接続する。詳細には、相互接続網１２７及び１２８は、チャネルアダプタ１２１、キャッシュメモリ１２３及びストレージアダプタ１２５、共有メモリ１２９を相互に接続する。同様に、相互接続網１２７、１２８は、チャネルアダプタ１２２、キャッシュメモリ１２４及びストレージアダプタ１２６、共有メモリ１２９を相互に接続する。

【００１９】

チャネルアダプタ１２１は、チャネル１１０、１１１、１１２、１１３を介してホスト計算機１００に接続されている。同様に、チャネルアダプタ１２２は、チャネル１１４、１１５、１１６、１１７を介してホスト計算機１００に接続されている。ホスト計算機１００は、パーソナルコンピュータ、ワークステーション、メインフレームコンピュータ等の計算機であり、ストレージシステム１０において、ストレージコントローラ１２０にデータの読み書きを要求する。ストレージコントローラ１２０は、チャネルアダプタ１２１、１２２を用いてそれらの要求を解釈し、それらの要求を満たすためにストレージアダプタ１２５、１２６を用いてフラッシュメモリ・モジュール１５１～１５４、１６１～１６４、１７１～１７４、１８１～１８４のデータを読み書きする。

【００２０】

その際、キャッシュメモリ１２３、１２４は、チャネルアダプタ１２１、１２２又はストレージアダプタ１２５、１２６から受信したデータを一時的に記憶したり、必要に応じて特定の受信データを永続的に記憶したりする。キャッシュメモリ１２３、１２４は、例えば、ダイナミック型ランダムアクセスメモリであり、高速に読み書きできる。共有メモリ１２９は、キャッシュメモリ１２３、１２４の記憶データを管理するためのテーブルを格納し、チャネルアダプタ１２１、１２２又はストレージアダプタ１２５、１２６がそれを参照・設定することができる。共有メモリ１２９は、例えば、ダイナミック型ランダムアクセスメモリであり、高速に読み書きできる。

【００２１】

ストレージアダプタ１２５は、フラッシュメモリ・モジュール１５１～１５４、１６１～１６４、１７１～１７４、１８１～１８４に接続されている。詳細には、ストレージアダプタ１２５は、チャネル１４０を介してフラッシュメモリ・モジュール１５１～１５４に接続されている。また、ストレージアダプタ１２５は、チャネル１４１を介してフラッシュメモリ・モジュール１６１～１６４に接続されている。また、ストレージアダプタ１２５は、チャネル１４２を介してフラッシュメモリ・モジュール１７１～１７４に接続されている。また、ストレージアダプタ１２５は、チャネル１４３を介してフラッシュメモリ・モジュール１８１～１８４に接続されている。

【００２２】

同様に、ストレージアダプタ１２６は、フラッシュメモリ・モジュール１５１～１５４、１６１～１６４、１７１～１７４、１８１～１８４に接続されている。詳細には、ストレージアダプタ１２６は、チャネル１４４を介してフラッシュメモリ・モジュール１５１～１５４に接続されている。また、ストレージアダプタ１２６は、チャネル１４５を介してフラッシュメモリ・モジュール１６１～１６４に接続されている。また、ストレージアダプタ１２６は、チャネル１４６を介してフラッシュメモリ・モジュール１７１～１７４に接続されている。また、ストレージアダプタ１２６は、チャネル１４７を介してフラッシュメモリ・モジュール１８１～１８４に接続されている。

【００２３】

チャネルアダプタ１２１、１２２及びストレージアダプタ１２５、１２６は、保守端末１３０に接続されている。保守端末１３０は、ストレージシステム１０の管理者から入力された設定情報を、チャネルアダプタ１２１、１２２及び／又はストレージアダプタ１２５、１２６に送信する。

【００２４】

なお、ストレージシステム１０は、ストレージアダプタ１２５及びチャネルアダプタ１

２１に替わって一つのアダプタを備えていても良い。この場合、その一つのアダプタが、ストレージアダプタ１２５及びチャネルアダプタ１２１の処理を行う。

【００２５】

１９０～１９３は、ＲＡＩＤ（Redundant Arrays of Inexpensive Disks）グループである。例えば、ＲＡＩＤグループ１９０は、フラッシュメモリ・モジュール１５１、１６１、１７１、１８１から成る。ＲＡＩＤグループ１９０に属するフラッシュメモリ・モジュールの一つ、例えば、フラッシュメモリ・モジュール１５１でエラーが発生してデータを読み出せなくなると、ＲＡＩＤグループ１９０に属する他のフラッシュメモリ・モジュール１６１、１７１、１８１からデータを再生できる。

【００２６】

図２はチャネルアダプタ１２１のハードウェア構成を示す図である。チャネルアダプタ１２１はホストチャネル・インタフェース２１４、キャッシュメモリ・インタフェース２１５、ネットワーク・インタフェース２１１、プロセッサ２１０、ローカルメモリ２１３、及びプロセッサ周辺制御部２１２を備える。

【００２７】

ホストチャネル・インタフェース２１４は、チャネルアダプタ１２１を、チャネル１１０、１１１、１１２、１１３を介してホスト計算機１００に接続するためのインタフェースである。ホストチャネル・インタフェース２１４は、チャネル１１０、１１１、１１２、１１３上のデータ転送プロトコルと、ストレージコントローラ１２０内部のデータ転送プロトコルとを相互に変換する。

【００２８】

キャッシュメモリ・インタフェース２１５は、チャネルアダプタ１２１を、相互結合網１２７、１２８に接続するためのインタフェースである。

【００２９】

ネットワーク・インタフェース２１１は、チャネルアダプタ１２１を、保守端末１３０に接続するためのインタフェースである。

【００３０】

なお、ホストチャネル・インタフェース２１４とキャッシュメモリ・インタフェース２１５とは、信号線２１６によって接続されている。

【００３１】

プロセッサ２１０は、ローカルメモリ２１３に記憶されているプログラムを実行することによって各種処理を行う。より詳細には、プロセッサ２１０は、ホスト計算機１００と相互結合網１２７、１２８との間のデータ転送を制御する。

【００３２】

ローカルメモリ２１３は、プロセッサ２１０によって実行されるプログラムを記憶する。また、ローカルメモリ２１３は、プロセッサ２１０によって参照されるテーブルを記憶する。プロセッサ２１０によって参照されるテーブルは、チャネルアダプタ１２１の動作を制御するための設定情報を含み、管理者によって設定又は変更される。この場合、管理者は、テーブルの設定又はテーブルの変更に関する情報を、保守端末１３０に入力する。保守端末１３０は、入力された情報を、ネットワーク・インタフェース２１１を介してプロセッサ２１０に送信する。プロセッサ２１０は、受信した情報に基づいて、テーブルを作成又は変更する。そして、プロセッサ２１０は、テーブルをローカルメモリ２１３に格納する。

【００３３】

プロセッサ周辺制御部２１２は、ホストチャネル・インタフェース２１４、キャッシュメモリ・インタフェース２１５、ネットワーク・インタフェース２１１、プロセッサ２１０、及びローカルメモリ２１３の間におけるデータ転送を制御する。プロセッサ周辺制御部２１２は、例えば、チップセット等である。

【００３４】

なお、チャネルアダプタ１２２のハードウェア構成は、チャネルアダプタ１２１のハー

ドウェア構成と同一であるので、チャネルアダプタ１２２のハードウェア構成についての説明を省略する。

【００３５】

図３はストレージアダプタ１２５のハードウェア構成を示す図である。ストレージアダプタ１２５は、キャッシュメモリ・インタフェース２２４、ストレージチャネルインタフェース２２５、ネットワーク・インタフェース２２１、プロセッサ２２０、ローカルメモリ２２３、及びプロセッサ周辺制御部２２２を備える。

【００３６】

キャッシュメモリ・インタフェース２２４は、ストレージアダプタ１２５を、相互結合網１２７、１２８に接続するためのインタフェースである。

【００３７】

ストレージチャネルインタフェース２２５は、ストレージアダプタ１２５を、チャネル１４０、１４１、１４２、１４３に接続するためのインタフェースである。ストレージチャネルインタフェース２２５は、チャネル１４０、１４１、１４２、１４３上のデータ転送プロトコルと、ストレージコントローラ１２０内部のデータ転送プロトコルとを相互に変換する。

【００３８】

なお、キャッシュメモリ・インタフェース２２４とストレージチャネルインタフェース２２５とは、信号線２２６によって接続されている。

【００３９】

ネットワーク・インタフェース２２１は、ストレージアダプタ１２５を、保守端末１３０に接続するためのインタフェースである。

【００４０】

プロセッサ２２０は、ローカルメモリ２２３に記憶されているプログラムを実行することによって各種処理を行う。

【００４１】

ローカルメモリ２２３は、プロセッサ２２０によって実行されるプログラムを記憶する。また、ローカルメモリ２２３は、プロセッサ２２０によって参照されるテーブルを記憶する。プロセッサ２２０によって参照されるテーブルは、ストレージアダプタ１２５の動作を制御するための設定情報を含み、管理者によって設定又は変更される。この場合、管理者は、テーブルの設定又はテーブルの変更に関する情報を、保守端末１３０に入力する。保守端末１３０は、入力された情報を、ネットワーク・インタフェース２２１を介してプロセッサ２２０に送信する。プロセッサ２２０は、受信した情報に基づいて、テーブルを作成又は変更する。そして、プロセッサ２２０はテーブルをローカルメモリ２２３に格納する。

【００４２】

プロセッサ周辺制御部２２２は、キャッシュメモリ・インタフェース２２４、ストレージチャネルインタフェース２２５、ネットワーク・インタフェース２２１、プロセッサ２２０及びローカルメモリ２２３の間のデータ転送を制御する。プロセッサ周辺制御部２２２は、例えば、チップセット等である。

【００４３】

なお、ストレージアダプタ１２６のハードウェア構成は、ストレージアダプタ１２５のハードウェア構成と同一であるので、ストレージアダプタ１２６のハードウェア構成についての説明を省略する。

【００４４】

図４はフラッシュメモリ・モジュール１５１のハードウェア構成を示す図である。フラッシュメモリ・モジュール１５１は、メモリコントローラ３１０及びフラッシュメモリ３２０を備える。フラッシュメモリ３２０は、データを記憶する。メモリコントローラ３１０は、フラッシュメモリ３２０のデータの「読み出し」、「書き込み」、及び「消去」を制御する。

【００４５】
メモリコントローラ３１０は、プロセッサ３１２、インタフェース３１１、データ転送部３１５、ＲＡＭ３１３、及びＲＯＭ３１４を備える。フラッシュメモリ３２０は、複数のフラッシュメモリ・チップ３２１を備える。
【００４６】
図５はフラッシュメモリ・チップ３２１の内部構成を示す図である。フラッシュメモリ・チップ３２１は、複数のブロック３３０を含み、それぞれのブロック３３０にデータを記憶する。ブロック３３０は、メモリコントローラ３１０がデータを消去する単位である。ブロック３３０は、複数のページ３４０を含む。ページ３４０は、メモリコントローラ３１０がデータを読み書きする単位である。フラッシュメモリ３２０では、１ページあたり２０μｓ程度の時間でデータを読み出し、１ページあたり０．２ｍｓ程度の時間でデータを書き込む。また、１ブロックあたり１．５ｍｓ程度の時間でデータを消去する。フラッシュメモリ３２０へのページ書き込みにかかる時間は、キャッシュメモリ１２３、１２４への同サイズのデータの書き込みにかかる時間よりも長い。なお、書き込みや消去はメモリセルを徐々に劣化させ、多数回（例えば、数１０万回）の書き換えを行うとエラーが発生することがある。
【００４７】
ページ３４０は、メモリコントローラ３１０により有効ページ、無効ページ、未書込ページ、又は不良ページの何れかに分類される。有効ページは、ストレージシステム１０として格納しておく必要がある有効なデータを記憶しているページ３４０である。無効ページは、ストレージシステム１０として格納する必要がなくなった無効なデータ（ガベッジ）を記憶しているページ３４０である。未書込ページは、所属するブロック３３０が消去されて以来データを記憶していないページ３４０である。不良ページは、ページ３４０内の記憶素子が壊れている等の理由によって、物理的に書き換えできないページ３４０である。ページ３４０が不良ページとなる要因には次の３つがある。
【００４８】
１つ目は、チップ製造段階の検査で不合格になること。２つ目は、ページ３４０の書込においてエラーが発生すること。なお、当該不良ページは、それ以降は読み出しのみ可能となる。このページを１つでも含むブロック３３０は不良ブロックと呼ばれ、ブロック消去やページ書込が禁止される。３つ目は、ブロック３３０の消去においてエラーが発生すること。なお、当該ブロック内の全ページは不良ページとなる。当該ブロックは不良ブロックと呼ばれ、ブロック消去やページ書込が禁止される。
【００４９】
インタフェース３１１は、チャネル１４０を介してストレージコントローラ１２０内のストレージアダプタ１２５に接続されている。また、インタフェース３１１は、チャネル１４４を介してストレージコントローラ１２０内のストレージアダプタ１２６に接続されている。インタフェース３１１は、ストレージアダプタ１２５及びストレージアダプタ１２６からの命令を受信する。ストレージアダプタ１２５及びストレージアダプタ１２６からの命令は、例えば、ＳＣＳＩコマンドである。
【００５０】
詳細には、インタフェース３１１は、ストレージアダプタ１２５及びストレージアダプタ１２６からデータを受信する。そして、インタフェース３１１は、受信したデータをＲＡＭ３１３にバッファする。また、インタフェース３１１は、ＲＡＭ３１３にバッファされているデータを、ストレージアダプタ１２５及びストレージアダプタ１２６へ送信する。
【００５１】
また、インタフェース３１１は、ハードディスクドライブとの互換性を有するインタフェース機能を有している。そのため、ストレージアダプタ１２５、１２６は、フラッシュメモリ・モジュール１５１～１８４を、ハードディスクドライブとして認識する。ストレージシステム１０は、データを格納するための記録媒体として、フラッシュメモリ・モジ

ュールとハードディスクドライブとを混載してもよい。

【００５２】

ＲＡＭ３１３は、例えば、ダイナミック型ランダムアクセスメモリであり、高速に読み書きできる。ＲＡＭ３１３は、インタフェース３１１が送受信するデータを一時的に記憶する。一方、ＲＯＭ３１４は、不揮発性メモリであり、プロセッサ３１２によって実行されるプログラムを記憶する。プロセッサ３１２によって実行されるプログラムは、プロセッサ３１２が実行可能となるように、ストレージシステム１０の起動時にＲＯＭ３１４からＲＡＭ３１３へロードされる。また、ＲＡＭ３１３は、プロセッサ３１２によって参照される管理情報を記憶する。

【００５３】

プロセッサ３１２によって参照される管理情報は、フラッシュメモリ３２０の論理ページアドレスと物理ページアドレスとを変換するためのアドレス変換テーブルを含む。論理ページアドレスは、フラッシュメモリ・モジュール１５１の外部から（例えば、ストレージアダプタ１２５から）、フラッシュメモリ３２０に読み書きする単位であるページを論理的に指示するためのアドレスである。物理ページアドレスは、メモリコントローラ３１０が、フラッシュメモリ３２０を読み書きする単位であるページを物理的にアクセスするためのアドレスである。プロセッサ３１２は、ページ対応関係の変化に応じてアドレス変換テーブルの内容を書き換える。なお、アドレス変換テーブルの具体的な例については口授する。

【００５４】

さらに、この管理情報は、フラッシュメモリ３２０の物理的なページ３４０の状態を管理するためのページ状態テーブルを含む。ページ状態テーブルは、予め定義されたページ状態を符号化したものを格納する。ページ状態は１６進数表記で以下の４通りを定義する。

・状態＝０…有効ページ
・状態＝８…無効ページ
・状態＝９…不良ページ
・状態＝Ｆ…未書込ページ

【００５５】

ページ状態テーブルはブロック単位で状態を保持する。例えば、あるブロックのページ状態が“８８０Ｆ”のとき、その第１・第２ページは無効データを含み、第３ページは有効データを含み、第４ページは未書込であることを意味する。プロセッサ３１２は、ページ状態の変化に応じてページ状態テーブルの内容を書き換える。なお、ページ情報テーブルの具体的な例については後述する。

【００５６】

データ転送部３１５は、例えばスイッチであり、プロセッサ３１２、インタフェース３１１、ＲＡＭ３１３、ＲＯＭ３１４、及びフラッシュメモリ３２０を相互に接続し、それらの間のデータ転送を制御する。

【００５７】

プロセッサ３１２は、ＲＡＭ３１３に記憶されているプログラムを実行することによって、各種処理を行う。例えば、プロセッサ３１２は、ＲＡＭ３１３に記憶されているアドレス変換テーブルを参照して、フラッシュメモリ３２０の論理ページアドレスとフラッシュメモリ３２０の物理ページアドレスとを変換し、フラッシュメモリ３２０にデータを読み書きする。また、プロセッサ３１２は、フラッシュメモリ・モジュール１５１内のブロック３３０について、ガベッジコレクション処理（ブロック再生処理）を行う。

【００５８】

ガベッジコレクション処理（ブロック再生処理）は、新しいデータを書くための未書込ページが少なくなったときに、未書込ページ数を増やすために、あるブロック３３０内の無効ページを未書込ページに再生する処理である。ガベッジコレクション処理の対象となるブロック（対象ブロック）３３０としては、最も多くの無効ページが存在しているブロ

ックを選択する。未書込ページを増加させるには、無効ページを消去する必要があるが、消去はブロック単位でしかできないため、プロセッサ３１２は、有効ページのデータを未書込ページのあるブロックに複写し、その後、対象ブロックを消去して、ブロックを再生する。このように、フラッシュメモリ・モジュール１５１に書き込まれたデータは、ストレージコントローラ１２０からの指示とは独立に、フラッシュメモリ・モジュール１５１内で移動することがある。メモリコントローラ３１０は、このデータ移動の結果を、アドレス変換テーブルやページ状態テーブルに正しく反映する。それにより、ストレージコントローラ１２０は、正しいデータをアクセスすることができる。

【００５９】

プロセッサ３１２は、フラッシュメモリ３２０におけるページ書き込みやガベッジコレクション処理を通じて変化する論理ページアドレスと物理ページアドレスの対応関係、ページ状態を、それぞれアドレス変換テーブル、ページ状態テーブルを用いて管理する。

【００６０】

なお、フラッシュメモリ・モジュール１５１のハードウェア構成について詳述したが、他のフラッシュメモリ・モジュール１５２～１８４についても、同様のハードウェア構成を有している。このため、他のフラッシュメモリもミュール１５２～１８４については、図示及び説明を省略する。

【００６１】

図５に示すように、各ブロックを構成する複数のページ３４０は、いずれもデータ部３５０及び冗長部３５１を含む。例えば、１ページにつき、ページ３４０は２１１２バイト、データ部３５０は２０４８バイト、冗長部３５１は６４バイトである。なお、本発明はそれらのページ・サイズを特に限定するものではない。

【００６２】

データ部３５０は、ユーザデータを記憶する。冗長部３５１は、ページ３４０自身に対応する論理ページアドレス、書込み時刻、及びエラー訂正コードを記憶する。論理ページアドレスは、ストレージシステム１０の起動時にＲＡＭ３１３上にアドレス変換テーブルを作成するときや、ガベッジコレクション処理のときに参照する。書込み時刻は、ストレージシステム１０の起動時にＲＡＭ３１３上にページ状態テーブルを作成するときに、ページ３４０が有効ページか無効ページであるかを知るために参照する。同じ論理ページアドレスが記録された複数のページが存在する場合、この時刻が最も遅いものが有効ページであり、それ以外は無効ページである。エラー訂正コードは、ページ３４０のエラーを検出及び訂正するための情報であり、例えば、ＢＣＨ（Bose-Chaud-huri-Hocquenghem）符号である。冗長部３５１は、通常、メモリコントローラ３１０のみがアクセス可能であり、ストレージアダプタ１２５、１２６からは、データ部３５０の内容のみがアクセス可能である。

【００６３】

（２）不良ブロックの代替方法の選択とその影響

【００６４】

図６～９は、フラッシュメモリ・チップ３２１を構成する複数のブロックの中に不良ブロックが発生したときの２種類のブロック代替方法と、それぞれがフラッシュメモリ・モジュール１５１～１５４、１６１～１６４、１７１～１７４及び１８１～１８４へのライト性能およびキャッシュメモリ１２３のヒット確率に及ぼす影響を説明するための図である。

【００６５】

説明を簡単化するため、一般的なフラッシュメモリ・チップよりも少ないブロック数、ブロック内ページ数とする。すなわち、ユーザデータの読み書きに使用するブロック数を７個とし、各ブロックは４ページで構成する。また、ユーザデータを格納する論理ページはＡ～Ｌの１２ページ（３ブロック分）とする。つまり、初期状態では、格納データ更新制御に使用される予備領域は４ブロックである。なお、図６～８で斜線の入ったページは無効ページ、空白のページは未書込ページである。

【００６６】
図６は、フラッシュメモリ・チップ３２１とキャッシュメモリ１２３を示すＴ１を示している。図６に示すように、Ｔ１は、フラッシュメモリ・チップ３２１内の１つのブロックが不良化し、２つ目の不良ブロックが発生する前の状態を示している。不良化したブロックは、４ページ全てが「ｂａｄ」と示されているブロック３３０Ａである。この１つ目の不良ブロック３３０Ａはフラッシュメモリ３２１内で代替される。このため、予備領域は４ブロックから３ブロックになる。また、未書込ページの残数はブロック３３０Ｂの４ページになっており、ガベッジコレクションが必要な状態である。
【００６７】
この時点で、利用可能な２４個の物理ページ（６ブロック分）内に１２個の論理ページが配置されている。したがって、フラッシュメモリ３２０の論理ページの冗長度は２４／１２＝２００パーセント（％）である。また、４個の未書込ページを除く２０個の物理ページ中に１２個の有効ページが含まれるため、平均的な無効ページ含有率は８／２０＝４０％（１ブロック辺り１．６ページ）である。したがって、ガベッジコレクション時に他ブロックへ退避すべき有効ページは平均６０％（１ブロック辺り２．４ページ）である。
【００６８】
図７は、フラッシュメモリ・チップ３２１とキャッシュメモリ１２３を示すＴ２を示している。図７に示すように、Ｔ２は、フラッシュメモリ・チップ３２１内の２つのブロック３３０Ａ、３３０Ｃが不良化した状態を示している。２つ目の不良化したブロックは、４ページ全てが「ｂａｄ」と示されているブロック３３０Ｃである。１つ目の不良ブロック３３０Ａと同様に、２つ目の不良ブロック３３０Ｃもフラッシュメモリ３２０内のブロック３３０Ｂで代替される。このため、予備領域は４ブロックから２ブロックになる。また、未書込ページの残数は４ページになっており、ガベッジコレクションが必要な状態である。
【００６９】
この時点で、利用可能な２０個の物理ページ（５ブロック分）内に１２個の論理ページが配置されている。したがって、フラッシュメモリ３２０の論理ページの冗長度は２０／１２＝１６７％である。また、４個の未書込ページを除く１６個の物理ページ中に１２個の有効ページが含まれるため、平均的な無効ページ含有率は４／１６＝２５％（１ブロック辺り１ページ）である。したがって、ガベッジコレクション時に他ブロックへ退避すべき有効ページは平均７５％（１ブロック辺り３ページ）である。このため、図６の場合に比べてガベッジコレクション時の有効ページ退避量は多くなり、このフラッシュメモリ・モジュール１５１のライト性能は低下する。なお、有効利用可能なキャッシュメモリ１２３容量は図６の場合と変わらないので、ストレージシステム１０のキャッシュヒット確率は不変である。
【００７０】
図８は、フラッシュメモリ・チップ３２１とキャッシュメモリ１２３を示すＴ３を示している。図８に示すように、Ｔ３は、フラッシュメモリ・チップ３２１内の２つのブロック３３０Ａ、３３０Ｃが不良化した状態を示す。１つ目の不良ブロック３３０Ａはフラッシュメモリ３２０内で代替されているが、２つ目の不良ブロック３３０Ｃはキャッシュメモリ１２３内の一部の領域４００で代替されている。キャッシュメモリ１２３内の一部の領域４００内で代替されるＥ～Ｈの４個の論理ページデータはキャッシュメモリ１２３上にのみ存在し、フラッシュメモリ３２０から消失している。その結果、ユーザデータを格納する論理ページはＡ～Ｄ、Ｉ～Ｌの８ページ（２ブロック分）となる。このため、予備領域は４ブロックから３ブロックになる。また、未書込ページの残数は４ページになっており、ガベッジコレクションが必要な状態である。
【００７１】
この時点で、利用可能な２０個の物理ページ（５ブロック分）内に８個の論理ページが配置されている。したがって、フラッシュメモリ３２０の論理ページの冗長度は２０／８＝２５０％である。また、４個の未書込ページを除く１６個の物理ページ中に８個の有効

ページが含まれるため、平均的な無効ページ含有率は８／１６＝５０％（１ブロック辺り２ページ）である。したがって、ガベッジコレクション時に他ブロックへ退避すべき有効ページは平均５０％（１ブロック辺り２ページ）である。このため、図６の場合に比べてガベッジコレクション時の有効ページ退避量は少なくなり、このフラッシュメモリ・モジュール１５１のライト性能は向上する。なお、有効利用可能なキャッシュメモリ１２３の容量は、４個の論理ページデータ分削減されるので、図６のＴ１の場合に比べてストレージシステム１０のキャッシュヒット確率は低下する。

【００７２】

図９は、以上の説明をまとめたテーブルＴ４である。図９に示すように、テーブルＴ４は、項目名Ｔ４１、Ｔ１～Ｔ３における結果Ｔ４１～Ｔ４３が対応付けられている。項目名Ｔ４１は、フラッシュメモリの論理ページの冗長度、ガベッジコレクション時の無効ページ含有率（平均）、ガベッジコレクション時の有効ページ移動量（平均）、フラッシュメモリのライト性能及びキャッシュメモリのヒット確立という項目名が配置されている。フラッシュメモリの論理ページの冗長度は、Ｔ４１（２００％）、Ｔ４２（１６７％）、Ｔ４３（２５０％）となっている。ガベッジコレクション時の無効ページ含有率（平均）は、Ｔ４１（４０％）、Ｔ４２（２５％）、Ｔ４３（５０％）となっている。ガベッジコレクション時の有効ページ移動量は、Ｔ４１（６０％）、Ｔ４２（７５％）、Ｔ４３（５０％）となっている。また、フラッシュメモリ３２０のライト性能は、Ｔ４２の場合は低下し、Ｔ４３の場合は向上する。キャッシュメモリ１２３のヒット確立は、Ｔ４２の場合は不変であるが、Ｔ４３の場合は低下する。

【００７３】

図９に示すように、不良ブロック３３０Ｃの代替先としてフラッシュメモリ３２０を選択すると、ストレージシステム１０のライト性能は低下するという欠点がある。一方、不良ブロック３３０Ｃの代替先としてキャッシュメモリ１２３を選択すると、ストレージシステム１０のライト性能は向上するという利点があるが、キャッシュヒット確率は低下するという欠点がある。キャッシュヒット確率が低下するということは、ミスヒットによるキャッシュメモリ１２３からフラッシュメモリ３２０へのライトバック頻度が多くなることである。上記のように、フラッシュメモリ３２０へのページ書き込みにかかる時間は、キャッシュメモリ１２３への同サイズのデータの書き込みにかかる時間よりも長い。したがって、キャッシュヒット確率が低下するということは、結果的にストレージシステム１０のライト性能を低下させる。

【００７４】

不良ブロック３３０Ｃの代替先としてキャッシュメモリ１２３を選択することの利点・欠点はトレードオフの関係にあり、ホスト計算機１００のフラッシュメモリ３２０へのアクセスパターンにより、ストレージシステム１０のライト性能が向上したり低下したりする。例えば、ランダムライトでは、キャッシュメモリ１２３の有用性が低いので、フラッシュメモリ・モジュール１５１のライト性能を向上させたほうが良く、不良ブロック（例えば、３３０Ｃ）の代替先としてキャッシュメモリ１２３を選択したほうが有利である。また、例えば、部分集中ライトでは、キャッシュメモリ１２３の有用性が高いので、有効利用可能なキャッシュメモリ１２３の容量を減らさずにヒット確率を維持したほうがよく、不良ブロック（例えば、３３０Ｃ）の代替先としてフラッシュメモリ３２０を選択したほうが有利である。

【００７５】

本発明を適用したストレージシステム１０は、不良ブロックの代替先としてキャッシュメモリ１２３を選択したときのストレージシステム１０のライト性能をホスト計算機１００のアクセスパターンに基づいて推定し、ストレージシステム１０のライト性能が向上すると判断される場合には不良ブロックの代替先としてキャッシュメモリ１２３を選択する。逆に、ストレージシステム１０のライト性能が低下すると判断される場合には不良ブロックの代替先として従来技術のようにフラッシュメモリ３２０を選択する。この処理については、後述する。

【００７６】
（３）キャッシュメモリおよびフラッシュメモリの管理手段
【００７７】
図１０、図１１を参照しながら、ストレージシステム１０におけるキャッシュメモリ１２３およびフラッシュメモリ・チップ３２１の管理手段を説明する。図１０及び図１１は、上記管理手段を説明するために、キャッシュメモリ１２３、共有メモリ１２９、ＲＡＭ３１３及びフラッシュメモリ・チップ３２１に記憶される内容を説明するための図である。なお、キャッシュメモリ１２４や、その他のフラッシュメモリ・チップについても同様であるため、キャッシュメモリ１２４や、その他のフラッシュメモリ・チップについては説明を省略する。
【００７８】
説明を簡単化するため、図６～図９の場合と同様にユーザデータの読み書きに使用するブロック数を７個とし、各ブロックは４ページで構成する。また、ユーザデータを格納する論理ページはＡｘ～Ｌｘの１２ページ（３ブロック分）とする。ここで、ｘは０以上の整数であり、各論理ページの更新回数を示す。例えば、Ｅ２とは２回更新された論理ページＥのデータを示す。図６～図９の場合と同様に、斜線の入ったページは無効ページ、空白のページは未書込ページである。また、キャッシュメモリ１２３の管理方式はフラッシュメモリ３２０のページ・サイズを単位とする２ｗａｙ－セットアソシアティブ方式とする。この方式は、１つのｗａｙが４つのｉｎｄｅｘ０～３を持ち、論理ページＡ、Ｅ、Ｉはインデックス（ｉｎｄｅｘ）＝０の２エントリを利用し、論理ページＢ、Ｆ、Ｊはインデックス＝１の２エントリを利用し、論理ページＣ、Ｇ、Ｋはインデックス＝２の２エントリを利用し、論理ページＤ、Ｈ、Ｌはインデックス＝３の２エントリを利用するものである。なお、本発明はキャッシュメモリ１２３の管理方式を特に限定するものではない。
【００７９】
図１０及び図１１に示すように、共有メモリ１２９には、キャッシュメモリ１２３を管理するためのキャッシュ管理テーブル５００を含む。フラッシュメモリ・モジュール１５１内のＲＡＭ３１３には、フラッシュメモリ・チップ３２１を管理するためのアドレス変換テーブル５１０、ページ状態テーブル５２０を含む。なお、図１０は、第５ブロックが不良化し、そのブロックがフラッシュメモリ・チップ３２１内で代替されている状態を示している。また、図１１は、図１０の状態から第４ブロックの第２ページにデータを書き込む際にライトエラーが発生し、第４ブロックが２つ目の不良ブロックになり、キャッシュメモリ１２３の一部をその代替先としたときの状態を示している。
【００８０】
図１０及び図１１に示すように、キャッシュ管理テーブル５００、は、４つのインデックス５００１とそれぞれ２つのウェイ（ｗａｙ）５００２、合計８エントリの使用状態を管理する。各エントリについてバリッド（ｖａｌｉｄ）フラグ５００３、キー（ｋｅｙ）レジスタ５００４、ダーティ（ｄｉｒｔｙ）レジスタ５００５を持つ。バリッドフラグ５００３はそのエントリが使用中か否かを記録し、
・ｖａｌｉｄ＝ｏｎ：使用中
・ｖａｌｉｄ＝ｏｆｆ：空き
と定義される。キーレジスタ５００４はそのエントリに格納しているキャッシュデータの論理ページを記録する。ダーティ（ｄｉｒｔｙ）レジスタ５００５はそのエントリのキャッシュデータがフラッシュメモリ３２０上のデータよりも新しいかどうか、あるいはそのエントリがフラッシュメモリ・チップ３２１の不良ブロック代替用に使用されているかを記録し、
・ｄｉｒｔｙ＝ｙｅｓ：フラッシュメモリ３２０のものより新しい更新データ保持
・ｄｉｒｔｙ＝ｎｏ：フラッシュメモリ３２０のものと同じオリジナルデータ保持
・ｄｉｒｔｙ＝ｓｕｂ：代替先として使用状態
と定義される。ｄｉｒｔｙ＝ｙｅｓのエントリは、別の論理ページのために使用する前にフラッシュメモリ３２０へライトバックして同期をとる必要がある。ｄｉｒｔｙ＝ｓｕｂ

のエントリは、フラッシュメモリ３２０上にライトバック先がないので別の論理ページのために使用しないように管理する必要がある。

【００８１】

図１０及び図１１に示すアドレス変換テーブル５１０は、上記のように論理ページアドレスと物理ページアドレスとを変換するためにそれらの対応状態を管理する。アドレス変換テーブル５１０は、ＬＰＡ５１０１とＰＰＡ５１０２が対応して構成される。このアドレス変換テーブル５１０において、ＬＰＡは論理ページアドレス、ＰＰＡは物理ページアドレスを意味する。なお、ブロック番号Ｘ内のページ番号Ｙの物理ページアドレスはＸＹで表している。

【００８２】

図１０及び図１１に示すページ状態テーブル５２０は、前述のようにブロック単位で各ページ状態を管理する。図１０及び図１１において、ブロック（ｂｌｏｃｋ）５２０１はブロック番号、ステータス（ｓｔａｔｕｓ）５２０２はそのブロックのページ状態を表す。

【００８３】

以下、図１０の状態から図１１の状態に遷移するときのキャッシュ管理テーブル５００アドレス変換テーブル５１０、ページ状態テーブル５２０の変化について説明する。

【００８４】

キャッシュメモリ１２３でのみ管理する対象論理ページとして、例えば、４個の論理ページＥ～Ｈを選択する。

【００８５】

キャッシュ管理テーブル５００を参照し、論理ページＥの最新データＥ２はすでにキャッシュメモリ１２３上にあるので、データ転送せずにダーティレジスタをｓｕｂ（代替状態）に設定する。

【００８６】

キャッシュ管理テーブル５００を参照し、論理ページＦの最新データＦ０はキャッシュメモリ１２３上にないので、インデックス＝１の空きエントリにデータ転送して、バリッドフラグを「ｏｎ」（使用中）、ダーティレジスタを「ｓｕｂ」（代替状態）に設定する。

【００８７】

キャッシュ管理テーブル５００を参照し、論理ページＧの最新データＧ０はキャッシュメモリ１２３上にないので、Ｋ０をキャッシュしているインデックス＝２のダーティ＝「ｎｏ」（オリジナルデータ保持）のエントリにデータ転送して、バリッドフラグを「ｏｎ」（使用中）、ダーティレジスタを「ｓｕｂ」（代替状態）に設定する。

【００８８】

キャッシュ管理テーブル５００を参照し、論理ページＨの最新データＨ０はキャッシュメモリ１２３上にないので、インデックス＝３の空きエントリにデータ転送して、バリッドフラグを「ｏｎ」（使用中）、ダーティレジスタを「ｓｕｂ」（代替状態）に設定する。

【００８９】

図１１に示すキャッシュ管理テーブル６００は、以上の設定結果を示している。

【００９０】

次に、アドレス変換テーブル５１０において、論理ページＥ～Ｈに対応する物理ページアドレス５１０２を調べ、ページ状態テーブル５２０において、その物理ページのページ状態を「８」（無効）に設定する。すなわち、ページ状態テーブル５２０の第２ブロックの欄を「８８８８」に設定する。そして、アドレス変換テーブル５１０において、論理ページＥ～Ｈに対応する物理ページアドレスをいずれもクリアして「対応なし」の状態に設定する。なお、「対応なし」は、アドレス変換テーブル５１０の実装上は、「対応なし」を意味する特別な値（例えば、「ＦＦＦＦＦＦＦＦ（１６進数）」）を定義し、その値をＲＡＭ３１３に書き込むことで「対応なし」状態を表現する。このＲＡＭ３１３に書き込

む値は実際に使用する論理ページアドレス範囲外の値であれば、どのような値でもよい。
【００９１】
次に、不良が発生した第４ブロックに含まれる有効ページのデータを退避する。つまり、第０ページのＢ１は未書込ページ（例えば、第６ブロックの第０ページ）にコピーする。当該コピー先ページが有効ページであることを示すため、ページ状態テーブル５２０の第６ブロックの欄を「０ＦＦＦ」に設定する。また、アドレス変換テーブル５１０のＬＰＡ５１０１「Ｂ」に対応するＰＰＡ５１０２を「６０」に設定する。
【００９２】
第１ページの「Ｅ２」はキャッシュメモリ１２３でのみ管理するので、すでに退避が完了しており、特に何もしない。
【００９３】
最後に、第４ブロックが不良ブロックとなったことを示すため、ページ状態テーブル５２０の第４ブロックのステータス５２０２の欄を「９９９９」に設定する。
【００９４】
図１１に示すアドレス変換テーブル５１０、ページ状態テーブル５２０は、以上の設定結果を示している。
【００９５】
（４）ストレージコントローラ１２０およびメモリコントローラ３１０の処理手順
【００９６】
図１２～図１６を参照しながら、ストレージシステム１０におけるストレージコントローラ１２０およびメモリコントローラ３１０の詳細な処理手順を説明する。以下、上記図１０、図１１に示したキャッシュメモリ１２３およびフラッシュメモリ３２０の管理手段に基づいて説明する。
【００９７】
図１２は、ホスト計算機１００からのデータライト要求について、ストレージコントローラ１２０とメモリコントローラ３１０が行う処理を示すフローチャートである。以下、その手順を説明する。
【００９８】
まず、ストレージコントローラ１２０はライト要求としてライト対象論理ページアドレスと新しいデータを受信し（Ｓ７０１）、当該論理ページアドレスのデータを含むエントリがキャッシュメモリ１２３内に存在するかを、キャッシュ管理テーブル５００で調べる（Ｓ７０２）。
【００９９】
その結果が真（存在する）ならば（Ｓ７０２：ＹＥＳ）、ストレージコントローラ１２０はキャッシュメモリ１２３内の当該データを、受信した新しいデータで更新する（Ｓ７０３）。そして、そのエントリのダーティレジスタ５００５が「ｓｕｂ（代替状態）」であるかを調べる（Ｓ７０４）。それが真ならば（Ｓ７０４：ＹＥＳ）、そのまま処理を終える。それが偽ならば（Ｓ７０４：ＮＯ）、当該ダーティレジスタ５００５を「ｙｅｓ」（更新データ保持）に設定し（Ｓ７０５）、処理を終える。
【０１００】
一方、ステップＳ７０２の結果が偽（存在しない）ならば（Ｓ７０２：ＮＯ）、ストレージコントローラ１２０はキャッシュ管理テーブル５００でバリッドフラグ５００３が「ｏｆｆ」のエントリ（空きエントリ）が存在するか調べる（Ｓ７０６）。空きエントリが存在すれば（Ｓ７０６：ＹＥＳ）、受信した新しいデータを当該エントリに書き込む（Ｓ７１０）。このとき、ストレージコントローラ１２０はそのエントリのバリッドフラグ５００３を「ｏｎ：（使用中）」に設定する。
【０１０１】
一方、空きエントリが存在しない場合は（Ｓ７０６：ＮＯ）、ストレージコントローラ１２０はダーティレジスタ５００５が「ｎｏ」のエントリ（オリジナルデータ保持エントリ）が存在するか調べる（Ｓ７０７）。オリジナルデータ保持エントリが存在すれば（Ｓ

７０７：ＹＥＳ）、そのエントリは上書きしても問題ないので、空きエントリ存在の場合と同様にステップＳ７１０へ遷移する。

【０１０２】

オリジナルデータ保持エントリが存在しなければ（Ｓ７０７：ＮＯ）、ストレージコントローラ１２０はダーティレジスタ５００５が「ｙｅｓ」のエントリ（更新データ保持エントリ）の格納データをフラッシュメモリ３２０へライトバックし（Ｓ７０８、図１４参照）、そのダーティレジスタ５００５を「ｎｏ」に設定する（Ｓ７０９）。その結果、そのエントリは上書きしても問題ないので、空きエントリ存在の場合と同様にステップＳ７１０へ遷移する。なお、ライトバック処理Ｓ７０８の対象となるエントリの選択には、例えば、ＬＲＵ（Least Recently Used）アルゴリズムを適用する。

【０１０３】

ステップＳ７１０の後、ストレージコントローラ１２０はそのダーティレジスタ５００５を「ｙｅｓ」（更新データ保持）に設定して（Ｓ７０５）、処理を終える。

【０１０４】

図１３は、ホスト計算機１００からのデータリード要求について、ストレージコントローラ１２０とメモリコントローラ３１０が行う処理を示すフローチャートである。以下、その手順を説明する。

【０１０５】

まず、ストレージコントローラ１２０はリード要求としてリード対象論理アドレスを受信し（Ｓ７１１）、当該論理ページアドレスのデータを含むエントリがキャッシュメモリ１２３内に存在するかを、キャッシュ管理テーブル５００で調べる（Ｓ７１２）。

【０１０６】

その結果が偽（存在しない）ならば（Ｓ７１２：ＮＯ）、ストレージコントローラ１２０はキャッシュ管理テーブル５００でバリッドフラグ５００３が「ｏｆｆ」のエントリ（空きエントリ）が存在するか調べる（Ｓ７１５）。空きエントリが存在すれば（Ｓ７１５：ＹＥＳ）、受信した論理アドレスからフラッシュメモリ・モジュール１５１～１５４、１６１～１６４、１７１～１７４及び１８１～１８４のいずれのフラッシュメモリ・モジュールの論理ページアドレスを特定し、メモリコントローラ３１０がアドレス変換テーブル５１０によりそれに対応する物理ページアドレス５１０２を調べ、そのページの格納データを読み出す（Ｓ７１９）。そして、メモリコントローラ３１０は、そのデータをキャッシュメモリ１２３の当該エントリに転送する（Ｓ７２０）。このとき、ストレージコントローラ１２０はそのエントリのバリッドフラグ５００３を「ｏｎ」（使用中）、ダーティレジスタ５００５を「ｎｏ」（オリジナルデータ保持）に設定する。その後、ステップＳ７１２に戻る。

【０１０７】

一方、空きエントリが存在しない場合は（Ｓ７１５：ＮＯ）、ストレージコントローラ１２０はダーティレジスタ５００５が「ｎｏ」のエントリ（オリジナルデータ保持エントリ）が存在するか調べる（Ｓ７１６）。オリジナルデータ保持エントリが存在すれば（Ｓ７１６：ＹＥＳ）、そのエントリは上書きしても問題ないので、空きエントリ存在の場合と同様にステップＳ７１９へ遷移する。オリジナルデータ保持エントリが存在しなければ（Ｓ７１６：ＮＯ）、ストレージコントローラ１２０はダーティレジスタ５００５が「ｙｅｓ」のエントリ（更新データ保持エントリ）の格納データをフラッシュメモリ３２０へライトバックし（Ｓ７１７、図１４参照）、そのダーティレジスタ５００５を「ｎｏ」に設定する（Ｓ７１８）。その結果、そのエントリは上書きしても問題ないので、空きエントリ存在の場合と同様にステップＳ７１９へ遷移する。なお、ライトバック処理Ｓ７１７の対象となるエントリの選択には、例えば、ＬＲＵアルゴリズムを適用する。

【０１０８】

一方、ステップＳ７１２の結果が真（存在する）ならば（Ｓ７１２：ＹＥＳ）、ストレージコントローラ１２０はキャッシュメモリ１２３内の当該データを読み出し（Ｓ７１３）、ホスト計算機１００へそれを送信して（Ｓ７１４）、処理を終える。

【０１０９】
図１４は、図１２におけるメモリコントローラ３１０によるキャッシュデータのライトバック処理Ｓ７０８、図１３におけるメモリコントローラ３１０によるキャッシュデータのライトバック処理Ｓ７１７、後述する図１５におけるメモリコントローラ３１０によるキャッシュデータのライトバック処理Ｓ９０５の詳細な処理手順を示すフローチャートである。以下、その手順を説明する。
【０１１０】
まず、メモリコントローラ３１０はキャッシュメモリ１２３からライトバック対象の論理ページアドレスとそのデータを取得する（Ｓ８０１）。そして、ページ状態テーブル５２０を用いて状態＝「Ｆ」（未書込）のページを選択し、そのページに取得したデータを書き込む（Ｓ８０２）。そして、メモリコントローラ３１０は、書き込みにエラーが発生したかを判定する（Ｓ８０３）。書き込みがエラーならば（Ｓ８０３：ＹＥＳ）、メモリコントローラ３１０は、不良ブロック代替処理（Ｓ８０４）を実施して、ステップＳ８０２に戻る。不良ブロック代替処理の詳細は図１５を参照して後述する。
【０１１１】
一方、書き込みが成功ならば（Ｓ８０３：ＮＯ）、メモリコントローラ３１０は、アドレス変換テーブル５１０を用いてライトバック対象の論理ページアドレス５１０１に対応する物理ページアドレス（旧アドレス）５１０２を調べ、ページ状態テーブル５２０においてその旧アドレスが示すページの状態を「８」（無効）に設定する（Ｓ８０５）。
【０１１２】
そして、アドレス変換テーブル５１０において、メモリコントローラ３１０は、ライトバック対象の論理ページアドレス５０１２に対応する物理ページアドレス５１０２に、ステップＳ８０２で書き込んだ物理ページのアドレス（新アドレス）を設定する（Ｓ８０６）。また、メモリコントローラ３１０は、ページ状態テーブル５２０においてその新アドレスが示すページの状態を「０」（有効）に設定する（Ｓ８０７）。最後に、メモリコントローラ３１０は、ガベッジコレクション処理（Ｓ８０８）を実施し、次回のライトバックのために十分な未書込ページを確保して、ライトバック処理を終える。
【０１１３】
図１５は、図１４におけるストレージコントローラ１２０とメモリコントローラ３１０による不良ブロックの代替処理Ｓ８０４の詳細な処理手順を示すフローチャートである。以下、その手順を説明する。
【０１１４】
まず、メモリコントローラ３１０は不良ブロックの代替先としてキャッシュメモリ１２３を選択した場合と、フラッシュメモリ３２０を選択した場合とで、ストレージシステム１０のライト性能の変化をそれぞれ推定し、比較する（Ｓ９００）。そのライト性能の具体的方法の一例は後述する。
【０１１５】
メモリコントローラ３１０は、比較の結果、キャッシュメモリ１２３へ代替したほうが有利か否かを判定し（Ｓ９０１）、キャッシュメモリ１２３へ代替したほうが有利（ライト性能が高い）であれば（Ｓ９０１：ＹＥＳ）、キャッシュメモリ１２３への代替処理（Ｓ９０２～Ｓ９１１及びＳ９１７）を実施し、さもなくば（Ｓ９０１：ＮＯ）、フラッシュメモリ３２０への代替処理（Ｓ９１２～Ｓ９１７）を実施する。
【０１１６】
キャッシュメモリ１２３への代替処理としては、まず、メモリコントローラ３１０はページ状態テーブル５２０を用いて不良ブロックの発生したフラッシュメモリ・チップ３２１から１ブロック分の有効ページ（図６～図１１の例では４ページ）を選択する（Ｓ９０２）。ストレージコントローラ１２０はキャッシュ管理テーブル５００でバリッドフラグ５００３が「ｏｆｆ」のエントリ（空きエントリ）が存在するか調べる（Ｓ９０３）。
【０１１７】
空きエントリが存在すれば（Ｓ９０３：ＹＥＳ）、メモリコントローラ３１０はステッ

プＳ９０２で選択した有効ページの１つの格納データを当該エントリに移動する（Ｓ９０７）。このとき、ストレージコントローラ１２０はそのエントリのバリッドフラグ５００３を「ｏｎ」（使用中）に設定する。そして、ストレージコントローラ１２０はそのダーティレジスタ５００５を「ｓｕｂ」（代替状態）に設定する（Ｓ９０８）。

【０１１８】

一方、空きエントリが存在しない場合は（Ｓ９０３：ＮＯ）、ストレージコントローラ１２０はダーティレジスタ５００５が「ｎｏ」のエントリ（オリジナルデータ保持エントリ）が存在するか調べる（Ｓ９０４）。オリジナルデータ保持エントリが存在すれば（Ｓ９０４：ＹＥＳ）、そのエントリは上書きしても問題ないので、空きエントリ存在の場合と同様にステップＳ９０７へ遷移する。

【０１１９】

オリジナルデータ保持エントリが存在しなければ（Ｓ９０４：ＮＯ）、ストレージコントローラ１２０はダーティレジスタ５００５が「ｙｅｓ」のエントリ（更新データ保持エントリ）の格納データをフラッシュメモリ３２０へライトバックし（Ｓ９０５、図１４参照）、そのダーティレジスタ５００５を「ｎｏ」に設定する（Ｓ９０６）。その結果、そのエントリは上書きしても問題ないので、空きエントリ存在の場合と同様にステップＳ９０７へ遷移する。なお、ライトバック処理Ｓ９０５の対象となるエントリの選択には、例えば、ＬＲＵアルゴリズムを適用する。ストレージコントローラ１２０は、ステップＳ９０２で選択した全ての有効ページの格納データの移動が終わるまでステップＳ９０３～Ｓ９０８を繰り返す（Ｓ９０９）。

【０１２０】

全ての移動が完了すれば（Ｓ９０９：ＹＥＳ）、メモリコントローラ３１０は不良ブロック内の各有効ページの物理ページアドレスをアドレス変換テーブル５１０のＰＰＡ５１０２から検出する（Ｓ９１０）。そして、メモリコントローラ３１０は、検出した物理ページアドレスのページ状態をページ状態テーブル５２０で「８」（無効）に設定し、アドレス変換テーブル５１０の当該ＰＰＡ５１０２をクリア（「対応なし」を設定）する（Ｓ９１１）。最後に、メモリコントローラ３１０は、不良ブロックを構成する各ページのページ状態を「９」（不良）に設定し（Ｓ９１７）、不良ブロック代替処理を終える。

【０１２１】

一方、フラッシュメモリへの代替処理としては、まず、メモリコントローラ３１０がガベッジコレクション処理を実施し（Ｓ９１２）、全てのページが未書込状態の消去済みブロックを確保する。

【０１２２】

そして、不良ブロック内の有効ページをページ状態テーブル５２０から検出し、その格納データを当該消去済みブロックに移動する（Ｓ９１３）。次に、メモリコントローラ３１０は、ページ状態テーブル５２０でそれぞれの移動先ページの状態を「０」（有効）に設定する（Ｓ９１４）。また、メモリコントローラ３１０は、不良ブロック内の各有効ページの物理ページアドレスをアドレス変換テーブル５１０のＰＰＡ５１０２から検出する（Ｓ９１５）。そして、そのＰＰＡ５１０２の欄にステップＳ９１３でのそれぞれの移動先ページの物理ページアドレスを設定する（Ｓ９１６）。最後に、メモリコントローラ３１０は、不良ブロックを構成する各ページのページ状態を「９」（不良）に設定し（Ｓ９１７）、不良ブロック代替処理を終える。

【０１２３】

なお、ステップＳ９０２で選択する有効ページとしては、書き換え頻度が高い論理ページのデータを格納しているものを選択することが好ましい。なぜなら、書き換えがほとんどない論理ページのデータをキャッシュメモリ１２３上で永続的に保持することは、性能面で非効率であるからである。

【０１２４】

図１６は、図１４におけるメモリコントローラ３１０によるガベッジコレクション処理Ｓ８０８、図１５におけるメモリコントローラ３１０によるガベッジコレクション処理Ｓ

９１２の詳細な処理手順を示すフローチャートである。以下、その手順を説明する。

【０１２５】

まず、メモリコントローラ３１０はページ状態テーブル５２０を見て、ページ状態値＝「Ｆ」（未書込）のページの残数が所定の値以下になったかを調べる（Ｓ１００１）。なお、ステップＳ８０８の時の所定数は「１ブロック分の総ページ数（図６～図１１の例では４ページ）」、ステップＳ９１２の時の所定数は「２ブロック分の総ページ数（図６～図１１の例では８ページ）」である。ステップＳ１００１の結果が偽（所定値より大きい）ならば（Ｓ１００１：ＮＯ）、メモリコントローラ３１０は、何もせずにそのまま処理を終える。

【０１２６】

一方、その結果が真（所定値以下）ならば（Ｓ１００１：ＹＥＳ）、ページ状態値＝「８」（無効）であるページを最も多く含むブロックを１つ選択し、その中にあるページ状態値＝「０」（有効）である有効ページを全て検出する（Ｓ１００２）。なお、検出した有効ページ数をＮとする。

【０１２７】

そしてメモリコントローラ３１０は、選択したブロック以外のブロック上にあるページ状態値＝Ｆの未書込ページを選択し、ステップＳ１００２で検出した有効ページの１つに格納されたデータを、それぞれ未書込ページの１つにコピー（退避）する（Ｓ１００３）。次に、メモリコントローラ３１０は、ページ状態テーブル５２０において、ステップＳ１００３でのそれぞれのコピー先ページのページ状態値を「０」に設定する（Ｓ１００４）。そして、メモリコントローラ３１０は、ステップＳ１００２で検出した有効ページのページ状態値を「８」（無効）に設定する（Ｓ１００５）。そして、アドレス変換テーブル５１０上に設定された全ての物理ページアドレス（ＰＰＡ）の中からステップＳ１００２で検出した有効ページのアドレスを検索することにより、そのアドレスが設定されたＰＰＡ５１０２を検出し、そこにステップＳ１００３でのコピー先ページの物理ページアドレスをコピーする（Ｓ１００６）。

【０１２８】

そして、メモリコントローラ３１０は、検出した全有効ページがコピー済みかどうかを判定する（Ｓ１００７）。そして、検出した全ての有効ページがコピー済みでなければ（Ｓ１００７：ＮＯ）、ステップＳ１００３に戻る。すなわち、検出したＮ個の有効ページの全てについてステップＳ１００３～Ｓ１００６を繰り返す。これにより、ステップＳ１００２で選択したブロック内の全ページが無効化され、そこにあった保存すべきページデータの退避も完了する。

【０１２９】

一方、検出した全ての有効ページがコピー済みであれば（Ｓ１００７：ＹＥＳ）、メモリコントローラ３１０は、ステップＳ１００２で選択したブロックを消去する（Ｓ１００８）。そして、メモリコントローラ３１０は、消去時にエラーが発生したか否かを判定する（Ｓ１００９）。この消去時にエラーが発生したならば（Ｓ１００９：ＹＥＳ）、ページ状態テーブル５２０において、選択したブロックの各ページの状態を「９」（不良）に設定して（Ｓ１０１０）、ステップＳ１００２に戻る。消去時にエラーが発生せず、消去が成功したならば（Ｓ１００９：ＮＯ）、ページ状態テーブル５２０において、そのブロック内の全ページのページ状態値を「Ｆ」（未書込）に設定する（Ｓ１０１１）。以上をステップＳ１００１の結果が偽となるまで繰り返し、処理を終える。

【０１３０】

（５）不良ブロック代替先によるライト性能の変化の推定・比較方法

【０１３１】

図１５のステップＳ９００におけるライト性能の変化の推定・比較方法の一例を以下で説明する。

【０１３２】

不良ブロックが発生した時点で、そのフラッシュメモリ・チップ３２１内で管理してい

る論理ブロック数をＭ、フラッシュメモリ・チップ３２１で利用可能な（つまり、不良でない）物理ブロック数をＮと定義する。このとき、フラッシュメモリ・チップ３２１の論理ページ冗長度はＮ／Ｍとなる。

【０１３３】

また、キャッシュメモリ１２３の総容量をＣ、フラッシュメモリ・チップ３２１のブロックサイズをＢ、キャッシュメモリ１２３で代替されている不良ブロック数をＳと定義する。このとき、キャッシュとして有効利用可能なキャッシュメモリ１２３の容量は「Ｃ－Ｂ＊Ｓ」（＊は乗算を示す）となる。

【０１３４】

さらに、図１２においてＰ７３０で囲まれた処理（Ｓ７０３～Ｓ７０５）に必要な時間をキャッシュメモリライト処理時間Ｔｃ、Ｐ７４０で囲まれた処理（Ｓ７０６～Ｓ７１０及びＳ７０５）に必要な時間をフラッシュメモリライト処理時間Ｔｆと定義する。

【０１３５】

キャッシュメモリライト処理時間Ｔｃは一定である。しかし、フラッシュメモリライト処理時間Ｔｆは、ライトバック処理Ｓ７０８に要する時間がフラッシュメモリ・チップ３２１の状態に依存して変化するため一定ではない。図１４に示したように、ライトバック処理Ｓ７０８はガベッジコレクション処理Ｓ８０８を含む。図１６に示したように、ガベッジコレクション処理Ｓ８０８では２ブロック間で有効ページの置換を行う。図６～図８を用いて説明したように、このときの置換ページ数の平均は論理ページ冗長度Ｎ／Ｍが小さいほど多くなる。ゆえに、論理ページ冗長度Ｎ／Ｍが小さいほどライトバック処理Ｓ７０８に要する時間は長くなる。したがって、フラッシュメモリライト処理時間Ｔｆは論理ページ冗長度Ｎ／Ｍの関数Ｔｆ（Ｎ／Ｍ）となる。

【０１３６】

次に、過去一定時間Δｔ内に、ホスト計算機１００がライトアクセスしたサイズをＡ、その間のキャッシュメモリ１２３のヒット率をＲとする。このときヒット率Ｒは、

・ＡがＣ－Ｂ＊Ｓ以上のとき、Ｒ＝（Ｃ－Ｂ＊Ｓ）／Ａ

・ＡがＣ－Ｂ＊Ｓ未満のとき、Ｒ＝１

と表される。そして、ホスト計算機１００のライトアクセス処理時間の期待値Ｔｗは、

・ＡがＣ－Ｂ＊Ｓ以上のとき、Ｔｗ＝Ｒ＊Ｔｃ＋（１－Ｒ）Ｔｆ

・ＡがＣ－Ｂ＊Ｓ未満のとき、Ｔｗ＝Ｔｃ

と表される。以上より、ライト処理時間期待値ＴｗはＳ、Ｍ、Ｎの関数Ｔｗ（Ｓ、Ｍ、Ｎ）となる。メモリコントローラ３１０は、ストレージコントローラ１２０の管理する「Ｒ」、「Ｓ」、「Ａ」といった情報に基づいて、この関数Ｔｗが今回発生した不良ブロックの代替先によってどのように変化するかを評価する。

【０１３７】

キャッシュメモリ１２３で代替した場合、「Ｓ」は１増加、「Ｍ」は１減少、「Ｎ」は１減少するため、ライト処理時間期待値はＴｗ（Ｓ＋１、Ｍ－１、Ｎ－１）に変化すると推定する。フラッシュメモリ３２０で代替した場合、「Ｓ」は不変、「Ｍ」は不変、「Ｎ」は１減少するため、ライト処理時間期待値はＴｗ（Ｓ、Ｍ、Ｎ－１）に変化すると推定する。

【０１３８】

以上より、Ｔｗ（Ｓ＋１、Ｍ－１、Ｎ－１）がＴｗ（Ｓ、Ｍ、Ｎ－１）未満ならば、フラッシュメモリ３２０を代替先とした方が有利である。また、Ｔｗ（Ｓ＋１、Ｍ－１、Ｎ－１）がＴｗ（Ｓ、Ｍ、Ｎ－１）以上ならば、キャッシュメモリ１２３を代替先とした方が有利である。このような比較により、不良ブロックの代替先を選択する。

【０１３９】

例えば、ランダムライトのようなキャッシュメモリ１２３のヒット率Ｒが非常に小さい状況下では、フラッシュメモリライト処理時間Ｔｆを短くすることによりライトアクセス処理時間の期待値Ｔｗを短くするのが有効であるため、キャッシュメモリ１２３が代替先として選択されやすくなる。また、例えば、部分集中ライトのようなキャッシュメモリ１

２３のヒット率Ｒが非常に大きい状況下では、キャッシュとして有効利用可能なキャッシュメモリ１２３の容量を維持することによりヒット率Ｒを維持してライトアクセス処理時間の期待値Ｔｗの増加を抑えるのが有効であるため、フラッシュメモリ３２０が代替先として選択されやすくなる。

【０１４０】

なお、上記評価において、キャッシュメモリ１２３の総容量Ｃに対するキャッシュとして有効利用可能なキャッシュメモリ１２３の容量の割合に下限を与えたり、キャッシュメモリ１２３で代替される不良ブロック数Ｓに上限を与えたりすることにより、キャッシュメモリ１２３が代替先として選択されるのを制限し、キャッシュとして有効利用可能なキャッシュメモリ１２３の容量が少なくなり過ぎてストレージシステム１０の性能が不安定になることを防止してもよい。ここでの不安定とは、ホスト計算機１００からのアクセスパターンの変動に対して性能が大きく変動し、保証性能が確保できない状況になるという意味である。

【０１４１】

なお、上記評価において、メモリコントローラ３１０がストレージコントローラ１２０の管理する「Ｒ」、「Ｓ」、「Ａ」といった情報を取得するため、ストレージコントローラ１２０はそれらを送信するコマンドを発行し、フラッシュメモリ・モジュール１５１～１５４、１６１～１６４、１７１～１７４及び１８１～１８４はそのコマンドを解釈するように構成されている。

【０１４２】

（６）代替データの移設によるライト性能最適化

【０１４３】

上記の実施形態では、不良ブロック発生時にその代替先の最適な場所を評価したが、通常の動作状況でもそれまでの不良ブロック代替先の最適な配分を評価して、その配分を調整してもよい。

【０１４４】

メモリコントローラ３１０はストレージコントローラ１２０の管理する「Ｒ」、「Ｓ」、「Ａ」といった情報を常時監視しながら、関数Ｔｗ（Ｓ、Ｍ、Ｎ）の最適化を実施する。

【０１４５】

すなわち、例えばキャッシュメモリ１２３上の代替データを１ブロック分フラッシュメモリ３２０へ移設した場合、「Ｓ」は１減少、「Ｍ」は１増加、「Ｎ」は不変のため、ライト処理時間期待値はＴｗ（Ｓ－１、Ｍ＋１、Ｎ）に変化すると推定する。逆に、フラッシュメモリ３２０上の代替データを１ブロック分キャッシュメモリ１２３へ移設した場合、「Ｓ」は１増加、「Ｍ」は１減少、「Ｎ」は不変のため、ライト処理時間期待値はＴｗ（Ｓ＋１、Ｍ－１、Ｎ）に変化すると推定する。

【０１４６】

以上より、ライト処理時間期待値Ｔｗ（Ｓ－１、Ｍ＋１、Ｎ）がライト処理時間期待値Ｔｗ（Ｓ、Ｍ、Ｎ）未満ならば、キャッシュメモリ１２３上の代替データを１ブロック分フラッシュメモリ３２０へ移設する。また、ライト処理時間期待値Ｔｗ（Ｓ＋１、Ｍ－１、Ｎ）がライト処理時間期待値Ｔｗ（Ｓ、Ｍ、Ｎ）未満ならば、フラッシュメモリ３２０上の代替データを１ブロック分キャッシュメモリ１２３へ移設する。

【０１４７】

以上のように、フラッシュメモリ３２０を主たる記憶媒体とし、キャッシュメモリ１２３を搭載し、本発明を適用したストレージシステム１０は、フラッシュメモリ３２０の不良ブロック数の増加に伴って格納データ更新時の作業効率が低下するのを抑制することができるため、フラッシュメモリおよびキャッシュメモリを搭載した従来技術のようなストレージシステムよりもライト性能の低下を抑制するという効果を奏する。

【０１４８】

以上（４）～（６）においては、キャッシュメモリ１２３で不良ブロックを代替する場

合を示したが、キャッシュメモリ１２４で代替するように構成しても良い。また、（５）や（６）の評価では、キャッシュメモリ１２３と１２４を合わせた総容量Ｃや両者の平均ヒット率Ｒ等を計算に用いても良い。

【０１４９】

（７）キャッシュメモリデータの信頼性向上

【０１５０】

上記の実施形態でのキャッシュメモリ１２３、１２４を不揮発性メモリによって構成してもよい。そうすれば、フラッシュメモリ・チップ３２１からキャッシュメモリ１２３に代替されたデータを無電源で永続的に保持することができる。

【０１５１】

不揮発性メモリの一例は、相変化ＲＡＭである。相変化ＲＡＭは、ダイナミック型ランダムアクセスメモリのキャパシタ部分を、光ディスクなどに使用されているＧＳＴ（Ga-Sb-Te）と呼ばれる相変化材料に置き換えた構造のものを利用することが好ましい。この相変化ＲＡＭはダイナミック型ランダムアクセスメモリとほぼ同程度のライト性能を持つため、このような実施形態も、上で述べたような本発明の効果を享受する。

【０１５２】

不揮発性メモリのもう一例は、フラッシュメモリである。このフラッシュメモリはフラッシュメモリ・チップ３２１と同じ種類のものでもよい。ただし、このフラッシュメモリはデータ更新作業効率を高めるために多くの予備領域を含み、フラッシュメモリ・チップ３２１よりも高速にデータを書き換えられる性能を持つ。したがって、このような実施形態も、上で述べたような本発明の効果を享受する。なお、このときの共有メモリ１２９には、ＲＡＭ３１３にあるようなアドレス変換テーブル５１０やページ状態テーブル５２０をさらに作成し、キャッシュメモリ１２３、１２４内のデータ格納位置を管理する。

【０１５３】

不揮発性のキャッシュメモリを搭載したストレージシステム１０は、フラッシュメモリからキャッシュメモリに代替されたデータを補助電源なしに保持することができるため、ストレージシステムの電源消費電力を削減することができ、そのデータを突然の電源遮断などの障害による揮発的消失から保護するという効果を奏する。

【０１５４】

また、上記の実施形態でのキャッシュメモリ１２３に含まれるフラッシュメモリ代替データ（ダーティ（ｄｉｒｔｙ）＝サブ（ｓｕｂ）のエントリの格納データ）を、もう１つのキャッシュメモリ１２４に複写し、２重に保持・管理してもよい。そうすれば、キャッシュメモリ１２３が故障しても、フラッシュメモリ代替データが消失しないので、ストレージシステム１０の信頼性が保全される。

【０１５５】

なお、フラッシュメモリ代替データの保持方式では上の２重化（ミラーリング）以外の方式を適用してもよい。例えば、ストレージシステム１０にキャッシュメモリをさらに追加し、複数のキャッシュメモリでＲＡＩＤグループを組み、ＲＡＩＤ５などの冗長化方式を適用してフラッシュメモリ代替データを保持してもよい。

【０１５６】

（８）他の実施形態

【０１５７】

上述の実施形態では本発明を、ページ単位でデータを書き込み、複数のページから構成されるブロックを単位としてデータを消去するとともに複数のブロックを有し、データの更新にページを含むブロックの消去を必要とするフラッシュメモリ３２０と、フラッシュメモリ３２０に書き込むべきデータをフラッシュメモリ３２０よりも高速に書き込むとともに一時的に記憶するキャッシュメモリ１２３と、フラッシュメモリ３２０のデータの読み出し、書き込み及び消去と、キャッシュメモリ１２３のデータの読み出し及び書き込みを制御し、フラッシュメモリ３２０内に不良なブロックが発生したことを検出するコントローラ１２０と、データのライト処理を要求するコマンドを発行するホスト計算機１００

とを含むストレージシステム１０において、チャネルアダプタ１２１，１２２及びストレージアダプタ１２５，１２６を含むストレージコントローラ１２０は、フラッシュメモリ３２０内に不良ブロックが発生したことを検出したときに、フラッシュメモリ３２０に格納された「Ｅ」、「Ｆ」、「Ｇ」及び「Ｈ」というデータをキャッシュメモリ１２３に移動し、その移動したデータを更新するためのコマンドをホスト計算機１００から受信しても、そのコマンドに基づくデータをフラッシュメモリ３２０へ書き込むことを禁止する場合について適用した場合について述べたが、本発明はこれに限られず、この他種々の構成のストレージシステムに広く適用することができる。

【０１５８】

また、ストレージシステム１０は、キャッシュメモリ１２３に記憶するデータを管理するキャッシュ管理テーブル５００を共有メモリ１２９内に備え、キャッシュ管理テーブル５００は、データをフラッシュメモリ３２０へ書き込むことを禁止する禁止情報として「ｓｕｂ」（ダーティレジスタ５００５の項目）を保持する場合について説明したが、禁止情報の保持形式は、これに限られるものではない。

【０１５９】

さらに、ストレージシステム１０は、フラッシュメモリ３２０に記憶するデータのＬＰＡ５１０１とＰＰＡ５１０２との対応関係を管理するアドレス変換テーブル５１０をＲＡＭ３１３内に備え、アドレス変換テーブル５１０は、データのＬＰＡ５１０１の「Ｅ」、「Ｆ」、「Ｇ」及び「Ｈ」に対応するＰＰＡ５１０２が存在しないことを表すアドレス不存在情報として「-」を保持する場合について説明したが、アドレス不在情報の保持形式は、これに限られるものではない。

【産業上の利用可能性】

【０１６０】

本発明は、種々のストレージシステムに広く適用することができる。

【図面の簡単な説明】

【０１６１】

【図１】本発明に係わるストレージシステムの構成を示す図である。

【図２】本発明に係わるストレージシステムを構成するチャネルアダプタの内部構成を示す図である。

【図３】本発明に係わるストレージシステムを構成するストレージアダプタの内部構成を示す図である。

【図４】本発明に係わるストレージシステムを構成するフラッシュメモリ・モジュールの内部構成を示す図である。

【図５】本発明に係わるストレージシステムを構成するフラッシュメモリ・モジュールに搭載されたフラッシュメモリ・チップの構造を示す図である。

【図６】本発明に係わるフラッシュメモリで発生した不良ブロックの代替方法の違いがもたらす影響を説明するための図である。

【図７】本発明に係わるフラッシュメモリで発生した不良ブロックの代替方法の違いがもたらす影響を説明するための図である。

【図８】本発明に係わるフラッシュメモリで発生した不良ブロックの代替方法の違いがもたらす影響を説明するための図である。

【図９】本発明に係わるフラッシュメモリで発生した不良ブロックの代替方法の違いがもたらす影響を説明するための図である。

【図１０】本発明に係わるキャッシュメモリおよびフラッシュメモリ・チップの管理に関する説明をするための図である。

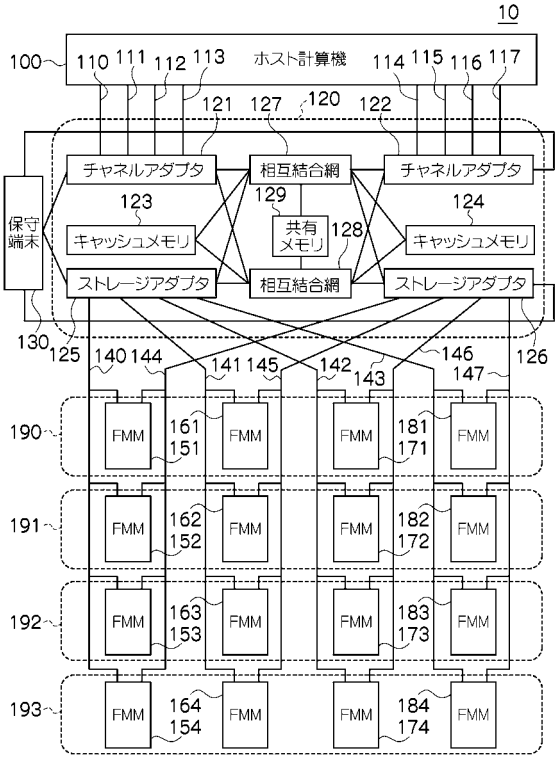
【図１１】本発明に係わるキャッシュメモリおよびフラッシュメモリ・チップの管理に関する説明をするための図である。

【図１２】本発明に係わるホスト計算機からのデータライト要求について、ストレージコントローラとメモリコントローラが行う処理を示すフローチャートである。

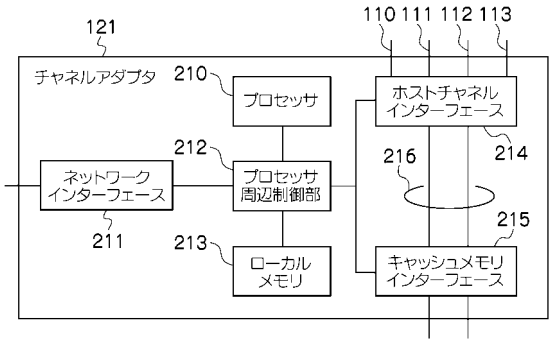
【図１３】本発明に係わるデータリード要求について、ストレージコントローラとメモリ

コントローラが行う処理を示すフローチャートである。
【図１４】本発明に係わるメモリコントローラによるキャッシュデータのライトバック処理を示すフローチャートである。
【図１５】本発明に係わるストレージコントローラとメモリコントローラによる不良ブロックの代替処理を示すフローチャートである。
【図１６】本発明に係わるメモリコントローラによるガベッジコレクション処理を示すフローチャートである。
【符号の説明】
【０１６２】
１０…ストレージシステム
１００…ホスト計算機
１２０…ストレージコントローラ
１２１，１２２…チャネルアダプタ
１２３、１２４…キャッシュメモリ
１２５，１２６…ストレージアダプタ
１２９…共有メモリ
１５１～１８４…フラッシュメモリ・モジュール
３１０…メモリコントローラ
３２０…フラッシュメモリ
３２１…フラッシュメモリ・チップ
５００…キャッシュ管理テーブル
５１０…アドレス管理テーブル
５２０…ページ情報テーブル
５００１…インデックス
５００２…ウェイ
５００３…バリッドフラグ
５００４…キーレジスタ
５００５…ダーティレジスタ
５１０１…ＬＰＡ（論理ページアドレス）
５１０２…ＰＰＡ（物理ページアドレス）
５２０１…ブロック
５２０２…ステータス

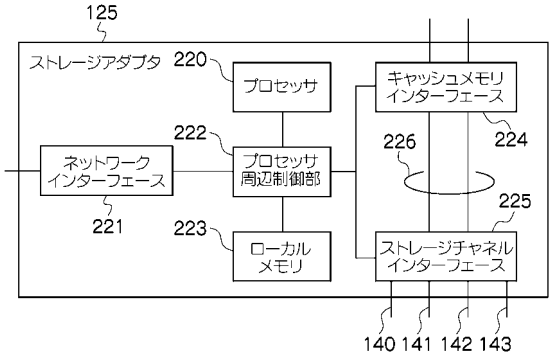
【図１】
図1



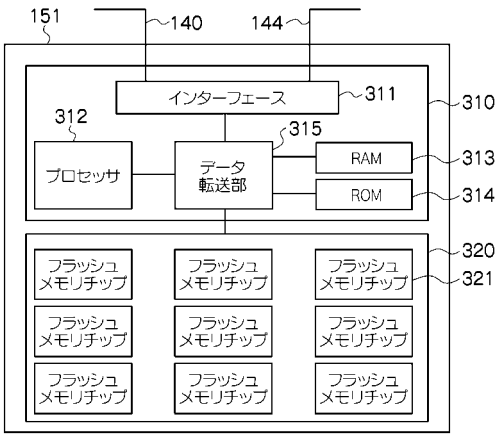

【図２】
図2



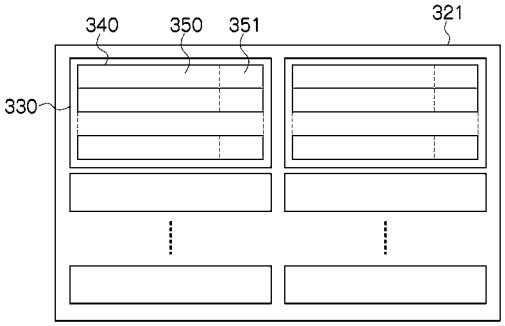

【図３】
図3



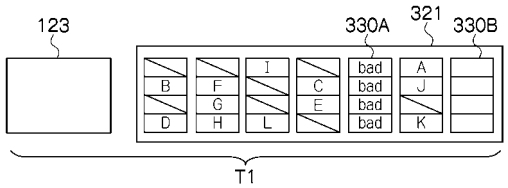

【図４】
図4



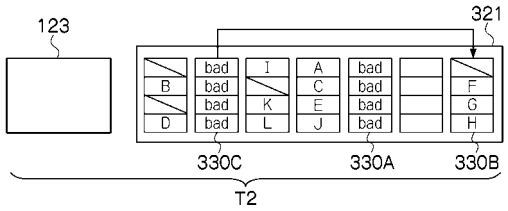

【図５】
図5



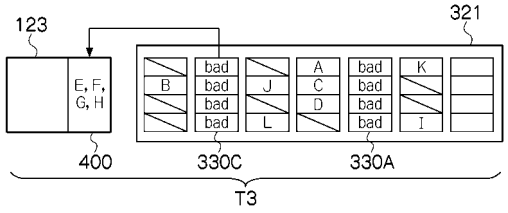

【図６】
図6




【図７】
図7




【図８】
図8



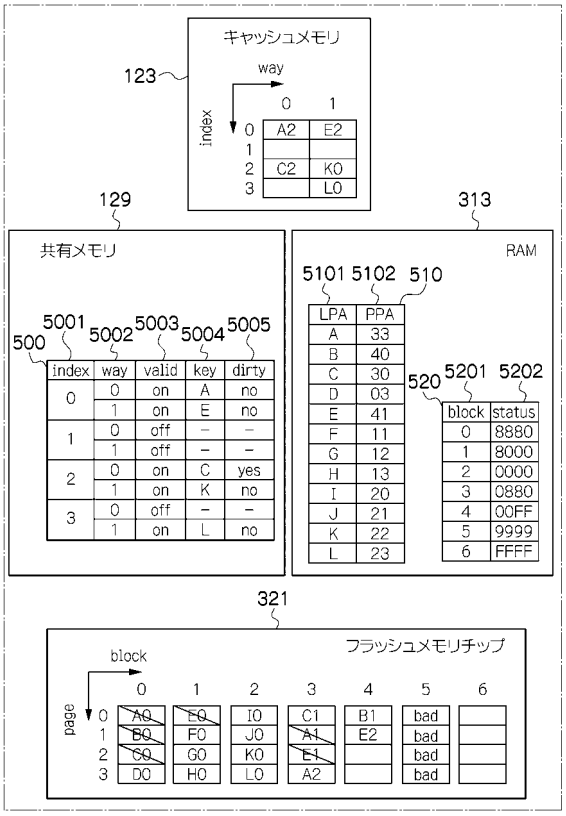
【図９】

図9

| 項目名 | T1 | T2 | T3 |
|---|---|---|---|
| フラッシュメモリの論理ページの冗長度 | 24/12 =200% | 20/12 =167% | 20/8 =250% |
| ガベッジコレクション時の無効ページ含有率(平均) | 8/20 =40% | 4/16 =25% | 8/16 =50% |
| ガベッジコレクション時の有効ページ移動量(平均) | 60% | 75% | 50% |
| フラッシュメモリのライト性能 | | 低下 | 向上 |
| フラッシュメモリのヒット確率 | | 不変 | 低下 |

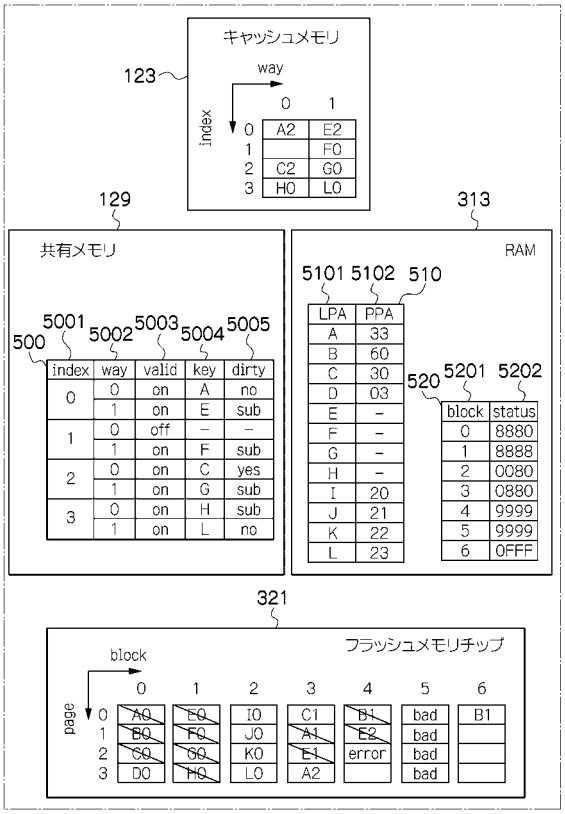
【図１０】

図10



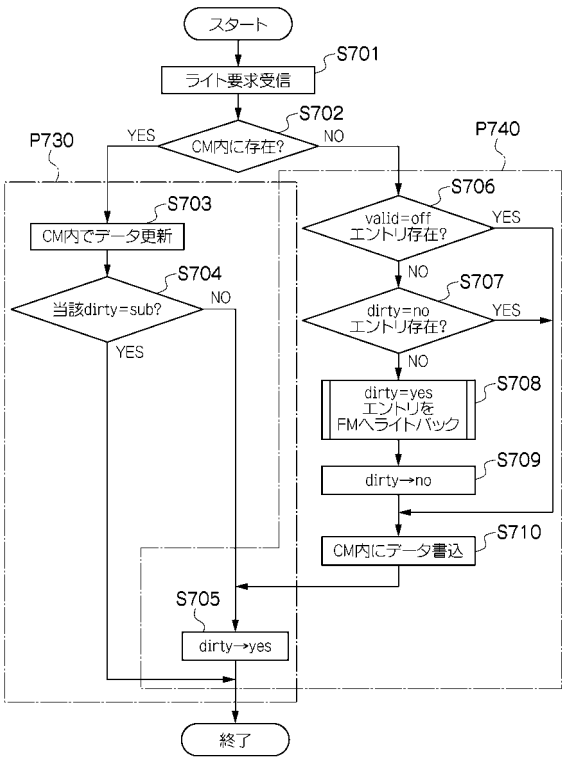

【図１１】

図11

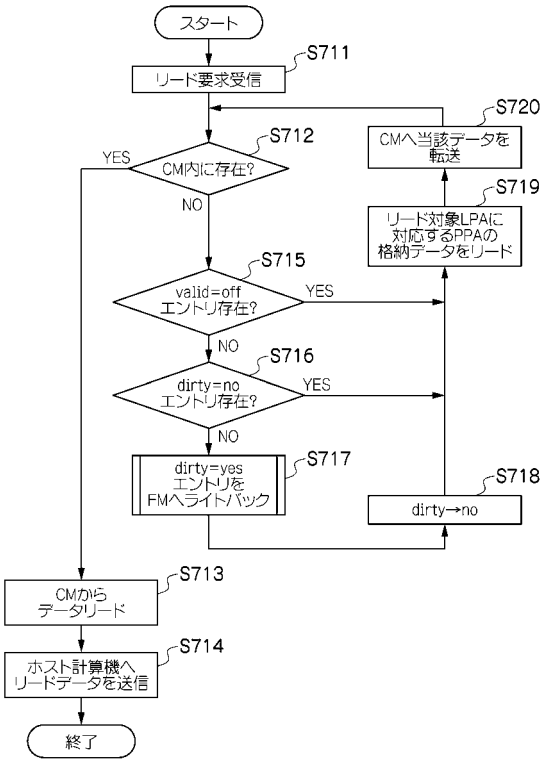



【図１２】

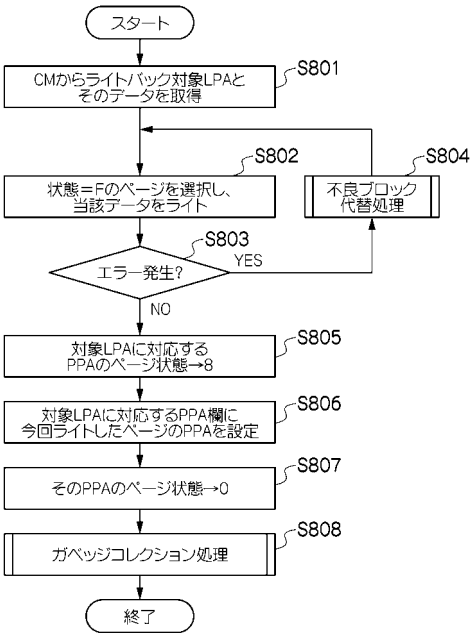
図12



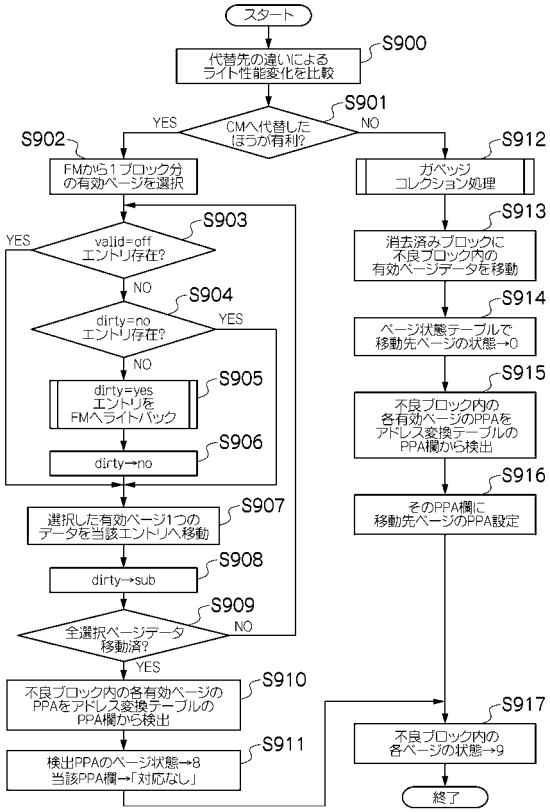
【図１３】
図13



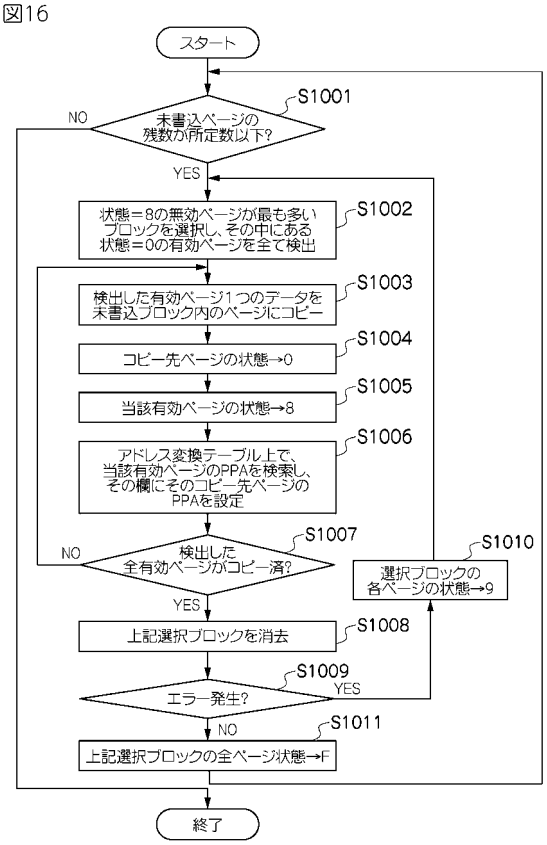

【図１４】
図14




【図１５】
図15




【図１６】
図16



フロントページの続き

| (51)Int.Cl. | FI | | テーマコード（参考） |
|---|---|---|---|
| | Ｇ０６Ｆ 12/16 | ３２０Ｌ | |
| | Ｇ０６Ｆ 12/08 | ５５７ | |
| | Ｇ０６Ｆ 12/08 | ５４１Ｂ | |
| | Ｇ０６Ｆ 12/08 | ５４１Ｚ | |

Ｆターム(参考) 5B005 JJ11 MM11 VV13 WW14
5B018 GA04 GA06 GA10 HA04 HA35 KA14 KA18 MA22 NA01 NA06
QA16 RA11
5B065 BA01 BA09 CH02 CS01
5B082 JA07