

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6957451号

(P6957451)

(45) 発行日 令和3年11月2日 (2021.11.2)

(24) 登録日 令和3年10月8日 (2021.10.8)

(51) Int. Cl.

F I

H O 4 L 12/931 (2013.01)

H O 4 L 12/931

H O 4 L 12/741 (2013.01)

H O 4 L 12/741

H O 4 L 12/70 (2013.01)

H O 4 L 12/70

D

G O 6 F 13/14 (2006.01)

G O 6 F 13/14

3 2 O B

G O 6 F 15/173 (2006.01)

G O 6 F 15/173

6 8 5 M

請求項の数 12 (全 61 頁)

(21) 出願番号 特願2018-510374 (P2018-510374)
 (86) (22) 出願日 平成29年1月27日 (2017.1.27)
 (65) 公表番号 特表2019-505103 (P2019-505103A)
 (43) 公表日 平成31年2月21日 (2019.2.21)
 (86) 国際出願番号 PCT/US2017/015464
 (87) 国際公開番号 W02017/132585
 (87) 国際公開日 平成29年8月3日 (2017.8.3)
 審査請求日 令和2年1月22日 (2020.1.22)
 (31) 優先権主張番号 62/288,101
 (32) 優先日 平成28年1月28日 (2016.1.28)
 (33) 優先権主張国・地域又は機関
 米国 (US)
 (31) 優先権主張番号 15/416,642
 (32) 優先日 平成29年1月26日 (2017.1.26)
 (33) 優先権主張国・地域又は機関
 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9406
 5レッドウッド・シティ、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、0687 オスロ、ビルベル
 クグレンダ、9
 (72) 発明者 スリニバサン、アルビンド
 アメリカ合衆国、95129 カリフォル
 ニア州、サン・ノゼ、ハッピー・バレー・
 アベニュー、1075

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での線形転送テーブル (L F T) 探索のためにグローバルルート
 ヘッダ (G R H) におけるサブネットプレフィックス値を用い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】

高性能コンピューティング環境でのネットワークスイッチ環境におけるパケット転送の
 ためにパケットヘッダを使用するための方法であって、

ネットワークファブリックを含むコンピュータ環境において、第1および第2のサブネ
 ットを提供することを備え、前記第1のサブネットは前記第2のサブネットとは異なり、
 前記第1のサブネットは、第1のセットのネットワークスイッチを含み、前記第1のセッ
 トのネットワークスイッチの各々は、線形転送テーブル (L F T) および複数の物理ポー
 トを含み、前記第2のサブネットは、第2のセットのネットワークスイッチを含み、前記
 第2のセットのネットワークスイッチの各々は、L F T および複数の物理ポートを含み、
 前記方法はさらに、

前記第1のセットのネットワークスイッチのうちの第1のネットワークスイッチが、前
 記第1のネットワークスイッチから前記第1のサブネットの前記第1のセットのネットワ
 ークスイッチのうちの1以上の他のネットワークスイッチへのパケットのサブネット内転
 送、ならびに前記第1のネットワークスイッチから前記第2のサブネットの前記第2のセ
 ットのネットワークスイッチの1以上のネットワークスイッチへのパケットのサブネット
 間転送の両方に対して用いられることを許可することと、

第1のパケットを、前記第1のサブネットの前記第1のセットのネットワークスイッチ
 のうちの前記第1のネットワークスイッチで受信することと、を含み、前記第1のパケッ
 トは、前記第1のサブネット内のアドレスを識別するローカルルートヘッダ (L R H) 部

分と、前記第2のサブネット内のアドレスを識別するグローバルルートヘッダ（GRH）とを含み、前記方法は、さらに、

前記第1のパケットの前記LRHの前記第1の部分を検査することと、

前記第1のパケットの前記LRHの前記第1の部分の第1および第2の条件を判断することとを含み、

前記第1および第2の条件を判断することは、前記第1のネットワークスイッチに格納されている所定値と前記LRHの前記第1の部分とを比較し、前記所定値と前記LRHの前記第1の部分との間の不一致に従って前記第1の条件を選択的に判断し、前記所定値と前記LRHの前記第1の部分との間の一致に従って、前記第1のネットワークスイッチの第1のLFTを索引付けして、前記第1のサブネットの境界を表わす特殊スイッチポート（SSP）を取得することによって、前記第2の条件を選択的に判断することを含み、前記方法は、さらに、

前記第1のパケットの前記LRHの前記第1の部分の前記第1の条件に選択的に基づいて、前記第1のネットワークスイッチのサブネット内ポートを決定するために前記第1のネットワークスイッチの前記第1のLFTを索引付け、前記第1のネットワークスイッチによって、前記第1のネットワークスイッチの決定されたサブネット内ポートと関連する前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの1つのネットワークスイッチへ前記第1のパケットをルーティングすることと、

前記第1のパケットの前記LRHの前記第1の部分の前記第2の条件に選択的に基づいて、前記LRHの前記第1の部分を前記GRHの選択された部分によって置換することによって、新たなLRHを形成し、前記第1のネットワークスイッチによって、前記GRHおよび前記形成された新たなLRHと関連する前記第2のサブネットの前記第2のセットのネットワークスイッチのうちの1つのネットワークスイッチへ前記第1のパケットをルーティングすることと、を含み、

前記ルーティングは、

前記GRHのサブネット間ルート番号（ISRN）を選択することによって行われ、前記ルーティングはさらに、

前記GRHの前記ISRNを使用して、前記第1のネットワークスイッチの前記第1のLFTを索引付けすることと、

前記GRHの前記ISRNを使用して前記第1のLFTを索引付けすることに従って、前記第1のパケットを前記第2のサブネットの前記第2のセットのネットワークスイッチのうちの前記選択されたネットワークスイッチにルーティングすることとによって行われ、前記第2のサブネットは前記ネットワークスイッチ環境の中間コアファブリックである、パケット転送のためにパケットヘッダを使用するための方法。

【請求項2】

前記第1のパケットの前記LRHの前記第1および第2の部分の条件を判断することは、前記LRHの前記第1の部分を前記第1のネットワークスイッチに格納される所定の値と比較することと、

前記LRHの前記第1の部分と前記所定の値との間の不一致に従って前記第1の条件を選択的に判断することと、

前記LRHの前記第1の部分と前記所定の値との間の一致に従って前記第2の条件を選択的に判断することとを含む、請求項1に記載の方法。

【請求項3】

前記LRHの前記第1の部分と前記所定の値との間の一致に従って前記第2の条件を選択的に判断することは、前記第1のLFTを索引付けして、前記第1のサブネットの境界を表わす特殊スイッチポート（SSP）を取得することを含む、請求項2に記載の方法。

【請求項4】

前記GRHおよび前記形成された新たなLRHと関連する前記第2のサブネットの前記第2のセットのネットワークスイッチのうちの1つのネットワークスイッチへ前記第1のパケットをルーティングすることは、

前記GRHのサブネット間ルート番号（ISRN）を選択することと、

前記GRHの前記ISRNを使用して、前記第1のネットワークスイッチの前記第1のLFTを索引付けすることと、

前記GRHの前記ISRNを使用して前記第1のLFTを索引付けすることに従って、前記第1のパケットを前記第2のサブネットの前記第2のセットのネットワークスイッチのうちの前記選択されたネットワークスイッチにルーティングすることとを含み、前記第2のサブネットは前記ネットワークスイッチ環境の中間コアファブリックである、請求項1～3のいずれか1項に記載の方法。

【請求項5】

前記ネットワークスイッチ環境の前記中間コアファブリックから、第2のパケットを、前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの前記第1のネットワークスイッチで受信することと、

中間コアファブリックから受信される前記第2のパケットが前記ISRNを使用して転送されていると判断することと、

前記ISRNを使用して、および前記第2のパケットのヘッダに含まれるターゲットローカル識別子（LID）を使用して、前記第2のパケットを転送することとをさらに備える、請求項4に記載の方法。

【請求項6】

前記第1のパケットのターゲットローカル識別子（LID）をフィルタリングすることによって前記第1のパケットの前記LRHの前記第1の部分の前記第1および第2の条件を判断して、前記第1のパケットが特殊スイッチポート（SSP）をターゲットとしているかどうかを判断することをさらに備える、請求項1または2に記載の方法。

【請求項7】

前記第1のパケットの前記LRHの前記第1の部分の前記第1の条件を判断することは、第1のパケットが前記SSPをターゲットとしていないと判断することを含み、

前記第1のパケットの前記LRHの前記第1の部分の前記第2の条件を判断することは、前記第1のパケットが前記SSPをターゲットとしていると判断することを含む、請求項6に記載の方法。

【請求項8】

高性能コンピューティング環境でのネットワークスイッチ環境におけるパケット転送のためにパケットヘッダを使用するためのシステムであって、

コンピュータ環境を備え、前記コンピュータ環境は、

ネットワークファブリックと、

第1のサブネットと、

第2のサブネットとを含み、前記第1のサブネットは前記第2のサブネットとは異なり、前記第1のサブネットは、第1のセットのネットワークスイッチを含み、前記第1のセットのネットワークスイッチの各々は、線形転送テーブル（LFT）および複数の物理ポートを含み、前記第2のサブネットは、第2のセットのネットワークスイッチを含み、前記第2のセットのネットワークスイッチの各々は、LFTおよび複数の物理ポートを含み、前記システムはさらに、

前記第1のセットのネットワークスイッチのうちの第1のネットワークスイッチの第1のLFTを備え、

前記第1のセットのネットワークスイッチのうちの第1のネットワークスイッチが、前記コンピュータ環境によって、前記第1のネットワークスイッチから前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの1以上の他のネットワークスイッチへのパケットのサブネット内転送、ならびに前記第1のネットワークスイッチから前記第2のサブネットの前記第2のセットのネットワークスイッチの1以上のネットワークスイッチへのパケットのサブネット間転送の両方に対して用いられ、前記コンピュータ環境は、

第1のパケットを、前記第1のサブネットの前記第1のセットのネットワークスイッチ

10

20

30

40

50

のうちの前記第1のネットワークスイッチで受信することによって行われ、前記第1のパケットは、前記第1のサブネット内のアドレスを識別するローカルルートヘッダ（LRH）部分と、前記第2のサブネット内のアドレスを識別するグローバルルートヘッダ（GRH）とを含み、前記コンピュータ環境は、

前記第1のパケットの前記LRHの第1の部分を検査することと、

前記第1のパケットの前記LRHの前記第1の部分の第1および第2の条件を判断することとを含み、

前記第1および第2の条件を判断することは、前記第1のネットワークスイッチに格納されている所定値と前記LRHの前記第1の部分とを比較し、前記所定値と前記LRHの前記第1の部分との間の不一致に従って前記第1の条件を選択的に判断し、前記所定値と前記LRHの前記第1の部分との間の一致に従って、前記第1のネットワークスイッチの第1のLFTを索引付けして、前記第1のサブネットの境界を表わす特殊スイッチポート（SSP）を取得することによって、前記第2の条件を選択的に判断することを含み、前記コンピュータ環境は、さらに、

前記第1のパケットの前記LRHの前記第1の部分の前記第1の条件に選択的に基づいて、前記第1のネットワークスイッチのサブネット内ポートを決定するために前記第1のネットワークスイッチの前記第1のLFTを索引付け、前記第1のネットワークスイッチによって、前記第1のネットワークスイッチの決定されたサブネット内ポートと関連する前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの1つのネットワークスイッチへ前記第1のパケットをルーティングすることと、

前記第1のパケットの前記LRHの前記第1の部分の前記第2の条件に選択的に基づいて、前記LRHの前記第1の部分を前記GRHの選択された部分によって置換することによって、新たなLRHを形成し、前記第1のネットワークスイッチによって、前記GRHおよび前記形成された新たなLRHと関連する前記第2のサブネットの前記第2のセットのネットワークスイッチのうちの1つのネットワークスイッチへ前記第1のパケットをルーティングすることと、によって行なわれ、

前記ルーティングは、

前記GRHのサブネット間ルート番号（ISRN）を選択することによって行われ、前記ルーティングはさらに、

前記GRHの前記ISRNを使用して、前記第1のネットワークスイッチの前記第1のLFTを索引付けすることと、

前記GRHの前記ISRNを使用して前記第1のLFTを索引付けすることに従って、前記第1のパケットを前記第2のサブネットの前記第2のセットのネットワークスイッチのうちの前記選択されたネットワークスイッチにルーティングすることとによって行われ、前記第2のサブネットは前記ネットワークスイッチ環境の中間コアファブリックである、高性能コンピューティング環境でのネットワークスイッチ環境におけるパケット転送のためにパケットヘッダを使用するためのシステム。

【請求項9】

前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの前記第1のネットワークスイッチは、

前記ネットワークスイッチ環境の前記中間コアファブリックから、第2のパケットを、前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの前記第1のネットワークスイッチで受信し、

中間コアファブリックから受信される前記第2のパケットが前記ISRNを使用して転送されていると判断し、

前記ISRNを使用して、および前記第2のパケットのヘッダに含まれるターゲットローカル識別子（LID）を使用して、前記第2のパケットを転送するよう動作可能である、請求項8に記載のシステム。

【請求項10】

前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの前記第1の

10

20

30

40

50

ネットワークスイッチは、

前記第1のパケットのターゲットローカル識別子（L I D）をフィルタリングすることによって前記第1のパケットの前記L R Hの前記第1の部分の前記第1および第2の条件を判断して、前記第1のパケットが特殊スイッチポート（S S P）をターゲットとしているかどうかを判断するよう動作可能である、請求項8記載のシステム。

【請求項11】

前記第1のサブネットの前記第1のセットのネットワークスイッチのうちの前記第1のネットワークスイッチは、

前記第1のパケットが前記S S Pをターゲットとしていないと判断することによって、前記第1のパケットの前記L R Hの前記第1の部分の前記第1の条件を判断し、

10

前記第1のパケットが前記S S Pをターゲットとしていると判断することによって、前記第1のパケットの前記L R Hの前記第1の部分の前記第2の条件を判断するよう動作可能である、請求項10に記載のシステム。

【請求項12】

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1～7のいずれか1項に記載の方法を実行させる、機械可読フォーマットにおけるプログラム命令を備える、コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

20

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。著作権所有者は、この特許文献または特許開示の、それが特許商標庁の特許ファイルまたは記録に現れているとおりの、何人による複製複製にも異議を唱えないが、それ以外の場合にはすべての著作権をどのようなものであろうと所有する。

【0002】

発明の分野：

本発明は、一般にコンピュータシステムに関し、特に、ネットワーク環境においてネットワークスイッチ機能を提供することに関する。

【背景技術】

30

【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：I B）技術などの高性能な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。

【0004】

単一のI Bサブネットにおける48Kユニキャストローカル識別（L I D）値空間は、エンドノードの数に関してサブネットのサイズの制限を表わす。仮想化されたホストチャネルアダプタ（H C A）が複数の仮想H C Aインスタンスを提供し、各そのような仮想H C Aインスタンスが仮想ポートごとに独立したL I Dとともに構成されてもよい場合、この制限は特に重要である。

40

【0005】

I B標準規格では、複数のサブネットを同じI Bファブリック内で接続して、パケット転送を、各中間サブネットごとおよび最終ターゲットサブネットについて、ファブリックワイドの128ビットの宛先グローバル識別子（D G I D）アドレス値から16ビットの宛先L I D（D L I D）アドレス値へのマッピングに基づいてサポートできるようにするために、ルータノードを定義する。

【0006】

50

しかしながら、I Bワイヤ速度で128ビットのD G I D値を16ビットのD L I D値にマッピングするには、一意に探索できる個々のD G I Dの数に関して（つまりハードウェア実装のための妥当なコスト／複雑さの制約内で）限られたスケーラビリティを有する複雑なコンテンツアドレス指定可能メモリ（C A M）ベースの探索ハードウェアが必要である。これは、多数の個々の宛先にスケーリングするために、128ビットのD G I D値を16ビットのD L I D値にマッピングすることは柔軟でなければならず、多数の個々のD G I Dアドレスを単一の探索エントリを介してマッピングすることができるように、階層的スキームを使用できなければならないことを意味する。

【0007】

しかしながら、階層的なマッピング構造は、表現できるサブネットとエンドノードの総数に関してスケーラビリティを表わすが、それは、また、マルチサブネットファブリック全体の異なるフローおよびワークロードに対して負荷均衡およびQ O S制約を維持するために個々の宛先を個別にルーティングする能力の厳しい制限も表わす。

【0008】

これは、本発明の実施形態が対処しようとす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】

【0009】

概要：

ここに開示されるのは、ミドルウェアおよびアプリケーション実行またはミドルウェアマシン環境のための設計されたシステムのような、ネットワーク環境においてサブネット内およびサブネット間アドレス解決をサポートできるシステムおよび方法であり、高性能コンピューティング環境での線形転送テーブル（L F T）探索のためにグローバルルートヘッダ（G R H）におけるサブネットプレフィックス値を用いる。例示のシステムおよび方法は、インフィニバンド（I B）アーキテクチャを有するネットワークのような高性能コンピューティング環境におけるネットワークスイッチ環境におけるL F T探索のためにG R Hにおけるサブネットプレフィックス値の使用を提供することができる。実施形態は、ネットワークファブリックを含むコンピュータ環境において、1つ以上のネットワークスイッチまたはホストに各々が関連付けられる1つ以上のサブネットを提供することができる。本明細書のシステムおよび方法は、サブネット内転送判断およびサブネット間転送判断の両方に同じ線形転送テーブルを使用できるようにすることができる。

【図面の簡単な説明】

【0010】

【図1】一実施形態によるインフィニバンド環境の一例を示す図である。

【図2】一実施形態による、パーティショニングされたクラスタ環境を示す図である。

【図3】一実施形態による、ネットワーク環境におけるツリートポロジを示す図である。

【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v S w i t c h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v P o r t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、L I Dが予めポピュレートされた例示的なv S w i t c h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L I D割当てがなされた例示的なv S w i t c h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L I D割当てがなされかつL I Dが予めポピュレートされているv S w i t c hを備えた例示的なv S w i t c hアーキテクチャを示す図である。

【図10】一実施形態による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11】一実施形態による、ネットワーク環境においてパケット転送ロジックにアクセスするためにインフィニバンド（I B）アドレス指定を使用するデータパケットフォーマ

10

20

30

40

50

ットの図を示す。

【図12】一実施形態による、サブネット内およびサブネット間転送のための線形転送テーブル（LFT）の例示的な部分を示す。

【図13】一実施形態による転送ドメインの図を示す。

【図14】一実施形態による、GRH／ISRNアドレス指定モードフォーマットからLRH／DLIDベースの転送フォーマットへのパケットヘッダの修正を示す図である。

【図15】一実施形態による、LRH／DLIDアドレス指定モードからGRH／ISRNアドレス指定モードにパケット転送を変更するための特殊スイッチポート境界を定義するメカニズムを提供するスイッチの図である。

【図16】一実施形態による、サブネット内転送およびサブネット間転送の両方について線形転送テーブル探索のためにパケットヘッダを使用する方法のフローチャートである。

【発明を実施するための形態】

【0011】

詳細な説明：

例示の実施形態は、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。

【0012】

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。

【0013】

高性能コンピューティング環境における線形転送テーブル（LFT）探索のためにグローバルルートヘッダ（GRH）におけるサブネットプレフィックス値を使用して仮想マシン（VM）のライブマイグレーションを可能にする、サブネット内およびサブネット間転送判断をサポートするシステムおよび方法がここに記載される。

【0014】

例示の実施形態の以下の説明は、高性能ネットワークのための例としてインフィニバンドTM（IB）ネットワークを使用する。以下の説明を通して、インフィニバンドTM規格（様々な、インフィニバンド規格、IB規格、またはレガシーIB規格とも呼ばれる）を参照することができる。このような参照は、引用によりその全体が本明細書に援用される、<http://www.infinibandta.org>で入手可能なインフィニバンド（登録商標）トレード・アソシエーション・アーキテクチャ規格、第1巻、バージョン1.3（2015年3月リリース）を参照すると理解される。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。

【0015】

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

10

20

30

40

50

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行うことが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。

【0016】

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。

10

【0017】

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。

【0018】

インフィニバンド (登録商標)

20

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。

【0019】

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。

30

【0020】

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。

40

【0021】

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。

【0022】

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

50

ッチに格納された線形転送テーブル（L F T）に基づき得る。L F Tは、使用中のルーティングメカニズムに従って、S M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：H C A）ポートおよびスイッチが、ローカル識別子（L I D）を使用してアドレス指定される。線形転送テーブル（L F T）における各エントリは、宛先L I D（destination LID：D L I D）と出力ポートとからなる。テーブルにおけるL I D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D L I D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（L I D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。

10

【0023】

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変化を検出し、それに応じてネットワークを再構成する。

【0024】

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（L I D）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4 9 1 5 1個のユニキャストL I D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L I Dの他に、各I Bデバイスは、6 4ビットのグローバル一意識別子（global unique identifier：G U I D）を有し得る。G U I Dは、I Bレイヤー3（L 3）アドレスであるグローバル識別子（global identifier：G I D）を形成するために使用され得る。

20

【0025】

S M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、S Mは、トポロジ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S Mが受信した場合、S M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。

30

【0026】

たとえば、S M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。

【0027】

一実施形態に従ったインフィニバンド環境1 0 0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A 1 0 1～E 1 0 5は、インフィニバンドファブリック1 2 0を使用して、それぞれのホストチャネルアダプタ1 1 1～1 1 5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A 1 0 1～E 1 0 5）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A 1 0 1～E 1 0 5）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もできる。

40

【0028】

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、I B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

50

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__キー)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。

10

【0029】

一実施形態によれば、パーティションは、あるグループのメンバが同じ論理グループの他のメンバとしか通信できないような、ポートの論理グループである。ホストチャネルアダプタ(HCA)およびスイッチでは、パーティションメンバーシップ情報を使用してパケットをフィルタリングして分離を実行できる。無効なパーティショニング情報を持つパケットは、そのパケットが着信ポートに到着するとすぐにドロップすることができる。パーティショニングされたIBシステムでは、パーティションを使用してテナントクラスタを作成できる。パーティションの適所における実施で、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はできない。このようにして、侵害されたテナントノードまたは悪意のあるテナントノードが存在する場合でも、システムのセキュリティを保証することができる。

20

【0030】

一実施形態によれば、ノード間の通信のために、管理キューペア(QP0およびQP1)を除き、キューペア(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。次に、P__キー情報を、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着すると、そのP__キー値を、SMによって構成されたテーブルに対して確認することができる。無効のP__キー値が見つかった場合、そのパケットは直ちに破棄される。このように、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。

30

【0031】

IBパーティションのある例が、図2に示されており、それは、一実施形態による、パーティショニングされたクラスタ環境を示している。図2に示す例では、ノードA101~E101~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。ノードA~Eは、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ため、ノードD104とノードE105とは、これらのノードがパーティションを共有しないので、通信することができない。一方、例えば、ノードA101とノードC103とは、これらのノードが両方ともパーティション2、140のメンバであるため、通信が許可される。

40

【0032】

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 R

50

oot 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。

【0033】

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。

10

【0034】

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。

20

【0035】

インフィニバンドにおける層およびアドレス指定

IB アーキテクチャは、複数の層に分割され、複数の層の各々は、他の層とは別々にかつ独立して動作する。IB 層抽象化の一端では、IB 物理層は IB システムの電気的および機械的特性を定義し、IB 層抽象化の他端では、IB 上位層はホストとリモートクライアントとの間でトランザクションを通信する。IB トランスポート層は、それぞれ、データの送受信時に、パーティショニング、チャネル多重化、トランスポートサービス、ならびにパケットセグメンテーションおよび再アセンブリを提供するように動作する。サブネット内におけるパケット転送およびスイッチングは IB リンク層で処理され、IB ネットワーク層はあるサブネットから別のサブネットへのパケットのルーティングを処理する。

30

【0036】

一般に、ネットワーク層は、単一のサブネット内および異なるサブネット間でパケットをルーティングするためのプロトコルを定義する。このために、IB アーキテクチャは 3 つの異なるタイプのアドレスを使用する。第 1 のタイプの IB 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 つの一意の LID が SM によって各 HCA ポートおよび各スイッチに割り当てられる。LID は、サブネット内においてトラフィックをルーティングするために使用され、リンクレベルスイッチングは、パケットを、パケットのローカルルートヘッダ (LRH) 内のソース LID (SLID) によって指定されるデバイスから、ローカルルートヘッダ LRH 内の宛先 LID (DLID) によって指定されるデバイスに転送する。LID が 16 ビット長であるので、65536 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 (0x0001-0xBFFF) 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。

40

【0037】

第 2 のタイプの IB 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

50

およびスイッチ)ならびに各々のH C Aポートに割当てられた64ビットのグローバル意識別子(G U I D)である。S Mは、H C Aポートに追加のサブネット固有G U I Dを割当ててもよく、これは、S R - I O Vが用いられる場合に有用となる。ネットワーク層で動作するルータは、異なるサブネット間でグローバルルートヘッダ(G R H)を含むパケットを送信する。ルータは、各デバイスの一意のG U I Dを使用して、サブネット間でパケットを転送する。このプロセスでは、宛先サブネットにおける宛先ポートに向かうパケット経路における最後のルータは、L R HにおけるソースL I Dを宛先ポートの適切なL I Dに置き換えることによって、パケットのL R Hを修正する。

【0038】

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(G I D)である。G I Dは有効なI P v 6ユニキャストアドレスであり、少なくとも1つが各々のH C Aポートに割当てられている。G I D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H C AポートのG U I Dアドレスとを組み合わせることによって形成される。G I DはL I Dから独立しているため、サブネットの再構成によって影響を受けないままである。

【0039】

ファットツリー(Fat Tree: F T r e e)トポロジーおよびルーティング

一実施形態によれば、I BベースのH P C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。

【0040】

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。

【0041】

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。

【0042】

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。

【0043】

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、I B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、O F E D (Open Fabric Enterprise Distribution: I B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O p e n S Mにおいて実現される。

【0044】

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

10

20

30

40

50

リンクにわたって最短経路ルートを均一に広げる L F T 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 L I D、ひいては対応するルートを各スイッチポートに割当てて、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。

【0045】

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 L F T 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。

【0046】

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ルリングすることができる階層ネットワークトポロジである。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: X G F T)、パラレルポート一般化ファットツリー (Parallel Ports Generalized Fat-Tree: P G F T) およびリアルライフファットツリー (Real Life Fat-Tree: R L F T) 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。

【0047】

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。X G F T 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。P G F T 定義はさらに、X G F T トポロジを拡張して、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は X G F T および P G F T 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 H P C 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、P G F T の制限バージョンである R L F T が導入されている。R L F T 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。

【0048】

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: I O V) 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、I O V により利用可能性をもたらすことができ、最新の C P U 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。

【0049】

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 I O V が所望される。I O V は、VM にエク

10

20

30

40

50

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-I/OV）などのさまざまなタイプの I/OV 技術が存在し得る。

【0050】

一実施形態に従うと、あるタイプの I/OV 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な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。

10

【0051】

一実施形態に従うと、別のタイプの I/OV 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。

20

【0052】

一実施形態に従うと、シングルルート I/OV（Single Root I/OV：SR-I/OV）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-I/OV は、純粹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-I/OV 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 I/OV 技術の中でも、SR-I/OV 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-I/OV 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。

30

【0053】

SR-I/OV 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トすることを可能にする。各々の SR-I/OV 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、固有のリクエスト ID が割当てられている。固有のリクエスト 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。

40

【0054】

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

50

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。

【0055】

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ーでもたらし可能性がある。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。

【0056】

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。

【0057】

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。

【0058】

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。

【0059】

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トされるが、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が破棄され）、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。

【0060】

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。

【0061】

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

10

20

30

40

50

）は、その L I D を宛先ハイパーバイザの L I D に変更させなければならない。さらに、Q P O アクセスが制限された結果、サブネットマネージャは V M の内部で実行させることができなくなる。

【0062】

インフィニバンド S R - I O V アーキテクチャモデル—仮想スイッチ (v S w i t c h)

図5は、一実施形態に従った例示的な v S w i t c h 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り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。

10

【0063】

一実施形態に従うと、v S w i t c h 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: v H C A) であり、これは、ハードウェアにおいて、V F に割り当てられた V M に、I B アドレス一式（たとえば G I D、G U I D、L I D）および専用の Q P スペースが割り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 S M については、H C A 400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 P F 420 を用いることができ、（仮想機能に付与された）V M は V F を用いることができる。

20

【0064】

一実施形態に従うと、v S w i t c h 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 L I D が割り当てられているので、利用可能な数の L I D が速やかに消費される。同様に、多くの L I D 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 S M によって演算しなければならず、それらの L F T を更新するために、より多くのサブネット管理パケット (S M P) 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。L I D スペースが49151個のユニキャスト L I D に制限されており、（V F を介する）各々の V M として、物理ノードおよびスイッチが L I D 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 V M の数が制限されてしまい、逆の場合も同様に制限される。

30

【0065】

インフィニバンド S R - I O V アーキテクチャモデル—仮想ポート (v P o r t)

図6は、一実施形態に従った例示的な v P o r t 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り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。

40

【0066】

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 v P o r t 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 S R I O V 専用とすべきであるとは規定されていない）、v P o r t の目的は、V M 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。v P o r t 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、S R - I O V 共有のポートのようなアーキテクチャおよび v S w i t c h 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、v P o r t はオプションの L I D をサポートするとともに、共有のポートとは異なり、S M は、v P o r t が専用の L I D 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 v P o r t を認識する。

50

【0067】

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。

【0068】

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。

【0069】

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。

【0070】

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、HCA 510、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。

【0071】

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。

【0072】

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。

【0073】

一実施形態に従うと、L I Dが予めポピュレートされたv S w i t c h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L I DをP Fを介して消費し、各々の追加のV Fごとに1つ以上のL I Dを消費することができる。I B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V F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I B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L I D（16個の仮想機能ごとに1つのL I Dと、物理機能のために1つのL I D）を消費する。このようなI B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L I Dの数によって規定されており、（49151個の利用可能なL I Dをハイパーバイザごとに17個のL I D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V Fを掛けて得られる）46256である（実質的には、I B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L I D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、v S w i t c hが、L I DをP Fと共有することができるので、付加的なL I Dを占有する必要がないことに留意されたい。

10

【0074】

一実施形態に従うと、L I Dが予めポピュレートされたv S w i t c hアーキテクチャにおいては、ネットワークが一旦ブートされると、すべてのL I D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L I D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。

20

【0075】

一実施形態に従うと、L I Dが予めポピュレートされたv S w i t c h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、L I Dを連続的にすることを必要とするL M C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L I Dマスク制御ライク（LID-Mask-Control-like：L M C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L I Dを宛先に送達する必要がある場合、不連続なL I Dを自由に使用できることは特に有用となる。

30

【0076】

一実施形態に従うと、L I Dが予めポピュレートされたv S w i t c h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、S R - I O V v S w i t c h対応のサブネットにおいてL I D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L I D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なる可能性がある。

40

【0077】

インフィニバンドS R - I O Vアーキテクチャモデルー動的L I D割当てがなされたv S w i t c h

一実施形態に従うと、本開示は、動的L I D割当てがなされたv S w i t c h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。

【0078】

図8は、一実施形態に従った、動的L I D割当てがなされた例示的なv S w i t c h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

50

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514～516、524～526、534～536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。

【0079】

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。

【0080】

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。

【0081】

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524～526および534～536はLIDの割当てを受けない。

【0082】

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。

【0083】

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

10

20

30

40

50

ットに接続され得る。

【0084】

一実施形態に従うと、動的L I D割当てがなされたv S w i t c h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L I Dも同様に発見される。しかしながら、新しく追加されたL I Dを処理するためのスイッチのL F T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I B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。

10

【0085】

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V FがP F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L F Tを繰返し、(VMが作成されている)ハイパーバイザのP Fに属するL I Dエントリから新しく追加されたL I Dにフォー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L F Tブロックを更新するために単一のS M P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。

【0086】

20

一実施形態に従うと、動的L I D割当てアーキテクチャを備えたv S w i t c hにおいて割当てられたL I D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L I DをL I Dが予めポピュレートされたv S w i t c hと動的L I D割当てがなされたv S w i t c hとで比較すると、動的L I D割当てアーキテクチャにおいて割当てられたL I Dが不連続であり、そこに予めポピュレートされたL I Dが本質的に連続的であることが分かるだろう。さらに、v S w i t c h動的L I D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L I Dが、VMの生存期間の間中ずっと用いられる。逆に、L I Dが予めポピュレートされたv S w i t c hにおいては、各々のVMは、対応するV Fに既に割当てられているL I D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V Fに連続的に付与されたVMが同じL I Dを得る。

30

【0087】

一実施形態に従うと、動的L I D割当てアーキテクチャを備えたv S w i t c hは、いくらかの追加のネットワークおよびランタイムS Mオーバーヘッドを犠牲にして、予めポピュレートされたL I Dアーキテクチャモデルを備えたv S w i t c h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L I Dで、サブネットにおける物理スイッチのL F T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(S M P)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、L M C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V Fの合計に対する制限はなく、V Fの数は、ユニキャストL I D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V F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V Fを備えることにより、ユニキャストL I D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。

40

【0088】

インフィニバンドS R-I O Vアーキテクチャモデルー動的L I D割当てがなされかつL I Dが予めポピュレートされたv S w i t c h

図9は、一実施形態に従った、動的L I D割当てがなされてL I Dが予めポピュレートされたv S w i t c hを備えた例示的なv S w i t c hアーキテクチャを示す。図に示さ

50

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。

10

【0089】

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。

20

【0090】

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。

【0091】

30

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。

40

【0092】

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v

50

Switch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8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） $0 \leq \text{予めポピュレートされたVF} \leq \text{総VF}$ 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。

【0093】

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。

【0094】

インフィニバンドーサブネット間通信（ファブリックマネージャ）

一実施形態によれば、本開示の実施形態は、単一サブネット内に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、2つ以上のサブネットにまたがるインフィニバンドファブリックを提供することもできる。

【0095】

図10は、一実施形態による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。図に示すように、サブネットA1000内では、ある数のスイッチ1001～1004が、サブネットA1000（例えば、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において通信を提供することができる。ファブリックは、例えば、ホストチャネルアダプタ1010などのある数のハードウェアデバイスを含むことができる。ホストチャネルアダプタ1010は、次いで、ハイパーバイザ1011と対話することができる。ハイパーバイザは、次いで、それが対話するホストチャネルアダプタと関連して、ある数の仮想機能1014をセットアップすることができる。ハイパーバイザは、加えて、仮想マシンを仮想機能の各々に割り当てることができる。仮想マシン10105が仮想機能1014に割り当てられるなどする。ハイパーバイザは、各ホストチャネルアダプタ上において、物理機能1013など、十分な機能を有する物理機能を介して、それらの関連付けられるホストチャネルアダプタにアクセスできる。サブネットB1040内では、ある数のスイッチ1021～1024が、サブネットB1040（例えば、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において通信を提供することができる。ファブリックは、例えば、ホストチャネルアダプタ1030などのある数のハードウェアデバイスを含むことができる。ホストチャネルアダプタ1030は、次いで、ハイパーバイザ1031と対話することができる。ハイパーバイザは、次いで、それが対話するホストチャネルアダプタと関連して、ある数の仮想機能1034をセットアップすることができる。ハイパーバイザは、加えて、仮想マシンを仮想機能の各々に割り当てることができる。仮想マシン21035が仮想機能21034に割り当てられるなどする。ハイパーバイザは、各ホストチャネルアダプタ上において、物理機能1033など、十分な機能を有する物理機能を介して、それらの関連付けられるホストチャネルアダプタにアクセスできる。各サブネット（すなわち、サブネットAおよびサブネットB）内には1つのホストチャネルアダプタしか示されていないが、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を各サブネット内に含めることができることを理解されたい。

【0096】

一実施形態によれば、ホストチャネルアダプタの各々は、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さらに関連付けることができ、各HCAは、上述したように、異なるアーキテクチャモデルでセットアップすることができる。図10内の両方のサブネットは、事前にポピュレートされたLIDアーキテクチャモデルを有するvSwitchを使用しているように示されているが、これはすべてのそのようなサブ

10

20

30

40

50

ネット構成が同様のアーキテクチャモデルに従わなければならないことを意味するものではない。

【0097】

一実施形態によれば、各サブネット内の少なくとも1つのスイッチはルータに関連付けられることができ、サブネットA1000内のスイッチ1002はルータ1005に関連付けられ、サブネットB1040内のスイッチ1021はルータ1006に関連付けられるなどする。

【0098】

一実施形態によれば、少なくとも1つのデバイス（例えば、スイッチ、ノードなど）をファブリックマネージャ（図示せず）に関連付けることができる。ファブリックマネージャは、例えば、サブネット間ファブリックトポロジを発見し、ファブリックプロファイル（例えば、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データベースオブジェクトを構築するために使用することができる。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パーティション番号を使用してどのルータポートを介して通信することが許可されているかに関して、法的なサブネット間接続を定義することができる。

【0099】

一実施形態によれば、サブネットA内の仮想マシン1などの発信元でのトラフィックが、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の宛先にアドレス指定されている場合、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、ルータ1005は次いでそのトラフィックをルータ1006とのそのリンクを介してサブネットBに渡すことができる。

【0100】

全体的に、例示的な実施形態は、非常に大きなファブリックに多数のノードを提供し、さらに、サブネット境界を各々が有する複数のサブネットを提供する。サブネット境界は、独立したサブネットマネージャ（SM）を可能にする。特に、サブネット境界は、非常に大きなファブリック内の複数のサブネットの各々において／について1つの独立したSMを可能にする。

【0101】

さらに、非常に効率的なパケットルーティングが、ファブリック全体にわたって独自の方法でパケットフローをルーティングするために、線形転送テーブル（LFT）探索プロトコルの単純な探索メカニズムの使用によって可能になる。

【0102】

ローカル／サブネット内ルーティング：

レイヤ2（L2）アドレス指定の場合、例示的な実施形態は、各サブネット内において「通常の」ローカルIDベースのルーティング／転送を使用する。つまり、IBパケットのローカルルートヘッダ（LRH）における16ビットの宛先ローカル識別子（DLID）は、スイッチポートでルーティングされているときに探索され、出力ポートは、IBパケットのLRHにおけるDLID値を使用してハードウェアでLFTを直接索引付けすることによって迅速に発見される。

【0103】

レイヤ3（L3）アドレス／アドレス指定：

例示的な実施形態におけるパケットのグローバルルートヘッダ（GRH）は、本質的に128ビットスキーム、すなわち一意の64ビットのノードアドレス - 「GUID」、および64ビットのサブネットプレフィックス番号を使用する。例示的な実施形態は、多数の独立したサブネットに分割された非常に大きなファブリックを提供し、同じルート／ルーティング能力が異なるサブネットにまたがって使用可能にされる。すなわち、例示的な実施形態では、エンドポイント間の経路は、あたかも大きな単一のサブネットがあるかのような場合と同じ方法でルーティングすることができる。

【0104】

仮想マシン（VM）では、物理エンドノードは、多数の仮想エンドポイントを表すことができ、各仮想エンドポイントは、16ビットのDLIDを含むそれら自体の完全なIBアドレスを有している（かつ、それを、レガシー互換性の理由のため、有するべきである）しかしながら、16ビットのDLID空間はすぐに使い尽くされるようになる。本質的に、各仮想サブネットには、16ビットDLIDを使用するアドレス空間ポインティング機能があるよりも多くの空間がある。

【0105】

したがって、例示的な実施形態によれば、全ファブリックは、独立したサブネットのセットに分割され、プライベートな「LID空間」が提供され、L3アドレスの64ビットのサブネットプレフィックス番号部分から指定されたビットフィールドを抽出するためのメカニズムが提供される。例示的な実施形態のメカニズムによれば、サブネット番号を符号化するためにL3アドレスの64ビット全体を使用する代わりに、本実施形態は、指定されたビットフィールド（ISRN）をL3アドレスの64ビットのサブネットプレフィックス番号部分から抽出するためのハードウェアサポートを提供する。したがって、中間インフラストラクチャを介して（仮想）サブネットAから（仮想）サブネットBへの通信がある場合、例示の実施形態は、L3アドレスの64ビットのサブネットプレフィックス番号部分の指定されたビットフィールドにおいて、

- a) 宛先サブネットにおけるLID／またはそのLIDおよび
- b) 「中間スイッチファブリック」インフラストラクチャの「中間」LIDの両方をエンコードする。

【0106】

グローバルルートヘッダ（GRH）ベースの線形転送テーブル（LFT）探索

IBファブリック全体にわたるルーティング／転送の調整された管理を伴う単一のファブリック構成内においてサブネット内トラフィックおよびサブネット間トラフィックの両方に対してスケーラブルで柔軟性があり決定論的なルーティング／転送を可能にするために必要なスイッチ／ルータオンチップリソースの量を削減するために、ある実施形態に従って、単一のサブネット内のLIDベースのテーブル探索と同じ方法で線形テーブル探索のために使用される番号を介してサブネット間のルートを識別するためのスキームが使用される。これに関して、図11は、一実施形態による、ネットワーク環境においてパケット転送ロジックにアクセスするためにインフィニバンド（IB）アドレス指定スキームを使用するデータパケットフォーマットの図を示す。図に示すように、第1のIBサブネット1100は、複数の物理（または仮想）ノード1101および1102と、サブネットアドミニストレータ（SA）1120とを含むことができる。第1のIBサブネット1100内のソースノード1101は、中間ノード1102を介して、第1のIBサブネット1100以外の第2のIBサブネットにおける宛先ノード1103にパケット（たとえば、IBパケット1110）を送信することができる。宛先ノード1103は、それが第2のIBサブネット内にあることを表すために点線で示されている。

【0107】

IBパケット1110は、ペイロード1114と、IBプロトコルによるさまざまなヘッダとを含むことができる。これらのヘッダは、グローバルルーティングヘッダ（GRH）1111、ローカルルーティングヘッダ（LRH）1112、および他のヘッダ1113を含むことができる。さらに、IBパケット1110は、さまざまな巡回冗長検査（CRC）1115とともに適用することができる。

【0108】

一実施形態によれば、システムは、第1のIBサブネット1100および第2のIBサブネットにおけるサブネット内およびサブネット間アドレス指定をサポートするために、GRH1111における宛先グローバル識別子（DGID）1121のサブネットプレフィックス部分およびLRH1112における宛先ローカル識別子（DLID）1122を利用できる。

【0109】

たとえば、システムは、（宛先ノード1103のためのDLIDの代わりに）中間ノード1102のためのDLIDになるように、IBパケット1110におけるDLID1122を設定することができる。IBサブネット1100内では、IBパケット1110は、SA1120によって解かれたDLID1122に基づいて、線形転送テーブルLFT探索を介して中間ノード1102にルーティングすることができる。したがって、IBパケット1110は、中間ノード1102上に設けられたパケット転送ロジックを使用して処理することができる。同様のパケット転送ロジックを、必要であるように、または所望のように、他のノード1101および1103に設けることができる。

【0110】

さらに、システムは、GRH1111におけるDGID1121の選択された部分を使用して、そのDLIDを宛先ノード1103のために示すことができる。したがって、中間ノード1102におけるパケット転送ロジックは、GRH1111におけるDGID1121情報に基づいて、宛先ノード1103のための実際のDLIDを解く（または取得する）ことができる。

【0111】

一実施形態によれば、中間ノード1102は、必要に応じて追加のパケットヘッダ1113および／またはペイロード1114の修正を行うことができる。たとえば、ファブリックレベルアクセス制御は、ソースノード1101および宛先ノード1103が、関連するパーティションの限定されたメンバーであるか、または同じパーティションのメンバーでないように設定され得る。そのような場合、中間ノード1102は、修正されたパケットを宛先ノード1103に転送する前に、IBパケット1110におけるP_Key値を変更することを許可されてもよい。

【0112】

図11は、一実施形態に従ってターゲットサブネット番号1142およびターゲットLID1143と共に使用されるサブネット間ルート番号（ISRN）1141を含むサブネットプレフィックス転送部分1131を含む例示的なパケット1100のDGID部分1121を示す。ISRN1141は、サブネット内転送判断を、サブネット間転送判断と並んで両方を行うために、同じ線形転送テーブル（LFT）が、実施形態に従って構成されたファブリックスイッチによって使用されることを可能にする。実施形態は、単一のファブリック構成内のサブネット間トラフィックが、異なるファブリック間のトラフィックとは異なるサブネットプレフィックス値を使用できるように、ファブリックローカル／内部サブネット番号および「グローバル」サブネット番号の両方をサポートする。

【0113】

ソースから最終的な宛先への経路に沿った任意のポイントにおけるGRH、DGIDからDLIDへのどのようなマッピングも行なう必要性なしに、エンドツーエンドルートをパケットヘッダ内において完全に定義できるようにするために、ある実施形態は、サブネット間ルート番号、ターゲット「ファブリックローカルサブネット番号」、およびターゲットサブネットにおける宛先LID値の両方を含むGRH、DGIDサブネットプレフィックスフォーマットを提供する。

【0114】

スイッチ／ルータポートに関連付けられる転送ドメインのスキームまたはプロトコルは、ある例示的な実施形態においてサポートされ、各転送ドメインタイプは、LFTにおける探索のためにどのアドレス情報が使用されるか、および転送ドメインタイプの変更が、ローカルルートヘッダをGRHからのDLID値で更新する必要があることを意味するかどうかを、定義する。

【0115】

ソースサブネット、中間スイッチファブリックおよびターゲットサブネットを有する3レベル転送ドメイン構造をサポートするために、実施形態に従って上記のDGID、サブネットプレフィックスフォーマットが使用される。ソースサブネット1100におけるソースノード1101および中間ノード1102は、線形転送テーブル（LFT）探索のた

10

20

30

40

50

めにパケット1100の最初のLRH、DLIDを用い、中間スイッチファブリックにおける宛先ノード1103への遷移は、サブネットプレフィックス転送部分1131のサブネット間ルート番号1141がLFT探索に使用されることを意味する。中間スイッチファブリックは、便宜上、本明細書では「コアファブリック」とも呼ばれ得る。

【0116】

パケットが中間スイッチファブリックタイプ転送ドメイン内で転送されていることを示すために、一実施形態によれば、特殊なDLID値が使用され、特殊なDLID値は、入口スイッチポートにおけるレジスタに格納される値と照合される。特殊なDLID値は、「特殊スイッチポート」を表し、特殊スイッチポートはローカル仮想サブネットの「終わり」を定義する。これにより、単一のLFT探索操作を、LRH、DLID値ではなく、GRH、DGID、サブネットプレフィックスの関連セクションに基づかせることができる。

10

【0117】

一実施形態によれば、転送ドメインの変更がいつ行われるかを判断するために、各入口スイッチポートは、同じスイッチノードにおける各可能な出口ポートについてのタイプステータスを伴う1つ以上のレジスタを含み、LFTで探索されるターゲットポートが転送ドメインタイプの変更を表わす場合に、進入経路におけるパケット処理が次の転送ドメインのためのパケットを準備できるようにする。

【0118】

例示的な実施形態によれば、IBパケット1110が、ソースノード1101のポートからなどのような、ソースサブネット1100内のポートから、中間ノード1102のポートへなどのような、同様にソースサブネット1100内のポートへのサブネット内転送である場合、パケットのLRH、DLIDを直接使用して、標準的なサブネット内パケット転送のために各スイッチでLFTを索引付けする。

20

【0119】

さらに、例示的な実施形態によれば、IBパケット1110が、ソースノード1101のポートからなどのような、ソースサブネット1100内のポートから、ソースサブネット1100外の中間スイッチファブリックへの、サブネット間転送である場合、パケット1110のローカルルートヘッダ1112のDLID1122部分は、ソースサブネット1100外の関連する中間スイッチファブリックにおける中間スイッチファブリック転送を表すように構成された値にローカルに更新される。

30

【0120】

さらに別の例示的な実施形態によれば、図11を続けて参照すると、IBパケットが宛先ノード1103などのソースサブネット1100外の中間スイッチファブリックからソースノード1101または中間ノード1102などのターゲットソースサブネット1100へのサブネット間転送である場合、LRH、DLIDはDGID、サブネットプレフィックスからのDLIDフィールドで更新され、次いで、それを用いて、パケットを、最終的な宛先ポートに、ターゲットソースサブネット1100におけるLFT探索を用いて転送する。この遷移では、パケットGRH、DGID、サブネットプレフィックスにおけるターゲットローカルサブネット番号が、このポートが表すローカルサブネット番号と一致しているかどうかを検証することもできる。このチェックは、好ましくは、そのような各ポートにおけるレジスタに基づく。

40

【0121】

例示的な実施形態では、任意のスイッチポートは、ローカル（ソース）サブネットと中間スイッチファブリックとの間の「ゲートウェイ」を表すことができ、そのような各ゲートウェイスイッチポートは、ローカルサブネットにおける1つ以上の特定のDLIDによって識別される。したがって、複数の並列コアファブリックが存在することができるファットツリーベースのトポロジを実現することが可能である。このようにして、各並列コアファブリックは独立してルーティングされるドメインを表すことができるため、完全なファブリックのアドレス指定スケーラビリティはトポロジでスケーリングできる。48Kの

50

L F Tサイズは、これらの並列コアファブリックの各々内での制限に過ぎず、完全なコアファブリック内の個々の「サブネット間ルート番号」の数は、並列コアファブリックの数の48K倍である。

【0122】

加えて、一実施形態によれば、内部ルートの独立した管理を伴う異なるファブリック構成間の「ランダムな」接続性を処理できるように、ある実施形態は、ファブリックローカル／内部サブネット番号を「グローバル」サブネット番号と並んで両方サポートするためのサポートスキームを提供して、単一のファブリック構成内のサブネット間トラフィックが、異なるファブリック間のトラフィックとは異なるサブネットプレフィックス値を使用できるようにする。サポートスキームは、また、任意のサブネットプレフィックスおよびG U I D値の転送を可能にするために、完全なD G I D値のT C A Mベースの探索およびマッピングを提供する。

10

【0123】

ファブリックモデル

例示的な実施形態によるコアファブリックモデルの概要として、I Bファブリック構成は、個々のサブネットに分解することができる一方で、個々の物理的または仮想的な宛先に対する個々の経路を明示的にルーティングできる柔軟性を依然として維持することができる、非常に大きなファットツリー構成に対するスケーリングを容易にする。システムは、階層的な態様で管理される単一のファブリック構成として構築されるが、最上位の管理エンティティは完全なファブリックの完全な可視性を有する。

20

【0124】

異なる独立して管理されるシステム間のネイティブなI B接続性を依然としてサポートしながら、大きなシステム構築柔軟性を提供するために、例示的な実施形態のファブリックモデルは、同一の単一の最上位のエンティティによって管理されない異なるファブリック構成／トポロジ間における「任意の」I B接続性を可能にする。

【0125】

例示的な実施形態は、上記の目標を達成するため、および単一のファブリック構成内でサブネット間通信を提供するために、経路／ルート情報をパケットG R HにおけるD G I Dフィールドのサブネットプレフィックス部分1131の一部としてエンコードする。例示的な実施形態は、独立したI Bファブリック間のI B—I Bルーティングを可能にする標準的なサブネットプレフィックススキームもサポートする。

30

【0126】

数百または数千の物理ホストを有する非常に大きな構成にスケーリングすることに加えて、例示的な実施形態のシステムは、さらに、非常に小さなサブネット構成の使用も容易にし、単一の集積回路チップを使用して、異なるサブネットマネージャ（S M）のセットによって管理される異なるI Bサブネットを各々が表わす複数の仮想スイッチを実現してもよい。加えて、非常に大きな全体ファブリック構成であっても、このような小さなサブネット構成を必要または所望に応じて依然として使用して、論理システム単位への所望の分解を反映させてもよい。一方、一部の構成では、依然としてより大きな単一のファブリックの一部である非常に大きな単一のサブネットトポロジを有してもよい。

40

【0127】

一般に、アップスケーリングおよびダウンスケーリングにおける柔軟性の必要性は、パケットルーティングおよび転送判断を処理するためのシステムオンチップリソースの点で課題を課す。これには、サブネット内およびサブネット間転送の両方、特に、異なるサブネットにおける個々の仮想エンドポイントに対処する能力が含まれる。

【0128】

ここにおける実施形態は、単一のファブリック構成内においてサブネット内トラフィックおよびサブネット間トラフィックの両方に対してスケーラブルで柔軟性があり決定論的なルーティング／転送を可能にするために必要なシステムオンチップリソースの量を削減するためのスキームを提供する。特に、「サブネット間ルート番号」（I S R N）を、単

50

一のサブネット内のL I Dベースのテーブル探索に類似する態様でテーブル探索のために使用され得る番号を介してサブネット間ルートを識別するために使用できる。さらに、それ以外の場合は、標準または「レガシー」線形転送テーブルが、サブネット内およびサブネット間の両方の転送の判断に使用される。たとえば、図12は、L R H. D L I D値またはI S R N値のいずれかによって等価的に0~48kから索引付けることができる線形転送テーブル(L F T) 1200の一部を示す。図示したL F T 1200において、テーブルにおける各エントリは、サポートされているポート番号または使用されていないエントリを表す「0×F F」エントリのいずれかを含む。更に加えて、単一のファブリック構成内のサブネット間トラフィックが、異なるファブリック間のトラフィックとは異なるサブネットプレフィックス値を使用できるように、ファブリックローカル／内部サブネット番号および「グローバル」サブネット番号の両方をサポートする。さらに、G R H. D G I D. サブネットプレフィックスフォーマットが提供され、サポートされる。例示的な実施形態では、G R H. D G I Dサブネットプレフィックスフォーマットは、ソースから最終的な宛先への経路に沿った任意のポイントにおけるG R H. D G I DからD L I Dへどのようなマッピングも行なう必要性なしに、エンドツーエンドルートをパケットヘッダ内において完全に定義できるようにするために、「サブネット間ルート番号」、ターゲット「ファブリックローカルサブネット番号」、およびターゲットサブネットにおける宛先L I D値の両方を含む。

【0129】

個々のサブネット間における「ランダム化された」マルチ経路転送を伴う階層的トポロジを可能にするために、リーフスイッチベースの転送が提供される。さらに、異なるファブリック構成間の「ランダム」接続を扱うことができるように、例示的な実施形態は、単一のファブリック構成内のサブネット間トラフィックが、異なるファブリック間のトラフィックとは異なるサブネットプレフィックス値を使用できるように、ファブリックローカル／内部サブネット番号および「グローバル」サブネット番号の両方をサポートし、任意のサブネットプレフィックスおよびG U I D値の転送を可能にするために、完全なD G I D値のT C A Mベースの探索およびマッピングを提供する。

【0130】

ローカルファブリック

一実施形態によれば、「ローカルファブリック」は、1つ以上の物理的な「リーフサブネット」および／または「コアファブリック」トポロジ内ならびにそのようなトポロジ間の接続性を表す物理的なI Bファブリックインフラストラクチャである。ローカルファブリックは階層的な態様で管理されるが、たとえば制御プレーンソフトウェアなどの最上位の中央管理エンティティが任意のI D値の割り当てを管理することができるという制約を伴い、物理接続およびルーティング（パケット転送）の両方の制約を完全なローカルファブリック（すなわち、ローカルファブリックの”全体的なビュー”）全体にわたって遵守する。

【0131】

リモートファブリック

一実施形態によれば、「リモートファブリック」は、1つ以上のI Bリンクを介してローカルファブリックに接続される任意のトポロジ／インフラストラクチャであるが、リンクのリモート側のポートは、ローカルファブリックと同じ最上位の中央管理エンティティによって制御されない。2つの独立したローカルファブリックを接続するポートは、典型的には、ルータポートとして設定される。したがって、これら2つのポートは、一実施形態によれば、最小I Bサブネットを形成する。

【0132】

ゲートウェイサブネット

一実施形態によれば、「ゲートウェイサブネット」は、典型的には、1つのローカルファブリックからの1つのルータポートを他のローカルファブリックからのちょうど1つのルータポート（すなわち、中間スイッチなし）と接続することによって形成される「単純

10

20

30

40

50

な」物理 I B サブネットである。ゲートウェイサブネットは、リンクアップ時に確立される単純な構成を持ち、パケット転送ポリシーは各側のルータによって独立して制御される。2つのルータポートは、リモートファブリックについて学習するために、相互発見プロトコルを実現する。これに代えて、および／または加えて、各ローカルファブリックにおける管理入力を、必要または所望に応じて選択的に用いて、異なるゲートウェイサブネットを介してルーティングされるべき1つ以上のリモートサブネット番号および個々の G I D のセットを定義することができる。ここでのゲートウェイサブネット定義の重要な局面は、管理すべきルータまたはスイッチデバイスではなく、I B リンク上の2つの最上位の管理ドメイン間の境界を配置することである。

【0133】

より複雑なゲートウェイサブネットは、ゲートウェイサブネット内のスイッチも管理を必要とし、この管理によって、直接リンクの場合には存在しないか、またはあまり存在しない複雑さおよびポリシーの問題が導入されることを暗示する。したがって、単一リンクの各側の十分に定義された所有者を有することよりも、ゲートウェイサブネットの所有権の問題が問題になる。しかしながら、大規模なネットワーク化の場合と同様に、複数のローカルファブリックが1つ以上の他のローカルファブリックを介して接続されている、より複雑な構成を想像することが可能である（つまり、中間ローカルファブリックのアドミニストレータは、接続された異なるリモートファブリックがこのローカルファブリックを介して通信することを許される方法を判断する）。本明細書の実施形態は、各ローカルファブリックが複数の他のローカルファブリックへの接続性を有することができるので、このより複雑なゲートウェイモデルに適用可能である。いずれにしても、これらのタイプの中間ローカルファブリックであっても、ローカルファブリックの任意の対の間の境界は、好ましくは、上で定義した最小ゲートウェイサブネットである。

【0134】

リーフサブネット

「リーフサブネット」は、それ自体のユニキャストおよびマルチキャスト L I D 空間を有する物理的な I B サブネットである。計算およびストレージサーバなどのホストタイプのエンドノードは、典型的には、単一のリーフサブネットを介してローカルファブリックに接続される。しかしながら、単一のホストが、さらに所望の通りに複数のリーフサブネットと接続されてもよい。リーフサブネットは、1つ以上の協働サブネットマネージャ（S M）インスタンスによって構成および制御され、リーフサブネット内では、パケットはパケットヘッダの L R H . D L I D 部分に基づいて転送される。リーフサブネットは、レガシースイッチで構成されてもよく、中央の最上位の管理エンティティと並んで、リーフサブネットを制御しているサブネットマネージャのセットに入力されるポリシーを表す管理上定義された「識別サブネット番号」（I S N）によって識別される。I S N は、グローバルに一意であってもよく、好ましくは、I B リンクを介して、すなわちゲートウェイサブネットを介して直接接続できるすべてのローカルファブリックにわたって一意である。各リーフサブネットは、ポートを介して1つ以上の「コアファブリック」に接続されてもよく、各リーフサブネットは、少なくとも1つのコアファブリックに接続されると、少なくとも1つの「ファブリックローカルサブネット番号」（F L S N）に関連付けられる。F L S N 値は、ローカルファブリックについての範囲を持ち、ローカルファブリック管理インフラストラクチャによって割り当てられる。加えて、リーフサブネットは、コアファブリック接続性を表さないルータポートを介して他の I B サブネットに接続されてもよい。一般的な場合において、このような「グローバル」接続は、ゲートウェイサブネットを介したリモートファブリックへの接続を表し、この種のグローバル接続の場合、I S N 値は、リモートサブネット／ファブリックからのこのリーフサブネットの D G I D ベースのアドレス指定に使用される。しかしながら、2つのリーフサブネットを同じローカルファブリックにおいてこのように、すなわちローカルコアファブリックを使用せずに、接続することも可能である。この種の接続は、I S N または F L S N ベースの D G I D アドレス指定のどちらを使用してもよい。直接、リーフサブネットからの、ローカルファブリッ

ク内またはリモートファブリックへの、ルータポートベースの接続には、システムポートまたは他のルータ実現例のどちらが関係してもよい。

【0135】

コアファブリック

一実施形態によれば、「コアファブリック」は、単一のローカルファブリックのコンテキスト内で1つ以上のリーフサブネットを接続する例示的な実施形態によるトポロジである。ここでは、「中間スイッチファブリック」という表現を代わりに使用してもよい。一実施形態によれば、コアファブリックは、たとえば、同じ物理サブネット内に既に直接的な接続性を有するセクションなど、単一のリーフサブネットの異なるセクション間に追加の接続性を提供する。例示的な実施形態では、単一のコアファブリックは、単一の「サブ
10 ネット間ルート番号」(ISRN)値空間を表すことが好ましい。単一のコアファブリックは直接接続されたエンドノードを有することができる。エンドノードの特殊なケースは、スイッチ管理ポート(ポート0)である。これらは専用のISRN値を介して到達することができる、GRHベースの通信が必要である。物理IBリンクおよび/またはシステムポートを介して接続される物理エンドノードは、一般に、別個のFLSN(またはISN)
20 値を表し、通信は、好ましくはGRHベースである。このような直接接続されたエンドノードは、一般に、わずか2つのポートおよび単一のリンク、すなわちエンドノードポートおよび関連するシステムポートを有する、「単純な」リーフサブネットを表す。例示的な実施形態の環境では、直接接続された物理エンドノードは、DLID値が一意でないかもしれないことをエンドノード実現例が処理できる限り、専用ISRN値を介してアドレス指定することもできることを理解されたい。一般に、FLSN値はローカルファブリック全体に対してグローバルでなければならないので、必要な値の空間を減らすために、別
個のFLSN値をそのようなエンドポートについて選択的に割り当てないかもしれない。

【0136】

ある例示的な実施形態による単一のコアファブリックは、従来の物理的IBサブネット内での他の処理または取り扱いと同様の態様で直接ルート(DR)サブネット管理パケット(SMP)を処理または他の態様で取り扱う。しかしながら、一実施形態によれば、単一のコアファブリックは、LIDベースの転送をISRNベースの転送と並んでサポートするように構成されている。したがって、使用される総LID値空間は、使用される総ISRN値空間と調整される。調整には、両方のタイプの転送に同じ物理的LFTリソース
30 の異なるセクションを使用すること、およびこれらのセクションのサイズを各システムポートについて実行時設定パラメータとして設定することが含まれる。

【0137】

例示的な実施形態では、重複するLID値およびISRN値について、2つの使用される値空間の合計が、関与するハードウェアコンポーネントによってサポートされ得る限り、制限はない。十分なLID値空間が割り当てられていると仮定すると、システムのポート0インスタンスおよび直接接続されたエンドノードの両方に、LIDベースのアドレス指定を介して、到達することができる。このように、単一のコアファブリックインスタンスは、従来のIBサブネットを表し、同時に、1つ以上の(他の)リーフサブネット間の
40 コアファブリックとして機能しているので、ローカルファブリック内の宛先リーフサブネットを表すこともできる。

【0138】

加えて、例示的な実施形態では、単一コアファブリックは、1つ以上の他の各単一コアファブリックごとに独立したISRN値空間を依然として維持しながら、そのような単一
コアファブリックへの直接的な物理的接続性を有することができる。DR SMPを、サポートされている任意の他のタイプの転送に依存することなく、2つのコアファブリック
インスタンス間で通信するよう使用してもよい。値空間の調整と、LIDまたはISRN空間の特定のサブセットの予約とに基づいて、2つのコアファブリックにおけるエンド
ノードのすべてまたはサブセット間の通信を、LIDまたはISRNベースのアドレス指定
50 を使用して許可し得るかまたは他の態様で可能にする。

【0139】

単一のコアファブリックは、例示的な実施形態では、リモートファブリックへの接続性を表す1つ以上のルータポートを有してもよい。つまり、それらは、調整されたFLSNまたはISRN空間のない独立したローカルファブリックである。このようなリモート接続性は、好ましくはゲートウェイサブネットを使用して実現される。この場合、ローカルエンドノードまたは接続されたリーフサブネットとの間でパケットを転送することは、ISNベースのサブネットプレフィックス値を伴うGRH、およびDGIDフィールドにおける関連付けられるGUID値に基づく。そのような転送は、DGIDの任意の部分の任意の特定のフォーマットに依存せず、代わりにTCAMベースの転送に依存する。

【0140】

アドレス指定モード

一実施形態によれば、パケットの「アドレス指定モード」は、パケットがソースから宛先へどのように送られるべきかを判断するために使用されるべき特定のヘッダ情報を定義する。IBパケットは、パケットがソースから宛先へどのように送られるべきかに関する情報を含む複数のヘッダ（すなわち、LRHおよびGRH）を含んでもよい。したがって、パケットのアドレス指定モードは、どのようなそのような情報が次の転送判断に使用されるべきかを定義する。

【0141】

例示的な実施形態では、この文脈における関連するアドレス指定モードは、以下を含む：

- LRH/DLIDアドレス指定モード；
- GRH/ISRNアドレス指定モード；
- GRH/リーフスイッチアドレス指定モード；および
- GRH/DGIDアドレス指定モード。

【0142】

転送モード

一実施形態によれば、「転送モード」は、パケット転送判断を実行するスイッチポートに適用される。「アドレス指定モード」の概念はIBパケットのヘッダに適用されるが、転送の判断を行うスイッチポートは、あるパケットアドレス指定モードに対応する「転送モード」で動作する。スイッチポートは1つより多いアドレス指定モードを同時にサポートしてもよい。したがって、スイッチポートが特定のパケットのために正しい転送モードを選択することを可能にするために、パケットは、スイッチポートが決定論的にアドレス指定モードを判断することを可能にする情報を含まなければならないかもしれない。そのような情報には、DGIDタイプ情報および特定の転送モード（たとえば、GRH/ISRN）を示すために予約された特殊なDLID値の両方が含まれる。また、場合によっては、次ホップアドレス指定モードは、パケットにおけるアドレス情報およびファブリックにおけるポートに関連付けられる構成情報の両方の関数である。宛先リーフサブネットに到達した時点进行判断するための例示的な実施形態における能力は、次ホップアドレス指定モードがパケットにおけるアドレス情報およびファブリックにおけるポートに関連付けられる構成情報の両方の関数であることの例である。

【0143】

転送ドメイン

一実施形態によれば、任意の特定のリーフサブネット、コアファブリックまたはゲートウェイサブネットインスタンスは、「転送ドメイン」を表わしてもよい。IBパケットが完全なIBファブリックを通じて転送されると、図13に示すような1つ以上の「転送ドメイン」を通過し、そのような各転送ドメインは特定の転送モードおよび特定のアドレス値空間を表す。一実施形態によれば、「転送ドメイン」は、複数のリーフサブネットおよび複数のコアファブリックの間の境界である。最も単純なスキームでは、個々のシステムインスタンスは、単一の転送ドメイン、たとえば単一のリーフサブネットまたは単一のコアファブリックインスタンスに属する。この場合、リーフサブネットまたはコアファブリ

10

20

30

40

50

ックインスタンスの任意の対間のすべての境界は、異なるシステムインスタンスを接続する1つ以上のリンク上に存在する。リーフサブネットおよびコアファブリックインスタンスの両方が任意のサイズを有することを可能にするために、個々のシステムインスタンスは、2つ以上の「仮想スイッチ」インスタンスに選択的に分割され、各そのような仮想スイッチは、すべてが単一のリーフサブネットまたは単一のコアファブリックへの、またはその内における接続性を表す、1つ以上の「仮想スイッチポート」の集まりを表す。この場合、転送ドメインの任意の対間の境界は、2つ以上の仮想スイッチインスタンスを接続するシステムクロスバー内に存在する。

【0144】

図13に示すように、ソースサブネット内では、第1の転送ドメインで元の（またはレガシー）LRH、DLIDスキームを使用してLFT探索が実行される。第2の転送ドメインにおいては、図示のように、コアファブリック内における、ソースサブネットからコアファブリックへパケットを転送する際のLFT探索が、本明細書の例示的な実施形態によるISRNスキームを使用して実行される。さらに、図示のように、第3の転送ドメインにおける、コアファブリックからターゲットサブネットにパケットを転送する際のLFT探索が、ターゲットLRH、DLID（図11）に基づくLRH、DLIDを使用して実行される。

【0145】

ゲートウェイポート

一実施形態によれば、転送ドメインへの接続性を表す物理的なシステムポートは、本明細書では「ゲートウェイポート」と呼ばれる。以下に述べる理由から、ゲートウェイポートは、パケットをゲートウェイポートに送信する同じシステムインスタンス上の別のポートの観点から定義される。まず、ゲートウェイポートにパケットを送信することは、送信が次のいずれかであるため、パケットに対するアドレス指定モードが変更されることを意味する。

【0146】

- I. ソースリーフをコアファブリックに送信する。
- II. コアファブリックを宛先リーフに送信する、または
- III. ソースリーフを宛先リーフに送信する。

【0147】

パケットのアドレス指定モードは、IBリンクから受信されるパケットに対して転送判断がどのように行われるかに適用されるので、図13に示される1つの転送ドメインから他の転送ドメインへと交差するときのアドレス指定モードの変更は、一実施形態によれば、パケットがシステム入口ポートから出口ポート、たとえばゲートウェイポートなどに移動されるときに実施される。アドレス指定モードは、次いで、パケットがゲートウェイ／出口ポートからIBリンクに転送されるときにアクティブになる。

【0148】

この理由から、ゲートウェイポートは、パケットをゲートウェイポートに送信する同じシステムインスタンス上の別のポートの観点から定義される。

【0149】

異なる転送ドメイン間の境界を表すリンクは、通常、リンクの両側にゲートウェイポートを有することになることを理解されたい。リンクの両側にゲートウェイポートを持つ動機は、転送ポリシーの更新がリンクの各側に確実に含まれるようにすることで、リンクにわたる書き込み型管理操作の必要性を排除することである。たとえば、リンクの一方の側がコアファブリックを表し、他方の側がリーフサブネットを表す場合、リーフサブネット側のサブネットマネージャは、例示的な実施形態に従って、IBリンク上で受信されるパケットの転送の完全な制御を有する。

【0150】

例示的な実施形態では、リンク上のパケットは、典型的には、各方向において異なるアドレス指定モードを有し、IBリンクの各側のポートに対する転送モードは、リンクのそ

10

20

30

40

50

の側の「転送ドメイン」に対応することになる。たとえば、リンクの一方の側がコアファブリックを表し、リンクの他方の側がリーフサブネットを表す場合、コアファブリックからリーフサブネットに転送されているパケットは、図13の右側に示すようにLRH/DLIDベースのアドレス指定モードを有することになり、リーフサブネットからコアファブリックに転送されているパケットは、図13の左側に示すようにGRH/ISRNベースのアドレス指定モードを有することになる。

【0151】

リンクの一方の側のみに位置する単一のゲートウェイポートの場合、その単一のゲートウェイポートは、リンクのリモート側の転送ドメインの一部として管理されることに注目されたい。例として、リンクの一方の側がコアファブリックを表し、他方の側がリーフサブネットを表し、リーフサブネット側にゲートウェイポートがない場合は、リンクのコアファブリック側のポートはリーフサブネットの一部でもある必要があり、さらに、その転送ロジックが、そのリーフサブネットにおけるサブネットマネージャによって構成される必要もあるだろう。これは、このサブネットマネージャが、次いで、コアファブリック内の最初のホップ転送を効果的に実現することを意味するので、2つの転送ドメイン間の不要な依存性は、転送ドメイン間の境界をリンクからシステムクロスバーに移動させ、ゲートウェイはコアファブリック側に位置することによって、効果的に除去され得る。上述のことが、他の方向におけるパケットの転送に対して、対応して、行われる。

【0152】

一方、ゲートウェイポートをリンクの両側に有する場合、転送ドメイン間の唯一の依存性は、パケットをゲートウェイポートを介してIBリンクに転送することが、パケットが正しいアドレス指定モードを有するかまたはそれを提供されるという保証とともに行われる、ということである。アドレス指定モードの変更が行われるときにパケットの出力ポートが既に判断されているので、アドレス指定モードの変更はこの点で転送/ルーティング判断に影響を与えない。

【0153】

受信側では、入力パケット処理および転送判断は、その転送ドメインにおける任意のスイッチポートに対するのと同様に、標準のサブネット内プロトコルに従って実行される。

【0154】

単一のシステムインスタンス内のクロスバーベースの境界の場合、同じクロスバー上の、すなわち同じシステムインスタンス内の)任意の仮想スイッチインスタンスに対するゲートウェイポートである任意の物理ポートは、また、同じシステムインスタンス内の別の仮想スイッチ内の仮想スイッチポートであってもよい。

【0155】

特殊なケースは、システムインスタンスをあるリーフサブネットまたはコアファブリックに接続する単一のポートを含む。技術的には、単一のポートは定義上、ローカルサブネット/転送ドメイン内の他のソースエンドポートとターゲットエンドポートとの間のパケット転送を行うことができるという意味において、わずかに単一の外部ポートを伴うスイッチは、スイッチ管理ポートへのアクセスを提供し、それによってスイッチ管理ポートをサブネット内のエンドポイントとして機能させるという点で、依然として有効な構成であるというIBTAの観点から、スイッチではない。本例示的实施形態では、このような単一のポートは、スイッチングについて依然として関連性があり、なぜならば、クロスバーベースの転送ドメイン境界を介する他のリーフサブネットおよび/またはコアファブリックへのアクセスを提供するからである。したがって、リモートリーフサブネットまたはコアファブリックの観点から、それはある数のゲートウェイポートに対するアクセスを提供するが、ローカル転送ドメイン内の仮想スイッチポートに対するアクセスは提供しない。

【0156】

上記に加えて、1つより多い物理ポートがシステムインスタンスを同じコアファブリックまたはリーフサブネットに接続する場合、このポートの集まりは、デフォルトで仮想スイッチインスタンスとして表されるであろうことに注目されたい。しかしながら、ファブ

10

20

30

40

50

リックトポロジの観点からは、この切換えられる接続性に対する必要性はないかもしれず、ポートの集まりを1つ以上の仮想スイッチとして表現することは、すべての個々のポートが、同一転送ドメインにおける他のポートへの接続性を許可せずに、他のリーフサブネットまたはコアファブリックにおけるゲートウェイポートへの接続性を与えるにすぎない場合を含んで、ファブリック定義ポリシーの問題になる。

【0157】

リーフサブネット仮想スイッチ

一実施形態によれば、「リーフサブネット仮想スイッチ」は、同じリーフサブネットに接続する単一のシステムインスタンス上の1つ以上の物理ポートの集まりをベースラインとして定義する。ファブリックローカルサブネット番号（FLSN）が使用されている場合、同じリーフサブネット仮想スイッチにおけるすべてのポートは、通常、同じFLSN値を表わすことになる。概念的には、同じリーフサブネット仮想スイッチ内の異なるポートは異なるFLSN値を表してもよく、各ポートは、すべて同じ物理サブネットを表すにもかかわらず、1つより多い値を表してもよいことに注目されたい。また、同じFLSN値が、異なる物理リーフサブネットにおけるリーフサブネット仮想スイッチポートに関連付けられてもよい。

10

【0158】

上記スキームに対する1つの動機は、経路情報の変更を暗示することなく、リーフサブネット間のVMのマイグレーションを可能にすることである。しかしながら、完全な経路情報を維持するために、DLIDを保存するか、またはDLIDについてそのような例外の場合に使用され得るマッピング機能を提供することができる。

20

【0159】

制限された数のマッピングリソース、およびポート当たり1つより多い制限された数のFLSN値で、さらなる例示的な実施形態による高度なスキームは、2段階のマイグレーションスキームを使用し、マイグレーションは、第1ラウンドでは、制限されたハードウェアリソースの一部を使用して、経路更新なしに行なわれ、一方で、同時の経路更新プロセスを提供し、関連するすべてのピアによって経路情報が更新され、元の経路／アドレスはもはや必要ではなくなると、制限されたハードウェアリソースが再び解放され得ようする。

【0160】

例示的な実施形態では、リーフサブネット仮想スイッチは、それ自体を、標準的なSMA属性を介して標準的なIBスイッチとして提示する。したがって、レガシーサブネットマネージャ（SM）は、ローカルサブネットの一部である任意の他の標準的IBレガシースイッチと同様に、システムベースのリーフサブネット仮想スイッチを発見および構成できる。

30

【0161】

さらに、例示的な実施形態では、リーフサブネット仮想スイッチを拡張して、ローカルファブリックの他の部分への接続性またはリモートファブリックへの接続性を提供するゲートウェイサブネットへの接続性を表してもよい。このような接続性は、1つ以上の「特殊な」ポートがリーフサブネット仮想スイッチに含まれ、たとえば、特殊なSMA属性を介する同じ仮想スイッチ上の追加のポート番号などとして直接確認できることを意味する。つまり、1つ以上のコアファブリックゲートウェイポートは1つ以上のコアファブリックへの接続性を提供でき、1つ以上の仮想ルータポートは、物理ルータポート、およびリーフサブネット仮想スイッチの範囲外のさまざまな他のポートへの接続性を提供できる。

40

【0162】

たとえば、仮想ルータポートは、標準的なSMA属性を介して間接的に確認され得るが、ルータポートとして識別される別のポートへの（仮想）リンクを有する別のスイッチポートによって表され得る。

【0163】

このようにして、レガシーサブネットマネージャ（SM）は、他のサブネットへのルー

50

タベースの接続性を発見し、利用することができる。しかしながら、システム認識サブネットワークマネージャ（SM）は特殊なSMA属性を利用して特殊なポートを直接発見することになるので、「レガシーSM」サポートの必要性は減少することを理解されたい。

【0164】

さらに例示的な実施形態に関して、1つ以上のリーフサブネットワークゲートウェイポートは、1つ以上の他のリーフサブネットワークへの接続性を提供することができる。

【0165】

リーフサブネットワーク仮想スイッチポート

例示的な実施形態によれば、リーフサブネットワーク仮想スイッチポートは、リーフサブネットワーク仮想スイッチインスタンスに属することができ、たとえばIBリンクに到着するパケットについて出力ポート番号を探索するためにLRH、DLID値を使用する。このリーフサブネットワーク仮想スイッチインスタンスに対して出力ポート番号が表すポートのタイプは、例示的な実施形態では、メンバーの物理ポートに利用可能な構成情報によって定義される。出力ポートは、同じリーフサブネットワーク仮想スイッチにおける別の仮想スイッチポート、コアファブリックゲートウェイポート、仮想ルータポート、または別のリーフサブネットワークへの直接接続性を提供するリーフサブネットワークゲートウェイポートを表してもよい。出力ポートは一般に物理ルータポートではない可能性があり、なぜならば、これは、新たなLRHが、パケットGRHに基づく動的な態様で構築されることになることを意味し、そして、これは、仮想ルータポートを介してしか行われ得ないからである。

【0166】

わずか2つのポート（すなわち、ローカルおよびリモートルータポート）を有する単純化されたゲートウェイサブネットワークの定義が与えられている場合、次のホップに使用するDLIDは、個定値に基づくか、またはローカル物理ルータポートにリモートルータポートのLID値を記録させることに基づくであろう単純化されたスキームを提供し、これを次ホップDLIDとして用いることを保証することが可能であることを理解されたい。

【0167】

リーフサブネットワーク仮想スイッチポートである物理ポートは、さらに、別のリーフサブネットワーク仮想スイッチまたはコアファブリック仮想スイッチの観点から、リーフサブネットワークゲートウェイポートとして動作してもよいことをさらに理解されたい。

【0168】

コアファブリック仮想スイッチ

一実施形態によれば、「コアファブリック仮想スイッチ」は、同じコアファブリックに接続する単一のシステムインスタンス上の1つ以上の物理ポートの集まりをベースラインとして定義する。

【0169】

例示的な実施形態では、コアファブリック仮想スイッチは、それ自体を、選択された特殊なSMA属性を介して特殊なIBスイッチとして提示する。したがって、レガシーサブネットワークマネージャ（SM）は、例示的な実施形態のシステムのコアファブリック仮想スイッチを発見および構成することはできないことになる。

【0170】

コアファブリック仮想スイッチを拡張して、ローカルファブリックの他の部分への接続性またはリモートファブリックへの接続性を提供するゲートウェイサブネットワークへの接続性を表してもよい。このような接続性は、1つ以上の他の「特殊な」ポートがコアファブリック仮想スイッチに含まれ、たとえば、特殊な選択されたSMA属性を介して同じ仮想スイッチ上の追加のポート番号として直接確認できることを意味し、1つ以上のリーフサブネットワークゲートウェイポートは1つ以上のリーフサブネットワークへの接続性を提供でき、1つ以上の仮想ルータポートは、物理ルータポート、およびリーフサブネットワーク仮想スイッチの範囲外のさまざまな他のポートへの接続性を提供でき、1つ以上のコアファブリックゲートウェイポートは、1つ以上の他のコアファブリックへの接続性を提供することができる。

【0171】

コアファブリック仮想スイッチポート

例示的な実施形態によれば、コアファブリック仮想スイッチインスタンスのコアファブリック仮想スイッチポートは、I Bリンクに到着するパケットについて出力ポート番号を探索するためにGRH、DGID、ISRN値を使用する。探索されたポート番号がこのコアファブリック仮想スイッチインスタンスに対して表す出力ポートのタイプは、メンバーの物理ポートに利用可能な構成情報によって定義される。出力ポートは、同じコアファブリック仮想スイッチにおける別のコアファブリック仮想スイッチポート、リーフサブネットゲートウェイポート、仮想ルータポート、または別のコアファブリックへの直接接続性を提供するコアファブリックゲートウェイポートを表してもよい。

【0172】

出力ポートは一般に物理ルータポートではない可能性があり、なぜならば、これは、新たなLRHが、パケットGRHに基づく動的な態様で構築されなければならないことを意味し、そして、これは、仮想ルータポートを介してしか行われ得ないからである。

【0173】

コアファブリック仮想スイッチポートである物理ポートは、さらに、別のリーフサブネット仮想スイッチまたはコアファブリック仮想スイッチの観点から、コアファブリックゲートウェイポートとして動作してもよいことを理解されたい。

【0174】

コアファブリックゲートウェイポート

システムクロスバーベースの仮想スイッチおよび転送ドメイン境界の場合、コアファブリックゲートウェイポートである物理ポートは、関連するコアファブリック接続性を表わすコアファブリック仮想スイッチの観点から、コアファブリック仮想スイッチポートとして動作してもよいことが理解されるべきである。したがって、物理スイッチポートの観点からは、パケット転送代替策は、このセクションで説明するゲートウェイポート関連の転送に加えて、上で「コアファブリック仮想スイッチポート」に関して説明したものを含む。

【0175】

例示的な実施形態によれば、コアファブリックゲートウェイポートは、それらの仮想コアファブリックスイッチが他のコアファブリックインスタンス／転送ドメインも表すなど、同じシステムインスタンスにおける他のコアファブリック仮想スイッチに属する1つ以上のリーフサブネット仮想スイッチポートまたはコアファブリック仮想スイッチポートに対するコアファブリックインスタンスに対する入力またはそれからの出力を処理するポートを表す。

【0176】

システムクロスバーを介してパケットを受信するコアファブリックゲートウェイポートは、一般に、いかなるアドレス指定モードも変更できず、代わりに、受信されるパケットが、それを既に転送したシステム入口ポートからの正しいアドレス指定で設定されていることに依存し得る。

【0177】

例示的な実施形態では、物理ポートは複数の役割を有し得、特定の役割が入口ポートからパケットを転送する時点で判断されるので、一部のハードウェア設計上の制約により、この種のパイプライン化がより効率的になるのであれば、一般に、アドレス情報を出力ポイントで更新する必要はないことに注目されたい。

【0178】

リンクベースの転送ドメイン境界の場合、コアファブリックゲートウェイポートは、ローカルシステムインスタンスが表すリーフサブネットの転送ドメインの一部である。したがって、I Bリンクに到着するパケットについての転送モードは、「LRH／DLID」転送モードである。

【0179】

システムクロスバーベースの転送ドメイン境界の場合、コアファブリックゲートウェイ

10

20

30

40

50

ポートは、リモートコアファブリックインスタンスによって定義された転送ドメインの一部である。したがって、I Bリンク上に到着するパケットについての転送モードは、「GRH/I SRN」転送モードである。

【0180】

探索されるポート番号が関連する転送ドメインに対して表わす出力ポートのタイプ（すなわち、ゲートウェイポートが表すコアファブリックインスタンス）は、物理ポートに利用可能な構成情報（すなわち、物理ポートの仮想スイッチポートのパーソナリティ（もしあれば）について略述されるのと同じ情報）によって定義される。

【0181】

出力ポートは、リーフサブネットゲートウェイポート、または2つのコアファブリックが単一のシステムクロスバーを介して接続されている場合にそうであるかもしれないような別のコアファブリックゲートウェイポートを表してもよく、また、仮想ルータポートを表してもよい。

【0182】

図14は、一実施形態による、GRH/I SRNアドレス指定モードフォーマットからLRH/DLIDベースの転送フォーマットへのパケットヘッダの変更を示す図である。パケットがコアファブリックゲートウェイポートからシステムクロスバー経由でリーフサブネットゲートウェイポートに移動するとき、この遷移はGRH/I SRNアドレス指定モードからLRH/DLIDベースの転送への変更を表わす。これは、図に示すように、新たなパケット1400のヘッダ内に、パケットGRH1111から抽出されたDLID 1143（すなわち、システム特有のGRH、DLID、リモートDLIDフィールド）を有する新たなLRHが構築されることを意味する。一般に、この転送は、パケットFLSN値を、リーフサブネットゲートウェイポートに関連付けられたFLSN値と関連付けることとは独立して、十分に定義されることに注目されたい。つまり、アドレス指定モードの変更を達成するために必要な唯一の情報は、出口ポートが入口ポートの観点からリーフサブネットゲートウェイポートとして定義されているということであり、これはどのようなFLSN値が関係するかとは無関係である。この観察に基づいて、パケットFLSN値とポートFLSN値との相関は、本質的に、ルーティングが正しく設定されていることの「アサート」である。

【0183】

この問題は、エンドノードに関連付けられるLIDの問題と同様である。ローカルLIDは、パケットを送信するときに正しいSLIDを生成するために必要であるが、サブネットがLID割当てに従って正しくルーティングされる限り、一旦着信パケットが宛先エンドノードに到着すれば、パケットでDLIDを処理する必要は本質的にない。スイッチ転送テーブルが、初期または新たなLID値をエンドノードに割り当てることに正しく正しいシーケンスで更新される限り、エンドノードは、受信したDLIDが割り当てられたLIDと競合しているケースを決して見ない。

【0184】

一方、エンドノードにおいてDLIDチェックを任意にすることは、ファブリックは、原則として、単一のエンドノードへの転送のために、任意の数のLIDおよび関連付けられるルーティングを用いることができることを意味する。宛先がリーフサブネットであるこの場合、FLSN値チェックを任意にすることは、異なるリーフサブネット間のVMマイグレーションをより柔軟に実現できることを意味する。

【0185】

図15は、一実施形態による、LRH/DLIDアドレス指定モードからGRH/I SRNアドレス指定モードにパケット転送を変更するための特殊スイッチポート境界を定義するメカニズムを提供するスイッチの図である。パケットがリーフサブネットゲートウェイポートからシステムクロスバー経由でコアファブリックゲートウェイポートに移動するとき、この遷移はLRH/DLIDアドレス指定モードからGRH/I SRNアドレス指定モードへの変更を表わす。これは、図に示すように、I SRNベースの転送を識別する

固定されたDLIDを有する新たなLRHが構築されることを意味する。これに関して、スイッチ1500は、LRH/DLIDアドレス指定モードからGRH/ISRNアドレス指定モードにパケット転送を変更するための特殊スイッチポート境界を定義するメカニズムを提供する。中間スイッチファブリックへの進入をサポートする実施形態によれば、あるメカニズムが、ファブリックの選択されたスイッチ1500に設けられ、それによって境界が画定され得る。一実施形態では、「特殊スイッチポート」(SSP)1510が定義され、SSPはローカルサブネットの「終わり」を表す。一実施形態では、パケットは、元の/通常のローカルルートヘッダ1112におけるDLID1122(図11)を使用して、LFT1200(図12)を介して仮想サブネットの仮想エンドポートを表すSSP1510を指す。SSPには、この点でこのサブネットにおける「エンドポート」を表すことを示す属性が与えられる。特殊属性は、転送のモードにおける変更を示し、次の転送レベルでは、ローカルルートヘッダ1112におけるDLID1122を使用する代わりに、GRH/L3アドレスの64ビットのサブネットプレフィックス番号部分からの指定/専用ビットフィールド1131を使用する。本質的に、元のDLID1122は、指定/専用のビットフィールドまたはその一部で置き換えられて、例示的な実施形態による、ISRNベースの転送で使用するための新たなDLID1110'を形成する。

10

【0186】

一実施形態では、スイッチ1500は、LIDベースのフィルタリング技術を使用することができる。この実施形態では、スイッチ1500は、SSP1510をターゲットとするパケットを識別することができる受信(Rx)フィルタ1520を含むことができる。したがって、スイッチ1500は、SSP1510をターゲットとするデータフロートラフィックを、通常/レガシー転送モダリティポート1530をターゲットとするデータフロートラフィックから分離することができる。たとえば、Rxフィルタ1520は、(たとえば、DLIDベースのフィルタリングを使用して)サービスDLIDに基づいて混合データフロートラフィックを分離することができる。以下は、例示的なパケット転送ロジックDLIDテーブルである。

20

【0187】

DLID = 0 x F

DLID = 0 x F F

着信パケットが一致するDLIDを有する場合(たとえば、0 x Fまたは0 x F F)、Rxフィルタ1520は、パケットをSSP1510に向けることができる。一方、着信パケットが一致するDLIDを有さない(すなわち、0 x Fおよび0 x F F以外のDLIDを伴う)場合、Rxフィルタ1520は、着信パケットを通常/レガシー転送モダリティポート1530に向けることができ、それは、IBプロトコルエンジンを使用してIBパケットを標準的なIBプロトコルに従って処理できる。

30

【0188】

例示的な実施形態によれば、パケットがコアファブリックゲートウェイポートと別の(すなわち異なるコアファブリックインスタンスに接続される)コアファブリックゲートウェイポートとの間で移動する場合、同じGRH、DGID、ISRN値が依然として転送に使用される。したがって、管理ソフトウェアは、1つのコアファブリックインスタンスから別のコアファブリックインスタンスに交差するパケットトラフィックが、2つ以上のコアファブリックの両方で一貫して処理されるGRH、DGID、ISRN値を表していることを保証し、なぜならば、デフォルトでは、各コアファブリックインスタンスはそれ自体のISRN値空間を表わすからである。

40

【0189】

パケットがコアファブリックゲートウェイポートから仮想ルータポートに移動するとき、次ステップ転送スキームは仮想ルータポートによって実現されるポリシーの機能である。

【0190】

出力ポートは一般に物理ルータポートではない可能性があり、なぜならば、これは、新

50

たなLRHが、パケットGRHに基づく動的な態様で構築されることになることを意味し、そして、これは、仮想ルータポートを介して行われるからである。

【0191】

リーフサブネットゲートウェイポート

システムクロスバーベースの転送ドメイン境界の場合、リーフサブネットゲートウェイポートである物理ポートは、関連するリーフサブネット接続性を表わすリーフサブネット仮想スイッチの観点から、リーフサブネット仮想スイッチポートとして動作してもよい。したがって、物理スイッチポートの観点からは、パケット転送代替策は、このセクションで説明するゲートウェイポート関連の転送に加えて、上で「リーフサブネット仮想スイッチポート」に関して説明したものを含む。

10

【0192】

例示的な実施形態によれば、リーフサブネットゲートウェイポートは、同じシステムインスタンスにおける他のリーフサブネット仮想スイッチに属する1つ以上のコアファブリックゲートウェイポートおよび／または1つ以上のリーフサブネット仮想スイッチポートに対するリーフサブネットインスタンスへの入力またはそれからの出力を処理するポートを表す。これに関して、システムクロスバーを介してパケットを受信するリーフサブネットゲートウェイポートは、一般に、いかなるアドレス指定モードも変更せず、代わりに、受信されるパケットが、それを転送したシステム入口ポートからの正しいアドレス指定で既に設定されていることに依存する。

20

【0193】

リンクベースの転送ドメイン境界の場合、リーフサブネットゲートウェイポートは、特定のシステムインスタンスが表すコアファブリックの転送ドメインの一部である。したがって、IBリンク上に到着するパケットについての転送モードは、「GRH／ISRN」転送モードである。

【0194】

システムクロスバーベースの転送ドメイン境界の場合、リーフサブネットゲートウェイポートは、リモートリーフサブネットインスタンスによって定義された転送ドメインの一部である。したがって、IBリンクに到着するパケットについての転送モードは、「LRH／DLID」転送モードである。

【0195】

出力ポートは、コアファブリックゲートウェイポート、または2つのリーフサブネットが単一のシステムクロスバーを介して接続されているときなどのような別のリーフサブネットゲートウェイポートを表してもよく、また、仮想ルータポートを表してもよい。

30

【0196】

例示的な実施形態によれば、パケットがリーフサブネットゲートウェイポートからシステムクロスバー経由でコアファブリックゲートウェイポートに移動するとき、この遷移はLRH／DLIDベースの転送モードからGRH／ISRNベースの転送モードへの変更を表わす。

【0197】

例示的な実施形態によれば、パケットがコアファブリックゲートウェイポートからシステムクロスバー経由でリーフサブネットゲートウェイポートに移動するとき、この遷移はGRH／ISRNアドレス指定モードからLRH／DLIDアドレス指定モードへの変更を表わす。

40

【0198】

例示的な実施形態によれば、パケットがリーフサブネットゲートウェイポートと別の（すなわち異なるリーフサブネットに接続される）リーフサブネットゲートウェイポートとの間で移動する場合、元のパケットLRH、DLID値は、出力ポートを探索するために用いられるが、以下に示すようにLRHが次のホップのためにどのように構築されるかについて2つの代替策が存在する。

【0199】

50

第1の代替策として、パケットにGRHがなく、宛先ポートが「修正無LID転送」を許可する場合、パケットは同じLRH、DLIDで転送される。したがって、この場合、新たなLRH、DLIDの生成を伴わずに1つのリーフサブネットインスタンスから別のリーフサブネットインスタンスに交差するパケットトラフィックが、両方のリーフサブネットで一貫して処理されるLRH、DLID値を表していることを保証することは、管理ソフトウェア次第であり、なぜならば、デフォルトでは、各リーフサブネットインスタンスはそれ自体のLID値空間を表わすからである。

【0200】

第2の代替策として、パケットがISRNベースの転送を表すGRHを有する場合、出力ポートとのFLSN一致がある場合には、パケットは転送される。上記のコアファブリックゲートウェイポートの議論で指摘したのと同様に、ルーティングおよび転送テーブルの設定が正しく行われている限り、この場合でも一般的にFLSNマッチングを実行する必要はないことに注目されたい。

【0201】

さらに、一実施形態によれば、異なる転送ドメイン間でのパケットの転送がFLSN割当てと競合していないことを実際に検証するために、FLSNチェックの使用を選択的に有効にすることができる。この場合にも、FLSN値チェックを任意にすることは、VM移行に関して以下に述べるように、異なるリーフサブネット間でのVMマイグレーションをより柔軟に実現できることを意味する。

【0202】

さらなる代替策として、パケットがISRNベースの転送を表さないGRHを有するが、宛先ポートが「修正無LID転送」を許可する場合、パケットは同じLRH、DLIDで転送される。したがって、この場合も、管理ソフトウェアが、新たなLRH、DLIDの生成を伴わずに1つのリーフサブネットインスタンスから別のリーフサブネットインスタンスに交差するパケットトラフィックが、両方のリーフサブネットで一貫して処理されるLRH、DLID値を表していることを保証し、なぜならば、デフォルトでは、各リーフサブネットインスタンスはそれ自体のLID値空間を表わすからである。

【0203】

パケットがリーフサブネットゲートウェイポートから仮想ルータポートに移動するとき、次ステップ転送スキームは仮想ルータポートによって実現されるポリシーの機能である。出力ポートは一般に物理ルータポートではない可能性があり、なぜならば、これは、新たなLRHが、パケットGRHに基づく動的な態様で構築されることになることを意味し、そして、これは、仮想ルータポートを介して行われるからである。

【0204】

組み合わせられたコアファブリックおよびリーフサブネット仮想スイッチ

例示的な実施形態によれば、組み合わせられたコアファブリックおよびリーフサブネット仮想スイッチ構成は、同じセットのポートがコアファブリックおよびリーフサブネット仮想スイッチの両方の特性を同時に提供することを可能にする。この構成の1つの動機は、非GRHベースのデータおよび制御トラフィックが、コアファブリックにおける1つ以上のシステムインスタンスと直接接続されたノード間で生ずることを可能にすることである。コアファブリック内の転送オプションについては上で説明した。

【0205】

この実施形態の実現の観点からは、物理的なシステムポートが、このポートの役割に対して出力ポートとして使用できる他のポートの役割を定義する構成情報を有する限り、異なる仮想スイッチインスタンス間の物理ポートメンバーシップの部分的な重複を許可することとは対照的に、特定のポートのセットを、組み合わせられたコアファブリックおよびリーフサブネット仮想スイッチとして定義する制約を実施する必要がある。

【0206】

組み合わせられたコアファブリックおよびリーフサブネット仮想スイッチポート

例示的な実施形態による組み合わせられたコアファブリックおよびリーフサブネット仮想

10

20

30

40

50

スイッチポートは、GRH／ISRN転送モードとLRH／DLID転送モードとの両方をサポートする。

【0207】

GRH／ISRNベースの転送を実行するために、LRH、DLIDは、ISRNベースの経路におけるすべてのポートについて共通の特定の事前定義された値である。したがって、このDLID値は、システムチップレベルの定数であってもよく、または好ましくは物理ポートレジスタごとに定義され得る。

【0208】

結合ポートは、論理線形転送テーブル（LFT）容量に関して、構成可能な数のDLIDエントリと、別の構成可能なISRN値範囲とをサポートすることができる。

10

【0209】

一般に、値のセットが任意のポートでサポートできるように、同じ仮想スイッチにおけるすべてのポートが同じDLIDおよび／またはISRN LFT容量を有することが好ましいことに注目されたい。しかしながら、これは単なる構成ポリシーであり、例示的な実施形態の実現の制約ではない。

【0210】

同様に、組み合わされた仮想スイッチインスタンスタイプおよび単一の仮想スイッチインスタンスタイプを含む各仮想スイッチインスタンスは、例示的な実施形態では、ポートごとの線形転送テーブル（LFT）または論理的に共通のLFTをすべての仮想スイッチポートに対して論理的に表すように選択的に定義される。しかしながら、これはファームウェアオプションであり、なぜならば、物理的なLFT実装は、例示的な実施形態では、ポートごとまたはデュアルポートのセットごとに分散されるからである。このファームウェアオプションの1つの理由は、明らかに、IB規格に準拠したIBスイッチモードをサポートするとともに、サブネットマネージャならびに分散されたLFTをサポートしないルーティングエンジンおよび初期化エンジンを使用することである。

20

【0211】**仮想ルータポート**

仮想ルータポートは、例示的な実施形態では、物理ポートの範囲外のシステムポート番号によって表される。したがって、仮想ルータポートには直接関連付けられる物理リンクはない。どの仮想スイッチポートでも仮想ルータポートを探索でき、仮想ルータポートは次いで物理出力ポートおよび次ホップLRHを定義するためにパケットGRH上で構成可能なルーティング機能を使用する。

30

【0212】

例示的な実施形態におけるルーティング機能は、同じローカルファブリックにおけるエンドポイント間の階層的転送と、独立したローカルファブリックインスタンス間のジェネリックな転送との両方を含むことに注目されたい。

【0213】

GRHのないパケットが仮想ルータポートに送信される場合、好ましくは、パケットはドロップされるか、または新たなLRHおよび物理的な出力ポートが構成可能な機能に基づいて生成される。

40

【0214】

各仮想スイッチポートは、パケットからのDLIDまたはISRN値がポートによってサポートされるLFT範囲外であるためにLFTベースの探索が不可能な場合に中間出力ポートとして使用される関連付けられるデフォルト仮想ルータポートを有する。この場合、仮想ルータポートは、選択的に、パケットをIB規格に従ってデフォルトとしてドロップするか、または物理出力ポートおよび次ホップLRHを定義するためにパケットヘッダ上で構成可能な機能を実行するように動作可能である。

【0215】**物理ルータポート**

物理ルータポートは、例示的な実施形態では、ローカルファブリックインスタンスへの

50

進入またはローカルファブリックインスタンスから外に出ることを処理するポートである。物理ルータポートは、リンク側から I B 規格に準拠したルータポートとして表される。

【0216】

例示的な実施形態では、物理ルータポートは、リンク側から到着するパケットに使用される仮想ルータポートと同じパケット転送ロジックに関連付けられる。

【0217】

例示的な実施形態では、物理ルータポートは、パケットのパケット G R H 上で選択された機能を実行し、次いで、物理ルータポートのリンク上で出るよう意図される新たな L R H を生成した仮想ルータポートまたは別の物理ルータポートによって探索されるかまたは他の態様で参照される。

10

【0218】

単純な L R H. D L I D 挿入は、最小限のゲートウェイサブネットにおけるリモートルータポートに関連付けられるリモート D L I D 値に基づいて与えられる。

【0219】

サブネット間ルート番号 (I S R N)

例示的な実施形態によれば、サブネット間ルート番号 (I S R N) は、コアファブリックインスタンス内において L F T エントリを索引付けするために使用されるパケット G R H. D G I D. サブネットプレフィックスの可変部分によって定義される値である。 I S R N 値空間はコアファブリックインスタンスごとであり、ファブリックグローバルマネージャ (F G M) は I S R N 値の使用を調整する。

20

【0220】

例示的な実施形態では、 I S R N 値を使用して転送されるべきパケットは、 I S D N 値が L F T ベースの転送に使用されるべきであることを受信側ポートに示す特殊な予約済み L R H. D L I D 値を有する。

【0221】

一実施形態では、システムポートが「複数のパーソナリティ」を有する能力は、パケットが、使用するべき特定のタイプの転送メカニズムに関する情報を、ポートごとの構成情報としてこれ (のみ) を有することとは対照的に、含むべきであることを意味する。

【0222】

ファブリックローカルサブネット番号 (F L S N)

30

例示的な実施形態によれば、ファブリックローカルサブネット番号 (F L S N) は、ローカルファブリック内で宛先サブネットを識別する値である。すべての F L S N 番号は、ファブリックグローバルマネージャ (F G M) によって割り当てられ、管理されることが好ましい。 I S R N 値に基づいて転送されるパケットは、宛先 F L S N 値も有する。

【0223】

F L S N 値は、論理ポートの役割を担うリーフサブネットゲートウェイポートに関連付けられ、そのようなポートに転送されるパケットは、宛先サブネットに到達したことを判断するために、パケット F L S N 番号がポート F L S N 番号と照合されることになる。

【0224】

一実施形態では、システムポートが「複数のパーソナリティ」を有する能力は、ポートごとの構成情報に単に依存することとは対照的に、パケットおよびポート情報の両方の組み合わせに基づいてパケットがそれらの宛先サブネットに到達したものとして認められることを意味する。

40

【0225】

ファブリックグローバルマネージャ (F G M)

例示的な実施形態によれば、ファブリックグローバルマネージャ (F G M) は、完全なローカルファブリックインスタンスの発見および構成を実行することができる「スーパーサブネットマネージャ」である。 F G M は、個々のリーフサブネットにおける「 F G M 認識」サブネットマネージャと協働する。

【0226】

50

各リーフサブネットサブネットマネージャは、F G Mとは独立してそれ自体のローカルリーフサブネットを管理することができるが、ローカルファブリックにおける任意のコアファブリックまたは他のリーフサブネットを含む任意のサブネット間トラフィックを容易にするために、F G Mに依存することになることに注目されたい。

【0227】

識別サブネット番号 (I S N)

例示的な実施形態によれば、識別サブネット番号 (I S N) は、グローバルまたはサイトに特定の一意性を有してもよい、管理上割り当てられたサブネット番号である。I S N番号は、異なるローカルファブリックインスタンス間においてサブネット間トラフィックに選択的に使用されるG I D値を形成するために、G U I Dとともに使用される。

10

【0228】

ハードウェアG U I D (H W G U I D)

例示的な実施形態によれば、ハードウェアG U I Dは、ハードウェアリソースの64ビットのグローバルに一意の識別情報である。例示的な実施形態における各物理システムインスタンスは、ベースH W G U I D値を有する。各システムベースの仮想スイッチインスタンスは、予め定義された方法でベースH W G U I Dから導出されるH W G U I Dを有する。各物理H C Aは、ベースH W G U I Dを有することになる。各仮想H C Aポートは、予め定義された方法でベースH W G U I Dから導出されるH W G U I Dを有することになる。

20

【0229】

仮想G U I D (v G U I D)

例示的な実施形態によれば、仮想G U I D (v G U I D) は、グローバルに一意またはローカルファブリック内で一意である64ビットのG U I D値である。仮想G U I Dは、ローカルファブリックに存在する任意のH W G U I Dに対して一意であることも保証される。

【0230】

仮想G U I Dは、典型的には仮想マシンインスタンスに関連付けられ、同じv G U I D値 (V Mが使用している各v H C Aポートに対して1つ) は、通常、物理サーバとv H C Aインスタンスとの間において複数のマイグレーション含んでもよいV Mライフサイクル全体にわたるV Mに関連付けられることになる。

30

【0231】

特定のv H C AインスタンスはそのH W G U I Dによって識別されるが、V Mがそれを使用しているときはいつでも、v H C Aポートは関連するV M v G U I D値を使用してアドレス指定することもできることに注目されたい。

【0232】

パケットヘッダフォーマットおよび規約：

パケットローカルおよびグローバルルートヘッダ (L R H / G R H) フォーマットおよび規約のための基準について以下に説明する。第1に、本明細書の例示的な実施形態で使用されるローカルルートヘッダ (L R H) は、インフィニバンド (I B) 規格によって定義される通りである。同様に、本明細書の例示的な実施形態で使用されるグローバルルートヘッダ (G R H) は、インフィニバンド (I B) 規格によって定義される通りであるが、2つの追加／例外のセットを有する：1) G R H.DGID.SUBNET_PREFIXはタイプ「I S R N」を有することができ、2) G R H.DGID.SUBNET_PREFIXはタイプ「LeafSwitchHierarchy (リーフスイッチ階層)」を有することができる。

40

【0233】

G R H.DGID.SUBNET_PREFIXがタイプ「I S R N」である場合、G R H.DGID.SUBNET_PREFIXは、例示的な実施形態では、以下のフォーマットを有する (I S R Nタイプに加えて：

(宛先) ファブリックローカルサブネット番号

サブネット間ルート番号 (すなわち、ソースリーフサブネットにおいて最初のL R H.D L I D値によって定義されるコアファブリックインスタンスにおける)

50

D L I D（ファブリックローカルサブネット番号で定義される宛先リーフサブネットにおける）。

【0234】

GRH.DGID.SUBNET_PREFIXがタイプ「LeafSwitchHierarchy」である場合、GRH.DGID.SUBNET_PREFIXは、例示的な実施形態では、以下のフォーマットを有する（LeafSwitchHierarchyタイプに加えて：

（宛先）ファブリックローカルサブネット番号

（宛先）リーフスイッチ番号

すべての特殊なサブフィールドは、ランタイムで定義されるサイズおよびオフセットを有する。（すなわち、サブフィールド抽出を実行するために専用TCAMエントリを使用することによって可能にされる）。

【0235】

ローカルファブリック全体にわたるSL値およびVL値のマッピングならびに選択：

本明細書の例示的な実施形態によれば、単一のリーフサブネット内において、LRHのみを有し、GRHを有さないパケットは、レガシーのSLからVLへのマッピングを使用していることになる。例示的な実施形態は、予約ビットおよび新たなヘッダバージョン番号を使用することによって、ローカルLIDおよび／またはSL／VL値空間を所望または必要に応じて選択的に拡張してもよい。LFTによって探索可能なポート番号の数を拡張することは、より多くのそのような「エイリアス」ポート番号が、関連付けられる物理ポートについて他のSL／VLマッピングを表すことを可能にする。（すなわち、SL／VL処理と同じ効果であるが、予約されたLRHビットの所有権を主張できる使用に頼るのではなく、LFTメモリを消費する。）これは、トラフィック分離を達成するために異なるSL値ではなく複数のDLID値を使用することを意味する。

【0236】

LRHおよびGRH無しの両方を伴う単一のリーフサブネットパケット内での値のマッピングならびに選択は、例示的な実施形態では、デフォルトで、レガシーのSLからVLへのマッピングを使用することになる。GRHフィールドは、好ましくは、コアファブリックにおけるのと同じ態様で使用される。

【0237】

リーフサブネット間およびコアファブリック内における値のマッピングならびに選択では、LRHのみを有するパケットは、好ましくは、単一のリーフサブネットにおけるように処理される。

【0238】

リーフサブネット間およびコアファブリック内における値のマッピングならびに選択では、LRHおよびGRHの両方を有するパケットは、以下のスキームに基づいてSL選択を有することになる：各トラフィッククラス値は、可能なSL値のセットを表す；フローラベル値のハッシュは、トラフィッククラスによって定義された利用可能なSL値の中から選択し、以下の「ゲートウェイサブネットを介したSL値およびVL値のマッピングならびに選択」で定義されているように、「積極的な」クレジット待機ポリシーに対してサポートが与えられる。

【0239】

ゲートウェイサブネットを介したSL値およびVL値のマッピングならびに選択：

デフォルトでは、ローカルファブリック内で使用するために定義されたものと同じGRHベースのスキームが、例示的な実施形態によってサポートされる。基本的なマッピングに加えて、たとえば、リモートファブリックにおいてゲートウェイサブネット／ルータポートにパケットを送信するルータポートなどの出口ルータポートは、すべてのそのようなパケットまたはサブセットを、定義された基準に基づいて、ローカルファブリックに輻輳ツリーが蓄積されないようにパケットを「早期に」ドロップする、より「積極的な」クレジット待機ポリシーを有するとして分類してもよい。

【0240】

ローカルファブリック全体にわたる P_Key 値のマッピングおよび選択：

例示的な実施形態における各転送ドメインは、ローカルファブリックレベルでの調整なしにリーフサブネットの内部のトラフィックに使用できる P_Key 値の範囲を有する。つまり、すべての転送ドメインによって、同じプライベート範囲を使用できる。ローカル範囲外の P_Key 値は、FGM を介して割当てられる。FGM は、ゲートウェイポートが、関連する転送ドメインに転送されることを許される P_Key 値を持つパケットの転送のみを受け入れることを保証する。

【0241】

さらなる例では、FGM は、異なる転送ドメイン間のような特定のゲートウェイポートでの P_Key 値のマッピングを含む特殊な経路を定義することができる。一般に、そのようなマッピングは、パケット不変 CRC (ICRC) が再計算されることを意味する。HCA などのエンドノードからのサポートを含む代替スキームとして、あるスキームが提供され、通常のパケット ICRC は、0 のパケット P_Key 値に基づいて計算されるが、加えて、GRH において、ICRC に含まれない予約ビットフィールドを有し、しかし、それは、現在の P_Key 値に対する CRC パリティ値を含む。

【0242】

ゲートウェイサブネットを介する P_Key 値のマッピングおよび選択：

本明細書で説明される例示的な実施形態は、一般に、異なるローカルファブリックインスタンス間の調整された P_Key 値割り当てについての仮定をなさない。ネイティブ IB 通信は、この場合、関与するローカルファブリックインスタンス間の動的に交渉される P_Key 値に基づくか、または上記のローカルファブリックの場合について概説された P_Key マッピングスキームに基づく。

【0243】

ローカルファブリック全体にわたるマルチキャストパケットの転送：

例示的な実施形態では、各転送ドメインは、各転送ドメインに対してプライベートであるとみなされる MLID 値の範囲を有し、すべての転送ドメインが同じプライベート範囲を使用することができる。プライベート MLID にマッピングされるマルチキャストグループは、ローカル転送ドメイン内でのみ転送できる。FGM は、プライベート MLID 転送が転送ドメイン境界を超えて行われないようにし、完全なローカルファブリックインスタンス全体にわたるマルチキャスト転送を実現するために使用できる共有 MLID 値のセットを調整する。完全なローカルファブリック全体にわたって共有 MLID 値を介して使用される MGID 値は、FGM によって割当てられる。そのような MGID 値は、それを FGM 制御される MGID であるよう定義する選択されたビットフィールドを有する。各転送ドメイン内におけるローカル SM は、FGM 制御される MGID の競合する使用がないことを保証することを担う。

【0244】

さらなる例では、仮想ルータポートが使用される。この例では、1 つの転送ドメインにおいて特定の MGID および MLID 値を有するマルチキャストパケットは、別の転送ドメインにおいては、同じ MGID 値で転送できるが、異なる MLID 値を有する。

【0245】

ゲートウェイサブネットを介したマルチキャストパケットの転送：

本明細書で説明される例示的な実施形態は、一般に、異なるローカルファブリックインスタンス間の調整された MGID 値割り当てについての仮定をなさない。純粋なハードウェアベースの転送を伴うネイティブ IB マルチキャスト通信は、この場合、関与するローカルファブリックインスタンス間の動的に交渉される MGID 値に基づく。

【0246】

例示的な実施形態では、デフォルトでは、異なるローカルファブリックインスタンス間のマルチキャスト転送は、ソフトウェアベースのプロキシサービスを構成することに基づき、ソフトウェアベースのプロキシサービスは、1 つ以上の物理ルータポートに関連して動作し、ローカルファブリックまたは物理ルータポートからマルチキャストパケットを受

10

20

30

40

50

信し、ペイロードを新たなマルチキャストパケットにおいて反対方向に転送する。プロキシは、好ましくは、デフォルトでは、ローカルファブリック側において通常のマルチキャストグループメンバーであり、たとえばゲートウェイサブネットを介するなどしてリモートファブリックから到着するすべてのマルチキャストパケットを傍受するために、物理ルータポートとの特殊な関係を有する。

【0247】

上記に加えて、リモートファブリックからのマルチキャストパケットがローカルファブリックにおけるマルチキャスト転送を妨げない態様でプロキシに転送されることを確実にするよう、あるメカニズムが選択的に提供される。これは、ローカルファブリックにおいて専用MLIDを使用する「デフォルトMC宛先」のスキームまたは概念を使用し得る。これは、受信機が任意の数のMGIDを受信し、プロキシインスタンスに関連付けられるQPに転送できる限り、実行可能である。これは、多数のパーティションも含んでもよい。この機能性が関連するHCAに対して保証することが困難である場合、本明細書の別の実施形態は、EnetパケットがEnetGWポートで受信されたときにトンネリングされる方法と同様の方法で、システムがIBMCパケットをIBユニキャストパケットでトンネリングするように機能することである。したがって、プロキシ実装は、着信リモートMCパケットを、ローカルファブリック上でどのようにトンネリングするかとは無関係に、解釈して処理することができる。

【0248】

任意の数のパーティションを処理するために、ローカルファブリックからリモートファブリックへの方向に1つより多いSWプロキシが必要とされてもよいことに注目されたい。さらに、物理ルータポートについては、双方向のトンネリングが実現されることに注目されたい。また、プロキシ実装が両方のファブリックに存在する場合、専用のハードウェアサポートを必要とせずプロキシ実装を協働させることによって、トンネリングが実現される。これに関連して、トンネリングスキームは、元のIBヘッダをトンネリングパケットにIBペイロードとして含めなければならないため、最大IBMTUをトンネリングされるMCパケットとして転送することができないことを意味する。プロキシは、好ましくは、同じMGID値が異なるローカルファブリックにおける2つの異なる論理マルチキャストグループに対して使用される場合を処理することができる。

【0249】

ローカルファブリックの初期発見および構成：

例示的な実施形態によれば、デフォルトでは、各システムインスタンスは、単一の従来のインフィニバンド（IB）スイッチとして動作する。このデフォルト構成では、既存のIBサブネットマネージャまたはSMPベースの診断および監視ツールは、完全な接続されたファブリックを単一のIBサブネットとして「見る」かまたは他の態様で学習することができる。

【0250】

システムインスタンスを複数の仮想スイッチに分割し、ポートタイプおよび関連付けられる構成情報を定義することは、例えば、通常のSMA動作の一部ではない「バンド外」構成ポリシーを表す。しかしながら、すべてまたはほとんどのバンド外構成操作は、「純粋な」バンド外構成の代替として、ベンダー固有のSMP操作を介して実装できる。例示的な実施形態では、各システムインスタンスは、任意のIBリンク状態から独立して利用可能であるEネットベースの管理インターフェイスを有し、IPoIB/EoIBをサポートしもするので、「バンド外」構成は原則的に任意のインターフェイスを介したIPベースのアクセスを表すことに注目されたい。

【0251】

さらに、ローカル管理プロセッサに元来関連付けられるPCI-Eのような純粋な「ローカル」管理インターフェイスの概念はないことに注目されたい。それでも、セキュリティに対する「鶏と卵」構成問題を解決するために、Enetインターフェイスは、例示の実施形態のシステムの一部である、例えば特定の独立した物理スイッチシャーシのような

10

20

30

40

50

物理システムは、好ましくは、ポートベースのVLANまたは他のアクセス制御規則を介してEnetアクセスを確保している、という仮定を伴う構成ポリシーの最初のソースを表わしてもよい。この知識は、好ましくは、「先験的」に定義された生産時間であり、システムインスタンスについての永続的な構成情報である。

【0252】

シャーシ内部管理プロセッサを表すHCAポートが存在し、ハードウェア設計によって定義されるような特定のシステムインスタンス上の特定のポートに接続される構成では、例示的な実施形態によるシステムインスタンスも、この接続性を有するローカルシステムポートに関する先験的情報を有するべきである。

【0253】

最初にEnetインターフェイスまたは管理ホストチャネルアダプタ（HCA）を介して供給される構成情報に基づいて、例示的な実施形態のシステムのファームウェアは、他のインターフェイスおよび／または管理Enetインターフェイス上の他のソースからの構成操作も同様に認証することができる。このような構成情報は、例示的な実施形態のファームウェアによって确实かつ永続的に記憶される。

【0254】

さらなる例示的な実施形態によるファームウェアは、さらに、ファームウェアが初期展開時に他のインターフェイス／ソースからのアクセスを認証することを可能にする追加の「先験的な」永続構成情報にアクセスを有してもよい。これには、ベンダー固有の管理ソフトウェアなどのための公開鍵を含めることができる。加えて、タイプ、およびおそらくはこのベンダーインスタンスのハードウェアGUIDに関連するような特定のアイデンティティも認証するために使用される秘密鍵が必要または所望され得るように存在する可能性がある。

【0255】

例示的な実施形態によれば、ポート役割および仮想スイッチが構成されているとき、例えばサブネットマネージャ（SM）およびFGMインスタンスなどのバンド内管理コンポーネントは、レガシーSMPベースのサブネット発見方法および1つ以上のベンダー固有のSMP方法の両方を使用して、ローカルの物理および／または仮想ファブリックのそれらの関連部分を発見することができる。

【0256】

例示的な実施形態によれば、各リーフサブネットは、好ましくは、独立したSMインスタンスによって他のリーフサブネットと並行して発見され構成される。

【0257】

コアファブリックにおいては、1つ以上の「マスタ」FGMインスタンスは、一般に、コアファブリックインスタンスごとに1つのマスタFGMインスタンスと同時に動作してもよい。

【0258】

さらに、リーフサブネットにおけるサブネットマネージャは、それらのリーフサブネットの外部に、それらの転送ドメイン外であるSMPアクセス（読出しまたは書込み）を有さないことが好ましい。リーフサブネットにおけるレガシーサブネットマネージャは、リーフサブネット内における普通の（仮想）スイッチポートおよびそれらのローカル接続性を検出できるのみであることになる。リーフサブネットにおけるファブリック認識サブネットマネージャは、好ましくは、ゲートウェイポートおよびそれらが表す接続性を発見することはできるが、ゲートウェイポートの構成を変更することはできない。リーフサブネットにおけるファブリック認識サブネットマネージャは、デフォルトでは、好ましくは、ゲートウェイポートを越えたりモートポート接続性やそれ以上のホップを発見できず、ゲートウェイポートを越えて構成情報を修正することもできない。

【0259】

例示的な実施形態では、FGMインスタンスは、デフォルトで、完全なローカルファブリックを発見することができることになる。FGMインスタンスは、例えば、特定のコア

10

20

30

40

50

ファブリックインスタンスの外部または完全なコアファブリックの外部などローカル転送ドメインの外部でのトポロジのホップ毎DR SMPベースの発見の実行を、たとえばFGMインスタンスに対する構成オプションによって回避するか、または、そのような発見の実行を、リモート仮想システムにおけるアクセス制限によって、妨げられてもよい。アクセス制限は、典型的には、異なるファブリックレベルの所有者を持つ異なるリーフサブネットに適用される。この場合、中央FGMは、各コアファブリックインスタンスからのトポロジ情報および各リーフサブネットからのトポロジ情報を関連するサブネットマネージャまたはFGMインスタンスを介して収集し、この情報を使用して完全なローカルファブリックの最適化されたルーティングを評価する。

【0260】

10

デフォルトでは、ローカルファブリック内における完全なエンド・ツー・エンド接続性は、各リーフサブネットにおけるすべてのエンドポート間、各リーフサブネットにおけるすべてのエンドポートとすべてのゲートウェイポートとの間、ならびに各コアファブリックインスタンスにおける各ゲートウェイポート間、および特にリーフサブネットゲートウェイポートの任意の対の間に完全な接続性が確立されている各転送ドメインのローカルルーティングの後に存在することに注目されたい。

【0261】

上記の完全な接続性に基づいて、中央FGMは、ソースリーフサブネットにおける経路と、もしあれば関連するコアファブリックインスタンスにおける経路と、宛先リーフサブネットにおける関連の経路とを組み合わせることに基づいて、ローカルファブリックにおけるすべての潜在的な「ピアの対」に対するエンド・ツー・エンド経路を選択することができる。中央FGMは、異なるリーフサブネット間のエンド・ツー・エンド経路がクレジットループを表さないことを保証する必要があるだろう。しかしながら、次の条件が満たされている限り、クレジットループ回避はトリビアルに保証される：各リーフサブネットは、ファットツリーを表し、そこでは、コアファブリックインスタンスおよび／または他のリーフサブネットへのゲートウェイポートはローカルリーフサブネットにおけるルートスイッチノードを表し、各コアファブリックインスタンスは、すべてのリーフサブネットゲートウェイポートを、通常ファットツリー内におけるリーフノード／ポートとして、ルーティングされる。

20

【0262】

30

例示的な実施形態によれば、完全なローカルファブリック内において専用または「主要な」ダウン経路を介して静的最適化を保証するために、中央FGMは、ターゲットリーフサブネットにおける宛先ノードのための専用または「主要な」ダウン経路を表すルートスイッチにおける出口ゲートウェイポートを使用することになるターゲットリーフサブネットのための経路を選択する。ルートスイッチに関する情報は、一般に、リーフサブネット内におけるマスタサブネットマネージャ(SM)から入手可能になる。リーフサブネットにおけるマスタサブネットマネージャ(SM)は、リーフサブネットのローカルルーティングの一部として宛先ノードについて本実施形態のシステムの仮想スイッチまたは完全な物理的インスタンスであってもよいルートスイッチを選択している。

【0263】

40

ゲートウェイポートの選択は、本来、現在ソースリーフサブネットからの接続性を有し、この時点でのリーフサブネット間トラフィックの「合理的な」均衡を表すコアファブリックインスタンスの選択も表す。

【0264】

次いで、中央FGMは、ソースリーフサブネットから選択されたコアファブリックインスタンスへの入口ゲートウェイポートと、選択されたコアファブリックインスタンスにおける選択された宛先出口ゲートウェイポートのための主要経路を表すISRN値とを選択する。

【0265】

コアファブリックインスタンスへの入口ゲートウェイポートの選択も、この時点にお

50

るリーフサブネット間トラフィックの「合理的な」均衡を表す。

【0266】

例示的な実施形態によれば、ソースリーフサブネットにおけるソースポートから選択されたコアファブリックインスタンスへの選択された入口ゲートウェイポートへの経路は、ソースリーフサブネットにおけるマスタサブネットマネージャ（SM）によって定義される。この経路は、好ましくは、リーフサブネット間トラフィックについての潜在的な使用についての知識なしに、マスタサブネットマネージャ（SM）によって構築される。中央FGMは、ソースおよびターゲットリーフサブネットの両方におけるマスタサブネットマネージャからの情報を、選択されたコアファブリックインスタンスについての選択されたISR値と組み合わせることに基づいて、完全な経路を構築できる。

10

【0267】

中央FGMは、LID値およびISR値によって定義される純粋な経路情報に加えて、関与する各マスタサブネットマネージャからSL値も取得し、さらにトラフィッククラスおよびフローラベルも選択し、コアファブリック内においてSL/VL選択を反映し、ターゲットリーフサブネットが入るとSL選択を反映する。

【0268】

最も単純なデフォルトの場合、単一のSL値が選択され得、それは、完全なローカルファブリック全体を通じて、例えばレガシーの単一サブネットファットツリートポロジに対してと同様のようにすべての転送ドメインにおいて固定されたVLへの一貫したマッピングによって、保存されることになることに注目されたい。

20

【0269】

DLIDが最終的な宛先を表し、ファブリック内におけるさまざまなリンクおよび／またはスイッチ障害の結果としてファブリックを介して異なる経路を使用するよう、同じDLIDを再ルーティングできる、従来の単一のサブネットファブリックとは異なり、ファブリック内におけるDLID-ISR-DLIDベースのエンド・ツー・エンド経路は、ソースサブネットからコアファブリックへの入口ゲートウェイポートについてのDLID、およびコアファブリックインスタンスから宛先リーフサブネットへの出口ゲートウェイポートを表すISR値の、調整された再ルーティングがない限り、障害の結果として更新する必要があるかもしれないことに注目されたい。したがって、例示的な実施形態では、ゲートウェイポートがもはや利用可能でない場合、同等の順方向接続性を提供できる代替ゲートウェイポートを表すよう、関連するDLID値またはISR値の再ルーティングが行われるか、または影響を受けたフロー／通信ピアについての経路情報の更新が行われる。

30

【0270】

また、例示的な実施形態では、全体的なリーフサブネット間経路の再平衡化を実行するために、中央FGMは、1つ以上のフローのための新たな経路情報を選択的に生成し、関与するピアに通知するか、または関連するすべてのマスタサブネットマネージャおよびFGMインスタンスを含む現在のDLID-ISR-DLIDベースの経路のエンド・ツー・エンド再ルーティングを調整する。

【0271】

物理ノードおよび仮想ノードポピュレーションのグローバルなオーバービューの維持：
例示的な実施形態によれば、各リーフサブネットにおける物理エンドノードおよび仮想エンドノードは、ローカルマスタサブネットマネージャ（SM）によって発見され、構成される。各マスタサブネットマネージャは、照会およびイベントによって、中央FGMに、そのローカルノードポピュレーションに関して通知する。物理および仮想管理サーバならびにスイッチ管理ポートを含む通信エンドノードを伴うコアファブリックの場合、関連するFGMインスタンスは、リーフサブネットにおけるマスタSMと同じ態様で、ポピュレーションに関する情報を追跡または他の態様で維持する。これにより、中央FGMは、GUIDに基づいて任意のエンドノードの所在を追跡することができ、したがって任意の必要な通信に関する関連する経路情報を構築することができる。

40

50

【0272】

一般に、単一のリーフサブネット内でのVMマイグレーションはリーフサブネット内でローカルに処理されるが、ルーティングまたは経路情報の変更は中央FGMインスタンスに伝達される。リーフサブネット境界を越えるVMマイグレーションは、一般に、中央FGMを介して調整される。

【0273】

中央FGMは、一般に、vGUIDの調整を担い、P_Key値は、さまざまなリーフサブネットおよびコアファブリックインスタンスの境界にわたる範囲を有することに注目されたい。したがって、現在のグローバルポリシーと同期していない転送ドメイン内における値の割当ては、関連する転送ドメイン外で可視になるよう許されもしなければ、他の転送ドメインに悪影響を与えもしないことになることに注目されたい。

10

【0274】

仮想マシン（VM）マイグレーション：

例示的な実施形態によれば、単一リーフサブネット内での仮想マシン（VM）マイグレーションは、一意のDLID値を用いて、VMが表わしている仮想HCAポートを示す限り、アドレス情報の変更なしに生じ得る。

【0275】

VMがリーフサブネットからマイグレーションされるたびに、一般に、アドレス指定情報は、VMに到達するのにLRHのみ（すなわち、GRHなし）に基づくアドレス指定を使用していた元のリーフサブネットにおけるすべてのピアについて更新され、なぜなら、GRHの使用がここで必要となるからである。同様に、VMがリーフサブネットからマイグレーションされるたびに、アドレス指定情報は、現在GRH/ISRNベースのアドレス指定を使用している他の転送ドメインにおけるすべてのピアについて、更新される。これがあてはまるのは、元のサブネットで使用されるVMのLID値を保持することが、最初のコアファブリックゲートウェイポートについてのDLIDおよび後のリーフサブネットゲートウェイポートについてのISRN値の両方の再ルーティングを実行することに加えて、可能でない限りにおいてである。

20

【0276】

さらなる例示的な実施形態によれば、VMがマイグレーションされる新たなリーフサブネットを直接接続する、システムインスタンスにおける仮想ルータポートを介したTCAMベースの転送およびLRH生成を使用して、最初のコアファブリックゲートウェイポートについてのLIDおよび後のリーフサブネットゲートウェイポートについてのISRN値が、関連する仮想ルータポートへの転送を提供するために再ルーティングされた、中間状態を有することが可能であり、仮想ルータポートは新たなDLIDを伴う最終的なLRHの生成を行うことができる可能性があり、マイグレーションに続いて完全な古いアドレス指定情報が有効になるであろう。

30

【0277】

上記のスキームは、関連するすべてのピアについてのアドレス情報を更新するための「遅延」スキームと必要または所望に応じて組み合わせることができ、古い（元の）アドレス情報および新たなアドレス情報の両方がそのときVMに対して同時に有効となり得る。例えば、中央FGMによって、いずれのピアももはや古いアドレス情報に依存しないことが確認された場合、古いアドレス情報は「リサイクル」できる。

40

【0278】

このスキームは、少なくとも元のISRN値が、専らVMに対して（すなわち、リーフサブネットゲートウェイポートに対してそのようなものとしてではなく）割当てられ、それを、中間コアファブリック経路について同じISRNを使用して他の宛先について競合を引き起こすことなく再ルーティングできるようにすることを前提とする。

【0279】

マイグレーション後にVMに到達するよう新たなコアファブリックインスタンスを使用しなければならないことを必要とする態様でVMマイグレーションが生じた場合、すべて

50

の通信リーフサブネットにおいてコアファブリックゲートウェイポートをアドレス指定するために使用されるDLID値について同じ制約が存在する。

【0280】

特別なケースとして、元のリーフサブネット内における通信ピアがある。—これらのピアが最初にGRH/ISRNベースのアドレス指定を使用するように指示されている限り、元のDLID値およびISRN値の再ルーティングに基づいてVMマイグレーションを透過的に処理することができる（つまり、まず、元のLRH、DLIDはVMを直接指していることになるが、マイグレーション後、これは、関連するコアファブリックゲートウェイポートを指すように更新されていることになる。

【0281】

リーフサブネットインスタンスからマイグレーションされているVMの観点からは、一般に、すべてのその通信ピアについてのアドレス情報を更新しなければならず、なぜならば、特に、最初のホップアドレス指定は、異なるリーフサブネットでは同じではないからである。これは、上のセクションで概説した通信ピアについてのアドレス透過性を実現するためのすべての拡張機能とは無関係である。

【0282】

上記のすべての場合において、前提は、FLSNチェックがイネーブルにされていないこと、または同じFLSN値を有するリーフサブネットインスタンス間でのみマイグレーションが行われること、またはシステム実装によっていくつかのFLSN値が上述したように各ポートに関連付けられ、競合するFLSN値を有する少数のマイグレーションを、任意の時点で、つまりすべてのピアが上述のようにアドレス情報を更新してしまうまで、処理することができること、であることに注目されたい。

【0283】

リーフサブネット/コアファブリックモデルに対するハードウェアサポート：

単一のリーフサブネット内では、例示的な実施形態のシステムは、インフィニバンド（IB）規格に従って線形転送テーブル（LFT）を使用してLRH、DLIDベースのパケット転送を実現する。デフォルトでは、各仮想スイッチは、完全な48KのユニキャストDLID空間をサポートするLFTと、完全な16KのマルチキャストDLID空間をサポートするマルチキャスト転送テーブルとを有する。システムに対するPORが、物理ポートの対によって共有されるプライベートユニキャストおよびマルチキャストLFTを有することになる限り、仮想スイッチポートメンバーシップを構成する最小限の細分性は、LFTを共有する2つのポートのグループである。次に、拡張されたSMおよびFGMルーティングロジックで、仮想スイッチごとの複数の独立したLFTをサポートしてもよい。しかしながら、LFTが1つより多い物理ポートによって共有されている限り、依存性は、実際の物理トポロジがそれ自体を複数の独立したLFTの処理に供することである。

【0284】

物理ポート間でのLFTの共有は、仮想スイッチごとに1つより多いLFTを利用することが有益となるために、ケーブルコネクタ当たりのポートのグループ化（4xより大きい場合）および大きなローカルファブリックのケーブル接続された物理的接続性がどのように見えるべきかに対して影響を有することに注目されたい。また、さまざまな所有権を主張できる技術またはプレ標準技術を使用して、本システムの実施形態は、潜在的に、LRH、DLIDユニキャスト空間を48Kを超えて拡張する可能性がある。

【0285】

例えば、異なるリーフサブネット間および/またはコアファブリックインスタンス間などの転送ドメイン間の境界を識別するために、例示的な実施形態のシステムは、以下に説明する特徴のセットを実現する。

【0286】

第1に、リーフサブネット内から、任意のコアファブリックゲートウェイポートが1つ以上のDLIDによって識別され、転送が通常のLRH、DLIDベースの探索を使用し

10

20

30

40

50

て行われる。探索された出力ポートがLRH/DLID（元のLRHに基づく）からGRH/ISRNへのアドレス指定モードにおける変更を必要とするコアファブリックゲートウェイポートを表すことを識別するために、パケットが出力ポートの出口リンクに到達する前にモードのこの変更を判断できる情報が提供される。パケットのストールを避けるために、この情報は、パケットが入口ポートで受信された後、可能な限り早く利用可能にされ、アービターがパケットを出力ポートに転送できるようにすると、新たなLRHが準備され、十分に定義される。したがって、情報は、入口ポートにおける出力ポートの探索とともに利用できる。さらに、同じ出力ポートが、どの他のローカルシステムポートがパケットをそれにクロスバーを介して送信しているかに依って、1つより多い「パーソナリティ」を有することを可能にするために、各入力ポートに関連付けられるレジスタを介して利用可能なタイプの出力ポートを有することは最も柔軟なスキームである。これは、LFTエントリにおいてエンコードされるポートタイプを有することに対して空間を節約し、それは、柔軟性を与え、共有LFTエントリ、および出力ポートのタイプがチップ全体に対して一回定義される集中型アプローチを使用することの両方に対して、アクセス調停の必要性がない。新たなアドレス指定モードで新たなパケットを構築するためのハードウェアサポートは、内部レジスタに基づく固定されたLRH、DLIDの生成、およびGRHから正しいISRNビットフィールドを識別して抽出するロジックを含む。

【0287】

第2に、コアファブリックインスタンス内では、入口ポートにおけるGRH/ISRN転送モードの識別は、着信LRH、DLID値を内部レジスタの内容と照合することに基づく。次いで、パケットLRH、DLID値がレジスタに含まれる値と一致する場合、転送モードは、例示的な実施形態の「GRH/ISRN」転送モードである。それ以外の場合、パケットLRH、DLID値がレジスタに含まれる値と一致しない場合には、転送モードはインフィニバンド（IB）レガシー「LRH/DLID」転送モードである。

【0288】

例示的な実施形態では、性能および柔軟性のために、このレジスタは、好ましくは、すべてのポートに対して複製される。これは、サブネットマネージャ（SM）はこのシステム特徴を認識していないが、そのリーフサブネット内において利用可能なLFTサイズを低減することを伴わない、リーフサブネット仮想スイッチインスタンスを有することも可能であることを意味する。コアファブリックインスタンス内からは、ゲートウェイポート（またはエンドポート）に向かうルートは、各「ルート内」システム入口ポートのLFTから出力ポートを探索するために使用されるISRN値によって定義される。

【0289】

出力ポートが宛先リーフサブネットへのゲートウェイであることを判断するために、例示的な実施形態の単純な解決策は、パケットFLSN番号を、探索された出力ポートに関連付けられるFLSN番号と比較することである。しかしながら、入口ポートが1つより多いモードで動作することを依然として可能にしながらアドレス指定モードにおける必要な変更を識別するためには、入口ポートが、可能な出力ポートごとに、フラグを有して、出力ポートがアドレス指定モードにおける変更を表すかどうか、または出力ポートがアドレス指定モードの変更を表さないことを告げれば十分である。したがって、入力転送モードがGRH/ISRNであり、出力ポートがアドレス指定モードにおける変更を表すものとして（すなわち、この入口ポートに対して）フラグを立てられる場合、アドレス指定モードはGRH.DGID.prefix.dlidに含まれるDLID値に基づいて決定論的にLRH/DLIDに変更することができる。この場合、FLSN値を一致させることは、コアファブリックインスタンス内におけるISRNベースのルーティングが正しい限り、追加の保全性チェックを表すに過ぎないであろうことに注目されたい。

【0290】

新たなアドレス指定モードを伴う新たなパケットを構築するためのHWサポートは、GRHから正しいDLIDビットフィールドを識別し抽出するロジックと、抽出されたDLID値に基づいた新たなLRHの生成とを含む。

【0291】

第3に、2つのリーフサブネットが同じシステムインスタンス上の異なる仮想スイッチによって直接接続されている場合、典型的には、2つの仮想スイッチにおけるポート間に相互リーフサブネットゲートウェイポート関係が存在する。したがって、リーフサブネットA内からは、リーフサブネットBへの直接接続性を提供する特定のリーフサブネットゲートウェイポートを識別する1つ以上のDLID値が存在することになる。次いで、出力ポートが宛先リーフサブネットへのゲートウェイであることを判断するために、例示的な実施形態の単純な解決策は、パケットFLSN番号を、探索された出力ポートに関連付けられるFLSN番号と比較することである。しかしながら、入口ポートが1つより多いモードで動作することを依然として可能にしながらアドレス指定モードにおける必要な変更を識別するためには、上記のケースに対してのように、入口ポートが、可能な出力ポートごとに、フラグを有して、出力ポートがアドレス指定モードにおける変更を表すかどうか、または出力ポートがアドレス指定モードの変更を表さないかどうかを示せば十分である。

10

【0292】

それでも、アドレス指定モードの変更が、LRH/DLIDからGRH/ISRNまたはLRH/DLIDからLRH/DLIDのいずれか（すなわち、この例におけるように）であり得る場合、出力ポートごとに単一のフラグ値のみを有することは十分ではない。したがって、出力ポートごとのフラグは、好ましくは、アドレス指定モードの変更だけでなく、ゲートウェイポートのタイプを識別する。しかしながら、ここでも、FLSNマッ

20

【0293】

リーフサブネット／コアファブリックモデルの追加ハードウェアサポート：

任意の数の入力DGID、サブネットプレフィックス値をサポートするが、依然としてDGID、GUID値をターゲット物理または仮想HCAポートについて検証できるようにするために、そのようなターゲット（v）HCAは、実施形態に従って、あるモードを含み、着信GRH、DGIDを物理または仮想のHCAポートについてハードウェアおよびエイリアスGUID値に対してチェックすることはGUID部分のみに制限され、サブネットプレフィックス部分は無視される。

30

【0294】

このモードは、各GUIDテーブルエントリについて個々に選択的に指定してもよい。しかしながら、HWの実現の観点からは、いくつかのテーブルエントリに対して連想的な一致を有し、次いで、異なる一致基準に基づいて複数の一致も有することは、困難である。したがって、一実施形態によれば、チェックが常にGUID部分のみについて行われるモードで十分である。

【0295】

仮定は、その場合、各仮想HCAポートは、GUID部分のみが考慮されるときにも一意である2つ以上のGUIDテーブルエントリとともに設定される、というものである。通常またはデフォルトの場合は、関連するhwGUIDは、現在アクティブなVMを表すvGUIDとともに存在する、ということである。2つの対応するGUIDテーブルエントリは、各々、割り当てられたISN値に基づくサブネットプレフィックス値を有することになる。

40

【0296】

（LFT）探索のために（GRH）におけるサブネットプレフィックス値を使用する方法

図16は、高性能コンピューティング環境でのネットワークスイッチ環境における線形転送テーブル探索のためにパケットヘッダを使用するための一実施形態による方法1600を示す。ここで、図の方法について説明する。第1および第2のサブネットは、ネットワークファブリックを含むコンピュータ環境で提供される。第1のサブネットは第2のサ

50

ブネットとは異なる。第1のサブネットは、第1のセットのネットワークスイッチを含み、第1のセットのネットワークスイッチの各々は、線形転送テーブル（LFT）および複数の物理ポートを含む。第2のサブネットは、第2のセットのネットワークスイッチを含み、第2のセットのネットワークスイッチの各々は、LFTおよび複数の物理ポートを含む。これは、本方法の例示的な実施形態の環境である。全体として、方法1600は、第1のセットのネットワークスイッチのうちの第1のネットワークスイッチの同じ第1のLFTが、第1のサブネットの第1のセットのネットワークスイッチのネットワークスイッチ間で転送されるべきパケットのサブネット内転送判断、ならびに第1および第2のサブネットのそれぞれの第1のセットのネットワークスイッチおよび第2のセットのネットワークスイッチの間で転送されるべきパケットのサブネット間転送判断の両方に対して用いられることを可能にする。

10

【0297】

第1のステップ1610において、第1のパケットが、第1のサブネットの第1のセットのネットワークスイッチのうちの第1のネットワークスイッチで受信される。ステップ1620において、第1のパケットのヘッダの第1の部分が検査される。第1のパケットのヘッダの第1の部分の第1の条件がステップ1630で判断され、第1のパケットのヘッダの第1の部分の第2の条件がステップ1640で判断される。

【0298】

サブネット内転送判断は、第1のパケットのヘッダの第1の部分が第1の条件を有するか、そうでなければ第1の条件を示すと判断された場合に、ステップ1630において選択的に行われる。そうでない場合、第1のパケットのヘッダの第1の部分が第1の条件を有していないか、そうでなければ第1の条件を示していないと判断された場合、方法はサブネット間転送判断ステップ1640に進む。

20

【0299】

第1のパケットのヘッダの第1の部分が第1の条件を有するか、さもなければ第1の条件を示すと判断された場合など、サブネット内転送判断がステップ1630において選択的に行われる場合、パケットは、例示的な実施形態に従って、ステップ1635において、サブネット内（LRH/DLID）転送を用いて処理される。

【0300】

同様に、第1のパケットのヘッダの第1の部分が第2の条件を有するか、さもなければ第2の条件を示すと判断された場合など、サブネット間転送判断がステップ1640において選択的に行われる場合、パケットは、例示的な実施形態に従って、ステップ1645において、サブネット間（GRH/ISRN）転送を用いて処理される。

30

【0301】

ステップ1645において、サブネット間転送が、第2のサブネットの第2のネットワークスイッチのセットのうちの選択されたネットワークスイッチに第1のパケットをルーティングするために、第1のパケットのヘッダの第1の部分の第2の条件に従って、第1のネットワークスイッチの第1のLFTを使用して、選択的に行われる。

【0302】

例示的な実施形態では、第1のパケットのヘッダの第1の部分を検査することは、第1のパケットのヘッダのローカルルートヘッダ（LRH）部分を検査することを含む。さらに、例示的な実施形態によれば、第1のパケットのヘッダの第1の部分の条件を判断することは、LRHの選択された部分を第1のネットワークスイッチに格納される所定の値と比較することと、LRHの選択された部分と所定の値との間の不一致に従って第1の条件を判断することと、LRHの選択された部分と所定の値との間の一致に従って第2の条件を判断することを含む。

40

【0303】

さらに、例示的な実施形態では、LRHの選択された部分と所定の値との間の一致に従って第2の条件を選択的に判断することは、LFTを使用して、LRHの選択された部分を使用して特殊スイッチポート（SSP）を索引付することを含む。

50

【0304】

特に、例示の実施形態に関して、第1のパケットを、第2のサブネットの第2のネットワークスイッチのセットのうちの選択されたネットワークスイッチにルーティングすることは、第1のパケットのヘッダのグローバルルートヘッダ（GRH）部分のセクションを選択することを含み、ヘッダのGRH部分の選択されたセクションは、サブネット間ルート番号（ISRN）を格納し、ルーティングすることはさらに、GRHのISRNを使用して、第1のネットワークスイッチの第1のLFTを索引付けすることと、GRHのISRNを使用して第1のLFTを索引付けすることに従って、第1のパケットを第2のサブネットの第2のネットワークスイッチのセットのうちの選択されたネットワークスイッチにルーティングすることとを含み、第2のサブネットはネットワークスイッチ環境の中間コアファブリックである。

10

【0305】

さらに、例示的な実施形態によれば、第1のパケットのヘッダの第1の部分の第1または第2の条件のいずれもステップ1630およびステップ1640で判断されないときは、パケットはステップ1650で破棄される。すなわち、例示的な実施形態では、転送条件のいずれも満たされないため、パケットはステップ1650で破棄される。

【0306】

例示の実施形態の特徴は、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製品において、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。

20

【0307】

例示の実施形態の特徴は、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例示の実施形態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。

30

【0308】

例示の実施形態の特徴はまた、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。

【0309】

加えて、例示の実施形態は、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。

40

【0310】

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。

50

【0311】

例示的实施形態は、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。

【0312】

例示的实施形態の前述の説明は、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限定するよう意図されていない。例示的实施形態の幅および範囲は、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。

10

【図1】

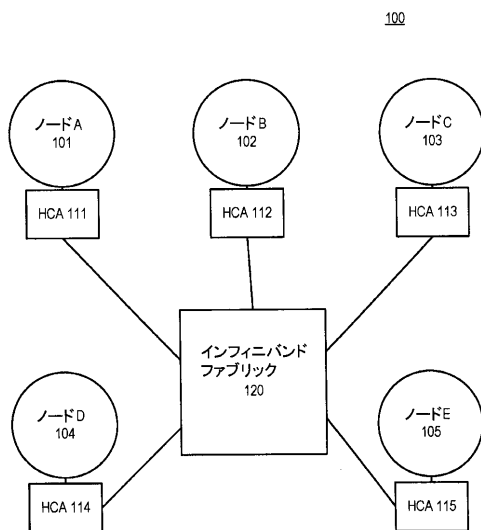


FIGURE 1

【図2】

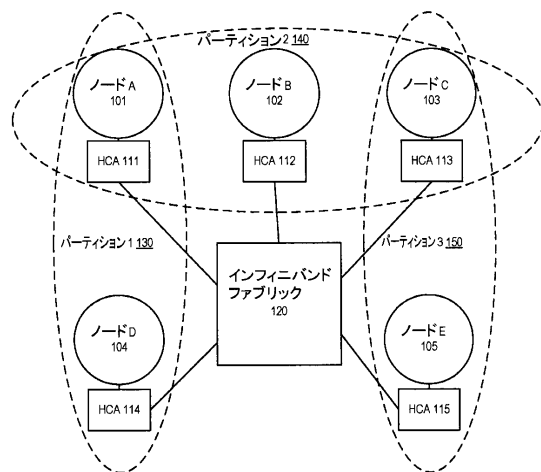


FIGURE 2

【図 3】

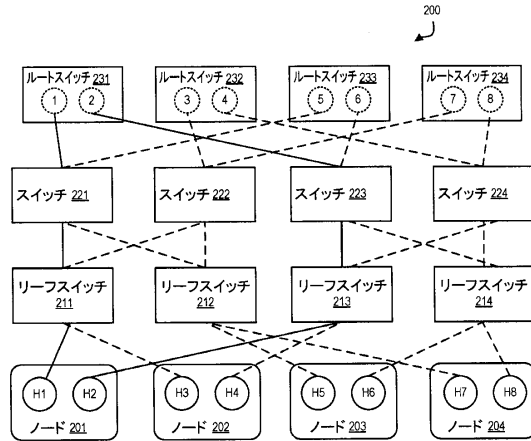


FIGURE 3

【図 4】

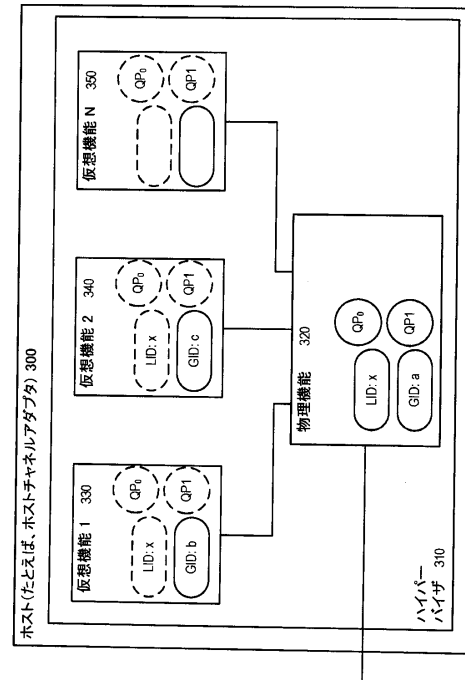


FIGURE 4

【図 5】

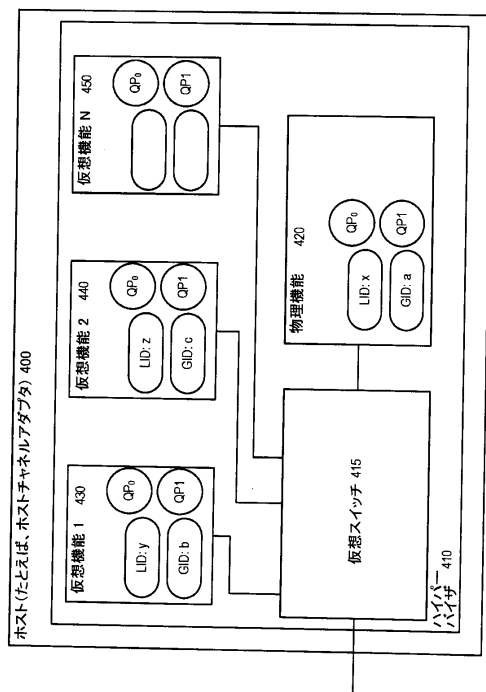


FIGURE 5

【図 6】

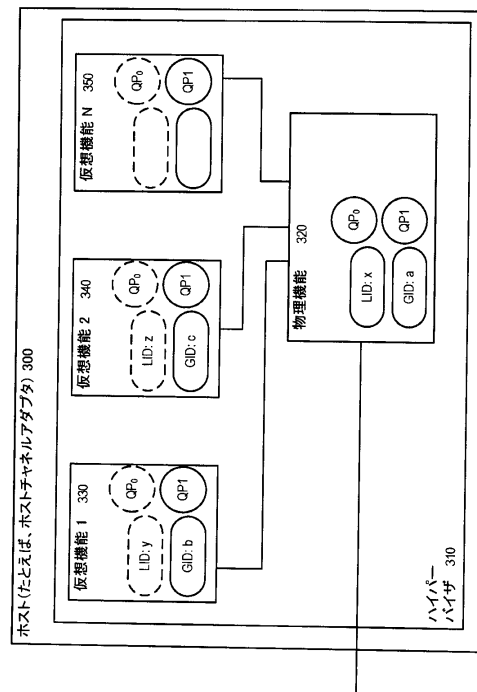


FIGURE 6

【図 7】

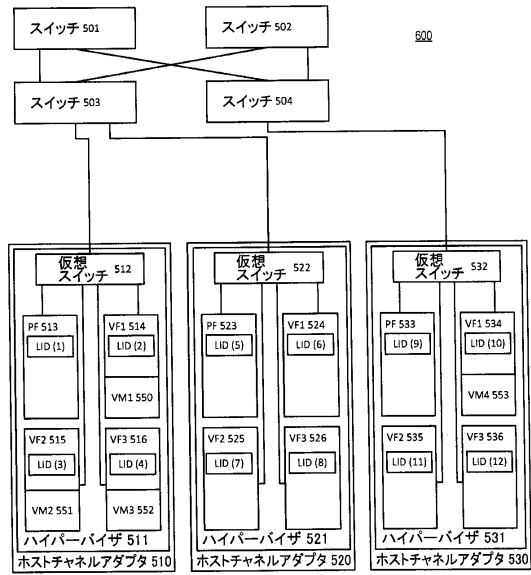


FIGURE 7

【図 8】

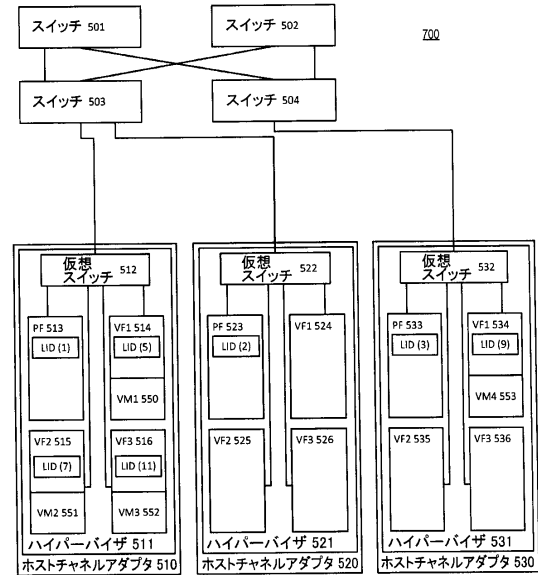


FIGURE 8

【図 9】

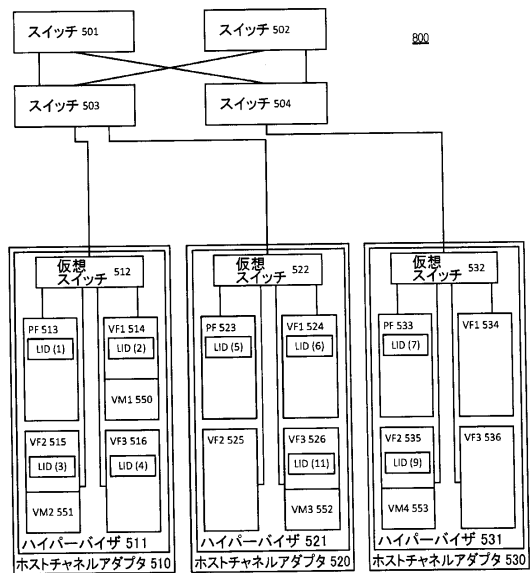


FIGURE 9

【図 10】

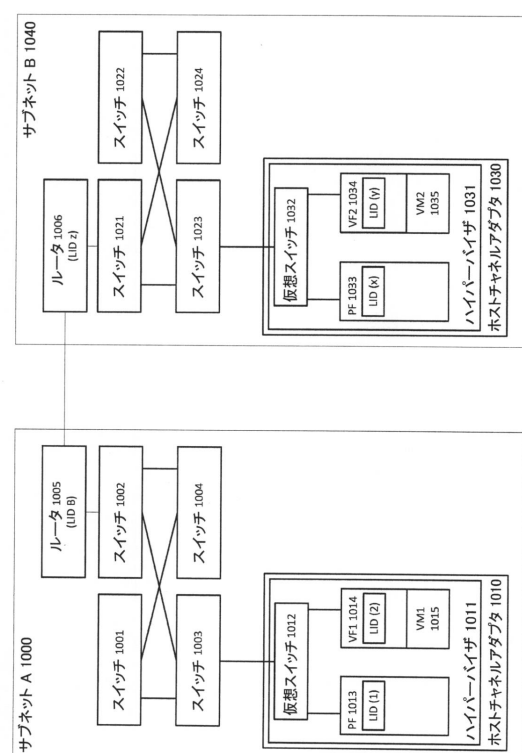


FIGURE 10

【図 1 1】

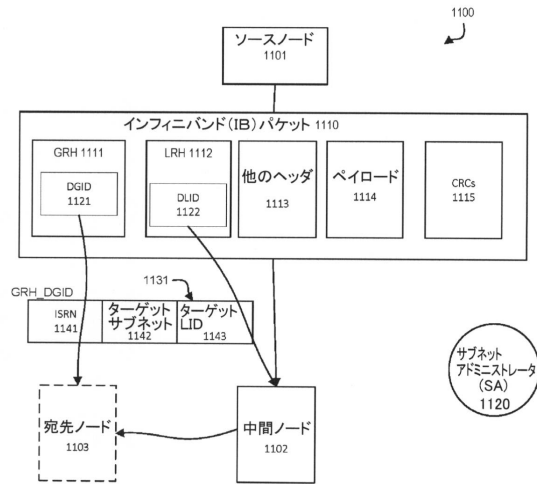


FIGURE 11

【図 1 2】

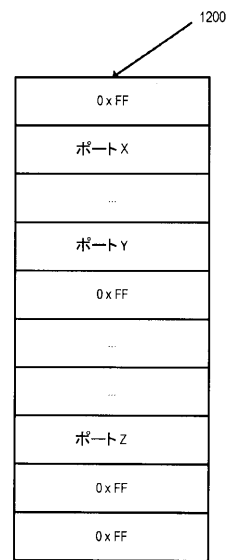


FIGURE 12

【図 1 3】

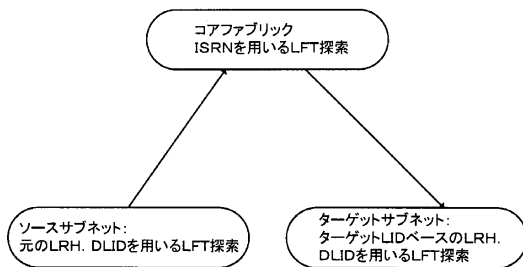


FIGURE 13

【図 1 4】

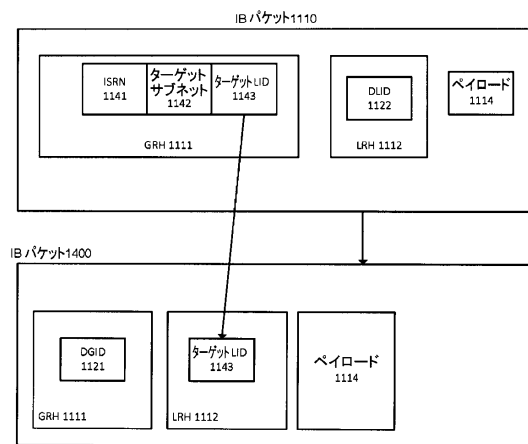


FIGURE 14

【図 16】

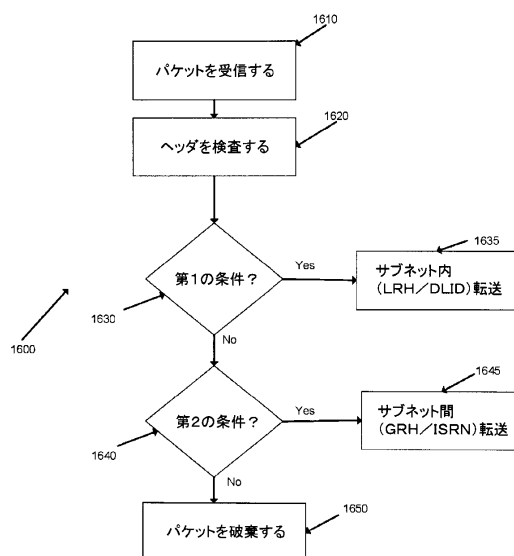


FIGURE 16

FIGURE 15

フロントページの続き

(72)発明者 ミュラー, シモン

アメリカ合衆国、94086 カリフォルニア州、サニーベール、ラ・メサ・テラス、983、ア
パートメント・ディ

審査官 佐々木 洋

(56)参考文献 米国特許出願公開第2013/0259033 (US, A1)

中国特許出願公開第103765833 (CN, A)

(58)調査した分野(Int.Cl., DB名)

H04L 12/931

H04L 12/741

H04L 12/70

G06F 13/14

G06F 15/173