(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-21556
(P2004-21556A)
(43)公開日　平成16年1月22日(2004.1.22)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 11/16	G06F 11/16	310C	5B034
G06F 3/06	G06F 3/06	304N	5B042
G06F 11/20	G06F 11/20	310E	5B065
G06F 11/30	G06F 11/30	310H	

審査請求 未請求　請求項の数 21　ＯＬ　（全 18 頁）

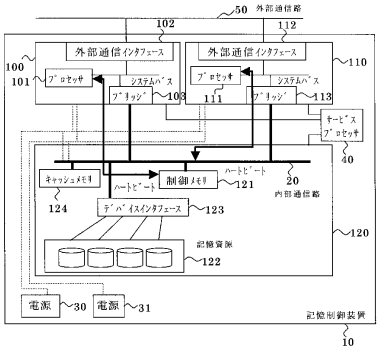
(21)出願番号　特願2002-174944(P2002-174944)
(22)出願日　平成14年6月14日(2002.6.14)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(74)代理人　100071283
弁理士　一色　健輔
(74)代理人　100084906
弁理士　原島　典孝
(74)代理人　100098523
弁理士　黒川　恵
(74)代理人　100112748
弁理士　吉田　浩二
(74)代理人　100110009
弁理士　青木　康

最終頁に続く

(54)【発明の名称】記憶制御装置およびその制御方法

(57)【要約】　（修正有）
【課題】内部通信路を用いて稼動情報を効率よく管理し、性能や信頼性が高く構築が容易で安価に移動情報の監視機能を実現する記憶制御装置を提供する。
【解決手段】本発明の記憶制御装置は、外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段に対するデータ入出力指示を送信する複数の制御部と、前記制御部間のデータ入出力指示及びデータ入出力を行うための内部通信路と、前記制御部が前記内部通信路を通じて他の前記制御部に自身の稼働情報を送信することで前記制御部が互いに他の前記制御部の稼働状態を監視する手段と、を備える。
【選択図】　図１



【特許請求の範囲】
【請求項１】
外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段に対するデータ入出力指示を送信する複数の制御部と、
前記制御部間のデータ入出力指示及びデータ入出力を行うための内部通信路と、
前記制御部が前記内部通信路を通じて他の前記制御部に自身の稼働情報を送信することで前記制御部が互いに他の前記制御部の稼働状態を監視する手段と、
を備えることを特徴とする記憶制御装置。
【請求項２】
外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段にデータ入出力指示を送信する複数の制御部と、
前記制御部間におけるデータ入出力指示及びデータ入出力を行い前記記憶手段が接続する内部通信路と、
前記制御部の稼働情報を前記制御部から前記内部通信路を通じて当該内部通信路に接続するメモリに送信しこれを当該メモリに記憶する手段と、
前記制御部が前記内部通信路を通じて前記メモリにアクセスし前記稼働情報に基づいて他の前記制御部の稼働状況を監視する手段と、
を備えることを特徴とする記憶制御装置。
【請求項３】
前記制御部が、
中央処理装置と、前記外部通信路に接続するための外部通信インタフェースと、
前記内部通信路と接続するための内部通信インタフェースと、
を備えることを特徴とする請求項２に記載の記憶制御装置。
【請求項４】
前記稼働情報には前記制御部の障害有無を示す情報が含まれており、
前記制御部が前記監視手段により他の前記制御部に障害が生じていることを認知した場合、障害が生じている前記制御部が行っていた処理を他の前記制御部に引き継ぐ手段を備えることを特徴とする請求項１または２のいずれかに記載の記憶制御装置。
【請求項５】
前記メモリは前記制御部が担当する処理に関するリソース情報を記憶する手段を備え、
前記処理の引き継ぎ先となる前記制御部が、前記リソース情報にアクセスし当該制御部自身が引き継ぐべき前記処理を認知する手段を備えることを特徴とする請求項４に記載の記憶制御装置。
【請求項６】
前記リソース情報には、前記制御部が前記外部通信路もしくは前記内部通信路による通信に必要な情報が含まれることを特徴とする請求項５に記載の記憶制御装置。
【請求項７】
前記リソース情報には、前記データ入出力指示において指定する記憶領域指定情報が含まれることを特徴とする請求項５に記載の記憶制御装置。
【請求項８】
前記リソース情報には、外部通信路上における前記制御部のネットワークアドレスが含まれることを特徴とする請求項５に記載の記憶制御装置。
【請求項９】
前記稼働情報には、前記中央処理装置、前記外部通信インタフェース、前記内部通信インタフェースのいずれかについての障害の有無を示す情報が含まれることを特徴とする請求項３に記載の記憶制御装置。
【請求項１０】
前記障害の有無を示す情報が、前記制御部が前記内部通信路を介して前記メモリに一定間隔で送信され、前記制御部に対応づけて記憶されるタイムスタンプであり、
前記監視手段は、ある前記制御部に対応する前記タイムスタンプが一定時間以上更新され

ていない場合に、そのタイムスタンプに対応づけられている前記制御部に障害が生じていると認知することを特徴とする請求項９に記載の記憶制御装置。
【請求項１１】
前記監視手段が、前記稼働情報にアクセスできない場合に、前記内部通信路に障害が生じていると判断する手段を備えることを特徴とする請求項４に記載の記憶制御装置。
【請求項１２】
前記監視手段が、前記稼働情報にアクセスできかつ前記のある制御部の前記タイムスタンプが一定時間以上更新されていない場合に、その制御部自体に障害が生じていると判断する手段を備えることを特徴とする請求項４に記載の記憶制御装置。
【請求項１３】
前記記憶手段が、中央処理装置と、前記メモリと、ディスクドライブなどの記憶資源に対するデータ入出力を行うデバイスインタフェースとを備えることを特徴とする請求項２に記載の記憶制御装置。
【請求項１４】
前記稼働情報、もしくは、前記リソース情報は、前記メモリに記憶されることを特徴とする請求項１３に記載の記憶制御装置。
【請求項１５】
前記稼働情報、もしくは、前記リソース情報は、前記記憶資源に記憶されることを特徴とする請求項１３に記載の記憶制御装置。
【請求項１６】
前記制御部が、データ入出力要求をファイル名単位で行うファイルシステムを備え、ファイル名により指定されるデータを単位として前記記憶手段に対する前記データ入出力指示を送信する手段を備えることを特徴とする請求項２に記載の記憶制御装置。
【請求項１７】
前記記憶手段が、ハードディスクなどの記憶資源に対するデータ入出力を制御するデバイスインタフェースと、
キャッシュメモリと、
前記キャッシュメモリを介して前記記憶資源に対するデータの読み書きを行う手段と、
を備えることを特徴とする請求項２に記載の記憶制御装置。
【請求項１８】
前記内部通信路が、複数の通信経路により冗長構成されることを特徴とする請求項１に記載の記憶制御装置。
【請求項１９】
前記制御部が、前記内部通信路とは別に設けられ前記制御部間を接続する他の通信路を通じて他の前記制御部に自身の稼働情報を送信する手段を備え、
前記制御部が前記他の通信路を通じて送信しようとする前記稼働情報を前記内部通信路に対して送信する変換手段を備えることを特徴とする請求項２に記載の記憶制御装置。
【請求項２０】
外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段に対するデータ入出力指示を送信する複数の制御部と、前記制御部間のデータ入出力指示及びデータ入出力を行うための内部通信路と、を備える記憶制御装置の制御方法であって、
前記制御部が前記内部通信路を通じて他の前記制御部に自身の稼働情報を送信することで前記制御部同士が他の前記制御部の稼働状態を監視することを特徴とする記憶制御装置の制御方法。
【請求項２１】
外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段にデータ入出力指示を送信する複数の制御部と、前記制御部間におけるデータ入出力指示及びデータ入出力を行い前記記憶手段が接続する内部通信路と、を備える記憶制御装置の制御方法であって、
前記制御部の稼働情報を前記制御部から前記内部通信路を通じて当該内部通信路に接続するメモリに送信しこれを当該メモリに記憶し、

前記制御部が前記内部通信路を通じて前記メモリにアクセスし前記稼働情報に基づいて他の前記制御部の稼働状況を監視することを特徴とする記憶制御装置の制御方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明は、外部通信路から入力されるデータ入出力指示を受信して記憶デバイスに対するデータ入出力制御を行い内部通信路で接続された複数の制御部を備える記憶制御装置に関し、とくに制御部間で交換する稼働情報の通信路として内部通信路を用い、稼働情報を効率よく管理し、性能や信頼性が高く構築が容易で安価に制御部間での稼働情報の監視機能を提供する技術に関する。

【０００２】

【従来の技術】

近年、ストレージ製品の一つとして、内部にファイルシステムが実装され、ファイル指定によるデータ入出力要求を取り扱うことができるようにした記憶制御装置である、いわゆる、ＮＡＳ（Ｎｅｔｗｏｒｋ　Ａｔｔａｃｈｅｄ　Ｓｔｏｒａｇｅ）サーバに対するニーズが高まっている。また、このようなＮＡＳサーバとしては、処理能力の向上等を目的として、それぞれがＬＡＮ等の外部通信路に接続し、それぞれが個別に外部通信路を通じて送られてくる処理要求に応答できるようにした複数の制御部を、同一の筐体内に収容する構成の製品が存在する。

【０００３】

【発明が解決しようとする課題】

ところで、記憶制御装置はミッションクリティカルな状況で使用される場合が多く、一般に高い可用性が要求される。

ここで、ＬＡＮ上のコンピュータ同士の障害検知に際しては、従来から、ＬＡＮ上のコンピュータ間でハートビートメッセージを交換して互いの稼働状態を監視し合うことで、可用性を向上させる仕組みが知られている。

【０００４】

例えば、特開２０００－２２２３７３号公報には、クラスタ化コンピュータシステムにおいて、クラスタを構成するコンピュータ同士でＬＡＮを通じて定期的にハートビートメッセージを交換し、お互いに稼働状態を監視し合うようにし、ハートビートメッセージの交換が正常に行われていない場合には、適宜制御部のデータサービスは他の正常な制御部へ引き継ぐようにした仕組みが開示されている。また、特開２００１－１００９４３号公報には、２台のＰＣサーバが、ハートビートを行うための通信を、ディスク装置が接続してあるＳＣＳＩバスを用いて行うようにしたクラスタシステムが開示されている。

【０００５】

一方、前記図１の構成の記憶制御装置のように、複数の制御部を備える構成の記憶制御装置においては、各制御部は同一筐体内に実装されてはいるものの、外部通信路からの要求に対してそれぞれが個別にサービスを提供している。

【０００６】

そこで、このような構成の記憶制御装置においても、前記公報のように、制御部間で稼働状態を監視し、また、ある制御部における障害を検知した場合には、その業務を他の正常な制御部に引き継ぐ仕組みを設けることは、可用性の向上に有効であると考えられる。

【０００７】

ここでこのような仕組みを設ける場合、稼働情報を伝達する通信路に何を用いるかが問題となる。例えば、前記公報のクラスタ化コンピュータシステムの場合には、このような稼働情報（この場合はハートビートメッセージ）の通信路として、コンピュータ間を結ぶＬＡＮ等の外部通信路や、専用の通信路を用いている。

【０００８】

しかしながら、外部通信路は、経路途中の信頼性が充分で無く、また、通信速度も充分でない、もしくは、一定しないといった問題がある。また、専用の通信路を設ける場合は、

専用の設備が必要となり、余分なコストが発生することになる。
【０００９】
一方、前述した図１の構成の記憶制御装置においては、制御部間は、例えば、回路基板上に形成されＣＰＵとメモリを結ぶ制御バスのようにデータ入出力指示及びデータ入出力を行うための内部通信路により互いに接続されている。そして、内部通信路は一般にＬＡＮなどの外部通信路よりも伝送能力が高く信頼性も高い。従って、内部通信路を稼働情報の通信路として用いることで、信頼性の高い制御部間で稼働状態を監視する仕組みを、容易かつ安価に実現することは可能であると考えられる。
【００１０】
本発明は、このような観点に基づいてなされたもので、内部通信路を用いて稼働情報を効率よく管理し、性能や信頼性が高く構築が容易で安価に稼働情報の監視機能を実現する記憶制御装置を提供することを目的とする。
【００１１】
【課題を解決するための手段】
この目的を達成する本発明のうち主たる発明は、
外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段に対するデータ入出力指示を送信する複数の制御部と、
前記制御部間のデータ入出力指示及びデータ入出力を行うための内部通信路と、
前記制御部が前記内部通信路を通じて他の前記制御部に自身の稼働情報を送信することで前記制御部が互いに他の前記制御部の稼働状態を監視する手段と、
を備えることを特徴とする。
【００１２】
すなわち、一般に外部通信路に比べて信頼性が高く、高速大容量通信が可能な、例えば、回路基板上に形成されＣＰＵとメモリを結ぶ制御バスのようにデータ入出力指示及びデータ入出力を行うための内部通信路を通じて稼働情報を通知するようにしたことで、記憶制御装置の部品点数の削減が図られ、装置コストを抑えて信頼性の高いシステムを実現できる。
【００１３】
【発明の実施の形態】
＝＝＝開示の概要＝＝＝
本明細書および添付図面の記載により、少なくとも、以下の事項が明らかとなる。
外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段に対するデータ入出力指示を送信する複数の制御部と、
前記制御部間のデータ入出力指示及びデータ入出力を行うための内部通信路と、
前記制御部が前記内部通信路を通じて他の前記制御部に自身の稼働情報を送信することで前記制御部が互いに他の前記制御部の稼働状態を監視する手段と、
を備えることを特徴とする記憶制御装置。
この記憶制御装置によれば、例えば、回路基板上に形成されＣＰＵとメモリを結ぶ制御バスのようにデータ入出力指示及びデータ入出力を行うための内部通信路を通じて稼働情報（例えば、ハートビートメッセージ）を通知することで、記憶制御装置の部品点数の削減が図られ、装置コストを抑えて信頼性の高いシステムを実現できる。
【００１４】
外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段にデータ入出力指示を送信する複数の制御部と、前記制御部間におけるデータ入出力指示及びデータ入出力を行い前記記憶手段が接続する内部通信路と、前記制御部の稼働情報を前記制御部から前記内部通信路を通じて当該内部通信路に接続するメモリに送信しこれを当該メモリに記憶する手段と、前記制御部が前記内部通信路を通じて前記メモリにアクセスし前記稼働情報に基づいて他の前記制御部の稼働状況を監視する手段と、を備えることを特徴とする記憶制御装置。
このような構成の記憶制御装置によれば、制御部とは別体の前記メモリ（例えば、後述す

る制御メモリ）に存在する稼働情報（例えば、後述する「稼働状態管理テーブル」）に障害の状態を管理することが可能となるため、内部通信路の稼働状態に関する情報などと組み合わせて、障害の原因や発生部分をより細かく特定することができる。

【００１５】

また、かかる記憶制御装置においては、例えば、前記制御部は、中央処理装置と、前記外部通信路に接続するための外部通信インタフェースと、前記内部通信路と接続するための内部通信インタフェースと、を備えることとする。

【００１６】

また、かかる記憶制御装置においては、前記稼働情報には前記制御部の障害有無を示す情報が含まれており、前記制御部が前記監視手段により他の前記制御部に障害が生じていることを認知した場合、障害が生じている前記制御部が行っていた処理を他の前記制御部に引き継ぐ手段を備えることとする。これにより、記憶制御装置の可用性が確保されることになる。

【００１７】

また、かかる記憶制御装置において、前記メモリは前記制御部が担当する処理に関するリソース情報を記憶する手段を備え、前記処理の引き継ぎ先となる前記制御部が、前記リソース情報にアクセスし当該制御部自身が引き継ぐべき前記処理を認知する手段を備えることとする。

【００１８】

また、かかる記憶制御装置において、前記リソース情報には、例えば、前記制御部が前記外部通信路もしくは前記内部通信路による通信に必要な情報、前記データ入出力指示において指定する記憶領域指定情報、外部通信路上における前記制御部のネットワークアドレスが含まれることとする。

【００１９】

また、かかる記憶制御装置において、前記稼働情報には、前記中央処理装置、前記外部通信インタフェース、前記内部通信インタフェースのいずれかについての障害の有無を示す情報が含まれることとする。

【００２０】

また、かかる記憶制御装置において、前記障害の有無を示す情報が、前記制御部が前記内部通信路を介して前記メモリに一定間隔で送信され、前記制御部に対応づけて記憶されるタイムスタンプであり、前記監視手段は、ある前記制御部に対応する前記タイムスタンプが一定時間以上更新されていない場合に、そのタイムスタンプに対応づけられている前記制御部に障害が生じていると認知することとする。

【００２１】

また、かかる記憶制御装置において、前記監視手段が、前記稼働情報にアクセスできない場合に、前記内部通信路に障害が生じていると判断する手段を備えることとする。また、前記監視手段が、前記稼働情報にアクセスできかつ前記のある制御部の前記タイムスタンプが一定時間以上更新されていない場合に、その制御部自体に障害が生じていると判断する手段を備えることとする。これにより、障害を細かく特定することが可能となる。

【００２２】

また、かかる記憶制御装置において、前記記憶手段が、中央処理装置と、メモリと、ディスクドライブなどの記憶資源に対するデータ入出力を行うデバイスインタフェースとを備えることとする。

【００２３】

また、前記稼働情報、もしくは、前記リソース情報は、前記メモリに記憶されることとする。

【００２４】

また、かかる記憶制御装置において、前記稼働情報、もしくは、前記リソース情報は、前記記憶資源に記憶されることとしてもよい。

【００２５】

また、かかる記憶制御装置において、前記制御部が、データ入出力要求をファイル名単位で行うファイルシステムを備え、ファイル名により指定されるデータを単位として前記記憶手段に対する前記データ入出力指示を送信する手段を備えることとする。すなわち、記憶制御装置が、例えば、ＮＡＳサーバとして用いられる場合である。

【００２６】

また、かかる記憶制御装置は、前記記憶手段が、ハードディスクなどの記憶資源に対するデータ入出力を制御するデバイスインタフェースと、キャッシュメモリと、キャッシュメモリを介して前記記憶資源に対するデータの読み書きを行う手段と、を備えることを特徴とする。また、かかる記憶制御装置は、前記内部通信路が、複数の通信経路により冗長構成されることとする。

【００２７】

また、かかる記憶制御装置は、前記制御部が、前記内部通信路とは別に設けられ前記制御部間を接続する他の通信路（例えば、後述する専用通信路）を通じて他の前記制御部に自身の稼働情報を送信する手段を備え、前記他の通信路に対して送信される稼働情報を前記内部通信路に対して送信する変換手段（例えば、後述するエミュレーションドライバ）を備えることを特徴とする。これにより汎用のクラスタソフトウエアなどを用いて、簡単かつ安価に稼働状態監視の仕組みを実現することができる。

【００２８】

また、本発明の記憶制御装置の制御方法は、外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段に対するデータ入出力指示を送信する複数の制御部と、前記制御部間のデータ入出力指示及びデータ入出力を行うための内部通信路とを備える記憶制御装置の制御方法であって、前記制御部が前記内部通信路を通じて他の前記制御部に自身の稼働情報を送信することで前記制御部同士が他の前記制御部の稼働状態を監視することを特徴とする。

【００２９】

また、本発明の他の記憶制御装置の制御方法は、外部通信路を通じて入力されるデータ入出力要求を受信して記憶手段にデータ入出力指示を送信する複数の制御部と、前記制御部間におけるデータ入出力指示及びデータ入出力を行い前記記憶手段が接続する内部通信路と、を備える記憶制御装置の制御方法であって、前記制御部の稼働情報を前記制御部から前記内部通信路を通じて当該内部通信路に接続するメモリに送信しこれを当該メモリに記憶し、前記制御部が前記内部通信路を通じて前記メモリにアクセスし前記稼働情報に基づいて他の前記制御部の稼働状況を監視することとする。

【００３０】

＝＝第１実施例＝＝

＜装置構成＞

まず、本発明を前述の図１に示す記憶制御装置１０に適用した実施例について説明する。

【００３１】

記憶制御装置１０は、例えば、回路基板上に形成されＣＰＵとメモリを結ぶ制御バス（システムバス）のようにデータ入出力指示及びデータ入出力を行うための内部通信路２０と、これに接続する複数の制御部１００，１１０、記憶装置１２０、冗長構成された電源装置３０，３１を備える。制御部１００，１１０や記憶装置１２０に接続するサービスプロセッサ４０は、制御部１００，１１０や記憶装置１２０の動作制御や各種設定、稼働状態監視などを行う。

【００３２】

制御部１００，１１０は、ＣＰＵなどで構成される中央処理装置としてのプロセッサ１０１，１１１、ＬＡＮなどの外部通信路５０に接続するための外部通信インタフェース１０２，１１２、内部通信路２０に接続するブリッジなどで構成された内部通信インタフェース１０３，１１３などを備える。

【００３３】

制御部１００，１１０では、ファイルシステム（不図示）が稼働し、制御部１００，１１

０は外部通信路５０からファイル名指定によるデータ入出力要求を取り扱う。つまり、制御部１００，１１０は、それぞれＬＡＮ上のファイルサーバとして機能するコンピュータとしての機能を備え、記憶制御装置１０は前述のＮＡＳサーバとして機能している。

【００３４】

一方、記憶装置１２０は、システム管理情報などが記憶される制御メモリ１２１、ハードディスクなどの記憶資源１２２、制御部１００，１１０から送信されてくる命令などに応じて記憶資源１２２に対するデータの書き込み／読み出しを実行するデバイスインタフェース１２３、キャッシュメモリ１２４などを備える。

【００３５】

なお、記憶資源１２２は、図１に示すように記憶制御装置１０に内蔵されている場合もあるし、また、記憶制御装置２０の外部の別筐体内に存在し、デバイスインタフェース１２３と適宜なインタフェースで接続していることもある。

【００３６】

＜基本動作＞

記憶制御装置１０の基本的な動作について説明する。

記憶制御装置１０がホストコンピュータなどの外部装置（不図示）から外部通信路５０を通じて入力されるデータ入出力要求を受信すると、プロセッサ１０１，１１１は、この要求に対応する指示コマンドやデータなどからなるデータ入出力指示を、内部通信路２０を介して制御メモリ１２１に送信する。制御メモリ１２１はこれを受信して記憶する。

【００３７】

ここで、例えば、前記データ入出力指示に含まれる前記指示コマンドがライト（Ｗｒｉｔｅ）コマンドであった場合、デバイスインタフェース１２３は、内部通信路２０を介してプロセッサ１０１，１１１にデータ送信要求を送信する。この要求を受信したプロセッサ１０１，１１１は、キャッシュメモリ１２４にライトデータを格納し、また、デバイスインタフェース１２３に対して割込み要求を送信する。この割り込み要求を受信したデバイスインタフェース１２３は、適宜な機会に、キャッシュメモリ１２４上の前記ライトデータを記憶資源１２２に書き込む。

【００３８】

一方、前記指示コマンドがリード（Ｒｅａｄ）コマンドであった場合、デバイスインタフェース１２３は、この指示コマンドに付帯指定される記憶資源１２２上の記憶領域に格納されているデータを読み出し、これをキャッシュメモリ１２４に格納し、さらに、読み出したデータを、内部通信路２０を介してプロセッサ１０１，１１１に送信する。

【００３９】

割り込み要求を受信したデバイスインタフェース１２３は、制御メモリ１２１に格納されている前記指示コマンドを参照し、キャッシュメモリ１２４内のデータと冗長データとを記憶資源１２２に転送する。

【００４０】

＜監視機能＞

各制御部１００，１１０は、互いに他の制御部１００，１１０に障害が発生したかどうかを監視している。各制御部１００，１１０がどの他のどの制御部１００，１１０の監視を担当するかは、例えば、サービスプロセッサ４０を介してオペレータなどが設定し、設定された情報は、制御メモリ１２１上に存在する図２に示す稼働状態管理テーブルの「監視対象の制御部」の項目２５１，２５４に登録される。

【００４１】

各制御部１００，１１０は、自身の監視対象として割り当てられている他の制御部１００，１１０の稼働状態を監視する。この監視により自身が担当する制御部１００，１１０に何らかの障害が生じていることを検知した場合、制御部１００，１１０は、検知した障害の内容に応じた処理を実行する。

【００４２】

以下、制御部１００，１１０による監視機能と、異常を検知した場合に実行される記憶制

御装置１０の機能について、制御部１１０の障害を制御部１００が検知する場合を例として図３のフローチャートとともに説明する。

【００４３】

制御部１１０は、内部通信路２０を通じて定期的（タイミングは任意に変更できる）に制御メモリ１２１にアクセスする（もしくは、ハートビートメッセージを送信する）。一方、記憶装置１２０は、前記アクセスがあると、アクセスのあった時刻をタイムスタンプとして稼働状態管理テーブルの制御部１１０のタイムスタンプの項目２５５に書き込む（Ｓ３１１）。すなわち、制御部１１０に異常が無ければタイムスタンプは定期的に更新されることになる。

【００４４】

一方、制御部１００は、稼働状態管理テーブルを参照するため、内部通信路２０を通じて定期的（定期的以外にも任意に設定してもよい）に制御メモリ１２１にアクセスする（Ｓ３１２）。

【００４５】

ここで制御部１００は、制御メモリ１２１にアクセスできなかった場合（Ｓ３１３）、内部通信路２０に何らかの障害が発生していると判断し、また、内部通信路２０に障害が発生している場合は、記憶装置１２０へのデータ入出力が正常に行えない状態にある可能性があるので、制御部１００は、例えば、サービスプロセッサ４０に指示を出すなどして、制御部１１０の記憶装置１２０に対するデータ入出力処理を停止させるとともに、この処理に関して制御部１１０が取得中のリソースを開放させる（Ｓ３１４）。

【００４６】

ここで、リソースとは、例えば、制御部１００，１１０が外部通信路５０もしくは内部通信路２０による通信に際し必要となる、ネットワークアドレス（例えば、ＩＰアドレス）などの情報、制御部１００，１１０が記憶手段に対して送信するデータ入出力指示において指定する記憶領域指定情報（例えば、制御部１１０がマウントしていた記憶領域に関する情報）などである。

【００４７】

一方、制御部１００は、制御メモリ１２１にアクセスできた場合、稼働状態管理テーブルの制御部１１０のタイムスタンプ２５５を参照し（Ｓ３１５）、アクセスした時刻とタイムスタンプとの差が一定時間以上であるかどうかを調べる（Ｓ３１６）。

【００４８】

ここで差が一定時間以上の場合には、制御部１００は、制御部１１０に障害が発生していると判断し、制御部１１０が担当している処理やリソースを制御部１００が引き継ぐ（Ｓ３１７）。

【００４９】

一方、差が一定時間に満たない場合には、さらに、稼働状態管理テーブル中の監視対象の制御部１１０の「状態」項目の内容２５６を参照する（Ｓ３１８）。ここでその内容が『正常』である場合には、制御部１００は、制御部１１０は正常に動作していると認知する（Ｓ３１９，Ｓ３２０）。他方、「状態」項目の内容２５６が『異常』であった場合には、制御部１１０に障害が発生したと判断し、前記と同様の方法により制御部１１０が担当している処理を引き継ぐ（Ｓ３１９，Ｓ３１７）。なお、稼働状態管理テーブルにおいて、『正常』もしくは『異常』は、ビット表現等の適宜な形式で記述される。

【００５０】

ところで、記憶制御装置１０において、稼働状態管理テーブルにおける「状態」項目の内容は、つぎのように管理されている。

例えば、制御部１００，１１０は、図４に示すように外部通信インタフェース１０２に障害が発生している場合、そのことを外部通信インタフェース１０２，１１２からプロセッサ１０１，１１１への直接の障害報告、もしくは、外部通信インタフェース５０に出した処理命令がタイムアウトする、といったことで認知する。

【００５１】

制御部１００，１１０は、外部通信インタフェース１０２，１１２に障害が発生していることを認知すると、内部通信路２０を通じて制御メモリ１２１上の稼働状態管理テーブルにアクセスし、障害が発生している制御部１００，１１０に対応する「状態」項目２５３，２５６に『異常』を書き込む。

【００５２】

一方、制御部１００，１１０は、図５に示すように内部通信インタフェース１０３，１１３に障害が発生している場合、そのことを内部通信インタフェース１０３，１１３からプロセッサ１０１，１１０への障害報告、もしくは、内部通信インタフェース１０３，１１３に出した処理命令のタイムアウトなどにより認知する。

【００５３】

制御部１００，１１０は、内部通信インタフェース１０３，１１３に障害が発生していることを認知すると、内部通信路２０を通じて制御メモリ１２１上の稼働状態管理テーブルにアクセスし、障害が発生している制御部１００，１１０に対応する「状態」項目に『異常』を書き込む。

【００５４】

図６はプロセッサ１０１に障害が発生している場合である。この場合、制御部１００は稼働状態管理テーブルに状態を書き込むことができない。しかしながら、制御部１１０のプロセッサ１１１が制御メモリ１２１を参照するため、プロセッサ１０１に障害している場合でも、その障害を検知することができる。

【００５５】

以上の実施例では、制御部１００が制御部１１０の監視を行う場合について説明したが、当然のことながら制御部１１０が制御部１００を監視する場合も同様の処理により行われる。

【００５６】

以上の実施例においては、制御メモリ１２１の稼働状態管理テーブルへの書き込みや参照などの制御部１００，１１０における障害監視のための通信を、内部通信路２０を介して行っている。このため、従来の外部通信路５０や専用の通信路による方式に比べ、障害監視のための通信を高速に行える。また、一般に内部通信路２０は、ＬＡＮなどの外部通信路５０に比べ伝送能力や信頼性に優れるため、障害監視のための通信を迅速かつ確実に行うことができる。

【００５７】

また、本来、制御部１００，１１０間やこれらと記憶装置１２０間で行われるデータ入出力処理等のために設けられている内部通信路２０を、障害監視のための通信に流用しているので、専用の通信路を設ける場合のように余分なハードウエアを増設する必要が無く、障害監視のための通信の仕組みを、容易かつ安価に構築できる。

【００５８】

内部通信路２０を介して制御部１００，１１０に接続する、制御部１００，１１０とは別体の、記憶装置１２０の制御メモリに存在する稼働状態管理テーブルに障害の状態を管理する仕組みであるため、内部通信路２０の稼働状態に関する情報などと組み合わせて、障害の原因や発生部分をより細かく特定することができる。

【００５９】

また、制御部１００，１１０が障害の発生を認知した場合における、処理引き継ぎのための各制御部１００，１１０が取得しているリソースに関する情報についても制御メモリ１２１に管理されるため、制御部１００，１１０に障害が発生した場合でも、記憶装置１２０は引き継ぐリソースを確認することができる。また、リソースの一元管理により管理負荷の軽減等が図られる。

【００６０】

また、内部通信路が冗長構成されている場合には、障害監視のための通信の安全性や確実性がさらに担保される。

【００６１】

なお、以上の実施例は、制御部が２つの場合であったが、これに限定されるものではなく、制御部が３つ以上の場合にも容易に拡張することができる。

【００６２】

＝＝＝既存クラスタシステムの利用＝＝＝

前述の公報にも記載されているように、従来から、外部通信路や専用の通信路を利用してコンピュータ間で障害監視のための通信を行うソフトウエアが存在する。ここでは、このようなソフトウエアを用いて障害監視のための通信に、内部通信路２０を利用する本発明の仕組みを実現する場合について説明する。

【００６３】

図７は、本発明の制御部に対応する、コンピュータ７２０，７３０が、専用の通信路により接続され、この専用通信路７５０を介してハートビートメッセージを伝送することで、コンピュータ７２０，７３０がお互いに稼働状態を監視し合う、従来のシステム構成である。

【００６４】

コンピュータ７２０，７３０は、それぞれ、クライアントからのファイルサービスの要求をネットワークインタフェース７２１、７３１、ネットワークドライバ７２２，７３２を経由しファイルシステム７２３，７３３で受信する。

【００６５】

ファイルシステム７２３，７３３は、記憶装置７２６に対するデータ転送が必要な場合、ストレージドライバ７２４，７３４、ストレージインタフェース７２５，７３５を経由し記憶装置７２６とデータ転送を行う。クラスタソフトウエア７２７，７３７は、ハートビート用ネットワークドライバ７２８，７３８、ハートビートネットワークインタフェース７２９，７３９を経由し、コンピュータ７３０上のクラスタソフトウエア７２７，７３７に対しハートビートメッセージを送信する。

【００６６】

このクラスタソフト７２７，７３７が導入されている図７の構成に、障害監視に内部通信路を利用する本発明を適用した場合が図８である。この図において、記憶制御装置８１０は、クラスタソフトウエア８４７，８５７からのアクセスを受領し、このアクセスをストレージドライバ８４４，８５４へのアクセスに変換するエミュレーションドライバ８４８，８５８を備えている。この構成により、汎用のクラスタソフトウエア８４７，８５７が発行したハートビートメッセージを内部通信路８４９を通じて伝送される稼働状態情報に変換すること、および、その逆の変換をすることができる。具体的には、例えば、エミュレーションドライバ８４８，８５８は、ＲＳ－２３２ＣやＬＡＮなどの他の通信路の伝送手順に従った通信によりクラスタソフトウエア８４７，８５７からのアクセスを受領して、これを内部通信路８４９上での通信に変換したり、逆に内部通信路８４９上での通信を前記他の通信路の伝送手順に変換してクラスタソフトウエア８４７，８５７に伝えるといった役割を果たす。なお、ＲＳ－２３２ＣやＬＡＮなどの他の通信路は、記憶制御装置１０がハードウエア／ソフトウエアとして実際に備えていてもよいし、備えていなくてもよい。また、前記変換機能部分は、エミュレーションドライバ８４８，８５８の内部に設けるのでは無く、ストレージドライバ８４４，８５４や記憶装置８４６における内部通信インタフェース（不図示）の機能を提供するファームウエア（不図示）などに設けるようにしてもよい。

【００６７】

以上に説明したように、既存のクラスタソフトウエアが導入されているシステムにおいては、エミュレーションドライバを導入するだけで本発明を実施することができる。

【００６８】

また、既存のクラスタソフトウエアが導入されていない場合には、既存のクラスタソフトウエアにエミュレーションドライバを組み合わせて導入することで、クラスタソフトウエアの機能部分に対する開発費が抑えられるため、低廉なコストで本発明を実施することができる。

【００６９】
＝＝＝内部通信路の他の構成＝＝＝
内部通信路が制御バスで無く、ファイバチャネル（ＦｉｂｒｅＣｈａｎｎｅｌ）、Ｉｎｔｅｌｌｉｇｅｎｔ　Ｉ／Ｏ、Ｒａｐｉｄ　Ｉ／Ｏなどの他のプロトコルに準拠した通信路で構成される場合を説明する。
図９は内部通信路をファイバチャネル（ＦｉｂｒｅＣｈａｎｎｅｌ）プロトコルによる通信路で構成した記憶制御装置９７１の構成を示す図である。制御部９５０，９６０は、内部通信インタフェースとしてファイバチャネルインタフェース９５３，９６３を備える。
【００７０】
一方、記憶装置９８０もファイバチャネルインタフェース９８６，９９６を備え、制御部９５０、９６０とはファイバチャネルスイッチ９５６，９６６、ファイバチャネルインタフェース９８６，９９２を介して接続する。
【００７１】
制御部９５０，９６０は、ファイバチャネルスイッチ９５６、９６６、ファイバチャネルインタフェース９５３，９６３を介して制御部９５０，９６０に接続され、制御メモリ９８２，９９２、書き込みデータ及びディスクドライブからの読出しデータを一時バッファリングするキャッシュメモリ９８３，９９３を備える。
【００７２】
制御部９５０，９６０は、デバイスインタフェース９９４を介してディスクドライブ群９８５へ接続される。制御部９５０、９６０、記憶装置９８０は、同一の筐体に実装される冗長構成の電源９７２，９７３より給電され、保守機構９７４により、各種動作設定や稼働管理が行われる。
【００７３】
制御部９５０がディスクドライブ群からなる記憶資源９８５とライト（Ｗｒｉｔｅ）データ転送を行う場合を例として、制御の流れ、データの流れを説明する。制御部９５０がサービスネットワーク９７０によりデータサービスの依頼を受けると、プロセッサ９５２は、サービス依頼を記憶装置９８０に対するＩ／Ｏコマンドに変換し、ファイバチャネルインタフェース９５３、ファイバチャネルスイッチ９５６、ファイバチャネルインタフェース９８６を介して、記憶装置９８０へＩ／Ｏコマンドを送信する。
【００７４】
送信されたＩ／Ｏコマンドは、制御メモリ９８２に格納される。Ｉ／Ｏコマンドがライトコマンドである場合、ファイバチャネルインタフェース９８６は、プロセッサ９５２にデータ転送を指示する。この指示によりプロセッサ９５２は、ファイバチャネルプロトコルに従い、記憶装置９８０に書き込みデータを送信する。記憶装置９８０は、送信されてきた書き込みデータを受信して、これを一旦キャッシュメモリ９８３に格納する。この書き込みデータは、デバイスインタフェース９８４により冗長データと共に記憶資源９８５に転送される。
【００７５】
つぎに、　図９のシステムにおいて、制御部９５０が障害になった場合には制御部９６０で、制御部９６０が障害になった場合には制御部９５０で、それぞれ障害が発生している制御部のリソースを引継ぎ、自動的にデータサービスなどの記憶制御装置１０の業務を続行できるようにした、記憶制御装置９７１の仕組みについて説明する。
【００７６】
記憶資源９８５には、図１０に示す稼働状態管理テーブルが格納されている。稼働状態管理テーブルには、各制御部を特定する情報１００１，１００２、監視すべき対象の制御部を特定する情報１００３，１００４についての項目がある。これらの項目は、例えば、フェイルオーバポリシーに従って設定される。タイムスタンプ１００５，１００６は、制御部９５０，９６０が稼働情報（ハートビートメッセージ）のＩ／Ｏを発行する際の制御部上の時刻を示す。また、この稼働状態管理テーブルの「状態」１００７，１００８項目には、制御部９５０，９６０の『正常』、『異常』を示す情報がセットされている。これら

の項目にセットされる情報には、必要に応じてクラスタシステムを構築する際のサービスネットワーク１０７０上の制御部に割り当てられた名称などの固有の識別子が用いられる。

制御部９５０上のプロセッサ９５２は、適宜なタイミング（例えば、定期的に）で記憶資源９８５の稼働情報管理テーブルの読み出し命令を送信し、稼働情報管理テーブルに示す全データもしくは一部のデータを取得する。稼働情報管理テーブルを読み出すと、プロセッサ９５２は、全データから制御部ネーム領域１００１，１００２を参照し、制御部９５０が有する制御部ネーム「ＳｅｒｖｅｒＡ」に対応づけられている監視対象制御部の制御部ネームを参照する。そして、この例では、「ＳｅｒｖｅｒＡ」の監視担当制御部として「ＳｅｒｖｅｒＢ」が割り当てられているので「ＳｅｒｖｅｒＢ」のタイムスタンプ１００６と状態情報１００８とを参照し、フェイルオーバ処理を実行する必要性の有無を判断する。

【００７７】

プロセッサ９５２は、「ＳｅｒｖｅｒＢ」のタイムスタンプ１００６と現在時刻との差が一定時間以上であった場合、もしくは、「状態」項目１００８に『異常』を示す情報がセットされていた場合には、制御部９６０に何らかの障害が発生していると認知する。また、プロセッサ９５２は、稼働状態管理テーブルにアクセスできることで、内部通信路に障害が発生していないと認知し、障害が制御部９６０におけるものであると判断して制御部９６０が行っていた処理やリソースの引き継ぎを開始する。

【００７８】

なお、プロセッサ９５２は、制御部９６０の障害を検出するかどうかに関わらず、「ＳｅｒｖｅｒＡ」のタイムスタンプと状態情報である図１０の１００５，１００７の項目の内容を適宜なタイミング（例えば定期的に）更新する。

【００７９】

制御部９５０，９６０が定期的に以上の処理を行うことにより汎用Ｉ／Ｏインタフェースを通信路として用いた障害監視等の仕組みが実現されることになる。

【００８０】

つぎに、ファイバチャネルスイッチ９５６に障害が生じた場合についての記憶制御装置の動作について説明する。

ファイバチャネルスイッチ９５６が障害となった場合、制御部９５０はＩ／Ｏ入出力が不可となるが、通常のフェイルオーバ方式により制御部９６０の処理を継続させた場合には制御部９６０に処理が集中し、性能が劣化する可能性がある。

【００８１】

そこで、このような場合には、汎用Ｉ／Ｏ切り替えソフトウエア等により、制御部９５０の入出力経路を、ファイバチャネルインタフェース９５３→ファイバチャネルスイッチ９５６→ファイバチャネルインタフェース９８６という通常の経路から、ファイバチャネルインタフェース９５４→ファイバチャネルスイッチ９６６→ファイバチャネルインタフェース９９６という経路に変更してＩ／Ｏを継続する。また、同時に、ディスクボリューム群９８５に制御部９５０から行われていた制御部９５０のアクセス状態情報を格納、確認するためのＩ／Ｏも経路を切り替える。

【００８２】

この例では、制御部９５０は、ファイバチャネルインタフェース９５４→ファイバチャネルスイッチ９６６→ファイバチャネルインタフェース９９６を順に経由して、図１０におけるタイムスタンプ１００５、制御部９５０の「状態」項目の内容１００７をディスクボリューム群９８５に格納する。

【００８３】

なお、以上はファイバチャネルスイッチ９５６が障害となった場合を例として説明したが、ファイバチャネルインタフェース９５３、ファイバチャネルスイッチ９５６、ファイバチャネルインタフェース９８６等に障害が発生した場合においても以上の処理を拡張できる。

【００８４】
以上のように、この実施例では、制御部９５０，９６０の状態情報を搭載した稼働情報の伝送を、外部通信路や専用の通信路では無く、記憶装置９８０への汎用Ｉ／Ｏインタフェース上で行っている。これによりハートビート用ネットワークインタフェースの導入が必要でなく、その分、手間やコストが削減され、また、信頼性の高い汎用Ｉ／Ｏインタフェースの利用により、信頼性の高い稼働状態監視機能を実現できる。
【００８５】
また、制御部の稼働状態を格納するディスクドライブ上に、クラスタ動作に必要な制御部ネットワークアドレス等のネットワーク属性、データが格納されているディスクドライブ等のリソース引継ぎ情報を格納することにより、各制御部より読み出し可能となるため、リソース引継ぎの設定管理を一元化し、管理コストの削減を図ることができる。
【００８６】
また、図９では２つのファイバチャネルスイッチ９５６，９６６を含む構成であったが、ファイバチャネルスイッチを全く含まない構成や、３つ以上のファイバチャネルスイッチを含む構成である場合にも本発明を拡張して適用することができる。また、ファイバチャネル以外の汎用Ｉ／Ｏインタフェースを用いる場合にも適用できる。
【００８７】
＝＝＝稼働情報の格納方式＝＝＝
つぎに、汎用クラスタソフトを使用して本発明のシステムを構築する際に、汎用クラスタソフトが送信する稼働情報を、記憶装置のディスクドライブなどの記憶資源（例えば、ディスクドライブ）上に格納するようにした実施例について説明する。
【００８８】
図１１はこのような構成の記憶制御装置１０００の一例である。制御部１１１０，１１５０は、クライアントからのファイルサービスの依頼をネットワークインタフェース１１１１，１１５１、ネットワークドライバ１１１２，１１５２を経由してファイルシステム１１１３，１１５３で受信する。
【００８９】
ファイルシステム１１１３，１１５３は、ストレージドライバ１１１４，１１５４、ファイバチャネルインタフェース１１１５，１１５５、ファイバチャネルスイッチ１１１６，１１５６を経由して記憶装置１１１９とデータ転送を行う。ストレージドライバ１１１４，１１５４内には、汎用Ｉ／Ｏ切り替えソフトウエアが組み込まれている。
【００９０】
データの転送経路に障害が発生した場合、ストレージドライバ１１１４，１１５４は、データ転送の経路をファイバチャネルインタフェース１１１７，１１５７、ファイバチャネルスイッチ１１１６，１１５６を経由する経路に切り替える。
【００９１】
このシステムにおいて、汎用クラスタソフトウエアは、ネットワークに対し制御部の状態情報を含む稼働情報を送信する。ここで当該稼働情報の要求をネットワークドライバとして受信し、汎用Ｉ／Ｏインタフェース経由で記憶装置へデータ転送を行う要求に変換するエミュレーションドライバ１１２０，１１６０を、各制御部において稼働させることで、汎用クラスタソフトウエアに対し、透過的に汎用Ｉ／Ｏインタフェース上でのハートビートの交換を行えるようにしている。
【００９２】
また、ストレージドライバ１０１４は、汎用Ｉ／Ｏインタフェース上のハートビートであるＩ／Ｏに対しても、データ転送の経路障害時に経路をファイバチャネルインタフェース１１１７，１１５７、ファイバチャネルスイッチ１１１６，１１５６を経由する経路に切り替えてデータ転送を継続する。これにより信頼性の高い稼働状態監視のための仕組みが実現される。
【００９３】
【発明の効果】

本発明の記憶制御装置にあっては、制御部間で交換する稼働情報の通信路として内部通信路を用い、稼働情報を効率よく管理し、性能や信頼性が高く構築が容易で安価に稼働情報の監視機能を提供することができる。

【図面の簡単な説明】

【図１】本発明の一実施例による記憶制御装置の構成を示す図である。

【図２】本発明の一実施例による稼働情報管理テーブルを示す図である。

【図３】本発明の一実施例として説明する記憶制御装置における、制御部が他の制御部が検知する処理を説明するフローチャートである。

【図４】本発明の一実施例として説明する記憶制御装置における、外部通信インタフェースに障害が発生している場合の障害認知等の処理を説明する図である。

【図５】本発明の一実施例として説明する記憶制御装置における、内部通信インタフェースに障害が発生している場合の障害認知等の処理を説明する図である。

【図６】本発明の一実施例として説明する記憶制御装置における、プロセッサに障害が発生している場合の障害認知等の処理を説明する図である。

【図７】従来における、コンピュータがお互いに稼働状態を監視し合う仕組みを説明する図である。

【図８】本発明の一実施例による、障害監視に内部通信路を用いた記憶制御装置の構成を示す図である。

【図９】本発明の一実施例による、内部通信路をファイバチャネルプロトコルの通信路で構成した記憶制御装置の構成を示す図である。

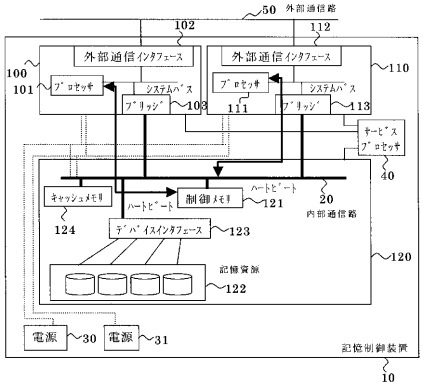
【図１０】本発明の一実施例による稼働情報管理テーブルを示す図である。

【図１１】本発明の一実施例による、汎用クラスタソフトを使用し、また、稼働情報をディスクドライブなどの記憶資源に格納するようにした記憶制御装置の構成を示す図である。

【符号の説明】

２０　内部通信路
５０　外部通信路
１０１，１１０　制御部
１２１　制御メモリ
１２４　キャッシュメモリ
１２２　記憶資源
１２３　デバイスインタフェース

【図１】



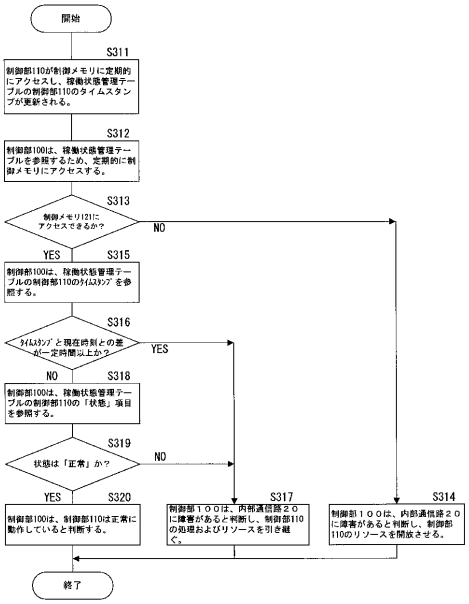
【図２】

251

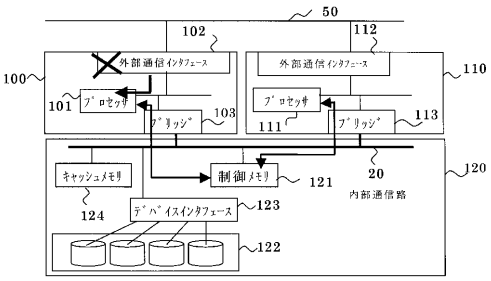
制御部	監視対象の制御部	タイムスタンプ	状態
100	110	2001/11/13 10:05:02	正常
110	100	2001/11/13 10:02:15	正常

254 252 255 253 256

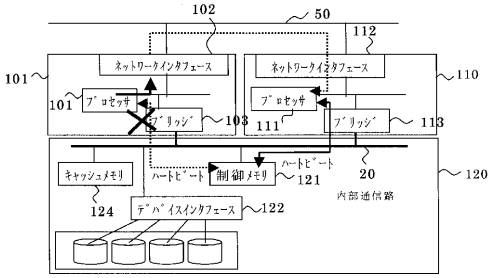
【図３】



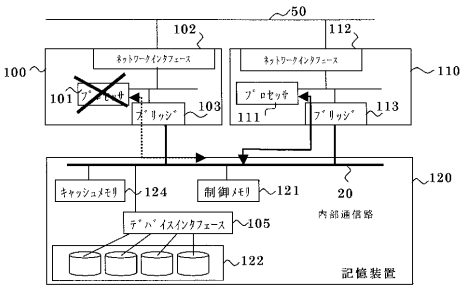
【図４】



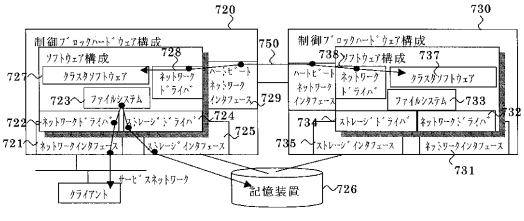
【図５】



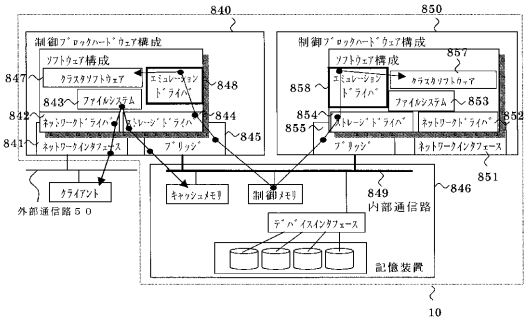
【図６】



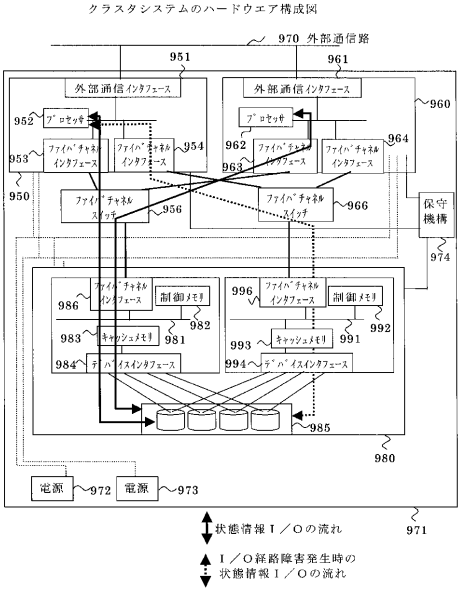
【図７】



【図８】



【図９】



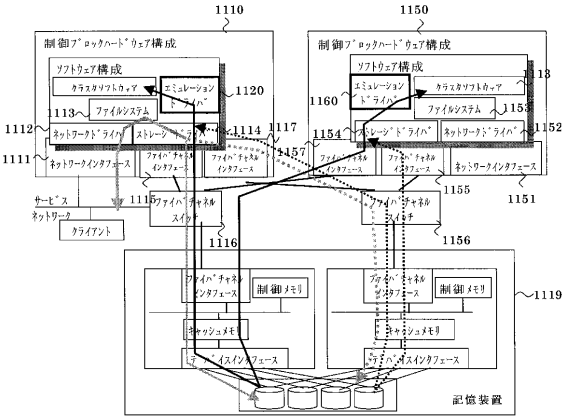
【図１０】

記憶装置内ハートビート用データ例

制御部ネーム	監視対象制御部	タイムスタンプ	状態
ServerA	ServerB	2001/11/13 10:05:02	正常
ServerB	ServerA	2001/11/13 10:02:15	正常

1001 1002 1003 1004 1005 1006 1007 1008

【図１１】



フロントページの続き

(72)発明者　室谷　暁
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
(72)発明者　中野　俊夫
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
(72)発明者　横畑　静生
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
(72)発明者　▲高▼本　賢一
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

Ｆターム(参考) 5B034 BB01　CC05　DD02
5B042 GA11　JJ04
5B065 BA01　CA11　EK02