(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4338068号
(P4338068)
(45)発行日　平成21年9月30日(2009.9.30)　(24)登録日　平成21年7月10日(2009.7.10)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 304B
G06F 3/06 302A
G06F 3/06 302B

請求項の数 12　(全 30 頁)

(21)出願番号　特願2002-77499(P2002-77499)
(22)出願日　平成14年3月20日(2002.3.20)
(65)公開番号　特開2003-280824(P2003-280824A)
(43)公開日　平成15年10月2日(2003.10.2)
審査請求日　平成16年9月24日(2004.9.24)

前置審査

(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　100093861
弁理士　大賀　眞司
(72)発明者　藤本　和久
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内

審査官　梅景　篤

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】
ディスク装置と、
個々が、ホストコンピュータとのインターフェースを有するチャネルインターフェース部及び前記ディスク装置とのインターフェースを有するディスクインターフェース部とを有する複数のディスク制御クラスタと、
前記ディスク装置に対しリード／ライトされるデータと、前記データの転送に関する制御情報と、前記ディスク装置の管理情報と、前記複数のディスク制御クラスタの負荷情報若しくは障害情報を含む管理情報を格納するグローバル情報制御部と、
前記複数のディスク制御クラスタを相互接続する相互結合網と、
前記複数のディスク制御クラスタの個々が有する前記チャネルインターフェース部の相互を接続するスイッチと、を有し、
前記複数のディスク制御クラスタにおいて、前記ホストコンピュータからのデータのリード／ライト要求に対し、前記チャネルインターフェース部は前記ホストコンピュータとのインターフェースと前記グローバル情報制御部との間のデータ転送を実行し、前記ディスクインターフェース部は前記ディスク装置と前記グローバル情報制御部との間のデータ転送を実行し、
前記スイッチは、定期的に若しくは必要に応じて、前記グローバル情報制御部に格納された前記負荷情報若しくは障害情報を含む管理情報の複製を格納するメモリを有し、
前記グローバル情報制御部は前記相互結合網と前記スイッチに接続され、

前記スイッチは、自身が有する複数のポート間の接続切換え情報を前記メモリの複製された管理情報に基づいて更新し、当該接続切換え情報に基づいて前記複数のポート間の接続切換えを行うことを特徴とするストレージシステム。

【請求項２】

個々が、ホストコンピュータとのインターフェースを有するチャネルインターフェース部と、ディスク装置とのインターフェースを有するディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータ、前記データの転送に関する制御情報及び前記ディスク装置の管理情報を格納するローカル共有メモリ部とを有し、前記ホストコンピュータからのデータのリード／ライト要求に対し、前記チャネルインターフェース部は前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、

前記ディスクインターフェース部と接続されるディスク装置と、

前記複数のディスク制御クラスタ負荷情報若しくは障害情報を含む管理情報を格納するグローバル情報制御部と、

前記複数のディスク制御クラスタを相互接続する相互結合網と、

前記複数のディスク制御クラスタの個々が有するチャネルインターフェース部を相互に接続するスイッチと、を有し、

前記スイッチは、定期的に若しくは必要に応じて、負荷情報若しくは障害情報を含む管理情報の複製を格納するメモリを有し、自身が有する複数のポート間の接続切換え情報を前記メモリに格納された前記管理情報に基づいて更新し、当該更新後の接続切換え情報に基づいて前記複数のポート間の接続切換えを行うことを特徴とするストレージシステム。

【請求項３】

前記チャネルインターフェース部、前記ディスクインターフェース部及び前記ローカル共有メモリ部とを相互に接続する接続部を前記複数のディスク制御クラスタの個々が有し、

前記複数のディスク制御クラスタの個々が有する前記制御部は、他のディスク制御クラスタが有する接続部と前記相互結合網を介して接続され、

前記グローバル情報制御部は前記相互結合網及び前記スイッチに接続され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】

前記複数のディスク制御クラスタの個々が有する前記ローカル共有メモリ部が前記相互結合網を介して相互に接続され、

前記グローバル情報制御部は前記相互結合網と前記スイッチに接続され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５】

ディスク装置と、

個々が、ホストコンピュータとのインターフェースを有するチャネルインターフェース部と、ディスク装置とのインターフェースを有するディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータを格納する第１のメモリと前記チャネルインターフェース部及び前記ディスクインターフェース部と前記第１のメモリとの間のデータ転送に関する制御情報及び前記ディスク装置の管理情報を格納する第２のメモリとを有するローカル共有メモリ部とを有し、前記ホストコンピュータからのデータのリード／ライト要求に対し、前記チャネルインターフェース部は前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部内の前記第１のメモリとの間のデータ転送を実行し、前記ディスクインターフェース部は前記ディスク装置と前記ローカル共有メモリ部内の前記第１のメモリとの間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、

前記複数のディスク制御クラスタの負荷情報若しくは障害情報を含む管理情報を格納す

るグローバル情報制御部と、
前記複数のディスク制御クラスタを相互に接続する第１、第２の相互結合網と、
前記複数のディスク制御クラスタの個々が有する前記チャネルインターフェース部を相互に接続するスイッチと、を有し、
前記スイッチは、定期的に若しくは必要に応じて、前記グローバル情報制御部に格納された前記負荷情報若しくは障害情報を含む管理情報の複製を格納するメモリを有し、自身が有する複数のポート間の接続切換え情報を前記メモリに格納された前記管理情報に基づいて更新し、当該更新後の接続切換え情報に基づいて前記複数のポート間の接続切換えを行うことを特徴とするストレージシステム。
【請求項６】
前記複数のディスク制御クラスタの個々が有する前記チャネルインターフェース部と前記ディスクインターフェース部が、他のディスク制御クラスタが有する前記チャネルインターフェース及び前記ディスクインターフェース部と前記第１の相互結合網を介して接続され、
前記グローバル情報制御部は前記第１の相互結合網と前記スイッチに接続されることを特徴とする請求項５記載のストレージシステム。
【請求項７】
ディスク装置と、
個々が、ホストコンピュータとのインターフェースを有するチャネルインターフェース部と、前記ディスク装置とのインターフェースを有するディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報と前記ディスク装置の管理情報を格納するローカル共有メモリ部と、前記チャネルインターフェース部、前記ディスクインターフェース部及び前記ローカル共有メモリ部とを相互に接続する第１の接続部とを有し、前記ホストコンピュータからのデータのリード／ライト要求に対し、前記チャネルインターフェース部は前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、
前記複数のディスク制御クラスタの負荷情報若しくは障害情報を含む管理情報を格納するグローバル情報制御部と、
前記複数のディスク制御クラスタの個々が有する前記チャネルインターフェース部を相互に接続するスイッチと、
前記複数のディスク制御クラスタの相互を接続する第２の接続部を有し、
前記グローバル情報制御部は、前記第２の接続部と前記スイッチに接続パスで接続され、
前記複数のディスク制御クラスタの個々が有する前記第１の接続部は前記第２の接続部に接続パスで接続され、
前記スイッチは、定期的に若しくは必要に応じて、前記グローバル情報制御部に格納された前記負荷情報若しくは障害情報を含む管理情報の複製を格納するメモリを有し、自身が有する複数のポート間の接続切換え情報を前記メモリに格納された前記管理情報に基づいて更新し、当該更新後の接続切換え情報に基づいて前記複数のポート間の接続切換えを行うことを特徴とするストレージシステム。
【請求項８】
前記複数のディスク制御クラスタの個々が有するローカル共有メモリ部は、自身が属するディスク制御クラスタ内の各部位の負荷情報及び障害情報、並びに前記自身が属するディスク制御クラスタが管理する記憶領域情報を格納しており、
前記グローバル情報制御部は第２のメモリを有し、前記第２のメモリは前記複数のディスク制御クラスタの前記負荷情報、障害情報、記憶領域情報を格納しており、
前記グローバル情報制御部は、前記ローカル共有メモリ部内の前記負荷情報、障害情報、記憶領域情報をある時間間隔で参照し、前記第２のメモリ内の該負荷情報、障害情報、

記憶領域情報を更新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９】
前記グローバル情報制御部は第２のメモリを有し、前記第２のメモリは前記複数のディスク制御クラスタの各部位の障害情報、負荷情報を格納しており、
前記複数のディスク制御クラスタの各々は、自身内のある部位で障害が発生した時点、あるいはある部位の負荷が予め設定した値より高くなった時点で、前記グローバル情報制御部の前記第２のメモリ内の前記障害情報あるいは負荷情報を更新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１０】
前記グローバル情報制御部は第２のメモリを有し、前記第２のメモリは前記複数のディスク制御クラスタが管理する記憶領域情報を格納しており、
前記複数のディスク制御クラスタの各々は、自身内のある記憶領域のデータを、他のディスク制御クラスタ内のある記憶領域にコピーあるいは移動した時点で、前記グローバル情報制御部の前記第２のメモリ内の前記記憶領域情報を更新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１１】
前記グローバル情報制御部は、該メモリ内の前記負荷情報、障害情報、あるいは記憶領域情報が更新された時点で、前記負荷情報、障害情報、あるいは記憶領域情報を前記スイッチ内のメモリにコピーすることを特徴とする請求項８、９及び１０に記載のうちいずれか一つのストレージシステム。
【請求項１２】
前記スイッチは、前記接続切換え情報としての接続切換えテーブルを有し、前記管理情報、あるいは、記憶領域情報が更新された時点で、前記接続切換えテーブルに登録された情報を更新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。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、データを複数のディスク装置に格納するストレージシステムとそれを構成するディスク制御クラスタに関する。
【０００２】
【従来の技術】
半導体記憶装置を記憶媒体とするコンピュータの主記憶のＩ／Ｏ性能に比べて、磁気ディスクを記憶媒体とするディスクサブシステム（以下「サブシステム」という。）のＩ／Ｏ性能は３～４桁程度小さく、従来からこの差を縮めること、すなわちサブシステムのＩ／Ｏ性能を向上させる努力がなされている。サブシステムのＩ／Ｏ性能を向上させるための１つの方法として、複数のディスク装置でサブシステムを構成し、データを複数のディスク装置に格納する、いわゆるディスクと呼ばれるシステムが知られている。
【０００３】
例えば、従来技術では、図２に示すようにホストコンピュータ３とディスク制御装置４との間のデータ転送を実行する複数のチャネルＩＦ部１１と、ディスク装置２とディスク制御装置４間のデータ転送を実行する複数のディスクＩＦ部１６と、ディスク装置２のデータとディスク制御装置４に関する制御情報（例えば、ディスク制御装置４内のデータ転送制御に関する情報、ディスク装置２に格納するデータの管理情報）を格納する共有メモリ部２０とを備え、１つのディスク制御装置４内において、共有メモリ部２０は全てのチャネルＩＦ部１１及びディスクＩＦ部１６からアクセス可能な構成となっている。このディスク制御装置４では、チャネルＩＦ部１１及びディスクＩＦ部１６と共有メモリ部２０との間は相互結合網３０で接続される。
【０００４】
チャネルＩＦ部１１は、ホストコンピュータ３と接続するためのインターフェース及びホストコンピュータ３に対する入出力を制御するマイクロプロセッサ（図示せず）を有して

いる。また、ディスクＩＦ部１６は、ディスク装置２と接続するためのインターフェース及びディスク装置２に対する入出力を制御するマイクロプロセッサ（図示せず）を有している。また、ディスクＩＦ部１６は、ＲＡＩＤ機能の実行も行う。
インターネットの普及等により企業で扱うデータは爆発的に増大しており、データセンタ等では一台のディスク制御装置で扱えるデータ量以上のデータを記憶する必要がある。このため、図４に示すようにディスク制御装置４を複数台設置し、それらのホストコンピュータ３とのインターフェースをＳＡＮスイッチ５を介して、ホストコンピュータ３に接続していた。
また、データ量の増大に伴いＳＡＮスイッチ５に接続するディスク制御装置４の台数が増えると、ホストコンピュータ３とＳＡＮスイッチ５を含めたシステム全体（このシステムをストレージ・エリア・ネットワーク（ＳＡＮ）と呼ぶ）の管理が複雑化する。それに対処するため、ＳＡＮスイッチ５にＳＡＮアプライアンス６を接続し、ＳＡＮアプライアンス６においてＳＡＮスイッチ５に繋がる全てのディスク制御装置４が管理するデータのディレクトリサービスを行い、ホストコンピュータ３に対して複数のディスク制御装置４を１つのストレージシステムに見せる処理、言い換えると、個々のディスク制御装置４が提供する記憶領域を１つの大きな記憶領域の固まりに見せ、その中から必要な量の記憶領域をホストコンピュータ３に割当てるという処理を行っていた。
また、特開２００１－２５６００３に開示されている他の従来技術では、図２３に示すように、１つのストレージシステム１は複数のディスク制御クラスタ１－１乃至１－ｎから構成される。各ディスク制御クラスタは、ホストコンピュータ３と該ディスク制御クラスタの間のデータ転送を実行する複数のチャネルＩＦ部１１と、ディスク装置２と該ディスク制御クラスタの間のデータ転送を実行するディスクＩＦ部１６と、共有メモリ部２５を有し、チャネルＩＦ部１１及びディスクＩＦ部１６と共有メモリ部２５の間は複数のディスク制御クラスタ１－１乃至１－ｎに跨る相互結合網３１を介して接続されている。共有メモリ部２５にストレージシステムの制御情報が格納されている。共有メモリ部２５は、相互結合網３１を介して、全てのチャネルＩＦ部１１及びディスクＩＦ部１２からアクセス可能な構成となっており、共有メモリ部２５を介して制御情報のやりとりをすることにより、複数のディスク制御クラスタが１つのストレージシステムとして動作していた。
【０００５】
【発明が解決しようとする課題】
銀行、証券、電話会社等に代表される大企業では、従来各所に分散していたコンピュータ及びストレージを、データセンターの中に集中化してコンピュータシステム及びストレージシステム構成することにより、コンピュータシステム及びストレージシステムの運用、保守、管理に要する費用を削減する傾向にある。
【０００６】
このような傾向の中で、大型／ハイエンドのディスク制御装置には、数百台以上のホストコンピュータへ接続するためのチャネルインターフェースのサポート(コネクティビティ)、数百テラバイト以上の記憶容量のサポートが要求されている。
【０００７】
一方、近年のオープン市場の拡大、ストレージ・エリア・ネットワーク(SAN)の普及により、大型／ハイエンドのディスク制御装置と同様の高機能・高信頼性を備えた小規模構成(小型筐体)のディスク制御装置への要求が高まっている。
【０００８】
前者の要求に対しては、従来の大型／ハイエンドのディスク制御装置を複数接続して超大規模なストレージシステムを構成する方法が考えられる。
【０００９】
また後者の要求に対しては、従来の大型／ハイエンドのディスク制御装置の最小構成のモデルにおいて筐体を小型化した装置を構成する方法が考えられる。また、この小型化した装置を複数台接続することにより、従来のディスク制御装置がサポートしている中規模から大規模の構成をサポートするストレージシステムを構成する方法が考えられる。

【００１０】
ストレージシステムでは、上記のように、小規模な構成から超大規模な構成まで、同一の高機能・高信頼なアーキテクチャで対応可能な、スケーラビリティのある構成のシステムが必要となってきており、そのためには、複数のディスク制御装置をクラスタリングし、１つのシステムとして運用できるストレージシステムが必要となる。
図２に示す従来技術では、複数のディスク制御装置４をＳＡＮスイッチ５を介してホストコンピュータ３に接続し、ＳＡＮアプライアンス６によりホストコンピュータ３に対して複数のディスク制御装置４を１つのストレージシステムに見せていた。
しかし、ＳＡＮアプライアンス６上で動作するソフトウェアで複数のディスク制御装置４を１つのシステムとして運用するため、従来の単体の大型のディスク制御装置に比べて信頼性、可用性が低いという問題があった。また、ＳＡＮアプライアンス６上でホストコンピュータ３から要求されたデータが存在するディスク制御装置４を検索するため、性能が低下するという問題があった。
また、あるディスク制御装置４のチャネルＩＦ部１１に障害が発生し、このディスク制御装置に繋がるディスク装置のデータにアクセスできなくなった場合、ホストコンピュータ３からそのディスク制御装置４へのアクセスをいったん停止してから、チャネルＩＦ部１１を交換する必要があるため、ホストコンピュータ３上で動作しているアプリケーションプログラムに影響を及ぼすという問題があった。
図２３に示す従来技術では、複数のディスク制御クラスタが共有メモリ部２５を介して制御情報のやり取りをすることにより、１つのストレージシステムとして動作し、スケーラビリティの高いストレージシステムを提供していた。
しかし、以下のように、ユーザにとって使い勝手が悪い点があった。すなわち、性能を上げるためには、ホストコンピュータ３に割当てる記憶領域は、そのホストコンピュータ３が繋がるディスク制御クラスタに接続されたディスク装置２上の記憶領域にする必要があった。また、ホストコンピュータ３とディスク制御クラスタの間のインターフェースの障害のためにディスク制御クラスタへアクセスすることが不可能になることを防ぐため、１つのホストコンピュータから複数のディスク制御クラスタに接続パスを繋ぐ必要があった。また、複数の接続パスを接続した場合にもホストコンピュータに通知せずに接続パスの切換えを行うことは不可能であった。
【００１１】
本発明の目的は、小規模な構成から超大規模な構成まで、同一の高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティのある構成のストレージシステムを提供することにある。
【００１２】
より具体的には、本発明の目的は、複数台のディスク制御装置をまとめて１つのシステムとしたストレージシステムにおいて高信頼・高性能で使い勝手の良いシステムを提供することにある。
【００１３】
【課題を解決するための手段】
上記目的は、ホストコンピュータとのインターフェースを有する１または複数のチャネルインターフェース部と、ディスク装置とのインターフェースを有する１または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報と前記ディスク装置の管理情報を格納するローカル共有メモリ部とを有し、前記ホストコンピュータからのデータのリード／ライト要求に対し、前記チャネルインターフェース部は前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、前記ディスク装置に対しリード／ライトされるデータと前記各ディスク制御クラスタの管理情報を格納するグローバル情報制御部と、前記複数のディスク制御クラスタを相互接続する相互結合網と、前記

複数のディスク制御クラスタ内のチャネルインターフェース部を接続するスイッチを有するストレージシステムであり、
該スイッチは前記グローバル情報制御部に格納された管理情報が複写されたメモリを有するストレージシステムによって達成される。
【００１４】
また、前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部と前記ローカル共有メモリ部とが接続された接続部と他の各ディスク制御クラスタ内の該接続部が前記相互結合網を介して接続され、前記グローバル情報制御部は前記相互結合網と前記スイッチに接続されたストレージシステムによって達成される。
【００１５】
また、前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部に前記ディスク制御クラスタ内で直接接続され、該各ディスク制御クラスタ内の該ローカル共有メモリ部と他の各ディスク制御クラスタ内の該ローカル共有メモリ部が前記相互結合網を介して接続され、前記グローバル情報制御部は前記相互結合網と前記スイッチに接続されたストレージシステムによって達成される。
また、前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部に前記ディスク制御クラスタ内で直接接続され、該各ディスク制御クラスタ内の該チャネルインターフェース部と該ディスクインターフェース部との接続部と他の各ディスク制御クラスタ内の該接続部が前記相互結合網を介して接続され、前記グローバル情報制御部は前記相互結合網と前記スイッチに接続されたストレージシステムによって達成される。
【００１６】
また、ホストコンピュータとのインターフェースを有する１または複数のチャネルインターフェース部と、ディスク装置とのインターフェースを有する１または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報と前記ディスク装置の管理情報と前記ディスク制御クラスタの管理情報を格納するグローバル情報制御部とを有し、前記ホストコンピュータからのデータのリード／ライト要求に対し、前記チャネルインターフェース部は前記ホストコンピュータとのインターフェースと前記グローバル情報制御部との間のデータ転送を実行し、前記ディスクインターフェース部は前記ディスク装置と前記グローバル情報制御部との間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、前記複数のディスク制御クラスタを相互接続する相互結合網と、前記複数のディスク制御クラスタ内のチャネルインターフェース部を接続するスイッチを有するストレージシステムであり、
該スイッチは前記グローバル情報制御部に格納された管理情報が複写されたメモリを有するストレージシステムによって達成される。
また、前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部との接続部と他の各ディスク制御クラスタ内の該接続部が前記相互結合網を介して接続され、前記グローバル情報制御部は前記相互結合網と前記スイッチに接続されたストレージシステムによって達成される。
【００１７】
また、ホストコンピュータとのインターフェースを有する１または複数のチャネルインターフェース部と、ディスク装置とのインターフェースを有する１または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータを格納する第１のメモリと、前記チャネルインターフェース部及び前記ディスクインターフェース部と前記第１のメモリとの間のデータ転送に関する制御情報及び前記ディスク装置の管理情報を格納する第２のメモリとを有するローカル共有メモリ部とを有し、前記ホストコンピュータからのデータのリード／ライト要求に対し、前記チャネルインターフェース部は前記

ホストコンピュータとのインターフェースと前記ローカル共有メモリ部内の前記第１のメモリとの間のデータ転送を実行し、前記ディスクインターフェース部は前記ディスク装置と前記ローカル共有メモリ部内の前記第１のメモリとの間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、前記各ディスク制御クラスタの管理情報を格納するグローバル情報制御部と、前記複数のディスク制御クラスタを相互接続する２つの異なる第１、第２の相互結合網と、前記複数のディスク制御クラスタ内のチャネルインターフェース部を接続するスイッチを有するストレージシステムであり、
該スイッチは前記グローバル情報制御部に格納された管理情報が複写されたメモリを有するストレージシステムによって達成される。

【００１８】
また、前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部内の第２のメモリに前記ディスク制御クラスタ内で直接接続され、該各ディスク制御クラスタ内の該チャネルインターフェース部と該ディスクインターフェース部との第１の接続部と他の各ディスク制御クラスタ内の該第１の接続部が前記第１の相互結合網を介して接続され、前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部と前記ローカル共有メモリ部内の第１のメモリとが接続された第２の接続部と他の各ディスク制御クラスタ内の該第２の接続部が前記第２の相互結合網を介して接続され、前記グローバル情報制御部は前記第１の相互結合網と前記スイッチに接続されたストレージシステムによって達成される。

【００１９】
その他、本願が開示する課題、及びその解決方法は、発明の実施形態の欄及び図面により明らかにされる。

【００２０】
【発明の実施の形態】
以下、本発明の実施例を図面を用いて説明する。
［実施例１］
図１、図３、図１２、及び図１３に、本発明の一実施例を示す。
以下の実施例において、相互結合網はスイッチを利用したものを例にして説明してあるが、相互に接続され制御情報やデータが転送されれば良いのであり、例えばバスで構成されても良い。
図１に示すように、ストレージシステム１は複数のディスク制御クラスタ１－１乃至１－ｎとフロントエンドスイッチ７から構成される。
ディスク制御クラスタ１－１は、ホストコンピュータ３とのインターフェース部（チャネルＩＦ部）１１と、ディスク装置２とのインターフェース部（ディスクＩＦ部）１６と、ローカル共有メモリ部２２を有し、チャネルＩＦ部１１及びディスクＩＦ部１６とローカル共有メモリ部２２の間は複数のディスク制御クラスタ１－１乃至１－ｎに跨る相互結合網３１を介して接続され、グローバル情報制御部２１は相互結合網３１に接続されている。すなわち、相互結合網３１を介して、全てのチャネルＩＦ部１１及びディスクＩＦ部１２から、グローバル情報制御部２１へアクセス可能な構成となっている。
ホストコンピュータ３は、フロントエンドスイッチ７を介してディスク制御クラスタに接続され、任意のホストコンピュータ３から任意のディスク制御クラスタへアクセス可能な構成となっている。
チャネルＩＦ部１１の具体的な一例を図１２に示す。
チャネルＩＦ部１１は、ホストコンピュータ３との２つのＩＦ（ホストＩＦ）２０２と、ホストコンピュータ３に対する入出力を制御する２つのマイクロプロセッサ２０１と、グローバル情報制御部２１あるいはローカル共有メモリ部２２へのアクセスを制御するアクセス制御部（メモリアクセス制御部）２０６を有し、ホストコンピュータ３とグローバル情報制御部２１あるいはローカル共有メモリ部２２間のデータ転送、及びマイクロプロセ

ッサ２０１とグローバル情報制御部２１あるいはローカル共有メモリ部２２間の制御情報の転送を実行する。マイクロプロセッサ２０１及びホストＩＦ２０２は内部バス２０５によって接続され、メモリアクセス制御部２０６は２つのホストＩＦ２０２に直接接続され、また内部バス２０５に接続されている。
ディスクＩＦ部１６の具体的な一例を図１３に示す。
ディスクＩＦ部１６は、ディスク装置２との２つのＩＦ（ドライブＩＦ）２０３と、ディスク装置２に対する入出力を制御する２つのマイクロプロセッサ２０１と、グローバル情報制御部２１あるいはローカル共有メモリ部２２へのアクセスを制御するアクセス制御部（メモリアクセス制御部）２０６を有し、ディスク装置２とグローバル情報制御部２１あるいはローカル共有メモリ部２２間のデータ転送、及びマイクロプロセッサ２０１とグローバル情報制御部２１あるいはローカル共有メモリ部２２間の制御情報の転送を実行する。マイクロプロセッサ２０１及びドライブＩＦ２０３は内部バス２０５によって接続され、メモリアクセス制御部２０６は２つのドライブＩＦ２０３に直接接続され、また、内部バス２０５に接続されている。ディスクＩＦ部１６はＲＡＩＤ機能の実行も行う。
１つのディスク制御クラスタは１つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体１つのディスク制御装置として機能を備えているものである。
ストレージシステムの具体的な一例を図３に示す。
ストレージシステム１は、２つのフロントエンドスイッチ７と、複数のディスク制御クラスタ１－１乃至１－ｎと、グローバル情報制御部２１と、２つのグローバルスイッチ（ＧＳＷ）１１５と、アクセスパス１３６と、アクセスパス１３７を有する。
グローバルスイッチ（ＧＳＷ）１１５は、グローバル情報制御部２１からのパスと複数のディスク制御クラスタからのパスを接続する接続部である。
ディスク制御クラスタ１－１乃至１－ｎは、ホストコンピュータ３との２つのチャネルＩＦ部１１と、ディスク装置２との２つのディスクＩＦ部１６と、２つのローカルスイッチ（ＬＳＷ）１１０と、２つのローカル共有メモリ部２２と、アクセスパス１３１と、アクセスパス１３２と、アクセスパス１３６を有する。
フロントエンドスイッチ７は、スイッチ７１と、スイッチ制御部７２と、メモリコントローラ７３と、メモリモジュール１０５を有する。
１つのホストコンピュータ３からは、２つのフロントエンドスイッチ７内のスイッチ７１にパスを１本ずつ接続し、スイッチ７１からは、ディスク制御クラスタ１－１乃至１－ｎの各チャネルＩＦ部１１に１本ずつパスを接続する。
【００２１】
グローバル情報制御部２１は、アクセス制御部１０１と管理機能部１０２とメモリモジュール１０５とを有し、ディスク制御クラスタ１－１乃至１－ｎの管理情報（例えば、各ディスク制御クラスタが管理する記憶領域の情報や、ディスク制御クラスタ内の各部位の負荷情報、障害情報、及び構成情報等）を格納する。ローカルスイッチ（ＬＳＷ）１１０は、チャネルＩＦ部１１からのパスと、ディスクＩＦ部１６からのパスと、ローカル共有メモリ部２２からのパスを接続する接続部である。
ローカル共有メモリ部２２は、メモリコントローラ１００とメモリモジュール１０５とを有し、ディスク制御クラスタの制御情報（例えば、チャネルＩＦ部１１及びディスクＩＦ部１６とローカル共有メモリ部２２との間のデータ転送制御に関する情報、ディスク装置２に記録するデータの管理情報等）とディスク装置２に記録するデータを格納する。
チャネルＩＦ部１１内のメモリアクセス制御部２０６には２本のアクセスパス１３１を接続し、それらを２つの異なるＬＳＷ１１０にそれぞれ接続する。
ＬＳＷ１１０には２本のアクセスパス１３２を接続し、それらを２つの異なるローカル共有メモリ部２２内のメモリコントローラ１００にそれぞれ接続する。
したがって、メモリコントローラ１００には、２つのＬＳＷ１１０から１本ずつ、計２本のアクセスパス１３２が接続される。
こうすることにより、１つのメモリアクセス制御部２０６から１つのメモリコントローラ

１００へのアクセスルートが２つとなる。
これにより、１つのアクセスパスまたはＬＳＷ１１０に障害が発生した場合でも、もう１つのアクセスルートによりローカル共有メモリ部２２へアクセスすることが可能となるため、耐障害性を向上させることができる。
ＬＳＷ１１０には、２つのチャネルＩＦ部１１と、２つのディスクＩＦ部１６からそれぞれ１本ずつ、計４本のアクセスパス１３１が接続される。
また、ＬＳＷ１１０には、２つのローカル共有メモリ部２２へのアクセスパス１３２が２本とＧＳＷ１１５へのアクセスパス１３６が１本接続される。
ＬＳＷ１１０には上記のようなアクセスパスが接続されるため、ＬＳＷ１１０内では、チャネルＩＦ部１１及びディスクＩＦ部１６からの４本のアクセスパスからの要求を、自ディスク制御クラスタ内のローカル共有メモリ部２２への２本のアクセスパスと、ＧＳＷ１１５への１本のアクセスパス１３６に振分ける機能を有する。
ＧＳＷ１１５には、各ディスク制御クラスタから１本ずつ、ディスク制御クラスタ数分の本数のアクセスパス１３６が接続される。
また、ＧＳＷ１１５には、２つのグローバル情報制御部２１内のアクセス制御部１０１へのアクセスパス１３７が１本ずつ、計２本接続される。
こうすることにより、１つのメモリアクセス制御部２０６から１つのアクセス制御部１０１へのアクセスルートが２つとなる。
これにより、１つのアクセスパスまたはＬＳＷ１１０またはＧＳＷ１１５に障害が発生した場合でも、もう１つのアクセスルートによりグローバル情報制御部２１へアクセスすることが可能となるため、耐障害性を向上させることができる。ＧＳＷ１１５を使わずに、アクセスパス１３６をメモリコントローラ１０１に直接接続しても本発明を実施する上で問題ない。そうすることにより、ＧＳＷ１１５で発生するデータ転送処理のオーバヘッドを削減することが可能となり、性能が向上する。
ＧＳＷ１１５を使わない場合、１つのメモリアクセス制御部２０６から１つのメモリコントローラ１０１へのアクセスルートを２つ確保し，耐障害性を向上するためには、ＬＳＷ１１０にアクセスパス１３６を２本接続し、それぞれを異なるメモリコントローラ１０１へ接続する。
また、ＧＳＷ１１５には、２つのフロントエンドスイッチ７内のメモリコントローラ７３へのパスが２本接続される。
フロントエンドスイッチ７内のメモリコントローラ７３には、スイッチ制御部７２とメモリモジュール１０５が接続される。
スイッチ制御部７２はスイッチ７１に接続され、メモリモジュール１０５に格納されたホストコンピュータ３のチャネルとディスク制御クラスタのチャネルの接続を示すルーティングテーブルを参照し、スイッチの切換えを制御する。
また、メモリモジュール１０５は、グローバル情報制御部２１内のメモリモジュール１０５に格納されたディスク制御クラスタ１－１乃至１－ｎ内の各部位の負荷情報、障害情報、および記憶領域情報のコピーを有し、スイッチ制御部７２は定期的あるいは必要に際してこれらの情報をもとにルーティングテーブルを変更する。
具体的な例としては、グローバル情報制御部２１は、図１８に示すようなシステム構成・稼動状況テーブル５００をそのメモリモジュール１０５に格納しており、フロントエンドスイッチ７は、テーブル５００のコピーをメモリモジュール１０５に格納している。
システム構成・稼動状況テーブル５００は、ディスク制御クラスタを識別するクラスタ番号５１１とディスク制御クラスタのチャネルを識別するチャネル番号５１２に対応する論理ボリューム番号５１３とチャネル稼動状況５１４を示す。
本実施例では、クラスタ番号５１１、チャネル番号５１２、及び論理ボリューム番号５１３を１６進数で示している。チャネル稼動状況５１４は、低負荷状態を‘０’、中負荷状態を‘１’、高負荷状態を‘２’、障害を‘３’で示している。
また、フロントエンドスイッチ７は、ホスト－論理ボリューム対応テーブル６００をそのメモリモジュール１０５に格納している。

ホスト－論理ボリューム対応テーブル６００は、個々のホストコンピュータを識別するホスト番号６１５に対応する論理ボリューム番号５１３、すなわち各ホストコンピュータに割当てられた論理ボリュームと、論理ボリューム番号５１３にアクセスするためのチャネル番号５１２と該当論理ボリュームを管理するディスク制御クラスタのクラスタ番号５１１を示す。
ホスト番号６１５は、例えば、ファイバチャネルのプロトコルで使われるワールドワイドネーム（ＷＷＮ）や、インターネットプロトコルで使われるマックアドレスやＩＰアドレスを充てることが考えられる。
スイッチ制御部７２は、ホスト－論理ボリューム対応テーブル６００（ルーティングテーブル）を参照し、スイッチ７１の切り替えを行う。
図３でＬＳＷ１１０はチャネルＩＦ部１１及びディスクＩＦ部１６とローカル共有メモリ部２２との接続部であり、ＧＳＷ１１５はディスク制御クラスタ１－１乃至１－ｎとグローバル情報制御部２１との接続部である。
図３において，ＧＳＷ１１５とグローバル情報制御部２１をボックスに実装し、フロントエンドスイッチ７を別のボックスに実装し、モジュール化した各ディスク制御クラスタ１－１乃至１－ｎといっしょに，１つの筐体の中に実装しても良い。また，各ディスク制御クラスタ１－１乃至１－ｎを別個の筐体として，距離的に離れた場所に分散しても良い。

【００２２】

図３において、ホストコンピュータ３からストレージシステム１に記録されたデータを読み出す場合の一例を述べる。
まず、ホストコンピュータ３は、ストレージシステム１に対してデータの読出し要求を発行する。
要求は、フロントエンドスイッチ７に受け取られ、フロントエンドスイッチ７内のスイッチ制御部７２は、要求パケットのヘッダを解析する。
要求パケットのヘッダには、要求を発行したホストコンピュータの番号（ホスト番号）と要求データが記録されている論理ボリューム番号が格納されており、スイッチ制御部７２はホスト－論理ボリューム対応テーブル６００を参照し、要求を発行したホストコンピュータが接続されたスイッチ７１のポートを、該当する論理ボリューム番号に割当てられているチャネル番号のスイッチ７１のポートに接続し、要求パケットをディスク制御クラスタに送る。
要求が送られたチャネルが接続されたチャネルＩＦ部１１内のマイクロプロセッサ２０１は、自ディスク制御クラスタ１－１内のローカル共有メモリ部２２にアクセスし、要求されたデータがどのディスク装置２内に格納されているかを調べる。
ローカル共有メモリ部２２には、要求データのアドレスとそのデータが実際に記録されているディスク装置２内のアドレスを対応させる変換テーブルが格納されており、要求されたデータがどのディスク装置２内に格納されているかを調べることができる。
さらに、要求を受けたチャネルＩＦ部１１内のマイクロプロセッサ２０１は、自ディスク制御クラスタ１－１内のローカル共有メモリ部２２にアクセスし、要求されたデータがローカル共有メモリ部２２内に格納されているかどうかを確認する。
ローカル共有メモリ部２２にはディスク装置２に格納するデータとともにそのデータのディレクトリ情報が格納されており、ローカル共有メモリ部２２内に要求データが存在するかどうかを確認できる。

【００２３】

それにより自ディスク制御クラスタ１－１のローカル共有メモリ部２２内にデータがあった場合は、ローカル共有メモリ部２２にアクセスしてそのデータを自身のＬＳＷ１１０を介してチャネルＩＦ部１１まで転送し、フロントエンドスイッチ７を介して、ホストコンピュータ３に送る。
自ディスク制御クラスタ１－１のローカル共有メモリ部２２内にデータが存在しなかった場合は、チャネルＩＦ部１１のマイクロプロセッサ２０１は、要求データが格納されているディスク装置２が接続されているディスクＩＦ部１６内のマイクロプロセッサ２０１に

対し、要求データを読出しローカル共有メモリ部２２に格納するというデータ要求の処理内容を示す制御情報を発行し、この制御情報の発行を受けたディスクＩＦ部１６内のマイクロプロセッサ２０１は、要求データが格納されているディスク装置２からデータを読出し、ＬＳＷ１１０を介して、自ディスク制御クラスタ１－１内のローカル共有メモリ部２２に要求データを転送し格納する。
すなわち、チャネルＩＦ部１１のマイクロプロセッサ２０１は、上記データ要求の処理内容を示す制御情報を発行し、ローカル共有メモリ部２２の制御情報領域（ジョブ制御ブロック）に格納する。
ディスクＩＦ部１６のマイクロプロセッサ２０１は、ローカル共有メモリ部２２の制御情報領域をポーリングで監視し、上記発行された制御情報が上記制御情報領域（ジョブ制御ブロック）に存在した場合は、要求データが格納されているディスク装置２からデータを読出し、ＬＳＷ１１０を介して、自ディスク制御クラスタ１－１内のローカル共有メモリ部２２に要求データを転送し格納する。
ディスクＩＦ部１６のマイクロプロセッサ２０１は、要求データをローカル共有メモリ部２２へ格納した後、前記制御情報を発行したチャネルＩＦ部１１のマイクロプロセッサ２０１にローカル共有メモリ部２２内のデータを格納したアドレスを、ローカル共有メモリ部２２内の制御情報を介して伝える。それを受けたチャネルＩＦ部１１のマイクロプロセッサ２０１は、ローカル共有メモリ部２２からデータを読出し、フロントエンドスイッチ７を介して、ホストコンピュータ３へ送る。
すなわち、ディスクＩＦ部１６のマイクロプロセッサ２０１は、要求データをローカル共有メモリ部２２へ格納した後、処理の実行の終了とデータを格納したアドレスを示す制御情報を発行し、ローカル共有メモリ部２２の制御情報領域に格納する。
前記制御情報を発行したチャネルＩＦ部１１のマイクロプロセッサ２０１は、ローカル共有メモリ部２２の制御情報領域をポーリングで監視し、ディスクＩＦ部１６のマイクロプロセッサ２０１から発行された制御情報が上記制御情報領域に存在した場合は、ローカル共有メモリ部内のデータを格納したアドレスによりローカル共有メモリ部２２からデータを読出し、チャネルＩＦ部１１まで転送し、さらにフロントエンドスイッチ７を介して、ホストコンピュータ３へ送る。
本実施例によれば、ホストコンピュータ３はフロントエンドスイッチ７のどの接続ポートに接続しても、ストレージシステム１を構成するディスク制御クラスタを意識することなく、その接続ポートにアクセス要求を発行するだけで、データの書き込み及び読出しを行うことが可能になり、ホストコンピュータ３に対して、複数台のディスク制御クラスタ１－１乃至１－ｎを１つのストレージシステムに見せることが可能となる。
そして，ディスク制御クラスタが１個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティがあり使い勝手の良い構成のストレージシステムを提供することが可能となる。
［実施例２］
図４、図５、図１２、及び図１３に、本発明の一実施例を示す。
【００２４】
図４に示すように、ディスク制御ユニット１－１乃至１－ｎとフロントエンドスイッチ７からなるストレージシステム１の構成は、チャネルＩＦ部１１及びディスクＩＦ部１６とローカル共有メモリ部２２及び相互結合網３１の間の接続構成を除いて、実施例１の図1に示す構成と同様である。
チャネルＩＦ部１１及びディスクＩＦ部１６とローカル共有メモリ部２２の間は、ディスク制御クラスタ内では直接接続されている。
また、複数のディスク制御クラスタ１－１乃至１－ｎ間では、ローカル共有メモリ部２２は相互結合網３１を介して接続されており、その相互結合網３１にグローバル情報制御部２１が接続されている。
上記のように、この実施例ではディスク制御ユニット１－１乃至１－ｎ内においてチャネ

ルＩＦ部１１及びディスクＩＦ部１６とローカル共有メモリ部２２を直接接続することにより、実施例1で示した相互結合網３１を介して接続する場合に比べ、ローカル共有メモリ部２２へのアクセス時間を短縮することが可能になる。
チャネルＩＦ部１１及びディスクＩＦ部１６の構成は、それぞれ図１２、図１３に示す実施例１の構成と同様である。
１つのディスク制御クラスタは１つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体１つのディスク制御装置として機能を備えているものである。
ストレージシステム１の具体的な一例を図５に示す。

【００２５】

ディスク制御クラスタ１－１乃至１－ｎ内の構成も、チャネルＩＦ部１１及びディスクＩＦ部１６とローカル共有メモリ部２２の間の接続構成とディスク制御クラスタ１－１乃至１－ｎとＧＳＷ１１５の接続構成を除いて、実施例1の図3に示す構成と同様である。
ストレージシステム１は、複数のディスク制御クラスタ１－１乃至１－ｎと、フロントエンドスイッチ７と、グローバル情報制御部２１と、２つのグローバルスイッチ（ＧＳＷ）１１５と、アクセスパス１３６と、アクセスパス１３７を有する。
ディスク制御クラスタ１－１乃至１－ｎは、ホストコンピュータ３との２つのチャネルＩＦ部１１と、ディスク装置２との２つのディスクＩＦ部１６と、２つのローカル共有メモリ部２２と、アクセスパス１３３と、アクセスパス１３６を有する。
チャネルＩＦ部１１内のメモリアクセス制御部２０６には２本のアクセスパス１３３を接続し、それらを２つの異なるメモリコントローラ１００にそれぞれ接続する。
したがって、メモリコントローラ１００には、２つのチャネルＩＦ部１１と２つのディスクＩＦ部１６から１本ずつ、計４本のアクセスパス１３３が接続される。また、ＧＳＷ１１５へのアクセスパス１３６が1本接続される。
メモリコントローラ１００には上記のようなアクセスパスが接続されるため、メモリコントローラ１００内では、チャネルＩＦ部１１及びディスクＩＦ部１６からの４本のアクセスパス１３３からの要求を、メモリモジュール１０５への１本のアクセスパスと、ＧＳＷ１１５への１本のアクセスパス１３６に振分ける機能を有する。
実施例1と同様にＧＳＷ１１５を使わずに、アクセスパス１３６をアクセス制御部１０１に直接接続しても本発明を実施する上で問題ない。そうすることにより、ＧＳＷ１１５で発生するデータ転送処理のオーバヘッドを削減することが可能となり、性能が向上する。
ＧＳＷ１１５を使わない場合、１つのメモリコントローラ１００から１つのアクセス制御部１０１へのアクセスルートを２つ確保し，耐障害性を向上するためには、メモリコントローラ１００にアクセスパス１３６を２本接続し、それぞれを異なるアクセス制御部１０１へ接続する。
また実施例１と同様に、図５において，ＧＳＷ１１５とグローバル情報制御部２１をボックスに実装し、フロントエンドスイッチ７を別のボックスに実装し、モジュール化した各ディスク制御クラスタ１－１乃至１－ｎといっしょに，１つの筐体の中に実装しても良い。また，各ディスク制御クラスタ１－１乃至１－ｎを別個の筐体として，距離的に離れた場所に分散しても良い。

【００２６】

本実施例において、ホストコンピュータ３からストレージシステム１へのデータの読み出し／書き込みを行う場合の、ストレージシステム１内の各部の動作は、チャネルＩＦ部１１及びディスクＩＦ部１６からローカル共有メモリ部２２へのアクセスが直接になることと、チャネルＩＦ部１１及びディスクＩＦ部１６からグローバル情報制御部２１へのアクセスがメモリコントローラ１００を介して行われることを除いて、実施例１と同様である。
本実施例によれば、ホストコンピュータ３はフロントエンドスイッチ７のどの接続ポートに接続しても、ストレージシステム１を構成するディスク制御クラスタを意識することなく、その接続ポートにアクセス要求を発行するだけで、データの書き込み及び読出しを行

うことが可能になり、ホストコンピュータ３に対して、複数台のディスク制御クラスタ１－１乃至１－ｎを１つのストレージシステムに見せることが可能となる。
そして，ディスク制御クラスタが１個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティがあり使い勝手の良い構成のストレージシステムを提供することが可能となる。
［実施例３］
図６、図７、図１２、及び図１３に、本発明の一実施例を示す。
【００２７】
図６に示すように、ディスク制御ユニット１－１乃至１－ｎとフロントエンドスイッチ７からなるストレージシステム１の構成は、チャネルＩＦ部１２及びディスクＩＦ部１７とローカル共有メモリ部２２の間の接続構成を除いて、実施例１の図1に示す構成と同様である。
チャネルＩＦ部１２及びディスクＩＦ部１７とローカル共有メモリ部２２の間は、ディスク制御クラスタ内では直接接続されている。
また、複数のディスク制御クラスタ１－１乃至１－ｎ間では、チャネルＩＦ部１２及びディスクＩＦ部１７が相互結合網３１を介して接続されており、その相互結合網３１にグローバル共有メモリ２１が接続されている。
上記のように、この実施例ではディスク制御ユニット１－１乃至１－ｎ内においてチャネルＩＦ部１２及びディスクＩＦ部１７とローカル共有メモリ部２２を直接接続することにより、実施例１で示した相互結合網３１を介して接続する場合に比べ、ローカル共有メモリ部２２へのアクセス時間を短縮することが可能になる。
チャネルＩＦ部１２及びディスクＩＦ部１７の構成は、それぞれ図１２、図１３に示すチャネルＩＦ部１１及びディスクＩＦ部１６の構成において、メモリアクセス制御部２０６のアクセスパスを４本に増やした構成となる。
ここで、４本のアクセスパスの内、２本はアクセスパス１３１、もう２本がアクセスパス１３３となる。
１つのディスク制御クラスタは１つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体１つのディスク制御装置として機能を備えているものである。
ストレージシステム１の具体的な一例を図７に示す。
【００２８】
ディスク制御クラスタ１－１乃至１－ｎ内の構成も、チャネルＩＦ部１２及びディスクＩＦ部１７とローカル共有メモリ部２２の間の接続構成を除いて、実施例１の図3に示す構成と同様である。
ストレージシステム１は、複数のディスク制御クラスタ１－１乃至１－ｎと、フロントエンドスイッチ７と、グローバル情報制御部２１と、２つのグローバルスイッチ（ＧＳＷ）１１５と、アクセスパス１３６と、アクセスパス１３７を有する。
ディスク制御クラスタ１－１乃至１－ｎは、ホストコンピュータ３との２つのチャネルＩＦ部１２と、ディスク装置２との２つのディスクＩＦ部１７と、２つのローカルスイッチ（ＬＳＷ）１１０と、２つのローカル共有メモリ部２２と、アクセスパス１３１と、アクセスパス１３３と、アクセスパス１３６を有する。
ＬＳＷ１１０は、チャネルＩＦ部１２からのパスと、ディスクＩＦ部１７からのパスを接続する接続部である。
【００２９】
チャネルＩＦ部１２及びディスクＩＦ部１７内のメモリアクセス制御部２０６には２本のアクセスパス１３３を接続し、それらを２つの異なるメモリコントローラ１００にそれぞれ接続する。したがって、メモリコントローラ１００には、２つのチャネルＩＦ部１１と２つのディスクＩＦ部１６から１本ずつ、計４本のアクセスパス１３３が接続される。
さらに、チャネルＩＦ部１２及びディスクＩＦ部１７内のメモリアクセス制御部２０６に

は２本のアクセスパス１３１を接続し、それらを２つの異なるＬＳＷ１１０にそれぞれ接続する。したがって、ＬＳＷ１１０には、２つのチャネルＩＦ部１２と２つのディスクＩＦ部１７から１本ずつ、計４本のアクセスパス１３１が接続される。また、ＧＳＷ１１５へのアクセスパス１３６が１本接続される。実施例１と同様にＧＳＷ１１５を使わずに、アクセスパス１３６をアクセス制御部１０１に直接接続しても本発明を実施する上で問題ない。そうすることにより、ＧＳＷ１１５で発生するデータ転送処理のオーバヘッドを削減することが可能となり、性能が向上する。
ＧＳＷ１１５を使わない場合、１つのＬＳＷ１１０から１つのアクセス制御部１０１へのアクセスルートを２つ確保し，耐障害性を向上するためには、ＬＳＷ１１０にアクセスパス１３６を２本接続し、それぞれを異なるメモリコントローラ１０１へ接続する。
また実施例１と同様に、図７において，ＧＳＷ１１５とグローバル情報制御部２１をボックスに実装し、フロントエンドスイッチ７を別のボックスに実装し、モジュール化した各ディスク制御クラスタ１－１乃至１－ｎといっしょに，１つの筐体の中に実装しても良い。また，各ディスク制御クラスタ１－１乃至１－ｎを別個の筐体として，距離的に離れた場所に分散しても良い。

【００３０】

本実施例において、ホストコンピュータ３からストレージシステム１へのデータの読み出し／書き込みを行う場合の、ストレージシステム１内の各部の動作は、チャネルＩＦ部１２及びディスクＩＦ部１７からローカル共有メモリ部２２へのアクセスが直接になることを除いて、実施例１と同様である。
本実施例によれば、ホストコンピュータ３はフロントエンドスイッチ７のどの接続ポートに接続しても、ストレージシステム１を構成するディスク制御クラスタを意識することなく、その接続ポートにアクセス要求を発行するだけで、データの書き込み及び読出しを行うことが可能になり、ホストコンピュータ３に対して、複数台のディスク制御クラスタ１－１乃至１－ｎを１つのストレージシステムに見せることが可能となる。
そして，ディスク制御クラスタが１個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティがあり使い勝手の良い構成のストレージシステムを提供することが可能となる。

［実施例４］

図８、図９、図１２、及び図１３に、本発明の一実施例を示す。

【００３１】

図８に示すように、ディスク制御ユニット１－１乃至１－ｎとフロントエンドスイッチ７からなるストレージシステム１の構成は、実施例１においてローカル共有メモリ２２を除いた構成である。

【００３２】

このため、実施例１の各ディスク制御ユニット１－１乃至１－ｎのローカル共有メモリ２２に格納する情報を全てグローバル情報制御部２１に格納する。

【００３３】

複数のディスク制御クラスタ１－１乃至１－ｎ間では、チャネルＩＦ部１１及びディスクＩＦ部１６が相互結合網３１を介して接続されており、その相互結合網３１にグローバル情報制御部２１が接続されている。
チャネルＩＦ部１１及びディスクＩＦ部１６の構成は、それぞれ図１２、図１３に示す実施例１の構成と同様である。
１つのディスク制御クラスタは１つの筐体として構成されるか、またはモジュールとして構成されても良い。
ストレージシステム１の具体的な一例を図９に示す。

【００３４】

ディスク制御クラスタ１－１乃至１－ｎ内の構成も、ローカル共有メモリ部２２が無いことを除いて、実施例１の図3に示す構成と同様である。

ストレージシステム１は、複数のディスク制御クラスタ１－１乃至１－ｎと、フロントエンドスイッチ７と、グローバル情報制御部２１と、２つのグローバルスイッチ（ＧＳＷ）１１５と、アクセスパス１３６と、アクセスパス１３７を有する。
ディスク制御クラスタ１－１乃至１－ｎは、ホストコンピュータ３との２つのチャネルＩＦ部１１と、ディスク装置２との２つのディスクＩＦ部１６と、２つのローカルスイッチ（ＬＳＷ）１１０と、アクセスパス１３１と、アクセスパス１３６を有する。
ＬＳＷ１１０は、チャネルＩＦ部１１からのパスとディスクＩＦ部１６からのパスを接続する接続部である。
チャネルＩＦ部１１及びディスクＩＦ部１６内のメモリアクセス制御部２０６には２本のアクセスパス１３１を接続し、それらを２つの異なるＬＳＷ１１０にそれぞれ接続する。
したがって、ＬＳＷ１１０には、２つのチャネルＩＦ部１１と２つのディスクＩＦ部１６から１本ずつ、計４本のアクセスパス１３１が接続される。また、ＧＳＷ１１５へのアクセスパス１３６が１本接続される。
実施例１と同様にＧＳＷ１１５を使わずに、アクセスパス１３６をアクセス制御部１０１に直接接続しても本発明を実施する上で問題ない。そうすることにより、ＧＳＷ１１５で発生するデータ転送処理のオーバヘッドを削減することが可能となり、性能が向上する。
ＧＳＷ１１５を使わない場合、１つのＬＳＷ１１０から１つのアクセス制御部１０１へのアクセスルートを２つ確保し，耐障害性を向上するためには、ＬＳＷ１１０にアクセスパス１３６を２本接続し、それぞれを異なるアクセス制御部１０１へ接続する。
また実施例１と同様に、図９において，ＧＳＷ１１５とグローバル情報制御部２１をボックスに実装し、フロントエンドスイッチ７を別のボックスに実装し、モジュール化した各ディスク制御クラスタ１－１乃至１－ｎといっしょに，１つの筐体の中に実装しても良い。また，各ディスク制御クラスタ１－１乃至１－ｎを別個の筐体として，距離的に離れた場所に分散しても良い。

【００３５】

本実施例において、ホストコンピュータ３からストレージシステム１へのデータの読み出し／書き込みを行う場合の、ストレージシステム１内の各部の動作は、実施例１の処理においてローカル共有メモリ２２における処理を全てグローバル情報制御部２１で行うことを除いて、実施例１と同様である。
本実施例によれば、ホストコンピュータ３はフロントエンドスイッチ７のどの接続ポートに接続しても、ストレージシステム１を構成するディスク制御クラスタを意識することなく、その接続ポートにアクセス要求を発行するだけで、データの書き込み及び読出しを行うことが可能になり、ホストコンピュータ３に対して、複数台のディスク制御クラスタ１－１乃至１－ｎを１つのストレージシステムに見せることが可能となる。
そして，ディスク制御クラスタが１個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティがあり使い勝手の良い構成のストレージシステムを提供することが可能となる。

［実施例５］

図１０に、本発明の一実施例を示す。
以下の実施例において、相互結合網はスイッチを利用したものを例にして説明してあるが、相互に接続され制御情報やデータが転送されれば良いのであり、例えばバスで構成されても良い。
図１０に示すように、ストレージシステム１は複数のディスク制御クラスタ１－１乃至１－ｎとフロントエンドスイッチ７から構成される。
ディスク制御クラスタ１－１乃至１－ｎは、ホストコンピュータ３とのインターフェース部（チャネルＩＦ部）１３と、ディスク装置２とのインターフェース部（ディスクＩＦ部）１８と、メモリ１：２５とメモリ２：２６を有するローカル共有メモリ部２２を有し、チャネルＩＦ部１３及びディスクＩＦ部１８とメモリ２の間は、ディスク制御クラスタ内部では直接接続される。

また、チャネルＩＦ部１３及びディスクＩＦ部１８は複数のディスク制御クラスタ１－１乃至１－ｎに跨る相互結合網１：３２を介して接続され、グローバル情報制御部２１は相互結合網１：３２に接続される。すなわち、相互結合網１：３２を介して、全てのチャネルＩＦ部１３及びディスクＩＦ部１８から、グローバル情報制御部２１へアクセス可能な構成となっている。
また、チャネルＩＦ部１３及びディスクＩＦ部１８とメモリ１の間は、複数のディスク制御クラスタ１－１乃至１－ｎに跨る相互結合網２：３３を介して接続される。
ホストコンピュータ３は、フロントエンドスイッチ７を介してディスク制御クラスタに接続され、任意のホストコンピュータ３から任意のディスク制御クラスタへアクセス可能な構成となっている。
チャネルＩＦ部１３の具体的な一例を図１４に示す。
チャネルＩＦ部１３は、ホストコンピュータ３との２つのＩＦ（ホストＩＦ）２０２と、ホストコンピュータ３に対する入出力を制御する２つのマイクロプロセッサ２０１と、グローバル情報制御部２１あるいはメモリ２：２６へのアクセスを制御するアクセス制御部１（メモリアクセス制御部１）２０７と、メモリ１：２５へのアクセスを制御するアクセス制御部２（メモリアクセス制御部２）２０８とを有し、ホストコンピュータ３とメモリ１間のデータ転送、及びマイクロプロセッサ２０１とグローバル情報制御部２１あるいはメモリ２間の制御情報の転送を実行する。
マイクロプロセッサ２０１及びホストＩＦ２０２は内部バス２０５によって接続され、メモリアクセス制御部１：２０７は内部バス２０５に接続され、メモリアクセス制御部２：２０８は２つのホストＩＦ２０２に直接接続され、また内部バス２０５に接続されている。
ディスクＩＦ部１８の具体的な一例を図１５に示す。
ディスクＩＦ部１８は、ディスク装置２との２つのＩＦ（ドライブＩＦ）２０３と、ディスク装置２に対する入出力を制御する２つのマイクロプロセッサ２０１と、グローバル情報制御部２１あるいはメモリ２へのアクセスを制御するアクセス制御部１（メモリアクセス制御部１）２０７と、メモリ１へのアクセスを制御するアクセス制御部２（メモリアクセス制御部２）２０８とを有し、ディスク装置２とメモリ１間のデータ転送、及びマイクロプロセッサ２０１とグローバル情報制御部２１あるいはメモリ２間の制御情報の転送を実行する。
マイクロプロセッサ２０１及びドライブＩＦ２０３は内部バス２０５によって接続され、メモリアクセス制御部１：２０７は内部バス２０５に接続され、メモリアクセス制御部２：２０８は２つのドライブＩＦ２０３に直接接続され、また内部バス２０５に接続されている。ディスクＩＦ部１８はＲＡＩＤ機能の実行も行う。
１つのディスク制御クラスタは１つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体１つのディスク制御装置として機能を備えているものである。
ストレージシステムの具体的な一例は、チャネルＩＦ部１３及びディスクＩＦ部１８とメモリ２：２６と相互結合網１：３２とグローバル共有メモリ２１との接続構成は、実施例３の図７に示す構成と同様になる。また、チャネルＩＦ部１３及びディスクＩＦ部１８とメモリ１と相互結合網２：３３との接続構成は、実施例１の図3に示す構成においてグローバル情報制御部２１を除いた構成と同様になる。
フロントエンドスイッチ７は、スイッチ７１と、スイッチ制御部７２と、メモリコントローラ７３と、メモリモジュール１０５を有する。
１つのホストコンピュータ３からは、２つのフロントエンドスイッチ７内のスイッチ７１にパスを１本ずつ接続し、スイッチ７１からは、ディスク制御クラスタ１－１乃至１－ｎの各チャネルＩＦ部１１に１本ずつパスを接続する。
【００３６】
グローバル情報制御部２１は、アクセス制御部１０１と管理機能部１０２とメモリモジュール１０５とを有し、ディスク制御クラスタ１－１乃至１－ｎの管理情報（例えば、各デ

ィスク制御クラスタが管理する記憶領域の情報や、ディスク制御クラスタ内の各部位の負荷情報、障害情報、及び構成情報等）を格納する。メモリ1は、ディスク装置２に記録するデータを一時的に格納する。また、メモリ２は、ディスク制御クラスタの制御情報（例えば、チャネルＩＦ部１３及びディスクＩＦ部１８とメモリ１：２５との間のデータ転送制御に関する情報、ディスク装置２に記録するデータの管理情報等）を格納する。
フロントエンドスイッチ７は、そのメモリモジュール１０５内に実施例１と同様の情報、テーブルを格納しており、同様の制御を行う。
また、グローバル情報制御部２１も、そのメモリモジュール１０５内に実施例１と同様の情報、テーブルを格納している。
図１０において，相互結合網１：３２を形成するディスク制御クラスタ外のスイッチ及び相互結合網２：３３を形成するディスク制御クラスタ外のスイッチとグローバル情報制御部２１をボックスに実装し、フロントエンドスイッチ７を別のボックスに実装し、モジュール化した各ディスク制御クラスタ１－１乃至１－ｎといっしょに，１つの筐体の中に実装しても良い。また，各ディスク制御クラスタ１－１乃至１－ｎを別個の筐体として，距離的に離れた場所に分散しても良い。

【００３７】

図１０において、ホストコンピュータ３からストレージシステム１に記録されたデータを読み出す場合の一例を述べる。
まず、ホストコンピュータ３は、ストレージシステム１に対してデータの読出し要求を発行する。
要求は、フロントエンドスイッチ７に受け取られ、フロントエンドスイッチ７内のスイッチ制御部７２は、要求パケットのヘッダを解析する。
要求パケットのヘッダには、要求を発行したホストコンピュータの番号（ホスト番号）と要求データが記録されている論理ボリューム番号が格納されており、スイッチ制御部７２はホスト－論理ボリューム対応テーブル６００を参照し、要求を発行したホストコンピュータが接続されたスイッチ７１のポートを、該当する論理ボリューム番号に割当てられているチャネル番号のスイッチ７１のポートに接続し、要求パケットをディスク制御クラスタに送る。
要求が送られたチャネルが接続されたチャネルＩＦ部１１内のマイクロプロセッサ２０１は、自ディスク制御クラスタ１－１内のメモリ２：２６にアクセスし、要求されたデータがどのディスク装置２内に格納されているかを調べる。メモリ2：２６には、要求データのアドレスとそのデータが実際に記録されているディスク装置２内のアドレスを対応させる変換テーブルが格納されており、要求されたデータがどのディスク装置２内に格納されているかを調べることができる。
さらに、要求を受けたチャネルＩＦ部１３内のマイクロプロセッサ２０１は、自ディスク制御クラスタ１－１内のメモリ２：２６にアクセスし、要求されたデータがメモリ１：２５に格納されているかどうかを確認する。メモリ２：２６にはメモリ１：２５に格納されているデータのディレクトリ情報が格納されており、メモリ１：２５に要求データが存在するかどうかを確認できる。

【００３８】

それにより自ディスク制御クラスタ１－１のメモリ１：２５にデータがあった場合は、そのデータをチャネルＩＦ部１３まで転送し、フロントエンドスイッチ７を介して、ホストコンピュータ３に送る。

【００３９】

自ディスク制御クラスタ１－１のメモリ１：２５にデータが存在しなかった場合は、チャネルＩＦ部１３内のマイクロプロセッサ２０１は要求データが格納されているディスク装置２が接続されているディスクＩＦ部１８内のマイクロプロセッサ２０１に対し、要求データを読出し、メモリ１：２５に格納するというデータ要求の処理内容を示す制御情報を発行し、この制御情報の発行を受けたディスクＩＦ部１８内のマイクロプロセッサ２０１は、要求データが格納されているディスク装置２からデータを読出し、自ディスク制御ク

ラスタ１－１内のメモリ１：２５に要求データを転送し格納する。
すなわち、チャネルＩＦ部１３のマイクロプロセッサ２０１は、上記データ要求の処理内容を示す制御情報を発行し、メモリ2:２６の制御情報領域（ジョブ制御ブロック）に格納する。
ディスクＩＦ部１８のマイクロプロセッサ２０１は、メモリ２：２６の制御情報領域をポーリングで監視し、上記発行された制御情報が上記制御情報領域（ジョブ制御ブロック）に存在した場合は、要求データが格納されているディスク装置２からデータを読出し、自ディスク制御クラスタ１－１内のメモリ1:２５に要求データを転送し格納する。
ディスクＩＦ部１８内のマイクロプロセッサ２０１は、要求データをメモリ１：２５へ格納した後、制御情報を発行したチャネルＩＦ部１３内のマイクロプロセッサ２０１に、メモリ１：２５内のデータを格納したアドレスを，メモリ２：２６内の制御情報を介して伝える。それを受けたチャネルＩＦ部１３内のマイクロプロセッサ２０１は、メモリ１：２５からデータを読出し、フロントエンドスイッチ７を介して、ホストコンピュータ３へ送る。
すなわち、ディスクＩＦ部１８のマイクロプロセッサ２０１は、要求データをメモリ1:２５へ格納した後、処理の実行の終了とデータを格納したアドレスを示す制御情報を発行し、ローカル共有メモリ部２２の制御情報領域に格納する。
前記制御情報を発行したチャネルＩＦ部１３のマイクロプロセッサ２０１は、メモリ2:２６の制御情報領域をポーリングで監視し、ディスクＩＦ部１８のマイクロプロセッサ２０１から発行された制御情報が上記制御情報領域に存在した場合、メモリ1:２５内のデータを格納したアドレスによりメモリ1:２５からデータを読出し、チャネルＩＦ部１３まで転送し、さらにフロントエンドスイッチ７を介して、ホストコンピュータ３へ送る。
【００４０】
制御情報とデータはデータ長が数千倍異なるため、１回のデータ転送時間がかなり異なる。このため、同じ相互結合網及びメモリを用いた場合、両者が互いの転送を妨げる。本実施例によれば、制御情報を転送する相互結合網１：３２とデータを転送する相互結合網２：３３を分けることができるため、両者が互いの転送を妨げることがなくなるため、性能が向上する。
［実施例６］
図１１、図１６、図１７に，実施例１のストレージシステム１におけるディスク制御クラスタの増設手順の一例を示す。
図１１に示すように、グローバルスイッチボックス３１０と、フロントエンドスイッチボックス３１５と、クラスタ筐体３０２は、それぞれ別筐体として、筐体３０１内に実装されている。
グローバルスイッチボックス３１０内には，ＧＳＷ１１５とグローバル情報制御部２１が実装されている。
グローバルスイッチボックス３１０はコネクタ３２１、コネクタ３２２をそれぞれ８個有し、ディスク制御クラスタを8クラスタ接続することができる。図ではディスク制御クラスタを３クラスタ接続した場合について示している。
ＧＳＷ１１５のアクセスパス１３６は１本ずつコネクタ３２１，コネクタ３２２に接続される。
また、グローバルスイッチボックス３１０はフロントエンドスイッチボックス３１５を接続するためのコネクタ３２６，コネクタ３２７をそれぞれ２個有する。グローバル情報制御部２１の１つのアクセス制御部１０１をフロントエンドスイッチ７の２つのメモリコントローラ７３に接続するための２本の接続パスは２つのコネクタ３２６に接続される。また、もう１つのアクセス制御部１０１をフロントエンドスイッチ７の２つのメモリコントローラ７３に接続するための２本の接続パスは２つのコネクタ３２７に接続される。
フロントエンドスイッチボックス３１５内には、フロントエンドスイッチ７が実装されている。
フロントエンドスイッチボックス３１５はグローバルスイッチボックス３１０を接続する

ためのコネクタ３２８、コネクタ３２９をそれぞれ２個有する。
フロントエンドスイッチ７の１つのメモリコントローラ７３をグローバル情報制御部２１の２つのアクセス制御部１０１に接続するための２本の接続パスは２つのコネクタ３２８に接続される。また、もう１つのメモリコントローラ７３をグローバル情報制御部２１の２つのアクセス制御部１０１に接続するための２本の接続パスは２つのコネクタ３２９に接続される。
上記個数は一実施例に過ぎず、個数を上記に限定するものではない。
各ディスク制御クラスタ１－１乃至１－３は，それぞれクラスタ筐体３０２に実装されている。クラスタ筐体３０２はコネクタ３２１，コネクタ３２２を有し、２本のアクセスパス１３６がそれぞれに１本ずつ接続されている。
グローバルスイッチボックス３１０に，ケーブル３３１，ケーブル３３２を介してそれぞれのコネクタ３２１，コネクタ３２２により３つのクラスタ筐体３０２が接続される。
また、グローバルスイッチボックス３１０のコネクタ３２６、コネクタ３２７に、ケーブル３３６、ケーブル３３７を介して、フロントエンドスイッチボックス３１５のコネクタ３２８、コネクタ３２９が接続される。
ここで、２つのコネクタ３２６に接続された２本のケーブル３３６は、１本ずつコネクタ３２８とコネクタ３２９に接続される。同様に、２つのコネクタ３２７に接続された２本のケーブル３３７は、１本ずつコネクタ３２８とコネクタ３２９に接続される。こうすることにより、アクセス制御部１０１の２本の接続パスは、１本ずつ２つのメモリコントローラ７３に接続される。
ストレージシステム１において，ディスク制御クラスタを増設する場合は、次の手順による。グローバルスイッチボックス３１０にディスク制御クラスタを増設するコネクタに余分があれば、そのコネクタにケーブル３３１，ケーブル３３２を接続する。
余分がなければ，ＧＳＷのみを実装したグローバルスイッチボックスを用意し、グローバルスイッチボックスを多段に接続した上でそのコネクタにケーブル３３１，ケーブル３３２を接続する。
それと共に、図１６に示すＧＳＷ１１５のポートに接続されるディスク制御クラスタを示す、言い換えるとストレージシステム１を構成しているディスク制御クラスタを示すＧＳＷポート－クラスタ対応テーブル４００と，図１７に示すディスク制御クラスタが管理する論理ボリュームを示すＧＳＷポート－クラスタ対応テーブル４０５とを書き換える。
ＧＳＷポート－クラスタ対応テーブル４００とＧＳＷポート－クラスタ対応テーブル４０５はグローバル情報制御部２１に格納されており、サービスプロセッサ（ＳＶＰ）により書き換えることが可能である。
ＳＶＰは通常ノートパソコンであることが多く、ノートパソコンのディスプレイ上に図１６及び図１７に示すテーブルが表示され、そこで内容を書き換える。
図１６及び図１７は，それぞれディスク制御クラスタの増設前、増設後のＧＳＷポート－クラスタ対応テーブル４００及びＧＳＷポート－クラスタ対応テーブル４０５を示している。
ここでは，ストレージシステム１が増設前に５台のディスク制御クラスタで構成されており、そこに１台のディスク制御クラスタを増設する例を示している。
図１６に示すように、ＧＳＷポート番号４０１の４番ポートが未接続となっており、そのポートにクラスタ５のケーブルを接続した後、ＳＶＰのディスプレイ上でポート番号４の行のクラスタ番号４０２の列の未接続表示を５に書き換える。その後、図１７に示すように、クラスタ番号４０２のクラスタ５の行の論理ボリューム番号４０６の列の未接続表示を１６６４０～２０７３５に書き換える。
ここで、論理ボリューム番号４０６は各クラスタが管理する論理ボリュームの範囲を示している。
増設前の論理ボリューム番号の最大値は１６６３９で、ディスク制御クラスタは４０９６個の論理ボリュームを持っているため、ディスク制御クラスタ５の管理する論理ボリュームの範囲は１６６４０～２０７３５となる。論理ボリューム番号は連続せず飛び飛びにな

っていても問題ない。
上記のようにすることで、ストレージシステムに新たにディスク制御クラスタを増設することができる。
［実施例７］
図１８～図２１に、ストレージシステム１において、あるチャネルＩＦ部のホストコンピュータとの１つのインタフェース(チャネル)の負荷が高くなった場合の、グローバル制御情報部及びフロントエンドスイッチ７の動作の一例を示す。実施例１に示すように、一例として、グローバル情報制御部２１は、図１８に示すようなシステム構成・稼動状況テーブル５００をそのメモリモジュール１０５に格納しており、フロントエンドスイッチ７は、テーブル５００のコピーをメモリモジュール１０５に格納している。
グローバル情報制御部２１内の管理機能部１０２は、各ディスク制御クラスタのローカル共有メモリ部２２内に格納された該ディスク制御クラスタが管理する記憶領域の情報や、ディスク制御クラスタ内の各部位の負荷情報、障害情報を定期的に参照し、システム構成・稼動状況テーブル５００を更新する。
また、フロントエンドスイッチ７内のスイッチ制御部７２は、テーブル５００を定期的に自メモリモジュール１０５にコピーする。
一方、各ディスク制御クラスタのローカル共有メモリ部２２内に格納された該ディスク制御クラスタが管理する記憶領域の情報、ディスク制御クラスタ内の各部位の負荷情報、あるいは障害情報が更新された時点で、チャネルＩＦ部またはディスクＩＦ部内のマイクロプロセッサが、グローバル情報制御部のメモリモジュール１０５内の該当情報を更新し、さらに、その更新の時点で、管理機能部１０２が該当情報をフロントエンドスイッチ７のメモリモジュール１０５にコピーする場合もある。これは、特にディスク制御部内のある部位に障害が起こった場合において有効である。
システム構成・稼動状況テーブル５００は、ディスク制御クラスタを識別するクラスタ番号５１１とディスク制御クラスタのチャネルを識別するチャネル番号５１２に対応する論理ボリューム番号５１３とチャネル稼動状況５１４を示す。
本実施例では、クラスタ番号５１１、チャネル番号５１２、及び論理ボリューム番号５１３を１６進数で示している。チャネル稼動状況５１４は、低負荷状態を‘０’、中負荷状態を‘１’、高負荷状態を‘２’、障害を‘３’で示している。
また、フロントエンドスイッチ７は、ホスト－論理ボリューム対応テーブル６００をそのメモリモジュール１０５に格納している。
ホスト－論理ボリューム対応テーブル６００は、個々のホストコンピュータを識別するホスト番号６１５に対応する論理ボリューム番号５１３、すなわち各ホストコンピュータに割当てられた論理ボリュームと、論理ボリューム番号５１３にアクセスするためのチャネル番号５１２と該当論理ボリュームを管理するディスク制御クラスタのクラスタ番号５１１を示す。
ホスト番号６１５は、例えば、ファイバチャネルのプロトコルで使われるワールドワイドネーム（ＷＷＮ）や、インターネットプロトコルで使われるマックアドレスやＩＰアドレスを充てることが考えられる。
図１９に示すように、１つのホストコンピュータに対し、連続しない複数の論理ボリュームの範囲を割当てても問題ない。
スイッチ制御部７２は、ホスト－論理ボリューム対応テーブル６００（ルーティングテーブル）を参照し、スイッチ７１の切り替えを行う。
図２０に示すように、チャネル番号５１２の２番チャネルの稼動状況５５１が高負荷を示す‘２’となった場合、２番チャネルに割当てられた論理ボリューム０１００～０１ＦＦの一部を、稼動状況５５２が低負荷を示す‘０’となっている１Ｆ番チャネルに割当て替える。
図２１は、上記割当て替えを行った後のテーブル５００の内容を示している。
本実施例では、２番チャネルに割当てられていた論理ボリューム０１００～０１ＦＦの半分の０１８０～０１ＦＦを１Ｆ番チャネルに割当て替えた。これにより、２番チャネルの

稼動状況５５１は中負荷を示す‘１’となり、１Ｆ番チャネルの稼動状況５５２も中負荷を示す‘１’となる。
図２１に示すテーブル５００がフロントエンドスイッチ７のメモリモジュール１０５にコピーされると、スイッチ制御部７２は、テーブル５００を参照し図１９に示すホスト－論理ボリューム対応テーブル６００を変更する。
図２２は変更後のテーブル６００を示す。
テーブル５００において、論理ボリュームの０１８０～０１ＦＦは１Ｆ番チャネルに割当て返られているので、２番のホストコンピュータ６５２に新たに１Ｆ番チャネルが割当てられる。
これにより、２番のホストコンピュータ６５２から論理ボリューム０１８０～０１ＦＦへのアクセス要求が発行された場合、スイッチは１Ｆ番チャネルに切り替えられる。
この場合、アクセス要求を受けたチャネルＩＦ部のマイクロプロセッサは、ローカル共有メモリ部にアクセスし、自ディスク制御クラスタ内に要求データがないことを知るが、その場合、グローバル情報制御部にアクセスすることにより、要求データのあるディスク制御クラスタを知り、該当ディスク制御クラスタから要求データを読み出すことができる。
これにより、ホストコンピュータに意識させること無しに、ストレージシステム１の各チャネルＩＦ部のチャネルの負荷を均等にすることができ、それによりシステム全体の性能向上が可能となる。

【００４１】

また、あるチャネル番号のチャネルが障害で使用できなくなった場合にも、上記の方法で論理ボリュームの割当て替えを行うことにより、別のチャネルから目的のディスク制御クラスタのディスク制御装置に格納されたデータへアクセスすることが、チャネルの障害をホストコンピュータに意識させること無しに、可能となる。

【００４２】

また、あるディスク制御クラスタに繋がるディスク装置上の論理ボリュームの負荷が高くなった場合、負荷を均等にするために該当論理ボリュームを他のディスク制御クラスタのディスク装置にコピーまたは移動する場合があるが、その場合も上記の方法でテーブル５００の論理ボリューム番号とチャネル番号を割当て替えることにより、ホストコンピュータに意識させること無しに、論理ボリュームの負荷を均等にすることが可能となる。

［実施例８］

図２４に、本発明の一実施例を示す。
以下の実施例において、相互結合網はスイッチを利用したものを例にして説明してあるが、相互に接続され制御情報やデータが転送されれば良いのであり、例えばバスで構成されても良い。
図２４に示すように、ストレージシステム１は複数のディスク制御クラスタ１－１乃至１－ｎとフロントエンドスイッチ７から構成される。
本実施例のストレージシステム１は、グローバル制御部２１とフロントエンドスイッチ７が接続されていない点を除いて、実施例１の図１に示すストレージシステムと同様である。
また、その具体的な一例は、図３に示すストレージシステム１においてアクセス制御部１０１とメモリコントローラ７３が接続されていない構成となる。
上記のように、アクセス制御部１０１とメモリコントローラ７３が接続されないため、グローバル制御情報部２１内の各ディスク制御クラスタが管理する記憶領域の情報、ディスク制御クラスタ内の各部位の負荷情報、および障害情報をチャネルＩＦ部１１とスイッチ７１の接続パスを介してフロントエンドスイッチ７のメモリモジュール１０５にコピーする処理を行う。
この処理は、フロントエンドスイッチ７内のスイッチ制御部２１がチャネルＩＦ部１１とスイッチ７１の接続パスを介して、チャネルＩＦ部１１内のマイクロプロセッサに処理要求を出すことにより行う。
例えば、上記接続パスが、インターネットプロトコルを流すことが可能である場合、シン

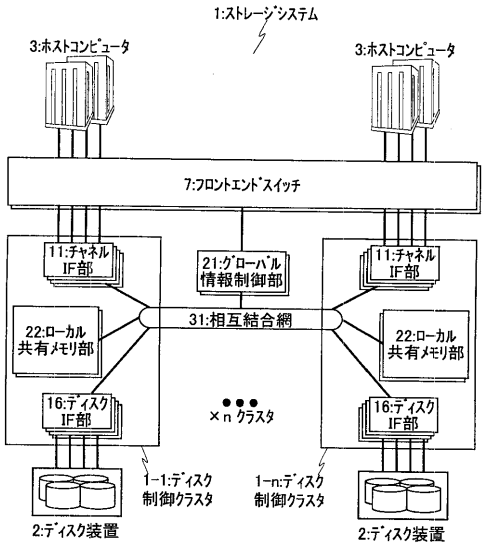
プル・ネットワーク・マネージメント・プロトコル（ＳＮＭＰ）により、上記情報を取得することが可能である。
本実施例によれば、フロントエンドスイッチ７とグローバル情報制御部２１を接続する必要がなくなり、ストレージシステムを構成する筐体の実装を簡素化することが可能となる。

【００４３】
【発明の効果】
本発明によれば、複数台のディスク制御クラスタを１つのシステムとして運用するストレージシステムにおいて、ディスク制御クラスタが１個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティがあり使い勝手の良い構成のストレージシステムを提供することが可能となる。
【図面の簡単な説明】
【図１】本発明によるストレージシステムの実施例１の構成を示す図。
【図２】従来の複数のディスク制御装置の構成を示す図。
【図３】図1に示す実施例１のストレージシステムの詳細構成を示す図。
【図４】本発明によるストレージシステムの実施例２の構成を示す図。
【図５】図４に示す実施例２のストレージシステムの詳細構成を示す図。
【図６】本発明によるストレージシステムの実施例３の構成を示す図。
【図７】図６に示す実施例３のストレージシステムの詳細構成を示す図。
【図８】本発明によるストレージシステムの実施例４の構成を示す図。
【図９】図８に示す実施例４のストレージシステムの詳細構成を示す図。
【図１０】本発明によるストレージシステムの実施例５の構成を示す図。
【図１１】本発明によるディスク制御クラスタの増設方法を説明するための図。
【図１２】本発明によるストレージシステムを構成するチャネルインターフェース部の構成を示す図。
【図１３】本発明によるストレージシステムを構成するディスクインターフェース部の構成を示す図。
【図１４】本発明によるストレージシステムを構成するチャネルインターフェース部の他の構成を示す図。
【図１５】本発明によるストレージシステムを構成するディスクインターフェース部の他の構成を示す図。
【図１６】グローバル情報制御部内に格納されたストレージシステムの構成情報の一例を示す図。
【図１７】グローバル情報制御部内に格納されたストレージシステムの構成情報の他の一例を示す図。
【図１８】グローバル情報制御部内に格納されたストレージシステムの構成・稼動状況の一例を示す図。
【図１９】フロントエンドスイッチ内に格納されたスイッチ切換え制御のためのテーブルの一例を示す図。
【図２０】グローバル情報制御部内に格納されたストレージシステムの構成・稼動状況の他の一例を示す図。
【図２１】グローバル情報制御部内に格納されたストレージシステムの構成・稼動状況の他の一例を示す図。
【図２２】フロントエンドスイッチ内に格納されたスイッチ切換え制御のためのテーブルの他の一例を示す図。
【図２３】従来の複数のディスク制御クラスタから成るストレージシステムの構成を示す図。
【図２４】本発明によるストレージシステムの実施例８の構成を示す図。
【符号の説明】

１：ストレージシステム、１－１、１－ｎ…ディスク制御クラスタ、２…ディスク装置、３…ホストコンピュータ、７…フロントエンドスイッチ、１１…チャネルＩＦ部、１６…ディスクＩＦ部、２１…グローバル情報制御部、２２…ローカル共有メモリ部、３１…相互結合網。

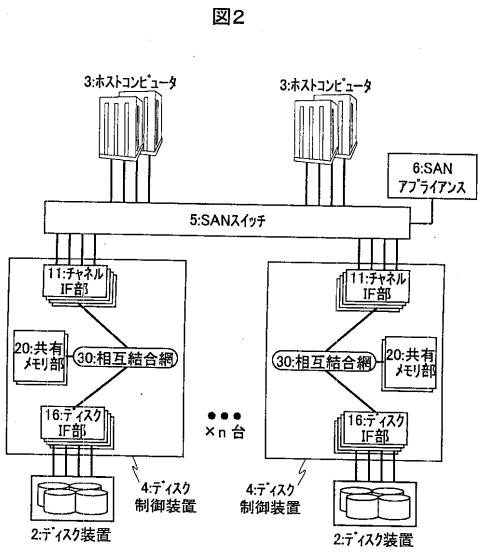
【図１】

図1



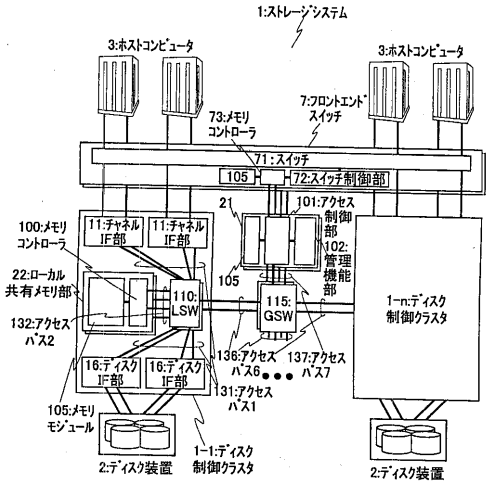
【図２】

図2



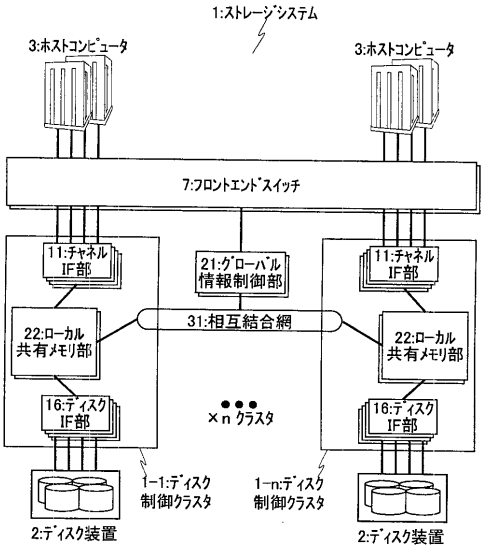
【図３】

図3



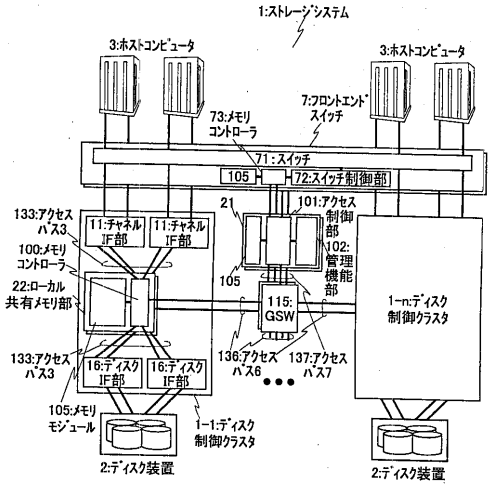
【図４】

図4



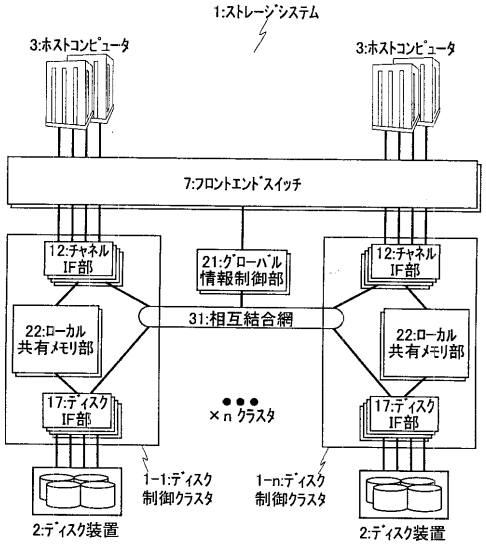
【図５】

図5



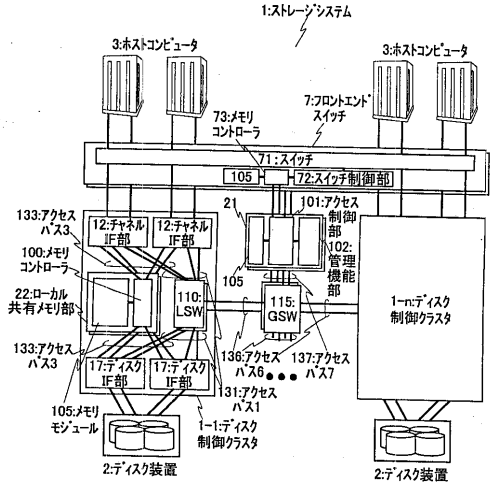
【図６】

図6



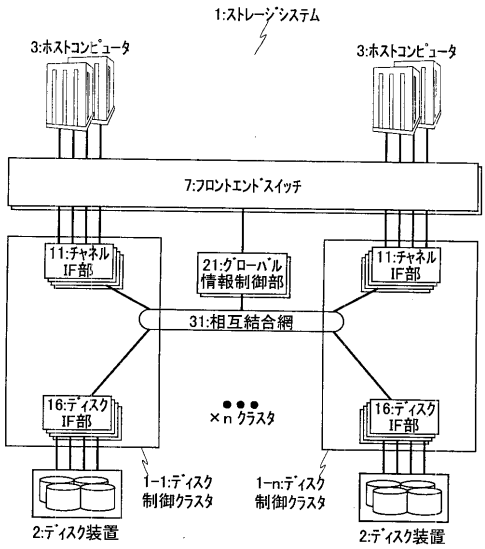
【図７】

図7



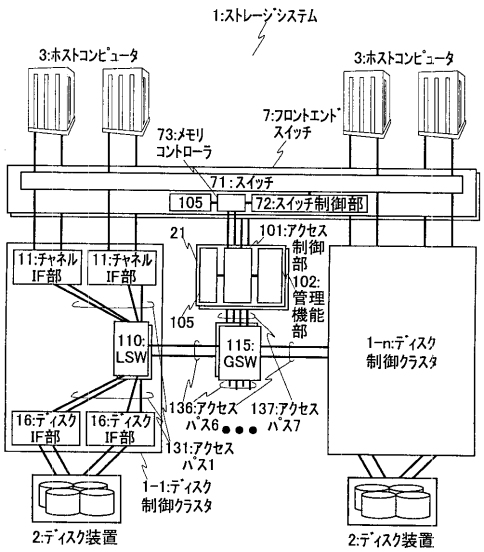
【図８】

図8



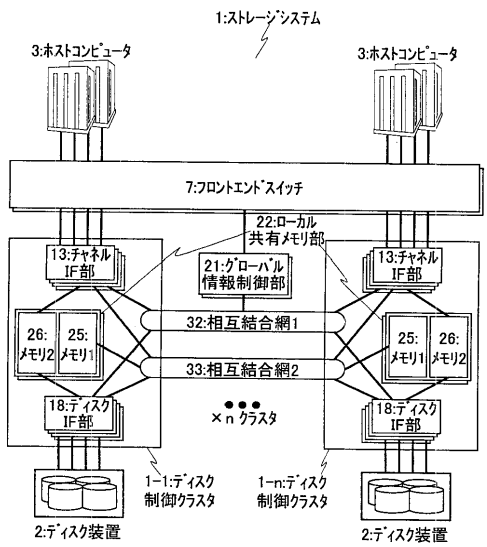
【図９】

図9



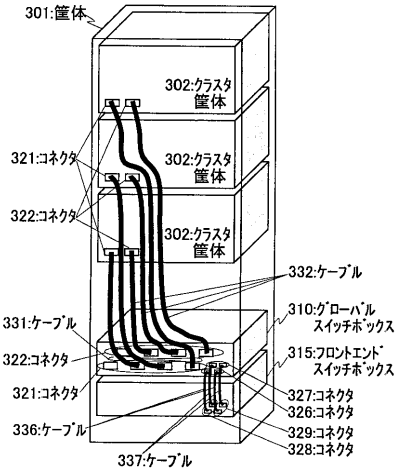
【図１０】

図10



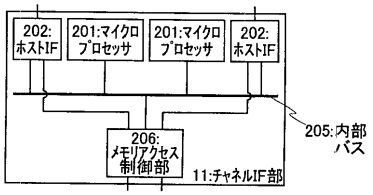
【図１１】

図11



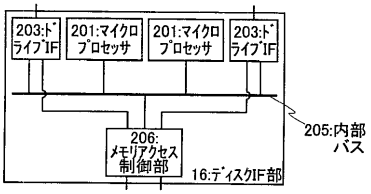
【図１２】

図12



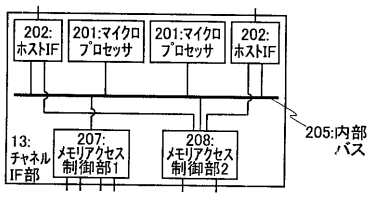
【図１３】

図13



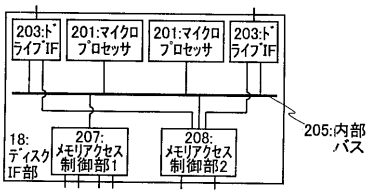
【図１４】

図14



【図１５】

図15



【図１６】

図16

(増設前)

401 GSWポート番号	402 クラスタ番号
0	0
1	1
2	3
3	2
4	未接続
5	4
6	未接続
7	未接続

400:GSWポート-クラスタ対応テーブル

(増設後)

401 GSWポート番号	402 クラスタ番号
0	0
1	1
2	3
3	2
4	5
5	4
6	未接続
7	未接続

400:GSWポート-クラスタ対応テーブル

【図１７】

図17

（増設前）

クラスタ番号	論理ﾎﾞﾘｭｰﾑ番号
0	0～4095
1	4096～6143
2	12288～16383
3	6144～12287
4	16384～16639
5	未実装
6	未実装
7	未実装

402 406

405:ｸﾗｽﾀｰ-論理ﾎﾞﾘｭｰﾑ対応ﾃｰﾌﾞﾙ

（増設後）

クラスタ番号	論理ﾎﾞﾘｭｰﾑ番号
0	0～4095
1	4096～6143
2	12288～16383
3	6144～12287
4	16384～16639
5	16640～20735
6	未実装
7	未実装

402 406

405:ｸﾗｽﾀｰ-論理ﾎﾞﾘｭｰﾑ対応ﾃｰﾌﾞﾙ

【図１８】

図18

クラスタ番号	チャネル番号	論理ﾎﾞﾘｭｰﾑ番号	チャネル稼動状況
0	0	0000～007F	1
0	1	0080～00FF	1
0	2	0100～01FF	1
0	3	0200～02FF	1
1	4	0300～033F	1
1	5	0340～03FF	1
1	6	0400～04FF	1
1	7	0300～033F	1
⋮	⋮	⋮	⋮
7	1C	1000～10FF	1
7	1D	1100～12FF	1
7	1E	1300～14FF	1
7	1F	1500～17FF	0

511 512 513 514

500:ｼｽﾃﾑ構成・稼動状況ﾃｰﾌﾞﾙ

【図１９】

図19

ホスト番号	論理ﾎﾞﾘｭｰﾑ番号	チャネル番号	クラスタ番号
0	0000～007F	0	0
1	0080～00FF	1	0
1	0340～03FF	5	1
2	0100～01FF	2	0
3	0200～02FF	3	0
4	0300～033F	4	1
5	0400～04FF	6	1
6	0300～033F	7	1
⋮	⋮	⋮	⋮
1C	1000～10FF	1C	7
1D	1100～12FF	1D	7
1E	1300～14FF	1E	7
1F	1500～17FF	1F	7

615 513 512 511

651

652

600:ﾎｽﾄｰ論理ﾎﾞﾘｭｰﾑ対応ﾃｰﾌﾞﾙ

【図２０】

図20

クラスタ番号	チャネル番号	論理ﾎﾞﾘｭｰﾑ番号	チャネル稼動状況
0	0	0000～007F	1
0	1	0080～00FF	1
0	2	0100～01FF	2
0	3	0200～02FF	1
1	4	0300～033F	1
1	5	0340～03FF	1
1	6	0400～04FF	1
1	7	0300～033F	1
⋮	⋮	⋮	⋮
7	1C	1000～10FF	1
7	1D	1100～12FF	1
7	1E	1300～14FF	1
7	1F	1500～17FF	0

511 512 513 514

551

552

500:ｼｽﾃﾑ構成・稼動状況ﾃｰﾌﾞﾙ

【図２１】

図21

511 クラスタ番号	512 チャネル番号	513 論理ボリューム番号	514 チャネル稼動状況
0	0	0000～007F	1
0	1	0080～00FF	1
0	2	0100～017F	1
0	3	0200～02FF	1
1	4	0300～033F	1
1	5	0340～03FF	1
1	6	0400～04FF	1
1	7	0300～033F	1
⋮	⋮	⋮	⋮
7	1C	1000～10FF	1
7	1D	1100～12FF	1
7	1E	1300～14FF	1
7	1F	0180～01FF	1
7	1F	1500～17FF	1

551, 552

500:システム構成・稼動状況テーブル

【図２２】

図22

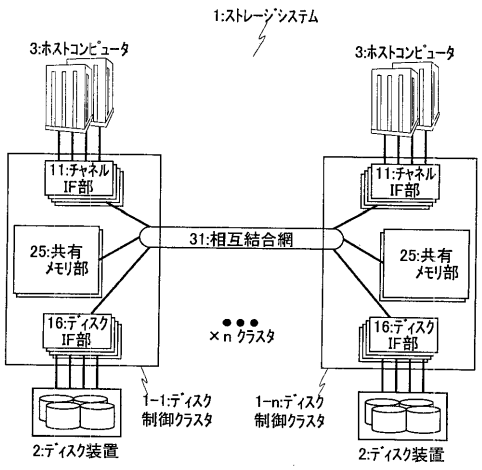
615 ホスト番号	613 論理ボリューム番号	612 チャネル番号	611 クラスタ番号
0	0000～007F	0	0
1	0080～00FF	1	0
1	0340～03FF	5	1
2	0100～017F	2	0
2	0180～01FF	1F	7
3	0200～02FF	3	0
4	0300～033F	4	1
5	0400～04FF	6	1
6	0300～033F	7	1
⋮	⋮	⋮	⋮
1C	1000～10FF	1C	7
1D	1100～12FF	1D	7
1E	1300～14FF	1E	7
1F	1500～17FF	1F	7

651, 652

600:ホスト－論理ボリューム対応テーブル

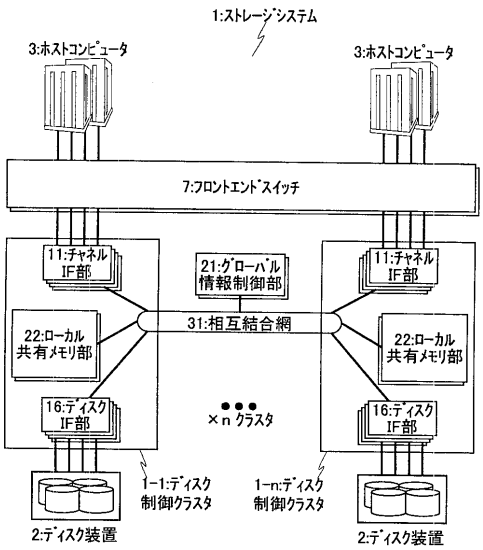
【図２３】

図23



【図２４】

図24



フロントページの続き

(56)参考文献 特開２０００－２４２４３４（ＪＰ，Ａ）
特開平１１－００７３５９（ＪＰ，Ａ）
特開２００１－２５６００３（ＪＰ，Ａ）
特開２００１－２２９０４２（ＪＰ，Ａ）
特開平０８－３２８７６０（ＪＰ，Ａ）
特開２００１－２９０６０８（ＪＰ，Ａ）
特開２００１－１４２６４８（ＪＰ，Ａ）
特開平１１－２９６３１３（ＪＰ，Ａ）
特開平０７－０５６６９１（ＪＰ，Ａ）

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