

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4874515号
(P4874515)

(45) 発行日 平成24年2月15日(2012.2.15)

(24) 登録日 平成23年12月2日(2011.12.2)

(51) Int.Cl.

F I

G 0 6 F 3/06 (2006.01)

G 0 6 F 3/06 3 0 1 A

G 0 6 F 3/06 5 4 0

請求項の数 12 (全 33 頁)

(21) 出願番号 特願2003-426165 (P2003-426165)
 (22) 出願日 平成15年12月24日(2003.12.24)
 (62) 分割の表示 特願平11-344260の分割
 原出願日 平成11年12月3日(1999.12.3)
 (65) 公開番号 特開2004-145901 (P2004-145901A)
 (43) 公開日 平成16年5月20日(2004.5.20)
 審査請求日 平成18年11月30日(2006.11.30)
 審査番号 不服2010-21471 (P2010-21471/J1)
 審査請求日 平成22年9月24日(2010.9.24)
 (31) 優先権主張番号 特願平10-364079
 (32) 優先日 平成10年12月22日(1998.12.22)
 (33) 優先権主張国 日本国(JP)

(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 110000279
 特許業務法人ウィルフォート国際特許事務所
 (72) 発明者 松並 直人
 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 大枝 高
 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】 記憶装置システム

(57) 【特許請求の範囲】

【請求項 1】

複数のコンピュータからアクセスされるデータを格納するシステムであって、
 データを格納する複数のディスクドライブと、前記複数のディスクドライブを用いて複
 数のロジカルユニット(LU)を構成するよう制御する制御部と、をそれぞれに有する複
 数のストレージシステムと、

前記複数のコンピュータと接続される複数の第一のノードと、各々前記ストレージシ
 ステムと接続される複数の第二のノードと、を有するスイッチと、を含み、

前記スイッチは、

前記複数のストレージシステムのうちの第一のストレージシステムの有する複数のLU
 のうちの第一のLUと、前記複数のストレージシステムのうちの第二のストレージシ
 ステムの有する複数のLUのうちの第二のLUと、から構成される第三のLUを前記コンピ
 ュータに提供し、

前記コンピュータに使用される前記第三のLU内のアドレスと、前記第一のストレ
 ジシステムまたは前記第二のストレージシステムに使用される第一のLU内のアドレ
 スおよび第二のLU内のアドレスと、を対応づける情報を有し、

前記コンピュータから、前記第一のノードを介して、前記コンピュータが使用する前
 記第三のLU内のアクセス対象データの格納位置を示す第一のアドレスを含む第一のアク
 セス要求を受信し、

前記情報を用いて、前記受信した第一のアクセス要求に含まれる前記第一のアドレスを

10

20

、前記アクセス対象データが含まれる前記第一のＬＵ又は前記第二のＬＵ内の、アクセス対象データの格納位置を示す、前記ストレージシステムが使用する第二のアドレスに変換し、

前記受信した第一のアクセス要求を、前記第二のアドレスを含む第二のアクセス要求に変換し、

前記第二のアクセス要求を、前記アクセス対象データを有する前記第一ストレージシステム又は前記第二ストレージシステムが接続されている第二のノードから、該ストレージシステムに送信することを特徴とするシステム。

【請求項２】

請求項１記載のシステムであって、

前記情報には、前記第三のＬＵのアドレス空間と、前記第三のＬＵを構成する前記第一のＬＵ及び前記第二のＬＵのそれぞれが有するアドレス空間との対応が含まれており、

前記第一のアドレスには、前記第三のＬＵを識別する論理ユニット番号と、前記第三のＬＵ内での前記アクセス対象であるデータが格納される格納位置を示す論理ブロックアドレスとが含まれ、

前記第二のアドレスには、前記アクセス対象データを有する第一のＬＵ又は第二のＬＵを識別する論理ユニット番号と、前記第一のＬＵ又は前記第二のＬＵ内での前記アクセス対象であるデータが格納される格納位置を示す論理ブロックアドレスとが含まれることを特徴とするシステム。

【請求項３】

請求項２記載のシステムであって、

前記第一のアクセス要求はフレームペイロードを有しており、前記第一のアドレスはフレームペイロードに格納されていることを特徴とするシステム。

【請求項４】

請求項２記載のシステムであって、

前記第一のアクセス要求及び前記第二のアクセス要求は、フレーム転送先ＩＤが格納されるフレームヘッダを有しており、

前記第一のアクセス要求には、前記第一のノードを示す第一フレーム転送先ＩＤが含まれており、

前記スイッチは、前記第一フレーム転送先ＩＤを、前記第二のアクセス要求の送信先であるストレージシステムを示す第二フレーム転送先ＩＤに変換し、変換後の第二フレーム転送先ＩＤを有する前記第二のアクセス要求を、前記第二のアクセス要求の送信先であるストレージシステムに送信することを特徴とするシステム。

【請求項５】

請求項１記載のシステムであって、さらに、

前記複数のストレージシステムと前記スイッチに接続され、前記情報を前記スイッチに設定する管理手段とを含み、

前記管理手段が前記スイッチに設定する前記情報は、前記複数のストレージシステムが有する論理ユニットの情報を含むことを特徴とするシステム。

【請求項６】

請求項１記載のシステムであって、さらに、

前記複数のストレージシステムと前記スイッチに接続され、前記情報を前記スイッチに設定する管理手段とを含み、

前記管理手段は、前記複数のストレージシステムの構成情報と前記スイッチの構成情報を獲得し、獲得した構成情報を基に前記対応付ける情報を前記スイッチに設定することを特徴とするシステム。

【請求項７】

請求項６記載のシステムであって、前記管理手段から前記スイッチに設定される情報は、前記ストレージシステムのディスクの数と容量が含まれることを特徴とするシステム。

【請求項８】

請求項 1 記載のシステムであって、さらに、

前記複数のストレージシステムと前記スイッチに接続され、前記情報を前記スイッチに設定する管理手段とを含み、

前記管理手段が設定する前記情報は、前記複数のコンピュータの一つから前記ストレージシステムの論理ユニットへのアクセスの可否を示すアクセス制限情報を含むことを特徴とするシステム。

【請求項 9】

請求項 1 記載のシステムであって、さらに、

前記複数のストレージシステムと前記スイッチに接続され、前記情報を前記スイッチに設定する管理手段とを含み、

前記管理手段が設定する前記情報は、前記コンピュータの一つから前記ストレージシステムの論理ユニットの特定アドレスへのアクセスの可否を示すアクセス制限情報を含むことを特徴とするシステム。

【請求項 10】

請求項 1 記載のシステムであって、さらに、

前記複数のストレージシステムと前記スイッチに接続され、前記情報を前記スイッチに設定する管理手段とを含み、

前記管理手段は、前記複数のストレージシステムが有する論理ユニットの RAID レベルを設定することを特徴とするシステム。

【請求項 11】

請求項 1 記載のシステムであって、さらに、

前記複数のストレージシステムと前記スイッチに接続され、前記情報を前記スイッチに設定する管理手段とを含み、

前記管理手段は、前記複数のストレージシステムの一つから障害情報を受信することを特徴とするシステム。

【請求項 12】

請求項 1 記載のシステムであって、さらに、

前記複数のストレージシステムと前記スイッチに接続され、前記情報を前記スイッチに設定する管理手段とを含み、

前記管理手段は、前記複数のストレージシステムの構成情報と前記スイッチの構成情報を獲得し、当該管理手段に接続された表示装置に、前記コンピュータと前記ストレージシステムの論理ユニットとの接続関係を表示することを特徴とするシステム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、複数のディスク装置を制御するディスク制御システムの実現方法に関し、特に、ディスク制御システムの高速化、低コスト化、コストパフォーマンスの向上の方法に関する。

【背景技術】

【0002】

計算機システムに用いられる記憶装置システムとして、複数のディスク装置を制御するディスクアレイシステムがある。ディスクアレイシステムについては、例えば、非特許文献 1 に開示されている。

【0003】

ディスクアレイは、複数のディスク装置を並列に動作させることで、ディスク装置を単体で用いた記憶装置システムに比べ高速化を実現する技術である。

【0004】

複数のディスクアレイシステムを、複数のホストと相互に接続する方法として、ファイバチャネル (Fibre Channel) の Fabric を使用した方法がある。この方法を適用した計算

10

20

30

40

50

機システムの例が、非特許文献 2 に示されている。

【 0 0 0 5 】

ここに開示される計算機システムでは、複数のホストコンピュータ（以下では単にホストと呼ぶ）と複数のディスクアレイシステムが、それぞれ、ファイバチャネルを介してファブリック装置に接続される。ファブリック装置は、ファイバチャネルのスイッチであり、ファブリック装置に接続する任意の装置間の転送路の接続を行う。ファブリック装置はファイバチャネルのパケットである「フレーム」の転送に対し透過であり、ホストとディスクアレイシステムは、互いにファブリック装置を意識することなく 2 点間で通信を行う。

【 0 0 0 6 】

【非特許文献 1】デビッド・エー・パターソン (David A. Patterson) 他 2 名著、「ケースオブリダンダントアレイオブインエクスペンシブディスク (A Case for Redundant Arrays of Inexpensive Disks (RAID))」、米国、エーシーエムシグモッドプロシーディング (In Proc. ACM SIGMOD)、1988 年 6 月、p.109-116

【 0 0 0 7 】

【非特許文献 2】「シリアル SCSI がいよいよ市場へ」、日経エレクトロニクス、no.639、1995 年 7 月 3 日、P.79 図 3

【発明の開示】

【発明が解決しようとする課題】

【 0 0 0 8 】

従来のディスクアレイシステムでは、大容量化のためディスク装置の台数を増やし、高性能化のため台数に見合った性能を有するコントローラを実現しようとする、コントローラの内部バスの性能限界や、転送制御を行うプロセッサの性能限界が顕在化する。このような問題に対処するために、内部バスを拡張し、プロセッサ数を増加することが行われている。しかし、このような対処の仕方は、多数のバス制御によるコントローラ構成の複雑化や、プロセッサ間の共有データの排他制御等による制御ソフトの複雑化とオーバヘッドの増加を招く。このため、コストを非常に上昇させるとともに、性能は頭打ちになり、その結果、コストパフォーマンスが悪化する。また、このような装置は、大規模なシステムでは、そのコストに見合った性能が実現できるものの、規模がそれほど大きくないシステムには見合わない、拡張性が制限される、開発期間の増大と開発コストの上昇を招くといった課題がある。

【 0 0 0 9 】

複数のディスクアレイシステムを並べファブリック装置で相互接続することによって、システム全体としての大容量化、高性能化を行うことが可能である。しかし、この方法では、ディスクアレイシステム間に関連性は全くなく、特定のディスクアレイシステムにアクセスが集中したとしてもそれを他の装置に分散することができないので、実使用上の高性能化が実現できない。また、ホストから見た論理的なディスク装置（論理ユニットと呼ぶ）の容量は、1 台のディスクアレイシステムの容量に制限されるので、論理ユニットの大容量化は実現できない。

【 0 0 1 0 】

ディスクアレイシステム全体を高信頼化しようとした際に、ホストが備えているミラーリング機能を用いて 2 台のディスクアレイシステムによるミラー構成を実現することができるが、ホストによるミラーリングのための制御オーバヘッドが発生し、システム性能が制限されるという課題がある。また、多数のディスクアレイシステムがシステム内に個別に存在すると、システム管理者が管理するための負荷が増加する。このため、多数の保守人員、複数台分の保守費用が必要になる等、管理コストが増加する。さらに、複数のディスクアレイシステム、ファブリック装置は、それぞれ独立した装置であるので、各種設定は、それぞれの装置毎に異なる方法で実施する必要がある。このため、管理者のトレーニングや、操作時間の増大にともない運用コストが増大する。

【 0 0 1 1 】

本発明の目的は、これら従来技術における課題を解決し、計算機システムの規模、要求などに応じた記憶装置システムを構築でき、将来における記憶装置システムの拡張、信頼性の向上などに容易に対応することのできる記憶装置システムを実現することにある。

【課題を解決するための手段】

【0012】

本発明の記憶装置システムは、データを保持する記憶媒体を有する記憶装置と、この記憶装置を制御する制御装置とを有する複数の記憶装置サブシステム、複数の記憶装置サブシステムに保持されるデータを使用する計算機に接続された第1のインタフェースノード、各々が記憶装置サブシステムのいずれかに接続された複数の第2のインタフェースノード、及び第1のインタフェースノード及び複数の第2のインタフェースノードが接続され、第1のインタフェースノードと複数の第2のインタフェースノードとの間でフレームの転送を行う転送手段を有する。

10

【0013】

好ましくは、第1のインタフェースノードは、記憶装置システムの構成情報を格納した構成管理テーブルと、計算機から送られてくるフレームに应答して、該フレームを解析し、構成管理テーブルに保持された構成情報に基づいてそのフレームの転送先に関する情報変換して転送手段に転送する。

【0014】

また、フレームの転送に際して、第1のインタフェースノードは、そのフレームを受け取るべきノードのノードアドレス情報をフレームに付加する。転送手段はフレームに付加されたノードアドレス情報に従ってフレームを転送する。第2のインタフェースノードは、転送手段から受け取ったフレームからノードアドレス情報を除いてフレームを再形成し、目的の記憶装置サブシステムに転送する。

20

【0015】

本発明のある態様において、記憶装置システムは、転送手段に接続する管理プロセッサを有する。管理プロセッサは、オペレータからの指示に従って、構成管理テーブルに構成情報を設定する。構成情報には、計算機からのアクセスを制限する情報が含まれる。

【発明の効果】

【0016】

本発明によれば、計算機システムの規模、要求などに応じた記憶装置システムの拡張、信頼性の向上などを容易に実現することのできる記憶装置システムを実現することができる。

30

【発明を実施するための最良の形態】

【0017】

〔第1実施形態〕

図1は、本発明が適用されたディスクレイシステムを用いたコンピュータシステムの一実施形態における構成図である。

【0018】

1はディスクレイシステム、30はディスクレイシステムが接続されるホストコンピュータ（ホスト）である。ディスクレイシステム1は、ディスクレイサブセット10、ディスクレイスイッチ20、ディスクレイシステム全体の設定管理を行うディスクレイシステム構成管理手段70、ディスクレイスイッチ20とディスクレイシステム構成管理手段70との間、およびディスクレイサブセット10とディスクレイシステム構成管理手段70との間の通信インタフェース（通信I/F）80を有する。ホスト30とディスクレイシステム1とは、ホストインタフェース（ホストI/F）31で接続されており、ホストI/F31はディスクレイシステム1のディスクレイスイッチ20に接続する。ディスクレイシステム1の内部において、ディスクレイスイッチ20とディスクレイサブセット10は、ディスクレイインタフェース（ディスクレイI/F21）で接続される。

40

【0019】

50

ホスト30、ディスクアレイサブセット10は、図では、各々4台示されているが、この台数に関しては制限はなく任意である。ホスト30とディスクアレイサブセット10の台数が異なっても構わない。また、ディスクアレイスイッチ20は、本実施形態では図示の通り二重化されている。各ホスト30および各ディスクアレイサブセット10は、それぞれ別々のホストI/F31、ディスクアレイI/F21で二重化されたディスクアレイスイッチ20の双方に接続されている。これは、一方のディスクアレイスイッチ20、ホストI/F31、あるいはディスクアレイI/F21が故障しても他方を使用することでホスト30からディスクアレイシステム1へのアクセスを可能とし、高い可用性を実現するためである。しかし、このような二重化は必ずしも必須ではなく、システムに要求される信頼性レベルに応じて選択可能である。

10

【0020】

図2は、ディスクアレイサブセット10の一構成例を示す構成図である。101は上位システム(ホスト10)からのコマンドを解釈してキャッシュヒットミス判定を実施し、上位システムとキャッシュ間のデータ転送を制御する上位アダプタ、102はディスクデータアクセス高速化のためのキャッシュ、および、マルチプロセッサ間の共有データを格納する共有メモリ(以下キャッシュ・共有メモリと呼ぶ)、104はディスクアレイサブセット10内に格納される複数のディスクユニットである。103はディスクユニット104を制御し、ディスクユニット104とキャッシュ間のデータ転送を制御する下位アダプタである。106はディスクアレイサブセット構成管理手段であり、ディスクアレイシステム1全

20

体を管理するディスクアレイシステム構成管理手段70と通信I/F80を介して通信し、構成パラメータの設定や、障害情報の通報等の管理を行う。

【0021】

上位アダプタ101、キャッシュ・共有メモリ102、下位アダプタ103はそれぞれ二重化されている。この理由は上記ディスクアレイスイッチ20の二重化と同様、高可用性を実現するためであり必須ではない。また、各ディスクユニット104は、二重化された下位アダプタ103のいずれからも制御可能である。本実施形態では、低コスト化の観点から同一のメモリ手段をキャッシュと共有メモリに共用しているが、これらは勿論分離することも可能である。

【0022】

30

上位アダプタ101は、上位アダプタ101の制御を実行する上位MPU1010、上位システム、すなわちディスクアレイスイッチ20との接続I/FであるディスクアレイI/F21を制御するディスクアレイI/Fコントローラ1011、キャッシュ・共有メモリ102と上位MPU1010とディスクアレイI/Fコントローラ1011との間の通信、データ転送を行う上位バス1012を含む。

【0023】

図では各上位アダプタ101毎に1台のディスクアレイI/Fコントローラ1011が示されているが、1つの上位アダプタに対し、複数のディスクアレイI/Fコントローラ1011を設けてもよい。

【0024】

40

下位アダプタ103は、下位アダプタ103の制御を実行する下位MPU1030、ディスク104とのインタフェースであるディスクI/Fを制御するディスクI/Fコントローラ1031、キャッシュ・共有メモリ102と下位MPU1030とディスクI/Fコントローラ1031との間の通信、データ転送を行う下位バス1032を含む。

【0025】

図では各下位アダプタ103毎に4台のディスクI/Fコントローラ1031が示されているが、その数は任意であり、ディスクアレイの構成や、接続するディスク台数に応じて変更可能である。

【0026】

図3は、ディスクアレイスイッチ20の一構成例を示す構成図である。200はディス

50

クアレイスイッチ全体の制御および管理を行うプロセッサである管理プロセッサ(MP)、201は $n \times n$ の相互スイッチ経路を構成するクロスバススイッチ、202はディスクアレイI/F21毎に設けられるディスクアレイI/Fノード、203はホストI/F31毎に設けられるホストI/Fノード、204はディスクアレイシステム構成管理手段70との間の通信を行う通信コントローラである。2020はディスクアレイI/Fノード202とクロスバススイッチ201を接続するパス、2030はホストI/Fノード203とクロスバススイッチ201を接続するパス、2040は他のディスクアレイスイッチ20と接続し、クラスタを構成するためのクラスタ間I/F、2050はMP200とクロスバススイッチ201を接続するためのパスである。

【0027】

10

図4はクロスバススイッチ201の構造を示す構成図である。2010はクロスバススイッチ201に接続するパス2020、2030、2050、およびクラスタ間I/F2040を接続するポートであるスイッチングポート(SWP)である。SWP2010はすべて同一の構造を有し、あるSWPから他のSWPへの転送経路のスイッチング制御を行う。図では1つのSWPについてのみ転送経路を示しているが、すべてのSWP間で同様の転送経路が存在する。

【0028】

図5は、ホストI/Fノード203の一構成例を示す構成図である。本実施形態では、具体的に説明をするためにホストI/F31とディスクアレイI/F21の両方にファイバチャネルを使用するものと仮定する。もちろんホストI/F31とディスクアレイI/F21として、ファイバチャネル以外のインタフェースを適用することも可能である。ホストI/Fノード203とディスクアレイI/Fノード202の両方に同一のインタフェースを使用することで、両者を同一構造にできる。本実施形態においては、ディスクアレイI/Fノード202も図に示すホストI/Fノード203と同様に構成される。以下では、ホストI/Fノード203を例に説明を行う。

20

【0029】

2021は受信したファイバチャネルフレーム(以下単にフレームと呼ぶ)をどのノードに転送するかを検索する検索プロセッサ(SP)、2022はホスト30(ディスクアレイI/Fノード202の場合は、ディスクアレイサブセット10)との間でフレームを送受信するインタフェースコントローラ(IC)、2023はIC2023が受信したフレームに対しSP2021が検索した結果に基づいて変換を施すスイッチングコントローラ(SC)、2024はSC2021が変換したフレームを他のノードに転送するためにクロスバススイッチ201を通過できる形式にパケット化するパケット生成部(SPG)、2025は受信したフレームを一時的に格納するフレームバッファ(FB)、2026は一つのホストからのディスクアレイアクセス要求コマンド(以下単にコマンドと呼ぶ)に対応した複数のフレーム列であるエクスチェンジ(Exchange)を識別するためのエクスチェンジ番号を管理するエクスチェンジテーブル(ET)、2027は複数のディスクアレイサブセット10の構成情報を格納するディスクアレイ構成管理テーブル(DCT)である。

30

【0030】

40

ディスクアレイスイッチ20の各構成部は、すべてハードウェアロジックで構成されることが性能上望ましい。しかし、求められる性能を満足できるならば、汎用プロセッサを用いたプログラム制御によりSP2021やSC2022の機能を実現することも可能である。

【0031】

各ディスクアレイサブセット10は、各々が有するディスクユニット104を1または複数の論理的なディスクユニットとして管理している。この論理的なディスクユニットを論理ユニット(LU)と呼ぶ。LUは、物理的なディスクユニット104と1対1で対応する必要はなく、1台のディスクユニット104に複数のLUが構成され、あるいは、複数のディスクユニット104で1つのLUが構成されても構わない。

50

【 0 0 3 2 】

ディスクアレイサブセット 1 0 の外部から見た場合、1 つの L U は、1 台のディスク装置として認識される。本実施形態では、ディスクアレイスイッチ 2 0 によりさらに論理的な L U が構成され、ホスト 3 0 は、この L U に対してアクセスするように動作する。本明細書では、1 つの L U でホスト 3 0 から認識される 1 つの L U が構成される場合、ホスト 3 0 により認識される L U を独立 L U (I L U)、複数の L U でホスト 3 0 から認識される 1 つの L U が構成される場合、ホスト 3 0 により認識される L U を統合 L U (C L U) と呼ぶ。

【 0 0 3 3 】

図 1 2 に、4 つのディスクアレイサブセットの L U で 1 つの統合 L U が構成される場合における各階層間でのアドレス空間の対応関係を示す。図において、1 0 0 0 は、一例として、ホスト “ # 2 ” からみたディスクアレイシステム 1 の 1 つの統合 L U におけるアドレス空間、1 1 0 0 は、ディスクアレイサブセット 1 0 の L U のアドレス空間、1 2 0 0 はディスクユニット 1 0 4 (ここでは、ディスクアレイサブセット “ # 0 ” についてのみ図示されている) のアドレス空間を示している。

【 0 0 3 4 】

各ディスクアレイサブセット 1 0 の L U は、ここでは、4 台のディスクユニット 1 0 4 により R A I D 5 (Redundant Arrays of Inexpensive Disks Level 5) 型ディスクアレイとして構成されるものとする。各ディスクアレイサブセット 1 0 は、それぞれ n_0 、 n_1 、 n_2 、 n_3 の容量を有する L U を持つ。ディスクアレイスイッチ 2 0 は、これら 4 つの L U の持つアドレス空間を $(n_0 + n_1 + n_2 + n_3)$ の容量を有するアドレス空間に統合し、ホスト 3 0 から認識される統合 L U を実現する。

【 0 0 3 5 】

本実施形態では、例えば、ホスト # 2 が領域 A 1 0 0 1 をアクセスする場合、領域 A 1 0 0 1 を指定したアクセス要求は、ディスクアレイスイッチ 2 0 によりディスクアレイサブセット # 0 の L U の領域 A ' 1 1 0 1 をアクセスするための要求に変換されてディスクアレイサブセット # 0 に転送される。ディスクアレイサブセット # 0 は、領域 A ' 1 1 0 1 をさらに、ディスクユニット 1 0 4 上の領域 A " 1 2 0 1 にマッピングしてアクセスを行う。アドレス空間 1 0 0 0 とアドレス空間 1 1 0 0 との間のマッピングは、ディスクアレイスイッチ 2 0 が有する D C T 2 0 7 に保持された構成情報に基づき行われる。この処理の詳細については後述する。なお、ディスクアレイサブセット内におけるマッピングについては、既によく知られた技術であり、本明細書では詳細な説明については省略する。

【 0 0 3 6 】

本実施形態において、D C T 2 0 7 は、システム構成テーブルとサブセット構成テーブルを含む。図 6 は、システム構成テーブルの構成を、図 7 は、サブセット構成テーブルの構成を示す。

【 0 0 3 7 】

図 7 に示すように、システム構成テーブル 20270 は、ホスト L U の構成を示す情報を保持するホスト L U 構成テーブル 20271、及びディスクアレイスイッチ 2 0 のディスクアレイ I / F ノード 2 0 2 とディスクアレイサブセット 1 0 との接続関係を示すディスクアレイ I / F ノード構成テーブル 20272 を有する。

【 0 0 3 8 】

ホスト L U 構成テーブル 20271 は、ホスト 3 0 からみた L U ごとに、その L U を識別する番号である Host-LU No.、L U の属性を示す LU Type、CLU Class、及び CLU Stripe Size、ホスト L U の状態を示す情報である Condition、ホスト L U を構成するディスクアレイサブセット 1 0 の L U に関する情報である L U 情報 (LU Info.) を有する。

【 0 0 3 9 】

LU Type は、このホスト L U が C L U であるか、I L U であるかといった L U の種類を示す情報である。CLU Class は、LU Type によりこのホスト L U が C L U であることが示される場合に、そのクラスが “Joined”、“mirrored”、及び “Striped” のいずれである

10

20

30

40

50

かを示す情報である。“Joined”は、図11により説明したように、いくつかのLUを連結して1つの大きな記憶空間を持つCLUが構成されていることを示す。“Mirrored”は、第6実施形態として後述するように、2つのLUにより二重化されたLUであることを示す。“Striped”は、第7実施形態として後述するように、複数のLUで構成され、データがこれら複数のLUに分散して格納されたLUであることを示す。CLU Stripe Sizeは、CLU Classにより「Striped」であることが示される場合に、ストライピングサイズ(データの分散の単位となるブロックのサイズ)を示す。

【0040】

Conditionにより示される状態には、“Normal”、“Warning”、“Fault”、及び“Not Defined”の4種類がある。“Normal”はこのホストLUが正常な状態であることを示す。
 “Warning”は、このホストLUを構成するLUに対応するいずれかのディスクユニットに障害が発生している等の理由により縮退運転が行われていることを示す。
 “Fault”は、ディスクアレイサブセット10の故障などによりこのホストLUを運転することができないことを示す。
 “Not Defined”は、対応するHost-LU No.のホストLUが定義されていないことを示す。

10

【0041】

LU Infoは、このホストLUを構成するLUについて、そのLUが属するディスクアレイサブセット10を特定する情報、ディスクアレイサブセット内でのLUN、及びそのサイズを示す情報を含む。ホストLUがILUの場合には、唯一のLUに関する情報が登録される。ホストLUがCLUの場合には、それを構成する全てのLUについて、それぞれのLUに関する情報が登録される。例えば、図において、Host-LU No.が“0”であるHost-LUは、ディスクアレイサブセット“#0”のLUN“0”、ディスクアレイサブセット“#1”のLUN“0”、ディスクアレイサブセット“#2”のLUN“0”、ディスクアレイサブセット“#3”のLUN“0”の4つのLUから構成されるCLUであり、そのCLUクラスが“Joined”であるCLUであることが分かる。

20

【0042】

ディスクアレイI/Fノード構成テーブル2072は、ディスクアレイI/F21が接続するディスクアレイサブセット10のポートごとに、どのディスクアレイスイッチ20のディスクアレイI/Fノード202が接続されるかを示す情報を保持する。

30

【0043】

具体的には、ディスクアレイサブセット10を特定するSubset No.、ポートを特定するSubset Port No.、そのポートに接続するディスクアレイスイッチ20を特定するSwitch No.、及びそのディスクアレイスイッチ20のディスクアレイI/Fノード202を特定するI/F Node No.を有する。ディスクアレイサブセット10が複数のポートを備えている場合には、そのポート毎に情報が設定される。

【0044】

サブセット構成テーブルは、図7に示すように、各ディスクアレイサブセット10に対応する複数のテーブル20720~20723を有する。各テーブルは、ディスクアレイサブセット10内で構築されたRAIDグループの構成を示す情報を保持するRAIDグループ構成テーブル20730と、ディスクアレイサブセット10内に構築されたLUの構成を示す情報を保持するLU構成テーブル20740を含む。

40

【0045】

RAIDグループ構成テーブル20730は、RAIDグループに付加された番号を示すGroup No.、そのRAIDグループのレベルを示すLevel、そのRAIDグループを構成するディスクの数を示す情報であるDisks、そのRAIDグループがRAIDレベル0、5等のストライピングされた構成の場合、そのストライプサイズを示すStripe Sizeを情報として含む。例えば、図に示されるテーブルにおいて、RAIDグループ“0”は、4台のディスクユニットにより構成されたRAIDグループであり、RAIDレベルが5、ストライプサイズがS0である。

50

【 0 0 4 6 】

L U 構成テーブル202740は、L U に付加された番号 (L U N) を示すLU No.、このL U がどのR A I D グループに構成されているのかを示すRAID Group、L U の状態を示すCondition、このL U のサイズ (容量) を示すSize、このL U がディスクアレイサブセット10のどのポートからアクセス可能なのかを示すPort、及びその代替となるポートを示すAlt. Portを情報として含む。Conditionで示される状態は、ホストL U についてのConditionと同様、“Normal”、“Warning”、“Fault”、“Not Defined”の4種類がある。Alt. Portに設定された情報により特定されるポートは、Portに設定された情報で特定されるポートに障害が発生したときに用いられるが、単に複数のポートから同一のL U をアクセスするために用いることもできる。

10

【 0 0 4 7 】

図8は、ファイバチャネルにおけるフレームの構成図である。ファイバチャネルのフレーム40は、フレームの先頭を示すS O F (Start Of Frame) 400、フレームヘッダ401、転送の実態データを格納する部位であるフレームペイロード402、32ビットのエラー検出コードであるC R C (Cyclic Redundancy Check) 403、フレームの最後尾を示すE O F (End Of Frame) 404を含む。フレームヘッダ401は、図9に示すような構造になっており、フレーム転送元のI D (S_ID)、フレーム転送先のI D (D_ID)、エキスチェンジの起動元、応答先が指定するそれぞれのエキスチェンジI D (OX_ID、RX_ID)、エキスチェンジ中のフレームグループを指定するシーケンスのI D (SEQ_ID) 等が格納されている。

20

【 0 0 4 8 】

本実施形態では、ホスト30により発行されるフレームには、S_IDとしてホスト30に割り当てられたI D が、また、D_IDとしてディスクアレイスイッチ20のポートに割り当てられたI D が使用される。一つのホストコマンドに対し、1ペアのエキスチェンジI D (OX_ID、RX_ID) が割り当てられる。複数のデータフレームを同一のエキスチェンジに対し発行する必要があるときは、その全データフレームに対して同一のSEQ_IDが割り当てられ、おのおのはシーケンスカウント (SEQ_CNT) で識別される。フレームペイロード402の最大長は2110バイトであり、フレーム種毎に格納される内容が異なる。例えば、後述するFCP_CMDフレームの場合、図10に示すように、S C S I のLogical Unit Number (L U N)、Command Description Block (C D B) 等が格納される。C D B は、ディスク (ディスクアレイ) アクセスに必要なコマンドバイト、転送開始論理アドレス (L B A)、転送長 (L E N) を含む。

30

【 0 0 4 9 】

以下、本実施形態のディスクアレイシステムの動作を説明する。

【 0 0 5 0 】

ディスクアレイシステムを使用するのに先立ち、ディスクアレイスイッチ20に対して、ディスクアレイサブセット10の構成情報を設定する必要がある。システム管理者は、管理端末5からディスクアレイシステム構成手段70を介して、すべてのディスクアレイサブセット10およびディスクアレイスイッチ20の構成設定情報を獲得する。管理者は、管理端末5から所望のシステム構成になるよう論理ユニットの構成設定、RAIDレベルの設定、障害発生時の交代パスの設定等、各種設定に必要な設定情報を入力する。ディスクアレイシステム構成管理手段70は、その設定情報を受け、各ディスクアレイサブセット10およびディスクアレイスイッチ20に設定情報を転送する。なお、管理端末5における設定情報の入力については第5実施形態にて別途説明する。

40

【 0 0 5 1 】

ディスクアレイスイッチ20では、通信コントローラ204が設定情報を獲得し、M P 200により各ディスクアレイサブセット10のアドレス空間情報等の構成情報が設定される。M P 200は、クロスバススイッチ201経由で各ホストI / F ノード203およびディスクアレイI / F ノード202に、ディスクアレイサブセット10の構成情報を配信する。

50

【 0 0 5 2 】

各ノード203、および202はこの情報を受信すると、SP2021により構成情報をDCT2027に格納する。ディスクアレイサブセット10では、ディスクアレイサブセット構成管理手段106が、設定情報を獲得し、共有メモリ102に格納する。各上位MPU1010および下位MPU1030は、共有メモリ102上の設定情報を参照し、各々の構成管理を実施する。

【 0 0 5 3 】

以下では、ホスト“ # 2 ”がディスクアレイシステム1に対し、リードコマンドを発行した場合の動作を説明する。図11に、ホストからのリード動作時にファイバチャネルを通して転送されるフレームのシーケンスを示す模式図を、図13にこのときのディスクアレイスイッチのホストI/Fノード203における動作のフローチャートを示す。

10

【 0 0 5 4 】

なお、以下の説明では、ホスト“ # 2 ”が、図12における記憶領域A1001をアクセスすることを仮定する。記憶領域A1001に対応する実際の記憶領域Aは、ディスクアレイサブセット“ # 0 ”のLUN = 0のLUを構成するディスクユニット#2のアドレス空間内に存在するものとする。また、アドレス空間1000を構成するLUを定義しているホストLU構成テーブル20271のLU Typeには「CLU」が、CLU Classには「Joined」が設定されているものとする。

【 0 0 5 5 】

データのリード時、ホスト30は、リードコマンドを格納したコマンドフレーム「FCP_CMD」をディスクアレイスイッチ20に発行する（図11矢印（a））。ディスクアレイスイッチ20のホストI/Fノード“ # 2 ”は、IC2023によりホストI/F31経由でコマンドフレーム「FCP_CMD」を受信する（ステップ20001）。IC2023は、SC2022にコマンドフレームを転送する。SC2022は、受け取ったコマンドフレームを一旦FB2025に格納する。この際、SC2022は、コマンドフレームのCRCを計算し、受信情報が正しいことを検査する。CRCの検査に誤りがあれば、SC2022は、その旨をIC2023に通知する。IC2023は、誤りの通知をSC2022から受けると、ホストI/F31を介してホスト30にCRCエラーを報告する。（ステップ20002）。

20

【 0 0 5 6 】

CRCが正しい場合、SC2022は、FB2025に保持したフレームをリードし、それがコマンドフレームであることを認識してフレームヘッダ401を解析する（ステップ20003）。そして、SC2022は、SP2021に指示し、S_ID、D_ID、OX_ID等のエクステンジ情報をET2026に登録する（ステップ20004）。

30

【 0 0 5 7 】

次に、SC2022は、フレームペイロード402を解析し、ホスト30により指定されたLUNおよびCDBを取得する（ステップ20005）。SP2021は、SC2022の指示により、DCT2027を検索し、ディスクアレイサブセット10の構成情報を得る。具体的には、SP2021は、ホストLU構成テーブル20271を検索し、受信したフレームペイロード402に格納されたLUNと一致するHost-LU No.を有する情報を見つける。SP2021は、LU Type、CLU Classに設定された情報からホストLUの構成を認識し、LU Info.に保持されている情報に基づきアクセスすべきディスクサブセット10とその中のLUのLUN、及びこのLU内でのLBAを判別する。次に、SP2021は、サブセット構成テーブル202720のLU構成テーブル202740を参照し、目的のディスクアレイサブセット10の接続ポートを確認し、ディスクアレイI/Fノード構成テーブル20272からそのポートに接続するディスクアレイI/Fノード202のノードNo.を得る。SP2021は、このようにして得たディスクアレイサブセット10を識別する番号、LUN、LBA等の変換情報をSC2022に報告する。（ステップ20006）。

40

【 0 0 5 8 】

次に、SC2022は、獲得した変換情報を使用しフレームペイロード402のLUN

50

とC D BのなかのL B Aを変換する。また、フレームヘッダ4 0 1のD_IDを対応するディスクアレイサブセット1 0のホストI / Fコントローラ1 0 1 1のD_IDに変換する。なお、この時点ではS_IDは書き換えない(ステップ20007)。

【0 0 5 9】

S C 2 0 2 2は、変換後のコマンドフレームと、対象ディスクアレイサブセット1 0に接続するディスクアレイI / Fノード番号を、S P G 2 0 2 4に転送する。S P G 2 0 2 4は、受け取った変換後のコマンドフレームに対し、図1 4に示すような簡単な拡張ヘッダ6 0 1を付加したパケットを生成する。このパケットをスイッチングパケット(S Packet) 6 0と呼ぶ。S Packet 6 0の拡張ヘッダ6 0 1には、転送元(自ノード)番号、転送先ノード番号、及び転送長が付加含まれる。S P G 2 0 2 4は、生成したS Packet 6 0をクロスバスイッチ2 0 1に送信する(ステップ20008)。

10

【0 0 6 0】

クロスバスイッチ2 0 1は、ホストI / Fノード“ # 2 ”と接続するS W P 2 0 1 0によりS Packet 6 0を受信する。S W P 2 0 1 0は、S Packet 6 0の拡張ヘッダ6 0 1を参照し、転送先のノードが接続するS W Pへのスイッチ制御を行って経路を確立し、S Packet 6 0を転送先のディスクアレイI / Fノード2 0 2(ここでは、ディスクアレイI / Fノード“ # 0 ”)に転送する。S W P 2 0 1 0は、経路の確立をS Packet 6 0の受信の度を実施し、S Packet 6 0の転送が終了したら、その経路を解放する。ディスクアレイI / Fノード“ # 0 ”では、S P G 2 0 2 4がS Packet 6 0を受信し、拡張ヘッダ6 0 1を外してコマンドフレームの部分をS C 2 0 2 2に渡す。

20

【0 0 6 1】

S C 2 0 2 2は、受け取ったコマンドフレームのフレームヘッダのS_IDに自分のI Dを書き込む。次にS C 2 0 2 2は、S P 2 0 2 1に対し、コマンドフレームのS_ID、D_ID、OX_ID等のエクスチェンジ情報、及びフレーム転送元ホストI / Fノード番号をE T 2 0 2 6に登録するよう指示し、I C 2 0 2 3にコマンドフレームを転送する。I C 2 0 2 3は、フレームヘッダ4 0 1の情報に従い、接続するディスクアレイサブセット1 0(ここでは、ディスクアレイサブセット“ # 0 ”)にコマンドフレームを転送する(図1 1矢印(b))。

【0 0 6 2】

ディスクアレイサブセット“ # 0 ”は、変換後のコマンドフレーム「FCP_CMD」をディスクアレイI / Fコントローラ1 0 1 1で受信する。上位M P U 1 0 1 0は、コマンドフレームのフレームペイロード4 0 2に格納されたL U NとC D Bを取得し、指定された論理ユニットのL B AからL E N長のデータをリードするコマンドであると認識する。

30

【0 0 6 3】

上位M P U 1 0 1 0は、共有メモリ1 0 2に格納されたキャッシュ管理情報を参照し、キャッシュヒットミス/ヒット判定を行う。ヒットすればキャッシュ1 0 2からデータ転送を実施する。ミスの場合、ディスクユニットからデータをリードする必要があるので、R A I D 5の構成に基づくアドレス変換を実施し、キャッシュ空間を確保する。そして、ディスクユニット2からのリード処理に必要な処理情報を生成し、下位M P U 1 0 3 0に処理を引き継ぐべく、共有メモリ1 0 2に処理情報を格納する。

40

【0 0 6 4】

下位M P U 1 0 3 0は、共有メモリ1 0 2に処理情報が格納されたことを契機に処理を開始する。下位M P U 1 0 3 0は、適切なディスクI / Fコントローラ1 0 3 1を特定し、ディスクユニット2へのリードコマンドを生成して、ディスクI / Fコントローラ1 0 3 1にコマンドを発行する。ディスクI / Fコントローラ1 0 3 1は、ディスクユニット2からリードしたデータをキャッシュ1 0 2の指定されたアドレスに格納して下位M P U 1 0 3 0に終了報告を通知する。下位M P U 1 0 3 0は、処理が正しく終了したことを上位M P U 1 0 1 0に通知すべく共有メモリ1 0 2に処理終了情報を格納する。

【0 0 6 5】

上位M P U 1 0 1 0は、共有メモリ1 0 2に処理終了情報が格納されたことを契機に処

50

理を再開し、ディスクアレイ I / F コントローラ 1 0 1 1 にリードデータ準備完了を通知する。ディスクアレイ I / F コントローラ 1 0 1 1 は、ディスクアレイスイッチ 2 0 の当該ディスクアレイ I / F ノード “ # 0 ” に対し、ファイバチャネルにおけるデータ転送準備完了フレームである「FCP_XFER_RDY」を発行する（図 1 1 矢印（c））。

【 0 0 6 6 】

ディスクアレイ I / F ノード “ # 0 ” では、データ転送準備完了フレーム「FCP_XFER_RDY」を受信すると、S C 2 0 2 2 が、ディスクアレイサブセット 2 0 から受信した応答先エクステンジ ID（RX_ID）を獲得し、S_ID、D_ID、OX_IDを指定して、S P 2 0 2 1 に指示し E T 2 0 2 6 の当該エクステンジ情報にRX_IDを登録する。S C 2 0 2 2 は、データ転送準備完了フレームの転送先（コマンドフレームの転送元）のホスト I / F ノード番号を獲得する。S C 2 0 2 2 は、このフレームのS_IDを無効化し、S P G 2 0 2 4 に転送する。S P G 2 0 2 4 は、先に述べたようにしてS Packetを生成し、クロスバスイッチ 2 0 1 経由で対象ホスト I / F ノード “ # 2 ” に転送する。

10

【 0 0 6 7 】

ホスト I / F ノード “ # 2 ” では、S P G 2 0 2 4 がデータ転送準備完了フレームのS Packetを受信すると、S Packetの拡張ヘッダを外し「FCP_XFER_RDY」を再生してS C 2 0 2 2 に渡す（ステップ20011）。S C 2 0 2 2 は、S P 2 0 2 1 に指示し E T 2 0 2 6 をサーチして該当するエクステンジを特定する（ステップ20012）。

【 0 0 6 8 】

次に、S C 2 0 2 2 は、フレームが「FCP_XFER_RDY」であるかどうか調べ（ステップ20013）、「FCP_XFER_RDY」であれば、E T 2 0 2 6 の応答先エクステンジ ID（RX_ID）の更新をS P 2 0 2 1 に指示する。応答先エクステンジ IDとしては、このフレームに付加されていた値が使用される（ステップ20014）。そして、S C 2 0 2 2 は、フレームヘッダ 4 0 1 のS_ID、D_IDをホスト I / F ノード 2 0 3 の ID とホスト 3 0 の ID を用いた適切な値に変換する（ステップ20015）。これらの処理によりフレームヘッダ 4 0 1 は、ホスト “ # 2 ” に対するフレームに変換される。I C 2 0 2 3 は、ホスト “ # 2 ” に対し、このデータ転送準備完了フレーム「FCP_XFER_RDY」を発行する（図 1 1 の矢印（d）：ステップ20016）。

20

【 0 0 6 9 】

ディスクアレイサブセット “ # 0 ” のディスクアレイ I / F コントローラ 1 0 1 1 は、データ転送を行うため、データフレーム「FCP_DATA」を生成し、ディスクアレイスイッチ 2 0 に転送する（図 1 1 矢印（e））。フレームペイロードの転送長には制限があるため、1 フレームで転送できる最大のデータ長は 2 K B である。データ長がこれを越える場合は、必要数だけデータフレームを生成し発行する。すべてのデータフレームには同一のSEQ_IDが割り当てられる。データフレームの発行は、同一のSEQ_IDに対し複数のフレームが生成されることを除き（すなわちSEQ_CNTが変化する）、データ転送準備完了フレームの場合と同様である。

30

【 0 0 7 0 】

ディスクアレイスイッチ 2 0 は、データ転送準備完了フレームの処理と同様に、データフレーム「FCP_DATA」のフレームヘッダ 4 0 1 の変換を実施する。ただし、データフレームの転送の場合、RX_IDが既に確立されているので、データ転送準備完了フレームの処理におけるステップ20014の処理はスキップされる。フレームヘッダ 4 0 1 の変換後、ディスクアレイスイッチ 2 0 は、ホスト “ # 2 ” にデータフレームを転送する（図 1 1 矢印（f））。

40

【 0 0 7 1 】

次に、ディスクアレイサブセット “ # 0 ” のディスクアレイ I / F コントローラ 1 0 1 1 は、終了ステータス転送を行うため、ステータスフレーム「FCP_RSP」を生成し、ディスクアレイスイッチ 2 0 に対し発行する（図 1 1 矢印（g））。ディスクアレイスイッチ 2 0 では、データ転送準備完了フレームの処理と同様に、S P G 2 0 2 4 がS Packetから拡張ヘッダを外し「FCP_RSP」ステータスフレームを再現し（ステップ20021）、S P 2 0

50

21によりET2026を検索しエクスチェンジ情報を獲得する(ステップ20022)。SC2022は、その情報に基づきフレームを変換する(ステップの20023)。変換されたフレームは、IC2023によりホスト“#2”に転送される(図11矢印(h):ステップ20024)。最後にSP2021は、ET2026からエクスチェンジ情報を削除する(ステップ20025)。

【0072】

以上のようにしてディスクアレイからのリード処理が行われる。ディスクアレイシステム1に対するライト処理についてもデータフレームの転送方向が逆転するのみで、上述したリード処理と同様の処理が行われる。

【0073】

図3に示したように、ディスクアレイスイッチ20は、クロスバスイッチ201にクラスタ間I/F2040を備えている。図1に示したシステム構成では、クラスタ間I/F2040は使用されていない。本実施形態のディスクアレイスイッチ20は、クラスタ間I/F2040を利用して図15に示すように、他のディスクアレイスイッチと相互に接続されることができる。

【0074】

本実施形態におけるディスクアレイスイッチ20単独では、ホスト30とディスクアレイサブセット10を合計8台までしか接続できないが、クラスタ間I/F2040を利用して複数のディスクアレイスイッチを相互接続し、接続できるホスト10とディスクアレイの数を増やすことができる。例えば、図15に示すシステムでは、4台のディスクアレイスイッチ20を使ってホスト30とディスクアレイサブセット10を合計32台まで接続でき、これらの間で相互にデータ転送が可能になる。

【0075】

このように、本実施形態では、ディスク容量や性能の必要性に合わせて、ディスクアレイサブセットやホストの接続台数を増加していくことができる。また、必要な転送帯域分のホストI/Fを用いてホスト-ディスクアレイシステム間を接続することができるので、容量、性能、接続台数の拡張性を大幅に向上させることができる。

【0076】

以上説明した実施形態によれば、1台のディスクアレイサブセットの性能が、内部のMPUや内部バスで制限されたとしても、複数のディスクアレイサブセットを用いて、ディスクアレイスイッチによりホストとディスクアレイサブセット間を相互接続することができる。これにより、ディスクアレイシステムトータルとして高い性能を実現することができる。ディスクアレイサブセットの性能が比較的低いものであっても、複数のディスクアレイサブセットを用いることで高性能化を実現できる。したがって、低コストのディスクアレイサブセットをコンピュータシステムの規模に合わせて必要な台数だけ接続することができ、規模に応じた適切なコストでディスクアレイシステムを構築することが可能となる。

【0077】

また、ディスク容量の増大や性能の向上が必要になったときは、ディスクアレイサブセットを必要なだけ追加すればよい。さらに、複数のディスクアレイスイッチを用いて任意の数のホスト及びディスクアレイサブセットを接続できるので、容量、性能、接続台数のいずれをも大幅に向上させることができ、高い拡張性を有するシステムが実現できる。

【0078】

さらにまた、本実施形態によれば、ディスクアレイサブセットとして、従来のディスクアレイシステムそのものの縮小機を用いることができるので、既に開発した大規模な制御ソフトウェア資産をそのまま利用でき、開発コストの低減と開発期間の短縮を実現することができる。

[第2実施形態]

図16は、本発明の第2の実施形態におけるコンピュータシステムの構成図である。本実施形態は、ディスクアレイスイッチのホストI/Fノードにおいて、フレームヘッダ4

10

20

30

40

50

01のみを変換し、フレームペイロード402は操作しない点、及び、ディスクアレイスイッチ、ホストI/F、ディスクアレイI/Fが二重化されていない点で第1実施形態と構成上相違する。したがって、各部の構成は、第1実施形態と大きく変わるところがなく、その詳細については説明を省略する。

【0079】

図16において、各ディスクアレイサブセット10は、複数の論理ユニット(LU)110で構成されている。各LU110は、独立LUとして構成される。一般に、各ディスクアレイサブセット10内のLU110に割り当てられるLUNは、0から始まる連続番号である。このため、ホスト30に対して、ディスクアレイシステム1内のすべてのLU110のLUNを連続的に見せる場合には、第1実施形態と同様に、フレームペイロード402のLUNフィールドを変換する必要がある。本実施形態では、各ディスクアレイサブセット10のLUNをそのままホスト30に見せることで、フレームペイロード402の変換を不要とし、ディスクアレイスイッチの制御を簡単なものとしている。

【0080】

本実施形態のディスクアレイスイッチ20は、ホストI/Fノード203ごとに特定のディスクアレイサブセット10をアクセスできるものと仮定する。この場合、一つのホストI/F31を使うと、1台のディスクアレイサブセット10にあるLU110のみがアクセス可能である。1台のホストから複数のディスクアレイサブセット10のLU110をアクセスしたい場合には、そのホストを複数のホストI/Fノード203に接続する。また、複数のホスト30から1台のディスクアレイサブセット10のLU110をアクセスできるようにする場合は、同一のホストI/Fノード203にループトポロジや、ファブリックトポロジ等を用い、複数のホスト30を接続する。このように構成すると、1台のホスト30から1つのLU110をアクセスする際に、ホストI/Fノード203のD_ID毎にディスクアレイサブセット10が確定することになるため、各LUのLUNをそのままホスト30に見せることが可能である。

【0081】

本実施形態では、上述した理由により、ホスト30に、各ディスクアレイサブセット10内のLU110のLUNをそのままホスト30に見せているため、ディスクアレイスイッチ20におけるLUNの変換は不要となる。このため、ディスクアレイスイッチ20は、ホスト30からフレームを受信すると、フレームヘッダ401のみを第1実施例と同様にして変換し、フレームペイロード402は変換せずにディスクアレイサブセット10に転送する。本実施形態における各部の動作は、フレームペイロード402の変換が行われないことを除くと第1実施形態と同様であるので、ここでは詳細な説明を省略する。本実施形態によれば、ディスクアレイスイッチ20の開発を容易にできる。

〔第3実施形態〕

第2実施形態では、ディスクアレイスイッチのホストI/Fノードにおいて、フレームヘッダのみを変換しているが、以下に説明する第3実施形態ではフレームヘッダも含め、フレームの変換を行わない形態について説明する。本実施形態のコンピュータシステムは、図1に示す第1実施形態におけるコンピュータシステムと同様に構成される。

【0082】

第1、および第2実施形態では、ホスト30に対し、ディスクアレイサブセット10の台数や、LU110の構成等、ディスクアレイシステム1の内部構成を隠蔽している。このため、ホスト30からはディスクアレイシステム1が全体で1つの記憶装置として見える。これに対し、本実施形態では、ディスクアレイサブセット10をそのままホスト30に公開し、ホスト30がフレームヘッダのD_IDとして直接ディスクアレイサブセットのポートのIDを使うようにする。これにより、ディスクアレイスイッチは、フレームヘッダの情報に従ってフレームの転送を制御するだけで済み、従来技術におけるファイバチャネルのファブリック装置と同等のスイッチ装置をディスクアレイスイッチ20に替えて利用することができる。

【0083】

ディスクアレイシステム構成管理手段 70 は、ディスクアレイサブセット 10 の通信コントローラ 106、及びディスクアレイスイッチ 20 の通信手段 204 と通信して各ディスクアレイサブセット 10 及びディスクアレイスイッチ 20 の構成情報を獲得し、あるいは、設定する。

【0084】

ディスクアレイスイッチ 20 は、基本的には図 3 に示す第 1 実施形態におけるディスクアレイスイッチと同様の構成を有する。しかし、本実施形態では、ホスト 30 が発行するフレームのフレームヘッダの情報をそのまま使ってフレームの転送を制御するため、第 1 実施形態、あるいは第 2 実施形態でディスクアレイスイッチ 20 のホスト I/F ノード 203、ディスクアレイ I/F ノード 202 が有する DCT 2027 や、SC 2022、SPG 2024 等により実現されるフレームヘッダ等の変換の機能は不要となる。ディスクアレイスイッチ 20 が有するクロスバススイッチ 201 は、フレームヘッダの情報に従ってホスト I/F ノード 203、及びディスクアレイ I/F ノード 202 の間でファイバチャネルのフレームの転送を行う。

【0085】

本実施形態では、ディスクアレイシステムの構成をディスクアレイシステム構成管理手段 70 で一括して管理するために、ディスクアレイ管理用テーブル（以下、このテーブルも DCT と呼ぶ）をディスクアレイシステム構成管理手段 70 に備える。ディスクアレイシステム構成管理手段 70 が備える DCT は、図 6、7 に示す、システム構成テーブル 20270 とサブセット構成テーブル 202720 ~ 202723 の 2 つのテーブル群を含む。なお、本実施形態では、ホスト LU は全て I LU として構成されるため、ホスト LU 構成テーブル 20271 の LU Type は全て「I LU」となり、CLU Class、CLU Stripe Size は意味をなさない。

【0086】

管理者は、管理端末 5 を操作してディスクアレイシステム構成管理手段 70 と通信し、ディスクアレイサブセット 10 のディスク容量、ディスクユニットの台数等の情報を得て、ディスクアレイサブセット 10 の LU 110 の設定、RAID レベルの設定等を行う。次に管理者は、管理端末 5 によりディスクアレイシステム構成管理手段 70 と通信し、ディスクアレイスイッチ 20 を制御して、各ホスト 30 とディスクアレイサブセット 20 間の関係情報を設定する。

【0087】

以上の操作により、ディスクアレイシステム 1 の構成が確立し、ホスト 30 から管理者が望む通りに LU 110 が見えるようになる。ディスクアレイ構成管理手段 70 は以上の設定情報を保存し、管理者からの操作に応じ構成の確認や、構成の変更を行うことができる。

【0088】

本実施形態によれば、ひとたびディスクアレイシステム 1 を構成すれば、管理者からディスクアレイスイッチ 20 の存在を認識させることが無く、複数のディスクアレイサブシステムを 1 台のディスクアレイシステムと同様に扱うことができる。また、本実施形態によれば、ディスクアレイスイッチ 20 とディスクアレイサブセット 10 は、同一の操作環境によって統一的に操作することができ、その構成確認や、構成変更も容易になる。さらに、本実施形態によれば、従来使用していたディスクアレイシステムを本実施形態におけるディスクアレイシステムに置き換える場合に、ホスト 30 の設定を変更することなく、ディスクアレイシステム 1 の構成をそれまで使用していたディスクアレイシステムの構成に合わせることができ、互換性を維持できる。

〔第 4 実施形態〕

以上説明した第 1 から第 3 の実施形態では、ホスト I/F にファイバチャネルを使用している。以下に説明する実施形態では、ファイバチャネル以外のインタフェースが混在した形態について説明する。

【0089】

図 17 は、ホスト I/F がパラレル SCSI である場合のホスト I/F ノード 203 内

10

20

30

40

50

部の I C 2 0 2 3 の一構成例を示す。20230はパラレル S C S I のプロトコル制御を行う S C S I プロトコルコントローラ (S P C)、20233はファイバチャネルのプロトコル制御を行うファイバチャネルプロトコルコントローラ (F P C)、20231はパラレル S C S I とファイバチャネルのシリアル S C S I をプロトコル変換するプロトコル変換プロセッサ (P E P)、20232はプロトコル変換中データを一時保存するバッファ (B U F) である。

【 0 0 9 0 】

本実施形態において、ホスト 3 0 は、ディスクアレイ I / F ノード 2 0 3 に対して S C S I コマンドを発行する。リードコマンドの場合、S P C 20230は、これを B U F 20232に格納し、P E P 20231に割り込みでコマンドの受信を報告する。P E P 20231は、B U F 20232に格納されたコマンドを利用し、F P C 20233へのコマンドに変換し、F P C 20233に送る。F P C 20233は、このコマンドを受信すると、フレーム形式に変換し、S C 2 0 2 2 に引き渡す。この際、エクスチェンジ I D、シーケンス I D、ソース I D、デスティネーション I D は、以降の処理が可能ないように P E P 20231により付加される。あとのコマンド処理は、第 1 実施形態と同様に行われる。

【 0 0 9 1 】

ディスクアレイサブセット 1 0 は、データの準備が完了すると、データ転送準備完了フレームの発行、データ転送、正常終了後ステータスフレームの発行を実施する。ディスクアレイサブセット 1 0 から I C 2 0 2 3 までの間では、フレームヘッダ 4 0 1 やフレームペイロード 4 0 2 が必要に応じ変換されながら、各種フレームの転送が行われる。I C 2 0 2 3 の F P C 20233は、データ転送準備完了フレームを受信し、続いてデータを受信して B U F 20232に格納し、続けて正常に転送が終わったならば、ステータスフレームを受信し、P T P 20231に割り込みをかけてデータの転送完了を報告する。P T P 20231は、割り込みを受けると、S P C 20230を起動し、ホスト 3 0 に対しデータ転送を開始するよう指示する。S P C 20230はホスト 3 0 にデータを送信し、正常終了を確認すると P T P 20231に対し割り込みで正常終了を報告する。

【 0 0 9 2 】

ここでは、ファイバチャネル以外のホスト I / F の例としてパラレル S C S I を示したが、他のインタフェース、例えば、メインフレームへのホスト I / F である ESCON 等に対しても同様に適用することが可能である。ディスクアレイスイッチ 2 0 のホスト I / F ノード 2 0 3 として、例えば、ファイバチャネル、パラレル S C S I、及び ESCON に対応したホスト I / F ノードを設けることで、1 台のディスクアレイシステム 1 に、メインフレームと、パーソナルコンピュータ、ワークステーション等のいわゆるオープンシステムの両方を混在させて接続することが可能である。本実施形態では、ディスクアレイ I / F としては、第 1 から第 3 実施形態と同様、ファイバチャネルを用いているが、ディスクアレイ I / F に対しても任意の I / F を使用することが可能である。

〔 第 5 実施形態 〕

次に、ディスクアレイシステム 1 の構成管理の方法について、第 5 実施形態として説明する。図 1 8 は、本実施形態のシステム構成図である。本実施形態では、ホスト 3 0 が 4 台設けられている。ホスト “ # 0 ”、“ # 1 ”とディスクアレイシステム 1 の間の I / F 3 0 はファイバチャネル、ホスト “ # 2 ”とディスクアレイシステム 1 の間は、パラレル S C S I (Ultra S C S I)、ホスト “ # 3 ”とディスクアレイシステム 1 の間は、パラレル S C S I (Ultra2 S C S I) で接続されている。

【 0 0 9 3 】

パラレル S C S I のディスクアレイスイッチ 2 0 への接続は第 4 実施形態と同様に行われる。ディスクアレイシステム 1 は、4 台のディスクアレイサブセット 3 0 を有する。ディスクアレイサブセット “ # 0 ”には 4 つの独立 L U、ディスクアレイサブセット “ # 1 ”には 2 つの独立 L U がそれぞれ構成されている。ディスクアレイサブセット “ # 2 ”と “ # 3 ”で 1 つの統合 L U が構成されている。本実施形態では、第 1 実施形態と同様、ホスト 3 0 に対しディスクアレイサブセット 1 0 を隠蔽し、ファイバチャネルのフレームを

変換するものとする。各LUに割り当てられるLUNは、ディスクアレイサブセット“#0”のLUから順に、LUN=0、1、2、・・・6までの7つである。

【0094】

図19は、管理端末5の表示画面上に表示される画面の一例である。図は、ホストI/F31と各論理ユニット(LU)との対応を示した論理接続構成画面である。

【0095】

論理接続構成画面50には、各ホストI/F31に関する情報3100、各LU110に関する情報11000、ディスクアレイサブセット10とLU110の関係等が表示される。ホストI/F31に関する情報としては、I/F種類、I/F速度、ステータス等が含まれる。LU110に関する情報としては、格納サブセット番号、LUN、容量、RAIDレベル、ステータス、情報、等が表示される。管理者はこの画面を参照することで、容易にディスクアレイシステム1の構成を管理することができる。

10

【0096】

論理接続構成画面50上で、ホストI/FとLUの間に引かれている線は、各ホストI/F31を経由してアクセス可能なLU110を示している。ホストI/Fから線の引かれていないLU110に対して、そのホストI/Fに接続するホスト30からはアクセスできない。ホスト30によって、扱うデータ形式が異なり、また使用者も異なることから、セキュリティ維持上、適切なアクセス制限を設けることが不可欠である。そこで、システムを設定する管理者が、この画面を用いて、各LU110とホストI/Fとの間のアクセス許可をあたえるか否かによって、アクセス制限を実施する。図において、例えば、LU“#0”は、ホストI/F“#0”および“#1”からアクセス可能であるが、ホストI/F“#2”、“#3”からはアクセスできない。LU“#4”は、ホストI/F“#2”からのみアクセス可能である。

20

【0097】

このようなアクセス制限を実現するためアクセス制限情報は、ディスクアレイシステム構成管理手段70からディスクアレイスイッチ20に対して送信される。ディスクアレイスイッチ20に送られたアクセス制限情報は、各ホストI/Fノード203に配信され、各ホストI/Fノード203のDCT2027に登録される。ホストにより、アクセスが制限されたLUに対するLU存在有無の検査コマンドが発行された場合、各ホストI/Fノード203は、DCT2027の検査を行い、検査コマンドに対し応答しないか、あるいは、エラーを返すことで、そのLUは、ホストからは認識されなくなる。LU存在有無の検査コマンドとしては、SCSIプロトコルの場合、Test Unit Readyコマンドや、Inquiryコマンドが一般に用いられる。この検査なしに、リード/ライトが実施されることはないため、容易にアクセスの制限をかけることが可能である。

30

【0098】

本実施形態ではホストI/F31毎にアクセス制限をかけているが、これを拡張することで、ホスト30毎にアクセス制限をかけることも容易に実現できる。また、ホストI/F31、ホスト30、あるいは、アドレス空間を特定して、リードのみ可、ライトのみ可、リード/ライトとも可、リード/ライトとも不可といった、コマンドの種別に応じたアクセス制限をかけることもできる。この場合、アクセス制限情報としてホストI/F番号、ホストID、アドレス空間、制限コマンド等を指定してディスクアレイスイッチ20に制限を設定する。

40

【0099】

次に、新たなディスクアレイサブセット10の追加について説明する。ディスクアレイサブセット10を新規に追加する場合、管理者は、ディスクアレイスイッチ20の空いているディスクアレイI/Fノード202に追加するディスクアレイサブセット10を接続する。つづけて、管理者は、管理端末5を操作し、論理接続構成画面50に表示されている「最新状態を反映」ボタン5001を押下する。この操作に応答して、未設定のディスクアレイサブセットを表す絵が画面上に表示される(図示せず)。このディスクアレイサブセットの絵が選択されるすると、ディスクアレイサブセットの設定画面が現れる。管理

50

者は、表示された設定画面上で、新規に追加されたディスクアレイサブセットの各種設定を実施する。ここで設定される項目にはLUの構成、RAIDレベル等がある。続けて、図19の論理接続構成図の画面に切り替えると、新規ディスクアレイサブセットとLUが現れる。以降、ホストI/F31毎に対するアクセス制限を設定し、「設定実行」ボタン5002を押下すると、ディスクアレイスイッチ20に対し、アクセス制限情報、およびディスクアレイサブセット、LUの情報が転送され、設定が実行される。

【0100】

各ディスクアレイサブセット10にLU110を追加する際の手順も上述した手順で行われる。また、ディスクアレイサブセット、およびLUの削除についてもほぼ同様の手順で行われる。異なる点は、管理者が各削除部位を画面上で選択して「削除」ボタン5003を押下し、適切な確認が行われたのち、実行される点である。以上のように、管理端末70を用いることで、管理者はディスクアレイシステム全体を一元的に管理できる。

〔第6実施形態〕

次に、ディスクアレイスイッチ20によるミラーリングの処理について、第6実施形態として説明する。ここで説明するミラーリングとは、2台のディスクアレイサブセットの2つの独立LUにより二重書きをサポートする方法であり、ディスクアレイサブセットのコントローラまで含めた二重化である。従って、信頼性は、ディスクのみの二重化とは異なる。

【0101】

本実施形態におけるシステムの構成は図1に示すものと同じである。図1に示す構成において、ディスクアレイサブセット“#0”と“#1”は全く同一のLU構成を備えており、この2つのディスクアレイサブセットがホスト30からは1つのディスクアレイとして見えるものとする。便宜上、ミラーリングされたディスクアレイサブセットのペアの番号を“#01”と呼ぶ。また、各ディスクアレイサブセットのLU“#0”とLU“#1”によってミラーリングペアが形成され、このLUのペアを便宜上、LU“#01”と呼ぶ。DCT2027のホストLU構成テーブル20271上でLU#01を管理するための情報は、CLU Classに「Mirrored」が設定され、LU Info.として、LU#0とLU#1に関する情報が設定される。その他の各部の構成は第1実施形態と同様である。

【0102】

本実施形態における各部の動作は、第1実施例とほぼ同様である。以下、第1実施形態と相違する点について、ディスクアレイスイッチ20のホストI/Fノード203の動作を中心に説明する。図20は、本実施形態におけるライト動作時に転送されるフレームのシーケンスを示す模式図、図21、22は、ライト動作時におけるホストI/Fノード203による処理の流れを示すフローチャートである。

【0103】

ライト動作時、ホスト30が発行したライトコマンドフレーム（FCP_CMD）は、IC2023により受信される（図20の矢印（a）：ステップ21001）。IC2023により受信されたライトコマンドフレームは、第1実施形態で説明したリード動作時におけるステップ20002 20005と同様に処理される（ステップ21002 - 21005）。

【0104】

SC2022は、SP2021を使ってDCT2027を検索し、ミラー化されたディスクアレイサブセット“#01”のLU“#01”へのライトアクセス要求であることを認識する（ステップ21006）。SC2022は、FB2025上に、受信したコマンドフレームの複製を作成する（ステップ21007）。SC2022は、DCT2027に設定されている構成情報に基づいてコマンドフレームの変換を行い、LU“#0”とLU“#1”の両者への別々のコマンドフレームを作成する（ステップ21008）。ここで、LU“#0”を主LU、LU“#1”を従LUと呼び、コマンドフレームにもそれぞれ主コマンドフレーム、従コマンドフレームと呼ぶ。そして、両者別々にET2026にエクステンジ情報を格納し、ディスクアレイサブセット“#0”およびディスクアレイサブセット“#1”に対し作成したコマンドフレームを発行する（図20の矢印（b0）（b1）：ス

10

20

30

40

50

テップ21009)。

【0105】

各ディスクアレイサブセット“#0”、“#1”は、コマンドフレームを受信し、それぞれ独立にデータ転送準備完了フレーム(FCP_XFER_RDY)をディスクアレイスイッチ20に送信する(図20の矢印(c0)(c1))。ディスクアレイスイッチ20では、ホストI/Fノード203が、第1実施形態におけるリード動作のステップ20011 20013と同様の処理により転送されてきたデータ転送準備完了フレームを処理する(ステップ21011 - 21013)。

【0106】

各ディスクアレイサブセットからのデータ転送準備完了フレームがそろった段階で(ステップ21014)、SC2022は、主データ転送準備完了フレームに対する変換を実施し(ステップ21015)、IC2023により変換後のフレームをホスト30に送信する(図20の矢印(d):ステップ21015)。

【0107】

ホスト30は、データ転送準備完了フレームを受信した後、ライトデータ送信のため、データフレーム(FCP_DATA)をディスクアレイスイッチ20に送信する(図20の矢印(e))。ホスト30からのデータフレームは、IC2023により受信されると(ステップ21031)、リードコマンドフレームやライトコマンドフレームと同様に、FB2025に格納され、CRC検査、フレームヘッダの解析が行われる(ステップ21032、21033)。フレームヘッダの解析結果に基づき、ET2026がSP2021により検索され、エクスチェンジ情報が獲得される(ステップ21034)。

【0108】

SC2022は、ライトコマンドフレームのときと同様に複製を作成し(ステップ21035)、その一方をディスクアレイサブセット“#0”内のLU“#0”に、他方をディスクアレイサブセット“#1”内のLU“#1”に向けて送信する(図20の矢印(f0)(f1):ステップ21037)。

【0109】

ディスクアレイサブセット“#0”、“#1”は、各々、データフレームを受信し、ディスクユニット104に対しそれぞれライトし、ステータスフレーム(FCP_RSP)をディスクアレイスイッチ20に送信する。

【0110】

SC2022は、ディスクアレイサブセット“#0”、“#1”それぞれからステータスフレームを受信すると、それらのステータスフレームから拡張ヘッダを外してフレームヘッダを再現し、ET2026からエクスチェンジ情報を獲得する(ステップ21041、21042)。

【0111】

ディスクアレイサブセット“#0”、“#1”の両者からのステータスフレームが揃うと(ステップ21043)、ステータスが正常終了であることを確認のうえ、LU“#0”からの主ステータスフレームに対する変換を行い(ステップ21044)、従ステータスフレーム消去する(ステップ21045)。そして、IC2023は、正常終了を報告するためのコマンドフレームをホストに送信する(図20の矢印(h):ステップ21046)。最後にSP2021は、ET2026のエクスチェンジ情報を消去する(ステップ21047)。

【0112】

以上でミラーリング構成におけるライト処理が終了する。ミラーリングされたLU“#01”に対するリード処理は、データの転送方向が異なるだけで、上述したライト処理とほぼ同様に行われるが、ライトとは異なり、2台のディスクアレイサブセットにリードコマンドを発行する必要はなく、どちらか一方に対してコマンドフレームを発行すればよい。たとえば、常に主LUに対してコマンドフレームを発行してもよいが、高速化のため、主/従双方のLUに対して、交互にコマンドフレームを発行するなどにより、負荷を分散すると有効である。

10

20

30

40

50

【 0 1 1 3 】

上述した処理では、ステップ21014、及びステップ21043で2台のディスクアレイサブセット“ # 0 ”、“ # 1 ”の応答を待ち、両者の同期をとって処理が進められる。このような制御では、双方のディスクアレイサブセットでの処理の成功が確認されてから処理が進むため、エラー発生時の対応が容易になる。その一方で、全体の処理速度が、どちらか遅いほうの応答に依存してしまうため、性能が低下するという欠点がある。

【 0 1 1 4 】

この問題を解決するため、ディスクアレイスイッチにおいて、ディスクアレイサブセットの応答を待たずに次の処理に進んだり、ディスクアレイサブセットのどちらか一方からの応答があった時点で次の処理に進む「非同期型」の制御をすることも可能である。非同期型の制御を行った場合のフレームシーケンスの一例を、図20において破線矢印で示す。

10

【 0 1 1 5 】

破線矢印で示されるフレームシーケンスでは、ステップ21016で行われるホストへのデータ転送準備完了フレームの送信が、ステップ21009の処理の後、ディスクアレイサブセット10からのデータ転送準備完了フレームを待たずに実施される。この場合、ホストに送信されるデータ転送準備完了フレームは、ディスクアレイスイッチ20のSC2022により生成される（破線矢印（d'））。

【 0 1 1 6 】

ホスト30からは、破線矢印（e'）で示されるタイミングでデータフレームがディスクアレイスイッチ20に転送される。ディスクアレイスイッチ20では、このデータフレームが一旦FB2025に格納される。SC2022は、ディスクアレイサブセット10からのデータ転送準備完了フレームの受信に応答して、データ転送準備完了フレームが送られてきたディスクアレイサブセット10に対し、FB2025に保持されたデータフレームを転送する（破線矢印（f0'）、（f1'））。

20

【 0 1 1 7 】

ディスクアレイスイッチ20からホスト30への終了報告は、双方のディスクアレイサブシステム10からの報告（破線矢印（g0'）、（g0'））があった時点でおこなわれる（破線矢印（h'））。このような処理により、図20に示される時間Taの分だけ処理時間を短縮することが可能である。

30

【 0 1 1 8 】

ディスクアレイスイッチ20とディスクアレイサブセット10間のフレーム転送の途中でエラーが発生した場合、以下の処理が実施される。

【 0 1 1 9 】

実行中の処理がライト処理の場合、エラーが発生したLUに対し、リトライ処理が行われる。リトライが成功すれば、処理はそのまま継続される。あらかじめ設定された規定の回数のリトライが失敗した場合、ディスクアレイスイッチ20は、このディスクアレイサブセット10（もしくはLU）に対するアクセスを禁止し、そのことを示す情報をDC2027に登録する。また、ディスクアレイスイッチ20は、MP200、通信コントローラ204を経由して、ディスクシステム構成手段70にそのことを通知する。

40

【 0 1 2 0 】

ディスクシステム構成手段70は、この通知に응答して管理端末5にアラームを発行する。これにより管理者は、トラブルが発生したことを認識できる。その後、ディスクアレイスイッチ20は、正常なディスクアレイサブセットを用いて運転を継続する。ホスト30は、エラーが発生したことを認識することなく、処理を継続できる。

【 0 1 2 1 】

本実施形態によれば、2台のディスクアレイサブシステムでミラー構成を実現できるので、ディスクの耐障害性を上げることができ、また、ディスクアレイコントローラ、ディスクアレイI/F、及びディスクアレイI/Fノードの耐障害性を上げることができ、内部バスの二重化等することなくディスクアレイシステム全体の信頼性を向上させる

50

ことができる。

【第7実施形態】

次に、3台以上のディスクアレイサブセット10を統合し、1台の論理的なディスクアレイサブセットのグループを構成する方法について説明する。本実施形態では、複数のディスクアレイサブセット10にデータを分散して格納する。これにより、ディスクアレイサブセットへのアクセスを分散させ、特定のディスクアレイサブセットへのアクセスの集中を抑止することで、トータルスループットを向上させる。本実施形態では、ディスクアレイスイッチによりこのようなストライピング処理を実施する。

【0122】

図23は、本実施形態におけるディスクアレイシステム1のアドレスマップである。ディスクアレイサブセット10のアドレス空間は、ストライプサイズSでストライピングされている。ホストから見たディスクアレイシステム1のアドレス空間は、ストライプサイズS毎に、ディスクアレイサブセット“#0”、“#1”、“#2”、“#3”に分散されている。ストライプサイズSのサイズは任意であるが、あまり小さくない方がよい。ストライプサイズSが小さすぎると、アクセスすべきデータが複数のストライプに属するストライプまたぎが発生したときに、その処理にオーバーヘッドが発生するおそれがある。ストライプサイズSを大きくすると、ストライプまたぎが発生する確率が減少するので性能向上のためには好ましい。LUの数は任意に設定することができる。

【0123】

以下、本実施形態におけるホストI/Fノード203の動作について、図24に示す動作フローチャートを参照しつつ第1実施形態との相違点に着目して説明する。なお、本実施形態では、DCT2027のホストLU構成テーブル20271上で、ストライピングされたホストLUに関する情報のCLU Classには「Striped」が、CLU Stripe Sizeにはストライプサイズ「S」が設定される。

【0124】

ホスト30がコマンドフレームを発行すると、ディスクアレイスイッチ20は、ホストI/Fノード203のIC2023でこれを受信する(ステップ22001)、SC2022は、IC2023からこのコマンドフレームを受け取り、SP2021を使ってDCT2027を検索し、ストライピングする必要があることを認識する(ステップ22005)。

【0125】

次に、SC2022は、SP2021によりDCT2027を検索し、ストライプサイズSを含む構成情報から、アクセスの対象となるデータが属するストライプのストライプ番号を求め、このストライプがどのディスクアレイサブセット10に格納されているか特定する(ステップ22006)。この際、ストライプまたぎが発生する可能性があるが、この場合の処理については後述する。ストライプまたぎが発生しない場合、SP2021の計算結果に基づき、SC2022はコマンドフレームに対し変換を施し(ステップ22007)、エクステンジ情報をET2026に格納する(ステップ22008)。以降は、第1実施形態と同様の処理が行われる。

【0126】

ストライプまたぎが発生した場合、SP2021は、2つのコマンドフレームを生成する。この生成は、例えば、ホスト30が発行したコマンドフレームを複製することで行われる。生成するコマンドフレームのフレームヘッダ、フレームペイロード等は、新規に設定する。第6実施形態と同様、SC2022でコマンドフレームの複製を作成した後、変換を実施することも可能であるが、ここでは、SP2021により新規に作成されるものとする。SC2022は、2つのコマンドフレームが生成されると、これらを各ディスクアレイサブセット10に送信する。

【0127】

この後、第1実施形態と同様にデータ転送が実施される。ここで、本実施形態では、第1実施形態、あるいは第6実施形態と異なり、データ自体を1台のホスト30と2台のディスクアレイサブセット10間で転送する必要がある。たとえば、リード処理の場合、2

10

20

30

40

50

台のディスクアレイサブセット10から転送されるデータフレームは、すべてホスト30に転送する必要がある。この際SC2022は、各ディスクアレイサブセット10から転送されてくるデータフレームに対し、ET2026に登録されたエクスチェンジ情報に従い、適切な順番で、適切なエクスチェンジ情報を付加してホスト30に送信する。

【0128】

ライト処理の場合は、コマンドフレームの場合と同様、2つのデータフレームに分割して、該当するディスクアレイサブセット10に転送する。なお、データフレームの順序制御は、ホスト、あるいはディスクアレイサブセットがアウトオブオーダー（Out of Order）機能と呼ばれる、順不同処理に対応しているならば必須ではない。

【0129】

最後に、すべてのデータ転送が完了し、ディスクアレイスイッチ20が2つのステータスフレームをディスクアレイサブセット10から受信すると、SP2021（あるいはSC2022）は、ホスト30へのステータスフレームを作成し、これをIC2023によりホスト30に送信する。

【0130】

本実施形態によれば、アクセスを複数のディスクアレイサブセットに分散することができるので、トータルとしてスループットを向上させることができるとともに、アクセスレイテンシも平均的に低減させることが可能である。

[第8実施形態]

次に、2台のディスクアレイシステム（またはディスクアレイサブセット）間における複製の作成について、第8実施形態として説明する。ここで説明するようなシステムは、2台のディスクアレイシステムの一方を遠隔地に配置し、天災等による他方のディスクアレイシステムの障害に対する耐性を備える。このような災害に対する対策をディザスタリカバリと呼び、遠隔地のディスクアレイシステムとの間で行われる複製の作成のことをリモートコピーと呼ぶ。

【0131】

第6実施形態で説明したミラーリングでは、地理的にほぼ同一の場所に設置されたディスクアレイサブセット10でミラーを構成するので、ディスクアレイE/F21はファイバチャネルでよい。しかし、リモートコピーを行うディスクアレイ（ディスクアレイサブセット）が10kmを越える遠隔地に設置される場合、中継なしでファイバチャネルによりフレームを転送する事ができない。ディザスタリカバリに用いられる場合、お互いの間の距離は通常数百km以上となる、このため、ファイバチャネルでディスクアレイ間を接続することは実用上不可能であり、ATM（Asynchronous Transfer Mode）等による高速公衆回線や衛星通信等が用いられる。

【0132】

図25は、本実施形態におけるディザスタリカバリシステムの構成例である。

【0133】

81はサイトA、82はサイトBであり、両サイトは、地理的な遠隔地に設置される。9は公衆回線であり、ATMパケットがここを通過する。サイトA81、およびサイトB82は、それぞれディスクアレイシステム1を有する。ここでは、サイトA81が通常使用される常用サイトであり、サイトB82はサイトA81が災害等でダウンしたときに使用されるリモートディザスタリカバリサイトである。

【0134】

サイトA81のディスクアレイシステム10のディスクアレイサブセット“#0”、“#1”の内容は、サイトB82のディスクアレイシステム10のリモートコピー用ディスクアレイサブセット“#0”、“#1”にコピーされる。ディスクアレイスイッチ20のI/Fノードのうち、リモートサイトに接続するものはATMを用いて公衆回線9に接続されている。このノードをATMノード205と呼ぶ。ATMノード205は、図5に示すホストI/Fノードと同様に構成され、IC2023がATM-ファイバチャネルの変換を行う。この変換は、第4実施形態におけるSCSI-ファイバチャネルの変換と同様

10

20

30

40

50

の方法により実現される。

【 0 1 3 5 】

本実施形態におけるリモートコピーの処理は、第 6 実施形態におけるミラーリングの処理と類似する。以下、第 6 実施形態におけるミラーリングの処理と異なる点について説明する。

【 0 1 3 6 】

ホスト 3 0 がライトコマンドフレームを発行すると、サイト A 8 1 のディスクアレイシステム 1 0 は、第 6 実施形態における場合と同様にフレームの二重化を実施し、その一方を自身のディスクアレイサブセット 10 に転送する。他方のフレームは、A T M ノード 2 0 5 によりファイバチャネルフレームから A T M パケットに変換され、公衆回線 9 を介して

10

【 0 1 3 7 】

サイト B 8 2 では、ディスクアレイスイッチ 2 0 の A T M ノード 2 0 5 がこのパケットを受信する。A T M ノード 2 0 5 の I C 2 0 2 3 は、A T M パケットからファイバチャネルフレームを再現し、S C 2 0 2 2 に転送する。S C 2 0 2 2 は、ホスト 3 0 からライトコマンドを受信したときと同様にフレーム変換を施し、リモートコピー用のディスクアレイサブセットに転送する。以降、データ転送準備完了フレーム、データフレーム、ステータスフレームのすべてにおいて、A T M ノード 2 0 5 においてファイバチャネル - A T M 変換を行い、同様のフレーム転送処理を実施することにより、リモートコピーが実現できる。

20

【 0 1 3 8 】

ホスト 3 0 がリードコマンドフレームを発行した際には、ディスクアレイスイッチ 2 0 は、自サイトのディスクアレイサブセット 1 0 に対してのみコマンドフレームを転送し、自サイトのディスクアレイサブセット 1 0 からのみデータをリードする。このときの動作は、第 1 実施形態と同一となる。

【 0 1 3 9 】

本実施形態によれば、ユーザデータをリアルタイムでバックアップし、天災等によるサイト障害、ディスクアレイシステム障害に対する耐性を備えることができる。

[第 9 実施形態]

次に、一台のディスクアレイサブセット 1 0 に包含される複数の L U の統合について説明する。例えば、メインフレーム用のディスク装置は、過去のシステムとの互換性を維持するために、論理ボリュームのサイズの最大値が 2 G B に設定されている。このようなディスクアレイシステムをオープンシステムでも共用する場合、L U は論理ボリュームサイズの制限をそのまま受けることになり、小サイズの L U が多数ホストから見えることになる。このような方法では、大容量化が進展した場合に運用が困難になるという問題が生じる。そこで、ディスクアレイスイッチ 2 0 の機能により、この論理ボリューム（すなわち L U ）を統合して一つの大きな統合 L U を構成することを考える。本実施形態では、統合 L U の作成をディスクアレイスイッチ 2 0 で実施する。

30

【 0 1 4 0 】

本実施形態における L U の統合は、第 1 実施形態における複数のディスクアレイサブセット 1 0 による統合 L U の作成と同一である。相違点は、同一のディスクアレイサブセット 1 0 内の複数 L U による統合であることだけである。ディスクアレイシステムとしての動作は、第 1 実施形態と全く同一となる。

40

【 0 1 4 1 】

このように、同一のディスクアレイサブセット 1 0 に包含される複数の L U を統合して一つの大きな L U を作成することで、ホストから多数の L U を管理する必要がなくなり、運用性に優れ、管理コストを低減したディスクアレイシステムを構築できる。

[第 1 0 実施形態]

次に、ディスクアレイスイッチ 1 0 による交代パスの設定方法について、図 2 6 を参照しつつ説明する。

50

【 0 1 4 2 】

図 2 6 に示された計算機システムにおける各部の構成は、第 1 の実施形態と同様である。ここでは、2 台のホスト 3 0 が、各々異なるディスクアレイ I / F 2 1 を用いてディスクアレイサブセット 1 0 をアクセスするとように構成していると仮定する。図では、ディスクアレイサブセット、ディスクアレイスイッチ 2 0 のホスト I / F ノード 2 0 3 およびディスクアレイ I / F ノード 2 0 2 は、ここでの説明に必要な数しか示されていない。

【 0 1 4 3 】

ディスクアレイサブセット 1 0 は、図 2 と同様の構成を有し、2 つのディスクアレイ I / F コントローラはそれぞれ 1 台のディスクアレイスイッチ 2 0 に接続している。ディスクアレイスイッチ 2 0 の各ノードの D C T 2 2 7 には、ディスクアレイ I / F 2 1 の交代パスが設定される。交代パスとは、ある一つのパスに障害が発生した場合にもアクセス可能になるように設けられる代替のパスのことである。ここでは、ディスクアレイ I / F “ # 0 ” の交替パスをディスクアレイ I / F “ # 1 ”、ディスクアレイ I / F “ # 1 ” の交替パスをディスクアレイ I / F “ # 0 ” として設定しておく。同様に、ディスクアレイサブセット 1 0 内の上位アダプタ間、キャッシュ・交代メモリ間、下位アダプタ間のそれぞれについても交代パスを設定しておく。

【 0 1 4 4 】

次に、図 2 6 に示すように、ディスクアレイサブセット 1 の上位アダプタ “ # 1 ” に接続するディスクアレイ I / F 2 1 が断線し、障害が発生したと仮定して、交替パスの設定動作を説明する。このとき、障害が発生したディスクアレイ I / F 2 1 を利用しているホスト “ # 1 ” は、ディスクアレイサブセット 1 0 にアクセスできなくなる。ディスクアレイスイッチ 2 0 は、ディスクアレイサブセット 1 0 との間のフレーム転送の異常を検出し、リトライ処理を実施しても回復しない場合、このパスに障害が発生したと認識する。

【 0 1 4 5 】

パスの障害が発生すると、S P 2 0 2 1 は、D C T 2 0 2 7 にディスクアレイ I / F “ # 1 ” に障害が発生したことを登録し、交代パスとしてディスクアレイ I / F “ # 0 ” を使用することを登録する。以降、ホスト I / F ノード 2 0 3 の S C 2 0 2 2 は、ホスト “ # 1 ” からのフレームをディスクアレイ I / F “ # 0 ” に接続するディスクアレイ I / F ノード 2 0 2 に転送するように動作する。

【 0 1 4 6 】

ディスクアレイサブセット 1 0 の上位アダプタ 1 0 1 は、ホスト “ # 1 ” からのコマンドを引き継いで処理する。また、ディスクアレイスイッチ 2 0 は、ディスクアレイシステム構成管理手段 7 0 に障害の発生を通知し、ディスクアレイシステム構成管理手段 7 0 により管理者に障害の発生が通報される。

【 0 1 4 7 】

本実施形態によれば、パスに障害が発生した際の交替パスへの切り替えを、ホスト側に認識させることなく行うことができ、ホスト側の交代処理設定を不要にできる。これにより、システムの可用性を向上させることができる。

【 0 1 4 8 】

以上説明した各実施形態では、記憶メディアとして、すべてディスク装置を用いたディスクアレイシステムについて説明した。しかし、本発明は、これに限定されるものではなく、記憶メディアとしてディスク装置に限らず、光ディスク装置、テープ装置、D V D 装置、半導体記憶装置等を用いた場合にも同様に適用できる。

【 図面の簡単な説明 】

【 0 1 4 9 】

【 図 1 】 第 1 実施形態のコンピュータシステムの構成図である。

【 図 2 】 第 1 実施形態のディスクアレイサブセットの構成図である。

【 図 3 】 第 1 実施形態のディスクアレイスイッチの構成図である。

【 図 4 】 第 1 実施形態におけるディスクアレイスイッチのクロスバスイッチの構成図である。

10

20

30

40

50

【図 5】第 1 実施形態におけるディスクアレイスイッチのホスト I / F ノードの構成図である。

【図 6】システム構成テーブルの構成図である。

【図 7】サブセット構成テーブルの構成図である。

【図 8】ファイバチャネルのフレームの構成図である。

【図 9】ファイバチャネルのフレームヘッダの構成図である。

【図 10】ファイバチャネルのフレームペイロードの構成図である。

【図 11】ホストからのリード動作時にファイバチャネルを通して転送されるフレームのシーケンスを示す模式図である。

【図 12】ホスト LU、各ディスクアレイサブセットの LU、及び各ディスクユニットの対応関係を示す模式図である。 10

【図 13】ライト処理時のホスト I / F ノードにおける処理のフローチャートである。

【図 14】スイッチングパケットの構成図である。

【図 15】複数のディスクアレイスイッチをクラスタ接続したディスクアレイシステムの構成図である。

【図 16】第 2 実施形態におけるコンピュータシステムの構成図である。

【図 17】第 4 実施形態におけるディスクアレイスイッチのインタフェースコントローラの構成図である。

【図 18】第 5 実施形態におけるコンピュータシステムの構成図である。

【図 19】論理接続構成画面の表示例を示す画面構成図である。 20

【図 20】第 6 実施形態におけるフレームシーケンスを示す模式図である。

【図 21】第 6 実施形態のミラーリングライト処理時のホスト I / F ノードにおける処理のフローチャートである。

【図 22】第 6 実施形態のミラーリングライト処理時のホスト I / F ノードにおける処理のフローチャートである。

【図 23】第 7 実施形態におけるホスト LU と各ディスクアレイサブセットの LU との対応関係を示す模式図である。

【図 24】第 7 実施形態におけるホスト I / F ノードの処理を示すフローチャートである。

。

【図 25】第 8 実施形態におけるディザスタリカバリシステムの構成図である。 30

【図 26】交替パスの設定についての説明図である。

【符号の説明】

【 0 1 5 0 】

1 ... ディスクアレイシステム、

5 ... 管理端末、

1 0 ... ディスクアレイサブセット、

2 0 ... ディスクアレイスイッチ、

3 0 ... ホストコンピュータ、

7 0 ... ディスクアレイシステム構成管理手段、

2 0 0 ... 管理プロセッサ、

2 0 1 ... クロスバスイッチ、

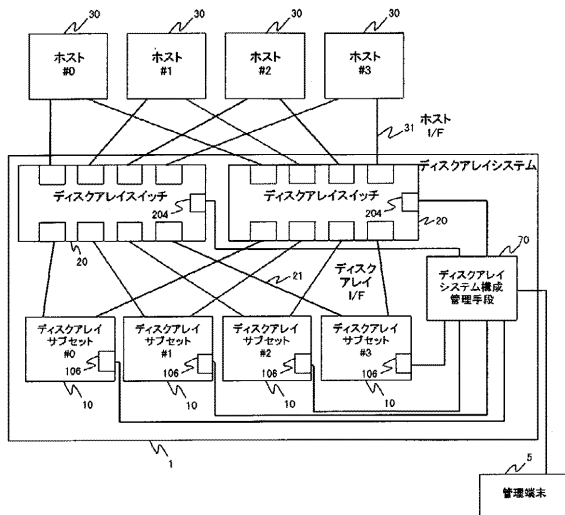
2 0 2 ... ディスクアレイ I / F ノード、

2 0 3 ... ホスト I / F ノード、

2 0 4 ... 通信コントローラ。 40

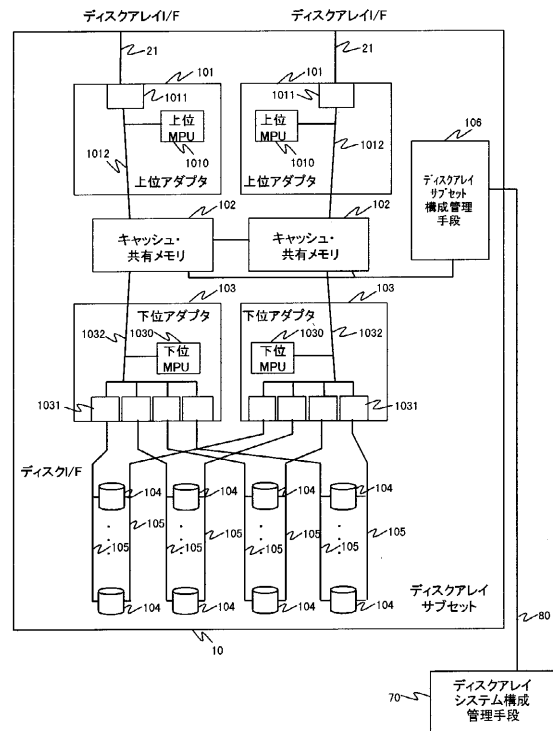
【図 1】

図1



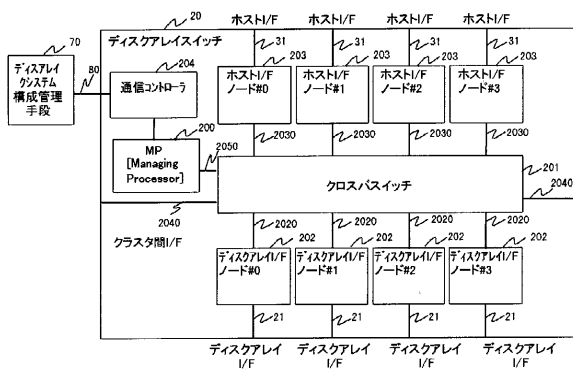
【図 2】

図2



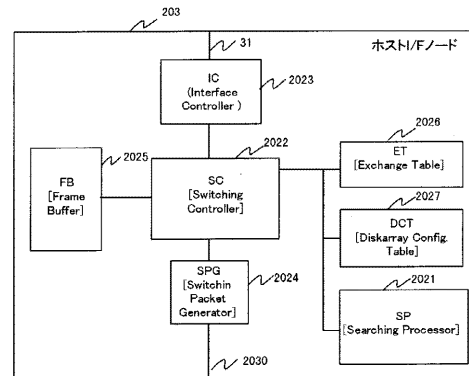
【図 3】

図3



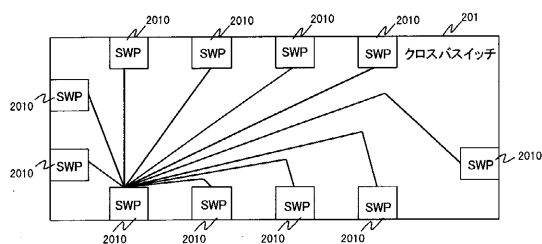
【図 5】

図5

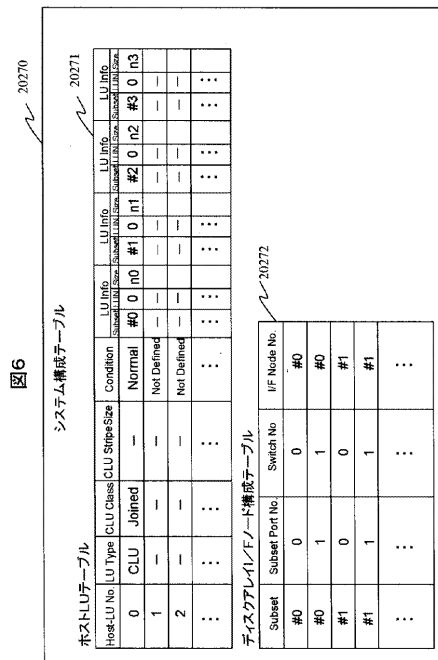


【図 4】

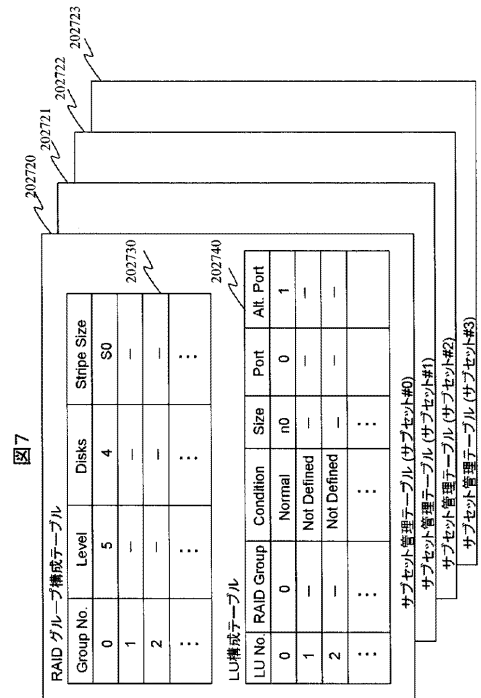
図4



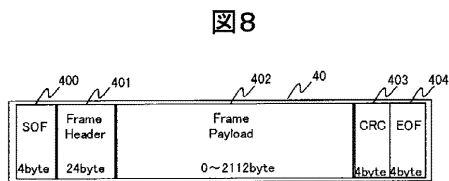
【図 6】



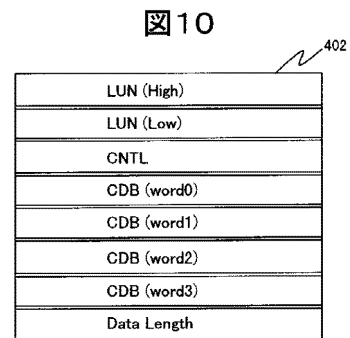
【図 7】



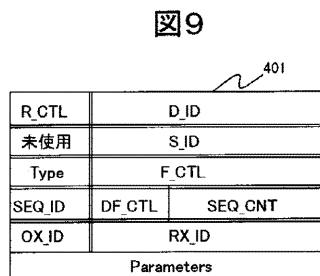
【図 8】



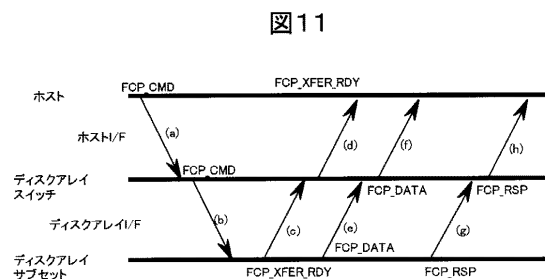
【図 10】



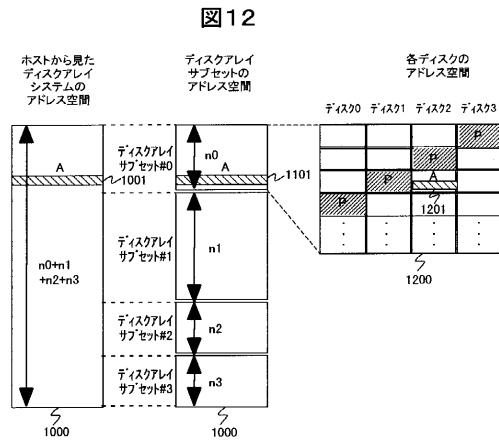
【図 9】



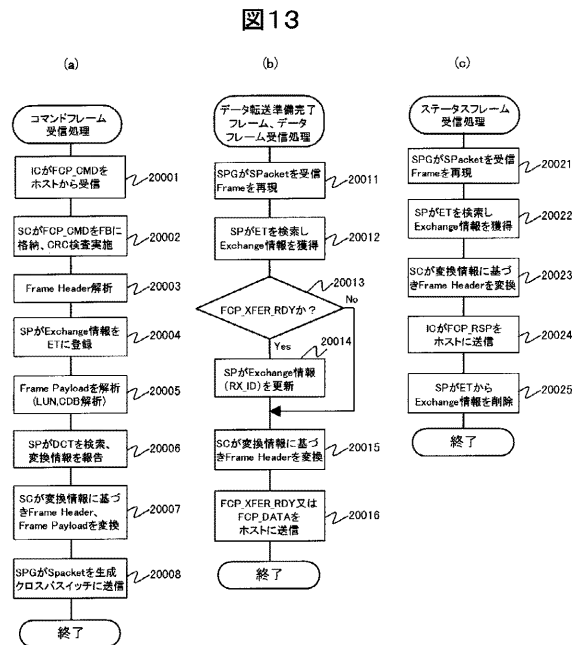
【図 11】



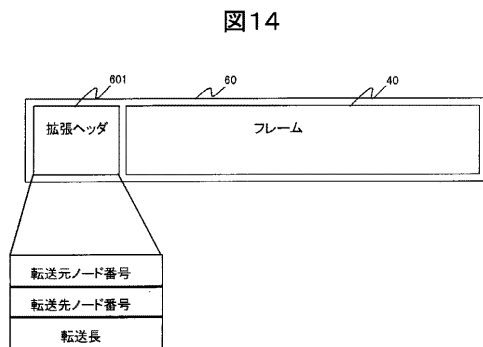
【図12】



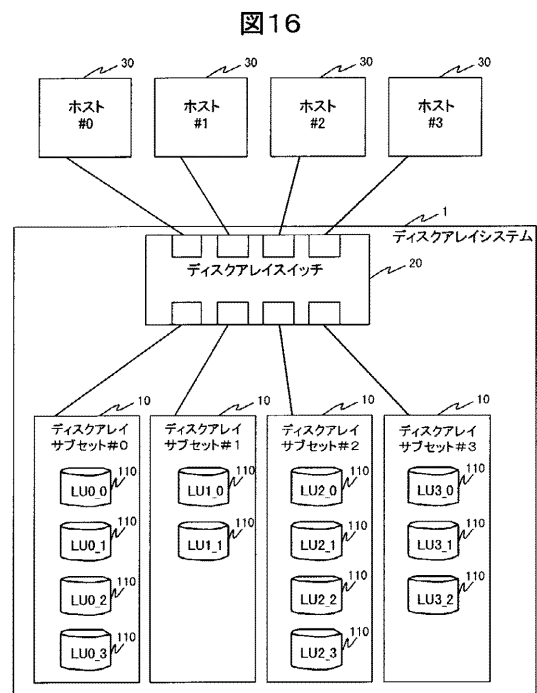
【図13】



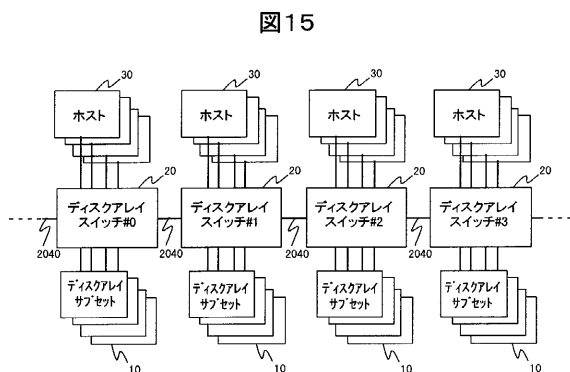
【図14】



【図16】

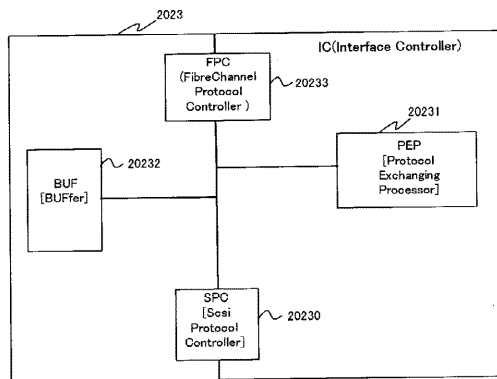


【図15】



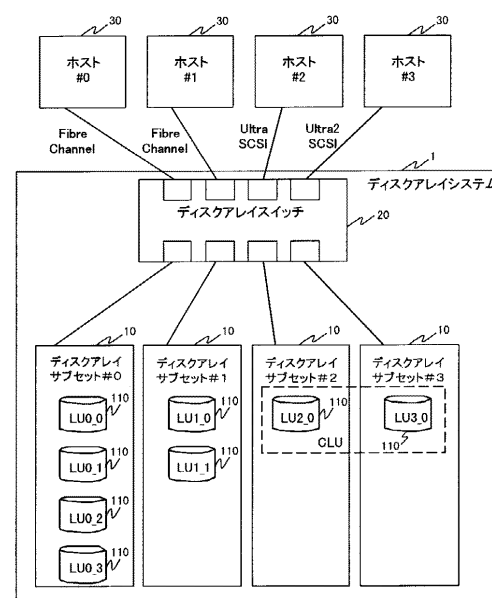
【図 17】

図17



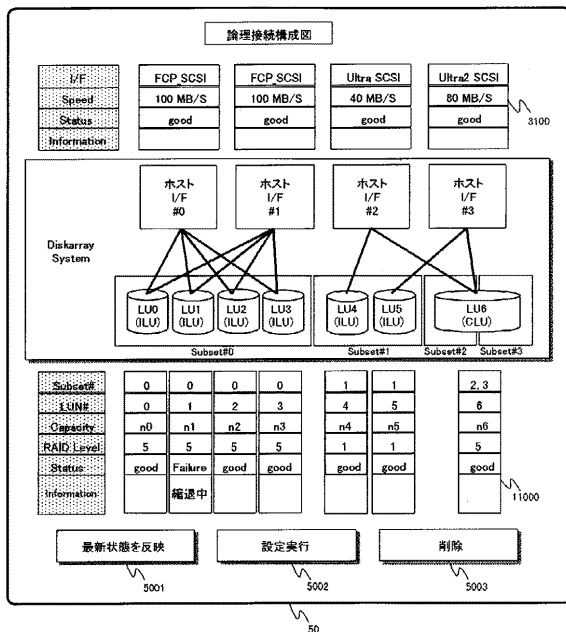
【図 18】

図18



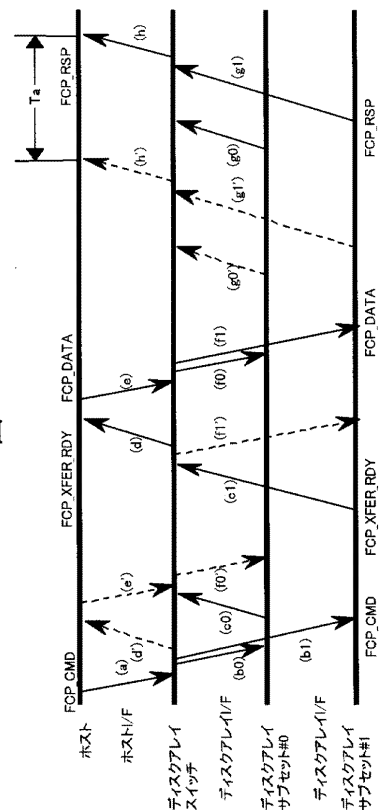
【図 19】

図19



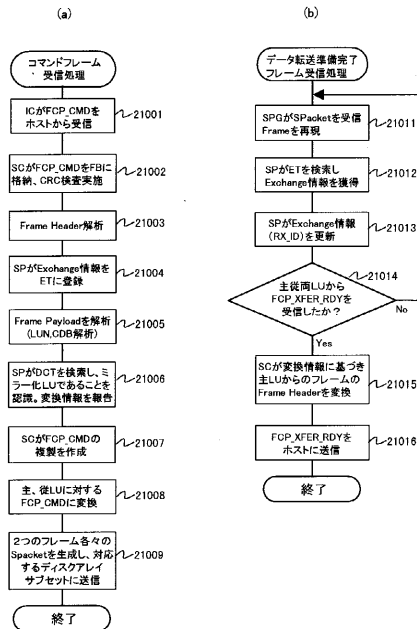
【図 20】

図20



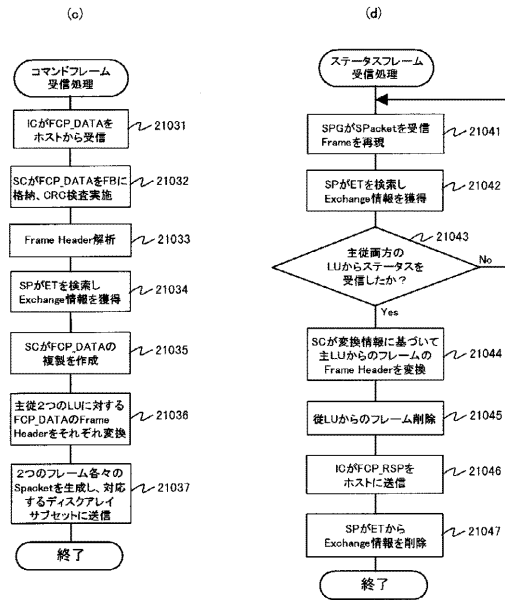
【図 2 1】

図21



【図 2 2】

図22



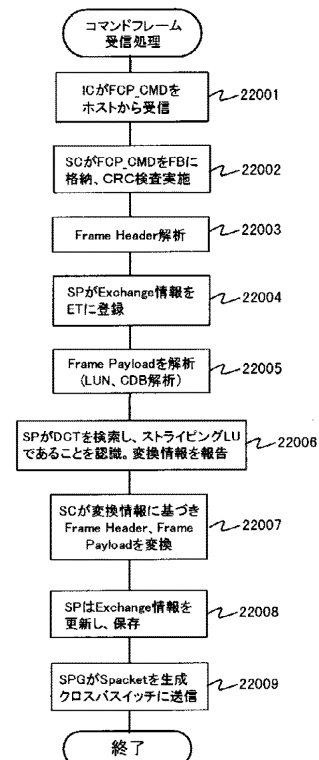
【図 2 3】

図23

ホストから見た ディスクアレイ システムの アドレス空間	ディスクアレイ サブセット#0の アドレス空間	ディスクアレイ サブセット#1の アドレス空間	ディスクアレイ サブセット#2の アドレス空間	ディスクアレイ サブセット#3の アドレス空間
0	0	1	2	3
1	4	5	6	7
2	8	9	10	11
3	12	13	14	15
4	*	*	*	*
5	*	*	*	*
6	*	*	*	*
7				
8				
9				
10				
11				
12				
13				
14				
15				
*				
*				

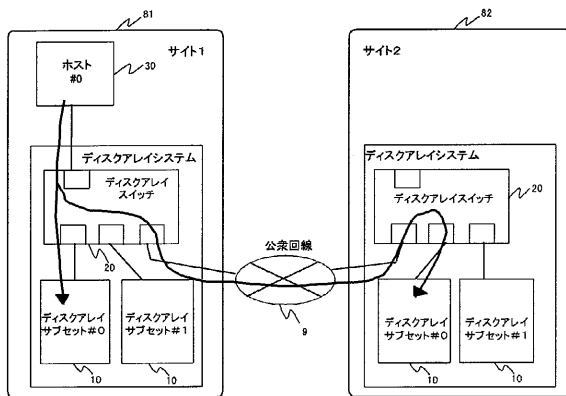
【図 2 4】

図24



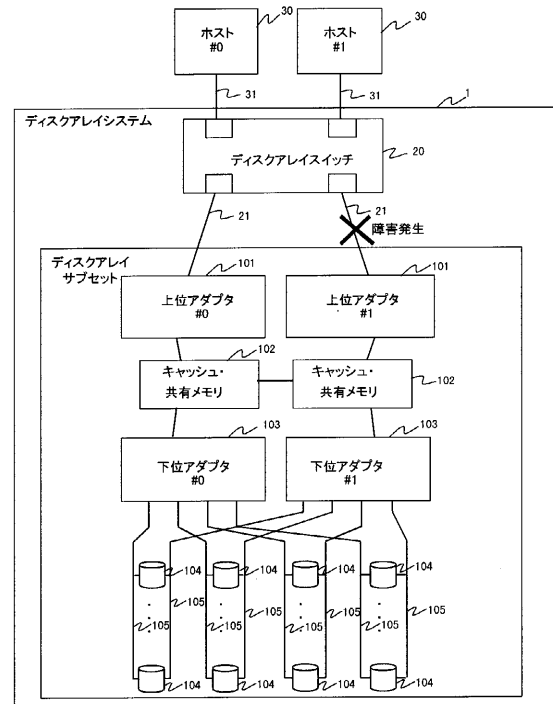
【図 25】

図25



【図 26】

図26



フロントページの続き

(72)発明者 山本 彰

神奈川県川崎市麻生区王禅寺 1 0 9 9 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 味松 康行

神奈川県川崎市麻生区王禅寺 1 0 9 9 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 佐藤 雅彦

神奈川県小田原市国府津 2 8 8 0 番地 株式会社日立製作所 ストレージシステム事業部内

合議体

審判長 清水 稔

審判官 丸山 高政

審判官 近藤 聡

(56)参考文献 特開平 8 - 3 2 8 7 6 0 (J P , A)

特開平 5 - 2 6 5 6 6 1 (J P , A)

特開平 6 - 1 8 7 2 0 1 (J P , A)

特開平 5 - 1 9 7 4 9 8 (J P , A)

特開平 9 - 2 9 2 9 5 4 (J P , A)

特開平 1 0 - 3 3 3 8 3 9 (J P , A)

特開平 1 - 2 9 2 5 2 2 (J P , A)

(58)調査した分野(Int.Cl. , D B 名)

G06F 3/06