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

特許第5116497号
(P5116497)

(45)発行日　平成25年1月9日(2013.1.9)　(24)登録日　平成24年10月26日(2012.10.26)

(51)Int.Cl.　　F I
G06F 13/14 (2006.01)　　G06F 13/14 310D
G06F 13/10 (2006.01)　　G06F 13/10 330C

請求項の数 12　(全 26 頁)

(21)出願番号　特願2008-20923(P2008-20923)
(22)出願日　平成20年1月31日(2008.1.31)
(65)公開番号　特開2009-181418(P2009-181418A)
(43)公開日　平成21年8月13日(2009.8.13)
審査請求日　平成22年10月13日(2010.10.13)

(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　100114236
弁理士　藤井　正弘
(74)代理人　100075513
弁理士　後藤　政喜
(74)代理人　100120260
弁理士　飯田　雅昭
(72)発明者　沖津　潤
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所　中央研究所内
(72)発明者　保田　淑子
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所　中央研究所内
最終頁に続く

(54)【発明の名称】情報処理システム、Ｉ／Ｏスイッチ及びＩ／Ｏパスの交替処理方法

(57)【特許請求の範囲】
【請求項１】
プロセッサとメモリを備えた物理サーバと、
前記物理サーバの計算機資源を仮想化して仮想サーバを実行する仮想化部と、
前記仮想サーバを実行する物理サーバを複数備え、これら物理サーバと１つ以上のＩ／Ｏデバイスを接続するＩ／Ｏスイッチと、を備えて、前記仮想化部が前記仮想サーバのマイグレーションを行う情報処理システムであって、
前記Ｉ／Ｏスイッチは、
前記Ｉ／Ｏデバイスから仮想サーバへのトランザクション発行の抑止を指示する抑止指示情報を格納するレジスタと、
前記レジスタに前記抑止指示情報が格納されたときに、前記Ｉ／Ｏデバイスからマイグレーションを行う仮想サーバへのトランザクションの発行を抑止し、前記トランザクションの抑止前に前記Ｉ／Ｏデバイスから発行されたトランザクションの完了を保証するトランザクション抑止制御部と、
前記仮想サーバのメモリアドレスと、当該仮想サーバを実行する物理サーバのメモリ上のアドレスとの対応関係を保持するアドレス変換部を備えて、前記仮想サーバのメモリアドレスを前記物理サーバのメモリ上のアドレスに変換する仮想化アシスト部と、
当該Ｉ／Ｏスイッチの構成を管理するスイッチ管理部と、
前記Ｉ／Ｏデバイスから該物理サーバへのトランザクションを保持する第１のバッファと、

前記物理サーバから前記Ｉ／Ｏデバイスへのトランザクションを保持する第２のバッファと、
を備え、
前記仮想化部は、
前記物理サーバ上の仮想サーバと他の仮想サーバに割り付けられたＩ／Ｏデバイスの対応を管理するＩ／Ｏデバイス管理部と、
前記Ｉ／Ｏデバイス管理部からマイグレーション対象の仮想サーバに割り付けられたＩ／Ｏデバイスを特定し、当該Ｉ／Ｏデバイスに対応する前記Ｉ／Ｏスイッチの前記レジスタに対して、前記抑止指示情報を設定および解除するトランザクション指示部と、
前記スイッチ管理部に対して当該Ｉ／Ｏスイッチの構成変更を指示し、前記仮想化アシスト部で管理されるアドレス変換部に対して物理サーバの前記アドレスの変更を指令する構成変更指示部と、を備え、
前記トランザクション指示部は、マイグレーションの開始時に前記抑止指示情報を前記レジスタに設定して、前記Ｉ／Ｏデバイスからマイグレーション対象の仮想サーバへのトランザクション発行を抑止し、前記マイグレーションが完了すると前記レジスタに設定した抑止指示情報を解除して前記レジスタに設定して前記Ｉ／Ｏデバイスからマイグレーションが完了した前記仮想サーバへのトランザクション発行を許可し、
前記トランザクション抑止制御部は、
前記第１のバッファからのトランザクションの発行を抑止するトランザクション抑止部と、
レスポンス付きトランザクションを生成して物理サーバに発行するレスポンス付きトランザクション発行部と、
前記第２のバッファを監視し、前記レスポンス付きトランザクションの完了を確認する応答確認部と、
前記応答確認部が前記レスポンス付きトランザクションの完了を確認したときに、トランザクション抑止の完了を前記仮想化部に通知する完了通知部と、
を含むことを特徴とする情報処理システム。
【請求項２】
前記情報処理システムは、
前記スイッチ管理部を制御するＩ／Ｏマネージャをさらに有し、
前記Ｉ／Ｏマネージャは、前記仮想化部と通信を行う設定インタフェースを備え、
前記仮想化部が当該設定インタフェースを介して前記Ｉ／Ｏマネージャに対して構成変更を指示すると、前記Ｉ／Ｏマネージャが前記スイッチ管理部を制御してＩ／Ｏスイッチの構成を変更することを特徴とする請求項１に記載の情報処理システム。
【請求項３】
前記レジスタは、
前記抑止指示情報を格納するフィールドと、アドレス情報を格納するフィールドと、を有することを特徴とする請求項１に記載の情報処理システム。
【請求項４】
前記アドレス情報を格納するフィールドは、前記マイグレーション対象の仮想サーバのメモリのアドレスを格納することを特徴とする請求項３に記載の情報処理システム。
【請求項５】
前記情報処理システムは、
前記物理サーバを管理するサーバマネージャをさらに有し、
前記サーバマネージャは、
前記仮想化部に対して、前記仮想サーバのマイグレーションの開始を指示する開始指示部を備えたことを特徴とする請求項１記載の情報処理システム。
【請求項６】
前記レスポンス付きトランザクションは、先行するトランザクションを追い抜かないことを特徴とする請求項１に記載の情報処理システム。

【請求項７】
１つ以上の物理サーバと１枚以上のＩ／Ｏデバイスを接続するＩ／Ｏスイッチにおいて、
前記Ｉ／Ｏスイッチは、
前記物理サーバを接続する１つ以上の第１のポートと、前記Ｉ／Ｏデバイスを接続する１つ以上の第２のポートを備え、
前記第２のポートは、
前記Ｉ／Ｏデバイスからのトランザクション発行の抑止を指示する抑止指示情報を保持するレジスタと、
当該第２のポートに接続する前記Ｉ／Ｏデバイスから物理サーバへのトランザクションを保持する第１のバッファと、
前記物理サーバから該第２のポートに接続するＩ／Ｏデバイスへのトランザクションを保持する第２のバッファと、
前記レジスタに前記抑止指示情報が設定されたことを契機に前記第１のバッファからのトランザクション発行を抑止するトランザクション抑止部と、
該トランザクション抑止部によるトランザクション発行の抑止後に、前記抑止指示情報に基づきレスポンス付きトランザクション要求を発行し、前記レスポンス付きトランザクションの完了を確認する滞留トランザクション完了確認部と、を備え、
前記滞留トランザクション完了確認部は、
前記レスポンス付きトランザクション要求を生成して前記物理サーバに発行するレスポンス付きトランザクション発行部と、
前記第２のバッファを監視し、前記レスポンス付きトランザクションの完了を確認する応答確認部と、
を含んで、前記Ｉ／Ｏデバイスから前記物理サーバへ発行されたトランザクションの完了を保証することを特徴とするＩ／Ｏスイッチ。
【請求項８】
前記レジスタは、
前記抑止指示情報を格納するフィールドと、
アドレス情報を保持するフィールドと、
を有することを特徴とする請求項７に記載のＩ／Ｏスイッチ。
【請求項９】
前記物理サーバは、
該物理サーバ上で稼動する１つ以上の仮想サーバと、
該仮想サーバに割当てられたＩ／Ｏデバイスの対応を管理する仮想化部と、を備え、
該仮想化部が、前記レジスタに対して前記抑止指示情報を設定することを特徴とする請求項７に記載のＩ／Ｏスイッチ。
【請求項１０】
前記第２のポートは、
前記レジスタの前記抑止指示情報が解除されたのを契機に、前記Ｉ／Ｏデバイスから該物理サーバへのトランザクションを保持する第１のバッファからのトランザクション発行を再開するトランザクション再開部を含むことを特徴とする請求項７に記載のＩ／Ｏスイッチ。
【請求項１１】
前記レスポンス付きトランザクションは、先行するトランザクションを追い抜かないことを特徴とする請求項７に記載のＩ／Ｏスイッチ。
【請求項１２】
１つ以上のサーバと１枚以上のＩ／Ｏデバイスを１つ以上のＩ／Ｏスイッチで接続する情報処理システムにおいて、前記Ｉ／Ｏスイッチに設定したＩ／Ｏパスを切り替えるＩ／Ｏパスの交替処理方法であって、
前記Ｉ／Ｏスイッチは、

前記Ｉ／Ｏデバイスからのトランザクション発行の抑止を指示する抑止指示情報を保持するレジスタと、
前記Ｉ／Ｏデバイスから前記サーバへのトランザクションを抑止し、前記トランザクションの抑止前に前記Ｉ／Ｏデバイスから発行されたトランザクションの完了を通知するトランザクション抑止制御部と、
前記Ｉ／Ｏスイッチの構成制御を管理するスイッチ管理部と、
を備え、
前記サーバを停止するステップと、
前記レジスタに対して前記抑止指示情報を設定するステップと、
前記Ｉ／Ｏデバイスから該物理サーバへのトランザクションを保持する第１のバッファからのトランザクションの発行を抑止するステップと、
レスポンス付きトランザクションを生成して物理サーバに発行するステップと、
前記物理サーバから前記Ｉ／Ｏデバイスへのトランザクションを保持する第２のバッファを監視して、前記レスポンス付きトランザクションの完了を確認するステップと、
前記レスポンス付きトランザクションの完了を確認したときに、トランザクション抑止の完了を通知するステップと、
前記完了通知を受け、前記サーバに割り付けられたＩ／Ｏデバイスの構成変更を行うステップと、
前記レジスタに設定された該抑止指示情報を解除するステップと、
前記サーバの停止を解除するステップと、
を含むことを特徴とするＩ／Ｏパスの交替処理方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、サーバ計算機とＩ／ＯデバイスをＰＣＩスイッチで接続する情報処理システムに関し、特に、サーバ計算機を仮想化して他の物理計算機へ移動させる技術に関する。

【背景技術】

【０００２】

近年、ＩＴシステムを構成するサーバ台数の増加による運用管理コスト増大と、ＣＰＵマルチコア化等による物理サーバ（物理計算機）の高性能化を背景として、１台の物理サーバを論理的に分割して仮想サーバとして運用するサーバ仮想技術を用いて、物理サーバ台数を削減するサーバ統合が注目を集めるようになっている。

【０００３】

上記のサーバ仮想化技術には幾つかの実現方法が存在するが、中でもハイパバイザを用いた仮想化方式は、仮想化に伴うオーバヘッドが小さいことが特徴として知られている（例えば、非特許文献１）。この非特許文献１に示されるサーバ仮想化方法は、物理サーバのファームウエアとして実装されるハイパバイザと、ハードウエアとして実装される仮想化アシスト機能から構成される。ハイパバイザは、物理サーバ上の各仮想サーバ（仮想計算機）を制御し、仮想化アシスト機能はＩ／Ｏデバイスからのメモリアクセスのアドレスを変換する。非特許文献１に示されるサーバ仮想化方法では、ハイパバイザが仮想サーバにＩ／Ｏデバイスを直接割当てることから、物理サーバと仮想サーバ間で同一のＯＳ環境を利用できる。また、ハードウエアベースの仮想化アシスト機能がＩ／Ｏデバイスのアドレス変換処理を行うため、Ｉ／Ｏ処理に伴うアドレス変換処理オーバヘッドを削減して、Ｉ／Ｏスループットを向上できる。

【０００４】

一方で、ＣＰＵマルチコア化等による物理サーバの高性能化が進むと、物理サーバの性能に応じたＩ／Ｏスループットが求められる。物理サーバの性能に応じたＩ／Ｏスループットを向上する技術としては、複数の物理サーバに対して複数のＩ／Ｏデバイスを接続可能なＰＣＩスイッチが有望である。

【０００５】

さらに、上記サーバ仮想化技術とＰＣＩスイッチを組合せ、仮想サーバとＩ／Ｏデバイスの対応付けを柔軟化するＩ／Ｏ仮想化技術の標準化も進められてきている（ＳＲ／ＭＲ－ＩｏＶ（非特許文献２））。

【０００６】

このような背景のもと、今後、サーバ仮想化技術とＰＣＩスイッチを組み合わせた情報処理システムが主流になると考えられる。

【０００７】

ところで、サーバ仮想化技術を適用した計算機システムの柔軟性や可用性を高める機能の一つにライブマイグレーションがある（非特許文献３）。

【０００８】

ライブマイグレーションは、稼動中の仮想サーバを他の物理サーバに移動させる機能である。ライブマイグレーションによって、物理サーバの負荷に応じて稼動中の仮想サーバの配置を変更したり、メンテナンス対象の物理サーバから稼動中の仮想サーバを退避できる。その結果、システムの柔軟性や可用性を高めることができる。

【０００９】

上記に述べたようなサーバ仮想化技術とＰＣＩスイッチを組み合わせた情報処理システムにおいてライブマイグレーションを実現する場合、稼動中の仮想サーバの３つの状態を保持して引き継ぐ必要がある。３つの状態は、（ａ）ＣＰＵ稼動状態、（ｂ）メモリ内容、（ｃ）Ｉ／Ｏデバイス状態である。（ａ）のＣＰＵ稼動状態は、ハイパバイザが仮想サーバの動作を停止させることで保持することができる。

【００１０】

しかしながら、通常のハイパバイザは、仮想サーバに割当てられたＩ／ＯデバイスからのＤＭＡ等のメモリアクセスや外部Ｉ／Ｏデバイス発のトランザクション処理を止められないため、（ｂ）メモリ内容や（ｃ）Ｉ／Ｏデバイス状態を保持することができない。そのため、上記に述べたようなサーバ仮想化技術においてライブマイグレーションを実現する場合は、稼動するＩ／Ｏデバイスの影響を排除し（ｂ）メモリ内容と、（ｃ）Ｉ／Ｏデバイス状態を保持する必要がある。

【００１１】

これらの状態を保持する技術としては幾つか存在する。

【００１２】

例えば、特許文献１には、物理サーバとＩ／ＯデバイスがＰＣＩスイッチで接続されＰＣＩマネージャによって管理されたシステムにおいて、仮想Ｉ／Ｏデバイスのマイグレーションを実現するために、マイグレーション中にＩ／Ｏデバイスからのトランザクション処理を停止させる方法が開示されている。

【００１３】

この特許文献１では、Ｉ／Ｏデバイスのコンフィギュレーション空間上にマイグレーションビットを持ち、Ｉ／Ｏデバイスが該マイグレーションビットを参照して仮想Ｉ／Ｏデバイスがマイグレーション中かどうかを判断する。

【００１４】

Ｉ／Ｏデバイスは、マイグレーション中であれば自身の処理を停止し、Ｉ／Ｏデバイス発メモリ行きのトランザクションの発行を抑止する。また、ＰＣＩマネージャは、マイグレーション中であればＰＣＩスイッチ内のＩ／Ｏデバイス発メモリ行きのトランザクションを特定のストレージ領域に退避し、マイグレーション完了後にストレージ領域に退避したトランザクションをリストアする。

【００１５】

これにより、マイグレーション中のＩ／Ｏデバイス発トランザクションがメモリ内容を書き替えたり、Ｉ／Ｏデバイス状態が変更されることを防止する。

【００１６】

また、特許文献２では、ホストＯＳがＩ／Ｏデバイスをエミュレーションしたデバイスエミュレータを仮想サーバに提供し、仮想サーバのゲストＯＳが該デバイスエミュレータ

を利用してＩ／Ｏデバイスに間接的にアクセスする方法が開示されている。
【００１７】
ホストＯＳが、マイグレーション中である事を認識し、マイグレーション中であれば、該エミュレーションデバイスの処理を停止させることで、マイグレーション中の仮想サーバのメモリ内容とＩ／Ｏデバイス状態を保持できる。
【特許文献１】米国特許出願公開第２００７／０１８６０２５号明細書
【特許文献２】米国特許第６４９６８４７号明細書
【非特許文献１】上野仁　他共著、"情報システムの運用効率を向上する「BladeSymphony」のサーバ仮想化機構「Virtage」"、2007年7月、インターネット＜http://www.hitachihyoron.com/2007/07/pdf/07a10.pdf＞
【非特許文献２】Michael Krause他　共著、"I/O Virtualization and Sharing"、2006年11月、http://www.pcisig.com/developers/main/training_materials/
【非特許文献３】Christopher Clark他　共著、"Live Migration of Virtual Machines"、2005年5月、NSDI(Networked Systems Design and Implementation)'05
【発明の開示】
【発明が解決しようとする課題】
【００１８】
しかしながら、上記従来例のような　前述の特許文献１を用いた方法では、Ｉ／Ｏデバイスが仮想サーバのマイグレーション中か否かを判定する機能を有している場合にのみ適用可能である。したがってＰＣ用途などに広く流通している汎用Ｉ／Ｏカードを対象に出来ないという問題がある。
【００１９】
また、前述の特許文献２を用いた方法では、仮想サーバがＩ／Ｏデバイスにアクセスを行うために、ゲストＯＳとホストＯＳの切り替えが必要となるため、この切り換えがオーバヘッドとなって処理性能が低下するという問題がある。さらには、仮想サーバにＩ／Ｏデバイスを直接割り付ける場合には、デバイスエミュレータの方式は適用できないという問題もある。
【００２０】
上記問題を解決するため、物理サーバとＩ／ＯデバイスがＰＣＩスイッチを介して接続された情報システムにおいて、汎用的なＩ／Ｏデバイスが仮想サーバに直接割当てられている場合であっても、処理のオーバヘッドを削減しつつ、マイグレーション中仮想サーバのメモリ内容とＩ／Ｏデバイス状態を保持することが必要である。
【００２１】
本発明の課題は、汎用的なＩ／Ｏデバイスが仮想サーバに直接割当てられている場合であっても、処理オーバヘッドを削減しつつ、マイグレーション中の仮想サーバのメモリ内容とＩ／Ｏデバイス状態を保持する機構を提供する事である。
【課題を解決するための手段】
【００２２】
本発明の一実施形態によれば、複数の物理サーバと１枚以上のＩ／Ｏデバイスを接続するＰＣＩスイッチを有する情報処理システムにおいて、物理サーバは、仮想サーバと該仮想サーバに割り付けられたＩ／Ｏデバイスの対応を管理する仮想化部を備える。
【００２３】
Ｉ／Ｏスイッチは、Ｉ／Ｏデバイスから仮想サーバへのトランザクション発行の抑止要求を指示するレジスタと、Ｉ／Ｏデバイスから仮想サーバへのトランザクションを抑止し、抑止前に発行されたＩ／Ｏデバイスからのトランザクションの完了を保証するトランザクション抑止制御部と、仮想サーバのアドレスを物理サーバのメモリ上のアドレスに変換する仮想化アシスト部と、Ｉ／Ｏスイッチの構成を管理するスイッチ管理部を備える。
【００２４】
仮想化部は、Ｉ／Ｏスイッチのレジスタに対して、トランザクション抑止要求（抑止指示情報）と仮想サーバのメモリアドレスを設定するトランザクション指示部と、Ｉ／Ｏス

イッチからの完了通知を受け、スイッチ管理部に対する構成変更の指示と、仮想化アシスト部にアドレス変換部の変更を指示する構成変更指示部を備える。

【００２５】

トランザクション指示部によって、Ｉ／Ｏデバイスから仮想サーバへのトランザクションの抑止と抑止前に発行されたＩ／Ｏデバイスからトランザクションの完了保証処理を行い、Ｉ／Ｏデバイスからのトランザクションが仮想サーバのメモリ状態を書き換えないようにする。また、構成変更指示部によって、Ｉ／ＯデバイスがＤＭＡ等のメモリアドレスが仮想サーバの移動後も有効になるよう仮想化アシスト部のアドレス変換部を更新し、Ｉ／Ｏデバイスの状態を維持する。

【発明の効果】

【００２６】

したがって、本発明は、汎用的なＩ／Ｏデバイスが仮想サーバに直接割当てられている場合であっても、処理のオーバヘッドを削減しつつマイグレーション中の仮想サーバのメモリ内容とＩ／Ｏデバイス状態を保持できる。これにより、ＰＣＩスイッチ等のＩ／Ｏスイッチを介して汎用的なＩ／Ｏデバイスを仮想計算機に割り当てた状態で、仮想計算機を他の物理計算機へ円滑に移動させることが可能となる。

【発明を実施するための最良の形態】

【００２７】

以下、本発明の一実施形態を添付図面に基づいて説明する。

【００２８】

以下に、本発明の実施の形態を添付の図面に基づいて説明する。

【００２９】

図１は、本発明の第１の実施の形態である仮想サーバの状態を保持する機構を備えた情報処理システム１００の構成の一例を示すブロック図である。

【００３０】

情報処理システム１００は、１つ以上の物理サーバ１１０ａ～１１０ｂと、１つ以上のＩ／Ｏデバイス１２０ａ～１２０ｂと、物理サーバ上の仮想サーバを制御するサーバマネージャ１４０と、Ｉ／Ｏデバイス１２０ａ～１２０ｂと物理サーバ１１０ａ～１１０ｂを接続するＰＣＩスイッチ１５０と、ＰＣＩスイッチ１５０を管理するＰＣＩマネージャ１３０とを含む。物理サーバ１１０ａ～１１０ｂと、ＰＣＩマネージャ１３０と、サーバマネージャ１４０と、Ｉ／Ｏデバイス１２０ａ～１２０ｂは、ＰＣＩスイッチ１５０を介して接続されている。また、物理サーバ１１０ａ～１１０ｂと、ＰＣＩマネージャ１３０と、サーバマネージャ１４０はＥｔｈｅｒｎｅｔ（登録商標）あるいはＩ２Ｃ（Inter - Integrated Circuit）などの管理用ネットワーク１０２により接続されている。あるいは、ＰＣＩスイッチ１５０を介してインバウンドでアクセスするための仕組みを備えても良い。物理サーバ１１０ａ～１１０ｂと、ＰＣＩマネージャ１３０と、サーバマネージャ１４０の間で情報の受け渡しが出来ればどのような接続方法でも良い。また、本発明の第１の実施の形態では、２台の物理サーバと、２枚のＩ／ＯデバイスがＰＣＩスイッチに接続されている例を示しているが、この数には限定されない。

【００３１】

物理サーバ１１０ａはハードウエア１１６ａとハイパバイザ１１１ａを備え、物理サーバ１１０ａ上では仮想サーバ１１５ａが動作する。

【００３２】

ハードウエア１１６ａは、物理サーバ上のハードウエアリソースであるＣＰＵ（プロセッサ）、チップセット、メモリ等を備える。ハードウエア１１６ａの物理的な接続構成は後述する（図３参照）。なお、ＰＣＩマネージャ１３０とサーバマネージャ１４０も物理サーバ１１０ａのハードウェア１１６ａと同様にＣＰＵとチップセット及びメモリ等から構成された計算機である。

【００３３】

ハイパバイザ１１１ａは、物理サーバ１１０ａ上に実装されたファームウエアもしくは

アプリケーションであり、物理サーバ１１０ａ上の仮想サーバ１１５ａを管理する。仮想サーバ１１５ａには、ハイパバイザ１１１ａにより管理されるハードウエア１１６ａのＣＰＵやメモリ等のリソースが割当てられる。仮想サーバ１１５ａへのＣＰＵリソースの割当を行うために、ハイパバイザ１１１ａはＣＰＵリソースの仮想サーバ１１５ａへの割当を管理するＣＰＵスケジューラ（図示省略）を保持する。ＣＰＵスケジューラは周知または公知の技術を用いれば良く、本発明の実施の形態では詳細説明を省略する。

【００３４】

本発明の第１の実施の形態では、仮想サーバ１１５ａのみを示すが、仮想サーバ数は一つに制限されない。ハイパバイザ１１１ａが必要に応じて仮想サーバを生成する。

【００３５】

ハイパバイザ１１１ａは、Ｉ／Ｏデバイス管理表１１７ａと、Ｉ／Ｏ発Ｔｘ抑止指示部１１２ａと、Ｉ／Ｏ構成変更指示部１１３ａと、Ｉ／Ｏ発Ｔｘ再開指示部１１４ａと、設定インタフェース１０１ａを備える。

【００３６】

Ｉ／Ｏデバイス管理表１１７ａは、仮想サーバ１１５ａと仮想サーバ１１５ａに割り付けられたＩ／Ｏデバイス１２０ａ、１２０ｂの対応を管理する。Ｉ／Ｏデバイス管理表１１７ａの構成は後述する（図２参照）。

【００３７】

Ｉ／Ｏ発Ｔｘ（トランザクション）抑止指示部１１２ａは、ＰＣＩスイッチ１５０に対して、仮想サーバ１１５ａに割当てられたＩ／Ｏデバイス１２０ａ、１２０ｂが発行するトランザクションを抑止するよう指示する。Ｉ／Ｏ発Ｔｘ抑止指示部１１２ａは、例えば、ＰＣＩスイッチ１５０のコンフィギュレーションレジスタ１５８に設けた設定レジスタ１６１に対して書込みトランザクションを発行する。

【００３８】

Ｉ／Ｏ構成変更指示部１１３ａは、ＰＣＩマネージャ１３０に対してＰＣＩスイッチ１５０の構成を変更するよう指示する。また、Ｉ／Ｏ構成変更指示部１１３ａは、ＰＣＩスイッチ１５０の仮想化アシスト部１５３に対してアドレス変換表１５２を変更するよう指示する。

【００３９】

Ｉ／Ｏ発Ｔｘ再開指示部１１４ａは、ＰＣＩスイッチ１５０に対して、仮想サーバ１１５ａに割当てられたＩ／Ｏデバイス１２０ａ、１２０ｂからのトランザクション（Ｉ／Ｏ発トランザクション）を発行するよう指示する。具体的には、ＰＣＩスイッチ１５０に設けた設定レジスタ１６１に対して書込みトランザクションを発行する。

【００４０】

設定インタフェース１０１ａは、サーバマネージャ１４０およびＰＣＩマネージャ１３０とハイパバイザ１１１ａとの間で設定情報をやり取りするためのインタフェースである。設定情報は、通常のネットワークを用いて情報のやり取りをしても良いし、情報共有用のレジスタをハイパバイザ１１１ａに用意し、レジスタアクセスを介して情報をやり取りしても良い。やり取りするための情報とは、例えばＩ／Ｏ構成変更をするためのトリガとなる情報である。

【００４１】

物理サーバ１１０ａは、以上のように構成されており、物理サーバ１１０ｂも物理サーバ１１０ａと同様に構成されたハイパバイザ１１１ｂ、仮想サーバ１１５ｂ、ハードウェア１１６ｂを備える。

【００４２】

Ｉ／Ｏデバイス１２０ａ～１２０ｂは、ＮＩＣやファイバチャネルのＨＢＡなどの汎用的なＩ／Ｏインタフェースカードである。

【００４３】

ＰＣＩマネージャ１３０は、ＰＣＩスイッチ１５０を管理する管理プログラムを含む計算機である。ＰＣＩマネージャ１３０は、ＣＰＵおよびメモリなどを備えたハードウエア

上に実装されている。ＰＣＩマネージャ１３０は、ＰＣＩスイッチ１５０の構成を変更するためのＩ／Ｏ構成変更部１３１と、ハイパバイザ１１１ａ～１１１ｂやサーバマネージャ１４０との間で設定情報のやり取りを行う設定インタフェース１０１ｄを備える。また、本発明の第１の実施の形態では、単一のＰＣＩマネージャ１３０のみ示すが、信頼性向上のため複数のＰＣＩマネージャを備えても良い。その場合、複数のＰＣＩマネージャ間では整合性が取れるように情報の制御を行う。

【００４４】

サーバマネージャ１４０は、情報処理システム１００全体を管理するプログラムを含む計算機であり、ＣＰＵおよびメモリなどを備えたハードウエア上に実装されている。サーバマネージャ１４０は、仮想サーバ１１５ａ（１１５ｂ）のマイグレーションの処理開始を指示する処理開始指示部１４１と、ハイパバイザ１１１ａ～１１１ｂやＰＣＩマネージャ１３０との間で設定情報のやり取りを行う設定インタフェース１０１ｅとを備える。本発明の第１の実施の形態では、単一のサーバマネージャ１４０のみを示すが、信頼性向上のため複数のサーバマネージャを備えても良い。この場合、複数のサーバマネージャ間では整合性が取れるように情報を制御する。

【００４５】

ＰＣＩスイッチ１５０は、複数の物理サーバ１１０ａ～１１０ｂとＩ／Ｏデバイス１２０ａ～１２０ｂとを接続するスイッチファブリックであり、物理サーバ１１０ａ、１１０ｂとＩ／Ｏデバイス１２０ａ、１２０ｂの間でトランザクションの送受信を行う。トランザクションの詳細については後述する（図１０参照）。

【００４６】

本発明の第１の実施の形態で示す情報処理システム１００では、単一のＰＣＩスイッチ１５０のみを示しているが、複数のＰＣＩスイッチを備えていてもよい。それぞれのＰＣＩスイッチが別々のＰＣＩマネージャで管理されていても良いし、単一のＰＣＩマネージャで管理されていても良い。また、本発明の第１の実施の形態では、ＰＣＩスイッチ１５０はＰＣＩ－Ｅｘｐｒｅｓｓプロトコルを対象としたスイッチファブリックとする。ただし、ＰＣＩスイッチ１５０は、ＰＣＩプロトコルやＰＣＩ－Ｘプロトコルといった他のプロトコルを対象としたスイッチファブリック（Ｉ／Ｏスイッチ）であってもよい。

【００４７】

ＰＣＩスイッチ１５０は、１つ以上のＵｐｓｔｒｅａｍポート１５１ａ～１５１ｃと、１つ以上のＤｏｗｎｓｔｒｅａｍポート１６０ａ～１６０ｂと、スイッチング部１５７と、ＰＣＩスイッチ管理部１５４と、設定レジスタ１６１及びルーティングテーブルなどを格納するコンフィギュレーションレジスタ１５８とを備え、スイッチング部１５７は仮想化アシスト部１５３を含む。本発明の第１の実施の形態では、３つのＵｐｓｔｒｅａｍポートと２つのＤｏｗｎｓｔｒｅａｍポートを示すが、ポート数はこの数には限定されない。

【００４８】

Ｕｐｓｔｒｅａｍポート１５１ａ～１５１ｃは物理サーバ１１０ａ、１１０ｂ、ＰＣＩマネージャ１３０を接続するポートである。

【００４９】

Ｄｏｗｎｓｔｒｅａｍポート１６０ａ～１６０ｂはＩ／Ｏデバイス１２０ａ、１２０ｂを接続するポートである。Ｄｏｗｎｓｔｒｅａｍポート１６０ａは、Ｔｘ抑止制御部１６２と、設定レジスタ１６１を備える。Ｔｘ抑止制御部１６２は、設定レジスタ１６１に設定された情報に従い、仮想サーバ１１５ａ、１１５ｂに割当てられたＩ／Ｏデバイス１２０ａ、１２０ｂからのトランザクションの抑止と再開を行う。設定レジスタ１６１は、ハイパバイザ１１１ａのＩ／Ｏ発Ｔｘ抑止指示部１１２ａから発行される抑止指示要求を保持する。Ｄｏｗｎｓｔｒｅａｍポート１６０ａの詳細は後述する（図５参照）。

【００５０】

ＰＣＩスイッチ管理部１５４は、ＰＣＩスイッチ１５０のスイッチング部１５７を制御する機能要素であり、ＰＣＩマネージャ１３０と連携して動作する。ＰＣＩスイッチ管理

部１５４はＰＣＩスイッチ１５０を１つ又は複数のＰＣＩツリーに分割して管理する。ＰＣＩツリーは、１つのＵｐｓｔｒｅａｍポートと１つ又は複数のＤｏｗｎｓｔｒｅａｍポートの組からなる。ＰＣＩスイッチ管理部１５４は、あるＰＣＩツリーのポートに接続する物理サーバ１１０ａ、１１０ｂやＩ／Ｏデバイス１２０ａ、１２０ｂが該ＰＣＩツリーに接続していない物理サーバやＩ／Ｏデバイスにアクセスできないよう、ＵｐｓｔｒｅａｍポートとＤｏｗｎｓｔｒｅａｍポート間のトランザクションのルーティングを行う。ＰＣＩスイッチ管理部１５４は、複数のＰＣＩツリーが存在する場合は、ＰＣＩツリー識別子を用いてＰＣＩツリーを一意に特定する。また、ＰＣＩスイッチ管理部１５４は、ＰＣＩマネージャ１３０のＩ／Ｏ構成変更部１３１からの指示により、ＰＣＩスイッチ１５０のコンフィギュレーションレジスタ１５８を書き換えて、ＰＣＩツリーに属するＵｐｓｔｒｅａｍポートと、Ｄｏｗｎｓｔｒｅａｍポートの設定変更を行う。

【００５１】

スイッチング部１５７は、ＰＣＩスイッチ管理部１５４が管理するツリー構造に従って、ＵｐｓｔｒｅａｍポートとＤｏｗｎｓｔｒｅａｍポートとの間でトランザクションを転送する。

【００５２】

仮想化アシスト部１５３は、Ｕｐｓｔｒｅａｍポート１５１ａ～１５１ｃとＤｏｗｎｓｔｒｅａｍポート１６０ａ～１６０ｂを接続する経路１５５、１５６上（スイッチング部１５７）に位置する制御回路であり、アドレス変換表１５２を含む。アドレス変換表１５２は、仮想サーバ１１５ａ、１１５ｂのアドレス（仮想アドレス）と物理サーバ１１０ａ、１１０ｂのアドレス（物理アドレス）の対応関係を管理する表である。アドレス変換表１５２の詳細は後述する（図４参照）。仮想化アシスト部１５３は、アドレス変換表１５２を参照して、Ｉ／Ｏデバイス１２０ａ、１２０ｂから仮想サーバ１１０ａ、１１０ｂに発行されるトランザクションのアドレスを変換して、物理サーバ１１０ａ、１１０ｂのメモリにトランザクションを発行する。また、仮想化アシスト部１５３は、ハイパバイザ１１１ａ、１１１ｂと連携して動作し、アドレス変換表１５２に設定されている仮想アドレスと物理アドレスの対応を設定し、必要に応じて変更する。本発明の第１の実施の形態では、仮想化アシスト部１５３を単一の制御要素として示したが、Ｕｐｓｔｒｅａｍポートもしくは Ｄｏｗｎｓｔｒｅａｍポート毎に分散して設けても良い。

【００５３】

図３は、物理サーバ１１０ａのハードウエア１１６ａの構成を説明した図である。ハードウエア１１６ａは、ＣＰＵ３０１、チップセット３０２、メモリ３０３、Ｒｏｏｔポート３０４を含む。ＣＰＵ３０１は、プログラムを実行するプロセッサである。ＣＰＵ３０１は、メモリアドレス領域の一部にＰＣＩスイッチ１５０のレジスタをマッピングしたＭＭＩ／Ｏ（メモリマップドＩ／Ｏ）を介して、ＰＣＩスイッチ１５０のレジスタにアクセスする。また、ＣＰＵ３０１およびチップセット３０２は、ＣＰＵ発Ｔｘ抑止要求（Ｑｕｉｅｓｃｅｎｃｅ要求）とＣＰＵ発Ｔｘ抑止解除要求（Ｄｅｑｕｉｅｓｃｅ要求）の処理をサポートし、ＣＰＵ３０１発トランザクションの抑止と抑止解除が可能である。また、チップセット３０２は、ＣＰＵ３０１、メモリ３０３、Ｒｏｏｔポート３０４を接続する。メモリ３０３は、物理サーバ１１０ａの主記憶領域であり、一部の領域が仮想サーバ１１５ａのメモリ領域として利用される。Ｒｏｏｔポート３０４は、ＰＣＩスイッチ１５０のＵｐｓｔｒｅａｍポート１５１ｂと接続し、接続するＰＣＩツリーのルートとなる。

【００５４】

図１０は、物理サーバ１１０ａ、１１０ｂとＩ／Ｏデバイス１２０ａ、１２０ｂ間で送受信されるトランザクションの構成である。トランザクションは、ヘッダ１２０１とペイロード１２０２で構成されるパケットである。

【００５５】

ヘッダ１２０１は、ＰＣＩスイッチ１５０がトランザクションをルーティングするために必要な情報であり、トランザクション送信元識別子１２０３（Ｒｅｑｕｅｓｔｅｒ ＩＤ）と、送信先アドレス１２０４と、トラフィッククラス１２０５を含む。また、複数の

ＰＣＩツリーがある場合には、ヘッダ１２０１にＰＣＩツリー識別子が追加される。ヘッダ情報は、標準化仕様で決まるため、本発明の実施の形態では、詳細説明を省略する。ここでは本発明に関連するヘッダ情報のみ説明する。

【００５６】

トランザクション送信元識別子１２０３は、送信元のＩ／Ｏデバイスや ＲｏｏｔポートのＰＣＩツリーにおけるバス番号、デバイス番号、ファンクション番号から構成される識別子であり、送信元Ｉ／Ｏデバイスに接続するＤｏｗｎｓｔｒｅａｍポートや Ｒｏｏｔポートに接続するＵｐｓｔｒｅａｍポートを一意に識別できる。

【００５７】

送信先アドレス１２０４は、トランザクション送信先のメモリのアドレスである。トラフィッククラス１２０５は、トランザクションが通過するＰＣＩスイッチ１５０内のＶｉｒｔｕａｌ Ｃｈａｎｎｅｌを一意に特定するための情報である。本発明の実施の形態では、トラフィッククラスは０から７までの値を設定可能である。

【００５８】

ペイロード１２０２は、トランザクションが保持するデータを格納する。例えば、メモリライトトランザクションの場合は、メモリに書き込むデータを格納する。

【００５９】

図２は、Ｉ／Ｏデバイス管理表１１７ａの構成を示す図である。なお、物理サーバ１１０ｂのＩ／Ｏデバイス管理表１１７ｂも同様に構成される。

【００６０】

Ｉ／Ｏデバイス管理表１１７ａは、ハイパバイザ１１１ａにより管理され、仮想サーバ識別子２０１と、Ｉ／Ｏデバイス識別子２０２を含む。仮想サーバ識別子２０１はＩ／Ｏデバイス管理表１１７ａを保持する物理サーバ１１０ａ内の仮想サーバ１１５ａを一意に識別する番号である。Ｉ／Ｏデバイス識別子２０２は、物理サーバ１１０ａに割当てられたＩ／Ｏデバイスを一意に識別する識別子である。本発明の第１の実施の形態では、Ｉ／Ｏデバイス識別子２０２として図１０に示したトランザクション送信元識別子１２０３を用いる。トランザクション送信元識別子は、トランザクションのヘッダ１２０１に含まれており、Ｉ／Ｏデバイスが接続するＤｏｗｎｓｔｒｅａｍポートを一意に特定可能である。ただし、トランザクションに含まれ、Ｉ／Ｏデバイスの接続するＤｏｗｎｓｔｒｅａｍポートを一意に識別できるならば他の識別子を用いても良い。

【００６１】

図４は、アドレス変換表１５２の詳細を示した図である。アドレス変換表１５２は、ＰＣＩツリー毎に管理され、Ｉ／Ｏデバイス識別子４０１、変換オフセット４０２から成る組を保持する。Ｉ／Ｏデバイス識別子４０１は、Ｉ／Ｏ発トランザクションのヘッダに含まれているトランザクション送信元識別子１２０３を用い、ＰＣＩツリー内のＩ／Ｏデバイスを一意に識別する。変換オフセット４０２は、該Ｉ／Ｏデバイスが割当てられた仮想サーバ１１５ａのメモリ領域の、物理サーバ１１０ａのメモリ領域における開始アドレスを示す。

【００６２】

仮想化アシスト部１５３は、アドレス変換表１５２を参照してＩ／Ｏ発トランザクションに含まれるトランザクション送信元識別子１２０３に対応する変換オフセットを求め、該変換オフセットを用いてＩ／Ｏ発トランザクションの送信先アドレス１２０４を仮想アドレスから物理アドレスに変換する。本発明の第１の実施の形態では、アドレス変換表１５２はＰＣＩツリー毎に分割して管理されているが、複数のＰＣＩツリーをまとめて一つのアドレス変換表で管理しても良い。

【００６３】

図５に、Ｄｏｗｎｓｔｒｅａｍポート１６０ａのブロック構成を示す。Ｄｏｗｎｓｔｒｅａｍポート１６０ａは、設定レジスタ１６１と、Ｔｘ抑止制御部１６２と、接続するＩ／ＯデバイスからのトランザクションをＵｐｓｔｒｅａｍポートへ送信する受信バッファ５０７と、Ｕｐｓｔｒｅａｍポートからのトランザクションを、このＤｏｗｎｓｔｒｅａ

ｍポートに接続されたＩ／Ｏデバイスへ送信する送信バッファ５０８を備え、Ｉ／Ｏデバイス１２０ａ、１２０ｂと物理サーバ１１０ａ間を流れるトランザクションの送受信を行う。

【００６４】

設定レジスタ１６１は、図１１で示すようにＰＣＩスイッチ１５０のコンフィギュレーションレジスタ１５８に設けたもので、Ｄｏｗｎｓｔｒｅａｍポート１６０ａ、１６０ｂ毎に設けたものであり、設定レジスタ１６１がＤｏｗｎｓｔｒｅａｍポート１６０ａに対応し、設定レジスタ１６１ｂがＤｏｗｎｓｔｒｅａｍポート１６０ｂに対応する。各設定レジスタ１６１、１６１ｂは、物理サーバ１１０ａ、１１０ｂやＰＣＩマネージャ１３０からＭＭＩ／Ｏ経由でアクセス可能なレジスタであり、Ｄｏｗｎｓｔｒｅａｍポート毎に抑止ビット５０９とアドレス５１０を格納するフィールドを含む。抑止ビット５０９は、ハイパバイザ１１１ａのＩ／Ｏ発Ｔｘ抑止指示部１１２ａおよびＩ／Ｏ発Ｔｘ再開指示部１１４ａからの指示に従い、設定、解除される。本発明の第１の実施の形態では、ＰＣＩのコンフィギュレーション空間への書込みを行うトランザクションを用いる。例えば、Ｉ／Ｏ発Ｔｘ抑止指示部１１２は、コンフィギュレーション空間への書込みトランザクションにより、抑止ビット５０１に１をセットし、Ｉ／Ｏ発Ｔｘ再開指示部１１４ａは、コンフィギュレーション空間への書込みトランザクション用いて、抑止ビット５０９の値を０クリアする。設定レジスタ１６１のアドレス５１０には、Ｉ／Ｏ発Ｔｘ抑止指示部１１２ａからの指示に従い、マイグレーションの対象となる仮想サーバの仮想アドレスが設定される。また、Ｉ／Ｏ発Ｔｘ再開指示部１１２ａは、コンフィギュレーション空間への書込みトランザクション用いて、アドレス５１０の値をクリアする。

【００６５】

Ｔｘ抑止制御部１６２は、Ｔｘ抑止部５０１と、滞留Ｔｘ完了確認部５０２と、Ｔｘ再開部５０３を備える。

【００６６】

Ｔｘ抑止部５０１は、受信バッファ５０７内のＩ／Ｏ発トランザクションをＵｐｓｔｒｅａｍポート５０７へ発行されないように制御する。例えば、ＰＣＩスイッチ１５０内の経路１５６（スイッチング部１５７）と受信バッファ５０７間で行われるフロー制御の仕組みを利用し、Ｔｘ抑止部５０１は受信バッファ５０７が発行したトランザクションに対するＡＣＫ（応答）を返送しないように制御することで、受信バッファ５０７のトランザクションの発行を抑止する。

【００６７】

滞留Ｔｘ完了確認部５０２は、レスポンス付きＴｘ発行部５０４と、Ｔｘ応答確認部５０５と、Ｔｘ完了通知部５０６を備え、抑止前に発行されたＩ／Ｏ発トランザクションの完了を保証する。

【００６８】

レスポンス付きＴｘ発行部５０４は、設定レジスタ１６１のアドレス５１０を送信先アドレスとするメモリリードトランザクションを生成し、生成したメモリリードトランザクションを発行する。メモリリードトランザクションは、レスポンス付きＴｘの一種である。

【００６９】

Ｔｘ応答確認部５０５は、レスポンス付きＴｘ発行部５０４が発行したメモリリードトランザクションの応答の受信を確認する。本発明の第１の実施の形態では、Ｔｘ応答確認部５０５が送信バッファ５０８を監視し、レスポンス付きＴｘ発行部５０４が発行したメモリリードトランザクションの応答を送信バッファ５０８が受信したことを確認する。

【００７０】

Ｔｘ完了通知部５０６は、Ｔｘ応答確認部５０５が応答を確認した事をうけて、Ｄｏｗｎｓｔｒｅａｍポート１６０ａと接続したＩ／Ｏデバイスに割当てられた物理サーバのハイパバイザ１１１ａに、Ｔｘ抑止制御の完了を通知する。

【００７１】

Ｔｘ再開部５０３は、設定レジスタ１６１の抑止ビット５０９がクリアされた事を検出すると受信バッファ５０７内のＩ／Ｏ発トランザクションの発行を再開する。例えば、Ｔｘ再開部５０３が、受信バッファ５０７が発行したトランザクションに対するＡＣＫの返送の抑止を解除することで、受信バッファ５０７のトランザクションの発行は再開される。

【００７２】

なお、Ｄｏｗｎｓｔｒｅａｍポート１６０ｂも上記Ｄｏｗｎｓｔｒｅａｍポート１６０ａと同様に構成され、設定レジスタ１６１ｂとＴｘ抑止制御部１６２ｂを備える。また、計算機と接続されるＵｐｓｔｒｅａｍポート１５１ａ～１５１ｃは、少なくとも送信バッファと受信バッファを備えていれば良く、設定レジスタやＴｘ抑止制御部が設定されていなくても良い。

【００７３】

図６は、図５で説明したＴｘ抑止制御部１６２（または１６２ｂ）で行われる処理の一例を示すフローチャートである。また、図７は、Ｔｘ抑止制御部１６２の処理に関するトランザクションの流れを示した図である。以下、図６および図７を参照してＴｘ抑止制御部の処理フローを説明する。なお、以下の説明では、Ｄｏｗｎｓｔｒｅａｍポート１６０ａに関する処理の一例を示すが、Ｄｏｗｎｓｔｒｅａｍポート１６０ｂも同様の処理を行うことができる。

【００７４】

図６、図７の本処理は、Ｉ／ＯデバイスＩ／Ｏデバイス１２０ａがＤｏｗｎｓｔｒｅａｍポート１６０ａに接続されている場合に開始される（Ｓ６００）。

【００７５】

先ず、Ｔｘ抑止制御部１６２のＴｘ抑止部４０１は、設定レジスタ１６１の抑止ビット５０９を監視し、抑止ビットが設定されたかどうかをチェックする（Ｓ６０１）。具体的には、Ｔｘ抑止制御部１６２は、設定レジスタ１６１の抑止ビット５０９の０から１への遷移を検出するまで、設定レジスタ１６１から抑止ビット５０９の値を繰り返して読み出す（図７の７０１）。抑止ビット５０９に１が設定された場合、次のステップＳ６０２を行う。抑止ビット５０９が０である場合には、ステップＳ６０１を繰り返す。

【００７６】

次に、受信バッファ５０７内のトランザクションの発行を抑止する（Ｓ６０２）。具体的には、Ｔｘ抑止部５０１が、受信バッファ５０７がＵｐｓｔｒｅａｍポート１５１ａ～１５１ｃへ向けて発行したトランザクションに対するＡＣＫ（応答）の返送を抑止する（図７の７０２）。

【００７７】

次に、Ｓ６０２によるトランザクションの発行抑止前に発行されたＩ／Ｏ発トランザクションの完了を保証するために、設定レジスタ１６１が保持するアドレス５１０に対してレスポンス付Ｔｘを生成して発行する（Ｓ６０３）。Ｉ／Ｏデバイスと物理サーバ間のパス（Ｉ／Ｏパス）が複数のパスに分かれている場合は、全てのＩ／Ｏパスについてレスポンス付きトランザクションを発行する。具体的には、滞留Ｔｘ完了確認部５０２のレスポンス付きＴｘ発行部５０４が、設定レジスタ１６１からアドレス５１０を取得し（図７の７０３）、該アドレスに対してレスポンス付きＴｘの１つであるメモリリードトランザクションを生成し（図７の７０４）、Ｕｐｓｔｒｅａｍポートへ発行する（図７の７０５）。また、ＰＣＩスイッチ１５０が複数のＶｉｒｔｕａｌ　Ｃｈａｎｎｅｌを備える場合には、ＰＣＩスイッチ１５０の全ての利用可能なＶｉｒｕｔａｌ　Ｃｈａｎｎｅｌについて Ｉ／Ｏ発トランザクションの完了を保証するために、トラフィッククラスが０から７のヘッダを付加した８種類のメモリリードトランザクションを発行する。全てのトラフィッククラスに対してメモリリードトランザクションを発行することで、利用可能な全てのＶｉｒｔｕａｌ　Ｃｈａｎｎｅｌについて Ｉ／Ｏ発トランザクションの保証が行える。

【００７８】

次に、Ｓ６０３で発行したレスポンス付きトランザクションの完了を確認する（Ｓ６０

４）。ＰＣＩ－Ｅｘｐｒｅｓｓのオーダリングルールでは、レスポンス付きトランザクションは、先行するメモリライトトランザクションを追い抜かないため、発行したレスポンス付きトランザクションが完了すれば、抑止前に発行されたＩ／Ｏ発メモリライトトランザクションが完了していることが保証される。すなわち、Ｓ６０３で発行したレスポンス付きトランザクションの完了を確認した後は、仮想サーバ１１５ａのメモリ内容が保持される。滞留Ｔｘ完了確認部５０２のＴｘ応答確認部５０５が、メモリリードトランザクションの応答（図７の７０６）を全て確認するまで、ステップＳ６０４を繰り返す（図７の７０７）。

【００７９】

次に、Ｔｘ抑止制御部１６２がハイパバイザ１１１ａにトランザクション抑止完了通知を送信する（Ｓ６０５）。具体的には、滞留Ｔｘ完了確認部５０２のＴｘ完了通知部５０６が、ハイパバイザ１１１ａに処理完了通知７０８を発行する。処理完了通知７０８の発行手段は、物理サーバ１１０ａに対する割込みでも、物理サーバ１１０ａのメモリ３０３に対する書込みでもよい。

【００８０】

次に、Ｉ／Ｏ発トランザクションの抑止指示情報が解除されるまで待つ（Ｓ６０６）。具体的には、Ｔｘ再開部５０３は、設定レジスタ１６１の抑止ビット５０９が０にクリアされたかどうかをチェックし（図７の７０１）、抑止ビット５０９が０に遷移したことを受けて次の処理に進む。

【００８１】

最後に、受信バッファ５０７内のＩ／Ｏ発トランザクションの抑止を解除する（Ｓ６０７）。具体的には、Ｔｘ再開部５０３が、受信バッファ５０７が発行したトランザクションに対するＡＣＫの返送の抑止を解除することで、受信バッファ５０７からＵｐｓｔｒｅａｍポート１５１ａ～１５１ｃへ向けたトランザクションの発行を再開させる（図７の７０９）。

【００８２】

以上の一連の処理を持って、Ｔｘ抑止制御部１６２の処理を完了する（Ｓ６０８）。以上のように、Ｔｘ抑止制御部１６２がＩ／Ｏ発トランザクションの抑止と完了保証を行うことで、物理サーバ１１０ａのメモリ内容をＩ／Ｏ発トランザクションから保護できる。すなわち、Ｉ／Ｏ発トランザクションの抑止を開始した後に、設定レジスタ１６１のアドレス５１０にセットした移動元の仮想サーバのアドレスに対してレスポンス付きトランザクションを発行し、発行したレスポンス付きトランザクションの完了を確認することで、抑止開始以前に発行したトランザクションが完了していることを保証することができるのである。

【００８３】

図８は、情報処理システム１００で仮想サーバのライブマイグレーションを実現する処理のフローチャートである。また、図９は、ライブマイグレーションの処理に関するトランザクションの流れを示した情報処理システム１００の要部を示すブロック図である。

【００８４】

図８のフローチャートは、Ｉ／Ｏデバイス１２０ａが割当てられている仮想サーバ１１５ａを、物理サーバ１１０ａから物理サーバ１１０ｂにマイグレーションするという想定で説明する。

【００８５】

マイグレーションの対象となる仮想サーバ１１５ａをマイグレーション対象仮想サーバと呼ぶ。マイグレーション対象仮想サーバが存在する物理サーバ１１０ａをマイグレーション元物理サーバと呼び、マイグレーション元物理サーバ１１０ａ上のハイパバイザ１１１ａをマイグレーション元ハイパバイザと呼ぶ。また、マイグレーション先の物理サーバ１１０ｂをマイグレーション先物理サーバと呼び、マイグレーション先物理サーバ上１１０ｂのハイパバイザ１１１ｂをマイグレーション先ハイパバイザと呼ぶ。

【００８６】

図８のフローチャートは、サーバマネージャ１４０の処理開始指示部１４１がマイグレーション元ハイパバイザ１１１ａに対してマイグレーション対象仮想サーバ１１５ａのマイグレーション開始要求を発行してから、マイグレーション先ハイパバイザ１１１ｂがサーバマネージャ１４０にマイグレーション完了を通知するまでの処理フローを示す。

【００８７】

ライブマイグレーションの処理は、マイグレーション元ハイパバイザ１１１ａがサーバマネージャ１４０の処理開始指示部１４１によるマイグレーション開始要求９０１を受付けたことを契機に開始される（Ｓ８０１）。

【００８８】

図９のマイグレーション開始要求９０１には、マイグレーション元仮想サーバ１１５ａの仮想サーバ識別子および、マイグレーション先物理サーバ１１０ｂの識別子を含む。マイグレーション元ハイパバイザ１１１ａは、マイグレーション開始要求９０１を受けて以下の処理を行う。

【００８９】

先ず、マイグレーション元ハイパバイザ１１１ａは、マイグレーション対象仮想サーバ１１５ａを停止する（Ｓ８０２）。実現にあたっては既知の手段を用いる。本発明の第１の実施の形態では、先ず、マイグレーション元ハイパバイザ１１１ａは、ハードウエア１１６ａのＣＰＵ３０１およびチップセット３０２に対してＣＰＵ発Ｔｘ抑止要求を。次に、ハイパバイザ１１１ａが、仮想サーバ１１５ａにＣＰＵリソースを割当てないようにＣＰＵスケジューラの設定を変更し、仮想サーバ１１５ａの動作を停止させる。最後にＣＰＵ発Ｔｘ抑止解除要求を行う。以上の処理により、仮想サーバ１１５ａの停止と仮想サーバ１１５ａ発のトランザクションの完了が保証される。

【００９０】

次に、マイグレーション元ハイパバイザ１１１ａは、ＰＣＩスイッチ１５０に対して、マイグレーション対象仮想サーバ１１５ａに接続するＩ／Ｏデバイス１２０ａからトランザクションの抑止を指示する（Ｓ８０３）。具体的には、マイグレーション元ハイパバイザ１１１ａのＩ／Ｏ発Ｔｘ抑止指示部１１２ａが、Ｉ／Ｏデバイス管理表１１７ａを参照して、仮想サーバ１１５ａに割当てられているＩ／Ｏデバイス１２０ａを抽出する。次に、Ｉ／Ｏ発Ｔｘ抑止指示部１１２ａは、Ｉ／Ｏデバイス１２０ａに接続するＤｏｗｎｓｔｒｅａｍポート１６０ａの設定レジスタ１６１に対して、ＰＣＩスイッチ１５０のＭＭＩ／Ｏを経由してコンフィギュレーションレジスタ１５８への書込みを行う（図９の９０２）。具体的には、設定レジスタ１６１の抑止ビット５０９に１をセットし、アドレス５１０に仮想サーバ１１５ａが用いるメモリアドレスの一部をセットする。その結果、Ｄｏｗｎｓｔｒｅａｍポート１６０ａのＴｘ抑止制御部１６２により、仮想サーバ１１５ａに割当てられたＩ／Ｏデバイス１２０ａからのトランザクションの抑止処理が開始される。このＩ／Ｏデバイスからのトランザクションの抑止処理フローは図６で説明した通りである。

【００９１】

次に、マイグレーション元ハイパバイザ１１１ａは、ＰＣＩスイッチ１５０のＴｘ抑止制御部１６２からのＩ／Ｏデバイス１２０ａ発のトランザクションの抑止完了通知を待つ（Ｓ８０４）。マイグレーション元ハイパバイザ１１１ａが、抑止完了通知７０８を受け取ることで、マイグレーション対象の仮想サーバ１１５ａ宛の全てのトランザクションが完了したことを確認でき、Ｉ／Ｏ発トランザクションによって仮想サーバ１１５ａのメモリ内容が書き換わらないことを保証できる。

【００９２】

次に、マイグレーション元ハイパバイザ１１１ａは、マイグレーション対象仮想サーバ１１５ａをマイグレーション元物理サーバ１１０ａからマイグレーション先物理サーバ１１０ｂに移動させる（Ｓ８０５）。実現にあたっては既知の手段を用いる。具体的には、ハイパバイザ１１１ａが、仮想サーバ１１５ａのＯＳ及びアプリケーションのイメージをマイグレーション先物理サーバ１１０ｂにコピーする。本発明の第１の実施の形態では、

ハイパバイザ１１１ａが、管理用ネットワーク１０２を利用したアウトバウンド通信を利用して、仮想サーバ１１５ａのメモリ内容とハイパバイザ１１１ａが保持する仮想サーバ１１５ａの構成情報を、マイグレーション先物理サーバ１１０ｂにコピーする。ただし、ＰＣＩスイッチ１５０を介したインバウンド通信による仮想サーバの移動を行っても良い。

【００９３】

次に、マイグレーション元ハイパバイザ１１１ａは、マイグレーション対象仮想サーバ１１５ａに割当てられたＩ／Ｏデバイス１２０ａの引継ぎを指示する（Ｓ８０６）。具体的には、ハイパバイザ１１１ａのＩ／Ｏ構成変更指示部１１３ａは、ＰＣＩスイッチ管理部１５４の設定を変更し、Ｉ／Ｏデバイス１２０ａが接続するＰＣＩツリーを変更する。すなわち、Ｉ／Ｏデバイス１２０ａの割り付けを物理サーバ１１０ａから物理サーバ１１０ｂの仮想サーバ１１５ａへ切り替える。また、仮想化アシスト部１５３にアクセスし、アドレス変換表１５２における物理アドレスと仮想アドレスの対応を変更する。この変更は、アドレス変換表１５２のＩ／Ｏデバイス識別子４０１がＩ／Ｏデバイス１２０ａに一致するエントリの変換オフセット４０２を、仮想サーバ１１５ａの移動先である物理サーバ１１０ｂのメモリの物理アドレスに対応するオフセット値に更新する。つまり、ハイパバイザ１１１ａの構成変更指示部１１３ａがＩ／Ｏデバイス識別子４０１と新たな物理サーバ１１０ｂの変換オフセット４０２をＰＣＩスイッチ管理部１５４に指令する。その後、ハイパバイザ１１１ａは仮想サーバ１１５ａが物理サーバ１１０ａから削除されたので、Ｉ／Ｏデバイス管理表１１７ａからＩ／Ｏデバイス１２０ａに関する情報を削除する。

【００９４】

ＰＣＩスイッチ管理部１５４の設定を変更するためには、既知の手段を用いる。具体的には、ハイパバイザ１１１ａのＩ／Ｏ構成変更指示部１１３ａが、ＰＣＩマネージャ１３０のＩ／Ｏ構成変更部１３１に対して、Ｉ／Ｏ構成変更要求９０６を発行する。Ｉ／Ｏ構成変更要求９０６は、ＰＣＩスイッチ１５０もしくは設定インタフェース１０１ａ、１０１ｄを介してＰＣＩマネージャ１３０に送られる。Ｉ／Ｏ構成変更要求９０６には、ＰＣＩスイッチの識別子、引継ぎ対象Ｉ／ＯデバイスのＩ／Ｏデバイス識別子が含まれる。また、ＰＣＩスイッチ１５０内に複数のＰＣＩツリーが含まれる場合は、Ｉ／Ｏ構成変更要求９０６には、引継ぎ元ＰＣＩツリーのＰＣＩツリー識別子、引継ぎ先ＰＣＩツリーのＰＣＩツリー識別子が含まれる。Ｉ／Ｏ構成変更部１３１は、Ｉ／Ｏ構成変更要求９０６を受けて、設定変更要求９０７をＰＣＩスイッチ管理部１５４に発行する。

【００９５】

また、アドレス変換表１５２の設定を変更するために、Ｉ／Ｏ構成変更指示部１１３ａは、仮想化アシスト部１５３に対して、アドレス変換表更新要求９０４を発行する。アドレス変換表更新要求９０４には、Ｉ／Ｏデバイス識別子、変換オフセットの情報が含まれる。ＰＣＩスイッチ１５０内に複数のＰＣＩツリーが含まれる場合は、アドレス変換表更新要求９０４には、ＰＣＩツリー識別子が含まれる。仮想化アシスト部１５３は、アドレス変換表更新要求９０４を受けて、アドレス変換表１５２を更新する。

【００９６】

ＰＣＩスイッチ管理部１５４の設定と仮想化アシスト部１５３の設定を変更することで、Ｄｏｗｎｓｔｒｅａｍポート１６０ａの受信バッファ５０７に存在するトランザクションは、マイグレーション先物理サーバ１１０ｂ上の仮想サーバ１１５ａに割当てられたメモリ領域に書き込まれる。

【００９７】

次に、マイグレーション先ハイパバイザ１１１ｂは、Ｉ／Ｏ構成変更の完了通知を待つ（Ｓ８０７）。完了通知は、ハイパバイザ１１１ｂがＰＣＩマネージャ１３０のＩ／Ｏ構成変更部１３１から明示的に受け取っても良いし、ハイパバイザ１１１ｂがマイグレーション先物理サーバ１１０ｂにＩ／Ｏデバイス１２０ａが追加されたことを共有レジスタなどを介して検知するのでもよい。ハイパバイザ１１１ｂがＩ／Ｏ構成変更が完了した事を認識できれば、何れの方法であっても良い。ハイパバイザ１１１ｂは、Ｉ／Ｏ構成変更の

完了通知を受け、Ｉ／Ｏデバイス管理表１１７ｂにＩ／Ｏデバイス１２０ａに関する情報を追加する。

【００９８】

次に、マイグレーション先ハイパバイザ１１１ｂは、マイグレーション対象仮想サーバ１１５ａの処理を再開させる（Ｓ８０８）。実現にあたっては既知の手段を用いる。例えば、ハイパバイザ１１１ｂが、仮想サーバ１１５ａにＣＰＵリソースを割当てるようハイパバイザ１１１ｂのＣＰＵスケジューラの設定を変更し、仮想サーバ１１５ａの動作を再開させる。

【００９９】

次に、マイグレーション先ハイパバイザ１１１ｂは、ＰＣＩスイッチ１５０に対して、Ｉ／Ｏデバイス発トランザクションの再開を指示する（Ｓ８０９）。具体的には、ハイパバイザ１１１ｂのＩ／Ｏ発Ｔｘ再開指示部１１４ｂは、Ｉ／Ｏデバイス管理表１１７ｂを参照して、マイグレーション対象仮想サーバ１１５ａに割当てられているＩ／Ｏデバイス１２０ａが接続するＤｏｗｎｓｔｒｅａｍポート１６０ａを抽出する。次に、Ｄｏｗｎｓｔｒｅａｍポート１６０ａに対応する設定レジスタ１６１に対して、コンフィギュレーション空間への書込みトランザクションを用いてレジスタアクセスを行う（図９の９０８）。レジスタアクセス９０８では、設定レジスタ１６１の抑止ビット５０９とアドレス５１０を０にクリアする。Ｄｏｗｎｓｔｒｅａｍポート１６０ａのＴｘ再開部５０３は、設定レジスタ１６１の各種設定情報が０にクリアされた事を検知し、Ｉ／Ｏデバイス１２０ａからのトランザクションの送付を再開させる。

【０１００】

最後に、ステップ８１０（Ｓ８１０）ではマイグレーション先ハイパバイザ１１１ｂは、サーバマネージャ１４０にマイグレーションが完了した事を通知する（図９の９０９）。

【０１０１】

以上の一連の処理を持って、ライブマイグレーションを完了する（Ｓ８１１）。以上のように、ライブマイグレーションの処理を行うことで、マイグレーション対象仮想サーバ１１５ａに割当てられたＩ／Ｏデバイス１２０ａからのトランザクションがマイグレーション中の仮想サーバのメモリ領域に書き込まれることを防止できる。そのため、ライブマイグレーション対象仮想サーバ１１５ａのメモリ状態とＩ／Ｏデバイス状態の保持を実現できる。

【０１０２】

＜変形例１＞

本発明の第１の実施の形態で述べた仮想サーバの状態を保持する機構は、仮想サーバのライブマイグレーション用途の他に、Ｉ／Ｏパス交替機能にも適用可能である。Ｉ／Ｏパス交替機能は、物理サーバと物理サーバに割り付けられたＩ／Ｏデバイスの間の経路（Ｉ／Ｏパス）を現用系と待機系の複数を用意し、Ｉ／Ｏパス上のポート等に障害が発生した際にＩ／Ｏパスを現用系から待機系にフェールオーバする機能である。Ｉ／Ｏパス交替機能によって、ＰＣＩスイッチのポート故障による情報処理システムの停止を回避できるため、情報処理システムの可用性が向上する。

【０１０３】

図１２は、本発明の変形例１におけるＩ／Ｏパス交替機能を備えた情報処理システム１０００の構成を示すブロック図である。

【０１０４】

情報処理システム１０００は、１つ以上の物理サーバ１０１０と、１つ以上のＩ／Ｏデバイス１２０ａと、ＰＣＩマネージャ１０３０と、サーバマネージャ１０４０と、１つ以上のＰＣＩスイッチ１５０ａ～１５０ｂとを含む。物理サーバ１０１０と、ＰＣＩマネージャ１０３０と、サーバマネージャ１０４０と、Ｉ／Ｏデバイス１２０ａは、ＰＣＩスイッチ１５０ａ～１５０ｂを介して接続されている。ＰＣＩスイッチ１５０ａとＰＣＩスイッチ１５０ｂは、Ｕｐｓｔｒｅａｍポート１５１ａとＤｏｗｎｓｔｒｅａｍポート１６０

ｃ、Ｕｐｓｔｒｅａｍポート１５１ｂとＤｏｗｎｓｔｒｅａｍポート１６０ｄという２つの経路で接続されている。また、物理サーバ１０１０と、ＰＣＩマネージャ１０３０と、サーバマネージャ１０４０は管理用ネットワーク１０２により接続されている。なお、ＰＣＩスイッチ１５０ａ、１５０ｂは、前記第１実施形態と同様に、コンフィギュレーションレジスタとスイッチング部とを備えるが図１２においては図示を省略した。また、Ｄｏｗｎｓｔｒｅａｍポート１６０ａ～１６０ｄはＴｘ抑止制御部とコンフィギュレーションレジスタに設定された設定レジスタ１６１を備えるが、Ｄｏｗｎｓｔｒｅａｍポート１６０ｂ～１６０ｄについては設定レジスタ等を省略した。

【０１０５】

物理サーバ１０１０には、Ｉ／Ｏデバイス１２０ａが割当られており、物理サーバ１０１０とＩ／Ｏデバイス１２０ａの間で現用系のＩ／ＯパスとしてＵｐｓｔｒｅａｍポート１５１ｄ、Ｄｏｗｎｓｔｒｅａｍポート１６０ｄ、Ｕｐｓｔｒｅａｍポート１５１ｂ、Ｄｏｗｎｓｔｒｅａｍポート１６０ａを経由するＩ／Ｏパスが設定されている。また、待機系のＩ／ＯパスとしてＵｐｓｔｒｅａｍポート１５１ｄ、Ｄｏｗｎｓｔｒｅａｍポート１６０ｃ、Ｕｐｓｔｒｅａｍポート１５１ａ、Ｄｏｗｎｓｔｒｅａｍポート１６０ａを経由するＩ／Ｏパスが設定されている。

【０１０６】

情報処理システム１０００を構成するコンポーネントの構成は本発明の第１の実施の形態で示した情報処理システム１００を構成するコンポーネントの構成と類似するため、以下、情報処理システム１０００と情報処理システム１００の差分について説明する。

【０１０７】

物理サーバ１０１０は、ＣＰＵ、チップセット、メモリ等から成るハードウエア１１６を備える。物理サーバ１０１０では、ＯＳ１０１５が動作し、ＯＳ１０１５ではアプリケーション１０１６が動作する。

【０１０８】

ＯＳ１０１５は、ドライバモジュール１０１７と、Ｉ／Ｏ障害検出部１０１１と、サーバ発Ｔｘ抑止部１０１２と、Ｉ／Ｏパス交替指示部１０１３と、サーバ発Ｔｘ再開部１０１４を備える。ドライバモジュール１０１７はＩ／Ｏデバイス１２０ａのドライバである。Ｉ／Ｏ障害検出部１０１１と、サーバ発Ｔｘ抑止部１０１２と、Ｉ／Ｏパス交替指示部１０１３と、サーバ発Ｔｘ再開部１０１４は、ＯＳ１０１５の一機能として実装されていているため、アプリケーション１０１６が意識することなく、Ｉ／Ｏパス交替処理を行える。

【０１０９】

Ｉ／Ｏ障害検出部１０１１は、ＯＳ１０１５が利用するＩ／ＯデバイスやＰＣＩスイッチからの障害通知を検出し、障害がＩ／Ｏパスに関する場合、Ｉ／Ｏパス交替処理を開始する。具体的には、Ｉ／Ｏ障害検出部１０１１は、例えばＰＣＩ－Ｅｘｐｒｅｓｓスイッチが備えるＡｄｖａｎｃｅｄ　Ｅｒｒｏｒ　Ｒｅｐｏｒｔｉｎｇ機能により受け取ったＩ／Ｏ障害通知を解析し、障害がＩ／Ｏパスに関する場合は、物理サーバ１０１０をリセットせずに、Ｉ／Ｏパス交替処理を開始する。

【０１１０】

サーバ発Ｔｘ抑止部１０１２は、障害が発生したＩ／Ｏパスを利用するＩ／Ｏデバイスへのトランザクション（サーバ発トランザクション）の発行を抑止する。実現にあたっては既知の手段を用いる。例えば、物理サーバ１０１０上のＯＳ１０１５は、Ｉ／ＯデバイスのＨｏｔ　Ｐｌｕｇ機能を備え、Ｈｏｔ　Ｐｌｕｇの仕組みを用いて物理サーバ１０１０に割当てられているＩ／Ｏデバイスを切り離す。

【０１１１】

Ｉ／Ｏパス交替指示部１０１３は、ＰＣＩマネージャ１０３０に対して、障害が発生したＩ／Ｏパスの交替を指示する。具体的には、Ｉ／Ｏパス交替指示部１０１３は、ＰＣＩマネージャ１０３０に対してＩ／Ｏパス交替要求を発行する。Ｉ／Ｏパス交替要求は、障害が発生したＰＣＩスイッチの識別子と、障害発生ポートの識別子が含まれる。Ｉ／Ｏパ

ス交替要求は、物理サーバ１０１０からＰＣＩスイッチ１５０ｂを介してＰＣＩマネージャ１０３０に通知されても良いし、物理サーバ１０１０から管理用ネットワーク１０２を介してサーバマネージャ１０４０を経由してＰＣＩマネージャ１０３０に通知されても良い。

【０１１２】

サーバ発Ｔｘ再開部１０１４は、サーバ発Ｔｘ抑止部１０１２によって抑止されたＩ／Ｏデバイスへのトランザクションの発行を再開させる。実現にあたっては既知の手段を用いる。例えば、ＯＳ１０１５は、Ｉ／ＯデバイスのＨｏｔ　Ｐｌｕｇ機能を備え、Ｈｏｔ　Ｐｌｕｇの仕組みを用いて物理サーバ１０１０に割当てられているＩ／Ｏデバイスを接続する。

【０１１３】

サーバ発Ｔｘ抑止部１０１２およびサーバ発Ｔｘ再開部１０１４は、サーバ発トランザクションの制御を行うための機構であればＨｏｔ　Ｐｌｕｇ機能に限定されない。例えば、物理サーバ１０１０がハイパバイザを備える場合、Ｓ８０２やＳ８０８で述べた方法を用いて、ハイパバイザがサーバ発トランザクションの制御を行っても良い。また、物理サーバ１０１０に接続するＰＣＩスイッチ１５０ｂが物理サーバ１０１０からのトランザクションを制御する機構を備えるならば、物理サーバ１０１０はサーバ発トランザクションの制御を行うための機構を備えなくても良い。

【０１１４】

ＰＣＩマネージャ１０３０は、Ｉ／Ｏ構成変更部１３１と、Ｉ／Ｏ発Ｔｘ抑止指示部１１２ｄ、Ｉ／Ｏ構成変更指示部１１３ｄ、Ｉ／Ｏ発Ｔｘ再開指示部１１４ｄと、Ｉ／Ｏパス交替完了通知部１０３１を備える。Ｉ／Ｏ発Ｔｘ抑止指示部１１２ｄと、Ｉ／Ｏ構成変更指示部１１３ｄと、Ｉ／Ｏ発Ｔｘ再開指示部１１４ｄは、本発明の第１の実施の形態で示した物理サーバ１１０ａが備えるＩ／Ｏ発Ｔｘ抑止指示部１１２ａと、Ｉ／Ｏ構成変更指示部１１３ａと、Ｉ／Ｏ発Ｔｘ再開指示部１１４ａと同一である。

【０１１５】

Ｉ／Ｏパス交替完了通知部１０３１は、Ｉ／Ｏパス交替を指示した物理サーバ１０１０に対して、Ｉ／Ｏパス交替完了を通知する。物理サーバ１０１０は、Ｉ／Ｏパス交替完了の通知を受けて、サーバ発トランザクションを再開させる。

【０１１６】

図１３は、ＰＣＩスイッチ１５０ａ、１５０ｂ上でＩ／Ｏパス交替を実現する処理のフローチャートである。以下、物理サーバ１０１０と物理サーバ１０１０に割当てられているＩ／Ｏデバイス１２０ａ間のＩ／ＯパスのＵｐｓｔｒｅａｍポート１５１ｂに障害が発生し、Ｉ／Ｏパスを現用系から待機系にフェールオーバするという想定で、図１２に示すＩ／Ｏパス交替処理を説明する。

【０１１７】

Ｉ／Ｏパス交替処理は、ＰＣＩスイッチ間のパスで障害が発生した際などに開始される（Ｓ１１００）。

【０１１８】

先ず、物理サーバ１０１０において、Ｉ／Ｏパスで障害が発生した事を検出する（Ｓ１１０１）。具体的には、Ｉ／Ｏ障害検出部１０１１は、例えばＰＣＩ－Ｅｘｐｒｅｓｓスイッチが備えるＡｄｖａｎｃｅｄ　Ｅｒｒｏｒ　Ｒｅｐｏｒｔｉｎｇ機能により、ＰＣＩスイッチ１５０ａのＵｐｓｔｒｅａｍポート１５１ｂで障害が発生した事を検出する。

【０１１９】

次に、物理サーバ１０１０において、障害が発生したＩ／Ｏパスを利用するＩ／Ｏデバイス１２０ａへのトランザクション（サーバ発トランザクション）の発行を抑止し、ＰＣＩマネージャ１０３０にＩ／Ｏパス交替を指示する（Ｓ１１０２）。具体的には、サーバ発Ｔｘ抑止部１０１２は、Ｈｏｔ　Ｐｌｕｇの仕組みを用いてＩ／Ｏデバイス１２０ａを切り離す。その結果、物理サーバ１０１０から、Ｉ／Ｏパス交替対象のＩ／Ｏパスへのトランザクションが抑止される。また、Ｉ／Ｏパス交替指示部１０１３は、ＰＣＩマネージ

ャ１０３０に対してＩ／Ｏパス交替要求を発行する。Ｉ／Ｏパス交替要求には、障害発生ＰＣＩスイッチ１５０ａの識別子と、障害発生ポート１５１ｂの識別子が含まれる。

【０１２０】

ＰＣＩマネージャ１０３０は、Ｉ／Ｏパス交替要求を受け取ると、先ず、障害発生Ｉ／Ｏパスに接続するＩ／Ｏデバイス１２０ａからのトランザクションを抑止する（Ｓ１１０３）。具体的には、ＰＣＩマネージャ１０３０のＩ／Ｏ発Ｔｘ抑止指示部１１２ｂは、Ｉ／Ｏデバイス１２０ａが接続するＤｏｗｎｓｔｒｅａｍポート１６０ａの設定レジスタ１６１に対して、ＰＣＩスイッチのＭＭＩ／Ｏを経由してコンフィギュレーションレジスタへの書込みを行う。その結果、上述と同様にＤｏｗｎｓｔｒｅａｍポート１６０ａのＴｘ抑止制御部１６２により、Ｉ／Ｏデバイス１２０ａからのトランザクションの抑止処理が開始される。Ｉ／Ｏデバイスからのトランザクションの抑止処理フローは図６で説明した通りである。

【０１２１】

次に、ＰＣＩマネージャ１０３０は、Ｔｘ抑止制御部１６２からの抑止完了通知を待つ（Ｓ１１０４）。ＰＣＩマネージャ１０３０が、抑止完了通知を受け取ることでＩ／Ｏデバイス１２０ａからのトランザクションが抑止されたことを保証できる。

【０１２２】

次に、ＰＣＩマネージャ１０３０は、ＰＣＩスイッチ１５０ａ～１５０ｂに対して、Ｉ／Ｏ構成変更を指示する（Ｓ１１０５）。具体的には、ＰＣＩマネージャ１０３０のＩ／Ｏ構成変更指示部１１３ａは、障害発生ポート１５１ｂを避ける待機系Ｉ／Ｏパス情報を元に、ＰＣＩスイッチ１５０ａ～１５０ｂに関するＩ／Ｏパス交替情報を生成する。ＰＣＩスイッチ１５０ａに関するＩ／Ｏパス交替情報は、交替前パス情報「Ｕｐｓｔｒｅａｍポート１５１ｂ、Ｄｏｗｎｓｔｒｅａｍポート１６０ａ」、交替後パス情報「Ｕｐｓｔｒｅａｍポート１５１ａ、Ｄｏｗｎｓｔｒｅａｍポート１６０ａ」を含む。また、ＰＣＩスイッチ１５０ｂに関するＩ／Ｏパス交替情報は、交替前パス情報「Ｕｐｓｔｒｅａｍポート１５１ｂ、Ｄｏｗｎｓｔｒｅａｍポート１６０ａ」、交替後パス情報「Ｕｐｓｔｒｅａｍポート１５１ａ、Ｄｏｗｎｓｔｒｅａｍポート１６０ａ」を含む。次に、Ｉ／Ｏ構成変更指示部１１３ａは、Ｉ／Ｏ構成変更部１３１に対してＩ／Ｏ構成変更要求発行する。Ｉ／Ｏ構成変更要求には、前述のＰＣＩスイッチ１５０ａ～１５０ｂに関するＩ／Ｏパス交替情報が含まれる。

【０１２３】

Ｉ／Ｏ構成変更部１３１は、Ｉ／Ｏ構成変更要求を受けて、設定変更要求をＰＣＩスイッチ管理部１５４ａ～１５４ｂに発行する。ＰＣＩスイッチ管理部１５４ａに対する設定変更要求は、前述のＰＣＩスイッチ１５０ａに関するＩ／Ｏパス交替情報を含む。また、ＰＣＩスイッチ管理部１５４ｂに対する設定変更要求は、前述のＰＣＩスイッチ１５０ｂに関するＩ／Ｏパス交替情報を含む。ＰＣＩスイッチ管理部１５４ａ～１５４ｂは、設定変更要求に従い、該当するＰＣＩツリーに属するポートの構成を変更する。なお、Ｉ／Ｏパス交替処理では、仮想サーバの移動を伴わないため、アドレス変換表の設定の変更は発生しない。

【０１２４】

次に、ＰＣＩマネージャ１０３０は、ＰＣＩスイッチ１５０ａに対して、Ｉ／Ｏパス交替対象Ｉ／Ｏデバイス１２０ａからのトランザクションの再開を指示する（Ｓ１１０６）。具体的には、ＰＣＩマネージャ１０３０のＩ／Ｏ発Ｔｘ再開指示部１１４ｂは、Ｉ／Ｏデバイス１２０ａが接続するＤｏｗｎｓｔｒｅａｍポート１６０ａの設定レジスタ１６０に対して、ＰＣＩスイッチ１５０ａのＭＭＩ／Ｏを経由してコンフィギュレーションレジスタへの書込みを行う。その結果、Ｄｏｗｎｓｔｒｅａｍポート１６０ａのＴｘ抑止制御部１６２により、Ｉ／Ｏデバイス１２０ａからのトランザクションが再開される。

【０１２５】

次に、ＰＣＩマネージャ１０３０は、物理サーバ１０１０に対し、Ｉ／Ｏパス交替完了を通知する（Ｓ１１０７）。具体的には、サーバマネージャ１０４０のＩ／Ｏパス交替完

了通知部１０３１は、管理用ネットワーク１０２を介して、物理サーバ１０１０に対して、Ｉ／Ｏパス交替完了を通知する。

【０１２６】

物理サーバ１０１０では、Ｉ／Ｏパス交替完了の通知を受けて、サーバ発トランザクションを再開させる（Ｓ１１０８）。具体的には、サーバ発Ｔｘ再開部１０１４は、Ｉ／Ｏパス交替完了の通知を受けて、Ｈｏｔ　Ｐｌｕｇの仕組みを用いてＩ／Ｏデバイス１２０ａを接続する。その結果、物理サーバ１０１０からＩ／Ｏデバイス１２０ａへのトランザクションの発行が再開される。

【０１２７】

以上の一連の処理を持って、Ｉ／Ｏパス交替処理を完了する（Ｓ１１０９）。以上のように、Ｉ／Ｏパス交替処理を行うことで、交替対象のＩ／Ｏパス上にトランザクションが存在しないことが保証された状態でＩ／Ｏパス交替を実現でき、Ｉ／Ｏパス交替処理に伴うトランザクションの喪失などを回避できる。

【０１２８】

＜変形例２＞

変形例２は、上記変形例１と比べ、Ｉ／Ｏデバイス１２０ａのドライバモジュール１０１７が、図１２に示した変形例１のＩ／Ｏ障害検出部１０１１と、サーバ発Ｔｘ抑止部１０１２と、Ｉ／Ｏパス交替指示部１０１３と、サーバ発Ｔｘ再開部１０１４を備え、さらに、Ｉ／Ｏパス交替を実現する機構を備える点が異なる。ドライバモジュール１０１７が備えるＩ／Ｏパス交替を実現する機構は、例えば、Ｉ／Ｏデバイス１２０ａと連携することでＩ／Ｏデバイス１２０ａに関するＩ／Ｏパスを複数管理し、それらＩ／Ｏパスを任意に変更する機構である。

【０１２９】

ドライバモジュール１０１７のＩ／Ｏパス交替指示部１０１３は、ＰＣＩマネージャ１０３０に対してＩ／Ｏ発トランザクションの抑止を指示すると共に、ドライバモジュール１０１７が備えるＩ／Ｏパス交替を実現する機構を用いて、Ｉ／Ｏパスの交替を行う。

【０１３０】

ドライバモジュール１０１７がＩ／Ｏパス交替処理の機構を備えることで、ＯＳ１０１５やアプリケーション１０１６がＩ／Ｏパス交替処理の機構を備えなくても、Ｉ／Ｏパス交替を行える。一方で、Ｉ／Ｏデバイス１２０ａおよびドライバモジュール１０１７は、Ｉ／Ｏパス交替処理のための機構を備える必要があるため、汎用的なＩ／Ｏデバイスに適用できない。

【０１３１】

＜変形例３＞

変形例３は、上記変形例１、変形例２と比べ、ＰＣＩマネージャ１０３０がＩ／Ｏ障害検出部１０１１を備える点が異なる。

【０１３２】

本変形例３では、ＰＣＩマネージャ１０３０が図１２に示した変形例１のＩ／Ｏ障害検出部１０１１を有し、Ｉ／Ｏ障害検出部１０１１は、Ｉ／Ｏパス障害を検出したら、物理サーバ１０１０に対してＩ／Ｏパス障害を通知する。物理サーバ１０１０のサーバ発Ｔｘ抑止部１０１２は、Ｉ／Ｏパス障害の通知を受けて、サーバ発トランザクションの発行を抑止する。

【産業上の利用可能性】

【０１３３】

以上のように、本発明では、Ｉ／Ｏスイッチを備えて計算機とＩ／Ｏデバイスの接続を動的に変更する計算機システムや、Ｉ／Ｏスイッチ内の経路を動的に変更する計算機システムに適用することができる。

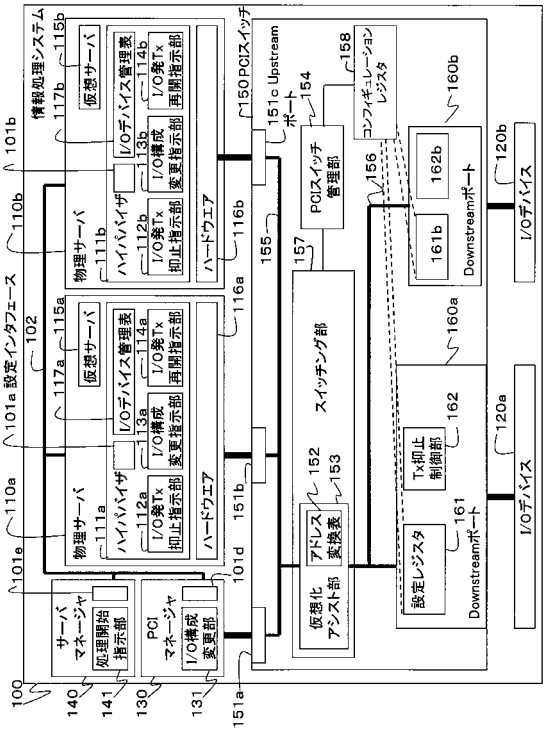
【図面の簡単な説明】

【０１３４】

【図１】第１の実施の形態を示し仮想サーバを実行する情報処理システムの構成の一例を

示すブロック図である。
【図２】第１の実施の形態を示し、Ｉ／Ｏデバイス管理表の一例を示す説明図である。
【図３】第１の実施の形態を示し、物理サーバのハードウェア構成を示すブロック図である。
【図４】第１の実施の形態を示し、アドレス変換表の一例を示す説明図。
【図５】第１の実施の形態を示し、Ｄｏｗｎｓｔｒｅａｍポートの構成を示すブロック図である。
【図６】第１の実施の形態を示し、Ｔｘ抑止制御部で行われる処理の一例を示すフローチャートである。
【図７】第１の実施の形態を示し、Ｔｘ抑止制御部の処理に関するトランザクションの流れを示すブロック図である。
【図８】第１の実施の形態を示し、情報処理システムで行われる仮想サーバのライブマイグレーション処理の一例を示すフローチャートである。
【図９】第１の実施の形態を示し、ライブマイグレーションの処理に関するトランザクションの流れを示した情報処理システムの要部を示すブロック図である。
【図１０】第１の実施の形態を示し、トランザクションの構成を示すブロック図である。
【図１１】第１の実施の形態を示し、コンフィギュレーションレジスタに設定された設定された設定レジスタを示すブロック図。
【図１２】変形例１を示し、Ｉ／Ｏパス交替機能を備えた情報処理システムの構成を示すブロック図である。
【図１３】計算機システムで行われるＰＣＩスイッチ１５０ａ、１５０ｂのＩ／Ｏパス交替処理のフローチャートである。
【符号の説明】
【０１３５】
１００、１０００　情報処理システム
１１０ａ、１１０ｂ、１１０ｃ　物理サーバ
１１１ａ、１１１ｂ、１１１ｃ　ハイパバイザ
１１２ａ、１１２ｂ、１１２ｄ　Ｉ／Ｏ発Ｔｘ抑止指示部
１１３ａ、１１３ｂ、１１３ｄ　Ｉ／Ｏ構成変更指示部
１１４ａ、１１４ｂ、１１４ｄ　Ｉ／Ｏ発Ｔｘ再開指示　部
１１５ａ、１１５ｂ　仮想サーバ
１２０ａ、１２０ｂ　Ｉ／Ｏデバイス
１３０　ＰＣＩマネージャ
１４０　サーバマネージャ
１５０　ＰＣＩスイッチ
１５３　仮想化アシスト部
１５４　ＰＣＩスイッチ管理部
１６０ａ、１６０ｂ　Ｄｏｗｎｓｔｒｅａｍポート
１６１　設定レジスタ
１５２　Ｔｘ抑止制御部

【図１】

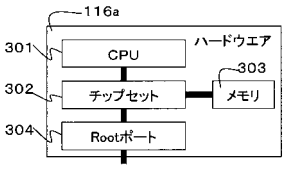


【図２】

I/Oデバイス管理表117a

201 仮想サーバ識別子	202 I/Oデバイス識別子
1	10012
2	10011
2	11012
…	…

【図３】

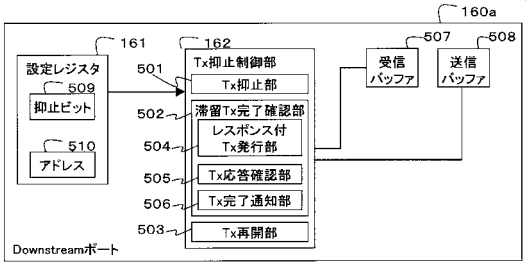


【図４】

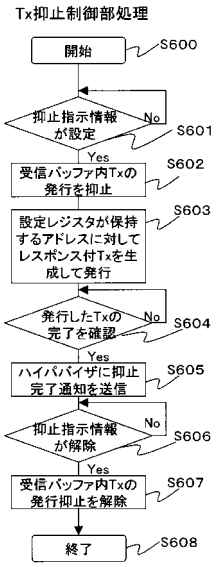
アドレス変換表152

401 I/Oデバイス識別子	402 変換オフセット
10012	0x1000000
10011	0x2000000
…	…

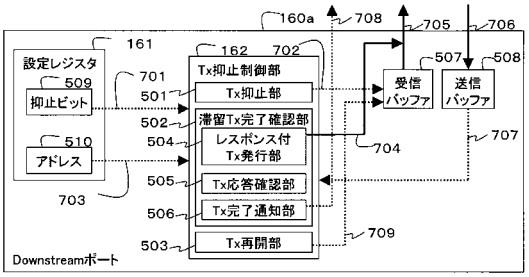
【図５】



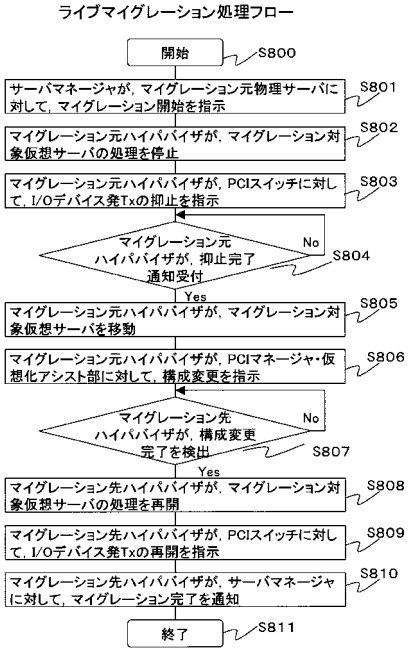
【図６】



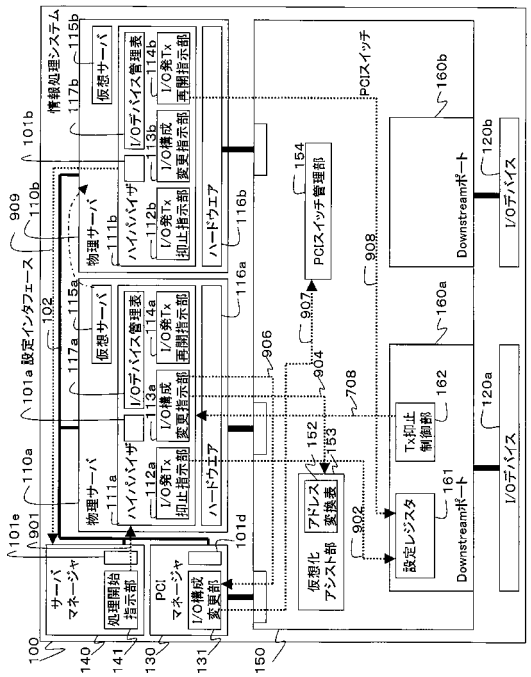
【図７】


160a
708
705
706
161
162
702
設定レジスタ
509
抑止ビット
701
501
510
アドレス
703
502
504
505
506
503
Tx抑止制御部
Tx抑止部
滞留Tx完了確認部
レスポンス付
Tx発行部
Tx応答確認部
Tx完了通知部
Tx再開部
Downstreamポート
507
508
受信
バッファ
送信
バッファ
704
707
709

【図８】


ライブマイグレーション処理フロー
開始
S800
サーバマネージャが，マイグレーション元物理サーバに対して，マイグレーション開始を指示
S801
マイグレーション元ハイパバイザが，マイグレーション対象仮想サーバの処理を停止
S802
マイグレーション元ハイパバイザが，PCIスイッチに対して，I/Oデバイス発Txの抑止を指示
S803
マイグレーション元ハイパバイザが，抑止完了通知受付
No
Yes
S804
マイグレーション元ハイパバイザが，マイグレーション対象仮想サーバを移動
S805
マイグレーション元ハイパバイザが，PCIマネージャ・仮想化アシスト部に対して，構成変更を指示
S806
マイグレーション先ハイパバイザが，構成変更完了を検出
No
Yes
S807
マイグレーション先ハイパバイザが，マイグレーション対象仮想サーバの処理を再開
S808
マイグレーション先ハイパバイザが，PCIスイッチに対して，I/Oデバイス発Txの再開を指示
S809
マイグレーション先ハイパバイザが，サーバマネージャに対して，マイグレーション完了を通知
S810
終了
S811

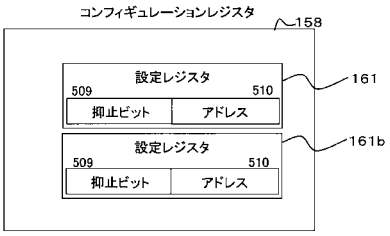
【図９】


情報処理システム
101b
909
110b
101a 設定インタフェース
102
101e
901
110a
101d
100
サーバマネージャ
処理開始指示部
140
141
PCIマネージャ
I/O構成変更部
130
131
150
物理サーバ
111a
仮想サーバ
117a
115a
I/Oデバイス管理表
114a
I/O発Tx再開指示部
113a
I/O構成変更指示部
112a
I/O発Tx抑止指示部
ハイパバイザ
ハードウェア
116a
物理サーバ
111b
仮想サーバ
117b
115b
I/Oデバイス管理表
114b
I/O発Tx再開指示部
113b
I/O構成変更指示部
112b
I/O発Tx抑止指示部
ハイパバイザ
ハードウェア
116b
906
907
904
708
152
153
154
PCIスイッチ管理部
908
PCIスイッチ
160a
160b
162
Tx抑止制御部
902
161
設定レジスタ
Downstreamポート
120a
120b
I/Oデバイス
仮想化アシスト部
アドレス変換表

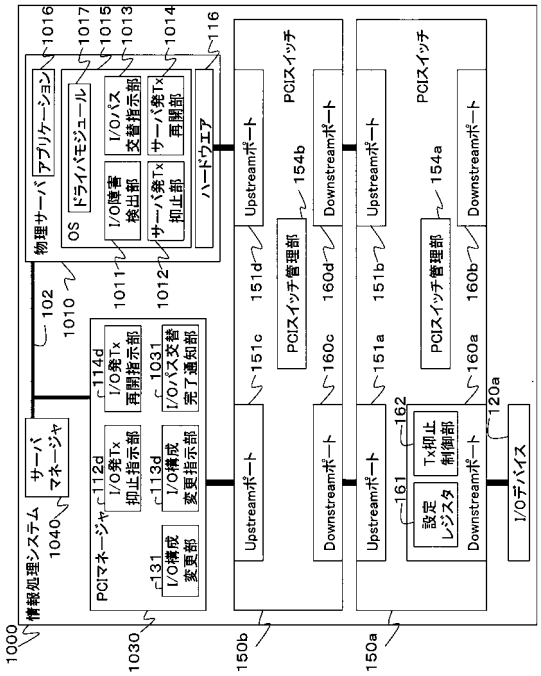
【図１０】


1203
1204
1205
トランザクション送信元識別子
送信先メモリアドレス
トラフィッククラス
ヘッダ
1201
ペイロード
1202

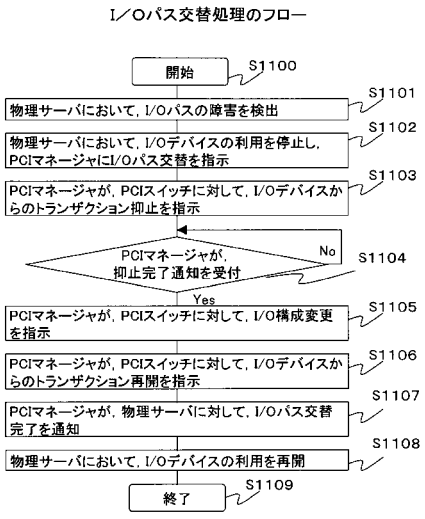
【図１１】


コンフィギュレーションレジスタ
158
設定レジスタ
509
510
抑止ビット
アドレス
161
設定レジスタ
509
510
抑止ビット
アドレス
161b

【図１２】


情報処理システム
1040
物理サーバ
アプリケーション
1016
OS
ドライバモジュール
1017
I/O障害検出部
1015
I/Oパス交替指示部
1013
サーバ発Tx抑止部
1012
サーバ発Tx再開部
1014
ハードウエア
116
102
1010
1011
サーバマネージャ
PCIマネージャ
I/O発Tx再開指示部
114d
I/O発Tx抑止指示部
112d
I/O構成変更指示部
113d
I/Oパス交替完了通知部
1031
I/O構成変更部
131
1030
PCIスイッチ
154b
PCIスイッチ管理部
160d
160c
Upstreamポート
151d
151c
Downstreamポート
Upstreamポート
Downstreamポート
1000
150b
PCIスイッチ
154a
PCIスイッチ管理部
160b
160a
151b
151a
Upstreamポート
Downstreamポート
Upstreamポート
Downstreamポート
162
Tx抑止制御部
161
設定レジスタ
120a
I/Oデバイス
150a

【図１３】


I／Oパス交替処理のフロー
開始
S1100
物理サーバにおいて，I/Oパスの障害を検出
S1101
物理サーバにおいて，I/Oデバイスの利用を停止し，PCIマネージャにI/Oパス交替を指示
S1102
PCIマネージャが，PCIスイッチに対して，I/Oデバイスからのトランザクション抑止を指示
S1103
PCIマネージャが，抑止完了通知を受付
No
S1104
Yes
PCIマネージャが，PCIスイッチに対して，I/O構成変更を指示
S1105
PCIマネージャが，PCIスイッチに対して，I/Oデバイスからのトランザクション再開を指示
S1106
PCIマネージャが，物理サーバに対して，I/Oパス交替完了を通知
S1107
物理サーバにおいて，I/Oデバイスの利用を再開
S1108
終了
S1109

フロントページの続き

(72)発明者　馬場　貴成
東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内
(72)発明者　上原　敬太郎
東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内
(72)発明者　對馬　雄次
東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内

審査官　木村　貴俊

(56)参考文献　米国特許出願公開第２００７／０１８６０２５（ＵＳ，Ａ１）
特開平０６－２９００６７（ＪＰ，Ａ）
特開２００４－０３２２２４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　３／０６－　３／０８、１１／２０
Ｇ０６Ｆ　１３／００－１３／４２