(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2019-503595
(P2019-503595A)
(43)公表日　平成31年2月7日(2019.2.7)

(51)Int.Cl.		F I			テーマコード（参考）
H04L 12/721	***(2013.01)***	H04L	12/721	Z	5K030
H04L 12/931	***(2013.01)***	H04L	12/931		5K033
H04L 12/28	***(2006.01)***	H04L	12/28	200Z	

審査請求 未請求　予備審査請求 未請求　（全 42 頁）

(21)出願番号　特願2018-502700 (P2018-502700)
(86)(22)出願日　平成29年1月25日 (2017.1.25)
(85)翻訳文提出日　平成30年1月19日 (2018.1.19)
(86)国際出願番号　PCT/US2017/014959
(87)国際公開番号　WO2017/132268
(87)国際公開日　平成29年8月3日 (2017.8.3)
(31)優先権主張番号　62/287,720
(32)優先日　平成28年1月27日 (2016.1.27)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　15/413,149
(32)優先日　平成29年1月23日 (2017.1.23)
(33)優先権主張国　米国 (US)

(71)出願人　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　ホレン，リネ
ノルウェー、１９００　フェツンド、ビタセン、１７
(72)発明者　ヨンセン，ビョルン・ダグ
ノルウェー、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】高性能コンピューティング環境における仮想ルータポートにわたるＳＭＰ接続性チェックのためのルータＳＭＡ抽象化をサポートするためのシステムおよび方法

(57)【要約】
高性能コンピューティング環境における仮想ルータにわたるＳＭＰ接続性チェックをサポートするためのシステムおよび方法を提供する。一実施形態に従うと、ＳＭＡモデル拡張は、ローカルルータポートにアドレス指定されたパケット（すなわち、ＳＭＰ）を送信する可能性を考慮に入れている。パケットがアドレス指定されているＳＭＡはパケットを受取り得るとともに、要求された情報が（たとえば、サブネットにわたる物理リンクによって接続された）リモートノード上にあることを定義する新しい属性を適用し得る。一実施形態に従うと、ＳＭＡは、（ＳＭＰを受取って別の要求を送信する）プロキシとして動作し得るか、または、ＳＭＡは、元のパケットを変更してサブネット間パケットとして送出し得る。

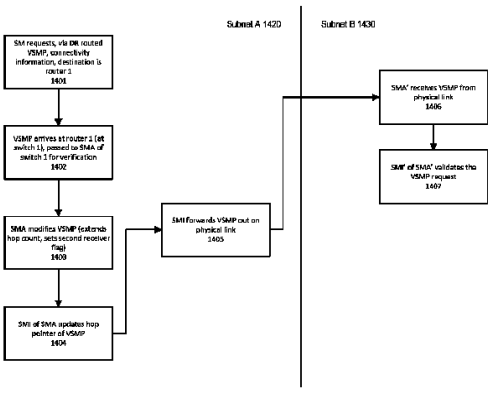

FIGURE 14

【特許請求の範囲】
【請求項１】
高性能コンピューティング環境における仮想ルータポートにわたるＳＭＰベースの接続性チェック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、
第１のサブネットとを含み、前記第１のサブネットは、
複数のスイッチを含み、前記複数のスイッチは少なくともリーフスイッチを含み、前記複数のスイッチの各々は複数のスイッチポートを含み、前記第１のサブネットはさらに、
複数のホストチャネルアダプタを含み、前記複数のホストチャネルアダプタの各々は、少なくとも１つのホストチャネルアダプタポートを含み、前記第１のサブネットはさらに、
複数のエンドノードを含み、前記複数のエンドノードの各々は、前記複数のホストチャネルアダプタのうち少なくとも１つのホストチャネルアダプタに関連付けられており、前記第１のサブネットはさらに、
サブネットマネージャを含み、前記サブネットマネージャは、前記複数のスイッチおよび前記複数のホストチャネルアダプタのうちの１つの上で実行しており、
前記複数のスイッチのうちの１スイッチ上の前記複数のスイッチポートのうちの１スイッチポートはルータポートとして構成されており、
前記ルータポートとして構成された前記１スイッチポートは仮想ルータに論理的に接続されており、
前記サブネットマネージャは、前記ルータポートにアドレス指定された要求パケットを、前記複数のスイッチのうち前記１スイッチ上にあるとともに前記ルータポートとして構成された前記複数のスイッチポートのうち前記１スイッチポートに送信し、前記要求パケットは、前記ルータポートを越えて接続性情報を要求し、
前記複数のスイッチのうちの前記１スイッチ上に常駐するサブネット管理エージェントは前記要求パケットを変更する、システム。
【請求項２】
前記仮想ルータは少なくとも２つの仮想ルータポートを含み、
前記ルータポートとして構成された前記スイッチポートは、前記少なくとも２つの仮想ルータポートのうち第１の仮想ルータポートに論理的に接続される、請求項１に記載のシステム。
【請求項３】
前記サブネット管理エージェントは、アドレスされた前記ルータポートにおいて前記要求パケットが受取られると、前記要求パケットを確認する、請求項１または２に記載のシステム。
【請求項４】
第２のサブネットをさらに含み、前記第２のサブネットは、
前記第２のサブネットの複数のスイッチを含み、前記第２のサブネットの前記複数のスイッチは、前記第２のサブネットの少なくともリーフスイッチを含み、前記第２のサブネットの前記複数のスイッチの各々は、前記第２のサブネットの複数のスイッチポートを含み、前記第２のサブネットはさらに、
前記第２のサブネットの複数のホストチャネルアダプタを含み、前記第２のサブネットの前記複数のホストチャネルアダプタの各々は、前記第２のサブネットの少なくとも１つのホストチャネルアダプタポートを含み、前記第２のサブネットはさらに、
前記第２のサブネットの複数のエンドノードを含み、前記第２のサブネットの前記複数のエンドノードの各々は、前記第２のサブネットの前記複数のホストチャネルアダプタのうち前記第２のサブネットの少なくとも１つのホストチャネルアダプタに関連付けられており、前記第２のサブネットはさらに、
前記第２のサブネットのサブネットマネージャを含み、前記第２のサブネットの前記

サブネットマネージャは、前記第２のサブネットの前記複数のスイッチおよび前記第２のサブネットの前記複数のホストチャネルアダプタのうちの１つの上で実行しており、
前記第２のサブネットの別の複数のスイッチのうち１スイッチ上における前記第２のサブネットの前記複数のスイッチポートのうち前記第２のサブネットの１スイッチポートは、前記第２のサブネットのルータポートとして構成されており、
前記第２のサブネットの前記ルータポートとして構成された前記第２のサブネットの前記１スイッチポートは前記第２のサブネットの仮想ルータに論理的に接続されており、前記第２のサブネットの前記仮想ルータは、前記第２のサブネットの少なくとも２つの仮想ルータポートを含み、
前記第１のサブネットは、物理リンクを介して前記第２のサブネットと相互接続されている、請求項１から３のいずれか１項に記載のシステム。
【請求項５】
前記要求パケットは、ＤＲ（ダイレクティットルーティング）ルーティングされたパケットであり、
前記サブネット管理エージェントによって実行された変更は、
１つ以上のホップによって、前記要求に関連付けられたホップ・カウンタを拡張することと、
リモートエンド宛先を示す受取りフラグを設定することとを含む、請求項１から４のいずれか１項に記載のシステム。
【請求項６】
前記サブネット管理エージェントに関連付けられたサブネット管理インターフェイス（ＳＭＩ）はさらに前記要求パケットを変更し、前記ＳＭＩによる前記変更はホップポインタ属性を更新することを含み、
前記ＳＭＩは、前記物理リンクを介して前記第２のサブネットの前記ルータポートに前記要求パケットを転送する、請求項４または５に記載のシステム。
【請求項７】
前記要求パケットはＬＩＤルーティングされたパケットであり、
前記サブネット管理エージェントによって実行された前記変更は、
１つ以上のホップによって、前記要求に関連付けられたホップ・カウンタを拡張することと、
リモートエンド宛先を示す受取りフラグを設定することと、
前記ＬＩＤルーティングされたパケットのアドレスに１ホップＤＲ（ダイレクティットルーティング）ルーティングされた経路を追加することとを含む、請求項１から４のいずれか１項に記載のシステム。
【請求項８】
前記サブネット管理エージェントに関連付けられたサブネット管理インターフェイス（ＳＭＩ）はさらに前記要求パケットを変更し、前記ＳＭＩによる前記変更はホップポインタ属性を更新することを含み、
前記ＳＭＩは、前記物理リンクを介して前記第２のサブネットの前記ルータポートに前記要求パケットを転送する、請求項４または５に記載のシステム。
【請求項９】
高性能コンピューティング環境における仮想ルータポートにわたるＳＭＰベースの接続性チェックをサポートするための方法であって、
１つ以上のマイクロプロセッサを含む１つ以上のコンピュータにおいて、第１のサブネットを提供するステップを含み、前記第１のサブネットは、
複数のスイッチを含み、前記複数のスイッチは少なくともリーフスイッチを含み、前記複数のスイッチの各々は複数のスイッチポートを含み、前記第１のサブネットはさらに、
複数のホストチャネルアダプタを含み、前記複数のホストチャネルアダプタの各々は少なくとも１つのホストチャネルアダプタポートを含み、前記第１のサブネットはさらに

、
複数のエンドノードを含み、前記複数のエンドノードの各々は、前記複数のホストチャネルアダプタのうち少なくとも１つのホストチャネルアダプタに関連付けられており、前記第１のサブネットはさらに、
サブネットマネージャを含み、前記サブネットマネージャは、前記複数のスイッチおよび前記複数のホストチャネルアダプタのうちの１つの上で実行しており、前記方法はさらに、
前記複数のスイッチのうち１スイッチ上において前記複数のスイッチポートのうち１スイッチポートをルータポートとして構成するステップと、
前記ルータポートとして構成された前記１スイッチポートを仮想ルータに論理的に接続するステップとを含み、前記仮想ルータは少なくとも２つの仮想ルータポートを含み、前記方法はさらに、
前記ルータポートにアドレス指定された要求パケットを、前記サブネットマネージャにより、前記複数のスイッチのうち前記１スイッチ上にあり前記ルータポートとして構成された前記複数のスイッチポートのうち前記１スイッチポートに送信するステップを含み、前記要求パケットは、前記ルータポートを越えて接続性情報を要求し、前記方法はさらに、
前記複数のスイッチのうち前記１スイッチ上に常駐するサブネット管理エージェントによって、前記要求パケットを変更するステップを含む、方法。
【請求項１０】
前記仮想ルータは少なくとも２つの仮想ルータポートを含み、
前記仮想ルータは少なくとも２つの仮想ルータポートを含み、
前記ルータポートとして構成された前記スイッチポートは、前記少なくとも２つの仮想ルータポートのうち第１の仮想ルータポートに論理的に接続される、請求項９に記載の方法。
【請求項１１】
前記サブネット管理エージェントは、アドレスされた前記ルータポートにおいて前記要求パケットが受取られると、前記要求パケットを確認する、請求項９または１０に記載の方法。
【請求項１２】
前記１つ以上のマイクロプロセッサを含む前記１つ以上のコンピュータにおいて、第２のサブネットを提供するステップをさらに含み、前記第２のサブネットは、
前記第２のサブネットの複数のスイッチを含み、前記第２のサブネットの前記複数のスイッチは、前記第２のサブネットの少なくともリーフスイッチを含み、前記第２のサブネットの前記複数のスイッチの各々は、前記第２のサブネットの複数のスイッチポートを含み、前記第２のサブネットはさらに、
前記第２のサブネットの複数のホストチャネルアダプタを含み、前記第２のサブネットの前記複数のホストチャネルアダプタの各々は、前記第２のサブネットの少なくとも１つのホストチャネルアダプタポートを含み、前記第２のサブネットはさらに、
前記第２のサブネットの複数のエンドノードを含み、前記第２のサブネットの前記複数のエンドノードの各々は、前記第２のサブネットの前記複数のホストチャネルアダプタのうち前記第２のサブネットの少なくとも１つのホストチャネルアダプタに関連付けられており、前記第２のサブネットはさらに、
前記第２のサブネットのサブネットマネージャを含み、前記第２のサブネットの前記サブネットマネージャは、前記第２のサブネットの前記複数のスイッチおよび前記第２のサブネットの前記複数のホストチャネルアダプタのうちの１つの上で実行しており、前記方法はさらに、
前記第２のサブネットの別の複数のスイッチのうちの１スイッチ上で、前記第２のサブネットの前記複数のスイッチポートのうち前記第２のサブネットの１スイッチポートを前記第２のサブネットのルータポートとして構成するステップを含み、

前記第２のサブネットの前記ルータポートとして構成された前記第２のサブネットの前記１スイッチポートは、前記第２のサブネットの仮想ルータに論理的に接続されており、前記第２のサブネットの前記仮想ルータは、前記第２のサブネットの少なくとも２つの仮想ルータポートを含み、

前記第１のサブネットは、物理リンクを介して前記第２のサブネットに相互接続されている、請求項９から１１のいずれか１項に記載の方法。

【請求項１３】

前記要求パケットは、ＤＲ（ダイレクティットルーティング）ルーティングされたパケットであり、

前記サブネット管理エージェントによって実行される変更は、

１つ以上のホップによって前記要求に関連付けられたホップ・カウンタを拡張することと、

リモートエンド宛先を示す受取りフラグを設定することとを含む、請求項９から１２のいずれか１項に記載の方法。

【請求項１４】

前記サブネット管理エージェントに関連付けられたサブネット管理インターフェイス（ＳＭＩ）はさらに、前記要求パケットを変更し、前記ＳＭＩによる前記変更はホップポインタ属性を更新することを含み、

前記ＳＭＩは、前記物理リンクを介して、前記要求パケットを前記第２のサブネットの前記ルータポートに転送する、請求項１２または１３に記載の方法。

【請求項１５】

前記要求パケットはＬＩＤルーティングされたパケットであり、

前記サブネット管理エージェントによって実行された前記変更は、

１つ以上のホップによって前記要求に関連付けられたホップ・カウンタを拡張することと、

リモートエンド宛先を示す受取りフラグを設定することと、

前記ＬＩＤルーティングされたパケットのアドレスに１ホップＤＲルーティングされた経路を追加することとを含む、請求項９から１２のいずれか１項に記載の方法。

【請求項１６】

前記サブネット管理エージェントに関連付けられたサブネット管理インターフェイス（ＳＭＩ）はさらに、前記要求パケットを変更し、前記ＳＭＩによる前記変更はホップポインタ属性を更新することを含み、

前記ＳＭＩは、前記物理リンクを介して前記第２のサブネットの前記ルータポートに前記要求パケットを転送する、請求項１２または１３に記載の方法。

【請求項１７】

高性能コンピューティング環境における仮想ルータポートにわたるＳＭＰベースの接続性チェックをサポートするための命令が格納されている非一時的なコンピュータ読取り可能記憶媒体であって、前記命令が、１つ以上のコンピュータによって読出されて実行されると、前記１つ以上のコンピュータに以下のステップを実行させ、前記以下のステップは、

１つ以上のマイクロプロセッサを含む１つ以上のコンピュータにおいて、

第１のサブネットを提供するステップを含み、前記第１のサブネットは、

複数のスイッチを含み、前記複数のスイッチは少なくともリーフスイッチを含み、前記複数のスイッチの各々は複数のスイッチポートを含み、前記第１のサブネットはさらに、

複数のホストチャネルアダプタを含み、前記複数のホストチャネルアダプタの各々は、少なくとも１つのホストチャネルアダプタポートを含み、前記第１のサブネットはさらに、

複数のエンドノードを含み、前記複数のエンドノードの各々は、前記複数のホストチャネルアダプタのうち少なくとも１つのホストチャネルアダプタに関連付けられており、

前記第１のサブネットはさらに、
サブネットマネージャを含み、前記サブネットマネージャは、前記複数のスイッチおよび前記複数のホストチャネルアダプタのうちの１つの上で実行しており、前記以下のステップはさらに、
前記複数のスイッチのうち１スイッチ上に、前記複数のスイッチポートのうち１スイッチポートをルータポートとして構成するステップと、
前記ルータポートとして構成された前記１スイッチポートを仮想ルータに論理的に接続するステップとを含み、前記仮想ルータは少なくとも２つの仮想ルータポートを含み、前記以下のステップはさらに、
前記ルータポートにアドレス指定された要求パケットを、前記サブネットマネージャによって、前記複数のスイッチのうち前記１スイッチ上にあり前記ルータポートとして構成された前記複数のスイッチポートのうち前記１スイッチポートに送信するステップを含み、前記要求パケットは、前記ルータポートを越えて接続性情報を要求し、前記以下のステップはさらに、
前記複数のスイッチのうち前記１スイッチ上に常駐するサブネット管理エージェントによって、前記要求パケットを変更するステップを含む、非一時的なコンピュータ読取り可能記憶媒体。
【請求項１８】
前記仮想ルータは少なくとも２つの仮想ルータポートを含み、
前記仮想ルータは少なくとも２つの仮想ルータポートを含み、
前記ルータポートとして構成された前記スイッチポートは、前記少なくとも２つの仮想ルータポートのうち第１の仮想ルータポートに論理的に接続される、請求項１７に記載の非一時的なコンピュータ読取り可能記憶媒体。
【請求項１９】
前記サブネット管理エージェントは、アドレス指定された前記ルータポートにおいて前記要求パケットが受取られると、前記要求パケットを確認する、請求項１７または１８に記載の非一時的なコンピュータ読取り可能記憶媒体。
【請求項２０】
前記以下のステップはさらに、
前記１つ以上のマイクロプロセッサを含む前記１つ以上のコンピュータにおいて第２のサブネットをさらに提供するステップを含み、前記第２のサブネットは、
前記第２のサブネットの複数のスイッチを含み、前記第２のサブネットの前記複数のスイッチは、前記第２のサブネットの少なくともリーフスイッチを含み、前記第２のサブネットの前記複数のスイッチの各々は、前記第２のサブネットの複数のスイッチポートを含み、前記第２のサブネットはさらに、
前記第２のサブネットの複数のホストチャネルアダプタを含み、前記第２のサブネットの前記複数のホストチャネルアダプタの各々は、前記第２のサブネットの少なくとも１つのホストチャネルアダプタポートを含み、前記第２のサブネットはさらに、
前記第２のサブネットの複数のエンドノードを含み、前記第２のサブネットの前記複数のエンドノードの各々は、前記第２のサブネットの前記複数のホストチャネルアダプタのうち前記第２のサブネットの少なくとも１つのホストチャネルアダプタに関連付けられており、前記第２のサブネットはさらに、
前記第２のサブネットのサブネットマネージャを含み、前記第２のサブネットの前記サブネットマネージャは、前記第２のサブネットの前記複数のスイッチおよび前記第２のサブネットの前記複数のホストチャネルアダプタのうちの１つの上で実行しており、前記以下のステップはさらに、
前記第２のサブネットの別の複数のスイッチうち１スイッチ上において、前記第２のサブネットの前記複数のスイッチポートのうち前記第２のサブネットの１スイッチポートを前記第２のサブネットのルータポートとして構成するステップを含み、
前記第２のサブネットの前記ルータポートとして構成された前記第２のサブネットの前

記１スイッチポートは前記第２のサブネットの仮想ルータに論理的に接続されており、前記第２のサブネットの前記仮想ルータは、前記第２のサブネットの少なくとも２つの仮想ルータポートを含み、

前記第１のサブネットは、物理リンクを介して前記第２のサブネットに相互接続されている、請求項１７から１９のいずれか１項に記載の非一時的なコンピュータ読取り可能記憶媒体。

【請求項２１】

機械読取り可能フォーマットのプログラム命令を含むコンピュータプログラムであって、前記プログラム命令がコンピュータシステムによって実行されると、前記コンピュータシステムに請求項９から１６のいずれか１項に記載の方法を実行させる、コンピュータプログラム。

【請求項２２】

非一時的な機械読取り可能データ記憶媒体に格納された、請求項２１に記載のコンピュータプログラムを含むコンピュータプログラムプロダクト。

【発明の詳細な説明】

【技術分野】

【０００１】

著作権表示：

この特許文献の開示の一部は、著作権保護の対象となる資料を含む。この特許文献または特許開示は特許商標庁の特許ファイルまたは記録に記載されているため、著作権保有者は、何人によるその複写複製に対しても異議はないが、その他の場合には如何なるときもすべての著作権を保有する。

【０００２】

発明の分野：

本発明は、概して、コンピュータシステムに関し、特に、高性能コンピューティング環境におけるデュアルポート仮想ルータポートにわたるＳＭＰ接続性チェックのためのＳＭＡ抽象化をサポートすることに関する。

【背景技術】

【０００３】

背景：

導入されるクラウドコンピューティングアーキテクチャがより大規模になるのに応じて、従来のネットワークおよびストレージに関する性能および管理の障害が深刻な問題になってきている。クラウドコンピューティングファブリックのための基礎としてインフィニバンド（登録商標）（InfiniBand：ＩＢ）技術などの高性能無損失相互接続を用いることへの関心がますます高まってきている。これは、本発明の実施形態が対応するように意図された一般領域である。

【発明の概要】

【課題を解決するための手段】

【０００４】

概要：

高性能コンピューティング環境における仮想ルータポートにわたるＳＭＰ接続性チェックをサポートするためのシステムおよび方法がここに記載される。例示的な方法は、１つ以上のマイクロプロセッサを含む１つ以上のコンピュータにおいて、第１のサブネットを提供し得る。第１のサブネットは、複数のスイッチを含み、複数のスイッチは少なくともリーフスイッチを含む。複数のスイッチの各々は複数のスイッチポートを含む。第１のサブネットは、複数のホストチャネルアダプタを含み、各々のホストチャネルアダプタは、少なくとも１つのホストチャネルアダプタポートを含む。第１のサブネットは、複数のエンドノードを含む。複数のエンドノードの各々は、複数のホストチャネルアダプタのうち少なくとも１つのホストチャネルアダプタに関連付けられている。第１のサブネットはさらに、サブネットマネージャを含む。サブネットマネージャは、複数のスイッチおよび複

数のホストチャネルアダプタのうちの１つの上で実行している。当該方法は、複数のスイッチのうち１スイッチ上において複数のスイッチポートのうち１スイッチポートをルータポートとして構成し得る。当該方法は、ルータポートとして構成されたスイッチポートを仮想ルータに論理的に接続し得る。仮想ルータは少なくとも２つの仮想ルータポートを含む。当該方法は、ルータポートにアドレス指定されたリクエストパケットを、サブネットマネージャによって、ルータポートとして構成された複数のスイッチのうち１スイッチ上における複数のスイッチポートのうち１スイッチポートに送信し得る。要求パケットは、ルータポートを越えて接続性情報を要求する。最後に、当該方法は、複数のスイッチのうち１スイッチ上に常駐するサブネット管理エージェントによって、要求パケットを変更し得る。

【０００５】

一実施形態に従うと、（第１のサブネットまたは第２のサブネットの）複数のホストチャネルアダプタのうち１つ以上は、少なくとも１つの仮想機能、少なくとも１つの仮想スイッチおよび少なくとも１つの物理機能を含み得る。（第１のサブネットまたは第２のサブネットの）複数のエンドノードは、物理ホスト、仮想マシンまたは物理ホストと仮想マシンとの組合せを含み得る。仮想マシンは、少なくとも１つの仮想機能に関連付けられている。

【図面の簡単な説明】

【０００６】

【図１】一実施形態に従ったインフィニバンド環境の一例を示す図である。

【図２】一実施形態に従った、パーティショニングされたクラスタ環境の一例を示す図である。

【図３】一実施形態に従った、ネットワーク環境におけるツリートポロジーの一例を示す図である。

【図４】一実施形態に従った例示的な共有ポートアーキテクチャを示す図である。

【図５】一実施形態に従った例示的なｖＳｗｉｔｃｈアーキテクチャを示す図である。

【図６】一実施形態に従った、例示的なｖＰｏｒｔアーキテクチャを示す図である。

【図７】一実施形態に従った、ＬＩＤが予めポピュレートされた例示的なｖＳｗｉｔｃｈアーキテクチャを示す図である。

【図８】一実施形態に従った、動的ＬＩＤ割当てがなされた例示的なｖＳｗｉｔｃｈアーキテクチャを示す図である。

【図９】一実施形態に従った、動的ＬＩＤ割当てがなされかつＬＩＤが予めポピュレートされているｖＳｗｉｔｃｈを備えた例示的なｖＳｗｉｔｃｈアーキテクチャを示す図である。

【図１０】一実施形態に従った、例示的なマルチサブネットインフィニバンドファブリックを示す図である。

【図１１】一実施形態に従った、高性能コンピューティング環境における２つのサブネット間の相互接続を示す図である。

【図１２】一実施形態に従った、高性能コンピューティング環境におけるデュアルポート仮想ルータ構成を介する２つのサブネット間の相互接続を示す図である。

【図１３】一実施形態に従った、高性能コンピューティング環境におけるデュアルポート仮想ルータをサポートするための方法を示すフローチャートである。

【図１４】一実施形態に従った、ＤＲルーティングされたＶＳＭＰパケットを示すフローチャートである。

【図１５】一実施形態に従った、ＤＲルーティングされたＶＳＭＰパケットに対する応答を示すフローチャートである。

【図１６】一実施形態に従った、ＬＩＤルーティングされたＶＳＭＰパケットを示すフローチャートである。

【図１７】一実施形態に従った、ＬＩＤルーティングされたＶＳＭＰパケットに対する応答を示すフローチャートである。

【発明を実施するための形態】
【０００７】
詳細な説明：
本発明は、同様の参照番号が同様の要素を指している添付図面の図において、限定のためではなく例示のために説明されている。なお、この開示における「ある」または「１つの」または「いくつかの」実施形態への参照は必ずしも同じ実施形態に対するものではなく、そのような参照は少なくとも１つを意味する。特定の実現例が説明されるが、これらの特定の実現例が例示的な目的のためにのみ提供されることが理解される。当業者であれば、他の構成要素および構成が、この発明の範囲および精神から逸脱することなく使用され得ることを認識するであろう。
【０００８】
図面および詳細な説明全体にわたって同様の要素を示すために、共通の参照番号が使用され得る。したがって、ある図で使用される参照番号は、要素が別のところで説明される場合、そのような図に特有の詳細な説明において参照される場合もあり、または参照されない場合もある。
【０００９】
高性能コンピューティング環境における仮想ルータポートにわたるＳＭＰ接続性チェックをサポートするためのシステムおよび方法がこの明細書中に記載される。
【００１０】
この発明の以下の説明は、高性能ネットワークについての一例として、インフィニバンド（ＩＢ）ネットワークを使用する。以下の記載全体にわたり、インフィニバンド（InfiniBand™）規格（インフィニバンド規格、ＩＢ規格またはレガシーＩＢ規格ともさまざまに称される）を参照することができる。このような参照は、全体がこの明細書中に引用により援用されている、http://www.inifinibandta.orgから入手可能な、２０１５年３月に公開されたインフィニバンドトレードアーキテクチャ規格；第１号；バージョン１．３を参照するものと理解される。他のタイプの高性能ネットワークが何ら限定されることなく使用され得ることが、当業者には明らかであるだろう。以下の説明ではまた、ファブリックトポロジーについての一例として、ファットツリートポロジーを使用する。他のタイプのファブリックトポロジーが何ら限定されることなく使用され得ることが当業者には明らかであるだろう。
【００１１】
現代（たとえばExascale（エクサスケール）時代）におけるクラウドの要求を満たすために、仮想マシンがリモート・ダイレクト・メモリ・アクセス（Remote Direct Memory Access：ＲＤＭＡ）などの低オーバーヘッドネットワーク通信パラダイムを利用できることが望ましい。ＲＤＭＡはＯＳスタックをバイパスし、ハードウェアと直接通信することで、シングルルートＩ／Ｏ仮想化（Single-Root I/O Virtualization：ＳＲ－ＩＯＶ）ネットワークアダプタのようなパススルー技術が使用可能となる。一実施形態に従うと、高性能な無損失相互接続ネットワークにおける適用可能性のために、仮想スイッチ（virtual switch：ｖＳｗｉｔｃｈ）ＳＲ－ＩＯＶアーキテクチャを提供することができる。ライブマイグレーションを実際に選択できるようにするためにネットワーク再構成時間が重要となるので、ネットワークアーキテクチャに加えて、スケーラブルであるとともにトポロジーに依存しない動的な再構成メカニズムを提供することができる。
【００１２】
一実施形態に従うと、さらには、ｖＳｗｉｔｃｈを用いる仮想化された環境のためのルーティング戦略を提供することができ、ネットワークトポロジー（たとえばファットツリートポロジー）のための効率的なルーティングアルゴリズムを提供することができる。動的な再構成メカニズムは、ファットツリーにおいて課されるオーバーヘッドを最小限にするためにさらに調整することができる。
【００１３】
本発明の一実施形態に従うと、仮想化は、クラウドコンピューティングにおける効率的

なリソース利用および融通性のあるリソース割当てに有益であり得る。ライブマイグレーションは、アプリケーションにトランスペアレントな態様で物理サーバ間で仮想マシン（virtual machine：ＶＭ）を移動させることによってリソース使用を最適化することを可能にする。このため、仮想化は、ライブマイグレーションによる統合、リソースのオン・デマンド・プロビジョニングおよび融通性を可能にし得る。

【００１４】

インフィニバンド（登録商標）

インフィニバンド（ＩＢ）は、インフィニバンド・トレード・アソシエーション（InfiniBand™ Trade Association）によって開発されたオープン標準無損失ネットワーク技術である。この技術は、特に高性能コンピューティング（high-performance computing：ＨＰＣ）アプリケーションおよびデータセンタを対象とする、高スループットおよび少ない待ち時間の通信を提供するシリアルポイントツーポイント全二重相互接続（serial point-to-point full-duplex interconnect）に基づいている。

【００１５】

インフィニバンド・アーキテクチャ（InfiniBand Architecture：ＩＢＡ）は、２層トポロジー分割をサポートする。低層では、ＩＢネットワークはサブネットと呼ばれ、１つのサブネットは、スイッチおよびポイントツーポイントリンクを使用して相互接続される一組のホストを含み得る。より高いレベルでは、１つのＩＢファブリックは、ルータを使用して相互接続され得る１つ以上のサブネットを構成する。

【００１６】

１つのサブネット内で、ホストは、スイッチおよびポイントツーポイントリンクを使用して接続され得る。加えて、サブネットにおける指定されたデバイス上に存在する、１つのマスター管理エンティティ、すなわちサブネットマネージャ（subnet manager：ＳＭ）があり得る。サブネットマネージャは、ＩＢサブネットを構成し、起動し、維持する役割を果たす。加えて、サブネットマネージャ（ＳＭ）は、ＩＢファブリックにおいてルーティングテーブル計算を行なう役割を果たし得る。ここで、たとえば、ＩＢネットワークのルーティングは、ローカルサブネットにおけるすべての送信元と宛先とのペア間の適正な負荷バランシングを目標とする。

【００１７】

サブネット管理インターフェイスを通して、サブネットマネージャは、サブネット管理パケット（subnet management packet：ＳＭＰ）と呼ばれる制御パケットを、サブネット管理エージェント（subnet management agent：ＳＭＡ）と交換する。サブネット管理エージェントは、すべてのＩＢサブネットデバイス上に存在する。ＳＭＰを使用することにより、サブネットマネージャは、ファブリックを発見し、エンドノードおよびスイッチを構成し、ＳＭＡから通知を受信することができる。

【００１８】

一実施形態によれば、ＩＢネットワークにおけるサブネット内のルーティングは、スイッチに格納されたリニアフォワーディングテーブル（linear forwarding table：ＬＦＴ）に基づき得る。ＬＦＴは、使用中のルーティングメカニズムに従って、ＳＭによって計算される。サブネットでは、エンドノード上のホストチャネルアダプタ（Host Channel Adapter：ＨＣＡ）ポートおよびスイッチが、ローカル識別子（local identifier：ＬＩＤ）を使用してアドレス指定される。ＬＦＴにおける各エントリは、宛先ＬＩＤ（destination LID：ＤＬＩＤ）と出力ポートとからなる。テーブルにおけるＬＩＤごとに１つのエントリのみがサポートされる。パケットがあるスイッチに到着すると、その出力ポートは、そのスイッチのフォワーディングテーブルにおいてＤＬＩＤを検索することによって判断される。所与の送信元－宛先ペア（ＬＩＤペア）間のネットワークにおいてパケットは同じ経路を通るため、ルーティングは決定論的である。

【００１９】

一般に、マスターサブネットマネージャを除く他のすべてのサブネットマネージャは、耐故障性のために待機モードで作動する。しかしながら、マスターサブネットマネージャ

が故障した状況では、待機中のサブネットマネージャによって、新しいマスターサブネットマネージャが取り決められる。マスターサブネットマネージャはまた、サブネットの周期的なスイープ（sweep）を行なってあらゆるトポロジー変化を検出し、それに応じてネットワークを再構成する。

【００２０】

さらに、サブネット内のホストおよびスイッチは、ローカル識別子（ＬＩＤ）を用いてアドレス指定され得るとともに、単一のサブネットは４９１５１個のユニキャストＬＩＤに制限され得る。サブネット内で有効なローカルアドレスであるＬＩＤの他に、各ＩＢデバイスは、６４ビットのグローバル一意識別子（global unique identifier：ＧＵＩＤ）を有し得る。ＧＵＩＤは、ＩＢレイヤー３（Ｌ３）アドレスであるグローバル識別子（global identifier：ＧＩＤ）を形成するために使用され得る。

【００２１】

ＳＭは、ネットワーク初期化時間に、ルーティングテーブル（すなわち、サブネット内のノードの各ペア間の接続／ルート）を計算し得る。さらに、トポロジーが変化するたびに、ルーティングテーブルは、接続性および最適性能を確実にするために更新され得る。通常動作中、ＳＭは、トポロジー変化をチェックするためにネットワークの周期的なライトスイープ（light sweep）を実行し得る。ライトスイープ中に変化が発見された場合、または、ネットワーク変化を信号で伝えるメッセージ（トラップ）をＳＭが受信した場合、ＳＭは、発見された変化に従ってネットワークを再構成し得る。

【００２２】

たとえば、ＳＭは、リンクがダウンした場合、デバイスが追加された場合、またはリンクが除去された場合など、ネットワークトポロジーが変化する場合に、ネットワークを再構成し得る。再構成ステップは、ネットワーク初期化中に行なわれるステップを含み得る。さらに、再構成は、ネットワーク変化が生じたサブネットに制限されるローカルスコープを有し得る。また、ルータを用いる大規模ファブリックのセグメント化は、再構成スコープを制限し得る。

【００２３】

一実施形態に従ったインフィニバンド環境１００の例を示す図１に、インフィニバンドファブリックの一例を示す。図１に示す例では、ノードＡ１０１～Ｅ１０５は、インフィニバンドファブリック１２０を使用して、それぞれのホストチャネルアダプタ１１１～１１５を介して通信する。一実施形態に従うと、さまざまなノード（たとえばノードＡ１０１～Ｅ１０５）はさまざまな物理デバイスによって表わすことができる。一実施形態に従うと、さまざまなノード（たとえばノードＡ１０１～Ｅ１０５）は仮想マシンなどのさまざまな仮想デバイスによって表わすことができる。

【００２４】

インフィニバンドにおけるパーティショニング

一実施形態によれば、ＩＢネットワークは、ネットワークファブリックを共有するシステムの論理グループの分離をもたらすためにセキュリティメカニズムとしてパーティショニングをサポートし得る。ファブリックにおけるノード上の各ＨＣＡポートは、１つ以上のパーティションのメンバであり得る。パーティションメンバーシップは、ＳＭの一部であり得る集中型パーティションマネージャによって管理される。ＳＭは、各ポートに関するパーティションメンバーシップ情報を、１６ビットのパーティションキー（partition key：Ｐ＿Ｋｅｙ）のテーブルとして構成することができる。ＳＭはまた、これらのポートを介してデータトラフィックを送信または受信するエンドノードに関連付けられたＰ＿Ｋｅｙ情報を含むパーティション実施テーブルを用いて、スイッチポートおよびルータポートを構成することができる。加えて、一般的な場合には、スイッチポートのパーティションメンバーシップは、（リンクに向かう）出口方向に向かってポートを介してルーティングされたＬＩＤに間接的に関連付けられたすべてのメンバーシップの集合を表わし得る。

【００２５】

一実施形態に従うと、パーティションは、１グループのメンバが単に同じ論理グループの他のメンバと通信することができるようなポートの論理グループである。ホストチャネルアダプタ（host channel adapter：ＨＣＡ）およびスイッチにおいて、分離を実施するためにパーティションメンバーシップ情報を用いてパケットをフィルタリングすることができる。無効なパーティショニング情報を備えたパケットは、パケットが受信ポートに到達すると直ちに削除され得る。パーティショニングされたＩＢシステムにおいては、パーティションはテナントクラスタを作成するために用いることができる。パーティションが適所で実施されていれば、ノードは、異なるテナントクラスタに属する他のノードと通信することができない。このようにして、安全性が損なわれたテナントノードまたは悪意あるテナントノードが存在する場合であっても、システムのセキュリティを保証することができる。

【００２６】

一実施形態に従うと、ノード間における通信のために、管理キュー対（（Queue Pair：ＱＰ）ＱＰ０およびＱＰ１）を除いて、キュー対（Queue Pair：ＱＰ）およびエンドツーエンドコンテキスト（End-to-End context：ＥＥＣ）を特定のパーティションに割当てることができる。次いで、Ｐ＿Ｋｅｙ情報は、送信されたすべてのＩＢトランスポートパケットに追加することができる。パケットがＨＣＡポートまたはスイッチに到達すると、そのＰ＿Ｋｅｙ値はＳＭによって構成されたテーブルと突き合わせて確認することができる。無効なＰ＿Ｋｅｙ値が発見された場合、パケットは直接廃棄される。このようにして、通信はパーティションを共有するポート間でのみ許可される。

【００２７】

一実施形態に従った、パーティショニングされたクラスタ環境の一例を示すＩＢパーティションの例が図２に示される。図２に示される例において、ノードＡ１０１～Ｅ１０５は、インフィニバンドファブリック１２０を用いて、それぞれのホストチャネルアダプタ１１１～１１５を介して通信する。ノードＡ～Ｅは、複数のパーティション、すなわちパーティション１　１３０、パーティション２　１４０およびパーティション３　１５０に配置される。パーティション１はノードＡ１０１およびノードＤ１０４を含む。パーティション２はノードＡ１０１、ノードＢ１０２およびノードＣ１０３を含む。パーティション３はノードＣ１０３およびノードＥ１０５を含む。パーティションの配置のせいで、ノードＤ１０４およびノードＥ１０５は、これらのノードがパーティションを共有していないために通信することが許可されない。一方で、たとえば、ノードＡ１０１およびノードＣ１０３は、これらのノードがともにパーティション２　１４０のメンバであるので、通信することが許可される。

【００２８】

インフィニバンドにおける仮想マシン

過去１０年の間に、ハードウェア仮想化サポートによってＣＰＵオーバーヘッドが実質的に排除され、メモリ管理ユニットを仮想化することによってメモリオーバーヘッドが著しく削減され、高速ＳＡＮストレージまたは分散型ネットワークファイルシステムの利用によってストレージオーバーヘッドが削減され、シングルルートＩ／Ｏ仮想化（Single Root Input/Output Virtualization：ＳＲ－ＩＯＶ）のようなデバイス・パススルー技術を使用することによってネットワークＩ／Ｏオーバーヘッドが削減されてきたことに応じて、仮想化された高性能コンピューティング（High Performance Computing：ＨＰＣ）環境の将来見通しが大幅に改善されてきた。現在では、クラウドが、高性能相互接続ソリューションを用いて仮想ＨＰＣ（virtual HPC：ｖＨＰＣ）クラスタに対応し、必要な性能を提供することができる。

【００２９】

しかしながら、インフィニバンド（ＩＢ）などの無損失ネットワークと連結されたとき、仮想マシン（ＶＭ）のライブマイグレーションなどのいくつかのクラウド機能は、これらのソリューションにおいて用いられる複雑なアドレス指定およびルーティングスキームのせいで、依然として問題となる。ＩＢは、高帯域および低レイテンシを提供する相互接

続ネットワーク技術であり、このため、ＨＰＣおよび他の通信集約型の作業負荷に非常によく適している。

【００３０】

ＩＢデバイスをＶＭに接続するための従来のアプローチは直接割当てされたＳＲ－ＩＯＶを利用することによるものである。しかしながら、ＳＲ－ＩＯＶを用いてＩＢホストチャネルアダプタ（ＨＣＡ）に割当てられたＶＭのライブマイグレーションを実現することは難易度の高いものであることが判明した。各々のＩＢが接続されているノードは、３つの異なるアドレス（すなわちＬＩＤ、ＧＵＩＤおよびＧＩＤ）を有する。ライブマイグレーションが発生すると、これらのアドレスのうち１つ以上が変化する。マイグレーション中のＶＭ（VM-in-migration）と通信する他のノードは接続性を失う可能性がある。これが発生すると、ＩＢサブネットマネージャ（Subnet Manager：ＳＭ）にサブネット管理（Subnet Administration：ＳＡ）経路記録クエリを送信することによって、再接続すべき仮想マシンの新しいアドレスを突きとめることにより、失われた接続を回復させるように試みることができる。

【００３１】

ＩＢは３つの異なるタイプのアドレスを用いる。第１のタイプのアドレスは１６ビットのローカル識別子（ＬＩＤ）である。少なくとも１つの固有のＬＩＤは、ＳＭによって各々のＨＣＡポートおよび各々のスイッチに割当てられる。ＬＩＤはサブネット内のトラフィックをルーティングするために用いられる。ＬＩＤが１６ビット長であるので、６５３６個の固有のアドレス組合せを構成することができ、そのうち４９１５１個（０×０００１－０×ＢＦＦＦ）だけをユニキャストアドレスとして用いることができる。結果として、入手可能なユニキャストアドレスの数は、ＩＢサブネットの最大サイズを定義することとなる。第２のタイプのアドレスは、製造業者によって各々のデバイス（たとえば、ＨＣＡおよびスイッチ）ならびに各々のＨＣＡポートに割当てられた６４ビットのグローバル一意識別子（ＧＵＩＤ）である。ＳＭは、ＨＣＡポートに追加のサブネット固有ＧＵＩＤを割当ててもよく、これは、ＳＲ－ＩＯＶが用いられる場合に有用となる。第３のタイプのアドレスは１２８ビットのグローバル識別子（ＧＩＤ）である。ＧＩＤは有効なＩＰｖ６ユニキャストアドレスであり、少なくとも１つが各々のＨＣＡポートに割当てられている。ＧＩＤは、ファブリックアドミニストレータによって割当てられたグローバルに固有の６４ビットプレフィックスと各々のＨＣＡポートのＧＵＩＤアドレスとを組合わせることによって形成される。

【００３２】

ファットツリー（Fat Tree：ＦＴｒｅｅ）トポロジーおよびルーティング

一実施形態によれば、ＩＢベースのＨＰＣシステムのいくつかは、ファットツリートポロジーを採用して、ファットツリーが提供する有用な特性を利用する。これらの特性は、各送信元宛先ペア間の複数経路の利用可能性に起因する、フルバイセクション帯域幅および固有の耐故障性を含む。ファットツリーの背後にある初期の概念は、ツリーがトポロジーのルート（root）に近づくにつれて、より利用可能な帯域幅を用いて、ノード間のより太いリンクを採用することであった。より太いリンクは、上位レベルのスイッチにおける輻輳を回避するのに役立てることができ、バイセクション帯域幅が維持される。

【００３３】

図３は、一実施形態に従った、ネットワーク環境におけるツリートポロジーの例を示す。図３に示すように、ネットワークファブリック２００において、１つ以上のエンドノード２０１～２０４が接続され得る。ネットワークファブリック２００は、複数のリーフスイッチ２１１～２１４と複数のスパインスイッチまたはルート（root）スイッチ２３１～２３４とを含むファットツリートポロジーに基づき得る。加えて、ネットワークファブリック２００は、スイッチ２２１～２２４などの１つ以上の中間スイッチを含み得る。

【００３４】

また、図３に示すように、エンドノード２０１～２０４の各々は、マルチホームノード、すなわち、複数のポートを介してネットワークファブリック２００のうち２つ以上の部

分に接続される単一のノード、であり得る。たとえば、ノード２０１はポートＨ１およびＨ２を含み、ノード２０２はポートＨ３およびＨ４を含み、ノード２０３はポートＨ５およびＨ６を含み、ノード２０４はポートＨ７およびＨ８を含み得る。

【００３５】

加えて、各スイッチは複数のスイッチポートを有し得る。たとえば、ルートスイッチ２３１はスイッチポート１～２を有し、ルートスイッチ２３２はスイッチポート３～４を有し、ルートスイッチ２３３はスイッチポート５～６を有し、ルートスイッチ２３４はスイッチポート７～８を有し得る。

【００３６】

一実施形態によれば、ファットツリールーティングメカニズムは、ＩＢベースのファットツリートポロジーに関して最も人気のあるルーティングアルゴリズムのうちの１つである。ファットツリールーティングメカニズムはまた、ＯＦＥＤ（Open Fabric Enterprise Distribution：ＩＢベースのアプリケーションを構築しデプロイするための標準ソフトウェアスタック）サブネットマネージャ、すなわちＯｐｅｎＳＭにおいて実現される。

【００３７】

ファットツリールーティングメカニズムの目的は、ネットワークファブリックにおけるリンクにわたって最短経路ルートを均一に広げるＬＦＴを生成することである。このメカニズムは、索引付け順序でファブリックを横断し、エンドノードの目標ＬＩＤ、ひいては対応するルートを各スイッチポートに割当てる。同じリーフスイッチに接続されたエンドノードについては、索引付け順序は、エンドノードが接続されるスイッチポートに依存し得る（すなわち、ポートナンバリングシーケンス）。各ポートについては、メカニズムはポート使用カウンタを維持することができ、新しいルートが追加されるたびに、ポート使用カウンタを使用して使用頻度が最小のポートを選択することができる。

【００３８】

一実施形態に従うと、パーティショニングされたサブネットでは、共通のパーティションのメンバではないノードは通信することを許可されない。実際には、これは、ファットツリールーティングアルゴリズムによって割当てられたルートのうちのいくつかがユーザトラフィックのために使用されないことを意味する。ファットツリールーティングメカニズムが、それらのルートについてのＬＦＴを、他の機能的経路と同じやり方で生成する場合、問題が生じる。この動作は、リンク上でバランシングを劣化させるおそれがある。なぜなら、ノードが索引付けの順序でルーティングされているからである。パーティションに気づかずにルーティングが行なわれ得るため、ファットツリーでルーティングされたサブネットにより、概して、パーティション間の分離が不良なものとなる。

【００３９】

一実施形態に従うと、ファットツリーは、利用可能なネットワークリソースでスケーリングすることができる階層ネットワークトポロジーである。さらに、ファットツリーは、さまざまなレベルの階層に配置された商品スイッチを用いて容易に構築される。さらに、ｋ－ａｒｙ－ｎ－ｔｒｅｅ、拡張された一般化ファットツリー（Extended Generalized Fat-Tree：ＸＧＦＴ）、パラレルポート一般化ファットツリー（Parallel Ports Generalized Fat-Tree：ＰＧＦＴ）およびリアルライフファットツリー（Real Life Fat-Tree：ＲＬＦＴ）を含むファットツリーのさまざまな変形例が、一般に利用可能である。

【００４０】

また、ｋ－ａｒｙ－ｎ－ｔｒｅｅは、ｎレベルのファットツリーであって、k^nエンドノードと、$n \cdot k^{n-1}$スイッチとを備え、各々が２ｋポートを備えている。各々のスイッチは、ツリーにおいて上下方向に同数の接続を有している。ＸＧＦＴファットツリーは、スイッチのための異なる数の上下方向の接続と、ツリーにおける各レベルでの異なる数の接続とをともに可能にすることによって、ｋ－ａｒｙ－ｎ－ｔｒｅｅを拡張させる。ＰＧＦＴ定義はさらに、ＸＧＦＴトポロジーを拡張して、スイッチ間の複数の接続を可能にする。多種多様なトポロジーはＸＧＦＴおよびＰＧＦＴを用いて定義することができる。しかしながら、実用化するために、現代のＨＰＣクラスタにおいて一般に見出されるファ

ットツリーを定義するために、ＰＧＦＴの制限バージョンであるＲＬＦＴが導入されている。ＲＬＦＴは、ファットツリーにおけるすべてのレベルに同じポートカウントスイッチを用いている。

【００４１】

入出力（Input/Output：Ｉ／Ｏ）仮想化

一実施形態に従うと、Ｉ／Ｏ仮想化（I/O Virtualization：ＩＯＶ）は、基礎をなす物理リソースに仮想マシン（ＶＭ）がアクセスすることを可能にすることによって、Ｉ／Ｏを利用可能にすることができる。ストレージトラフィックとサーバ間通信とを組合せると、シングルサーバのＩ／Ｏリソースにとって抗し難い高い負荷が課され、結果として、データの待機中に、バックログが発生し、プロセッサがアイドル状態になる可能性がある。Ｉ／Ｏ要求の数が増えるにつれて、ＩＯＶにより利用可能性をもたらすことができ、最新のＣＰＵ仮想化において見られる性能レベルに匹敵するように、（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融通性を向上させることができる。

【００４２】

一実施形態に従うと、Ｉ／Ｏリソースの共有を可能にして、ＶＭからリソースへのアクセスが保護されることを可能にし得るようなＩＯＶが所望される。ＩＯＶは、ＶＭにエクスポーズされる論理装置を、その物理的な実装から分離する。現在、エミュレーション、準仮想化、直接的な割当て（direct assignment：ＤＡ）、およびシングルルートＩ／Ｏ仮想化（ＳＲ－ＩＯＶ）などのさまざまなタイプのＩＯＶ技術が存在し得る。

【００４３】

一実施形態に従うと、あるタイプのＩＯＶ技術としてソフトウェアエミュレーションがある。ソフトウェアエミュレーションは分離されたフロントエンド／バックエンド・ソフトウェアアーキテクチャを可能にし得る。フロントエンドはＶＭに配置されたデバイスドライバであり得、Ｉ／Ｏアクセスをもたらすためにハイパーバイザによって実現されるバックエンドと通信し得る。物理デバイス共有比率は高く、ＶＭのライブマイグレーションはネットワークダウンタイムのわずか数ミリ秒で実現可能である。しかしながら、ソフトウェアエミュレーションはさらなる不所望な計算上のオーバーヘッドをもたらしてしまう。

【００４４】

一実施形態に従うと、別のタイプのＩＯＶ技術として直接的なデバイスの割当てがある。直接的なデバイスの割当てでは、Ｉ／ＯデバイスをＶＭに連結する必要があるが、デバイスはＶＭ間では共有されない。直接的な割当てまたはデバイス・パススルーは、最小限のオーバーヘッドでほぼ固有の性能を提供する。物理デバイスはハイパーバイザをバイパスし、直接、ＶＭに取付けられている。しかしながら、このような直接的なデバイスの割当ての欠点は、仮想マシン間で共有がなされないため、１枚の物理ネットワークカードが１つのＶＭと連結されるといったように、スケーラビリティが制限されてしまうことである。

【００４５】

一実施形態に従うと、シングルルートＩＯＶ（Single Root IOV：ＳＲ－ＩＯＶ）は、ハードウェア仮想化によって、物理装置がその同じ装置の複数の独立した軽量のインスタンスとして現われることを可能にし得る。これらのインスタンスは、パススルー装置としてＶＭに割当てることができ、仮想機能（Virtual Function：ＶＦ）としてアクセスすることができる。ハイパーバイザは、（１つのデバイスごとに）固有の、十分な機能を有する物理機能（Physical Function：ＰＦ）によってデバイスにアクセスする。ＳＲ－ＩＯＶは、純粋に直接的に割当てする際のスケーラビリティの問題を軽減する。しかしながら、ＳＲ－ＩＯＶによって提示される問題は、それがＶＭマイグレーションを損なう可能性があることである。これらのＩＯＶ技術の中でも、ＳＲ－ＩＯＶは、ほぼ固有の性能を維持しながらも、複数のＶＭから単一の物理デバイスに直接アクセスすることを可能にする手段を用いてＰＣＩ　Ｅｘｐｒｅｓｓ（ＰＣＩｅ）規格を拡張することができる。これにより、ＳＲ－ＩＯＶは優れた性能およびスケーラビリティを提供することができる。

【００４６】
ＳＲ－ＩＯＶは、ＰＣＩｅデバイスが、各々のゲストに１つの仮想デバイスを割当てることによって複数のゲスト間で共有することができる複数の仮想デバイスをエクスポーズすることを可能にする。各々のＳＲ－ＩＯＶデバイスは、少なくとも１つの物理機能（ＰＦ）と、１つ以上の関連付けられた仮想機能（ＶＦ）とを有する。ＰＦは、仮想マシンモニタ（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機能であるのに対して、ＶＦは軽量のＰＣＩｅ機能である。各々のＶＦはそれ自体のベースアドレス（base address：ＢＡＲ）を有しており、固有のリクエスタＩＤが割当てられている。固有のリクエスタＩＤは、Ｉ／Ｏメモリ管理ユニット（I/O memory management unit：ＩＯＭＭＵ）がさまざまなＶＦへの／からのトラフィックストリームを区別することを可能にする。ＩＯＭＭＵはまた、メモリを適用して、ＰＦとＶＦとの間の変換を中断する。

【００４７】
しかし、残念ながら、直接的デバイス割当て技術は、仮想マシンのトランスペアレントなライブマイグレーションがデータセンタ最適化のために所望されるような状況においては、クラウドプロバイダにとって障壁となる。ライブマイグレーションの本質は、ＶＭのメモリ内容がリモートハイパーバイザにコピーされるという点である。さらに、ＶＭがソースハイパーバイザにおいて中断され、ＶＭの動作が宛先において再開される。ソフトウェアエミュレーション方法を用いる場合、ネットワークインターフェイスは、それらの内部状態がメモリに記憶され、さらにコピーされるように仮想的である。このため、ダウンタイムは数ミリ秒にまで減らされ得る。

【００４８】
しかしながら、ＳＲ－ＩＯＶなどの直接的デバイス割当て技術が用いられる場合、マイグレーションはより困難になる。このような状況においては、ネットワークインターフェイスの内部状態全体は、それがハードウェアに結び付けられているのでコピーすることができない。代わりに、ＶＭに割当てられたＳＲ－ＩＯＶ　ＶＦが分離され、ライブマイグレーションが実行されることとなり、新しいＶＦが宛先において付与されることとなる。インフィニバンドおよびＳＲ－ＩＯＶの場合、このプロセスがダウンタイムを数秒のオーダでもたらす可能性がある。さらに、ＳＲ－ＩＯＶ共有型ポートモデルにおいては、ＶＭのアドレスがマイグレーション後に変化することとなり、これにより、ＳＭにオーバーヘッドが追加され、基礎をなすネットワークファブリックの性能に対して悪影響が及ぼされることとなる。

【００４９】
インフィニバンドＳＲ－ＩＯＶアーキテクチャ－共有ポート
さまざまなタイプのＳＲ－ＩＯＶモデル（たとえば共有ポートモデル、仮想スイッチモデルおよび仮想ポートモデル）があり得る。

【００５０】
図４は、一実施形態に従った例示的な共有ポートアーキテクチャを示す。図に示されるように、ホスト３００（たとえばホストチャネルアダプタ）はハイパーバイザ３１０と対話し得る。ハイパーバイザ３１０は、さまざまな仮想機能３３０、３４０および３５０をいくつかの仮想マシンに割当て得る。同様に、物理機能はハイパーバイザ３１０によって処理することができる。

【００５１】
一実施形態に従うと、図４に示されるような共有ポートアーキテクチャを用いる場合、ホスト（たとえばＨＣＡ）は、物理機能３２０と仮想機能３３０、３５０、３５０との間において単一の共有ＬＩＤおよび共有キュー対（Queue Pair：ＱＰ）のスペースがあるネットワークにおいて単一のポートとして現われる。しかしながら、各々の機能（すなわち、物理機能および仮想機能）はそれら自体のＧＩＤを有し得る。

【００５２】
図４に示されるように、一実施形態に従うと、さまざまなＧＩＤを仮想機能および物理

機能に割当てることができ、特別のキュー対であるＱＰ０およびＱＰ１（すなわちインフィニバンド管理パケットのために用いられる専用のキュー対）が物理機能によって所有される。これらのＱＰはＶＦにも同様にエクスポーズされるが、ＶＦはＱＰ０を使用することが許可されておらず（ＶＦからＱＰ０に向かって入来するすべてのＳＭＰが廃棄され）、ＱＰ１は、ＰＦが所有する実際のＱＰ１のプロキシとして機能し得る。

【００５３】

一実施形態に従うと、共有ポートアーキテクチャは、（仮想機能に割当てられることによってネットワークに付随する）ＶＭの数によって制限されることのない高度にスケーラブルなデータセンタを可能にし得る。なぜなら、ネットワークにおける物理的なマシンおよびスイッチによってＬＩＤスペースが消費されるだけであるからである。

【００５４】

しかしながら、共有ポートアーキテクチャの欠点は、トランスペアレントなライブマイグレーションを提供することができない点であり、これにより、フレキシブルなＶＭ配置についての可能性が妨害されてしまう。各々のＬＩＤが特定のハイパーバイザに関連付けられており、かつハイパーバイザ上に常駐するすべてのＶＭ間で共有されているので、マイグレートしているＶＭ（すなわち、宛先ハイパーバイザにマイグレートする仮想マシン）は、そのＬＩＤを宛先ハイパーバイザのＬＩＤに変更させなければならない。さらに、ＱＰ０アクセスが制限された結果、サブネットマネージャはＶＭの内部で実行させることができなくなる。

【００５５】

インフィニバンドＳＲ－ＩＯＶアーキテクチャモデル－仮想スイッチ（ｖＳｗｉｔｃｈ）

図５は、一実施形態に従った例示的なｖＳｗｉｔｃｈアーキテクチャを示す。図に示されるように、ホスト４００（たとえばホストチャネルアダプタ）はハイパーバイザ４１０と対話することができ、当該ハイパーバイザ４１０は、さまざまな仮想機能４３０、４４０および４５０をいくつかの仮想マシンに割当てることができる。同様に、物理機能はハイパーバイザ４１０によって処理することができる。仮想スイッチ４１５もハイパーバイザ４０１によって処理することができる。

【００５６】

一実施形態に従うと、ｖＳｗｉｔｃｈアーキテクチャにおいては、各々の仮想機能４３０、４４０、４５０は完全な仮想ホストチャネルアダプタ（virtual Host Channel Adapter：ｖＨＣＡ）であり、これは、ハードウェアにおいて、ＶＦに割当てられたＶＭに、ＩＢアドレス一式（たとえばＧＩＤ、ＧＵＩＤ、ＬＩＤ）および専用のＱＰスペースが割当てられていることを意味する。残りのネットワークおよびＳＭについては、ＨＣＡ４００は、仮想スイッチ４１５を介して追加のノードが接続されているスイッチのように見えている。ハイパーバイザ４１０はＰＦ４２０を用いることができ、（仮想機能に付与された）ＶＭはＶＦを用いる。

【００５７】

一実施形態に従うと、ｖＳｗｉｔｃｈアーキテクチャは、トランスペアレントな仮想化を提供する。しかしながら、各々の仮想機能には固有のＬＩＤが割当てられているので、利用可能な数のＬＩＤが速やかに消費される。同様に、多くのＬＩＤアドレスが（すなわち、各々の物理機能および各々の仮想機能ごとに１つずつ）使用されている場合、より多くの通信経路をＳＭによって演算しなければならず、それらのＬＦＴを更新するために、より多くのサブネット管理パケット（ＳＭＰ）をスイッチに送信しなければならない。たとえば、通信経路の演算は大規模ネットワークにおいては数分かかる可能性がある。ＬＩＤスペースが４９１５１個のユニキャストＬＩＤに制限されており、（ＶＦを介する）各々のＶＭとして、物理ノードおよびスイッチがＬＩＤを１つずつ占有するので、ネットワークにおける物理ノードおよびスイッチの数によってアクティブなＶＭの数が制限されてしまい、逆の場合も同様に制限される。

【００５８】

インフィニバンドＳＲ－ＩＯＶアーキテクチャモデル－仮想ポート（ｖＰｏｒｔ）

図６は、一実施形態に従った例示的なｖＰｏｒｔの概念を示す。図に示されるように、ホスト３００（たとえばホストチャネルアダプタ）は、さまざまな仮想機能３３０、３４０および３５０をいくつかの仮想マシンに割当てることができるハイパーバイザ４１０と対話することができる。同様に、物理機能はハイパーバイザ３１０によって処理することができる。

【００５９】

一実施形態に従うと、ベンダーに実装の自由を与えるためにｖＰｏｒｔ概念は緩やかに定義されており（たとえば、当該定義では、実装がＳＲＩＯＶ専用とすべきであるとは規定されていない）、ｖＰｏｒｔの目的は、ＶＭがサブネットにおいて処理される方法を標準化することである。ｖＰｏｒｔ概念であれば、空間ドメインおよび性能ドメインの両方においてよりスケーラブルであり得る、ＳＲ－ＩＯＶ共有ポートのようなアーキテクチャおよびｖＳｗｉｔｃｈのようなアーキテクチャの両方、または、これらのアーキテクチャの組合せが規定され得る。また、ｖＰｏｒｔはオプションのＬＩＤをサポートするとともに、共有ポートとは異なり、ＳＭは、ｖＰｏｒｔが専用のＬＩＤを用いていなくても、サブネットにおいて利用可能なすべてのｖＰｏｒｔを認識する。

【００６０】

インフィニバンドＳＲ－ＩＯＶアーキテクチャモデル－ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈ

一実施形態に従うと、本開示は、ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈアーキテクチャを提供するためのシステムおよび方法を提供する。

【００６１】

図７は、一実施形態に従った、ＬＩＤが予めポピュレートされた例示的なｖＳｗｉｔｃｈアーキテクチャを示す。図に示されるように、いくつかのスイッチ５０１～５０４は、ネットワーク切替環境６００（たとえばＩＢサブネット）内においてインフィニバンドファブリックなどのファブリックのメンバ間で通信を確立することができる。ファブリックはホストチャネルアダプタ５１０、５２０、５３０などのいくつかのハードウェアデバイスを含み得る。さらに、ホストチャネルアダプタ５１０、５２０および５３０は、それぞれ、ハイパーバイザ５１１、５２１および５３１と対話することができる。各々のハイパーバイザは、さらに、ホストチャネルアダプタと共に、いくつかの仮想機能５１４、５１５、５１６、５２４、５２５、５２６、５３４、５３５および５３６と対話し、設定し、いくつかの仮想マシンに割当てることができる。たとえば、仮想マシン１　５５０はハイパーバイザ５１１によって仮想機能１　５１４に割当てることができる。ハイパーバイザ５１１は、加えて、仮想マシン２　５５１を仮想機能２　５１５に割当て、仮想マシン３　５５２を仮想機能３　５１６に割当てることができる。ハイパーバイザ５３１は、さらに、仮想マシン４　５５３を仮想機能１　５３４に割当てることができる。ハイパーバイザは、ホストチャネルアダプタの各々の上で十分な機能を有する物理機能５１３、５２３および５３３を介してホストチャネルアダプタにアクセスすることができる。

【００６２】

一実施形態に従うと、スイッチ５０１～５０４の各々はいくつかのポート（図示せず）を含み得る。いくつかのポートは、ネットワーク切替環境６００内においてトラフィックを方向付けるためにリニアフォワーディングテーブルを設定するのに用いられる。

【００６３】

一実施形態に従うと、仮想スイッチ５１２、５２２および５３２は、それぞれのハイパーバイザ５１１、５２１、５３１によって処理することができる。このようなｖＳｗｉｔｃｈアーキテクチャにおいては、各々の仮想機能は完全な仮想ホストチャネルアダプタ（ｖＨＣＡ）であり、これは、ハードウェアにおいて、ＶＦに割当てられたＶＭに、ＩＢアドレス一式（たとえばＧＩＤ、ＧＵＩＤ、ＬＩＤ）および専用のＱＰスペースが割当てられていることを意味する。残りのネットワークおよびＳＭ（図示せず）については、ＨＣＡ５１０、５２０および５３０は、仮想スイッチを介して追加のノードが接続されている

スイッチのように見えている。

【００６４】

一実施形態に従うと、本開示は、ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈアーキテクチャを提供するためのシステムおよび方法を提供する。図７を参照すると、ＬＩＤは、さまざまな物理機能５１３、５２３および５３３に、さらには、仮想機能５１４～５１６、５２４～５２６、５３４～５３６（その時点でアクティブな仮想マシンに関連付けられていない仮想機能であっても）にも、予めポピュレートされている。たとえば、物理機能５１３はＬＩＤ１が予めポピュレートされており、仮想機能１　５３４はＬＩＤ１０が予めポピュレートされている。ネットワークがブートされているとき、ＬＩＤはＳＲ－ＩＯＶ　ｖＳｗｉｔｃｈ対応のサブネットにおいて予めポピュレートされている。ＶＦのすべてがネットワークにおけるＶＭによって占有されていない場合であっても、ポピュレートされたＶＦには、図７に示されるようにＬＩＤが割当てられている。

【００６５】

一実施形態に従うと、多くの同様の物理的なホストチャネルアダプタが２つ以上のポートを有することができ（冗長性のために２つのポートが共用となっている）、仮想ＨＣＡも２つのポートで表わされ、１つまたは２つ以上の仮想スイッチを介して外部ＩＢサブネットに接続され得る。

【００６６】

一実施形態に従うと、ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈアーキテクチャにおいては、各々のハイパーバイザは、それ自体のための１つのＬＩＤをＰＦを介して消費し、各々の追加のＶＦごとに１つ以上のＬＩＤを消費することができる。ＩＢサブネットにおけるすべてのハイパーバイザにおいて利用可能なすべてのＶＦを合計すると、サブネットにおいて実行することが可能なＶＭの最大量が得られる。たとえば、サブネット内の１ハイパーバイザごとに１６個の仮想機能を備えたＩＢサブネットにおいては、各々のハイパーバイザは、サブネットにおいて１７個のＬＩＤ（１６個の仮想機能ごとに１つのＬＩＤと、物理機能のために１つのＬＩＤ）を消費する。このようなＩＢサブネットにおいては、単一のサブネットについて理論上のハイパーバイザ限度は利用可能なユニキャストＬＩＤの数によって規定されており、（４９１５１個の利用可能なＬＩＤをハイパーバイザごとに１７個のＬＩＤで割って得られる）２８９１であり、ＶＭの総数（すなわち限度）は（ハイパーバイザごとに２８９１個のハイパーバイザに１６のＶＦを掛けて得られる）４６２５６である（実質的には、ＩＢサブネットにおける各々のスイッチ、ルータまたは専用のＳＭノードが同様にＬＩＤを消費するので、これらの数は実際にはより小さくなる）。なお、ｖＳｗｉｔｃｈが、ＬＩＤをＰＦと共有することができるので、付加的なＬＩＤを占有する必要がないことに留意されたい。

【００６７】

一実施形態に従うと、ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈアーキテクチャにおいては、ネットワークが最初にブートされるときに、すべてのＬＩＤについて通信経路が計算される。新しいＶＭを始動させる必要がある場合、システムは、サブネットにおいて新しいＬＩＤを追加する必要はない。それ以外の場合、経路の再計算を含め、ネットワークを完全に再構成させ得る動作は、最も時間を消費する要素となる。代わりに、ＶＭのための利用可能なポートはハイパーバイザのうちの１つに位置し（すなわち利用可能な仮想機能）、仮想マシンは利用可能な仮想機能に付与されている。

【００６８】

一実施形態に従うと、ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈアーキテクチャはまた、同じハイパーバイザによってホストされているさまざまなＶＭに達するために、さまざまな経路を計算して用いる能力を可能にする。本質的には、これは、ＬＩＤを連続的にすることを必要とするＬＭＣの制約によって拘束されることなく、１つの物理的なマシンに向かう代替的な経路を設けるために、このようなサブネットおよびネットワークがＬＩＤマスク制御ライク（LID-Mask-Control-like：ＬＭＣライク）な特徴を用いることを可能にする。ＶＭをマイグレートしてその関連するＬＩＤを宛先に送達する必要がある

場合、不連続なＬＩＤを自由に使用できることは特に有用となる。

【００６９】

一実施形態に従うと、ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈアーキテクチャについての上述の利点と共に、いくつかの検討事項を考慮に入れることができる。たとえば、ネットワークがブートされているときに、ＳＲ－ＩＯＶ　ｖＳｗｉｔｃｈ対応のサブネットにおいてＬＩＤが予めポピュレートされているので、（たとえば起動時の）最初の経路演算はＬＩＤが予めポピュレートされていなかった場合よりも時間が長くかかる可能性がある。

【００７０】

インフィニバンドＳＲ－ＩＯＶアーキテクチャモデル－動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈ

一実施形態に従うと、本開示は、動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈアーキテクチャを提供するためのシステムおよび方法を提供する。

【００７１】

図８は、一実施形態に従った、動的ＬＩＤ割当てがなされた例示的なｖＳｗｉｔｃｈアーキテクチャを示す。図に示されるように、いくつかのスイッチ５０１～５０４は、ネットワーク切替環境７００（たとえばＩＢサブネット）内においてインフィニバンドファブリックなどのファブリックのメンバ間で通信を確立することができる。ファブリックは、ホストチャネルアダプタ５１０、５２０、５３０などのいくつかのハードウェアデバイスを含み得る。ホストチャネルアダプタ５１０、５２０および５３０は、さらに、ハイパーバイザ５１１、５２１および５３１とそれぞれ対話することができる。各々のハイパーバイザは、さらに、ホストチャネルアダプタと共に、いくつかの仮想機能５１４、５１５、５１６、５２４、５２５、５２６、５３４、５３５および５３６と対話し、設定し、いくつかの仮想マシンに割当てることができる。たとえば、仮想マシン１　５５０はハイパーバイザ５１１によって仮想機能１　５１４に割当てることができる。ハイパーバイザ５１１は、加えて、仮想マシン２　５５１を仮想機能２　５１５に割当て、仮想マシン３　５５２を仮想機能３　５１６に割当てることができる。ハイパーバイザ５３１はさらに、仮想マシン４　５５３を仮想機能１　５３４に割当てることができる。ハイパーバイザは、ホストチャネルアダプタの各々の上において十分な機能を有する物理機能５１３、５２３および５３３を介してホストチャネルアダプタにアクセスすることができる。

【００７２】

一実施形態に従うと、スイッチ５０１～５０４の各々はいくつかのポート（図示せず）を含み得る。いくつかのポートは、ネットワーク切替環境７００内においてトラフィックを方向付けるためにリニアフォワーディングテーブルを設定するのに用いられる。

【００７３】

一実施形態に従うと、仮想スイッチ５１２、５２２および５３２は、それぞれのハイパーバイザ５１１、５２１および５３１によって処理することができる。このようなｖＳｗｉｔｃｈアーキテクチャにおいては、各々の仮想機能は完全な仮想ホストチャネルアダプタ（ｖＨＣＡ）であり、これは、ハードウェアにおいて、ＶＦに割当てられたＶＭに、ＩＢアドレス一式（たとえばＧＩＤ、ＧＵＩＤ、ＬＩＤ）および専用のＱＰスペースが割当てられていることを意味する。残りのネットワークおよびＳＭ（図示せず）については、ＨＣＡ５１０、５２０および５３０は、仮想スイッチを介して、追加のノードが接続されているスイッチのように見えている。

【００７４】

一実施形態に従うと、本開示は、動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈアーキテクチャを提供するためのシステムおよび方法を提供する。図８を参照すると、ＬＩＤには、さまざまな物理機能５１３、５２３および５３３が動的に割当てられており、物理機能５１３がＬＩＤ１を受取り、物理機能５２３がＬＩＤ２を受取り、物理機能５３３がＬＩＤ３を受取る。アクティブな仮想マシンに関連付けられたそれらの仮想機能はまた、動的に割当てられたＬＩＤを受取ることもできる。たとえば、仮想マシン１　５５０がアクテ

ィブであり、仮想機能１　５１４に関連付けられているので、仮想機能５１４にはＬＩＤ５が割当てられ得る。同様に、仮想機能２　５１５、仮想機能３　５１６および仮想機能１　５３４は、各々、アクティブな仮想機能に関連付けられている。このため、これらの仮想機能にＬＩＤが割当てられ、ＬＩＤ７が仮想機能２　５１５に割当てられ、ＬＩＤ１１が仮想機能３　５１６に割当てられ、ＬＩＤ９が仮想機能１　５３４に割当てられている。ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈとは異なり、アクティブな仮想マシンにその時点で関連付けられていない仮想機能はＬＩＤの割当てを受けない。

【００７５】

一実施形態に従うと、動的ＬＩＤ割当てがなされていれば、最初の経路演算を実質的に減らすことができる。ネットワークが初めてブートしており、ＶＭが存在していない場合、比較的少数のＬＩＤを最初の経路計算およびＬＦＴ分配のために用いることができる。

【００７６】

一実施形態に従うと、多くの同様の物理的なホストチャネルアダプタが２つ以上のポートを有することができ（冗長性のために２つのポートが共用となっている）、仮想ＨＣＡも２つのポートで表わされ、１つまたは２つ以上の仮想スイッチを介して外部ＩＢサブネットに接続され得る。

【００７７】

一実施形態に従うと、動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈを利用するシステムにおいて新しいＶＭが作成される場合、どのハイパーバイザ上で新しく追加されたＶＭをブートすべきであるかを決定するために、自由なＶＭスロットが発見され、固有の未使用のユニキャストＬＩＤも同様に発見される。しかしながら、新しく追加されたＬＩＤを処理するためのスイッチのＬＦＴおよびネットワークに既知の経路が存在しない。新しく追加されたＶＭを処理するために新しいセットの経路を演算することは、いくつかのＶＭが毎分ごとにブートされ得る動的な環境においては望ましくない。大規模なＩＢサブネットにおいては、新しい１セットのルートの演算には数分かかる可能性があり、この手順は、新しいＶＭがブートされるたびに繰返されなければならないだろう。

【００７８】

有利には、一実施形態に従うと、ハイパーバイザにおけるすべてのＶＦがＰＦと同じアップリンクを共有しているので、新しいセットのルートを演算する必要はない。ネットワークにおけるすべての物理スイッチのＬＦＴを繰返し、（ＶＭが作成されている）ハイパーバイザのＰＦに属するＬＩＤエントリから新しく追加されたＬＩＤにフォワーディングポートをコピーし、かつ、特定のスイッチの対応するＬＦＴブロックを更新するために単一のＳＭＰを送信するだけでよい。これにより、当該システムおよび方法では、新しいセットのルートを演算する必要がなくなる。

【００７９】

一実施形態に従うと、動的ＬＩＤ割当てアーキテクチャを備えたｖＳｗｉｔｃｈにおいて割当てられたＬＩＤは連続的である必要はない。各々のハイパーバイザ上のＶＭ上で割当てられたＬＩＤをＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈと動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈとで比較すると、動的ＬＩＤ割当てアーキテクチャにおいて割当てられたＬＩＤが不連続であり、そこに予めポピュレートされたＬＩＤが本質的に連続的となっていることが分かるだろう。さらに、ｖＳｗｉｔｃｈ動的ＬＩＤ割当てアーキテクチャにおいては、新しいＶＭが作成されると、次に利用可能なＬＩＤが、ＶＭの生存期間の間中ずっと用いられる。逆に、ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈにおいては、各々のＶＭは、対応するＶＦに既に割当てられているＬＩＤを引継ぎ、ライブマイグレーションのないネットワークにおいては、所与のＶＦに連続的に付与されたＶＭが同じＬＩＤを得る。

【００８０】

一実施形態に従うと、動的ＬＩＤ割当てアーキテクチャを備えたｖＳｗｉｔｃｈは、いくらかの追加のネットワークおよびランタイムＳＭオーバーヘッドを犠牲にして、予めポピュレートされたＬＩＤアーキテクチャモデルを備えたｖＳｗｉｔｃｈの欠点を解決する

ことができる。ＶＭが作成されるたびに、作成されたＶＭに関連付けられた、新しく追加されたＬＩＤで、サブネットにおける物理スイッチのＬＦＴが更新される。この動作のために、１スイッチごとに１つのサブネット管理パケット（ＳＭＰ）が送信される必要がある。各々のＶＭがそのホストハイパーバイザと同じ経路を用いているので、ＬＭＣのような機能も利用できなくなる。しかしながら、すべてのハイパーバイザに存在するＶＦの合計に対する制限はなく、ＶＦの数は、ユニキャストＬＩＤの限度を上回る可能性もある。このような場合、当然、アクティブなＶＭ上でＶＦのすべてが必ずしも同時に付与されることが可能になるわけではなく、より多くの予備のハイパーバイザおよびＶＦを備えることにより、ユニキャストＬＩＤ限度付近で動作する際に、断片化されたネットワークの障害を回復および最適化させるための融通性が追加される。

【００８１】

インフィニバンドＳＲ－ＩＯＶアーキテクチャモデル－動的ＬＩＤ割当てがなされかつＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈ

図９は、一実施形態に従った、動的ＬＩＤ割当てがなされてＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈを備えた例示的なｖＳｗｉｔｃｈアーキテクチャを示す。図に示されるように、いくつかのスイッチ５０１～５０４は、ネットワーク切替環境８００（たとえばＩＢサブネット）内においてインフィニバンドファブリックなどのファブリックのメンバ間で通信を確立することができる。ファブリックはホストチャネルアダプタ５１０、５２０、５３０などのいくつかのハードウェアデバイスを含み得る。ホストチャネルアダプタ５１０、５２０および５３０は、それぞれ、さらに、ハイパーバイザ５１１、５２１および５３１と対話することができる。各々のハイパーバイザは、さらに、ホストチャネルアダプタと共に、いくつかの仮想機能５１４、５１５、５１６、５２４、５２５、５２６、５３４、５３５および５３６と対話し、設定し、いくつかの仮想マシンに割当てることができる。たとえば、仮想マシン１　５５０は、ハイパーバイザ５１１によって仮想機能１　５１４に割当てることができる。ハイパーバイザ５１１は、加えて、仮想マシン２　５５１を仮想機能２　５１５に割当てることができる。ハイパーバイザ５２１は、仮想マシン３　５５２を仮想機能３　５２６に割当てることができる。ハイパーバイザ５３１は、さらに、仮想マシン４　５５３を仮想機能２　５３５に割当てることができる。ハイパーバイザは、ホストチャネルアダプタの各々の上において十分な機能を有する物理機能５１３、５２３および５３３を介してホストチャネルアダプタにアクセスすることができる。

【００８２】

一実施形態に従うと、スイッチ５０１～５０４の各々はいくつかのポート（図示せず）を含み得る。これらいくつかのポートは、ネットワーク切替環境８００内においてトラフィックを方向付けるためにリニアフォワーディングテーブルを設定するのに用いられる。

【００８３】

一実施形態に従うと、仮想スイッチ５１２、５２２および５３２は、それぞれのハイパーバイザ５１１、５２１、５３１によって処理することができる。このようなｖＳｗｉｔｃｈアーキテクチャにおいては、各々の仮想機能は、完全な仮想ホストチャネルアダプタ（ｖＨＣＡ）であり、これは、ハードウェアにおいて、ＶＦに割当てられたＶＭに、ＩＢアドレス一式（たとえばＧＩＤ、ＧＵＩＤ、ＬＩＤ）および専用のＱＰスペースが割当てられていることを意味する。残りのネットワークおよびＳＭ（図示せず）については、ＨＣＡ５１０、５２０および５３０は、仮想スイッチを介して、追加のノードが接続されているスイッチのように見えている。

【００８４】

一実施形態に従うと、本開示は、動的ＬＩＤ割当てがなされＬＩＤが予めポピュレートされたハイブリッドｖＳｗｉｔｃｈアーキテクチャを提供するためのシステムおよび方法を提供する。図９を参照すると、ハイパーバイザ５１１には、予めポピュレートされたＬＩＤアーキテクチャを備えたｖＳｗｉｔｃｈが配置され得るとともに、ハイパーバイザ５２１には、ＬＩＤが予めポピュレートされて動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈ

が配置され得る。ハイパーバイザ５３１には、動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈが配置され得る。このため、物理機能５１３および仮想機能５１４～５１６には、それらのＬＩＤが予めポピュレートされている（すなわち、アクティブな仮想マシンに付与されていない仮想機能であってもＬＩＤが割当てられている）。物理機能５２３および仮想機能１　５２４にはそれらのＬＩＤが予めポピュレートされ得るとともに、仮想機能２　５２５および仮想機能３　５２６にはそれらのＬＩＤが動的に割当てられている（すなわち、仮想機能２　５２５は動的ＬＩＤ割当てのために利用可能であり、仮想機能３　５２６は、仮想マシン３　５５２が付与されているので、１１というＬＩＤが動的に割当てられている）。最後に、ハイパーバイザ３　５３１に関連付けられた機能（物理機能および仮想機能）にはそれらのＬＩＤを動的に割当てることができる。これにより、結果として、仮想機能１　５３４および仮想機能３　５３６が動的ＬＩＤ割当てのために利用可能となるとともに、仮想機能２　５３５には、仮想マシン４　５５３が付与されているので、９というＬＩＤが動的に割当てられている。

【００８５】

ＬＩＤが予めポピュレートされたｖＳｗｉｔｃｈおよび動的ＬＩＤ割当てがなされたｖＳｗｉｔｃｈがともに（いずれかの所与のハイパーバイザ内で独立して、または組合わされて）利用されている、図９に示されるような一実施形態に従うと、ホストチャネルアダプタごとの予めポピュレートされたＬＩＤの数はファブリックアドミニストレータによって定義することができ、（ホストチャネルアダプタごとに）０＜＝予めポピュレートされたＶＦ＜＝総ＶＦの範囲内になり得る。動的ＬＩＤ割当てのために利用可能なＶＦは、（ホストチャネルアダプタごとに）ＶＦの総数から予めポピュレートされたＶＦの数を減じることによって見出すことができる。

【００８６】

一実施形態に従うと、多くの同様の物理的なホストチャネルアダプタが２つ以上のポートを有することができ（冗長性のために２つのポートが共用となっている）、仮想ＨＣＡも２つのポートで表わされ、１つまたは２つ以上の仮想スイッチを介して外部ＩＢサブネットに接続され得る。

【００８７】

インフィニバンド－サブネット間通信（ファブリックマネージャ）

一実施形態に従うと、単一のサブネット内においてインフィニバンドファブリックを提供することに加えて、本開示の実施形態はまた、２つ以上のサブネットにわたるインフィニバンドファブリックを提供することができる。

【００８８】

図１０は、一実施形態に従った、例示的なマルチサブネットインフィニバンドファブリックを示す。図に示されるように、サブネットＡ１０００内において、いくつかのスイッチ１００１～１００４は、インフィニバンドファブリックなどのファブリックのメンバ間でサブネットＡ　１０００（たとえばＩＢサブネット）内における通信を提供することができる。ファブリックは、たとえばチャネルアダプタ１０１０などのいくつかのハードウェアデバイスを含み得る。ホストチャネルアダプタ１０１０は、さらに、ハイパーバイザ１０１１と対話し得る。ハイパーバイザは、さらに、それが対話するホストチャネルアダプタと共に、いくつかの仮想機能１０１４を設定し得る。ハイパーバイザは、加えて、仮想機能１　１０１４に割当てられている仮想マシン１　１０１５などの仮想マシンを仮想機能の各々に割当て得る。ハイパーバイザは、ホストチャネルアダプタの各々の上で、物理機能１０１３などの十分な機能を有する物理機能によって、それらの関連付けられたホストチャネルアダプタにアクセスし得る。サブネットＢ１０４０内では、いくつかのスイッチ１０２１～１０２４は、インフィニバンドファブリックなどのファブリックのメンバ間でサブネットＢ１０４０（たとえばＩＢサブネット）内において通信を提供し得る。ファブリックは、たとえばチャネルアダプタ１０３０などのいくつかのハードウェアデバイスを含み得る。ホストチャネルアダプタ１０３０は、さらに、ハイパーバイザ１０３１と対話し得る。ハイパーバイザは、さらに、それが対話するホストチャネルアダプタと共に

、いくつかの仮想機能１０３４を設定し得る。ハイパーバイザは、加えて、仮想機能２　１０３４に割当てられている仮想マシン２　１０３５などの仮想マシンを仮想機能の各々に割当て得る。ハイパーバイザは、各々のホストチャネルアダプタ上で、物理機能１０３３などの十分な機能を有する物理機能によってそれらの関連付けられたホストチャネルアダプタにアクセスし得る。なお、１つのホストチャネルアダプタだけが各々のサブネット（すなわちサブネットＡおよびサブネットＢ）内において図示されているが、複数のホストチャネルアダプタおよびそれらの対応するコンポーネントが各々のサブネット内に含まれ得ることが理解されるはずである。

【００８９】

一実施形態に従うと、上述のとおり、ホストチャネルアダプタの各々は、加えて、仮想スイッチ１０１２および仮想スイッチ１０３２などの仮想スイッチに関連付けられ得るとともに、各々のＨＣＡはさまざまなアーキテクチャモデルで設定され得る。図１０内における両方のサブネットは予めポピュレートされたＬＩＤアーキテクチャモデルを備えたｖＳｗｉｔｃｈを用いるものとして示されているが、これは、このようなすべてのサブネット構成が同様のアーキテクチャモデルに従うべきであることを示唆するよう意図したものではない。

【００９０】

一実施形態に従うと、サブネットＡ１０００内のスイッチ１００２がルータ１００５に関連付けられ、サブネットＢ１０４０内のスイッチ１０２１がルータ１００６に関連付けられているなどのように、各々サブネット内の少なくとも１つのスイッチがルータに関連付けられ得る。

【００９１】

一実施形態に従うと、少なくとも１つのデバイス（たとえばスイッチ、ノード…など）は、ファブリックマネージャ（図示せず）に関連付けられ得る。ファブリックマネージャを用いることにより、たとえば、サブネット間ファブリックトポロジーを発見し、ファブリックプロファイル（たとえば仮想マシンファブリックプロファイル）を作成し、仮想マシンファブリックプロファイルを構築するための基礎を形成する仮想マシン関連のデータベースオブジェクトを構築することができる。加えて、ファブリックマネージャは、どのサブネットが、どのルータポートを介して、どのパーティション番号を用いて通信することが許可されているかについて、法的なサブネット間接続性を定義し得る。

【００９２】

一実施形態に従うと、サブネットＡ内における仮想マシン１などの送信元におけるトラフィックが、サブネットＢ内の仮想マシン２などの異なるサブネットにおける宛先にアドレス指定されると、トラフィックは、サブネットＡ内のルータ（すなわち、ルータ１００５）にアドレス指定され得る、次いで、ルータ１００５は、ルータ１００６とのリンクを介してサブネットＢにトラフィックを渡し得る。

【００９３】

仮想デュアルポートルータ

一実施形態に従うと、デュアルポートルータの抽象化は、サブネットとサブネットとの間のルータ機能が、通常のＬＲＨベースの切替えを実行することに加えて、グローバル・ルート・ヘッダ（global route header：ＧＲＨ）とローカル・ルート・ヘッダ（local route header：ＬＲＨ）との変換を実行する能力を有するスイッチハードウェア実装に基づいて定義されることを可能にするための単純な方法を提供することができる。

【００９４】

一実施形態に従うと、仮想デュアルポートルータは、対応するスイッチポートの外部に論理的に接続され得る。この仮想デュアルポートルータは、サブネットマネージャなどの標準的な管理エンティティにインフィニバンド規格対応の概念を提供することができる。

【００９５】

一実施形態に従うと、デュアルポート・ルータモデルは、各々のサブネットが、サブネットに対する入口経路におけるアドレスマッピングおよびパケットの転送を十分に制御す

るような態様で、かつ、不正確に接続されたサブネットのいずれかの内部におけるルーティングおよび論理接続性に影響を及ぼすことなく、さまざまなサブネットが接続可能であることを示唆している。

【００９６】

一実施形態に従うと、不正確に接続されたファブリックを伴う状況においては、仮想デュアルポートルータ抽象化を用いることにより、リモートサブネットに対して意図せず物理的に接続性している状態でサブネットマネージャおよびＩＢ診断ソフトウェアなどの管理エンティティが正確に機能することをも可能にし得る。

【００９７】

図１１は、一実施形態に従った、高性能コンピューティング環境における２つのサブネット間の相互接続を示す。仮想デュアルポートルータを備えた構成に先だって、サブネットＡ１１０１におけるスイッチ１１２０は、スイッチ１１２０のスイッチポート１１２１を通じて、物理的接続１１１０を介し、サブネットＢ１１０２におけるスイッチ１１３０に、スイッチ１１３０のスイッチポート１１３１を介して接続され得る。このような実施形態においては、各々のスイッチポート１１２１および１１３１は、スイッチポートおよびルータポートの両方の機能を果たし得る。

【００９８】

一実施形態に従うと、この構成に関する問題は、インフィニバンドサブネットにおけるサブネットマネージャなどの管理エンティティが、スイッチポートおよびルータポートの両方である物理ポートを区別することができない点である。このような状況においては、ＳＭは、スイッチポートを、そのスイッチポートに接続されたルータポートを有するものとして取扱うことができる。しかしながら、スイッチポートが、別のサブネットマネージャとともに、たとえば物理リンクを介して別のサブネットに接続される場合、サブネットマネージャは物理リンク上で発見メッセージを送出することができる。しかしながら、このような発見メッセージは他のサブネットにおいて許可することはできない。

【００９９】

図１２は、一実施形態に従った、高性能コンピューティング環境におけるデュアルポート仮想ルータ構成を介する２つのサブネット間の相互接続を示す。

【０１００】

一実施形態に従うと、構成後、あるサブネットマネージャが担当しているサブネットの端部を表わしている適切なエンドノードを当該サブネットマネージャが認識するように、デュアルポート仮想ルータ構成が提供され得る。

【０１０１】

一実施形態に従うと、サブネットＡ１２０１におけるスイッチ１２２０において、スイッチポートは、仮想リンク１２２３を介して仮想ルータ１２１０におけるルータポート１２１１に接続（すなわち、論理的に接続）され得る。仮想ルータ１２１０（たとえばデュアルポート仮想ルータ）は、スイッチ１２２０の外部にあるものとして示されており、実施形態においては、スイッチ１２２０内に論理的に含まれ得るものであるが、第２のルータポート（ルータポートＩＩ１２１２）を含み得る。一実施形態に従うと、２つの端部を有し得る物理リンク１２０３は、ルータポートＩＩ１２１２と、サブネットＢ１２０２における仮想ルータ１２３０に含まれるルータポートＩＩ１２３２とを介して、物理リンクの第１の端部を介するサブネットＡ１２０１を、物理リンクの第２の端部を介するサブネットＢ１２０２と、接続することができる。仮想ルータ１２３０は、加えて、ルータポート１２３１を含み得る。ルータポート１２３１は、仮想リンク１２３３を介して、スイッチ１２４０上のスイッチポート１２４１に接続（すなわち、論理的に接続）され得る。

【０１０２】

一実施形態に従うと、サブネットＡ上のサブネットマネージャ（図示せず）は、サブネットマネージャが制御するサブネットのエンドポイントとして、仮想ルータ１２１０上においてルータポート１２１１を検出し得る。デュアルポート仮想ルータ抽象化は、サブネットＡ上のサブネットマネージャが、（たとえば、インフィニバンド規格に準拠して定義

されるような）通常の態様でサブネットＡを処理することを可能にし得る。サブネット管理エージェントレベルにおいて、ＳＭが通常のスイッチポートを認識するように、デュアルポート仮想ルータ抽象化が提供され得るとともに、ＳＭＡレベルにおいては、スイッチポートに接続された別のポートが存在する抽象化が行なわれ、このポートはデュアルポート仮想ルータ上のルータポートとなる。ローカルＳＭにおいては、従来のファブリックトポロジーを使用し続けることができる（ＳＭは当該ポートをトポロジーにおける標準スイッチポートとして認識する）。このため、ＳＭはルータポートをエンドポートとして認識する。物理的接続は、２つの異なるサブネットにおけるルータポートとしても構築される２つのスイッチポート間において構成され得る。

【０１０３】

一実施形態に従うと、デュアルポート仮想ルータはまた、物理リンクが同じサブネットにおける他の何らかのスイッチポートに、または、別のサブネットに接続されるよう意図されていなかったスイッチポートに、間違って接続され得るという問題を解決することもできる。したがって、この明細書中に記載される方法およびシステムはまた、サブネットの外部に存在するものも表わしている。

【０１０４】

一実施形態に従うと、サブネットＡなどのサブネット内においては、ローカルＳＭがスイッチポートを決定し、次いで、そのスイッチポートに接続されるルータポート（たとえば、仮想リンク１２２３を介してスイッチポート１２２１に接続されるルータポート１２１１）を決定する。ＳＭが、当該ＳＭが管理しているサブネットの端部としてルータポート１２１１を認識するので、ＳＭは、この点（たとえばルータポートＩＩ１２１２）を越えて発見メッセージおよび／または管理メッセージを送信することができない。

【０１０５】

一実施形態に従うと、上述のデュアルポート仮想ルータは、デュアルポート仮想ルータが属するサブネット内において、デュアルポート仮想ルータ抽象化が管理エンティティ（たとえばＳＭまたはＳＭＡ）によって完全に管理されるという利点を提供する。ローカル側でのみ管理を許可することにより、システムは、外部の独立した管理エンティティを提供する必要がなくなる。すなわち、サブネット接続に対するサブネットの両サイドが、それ自体のデュアルポート仮想ルータを構成する役割を果たし得る。

【０１０６】

一実施形態に従うと、離れた宛先（すなわち、ローカルサブネットの外側）にアドレス指定されているＳＭＰなどのパケットが上述のデュアルポート仮想ルータとして構成されていないローカルターゲットポートに到達する状況においては、ローカルポートは、それがルータポートではないことを規定するメッセージを戻すことができる。

【０１０７】

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において、それらを用いて、またはそれらの支援により、実行可能である。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現され得る。

【０１０８】

図１３は、一実施形態に従った、高性能コンピューティング環境におけるデュアルポート仮想ルータをサポートするための方法を示す。ステップ１３１０において、当該方法は、１つ以上のマイクロプロセッサを含む１つ以上のコンピュータにおいて、第１のサブネットを提供し得る。第１のサブネットは複数のスイッチを含み、複数のスイッチは少なくともリーフスイッチを含む。複数のスイッチの各々は複数のスイッチポートを含む。第１のサブネットはさらに、複数のホストチャネルアダプタを含む。各々のホストチャネルアダプタは少なくとも１つのホストチャネルアダプタポートを含む。第１のサブネットはさらに、複数のエンドノードを含む。エンドノードの各々は、複数のホストチャネルアダプタのうち少なくとも１つのホストチャネルアダプタに関連付けられている。第１のサブネットはさらに、複数のスイッチおよび複数のホストチャネルアダプタのうち１つの上で実

行しているサブネットマネージャを含む。

【０１０９】

ステップ１３２０において、当該方法は、複数のスイッチのうち１スイッチ上において複数のスイッチポートのうち１スイッチポートをルータポートとして構成し得る。

【０１１０】

ステップ１３３０において、当該方法は、ルータポートとして構成されたスイッチポートを仮想ルータに論理的に接続し得る。仮想ルータは少なくとも２つの仮想ルータポートを含む。

【０１１１】

ＳＭＰベースの接続性チェックを可能にするためのルータＳＭＡ抽象化

一実施形態に従うと、サブネット管理パケット（Subnet Management Packet：ＳＭＰ）は、パケットがルータポートを越えて送信されるであろうことを示唆するアドレス指定情報を有することが許可されていない。しかしながら、（仮想）ルータポートの遠隔側での物理的接続性の発見（すなわち、遠隔接続性のローカルな発見）を可能にするために、ＳＭＡ属性の新しいセットが定義され得る。この場合、このような各々の新しい属性は、標準的なＳＭＡ属性またはベンダー特有のＳＭＡ属性とともに「遠隔情報」を表わしている。

【０１１２】

一実施形態に従うと、ルータＳＭＡが「遠隔」情報／属性を表わす属性を処理すると、次いで、対応するＳＭＰ要求は、元の要求の送信側に完全にトランスペアレントな態様で外部の物理リンク上で送信することができる。

【０１１３】

一実施形態に従うと、ローカルＳＭＡは、受信要求とは無関係に遠隔からの発見を実行して関連情報をローカルにキャッシュすることを選択し得るか、または、単純なプロキシのように動作して、「遠隔」情報／属性を指定する要求を受取るたびに外部リンクに対する対応する要求を生成し得る。

【０１１４】

一実施形態に従うと、「遠隔」属性を要求するＳＭＰがローカルサブネット側（すなわち、ローカルスイッチポートに論理的に接続されている仮想ルータポート）から、または外部リンク（すなわち、仮想ルータの遠隔側）から、受取られたかどうかを追跡することによって、ＳＭＡ実装は、ローカルサブネットにおける元の要求を表現する観点、または、ピアルータポートからのプロキシ要求を表現する観点から、遠隔要求がどの程度まで有効であるかを追跡することができる。

【０１１５】

一実施形態に従うと、ＩＢ規格に準拠したＳＭの場合、ルータポートはサブネットにおけるエンドポートである。このため、（発見に用いられてサブネットを構成する）低レベルＳＭＰはルータポートにわたって送信することができない。しかしながら、サブネット間トラフィックのためのルートを維持するために、ローカルＳＭまたはファブリックマネージャは、ローカルリソースのいずれかの構成を構築する前に、物理リンクの遠隔側での物理的接続性を観察可能である必要がある。しかしながら、遠隔接続性を認識したいとの要望に関して、ＳＭは、物理リンクの遠隔側を構成することを許可することができない（すなわち、ＳＭの構成はそれ自体のサブネット内に含まれていなければならない）。

【０１１６】

一実施形態に従うと、ＳＭＡモデル拡張は、ローカルルータポートにアドレス指定されているパケット（すなわちＳＭＰ）を送信する可能性を考慮に入れている。パケットがアドレス指定されているＳＭＡは、パケットを受取り得るとともに、次いで、要求された情報が（たとえば、サブネットにわたる物理リンクによって接続されている）リモートノード上にあることを定義している新しい属性を適用し得る。

【０１１７】

一実施形態に従うと、ＳＭＡは、プロキシとして動作する（ＳＭＰを受取って別の要求

を送信する）ことができるか、または、ＳＭＡは元のパケットを変更して、これをサブネット間パケットとして送信することができる。ＳＭＡは、パケットにおけるアドレス情報を更新することができる。これにより、ＳＭＰに対する１ホップ・ダイレクトルート経路の追加が更新される。次いで、リモートノード（ルータポート）によってＳＭＰを受取ることができる。ＳＭＰは、リモートエンド上のノードが同じ態様で（たとえば仮想ルータとして）構成されているか、または、基本的なスイッチポート（たとえば、レガシースイッチ実装に対する物理的接続性）として構成されているかどうかとは無関係に、機能し得る。次いで、受信ノードが認識し得る要求パケットは基本的な要求となり、通常の態様で応答することとなる。実際には、サブネットを越えて発信された要求は受信ノードに対してトランスペアレント（不可視）である。

【０１１８】

一実施形態に従うと、これは、抽象化を利用することによって、ローカルサブネットによるリモートサブネットの発見を可能にする。提供されるＳＭＡ抽象化は、遠隔側で、ローカルな（すなわち、実行中のサブネット発見からは離れている）サブネットから問合わせされていたことを認識することなく、（たとえば物理リンクにわたって）リモートサブネットから情報を検索することを可能にする。

【０１１９】

アドレス指定スキーム

一実施形態に従うと、ＩＢ規格と準拠させたままにするために、ＳＭＰパケットは、サブネットの境界によって連結されている（すなわち、ＳＭは、それが関連付けられているサブネットの外にある情報を「認識する」かまたは発見することが許可されていない）。しかしながら、仮想ルータポートのリモートエンドからの接続性情報などの情報（すなわち、サブネット境界を越えた１「ホップ」）を検索する必要性が依然として存在する。この情報は、ベンダー特有のＳＭＰ（vendor specific SMP：ＶＳＭＰ）と称され得る特別なタイプのＳＭＰに符号化することができる。

【０１２０】

一実施形態に従うと、ＶＳＭＰは、概して、一般的なＳＭＰ（サブネット向けのＳＭＰ）と同様のアドレス指定方式を利用することができ、ＤＲ（Directed Routing（ダイレクティッドルーティング）：この場合、ＳＭＰは、スイッチ間を移動しているときにどのポートから生じているかを明確に示すことができる）およびＬＩＤの両方のルーティングを用いることができる。しかしながら、ルータポートのリモートエンドに適用され得る属性のために、属性修飾子におけるシングルビットを用いてローカルポート対リモートポートを示すことができる。属性のリモートビットが設定されている場合、付加的な処理がルータポートを表わすＳＭＡにおいて生じ得る。

【０１２１】

一実施形態に従うと、アドレス指定方式の重要な局面は、システムの構成または配線が間違ってなされていたとしても、ルータポートのリモート・ピア・ポートが要求に応答することができるという点である。たとえば、仮想ルータが、リモートサブネットの汎用のスイッチポートにおいて、物理リンクを介して、たとえばリモートサブネットに、接続されている場合、リモートサブネットの汎用スイッチポートを処理するＳＭＡは、リモート属性がサポートされていないことを示す状態値とともに要求に応答することができ、さらに、当該応答は要求しているＳＭに到達することができる。

【０１２２】

一実施形態に従うと、ＳＭＰ（すなわち、ＶＳＭＰ）は、ＤＲ経路を用いることによって、またはルータポートのＬＩＤを介して、ローカルルータポートに送信することができる。要求された情報がルータポートのリモート・ピア・ポートのためのものである場合、属性修飾子のリモートフラグ／ビットを設定することができる。ＳＭＡは、このようなＶＳＭＰを受取ると、パケットを変更し、リモートエンドをアドレス指定する付加的なＤＲステップを追加することができる。

【０１２３】

一実施形態に従うと、パケット属性の一部分（たとえば１６ビット属性修飾子のビット）はＶＳＭＰがローカルであるかリモートであるかどうかを信号で伝えるために用いることができる。たとえば、この部分を１の値に設定することにより、リモート・ピア・ポートがＶＳＭＰのための最終宛先であることを示唆することができる。属性のこの部分はリモートフラグとも称され得る。

【０１２４】

一実施形態に従うと、リモートフラグに加えて、どの宛先インスタンスが、最終宛先への経路に沿ってＶＳＭＰを処理するべきかを示すことができる追加のフラグが存在し得る。パケット属性のうち２つの付加的部分（たとえば属性修飾子の２ビット）は、この目的のために用いることができる。第１の受取りフラグと称される第１の部分（たとえばビット２０）は、（元の要求の宛先アドレスと一致するはずである）ローカルルータポートによって受取られたとき、パケットの処理が予想されることを示し得る。第１の受信側における予想される処理がすべて実行されると、パケット属性の第２の部分（たとえば属性修飾子のビット２１）を設定して、パケットをリモートエンドに転送することができる。この第２の部分は第２の受取りフラグと称され得る。

【０１２５】

ＤＲルーティングされたパケット

一実施形態に従うと、一例として、ＤＲルーティングされたパケットは例示的なフローに追従し得る。ソースノードは、ＬＩＤ　Ａにおいて、ルータ１を宛先ノードとして指定する要求パケットを開始することができる。例示的なＤＲルーティングパケット構成は以下のとおりである。

【０１２６】

・MADHdr.Class = 0x81（ＤＲルーティングされたＳＭＰ）
・MADHdr.Method = 0x1（取得）
・LRH.SLID = LRH.DLID = 0xffff（許可LID）
・MADHdr.DrSLID = MADHdr.DrDLID = 0xffff
・MADHdr.AttrID = <VSMP attrID>
・MADHdr.AttrMod.remote = 1
・MADHdr.AttrMod.first_receiver = 1
・MADHdr.InitPath = DR経路からRtr1（LID B）
・MADHdr.HopCnt = N
・MADHdr.HopPtr = 0

要求パケットがルータ１に到達すると、当該要求パケットは、対応するＳＭＡに渡され得るとともに、さらにこの対応するＳＭＡが、要求が有効であることを確認し得る。要求の有効性を確認した後、ＳＭＡはパケットを変更し得る。この変更は、１つの余分なホップを用いてＤＲ経路を拡張することと、第２の受取りフラグを設定することとを含み得る。このような例示的な構成変更を以下に示す。

【０１２７】

・MADHdr.HopCnt = N+1（１つの余分なホップを用いてＤＲ経路を拡張）
・MADHdr.InitPath[N+1] =（仮想ルータ外部ポート番号（すなわち、２））
・MADHdr.AttrMod.second_receiver = 1

一実施形態に従うと、ＳＭＡのサブネット管理インターフェイス（subnet management interface：ＳＭＩ）層は、ＳＭＡが物理リンク上でＶＳＭＰを転送する前に、余分なホップを用いてＤＲを更新することができる。

【０１２８】

一実施形態に従うと、物理リンクの遠隔側でＶＳＭＰを受取るＳＭＡのために、ＶＳＭＰによって用いられるアドレス指定方式は、正常にＩＢ規定されたパケットの方式のように見える場合もある。

【０１２９】

一実施形態に従うと、物理リンクの遠隔側でＶＳＭＰを受取るＳＭＡがＶＳＭＰの最終

宛先である場合、ＶＳＭＰは受取られたＳＭＡによって確認することができる。確認はフラグ設定に対して入力ポートをチェックすることを含み得る。リモートフラグならびに第１の受取りフラグおよび第２の受取りフラグが設定される場合、パケットは物理ポート上で（すなわち、仮想ルータで構成されたポートの外部リンク側から）受取ることができる。遠隔にある第１の受取りフラグだけが設定される場合、パケットは、仮想リンク上で（すなわち、仮想ルータポートの内部スイッチ側から）到達し得る。確認が失敗すれば、状態が適切なエラーメッセージに設定され得る。

【０１３０】

図１４は、一実施形態に従った、ＤＲルーティングされたＶＳＭＰパケットを示すフローチャートである。

【０１３１】

ステップ１４０１において、一実施形態に従うと、サブネットＡ内において、サブネットＡのサブネットマネージャなどのエンティティは、ＤＲルーティングされたＶＳＭＰを介して接続性情報を要求することができる。ＤＲルーティングされたＶＳＭＰの宛先は、サブネットＡ内におけるルータ１などのルータであり得る。

【０１３２】

ステップ１４０２において、ＤＲルーティングされたＶＳＭＰがルータ１に到達し得る。ルータ１は、一実施形態においては、上述のとおり、スイッチ１などのスイッチに含まれ得る。ＤＲルーティングされたＶＳＭＰは、確認のためにスイッチ１のＳＭＡに渡され得る。

【０１３３】

ステップ１４０３において、一実施形態に従うと、ＳＭＡはＤＲルーティングされたＶＳＭＰを変更し得る。この変更は、（第２のサブネットに対する物理リンクにわたるホップのための）ホップ・カウンタを拡張することと、第２の受取りフラグを設定することとを含み得る。

【０１３４】

ステップ１４０４において、一実施形態に従うと、ＳＭＩ（ＳＭＩはスイッチ／ルータのＳＭＡに関連付けられている）は、ＶＳＭＰのホップポインタを更新し得る。

【０１３５】

ステップ１４０５において、一実施形態に従うと、ＳＭＩは、物理リンク上で、サブネットＡ１４２０とサブネットＢ１４３０との間のサブネット境界にわたってＤＲ　ＶＳＭＰを転送することができる。この場合、物理リンクの第１の端部は、サブネットＡ内のルータに接続され得るとともに、物理リンクの第２の端部は、サブネットＢ内のルータに接続され得る。

【０１３６】

ステップ１４０６において、一実施形態に従うと、サブネットＢ１４３０内のルータ２のＳＭＡ′は物理リンクからＶＳＭＰを受取り得る。

【０１３７】

ステップ１４０７において、一実施形態に従うと、（ＳＭＡ′に関連付けられた）ＳＭＩ′がＶＳＭＰ要求を確認し得る。

【０１３８】

一実施形態に従うと、応答エンティティ（たとえば、物理リンクの遠隔側のルータ）は、ＳＭＰ応答を完了し、応答を示す方向属性を設定し得る。このような例示的な構成を以下に示す。

【０１３９】

・MADHdr.Method = 0x81 (GetResp)
・MADHdr.Direction = 1 （応答を示す）

一実施形態に従うと、ＳＭＩ層は、ホップポインタをデクリメントし、物理リンク上で、ローカル側に応答を転送し得る。ローカルルータにおけるＳＭＡは、要求に対してなされた変更を元に戻し得る。ローカル側ルータにおけるＳＭＩは、ホップをデクリメントす

ることと、仮想リンク上で（すなわち、スイッチポートと仮想ルータポートとの間の内部仮想リンク上で）スイッチに応答を送出することとを含む通常の処理を実行し得る。次のホップのために、ＳＭＩが再びホップをデクリメントし、要求が本来到達していた物理スイッチポート上でパケットを送出し得る。

【０１４０】

図１５は、一実施形態に従った、ＤＲルーティングされたＶＳＭＰパケットに対する応答を示すフローチャートである。

【０１４１】

ステップ１５０１において、一実施形態に従うと、ルータ２のＳＭＡ′は、ＶＳＭＰに応答情報（たとえば接続性情報）を記入し、応答を示すために方向性フラグを設定し得る。

【０１４２】

ステップ１５０２において、一実施形態に従うと、ＳＭＩ′は、応答のホップポインタをデクリメントし得るとともに、たとえば、２つの端部を有する物理リンク上でサブネットＢ１５３０からサブネットＡ１５２０に応答を転送し返し得る。物理リンクの第１の端部は、サブネットＡ内のルータに接続され得るとともに、物理リンクの第２の端部は、サブネットＢ内のルータに接続され得る。

【０１４３】

ステップ１５０３において、一実施形態に従うと、ルータ１におけるＳＭＡは応答を受取って、ＶＳＭＰ上でなされた変更を元に戻し得る。

【０１４４】

ステップ１５０４において、一実施形態に従うと、ルータ１におけるＳＭＩは、ＶＳＭＰに対する応答を処理し、物理スイッチに対してリンク（たとえば、仮想ルータ１と物理スイッチとの間の仮想リンク）上で応答を送信し得る。

【０１４５】

ステップ１５０５において、一実施形態に従うと、ＳＭＩは、ホップポイントカウンタをデクリメントし、ＶＳＭＰが元々受取られていた物理スイッチポート上でパケットを送出し得る。

【０１４６】

ＬＩＤルーティングされたパケット

一実施形態に従うと、一例として、ＬＩＤルーティングされたパケットは例示的なフローに追従し得る。ソースノードは、ＬＩＤ　Ａにおいて、宛先ノードとしてルータ１を指定する要求パケットを開始させ得る。このような例示的なパケットは以下の構成を有し得る。

【０１４７】

・MADHdr.Class = 0x01 （ＬＩＤルーティングされたＳＭＰ）
・MADHdr.Method = 0x1（取得）
・LRH.SLID = LID A
・LRH.DLID = LID B
・MADHdr.AttrID = <VSMP attrID>
・MADHdr.AttrMod.remote = 1
・ADHdr.AttrMod.first_receiver = 1

一実施形態に従うと、要求パケットがルータ１に到達すると、当該要求パケットは、対応するＳＭＡに渡され、当該対応するＳＭＡは、要求が有効であることを確認し得る。要求の有効性を確認した後、ＳＭＡはパケットを変更し得る。この変更は、端部における単一のＤＲ経路を追加することを含み得る。結果として、これは、アドレスの合計がＬＩＤルーティングされたパケットとＤＲルーティングされたホップとの組合せになり得ることを意味する。例示的な構成は以下のとおりである。

【０１４８】

・MADHdr.Class = 0x81（ＤＲルーティングされたＳＭＰ）

・MADHdr.HopCnt = 1
・MADHdr.HopPtr = 0
・MADHdr.Direction = 0（アウトバウンド）
・MADHdr.InitPath[1] = （仮想ルータ外部ポート番号（すなわち、２））
・MADHdr.DrSLID = LRH.SLID （すなわち、LRH.SLIDは元のリクエスタからのＬＩＤＡを含む）
・MADHdr.DrDLID = 0xffff
・MADHdr.AttrMod.second_receiver = 1

一実施形態に従うと、ルータにおけるＳＭＩ層は、アウトバウンドのパケットを正常に処理し、これを物理リンク上で送出し得る。例示的な構成は以下のとおりである。

【０１４９】

・LRH.SLID = LRH.DLID = 0xffff
・MADHdr.HopPtr = 1 （１ずつインクリメント）

一実施形態に従うと、宛先ルータにおける（物理リンクの他方の端部における）ＳＭＩは、それが要求についての宛先であると判断し、関連付けられたＳＭＡに対してＶＳＭＰを渡し得る。

【０１５０】

図１６は、一実施形態に従った、ＤＲルーティングされたＶＳＭＰパケットを示すフローチャートである。

【０１５１】

ステップ１６０１において、一実施形態に従うと、サブネットＡ内において、サブネットＡのサブネットマネージャなどのエンティティは、ＬＩＤルーティングされたＶＳＭＰを介して接続性情報を要求し得る。ＬＩＤルーティングされたＶＳＭＰの宛先は、ルータ１などの、サブネットＡ内のルータであり得る。

【０１５２】

ステップ１６０２において、ＬＩＤルーティングされたＶＳＭＰはルータ１に到達し得る。ルータ１は、一実施形態においては、上述のとおり、スイッチ１などのスイッチに含まれ得る。ＬＩＤルーティングされたＶＳＭＰは、確認のためにスイッチ１のＳＭＡに渡され得る。

【０１５３】

ステップ１６０３において、一実施形態に従うと、ＳＭＡはＤＲルーティングされたＶＳＭＰを変更し得る。この変更は、単一のホップＤＲルーティングされた経路をアドレスの最後に追加することと、（第２のサブネットに対する物理リンクにわたるホップのために）ホップ・カウンタを拡張することと、第２の受取りフラグを設定することとを含み得る。

【０１５４】

ステップ１６０４において、一実施形態に従うと、ＳＭＩ（ＳＭＩはスイッチ／ルータのＳＭＡに関連付けられている）は、ＶＳＭＰのホップポインタを更新し得る。

【０１５５】

ステップ１６０５において、一実施形態に従うと、ＳＭＩは、物理リンク上で、サブネットＡ１６２０とサブネットＢ１６３０との間におけるサブネット境界にわたって、元々ＬＩＤにルーティングされていたＶＳＭＰを転送し得る。この場合、物理リンクの第１の端部はサブネットＡ内のルータに接続され得るとともに、物理リンクの第２の端部は、サブネットＢ内のルータに接続され得る。

【０１５６】

ステップ１６０６において、一実施形態に従うと、サブネットＢ１６３０内のルータ２のＳＭＡ′は、このとき、ＤＲルーティングされたＶＳＭＰを物理リンクから受取り得る。

【０１５７】

ステップ１６０７において、一実施形態に従うと、（ＳＭＡ′に関連付けられた）ＳＭ

Ｉ′はＶＳＭＰ要求を確認し得る。

【０１５８】

一実施形態に従うと、応答フローのために、ルータ２におけるＳＭＡがＶＳＭＰを確認し得る。この確認は、フラグ設定に対して入力ポートをチェックすることを含み得る。リモートフラグならびに第１の受取りフラグおよび第２の受取りフラグが設定される場合、パケットは物理ポート上で（すなわち、仮想ルータで構成されたポートの外部リンク側から）受取ることができる。遠隔側である第１の受取りフラグだけが設定される場合、パケットは、仮想リンク上で（すなわち、仮想ルータポートの内部スイッチ側から）到達し得る。確認が失敗すれば、状態が適切なエラーメッセージに設定され得る。

【０１５９】

一実施形態に従うと、ルータ２におけるＳＭＡ′は、ＳＭＰ応答を完了し、応答を示す方向属性を設定し得る。このような例示的な構成を以下に示す。

【０１６０】

・MADHdr.Method = 0x81 (GetResp)
・MADHdr.Direction = 1（応答を示す）
・LRH.SLID = 0xffff
・LRH.DLID = ReqMAD.MADHdr.SLID = 0xffff

ルータ２におけるＳＭＩ′は、ホップポインタをデクリメントし、物理リンク上で応答をルータ１に転送し返し得る。次いで、ルータ１におけるＳＭＡは、応答を物理スイッチポートから送信元に転送し返す前に、元の要求上でなされた変更を元に戻し得る。このように元に戻した後の例示的な構成を以下に示す。

【０１６１】

・MADHdr.Class = 0x01
・LRH.DLID = MADHdr.DrSLID （すなわち、元のリクエスタからのＬＩＤ　Ａを含む）
・LRH.SLID = local LID = B

一実施形態に従うと、加えて、ＳＭＡがＶＳＭＰのＤＲ特有のフィールドをクリアし得るため、応答は、当該応答が元のリクエスタに到達したときに元のＶＳＭＰと完全に一致しているように見えることとなる。

【０１６２】

一実施形態に従うと、応答が送信元に戻ってくると、送信元は、正常なルータＬＩＤ応答を、それがルータ１において完全に処理されてしまっているかのように認識するだろう。

【０１６３】

図１７は、一実施形態に従った、ＬＩＤルーティングされたＶＳＭＰパケットに対する応答を示すフローチャートである。

【０１６４】

ステップ１７０１において、一実施形態に従うと、ルータ２のＳＭＡ′は、ＶＳＭＰに対する応答情報（たとえば接続性情報）を記入し得るとともに、応答を示すために方向性フラグを設定し得る。

【０１６５】

ステップ１７０２において、一実施形態に従うと、ＳＭＩ′は、応答のホップポインタをデクリメントし、たとえば、２つの端部を有する物理リンク上で、サブネットＢ１７３０からサブネットＡ１７２０に応答を転送し返し得る。物理リンクの第１の端部は、サブネットＡ内のルータに接続され得るとともに、物理リンクの第２の端部は、サブネットＢ内のルータに接続され得る。

【０１６６】

ステップ１７０３において、一実施形態に従うと、ルータ１におけるＳＭＡは、応答を受取り得るとともに、ＤＲルーティングされたホップを取り除くことを含む、ＶＳＭＰ上でなされた変更を元に戻し得る。

【０１６７】
ステップ１７０４において、一実施形態に従うと、ルータ１におけるＳＭＩは、ＶＳＭＰに対する応答を処理し得るとともに、ＶＳＭＰが元々受取られていた物理スイッチポート上でＬＩＤルーティングされた応答を送出し得る。
【０１６８】
この発明の特徴は、ここに提示された特徴のうちのいずれかを行なうように処理システムをプログラミングするために使用可能な命令を格納した記憶媒体またはコンピュータ読取り可能媒体であるコンピュータプログラムプロダクトにおいて、それを使用して、またはその助けを借りて実現され得る。記憶媒体は、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、フラッシュメモリ装置、磁気カードもしくは光カード、ナノシステム（分子メモリＩＣを含む）、または、命令および／もしく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ものの、それらに限定されない。
【０１６９】
この発明の特徴は、機械読取り可能媒体のうちのいずれかに格納された状態で、処理システムのハードウェアを制御するために、および処理システムがこの発明の結果を利用する他の機構とやり取りすること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取込まれ得る。そ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、装置ドライバ、オペレーティングシステム、および実行環境／コンテナを含み得るものの、それらに限定されない。
【０１７０】
この発明の特徴はまた、たとえば、特定用途向け集積回路（application specific integrated circuit：ＡＳＩＣ）などのハードウェアコンポーネントを使用して、ハードウェアにおいて実現されてもよい。ここに説明された機能を行なうようにハードウェアステートマシンを実現することは、関連技術の当業者には明らかであろう。
【０１７１】
加えて、この発明は、この開示の教示に従ってプログラミングされた１つ以上のプロセッサ、メモリおよび／またはコンピュータ読取り可能記憶媒体を含む、１つ以上の従来の汎用または特殊デジタルコンピュータ、コンピューティング装置、マシン、またはマイクロプロセッサを使用して都合よく実現され得る。ソフトウェア技術の当業者には明らかであるように、この開示の教示に基づいて、適切なソフトウェアコーディングが、熟練したプログラマによって容易に準備され得る。
【０１７２】
この発明のさまざまな実施形態が上述されてきたが、それらは限定のためではなく例示のために提示されたことが理解されるべきである。この発明の精神および範囲から逸脱することなく、形状および詳細のさまざまな変更を行なうことができることは、関連技術の当業者には明らかであろう。
【０１７３】
この発明は、特定された機能およびそれらの関係の実行を示す機能的構築ブロックの助けを借りて上述されてきた。説明の便宜上、これらの機能的構築ブロックの境界は、この明細書中ではしばしば任意に規定されてきた。特定された機能およびそれらの関係が適切に実行される限り、代替的な境界を規定することができる。このため、そのようないかなる代替的な境界も、この発明の範囲および精神に含まれる。
【０１７４】
この発明の前述の説明は、例示および説明のために提供されてきた。それは、網羅的であるよう、またはこの発明を開示された形態そのものに限定するよう意図されてはいない。この発明の幅および範囲は、上述の例示的な実施形態のいずれによっても限定されるべきでない。多くの変更および変形が、当業者には明らかになるだろう。これらの変更および変形は、開示された特徴の関連するあらゆる組合せを含む。実施形態は、この発明の原

理およびその実用的応用を最良に説明するために選択され説明されたものであり、それにより、考えられる特定の使用に適したさまざまな実施形態についての、およびさまざまな変更例を有するこの発明を、当業者が理解できるようにする。この発明の範囲は、請求項およびそれらの同等例によって定義されるよう意図されている。

【図１】

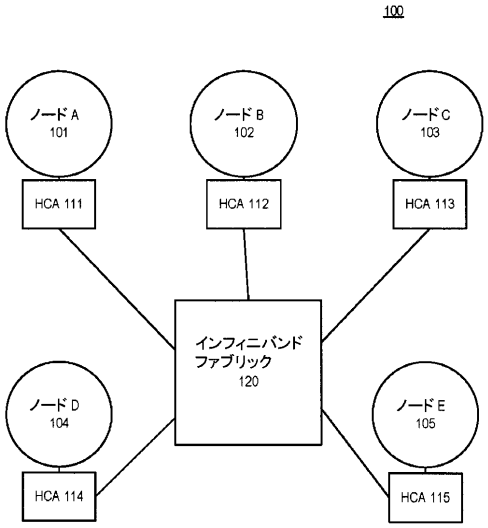

FIGURE 1

【図２】

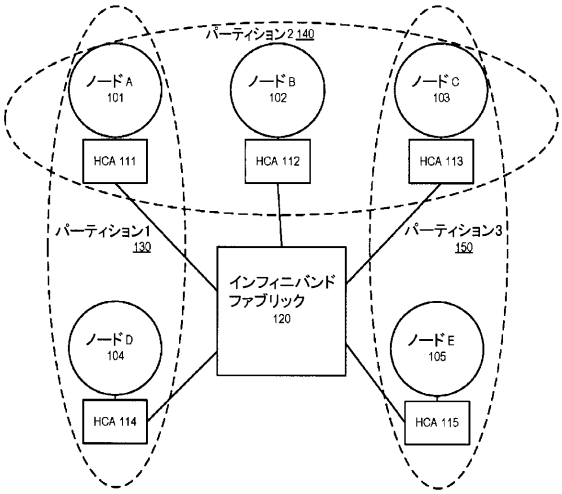

FIGURE 2

【図３】

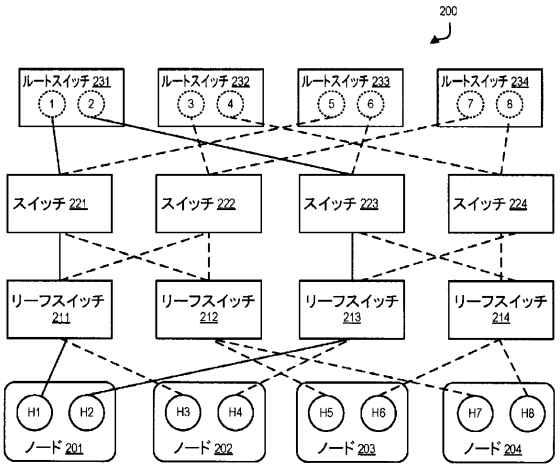

FIGURE 3

【図４】

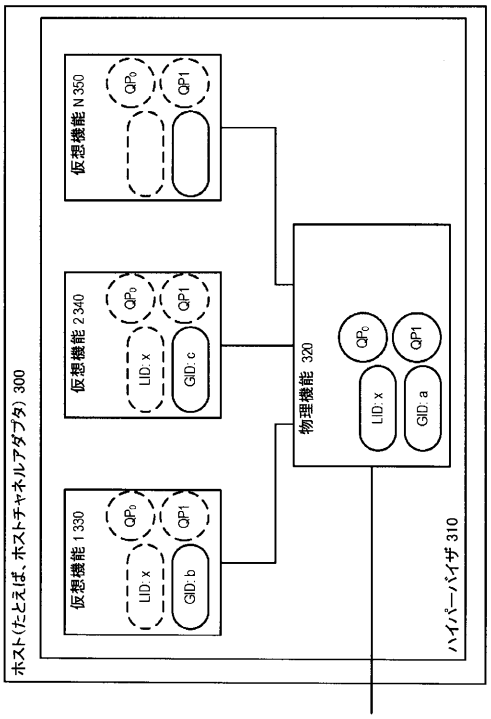

FIGURE 4

【図５】

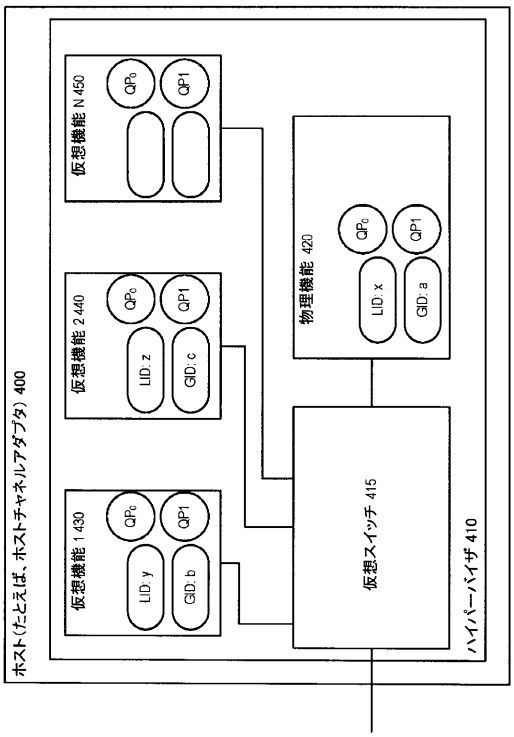

FIGURE 5

【図６】

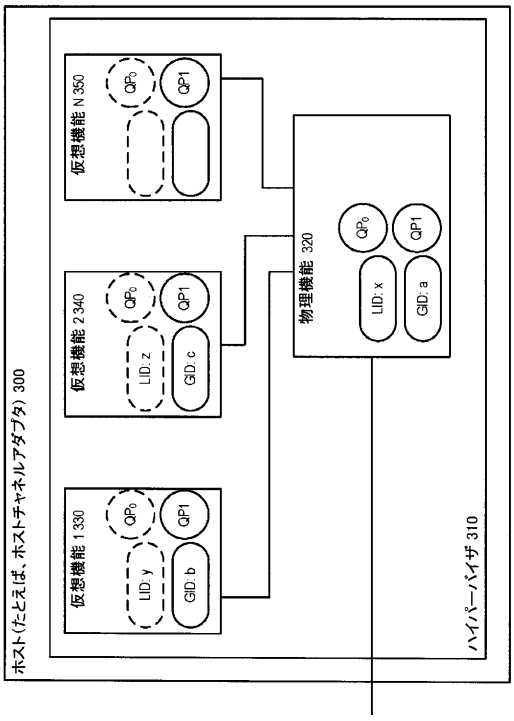

FIGURE 6

【図７】

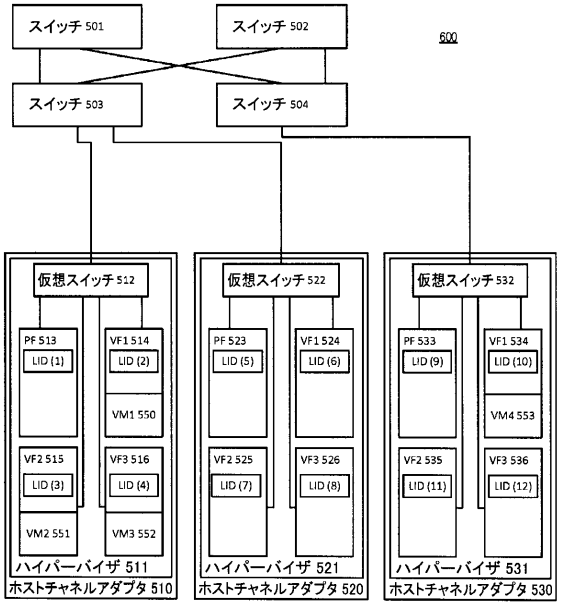

FIGURE 7

【図８】

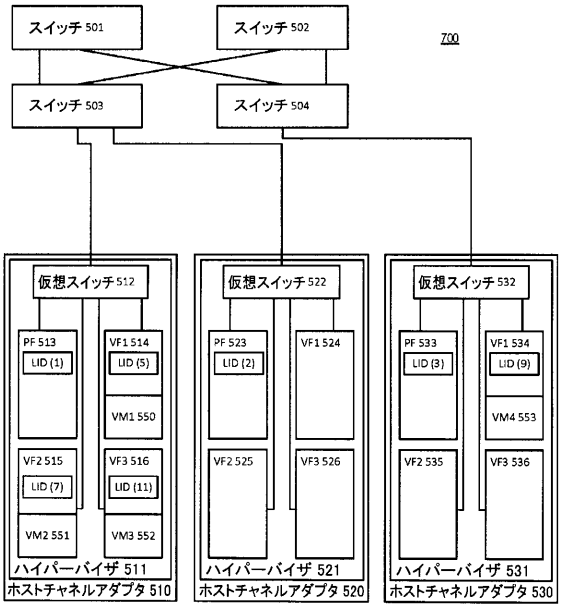

FIGURE 8

【図９】

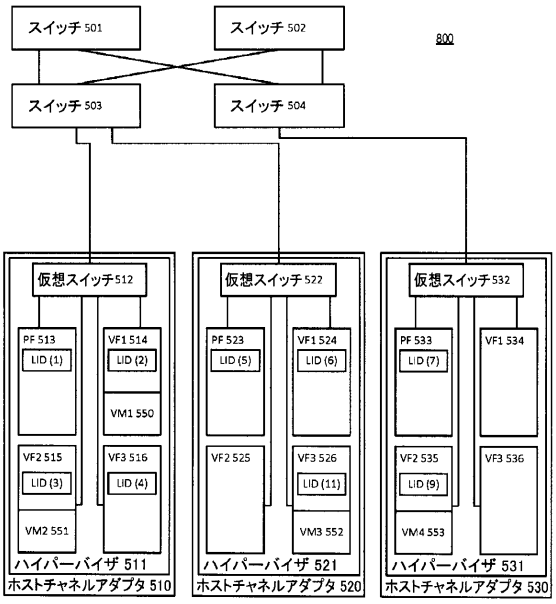

FIGURE 9

【図１０】

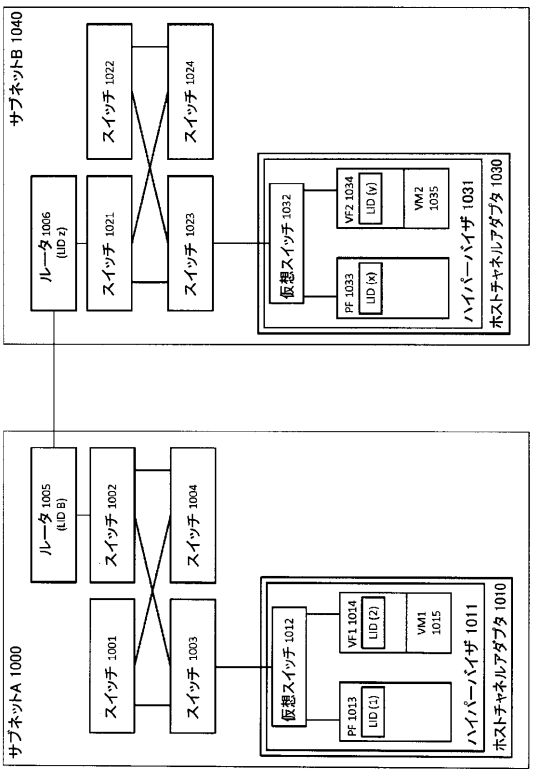

FIGURE 10

【図１１】

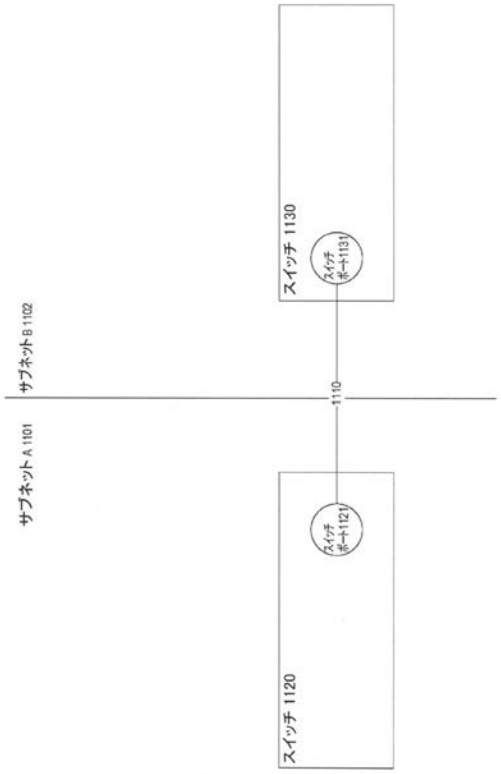

FIGURE 11

【図１２】

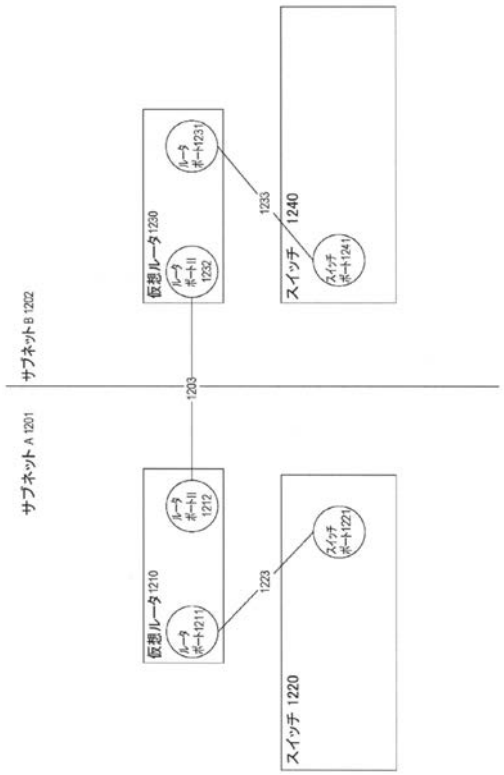

FIGURE 12

【図１３】

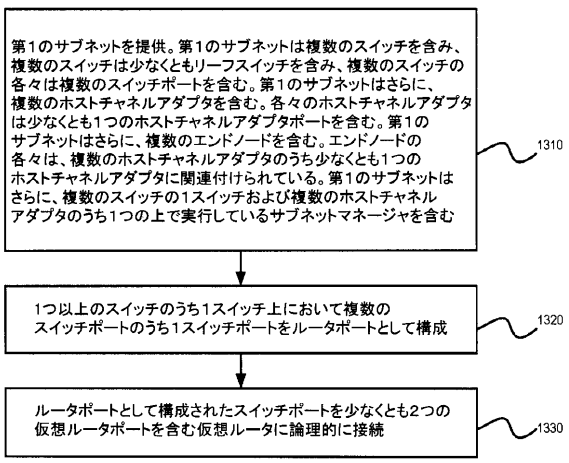

FIGURE 13

【図１４】

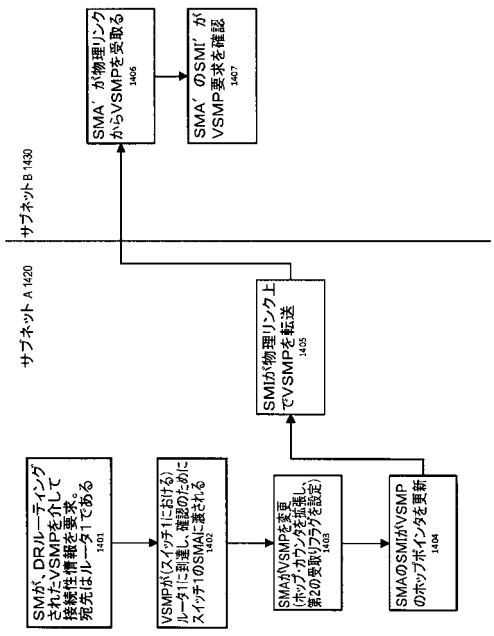

FIGURE 14

【図１５】

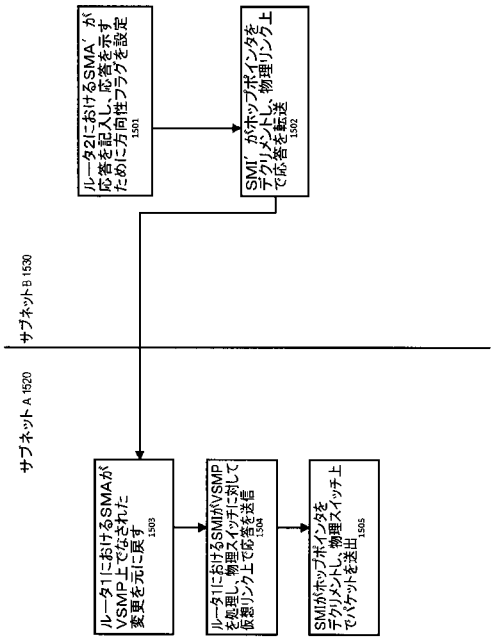

FIGURE 15

【図１６】

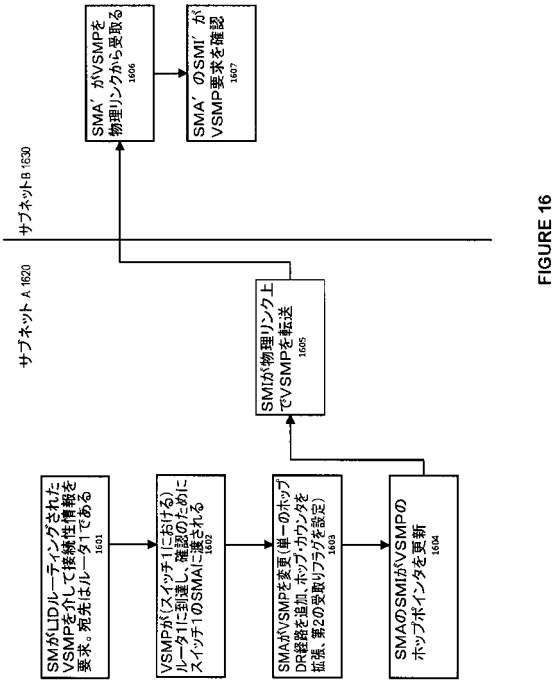

FIGURE 16

【図１７】

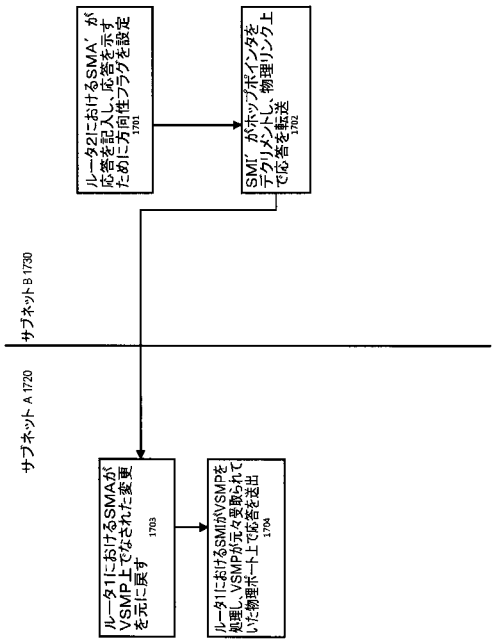

FIGURE 17

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2017/014959

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L12/931 H04L12/733 H04L12/713 H04L12/751 H04L12/721
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 9 219 718 B2 (JOHNSEN BJÖRN DAG [NO] ET AL) 22 December 2015 (2015-12-22) figures 1-3 column 1, line 39 - line 43 column 1, line 61 - column 2, line 4 column 2, line 23 - line 67 column 3, line 27 - line 43 column 5, line 4 - line 8 column 5, line 24 - line 48 column 6, line 23 - line 63	1-22
A	US 2009/213753 A1 (BURROW STEPHEN R [US] ET AL) 27 August 2009 (2009-08-27) paragraphs [0029] - [0032], [0052] - [0057]	1-22
A	US 2014/269686 A1 (SRINIVASAN ARVIND [US] ET AL) 18 September 2014 (2014-09-18) the whole document	1-22

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 April 2017	04/05/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Tyszka, Krzysztof

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2017/014959

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 9219718 B2	22-12-2015	CN 103597795 A	19-02-2014
		EP 2716003 A1	09-04-2014
		JP 5965478 B2	03-08-2016
		JP 2014517406 A	17-07-2014
		US 2012307682 A1	06-12-2012
		US 2012311122 A1	06-12-2012
		US 2012311123 A1	06-12-2012
		US 2012311143 A1	06-12-2012
		US 2012311182 A1	06-12-2012
		US 2012311332 A1	06-12-2012
		US 2012311333 A1	06-12-2012
		US 2012311670 A1	06-12-2012
		US 2014241208 A1	28-08-2014
		WO 2012167268 A1	06-12-2012
US 2009213753 A1	27-08-2009	NONE	
US 2014269686 A1	18-09-2014	NONE	

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ

(72)発明者 ボグダンスキー, バルトシュ
ノルウェー、０２７５ オスロ、ホフ・テラッセ、１５、エイチ・０２０３

Ｆターム(参考) 5K030 GA11 HB11 HD03 KX17 LB05 MA14
5K033 AA05 CB01 CB08 DB03