

(19)日本国特許庁(JP)

## (12)特許公報(B2)

(11)特許番号  
特許第7065928号  
(P7065928)

(45)発行日 令和4年5月12日(2022.5.12)

(24)登録日 令和4年4月28日(2022.4.28)

(51)国際特許分類

F I

G 0 6 F	3/06 (2006.01)	G 0 6 F	3/06	3 0 1 W
G 0 6 F	11/20 (2006.01)	G 0 6 F	3/06	3 0 2 A
G 0 6 F	12/0866(2016.01)	G 0 6 F	3/06	3 0 4 B
G 0 6 F	12/0868(2016.01)	G 0 6 F	3/06	3 0 5 C
G 0 6 F	16/185(2019.01)	G 0 6 F	11/20	6 8 9

請求項の数 3 (全25頁) 最終頁に続く

(21)出願番号 特願2020-185556(P2020-185556)  
 (22)出願日 令和2年11月6日(2020.11.6)  
 (62)分割の表示 特願2018-60662(P2018-60662)の  
 分割  
 原出願日 平成30年3月27日(2018.3.27)  
 (65)公開番号 特開2021-39771(P2021-39771A)  
 (43)公開日 令和3年3月11日(2021.3.11)  
 審査請求日 令和2年11月6日(2020.11.6)

(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 110001678  
 特許業務法人藤央特許事務所  
 (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  
 (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 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 川口 智大

最終頁に続く

(54)【発明の名称】 ストレージシステム及びその制御方法

(57)【特許請求の範囲】

【請求項1】

第1のストレージ制御部と、第2のストレージ制御部と、少なくとも前記第1のストレージ制御部に接続され、不揮発性の記憶媒体を有するストレージドライブと、を有するストレージシステムであって、

前記第1のストレージ制御部は、データを格納する第1のキャッシュ領域と、データを格納する第1のバッファ領域と、を有しており、

前記第2のストレージ制御部は、それぞれ、データを格納する第2のキャッシュ領域と、データを格納する第2のバッファ領域と、を有しており、

前記第1のストレージ制御部は、前記第1のキャッシュ領域に格納されたデータを前記第2のキャッシュ領域にも格納して二重化を行うようになっており、

前記第1のストレージ制御部は、ホスト計算機からデータの書き込み命令を受信すると、前記書き込み命令の対象のデータを、前記第1のストレージ制御部の前記第1のキャッシュ領域に格納するとともに、前記第1のキャッシュ領域に格納したデータを前記第2のストレージ制御部の前記第2のキャッシュ領域に格納して二重化を行い、前記二重化が完了したら、前記ホスト計算機に、前記データの書き込みの終了を示す応答を送信し、

前記第1のストレージ制御部は、前記書き込み命令の対象であるいずれかの前記キャッシュ領域に格納された二重化されたデータのうち前記ストレージドライブに格納されるデータについては、前記キャッシュ領域から読み出した前記書き込み命令の対象のデータに圧縮処理を行って前記第1のキャッシュ領域に格納せずに前記第1のバッファ領域に格納し

、前記第1のバッファ領域に前記圧縮処理して格納されたデータに基づいてパリティを生成して前記第1のバッファ領域に格納し、前記第1のバッファ領域に格納したデータ及びパリティを読み出して前記ストレージドライブに送信して格納させ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に記載のストレージシステムであって、

前記第1のストレージ制御部に接続された複数の前記ストレージドライブを有し、

前記第1のストレージ制御部は、前記第1のバッファ領域に前記圧縮処理して格納したデータの量がパリティを生成するための所定のデータ量に達した場合、前記第1のバッファ領域から読み出したデータに基づいてパリティを作成し、前記第1のバッファ領域から読み出したデータ及び前記パリティを前記複数のストレージドライブに送信することを特徴とするストレージシステム。

10

【請求項3】

ストレージシステムの制御方法であって、

前記ストレージシステムは、第1のストレージ制御部と、第2のストレージ制御部と、少なくとも前記第1のストレージ制御部に接続され、不揮発性の記憶媒体を有するストレージドライブと、を有し、

前記第1のストレージ制御部は、データを格納する第1のキャッシュ領域と、データを格納する第1のバッファ領域と、を有しており、前記第2のストレージ制御部は、それぞれ、データを格納する第2のキャッシュ領域と、データを格納する第2のバッファ領域と、を有しており、前記第1のストレージ制御部は、前記第1のキャッシュ領域に格納されたデータを前記第2のキャッシュ領域にも格納して二重化を行うようになっており、

20

前記ストレージシステムの制御方法は、

前記第1のストレージ制御部が、ホスト計算機からデータの書き込み命令を受信すると、前記書き込み命令の対象のデータを、前記第1のストレージ制御部の前記第1のキャッシュ領域に格納するとともに、前記第1のキャッシュ領域に格納したデータを前記第2のストレージ制御部の前記第2のキャッシュ領域に格納して二重化を行い、前記二重化が完了したら、前記ホスト計算機に、前記データの書き込みの終了を示す応答を送信する手順と、

前記第1のストレージ制御部が、前記書き込み命令の対象であるいずれかの前記キャッシュ領域に格納された二重化されたデータのうち前記ストレージドライブに格納されるデータについては、前記キャッシュ領域から読み出した前記書き込み命令の対象のデータに圧縮処理を行って前記第1のキャッシュ領域に格納せずに前記第1のバッファ領域に格納し、前記第1のバッファ領域に前記圧縮処理して格納されたデータに基づいてパリティを生成して前記第1のバッファ領域に格納し、前記第1のバッファ領域に格納したデータ及びパリティを読み出して前記ストレージドライブに送信して格納させる手順と、を含むことを特徴とするストレージシステムの制御方法。

30

【発明の詳細な説明】

【技術分野】

【0001】

本発明はストレージシステムに関する。

40

【背景技術】

【0002】

ストレージシステムは、一般的に、1以上のストレージ装置を備える。1以上のストレージ装置の各々は、一般的に、記憶デバイスとして、例えば、HDD (Hard Disk Drive) 又はSSD (Solid State Drive) を備える。ストレージシステムが、SAN (Storage Area Network) 又はLAN (Local Area Network) といったネットワーク経由で、1又は複数の上位装置 (例えば、ホスト計算機) からアクセスされる。一般的に、ストレージ装置は、RAID (Redundant Array of Independent (or Inexpensive) Disks) 技術に従う高信頼化方法を用いることで信頼性を向上している。

50

## 【 0 0 0 3 】

特許文献 1 には、ホスト計算機からのデータ書き込み速度を維持しながら、データを圧縮させることが出来る情報システムが開示されている。特許文献 1 によれば、ストレージ装置においてホスト計算機からのデータ書き込みを受け付ける第 1 ボリュームと、第 1 ボリューム上のデータを圧縮して管理する第 2 ボリュームを提供する。ホスト計算機から第 1 ボリュームに対するデータ書き込みを終えると、ストレージ装置はホスト計算機に対して書き込み処理が完了したとして応答を返す。その後ストレージ装置は、ホスト計算機からのデータ書き込みとは非同期的な契機にデータを圧縮して第 2 ボリュームに格納する。

## 【 0 0 0 4 】

非特許文献 1 には、ホスト計算機から書き込まれた重複するデータを一つにまとめる重複排除処理について、ストレージ装置の稼働率に応じて処理契機を切り替えることで、レスポンスとスループットを両立させる方法について開示されている。

10

## 【 0 0 0 5 】

例えば、非特許文献 1 には、「方式の違いによって I O P S やレイテンシーに関する特性が異なり、これらを使い分けることで dedup-back 方式の低レイテンシー、dedup-through 方式の高 I O P S を実現するのが本稿で提案するハイブリッド方式である。」及び「本稿では、従来の同期的に重複除去を行う dedup-through 方式に加えて、非同期に重複除去を行う dedup-back 方式の 2 つを比較して、dedup-through 方式の高い I O P S 性能と同期的な重複除去処理のオーバーヘッドによる高レイテンシー、dedup-back 方式の低レイテンシーと tail latency の増加に伴う I O P S 低下を明らかにして、この 2 つの方式を組み合わせることで高 I O P S と低レイテンシーの両立を目指すハイブリッド方式を提案した。」と記載されている。

20

## 【 0 0 0 6 】

すなわち、非特許文献 1 によれば、ストレージ装置の稼働率が低い場合、ホスト計算機からのデータ書き込みを終えてから重複排除処理を実施することで応答時間を短くし、稼働率が高い場合はデータ書き込みと同時に重複排除処理を実施する。

## 【 先行技術文献 】

## 【 特許文献 】

## 【 0 0 0 7 】

【 文献 】 米国特許出願公開第 2 0 0 9 / 0 1 4 4 4 9 6 号明細書

30

## 【 非特許文献 】

## 【 0 0 0 8 】

【 文献 】 加藤 純，大辻 弘貴，鈴木 康介，佐藤 充，吉田 英司：「インメモリー重複除去における書き込み高速化」，研究報告コンピュータシステム・シンポジウム，2 0 1 6 年 1 1 月 2 8 日，p . 5 1 - 5 9

## 【 発明の概要 】

## 【 発明が解決しようとする課題 】

## 【 0 0 0 9 】

データ書き込みにおいて R A I D 技術に従ったデータ保護を行うには、冗長化に必要なデータ量（パリティサイクル）を集める必要がある。パリティサイクル分のデータが集まるまでキャッシュメモリ上でのデータ保護が必要なため、キャッシュメモリ上のデータは二重化される。これは、ホスト計算機から書き込まれたデータ及び圧縮されたデータについても同様に行われる。このような場合、データ書き込みの最大速度は、データの読み出し及び二重化によるキャッシュアクセス量によって制限される。

40

## 【 0 0 1 0 】

キャッシュアクセス量を低減する方法として、書き込みと同期してデータを圧縮することによって圧縮前のデータを二重化する処理を省略する方法が考えられる。しかし、ホスト計算機に対しての処理完了の応答を返すには、圧縮データを二重化するため、圧縮処理の時間だけ応答速度が遅くなる。

## 【 0 0 1 1 】

50

このような課題は、圧縮機能を有するストレージシステムに限らず、重複排除などの他のデータ量削減機能を有するストレージシステム、及び、暗号化又は冗長化などを行うストレージシステムについてもあり得る。

【課題を解決するための手段】

【0012】

上記の課題の少なくとも一つを解決するための本発明の代表的な一例を示せば、次の通りである。すなわち、第1のストレージ制御部と、第2のストレージ制御部と、少なくとも前記第1のストレージ制御部に接続され、不揮発性の記憶媒体を有するストレージドライブと、を有するストレージシステムであって、前記第1のストレージ制御部は、データを格納する第1のキャッシュ領域と、データを格納する第1のバッファ領域と、を有しており、前記第2のストレージ制御部は、それぞれ、データを格納する第2のキャッシュ領域と、データを格納する第2のバッファ領域と、を有しており、前記第1のストレージ制御部は、前記第1のキャッシュ領域に格納されたデータを前記第2のキャッシュ領域にも格納して二重化を行うようになっており、前記第1のストレージ制御部は、ホスト計算機からデータの書き込み命令を受信すると、前記書き込み命令の対象のデータを、前記第1のストレージ制御部の前記第1のキャッシュ領域に格納するとともに、前記第1のキャッシュ領域に格納したデータを前記第2のストレージ制御部の前記第2のキャッシュ領域に格納して二重化を行い、前記二重化が完了したら、前記ホスト計算機に、前記データの書き込みの終了を示す応答を送信し、前記第1のストレージ制御部は、前記書き込み命令の対象であるいずれかの前記キャッシュ領域に格納された二重化されたデータのうち前記ストレージドライブに格納されるデータについては、前記キャッシュ領域から読み出した前記書き込み命令の対象のデータに圧縮処理を行って前記第1のキャッシュ領域に格納せずに前記第1のバッファ領域に格納し、前記第1のバッファ領域に前記圧縮処理して格納されたデータに基づいてパリティを生成して前記第1のバッファ領域に格納し、前記第1のバッファ領域に格納したデータ及びパリティを読み出して前記ストレージドライブに送信して格納させることを特徴とする。

10

20

【発明の効果】

【0013】

本発明の一態様によれば、圧縮処理から記憶デバイスへの格納までを一括で行うことによって、圧縮データの二重化処理が省略される。圧縮データの二重化が不要になることで、キャッシュアクセス量を削減し、データ書き込みの最大速度が向上できる。また、記憶デバイスへの圧縮データの格納が完了するまでキャッシュメモリ上に圧縮前のデータを保持することによって、圧縮処理や記憶デバイスへの格納などの処理中に装置障害が発生してもデータを保護することが出来る。

30

【0014】

上記した以外の課題、構成及び効果は、以下の実施形態の説明によって明らかにされる。

【図面の簡単な説明】

【0015】

【図1】本発明の実施例1のストレージシステムが実行する、データ圧縮処理を伴うデータライト手順を示す説明図である。

40

【図2】本発明の実施例1のストレージ装置の構成を示すブロック図である。

【図3】本発明の実施例1のストレージ装置が保持するVOL管理テーブルの構成例を示す説明図である。

【図4】本発明の実施例1のストレージ装置が保持するプール構成管理テーブルの構成例を示す説明図である。

【図5】本発明の実施例1のストレージ装置が保持するRAID構成管理テーブルの構成例を示す説明図である。

【図6】本発明の実施例1のストレージ装置が保持するプール割当管理テーブルの構成例を示す説明図である。

【図7】本発明の実施例1のストレージ装置が保持するドライブ割当管理テーブルの構成

50

例を示す説明図である。

【図 8】本発明の実施例 1 のストレージ装置によって管理される論理記憶階層の構成例を示す説明図である。

【図 9】本発明の実施例 1 のストレージ装置が保持するメモリ割当管理テーブルの構成例を示す説明図である。

【図 10】本発明の実施例 1 のストレージ装置におけるメモリ割当の構成例を示す図である。

【図 11】本発明の実施例 1 のストレージ装置が実行するリード処理を示すフローチャートである。

【図 12】本発明の実施例 1 のストレージ装置が実行するライト処理を示すフローチャートである。

10

【図 13】本発明の実施例 1 のストレージ装置が実行するデステージ処理を示すフローチャートである。

【図 14】本発明の実施例 1 のストレージ装置が実行する、排他手順を変更したデステージ処理を示すフローチャートである。

【発明を実施するための形態】

【0016】

以下の説明では、「インターフェース部」は、ユーザインターフェース部と、通信インターフェース部とのうちの少なくとも 1 つを含んでよい。ユーザインターフェース部は、1 以上の I/O デバイス（例えば入力デバイス（例えばキーボード及びポインティングデバイス）と出力デバイス（例えば表示デバイス））と表示用計算機とのうちの少なくとも 1 つの I/O デバイスを含んでよい。通信インターフェース部は、1 以上の通信インターフェースデバイスを含んでよい。1 以上の通信インターフェースデバイスは、1 以上の同種の通信インターフェースデバイス（例えば 1 以上の NIC（Network Interface Card））であってもよいし 2 以上の異種の通信インターフェースデバイス（例えば NIC と HBA（Host Bus Adapter））であってもよい。

20

【0017】

また、以下の説明では、「メモリ部」は、1 以上のメモリを含む。少なくとも 1 つのメモリは、揮発性メモリであってもよいし不揮発性メモリであってもよい。メモリ部は、主に、プロセッサ部による処理の際に使用される。

30

【0018】

また、以下の説明では、「プロセッサ部」は、1 以上のプロセッサを含む。少なくとも 1 つのプロセッサは、典型的には、CPU（Central Processing Unit）である。プロセッサは、処理の一部又は全部を行うハードウェア回路を含んでもよい。

【0019】

また、以下の説明では、「xxx テーブル」といった表現にて情報を説明することがあるが、情報は、どのようなデータ構造で表現されていてもよい。すなわち、情報がデータ構造に依存しないことを示すために、「xxx テーブル」を「xxx 情報」と言うことができる。また、以下の説明において、各テーブルの構成は一例であり、1 つのテーブルは、2 以上のテーブルに分割されてもよいし、2 以上のテーブルの全部又は一部が 1 つのテーブルであってもよい。

40

【0020】

また、以下の説明では、同種の要素を区別しないで説明する場合には、参照符号のうちの共通符号を使用し、同種の要素を区別する場合は、参照符号（又は要素の ID（例えば識別番号））を使用することがある。例えば、複数のストレージコントローラを区別しない場合には、「ストレージコントローラ 22」と記載し、各ストレージコントローラを区別する場合には、「ストレージコントローラ 1\_\_22A」、「ストレージコントローラ 2\_\_22B」のように記載する。他の要素（例えばキャッシュ領域 203、バッファ領域 202、アドレス 1100、1101、1104 等）も同様である。

【0021】

50

また、以下の説明では、「ストレージシステム」は、1以上のストレージ装置を含む。少なくとも1つのストレージ装置は、汎用的な物理計算機であってもよい。また、少なくとも1つのストレージ装置が、仮想的なストレージ装置であってもよいし、SDx (Software-Defined anything) を実行してもよい。SDxとしては、例えば、SDS (Software Defined Storage) (仮想的なストレージ装置の一例) 又はSDDC (Software-defined Datacenter) を採用することができる。

【0022】

以下、本発明の実施例を図面に基づいて説明する。

【実施例1】

【0023】

以下、本発明の実施例1を説明する。

【0024】

<記憶デバイスへの圧縮データの格納手順>

図1は、本発明の実施例1のストレージシステム100が実行する、データ圧縮処理を伴うデータライト手順を示す説明図である。

【0025】

ストレージシステム100は、ホスト計算機30及びストレージ装置11によって構成される。ホスト計算機30は、ネットワーク31を介してストレージ装置11に接続され、管理計算機(図示せず)によって管理される。ストレージ装置11は、1以上のボリューム(論理的な記憶領域)を有する。ホスト計算機30は、物理的な計算機でもよいし、物理的な計算機で実行される仮想的な計算機でもよい。ホスト計算機30は、ストレージシステムにおいて実行される仮想的な計算機でもよい。

【0026】

ホスト計算機30からは、ストレージ装置11のストレージコントローラ1\_\_22A又はストレージコントローラ2\_\_22Bに対してデータの書き込みが行われる。このストレージシステム100において、ホスト計算機30からの圧縮処理を伴うデータのライト処理について説明する。

【0027】

本実施例では、ホスト計算機30からのライト命令をストレージコントローラ1\_\_22Aが受領した場合について示す。

【0028】

具体例は、下記に示す通りである。

【0029】

(S1) ストレージ装置11は、ホスト計算機30からネットワーク31を介してライト命令を受信する。ライト命令は、データとデータの割当先アドレス1100とを含んでいる。ライト命令を受信した場合に、S2以降のライト処理が開始する。

【0030】

(S2) ストレージ装置11は、ライト命令に応答して、割当先アドレス1100が示すスロットの排他を確保する。これによって、そのスロット内のデータが他のライト命令によって更新されることを防ぐ。「スロット」とは、ボリューム(VOL)における領域である。具体的には、本実施例のスロットは、ドライブ29への書き込みが行われたか否か、及び、バッファ領域202への転送が行われたか否か等の管理の単位となる領域である。本実施例ではこの領域を「スロット」と呼ぶが、他の名称で呼ばれてもよい。

【0031】

「スロットの排他を確保」とは、ホスト計算機30からのリード命令及びライト命令で指定されたアドレスが示すスロットに対するリード及びライトを防ぐ操作であり、排他を確保したことをホスト計算機30が認識するための情報が管理される。なお、この情報はビットマップ又は時間情報など識別できるものであれば種別は問わない。また、本実施例において、「スロット」が、VOL(例えば、シンプロビジョニングに従うVOLであるTP-VOL)における領域であるのに対し、「データ領域」は、スロットに割り当てられ

10

20

30

40

50

る領域（例えば、プール内の領域であるプール領域）である。

【0032】

(S3) ストレージ装置11のストレージコントローラ1\_\_22A内、キャッシュ領域203Aにおいて、データの割当先アドレス1100に対応するアドレス1100Aにデータを格納する。

【0033】

(S4) ストレージコントローラ1\_\_22Aは、キャッシュ領域203A内に格納されたデータをストレージコントローラ2\_\_22Bに転送する。ストレージコントローラ2\_\_22Bは、割当先アドレス1100に対応するキャッシュ領域203B内のアドレス1100Bに受領したデータを格納して、ストレージコントローラ1\_\_22Aへ応答を返すことでストレージ装置11内での二重化を完了する。

10

【0034】

(S5) 二重化を完了した後にストレージ装置11からホスト計算機30に対してネットワーク31を介してライト完了を応答する。なお、この時点でホスト計算機30はライトが完了したと認識する。

【0035】

(S6) ストレージコントローラ1\_\_22Aは、キャッシュ領域203Aからドライブへ書き出すデータを選択し、選択したデータを圧縮してバッファ領域202A内のアドレス1101Aに格納する。なお、この処理はバッファ領域202A内にパリティサイクル分のデータが溜まるまで実施される。

20

【0036】

また、後述するように、ストレージコントローラ1\_\_22Aは、選択したデータを圧縮せずにそのままアドレス1101Aに格納してもよいし、圧縮以外の処理（例えば重複排除又は暗号化等）を行って、処理後のデータをアドレス1101Aに格納してもよい。

【0037】

(S7) ストレージコントローラ1\_\_22Aは、バッファ領域202A内のデータ量がパリティサイクル分に達すると、格納したデータからパリティデータを生成し、バッファ領域202A内のアドレス1104Aへ格納する。

【0038】

(S8) ストレージコントローラ1\_\_22Aは、バッファ領域202A内の圧縮データ及びパリティデータをドライブ29へ書き出す（デステージ処理）。

30

【0039】

(S9) ストレージコントローラ1\_\_22Aは、デステージ処理が完了すると、(S2)において確保したスロットの排他を解放する。

【0040】

以上が、ライト処理の一例である。

【0041】

<ストレージ装置>

図2は、本発明の実施例1のストレージ装置11の構成を示すブロック図である。

【0042】

ストレージ装置11は、1以上のストレージコントローラ22と、1以上のストレージコントローラ22に接続された複数のドライブ29とを有する。

40

【0043】

ストレージコントローラ22は、ホスト計算機30との通信を行うFE\_\_I/F（フロントエンドインターフェースデバイス）23、ストレージ装置間での通信を行うためのストレージI/F（ストレージインターフェースデバイス）28、装置全体を制御するプロセッサ24、プロセッサ24で使用するプログラム及び情報を格納するメモリ25、ドライブ29との通信を行うBE\_\_I/F（バックエンドインターフェースデバイス）27、及びそれらをつなぐ内部ネットワーク26を備える。

【0044】

50

メモリ 25 は、プログラムを管理するプログラム領域 201、データの転送及びコピーの時の一時的な保存領域であるバッファ領域 202、ホスト計算機 30 からのライトデータ（ライト命令に 응답して書き込まれるデータ）及びドライブ 29 からのリードデータ（リード命令に 응답して読み出されたデータ）を一時的に格納するキャッシュ領域 203、及び、種々のテーブルを格納するテーブル管理領域 206 を有する。

【0045】

キャッシュ領域 203 は、ホスト計算機 30 からのライトデータを一時的に格納する非圧縮データ格納領域 204、及び、圧縮したデータを格納する圧縮データ格納領域 205 を有する。テーブル管理領域 206 は、VOL に関する情報を保持する VOL 管理テーブル 207、プールに関する情報を保持するプール構成管理テーブル 208、RAID 構成に関する情報を保持する RAID 構成管理テーブル 209、プール割当てに関する情報を保持するプール割当て管理テーブル 210、ドライブ割当てに関する情報を保持するドライブ割当て管理テーブル 211、及び、メモリ割当てに関する情報を保持するメモリ割当て管理テーブル 212 を格納する。

10

【0046】

ドライブ 29 は、不揮発性のデータ記憶媒体を有する装置であり、例えば SSD (Solid State Drive) でも HDD (Hard Disk Drive) でもよい。複数のドライブ 29 が、複数の RAID グループ（パリティグループとも呼ばれる）を構成してよい。各 RAID グループは、1 以上のドライブ 29 から構成される。

【0047】

FE\_I / F23、BE\_I / F27 及びストレージ I / F28 が、インターフェース部の一例である。メモリ 25 が、メモリ部の一例である。プロセッサ 24 が、プロセッサ部の一例である。

20

【0048】

<VOL 管理テーブル>

図 3 は、本発明の実施例 1 のストレージ装置 11 が保持する VOL 管理テーブル 207 の構成例を示す説明図である。

【0049】

VOL 管理テーブル 207 は、VOL 毎にエントリを有する。各エントリは、VOL\_ID 41、VOL 属性 42、VOL 容量 43 及びプール ID 44 といった情報を格納する。以下、1 つの VOL（図 3 の説明において「対象 VOL」）を例に取る。

30

【0050】

VOL\_ID 41 は、対象 VOL の ID である。VOL 属性 42 は、対象 VOL の属性（例えば、対象 VOL がシンプロビジョニングを適用される VOL であるか、通常の VOL であるか、また、圧縮が有効であるか否かなど）を示す。VOL 容量 43 は、対象 VOL の容量を示す。プール ID 44 は、対象 VOL に関連付けられているプールの ID である。

【0051】

プロセッサ 24 は、デステージ処理において、VOL 管理テーブル 207 の VOL 属性 42 を参照することで、データ圧縮を必要とする VOL か否かを判定できる。例えば、VOL 属性 42 “圧縮有効” ならばデータ圧縮処理を行う。

40

【0052】

<構成管理テーブル>

図 4 は、本発明の実施例 1 のストレージ装置 11 が保持するプール構成管理テーブル 208 の構成例を示す説明図である。

【0053】

プールは、1 以上の RAID グループを基に構成された論理記憶領域である。プール構成管理テーブル 208 は、プール毎にエントリを有する。各エントリは、プール ID 51、RAID グループ ID 52、プール容量 53 及びプール使用容量 54 といった情報を格納する。以下、1 つのプール（図 4 の説明において「対象プール」）を例に取る。

【0054】

50

プールID 5 1は、対象プールのIDである。RAIDグループID 5 2は、対象プールの基になっている1以上のRAIDグループの各々のIDである。プール容量5 3は、対象プールの容量を示す。プール使用容量5 4は、対象プールのプール容量のうちVOLに割り当てられている領域の総量を示す。

【0055】

図5は、本発明の実施例1のストレージ装置11が保持するRAID構成管理テーブル209の構成例を示す説明図である。

【0056】

RAID構成管理テーブル209は、RAIDグループ毎にエントリを有する。各エントリは、RAIDグループID 6 1、RAIDレベル6 2、ドライブID 6 3、ドライブ種別6 4、容量6 5及び使用容量6 6といった情報を格納する。以下、1つのRAIDグループ(図5の説明において「対象RAIDグループ」)を例に取る。

10

【0057】

RAIDグループID 6 1は、対象RAIDグループのIDである。RAIDレベル6 2は、対象RAIDグループに適用されるRAIDアルゴリズムの種別を示す。ドライブID 6 3は、対象RAIDグループを構成する1以上のドライブの各々のIDである。ドライブ種別6 4は、対象RAIDグループを構成するドライブの種別(例えばHDDかSSDか)を示す。容量6 5は、対象RAIDグループの容量を示す。使用容量6 6は、対象RAIDグループの容量のうちの使用されている容量を示す。

【0058】

20

<割当管理テーブル>

図6は、本発明の実施例1のストレージ装置11が保持するプール割当管理テーブル210の構成例を示す説明図である。

【0059】

プール割当管理テーブル210は、VOLアドレス(VOL内のスロットを示すアドレス)毎にエントリを有する。各エントリは、VOL\_ID 7 1、VOLアドレス7 2、プールID 7 3、プールアドレス7 4、圧縮前サイズ7 5、圧縮後サイズ7 6、及び圧縮率7 7といった情報を格納する。以下、1つのVOLアドレス(図6の説明において「対象VOLアドレス」)を例に取る。

【0060】

30

VOL\_ID 7 1は、対象VOLアドレスによって識別されるスロットが属するVOLのIDである。VOLアドレス7 2は、対象VOLアドレスである。プールID 7 3は、対象VOLアドレスに割り当てられているデータ領域を含むプールのIDである。プールアドレス7 4は、対象VOLアドレスに割り当てられているデータ領域のアドレス(プールに属するアドレス)である。圧縮前サイズ7 5は、対象プールアドレスを指定したライト命令に従うデータの圧縮前サイズを示す。圧縮後サイズ7 6は、対象プールアドレスを指定したライト命令に従うデータの圧縮後のサイズを示す。圧縮率7 7は、圧縮後サイズ7 6 / 圧縮前サイズ7 5の値である。

【0061】

図7は、本発明の実施例1のストレージ装置11が保持するドライブ割当管理テーブル211の構成例を示す説明図である。

40

【0062】

ドライブ割当管理テーブル211は、プールアドレス毎にエントリを有する。各エントリは、プールID 8 1、プールアドレス8 2、RAIDグループID 8 3、ドライブID 8 4及びドライブアドレス8 5といった情報を格納する。以下、1つのプールアドレス(図7の説明において「対象プールアドレス」)を例に取る。

【0063】

プールID 8 1は、対象プールアドレスが属するプールのIDである。プールアドレス8 2は、対象プールアドレスである。RAIDグループID 8 3は、対象プールアドレスが示すデータ領域の基になっているRAIDグループのIDである。ドライブID 8 4は、

50

対象プールアドレスが示すデータ領域の基になっているドライブのIDである。ドライブアドレス85は、対象プールアドレスに対応したドライブアドレスである。

【0064】

<論理記憶階層>

図8は、本発明の実施例1のストレージ装置11によって管理される論理記憶階層の構成例を示す説明図である。

【0065】

VOL1000は、ホスト計算機30に提供される。また、コピー処理又は重複排除処理によって、VOL1000内の複数のスロットから1つのプールアドレスを指すことがあり、複数のVOLのスロットから一つのプールアドレスを指すこともある。図8の例では、異なる2つのスロット(VOLアドレス)1100及び1103が、同一のプールアドレス1101を指している。なお、VOL1000からプール1001の割当ては、プール割当管理テーブル210を基に管理される。また、プール1001からドライブアドレス空間1003(すなわちRAIDグループ1002を構成する複数のドライブ29が提供する複数のドライブアドレス空間)への割当ては、ドライブ割当管理テーブル211を基に管理される。

10

【0066】

<メモリ割当管理テーブル>

図9は、本発明の実施例1のストレージ装置11が保持するメモリ割当管理テーブル212の構成例を示す説明図である。

20

【0067】

メモリ割当管理テーブル212は、VOLアドレス(スロットを示すアドレス)毎にエントリを有する。各エントリは、VOL\_ID91、VOLアドレス92、バッファ(BF)アドレス93、圧縮後VOLアドレス94、キュー状態95及びBF転送状態96といった情報を格納する。以下、1つのVOLアドレス(図9の説明において「対象VOLアドレス」)を例に取る。

【0068】

VOL\_ID91は、対象VOLアドレスによって識別されるスロットが属するVOLのIDである。VOLアドレス92は、対象VOLアドレスである。BFアドレス93は、対象VOLアドレスを指定してライトされたデータの転送先BFアドレスを示す。圧縮後VOLアドレス94は、対象VOLアドレスを指定してライトされたデータの内、BFへの転送の対象外となったデータの転送先VOLアドレスを示す。キュー状態95は、対象VOLアドレスを指定してライトされたデータのドライブ29へのデータ格納が完了しているかを示す。図9では、キュー状態95の値のうち“Dirty”はドライブ29への格納が出来ていないことを、“Clean”はドライブ29への格納が済んでいることを表す。BF転送状態96は、対象VOLアドレスを指定してライトされたデータが圧縮されてBFへ転送されているか否かを示す。BFへの転送が完了している場合、BF転送状態96の値は“転送済み”となり、転送が行われていない場合は“無し”となる。

30

【0069】

図10は、本発明の実施例1のストレージ装置11におけるメモリ割当の構成例を示す図である。

40

【0070】

キャッシュ領域203は、VOLに対応した仮想的なアドレス空間である非圧縮データ格納領域204、及び、プールアドレスに対応した圧縮データ格納領域205をストレージコントローラ22へ提供している。ホスト計算機30からストレージコントローラ22へのライト命令によって、VOLアドレスに対応した非圧縮データ格納領域204が割当てられる。ストレージコントローラ22は、ライト命令と非同期でデータを圧縮すると、圧縮したデータを、バッファ領域202、又は、キャッシュ領域203内圧縮データ格納領域205に、プールアドレスに対応させて格納する。

【0071】

50

図 10 の例では、ライトされたデータが格納されているVOL内のスロット1100が、プールアドレスに対応したバッファ領域202上の領域1101を指している。VOLアドレスとプールアドレスの割当ては、プール割当管理テーブル210で管理される。また、バッファ領域202への割当てはメモリ割当管理テーブル212のBFアドレス93で、圧縮データ格納領域への割当てはメモリ割当管理テーブル212の圧縮後VOLアドレス94で、それぞれ管理される。

【0072】

バッファ領域202では、バッファ領域内のデータ量がパリティサイクルのサイズに達すると、プロセッサ24を介して非圧縮データ格納領域204とは対応しないパリティ1104が生成される。

【0073】

以下、本実施例で行われる処理の例を説明する。

【0074】

<リード処理>

図 11 は、本発明の実施例1のストレージ装置11が実行するリード処理を示すフローチャートである。

【0075】

リード処理は、ホスト計算機30からネットワーク31を介してストレージ装置11がリード命令を受けた場合に開始する。リード命令では、例えば、仮想ID（例えば、仮想VOL\_ID）、アドレス、及びデータサイズが指定される。

【0076】

S1201で、プロセッサ24は、リード命令から特定されるスロットの排他を確保する。なお、スロット排他確保時に他の処理がスロットの排他を確保している場合、プロセッサ24は、一定の時間待ってから、S1201を行う。

【0077】

S1202で、プロセッサ24は、リードデータがキャッシュ領域203に存在するか否かを判定する。S1202の判定結果が真の場合、S1204に進む。S1202の判定結果が偽の場合、プロセッサ24は、S1203で、RAIDグループからリードデータをバッファ領域202に転送する。なお、この際、プロセッサ24は、ホスト計算機30が指定したVOL\_IDとVOLアドレスから、プール割当管理テーブル210のプールID73、プールアドレス74及び圧縮後サイズ76を特定し、ドライブ割当管理テーブル211からドライブID84及びドライブアドレス85を参照し、データの格納場所及びデータサイズを特定する。

【0078】

S1204で、プロセッサ24はバッファ領域202上のリードデータが圧縮されているか否かを圧縮後サイズ76から判定し、圧縮済みのデータであればS1205において伸長し、圧縮データで無い場合はS1205をスキップする。

【0079】

S1206で、プロセッサ24はバッファ領域202上のリードデータをホスト計算機30に転送する。ホスト計算機30は、S1206のデータ転送が完了した時点でリード処理が終了したと認識する。

【0080】

その後、プロセッサ24は、S1205で、確保していたスロット排他を解除する。

【0081】

<ライト処理>

図 12 は、本発明の実施例1のストレージ装置11が実行するライト処理を示すフローチャートである。

【0082】

ライト処理は、ホスト計算機30からストレージ装置11がライト命令を受信した場合に開始する。なお、以下の説明では、例えば、ストレージコントローラ2\_22Aのプロセ

10

20

30

40

50

ッサ 24 をプロセッサ 24 A と記載するなど、ストレージコントローラ 2\_\_22 A 及びストレージコントローラ 2\_\_22 B に属するものをそれぞれ参照符号に付した「A」及び「B」によって区別する。

【0083】

ホスト計算機 30 からのライト命令には、割当て先アドレスが付随している。ストレージ装置 11 は、S1301 において割当て先アドレスが示すスロットの排他を確保する。なお、スロット排他確保と同時に、プロセッサ 24 A は、データのライト先とするキャッシュ領域 203 A のスロット領域を割当てる。

【0084】

S1302 で、プロセッサ 24 A は、ホスト計算機 30 に対してライト処理の準備ができたことを示す「Ready」を応答する。プロセッサ 24 A は、「Ready」を受け取ったホスト計算機 30 から、ライトデータを受ける。その後、S1303 でプロセッサ 24 はライト命令と同期して圧縮処理を実行する必要があるかを判定する。なお、プロセッサ 24 A の負荷、ストレージ装置 11 に対するライト量、及びライトデータのデータ長から、ストレージシステム 100 においてレスポンス性能を優先するケース 1 及びスループット性能を優先するケース 2 のいずれかへ分岐する。例えば、ストレージ装置 11 は、以下のような条件を保持し、プロセッサ 24 A は、ライト命令を受信すると、保持している条件に基づいてレスポンス性能及びスループット性能のいずれを優先するかを判定してもよい。

【0085】

<ケース 1> レスポンス優先

レスポンス性能を優先する条件として以下のものがある。例えば、以下の複数の条件のうちいずれか一つのみ、又は、複数の組合せに基づいて、レスポンス性能を優先するか否かが判定されてもよい。後述するスループット性能に関する条件についても同様である。

【0086】

(1) ストレージコントローラ 22 の(すなわちプロセッサ 24 の)負荷が所定の基準より低い

【0087】

(2) ライトデータを圧縮した場合の圧縮率が所定の基準より低くなることが予想される

【0088】

(3) 書き込み先のボリュームに圧縮データを格納できない

【0089】

ここで、上記(1)は、所定の基準近傍で判定結果が頻繁に切り替わると負荷の変動が不安定になるため、これを防ぐために多段階で基準を変動させてもよい。また、上記(1)は、例えばストレージ装置 11 に対する I/O 命令の量に基づいて判定されてもよい。例えば、単位時間当たりの I/O 命令の回数、又は、I/O 命令によって書き込み/読み出しが行われるデータ量が所定の基準より少ない場合に、負荷が低いと判定されてもよい。

【0090】

上記(2)は、例えば、ライトデータのサイズが所定の基準より小さい場合に、ライトデータの圧縮率が低い、すなわち圧縮によるデータ削減が見込めないと判定されてもよい。

上記(3)は、例えば、ライトデータの書き込み先の VOL に対応する VOL 管理テーブル 207 の VOL 属性 42 が“圧縮有効”でない場合に、書き込み先のボリュームに圧縮データを格納できないと判定されてもよい。

【0091】

例えばプロセッサ 24 A が低負荷であり、レスポンス性能を優先する場合、S1303 の判定において偽となる。この場合、プロセッサ 24 A は、S1306 において受け取ったライトデータを割当てたキャッシュ領域 203 A へ格納する。S1307 において、ストレージコントローラ 1\_\_22 A からストレージコントローラ 2\_\_22 B に対してキャッシュ領域 203 A に格納したライトデータを転送し、キャッシュ領域 203 B に格納することで二重化を行う。

10

20

30

40

50

## 【 0 0 9 2 】

S 1 3 0 8において、プロセッサ 2 4 Aは、メモリ割当管理テーブル 2 1 2を更新する。なお、本ケースにおいてライトデータは未だ圧縮されていない。このため、データのライト先として割当てられたスロットのVOLアドレスに対応するBFアドレス 9 3及び圧縮後VOLアドレス 9 4の値は無く、プロセッサ 2 4 Aは、キュー状態 9 5を“Dirty”に更新する。

## 【 0 0 9 3 】

次に、S 1 3 0 9において、ストレージ装置 1 1から、ネットワーク 3 1を介してホスト計算機 3 0に対してライト処理が完了したとして完了応答を返却する。完了応答を返却すると、S 1 3 1 0においてストレージ装置 1 1は確保していたスロットの排他を解放してライト処理を終了する。

10

## 【 0 0 9 4 】

<ケース 2>スループット優先

スループット性能を優先する条件として以下のものがある。

## 【 0 0 9 5 】

( 4 )ストレージコントローラ 2 2の(すなわちプロセッサ 2 4の)負荷が所定の基準より高い

## 【 0 0 9 6 】

( 5 )ライトデータを圧縮した場合の圧縮率が所定の基準より高くなることが予想される

## 【 0 0 9 7 】

ここで、上記( 4 )は、上記( 1 )と同様に、例えばストレージ装置 1 1に対するI/O命令の量に基づいて判定することができる。例えば、単位時間当たりのI/O命令の回数等が所定の基準より多い場合に、負荷が高いと判定されてもよい。

20

## 【 0 0 9 8 】

上記( 5 )は、例えば、ライトデータのサイズが所定の基準より大きい場合に、ライトデータの圧縮率が高い、すなわち圧縮によるデータ削減が見込まれると判定されてもよい。

## 【 0 0 9 9 】

例えばプロセッサ 2 4が高負荷であり、スループット性能を優先する場合、S 1 3 0 3の判定において真となる。この場合、プロセッサ 2 4 Aは、S 1 3 0 4において受け取ったライトデータをバッファ領域 2 0 2 Aへ転送する。次に、S 1 3 0 5で、プロセッサ 2 4 Aは、バッファ内のデータを圧縮する。

30

## 【 0 1 0 0 】

なお、S 1 3 0 4及びS 1 3 0 5において、ライトデータのバッファ領域 2 0 2 Aへの格納時に圧縮が行われても良い(すなわち、バッファ領域 2 0 2 Aへの格納前に圧縮が行われ、圧縮されたデータがバッファ領域 2 0 2 Aへ格納されても良い)し、バッファ領域 2 0 2 Aへの格納後にバッファ領域 2 0 2 A内で圧縮が行われても良い。いずれの場合も、最終的には、圧縮後のデータがバッファ領域 2 0 2 Aに格納される。

## 【 0 1 0 1 】

また、この圧縮は、バッファ領域 2 0 2 A以外の記憶領域(例えばプロセッサ 2 4 A内のメモリ)において行われてもよい。

40

## 【 0 1 0 2 】

ここで、圧縮は、ライトデータに対して行われる所定の処理の一例である。プロセッサ 2 4は、圧縮以外の処理、例えば、重複排除、暗号化又は冗長化等を行い、処理後のデータをバッファ領域 2 0 2 Aに格納してもよい。後述する図 1 4のS 1 4 1 1についても同様である。

## 【 0 1 0 3 】

次に、S 1 3 0 6において、プロセッサ 2 4 Aは、バッファ領域 2 0 2 A内の圧縮データを、割当てたキャッシュ領域 2 0 3 Aへ格納する。S 1 3 0 7において、ストレージコントローラ 1 \_\_ 2 2 Aからストレージコントローラ 2 \_\_ 2 2 Bに対してキャッシュ領域 2 0 3 Aに格納したライトデータを転送し、キャッシュ領域 2 0 3 Bに格納することで圧縮デ

50

ータの二重化を行う。

【0104】

S1308において、プロセッサ24Aは、メモリ割当管理テーブル212を更新する。なお、本ケースにおいてライトデータは圧縮されており、圧縮データに対してアドレスが割当てられる。このため、データのライト先として割当てられたスロットのVOLアドレスに対応する圧縮後VOLアドレス94が更新される。また、BFアドレス93の値は無く、プロセッサ24Aは、キュー状態95を“Dirty”に更新する。

【0105】

次に、S1309において、ストレージ装置11から、ネットワーク31を介してホスト計算機30に対してライト処理が完了したとして完了応答を返却する。完了応答を返却すると、S1310においてストレージ装置11は確保していたスロットの排他を解放してライト処理を終了する。

10

【0106】

< デステージ処理 >

図13は、本発明の実施例1のストレージ装置11が実行するデステージ処理を示すフローチャートである。

【0107】

デステージ処理は、ホスト計算機30からストレージ装置11へのライト命令が完了した後、非同期的に行われる。なお、デステージは、ライト命令が完了したことを契機として開始されても良いし、周期的に起動しても良いし、キャッシュ領域203の消費量などからライト量を判定して選択しても良い。

20

【0108】

デステージ処理が開始されると、ストレージ装置11は、S1401において、デステージ処理の対象領域がキャッシュ領域上の圧縮データ格納領域205に属しているか否かを判定する。判定が真の場合（すなわち対象領域が圧縮データ格納領域205に属している場合）はケース2-1、判定が偽の場合（すなわち対象領域が非圧縮データ格納領域204に属している場合）はケース1-1の処理が行われる。

【0109】

< ケース2-1 > 圧縮済データのデステージ

S1401の判定が真の場合、キャッシュ領域203内の圧縮データ格納領域205に対してデステージ処理（S1402～S1406）が行われる。S1402では、プロセッサ24Aは、圧縮データ格納領域205からデステージ処理を実行するデータを選択する。通常、パリティサイクル分のデータが並ぶデータ列（ストライプ列）が選択され、それに対してデステージが行われる。

30

【0110】

S1403で、プロセッサ24Aは、デステージするデータが属するスロットの排他を確保する。排他を確保した後、プロセッサ24Aは、S1404で対象のデータ列からパリティデータを生成する。S1405で、プロセッサ24Aは、対象のデータ列及び生成したパリティデータをドライブに書き出す。S1406において、プロセッサ24Aは、メモリ割当管理テーブル212を更新する。なお、本ケースにおいて、キュー状態95が“Clean”に更新される。S1407で、プロセッサ24Aは、デステージされた範囲のスロットの排他を解放し、処理を終了する。

40

【0111】

< ケース1-1 > 圧縮及びデステージ一括処理（デステージ中排他保持）

S1401の判定が偽の場合、キャッシュ領域203内の非圧縮データ格納領域204に対してデステージ処理（S1408～S1415）が行われる。S1408では、プロセッサ24Aは、非圧縮データ格納領域204に格納されているデータのうち、キュー状態95が“Dirty”であるスロットに属するデータから、デステージ処理を実行するデータを選択する。通常、パリティサイクル分のデータが並ぶデータ列（ストライプ列）が選択され、それに対してデステージが行われる。

50

## 【 0 1 1 2 】

S 1 4 0 9 で、プロセッサ 2 4 は、デステージするデータが属するスロットの排他を確保する。なお、図 1 3 に示すデステージ処理が、図 1 2 に示したライト処理の終了を契機として（すなわちライト処理の直後に）行われる場合には、S 1 3 1 0 及び S 1 4 0 9 を省略してもよい。

## 【 0 1 1 3 】

排他を確保した後、プロセッサ 2 4 A は、S 1 4 1 0 で対象のデータを読み出して、バッファ領域 2 0 2 へ転送する。なお転送の際、プロセッサ 2 4 は、メモリ割当管理テーブル 2 1 2 の B F アドレス 9 3 及び圧縮後 V O L アドレス 9 4 を割当てて。また、プロセッサ 2 4 A は、バッファ領域 2 0 2 への転送完了後、B F 転送状態 9 6 を“転送済”に更新する。なお、圧縮後 V O L アドレス 9 4 の割当ては、パリティサイクル分を割当てることが明らかかなため、あらかじめパリティサイクル分の領域を割当ててすることで、マッピング情報の更新回数を削減できる。

10

## 【 0 1 1 4 】

S 1 4 1 1 で、プロセッサ 2 4 A は、転送したデータを圧縮する。なお、圧縮処理はバッファ転送時に行っても良い（すなわち、バッファ領域 2 0 2 への格納前に圧縮が行われ、圧縮されたデータがバッファ領域 2 0 2 へ格納されても良い）し、転送後バッファ内で行っても良い。

## 【 0 1 1 5 】

S 1 4 1 2 において、プロセッサ 2 4 A は、バッファ内の圧縮データの量を判定する。圧縮データ量がパリティサイクル分よりも小さい場合、プロセッサ 2 4 は、S 1 4 0 8 に戻ってデステージするデータを追加で選択する。パリティサイクル分のデータがバッファ領域 2 0 2 内に溜まった場合、S 1 4 1 2 の判定を真として S 1 4 1 3 に進む。なお、圧縮データサイズは可変長であるため、バッファ領域 2 0 2 内のデータが必ずしもパリティサイクル分揃うとは限らないことから、パリティサイクルを超える前に S 1 4 1 3 へ処理を進めることもありえる。

20

## 【 0 1 1 6 】

S 1 4 1 3 において、プロセッサ 2 4 A は、バッファ領域 2 0 2 内の圧縮データからパリティデータを生成する。S 1 4 1 4 で、プロセッサ 2 4 A は、対象のデータ列及び生成したパリティデータを、R A I D グループを構成するドライブ 2 9 に書き出す。S 1 4 1 5 において、プロセッサ 2 4 A は、メモリ割当管理テーブル 2 1 2 の更新を確定する。なお、本ケースにおいて、キュー状態 9 5 が“Clean”に更新される。S 1 4 0 7 で、プロセッサ 2 4 A は、デステージされた範囲のスロットの排他を解放し、処理を終了する。

30

## 【 0 1 1 7 】

上記の例では、S 1 4 1 2 において、バッファ内の圧縮データの量がパリティサイクルのデータ量に達したか否かが判定されている。しかし、ドライブ 2 9 が R A I D を構成するか否かにかかわらず、所定の量のデータをまとめてドライブ 2 9 に格納する場合には、プロセッサ 2 4 A は、S 1 4 1 2 においてバッファ内の圧縮データの量が当該所定の量に達したか否かを判定する。本実施例の S 1 4 1 2 におけるパリティサイクルのデータ量は、上記の所定のデータ量の一例である。

40

## 【 0 1 1 8 】

なお、プロセッサ 2 4 A は、S 1 4 0 1 の判定が偽の場合であっても、S 1 4 0 8 ~ S 1 4 1 5 ではなく、S 1 4 0 2 ~ S 1 4 0 6 を実行する場合がある。例えば、ライトデータの書き込み先の V O L 属性 4 2 が圧縮有効でないために、図 1 2 の S 1 3 0 3 の判定が偽であった場合、非圧縮データがキャッシュ領域 2 0 3 A に格納されている。この場合、S 1 4 0 1 の判定は偽となるが、データの圧縮は行わないため、S 1 4 0 2 ~ S 1 4 0 6 が実行される。

## 【 0 1 1 9 】

上記の例では、スループット性能が優先される場合に、ライト処理時には圧縮後のデータがキャッシュ領域 2 0 3 で二重化された時点でホスト計算機 3 0 に応答が返され、デステ

50

ージ処理ではデータの圧縮が不要となる。これによって、レスポンス性能は低下するが、デステージ処理の際のキャッシュアクセスが削減されるため、スループット性能が向上する。このような処理は一例であり、スループット性能が優先される場合に、ライト処理の際にさらに多くの処理が行われてもよい。

【0120】

例えば、プロセッサ24Aは、S1303（図12）の判定が真である場合に、S1304～S1308を実行し、続いて、S1412、S1404～S1406（図13）と同様の処理を実行し、その後にS1309、S1310を実行してもよい。すなわち、ライト命令に対して圧縮処理及びデステージまで一括して行われるため、レスポンス性能はさらに低下するが、スループット性能は向上する。

10

【0121】

この場合も、S1303（図12）の判定が偽であるときの処理は、上記の図12及び図13を示して説明した通りである。すなわち、プロセッサ24Aは、S1304～S1305を実行せずに、S1306～S1310を実行する。さらに、プロセッサ24Aは、S1408～S1415及びS1407を実行する。

【0122】

上記の例によれば、デステージが開始されるとスロットの排他が確保され（S1409）、その後、データのドライブ29への転送が終了して（S1414）マッピング情報が更新される（S1415）まで、スロットの排他が確保される（S1407）。このように長時間排他を確保することによって、必要なI/O命令が実行できないといったトラブルが発生する可能性がある。このようなトラブルを回避するために、ケース1-1における排他手順を変更した実施例として、以下のケース1-2を示す。

20

【0123】

図14は、本発明の実施例1のストレージ装置11が実行する、排他手順を変更したデステージ処理を示すフローチャートである。

【0124】

<ケース1-2> 圧縮及びデステージ一括処理（デステージ中排他解放）

S1501において、ストレージ装置11は、図13のS1401と同様の判定を行う。S1501の判定が真の場合、キャッシュ領域203内の圧縮データ格納領域205に対してデステージ処理（S1502～S1507）が行われる。これらの処理は、図13のS1402～S1407と同様であるため、説明を省略する。

30

【0125】

S1501の判定が偽の場合、キャッシュ領域203内の非圧縮データ格納領域204に対してデステージ処理が行われる（S1508～S1519）。S1508では、プロセッサ24は、非圧縮データ格納領域204に格納されているデータのうち、キュー状態95が“Dirty”であるスロットに属するデータからデステージ処理を実行するデータを選択する。通常、パリティサイクル分のデータが並ぶデータ列（ストライプ列）が選択され、それに対してデステージが行われる。

【0126】

先述のケース1-1ではデステージ処理が完了するまでデステージ対象となるスロット範囲が保持されている。しかし、圧縮後のデータサイズがパリティサイクル分に達する広範囲の排他を保持し続けると、ホスト計算機30からのライト命令が排他範囲に生じることによってデステージ待ちを生じる可能性が高くなる。そこで、プロセッサ24は、S1509でデステージするデータが属するスロットの排他を確保した後、S1510のバッファ転送及びS1511の圧縮処理を行う。そして、プロセッサ24は、圧縮処理が完了した後のS1512でメモリ割当管理テーブル212のBF転送状態96を“転送済”に更新する。更新が完了すると、プロセッサ24は、S1513においてスロット排他を解放する。

40

【0127】

以後、プロセッサ24は、S1514のドライブ転送可否の判定、S1515のパリティ

50

生成、S 1 5 1 6 のドライブ転送を、それぞれケース 1 - 1 の S 1 4 1 2、S 1 4 1 3 及び S 1 4 1 4 と同様に行う。

【 0 1 2 8 】

S 1 5 1 7 において、プロセッサ 2 4 は、デステージ範囲のスロット排他を再度確保し、S 1 5 1 8 でメモリ割当管理テーブル 2 1 2 のキュー状態 9 5 を “ Clean ” に更新する。

【 0 1 2 9 】

なお、S 1 5 1 7 までの間に、上記のデステージ範囲のスロットに対してホスト計算機 3 0 からの更新ライトが発生した場合、プロセッサ 2 4 は、S 1 3 0 8 においてメモリ割当管理テーブル 2 1 2 の B F 転送状態 9 6 を “ 無し ” に更新する。この場合、S 1 5 1 8 でプロセッサ 2 4 がキュー状態 9 5 を更新する際に B F 転送状態 9 6 が切り替わったことを判定することによって、更新ライトが発生したことに気づくことができる。

10

【 0 1 3 0 】

なお、更新ライトの発生に気づいた（すなわち S 1 5 1 2 で “ 転送済 ” に更新した B F 転送状態 9 6 が S 1 5 1 7 の時点で “ 無し ” となっていた）場合、プロセッサ 2 4 は、処理をやり直すか又は対象箇所のマッピング情報更新をスキップする。具体的には、プロセッサ 2 4 は、S 1 5 1 8 に進まずに S 1 5 0 8 に戻り、更新ライトが行われたスロットを対象とするデステージ処理をやり直してもよい。あるいは、プロセッサは、そのまま S 1 5 0 8 に進み、更新ライトが行われたスロットのキュー状態 9 5 を “ Clean ” に更新せずに、S 1 5 1 9 に進んでもよい。その場合、当該スロットは次回以降のデステージ処理の対象となる。

20

【 0 1 3 1 】

最後に S 1 5 1 9 で、プロセッサ 2 4 は、デステージされた範囲のスロットの排他を解放し、処理を終了する。

【 0 1 3 2 】

以上の本発明の実施例によれば、キャッシュ領域に格納されたデータをデステージする際に、圧縮処理から記憶デバイス（ドライブ）への格納までを一括で行うことによって、圧縮データの二重化処理が省略される。キャッシュ領域における圧縮データの二重化が不要になることで、キャッシュアクセス量を削減し、データ書き込みの最大速度が向上できる。

【 0 1 3 3 】

また、記憶デバイスへの圧縮データの格納が完了するまでキャッシュメモリ上に圧縮前のデータを二重化して保持することによって、圧縮処理及び記憶デバイスへの格納などの処理中に装置障害が発生してもデータを保護することができる。ストレージ装置が圧縮以外の処理（例えば重複排除、暗号化又は冗長化等）を行う場合にも、同様の効果が得られる。

30

【 0 1 3 4 】

また、デステージの際に圧縮処理を行う場合、例えばパリティサイクル等の所定の大きさの領域を予め割り当てることができるため、マッピング情報の更新回数を削減することができる。

【 0 1 3 5 】

また、本発明の実施例によれば、ストレージ装置は、所定の条件に基づいてレスポンス性能及びスループット性能のいずれを優先するかを判定する。そして、レスポンス性能を優先する場合にはキャッシュメモリ上に圧縮前のデータを二重化して保持したところでホストに応答する。これによって、レスポンス性能が向上する。一方、スループット性能を優先する場合には圧縮を行い、圧縮後のデータを二重化して保持したところでホストに応答する。これによってレスポンス性能は低下するが、デステージの際のキャッシュアクセス量が削減されるため、スループット性能は向上する。

40

【 0 1 3 6 】

例えば、I/O 命令の量、予想される圧縮率又は書き込み先のボリュームの属性などに基づいてレスポンス性能又はスループット性能のいずれを優先するかを判定することによって、状況に応じて最適な性能を実現することができる。

【 0 1 3 7 】

50

また、キャッシュ領域に格納された圧縮前のデータをデステージする場合に、当該データをキャッシュ領域から読み出すときから記憶デバイスへの圧縮後のデータの格納が完了し、キュー状態を“Clean”に変更するまで(S1409~S1415、S1407)、当該データの領域の排他を確保してもよい。これによって、まだデステージされていないデータがデステージされたと誤って判定することが防止される。

【0138】

あるいは、当該データを読み出して、圧縮を行い、バッファ領域に転送した時点で排他を一旦解除してもよい(S1513)。これによって、排他が確保される時間が短縮し、必要なI/Oが実行できないというトラブルが軽減される。この場合、排他を一旦解除(S1513)してから当該データの記憶デバイスへの転送が終了(S1516)するまでの間に新たな書き込みが行われると、そのことが記録される(すなわちBF転送状態が“転送済み”から“なし”に更新される)。これによって、まだデステージされていないデータがデステージされたと誤って判定することが防止される。

10

【0139】

特許請求の範囲に記載したもののほか、本発明の観点の代表的なものとして、次のものが挙げられる。

(1) 第1のストレージ制御部と、第2のストレージ制御部と、少なくとも前記第1のストレージ制御部に接続され、不揮発性の記憶媒体を有するストレージドライブと、を有するストレージシステムであって、

前記第1のストレージ制御部と前記第2のストレージ制御部とは、それぞれ、データを格納するキャッシュ領域と、データを格納するバッファ領域と、を有しており、前記キャッシュ領域に格納されたデータを互いのキャッシュ領域にも格納して二重化を行い、

20

前記第1のストレージ制御部は、ホスト計算機からデータの書き込み命令を受信すると、前記書き込み命令の対象のデータを、前記第1のストレージ制御部の前記キャッシュ領域である第1のキャッシュ領域に格納するとともに、前記データを前記第2のストレージ制御部の前記キャッシュ領域である第2のキャッシュ領域に格納して二重化を行い、前記二重化が完了したら、前記ホスト計算機に、前記データの書き込みの終了を示す応答を送信し、

前記書き込み命令の対象のデータに所定の処理を行い前記バッファ領域に格納し、前記バッファ領域に格納したデータを読み出して前記ストレージドライブに送信することを特徴とするストレージシステム。

30

(2) 上記(1)に記載のストレージシステムであって、

前記第1のストレージ制御部は、前記第1のキャッシュ領域から読み出したデータに前記所定の処理を行い、前記所定の処理後のデータを前記バッファ領域に格納することを特徴とするストレージシステム。

(3) 上記(1)に記載のストレージシステムであって、

前記所定の処理は、前記第1のキャッシュ領域から読み出したデータの圧縮、重複排除、暗号化又は冗長化のいずれかであることを特徴とするストレージシステム。

(4) 上記(1)に記載のストレージシステムであって、

前記第1のストレージ制御部に接続された複数の前記ストレージドライブを有し、前記第1のストレージ制御部は、前記バッファ領域に格納したデータの量がパリティを生成するための所定のデータ量に達した場合、前記バッファ領域から読み出したデータに基づいてパリティを作成し、前記バッファ領域から読み出したデータ及び前記パリティを前記複数のストレージドライブに送信することを特徴とするストレージシステム。

40

(5) 上記(2)に記載のストレージシステムであって、

前記第1のストレージ制御部は、前記ストレージシステムにおいてレスポンス性能及びスループット性能のいずれが優先されるかを判定するための所定の条件を保持して、前記所定の条件に基づいて、レスポンス性能及びスループット性能のいずれが優先されるかを判定し、

レスポンス性能が優先される場合、前記データを前記第1及び第2のキャッシュ領域に格

50

納してから、前記所定の処理を行い、

スループット性能が優先される場合、前記所定の処理を行ったデータを前記第 1 及び第 2 のキャッシュ領域に格納することを特徴とするストレージシステム。

(6) 上記(5)に記載のストレージシステムであって、

前記第 1 のストレージ制御部は、前記第 1 のストレージ制御部の処理の負荷が所定の基準より低い場合に、レスポンス性能が優先されると判定することを特徴とするストレージシステム。

(7) 上記(5)に記載のストレージシステムであって、

前記所定の処理は、前記データの圧縮であり、

前記第 1 のストレージ制御部は、前記データの圧縮率が所定の基準より低くなることが予測される場合、又は、前記データの書き込み対象として指定されたボリュームに圧縮データを格納することができない場合に、レスポンス性能が優先されると判定することを特徴とするストレージシステム。

10

(8) 上記(5)に記載のストレージシステムであって、

前記第 1 のストレージ制御部は、

前記データが書き込まれるボリュームの管理単位領域ごとに、当該管理単位領域に書き込まれたデータが前記ストレージドライブに格納されたかを示すキュー状態を保持し、

前記ホスト計算機から前記データの書き込み命令を受信すると、前記データの書き込み対象である前記管理単位領域の排他を確保した後に、前記第 1 のキャッシュ領域に前記データを格納し、

20

前記ホスト計算機に、前記データの書き込みの終了を示す応答を送信した後に、前記データの書き込み対象である前記管理単位領域の排他を解除し、

前記管理単位領域のうち、前記キュー状態が、書き込まれたデータが前記ストレージドライブに格納されていないことを示す前記管理単位領域の排他を確保した後に、当該管理単位領域に書き込まれたデータを前記第 1 のキャッシュ領域から読み出して、前記所定の処理後のデータを前記バッファ領域に格納し、

前記バッファ領域から読み出した前記所定の処理後のデータの前記ストレージドライブへの格納が終了すると、前記キュー状態を、書き込まれたデータが前記ストレージドライブに格納されたことを示す値に更新し、その後、当該管理単位領域の排他を解除することを特徴とするストレージシステム。

30

(9) 上記(5)に記載のストレージシステムであって、

前記第 1 のストレージ制御部は、

前記データが書き込まれるボリュームの管理単位領域ごとに、当該管理単位領域に書き込まれたデータが前記ストレージドライブに格納されたかを示すキュー状態、及び、当該管理単位領域に書き込まれたデータが前記バッファ領域に格納されたかを示すバッファ転送状態を保持し、

前記ホスト計算機から前記データの書き込み命令を受信すると、前記データの書き込み対象である前記管理単位領域の排他を確保した後に、前記第 1 のキャッシュ領域に前記データを格納し、

前記ホスト計算機に、前記データの書き込みの終了を示す応答を送信した後に、前記データの書き込み対象である前記管理単位領域の排他を解除し、

40

前記管理単位領域のうち、前記キュー状態が、書き込まれたデータが前記ストレージドライブに格納されていないことを示す前記管理単位領域の排他を確保した後に、当該管理単位領域に書き込まれたデータを前記第 1 のキャッシュ領域から読み出して、前記所定の処理後のデータを前記バッファ領域に格納し、

当該管理単位領域の前記バッファ転送状態を、格納されたデータが前記バッファ領域に格納されたことを示す値に更新した後に、当該管理単位領域の排他を解除し、

当該管理単位領域の排他が解除されている間に、当該管理単位領域に対するデータの書き込みを行った場合、当該管理単位領域の前記バッファ転送状態を、書き込まれたデータが前記バッファ領域に格納されていないことを示す値に更新し、

50

前記バッファ領域から読み出した前記所定の処理後のデータが前記ストレージドライブに格納された後に、当該管理単位領域の排他を確保し、

当該管理単位領域の前記バッファ転送状態が、書き込まれたデータが前記バッファ領域に格納されていることを示す場合、前記キュー状態を、書き込まれたデータが前記ストレージドライブに格納されたことを示す値に更新した後に、当該管理単位領域の排他を解除することを特徴とするストレージシステム。

(10) 上記(2)に記載のストレージシステムであって、

前記第1のストレージ制御部は、データを格納する第3のキャッシュ領域をさらに有し、

前記第2のストレージ制御部は、データを格納する第4のキャッシュ領域をさらに有し、

前記第1のストレージ制御部は、前記所定の条件に基づいて、スループット性能が優先されると判定した場合、前記データに前記所定の処理を行い、前記所定の処理後のデータを前記第3のキャッシュ領域に格納して、前記所定の処理後のデータを前記第2のストレージ制御部に送信し、

前記第2のストレージ制御部は、前記第1のストレージ制御部から受信した前記所定の処理後のデータを前記第4のキャッシュ領域に格納して二重化を行い、

前記第1のストレージ制御部は、

前記第2のストレージ制御部による前記第4のキャッシュ領域への前記所定の処理後のデータの格納が終了すると、前記ホスト計算機に、前記データの書き込みの終了を示す応答を送信し、

前記第3のキャッシュ領域に格納したデータを読み出して前記ストレージドライブに送信することを特徴とするストレージシステム。

【0140】

なお、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施例は本発明のより良い理解のために詳細に説明したのであり、必ずしも説明の全ての構成を備えるものに限定されるものではない。

【0141】

また、上記の各構成、機能、処理部、処理手段等は、それらの一部又は全部を、例えば集積回路で設計する等によってハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによってソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、不揮発性半導体メモリ、ハードディスクドライブ、SSD(Solid State Drive)等の記憶デバイス、または、ICカード、SDカード、DVD等の計算機読み取り可能な非一時的データ記憶媒体に格納することができる。

【0142】

また、制御線及び情報線は説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線及び情報線を示しているとは限らない。実際にはほとんど全ての構成が相互に接続されていると考えてもよい。

【符号の説明】

【0143】

100 ストレージシステム

11 ストレージ装置

22、22A、22B ストレージコントローラ

202 バッファ領域

203、203A、203B キャッシュ領域

204 非圧縮データ格納領域

205 圧縮データ格納領域

29 ドライブ

30 ホスト計算機

31 ネットワーク

10

20

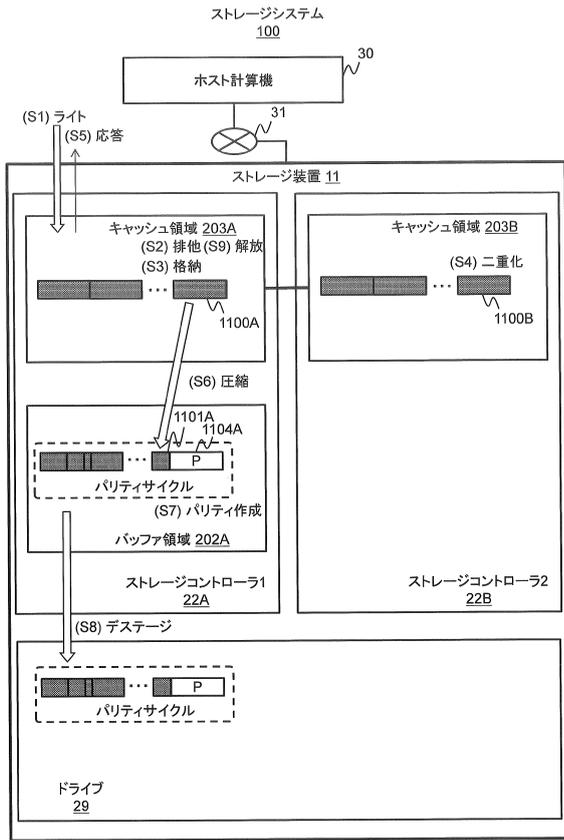
30

40

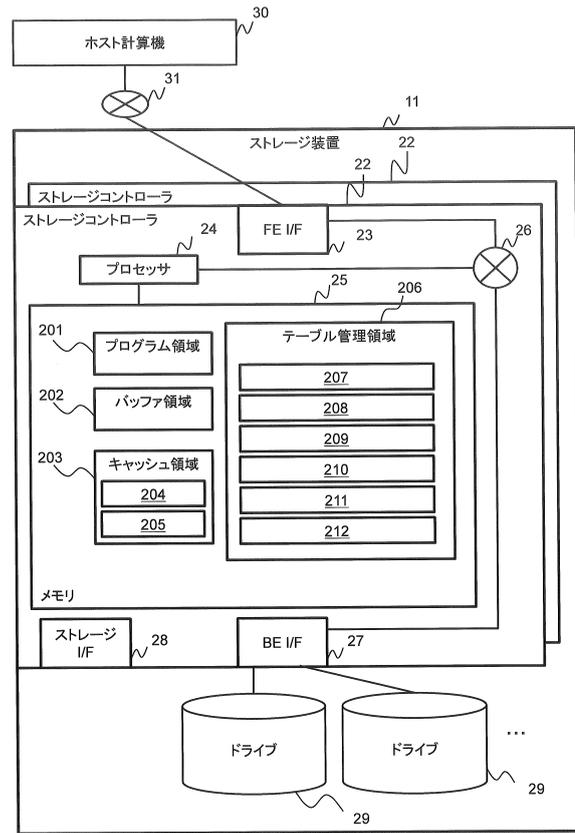
50

【図面】

【図 1】



【図 2】



10

20

【図 3】

ボリューム管理テーブル 207

VOL ID	VOL属性	VOL容量	プールID
0	シンプロビジョニング	100GB	0
10	圧縮有効	200GB	0
20	通常VOL	500GB	1
...	...	...	...

【図 4】

プール構成管理テーブル 208

プールID	RAIDグループID	プール容量	プール使用容量
0	0	10TB	5TB
...	...	...	...

30

40

50

【図5】

RAID構成管理テーブル  
209

RAIDグループID	RAIDレベル	ドライブID	ドライブ種別	容量	使用容量
0	RAID5	0 1 2 3	HDD	5TB	3TB
...	...	...	...	...	...

【図6】

プール割当管理テーブル  
210

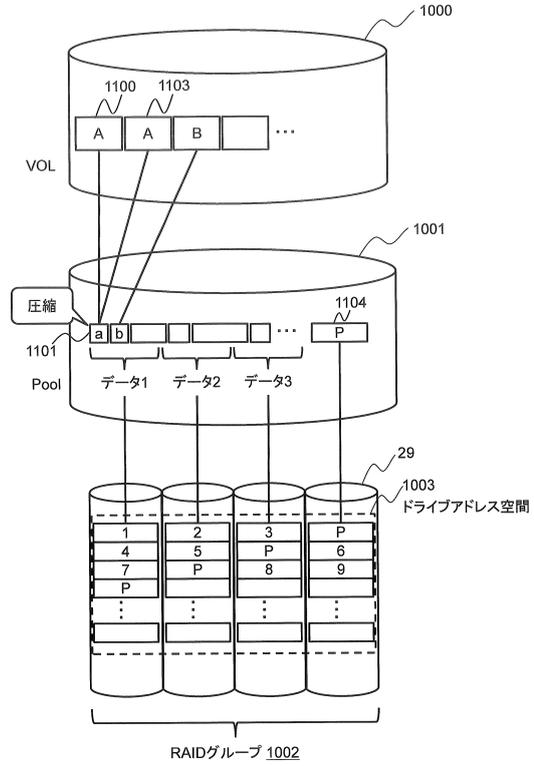
VOL ID	VOL アドレス	プールID	プール アドレス	圧縮前 サイズ	圧縮後 サイズ	圧縮率
0	100	0	10	8KB	4KB	1/2
0	200	0	10000	8KB	2KB	1/4
...	...	...	...	...	...	...

【図7】

ドライブ割当管理テーブル  
211

プールID	プールアドレス	RAIDグループID	ドライブID	ドライブアドレス
0	10	0	0	100
0	10000	1	5	1000
...	...	...	...	...

【図8】



10

20

30

40

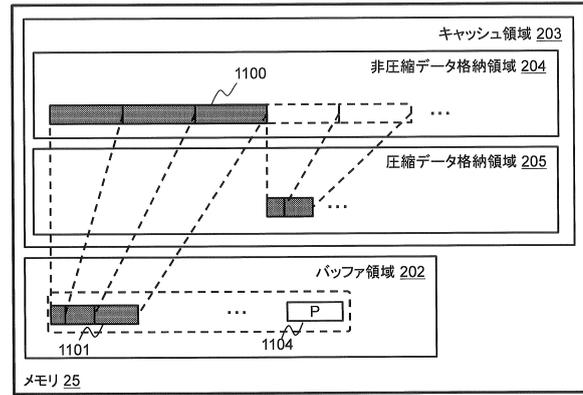
50

【図 9】

メモリ割当管理テーブル  
212

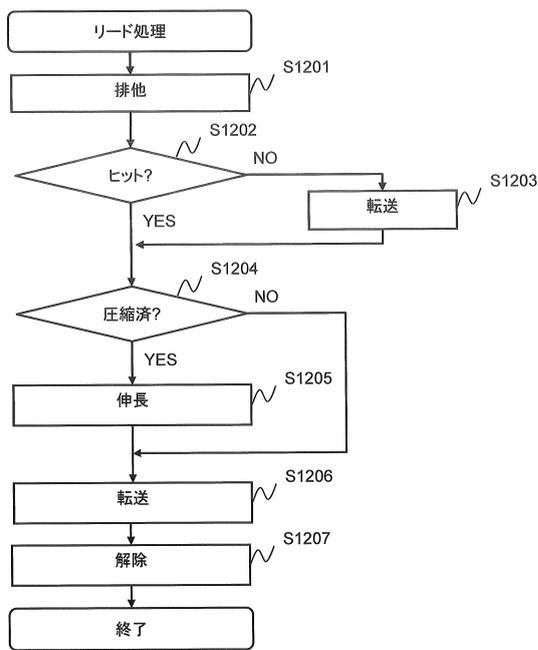
VOL ID	VOLアドレス	BFアドレス	圧縮後VOL アドレス	キュー状態	BF転送状態
0	100	-	-	Dirty	無し
0	200	50	1200	Dirty	転送済
0	300	-	1500	Dirty	無し
0	400	-	1000	Clean	転送済
...	...	...	...	...	...

【図 10】

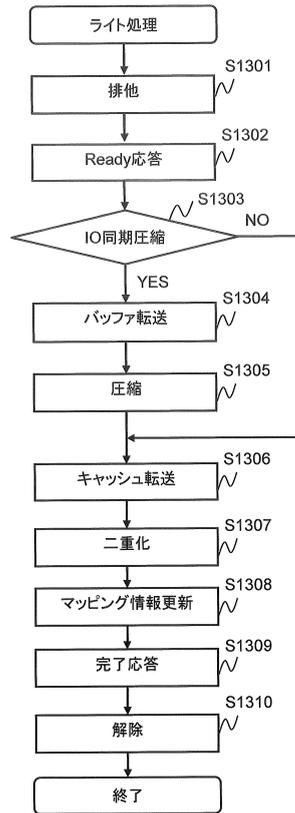


10

【図 11】



【図 12】



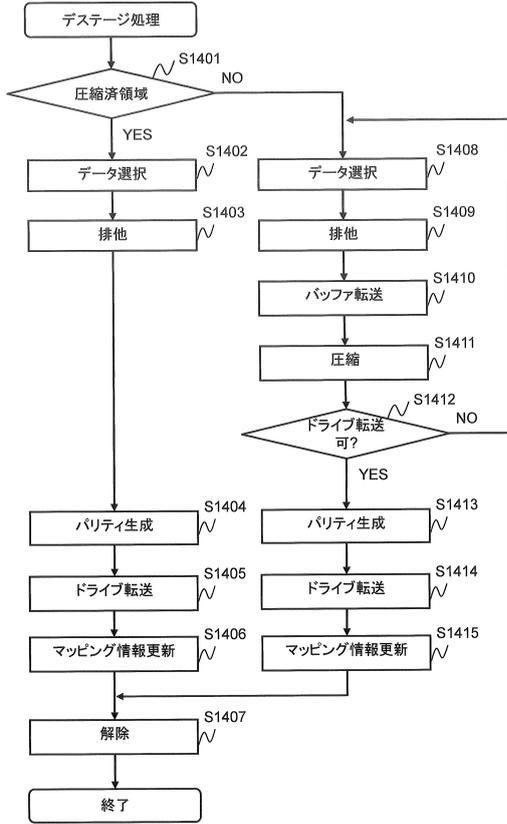
20

30

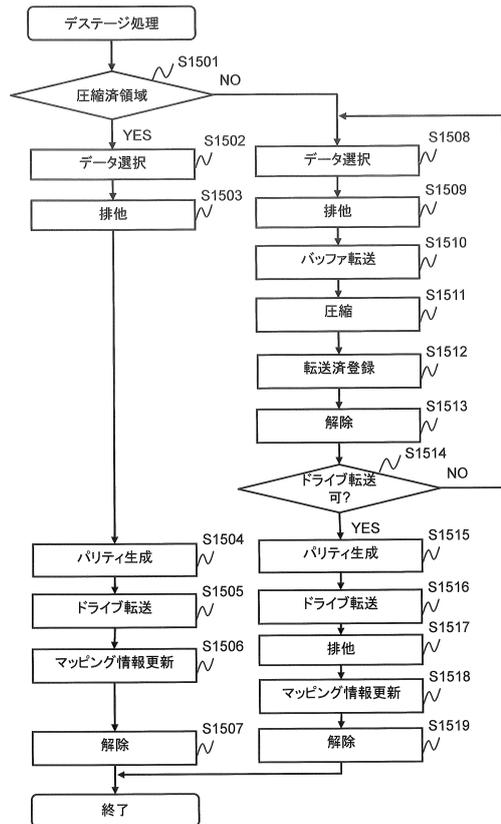
40

50

【図 1 3】



【図 1 4】



10

20

30

40

50

フロントページの続き

(51)国際特許分類	F I			
<b>G 0 6 F 16/172(2019.01)</b>	G 0 6 F	12/0866	1 0 0	
	G 0 6 F	12/0868	1 0 5	
	G 0 6 F	16/185		
	G 0 6 F	16/172		

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 松平 英

(56)参考文献 国際公開第2017/056219(WO, A1)  
 特開2005-157815(JP, A)  
 特開2010-224845(JP, A)  
 特表2016-510440(JP, A)  
 国際公開第2015/181902(WO, A1)  
 特開2004-318484(JP, A)  
 特開2007-293651(JP, A)  
 特開2009-048497(JP, A)  
 特表2015-517697(JP, A)

(58)調査した分野 (Int.Cl., DB名)  
 G 0 6 F 3 / 0 0  
 3 / 0 6 - 3 / 0 8  
 3 / 1 8  
 1 1 / 1 6 - 1 1 / 2 0  
 1 2 / 0 8 - 1 2 / 1 2 8  
 1 3 / 1 0 - 1 3 / 1 4  
 1 3 / 2 0 - 1 3 / 4 2  
 1 6 / 0 0 - 1 6 / 9 5 8  
 1 7 / 3 0