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

特許第5869135号
(P5869135)

(45)発行日　平成28年2月24日(2016.2.24)　(24)登録日　平成28年1月15日(2016.1.15)

(51)Int.Cl.		F I		
G06F 13/36	(2006.01)	G06F	13/36	310E
G06F 13/38	(2006.01)	G06F	13/38	340A
G06F 13/28	(2006.01)	G06F	13/38	350
		G06F	13/28	310Z

請求項の数 20　(全 16 頁)

(21)出願番号	特願2014-533274 (P2014-533274)
(86)(22)出願日	平成23年9月30日 (2011.9.30)
(65)公表番号	特表2014-531685 (P2014-531685A)
(43)公表日	平成26年11月27日 (2014.11.27)
(86)国際出願番号	PCT/US2011/054394
(87)国際公開番号	WO2013/048477
(87)国際公開日	平成25年4月4日 (2013.4.4)
審査請求日	平成26年5月13日 (2014.5.13)

(73)特許権者　593096712
インテル　コーポレイション
アメリカ合衆国　９５０５４　カリフォルニア州　サンタ　クララ　ミッション　カレッジ　ブールバード　２２００
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(74)代理人　100091214
弁理士　大貫　進介
(72)発明者　マーゴ，ウィリアム，アール．
アメリカ合衆国　５１８２２－９２２８
イリノイ州　シャンペーン　コブルフィールド　ロード　１６０３

最終頁に続く

(54)【発明の名称】コプロセッサのためのダイレクトＩ／Ｏアクセス

(57)【特許請求の範囲】

【請求項１】

リモートダイレクトメモリーアクセス（ＲＤＭＡ）デバイスと；

中央処理装置（ＣＰＵ）とメモリーを有する周辺機器と；

前記ＲＤＭＡデバイスと前記周辺機器に対して通信可能に接続されたホストコントローラーと、を含み、

前記ホストコントローラーは、

前記ホストコントローラーの中に含まれ、前記周辺機器のメモリー又は前記ＲＤＭＡデバイスのメモリーのうちの一つに対してマップされた仮想アドレスにおけるアクセスリクエストメッセージに係る通知を受け取り；

前記ＲＤＭＡデバイスの前記メモリーに対してマップされた仮想アドレスにおける前記アクセスリクエストメッセージの受け取りに応じて、前記周辺機器からの前記ＲＤＭＡデバイスのメモリーに対する前記アクセスリクエストメッセージに基づき、ＲＤＭＡアクセスリクエストを前記ＲＤＭＡデバイスに対して伝送し、前記ＲＤＭＡアクセスリクエストメッセージは仮想アドレスＲＤＭＡパラメーターを含み；かつ

前記周辺機器の前記メモリーに対してマップされた仮想アドレスにおける前記アクセスリクエストメッセージの受け取りに応じて、アクセスリクエストメッセージを前記周辺機器の前記メモリーに対して伝送する、

ように構成されており、

前記ＲＤＭＡデバイスは、リクエストの完了後、前記周辺機器のメモリーにマップされ

た仮想アドレスに対して完了メッセージを送付し、前記仮想アドレスは、前記完了メッセージを保管するための完了キューを表している、

ことを特徴とするシステム。

【請求項２】

前記ＲＤＭＡアクセスリクエストは、ＲＤＭＡ読み出しリクエストを有し、

前記ホストコントローラーは、さらに；

前記ＲＤＭＡデバイスからの前記周辺機器によりリクエストされたデータを受け取り、かつ、

前記周辺機器の前記メモリーに対して、前記データを含んでいるデータメッセージを伝送する、

請求項１に記載のシステム。

【請求項３】

前記ＲＤＭＡアクセスリクエストは、ＲＤＭＡ書き込みリクエストを有し、

前記ホストコントローラーは、さらに；

前記周辺機器のメモリーからの前記ＲＤＭＡデバイスによりリクエストされたデータを受け取り、かつ、

前記ＲＤＭＡデバイスに対して、前記データを含んでいるデータメッセージを伝送する、

請求項１に記載のシステム。

【請求項４】

前記周辺機器の前記ＣＰＵは、複数のプロセッサコアのうちの一つ、または、前記周辺機器の中に含まれているマルチコアプロセッサを有する、

請求項１に記載のシステム。

【請求項５】

前記周辺機器と前記ＲＤＭＡデバイスは、周辺コンポーネント相互接続エクスプレス（ＰＣＩｅ）リンクを介して、前記ホストコントローラーに対して通信可能に接続されている、

請求項１に記載のシステム。

【請求項６】

前記ＲＤＭＡデバイスと前記ホストコントローラーは、ＰＣＩｅルートポートを介して通信可能に接続されている、

請求項５に記載のシステム。

【請求項７】

前記周辺機器は、グラフィックサブシステムを有し、かつ

前記ＣＰＵは、グラフィックプロセッサユニット（ＧＰＵ）を有する、

請求項１に記載のシステム。

【請求項８】

前記アクセスリクエストメッセージは、インフィニバンド仕様に準じている、

請求項１に記載のシステム。

【請求項９】

前記アクセスリクエストメッセージは、ＲＤＭＡコンソーシアム仕様に準じている、

請求項１に記載のシステム。

【請求項１０】

リモートダイレクトメモリーアクセス（ＲＤＭＡ）デバイスのメモリーと周辺機器のメモリーのうちの一つに対してマップされた仮想アドレスにおけるアクセスリクエストメッセージに係る通知を受け取るステップであり、前記周辺機器は、さらに、中央処理装置（ＣＰＵ）とメモリーを有するステップと；

前記ＲＤＭＡデバイスの前記メモリーに対してマップされた仮想アドレスにおける前記アクセスリクエストメッセージの受け取りに応じて：

ＲＤＭＡパラメーターとして前記ＲＤＭＡデバイスの前記メモリーに対してマップさ

れた前記仮想アドレスを利用するステップと；
　前記周辺機器からの前記ＲＤＭＡデバイスのメモリーに対する前記アクセスリクエストメッセージに基づき、ＲＤＭＡアクセスリクエストを前記ＲＤＭＡデバイスに対して伝送するステップであり、前記ＲＤＭＡアクセスリクエストメッセージは仮想アドレスＲＤＭＡパラメーターを含んでいるステップと；
　前記周辺機器の前記メモリーに対してマップされた仮想アドレスにおける前記アクセスリクエストメッセージの受け取りに応じて、
　アクセスリクエストメッセージを前記周辺機器の前記メモリーに対して伝送するステップと、を含み、
　前記ＲＤＭＡデバイスは、リクエストの完了後、前記周辺機器のメモリーにマップされた仮想アドレスに対して完了メッセージを送付し、前記仮想アドレスは、前記完了メッセージを保管するための完了キューを表している、
　ことを特徴とする方法。
【請求項１１】
　前記ＲＤＭＡアクセスリクエストは、ＲＤＭＡ読み出しリクエストを有し、
　前記方法は、さらに；
　前記ＲＤＭＡデバイスからの前記周辺機器によりリクエストされたデータを受け取るステップと、
　前記周辺機器の前記メモリーに対して、前記データを含んでいるデータメッセージを伝送するステップと、を含む、
　請求項１０に記載の方法。
【請求項１２】
　前記ＲＤＭＡアクセスリクエストは、ＲＤＭＡ書き込みリクエストを有し、
　前記方法は、さらに；
　前記周辺機器のメモリーからの前記ＲＤＭＡデバイスによりリクエストされたデータを受け取るステップと、
　前記ＲＤＭＡデバイスに対して、前記データを含んでいるデータメッセージを伝送するステップと、を含む、
　請求項１０に記載の方法。
【請求項１３】
　前記周辺機器の前記ＣＰＵは、複数のプロセッサコアのうちの一つ、または、前記周辺機器の中に含まれているマルチコアプロセッサを有する、
　請求項１０に記載の方法。
【請求項１４】
　前記周辺機器と前記ＲＤＭＡデバイスは、周辺コンポーネント相互接続エクスプレス（ＰＣＩｅ）リンクを介して、ホストコントローラーに対して通信可能に接続されている、
　請求項１０に記載の方法。
【請求項１５】
　前記ＲＤＭＡデバイスと前記ホストコントローラーは、ＰＣＩｅルートポートを介して通信可能に接続されている、
　請求項１４に記載の方法。
【請求項１６】
　前記周辺機器は、グラフィックサブシステムを有し、かつ
　前記ＣＰＵは、グラフィックプロセッサユニット（ＧＰＵ）を有する、
　請求項１０に記載の方法。
【請求項１７】
　リモートダイレクトメモリーアクセス（ＲＤＭＡ）デバイスに対して通信可能に接続された第１の相互接続リンクと；
　中央処理装置（ＣＰＵ）とメモリーを有する周辺機器に対して通信可能に接続された第２の相互接続リンクと；

ホストコントローラーと、を含み、
前記ホストコントローラーは、
前記ホストコントローラーの中に含まれ、前記周辺機器のメモリー又は前記ＲＤＭＡデバイスのメモリーのうちの一つに対してマップされた仮想アドレスにおけるアクセスリクエストメッセージに係る通知を受け取り；
前記ＲＤＭＡデバイスの前記メモリーに対してマップされた仮想アドレスにおける前記アクセスリクエストメッセージの受け取りに応じて、前記周辺機器からの前記ＲＤＭＡデバイスのメモリーに対する前記アクセスリクエストメッセージに基づき、ＲＤＭＡアクセスリクエストを前記ＲＤＭＡデバイスに対して伝送し、前記ＲＤＭＡアクセスリクエストメッセージは仮想アドレスＲＤＭＡパラメーターを含み；かつ
前記周辺機器の前記メモリーに対してマップされた仮想アドレスにおける前記アクセスリクエストメッセージの受け取りに応じて、アクセスリクエストメッセージを前記周辺機器の前記メモリーに対して伝送し、
前記ＲＤＭＡデバイスは、リクエストの完了後、前記周辺機器のメモリーにマップされた仮想アドレスに対して完了メッセージを送付し、前記仮想アドレスは、前記完了メッセージを保管するための完了キューを表している、
ことを特徴とする装置。
【請求項１８】
前記ＲＤＭＡアクセスリクエストは、ＲＤＭＡ読み出しリクエストを有し、
前記ホストコントローラーは、さらに；
前記ＲＤＭＡデバイスからの前記周辺機器によりリクエストされたデータを受け取り、かつ、
前記周辺機器の前記メモリーに対して、前記データを含んでいるデータメッセージを伝送する、
請求項１７に記載の装置。
【請求項１９】
前記ＲＤＭＡアクセスリクエストは、ＲＤＭＡ書き込みリクエストを有し、
前記ホストコントローラーは、さらに；
前記周辺機器のメモリーからの前記ＲＤＭＡデバイスによりリクエストされたデータを受け取り、かつ、
前記ＲＤＭＡデバイスに対して、前記データを含んでいるデータメッセージを伝送する、
請求項１７に記載の装置。
【請求項２０】
前記第１および第２の相互接続リンクのそれぞれは、周辺コンポーネント相互接続エクスプレス（ＰＣＩｅ）リンクを有しており、かつ、
前記装置は、さらに、前記ＲＤＭＡデバイスに対して通信可能に接続されたルートポートを含んでいる、
請求項１７に記載の装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施例は、一般的に、コンピューターデバイスに適用される。より特定的には、周辺機器のためのメモリーアクセス管理に関する。
【背景技術】
【０００２】
コンピューターシステムは、ネットワーク又はバス構成といった相互接続構造を介してお互いに接続されている種々のデバイスを含んでいる。これらのデバイスは、典型的にはローカルメモリーを有しており、コンピューター環境の中で処理速度と適応性を提供するために複数のデバイスが並行して動作される。

【０００３】
リモートダイレクトメモリーアクセス（ＲＤＭＡ）は、ネットワークインターフェイスカード（ＮＩＣ）の機能であり、コンピューターデバイスが、別のコンピューターデバイスのメモリーの情報にアクセスできるようにする。特に、ＲＤＭＡ技術を介して、コンピューターデバイスは、ホストオペレーティングシステム（ＯＳ）を巻き込むことなく、別のコンピューターデバイスのメモリーから情報を読み出すことができる。別のコンピューターデバイスのメモリーに情報を書き込むことも同様である。
【０００４】
図１は、ＣＰＵとメモリーの複合物を有している周辺機器を含んだ従来技術に係るシステムを図示している。システム１００は、システムＣＰＵ１１０、システムメモリー１２０、周辺機器コントローラー１３０、周辺機器１４０、および、ＲＤＭＡデバイス１５０を含んでいる。周辺機器１４０は、プロセッサ１４１及びメモリー１４２を含んでいる。周辺機器１４０及びＲＤＭＡデバイス１５０は、「ピア（”ｐｅｅｒ”）」デバイスとして参照されてよい。
【発明の概要】
【発明が解決しようとする課題】
【０００５】
周辺機器１４０は、ＲＤＭＡ１５０の中に保管されているデータにアクセスする必要があり、その逆もまた同様である。現在、インフィニバンド（ＩｎｆｉｎｉＢａｎｄ）仕様又はＲＤＭＡコンソーシアム仕様、等といった多くの相互接続構造規格の実施は、ピアデバイスが、別のピアデバイスのアドレス空間において保管されているデータに直接的にアクセスすることをできなくしている。
【０００６】
現在のソリューションは、ピアデバイスが、リクエストされたデータを共通に利用可能なメモリーに対して書き込むことを要求する。この説明図においては、システムメモリー１２０であり、相互接続性構造に接続されたあらゆる周辺機器によってアクセス可能である。しかしながら、そうしたデータ転送のために共通のシステムメモリーを使用することは時間の浪費であり、処理のオーバーヘッドを増加させるものである。さらに、共通システムのメモリーを使用することは、周辺機器の処理動作を遅くしてしまう。
【課題を解決するための手段】
【０００７】
本発明の実施例は、ホストと、ＣＰＵ及びメモリーの複合体を含む周辺機器（代替的には、ここにおいてプロセッサアドインカードとして言及されるもの）との間で、リモートダイレクトメモリーアクセス（ＲＤＭＡ）デバイスハードウェアを共有することができるシステム、装置、および、方法を説明する。本発明の実施例は、プロセッサアドインカード（ａｄｄ－ｉｎ　ｃａｒｄ）とＲＤＭＡデバイスとの間でのピアツーピア（ｐｅｅｒ－ｔｏ－ｐｅｅｒ）のデータ転送のための周辺コンポーネント相互接続エクスプレス（ＰＣＩｅ）といった相互接続ハードウェアを利用する。ホストシステムは、メモリー及びレジスターをＲＤＭＡデバイスに対して、及び/又は、ＲＤＭＡデバイスからマップするためのモジュール又はロジックを含んでよい。これにより、ホストシステムのＩ／Ｏオペレーティングシステムと同時に、アドインカードのプロセッサ上のユーザーモードアプリケーションに対して、または、アプリケーションから直接的に実行されるべきＩ／Ｏをできるようにしている。
【図面の簡単な説明】
【０００８】
以降の記述は、本発明の具体化に係る実施例によって与えられる例解を有する図面の説明を含んでいる。図面は、例示として理解されるべきであり、本発明を限定するものとして理解されるべきではない。ここにおいて使用されるように、一つまたはそれ以上の「実施例」は、本発明に係る少なくとも一つの実施に含まれる所定の機能、構成、または、特性を表しているものとして理解されるべきである。このように、ここにおいて表れている

「一つの実施例において」または「代替的な実施例において」といったフレーズは、本発明に係る種々の具体化及び実施を説明するものであり、必ずしも全てが同一の実施例を参照する必要はない。しかしながら、それらはお互いに排他的であることも要しない。
【図１】図１は、ＣＰＵ及びメモリーの複合体を有する周辺機器を含む従来技術のシステムを示している。
【図２】図２は、本発明の一つの実施例に従ったブロックダイヤグラムである。
【図３】図３は、本発明の一つの実施例に従ったシステムコンポーネントのブロックダイヤグラムである。
【図４】図４は、本発明の一つの実施例に従ったプロセスのフローチャートである。
【図５】図５は、本発明の一つの実施例に従ったプロセスのフローチャートである。
【図６】図６は、本発明の一つの実施例に従ったホストとアドインカードモジュールのブロックダイヤグラムである。
【図７】図７は、本発明の一つの実施例を利用するシステムのブロックダイヤグラムである。
【０００９】
所定の詳細及び実施例に係る説明が後に続く。図面の説明を含むものであり、図面は以降に記述される実施例のいくつか又は全てを表している。他の可能性のある具体化、または、ここにおいて表される本発明の概念に係る実施例の説明も同様に後に続く。本発明の実施例の概要が以下に提供され、図面に関してより詳細な説明が続く。
【発明を実施するための形態】
【００１０】
本発明の具体例を利用するプロセッサアドインカードの実施例は、グラフィックプロセッサユニット（ＧＰＵ）を有するグラフィックプロセッササブシステムと、高度に並列なアプリケーションのパフォーマンスを改善するための複数の小さく、低消費電力なプロセッサコア又はマルチコアプロセッサを有するプロセッサアドインカードを含んでいる。異種混在（ｈｅｔｅｒｏｇｅｎｅｏｕｓ）が加速されたコンピューター環境においては、しかしながら、共通で、規格ベースのプログラミング及びコミュニケーションモデルを提供することは難しいと証明されている。このことは、クラスター（ｃｌｕｓｔｅｒ）において特に真実であり、全てのプロセッサからの効率的なコミュニケーションメカニズムを有することが望ましい。プロセッサが、プライマリーシステムのＣＰＵであるか、アドインカードのコプロセッサコアであるかにかかわらずである。共通で、規格ベースのプログラミングモデルは、アプリケーションの開発とメインテナンスを簡素化するだけでなく、システムを使用するためのより大きな適応性を与えてシステムのパフォーマンスの全ての利点を得ることができる。
【００１１】
クラスターは、一般的に、お互いに密接に動作するようにリンク又は相互接続されたコンピューターシステムのグループを参照するものである。それらは、多くの観点で一つのコンピューターを形成するようにリンク等されている。クラスターは、一般的に、単独のコンピューターによって提供されるものよりも大幅に改善されたパフォーマンス及び/又は可用性を提供する。クラスターは、また、典型的には、同等な速度と可用性のある単独のコンピューターよりも、コスト効率が良い。
【００１２】
大規模なクラスターシステムの構築に対する重要な見地は、相互接続である。相互接続は、システムの全てを一緒に接続する「ファブリック（”ｆａｂｒｉｃ”）」を含んでよい。システムをファブリックに対してインターフェイスするホストアダプターも同様である。クラスターは、インフィニバンド仕様又はＲＤＭＡコンソーシアム仕様と調和した相互接続を利用してよい。インフィニバンドは、本来高性能コンピューターにおいて使用されたスイッチファブリックコミュニケーションリンク（ｓｗｉｔｃｈｅｄ　ｆａｂｒｉｃ　ｃｏｍｍｕｎｉｃａｔｉｏｎ　ｌｉｎｋ）であり、スケーラビリティと同様に、サービス品質とフェイルオーバー機能を提供する。インフィニバンド相互接続は、一般的に、よ

り小さなレイテンシー（ｌａｔｅｎｃｙ）、より高いバンド幅、および、改善された信頼性を提供する。

【００１３】

インフィニバンドといったＲＤＭＡアーキテクチャーは、メッセージ受渡しオペレーティングシステムのレイテンシーを削減し、かつ、バンド幅を増加させることによって、高パフォーマンスコンピューティング（ＨＰＣ）クラスターアプリケーションのパフォーマンス改善において、非常に成功してきた。ＲＤＭＡアーキテクチャーは、カーネルバイパス（ｋｅｒｎｅｌ－ｂｙｐａｓｓ）、ダイレクトデータ配置を通じて、ネットワークインターフェイスをアプリケーションに対して大変近くに移動することによってパフォーマンスを改善し、かつ、アプリケーションの要求に合せるようにＩ／Ｏオペレーティングシステムをより広くコントロールすることができる。

【００１４】

ＲＤＭＡアーキテクチャーは、ハードウェアにおいて、プロセス分離、保護、および、アドレス変換をすることができる。このアーキテクチャーは、ホストとコプロセッサアプリケーションが分離されたアドレスドメインにおいて実行されるコプロセッサコア環境に対して相性がいい。しかしながら、従来技術においては、取り付けられたプロセッサ（つまり、プロセッサアドインカード）に対してＲＤＭＡの利益は利用可能ではなかった。本発明の実施例は、ＲＤＭＡアーキテクチャーの利益を取り付けられたプロセッサに対して直接的に提供し、図１に示すようなソリューションに対する必要性を取り除く。図１では、取り付けられたプロセッサに係る内外のコミュニケーションは、ホストメモリーの中への追加のデータコピーを招くことを要し、実質的に、メッセージレイテンシーと達成可能なバンド幅の両方に影響を与えている。

【００１５】

図２は、本発明の一つの実施例に従ったシステムのブロックダイヤグラムである。システム２００は、システムＣＰＵ２１０、システムメモリー２２０、プラットフォームコントローラーハブ（ＰＣＨ）２３０、周辺機器２４０、および、ＲＤＭＡデバイス２５０を含んでいる。この実施例において、周辺機器２４０は、プロセッサアドインカードであり、プロセッサ２４１とメモリー２４２を含んでいる。周辺機器２４０とＲＤＭＡデバイス２５０は、「ピア（”ｐｅｅｒ”）」デバイスとして参照されてよい。

【００１６】

周辺機器２４０は、ＲＤＭＡデバイス２５０の中に保管されているデータにアクセスをリクエストし得る。周辺機器が、ＰＣＩｅリンクを介してＰＣＨ２３０と通信可能に接続されているものと示されている一方で、ＲＤＭＡデバイスは、ＰＣＩｅルートポート２３１を介してＰＣＨ１３０に対して動作可能に接続されているものと示されている。この実施例において、周辺機器２４０は、ＰＣＨを含んでおらず、従って、ホストチャンネルアダプター／ネットワークインターフェイスコントローラー（ＨＣＡ／ＮＩＣ）カードをＲＤＭＡデバイス２５０に専念させる能力を有していない（しかしながら、ＰＣＨ又はオンボードのＲＤＭＡデバイスを伴う内部ＰＣＩｅバスを含んでいる周辺機器も、また、以下に説明される本発明の実施例を利用し得ることが理解されるべきである）。ＰＣＩｅ相互接続は、単独では、ピアデバイスが、別のピアデバイスのアドレス空間の中に保管されているデータにアクセスするようにはできない。

【００１７】

本発明の実施例は、システム２００、ＰＣＨ２３０、および、周辺機器２４０に含まれているロジック及び/又はモジュールについて説明する。周辺機器が、ＲＤＭＡデバイス２５０の中に含まれているデータに対して直接的にアクセスできるようにするものである。つまり、システムメモリー２２０は、図１のシステム１００といった従来技術のソリューションにおいて必要とされたようには必要とされない。リクエストされたデータをホストメモリーの中に受け取るための、ホストメモリーに対する必要性を取り除くことによって、本発明の実施例は、著しく、メッセージレイテンシーを減少させ、達成可能なバンド幅を増加させる。

【００１８】
図３は、本発明の一つの実施例に従ったシステムコンポーネントのブロックダイヤグラムである。この実施例において、ＰＣＨ３１０、周辺機器３２０、および、ＲＤＭＡ３３０は、システムバス３９０（ここにおいては、代替的にイントラノード（ｉｎｔｒａ－ｎｏｄｅ）バスとして参照される）を介してお互いに通信可能に接続されている。周辺機器３２０は、図２に係る機器２４０と類似のものであり、ＲＤＭＡデバイス３３０を専念させる能力を有していない。
【００１９】
この実施例において、周辺機器３２０のメモリー３２１はリクエストキュー（ｑｕｅｕｅ）３２２を含んでいる。リクエストキューは、ＲＭＤＡデバイス３３０のメモリー３３１に向けて指示されたリクエストを保管している。周辺機器のＣＰＵ３２３は、リクエストキューがメモリー３３１へのアクセスに対する未解決のリクエストを有していることを、ＲＤＭＡデバイス３３０に通知する。一つの実施例において、ＲＤＭＡデバイス３３０は、キュー３２２の中に未解決のリクエストがいくらかでもある場合に通知される。他の実施例において、ＲＤＭＡデバイス３３０は、キュー３２２の中の未解決のリクエストの数量が閾値を超えた場合に通知される。
【００２０】
ＣＰＵ３２３は、ＰＣＨ３１０の中に含まれる仮想アドレス３１１に対して書き込むことによって、未解決のリクエストについてＲＤＭＡデバイス３５０に通知する。仮想アドレスは、ＲＤＭＡデバイスのメモリー３３１に対してマップされたメモリーである。未解決のリクエストは、実行されるべきオペレーションを記述することができる。つまり、ｓｅｎｄ、ｒｅｃｅｉｖｅ、ｗｒｉｔｅ、ｒｅａｄ、ａｔｏｍｉｃ　ｃｏｍｐａｒｅ／ｅｘｃｈａｎｇｅ、ａｔｏｍｉｃ　ｆｅｔｃｈ／ａｄｄ、等である。このように、本発明の実施例は、周辺機器３２０からホストシステムのＰＣＨ３１０への「プロキシー（”ｐｒｏｘｙ”）」リクエストに対して記述されてよい。
【００２１】
図４は、本発明の一つの実施例に従ったプロセスのフローチャートである。ここにおいて説明されるように、フローチャートは、種々のプロセス動作のシーケンスに係る実施例を提供する。所定のシーケンス又は順序において示されてはいるが、別に指定がなければ、動作の順序は変更することができる。従って、図示された実施例は、単なる例示として理解されるべきであり、図示されたプロセスは、異なる順序で実行することができ、いくつかの動作は並行に実行され得る。加えて、一つまたはそれ以上の動作は、本発明に種々の実施例において除外することができる。従って、全ての動作が、それぞれの実施において必要とされるわけではない。他のプロセスフローも可能である。
【００２２】
プロセス４００は、プロセッサコアとメモリーを有する周辺機器による、ＲＤＭＡデバイスのメモリーへのアクセスに対するリクエストを受け取るオペレーションを含んでいる、４１０。そのリクエストは、周辺機器のメモリーにおけるキューの中に保管されてよい。いくつかの実施例において、周辺機器は複数のプロセッサコアを含んでおり、それぞれのプロセッサコアに対して分離されたキューを維持している。
【００２３】
周辺機器は、ＰＣＨの中に含まれている仮想アドレスに対してデータを送付することによって、キューの中のアクセスリクエストメッセージをＲＤＭＡデバイスに通知する、４２０。その仮想アドレスは、ＰＣＨの中に含まれており、ＲＤＭＡデバイスのメモリーにマップされている。ＰＣＨは、ＲＤＭＡデバイスにマップされた仮想アドレスを利用する、４３０。そして、ＲＤＭＡデバイスのメモリーに対してＲＤＭＡアクセスリクエストメッセージを伝送する、４４０。そのＲＤＭＡアクセスリクエストメッセージは、周辺機器の中にキューされたアクセスリクエストメッセージに基づくものである。
【００２４】
ＲＤＭＡデバイスは、ＲＤＭＡアクセスリクエストメッセージの受け取りに応じて、周

辺機器からのリクエストを完了する、４５０。いくつかの実施例において、ＲＤＭＡデバイスは、ＰＣＨの中に含まれ、かつ、周辺機器のメモリーにマップされている仮想アドレスに対して完了メッセージを送付する。その仮想アドレスは、完了メッセージを保管するための完了キューを表してよい。例えば、ＲＤＭＡアクセスリクエストメッセージがＲＤＭＡ読み出しリクエストを有する場合、周辺機器によってリクエストされたデータを含んでいるデータメッセージが、ＰＣＨを介して周辺機器に対して送付される。ＲＤＭＡアクセスリクエストメッセージがＲＤＭＡ書き込みリクエストを有する場合、ＰＣＨは、書き込まれるべきデータを含んでいるデータメッセージをＲＤＭＡデバイスに対して伝送する。いくつかの実施例において、ＲＤＭＡ動作の完了は、ＲＤＭＡデバイスによって完了メッセージがいつ送付されるかを決定するものではない。例えば、インフィニバンド仕様は、リクエストは、それらが投稿された順序で実行されることを要求する。先の全てのＲＤＭＡ動作が完了するまで、周辺機器は完了メッセージを受け取らないことを意味するものである。

【００２５】

図５は、本発明の一つの実施例に従ったプロセスのフローチャートである。プロセス５００は、周辺機器（プロセッサコアも含んでいる）のメモリーへのアクセスに対するリクエストを受け取るオペレーションを含んでいる、５１０。そのリクエストは、ＲＤＭＡデバイスのメモリーにおけるキューの中に保管されてよい。

【００２６】

ＲＤＭＡデバイスは、ＰＣＨの中に含まれている仮想アドレスに対してデータを送付することによって、キューの中のアクセスリクエストメッセージをホストシステムに通知する、５２０。その仮想アドレスは、周辺機器のメモリーにマップされている。ＰＣＨは、アクセスリクエストメッセージパラメーターとして、ＲＤＭＡデバイスにマップされた仮想アドレスを利用する、５３０。そして、周辺機器のメモリーに対してアクセスリクエストメッセージを伝送する、５４０。そのアクセスリクエストメッセージは、ＲＤＭＡデバイスの中にキューされたアクセスリクエストメッセージに基づいて、周辺機器に対して伝送される。

【００２７】

周辺機器は、ＰＣＨからのアクセスリクエストメッセージの受け取りに応じて、ＲＤＭＡデバイスからのリクエストを完了する。例えば、アクセスリクエストメッセージが読み出しリクエストを有する場合、ＲＤＭＡデバイスによってリクエストされたデータを含んでいるデータメッセージが、ＰＣＨを介してＲＤＭＡデバイスに対して送付される。アクセスリクエストメッセージが書き込みリクエストを有する場合、ＲＤＭＡデバイスは、ＰＣＨを介して、書き込まれるべきデータを含んでいるデータメッセージを周辺機器に対して伝送する。ＲＤＭＡデバイスは、ＲＤＭＡ動作及び/又はパラメーターのタイプに応じて、ＰＣＨの中に含まれ、かつ、周辺機器のメモリーにマップされている仮想アドレスに対して完了メッセージを送付する、５５０。その仮想アドレスは、完了メッセージを保管するための完了キューを表してよい。

【００２８】

図６は、本発明の一つの実施例に従ったホストとプロセッサアドインカードモジュールのブロックダイヤグラムである。上述のように、本発明の実施例は、インフィニバンド仕様（例えば、規格リリース１．０．ａ、２００１年7月１９日発行）に準じた相互接続を利用してよい。インフィニバンドは、本来高性能コンピューターにおいて使用されるスイッチファブリックコミュニケーションリンクであり、スケーラビリティと同様に、サービス品質とフェイルオーバー機能を提供する。インフィニバンド相互接続は、一般的に、より小さなレイテンシー、より高いバンド幅、および、改善された信頼性を提供する。インフィニバンドは、コンピューターシステムのコンポーネントの中、および、コンピューターの中で情報を移動する方法を提供する。インフィニバンドは、コンピューターのＣＰＵが、非常に高いパフォーマンスでＩ／Ｏデバイス及び他のＣＰＵと直接的にコミュニケーションできるようにする。インフィニバンド技術は、あらゆるネットワークに係るデータ

センターのバックエンド（ｂａｃｋ　ｅｎｄ）を対象としている。ネットワークインフラストラクチャーのフロントエンド及びミドルエンドは、典型的には、従来のイーサネット（登録商標）（Ｅｔｈｅｒｎｅｔ）技術を含んでいる。別の言葉で言えば、インフィニバンド及びイーサネット技術は、両方とも同一のホストによって使用され得るものである。

【００２９】

ホスト６１０と周辺機器カード６３０は、種々のインフィニバンドモジュール（以下に説明される）を含むように示されており、ＰＣＩｅ相互接続６６０を介してＲＤＭＡデバイス６５０と通信可能に接続されている。以下に説明するように、ＲＤＭＡデバイス６５０によって受け取られるＲＤＭＡメッセージは、ＲＤＭＡコンソーシアム仕様（例えば、ＲＤＭＡプロトコル規格（バージョン１．０）、２００２年１０月２１日発行）に準じてよい。ＲＤＭＡコンソーシアムのＲＤＭＡプロトコルは、ＴＣＰ／ＩＰプロトコルにおけるＴＣＰ層の上で規定されている。従って、ＲＤＭＡ動作は、プロトコルスタックのトップから送信器側のボトムまで進み、次に、プロトコルスタックを越えて受信器側のトップまで進む。本発明の実施例によって利用されるＲＤＭＡプロトコルは、基本のＴＣＰ／ＩＰ処理ハードウェアを構築することを含んでよい。パケットを受け取り、ＴＣＰ／ＩＰを停止し、ＴＣＰ／ＩＰを通してアプリケーションに対してパケットを処理し、そして、データ及びメモリーに書き込むためのアプリケーション層でのアドレスを抽出することである。ＲＤＭＡプロトコルは、データ転送をより効率的にする（特により大きなデータペイロードに対して）ことによって、パケットをメモリーにコピーすることを防ぐことができる（そして、その後にデータペイロードをメモリーにコピーする）。

【００３０】

この実施例において、ホスト６１０上のモジュールと周辺機器６３０はお互いにコミュニケーションし、ＰＣＩｅ相互接続６６０をわたりＲＤＭＡデバイス６５０に対して直接的なアクセスを有している。モジュールは、ＲＤＭＡデバイスのリソースを管理するためにＰＣＩｅ相互接続６６０をわたりプロキシーオペレーションに対するスプリットドライバー（ｓｐｌｉｔ－ｄｅｒｉｖｅｒ）モデルを使用する。

【００３１】

ホスト６１０は、インフィニバンドベースのソフトウェアコミュニケーションスタックを含むように図示されている。メッセージパッシングインターフェイス（ＭＰＩ）アプリケーション６１１、ＲＤＭＡ　ＡＰＩであるユーザーモードダイレクトアクセスプロバイダーライブラリー（ｕＤＡＰＬ）６１２、ＩＢバーブ（ｖｅｒｂ）（つまり、機能）ライブラリー６１３、ベンダーライブラリー６１４、ＩＢユーバーブ（ｕｖｅｒｂ）６３５、および、ＩＢコア６３６を含んでいる。ホストと周辺機器は、さらに、以下に説明されるモジュールとして実行される本発明の実施例を利用する。

【００３２】

ホスト６１０は、ＩＢプロキシーデーモン（ｄａｅｍｏｎ）６１８を含んでいる。ＩＢプロキシーデーモンは、ホストユーザーモードのアプリケーションであり、基底にあるベンダードライバー６１７に対するコール（ｃａｌｌ）のためにＩＢプロキシーサーバー６１９（以下に説明される）に対してユーザーモードプロセスコンテクスト（ｃｏｎｔｅｘｔ）を提供する。ユーザーモードプロセスコンテクストは、ベンダードライバー６１７を変更することなくＲＤＭＡデバイス６５０のメモリーの仮想アドレスマッピングを実行するために使用され得る。

【００３３】

ホスト６１０は、さらに、ＩＢプロキシーサーバー６１９を含んでおり、ＩＢプロキシーサーバーはホストカーネルモジュールを有している。この実施例において、そのＩＢプロキシーサーバーは、コミュニケーションと周辺機器６３０（以下に説明される）のＩＢプロキシークライアント６３８のためのコマンドサービスを提供する。この実施例において、ＩＢプロキシーサーバー６１９は、クライアント接続をリスン（ｌｉｓｔｅｎ）し、ＲＤＭＡデバイスの追加、除去、および、イベント通知メッセージをリレーする。ＩＢプロキシーサーバー６１９は、さらに、ＩＢプロキシークライアント６３８のためにＩＢコ

アレイヤー６１６に対するカーネルモードＩＢバーブコールを開始して、その結果を戻すことができる。

【００３４】

周辺機器６３０は、ＩＢプロキシークライアント６３８を含んでおり、ＩＢプロキシークライアントはカーネルモジュールを有している。そのＩＢプロキシークライアントは、ホスト６１０上でカーネルモードＩＢバーブを実行するために、ベンダープロキシードライバー６３７（以下に説明される）に対してプログラミングインターフェイスを提供する。インターフェイスは、さらに、コマンドのフォーマット及びコミュニケーションの実行に係る詳細を抽出し得る。ＩＢプロキシークライアント６３８は、所定のデバイスの追加、除去、および、ベンダープロキシードライバー６３７へのイベント通知に対するコールバック（ｃａｌｌｂａｃｋ）を呼び出す。

【００３５】

周辺機器６３０は、さらに、ベンダープロキシードライバー６３７を含んでおり、ベンダープロキシードライバーはカーネルモジュールを有している。所定のＲＤＭＡデバイスをサポートするために異なるベンダープロキシードライバーが使用されてよい。そのベンダープロキシードライバーそれぞれは、ＲＤＭＡのデバイスの追加、除去、および、所定のＰＣＩｅドライバーからのイベント通知をＩＢプロキシークライアント６３８に登録し得る。ベンダープロキシードライバー６３７は、カーネルモードＩＢバーブコールを実行するために、ＩＢプロキシークライアント６３８によって提供されるプログラミングインターフェイスを使用してよい。そのベンダープロキシードライバーは、さらに、ベンダーライブラリー６３４とホスト６１０上のベンダードライバー６１７との間で共有されるあらゆるプライベートデータの解釈及び変換を取り扱う。

【００３６】

この実施例において、ホスト６１０と周辺機器６３０の両方は、シンメトリックコミュニケーションインターフェイス（ＳＣＩＦ）モジュール６２０と６４０を、それぞれに含んでいる。そのＳＣＩＦモジュールは、シングルプラットフォームの中でイントラノード（ｉｎｔｒａ－ｎｏｄｅ）コミュニケーションのためのメカニズムを提供する。ＳＣＩＦは、ホスト６１０と周辺機器６３０との間で対称なＡＰＩを提供する一方で、ＰＣＩｅにわたるコミュニケーション（および、関連する周辺機器ハードウェアのコントロール）の詳細を抽出する。

【００３７】

上述のモジュールに加えて、本発明の実施例は、ＩＢコアレイヤー６１６におけるコールを利用して、ベンダーライブラリー６１４及び６３４とベンダードライバー６１７との間でプライベートデータを転送する。メモリーをＲＤＭＡデバイス６５０に対してマッピングすることも同様である。

【００３８】

上記の「バーブ（”ｖｅｒｂ”）」（つまり、機能）は、ＲＤＭＡデバイス６５０に向けたＰＤＭＡオペレーションを実行する。バーブは、特権クラスと非特権クラスに分類されてよい。特権バーブは、典型的にＲＤＭＡハードウェアのリソースを割り当て、管理するために使用され、ベンダードライバー６１７によって実施される。周辺機器６３０上で稼働しているアプリケーションのために、これらの特権バーブが、ベンダープロキシーサーバー６３７を通じてホスト６１０上のベンダードライバー６１７に対してフォワードされる。一旦、ハードウェアが割り当てられ開始されると、非特権バーブは、カーネルをバイパスし、リソース割り当ての最中にアプリケーションアドレス空間の中にマップされたメモリーを使用して、ユーザーモードからハードウェアに対する直接的なアクセスを許可する。同様に、ＲＤＭＡデバイスは、キューにアクセスすることができ、プロセスアドレス空間へ、または、プロセスアドレス空間から直接的にデータ転送を実行することができる。このように、本発明の実施例は、周辺機器６３０上のクライアントプロセスを、まるでホスト６１０上の別の「ユーザーモード」プロセスであるかのようにする。

【００３９】

従って、上記のモジュールにより、ホスト６１０は、ＲＤＭＡデバイス６５０のメモリーにアクセスするための、周辺機器６３０のプロセッサコアからのアクセスリクエストメッセージに係る通知を受け取ることができる。その通知は、ホスト６１０及びＲＤＭＡデバイス６５０のメモリーにマップされたメモリー（例えば、図３に示されるように）の中に含まれる仮想アドレスにおいて受け取られる。ホストデバイス６１０は、そのアクセスリクエストメッセージに基づいて、ＲＤＭＡデバイスに対してＲＤＭＡアクセスリクエストを転送する。ここで、リクエストは仮想アドレスＲＤＭＡパラメーターを含んでいる。従って、データリクエストのタイプが何であれ、アクセスリクエストメッセージは、例えば、ｓｅｎｄ、ｒｅｃｅｉｖｅ、ｗｒｉｔｅ、ｒｅａｄ、ａｔｏｍｉｃ　ｃｏｍｐａｒｅ／ｅｘｃｈａｎｇｅ、ａｔｏｍｉｃ　ｆｅｔｃｈ／ａｄｄ、等を含んでいる。データリクエストは、ＲＤＭＡデバイス６５０に対して、周辺機器６３０からというよりむしろ、まるでホスト６１０上の「ユーザーモード」から生じたものであるように見える。

【００４０】

図７は、本発明の一つの実施例を利用するシステムのブロックダイヤグラムである。システム７００は、サーバープラットフォームを記述しているが、例えば、以下のものに含まれてのよい。デスクトップコンピューター、ラップトップコンピューター、タブレットコンピューター、ネットブック、ノートブックコンピューター、パーソナルデジタルアシスタント（ＰＤＡ）、サーバー、ワークステーション、携帯電話、モバイルコンピューター機器、インターネット機器、ＭＰ３又はメディアプレーヤー、または、あらゆる他のタイプのコンピューターデバイス、である。

【００４１】

システム７００は、システムバス７２０を介して、データを交換するためのプロセッサ７１０、ユーザーインターフェイス７６０、システムメモリー７３０、周辺機器コントローラー７４０、および、ネットワークコネクター７５０を含んでいる。その周辺機器コントローラーは、周辺機器及びＲＤＭＡデバイスとコミュニケーション可能に接続され、上記の本発明の実施例のいずれかに従ってデバイ患者関連情報のＩ／Ｏリクエストを管理する。

【００４２】

システム７００は、さらに、システム７００の種々のエレメントによって処理されるべき信号を送信及び受信するためのアンテナとＲＦ回路７７０を含んでいる。上記のアンテナは、指向性アンテナ又は無指向性アンテナであってよい。ここにおいて使用されるように、無指向性アンテナという用語は、少なくとも一平面において実質的に均一なパターンを有するあらゆるアンテナを参照するものである。例えば、いくつかの実施例において、アンテナは、ダイポール（ｄｉｐｏｌｅ）アンテナ、または、四分の一波長（ｑｕａｒｔｅｒ　ｗａｖｅ）アンテナといった、無指向性アンテナであってよい。例えば、いくつかの実施例において、アンテナは、パラボラアンテナ、パッチアンテナ、または、八木アンテナといった、指向性アンテナであってもよい。いくつかの実施例において、システム７００は、複数の物理的なアンテナを含んでよい。

【００４３】

ネットワークコネクター７５０から離れているように示されているが、他の実施例において、アンテナ及びＲＦ回路７７０は、無線インターフェイスを含んでよいことが理解されるべきである。無線インターフェイスは、これらに限定されるわけではないが、ＩＥＥＥ８０２．１１規格及び関連ファミリー規格、Ｈｏｍｅ　Ｐｌｕｇ　ＡＶ（ＨＰＡＶ）、ウルトラワイドバンド（ＵＷＢ）、Ｂｌｕｅｔｏｏｔｈ（登録商標）、ＷｉＭａｘ、または、無線通信プロトコルの他の形式に従って動作するものである。

【００４４】

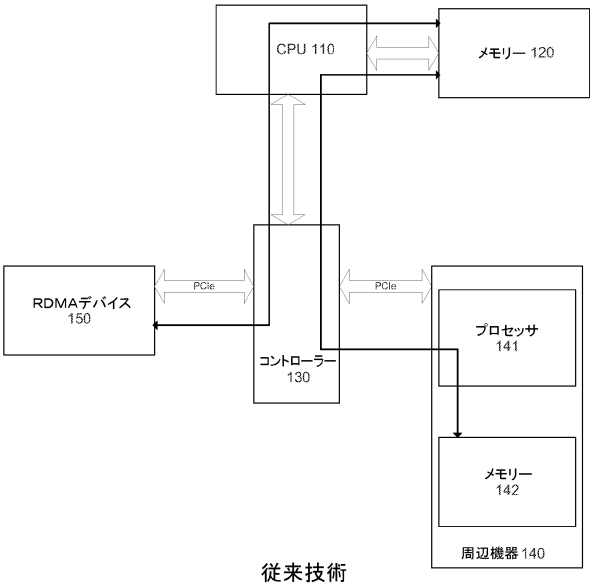
ここにおいて、プロセス、サーバー、または、ツールとして説明され、上記に参照された種々のコンポーネントは、説明された機能を実行するための手段であってよい。ここにおいて説明されたそれぞれのコンポーネントは、ソフトウェア又はハードウェア、または、これらの組み合わせを含んでいる。それぞれ及び全てのコンポーネントは、ソフトウェ

アモジュール、ハードウェアモジュール、特定用途ハードウェア（例えば、アプリケーション所定のハードウェア、ＡＳＩＣ、ＤＳＰ、等）、エンベッドコントローラー、ハードワイヤード回路、ハードウェアロジック、等として実施されてよい。ソフトウェアコンテンツ（例えば、データ、インストラクション、コンフィグレーション）は、固定で有形なコンピューター又はマシンで読取り可能な媒体を含む製品を介して提供されてよい。実行され得るインストラクションを表すコンテンツを提供するものである。コンテンツは、結果として、ここにおいて説明された種々の機能／動作のコンピューターによる実行を生じる。

【００４５】

コンピューターで読取り可能な固定記録媒体は、コンピューター（例えば、計算機器、電子システム、等）によってアクセス可能な形式で情報を提供（つまり、保管及び/又は送信）するあらゆるメカニズムを含んでいる。記録可能／記録不能媒体（例えば、読み出し専用メモリー（ＲＯＭ）、ランダムアクセスメモリー（ＲＡＭ）、磁気ディスク記録媒体、光記録媒体、フラッシュメモリーデバイス、等）といったものである。コンテンツは、直接的に実行可能な（「オブジェクト」又は「実行可能」形式）ソースコード、または、異なるコード（「デルタ」又は「パッチ」コード）であってよい。コンピューターで読取り可能な固定記録媒体は、また、ストレージ又はデータベースを含んでよく、そこからコンテンツをダウンロードすることができる。コンピューターで読取り可能な媒体は、また、販売時又は引き渡し時に、媒体上にコンテンツが保管されているデバイス又はプロダクトを含んでよい。従って、コンテンツが保管されたデバイスを引き渡すこと、または、コミュニケーション媒体にわたるダウンロードのためにコンテンツを提供することは、ここにおいて説明されたようなコンテンツを伴う製品を提供することとして理解されてよい。

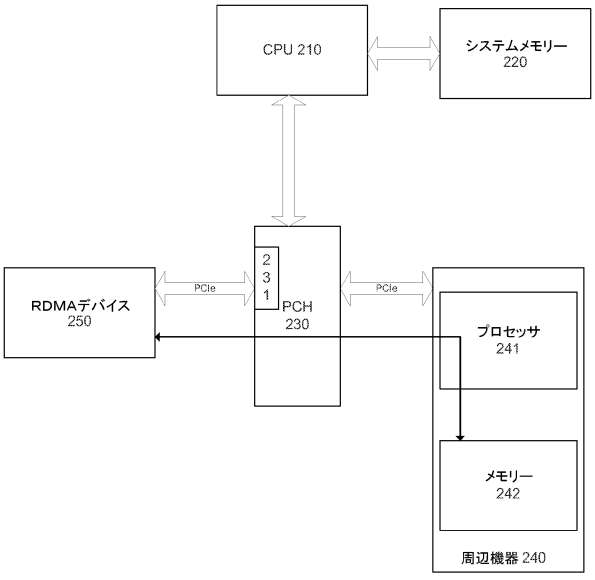
【図１】



従来技術

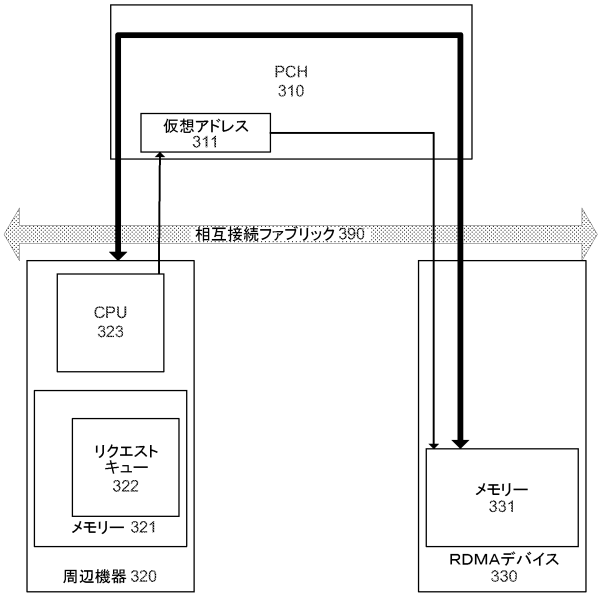
100

【図２】

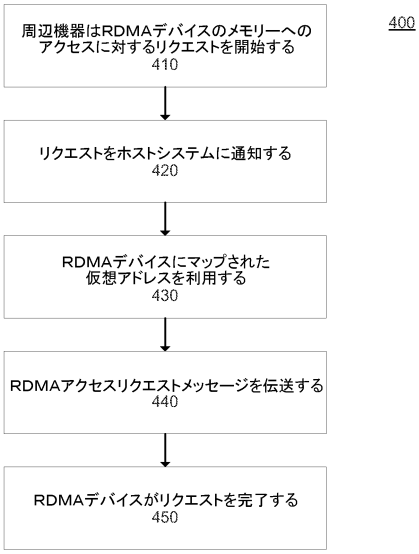


200

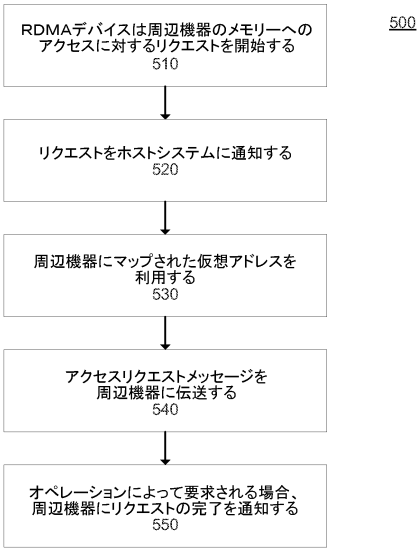
【図３】


PCH
310
仮想アドレス
311
相互接続ファブリック 390
CPU
323
リクエスト
キュー
322
メモリー 321
周辺機器 320
メモリー
331
RDMAデバイス
330

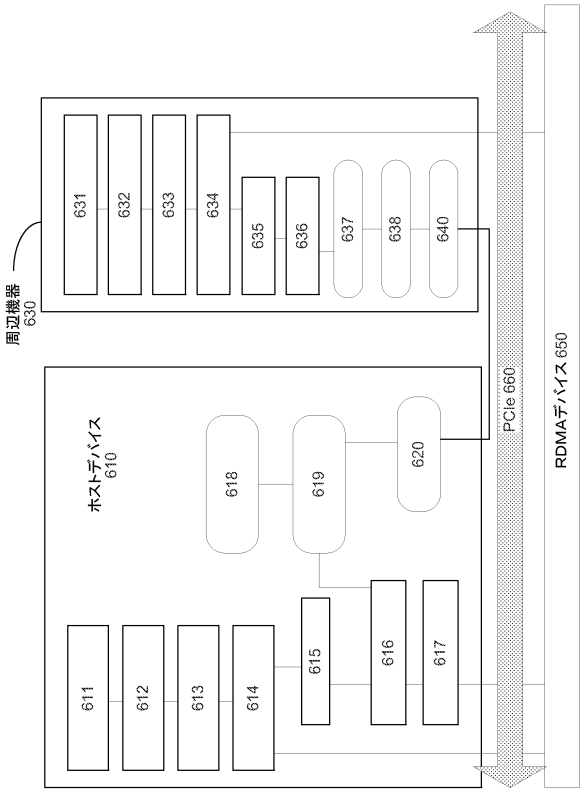
【図４】


400
周辺機器はRDMAデバイスのメモリーへの
アクセスに対するリクエストを開始する
410
リクエストをホストシステムに通知する
420
RDMAデバイスにマップされた
仮想アドレスを利用する
430
RDMAアクセスリクエストメッセージを伝送する
440
RDMAデバイスがリクエストを完了する
450

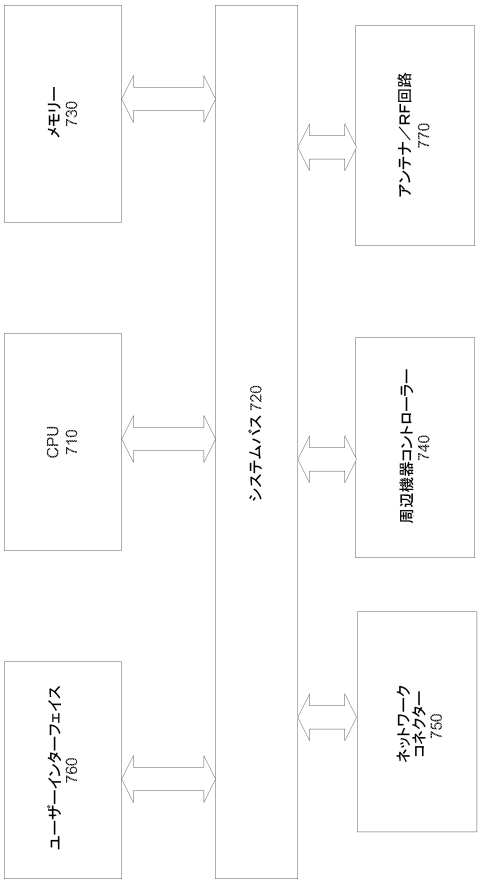
【図５】


500
RDMAデバイスは周辺機器のメモリーへの
アクセスに対するリクエストを開始する
510
リクエストをホストシステムに通知する
520
周辺機器にマップされた仮想アドレスを
利用する
530
アクセスリクエストメッセージを
周辺機器に伝送する
540
オペレーションによって要求される場合、
周辺機器にリクエストの完了を通知する
550

【図６】


周辺機器
630
631
632
633
634
635
636
637
638
640
ホストデバイス
610
618
619
620
611
612
613
614
615
616
617
PCIe 660
RDMAデバイス 650

【図７】


700
メモリー
730
CPU
710
ユーザーインターフェイス
760
システムバス 720
アンテナ／RF回路
770
周辺機器コントローラー
740
ネットワーク
コネクター
750

フロントページの続き

(72)発明者 ウッドラフ, ロバート, ジェイ.
アメリカ合衆国 ９７１０６ オレゴン州 バンクス ノースウエスト グランドヴュー １３９８３
(72)発明者 リー, デイヴィッド, エム.
アメリカ合衆国 ９７２２９ オレゴン州 ポートランド ノースウエスト クレストヴュー ウェイ ２２０３
(72)発明者 デイヴィス, アーリン, アール.
アメリカ合衆国 ９７１４８ オレゴン州 ヤムヒル ノースイースト コーヴ オーチャード ロード ２３１９０
(72)発明者 ヘフティ, マーク, ショーン
アメリカ合衆国 ９７００７ オレゴン州 アロハ サウスウエスト ハート ロード １８５６０
(72)発明者 コフマン, ジェリー, エル.
アメリカ合衆国 ９７１２４ オレゴン州 ヒルズボロ ノースイースト １１ス コート ３１５３

審査官 寺谷 大亮

(56)参考文献 米国特許第０７７１１７９３（ＵＳ，Ｂ１）
米国特許出願公開第２００４／００３４７２５（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １３／３６
Ｇ０６Ｆ １３／２８
Ｇ０６Ｆ １３／３８