(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2014-170394
(P2014-170394A)
(43)公開日　平成26年9月18日(2014.9.18)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 11/20	(2006.01)	G06F	11/20	310E	5B034
G06F 3/06	(2006.01)	G06F	3/06	301A	
		G06F	3/06	306B	

審査請求 未請求　請求項の数 10　ＯＬ　（全 21 頁）

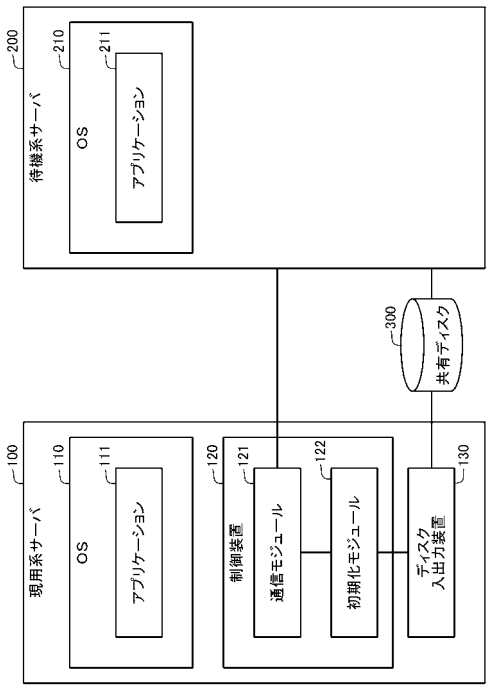
(21)出願番号　特願2013-42016(P2013-42016)
(22)出願日　平成25年3月4日(2013.3.4)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100124811
弁理士　馬場　資博
(74)代理人　100088959
弁理士　境　廣巳
(72)発明者　丹生谷　吉紀
東京都港区芝五丁目７番１号　日本電気株式会社内

Ｆターム(参考)　5B034 BB02　CC01　DD05

(54)【発明の名称】クラスタシステム

(57)【要約】
【課題】共有データを保護しつつ、迅速にフェイルオーバを行うこと。
【解決手段】本発明のクラスタシステムは、フェイルオーバ機能を有する現用系サーバ及び待機系サーバと、共有ディスクと、を備え、上記現用系サーバは、OSの影響を受けずに作動する制御装置と、共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えており、上記現用系サーバの制御装置は、待機系サーバと通信を行う通信モジュールと、現用系サーバの障害発生時にディスク入出力装置を初期化して通信モジュールを介して待機系サーバに通知する初期化モジュールと、を備えた。
【選択図】図１０



【特許請求の範囲】
【請求項１】
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備え、
前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えており、
前記現用系サーバの前記制御装置は、前記待機系サーバと通信を行う通信モジュールと、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールと、を備え、
前記待機系サーバは、前記現用系サーバの前記制御装置から前記Ｉ／Ｏフェンシングが完了した旨の通知を受けて、前記待機系サーバ上における前記アプリケーションにて前記現用系サーバ上における前記アプリケーションによる処理を引き継ぐフェイルオーバ処理機能を有する、
クラスタシステム。
【請求項２】
請求項１に記載のクラスタシステムであって、
前記現用系サーバは、当該現用系サーバにて作動しているソフトウェアの障害発生を検出して前記制御装置の前記通知モジュールに通知する障害検出モジュールを備え、
前記制御装置の前記初期化モジュールは、前記障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
クラスタシステム。
【請求項３】
請求項１又は２に記載のクラスタシステムであって、
前記現用系サーバの前記制御装置は、当該現用系サーバに装備されたハードウェアの障害発生を検出して前記制御装置の前記初期化モジュールに通知するハードウェア障害検出モジュールを備え、
前記制御装置の前記初期化モジュールは、前記ハードウェア障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
クラスタシステム。
【請求項４】
請求項１乃至３のいずれかに記載のクラスタシステムであって、
前記待機系サーバは、前記現用系サーバにおける障害発生を検出して、前記現用系サーバの前記初期化モジュールに通知する待機系側障害検出モジュールを備え、
前記現用系サーバの前記初期化モジュールは、前記待機系サーバの前記待機系側障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
クラスタシステム。
【請求項５】
請求項１乃至４のいずれかに記載のクラスタシステムであって、
前記初期化モジュールは、前記ディスク入出力装置にアクセスするためのアクセス情報

と、前記待機系サーバに情報を通知するためのアドレス情報と、を記憶保持しており、前記アクセス情報に基づいて前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記アドレス情報に基づいて前記通信モジュールを介して前記待機系サーバに通知する、
クラスタシステム。
【請求項６】
請求項５に記載のクラスタシステムであって、
前記現用系サーバは、当該現用系サーバの起動時に、前記アクセス情報及び前記アドレス情報が予め設定された設定ファイルを読み出し、当該設定ファイルに設定されている前記アクセス情報及び前記アドレス情報を前記制御装置の前記初期化モジュールに通知し、
前記初期化モジュールは、通知された前記アクセス情報及び前記アドレス情報を記憶保持する、
クラスタシステム。
【請求項７】
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムにおける前記現用系サーバであって、
前記現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えており、
前記制御装置は、前記待機系サーバと通信を行う通信モジュールと、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールと、を備えた、
現用系サーバ。
【請求項８】
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムにおける前記現用系サーバが、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、
当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、
前記制御装置に、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールによる処理、
を実行させるためのプログラム。
【請求項９】
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムによるフェイルオーバ方法であって、
前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、
前記現用系サーバの前記制御装置は、前記現用系サーバの障害発生時に、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行い

、当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知し、
　前記待機系サーバは、前記現用系サーバの前記制御装置から前記Ｉ／Ｏフェンシングが完了した旨の通知を受けて、前記待機系サーバ上における前記アプリケーションにて前記現用系サーバ上における前記アプリケーションによる処理を引き継ぐ、
フェイルオーバ方法。
【請求項１０】
　相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムによるフェイルオーバ方法であって、
　前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、
　前記現用系サーバの前記制御装置は、前記現用系サーバの障害発生時に、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行い、当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
フェイルオーバ方法。

【発明の詳細な説明】
【技術分野】
【０００１】
　本発明は、クラスタシステムにかかり、特に、フェイルオーバ機能を有するクラスタシステムに関する。
【背景技術】
【０００２】
　コンピュータの障害発生による処理の停止を抑制するために、例えば、特許文献１に示すように、現用系及び待機系の両サーバと共有ディスクを備えたフェイルオーバ型クラスタシステムが知られている。このようなフェイルオーバ型のクラスタシステムでは、両サーバにて作動する業務アプリケーションのデータは、現用系サーバからも待機系サーバからもアクセス可能な共有ディスクに格納される。そして、現用系サーバに障害が発生した場合、業務アプリケーションの切り替え（以降「フェイルオーバ」と記す）を行うことで、クラスタソフトウェアは待機系サーバで業務アプリケーションを起動する。待機系サーバで起動した業務アプリケーションは、共有ディスクに格納されたデータを使用し、現用系サーバの業務アプリケーションが停止した時点から業務処理を再開する。
【０００３】
　上述したように、共有ディスク構成のフェイルオーバ型クラスタシステムでは、データを共有ディスクに格納するため、現用系と待機系の両サーバから同時に共有ディスクに書込を行うと、データが破壊するおそれがある。そのため、通常は現用系サーバからの書込のみを行うように排他制御を行う。以下に排他制御を行う例を記載する。
【０００４】
　（１：フェイルオーバ時の排他制御）
　現用系サーバから待機系サーバにフェイルオーバを実行する場合は、現用系サーバからの共有ディスクアクセスを確実に停止した後に、待機系サーバで業務アプリケーションを起動する必要がある。
【０００５】
　（２：スプリットブレイン状態の排他制御）
　サーバ間のネットワークの障害によりクラスタがお互いの状態を認識できないスプリッ

トブレイン状態になった場合は、共有ディスクへのアクセスの有無により、業務アプリケーションを切り替えるサーバを決定する仕組み（スプリットブレイン解決）がある。このスプリットブレイン解決方式では、ネットワーク機器に障害が発生したサーバが共有ディスクへアクセスを行うと、障害サーバが健全に作動していると誤認識するおそれがあり、障害サーバから確実に共有ディスクにアクセスできないようにする必要がある。

【０００６】

上記のように、クラスタシステムにおいてフェイルオーバを行う場合には、正しく排他制御を行なって障害サーバから共有ディスクへのアクセスを確実に停止する必要がある。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】特開２０１２－１７３７５２号公報

【発明の概要】

【発明が解決しようとする課題】

【０００８】

ここで、クラスタシステムにおいて共有ディスクへのアクセスを停止する手法として、ディスクのアンマウントやFC（Fibre Channel）スイッチのポート閉塞、OS（Operating System）パニック、などが用いられている。ところが、上述した各手法には以下の様な課題がある。

【０００９】

ディスクのアンマウントの場合は、処理に時間がかかり、書込み中のプロセスが存在する場合にはアンマウント処理が失敗する、という問題がある。FCスイッチのポート閉塞の場合は、サーバ外のモジュールに接続するため接続時間がかかり、障害の種類によってはFCスイッチに接続できない、という問題がある。OSパニックの場合は、HBA（Host Bus Adapter）カードのキャッシュに残っているI/Oが書き込まれる可能性がある、という問題が生じる。

【００１０】

また、HW（Hardware）レイヤで共有ディスクへの書き込みを防ぐI/Oフェンシングを実施する場合は、カードへの電源供給を断つ方式などがある。HWレイヤでI/Oフェンシングを実施した場合は、確実なI/Oフェンシングが可能であるが、I/Oフェンシングの完了を待機系サーバに通知する手段がない。このため、例えば、待機系サーバは、ハートビートの通信タイムアウトを待ってフェイルオーバを開始しなければならず、迅速にフェイルオーバを行うことができない、という問題がある。

【００１１】

以上のように、フェイルオーバ型のクラスタシステムでは、共有データの保護を行いつつ、迅速にフェイルオーバを行うことができない、という問題があった。

【００１２】

このため、本発明の目的は、上述した課題である、共有データの保護を行いつつ、迅速にフェイルオーバを行うことができない、ということを解決することができるクラスタシステムを提供することにある。

【課題を解決するための手段】

【００１３】

本発明の一形態であるクラスタシステムは、

相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備え、

前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えており、

前記現用系サーバの前記制御装置は、前記待機系サーバと通信を行う通信モジュールと

、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールと、を備え、

前記待機系サーバは、前記現用系サーバの前記制御装置から前記Ｉ／Ｏフェンシングが完了した旨の通知を受けて、前記待機系サーバ上における前記アプリケーションにて前記現用系サーバ上における前記アプリケーションによる処理を引き継ぐフェイルオーバ処理機能を有する、

という構成をとる。

【００１４】

また、本発明の他の形態である現用系サーバは、

相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムにおける前記現用系サーバであって、

前記現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えており、

前記制御装置は、前記待機系サーバと通信を行う通信モジュールと、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールと、を備えた、

という構成をとる。

【００１５】

また、本発明の他の形態であるプログラムは、

相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムにおける前記現用系サーバが、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、

前記制御装置に、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールによる処理、

を実行させるためのプログラムである。

【００１６】

また、本発明の他の形態であるフェイルオーバ方法は、

相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムによるフェイルオーバ方法であって、

前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、

前記現用系サーバの前記制御装置は、前記現用系サーバの障害発生時に、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行い、当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバ

に通知し、
前記待機系サーバは、前記現用系サーバの前記制御装置から前記Ｉ／Ｏフェンシングが完了した旨の通知を受けて、前記待機系サーバ上における前記アプリケーションにて前記現用系サーバ上における前記アプリケーションによる処理を引き継ぐ、
という構成をとる。
【００１７】
また、本発明の他の形態であるフェイルオーバ方法は、
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムによるフェイルオーバ方法であって、
前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、
前記現用系サーバの前記制御装置は、前記現用系サーバの障害発生時に、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行い、当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
という構成をとる。
【発明の効果】
【００１８】
本発明は、以上のように構成されることにより、共有データの保護を行いつつ、迅速にフェイルオーバを行うことができる。
【図面の簡単な説明】
【００１９】
【図１】本発明の実施形態１におけるクラスタシステムの構成を示すブロック図である。
【図２】図１に開示したクラスタシステムにおけるクラスタ起動時の動作を示すフローチャートである。
【図３】図１に開示したクラスタシステムにおけるクラスタ起動時の動作を説明する図である。
【図４】図１に開示したクラスタシステムにおけるソフトウェア障害検出時の動作を示すフローチャートである。
【図５】図１に開示したクラスタシステムにおけるソフトウェア障害検出時の動作を説明する図である。
【図６】図１に開示したクラスタシステムにおけるハードウェア障害検出時の動作を示すフローチャートである。
【図７】図１に開示したクラスタシステムにおけるハードウェア障害検出時の動作を説明する図である。
【図８】図１に開示したクラスタシステムにおける待機系サーバによる現用系サーバの障害検出時の動作を示すフローチャートである。
【図９】図１に開示したクラスタシステムにおける待機系サーバによる現用系サーバの障害検出時の動作を説明する図である。
【図１０】本発明の付記１におけるクラスタシステムの構成を示すブロック図である。
【発明を実施するための形態】
【００２０】
＜実施形態１＞
本発明の第１の実施形態を、図１乃至図９を参照して説明する。図１は、クラスタシステムの構成を示す図であり、図２乃至図９は、クラスタシステムの動作を示す図である。
【００２１】

［構成］
本実施形態におけるクラスタシステム１は、共有ディスクを備えたフェイルオーバ型のクラスタシステムである。このクラスタシステム１は、図１に示すように、現用系サーバ１０と待機系サーバ２０とを備えており、これら両サーバ１０，２０にてアクセス可能であり共有される共有ディスク５０を備えて構成されている。
【００２２】
現用系サーバ１０と待機系サーバ２０とは、それぞれほぼ同一の構成を取っており、同一の業務アプリケーションを実行することが可能である。そして、クラスタシステム１は、現用系サーバ１０に障害が生じた場合には、当該現用系サーバ１０の業務アプリケーションによる処理を、待機系サーバ２０の業務アプリケーションが引き継ぐフェイルオーバ機能を有している。以下、現用系サーバ１０と待機系サーバ２０の構成はほぼ同一であるため、主に現用系サーバ１０の構成について詳述する。
【００２３】
現用系サーバ１０は、演算装置と記憶装置とを備えた情報処理装置にて構成されている。そして、現用系サーバ１０は、演算装置にプログラムが組み込まれることで構築されたOS（Operating System）２０を備えており、当該OS２０は、クラスタソフトウェア３０とIPMIドライバ２１とを備えている。また、OS２０は、図示しない業務アプリケーションを備えている。
【００２４】
また、現用系サーバ１０は、当該サーバ１０に装備されたハードウェア（HW）を管理するBMC（Base Management Controller）４０（制御装置）と、共有ディスク５０へのアクセスを行うPCIカード１１（ディスク入出力装置）を備えている。なお、上記BMC４０は、上記OS２０の影響を受けずに作動するよう構成されている。以下、各構成についてさらに詳述する。
【００２５】
上記OS２０が備えるクラスタソフトウェア３０は、サーバ管理モジュール３１、HB（Heart Beat）モジュー３２、障害検出モジュール３３、サーバ制御モジュール３４、通知受信モジュール３５を備えている。
【００２６】
上記サーバ管理モジュール３１は、クラスタシステムを構成する各サーバの状態（起動状態／停止状態）や、業務アプリケーションの状態（起動状態／停止状態、起動しているサーバ）を管理する。また、サーバ管理モジュール３１は、HBモジュール３２や障害検出モジュール３３から障害情報や他サーバの停止情報を受信した場合に、業務アプリケーションの状態を確認し、アプリケーションが停止している、または、正常に動作できないと判断した場合に、業務アプリケーションのフェイルオーバを指示する。
【００２７】
上記HBモジュール３２は、定期的なネットワークパケットの送受信やディスクへの書込などの手段によりクラスタシステムを構成する各サーバが起動しているか停止しているかを判断する。他のサーバが起動状態から停止状態、または、停止状態から起動状態に移行した場合は、サーバ管理モジュール３１にサーバの状態変化を通知する。
【００２８】
上記障害検出モジュール３３は、自サーバの内蔵・共有ディスクへのアクセス可否や業務アプリケーションの死活、ネットワーク通信が正常に行えるかなど、自サーバのコンピュータ資源が正常に動作しているかどうかの監視を行う。コンピュータ資源に異常を検出した場合は、サーバ管理モジュール３１に異常を通知する。
【００２９】
上記サーバ制御モジュール３４は、サーバ管理モジュール３１からの指示に従い自サーバ１０のシャットダウン、自サーバ１０のBMC４０や相手サーバ２０のBMC９０への指示を行い、自サーバ１０や相手サーバ２０の起動停止を制御する。
【００３０】

通知受信モジュール３５は、相手サーバ２０のBMC９０からのI/Oフェンシング完了通知を受信し、サーバ管理モジュール３１に通知を行う。

【００３１】

そして、上述したOS２０及びクラスタソフトウェア３０の構成を、待機系サーバ６０のOS７０及びクラスタソフトウェア８０も備えている。つまり、待機系サーバ６０のOS７０は、IPMIドライバ７１を備えており、クラスタソフトウェア８０は、サーバ管理モジュール８１、HB（Heart Beat）モジュール８２、障害検出モジュール８３、サーバ制御モジュール８４、通知受信モジュール８５を備えている。

【００３２】

次に、現用系サーバ１０のBMC４０の構成について詳述する。BMC４０は、当該BMC４０が備える演算装置にプログラムが組み込まれることで構築された、BMC制御モジュール４１、IPMI受信モジュール４２、フェンシング制御モジュール４３、通知送信モジュール４４、HW障害検出モジュール４５を備えている。

【００３３】

BMC制御モジュール４１（初期化モジュール）は、起動時にクラスタソフトウェア３０から通知されるI/Oフェンシング対象となるPCIカード１１のスロット番号や、フェンシング完了通知先となる待機系サーバ２０のIPアドレスの管理を行う。なお、PCIカード１１のスロット番号は、PCIカード１１にアクセスするためのアクセス情報であり、待機系サーバ２０のIPアドレスは、情報を通知する際に当該待機系サーバ２０の宛先となるアドレス情報である。

【００３４】

また、BMC制御モジュール４１は、クラスタソフトウェア３０からI/Oフェンシング命令を受信した場合や、HW障害検出モジュール４５からHW障害の通知を受信した場合に、I/Oフェンシングの実行をフェンシング制御モジュール４３に指示する。

【００３５】

上記IPMI受信モジュール４２は、自サーバ１０のクラスタソフトウェア３０や他サーバ２０のクラスタソフトウェア８０からのIPMI制御コマンド（I/Oフェンシング命令や設定情報）を受信し、BMC制御モジュール４１に通知する。

【００３６】

上記フェンシング制御モジュール４３（初期化モジュール）は、BMC制御モジュール４１からのI/Oフェンシング命令を受信し、指定されたPCIカード１１が接続されているPCIスロットのリセットを行い、I/Oフェンシングを行う。また、フェンシング制御モジュール４３は、I/Oフェンシング完了後に、通知送信モジュール４４にI/Oフェンシングの完了を通知する。なお、上記I/Oフェンシングとは、クラスタシステムを構成するサーバが停止する場合に、サーバをFile System（共有ディスク）から切り離してI/Oが発生しない状態にすることである。

【００３７】

上記通知送信モジュール４４（通信モジュール）は、I/Oフェンシングの完了を受信すると、登録されたIPアドレスに対して、I/Oフェンシングの完了をネットワーク経由で通知する。

【００３８】

上記HW障害検出モジュール４５は、サーバ１０を構成するHWモジュールを監視し、障害検出時にBMC制御モジュール４１に障害を通知する。

【００３９】

そして、上述したBMC４０の構成を、待機系サーバ６０のBMC９０も備えている。つまり、待機系サーバ６０のBMC９０は、当該BMC９０が備える演算装置にプログラムが組み込まれることで構築された、BMC制御モジュール９１、IPMI受信モジュール９２、フェンシング制御モジュール９３、通知送信モジュール９４、HW障害検出モジュール９５を備えている。

【００４０】

［動作］
次に、上述したクラスタシステムの動作を、図２乃至図９を参照して説明する。以下では、クラスタシステムの動作を、次の４種類のケースに分けて説明する。
（１）クラスタ起動
（２）クラスタソフトウェアが自サーバの障害を検出したケース
（３）BMCが自サーバのHW障害を検出したケース
（４）クラスタソフトウェアが、他サーバの停止を検出したケース
【００４１】
（１）クラスタ起動
まず、クラスタシステムが起動したときの動作を、図２のフローチャート、及び、図３の点線矢印を参照して説明する。
【００４２】
まず、現用系サーバ１０の起動時に、サーバ制御モジュール３４は、予め登録された設定ファイルから、共有ディスク５０と接続しているPCIカード１１のスロット番号と、待機系サーバ２０のIPアドレスと、をロードし（ステップＳ１）、IPMIドライバ２１経由でBMC４０に通知する（ステップＳ２）。通知を受信したBMC４０のIPMI受信モジュール４２は、BMC制御モジュール４１にPCIカード１１のスロット番号と待機系サーバのIPアドレスを通知する。そして、BMC制御モジュール４１は、通知を受けたPCIカード１１のスロット番号と待機系サーバ２０のIPアドレスとを、モジュール内の揮発性メモリに保存する（ステップＳ３）。なお、待機系サーバ２０でも同様の処理が行われる。
【００４３】
上記の初期設定が行われない場合は、HBタイムアウトによる他サーバの停止検出と、OSパニックやディスクのアンマウントにより共有ディスクへのI/Oを停止する一般的なクラスタソフトウェアとして動作する。
【００４４】
（２）クラスタソフトウェアが自サーバの障害を検出したケース
次に、クラスタソフトウェアが自サーバの障害を検出したときの動作を、図４のフローチャート、及び、図５の点線矢印を参照して説明する。本障害ケースでは、現用系サーバ１０上のクラスタソフトウェア３０を構成する障害検出モジュール３３が障害を検出する。本ケースにおける障害は、業務アプリケーションの消滅、NICドライバやディスクドライバなどのドライバレベルで検出される障害など、ソフトウェアレベルで検出可能な障害である。
【００４５】
まず、障害検出モジュール３３が障害を検出すると、サーバ管理モジュール３１に障害の発生を通知する（ステップＳ１１）。サーバ管理モジュール３１は、BMC４０によるI/Oフェンシング機能が有効であることを確認し、速やかに業務アプリケーションのフェイルオーバ処理を開始する。
【００４６】
サーバ管理モジュール３１は、共有ディスク５０へのI/Oの停止を行うために、サーバ制御モジュール３４に共有ディスク５０へのI/Oフェンシング処理を指示する（ステップＳ１２）。サーバ制御モジュール３４は、サーバ管理モジュール３１からのI/Oフェンシング指示を受け取ると直ちに、I/Oフェンシング命令と対象のPCIカード１１のスロット番号を、IPMIドライバ２１経由でBMC４０に通知する。
【００４７】
続いて、IPMI受信モジュール４２は、IPMIドライバ２１経由でI/Oフェンシング命令と対象PCIカードのスロット番号を受信する。IPMI受信モジュール４２は、BMC制御モジュール４１にI/Oフェンシング命令と対象となるPCIカード１１のスロット番号を通知する。BMC制御モジュール４１は、フェンシング制御モジュール４３にI/Oフェンシング命令と対象となるPCIカード１１のスロット番号を通知する。なお、BMC制御モジュール４１は、起動時に登録されたPCIカード１１のスロット番号を、フェンシング制御モジュール４３に通

知してもよい。

【００４８】

そして、フェンシング制御モジュール４３は、I/Oフェンシングの命令を受けると、指定されたPCIカード１１のスロット番号に対して当該PCIカード１１をリセット（初期化）する（ステップＳ１３）。これにより、ソフトウェアレベルでI/Oフェンシングを制御するよりも、確実に共有ディスク５０に対するI/Oを停止することができる。

【００４９】

上述したように、PCIカード１１のリセットが完了すると、フェンシング制御モジュール４３は、通知送信モジュール４４にI/Oフェンシングの完了を通知するように指示を行う。通知送信モジュール４４は、あらかじめ登録されている通知先（待機系サーバ２０のIPアドレス）に対して、I/Oフェンシングの完了通知を送信する（ステップＳ１４）。通知の方法は、ネットワーク経由のSNMP Trapなどを用いる。なお、上述したBMC４０がI/Oフェンシング命令を受信してから完了通知を待機系サーバ２０に送信する一連の処理を、BMCによるI/Oフェンシング処理と呼ぶ。

【００５０】

その後、待機系サーバ２０上のクラスタウェア８０を構成する通知受信モジュール８５は、ネットワーク経由で、現用系サーバ１０のBMC４０から当該現用系サーバ１０のI/Oフェンシングの完了通知を受信し、I/Oフェンシングの完了したサーバの情報をサーバ管理モジュール８１に通知する。サーバ管理モジュール８１は、I/Oフェンシングの完了通知と対象サーバである現用系サーバ１０の情報を受信すると、当該サーバ１０で動作していた業務アプリケーションのフェイルオーバを開始する。これにより、業務アプリケーションのフェイルオーバが完了し、待機系サーバ２０にて業務が再開する（ステップＳ１５）。

【００５１】

ここで、クラスタシステムを構成するサーバが３台以上の場合は、業務アプリケーションのフェイルオーバ先を、予め登録された優先順位や、リソースの空き状況などのルールに従って１台のサーバに決定する。なお、待機系サーバ２０の通知受信モジュール８５がI/Oフェンシングの完了通知を受信してからフェイルオーバを行う一連の処理を、クラスタ再構成処理と呼ぶ。

【００５２】

（３）BMCが自サーバのHW障害を検出したケース

次に、現用系サーバ１０のBMC４０が、自サーバのHW障害を検出したときの動作を、図６のフローチャート、及び、図７の点線矢印を参照して説明する。

【００５３】

本ケースでは、現用系サーバ１０のBMC４０を構成するHW障害検出モジュール４５が障害を検出する（ステップＳ２１）。本ケースにおける障害は、ハードウェア（HW）のセンサ異常や温度上昇、電圧低下などのハードウェアレベルで検出可能な障害である。

【００５４】

まず、HW障害検出モジュール４５は、障害を検出した場合、BMC制御モジュール４１に障害の発生を通知する。BMC制御モジュール４１は、障害内容を確認し、現用系サーバ１０の停止が必要な障害と判断した場合は、フェンシング制御モジュール４３にI/Oフェンシング命令と起動処理で保存したPCIカード１１のスロット番号を、フェンシング制御モジュール４３に通知する（ステップＳ２２）。

【００５５】

フェンシング制御モジュール４３は、指定された番号のPCIカード１１のスロットをリセットする（ステップＳ２２）。これにより、共有ディスク５０に対するI/Oを停止するI/Oフェンシングを行う。

【００５６】

PCIカード１１のスロットのリセットが完了すると、フェンシング制御モジュール４３は通知送信モジュール４４に、I/Oフェンシングの完了を通知するように指示を行う。通

知送信モジュール４４は、あらかじめ登録されている通知先（待機系サーバ２０のIPアドレス）に対して、I/0フェンシングの完了通知を送信する（ステップＳ２３）。

【００５７】

待機系サーバ２０のクラスタウェア８０は、I/0フェンシングの完了通知を受信し、クラスタ再構成処理を行う（ステップＳ２４）。

【００５８】

（４）クラスタソフトウェアが、他サーバの停止を検出したケース

次に、待機系サーバ６０のクラスタソフトウェア８０が、現用系サーバの障害を検出したときの動作を、図８のフローチャート、及び、図９の点線矢印を参照して説明する。

【００５９】

本ケースでは、待機系サーバ２０のクラスタソフトウェア８０を構成するHBモジュール８２が、現用系サーバ１０の障害を検出する。本ケースにおける障害は、現用系サーバ１０のOS２０のストールやパニック、クラスタソフトウェア３０の停止である。

【００６０】

まず、HBモジュール８２は、現用系サーバ１０のHBモジュール３２と定期的にハートビート通信を行なっている。このとき、HBモジュール８２が一定期間ハートビート通信ができない状態になると、当該HBモジュール８２は現用系サーバ１０が停止したと判断し（ステップＳ３１）、サーバ管理モジュール８１にハートビート通信異常を通知する。サーバ管理モジュール８１は、スプリットブレイン解決を行い、現用系サーバ１０が停止していると判断した場合に、サーバ制御モジュール８４に現用系サーバ１０のI/0フェンシング実行を指示する。

【００６１】

本処理は、現用系サー１０のOS２０がストールによりハートビート通信できていない場合、共有ディスク５０に対するI/0が完全に停止していないことになるため、確実なI/0フェンシングを実施するための処理である。サーバ制御モジュール８４は、サーバ管理モジュール８１からのI/0フェンシング命令を受信すると直ちに、ネットワーク経由でI/0フェンシング命令と対象のPCIカード１１のスロット番号を、現用系サーバ１０のBMC４０に通知する（ステップＳ３２）。

【００６２】

現用系サーバ１０のBMC４０のIPMI受信モジュール４２は、待機系サーバ６０のサーバ制御モジュール８４からのI/0フェンシング命令を受信すると、BMC４０によるI/0フェンシング処理を行う。つまり、上述同様に、共有ディスク５０に接続されたPCIカード１１のスロットをリセットし（ステップＳ３３）、あらかじめ登録されている通知先（待機系サーバ２０のIPアドレス）に対して、I/0フェンシングの完了通知を送信する（ステップＳ３４）。

【００６３】

その後、待機系サーバ６０のクラスタウェア８０は、I/0フェンシングの完了通知を受信し、クラスタ再構成処理を行う（ステップＳ３５）。

【００６４】

本発明のクラスタシステムは以上のように作動するため、以下の効果を発揮しうる。

まず、共有ディスクアクセスの排他制御の確実性が向上する。これは、上述したように、BMCおよびHWレベルでPCIカードをリセットするためであり、ソフトウェアレベルでI/0フェンシングを制御するよりも確実にI/0を停止可能である。たとえば、ソフトウェアレベルでは、カードにキューイングされているI/0を障害発生時に停止することは困難である。また、OSがストールまたは停止していると判断される障害サーバに対して、外部のサーバがI/0フェンシングの命令を行うことで、確実にI/0フェンシングを実行できる。

【００６５】

また、高速なフェイルオーバ実現できる。これは、I/0フェンシング完了後に、現用系サーバのBMCから待機系サーバのクラスタウェアにネットワーク通知を行っていることによる。通常のクラスタウェアは、ハートビートの通信異常（タイムアウト）を検知して現

用系サーバが停止したと判断するため、必ずタイムアウト時間が必要となる。本発明では、タイムアウトではなく、BMCからの通知を契機にフェイルオーバを実行するため、フェイルオーバ開始までの時間短縮が可能となる。

【００６６】

ここで、上記では、２台のクラスタ構成であったが、I/Oフェンシング通知対象IPアドレスを複数登録することで、３台以上のクラスタ構成でも同様の方式が実現可能である。また、両サーバアクティブのクラスタ構成でもアクティブ-スタンバイ型と同用の構成で、提案の方式が実現可能である。

【００６７】

＜付記＞

上記実施形態の一部又は全部は、以下の付記のようにも記載されうる。以下、本発明におけるクラスタシステム（図１０参照）、現用系サーバ、プログラム、フェイルオーバ方法の構成の概略を説明する。但し、本発明は、以下の構成に限定されない。

【００６８】

（付記１）

相互に同一のアプリケーション１１１，２１１を実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ１００及び待機系サーバ２００と、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスク３００と、を備え、

前記現用系サーバ１００は、当該現用系サーバに組み込まれたオペレーティングシステム１１０の影響を受けずに作動する制御装置１２０と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置１３０と、を備えており、

前記現用系サーバ１００の前記制御装置１２０は、前記待機系サーバと通信を行う通信モジュール１２１と、前記現用系サーバの障害発生時に前記ディスク入出力装置１３０を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュール１２１を介して前記待機系サーバ２００に通知する初期化モジュール１２２と、を備え、

前記待機系サーバ２００は、前記現用系サーバの前記制御装置から前記Ｉ／Ｏフェンシングが完了した旨の通知を受けて、前記待機系サーバ上における前記アプリケーションにて前記現用系サーバ上における前記アプリケーションによる処理を引き継ぐフェイルオーバ処理機能を有する、

クラスタシステム。

【００６９】

（付記２）

付記１に記載のクラスタシステムであって、

前記現用系サーバは、当該現用系サーバにて作動しているソフトウェアの障害発生を検出して前記制御装置の前記通知モジュールに通知する障害検出モジュールを備え、

前記制御装置の前記初期化モジュールは、前記障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、

クラスタシステム。

【００７０】

（付記３）

付記１又は２に記載のクラスタシステムであって、

前記現用系サーバの前記制御装置は、当該現用系サーバに装備されたハードウェアの障害発生を検出して前記制御装置の前記初期化モジュールに通知するハードウェア障害検出モジュールを備え、

前記制御装置の前記初期化モジュールは、前記ハードウェア障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化す

ることにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
クラスタシステム。
【００７１】
（付記４）
付記１乃至３のいずれかに記載のクラスタシステムであって、
前記待機系サーバは、前記現用系サーバにおける障害発生を検出して、前記現用系サーバの前記初期化モジュールに通知する待機系側障害検出モジュールを備え、
前記現用系サーバの前記初期化モジュールは、前記待機系サーバの前記待機系側障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
クラスタシステム。
【００７２】
（付記５）
付記１乃至４のいずれかに記載のクラスタシステムであって、
前記初期化モジュールは、前記ディスク入出力装置にアクセスするためのアクセス情報と、前記待機系サーバに情報を通知するためのアドレス情報と、を記憶保持しており、前記アクセス情報に基づいて前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記アドレス情報に基づいて前記通信モジュールを介して前記待機系サーバに通知する、
クラスタシステム。
【００７３】
（付記６）
付記５に記載のクラスタシステムであって、
前記現用系サーバは、当該現用系サーバの起動時に、前記アクセス情報及び前記アドレス情報が予め設定された設定ファイルを読み出し、当該設定ファイルに設定されている前記アクセス情報及び前記アドレス情報を前記制御装置の前記初期化モジュールに通知し、
前記初期化モジュールは、通知された前記アクセス情報及び前記アドレス情報を記憶保持する、
クラスタシステム。
【００７４】
（付記７）
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムにおける前記現用系サーバであって、
前記現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えており、
前記制御装置は、前記待機系サーバと通信を行う通信モジュールと、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールと、を備えた、
現用系サーバ。
【００７５】
（付記７－２）
付記７に記載の現用系サーバであって、
前記現用系サーバは、当該現用系サーバにて作動しているソフトウェアの障害発生を検

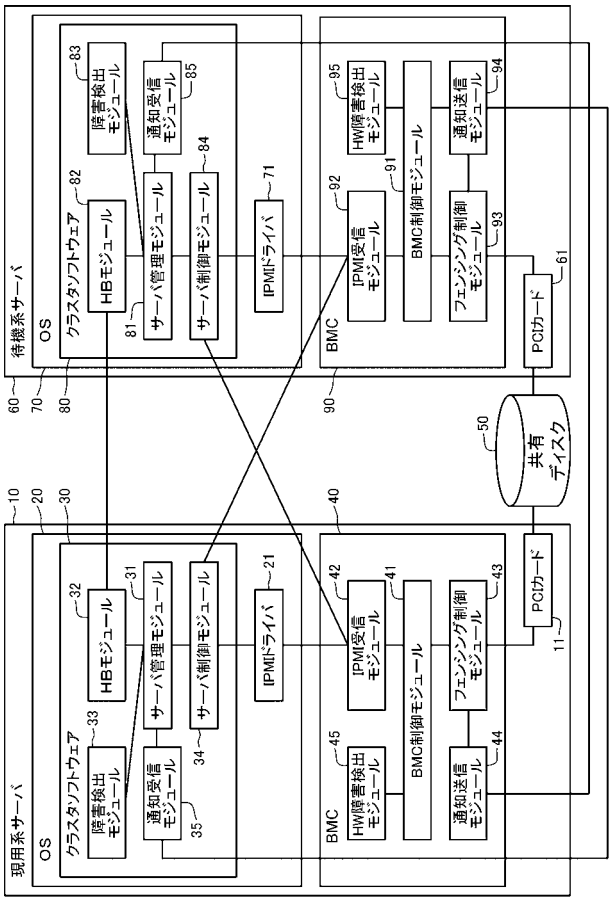
出して前記初期化モジュールに通知する障害検出モジュールを備え、
　前記初期化モジュールは、前記障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
現用系サーバ。
【００７６】
（付記７－３）
　付記７又は７－２に記載の現用系サーバであって、
　前記制御装置は、当該現用系サーバに装備されたハードウェアの障害発生を検出して前記初期化モジュールに通知するハードウェア障害検出モジュールを備え、
　前記初期化モジュールは、前記ハードウェア障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
現用系サーバ。
【００７７】
（付記７－４）
　付記７乃至７－３のいずれかに記載の現用系サーバであって、
　前記待機系サーバが、前記現用系サーバにおける障害発生を検出して、前記現用系サーバの前記初期化モジュールに通知する待機系側障害検出モジュールを備えており、
　前記初期化モジュールは、前記待機系サーバの前記待機系側障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
現用系サーバ。
【００７８】
（付記８）
　相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムにおける前記現用系サーバが、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、
当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、
　前記制御装置に、前記現用系サーバの障害発生時に前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する初期化モジュールによる処理、
を実行させるためのプログラム。
【００７９】
（付記８－２）
　付記８に記載のプログラムであって、
　前記現用系サーバは、当該現用系サーバにて作動しているソフトウェアの障害発生を検出して前記初期化モジュールに通知する障害検出モジュールを備えており、
　前記初期化モジュールは、前記障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
プログラム。

【００８０】
（付記８－３）
付記８又は８－２に記載のプログラムであって、
前記制御装置は、当該現用系サーバに装備されたハードウェアの障害発生を検出して前記初期化モジュールに通知するハードウェア障害検出モジュールを備えており、
前記初期化モジュールは、前記ハードウェア障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
プログラム。
【００８１】
（付記８－４）
付記８乃至８－３のいずれかに記載のプログラムであって、
前記待機系サーバは、前記現用系サーバにおける障害発生を検出して、前記現用系サーバの前記初期化モジュールに通知する待機系側障害検出モジュールを備えており、
前記現用系サーバの前記初期化モジュールは、前記待機系サーバの前記待機系側障害検出モジュールから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
プログラム。
【００８２】
（付記９）
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムによるフェイルオーバ方法であって、
前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、
前記現用系サーバの前記制御装置は、前記現用系サーバの障害発生時に、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行い、当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知し、
前記待機系サーバは、前記現用系サーバの前記制御装置から前記Ｉ／Ｏフェンシングが完了した旨の通知を受けて、前記待機系サーバ上における前記アプリケーションにて前記現用系サーバ上における前記アプリケーションによる処理を引き継ぐ、
フェイルオーバ方法。
【００８３】
（付記９－２）
付記９に記載のフェイルオーバ方法であって、
前記現用系サーバは、当該現用系サーバにて作動しているソフトウェアの障害発生を検出して前記制御装置に通知し、
前記現用系サーバの前記制御装置は、前記現用系サーバにて作動しているソフトウェアの障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
フェイルオーバ方法。
【００８４】

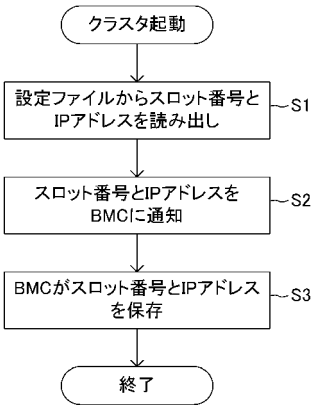
（付記９－３）
付記９又は９－２に記載のフェイルオーバ方法であって、
前記現用系サーバの前記制御装置は、当該現用系サーバに装備されたハードウェアの障害発生を検出すると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
フェイルオーバ方法。
【００８５】
（付記９－４）
付記９乃至９－３のいずれかに記載のフェイルオーバ方法であって、
前記待機系サーバは、前記現用系サーバにおける障害発生を検出して、前記現用系サーバの前記制御装置に通知し、
前記現用系サーバの前記制御装置は、前記待機系サーバから前記現用系サーバにおける障害発生の通知を受けると、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行うと共に当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
フェイルオーバ方法。
【００８６】
（付記１０）
相互に同一のアプリケーションを実行し当該アプリケーションによる他方の処理を引き継ぐことが可能なフェイルオーバ機能を有する現用系サーバ及び待機系サーバと、当該現用系サーバ及び待機系サーバにて共有可能なデータを記憶する共有ディスクと、を備えたクラスタシステムによるフェイルオーバ方法であって、
前記現用系サーバは、当該現用系サーバに組み込まれたオペレーティングシステムの影響を受けずに作動する制御装置と、前記共有ディスクに接続し当該共有ディスクに対するデータの入出力を行うディスク入出力装置と、を備えると共に、当該制御装置は、前記待機系サーバと通信を行う通信モジュールを備えており、
前記現用系サーバの前記制御装置は、前記現用系サーバの障害発生時に、前記ディスク入出力装置を初期化することにより前記共有ディスクに対するＩ／Ｏフェンシングを行い、当該Ｉ／Ｏフェンシングが完了した旨を前記通信モジュールを介して前記待機系サーバに通知する、
フェイルオーバ方法。
【００８７】
なお、上述したプログラムは、記憶装置に記憶されていたり、コンピュータが読み取り可能な記録媒体に記録されている。例えば、記録媒体は、フレキシブルディスク、光ディスク、光磁気ディスク、及び、半導体メモリ等の可搬性を有する媒体である。
【００８８】
以上、上記実施形態等を参照して本願発明を説明したが、本願発明は、上述した実施形態に限定されるものではない。本願発明の構成や詳細には、本願発明の範囲内で当業者が理解しうる様々な変更をすることができる。
【符号の説明】
【００８９】
１０　現用系サーバ
１１　PCIカード
２０　OS
３０　クラスタソフトウェア
３１　サーバ管理モジュール
３２　HBモジュール
３３　障害検出モジュール
３４　サーバ制御モジュール

３５　通知受信モジュール
４０　BMC
４１　BMC制御モジュール
４２　IPMI受信モジュール
４３　フェンシング制御モジュール
４４　通知送信モジュール
４５　HW障害検出モジュール
５０　共有ディスク
６０　待機系サーバ
６１　PCIカード
７０　OS
８０　クラスタソフトウェア
８１　サーバ管理モジュール
８２　HBモジュール
８３　障害検出モジュール
８４　サーバ制御モジュール
８５　通知受信モジュール
９０　BMC
９１　BMC制御モジュール
９２　IPMI受信モジュール
９３　フェンシング制御モジュール
９４　通知送信モジュール
９５　HW障害検出モジュール
１００　現用系サーバ
１１０　OS
１１１　アプリケーション
１２０　制御装置
１２１　通信モジュール
１２２　初期化モジュール
１３０　入出力装置
２００　待機系サーバ
２１０　OS
２１１　アプリケーション
３００　共有ディスク

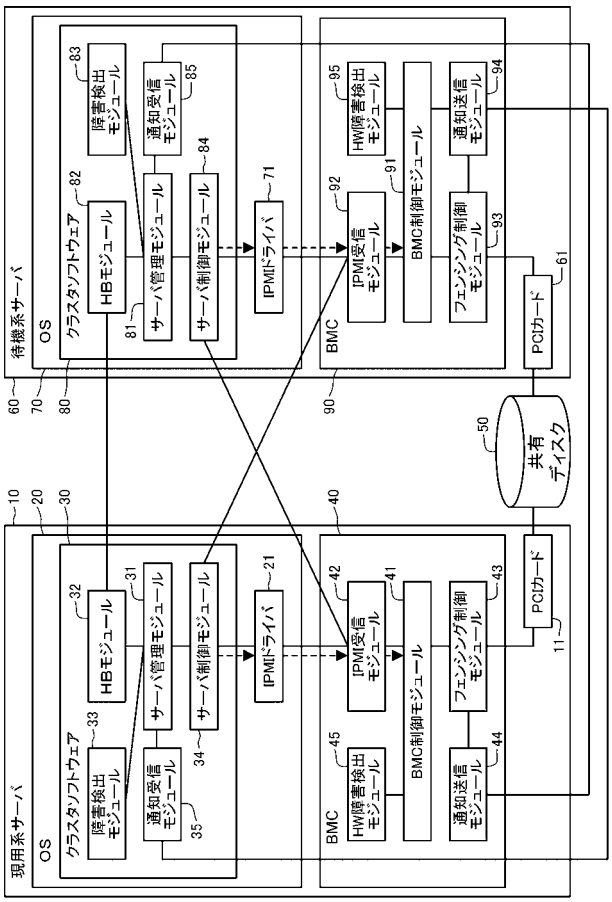
【図１】


待機系サーバ
現用系サーバ
OS
クラスタソフトウェア
HBモジュール
障害検出モジュール
サーバ管理モジュール
通知受信モジュール
サーバ制御モジュール
IPMIドライバ
BMC
IPMI受信モジュール
BMC制御モジュール
HW障害検出モジュール
フェンシング制御モジュール
通知送信モジュール
PCIカード
共有ディスク

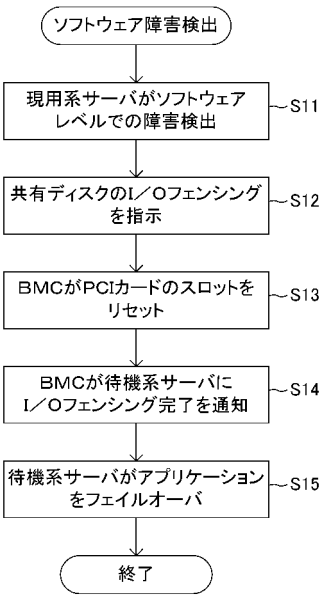
【図２】


クラスタ起動
設定ファイルからスロット番号とIPアドレスを読み出し
S1
スロット番号とIPアドレスをBMCに通知
S2
BMCがスロット番号とIPアドレスを保存
S3
終了

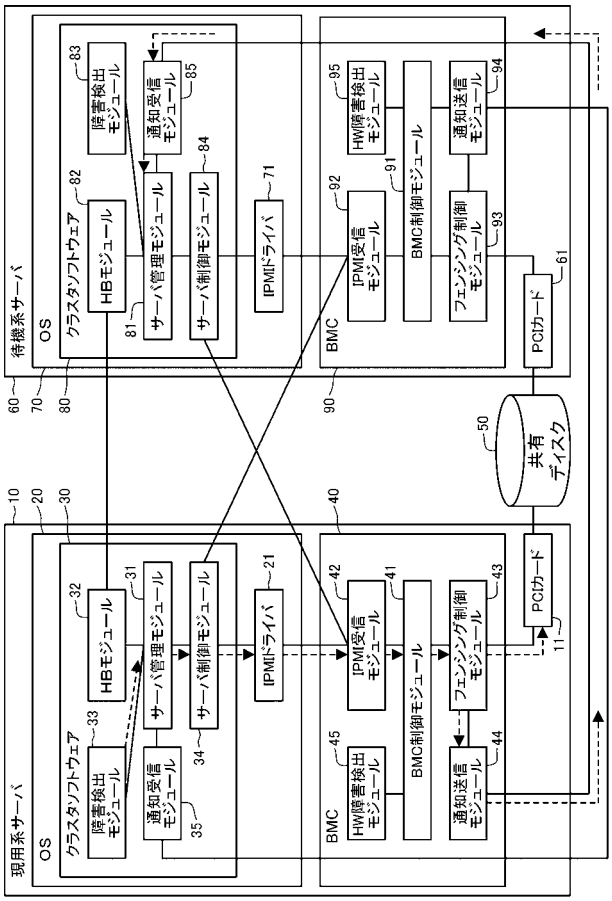
【図３】


待機系サーバ
現用系サーバ
OS
クラスタソフトウェア
HBモジュール
障害検出モジュール
サーバ管理モジュール
通知受信モジュール
サーバ制御モジュール
IPMIドライバ
BMC
IPMI受信モジュール
BMC制御モジュール
HW障害検出モジュール
フェンシング制御モジュール
通知送信モジュール
PCIカード
共有ディスク

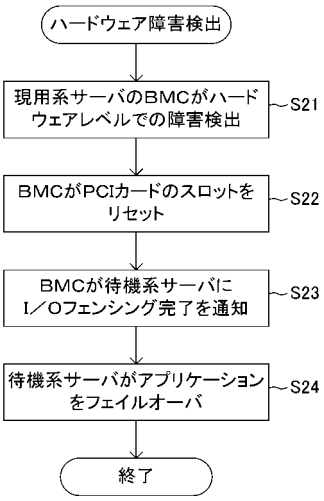
【図４】


ソフトウェア障害検出
現用系サーバがソフトウェアレベルでの障害検出
S11
共有ディスクのI／Oフェンシングを指示
S12
BMCがPCIカードのスロットをリセット
S13
BMCが待機系サーバにI／Oフェンシング完了を通知
S14
待機系サーバがアプリケーションをフェイルオーバ
S15
終了

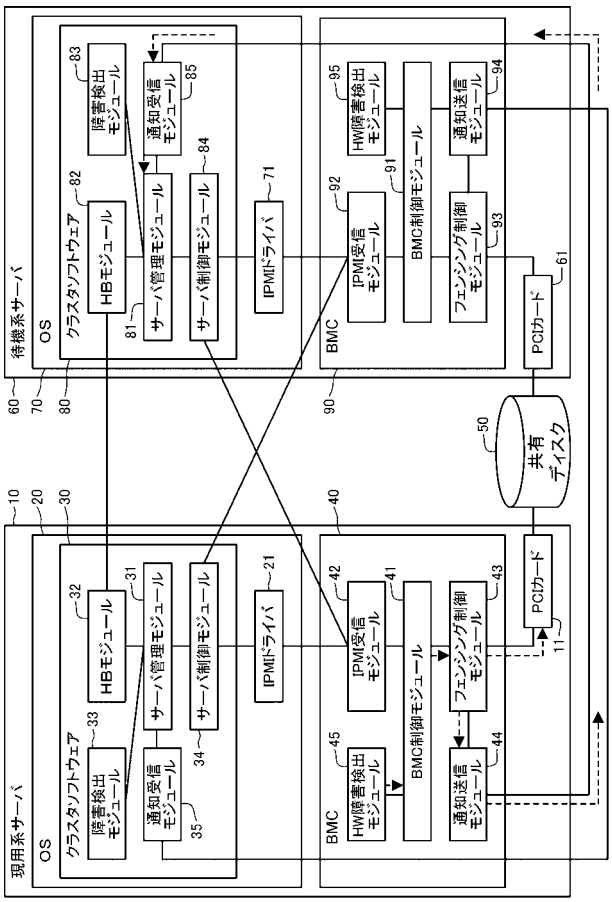
【図５】


待機系サーバ
現用系サーバ
OS
クラスタソフトウェア
HBモジュール
障害検出モジュール
サーバ管理モジュール
通知受信モジュール
サーバ制御モジュール
IPMIドライバ
BMC
IPMI受信モジュール
BMC制御モジュール
フェンシング制御モジュール
通知送信モジュール
HW障害検出モジュール
PCIカード
共有ディスク

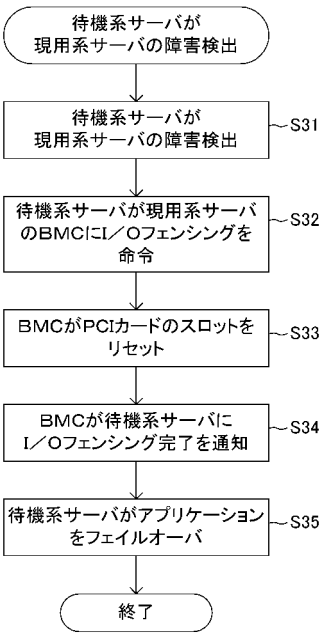
【図６】


ハードウェア障害検出
現用系サーバのBMCがハードウェアレベルでの障害検出
S21
BMCがPCIカードのスロットをリセット
S22
BMCが待機系サーバにI／Oフェンシング完了を通知
S23
待機系サーバがアプリケーションをフェイルオーバ
S24
終了

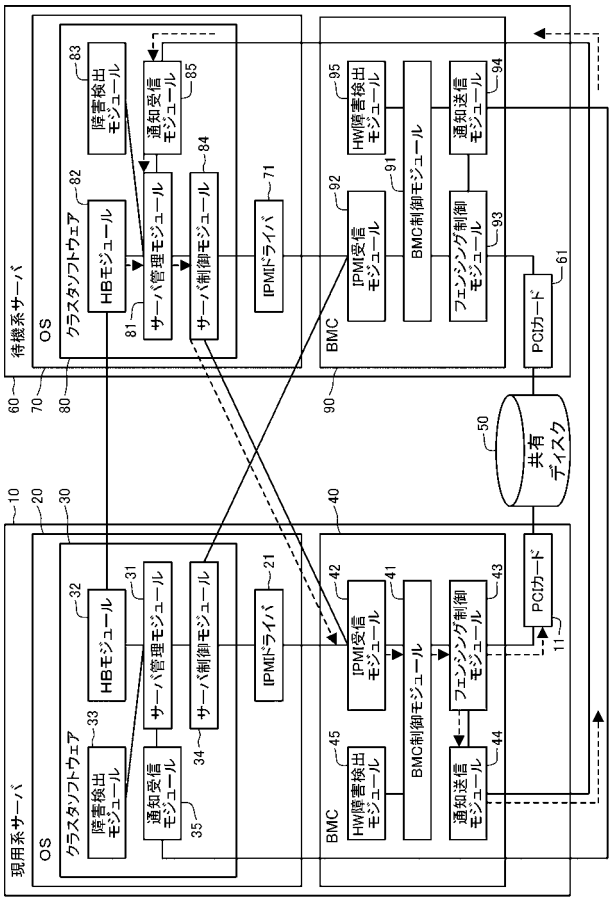
【図７】


待機系サーバ
現用系サーバ
OS
クラスタソフトウェア
HBモジュール
障害検出モジュール
サーバ管理モジュール
通知受信モジュール
サーバ制御モジュール
IPMIドライバ
BMC
IPMI受信モジュール
BMC制御モジュール
フェンシング制御モジュール
通知送信モジュール
HW障害検出モジュール
PCIカード
共有ディスク

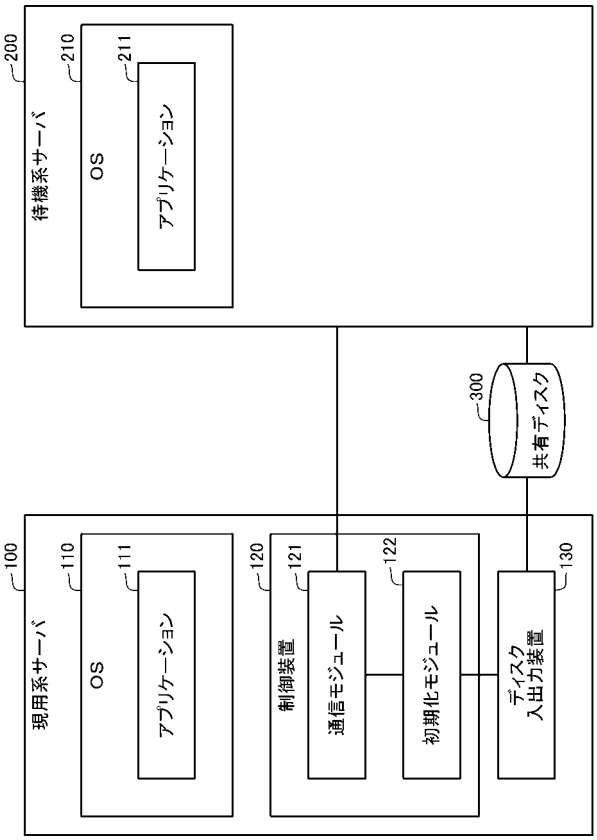
【図８】


待機系サーバが現用系サーバの障害検出
待機系サーバが現用系サーバの障害検出
S31
待機系サーバが現用系サーバのBMCにI／Oフェンシングを命令
S32
BMCがPCIカードのスロットをリセット
S33
BMCが待機系サーバにI／Oフェンシング完了を通知
S34
待機系サーバがアプリケーションをフェイルオーバ
S35
終了

【図９】


待機系サーバ
現用系サーバ
OS
クラスタソフトウェア
HBモジュール
障害検出モジュール
サーバ管理モジュール
通知受信モジュール
サーバ制御モジュール
IPMIドライバ
BMC
IPMI受信モジュール
BMC制御モジュール
フェンシング制御モジュール
通知送信モジュール
HW障害検出モジュール
PCIカード
共有ディスク

【図１０】


待機系サーバ
現用系サーバ
OS
アプリケーション
制御装置
通信モジュール
初期化モジュール
ディスク入出力装置
共有ディスク