(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2009-282917
(P2009-282917A)
(43)公開日　平成21年12月3日(2009.12.3)

(51)Int.Cl.	F I	テーマコード（参考）
G06F 15/173 (2006.01)	G06F 15/173 650S	5B045

審査請求 未請求　請求項の数 6　ＯＬ　（全 12 頁）

(21)出願番号　特願2008-136943 (P2008-136943)
(22)出願日　平成20年5月26日 (2008.5.26)

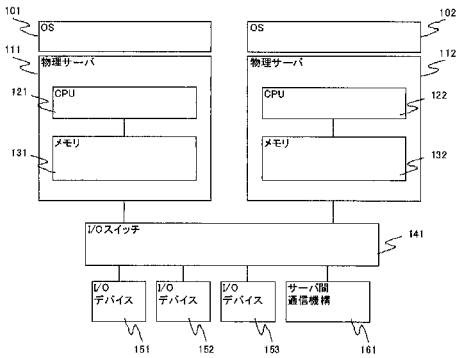
(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110000442
特許業務法人　武和国際特許事務所
(72)発明者　▲高▼瀬　亮
神奈川県秦野市堀山下１番地　株式会社日立製作所エンタープライズサーバ事業部内
(72)発明者　情野　雄太郎
神奈川県秦野市堀山下１番地　株式会社日立製作所エンタープライズサーバ事業部内
(72)発明者　藤原　至誠
神奈川県秦野市堀山下１番地　株式会社日立製作所エンタープライズサーバ事業部内

Ｆターム(参考)　5B045 BB15　BB28　BB34

(54)【発明の名称】サーバ間通信機構及びコンピュータシステム

(57)【要約】
【課題】物理サーバ毎にサーバ間通信を行うための外付けのＩ／Ｏデバイスを用意する必要がなく、プロトコル変換によるオーバヘッドも生じることのないようにする。
【解決手段】サーバ間通信機構161には、I/0リンク201、図示しないI/0スイッチを介して複数の物理サーバに接続されている。サーバ間通信機構161は、各物理サーバのデータをアクセスするための命令を発行する読み出し命令生成部203、読み出したデータを他のサーバに送信する書き込み命令生成部204を有し、サーバ間通信機構161の内部でデータの送信元からデータを読み出し、読み出したデータをそのままデータの送信先に書き出して、サーバ間通信機構161でデータを直接折り返すことにより物理サーバ相互間でのデータ転送を行う。
【選択図】図１

図１



【特許請求の範囲】
【請求項１】
メモリの内容を読み出すための命令を生成する読み出し命令生成手段と、前記読み出し命令の結果として返却されるメモリデータ返却命令を受信する返却命令受信手段と、メモリデータ返却命令と共に返却されるメモリデータをバッファリングするデータバッファと、バッファリングされたメモリデータを書き込むための命令を生成する書き込み命令生成手段と、前記読み出し命令及び前記書き込み命令の宛先情報を付加する宛先情報付加手段とを備え、Ｉ／Ｏスイッチを介して接続されている複数の物理サーバにおける、データ送信元の物理サーバのメモリ上にあるデータを送信先の物理サーバのメモリ上に転送することを特徴とするサーバ間通信機構。
【請求項２】
請求項１記載のサーバ間通信機構において、前記サーバ間通信機構は、制御レジスタを内蔵しており、前記制御レジスタは、複数の物理サーバから共通に読み出し、あるいは、書き込みが行われることを特徴とするサーバ間通信機構。
【請求項３】
請求項１記載のサーバ間通信機構であって、該サーバ間通信機構は、複数のアップストリームポートと１つ以上のダウンストリームポートとを有するＩ／Ｏスイッチに内蔵されていることを特徴とするサーバ間通信機構。
【請求項４】
それぞれが１つ以上のＣＰＵとメモリとを有する複数の物理サーバと、複数のアップストリームポートと１つ以上のダウンストリームポートを持つＩ／Ｏスイッチとからなり、前記複数の物理サーバのそれぞれが前記Ｉ／Ｏスイッチのアップストリームポートを経由して接続されたコンピュータシステムにおいて、
前記Ｉ／Ｏスイッチのダウンストリームポートには、請求項１記載のサーバ間通信機構が接続されていることを特徴とするコンピュータシステム。
【請求項５】
それぞれが１つ以上のＣＰＵとメモリとを有する複数の物理サーバと、複数のアップストリームポートと１つ以上のダウンストリームポートを持つＩ／Ｏスイッチとからなり、前記複数の物理サーバのそれぞれが前記Ｉ／Ｏスイッチのアップストリームポートを経由して接続されたコンピュータシステムにおいて、
前記Ｉ／Ｏスイッチには、請求項１記載のサーバ間通信機構が内蔵されており、前記ダウンストリームポートにはＩ／Ｏデバイスが接続されていることを特徴とするコンピュータシステム。
【請求項６】
それぞれが１つ以上のＣＰＵとメモリとを有する複数の物理サーバと、複数のアップストリームポートと１つ以上のダウンストリームポートを持つＩ／Ｏスイッチとからなり、前記複数の物理サーバのそれぞれが前記Ｉ／Ｏスイッチのアップストリームポートを経由して接続されたコンピュータシステムにおいて、
前記Ｉ／Ｏスイッチが多段に接続され、最上位の前記Ｉ／Ｏスイッチのアップストリームポートに前記物理サーバが接続され、多段に接続された前記Ｉ／Ｏスイッチのそれぞれには、請求項１記載のサーバ間通信機構が内蔵されていることを特徴とするコンピュータシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、サーバ間通信機構及びコンピュータシステムに係り、特に、２台以上の物理サーバがＩ／Ｏスイッチを介して接続されて構成されたコンピュータシステムにおける物理サーバ間の通信を行うためのサーバ間通信機構及びコンピュータシステムに関する。
【背景技術】
【０００２】

近年のコンピュータシステムは、ＣＰＵ単体の高性能化及びＣＰＵのマルチコア化等のＣＰＵの処理性能の向上に伴って、１台の物理サーバとしてのコンピュータ上に複数の仮想サーバを稼働させるサーバ統合のニーズが高まってきている。このようなサーバ統合を行うことにより、１台の物理サーバで稼働するＯＳやアプリケーションの数を増大させることが可能になり、コンピュータシステムの高性能化を図ることができる。その結果、そのようなコンピュータシステムは、１台の物理サーバとしてのコンピュータに接続しなければならないＩ／Ｏデバイスの数が増加していくことが予測される。そして、より多くのＩ／Ｏデバイスを搭載するために、この種のコンピュータシステムは、サーバとＩ／Ｏデバイスとの間をＰＣＩ－Ｅｘｐｒｅｓｓ（Ｒ）スイッチ等のＩ／Ｏスイッチを介して接続して構成する必要性が高まっている。

【０００３】

前述したようなＩ／Ｏスイッチを用いてサーバに接続する物理的なＩ／Ｏデバイスの数を増加させるというアプローチの一方で、物理サーバ相互間あるいは仮想サーバ相互間でＩ／Ｏデバイスを共有するＩ／Ｏ仮想化の普及も予測される。Ｉ／Ｏ仮想化とは、物理的なＩ／Ｏデバイスの上に複数の仮想Ｉ／Ｏデバイスを構築し、各物理サーバあるいは各仮想サーバのそれぞれに仮想Ｉ／Ｏデバイスを割り当てることにより、Ｉ／Ｏデバイスを物理サーバ相互間あるいは仮想サーバ相互間で共有するという方法である。

【０００４】

Ｉ／Ｏデバイスを複数の物理サーバで共有する場合、そのコンピュータシステムは、複数のアップストリームポートと複数のダウンストリームポートとを有するＩ／Ｏスイッチを用意し、このＩ／Ｏスイッチのアップストリームポートのそれぞれに物理サーバを接続し、ダウンストリームポートにＩ／Ｏデバイスを接続する構成とされる。これにより、Ｉ／Ｏデバイスを複数の物理サーバ相互間で共有することができる。このようなＩ／Ｏ仮想化を用いることにより、Ｉ／Ｏデバイスの物理的な数を増やすことなく、統合されたサーバ上で稼働するＯＳやアプリケーションのそれぞれから多くのＩ／Ｏデバイスを使用することが可能になる。

【０００５】

また、サーバ統合を行って１台の物理サーバ上に複数の仮想サーバを稼働させるようにしたコンピュータシステムは、ある物理サーバ上で稼働する仮想サーバを、稼働中に別の物理サーバに集約したり再構成したりしたいという要求がある。仮想サーバをある物理サーバから別の物理サーバに移動させるためには、その仮想サーバが使用しているメモリの内容や稼働状態を、そのまま別の物理サーバに引き継ぐということを意味する。すなわち、仮想サーバをある物理サーバから別の物理サーバに移動させるためには、このような稼働に関するメモリ上の大量のデータを高速に転送することが必要となり、物理サーバ間での高速な通信手段が必要となる。

【０００６】

コンピュータ相互間での高速な通信を行うことを可能とした通信手段に関する従来技術として、例えば、特許文献１等に記載された技術が知られている。この従来技術は、マルチノードコンピュータシステムにおいて、ノード間通信に汎用Ｉ／Ｏインタフェースを用い、各ノードに存在する通信制御装置がデータ転送のための転送コマンドを解釈して、汎用Ｉ／Ｏインタフェースの制御を行ってノード間のデータ転送を高速に行うことを可能にするというものである。

【特許文献１】特開２００６－５８９５６号公報

【発明の開示】

【発明が解決しようとする課題】

【０００７】

前述した従来技術は、物理サーバ相互間の通信に適用した場合、各物理サーバにサーバ間通信用のＩ／Ｏデバイスを接続し、各物理サーバの通信用Ｉ／Ｏデバイス相互間で通信を行うことになり、これにより、サーバ間での高速通信を実現することが可能になる。このような手法を用いて、複数の物理サーバの全てがその相互間で通信を行うためには、各

物理サーバのそれぞれに通信用のＩ／Ｏデバイスを接続する必要がある。
【０００８】
このため、前述した従来技術は、ブレードサーバに代表される物理サーバの集積度が高いサーバに適用した場合に、各物理サーバ間の通信を実現するために必要な通信用のＩ／Ｏデバイスの数が非常に多くなってしまうという問題点を生じる。
【０００９】
また、前述した従来技術は、それを適用し、通信用のＩ／Ｏデバイスを用いてサーバ間通信を実現した場合、物理サーバと通信用Ｉ／Ｏデバイスとの間のインタフェースと、通信用Ｉ／Ｏデバイス相互間のインタフェースとで用いられる通信プロトコルが異なるため、サーバ間通信を行うためにプロトコル変換のためのオーバヘッドが生じ、通信スループットが低下したり、通信レイテンシが大きくなったりするという問題点が生じてしまう。
【００１０】
前述した通信用のＩ／Ｏデバイスの数が非常に多くなるという問題点は、Ｉ／Ｏ仮想化の技術を利用してサーバ間通信用のＩ／Ｏデバイスを物理サーバ間で共有するようにすることにより、Ｉ／Ｏデバイスの数を低減することができ、問題をある程度解決することが可能であるが、プロトコル変換によるオーバヘッドが生じるという問題を解決することができない。
【００１１】
本発明の目的は、前述した従来技術の問題点を解決し、物理サーバ毎にサーバ間通信を行うための外付けのＩ／Ｏデバイスを用意する必要がなく、プロトコル変換によるオーバヘッドが生じることのないサーバ間通信機構及び該サーバ間通信機構を用いたコンピュー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】
【００１２】
本発明によれば前記目的は、メモリの内容を読み出すための命令を生成する読み出し命令生成手段と、前記読み出し命令の結果として返却されるメモリデータ返却命令を受信する返却命令受信手段と、メモリデータ返却命令と共に返却されるメモリデータをバッファリングするデータバッファと、バッファリングされたメモリデータを書き込むための命令を生成する書き込み命令生成手段と、前記読み出し命令及び前記書き込み命令の宛先情報を付加する宛先情報付加手段とを備え、Ｉ／Ｏスイッチを介して接続されている複数の物理サーバにおける、データ送信元の物理サーバのメモリ上にあるデータを送信先の物理サーバのメモリ上に転送することにより達成される。
【発明の効果】
【００１３】
本発明によれば、物理サーバ毎にサーバ間通信を行うための外付けのＩ／Ｏデバイスを用意する必要がなく、プロトコル変換によるオーバヘッドが生じさせることを防止して、通信のスループットを増大させることができ、通信によるレイテンシの増大を防止することができる。
【発明を実施するための最良の形態】
【００１４】
以下、本発明によるサーバ間通信機構及びコンピュータシステムの実施形態を図面により詳細に説明する。
【００１５】
図１は本発明の第１の実施形態によるコンピュータシステムの構成を示すブロック図である。
【００１６】
本発明の第１の実施形態によるコンピュータシステムは、複数の物理サーバ１１１、１１２が、複数のアップストリームポートと複数のダウンストリームポートとを有するＩ／Ｏスイッチ１４１のアップストリームポートに接続され、Ｉ／Ｏスイッチ１４１のダウンストリームポートに複数のＩ／Ｏデバイス１５１～１５３とサーバ間通信機構１６１とが

接続され、物理サーバ１１１、１１２の上でＯＳ１０１、１０２が稼働するように構成されている。物理サーバ１１１は、ＣＰＵ１２１及びメモリ１３１を有して構成され、また、物理サーバ１１２も、同様で、ＣＰＵ１２２及びメモリ１３２を有して構成されている。

【００１７】

すなわち、本発明の第１の実施形態によるコンピュータシステムは、Ｉ／Ｏスイッチ１４１を経由して物理サーバ１１１、１１２と、Ｉ／Ｏデバイス１５１～１５３とが接続されて構成されている。Ｉ／Ｏデバイス１５１～１５３のそれぞれは、物理サーバ１１１と１１２との両者で共有して使用されるＩ／Ｏデバイスであっても、物理サーバ１１１または１１２のどちらかで占有して使用されるＩ／Ｏデバイスであってもよい。また、Ｉ／Ｏスイッチ１４１のダウンストリームポートには、本発明によるサーバ相互間の通信に利用するサーバ間通信機構１６１が接続されており、Ｉ／Ｏスイッチ１４１を経由して物理サーバ１１１及び１１２と接続されている。

【００１８】

前述において、本発明の実施形態では、ＯＳが稼働している物理サーバを２台、サーバ間通信機構を１台備えるとして示しているが、本発明は、物理サーバがさらに多数設けられてもよく、また、サーバ間通信機構を２台以上設けてもよい。これにより、２台のサーバの組でサーバ相互間の通信を行っている場合にも、他の２台のサーバの組でサーバ相互間の通信を並行して行うことが可能となる。

【００１９】

図２はサーバ間通信機構１６１の構成を示すブロック図である。サーバ間通信機構１６１は、Ｉ／Ｏリンク２０１によってＩ／Ｏスイッチ１４１と接続されている。Ｉ／Ｏリンク２０１とサーバ間通信機構１６１との内部は、Ｉ／Ｏインタフェース２０２により接続されている。

【００２０】

サーバ間通信機構１６１は、物理サーバ内のメモリデータを取り込むためのメモリ読み出し命令生成部２０３と、メモリデータを他の物理サーバに送り出すためのメモリ書き込み命令生成部２０４と、割り込みを発生させるための割り込み命令生成部２０５とを有している。メモリ読み出し命令生成部２０３、メモリ書き込み命令生成部２０４、割り込み命令生成部２０５には、命令を正しい宛先の物理サーバに送出するための宛先情報付加機構として、それぞれ送信側サーバ宛先情報付加部２０６と受信側サーバ宛先情報付加部２０７とが接続されている。また、サーバ間通信機構１６１は、メモリ読み出し命令生成部２０３、メモリ書き込み命令生成部２０４及び割り込み命令生成部２０５の動作を制御するために命令発行用シーケンサ２０８を備えている。また、サーバ間通信機構１６１は、メモリ読み出し命令により、物理サーバから読み出されたデータを取り込むためのメモリデータ返却命令受信部２０９を備えると共に、ここで受信したメモリデータを格納するためのメモリデータバッファ２１０を備えている。さらに、サーバ間通信機構１６１は、ソフトウェアからサーバ間通信機構を制御するための機構として、サーバ間通信機構レジスタ２１１を備えている。このサーバ間通信機構レジスタ２１１には、送信メモリアドレスレジスタ２１２、受信メモリアドレスレジスタ２１３、送信メモリ領域長レジスタ２１４、起動レジスタ２１５が含まれている。

【００２１】

図３はサーバ間通信機構が物理サーバ相互間でメモリデータの転送を行う制御を説明するシーケンスチャートであり、次に、これについて説明する。ここでのメモリデータの転送は、図１に示すコンピュータシステムの例における物理サーバ１１１内のメモリ１３１に存在するデータを、物理サーバ１１２内のメモリ１３２に転送する場合を例としている。

【００２２】

（１）まず、データ送信側となる物理サーバ１１１上で稼働しているＯＳ１０１が、送信するメモリ領域の先頭アドレスを設定する。この設定は、データ送信側となる物理サー

バ１１１からサーバ間通信機構１６１の送信メモリアドレスレジスタ２１２に対して書き込み命令を送信することにより行われる（ステップ３０１）。

【００２３】

（２）同様に、データ送信側となる物理サーバ１１１上で稼働しているＯＳ１０１は、サーバ間通信機構１６１の送信メモリ領域長レジスタ２１４に対して書き込みを行い、送信するメモリ領域の大きさを設定する（ステップ３０２）。

【００２４】

（３）また、同様に、データ受信側となる物理サーバ１１２上で稼働しているＯＳ１０２が、サーバ間通信機構１６１の受信メモリアドレスレジスタ２１３に対して書き込みを行い、受信するメモリ領域の先頭アドレスを設定する（ステップ３０３）。

【００２５】

前述までの処理でサーバ間通信を行うための初期設定処理が完了する。

【００２６】

（４）次に、データ送信側となる物理サーバ１１１上で稼働しているＯＳ１０１が、サーバ間通信機構１６１の起動レジスタ２１５に書き込みを行うことにより、サーバ間通信機構１６１を起動させる（ステップ３０４）。

【００２７】

（５）サーバ間通信機構１６１は、起動されると命令発行用シーケンサ２０８が動作を開始し、メモリ読み出し命令生成部２０３から送信側サーバ宛先情報付加部２０６を経由してデータ送信側の物理サーバ１１１の宛先情報を付加し、メモリ読み出し命令を発行する。このメモリ読み出し命令は、Ｉ／Ｏスイッチ１４１を通過する際に前記の送信側サーバ宛先情報を用いてデータ送信側の物理サーバ１１１へと正しく送信される（ステップ３０５ａ）。

【００２８】

（６）メモリ読み出し命令を受け取ったデータ送信側の物理サーバ１１１は、サーバ間通信機構１６１に対して、送信するメモリデータを含むデータ返却命令を送信することにより、ステップ３０５ａでのメモリ読み出し命令に対するメモリデータ返却を行う。なお、１回のデータ返却命令で送信するメモリデータの量は、予め定められた量であり、送信すべきメモリ領域内のデータ量が大きい場合には、複数回に分けて送信される（ステップ３０６ａ）。

【００２９】

（７）サーバ間通信機構１６１は、データ返却命令をメモリデータ返却命令受信部２０９で受信し、メモリデータ部分をメモリデータバッファ２１０に格納する。そして、サーバ間通信機構１６１は、メモリデータを受け取ると、メモリ書き込み命令生成部２０４がメモリデータバッファ２１０からメモリデータを取り出し、受信側サーバ宛先情報付加部２０７を経由してデータ受信側の物理サーバ１１２の宛先情報を付加し、転送すべきメモリデータを含むメモリ書き込み命令を発行する。このメモリ書き込み命令は、Ｉ／Ｏスイッチ１４１を通過する際に前記の受信側サーバ宛先情報を用いてデータ受信側の物理サーバ１１２へと正しく送信される（ステップ３０７ａ）。

【００３０】

（８）命令発行用シーケンサ２０８は、前述したステップ３０５ａでのメモリ読み出し命令の送信、ステップ３０６ａでのメモリ返却命令の受信、ステップ３０７ａでのメモリ書き込み命令の送信の一連の動作を、送信メモリ領域長レジスタ２１４で指定されたデータ長を転送し終えるまで繰り返すことにより、データ送信側の物理サーバ１１１からデータ受信側の物理サーバ１１２へデータの転送を行う（ステップ３０５ｂ、３０６ｂ、３０７ｂ）。

【００３１】

前述したステップ３０５ｂ、３０６ｂ、３０７ｂの処理を指定されたデータ長を転送し終えるまで繰り返すことにより、データ送受信の処理が完了する。

【００３２】

（９）データの転送が終了すると、サーバ間通信機構１６１の命令発行用シーケンサ２０８は、割り込み命令生成部２０５を起動する。割り込み命令生成部２０５は、データ送信側の物理サーバ１１１及びデータ受信側の物理サーバ１１２のそれぞれに対して、データ送信完了を通知する割り込み命令を発行する。割り込み命令生成部２０５で生成された割り込み命令は、送信側サーバ宛先情報付加部２０６あるいは受信側サーバ宛先情報付加部２０７を経由して正しい宛先情報が付加され、それぞれデータ送信側の物理サーバ１１１、データ受信側の物理サーバ１１２へ送信される（ステップ３０８、３０９）。

【００３３】

前述したステップ３０８、３０９の処理を行うことにより、サーバ間でのデータ転送処理の全てが完了する。

【００３４】

前述した本発明の第１の実施形態では、サーバ間通信機構１６１に設定指示をしたり、サーバ間通信機構を起動したりする動作主体をＯＳであるとして説明したが、動作主体は、サーバ間通信機構用のデバイスドライバであっても、アプリケーションプログラム等であっても、仮想サーバを管理するハイパバイザ等であってもよい。

【００３５】

また、図３に示す処理を開始する契機となるのは、図１に示していないが、物理サーバ１１１、１１２に接続されている管理用のコンピュータからの指示であっても、また、物理サーバ１１１、１１２の中に構築されているサービスプロセッサからの指示であってもよい。また、この指示は、管理用のコンピュータ、サービスプロセッサが、物理サーバの状態から自動的に行われるものであっても、システム管理者等から行われるものであってもよい。

【００３６】

図４は本発明の第２の実施形態によるコンピュータシステムの構成を示すブロック図である。図４に示す本発明の第２の実施形態は、Ｉ／Ｏスイッチ内にサーバ間通信機構を設けるようにした例である。

【００３７】

図１～図３により説明した本発明の第１の実施形態は、サーバ間通信機構１６１を外付けのＩ／Ｏデバイスとして、Ｉ／Ｏスイッチ１４１のダウンストリームポートに接続するものとして説明したが、本発明の第２の実施形態は、図４に示すように、Ｉ／Ｏスイッチにサーバ間通信機構４２１を内蔵させて通信機構内蔵Ｉ／Ｏスイッチ４１１として構成し、物理サーバ４０１及び４０２相互間のデータ転送を行うようにしたものである。

【００３８】

本発明の第２の実施形態は、このように構成することにより、通信機構内蔵Ｉ／Ｏスイッチ４１１内のサーバ間通信機構４２１により、図３により説明した場合と同様に、物理サーバ４０１、４０２相互間の通信を行うことが可能となる。

【００３９】

前述した本発明の第２の実施形態は、サーバ間通信機構４２１を接続するためにのために、Ｉ／Ｏスイッチのダウンストリームスロットを占有する必要がなくなり、他のデバイスのためにＩ／Ｏスイッチのダウンストリームスロットを開放することができるという利点を得ることができる。また、サーバ間通信機構４２１を外付けのデバイスとして用意する必要がないため、サーバ間通信を導入するコストをより低減することができるという利点を得ることができる。

【００４０】

図５は本発明の第３の実施形態によるコンピュータシステムの構成を示すブロック図である。図５に示す本発明の第３の実施形態は、物理サーバに接続するＩ／Ｏデバイスの数を増加するために、通信機構内蔵Ｉ／Ｏスイッチを多段に設けて構成した場合のコンピュータシステムの例である。

【００４１】

図５に示す本発明の第３の実施形態によるコンピュータシステムは、複数の物理サーバ

５０１、５０２が１段目の通信機構内蔵Ｉ／Ｏスイッチのアップストリームポートに接続され、複数の２段目の通信機構内蔵Ｉ／Ｏスイッチ５１２、５１３が１段目の通信機構内蔵Ｉ／Ｏスイッチのダウンストリームポートに接続され、２段目の通信機構内蔵Ｉ／Ｏスイッチ５１２、５１３のそれぞれの複数のダウンストリームポートにＩ／Ｏデバイスが接続されて構成されている。そして、物理サーバ５０１と５０２との間のサーバ間通信は、初段の通信機構内蔵Ｉ／Ｏスイッチ５１１に内蔵したサーバ間通信機構５２１を用いて行うことができる。

【００４２】

前述の本発明の第３の実施形態によれば、初段のＩ／Ｏスイッチ内のサーバ間通信機構を用いてサーバ間通信を行うことができるので、図６により後述するＩ／Ｏスイッチにサーバ間通信機構を内蔵していないＩ／Ｏスイッチを多段に設けて構成した例に比較して、サーバ間通信に要する通信レイテンシを低減することができる。

【００４３】

図５に示す本発明の第３の実施形態によるコンピュータシステムは、初段の通信機構内蔵Ｉ／Ｏスイッチ５１１を１台だけ備えて構成されているが、この例では、初段の通信機構内蔵Ｉ／Ｏスイッチ５１１を複数台設けることも可能であり、また、２段目の通信機構内蔵Ｉ／Ｏスイッチ５１２、５１３を３台以上設けて構成することができる。このように構成された本発明の第３の実施形態は、２段目の通信機構内蔵Ｉ／Ｏスイッチの１つに複数の初段の通信機構内蔵Ｉ／Ｏスイッチを接続するような構成をとることにより、異なる初段の通信機構内蔵Ｉ／Ｏスイッチ内のサーバ間通信機構を用いては通信を行うことができない物理サーバ間の通信を、２段目の通信機構内蔵Ｉ／Ｏスイッチに内蔵したサーバ間通信機構を用いて行うことができる。

【００４４】

図６は本発明の第４の実施形態によるコンピュータシステムの構成を示すブロック図である。図６に示す本発明の第４の実施形態は、物理サーバに接続するＩ／Ｏデバイスの数を増加させるために、Ｉ／Ｏスイッチとして、サーバ間通信機構を内蔵していないＩ／Ｏスイッチを多段（図示例では２段）に設け、２段目のＩ／Ｏスイッチにサーバ間通信機構を接続して構成した例である。

【００４５】

すなわち、図６に示す例は、複数の物理サーバ５０１、５０２を１段目のＩ／Ｏスイッチ６１１に接続し、このＩ／Ｏスイッチ６１１に複数の２段目のＩ／Ｏスイッチ６１２、６１３を接続し、２段目のＩ／Ｏスイッチ６１２、６１３にＩ／Ｏデバイスとサーバ間通信機構を接続して構成されている。

【００４６】

図６に示す本発明の第４の実施形態によるコンピュータシステムは、Ｉ／Ｏスイッチにサーバ間通信機構を内蔵していないＩ／Ｏスイッチを多段に設けて構成しているため、物理サーバ５０１と５０２との間のサーバ間通信は、２段目のＩ／Ｏスイッチ６１２に接続されたサーバ間通信機構６２２で行う必要がある。なお、図６には示していないが、２段目のＩ／Ｏスイッチ６１３にもサーバ間通信機構を接続することができる。

【００４７】

前述した本発明の各実施形態は、サーバ間通信機構によるデータ送受信の処理の完了を、割り込み命令生成部からデータ送信側の物理サーバ及びデータ受信側の物理サーバに割り込み命令を発行するすることにより、各物理サーバに通知するものであったが、本発明は、データ送信処理の完了を他の方法により、各物理サーバに通知するようにすることもできる。

【００４８】

図７はサーバ間通信機構が物理サーバ相互間でメモリデータの転送を制御する他の例を説明するシーケンスチャートであり、次に、これについて説明する。ここでのメモリデータの転送は、図３に示して説明した場合と同様に、図１に示すコンピュータシステムの例における物理サーバ１１１内のメモリ１３１に存在するデータを、物理サーバ１１２内の

メモリ１３２に転送する場合を例としている。

【００４９】

そして、この例においては、サーバ間通信機構１６１内のサーバ間通信機構レジスタ２１１に、データ送信側の物理サーバからもデータ受信側の物理サーバからも共通して読み出すことが可能な完了ステータスレジスタを設け、データ送受信処理を終了すると、サーバ間通信機構１６１が、サーバ間通信機構レジスタ２１１内に設けた完了ステータスレジスタに完了を登録するようにし、物理サーバが、完了ステータスレジスタへのポーリングにより、完了ステータスレジスタを読み出すことにより、送信側及び受信側の物理サーバがデータの転送の終了を知るようにコンピュータシステムが構成される。

【００５０】

図７に示すシーケンスにおけるステップ３０１～３０７ｂまでの処理は、図３に示して説明したと同一であるので、ここまでの説明については省略する。

【００５１】

（１）ステップ３０１～３０７ｂまでの処理でデータの送受信処理が完了すると、サーバ間通信機構１６１が、サーバ間通信機構レジスタ２１１内に設けた完了ステータスレジスタに完了を登録する。その後、サーバ間通信機構１６１は、データ送信側物理サーバ１１１から完了ステータスレジスタの読み出し命令を受け取って、完了ステータスレジスタの内容をデータ送信側物理サーバ１１１に返却する（ステップ７０１、７０２）。

【００５２】

（２）同様に、サーバ間通信機構１６１は、データ受信側物理サーバ１１２からも完了ステータスレジスタの読み出し命令を受け取って、完了ステータスレジスタの内容をデータ受信側物理サーバ１１２に返却する（ステップ７０３、７０４）。

【００５３】

前述では、完了ステータスレジスタを用いる例を説明したが、本発明は、サーバ通信機構１６１内の単一のレジスタに対して、データ送信側物理サーバ１１１からもデータ受信側物理サーバ１１２からも同様に読み出しまたは書き込みといったアクセスを可能とすることにより、サーバ間通信機構で単一のレジスタの状態を変更することにより、データ送信側物理サーバ及びデータ受信側物理サーバの両者にその状態を通知するようにすることができる。

【００５４】

前述した本発明の実施形態での各処理は、プログラムにより構成し、本発明が備えるサーバ間通信機構に実行させることができ、また、それらのプログラムは、ＦＤ、ＣＤＲＯＭ、ＤＶＤ等の記録媒体に格納して提供することができ、また、ネットワークを介してディジタル情報により提供することができる。

【００５５】

前述した本発明の実施形態によれば、物理サーバ間のデータ転送を行う際に、従来は物理サーバ毎に用意していたサーバ間通信用の外付けＩ／Ｏデバイスの数を削減することができ、また、サーバ間通信機構をＩ／Ｏスイッチに内蔵することにより、サーバ間通信用の外付けＩ／Ｏデバイスを不要とすることができる。

【００５６】

また、本発明の実施形態によれば、物理サーバ間のデータ通信は、サーバ間通信機構の内部で行うことができるため、プロトコル変換のオーバヘッドが生じることがなく、通信スループットやレイテンシの面で有利となる。

【００５７】

さらに、本発明の実施形態によれば、サーバ間通信機構をＩ／Ｏスイッチに内蔵することにより、物理サーバ毎にサーバ間通信を行うための外付けのＩ／Ｏデバイスを用意する必要がなく、サーバ間通信を行うための導入コストを抑えることができる。また、Ｉ／Ｏスイッチのスロットがサーバ間通信用のＩ／Ｏデバイスで占有されることがなく、Ｉ／Ｏスロットを有効に活用することが可能となる。また、Ｉ／Ｏスイッチを多段の構成にしてＩ／Ｏデバイスの数を増加させたシステムにおいて、中継段のＩ／Ｏスイッチにもサーバ

間通信機構を内蔵させることができるため、中継段のＩ／Ｏスイッチに接続された物理サーバ間のサーバ間通信は、中継段のＩ／Ｏスイッチ内蔵のサーバ間通信機構で折り返すことが可能であり、通信レイテンシをより低減することができる。

【００５８】

また、本発明の実施形態によれば、アドレス変換を必要とせずに大量のメモリデータをサーバ間で転送するような用途で特に効果がある。例えば、仮想サーバのメモリイメージは、大容量であり、かつ、物理アドレスで連続領域であることがある。このような場合に、ハイパバイザが本発明によるサーバ間通信機構を使用して、仮想サーバのマイグレーションを物理サーバ間で行うといった用途に本発明を適用して大きな効果を得ることができる。

【図面の簡単な説明】

【００５９】

【図１】本発明の第１の実施形態によるコンピュータシステムの構成を示すブロック図である。

【図２】サーバ間通信機構の構成を示すブロック図である。

【図３】サーバ間通信機構が物理サーバ相互間でメモリデータの転送を行う制御を説明するシーケンスチャートである。

【図４】本発明の第２の実施形態によるコンピュータシステムの構成を示すブロック図である。

【図５】本発明の第３の実施形態によるコンピュータシステムの構成を示すブロック図である。

【図６】本発明の第４の実施形態によるコンピュータシステムの構成を示すブロック図である。

【図７】サーバ間通信機構が物理サーバ相互間でメモリデータの転送を制御する他の例を説明するシーケンスチャートである。

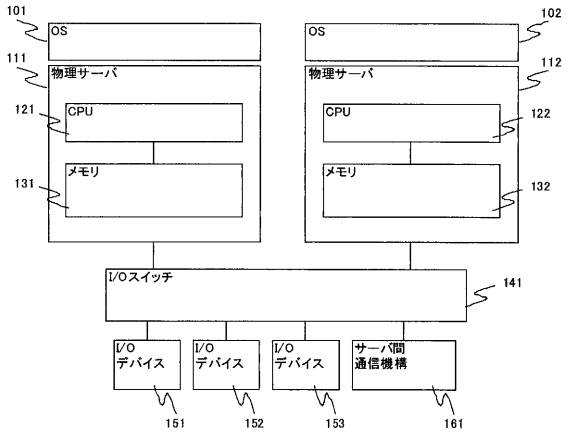
【符号の説明】

【００６０】

１０１、１０２　ＯＳ
１１１、１１２、４０１、４０２、５０１、５０２　物理サーバ
１２１、１２２　ＣＰＵ
１３１、１３２　メモリ
１４１、６１１、６１２、６１３　Ｉ／Ｏスイッチ
１５１～１５３　Ｉ／Ｏデバイス
１６１、４２１、５１１、５２２、５２３、６２２　サーバ間通信機構
２０１　Ｉ／Ｏリンク
２０２　Ｉ／Ｏインタフェース
２０３　メモリ読み出し命令生成部
２０４　メモリ書き込み命令生成部
２０５　割り込み命令生成部
２０６　送信側サーバ宛先情報付加部
２０７　受信側サーバ宛先情報付加部
２０８　命令発行用シーケンサ
２０９　メモリデータ返却命令受信部
２１０　メモリデータバッファ
２１１　サーバ間通信機構レジスタ
２１２　送信メモリアドレスレジスタ
２１３　受信メモリアドレスレジスタ
２１４　送信メモリ領域長レジスタ
２１５　起動レジスタ
４１１、５１１、５１２、５１３　通信機構内蔵Ｉ／Ｏスイッチ

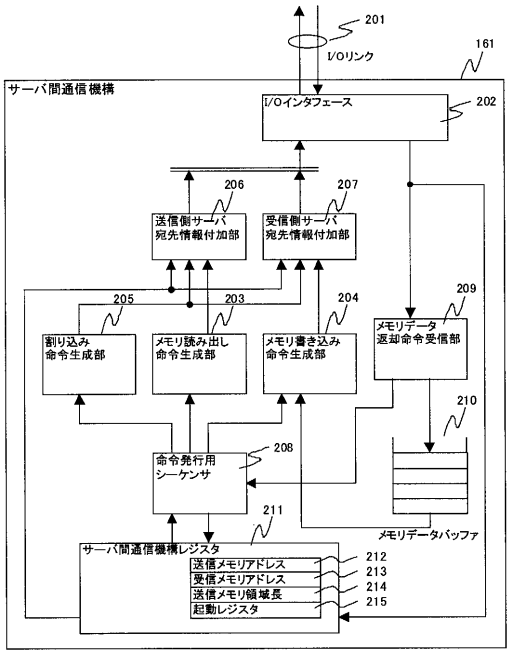
【図１】

図１



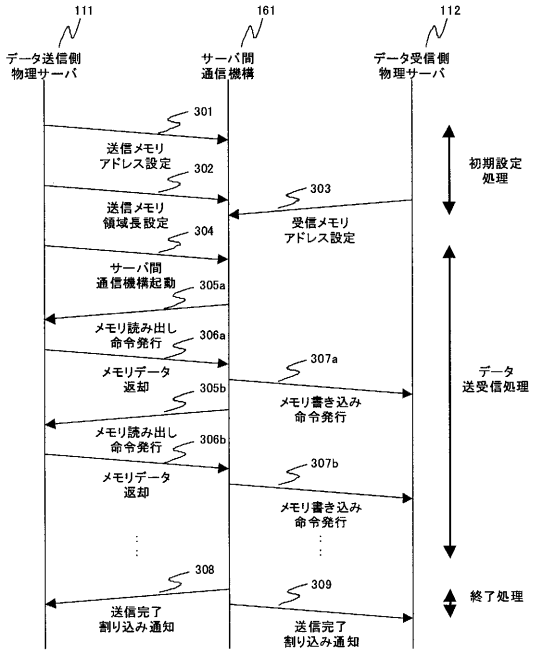
【図２】

図２



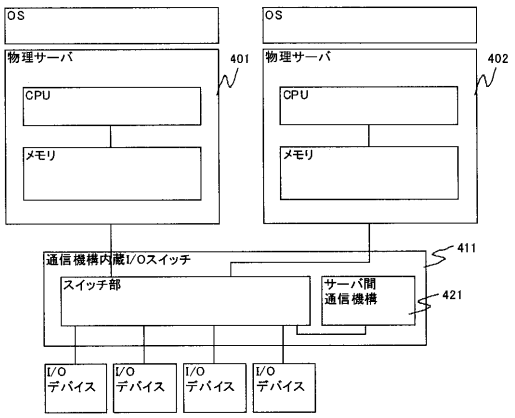
【図３】

図３



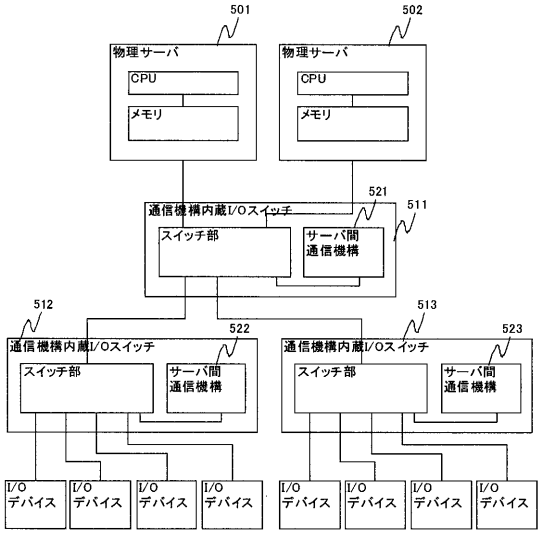
【図４】

図４



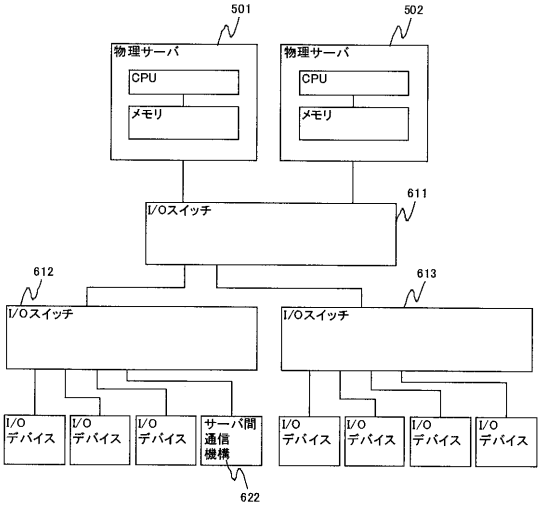
【図５】

図５



【図６】

図６



【図７】

図７