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

特許第3964212号
(P3964212)

(45)発行日　平成19年8月22日(2007.8.22)　(24)登録日　平成19年6月1日(2007.6.1)

(51)Int.Cl.		F I		
G06F 13/00	(2006.01)	G06F	13/00	301P
G06F 3/06	(2006.01)	G06F	3/06	301M
		G06F	3/06	302A
		G06F	3/06	304B

請求項の数 3　(全 18 頁)

(21)出願番号　特願2002-6873(P2002-6873)
(22)出願日　平成14年1月16日(2002.1.16)
(65)公開番号　特開2003-208362(P2003-208362A)
(43)公開日　平成15年7月25日(2003.7.25)
審査請求日　平成16年9月24日(2004.9.24)

(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　100079108
弁理士　稲葉　良幸
(74)代理人　100093861
弁理士　大賀　眞司
(74)代理人　100100310
弁理士　井上　学
(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

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】

ブロックＩ／Ｏインタフェースを制御する第１のインタフェース制御装置を有する第１のアダプタボードと、

ファイルＩ／Ｏインタフェースを制御する第２のインタフェース制御装置を有する第２のアダプタボードと、

複数の前記第１のインタフェース制御装置及び複数の前記第２のインタフェース制御装置に接続される共有メモリを有する第１のコントローラボードと、

前記第１のインタフェース制御装置、前記第２のインタフェース制御装置、前記共有メモリ及びディスクアダプタに接続されるキャッシュメモリを有する第２のコントローラボードと、

前記第１のアダプタボード、前記第２のアダプタボード、前記第１のコントローラボード、及び前記第２のコントローラボードが格納され、前記第１のアダプタボードが格納されるスロット及び前記第２のアダプタボードが格納されるスロットが同一形状である複数のスロット、及び前記共有メモリ及び前記キャッシュメモリと接続される前記ディスクアダプタを有するディスクコントローラと、

前記ディスクコントローラと接続される複数のディスク装置を有するディスクユニットと

を備え、

前記複数の第１のインタフェース制御装置のうちの一部は、同一のドメインで管理され

るネットワークに接続され、他の一部は、前記ドメインとは異なるドメインで管理されるネットワークに接続され、

前記ディスクコントローラは、前記同一のドメインで管理されるネットワークに接続される前記複数の第１のインタフェース制御装置の一部に含まれるインタフェース制御装置に障害が発生した場合、複数の前記第１のインタフェース制御装置の一部に含まれる他のインタフェース制御装置に前記障害が発生したインタフェース制御装置の行っていた処理を移す第１のフェイルオーバ手段と、前記他のインタフェース制御装置に障害が発生した場合、前記複数の第１のインタフェース制御装置の一部に含まれ、かつ障害が発生していないインタフェース制御装置に前記他のインタフェース制御装置が行っていた処理を移す第２のフェイルオーバ手段とを有し、

前記共有メモリは、障害が発生したインタフェース制御装置が行っていた処理を他のインタフェース制御装置へ移す手順を格納し、

前記第１及び第２のフェイルオーバ手段は、前記手順に従って実行される

ことを特徴とする記憶装置システム。

【請求項２】

前記複数の第１及び第２のインタフェース制御装置は、

ハートビートマークを一定時間間隔で前記共有メモリが有するハートビート格納領域の定められた領域に保存する手段と、前記ハートビート格納領域に格納された前記ハートビートマークを用いてお互いに他の制御装置の状態を監視する手段とを有する

ことを特徴とする請求項１に記載の記憶装置システム。

【請求項３】

前記第１のフェイルオーバ手段は、

前記複数の第１のインタフェース制御装置の一部に含まれる他のインタフェース制御装置のうち、稼働率の最も低いインタフェース制御装置を前記障害が発生したインタフェース制御装置の行っていた処理を移す対象として選択する

ことを特徴とする請求項１に記載の記憶装置システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、計算機システムで用いられる記憶装置システムに係り、特に、複数の入出力インタフェースを有する記憶装置システムに関する。

【０００２】

【従来の技術】

記憶装置システムと計算機との間のインタフェース（以下、「Ｉ／Ｆ」と記す）には、大きく分けて二つのタイプが存在する。

【０００３】

第一は、記憶装置におけるデータの管理単位であるブロックを単位としてデータの入出力（以下、「Ｉ／Ｏ」と記す）が行われる「ブロックＩ／Ｏインタフェース」である。ファイバチャネル、ＳＣＳＩ（Small Computer Systems Interface）、ＥＳＣＯＮ（ESCONは、米国IBM社の登録商標である）等は、ブロックＩ／Ｏインタフェースに属する。複数の計算機と複数の記憶装置システムとがブロックＩ／Ｏインタフェースを用いて相互に接続されたシステムは、ストレージエリアネットワーク（ＳＡＮ）と呼ばれる。ＳＡＮの構築にはファイバチャネルが使用されることが多い。

【０００４】

第二は、ファイル単位にデータのＩ／Ｏが行われる「ファイルＩ／Ｏインタフェース」である。ファイルサーバとクライアントサーバとの間のファイル転送プロトコルであるNetwork File System（ＮＦＳ）やCommon Internet File System（ＣＩＦＳ）に基づいてＩ／Ｏが行われるインタフェースは、ファイルＩ／Ｏインタフェースである。ファイルＩ／Ｏインタフェースを備え、ローカルエリアネットワーク（ＬＡＮ）などのネットワークに接続できる記憶装置システムは、ネットワークアタッチドストレージ（ＮＡＳ）と呼

ばれる。
【０００５】
又、ファイルサーバの耐障害性を保証するフェイルオーバと呼ばれる技術がある。本技術はアメリカ特許第５６９６８９５号に詳しい。具体的には、第一の記憶装置を利用する第一のサーバと、第二の記憶装置を利用する第二のサーバとの間でハートビート信号が交換される。第一のサーバに障害が発生するとハートビート信号がとぎれる。ハートビート信号がとぎれたことを検出した第二のサーバは、第一のサーバが使用していた第一の記憶装置へアクセスし、第一のサーバの処理を引き継ぐ（フェイルオーバする）。
【０００６】
【発明が解決しようとする課題】
従来技術を用いてＳＡＮとＮＡＳの両方が混在する計算機システムを構成する場合、ＳＡＮの機能を備える記憶装置システム及びＮＡＳの機能を備える記憶装置システム個々を有するシステムしか構成できない。その結果、複数の記憶装置システムを別個に管理する必要があり、システムを管理するためのコストが増大してしまう。
【０００７】
また、記憶装置システムを複数有するシステムでは、一つの記憶装置システムに障害が発生した場合、従来技術を用いて他の記憶装置システムがフェイルオーバを行い、システム全体の耐障害性を確保できる。しかし、従来技術では、フェイルオーバを行った記憶装置システムに障害が発生した際の耐多重障害について考慮されていない。
【０００８】
さらに、従来技術では、複数のネットワークドメインに接続できる記憶装置システムについて何ら考慮されていない。また、このような環境におけるフェイルオーバについても考慮されていない。
【０００９】
本発明の目的は、システムが有する複数のインターフェースを一括して管理することで管理コストを低減できる記憶装置システムを提供することである。
【００１０】
また、本発明の他の目的は、耐多重障害性を有し、複数のネットワークドメインに対応できる記憶装置システムを提供することである。
【００１１】
【課題を解決するための手段】
上記目的を達成するために、本発明の記憶装置システムは、ブロックＩ／Ｏインタフェース制御装置、ファイルＩ／Ｏインタフェース制御装置等の複数のインタフェース制御装置の各々に対応する複数のスロット、及び、任意のインタフェース制御装置からアクセス可能な複数のディスク装置を制御する複数のディスク制御部とを有するディスクコントローラを有する。
【００１２】
また、好ましい実施の形態としては、各インタフェース制御装置は、同一の形状のボードとして実装され、スロットの場所に依存せず装填可能とすることが考えられる。
【００１３】
さらに、本発明の好ましい実施形態としては、本発明に係る記憶装置システムは、上記の構成に加え、フェイルオーバ可能なインターフェース制御装置群を一括して管理するための管理テーブル、フェイルオーバの手順を規定する情報テーブル、及び規定されたフェイルオーバ手順に従い、同一のフェイルオーバ可能なグループに属するインタフェース制御装置間の処理の引継ぎを制御するフェイルオーバ制御手段とを有する構成が考えられる。
【００１４】
【発明の実施の形態】
図１は、本発明を適用した記憶装置システムの実施形態を示した図である。以下、xは任意の整数を表す。
【００１５】

記憶装置システム１は、ディスクコントローラ（以下、「DKC」と称する）11及び記憶装置1700を有する。DKC11において、NASチャネルアダプタ（以下CHN）110xは、ファイルＩ／Ｏインタフェースで NASクライアント400と接続されるインターフェース制御装置である。ファイバチャネルアダプタ（以下CHF）111xは、ブロックＩ／ＯインタフェースでSANクライアント500と接続されるインターフェース制御装置である。以下、CHN及びCHFをまとめてチャネルアダプタ（以下CH）と称する。記憶装置1700は、ディスクアダプタ120に接続されている。各ディスクアダプタ120は、接続されている記憶装置1700を制御する（以下、「DKA」と称する）。13は共有メモリ（以下SM）である。14はキャッシュメモリ（以下CM）である。共有メモリコントローラ（以下SMC）15は、CHN110x、CHF111x、DKA120及びSM13と接続される。SMC15は、CHN110x、CHF111x、及びDKA120とSM13との間のデータ転送を制御する。キャッシュメモリコントローラ（以下CMC）16は、CHN110x、CHF111x、DKA120及びCM14と接続される。CMC16は、CHN110x、CHF111x、及びDKA120とCM14との間のデータ転送を制御する。

【００１６】

LAN20及び21は、CHN110xとNASクライアント400とを接続する。一般的に、LANにはEthernet（Ethernetは、富士ゼロックス社の登録商標である）が用いられる。LAN20及びLAN21には、各々異なるドメインが割り当てられる。ドメインとは、ネットワークにおける管理範囲を指す。本実施形態においては、各LANにDOM-LAN0、DOM-LAN1のドメイン名を付与する。SAN30は、CHF111x及びSANクライアント500とを接続する。本実施形態においては、SAN30にDOM-FC0のドメイン名を付与する。

【００１７】

記憶装置システム１では、すべてのCHは、CMC16、SMC15を介して、CM14及びすべての記憶装置1700へアクセスすることが出来る。

また、記憶装置システム1は、SANとNASの双方のインターフェースを有するシステムである。また、本実施形態においては、複数のCHNが幾つかのグループに分割され、その各々のグループが異なるドメインに管理されるLANと接続されることができる。

尚、記憶装置システム１が有するインターフェースは、SANのみ又はNASのみでも構わない。

【００１８】

図２は、記憶装置システム１の外観図である。DKC11は、CHN110x及びCHF111x、DKA120、SM13、及びCM14を格納する。ディスクユニット(以下、「DKU」と称する)180及び181は、記憶装置1700を格納する。SM13は、実際には複数のコントローラボード130で構成される。また、CM14も、複数のキャッシュボード140で構成される。ボード130及び140はスロット190に格納される。記憶装置システム１の使用者は、これらのボードの枚数を増減して、所望の記憶容量を構成する。尚、図２は、装填の具体例として、ボード130及び140が各々４枚ずつスロット１９０に装填されている状態を示している。

【００１９】

スロット190には、CHN110x等が作りこまれたアダプタボードが格納される。本実施形態においては、スロット190の形状、アダプタボードのサイズ及びコネクタの形状を、インターフェースの種類を問わず一定にし、互換性を保つようにする。したがって、DKC11には、インターフェースの種類を問わず、任意のスロット190に任意のアダプタボードを装填することができる。また、記憶装置システム１の使用者は、CHN110x及びCHF111xのアダプタボードの数を自由に組み合わせて記憶装置システム１のスロット190に装填することが出来る。

【００２０】

図３は、CHN110xが作りこまれたアダプタボードの構成を示す図である。コネクタ11007は、DKC19が有するコネクタと接続される。本実施形態においては、上述したように、CHN110x及びCHF111xは同一形状のコネクタを有する。インタフェースコネクタ2001は、イーサネット（イーサネットは、富士ゼロックスの登録商標である）に対応している。尚、アダプタボードがCHF111xの場合、インタフェースコネクタ2001は、ファイバチャネルに対応

する。
【００２１】
図４は、CHN110xの内部構成を示す図である。11001は中央制御部である。LANコントローラ11002は、インターフェースコネクタ2001を介して、LANと接続される。メモリ1004は、中央制御部11001と接続される。メモリ11004には、中央制御部11001が実行するプログラムや、制御データが格納される。SM I/F制御部11005は、CHN110xからSM13へのアクセスを制御する。CM I/F制御手段11006は、CHN110xからCM14へのアクセスを制御する。
【００２２】
なお、中央制御部11001は単一のプロセッサでも、複数のプロセッサの集合体でも良い。例えば、制御に伴う処理を水平負荷分散させる対称型マルチプロセッサの構成でもよい。また、複数のプロセッサのうち、一方がＩ／Ｏインタフェースのプロトコル処理を行い、他方のプロセッサがディスクボリュームの制御を行うという非対称型マルチプロセッサの構成でもよい。
尚、CHF111xの構成は、LANコントローラ11002がファイバチャネルコントローラに置き換えられる以外は本図と同一である。
【００２３】
図５は、CHN110xが有するメモリ11004の内容を示す図である。オペレーティングシステムプログラム110040は、プログラム全体の管理や入出力制御に用いられる。LANコントローラドライバプログラム110041は、LANコントローラ11002の制御に用いられる。TCP/IPプログラム110042は、LAN上の通信プロトコルであるTCP/IPの制御に用いられる。ファイルシステムプログラム110043は、記憶装置に格納されるファイルの管理に用いられる。ネットワークファイルシステムプログラム110044は、記憶装置に格納されるファイルをNASクライアント400に提供するためのプロトコルであるNFSやCIFS等の制御に用いられる。ディスクボリューム制御プログラム110045は、記憶装置1700に設定されたディスクボリュームへのアクセス制御に用いられる。キャッシュ制御プログラム110046は、CM14のデータ管理やヒット／ミス判定等の制御に用いられる。フェイルオーバプログラム110047は、障害が発生したCHNから他の正常なCHNへ処理を移す等の処理を制御するために用いられる。フェイルオーバプログラム110047の詳細は後述する。
【００２４】
以下、記憶装置システム１における処理について説明する。本実施形態においては、記憶装置システム１の管理を容易にするために、記憶装置システム１が有するCHを階層的に管理する。具体的には、物理I/F、論理I/F、ドメイン、フェイルオーバグループという４つの指標に基づいて、Ch群を階層化する。尚、指標は上記４つの指標に限られない。
【００２５】
図６は、階層化されたCH群の具体例を示す図である。記憶装置システム１に該当する部分は、図中、網掛けで示してある。
図６の各々において、最も外側の円は、物理I/Fの階層を示す。本階層では、CHは、CH自身のホスト接続インタフェースの物理媒体に基づいてグループ分けされる。具体的には、ファイバチャネル、UltraSCSI、ESCON、及びイーサネットの４種類の物理媒体でグループ分けが行われる。
【００２６】
外側から２番目の円は、論理I/Fに基づく階層を示す。本階層では、CHは、CH自身のホスト接続インタフェースの論理プロトコルに基づいてグループ分けされる。具体的には、ファイバチャネルプロトコル（FCP）、SCSI、ESCON、NAS（すなわちファイルＩ／Ｏインタフェース）及びiSCSIの論理プロトコルでグループ分けが行われる。
【００２７】
外側から３番目の円は、ドメインに基づく階層を示す。本階層では、CHに割り当てられたドメイン（イーサネットの場合にはIPネットワークのネットワークドメイン（サブネット）、SCSIの場合には１本のSCSIバス、ファイバチャネルの場合には、一つのループ又は単一アドレス空間で構成されたSAN全体）に基づいてグループ分けされる。尚、いずれドメ

インの場合でも、１つのドメイン内では、アドレス空間が同一であることが必要となる。
【００２８】
外側から４番目、すなわち一番内側の円は、フェイルオーバグループによる階層を示している。本階層では、相互にフェイルオーバ可能なCHを１つのグループとする。
【００２９】
尚、CH間でフェイルオーバを行うためには、CH間でアドレス情報を送受信する必要がある。したがって、１つのフェイルオーバグループを構成するCHの各々には、同一のドメインが付与されなければならない。同一のドメイン内であれば、フェイルオーバグループは、DOM-LAN0ドメインのように同一ドメインの中に１つであってもよいし、DOM-LAN2ドメインのように２つ以上であっても構わない。また、フェイルオーバグループに含まれるCHの数も、FOG-LN1のように２つであっても、FOG-LN0のように３つ以上であっても良い。
【００３０】
最も内側に位置する四角は、CHを示す。同図では、合計27台のCHが存在している。
【００３１】
図７は、SM13の内容を示す図である。CHを管理するための管理情報は、SM13に格納される。構成管理情報格納エリア131には、記憶装置システム１が有するインターフェース等の構成を示す管理情報が格納される。構成情報格納エリア131には、チャネルアダプタ管理テーブル1310、フェイルオーバ管理情報1311、ハートビートマーク格納エリア1312、及び引継情報格納エリア1313の各エリアが用意される。
【００３２】
図８は、チャネルアダプタ管理テーブル1310を示す図である。チャネルアダプタ管理テーブル1310は、CHのグループを管理するためのテーブルである。本図では、記憶装置システム1の構成に対応するテーブルを示している。
【００３３】
チャネルアダプタエントリ13101には、CHに付与された識別子が登録される。物理I/Fグループエントリ13102には、登録されたCHが属する物理I/Fグループを示す情報が登録される。論理I/Fグループエントリ13103には、登録されたCHが属する論理I/Fグループを示す情報が登録される。ドメインエントリ13104には、登録されたCHが属するドメインを示す情報が登録される。フェイルオーバグループエントリ13105には、登録されたチャネルアダプタが属するフェイルオーバグループを示す情報が登録される。ステータスエントリ13106には、登録されたCHの状態を示す情報（正常、異常、どのCHの処理を引き継いでいるか等）が登録される。稼働率エントリ13107には、登録されたCHの稼働状況を示す情報、具体的にはCHの稼働率が登録される。
【００３４】
尚、図８は、CHN1及びCHN2に障害が発生（Failed）し、同一ドメイン（DOM-LAN0）の同一フェイルオーバグループ（FOG-LN0）に属する正常なチャネルアダプタCHN3に処理が引き継がれている（TakingOver）状態を示している。この場合、CHN3は、自らの処理だけでなく、他の２台のCHNの処理も実行しているため、CHN3の稼働率は86%と高い値となる。
【００３５】
図７の説明に戻る。ハートビートマーク格納エリア1312には、CHの状態に関する情報が格納される。以下、CHの状態に関する情報をハートビートマーク（以下、「HBM」と称する）という。HBMには、CHN識別子、正常を示す符号、更新時刻等の情報が含まれる。
【００３６】
引継情報格納エリア1313には、各CHに関する引継情報が格納される。障害が発生したCHが行っている処理を他のCHが代替できるように、各CHは、引継情報格納エリア1313に自らの引継情報を格納する。引継情報には、LANコントローラ11002のMACアドレス及びIPアドレス、ファイルシステム110043を構成するデバイス情報又はマウントポイント情報、及びネットワークファイルシステム110044のエクスポート情報等が含まれる。
【００３７】
フェイルオーバ管理情報1311には、各CH間の引継関係や、監視関係に関する情報、具体的

には、後に説明される引継対象チャネルアダプタ、監視対象チャネルアダプタ等の情報が格納される。各情報の具体的な内容については、図１２から図１４を用いて後述する。

【００３８】

以下、本実施形態における記憶装置システム１の動作を説明する。はじめに、CHに障害が発生した際の動作の概要を説明する。尚、ここでいう「障害」とは、CHに再起不能な障害が発生し、他の正常なCHへ処理を引継がなければならない故障であるとする。

【００３９】

障害が発生したCHをCH－A、CH－Aの処理を引継ぐアダプタをCH－Bとする。

CH－Aに障害が発生したことをCH－A自身が検出した場合には、以下の手順でフェイルオーバ処理、復旧処理及びテイクバック処理を行う。

（１）CH－Aが自己の障害を発見し、CH－Aが自分自身の閉塞処理を実施する。

これに伴い、CH－AのHBMの更新が停止される。尚、閉塞処理とは、CHの動作を停止することを言う。

（２）CH－BがCH－AのHBM更新が停止したことを確認する。

（３）CH－BがCH－Aの処理を引き継ぐ（フェイルオーバ）。

（４）CH－Aの復旧作業が実施される。復旧処理とは、具体的には、保守員によるCH－Aのボード交換等が挙げられる。復旧処理は、記憶装置システム１が行う、障害の通知に基づいて行われる。通知の具体例としては、管理端末画面表示、SNMP、Email、Syslog、ポケベル、Assist通報（センタへのホットライン）、等がある。

【００４０】

（５）CH－Aが復旧し、CH－AのHBMの更新が再開される。

（６）CH－Bが、CH－AのHBMの更新を確認する。

（７）CH－Aが、CH－Bにフェイルオーバされた処理をCH－Bから引き戻す（テイクバック）。

【００４１】

また、CH－Aに障害が発生したが、CH－A自身で閉塞処理を実施できない場合、以下の手順で処理を行う。

（１’）CH－Aに障害が発生する（中央制御部が機能しないので、HBM更新も同時に停止する）

（２）CH－BがCH－AのHBM更新が停止したことを確認する。

（３’）CH－BがCH－Aを強制的に閉塞する。

（３’）以降の手順は、上記（３）以降の手順と同一であるので省略する。

【００４２】

（１）、及び（１’）の処理の詳細を説明する。以下、CHNについてのみ説明するが、CHFでも同様である。

図９は、CH、ここではCHN1101の中央制御部11001の（１）の動作を説明するフローチャートである。ここでは、CHN1101がCH-Aに該当する。

【００４３】

中央制御部11001は、フェイルオーバ制御プログラム110047を用いて、中央制御部11001が属するCHN1101に障害が発生していないか監視する。フェイルオーバ制御プログラム110047は、CHN1101の電源がＯＮされるとともに中央制御部11001で実行が開始される（ステップ4700）。その後、フェイルオーバ制御プログラム110047にしたがって、中央制御部11001は、障害の有無の判定を開始する（ステップ4701）。

【００４４】

自身の障害が検出されない場合、中央制御部11001は、SM13のハートビートマーク格納エリア1312にHBMを保存するよう制御を行う（ステップ4702）。HBMの保存（又は更新）が終了すると、フェイルオーバ制御プログラム110047の実行が一定時間停止される（ステップ4703）。その後、中央制御部11001は、ステップ4701～4703の処理を繰り返す。

【００４５】

ステップ4701で障害の発生が検出された場合、中央制御部11001は以下の処理を実行する

。なお、ステップ4701以外でも、任意の中央制御部11001へのハードウェア割り込みでハードウエアの障害が検出される場合がある。この場合も、中央制御部11001は、以下の処理を実行する。

【００４６】

中央制御部11001が動作可能である場合、中央制御部11001は、HBMの更新を停止する。このとき、CHN1101が障害発生により停止したことを示す情報をHBMに含めるようにHBMの更新を制御することも出来る（ステップ4704）。

【００４７】

中央制御部11001は、チャネルアダプタ管理テーブル1310のCHN1101に該当するステータスエントリ13106に、障害発生(Failed)を設定する（ステップ4705）。その後、中央制御部11001は閉塞処理を実行する（ステップ4706）。

【００４８】

尚、中央制御部11001が動作不能な場合、ステップ4704～4706の処理は実行できない。しかし、中央制御部11001が動作不能に陥ると、HBM更新時刻を越えてもHBMが更新されない（（１’）に該当する）。したがって、HBMの交信状況を他のCHが監視することで、CHに障害が発生していることを検出することができる（（２）に該当する）。さらに、他のCHが、障害が発生したCHに替わってステップ4705及び4706の処理、すなわち（３’）の処理を行うことで、フェイルオーバの処理を続行する。

【００４９】

図１０は、CH－Aの処理を引き継ぐCH-B、具体的にはCHN1102の（２）及び（３）の動作を示すフローチャートである。

CHN1102の中央制御部11001は、電源がＯＮされると、フェイルオーバ制御プログラム110047の実行を開始する（ステップ4800）。

【００５０】

中央制御部11001は、同一フェイルオーバグループ内の監視対象チャネルアダプタ（監視対象CH）における障害の発生の有無を、監視対象CH（ここではCHN1101）のHBMを確認することで監視している。監視対象CHとは、あるCHに割り当てられた、そのCHが監視すべき他のCHである。あるCHの監視対象CHは、SM13に格納されているフェイルオーバ管理情報1311に登録されている。尚、監視対象CHの設定は、製品が工場から出荷されれる際にあらかじめ行われる場合や、製品にインストールされるソフトウエアによってユーザが自由に行う場合がある。

【００５１】

監視対象CHのHBMが一定時刻をすぎても更新されない、又はHBMに障害発生を示す符号が記載されていることを確認すると、中央制御部11001は、監視対象CHに障害が発生していると判断する（ステップ4801）。

【００５２】

障害を検出しなかった場合、中央制御部11001は一定時間スリープし（ステップ4802、4803）、その後、ステップ4801～4803の処理を繰り返す。

【００５３】

障害を検出した中央制御部11001は、障害が発生した監視対象CH、ここではCHN1101の状態を確認する（ステップ4804）。CHN1101で閉塞処理が行われていない場合、すなわち、CHN1101が（１’）の状態であった場合、CHN1102はCHN1101の障害後処理を実行する。障害後処理とは、障害が発生したCHの中央制御部11001に代わって、障害を発見したCHが、チャネルアダプタ管理テーブル1310の障害が発生したCHに対応するステータスエントリに障害発生(Failed)を設定したり、強制的に障害が発生したCHを閉塞させる処理である。これは、（３’）の処理に該当する（ステップ4810）。

【００５４】

その後、中央制御部11001は、引継対象チャネルアダプタ（引継対象CH）を特定する。引継対象CHの情報は、フェイルオーバ管理情報1131に格納されている。

【００５５】

引継対象CHとは、あるCHに割り当てられる、あるCHが処理を引継ぐべきCHのことである。例えば、CHN1101がCHN1102の引継対象CHに割り当てられる場合、CHN1101に障害が発生したら、CHN1102がCHN1101の処理を引継ぐ。尚、引継対象CHには、障害が検出されたCHだけでなく、障害が検出されたCHが処理を引き継いでいた他のCHも含まれる。このような場合、処理を引き継ぐCHは、全てのCHの処理を引き継ぐ必要がある。したがって、中央制御部11001は、このようなCHの有無を、フェイルオーバ管理情報1311で確認する。

【００５６】

本実施形態では、CHN1102には、CHN1101が引継対象CHとして割り当てられると仮定する。したがって、本ステップにおいては、中央制御部11001は、引継対象CHとしてCHN1101を特定する。引継対象CHの確認の仕方については後述する（ステップ4805）。

【００５７】

中央制御部11001は、フェイルオーバ管理情報1311に含まれる情報を更新する。更新の仕方については後述する（ステップ4806）。

【００５８】

中央制御部11001は、監視対象CHを更新する。これはフェイルオーバ管理情報1311内の情報が更新されることにより、他にも監視すべきCHNが割り当てられるかもしれないからである。更新の仕方については後述する（ステップ4807）。

【００５９】

監視対象CHでありかつ引継対象CHとなるCHN1101の障害を発見したCHN1102の中央制御部11001は、引継処理を実行する。引継処理は以下の手順に従う。

【００６０】

中央制御部11001は、障害が発生したCHN1101に関する引継情報をSM13の引継情報格納エリア1313から獲得する。中央制御部11001は、獲得した引継情報に基づいて、障害が発生したCHN1101が有するLANコントローラ11002のMACアドレス及びIPアドレスを、CHN1102のLANコントローラ11002に設定する。この設定により、CHN1102は、障害が発生したCHN1101へのLANアクセス及びCHN1102へのLANアクセスの双方に応答できる。また、中央制御部11001は、CHN1101のファイルシステム110043を構成するデバイス情報やマウントポイント情報に基づいて、CHN1101にマウントされていたファイルシステムをCHN1102にマウントする。中央制御部11001は、ファイルシステムの回復処理としてジャーナルのリプレイを行う。その後、中央制御部11001は、ネットワークファイルシステム110044のエクスポート情報に基づいて、回復したファイルシステムを定められたエクスポートポイントで公開する。又、中央制御部11001は、必要に応じNASクライアントからCHN1101に要求された実行途中の処理を継続する（ステップ4808）。

【００６１】

以上で引継処理が終了する（ステップ4809）。この後、中央制御部11001は、改めてステップ4800から監視処理を実行する。

【００６２】

図１１は、障害が発生したCHの処理を引き継いでいるCH、ここではCHN1102における復旧処理、すなわち（６）及び（７）の動作を説明するフローチャートである。

【００６３】

中央制御部11001は、復旧処理を開始する（ステップ4900）。

中央制御部11001は、全ての監視対象CHのHBMを確認する。尚、この処理は、ステップ4801の処理と同一である（ステップ4901）。

【００６４】

障害が発生していたCH、ここではCNH1101の復旧を検出した中央制御部11001は、ステップ4904以降の処理を行う（ステップ4902）。復旧が検出されない場合、中央制御部11001は一定時間スリープし（ステップ4903）、4901～4903の処理を繰り返す。

【００６５】

中央制御部11001は、CNH1101をフェイルオーバの対象から外すため、フェイルオーバ管理情報1311を更新する。更新の仕方については後述する（ステップ4904）。

【００６６】
中央制御部11001は、引継対象CHを更新する。具体的には、復旧されたCHを引継対象CHから外すような更新を行う。CNH1102が、CNH1101だけでなく、CNH1101が引継いでいた他のCHNの処理も引継いでいる場合、このようなCHNも引継対象CHから外すことができる。具体例を用いて説明する。まずCHN1101に障害が発生し、CHN1102が処理を引継ぐ。その後、CHN1102に障害が発生し、CHN1103がCHN1102とCHN1101の両方の処理を引継いだとする。その後、CHN1102が復旧すると、CHN1103は、CHN1102だけでなく、CHN1101も引継対象CHから外すことができる。尚、具体的な更新の仕方については、図１２から図１４で説明する（ステップ4905）。

【００６７】
中央制御部11001は、監視対象CHを更新する。フェイルオーバ管理情報等の更新により、監視対象CHも変更される可能性があるからである（ステップ4906）。

【００６８】
中央制御部11001は、差し戻し処理を実行する。差し戻し処理とは、引継処理で引き継いだ処理を元のCHNに返却する処理のことである。具体的には、引継処理が行われた際に引継がれた引継情報を復旧が行われたCHに送り返す処理を行う（ステップ4907）。

【００６９】
以上で、復旧処理が完了する（ステップ4908）。CHN1102が引継いでいるCHNが他に存在しているならば、再び以上の処理を繰り返す。

【００７０】
図１２から図１４は、一連のフェイルオーバの動作の具体例を示した図である。ここでは、ドメインDOM-LAN0のフェイルオーバグループFOG-LN0に属するCHN1100、CHN1101、CHN1102、及びCHN1103の４台のCHNがある場合に、CHN1及びCHN2に障害が連続して発生する場合を考える。

【００７１】
図１２（ａ）の左図は、各CHNが正常に動作していることを示す。各CHNは、ハートビート格納エリア1312に格納された各々のHBMを定期的に更新している（HBMが定期的に更新されている状態をONと表現する）。
この場合のフェイルオーバ管理情報1311の内容を図１２（ａ）の右図に示す。実際には、リスト構造を用いて図１２右図のような情報がフェイルオーバ管理情報1311に格納される。

【００７２】
尚、矢印の終端にあるCHNが始端にあるCHNを監視する。又、始端のCHNに障害が発生した際には、矢印の終端にあるCHNがフェイルオーバを行う（図１２（ａ）左図の点線矢印も同じ関係を示す）。例えば、CHN1101はCHN1100を監視している。すなわち、CHN1100はCHN1101の監視対象CHである。この矢印で示される関係を「current」の引継関係と称する。

【００７３】
図１２（ｂ）は、CHN1101に障害が発生した状態を示す。CHN1101に障害が発生すると、CHN1101のHBM更新が停止する（HBMの更新が停止された状態をOFFと表現する）。CHN1101のHBM更新の停止をCHN1102が検出する。右図のフェイルオーバ管理情報1311は、この時点においては、まだ変更されない。

【００７４】
図１２（ｃ）は、CHN110１の処理がCHN1102に引き継がれた状態を示す。引継が行われる前のフェイルオーバ管理情報1311において、CHN1101の引継先（以下「テークオーバ先」という）はCHN1102に設定されている。したがって、CHN1101の障害を検出したCHN1102は、CHN1101を引継対象CHとして特定し、フェイルオーバ管理情報1311を右図の通り変更する。

【００７５】
図１２（ｃ）右図は、CHN1101がCHN1102の配下におかれ、テークオーバされていることを示す（同図上向き矢印）。この上向き矢印で示される関係を「takeover」の引継関係と称

する。又、CHN1100とCHN1102との間に、「current」の引継関係が設定される。これは、CHN1102の監視対象CHにCHN1100が加わったことを示す。一方、CHN1100とCHN1101との間の「current」の引継関係及びCHN1101とCHN1102との間の「current」の引継関係は、「default」の引継関係（同図点線矢印）に変更される。尚、実線の引継関係は「アクティブな」引継関係を示し、お互いが監視対象CH（または引継対象CH）であることを示す。一方、点線の引継関係は「インアクティブな」引継関係を示す。インアクティブな関係は、監視又は引継が実行される関係ではないことを示す。

【００７６】

以上のフェイルオーバ管理情報1311の変更により、CHN1102は「takeover」の引継関係と、「current」の引継関係の２つのアクティブな引継関係を保持することになる。これにより、CHN1102の監視対象CHはCHN1101とCHN1100の２つになる（図１２（ｃ）左図）。

【００７７】

図１３（ａ）は、CHN1102に障害が発生した場合を示している。CHN1102に障害が発生すると、CHN1102のHBM更新が停止する。CHN1102のHBM更新の停止をCHN1103が検出する。右図のフェイルオーバ管理情報1311は、この時点においては、まだ変更されない。

図１３（ｂ）は、CHN1102の処理がCHN1103に引継がれた状態を示す。引継が発生する前のフェイルオーバ管理情報1311において、CHN1102のテークオーバ先はCHN1103に設定されている。したがって、CHN1102の障害を検出したCHN1103は、CHN1102を引継対象CHとして特定し、フェイルオーバ管理情報1311を右図の通り変更する。同図は、障害が発生したCHN1102がCHN1103の配下におかれ、かつCHN1102の配下におかれていたCHN1101もCHN1103の配下におかれ、共にCHN3にテークオーバされることを示している（同図２本の上向き矢印）。すなわち、CHN1101とCHN1103との間、及びCHN1102とCHN1103との間に「takeover」の引継関係が設定される。一方、CHN1101とCHN1102との間の「takeover」の関係及びCHN1100とCHN1102との間の「current」の引継関係は解消される。又、CHN1100とCHN1103との間に新たに「current」の引継関係が設定される。さらに、CHN1102とCHN1103との間の「current」の引継関係は「default」に変更される。尚、CHN1100とCHN1101との間、及びCHN1101とCHN1102との間の「default」の引継関係は維持される。

【００７８】

以上のフェイルオーバ管理情報1311の変更により、CHN1103は２つの「takeover」の引継関係と、１つの「current」の引継関係の３つのアクティブな引継関係を保持することになる。これにより、CHN1103の監視対象CHは、CHN1101、CHN1102及びCHN1100の３つになる（左図）。

【００７９】

図１３（ｃ）は、CHN1101の障害が復旧された状態を示す。CHN1101の障害が復旧されると、CHN1101のHBM更新が再開される。CHN1101のHBM更新の再開をCHN1103が検出する。右図のフェイルオーバ管理情報1311は、この時点においては、まだ変更されない。

【００８０】

図１４（ａ）は、CHN1103からCHN1101へ処理が戻された状態を示す。CHN1101が復旧する前のフェイルオーバ管理情報1311において、CHN1101のテークオーバ先はCHN1103に設定されている。したがって、CHN1101の復旧を検出したCHN1103は、フェイルオーバ管理情報1311を右図の通り変更する。同図において、復旧されたCHN1101においては、CHN1103との間の「takeover」の引継関係が解消される。又、CHN1101とCHN1103との間の「default」の関係が「current」な引継関係へ変更される。又、CHN1100とCHN1101との間の関係は、「default」から「current」の関係に変更される。これは、CHN1103からCHN1101に処理が差し戻し（テークバック）されたことを示している。さらに、CHN1100とCHN1103との間の「current」な引継関係は解消される。尚、CHN1102とCHN1103との間の状態に変化はない。

【００８１】

以上のフェイルオーバ管理情報1311の変更により、CHN1103は、１つの「takeover」の引継関係と、１つの「current」の引継関係の２つのアクティブな引継関係を保持することになる。これにより、CHN1103の監視対象CHはCHN1101とCHN1102の２つになる（図１４（

ａ）左図）。

【００８２】

図１４（ｂ）は、CHN1102の障害が復旧された状態を示す。CHN1102の障害が復旧されると、CHN1102のHBM更新が再開される。CHN1102のHBM更新の再開をCHN1103が検出する。右図のフェイルオーバ管理情報1311は、この時点においては、まだ変更されない。

【００８３】

図１４（ｃ）は、CHN1103からCHN1102へ処理が戻された状態を示す。CHN1102が復旧する前のフェイルオーバ管理情報1311において、CHN1102のテークオーバ先はCHN1103に設定されている。したがって、CHN1102の復旧を検出したCHN1103は、フェイルオーバ管理情報1311を右図の通り変更する。同図において、復旧されたCHN1102においては、CHN1103との間の「takeover」の関係が解消される。又、CHN1102とCHN1103との間の「default」の引継関係が「current」な引継関係へ変更される。又、CHN1101とCHN1102との間の関係は、「default」から「current」の関係に変更される。これは、CHN1103からCHN1102に処理がテークバックされたことを示している。また、CHN1101とCHN1103との間の「current」な引継関係は解消される。

以上の引継関係の変更により、図１２（ａ）の状態に復帰する。

【００８４】

本実施形態によれば、一台の記憶装置システムに複数の種類のブロックＩ／Ｏインタフェースを備えるCHとファイルＩ／Ｏインタフェースを備えるCHを混在させることができ、さらに複数のネットワークのドメインに接続することができる。また、このような構成の下で、適切なフェイルオーバグループを構成することができ、フェイルオーバグループ内で複数のCHに障害が連続的に発生しても、他の正常なCHがその処理を引き継ぐことができる。

【００８５】

尚、本実施形態では、監視対象CHと引継対象CHが同一であるとして説明したが、監視対象CHと引継対象CHが異なる、例えば、CHN1101を監視するのはCHN1102だが、処理を引継ぐのがCHN1103というような構成とすることも可能である。本構成の場合、CHN1102とCHN1103との間で情報の交換を行う必要がある。この点に付いては、第２の実施形態を説明する際に説明する。

【００８６】

また、上述した実施形態においては、記憶装置システム１は、予め規定されたフェイルオーバ管理情報に従い、静的に引継先CH（又は引継対象CH）を決定していた。しかし、一つのCHに複数のCHの処理をフェイルオーバすると、フェイルオーバを引き受けたCHの稼働率が著しく高くなってしまう。

【００８７】

本発明の第２の実施形態においては、記憶装置システム自身が、各CHの稼働率を収集、記録し、同一のフェイルオーバグループの中で稼働率が最も低いCHを引継先CHとして決定する。そして、引継先CHとして決定されたCHに、障害が起こったCHの処理を引継ぐような方法を採用する。

【００８８】

具体的には、まず、図１２（ｂ）に示した第１の実施形態の「引継関係」を「監視関係」と定義しなおす。具体的には、図の矢印の引線は引継の関係を表現するのではなく、あるCHに障害が発生しているかどうかを監視するだけの関係を表現するものとする。

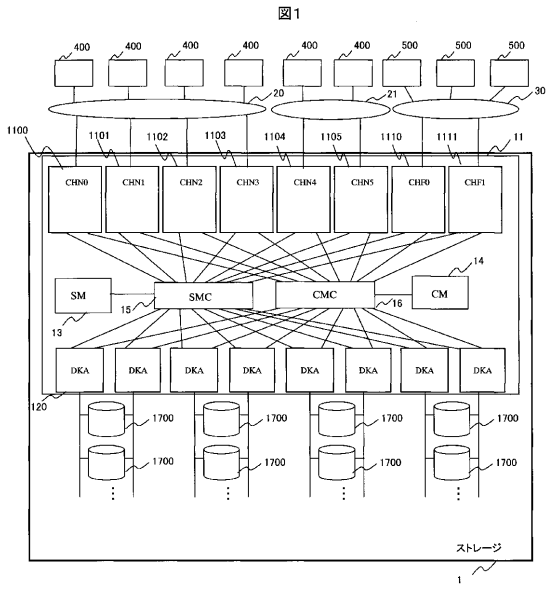
【００８９】

また、各CHは、各々の中央制御部11001の稼働率を測定し、その測定結果を定期的にチャネルアダプタ管理テーブル1310に保存する。具体的には、中央制御部11001が実行する作業が無いときにアイドルプロセスを実行するようにする。そして、一定期間のうち、アイドルプロセスが実行された時間を計測することで、一定期間のうちの中央制御部11001の稼働率を算出する。なお、一定時間は任意の値でかまわないが、測定オーバヘッドを加味し、プロセッサクロックに対して十分に大きな時間間隔、例えば１秒程度であることが望

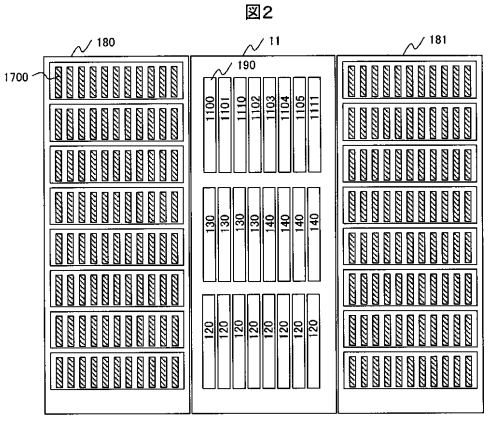
ましい。
【００９０】
引継先CHの特定は、以下のように行う。あるCHが第１の実施形態と同様に、ハートビートマーク格納エリア1312を監視し、監視対象CHであるCHの障害を検出する。障害を検出したCHは、チャネルアダプタ管理テーブル1310を参照し、同一フェイルオーバグループに属する正常なCHのうち、その時点で、もっとも稼働率の低いCHを特定する。そして、障害を検出したCHは、もっとも稼働率の低いCHを引継先CHとして決定する。その後、障害を検出したCHは、フェイルオーバ管理情報1311を変更する。障害が発生したCHと引継先CHとして決定されたCHNとの間に「takeover」の関係が設定される。「default」および「current」の監視関係は第１の実施形態の図１２と同一である。
【００９１】
監視を行っているCHは、引継先CHとして決定されたCHに信号を送信する。信号を受信したCHはフェイルオーバ管理情報1311を参照することで、障害が発生したCHの引継先CHになったことを判断する。その後、引継先となったCHは、第１の実施形態と同様にフェイルオーバ処理を実施する。
本実施形態によれば、フェイルオーバを行うCHへの負荷の集中を避けることができる。
【００９２】
上記の説明では、ある時点の稼働率で引継先を決定するとしたが、稼働率の時間変動等も記録し、より長い時間範囲で負荷を分散するように引継先を決定することもできる。このようにすると、時間的に負荷変動が大きいシステムの場合、負荷均衡化の効果はさらに大きくなる。
【００９３】
その他、稼働率以外の決定方法として、CH当たりの接続クライアント数が平均化するようにフェイルオーバする方法や、CH当たりのアクセスディスク台数が平均化するようにフェイルオーバする方法等がある。
【００９４】
【発明の効果】
本発明によれば、ＮＡＳやＳＡＮに対応する複数種のインタフェースを１つの記憶装置システムに混在させることができる。これにより、システム構成に自由度を有し、かつ管理コストを低減できる記憶装置システムを提供できる。
【００９５】
また、複数のＮＡＳ又はＳＡＮを構成するインタフェースに対して、耐多重障害性を備える記憶装置システムを提供できる。
【図面の簡単な説明】
【図１】本発明を適用したシステムの実施形態の構成を示す図である。
【図２】本実施形態における記憶装置システムの概観を示す図である。
【図３】チャネルアダプタの外観を示す図である。
【図４】チャネルアダプタの構成を示す図である。
【図５】チャネルアダプタが有するメモリの内容構成を示す図である。
【図６】チャネルアダプタのグループ化の概念を示す図である。
【図７】共有メモリの内容構成を示す図である。
【図８】チャネルアダプタ管理テーブルの内容を示す図である。
【図９】障害発生及び引継元アダプタ側処理を示すフローチャートである。
【図１０】障害検出及び引継先アダプタ側処理を示すフローチャートである。
【図１１】復旧及び引継先アダプタ側処理を示すフローチャートである。
【図１２】フェイルオーバの動作の一例を示す図である。
【図１３】フェイルオーバの動作の一例を示す図である。
【図１４】フェイルオーバの動作の一例を示す図である。
【符号の説明】
1…記憶装置システム、1100～1105…ＮＡＳチャネルアダプタ、1110、1111…ファイバチ

ャネルアダプタ、120…ディスク制御アダプタ（DKA）、13…共有メモリ、14…キャッシュメモリ、15…共有メモリコントローラ、16…キャッシュメモリコントローラ、1700…記憶装置、400…ＮＡＳクライアント、500…ＳＡＮクライアント、20、21…LAN、30…SAN。

【図１】

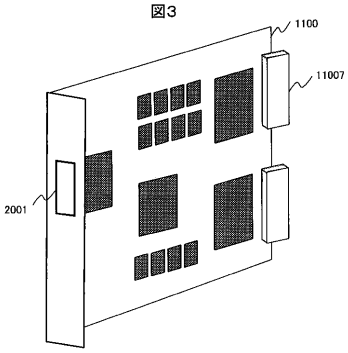


【図２】



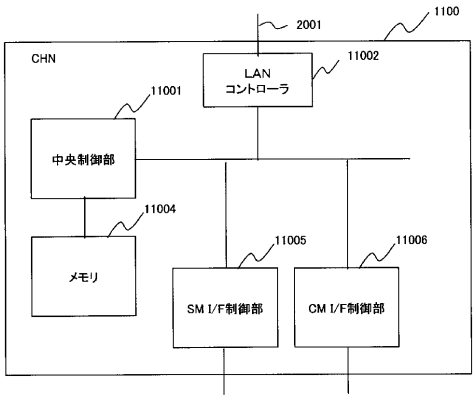
【図３】

図3



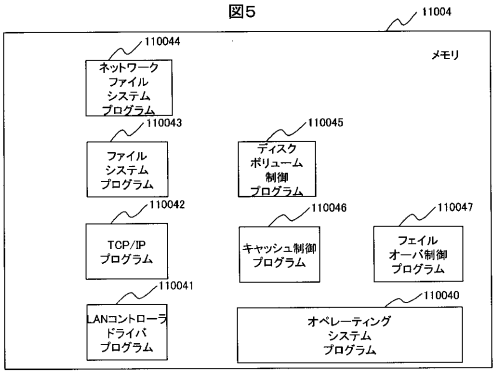
【図４】

図4



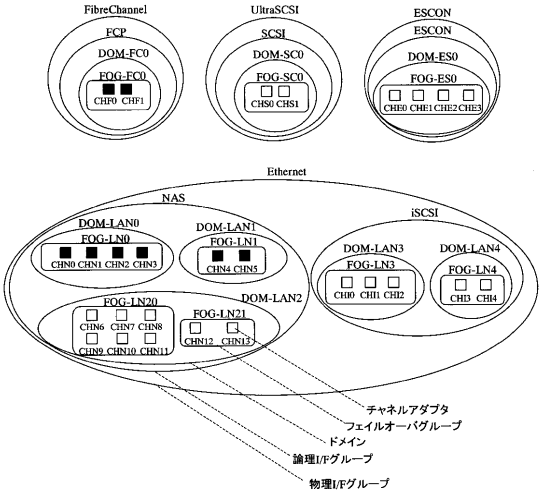
【図５】

図5



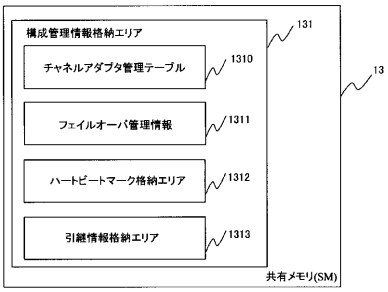
【図６】

図6



【図７】

図7



【図８】

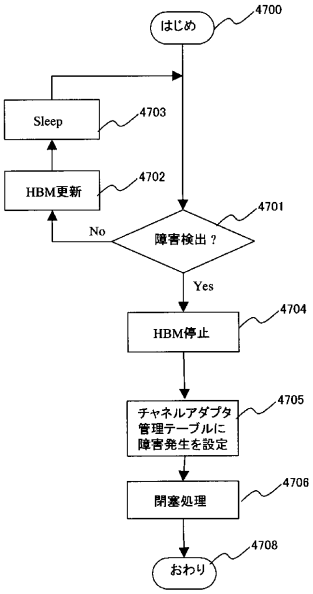
図8

1310

13101 チャネルアダプタ	13102 物理I/F グループ	13103 論理I/F グループ	13104 ドメイン	13105 フェイルオーバ グループ	13106 ステータス	13107 稼働率
CHN0	Ethernet	NAS	DOM-LAN0	FOG-LN0	Normal	48%
CHN1	Ethernet	NAS	DOM-LAN0	FOG-LN0	Failed	0%
CHN2	Ethernet	NAS	DOM-LAN0	FOG-LN0	Failed	0%
CHN3	Ethernet	NAS	DOM-LAN0	FOG-LN0	Normal TakingOver(CHN1) TakingOver(CHN2)	86%
CHN4	Ethernet	NAS	DOM-LAN1	FOG-LN1	Normal	32%
CHN5	Ethernet	NAS	DOM-LAN1	FOG-LN1	Normal	12%
CHF0	FibreChannel	FCP	DOM-FC0	FOG-FC0	Normal	74%
CHF1	FibreChannel	FCP	DOM-FC0	FOG-FC0	Normal	39%

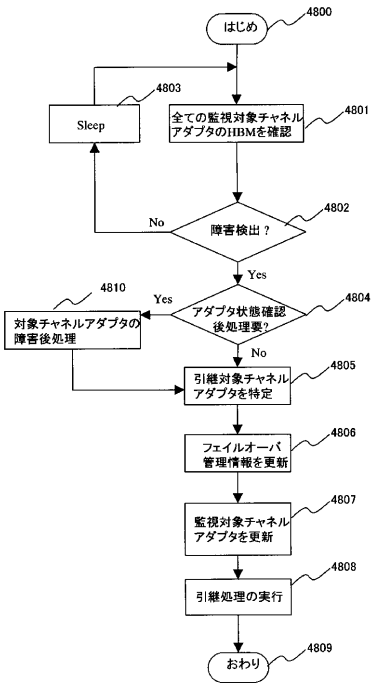
【図９】

図9



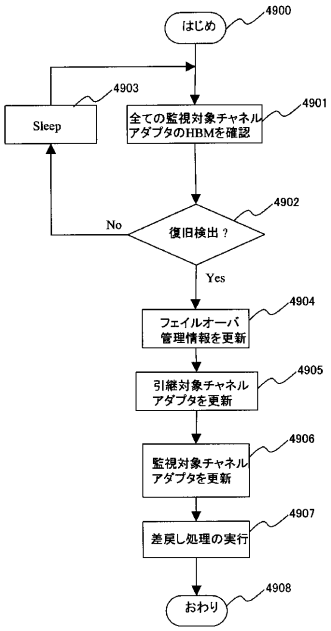
【図１０】

図10



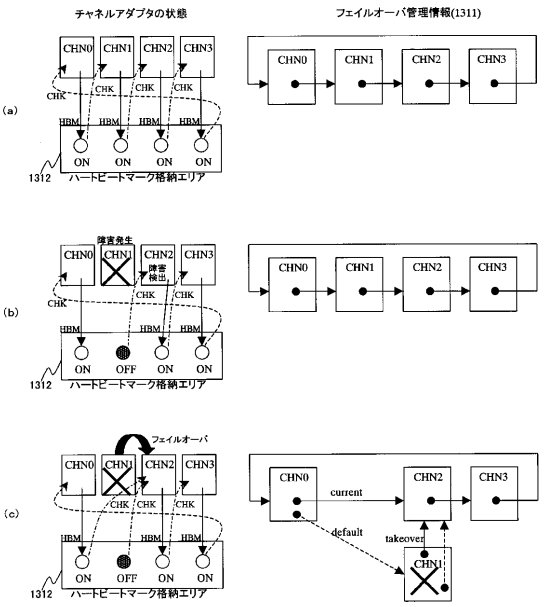
【図１１】

図11



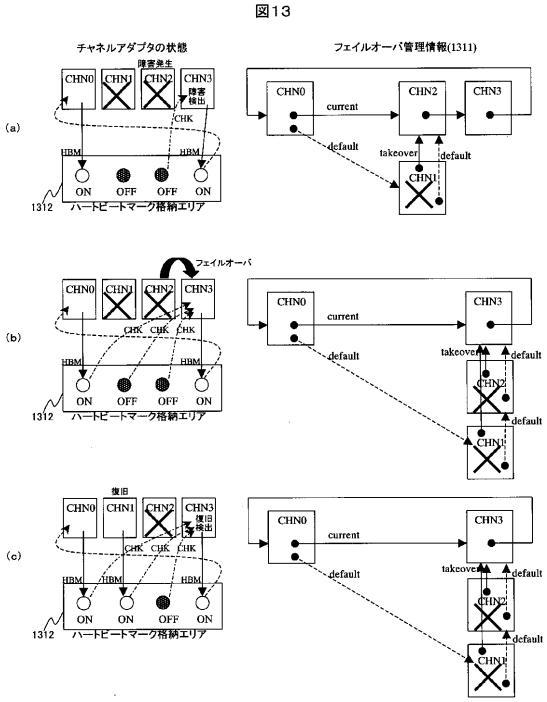
【図１２】

図12



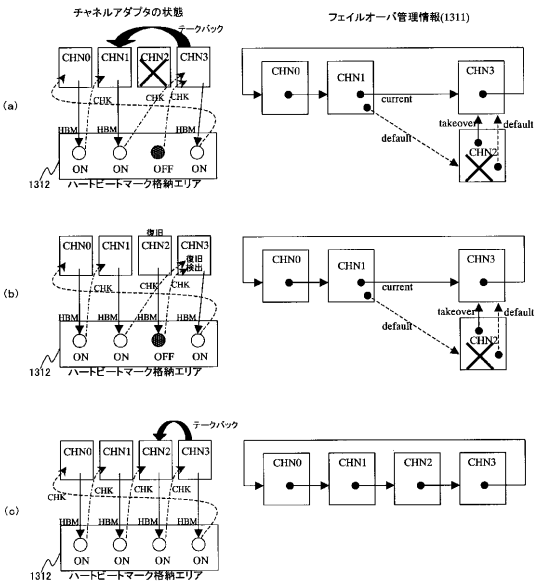
【図１３】

図13



【図１４】

図14



フロントページの続き

(72)発明者　薗田　浩二
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　北村　学
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
(72)発明者　▲高▼田　豊
神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

審査官　藤井　浩

(56)参考文献　特開２００１－２５６００３（ＪＰ，Ａ）
特開２００１－３２５２０７（ＪＰ，Ａ）
特開２０００－２７６３０６（ＪＰ，Ａ）
特開平１１－０３９１０３（ＪＰ，Ａ）
特表２００１－５０８２０８（ＪＰ，Ａ）
特開平０６－２８２３８５（ＪＰ，Ａ）

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