(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7316242号
(P7316242)

(45)発行日　令和5年7月27日(2023.7.27)

(24)登録日　令和5年7月19日(2023.7.19)

(51)国際特許分類

		FI		
G06F	*3/06* *(2006.01)*	G06F	3/06	304E
G06F	*13/10* *(2006.01)*	G06F	3/06	302A
G06F	*11/16* *(2006.01)*	G06F	13/10	340A
		G06F	11/16	666

請求項の数　11　（全48頁）

(21)出願番号　特願2020-47216(P2020-47216)
(22)出願日　令和2年3月18日(2020.3.18)
(65)公開番号　特開2021-149366(P2021-149366A)
(43)公開日　令和3年9月27日(2021.9.27)
審査請求日　令和3年2月8日(2021.2.8)

(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　110001689
青稜弁理士法人
(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　達見 良介
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　大平 良徳
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　小川 純司
最終頁に続く

(54)【発明の名称】　ストレージシステムおよびデータ転送方法

(57)【特許請求の範囲】

【請求項１】

キャッシュと、ストレージコントローラと、データを格納するストレージ装置を含む一つ以上のストレージシステムにおいて、

前記ストレージ装置は、記憶媒体を含む一つ以上のストレージボックスを含み、

前記ストレージ装置は、サーバからのライト要求にかかるデータを、前記キャッシュに格納してから前記ストレージボックスに格納する第１のライト処理と、前記キャッシュに格納せずに前記ストレージボックスに格納する第２のライト処理と、を実行可能であり、

前記ストレージコントローラは、前記第２のライト処理を行う場合に、サーバから受信したライト要求に基づき、前記ライト要求にかかるデータを格納するアドレスを設定して前記ストレージボックスに通知し、

前記ストレージボックスは、

サーバから受信したライトデータを受信し、受信したライトデータから冗長データを生成し、

前記ライトデータ及び前記冗長データを、前記ストレージコントローラから通知されたアドレスに従って前記記憶媒体に書き込み、

前記記憶媒体への書き込みに際して異常が発生したとき、前記ライトデータを、前記ストレージコントローラに送信し、

前記ストレージコントローラは、

受信した前記ライトデータを前記キャッシュに書き込む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において、
前記ストレージボックスは、
前記ライトデータの中で、前記ストレージコントローラに送ることができないデータが存在する場合、その旨とそのデータを書き込むよう指定された領域のアドレスを、前記ストレージコントローラに伝え、
前記ストレージコントローラは、
前記データを書き込むよう指定された領域のアドレスに対応するビットマップをオンに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２に記載のストレージシステムにおいて、
前記ストレージコントローラは、
前記ビットマップにおいてオンになっているビットに対応する領域に対し、前記サーバからリード要求を受けつけたとき、前記リード要求を、異常終了させ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において、
前記ストレージボックスは、前記冗長データを生成できない場合に、前記ライトデータを前記ストレージコントローラに送り、
前記ストレージコントローラは、前記ストレージボックスから受信した前記ライトデータを前記キャッシュに、二重に書き込む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において、
前記ストレージボックスは、
前記ライトデータ及び前記冗長データを、前記記憶媒体の指定された領域に書き込めないとき前記ライトデータを前記ストレージコントローラに送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において、
前記ストレージ装置は、
前記サーバから、書込み領域を指定したライト要求を受けつけ、
前記指定された書込み領域がシーケンシャルになっているかを判別し、
シーケンシャルになっている場合、前記ストレージボックスに、前記サーバからライトデータを受け付けるよう指示することを特徴とするストレージシステム。
【請求項７】
請求項１に記載のストレージシステムにおいて、
前記ストレージ装置は、
前記サーバから、書込み領域を指定したライト要求を受け、
前記ライト要求で書き込まれるデータを前記記憶媒体にログストラクチャ形式で書き込むよう制御し、
前記ストレージボックスに、前記サーバからのライト要求で書き込まれるデータを受け取るよう指示する
ことを特徴とするストレージシステム。
【請求項８】
請求項７に記載のストレージシステムにおいて、
前記ストレージボックスは、
受け取ったライトデータのハッシュ値を計算し、前記ハッシュ値と対応するライトデータの識別子を、前記ストレージコントローラに送り、
前記ストレージコントローラは、前記ハッシュ値と前記対応するライトデータの識別子

を記憶することを特徴とするストレージシステム。
【請求項９】
請求項８に記載のストレージシステムにおいて、
前記記憶媒体への書き込みに際して異常が発生した場合に、
前記ストレージボックスは、
ハッシュ値が一致するものがある場合、指定された領域のデータを読み出し、ライトデータの値と一致するかどうかをチェックし、
ハッシュ値が一致するものがなかったライトデータと、一致するものがあっても読み出した領域のデータが一致しなかったライトデータから、冗長データを生成して、前記ライトデータと前記冗長データとを、前記ストレージコントローラに送信し、
前記ストレージコントローラは、前記ストレージボックスから受信した前記ライトデータと前記冗長データを、前記キャッシュに記憶し、
前記ストレージボックスは、
前記読み出した領域データとライトデータが一致した場合、その旨と前記ライトデータの識別子と読み出した領域を、前記ストレージコントローラに送信し、
前記ストレージコントローラは、
受け取ったライトデータの識別子と読み出した領域を記憶することを特徴とするストレージシステム。
【請求項１０】
請求項１に記載のストレージシステムにおいて、
前記ストレージコントローラは、前記第２のライト処理を行う場合に、前記ストレージボックスが前記ライトデータを受信した報告を受領したら、前記サーバに完了報告を送信し、
前記完了報告の送信後に、前記ストレージボックスは、前記冗長データの生成と、前記ライトデータ及び前記冗長データの前記記憶媒体への書き込みと、を行うことを特徴とするストレージシステム。
【請求項１１】
キャッシュと、ストレージコントローラと、データを格納するストレージ装置とを含む一つ以上のストレージシステムにおけるデータ転送方法において、
前記ストレージ装置は、記憶媒体を含む一つ以上のストレージボックスを含み、
前記ストレージ装置は、サーバからのライト要求にかかるデータを、前記キャッシュに格納してから前記ストレージボックスに格納する第１のライト処理と、前記キャッシュに格納せずに前記ストレージボックスに格納する第２のライト処理と、を実行可能であり、
前記ストレージコントローラは、前記第２のライト処理を行う場合に、サーバから受信したライト要求に基づき、前記ライト要求にかかるデータを格納するアドレスを設定して前記ストレージボックスに通知し、
前記ストレージボックスは、
サーバから受信したライトデータから冗長データを生成し、
前記ライトデータ及び前記冗長データを前記、前記ストレージコントローラから通知されたアドレスに従って前記記憶媒体に書き込み、
前記記憶媒体への書き込みに際して異常が発生したとき、前記ライトデータを、前記ストレージコントローラに送信し、
前記ストレージコントローラは、
受信した前記ライトデータを前記キャッシュに格納することを特徴とするデータ転送方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、フラッシュメモリや磁気ディスクをストレージ（記憶媒体）として用いるスケールアウト型のストレージシステムライト処理の高性能化、高信頼化に関する。

【背景技術】
【０００２】
近年、SSDをはじめ、不揮発性メモリを使用したフラッシュストレージのために最適化された通信プロトコルであるNVMe（Non-Volatile Memory Express）を利用することで、ストレージシステムの大幅な処理高速化が図られている。このようなストレージシステムでは、ストレージコントローラが性能のボトルネックとなることが考えられる。そのため、ストレージシステムに接続されるドライブボックスにホストコンピュータをファブリックネットワーク（NVMeoF:NVMe over Fabric）により直接接続し、ホストコンピュータとドライブボックス間でストレージシステムを介さず直接データ転送をすることによりストレージコントローラの性能ボトルネックを解消し、高速なデータ転送を実現することが考えられる。
【０００３】
（直接データ転送実現の課題）
直接データ転送を実現するにあたり、以下の２つの課題がある。
【０００４】
（課題１）ストレージシステムの提供する論理ボリュームについて、ホストコンピュータから見たアドレス空間と、FBOF内のドライブのアドレス空間が異なり、ホストコンピュータは所望のデータがFBOF内のドライブのどのアドレスに格納されているか識別できない。
【０００５】
（課題２）ストレージシステムのキャッシュを使うことでデータアクセスの高性能化を行う場合、キャッシュに新データがある時は、ストレージのキャッシュからデータを読み込む必要があるが、ホストコンピュータは新データの有無を判断できない。
【０００６】
このような課題に対して、ホストコンピュータで動作するAgentソフトウエアが、ホストコンピュータのアクセス先データに対応するFBOF内のドライブとそのアドレスをストレージコントローラに問い合わせ、得られた情報を元に直接FBOF内のドライブにアクセスする発明が開示されている。
【０００７】
特許文献１に開示された発明では、ホストコンピュータが直接FBOFのドライブにアクセスできる一方で、AgentソフトウエアでRAIDなどのデータ保護のための計算をしなくてはならず、ホストコンピュータ側に高信頼な処理を行うための計算負荷がかかる問題がある。
【先行技術文献】
【特許文献】
【０００８】
【文献】US9,800,661号明細書
【発明の概要】
【発明が解決しようとする課題】
【０００９】
特許文献１に開示された発明では、ホストコンピュータが直接FBOFのドライブにアクセスできる一方で、AgentソフトウエアでRAIDなどのデータ保護のための計算をしなくてはならず、ホストコンピュータ側に高信頼な処理を行うための計算負荷がかかる問題がある。
【００１０】
また、サーバからストレージボックスに、データを直接転送するライト処理を実行する場合、以下の三つの課題を解決する必要がある。
【００１１】
第１は、従来のライト処理は、ストレージ制御装置が有するキャッシュに、サーバから受け取ったデータを書き込んだ段階で、サーバに対する終了報告を行っていた。本発明に

おいても、サーバから見て、従来のライト処理の同等の時間で、ライト処理を同程度の時間で完了させる必要がある。

【００１２】

第２は、冗長データの作成が課題となる。従来のストレージ制御装置は、データを受け取った後、受け取ったデータから冗長データを生成して、受け取ったデータ、冗長データを記憶装置に格納した。記憶装置が故障しても、他の記憶装置のデータと冗長データから、故障した記憶装置のデータを復元する機能を有していた。受け取ったデータ、冗長データを記憶装置に格納するまでの間、ストレージ制御装置が、受け取ったデータを喪失しないように、キャッシュにはバッテリーなどが装備され、データもキャッシュに２重化して格納していた。本発明では、データをストレージ制御装置が受け取らないので、データの喪失を回避しつつ、冗長データをどのように作成するかが課題となる。

【００１３】

第３は、障害が発生したときの対応である。サーバに対して、完了を報告したデータについては、ストレージ制御装置は、責任を持って保管する必要があった。このため、障害などが発生しても、データの喪失等を極力抑える必要があるため、ストレージ制御装置は、これらを防ぐため、膨大な論理を有している。同様の論理をストレージボックスで開発しようとすると、開発工数が多大となり現実的でない。

【００１４】

本発明は、複数のサーバ、複数のストレージシステム、フラッシュメモリや磁気ディスクなどのストレージ装置を格納するストレージボックスをネットワークで共有する構成において、ライト処理を、サーバからストレージボックスに、データを直接転送する際の課題を解決する複合的システムおよびデータ転送方法を提供することを目的とする。

【課題を解決するための手段】

【００１５】

上記課題を解決する本発明の一側面は、キャッシュとストレージコントローラを含む一つ以上のストレージシステムと、記憶媒体を含む一つ以上のストレージボックスと、を含む複合的システムにおいて、前記ストレージボックスは、サーバから受信したライトデータから冗長データを生成し、前記ライトデータ及び前記冗長データを前記記憶媒体に書き込み、前記ストレージボックスは、前記冗長データを生成できない、または前記ライトデータ及び前記冗長データを前記記憶媒体に書き込みできないとき、前記ライトデータを、前記ストレージシステムに送信し、前記ストレージシステムは、受信した前記ライトデータを前記キャッシュに格納する。

【発明の効果】

【００１６】

本発明により、複数のサーバ、複数のストレージシステム、フラッシュメモリや磁気ディスクなどのストレージ装置を格納するストレージボックスをネットワークで共有する構成において、ライト処理を実行する際に、ストレージ制御装置の指示により、サーバから送られたデータを、直接ストレージボックスが受け取ることで、ネットワーク負荷の低減を図ることができる。

【図面の簡単な説明】

【００１７】

【図１】実施例１における情報システムの構成を示す図である。

【図２】実施例１におけるサーバポート情報の構成を示す図である。

【図３】実施例１における実ストレージシステムの構成を示す図である。

【図４】実施例１におけるキャッシュの構成を示す図である。

【図５】実施例１におけるストレージシステムの共有メモリに格納された情報を示す図である。

【図６】実施例１における仮想ストレージシステム情報を示す図である。

【図７】実施例１における他ストレージシステムの情報の形式を示す図である。

【図８】実施例１における仮想論理ボリューム情報の形式を示す図である。

【図９】実施例１における論理ボリューム情報の形式を示す図である。
【図１０】実施例１におけるキャッシュ管理情報の形式を示す図である。
【図１１】実施例１における空きキャッシュ管理情報キューの形式を示す図である。
【図１２】実施例１におけるストレージボックス情報の形式を示す図である。
【図１３】実施例１におけるストレージグループ情報の形式を示す図である。
【図１４】実施例１におけるストレージ装置情報の形式を示す図である。
【図１５】実施例１におけるストレージコントローラが実行するプログラムの構成を示す図である。
【図１６Ａ】実施例１におけるライト要求受付部の処理フローを示す図である。
【図１６Ｂ】実施例１におけるライト要求受付部の処理フローを示す図である。
【図１７】実施例１におけるライト異常処理対応処理部の処理フローを示す図である。
【図１８】実施例１におけるリード処理実行部の処理フローを示す図である。
【図１９】実施例１におけるストレージボックスが実行するプログラムの構成を示す図である。
【図２０】実施例１におけるライトデータ受領部の処理フローを示す図である。
【図２１】実施例１におけるライトデータ書込み部の処理フローを示す図である。
【図２２】実施例１における一時書込み領域転送部の処理フローを示す図である。
【図２３】実施例１におけるリードデータ直接転送部の処理フローを示す図である。
【図２４】実施例２における情報システムの構成を示す図である。
【図２５】実施例２におけるストレージシステムの共有メモリに格納された情報を示す図である。
【図２６】実施例２における論理ボリューム情報の形式を示す図である。
【図２７】実施例２におけるストレージグループ情報の形式を示す図である。
【図２８】実施例２における実ページ情報の形式を示す図である。
【図２９】実施例２における空きページ管理情報キューの構成を示す図である。
【図３０】実施例２におけるトレージコントローラが実行するプログラムの構成を示す図である。
【図３１Ａ】実施例１におけるライト要求受付部の処理フローを示す図である。
【図３１Ｂ】実施例１におけるライト要求受付部の処理フローを示す図である。
【図３２】実施例２におけるライト異常処理対応処理部の処理フローを示す図である。
【図３３】実施例２におけるリード処理実行部の処理フローを示す図である。
【図３４】実施例２における重複排除処理実行部の処理フローを示す図である。
【図３５】実施例２におけるストレージボックスが実行するプログラムの構成を示す図である。
【図３６】実施例２におけるデータ移動部の処理フローを示す図である。
【図３７】実施例２におけるライトデータ受領部の処理フローを示す図である。
【図３８】実施例２におけるライトデータ書込み部の処理フローを示す図である。
【図３９】実施例２における一時書込み領域転送部の処理フローを示す図である。
【図４０】実施例２におけるリードデータ直接転送部の処理フローを示す図である。
【図４１】実施例２におけるボックス間データ転送部の処理フローを示す図である。
【発明を実施するための形態】
【実施例１】
【００１８】
＜発明の概要＞
サーバからストレージボックスに、データを直接転送するライト処理を実行する場合、以下の三つの課題を解決する必要がある。
【００１９】
第１は、従来のライト処理は、ストレージ制御装置が有するキャッシュに、サーバから受け取ったデータを書き込んだ段階で、サーバに対する終了報告を行っていた。本発明においても、サーバから見て、従来のライト処理の同等の時間で、ライト処理を同程度の時

間で完了させる必要がある。

【００２０】

第２は、冗長データの作成が課題となる。従来のストレージ制御装置は、データを受け取った後、受け取ったデータから冗長データを生成して、受け取ったデータ、冗長データを記憶装置に格納した。記憶装置が故障しても、他の記憶装置のデータと冗長データから、故障した記憶装置のデータを復元する機能を有していた。受け取ったデータ、冗長データを記憶装置に格納するまでの間、ストレージ制御装置が、受け取ったデータを喪失しないように、キャッシュにはバッテリーなどが装備され、データもキャッシュに２重化して格納していた。本発明では、データをストレージ制御装置が受け取らないので、データの喪失を回避しつつ、冗長データをどのように作成するかが課題となる。

【００２１】

第３は、障害が発生したときの対応である。サーバに対して、完了を報告したデータについては、ストレージ制御装置は、責任を持って保管する必要があった。このため、障害などが発生しても、データの喪失等を極力抑える必要があるため、ストレージ制御装置は、これらを防ぐため、膨大な論理を有している。同様の論理をストレージボックスで開発しようとすると、開発工数が多大となり現実的でない。

【００２２】

本発明では、基本的には、ストレージ制御装置とストレージボックスが、機能分担を行い、これらの課題を解決する。

【００２３】

第１と第２の課題を解決するための方法を説明する。なお、本発明では、ストレージ制御装置のように、データはストレージ制御装置を通過しないので、ストレージボックスが、冗長データを生成する。冗長データは、主に、データを２重化する方法と、複数のデータから、冗長データを生成し、失われたデータを、他のデータと冗長データから、復元する方法がある。後者を詳細に説明する。後者において、冗長データを生成する方法を説明する。冗長データを生成する、複数のデータの組をパリティグループと呼ぶ。第１の方法は、パリティグループ全体のデータの書込みが行われたとき、全体のデータから冗長データを生成する方法である。もう一つの方法は、比較的少量のデータが書き込まれたときに適用される方法である。本発明は、ネットワークの負荷を低減することを目的としているので、前者を前提とし、ストレージボックスにデータを直接転送するようにする。つまり、このデータの更新前の値と更新後の値と更新前の冗長データとから、更新後のパリティを生成する方法である。

【００２４】

パリティグループ全体のデータを更新するアクセスパターンは、データをアドレス順に更新するシーケンシャルライトである。また、１回のシーケンシャルライトでパリティグループ全体が更新されるわけではない。パリティグループ全体のデータが揃うまで、冗長データは生成できないので、この間、ストレージボックスが受け取ったデータを喪失しないように、本発明では、ストレージボックスがボックス内の記憶装置に、受け取ったデータを一時的に２重に書き込んでおくようにする。ストレージボックスは、記憶装置に２重にデータを書き終えた契機で、ライト処理を完了させる。これにより、第１の課題である、サーバから見たライト処理を、従来と同等の時間で完了させることができる。

【００２５】

記憶装置にパリティグループ全体のデータが２重に書きこまれると、ストレージボックスが冗長データを生成し、パリティグループ全体のデータと冗長データを、ストレージボックスの中の本来書き込むべき領域に、書き込む。この後、一時的に２重にデータを書き込んでおいた領域を開放する。パリティグループ全体のデータと冗長データを書き込むまで、別の領域に、データを２重書き込まれているので、データの喪失を回避しつつ、冗長データを生成するという第２の課題を解決できる。

【００２６】

第３の課題の解決方法を説明する。ストレージ制御装置では、キャッシュにライトデー

タを２重化して保持した状態で、障害等が発生した場合の処理を実行していた。このため、本発明では、パリティグループ全体のデータと冗長データを、書き込む前に、何か障害が発生すると、ストレージボックスは、それまで受け取っていたデータをストレージ制御装置に転送し、ストレージ制御装置は、このデータを、キャッシュに２重に格納する。この状態にすると、従来ストレージ制御装置がもっていた障害に対応する論理をそのまま適用することが可能となるので、第３の課題を解決できる。

【００２７】

図１は、実施例１における情報システムの構成を示す。情報システムは、一つ以上のストレージボックス１３０、一つ以上の実ストレージシステム１００、一つ以上のサーバ１１０と、実ストレージシステム１００、サーバ１１０、ストレージボックス１３０とを接続するネットワーク１２０と、ストレージ管理サーバ”１５０とを有する。

【００２８】

サーバ１１０は、一つ以上のサーバポート１９５によって、実ストレージシステム１００は、一つ以上のストレージポート（ストレージパス）１９７によって、ストレージボックス１３０は、ボックスポート１８０によって、ネットワーク１２０に接続される。

【００２９】

サーバ１１０は、内部に、サーバポート情報１９８をもつ。サーバ１１０は、ユーザアプリケーションが動作するシステムで、実ストレージステム１００に対して、ネットワーク１２０経由で、データの読み書き要求を発行する。また、実ストレージシステム１００は、ネットワーク１２０経由で、サーバ１１０から受け取った読み書き要求で指定されたデータを格納したストレージボックス１３０に要求を伝える。

【００３０】

実ストレージシステム１００同士も、互いに、ネットワーク１２０経由で、データの送受信を行う。ネットワーク１２０は例えば、NVM-e over Ethernet等のプロトコルを用いることができる。実施例１では、一つ以上の実ストレージシステム１００から構成される仮想ストレージシステム１９０が存在する。サーバ１１０からは、仮想ストレージシステム１９０が、一台のストレージシステムに見える。ただし、本実施の形態は、仮想ストレージステム１９０が存在しない場合も有効である。仮想ストレージステム１９０が存在しない場合、実ストレージシステム１００が、サーバ１１０からは、ストレージステムに見える。

【００３１】

ストレージボックス１３０は、ＨＤＤ（Hard Disk Drive）、フラッシュメモリを記憶媒体とするフラッシュストレージなどのストレージ装置１６０（共有ストレージ装置１７５）とネットワーク１２０に接続するための１つ以上のボックスポート１８０、実ストレージシステム１００の指示にしたがって、ストレージ装置１６０との間で転送を行うひとつ以上のボックスプロセッサ１７０、ボックスプロセッサ１７０が使用するボックスメモリ１８１などが含まれる。また、一つのストレージ装置１６０は、一つ以上のストレージパス１９７に接続されているものとする。

【００３２】

また、フラッシュストレージも、いくつかの種類があり、高価格、高性能、消去可能回数の多いSLCと、これに対し、低価格、低性能、消去可能回数の少ないMLCとがある。さらに、相変化メモリなどの新しい記憶媒体が含まれていてもよい。

【００３３】

実施例１においては、ストレージボックス１３０の中のストレージ装置１６０は、複数の実ストレージシステム１００によって共有される。したがって、実施例１では、ストレージボックス１３０の中のストレージ装置１６０を、共有ストレージ装置１７５と呼ぶ。

【００３４】

実ストレージシステム１００とストレージボックス１３０は、ネットワーク１２０経由で、仮想ストレージシステム１９０の中の一つ以上の実ストレージシステム１００に接続される。ストレージボックス１３０は、必ずしも、仮想ストレージシステム１９０の中の

すべての実ストレージシステム１００と接続されている必要はない。同様に、ある実ストレージシステム１００と別の実ストレージシステム１００に接続されているストレージボックス１３０の集合は、まったく同じである必要はない。ストレージボックス１３０は、ネットワーク１２０経由で、実ストレージシステム１００のいくつかによって、共有されている。

【００３５】

ストレージ管理サーバ１５０は、ネットワーク１２０などを経由して、実ストレージシステム１００と接続されていて、実ストレージステム１００、ストレージボックス１３０の管理を行うため、ストレージ管理者が利用する装置である。また、実施例１では、実ストレージシステム１００は、容量仮想化機能をもたないものとする。ただし、実施例１は、実ストレージシステム１００が、容量仮想化機能をもつ場合でも、有効である。

【００３６】

図２は、サーバポート情報１９８のフォーマットである。サーバポート情報１９８は、サーバポート１９５ごとに持つ情報である。サーバ１１０は、実ストレージシステム１００に読み書き要求を発行する際、ストレージシステムの識別子、論理ボリュームの識別子、ストレージポート（ストレージパス１９７）の識別子を設定する。このため、サーバポート情報１９８は、サーバポート識別子２４０００と、当該サーバポート１９５からアクセスする一つ以上の論理ボリューム識別子２４００１、この論理ボリュームを含むストレージシステム識別子２４００２、ストレージポート識別子２４００３が含まれる。当該論理ボリュームが複数のストレージパス１９７に接続されている場合、複数のストレージポート識別子（ストレージパス識別子）２４００３が設定される。なお、実施例１では、ストレージシステム識別子２４００２には、仮想ストレージシステム１９０の識別子が設定される。ただし、本発明は、仮想ストレージシステム１９０が存在しない場合も有効で、その場合には、ストレージシステム識別子２４００２には、実ストレージステム１００の識別子が設定される。ストレージポート識別子２４００３には、ストレージパス１９７の識別子が設定される。

【００３７】

実施例１では、論理ボリューム識別子２４００１には、仮想論理ボリュームの識別子が設定される。仮想論理ボリュームの識別子は、仮想ストレージステム１９０内で、ユニークな値である。一方、それぞれの実ストレージススステム１００も論理ボリュームをもつ。論理ボリュームの識別子は、実ストレージシステム１００内では、ユニークである。サーバ１１０の読み書き要求は、仮想ストレージシステムの識別子、仮想論理ボリュームの識別子、ストレージパス１９７の識別子を含む。ストレージパス１９７の識別子は、リアルな値であるため、この要求を受け取る、実ストレージシステム１００が決定される。なお、本発明は、仮想ストレージシステム１９０が存在しない場合も有効で、その場合には、論理ボリューム識別子２４００１には、実ストレージステム１００の論理ボリュームの識別子が設定される。また、実施例１では、仮想論理ボリュームを読み書きする実ストレージシステム１００を変更するが、その際にも、仮想論理ボリュームの識別子は、変化せず、接続されたストレージパス１９７を変更する。

【００３８】

図３は、実ストレージシステム１００の構成である。実ストレージシステム１００は、一つ以上のストレージコントローラ２００、キャッシュ２１０、共有メモリ２２０、ストレージ装置１６０（内部ストレージ装置２３０）と、これらの構成要素を接続する一つ以上の接続装置２５０、ネットワーク１２０とのインターフェイスであるストレージポート１９７とを有する。

【００３９】

実ストレージシステム１００内に含まれる内部ストレージ装置２３０も、ストレージボックス１３０内に含まれるストレージ装置１６０も、データを格納する装置である。特に、実施例１では、実ストレージシステム１００に含まれるストレージ装置１６０を内部ストレージ装置２３０、ストレージボックス１３０内のストレージ装置１６０を共有ストレ

ージ１７５と呼ぶ。ただし、本発明の対象は、ストレージボックス１３０内のストレージ装置であるため、基本的には、ストレージ装置１６０は、共有ストレージ１７５を指すものとする。また、本発明においては、ストレージシステム１００は、必ずしも、内部ストレージ装置２３０をもつ必要はない。また、本発明は、複数の実ストレージシステム１００によって共有される一つ以上のストレージボックス１３０によって構成される複合ストレージシステムに関するものなので、内部ストレージ装置２３０に対し、ストレージコントローラ２００が実行する処理、共有メモリ２２０に保有する情報などについては、説明を省略する。

【００４０】

ストレージコントローラ２００は、サーバ１１０から発行されたリード・ライト要求を処理するプロセッサ２６０、プログラムや情報を保管するメモリ２７０、バッファ２７５と、から構成される。特に、本発明は、バッファ２７５は、（１）後述する冗長データを生成する際、生成に必要な情報や、生成した冗長データを格納するため、（２）実ストレージシステム１００のキャッシュ領域に格納されたデータを、恒久的に格納するストレージ装置１６０に書き込む際の一時的な格納領域とし使用される。

【００４１】

接続装置２５０は、実ストレージシステム１００内の各構成要素を接続する機構である。本発明の特徴は、接続装置２５０経由で、一つ以上のストレージボックス１３０が接続されている点である。これにより、ストレージコントローラ２００は、ストレージボックス１３０内のストレージ装置１６０に、読み書きできる。また、実施例１では、ストレージボックス１３０には、実ストレージシステム１００内の一つ以上のストレージコントローラ２００に接続されているものとする。

【００４２】

キャッシュ２１０、共有メモリ２２０は、通常DRAMなどの揮発メモリで構成されるが、バッテリーなどにより不揮発化されているものとする。また、実施例１では、高信頼化のため、それぞれが２重化されているものとする。ただ、本発明は、キャッシュ２１０、共有メモリ２２０が不揮発化されていなくても、２重化されていなくとも、有効である。また、実施例１では、キャッシュ２１０、共有メモリ２２０は、ストレージコントローラ２００の外側にあるが、ストレージコントローラ２００の内部のメモリ２７０上に、構成されていても良い。

【００４３】

キャッシュ２１０には、内部ストレージ装置２３０、ストレージボックス１３０の中のストレージ装置１６０に格納されたデータの中で、ストレージコントローラ２００からよくアクセスされるデータが格納される。

【００４４】

ストレージコントローラ２００は、サーバ１１０から、リード・ライト要求を受け付ける。通常は、ストレージ装置１６０に書き込むよう受け取ったデータを、キャッシュ２１０に書き込んで、該当するライト要求を完了させた後から、このデータを、ストレージボックス１３０のストレージ装置１６０に書き込むことが多い。しかし、この方法では、データは２回、ネットワーク１２０を通過することになるので、ネットワーク１２０に対する負荷が大きい。

【００４５】

本発明では、書き込むデータを、サーバ１１０から直接、ストレージボックス１３０に転送することにより、ネットワーク１２０の負荷を軽減する。

【００４６】

図４は、キャッシュ２１０の構成を表しているものとする。キャッシュ２１０は、固定長のスロット２１１００に分割されている。スロット２１１００が、リード・ライトデータの割り当て単位となる。

【００４７】

なお、実施例１では、実ストレージシステム１００のストレージ装置１６０の中の一台

の装置が故障しても、その装置のデータを回復できるRedundancy Array Independent Device（RAID）機能をもっているものとする。RAID機能もった場合、複数の同一種類の記憶装置が、一つのRAID構成をとる。これをストレージグループ２８０（図３参照）と呼ぶ。

【００４８】

実施例１では、RAID構成は、一つのストレージボックス１３０内の共有ストレージ１７５の集合、あるいは、一つの実ストレージシステム１００内の内部ストレージ装置２３０の集合から構成されるものとする。RAIDには、いくつかのタイプがある。例えば、RAID1は、データを２重に書く。RAID4やRAID5は、複数のデータから、パリティ（以下、パリティ）を生成する。RAIDのタイプは、ストレージ装置グループごとに決まっている。なお、実施例１では、RAIDのタイプは、RAID5とする。もちろん、本発明は、RAID５以外のRAIDタイプでも有効である。

【００４９】

本発明では、ストレージボックス１３０がサーバ１１０からデータを直接受け取り、データはストレージ制御装置を通過しないので、ストレージボックス１３０が、パリティを生成する。

【００５０】

RAID5おいて、パリティを生成する方法を説明する。パリティを生成する、複数のデータの組をパリティグループと呼ぶ。

【００５１】

第１の方法は、パリティグループ全体のデータの書込みが行われたとき、全体のデータから冗長データを生成する方法である。もう一つの方法は、比較的少量のデータが書き込まれたときに適用される方法である。本発明は、ネットワークの負荷を低減することを目的としているので、前者を前提とし、ストレージボックスにデータを直接転送するようにする。

【００５２】

パリティグループ全体のデータを更新するアクセスパターンは、データをアドレス順に更新するシーケンシャルライトである。また、１回のシーケンシャルライトでパリティグループ全体が更新されるわけではない。パリティグループ全体のデータが揃うまで、パリティは生成できないので、この間、ストレージボックス１３０が受け取ったデータを喪失しないように、実施例１では、ストレージボックス１３０がボックス内のストレージ装置１６０に、受け取ったデータを一時的に２重に書き込んでおくようにする。ただし、本発明は、ストレージボックス１３０が、ストレージコントローラ２００と同様に、ストレージ装置１６０以外に、不揮発化されたキャッシュを持ち、このキャッシュに、受け取ったデータを２重に書き込んでも有効である。キャッシュの記憶媒体は、バッテリバックアップされたDRAMでも、新しい不揮発の半導体メモリでも、本発明は有効である。ストレージボックス１３０は、ストレージ装置１６０に２重にデータを書き終えた契機で、ライト処理を完了させる。

【００５３】

ストレージ装置１６０にパリティグループ全体のデータが２重に書きこまれると、ストレージボックス１３０がパリティを生成し、パリティグループ全体のデータとパリティを、ストレージボックス１３０の中の本来書き込むべき領域（サーバが書込みを指示した領域）に、書き込む。この後、一時的に２重にデータを書き込んでおいた領域を開放する。

【００５４】

パリティグループ全体のデータとパリティを書き込むまで、データは、別の領域に２重書き込まれているので、データの喪失を回避しつつ、パリティを生成することができる。ただし、２重に書き込む一方を、本来書き込むべき領域（サーバが書込みを指示した領域）に書き込んでおき、パリティグループ全体のデータが２重に書きこまれると、ストレージボックス１３０がパリティを生成し、パリティを書き込むようにしても、本発明は有効である。

【００５５】
ストレージコントローラ２００では、キャッシュ２１０にライトデータを２重化して保持した状態で、障害等が発生した場合の処理を実行していた。このため、本発明では、パリティグループ全体のデータとパリティを、書き込む前に、何か障害が発生すると、ストレージボックス１３０は、それまで受け取っていたデータをストレージコントローラ２００に転送し、ストレージコントローラ２００は、このデータを、キャッシュ２１０に２重に格納する。この状態にすると、ストレージコントローラ２００が従来からもっていた障害に対応する論理をそのまま適用することが可能となる。
【００５６】
図５は、実施例１における実ストレージシステム１００の共有メモリ２２０の中の実施例１に関する情報を示している。共有メモリ２２０には、ストレージシステム情報２０６０、他ストレージシステム情報２０７０、仮想論理ボリューム情報２０８５、論理ボリューム情報２０００、ストレージボックス情報２０５０、ストレージグループ情報２３００、ストレージ装置情報２５００、キャッシュ管理情報２７５０、空きキャッシュ管理情報ポインタ（キュー）２６５０、によって構成される。
【００５７】
この中で、ストレージシステム情報２０６０は、図６に示すように、実ストレージシステム１００に関する情報で、実施例１では、仮想ストレージシステム識別子２０６１、実ストレージシステム識別子２０６２から構成される。仮想ストレージシステム識別子２０６１は、当該実ストレージシステム１００が含まれる仮想ストレージシステム１９０の識別子である。実ストレージシステム識別子２０６２は、当該実ストレージシステム１００の識別子である。
【００５８】
図７は、他ストレージシステム情報２０７０に関する情報を示した図であり、仮想ストレージシステム識別子２０７１、他実ストレージシステム識別子２０７２から構成される。仮想ストレージシステム識別子２０７１は、図６に含まれる仮想ストレージシステム識別子２０５１と同じで、当該実ストレージシステム１００が含まれる仮想ストレージシステム１９０の識別子である。他実ストレージシステム識別子２０７２は、当該実ストレージシステム１００を含む仮想ストレージシステム１９０に含まれる他の実ストレージシステムである。
【００５９】
図８は、仮想論理ボリューム情報２０８５に関する情報を示した図である。論理ボリューム対応に作成される仮想論理ボリューム情報２０８５は、仮想論理ボリューム識別子２０８６、制御権情報２０８７、実ストレージシステム情報２０８８、ストレージポート識別子（ストレージパス識別子）２０８９、サーバポート識別子２０９１、論理ボリューム識別子２０９０等から構成される。
【００６０】
仮想論理ボリューム識別子２０８６は、当該仮想論理ボリュームの識別子である。実施例１では、仮想論理ボリュームは、どれか一つの実ストレージシステム１００が読み書きを行う権限をもつ。制御権情報２０８７は、当該実ストレージシステム１００が制御権をもっているか、そうでないかを表す。制御権をもっていない場合、実ストレージシステム情報２０８８、ストレージパス識別子２０８９に、どの実ストレージシステム１００が制御権をもっているか、その仮想論理ボリュームが接続されている一つ以上のストレージパス１９７が示される。
【００６１】
論理ボリューム識別子２０９０は、当該実ストレージシステム１００内の対応する論理ボリュームの識別子が格納され、そうでない場合、制御権をもつ実ストレージシステム１００内の対応する論理ボリュームの識別子が格納される。
【００６２】
図９は、本発明に関する論理ボリューム情報２０００に関する情報を示した図である。

実施例１では、サーバ１１０がデータをリード・ライトするストレージ装置１６０は、仮想論理ボリュームである。また、サーバ１１０は、仮想論理ボリュームのID、仮想論理ボリューム内のアドレス、読み書きしたいデータの長さを指定して、リード要求やライト要求を発行する。実ストレージシステム１００は、サーバ１１０から読み書き要求を受け取ると、仮想論理ボリューム情報２０８５により、対応する論理ボリュームの識別子を認識する。実施例１では、論理ボリュームの識別子は、実ストレージシステム１００内でユニークな情報である。論理ボリューム情報２０００は、論理ボリュームごとに存在する情報である。この情報は、論理ボリューム識別子２００１、論理容量２００２、論理ボリュームタイプ２００５、論理ボリュームRAIDタイプ２００３、割り当て範囲２０２６、第１キャッシュ管理情報ポインタ２０２２、第２キャッシュ管理情報ポインタ２０２３、第１ストレージポインタ２０２４と第２ストレージポインタ２０２５等より構成される。

【００６３】

論理ボリューム識別子２００１は、対応する論理ボリュームのIDを示す。論理容量２００２は、この論理ボリュームの容量である。論理ボリュームタイプ２００５は、論理ボリュームのタイプを表す。実施例１では、当該論理ボリュームが、内部ストレージ装置２３０に格納されているか、共有ストレージ装置１７５に格納されているかを示す。論理ボリュームRAIDタイプ２００３は、該当する論理ボリュームのRAIDタイプ、RAID0、RAID1などを指定する。RAID5のように、N台の容量に対し、１台の容量の冗長データを格納する場合、Nの具体的数値を指定するものとする。ただし、任意のRAIDタイプが指定できるわけでなく、少なくとも一つストレージグループがもつRAIDタイプである必要がある。割り当て範囲２０２６は、当該論理ボリュームに割り当てたストレージグループの識別子と、最も小さい番号のセグメントの番号を示している。セグメントビットマップは、最も小さい番号のセグメントを最初１ビットとして、以降の各セグメントに、当該論理ボリュームに割り当てられたか、そうでないかを表す。

【００６４】

第１キャッシュ管理情報ポインタ２０２２は、当該論理ボリュームを、スロット２１１００に相当する容量で分割したそれぞれの領域が、スロット２１１００が割り当てられている（キャッシュ２１０に格納されている）かを表す。割り当てられている場合、対応するキャッシュ管理情報２７５０がポイントされる。割り当てられていない場合、ヌル状態になる。実施例１では、ストレージ装置１６０に未書込みデータは、高信頼化のために、２重化してキャッシュ２１０に格納する。このため、第２キャッシュ管理情報ポインタ２０２３は、未書込みデータを２重化して格納するキャッシュ管理情報２７５０へのポインタである。

【００６５】

第１ストレージポインタ２０２４と第２ストレージポインタ２０２５は、当該論理ボリュームを、パリティグループに相当する容量で分割したそれぞれの領域が（ストレージボックス１３０に、ライトデータを直接転送している場合）、一時的にストレージ装置１６０に、二重に、データを書き込むために、確保した領域のアドレスを示す。アドレスには、ストレージボックス１３０の識別子、ストレージ装置１６０の識別子、ストレージ装置１６０のアドレスが示される。

【００６６】

図１０は、キャッシュ管理情報２７５０の構成である。キャッシュ管理情報２７５０は、スロット２１１００対応に存在する情報である。次キャッシュ管理情報ポインタ２７５１、割り当て論理ボリュームアドレス２７５２、ブロックビットマップ２７５３、更新ビットマップ２７５４、アクセス不可ビットマップ２７５５、最終アクセスアドレスフラグ２７５６、直接転送フラグ２７５７、直接転送中断フラグ２７５８、旧データ破壊フラグ２７５９から構成される。

【００６７】

次キャッシュ管理情報ポインタ２７５１は、データを格納していない状態のあるスロットに対応するキャッシュ管理情報２７５０をポイントし、当該スロットは次のデータを格

納する。
【００６８】
割り当て論理ボリュームアドレス２７５２は、対応するスロットにデータを格納したとき、どの論理ボリュームのどのアドレスから割当を開始する領域のデータを格納したかを示す。ブロックビットマップ２７５３は、この領域の中で、キャッシュ２１０、あるいは、ストレージ装置１６０のバッファに格納されたブロック（読み書きの最小単位）を示す。格納時にはＯＮとなる。
【００６９】
更新ビットマップ２７５４は、サーバ１１０から書込み要求を受け、まだ、ストレージ装置１６０に書き込んでいないブロックを示す。未書きこみは、ＯＮとなる。
【００７０】
アクセス不可ビットマップ２７５５は、障害などが発生していて、リード要求がきた場合、エラーを返すブロックを示す。ライト要求がきた場合、データを受付、正常終了すれば、読んでもよいデータとなるため、アクセス不可ビットマップ２７５５の対応するビットをオフにする。実施例１では、シーケンシャルライトに対して、データを直接、ストレージボックス１３０に転送するようにする。シーケンシャルライトは、アドレス順に、データが書き込まれる。
【００７１】
最終アクセスアドレスフラグ２７５６は、前回の要求で、アクセスされたアドレスを示しており、これにより、アクセスがアドレス順になっているかをチェックできる。また、ストレージボックス１３０に直接データの送信、あるいは送信の中止することができる。ビットマップ情報や最終アクセスアドレスフラグ２７５６については、直接データ転送を行わない従来のキャッシュに書き込む処理を実行していても、正しい値が格納されているものとする。
【００７２】
直接転送フラグ２７５７は、現在、このスロットに対応する領域は、サーバ１１０から、ストレージボックス１３０に直接データ転送が行われていることを示す。直接転送中止フラグ２７５８は、なんらかの理由により、直接転送が中止されたことを示すフラグである。旧データ破壊フラグ２７５９は、ストレージボックス１３０がパリティグループ全体のデータを受領後、パリティを生成し、パリティグループ全体のデータとパリティをストレージ装置１６０に書き込んでいる際に、障害が発生した場合、すでに、旧データを破壊してしまっている可能性がある際、当該パリティグループのすべてのキャッシュ管理情報２７５０の当該フラグをオンにする。
【００７３】
図１１は、空きキャッシュ管理情報ポインタ２６５０によって管理されるデータを格納されていないスロットに対応するキャッシュ管理情報２７５０の管理方式である。空きキャッシュ管理情報ポインタ２６５０が、データが格納されていないスロットに対応する先頭のキャッシュ管理情報２７５０をポイントし、次のキャッシュ管理情報２７５０を、次ぎキャッシュ管理情報ポインタ２７５１がポイントする。
【００７４】
図１２は、ストレージボックス情報２０５０の構成である。ストレージボックス情報２０５０は、ストレージボックス１３０ごとに設ける情報である。
【００７５】
ストレージボックス情報２０５０は、ストレージボックス識別子７０００、接続情報７００１、ストレージ台数７００２、接続ストレージ台数７００３、ボックスポート数７００４、ボックスポート識別子７００５、ボックスプロセッサ数７００６、ボックスプロセッサ識別子７００７、パス数７００８、パス識別子７００９、ストレージグループ数７０１０、ストレージグループ識別子７０１１から構成される。
【００７６】
ストレージボックス識別子７０００は、対応するストレージボックス１３０の識別子で

ある。接続情報７００１は、対応するストレージボックス１３０が、それぞれの実ストレージシステム１００に接続されているか、いないかを示す情報である。

【００７７】

ストレージ台数７００２は、当該ストレージボックス１３０に接続可能なストレージ装置１６０の台数である。接続ストレージ台数７００３は、実際に接続されているストレージ装置１６０の台数である。ボックスポート数７００４は、ネットワーク１２０に接続された当該ストレージボックス１３０のボックスポート１８０の数で、ボックスポート識別子７００５は、それぞれのボックスポート１８０ごとの識別子である。

【００７８】

ボックスプロセッサ数７００６は、当該ストレージボックス１３０内のボックスプロセッサ１７０の数である。ボックスプロセッサ識別子７００７は、それぞれのボックスプロセッサ１７０との識別子である。パス数７００８は、当該、ストレージボックス１３０内のストレージ装置１６０とボックスプロセッサ１７０の間のパスの数である。パス識別子７００９は、それぞれのストレージパス１９７ごとの識別子である。ストレージグループ数７０１０は、当該ストレージボックス１３０に含まれるストレージグループの数である。ストレージグループ識別子７０１１は、当該ストレージボックス１３０に含まれるストレージグループの識別子である。実施例１では、ストレージグループを構成するストレージ装置１６０は、一つのストレージボックス１３０に含まれるとする。ただし、本発明は、ストレージグループが複数のストレージボックス１３０のストレージ装置１６０で構成されても有効である。

【００７９】

図１３は、ストレージグループ情報２３００の形式を示す。ストレージグループ情報２３００は、ストレージグループID２３０１、ストレージグループRAIDタイプ２３０２、セグメント数２３０３、割り当て権限セグメント数２３０９、空きセグメント数２３０４、割り当て権限セグメントビットマップ２３０８、空きセグメントビットマップ２３０７、ストレージ装置ポインタ２３０５から構成される。

【００８０】

ストレージグループID２３０１は、当該ストレージグループ情報２３００の識別子である。この識別子は、ストレージボックス１３０の識別子も含む。ストレージグループRAIDタイプ２３０２は、当該ストレージグループのRAIDタイプである。実施例１におけるRAIDタイプは、論理ボリュームRAIDタイプ２００３を説明したときに述べたとおりである。実施例１では、ストレージコントローラ２００は、容量仮想化機能をもたないので、論理ボリュームを定義したときに、その容量分の領域が確保される。実施例１では、ストレージグループ情報２３００の容量は、セグメントという単位に分割されているとする。論理ボリュームの容量が定義されると、その容量以上で最小の数のセグメントが、確保されることになる。セグメント数２３０３は、当該ストレージグループのセグメントの数を示す。割り当て権限セグメント数２３０９は、その中で、当該実ストレージシステム１００が、割り当て権限をもっているセグメントの数を示す。ストレージボックス１３０の中のストレージグループの場合、複数の実ストレージシステムにシェアされるので、それぞれ、実ストレージシステム１００ごとに、割り当て権限をもつセグメントの集合を決めておく。割り当て権限ビットマップ２３０８は、それぞれのセグメントについて、当該実ストレージシステム１００が、割り当て権限をもっているか、いないかを表す。空きセグメント数２３０４は、該ストレージグループが割り当て権限をもっているセグメントの中で、空いた状態のセグメントの数を示す。空きセグメントビットマップ２３０７は、該ストレージグループが割り当て権限をもっているセグメントの中で、それぞれのセグメントが、空いた状態か、割り当て済みであるかを示す情報である。ストレージ装置ポインタ２３０５の数は、当該ストレージグループに属するストレージ装置１６０の数であるが、これは、ストレージグループRAIDタイプ２３０２によって決まる値である。ストレージ装置ポインタ２３０５は、当該ストレージグループに属するストレージ装置１６０の識別子である。

【００８１】

図１４は、ストレージ装置情報２５００のフォーマットである。ストレージ装置情報２５００のフォーマットは、ストレージ装置識別子２５１０、接続パス数２５０１、接続パス識別子（接続パス）２５０２、ストレージタイプ２５０３、容量２５０４、一時書込み領域数２５０５、一時書込み領域アドレス２５０６、使用中フラグ２５０７である。

【００８２】

ストレージ装置識別子２５１０は、当該ストレージ装置１６０の識別子である。接続パス数２５０１は、当該ストレージ装置１６０とボックスプロセッサ１７０との間のストレージパス１９７の数である。接続パス識別子２５０２は、接続されているパスの識別子を示す。ストレージタイプ２５０３は、当該ストレージ装置１６０が、ＨＤＤ、フラッシュメモリ、などのどんな記憶媒体なのかを示す。容量２５０４は、当該ストレージ装置の容量である。なお、実施例１では、ストレージグループを構成するストレージ装置１６０のストレージタイプ２５０３と容量２５０４は、等しいとする。一時書込み領域数２５０５は、ストレージボックス１３０にデータを直接書き込んだときに、パリティグループ全体のデータを受領するため一時的に書き込んでおくために確保している領域の数である。実施例１においては、領域の大きさは、一つのパリティグループの容量である。一時書込み領域アドレス２５０６は、一時書込み領域のアドレスである。実施例１では、二つのストレージ装置の一時書込み領域を選択し、それぞれの領域にパリティグループのデータを書き込む。ただし、本発明は、一つの一時書込み領域を複数のｎ台のストレージ装置１６０から構成しておき、二つの書込み領域を選択し、２ｎ台のストレージ装置に書き込んでも有効である。なお、一時書込み領域は、サーバ１１０からの書込みが、直接指定されない領域であるものとする。使用中フラグ２５０７は、一時書込み領域対応に存在するフラグで、当該一時書込み領域が、使用中であることを示す。

【００８３】

次に、上記に説明した管理情報を用いて、ストレージコントローラ２００が実行する動作の説明を行う。まず、ストレージコントローラ２００の動作を説明する。ストレージコントローラ２００の動作は、ストレージコントローラ２００内のプロセッサ２６０が実行し、そのプログラムは、メモリ２７０に格納されている。図１５は、メモリ２７０内に、格納された実施例１に関するプログラムが示されている。実施例１に関するプログラムは、ライト要求受付部４１００、ライト異常終了対応処理部４２００、リード処理実行部４３００である。

【００８４】

図１６Ａ、Ｂは、ライト要求受付部４１００の処理フローである。ライト要求受付部４１００は、ストレージコントローラ２００が、サーバ１１０からライト要求を受け付けたときに実行される。

【００８５】

図１６Ａから説明する。

ステップＳ６０００：プロセッサ２６０は、受け取ったライト要求で指定された仮想論理ボリュームを、仮想論理ボリューム情報２０５５によって、論理ボリュームに変換する。

【００８６】

ステップＳ６００１ａ：受け取ったライト要求のアドレスが、パリティグループの先頭のアドレスかを判断する。そうでなければ、ステップＳ６００７へジャンプする。

ステップＳ６００１ｂ： キャッシュ管理情報を割り当てる。

【００８７】

ステップＳ６００２：先頭の場合、一つ前の領域にキャッシュ管理情報２７５０が割りあてられているかをチェックする。割り当てられていない場合、直接転送を行わないので、キャッシュ２１０にライトデータを転送する従来の処理を実行するので、ステップＳ６０２７へジャンプする。

【００８８】

ステップＳ６００３：当該パリティグループ全体のキャッシュ管理情報２７５０がミスしているかをチェックする。割り当てられているものが一つでも存在する場合、実施例１

では直接転送に入らないので、ステップＳ６０２７へジャンプする。なお、割り当てられているものがあった場合にも、直接転送を実行しても、本発明は有効である。

【００８９】

ステップＳ６００４：一つ前のキャッシュ管理情報２７５０の最終アクセスアドレスをチェックし、今回のアクセスが、シーケンシャルアクセスになっているかをチェックする。なっていない場合、直接転送を行わないので、キャッシュ２１０にライトデータを転送する従来の処理を実行するので、ステップＳ６０２７へジャンプする。

【００９０】

ステップＳ６００５：一つ前のキャッシュ管理情報２７５０のすべての更新ビットマップ２７５４、アクセス不可ビットマップ２７５５がオフかをチェックする。すべて、オフであれば、このキャッシュ管理情報２７５０を、空きキャッシュ管理情報ポインタ（キュー）２６５０に戻す。

【００９１】

ステップＳ６００６：ここで、ストレージボックス１３０に直接ライトデータを転送する準備処理に入る。ここでは、空きキャッシュ管理情報ポインタ２６５０から、キャッシュ管理情報２７５０を、パリティグループに相当する領域の、第１キャッシュ管理情報ポインタ２０２２と、第２キャッシュ管理情報ポインタ２０２３に割り当てる。そして、それぞれのキャッシュ管理情報２７５０の直接転送フラグ２７５７をオンにする。また、ブロックビットマップ２７５３，更新ビットマップ２７５４、アクセス不可ビットマップ２７５５は、オールクリアする。次に、第１ストレージポインタ２０２４と第２ストレージポインタ２０２５に、当該パリティグループを格納するストレージボックス１３０のストレージ装置１６０の一時書込み領域アドレス２５０６の中で、使用中フラグ２５０７がオフのもののなかから、異なるストレージ装置１６０の一時書込み領域アドレス２５０６のふたつを選んで、設定する。選んだら、使用中フラグ２５０７をオンにする。（なお、空き状態にあるスロット管理情報や未使用の一時書込み領域が不足した場合、割り当てをやめ、直接転送は実行せず、ステップＳ６０２７へジャンプし、従来の処理を実行する。）この後、ステップＳ６０１５にジャンプする。

【００９２】

ステップＳ６００７：このステップは、パリティグループの先頭以外のアドレスのライト要求を受け付けた場合に実行される。まず、対応する領域に、キャッシュ管理情報２７５０が割り当てられているかをチェックする。割り当てられていない場合、直接データ転送は実行しないので、ステップＳ６０１２を判定し、ステップＳ６０２７へジャンプする。

【００９３】

ステップＳ６００８：以下では、アクセスがシーケンシャルになっているかをチェックする。本ステップでは、アクセスが対応するスロットの先頭かをチェックする。先頭でない場合、ステップＳ６０２７へジャンプする。

【００９４】

ステップＳ６００９：先頭のアドレスの場合、一つ前の領域に、キャッシュ管理情報２７５０が割り当てられているかをチェックする。割り当てられていない場合、直接データ転送は実行しないので、ステップＳ６０２７へジャンプする。

【００９５】

ステップＳ６０１０：割り当てられている場合、一つ前のキャッシュ管理情報２７５０の直接転送フラグ２７５７をチェックする。オフの場合、直接データ転送は実行しないので、ステップＳ６０２７へジャンプする。

【００９６】

ステップＳ６０１１：次に、アクセスがシーケンシャルかを確認するため、一つ前のキャッシュ管理情報２７５０の最終アクセスアドレスフラグ２７５６を、チェックする。シーケンシャルになっていない場合、直接転送を中止するため、ステップＳ６０２７にジャンプする。加えて、直接転送中止フラグ２７５８がオンになっているかをチェックする。オンの場合、直接転送を中止するため、ステップＳ６０２７にジャンプする。オフの場合

、直接転送を実行するため、ステップＳ６０１５へジャンプする。
【００９７】
ステップＳ６０１２：当該キャッシュ管理情報２７５０の直接転送フラグ２７５７をチェックする。オフの場合、直接データ転送は実行しないので、ステップＳ６０２７へジャンプする。
【００９８】
ステップＳ６０１３：次に、アクセスがシーケンシャルかを確認するため、当該キャッシュ管理情報２７５０の最終アクセスアドレスフラグ２７５６を、チェックする。シーケンシャルになっていない場合、直接転送を中止するため、ステップＳ６０２７にジャンプする。
【００９９】
ステップＳ６０１４：次に、直接転送中止フラグ２７５８がオンになっているかをチェックする。オンの場合、直接転送を中止するため、ステップＳ６０２７にジャンプする。オフの場合、以下のステップで直接転送に入る。
【０１００】
ステップＳ６０１５：本ステップでは、ストレージボックス１３０に、直接転送を行うように指示する。この際、ライトデータを、本来書き込むアドレスと二つの一時的に書き込むアドレスを指示する。また、本来書き込むアドレスをサーバ１９８に送る。
【０１０１】
図１６Ｂの説明を行う。
ステップＳ６０１６：完了をまつ。
ステップＳ６０１７：正常終了か異常終了かを判別する。異常終了の場合、まず、サーバ１１０に異常終了報告を行う。また、パリティグループの先頭以外のアクセスの場合、ステップＳ６０２６へジャンプする。（パリティグループの先頭のアクセスの場合、まだ、正常終了報告を返したライトデータはないので、後処理を行う必要がない。）なお、異常終了の場合、ある時間たっても、ストレージボックス１３０からの完了報告がなく、ストレージコントローラ２００の時間監視で、タイムオーバを起こした場合も含めるとする。
【０１０２】
ステップＳ６０１８：一時領域に格納したデータに対応する、ブロックビットマップ２７５３，更新ビットマップ２７５４をセットする。
【０１０３】
ステップＳ６０１９：当該ライト要求のアクセスアドレスを、最終アクセスアドレスに記憶する。
ステップＳ６０２０：サーバ１１０に正常終了を報告する。
ステップＳ６０２１：パリティグループ全体の書込みが完了したかをチェックする。完了していない場合、処理を完了する。
ステップＳ６０２２：完了した場合、ストレージボックス１３０に、パリティを生成して、ライトデータとパリティを書き込むべき領域に書き込むよう指示する。
ステップＳ６０２３：完了をまつ。
【０１０４】
ステップＳ６０２４：正常終了か、異常終了かを判別する。異常終了の場合、直接転送を中止するための処理を実行するため、ステップＳ６０２６へジャンプする。なお、異常終了の場合、ある時間たっても、ストレージボックス１３０からの完了報告がなく、ストレージコントローラ２００の時間監視で、タイムオーバを起こした場合も含めるとする。また、本ステップで異常終了した場合、パリティグループ全体のデータとパリティをストレージ装置１６０に書き込んでいる際に、障害が発生した場合、すでに、旧データを破壊してしまっている可能性があるので、当該パリティグループのすべてのキャッシュ管理情報２７５０の旧データ破壊フラグ２７５９をオンにする。
【０１０５】
ステップＳ６０２５：ここでは、当該パリティグループに割り当てたキャッシュ管理情

報２７５０と一時書込み領域を開放する。具体的には、最後のキャッシュ管理情報２７５０以外のキャッシュ管理情報２７５０は、空きキャッシュ管理情報キュー２６５０に戻す。また、キューに戻すキャッシュ管理情報２７５０のブロックビットマップ等のビットマップ情報や最終アクセスアドレス、フラグ情報は、すべてオフにする。加えて、最終のキャッシュ管理情報２７５０は、最終アドレスアドレスフラグ２７５６以外の情報は、すべてオフにする。この後、処理を終了する。

【０１０６】

ステップＳ６０２６：異常処理が発生したので、直接転送の後処理を実行するため、直接転送異常処理をコールする。この後、処理を終了する。

【０１０７】

ステップＳ６０２７：キャッシュにライトデータを受け取る処理を実行する。この際、当該領域に、キャッシュ管理情報２７５０が割り当てられていない場合、キャッシュ管理情報２７５０を割り当てることになる。また、今回アクセスされたアドレスを最終アクセスアドレスに記憶する。加えて、ライト要求が正常終了した場合、当該要求でライトされた領域に対応するアクセス不可ビットマップ２７５５がオンの場合、オフする。さらに、当該キャッシュ管理情報２７５０のすべてのアクセス付加ビットマップがオフになったら、直接転送中止フラグ２７５８をオフする。

【０１０８】

図１７は、ライト異常終了対応処理部４２００の処理フローである。本処理も、プロセッサ２６０が、適宜実行する処理である。

【０１０９】

ステップＳ７０００：ストレージボックス１３０が、無応答で異常終了したのかをチェックする。無応答で処理を終了していない場合、ステップＳ７００４へジャンプする。

【０１１０】

ストレージボックス１３０からストレージコントローラ２００へ、障害通知があった場合は、このステップで、「ＮＯ」となり、ステップＳ７００４にジャンプする。障害通知としては、ストレージボックス１３０で、冗長データを生成することができないとき、ライトデータをストレージコントローラ２００に送れる場合が含まれる。

【０１１１】

また、ライトデータの中で、障害により、ストレージボックス１３０がストレージコントローラ２００に送ることができないデータが存在する場合、その旨とそのデータを書き込むよう指定された領域のアドレスを、ストレージコントローラ２００に伝える場合も含まれる。

【０１１２】

また、ライトデータ或いは冗長データを、障害により、記憶媒体に書き込むよう指定された領域に書き込めないとき、ストレージコントローラ２００に送る場合も含まれる。

【０１１３】

ステップＳ７００１：無応答で処理を終了した場合、ストレージボックス１３０が電源ダウン等で、通信できない可能性がある。本ステップでは、一度、ストレージボックスにリセット要求を発行する。

【０１１４】

ステップＳ７００２：リセット要求が完了した場合、ステップＳ７００４へジャンプする。

【０１１５】

ステップＳ７００３：ストレージボックス１３０がダウンしている等の原因があるので、ストレージ管理者、保守員等に、連絡して、ストレージボックス１３０が立ち上がるのをまつ。立ち上がった後、ステップＳ７００４から実行を開始する。

【０１１６】

ステップＳ７００４：当該パリティグループに対応したキャッシュ管理情報２７５０のブロックビットマップ２７５３，更新ビットマップ２７５４がオンのデータを格納した、

一時書込み領域のアドレスを算出する。また、当該パリティグループに対応したキャッシュ管理情報２７５０の直接転送中止フラグ２７５８をオンにする。

【０１１７】

ステップＳ７００５：ストレージボックス１３０に、このアドレス（ステップＳ７００４で計算したアドレス）を通知して、これらのデータを、ストレージコントローラ２００に送るよう指示する。なお、実施例１では、ストレージボックス１３０には、サーバ１１０から受け取ったデータを転送するよう、指示しているが、ストレージボックス１３０で、すでにパリティを生成している場合、パリティもストレージボックス１３０から受け取って、キャッシュ２１０に格納するようにしてもよい。

【０１１８】

ステップＳ７００６：完了をまつ。ある時間待って、完了が返って、こなかった場合、ステップＳ７０００へジャンプする。

【０１１９】

ステップＳ７００７：正常終了か異常終了かをチェックする。ストレージボックス１３０との間で、一部のデータが転送に失敗した場合や、ストレージボックス１３０から、一時書込み領域から読み出せなかったデータがあった旨の報告を受けた場合も、ステップＳ７００９へジャンプする。

【０１２０】

ステップＳ７００８：ストレージボックスとの間で、すべてのデータの転送が正常終了して、かつ、ストレージボックスから一時書込み領域のすべてのデータを読み出せた旨をうけとった場合、受け取ったデータを対応するスロット管理情報２７０５に対応したキャッシュ２１０の領域に２重に書き込む。この後、処理を終了する。

【０１２１】

ステップＳ７００９：転送に成功して、かつ、一時書込み領域から正常に読み出せたデータのみ、キャッシュ管理情報２７５０に対応した領域に、２重に書き込む。

【０１２２】

ステップＳ７０１０：転送に失敗したデータあるいは、一時書込み領域から読み出せなかったデータに対応するアクセス不可ビットマップ２７５５をオンにしする。また、ライトデータ、データを書き込むよう指定された領域を記憶する。

【０１２３】

当該パリティグループに対応するアクセス不可ビットマップ２７５５アクセス不可ビットマップ２７５５にオンがない場合、（すべてのデータがキャッシュ２１０に正常に格納できた場合）、ストレージコントローラ２００がパリティを生成して、データとパリティを、ストレージ装置１６０に書き込むことができる。アクセス不可ビットマップ２７５５がオンのビットをもつスロットの場合、このアクセス不可ビットマップ２７５５がオンに対応する領域に、新しいライト要求が入り、キャッシュ２１０にデータを正常に格納できれば、対応するアクセス不可ビットマップ２７５５をオフにできる。当該スロットに対応するキャッシュ管理情報２７５０にすべてのアクセス不可ビットマップ２７５５がオフになった場合、旧データ破壊フラグ２７５９がオフであれば、ストレージコントローラ２００が持つ、キャッシュ２１０に旧データや旧パリティを格納して、これらのデータから、新パリティを生成して、データとパリティをストレージ装置１６０に格納する論理を適用できる。旧データ破壊フラグ２７５９がオンの場合、パリティの生成に旧データが使用できないので、当該パリティグループに対応するすべてのキャッシュ管理情報２７５０のすべてのアクセス不可ビットマップ２７５５がオフになった場合、ストレージコントローラ２００が、残りのパリティグループのデータをキャッシュ２１０に格納し、パリティグループ全体のデータから、パリティを生成して、更新データとパリティを、ストレージ装置１６０に書き込むことができる。また、旧データ破壊２７５９がオフの場合にも、残りのパリティグループのデータをキャッシュ２１０に格納し、パリティグループ全体のデータから、パリティを生成してもよい。

【０１２４】

図１８は、リード処理実行部４３００の処理フローである。実施例１ではストレージボックス１３０からサーバに、要求されたデータを直接転送するようにする。もちろん、本発明は、リード要求で指定されたデータを、ストレージボックス１３０からストレージコントローラ２００を経由して、サーバに送っても、本発明は有効である。

【０１２５】

ステップＳ５０００：プロセッサ２６０は、受け取ったリード要求で指定された仮想論理ボリュームを、仮想論理ボリューム情報２０５５によって、論理ボリュームに変換し、対応する論理ボリューム情報２０００を獲得する。

【０１２６】

ステップＳ５００１：受け取ったリード要求のアドレスから、当該領域にキャッシュ管理情報２７５０が割り当てられているかをチェックする。割り当てられている場合、ステップＳ５００６へジャンプする。

【０１２７】

ステップＳ５００２：ストレージボックス１３０に、指定されたデータを、サーバ１１０に送るよう指示する。

ステップＳ５００３：完了をまつ。

ステップＳ５００４：サーバ１１０に完了報告を行い、処理を終了する。

ステップＳ５００６：要求されたデータに対応するアクセス不可ビットマップ２７５５のビットがオンになっていないかをチェックする。オフであれば、ステップＳ５００７へジャンプする。

ステップＳ５００７：サーバ１１０に対し、当該データがアクセスできないという異常報告を行う。この後処理を終了する。

ステップＳ５００８：直接転送フラグ２７５７がオンで、直接転送中止フラグ２７５８がオフかをチェックし、条件が満足した場合、ステップＳ５０１２へジャンプする。ストレージコントローラ２００が、記憶した領域に対し、サーバ１１０からリード要求を受けつけたとき、リード要求を、異常終了させる。

ステップＳ５００８：対応するブロックビットマップ２７５３がオンかをチェックする。オフの場合、ステップＳ５０１２へジャンプする。

ステップＳ５００９：キャッシュから、要求されたデータを、サーバに転送する。

ステップＳ５０１０：転送が完了するのを待つ。

ステップＳ５０１１：サーバ１１０に対し、完了報告を行う。

ステップＳ５０１２：ストレージボックス１３０に、一時書込み領域の中の対応する領域から、サーバ１１０に送るよう指示する。

ステップＳ５０１３：完了をまつ。

ステップＳ５０１４：サーバ１１０に完了報告を行い、処理を終了する。

【０１２８】

次に、ストレージコントローラ２００の指示にしたがって、ストレージボックス１３０のボックスプロセッサ１７０が実行する動作の説明を行う。実行するプログラムは、ボックスメモリ１８１に格納されている。図１９は、ボックスメモリ１８１内に、格納された実施例１に関するプログラムが示されている。実施例１に関するプログラムは、ライトデータ受領部４４００、ライトデータ書込み部４５００，一時書込み領域転送部４６００，リードデータ直接転送部４７００である。

【０１２９】

図２０は、ライトデータ受領部４４００の処理フローである。ライトデータ受領部４４００は、ストレージコントローラ２００の指示にしたがって、サーバ１１０から、ライトデータを受け取り、一時書込み領域にデータを書き込む処理である。

ステップＳ８０００：サーバ１１０に、データを送るよう指示する。

ステップＳ８００１：転送の完了を待つ。

ステップＳ８００２：正常終了であれば、ステップＳ８００４へジャンプする。

ステップＳ８００３：異常終了報告を、ストレージコントローラに返し、処理を終了す

る。
ステップＳ８００４：受け取ったライトデータを、ストレージコントローラ２００から指定された、二つの一時書込み領域のアドレスに書き込む。
ステップＳ８００５：書込みが完了するのをまつ。
ステップＳ８００６：正常終了であれば、ステップＳ８００８へジャンプする。
ステップＳ８００７：異常終了報告を、ストレージコントローラに返し、処理を終了する。
ステップＳ８００８：正常終了報告を、ストレージコントローラに返し、処理を終了する。
【０１３０】
図２１は、ライトデータ書込み部４５００の処理フローである。ライトデータ書込み部４５００は、ストレージコントローラ２００の指示にしたがって、一時書込み領域からライトデータを読み出し、パリティを生成し、ライトデータとパリティを、ストレージコントローラ２００から、指示されたストレージ装置１６０の領域に書き込む。
ステップＳ９０００：ストレージコントローラ２００から指定された二つの一時書込み領域の一つから、データを読み出す。
ステップＳ９００１：処理の完了をまつ。
ステップＳ９００２：すべてのデータの読み出しが正常終了したかをチェックする。正常終了していれば、ステップＳ９００７へジャンプする。
ステップＳ９００３：もう一つの一時書込み領域から、異常終了したデータだけを読み出す。
ステップＳ９００４：完了をまつ。
ステップＳ９００５：指定したすべてのデータの読み出しが正常終了したかをチェックし、正常終了した場合、ステップＳ９００７へジャンプする。
ステップＳ９００６：ストレージコントローラ２００に、異常終了を報告し、処理を完了する。
ステップＳ９００７：読み出したライトデータから、パリティを生成する。
ステップＳ９００８：ライトデータとパリティを、ストレージコントローラ２００が指定したストレージ装置１６０の領域に書き込む。
ステップＳ９００９：完了をまつ。
ステップＳ９０１０：すべての書込みが正常終了した場合、ストレージコントローラ２００に、正常終了報告を行い、一つでも、異常終了があった場合、異常終了報告をストレージコントローラ２００に行う。
【０１３１】
図２２は、一時書込み領域転送部４６００の処理である。一時書込み領域転送部４６００は、ストレージコントローラ２００の指示にしたがって、一時書込み領域からライトデータを読み出し、ストレージコントローラ２００に送る。一時書込み領域から、データを読み出す処理は、ライトデータ書込み処理部のステップＳ９０００からステップＳ９００５までの処理と同様である。読み出したデータをストレージコントローラ２００に送る以下の処理のみが異なる。
ステップＳ１００００：正常に読み出せなかったデータが存在する処理である。正常に読み出せたデータ（ライトデータと冗長データ）とそのアドレスをすべてストレージコントローラ２００に送る。また、読み出せなかったデータについては、読み出せなかった旨とそのアドレスをストレージコントローラ２００に通知する。
【０１３２】
その後、処理を完了する。
ステップＳ１０００１：すべてのデータを読み出せたことになるので、データとそのアドレスを、ストレージコントローラ２００に送る。この後、処理を終了する。
【０１３３】
以上の通り、障害により、ストレージボックス１３０で、冗長データを生成することが

できないとき、ライトデータをストレージコントローラ２００に送る。
【０１３４】
　また、ライトデータの中で、障害により、ストレージボックス１３０がストレージコントローラ２００に送ることができないデータが存在する場合、その旨とそのデータを書き込むよう指定された領域のアドレスを、ストレージコントローラ２００に伝える。
【０１３５】
　また、ライトデータ或いは冗長データを、障害により、記憶媒体に書き込むよう指定された領域に書き込めないとき、ストレージコントローラ２００に送る。
【０１３６】
　図２３は、リードデータ直接転送部４７００の処理フローである。リードデータ直接転送部４７００は、ストレージコントローラ２００からの指示によって、指定されたデータをサーバ１１０に送る処理である。実施例１では、一時書込み領域の管理を、ストレージコントローラ２００が行っているため、サーバ１１０が読み書きするデータを恒久的に格納した領域か、パリティを作成するための一時的な書込み領域の区別を、ストレージボックス１３０は、認識していない。このため、サーバ１１０が読み書きするデータを恒久的に格納した領域か、パリティを作成するための一時的な書込み領域に係わらず、ストレージボックス１３０は、指定された領域のデータを送るだけである。以上から、サーバ１１０が読み書きするデータを恒久的に格納した領域のデータを送る処理と、パリティを作成するための一時的な書込み領域のデータを送る処理の処理フローは、同一となり、いずれも、リードデータ直接転送部４７００が実行する。
　ステップＳ１１０００：指定された領域のデータをサーバ１１０に送る。
　ステップＳ１１００１：処理が完了するのを待つ。
　ステップＳ１１００２：正常終了した場合、正常終了報告を、異常終了した場合、異常終了報告を、ストレージコントローラ２００に返す。この後、処理を終了する。
【実施例２】
【０１３７】
　次に、実施例２について説明する。実施例２では、ストレージコントローラ２００が容量仮想化機能（Ｔｈｉｎ　Ｐｒｏｖｉｓｉｏｎｉｎｇ）をもつとする。また、ストレージコントローラ２００とストレージボックス１３０が連携して、圧縮・重複排除機能を実現する。圧縮・重複排除機能を実現する場合、ライト処理は、ログストラクチャ方式を採る。この場合、ライトデータは、追記型となり、新しいアドレスに書き込まれる。
【０１３８】
　図２４は、実施例２における情報システムの構成を示す。実施例２における情報システムと実施例１の情報システムの相違は、ストレージボックス１３０の構成が異なる点である。実施例２におけるストレージボックス１３０は、ストレージコントローラ２００と連携して、実施例１におけるストレージボックス１３０の構成要素に、複数の圧縮・伸長回路１９５０、ハッシュ回路１９６０を備える点である。ただし、ボックスプロセッサ１７０で、これらの処理を実行しても、本発明は有効である。また、実施例２では、圧縮・伸長回路１９５０、ハッシュ回路１９６０は、ボックスプロセッサ１７０対応に存在するものとする。ただし、圧縮・伸長回路１９５０、ハッシュ回路１９６０は、ボックスプロセッサ１７０対応に存在しない場合にも、本発明は有効である。
【０１３９】
　実施例２のサーバポート情報１９８のフォーマットは、第１のサーバポート情報と同じである。第２の実ストレージシステム１００の構成は、実施例１と同じである。
【０１４０】
　図２５は、実施例２における実ストレージシステム１００の共有メモリ２２０の中の実施例２に関する情報を示しており、ストレージシステム情報２０６０、仮想論理ボリューム情報２０８５、他ストレージシステム情報２０７０、論理ボリューム情報２０００、ストレージボックス情報２０５０、ストレージグループ情報２３００、ストレージ装置情報２５００、キャッシュ管理情報２７５０、空きキャッシュ管理情報ポインタ２６５０、に

よって構成される。また、実施例２では、新たに、実ページ情報２１００、空きページ管理情報ポインタ２２００、仮想ページ容量２６００が新しく加わる。これらの情報は、容量仮想化機能をサポートしたために必要となる情報である。この中で、ストレージシステム情報２０６０は、実施例１と同様である。他ストレージシステム情報２０７０の形式も、実施例１と同様である。仮想論理ボリューム情報２０８５の形式も、実施例１と同様である。

【０１４１】

図２６は、実施例２における論理ボリューム情報２０００の形式を示したものである。実施例２でも、サーバ１１０がデータをリード・ライトする記憶装置は、仮想論理ボリュームである。また、サーバ１１０は、仮想論理ボリュームのID、仮想論理ボリューム内のアドレス、読み書きしたいデータの長さを指定して、リード要求やライト要求を発行する。実ストレージシステム１００は、サーバ１１０から読み書き要求を受け取ると、仮想ボリューム情報２０８５により、対応する論理ボリュームの識別子を認識する。実施例２でも、論理ボリュームの識別子は、実ストレージシステム１００内でユニークな情報である。

【０１４２】

論理ボリューム情報２０００は、論理ボリュームごとに存在する情報である。この情報は、論理ボリューム識別子２００１、論理容量２００２、論理ボリュームタイプ２００５、論理ボリュームRAIDタイプ２００３、第１キャッシュ管理情報ポインタ２０２２、第２キャッシュ管理ポインタ２０２３、マッピング情報（マッピングアドレス）２００４、圧縮後データ長２０１８、ハッシュ値２０１９、ログストラクチャポインタ２００５、ログストラクチャ書込みポインタ２００６、第１圧縮前データ格納領域ポインタ２００７、第２圧縮前データ格納領域ポインタ２００８、第１圧縮後データ格納領域ポインタ２００９、第２圧縮後データ格納領域ポインタ２０１０、第１無効圧縮前ビットマップ２０２０、第２圧縮前無効ビットマップ２０２１、第１無効圧縮後ビットマップ２０１１、第２圧縮後無効ビットマップ２０１２、第１圧縮前書込みポインタ２０１３、第２圧縮前書込みポインタ２０１４、第１圧縮後書込みポインタ２０１５、第２圧縮後書込みポインタ２０１６、論理ボリューム直接転送中止フラグ２０３０より構成される。

【０１４３】

論理ボリューム識別子２００１、論理容量２００２、論理ボリュームタイプ２００５、論理ボリュームRAIDタイプ２００３、第１キャッシュ管理情報ポインタ２０２２、第２キャッシュ管理情報ポインタ２０２３までの情報は、第１の実施例と同様である。

【０１４４】

第２の実施例では、ストレージコントローラ２００は、圧縮・重複排除機能をもつ。圧縮・重複排除機能をもつ場合、ライトデータは、ログストラクチャストリームで、書込み順に追記される。実施例２では、論理ボリューム対応に、ログストラクチャストリームをもつ。実施例２では、論理ボリュームごとにログストラクチャストリームは一つであるが、高速化のために、論理ボリュームごとに複数もっても、本発明は有効である。また、本発明は、ログストラクチャストリームが、論理ボリューム対応でなくとも有効である。マッピング情報（マッピングアドレス）２００４は、論理ボリュームの圧縮・重複排除単位が、どの実ページのどの相対アドレスに割り当てられているかを示す情報である。また、本発明では、ストレージボックス１３０にライトデータが直接転送され、パリティが生成されるまでは、一時的な領域に二重に格納されている。その場合、そのアドレスが二つ示される。割り当てられていない場合、ヌル状態となる。

【０１４５】

また、実施例２では、圧縮後データ長２０１８はその管理単位の圧縮後のデータ長、ハッシュ値２０１９はその管理単位のハッシュ値を示す。

【０１４６】

実施例２では、ストレージコントローラ２００は、容量仮想化機能をもつ。この場合、データはページに書き込まれる。ログストラクチャ書込みポインタ２００６には、ログストラクチャストリームに割り当てられた実ページ情報２１００が、キュー状に接続されて

いるものである。ログストラクチャ書込みポインタ２００６は、次に、データを書き込むページのアドレスを示すものである。実施例２では、ライトデータは、直接、ストレージボックス１３０に転送される。この際、ストレージボックス１３０では、データが二重化され、パリティグループのデータが揃うまで、パリティが生成されない。また、実施例２では、ストレージコントローラ２００とストレージボックス１３０が、連携して圧縮・重複排除を行うため、実際に記憶装置に書き込むデータの量は、受け取ったライトデータに比べ小さくなる。

【０１４７】

また、データはすべて２重化するので、実施例２では、圧縮重複・排除前のデータを格納するための領域をｍ×２面で２ｍ個割り当てているものとする。第１圧縮前データ格納領域ポインタ２００７、第２圧縮前データ格納領域ポインタ２００８は、二重に圧縮前データを格納する領域で、それぞれｍ個の領域に対応し、それぞれがｍ個存在する。第１圧縮後データ格納領域ポインタ２００９、第２圧縮後データ格納領域ポインタ２０１０は、二重に圧縮前データを格納する領域で、それぞれ２個の領域に対応し、それぞれが２個存在する。

【０１４８】

また、実施例２では、ログストラクチャストリームで、ライトデータは新しい領域に書き込まれるので、それまでのデータは無効になる。これは、まだ、第１圧縮前データ格納領域ポインタ２００７、第２圧縮前データ格納領域ポインタ２００８，第１圧縮後データ格納領域ポインタ２００９、第２圧縮後データ格納領域ポインタ２０１０が示す領域に存在する間にも発生する可能性がある。この場合、対応する領域の第１無効圧縮前ビットマップ２０２０、第２圧縮前無効ビットマップ２０２１、第１無効圧縮後ビットマップ２０１１、第２圧縮後無効ビットマップ２０１２のビットなどをオンにすることになる。

【０１４９】

また、次にサーバ１１０から受け取ったライトデータを書き込むべきアドレスが、第１圧縮前書込みポインタ２０１３、第２圧縮前書込みポインタ２０１４である。また、次に圧縮・重複排除後のデータを書き込むべきデータのアドレスが、第１圧縮後書込みポインタ２０１５、第２圧縮後書込みポインタ２０１６である。実施例２では、ライトデータは論理ボリューム単位に、ログストラクチャストリームに書き込まれ、基本的にはストレージボックス１３０に直接転送される。このため、障害が発生すると、論理ボリューム単位で、直接転送が中止になる。直接転送中止フラグ２７５８は、当該論理ボリュームのライトデータの直接転送が中止になったことを示す。

【０１５０】

キャッシュ管理情報２７５０は、基本的には、実施例１と同様である。ただし、最終アクセスアドレスフラグ２７５６はなくともよい。

【０１５１】

空きキャッシュ管理情報ポインタ２６５０の形式は、実施例１と同様である。

【０１５２】

実施例２の特徴は、実ストレージステム１００は、容量仮想化機能をサポートしている点である。通常、容量仮想化機能において、記憶領域の割り当て単位は、ページと呼ばれる。また、論理ボリュームは、通常、サーバ１１０が読み書きをする論理的な記憶装置である。ただし、本発明では、キャッシングのために使用する記憶装置の領域を論理ボリュームとして定義するのが特徴である。そして、この論理ボリュームにも、容量仮想化機能を適用して、ページを割り当てることで、実記憶領域を確保する。

【０１５３】

なお、実施例２では、論理ボリュームの空間は、仮想ページという単位で、分割されているものとし、実際の記憶装置グループは、実ページという単位で分割されているものとする。容量仮想化においては、論理ボリュームの記憶容量を、実際の記憶媒体の容量よりも大きく見せる。このため、仮想ページの数のほうが、実ページの数より大きいのが、一般的である。容量仮想化機能を実現した場合、ストレージコントローラ２００がサーバ１

１０からのライト不要で書込みを指示されたアドレスを含む仮想ページに実ページが割り当てていないとき、実ページを割り当てる。仮想ページ容量２６００は、仮想ページの容量である。

【０１５４】

しかし、実施例２では、仮想ページ容量２６００と実ページの容量は等しいというわけではない。というのは、実ページの容量は、RAIDのタイプにより異なってくる冗長データを含むためである。したがって、実ページの容量は、その実ページが割り当てられた記憶装置グループのRAIDタイプにより決まる。たとえば、RAID1のようにデータを２重に書き込む場合、実ページの容量は、仮想ページ容量２６００の２倍になる。RAID5のように、N台の記憶装置の容量に対し、１台分の記憶装置の容量の冗長データを格納する場合、仮想ページ容量２６００の（N＋1）/Nの容量が確保される。当然、RAID０のように、冗長性がない場合、仮想ページ容量２６００と等しい容量が実ページの容量ということになる。

【０１５５】

なお、実施例２においては、仮想ページ容量２６００はストレージシステム１００の中で共通であるが、ストレージシステム１００に仮想ページ容量２６００に異なったものがあっても、本発明は有効である。なお、実施例２では、それぞれの記憶装置グループは、RAID5で構成されているものとする。もちろん、本発明は、記憶装置グループが任意のRAIDグループで構成されていても有効である。

【０１５６】

ストレージボックス情報２０５０の構成は、実施例１と同様である。

【０１５７】

図２７に、実施例２のストレージグループ情報２３００の形式を示す。第３のストレージグループ情報２３００は、実施例１では含まれるセグメント数２３０３、割り当て権限セグメント数３２０９、空きセグメント数２３０４、割り当て権限セグメントビットマップ３２０８、空きセグメントビットマップ２３０７を含まないかわりに、空きページ情報ポインタ２３１０を含む。

【０１５８】

図２８は、実ページ情報２１００の形式である。実ページ情報２１００は、実ページごとに存在する該当する実ページの管理情報である。実ページ情報２１００は、ストレージグループ識別子２１０１、実ページアドレス２１０２、空きページ情報ポインタ２３１０、ページ無効ビットマップ２１０４、第１キャッシュ管理情報ポインタ２１０５、第２キャッシュ管理情報ポインタ２１０６から構成される。

【０１５９】

実施例２では、ログストラクチャストリームで、ライトデータは新しい領域に書き込まれるので、それまでのデータは無効になる。この場合、対応する領域のページ無効ビットマップ２１０４のビットをオンにする。実施例２では、当該ログストリームに書き込むデータを、パリティを生成し、ストレージ装置１６０に書き込むまでの間、キャッシュに２重に格納する。第１キャッシュ管理情報ポインタ２１０５、第２キャッシュ管理情報ポインタ２１０６は、このデータを格納したキャッシュ管理情報である。

【０１６０】

また、ここでは、生成したパリティもキャッシュ２１０に格納するため、このためのキャッシュ管理情報２７５０もポイントされる。第１キャッシュ管理情報ポインタ２１０５、第２キャッシュ管理情報ポインタ２１０６の数は、実ページ容量（（パリティも含む）／一つのキャッシュ管理２７５０で管理するキャッシュ容量）である。

【０１６１】

ストレージグループ識別子２１０１は、該当する実ページが、どのストレージグループ２８０に割り当てられているかを示す。なお、実施例２でも、ストレージグループ２８０を構成するするストレージ装置１６０は、一つのストレージボックス１３０に含まれるものとする。なお、本発明は、ストレージグループ２８０を構成するストレージ装置１６０

が、複数のストレージボックス１３０に含まれていても有効である。実ページアドレスは、対応ストレージグループ２８０の中で、どの相対的なアドレスに割り当てられているかを示す情報である。空きページポインタ２１０３は、この実ページに、仮想ページが割り当てられていない場合、有効な値となる。この場合、その値は、対応するストレージグループ２８０の中で、次の仮想ページが割り当てられていない空きページ情報をさす。

【０１６２】

空きページ情報ポインタ２３１０は、ストレージグループごとに設けられる情報である。図２９は、空きページ情報ポインタ２３１０によって管理される空き実ページの集合を表している。空き実ページとは、仮想ページに割り当てられていない実ページを意味する。また、空き実ページに対応した実ページ情報２１００を空き実ページ情報２１００と呼ぶ。全体の構造を、空き実ページ情報キュー２９００と呼ぶ。空きページ管理情報ポインタ２２００（図２５参照）は、先頭の空き実ページ情報２１００のアドレスを指す。次に、先頭の空き実ページ情報２１００の中の空きページポインタ２１０３が次の空き実ページ情報２１００を指す。図２９では、最後の空き実ページ情報２１００の空きページポインタ２１０３は、空きページ管理情報ポインタ２２００を示しているが、ヌル値でもよい。ストレージコントローラ２００は、実ページを割り当てていない仮想ページに書込み要求を受け付けると、論理ボリュームRAIDタイプ２００３と割り当て範囲２０２６（図９参照）に該当する、ストレージグループのいずれか、例えば、該当する記憶装置グループの中の空き実ページ数の最も多いストレージグループに対応する空きページ管理情報ポインタ２２００から、空き実ページを探し、仮想ページに割り当てる。仮想ページ容量２６００は、仮想ページの容量である。

【０１６３】

ストレージ装置情報２５００のフォーマットは、実施例１と同様である。ただし、一時書込み領域アドレス２５０６が示す領域は、実ページを割り当てていない領域とする。

【０１６４】

次に、上記に説明した管理情報を用いて、ストレージコントローラ２００が実行する動作の説明を行う。まず、ストレージコントローラ２００の動作を説明する。ストレージコントローラ２００の動作は、ストレージコントローラ２００内のプロセッサ２６０が実行し、そのプログラムは、メモリ２７０に格納されている。図３０は、メモリ２７０内に、格納された実施例２に関するプログラムが示されている。実施例２に関するプログラムは、実施例１に加えて、重複排前処理部４８００が加わる。

【０１６５】

次に、上記に説明した管理情報を用いて、ストレージコントローラ２００が実行する動作の説明を行う。

【０１６６】

図３１Ａ、Ｂは、実施例２のライト要求受付部４１００の処理フローである。ライト要求受付部４１００は、ストレージコントローラ２００が、サーバ１１０からライト要求を受け付けたときに実行される。

ステップＳ１２０００：プロセッサ２６０は、受け取ったライト要求で指定された仮想論理ボリュームを、仮想論理ボリューム情報２０８５によって、論理ボリュームに変換する。

ステップＳ１２００１：論理ボリューム直接転送中止フラグ２０３０がオンになっているかをチェックする。この場合、直接転送を中止するので、ステップＳ１２０２６へジャンプする。

ステップＳ１２００２：指定されたライトを行う領域に、キャッシュ管理情報２７５０が、割り当てられているかをチェックする。割り当てられている場合、ステップＳ１２００４へジャンプする。

【０１６７】

ステップＳ１２００３：割り当てられていない場合、当該領域に、キャッシュ管理情報２７５０を割り当てる。ここでは、空きキャッシュ管理情報ポインタ２６５０から、キャ

ッシュ管理情報２７５０を、パリティグループに相当する領域の、第１キャッシュ管理情報ポインタ２０２２と、第２キャッシュ管理情報ポインタ２０２３に割り当てる。そして、それぞれのキャッシュ管理情報２７５０の直接転送フラグ２７５７をオンにする。また、ブロックビットマップ２７５３，更新ビットマップ２７５４、アクセス不可ビットマップ２７５５は、オールクリアする。なお、空き状態にあるスロット管理情報や未使用の一時書込み領域が不足した場合、割り当てをやめ、直接転送は実行せず、従来の処理を実行する。この後、ステップＳ１２００５にジャンプする。

【０１６８】

ステップＳ１２００４：マッピング情報（マッピングアドレス）２００４を参照して、ライトを行う領域が、現在は、どこに格納されているかをチェックする。実ページに格納されている場合、対応する実ページ情報２１００の対応するページ無効ビットマップ２１０４をオンにする。圧縮前格納領域存在する場合、第１圧縮前無効ビットマップ２０２０、第２圧縮前無効ビットマップ２０２１の対応する無効ビットをオンにする。圧縮後格納領域存在する場合、第１無効圧縮後ビットマップ２０１１、第２圧縮後無効ビットマップ２０１２の対応する無効ビットをオンにする。マッピング情報（マッピングアドレス）２００４がヌルの場合、特に何もしない。

【０１６９】

ステップＳ１２００５：ここでは、ライトを指定されたアドレスに対応する更新ビットマップ２７５４、ブロックビットマップ２７５３をオンにする。また、マッピング情報（マッピングアドレス）２００４には、第１圧縮前書込みポインタ２０１３、第２圧縮前書込みポインタ２０１４が示すアドレスを設定する。

【０１７０】

ステップＳ１２００６：本ステップでは、ストレージボックス１３０に、直接転送を行うように指示する。この際、ライトデータを書き込むアドレスとして、第１圧縮前書込みポインタ２０１３、第２圧縮前書込みポインタ２０１４が示すアドレスと二つのアドレスを、一時的に書き込むアドレスとして指示する。また、第２の実施例では、ログストラクチャストリームでデータが書き込まれるので、論理ボリュームのバラバラなアドレスのデータが圧縮前格納領域に書き込まれることになる。実施例２では、ストレージボックス１３０に、論理ボリュームの識別子と書込みが発生した相対アドレス（実ページ内のオフセット）を送るものとする。もちろん、ストレージボックス１３０に、論理ボリュームの識別子と書込みが発生した相対アドレスをストレージボックス１３０にこれらの情報を送らなくとも、本発明は有効である。

ステップＳ１２００７：完了をまつ。

ステップＳ１２００８：正常終了か異常終了かを判別する。異常終了の場合、直接転送の中止を行うため、ステップＳ１２０２５へジャンプする。なお、異常終了の場合、ある時間たっても、ストレージボックス１３０からの完了報告がなく、ストレージコントローラ２００の時間監視で、タイムオーバを起こした場合も含める。

【０１７１】

ステップＳ１２００９：ストレージボックス１３０から受け取ったハッシュ値を、対応するハッシュ値２０１９に設定する。また、第１圧縮前書込みポインタ２０１３、第２圧縮前書込みポインタ２０１４を更新する。

ステップＳ１２０１０：サーバ１１０に正常終了を報告する。

ステップＳ１２０１１：第１圧縮前格納領域全体の書込みが完了したかをチェックする。完了していない場合、処理を完了する。

ステップＳ１２０１２：完了した場合、ログストラクチャ書込みポインタ２００６が示す実ページの相対アドレスに対応するキャッシュ管理情報２７５０が割り当てられているかをチェックする。割り当てられていれば、ステップＳ１２０１４へジャンプする。

ステップＳ１２０１３：割り当てられていなければ、キャッシュ管理情報を割り当てる。割り当てた場合、また、ブロックビットマップ２７５３，更新ビットマップ２７５４、アクセス不可ビットマップ２７５５をオールクリアする。

ステップＳ１２０１４：実施例２では、第１圧縮前格納領域全体の書込みが完了した場合、ストレージボックス１３０から、第１圧縮前格納領域に格納されていたデータの論理ボリュームの相対アドレスが送られてくるものとする。ここでは、第１圧縮前格納領域に格納されていたデータの論理ボリュームの相対アドレスをパラメータにして、重複排除前処理部４８００をコールする。

【０１７２】

図３１Ｂの説明を行う。

ステップＳ１２０１５：処理終了後以下の処理を実行する。ストレージボックス１３０に、それぞれの重複排除単位の論理ボリュームの相対アドレスと、その単位が、重複排除できないか、できる可能性があるかを示す情報をおくる。加えて、できる可能性がある場合、データを読み出し、チェックする必要のあるアドレスを通知する。さらに、重複排除できなかったデータを、圧縮し、第１圧縮後書込みポインタ２０１５、第２圧縮後書込みポインタ２０１６が示す、アドレスから格納していくように指示する。

ステップＳ１２０１６：完了をまつ。

【０１７３】

ステップＳ１２０１７：正常終了か、異常終了かを判別する。異常終了の場合、直接転送を中止するための処理を実行するため、ステップＳ１２０２５へジャンプする。なお、異常終了の場合、ある時間たっても、ストレージボックス１３０から　の完了報告がなく、ストレージコントローラ２００の時間監視で、タイムオーバを起こした場合も含める。

【０１７４】

ステップＳ１２０１８：対応する圧縮後データ長２０１８に、ストレージボックス１３０から受け取ったデータ長を設定する。重複排除できたデータに対応するマッピング情報（マッピングアドレス）２００４に、同じデータを格納していたアドレスを設定する。また、重複排除できなかったデータについては、対応するマッピング情報（マッピングアドレス）２００４に、ストレージボックス１３０から受け取った、このデータを格納した第１圧縮後データ格納領域ポインタ２００９と第２圧縮後データ格納領域２０１２のアドレスを設定する。また、圧縮を行ったデータに対応する論理ボリュームの相対アドレスに割り当てたキャッシュ管理情報の更新ビットマップ２７５４、ブロックビットマップ２７５３をオフする。キャッシュ管理情報のすべての更新ビットマップ２７５４、ブロックビットマップ２７５３がオフになった場合、キャッシュ管理情報２７５０を、空きキャッシュ管理情報ポインタ２６５０に戻す。

【０１７５】

ステップＳ１２０１９：圧縮後データに対応する実ページに割り当てたキャッシュ管理情報２７５０の対応するブロックビットマップ２７５３、更新ビットマップ２７５４をオンにする。

ステップＳ１２０２０：第１圧縮後データ格納領域ポインタ２００９が満杯になった（格納したデータ長が、パリティグループ長になった）かを、チェックする。なっていなければ、処理を終了する。

ステップＳ１２０２１：満杯になっている場合、ストレージボックス１３０に、パリティを生成し、ログストラクチャ書込みポインタ２００６が示す領域からデータと生成したパリティを記憶装置に書き込むよう、指示する。

ステップＳ１２０２２：完了をまつ。

【０１７６】

ステップＳ１２０２３：正常終了か、異常終了かを判別する。異常終了の場合、直接転送を中止するための処理を実行するため、ステップＳ１２０２５へジャンプする。なお、異常終了の場合、ある時間たっても、ストレージボックス１３０から　の完了報告がなく、ストレージコントローラ２００の時間監視で、タイムオーバを起こした場合も含める。

【０１７７】

ステップＳ１２０２４：圧縮後データを格納した実ページと相対アドレスを、対応するマッピング情報２００４に格納する。また、ここでは、当該パリティグループに割り当て

たキャッシュ管理情報２７５０を開放する。具体的には、最後のキャッシュ管理情報２７５０以外のキャッシュ管理情報２７５０は、空きキャッシュ管理情報キュー２６５０に戻す。また、キューに戻すキャッシュ管理情報２７５０のブロックビットマップ２７５３等のビットマップ情報やフラグ情報は、すべてオフにする。この後、処理を終了する。

【０１７８】

ステップＳ１２０２５：異常処理が発生したので、直接転送の後処理を実行するため、直接転送異常処理をコールする。この後、処理を終了する。

ステップＳ１２０２６：キャッシュにライトデータを受け取る処理に入る。この処理は、公知の技術であるため、詳細を説明しない。終了後、処理を完了させる。

【０１７９】

図３２は、実施例２のライト異常終了対応処理部４２００の処理フローである。本処理も、プロセッサ２６０が、適宜実行する処理である。

【０１８０】

ステップＳ１３０００：ストレージボックス１３０が、無応答で異常終了したのかをチェックする。無応答で処理を終了していない場合、ステップＳ１３００４へジャンプする。

【０１８１】

ステップＳ１３００１：無応答で処理を終了した場合、ストレージボックス１３０が電源ダウン等で、通信できない可能性がある。本ステップでは、一度、ストレージボックスにリセット要求を発行する。

ステップＳ１３００２：リセット要求が完了した場合、ステップＳ１３００４へジャンプする。

ステップＳ１３００３：ストレージボックス１３０がダウンしている等の原因があるので、ストレージ管理者、保守員等に、連絡して、ストレージボックス１３０が立ち上がるのをまつ。立ち上がった後、ステップＳ１３００４から実行を開始する。

ステップＳ１３００４：当該ログストラクチャストリームに対応する論理ボリューム直接転送中止フラグ２０３０をオンにする。

【０１８２】

ステップＳ１３００５：ストレージボックス１３０に、m個の第１圧縮前データ格納領域ポインタ２００７、m個の第２圧縮前データ格納領域ポインタ２００８、2個の第１圧縮後データ格納領域ポインタ２００９、2個の第２圧縮後データ格納領域ポインタ２０１０が示すアドレスと、それぞれの格納領域に対応した第１無効圧縮前ビットマップ２０２０、第２圧縮前無効ビットマップ２０２１、第１無効圧縮後ビットマップ２０１１、第２圧縮後無効ビットマップ２０１２を通知して、無効ビットがオフの領域に格納されたデータとその論理ボリュームのアドレスを、ストレージコントローラ２００に送るよう指示する。

【０１８３】

なお、実施例２では、ストレージボックス１３０には、サーバ１１０から受け取ったデータを転送するよう、指示しているが、ストレージボックス１３０で、すでにパリティを生成している場合、パリティもストレージボックス１３０から受け取って、キャッシュ２１０に格納するようにしてもよい。

【０１８４】

ステップＳ１３００６：完了をまつ。ある時間待って、完了が返って、こなかった場合、ステップＳ１３０００へジャンプする。

ステップＳ１３００７：一部のデータが転送に失敗した場合、ステップＳ１３００８へジャンプする。

【０１８５】

ステップＳ１３００８：すべてのデータの転送が正常終了した場合、以下の処理を実行する。第１圧縮前データ格納領域、あるいは、第２圧縮前データ格納領域　から受け取ったデータは、対応する論理ボリュームの相対アドレスから論理ボリュームの空間に割り当てた対応するスロット管理情報２７０５に対応したキャッシュ２１０の領域に２重に書き

込む。また、第１圧縮後データ格納領域、あるいは、第２圧縮後データ格納領域　から受け取ったデータは、実ページに空間に割り当てた対応するスロット管理情報２７０５に対応したキャッシュ２１０の領域に２重に書き込む。この後、処理を終了する。

【０１８６】

ステップＳ１３００９：転送に成功したデータのみ、キャッシュ管理情報２７５０に対応した領域に、２重に書き込む。

ステップＳ１３１０：転送に失敗したデータに対応するアクセス不可ビットマップ２７５５をオンにする。

【０１８７】

当該パリティグループに対応するアクセス不可ビットマップ２７５５にオンがない場合、（すべてのデータがキャッシュ２１０に正常に格納できた場合）、ストレージコントローラ２００がパリティを生成して、データとパリティを、ストレージ装置１６０に書き込むことができる。

【０１８８】

アクセス不可ビットマップ２７５５がオンのビットをもつスロットの場合、このアクセス不可ビットマップ２７５５がオンに対応する領域に、新しいライト要求が入り、キャッシュ２１０にデータを正常に格納できれば、対応するアクセス不可ビットマップ２７５５をオフにできる。当該スロットに対応するキャッシュ管理情報２７５０にすべてのアクセス不可ビットマップ２７５５がオフになった場合、旧データ破壊フラグ２７５９がオフであれば、ストレージコントローラ２００が、キャッシュ２１０に旧データや旧パリティを読み込み、これらから、パリティを生成して、データとパリティをストレージ装置１６０に格納する論理を適用できる。

【０１８９】

旧データ破壊フラグ２７５９がオンの場合、パリティの生成に旧データが使用できないので、当該パリティグループに対応するすべてのキャッシュ管理情報２７５０のすべてのアクセス不可ビットマップ２７５５がオフになった場合、ストレージコントローラ２００が、残りのパリティグループのデータをキャッシュ２１０に読み込み、パリティを生成して、パリティグループ全体のデータとパリティを、ストレージ装置１６０に書き込むことができる。

【０１９０】

図３３は、実施例２のリード処理実行部４３００の処理フローである。実施例２では、ストレージボックス１３０からサーバに、要求されたデータを直接転送するようにする。もちろん、本発明は、リード要求で指定されたデータを、ストレージボックス１３０からストレージコントローラ２００を経由して、サーバに送っても、本発明は有効である。

【０１９１】

ステップＳ１４０００：プロセッサ２６０は、受け取ったリード要求で指定された仮想論理ボリュームを、仮想論理ボリューム情報２０８５によって、論理ボリュームに変換し、対応する論理ボリューム情報２０００を獲得する。

【０１９２】

ステップＳ１４００１：受け取ったリード要求のアドレスから、マッピング情報（マッピングアドレス）２００４，キャッシュ管理情報ポインタ２０２２，実ページ情報２１００、キャッシュ管理情報２７５０のブロックビットマップ２７５３などから、当該要求で、指定されたデータが、キャッシュ２１０にヒットしているかをチェックする。ミスであれば、ステップＳ１４００２へジャンプする。ヒットであれば、ステップＳ１４０１１へジャンプする。

【０１９３】

ステップＳ１４００２：マッピング情報（マッピングアドレス）２００４を参照して、当該データが、圧縮前データ格納領域、あるいは、圧縮後データ格納領域に格納されているかをチェックする。そうであれば、ステップＳ１４００６へジャンプする。

【０１９４】

ステップＳ１４００３：マッピング情報（マッピングアドレス）２００４に格納された実ページ上のアドレスを指定して、ストレージボックス１３０に、格納されたデータを、サーバ１１０に送るよう指示する。
ステップＳ１４００４：完了をまつ。
ステップＳ１４００５：サーバ１１０に完了報告を行い、処理を終了する。
ステップＳ１４００６：要求されたデータに対応するアクセス不可ビットマップ２７５５のビットがオンになっていないかをチェックする。オフであれば、ステップＳ１４００８へジャンプする。
ステップＳ１４００７：サーバ１１０に対し、当該データがアクセスできないという異常報告を行う。この後処理を終了する。
ステップＳ１４００８：ストレージボックス１３０に、マッピング情報に示された第１圧縮前デー格納領域と第２圧縮前データ格納領域の二つのアドレス、あるいは、第１圧縮後データ格納領域と第２圧縮後データ格納領域の二つのアドレスから、サーバ１１０に送るよう指示する。
ステップＳ１４００９：完了をまつ。
ステップＳ１４０１０：サーバ１１０に完了報告を行い、処理を終了する。
ステップＳ１４０１１：キャッシュから、要求されたデータを、サーバに転送する。
ステップＳ１４０１２：転送が完了するのを待つ。
ステップＳ１４０１３：サーバ１１０に対し、完了報告を行い、処理を終了する。

【０１９５】

図３４は、重複排除前処理部４８００の処理フローである。重複排除処理とは、まったく同じ内容のデータを格納しないことにより、容量を削減する機能である。効率よく行うため、以下のステップを踏む。（１）データのハッシュ値をとり、ハッシュ値が同じ値になるデータの集合を見つける。（２）ハッシュ値が同じ値になるデータに関しては、データそのものを比較し、一致しているものがみつかった場合、そのデータを格納しない。実施例２では、（１）を、ストレージコントローラで実現し、（２）をストレージボックスで実現する。重複排除前処理部４８００は、（１）の処理を実行する。この処理自体は、公知であるため、簡単に説明する。

【０１９６】

ステップＳ１５０００：マッピングテーブルを参照して、満杯になった第１圧縮前データ格納領域に格納されたデータのハッシュ値と同じハッシュ値をもつものを探す。それぞれのデータごとに、同じハッシュ値をもつ場合、そのデータが格納されているアドレスの集合を獲得する。以上を実行して、処理を完了する。

【０１９７】

次に、ストレージコントローラ２００の指示にしたがって、ストレージボックス１３０のボックスプロセッサ１７０が実行する動作の説明を行う。実行するプログラムは、ボックスメモリ１８１に格納されている。

【０１９８】

図３５は、ボックスメモリ１８１内に、格納された実施例２に関するプログラムが示されている。実施例２に関するプログラムは、実施例１に加えて、データ移動部４９００、ボックス間データ転送部４９５０が含まれる。

【０１９９】

図３６は、実施例２におけるライトデータ受領部４４００の処理フローである。ライトデータ受領部４４００は、ストレージコントローラ２００の指示にしたがって、サーバ１１０から、ライトデータを受け取り、一時書込み領域にデータを書き込む処理である。
ステップＳ１６０００：サーバ１１０に、データを送るよう指示する。
ステップＳ１６００１：転送の完了を待つ。
ステップＳ１６００２：正常終了であれば、ステップＳ１６００４へジャンプする。
ステップＳ１６００３：異常終了報告を、ストレージコントローラに返し、処理を終了する。

ステップＳ１６００４：受け取ったライトデータのハッシュ値を計算する。
ステップＳ１６００５：ストレージコントローラ２００から指定された、二つの一時書込み領域のアドレスに書き込む。
ステップＳ１６００６：書込みが完了するのをまつ。
ステップＳ１６００７：正常終了であれば、ステップＳ１６００９へジャンプする。
ステップＳ１６００８：異常終了報告を、ストレージコントローラに返し、処理を終了する。
ステップＳ１６００９：ハッシュ値と正常終了報告を、ストレージコントローラに返し、処理を終了する。

【０２００】

図３７は、実施例２におけるデータ移動部４９００の処理フローである。データ移動部は、まず、重複排除処理の（２）の部分に関する処理を実行する。具体的には、ストレージコントローラから受け取った、同じハッシュ値をもつデータのアドレスのデータを読み出し内容が一致するかどうかをチェックする。一致したものがない場合、圧縮処理を行い、圧縮後のデータを、第１圧縮後データ格納領域と第２圧縮後データ格納領域の指定されたアドレスから格納していく。この圧縮後のデータの長さは、ストレージコントローラに返す。また、第１圧縮後データ格納領域と第２圧縮後データ格納領域が満杯になった場合、次の第１圧縮後データ格納領域と第２圧縮後データ格納領域の先頭からデータを格納する。加えて、第１圧縮後データ格納領域と第２圧縮後データ格納領域が満杯になった場合、この旨をストレージコントローラに報告する。

【０２０１】

ステップＳ１７０００：最初のデータを実行対象とする。
ステップＳ１７００１：対象としたデータが、重複排除できる可能性のあるデータかどうかをチェックする。具体的には、同じハッシュ値をもつデータがあるかどうかをチェックする。同じハッシュ値をもつデータがない場合、ステップＳ１７００６へジャンプする。

【０２０２】

ステップＳ１７００２：指定されたアドレスに該当するデータを読み出す要求を発行する。この場合、複数のアドレスが指定された場合、複数の読み出し要求を並行して発行する。また、指定されたアドレスが当該ストレージボックス１３０うちストレージ装置１６０である場合、そのストレージに読み出す領域のアドレスを指定して、要求を発行する。また、当該ストレージボックス１３０のストレージ装置１６０でない場合、そのストレージ装置を含むストレージボックス１３０に要求を発行する。

【０２０３】

ステップＳ１７００３：要求が完了するのをまつ。
ステップＳ１７００４：読み出したデータの中に、当該データと一致するものがないかをチェックする。一致するものがない場合、ステップＳ１７００６にジャンプする。
ステップＳ１７００５：当該データが重複排除できたことを記憶しておく。この後、ステップＳ１７００８へジャンプする。
ステップＳ１７００６：扱っているデータを圧縮し、第１圧縮後データ格納領域と第２圧縮後データ格納領域に格納する。
ステップＳ１７００７：第１圧縮後データ格納領域と第２圧縮後データ格納領域が満杯になったかをチェックする。満杯でなければ、ステップＳ１７００９にジャンプする。
ステップＳ１７００８満杯になった場合、これを記憶し、次の第１圧縮後データ格納領域と第２圧縮後データ格納領域を使用するようにする。

【０２０４】

ステップＳ１７００９：指定された第１圧縮前データ格納領域と第２圧縮前データ格納領域のすべてのデータの処理が実行されたかをチェックする。実行されていない場合、次の第１圧縮前データ格納領域と第２圧縮前データ格納領域のデータを実行対象として、ステップＳ１７００１へジャンプする。

【０２０５】

ステップＳ１７０１０：コントローラに完了報告を通知する。この際、重複排除できたデータのアドレスを通知する。また、第１圧縮後データ格納領域と第２圧縮後データ格納領域が満杯になった場合、この旨を報告する。

【０２０６】

図３８は、実施例２におけるライトデータ書込み部４５００の処理フローである。ライトデータ書込み部４５００は、ストレージコントローラ２００の指示にしたがって、第１圧縮後データ格納領域と第２圧縮後データ格納領域からライトデータを読み出し、パリティを生成し、ライトデータとパリティを、ストレージコントローラ２００から、指示されたストレージ装置１６０の領域に書き込む。

【０２０７】

ステップＳ１８０００：ストレージコントローラ２００から指定された二つの第１圧縮後データ格納領域と第２圧縮後データ格納領域の一つから、データを読み出す。

ステップＳ１８００１：処理の完了をまつ。

ステップＳ１８００２：すべてのデータの読み出しが正常終了したかをチェックする。正常終了していれば、ステップＳ１８００５へジャンプする。

ステップＳ１８００３：もう一つの圧縮後データ格納領域から、異常終了したデータだけを読み出す。

ステップＳ１８００４：完了をまつ。

ステップＳ１８００５：指定したすべてのデータの読み出しが正常終了したかをチェックし、正常終了した場合、ステップＳ１８００７へジャンプする。

ステップＳ１８００６：ストレージコントローラ２００に、異常終了を報告し、処理を完了する。

ステップＳ１８００７：読み出したライトデータから、パリティを生成する。

ステップＳ１８００８：ライトデータとパリティを、ストレージコントローラ２００が指定したストレージ装置１６０の領域に書き込む。

ステップＳ１８００９：完了をまつ。

ステップＳ１８０１０：すべての書込みが正常終了した場合、ストレージコントローラ２００に、正常終了報告を行い、一つでも、異常終了があった場合、異常終了報告をストレージコントローラ２００に行う。

【０２０８】

図３９は、実施例２における一時書込み領域転送部４６００の処理フローである。一時書込み領域転送部４６００は、ストレージコントローラ２００の指示にしたがって、第１圧縮前データ格納領域、第２圧縮前データ格納領域、第１圧縮後データ格納領域、第２圧縮後データ格納領域から、データを読み出し、ストレージコントローラ２にデータを転送する。

【０２０９】

ステップＳ１９０００：ストレージコントローラ２が指定した、第１圧縮前データ格納領域、第１圧縮後データ格納領域のデータを読み出すため、対応する記憶装置に、データ読み出す要求を発行する。

ステップＳ１９００１：要求が完了するのをまつ。

ステップＳ１９００２：すべてのデータが正常に読みだせたかをチェックする。読み出せていれば、ステップＳ１９００７へジャンプする。

ステップＳ１９００３：読み出せなかったデータと同じデータを格納した第２圧縮前データ格納領域、第２圧縮後データ格納領域からデータを読み出すため、対応する記憶装置から、データを読み出す要求を発行する。

ステップＳ１９００４：要求が完了するのをまつ

ステップＳ１９００５：すべてのデータが正常に読みだせたかをチェックする。読み出せていれば、ステップＳ１９００７へジャンプする。

【０２１０】

ステップＳ１９００６：正常に読み出せたデータとそのアドレスをすべてストレージコ

ントローラ２００に送る。また、読み出せなかったデータについては、読み出せなかった旨とそのアドレスをストレージコントローラ２００に通知する。その後、処理を完了する。
ステップＳ１９００７：すべてのデータを読み出せたことになるので、データとそのアドレスを、ストレージコントローラ２００に送る。この後、処理を終了する。
【０２１１】
図４０は、実施例２におけるリードデータ直接転送部４７００の処理フローである。リードデータ直接転送部４７００は、ストレージコントローラ２００からの指示によって、指定されたデータをサーバ１１０に送る処理である。実施例２では、圧縮前データ格納領域と圧縮後データ格納領域の管理を、ストレージコントローラ２００が行っているため、サーバ１１０が読み書きするデータを恒久的に格納した領域か、パリティを作成するための一時的な書込み領域の区別を、ストレージボックス１３０は、認識していない。このため、サーバ１１０が読み書きするデータを恒久的に格納した領域か、パリティを作成するための一時的な書込み領域に係わらず、ストレージボックス１３０は、指定された領域のデータを送るだけである。以上から、サーバ１１０が読み書きするデータを恒久的に格納した領域のデータを送る処理と、パリティを作成するための一時的な書込み領域のデータを送る処理の処理フローは、同一となり、いずれも、リードデータ直接転送部４７００が実行する。
【０２１２】
ステップＳ２００００：指定された領域のデータをサーバ１１０に送る。
ステップＳ２０００１：処理が完了するのを待つ。
ステップＳ２０００２：正常終了した場合、正常終了報告を、異常終了した場合、異常終了報告を、ストレージコントローラ２００に返す。この後、処理を終了する。
【０２１３】
図４１は、実施例２におけるボックス間データ転送部４９５０の処理フローである。ボックス間データ転送部４９５０は、他のストレージボックス１３０から、要求されたデータを送る際に、実行される処理フローである。
【０２１４】
ステップＳ２１０００：指定されたストレージ装置１６０のアドレスから、データを読み出すよう、ストレージ装置１６０に要求を発行する。
ステップＳ２１００１：完了をまつ。
ステップＳ２２００１：読み出し要求を受け取ったストレージボックス１３０に読み出したデータを送信する。完了後、処理を終了する。
【０２１５】
本発明は、サーバ、ストレージコントローラとストレージボックスがネットワークを介して、接続された環境で、サーバからストレージボックスにデータを書き込んだ場合、障害が発生した場合、書き込んだデータをストレージコントローラのキャッシュに格納することにより、ストレージコントローラのもつ障害に対応した各種の処理を適応することができる。
【符号の説明】
【０２１６】
１００　実ストレージシステム
１１０　サーバ
１２０　ネットワーク
１３０　ストレージボックス
１５０　ストレージ管理サーバ
１６０　ストレージ装置
１７５　共有ストレージ装置
１７０　ボックスプロセッサ
１８０　ボックスポート
１８１　ボックスメモリ

１９０　仮想ストレージシステム
１９５　サーバポート
１９７　ストレージパス
１９８　サーバポート情報
２００　ストレージコントローラ
２１０　キャッシュ
２２０　共有メモリ
２３０　内部ストレージ装置
２５０　接続装置
２６０　プロセッサ
２７０　メモリ
２７５　バッファ
２０５０　ストレージボックス情報
２０５５　仮想論理ボリューム情報
２０７０　他ストレージシステム情報
２０８４　仮想論理ボリューム情報
２０００　論理ボリューム情報
２０６０　ストレージシステム情報
２３００　ストレージグループ情報
２３５０　ストレージ装置情報
２７５０　キャッシュ管理情報
２６５０　空きキャッシュ管理情報ポインタ
４１００　ライト要求受付部
４２００　ライト異常終了対応処理部
４３００　リード処理実行部
４４００　ライトデータ受領部
４５００　ライトデータ書込み部
４６００　一時書込み領域転送部
４７００　リードデータ直接転送部
１９６　圧縮/伸長回路
１９７　ハッシュ回路
２１００　実ページ管理情報
２２００　空きページ管理情報ポインタ
２６００　仮想ページ容量
４８００　重複排除前処理部
４９００　データ移動部
４９５０　ボックス間データ転送部

【図面】

【図１】

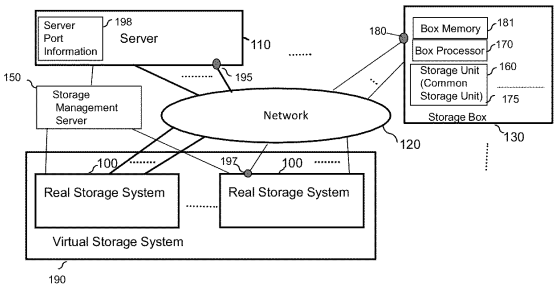
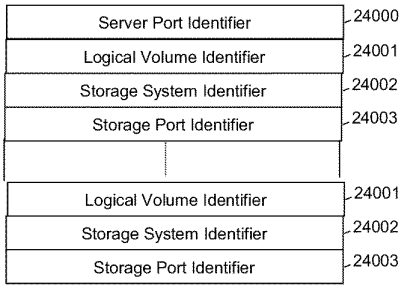


Fig. 1

【図２】



Sever port Information 198

Fig. 2

【図３】

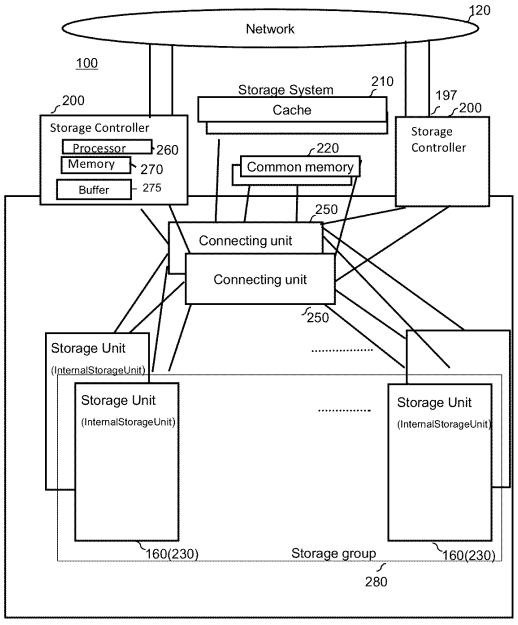
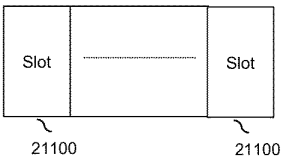


Fig. 3

【図４】



Cache Memory 210

Fig.4

【図５】

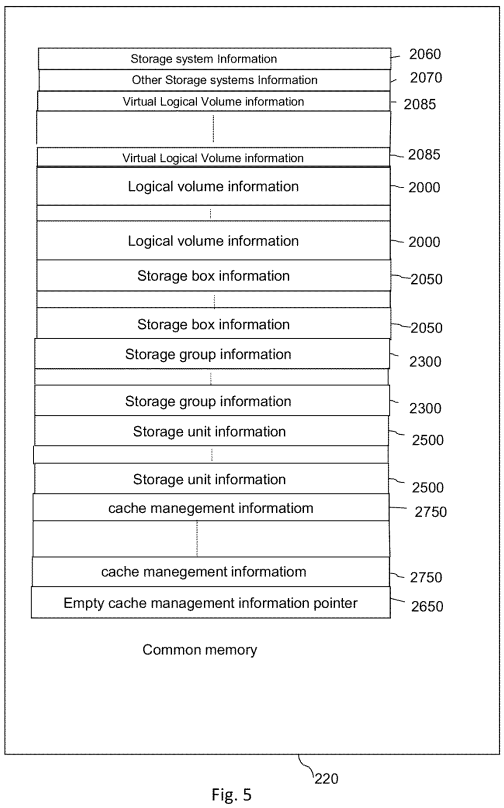
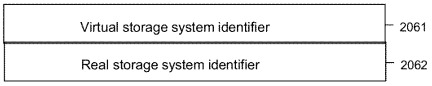


Fig. 5

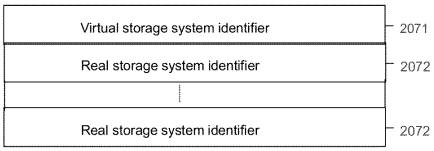
【図６】



Virtual Storage system information 2060

Fig.6

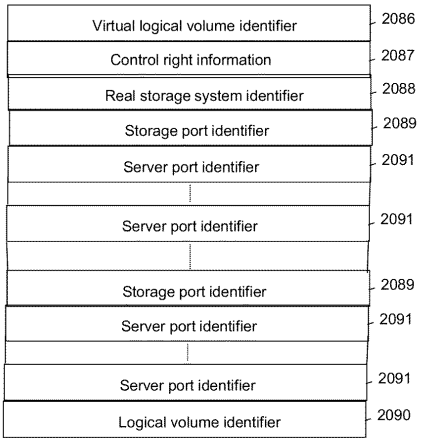
【図７】



Other Storage systems information 2070

Fig.7

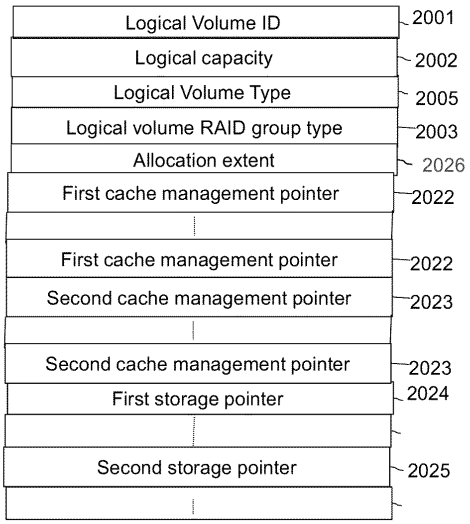
【図８】



Virtual logical volume information 2085

Fig.8

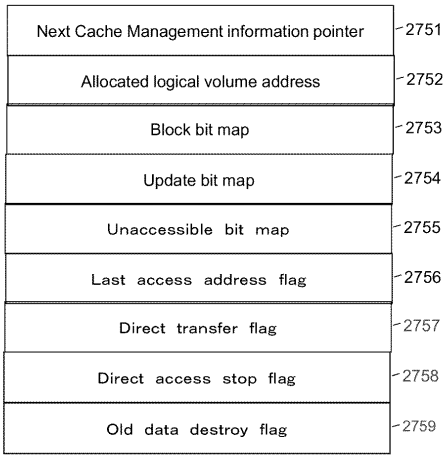
【図９】



Logical volume information 2000

Fig. 9

【図１０】



Cache Management information 2750

Fig.10

【図１１】

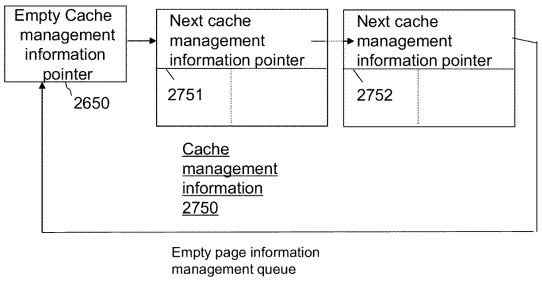
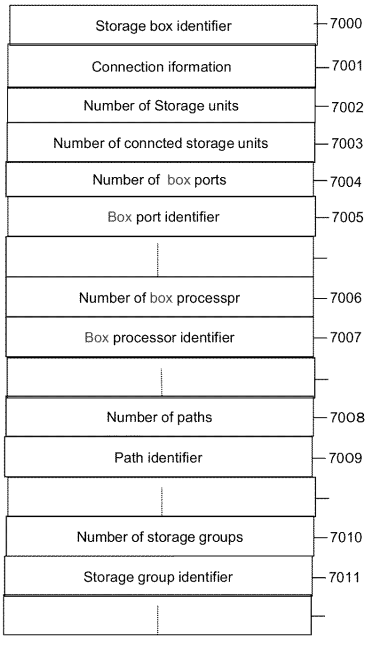


Fig. 11

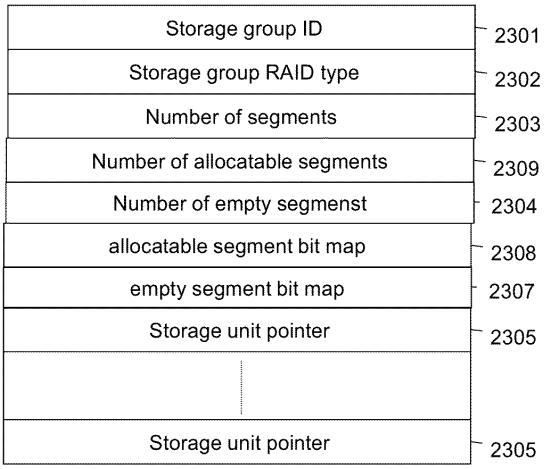
【図１２】



Storage box information 2050

Fig.12

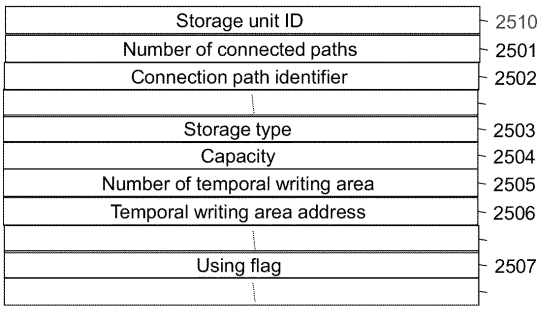
【図１３】



Storage group information 2300

Fig. 13

【図１４】



Storage unit information 2500

Fig. 14

【図１５】

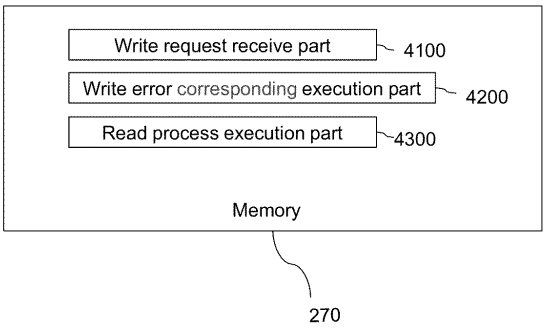


Fig. 15

【図１６Ａ】

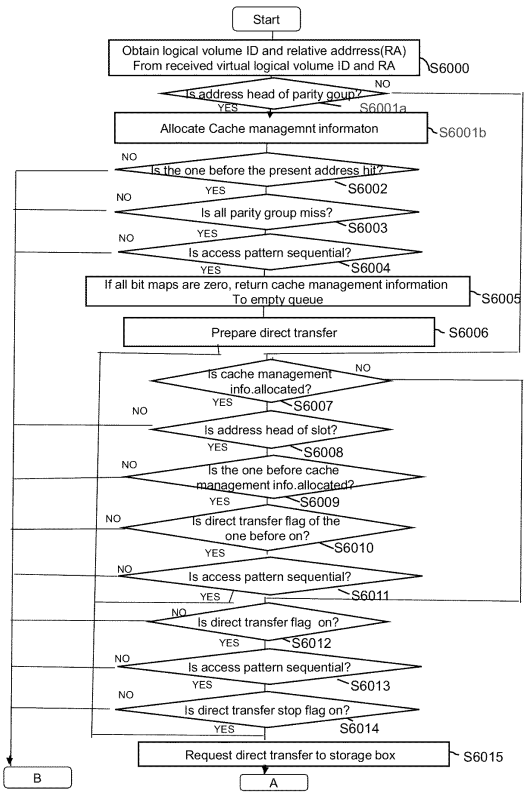


Fig. 16A

【図１６Ｂ】

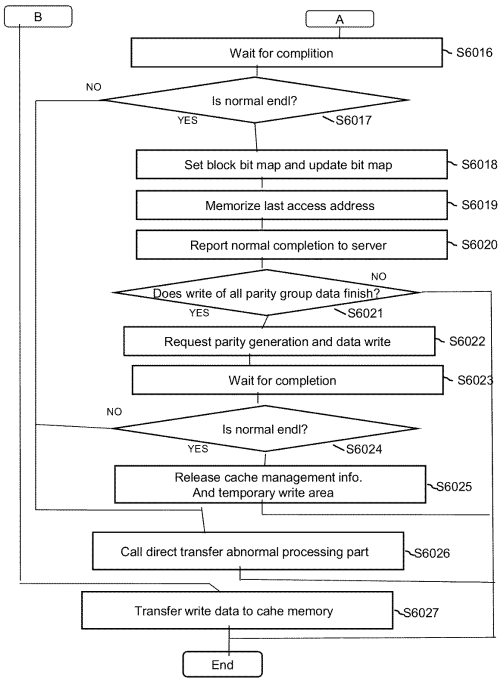


Fig. 16B

【図１７】

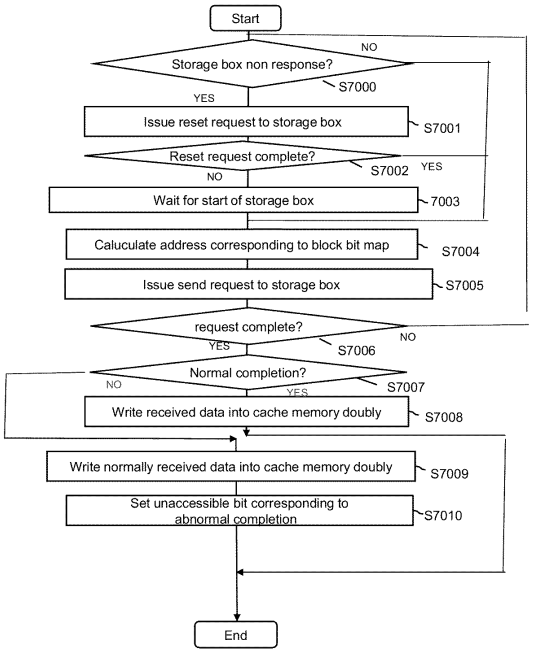


Fig. 17

【図１８】

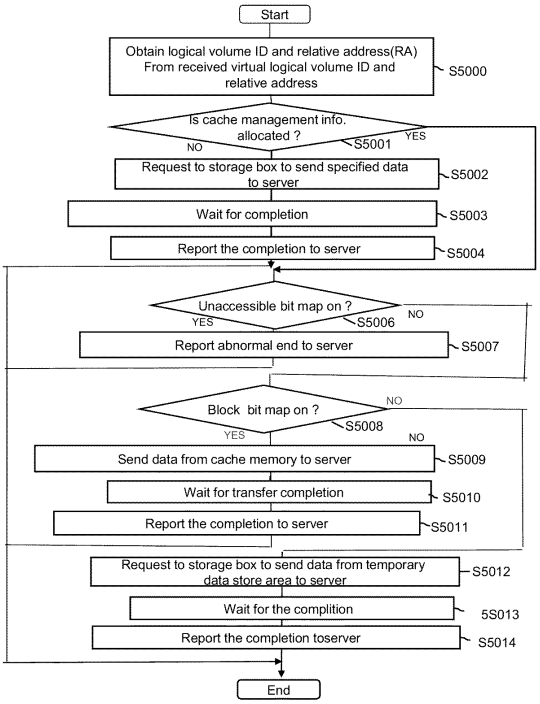


Fig. 18

【図１９】

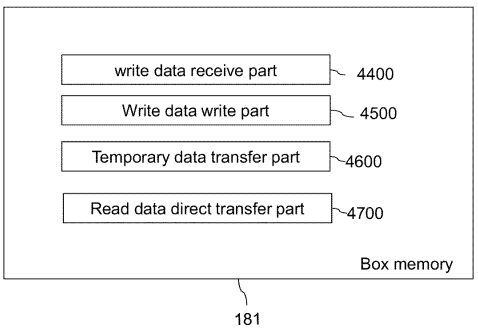


Fig. 19

【図２０】

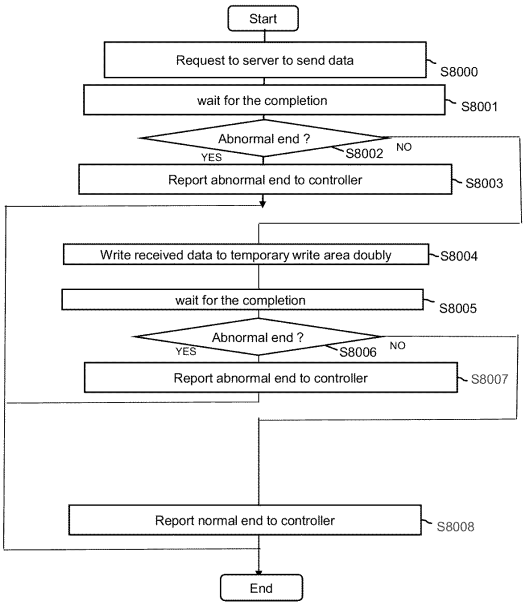


Fig. 20

【図２１】

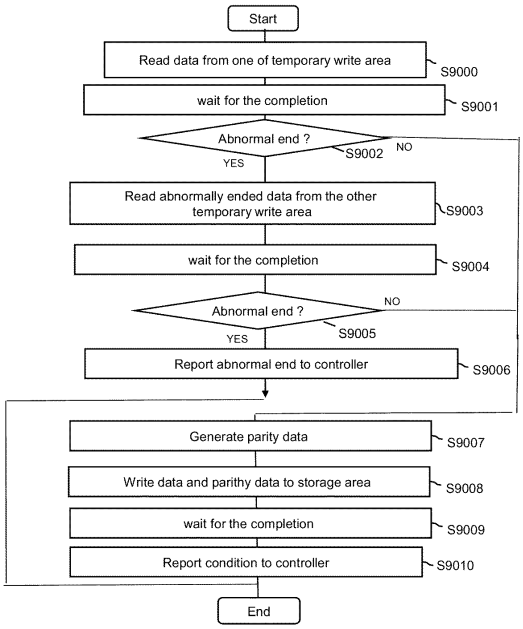


Fig. 21

【図２２】

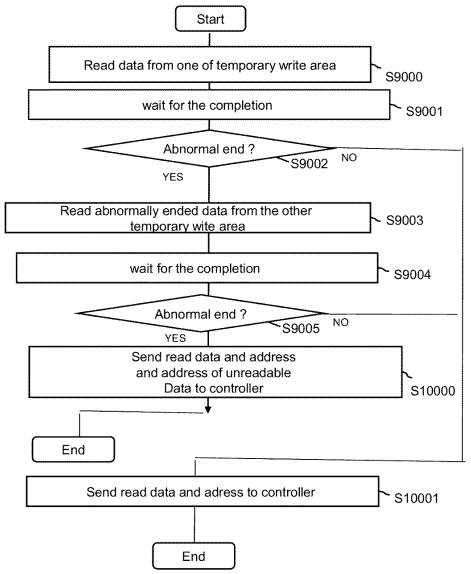


Fig. 22

【図２３】

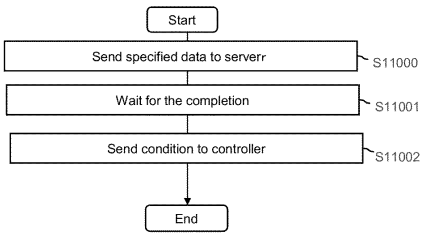


Fig. 23

【図２４】

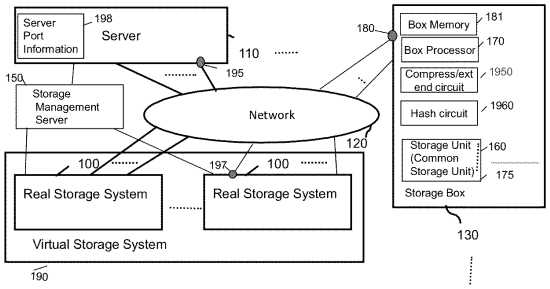


Fig. 24

【図２５】

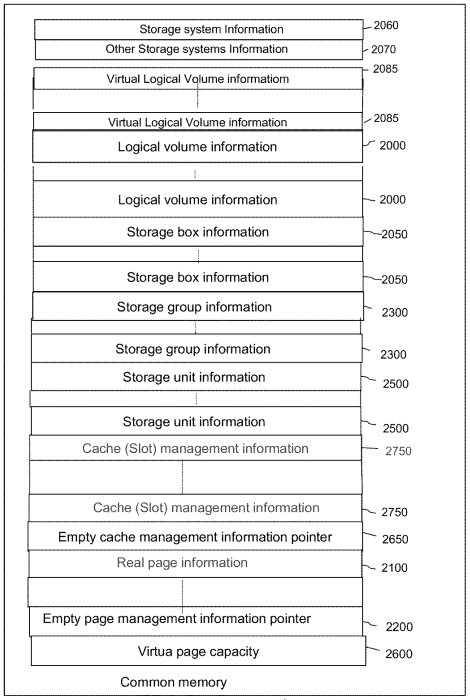
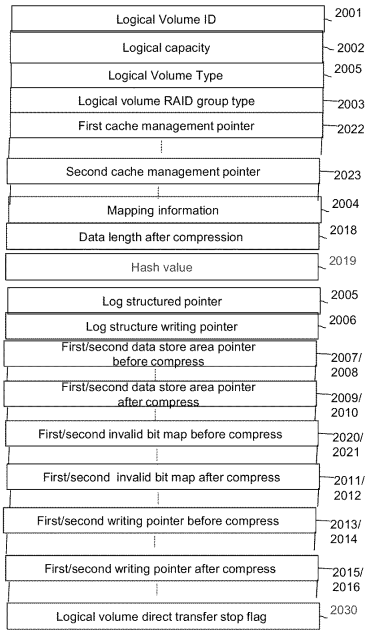


Fig.25

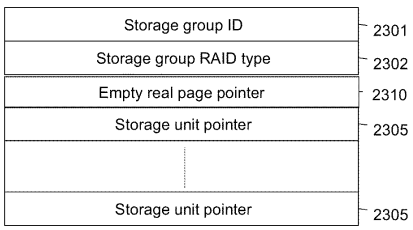
【図２６】



Logical volume information 2000

Fig. 26

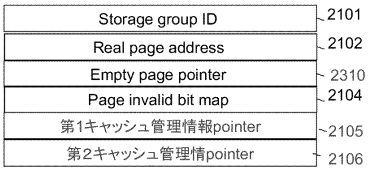
【図２７】



Storage group information 2300

Fig. 27

【図２８】



Real page information 2100

Fig. 28

【図２９】

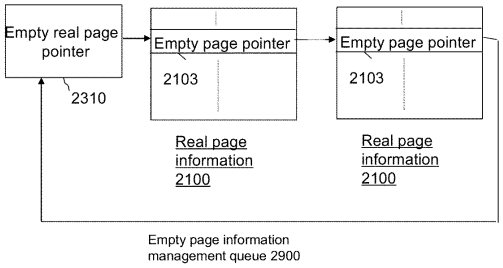


Fig.29

【図３０】

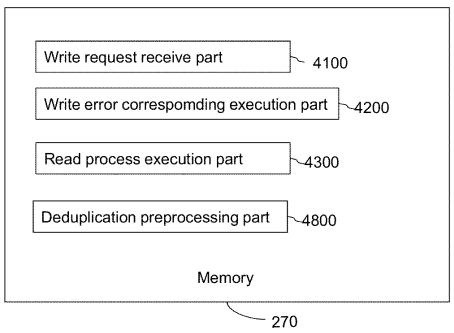


Fig. 30

【図３１Ａ】

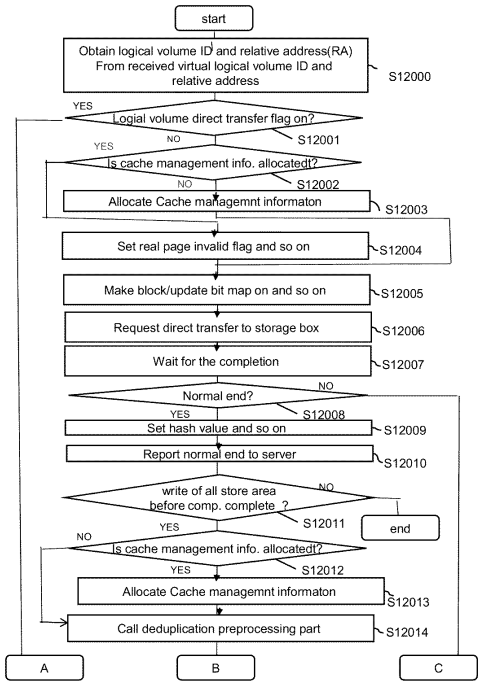


Fig. 31A

【図３１Ｂ】

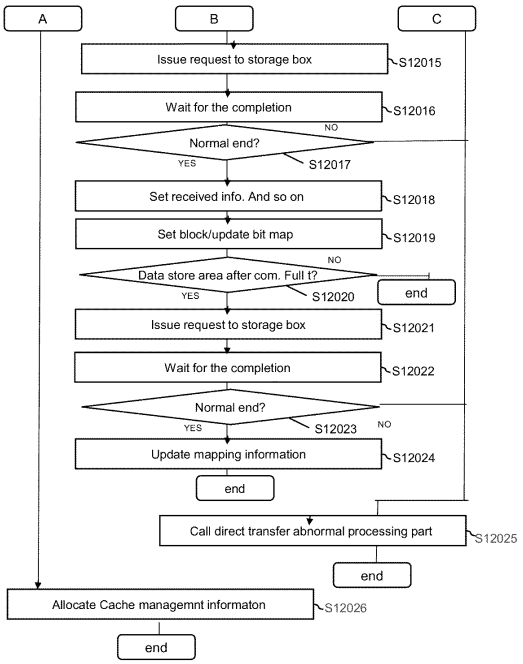


Fig. 31B

【図３２】

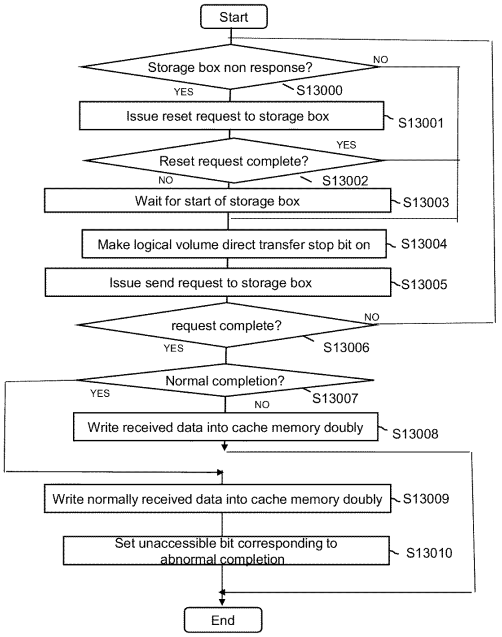


Fig.32

【図３３】

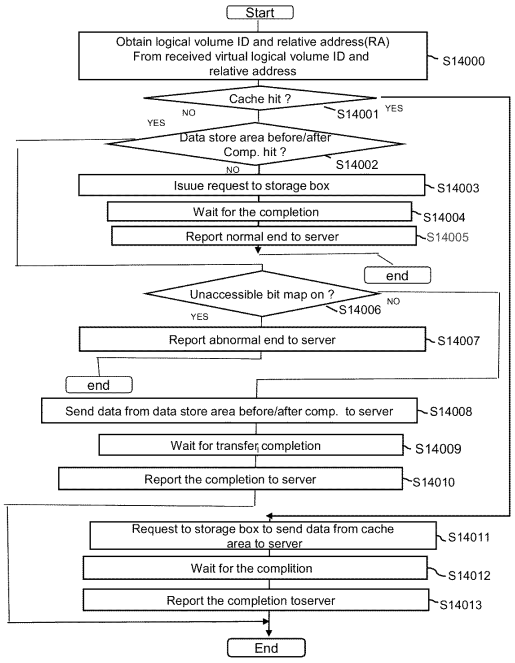


Fig. 33

【図３４】

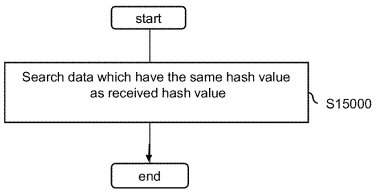


Fig.34

【図３５】

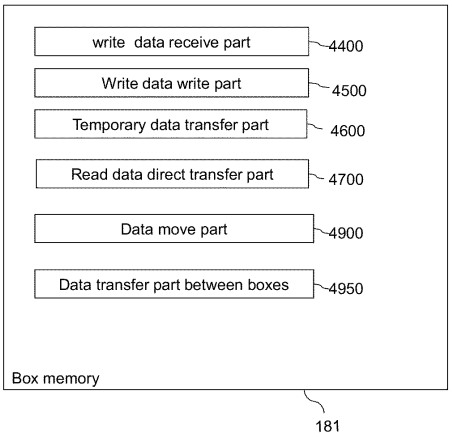


Fig. 35

【図３６】

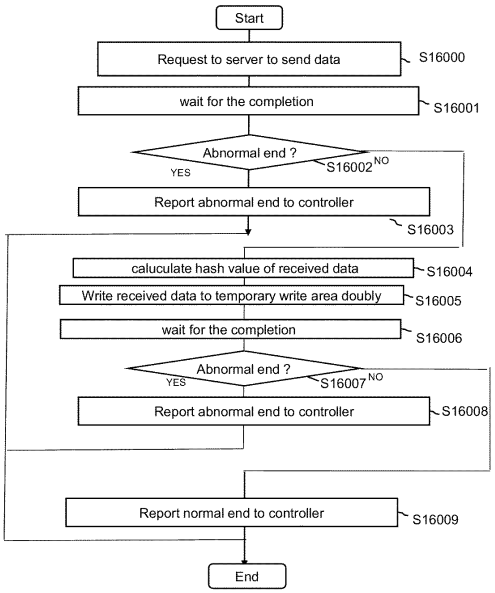


Fig. 36

【図３７】

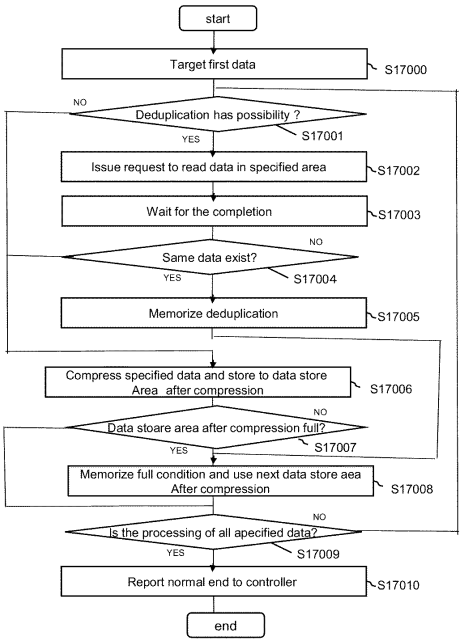


Fig. 37

【図３８】

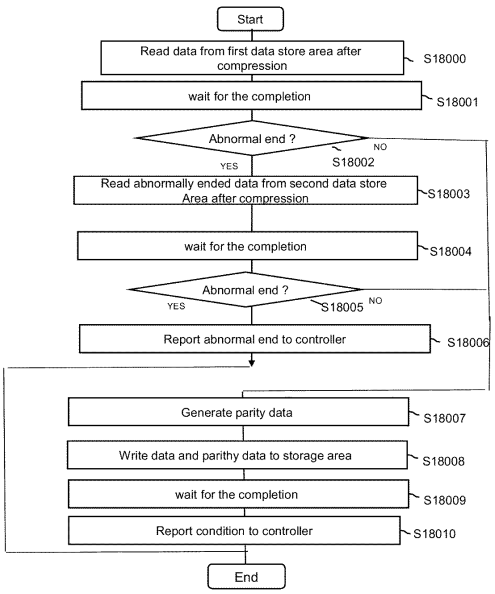


Fig. 38

【図３９】

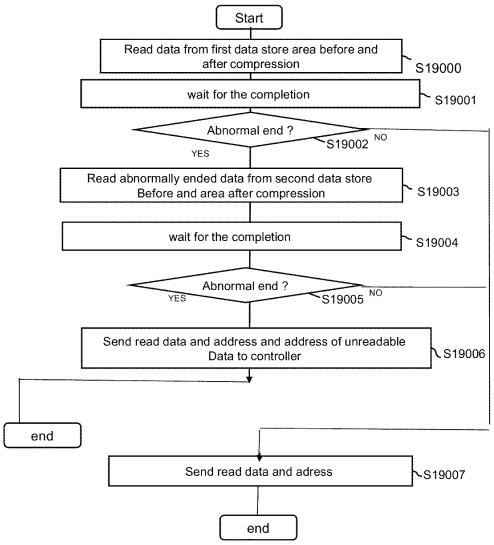


Fig. 39

【図４０】

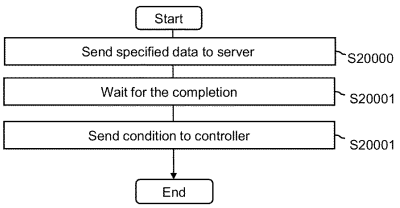


Fig. 40

【図４１】

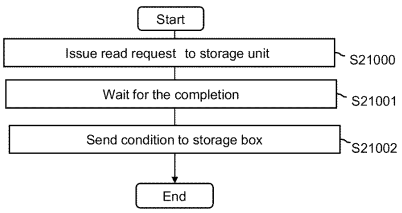


Fig. 41

フロントページの続き

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 田名網 忠雄

(56)参考文献 特開２００８－１０２９６７（ＪＰ，Ａ）
特開２００９－０９３２２５（ＪＰ，Ａ）
米国特許出願公開第２００２／００８７７５１（ＵＳ，Ａ１）
米国特許出願公開第２０１５／０２８６４３８（ＵＳ，Ａ１）
再公表特許第２０１５／０５２７９８（ＪＰ，Ａ１）

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ３／０６－３／０８
Ｇ０６Ｆ １３／１０－１３／１４
Ｇ０６Ｆ １１／１６