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

特許第6492083号
(P6492083)

(45)発行日　平成31年3月27日(2019.3.27)　(24)登録日　平成31年3月8日(2019.3.8)

(51)Int.Cl.		F I		
G06F 13/38	(2006.01)	G06F	13/38	320Z
G06F 13/28	(2006.01)	G06F	13/28	310A
G06F 13/10	(2006.01)	G06F	13/10	310A
G06F 13/00	(2006.01)	G06F	13/00	351A
H04L 12/861	(2013.01)	G06F	13/38	350

請求項の数 11　(全 29 頁)　最終頁に続く

(21)出願番号	特願2016-536668(P2016-536668)
(86)(22)出願日	平成26年10月21日(2014.10.21)
(65)公表番号	特表2017-501492(P2017-501492A)
(43)公表日	平成29年1月12日(2017.1.12)
(86)国際出願番号	PCT/US2014/061640
(87)国際公開番号	WO2015/084506
(87)国際公開日	平成27年6月11日(2015.6.11)
審査請求日	平成29年10月11日(2017.10.11)
(31)優先権主張番号	14/096,987
(32)優先日	平成25年12月4日(2013.12.4)
(33)優先権主張国	米国(US)
(31)優先権主張番号	14/096,949
(32)優先日	平成25年12月4日(2013.12.4)
(33)優先権主張国	米国(US)

(73)特許権者　502303739
オラクル・インターナショナル・コーポレイション
アメリカ合衆国カリフォルニア州９４０６５レッドウッド・シティー，オラクル・パークウェイ５００
(74)代理人　110001195
特許業務法人深見特許事務所
(72)発明者　アガーワル，ウッタム
アメリカ合衆国、９４５８３　カリフォルニア州、サン・ラモン、サマークリーク・レーン、４２０

審査官　田名網　忠雄

最終頁に続く

(54)【発明の名称】インフィニバンド（ＩＢ）上で仮想ホストバスアダプタ（ｖＨＢＡ）を管理およびサポートするためのシステムおよび方法、ならびに単一の外部メモリインターフェイスを用いてバッファの効

(57)【特許請求の範囲】

【請求項１】

コンピューティング環境において、入力／出力（Ｉ／Ｏ）仮想化をサポートするためのシステムであって、

ネットワークファブリック上のサーバに関連付けられたチップを備え、前記チップは、複数のパケットバッファを含む外部メモリに関連付けられており、

前記チップは、物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを含む１つ以上のパケットの状態を保存するオンチップメモリを含み、

前記チップは、

前記外部メモリ上の前記複数のパケットバッファ内の前記１つ以上のパケットをキューに入れ、

前記１つ以上のパケットの状態に基づいて、前記外部メモリから前記１つ以上のパケットを読出し、および

前記１つ以上のパケットを前記ネットワークファブリック上の前記サーバに送信するように動作する、システム。

【請求項２】

前記ネットワークファブリックは、インフィニバンド（ＩＢ）ファブリックであり、

前記サーバは、１つ以上のキューペア（ＱＰ）に関連付けられる、請求項１に記載のシステム。

【請求項３】

前記複数のパケットバッファは、１つ以上のバッファリストに格納され、
各バッファリストは、前記サーバに関連付けられたキューペア（ＱＰ）に対応する、請求項２に記載のシステム。
【請求項４】
前記チップは、１つ以上のＩＢヘッダとシーケンス番号とを前記物理ＨＢＡから受信した各パケットに追加するように動作する、請求項２または３に記載のシステム。
【請求項５】
少なくとも１つのＩＢヘッダは、ＩＢ命令を含む、請求項４に記載のシステム。
【請求項６】
前記ＩＢファブリック内のキューペア（ＱＰ）をターゲティングする複数のパケットは、パケットシーケンス番号スペースを共有するように配置される、請求項２～５のいずれか１項に記載のシステム。
【請求項７】
前記サーバに関連付けられたＱＰのために、複数のディスク読取命令、リモートダイレクトメモリアクセス（ＲＤＭＡ）読取リクエストおよび送信命令を多重化することをサポートするように、複数のコンテキストが開かれる、請求項２～６のいずれか１項に記載のシステム。
【請求項８】
前記１つ以上のパケットの状態は、前記外部メモリに格納された前記１つ以上のパケットがすべてキューに入れられたことを示す、および／または、１つ以上の関連ＩＢ命令が更新されたことを示す、請求項１～７のいずれか１項に記載のシステム。
【請求項９】
前記オンチップメモリ内の各エントリは、２ビットの幅を有しており、
前記オンチップメモリ内の各エントリの一方のビットは、関連ＩＢ命令を更新する必要があるか否かを示し、
他方のビットは、チップがキューに入れられたパケットを前記外部メモリから読出すことができるか否かを示す、請求項１～８のいずれか１項に記載のシステム。
【請求項１０】
コンピューティング環境において、効率的なパケット処理をサポートするための方法であって、
前記コンピューティング環境は、ネットワークファブリック上のサーバに関連付けられたチップを備え、前記チップは、複数のパケットバッファを含む外部メモリに関連付けられており、
物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを含む１つ以上のパケットの状態を、前記チップに設けられるオンチップメモリ上に保存するステップとを含み、
前記チップは、
前記外部メモリ上の前記複数のパケットバッファ内の前記１つ以上のパケットをキューに入れ、
前記１つ以上のパケットの状態に基づいて、前記外部メモリから前記１つ以上のパケットを読出し、および
前記１つ以上のパケットを前記ネットワークファブリック上の前記サーバに送信するように動作する、方法。
【請求項１１】
命令を含む機械読取可能プログラムであって、前記命令は、実行されると、複数のステップをシステムに実行させ、
前記システムは、ネットワークファブリック上のサーバに関連付けられたチップを備え、前記チップは、複数のパケットバッファを含む外部メモリに関連付けられており、
前記複数のステップは、
物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを含む１つ以上の

パケットの状態を、前記チップに設けられるオンチップメモリ上に保存するステップと、
前記外部メモリ上の前記複数のパケットバッファ内の前記１つ以上のパケットをキューに入れ、前記１つ以上のパケットの状態に基づいて、前記外部メモリから前記１つ以上のパケットを読出し、および前記１つ以上のパケットを前記ネットワークファブリック上の前記サーバに送信するように、前記チップを動作させるステップとを含む、機械読取可能プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

著作権表示

この特許文書の開示の一部には、著作権保護の対象となるものが含まれている。著作権者は、特許商標庁の特許ファイルまたは記録に掲載された特許文書または特許開示の複製に対しては異議を唱えないが、その他の場合、すべての著作権を留保する。

【０００２】

発明の分野

本発明は、一般に、コンピュータシステムに関連し、特に、コンピューティング環境において、入力／出力（Ｉ／Ｏ）仮想化のサポートに関連する。

【背景技術】

【０００３】

背景

インフィニバンド（ＩＢ）技術は、クラウドコンピューティングファブリックの基盤として、その応用が増加しつつである。より大きなクラウドコンピューティングアーキテクチャが導入されるため、従来のネットワークおよびストレージに関連する性能上および管理上のボトルネックが重大な問題となっている。このような問題を対処することは、本発明の実施形態の一般的な意図である。

【発明の概要】

【課題を解決するための手段】

【０００４】

概要

システムおよび方法は、コンピューティング環境において、入力／出力（Ｉ／Ｏ）仮想化をサポートすることができる。システムは、チップを含むことができる。このチップは、ネットワークファブリック上のサーバに関連付けられている。さらに、このチップは、複数のパケットバッファを含む外部メモリに関連付けられている。また、オンチップメモリは、物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを含む１つ以上のパケットの状態を保存する。さらに、チップは、外部メモリ上の複数のパケットバッファ内の１つ以上のパケットをキューに入れ、１つ以上のパケットの状態に基づいて、外部メモリから１つ以上のパケットを読出し、１つ以上のパケットをネットワークファブリック上のサーバに送信するように動作する。

【０００５】

システムおよび方法は、コンピューティング環境において、入力／出力（Ｉ／Ｏ）仮想化をサポートすることができる。システムは、１つ以上の仮想ホストバスアダプタ（ｖＨＢＡ）に関連付けられた複数のパケットバッファを含む空きバッファプールを含み、ｖＨＢＡの各々は、１つ以上のパケットバッファに指向するバッファポインタの主要リンクリストを空きバッファプールに保存する。また、入力／出力（Ｉ／Ｏ）装置に関連付けられたオンチップメモリ上で、コンテキストテーブルを定義することができる。このコンテキストテーブルは、ディスク読取操作のために、空きバッファプールから割当てられた１つ以上のパケットバッファに指向するバッファポインタの一時リンクリストを保存する。Ｉ／Ｏ装置は、ディスク読取操作を実行する物理ホストバスアダプタ（ＨＢＡ）からすディスク読取データを受信すると、コンテキストテーブルを開き、バッファポインタの一時リンクリストを更新し、およびコンテキストテーブルが閉じられると、バッファポインタの

一時リンクリストをバッファポインタの主要リンクリストに合併するように、動作する。
【０００６】
本明細書に記載のシステムおよび方法は、コンピューティング環境において、入力／出力（Ｉ／Ｏ）仮想化をサポートすることができる。システムは、メモリにおいて、空きバッファプールを備えることができる。Ｉ／Ｏ装置は、空きバッファプールを用いて、物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを保存するように動作する。空きバッファプールは、２次元リンクリストおよび１次元リンクリストを含むことができる。２次元リンクリストの各エントリは、連続したメモリ位置で複数のパケットバッファを含み、１次元リンクリストの各エントリは、単一のパケットバッファを含む。
【図面の簡単な説明】
【０００７】
【図１】さまざまなメモリインターフェイスを用いて、入力／出力（Ｉ／Ｏ）仮想化をサポートすることを示す図である。
【図２】本発明の一実施形態に従って、単一のメモリインターフェイスを用いて、入力／出力（Ｉ／Ｏ）仮想化をサポートすることを示す図である。
【図３】本発明の一実施形態に従って、単一のメモリインターフェイスを用いて、入来トラフィックを処理するための体系をサポートすることを示す図である。
【図４】本発明の一実施形態に従って、Ｉ／Ｏ装置上でディスク読取操作を開始することを示す図である。
【図５】本発明の一実施形態に従って、Ｉ／Ｏ装置上で要求したＩＯＣＢを抽出することを示す図である。
【図６】本発明の一実施形態に従って、Ｉ／Ｏ装置上でディスク読取データを処理することを示す図である。
【図７】本発明の一実施形態に従って、Ｉ／Ｏ装置上でディスク読取操作を完了する処理を示す図である。
【図８】本発明の一実施形態に従って、複数の仮想ホストバスアダプタ（ｖＨＢＡ）を用いて、Ｉ／Ｏ仮想化をサポートすることを示す図である。
【図９】本発明の一実施形態に従って、仮想ホストバスアダプタ（ｖＨＢＡ）において、複数のコンテキストをサポートすることを示す図である。
【図１０】本発明の一実施形態に従って、オンチップメモリを用いて、外部メモリ上でキューに入れられたパケットの状態を保存することを示す図である。
【図１１】本発明の一実施形態に従って、単一のメモリインターフェイスを用いて、入力／出力（Ｉ／Ｏ）仮想化をサポートすることを示す例示的なフローチャートである。
【図１２】本発明の一実施形態に従って、空きバッファプールを用いて、複数の仮想ホストバスアダプタ（ｖＨＢＡ）をサポートすることを示す図である。
【図１３】本発明の一実施形態に従って、ハイブリッドリンクリスト構造を用いて、ディスク読取操作をサポートすることを示す図である。
【図１４】本発明の一実施形態に従って、ハイブリッドのリンクリスト構造を用いて、ヘッドラインブロッキングを回避することを示す図である。
【図１５】本発明の一実施形態に従って、ハイブリッドリンクリスト構造を用いて、ヘッドラインブロッキングを回避することを示す例示的なフローチャートである。
【図１６】本発明の一実施形態に従って、Ｉ／Ｏ装置のために２次元リンクリスト構造をサポートすることを示す図である。
【図１７】本発明の一実施形態に従って、Ｉ／Ｏ装置のためにメモリの効率的な使用をサポートすることを示す図である。
【図１８】本発明の一実施形態に従って、コンピューティング環境において、効率的なパケット処理をサポートすることを示す例示的なフローチャートである。
【発明を実施するための形態】
【０００８】
詳細な説明

本発明は、限定することなく、例示として、添付の図面に示される。図面において、同様の参照番号は、同様の要素を標記する。本開示において、「一実施形態」または「１つの実施形態」または「いくつかの実施形態」を言及する場合、必ずしも同様の実施形態に限定されず、少なくとも１つの実施形態を意味することに留意すべきである。

【０００９】

本発明の以下の説明において、高性能ネットワークの一例として、インフィニバンド（ＩＢ）ネットワークを使用する。限定することなく、他の種類の高性能ネットワークを使用できることは、当業者には明らかであろう。また、本発明の以下の説明において、ストレージネットワークの一例として、ファイバチャネル（ＦＣ）ストレージネットワークを使用する。限定することなく、他の種類のストレージネットワークを使用できることは、当業者には明らかであろう。

【００１０】

本明細書に記載のシステムおよび方法は、１つ以上の仮想ホストバスアダプタ（ｖＨＢＡ）を用いて、入力／出力（Ｉ／Ｏ）仮想化をサポートすることができる。

【００１１】

入力／出力（Ｉ／Ｏ）仮想化

２つの異なるメモリインターフェイスに基づいて、ＩＢファブリック上で、Ｉ／Ｏ仮想化をサポートすることができる。

【００１２】

図１は、異なるメモリインターフェイスを用いて、入力／出力（Ｉ／Ｏ）仮想化をサポートすることを示す図である。図１に示すように、Ｉ／Ｏ装置１００は、ファイバチャネル（ＦＣ）ドメイン１０１およびインフィニバンド（ＩＢ）ドメイン１０２を用いて、入来トラフィック、たとえばストレージネットワーク１０５からＩＢファブリック１０４へのディスク読取データを処理することができる。

【００１３】

図１に示すように、ファイバチャネル（ＦＣ）ドメイン１０１は、物理ホストバスアダプタ（ＨＢＡ）１０３に接続することができる。物理ＨＢＡ１０３は、たとえばＦＣ命令を用いて、ディスク読取操作を実行することができ、たとえば周辺機器相互接続エクスプレス（ＰＣＩエクスプレスまたはＰＣＩｅ）命令を用いて、データおよびコンテキストをＦＣドメイン１０１に送信することができる。

【００１４】

ＦＣドメイン１０１は、ＦＣコンテキストリスト１２１を保存することができる。ＦＣコンテキストリスト１２１は、さまざまな仮想ホストバスアダプタ（ｖＨＢＡ）に関連する情報およびコンテキストを含むことができる。また、ＦＣドメイン１０１は、受信したディスク読取データおよび／またはコンテキストを外部メモリ、たとえば、シンクロナスダイナミックランダムアクセスメモリ（ＳＤＲＡＭ）１１１に記憶することができる。

【００１５】

図１に示すように、ＦＣドメイン１０１とＩＢドメイン１０２とは、シリアル相互接続を介して直接に接続されている。ＩＢドメイン１０２は、ＦＣドメイン１０１からＦＣデータおよびコンテキストを受信することができ、シーケンスの順序付けおよびコンテキストの管理のために、受信したＦＣデータおよびコンテキストをＩＢコンテキストリスト１２２内の異なるキューペア（ＱＰ）にマッピングすることができる。また、ＩＢドメイン１０２は、受信したディスク読取データおよびコンテキストを外部メモリ、たとえばＳＤＲＡＭ１１２に保存することができる。受信したディスク読取データおよびコンテキストは、ＩＢフォーマットであってもよい。次に、ＩＢドメイン１０２は、これらの情報をＩＢファブリック１０４に転送することができる。

【００１６】

このように、システムは、複数の異なるメモリインターフェイスを用いて、ストレージネットワーク１０５からＩＢファブリック１０４への入来トラフィックを処理することができる。

【００１７】
単一のメモリインターフェイス
本発明の一実施形態によれば、システムは、単一のメモリインターフェイスを用いて、Ｉ／Ｏ仮想化をサポートする、たとえば入来トラフィックおよび送出トラフィックの両方のためにＩＢファブリック上に作成された異なる仮想ＨＢＡ用の並列ＦＣコンテキストを管理することができる。
【００１８】
図２は、本発明の一実施形態に従って、単一のメモリインターフェイスを用いて、入力／出力（Ｉ／Ｏ）仮想化をサポートすることを示す図である。図２に示すように、Ｉ／Ｏ装置２００は、単一のＦＣ／ＩＢドメイン２０１を表すチップを用いて、ストレージネットワーク２０５からＩＢファブリック２０４への入来トラフィック、たとえばディスク読取データを処理することができる。
【００１９】
単一のＦＣ／ＩＢドメイン２０１は、物理ホストバスアダプタ（ＨＢＡ）２０３に直接に接続することができる。物理ＨＢＡ２０３は、ＦＣ命令を用いて、ディスク読取操作を実行することができる。物理ＨＢＡ２０３は、ＰＣＩｅ命令を用いて、ディスク読取データおよびコンテキストをＦＣ／ＩＢドメイン２０１に送信することもできる。その後、ＦＣ／ＩＢドメイン２０１は、ＩＢプロトコルを用いて、受信したディスク読取データおよびコンテキストをＩＢファブリック２０４に送信することができる。
【００２０】
図２に示すように、ＦＣ／ＩＢドメイン２０１は、ｖＨＢＡ／ＱＰ情報リスト２２０を保存することができる。ｖＨＢＡ／ＱＰ情報リスト２２０は、受信したＦＣデータおよびコンテキストをＩＢコンテキストリスト内の異なるキューペア（ＱＰ）にマッピングすることができる。また、ＦＣ／ＩＢドメイン２０１は、たとえばＩＢフォーマットのディスク読取データおよびコンテキストを外部メモリ、たとえばＳＤＲＡＭ２１０に保存することができる。
【００２１】
本発明の一実施形態によれば、ＦＣコンテキストリストをＩＢ信頼性のある接続（ＲＣ）キューペア（ＱＰ）リストに合併することによって、２つの異なるメモリインターフェイスの代わりに、単一のメモリインターフェイスを使用することができる。たとえば、システムは、一時コンテキストリストをＩＢドメインにマッピングする前に、外部メモリバッファのために、この一時コンテキストリストの動的リストを保有することができる。この手法は、２つの異なる外部メモリを使用することを回避することができ、ＩＢドメインからバックプレシャーメッセージ（back pressure message）をＦＣドメインに送信することを回避することができる。したがって、システムは、同一のデータおよび／またはコンテキストを複数回格納することを回避することができ、遅延を改善する。さらに、２つの異なるチップおよびメモリインターフェイスの代わりに、単一のチップおよびメモリインターフェイスを使用することは、システムのコストを低減することができる。
【００２２】
また、システムは、２つの異なるドメイン間に通信を行うために、外部（たとえば、ベンダ固有）インターフェイスに依存しない。単一のメモリインターフェイスが使用されるため、ＦＣ／ＩＢドメイン２０１は、バッファサイズを知り、外部メモリ、たとえばＳＤＲＡＭ２１０のバッファを超過することを回避することができる。単一のメモリインターフェイス手法によって、ｖＨＢＡがダウンした場合、より良いフラッシュ操作を行うことができる。ＩＢドメインとＦＣドメインとの間にメッセージの送受信がないため、フラッシュ操作を迅速かつ奇麗に行うことができる。
【００２３】
図３は、本発明の一実施形態に従って、単一のメモリインターフェイスを用いて、入来トラフィックを処理するための体系３００をサポートすることを示す図である。図３に示すように、単一のメモリインターフェイスに関連付けられているＦＣ／ＩＢドメイン３２

０を用いて、ストレージネットワークに接続している物理ホストバスアダプタ（ＨＢＡ）３３０からＩＢファブリック上のサーバ３１０への入来トラフィックを処理することができる。

【００２４】

ステップ３０１において、サーバ３１０は、たとえば、初期化ブロックをＲＣ送信メッセージとしてＦＣ／ＩＢドメイン３２０に送信することによって、ディスク読取操作を始動することができる。次に、ステップ３０２において、ＦＣ／ＩＢドメイン３２０は、メッセージを受信したことをサーバ３１０に知らせることができる。

【００２５】

続いて、ステップ３０３において、サーバ３１０は、記述子リングの書込インデックスを更新することができ、１つ以上の新たな入力／出力制御ブロック（ＩＯＣＢ）が存在していることをＦＣ／ＩＢドメイン３２０に知らせることができる。次に、ステップ３０４において、ＦＣ／ＩＢドメイン３２０は、メッセージを受信したことをサーバ３１０に知らせることができる。

【００２６】

また、ＦＣ／ＩＢドメイン３２０は、受信した書込インデックス値を読取インデックス値と比較することができる。値が異なる場合、ＦＣ／ＩＢドメイン３２０は、ステップ３０５において、ＲＤＭＡ読取命令を用いて、サーバ３１０から１つ以上のＩＯＣＢを取得しようとする。よって、ステップ３０６において、サーバ３１０は、１つ以上のＩＯＣＢをＲＤＭＡ読取応答データとしてＦＣ／ＩＢドメイン３２０に送信することができる。

【００２７】

ＦＣ／ＩＢドメイン３２０は、サーバ３１０からＩＯＣＢを受信すると、利用可能なコンテキストが存在する場合に、このコンテキストを開くことができる。本明細書において、コンテキストは、オンチップメモリを用いてチップ上に保存された特定命令の状態を示すものである。その後、ＦＣ／ＩＢドメイン３２０は、このＩＯＣＢ命令を物理ＨＢＡ３３０にプッシュすることができる。

【００２８】

たとえば、ステップ３０７において、ＦＣ／ＩＢドメイン３２０は、ポインタ、たとえば応答書込インデックスを更新することができる。この応答書込インデックスは、要求したＩＯＣＢが利用可能であることをＨＢＡ３３０に表示する。次に、ステップ３０８において、ＨＢＡ３３０は、ＩＯＣＢリクエストの読取を試行することができ、ステップ３０９において、ＦＣ／ＩＢドメイン３２０は、ＩＯＣＢリクエスト読取データをＨＢＡ３３０に送信することができる。それに応じて、ＨＢＡ３３０は、ディスク読取操作を実行することができる。

【００２９】

本発明の一実施形態によれば、並列サーバのＩＯＣＢ命令を処理するために、上記のステップ３０１～３０９は、同時に行うことができる。すなわち、ＦＣ／ＩＢドメイン３２０は、複数の並列コンテキストを同時に保存および処理することができる。

【００３０】

また、ＨＢＡ３３０は、ステップ３１１～３１９において、ディスク読取データをＦＣ／ＩＢドメイン３２０に送信することができる。これに応じて、ＦＣ／ＩＢドメイン３２０は、ステップ３２１～３２９において、ＲＤＭＡ書込操作を実行することによって、ディスク読取データをＩＢファブリック上のサーバ３１０送信することができる。

【００３１】

本発明の一実施形態によれば、システムは、ディスク読取データが完全にサーバ３１０またはホストに転送されたことを確保することができる。ステップ３３１において、サーバ３１０は、ディスク読取データの受信を確認するために、メッセージをＦＣ／ＩＢドメイン３２０に送信することができる。

【００３２】

さらに、ステップ３３２において、ディスク読取データが完全に送信された場合、物理

ＨＢＡ３３０は、対応するＩＯＣＢリクエストが完全に処理されたことを示すＩＯＣＢ応答をＦＣ／ＩＢドメイン３２０に送信することができる。これに応じて、ステップ３３３において、ＦＣ／ＩＢドメイン３２０は、ＲＣ送信メッセージを用いて、サーバ３１０にＩＯＣＢ応答を送信することができる。

【００３３】

最後に、ステップ３３４において、サーバは、ＩＯＣＢ応答の受信を確認することができ、ステップ３３５において、ＦＣ／ＩＢドメイン３２０は、ポインタ、たとえばＩＯＣＢ応答がサーバ３１０に送信されたことをＨＢＡ３３０に示す応答読取インデックスを更新することができる。

【００３４】

本発明の一実施形態によれば、ＦＣ／ＩＢドメイン３２０は、データパス内の異なる種類の入来トラフィック、たとえば、ｖＨＢＡ上のコンテキストのためのＲＤＭＡ読取リクエスト、物理ＨＢＡからのディスク読取データ、およびｖＨＢＡ上のコンテキストのために、物理ＨＢＡから受信したＩＯＣＢ応答を処理することができる。本実施形態において、ディスク書込データを取得するためのＲＤＭＡ読取リクエストは、ＦＣ／ＩＢドメイン３２０によって内部で生成することができ、ディスク読取データおよびＩＯＣＢ応答は、ＰＣＩｅバス（PCI-Express bus）を介して、物理ＨＢＡから受信することができる。

【００３５】

図４は、本発明の一実施形態に従って、Ｉ／Ｏ装置上でディスク読取操作を開始することを示す図である。図４に示すように、Ｉ／Ｏ装置４００、たとえばＦＣ／ＩＢドメイン４０１を表すチップは、ＩＢファブリック上のサーバ４０２から書込インデックス４１２を取得することができる。

【００３６】

ＦＣ／ＩＢドメイン４０１は、取得した書込インデックス４１２の値を読取インデックス値のコピーと比較することができる。値が異なる場合、ＦＣ／ＩＢドメイン４０１は、ＲＤＭＡ読取命令４１３を用いて、サーバ４０２から１つ以上の要求したＩＯＣＢ４１１を取得することができる。これらのＲＤＭＡ読取命令４１３は、ＩＢフォーマットに変換することができ、ＦＣ／ＩＢドメイン４０１に関連付けられた外部入来メモリ４１０内の空きバッファプール４２０に保存することもできる。

【００３７】

本実施形態において、ＲＤＭＡ読取命令４２１を入来ＤＲＡＭ４１０に格納する前に、キューロジックは、ＲＤＭＡ読取リクエストに利用可能なバッファが外部入来メモリ４１０に存在することを確認することができる。その後、ＦＣ／ＩＢドメイン４０１は、物理ＨＢＡ４０３からサーバ４０２への入来トラフィックを処理することができる。

【００３８】

図５は、本発明の一実施形態に従って、Ｉ／Ｏ装置上で要求したＩＯＣＢを抽出することを示す図である。図５に示すように、Ｉ／Ｏ装置５００、たとえばＦＣ／ＩＢドメイン５０１を表すチップは、ＩＢファブリック上のサーバ５０２からＲＤＭＡ読取応答データを受信することができる。

【００３９】

ＩＢプロトコルを用いて、予期通りにサーバ５０２からＲＤＭＡ応答読取データを完全に受信すると、ＦＣ／ＩＢドメイン５０１は、外部入来メモリ５１０内の空きバッファプール５２０に保存されたＲＤＭＡ読取リクエスト５２１をキューから除外することができる。その後、ＦＣ／ＩＢドメイン５０１は、受信したＲＤＭＡ読取応答データ５１２を保存されたＲＤＭＡ読取リクエスト５２１と比較することができる。

【００４０】

また、ＦＣ／ＩＢドメイン５０１は、受信したＲＤＭＡ読取応答データ５１２を解析することができ、ＩＯＣＢリクエストをＨＢＡ５０３に転送する前に、ＲＤＭＡ読取応答データ５１２に含まれたＩＯＣＢリクエスト５１１を抽出することができる。

【００４１】

図６は、本発明の一実施形態に従って、Ｉ／Ｏ装置上でディスク読取データを処理することを示す図である。図６に示すように、Ｉ／Ｏ装置６００、たとえばＦＣ／ＩＢドメイン６０１を表すチップは、ＩＯＣＢリクエスト６１３をＨＢＡ６０３に転送する前に、ＩＯＣＢリクエスト６１３用のコンテキスト６１２を開くことができる。

【００４２】

本発明の一実施形態によれば、ＦＣ／ＩＢドメイン６０１は、ＩＯＣＢリクエスト６１３用のコンテキスト６１２を開く前に、ＩＯＣＢリクエスト命令６１３用のＨＢＡ６０３からのディスク読取データ６１１を格納するのに十分なスペース（たとえば、外部入来メモリ６１０に保留したＤＲＡＭスペース６２１）を有することを確認することができる。したがって、システムは、（たとえば、ＩＣＯＢリクエスト命令６１３内の）ディスク読取命令が発行されると、ＦＣ／ＩＢドメイン６０１が物理ＨＢＡ６０３にバックプレシャーを与えないことを確認することができる。

【００４３】

ＦＣ／ＩＢドメイン６０１から、ＩＣＯＢリクエスト命令６１３内のディスク読取命令を受信した後、ＨＢＡ６０３は、（たとえば、ＦＣプロトコルを用いて）ストレージ上で、実際のディスク読取操作を実行することができる。ＨＢＡ６０３は、ＰＣＩ／ＰＣＩｅ書込トランザクションを用いて、ディスク読取データ６１１をＦＣ／ＩＢドメイン６０１に返送することができる。

【００４４】

ＦＣ／ＩＢドメイン６０１は、コンテキストを開くときに、ディスク読取命令のために外部入来メモリ６１０の空きバッファプール６２０にスペース６２１を保留しているため、受信したディスク読取データ６１１を外部入来メモリ６１０内のパケットバッファに書込む操作を開始することができる。また、ＦＣ／ＩＢドメイン６０１は、ディスク読取データ６１１を外部入来メモリ６１０内のパケットバッファに書込む前に、ＩＢヘッダおよびシーケンス番号をディスク読取データ６１１のために受信したパケットに追加することができる。よって、ディスク読取データ６１１のために受信した格納パケットは、ＩＢフォーマットであってもよい。

【００４５】

また、ＦＣ／ＩＢドメイン６０１は、完全メッセージ（たとえば、ＲＤＭＡ読取リクエスト）またはＩＢ最大伝送ユニット（ＭＴＵ）パケット（たとえば、ＲＤＭＡ書込専用パケット）が利用可能な場合、格納されたディスク読取データ６１１を外部入来メモリ６１０に読出すことができる。その後、ＦＣ／ＩＢドメイン６０１は、外部入来メモリ６１０内の空きバッファプール６２０から読出されたＩＢパケットをディスク読取データ６３１としてＩＢファブリック上のサーバ６０２に送信することができる。

【００４６】

図７は、本発明の一実施形態に従って、Ｉ／Ｏ装置上でディスク読取操作を完了する処理を示す図である。図７に示すように、たとえば、Ｉ／Ｏ装置７００、たとえばＦＣ／ＩＢドメイン７０１を表すチップを用いて、ＩＢファブリック上の物理ＨＢＡ７０３からサーバ７０２への入来トラフィックを処理することができる。

【００４７】

ディスク読取データが完全に送信された場合、ＨＢＡ７０３は、コンテキスト７１２に関連付けられた対応のＩＯＣＢリクエストが完全に処理されたことを示すＩＯＣＢ応答７１１をＦＣ／ＩＢドメイン７０１に送信することができる。その後、ＦＣ／ＩＢドメイン７０１は、ＩＢヘッダおよびシーケンス番号をＩＯＣＢ応答７１１に追加することができ、ＩＯＣＢ応答７２１を外部入来メモリ７１０内の空きバッファプール７２０に格納することができる。

【００４８】

メッセージまたはパケットを送信する準備ができると、ＦＣ／ＩＢドメイン７０１は、ＩＢプロトコルを用いて、ホスト／サーバ７０２にＩＯＣＢ応答７２１を送信することができる。ＩＯＣＢ７１３応答を受信すると、ホスト／サーバ７０２は、ディスク読取ＩＯ

ＣＢリクエスト命令７３１がハードウェアによって完全に処理されたことを確認することができる。

【００４９】

また、ＦＣ／ＩＢドメイン７０１は、コンテキスト用のＩＯＣＢ応答７２１を送信した後、関連するコンテキスト７１２を閉じることができる（すなわち、状態メモリを消去し、外部入来メモリ７１０内の保留スペースを削除することができる）。

【００５０】

複数のコンテキスト

図８は、本発明の一実施形態に従って、複数の仮想ホストバスアダプタ（ｖＨＢＡ）を用いて、Ｉ／Ｏ仮想化をサポートすることを示す図である。図８に示すように、Ｉ／Ｏ装置８００、たとえばＦＣ／ＩＢドメイン８０１を表すチップを用いて、入来トラフィック８３０を処理することができる。入来トラフィック８３０は、物理ＨＢＡ８０３からＩＢファブリック上のサーバ８０２に転送された複数のパケット、たとえばパケット８３１～８３９を含むことができる。

【００５１】

また、ＦＣ／ＩＢドメイン８０１は、１つ以上のｖＨＢＡ、たとえば、ｖＨＢＡ　Ａ８５１、ｖＨＢＡ　Ｂ８５２およびｖＨＢＡ　Ｃ８５３をサポートすることができる。ｖＨＢＡ　Ａ８５１、ｖＨＢＡ　Ｂ８５２およびｖＨＢＡ　Ｃ８５３は、ＩＢサーバ８０２に関連付けられたキューペア（ＱＰ）、たとえばＱＰ　Ａ８４１、ＱＰ　Ｂ８４２およびＱＰ　Ｃ８４３にそれぞれ対応することができる。

【００５２】

さらに、ＦＣ／ＩＢドメイン８０１は、外部入来メモリ８１０を用いて、１つ以上の受信パケットを格納することができる。ＦＣ／ＩＢドメイン８０１は、単一のメモリインターフェイスの使用をサポートするために、ＦＣコンテキスト情報、たとえばｖＨＢＡ　Ａ８５１、ｖＨＢＡ　Ｂ８５２およびｖＨＢＡ　Ｃ８５３をＩＢコンテキストリスト、たとえばＱＰ　Ａ８４１、ＱＰ　Ｂ８４２およびＱＰ　Ｃ８４３に合併することができる。

【００５３】

図８に示すように、外部入来メモリ８１０は、空きバッファプール８２０を提供することができる。空きバッファプール８２０は、１つ以上のバッファリスト、たとえばバッファリストＡ８２１、バッファリストＢ８２２およびバッファリストＣ８２３を含む。バッファリストＡ８２１、バッファリストＢ８２２およびバッファリストＣ８２３の各々を用いて、特定のＱＰ（またはｖＨＢＡ）をターゲットする１つ以上の受信パケットを格納することができる。

【００５４】

たとえば、ＦＣ／ＩＢドメイン８０１は、ｖＨＢＡ　Ａ８５１に関連付けられたバッファリストＡ８２１内のＱＰ　Ａ８４１をターゲットするパケット８３２および８３９をキューに入れることができる。同様に、ＦＣ／ＩＢドメイン８０１は、ｖＨＢＡ　Ｂ８５２に関連付けられたバッファリストＢ８２２内のＱＰ　Ｂ８４２をターゲットするパケット８３３および８３８をキューに入れることができ、ｖＨＢＡ　Ｃ８５３に関連付けられたバッファリストＣ８２３内のＱＰ　Ｃ８４３をターゲットするパケット８３１をキューに入れることができる。

【００５５】

また、ＦＣ／ＩＢドメイン８０１は、受信した複数のパケット８３１～８３９の状態を保存することができる制御構造８１１を含むことができる。さらに、ＦＣ／ＩＢドメイン８０１は、読取ロジック８１２を用いて、格納されたパケット８３１～８３９のうち１つ以上を読出すことができる。

【００５６】

本発明の一実施形態によれば、ＦＣ／ＩＢドメイン８０１は、ＩＢドメイン内のＱＰのために、複数のディスク読取命令、ＲＤＭＡ読取リクエストおよびＲＣ送信命令の多重化をサポートするために、ｖＨＢＡ　Ａ８５１～ｖＨＢＡ　Ｃ８５３内の複数のコンテキスト

を開くことができる。

【００５７】

図９は、本発明の一実施形態に従って、仮想ホストバスアダプタ（ｖＨＢＡ）において、複数のコンテキストをサポートすることを示す図である。図９に示すように、Ｉ／Ｏ装置、たとえばＦＣ／ＩＢドメイン９００を表すチップは、たとえば、ＱＰ９０４のために物理ＨＢＡ９０３上で複数のディスク読取命令を実行するために、単一のｖＨＢＡ９０１内の複数のコンテキスト、たとえばコンテキストＩ　９１０およびコンテキストＩＩ　９２０を開くことができる。

【００５８】

たとえば、コンテキストＩ　９１０は、物理ＨＢＡ９０３から受信したいくつかのパケット、たとえばＣ１Ｄ１　９１１、Ｃ１Ｄ２　９１２およびＣ１Ｄ３　９１３を含むことができる。Ｃ１Ｄ１　９１１は、コンテキストＩ　９１０用のディスク読取データＤ１を含み、Ｃ１Ｄ２　９１２は、コンテキストＩ　９１０用のディスク読取データＤ２を含み、Ｃ１Ｄ３　９１３は、コンテキストＩ　９１０用のディスク読取データＤ３を含むことができる。

【００５９】

また、コンテキストＩＩ　９２０は、物理ＨＢＡ９０３から受信したいくつかのパケットＣ２Ｄ１　９２１およびＣ２Ｄ２　９２２を含むことができる。Ｃ２Ｄ１　９２１は、コンテキストＩＩ　９２０用のディスク読取データＤ１を含み、Ｃ２Ｄ２　９２２は、コンテキストＩＩ　９２０用のディスク読取データＤ２を含むことができる。

【００６０】

また、ＦＣ／ＩＢドメイン９００は、物理ＨＢＡ９０３から受信したパケットをＩＢファブリック上のＱＰ９０４に送信する前に、このパケットに対応のシーケンス番号（ＰＳＮ）および異なるＩＢヘッダを追加することができる。

【００６１】

本発明の一実施形態によれば、同一のＱＰたとえば（ｖＨＢＡ９０１に関連付けられた）ＱＰ９０４をターゲットするすべてのパケットは、ＩＢドメイン内で単一ＰＳＮスペース９０２を共有することができる。図９に示すように、ＰＳＮスペース９０２において、パケットをＰ０、Ｐ１、・・・、Ｐ（Ｎ）の順序に編成することができる。ここでは、Ｐ１＝Ｐ０＋１、Ｐ２＝Ｐ１＋１、・・・、Ｐ（Ｎ）＝Ｐ（Ｎ－１）＋１。

【００６２】

一方、ＩＢドメイン内でＰＳＮスペース９０２を共有する場合、動作中に、ＩＢドメイン内のＰＳＮ番号割当体系を用いて発信パケットの順序を変更することができないため、異なるコンテキストにおいて、ＩＢヘッダおよびシーケンス番号を単一のメモリインターフェイスに基づいてＨＢＡ８０３から受信したパケットに追加することを複雑化する可能性がある。

【００６３】

図９に示すように、コンテキストＩ　９１０用のディスク読取データが完全に処理される前に、コンテキストＩＩ　９２０用のディスク読取データが入来すると、ｖＨＢＡ９０１にヘッドラインブロッキングという問題が生じる可能性がある。たとえば、システムが別のディスク書込操作を行っている最中に、ＦＣ／ＩＢドメイン９００がディスク書込操作のためにＲＤＭＡ読取リクエストをスケジュールしようとするときに、このような問題が生じる。

【００６４】

図１０は、本発明の一実施形態に従って、オンチップメモリを用いて、外部メモリ上でキューに入れられたパケットの状態を保存することを示す図である。図１０に示すように、Ｉ／Ｏ装置１０００、たとえばＦＣ／ＩＢドメイン１０００を表すチップは、単一のｖＨＢＡ／ＱＰ内の複数のコンテキスト、たとえばコンテキストＩ　１０１０およびコンテキストＩＩ　１０２０を開くことができる。各コンテキストは、１つ以上のパケットを含むことができる。たとえば、コンテキストＩ　１０１０は、パケットＣ１Ｄ１　１０１１

、Ｃ１Ｄ２　１０１２およびＣ１Ｄ３　１０１３を含み、コンテキストＩＩ　１０２０は、パケットＣ２Ｄ１　１０２１およびＣ２Ｄ２　１０２２を含むことができる。

【００６５】

本発明の一実施形態によれば、ＦＣは／ＩＢドメイン１０００は、シーケンス番号およびさまざまなＩＢヘッダを物理ＨＢＡ１００５から受信した各パケットに追加することができる。ＩＢヘッダは、特定のパケットをＩＢファブリック１００４に送信するときに、この特定のパケットに適用することができるＩＢ命令を含むことができる。

【００６６】

たとえば、コンテキスト用のディスク読取データのサイズがＩＢ最大伝送ユニット（ＭＴＵ）のサイズと同一である場合、ＩＢヘッダ内のキューに入れられたＩＢ命令は、コンテキストメモリに指定された仮想アドレス（ＶＡ）を有する「ＲＤＭＡ書込専用」命令にすることができる。逆に、所定のコンテキスト用のディスク読取データのサイズがＩＢＭＴＵのサイズよりも大きい場合、ディスク読取データを複数のパケットに分割することができる。各パケットのＩＢヘッダ内のキューに入れられた命令は、コンテキストメモリにより指定されたＶＡを有する「最初のＲＤＭＡ書込」命令、「中間のＲＤＭＡ書込」命令および「最後のＲＤＭＡ書込」命令のいずれか１つであってもよい。ディスク読取データのサイズに応じて、「中間のＲＤＭＡ書込」命令を有するようにパケットをキューに入れなくてもよく、または「中間のＲＤＭＡ書込」命令を有するように複数のパケットをキューに入れてもよい。

【００６７】

図１０に示すように、初期ＰＳＮは、Ｐ０であり、ディスク読取データは、ＩＢ　ＭＴＵよりも大きい。第１のパケット（Ｃ１Ｄ１　１０１１）を受信すると、システムは、ＰＳＮとしてのＰ０および「最初のＲＤＭＡ書込」命令をパケットに追加することができる。次のパケット（Ｃ１Ｄ２　１０１２）を受信すると、システムは、ＰＳＮとしてのＰ１（すなわち、Ｐ０＋１）および「中間のＲＤＭＡ書込」命令をパケットに追加することができる。第３のパケット（Ｃ２Ｄ１　１０２１）を受信すると、システムは、ＰＳＮとしてのＰ２（すなわち、Ｐ１＋１）および「最初のＲＤＭＡ書込」命令をパケットに追加することができる。さらに、パケット（Ｃ２Ｄ２　１０２２）を受信すると、システムは、ＰＳＮとしてのＰ３（すなわち、Ｐ２＋１）および「最後のＲＤＭＡ書込」命令をパケットに追加することができる。

【００６８】

しかしながら、上記一連の操作には、整合性問題が存在している。パケットＣ１Ｄ２　１０１２に追加されたＩＢ命令は、「中間のＲＤＭＡ書込」命令である。システムは、ディスク読取データのサイズに応じて、次のパケットに追加されたＩＢ命令が「中間のＲＤＭＡ書込」命令または「最後のＲＤＭＡ書込」命令のいずれかであると予期している。しかしながら、図１０に示すように、次のパケットＣ２Ｄ１　１０２１が異なるコンテキストＩＩ　１０２０からのものであるため、ＦＣ／ＩＢドメイン１０００は、新たな命令、たとえば、「最初のＲＤＭＡ書込」命令または「ＲＤＭＡ書込専用」命令をパケット（そのＰＳＮが正確であっても）に追加することができる。また、ＦＣ／ＩＢドメイン１０００が、別のコンテキストのために記述子の取得またはＲＣ送信ＩＯＣＢ応答を行うため、ＲＤＭＡ読取リクエストをキューに入れようとする際に、同様の問題が生じる可能性もある。

【００６９】

この問題を解決するために、システムは、既にキューに入れられたパケットに関連付けられたＩＢ命令を更新することができる。たとえば、ＦＣ／ＩＢドメイン１０００がパケットＣ２Ｄ１　１０２１を受信した後、システムは、Ｃ１Ｄ２　１０１２のＩＢ命令、すなわち「中間のＲＤＭＡ書込」命令１００７を「最後のＲＤＭＡ書込」命令１００８に変更することができる。

【００７０】

本発明の一実施形態によれば、整合性を確保するために、ＦＣ／ＩＢドメイン１０００

は、オンチップメモリ１００２を用いて、外部メモリ１００１上でキューに入れられたパケット１００６の状態１００９を格納することができる。

【００７１】

このオンチップメモリ１００２は、異なるパースペクティブ（perspective）に有益であり得る。まず、パケットがキューに入れられ、必要に応じてパケットに関連付けられたＩＢ命令が更新された場合に限り、外部メモリからパケットを読取りおよびパケットをホストに送信するように、読取ロジックを確保することができる。次に、複数のコンテキストをサポートするために、既にキューに入れられたパケットに関連付けられたＩＢ命令を必要に応じて更新することができる。

【００７２】

たとえば、オンチップメモリ１００２は、２ビットの幅（および６４Ｋの深さ）を有することができる。オンチップメモリ１００２内のエントリの第１ビット、たとえばビット０は、ＩＢ命令を変更または更新する必要があるか否かを示し、第２ビット、たとえばビット１は、ＦＣ／ＩＢドメイン１０００の読取ロジックが外部メモリ１００１からキューに入れられたパケットを取出すことができるか否かを示すことができる。

【００７３】

以下の表１は、一連のパケットが到着する場合に、例示のオンチップメモリに格納された異なるパケットの状態を示している。

【００７４】

【表１】

物理ＨＢＡからのパケット	外部メモリ上で、パケットをキューに入れる	現在のパケットをキューに入れるときのオンチップメモリ	次のパケットをキューに入れるときのオンチップメモリ
C2D1	最初の RDMA 書込 PSN P0, C2D1	ビット 0 : 0 ビット 1 : 0	ビット 0 : 0 ビット 1 : 1
C2D2	中間の RDMA 書込 PSN P1, C2D2	ビット 0 : 0 ビット 1 : 0	ビット 0 : 1 ビット 1 : 1
C1D1	最初の RDMA 書込 PSN P2, C1D1	ビット 0 : 0 ビット 1 : 0	ビット 0 : 0 ビット 1 : 1
C1D2	最後の RDMA 書込 PSN P3, C1D2	ビット 0 : 0 ビット 1 : 1	ビット 0 : 0 ビット 1 : 1

【００７５】

上記の表１に示すように、第１のパケット、すなわちＣ２Ｄ１をキューに入れるときに、オンチップステートメモリは、２＇ｂ００である。このとき、読取ロジックは、このパケットを読取ることができない。その理由は、パケットが他のコンテキストから来る場合、システムがこのパケットの命令を変更する必要があるからである。

【００７６】

次のパケットＣ２Ｄ２が来るときに、前のパケット、すなわちＣ２Ｄ１のオンチップ状態は、２＇ｂ１０に変更されている。このとき、パケットは、正常にキューに入れられ、読取ロジックは、このパケットを読出すことができる。この場合、Ｃ２Ｄ２が同様のコンテキストＩＩ（Ｃ２）から来たため、命令を変更する必要がない。

【００７７】

さらに、第３のパケット、すなわちＣ１Ｄ１が来るときに、Ｃ２Ｄ２のオンチップメモリの状態は、２＇ｂ１１に変更されている。このとき、パケットは、キューに入れられ読取側で命令を変更する必要がある。読取ロジックは、このパケットを読出すことができ、このパケットを送信する前に、命令を「中間のＲＤＭＡ書込」命令から「最後のＲＤＭＡ書込」命令に変更することができる。

【００７８】

図１１は、本発明の一実施形態に従って、単一のメモリインターフェイスを用いて、入

力／出力（Ｉ／Ｏ）仮想化をサポートすることを示す例示的なフローチャートである。図１１に示すように、ステップ１１０１において、システムは、ネットワークファブリック上のサーバに関連付けられ、且つ、複数のパケットバッファを含む外部メモリに関連付けられているチップを提供することができる。また、ステップ１１０２において、システムは、物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを含む１つ以上のパケットの状態をオンチップメモリに保存することができる。また、ステップ１１０３において、システムは、外部メモリ上の複数のパケットバッファ内の１つ以上のパケットをキューに入れ、１つ以上のパケットの状態に基づいて、外部メモリから１つ以上のパケットを読出し、およびネットワークファブリック上のサーバに１つ以上のパケットを送信するように、チップを動作させることができる。

【００７９】

ハイブリッドリンクリスト構造

本発明の一実施形態によれば、システムは、ハイブリッドリンクリスト構造を用いて、仮想ホストバスアダプタ（ｖＨＢＡ）内の複数のコンテキストに関連付けられた入来トラフィックを処理することができる。このハイブリッドリンクリスト構造は、主要リンクリストおよび一時リンクリストを含むことができる。

【００８０】

図１２は、本発明の一実施形態に従って、空きバッファプールを用いて、複数の仮想ホストバスアダプタ（ｖＨＢＡ）をサポートすることを示す図である。図１２に示すように、Ｉ／Ｏ装置１２００、たとえばＦＣ／ＩＢドメイン１２０４を表すチップは、たとえば、空きバッファプール１２１０を用いて、異なる仮想ホストバスアダプタ（ｖＨＢＡ）、たとえばｖＨＢＡ　Ｉ　１２０１およびｖＨＢＡ　ＩＩ　１２０２に関連付けられ得る入来トラフィック１２０３を処理することができる。

【００８１】

また、各ｖＨＢＡは、入来トラフィック１２０３から受信した各種パケットを空きバッファプール１２１０内のパケットバッファに格納するために、バッファポインタの１つ以上のリンクリストを保存することができる。たとえば、ｖＨＢＡ　Ｉ　１２０１は、主要リンクリストＩ　１２１１を保存することができ、ｖＨＢＡ　ＩＩ　１２０２は、主要リンクリストＩＩ　１２１２を保存することができる。

【００８２】

図１３は、本発明の一実施形態に従って、ハイブリッドリンクリスト構造を用いて、さまざまなディスク読取操作をサポートすることを示す図である。図１３に示すように、Ｉ／Ｏ装置１３００、たとえばＦＣ／ＩＢドメインを表すチップは、複数のディスク読取操作を並列で実行するために、単一のｖＨＢＡ１３０３内の複数のコンテキスト、たとえばコンテキストＡ　１３０１およびコンテキストＢ　１３０２を開くことができる。

【００８３】

本発明の一実施形態によれば、ＦＣ／ＩＢドメインは、各ディスク読取操作のために、オンチップメモリ上で、コンテキストテーブルを定義することができる。たとえば、ＦＣ／ＩＢドメインは、コンテキストＡ　１３０１のディスク読取データを受信した場合、オンチップメモリ１３１０上で、コンテキストテーブルＡ　１３１１を開くことができる。コンテキストテーブルＡ　１３１１は、空きバッファプール１３２０から割当てられた１つ以上のパケットバッファに指向するバッファポインタの一時リンクリストを保存することができる。また、コンテキストテーブルＡ　１３１１は、所定のトランザクション用の仮想アドレスを保存することもできる。

【００８４】

コンテキストＡ　１３０１の入来データが来るときに、ディスクデータをＩＢヘッダおよびＰＳＮ番号とともに外部ＳＤＲＡＭメモリ内の空きバッファプール１３２０に書込むことができる。図１３に示すように、コンテキストメモリ内のバッファポインタの一時リンクリスト１３２１が更新されるが、所定のｖＨＢＡ用のバッファポインタの主要リンクリスト１３３０が変更されない。

【００８５】
本発明の一実施形態によれば、異なるディスク読取操作のために、複数のコンテキストを開くことができる。新たに開いたコンテキストのディスク読取データが来るときに、システムは、外部ＳＤＲＡＭメモリから以前に開かれたコンテキストに格納されたデータを読出すことができ、必要に応じて、コンテキストリスト末尾のＩＢヘッダ内の命令を更新することができる。たとえば、所定のコンテキストのためのディスク読取データがＩＢＭＴＵよりも大きいときに、ＩＢヘッダにキューに入れられた命令が「最初のＲＤＭＡ書込」命令である場合、その命令を「ＲＤＭＡ書込専用」命令に変更することができ、ＩＢヘッダにキューに入れられた命令が「中間のＲＤＭＡ書込」命令である場合、その命令を「最後のＲＤＭＡ書込」命令に変更することができる。

【００８６】
図１３に示すように、ＦＣ／ＩＢドメインが異なるコンテキストＢ　１３０２からデータを受信する場合、一時リンクリスト１３２１をｖＨＢＡ１３０３の主要リンクリスト１３３０に合併することができる。たとえば、ｖＨＢＡ１３０３の主要リンクリスト１３３０のテールポインタを一時リンクリスト１３２１のヘッドポインタに変更することができ、一時リンクリスト１３２１のテールポインタを主要リンクリスト１３３０の新たなテールポインタになる。よって、新たなコンテキストのデータは、新たなＰＳＮを有する新たなコンテキストメモリに書込まれることができる。それに応じて、そのコンテキストの一時ポインタは、更新されることができる。

【００８７】
同様に、システムは、新たなコンテキスト内の「ＲＤＭＡ書込専用」命令、「最後のＲＤＭＡ書込」命令、「送信専用」命令、およびＲＤＭＡ読取リクエストなどの命令を実行する必要がある場合、以前に開かれたコンテキストを閉じ、一時リンクリスト１３２１を主要リンクリスト１３３０に合併することができる。

【００８８】
図１４は、本発明の一実施形態に従って、ハイブリッドのリンクリスト構造を用いて、ヘッドラインブロッキングを回避することを示す図である。図１４に示すように、コンテキスト１４０１を閉じ、ｖＨＢＡに１４０３の主要リンクリスト１４３０を更新した後、Ｉ／Ｏ装置１４００、たとえばＦＣ／ＩＢドメインを表すチップは、コンテキストＢ　１４０２のために、オンチップメモリ１４１０内の新たなコンテキストテーブルＢ　１４１２を開くことができる。コンテキストテーブルＢ　１４１２は、空きバッファプール１４２０から割当てられたパケットバッファに指向するバッファポインタを含む新たな一時リンクリストＢ　１４２２を保存することができる。

【００８９】
図１４に示すように、コンテキストＢ　１４０２内の受信したディスク読取データ（またはＲＤＭＡ読取リクエスト）に対する処理は、コンテキストＡ　１４０１内の受信したディスク読取データの処理によってブロックされるため、システムは、ヘッドラインブロッキングを回避することができる。よって、所定のｖＨＢＡ内の異なるコンテキストのディスク読取データを並列に処理することができる。

【００９０】
図１５は、本発明の一実施形態に従って、ハイブリッドリンクリスト構造を用いて、ヘッドラインブロッキングを回避することを示す例示的なフローチャートである。図１５に示すように、ステップ１５０１において、システムは、ネットワーク環境内の１つ以上の仮想ホストバスアダプタ（ｖＨＢＡ）に関連付けられた複数のパケットバッファを含む空きバッファプールを提供することができる。各々のｖＨＢＡは、空きバッファプール内の１つ以上のパケットバッファに指向するバッファポインタの主要リンクリストを保存する。ステップ１５０２において、システムは、入力／出力（Ｉ／Ｏ）装置に関連付けられたオンチップメモリ上で、コンテキストテーブルを定義することができる。コンテキストテーブルは、ディスク読取操作のために空きバッファプールから割当てられた１つ以上のパケットバッファに指向するバッファポインタの一時リンクリストを保存する。ステップ１

５０３において、システムは、Ｉ／Ｏ装置がディスク読取操作を実行する物理ホストバスアダプタ（ＨＢＡ）からディスク読取データを受信すると、コンテキストテーブルを開き、一時リンクリストを更新し、およびコンテキストテーブルを閉じたと、バッファポインタの一時リンクリストをバッファポインタの主要リンクリストに合併するように、Ｉ／Ｏ装置を動作させることができる。

【００９１】

一体化メモリ構造

本発明の一実施形態によれば、ディスク読取データがＨＢＡチップから送信されるときに、システムは、ディスク読取データを一体化メモリ構造内のさまざまなデータバッファに格納することができる。

【００９２】

図１６は、本発明の一実施形態に従って、Ｉ／Ｏ装置のために２次元リンクリスト構造をサポートすることを示す図である。図１６に示すように、システムは、空きバッファプール１６００内の２次元リンクリスト１６１０を用いて、入来パケットバッファを管理することができる。２次元リンクリスト１６１０は、複数のエントリを含むことができる。各エントリは、スーパーブロック（たとえば、スーパーブロック１６０１～１６０９）であってもよい。

【００９３】

本発明の一実施形態によれば、スーパーブロック１６０１～１６０９は、連続したメモリ位置に格納された複数のパケットを表すことができる。また、スーパーブロック１６０１～１６０９の各々は、バッファ管理のために、内部でパケットバッファリストに指向することができる。したがって、オンチップリソース使用の観点から、２次元リンクリスト構造は、非常に効率的である。システムは、バッファされるパケットの数を最大にしながら、オンチップメモリ上のリンクリストのサイズを最小にすることができる。

【００９４】

たとえば、さまざまなサイズの（オーバーヘッドを含む）ＩＢパケットを収容するために、２次元リンクリスト１６１０は、８Ｋのスーパーブロックを含むことができる。また、各スーパーブロックは、（各々が８ＫＢのサイズを有する）８つのパケットを保存することができる６４ＫＢ（５１２Ｋｂ）のサイズを有することができる。図１６に示すように、スーパーブロック１６０１は、８つのパケットバッファ、すなわち、パケットバッファ１６１１～１６１８を含むことができる。

【００９５】

２次元リンクリスト１６１０によって、ＦＣ／ＩＢドメインは、異なるＱＰをターゲットする読取データディスクをＩＢドメインに格納することができる。図１６に示すように、ＦＣ／ＩＢドメインは、異なるポインタを用いて、２次元リンクリスト１６１０内のさまざまなスーパーブロックのリンクリストにアクセスすることができる。たとえば、ＦＣ／ＩＢドメインは、スーパーブロック１６０２、スーパーブロック１６０４およびスーパーブロック１６０８を含むスーパーブロックのリンクリストに指向するＱＰ　Ａヘッドポインタ１６２１（および／またはＱＰ　Ａテールポインタ１６２２）を保存することができる。また、ＦＣ／ＩＢドメインは、スーパーブロック１６０６、スーパーブロック１６０５およびスーパーブロック１６０９を含むスーパーブロックのリンクリストに指向するＱＰ　Ｂヘッドポインタ１６２３（および／またはＱＰ　Ｂテールポインタ１６２４）を保存することができる。

【００９６】

本発明の一実施形態によれば、システムは、２次元ハイブリッドリンクリスト１６１０を所定のインフィニバンド（ＩＢ）ＲＣ　ＱＰ接続の１次元リンクリストと動的に合併することによって、外部ＤＲＡＭメモリの効率的な使用をサポートすることができる。したがって、システムは、小さいサイズのパケットを固定サイズのスーパーブロックに格納することによるメモリスペースの浪費を避けることができる。

【００９７】

たとえば、ＦＣ／ＩＢドメインは、ディスク読取リクエストを実行するために、空きバッファプール１６００にバッファの有無を照会することができる。空きバッファプール１６００に十分なパケットバッファがある場合に、ＦＣ／ＩＢドメインは、物理ＨＢＡにディスク読取ＩＯＣＢリクエストを発行することができる。ディスク読取リクエストに要求されたバッファは、空きバッファプール１６００に保留され、現在のコンテキストがＦＣ／ＩＢドメインによって解放されるまで、他の後続のリクエストに使用されない。

【００９８】

また、システムは、ＲＤＭＡ読取リクエストを保存するバッファ（たとえば、４Ｋバッファ）のリストを定義することができる。システムは、ＲＤＭＡ読取リクエストが発行されるたびに、ＲＤＭＡ読取リクエストに利用可能なスペースが外部メモリに保留されており、ＲＤＭＡ読取リクエストがＲＤＭＡ書込操作にブロックされないことを保証することができる。

【００９９】

２次元リンクリスト１６１０のみを使用する場合、システムは、ＲＤＭＡ読取リクエストに１２８個のキューペア（またはｖＨＢＡ）に共有される４Ｋのパケットバッファを提供するために、６４Ｋ（スーパーブロックのサイズ）×４Ｋ×１２８バイトのスペースをメモリに確保する必要がある。このスペースが８Ｋ（パケットバッファのサイズ）×４Ｋ×１２８バイトであるパケットバッファのメモリ使用量よりも実質的に多いため、この手法は、メモリを浪費する。

【０１００】

図１７は、本発明の一実施形態に従って、Ｉ／Ｏ装置のためにメモリの効率的な使用をサポートすることを示す図である。図１７に示すように、Ｉ／Ｏ装置１７００、たとえばＦＣ／ＩＢドメインを表すチップは、空きバッファプール１７０１を用いて、さまざまなパケットのキュー入れ（１７３０）をサポートすることができる。空きバッファプール１７０１は、スーパーブロック１７１１～１７１９を包含する２次元リンクリスト１７１０、およびパケットバッファ１７２１～１７２９を包含する１次元リンクリスト１７２０を含むことができる。この２次元リンクリスト１７１０は、図１６に示された２次元リンクリスト１６１０と類似してもよい。

【０１０１】

本発明の一実施形態によれば、空きバッファプール１７０１において、異なる種類のトランザクションをキューに入れることができる。たとえば、これらのトランザクションを用いて、ＲＤＭＡ書込命令１７４２およびＲＤＭＡ読取リクエスト１７４１を実行することができる。

【０１０２】

パケットをキューに入れた（１７３０）場合、トランザクションの種類に基づいて、２次元リンクリスト１７１０からまたは単一の１次元リンクリスト１７２０から、空きバッファを割当てることができる。

【０１０３】

また、システムは、さまざまなバッファされたパケットの状態を保存するために、チップ上でリンクリストの制御構造１７４０を保有することができる。制御構造１７４０は、たとえば、メモリスーパーブロック位置のヘッドポインタ（たとえば、１３ビットのＳＢＬＫＨＥＡＤ）、スーパーブロック内のパケットオフセット位置のヘッドポインタ（たとえば、３ビットのＰＫＴＨＥＡＤ）、メモリスーパーブロック位置のテールポインタ（たとえば、１３ビットのＳＢＬＫＴＡＩＬ）、スーパーブロック内のパケットオフセット位置のテールポインタ（たとえば、３ビットのＰＫＴＴＡＩＬ）、およびパケットバッファが２次元リンクリストから割当てられているかまたは１次元リンクリストから割当てられているかを示すフラグ（たとえば、１ビットのＬＩＳＴＴＹＰＥ）を格納することができる。また、制御構造１７４０は、ＱＰ／ｖＨＢＡの数に基づいて深さ情報を格納することができ、必要な制御情報に基づいて幅情報を格納することができる。

【０１０４】

本発明の一実施形態によれば、システムは、異なるキュー入れシナリオをサポートすることができる。

【０１０５】

キューに入れられたトランザクションがＲＤＭＡ書込命令用のものである場合、システムは、２次元リンクリスト１７１０からバッファまたはスーパーブロックを取得することができる。

【０１０６】

逆に、キューに入れられたトランザクションがＲＤＭＡ読取命令用のものである場合、以前にキューに入れられたトランザクションに割当てられたスーパーブロックにパケットバッファが残されていないときに、システムは、１次元リンクリスト１７２０からバッファを取得することができる。

【０１０７】

一方、ＲＤＭＡ書込操作が進行中に、ＲＤＭＡ読取リクエスト用のトランザクションがキューに入れられる場合もある。利用可能なパケットバッファが存在する場合、システムは、ＲＤＭＡ書込動作に割当てられたスーパーブロック内の現在のパケット位置で、ＲＤＭＡ読取リクエストをキューに入れることができる。

【０１０８】

また、特定のＱＰ／ｖＨＢＡのために、単一リンクリスト１７２０からのパケットバッファを保留することができる。単一リンクリスト１７２０から保留されたバッファは、ＲＤＭＡ書込パケットまたはＲＤＭＡ読取リクエストパケットのいずれかに使用されることができる。さらに、システムは、制御メモリ内のＬＩＳＴＴＹＰＥフィールドにフラグを付けることができる。よって、デキュー（dequeue）ロジックおよび／または読取ロジックは、１つのパケットが単一リンクリスト１７２０のキューに入れられたことを知ることができる。

【０１０９】

このように、システムは、効率的なパケット処理を実現することができ、外部メモリの浪費を回避することができる。

【０１１０】

図１８は、本発明の一実施形態に従って、コンピューティング環境において、効率的なパケット処理をサポートすることを示す例示的なフローチャートである。図１８に示すように、ステップ１８０１において、システムは、２次元リンクリストおよび１次元リンクリストを含む空きバッファプールをメモリに提供することができる。また、ステップ１８０２において、システムは、２次元リンクリストの各エントリが連続したメモリ位置に複数のパケットバッファを有すること、および１次元リンクリストの各エントリが単一のパケットバッファを有することを可能にする。次に、１８０３において、Ｉ／Ｏ装置は、物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを空きバッファプールに格納することができる。

【０１１１】

本発明の一実施形態は、コンピューティング環境において、Ｉ／Ｏ仮想化をサポートするためのシステムを提供する。このシステムは、コンピューティング環境内の１つ以上の仮想ホストバスアダプタ（ｖＨＢＡ）に関連付けられた複数のパケットバッファを含む空きバッファプールを含み、ｖＨＢＡの各々は、１つ以上のパケットバッファに指向するバッファポインタの主要リンクリストを空きバッファプールに保存し、入力／出力（Ｉ／Ｏ）装置に関連付けられたオンチップメモリ上で定義されたコンテキストテーブルを含み、コンテキストテーブルは、ディスク読取操作のために、空きバッファプールから割当てられた１つ以上のパケットバッファに指向するバッファポインタの一時リンクリストを保存する。Ｉ／Ｏ装置は、ディスク読取操作を実行する物理ホストバスアダプタ（ＨＢＡ）からすディスク読取データを受信すると、コンテキストテーブルを開き、バッファポインタの一時リンクリストを更新し、およびコンテキストテーブルが閉じられると、バッファポインタの一時リンクリストをバッファポインタの主要リンクリストに合併するように動作

する。
【０１１２】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、ディスク読取操作を開始するように、インフィニバンド（ＩＢ）ファブリック上のサーバを動作させる。
【０１１３】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、ＩＢヘッダとシーケンス番号とを物理ＨＢＡから受信した各パケットに追加する。
【０１１４】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、完全メッセージまたはＩＢ最大伝送ユニット（ＭＴＵ）パケットを受信すると、外部メモリに格納されたディスク読取データを読出すように動作する。
【０１１５】
上記に提供されたシステムにおいて、各仮想ＨＢＡは、ＩＢドメインにおいて、異なるパケットシーケンス番号（ＰＳＮ）スペースを保留する。
【０１１６】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、ｖＨＢＡに関連付けられた異なるディスク読取操作のために、異なるコンテキストテーブルを保存する。
【０１１７】
上記に提供されたシステムにおいて、コンテキストテーブルは、Ｉ／Ｏ装置が仮想ＨＢＡに関連付けられた別のコンテキストテーブルを開くと、閉じられる。
【０１１８】
上記に提供されたシステムにおいて、別のコンテキストテーブルは、空きバッファプールから割当てられた１つ以上のパケットバッファに指向するバッファポインタの新たな一時リンクリストを保存する。
【０１１９】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、仮想ＨＢＡが別のディスク読取操作からデータを受信すると、別のコンテキストテーブルを開く。
【０１２０】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、仮想ＨＢＡが書込専用命令、ＲＤＭＡ書込最終命令、送信専用命令およびＲＤＭＡ読取リクエスト命令のうち１つの命令を受信すると、別のコンテキストテーブルを開く。
【０１２１】
本発明の一実施形態は、コンピューティング環境において、効率的なパケット処理をサポートするための方法を提供する。この方法は、コンピューティング環境内の１つ以上の仮想ホストバスアダプタ（ｖＨＢＡ）に関連付けられた複数のパケットバッファを含む空きバッファプールを提供するステップを含み、ｖＨＢＡの各々は、１つ以上のパケットバッファに指向するバッファポインタの主要リンクリストを空きバッファプールに保存し、入力／出力（Ｉ／Ｏ）装置に関連付けられたオンチップメモリにおいて、コンテキストテーブルを定義するステップを含み、コンテキストテーブルは、ディスク読取操作のために、空きバッファプールから割当てられた１つ以上のパケットバッファに指向するバッファポインタの一時リンクリストを保存し、ディスク読取操作を実行する物理ホストバスアダプタ（ＨＢＡ）からディスク読取データを受信すると、コンテキストテーブルを開き、バッファポインタの一時リンクリストを更新し、およびコンテキストテーブルが閉じられると、バッファポインタの一時リンクリストをバッファポインタの主要リンクリストに合併するように、Ｉ／Ｏ装置を動作させるステップを含む。
【０１２２】
上記に提供された方法は、ディスク読取操作を開始するように、インフィニバンド（ＩＢ）ファブリック上のサーバを動作させる。
【０１２３】
上記に提供された方法は、ＩＢヘッダとシーケンス番号とをＨＢＡから受信した各パケ

ットに追加するステップをさらに含む。
【０１２４】
上記に提供された方法は、完全メッセージまたはＩＢ最大伝送ユニット（ＭＴＵ）パケットを受信すると、外部メモリに格納されたディスク読取データを読出すステップをさらに含む。
【０１２５】
上記に提供された方法は、ＩＢドメインにおいて、異なるパケットシーケンス番号（ＰＳＮ）スペースを保留するように、各仮想ＨＢＡを構成するステップをさらに含む。
【０１２６】
上記に提供された方法は、ｖＨＢＡに関連付けられた異なるディスク読取操作のために、異なるコンテキストテーブルを保存するステップをさらに含む。
【０１２７】
上記に提供された方法は、Ｉ／Ｏ装置が仮想ＨＢＡに関連付けられた別のコンテキストテーブルを開くと、コンテキストテーブルを閉じるステップをさらに含む。
【０１２８】
上記に提供された方法は、空きバッファプールから割当てられた１つ以上のパケットバッファに指向するバッファポインタの新たな一時リンクリストを保存するように、別のコンテキストテーブルを構成するステップをさらに含む。
【０１２９】
上記に提供された方法は、仮想ＨＢＡが別のディスク読取操作からデータ、または書込専用命令、ＲＤＭＡ書込最終命令、送信専用命令およびＲＤＭＡ読取リクエスト命令のうち１つの命令を受信すると、別のコンテキストテーブルを開くステップをさらに含む。
【０１３０】
本発明の一実施形態は、命令を格納する非一時的な機械読取可能記憶媒体を提供する。これらの命令は、実行されると、以下のステップをシステムに実行させ、当該以下のステップは、ネットワーク環境内のつ以上の仮想ホストバスアダプタ（ｖＨＢＡ）に関連付けられた複数のパケットバッファを含む空きバッファプールを提供するステップを含み、ｖＨＢＡの各々は、１つ以上のパケットバッファに指向するバッファポインタの主要リンクリストを空きバッファプールに保存し、入力／出力（Ｉ／Ｏ）装置に関連付けられたオンチップメモリにおいて、コンテキストテーブルを定義するステップを含み、コンテキストテーブルは、ディスク読取操作のために、空きバッファプールから割当てられた１つ以上のパケットバッファに指向するバッファポインタの一時リンクリストを保存し、ディスク読取操作を実行する物理ホストバスアダプタ（ＨＢＡ）からディスク読取データを受信すると、コンテキストテーブルを開き、バッファポインタの一時リンクリストを更新し、およびコンテキストテーブルが閉じられると、バッファポインタの一時リンクリストをバッファポインタの主要リンクリストに合併するように、Ｉ／Ｏ装置を動作させるステップを含む。
【０１３１】
本発明の一実施形態は、コンピューティング環境において、入力／出力（Ｉ／Ｏ）仮想化をサポートするためのシステムを提供する。このシステムは、メモリ内の空きバッファプールを含み、空きバッファプールは、２次元リンクリストおよび１次元リンクリストを備え、２次元リンクリストの各エントリは、連続したメモリ位置で複数のパケットバッファを含み、１次元リンクリストの各エントリは、単一のパケットバッファを含み、Ｉ／Ｏ装置は、空きバッファプールを用いて、物理ホストバスアダプタ（ＨＢＡ）から受信したディスク読取データを保存するように動作する。
【０１３２】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、ディスク読取操作を開始するように、インフィニバンド（ＩＢ）ファブリック上のサーバを動作させる。
【０１３３】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、１つ以上のＩＢヘッダとシーケン

ス番号とを物理ＨＢＡから受信した各パケットに追加する。
【０１３４】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、完全メッセージまたはＩＢ最大伝送ユニット（ＭＴＵ）パケットを受信すると、外部メモリに格納されたディスク読取データを読出すように動作する。
【０１３５】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、１つ以上の仮想ホストバスアダプタ（ｖＨＢＡ）をサポートしており、各ｖＨＢＡは、ＩＢドメインにおいて、異なるパケットシーケンス番号（ＰＳＮ）スペースを保留する。
【０１３６】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、パケットがリモートダイレクトメモリアクセス（ＲＤＭＡ）書込トランザクションまたはＲＤＭＡ読取リクエストトランザクションのいずれかを実行する場合、このパケットを外部メモリにキュー入れするように動作する。
【０１３７】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、キューに入れられたパケットがＲＤＭＡ書込トランザクション用のものである場合、２次元リンクリストからスーパーブロックを割当てるように動作する。
【０１３８】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、Ｉ／Ｏ装置は、キューに入れられたパケットがＲＤＭＡ読取リクエストトランザクション用のものであり、且つ、スーパーブロックに１つ以上のパケットバッファが残された場合、１次元リンクリストからパケットバッファを割当てるように動作する。
【０１３９】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、キューに入れられたパケットがＲＤＭＡ読取リクエストトランザクション用のものであり、且つ、スーパーブロックに１つ以上のパケットバッファが残された場合、２次元リンクリストからスーパーブロックを割当てるように動作する。
【０１４０】
上記に提供されたシステムにおいて、Ｉ／Ｏ装置は、ｖＨＢＡのために、単純リンクリストにおいてパケットバッファを保留するように動作する。
【０１４１】
本発明の一実施形態は、ネットワーク環境において、効率的なパケット処理をサポートするための方法を提供する。この方法は、メモリに空きバッファプールを提供するステップを含み、空きバッファプールは、２次元リンクリストおよび１次元リンクリストを含み、２次元リンクリスト各エントリを連続したメモリ位置で複数のパケットバッファを含み、および１次元リンクリストの各エントリを単一のパケットバッファを含むようにするステップを含み、Ｉ／Ｏ装置を介して、空きバッファプールを用いて、物理ホストバスアダプタ（ＨＢＡ）から受信したデータ読取ディスクを保存するステップを含む。
【０１４２】
上記に提供された方法は、ディスク読取操作を開始するように、インフィニバンド（ＩＢ）ファブリック上のサーバを動作させるステップをさらに含む。
【０１４３】
上記に提供された方法は、１つ以上のＩＢヘッダとシーケンス番号とを物理ＨＢＡから受信した各パケットに追加するステップをさらに含む。
【０１４４】
上記に提供された方法は、完全メッセージまたはＩＢ最大伝送ユニット（ＭＴＵ）パケットを受信すると、外部メモリに格納されたディスク読取データを読出すステップをさらに含む。
【０１４５】

上記に提供された方法は、１つ以上の仮想ホストバスアダプタ（ｖＨＢＡ）をサポートするステップをさらに含み、各ｖＨＢＡは、ＩＢドメインにおいて、異なるパケットシーケンス番号（ＰＳＮ）スペースを保留する。

【０１４６】

上記に提供された方法は、パケットがリモートダイレクトメモリアクセス（ＲＤＭＡ）書込トランザクションまたはＲＤＭＡ読取リクエストトランザクションのいずれかを実行する場合、このパケットを外部メモリにキュー入れするステップをさらに含む。

【０１４７】

上記に提供された方法は、キューに入れられたパケットがＲＤＭＡ書込トランザクション用のものである場合、２次元リンクリストからスーパーブロックを割当てるステップをさらに含む。

【０１４８】

上記に提供された方法は、キューに入れられたパケットがＲＤＭＡ読取リクエストトランザクション用のものであり、且つ、スーパーブロックに１つ以上のパケットバッファが残された場合、１次元リンクリストからパケットバッファを割当てるステップをさらに含む。

【０１４９】

上記に提供された方法は、キューに入れられたパケットがＲＤＭＡ読取リクエストトランザクション用のものであり、且つ、スーパーブロックに１つ以上のパケットバッファが残された場合、２次元リンクリストからスーパーブロックを割当てるステップと、

ｖＨＢＡのために、単純リンクリストにおいてパケットバッファを保留するステップとをさらに含む。

【０１５０】

本発明の一実施形態は、命令を格納する非一時的な機械読取可能記憶媒体を提供する。命令は、実行されると、以下のステップをシステムに実行させ、当該以下のステップは、メモリに空きバッファプールを提供するステップを含み、空きバッファプールは、２次元リンクリストおよび１次元リンクリストを含み、２次元リンクリスト各エントリを連続したメモリ位置で複数のパケットバッファを含み、および１次元リンクリストの各エントリを単一のパケットバッファを含むようにするステップを含み、Ｉ／Ｏ装置を介して、空きバッファプールを用いて、物理ホストバスアダプタ（ＨＢＡ）から受信したデータ読取ディスクを保存するステップを含む。

【０１５１】

本発明の多くの特徴は、ハードウェア、ソフトウェア、ファームウェア、またはそれらの組合せの内部で、またはそれらを用いて、またはそれらの援助をもって実現されることができる。したがって、本発明の特徴は、（たとえば、１つ以上のプロセッサを含む）処理システムを用いて実施されることができる。

【０１５２】

本発明の特徴は、コンピュータプログラム製品の内部で、またはそれを用いて、またはその援助をもって実現されることができる。コンピュータプログラム製品は、本明細書に記載の特徴のいずれかを実現するように、処理システムをプログラムさせるために使用することができる命令をその上に／中に格納する記憶媒体またはコンピュータ読取可能媒体である。記憶媒体は、フロッピーディスク（登録商標）、光ディスク、ＤＶＤ、ＣＤ－ＲＯＭ、マイクロドライブおよび光磁気ディスクを含む任意の種類のディスク、ＲＯＭ、ＲＡＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、ＤＲＡＭ、ＶＲＡＭ、フラッシュメモリデバイス、磁気または光カード、（分子メモリＩＣを含む）ナノシステム、または指令および／またはデータの格納に適した任意の種類の媒体またはデバイスを含むことができるが、これらに限定されない。

【０１５３】

機械読取可能媒体のいずれかに格納された本発明の特徴は、処理システムのハードウェアを制御するため、および本発明の結果を利用する他の機構と対話できる処理システムを

可能にするために、ソフトウェアおよび／またはファームウェアに組込むことができる。このようなソフトウェアまたはファームウェアは、アプリケーションコード、デバイスドライバ、オペレーティングシステム、実行環境／コンテナ含むことができるが、これらに限定されない。

【０１５４】

また、本発明の特徴は、たとえば特定用途向け集積回路（ＡＳＩＣ）のようなハードウェア部品を用いて、ハードウェアにおいて実現されてもよい。本明細書に記載の機能を実行するようにハードウェア状態マシンを実装することは、当業者には明らかであろう。

【０１５５】

さらに、本発明は、１つ以上の従来の汎用または専用デジタルコンピュータ、コンピューティング装置、コンピューティング機械、または１つ以上のプロセッサ、メモリおよび／または本開示の教示に従ってプログラムされたコンピュータ読取可能記憶媒体を含むマイクロプロセッサを用いて、簡便に実施することができる。ソフトウェア分野の当業者には明らかなように、本開示の教示に基づいて、熟練したプログラマは、適切なソフトウェアコーディングを容易に用意することができる。

【０１５６】

上記で本発明のさまざまな実施形態を説明したが、これらの実施形態は、限定の目的ではなく、例示として提示されていることが理解すべきである。本発明の精神および範囲から逸脱することなく、本発明に形式上および詳細上のさまざまな変更を行うことができることは、当業者には明らかであろう。

【０１５７】

本発明は、特定の機能およびそれらの関係を示す機能的構造ブロックを用いて説明しました。説明の便宜のために、これらの機能的構造ブロックは、多くの場合、本明細書において任意に定義されている。特定の機能およびそれらの関係が適切に実行される限り、代替的な構造ブロックを定義することができる。任意の代替的な構造ブロックは、本発明の範囲および精神に含まれる。

【０１５８】

本発明の上記説明は、例示および説明のために提供されている。本発明を網羅的であることにまたは開示された形態に厳密に限定することを意図するものではない。本発明の幅および範囲は、上述した例示的な実施形態のいずれかに限定されない。多くの修正および変更は、当業者にとって明らかであろう。修正および変更は、開示された特徴の任意の適切な組合せを含む。実施形態は、本発明の原理およびその実際の応用を最善に説明するために選択され説明された。よって、当業者は、さまざまな実施形態により本発明を理解し、考えられる特定の用途に適したさまざまな修正を行うことができる。なお、本発明の範囲が添付の特許請求の範囲およびその等価物によって定義されることが意図される。

【図１】

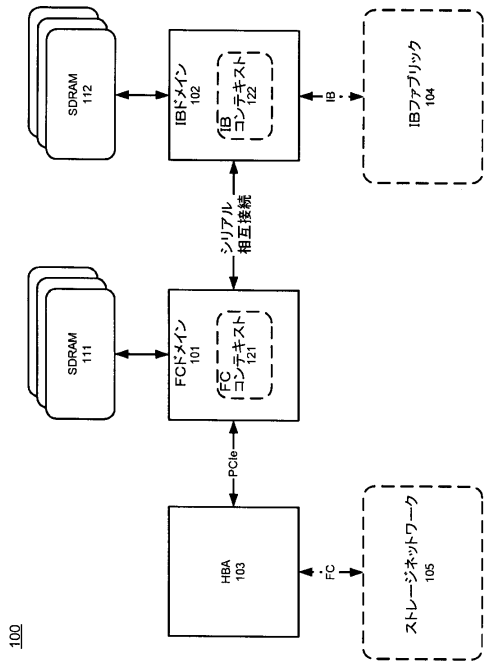

FIGURE 1

【図２】

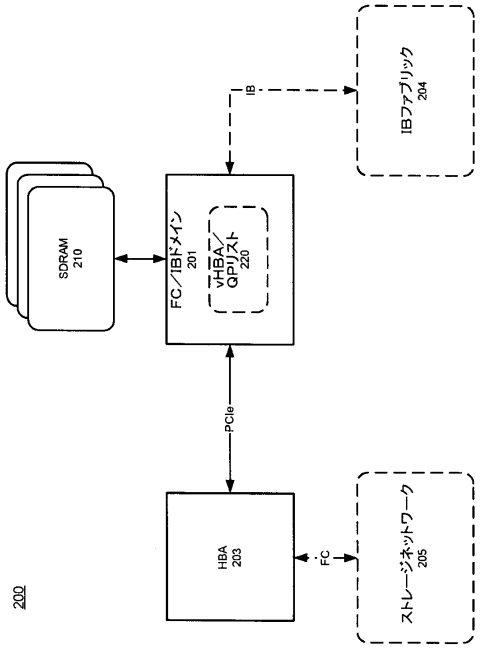

FIGURE 2

【図３】

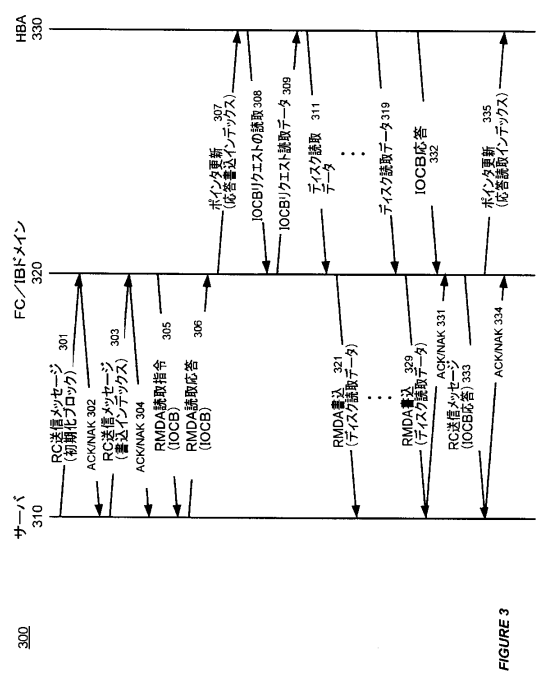

FIGURE 3

【図４】

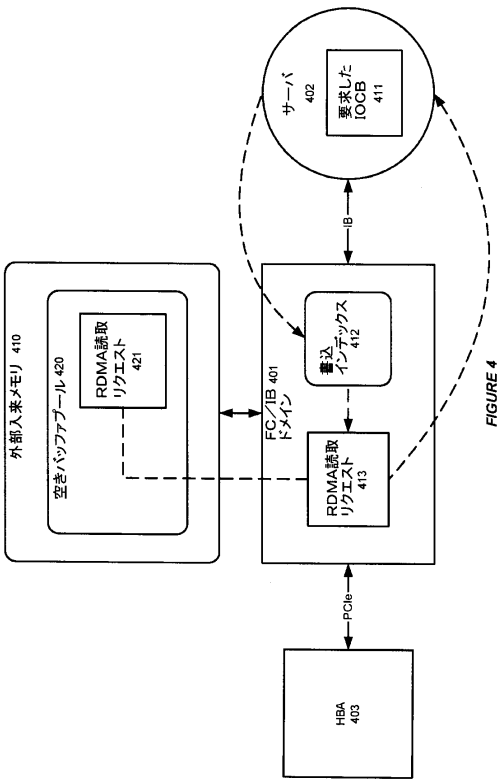

FIGURE 4

【図５】

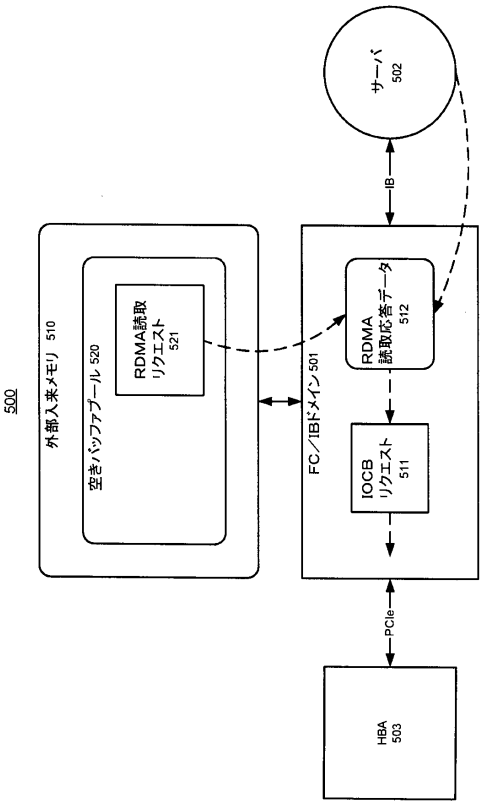

FIGURE 5

【図６】

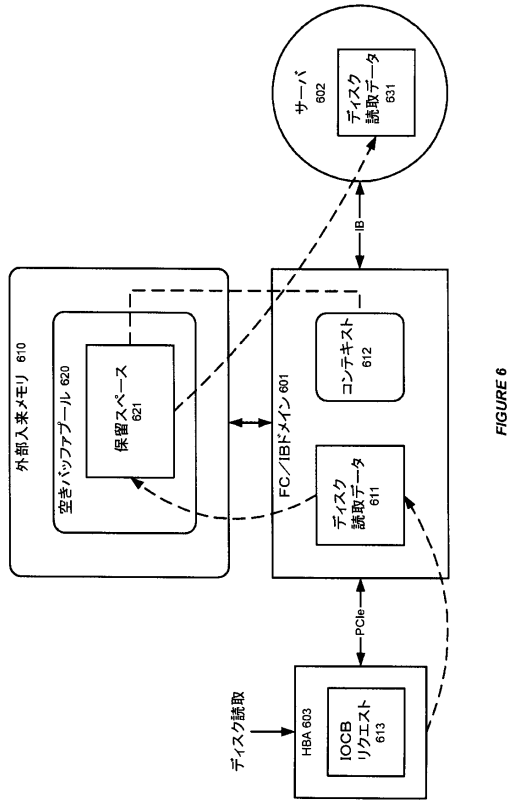

FIGURE 6

【図７】

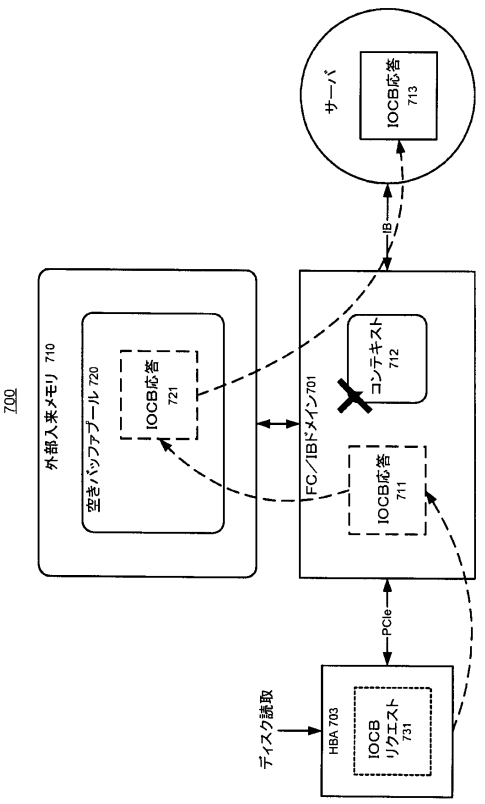

FIGURE 7

【図８】

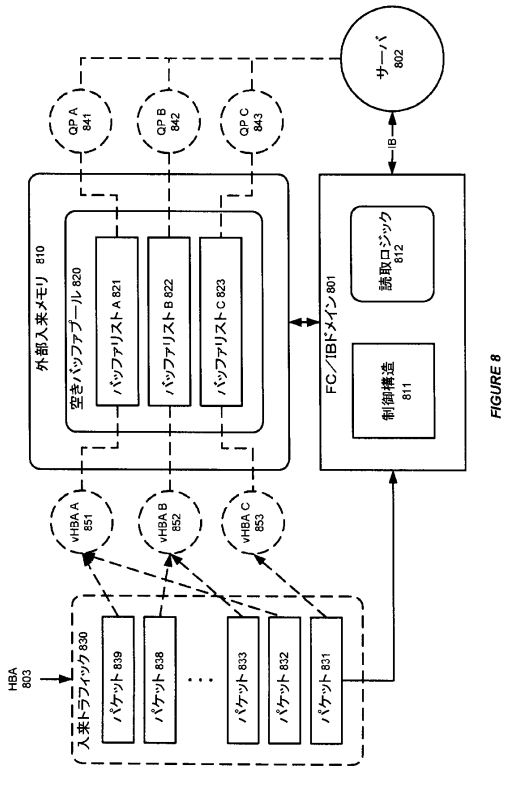

FIGURE 8

【図９】

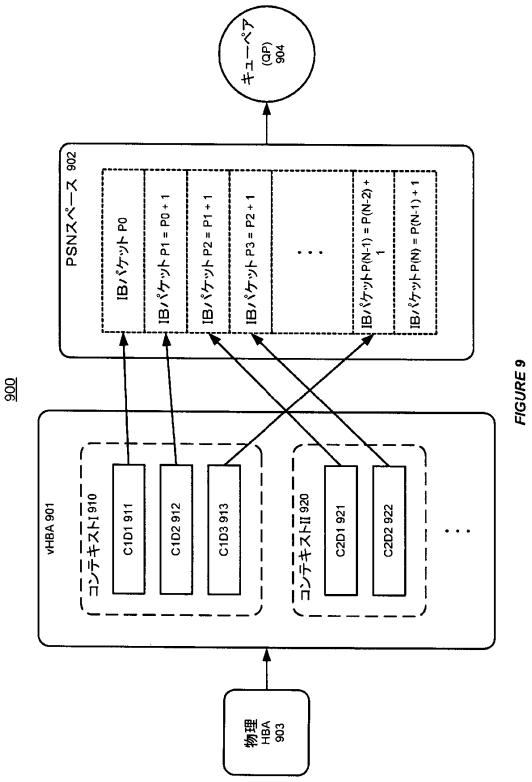

FIGURE 9

【図１０】

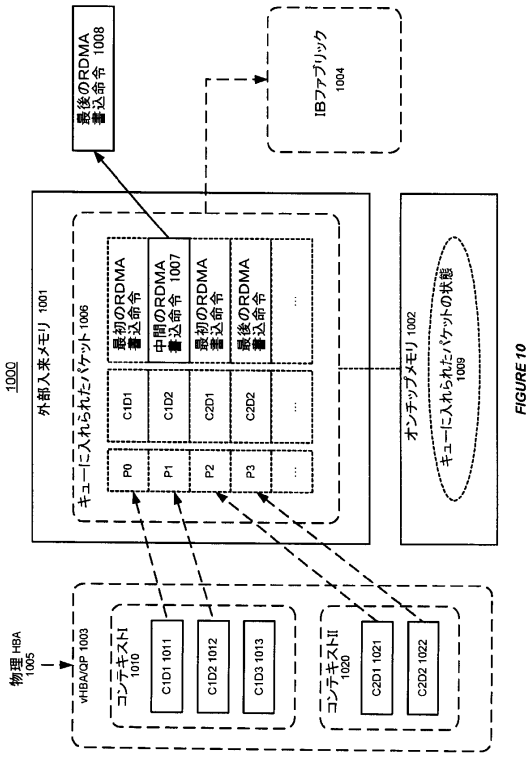

FIGURE 10

【図１１】

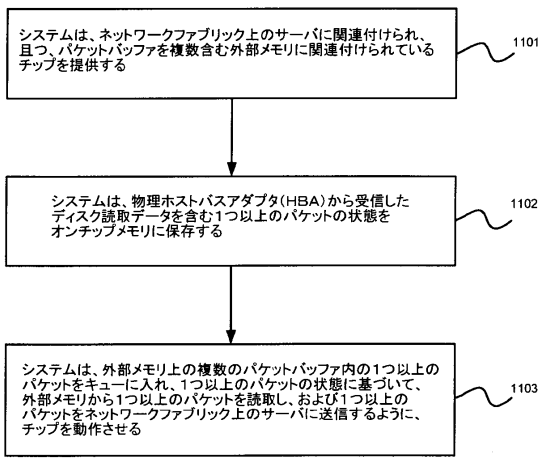

FIGURE 11

【図１２】

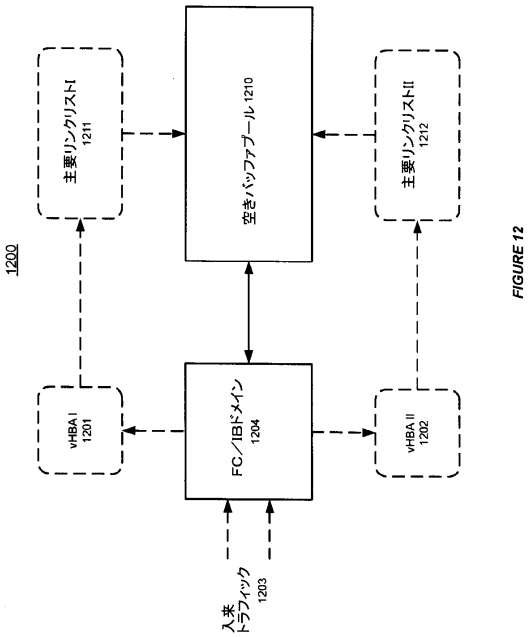

FIGURE 12

【図１３】

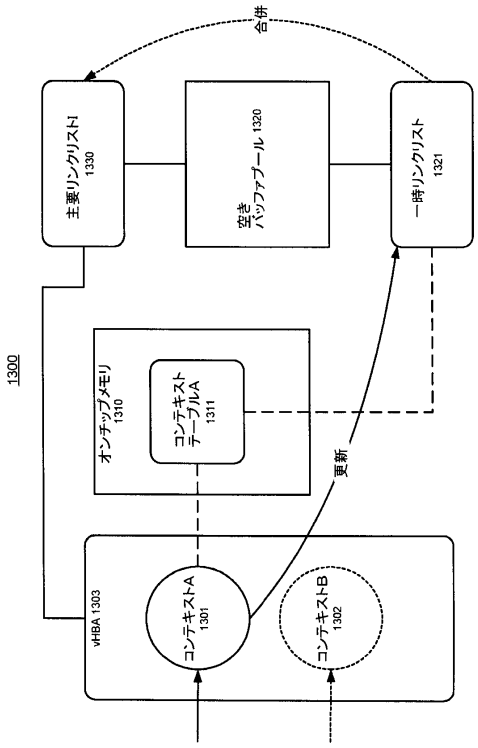

FIGURE 13

【図１４】

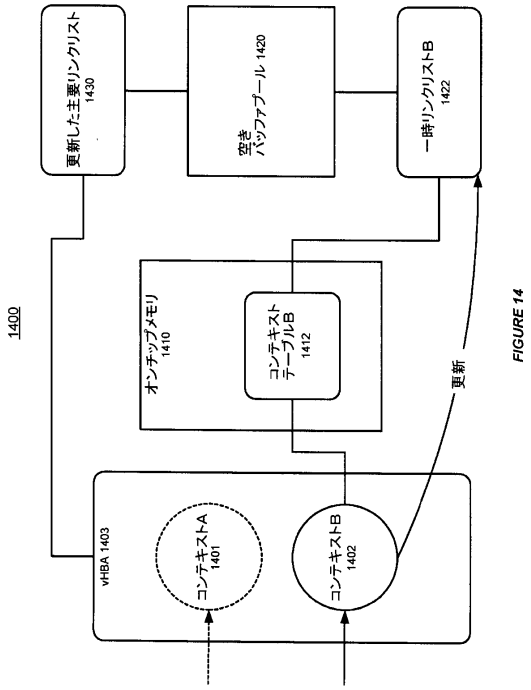

FIGURE 14

【図１５】

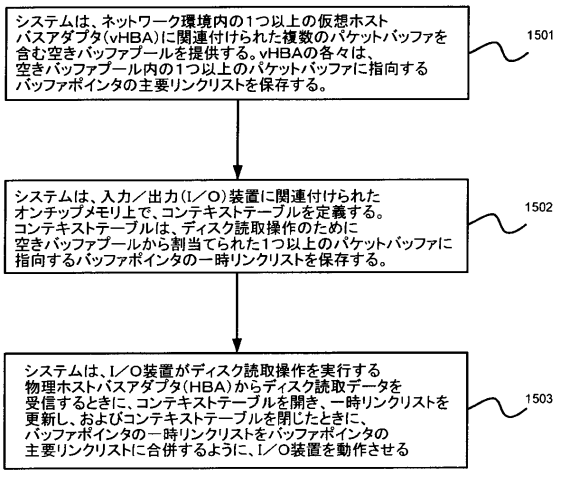

FIGURE 15

【図１６】

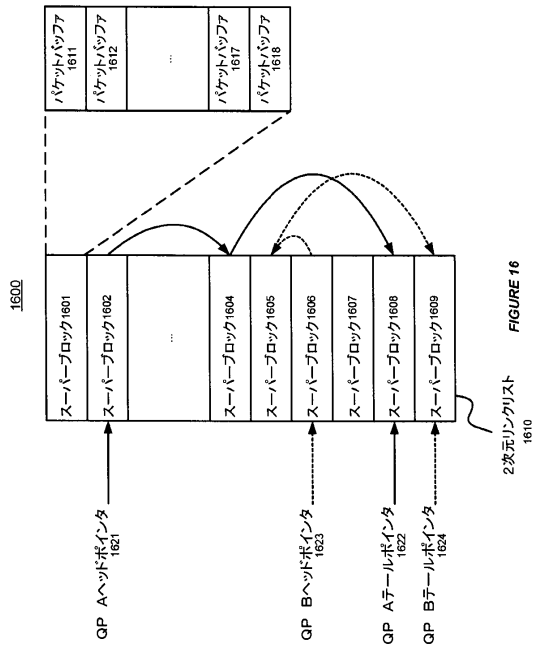

FIGURE 16

【図１７】

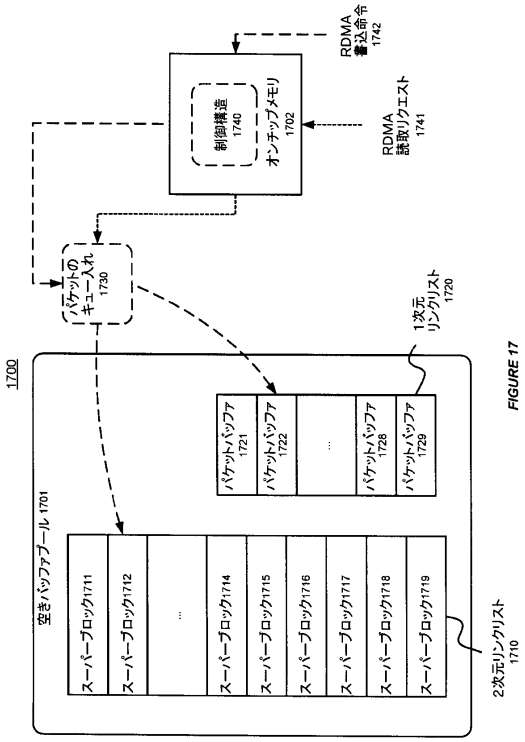

FIGURE 17

【図１８】

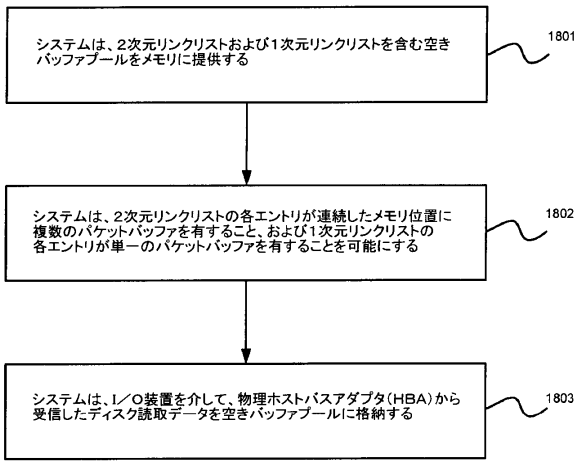

FIGURE 18

フロントページの続き

(51)Int.Cl. FI
H04L 12/861

(31)優先権主張番号 14/097,009
(32)優先日 平成25年12月4日(2013.12.4)
(33)優先権主張国 米国(US)

(56)参考文献 米国特許出願公開第２０１３／０１３８８３６（ＵＳ，Ａ１）
米国特許第０８４５８３０６（ＵＳ，Ｂ１）
特開２００５－２１６２８３（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １３／００－１３／１４
Ｇ０６Ｆ １３／２０－１３／４２
Ｇ０６Ｆ １５／１６－１５／１７７
Ｈ０４Ｌ １２／８６１

(54)【発明の名称】インフィニバンド（ＩＢ）上で仮想ホストバスアダプタ（ｖＨＢＡ）を管理およびサポートするためのシステムおよび方法、ならびに単一の外部メモリインターフェイスを用いてバッファの効率的な使用をサポートするためのシステムおよび方法