(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6802209号
(P6802209)

(45)発行日　令和2年12月16日(2020.12.16)　(24)登録日　令和2年11月30日(2020.11.30)

(51)Int.Cl.		F I		
G06F 3/06	(2006.01)	G06F	3/06	302A
G06F 16/185	(2019.01)	G06F	3/06	304E
G06F 16/172	(2019.01)	G06F	3/06	305C
G06F 16/174	(2019.01)	G06F	16/185	
G06F 12/0866	(2016.01)	G06F	16/172	

請求項の数 6　(全 23 頁)　最終頁に続く

(21)出願番号　特願2018-60662(P2018-60662)
(22)出願日　平成30年3月27日(2018.3.27)
(65)公開番号　特開2019-174994(P2019-174994A)
(43)公開日　令和1年10月10日(2019.10.10)
審査請求日　平成31年2月5日(2019.2.5)

(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

最終頁に続く

(54)【発明の名称】ストレージシステム

(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、それぞれ、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のデータに所定の処理を行い、前記所定の処理を行った前記書き込み命令の対象のデータを前記ストレージドライブに送信して格納させ、
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、前記データを前記二重化により前記第１及び第２のキャッシュ領域に格納して前記データの書き込みの終了を示す応答を送信してから、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い、前記所定の処理を行ったデータを前記二重化により前記第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。
【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。
【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。
【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、
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第１のバッファ領域に格納し、
前記第１の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。
【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第１の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、
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第１の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第１の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第１の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、
前記第１の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第１の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。
【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。
【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。
【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。
【０００３】
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。

【０００４】

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。

【０００５】

例えば、非特許文献１には、「方式の違いによってＩＯＰＳやレイテンシーに関する特性が異なっており、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ＩＯＰＳ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ＩＯＰＳ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ＩＯＰＳ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ＩＯＰＳ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。

【０００６】

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】

【０００７】

【特許文献１】米国特許出願公開第２００９／０１４４４９６号明細書

【非特許文献】

【０００８】

【非特許文献１】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，２０１６年１１月２８日，ｐ．５１－５９

【発明の概要】

【発明が解決しようとする課題】

【０００９】

データ書き込みにおいてＲＡＩＤ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。

【００１０】

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必要があるため、圧縮処理の時間だけ応答速度が遅くなる。

【００１１】

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、それぞれ、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のデータに所定の処理を行い、前記所定の処理を行った前記書き込み命令の対象のデータを前記ストレージドライブに送信して格納させ、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、前記データを前記二重化により前記第１及び第２のキャッシュ領域に格納して前記データの書き込みの終了を示す応答を送信してから、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い、前記所定の処理を行ったデータを前記二重化により前記第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。
【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。
【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。
【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。
【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。
【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。
【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。
【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】
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。

【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。

【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）。

【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。

【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で使用され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】
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（ライト命令に応答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（リード命令に応答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管理テーブル２１２を格納する。
【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。
【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＿ＩＤ４１、ＶＯＬ属性４２、ＶＯＬ容量４３及びプールＩＤ４４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。
【００５０】
ＶＯＬ＿ＩＤ４１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ＩＤ４４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。
【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ＩＤ５１、ＲＡＩＤグループＩＤ５２、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】
プールＩＤ５１は、対象プールのＩＤである。ＲＡＩＤグループＩＤ５２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５３は、対象プールの容量を示す。プール使用容量５４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ＩＤ６１、ＲＡＩＤレベル６２、ドライブＩＤ６３、ドライブ種別６４、容量６５及び使用容量６６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。
【００５７】
ＲＡＩＤグループＩＤ６１は、対象ＲＡＩＤグループのＩＤである。ＲＡＩＤレベル６２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ＩＤ６３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６４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６５は、対象ＲＡＩＤグループの容量を示す。使用容量６６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】
＜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＿ＩＤ７１、ＶＯＬアドレス７２、プールＩＤ７３、プールアドレス７４、圧縮前サイズ７５、圧縮後サイズ７６、及び圧縮率７７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】
ＶＯＬ＿ＩＤ７１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７２は、対象ＶＯＬアドレスである。プールＩＤ７３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７４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７５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７６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７７は、圧縮後サイズ７６／圧縮前サイズ７５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。
【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ＩＤ８１、プールアドレス８２、ＲＡＩＤグループＩＤ８３、ドライブＩＤ８４及びドライブアドレス８５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ＩＤ８１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

８２は、対象プールアドレスである。ＲＡＩＤグループＩＤ８３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ＩＤ８４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。

【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。

【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。

【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。

【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】
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のプロセッサ２４をプロセッサ２４Ａと記載するなど、ストレージコントローラ２＿２２Ａ及びストレージコントローラ２＿２２Ｂ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２４Ａは、データのライト先とするキャッシュ領域２０３Ａのスロット領域を割当てる。

【００８４】

Ｓ１３０２で、プロセッサ２４Ａ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２４Ａ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２４Ａ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２４Ａ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ＩＯ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ＩＯ命令の回数、又は、ＩＯ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２４Ａ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２４Ａ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２０３Ａへ格納する。Ｓ１３０７において、ス

トレージコントローラ１＿２２Ａからストレージコントローラ２＿２２Ｂに対してキャッシュ領域２０３Ａに格納したライトデータを転送し、キャッシュ領域２０３Ｂに格納することで二重化を行う。

【００９２】

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９３及び圧縮後ＶＯＬアドレス９４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。

【００９３】

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。

【００９４】

＜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。

【００９５】

（４）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高い

【００９６】

（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

【００９７】

ここで、上記（４）は、上記（１）と同様に、例えばストレージ装置１１に対するＩＯ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ＩＯ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。

【００９８】

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。

【００９９】

例えばプロセッサ２４が高負荷であり、スループット性能を優先する場合、Ｓ１３０３の判定において真となる。この場合、プロセッサ２４Ａは、Ｓ１３０４において受け取ったライトデータをバッファ領域２０２Ａへ転送する。次に、Ｓ１３０５で、プロセッサ２４Ａは、バッファ内のデータを圧縮する。

【０１００】

なお、Ｓ１３０４及びＳ１３０５において、ライトデータのバッファ領域２０２Ａ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２０２Ａ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２０２Ａへ格納されても良い）し、バッファ領域２０２Ａへの格納後にバッファ領域２０２Ａ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２０２Ａに格納される。

【０１０１】

また、この圧縮は、バッファ領域２０２Ａ以外の記憶領域（例えばプロセッサ２４Ａ内のメモリ）において行われてもよい。

【０１０２】

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２４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２０２Ａに格納してもよい。後述する図１４のＳ１４１１についても同様である。

【０１０３】
次に、Ｓ１３０６において、プロセッサ２４Ａは、バッファ領域２０２Ａ内の圧縮データを、割当てたキャッシュ領域２０３Ａへ格納する。Ｓ１３０７において、ストレージコントローラ１＿２２Ａからストレージコントローラ２＿２２Ｂに対してキャッシュ領域２０３Ａに格納したライトデータを転送し、キャッシュ領域２０３Ｂに格納することで圧縮デ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。
【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。
【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。
【０１１０】
Ｓ１４０３で、プロセッサ２４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。
【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。

【０１１２】

Ｓ１４０９で、プロセッサ２４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１３に示すデステージ処理が、図１２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、Ｓ１３１０及びＳ１４０９を省略してもよい。

【０１１３】

排他を確保した後、プロセッサ２４Ａは、Ｓ１４１０で対象のデータを読み出して、バッファ領域２０２へ転送する。なお転送の際、プロセッサ２４は、メモリ割当管理テーブル２１２のＢＦアドレス９３及び圧縮後ＶＯＬアドレス９４を割当てる。また、プロセッサ２４Ａは、バッファ領域２０２への転送完了後、ＢＦ転送状態９６を“転送済”に更新する。なお、圧縮後ＶＯＬアドレス９４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なため、あらかじめパリティサイクル分の領域を割当てることで、マッピング情報の更新回数を削減できる。

【０１１４】

Ｓ１４１１で、プロセッサ２４Ａ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２０２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２０２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。

【０１１５】

Ｓ１４１２において、プロセッサ２４Ａ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２４は、Ｓ１４０８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２０２内に溜まった場合、Ｓ１４１２の判定を真としてＳ１４１３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２０２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Ｓ１４１３へ処理を進めることもありえる。

【０１１６】

Ｓ１４１３において、プロセッサ２４Ａは、バッファ領域２０２内の圧縮データからパリティデータを生成する。Ｓ１４１４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、ＲＡＩＤグループを構成するドライブ２９に書き出す。Ｓ１４１５において、プロセッサ２４Ａは、メモリ割当管理テーブル２１２の更新を確定する。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。

【０１１７】

上記の例では、Ｓ１４１２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２９がＲＡＩＤ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２９に格納する場合には、プロセッサ２４Ａは、Ｓ１４１２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Ｓ１４１２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。

【０１１８】

なお、プロセッサ２４Ａは、Ｓ１４０１の判定が偽の場合であっても、Ｓ１４０８～Ｓ１４１５ではなく、Ｓ１４０２～Ｓ１４０６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ＶＯＬ属性４２が圧縮有効でないために、図１２のＳ１３０３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２０３Ａに格納されている。この場合、

Ｓ１４０１の判定は偽となるが、データの圧縮は行わないため、Ｓ１４０２～Ｓ１４０６が実行される。

【０１１９】

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２０３で二重化された時点でホスト計算機３０に応答が返され、デステ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。

【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ＩＯ命令が実行できないといったトラブルが発生する場合がある。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。

【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。

【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。

【０１２７】

以後、プロセッサ２４は、Ｓ１５１４のドライブ転送可否の判定、Ｓ１５１５のパリティ生成、Ｓ１５１６のドライブ転送を、それぞれケース１－１のＳ１４１２、Ｓ１４１３及びＳ１４１４と同様に行う。

【０１２８】

Ｓ１５１７において、プロセッサ２４は、デステージ範囲のスロット排他を再度確保し、Ｓ１５１８でメモリ割当管理テーブル２１２のキュー状態９５を“Clean”に更新する。

【０１２９】

なお、Ｓ１５１７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３０からの更新ライトが発生した場合、プロセッサ２４は、Ｓ１３０８においてメモリ割当管理テーブル２１２のＢＦ転送状態９６を“無し”に更新する。この場合、Ｓ１５１８でプロセッサ２４がキュー状態９５を更新する際にＢＦ転送状態９６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出来る。

【０１３０】

なお、更新ライトの発生に気づいた（すなわちＳ１５１２で“転送済”に更新したＢＦ転送状態９６がＳ１５１７の時点で“無し”となっていた）場合、プロセッサ２４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２４は、Ｓ１５１８に進まずにＳ１５０８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Ｓ１５０８に進み、更新ライトが行われたスロットのキュー状態９５を“Clean”に更新せずに、Ｓ１５１９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。

【０１３１】

最後にＳ１５１９で、プロセッサ２４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。

【０１３２】

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。

【０１３３】

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。

【０１３４】

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。

【０１３５】

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。

【０１３６】

例えば、ＩＯ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。

【０１３７】

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ＩＯ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。

【０１３９】

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ものではない。

【０１４０】

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。

【０１４１】

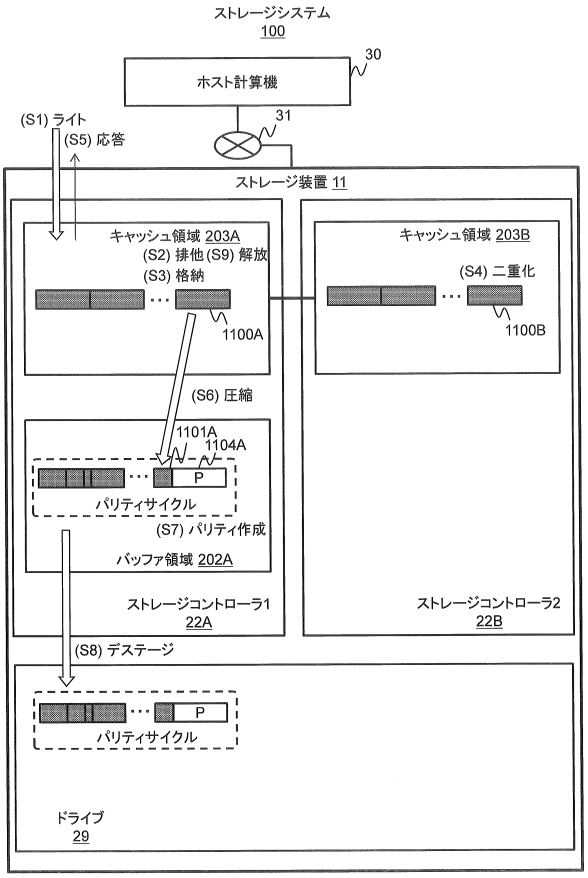
また、制御線及び情報線は説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線及び情報線を示しているとは限らない。実際にはほとんど全ての構成が相互に接続されていると考えてもよい。

【符号の説明】

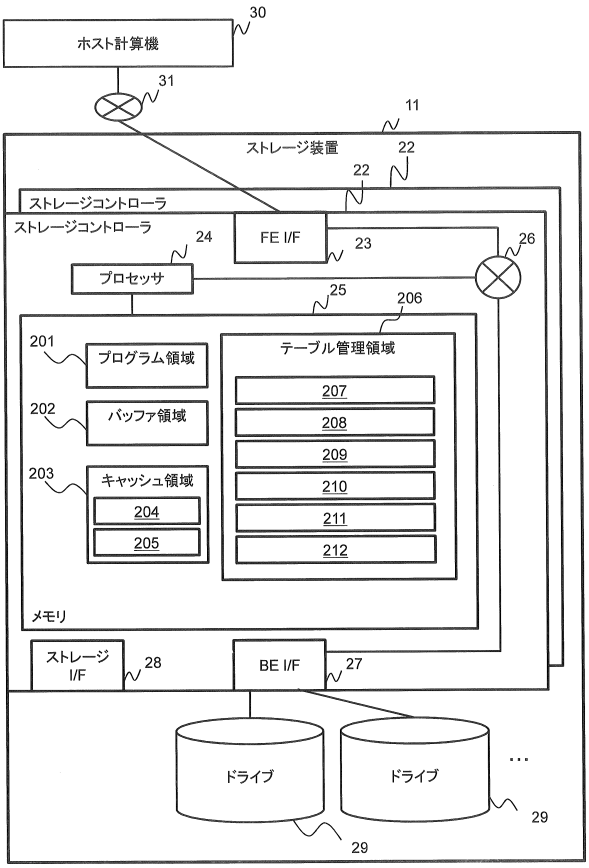
【０１４２】

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

【図１】



【図２】



【図３】

ボリューム管理テーブル
207

41 VOL ID	42 VOL属性	43 VOL容量	44 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

51 プールID	52 RAIDグループID	53 プール容量	54 プール使用容量
0	0 1	10TB	5TB
…	…	…	…

【図５】

RAID構成管理テーブル
209

61 RAIDグループID	62 RAIDレベル	63 ドライブID	64 ドライブ種別	65 容量	66 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

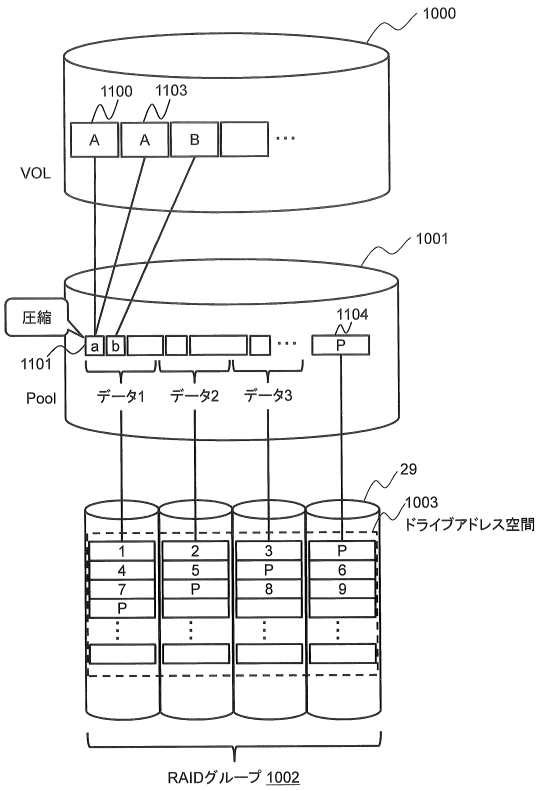
71 VOL ID	72 VOL アドレス	73 プールID	74 プール アドレス	75 圧縮前 サイズ	76 圧縮後 サイズ	77 圧縮率
0	100	0	10	8KB	4KB	1/2
0	200	0	10000	8KB	2KB	1/4
…	…	…	…	…	…	…

【図７】

ドライブ割当管理テーブル
211

81 プールID	82 プールアドレス	83 RAIDグループID	84 ドライブID	85 ドライブアドレス
0	10	0	0	100
0	10000	1	5	1000
…	…	…	…	…

【図８】

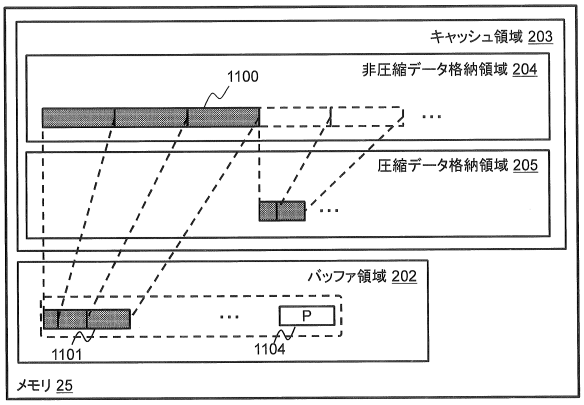


【図９】

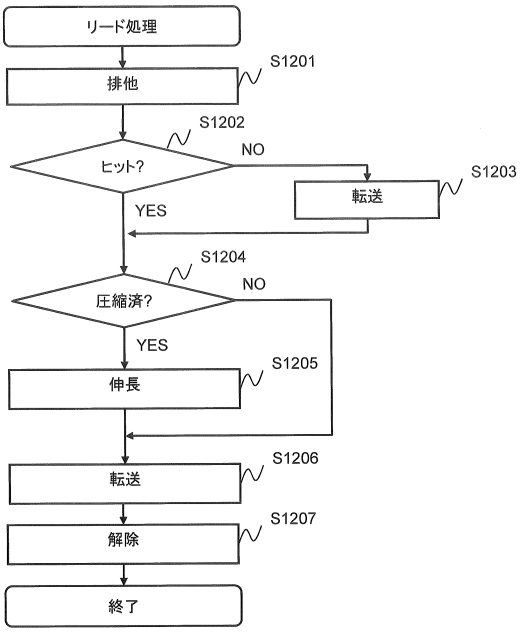
メモリ割当管理テーブル
212

91 VOL ID	92 VOLアドレス	93 BFアドレス	94 圧縮後VOLアドレス	95 キュー状態	96 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
…	…	…	…	…	

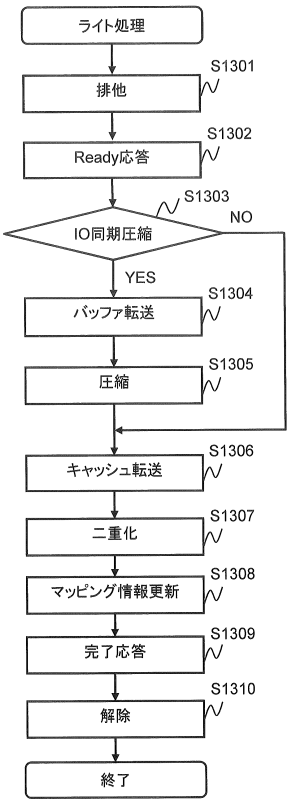
【図１０】



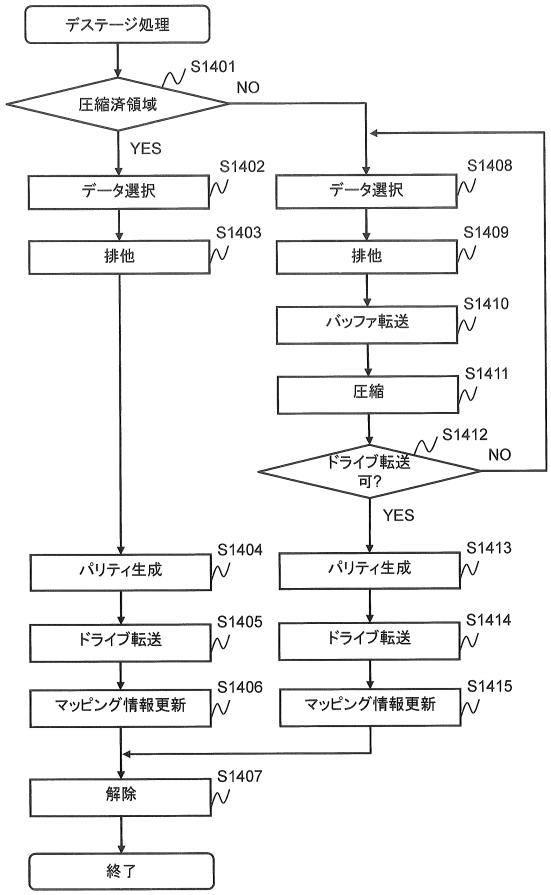
【図１１】



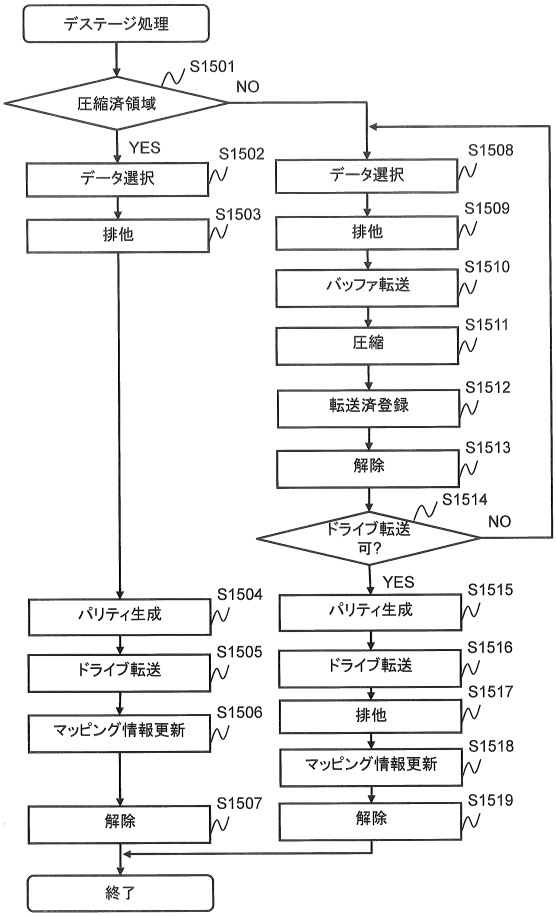
【図１２】



【図１３】


デステージ処理
S1401
圧縮済領域
NO
YES
S1402
データ選択
S1403
排他
S1404
パリティ生成
S1405
ドライブ転送
S1406
マッピング情報更新
S1407
解除
終了
S1408
データ選択
S1409
排他
S1410
バッファ転送
S1411
圧縮
S1412
ドライブ転送可?
NO
YES
S1413
パリティ生成
S1414
ドライブ転送
S1415
マッピング情報更新

【図１４】


デステージ処理
S1501
圧縮済領域
NO
YES
S1502
データ選択
S1503
排他
S1504
パリティ生成
S1505
ドライブ転送
S1506
マッピング情報更新
S1507
解除
終了
S1508
データ選択
S1509
排他
S1510
バッファ転送
S1511
圧縮
S1512
転送済登録
S1513
解除
S1514
ドライブ転送可?
NO
YES
S1515
パリティ生成
S1516
ドライブ転送
S1517
排他
S1518
マッピング情報更新
S1519
解除

フロントページの続き

(51)Int.Cl.			FI		
G06F 12/0868		*(2016.01)*	G06F	16/174	
G06F 11/20		*(2006.01)*	G06F	12/0866	100
			G06F	12/0868	105
			G06F	11/20	689

(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

審査官 松平 英

(56)参考文献 特開２０１１－１９２０５３（ＪＰ，Ａ）
特開２００７－２９３６５１（ＪＰ，Ａ）
特開２００４－３１８４８４（ＪＰ，Ａ）
特開２００９－０４８４９７（ＪＰ，Ａ）
特表２０１５－５１７６９７（ＪＰ，Ａ）
米国特許出願公開第２００７／０２５５９１４（ＵＳ，Ａ１）
米国特許出願公開第２００４／０２１０７１３（ＵＳ，Ａ１）
米国特許出願公開第２００９／００５５５９３（ＵＳ，Ａ１）
国際公開第２０１７／０５６２１９（ＷＯ，Ａ１）
特開２００５－１５７８１５（ＪＰ，Ａ）
特表２０１６－５１０４４０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ３／０６－３／０８
１１／１６－１１／２０
１２／０８－１２／１２８
Ｇ０６Ｆ１６／００－１６／９５８