(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6134720号
(P6134720)
(45)発行日　平成29年5月24日(2017.5.24)　(24)登録日　平成29年4月28日(2017.4.28)

(51)Int.Cl.　F I
G06F 13/00 (2006.01)　G06F 13/00 301J

請求項の数 15 (全 15 頁)

(21)出願番号　特願2014-533980(P2014-533980)
(86)(22)出願日　平成24年9月28日(2012.9.28)
(65)公表番号　特表2014-532236(P2014-532236A)
(43)公表日　平成26年12月4日(2014.12.4)
(86)国際出願番号　PCT/GB2012/052413
(87)国際公開番号　WO2013/050745
(87)国際公開日　平成25年4月11日(2013.4.11)
審査請求日　平成27年7月14日(2015.7.14)
(31)優先権主張番号　1117230.1
(32)優先日　平成23年10月5日(2011.10.5)
(33)優先権主張国　英国(GB)

(73)特許権者　512313285
マイクロン　テクノロジー，インコーポレイティド
アメリカ合衆国，アイダホ　８３７１６－９６３２，ボイジー，サウス　フェデラル　ウェイ　８０００
(74)代理人　100083806
弁理士　三好　秀和
(74)代理人　100095500
弁理士　伊藤　正和
(74)代理人　100111235
弁理士　原　裕子

最終頁に続く

(54)【発明の名称】接続方法

(57)【特許請求の範囲】
【請求項１】
第１のデバイスと第２のデバイス間の通信の障害を管理する方法であって、
上記第１のデバイスと上記第２のデバイスの中間の第３のデバイスにおいて、上記第１のデバイスから上記第２のデバイスへ供給される第１のデータを傍受し、上記第１のデータは上記第１のデータが属するトランザクションを示すデータフィールドを含み、
上記第３のデバイスにおいて、上記第１のデータに対する応答が上記第１のデバイスによって受信されないという上記第１のデバイスと第２のデバイス間の通信の障害を検出し、
上記第１のデータに対する応答が上記第２のデバイスから上記第１のデバイスによって受信されないことに起因する上記第１のデバイスのサービスの中断を防止するために、上記第２のデバイスが該第１のデバイスによって使えないことを示す第２のデータが、上記第３のデバイスにおいて、生成され、上記第３のデバイスから上記第１のデバイスに送信されることによって上記第２のデバイスからの連続した通信をエミュレートし、上記第２のデータは上記第２のデータの送信元を示す一部を含み、上記第２のデータの上記一部は上記第２のデータが上記第２のデバイスに由来するように見えるように上記第２のデバイスを示す方法。
【請求項２】
上記エミュレーションは、上記第３のデバイスにおいて、上記第１のデータの上記データフィールドによって示された上記トランザクションに少なくとも基づいて複数の格納さ

れたエラーメッセージから一つのエラーメッセージを選択し、上記選択されたエラーメッセージは上記第２のデータに含まれる請求項１に記載の方法。

【請求項３】

上記第２のデータは、該第２のデータが上記第１のデータに対する応答であることを示すデータを含む請求項１記載の方法。

【請求項４】

上記第１及び第２のデバイスの少なくとも１つは、サーバである請求項１乃至３のいずれか１項記載の方法。

【請求項５】

上記第１及び第２のデバイスの少なくとも１つは、Ｉ／Ｏデバイスである請求項１乃至４のいずれか１項記載の方法。

【請求項６】

上記第１のデバイスは、PCI Express接続によって上記第２のデバイスに接続される請求項１乃至５のいずれか１項記載の方法。

【請求項７】

上記第２のデータは、上記第２のデバイスが連絡できないことを示す請求項１乃至６のいずれか１項記載の方法。

【請求項８】

上記第２のデータは、上記第２のデバイスが故障したことを示す請求項１乃至７のいずれか１項記載の方法。

【請求項９】

上記第２のデータは、該第２のデータが破損していることを示す請求項１乃至８のいずれか１項記載の方法。

【請求項１０】

上記第１のデバイスと第２のデバイス間の通信の障害の検出では、上記第１のデバイスと上記第２のデバイスを接続しているケーブルが切断されたことを検出することを含む請求項１乃至９のいずれか１項記載の方法。

【請求項１１】

上記第１のデバイスと第２のデバイス間の通信の障害の検出では、

上記第３のデバイスから上記第２のデバイスに、第３のデータを送信し、

上記第２のデバイスから、上記第３のデータに対する応答が所定の時間内に受信されないときに、当該第１のデバイスと第２のデバイス間の通信の障害を決定することを含む請求項１乃至１０のいずれか１項記載の方法。

【請求項１２】

コンピュータが読取り可能なプログラムを有するコンピュータが使えるストレージデバイスを含むコンピュータプログラム製品であって、上記コンピュータが読取り可能プログラムは、第１のデバイスと第２のデバイスの中間で実行されると、上記コンピュータに、

上記第１のデバイスから上記第２のデバイスへの第１のデータを傍受させ、上記第１のデータは上記第１のデータが属するトランザクションを示すデータフィールドを含み、

上記第１のデータに対する応答が上記第１のデバイスによって受信されないという上記第１のデバイスと上記第２のデバイス間の通信の障害を検出させ、

上記第１のデータに対する応答が上記第２のデバイスから上記第１のデバイスによって受信されないことに起因する上記第１のデバイスのサービスの中断を防止するために、第２のデータを生成して上記第１のデバイスに送信することによって上記第２のデバイスからの連続した通信をエミュレートさせ、上記第２のデータは上記第２のデバイスが上記第１のデバイスによって使えないことを示し、上記第２のデータは上記第２のデータの送信元を示す一部を含み、上記第２のデータの上記一部は上記第２のデータが上記第２のデバイスに由来するように見えるように上記第２のデバイスを示すようにさせるコンピュータプログラム製品。

【請求項１３】

コンピュータが読取り可能なプログラムを格納したメモリデバイスを含むコンピュータプログラム製品であって、上記コンピュータが読取り可能プログラムは、第１のデバイスと第２のデバイスの中間で実行されると、上記コンピュータに、
上記第１のデバイスから上記第２のデバイスへの第１のデータを傍受させ、上記第１のデータは上記第１のデータが属するトランザクションを示すデータフィールドを含み、
上記第１のデータに対する応答が上記第１のデバイスによって受信されないという上記第１のデバイスと上記第２のデバイス間の通信の障害を検出させ、
上記第１のデータに対する応答が上記第２のデバイスから上記第１のデバイスによって受信されないことに起因する上記第１のデバイスのサービスの中断を防止するために、第２のデータを生成して上記第１のデバイスに送信することによって上記第２のデバイスからの連続した通信をエミュレートさせ、上記第２のデータは上記第２のデバイスが上記第１のデバイスによって使えないことを示し、上記第２のデータは上記第２のデータの送信元を示す一部を含み、上記第２のデータの上記一部は上記第２のデータが上記第２のデバイスに由来するように見えるように上記第２のデバイスを示すようにさせるコンピュータプログラム製品。
【請求項１４】
第１のデバイスと第２のデバイス間の通信の障害を管理するコンピュータ装置であって、
プロセッサで読取り可能な命令を記憶するメモリと、
上記メモリに記憶された命令を読み出して、実行するプロセッサとを含み、
上記プロセッサで読取り可能な命令は、上記コンピュータを制御するように構成され、
上記第１のデバイスから上記第２のデバイスへの第１のデータを傍受させ、上記第１のデータは上記第１のデータが属するトランザクションを示すデータフィールドを含み、
上記第１のデータに対する応答が上記第１のデバイスによって受信されないという上記第１のデバイスと上記第２のデバイス間の通信の障害を検出させ、
上記第１のデータに対する応答が上記第２のデバイスから上記第１のデバイスによって受信されないことに起因する上記第１のデバイスのサービスの中断を防止するために、第２のデータを生成して上記第１のデバイスに送信することによって上記第２のデバイスからの連続した通信をエミュレートさせ、上記第２のデータは上記第２のデバイスが上記第１のデバイスによって使えないことを示し、上記第２のデータは上記第２のデータの送信元を示す一部を含み、上記第２のデータの上記一部は上記第２のデータが上記第２のデバイスに由来するように見えるように上記第２のデバイスを示すようにさせるコンピュータ装置。
【請求項１５】
自己修復ケーブルアダプタであって、接続された第１のデバイスから接続された第２のデバイスへの第１のデータを傍受し、上記第１のデータに対する応答が上記接続された第１のデバイスによって受信されないという上記接続された第１のデバイスと上記接続された第２のデバイス間の通信の障害を検出し、上記第１のデータに対する応答が上記接続された第２のデバイスから上記接続された第１のデバイスによって受信されないことに起因する上記接続された第１のデバイスのサービスの中断を防止するように、第２のデータを生成して上記接続された第１のデバイスに送信することによって上記接続された第２のデバイスからの連続した通信をエミュレートし、上記第２のデータは上記第２のデータの送信元を示す一部を含み、上記第２のデータの上記一部は上記第２のデータが上記第２のデバイスに由来するように見えるように上記接続された第２のデバイスを示し、上記第２のデータは上記接続された第２のデバイスが上記第１のデバイスによって使えないことを示し、上記第１のデータは上記第１のデータが属すトランザクションを示すデータフィールドを含む自己修復ケーブルアダプタを含む装置。
【発明の詳細な説明】
【技術分野】
【０００１】

本発明は、第１のデバイスと第２のデバイス間の通信の障害を管理する方法に関する。
【背景技術】
【０００２】
コンピュータシステムでは、多くの場合、デバイス間でデータを送信することが必要であり、例えば、多くの場合、処理デバイスを複数の入出力デバイスに接続することが必要である。適切なデータ通信は、デバイスがリンクを介して互いにデータパケットを送信できるように、デバイスを接続することによって達成され、そして、リンクは有線リンク又は無線リンクであってもよい。
【０００３】
被接続デバイスのシステムにとって、それらの被接続デバイスの１つの予期しない故障又は不稼働率を管理できることが、ますます重要になっている。例えば、システムは、ケーブルが間違って切断（それによって、例えば、被接続デバイスがサーバから切断）されたときに、あるいは遠隔デバイスが故障した場合に、自己修復する（それらの指定されたサービスを提供し続ける）必要がある。一般的には、オペレーティングシステムは、動作中の（「ホット」）システムにおける入出力（Ｉ／Ｏ）アダプタの管理（すなわち計画）された挿抜（すなわちホットプラグ／ホットスワッピング）を幾分サポートしているが、ホットシステムにおけるデバイスの予期しない取外しは、殆ど又は全くサポートしていない。
【０００４】
例えば、サーバ（すなわち「インボックス」）内にＩ／Ｏアダプタを永久接続するために、PCI Expressだけが既に用いられている。Ｉ／Ｏデバイスをサーバ内に永久的にインストールするこのような構成の場合、Ｉ／Ｏデバイスの故障は、サーバ全体の故障として現れ、そのサーバ自体は、引き続き動作する必要があるとは考えられない。遠隔被接続PCI ExpressＩ／Ｏデバイス（すなわちサーバの外で接続されるＩ／Ｏデバイス）に対する関心が高まり、被接続デバイスの予期しない活線取外しをサーバが段階的にサポートする要求が、システムの障害許容力に対して重要になることを意味している。
【０００５】
データパケットの送受信は、多くの場合、トランザクションの観点から説明される。トランザクションは、デバイス間で送信される１つ以上のデータパケットを含む。PCI Expressは、スプリットトランザクションモデルを実行し、そこでは、送信元デバイスは、要求データパケットを宛先デバイスに送信し、そして、応答における宛先デバイスからの完了データパケットを待つ。一般的に、オペレーティングシステムは、失敗したPCI Expressトランザクションを段階的に取り扱うように構成されていない。例えば、サーバが要求データパケットを被接続デバイスに送信し、予想に反して、その要求に対する応答において完了データパケットを受信しない場合、サーバのオペレーティングシステムは、クラッシュする虞がある。PCI Expressに基づく現在の被接続システム自体は、PCI Expressに接続された送信元デバイスが予想に反して利用できなくなったとき、クラッシュする虞がある。
【０００６】
PCI Expressの標準的な実装は、PCI Expressサブシステム、すなわち被接続PCI Expressデバイスの故障を取り扱う十分な手段を提供していない。これは、結果として、満足できるレベルの障害許容力を有する遠隔又は共用のＩ／Ｏシステムの開発を難しくしている。
【発明の概要】
【発明が解決しようとする課題】
【０００７】
本発明の実施の形態の目的は、上で概説した問題の１つ以上を取り除く又は軽減することである。
【課題を解決するための手段】
【０００８】

本発明の第１の特徴に基づいて、第１のデバイスと第２のデバイス間の通信の障害を管理する方法を提供し、この方法は、第１のデバイスと第２のデバイスの中間の第３のデバイスにおいて、第１のデバイスと第２のデバイスの通信の障害を検出し、第３のデバイスから第１のデバイスに、第２のデバイスが第１のデバイスによって使えないことを示す第１のデータを伝送する。

【０００９】

このように、第１のデバイスは、第１のデバイスが故障する原因となる、第２のデバイスによる予期しない通信の損失にさらされることはない。第１のデバイスと第２のデバイス間の通信の損失を検出することによって、第３のデバイスは、適切なデータを第１のデバイスに送信して、第１のデバイスの故障を防止することができる。例えば、第２のデバイスが状態を変えたことを第１のデバイスに通知することによって、第１のデータは、通信の障害を第１のデバイス上で動作しているソフトウェア（ＯＳ、ドライバ、アプリケーション）から隠すことができる。例えば、第３のデバイスは、第２のデバイスの連続した存在を、第１のデバイスが「利用可能であるが、使えない」として解釈する状態にエミュレートすることができる。第１のデバイスは、（例えば第２のデバイスに代わるもので通信することによって）、サービスが中断することなく、「利用可能であるが、使えない」状態を段階的に取り扱うように構成することができる。

【００１０】

第３のデバイスは、第１のデバイスによって定義される故障ユニット内にあってもよく、一方、第２のデバイスは、第１のデバイスによって定義される故障ユニットの外にあってもよい。

【００１１】

第２のデバイスが使えないことを示す際に、第１のデータは、一時的かもしれない外部の問題のために、第２のデバイスが局所的に機能するが、その指定サービスを実行できないことを示すことができる。

【００１２】

方法は、更に、第３のデバイスにおいて、第１のデバイスから第２のデバイスへの第２のデータを傍受してもよい。

【００１３】

第１のデバイスと第２のデバイス間の通信が第３のデバイスを通るように、第３のデバイスは、第１のデバイスと第２のデバイスの間にあってもよい。例えば、第３のデバイスは、ケーブルアダプタの形をとってもよい。あるいは、第１のデバイスから第２のデバイスに送信されるデータは、第１のデバイスと第２のデバイス間の更なるデバイスによって、第３のデバイスに転送されてもよい。

【００１４】

第１のデバイスと第２のデバイス間の通信の障害の検出では、第２のデータに対する応答が第１のデバイスによって受信されないことを検出することを含んでもよい。第１のデバイスと第２のデバイス間の通信の障害の検出では、第１のデバイスと第２のデバイスを接続しているケーブルが切断されたことを検出することを含んでもよい。例えば、ケーブルの切断の検出では、以前に存在していた「ケーブル検出」信号がないことを検出することを含んでもよい。あるいは、第１のデバイスと第２のデバイス間の接続が無線接続の場合、通信の障害の検出では、無線接続の中断又は干渉を検出することを含んでもよい。

【００１５】

第１のデータの送信元を示すことを目的とした第１のデータの一部は、第２のデバイスを示すことができる。第１のデータは、第１のデータが第２のデータに対する応答であることを示すデータを含むことができる。例えば、第２のデータは、第２のデータが属するトランザクションを示すデータフィールドを含むことができ、第１のデータは、それが同じトランザクションに属することを示すデータを含むことができる。

【００１６】

第１及び第２のデバイスの１つは、サーバであってもよく、及び／又は第１及び第２の

デバイスの１つは、第１又は第２のデバイスの遠隔被接続リソースであってもよい。例えば、第１及び第２のデバイスの１つは、Ｉ／Ｏデバイス又は他の遠隔リソースであってもよい。

【００１７】

第１のデバイスは、PCI Express接続によって第２のデバイスに接続されていてもよい。

【００１８】

第１のデータは、第２のデバイスが連絡できないことを示してもよく、第１のデータは、第２のデバイスが故障したことを示してもよく、及び／又は第１のデータは、第１のデータが破損していることを示してもよい。例えば、第１のデータは、第２のデバイスが故障し、及び／又は第１のデータが破損したことを示す値が「１」を有する状態ビットを含むことができる。

【００１９】

第１のデバイスと第２のデバイス間の通信の障害の検出では、第３のデバイスから第２のデバイスに、第３のデータを伝送し、第２のデバイスから、第３のデータに対する応答が所定の時間内に受信されないときに、第１のデバイスと第２のデバイス間の通信のその障害を決定することを含んでもよい。

【００２０】

本発明の第２の特徴に基づいて、第１のデバイスと第２のデバイス間の通信の障害を管理する装置を提供し、この装置は、第１のデバイスと第２のデバイスの中間の第３のデバイスにおいて、第１のデバイスと第２のデバイス間の通信の障害を検出する手段と、第３のデバイスから第１のデバイスに、第２のデバイスが第１のデバイスによって使えないことを示す第１のデータを伝送する手段とを含む。

【００２１】

本発明の第３の特徴に基づいて、第１のデバイスと第２のデバイス間を接続する自己修復ケーブルアダプタを提供し、自己修復ケーブルアダプタは、被接続第１のデバイスと被接続第２のデバイス間の通信の障害を検出する検出器と、被接続第１のデバイスに、被接続第２のデバイスが被接続第１のデバイスによって使えないことを示す第１のデータを伝送する送信機とを含む。

【００２２】

発明の１つの特徴に関連して説明した多くの特徴が発明の他の特徴に関連して適用できることは言うまでもない。

【００２３】

本発明の特徴が適切なハードウェア及び／又はソフトウェアを含むあらゆる便利な方法で実現できることは言うまでもない。例えば、発明を実施するように配置されたスイッチングデバイスは、適切なハードウェアコンポーネントを用いて形成することができる。あるいは、プログラマブルデバイスは、発明の実施の形態を実現するようにプログラムすることができる。したがって、また、発明は、発明の特徴を実現する適切なコンピュータプログラムを提供する。このようなコンピュータプログラムは、有形のキャリア媒体（例えばハードディスク、ＣＤ－ＲＯＭ等）と、無形のキャリア媒体（例えば通信信号）とを含む適切なキャリア媒体で伝達することができる。

【図面の簡単な説明】

【００２４】

つぎに、本発明の実施の形態を、図面を参照し、一例として説明する。

【００２５】

【図１】Ｉ／Ｏデバイスがサーバのコンポーネントとして設けられた従来のＩ／Ｏ構成のブロック図である

【００２６】

【図１ａ】Ｉ／Ｏデバイスがサーバのコンポーネントとして設けられた従来のＩ／Ｏ構成のブロック図である

【００２７】
【図２】Ｉ／Ｏデバイスがサーバに遠隔で接続された従来のＩ／Ｏ構成のブロック図である。
【００２８】
【図３】複数のＩ／Ｏデバイスがサーバに遠隔で接続された従来のＩ／Ｏ構成のブロック図である。
【００２９】
【図４】Ｉ／Ｏデバイスが本発明の実施の形態に基づいてサーバに接続されたＩ／Ｏ構成のブロック図である。
【００３０】
【図５】データパケットヘッダの略図である。
【００３１】
【図６】サーバが本発明の実施の形態に基づいてＩ／Ｏデバイスに接続されたＩ／Ｏ構成のブロック図である。
【発明を実施するための形態】
【００３２】
図１に示すように、サーバ１は、ＣＰＵ／PCI Expressルートコンプレックス（ＣＰＵ／ＲＣ）２と、ネットワークインタフェースコントローラ（ＮＩＣ）３とを含む。ＣＰＵ／ＲＣ２は、PCI Expressチップ間接続４によって、ＮＩＣ３に接続されている。ＮＩＣ３は、イーサネット（登録商標）接続６（例えば、ケーブル接続又は無線接続であってもよい）によって、ネットワーク５に接続されている。ＮＩＣ３及び適切なソフトウェアを使用することによって、サーバ１は、サーバ１のユーザに、ネットワーク５に対するアクセスを提供する。サーバ１の本発明に関係しない他の詳細は、説明を明確にするために省略されていることは言うまでもなく、それらの詳細は、当業者にとって容易に明らかである。
【００３３】
一般論として、図１のサーバ１は、単一の故障ユニットであると考えられる。サーバ１の内部コンポーネント（例えばＮＩＣ３又は接続４）の故障は、サーバ１全体の故障であると考えられ、サーバ１が最早その意図された機能（すなわちネットワーク５に対するアクセスを容易にすること）を提供できないと考えられる。ＮＩＣ３自体が故障し、又はＮＩＣ３とＣＰＵ／ＲＣ２間の通信に障害が生じると、サーバ１が機能し続ける必要性はなくなり、そして、サーバ１は、機能しなくなる。
【００３４】
内部コンポーネントの故障とは対照的に、外部コンポーネント（すなわちサーバ１の外部の単一の故障ユニット）の故障、例えばイーサネット接続６又は下流のスイッチ（図示せず）は、サーバ１の故障であると考えられず、サーバ１自体は、動作し続ける必要がある。一般論として、交換するデバイスは、簡単に接続する（あるいは、例えば間違って切断した場合に再接続する）ことができることを考えると、サーバ１に外的に接続されたコンポーネントの故障は、サーバ１の故障とすべきではない。したがって、サーバ１は、外部リソースが現在利用できないことを正確に通知する状況を段階的に取り扱うように構成されている。例えば、ＣＰＵ／ＲＣ２とＮＩＣ３間のその通信が維持されているならば、イーサネット接続６の故障により、ＮＩＣ３は、（適切なエラーメッセージを伝送することによって）イーサネット接続６が利用できないことをＣＰＵ／ＲＣ２に通知する。ＣＰＵ／ＲＣ２は、このメッセージを受信すると、サーバ１全体の故障を起こすことなく、イーサネット接続６上の伝送を用いるＮＩＣ３へのデータパケットの送信を止めることができる。
【００３５】
ＣＰＵ／ＲＣは、ＮＩＣからのエラーメッセージを受信すると、利用可能なバックアップリソースを利用することによって、その指定サービスを提供し続けることができる。図１ａに示すように、サーバ１ａは、PCI Expressチップ間接続４ａによってＮＩＣ３ａに

、PCI Expressチップ間接続４ｂによって ＮＩＣ３ｂに接続されたＣＰＵ／ＲＣ２ａを備える。ＮＩＣ３ａ、３ｂのそれぞれは、それぞれのイーサネット接続６ａ、６ｂによってネットワーク５に接続されている。ＮＩＣ３ａ、３ｂのそれぞれ自体は、ネットワーク５に対するアクセスを容易にすることができる。イーサネット接続６ｂが故障すると、例えば、対応するＮＩＣ３ｂは、イーサネット接続６ｂが利用できないことを示すエラーメッセージをＣＰＵ／ＲＣ２ａに送信する。ＣＰＵ／ＲＣ２ａがこのようなエラーメッセージを段階的に取り扱うように構成されているとき、ＣＰＵ／ＲＣ２ａは、ＮＩＣ３ａ及びイーサネット接続６ａを用いて、その指定機能を提供し続けることができる。

【００３６】

図２の構成において、ＮＩＣ７は、サーバ８に対して外的（すなわち、サーバ８によって定義される単一の故障ユニットの外）に収容されている。図２に示すように、サーバ８は、PCI Express接続１１によってＣＰＵ／ＲＣ１０に接続されたPCI Expressケーブルアダプタ９を含む。PCI Expressケーブルアダプタ９は、ＣＰＵ／ＲＣ１０をケーブル１４によって、遠隔Ｉ／Ｏ機器１３内に収容されたPCI Expressケーブルアダプタ１２に接続する。また、遠隔Ｉ／Ｏ機器１３は、ＮＩＣ７を収容し、ＮＩＣ７は、PCI Express接続１５によってPCI Expressケーブルアダプタ１２に接続されている。PCI Expressケーブルアダプタ９、１２は、ケーブル１４でPCI Express信号を運ぶのに必要なケーブル駆動及び信号調整機能を備えるだけである。すなわち、ケーブルアダプタ９、１２は、サーバ８上で動作するソフトウェアのような論理機能を全く備えないが、図２のシステムは、論理的には図１のシステムと同じである。

【００３７】

ケーブル１４が故障又はＮＩＣ７自体が故障すると、ＣＰＵ／ＲＣ１０とＮＩＣ７間の通信は切断される。ＣＰＵ／ＲＣ１０自体は、ＮＩＣ７によって処理されていないトランザクションに対する応答を受信しない。上述したように、未処理のトランザクションに対する応答の欠如は、結果的に、サーバ８全体の故障につながる。図１の構成におけるＮＩＣ３（又はＮＩＣ３とＣＰＵ／ＲＣ２間の接続４）の故障がサーバ１全体の故障となることは、通常、許容できるが、図２の構成におけるケーブル１４及びＮＩＣ７の遠隔の本質は、これらのコンポーネントを簡単に交換又は再接続できることを意味する。図２の構成自体において、ＮＩＣ７又はケーブル１４のどちらかの故障をサーバ８の故障の原因としなければならないことは、望ましくない。

【００３８】

さらに、図３の構成に示すように、サーバから離してＩ／Ｏリソースを配置することにより、複数のＩ／Ｏリソースを容易に配置でき、１つの被接続Ｉ／Ｏリソースが故障した場合に、次のリソースがバックアップとして機能することができる。図３に示すように、サーバ２０のＣＰＵ／ＲＣ１９は、ケーブルアダプタ２１によって、ケーブル２３により遠隔Ｉ／Ｏ機器２２に、ケーブル２５により遠隔Ｉ／Ｏ機器２４に接続されている。遠隔Ｉ／Ｏ機器２２は、ケーブル２３に接続されたケーブルアダプタ２６と、接続２８によってケーブルアダプタ２６に接続されたＮＩＣ２７とを含む。遠隔Ｉ／Ｏ機器２４は、ケーブル２５に接続されたケーブルアダプタ２９と、接続３１によってケーブルアダプタ２９に接続されたＮＩＣ３０とを含む。各ＮＩＣ２７、３０は、それぞれの接続３２、３３によってネットワーク５に接続されている。Ｉ／Ｏ機器２２、２４の１つが故障し、ＮＩＣ２７、３０の１つによってその通信が最早できない場合、あるいは接続２３、２５の１つが故障すると、サーバ２０は、別のＩ／Ｏ機器２２、２４を用いて、その指定サービス（すなわちネットワーク５に対する接続）を提供し続けることができることは言うまでもない。しかしながら、Ｉ／Ｏ機器２２、２４（又は接続２３、２５）のどちらかが故障すると、ＣＰＵ／ＲＣ１９は、ＮＩＣ２７、３０のうちの切断された１つに送信したデータパケットに対する応答を受信せず、結果的に、サーバ２０の望ましくない故障につながる。

【００３９】

図４は、本発明の実施の形態に基づいて変更された図２の一般的配置を示す。特に、図４の構成では、サーバによって定義される故障ユニット内のケーブルアダプタは、自己修

復ケーブルアダプタに置き換えられており、以下、その動作を更に詳細に説明する。しかしながら、一般論として、自己修復ケーブルアダプタは、致命的エラー（例えばＰＣＩケーブルの故障）を致命的でないエラーに変換する。

【００４０】

図４において、サーバ３５は、PCI Express接続３８によって自己修復ケーブルアダプタ３７に接続されたＣＰＵ／ＲＣ３６を含む。自己修復ケーブルアダプタ３７は、ＣＰＵ／ＲＣ３６をケーブル４０によって、遠隔Ｉ／Ｏ機器３９内に収容されたPCI Expressケーブルアダプタ３８に接続する。遠隔Ｉ／Ｏ機器３９は、ＮＩＣ４１を収容し、ＮＩＣ４１は、PCI Express接続４２によってPCI Expressケーブルアダプタ３８に、イーサネット接続４３によってネットワーク５に接続されている。

【００４１】

より詳細には、自己修復ケーブルアダプタ３７は、自己修復ケーブルアダプタ３７の下流のコンポーネント（例えばケーブル４０、ＮＩＣ４１等）の故障を監視するのに必要な論理及びハードウェアを備えている。例えば、自己修復ケーブルアダプタ３７は、（例えば、供給される「存在検出」信号又は他のあらゆる適切な方法によって）ケーブルが抜かれたときを検出し、遠隔Ｉ／Ｏ機器３９からのＮＩＣ４１に関する問題を示すメッセージを監視するように構成されている。さらに、自己修復ケーブルアダプタ３７は、サーバ３５によって開始されたトランザクションに関する情報を判定して、記録するために、データパケットを調べ、そして、適切な応答を待つように構成されている。自己修復ケーブルアダプタ３７は、オンボードタイマを用いて、その終了が故障の発生となる所定の時間待つことができる。例えば、時間の閾値は、応答データパケットを受信しないことの結果として、サーバ３５がクラッシュすることになる時間よりも前に終了するように、設定することができる。また、自己修復ケーブルアダプタ３７は、ＣＰＵ／ＲＣ３６とは関係なく、データパケットを生成して、ＮＩＣ４１に伝送し、そして、適切な応答を待つように構成されている（それによって、応答が所定の時間内にないことが、ＮＩＣ４１の故障となる）。自己修復ケーブルアダプタ３７の下流のコンポーネントの故障が、複数の異なる原因のうちの１つ以上を有する可能性がある場合、当業者にとって容易に明らかなように、自己修復ケーブルアダプタ３７が、このようなイベントを検出するあらゆる適切な手段を実装できることは言うまでもない。

【００４２】

自己修復ケーブルアダプタ３７は、下流のコンポーネントの故障を検出すると、適切な致命的でないエラーメッセージをＣＰＵ／ＲＣ３６に供給するように、ＮＩＣ４１をエミュレートする。致命的でないエラーメッセージは、ＮＩＣ４１が利用可能であるが、それを用いることができない状態にあることを示す（例えば、イーサネットケーブル４３が抜かれた状態をエミュレートする）。

【００４３】

ＮＩＣ４１をエミュレートするために、自己修復ケーブルアダプタ３７は、ＮＩＣ４１によって生成されたように見えるデータパケットを生成することができる。より詳細には、ＮＩＣ４１は、複数の独立したデバイス機能、すなわちPCI Expressプロトコルによってサポートされている最大８つの機能を有することができる。すなわち、ＮＩＣ４１は、ＣＰＵ／ＲＣ３６に対して、最大８つの別々のデバイスであるように見えることができる。ＮＩＣ４１の各デバイス機能は、一意的にそれらの機能を識別する対応した識別子を有する。ＮＩＣ４１の特定のデバイス機能から送信されたデータパケットは、トランザクション識別子を有し、トランザクション識別子は、データパケットを送信しているデバイス機能の識別子に対応したリクエスタ識別子を含んでいる。

【００４４】

PCI Expressプロトコルによって用いられるデータパケットヘッダのフォーマットを、図５を参照して説明する。リクエスタ識別子４５は、データパケットを始めたデバイス機能を識別し、０から１５までのインデックスが付けられた１６ビットを含む。リクエスタ識別子４５は、上位８ビットを占めるバス番号フィールド４６と、中央の５ビットを占め

るデバイス番号フィールド４７と、下位３ビットを占める機能番号フィールド４８とを含むことが分かる。PCI Expressプロトコルを用いるとき、バス番号４６、デバイス番号４７及び機能番号４８の組合せは、特定のデバイスによって提供される機能を一意的に識別する。

【００４５】

図５に示すパケットヘッダは、更に８ビットからなるタグフィールド４９を含む。上述したように、トランザクションは、要求データパケットと、１つ以上の対応する完了データパケットとで構成されている。各要求データパケットは、タグフィールド４９に記憶されている値に関連する。各対応する完了データパケットは、タグフィールド４９に記憶されている値と同じ値を有し、したがって、完了データパケットを適切な要求データパケットと関連させる。一意的なタグ値が、宛先デバイスから１つ以上の完了データパケットを要求する全ての未処理の要求に割り当てられる。タグフィールド４９は、８ビットを有する場合、２５６（2^8）個の考えられるタグ値を表すことができる。

【００４６】

機能番号フィールド４８は、要求を送信するデバイスに関連した機能の機能番号が設けられる。デバイスが有する機能が８つより少ない場合、機能番号フィールド４８には未使用ビットがあってもよい。したがって、デバイスの機能を表すのに機能番号フィールド４８の十分なビットだけを用い、タグフィールド４９と論理的に組み合わせる仮想機能番号として、機能番号フィールド４８の幾つかの未使用ビットを用いることが、理解される。１つの機能だけが設けられている場合、機能番号フィールド４８の全てのビットは、タグフィールド４９と論理的に組み合わせて、最大２０４８（2^{11}）個の未処理要求のサポートを行うことができる。

【００４７】

自己修復ケーブルアダプタ３７は、受信データパケットのトランザクション識別子を判定するために、そのデータパケット、特にタグフィールド及びリクエスタＩＤを調べることができる。このデータは、自己修復ケーブルアダプタが備えるオンボードメモリに記憶することができる。そして、自己修復ケーブルアダプタ３７は（供給するエラーメッセージの種類がトランザクションＩＤを必要とする場合）、ＣＰＵ／ＲＣ３６が、メッセージがＣＰＵ／ＲＣ３６とＮＩＣ４１間の未処理のトランザクションに関するものと信じるように、正しいトランザクション識別子を与えることができるデータパケットを生成する。同様に、自己修復ケーブルアダプタ３７は、ＮＩＣ４１のリクエスタＩＤを有するデータパケットを生成することができ、それによって、それらのデータパケットは、ＮＩＣ４１によって生成されたように見える。

【００４８】

適切なエラーメッセージは、自己修復ケーブルアダプタ３７によって登録されたトランザクションに関する情報に基づいて選択される。例えば、ＣＰＵ／ＲＣ３６がレジスタ読出し要求データパケットをＮＩＣ４１に送信した場合、下流のコンポーネントが故障して、ＣＰＵ／ＲＣ３６とＮＩＣ４１間の通信が切断されると、自己修復ケーブルアダプタ３７は、エラーステータスビットが「１」の値に設定されたデータパケットを供給することができる。サーバ１上で動作しているソフトウェアは、データパケットの受信を、エラーが発生したことを示すように状態ビットが設定されていると解釈するが、エラーは、サーバ３５の故障の原因とはならない。更なる例として、トランザクションがネットワーク５上で伝送されているデータに関する場合、自己修復ケーブルアダプタ３７は、イーサネット接続４３が切断されたことを示すメッセージを、ＣＰＵ／ＲＣ３６に送信することができる。各エラーメッセージは、自己修復ケーブルアダプタ３７のオンボードメモリにハードコードされていてもよい。そして、自己修復ケーブルアダプタ３７上で動作している論理は、完了データパケットを待っているトランザクションを考慮して、格納されているメッセージのうちのどれが適しているかを決定することができる。

【００４９】

さらに、ＣＰＵ／ＲＣ３６に供給するエラーメッセージの選択は、自己修復ケーブルア

ダプタ３７がエミュレートしている遠隔被接続デバイスの詳細によっても異なってもよい。例えば、遠隔被接続デバイスがストレージリソースの場合、イーサネット接続が切断されたことを示すエラーメッセージは、適切でないことになる。エラーメッセージ自体は、そのデバイスが局所的には機能するが、その指定サービスを実行することができないことを示す（すなわち、致命的エラーを致命的でないエラーに置き換える）ように、遠隔被接続デバイスの種類に合わせることができる。

【００５０】

本発明の上述した実施の形態は、２つの前提に基づいており、第１に、サーバ３５に遠隔で接続されたあらゆるデバイスは、サーバ３５が使えないデバイスと考えるが、サーバ３５がクラッシュする原因とはならない状態を有し、そして、第２に、使えない状態は、Ｉ／Ｏデバイスの完全な動作状態と比較して、簡単である。

【００５１】

さらに、あるいは、被接続サーバ（すなわちサーバへの接続）が故障したときに、遠隔被接続Ｉ／Ｏデバイスが故障することを防止することが、望ましい。これは、限定されないが、特に、遠隔Ｉ／Ｏデバイスが複数のサーバに接続し、複数のサーバによって共有されるシステム（例えば、マルチルートＩ／Ｏ仮想化を実装したシステム）に適している。１つのサーバの故障が、遠隔Ｉ／Ｏデバイスが故障する原因になってはならず、したがって、そのサービスを残りのサーバ（複数のサーバ）に提供することができないことは言うまでもない。

【００５２】

図６は、本発明の実施の形態に基づいて変更された図２の一般的配置を示す。特に、図６の構成では、遠隔Ｉ／Ｏ機器内のケーブルアダプタは、自己修復ケーブルアダプタに置き換えられている。

【００５３】

図６において、サーバ５０は、PCI Express接続５３によってケーブルアダプタ５２に接続されたＣＰＵ／ＲＣ５１を含む。ケーブルアダプタ５２は、ＣＰＵ／ＲＣ５１をケーブル５６によって、遠隔Ｉ／Ｏ機器５５内に収容された自己修復ケーブルアダプタ５４に接続する。また、遠隔Ｉ／Ｏ機器５５は、ＮＩＣ５７を収容し、ＮＩＣ５７は、PCI Express接続５８によって自己修復ケーブルアダプタ５４に、イーサネット接続５９によってネットワーク５に接続されている。

【００５４】

自己修復ケーブルアダプタ５４は、自己修復ケーブルアダプタ５４の上流のコンポーネントの故障を検出するように構成されている。自己修復ケーブルアダプタ５４は、上流のコンポーネントの故障を検出すると、サーバ５０の存在をエミュレートするように構成されている。例えば、自己修復ケーブルアダプタ５４は、ＮＩＣ５７によって発行された未処理のメモリ読出し／書込み要求に対する応答において、直ちに、致命的でないエラー状態を有する完了データパケットを発行するように構成することができる。

【００５５】

自己修復ケーブルアダプタ３７に関して上述したように、ＮＩＣ５７に供給されるエラーメッセージは、完了データパケットが必要とされるトランザクションによって異なる。

【００５６】

サーバ及び遠隔リソースの構成が、自己修復ケーブルアダプタ３７と、自己修復ケーブルアダプタ５４との両方を含むことができることは言うまでもない。すなわち、ある構成において、自己修復ケーブルアダプタは、１つ以上のサーバによって定義される１つ以上のそれぞれの故障ユニット内と、１つ以上のそれぞれの遠隔被接続リソースによって定義される１つ以上の故障ユニット内との両方に設けることができる。

【００５７】

上述の具体的な実施の形態では、サーバに遠隔で接続されたPCI Expressイーサネットネットワークインタフェースカード（ＮＩＣ）を含むシステムに関して説明している。しかしながら、本発明は、より一般的に応用できることは言うまでもない。実際、本発明は

、PCI Express以外の内部接続を利用し、ＮＩＣ以外の遠隔デバイス（例えば、ファイバチャンネルホストバスアダプタ、ストレージコントローラ等）を有するシステム内で用いることができる。

【００５８】

さらに、上述の説明は、サーバと、そのサーバの遠隔被接続リソースとを含むシステムに関するものであるが、本発明は、第１のサーバと第２のサーバ間の通信に用いることができることは言うまでもない。

【００５９】

さらに、本発明は、被接続デバイスの他の構成、例えばマルチサーバＩ／Ｏ仮想化システムに適用することができる。実際、多くの実施の形態において、本発明を利用したシステムが、複数の独立した遠隔デバイスに対する接続をそれぞれ有する複数のサーバを含むことは言うまでもない。このような構成により、１つの遠隔デバイスに関連した故障が生じた場合、各サーバは、他の遠隔被接続デバイスを用いることによって、有効な動作を続けることができる。複数のサーバ又は複数の遠隔リソースに設けられている場合、自己修復ケーブルアダプタ３７、５４は、（必要に応じて、故障したサーバ又は遠隔リソースのアクティブトランザクションを考慮して）適切な致命的でないエラーメッセージを選択するために、どのサーバ又はどの遠隔リソースが故障したかの記録を維持する。

【００６０】

さらに、上述では、自己修復ケーブルアダプタを、本発明の機能と、標準のケーブルアダプタの機能とを含む単一のデバイスとして説明したが、本発明は専用デバイスに実装することができることは言うまでもない。例えば、図２において、本発明を実装したデバイスは、ＣＰＵ／ＲＣ１０とケーブルアダプタ９間、又はケーブルアダプタ９とケーブル１４間に配置することができる。

【００６１】

自己修復ケーブルアダプタ３７、５４は、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）又は特定用途向け集積回路（ＡＳＩＣ）で実現することができる。しかしながら、自己修復ケーブルアダプタ３７、５４は、あらゆる適切な手段を用いて実現してもよいことは言うまでもない。

【００６２】

上述では、サーバとＩ／Ｏデバイス間でデータパケットを伝送する発明の実施の形態を説明した。用語サーバは、広範囲のものを意図するものであり、あらゆるコンピュータ装置を網羅することを意図することは言うまでもない。

【００６３】

当業者にとって容易に明らかなように、本発明は、請求項の要旨を逸脱することなく、ここに教示する内容から変更及び応用することができる。

【図１】

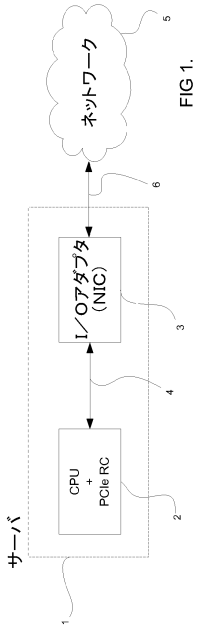

FIG 1.

【図１ａ】

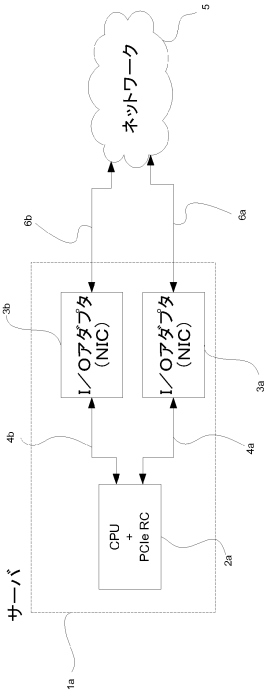

FIG 1a.

【図２】

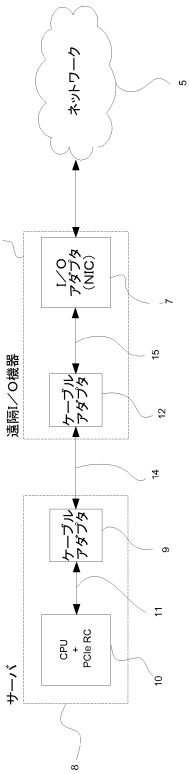

FIG 2.

【図３】

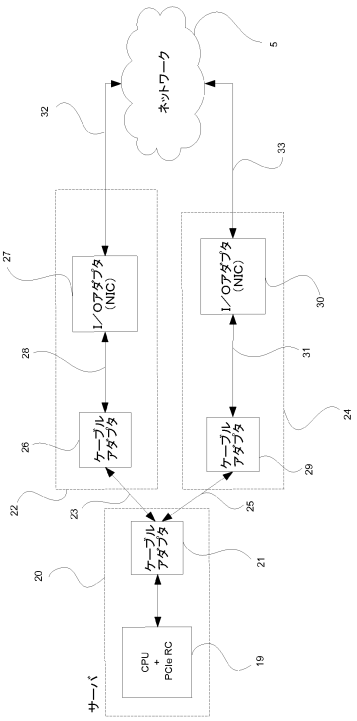

FIG 3.

【図４】

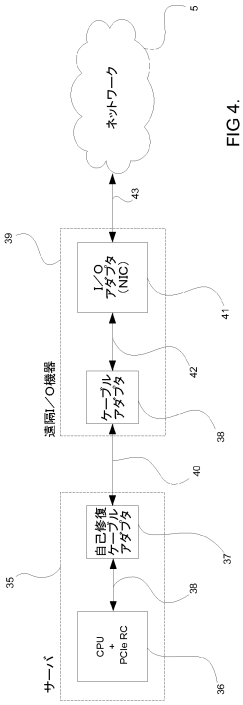

FIG 4.

【図５】

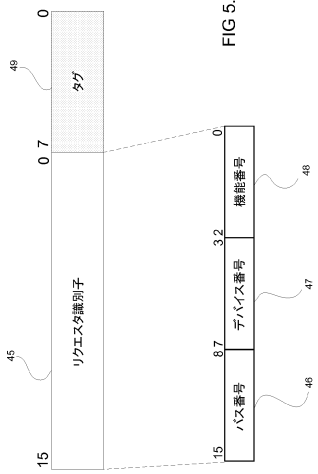

FIG 5.

【図６】

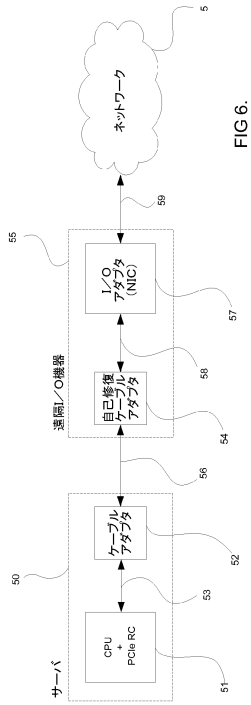

FIG 6.

フロントページの続き

(72)発明者　パイカルスキー、　マレク
イギリス国　エスケイ１１　８ユーイー　チェシャー　マックルズフィールド　ゴースワース　ロード　２０

審査官　田中　啓介

(56)参考文献　特開２００８－１６５２６９（ＪＰ，Ａ）
特開２０１０－２３８１５０（ＪＰ，Ａ）
特開２００７－０９４７０６（ＪＰ，Ａ）
特開２００６－１７２２１８（ＪＰ，Ａ）
特開２００５－１７２２１８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ３／００、３／１８
１３／００、１３／２０－１３／４２