(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6678230号
(P6678230)
(45)発行日　令和2年4月8日(2020.4.8)　(24)登録日　令和2年3月18日(2020.3.18)

(51)Int.Cl.　　F I
G06F 3/06 (2006.01)　　G06F 3/06 301Z
G06F 16/185 (2019.01)　　G06F 3/06 301W
G06F 16/185

請求項の数 12　(全 31 頁)

(21)出願番号　特願2018-502866(P2018-502866)
(86)(22)出願日　平成28年2月29日(2016.2.29)
(86)国際出願番号　PCT/JP2016/056018
(87)国際公開番号　WO2017/149592
(87)国際公開日　平成29年9月8日(2017.9.8)
審査請求日　平成30年6月4日(2018.6.4)

(73)特許権者　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　110000062
特許業務法人第一国際特許事務所
(72)発明者　高岡　伸光
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内
(72)発明者　山本　彰
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内
(72)発明者　川口　智大
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内

最終頁に続く

(54)【発明の名称】ストレージ装置

(57)【特許請求の範囲】
【請求項１】
ホスト計算機から書き込み要求のあったライトデータを格納するための１以上の記憶デバイスと、前記ホスト計算機に１以上のボリュームを提供するストレージコントローラと、を有し、
前記ストレージコントローラは、前記ホスト計算機から前記ボリューム内の区画に対するライト要求とライト対象データを受領すると、前記区画に、前記区画と同サイズの前記記憶デバイスの第１記憶領域を割り当て、前記割り当てられた前記第１記憶領域に前記ライト対象データを格納し、
前記ストレージコントローラは、前記ボリューム内の区画ごとに、前記ホスト計算機から最後にライト要求を受け付けた時刻である最終ライト時刻を保持しており、
前記ストレージコントローラは、前記最終ライト時刻から所定の期間以上ライト要求を受け付けていない前記区画について、前記ホスト計算機から前記区画に書き込まれたデータと同一のデータが前記記憶デバイスに格納済みの場合、前記区画に書き込まれたデータを前記記憶デバイスに格納しないようにする重複排除処理を実施し、
前記ストレージコントローラは、前記重複排除処理において、前記区画を複数の重複排除ブロックに区分し、前記重複排除ブロックごとに、前記重複排除ブロックと同サイズで前記第１記憶領域と異なる記憶領域であるデータブロックを割り当て、前記重複排除ブロックと前記データブロックとのマッピングを管理するよう構成されており、
前記ストレージコントローラは、前記重複排除ブロックのうち第２の重複排除ブロック

に前記データブロックを割り当てる時、前記第２の重複排除ブロックに書き込まれているデータが、第１の重複排除ブロックに割り当てられている第１データブロックに格納されているデータと同一の場合、前記第２の重複排除ブロックに前記第１データブロックを割り当てる、
ことを特徴とする、ストレージ装置。
【請求項２】
前記ストレージコントローラは、前記区画ごとに状態を管理しており、
前記区画の状態として、前記重複排除処理が実施されていない状態である第１状態と、前記重複排除処理が実施された状態である第２状態と、前記第２状態から前記重複排除処理が行われる前の状態に戻された状態である第３状態とがあり、
前記ストレージコントローラは、前記第１状態の前記区画に対して前記重複排除処理を行った結果、前記区画の重複排除率が所定の閾値未満だった場合、前記区画の状態を前記第３状態に変更する
ことを特徴とする、請求項１に記載のストレージ装置。
【請求項３】
前記第３状態の前記区画は、前記ホスト計算機からのライト要求が発行されるまでは前記状態が変更されない、
ことを特徴とする、請求項２に記載のストレージ装置。
【請求項４】
前記ストレージコントローラは、前記ホスト計算機から前記第２状態または前記第３状態の前記区画に対してライト要求が発行されると、前記区画の状態を前記第１状態に変更する、
ことを特徴とする、請求項３に記載のストレージ装置。
【請求項５】
前記ストレージコントローラは、前記重複排除ブロックに書き込まれたデータの特徴量を算出し、検索テーブルに、前記算出された特徴量と前記重複排除ブロックに割り当てられた前記データブロックとのマッピングを記録し、
前記ストレージコントローラは、前記第２の重複排除ブロックの特徴量を算出すると、前記第２の重複排除ブロックの特徴量と同一の値が前記検索テーブルに格納されていない場合、前記第２の重複排除ブロックに第２データブロックを割り当て、前記第２データブロックに前記第２の重複排除ブロックに書き込まれたデータを格納する、
ことを特徴とする、請求項４に記載のストレージ装置。
【請求項６】
ホスト計算機から書き込み要求のあったライトデータを格納するための１以上の記憶デバイスと、前記ホスト計算機に１以上のボリュームを提供するストレージコントローラと、を有するストレージ装置において、
前記ストレージコントローラが、前記ホスト計算機から前記ボリューム内の区画に対するライト要求とライト対象データを受領すると、前記区画に、前記区画と同サイズの前記記憶デバイスの第１記憶領域を割り当て、前記割り当てられた前記第１記憶領域に前記ライト対象データを格納するステップと、
前記ストレージコントローラが、前記ボリューム内の区画ごとに、前記ホスト計算機から最後にライト要求を受け付けた時刻である最終ライト時刻を記録するステップと、
前記最終ライト時刻から所定の期間以上ライト要求を受け付けていない前記区画を検出するステップと、
前記検出された区画について、前記区画に書き込まれたデータのうち、前記記憶デバイスに格納済みのデータと異なるデータのみを前記記憶デバイスに格納する重複排除処理を実施するステップと、を実行し、
前記ストレージコントローラは、前記重複排除処理において、前記区画を複数の重複排除ブロックに区分し、前記重複排除ブロックごとに、前記重複排除ブロックと同サイズで前記第１記憶領域と異なる記憶領域であるデータブロックを割り当て、前記重複排除ブロ

ックと前記データブロックとのマッピングを管理するよう構成されており、
前記ストレージコントローラが、前記重複排除ブロックのうち第２の重複排除ブロックに前記データブロックを割り当てる時、前記第２の重複排除ブロックに書き込まれているデータが、第１の重複排除ブロックに割り当てられている第１データブロックに格納されているデータと同一の場合、前記第２の重複排除ブロックに前記第１データブロックを割り当てるステップを実行する、
ことを特徴とする、ストレージ装置の制御方法。
【請求項７】
前記ストレージコントローラは、前記区画ごとに状態を管理しており、
前記区画の状態として、前記重複排除処理が実施されていない状態である第１状態と、
前記重複排除処理が実施された状態である第２状態と、前記第２状態から前記重複排除処理が行われる前の状態に戻された状態である第３状態とがあり、
前記重複排除処理を実施するステップでは、前記ストレージコントローラが、前記第１状態の前記区画に対して前記重複排除処理を行った結果、前記区画の重複排除率が所定の閾値未満だった場合、前記区画の状態を前記第３状態に変更する、
ことを特徴とする、請求項６に記載のストレージ装置の制御方法。
【請求項８】
前記第３状態の前記区画は、前記ホスト計算機からのライト要求が発行されるまでは前記状態が変更されない、
ことを特徴とする、請求項７に記載のストレージ装置の制御方法。
【請求項９】
前記ストレージコントローラは、前記ホスト計算機から前記第２状態または前記第３状態の前記区画に対してライト要求が発行されると、前記区画の状態を前記第１状態に変更するステップを実行する、
ことを特徴とする、請求項８に記載のストレージ装置の制御方法。
【請求項１０】
ライトデータを格納するための１以上の記憶デバイスを有する計算機のプロセッサで実行されるプログラムを記録した記憶媒体であって、前記プログラムは前記プロセッサに、
ボリュームに対するライト要求とライト対象データを受け付けるステップと、
前記プロセッサは、前記計算機から前記ボリューム内の区画に対するライト要求とライト対象データを受領すると、　前記区画に、前記区画と同サイズの前記記憶デバイスの第１記憶領域を割り当て、前記割り当てられた前記第１記憶領域に前記ライト対象データを格納するステップと、
前記ボリューム内の区画ごとに、最後にライト要求を受け付けた時刻である最終ライト時刻を保持するステップと、
前記最終ライト時刻から所定の期間以上ライト要求を受け付けていない前記区画を検出するステップと、
前記検出された区画について、前記区画に書き込まれたデータのうち、前記記憶デバイスに格納済みのデータと異なるデータのみを前記記憶デバイスに格納する重複排除処理を実施するステップと、を実行させ、
前記プロセッサは、前記重複排除処理において、前記区画を複数の重複排除ブロックに区分し、前記重複排除ブロックごとに、前記重複排除ブロックと同サイズで前記第１記憶領域と異なる記憶領域であるデータブロックを割り当て、前記重複排除ブロックと前記データブロックとのマッピングを管理するよう構成されており、
前記プロセッサが、前記重複排除ブロックのうち第２の重複排除ブロックに前記データブロックを割り当てる時、前記第２の重複排除ブロックに書き込まれているデータが、第１の重複排除ブロックに割り当てられている第１データブロックに格納されているデータと同一の場合、前記第２の重複排除ブロックに前記第１データブロックを割り当てるステップを実行させる、
ことを特徴とする、プログラムを記録した記憶媒体。

【請求項１１】
前記プロセッサは、前記区画ごとに状態を管理しており、
前記区画の状態として、前記重複排除処理が実施されていない状態である第１状態と、前記重複排除処理が実施された状態である第２状態と、前記第２状態から前記重複排除処理が行われる前の状態に戻された状態である第３状態とがあり、
前記重複排除処理を実施するステップでは、前記プロセッサが前記第１状態の前記区画に対して前記重複排除処理を行った結果、前記区画の重複排除率が所定の閾値未満だった場合、前記区画の状態を前記第３状態に変更させる、
ことを特徴とする、請求項１０に記載のプログラムを記録した記憶媒体。
【請求項１２】
前記プロセッサに、前記第２状態または前記第３状態の前記区画に対するライト要求を受領すると、前記区画の状態を前記第１状態に変更するステップを実行させる、
ことを特徴とする、請求項１１に記載のプログラムを記録した記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ストレージ装置に関する。
【背景技術】
【０００２】
ストレージ装置は、データを格納する複数の記憶デバイスと、記憶デバイスを制御するストレージコントローラとを有しており、ホスト計算機に大容量のデータ格納空間を提供することを目的としている。
【０００３】
ストレージ装置には、低コストで大量のデータを保存することが求められる。こうした要求を満たすために、ホストから受領したライトデータのサイズを縮小して記憶デバイスに記録する技術が知られている。ライトデータのサイズを縮小してから記憶デバイスに記録すると、データの保持コスト（記憶媒体のビットコスト、ストレージ装置の消費電力コスト等）を削減できる。データサイズを縮小するために、可逆圧縮アルゴリズムを用いて、データの意味を保ったままデータサイズを縮小する技術がある。この処理は「可逆圧縮」または「圧縮」と呼ばれる。
【０００４】
データを圧縮して記憶デバイスに記録する場合、ライト時には圧縮処理、リード時には圧縮データの伸長処理のオーバヘッドが発生するため、アクセス性能が低下することがある。つまり格納データ量の削減とアクセス性能はトレードオフの関係にある。これを避けるために、選択的にデータの圧縮を行う技術が存在する。たとえば特許文献１には、複数の記憶階層（ｔｉｅｒ）を管理するストレージシステムにおいて、例えば下位ｔｉｅｒに移動されるデータを圧縮して格納することで、アクセス性能の低下を抑止しつつ、格納データ量を削減する方法が開示されている。
【０００５】
また格納すべきデータ量を削減するもう一つの技術として、重複排除技術がある。たとえばストレージ装置に同内容のデータが複数個存在していることをストレージ装置が検出した時、そのうちの１つだけをストレージ装置内の記憶デバイスに残し、残りのデータは記憶デバイスに格納しないようにする技術である。重複排除技術と可逆圧縮技術の何れが用いられても、ホストから受領したライトデータの量より記憶デバイスに格納されるライトデータの量が小さくなる。そのため重複排除技術も広義には圧縮技術の１つといえる。
【先行技術文献】
【特許文献】
【０００６】
【特許文献１】米国特許８３５９４４４号明細書
【発明の概要】

【発明が解決しようとする課題】
【０００７】
データを圧縮する場合、データ内容に依存して圧縮率（あるいはデータの縮小量）が異なり得る。そのため、データに圧縮または重複排除処理が施されたが、記憶デバイスに格納されるデータ量が殆ど削減されないという事も起こり得る。その場合、アクセス性能も低下し、かつデータの保持コストも低下しないことになる。低ビットコストで高性能なストレージ装置を提供するためには、このような事態が発生することを防ぐ必要がある。
【課題を解決するための手段】
【０００８】
本発明の一観点に係るストレージ装置は、ホスト計算機から書き込み要求のあったライトデータを格納するための１以上の記憶デバイスと、ホスト計算機に１以上のボリュームを提供するストレージコントローラとを有し、ボリューム内の区画ごとに、ホスト計算機から最後にライト要求を受け付けた時刻である最終ライト時刻を保持する。そしてストレージコントローラは、最終ライト時刻から所定の期間以上ライト要求を受け付けていない区画を検出すると重複排除処理を実施する。またストレージコントローラは、区画の重複排除処理の結果、重複排除率が低い区画については、区画を重複排除処理の実施されていない状態へと戻す処理を実施する。
【発明の効果】
【０００９】
本発明によれば、低ビットコストで高性能なストレージ装置を提供することが可能になる。
【図面の簡単な説明】
【００１０】
【図１】実施例に係るストレージ装置を含む計算機システムの論理構成図である。
【図２】論理ページの状態遷移を表した図である。
【図３】ストレージ装置の構成図である。
【図４】論理ページと物理ページのマッピング関係を説明する図である。
【図５】論理ページ管理テーブルの構成例である。
【図６】マッピングテーブルの構成例である。
【図７】プール管理テーブルの構成例である。
【図８】検索テーブルの構成例である。
【図９】逆参照テーブルの構成例である。
【図１０】追記ポインタの構成例である。
【図１１】重複排除処理部のフローチャートである。
【図１２】重複排除処理の流れを表した図である。
【図１３】重複排除解除部のフローチャートである。
【図１４】移行処理のフローチャートである。
【図１５】ライト処理のフローチャートである。
【図１６】リード処理のフローチャートである。
【図１７】物理ページ解放処理のフローチャートである。
【図１８】共有データ判定・複製処理のフローチャートである。
【図１９】マッピング切替処理のフローチャートである。
【発明を実施するための形態】
【００１１】
以下、幾つかの実施例について、図面を用いて説明する。
【００１２】
なお、以下の実施例において、ストレージ装置内で実行される処理について、「プログラム」を主語として説明を行う場合がある。実際には、ストレージ装置が有するプロセッサ（ＣＰＵ）がプログラムを実行することによって、プログラムに記述された処理が行われるため、処理の主体はプロセッサ（ＣＰＵ）であるが、説明が冗長になることを防ぐた

め、プログラムを主語にして処理の内容を説明することがある。また、プログラムの一部または全ては専用ハードウェアによって実現されてもよい。また、以下で説明される各種プログラムは、プログラム配布サーバや計算機が読み取り可能な記憶メディアによって提供され、プログラムを実行する各装置にインストールされてもよい。計算機が読み取り可能な記憶メディアとは、非一時的なコンピュータ可読媒体で、例えばＩＣカード、ＳＤカード、ＤＶＤ等の不揮発性記憶媒体である。

【００１３】

実施例の説明に入る前に、実施例で用いられる各種用語について説明する。

【００１４】

「ボリューム」とは、ストレージ装置や記憶デバイス等のターゲットデバイスが、ホスト計算機等のイニシエータデバイスに提供する記憶空間のことを意味する。イニシエータデバイスが記憶空間上の領域に対するデータ書き込み要求を発行すると、その領域に対応付けられているターゲットデバイス上の領域にデータが格納される。本実施例に係るストレージ装置はボリュームとして、いわゆるＴｈｉｎ　Ｐｒｏｖｉｓｉｏｎｉｎｇ技術により形成される仮想ボリュームをホストに提供する。仮想ボリュームは、その初期状態（仮想ボリュームが定義された直後）では、記憶空間上の領域に記憶デバイスが対応付けられていない。イニシエータデバイス（ホスト）が記憶空間上の領域にデータ書き込み要求を発行した時点で、ストレージ装置はその領域に対応付けられる記憶デバイスを動的に決定する。

【００１５】

「重複排除処理」とは、ストレージ装置内に同内容のデータが複数存在する場合、１つだけをストレージ装置に残し、それ以外のデータをストレージ装置から削除する処理である。ストレージ装置内に同内容のデータが存在するか判定する処理のことを、「重複判定」処理と呼ぶ。なお、特に断りのない限り、重複排除処理は重複判定処理を含む処理である。

【００１６】

以下で説明する実施例に係るストレージ装置では、重複排除ブロックと呼ばれる所定サイズのデータ毎に重複判定を行う。以下の実施例では、重複排除ブロックのサイズが８ＫＢの例について説明されるが、重複排除ブロックのサイズは８ＫＢ以外のサイズであってもよい。同内容のデータのことを「重複データ」と呼ぶ。

【００１７】

重複判定の際、２つのデータをビット単位あるいはバイト単位で比較すると、判定処理に長時間を要することになる。そのため一般的には重複判定を行う装置は、比較対象のデータに所定の演算（たとえばハッシュ関数を用いた演算等）を行うことで、小サイズ（たとえば８バイト程度）の特徴量を生成し、それを用いて重複判定を行う。以下の実施例では、データから生成される特徴量のことを、「フィンガープリント」と呼ぶ。フィンガープリントは、ＦＰと略記されることもある。

【００１８】

以下で説明する実施例では、データＡから算出されたＦＰの値がＨであった場合、値ＨはデータＡのＦＰと呼ばれる。逆にデータＡのことを、「ＦＰ　Ｈを持つデータ」と呼ぶことがある。また、データＡの書き込まれる領域（重複排除ブロック）のことも、「ＦＰ　Ｈを持つ領域（重複排除ブロック）」と呼ぶことがある。

【００１９】

本実施例において「衝突」とは、複数の異なるデータそれぞれに対して所定の演算を施してＦＰを生成した時、生成されたそれぞれのＦＰが同一になることを意味する。ハッシュ関数などを用いて、小サイズの特徴量を算出する場合、衝突は発生し得る。

【００２０】

「重複排除率」とは、重複排除処理による記憶領域消費量の削減効率を表す指標値である。たとえば重複排除率は、ストレージ装置のボリュームに書き込まれたデータ量と、ストレージ装置がデータ格納のために使用（消費）した記憶領域の量の比で表される値であ

る。重複排除処理では、同内容のデータが多数格納されると、１つだけがストレージ装置の記憶領域に書き込まれ（記憶領域が消費され）、それ以外のデータは記憶領域に書き込まれない（記憶領域が消費されない）ので、記憶領域消費量の削減効率が高くなる。

【実施例】

【００２１】

（１）発明の概要

まず図１、図２、そして図４を用いて、本発明の実施例に係るストレージ装置が実施する、重複排除方法の概要を説明する。図１は、本発明の実施例に係るストレージ装置が計算機に提供する仮想ボリュームの構成を表した図である。

【００２２】

ストレージ装置１は、複数の記憶デバイス（図１では非図示）を有し、ホスト計算機５（以下では「計算機５」と略記する）からのライトデータを記憶デバイスに格納する。記憶デバイスは、所定サイズの記憶空間をストレージ装置１に提供するが、ストレージ装置１は、記憶デバイスの提供する記憶空間を直接計算機５には提供しない。計算機５には、記憶デバイスの有する記憶空間とは異なる、１以上の仮想的な記憶空間を提供する。この仮想的な記憶空間を「仮想ボリューム」と呼ぶ。図１では、２つの仮想ボリューム（仮想ボリューム１６、仮想ボリューム２０）が計算機５に提供される例が示されている。

【００２３】

ストレージ装置１は、仮想ボリュームの記憶空間を、複数の所定サイズ（一例として４２ＭＢ）の区画に分割して管理している。本実施例では、この区画のことを「論理ページ」と呼ぶ。各論理ページには仮想ボリューム内で一意な識別子が付されており、この識別子を論理ページ識別子（または論理ページ番号）と呼ぶ。

【００２４】

仮想ボリュームは、公知のＴｈｉｎ　Ｐｒｏｖｉｓｉｏｎｉｎｇ技術などを用いて形成されるボリュームであり、ストレージ装置１は仮想ボリュームの論理ページに対するアクセス要求を受け付けた時点で、記憶デバイスの記憶領域を動的に論理ページに割り当てる（マップする）。言い換えると、各論理ページには、計算機５からアクセス要求を受け付けるまでは、その論理ページに記憶領域が割り当てられていない。

【００２５】

ストレージ装置１はまた、論理ページに対して割り当てる記憶領域（これは記憶デバイスが提供する記憶領域である）を管理するための管理概念を有しており、これを「ストレージプール」または「プール」と呼ぶ。図１では、円筒状のオブジェクト１７としてストレージプールが表現されている。ストレージ装置１は、プールの記憶領域を、論理ページと同サイズの領域（あるいは論理ページより大きいサイズの領域でも良い）に区分し、区分された領域ごとに識別子を付して管理する。この区分された領域は「物理ページ」と呼ばれ、物理ページに付される識別子は「物理ページ識別子」または「物理ページ番号」と呼ばれる。

【００２６】

ストレージ装置１が計算機５から、仮想ボリュームに対するライト要求を受信すると、ライト要求に含まれているライト対象領域のアドレスを論理ページ番号に変換し、ライト対象領域を含む論理ページを特定する。特定された論理ページに物理ページが割り当てられていない場合、ストレージ装置１は、プール１７内の未使用の物理ページ（まだ論理ページに割り当てられていない物理ページ）を選択し、アクセス対象の論理ページに、選択された物理ページを割り当てる（マップする）。計算機５からのライトデータは、このアクセス対象論理ページにマップされた物理ページに格納される。

【００２７】

またストレージ装置１は、論理ページと、論理ページに割り当てられた物理ページとの対応関係（マッピング）を論理ページ管理テーブル１２６に記憶している。論理ページに対するリード要求を受け付けた時には、ストレージ装置１は論理ページ管理テーブル１２６を参照することで、論理ページに割り当てられた記憶領域を特定し、特定された記憶領

域からデータを読み出す。

【００２８】

本実施例に係るストレージ装置１は、計算機５からライトデータの書き込まれた論理ページのうち、所定の条件に該当する論理ページを重複排除処理対象として扱うことと決定し、その論理ページに対して重複排除処理を行う。所定の条件とはたとえばライト頻度の低い論理ページで、具体的には、所定時間以上ライトデータの書き込みがない論理ページである。重複排除処理の行われた論理ページのことを、「重複排除論理ページ」と呼ぶ。逆に、重複排除論理ページでない論理ページのことは、「通常論理ページ」と呼ばれることがある。

【００２９】

図４は、通常論理ページに物理ページが割り当てられた状態と、重複排除論理ページに物理ページが割り当てられた状態の概念図である。図４（ａ）は、通常論理ページに物理ページが割り当てられた状態を表している。１つの通常論理ページ１６１ｂには１つの物理ページ１７１ｃが割り当てられる。たとえば計算機５が通常論理ページ１６１ｂの先頭からｋバイト目の領域３６１にデータをライトする要求をストレージ装置１に発行すると、ストレージ装置１は、通常論理ページ１６１ｂに割り当てられた物理ページ１７１ｃの先頭からｋバイト目の領域３７１に、ライト要求で指定されたデータを格納する。この関係が維持されている為、ストレージ装置１は、通常論理ページと物理ページのマッピングを管理する際には、論理ページごとに、論理ページに割り当てられている物理ページの物理ページ番号だけを管理すればよい。

【００３０】

図４（ｂ）は、重複排除論理ページに物理ページが割り当てられた状態を表している。ストレージ装置１は、論理ページ内の領域を、所定サイズ（たとえば８ＫＢ）の部分領域に区分し、この部分領域ごとに物理ページの領域を割り当てる。本実施例ではこの部分領域ごとに重複判定が行われるので、これを「重複排除ブロック」と呼ぶ。一方、重複排除ブロックに割り当てられる物理ページ上の領域は、「データブロック」と呼ばれる。

【００３１】

なお、通常論理ページが重複排除論理ページとして扱われるようになると、これまで通常論理ページに割り当てられていた物理ページは削除される（割り当てられていない状態にされる）。そしてこれまで物理ページに格納されていたデータは、重複排除ブロックに割り当てられる物理ページ上の領域（データブロック）に移動される。データブロックは、プール１７に属する物理ページ上の領域である。重複排除ブロックとデータブロックのマッピング情報は、後述するマッピングテーブル１２７に格納される。

【００３２】

ストレージ装置１が行う重複排除処理では、重複排除ブロックごとにデータの比較を行う。図４（ｂ）を用いて重複排除処理について概説する。図４（ｂ）では、１６１ａは重複排除論理ページであり、重複排除ブロック１８ｂには物理ページ１７１ｄ内のデータブロック１９ｃが割り当てられている。ここでストレージ装置１が論理ページ２０１ｘに対して重複排除処理を行う場合を想定する。ストレージ装置１は、論理ページ２０１ｘの重複排除ブロック２１ｃに対して書き込まれたデータと、物理ページ１７１ｄの各データブロックに格納されているデータの内容を比較する。比較の結果、重複排除ブロック２１ｃに対して書き込まれたデータがデータブロック１９ｃに格納されているデータと同一だった場合、ストレージ装置１は重複排除ブロック２１ｃにデータブロック１９ｃを割り当てる。そしてストレージ装置１は、重複排除ブロック２１ｃに対して書き込まれたデータを、新たに物理ページ１７１ｄに書き込むことはしない。この結果、データブロック１９ｃは、２つの重複排除ブロック（１８ｂと２１ｃ）に割り当てられることになる。本実施例では、データブロックが複数の重複排除ブロックに割り当てられた状態のことを、「データブロックが複数の重複排除ブロックに共有されている」または「複数の重複排除ブロックが１つのデータブロックを共有している」と表現する。

【００３３】

これにより、ストレージ装置１は計算機５から重複排除ブロック１８ｂと重複排除ブロック２１ｃのいずれに対するリード要求を受け付けた場合でも、データブロック１９ｃからデータＡを読み出して、計算機５に返送する。また、重複排除ブロック１８ｂと重複排除ブロック２１ｃのそれぞれに対して異なるデータブロックを割り当てる必要がないため、実質的にストレージプール１７への格納データ量（言い換えると、記憶デバイスの記憶領域消費量）が１／２に削減される効果がある。もしｎ個の重複排除ブロックが１つのデータブロックを共有している場合、記憶領域消費量は１／ｎに削減されることになる。

【００３４】

このように、本実施例に係るストレージ装置１では、論理ページが、重複排除論理ページとして扱われる場合と通常論理ページとして扱われる場合がある。ストレージ装置１では、各論理ページがいずれの状態にあるかを管理するため、各論理ページの状態についての情報を保持する。図２を用いて、論理ページの各状態の説明を行う。

【００３５】

図２は、ある論理ページの状態遷移を表した図である。論理ページは、Ｐ，Ｘ，Ｑのいずれかの状態を有する。初期状態（仮想ボリュームが定義された直後で、論理ページに物理ページが割り当てられていない）では、論理ページの状態はＰである。論理ページに計算機５からデータのライトが行われると、ストレージ装置１はその論理ページの状態をＸに変更する。

【００３６】

状態がＸの論理ページに対して、一定以上の間計算機５からのライト要求が到来しなかった場合、ストレージ装置１はその論理ページの状態をＱに変更する。状態がＱの論理ページは、重複排除論理ページであり、重複排除処理が行われる。状態がＱの論理ページ（重複排除論理ページ）に対して、計算機５からのライト要求が到来すると、ストレージ装置１は再びその論理ページの状態をＸに変更する。重複排除論理ページに対してデータを格納する場合、重複判定等の処理のオーバヘッドが増加し、アクセス性能の低下を招くためである。

【００３７】

計算機５からのライト要求が到来しない場合には、重複排除論理ページは原則としてその状態（状態Ｑ）が維持される。ただし、その重複排除論理ページの記憶領域消費量の削減効果が薄い場合、たとえばその重複排除論理ページの各重複排除ブロックに書き込まれたデータの大半が、他のデータと異なっている場合は、ストレージ装置１のビットコスト削減に寄与せず、かつアクセス性能の低下を招くこともある。そのためストレージ装置１は、記憶領域消費量の削減効果が薄い重複排除論理ページがあった場合、その重複排除論理ページの状態をＱからＰへと遷移させる。

【００３８】

ストレージ装置１は、状態をＱからＰに変更させるべき重複排除論理ページを決定する際、論理ページの重複排除率を用いる。なお、本実施例における“論理ページの重複排除率”とは、以下の計算式により定義されるものである。

【００３９】

論理ページ内の重複排除ブロック数をＰ、論理ページ（重複排除論理ページ）の重複排除ブロックのうち、既に別の重複排除ブロックへ割り当てられているデータブロックが割り当てられた重複排除ブロックの数をＷとする（別の表現をすると、重複排除論理ページに割り当てられたデータブロックのうち、複数の重複排除ブロックに共有されているデータブロックの数がＷである）。なお論理ページ内の重複排除ブロック数Ｐは、論理ページサイズ（たとえば４２ＭＢ）÷重複排除ブロックのサイズ（たとえば８ＫＢ）で求まる固定値である。このとき論理ページの重複排除率Ｄは、以下の計算式で表現される。

Ｄ＝Ｗ÷Ｐ

【００４０】

Ｐは固定値のため、Ｗが大きいと重複排除率Ｄも大きくなる。そのため、Ｄが大きい（１に近い）場合、記憶デバイスの記憶領域の消費量が少ない（重複排除による記憶領域の

削減効果が高い）ことを意味し、逆にＤが小さい（０に近い場合）は、重複排除による記憶領域の削減効果が小さいことを意味する。本実施例ではストレージ装置１は、重複排除率が所定の閾値未満の重複排除論理ページがあった場合、その重複排除論理ページの状態をＱからＰへと遷移させる。

【００４１】

論理ページの状態がＱからＰに変更される場合、ストレージ装置はその論理ページのデータを、一旦ストレージ装置１内のキャッシュメモリ等の一時的な記憶領域上に読み出し、論理ページに新たな物理ページを割り当てて、割り当てられた物理ページにデータを書き戻すことで、重複排除処理が行われていない状態にする。本実施例に係るストレージ装置は、論理ページのアクセス頻度や記憶領域消費量の削減効率を考慮して、重複排除処理の対象にする論理ページを制限することで、高アクセス性能と低ビットコストの両立を図っている。

【００４２】

なお、状態Ｐの論理ページと状態Ｘの論理ページは、以下の点で異なる。状態Ｘの論理ページは、状態Ｑに変更されることがある。具体的には、状態Ｘのある論理ページに、計算機５からのライト要求が一定の時間以上到来しなかった場合に、その論理ページの状態はＱに変更される。一方状態Ｐの論理ページは、状態がＱに変更されることはない。状態Ｐの論理ページは、記憶領域消費量の削減効果が薄いと判定されたために、状態がＱからＰに変更された論理ページだからである（あるいは、まだ計算機５から一度もライトデータの書き込みが行われておらず、物理ページが割り当てられていない論理ページである）。そのような論理ページの状態をＱに変更しても、記憶領域消費量の削減効果が小さい（またはない）ことは明らかであるため、ストレージ装置１は状態Ｐの論理ページの状態をＱに変更することは行わない。これにより、頻繁に論理ページの状態の変更が発生しないようにしている（処理オーバヘッドの増加を抑制している）。

【００４３】

（２）システム構成

図３は、本実施例に係るストレージ装置１を含む計算機システムのハードウェア構成例を示している。ストレージ装置１は、ストレージコントローラ１０と、ストレージコントローラ１０に接続された複数の記憶デバイス１５を有する。

【００４４】

記憶デバイス１５は、ストレージ装置１が計算機５などの上位装置からのライトデータを記憶するために用いられる。記憶デバイスとしては、たとえば磁気ディスクを記憶媒体として用いるＨＤＤ（Ｈａｒｄ　Ｄｉｓｋ　Ｄｒｉｖｅ）や、フラッシュメモリ等の不揮発性半導体メモリを記憶媒体として採用したＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）が用いられる。本実施例では、記憶デバイス１５は「ドライブ１５」と表記されることもある。記憶デバイス１５は一例として、ＳＡＳ（Ｓｅｒｉａｌ　Ａｔｔａｃｈｅｄ　ＳＣＳＩ）規格に従う伝送線（ＳＡＳリンク）や、ＰＣＩ（Ｐｅｒｉｐｈｅｒａｌ　Ｃｏｍｐｏｎｅｎｔ　Ｉｎｔｅｒｃｏｎｎｅｃｔ）規格に従う伝送線（ＰＣＩリンク）などによって、ストレージコントローラ１０と接続される。

【００４５】

ストレージコントローラ１０は少なくとも、プロセッサ（ＣＰＵとも呼ばれる）１１、システムメモリ１２、キャッシュメモリ１４、そしてＳＡＮ（Ｓｔｏｒａｇｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）６に接続するためのインタフェース（非図示）を有する。ＳＡＮ６は、一例としてファイバチャネルを用いて形成されるネットワークである。

【００４６】

プロセッサ１１は、ストレージ装置１の各種制御を行う。システムメモリ１２は、ＣＰＵ１１が実行するプログラムや、ＣＰＵ１１がプログラム実行の際に使用する管理情報を格納するためのものである。一方、キャッシュメモリ１４は、計算機５から受領したライトデータや、記憶デバイス１５から読み出されたリードデータを一時的に記憶するために用いられる。

【００４７】
システムメモリ１２やキャッシュメモリ１４には、ＤＲＡＭ、ＳＲＡＭ等の揮発性記憶媒体が用いられるが、別の実施形態として、不揮発性メモリを用いてキャッシュメモリ１４を構成してもよい。また、キャッシュメモリ１４に揮発性記憶媒体が用いられる場合、ストレージ装置１にバッテリ等の補助電源を搭載し、停電時にキャッシュメモリ１４の記憶内容を維持できるように構成されていてもよい。
【００４８】
また、別の実施形態として、ストレージ装置１は２種類のメモリ（システムメモリ１２やキャッシュメモリ１４）を有さない構成でも良い。つまりストレージ装置１が１種類のメモリのみを有する構成でも良い。その場合、プログラム、管理情報、ライトデータなどは、同じメモリに格納される。
【００４９】
計算機５は、ストレージ装置１へのアクセス要求発行元となる装置である。計算機５は、ＰＣ（パーソナルコンピュータ）等の汎用のコンピュータであり、少なくともプロセッサとメモリ（非図示）を有する。プロセッサでは、ストレージ装置１が提供する仮想ボリュームにＩ／Ｏ要求を発行するアプリケーションプログラムなどが実行される。
【００５０】
（３）管理情報
続いてストレージ装置１の有する管理情報、プログラムの内容の説明を行う。ストレージ装置１のシステムメモリ１２には少なくとも、論理ページ管理テーブル１２６、マッピングテーブル１２７、プール管理テーブル１２８、検索テーブル１２９、逆参照テーブル１３０、追記ポインタ１３１の、６種類の管理情報が格納されている。以下では、これらの各種管理情報の内容について説明する。
【００５１】
図５に、論理ページ管理テーブル１２６の構成を示す。論理ページ管理テーブル１２６は、各論理ページの状態を管理するためのテーブルで、各行（レコード）に、それぞれの論理ページの状態や属性情報が格納される。以下、論理ページ管理テーブル１２６の各カラムに格納される情報について説明する。
【００５２】
仮想ボリューム１２６１、論理ページ１２６２にはそれぞれ、論理ページの属する仮想ボリュームの識別子、論理ページの識別子が格納される。
【００５３】
重複排除１２６３には、“有効”または“無効”が格納される。あるレコードの重複排除１２６３に、“有効”が格納されている場合、そのレコードで管理される論理ページは重複排除論理ページであることを意味し、“無効” が格納されている場合、そのレコードで管理される論理ページは通常論理ページであることを意味する。
【００５４】
物理ページ１２６４には、論理ページに割り当てられている物理ページの識別子が格納される。論理ページに物理ページが割り当てられていない場合には、ＮＵＬＬが格納される。
【００５５】
なお、一般に仮想ボリュームや論理ページの識別子には非負の整数値が使われるが、図５では説明の都合上、仮想ボリューム１２６１、論理ページ１２６２、物理ページ１２６４の各カラムには、図１等に記載の仮想ボリュームに付された参照番号、論理ページに付された参照番号、物理ページに付された参照番号が格納されている。また、これ以降で説明される管理情報についても同様に、仮想ボリューム、論理ページ、物理ページ等の識別子を格納するためのカラムには、図１等に記載の仮想ボリューム、論理ページ、物理ページ等に付された参照番号が格納されている。
【００５６】
状態１２６５には、論理ページの状態が格納される。論理ページの状態には、先に述べ

たとおり、Ｐ，Ｑ，Ｘの３つがある。最終ライト時刻１２６６には、論理ページに対して計算機５からライト要求を受領した最新の時刻が格納される。

【００５７】

排除ブロック数１２６７は、論理ページが重複排除論理ページである場合に有効な情報である。排除ブロック数１２６７には、論理ページ（重複排除論理ページ）内の重複排除ブロックのうち、重複排除処理によって、既に別の重複排除ブロックへ割り当てられているデータブロックがさらに割り当てられた重複排除ブロックの数が記録される。排除ブロック数１２６７が大きいほど、その論理ページの記憶領域消費量の削減効果が大きいといえる。以降、このように、重複排除処理によって既存のデータブロックが割り当てられることで、記憶領域を消費せずにデータを保持している重複排除ブロックを、削減済み重複排除ブロックと呼ぶことがある。

【００５８】

続いてマッピングテーブル１２７について、図６を参照しながら説明する。マッピングテーブル１２７は、重複排除ブロックとデータブロックのマッピング状態を管理するためのテーブルで、各行（レコード）には、重複排除ブロックの識別子や、重複排除ブロックに割り当てられているデータブロックの識別子等の情報が格納される。以下、マッピングテーブル１２７の各カラムに格納される情報について説明する。

【００５９】

仮想ボリューム１２７１、論理ページ１２７２にはそれぞれ、管理対象の重複排除ブロックの属する仮想ボリュームの識別子、論理ページの識別子が格納される。そして重複排除ブロック１２７３には、管理対象の重複排除ブロックの識別子が格納される。フィンガープリント１２７４には、重複排除処理において算出した、重複排除ブロックごとのフィンガープリントが格納される。

【００６０】

物理ページ１２７５とデータブロック１２７６にはそれぞれ、管理対象の重複排除ブロックに割り当てられているデータブロックの属する物理ページの識別子、データブロックの識別子が格納される。

【００６１】

なお、重複排除ブロックまたはデータブロックの識別子には、重複排除ブロックまたはデータブロックを一意に識別可能な情報であれば、任意の情報が用いられて良い。本実施例では、重複排除ブロックの識別子には、重複排除ブロックの属する仮想ボリューム内のアドレスが用いられる。一方、データブロックの識別子には、データブロックの属する物理ページ内の相対アドレス（物理ページの先頭のデータブロックの識別子を０とするアドレス）が用いられる。ただし図６では、図４に記載の重複排除ブロックとデータブロックのマッピング関係を説明するために、重複排除ブロックの識別子として、図４に記載の重複排除論理ブロックに付されたアルファベットを用い、またデータブロックの識別子としては、データブロックに付されたアルファベットを用いている。

【００６２】

削減フラグ１２７７は、管理対象の重複排除ブロックが削減済み重複排除ブロックか否かを示す情報である。管理対象の重複排除ブロックが削減済み重複排除ブロックである場合には、削減フラグ１２７７にはＴＲＵＥが格納され、そうでなければＦＡＬＳＥが格納される。

【００６３】

物理ページについての情報は、プール管理テーブル１２８に格納されて管理される。図７を参照しながらプール管理テーブル１２８の内容を説明する。

【００６４】

プール管理テーブル１２８の各レコードには、物理ページの状態等の情報が格納される。物理ページ１２８１には、管理対象の物理ページの識別子が格納され、論理ページ１２８３には、物理ページが割り当てられている論理ページの識別子が格納される。物理ページが論理ページ（通常論理ページ）に割り当てられていない場合、あるいは物理ページ内

の領域が重複排除ブロックに割り当てられている場合には、論理ページ１２８３にはＮＵＬＬが格納される。

【００６５】

使用状況１２８２には、物理ページの使用状態が格納される。物理ページの状態には、通常論理ページに割り当てられている状態、重複排除ブロックに割り当てられるために使用されている状態、そして未使用状態がありえる。物理ページが通常論理ページに割り当てられている場合、使用状況１２８２には“論理ページ”が格納される。物理ページ内の領域が重複排除ブロックに割り当てられている場合には、使用状況１２８２には“データブロック”が格納される。物理ページが論理ページにも割り当てられておらず、かつ物理ページ内の領域が重複排除ブロックに割り当てられてもいない場合、その物理ページの状態は“未使用状態”と呼ばれ、その場合には使用状況１２８２には“未使用”が格納される。

【００６６】

ストレージ装置１がたとえば、ある論理ページに物理ページを割り当てる際には、プール管理テーブル１２８のレコードのうち、使用状況１２８２が“未使用”のレコードを１つ特定し、特定されたレコードで管理される物理ページを論理ページに割り当てると決定する。またストレージ装置１は論理ページ管理テーブル１２６に、割り当てることを決定した物理ページの識別子（物理ページ１２８１）を格納する。

【００６７】

なお、物理ページは、実際には１または複数の記憶デバイス１５上の領域である。そのためストレージ装置１は、プール管理テーブル１２８に加えて、物理ページと、物理ページが存在する記憶デバイス１５及び記憶デバイス上のアドレスとのマッピングを管理するための情報も保持する。ストレージ装置１は、たとえば計算機５からのアクセス要求を受け付けると、アクセス要求で指定されている領域に対応する論理ページを特定し、その後論理ページに割り当てられている物理ページを特定する。さらにその後ストレージ装置１は、このマッピング情報を参照することで、アクセス対象の記憶デバイス１５上アドレスを特定する。ただし、このマッピング情報及びマッピング情報を用いたアクセス先記憶デバイス１５のアドレス特定方法は、Ｔｈｉｎ　Ｐｒｏｖｉｓｉｏｎｉｎｇ技術を用いるストレージ装置において公知のものであるため、本実施例では詳細説明を略す。

【００６８】

続いて検索テーブル１２９について説明する。検索テーブル１２９は重複排除処理で用いられる。重複排除処理では、重複判定対象のデータと同じデータが既にストレージプール１７にあるか判定する処理が行われるが、この判定処理の高速化のために、ストレージ装置１は重複排除ブロックごとにフィンガープリントを算出し、算出したフィンガープリントを検索テーブル１２９に格納する。

【００６９】

図８に検索テーブル１２９の例を示す。検索テーブル１２９のカラムのうち、フィンガープリント１２９１にはフィンガープリントが格納される。そして重複排除ブロック１２９２には、フィンガープリント１２９１に格納されたフィンガープリントを持つ重複排除ブロックの位置情報が格納される。位置情報としては、重複排除ブロックの属する仮想ボリュームの識別子と、重複排除ブロックの識別子（仮想ボリューム内アドレス）の組からなる値が格納される。検索テーブル１２９の各レコードは、フィンガープリント１２９１の値の小さい順にソートされて格納される。

【００７０】

あるレコードの重複排除ブロック１２９２には、複数の位置情報が格納されることもある（重複データが存在するケース）。また、フィンガープリント１２９１の値が同じレコードが複数存在することもありえる。内容の異なる複数のデータについてＦＰを算出した場合でも、それぞれのＦＰが同じになることがあり得るからである。

【００７１】

ストレージ装置１があるデータについて重複判定を行う場合、データのＦＰを算出し、

算出されたＦＰと同じものが、検索テーブル１２９のカラム“フィンガープリント１２９１”に格納されているレコードがあるか判定する。そのようなレコードがある場合、ストレージ装置１はさらに、そのレコードのカラム“重複排除ブロック１２９２”を参照して重複排除ブロック（の位置情報）を特定する。さらにストレージ装置１はマッピングテーブル１２７を参照して、この重複排除ブロックに割り当てられているデータブロックを特定し、当該データブロックからデータを読み出して、読み出されたデータと判定対象のデータとをバイト単位で比較することで、両者が一致しているかを判定する。

【００７２】

図９に、逆参照テーブル１３０の例を示す。逆参照テーブル１３０は、データブロックと重複排除ブロックのマッピング情報を管理するテーブルである。マッピングテーブル１２７とは異なり、逆参照テーブル１３０は、管理対象のデータブロックのアドレスから、そのデータブロックが割り当てられている重複排除ブロックのアドレスを特定するために用いられる。

【００７３】

逆参照テーブル１３０のカラムのうち、物理ページ１３０１とデータブロック１３０２にはそれぞれ、管理対象のデータブロックが属する物理ページの識別子とデータブロックの識別子が格納される。データブロックの識別子には先に述べたとおり、データブロックの属する物理ページ内の相対アドレスが用いられる。重複排除ブロック１３０３には、管理対象のデータブロックが割り当てられている重複排除ブロックの位置情報が格納される。重複排除ブロックの位置情報は、その重複排除ブロックの属する仮想ボリュームの識別子と重複排除ブロックの識別子の組からなる値である。

【００７４】

図１０に、追記ポインタ１３１の例を示す。追記ポインタ１３１は、重複排除ブロックへのデータブロックの割り当てが必要な場合に、割り当てるべきデータブロックのアドレスを管理するための情報である。追記ポインタ１３１は、物理ページの識別子と、データブロックの識別子の組からなる。ストレージ装置１は追記ポインタ１３１で指し示されているデータブロックを重複排除ブロックに割り当て、割り当てが終わった後には、追記ポインタ１３１の内容を、現在追記ポインタ１３１に格納されているアドレスの次のアドレスに更新する。物理ページの終端のデータブロックが重複排除ブロックに割り当てられた場合には、ストレージ装置１は未使用状態の物理ページを新たにデータブロック格納用に確保し、その物理ページの識別子と、その物理ページの先頭のデータブロックの識別子から成る情報で、追記ポインタ１３１を更新する。

【００７５】

（４）処理の流れ

続いて、ストレージ装置１で行われる各種処理の流れを説明する。ストレージ装置１のシステムメモリ１２には少なくとも、Ｉ／Ｏ処理部１２１、重複排除処理部１２３、重複排除解除部１２４、論理ページ変更部１２５、物理ページ解放処理部１３２の、５種類のプログラムが格納されている。なお、システムメモリ１２に格納されているこれらのプログラムのことを、「ストレージ制御プログラム」と呼ぶこともある。以下ではこれらのプログラムによって行われる処理の説明を行う。なお、本実施例に係る図面において、参照番号の前に付されている文字列“ＳＰ”は「ステップ」を意味する。

【００７６】

まずＩ／Ｏ処理部１２１により行われる主な処理の説明を行う。Ｉ／Ｏ処理部１２１は、計算機５等のイニシエータデバイスに仮想ボリュームを提供し、イニシエータデバイスから受け付けたＩ／Ｏ要求（リード要求やライト要求）の処理を行うプログラムである。

【００７７】

図１６を参照しながらリード処理の流れを説明する。ストレージ装置１が計算機５から仮想ボリュームに対するリード要求を受領すると、ストレージ装置１はリード要求で指定されたデータを計算機５に返送する。この処理をリード処理と呼ぶ。

【００７８】

ステップ７０２：ストレージ装置１が計算機５からリード要求を受領すると、Ｉ／Ｏ処理部１２１はリード要求に含まれるリード先アドレス（ＬＢＡ）から論理ページ番号を算出することで、リード先アドレスの含まれる論理ページを特定する。また同時にＩ／Ｏ処理部１２１は、リード先アドレスに対応する論理ページ内アドレスも算出する。以下では、ここで特定された論理ページを「アクセス対象論理ページ」と呼ぶ。

【００７９】

ステップ７０３：続いてＩ／Ｏ処理部１２１は論理ページ管理テーブル１２６を参照することで、アクセス対象論理ページが重複排除論理ページか判定する。この判定ではＩ／Ｏ処理部１２１は、論理ページ管理テーブル１２のレコードのうち、アクセス対象論理ページに対応するレコードの状態１２６５が“Ｑ”か否かを判定する。状態１２６５が“Ｑ”の場合（ＳＰ７０３：Ｙ）、次にステップ７０４が行われる。それ以外の場合には、Ｉ／Ｏ処理部１２１は次にステップ７０５を行う。

【００８０】

ステップ７０４：Ｉ／Ｏ処理部１２１はマッピングテーブル１２７を参照することで、リード先アドレスで指定されている領域に割り当てられているデータブロックの位置情報（物理ページ１２７５とデータブロック１２７６の組）を特定する。そしてこの情報を用いて、リード対象データの格納されている記憶デバイス１５のアドレスを求める。

【００８１】

ステップ７０５：Ｉ／Ｏ処理部１２１は論理ページ管理テーブル１２６を参照することで、アクセス対象論理ページに割り当てられている物理ページ（の物理ページ番号）を特定する。さらに特定された物理ページ番号及びステップ７０２で算出した論理ページ内アドレスをもとに、リード対象データの格納されている記憶デバイス１５のアドレスを求める。

【００８２】

ステップ７０６：Ｉ／Ｏ処理部１２１は、ステップ７０４またはステップ７０５で求められた記憶デバイス１５のアドレスからデータを読み出し、読み出したデータをキャッシュメモリ１４に格納する。またＩ／Ｏ処理部１２１は、キャッシュメモリ１４に格納されたデータを計算機５に返送し、処理を終了する。

【００８３】

続いてストレージ装置１が、計算機５から仮想ボリュームに対するライト要求及びライトデータを受領した時の処理（ライト処理）の流れを、図１５を用いて説明する。この処理もＩ／Ｏ処理部１２１が実行する。なお、計算機５が発行するライト要求には、ライトデータの書き込み先位置の情報（ＬＢＡ及びデータ長）が含まれている。

【００８４】

ステップ６０２：Ｉ／Ｏ処理部１２１はライト要求に含まれる書き込み先位置の情報から、ライトデータの書き込み先となる論理ページの識別子（論理ページ番号）、論理ページ内アドレスを算出する。またＩ／Ｏ処理部１２１は論理ページ管理テーブル１２６を参照し、ライトデータの書き込み先となる論理ページの状態１２６５がＰでかつ物理ページ１２６４がＮＵＬＬの場合には、この論理ページに物理ページを割り当てる処理を行う。具体的にはＩ／Ｏ処理部１２１はプール管理テーブル１２８を参照することで、使用状況１２８２が“未使用”のレコードを選択し、このレコードの論理ページ１２８３に、ライトデータの書き込み先となる論理ページの識別子を格納し、また使用状況１２８２を“論理ページ”に変更する。さらにここで選択されたレコードの物理ページ１２８１の値を、論理ページ管理テーブル１２６のレコードのうち、ライトデータの書き込み先の論理ページのレコードの物理ページ１２６４に格納する。

【００８５】

ステップ６０３：Ｉ／Ｏ処理部１２１はライトデータをキャッシュメモリ１４に格納する。この時、Ｉ／Ｏ処理部１２１はライトデータに、書き込み先論理ページの識別子と論理ページ内アドレスを付加してキャッシュメモリ１４に格納する。

【００８６】

ステップ６０４：Ｉ／Ｏ処理部１２１は、論理ページ管理テーブル１２６のレコードのうち、書き込み先論理ページに対応するレコードの最終ライト時刻１２６６に、現在時刻（ステップ６０４実行時点の時刻）を格納する。

【００８７】

ステップ６０５：Ｉ／Ｏ処理部１２１は、論理ページ管理テーブル１２６のレコードの状態１２６５を参照し、書き込み先論理ページが状態“Ｐ”であるか否かを判定する。書き込み先論理ページが状態“Ｐ”であった場合（ＳＰ６０５：Ｙ）、Ｉ／Ｏ処理部１２１は書き込み先論理ページの状態１２６５を“Ｘ”へ変更し（ステップ６０７）、次にステップ６０９を行う。

【００８８】

ステップ６０６：Ｉ／Ｏ処理部１２１は、論理ページ管理テーブル１２６のレコードの状態１２６５を参照し、書き込み先論理ページが重複排除論理ページかを判定する。状態１２６５が“Ｑ”だった場合、書き込み先論理ページが重複排除論理ページであったことを意味する。状態１２６５が“Ｑ”だった場合（ＳＰ６０６：Ｙ）、Ｉ／Ｏ処理部１２１は次にステップ６０８を行う。状態１２６５が“Ｑ”でない場合（ＳＰ６０６：Ｎ）、ステップ６０８はスキップされる。

【００８９】

ステップ６０８：Ｉ／Ｏ処理部１２１は重複排除解除部１２４を呼び出すことで、書き込み先論理ページの重複排除解除処理を行う。重複排除解除処理は重複排除解除部１２４が実行する。重複排除解除部１２４の処理は後述する。重複排除解除処理が行われることで、書き込み先論理ページは通常論理ページになる（状態１２６５がＸになる）。

【００９０】

ステップ６０９：Ｉ／Ｏ処理部１２１は、ステップ６０３でキャッシュメモリ１４に格納したデータを記憶デバイス１５に格納する。キャッシュメモリ１４に一時的に格納されたデータを記憶デバイス１５に格納する処理のことを「デステージ処理」と呼ぶ。Ｉ／Ｏ処理部１２１は、図１６のステップ７０５と同様に、論理ページ管理テーブル１２６を参照することで、書き込み先論理ページに割り当てられている物理ページを特定し、さらに特定された物理ページの存在する記憶デバイス１５のアドレスを特定する。そしてＩ／Ｏ処理部１２１は、この特定された記憶デバイス１５のアドレスに対して、キャッシュメモリ１４に格納されていたデータを書き込む。デステージ処理の後、Ｉ／Ｏ処理部１２１は計算機５にライト処理が完了した旨を通知し、ライト処理を終了する。

【００９１】

なお、上で説明した例では、ストレージ装置１はライトスルー処理を行っている。つまりデステージ処理（ステップ６０９）の後で、計算機５にライト処理が完了した旨が通知される。ただし別の実施形態として、キャッシュメモリ１４がライトバックキャッシュとして用いられてもよい。その場合、Ｉ／Ｏ処理部１２１は、キャッシュメモリ１４に計算機５から受領したデータを格納した時点（ステップ６０３）で、計算機５にライト処理が完了した旨を通知してもよい。またこの場合、ステップ６０５以降の処理は、かならずしもキャッシュメモリ１４にデータが格納された直後に行われなくてもよく、任意のタイミングで行われるようにしてもよい。

【００９２】

続いて、論理ページ変更部１２５が行う処理の流れを、図１４を参照しながら説明する。本実施例ではこの処理を「移行処理」と呼ぶ。移行処理は仮想ボリュームの各論理ページのうち、所定の条件を満足する通常論理ページの状態を変更する処理である。移行処理は各仮想ボリュームについて定期的に実行される。以下ではある特定の仮想ボリュームについて、ストレージ装置１が移行処理を行う時の処理の流れを説明する。

【００９３】

ステップ４０２：論理ページ変更部１２５は、仮想ボリュームの論理ページのうち、まだステップ４０３以降の処理が行われていない論理ページを１つ選択する。ここでの論理ページの選択方法は任意であるが、たとえば論理ページ変更部１２５は、仮想ボリューム

の先頭の論理ページから順に、ステップ４０３以降の処理を行うとよい。

【００９４】

ステップ４０３：論理ページ変更部１２５は論理ページ管理テーブル１２６を参照することで、選択された論理ページの状態１２６５が“Ｘ”か否かを判定する。状態１２６５が“Ｘ”の場合（ステップ４０３：Ｙ）、次にステップ４０４が行われる。状態１２６５が“Ｘ”でない場合（ステップ４０３：Ｎ）、ステップ４０４～ステップ４０７はスキップされる。

【００９５】

ステップ４０４：論理ページ変更部１２５は選択された論理ページの最終ライト時刻１２６６を参照し、現在時刻と最終ライト時刻１２６６との差を算出し、この差が所定値以上であるか判定する。この差が所定値以上の場合（ステップ４０４：Ｙ）、選択された論理ページには、一定期間以上の間、計算機５からの書き込みが発生していないことを意味する。この場合には、論理ページ変更部１２５は次にステップ４０５を行う。現在時刻と最終ライト時刻１２６６との差が所定値未満の場合、ステップ４０５～ステップ４０７はスキップされる。

【００９６】

ステップ４０５：論理ページ変更部１２５は重複排除処理部１２３を呼び出して、選択された論理ページに対して重複排除処理を行う。この処理は、通常論理ページの状態を、状態Ｑに変更する処理（つまり重複排除論理ページに変更する処理）である。詳細は後述する。

【００９７】

ステップ４０６：論理ページ変更部１２５は、選択された論理ページを重複排除論理ページとして維持しておくべきか否かを判定する。具体的には論理ページ変更部１２５はまず、論理ページ管理テーブル１２６内の各レコードのうち、重複排除１２６３が“有効”となっているレコードの数を計数することで、ストレージ装置１内の重複排除論理ページの数を特定し、この数が閾値以上であるか否かを判定する。この数が閾値未満の場合には、論理ページ変更部１２５は選択された論理ページは重複排除論理ページとして維持しておくべきと判断する。

【００９８】

ストレージ装置１内の重複排除論理ページの数が少ない（閾値未満の）場合、重複判定対象となる重複排除ブロック数が少ない。重複判定対象となる重複排除ブロック数が少ない場合、重複排除処理による記憶領域消費量の削減効果が出ないため、ストレージ装置１は、閾値以上の数の重複排除論理ページが存在するようになるまでは、選択された論理ページを重複排除論理ページとして維持する。

【００９９】

なお重複排除論理ページの数を特定する方法は、上で挙げた方法に限定されない。別の実施形態として、ストレージ装置１は重複排除論理ページの数を特定するために、システムメモリ１２に重複排除論理ページの数を記録する領域を予め設けておき、通常論理ページを重複排除論理ページへ、またはその逆へ変更する処理を行う度に、システムメモリ１２に記録された重複排除論理ページの数を増減させるようにしてもよい。

【０１００】

重複排除論理ページの数が閾値以上である場合、論理ページ変更部１２５はさらに、選択された論理ページの排除ブロック数１２６７を参照することで重複排除率を算出し、この値が所定の閾値以上か判定する。重複排除率は、“排除ブロック数１２６７÷論理ページ内重複排除ブロック数”で求められる。

【０１０１】

重複排除率が所定の閾値以上であれば、選択された論理ページは重複排除処理によって記憶領域消費量が少なくなっていることを意味するので、論理ページ変更部１２５は、選択された論理ページを重複排除論理ページとして維持しておくべきと判断する。

【０１０２】

一方重複排除率が閾値未満の場合には、選択された論理ページの記憶領域消費量の削減効果はあまり大きくないため、重複排除論理ページとして維持しておく必要性は低い。そのため論理ページ変更部１２５は、選択された論理ページを重複排除論理ページとして維持しないと判断する。選択された論理ページが、重複排除論理ページとして維持しておく必要がないと判断された場合（ステップ４０６：Ｎ）、次にステップ４０７が行われる。重複排除論理ページとして維持しておく必要があると判断された場合（ステップ４０６：Ｙ）、ステップ４０７はスキップされる。

【０１０３】

ステップ４０７：論理ページ変更部１２５は重複排除解除部１２４を呼び出して、選択された論理ページに対して重複排除解除処理を行う。この処理は、重複排除論理ページを、状態Ｐの論理ページに変更する処理である。この処理の詳細は後述する。

【０１０４】

ステップ４０８：論理ページ変更部１２５は、仮想ボリュームの論理ページのうち、まだステップ４０３～ステップ４０７の処理が行われていない論理ページがあるか判定する。まだステップ４０３～ステップ４０７の処理が行われていない論理ページがある場合（ステップ４０８：Ｙ）、論理ページ変更部１２５は再びステップ４０２を実行する。仮想ボリューム内の全論理ページに対してステップ４０３～ステップ４０７の処理が行われた場合には（ステップ４０８：Ｎ）、論理ページ変更部１２５は処理を終了する。

【０１０５】

なお、上で説明した例では、ステップ４０６で論理ページ変更部１２５は重複排除率を算出し、重複排除率に基づいて選択された論理ページを重複排除論理ページとして維持しておくか否かを判定している。ただし論理ページ変更部１２５は、重複排除率を算出する代わりに、排除ブロック数１２６７が所定の閾値以上か否かを判定することで、選択された論理ページを重複排除論理ページとして維持しておくか否か決定してもよい。排除ブロック数１２６７が大きければ、実質的に重複排除率も大きく、また排除ブロック数１２６７が小さければ、実質的に重複排除率も小さいことは明らかだからである。また、重複排除率や排除ブロック数１２６７以外にも、論理ページの記憶領域消費量の削減効果を推し量ることができるその他の指標値が、ステップ４０６の判定に用いられてもよい。

【０１０６】

続いて、ステップ４０５で行われる重複排除処理の流れを、図１１、図１２を参照しながら説明する。図１１、図１２は重複排除処理部１２３が実行する処理のフローチャートである。重複排除処理部１２３は、論理ページ変更部１２５から呼び出されることにより（図１４　ステップ４０５）、処理を開始する。この時論理ページ変更部１２５は重複排除処理部１２３に、処理対象の論理ページの情報（具体的には論理ページの属する仮想ボリュームの識別子、論理ページ識別子）を通知する。重複排除処理部１２３は、この通知された論理ページの状態を変更する処理を行う。以下では論理ページ変更部１２５から通知された論理ページのことを、「指定された論理ページ」と呼ぶ。

【０１０７】

ステップ１０２：重複排除処理部１２３は論理ページ管理テーブル１２６を参照することで、指定された論理ページに割り当てられている物理ページを特定し、この物理ページに格納されているデータを読み出してキャッシュメモリ１４に格納する。

【０１０８】

ステップ１０３：重複排除処理部１２３は重複排除処理を行う。ここでは、指定された論理ページ内の重複排除ブロックごとに重複排除処理を行う。以下、図１２を参照しながら説明する。

【０１０９】

まずステップ２０１で重複排除処理部１２３は、変数ｋを用意し、ｋが指定された論理ページの先頭の重複排除ブロックを指し示すようにする（具体的にはｋに、指定された論理ページの先頭の重複排除ブロックの識別子を代入する）。以下では、変数ｋによって指し示される重複排除ブロックのことを「選択された重複排除ブロック」と呼ぶ。

【０１１０】
続いて重複排除処理部１２３は、ステップ１０２においてキャッシュメモリ１４に格納されたデータのうち、選択された重複排除ブロックのデータを選択（特定）する（ステップ２０２）。続いて重複排除処理部１２３は、特定されたデータを読み出して、このデータのフィンガープリントを算出する（ステップ２０３）。
【０１１１】
次にステップ２０４で、重複排除処理部１２３は検索テーブル１２９を参照し、ステップ２０３で算出されたフィンガープリントと同じ値が、フィンガープリント１２９１に格納されているレコードを検索する。レコードがあった場合（ステップ２０５：Ｙ）、ステップ２０７が実行される。レコードがなかった場合（ステップ２０５：Ｎ）、ステップ２０６が実行される。
【０１１２】
ステップ２０６で重複排除処理部１２３は、ステップ２０３で算出されたフィンガープリントと、選択された重複排除ブロックの位置情報とで構成されるレコードを作成して、これを検索テーブル１２９に追加する。次いでステップ２０８が実行される。
【０１１３】
ステップ２０７で、重複排除処理部１２３は、ステップ２０４の検索の結果得られたレコードのカラム“重複排除ブロック１２９２”で特定される重複排除ブロックからデータを読み出す。より詳細には、重複排除処理部１２３は、特定された重複排除ブロックに割り当てられているデータブロックを、マッピングテーブル１２７の物理ページ１２７５とデータブロック１２７６を参照して特定し、特定されたデータブロックからデータを読み出す。そしてこの読み出されたデータと、ステップ２０２で選択されたデータとを、バイト単位で比較し、両者が一致するか判定する。両者が一致する場合（ステップ２０７：Ｙ）、次にステップ２１０が行われる。
【０１１４】
もし、ステップ２０４の検索により得られた検索テーブル１２９のレコードの重複排除ブロック１２９２に、複数の重複排除ブロックの識別子が格納されている場合は、重複排除処理部１２３は、それらの重複排除ブロックについても同様のデータの比較を行う。また、前記検索の結果、検索テーブル１２９のレコードが複数得られた場合は、重複排除処理部１２３は、各レコードにつき同様にデータの比較を行う。比較の結果、いずれの重複排除ブロックも、ステップ２０２で選択されたデータと一致しない場合には（ステップ２０７：Ｎ）、次にステップ２０６が行われる。
【０１１５】
ステップ２０８では、重複排除処理部１２３はステップ２０２で選択されたデータを書き込むデータブロックを確保する。具体的には、重複排除処理部１２３は、追記ポインタ１３１に記録されているデータブロックのアドレスをデータ格納先として選択し、追記ポインタ１３１にデータブロックのサイズを加算する。もし、選択した追記ポインタ１３１が物理ページの最後のアドレスであった場合には、重複排除処理部１２３は、プール管理テーブル１２８を参照して未使用状態の物理ページを選択し、この物理ページをデータブロック格納用の物理ページへ変更し、追記ポインタ１３１に選択した物理ページの先頭アドレスを格納する。
【０１１６】
次いでステップ２０９で、重複排除処理部１２３は確保したデータブロックを有する物理ページ（記憶デバイス１５）に、ステップ２０２で選択されたデータを書き込む。さらに、重複排除処理部１２３は、逆参照テーブル１３０に、ステップ２０８で確保したデータブロックが割り当てられる重複排除ブロックの情報を記録する。具体的には、データブロックに対応する逆参照テーブル１３０のレコードの重複排除ブロック１３０３に、処理対象の重複排除ブロックの識別子を格納する。
【０１１７】
ステップ２１０は、ステップ２０２で選択されたデータの重複データが、すでに他の重

複排除ブロックに割り当てられているデータブロックに存在する場合に実行される処理である。ステップ２１０では重複排除処理部１２３は、指定された論理ページの排除ブロック数１２６７に１を加算する。なお、この場合、ステップ２０８～ステップ２０９と異なり、ステップ２０２で選択されたデータは記憶デバイス１５に書き込まれない。

【０１１８】

ステップ２１１で重複排除処理部１２３は、ステップ２０７の結果一致するデータが見つかった検索テーブル１２９のレコードのカラム“重複排除ブロック１２９２”に、選択された重複排除ブロックの識別子を追加する。

【０１１９】

ステップ２１２で重複排除処理部１２３は、マッピングテーブル１２７のレコードのうち、選択された重複排除ブロックについてのレコードの更新を行う。もしステップ２０７の判定が肯定的な場合、つまりステップ２０２で選択されたデータの重複データが格納されたデータブロックがあった場合、そのデータブロックの物理ページ識別子およびデータブロックアドレスをマッピングテーブル１２７のレコードに登録する。この時重複排除処理部１２３は、このレコードのカラム“削減フラグ１２７７”には、重複排除ブロックが削減済み重複排除ブロックであることを示す値である“ＴＲＵＥ”を格納する。

【０１２０】

一方、ステップ２０２で選択されたデータの重複データが格納されたデータブロックがなかった場合（つまりステップ２０５またはステップ２０７の判定が否定的な場合）、ステップ２０６でデータを書き込んだ先のデータブロックの物理ページ識別子およびデータブロックの識別子をマッピングテーブル１２７のレコードに登録する。この登録されるレコードのカラム“削減フラグ１２７７”には、“ＦＡＬＳＥ”が格納される。

【０１２１】

また、ステップ２０７の判定が肯定的な場合も否定的な場合も、レコードのカラム“フィンガープリント１２７４”には、ステップ２０３で算出したフィンガープリント値が格納される。

【０１２２】

変数ｋが論理ページ内の終端アドレスに等しい場合、つまりすべての重複排除ブロックに対して、ステップ２１２までの処理が完了した場合（ステップ２１３：Ｙ）、この処理（ステップ１０３）は終了する。まだステップ２１２までの処理が完了していない重複排除ブロックが残っている場合（ステップ２１３：Ｎ）、重複排除処理部１２３は変数ｋが次の重複排除ブロックを指し示すように、変数ｋの更新を行い（変数ｋに重複排除ブロックのサイズを加算する）、再びステップ２０２から処理を行う。以上が、ステップ１０３で行われる処理の内容である。

【０１２３】

図１１の説明に戻る。ステップ１０４、ステップ１０５で、重複排除処理部１２３は指定された論理ページに割り当てられている物理ページの割り当てを解除する。

【０１２４】

ステップ１０４：重複排除処理部１２３はプール管理テーブル１２８のレコードのうち、割り当てられている物理ページのレコードの使用状況１２８２を“未使用”に変更し、論理ページ１２８３をＮＵＬＬに変更する。

【０１２５】

ステップ１０５：重複排除処理部１２３は論理ページ管理テーブル１２６のレコードのうち、指定された論理ページのレコードの内容を変更する。具体的にはこのレコードの物理ページ１２６４と最終ライト時刻１２６６がＮＵＬＬに変更され、重複排除１２６３には“無効”が、状態１２６５には“Ｑ”が格納される。ステップ１０５までの処理が完了すると、重複排除処理部１２３は処理を終了する。

【０１２６】

続いて、ステップ４０７またはステップ６０８で行われる重複排除解除処理の流れを、図１３を参照しながら説明する。図１３は重複排除解除部１２４が実行する処理のフロー

チャートである。重複排除解除部１２４は、論理ページ変更部１２５またはＩ／Ｏ処理部１２１から呼び出されることにより、処理を開始する。

【０１２７】

この時論理ページ変更部１２５またはＩ／Ｏ処理部１２１は、重複排除解除部１２４に処理対象の論理ページの情報（具体的には論理ページの属する仮想ボリュームの識別子、論理ページ識別子）を通知する。重複排除解除部１２４は、この通知された論理ページの状態を変更する処理を行う。以下では、重複排除解除部１２４に通知された情報により特定される論理ページのことを、「指定された論理ページ」と呼ぶ。

【０１２８】

ステップ３０２：重複排除解除部１２４はマッピングテーブル１２７を参照することで、指定された論理ページの各重複排除ブロックに割り当てられているデータブロックを特定し、特定された全てのデータブロックに格納されているデータを読み出してキャッシュメモリ１４に格納する。

【０１２９】

ステップ３０３：重複排除解除部１２４は指定された論理ページに、未使用の物理ページを割り当てる。物理ページを割り当てる処理は、ステップ６０２で説明したものと同様である。

【０１３０】

ステップ３０４：重複排除解除部１２４は指定された論理ページに対応する、論理ページ管理テーブル１２６のレコードの内容を更新する。具体的には以下の内容更新が行われる。まずこのレコードの重複排除１２６３は“無効”に変更され、排除ブロック数１２６７は０に更新される。またステップ３０３が実行された時点で、物理ページ１２６３には割り当てられた物理ページの識別子が格納されているので、ここでは物理ページ１２６３の更新は行われない。

【０１３１】

ステップ３０５：重複排除解除部１２４は、ステップ３０２でキャッシュメモリ１４に格納したデータを、ステップ３０３で割り当てられた物理ページにデステージする。

【０１３２】

ステップ３０６：重複排除解除部１２４はマッピングテーブル１２７のレコードのうち、指定された論理ページに含まれる各重複排除ブロックについての情報を管理しているレコードの内容を更新する。ここでは、各レコードの物理ページ１２７５およびデータブロック１２７６は、ＮＵＬＬに変更される。なお、この時点ではフィンガープリント１２７４および削減フラグ１２７７は更新されず、元の値を保持しつづける。

【０１３３】

ステップ３０７：重複排除処理部１２４は、指定された論理ページの各重複排除ブロックに割り当てられていたデータブロックを有する物理ページを特定する。具体的には、重複排除処理部１２４は、マッピングテーブル１２７のレコードのうち、指定された論理ページの識別子と“論理ページ１２７２”のカラムの情報とが一致し、かつ削減フラグ１２７７がＦＡＬＳＥであるレコードを検索し、検索の結果得られたレコードから物理ページ１２７５の情報を重複なく取り出す。ステップ３０７においては、１個以上の物理ページが特定される場合がある。またここでは、削減フラグ１２７７がＦＡＬＳＥであるレコードだけが検索されるので、削減済み重複排除ブロックは検索結果から除外される。

【０１３４】

ステップ３０８：重複排除処理部１２４は、ステップ３０７で特定した物理ページを未割当状態に変更する。本実施例では、データブロック格納用に使用されている物理ページを未割当状態に変更する処理を、「物理ページ解放処理」と呼ぶ。また、物理ページを未割当状態にすることを、「物理ページを解放する」と表現する場合がある。

【０１３５】

ステップ３０８の実行前時点において、解放する対象の物理ページには、依然として複数の重複排除ブロックに共有されているデータブロックがある場合がある。このようなデ

ータブロックを、以降では共有データブロックと呼ぶ。また、共有データブロックのデータを共有データと呼ぶ。

【０１３６】

物理ページ解放処理は、物理ページを未割当状態へ変更する前に、その物理ページの共有データを、別の物理ページのデータブロックへ移動（複製）する。そして、その共有データブロックが割り当てられている重複排除ブロックを、複製先のデータブロックが割り当てられた状態へ変更する。その後、対象の物理ページを未割当状態に変更する。

【０１３７】

物理ページ解放処理は、物理ページ解放処理部１３２により実行される。ステップ３０８において、重複排除処理部１２４は、ステップ３０７で特定した物理ページの識別子のうち、追記ポインタ１３１に記録されている物理ページを除いたものを、物理ページ解放処理部１３２に通知する。物理ページ解放処理部１３２は、通知された物理ページを対象に、物理ページ解放処理を実行する。

【０１３８】

物理ページ解放処理部１３２における処理の流れを、図１７を参照しながら説明する。図１７は、物理ページ解放処理部１３２が実行する処理のフローチャートである。

【０１３９】

ステップ８０２：物理ページ解放処理部１３２は、通知された物理ページの識別子の内のひとつを選択する。この選択した物理ページの識別子により特定される物理ページを、処理対象の物理ページと呼ぶ。

【０１４０】

ステップ８０３：物理ページ解放処理部１３２は、処理対象の物理ページに含まれるデータブロックを１つ選択する。以下では、ステップ８０３で選択したデータブロックを処理対象のデータブロックと呼ぶ。

【０１４１】

ステップ８０４：物理ページ解放処理部１３２は、処理対象のデータブロックに対応している逆参照テーブル１３０のレコードを参照し、重複排除ブロック１３０３に格納された重複排除ブロックの識別子より重複排除ブロックを特定する。ここで特定された重複排除ブロックは、ステップ２０９（重複排除処理）により処理対象のデータブロックが割り当てられた重複排除ブロックである。

【０１４２】

ステップ８０６：物理ページ解放処理部１３２は、ステップ８０４で特定された重複排除ブロックに対応しているマッピングテーブル１２７のレコードを参照し、フィンガープリント１２７４の値を得る。

【０１４３】

ステップ８０７：物理ページ解放処理部１３２は、処理対象のデータブロックが共有データブロックである場合には、そのデータを別の物理ページのデータブロックへ移動（複製）し、処理対象のデータブロックを共有している重複排除ブロックが複製先のデータブロックを共有するように変更する。ステップ８０７の詳細については後述する。

【０１４４】

ステップ８０８：物理ページ解放処理部１３２は、処理対象の物理ページに含まれる全てのデータブロックについて、ステップ８０４～８０７を実行したどうかを判定する。ステップ８０４～８０７を実行していないデータブロックがある場合（ステップ８０８：Ｙ）、物理ページ解放処理部１３２は、処理対象のデータブロックの次のデータブロックを選択し（ステップ８０３）、ステップ８０４～８０７を実行する。処理対象の物理ページに含まれる全てのデータブロックについて、ステップ８０４～８０７を実行済みであれば（ステップ８０８：Ｎ）、ステップ８０９が実行される。

【０１４５】

ステップ８０９：物理ページ解放処理部１３２は、処理対象の物理ページに対応するプール管理テーブル１２８のレコードの使用状況１２８２を、“未使用”に変更する。

【０１４６】
ステップ８１０：物理ページ解放処理部１３２は、処理対象の物理ページに含まれる全てのデータブロックについて、それぞれ逆参照テーブル１３０の対応するレコードの重複排除ブロック１３０３にＮＵＬＬを格納する。
【０１４７】
ステップ８１１：物理ページ解放処理部１３２は、重複排除解除部１２４から通知された全ての物理ページについて、ステップ８０３～８１１を実行したかどうかを判定する。ステップ８０３～８１１を実行していない物理ページがある場合（ステップ８１１：Ｙ）、物理ページ解放処理部１３２はそのような物理ページの一つを選択し（ステップ８０２）、ステップ８０３からのステップを実行する。通知された物理ページの全てについて、ステップ８０３～８１１を実行済みであれば（ステップ８１１：Ｎ）、物理ページ解放処理を終了する。
【０１４８】
図１８を参照しながら、ステップ８０７の共有データ判定・複製処理の流れを説明する。図１８は、共有データ判定・複製処理のフローチャートである。
【０１４９】
ステップ９０２：物理ページ解放処理部１３２は、内部変数「移動済み」を用意し、初期値としてＦＡＬＳＥを設定する。
【０１５０】
ステップ９０３：物理ページ解放処理部１３２は、検索テーブル１２９を検索し、ステップ８０６で得たフィンガープリントを“フィンガープリント１２９１”のカラムに含み、かつステップ８０４で得た重複排除ブロックの識別子を“重複排除ブロック１２９２”のカラムに含むレコードを特定する。
【０１５１】
ステップ９０４：物理ページ解放処理部１３２は、特定した検索テーブル１２９のレコードの重複排除ブロック１２９２に格納されている重複排除ブロックの識別子を選択する。
【０１５２】
ステップ９０５：物理ページ解放処理部１３２は、ステップ９０４で選択した重複排除ブロックの識別子に対応するマッピングテーブル１２７のレコードを参照し、レコードに格納された各種情報を得る。
【０１５３】
ステップ９０６：物理ページ解放処理部１３２は、処理対象のデータブロックが、ステップ９０４で選択した重複排除ブロックに割り当てられているか否かを判定する。割り当てられている場合には、処理対象のデータブロックは共有データブロックであることになる。判定の具体的な方法は以下のとおりである。すなわち、ステップ９０５で得られたレコードの物理ページ１２７５およびデータブロック１２７６で特定されるデータブロックが、処理対象のデータブロックと一致するならば、割り当てられていると判定される。一致する場合（ステップ９０６：Ｙ）、そのデータを別の物理ページのデータブロックへ移動（複製）し、重複排除ブロックに複製先のデータブロックを割り当てる処理を行う。この処理は、ステップ９０７のマッピング切替処理で実行される。マッピング切替処理の流れについては後述する。一致しない場合（ステップ９０６：Ｎ）、次にステップ９０８が実行される。
【０１５４】
なお、ステップ９０４で選択した重複排除ブロックが、重複排除解除処理を適用した論理ページに含まれる場合、ステップ９０６の判定は「一致しない」となる。なぜなら、重複排除解除処理のステップ３０６で、重複排除ブロックに対応するマッピングテーブル１２７のレコードの物理ページ１２７５およびデータブロック１２７６には、ＮＵＬＬが格納されるからである。
【０１５５】

ステップ９０８：物理ページ解放処理部１３２は、ステップ９０３で特定した検索テーブル１２９のレコードの重複排除ブロック１２９２から、ステップ９０４で選択した重複排除ブロックの識別子を削除する。この結果、重複排除ブロック１２９２に重複排除ブロックの識別子が登録されていない状態、つまり重複排除ブロック１２９２が空の状態となった場合は、当該レコードを検索テーブル１２９から削除する。

【０１５６】

ステップ９０９：物理ページ解放処理部１３２は、ステップ９０５で参照したマッピングテーブル１２７のレコードのフィンガープリント１２７４にＮＵＬＬを格納する。また、同じレコードの削減フラグ１２７７にＦＡＬＳＥを格納する。

【０１５７】

ステップ９１０：物理ページ解放処理部１３２は、ステップ９０３で特定した検索テーブル１２９のレコードの重複排除ブロック１２９２に格納されている重複排除ブロックの識別子のうち、ステップ９０５～９０９を実行していないものがあるかどうかを判定する。ステップ９０５～９０９をまだ実行していない重複排除ブロックの識別子がある場合（ステップ９１０：Ｙ）、物理ページ解放処理部１３２はそのうちの一つを選択し（ステップ９０４）、ステップ９０５からの処理を実行する。そうでない場合（ステップ９１０：Ｎ）、処理を終了する。

【０１５８】

図１９を参照しながら、ステップ９０７のマッピング切替処理の流れを説明する。図１９は、マッピング切替処理のフローチャートである。

【０１５９】

マッピング切替処理は、ステップ８０３で選択されたデータブロックのデータを、別の物理ページのデータブロックに移動（複製）する処理である。

【０１６０】

ステップ１００２：物理ページ解放処理部１３２は、ステップ９０２で用意した内部変数「移動済みフラグ」の値がＦＡＬＳＥ場合（ステップ１００２：Ｎ）、ステップ１００３を実行する。移動済みフラグがＴＲＵＥである場合（ステップ１００２：Ｙ）、ステップ１００３～ステップ１００６はスキップされる。

【０１６１】

ステップ１００３：物理ページ解放処理部１３２は、移動済みフラグにＴＲＵＥを設定する。

【０１６２】

ステップ１００４：物理ページ解放処理部１３２は、ステップ８０３で選択したデータブロックのデータの複製先となるデータブロックを確保する。データブロックの確保の処理内容は、ステップ２０８（重複排除処理）と同様である。

【０１６３】

ステップ１００５：物理ページ解放処理部１３２は、ステップ８０３で選択したデータブロックのデータを読み出し、ステップ１００３で確保したデータブロックに書き込む。この処理は、ステップ２０９（重複排除処理）と同様である。

【０１６４】

ステップ１００６：物理ページ解放処理部１３２は、ステップ９０４で選択した重複排除ブロックを、削減済み重複排除ブロックから、通常の重複排除ブロックへ変更する。このため、ステップ１００６では、物理ページ解放処理部１３２は、当該重複排除ブロックを含む論理ページの排除ブロック数（論理ページ管理テーブル１２６の排除ブロック数１２６７）を、１減じる。

【０１６５】

ステップ１００７：物理ページ解放処理部１３２は、ステップ９０４で選択した重複排除ブロックに対応するマッピングテーブル１２７のレコードを更新する。変更内容は以下のとおりである。まず、物理ページ１２７５とデータブロック１２７６には、ステップ１００４で確保したデータブロックの物理ページおよびデータブロックの識別子がそれぞれ

格納される。次に、ステップ１００２において移動済みフラグがＦＡＬＳＥであった場合、すなわち、削減済み重複排除ブロックを通常の重複排除ブロックへ変更した場合、削減フラグ１２７７の内容がＦＡＬＳＥに変更される。

【０１６６】

以上、本発明の実施例を説明したが、これらは、本発明の説明のための例示であって、本発明の範囲をこれらの実施例にのみ限定する趣旨ではない。すなわち、本発明は、他の種々の形態でも実施する事が可能である。

【０１６７】

上では、ストレージ装置が使用するいくつかの情報が、論理ページ管理テーブル１２６等のように、テーブル構造で管理される例を説明したが、テーブル構造で情報を管理する態様に限定されるわけではない。ストレージ装置はテーブル以外のデータ構造、例えばリスト構造などを用いて、情報を管理してもよい。

【０１６８】

また、ストレージ装置は重複排除処理に加えて圧縮処理を行ってもよい。たとえばストレージ装置に搭載される記憶デバイスとして、データを圧縮して格納する機能（圧縮機能）を有する記憶デバイスを用いることで、記憶デバイスに格納されるデータを圧縮するようにしてもよい。その際、通常論理ページに書き込まれたデータに対するアクセス性能を維持するために、重複排除論理ページに書き込まれたデータだけが、圧縮機能を有する記憶デバイスに格納されるようにするとよい。さらに別の実施形態として、ストレージ装置は、ストレージコントローラのＣＰＵが、記憶デバイスに格納されるデータの圧縮を行うように構成されていてもよい。

【０１６９】

また、上では重複排除論理ページへデータが書き込まれたとき、その重複排除論理ページを通常の論理ページへ変更した上で、データをドライブへ格納する例を説明したが、重複排除論理ページを通常の論理ページへ変更せずに、書き込まれたデータをドライブへ格納するようにしてもよい。例えば、重複排除論理ページへデータが書き込まれたとき、物理ページの領域を確保してデータを格納し、重複排除論理ページのアドレスと物理ページのアドレスのマッピング関係を更新するようにしてもよい。この場合、重複排除論理ページへライトされたデータの量、ライトの回数、それらの一定期間あたりの量などの指標から、重複排除論理ページを通常論理ページへ戻すか否かを判断するようにしてもよい。

【０１７０】

また、上では重複排除論理ページを通常の論理ページへ変更したときに、その論理ページのデータが変更前に格納されていた物理ページを未使用の状態に変更する例を説明したが、重複排除論理ページを通常の論理ページへ変更した後で、任意の契機で物理ページを未使用の状態に変更するようにしてもよい。例えば、ストレージ装置は、未使用の状態の物理ページの数、論理ページへ割り当てられている物理ページの数、重複排除論理ページのデータを格納するために割り当てられた物理ページの数、重複排除論理ページの数、ストレージ装置を構成するハードウェアの処理負荷などの指標から、物理ページを未使用の状態に変更する契機を決定するようにしてもよい。また、ストレージ装置は、物理ページごとに、格納したデータの量や、重複排除ブロックへの割り当てを解除した量等の情報を記録するようにしてもよい。そして、これらの物理ページごとの情報に基づいて、未使用の状態に変更する物理ページを選択するようにしてもよい。

【０１７１】

また、ストレージ装置のハードウェア構成は、上で説明した構成に限定されない。たとえば、上の実施例で述べた計算機に１以上の記憶デバイスを搭載した装置を、ストレージ装置として用いてもよい。そして、上の実施例で説明した各プログラムを計算機のプロセッサに実行させれば、上の実施例で説明したストレージ装置と同じことが実現できる。

【０１７２】

また、上で説明した実施例では、アクセス要求元の計算機５とストレージ装置１とが異なるハードウェアである例が説明されたが、別の実施形態として、計算機５とストレージ

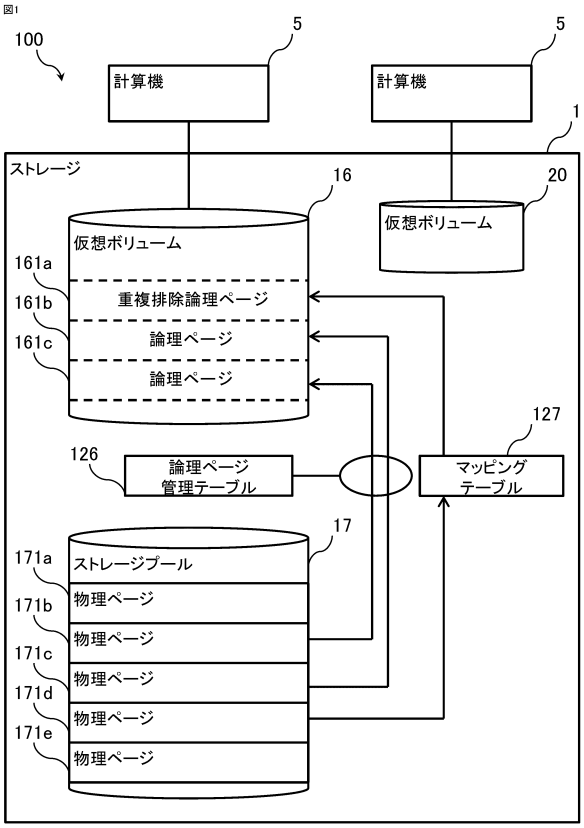
装置１とが単一のハードウェアとして実装されてもよい。つまり、上の実施例で説明した、ストレージ装置１で実行されるストレージ制御プログラムと、計算機５で実行されるアプリケーションプログラムとが、同一の計算機上で実行されてもよい。この場合、アプリケーションプログラムはアクセス要求元として、ストレージ制御プログラムに対してＩ／Ｏ要求を発行し、ストレージ制御プログラムはＩ／Ｏ要求に対する応答（Ｉ／Ｏ要求がリード要求の場合にはリードデータ）をアクセス要求元であるアプリケーションプログラムに返却するように構成されていても良い。好ましくは、計算機上で、仮想計算機を形成するためのプログラム（ハイパーバイザ等）を実行させることで、アプリケーションプログラムを実行する仮想計算機と、Ｉ／Ｏ処理部１２１や重複排除処理部１２３等のプログラムを実行する仮想計算機とを形成するとよい。

【符号の説明】

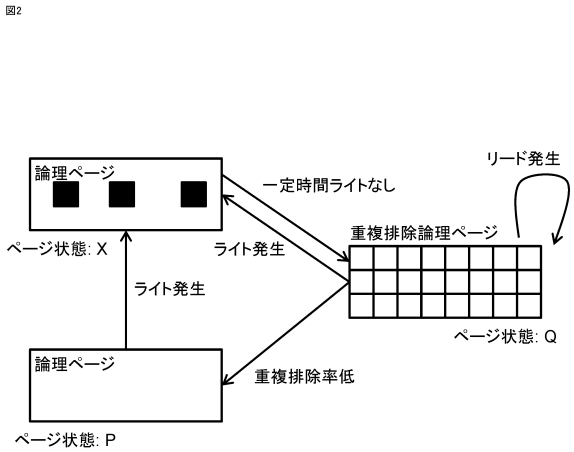
【０１７３】

１：ストレージ装置，　５：計算機，６：ＳＡＮ，　１０：ストレージコントローラ，　１１：ＣＰＵ，　１２：システムメモリ，　１４：キャッシュメモリ，　１５：記憶デバイス，　１２１：Ｉ／Ｏ処理部，　１２３：重複排除処理部，　１２４：重複排除解除部，　１２５：論理ページ変更部，　１２６：論理ページ管理テーブル，　１２７：マッピングテーブル，　１２８：プール管理テーブル，　１２９：検索テーブル，　１３０：逆参照テーブル，　１３１：追記ポインタ，　１３２：物理ページ解放処理部

【図１】

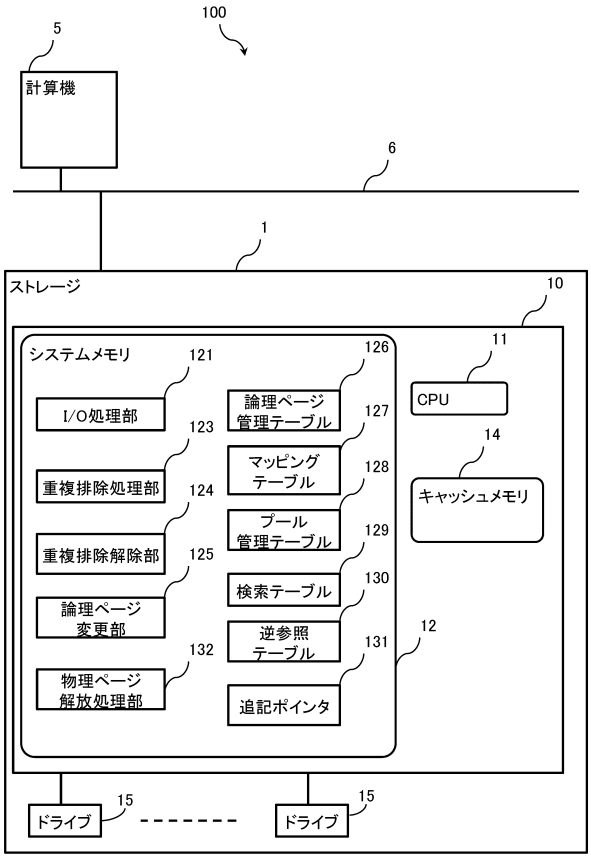


【図２】



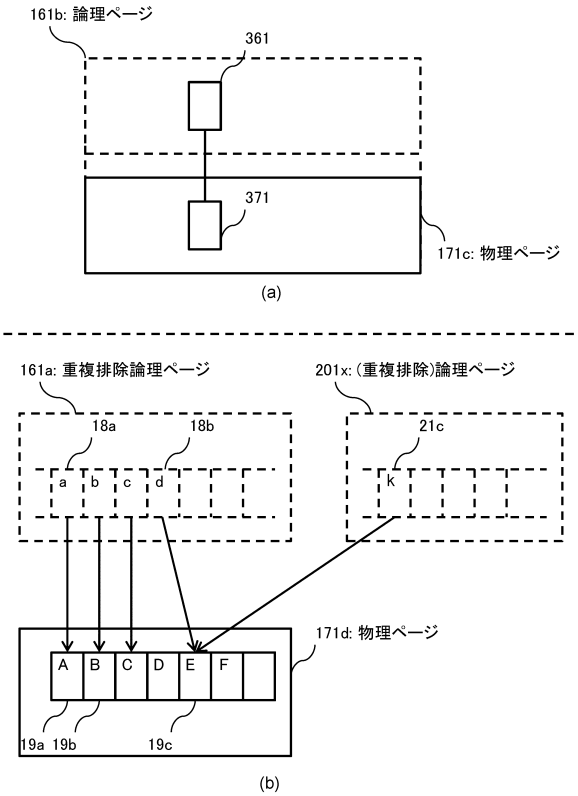
【図３】

図3



【図４】

図4



【図５】

図5

126: 論理ページ管理テーブル

仮想ボリューム	論理ページ	重複排除	物理ページ	状態	最終ライト時刻	排除ブロック数
16	161a	有効	NULL	Q	NULL	0
16	161b	無効	171b	X	12/10 14:30	0
16	161c	無効	171c	P	12/15 20:00	0
20	201a	有効	NULL	Q	NULL	100
1261	1262	1263	1264	1265	1266	1267

【図７】

図7

128: プール管理テーブル

物理ページ	使用状況	論理ページ
171a	未使用	NULL
171b	論理ページ	16b
171c	論理ページ	16c
171d	データブロック	NULL
171e	未使用	NULL
1281	1282	1283

【図６】

図6

127: マッピングテーブル

仮想ボリューム	論理ページ	重複排除ブロック	フィンガープリント	物理ページ	データブロック	削減フラグ
…	…	…	…	…	…	…
16	161a	a	11111111	171d	A	FALSE
16	161a	b	22222222	171d	B	FALSE
16	161a	c	33333333	171d	C	FALSE
16	161a	d	44444444	171d	E	FALSE
…	…	…	…	…	…	…
20	201x	k	44444444	171d	E	TRUE
…	…	…	…	…	…	…
1271	1272	1273	1274	1275	1276	1277

【図８】

図8

129: 検索テーブル

フィンガープリント	重複排除ブロック
…	…
11111111	16:a
…	…
22222222	16:b
…	…
33333333	16:c
…	…
44444444	16:d, 20:k
…	…
1291	1292

【図９】

図9

130: 逆参照テーブル

物理ページ	データブロック	重複排除ブロック
・・・	・・・	・・・
171d	A	16:a
171d	B	16:b
171d	C	16:c
171d	E	16:d
・・・	・・・	・・・

1301 1302 1303

【図１０】

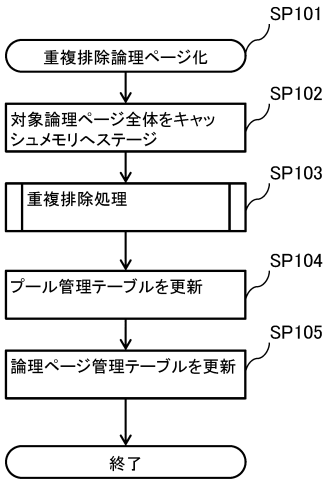
図10

131: 追記ポインタ

171d:F

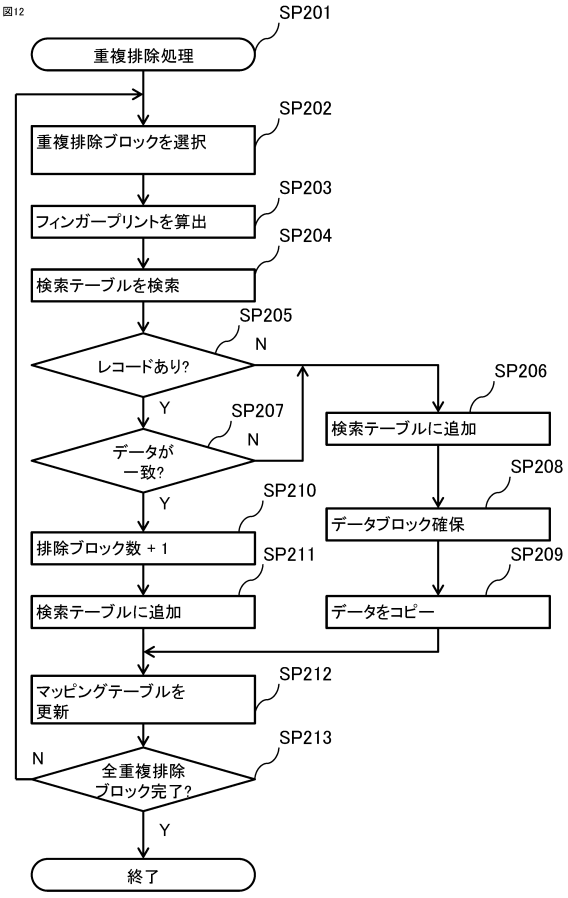
【図１１】

図11



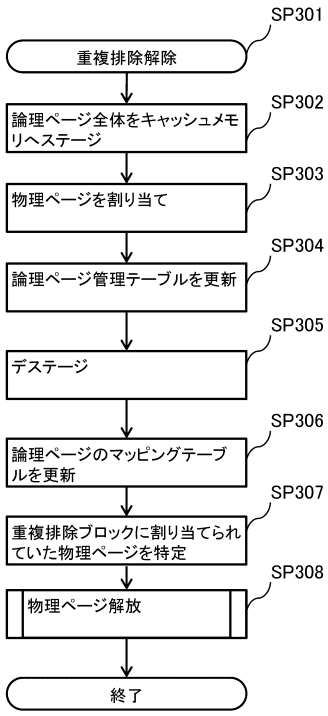
【図１２】

図12



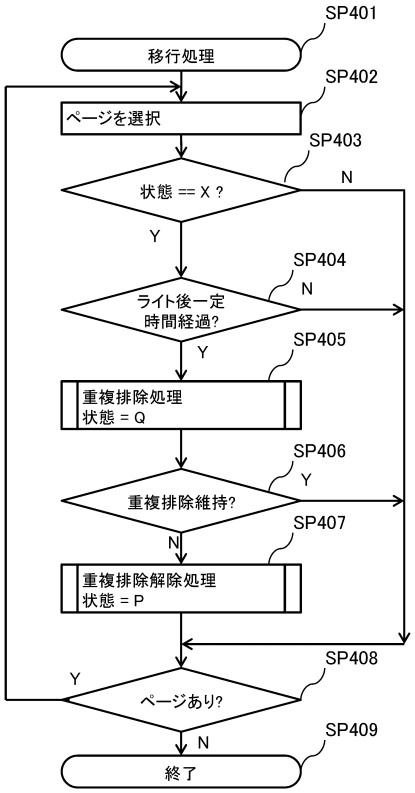
【図１３】

図13



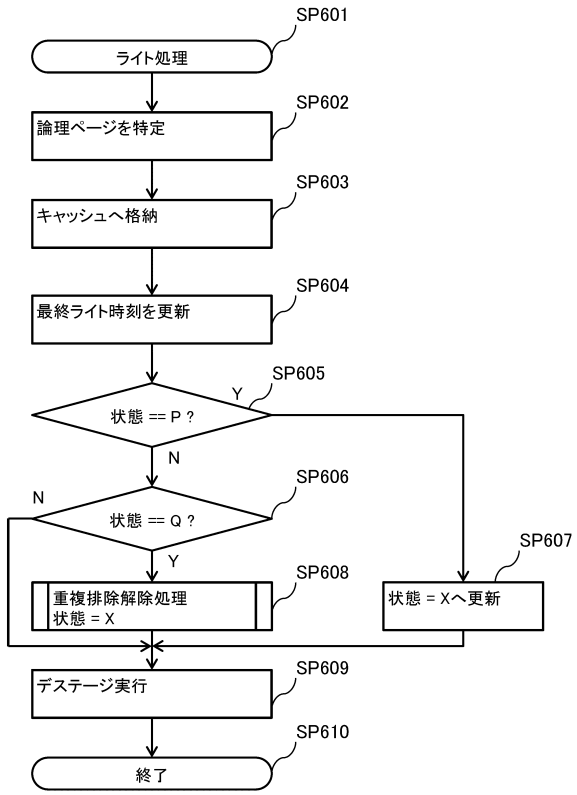
【図１４】

図14



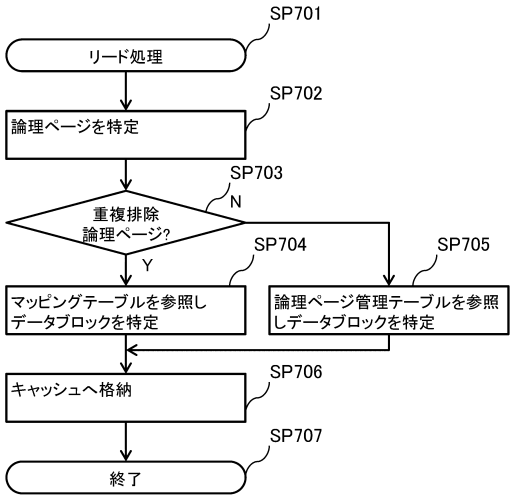
【図１５】

図15



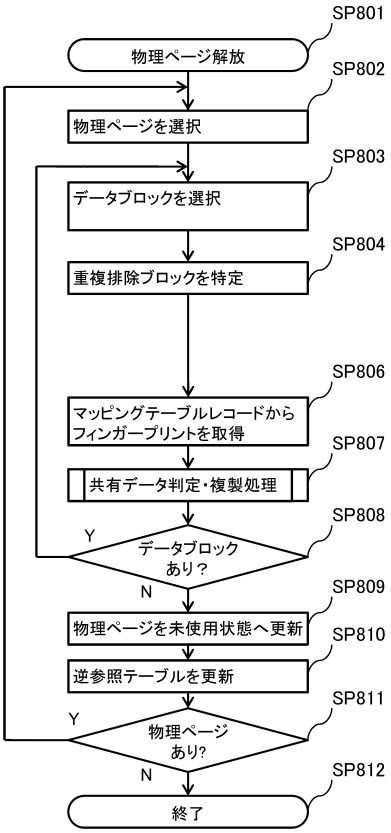
【図１６】

図16



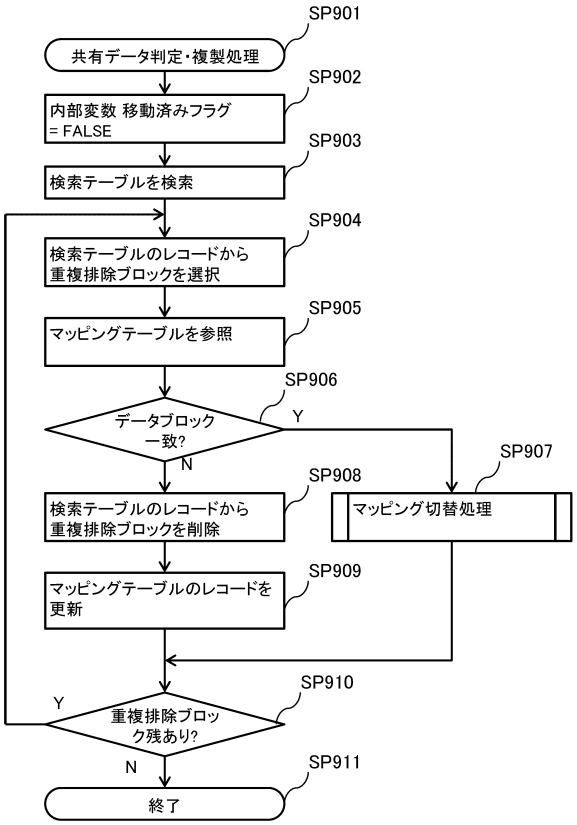
【図１７】

図17



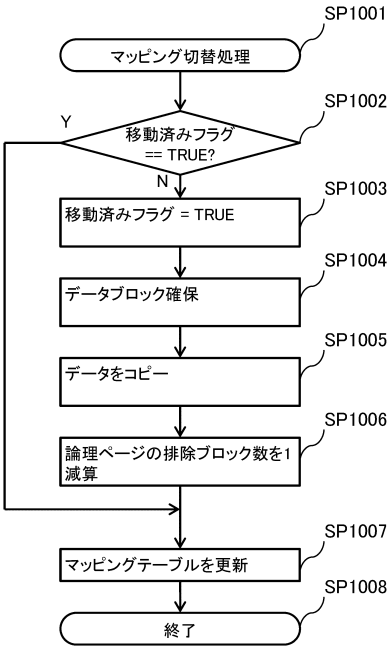
【図１８】

図18



【図１９】

図19



フロントページの続き

(72)発明者　渡邊　恭男
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内
(72)発明者　吉井　義裕
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内
(72)発明者　松上　一樹
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内

審査官　田中　啓介

(56)参考文献　特開２０１１－０６５３１４（ＪＰ，Ａ）
国際公開第２０１４／１５７２４３（WO，Ａ１）
特表２０１２－５２３０２３（ＪＰ，Ａ）
特表２０１５－５２８９２８（ＪＰ，Ａ）
特開２００９－０９３５７１（ＪＰ，Ａ）
特開２００８－２３４１５８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ３／０６－３／０８
Ｇ０６Ｆ１２／００－１２／０６
Ｇ０６Ｆ１３／１０－１３／１８
Ｇ０６Ｆ１６／００－１６／９５８