(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2020-140325
(P2020-140325A)
(43)公開日　令和2年9月3日(2020.9.3)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	302A	5B205
G06F 13/10	(2006.01)	G06F	3/06	302Z	
G06F 12/0866	(2016.01)	G06F	13/10	340A	
G06F 12/0868	(2016.01)	G06F	12/0866	100	
		G06F	12/0868	105	

審査請求 有　請求項の数 11　ＯＬ　（全 33 頁）

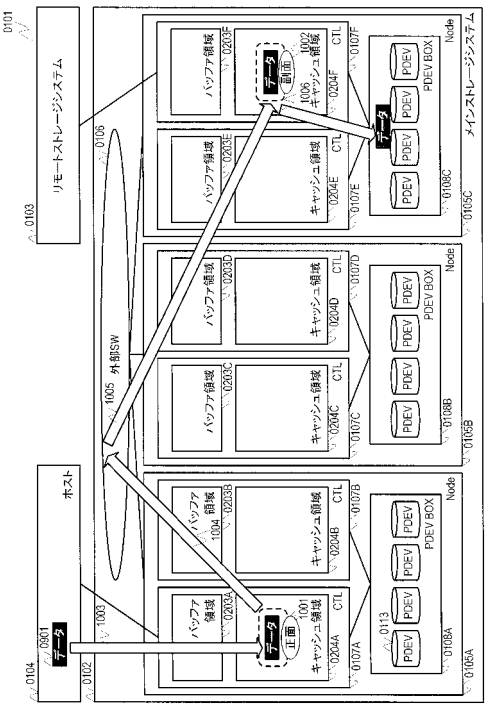
(21)出願番号　特願2019-33920(P2019-33920)
(22)出願日　平成31年2月27日(2019.2.27)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110001678
特許業務法人藤央特許事務所
(72)発明者　鶴谷　昌弘
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　吉原　朋宏
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　逢見　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

最終頁に続く

(54)【発明の名称】複数のストレージノードを含むストレージシステム

(57)【要約】
【課題】ストレージシステムの性能を向上する。
【解決手段】ストレージシステムは、ネットワークを介して通信する複数のストレージノードを含む。複数のストレージノードのそれぞれは、１以上のコントローラを含む。コントローラにおける少なくとも１つのコントローラは、ホストからライトデータを受信するコントローラと、ライトデータを処理するコントローラとに基づいて、ライトデータを格納するキャッシュサブ領域を確保する少なくとも二つのコントローラを指定し、指定したコントローラにおいて、キャッシュサブ領域を確保する。
【選択図】図１０



【特許請求の範囲】
【請求項１】
ネットワークを介して通信する複数のストレージノードを含むストレージシステムであって、
前記複数のストレージノードのそれぞれは、１以上のコントローラを含み、
前記コントローラにおける少なくとも１つのコントローラは、
ホストからライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとに基づいて、前記ライトデータを格納するキャッシュサブ領域を確保する少なくとも二つのコントローラを指定し、
指定したコントローラにおいて、キャッシュサブ領域を確保する
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、
前記少なくとも１つのコントローラは、さらに、各コントローラが属するストレージノードの、前記ライトデータを格納する記憶デバイスへの接続の有無に基づいて、前記二つのコントローラを指定する
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、
前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとは、別のコントローラであり、
前記少なくとも１つのコントローラは、前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定する
ストレージシステム。
【請求項４】
請求項３に記載のストレージシステムであって、
前記指定した二つのコントローラのうち、前記ライトデータを格納する記憶デバイスに接続されたコントローラは、その前記キャッシュサブ領域に記憶された前記ライトデータを前記記憶デバイスに格納する
ストレージシステム。
【請求項５】
請求項４に記載のストレージシステムであって、
前記指定した二つのコントローラのいずれもが、前記記憶デバイスに接続されていない場合、二つのコントローラのうちいずれかが、前記記憶デバイスに接続された他のコントローラのバッファ領域に前記ライトデータを転送し、
前記他のコントローラが、前記バッファ領域に格納された前記ライトデータを、前記記憶デバイスに格納する
ストレージシステム。
【請求項６】
請求項２に記載のストレージシステムであって、
前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとは、同じコントローラであり、
前記少なくとも１つのコントローラは、前記同じコントローラと、他の１つのコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定し、
前記指定する二つのコントローラのうち少なくとも１つは、前記記憶デバイスに接続されている
ストレージシステム。
【請求項７】
請求項２に記載のストレージシステムにおいて、

前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを格納する記憶デバイスへの接続された他のコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定する
ストレージシステム。
【請求項８】
請求項２に記載のストレージシステムであって、
前記少なくとも１つのコントローラは、前記ホストから前記ライトデータを受信するコントローラと、前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを格納する記憶デバイスへの接続された他のコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定する
ストレージシステム。
【請求項９】
請求項２に記載のストレージシステムであって、
前記記憶デバイスは、前記ストレージノードの内に設けられた記憶媒体、またはリモートストレージシステムである
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であって、
前記少なくとも１つのコントローラは、
前記ホストからのライト要求のアクセスパタンがシーケンシャルで場合に、前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを処理する前記コントローラとに基づいて、前記コントローラの指定を行い、
前記ホストからのライト要求のアクセスパタンがランダムで場合に、キャッシュ利用効率に基づいて、前記コントローラの指定を行う
ストレージシステム。
【請求項１１】
ネットワークを介して通信する複数のストレージノードを含むストレージシステムの制御方法であって、
前記複数のストレージノードのそれぞれは、１以上のコントローラを含み、
前記制御方法は、
ホストからライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとに基づいて、前記ライトデータを格納するキャッシュサブ領域を確保する少なくとも二つのコントローラを指定し、
指定したコントローラにおいて、キャッシュサブ領域を確保する
制御方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、複数のストレージノードを含むストレージシステムに関する。
【背景技術】
【０００２】
近年、ＩｏＴ（Ｉｎｔｅｒｎｅｔ ｏｆ Ｔｈｉｎｇｓ）やＡＩ（Ａｒｔｉｆｉｃａｌ Ｉｎｔｅｌｉｇｅｎｃｅ）に代表されるような膨大なデータを蓄積、及び分析することにより、新たな価値を生み出す技術の重要性が増している。これらの技術では、膨大なデータを蓄積可能な容量だけでなく、蓄積したデータを分析するために高性能なＩ／Ｏ（Ｉｎｐｕｔ／Ｏｕｔｐｕｔ）性能を有するストレージシステムが必要となる。
【０００３】
一方、ストレージシステムの導入にあたっては、導入コストを抑えるため、導入初期は小規模な構成のシステムを導入し、事業規模の拡大に合わせてシステムを拡張していくことが望ましい。これを実現する方式の一つとして、スケールアウト型ストレージシステム

がある。スケールアウト型ストレージシステムでは、事業規模が拡大し、必要となるＩ／Ｏ性能が増加してきたら、ストレージ装置（Ｎｏｄｅ）の台数を増やすことで、ストレージシステムのＩ／Ｏ性能を向上することが可能である。

【０００４】

スケールアウト型ストレージシステムのＩ／Ｏ性能を向上させる１つの手段として、スケールアウト型ストレージシステムのキャッシュ制御が考えられる。キャッシュ制御に関して、例えば特許文献１の技術が知られている。特許文献１では、キャッシュ制御として、キャッシュを割り当てるＮｏｄｅを制御することで、リード性能を向上させる方式が述べられている。

【先行技術文献】

【特許文献】

【０００５】

【特許文献１】特願２０１７－５２４２４８号公報

【発明の概要】

【発明が解決しようとする課題】

【０００６】

一般に、スケールアウト型ストレージシステムにおけるＮｏｄｅ間接続は、「疎結合」である。本明細書において、「疎結合」とは、一方のＮｏｄｅから他方のＮｏｄｅのキャッシュメモリ領域（以下単に、キャッシュ領域）にデータを入出力できないＮｏｄｅ間接続を意味する。疎結合のストレージシステムでは、一方のＮｏｄｅは、自分のキャッシュ領域を管理するが、他方のＮｏｄｅのキャッシュ領域を管理しない。

【０００７】

このため、一方のＮｏｄｅは、自分のキャッシュ領域からデータの格納先キャッシュセグメント（キャッシュサブ領域）を確保できるが、他のＮｏｄｅのキャッシュ領域からキャッシュセグメントを確保できない。結果として、疎結合のＮｏｄｅ間でＩ／Ｏデータ（Ｉ／Ｏ要求に従い入出力される対象のデータ）が転送された場合、そのＩ／Ｏデータは、転送元のＮｏｄｅのキャッシュ領域にも転送先のＮｏｄｅのキャッシュ領域にも格納されることになる。

【０００８】

そこで、スケールアウト型ストレージシステムにおけるＮｏｄｅ間接続を、「密結合」にすることを検討する。本明細書において、「密結合」とは、一方のＮｏｄｅから他方のＮｏｄｅのキャッシュ領域にＩ／Ｏデータを入出力できるＮｏｄｅ間接続を意味する。密結合のストレージシステムでは、一方のＮｏｄｅは、自分のキャッシュ領域だけでなく、他方のＮｏｄｅのキャッシュ領域を管理する。このため、一方のＮｏｄｅは、自分のキャッシュ領域と他のＮｏｄｅのキャッシュ領域のいずれからもキャッシュセグメントを確保できる。結果として、密結合のＮｏｄｅ間でＩ／Ｏデータが転送された場合、そのＩ／Ｏデータは、転送元のＮｏｄｅのキャッシュ領域と転送先のＮｏｄｅのキャッシュ領域とのいずれか一方にのみ格納されることになる。

【０００９】

このように、密結合では、１つのＩ／Ｏデータについて、キャッシュセグメントが確保されるキャッシュ領域は１つでよい。なお、ライトにおいて、キャッシュセグメントの二重化（冗長化）が行われることがある。その場合、メインのキャッシュセグメント（正面）が確保されるキャッシュ領域は１つでよく、サブのキャッシュセグメント（副面）が確保されるキャッシュ領域は１つ以上でよい。以降、明細書においては、「密結合」のスケールアウト型ストレージシステムを前提とする。

【００１０】

スケールアウト型ストレージシステムは、前述のとおり、性能を拡張性可能な点で、従来の非スケールアウト型ストレージに対する利点がある一方、Ｎｏｄｅ間でＩ／Ｏデータを複写する回数が多くなり、Ｎｏｄｅ間を接続する経路の帯域がボトルネックとなり、ストレージシステムのＩ／Ｏ性能が下がる場合がある。そこで、特許文献１では、Ｎｏｄｅ

間でのＩ／Ｏデータ複写回数が小さくなるよう、Ｉ／Ｏパタンと接続形態の一方又は両方を用いて、キャッシュを確保するＮｏｄｅ（Ｎｏｄｅに複数のストレージコントローラ（ＣＴＬ）が含まれる場合、ＣＴＬ）を決定する方式が述べられており、これによりストレージシステムのリード性能を向上させることができる。

【００１１】

一方、ライト性能については、特許文献１では改善方式が述べられていない。一般的にライトバック方式を採用するストレージシステムでは、キャッシュメモリにＩ／Ｏデータを格納した時点でＩ／Ｏ要求元のホストにライト完了を通知する。このため、単一のＮｏｄｅ障害でＩ／Ｏデータを消失しないよう、Ｉ／Ｏデータを最終記憶媒体に格納するまでの間、２つ以上のＮｏｄｅでＩ／Ｏデータを複製してキャッシュする方式が用いられる。このため、リードと比較し、ライトの場合はＮｏｄｅ間のデータ複写回数が多くなり、Ｎｏｄｅ間のデータパス帯域がボトルネックとなった場合、従来の非スケールアウト型ストレージよりも性能が低下する。

【００１２】

一方、ストレージシステムでは、Ｉ／Ｏのレスポンス性能を改善するため、高頻度でアクセスされるデータを可能な限りキャッシュに保持しておくことが望ましい。このため、単純にＣＴＬ間のデータ複写が少なくなるようにキャッシュを確保しても、ホストからのアクセスパタンによっては、特定のＮｏｄｅでキャッシュヒット率の高いデータがキャッシュに保持できず、キャッシュヒット率が低下することで、レスポンス性能の悪化を招いてしまう。よって、レスポンス性能の低下を招くことなく、ライトスループット性能を向上させることが課題である。

【課題を解決するための手段】

【００１３】

本発明の一態様は、ネットワークを介して通信する複数のストレージノードを含むストレージシステムであって、前記複数のストレージノードのそれぞれは、１以上のコントローラを含み、前記コントローラにおける少なくとも１つのコントローラは、ホストからライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとに基づいて、前記ライトデータを格納するキャッシュサブ領域を確保する少なくとも二つのコントローラを指定し、指定したコントローラにおいて、キャッシュサブ領域を確保する。

【発明の効果】

【００１４】

本発明の一態様により、ストレージシステムの性能を向上させる。

【図面の簡単な説明】

【００１５】

【図１】メインストレージシステムを含む情報システムの構成例を示す図である。
【図２】主記憶の構成例を示す図である。
【図３】プログラム領域の構成例を示す図である。
【図４】管理情報領域の構成例を示す図である。
【図５】ＰＤＥＶ管理テーブルの構成例を示す図である。
【図６】ＬＤＥＶ管理テーブルの構成例を示す図である。
【図７】バッファセグメント管理テーブルの構成例を示す図である。
【図８】キャッシュセグメント管理テーブルの構成例を示す図である。
【図９】メインストレージシステムにおいて、従来方式におけるシーケンシャルライト時のデータ転送の一例を示す図である。
【図１０】メインストレージシステムにおいて、本発明に基づく方式におけるシーケンシャルライト時のデータ転送の一例を示す図である。
【図１１】メインストレージシステムにおけるホストらのＩ／Ｏ要求に基づく処理流れの一例を示すシーケンス図である。
【図１２】メインストレージシステムにおけるライト要求を処理する過程でセグメントを確保するストレージコントローラを決定する処理の一例を示すシーケンス図である。

【図１３】メインストレージシステムにおける特定のストレージコントローラを指定してセグメントを確保するストレージコントローラを決定する処理の一例を示すシーケンス図である。
【図１４】メインストレージシステムにおけるストレージコントローラ間のデータ複写回数を最小化するようにセグメントを確保するストレージコントローラを決定する処理の一例を示すシーケンス図である。
【図１５】メインストレージシステムにおけるセグメントを確保する処理の一例を示すシーケンス図である。
【図１６】メインストレージシステムにおけるライト処理の一例を示すシーケンス図である。
【図１７】メインストレージシステムにおけるセグメントを解放する処理の一例を示すシーケンス図である。
【発明を実施するための形態】
【００１６】
以下、図面を用いて実施例を説明する。なお、実施例は本発明を実現するための一例に過ぎず、本発明の技術的範囲を限定するものではない。以下の説明では、「ｘｘｘテーブル」の表現にて各種情報を説明することがあるが、各種情報は、テーブル以外のデータ構造で表現されていてもよい。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」と呼ぶことができる。
【００１７】
以下の説明では、要素の識別情報として番号が使用されるが、他種の識別情報（例えば名前、識別子）が使用されて良い。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、参照符号における共通符号（又は参照符号）を使用し、同種の要素を区別して説明する場合は、参照符号（又は要素のＩＤ）を使用することがある。
【００１８】
以下の説明では、「主記憶」は、メモリを含んだ１以上の記憶デバイスでよい。例えば、主記憶は、主記憶デバイス（典型的に揮発性の記憶デバイス）及び補助記憶デバイス（典型的には不揮発性の記憶デバイス）のうちの少なくとも主記憶デバイスでよい。また、記憶部は、キャッシュ領域（例えばキャッシュメモリ又はその一部領域）とバッファ領域（例えばバッファメモリ又はその一部領域）とのうちの少なくとも１つを含んでもよい。
【００１９】
以下の説明では、「ＰＤＥＶ」は、物理的な記憶デバイスを意味し、典型的には、不揮発性の記憶デバイス（例えば補助記憶デバイス）でよい。ＰＤＥＶは、例えば、ＨＤＤ（Ｈａｒｄ Ｄｉｓｋ Ｄｒｉｖｅ）又はＳＳＤ（Ｄｏｌｉｄ Ｓｔａｔｅ Ｄｒｉｖｅ）でよい。以下の説明では、「ＲＡＩＤ」は、Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙ ｏｆ Ｉｎｄｅｐｅｎｄｅｎｔ （ｏｒ Ｉｎｅｘｐｅｎｓｉｖｅ） Ｄｉｓｋｓの略である。
【００２０】
ＲＡＩＤグループは、複数のＰＤＥＶで構成され、そのＲＡＩＤグループに関連付けられたＲＡＩＤレベルに従いデータを記憶する。ＲＡＩＤグループは、パリティグループと呼ばれてもよい。パリティグループは、例えば、パリティを格納するＲＡＩＤグループのことでよい。以下の説明では、「ＬＤＥＶ」は、論理的な記憶デバイスを意味し、ＲＡＩＤグループ、またはＰＤＥＶから構成され、ホストは「ＬＤＥＶ」に対してＩ／Ｏ要求を実行する。
【００２１】
以下、「プログラム」を主語として処理を説明することがある場合、プログラムは、ストレージコントローラに含まれるプロセサ（例えばＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ））によって実行されることで、定められた処理を、適宜に記憶資源（例えば主記憶）及び／又は通信インタフェース装置（例えばＨＣＡ）を用いながら行うため、処理の主語がストレージコントローラ或いはプロセサとされてもよい。また、ストレージコントローラは、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。

コンピュータプログラムは、プログラムソースからインストールされてもよい。プログラムソースは、例えば、プログラム配布サーバ、又は、コンピュータ読取可能な記憶メディアであってもよい。

【００２２】

以下の説明では、「ホスト」は、ストレージシステムにＩ／Ｏ要求を送信するシステムであり、インターフェースデバイスと、記憶部（例えばメモリ）と、それらに接続されたプロセッサとを有してよい。ホストシステムは、１以上のホスト計算機で構成されてよい。少なくとも１つのホスト計算機は、物理的な計算機でよく、ホストシステムは、物理的なホスト計算機に加えて仮想的なホスト計算機を含んでよい。

【００２３】

以下、Ｉ／Ｏパタン及び接続形態に応じてキャッシュ先ストレージ装置を決定するストレージシステムの例を説明する。一般的に、ストレージシステムにおいて、ランダムアクセス時のＩ／Ｏ処理は、プロセサ処理がボトルネックとなり、シーケンシャルアクセス時のＩ／Ｏ処理は、データパス帯域（Ｎｏｄｅ間を含む）がボトルネックとなる。また、シーケンシャルアクセスされるデータは、再度アクセスされる可能性が低い。

【００２４】

そこで、以下に説明する例において、ストレージシステムは、Ｎｏｄｅ間のデータパスがボトルネックとなるシーケンシャルライト時、Ｎｏｄｅ間のデータ複写回数が少なくなるようにキャッシュ領域を確保するＮｏｄｅを決定する。さらに、一例において、ストレージシステムは、最終記憶媒体にＩ／Ｏデータを格納した後、すぐにキャッシュ領域を解放する。以下に説明する例において、ストレージシステムは、ランダムライト時、Ｎｏｄｅ間でキャッシュ利用効率が高くなるようにキャッシュ領域を確保するＮｏｄｅを決定する。

【００２５】

なお、以下に説明する実施例は、請求の範囲にかかる発明を限定するものではなく、また実施例の中で説明されている特徴の組合せの全てが発明の解決手段に必須であるとは限らない。たとえば、ストレージシステムは、ライト要求のアクセスパタン（シーケンシャルライト／ラインダムライト）と独立に、以下に説明するようにＮｏｄｅ間のデータ複写回数が少なくなるようにキャッシュ領域を確保するＮｏｄｅを決定してもよい。

【００２６】

図１は、情報システム０１０１の構成例を示す図である。情報システム０１０１は、１つ以上のメインストレージシステム０１０２、１つ以上のリモートストレージシステム０１０３、１つ以上のホスト０１０４から構成されている。リモートストレージシステム０１０３は、含んでいなくてもよい。また、図１では、メインストレージシステム０１０２とリモートストレージシステム０１０３、及びメインストレージシステム０１０２とホスト０１０４は直接接続されているが、ＳＡＮ（ＳＡＮ：Ｓｔｏｒａｇｅ Ａｒｅａ Ｎｅｔｗｏｒｋ）、ＬＡＮ（ＬＡＮ：Ｌｏａｃａｌ Ａｒｅａ Ｎｅｔｗｏｒｋ）、及びＷＡＮ（ＷＡＮ：Ｗｏｒｌｄ Ａｒｅａ Ｎｅｔｗｏｒｋ）などのネットワークを介して接続されていてもよい。

【００２７】

メインストレージシステム０１０２は、１つ以上のＮｏｄｅ０１０５（ストレージノードとも呼ぶ）、及び１つ以上の外部ＳＷ（ＳＷ：ＳＷｉｔｃｈ）０１０６から構成されている。図１では、Ｎｏｄｅ０１０５は、外部ＳＷ０１０６を介して接続されているが、Ｎｏｄｅ０１０５間で直接接続されていてもよい。

【００２８】

複数のＮｏｄｅ０１０５の各々を、１つのＮｏｄｅ０１０５を例に取り説明する。Ｎｏｄｅ０１０５は、１つ以上のストレージコントローラ（ＣＴＬ）０１０７、及び１つ以上のＰＤＥＶ ＢＯＸ０１０８を含む。また、ＰＤＥＶ ＢＯＸ０１０８は、１つ以上のＰＤＥＶ０１１３を有し、ＣＴＬ０１０７とＰＤＥＶ０１１３の接続を仲介している。

【００２９】

図１では、ＣＴＬ０１０７は、直接ＰＤＥＶ　ＢＯＸ０１０８に接続されているが、Ｓｗｉｔｃｈなどを介して接続されていてもよい。また、ＰＤＥＶ　ＢＯＸ０１０８を介さず、ＣＴＬ０１０７とＰＤＥＶ０１１３は直接接続されていてもよい。さらに、同一Ｎｏｄｅ０１０５内のＣＴＬ０１０７間は、外部ＳＷ０１０６を介して接続されているが、直接接続されていてもよい。

【００３０】

複数のＣＴＬ０１０７の各々を、１つのＣＴＬ０１０７を例に取り説明する。ＣＴＬ０１０７は、プロセサ０１０６、主記憶０１０９、ＦＥ　Ｉ／Ｆ（ＦｒｏｎｔＥｎｄ　Ｉｎｔｅｒｆａｃｅ）０１１０、ＢＥ　Ｉ／Ｆ（ＢａｃｋＥｎｄ　Ｉｎｔｅｒｆａｃｅ）０１１１、及びＨＣＡ（Ｈｏｓｔ　Ｃｈａｎｎｅｌ　Ａｄａｐｔｅｒ）０１１２を含む。ＣＴＬ０１０７を構成する各種要素の数は、１以上でよい。

【００３１】

プロセサ０１０６は、ＣＴＬ０１０７全体を制御し、主記憶０１０９に格納されたマイクロプログラムに基づき動作する。ＦＥ　Ｉ／Ｆ０１１０は、プロセサ０１０６により制御され、ホスト０１０４、及びリモートストレージシステム０１０３とＩ／Ｏ要求やＩ／Ｏデータの送受信などを実施する。ＢＥ　Ｉ／Ｆ０１１１は、プロセサ０１０６により制御され、ＰＤＥＶ　ＢＯＸ０１０６を介し、ＰＤＥＶ０１１３とＩ／Ｏデータの送受信などを実施する。ＨＣＡ０１１１は、プロセサ０１０６により制御され、外部ＳＷ０１０６を介して他のＣＴＬ０１０７と制御情報やＩ／Ｏデータの送受信などを実施する。

【００３２】

本実施例では、Ｎｏｄｅ０１０５間接続は、密結合である。密結合のストレージシステムでは、Ｎｏｄｅ０１０５間の通信で使用される通信プロトコルと、ＣＴＬ０１０７内のデバイス（要素）間の通信で使用される通信プロトコルは、同じである。どちらの通信プロトコルも、例えば、ＰＣＩｅ（ＰＣＩ－ｅｘｐｒｅｓｓ）である。

【００３３】

一方、疎結合のストレージ装置では、Ｎｏｄｅ間の通信で使用される通信プロトコルと、ストレージ装置内のデバイス間の通信で使用される通信プロトコルは、異なる。前者の通信プロトコルは、例えば、ＦＣ（Ｆｉｂｒｅ　Ｃｈａｎｎｅｌ）又はＩＰ（Ｉｎｔｅｒｎｅｔ　Ｐｒｏｔｏｃｏｌ）であり、後者の通信プロトコルは、例えば、ＰＣＩｅである。なお、密結合及び疎結合の定義は、既に説明した通りである。

【００３４】

図２は、主記憶０１０９の構成例を示す図である。主記憶０１０９には、プログラム領域０２０１、管理情報領域０２０２、バッファ領域０２０３、及びキャッシュ領域０２０４が確保されている。プログラム領域０２０１は、同一ＣＴＬ０１０７のプロセサ０１０６が処理を実施するための各プログラムが格納されている領域である。管理情報領域０２０２は、メインストレージシステム０１０２内の全てのプロセサからアクセスされる領域で、各種管理テーブルが格納される領域である。

【００３５】

バッファ領域０２０３、及びキャッシュ領域０２０４は、同一ＣＴＬ０１０７のＦＥ　Ｉ／Ｆ０１１０、ＢＥ　Ｉ／Ｆ０１１１、及びＨＣＡ０１１２等によるデータ転送の再に、一時的にデータが格納される領域である。なお、バッファ領域０２０３とキャッシュ領域０２０４は、複数のセグメント（キャッシュ領域を区切った単位であり、キャッシュサブ領域とも呼ぶ）で構成され、セグメント単位で領域が確保される。また、バッファ領域０２０３から読みだされたデータは、バッファ領域０２０３には残らない。一方、キャッシュ領域０２０４から読み出されたデータは、キャッシュ領域０２０４に残る。

【００３６】

図３は、プログラム領域０２０１に格納されるプログラムの一例を示す図である。プログラム領域０２０１には、例えば、Ｉ／Ｏ要求のＣＴＬ間振り分けプログラム０３０１、Ｉ／Ｏプログラム０３０２、Ｉ／Ｏパタン判定プログラム０３０３、セグメント確保プログラム０３０４、セグメント解放プログラム０３０５、フリーセグメント確保プログラム

０３０６、及びストレージコントローラ間データ転送プログラム０３０７が格納される。
【００３７】
ホスト０１０４からのＩ／Ｏ要求を受けると、Ｉ／Ｏ要求のＣＴＬ間振り分けプログラム０３０１は、当該Ｉ／Ｏ要求を担当するＣＴＬ０１０７を決定し、振り分ける。振り分けは、あらかじめＬＤＥＶごとにＩ／Ｏ要求を処理するＣＴＬ０１０７を決めておいてもよいし、動的に決定してもよい。次に、Ｉ／Ｏ要求を割り振られたＣＴＬ０１０７のプロセサ０１０６は、Ｉ／Ｏプログラム０３０２を実行することにより、Ｉ／Ｏ要求に従い、対応する処理を実行する。
【００３８】
Ｉ／Ｏパタン判定プログラム０３０３は、例えばＩ／Ｏプログラム０３０２がＩ／Ｏ要求を処理する過程で呼び出され、Ｉ／Ｏ要求のアクセスパタン（アクセスパタン又はＩ／Ｏパタン）が、シーケンシャルであるか、ランダムであるかを判定する。セグメント確保プログラム０３０４は、例えばＩ／Ｏプログラム０３０２がＩ／Ｏ要求を処理する過程で呼び出され、バッファセグメント、及びキャッシュセグメントを確保する。セグメント解放プログラム０３０５は、例えばＩ／Ｏプログラム０３０２がＩ／Ｏ要求を処理する過程で呼び出され、バッファセグメント、及びキャッシュセグメントを解放する。
【００３９】
フリーセグメント確保プログラム０３０６は、Ｉ／Ｏ要求とは非同期的に実行され、バッファセグメント、及びキャッシュセグメントのうち、確保可能な状態のセグメント（フリーセグメント）の量を一定以上に保つ。ストレージコントローラ間データ転送プログラム０３０７は、例えばＩ／Ｏプログラム０３０２がＩ／Ｏ要求を処理する過程などで呼び出され、ＣＴＬ０１０７間でのデータ転送を実行する。
【００４０】
図４は、管理情報領域０２０２に格納される情報の一例を示す図である。管理情報領域０２０２には、ＰＤＥＶ管理テーブル０４０１、ＬＤＥＶ管理テーブル０４０２、バッファセグメント管理テーブル０４０３、及びキャッシュセグメント管理テーブル０４０４が格納される。
【００４１】
ＰＤＥＶ管理テーブル０４０１は、メインストレージシステム１０２内の全てのＰＤＥＶ０１１４の状態や、ＣＴＬ０１０７との対応関係を示す。ＬＤＥＶ管理テーブル０４０２は、メインストレージシステム１０２内の全てのＰＤＥＶ０１１４とＬＤＥＶの対応関係を示す。バッファセグメント管理テーブル０４０３は、メインストレージシステム１０２内の全てのバッファ領域０２０３を管理するのに用いられる。キャッシュセグメント管理テーブル０４０４は、メインストレージシステム１０２内の全てのキャッシュ領域０２０４を管理するのに用いられる。
【００４２】
図５は、ＰＤＥＶ管理テーブル０４０１の構成例を示す図である。ＰＤＥＶ管理テーブル０４０１は、ＰＤＥＶ＃０５０１、容量０５０２、状態０５０３、及び接続ＣＴＬ＃０５０４のエントリを有する。ＰＤＥＶ＃０５０１は、ＰＤＥＶ０１１４の識別子である。容量０５０２は、ＰＤＥＶ０１１４のデータを格納可能な容量を示す。状態０５０３は、ＰＤＥＶ０１１４が正常に動作中か否か（故障していないか）の状態を示す。接続ＣＴＬ＃０５０４は、ＰＤＥＶ０１１４に接続され、アクセスするＣＴＬ０１０７を示す。
【００４３】
図６は、ＬＤＥＶ管理テーブル０４０２の構成例を示す図である。ＬＤＥＶ管理テーブル０４０２は、ＬＤＥＶ＃０６０１、容量０６０２、状態０６０３、冗長構成０６０４、及び構成情報０６０５のエントリを有する。
【００４４】
ＬＤＥＶ＃０６０１は、ＬＤＥＶの識別子である。容量０６０２は、ＬＤＥＶにデータを格納可能な容量を示す。状態０５０３は、ＬＤＥＶへ正常にＩ／Ｏ可能か否かを示す。冗長構成０６０４は、ＲＡＩＤによる冗長化、またはリモートストレージシステム０１０

３とのストレージシステム冗長化の状態を示す。構成情報０６０５は、ＬＤＥＶに属するＰＤＥＶ０１１４、及びストレージシステム冗長化の対応リモートストレージシステム０１０３を示す。

【００４５】

図７は、バッファセグメント管理テーブル０４０３の構成例を示す図である。バッファセグメント管理テーブル０４０３は、セグメント＃０７０１、ＣＴＬ＃０７０２、状態０７０３を有する。セグメント＃０７０１は、バッファセグメントの識別子である。ＣＴＬ＃０７０２は、ＣＴＬ０１０７の識別子であり、当該バッファセグメントの利用権利を有したＣＴＬ０１０７を示している。状態０７０３は、バッファセグメントが確保されているか否か（ロック、フリー）を示す。

【００４６】

図８は、キャッシュセグメント管理テーブル０４０４の構成例を示す図である。キャッシュセグメント管理テーブル（０４０４）は、セグメント＃０８０１、ＬＤＥＶ＃０８０２、ＬＤＥＶ内セグメント＃０８０３、ＣＴＬ＃０８０４、状態０８０５、及び属性０８０６を有する。

【００４７】

セグメント＃０８０１は、キャッシュセグメントの識別子である。ＬＤＥＶ＃は、ＬＤＥＶの識別子である。ＬＤＥＶ内セグメント＃０８０３は、当該キャッシュセグメントがＬＤＥＶの論理アドレス空間上のどこに割り当てられているかを一意に確定する識別子である。ＣＴＬ＃０８０４は、ＣＴＬ０１０７の識別子であり、当該バッファセグメントの利用権利を有したＣＴＬ０１０７を示している。

【００４８】

状態０８０５は、キャッシュセグメントの状態を表し、「フリー」は未使用、「ロック」は確保済みだが未使用、「クリーン」確保済みでＰＤＥＶにデータ格納済み、及び「ダーティー」は確保済みでＰＤＥＶにデータ未格納であることを示す。属性０８０６は、当該キャッシュセグメントが正面、または副面のどちらで確保されているかを示す。

【００４９】

図９は、メインストレージシステム０１０２において、従来方式を用いた、シーケンシャルライトにおけるライトシーケンスの一例を示したものである。本図は、ライトシーケンスを説明するため、情報システム０１０１の構成要素を一部省略して記載している。また、メインストレージシステム０１０２を構成する要素のうち、複数存在し、かつ説明上識別する必要のあるものについて、４桁の通番に加え、アルファベット一文字を付加し、区別している。

【００５０】

ライト処理で必要となる、キャッシュセグメントの正面０９０２、及び副面０９０３は、それぞれＣＴＬ０１０７Ｄのキャッシュ領域０２０４Ｄ、及びＣＴＬ０１０７Ｂのキャッシュ領域０２０４Ｂに確保されている。データ０９０１の流れを示す矢印０９０４、０９０５、０９０６、０９０７、０９０８、０９０９、０９１０、及び０９１１より、データ０９０１のＣＴＬ間データ複写の回数は、３回である。

【００５１】

図１０は、メインストレージシステム０１０２において、本実施例の方式を用いた、シーケンシャルライトにおけるライトシーケンスの一例を示したものである。本図は、ライトシーケンスを説明するため、情報システム０１０１の構成要素を一部省略して記載している。

【００５２】

図９との差分は、正面１００１と副面１００２を確保するＣＴＬ０１０７である。データ０９０１の流れを示す矢印１００３、１００４、１００５、及び１００６に示すとおり、データ０９０１のＣＴＬ間複写回数は、図９の３回から１回に減少している。

【００５３】

図１１は、ストレージシステム（メイン）０１０２がホスト０１０４から受領したＩ／

要求を処理する流れの一例を示したフロー図である。ＣＴＬ０１０７は、ホスト０１０４からのＩ／Ｏ要求を受領する（１１０１）。ホスト０１０４からＩ／Ｏ要求を受領したＣＴＬ０１０７は、当該Ｉ／Ｏを担当するＣＴＬ０１０７を決定し、その担当ＣＴＬ０１０７が、処理を継続する（１１０２）。

【００５４】

担当ＣＴＬ０１０７は、Ｉ／Ｏ要求を解析し、アクセス先のＬＤＥＶ、Ｉ／Ｏ要求種別（リード要求、ライト要求、他）、及びアクセスパタン（ランダムアクセスまたはシーケンシャルアクセス）を判定する。例えば、アクセス先の連続アドレスが所定値より大きい場合に、シーケンシャルと判定される。アクセスパタンは、ＬＤＥＶごとに直近のＩ／Ｏ履歴を取得して判定することにより、ストレージシステム（メイン）が判定してもよいし、ホスト０１０４からアクセスパタンに関するヒント情報を取得して判定してもよい（１１０３）。

【００５５】

担当ＣＴＬ０１０７は、ホスト０１０４からのＩ／Ｏ要求の解析１１０３で明らかにしたＩ／Ｏ要求種別を用いて、Ｉ／Ｏ要求種別がライト要求であるか否かを判定する（１１０４）。Ｉ／Ｏ要求種別がライト要求の場合（１１０４：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１１０５へ進む（Ａ）。一方、Ｉ／Ｏ要求種別がライト要求でない場合（１１０４：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１１１０へ進む（Ｂ）。

【００５６】

（Ａ）

Ｉ／Ｏ要求種別がライト要求の場合（１１０４：Ｙｅｓ）、担当ＣＴＬ０１０７は、ＰＤＥＶ管理テーブル０４０１とＬＤＥＶ管理テーブル０４０２を参照し、当該Ｉ／ＯのＬＤＥＶに関する情報、及び当該ＬＤＥＶを構成するＰＤＥＶの情報を取得する（１１０５）。担当ＣＴＬ０１０７は、セグメント確保ＣＴＬ決定処理を呼び出し、ライト処理に必要なキャッシュセグメントを確保するＣＴＬ０１０５を決定する（１１０６）。なお、セグメント確保ＣＴＬ決定処理１１０６の詳細は、後述する。

【００５７】

担当ＣＴＬ０１０７は、セグメント確保ＣＴＬ決定処理１１０６での決定に基づき、キャッシュセグメント、及びバッファセグメントを確保する（１１０７）。なお、セグメント確保処理１１０７の詳細は、後述する。担当ＣＴＬ０１０７は、ライト処理を呼び出し、セグメント確保処理１１０７で確保したセグメントを使用し、ライト処理を実行する（１１０８）。なお、ライト処理１１０８の詳細は、後述する。

【００５８】

担当ＣＴＬ０１０７は、セグメント解放処理を呼びだし、セグメント確保処理１１０７で確保したセグメントの一部、または全てを解放し、処理を終了する（１１０９）。なお、セグメント解放処理１１０９の詳細は、後述する。

【００５９】

（Ｂ）

Ｉ／Ｏ要求種別がライト要求でない場合（１１０４：Ｎｏ）、担当ＣＴＬ０１０７は、Ｉ／Ｏ要求に基づき処理を実行した（１１１０）後、処理を終了する。当該処理は、本実施例の影響を受けないため、詳細な説明を省略する。

【００６０】

図１２は、図１１で説明したフローから呼び出される、セグメント確保ＣＴＬ決定処理１１０６の流れの一例を示したフロー図である。担当ＣＴＬ０１０７は、ホスト０１０４からのＩ／Ｏ要求の解析１１０３で明らかにしたアクセスパタンを参照し、アクセスパタンはシーケンシャルであるか否かを判定する（１２０１）。アクセスパタンがシーケンシャルである場合（１２０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１１０２に進む（Ａ）。一方、アクセスパタンがシーケンシャルでない場合（１２０１：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１２０８に進む（Ｇ）。

【００６１】

（Ａ）
アクセスパタンがシーケンシャルである場合（１２０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１１０５で取得したＬＤＥＶ情報を用いて、ストレージ間冗長化のＬＤＥＶに対するライトであるか否かを判定する（１２０２）。ストレージ間冗長化のＬＤＥＶに対するライトデータは、リモートストレージシステム０１０３に接続されているＣＴＬ０１０７によって、リモートストレージシステム０１０３に転送される。ストレージ間冗長化のＬＤＥＶに対するライトの場合（１２０２：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１２０３に進む（Ｂ）。一方、ストレージ間冗長化のＬＤＥＶに対するライトでない場合（１２０２：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１２０５に進む（Ｃ）。

【００６２】
（Ｂ）
ストレージ間冗長化のＬＤＥＶに対するライトの場合（１２０２：Ｙｅｓ）、担当ＣＴＬ０１０７は、当該ＬＤＥＶがストレージ間冗長化をしているストレージコントローラ（リモート）０１０３が接続されているＣＴＬのＣＴＬ＃を取得し（１２０３）、ステップ１２０４に進む（Ｅ）。なお、取得方法は、あらかじめストレージシステム同士の接続状態を管理するテーブルを作成しておき、これを参照してもよいし、ストレージシステム（メイン）０１０２を構成する全てのＣＴＬ０１０７に問い合わせてもよい。

【００６３】
（Ｃ）
ストレージ間冗長化のＬＤＥＶに対するライトでない場合（１２０２：Ｎｏ）、担当ＣＴＬ０１０７は、当該ＬＤＥＶＥへのライトについて、担当ＣＴＬ０１０７でホスト０１０４からライトされたデータを参照する処理（例えば、スナップショットなどライトに起因する処理）が存在するか否かを判定する（１２０５）。担当ＣＴＬ０１０７でデータを参照する処理が存在する場合（１２０５：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１２０６へ進む（Ｄ）。一方、担当ＣＴＬ０１０７でデータを参照する処理が存在しない場合（１２０５：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１２０７へ進む（Ｆ）。

【００６４】
（Ｄ）
担当ＣＴＬ０１０７でデータを参照する処理が存在する場合（１２０５：Ｙｅｓ）、担当ＣＴＬ０１０７は、担当ＣＴＬ自身のＣＴＬ＃を取得し（１２０６）、ステップ１２０４へ進む（Ｅ）。

【００６５】
（Ｅ）
担当ＣＴＬ０１０７は、ステップ１２０３、またはステップ１２０６で取得したＣＴＬ＃を引数とし、ＣＴＬ＃指定のセグメント確保ＣＴＬ決定処理１１０６を呼び出し、指定したＣＴＬ＃でセグメントを確保したうえで、ＣＴＬ間のデータ複写回数が最小となるように、ライト処理に必要なセグメントを確保し（１２０４）、処理を終了する。なお、ＣＴＬ＃指定のセグメント確保処理１２０４の詳細は、後述する。

【００６６】
（Ｆ）
担当ＣＴＬ０１０７でデータを参照する処理が存在しない場合（１２０５：Ｎｏ）、担当ＣＴＬ０１０７は、ＣＴＬ間データ複写回数最小セグメント確保ＣＴＬ決定処理を呼び出し、ＣＴＬ間のデータ複写回数が最小となるように、ライト処理に必要なセグメントを確保し（１２０７）、処理を終了する。なお、ＣＴＬ間データ複写回数最小セグメント確保ＣＴＬ決定処理１２０７の詳細は、後述する。

【００６７】
（Ｇ）
アクセスパタンがシーケンシャルでない場合（１２０１：Ｎｏ）、担当ＣＴＬ０１０７は、キャッシュセグメントを「ＣＴＬの指定なしで確保」することに決定し、結果を返却する（１２０８）。

【００６８】
図１３は、図１２で説明したフローから呼び出される、ＣＴＬ＃指定のセグメント確保ＣＴＬ決定処理１２０４の流れの一例を示したフロー図である。担当ＣＴＬ０１０７は、キャッシュセグメントの正面を「担当ＣＴＬで確保」することに決定する（１３０１）。
【００６９】
担当ＣＴＬ０１０７は、引数として渡された指定のＣＴＬ＃を用いて、指定のＣＴＬと担当ＣＴＬ０１０７が一致するか否かを判定する（１３０２）。指定のＣＴＬと担当ＣＴＬ０１０７が一致する場合（１３０２：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１３０３に進む（Ａ）。一方、指定のＣＴＬと担当ＣＴＬ０１０７が一致しない場合（１３０２：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１３０９に進む（Ｇ）。
【００７０】
（Ａ）
指定のＣＴＬと担当ＣＴＬが一致する場合（１３０２：Ｙｅｓ）、担当ＣＴＬ０１０７は、担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されているか否かを判定する（１３０３）。担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されている場合（１３０３：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１３０４に進む（Ｂ）。一方、担当ＣＴＬにライト先のＰＤＥＶが接続されていない場合（１３０３：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１３０６に進む（Ｄ）。
【００７１】
（Ｂ）
担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されている場合（１３０３：Ｙｅｓ）、担当ＣＴＬ０１０７は、担当ＣＴＬ０１０７にホスト０１０４が接続されているか否かを判定する（１３０４）。担当ＣＴＬにホスト０１０４が接続されている場合（１３０４：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１３０５に進む（Ｃ）。一方、担当ＣＴＬ０１０７にホスト０１０４が接続されていない場合（１３０４：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１３０７に進む（Ｅ）。
【００７２】
（Ｃ）
担当ＣＴＬにホスト０１０４が接続されている場合（１３０４：Ｙｅｓ）、担当ＣＴＬ０１０７は、キャッシュセグメントの副面を「担当ＣＴＬ以外で確保」することに決定し（１３０５）、結果を返却し、処理を終了する。
【００７３】
（Ｄ）
担当ＣＴＬにライト先のＰＤＥＶが接続されていない場合（１３０３：Ｎｏ）、担当ＣＴＬ０１０７は、担当ＣＴＬ０１０７にホスト０１０４が非接続であるか否かを判定する（１３０６）。担当ＣＴＬにホスト０１０４が非接続である場合（１３０６：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１３０７に進む（Ｅ）。一方、担当ＣＴＬ０１０７にホスト０１０４が非接続でない（接続されている）場合（１３０６：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１３０８に進む（Ｆ）。
（Ｅ）
【００７４】
担当ＣＴＬにホスト０１０４が非接続である場合（１３０６：Ｙｅｓ）、担当ＣＴＬ０１０７は、キャッシュセグメントの副面を「ホスト接続ＣＴＬで確保」することに決定し（１３０７）、結果を返却し、処理を終了する。
【００７５】
（Ｆ）
担当ＣＴＬにホスト０１０４が接続されている場合（１３０６：Ｎｏ）、担当ＣＴＬ０１０７は、キャッシュセグメントの副面を「ＰＤＥＶ接続ＣＴＬで確保」することに決定し（１３０８）、結果を返却し、処理を終了する。
【００７６】

（Ｇ）
指定のＣＴＬと担当ＣＴＬ０１０７が一致しない場合（１３０２：Ｎｏ）、担当ＣＴＬ０１０７は、キャッシュセグメントの副面を「指定ＣＴＬで確保」することに決定し（１３０９）、結果を返却し、処理を終了する。
【００７７】
上述のように、ＣＴＬ＃指定のセグメント確保ＣＴＬ決定処理１２０４の開始時、リモートストレージシステムに接続されたＣＴＬまたは担当ＣＴＬが、キャッシュセグメントを確保するＣＴＬとして指定されている。リモートストレージシステムに接続されたＣＴＬが指定されている場合（１３０２：Ｎｏ）、当該ＣＴＬのキャッシュセグメントが確保される（１３０９）。
【００７８】
担当ＣＴＬがライトデータを参照する処理が存在する場合（１２０５：Ｙｅｓ）、担当ＣＴＬのキャッシュセグメントが確保される（１３０１）。上記処理のために担当ＣＴＬはライトデータの複製を必要とする。また、担当ＣＴＬの自キャッシュセグメントの確保の負荷は、他のＣＴＬのキャッシュセグメントの確保よりも小さい。したがって、ＣＴＬ間のデータ複製の回数を低減しつつ、レスポンスを改善できる。
【００７９】
ストレージ間冗長化のライトのため、リモートストレージシステムに接続されているＣＴＬはライトデータの複製を必要とする。ストレージ間冗長化のライトは長い時間を必要とし、さらに、当該ライトが完了した後にホストに対してライト完了の応答が送信される。したがって、リモートストレージシステムに接続されているＣＴＬのキャッシュセグメントを確保することで、ＣＴＬ間のデータ複製の回数を低減しつつ、レスポンスを改善できる。
【００８０】
ＣＴＬ＃指定のセグメント確保ＣＴＬ決定処理１２０４の開始時、担当ＣＴＬが指定され（１３０２：Ｙｅｓ）、担当ＣＴＬがＰＤＥＶに接続されておらず（１３０３：Ｎｏ）、担当ＣＴＬがホストに接続されている場合（１３０６：Ｎｏ）、ＰＤＥＶに接続されているＣＴＬのキャッシュセグメントが確保される（１３０８）。ＰＤＥＶはライトデータの最終格納媒体であり、ＰＤＥＶに接続されているＣＴＬは、ホスト又は他のＣＴＬからライトデータを受信する。そのため、ＣＴＬ間でのライトデータの複製回数を低減できる。
【００８１】
ＣＴＬ＃指定のセグメント確保ＣＴＬ決定処理１２０４の開始時、担当ＣＴＬが指定され（１３０２：Ｙｅｓ）、担当ＣＴＬがホストに接続されていない場合（１３０４：Ｎｏまたは１３０６：Ｙｅｓ）、ホストに接続されているＣＴＬのキャッシュセグメントが確保される（１３０７）。ホストに接続されているＣＴＬがライトデータをホストから受信するため、ＣＴＬ間でのライトデータの複製回数を低減できる。
【００８２】
ＣＴＬ＃指定のセグメント確保ＣＴＬ決定処理１２０４の開始時、担当ＣＴＬが指定され（１３０２：Ｙｅｓ）、担当ＣＴＬがＰＤＥＶに接続されており（１３０３：Ｙｅｓ）、担当ＣＴＬがホストに接続されている場合（１３０４：Ｙｅｓ）、担当ＣＴＬ以外のＣＴＬのキャッシュセグメントが確保される（１３０５）。担当ＣＴＬ以外のＣＴＬは、ライトデータの複製を必要としておらず、キャッシュデータの二重化のために、任意のＣＴＬを選択可能である。
【００８３】
なお、ＣＴＬ＃指定のセグメント確保ＣＴＬ決定処理１２０４の開始時リモートストレージシステムに接続されたＣＴＬが指定されている場合、ステップ１３０１は、担当ＣＴＬと異なるＣＴＬを指定してもよい。上記フローにおいて、担当ＣＴＬのキャッシュセグメントが副面で、もう一方のＣＴＬのキャッシュセグメントが正面であってもよい。
【００８４】

図１４は、図１２で説明したフローから呼び出される、ＣＴＬ間データ複写回数最小セグメント確保ＣＴＬ決定処理１２０７の流れの一例を示したフロー図である。

【００８５】

担当ＣＴＬ０１０７は、担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されているか否かを判定する（１４０１）。担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されている場合（１４０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１４０２に進む（Ａ）。一方、担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されていない場合（１４０１：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１４０６に進む（Ｄ）。

【００８６】

（Ａ）

担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されている場合（１４０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されているか否かを判定する（１４０２）。ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されている場合（１４０２：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１４０３に進む（Ｂ）。一方、ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されていない場合（１４０２：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１４０７に進む（Ｅ）。

【００８７】

（Ｂ）

ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されている場合（１４０２：Ｙｅｓ）、担当ＣＴＬは、担当ＣＴＬ０１０７にホスト０１０４が接続されているか否かを判定する（１４０３）。担当ＣＴＬ０１０７にホスト０１０４が接続されている場合（１４０３：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１４０４に進む（Ｃ）。一方、担当ＣＴＬ０１０７にホスト０１０４が接続されていない場合（１３０４：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１４０７に進む（Ｅ）。

【００８８】

（Ｃ）

担当ＣＴＬ０１０７にホスト０１０４が接続されている場合（１４０３：Ｙｅｓ）、担当ＣＴＬ０１０７は、キャッシュセグメントの正面を「担当ＣＴＬで確保」することに決定する（１４０４）。担当ＣＴＬ０１０７は、キャッシュセグメントの副面を「担当ＣＴＬ以外で確保」することに決定し（１４０５）、結果を返却し、処理を終了する。

【００８９】

（Ｄ）

担当ＣＴＬ０１０７にライト先のＰＤＥＶが接続されていない場合（１４０１：Ｎｏ）、担当ＣＴＬ０１０７は、ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されているか否かを判定する（１４０６）。ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されている場合（１４０６：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１４０７に進む（Ｅ）。一方、ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されていない場合（１４０６：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１４０９に進む（Ｆ）。

【００９０】

（Ｅ）

ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されている場合（１４０６：Ｙｅｓ）、担当ＣＴＬ０１０７は、キャッシュセグメントの正面を「担当ＣＴＬで確保」することに決定する（１４０７）。担当ＣＴＬ０１０７は、キャッシュセグメントの副面を「ホスト接続ＣＴＬで確保」することに決定し（１４０８）、結果を返却し、処理を終了する。

【００９１】

（Ｆ）

ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されていない場合（１４０６：Ｎｏ）、担当ＣＴＬ０１０７は、キャッシュセグメントの正面を「ホスト接続ＣＴＬで確保」することに決定する（１４０９）。担当ＣＴＬ０１０７は、キャッシュセグメントの副面を「ＰＤＥＶ接続ＣＴＬで確保」することに決定し（１４１０）、結果を返却し、処理を終了

する。
【００９２】
図１４を参照した上記フローは、シーケンシャルアクセス時のライトに限定して、ホストが接続されたＣＴＬにキャッシュの正面及び副面の一方の面を配置し、ＰＤＥＶが接続されているＣＴＬに他方の面を配置する。一つのＣＴＬがホスト及びＰＤＥＶに接続されている場合、当該ＣＴＬ及び他のＣＴＬが選択される。ホストに接続されたＣＴＬはホストからライトデータを受信する。ＰＤＥＶに接続されたＣＴＬは、ライトデータをＰＤＥＶに格納する。したがって、キャッシュデータの二重化のためのＣＴＬ間の複製回数を低減することができる。
【００９３】
図１２、１３及び１４を参照して説明したように、リモートストレージシステムに接続されたＣＴＬ、ライトデータの処理を行う担当ＣＴＬ、ホストからライトデータを受信するＣＴＬ、及びＰＤＥＶにデータを格納するＣＴＬは、それぞれ、ライトデータのキャッシュと独立して、異なる目的のために、ライトデータをホストまたは他のＣＴＬから受信することが決まっている。したがって、これらがキャッシュセグメントを確保するＣＴＬとして指定可能であることで、ライトデータのキャッシュ多重化のためのデータ転送回数を低減できる。
【００９４】
図１５は、図１１で説明したフローから呼び出される、セグメント確保処理１１０７の流れの一例を示したフロー図である。担当ＣＴＬ０１０７は、キャッシュセグメント確保ＣＴＬ決定処理１１０６での決定に基づき、キャッシュセグメントがＣＴＬ指定されているか否かを判定する（１５０１）。キャッシュセグメントがＣＴＬ指定されている場合（１５０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１５０２に進む（Ａ）。一方、キャッシュセグメントがＣＴＬ指定されていない場合（１５０１：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１５０３に進む（Ｂ）。
【００９５】
（Ａ）
キャッシュセグメントがＣＴＬ指定されている場合（１５０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、指定されたＣＴＬにてキャッシュセグメントの正面と副面を確保し、担当ＣＴＬ０１０７は、ステップ１５０４に進む（Ｃ）。
【００９６】
（Ｂ）
キャッシュセグメントがＣＴＬ指定されていない場合（１５０１：Ｎｏ）、担当ＣＴＬ０１０７は、ＣＴＬごとのキャッシュセグメント使用量を算出し、ＣＴＬ間で使用量が偏らないように正面と副面を選択した異なるＣＴＬにそれぞれ確保（例えば、最も使用量の少ないＣＴＬから順に正面と副面を確保）し、処理１５０４に進む（Ｃ）。これにより、キャッシュヒット率を上げることができる。なお、キャッシュセグメントを確保するＣＴＬは、必ずしもＣＴＬごとのキャッシュセグメント使用量に基づいて決定する必要はなく、例えば、前回に確保したＣＴＬを記憶しておき、ラウンドロビンで正面と副面を確保するＣＴＬを決定してもよい。
【００９７】
（Ｃ）
担当ＣＴＬ０１０７は、ホスト０１０４が接続されたＣＴＬにおいて正面又は副面が確保されているか否かを判定する（１５０４）。ホスト０１０４が接続されたＣＴＬにおいて正面又は副面が確保されている場合（１５０４：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１５０６に進む（Ｅ）。一方、ホスト０１０４が接続されたＣＴＬにおいて正面又は副面が確保されていない場合（１５０４：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１５０５に進む（Ｄ）。
【００９８】
（Ｄ）

ホスト０１０４が接続されたＣＴＬにおいて正面又は副面が確保されていない場合（１５０４：Ｎｏ）、担当ＣＴＬ０１０７は、ホスト０１０４が接続されたＣＴＬにおいてバッファセグメントを確保し（１５０５）、担当ＣＴＬ０１０７は、ステップ１５０６に進む（Ｅ）。

【００９９】

（Ｅ）

ホスト０１０４が接続されたＣＴＬにおいて正面又は副面が確保されている場合（１５０４：Ｙｅｓ）、担当ＣＴＬ０１０７は、ライト先のＰＤＥＶが接続されたＣＴＬにおいて正面又は副面が確保されているか否かを判定する（１５０６）。ライト先のＰＤＥＶが接続されたＣＴＬにおいて正面又は副面が確保されている場合（１５０６：Ｙｅｓ）、処理を終了する。

【００１００】

一方、ライト先のＰＤＥＶが接続されたＣＴＬにおいて正面又は副面が確保されていない場合（１５０６：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１５０７に進む。担当ＣＴＬ０１０７は、ライト先のＰＤＥＶが接続されたＣＴＬにおいてバッファセグメントを確保し（１５０７）、処理を終了する。

【００１０１】

図１６は、図１１で説明したシーケンスから呼び出される、ライト処理１１０８の流れの一例を示したシーケンス図である。以下において、担当ＣＴＬ０１０７は、他のＣＴＬ０１０７に必要な指示を行うことで、各ステップを実行する。担当ＣＴＬ０１０７は、ホスト０１０４に接続されているＣＴＬ０１０７を介して、ホスト０１０４にデータ転送を要求する（１６０１）。

【００１０２】

担当ＣＴＬ０１０７は、ホスト０１０４に接続されているＣＴＬ０１０７に確保したセグメントがキャッシュであるか否かを判定する（１６０２）。ホスト０１０４に接続されているＣＴＬ０１０７に確保したセグメントがキャッシュである場合（１６０２：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１６０３に進む（Ａ）。一方、ホスト０１０４に接続されているＣＴＬに確保したセグメントがキャッシュでない（バッファセグメントである）場合（１６０２：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１６０５に進む（Ｂ）。

【００１０３】

（Ａ）

ホスト０１０４に接続されているＣＴＬ０１０７に確保したセグメントがキャッシュである場合（１６０２：Ｙｅｓ）、担当ＣＴＬ０１０７は、ホスト０１０４から転送されたＩ／Ｏデータを当該キャッシュセグメントに格納する（１６０３）。担当ＣＴＬ０１０７は、ステップ１６０３でデータを格納したキャッシュセグメントでない方のキャッシュセグメントへ、当該Ｉ／Ｏデータを複写し（１６０４）、担当ＣＴＬ０１０７は、ステップ１６０７に進む（Ｃ）。

【００１０４】

（Ｂ）

ホスト０１０４に接続されているＣＴＬに確保したセグメントがバッファセグメントである場合（１６０２：Ｎｏ）、担当ＣＴＬ０１０７は、ホスト０１０４から転送されたＩ／Ｏデータを当該バッファセグメントに格納する（１６０５）。担当ＣＴＬ０１０７は、キャッシュセグメントの正面と副面の両方に、当該Ｉ／Ｏデータを複写し（１６０６）、担当ＣＴＬ０１０７は、ステップ１６０７に進む（Ｃ）。

【００１０５】

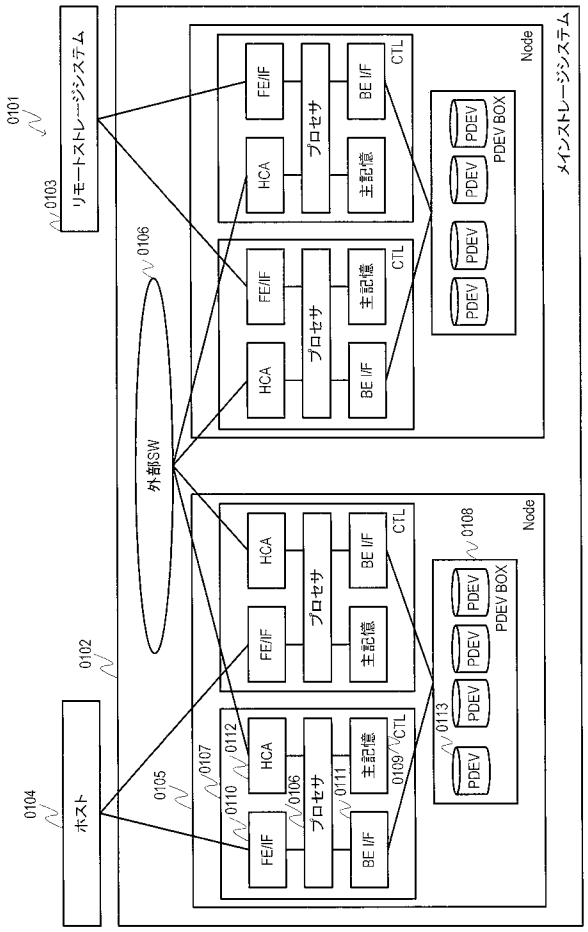
（Ｃ）

担当ＣＴＬ０１０７は、ストレージ冗長化のＬＤＥＶに対するライトである否かを判定する（１６０７）。ストレージ冗長化のＬＤＥＶに対するライトである場合（１６０７：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１６０８に進む（Ｄ）。一方、ストレージ冗長化のＬＤＥＶに対するライトでない場合（１６０７：Ｎｏ）、担当ＣＴＬ０１０７は、

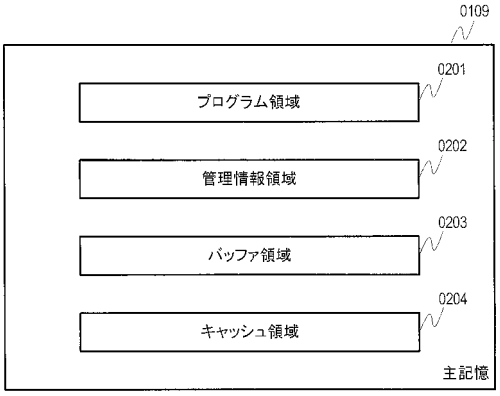
ステップ１６０９に進む（Ｅ）。
【０１０６】
（Ｄ）
ストレージ冗長化のＬＤＥＶに対するライトである場合（１６０７：Ｙｅｓ）、担当ＣＴＬ０１０７は、リモートストレージシステム０１０３に接続されているＣＴＬ０１０７より、当該Ｉ／Ｏデータを転送する（１６０８）。
【０１０７】
（Ｅ）
ストレージ冗長化のＬＤＥＶに対するライトでない場合（１６０７：Ｎｏ）、該当ＣＴＬは０１０７、ホスト０１０４にライトの完了を応答する（１６０９）。担当ＣＴＬ０１０７は、担当ＣＴＬ０１０７で固有の処理を実行する（１６１０）。例えば、スナップショットに関する処理であり、ＲＡＩＤグループを構成している場合、担当ＣＴＬ０１０７は、パリティの生成を実施する。
【０１０８】
担当ＣＴＬ０１０７は、前述の処理でデータを複写したバッファセグメントのほかに、確保したバッファセグメントが存在するか否かを判定する（１６１１）。他に確保したバッファセグメントが存在する場合（１６１１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１６１２に進む（Ｆ）。一方、他に確保したバッファセグメントが存在しない場合（１６１１：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１６１３に進む（Ｇ）。
【０１０９】
（Ｆ）
他に確保したバッファセグメントが存在する場合（１６１１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ホスト０１０４から転送されたＩ／Ｏデータを当該バッファセグメントに格納し（１６１２）し、担当ＣＴＬ０１０７は、ステップ１６１３に進む（Ｅ）。
【０１１０】
（Ｇ）
担当ＣＴＬ０１０７は、ライト先のＰＤＥＶにホスト０１０４から転送されたＩ／Ｏデータを転送し（１６１３）、処理を終了する。
【０１１１】
図１７は、図１１で説明したフローから呼び出される、セグメント解放処理１１０９の流れの一例を示したフロー図である。担当ＣＴＬ０１０７は、当該Ｉ／Ｏのアクセスパタンがシーケンシャルであるか否かを判定する（１７０１）。当該Ｉ／Ｏのアクセスパタンがシーケンシャルであった場合（１７０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、確保した全てのキャッシュセグメント、及びバッファセグメントを解放し（１７０２）、処理を終了する。
【０１１２】
このように、アクセスパタンがシーケンシャルである場合、ライト処理１１０８の終了（最終記憶媒体へのライトデータの格納）に応答して、確保した全てのキャッシュセグメントが解放される。シーケンシャルアクセスされるデータは、再度アクセスされる可能性が低く、効率的にキャッシュ領域を利用できる。
【０１１３】
一方、当該Ｉ／Ｏのアクセスパタンがシーケンシャルでなかった場合（１７０２：Ｎｏ）、担当ＣＴＬ０１０７は、正面を除く確保した副面のキャッシュセグメント、及び確保した全てのバッファセグメントを解放し（１７０３）、処理を終了する。これにより、キャッシュヒット率を高めることができる。
【０１１４】
上述のように、ストレージシステムは、ホストからライトデータを受信するコントローラと、ライトデータを処理するコントローラとに基づいて、ライトデータを格納するキャッシュサブ領域を確保する少なくとも二つのコントローラを指定する。ストレージシステムは、さらに、各コントローラが属するストレージノードの、前記ライトデータを格納す

る記憶デバイスへの接続の有無に基づいて、前記二つのコントローラを指定する。
【０１１５】
具体的には、上記例において、キャッシュセグメントを確保するＣＴＬとして指定されるＣＴＬの候補は、リモートストレージシステムにライトデータを転送するＣＴＬ（存在する場合）、ライトデータを処理する担当ＣＴＬ（存在する場合）、ライトデータをホストから受信するＣＴＬ及びＰＤＥＶにライトデータを格納するＣＴＬで構成されている。
【０１１６】
これらのＣＴＬの候補の一部、例えば、リモートストレージシステムにライトデータを転送するＣＴＬ（存在する場合）及びライトデータを処理する担当ＣＴＬ（存在する場合）の双方または一方は、省略されてもよく、ライトデータをホストから受信するＣＴＬ及びＰＤＥＶにライトデータを格納するＣＴＬの双方または一方が省略されてもよい。ＣＴＬの候補が一つであり、キャッシュセグメントの確保のために常に指定されてもよい。上記例は最大二つのＣＴＬを指定するが、他の例は、最大３以上のＣＴＬを指定してもよい。
【０１１７】
指定されるＣＴＬの候補の優先度は、設計により決定してもよい。上記例は、リモートストレージシステムにライトデータを転送するＣＴＬ（存在する場合）及びライトデータを処理する担当ＣＴＬ（存在する場合）を、ライトデータをホストから受信するＣＴＬ及びＰＤＥＶにライトデータを格納するＣＴＬよりも優先して選択している。他の例は、これと異なる優先度が候補に与えられていてもよい。
【０１１８】
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を分かりやすく説明するために詳細に説明したものであり、必ずしも説明したすべての構成を備えるものに限定されるものではない。また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。
【０１１９】
また、上記の各構成・機能・処理部等は、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り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、実行することにより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、メモリや、ハードディスク、ＳＳＤ（Ｓｏｌｉｄ Ｓｔａｔｅ Ｄｒｉｖｅ）等の記録装置、または、ＩＣカード、ＳＤカード等の記録媒体に置くことができる。また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしもすべての制御線や情報線を示しているとは限らない。実際には殆どすべ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】
【０１２０】
０１０１　情報システム、０１０２　メインストレージシステム、０１０３　リモートストレージシステム、０１０４　ホスト、０１０５　Ｎｏｄｅ、０１０６　外部スイッチ（ＳＷ）、０１０６　プロセサ、０１０７　コントローラ（ＣＴＬ）、０１０８　ＰＤＥＶ　ＢＯＸ、０１０９　主記憶、０１１０　ＦＥ　Ｉ／Ｆ（ＦｒｏｎｔＥｎｄ　Ｉｎｔｅｒｆａｃｅ）、０１１１　ＢＥ　Ｉ／Ｆ（ＢａｃｋＥｎｄ　Ｉｎｔｅｒｆａｃｅ）、０１１２　ＨＣＡ（Ｈｏｓｔ　Ｃｈａｎｎｅｌ Ａｄａｐｔｅｒ）、０１１３　ＰＤＥＶ、０２０１　プログラム領、０２０２　管理情報領域、０２０３　バッファ領域、０２０４　キャッシュ領域、０３０１　ＣＴＬ間振り分けプログラム、０３０２　Ｉ／Ｏプログラム、０３０３　Ｉ／Ｏパタン判定プログラム、０３０４　セグメント確保プログラム、０３０５　セグメント解放プログラム、０３０６　フリーセグメント確保プログラム、０３０７　ストレージコントローラ間データ転送プログラム

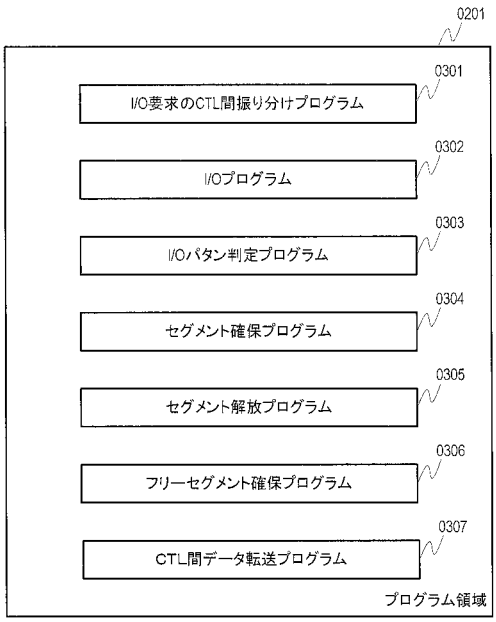
【図１】


リモートストレージシステム
0101
0103
ホスト
0104
0102
外部SW
0106
0105
メインストレージシステム
Node
CTL
FE/IF
HCA
プロセサ
BE I/F
主記憶
PDEV BOX
PDEV
0108
0107
0112
0110
0111
0109
0113

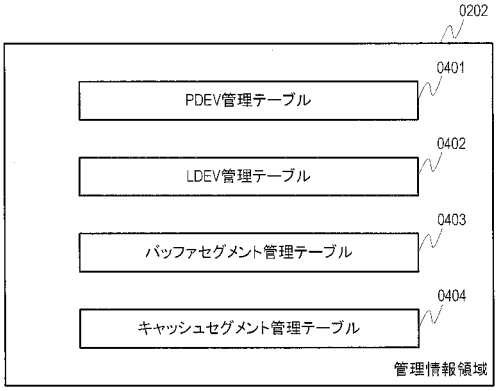
【図２】


0109
0201
プログラム領域
0202
管理情報領域
0203
バッファ領域
0204
キャッシュ領域
主記憶

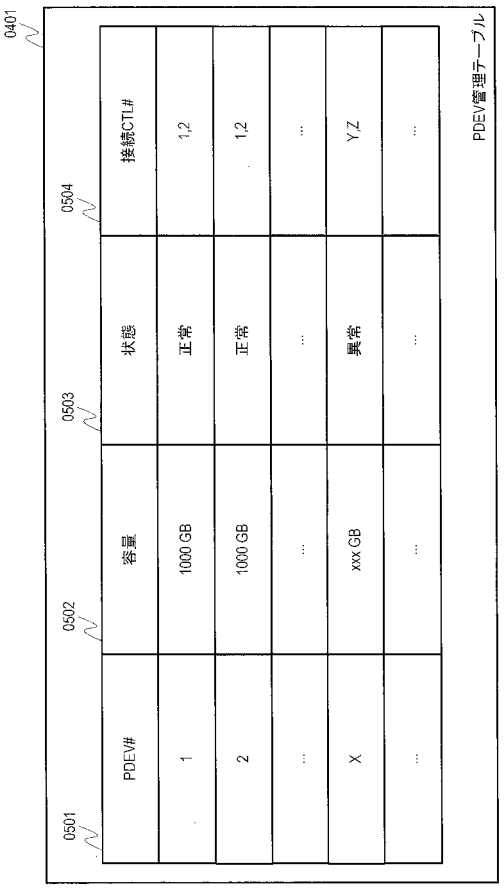
【図３】


0201
0301
I/O要求のCTL間振り分けプログラム
0302
I/Oプログラム
0303
I/Oパタン判定プログラム
0304
セグメント確保プログラム
0305
セグメント解放プログラム
0306
フリーセグメント確保プログラム
0307
CTL間データ転送プログラム
プログラム領域

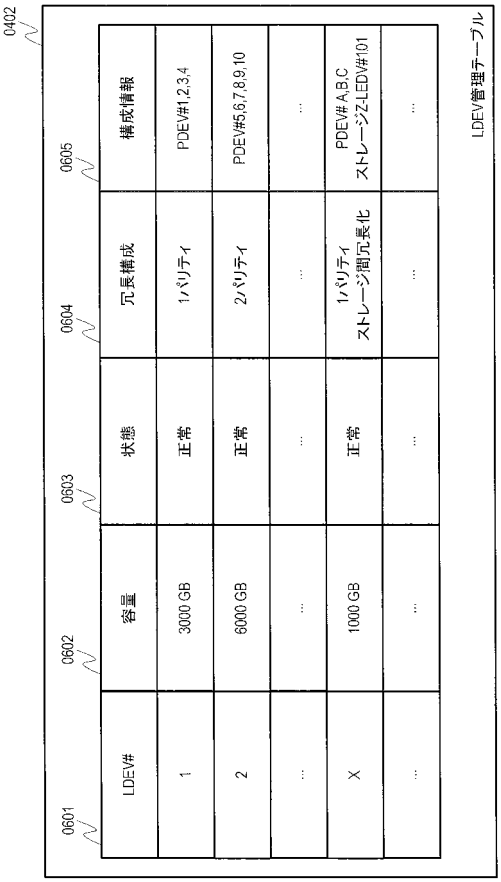
【図４】


0202
0401
PDEV管理テーブル
0402
LDEV管理テーブル
0403
バッファセグメント管理テーブル
0404
キャッシュセグメント管理テーブル
管理情報領域

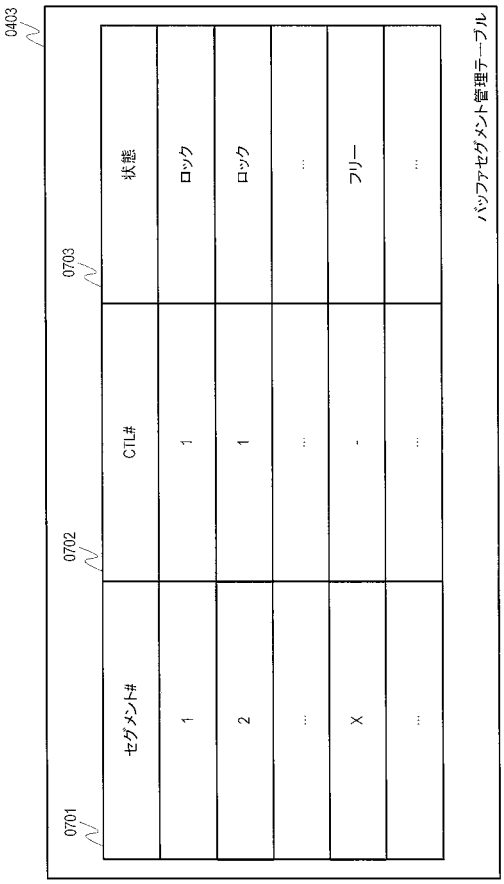
【図５】


0401
PDEV管理テーブル
0501
0502
0503
0504
PDEV#
容量
状態
接続CTL#
1
1000 GB
正常
1,2
2
1000 GB
正常
1,2
X
xxx GB
異常
Y,Z

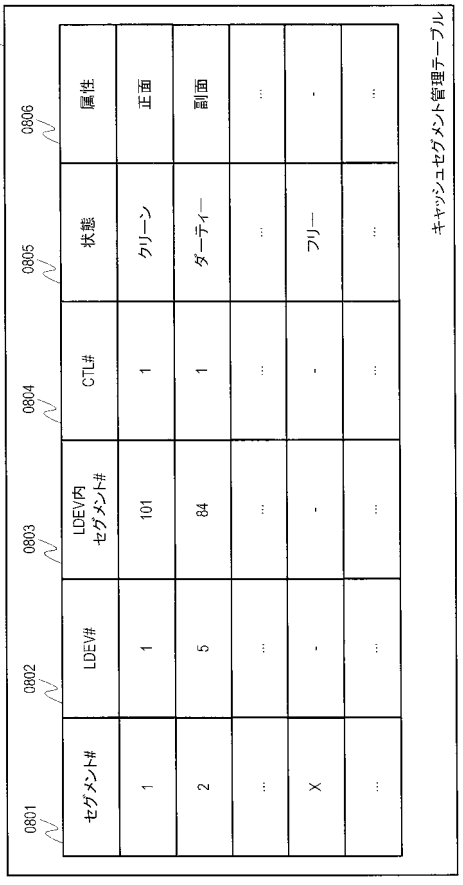
【図６】


0402
LDEV管理テーブル
0601
0602
0603
0604
0605
LDEV#
容量
状態
冗長構成
構成情報
1
3000 GB
正常
1パリティ
PDEV#1,2,3,4
2
6000 GB
正常
2パリティ
PDEV#5,6,7,8,9,10
X
1000 GB
正常
1パリティ
ストレージZ-LEDV#101
PDEV#A,B,C

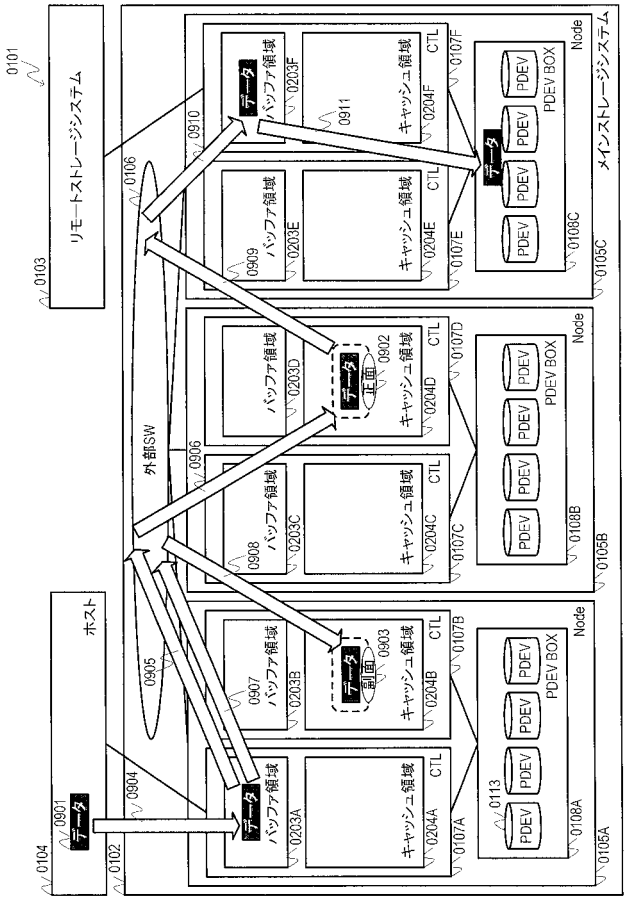
【図７】


0403
バッファセグメント管理テーブル
0701
0702
0703
セグメント#
CTL#
状態
1
1
ロック
2
1
ロック
X
-
フリー

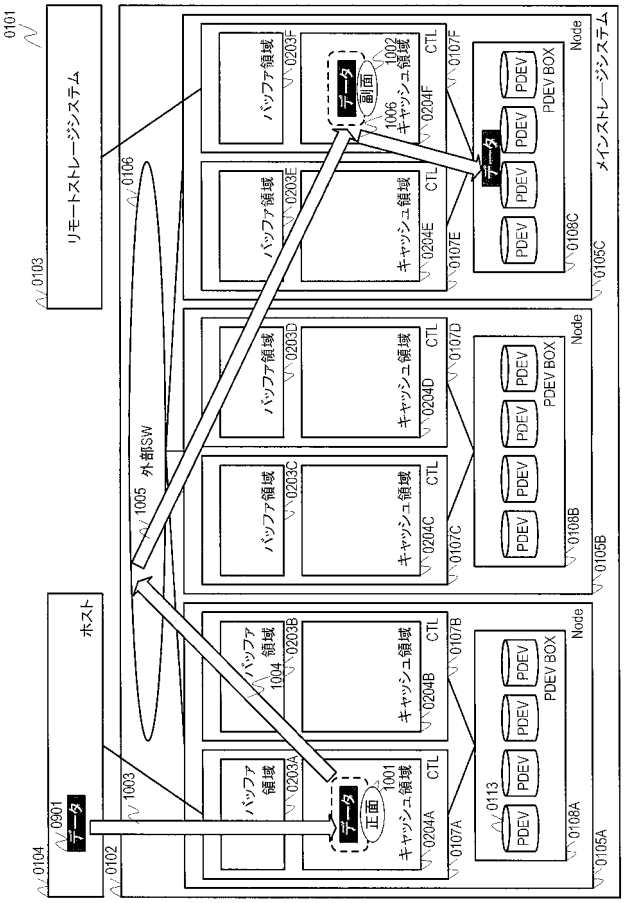
【図８】


0404
キャッシュセグメント管理テーブル
0801
0802
0803
0804
0805
0806
セグメント#
LDEV#
LDEV内セグメント#
CTL#
状態
属性
1
1
101
1
クリーン
正面
2
5
84
1
ダーティー
副面
X
-
-
-
フリー
-

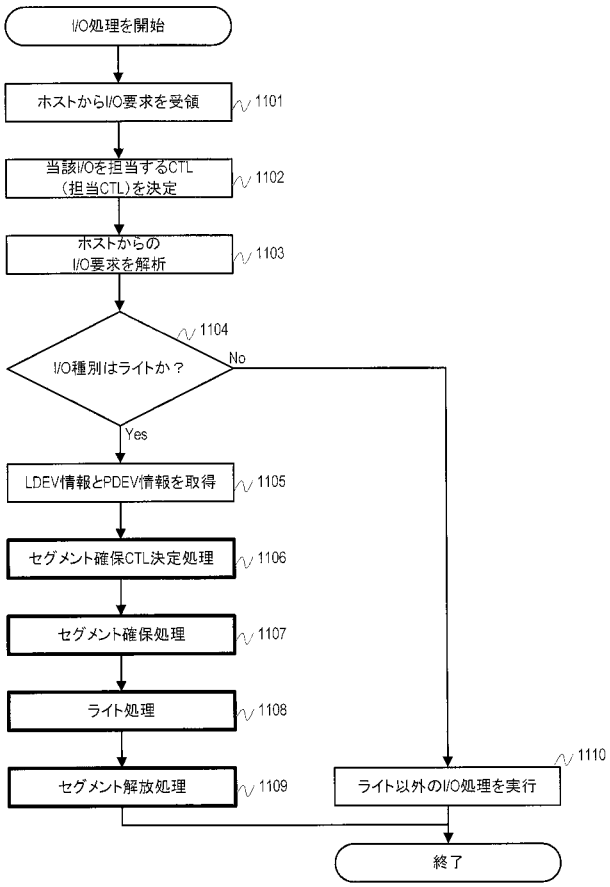
【図９】



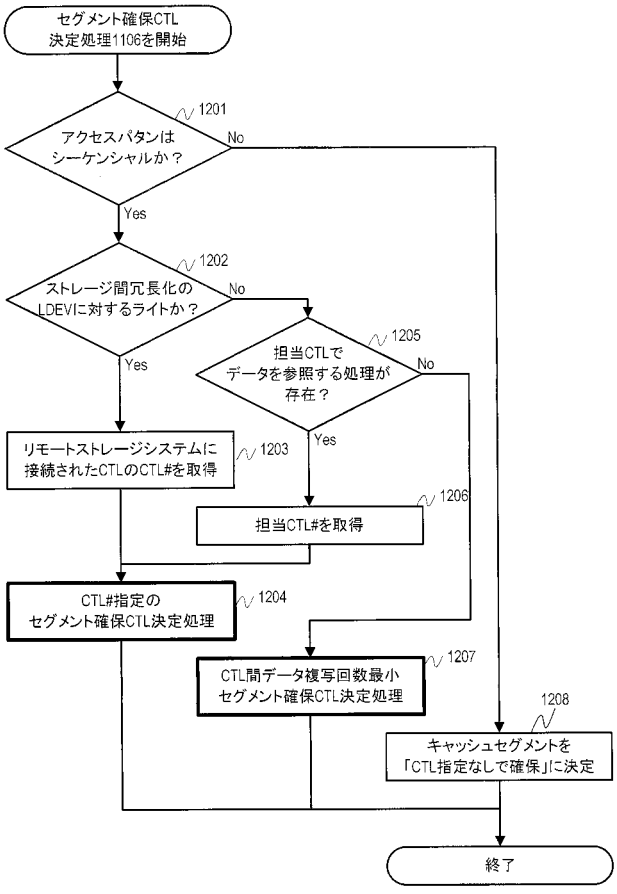
【図１０】



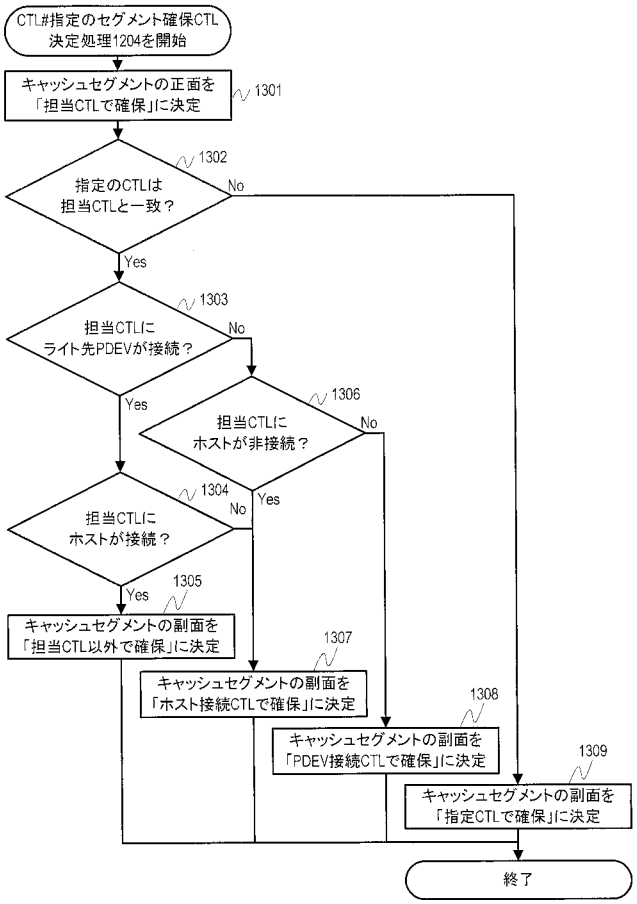
【図１１】



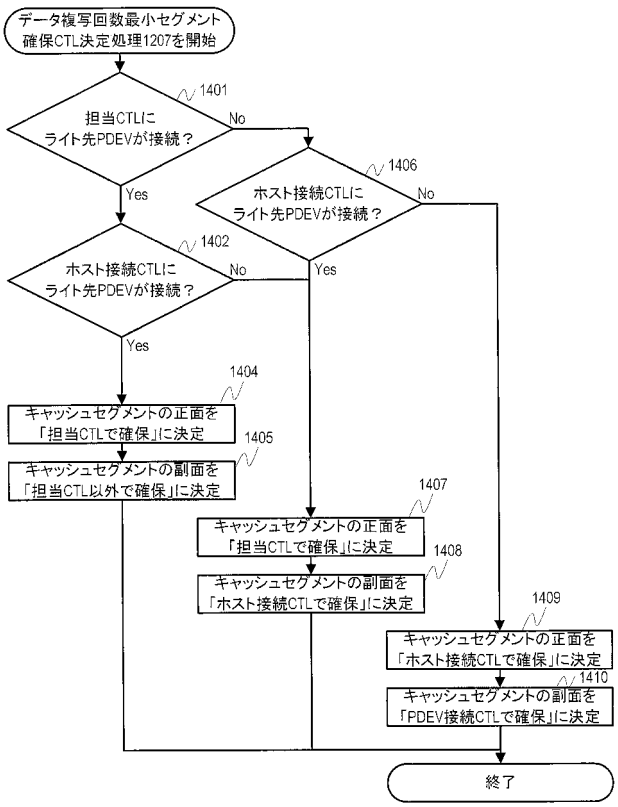
【図１２】



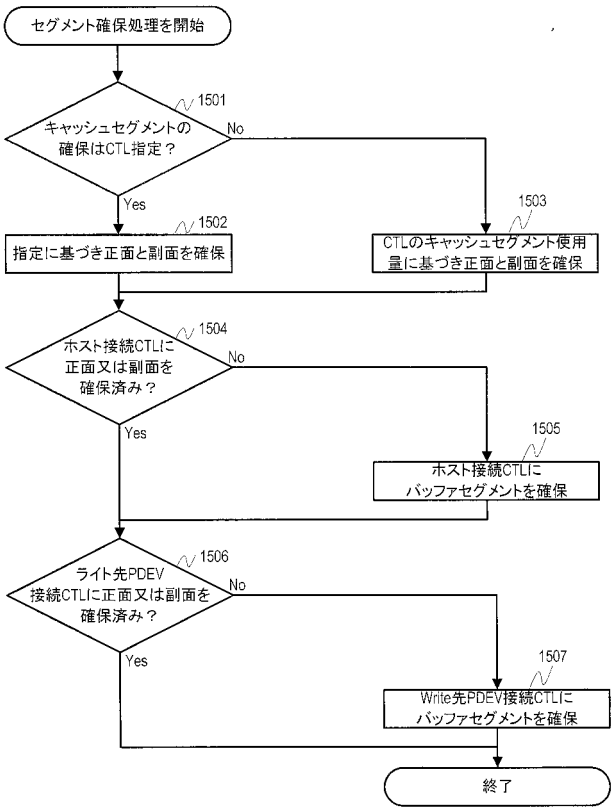
【図１３】



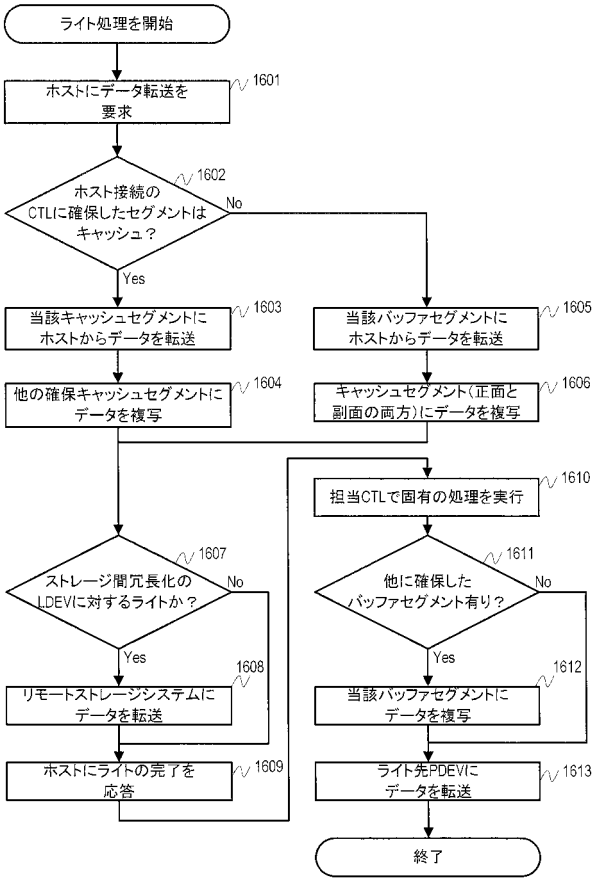
【図１４】



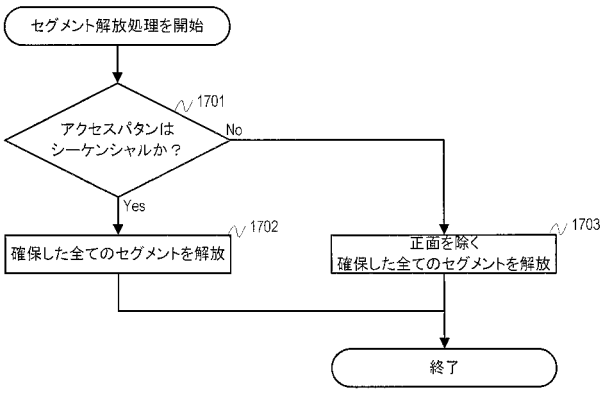
【図１５】



【図１６】



【図１７】



【手続補正書】
【提出日】令和2年1月24日(2020.1.24)
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】００１９
【補正方法】変更
【補正の内容】
【００１９】
以下の説明では、「ＰＤＥＶ」は、物理的な記憶デバイスを意味し、典型的には、不揮発性の記憶デバイス（例えば補助記憶デバイス）でよい。ＰＤＥＶは、例えば、ＨＤＤ（Ｈａｒｄ　Ｄｉｓｋ　Ｄｒｉｖｅ）又はＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）でよい。以下の説明では、「ＲＡＩＤ」は、Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙ　ｏｆ　Ｉｎｄｅｐｅｎｄｅｎｔ　（ｏｒ　Ｉｎｅｘｐｅｎｓｉｖｅ）　Ｄｉｓｋｓの略である。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】００２７
【補正方法】変更
【補正の内容】
【００２７】
メインストレージシステム０１０２は、１つ以上のＮｏｄｅ０１０５（ストレージノードとも呼ぶ）、及び１つ以上の外部ＳＷ（ＳＷ：ＳＷｉｔｃｈ）０１１６から構成されている。図１では、Ｎｏｄｅ０１０５は、外部ＳＷ０１０６を介して接続されているが、Ｎｏｄｅ０１０５間で直接接続されていてもよい。
【手続補正３】
【補正対象書類名】明細書

【補正対象項目名】００２９
【補正方法】変更
【補正の内容】
【００２９】
図１では、ＣＴＬ０１０７は、直接ＰＤＥＶ　ＢＯＸ０１０８に接続されているが、Ｓｗｉｔｃｈなどを介して接続されていてもよい。また、ＰＤＥＶ　ＢＯＸ０１０８を介さず、ＣＴＬ０１０７とＰＤＥＶ０１１３は直接接続されていてもよい。さらに、同一Ｎｏｄｅ０１０５内のＣＴＬ０１０７間は、外部ＳＷ０１１６を介して接続されているが、直接接続されていてもよい。
【手続補正４】
【補正対象書類名】明細書
【補正対象項目名】００３１
【補正方法】変更
【補正の内容】
【００３１】
プロセサ０１０６は、ＣＴＬ０１０７全体を制御し、主記憶０１０９に格納されたマイクロプログラムに基づき動作する。ＦＥ　Ｉ／Ｆ０１１０は、プロセサ０１０６により制御され、ホスト０１０４、及びリモートストレージシステム０１０３とＩ／Ｏ要求やＩ／Ｏデータの送受信などを実施する。ＢＥ　Ｉ／Ｆ０１１１は、プロセサ０１０６により制御され、ＰＤＥＶ　ＢＯＸ０１０８を介し、ＰＤＥＶ０１１３とＩ／Ｏデータの送受信などを実施する。ＨＣＡ０１１２は、プロセサ０１０６により制御され、外部ＳＷ０１１６を介して他のＣＴＬ０１０７と制御情報やＩ／Ｏデータの送受信などを実施する。
【手続補正５】
【補正対象書類名】明細書
【補正対象項目名】００４１
【補正方法】変更
【補正の内容】
【００４１】
ＰＤＥＶ管理テーブル０４０１は、メインストレージシステム１０２内の全てのＰＤＥＶ０１１３の状態や、ＣＴＬ０１０７との対応関係を示す。ＬＤＥＶ管理テーブル０４０２は、メインストレージシステム１０２内の全てのＰＤＥＶ０１１３とＬＤＥＶの対応関係を示す。バッファセグメント管理テーブル０４０３は、メインストレージシステム１０２内の全てのバッファ領域０２０３を管理するのに用いられる。キャッシュセグメント管理テーブル０４０４は、メインストレージシステム１０２内の全てのキャッシュ領域０２０４を管理するのに用いられる。
【手続補正６】
【補正対象書類名】明細書
【補正対象項目名】００４２
【補正方法】変更
【補正の内容】
【００４２】
図５は、ＰＤＥＶ管理テーブル０４０１の構成例を示す図である。ＰＤＥＶ管理テーブル０４０１は、ＰＤＥＶ＃０５０１、容量０５０２、状態０５０３、及び接続ＣＴＬ＃０５０４のエントリを有する。ＰＤＥＶ＃０５０１は、ＰＤＥＶ０１１３の識別子である。容量０５０２は、ＰＤＥＶ０１１３のデータを格納可能な容量を示す。状態０５０３は、ＰＤＥＶ０１１３が正常に動作中か否か（故障していないか）の状態を示す。接続ＣＴＬ＃０５０４は、ＰＤＥＶ０１１３に接続され、アクセスするＣＴＬ０１０７を示す。
【手続補正７】
【補正対象書類名】明細書
【補正対象項目名】００４４

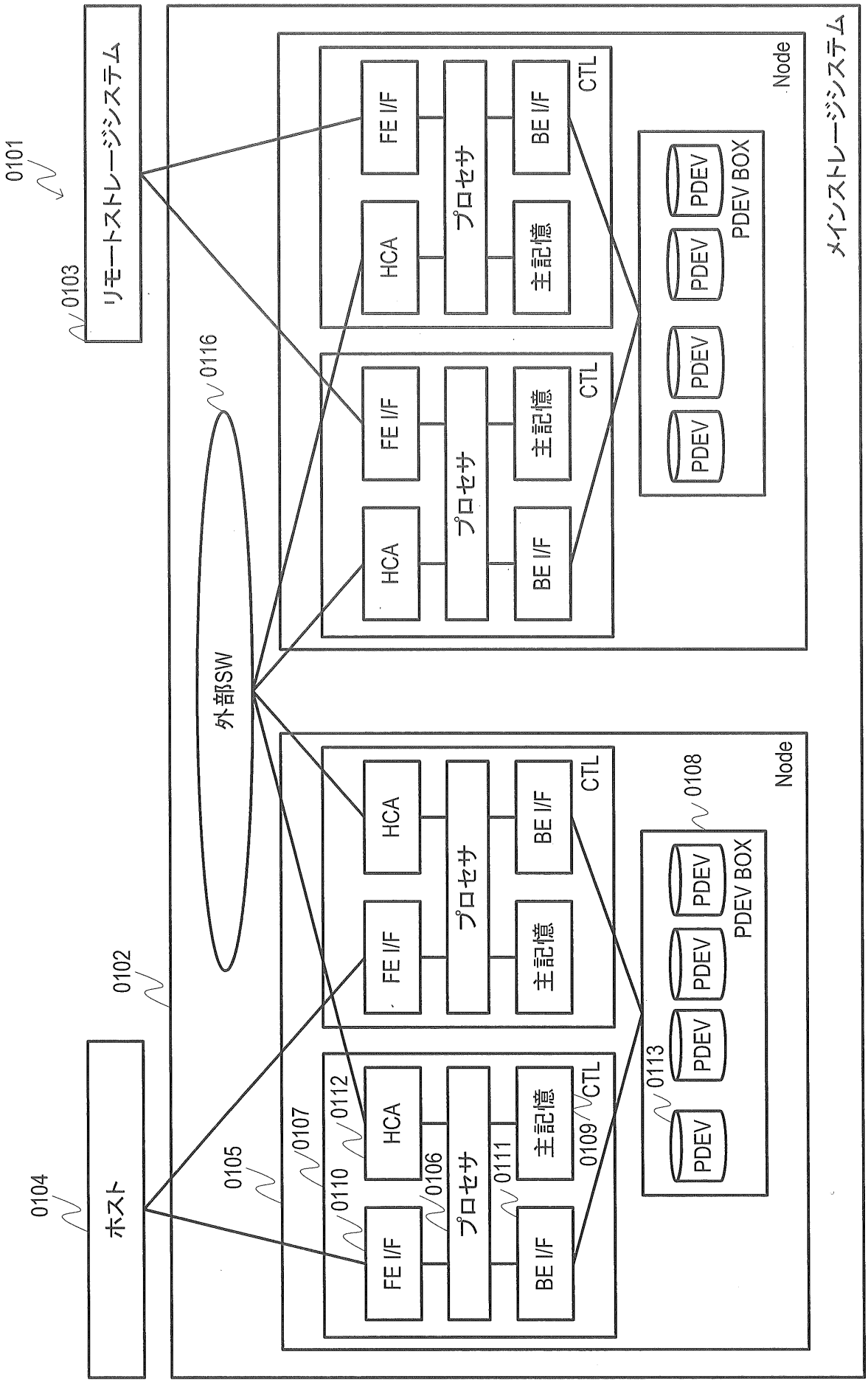
【補正方法】変更
【補正の内容】
【００４４】
ＬＤＥＶ＃０６０１は、ＬＤＥＶの識別子である。容量０６０２は、ＬＤＥＶにデータを格納可能な容量を示す。状態０５０３は、ＬＤＥＶへ正常にＩ／Ｏ可能か否かを示す。冗長構成０６０４は、ＲＡＩＤによる冗長化、またはリモートストレージシステム０１０３とのストレージシステム冗長化の状態を示す。構成情報０６０５は、ＬＤＥＶに属するＰＤＥＶ０１１３、及びストレージシステム冗長化の対応リモートストレージシステム０１０３を示す。
【手続補正８】
【補正対象書類名】明細書
【補正対象項目名】００５３
【補正方法】変更
【補正の内容】
【００５３】
図１１は、ストレージシステム（メイン）０１０２がホスト０１０４から受領したＩ／Ｏ要求を処理する流れの一例を示したフロー図である。ＣＴＬ０１０７は、ホスト０１０４からのＩ／Ｏ要求を受領する（１１０１）。ホスト０１０４からＩ／Ｏ要求を受領したＣＴＬ０１０７は、当該Ｉ／Ｏを担当するＣＴＬ０１０７を決定し、その担当ＣＴＬ０１０７が、処理を継続する（１１０２）。
【手続補正９】
【補正対象書類名】明細書
【補正対象項目名】００５６
【補正方法】変更
【補正の内容】
【００５６】
（Ａ）
Ｉ／Ｏ要求種別がライト要求の場合（１１０４：Ｙｅｓ）、担当ＣＴＬ０１０７は、ＰＤＥＶ管理テーブル０４０１とＬＤＥＶ管理テーブル０４０２を参照し、当該Ｉ／ＯのＬＤＥＶに関する情報、及び当該ＬＤＥＶを構成するＰＤＥＶの情報を取得する（１１０５）。担当ＣＴＬ０１０７は、セグメント確保ＣＴＬ決定処理を呼び出し、ライト処理に必要なキャッシュセグメントを確保するＣＴＬ０１０７を決定する（１１０６）。なお、セグメント確保ＣＴＬ決定処理１１０６の詳細は、後述する。
【手続補正１０】
【補正対象書類名】明細書
【補正対象項目名】００６０
【補正方法】変更
【補正の内容】
【００６０】
図１２は、図１１で説明したフローから呼び出される、セグメント確保ＣＴＬ決定処理１１０６の流れの一例を示したフロー図である。担当ＣＴＬ０１０７は、ホスト０１０４からのＩ／Ｏ要求の解析１１０３で明らかにしたアクセスパタンを参照し、アクセスパタンはシーケンシャルであるか否かを判定する（１２０１）。アクセスパタンがシーケンシャルである場合（１２０１：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１２０２に進む（Ａ）。一方、アクセスパタンがシーケンシャルでない場合（１２０１：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１２０８に進む（Ｇ）。
【手続補正１１】
【補正対象書類名】明細書
【補正対象項目名】００６３
【補正方法】変更

【補正の内容】
【００６３】
（Ｃ）
ストレージ間冗長化のＬＤＥＶに対するライトでない場合（１２０２：Ｎｏ）、担当ＣＴＬ０１０７は、当該ＬＤＥＶへのライトについて、担当ＣＴＬ０１０７でホスト０１０４からライトされたデータを参照する処理（例えば、スナップショットなどライトに起因する処理）が存在するか否かを判定する（１２０５）。担当ＣＴＬ０１０７でデータを参照する処理が存在する場合（１２０５：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１２０６へ進む（Ｄ）。一方、担当ＣＴＬ０１０７でデータを参照する処理が存在しない場合（１２０５：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１２０７へ進む（Ｆ）。
【手続補正１２】
【補正対象書類名】明細書
【補正対象項目名】００８７
【補正方法】変更
【補正の内容】
【００８７】
（Ｂ）
ホスト接続ＣＴＬにライト先のＰＤＥＶが接続されている場合（１４０２：Ｙｅｓ）、担当ＣＴＬは、担当ＣＴＬ０１０７にホスト０１０４が接続されているか否かを判定する（１４０３）。担当ＣＴＬ０１０７にホスト０１０４が接続されている場合（１４０３：Ｙｅｓ）、担当ＣＴＬ０１０７は、ステップ１４０４に進む（Ｃ）。一方、担当ＣＴＬ０１０７にホスト０１０４が接続されていない場合（１４０３：Ｎｏ）、担当ＣＴＬ０１０７は、ステップ１４０７に進む（Ｅ）。
【手続補正１３】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
ネットワークを介して通信する複数のストレージノードを含むストレージシステムであって、
前記複数のストレージノードのそれぞれは、１以上のコントローラを含み、
前記コントローラにおける少なくとも１つのコントローラは、
ホストからライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとに基づいて、前記ライトデータを格納するキャッシュサブ領域を確保する少なくとも二つのコントローラを指定し、
指定したコントローラにおいて、キャッシュサブ領域を確保する
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、
前記少なくとも１つのコントローラは、さらに、各コントローラが属するストレージノードの、前記ライトデータを格納する記憶デバイスへの接続の有無に基づいて、前記二つのコントローラを指定する
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、
前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとは、別のコントローラであり、
前記少なくとも１つのコントローラは、前記ホストから前記ライトデータを受信するコ

ントローラと、前記ライトデータを処理するコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定する
ストレージシステム。
【請求項４】
請求項３に記載のストレージシステムであって、
前記指定した二つのコントローラのうち、前記ライトデータを格納する記憶デバイスに接続されたコントローラは、前記ライトデータを格納する記憶デバイスに接続されたコントローラの前記キャッシュサブ領域に記憶された前記ライトデータを、前記記憶デバイスに格納する
ストレージシステム。
【請求項５】
請求項４に記載のストレージシステムであって、
前記指定した二つのコントローラのいずれもが、前記記憶デバイスに接続されていない場合、前記指定した二つのコントローラのうちいずれかが、前記記憶デバイスに接続された他のコントローラのバッファ領域に前記ライトデータを転送し、
前記他のコントローラが、前記バッファ領域に格納された前記ライトデータを、前記記憶デバイスに格納する
ストレージシステム。
【請求項６】
請求項２に記載のストレージシステムであって、
前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとは、同じコントローラであり、
前記少なくとも１つのコントローラは、前記同じコントローラと、他の１つのコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定し、
前記指定した二つのコントローラのうち少なくとも１つは、前記記憶デバイスに接続されている
ストレージシステム。
【請求項７】
請求項２に記載のストレージシステムであって、
前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを格納する記憶デバイスに接続された他のコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定する
ストレージシステム。
【請求項８】
請求項２に記載のストレージシステムであって、
前記少なくとも１つのコントローラは、前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを格納する記憶デバイスに接続された他のコントローラとを、前記キャッシュサブ領域を確保するコントローラに指定する
ストレージシステム。
【請求項９】
請求項２に記載のストレージシステムであって、
前記記憶デバイスは、前記ストレージノードの内に設けられた記憶媒体、またはリモートストレージシステムである
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であって、
前記少なくとも１つのコントローラは、
前記ホストからのライト要求のアクセスパタンがシーケンシャルである場合に、前記ホストから前記ライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとに基づいて、前記コントローラの指定を行い、

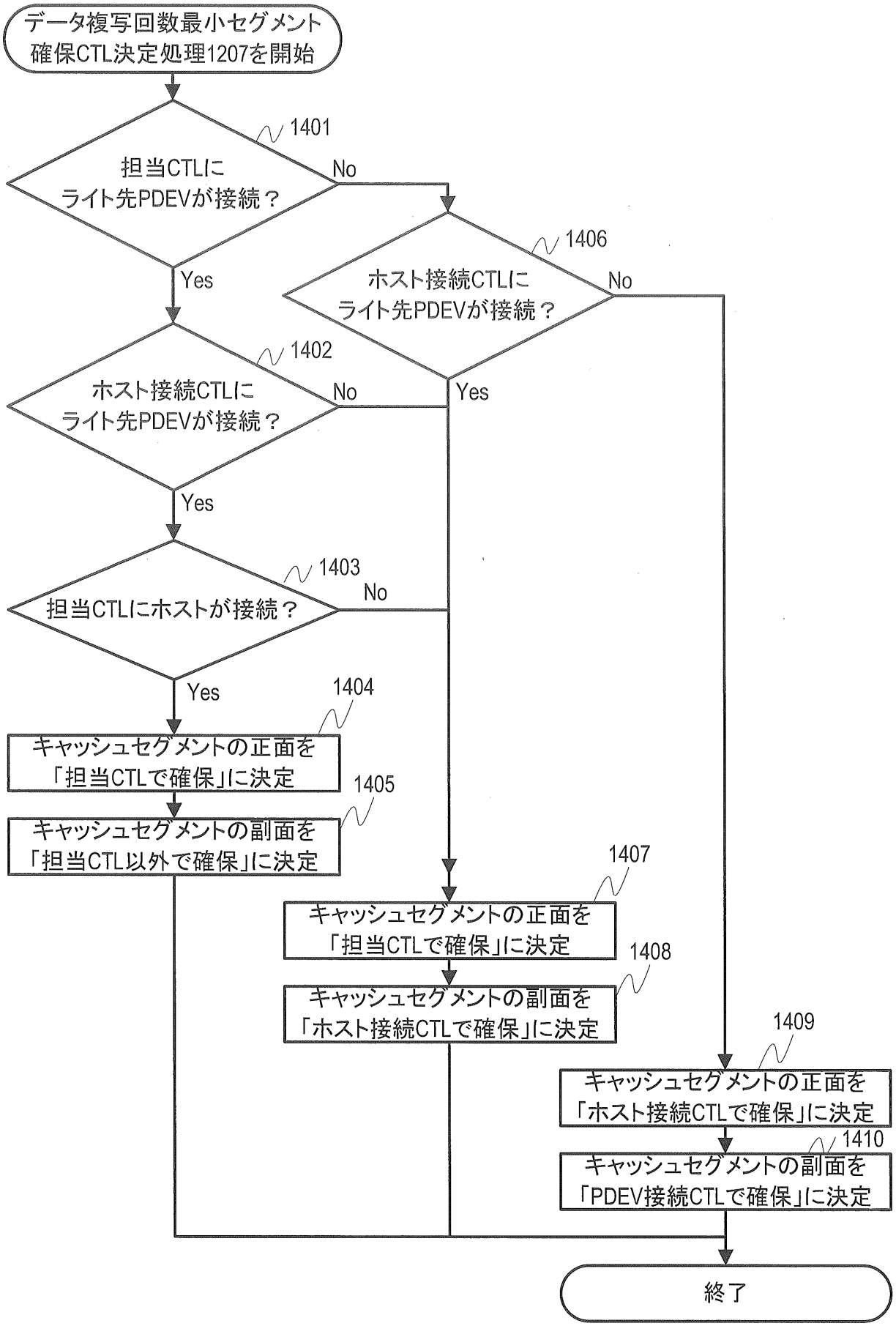
前記ホストからのライト要求のアクセスパタンがランダムである場合に、キャッシュ利用効率に基づいて、前記コントローラの指定を行う
ストレージシステム。
【請求項１１】
ネットワークを介して通信する複数のストレージノードを含むストレージシステムの制御方法であって、
前記複数のストレージノードのそれぞれは、１以上のコントローラを含み、
前記制御方法は、
ホストからライトデータを受信するコントローラと、前記ライトデータを処理するコントローラとに基づいて、前記ライトデータを格納するキャッシュサブ領域を確保する少なくとも二つのコントローラを指定し、
指定したコントローラにおいて、キャッシュサブ領域を確保する
制御方法。
【手続補正１４】
【補正対象書類名】図面
【補正対象項目名】図１
【補正方法】変更
【補正の内容】

【図１】


0101
リモートストレージシステム
0103
0116
外部SW
0102
メインストレージシステム
Node
CTL
FE I/F
HCA
プロセサ
主記憶
BE I/F
PDEV
PDEV BOX
0104
ホスト
0105
0107
0110
0112
0106
0111
0109
0113
0108

【手続補正１５】
【補正対象書類名】図面
【補正対象項目名】図１４
【補正方法】変更
【補正の内容】

【図１４】


データ複写回数最小セグメント確保CTL決定処理1207を開始
1401
担当CTLにライト先PDEVが接続？
No
Yes
1406
ホスト接続CTLにライト先PDEVが接続？
No
Yes
1402
ホスト接続CTLにライト先PDEVが接続？
No
Yes
1403
担当CTLにホストが接続？
No
Yes
1404
キャッシュセグメントの正面を「担当CTLで確保」に決定
1405
キャッシュセグメントの副面を「担当CTL以外で確保」に決定
1407
キャッシュセグメントの正面を「担当CTLで確保」に決定
1408
キャッシュセグメントの副面を「ホスト接続CTLで確保」に決定
1409
キャッシュセグメントの正面を「ホスト接続CTLで確保」に決定
1410
キャッシュセグメントの副面を「PDEV接続CTLで確保」に決定
終了

フロントページの続き

(72)発明者　井澤　信介
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

Ｆターム(参考) 5B205 MM12