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

特許第6122483号
(P6122483)

(45)発行日　平成29年4月26日(2017.4.26)　(24)登録日　平成29年4月7日(2017.4.7)

(51)Int.Cl.		F I		
G06F 9/46	(2006.01)	G06F	9/46	350
G06F 9/50	(2006.01)	G06F	9/46	462Z
G06F 13/10	(2006.01)	G06F	13/10	330C

請求項の数 12　(全 21 頁)

(21)出願番号	特願2015-503440 (P2015-503440)
(86)(22)出願日	平成25年3月25日 (2013.3.25)
(65)公表番号	特表2015-518602 (P2015-518602A)
(43)公表日	平成27年7月2日 (2015.7.2)
(86)国際出願番号	PCT/US2013/033754
(87)国際公開番号	WO2013/148600
(87)国際公開日	平成25年10月3日 (2013.10.3)
審査請求日	平成28年1月19日 (2016.1.19)
(31)優先権主張番号	61/615,731
(32)優先日	平成24年3月26日 (2012.3.26)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	61/693,703
(32)優先日	平成24年8月27日 (2012.8.27)
(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　ヨンセン，ビョルン・ダグ
ノルウェー、エヌ－０６８７　オスロ、ビルベルクグレンダ、９

最終頁に続く

(54)【発明の名称】拡張ホストチャネルアダプタ（ＨＣＡ）モデルに基づいてバーチャルマシンのライブマイグレーションをサポートするためのシステムおよび方法

(57)【特許請求の範囲】

【請求項１】

ネットワークにおいてバーチャルマシンライブマイグレーションをサポートするためのシステムであって、

１つ以上のマイクロプロセッサと、

前記１つ以上のマイクロプロセッサに関連付けられるファブリックアダプタとを含み、前記ファブリックアダプタは複数のバーチャルホストチャネルアダプタ（ＨＣＡ）に関連付けられており、前記バーチャルホストチャネルアダプタ（ｖＨＣＡ）の各々は、別個のキューペア（ＱＰ）スペースに関連付けられており、前記システムはさらに、

第１のホストから第２のホストにライブマイグレーションを行なうよう動作する少なくとも１つのバーチャルマシンを含み、前記少なくとも１つのバーチャルマシンには、前記キューペア（ＱＰ）スペースにおけるキューペア（ＱＰ）に関連付けられる前記バーチャルホストチャネルアダプタ（ｖＨＣＡ）が付与されており、

前記キューペア（ＱＰ）は、前記ライブマイグレーションについてピアＱＰに信号送信するよう動作し、マイグレーションの後に前記ピアＱＰにアドレス情報を提供する、システム。

【請求項２】

前記少なくとも１つのバーチャルマシンを管理する少なくとも１つのバーチャルマシンモニタをさらに含み、前記バーチャルマシンの各々は、プライベート仮想アドレススペースに関連付けられる、請求項１に記載のシステム。

【請求項３】
前記ファブリックアダプタは、仮想スイッチモデルおよび拡張ホストチャネルアダプタ（ＨＣＡ）モデルの少なくとも１つを実現する、請求項１または２に記載のシステム。
【請求項４】
中断状態は、マイグレートされたＱＰとマイグレートされたＱＰに接続されるリモートピアＱＰとの両方に適用可能である、請求項１～３のいずれか１項に記載のシステム。
【請求項５】
特別なヘッダ情報および／または非送信請求メッセージが、ピア同士の間の通信および通信状態の更新をサポートするために使用可能である、請求項１～４のいずれか１項に記載のシステム。
【請求項６】
無効なｖＨＣＡを目標とする入力パケットがホストにて受け取られると、前記ホストは、ローカル非同期イベントを生成し、前記入力パケットのソースに非送信請求パケットを送信可能であり、前記入力パケットのソースに関連付けられるキューペア（ＱＰ）は、自動的に中断状態への遷移を実行可能である、請求項１～５のいずれか１項に記載のシステム。
【請求項７】
マイグレートされたｖＨＣＡから送られるパケットは、マイグレートされた前記ｖＨＣＡが新しい位置で再び作動状態になる前に更新可能であるマイグレーションカウントフィールドを包含可能である、請求項１～６のいずれか１項に記載のシステム。
【請求項８】
マイグレートされたｖＨＣＡは、非中断ペンディング状態のＱＰで開始可能である、請求項１～７のいずれか１項に記載のシステム。
【請求項９】
ＬＩＤ値をＨＣＡポートに割り当て可能なサブネットマネージャをさらに含み、前記ＬＩＤ値は、前記ＨＣＡポートに既に関連付けられる任意の他のＬＩＤ値から独立して、関連するＨＣＡポートへのパケット転送を促進するよう前記サブネットマネージャが使用することを決定する、請求項１～８のいずれか１項に記載のシステム。
【請求項１０】
前記サブネットマネージャは、アドレシングのためのユニークＬＩＤを必要とするＶＭにａｄｍｉｎ　ＬＩＤを割り当て可能である、請求項９に記載のシステム。
【請求項１１】
ネットワークにおいてバーチャルマシンライブマイグレーションをサポートするための方法であって、
ファブリックアダプタを複数のバーチャルホストチャネルアダプタ（ＨＣＡ）に関連付けるステップを含み、前記バーチャルホストチャネルアダプタ（ｖＨＣＡ）の各々は、別個のキューペア（ＱＰ）スペースに関連付けられており、前記方法はさらに、
少なくとも１つのバーチャルマシンを介して、第１のホストから第２のホストにライブマイグレーションを行なうステップを含み、前記少なくとも１つのバーチャルマシンには、前記キューペア（ＱＰ）スペースにおけるキューペア（ＱＰ）に関連付けられる前記バーチャルホストチャネルアダプタ（ｖＨＣＡ）が付与されており、前記方法はさらに、
前記キューペア（ＱＰ）を介して、前記ライブマイグレーションについてピアＱＰに信号送信し、マイグレーションの後に前記ピアＱＰにアドレス情報を提供するステップを含む、方法。
【請求項１２】
コンピュータ読取可能なプログラムであって、実行されると、システムに、
ファブリックアダプタを複数のバーチャルホストチャネルアダプタ（ＨＣＡ）に関連付けるステップを行わせ、前記バーチャルホストチャネルアダプタ（ｖＨＣＡ）の各々は、別個のキューペア（ＱＰ）スペースに関連付けられており、さらに、
少なくとも１つのバーチャルマシンを介して、第１のホストから第２のホストにライブ

マイグレーションを行なうステップを行わせ、前記少なくとも１つのバーチャルマシンには、前記キューペア（ＱＰ）スペースにおけるキューペア（ＱＰ）に関連付けられる前記バーチャルホストチャネルアダプタ（ｖＨＣＡ）が付与されており、さらに、

前記キューペア（ＱＰ）を介して、前記ライブマイグレーションについてピアＱＰに信号送信し、マイグレーションの後に前記ピアＱＰにアドレス情報を提供するステップを行わせる、コンピュータ読取可能なプログラム。

【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は、著作権の保護下にある内容を含む。著作権所有者は、特許商標庁の特許ファイルまたはレコードに現れるので、誰でも当該特許文書または特許開示を複製することについて異議はないが、そうでなければ如何なる場合でもすべての著作権を留保する。

【背景技術】

【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し、特にコンピュータシステム仮想化をサポートすることに関する。

【０００３】

背景

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につれて、従来のネットワークおよびストレージに関連付けられる性能および管理上のボトルネックが重要な問題になっている。インフィニバンド（ＩｎｆｉｎｉＢａｎｄ（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礎として、展開が増加している。これは、本発明の実施例が対応することが意図される一般的な領域である。

【発明の概要】

【課題を解決するための手段】

【０００４】

概要

本願明細書では、ネットワークにおいてバーチャルマシンライブマイグレーションをサポートするためのシステムおよび方法が記載される。ファブリックアダプタは、複数のバーチャルホストチャネルアダプタ（ＨＣＡ）に関連付けられ得、上記バーチャルホストチャネルアダプタ（ｖＨＣＡ）の各々は、別個のキューペア（ＱＰ）スペースに関連付けられている。少なくとも１つのバーチャルマシンが第１のホストから第２のホストにライブマイグレーションを行なうよう動作し、上記少なくとも１つのバーチャルマシンには、上記キューペア（ＱＰ）スペースにおけるキューペア（ＱＰ）に関連付けられる上記バーチャルホストチャネルアダプタ（ｖＨＣＡ）が付与されており、上記キューペア（ＱＰ）は、上記ライブマイグレーションについてピアＱＰに信号送信するよう動作し、マイグレーションの後に上記ピアＱＰにアドレス情報を提供する。

【図面の簡単な説明】

【０００５】

【図１】本発明の実施例に従った仮想インターフェイスアーキテクチャ（ＶＩＡ）ネットワークデバイスの図を示す図である。

【図２】本発明の実施例に従った、仮想インターフェイス（ＶＩ）に関連付けられる異なる状態の図を示す図である。

【図３】本発明の実施例に従った、キューペア（ＱＰ）に関連付けられる異なる状態の図を示す図である。

【図４】本発明の実施例に従った、ネットワークおよびルーティングの観点からのバーチャルマシン（ＶＭ）のライブマイグレーションの図を示す図である。

【図５】本発明の実施例に従った、バーチャルマシン（ＶＭ）ライブマイグレーションの前の仮想化環境の図を示す図である。
【図６】本発明の実施例に従った、バーチャルマシン（ＶＭ）ライブマイグレーションの後の仮想化環境の図を示す図である。
【図７】本発明の実施例に従った、仮想化環境におけるバーチャルマシンのライブマイグレーションをサポートするための例示的なフローチャートを示す図である。
【図８】本発明の実施例に従った、インフィニバンドアーキテクチャ（ＩＢＡ）ネットワークデバイスの図を示す図である。
【図９】本発明の実施例に従った、マイグレートされたキューペア（ＱＰ）とリモートピアＱＰとの間の通信をサポートする図を示す図である。
【図１０】本発明の実施例に従った、拡張ｖＨＣＡモデルに基づいてバーチャルマシンのライブマイグレーションをサポートするための例示的なフローチャートを示す図である。
【図１１】本発明の実施例に従った、仮想化環境においてローカル識別子（ＬＩＤ）の割り当てをサポートする図を示す図である。
【発明を実施するための形態】
【０００６】
詳細な説明
本発明は、限定目的ではなく例示目的で、添付の図面の図において示される。当該図面においては、同様の参照符号が同様の要素を示す。なお、この開示における「ある」、「１つ」または「いくつか」の実施例への参照は、必ずしも同じ実施例に対してなされるものではなく、このような参照は、少なくとも１つの実施例を意味する。
【０００７】
以下の本発明の記載は、高性能ネットワークについての例として、インフィニバンド（ＩＢ）ネットワークを使用する。他のタイプの高性能ネットワークが限定なしで使用され得るということは、当業者には明らかであろう。さらに、以下の本発明の記載は、仮想化モデルについての例として、Ｘｅｎ仮想化モデルを使用する。他のタイプの仮想化モデルが限定なしで使用され得るということは、当業者には明らかであろう。
【０００８】
本願明細書において、ネットワークにおいてバーチャルマシン（virtual machine（ＶＭ））ライブマイグレーション（live migration）をサポートすることができるシステムおよび方法が記載される。
【０００９】
本発明の実施例に従うと、仮想化は、クラウドコンピューティングでの効率的なリソース活用および弾性的なリソースアロケーションに有益になり得る。ライブマイグレーションは、物理サーバ同士の間でバーチャルマシン（ＶＭ）をアプリケーショントランスペアレントな態様で移動させることによって、リソースの使用を最適化することを可能にする。したがって仮想化は、コンソリデーション、リソースのオンデマンドプロビジョニング、およびライブマイグレーションによる弾性を可能にし得る。
【００１０】
インフィニバンド（ＩＢ）アーキテクチャ
ＩＢアーキテクチャはシリアル・ポイント・ツー・ポイント全二重技術である。ＩＢネットワークは、サブネットとも称され得る。サブネットは、スイッチとポイント・ツー・ポイントリンクとを使用して相互接続されるホストのセットからなる。ＩＢサブネットは、サブネットにおけるすべてのスイッチ、ルータおよびホストチャネルアダプタ（ＨＣＡ）のコンフィギュレーションを含み、ネットワークを初期化および起動することを担う、少なくとも１つのサブネットマネージャ（subnet manager（ＳＭ））を含み得る。
【００１１】
ＩＢは、リモートダイレクトメモリアクセス（remote direct memory access（ＲＤＭＡ））および従来の送信／受信セマンティックスの両方を提供するために、伝送サービスの豊富なセットをサポートする。使用される伝送サービスから独立して、ＩＢ　ＨＣＡは

キューペア（ＱＰ）を使用して通信する。ＱＰは、通信セットアップの間に作り出され、ＱＰナンバー、ＨＣＡポート、宛先ＬＩＤ、キューサイズおよび伝送サービスといった提供される初期属性のセットを有し得る。ＨＣＡは、多くのＱＰを扱い得、各ＱＰは、送信キュー（send queue（ＳＱ））および受信キュー（receive queue（ＲＱ））のようなキューのペアからなっており、このような１つのペアが、通信に参加する各エンドノードに存在する。送信キューは、リモートノードに転送されるワーク要求を保持する一方、受信キューは、リモートノードから受け取ったデータで何を行なうべきかについての情報を保持する。ＱＰに加えて、各ＨＣＡは、送信キューおよび受信キューのセットに関連付けられる１つ以上の完了キュー（completion queue（ＣＱ））を有する。ＣＱは、送信および受信キューにポスティングしたワーク要求についての完了通知を保持する。通信の複雑性はユーザから隠されているが、ＱＰ状態情報はＨＣＡに維持される。

【００１２】

各物理的なＩＢデバイスには、ＬＩＤおよびグローバルユニーク識別子（globally unique identifier（ＧＵＩＤ））の２つのアドレスが割り当てられる。ＬＩＤは、サブネット内でＩＢパケットをルーティングするよう使用され、ＧＵＩＤは、物理的ＩＢデバイスを一意に示すハードウェアアドレスである。６４ビットＧＵＩＤは、ローカルの（６４ビット）サブネットプレフィックスと組み合わされてグローバル識別子を形成する。１２８ビットグローバル識別子は、ＩＢサブネット同士の間でＩＢパケットをルーティングするよう使用される。

【００１３】

入出力（Ｉ／Ｏ）仮想化

Ｉ／Ｏ仮想化（I/O Virtualization（ＩＯＶ））は、存在する物理リソースにＶＭがアクセスすることを可能にすることによって、Ｉ／Ｏの可用性を提供し得る。ストレージトラフィックおよびサーバ間通信の組合せによって、単一のサーバのＩ／Ｏリソースを圧倒し得るような増加した負荷が課され、これにより、データを待っている際に、バックログおよびアイドル状態になるプロセッサが引き起こされる。Ｉ／Ｏ要求の数の増加により、ＩＯＶは可用性を提供し得るとともに、現代のＣＰＵ仮想化において見られる性能のレベルに合致するよう、（仮想化された）Ｉ／Ｏリソースの性能、スケーラビリティおよびフレキシビリティを向上し得る。

【００１４】

エミュレーション、準仮想化、直接割当（direct assignment（ＤＡ））およびシングルルートＩ／Ｏ仮想化（single root-I/O virtualization（ＳＲ－ＩＯＶ））といった、異なるタイプのＩＯＶ技術が存在し得る。これらのＩＯＶ技術のうち、ＳＲ－ＩＯＶは、ほぼネイティブの性能を維持しつつ、複数のＶＭから単一の物理デバイスへの直接アクセスを可能にする手段により、ＰＣＩ　Ｅｘｐｒｅｓｓ（ＰＣＩｅ）規格を拡張し得る。したがってＳＲ－ＩＯＶは、良好な性能およびスケーラビリティを提供し得る。

【００１５】

ＳＲ－ＩＯＶは、各ゲストに１つの仮想デバイスを割り当てることによって複数のゲスト同士の間で共有され得る複数の仮想デバイスをＰＣＩｅデバイスが公開することを可能にする。各ＳＲ－ＩＯＶデバイスは、少なくとも１つの物理的機能（physical function（ＰＦ））と、１つ以上の関連付けられる仮想機能（virtual function（ＶＦ））とを有する。ＰＦは、バーチャルマシンモニタ（virtual machine monitor（ＶＭＭ））またはハイパーバイザによって制御される通常のＰＣＩｅ機能であり、ＶＦは、軽量のＰＣＩｅ機能である。各ＶＦは、自身のベースアドレス（base address（ＢＡＲ））を有しており、ユニークリクエスタＩＤが割り当てられている。ユニークリクエスタＩＤは、Ｉ／Ｏメモリ管理ユニット（I/O memory management unit（ＩＯＭＭＵ））が、異なるＶＦに対するトラフィックストリーム同士を区別することを可能にする。ＩＯＭＭＵはさらに、ＰＦとＶＦとの間でメモリおよび割込の変換を適用する。

【００１６】

たとえば共有ポートモデルおよび仮想スイッチモデルといった、異なるタイプのＳＲ－

ＩＯＶモデルが存在し得る。共有ポートモデルでは、すべてのＶＦは、単一のポートアドレスおよび単一のＱＰネームスペースを共有し得、単一のＨＣＡポートのみがネットワークに公開される。仮想スイッチモデルでは、各ＶＦは、ユニークポートアドレスおよびユニークＱＰネームスペースを含む仮想ＨＣＡであり、デバイス上のＶＦごとに１つのＨＣＡがネットワークに公開される。したがって、より複雑なハードウェアであっても、仮想スイッチモデルはＩＯＶを簡素化し得る。

【００１７】

ＳＲ－ＩＯＶ対応のデバイスの使用によって、ほぼネイティブの性能およびスケーラビリティの向上が与えられ得る。その一方、ＳＲ－ｌＯＶは、ライブマイグレーションおよびチェックポイント／リスタートメカニズムと完全には互換性がないわけではない。

【００１８】

ライブマイグレーションのためのハードウェアソリューション

本発明の実施例に従って、ハードウェアソリューションは、アプリケーショントランスペアレントな態様で、フレキシブルで高性能かつスケーラブルな仮想化されたＩＯアーキテクチャに基づき、物理サーバ同士間でのバーチャルマシン（ＶＭ）のライブマイグレーションをサポートし得る。さらに、ハードウェアソリューションは、マイグレーションのトランスペアレンシーを増加し得、ＩＢとＳＲ－ＩＯＶとの間の相互運用性を改善し得る。さらに、当該システムは、仮想インターフェイスアーキテクチャ（virtual interface architecture（ＶＩＡ））およびミリネット（Myrinet）のような他の高速ロスレスネットワーキング技術に適用可能であり得る。

【００１９】

ハードウェアソリューションを使用して、ＩＢホストチャネルアダプタ（ＨＣＡ）は、ＱＰについて自身のリソースプールを有する仮想エンドポイントとして、各物理的な機能（ＶＦ）を区別し得る。さらに、ＱＰＮのようなＱＰ属性は、マイグレーションの後に再使用され得る。さらに、ＩＢ　ＨＣＡは、通信を中断および再開するためにＱＰ状態に基づいて、ＶＭライブマイグレーションをサポートし得る。

【００２０】

さらに、システムは、マイグレートされたＱＰとマイグレートされたＱＰのピアとの間で休止するよう用いられ得るハンドシェイクメカニズムを回避し得る。したがって、当該システムは、バーチャルマシン（ＶＭ）のライブマイグレーションが、ＩＢにおける信頼性のある接続の再試行タイムアウトに相当するレベルにまでサービスダウンタイムを低減することを可能にする。

【００２１】

さらにシステムは、厳格なタイミング要件を有するシナリオおよびアプリケーションを含む、ＶＭライブマイグレーションをサポートする一般的なソリューションを示し得る。たとえば高い可用性のクラスタにおいて、欠陥のある物理サーバによってホストされたＶＭは、限られた時間内に別のサーバにマイグレートされ得る。マイグレーションは、欠陥が発見された後に行なわれ得るが、当該欠陥のあるサーバが回復不能なエラーになる前に行なわれ得る。ソフトウェアアプローチは、オーバーヘッドに苦しんでいるので、この低いサービスダウンタイムの要求を満たすのに不十分である。さらに、ソフトウェアアプローチは開発コストの点で高価である。

【００２２】

仮想エンドポイント

本発明の実施例に従うと、仮想インターフェイスアーキテクチャ（ＶＩＡ）ハードウェアは、各ＶＦから作り出されたＶＩを区別するよう、ＰＣＩ仮想スイッチを統合し得る。仮想スイッチの後ろでは、複数の独立した仮想エンドポイントがネットワークマネージャによって発見可能であり得る。

【００２３】

図１は、本発明の実施例に従った仮想インターフェイスアーキテクチャ（ＶＩＡ）ネットワークデバイスの図を示す。図１に示されるように、仮想インターフェイスアーキテク

チャ（ＶＩＡ）ネットワークデバイス１００は、ハードウェアレイヤー１１０、バーチャルマシンモニタ１２０、およびたとえばＶＭ１０１－１０２といった１つ以上のバーチャルマシンを含み得る。ここで、ＶＭ１０１は、カーネルエージェント１０７に基づいて、固定されたメモリ１０５に関連付けられるユーザプロセス１０３をサポートし得る。ＶＭ１０２は、カーネルエージェント１０８に基づいて、固定されたメモリ１０６に関連付けられるユーザプロセス１０４をサポートし得る。

【００２４】

仮想スイッチ１３０はハードウェアレイヤー１１０上に提供され得る。仮想スイッチ１３０は、たとえばＶＦ１１１－１１９といった１つ以上の仮想機能（ＶＦ）に関連付けられ得る。ＶＦ１１１－１１９の各々は、１つ以上の仮想インターフェイス（ＶＩ）を作成および管理し得る。

【００２５】

本発明の実施例に従うと、仮想スイッチモデルを使用して、ＶＦ１１１－１１９の各々が、たとえばＶＭ１０１またはＶＭ１０２といったゲストＶＭに割り当てられ得る専用のＰＣＩエンドポイントとして分離され得る。さらに、ＶＦ１１１－１１９の各々は、その専用のＶＩネームスペース１２１－１２９を有し得、完全なＶＩエンドポイントとして考えられ得る。ＶＩネームスペースの分離は、仮想化された環境において保護およびフレキシビリティを提供し得る。各ＶＦは、ＩＯ動作についてＶＭの保護および分離プロパティを拡張するパーティション内のＶＩリソースにアクセスするのが制限され得る。さらに、ＶＩ属性はマイグレーションの後に再使用され得る。

【００２６】

たとえば、ハードウェアレイヤー１１０は、ＩＢアーキテクチャのような高速ネットワークアーキテクチャに基づき得る。ＶＦ１１１－１１９の各々は、ユニークなポートアドレスおよびユニークなＱＰネームスペースを含むバーチャルホストチャネルアダプタ（virtual host channel adaptor（ｖＨＣＡ））であり得る。さらに、ＩＢデバイス上のＶＦごとに、１つのｖＨＣＡがネットワークに公開され得る。さらに、各ｖＨＣＡは、自身のＱＰネームスペースに関連付けられ得、ＩＢアーキテクチャにおけるＱＰ１３１および１３２のような１つ以上のキューペア（ＱＰ）を作成および管理し得る。図１に示されるように、キューペア（ＱＰ）１３１－１３２の各々は、センダーキューおよびレシーバーキューを含み得る。

【００２７】

さらに、ＶＩＡデバイス１００は、ＩＯメモリ管理ユニット（ＩＯＭＭＵ）の機能性を活用し得る。ＶＩ属性は、ＩＯＭＭＵによって変換（translate）されるＩＯ仮想メモリを介してアクセスされるメモリ領域に格納され得る。新しいＶＦが、ＶＭに再付与されると、ハイパーバイザは、ＩＯＭＭＵを変更して、明示的に新しいＶＩを再作成することなしに物理メモリの新しいセットにＩＯ仮想メモリを再マッピングし得る。共有されたＶＩネームスペースモデルと比較して、仮想スイッチモデルは、リソースマイグレーションを緩和するとともに、ＶＩ属性を再マッピングするためのマッピングテーブルの使用を回避して、動作中のアプリケーションについてトランスペアレンシーを維持する。

【００２８】

キューペア（ＱＰ）中断状態

図２は、本発明の実施例に従った、仮想インターフェイス（ＶＩ）に関連付けられる異なる状態の図を示す。図２に示されるように、ＶＩのライフサイクルは、アイドル状態２０１、ペンディング接続状態２０２、接続済状態２０３、エラー状態２０４および中断状態２０５といった複数の状態を含み得る。

【００２９】

中断状態２０５は、進行中の通信が、たとえばライブマイグレーションの間、ＶＩを休止状態にするために一時的に停止されることを示す。接続済状態２０３と中断状態２０５との間の遷移は、ソフトウェアが開始した要求またはＶＩイベント要求によって行なわれ得る。たとえば、マイグレーションの間に、ソフトウェアインターフェイスは、接続済状

態２０３から中断状態２０５まで、マイグレートするＶＩを修正し得る。
【００３０】
ＶＩが中断状態２０５に入ると、ＶＩは、如何なる以前に開始されたメッセージ送信も完了し得、それらを処理することなく、ハードウェアセンダーキュー上の如何なる残存するメッセージもフラッシュする。ハードウェアは、残存するメッセージが、フラッシュする前に、ユーザプロセスについてのセンダーキューによってキャッシュされることを保証し得る。中断状態２０５においては、如何なる新しいメッセージも、センダーキューおよびレシーバーキューの両方においてキューに入れられ得、処理されることなくポストされたままである。
【００３１】
さらに、入力メッセージが、中断されたＶＩを目標としている場合、中断状態２０５にピアＶＩを移行させるよう、中断されたＶＩによってSuspendRequestイベントが生成され得る。このような場合、ResumeRequestイベントを受け取った後、以前の開始されたメッセージが再送信され得、不完全な送信時にはタイムアウトエラーが生成され得ない。
【００３２】
新しいＶＦがマイグレートされたＶＭに付与された後およびマイグレートされたＶＭが新しい位置で再び開始される前に、ストップ・アンド・コピーステージ（stop-and-copy stage）が、ＩＯＭＭＵにおける新しいセットの物理メモリにＩＯ仮想メモリを再登録し得る。その後、ソフトウェアインターフェイスは、マイグレートされたＶＭのＶＩ状態を接続済状態２０３に修正し得る。その後、システムは、ピアＶＩへ転送されるResumeRequestイベントをトリガし得る。その後、ピアＶＩは、ResumeRequestイベントを生成することなく接続済状態２０３に移行し得る。したがって、中断状態２０５は、ＶＭのライブマイグレーションの間にスムーズな遷移を提供し得、ＶＩは休止状態にあり得る。
【００３３】
本発明の実施例に従うと、ＶＭライブマイグレーションの間にダウンタイムを低減するために、最小遅延のアプローチは、すべての未完了の動作が完了する直前に、ＶＩを中断またはＶＦを分離することである。したがって、延長によりタイムアウトが向上し得、中断状態２０５のハンドリングの部分として、信頼性のあるトランスポートプロトコルによって与えられるハンドリングが再試行され得る。延長された再試行およびタイムアウトロジックは、ピアＶＩまたはマイグレートされるＶＩのいずれかについて致命的なエラーを引き起こすことなくＶＩを中断状態２０５に移行し得る。
【００３４】
したがって、マイグレーションの後にＶｌの可用性に対してハードなリアルタイムの制約を課すことなく、ＶＩはマイグレートされ得る。このアプローチはさらに、ネットワークの遅延によりイベント報告（SuspendRequestおよびResumeRequest）がＶＩ同士の間で生成されない場合をカバーし得る。さらに、複数のＶＩを備えた場合において、各ＶＩは上記の動作を同時に行ない得る。
【００３５】
図３は、本発明の実施例に従った、キューペア（ＱＰ）に関連付けられる異なる状態の図を示す。図３に示されるように、ＱＰのライフサイクル３００は、リセット状態３０１、初期化（ＩＮＴ）状態３０２、レディ・ツー・レシーブ（ＲＴＲ）状態３０３、レディ・ツー・センド（ＲＴＳ）状態３０４、ＳＱエラー（ＳＱＥ）状態３０５、エラー状態３０６、ＳＱドレイン（ＳＱＤ）状態３０７、および中断（ＳＵＳ）状態３０８といった複数の状態を含む。
【００３６】
中断状態であるＳＵＳ３０８は、マイグレーションの間にＱＰを休止するよう進行中の通信を一時的に停止させ得る。ＲＴＳ３０４とＳＵＳ３０８との間の遷移は、ソフトウェアが開始した要求またはイベント要求によってトリガされ得る。マイグレーションの間に、ソフトウェアインターフェイスは、マイグレートするＱＰをＲＴＳ３０４状態からＳＵＳ３０８状態まで移行し得る。ＱＰは、中断状態に入ると、如何なる以前に開始されたメ

ッセージ送信も完了し得、ＳＱにおける如何なる残存するメッセージを、それらを処理することなく、フラッシュし得る。ハードウェアはさらに、残存するメッセージが、フラッシュする前にユーザプロセスバッファによってキャッシュされることを保証し得る。中断状態において、如何なる新しいメッセージも、ＳＱおよびＲＱの両方にキューに入れられ得るが、処理されることなくポストされたままである。

【００３７】

ＶＭライブマイグレーションのネットワーキングおよびルーティングの観点

図４は、本発明の実施例に従った、ネットワークおよびルーティングの観点からのバーチャルマシン（ＶＭ）のライブマイグレーションの図を示す。図４に示されるように、仮想化環境４００は、たとえばｖＨＣＡ　Ａ－Ｃ４０１－４０３といった複数のｖＨＣＡを含み得る。ｖＨＣＡ　Ａ－Ｃ４０１－４０３の各々は、仮想化についてトランスピアレンシーを提供し得る完全なＩＢエンドポイントとして機能し得る。

【００３８】

さらに、仮想化環境４００は、ネットワークマネージャ４１０を含み得る。ネットワークマネージャは、ＩＢスイッチＡ－Ｂ４３１－４３２の後ろにｖＨＣＡ　Ａ－Ｃ４０１－４０３を発見し得、異なるＬＩＤおよびｖＧＵＩＤに基づき、ｖＨＣＡ　Ａ－Ｃ４０１－４０３の各々を認識し得る。たとえば、ｖＨＣＡ　Ａ４０１には、ＬＩＤ　Ａ４１１およびｖＧＵＩＤ　Ａ４２１が割り当てられ得、ｖＨＣＡ　Ａ４０２にはＬＩＤ　Ｂ４１２およびｖＧＵＩＤ　Ｂ４２２が割り当てられ得、ｖＨＣＡ　Ｃ４０３には、ＬＩＤ　Ｃ４１３およびｖＧＵＩＤ　Ｃ４２３が割り当てられ得る。

【００３９】

本発明の実施例に従うと、ＶＭマイグレーションの間、ネットワークマネージャ４１０は、たとえばｖＨＣＡ　Ｂ４０２のようなＩＢエンドポイントがＩＢスイッチＡ４３１の後ろでダウン（すなわち中断）し、ネットワークの他方に（ｖＨＣＡ　Ｂ４０４として）ＩＢスイッチＢ４３２の後ろに、たとえばＬＩＤ　Ｂ４１２およびｖＧＵＩＤ　Ｂ４２２のような同じアドレスが新しい位置に引き継がれた状態で再び現れることを観察し得る。

【００４０】

マイグレーションの後、ネットワークマネージャ４１０は、そのシフトの後にＬＩＤの新しい再構成を反映するようルーティングテーブル４２０を更新し得る。さらに、全ネットワークのリルーティングタイムを低減するために、ＬＩＤをローカルに修正する新しいルーティングアルゴリズムが実現され得る。さらに、ネットワークマネージャ４１０はさらに、新しいＬＩＤ構成はデッドロックがないことを保証し得る。

【００４１】

さらに、高速ネットワークでは、ローカルアドレスリソースが制限されている場合がある。たとえば、１６ビットＬＩＤを使用し得るＩＢネットワークは、各ｖＨＣＡに自身のＬＩＤが割り当てられるとともに１０００ノードクラスタにおける各ノードが複数のＶＭをホストする場合、すぐにアドレススペースを使い果たし得る。

【００４２】

本発明の実施例に従うと、システムは、仮想化環境におけるスケーラビリティの問題を扱うために、異なるアプローチを使用し得る。たとえば、システムは、３２ビットまでＬＩＤアドレシングスペースを拡張し得るか、または、ｖＨＣＡ同士の間のルーティングを行なうようｖＧＵＩＤをＬＩＤに組み合わせ得る。第１のソリューションは、ネットワークマネージャにおける如何なるアーキテクチャの修正も必要としないが、より古いハードウェアとの後方互換性がない。

【００４３】

バーチャルマシンの観点

図５は、本発明の実施例に従った、バーチャルマシン（ＶＭ）ライブマイグレーションの前の仮想化環境の図を示す。図５に示されるように、ＩＢネットワーク５００は、たとえばホストＡ－Ｂ５０１－５０２といった複数のホストを含み得る。ホストＡ５０１は、ＨＣＡ５１７を使用するＶＭＭ　Ａ５１５を含み、たとえばＤｏｍＵ　５１３といったゲ

ストドメインを管理する特権ドメイン（または管理ドメイン）であるＤｏｍ０　５１１をサポートし得る。さらに、ホストＢ５０２は、ＨＣＡ　５１８を使用するＶＭＭ　Ｂ５１６を含み、ＤｏｍＵ　５１４を管理するＤｏｍ０　５１２をサポートし得る。

【００４４】

さらに、ホストＡ５０１上のＶＭａ５０３には、たとえばＱＰａ５０７といったキューペアに接続されるＶＦ５０５といったＶＦが付与され得る。さらに、ＶＭａ５０３は、ＱＰｂ５０８に関連付けられるＶＭｂ５０４と通信し得る。

【００４５】

本発明の実施例に従うと、たとえばホストＡ５０１上のＤｏｍ０　５１１といった特権ドメインは、たとえばマイグレーションの間に分離されるべきＶＦ５０５のようなデバイスを必要としない。システムは単に、たとえばＱＰａ５０７といったアクティブなＱＰがストップ・アンド・コピーステージで中断されることを必要とし得る。

【００４６】

図５に示されるように、ホストＡ５０１上の特権ドメインであるＤｏｍ０　５１１は、ホストＢ１０２へのＶＭａ５０３のマイグレーションを（図６におけるＶＭａ６０４として）開始し得る。ＶＭａ５０３のマイグレーションの前に、システムは、それを中断状態にする状態遷移動作を行ない得る。さらに、ＱＰｂ５０８から入力パケットを受け取った後、中断されたＱＰａ５０７は、ＱＰｂ５０８に同様に中断状態に入るように命令し得る。

【００４７】

その後、たとえばリモートダイレクトメモリアクセス（ＲＤＭＡ）接続５２０に基づき、Ｉ／Ｏ仮想メモリを含むダーティページ（メモリ）をソースサーバと宛先サーバとの間で同期させることによって、ＶＭマイグレーションが行なわれ得る。

【００４８】

図６は、本発明の実施例に従った、バーチャルマシン（ＶＭ）ライブマイグレーションの後の仮想化環境の図を示す。図６に示されるように、ＩＢネットワーク６００は、たとえばホストＡ－Ｂ　６０１－６０２といった複数のホストを含み得る。（さらに図５と同じ番号を有する図６における要素は同じ機能を果たす）。

【００４９】

ホストＢ６０２へのマイグレーションの後に、ＶＭａ６０３には、たとえばＱＰａ’６０７のようなキューペアに接続されるたとえばＶＦ６１６のような新しいＶＦが付与され得る。ここで、ＱＰａ’６０７は、図５におけるＱＰａ５０７と論理上同じＱＰであり、ＶＭａ５０３（またはＶＭａ６０３）の仮想メモリにおいて同じコンテキスト情報を有する。

【００５０】

さらに、ＶＭａ６０３が停止から解放される前に、特権ドメインであるＤｏｍ０　６１２は、ＱＰａ’６０７をＲＴＳ状態に戻し得る、ＱＰａ’６０７に対する状態遷移を行ない得る。さらに、ＱＰａ’６０７は、ホストＡ６０１上のＱＰｂ６０８との通信を再開し得る。

【００５１】

したがって、マイグレーションについてのサービスダウンタイムは、ストップ・アンド・コピーステージの期間と同等になり得る。さらに、マイグレーションを行なうようＲＤＭＡ動作を使用することは、サービスダウンタイムおよびトータルマイグレーションタイムを低減し得る。

【００５２】

図７は、本発明の実施例に従った、仮想化環境におけるバーチャルマシンのライブマイグレーションをサポートするための例示的なフローチャートを示す。図７に示されるように、ステップ７０１では、システムは、各々が別個のキューペア（ＱＰ）スペースに関連付けられる複数の仮想機能（ＶＦ）に仮想スイッチを関連付け得る。さらに、ステップ７０２では、システムは、少なくとも１つのバーチャルマシンに上記仮想機能（ＶＦ）を付

与し得る。上記仮想機能（ＶＦ）は、仮想インターフェイス（ＶＩ）に関連付けられる。次いで、ステップ７０３では、上記の少なくとも１つのバーチャルマシンは、上記仮想機能（ＶＦ）が付与された状態で第１のホストから第２のホストにライブマイグレーションを行ない得る。

【００５３】

拡張ＨＣＡモデル

図８は、本発明の実施例に従った、インフィニバンドアーキテクチャ（ＩＢＡ）ネットワークデバイスの図を示す。図８に示されるように、インフィニバンドアーキテクチャ（ＩＢＡ）ネットワークデバイス８００は、ハードウェアレイヤー８１０、バーチャルマシンモニタ８２０、およびたとえばＶＭ８０１－８０２といった１つ以上のバーチャルマシンを含み得る。ＶＭ８０１は、カーネル８０７およびユーザライブラリ８３７に基づき、固定されたメモリ８０５に関連付けられるアプリケーション８０３をサポートし得る。ＶＭ８０２は、カーネルエージェント８０８およびユーザライブラリ８３８に基づき、固定されたメモリ８０６に関連付けられるアプリケーション８０４をサポートし得る。

【００５４】

ハードウェアレイヤー８１０は、ＩＢ　ＨＣＡ８３０のような、ＩＢサブネット８４０に接続するファブリックアダプタを含み得る。さらに、ＩＢ　ＨＣＡ８３０は、たとえばｖＨＣＡ　８１１－８１９といった複数のｖＨＣＡインスタンスが関連付けられ得る。

【００５５】

本発明の実施例に従うと、たとえば物理ＨＣＡ８３０といったファブリックアダプタは、仮想スイッチモデル、拡張ＨＣＡモデルまたは両方のモデルの組合せを実現することができる。さらに、拡張ＨＣＡモデルを使用して、スイッチなしの物理ＨＣＡ８３０のみが存在し得る。バーチャルＨＣＡ８１１－８１９は、物理ＨＣＡモデルの拡張を示し、バーチャルＨＣＡインスタンス８１１－８１９の１つ以上は、物理ＨＣＡ８３０ポートを介して、たとえばＩＢサブネット８４０といったファブリックからアクセス可能であり得る。

【００５６】

本発明の実施例に従うと、ハードウェアレイヤー８１０は、ＶＭライブマイグレーションを行なうための割り当てられたＩＢリソースを保存するよう、以下の機能をサポートし得る。

【００５７】

１．　ＱＰコンテキスト（たとえばＱＰナンバー）は、各ｖＨＣＡがプライベートＱＰスペースを有することを可能にすることによって保存され得る。

【００５８】

２．　メモリ領域ナンバーは、各ｖＨＣＡがプライベートメモリ領域ナンバースペースを有することを可能にすることによって保存され得る。

【００５９】

３．　登録したメモリは、各ｖＨＣＡが、ローカルの物理的なホストのローカルメモリへのアクセスのためにプライベート仮想アドレススペースを実現することを可能にすることによって、保存され得る。

【００６０】

プライベートＱＰスペースは、上記のセクションで論じられた仮想スイッチベースのモデルを用いるか、または拡張ＨＣＡモデルを用いることのいずれかによって実現され得る。ＨＣＡ規格を拡張する拡張ＨＣＡモデルは、各々がプライベートＱＰスペースを有し得る複数のｖＨＣＡインスタンス８１１－８１９の定義を可能にする。図８に示されるように、ｖＨＣＡ８１１－８１９の各々は、たとえばＱＰ８２１－８２９のようなキューペア（ＱＰ）をサポートし得る。当該キューペア（ＱＰ）は、パケットを送信および受信するためのたとえばアプリケーション８０３または８０４といったアプリケーションに関連付けられるセンダー／レシーバーバッファによって使用され得る。

【００６１】

さらに、拡張ＨＣＡ規格は、より複雑なファブリックトポロジに依存しないという長所

を有し得る。さらに、プライベートメモリ領域ナンバースペースは、プライベートＱＰスペースに関する同じ代替例に基づいて、ｖＨＣＡごとに実現され得る。

【００６２】

本発明の実施例に従うと、ローカルの物理メモリについてのプライベート仮想アドレススペースは、ｖＨＣＡスキームが仮想スイッチモデルまたは拡張ＨＣＡモデルに基づくかどうかに関わらず実現され得る。

【００６３】

プライベート仮想アドレススペースを実現する１つの方法は、サーバプラットフォームを使用することによる。サーバプラットフォームでは、物理的なＨＣＡによって観察されたローカルメモリアドレススペースが仮想メモリを示す。この仮想メモリは、具体的にＳＲＩＯＶ　ＶＦごとに、すなわち、ｖＨＣＡ８１１－８１９ごとにホストプラットフォームによって実現され得る。

【００６４】

プライベート仮想アドレススペースを実現する別の方法は、たとえばＶＭＭ８２０といったプラットフォームハイパーバイザによって制御されるシステムＭＭＵのセット内の物理的なＨＣＡのメモリ管理ユニット（ＭＭＵ）を含むことによる。したがって、如何なる仮想アドレスも変更することなく、仮想アドレスから物理アドレスへのマッピングが必要に応じて更新される状態で、ＶＭ固有のアドレススペースが異なる物理的なホスト同士の間でマイグレートされ得る。（たとえばプライベート仮想アドレススペースは、物理的なＨＣＡとローカルの物理メモリとの間にシステムＭＭＵを有することによって実現することができる）。

【００６５】

本発明の実施例に従うと、接続状態は、マイグレートされているＶＭのローカルメモリに保存され得る。さらに、接続状態は、マイグレーションの後、新しいｖＨＣＡについて、如何なる新しいリソースも割り当てる必要なしで再確立（または再使用）され得る。

【００６６】

さらに、（図３に示されるような）中断状態は、マイグレートされたＱＰとマイグレートされたＱＰに接続されるリモートピアＱＰとの両方に適用され得る。したがって、システムは、ＶＭマイグレーションの間にタイムアウトが発生せず、システムが、マイグレートされたＶＭのピアによって実現されるＩＢ接続関連のタイムアウト値と無関係に、マイグレーション動作が任意の量の時間をかけることを可能にし得ることを保証し得る。さらに、接続に関連付けられるＩＢアドレスがマイグレーションの結果変化することを可能にするために、システムは、中断状態におけるＱＰについてアドレス情報を更新し得る。

【００６７】

リモートＱＰピアとの通信

本発明の実施例に従うと、システムは、ピア同士の間の通信をサポートするための特別のヘッダ情報および／または非送信請求（non-solicited）メッセージを使用し得る。システムは、リモートピア同士間の休止動作を行なうことなくＶＭライブマイグレーションが開始されることを可能にする。さらに、システムは、リモートピアがマイグレーションに関して学習すること、および非同期イベントドリブンの態様で中断状態に入ることを可能にする。

【００６８】

たとえば、ｖＨＣＡルートヘッダ（ＶＲＨ）が、サポートされ得、ＩＢネットワークにおけるグローバルルートヘッダ（ＧＲＨ）ベースのアドレシングの代わりに、ＶＲＨベースのｖＨＣＡアドレシングが使用され得る。ＶＲＨを含むパケットは、たとえばＶＲＨにおける宛先ｖＨＣＡタグ（ＤＶＴ）が０でなく、かつヘッダにおけるＤＶＴが、ＶＲＨにおけるＤＶＩによって識別されたｖＨＣＡタグテーブルエントリにおけるＤＶＴと同じ場合、ＶＲＨにおいて指定されたように宛先ｖＨＣＡインデックス（ＤＶＩ）に送達され得る。システムは、ＶＲＨの存在を規定するワイヤープロトコル拡張が存在しない場合、ＬＩＤポリシーを介して、どのパケットがＶＲＨを含むと予想されるかを特定し得る。代替

的には、システムは、すべてのｖＨＣＡに共通であるＱＰナンバー範囲を規定し得、異なる範囲がＶＲＨまたは他のプロトコルポリシーの使用に関連付けられ得る。さらに、ＶＲＨは、ｖＨＣＡインデックスがＶＲＨから抽出されるまでＱＰコンテキストが知られていないので、ＱＰコンテキストを介して特定され得ない。

【００６９】

図９は、本発明の実施例に従った、マイグレートされたキューペア（ＱＰ）とリモートピアＱＰとの間の通信をサポートする図を示す。図９に示されるように、仮想化環境９００は、各々がたとえばＳＷ９１１－９１３のうちの１つのようなソフトウェアスタックを含み得るホストＡ－Ｃ９０１－９０３といった複数のホストを含み得る。（ソフトウェアスタックは、カーネルスペース、ユーザスペースまたはその両方に存在し得る）。

【００７０】

さらに、ホストＡ９０１上のＱＰａ９１０は、ＱＰａ９１０がホストＣ９０３に（ＱＰａ９３０として）マイグレートする前に、ホストＢ９０２上のＱＰｂ９２０と通信状態にあり得る。ＱＰａ９１０は、その時間の間にＱＰｂ９２０がＱＰａ９１０にパケット（ＵＤ、ＵＣまたはＲＣ）を送信し得るマイグレーションを可能にするよう中断状態にセットされ得る。ホストＡ９０１上の無効なｖＨＣＡ９３１を目標とするこのパケットは、ＧＲＨまたはＶＲＨ／ＬＩＤのいずれかのマップベースのアドレシングを使用し得る。

【００７１】

無効なｖＨＣＡ９３１を目標とするパケットを受け取った後、ホストＡ９０１は、ローカルの非同期イベントを生成し得、随意にｖＨＣＡ９３２上のＱＰｂ９２０に非送信請求応答パケットを送信し得る。さらにＱＰｂ９２０は、非送信請求応答パケットを受け取ると、当該非送信請求応答から情報を含むローカルの非同期イベントを生成し得る。その後、ＱＰｂ９２０は、ＱＰｂ９２０が、有効にされた信頼性のある接続された（reliable connected（ＲＣ））ＱＰまたは信頼性のない接続（unreliable connection（ＵＣ））ＱＰである場合、（図３に示されるように）自動的に中断状態への遷移を行ない得る。

【００７２】

本発明の実施例に従うと、ＩＢＡによって提供されるバーブズインターフェイス（verbs interface）は、キューペア（ＱＰ）を中断状態におよび中断状態から移すためにModifyＱＰおよびModifyAddressHandleオプションを含み得る。さらに、中断状態がＲＣ　ＱＰについてリセットされると、再試行カウントおよびタイマーが初期値にリセットされ得る（すなわち中断状態に入る前に消費された如何なる「再試行バジェット（retry budget）」も「忘れられ（forgotten）」得る）。中断状態にある間、パス情報はＲＣまたはＵＣＱＰについて更新され得る。ＱＰおよびアドレスハンドルは、たとえばＳＷ　Ａ－Ｃ９１１－９１３のようなローカルのソフトウェアが明示的に中断状態をリセットするまで、中断状態のままであり得る。ここで、中断状態をリセットする前に、ローカルソフトウェアは、宛先ＬＩＤおよびＶＲＨ情報を含み得る更新されたパス情報を得ることができる。

【００７３】

本発明の実施例に従うと、そのｖＧＵＩＤによって識別されたｖＨＣＡポートがＩＢネットワークにおいて作動状態になるたびに、ｖＨＣＡポートは、適切な物理的なＨＣＡポートについて、ｖＧＵＩＤおよびｖＨＣＡ情報を介して報告され得る。ＶＭ／ｖＨＣＡマイグレーションの場合には、宛先ノードの物理的なＨＣＡポートのｖＧＵＩＤおよびｖＨＣＡ情報ならびにＳＭＡ情報が、マイグレートされるｖＨＣＡが作動状態になると、更新され得る。ＶＭのマイグレーションの後に、中断されたＱＰまたはアドレスハンドルを有するリモートピアは、関連するパス情報を更新した後、関連情報を観察し得、中断状態をリセットし得る。

【００７４】

さらに、マイグレートされたＶＭおよびｖＨＣＡが新しい位置にて再び開始される場合、通信ピアは、当該マイグレートされたｖＨＣＡを認識していない場合がある。たとえば、ＱＰａ９３０のようなマイグレートされたＲＣ　ＱＰがそのピアであるＱＰｂ９２０に要求を送信すると、ＱＰａ９３０は、ＱＰｂ９２０が同じ時間フレーム中にマイグレート

されなかったので、正しいパス情報を有し得る。さらに、ＱＰｂ９２０が更新されたパス情報を有していなければ、マイグレートされたＱＰａ９３０からの入力要求に対する承認および応答が、古い宛先へ送信され得る。その後、当該古い宛先は、ＱＰｂ９２０に非送信請求応答を送信し得、これにより、パス情報が既に更新されたことをＱＰｂ９２０に認識させる。

【００７５】

さらに、たとえばホストＣ９０３のような新しい位置にて開始する、ｖＨＣＡ９３３のようなマイグレートされたｖＨＣＡは、送出要求がタイムアウトするのを回避するために、タイムリーな態様でパス情報を更新するようピアに依存し得る。たとえば、ＱＰｂ９２０が正しいパス情報を有していない場合、マイグレートされたｖＨＣＡ９３３からの送出要求は、承認および応答が戻ってこないという事実により、タイムアウトし得る。

【００７６】

マイグレートされたＱＰａ９３０からのオリジナルの要求パケットが、マイグレーションが行われたことをピアＱＰｂ９２０に通知し得ることを保証するために、システムは、マイグレーション情報を、要求パケットにおけるヘッダコンテンツから検知できるようにし得る。たとえば、ソースアドレス情報における任意の関連する変更を検知するのをピアに依存する代わりに、ＶＲＨは、マイグレーションカウントフィールドを含み得る。マイグレーションカウントフィールドは、マイグレートされたｖＨＣＡ９３０が新しい位置であるホストＣ９０３で再び作動状態になる前に更新され得る。さらに、ピアＱＰ状態に記録されるものに対するマイグレーションカウントにおける変化によって、ピアノードに関する条件を識別する非同期イベントが引き起こされ得る。したがって、ＶＲＨマイグレーションカウントフィールドを介して明示的にリモートピアに通知することによって、システムは、ほとんどの場合、マイグレートされたＨＣＡ側上のＱＰについてタイムアウトを回避することができる。

【００７７】

本発明の実施例に従うと、パス情報を更新するのに完全にピアに対してリアルタイムに依存することを回避するために、マイグレートされたｖＨＣＡは、非中断ペンディング状態のＱＰで開始し得る。非中断ペンディングＱＰ状態は、送出要求が可能にされるという例外を有する中断状態に類似する。さらに、ｖＨＣＡがマイグレーションの後で再び開始された後に生成される要求についての応答または承認がリモートピアから受け取られた後、非中断ペンディング状態はリセットされ得、ＱＰは正常なものとして挙動し得る。

【００７８】

さらに、システムの実現例（ＨＷ、ＦＷまたはＳＷ）は、タイムアウトおよび再試行を使用してパケットのロスに対応する一方、ＱＰ自体のタイムアウト挙動は中断状態におけるものと同じであり得る（すなわち、タイムアウトによりエラー状態にならない）。さらに、非中断ペンディング状態は、マイグレートされないｖＨＣＡ／ＱＰについて中断状態からの遷移を適切に行なうよう使用され得る。

【００７９】

代替的には、ＶＲＨが使用されない場合、マイグレーションの信号送信は非送信請求メッセージを介して達成され得る。たとえば、マイグレーションの後のリスタートに続いて、正常なメッセージとの間でインターリーブされた非送信請求メッセージが、場合によっては送られ得る。さらに、リモートピアＱＰからの要求または応答の受信により、このような非送信請求メッセージを送信する必要性が除去され得る。ここで、非送信請求メッセージの使用により、ＱＰ／接続に対する如何なる他の通信アクティビティとも無関係に、信号送信が行われることを可能にする。

【００８０】

したがって、付加的なヘッダ情報および／または非送信請求メッセージを使用して、システムは、マイグレートされたＱＰが、行われるマイグレーションについてそのピアＱＰに自動的に信号送信することを可能にし、含まれる新しいアドレス情報をそのピアＱＰに提供することを可能にする。アドレス情報の関連する更新は、サブネットマネージャの関

連情報の通知および報告と並行して行われ得、接続を完全な作動状態にするためにレイテンシをさらに低減し得る。

【００８１】

図１０は、本発明の実施例に従った、拡張ｖＨＣＡモデルに基づいてバーチャルマシンのライブマイグレーションをサポートするための例示的なフローチャートを示す。図１０に示されるように、ステップ１００１では、システムは、複数のバーチャルホストチャネルアダプタ（ＨＣＡ）にファブリックアダプタを関連付け得、上記バーチャルホストチャネルアダプタ（ｖＨＣＡ）の各々は、別個のキューペア（ＱＰ）スペースに関連付けられ得る。ここで、ファブリックアダプタは、仮想スイッチモデル、拡張ホストチャネルアダプタ（ＨＣＡ）モデル、または両方のモデルの組合せを実現することができる。さらに、ステップ１００２では、少なくとも１つのバーチャルマシンが、第１のホストから第２のホストにライブマイグレーションを行なうよう動作し得、上記少なくとも１つのバーチャルマシンには、上記キューペア（ＱＰ）スペースにおけるキューペア（ＱＰ）が関連付けられる上記のバーチャルホストチャネルアダプタ（ｖＨＣＡ）が付与される。その後、ステップ１００３では、上記キューペア（ＱＰ）は、ライブマイグレーションについてピアＱＰに信号送信し得、マイグレーションの後に上記ピアＱＰにアドレス情報を提供し得る。

【００８２】

スケーラブルアドレシングスキーム

本発明の実施例に従うと、レイヤー３（Ｌ３）アドレスの一部であるＶＭに関連付けられる仮想ＧＵＩＤが、ライブマイグレーションを通じて保存され得、仮想ＧＵＩＤが、ピアに通信する観点と、サブネットマネージャの観点とから一意にＶＭを識別するよう使用され得る。

【００８３】

マイグレートされているリモート接続ピアに関連付けられるアドレス情報を更新する本質的な必要性が存在するソフトウェアアプローチと異なり、拡張ＨＣＡモデルは、ＶＭに関連付けられるアドレス情報がマイグレーションに亘って保存されることを可能にする。したがって、接続ピアについてアドレス情報を更新する本質的な必要性が存在せず、システムはＩＢレイヤー２（Ｌ２）またはローカルルートヘッダ（ＬＲＨ）を保存し得るので、Ｌ３またはグローバルルートヘッダ（ＧＲＨ）アドレスに加えてアドレス情報が保存され得る。

【００８４】

本発明の実施例に従うと、システムは、ＩＢサブネットにおいてスケーラビリティおよびフレキシビリティを達成することができる。たとえば、ＩＢサブネットは、４８ＫのＬＩＤ値に限定され得る。システムは、関連するＨＣＡポート（物理的なポートまたはｖＨＣＡポート）へのパケット転送を促進するよう用いることをＩＢサブネットにおけるサブネットマネージャが決定するどのようなＬＩＤ値にでも基づきＨＣＡがアドレス指定されることを可能にするｉｇｎｏｒｅ　ＬＭＣ機能をサポートする。たとえば、ＨＣＡポートは、専用のＬＩＤ、またはベースとなる２つの隣接するＬＩＤの範囲（ＨＣＡポートＬＩＤマスク、すなわちＬＭＣ値、によって規定されるような）についての権限を有することを強制され得ない。

【００８５】

さらに、システムは、物理的なポートについて静的に割り付けられたＬＩＤ／ＬＩＤ範囲としてサブネットマネージャが使用しないＬＩＤ値を示す範囲のａｄｍｉｎ　ＬＩＤの使用をサポートし得る。さらに、サブネットマネージャは、ＶＭ作成時または動的にＶＭライフタイム中のいずれかに、アドレシングのためにユニークなＬＩＤを必要とするＶＭにａｄｍｉｎ　ＬＩＤを割り当て得る。

【００８６】

図１１は、本発明の実施例に従った、仮想化環境においてローカル識別子（ＬＩＤ）の割り当てをサポートする図を示す。図１１に示されるように、仮想化環境１１００は、物

理的なＨＣＡポート１１１１－１１１２およびバーチャルＨＣＡポート１１２１－１１２３のようなさまざまなネットワークデバイスについてＬＩＤを割り当てることを担い得るサブネットマネージャ（ＳＭ）１１０１を含み得る。

【００８７】

ＳＭ１１０１は、ある範囲のａｄｍｉｎ　ＬＩＤ１１２０が保存され得るポリシーインターフェイスをサポートし得る。各ａｄｍｉｎ　ＬＩＤは、もしポリシー入力を介して命令が与えられなければ、ＳＭ１１０１がいずれのポートにもこの値を割り当て得ないという点で特別である。

【００８８】

さらに、ＳＭ１１０１は、ポリシーインターフェイスをサポートし得る。当該ポリシーインターフェイスでは、ｖＧＵＩＤが１つ以上のこのようなａｄｍｉｎ　ＬＩＤに関連付けられ得るとともに、ｖＧＵＩＤマイグレーションは、対応するＬＩＤについてのルーティングが、関連する物理的なＨＣＡポートに対してセットアップされることを示す。さらに、ｖＧＵＩＤがａｄｍｉｎ　ＬＩＤに関連付けられていれば、ＳＭ１１０１は、ｉｇｎｏｒｅ　ＬＭＣ機能をサポートしない物理的なＨＣＡポートにｖＧＵＩＤを関連付けることを許可し得ない。

【００８９】

本発明の実施例に従うと、ａｄｍｉｎ　ＬＩＤのルーティングは、ルーティングアルゴリズムと、その時のＳＭ１０１についてアクティブなポリシーに基づき得る。ａｄｍｉｎ　ＬＩＤに関連付けられるｖＧＵＩＤについてのパス情報は、ＳＭ１０１によって規定されるポリシーに従って、割り当てられるａｄｍｉｎ　ＬＩＤを反映し得る。

【００９０】

したがって、当該システムは、ｉｇｎｏｒｅ　ＬＭＣ機能およびａｄｍｉｎ　ＬＩＤ機能を使用して、ＩＢサブネット内でのアドレシングについてスケーラビリティおよびフレキシビリティを、ＶＭライブマイグレーション動作から独立して提供し得る。一般にこれらの機能は、ＩＢサブネットの内の任意の物理的またはバーチャルＨＣＡポートに到達するのに使用され得る１つ以上の付加的なＬＩＤの動的な生成を促進するよう用いられ得る。

【００９１】

さらに、動的なサービス品質（quality of service（ＱｏＳ））、高い可用性（high availability（ＨＡ））、および／または他のマルチアドレスベースのプロパティがランタイムにて動的に確立され得る。これらの機能は、サブネット再初期化を必要とすることなく、またはＩＢサブネット内での進行中の通信に対して如何なる他のネガティブなインパクトも引き起こすことなく、任意の特定の物理的または仮想のＨＣＡポートに対して適用され得る。

【００９２】

本発明は、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な記録媒体を含む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デバイス、マシン、またはマイクロプロセッサを用いて簡便に実施され得る。ソフトウェア技術の当業者には明らかであるように、適切なソフトウェアコーディングは、熟練したプログラマによって本開示の教示に基づき容易に用意され得る。

【００９３】

いくつかの実施例では、本発明は、本発明の処理のいずれかを実行するようコンピュータをプログラムするのに用いられ得る命令を格納した記録媒体またはコンピュータ読取可能媒体であるコンピュータプログラムプロダクトを含む。当該記録媒体は、フロッピー（登録商標）ディスク、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タイプのディスク、ＲＯＭ、ＲＡＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、ＤＲＡＭ、ＶＲＡＭ、フラッシュメモリ素子、磁気または光学カード、ナノシステム（分子メモリＩＣを含む）、または命令および／またはデータを格納するのに好適な任意

のタイプの媒体もしくは装置を含み得るが、これらに限定されない。

【００９４】

本発明の上記の記載は、例示および説明目的で与えられている。網羅的であることまたは開示されたそのものの形態に本発明を限定することを意図したものではない。当業者にとっては、多くの修正例および変形例が明確であろう。上記の実施例は、本発明の原理およびその実際的な適用をもっともよく説明するために選択および記載されたものであり、これにより他の当業者が、特定の使用に好適なさまざまな修正例を考慮して、さまざまな実施例について本発明を理解するのが可能になる。本発明の範囲は、添付の特許請求の範囲およびそれらの均等物によって定義されることが意図される。

【図１】

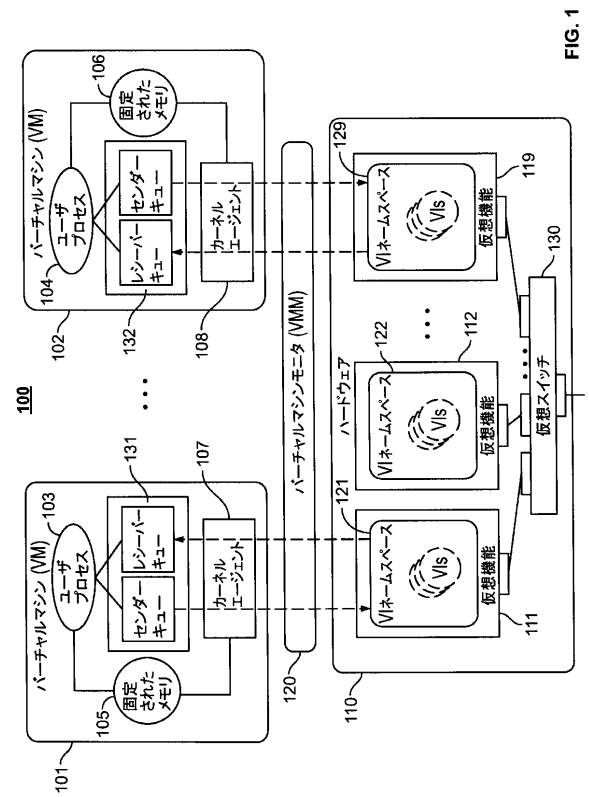

FIG. 1

【図２】

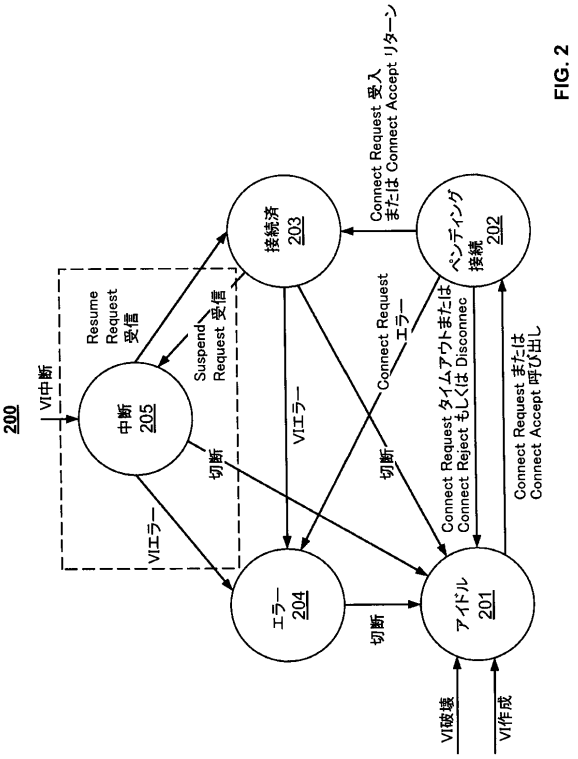

FIG. 2

【図３】

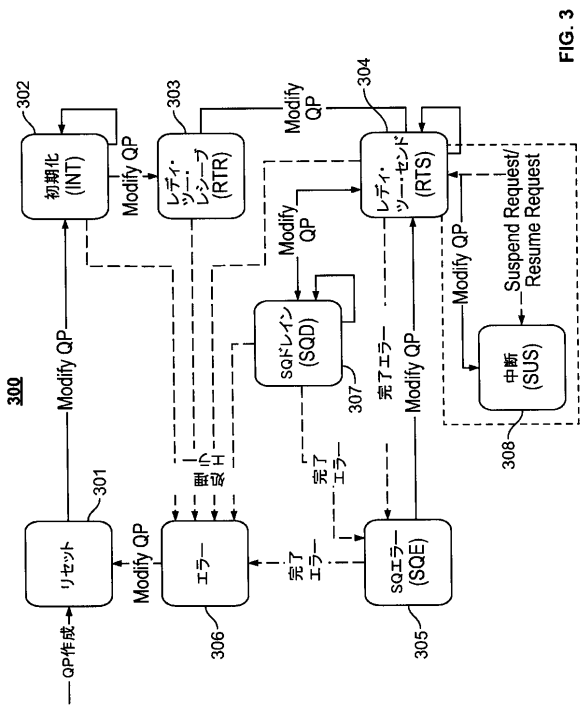

FIG. 3

【図４】

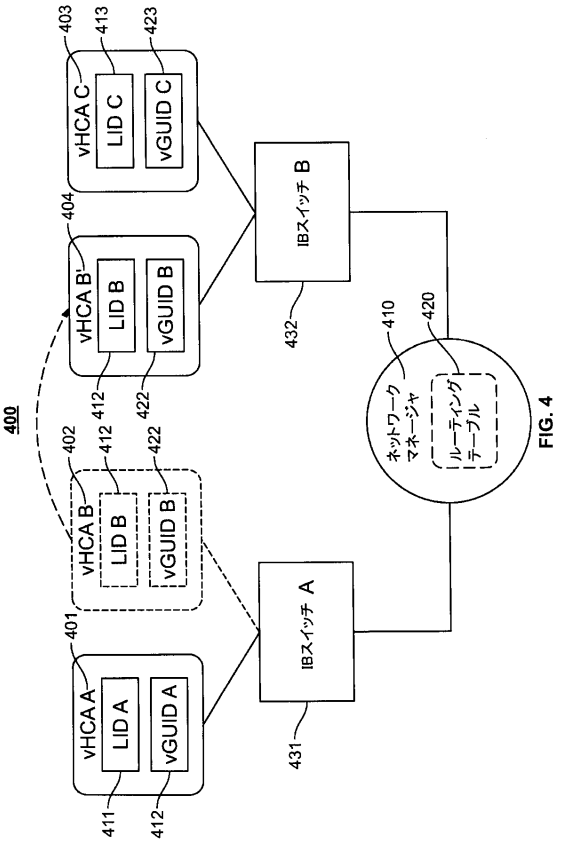

FIG. 4

【図５】

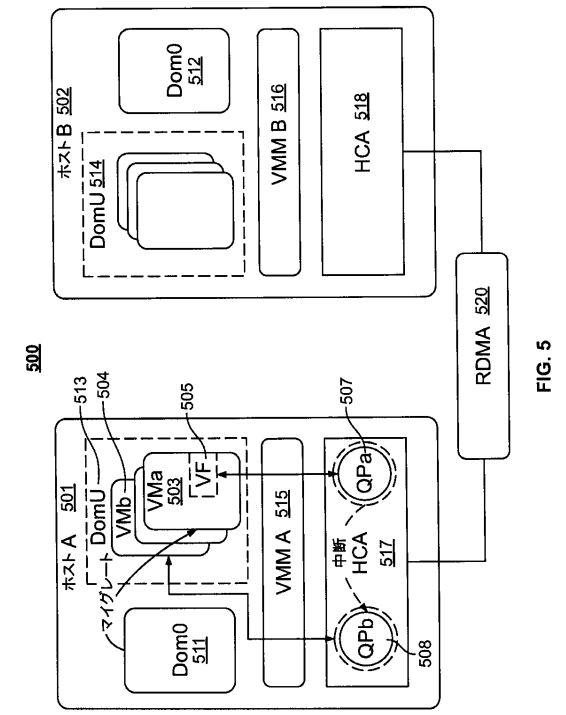

FIG. 5

【図６】

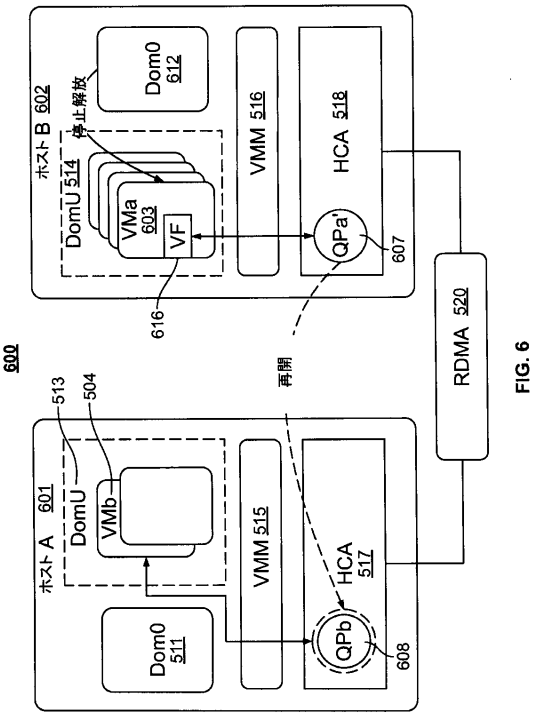

FIG. 6

【図７】

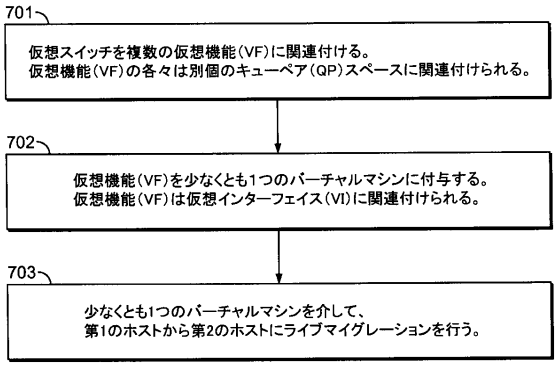

FIG. 7

【図８】

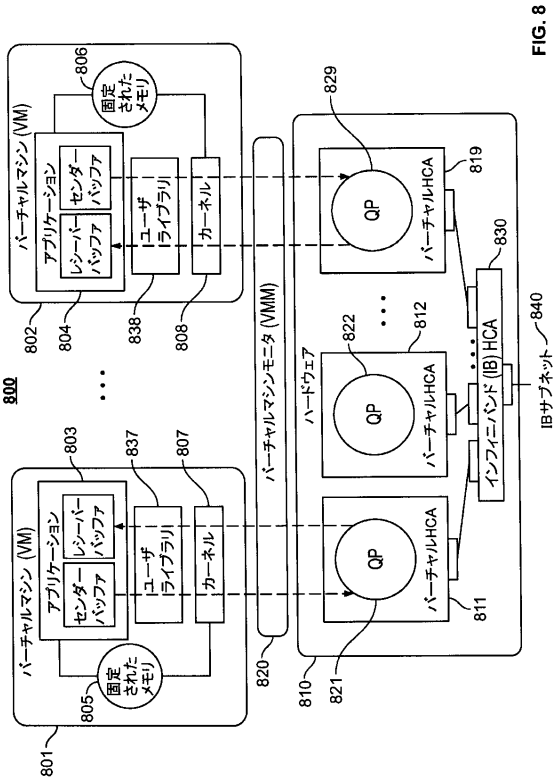

FIG. 8

【図９】

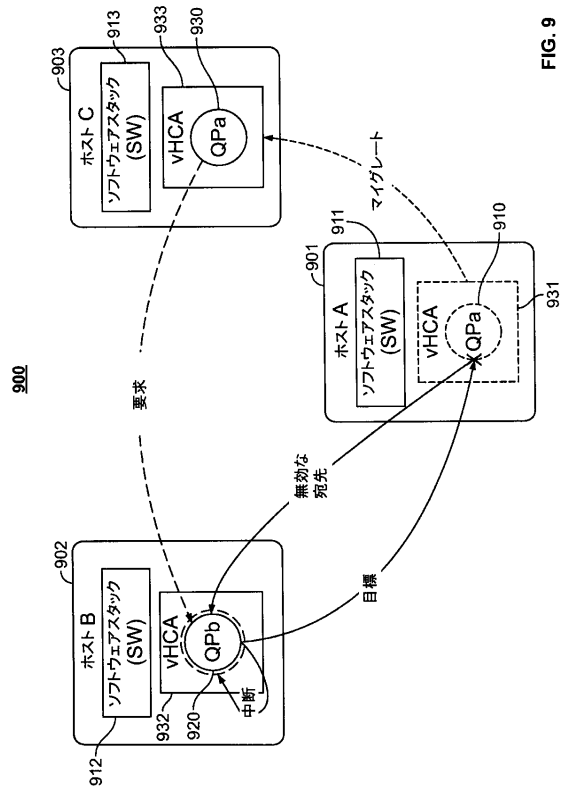

FIG. 9

【図１０】

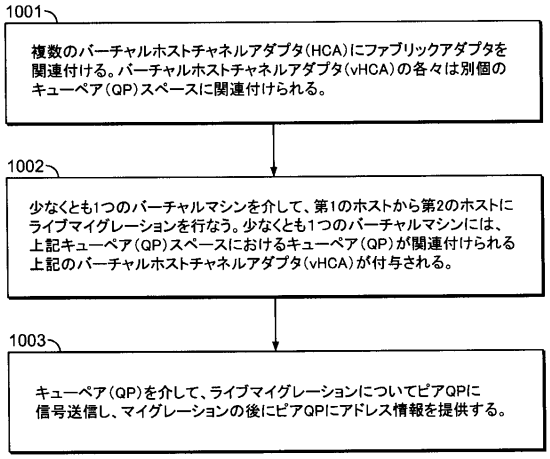

FIG. 10

【図１１】

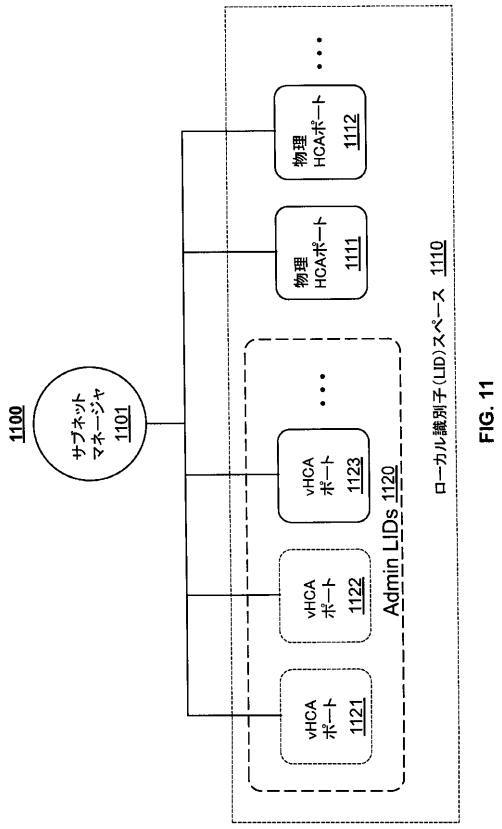
1100
サブネット
マネージャ
1101
ローカル識別子(LID)スペース 1110
物理
HCAポート
1111
物理
HCAポート
1112
Admin LIDs 1120
vHCA
ポート
1121
vHCA
ポート
1122
vHCA
ポート
1123

FIG. 11

フロントページの続き

(31)優先権主張番号 13/838,275
(32)優先日 平成25年3月15日(2013.3.15)
(33)優先権主張国 米国(US)
(31)優先権主張番号 13/837,922
(32)優先日 平成25年3月15日(2013.3.15)
(33)優先権主張国 米国(US)
(31)優先権主張番号 13/838,121
(32)優先日 平成25年3月15日(2013.3.15)
(33)優先権主張国 米国(US)
(31)優先権主張番号 13/838,502
(32)優先日 平成25年3月15日(2013.3.15)
(33)優先権主張国 米国(US)

(72)発明者 グアイ, ウェイ・リン
マレーシア、１１９５０ ペナン、バヤン・レパス、ティンカット・ブキット・ジャンブル、１、パークビュー・タワー、２５－１２－１

審査官 多賀 実

(56)参考文献 特表２０１１－５１９０８９（ＪＰ，Ａ）
特開２０１１－１８６９６７（ＪＰ，Ａ）
米国特許出願公開第２００８／０１８９４３２（ＵＳ，Ａ１）
米国特許出願公開第２０１２／００４２０３４（ＵＳ，Ａ１）
池田 宗広 外4名, 「Ｌｉｎｕｘカーネル Ｈａｃｋｓ」, 株式会社オライリー・ジャパン, ２０１１年 7月２２日, 第1版, ｐｐ. ２２５－２４１

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ ９／４６－９／５４
Ｇ０６Ｆ １３／１０