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

特許第4147198号
(P4147198)

(45)発行日　平成20年9月10日(2008.9.10)　(24)登録日　平成20年6月27日(2008.6.27)

(51)Int.Cl.		F I		
G06F 12/08	(2006.01)	G06F	12/08	557
G06F 3/06	(2006.01)	G06F	12/08	551Z
G06F 13/10	(2006.01)	G06F	12/08	511Z
		G06F	12/08	543B
		G06F	3/06	302A

請求項の数 20　(全 36 頁)　最終頁に続く

(21)出願番号　特願2004-84229(P2004-84229)
(22)出願日　平成16年3月23日(2004.3.23)
(65)公開番号　特開2005-275525(P2005-275525A)
(43)公開日　平成17年10月6日(2005.10.6)
審査請求日　平成18年10月27日(2006.10.27)

(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　110000350
ポレール特許業務法人
(74)代理人　100068504
弁理士　小川　勝男
(74)代理人　100086656
弁理士　田中　恭助
(74)代理人　100094352
弁理士　佐々木　孝
(72)発明者　渡辺　恭男
神奈川県川崎市麻生区王禅寺１０９９番地
株式会社日立製作所　システム開発研究所内

最終頁に続く

(54)【発明の名称】ストレージシステム

(57)【特許請求の範囲】
【請求項１】
クラスタ構成のストレージシステムにおいて、
ホストから入出力要求に従い、読み書きされるデータを格納するキャッシュメモリ、及び該キャッシュメモリに格納されるデータを保持するデバイスを有する複数のストレージ制御部と、
読み書きされるデータを扱う論理デバイス及びキャッシュメモリを持つ外部ストレージを該ストレージ制御部に接続する手段と、
該複数のストレージ制御部のキャッシュメモリの使用状況を監視して把握する手段と、
該把握手段によって得られるキャッシュメモリの使用状況に関する情報を参照して、キャッシュメモリの使用量を均等化させるように、いずれかの該ストレージ制御部を選択する手段と、を有し、
該選択手段により選択された該ストレージ制御部により該接続手段を介して該外部ストレージの論理デバイスを制御することを特徴とするストレージシステム。
【請求項２】
前記把握手段は、複数のストレージ制御部の該キャッシュメモリ毎にそのダーティデータ量（第1のダーティデータ量）を取得し、
前記選択手段は、得られた該ダーティデータ量に従い、ダーティデータ量のより少ないストレージ制御部を選択して、外部ストレージの論理デバイスを制御することを特徴とする請求項１記載のストレージシステム。

【請求項３】
更に、前記把握手段は、該外部ストレージの論理デバイスに格納されるデータであって、いずれかのストレージ制御部が有するキャッシュメモリに格納されているダーティデータの量（第２のダーティデータ量）を把握し、
前記選択手段は、該第1のダーティデータ量及び該第2のダーティデータ量に従って、該外部ストレージの論理デバイスを制御するためのストレージ制御部を選択することを特徴とする請求項２記載のストレージシステム。
【請求項４】
更に、該外部ストレージの論理デバイスに格納されるデータについて、他のストレージへ非同期のリモートコピーを行う手段と、
該ストレージ制御部の該キャッシュメモリに保持されていて、未だ該他のストレージへ送信されていないデータ（サイドファイルデータ）の量を把握する手段と、を有し、
前記選択手段は、該把握手段によって取得されたサイドファイルデータ量を参照して、該外部ストレージの論理デバイスを制御するためのストレージ制御部を選択することを特徴とする請求項２又は３記載のストレージシステム。
【請求項５】
前記把握手段により取得されたキャッシュメモリの使用状況、又はダーティデータ量、若しくはサイドファイルデータ量を管理者へ提示するために管理端末又は管理サーバへ転送する手段を有し、
前記選択手段は、該管理端末又は管理サーバから指定されたストレージ制御部を選択することを特徴とする請求項１乃至４のいずれか記載のストレージシステム。
【請求項６】
クラスタ構成のストレージシステムにおいて、
ホストから入出力要求に従い、読み書きされるデータを一時的に格納するキャッシュメモリ、及び該キャッシュメモリのデータを保持するデバイスを有する複数のストレージノードと、
該ホストから入出力要求に従い、読み書きされるデータを格納するキャッシュメモリ及びデバイスを有する外部ストレージと、該ストレージノードと接続するインターフェースと、
該複数のストレージノードの該キャッシュメモリの使用状況を監視して把握する手段と、
該把握手段によって得られるキャッシュメモリの使用状況に関する情報を参照して、キャッシュメモリの使用量を均等化させるように、あるストレージノードを選択する手段と、を有し、
該選択手段により選択された該ストレージノードにより該インターフェースを介して外部ストレージのデバイスを制御することを特徴とするストレージシステム。
【請求項７】
前記把握手段は、該キャッシュメモリ毎のダーティデータ量（第1のダーティデータ量）を取得し、
前記選択手段は、得られた該ダーティデータ量に従い、ダーティデータ量のより少ないストレージノードを選択して、外部ストレージのデバイスを制御することを特徴とする請求項６記載のストレージシステム。
【請求項８】
更に、前記複数のストレージノードを含む第２のストレージから第3のストレージに対して非同期のリモートコピーを行う手段と、
該ストレージノードの該キャッシュメモリに保持されていて、未だ該第３のストレージへ送信されていないデータ（サイドファイルデータ）の量を把握する手段と、を有し、
該選択手段は、該把握手段によって取得されたサイドファイルデータ量を参照して、ストレージノードを選択することを特徴とする請求項７記載のストレージシステム。
【請求項９】
前記把握手段により取得されたキャッシュメモリの使用状況、又はダーティデータ量、若

しくはサイドファイルデータ量を管理者へ提示するために管理端末又は管理サーバへ転送する手段を有し、
前記選択手段は、該管理端末又は管理サーバから指定されたストレージノードを選択することを特徴とする請求項６乃至８のいずれか記載のストレージシステム。
【請求項１０】
クラスタ構成のストレージシステムにおいて、
ホストとのインターフェースを有し、該ホストとの間で取り交わす入出力プロトコルをシステム内部のプロトコルに変換する第１のプロトコル変換部と、ホストとの間で読み書きされるデータを格納するキャッシュメモリ及びディスク装置を有する第1のストレージとのインターフェースを有し、該第1のストレージとの間で取り交わす入出力プロトコルをシステム内部のプロトコルに変換する第２のプロトコル変換部を含む複数のプロトコル変換部と、
該ホストと該第１のプロトコル変換部の間で読み書きされるデータを格納するキャッシュメモリ及びディスク装置を有し、かつ該キャッシュメモリと該ディスク装置へのアクセスを制御する制御部を有する複数のストレージ制御部と、
該複数のプロトコル変換部と該複数のストレージ制御部を管理する構成管理部と、
該構成管理部を介して接続されて該ストレージ制御部と通信を行う管理端末と、
該プロトコル変換部と該ストレージ制御部と該構成管理部とを接続する相互結合網と、
該第1のストレージのディスク装置及び該ストレージ制御部に保持されるディスク装置を該ストレージ制御部のディスク装置として該ホストへ提示する手段と、
該ホストから受け付けた入出力要求のアクセス対象が該ストレージ制御部のディスク装置もしくは該第1のストレージのディスク装置である場合、該入出力要求を該ストレージ制御部で処理する手段と、
該ストレージ制御部の該キャッシュメモリ上にあって、該ストレージ制御部のディスク装置若しくは該第1のストレージのディスク装置に未反映なライトデータの合計量である第1のダーティデータ量を含むキャッシュ使用量情報を取得する手段と、
該管理端末に対して該第1のダーティデータ量の情報を提示すると共に、該第１のストレージのディスク装置のための処理を行うストレージ制御部の選択指示を該管理端末から受け付ける手段と、
を有することを特徴とするストレージシステム。
【請求項１１】
該第１のストレージのディスク装置の入出力処理を行う該ストレージ制御部が該管理端末から指定されなかった場合、
該第１のストレージのディスク装置の入出力処理を行う該ストレージ制御部を、該第１のダーティデータ量情報を用いて決定することを特徴とする請求項１０記載のストレージシステム。
【請求項１２】
第１のストレージのディスク装置の入出力処理を行うストレージ制御部を、あるストレージ制御部（第１ストレージ制御部）から他のストレージ制御部（第２ストレージ制御部）に変更する変更手段を有することを特徴とする請求項１１記載のストレージシステム。
【請求項１３】
該第１ストレージ制御部のキャッシュメモリに保持した該第１のストレージの該ディスク装置のデータを検索し、第１のストレージのディスク装置へ未反映なデータは、第１のストレージのディスク装置へ書き込んで前記キャッシュメモリ領域を解放し、反映済みのデータについては前記キャッシュメモリ領域を解放することを特徴とする請求項１２記載のストレージシステム。
【請求項１４】
第1のプロトコル変換部が、第１のストレージの該ディスク装置に対応する該第１ストレージ制御部のディスク装置への前記ホストからの入出力要求に対して、該第１ストレージ制御部で変更処理が終わった部分への入出力要求は該第２ストレージ制御部へ行い、変更

処理が終わっていない部分への要求は該第１ストレージ制御部へ振り分けることを特徴とする請求項１３記載のストレージシステム。
【請求項１５】
前記第１ストレージ制御部の前記キャッシュメモリ上にあって、前記第1のストレージのディスク装置に未反映なライトデータの量である第２のダーティデータ量情報を取得する手段を有し、
該管理端末に、該第１のダーティデータ量情報および該第２のダーティデータ量情報を提示し、かつ前記変更処理の対象となる前記第１のストレージのディスク装置および変更処理による変更先である該第２ストレージ制御部の指定を該管理端末から受け付けることを特徴とする請求項１２記載のストレージシステム。
【請求項１６】
前記変更処理の対象となる前記第１のストレージのディスク装置および前記変更処理の変更先である該第２ストレージ制御部が該管理端末から指定されなかった場合、該第１のダーティデータ量および前記第２のダーティデータ量を用いて、変更処理の対象となる前記第１のストレージのディスク装置および変更先である該第２ストレージ制御部を決定する選択処理を行うことを特徴とする請求項１５記載のストレージシステム。
【請求項１７】
更に、第３のストレージシステムとのインターフェースを有し、該第３のストレージシステムとの間で取り交わす入出力プロトコルをシステム内部のプロトコルに変換する第３のプロトコル変換部と、
該第３のストレージシステムのディスク装置に前記第１のストレージのディスク装置の複製を作成し、前記第１のストレージのディスク装置に対するライトデータを該ストレージ制御部のキャッシュメモリに格納し、該ライトデータを該第３のストレージシステムのディスク装置に送信する手段と、
該ストレージ制御部のキャッシュメモリ上にあって、該第３のストレージシステムに未送信なライトデータの合計量である第１のサイドファイル量情報を取得する手段と、を有し、
前記管理端末に該第１のサイドファイル量情報を提示し、第１のストレージのディスク装置の処理を行う該ストレージ制御部の指定を該管理端末から受け付けることを特徴とする請求項１０記載のストレージシステム。
【請求項１８】
クラスタ構成のストレージシステムにおいてホストからの入出力要求を処理する方法であって、
データを格納するデバイス、及び該デバイスに格納されるデータを一時的に格納するキャッシュメモリとを有する複数のストレージ部分で、ホストからの入出力要求を処理するステップと、
データを格納するデバイス及びキャッシュメモリを有する外部ストレージの該デバイスを、あるストレージ部分によって制御するステップと、
該複数のストレージ部分におけるキャッシュメモリの使用状況を把握するステップと、
得られたキャッシュメモリの使用状況に関する情報を参照して、キャッシュメモリの使用量を均等化させるように、あるストレージ部分を選択するステップと、
選択された該ストレージ部分を用いて該外部ストレージに対する該ホストからの入出力要求を処理するステップと、
を有することを特徴とする入出力処理方法。
【請求項１９】
クラスタ構成のストレージシステムにおいて
ホストとのインターフェースを有し、該ホストとの間で取り交わす入出力プロトコルをシステム内部のプロトコルに変換する第１のプロトコル変換部と、ホストとの間で読み書きされるデータを格納するキャッシュメモリ及びディスク装置を有する外部ストレージとのインターフェースを有し、該外部ストレージとの間で取り交わす入出力プロトコルをシス

テム内部のプロトコルに変換する第２のプロトコル変換部を含む複数のプロトコル変換部と、
該ホストと該第１のプロトコル変換部の間で読み書きされるデータを格納するキャッシュメモリ及びディスク装置を有し、該キャッシュメモリ及び該ディスク装置へのアクセスを制御すると共に、いずれのストレージ制御部も該外部ストレージと交信可能である複数のストレージ制御部と、
該複数のプロトコル変換部と該複数のストレージ制御部を管理する構成管理部と、
該構成管理部を介して接続されて該ストレージ制御部と通信を行う管理端末と、
該プロトコル変換部と該ストレージ制御部と該構成管理部とを接続する相互結合網と、
該ストレージ制御部の該キャッシュメモリ内に在って、該ストレージ制御部のディスク装置及び該外部ストレージのディスク装置に未反映なデータの合計量であるダーティデータ量を含むキャッシュ使用量情報をメモリに保持して管理する手段と、
該管理端末からの指示に基づき、接続対象となる外部ストレージを探索し、正常な応答のあった外部ストレージに関するデバイス管理情報を該構成管理部内のメモリに登録する外部デバイス定義処理手段と、
論理デバイスの定義の対象が外部ストレージである場合、該ダーティデータ量を含むキャッシュ使用量情報を該管理端末に提供する手段と、
該管理端末からの指示により、該外部ストレージを割当てるためのストレージ制御部に関する情報を受付け、又は該キャッシュ使用量情報の中からダーティデータ量がより少ないストレージ制御部を該外部ストレージに割り当てるようにストレージ制御部を選択する選択手段と、
選択された該ストレージ制御部は、該ストレージ制御部内のディスク装置に関する物理デバイス情報、及び該外部ストレージに関するデバイス管理情報に従って論理デバイスの定義を設定する論理デバイス定義処理手段と、
該管理端末からの指示に基づき、該論理デバイス定義処理手段により定義された論理デバイスに対してパスの設定を行うパス定義処理手段と、
該ホストから受け付けた入出力要求のアクセス対象が、該ストレージ制御部のディスク装置もしくは該外部ストレージのディスク装置である場合、該入出力要求を該ストレージ制御部で処理する手段と、
を有することを特徴とするストレージシステム。
【請求項２０】
前記ストレージシステムは、更に、
該管理端末から外部ストレージの割当て変更の指示を受付ける手段と、
変更される先の該ストレージ制御部に関する予測ダーティデータ量を算出して、予測キャッシュ使用量の情報を該管理端末へ提供する手段と、を有し
該構成管理部は、該管理端末から変更指示を受け、変更先のストレージ制御部に割当て変更の指示を発行する手段を有し、
割当て変更の指示を受けた変更先のストレージ制御部は、指示された外部ストレージに対して論理デバイスの登録を行う手段を有し、
変更元のストレージ制御部は、外部ストレージのデータについてキャッシュミス化処理を行う手段を有する、
請求項１９に記載のストレージシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ストレージシステムに係り、特にクラスタ構成を有するストレージシステムにおける負荷分散のための制御、及びそのためのキャッシュメモリの制御に関する。
【背景技術】
【０００２】
企業のＩＴシステムにおいては大容量かつ高性能なストレージシステムが要求されてい

る。
この要求に応えるために、小容量のストレージを複数台導入することによって、大容量のデータを取り扱うことが実現できる。しかしストレージの台数の増加に伴い、障害・保守などによるストレージの管理コストの増加が問題となる。一方、1台のストレージにより大容量のシステムを提供するという方法もあるが、従来のような、メモリ、制御メモリ、内部転送機構といった計算機資源を共有するような形態のストレージにおいては、コストおよび技術的な要因により、現在求められているような大容量かつ高性能なストレージを実現するのは困難になってきている。

【０００３】

このような問題を解決するために、例えば特許文献１（ＵＳＰ６，２５６，７４０明細書）には、ストレージへのクラスタ技術の適用が開示されている。クラスタ技術はこれまで主にサーバなどホストコンピュータの分野で，大量の処理能力を実現する実装技術として用いられてきた。この技術をストレージに適用することで，大規模なストレージを比較的低コストで実装することが可能となる。このようなストレージをクラスタ構成ストレージと呼ぶ。
クラスタ構成ストレージは、比較的小構成の複数のストレージノードを相互結合網で結合し、大容量の単一ストレージを実現する。クラスタ構成ストレージに対する入出力要求は、対象データを保持するストレージノードに振り分けられ、各ストレージノードにおいて処理される。ストレージノードは、通常のストレージと同様に、ホストインターフェース、ディスク装置、制御プロセッサ、メモリ、制御メモリ、ディスクキャッシュなどが搭載され、これらの部位がストレージノード内で内部ネットワークにより結合される。各ストレージノードでは、これらの部位を用いてディスク装置への入出力要求処理を行う。
一般に、大規模なキャッシュメモリや制御メモリを共有メモリ空間として管理しようとすると、大量のアクセスに対応するために広帯域の相互結合網が必要となり、ストレージのコストが高くなるが、ストレージノードは規模が小さいためにキャッシュメモリや制御メモリに対するアクセス負荷が小さく、ストレージのコストを削減することができる。このため、クラスタ構成ストレージは大容量のストレージを低コストで実現することが可能である。

【０００４】

また、特許文献２（特開平１０－２８３２７２号公報）には、アクセスインターフェースの異なるＩ／Ｏサブシステム間において、オープン系のＩ／Ｏサブシステムで発生したデータを、そのサブシステムが直接接続されていないメインフレーム系のＩ／Ｏサブシステムに転送して、メインフレーム系のＩ／Ｏサブシステムの記憶装置内にバックアップする複合計算機システムが開示されている。メインフレーム系のディスク制御装置は自記憶装置のアドレスがオープン系のＩ／Ｏサブシステムに割当てられているかを示すテーブルを有し、このテーブルを参照することによりオープン系のＩ／Ｏサブシステムに対するアクセスを許容する。

【０００５】

【特許文献１】ＵＳＰ６，２５６，７４０明細書

【０００６】

【特許文献２】特開平１０－２８３２７２号公報

【発明の開示】

【発明が解決しようとする課題】

【０００７】

上記特許文献２には、オープン系のＩ／Ｏサブシステムがメインフレーム系Ｉ／Ｏシステムの記憶装置を外部記憶装置として使用するための技術は開示されているが、その記憶装置をクラスタ構成ストレージとして使用するための具体的な開示はない。また、オープン系のＩ／Ｏサブシステムとメインフレーム系Ｉ／Ｏシステムを２つのストレージサブシステムと想定した場合、これらサブシステムにおけるアクセス頻度に伴う負荷の偏りを如何に防止するかの示唆もされていない。

【０００８】
ところで、クラスタ構成ストレージに対して外部に位置付けられる外部ストレージを接続し、クラスタ構成ストレージが提供する機能を外部ストレージで使用することが考えられている。外部ストレージの外部デバイスに、ストレージノードの機能を提供するために、いずれかのストレージノードのキャッシュメモリを使用する。そのため、あるストレージノードに、多数の外部デバイスに関する処理が集中する場合や、アクセス頻度の高い外部デバイスの処理を行う場合には、外部デバイスのためにキャッシュメモリが大量に使用され、そのストレージノード内のデバイスに対するアクセス性能が低下するおそれがある。このキャッシュメモリの使用量の不均衡に伴うアクセス性能の低下に対する対策についても、上記特許文献のいずれにも記載されていない。

【０００９】
本発明の目的は、クラスタ構成ストレージにおいて、あるストレージノードに負荷が集中することを防止し、アクセス性能を向上させ得るストレージシステムを提供することにある。
本発明の他の目的は、外部ストレージ接続を行うクラスタ構成ストレージにおいて、各ストレージノードにおけるキャッシュ使用量を均等化させるストレージシステム及びキャシュメモリの制御方法を提供することにある。
【課題を解決するための手段】
【００１０】
本発明は、クラスタ構成のストレージシステムにおいて、ホストから入出力要求に従い、読み書きされるデータを格納するキャッシュメモリ及びキャッシュメモリに格納されるデータを保持するデバイスを有する複数のストレージ制御部と、読み書きされるデータを扱う論理デバイス及びキャッシュメモリを持つ外部ストレージをストレージ制御部に接続する手段と、複数のストレージ制御部のキャッシュメモリの使用状況を監視して把握する手段と、この把握手段によって得られるキャッシュメモリの使用状況に関する情報を参照して、キャッシュメモリの使用量を均等化させるように、いずれかのストレージ制御部を選択する手段と、を有し、選択手段により選択されたストレージ制御部により接続手段を介して外部ストレージの論理デバイスを制御するように構成したものである。
好ましくは、上記把握手段は、複数のストレージ制御部のキャッシュメモリ毎にそのダーティデータ量（第1のダーティデータ量）を取得し、選択手段は、得られたダーティデータ量に従い、ダーティデータ量のより少ない、例えば最も少ないストレージ制御部を選択して、外部ストレージの論理デバイスを制御する。
更に、前記把握手段は、外部ストレージの論理デバイスに格納されるデータであって、いずれかのストレージ制御部が有するキャッシュメモリに格納されているダーティデータの量（第２のダーティデータ量）を把握し、選択手段は、第1のダーティデータ量及び第2のダーティデータ量に従って、外部ストレージの論理デバイスを制御するためのストレージ制御部を選択する。
更に、外部ストレージの論理デバイスに格納されるデータについて、他のストレージへ非同期のリモートコピーを行う手段と、ストレージ制御部のキャッシュメモリに保持されていて、未だ他のストレージへ送信されていないデータ（サイドファイルデータ）の量を把握する手段とを有し、選択手段は、把握手段によって取得されたサイドファイルデータ量を参照して、外部ストレージの論理デバイスを制御するためのストレージ制御部を選択する。
更に、前記把握手段により取得されたキャッシュメモリの使用状況、又はダーティデータ量、若しくはサイドファイルデータ量を管理者へ提示するために管理端末又は管理サーバへ転送する手段を有し、前記選択手段は、管理端末又は管理サーバから指定されたストレージ制御部を選択する。
【００１１】
本発明に係るストレージシステムは、また、クラスタ構成のストレージシステムにおいて、ホストから入出力要求に従い、読み書きされるデータを一時的に格納するキャッシュ

メモリ、及びキャッシュメモリに格納されるデータを保持するデバイスを有する複数のストレージノードと、ホストから入出力要求に従い、読み書きされるデータを格納するキャッシュメモリ及びデバイスを有する外部ストレージと該ストレージノードと接続するインターフェースと、複数のストレージノードのキャッシュメモリの使用状況を監視して把握する手段と、把握手段によって得られるキャッシュメモリの使用状況に関する情報を参照して、キャッシュメモリの使用量を均等化させるように、あるストレージノードを選択する手段とを有し、選択手段により選択されたストレージノードによりインターフェースを介して外部ストレージのデバイスを制御するように構成される。
好ましくは、上記把握手段は、キャッシュメモリ毎のダーティデータ量（第1のダーティデータ量）を取得し、選択手段は、得られた該ダーティデータ量に従い、ダーティデータ量のより少ないストレージノードを選択して、外部ストレージを制御する。
更に、上記複数のストレージノードを含む第２のストレージから第3のストレージに対して非同期のリモートコピーを行う手段と、ストレージノードのキャッシュメモリに保持されていて、未だ該第３のストレージへ送信されていないデータ（サイドファイルデータ）の量を把握する手段とを有し、選択手段は、把握手段によって取得されたサイドファイルデータ量を参照して、ストレージノードを選択する。
また、上記把握手段により取得されたキャッシュメモリの使用状況、又はダーティデータ量、若しくはサイドファイルデータ量を管理者へ提示するために管理端末又は管理サーバへ転送する手段を有し、選択手段は、管理端末又は管理サーバから指定されたストレージノードを選択する。

【００１２】

本発明は、また、クラスタ構成のストレージシステムにおいてホストからの入出力要求を処理する方法として把握される。すなわち、本発明は、データを格納するデバイス及びデバイスに格納されるデータを一時的に格納するキャッシュメモリとを有する複数のストレージ部分でホストから入出力要求を処理するステップと、データを格納するデバイス及びキャッシュメモリを有する外部ストレージのデバイスを、あるストレージ部分によって制御するステップと、複数のストレージ部分におけるキャッシュメモリの使用状況を把握するステップと、得られたキャッシュメモリの使用状況に関する情報を参照して、キャッシュメモリの使用量を均等化させるように、あるストレージ部分を選択するステップと、選択されたストレージ部分を用いて外部ストレージに対するホストからの入出力要求を処理するステップと、を有する入出力処理方法である。

【００１３】

好ましい一例では、クラスタ構成ストレージにおいて、ストレージノードを、上位ホストインターフェースを制御するプロトコル変換部と、ディスク装置およびディスクキャッシュを制御するストレージ制御部に分割し、複数のプロトコル変換部と複数のストレージ制御部を相互結合機構で各部位が他の全ての部位と交信可能なように接続する。プロトコル変換部には、ホストインターフェースとそれを制御する制御プロセッサを搭載し、ホストインターフェースを介して受信した入出力要求を対象デバイス（ディスク装置）が属するストレージ制御部へ振り分ける処理を行う。ストレージ制御部には、ディスク装置とディスクキャッシュと制御プロセッサとメモリ、さらにデバイスへの入出力要求処理に必要な制御情報を格納する制御メモリを搭載し、プロトコル変換部より送信され、内部結合機構を介して転送された入出力要求を受信し、対象デバイスへの入出力処理を実行する。また、ストレージ制御部の制御プロセッサでは、データ複製やデータ再配置などのデータ連携機能を実現する各種処理を実行する。ストレージが提供する論理デバイス（以下、上位論理デバイス）と各ストレージ制御部が提供する論理デバイス（以下、下位論理デバイス）の対応は、同じく内部結合機構に接続された構成管理部で管理する。構成管理部では、こうしたデバイス管理に加えて、プロトコル変換部やストレージ制御部などの障害管理も行う。内部結合機構はプロトコル変換部、ストレージ制御部、構成管理部を接続するストレージ内部のネットワークで、各デバイスのアクセスデータや、各部位間の制御情報のやりとりに用いられる。

この例では、キャッシュメモリおよび制御メモリが各ストレージ制御部内の制御プロセッサ間でだけ共有されるため、メモリ帯域やメモリと制御プロセッサ間を接続するバックプレーン帯域を抑制し、製造コストを削減することができる。また、内部結合機構により、任意のプロトコル変換部が有する任意のホストインターフェースから、任意のストレージ制御部内の任意の論理デバイスへアクセスできる。

【００１４】

上記のようなクラスタ構成ストレージで、外部ストレージ接続を行う場合、外部ストレージをホストインターフェースに接続し、外部ストレージ内デバイス（外部デバイス）と下位論理デバイスの対応を構成管理部で管理する。また、ストレージ制御部でも下位論理デバイスと物理デバイスの対応と共に外部デバイスとの対応を管理できる情報を追加する。

外部ストレージである第一のストレージを上記のようなクラスタ構成を採る第二のストレージに接続し、外部デバイスを第二のストレージデバイスとして統合する場合、外部デバイスを特定のストレージ制御部の下位論理デバイスと対応づけてから下位論理デバイスを上位論理デバイスへ対応づける。これを、外部デバイスのストレージ制御部への割り当てと呼ぶ。そして、ホストからの上位論理デバイスに対する入出力要求を、ホストインターフェースを介して受信したプロトコル変換部で、上位論理デバイスが対応する下位論理デバイスが属するストレージ制御部へ入出力要求を転送する。

【００１５】

上記のような構成のクラスタ構成ストレージにおいて外部ストレージ接続を行う場合、プロトコル変換部に接続された外部ストレージは、任意のストレージ制御部と交信可能である。従って、外部デバイスに関する処理は任意のストレージ制御部により行うことが可能である。この際、クラスタ構成ストレージのシステム全体としての性能を向上させるためには、複数のストレージ制御部の中から、外部デバイスを割り当てるストレージ制御部を適切に選択する。この選択基準としてストレージ制御部におけるキャッシュ使用量を用い、複数のストレージ制御部のうち最もキャッシュ使用量の少ないものを選択する。

また、ストレージ制御部のキャッシュ使用状況は時間と共に変化するため、複数ストレージ制御部間にキャッシュ使用量の偏りが発生するが、複数ストレージ制御部間の、時間経過に対応したキャッシュ使用量の均等化を行うために、外部デバイスを割り当てるストレージ制御部を動的に変更する。

【発明の効果】

【００１６】

本発明によれば、クラスタ構成ストレージを構成する複数のストレージノードを有するストレージシステムにおいて、あるストレージノードに負荷が集中することを防止でき、アクセス性能を向上できる。また、各ストレージノードにおけるキャッシュ使用量を均等化させることができ、ストレージノードの負荷を分散させることができる。

【発明を実施するための最良の形態】

【００１７】

以下、本発明の複数の実施形態について説明される。最初に各実施形態の概要と参照される図面との関係について述べておく。

第１の実施形態は、図１～図１６を参照し、第１のストレージである外部ストレージ内のデバイスを、第２のストレージであるクラスタ構成ストレージの論理デバイスとして定義する際に、ダーティ属性を持つキャッシュ量に基づき、外部デバイスの割り当て対象とするストレージ制御部を選択し、割り当てる場合の例である。

【００１８】

第２の実施形態は、図１７～図１９を参照し、当該外部デバイスへの入出力処理を受け付けながら、外部デバイスの割り当て対象ストレージ制御部を変更する場合の例である。

第３の実施形態は、図２０～図２３を参照し、第１のストレージである外部ストレージ内のデバイスに対して後述する非同期リモートコピー機能を適用することを前提とし、外部デバイスをクラスタ構成ストレージである第２のストレージの論理デバイスとして定義す

る際に、サイドファイル属性を持つキャッシュ量に基づき、割り当て対象とするストレージ制御部を選択し、割り当てる場合の例である。

【００１９】

第４の実施形態は、図２４を参照し、第１のストレージである外部ストレージ内のデバイスに対して後述する非同期リモートコピー機能を適用することを前提とし、外部デバイスへのホスト入出力処理を受け付けながら、外部デバイスの割り当て対象ストレージ制御部を変更する場合の例である。

第５の実施形態は、図２５を参照し、ストレージノード自体を選択或いは切り替える例である。第１乃至第４の実施形態においては、ストレージノードをプロトコル変換部とストレージ制御部に分割し、ストレージ制御部を選択或いは切り替えるシステムであるのに対し、第５の実施形態ではプロトコル変換部とストレージ制御部の分割を行わないシステムを前提とし、負荷の分散を考慮してストレージノードそのものを選択又は切り替える。

【００２０】

第1の実施形態：

まず図１から図１６を参照して，第１の実施形態について説明する。

図１は、第１の実施形態における計算機システムのハードウェア構成を示す図である。

計算機システムは、１台以上のホストコンピュータ（単にホストという）１００と、管理サーバ１１０と、ファイバチャネルスイッチ１２０と、ストレージ１３０と、１台以上の外部ストレージ１８０ａ、１８０ｂ（１８０と総称する）と、管理端末１９０を含んで構成される。

【００２１】

ホスト１００、ストレージ１３０は、それぞれポート１０７、ポート１４１によりファイバチャネルスイッチ１２０のポート１２１に接続して、ＳＡＮ（Ｓｔｏｒａｇｅ Ａｒｅａ Ｎｅｔｗｏｒｋ）を構成する。さらに、外部ストレージ１８０ａと１８０ｂは各々ポート１８１によりストレージ１３０に接続し、ストレージ１３０によりストレージ１３０のデバイスとしてホスト１００へ提供される。また、ホスト１００やスイッチ１２０など全機器がＩＰネットワーク１７５を介して管理サーバ１１０に接続され、管理サーバ１１０で動作するＳＡＮ管理ソフトウェア（図示せず）によって統合管理される。なお、本実施形態では、ストレージ１３０は、管理端末１９０を介して管理サーバ１１０に接続する形態をとるものとする。

【００２２】

ホスト１００は、ＣＰＵ１０１やメモリ１０２などを有する計算機であり、ディスク装置や光磁気ディスク装置などの記憶装置１０３に格納されたオペレーティングシステムやアプリケーションプログラムなどのソフトウェアをメモリ１０２に読み上げ、ＣＰＵ１０１がメモリ１０２からそれらのプログラムを読み出して実行することで、所定の機能を達成する。キーボードやマウスなどの入力装置１０４やディスプレイ１０５を具備し、ホスト管理者などからの操作を受け付け、指示された情報を表示することが可能である。

【００２３】

管理サーバ１１０も、記憶装置１０３に格納されたＳＡＮ管理ソフトウェアなどをメモリ１１２に読み上げ、ＣＰＵ１１１がそれを読んで実行することで、計算機システム全体の運用・保守管理といった、所定の機能を達成する。また、インターフェース１１６からＩＰネットワーク１７５を介して、計算機システム内の各機器から構成情報、リソース利用率、性能監視情報などを収集し、それらの情報をストレージ管理者にディスプレイ１１５で提示し、入力装置１１４から受信した運用・保守指示を各機器に送信する。同処理は、図示しないＳＡＮ管理ソフトウェアにより行われる。

【００２４】

ファイバチャネルスイッチ１２０は、複数のポート１２１を有する。各ポート１２１には、ホスト１００のポート１０７、および、ストレージ１３０のポート１４１のいずれかが接続される。ファイバチャネルスイッチ１２０は、インターフェース１２３を有しており、ＩＰネットワーク１７５にも接続されている。ファイバチャネルスイッチ１２０は１

台以上のホスト１００がストレージ１３０を自由にアクセスできるようにするために使用される。この構成では、物理的には全てのホスト１００がファイバチャネルスイッチ１２０に接続されたストレージ１３０にアクセスすることが可能である。また、ファイバチャネルスイッチ１２０はゾーニングと呼ばれる特定ポートから特定ポートへの通信を制限する機能を有し、例えば、特定ストレージ１３０の特定ポート１４１へのアクセスを特定ホスト１００に制限する場合などに用いられる。接続元ポートと接続先ポートの組み合わせを制御する方法については、ファイバチャネルスイッチ１２０のポート１２１に割り当てられたポートＩＤを用いる方法、各ホスト１００のポート１０７やストレージ１３０のポート１４１が保持するＷＷＮ（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｎａｍｅ）を用いる方法などがある。

【００２５】

ストレージ１３０は、複数のプロトコル変換部１４０と複数のストレージ制御部１５０と構成管理部１６０を、相互結合網１７０で接続して構成される。

プロトコル変換部１４０は、複数のポート１４１と１つ以上の制御プロセッサ１４２とメモリ１４４を搭載し、ポート１４１から受信した入出力要求についてアクセス対象デバイスを特定し、転送制御部１４４より相互結合網１７０を介して適当なストレージ制御部１５０へ入出力要求やデータを転送する処理を行う。その際、制御プロセッサ１４２は入出力要求に含まれるポートＩＤおよびＬＵＮ（Ｌｏｇｉｃａｌ　Ｕｎｉｔ　Ｎｕｍｂｅｒ）からストレージ１３０がホスト１００へ提供する上位論理デバイス番号を算出し、さらに上位論理デバイスが対応するストレージ制御部１５０と下位論理デバイス番号を算出して、対象となるストレージ制御部１５０へ入出力要求を転送する。また、プロトコル変換部１４０は、外部ストレージ１８０など別のストレージを接続し、ストレージ制御部１５０からの入出力要求を外部ストレージ１８０へ送信し、外部ストレージ１８０へのリード／ライトを行うことができる。なお、本実施形態では、ポート１４１としては、ＳＣＳＩ（Ｓｍａｌｌ　Ｃｏｍｐｕｅｒ　Ｓｙｓｔｅｍ　Ｉｎｔｅｒｆａｃｅ）を上位プロトコルとしたファイバチャネルインターフェースを想定しているが、ＳＣＳＩを上位プロトコルとしたＩＰネットワークインターフェースなど、他のストレージ接続用ネットワークインターフェースであっても構わない。

【００２６】

ストレージ制御部１５０a～１５０c（１５０と総称する）は、１台以上のディスク装置１５７と、１つ以上の制御プロセッサ１５２とこれに対応するメモリ１５３、ディスクキャッシュ１５４、制御メモリ１５５、および転送制御部１５１を有する。制御プロセッサ１５２は、相互結合網１７０を介して転送制御部１５１により受信した当該ディスク装置１５７への入出力要求を処理する。制御プロセッサ１５２は、特に、ストレージ１３０がホスト１００に対してディスク装置１５７単体ではなく、ディスクアレイのように複数のディスク装置１５７を１つ又は複数の論理デバイスに見せかけている場合には、その処理や管理などを行う。ディスクキャッシュ１５４は、ホスト１００からのアクセス処理速度を高めるため、頻繁に読み出されるデータを格納したり、あるいはホスト１００からのライトデータを一時的に格納したりする。制御メモリ１５５は、ディスク装置１５７や複数のディスク装置１５７を組み合わせて構成される物理デバイスや同ストレージ１３０に接続した外部ストレージ１８０のデバイス（以下、外部デバイス）の管理や、外部/物理デバイスと下位論理デバイスの対応関係の管理を行うための情報を格納する。

【００２７】

なお、制御メモリ１５５に格納した制御情報の消失は、ディスク装置１５７に格納したデータへのアクセスが不可になる事態を引き起こすため、制御メモリはバッテリバックアップなどによる不揮発化や媒体障害への耐性向上のため二重化などの高可用化を行うことが望ましい。同様にディスクキャッシュ１５４を用いたライトアフタを行う場合、ディスクキャッシュ１５４に保持したディスク装置１５７未反映のデータ消失を避けるため、ディスクキャッシュ１５４も記録媒体の二重化や不揮発化により可用性を向上させることが望ましい。また、本実施形態におけるストレージ１３０は、複数のディスク装置１５７を

まとめて１つまたは複数の物理デバイスを定義し、１つの物理デバイスに１つの下位／上位論理デバイスを割り当て、ホスト１００に提供する。もちろん、個々のディスク装置１５７を１つの物理デバイスおよび１つの上位／下位論理デバイスとしてホスト１００に見せるようにしてもよい。

【００２８】

構成管理部１６０は、制御プロセッサ１６２、メモリ１６３、制御メモリ１６４、記憶装置１６５、転送制御部１６１、インターフェース１６６を有する。固定ディスク装置などの記憶装置１６５に格納した制御プログラムをメモリ１６３に読み上げ、これを制御プロセッサ１６２で実行することでストレージ１３０の構成管理という所定の動作を実現する。インターフェース１６６で接続した管理端末１９０からストレージ管理者への構成情報提示、管理者からの保守・運用指示の受領を行い、受領した指示に従い、ストレージ１３０の構成変更などを行う。ストレージ１３０の構成情報は制御メモリ１６４に保持する。制御メモリ１６４上の構成情報はプロトコル変換部１４０の制御プロセッサ１４２やストレージ制御部１５０の制御プロセッサ１５２から参照・更新することで、各部位間での構成情報の共有を実現する。なお、構成管理部１６０が障害などにより動作不可に陥った場合、ストレージ１３０全体がアクセス不可に陥るため、構成管理部１６０内の各部位の二重化もしくは、構成管理部１６０自体の二重化が望ましい。

【００２９】

相互結合網１７０は、例えばクロスバスイッチであり、プロトコル変換部１４０、ストレージ制御部１５０、構成管理部１６０を接続し、各部位間のデータおよび制御情報、構成情報のやりとりを実現する。この相互結合網１７０により、構成管理部１６０が全部位の構成管理と構成情報を配布したり、プロトコル変換部１４０の任意のポート１４１からストレージ制御部１５０の任意の下位論理デバイスへアクセスすることが可能となる。なお、可用性向上の観点から相互結合網も多重化されていることが望ましい。

【００３０】

管理端末１９０は、例えばＰＣ（パーソナルコンピュータ）であり、ストレージ管理プログラムを動作させる機能と、ストレージ管理者による入出力操作のための機能を有し、構成情報参照、構成変更指示、特定機能の動作指示など、ストレージ１３０の保守運用に関するストレージ管理者又は管理サーバ１１０からのインターフェースとなる。このために、プログラムやデータを格納する記憶装置１９４、記憶装置１９４から読み込まれたプログラムや種々のデータを格納するメモリ１９３、プログラムを実行するＣＰＵ１９２、及び入出力機能として入力装置１９５、ディスプレイ１９６を有する。

なお、変形例では、この管理端末１９０を省略して、ストレージ１３０を直接管理サーバ１１０へ接続し、管理サーバ１１０で動作する管理ソフトウェアで管理してもよい。

【００３１】

外部ストレージ１８０は、ストレージ１３０と同様に、ポート１８１から受信したディスク装置１８６への入出力要求を処理する機能を有する。即ち、内部インターフェースにポート１８５を介して接続された大容量のディスク１８６、ディスクキャッシュ１８４、メモリ１８３、および制御プロセッサ１８２を有する。

なおこの例では、外部ストレージ１８０をストレージ１３０より小規模構成としているが、ストレージ１３０と同じ構成、規模のストレージであってもかまわない。

【００３２】

次に、ストレージ１３０のソフトウェア構成について説明する。

図２は、ストレージ１３０および管理端末１９０の制御メモリやメモリに格納する制御情報とストレージ制御処理プログラムを示したソフトウェアの構成図である。

なお、以降の説明では，簡略化のため、プロトコル変換部１４０をＰＡ（Ｐｒｏｔｏｃｏｌ　Ａｄａｐｔｏｒ）、ストレージ制御部１５０をＳＡ（Ｓｔｏｒａｇｅ　Ａｄａｐｔｏｒ）、構成管理部１６０をＭＡ（Ｍａｎａｇｅｍｅｎｔ　Ａｄａｐｔｏｒ）、管理端末１９０をＳＴ（Ｓｅｒｖｉｃｅ　Ｔｅｒｍｉｎａｌ）と表記する。

【００３３】

この例において、ストレージ１３０内のデバイス階層は次の通りである。ＳＡ１５０内では複数のディスク装置１５７によりディスクアレイが構成されて物理デバイスが形成される。また、ＰＡ１４０に接続した外部ストレージ１８０の外部デバイスは、ＰＡ１４０で認識された後、ＭＡ１６０で管理される。ＳＡ１５０では物理デバイスおよび外部デバイスに対して下位論理デバイスが割り当てられる。下位論理デバイスは各ＳＡ１５０内における論理デバイスであり、その番号は各ＳＡ１５０で独立的に管理される。下位論理デバイスは、ＭＡ１６０で管理する上位論理デバイスに対応付き、ストレージ１３０のデバイスとしてホスト１００に提供される。ストレージ１３０のキャッシュ使用量情報としては、ＳＴ１９０のメモリ１９３にＣＡキャッシュ使用量情報２２２および外部デバイスキャッシュ使用量情報２２４を格納する。

これら種々の管理情報および各種の処理については、図３～図１６を参照して後述される。

【００３４】

図３及び図４はストレージ１３０内のキャッシュ使用量を管理するためのテーブルの構成を示す図である。

これらテーブルの説明の前に、まずディスクキャッシュ１５４の制御方法について説明する。

ホスト１００から受信した書き込みデータはディスクキャッシュ１５４に格納された時点で、ストレージ制御部１５０はホスト１００に対して書き込み完了報告を返し、この書き込み完了の報告後に適当なタイミングで物理／外部デバイスに書き込みデータを反映する処理を行う。この処理をライトアフタ処理と呼ぶ。ディスクキャッシュ１５４に格納されたデータであって物理／外部デバイスに未反映のデータをダーティデータという。ダーティデータは、物理／外部デバイスに反映されるまでの間、ディスクキャッシュ１５４上に保持し続ける必要がある。

【００３５】

しかし、ダーティデータの物理／外部デバイスへの書き込み処理速度とホスト１００からの書き込み要求の頻度によっては、ダーティデータがディスクキャッシュ１５４上に大量に滞留し、新たな入出力処理要求に対してディスクキャッシュ１５４を割り当てることができなくなる可能性がある。このような状態をなるべく回避し、ストレージ１３０全体としてのアクセス性能を向上させるためには、各ＳＡ１５０内デバイスへの書き込み要求頻度を均等化することが有効である。ＳＡ１５０内のデバイスに対する書き込み要求頻度はデバイスに対応するダーティ量の時間平均と相関があるので、各ＳＡ１５０内デバイスへの書き込み要求頻度の均等化のためにダーティ量の時間平均を指標として用いる。

【００３６】

ディスクキャッシュ１５４は、セグメントと呼ばれる、ディスクキャッシュ１５４を一定量に分割した単位により管理される。ダーティデータ量はダーティ属性を持つセグメントの個数により算出される。図示しないが、ＳＡ１５０は、制御メモリ１５５上にダーティ属性をもつセグメントの数を保持するダーティセグメントカウンタ、およびストレージ制御部１５０内で管理されている外部デバイスに対応する、ダーティ属性をもつセグメントの数を保持するダーティセグメントカウンタを持つ。

【００３７】

外部ストレージ接続においては、ストレージ１３０に接続されたホスト１００が外部デバイスにアクセスする際、ストレージ１３０のＳＡ１５０上のディスクキャッシュ１５４を外部デバイスに対して使用させることにより、外部デバイスに対するアクセス性能を向上させることが可能である。

【００３８】

図３は、ＳＡキャッシュ使用量情報２２２の構成を示す。

ＳＡキャッシュ使用量情報２２２は、ストレージ１３０内のＳＡ１５０の特定時間帯におけるダーティ量を管理するために、ＳＡ番号３０１および対応するダーティ量情報３０２を保持する。ダーティ量情報３０２は、ＳＡ１５０において、後述する集計時間情報５０

１に含まれる時間帯に、一定の時間間隔で各ＳＡ１５０に対応するダーティセグメントカウンタを参照し、それらの平均値を求めることにより算出される。このようにして得られたダーティ量情報３０２はＳＡ番号３０１と共にＳＴ１９０に送信され、ＳＴ１９０内のＣＡキャッシュ使用量情報２２２に格納される。

【００３９】

図４は、外部デバイスキャッシュ使用量情報２２４の構成を示す。

外部デバイスキャッシュ使用量情報２２４は、ストレージ１３０で管理されている外部デバイスのダーティ量に関する情報であり、外部デバイス番号４０１および対応するダーティ量情報４０２を保持する。即ち、外部デバイスキャッシュ使用量情報２２４は、外部デバイス番号４０１で示される外部デバイスを管理しているストレージ１３０内のSA１５０において、外部デバイスに格納されるデータとしてSA１５０内のキャッシュに格納しているダーティデータの量を保持するものである。ダーティ量情報４０２の値は、ＳＡ１５０において、後述する集計時間情報５０１に含まれる時間帯に、一定の時間間隔で外部デバイスに対応するダーティセグメントカウンタを参照し、それらの平均値を求めることにより算出される。このようにして得られたダーティ量情報４０２は外部デバイス番号４０１と共にＳＴ１９０に送信され、ＳＴ１９０内の外部デバイスキャッシュ使用量情報２２４に格納される。

【００４０】

図５は、管理端末制御情報２７０の構成を示す。

管理端末制御情報２７０は、ストレージ管理者により入力装置１９５から入力され、ＳＴ１９０のメモリ１９３上に保持される。またＳＡ１５０のメモリ１５３上には、管理端末制御情報の複製２７１が保持される。集計時間情報５０１は前述のＳＡキャッシュ使用量情報２２２および外部デバイスキャッシュ使用量情報２２４に格納するキャッシュ使用量情報を算出する際に、どの時間帯の時間平均を算出するかを指定するものである。割当変更時間情報５０２は、ＳＴ１９０が構成管理部１６０に対して外部デバイスの割り当て変更の指示をいつ行うかを指定する情報である。

【００４１】

ストレージ１３０の構成管理情報としては、ＳＡ１５０の制御メモリ１５５に下位論理デバイス管理情報２０１と物理デバイス管理情報２０２とキャッシュ管理情報２０３を、ＭＡ１６０の制御メモリ１６４に上位論理デバイス管理情報２０４と外部デバイス管理情報２０５とＬＵパス管理情報２０６を格納する。

【００４２】

図６は、上位論理デバイス管理情報２０４の構成を示す。

各上位論理デバイスにつき、上位論理デバイス番号６０１から接続ホスト名６０６までの情報組を保持する。

サイズ６０２には、上位論理デバイス番号６０１により特定される上位論理デバイスの容量が格納される。対応のＳＡ番号、下位論理デバイス番号６０３には、上位論理デバイスが対応する下位論理デバイスの番号と下位論理デバイスが属するＳＡ番号が格納される。上位論理デバイスが未定義の場合、このエントリには無効値が設定される。この下位論理デバイス番号が特定ＳＡ１５０の下位論理デバイス管理情報２０１のエントリ番号となる。

【００４３】

デバイス状態６０４には、上位論理デバイスの状態を示す情報が設定される。状態としては、「オンライン」、「オフライン」、「未実装」、「障害オフライン」が存在する。「オンライン」は、上位論理デバイスが正常に稼動し、ホスト１００からのアクセスが可能な状態であることを示す。「オフライン」は、上位論理デバイスは定義され、正常に稼動しているが、ＬＵパス未定義などでホスト１００からのアクセスはできない状態にあることを示す。「未実装」は、上位論理デバイスが定義されておらずホスト１００からのアクセスはできない状態にあることを示す。「障害オフライン」は、上位論理デバイスに障害が発生してホスト１００からのアクセスができないことを示す。

なお、本実施形態では、簡単のため、上位論理デバイスは、製品の出荷時にあらかじめディスク装置１５７上に作成された物理デバイスへ割り当てられた、下位論理デバイスへ割り当てられているものとする。このため、利用可能な上位論理デバイスについてはデバイス状態６０４の初期値は「オフライン」状態、その他は「未実装」状態となる。

【００４４】

エントリ６０５のポート番号には、上位論理デバイスが複数のポート１４１のどのポートに接続されているかを表す情報が設定される。各ポート１４１には、ストレージ１３０内で一意な番号が割り振られており、上位論理デバイスがＬＵＮ定義されているポート１４１の番号が記録される。また、同エントリのターゲットＩＤとＬＵＮは、上位論理デバイスを識別するための識別子である。ここでは、これらの識別子として、ＳＣＳＩ上でホスト１００からデバイスをアクセスする場合に用いられるＳＣＳＩ－ＩＤ、ＬＵＮが用いられる。

接続ホスト名６０６は、このデバイスにアクセスが許可されているホスト１００を識別するホスト名である。ホスト名としては、ホスト１００のポート１０７に付与されたＷＷＮ(Ｗｏｒｌｄ　Ｗｉｄｅ　Ｎａｍｅ)など、ホスト１００もしくはポート１０７を一意に識別可能な値であればよい。同じストレージ１３０には、このほかに、各ポート１４１のＷＷＮなどの属性に関する管理情報を保持する。

【００４５】

図７は、ＬＵパス管理情報２０６の構成を示す。

ストレージ１３０内の各ポート１４１につき、有効なＬＵＮ分の情報を保持する。ターゲットＩＤ／ＬＵＮ７０２は、ポート番号７０１に対応するＬＵＮのアドレスを格納する。対応上位論理デバイス番号７０３には、ＬＵＮを割り当てた上位論理デバイスの番号を格納する。接続ホスト名７０４は、ポート１４１のＬＵＮに対してアクセスを許可されているホスト１００を示す情報である。一つの上位論理デバイスに対して複数ポート１４１のＬＵＮが定義されている場合、それら全ＬＵＮの接続ホスト名７０４の和集合が上位論理デバイス管理情報２０３の接続ホスト名６０６に保持される。

【００４６】

図８は、下位論理デバイス管理情報２０１の構成を示す。

各ＳＡ１５０毎に下位論理デバイスにつき、下位論理デバイス番号８０１から対応上位論理デバイス番号８０５までの情報組を保持する。

サイズ８０２には、下位論理デバイス番号８０１により特定される下位論理デバイスの容量が格納されている。対応物理／外部デバイス番号８０３には、下位論理デバイスが対応するＳＡ１５０内の物理デバイス番号、もしくは外部デバイス番号が格納される。物理／外部デバイスへ未割り当ての場合、このエントリには無効値が設定される。このデバイス番号は、物理デバイス管理情報２０２、もしくは外部デバイス管理情報２０５のエントリ番号となる。

デバイス状態８０４には、下位論理デバイスの状態を示す情報が設定される。下位論理デバイスの状態を示す値は上位論理デバイス管理情報２０３のデバイス状態６０４と同様であるため、説明を省略する。対応上位論理デバイス番号８０５には、下位論理デバイスが対応する上位論理デバイス番号が設定される。

【００４７】

図９は、ＳＡ１５０内のディスク装置１５７から構成される物理デバイスを管理する物理デバイス管理情報２０２の構成を示す。

各ＳＡ１５０毎に物理デバイスにつき、物理デバイス番号９０１からディスク内サイズ９０９の情報組を保持する。サイズ９０２には、物理デバイス番号９０１により特定される物理デバイスの容量が格納されている。対応下位論理デバイス番号９０３には、物理デバイスが対応するＳＡ１５０内の下位論理デバイス番号が格納される。下位論理デバイスへ未割り当ての場合、当該エントリには無効値が設定される。

【００４８】

デバイス状態９０４には、物理デバイスの状態を示す情報が設定される。状態としては

、「オンライン」、「オフライン」、「未実装」、「障害オフライン」が存在する。「オンライン」は、物理デバイスが正常に稼動し、下位論理デバイスに割り当てられている状態であることを示す。「オフライン」は、物理デバイスは定義され、正常に稼動しているが、下位論理デバイスに未割り当てであることを示す。「未実装」は、物理デバイスがディスク装置１５７上に定義されていない状態にあることを示す。「障害オフライン」は、物理デバイスに障害が発生して下位論理デバイスに割り当てられないことを示す。なお、本実施形態では、簡単のため、物理デバイスは製品の工場出荷時にあらかじめディスク装置１５７上に作成されているものとする。このため、利用可能な物理デバイスについてはデバイス状態６４の初期値は「オフライン」状態、その他は「未実装」状態となる。

【００４９】

ＲＡＩＤ構成９０５には、物理デバイスが割り当てられたディスク装置１５７のＲＡＩＤレベル、データディスクとパリティディスク数などＲＡＩＤ構成に関連する情報が保持される。同じように、ストライプサイズ９０６には、ＲＡＩＤにおけるデータ分割単位(ストライプ)長が保持される。ディスク番号リスト９０７には、物理デバイスが割り当てられたＲＡＩＤを構成する複数のディスク装置１５７の番号が保持される。この番号はＳＡ１５０内でディスク装置１５７を識別するために付与した一意な値である。ディスク内開始オフセット９０８とディスク内サイズ９０９には、物理デバイスデータが各ディスク装置１５７内のどの領域に割り当てられているかを示す情報である。本実施形態では簡単のため、全物理デバイスについてＲＡＩＤを構成する各ディスク装置１５７内のオフセットとサイズを統一している。

【００５０】

図１０は、外部デバイス管理情報２０５の構成を示す。

この管理情報２０５は、ストレージ１３０に接続し、ストレージ１３０の下位論理デバイスと対応づける、外部ストレージ１８０内のデバイスを管理するために、ストレージ１３０全体で各外部デバイス毎に、外部デバイス番号１００１乃至ターゲットポートＩＤ／ターゲットＩＤ／ＬＵＮリスト１００８の組を保持する。

外部デバイス番号１００１にはストレージ１３０で割り当てた、ストレージ１３０内で一意な値を保持する。サイズ１００２には、外部デバイス番号１００１により特定される外部デバイスの容量が格納される。対応ＳＡ番号、下位論理デバイス番号１００３には、外部デバイスが対応するストレージ１３０内のＳＡ番号と下位論理デバイス番号が格納される。下位論理デバイスへ未割り当ての場合、このエントリには無効値が設定される。

デバイス状態１００４には、当該外部デバイスの状態を示す情報が設定されるが、各状態の意味は物理デバイス管理情報２０２内デバイス状態９０４と同じである。ストレージ１３０の初期状態は外部ストレージを接続していないため、デバイス状態１００４の初期値は「未実装」となる。

【００５１】

ストレージ識別情報１００５には、外部デバイスを搭載する外部ストレージ１８０の識別情報を保持する。ストレージ識別情報としては、同ストレージのベンダ識別情報と各ベンダが一意に割り振る製造シリアル番号の組み合わせ、などが考えられる。外部ストレージ内デバイス番号１００６には、外部デバイスが対応する外部ストレージ１８０内のデバイス番号を保持する。外部デバイスは外部ストレージ１８０の論理デバイスであるから、本エントリには外部ストレージ１８０の論理デバイス番号を保持する。

ＰＡ番号／イニシエータポート番号リスト１００７には、外部デバイスへアクセス可能なストレージ１３０のポート１４１とそれが属するＰＡ１４０の番号のリストが保持される。ターゲットポートＩＤ／ターゲットＩＤ／ＬＵＮリスト１００８には、外部デバイスが外部ストレージ１８０の１つ以上のポート１８１にＬＵＮ定義されている場合、それらのポート１８１のポートＩＤおよび外部デバイスが割り当てられたターゲットＩＤ／ＬＵＮが１つ又は複数個保持される。

【００５２】

次に、再び図２に戻り、各部位のメモリ内に格納される情報およびプログラムについて

説明する。
制御メモリ１５５，１６４に格納した各制御情報は各部位の制御プロセッサから参照・更新可能であるが、その際に相互結合網１７０などを介したアクセスが必要となる。よって、処理性能向上のため、各制御プロセッサで動作する処理に必要な制御情報の複製を各部位のメモリに保持する。構成変更により制御情報が更新された場合は、相互結合網１７０を介してその旨を他部位に通知し、最新情報を制御メモリから各部位メモリへ取り込ませる。
別の方法としては、例えば、制御メモリに保持した構成情報毎に制御メモリ上に更新有無を示すフラグを設け、各部位の制御プロセッサは処理開始時もしくは各構成情報を参照する毎に同フラグを参照して更新有無をチェックする方法などが考えられる。また、各部位のメモリにはこうした制御情報の複製に加えて、各制御プロセッサで動作する制御プログラムが格納される。
【００５３】
この実施例では、外部デバイスを含むデバイスを特定サーバへ割り当ててストレージ１３０において使用可能とする処理、外部デバイスを含むストレージ１３０のデバイスへの入出力要求処理と、割り当てられた外部デバイスの割当を変更する処理を例にしてその制御方法を説明する。
外部デバイスを含むデバイスを特定サーバへ割り当てて使用可能とする処理については、外部デバイス定義、論理デバイス定義、ＬＵパス定義の大きく３つの処理に分けることができる。
【００５４】
図１１は、外部デバイス定義処理２５３の処理フローを示す。
外部デバイス定義処理２５３は、ストレージ１３０管理下に外部ストレージ１８０のデバイスを外部デバイスとして導入する場合の処理である。
まず、管理端末１９０又は管理サーバ１１０からの外部ストレージ１８０接続指示を受け付けたＳＴ１９０がＭＡ１６０に接続指示を送信する（１１０１）。接続指示には、対象となる外部ストレージ１８０を特定する情報、例えば、外部ストレージ１８０のポート１８１のＷＷＮや外部ストレージデバイスへのＩｎｑｕｉｒｙコマンド送信などにより得られる装置識別情報もしくはその両方の情報と外部ストレージ１８０へ接続するポート１４１番号が付加される。ＭＡ１６０では外部ストレージ１８０の接続指示を受信し、指示された全ポート１４１番号に対応する全ＰＡ１４０へ外部ストレージ接続指示を送信する（１１０２）。
ＰＡ１４０では接続指示に付加された外部ストレージ１８０識別情報を用いて接続対象となる外部デバイスを探索する（１１０３）。具体的には、外部ストレージ識別情報としてポート１８１のＷＷＮをもらう場合は、指定されたポート１４１から外部ストレージ１８０のポート１８１の全ＬＵＮに対してＩｎｑｕｉｒｙコマンドを送信して、正常な応答のあったＬＵＮを外部デバイス登録候補とする。識別情報として装置識別情報しかもらわない場合には、ＰＡ１４０の全ポート１４１から検出した全ノードポートのそれぞれに対して（ノードポートログイン時に検出済み）、全ＬＵＮについてＩｎｑｕｉｒｙコマンドを送信し、正常に応答のあったデバイスについて、返却情報内の装置識別情報を接続指示に付加された値と比較する。そして、検出した外部デバイス登録候補についての情報リストをＰＡ１４０はＭＡ１６０へ返却する（１１０４）。このときの情報には外部デバイスについて外部デバイス管理情報２０５を設定するのに必要な情報を含んでいる。
【００５５】
ＭＡ１６０では、受け取った外部デバイスリストにある外部デバイスに採番し、外部デバイス管理情報２０５へデバイス情報を登録し、同情報に更新のあった旨を各部へ通知する（１１０５）。外部デバイス管理情報２０５への情報登録について、具体的には、割り当てた外部デバイス番号のエントリにおいて、Ｉｎｑｕｉｒｙ情報によりサイズ１００２、ストレージ識別情報１００５、外部ストレージ内デバイス番号１００６を、ＰＡ１４０からの情報によりエントリ１００７、１００８をそれぞれ設定する。エントリ１００３の

デバイス番号は未割当のため初期値である無効値を設定する。デバイス状態１００４には「オフライン」状態が設定される。
更新通知を受けた各ＳＡ１５０，ＰＡ１４０では、ＭＡ１６０の制御メモリ１６４の外部デバイス管理情報２０５を各々のメモリに取り込む。ＳＴ１９０では同情報のメモリ取込みに加えて、外部デバイス定義処理の完了を要求元である管理端末１９０又は管理サーバ１１０に報告する（１１０６）。

【００５６】

なお、本実施例では、ストレージ１３０に対して管理端末１９０又は管理サーバ１１０が接続指示と共に、導入対象となる外部ストレージ１８０を指示する形態をとる。しかし、外部ストレージ１８０の接続指示のみをストレージ１３０に指示し、ストレージ１３０で全ポート１４１から検出した全ストレージの全デバイスを外部デバイスとして登録しても構わない。また、特に明示的な接続指示を与えず、ストレージ１３０に外部ストレージ１８０が接続された契機にストレージ１３０が検出可能な全デバイスを外部デバイスとして登録してもよい。

【００５７】

図１２は、論理デバイス定義処理２５５の処理フロー図である。
論理デバイス定義処理２５５は、管理端末１９０又は管理サーバ１１０からの指示を受けて、ストレージ１３０が搭載する物理デバイスもしくは外部デバイス定義処理２５３で定義された外部デバイスに対して下位論理デバイスを定義する処理である。
まず、ＳＴ１９０は論理デバイス定義指示を受け付ける（１２０１）。この指示には、論理デバイス定義対象となる物理／外部デバイス番号と定義する上位論理デバイス番号が付加される。論理デバイス定義対象となるデバイスが外部デバイスでない場合は、さらに、ＳＡ番号と下位論理デバイス番号が付加される。

【００５８】

なお、本実施形態では、説明の簡略化のため１つの物理／外部デバイスに対して１つの論理デバイスを割り当てるものとするが、２つ以上の物理／外部デバイスからなるデバイスグループに対して１つの論理デバイスを、１つの物理／外部デバイスに対して２つ以上の論理デバイスを、２つ以上の物理／外部デバイスからなるデバイスグループに対して２つ以上の論理デバイスを定義しても構わない。ただし、それぞれの場合、下位論理デバイス管理情報２０１に、物理／外部デバイス内での当該論理デバイスの開始位置およびサイズなどの付加情報が必要となる。論理デバイス定義対象となるデバイスが物理デバイスの場合は、ＳＴ１９０はＭＡ１６０に対して論理デバイス定義指示を送信する（１２０７）。

【００５９】

論理デバイス定義対象となるデバイスが外部デバイスの場合は（１２０２）、ＳＴ１９０はストレージ管理者にSAキャッシュ使用量情報２２２を提示し、必要に応じて、外部デバイスを割り当てるＳＡ番号および下位論理デバイス番号を受け付ける（１２０３）。
ステップ１２０３において外部デバイスの割当対象ＳＡ番号および下位論理デバイス番号が指定されていた場合には（１２０４）、ＳＴ１９０はＭＡ１６０に対して論理デバイス定義指示を送信する（１２０７）。
ステップ１２０４において、外部デバイスの割当対象ＳＡ番号および下位論理デバイス番号が指定されていない場合には、ＳＴ１９０は外部デバイスの割当対象ＳＡ１５０を選択するために、SAキャッシュ使用量情報２２２を参照する（１２０５）。ＳＴ１９０はSAキャッシュ使用量情報２２２の中から最もダーティ量の少ないＳＡ１５０を割り当て対象として選択する（１２０６）。また、ＳＡ１５０における未使用下位論理デバイス番号を選択する。その後、ＭＡ１６０に対して論理デバイス定義指示を送信する（１２０７）。

【００６０】

論理デバイス定義指示を受信したＭＡ１６０は、定義指示情報より対象ＳＡ１５０を特定し、ＳＡ１５０に定義指示を送信する（１２０８）。対象ＳＡ１５０では、指示された物理／外部デバイスに対して下位論理デバイスを登録する（１２０９）。具体的には、下

位論理デバイス管理情報２０１の対象デバイスエントリについて、物理／外部デバイスのサイズとデバイス番号をエントリ８０２、８０３に、対応上位論理デバイス番号８０５に上位論理デバイス番号を、デバイス状態８０４に「オンライン」をそれぞれ設定する。また、物理／外部デバイスの対応ＳＡ番号／下位論理デバイス番号を設定し、デバイス状態を「オンライン」に更新する。登録が完了したらＳＡ１５０はＭＡ１６０にその旨通知する。

次にＭＡ１６０では、対象上位論理デバイスを下位論理デバイスへ定義し、情報更新の旨を各部位に通知する（１２１０）。具体的には、上位論理デバイス管理情報２０３のデバイスエントリについてサイズ６０２と対応ＳＡ番号、下位論理デバイス番号６０３を設定し、デバイス状態を「オフライン」状態に、エントリ６０５，６０６は未割当のため無効値を設定する。情報更新を通知されたＰＡ１４０では更新のあった管理情報をメモリ１４３へ取り込み、ＳＴ１６０では情報取込み後、論理デバイス定義処理の完了を要求元へ報告する（１２１１）。

【００６１】

図１３は、ＬＵパス定義処理２５２の処理フロー図である。

ＬＵパス定義指示を受け付けたＳＴ１９０がＭＡ１６０に対して同指示を転送する（１３０１）。同指示には、対象上位論理デバイス番号とＬＵを定義するポート１４１番号とＬＵＮに加えて、同ＬＵをアクセスするホスト１００の識別情報（ポート１０７のＷＷＮなど）が付加される。ＬＵパス定義指示を受け付けたＭＡ１６０は、上位論理デバイスに対してＬＵパス登録を行う（１３０２）。具体的には、上位論理デバイス管理情報２０３のデバイスエントリのポート番号、ターゲットＩＤ、ＬＵＮ６０５と接続ホスト名６０６に対応情報を設定し、ＬＵパス管理情報２０６のポート１４１に対応する空きエントリにターゲットＩＤ／ＬＵＮ７０１を始めとする構成情報を設定する。登録が完了したら、その旨各部位へ通知し、ＰＡ１４０で情報取込み、ＳＡ１９０で情報取込みと要求元への完了報告を行う。

【００６２】

以上の３つの処理により、外部デバイスをストレージ１３０のデバイスとして登録し、ストレージ１３０内のＳＡ１５０のキャッシュ使用量の均等化を考慮していずれかのＳＡ１５０へ割り当て、ホスト１００からのアクセスが可能となる。

【００６３】

次に、このように外部デバイスが割り当てられている状態での、ホスト１００からの入出力要求処理方法について、ＰＡ１４０、ＳＡ１５０でのリードコマンド処理、ライトコマンド処理、ライトアフタ処理の３つに分けて説明する。

【００６４】

図１４は、リードコマンド処理２６１の処理フロー図である。

ホスト１００がストレージ１３０内の外部デバイスを含むデバイスのデータを読み出す場合の処理を説明するものである。

ＰＡ１４０は、ホスト１００が発行したリードコマンドを特定ポート１４１で受信する（１４０１）。ＰＡ１４０は、受信したコマンドの解析を行い、要求データに対応する上位論理デバイス番号を割り出し、上位論理デバイス管理情報から対応するＳＡ番号、下位論理デバイス番号を算出し（１４０２）、ＳＡ番号に対応するＳＡ１５０にリード要求を転送する（１４０３）。ＳＡ１５０は転送制御部１５１よりリード要求を受け取り、要求データがディスクキャッシュ１５４上にあるかどうかを、キャッシュ管理情報２０３を参照することにより判定する。

【００６５】

要求データがディスクキャッシュ１５４上にある場合（キャッシュヒット）、ＳＡ１５０は該当するデータを要求の発行元であるＰＡ１４０に送信する（１４１０）。ＳＡ１５０よりデータを受信したＰＡ１４０は、ポート１４１を通じて、ホスト１００へデータを送信する（１４１１、１４１２）。

【００６６】

一方、要求データがディスクキャッシュ１５４上に無い場合（キャッシュミス）には、ＳＡ１５０はキャッシュ管理情報２０３を更新し、ディスクキャッシュ１５４上に要求データを格納する領域を確保する。要求が外部デバイスに対するものでない場合は、物理デバイスから要求データを読み出し、対応するディスクキャッシュ１５４領域にデータを格納する。以降の動作フローはキャッシュヒットした場合と同様である（１４１～１４１２）。要求が外部デバイスに対するものである場合、ＳＡ１５０は、ＰＡ１４０を介して外部ストレージ１８０からデータを読み込み（１４１３～１４１６）、対応するディスクキャッシュ１５４領域にデータを格納する（１４０９）。以降の動作フローはキャッシュヒットした場合と同様である（１４１０～１４１２）。以上のようにして、ホスト１００からのリード要求に対して、外部デバイスを含むデバイスからデータを読み出して、ホスト１００へ送信する。

【００６７】

図１５は、ライトコマンド処理２６２の処理フロー図である。

ホスト１００がストレージ１３０内のディスクキャッシュ１５４にデータを格納する場合の処理を説明するものである。

まず、ＰＡ１４０は、ホスト１００が発行したライトコマンドを特定ポート１４１で受信する（１５０１）。ＰＡ１４０は、受信したコマンドの解析を行い、要求データに対応する上位論理デバイス番号を割り出し、上位論理デバイス管理情報から対応するＳＡ番号、下位論理デバイス番号を算出し（１５０２）、ＳＡ番号に対応するＳＡ１５０にライト要求を転送する（１５０３）。ＳＡ１５０は転送制御部１５１よりライト要求を受け取り、要求データがディスクキャッシュ１５４上にあるかどうかを、キャッシュ管理情報２０３を参照することにより判定する（１５０４）。

【００６８】

要求データがディスクキャッシュ１５４上にあった場合（キャッシュヒット）、ＳＡ１５０は要求元のＰＡ１４０に対してライト準備完を通知する（１５０７）。ＳＡ１５０よりライト準備完の通知を受信したＰＡ１４０は、ポート１４１を介してホスト１００にライト準備完を通知する（１５０８、１５０９）。その後、ＰＡ１４０はポート１２１を介してホスト１００からデータを受信し、そのデータを該当するＳＡ１５０へ送信する（１５１０、１５１１）。ＰＡ１４０からデータを受け取ったＳＡ１５０は対応するディスクキャッシュ１５４領域にデータを格納し、ＰＡ１４０に対して完了報告を送信する（１５１２、１５１３、１５１４）。ＳＡ１５０より完了報告を受信したＰＡ１４０はホスト１００に対して、完了報告を送信する（１５１５、１５１６）。

【００６９】

一方、要求データがディスクキャッシュ１５４上に無かった場合（キャッシュミス）、ＳＡ１５０はキャッシュ管理情報２０３を更新し、ディスクキャッシュ１５４上に要求データを格納する領域を確保する（１５０６）。以降の動作フローは、キャッシュヒットした場合と同様である（１５０７から１５１６）。

以上のようにして、ホスト１００からのライト要求に対して、ホスト１００からのライトデータをディスクキャッシュ１５４へ格納する。

【００７０】

図１６は、ライトアフタ処理２５７の処理フロー図である。

この処理は、ＳＡ１５０におけるライトコマンド処理２６２の結果、ディスクキャッシュ１５４に格納されたライトデータをディスク装置１５７もしくは外部ストレージ１８０へ書き出す処理である。

ディスクキャッシュ１５４上に保持されたライトデータは、キャッシュ管理情報２０３で管理される。通常、ライトデータやディスクからリードしたデータはなるべく古いものからディスクキャッシュ１５４より追い出されるように、キューなどで管理されている。その様な従来の方法によって、管理されたデータの中からライトアフタ対象データを決定する（１６０１）。対象データが、物理デバイスへのライトデータの場合は、ディスク装置１５７の該当する領域にデータを書き込んだ後に、対象データのディスクキャッシュ１５

４領域を解放する（１６０３、１６０４）。対象データが外部デバイスへのライトデータの場合、ＳＡ１５０はＰＡ１４０を介して外部ストレージ１８０に対してデータを書き込み（１６０５から１６１０）、対象データのディスクキャッシュ１５４領域を解放する（１６０４）。

【００７１】

第２の実施形態：

次に図１７乃至図１９を参照して、第２の実施形態を説明する。

この例は、各ＳＡ１５０のダーティ属性を持つキャッシュ量の変動に応じて、論理デバイス定義時に外部デバイスに割り当てられたＳＡ１５０を、同デバイスへのホスト１００からの入出力要求を受け付けつつ、別のＳＡ１５０に変更するものである。第２の実施形態では、第１の実施形態と実質的に同様のハードウェアおよびソフトウェア構成を前提としているので、以下、第1の実施形態との差異について説明する。

【００７２】

図１７に、第２の実施形態における上位論理デバイス管理情報２０４の構成を示す。

図６に示した第1の実施形態のものに比べて、図１７の例では、デバイスアクセスモード６０７、切替進捗ポインタ６０８および変更先ＳＡ番号／下位論理デバイス番号６０９が追加される。

デバイスアクセスモード６０７には、上位論理デバイスへの入出力要求の処理形態を示す「正常」もしくは「割当変更中」の値をとる。特定ＳＡ１５０の下位論理デバイスに割り当てられ、通常の入出力処理が行われる上位論理デバイスには「正常」の値が設定され、実体が外部デバイスの上位論理デバイスで、かつ外部デバイスが特定SAの下位論理デバイスから別のＳＡ１５０の下位論理デバイスへ割り当て変更中である上位論理デバイスに対しては「割当変更中」が設定される。

【００７３】

切替進捗ポインタ６０８は、上位論理デバイスが「割当変更中」に使用され、上位論理デバイスのうち、外部デバイスの割当変更処理が未完了な部分の先頭アドレスを示す情報である。切替進捗ポインタ６０８は、後述するＳＡ１５０のキャッシュミス化処理の進捗に応じて更新される。変更先ＳＡ番号／下位論理デバイス番号６０９は上位論理デバイスが「割当変更中」に使用され、外部デバイスの割り当て変更先のＳＡ番号および下位論理デバイス番号を保持する。

【００７４】

図１８は、第２の実施形態における下位論理デバイス管理情報２０１の構成を示す。

図８に示した第1の実施形態のものに比べて、図１８の例では、切替進捗ポインタ８０６が追加される。その他のエントリの内容は上記上位論理デバイス管理情報２０１と同じである。ＳＡ１５０によるキャッシュミス化処理の進捗に応じて更新するが、２重に管理することで、ＭＡ１６０で保持する上位論理デバイス管理情報２０３への更新頻度を抑えることも可能となる。

上位論理デバイス管理情報２０３の更新時には、全ＰＡ１４０のメモリ１４３へ更新情報の取り込みが必要となるため、性能を考慮すれば更新頻度の抑制は有効である。ただし、その場合、ＰＡ１４０でのリードコマンド処理２６１、ライトコマンド処理２６２で古い切替進捗ポインタ６０８を参照することになるため、既にキャッシュミス化を終えた領域についてもＰＡ１４０からＳＡ１５０へ入出力要求が転送されることになる。

よって、ＳＡ１５０のコマンド処理２５４で、キャッシュミス化処理を終えた領域へのライト要求についてはライトアフタで処理せず、即物理／外部デバイスへ書き込むなどこの処理で利用したディスクキャッシュ１５４を処理後速やかに解放する論理が必要となる。

【００７５】

図１９は外部デバイス割り当て変更処理の処理フローである。

この処理は、管理端末１９０又は管理サーバ１１０からの外部デバイス割り当て変更指示を受け、外部デバイスに割り当てられたＳＡ１５０、即ち外部デバイスの入出力要求を処理するＳＡ１５０を変更する処理である。

【００７６】
ＳＴ１９０は、ＳＡキャッシュ使用量情報２２２および外部デバイスキャッシュ使用量情報２２４を参照し、全ＳＡ１５０に対するダーティ量情報３０２および全外部デバイスに対するダーティ量情報４０２を得て、管理端末１９０又は管理サーバ１１０に対してそれらを提示し、外部デバイス割り当て変更の指示を受け付ける（１９０１、１９０２）。管理端末１９０又は管理サーバ１１０はＳＡ１５０のキャッシュ使用量が均等化されるように、変更対象とする外部デバイスおよび変更先のＳＡ１５０を指定することができる。ここで、指定しないことも可能である。
変更対象とする外部デバイスおよび変更先のＳＡ１５０を指定された場合は（１９０３）、外部デバイスの割り当て変更によるキャッシュ使用量の均等化の効果を示すために、管理端末１９０又は管理サーバ１１０に対して割り当て変更後の予測ＣＡキャッシュ使用量情報を提示する（１９０６）。予測ＳＡキャッシュ使用量情報は、変更対象とする外部デバイスの現在のダーティ量を、変更元のＳＡ１５０のダーティ量から減じ、変更先のＳＡ１５０のダーティ量に加えることにより算出する。その後、ＳＴ１９０は、割当変更時間情報５０２により指定された時間帯に、ＭＡ１６０に対して割り当て変更の指示を行う（１９０７）。この指示には、変更対象の外部デバイス番号、変更先ＳＡ番号の情報が含まれる。割り当て変更の指示を受け付けたＭＡ１６０は外部デバイスのデバイスアクセスモードを「正常」から「割当変更中」に変更する（１９０８）。ＭＡ１６０は変更先ＳＡ１５０に対して割当変更指示を発行する（１９０９）。割当変更指示を受け取った変更先ＳＡ１５０では、指示された外部デバイスに対して下位論理デバイスを登録し、ＭＡ１６０に対して完了報告を送信する（１９１０）。
【００７７】
変更先ＳＡ１５０から完了報告を受け取ったＭＡ１６０は、変更元ＳＡ１５０に対して割当変更指示を送信する（１９１１）。割当変更指示を受け取った変更元ＳＡ１５０は、対応する下位論理デバイスの全データについてディスクキャッシュ１５４を検索してミス化、即ち、物理／外部デバイスへ未更新のデータは対応デバイスへの書き出し後、更新済みのデータについては、即そのデータに割り当てられたキャッシュ領域を解放する。検索はキャッシュ管理情報２０３を用いて、下位論理デバイスの先頭から順次行い、検索およびミス化処理が終わった部分については、上位／下位論理デバイス管理情報の切替進捗ポインタを進めることで進捗を管理する。全領域についてミス化処理が完了したら、その旨をＭＡ１６０へ報告し、ＭＡ１６０でデバイスアクセスモード６０７を「正常」へ変更して外部デバイス割当変更処理を完了する（１９１２、１９１３）。
【００７８】
変更対象とする外部デバイスおよび変更先のＳＡ１５０が指定されなかった場合は（１９０３）、ＳＴ１９０が各ＳＡ１５０のキャッシュ使用量が均等化するように、変更対象とする外部デバイスおよび変更先のＳＡ１５０を選択する（１９０４）。変更先のＳＡ１５０としては最もダーティ量の少ないＳＡ１５０とする。
キャッシュ使用量の均等化の指標としては、例えば、各ＳＡ１５０のキャッシュ使用量の標準偏差を用いることができる。外部デバイス管理情報を用いて、ストレージ１３０に割り当てられた各外部デバイスを変更対象候補として１つずつ選ぶ。選択された外部デバイスのダーティ量および各ＳＡ１５０のダーティ量が分かっているので、外部デバイスを割り当てるＳＡ１５０を変更した場合の各ＳＡ１５０の予測ダーティ量を算出でき、各ＳＡ１５０の予測ダーティ量を用いて予測標準偏差を求めることができる。変更対象として選択する外部デバイスは、予測標準偏差が最小となる外部デバイスとする。ただし、均等化されない場合、即ち、予測標準偏差が現在の標準偏差以上の値になる場合は、外部デバイスの割り当て変更は実施しない（１９０５）。以降の処理は、変更対象とする外部デバイスおよび変更先のＳＡ１５０を指定された場合と同様である。
【００７９】
第２の実施形態においては、第1の実施形態に比べてリードコマンド処理２６１、ライトコマンド処理２６２が一部変わる。

即ち、図１４のリードコマンド処理２６１の１４０２において、ＰＡ１４０は受信したコマンドの解析を行い、要求データに対応する上位論理デバイス番号を割り出した後、デバイスアクセスモードをチェックする。デバイスアクセスモードが「正常」の場合には、上位論理デバイス管理情報から対応するＳＡ番号、下位論理デバイス番号を算出する。その後、ＳＡ番号に対応するＳＡ１５０にリード要求を転送する（１４０３）。
デバイスアクセスモードが「割当変更中」の場合には、切替進捗ポインタを参照し、その要求のアクセス領域が切替進捗ポインタの前後どちらにあるかを判定する。切替進捗ポインタよりも前、即ち変更先ＳＡ１５０へ切替済みの領域については、変更先ＳＡ１５０を要求先ＳＡ１５０として決定する。もし、そうでない場合、即ち変更先ＳＡ１５０への切替が完了していない領域に関しては、変更元ＳＡ１５０を要求先ＳＡ１５０として決定する。図１５のステップ１５０２においても、同様の判定を行い要求先ＳＡ１５０を決定する。

【００８０】

第３の実施形態：

次に図２０乃至図２３を参照して、第３の実施形態を説明する。
この例は、第１のストレージである外部ストレージ１８０aのデバイス（外部デバイス）に対して、ストレージ１３０が非同期リモートコピー機能を適用し、ストレージ１３０が有する非同期リモートコピー機能によって、外部ストレージ１８０aのデバイス内に格納されるデータを他の外部ストレージ１８０bのデバイスにコピーすることを前提とする。この際、本実施形態においては、サイドファイル属性を持つキャッシュのキャッシュ使用量に基づき、外部ストレージ１８０aのデバイスを自身の論理デバイスとして管理し、外部ストレージ１８０aのデバイスについてリモートコピー処理を実行するストレージ１３０内のストレージ制御部を選択する。

【００８１】

第３の実施形態を説明する前に、まず、リモートコピー機能について説明する。
リモートコピーは、正サイトにあるストレージシステムのデバイス（正デバイス）のバックアップを副サイトにあるストレージシステムのデバイス（副デバイス）にリアルタイムで作成する機能である。正デバイスのコピーを遠隔地にある副デバイスに保持することにより、テロや災害等によるデータの損失を最小限に抑えることが可能である。リモートコピーの操作には、形成コピーと更新コピーの２種類がある。形成コピーはホスト１００からのリード／ライトとは別に正デバイスと副デバイスのデバイスペアを同期させる操作である。形成コピーにより、デバイスペアが同期された後の正デバイスに対するライトは、更新コピーにより副デバイスに対しても適用され、同期状態が維持される。リモートコピーは、更新コピーの方式により、同期リモートコピーと非同期リモートコピーに大別される。

【００８２】

非同期リモートコピーは、特に、ホスト１００からライトデータを受信しキャッシュメモリに書き込んだ時点で、ホスト１００に対して完了報告を送信する。非同期リモートコピー対象のライトデータのうち、正ストレージシステムのキャッシュ上に保持され、副ストレージシステムに未送信のデータはサイドファイルと呼ばれる。サイドファイルは副ストレージシステムに対して送信されるまで、キャッシュメモリ上に保持し続ける必要がある。しかし、サイドファイルの副ストーレジシステムへの送信速度とホスト１００からのライトの頻度によっては、サイドファイルがキャッシュメモリ上に大量に滞留し、新たなリード／ライトに対してキャッシュメモリを割り当てることができなくなる可能性がある。

【００８３】

サイドファイルは、正ストレージシステムのＳＡ１５０の制御メモリ１５５上のキャッシュ管理情報２０３により、サイドファイル属性を持つセグメントとして管理される。サイドファイル量はサイドファイル属性を持つセグメントの個数により算出される。キャッシュ管理情報２０３としては、ライトデータがディスクキャッシュ１５４に書き込まれた

時間を示すタイムスタンプ情報、ライトデータに対応する上位論理デバイス番号、ライトデータの上位論理デバイス内での位置情報、ライトデータのサイズのような制御情報が格納される。タイムスタンプ情報は、サイドファイルが副ストレージシステムに送信される際の付加情報としてライトデータと共に送信される。互いに依存関係のある複数正デバイスにより構成されるデバイスグループに対するライトを副デバイスに適用する際には、デバイスグループ内のデータの整合性を維持するために、ライトの順序を保障する必要がある。副ストレージシステムにおいては、タイムスタンプ情報に基づきライトの順序を維持しながら、複数の副デバイスに対する更新が行われる。

【００８４】

第三の実施形態においては、例えば第一の外部ストレージ１８０a内の外部デバイスと対応付けられているストレージ１３０内の上位論理デバイスと、遠隔地にある第2の外部ストレージ１８０ｂ内の論理デバイスとの間で非同期リモートコピーを行う。即ち、ホスト１００からストレージ１３０内の上位論理デバイスへの書き込み要求があった際には、上位論理デバイスを管理しているストレージ１３０内のストレージ制御部は、キャッシュにライトデータを書き込んだ後ホスト１００に完了報告を返送し、その後ライトデータを第2の外部ストレージ１８０ｂに転送すると共に、前記上位論理デバイスと対応付けられている第１の外部ストレージ１８０a内の外部デバイスにもライトデータ書き込む。

【００８５】

尚、本実施形態においては、リモートコピー先のストレージは第2の外部ストレージ１８０ｂであるものと仮定して説明するが、リモートコピー先のストレージは外部ストレージ１８０に限られるものではない。即ち、リモートコピー先のストレージは、ストレージ１３０のストレージ制御部によってストレージ１３０の上位論理デバイスとして管理される外部デバイスを有する外部ストレージであっても良いが、ストレージ１３０とネットワークを介して接続されるストレージであれば外部ストレージに限られることはなく、ホスト１００からストレージ１３０を介することなくアクセスされ得るデバイスを有するストレージであっても良い。

第３の実施形態も、第１の実施形態と実質的に同様のハードウェアおよびソフトウェア構成を前提としているので、以下、第1の実施形態との差異について説明する。

【００８６】

図２０はSAキャッシュ使用量情報２２２を示す図である。

この図２０は、図３に比べて、ストレージ１３０内のＳＡ１５０のサイドファイル量に関する情報を管理するサイドファイル量情報３０３が追加されている。サイドファイル量情報の値は、ＳＡ１５０において、集計時間情報５０１に含まれる時間帯に、一定の時間間隔でＳＡ１５０に対応するサイドファイルセグメントカウンタを参照し、それらの平均値を求めることにより算出される。

【００８７】

図２１は外部デバイスキャッシュ使用量情報２２４を示す図である。

この図２１では、図４に比べて、ストレージ１３０内の外部デバイスの特定の時間帯におけるサイドファイル量を管理するサイドファイル量情報４０３が追加されている。サイドファイル量情報４０３の値は、ＳＡ１５０において、集計時間情報５０１に含まれる時間帯に、一定の時間間隔で当該外部デバイスに対応するサイドファイルセグメントカウンタを参照し、それらの平均値を求めることにより算出される。

【００８８】

第３の実施形態においては、論理デバイス定義処理２５５が部分的に変わる。

第１の実施形態においては、図１２のステップ１２０６において、最もダーティ量の少ないＳＡ１５０を割り当て対象として選択したが、第３の実施形態においては、最もサイドファイル量の少ないＳＡ１５０を割り当て対象とする。この処理により、外部デバイスを、ストレージ１３０内のＳＡ１５０のキャッシュ使用量の均等化を考慮して、いずれかのＳＡ１５０へ割り当てることが可能となる。

【００８９】

外部デバイス定義処理２５３およびＬＵパス定義処理２５２は、第１の実施形態と同様であり、上記の論理デバイス定義処理２５５と合わせた３つの処理によりホスト１００からのアクセスが可能となる。

【００９０】

次に、第１の外部ストレージ１８０ａ内の外部デバイスを実体とする上位論理デバイスである正デバイスと、第２の外部ストレージ１８０ｂ内の論理デバイスである副デバイスの間で非同期リモートコピーを適用する場合における、リードコマンド処理／ライトコマンド処理について述べる。尚、リードコマンド処理２６１は第１の実施形態と同様なので、その説明を省略する。

【００９１】

次に、図２２を参照して、ライトコマンド処理２６２について説明する。

第３の実施形態におけるライトコマンド処理は、図１５におけるライトコマンド処理２６２とほぼ同じであるが、ステップ２２１７から２２１９が追加されている。

ステップ２２１７においては、ライトデータに対応するデバイスが非同期リモートコピー機能を使用中かを判定し、使用中なら、ディスクキャッシュ１５４上にライトデータ用のサイドファイル領域を確保し、管理メモリ上のキャッシュ管理情報を更新する（２２１８）。

ステップ２２１９においては、ライトデータをディスクキャッシュ１５４上のサイドファイル領域に格納し、キャッシュ管理情報にタイムスタンプ情報等の制御情報を格納する。

【００９２】

図２３はサイドファイル送信処理を説明するための図である。

この処理は、ＳＡ１５０におけるライトコマンド処理２６２の結果、ディスクキャッシュ１５４上のサイドファイル領域に格納されたライトデータを第２の外部ストレージ１８０ｂに書き出す処理である。

ＳＡ１５０は、管理メモリ１５５上のキャッシュ管理情報２０３から、サイドファイルが書き込まれた順番にサイドファイルを送信するように、サイドファイルを決定する（２３０１）。ＳＡ１５０は該当するＰＡ１４０にサイドファイルデータをタイムスタンプ情報等の制御情報と共に送信する（２３０２）。サイドファイルデータと制御情報を受信したＰＡ１４０は、該当する外部ストレージ１８０に対して、サイドファイルデータと制御情報を送信する（２３０３）。その後、ＳＡ１５０は対象データのサイドファイル領域を解放する（２３０４）。

【００９３】

第４の実施形態：

この例は、非同期リモートコピー機能が適用されている外部デバイスに対して割り当てられたＳＡ１５０をその後の各ＳＡ１５０のサイドファイル属性を持つキャッシュ量に応じて、そのデバイスへのホスト１００からの入出力要求を受け付けつつ、変更するものである。

第４の実施形態も、第１の実施形態と実質的に同様のハードウェアおよびソフトウェア構成を前提としているので、以下、第1の実施形態との差異について説明する。

第４の実施形態においては、第２の実施形態で説明した、図１７の上位論理デバイス管理情報２０３および図１８の下位論理デバイス管理情報２０１を用いる。

【００９４】

図２４は外部デバイス割当変更処理を説明する図である。

管理端末１９０又は管理サーバ１１０からの外部デバイス割り当て変更指示を受け、外部デバイスに割り当てられたＳＡ１５０、即ち、当該外部デバイスの入出力要求を処理するＳＡ１５０を変更する処理である。図２４は第２の実施形態で説明した図１９とほぼ同様であるが、ステップ２４０１、２４０２、２４０４、２４０５、２４０６が変わっている。

ステップ１９０１においてはダーティ量情報３０２およびダーティ量情報４０２を参照していたのに対して、ステップ２４０１においては、ＳＴ１９０はSAキャッシュ使用量情

報２２２および外部デバイスキャッシュ使用量情報２２４を参照して、全ＳＡ１５０に対するサイドファイル量情報３０３および全外部デバイスに対するサイドファイル量情報４０３を得る。この変更に伴い、ステップ２４０２、２４０４、２４０５、２４０６においては、ダーティ量情報をサイドファイル量情報に置き換えて、ステップ１９０２、１９０４、１９０５、１９０６と同様の処理を行う。

【００９５】

第４の実施形態におけるリードコマンド処理２６１は、第２の実施形態で説明したものと同様である。第４の実施形態におけるライトコマンド処理２６２としては、第３の実施形態で説明したライトコマンド処理２６２を一部変更したものを用いる。変更個所は図２２のステップ２２０２であり、変更内容は第２の実施形態で説明したリードコマンド処理２６１に対する変更と同様である。

【００９６】

第５の実施形態：

上記した実施形態では、ストレージ１３０はＰＡ１４０とＳＡ１５０とＭＡ１６０が相互結合網により結合された構成を持つクラスタ構成ストレージを例として挙げた。しかし本発明は、上記したクラスタ構成ストレージに限定されず、他の構成からなるクラスタ構成ストレージに関しても適用される。

【００９７】

図２５は、本発明の他の構成による計算機システム例を示す。

この例において、ストレージ２５３０は、複数のストレージノード２５５０とストレージノード２５５０間でのデータ連携を行うための相互結合網２５７０から構成されるクラスタ構成ストレージである。各ストレージノード２５５０は、通常のストレージと同様に、１または複数のディスク装置２５５７と、ディスクキャッシュ２５５４と、制御プロセッサ２５５２と、メモリ２５５３と、ポート２５５１から構成される。

本構成例においては、前述したストレージ２５３０が提供する高機能処理を担当する部位であるＳＡ１５０と入出力要求を振り分けるＰＡ１４０が、共にストレージノード２５５０として構成されている点が第１から第４の実施形態とは異なる。この例においても、複数のストレージノード２５５０はそれぞれ相互結合網２５７０により接続されているため、任意のストレージノード２５５０に対して外部デバイスを割り当てることができる。なお、この例において、第１乃至第４の実施形態における構成管理部１６０に相当する処理はストレージ２５３０内のいずれかのストレージノード２５５０が行う。

【００９８】

管理サーバ２５１０は、第１乃至第４の実施形態における管理端末１９０の役割も兼ねており、ＩＰネットワーク２５７５により計算機システム内の各機器とインターフェース２５１６を介して交信し、計算機システム内の各機器から構成情報、リソース利用率、性能監視情報などを収集する。またそれらの情報をディスプレイ２５１５に表示してストレージ管理者に提示する。さらに入力装置から入力され受信した運用・保守に関する指示を各機器に送信する。また、第１乃至第４の実施形態におけるＳＡ１５０と同様に、ストレージノード２５５０はディスクキャッシュ２５５４上のダーティ量やサイドファイル量に関する情報を採取する。管理サーバ２５１０は各ストレージノード２５５０から、ダーティ量やサイドファイル量に関する情報を収集し、メモリ２５１２上に保持する。さらに、管理サーバ２５１０は外部ストレージ２５８０を含めた計算機システム全体の管理も行う。

【００９９】

ファイバチャネルスイッチ２５２０は、ホスト２５００のポート２５０７およびストレージ２５３０のポート２５５１に加え、外部ストレージのポート２５８１にも接続されている。その他の機器は第１乃至第４の実施形態で説明したものと同等の役割を果たす。

【０１００】

次に、第５の実施形態と上記した第１乃至第４の実施形態との処理の類似点および相違点について説明する。この第５の実施形態では、ハードウェア構成が前述の第１乃至第４

の実施形態とは異なるので、それによりリードライト処理も相違してくる。
具体的には、あるストレージノード２５５０がホスト２５００からの入出力要求を受領した際、その入出力要求がそのストレージノード２５５０にて処理可能な入出力要求であった場合、そのストレージノード２５５０において処理し、他のストレージノード２５５０において処理すべき入出力要求であった場合には、相互結合網２５７０を介して他のストレージノード２５５０に入出力要求を転送する。さらに、第１乃至第４の実施形態におけるＳＡ１５０と同様に、ストレージノード２５５０においては、外部ストレージ２５８０のデバイスはストレージ２５３０の論理デバイスとして定義され、ホスト２５００からの入出力要求に対して、ストレージノード内部のディスク装置に対するアクセスか、外部ストレージに対するアクセスかを識別して、要求を振り分ける。

【０１０１】
この実施形態においては、第１の実施形態のように、外部デバイスの処理を行うストレージノード２５５０を選択する際に、ストレージノード２５５０のダーティ量を用いて、ストレージシステム全体としてのキャッシュ使用量の均等化を行うことができる。
具体的な処理内容は、ストレージノード２５５０がＰＡ１４０、ＳＡ１５０の両方の処理を行う点を除いては、第１の実施形態で説明したものと同様である。ただし、以下の点で処理が異なる。

【０１０２】
第一に、予め、全てのストレージノード２５５０から外部ストレージ２５８０にアクセスできるように、ファイバチャネルスイッチ２５２０のゾーニング設定を変更しておくものとする。これは、どのストレージノードも外部ストレージへの経路としての役割を果たすことができることを意味し、第２から第４の実施形態の構成例においても同様である。
第二に、論理デバイス定義処理２５５において、ＳＡ１５０としての役割を果たすストレージノード２５５０が同時にＰＡ１４０としての役割も果たすように設定を行う。これは第３の実施形態の構成例においても同様である。

【０１０３】
さらに、この例においては、第２の実施形態のように、各ストレージノード２５５０のダーティ属性を持つキャッシュ量の変動に応じて、論理デバイス定義時に外部デバイスに割り当てられたストレージノード２５５０を、同デバイスへのホスト２５００からの入出力要求を受け付けつつ，別のストレージノード２５５０に変更することができる。
具体的な処理内容は、ストレージノード２５５０がＰＡ１４０、ＳＡ１５０の両方の処理を行う点を除いては、第２の実施形態で説明したものと同様である。ただし、次の点で処理が異なる。即ち、外部デバイス割当変更処理２５６において、ＳＡ１５０としての役割を果たすストレージノード２５５０が同時にＰＡ１４０としての役割も果たすように設定を行う。これは第４の実施形態の構成例においても同様である。

【０１０４】
また、第３の実施形態のように、外部デバイスを実体とする、ストレージノード２５５０内の論理デバイスに対して非同期リモートコピーを適用する場合においても、各ストレージノード２５５０におけるサイドファイル量に関する情報を用いて、非同期リモートコピーの処理を行うストレージノード２５５０を適切に選択することにより、各ストレージノード２５５０のサイドファイル量の均等化を行うことが可能である。具体的な処理内容は、ストレージノード２５５０がＰＡ１４０、ＳＡ１５０の両方の処理を行う点を除いては、第３の実施形態で説明したものと同様である。

【０１０５】
さらに、本構成例においては、第４の実施形態のように、非同期リモートコピー機能が適用されている外部デバイスに対して割り当てられたストレージノード２５５０をその後の各ストレージノード２５５０のサイドファイル属性を持つキャッシュ量に応じて、同デバイスへのホスト２５００からの入出力要求を受け付けつつ、変更することができる。具体的な処理内容は、ストレージノード２５５０がＰＡ１４０、ＳＡ１５０の両方の処理を行う点を除いては、第４の実施形態で説明したものと同様である。

この他にも本発明の趣旨を逸脱しない範囲で種々変形して実施し得る。

【０１０６】

尚、上記第５の実施形態に関する態様を整理すると、例えば以下のように列挙できる。

（１）ホストとのインターフェースを有し、ホストから入出力要求に従って読み書きされるデータを格納するキャッシュメモリ及びディスク装置を有すると共に、第１のストレージシステムとのインターフェースを有して第１のストレージシステムに対して入出力要求を行う複数のストレージノードと、ストレージノードの間を互いに接続する相互結合網と、ストレージノードと通信を行う管理サーバとを有するストレージシステムにおいて、第１のストレージシステムのディスク装置及びストレージノードに保持するディスク装置をストレージシステムが持つディスク装置としてホストへ提示する手段と、ホストから受け付けた入出力要求のアクセス対象であるストレージシステムのディスク装置がストレージノードのディスク装置もしくは第１のストレージシステムのディスク装置である場合、入出力要求をストレージノードで処理する手段と、ストレージノードのキャッシュメモリ上にあって、ストレージノードのディスク装置と第１のストレージシステムのディスク装置に未反映なライトデータの合計量である第１のダーティ量情報を取得する手段と、管理サーバに第１のダーティ量情報を提示し、第１のストレージシステムのディスク装置の処理を行う該ストレージノードの指定を受け付ける手段と、を有するストレージシステム。

（２）上記（１）のストレージシステムにおいて、第１のストレージシステムのディスク装置の処理を行う第２のストレージシステム内のストレージノードが管理サーバで指定されなかった場合、第１のストレージシステムのディスク装置の処理を行うストレージノードを、第１のダーティ量情報を用いて決定するストレージシステム。

（３）上記（２）のストレージシステムにおいて、第１のストレージシステムのディスク装置の処理を行うストレージノードを、第１のストレージノードから第２のストレージノードに変更する切替処理手段を有するストレージシステム。

（４）上記（３）のストレージシステムにおいて、ストレージノードのキャッシュメモリ上にあって第１のストレージノードのディスク装置のデータを検索し、第１のストレージシステムのディスク装置へ未反映なデータは第１のストレージシステムのディスク装置へ書き込んでキャッシュメモリ領域を解放し、反映済みのデータはキャッシュメモリ領域を解放するストレージシステム。

（５）上記（４）のストレージシステムにおいて、第３のストレージノードが第１のストレージノードのディスク装置へのホストからの入出力要求に対して、第１のストレージノードのディスク装置で切替処理が終わった部分への入出力要求は第２のストレージノードへ、切替処理が終わっていない部分への要求は第１のストレージノードへ振り分けるストレージシステム。

（６）上記（３）のストレージシステムにおいて、ストレージノードのキャッシュメモリ上にあって、第１のストレージシステムのディスク装置に未反映なライトデータの量である第２のダーティ量情報を取得する手段を有し、上記管理サーバへ第１のダーティ量情報および第２のダーティ量情報を提示し、上記切替処理の対象となる第１のストレージシステムのディスク装置および切替処理の切替先である第２のストレージノードの指定を受け付ける手段を有するストレージシステム。

（７）上記（６）のストレージシステムにおいて、上記切替処理の対象となる第１のストレージシステムのディスク装置および切替処理の切替先である第２のストレージノードの指定が無かった場合、第１のダーティ量および第２のダーティ量を用いて、切替処理の対象となる第１のストレージシステムのディスク装置および切替処理の切替先である第２のストレージノードを決定するストレージシステム。

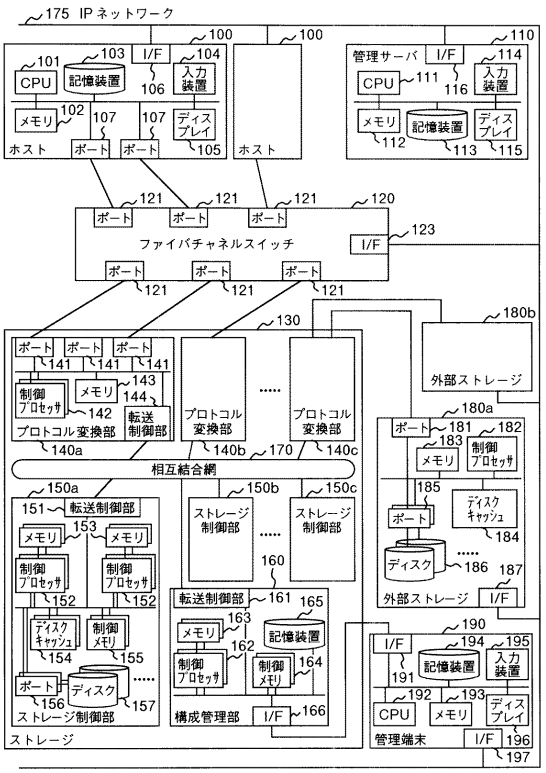
（８）上記（１）のストレージシステムにおいて、第３のストレージシステムとのインターフェースを有し、第３のストレージシステムに対してコピー処理を行う第４のストレージノードを含み、第３のストレージシステムのディスク装置にストレージノードにおけるディスク装置の複製を作成し、このストレージノードにおけるディスク装置に対するライトデータを逐次そのストレージノードのキャッシュメモリに格納し、書き込みデータを第

３のストレージシステムのディスク装置に送信する機能を有するストレージシステムであって、前記ストレージノードのキャッシュメモリ上にあって、第３のストレージシステムに未送信なライトデータの合計量である第１のサイドファイル量情報を取得する手段を有し、管理サーバに第１のサイドファイル量情報を提示し、第１のストレージシステムのディスク装置の処理を行うストレージノードの指定をから受け付けるストレージシステム。
（９）上記（８）のストレージシステムにおいて、第１のストレージシステムのディスク装置の処理を行うストレージノードの指定が無かった場合、第１のストレージシステムのディスク装置の処理を行うストレージノードを、第１のサイドファイル量情報を用いて決定するストレージシステム。
（１０）上記（８）のストレージシステムにおいて、第１のストレージシステムのディスク装置をホストに提示する処理を行うストレージノードを、第１のストレージノードから第２のストレージノードに変更するストレージシステム。
（１１）上記（１０）のストレージシステムにおいて、ストレージノードのキャッシュメモリ上にあって、第３のストレージシステムのディスク装置に未送信なデータ量である第２のサイドファイル量情報を取得する手段を有し、管理サーバに第１のサイドファイル量情報および前記第２のサイドファイル量情報を提示し、切替処理の対象となる第１のストレージシステムのディスク装置および切替処理の切替先である第２のストレージノードの指定を受け付けるストレージシステム。
（１２）上記（１１）のストレージシステムにおいて、切替処理の対象となる第１のストレージシステムのディスク装置および切替処理の切替先である第２のストレージノードの指定が無かった場合、第１のサイドファイル量情報および第２のサイドファイル量情報を用いて、切替処理の対象となる第１のストレージシステムのディスク装置および切替処理の切替先である第２のストレージノードを決定するストレージシステム。
【図面の簡単な説明】
【０１０７】
【図１】本発明の第１の実施形態による計算機システムのハードウェア構成を示す図である。
【図２】第１の実施形態におけるソフトウェアの構成を示す図である。
【図３】第１の実施形態におけるストレージ制御部のキャッシュ使用量情報の構成を示す図である。
【図４】第１の実施形態における外部デバイスキャッシュ使用量情報の構成を示す図である。
【図５】第１の実施形態における管理端末制御情報の構成を示す図である。
【図６】第１の実施形態における上位論理デバイス管理情報の構成を示す図である。
【図７】第１の実施形態におけるＬＵパス管理情報の構成を示す図である。
【図８】第１の実施形態における下位論理デバイス管理情報の構成を示す図である。
【図９】第１の実施形態における物理デバイス管理情報の構成を示す図である。
【図１０】第１の実施形態における外部デバイス管理情報の構成を示す図である。
【図１１】第１の実施形態における外部デバイス定義処理の処理フロー図である。
【図１２】第１の実施形態における論理デバイス定義処理の処理フロー図である。
【図１３】第１の実施形態におけるＬＵパス定義処理の処理フロー図である。
【図１４】第１の実施形態におけるリードコマンド処理の処理フロー図である。
【図１５】第１の実施形態におけるライトコマンド処理の処理フロー図である。
【図１６】第１の実施形態におけるライトアフタ処理の処理フロー図である。
【図１７】第２の実施形態における上位論理デバイス管理情報の構成を示す図である。
【図１８】第２の実施形態における上位論理デバイス管理情報の構成を示す図である。
【図１９】第２の実施形態における外部デバイス割当変更処理の処理フロー図である。
【図２０】第３実施形態におけるストレージ制御部キャッシュ使用量情報の構成を示す図である。
【図２１】第３の実施形態における外部デバイスキャッシュ使用量情報の構成を示す図で

ある。
【図２２】第３の実施形態におけるライトコマンド処理の処理フロー図である。
【図２３】第３の実施形態におけるサイドファイル送信処理の処理フロー図である。
【図２４】第４の実施形態における外部デバイス割当変更処理の処理フロー図である。
【図２５】第５の実施形態における計算機システムの構成を示す図である。
【符号の説明】
【０１０８】
１００…ホスト、　１１０…管理サーバ、
１２０…ファイバチャネルスイッチ、　１３０…ストレージ、
１４０…プロトコル変換部、　１５０…ストレージ制御部、
１５４…ディスクキャッシュ、　１５７…ディスク
１６０…構成管理部、　１７０…相互結合網、
１８０…外部ストレージ、　１９０…管理端末、
２０１…下位論理デバイス管理情報、　２０２…物理デバイス管理情報、
２０３…キャッシュ管理情報、　２０４…上位論理デバイス管理情報、
２０５…外部デバイス管理情報、　２０６…ＬＵパス管理情報、
２５２…ＬＵパス定義処理、　２５３…外部デバイス定義処理、
２５５…論理デバイス定義処理、　２５６…外部デバイス割当変更処理、
２５７…ライトアフタ処理、　２６１…リードコマンド処理、
２６２…ライトコマンド処理、　２６３…サイドファイル送信処理
２５００…ホスト、　２５１０…管理サーバ
２５３０…ストレージ、　２５５０…ストレージノード
２５５４…ディスクキャッシュ、　２５８０…外部ストレージ
２５７０…相互結合網。

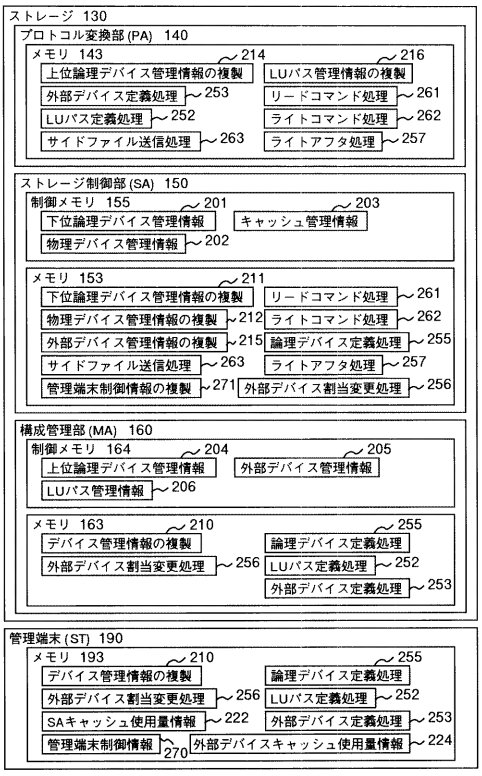
【図１】

図　1



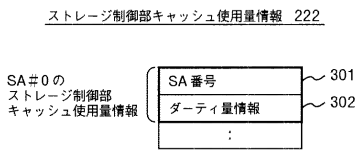
【図２】

図　2



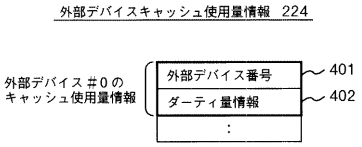
【図３】

図　3



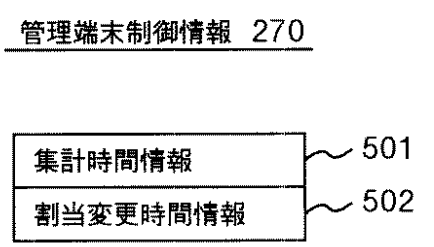
【図４】

図　4



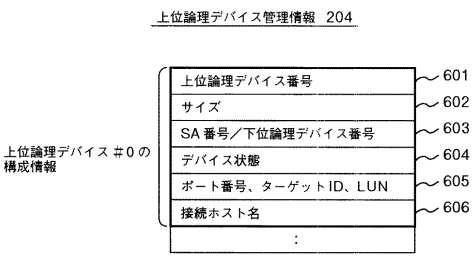
【図５】

図　5



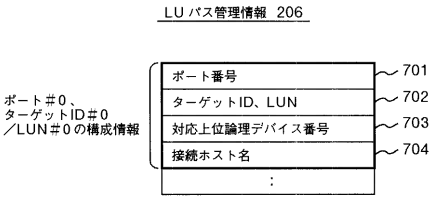
【図６】

図　6



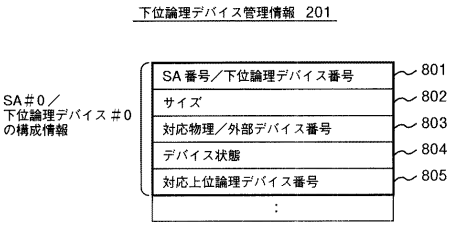
【図７】

図　7



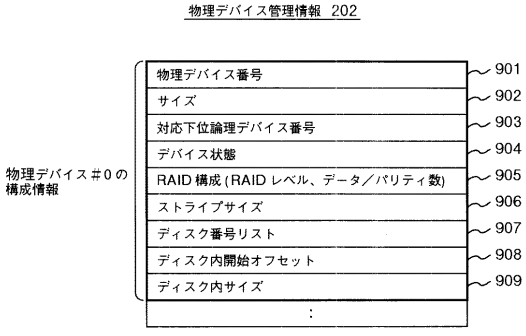
【図８】

図　8



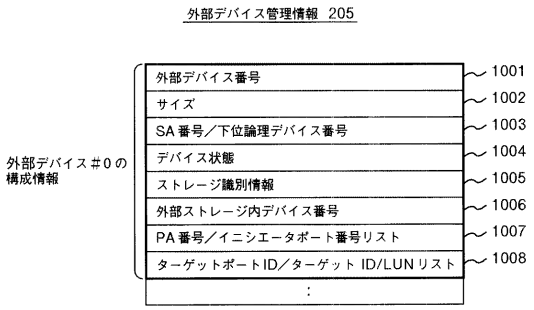
【図９】

図　9



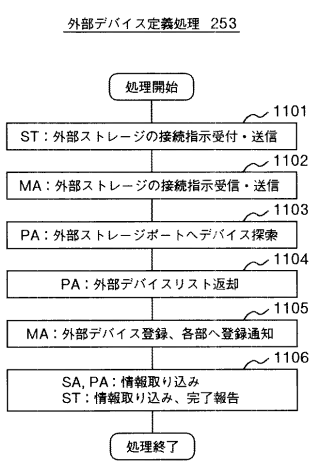
【図１０】

図　10



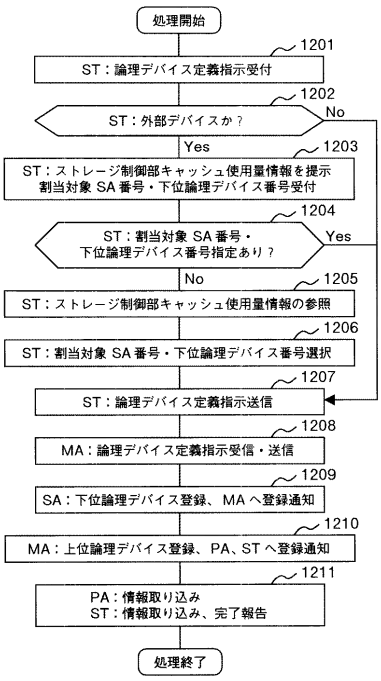
【図１１】

図　１１



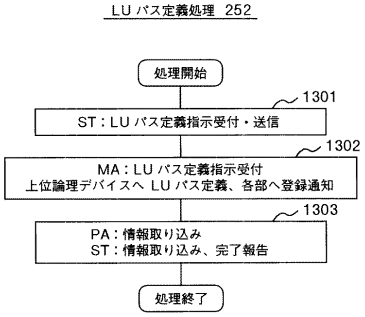
【図１２】

図　１２



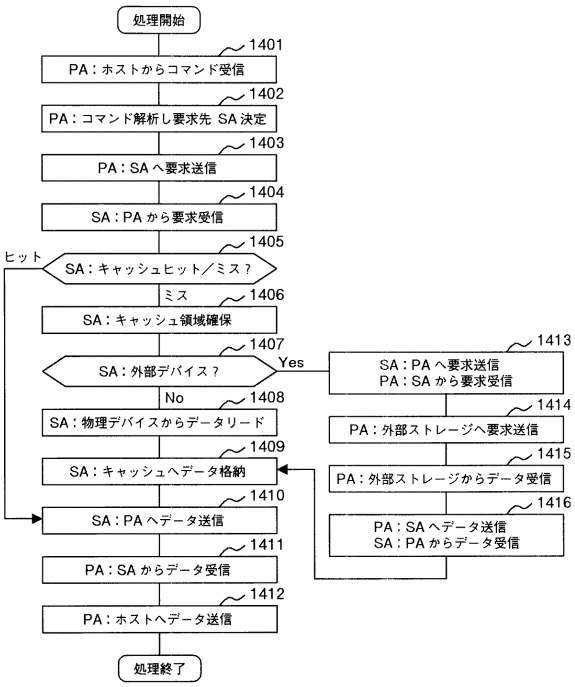
【図１３】

図　１３



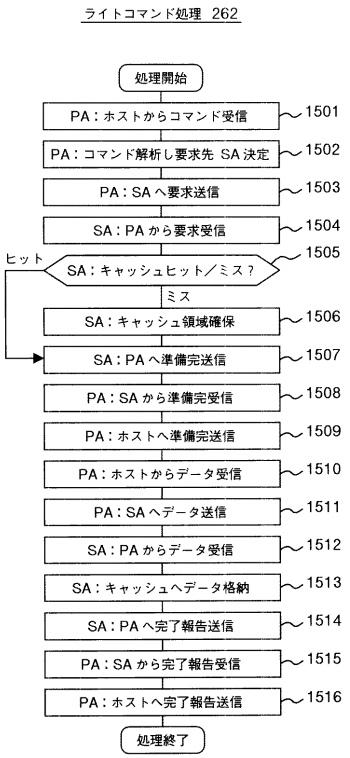
【図１４】

図　１４



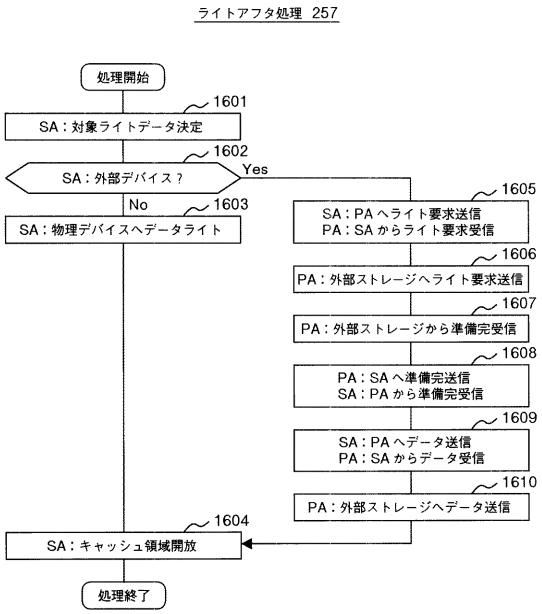
【図１５】

図　１５



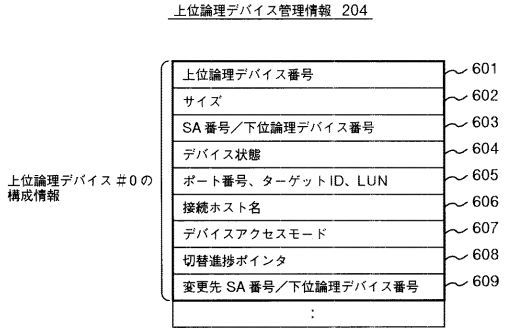
【図１６】

図　１６



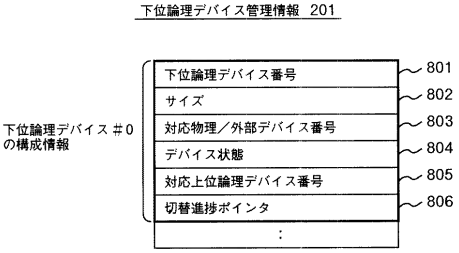
【図１７】

図　１７



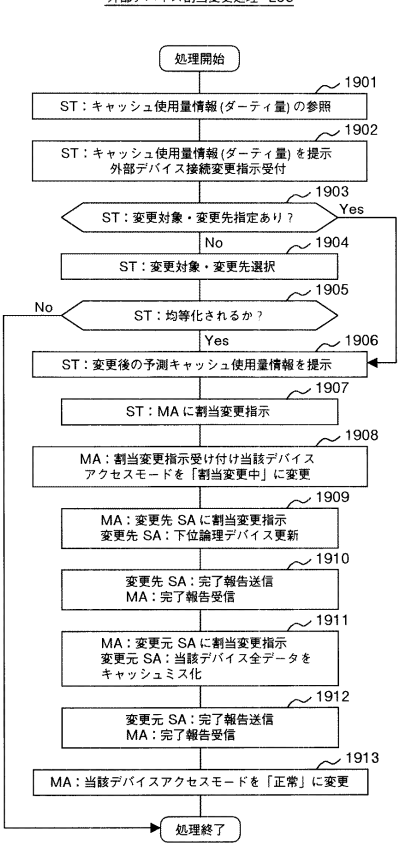
【図１８】

図　１８



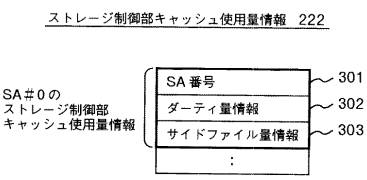
【図１９】

図　１９



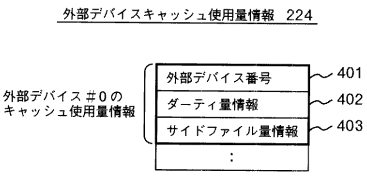
【図２０】

図　２０



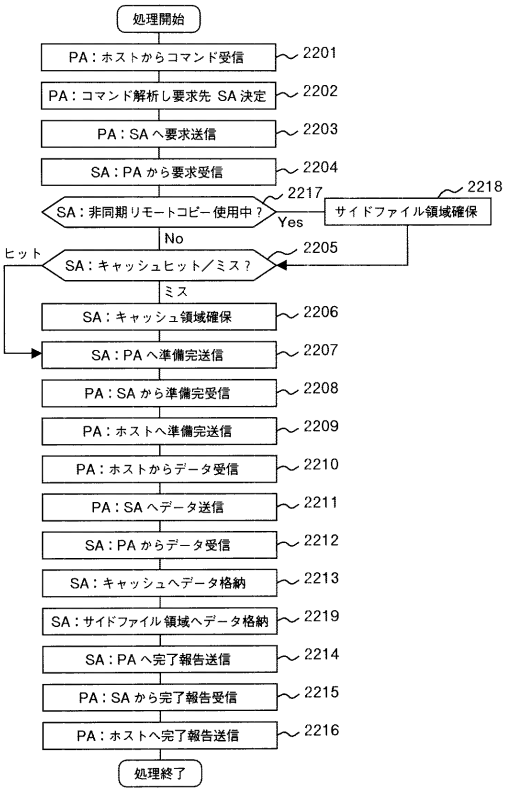
【図２１】

図　２１



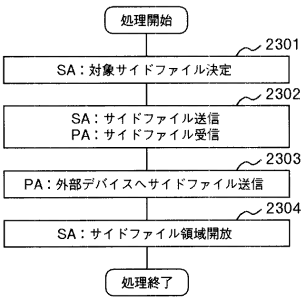
【図２２】

図　２２



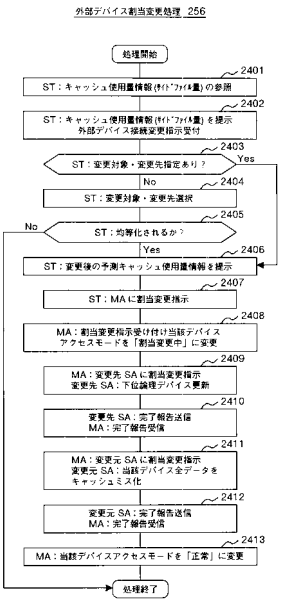
【図２３】

図　２３



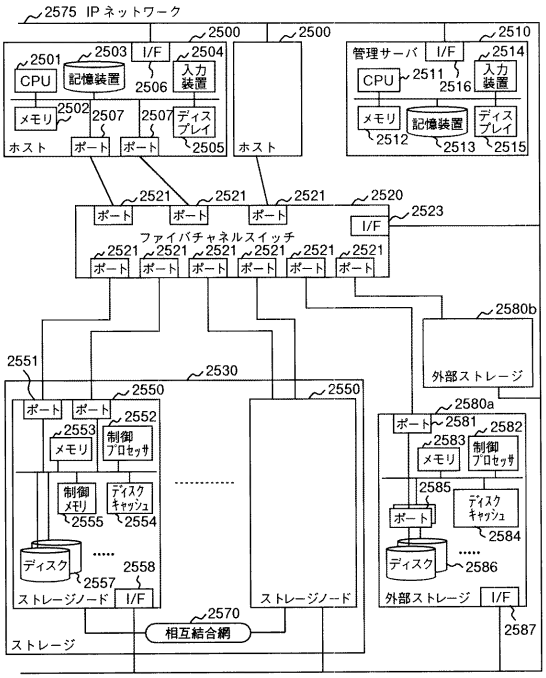
【図２４】

図　２４



【図２５】

図　２５


2575 IPネットワーク
2500
2500
2510
2501
2503
I/F
2504
管理サーバ
I/F
2514
CPU
記憶装置
2506
入力装置
CPU
2511
2516
入力装置
メモリ
2502
2507
2507
ディスプレイ
メモリ
記憶装置
ディスプレイ
ホスト
ポート
ポート
2505
ホスト
2512
2513
2515
2521
2521
2521
2520
ポート
ポート
ポート
I/F
2523
ファイバチャネルスイッチ
2521
2521
2521
2521
2521
2521
ポート
ポート
ポート
ポート
ポート
ポート
2580b
外部ストレージ
2551
2530
2550
2550
2580a
ポート
ポート
2553
2552
ポート
2581
2582
メモリ
制御プロセッサ
メモリ
2583
制御プロセッサ
制御メモリ
ディスクキャッシュ
2585
ディスクキャッシュ
2555
2554
ポート
2584
ディスク
2558
ディスク
2586
2557
ストレージノード
I/F
ストレージノード
外部ストレージ
I/F
2587
2570
相互結合網
ストレージ

フロントページの続き

(51)Int.Cl. FI
G06F 13/10 340A

(72)発明者 山本 康友
神奈川県川崎市麻生区王禅寺１０９９番地 株式会社日立製作所 システム開発研究所内
(72)発明者 藤本 和久
神奈川県川崎市麻生区王禅寺１０９９番地 株式会社日立製作所 システム開発研究所内

審査官 清木 泰

(56)参考文献 特開２００５－０５５９７０（ＪＰ，Ａ）
特開２００３－１３１９４４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ１２／０８－１２／１２
Ｇ０６Ｆ ３／０６－ ３／０８
Ｇ０６Ｆ１３／１０－１３／１４
Ｇ０６Ｆ１２／００－１２／００，５４９