

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号  
特許第7566900号  
(P7566900)

(45)発行日 令和6年10月15日(2024.10.15)

(24)登録日 令和6年10月4日(2024.10.4)

(51)国際特許分類

F I

H O 4 L 47/12 (2022.01)

H O 4 L 47/12

H O 4 L 47/70 (2022.01)

H O 4 L 47/70

請求項の数 7 (全96頁)

|                   |                               |          |                                                                                 |
|-------------------|-------------------------------|----------|---------------------------------------------------------------------------------|
| (21)出願番号          | 特願2022-529054(P2022-529054)   | (73)特許権者 | 502303739                                                                       |
| (86)(22)出願日       | 令和2年8月6日(2020.8.6)            |          | オラクル・インターナショナル・コーポレーション                                                         |
| (65)公表番号          | 特表2023-503063(P2023-503063 A) |          | アメリカ合衆国 9 4 0 6 5 カリフォルニア州 レッドウッド ショアーズ, メール ストップ 5 オーピー 7 オラクル パークウェイ 5 0 0    |
| (43)公表日           | 令和5年1月26日(2023.1.26)          | (74)代理人  | 110001195                                                                       |
| (86)国際出願番号        | PCT/US2020/045271             |          | 弁理士法人深見特許事務所                                                                    |
| (87)国際公開番号        | WO2021/101600                 | (72)発明者  | ヨンセン, ビョルン・ダグ                                                                   |
| (87)国際公開日         | 令和3年5月27日(2021.5.27)          |          | アメリカ合衆国、9 4 0 6 5 カリフォルニア州、レッドウッド・シティ、オラクル・パークウェイ、5 0 0、オラクル・インターナショナル・コーポレーション |
| 審査請求日             | 令和5年3月10日(2023.3.10)          |          |                                                                                 |
| (31)優先権主張番号       | 62/937,594                    |          |                                                                                 |
| (32)優先日           | 令和1年11月19日(2019.11.19)        |          |                                                                                 |
| (33)優先権主張国・地域又は機関 | 米国(US)                        |          |                                                                                 |
| (31)優先権主張番号       | 16/872,035                    |          |                                                                                 |
| (32)優先日           | 令和2年5月11日(2020.5.11)          |          |                                                                                 |
|                   | 最終頁に続く                        |          | 最終頁に続く                                                                          |

(54)【発明の名称】 高性能コンピューティング環境においてプライベートファブリックにおける帯域幅輻輳制御を提供するためのシステムおよび方法

(57)【特許請求の範囲】

【請求項1】

高性能コンピューティング環境においてプライベートファブリックにおける帯域幅輻輳制御を提供するためのシステムであって、

1つ以上のマイクロプロセッサを含むコンピュータと、

前記コンピュータにおいて提供される第1のサブネットとを備え、前記第1のサブネットは、

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

複数のホストチャネルアダプタを含み、前記ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、前記複数のホストチャネルアダプタは、前記複数のスイッチを介して相互接続され、前記第1のサブネットはさらに、

複数のエンドノードを含み、

エンドノードに関連付けられるエンドノードイングレス帯域幅割当が、ホストチャネルアダプタに提供され、前記エンドノードは、前記ホストチャネルアダプタに取り付けられ、

イングレス帯域幅は、前記ホストチャネルアダプタの前記エンドノードにおいて受信され、前記イングレス帯域幅は、前記エンドノードの前記イングレス帯域幅割当を超えており、

前記イングレス帯域幅は、少なくとも2つの遠隔エンドノードから受信され、

前記複数のスイッチのうちの2つのスイッチ間のスイッチ間リンクまたは前記複数のスイ

ツチのうちのあるスイッチのポートのうち少なくとも1つにおいて、ターゲットグループが定義され、前記ターゲットグループは、前記複数のスイッチのうち2つのスイッチ間のスイッチ間リンクまたは前記複数のスイッチのうちのあるスイッチのポートのうちの前記少なくとも1つにおいて帯域幅割当を定義し、

ターゲットグループリポジトリが、前記ホストチャネルアダプタのメモリに記憶され、前記定義されたターゲットグループは、前記ターゲットグループリポジトリに記録され、前記少なくとも2つの遠隔エンドノードから前記インGRES帯域幅を受信することに応答して、前記ターゲットグループの前記帯域幅割当が更新される、システム。

【請求項2】

前記インGRES帯域幅は遠隔エンドノードから受信され、

前記ホストチャネルアダプタに取り付けられた前記エンドノードと前記遠隔エンドノードとの間で帯域幅割当が交渉される、請求項1に記載のシステム。

【請求項3】

前記遠隔エンドノードから前記インGRES帯域幅を受信したことに応答して、前記ホストチャネルアダプタに取り付けられた前記エンドノードから前記遠隔エンドノードに応答メッセージが送信される、請求項2に記載のシステム。

【請求項4】

前記応答メッセージは、前記ホストチャネルアダプタに取り付けられた前記エンドノードの前記インGRES帯域幅割当を超えていることを示す、請求項3に記載のシステム。

【請求項5】

前記第1のサブネットはインフィニバンドサブネットを含む、請求項4に記載のシステム。

【請求項6】

高性能コンピューティング環境においてプライベートファブリックにおける帯域幅輻輳制御を提供するための方法であって、

1つ以上のマイクロプロセッサが、第1のサブネットを提供することを含み、前記第1のサブネットは、

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

複数のホストチャネルアダプタを含み、前記ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、前記複数のホストチャネルアダプタは、前記複数のスイッチを介して相互接続され、前記第1のサブネットはさらに、

複数のエンドノードを含み、前記方法はさらに、

ホストチャネルアダプタが、前記ホストチャネルアダプタに取り付けられたエンドノードに関連付けられるエンドノードインGRES帯域幅割当を提供することと、

前記ホストチャネルアダプタの前記エンドノードが、インGRES帯域幅を受信することを含み、前記インGRES帯域幅は前記エンドノードの前記インGRES帯域幅割当を超えており、

前記インGRES帯域幅は、少なくとも2つの遠隔エンドノードから受信され、

前記複数のスイッチのうち2つのスイッチ間のスイッチ間リンクまたは前記複数のスイッチのうちのあるスイッチのポートのうち少なくとも1つにおいて、ターゲットグループが定義され、前記ターゲットグループは、前記複数のスイッチのうち2つのスイッチ間のスイッチ間リンクまたは前記複数のスイッチのうちのあるスイッチのポートのうちの前記少なくとも1つにおいて帯域幅割当を定義し、

ターゲットグループリポジトリが、前記ホストチャネルアダプタのメモリに記憶され、前記定義されたターゲットグループは、前記ターゲットグループリポジトリに記録され、前記少なくとも2つの遠隔エンドノードから前記インGRES帯域幅を受信することに応答して、前記ターゲットグループの前記帯域幅割当が更新される、方法。

【請求項7】

高性能コンピューティング環境においてプライベートファブリックにおける帯域幅輻輳

10

20

30

40

50

制御を提供するための命令を搬送するコンピュータ可読プログラムであって、前記命令は、読み取られ、実行されると、コンピュータにステップを実行させ、前記ステップは、

1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することを含み、前記第1のサブネットは、

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

複数のホストチャネルアダプタを含み、前記ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、前記複数のホストチャネルアダプタは、前記複数のスイッチを介して相互接続され、前記第1のサブネットはさらに、

複数のエンドノードを含み、前記ステップはさらに、

ホストチャネルアダプタにおいて、前記ホストチャネルアダプタに取り付けられたエンドノードに関連付けられるエンドノードイングレス帯域幅割当を提供することと、

前記ホストチャネルアダプタの前記エンドノードにおいて、イングレス帯域幅を受信することとを含み、前記イングレス帯域幅は前記エンドノードの前記イングレス帯域幅割当を超えており、

前記イングレス帯域幅は、少なくとも2つの遠隔エンドノードから受信され、

前記複数のスイッチのうちの2つのスイッチ間のスイッチ間リンクまたは前記複数のスイッチのうちの2つのスイッチのうちの少なくとも1つにおいて、ターゲットグループが定義され、前記ターゲットグループは、前記複数のスイッチのうちの2つのスイッチ間のスイッチ間リンクまたは前記複数のスイッチのうちの2つのスイッチのうちの少なくとも1つにおいて帯域幅割当を定義し、

ターゲットグループリポジトリが、前記ホストチャネルアダプタのメモリに記憶され、前記定義されたターゲットグループは、前記ターゲットグループリポジトリに記録され、前記少なくとも2つの遠隔エンドノードから前記イングレス帯域幅を受信することに応答して、前記ターゲットグループの前記帯域幅割当が更新される、コンピュータ可読プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

著作権表示

本特許文献の開示の一部には、著作権保護の対象となるものが含まれている。著作権者は、この特許文献または特許開示の何者かによる複製が、特許商標庁の特許ファイルまたは記録にある限り、それに対して異議を唱えないが、そうでなければ、いかなる場合もすべての著作権を留保する。

【0002】

優先権の主張および関連出願への相互参照：

本出願は、2019年11月19日に出願された「SYSTEM AND METHOD FOR PROVIDING QUALITY-OF-SERVICE AND SERVICE-LEVEL AGREEMENTS IN A PRIVATE FABRIC IN A HIGH PERFORMANCE COMPUTING ENVIRONMENT（高性能コンピューティング環境においてプライベートファブリックにおけるサービス品質およびサービスレベルの取り決めを与えるためのシステムおよび方法）」と題される米国仮特許出願、出願番号第62／937，594号に対する優先権の恩恵を主張し、その全体をここに引用により援用する。

【0003】

本出願は、以下の特許出願に対する優先権の恩恵も主張し、それら全体もここに引用により援用する：2020年5月11日に出願された「SYSTEM AND METHOD FOR SUPPORTING RDMA BANDWIDTH RESTRICTIONS IN A PRIVATE FABRIC IN A HIGH PERFORMANCE COMPUTING ENVIRONMENT（高性能コンピューティング環境においてプライベートファブリックにおけるRDMA帯域幅制限をサポートするためのシステムおよび方法）」と題される米国特許出願、出願番号第16／872，035号；2020

年5月11日に出願された「SYSTEM AND METHOD FOR PROVIDING BANDWIDTH CONGESTION CONTROL IN A PRIVATE FABRIC IN A HIGH PERFORMANCE COMPUTING ENVIRONMENT（高性能コンピューティング環境においてプライベートファブリックにおける帯域幅輻輳制御を提供するためのシステムおよび方法）」と題される米国特許出願、出願番号第16／872，038号；2020年5月11日に出願された「SYSTEM AND METHOD FOR SUPPORTING TARGET GROUPS FOR CONGESTION CONTROL IN A PRIVATE FABRIC IN A HIGH PERFORMANCE COMPUTING ENVIRONMENT（高性能コンピューティング環境においてプライベートファブリックにおける輻輳制御のためにターゲットグループをサポートするためのシステムおよび方法）」と題される米国特許出願、出願番号第16／872，039号；および2020年5月11日に  
10  
出願された「SYSTEM AND METHOD FOR SUPPORTING USE OF FORWARD AND BACKWARD CONGESTION NOTIFICATIONS IN A PRIVATE FABRIC IN A HIGH PERFORMANCE COMPUTING ENVIRONMENT（高性能コンピューティング環境においてプライベートファブリックにおける順方向および逆方向輻輳通知の使用をサポートするためのシステムおよび方法）」と題される米国特許出願、出願番号第16／872，043号。

【0004】

#### 分野

本教示は、インフィニバンド（IB）およびRoCE（コンバージドイーサネット（登録商標）を介するRDMA（リモートダイレクトメモリアクセス））などのプライベート  
20  
高性能相互接続ファブリックにおいてサービス品質（QOS：quality of service）およびサービスレベル合意（SLA）を実施するためのシステムおよび方法に関する。

【背景技術】

【0005】

#### 背景

より大きなクラウドコンピューティングアーキテクチャが導入されるにつれ、従来のネットワークおよびストレージに関連する性能ならびに管理ボトルネックが重大な問題となってきた。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（InfiniBand：IB）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。これは、本教示の実施形態が対処することが意図される一般的な  
30  
領域である。

【発明の概要】

【課題を解決するための手段】

【0006】

#### 概要：

特定の局面は、独立請求項に記載される。様々な任意選択の実施形態が従属請求項に記載される。

【0007】

本明細書では、高性能コンピューティング環境においてプライベートファブリックにおける帯域幅輻輳制御を提供するためのシステムおよび方法を説明する。例示的な方法は、1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することができ、第1  
40  
のサブネットは、複数のスイッチと、複数のホストチャネルアダプタとを含み、ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、複数のホストチャネルアダプタは、上記複数のスイッチを介して相互接続され、第1のサブネットはさらに、複数のエンドノードを含む。本方法は、ホストチャネルアダプタにおいて、ホストチャネルアダプタに取り付けられたエンドノードに関連付けられるエンドノードイングレス帯域幅割当を提供することができる。本方法は、ホストチャネルアダプタのエンドノードにおいて、イングレス帯域幅を受信することができ、イングレス帯域幅はエンドノードのイングレス帯域幅割当を超える。

【図面の簡単な説明】

【0008】

10

20

30

40

50

- 【図 1】一実施形態に従うインフィニバンド環境の一例を示す図である。
- 【図 2】一実施形態に従う分割されたクラスタ環境の一例を示す図である。
- 【図 3】一実施形態に従うネットワーク環境におけるツリートポロジの一例を示す図である。
- 【図 4】一実施形態に従う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】一実施形態に従う例示的な `vSwitch` アーキテクチャを示す図である。
- 【図 6】一実施形態に従う例示的な `vPort` アーキテクチャを示す図である。
- 【図 7】一実施形態に従う `LID` が予めポピュレートされている例示的な `vSwitch` アーキテクチャを示す図である。
- 【図 8】一実施形態に従う動的 `LID` 割当てがなされた例示的な `vSwitch` アーキテクチャを示す図である。 10
- 【図 9】一実施形態に従う、`vSwitch` に動的 `LID` 割当てがなされかつ `LID` が予めポピュレートされている、例示的な `vSwitch` アーキテクチャを示す図である。
- 【図 10】一実施形態に従う例示的なマルチサブネットインフィニバンドファブリックを示す図である。
- 【図 11】一実施形態に従う、高性能コンピューティング環境における 2 つのサブネット間の相互接続を示す図である。
- 【図 12】一実施形態に従う、高性能コンピューティング環境におけるデュアルポート仮想ルータ構成を介した 2 つのサブネット間の相互接続を示す図である。
- 【図 13】一実施形態に従う、高性能コンピューティング環境においてデュアルポート仮想ルータをサポートする方法のフローチャートを示す図である。 20
- 【図 14】一実施形態による、高性能コンピューティング環境において、`RDMA` 読出要求を、制限された特徴として提供するためのシステムを示す。
- 【図 15】一実施形態による、高性能コンピューティング環境において、`RDMA` 読出要求を、制限された特徴として提供するためのシステムを示す。
- 【図 16】一実施形態による、高性能コンピューティング環境において、`RDMA` 読出要求を、制限された特徴として提供するためのシステムを示す。
- 【図 17】ある実施形態に係る、高性能コンピューティング環境において明示的な `RDMA` 読出帯域幅制限を提供するためのシステムを示す。
- 【図 18】ある実施形態に係る、高性能コンピューティング環境において明示的な `RDMA` 読出帯域幅制限を提供するためのシステムを示す。 30
- 【図 19】ある実施形態に係る、高性能コンピューティング環境において明示的な `RDMA` 読出帯域幅制限を提供するためのシステムを示す。
- 【図 20】高性能コンピューティング環境において、`RDMA`（リモートダイレクトメモリアクセス）読出要求を、制限された特徴として提供するための方法のフローチャートである。
- 【図 21】ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。
- 【図 22】ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。 40
- 【図 23】ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。
- 【図 24】ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。
- 【図 25】ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。
- 【図 26】一実施形態による、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるための方法のフローチャートである。
- 【図 27】一実施形態による、高性能コンピューティング環境においてターゲット固有の `RDMA` 書込帯域幅制限と `RDMA` 読出帯域幅制限とを組み合わせるためのシステムを示 50

す。

【図 28】一実施形態による、高性能コンピューティング環境においてターゲット固有の R D M A 書込帯域幅制限と R D M A 読出帯域幅制限とを組み合わせるためのシステムを示す。

【図 29】一実施形態による、高性能コンピューティング環境においてターゲット固有の R D M A 書込帯域幅制限と R D M A 読出帯域幅制限とを組み合わせるためのシステムを示す。

【図 30】一実施形態による、高性能コンピューティング環境においてターゲット固有の R D M A 書込帯域幅制限と R D M A 読出帯域幅制限とを組み合わせるためのシステムを示す。

10

【図 31】一実施形態による、高性能コンピューティング環境においてターゲット固有の R D M A 書込帯域幅制限と R D M A 読出帯域幅制限とを組み合わせるための方法のフローチャートである。

【図 32】ある実施形態に係る、高性能コンピューティング環境においてインGRESS帯域幅調停および輻輳フィードバックを組み合わせるためのシステムを示す。

【図 33】ある実施形態に係る、高性能コンピューティング環境においてインGRESS帯域幅調停および輻輳フィードバックを組み合わせるためのシステムを示す。

【図 34】ある実施形態に係る、高性能コンピューティング環境においてインGRESS帯域幅調停および輻輳フィードバックを組み合わせるためのシステムを示す。

【図 35】一実施形態による、高性能コンピューティング環境においてインGRESS帯域幅調停と輻輳フィードバックとを組み合わせる方法のフローチャートである。

20

【図 36】ある実施形態に係る、高性能コンピューティング環境において F E C N および B E C N の両方において複数の C E フラグを用いるためのシステムを示す。

【図 37】ある実施形態に係る、高性能コンピューティング環境において F E C N および B E C N の両方において複数の C E フラグを用いるためのシステムを示す。

【図 38】ある実施形態に係る、高性能コンピューティング環境において F E C N および B E C N の両方において複数の C E フラグを用いるためのシステムを示す。

【図 39】ある実施形態に係る、高性能コンピューティング環境において F E C N および B E C N の両方において複数の C E フラグを用い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】

30

【0009】

詳細な説明：

本教示は、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、本精神および範囲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。

【0010】

40

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。

【0011】

本明細書では、高性能コンピューティング環境においてプライベートファブリックにおけるサービス品質 (Q O S) およびサービスレベル合意 (S L A) を提供するシステムおよび方法について説明する。

【0012】

ある実施形態によれば、本教示の以下の説明は、高性能ネットワークの一例として、イ

50

インフィニバンド<sup>TM</sup>（IB）ネットワークを使用する。以下の説明全体にわたり、インフィニバンド<sup>TM</sup>の仕様（インフィニバンド仕様、IB仕様、またはレガシーIB仕様など、さまざまな呼ばれ方がある）を引用することがある。このような引用は、2015年3月に発表され、<http://www.infinibandta.org>から入手可能な、本明細書にその全体を引用により援用するInfiniBand Trade Association Architecture Specification, Volume 1, Version 1.3を引用することであると理解される。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。

10

#### 【0013】

実施形態によれば、以下の説明は、RoCE（コンバージドイーサネットを介するRDMA（リモートダイレクトメモリアクセス））を用いる。コンバージドイーサネットを介するRDMA（RoCE）は、ハードウェアRDMAエンジン実装によるオフロードトランスポートおよび優れた性能を可能にする、イーサネットネットワークを介したRDMAの効率的なデータ転送を可能にする標準プロトコルである。RoCEは、InfiniBand Trade Association（IBTA）規格において定義される標準プロトコルである。RoCEは、UDP（ユーザデータグラムプロトコル）カプセル化を利用して、それがレイヤ3ネットワークを越えることができるようにする。RDMAは、インフィニバンド相互接続技術によってネイティブに使用される重要な能力である。インフィニバンドおよびイーサネットRoCEは両方とも、共通のユーザAPIを共有するが、異なる物理層およびリンク層を有する。

20

#### 【0014】

一実施形態によれば、本明細書のさまざまな部分は、さまざまな実現例を説明する際にインフィニバンドファブリックへの言及を含むが、当業者は、本明細書で説明されるさまざまな実施形態はRoCEファブリックにおいても実現され得ることを容易に理解するであろう。

#### 【0015】

今の時代（たとえばエクサスケール（exascale）時代）のクラウドの要求を満たすためには、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）等の低オーバーヘッドの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ため、シングルルートI/O仮想化（SR-IOV）ネットワークアダプタのようなパス・スルー技術を使用することができる。一実施形態に従うと、仮想スイッチ（vSwitch）SR-IOVアーキテクチャを、高性能無損失相互接続ネットワークに適用することができる。ネットワーク再構成時間はライブマイグレーションを現実的な選択肢にするために重要なので、ネットワークアーキテクチャに加えて、スケーラブルでありトポロジーに依存しない動的再構成機構を提供することができる。

30

#### 【0016】

一実施形態に従い、さらに、vSwitchを使用する仮想化環境に対するルーティングストラテジー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）に対する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。動的再構成機構をさらに調整することにより、ファットツリーに課されるオーバーヘッドを最小にすることができる。

40

#### 【0017】

本教示の一実施形態に従うと、仮想化は、クラウドコンピューティングにおける効率的なリソースの利用および柔軟なリソースの割当てにとって有益になり得る。ライブマイグレーションは、アプリケーションにとってトランスペアレントになるように物理サーバ間で仮想マシン（VM）を移動させることでリソースの利用を最適化することを可能にする。このように、仮想化は、ライブマイグレーションにより、コンソリデーション、リソー

50

スのオンデマンドプロビジョニング、および柔軟性を可能にする。

【0018】

インフィニバンド™

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。

【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。

【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】

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

【0023】

一般に、マスタ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、新しいマスタサブネットマネージャが取り決められる。マスタ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。

10

20

30

40

50

## 【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49151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、64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。

## 【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、トポロジーが変化するたびに、ルーティングテーブルは、接続性および最適性能を確実にするために更新され得る。通常動作中、S Mは、トポロジー変化をチェックするためにネットワークの周期的なライトスイープ（light sweep）を実行し得る。ライトスイープ中に変化が発見された場合、または、ネットワーク変化を信号で伝えるメッセージ（トラップ）を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。

## 【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ーが変化する場合に、ネットワークを再構成し得る。再構成ステップは、ネットワーク初期化中に行なわれるステップを含み得る。さらに、再構成は、ネットワーク変化が生じたサブネットに制限されるローカルスコープを有し得る。また、ルータを用いる大規模ファブリックのセグメント化は、再構成スコープを制限し得る。

## 【0027】

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

## 【0028】

インフィニバンドにおけるパーティショニング

一実施形態に従うと、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テムの論理グループを分離するためのセキュリティメカニズムとしての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H C Aポートは、1つ以上のパーティションのメンバである可能性がある。パーティションメンバーシップは、S M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。S M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\_\_K e y）のテーブルとして構成することができる。S M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\_\_K e y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L I D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。

## 【0029】

一実施形態に従うと、パーティションはポートの論理グループであり、あるグループのメンバは同じ論理グループの他のメンバとしか通信できない。ホストチャネルアダプタ（H C A）およびスイッチにおいて、パーティションメンバーシップ情報を用いてパケットをフィルタリングすることにより、分離を実施することができる。無効なパーティショニ

ング情報を有するパケットは、当該パケットが入口ポートに達すると直ち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システムにおいて、パーティションを用いることにより、テナントクラスタを作成できる。パーティションを適所で実施すると、ノードは異なるテナントクラスタに属する他のノードと通信することができない。このようにして、欠陥があるまたは悪意があるテナントノードが存在していても、システムのセキュリティを保証することができる。

#### 【0030】

一実施形態に従うと、ノード間の通信のために、マネージメントキューペア（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り当てることができる。次に、P\_Key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\_Key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\_Key値が見つかった場合、そのパケットは直ちに廃棄される。このようにして、通信は、パーティションを共有するポート間でのみ許可される。

#### 【0031】

一実施形態に従い、パーティショニングされたクラスタ環境の一例を示す図2に、IBパーティションの一例が示される。図2に示す例では、ノードA101～E105は、インフィニバンドファブリック120を使用して、それぞれのホストチャネルアダプタ111～115を介して通信する。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に配置されている。パーティション1はノードA 101とノードD 104とを含む。パーティション2はノードA 101とノードB 102とノードC 103とを含む。パーティション3はノードC 103とノードE 105とを含む。パーティションのこの配置により、ノードD 104およびノードE 105は、1つのパーティションを共有していないので、通信することができない。一方、たとえばノードA 101およびノードC 103は、どちらもパーティション2 140のメンバなので、通信することができる。

#### 【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 Root Input/Output Virtualization：SR-IOV）のようなデバイス・パススルー技術を使用することによってネットワークI/Oオーバーヘッドが削減されてきたことに応じて、仮想化された高性能コンピューティング（High Performance Computing：HPC）環境の将来の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。

#### 【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り当てされた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）に割り当てられたVMのライブマイグレーションを実現すること

10

20

30

40

50

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。

#### 【0035】

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

#### 【0036】

##### ファットツリー（Fat Tree）トポロジーおよびルーティング

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

#### 【0037】

図3は、一実施形態に従う、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。

#### 【0038】

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。

#### 【0039】

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

10

20

30

40

50

し、ルートスイッチ 2 3 3 はスイッチポート 5 ～ 6 を有し、ルートスイッチ 2 3 4 はスイッチポート 7 ～ 8 を有し得る。

#### 【 0 0 4 0 】

実施形態に従うと、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。

#### 【 0 0 4 1 】

ファットツリールーティングメカニズムの目的は、ネットワークファブリックにおけるリンクにわたって最短経路ルートを均一に広げる 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。

#### 【 0 0 4 2 】

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。

#### 【 0 0 4 3 】

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、 $k$ -a r y - n - t r e e、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。

#### 【 0 0 4 4 】

また、 $k$ -a r y - n - t r e e 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、 $k$ -a r y - n - t r e e 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。

#### 【 0 0 4 5 】

入出力 (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み合わせると、シングルサーバのI/Oリソースにとって抗し難い高い負荷が課され、結果として、データの待機中に、バックログが発生し、プロセッサがアイドル状態になる可能性がある。I/O要求の数が増えるにつれて、IOVにより利用可能性をもたらすことができ、最新のCPU仮想化において見られる性能レベルに匹敵するように、（仮想化された）I/Oリソースの性能、スケーラビリティおよび融通性を向上させることができる。

#### 【0046】

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

#### 【0047】

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

#### 【0048】

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

#### 【0049】

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

#### 【0050】

SR-IOVは、PCIeデバイスが、各々のゲストに1つの仮想デバイスを割当てることによって複数のゲスト間で共有することができる複数の仮想デバイスをエクスポーズすることを可能にする。各々のSR-IOVデバイスは、少なくとも1つの物理機能（PF）と、1つ以上の関連付けられた仮想機能（VF）とを有する。PFは、仮想マシンモニタ（virtual machine monitor：VMM）またはハイパーバイザによって制御される通常のPCIe機能であるのに対して、VFは軽量のPCIe機能である。各々のVFはそ

10

20

30

40

50

れ自体のベースアドレス (base address : B A R) を有しており、固有のリクエスト I D が割当てられている。固有のリクエスト I D は、I / O メモリ管理ユニット (I / O memory management unit : I O M M U) がさまざまな V F 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。I O M M U はまた、メモリを適用して、P F と V F との間の変換を中断する。

【 0 0 5 1 】

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、V M のメモリ内容がリモートハイパーバイザにコピーされるという点である。さらに、V M がソースハイパーバイザにおいて中断され、V M 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。

【 0 0 5 2 】

しかしながら、S R - I O V 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、V M に割当てられた S R - I O V V F 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 V F 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 S R - I O V 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ず可能性がある。さらに、S R - I O V 共有型ポートモデルにおいては、V M 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、S M 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。

【 0 0 5 3 】

#### インフィニバンド S R - I O V アーキテクチャ共有ポート

さまざまなタイプの S R - I O V 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。

【 0 0 5 4 】

図 4 は、一実施形態に従う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 3 0 0 (たとえばホストチャネルアダプタ) はハイパーバイザ 3 1 0 と対話し得る。ハイパーバイザ 3 1 0 は、さまざまな仮想機能 3 3 0、3 4 0 および 3 5 0 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 3 1 0 によって処理することができる。

【 0 0 5 5 】

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 H C A) は、物理機能 3 2 0 と仮想機能 3 3 0、3 5 0、3 5 0 との間において単一の共有 L I D および共有キュー対 (Queue Pair : Q P) 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 G I D を有し得る。

【 0 0 5 6 】

図 4 に示されるように、一実施形態に従うと、さまざまな G I D 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 Q P 0 および Q P 1 (すなわちインフィニバンド<sup>T M</sup>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 Q P は V F にも同様にエクスポートされるが、V F は Q P 0 を使用することが許可されておらず (V F から Q P 0 に向かって入来するすべての S M P が廃棄され)、Q P 1 は、P F が所有する実際の Q P 1 のプロキシとして機能し得る。

【 0 0 5 7 】

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

10

20

30

40

50

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L I Dスペースが消費されるだけであるからである。

【0058】

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L I Dが特定のハイパーバイザに関連付けられており、かつハイパーバイザ上に常駐するすべてのVM間で共有されているので、マイグレートしているVM(すなわち、宛先ハイパーバイザにマイグレートする仮想マシン)は、その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VMの内部で実行させることができなくなる。

【0059】

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

図5は、一実施形態に従う例示的なvSwitchアーキテクチャを示す。図に示されるように、ホスト400(たとえばホストチャネルアダプタ)はハイパーバイザ410と対話することができ、当該ハイパーバイザ410は、さまざまな仮想機能430、440および450をいくつかの仮想マシンに割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。

【0060】

一実施形態に従うと、vSwitchアーキテクチャにおいては、各々の仮想機能430、440、450は完全な仮想ホストチャネルアダプタ(virtual Host Channel Adapter: vHCA)であり、これは、ハードウェアにおいて、VFに割り当てられた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SMについては、HCA400は、仮想スイッチ415を介して追加のノードが接続されているスイッチのように見えている。ハイパーバイザ410はPF420を用いることができ、(仮想機能に付与された)VMはVFを用いる。

【0061】

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

【0062】

インフィニバンドSR-IOVアーキテクチャモデル-仮想ポート(vPort)

図6は、一実施形態に従う例示的なvPortの概念を示す。図に示されるように、ホスト300(たとえばホストチャネルアダプタ)は、さまざまな仮想機能330、340および350をいくつかの仮想マシンに割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。

【0063】

一実施形態に従うと、ベンダーに実装の自由を与えるためにvPort概念は緩やかに

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

#### 【0064】

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

10

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。

#### 【0065】

図7は、一実施形態に従う、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1550はハイパーバイザ511によって仮想機能1514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2551を仮想機能2515に割当て、仮想マシン3552を仮想機能3516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4553を仮想機能1534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。

20

30

#### 【0066】

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ーワーディングテーブルを設定するのに用いられる。

#### 【0067】

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。

40

#### 【0068】

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1534はLID10

50

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、L I DはS R - I O V v S w i t c h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。V FのすべてがネットワークにおけるV M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V Fには、図7に示されるようにL I Dが割当てられている。

#### 【0069】

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H C A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I Bサブネットに接続され得る。

#### 【0070】

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V M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、V M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S Mノードが同様にL I Dを消費するので、実際これらの数はより小さくなる）。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。

#### 【0071】

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V M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、V M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。

#### 【0072】

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V M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（L I D-Mask-Control-like：L M Cライク）な特徴を用いることを可能にする。V M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。

#### 【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。

10

20

30

40

50

## 【0074】

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

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

## 【0075】

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

## 【0076】

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。

## 【0077】

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。

## 【0078】

一実施形態に従うと、本開示は、動的LID割当てがなされたvSwitchアーキテクチャを提供するためのシステムおよび方法を提供する。図8を参照すると、LIDには、さまざまな物理機能513、523および533が動的に割当てられており、物理機能513がLID1を受取り、物理機能523がLID2を受取り、物理機能533がLID3を受取る。アクティブな仮想マシンに関連付けられたそれらの仮想機能はまた、動的に割当てられたLIDを受取ることもできる。たとえば、仮想マシン1 550がアクティブであり、仮想機能1 514に関連付けられているので、仮想機能514にはLID5が割当てられ得る。同様に、仮想機能2 515、仮想機能3 516および仮想機能1 534は、各々、アクティブな仮想機能に関連付けられている。このため、これらの仮想機能にLIDが割当てられ、LID7が仮想機能2 515に割当てられ、LID11が仮想機能3 516に割当てられ、LID9が仮想機能1 534に割当てられている。LIDが予めポピュレートされたvSwitchとは異なり、アクティブな仮想マシンにその時点で関連付けられていない仮想機能はLIDの割当てを受けない。

## 【0079】

一実施形態に従うと、動的 L I D 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、V M が存在していない場合、比較的少数の L I D を最初の経路計算および L F T 分配のために用いることができる。

【0080】

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 H C A 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 I B サブネットに接続され得る。

【0081】

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 V M 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 V M をブートすべきであるかを決定するために、自由な V M 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 V M 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 V M 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 V M 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。

【0082】

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、（V M 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。

【0083】

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 V M 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 V M が作成されると、次に利用可能な L I D が、V M 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 V M 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 V M が同じ L I D を得る。

【0084】

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。V M が作成されるたびに、作成された V M 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 V M 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 V M 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

10

20

30

40

50

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。

【 0 0 8 5 】

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、一実施形態に従う、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。図に示されるように、いくつかのスイッチ 5 0 1 ~ 5 0 4 は、ネットワーク切替環境 8 0 0 (たとえば I B サブネット) 内においてインフィニバンド<sup>T M</sup>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 5 1 0、5 2 0、5 3 0 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 5 1 0、5 2 0 および 5 3 0 は、それぞれ、さらに、ハイパーバイザ 5 1 1、5 2 1 および 5 3 1 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 5 1 4、5 1 5、5 1 6、5 2 4、5 2 5、5 2 6、5 3 4、5 3 5 および 5 3 6 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 5 5 0 は、ハイパーバイザ 5 1 1 によって仮想機能 1 5 1 4 に割当てることができる。ハイパーバイザ 5 1 1 は、加えて、仮想マシン 2 5 5 1 を仮想機能 2 5 1 5 に割当てることができる。ハイパーバイザ 5 2 1 は、仮想マシン 3 5 5 2 を仮想機能 3 5 2 6 に割当てることができる。ハイパーバイザ 5 3 1 は、さらに、仮想マシン 4 5 5 3 を仮想機能 2 5 3 5 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 5 1 3、5 2 3 および 5 3 3 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。

【 0 0 8 6 】

一実施形態に従うと、スイッチ 5 0 1 ~ 5 0 4 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 8 0 0 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。

【 0 0 8 7 】

一実施形態に従うと、仮想スイッチ 5 1 2、5 2 2 および 5 3 2 は、それぞれのハイパーバイザ 5 1 1、5 2 1、5 3 1 によって処理することができる。このような v S w i t c h アーキテクチャにおいては、各々の仮想機能は、完全な仮想ホストチャネルアダプタ (v H C A) であり、これは、ハードウェアにおいて、V F に割当てられた 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 (図示せず) については、H C A 5 1 0、5 2 0 および 5 3 0 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。

【 0 0 8 8 】

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 L I D が予めポピュレートされたハイブリッド 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。図 9 を参照すると、ハイパーバイザ 5 1 1 には、予めポピュレートされた L I D アーキテクチャを備えた v S w i t c h が配置され得るとともに、ハイパーバイザ 5 2 1 には、L I D が予めポピュレートされて動的 L I D 割当てがなされた v S w i t c h が配置され得る。ハイパーバイザ 5 3 1 には、動的 L I D 割当てがなされた v S w i t c h が配置され得る。このため、物理機能 5 1 3 および仮想機能 5 1 4 ~ 5 1 6 には、それらの L I D 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 L I D が割当てられている)。物理機能 5 2 3 および仮想機能 1 5 2 4 にはそれらの L I D が予めポピュレートされ得るとともに、仮想機能 2 5 2 5 および仮想機能 3 5 2 6 にはそれらの L I D が動的に割当てられている (すなわち、仮想機能 2 5 2 5 は動的 L I D 割当てのために利用可能であり、仮想機能 3 5 2 6 は、仮想マシン 3 5 5 2 が付与されているので、1 1 という L I D が動的に割当てられている)。最後に、ハイパーバイザ 3 5 3 1 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L I D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L I D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L I Dが動的に割当てられている。

【0089】

L I Dが予めポピュレートされたv S w i t c hおよび動的L I D割当てがなされたv S w i t c hがともに(いずれかの所与のハイパーバイザ内で独立して、または組合わされて)利用されている、図9に示されるような一実施形態に従うと、ホストチャネルアダプタごとの予めポピュレートされたL I D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)  $0 \leq \text{予めポピュレートされたV F} \leq \text{総V F}$ の範囲内になり得る。動的L I D割当てのために利用可能なV Fは、(ホストチャネルアダプタごとに) V Fの総数から予めポピュレートされたV Fの数を減じることによって見出すことができる。

【0090】

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H C A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I Bサブネットに接続され得る。

【0091】

#### 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)

一実施形態に従うと、1つの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、本開示の実施形態は、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。

【0092】

図10は、一実施形態に従う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。この図に示されるように、サブネットA 1000の内部の多数のスイッチ1001~1004は、サブネットA 1000(たとえばI Bサブネット)内における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の通信を提供することができる。このファブリックは、たとえばチャネルアダプタ1010などの多数のハードウェアデバイスを含み得る。ホストチャネルアダプタ1010は、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、対話の相手であるホストチャネルアダプタとともに、多数の仮想機能1014をセットアップすることができる。加えて、ハイパーバイザは、仮想マシンを仮想機能各々に割当てることができる。たとえば、仮想マシン1 1015は仮想機能1 1014に割当てられる。ハイパーバイザは、その対応付けられたホストチャネルアダプタに、各ホストチャネルアダプタ上の物理機能1013などの十分な機能を有する物理機能を通して、アクセスすることができる。多数のスイッチ1021~1024は、サブネットB 1040(たとえばI Bサブネット)内における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の通信を提供することができる。このファブリックは、たとえばホストチャネルアダプタ1030などの多数のハードウェアデバイスを含み得る。ホストチャネルアダプタ1030は、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、対話の相手であるホストチャネルアダプタとともに、多数の仮想機能1034をセットアップすることができる。加えて、ハイパーバイザは、仮想マシンを仮想機能各々に割当てることができる。たとえば、仮想マシン2 1035は仮想機能2 1034に割当てられる。ハイパーバイザは、その対応付けられたホストチャネルアダプタに、各ホストチャネルアダプタ上の物理機能1033などの十分な機能を有する物理機能を通して、アクセスすることができる。なお、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)内に示されているホストチャネルアダプタは1つだけであるが、各サブネット内に複数のホストチャネルアダプタおよびそれらに対応するコンポーネントが含まれていてもよいことが、理解されるはずである。

【0093】

一実施形態に従うと、各ホストチャネルアダプタはさらに、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対応付けられていてもよく、上記のように各HCAは異なるアーキテクチャモデルでセットアップされてもよい。図10のサブネットはどちらもLIDが予めポピュレートされているvSwitchのアーキテクチャモデルを使用するものとして示されているが、これは、このようなサブネット構成すべてが同様のアーキテクチャモデルに従い得ることを示唆しようとしているのではない。

【0094】

一実施形態に従うと、各サブネット内の少なくとも1つのスイッチがルータに対応付けられていてもよい。たとえば、サブネットA 1000内のスイッチ1002はルータ1005に対応付けられ、サブネットB 1040内のスイッチ1021はルータ1006に対応付けられている。

10

【0095】

一実施形態に従うと、少なくとも1つのデバイス（たとえばスイッチ、ノード等）を、ファブリックマネージャ（図示せず）に対応付けることができる。ファブリックマネージャを使用して、たとえば、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（たとえば仮想マシンファブリックプロファイル）を作成し、仮想マシンファブリックプロファイルを構築するための基礎を形成する仮想マシン関連データベースオブジェクトを構築することができる。加えて、ファブリックマネージャは、どのサブネットがどのルータポートを介しどのパーティション番号を用いて通信することを許可されるかについて、法的なサブネット間接続性を規定することができる。

20

【0096】

一実施形態に従うと、サブネットA内の仮想マシン1などの発信ソースにおけるトラフィックを、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を宛先としてそれに向ける場合、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に向けられ、そうすると、ルータ1005はこのトラフィックをルータ1006とのリンクを介してサブネットBに送ることができる。

【0097】

#### 仮想デュアルポートルータ

一実施形態に従うと、デュアルポートルータアブストラクション（dual port router abstraction）は、GRH（グローバルルートヘッダ（global route header））からLRH（ローカルルートヘッダ（local route header））への変換を、通常のLRHベースのスイッチングの実行に加えて行なう機能を有するスイッチハードウェア実装に基づいてサブネット間ルータ機能を規定することを可能にする簡単な方法を提供することができる。

30

【0098】

一実施形態に従うと、仮想デュアルポートルータは、対応するスイッチポートの外部で論理的に接続することができる。この仮想デュアルポートルータは、サブネットマネージャ等の標準管理エンティティに対しインフィニバンド規格に準拠したビューを提供することができる。

【0099】

一実施形態に従うと、デュアルポートルータモデルは、異なるサブネットを、各サブネットがサブネットへの進入（ingress）経路におけるパケット転送とアドレスマッピングとを完全に制御し、かつ、間違って接続されたサブネットのうちいずれのサブネット内のルーティングおよび論理的接続にも影響を与えないように、接続できることを、示している。

40

【0100】

一実施形態に従うと、間違って接続されたファブリックを含む状況において、仮想デュアルポートルータアブストラクションを使用することにより、サブネットマネージャおよびIB診断ソフトウェア等の管理エンティティが、遠隔サブネットへの意図しない物理的接続の存在下で、正しく作用するようにすることもできる。

【0101】

50

図11は、一実施形態に従う、高性能コンピューティング環境における2つのサブネット間の相互接続を示す。仮想デュアルポートルータを用いて構成する前に、サブネットA 1101内のスイッチ1120を、スイッチ1120のスイッチポート1121を通し、物理接続1110を介して、サブネットB 1102内のスイッチ1130に、スイッチ1130のスイッチポート1131を通して接続することができる。このような実施形態において、スイッチポート1121および1131の各々は、スイッチポートとしてもルータポートとしても機能することができる。

#### 【0102】

一実施形態に従うと、この構成の問題は、インフィニバンドサブネット内のサブネットマネージャ等の管理エンティティが、スイッチポートでもありルータポートでもある物理ポートを区別できないことである。このような状況において、SMは、スイッチポートを、このスイッチポートに接続されたルータポートを有するものとして扱うことができる。しかしながら、スイッチポートがたとえば物理リンクを介して別のサブネットマネージャを有する別のサブネットに接続されている場合、サブネットマネージャはディスカバリメッセージを物理リンクに送ることができる。しかしながら、このようなディスカバリメッセージは他方のサブネットでは許可されない。

#### 【0103】

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

#### 【0104】

一実施形態に従うと、構成後に、デュアルポート仮想ルータ構成を、サブネットマネージャの責任であるサブネットの端部を示す適切なエンドノードが、サブネットマネージャにわかるように、提供することができる。

#### 【0105】

一実施形態に従うと、サブネットA 1201内のスイッチ1220におけるスイッチポートは、仮想リンク1223を介して仮想ルータ1210内のルータポート1211に接続（すなわち論理的に接続）することができる。仮想ルータ1210（たとえばデュアルポート仮想ルータ）は、実施形態ではスイッチ1220の外部にあるものとして示されているが、論理的にはスイッチ1220の中を含めることができ、第2のルータポートであるルータポートII 1212も含み得る。一実施形態に従うと、2つの端部を有し得る物理リンク1203は、サブネットA 1201を、サブネットB 1202に、物理リンクの第1の端部を介し、物理リンクの第2の端部を介し、ルータポートII 1212を介し、サブネットB 1202内の仮想ルータ1230に含まれるルータポートII 1232を介して、接続することができる。仮想ルータ1230はさらに、仮想リンク1233を介してスイッチ1240上のスイッチポート1241に接続（すなわち論理的に接続）することができるルータポート1231を含み得る。

#### 【0106】

一実施形態に従うと、サブネットA上のサブネットマネージャ（図示せず）は、仮想ルータ1210上のルータポート1211を、当該サブネットマネージャが制御するサブネットの終点として検出することができる。デュアルポート仮想ルータアブストラクションは、サブネットA上のサブネットマネージャが、サブネットAを通常のやり方で（たとえばインフィニバンド規格に規定されているように）扱うことを可能にする。サブネット管理エージェント（subnet management agent）レベルにおいて、デュアルポート仮想ルータアブストラクションを提供して通常のスイッチポートがSMにわかるようにし、その後、SMAレベルにおいて、当該アブストラクションを提供してこのスイッチポートに接続されている別のポートが存在しこのポートがデュアルポート仮想ルータ上のルータポートとなるようにすることができる。ローカルSMでは、従来のファブリックトポロジを引続き使用することができ（このトポロジにおいてSMはポートを標準スイッチポートとみなす）、したがって、SMはルータポートをエンドポートとみなす。物理的接続は、2つの異なるサブネット内のルータポートとしても構成されている2つのスイッチポート

10

20

30

40

50

間で行なうことができる。

【0107】

一実施形態に従うと、デュアルポート仮想ルータは、物理リンクが間違っ

て同じサブネット内の他のいずれかのスイッチポートに接続される、または、別のサブネットへの接続を提供することを意図していないスイッチポートに接続される可能性があるという問題を、解決することもできる。したがって、本明細書に記載の方法およびシステムは、サブネットの外側にあるものも表現する。

【0108】

一実施形態に従うと、サブネットA等のサブネット内のローカルSMは、スイッチポートを確定し、次に、このスイッチポートに接続されているルータポート（たとえば仮想リンク1223を介してスイッチポート1221に接続されているルータポート1211）を確定する。SMは、ルータポート1211を、当該SMが管理するサブネットの端部とみなすので、SMはディスカバリおよび／または管理メッセージをこのポイントよりも遠くに（たとえばルータポートII 1212に）送ることができない。

【0109】

一実施形態に従うと、上記デュアルポート仮想ルータは、当該デュアルポート仮想ルータが属するサブネット内の管理エンティティ（たとえばSMまたはSMA）によってデュアルポート仮想ルータアブストラクションが完全に管理されるという利点を提供する。管理をローカル側のみにすることにより、システムは外部の独立した管理エンティティを提供する必要がない。すなわち、サブネット間接続の各側は自身のデュアルポート仮想ルータを構成する役割を担う。

【0110】

一実施形態に従うと、遠隔の宛先（すなわちローカルサブネットの外部）に向けられたSMP等のパケットが、上記デュアルポート仮想ルータを介して構成されていないローカルターゲットポートに到着した場合、ローカルポートは、自身はルータポートではないことを示すメッセージを返すことができる。

【0111】

本開示の多数の特徴は、ハードウェア、ソフトウェア、ファームウェア、またはこれらを組合わせたものにおいて、これを用いて、またはこれに支援されて、実施することができる。したがって、本開示の特徴は、処理システム（たとえば1つ以上のプロセッサを含む）を用いて実現し得る。

【0112】

図13は、一実施形態に従う、高性能コンピューティング環境においてデュアルポート仮想ルータをサポートする方法を示す。ステップ1310において、1つ以上のマイクロプロセッサを含む1つ以上コンピュータに、この方法は第1のサブネットを設けることができる。第1のサブネットは複数のスイッチを含み、複数のスイッチは少なくともリーフスイッチを含み、複数のスイッチの各々は複数のスイッチポートを含む。第1のサブネットはさらに、各々が少なくとも1つのホストチャネルアダプタポートを含む複数のホストチャネルアダプタと、各々が複数のホストチャネルアダプタのうちの少なくとも1つのホストチャネルアダプタに対応付けられている複数のエンドノードと、サブネットマネージャとを含み、サブネットマネージャは、複数のスイッチおよび複数のホストチャネルアダプタの一方において実行される。

【0113】

ステップ1320において、この方法は、複数のスイッチのうちの1つのスイッチ上の複数のスイッチポートのうちの1つのスイッチポートを、ルータポートとして構成することができる。

【0114】

スイッチ1330において、この方法は、ルータポートとして構成したスイッチポートを仮想ルータに論理的に接続することができ、この仮想ルータは少なくとも2つの仮想ルータポートを含む。

## 【0115】

プライベートファブリックにおけるサービス品質およびサービスレベル合意

ある実施形態に係る、インフィニバンドまたはR o C E上で動作するスイッチドネットワークなどの、クラウド内ならびに顧客および構内設置におけるより大きいクラウド内の高性能コンピューティング環境は、仮想マシン（VM）ベースのワークロードを展開する能力があり、ある固有の要件は、異なるタイプの通信フローに対してサービス品質（Q O S）を定義および制御することが可能であることである。加えて、異なるテナントに属するワークロードは、そのようなワークロード間の干渉を最小限に抑え、異なる通信タイプに対するQ O S仮定を維持しながら、関連するサービスレベル合意（S L A）の境界内で実行されなければならない。

10

## 【0116】

制限された特徴としてのR D M A読出（O R A 2 0 Q 2 4 6 - U S - N P - 1）

ある実施形態に従うと、従来のネットワークインターフェイス（N I C）を用いるシステムにおいて帯域幅制限を定義する場合、一般に、各ノード／VMがネットワーク上に生成することを許可されるエグレス帯域幅を制御することで、充分である。

## 【0117】

しかしながら、ある実施形態に従うと、異なるノードがR D M A読出要求（すなわち、エグレス帯域幅）を生成することができるR D M Aベースのネットワーキングでは、これは、少量のエグレス帯域幅を表し得る。しかしながら、そのようなR D M A読出要求は、そのようなR D M A読出要求に応答して非常に大量のイングレスR D M Aトラフィックを表す潜在性があり得る。そのような状況では、システムにおいて全体のトラフィック生成を制御するためにすべてのノード／VMのエグレス帯域幅を制限することは、もはや充分ではない。

20

## 【0118】

ある実施形態に従うと、R D M A読出動作を、制限された特徴とし、過剰なR D M A読出ベースのイングレス帯域幅を生成しないと信頼できるノード／VMについてのみそのような読出要求を可能にすることによって、信頼できないノード／VMについての送信（エグレス）帯域幅の制限のみを行いながら総帯域幅利用を制限することが可能である。

## 【0119】

図14は、一実施形態による、高性能コンピューティング環境において、R D M A読出要求を、制限された特徴として提供するためのシステムを示す。

30

## 【0120】

より具体的には、ある実施形態に従うと、図14は、ハイパーバイザ1411を含むホストチャネルアダプタ1401を示す。ハイパーバイザは、V F 1 4 1 4 ~ 1 4 1 6などのいくつかの仮想機能（V F）、および物理機能（P F）1413をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク1400などのネットワークに接続するために使用されるポート1402および1403などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、H C A 1 4 0 1を、スイッチ、追加の別個のH C Aなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはR o C Eネットワークなどのスイッチドネットワークを備えることができる。

40

## 【0121】

ある実施形態に従うと、上述したように、仮想機能の各々は、V M 1 1 4 5 0、V M 2 1 4 5 1、およびV M 3 1 4 5 2のような仮想マシン（VM）をホストすることができる。

## 【0122】

ある実施形態に従うと、ホストチャネルアダプタ1401はさらに、ハイパーバイザを介して仮想スイッチ1412をサポートすることができる。これは、v S w i t c hアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（v P o r t）アーキテクチャをさらにサポートすることがで

50

きる。

【0123】

ある実施形態に従うと、ホストチャネルアダプタは、トラステッドRDMA読出制限1460を実現することができ、それにより、読出制限1460は、仮想マシン（例えば、VM1、VM2、および／またはVM3である）のいずれかが（たとえばポート1402または1403を介して）任意のRDMA読出要求をネットワークに送出することをブロックするように構成することができる。

【0124】

ある実施形態に従うと、トラステッドRDMA読出制限1460は、仮想マシンなどのあるエンドポイントからのあるタイプのパケットまたはネットワークに接続するためにHCA1401を利用する他の物理ノードからのあるタイプのパケットがRDMA読出要求パケットを生成（すなわちエグレス）することを、ホストチャネルアダプタレベルでブロックすることを実現することができる。この構成可能な制限コンポーネント1460は、たとえば、トラステッドノード（たとえば、VMまたは物理エンドノード）のみがそのようなタイプのパケットを生成することを可能にすることができる。

【0125】

ある実施形態に従うと、トラステッドRDMA読出制限コンポーネントは、たとえばホストチャネルアダプタによって受信された命令に基づいて構成されることができ、またはたとえばサブネットマネージャ（図示せず）によって直接構成されることができ。

【0126】

図15は、一実施形態による、高性能コンピューティング環境において、RDMA読出要求を、制限された特徴として提供するためのシステムを示す。

【0127】

より具体的には、ある実施形態に従うと、図15は、ハイパーバイザ1511を含むホストチャネルアダプタ1501を示す。ハイパーバイザは、VF1514～1516などのいくつかの仮想機能（VF）、および物理機能（PF）1513をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、加えて、ホストチャネルアダプタをネットワーク1500等のネットワークに接続するために使用される、ポート1502および1503等のいくつかのポートをサポートするかまたは備えることができる。ネットワークは、たとえば、HCA1501を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

【0128】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 1550、VM2 1551、およびVM3 1552のような仮想マシン（VM）をホストすることができる。

【0129】

ある実施形態に従うと、ホストチャネルアダプタ1501はさらに、ハイパーバイザを介して仮想スイッチ1512をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（vPort）アーキテクチャをさらにサポートすることができる。

【0130】

ある実施形態に従うと、ホストチャネルアダプタは、トラステッドRDMA読出制限1560を実現することができ、それにより、読出制限1560は、仮想マシン（例えば、VM1、VM2、および／またはVM3である）のいずれかが（たとえばポート1502または1503を介して）任意のRDMA読出要求をネットワークに送出することをブロックするように構成することができる。

【0131】

ある実施形態に従うと、トラステッドRDMA読出制限1560は、仮想マシンなどの

10

20

30

40

50

あるエンドポイントからのあるタイプのパケットまたはネットワークに接続するために H C A 1 5 0 1 を利用する他の物理ノードからのあるタイプのパケットが R D M A 読出要求パケットを生成（すなわちエグレス）することを、ホストチャネルアダプタレベルでブロッキングすることを実現することができる。この構成可能な制限コンポーネント 1 5 6 0 は、たとえば、トラステッドノード（たとえば、V M または物理エンドノード）のみがそのようなタイプのパケットを生成することを可能にすることができる。

【 0 1 3 2 】

ある実施形態に従うと、トラステッド R D M A 読出制限コンポーネントは、たとえばホストチャネルアダプタによって受信された命令に基づいて構成されることができるか、またはたとえばサブネットマネージャ（図示せず）によって直接構成されることができる。

10

【 0 1 3 3 】

ある実施形態に従うと、一例として、トラステッド R D M A 読出制限 1 5 6 0 は、V M 1 1 5 5 0 を信頼し、V M 2 1 5 5 1 を信頼しないように構成することができる。したがって、V M 1 から開始する R D M A 読出要求 1 5 5 4 は許可され得、V M 2 から開始する R D M A 読出要求 1 5 5 5 は、それが出てホストチャネルアダプタ 1 5 0 1 に着く前に阻止されることができる（図 1 5 内では H C A の外側に示されているが、これは単に図面の便宜上のものである）。

【 0 1 3 4 】

図 1 6 は、一実施形態による、高性能コンピューティング環境において、R D M A 読出要求を、制限された特徴として提供するためのシステムを示す。

20

【 0 1 3 5 】

ある実施形態に従うと、スイッチドネットワークまたはサブネット 1 6 0 0 のような高性能コンピューティング環境内では、いくつかのエンドノード 1 6 0 1 および 1 6 0 2 は、リーフスイッチ 1 6 1 1 および 1 6 1 2、スイッチ 1 6 2 1 および 1 6 2 2、ならびにルートスイッチ 1 6 3 1 および 1 6 3 2 のようないくつかのスイッチを介して相互接続されるいくつかの仮想マシン V M 1 ~ V M 4 1 6 5 0 ~ 1 6 5 3 をサポートすることができる。

【 0 1 3 6 】

ある実施形態に従うと、ノード 1 6 0 1 および 1 6 0 2 の接続のための機能を提供するさまざまなホストチャネルアダプタ、ならびにサブネットに接続されるべき仮想マシンは図示されていない。そのような実施形態の議論は S R - I O V に関して上記されており、各仮想マシンはホストチャネルアダプタ上のハイパーバイザの仮想機能に関連付けられ得る。

30

【 0 1 3 7 】

一実施形態によれば、一般的なシステムでは、エンドノードをサブネットに接続する任意のリンクの帯域幅を任意の 1 つの仮想マシンが独占することを防止するために、R D M A エグレス帯域幅がエンドノードからの任意の 1 つの仮想マシンから制限される。しかしながら、そのようなエグレス帯域幅制限は、一般的なケースでは効果的であるが、仮想マシンが R D M A 読出要求 1 6 5 4 および 1 6 5 5 などの R D M A 読出要求を発行することを妨げない。これは、そのような R D M A 読出要求が一般に小さいパケットであり、エグレス帯域幅をほとんど利用しないためである。

40

【 0 1 3 8 】

しかしながら、ある実施形態に従うと、このような R D M A 読出要求は、V M 1 および V M 3 などの発行エンティティへの大量のリターントラフィックの生成をもたらし得る。そのような状況では、次いで、R D M A 読出要求は、たとえば、読出要求 1 6 5 4 が、宛先における読出要求の実行の結果として V M 1 に流れ戻る大量のデータトラフィックをもたらすとき、リンク輻輳およびネットワーク性能の低下につながり得る。

【 0 1 3 9 】

ある実施形態に従うと、特に複数のテナントがサブネット 1 6 0 0 を共有する状況において、これは、サブネットのパフォーマンスの損失につながり得る。

50

## 【0140】

ある実施形態に従うと、各ノード（またはホストチャネルアダプタ）は、任意のVMが信頼されない場合に、そのVMに、RDMA読出要求を発行しないようブロックを配置するRDMA読出制限1660および1661を用いて構成することができる。そのようなRDMA読出制限は、RDMA読出要求の発行に対する常時ブロックから、RDMA読出要求制限とともに構成される仮想マシンがRDMA読出要求を発行することができる（たとえば、低速ネットワークトラフィック期間中）に時間フレームを置く制限まで、変動し得る。加えて、RDMA読出制限1660および1661は、さらに、トラステッドVMがRDMA読出要求を発行することを可能にし得る。

## 【0141】

一実施形態によれば、複数のVM／テナントが「新式の」HCA、すなわち関連する新たな特徴に対するサポートを有するHCAを共有しているが、そのようなサポートを持たないリモートの「旧式の」HCAに対してRDMA要求を実行しているシナリオを有することが考えられるので、そのようなVMが「旧式の」RDMA読出応答側HCA上の静的レート構成に頼ることなくRDMA読出応答に関して生成することができるインGRESS帯域幅を制限する方法を有することは意味があるであろう。VMが「任意の」RDMA読出サイズを生成することを許される限り、これを行うための単純な方法はない。また、原則としてある期間にわたって生成される複数のRDMA読出要求はすべて同時に応答データを受信する場合があるので、単一の要求において生成され得るRDMA読出サイズに対する制限、および同じvHCAポートからの未処理のRDMA読出要求の総数に対する制限の両方がない限り、インGRESS帯域幅が非常に限られた時間を超えて最大帯域幅を超えることはできないことを保証することは不可能である。

## 【0142】

したがって、ある実施形態に従うと、vHCAについて最大読出サイズが定義されている場合、帯域幅制御は、すべての未処理の読出サイズの合計に対する割当てに基づいてよいし、または、より単純なスキームは、「最悪の場合の」読出サイズに基づいて未処理のRDMA読出の最大数を単に制限することであってもよい。したがって、いずれの場合も、短い間隔内のピーク帯域幅に対する制限（HCAポート最大リンク帯域幅を除く）はないが、そのようなピーク帯域幅「ウィンドウ」の持続時間は制限されることになる。しかしながら、加えて、データを有する応答が同じレートで受信されると仮定すると、要求の送信レートが最大許容インGRESSレートを超えないように、RDMA読出要求の送信レートも、スロットリングされなければならない。言い換えれば、最大未処理要求制限は、最悪の場合の短い間隔帯域幅を定義し、要求送信レート制限は、新たな要求は、応答が受信されると直ちに生成されることはできず、RDMA読出応答のための許容可能な平均インGRESS帯域幅を表す関連の遅延の後にのみ、生成されることができることを保証するであろう。したがって、最悪の場合、許可された数の要求は、いかなる応答も伴わずに送信され、次いで、これらの応答はすべて、「同じ時間」に受信される。この時点で、次の要求は、第1の応答が到着したときに直ちに送信され得るが、次の要求は、指定された遅延期間の間遅延されなければならないことになる。したがって、経時的に、平均インGRESS帯域幅は、要求レートが定義するものを超えることはできない。しかしながら、未処理の要求の最大数が小さいほど、可能な「変動性」が低減されることになる。

## 【0143】

明示的なRDMA読出帯域幅制限の使用（ORA200246-US-NP-1）

ある実施形態に従うと、従来のネットワークインターフェイス（NIC）を用いるシステムにおいて帯域幅制限を定義する場合、一般に、各ノード／VMがネットワーク上に生成することを許可されるEGRESS帯域幅を制御することで、充分である。

## 【0144】

しかしながら、ある実施形態に従うと、異なるノードが、小さな要求メッセージを表すが潜在的に非常に大きな応答メッセージを表すRDMA読出要求を生成することができる、RDMAベースのネットワーキングでは、システムにおける総トラフィック生成を制御

10

20

30

40

50

するためにすべてのノード／VMのエグレス帯域幅を制限することはもはや充分ではない。

【0145】

ある実施形態に従うと、どのような送信／エグレス帯域幅制限からも独立して、どの程度のRDMA読出イングレス帯域幅をノード／VMが生成することが許されるかについての明示的な割当を定義することにより、信頼できないノード／VMについてのRDMA読出の使用を制限することに頼ることなく、システムにおける総トラフィック生成を制御することが可能である。

【0146】

ある実施形態に従うと、システムおよび方法は、ローカルに生成されたRDMA読出要求に起因する平均イングレス帯域幅利用をサポートすることに加えて、最悪の場合の最大リンク帯域幅バーストの持続時間／長さ（すなわち、RDMA読出応答が「山積する」結果としての）をサポートすることができる。

【0147】

図17は、ある実施形態に係る、高性能コンピューティング環境において明示的なRDMA読出帯域幅制限を提供するためのシステムを示す。

【0148】

より具体的には、ある実施形態に従うと、図17は、ハイパーバイザ1711を含むホストチャンネルアダプタ1701を示す。ハイパーバイザは、VF1714～1716などのいくつかの仮想機能（VF）、および物理機能（PF）1713をホストするかまたはそれらに関連付けられることができる。ホストチャンネルアダプタは、ホストチャンネルアダプタをネットワーク1700などのネットワークに接続するために使用されるポート1702および1703などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA1701を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

【0149】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 1750、VM2 1751、およびVM3 1752のような仮想マシン（VM）をホストすることができる。

【0150】

ある実施形態に従うと、ホストチャンネルアダプタ1701はさらに、ハイパーバイザを介して仮想スイッチ1712をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（vPort）アーキテクチャをさらにサポートすることができる。

【0151】

ある実施形態に従うと、ホストチャンネルアダプタは、RDMA読出制限1760を実現することができる。それにより、読出制限1760は、（HCA1701の）任意のVMが、特定のVMによって送出されるRDMA読出要求に対する応答に関して生成することができるイングレス帯域幅の量に、割当を課すように構成されることができる。そのようなイングレス帯域幅を制限することは、ホストチャンネルアダプタにおいてローカルに実行される。

【0152】

ある実施形態に従うと、RDMA読出制限コンポーネントは、たとえばホストチャンネルアダプタによって受信された命令に基づいて構成されることができるか、またはたとえばサブネットマネージャ（図示せず）によって直接構成されることができる。

【0153】

図18は、ある実施形態に係る、高性能コンピューティング環境において明示的なRDMA読出帯域幅制限を提供するためのシステムを示す。

【0154】

10

20

30

40

50

より具体的には、ある実施形態に従うと、図18は、ハイパーバイザ1811を含むホストチャネルアダプタ1801を示す。ハイパーバイザは、VF1814~1816などのいくつかの仮想機能(VF)、および物理機能(PF)1813をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク1800などのネットワークに接続するために使用されるポート1802および1803などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA1801を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

【0155】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 1850、VM2 1851、およびVM3 1852のような仮想マシン(VM)をホストすることができる。

【0156】

ある実施形態に従うと、ホストチャネルアダプタ1801はさらに、ハイパーバイザを介して仮想スイッチ1812をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート(vPort)アーキテクチャをさらにサポートすることができる。

【0157】

ある実施形態に従うと、ホストチャネルアダプタは、RDMA読出制限1860を実現することができる、それにより、読出制限1860は、(HCA1701の)任意のVMが、特定のVMによって送出されるRDMA読出要求に対する応答に関して生成することができるイングレス帯域幅の量に、割当を課すように構成されることができる。そのようなイングレス帯域幅を制限することは、ホストチャネルアダプタにおいてローカルに実行される。

【0158】

ある実施形態に従うと、RDMA読出制限コンポーネントは、たとえばホストチャネルアダプタによって受信された命令に基づいて構成されることができるか、またはたとえばサブネットマネージャ(図示せず)によって直接構成されることができる。

【0159】

ある実施形態に従うと、たとえば、VM1は、少なくとも2つのRDMA読出要求を以前に送出し、読出動作が、接続されたノード上で実行されるよう、要求することができる。これに回答して、VM1は、RDMA読出応答1855および1854として図に示されている、RDMA読出要求に対する複数の応答を受信するプロセスにあることができる。これらのRDMA読出応答は、特にVM1によって最初に送信されたRDMA読出要求と比較した場合に非常に大きくなり得るので、これらの読出応答1854および1855は、RDMA読出制限1860の対象となり得、イングレス帯域幅は制限またはスロットリングされ得る。このスロットリングは、明示的なイングレス帯域幅制限に基づくことができ、またはRDMA制限1860内に設定されたVM1のQoSおよび/またはSLAに基づくことができる。

【0160】

図19は、ある実施形態に係る、高性能コンピューティング環境において明示的なRDMA読出帯域幅制限を提供するためのシステムを示す。

【0161】

ある実施形態に従うと、スイッチドネットワークまたはサブネット1900のような高性能コンピューティング環境内では、いくつかのエンドノード1901および1902は、リーフスイッチ1911および1912、スイッチ1921および1922、ならびにルートスイッチ1931および1932のようないくつかのスイッチを介して相互接続されるいくつかの仮想マシンVM1~VM4 1950~1953をサポートすることがで

10

20

30

40

50

きる。

#### 【0162】

ある実施形態に従うと、ノード1901および1902の接続のための機能を提供するさまざまなホストチャネルアダプタ、ならびにサブネットに接続されるべき仮想マシンは図示されていない。そのような実施形態の議論はSRIOVに関して上記されており、各仮想マシンはホストチャネルアダプタ上のハイパーバイザの仮想機能に関連付けられ得る。

#### 【0163】

一実施形態によれば、一般的なシステムでは、エンドノードをサブネットに接続する任意のリンクの帯域幅を任意の1つの仮想マシンが独占することを防止するために、RDMAエグレス帯域幅がエンドノードからの任意の1つの仮想マシンから制限される。しかしながら、そのようなエグレス帯域幅制限は、一般的なケースでは効果的であるが、RDMA読出応答の流入が、要求元VMとネットワークとの間でリンクを独占することを防止することはできない。

#### 【0164】

別の言い方をすれば、ある実施形態によれば、VM1がいくつかのRDMA読出要求を送出する場合、VM1は、そのような読出要求に対する応答がVM1にいつ返されるかを制御することはできない。これは、RDMA読出要求に対する応答のバックアップ／山積をもたらし得、各々は、(RDMA読出応答1954を介して)要求された情報をVM1に戻すために同じリンクを用いることを試みる。これは、ネットワークにおけるトラフィックの輻輳および未処理分をもたらす。

#### 【0165】

ある実施形態に従うと、RDMA制限1960および1961は、あるVMが特定のVMによって送られるRDMA読出要求に対する応答に関して生成することができるイングレス帯域幅の量に割当を課することができる。そのようなイングレス帯域幅を制限することは、ローカルで実行される。

#### 【0166】

ある実施形態に従うと、vHCAについて最大読出サイズが定義されている場合、帯域幅制御は、すべての未処理の読出サイズの合計に対する割当に基づき得るか、または、より単純なスキームは、「最悪の場合の」読出サイズに基づいて未処理のRDMA読出の最大数を単に制限することになり得る。したがって、いずれの場合も、短い間隔内のピーク帯域幅に対する制限(HCAポート最大リンク帯域幅を除く)はないが、そのようなピーク帯域幅「ウィンドウ」の持続時間は制限されることになる。しかしながら、加えて、データを有する応答が同じレートで受信されると仮定すると、要求の送信レートが最大許容イングレスレートを超えないように、RDMA読出要求の送信レートも、スロットリングされなければならない。言い換えれば、最大未処理要求制限は、最悪の場合の短い間隔帯域幅を定義し、要求送信レート制限は、新たな要求は、応答が受信されると直ちに生成されることはできず、RDMA読出応答のための許容可能な平均イングレス帯域幅を表す関連の遅延の後にのみ、生成されることができるとを保証するであろう。したがって、最悪の場合、許可された数の要求は、いかなる応答も伴わずに送信され、次いで、これらの応答はすべて、「同じ時間」に受信される。この時点で、次の要求は、第1の応答が到着したときに直ちに送信され得るが、次の要求は、指定された遅延期間の間遅延されなければならないことになる。したがって、経時的に、平均イングレス帯域幅は、要求レートが定義するものを超えることはできない。しかしながら、未処理の要求の最大数が小さいほど、可能な「変動性」が低減されることになる。

#### 【0167】

図20は、一実施形態による、高性能コンピューティング環境において、RDMA(リモートダイレクトメモリアクセス)読出要求を、制限された特徴として提供するための方法のフローチャートである。

#### 【0168】

ある実施形態に従うと、ステップ2010において、方法は、1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することができ、第1のサブネットは、複数のスイッチと、複数のホストチャネルアダプタとを含み、ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、複数のホストチャネルアダプタは、複数のスイッチを介して相互接続される。

【0169】

ある実施形態に従うと、ステップ2020において、本方法は、複数の仮想マシンを含む複数のエンドノードを提供することができる。

【0170】

ある実施形態に従うと、ステップ2030において、本方法は、ホストチャネルアダプタを選択的RDMA制限に関連付けることができる。

【0171】

ある実施形態に従うと、ステップ2040において、本方法は、選択的RDMA制限を含むホストチャネルアダプタにおいて、複数の仮想マシンのうちのある仮想マシンをホストすることができる。

【0172】

複数の共有帯域幅セグメントを組み合わせる (ORA20Q246-US-NP-3)

ある実施形態に従うと、ネットワークインターフェイスのための従来の帯域幅／レート制限スキームは、典型的には、全体の合算された送信レートと、場合によっては個々の宛先に対する最大レートとの組み合わせに制限される。しかしながら、多くの場合、中間ネットワーク／ファブリックトポロジにおいて、共有ボトルネックがあり、ターゲットのセットに利用可能な総帯域幅がこの共有ボトルネックによって制限されることを意味する。したがって、どのようなレートで様々なデータフローを送信することができるかを決定するときにそのような共有ボトルネックが考慮されなければ、各ターゲットごとのレート制限が遵守されるにもかかわらず、共有ボトルネックは過負荷になる可能性が高い。

【0173】

ある実施形態に従うと、本明細書のシステムおよび方法は、複数の個々のフローに関連付けることができるオブジェクト「ターゲットグループ」を導入することができ、このターゲットグループは、フローが用いているネットワーク／ファブリック経路内の個々の（共有される潜在性のある）リンクまたは他のボトルネックのレート制限を表すことができる。さらに、本システムおよび方法は、各フローがそのようなターゲットグループの階層に関連することを可能にして、個々のフローについての送信元とターゲットとの間の経路におけるすべてのリンクセグメントおよび任意の他の（共有）ボトルネックを表すことができるようにすることができる。

【0174】

ある実施形態に従うと、エグレス帯域幅を制限するために、本システムおよび方法は、共有ISL（スイッチ間リンク）上の輻輳の可能性を低減するために、帯域幅割当を共有する宛先のグループを確立することができる。これは、どの宛先／経路が論理レベルでどのグループにマッピングするかに関して管理されることができる宛先／経路関連ルックアップ機構を必要とする。これが意味するところは、ハイパー特権通信インフラストラクチャが、ファブリックトポロジにおけるピアノードの実際の位置、ならびに関連付けられる帯域幅割当を有するローカルHCA内の「ターゲットグループ」（すなわち、HCAレベルのオブジェクトタイプ）にマッピングされ得る関連するルーティングおよび容量情報を認識しなければならないことである。しかしながら、関連するターゲットグループにマッピングするために、WQE（ワークキューエントリ）／パケットアドレス情報の直接ルックアップをHWに行わせることは実際的ではない。代わりに、HCA実現例は、発信トラフィックおよび関連するターゲットグループのための送信コンテキストを表す、RC（信頼できる接続された）QP（キューペア）とアドレスハンドルとの間の関連付けを提供することができる。このようにして、この関連付けは、バーブズ（verbs）レベルで透過的であってもよく、代わりに、ハイパー特権ソフトウェアレベルによってセットアップされ

10

20

30

40

50

、その後、HCA HW（およびファームウェア）レベルで実施されることができる。このスキームに関連付けられる重大な追加の複雑さは、関連するVMまたはvHCAポートアドレス情報がマイグレーションにわたって維持されるライブVMマイグレーションが、異なる通信ピアに対してターゲットグループの変更があることを依然として意味するかもしれないことである。しかしながら、本システムおよび方法が、関連する帯域幅割当が100%正確でないいくつかの過渡期間を許容する限り、ターゲットグループ関連付けは同期して更新される必要はない。したがって、論理的接続性および通信する能力はVMマイグレーションによって変化しない場合があるが、マイグレーションされるVMおよびその通信ピアVMの両方におけるRC接続およびアドレスハンドルに関連付けられるターゲットグループは、マイグレーション後に「完全に誤っている」場合がある。これは、（例えば、VMが遠隔位置からそのピアと同じ「リーフグループ」に移動されるときに）利用可能な帯域幅よりも少ない帯域幅が利用されることと、（例えば、VMが、そのピアと同じ「リーフグループ」から、制限された帯域幅を有する共有ISLを意味する遠隔位置に移動されるときに）過剰な帯域幅が生成されることとの両方を意味する場合がある。

【0175】

ある実施形態に従うと、ターゲットグループが表す、ファブリックにおける関連する経路内におけるさまざまな優先度についての期待される帯域幅使用を反映するために、ターゲットグループ固有の帯域幅割当は、原則として、特定の優先度（「QOSクラス」）についての割当にも分割され得る。

【0176】

ある実施形態に従うと、ターゲットグループは、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。

【0177】

ある実施形態に従うと、このシステムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループ（帯域幅割当）の階層を考慮することができる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ（最大レート）に関連付けられるべきであることを意味する。すなわち、例えば、ターゲット制限が30Gb/sであり、中間アップリンク制限が50Gb/sである場合、ターゲットに向かう最大レートは、30Gb/sを決して超えることはできない。他方、複数の30Gb/sターゲットが同じ50Gb/s中間制限を共有している場合、これらのターゲットに向かうフローに対する関連するターゲットレート制限の使用は、中間レート制限のオーバーランを意味し得る。したがって、関連する制限内で可能な限り最良の利用およびスループットを保証するために、関連する階層内のすべてのターゲットグループを関連する厳密な順序で考慮することができる。これは、階層内の各ターゲットグループが利用可能な帯域幅を表す場合にのみ、パケットを関連する宛先に向けて送信することができることを意味する。したがって、上記の例において単一のフローがターゲットのうちの1つに向かってアクティブである場合、このフローは30Gb/sで動作することが許可されることになる。しかしながら、別のフローが（共有中間ターゲットグループを介して）別のターゲットに向かってアクティブになると、各フローは25Gb/sに制限されることになる。次のラウンドにおいて、2つのターゲットのうちの1つに向かう追加のフローがアクティブになる場合、同じターゲットへの2つのフローは、各々（すなわち、平均して、およびそれらが何らかの追加の帯域幅割当／制限を有さない限り、）12.5Gb/sで動作しているであろう。

【0178】

ある実施形態に従うと、複数のテナントがサーバ／HCAを共有している場合、初期エグレス帯域幅および実際のターゲット帯域幅の両方を、任意の中間ISL帯域幅の共有に加えて共有してもよい。他方、テナントごとに専用サーバ／HCAを有するシナリオでは、中間のISL帯域幅は、唯一の可能な「テナント間」帯域幅共有を表す。

## 【0179】

ある実施形態に従うと、ターゲットグループは通常、HCAポートについてグローバルであるべきであり、HCAレベルでのVF／テナント割当は、グローバルに、または特定の優先度について、ターゲットの任意の組合せについてテナントが生成することができる最大のローカルトラフィックを表すであろう。さらに、同じ階層内の「グローバル」ターゲットグループとならんで、いくつかのテナントに特有のターゲットグループを用いることも可能であろう。

## 【0180】

ある実施形態に従うと、ターゲットグループを実現するとともに、特定のQPまたはアドレスハンドルについてのターゲットグループ関連付け（階層）を表すためのいくつかの考えられ得る方法がある。しかしながら、16ビットのターゲットグループID空間、ならびに各QPおよびアドレスハンドルに対する最大4つまたは8つのターゲットグループ関連付けに対するサポートを提供することができる。次いで、各ターゲットグループID値は、関連するレートに対する関連するIPD（パケット間遅延）値、ならびにこのターゲットグループに関連付けられる次のパケットがいつ送信され得るかを定義するタイマー情報を反映する、何らかのHW状態を表すであろう。

## 【0181】

ある実施形態に従うと、異なるフロー／経路は、同じ共有リンクセグメント上で異なる「QOS ID」（すなわちサービスレベル、優先度など）を用いることができるので、異なるターゲットグループがそのような異なるQOS IDの帯域幅割当を表すように、異なるターゲットグループを同じリンクセグメントに関連付けることも可能である。しかしながら、QOS ID固有のターゲットグループと、同じリンクセグメントの物理リンクを表す単一のターゲットグループとの両方を表すことも可能である。

## 【0182】

ある実施形態に従うと、同様に、本システムおよび方法は、さらに、明示的なフロータイプのパケットヘッダパラメータによって、および／または動作タイプ（たとえばRDMA読出／書込／送信）を考慮に入れることによって定義される異なるフロータイプを区別して、異なるそのようなフロータイプ間を調停するよう、異なる「サブ割当」を実現することができる。特に、これは、ローカルノード自体によって最初に開始された要求側モードトラフィックに対して応答側モード帯域幅を表すフロー（すなわち、典型的にはRDMA読出応答トラフィック）を区別するのに有用であり得る。

## 【0183】

ある実施形態に従うと、ターゲットグループの厳密な使用、および任意のターゲットまたは共有ISLセグメントの容量を超えない総最大レートまで加算されるすべての関連する送信側HCAに対するレート制限で、原則として、「任意の」輻輳を回避することが可能である。しかしながら、これは、異なるフローのための持続された帯域幅と、利用可能なリンク帯域幅の低い平均利用率との両方の厳しい制限を意味する場合がある。したがって、様々なレート制限は、異なるHCAがより楽観的な最大レートを用いることを可能にするように設定される場合がある。この場合、集計された合計は、持続可能な最大値よりも大きく、したがって、輻輳につながる場合がある。

## 【0184】

図21は、ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。

## 【0185】

より具体的には、ある実施形態に従うと、図21は、ハイパーバイザ2111を含むホストチャネルアダプタ2101を示す。ハイパーバイザは、VF2114～2116などのいくつかの仮想機能（VF）、および物理機能（PF）2113をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク2100などのネットワークに接続するために使用されるポート2102および2103などのいくつかのポートをさらにサポートするかまたは備えることが

10

20

30

40

50

できる。ネットワークは、たとえば、H C A 2 1 0 1を、スイッチ、追加の別個のH C Aなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはR o C Eネットワークなどのスイッチドネットワークを備えることができる。

【0186】

ある実施形態に従うと、上述したように、仮想機能の各々は、V M 1 2 1 5 0、V M 2 2 1 5 1およびV M 3 2 1 5 2のような仮想マシン（V M）をホストすることができる。

【0187】

ある実施形態に従うと、ホストチャネルアダプタ2 1 0 1は、さらに、ハイパーバイザを介して、仮想スイッチ2 1 1 2をサポートすることができる。これは、v S w i t c hアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（v P o r t）アーキテクチャをさらにサポートすることができる。

10

【0188】

ある実施形態に従うと、ネットワーク2 1 0 0は、図示のように、スイッチ2 1 4 0，2 1 4 1，2 1 4 2，および2 1 4 3など、相互接続され、たとえばリーフスイッチ2 1 4 0および2 1 4 1を介してホストチャネルアダプタ2 1 0 1に接続され得る、いくつかのスイッチを備え得る。

【0189】

ある実施形態に従うと、スイッチ2 1 4 0～2 1 4 3は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード（たとえば他のH C A）に接続されることができる。

20

【0190】

ある実施形態に従うと、ターゲットグループ2 1 7 0および2 1 7 1のようなターゲットグループは、リーフスイッチ2 1 4 0とスイッチ2 1 4 2との間およびリーフスイッチ2 1 4 1とスイッチ2 1 4 3との間のI S Lのようなスイッチ間リンク（I S L）に沿って定義することができる。これらのターゲットグループ2 1 7 0および2 1 7 1は、たとえば、レート制限コンポーネント2 1 6 0によってアクセス可能である、H C Aに関連付けられるターゲットグループリポジトリ2 1 6 1に記憶された、H C Aオブジェクトとしての帯域幅割当を表すことができる。

【0191】

ある実施形態に従うと、ターゲットグループ2 1 7 0および2 1 7 1は、特定の（および異なる）帯域幅割当を表すことができる。これらの帯域幅割当は、ターゲットグループが表す、ファブリックにおける関連する経路内のさまざまな優先度についての予想される帯域幅使用を反映するために、特定の優先度（「Q O Sクラス」）について割当に分割され得る。

30

【0192】

ある実施形態に従うと、ターゲットグループ2 1 7 0および2 1 7 1は、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。すなわち、例えば、V M 2 2 1 5 1に対するデフォルト／元のエグレス制限が1つの閾値に設定されるが、V M 2から送信されるパケットの宛先が、より低い帯域幅制限を設定するターゲットグループ2 1 7 0を通過するであろう場合、V M 2からのエグレス帯域幅は、V M 2に課されたデフォルト／元のエグレス制限よりも低いレベルに制限され得る。H C Aは、例えばV M 2からのパケットのルーティングに関与するターゲットグループに応じて、そのようなスロットリング／エグレス帯域幅制限調整を担うことができる。

40

【0193】

ある実施形態に従うと、ターゲットグループは、本質的に階層的であることもでき、それにより、本システムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループの階層（帯域幅割当）を考慮することがで

50

きる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ（最大レート）に関連付けられるべきであることを意味する。すなわち、例えば、ターゲットグループ 2170 がターゲットグループ 2171 よりも高い帯域幅制限を表し、パケットが 2 つのターゲットグループによって表される両方のスイッチ間リンクを介してアドレス指定される場合、ターゲットグループ 2171 の帯域幅制限は、制御的帯域幅制限ファクタである。

#### 【0194】

ある実施形態に従うと、ターゲットグループは、複数のフローによって共有することもできる。たとえば、各フローに関連付けられる QoS および SLA に応じて、ターゲットグループによって表される帯域幅割当を分割することができる。例として、VM1 および VM2 の両方が、例えば、10Gb/s の帯域幅割当を表すターゲットグループ 2170 を関与させるであろうフローを同時に送信し、それぞれのフローがそれに関連付けられる等しい QoS および SLA を有する場合、ターゲットグループ 2170 は、各フローについて 5Gb/s の制限を表すであろう。ターゲットグループ帯域幅割当のこの共有または分割は、それぞれのフローに関連付けられる QoS および SLA に基づいて変更され得る。

#### 【0195】

図 22 は、ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。

#### 【0196】

より具体的には、ある実施形態に従うと、図 22 は、ハイパーバイザ 2211 を含むホストチャネルアダプタ 2201 を示す。ハイパーバイザは、VF 2214～2216 などのいくつかの仮想機能（VF）、および物理機能（PF）2213 をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク 2200 などのネットワークに接続するために使用されるポート 2202 および 2203 などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA 2201 を、スイッチ、追加の別個の HCA などのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたは RoCE ネットワークなどのスイッチドネットワークを備えることができる。

#### 【0197】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 2250、VM2 2251 および VM3 2252 のような仮想マシン（VM）をホストすることができる。

#### 【0198】

ある実施形態に従うと、ホストチャネルアダプタ 2201 はさらに、ハイパーバイザを介して仮想スイッチ 2212 をサポートすることができる。これは、vSwitch アーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（vPort）アーキテクチャをさらにサポートすることができる。

#### 【0199】

ある実施形態に従うと、ネットワーク 2200 は、図示のように、スイッチ 2240、2241、2242、および 2243 など、相互接続され、たとえばリーフスイッチ 2240 および 2241 を介してホストチャネルアダプタ 2201 に接続され得る、いくつかのスイッチを備え得る。

#### 【0200】

ある実施形態に従うと、スイッチ 2240～2243 は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード（たとえば他の HCA）に接続されることができる。

#### 【0201】

ある実施形態に従うと、ターゲットグループ 2270 および 2271 のようなターゲットグループを、たとえばスイッチポートにおいて定義することができる。図に示すように、ターゲットグループ 2270 および 2271 は、それぞれスイッチ 2242 および 22

10

20

30

40

50

43のスイッチポートにおいて定義される。これらのターゲットグループ2270および2271は、たとえば、レート制限コンポーネント2260によってアクセス可能である、HCAに関連付けられるターゲットグループリポジトリ2261に記憶された、HCAオブジェクトとしての帯域幅割当を表すことができる。

#### 【0202】

ある実施形態に従うと、ターゲットグループ2270および2271は、特定の（および異なる）帯域幅割当を表すことができる。これらの帯域幅割当は、ターゲットグループが表す、ファブリックにおける関連する経路内のさまざまな優先度についての予想される帯域幅使用を反映するために、特定の優先度（「QoSクラス」）について割当に分割され得る。

#### 【0203】

ある実施形態に従うと、ターゲットグループ2270および2271は、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。すなわち、例えば、VM2 2251に対するデフォルト／元のエグレス制限が1つの閾値に設定されるが、VM2から送信されるパケットの宛先が、より低い帯域幅制限を設定するターゲットグループ2270を通過するであろう場合、VM2からのエグレス帯域幅は、VM2に課されたデフォルト／元のエグレス制限よりも低いレベルに制限され得る。HCAは、例えばVM2からのパケットのルーティングに関与するターゲットグループに応じて、そのようなスロットリング／エグレス帯域幅制限調整を担うことができる。

#### 【0204】

ある実施形態に従うと、ターゲットグループは、本質的に階層的であることもでき、それにより、本システムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループの階層（帯域幅割当）を考慮することができる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ（最大レート）に関連付けられるべきであることを意味する。すなわち、例えば、ターゲットグループ2270がターゲットグループ2271よりも高い帯域幅制限を表し、パケットが2つのターゲットグループによって表される両方のスイッチ間リンクを介してアドレス指定される場合、ターゲットグループ2271の帯域幅制限は、制御的帯域幅制限ファクタである。

#### 【0205】

ある実施形態に従うと、ターゲットグループは、複数のフローによって共有することもできる。たとえば、各フローに関連付けられるQoSおよびSLAに応じて、ターゲットグループによって表される帯域幅割当を分割することができる。例として、VM1およびVM2の両方が、例えば、10Gb/sの帯域幅割当を表すターゲットグループ2270を関与させるであろうフローを同時に送信し、それぞれのフローがそれに関連付けられる等しいQoSおよびSLAを有する場合、ターゲットグループ2270は、各フローについて5Gb/sの制限を表すであろう。ターゲットグループ帯域幅割当のこの共有または分割は、それぞれのフローに関連付けられるQoSおよびSLAに基づいて変更され得る。

#### 【0206】

ある実施形態に従うと、図21および図22は、それぞれスイッチ間リンクおよびスイッチポートにおいて定義されるターゲットグループを示す。当業者は、ターゲットグループがサブネット内の様々な場所で定義され得ること、およびいずれの所与のサブネットも、ターゲットグループがISLおよびスイッチポートにおいてのみ定義されるように限定されはしないが、概してそのようなターゲットグループは任意の所与のサブネット内のISLおよびスイッチポートの両方において定義され得ることを容易に理解するであろう。

#### 【0207】

図23は、ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。

## 【0208】

より具体的には、ある実施形態に従うと、図23は、ハイパーバイザ2311を含むホストチャンネルアダプタ2301を示す。ハイパーバイザは、VF2314～2316などのいくつかの仮想機能(VF)、および物理機能(PF)2313をホストするかまたはそれらに関連付けられることができる。ホストチャンネルアダプタは、ホストチャンネルアダプタをネットワーク2300などのネットワークに接続するために使用されるポート2302および2303などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA2301を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

10

## 【0209】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 2350、VM2 2351、およびVM3 2352のような仮想マシン(VM)をホストすることができる。

## 【0210】

ある実施形態に従うと、ホストチャンネルアダプタ2301は、さらに、ハイパーバイザを介して仮想スイッチ2312をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート(vPort)アーキテクチャをさらにサポートすることができる。

20

## 【0211】

ある実施形態に従うと、ネットワーク2300は、図示のように、スイッチ2340、2341、2342、および2343など、相互接続され、たとえばリーフスイッチ2340および2341を介してホストチャンネルアダプタ2301に接続され得る、いくつかのスイッチを備え得る。

## 【0212】

ある実施形態に従うと、スイッチ2340～2343は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード(たとえば他のHCA)に接続されることができる。

## 【0213】

ある実施形態に従うと、ターゲットグループ2370および2371のようなターゲットグループは、リーフスイッチ2340とスイッチ2342との間およびリーフスイッチ2341とスイッチ2343との間のISLのようなスイッチ間リンク(ISL)に沿って定義することができる。これらのターゲットグループ2370および2371は、たとえば、レート制限コンポーネント2360によってアクセス可能である、HCAに関連付けられるターゲットグループリポジトリ2361に記憶された、HCAオブジェクトとしての帯域幅割当を表すことができる。

30

## 【0214】

ある実施形態に従うと、ターゲットグループ2370および2371は、特定の(および異なる)帯域幅割当を表すことができる。これらの帯域幅割当は、ターゲットグループが表す、ファブリックにおける関連する経路内のさまざまな優先度についての予想される帯域幅使用を反映するために、特定の優先度(「QOSクラス」)について割当に分割され得る。

40

## 【0215】

ある実施形態に従うと、ターゲットグループ2370および2371は、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。すなわち、例えば、VM2 2351に対するデフォルト/元のエグレス制限が1つの閾値に設定されるが、VM2から送信されるパケットの宛先が、より低い帯域幅制限を設定するターゲットグループ2370を通過するである

50

う場合、VM 2からのエグレス帯域幅は、VM 2に課されたデフォルト／元のエグレス制限よりも低いレベルに制限され得る。HCAは、例えばVM 2からのパケットのルーティングに関与するターゲットグループに応じて、そのようなスロットリング／エグレス帯域幅制限調整を担うことができる。

#### 【0216】

ある実施形態に従うと、ターゲットグループは、本質的に階層的であることもでき、それにより、本システムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループの階層（帯域幅割当）を考慮することができる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ（最大レート）に関連付けられるべきであることを意味する。すなわち、例えば、ターゲットグループ2370がターゲットグループ2371よりも高い帯域幅制限を表し、パケットが2つのターゲットグループによって表される両方のスイッチ間リンクを介してアドレス指定される場合、ターゲットグループ2371の帯域幅制限は、制御的帯域幅制限ファクタである。

#### 【0217】

ある実施形態に従うと、ターゲットグループは、複数のフローによって共有することもできる。たとえば、各フローに関連付けられるQoSおよびSLAに応じて、ターゲットグループによって表される帯域幅割当を分割することができる。例として、VM 1およびVM 2の両方が、例えば、10Gb/sの帯域幅割当を表すターゲットグループ2370を関与させるであろうフローを同時に送信し、それぞれのフローがそれに関連付けられる等しいQoSおよびSLAを有する場合、ターゲットグループ2370は、各フローについて5Gb/sの制限を表すであろう。ターゲットグループ帯域幅割当のこの共有または分割は、それぞれのフローに関連付けられるQoSおよびSLAに基づいて変更され得る。

#### 【0218】

ある実施形態に従うと、ターゲットグループリポジトリは、ターゲットグループ2370にクエリ2375を行い、たとえばターゲットグループの帯域幅割当を判断することができる。ターゲットグループの帯域幅割当を判断すると、ターゲットグループリポジトリは、ターゲットグループに関連付けられる割当値を記憶し得る。この割当は、次いで、レート制限コンポーネントによって以下のように使用され得る：a) ターゲットグループの帯域幅割当がVMの帯域幅割当のそれよりも低いかどうかをQoSまたはSLAに基づいて判定し、およびb) そのような判定で、ターゲットグループ2370を横断する経路に基づいてVMの帯域幅割当を更新2376する。

#### 【0219】

図24は、ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。

#### 【0220】

より具体的には、ある実施形態に従うと、図24は、ハイパーバイザ2411を含むホストチャネルアダプタ2401を示す。ハイパーバイザは、VF2414～2416などのいくつかの仮想機能（VF）、および物理機能（PF）2413をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク2400などのネットワークに接続するために使用されるポート2402および2403などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA2401を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

#### 【0221】

ある実施形態に従うと、上記のように、仮想機能の各々は、VM1 2450、VM2 2451、VM3 2453などの仮想マシン（VM）をホストすることができる。

#### 【0222】

ある実施形態に従うと、ホストチャネルアダプタ2401は、さらに、ハイパーバイザ

10

20

30

40

50

を介して仮想スイッチ 2412 をサポートすることができる。これは、vSwitch アーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート (vPort) アーキテクチャをさらにサポートすることができる。

#### 【0223】

ある実施形態に従うと、ネットワーク 2400 は、図示のように、スイッチ 2440, 2441, 2442, および 2443 など、相互接続され、たとえばリーフスイッチ 2440 および 2441 を介してホストチャネルアダプタ 2401 に接続され得る、いくつかのスイッチを備え得る。

#### 【0224】

ある実施形態に従うと、スイッチ 2440 ~ 2443 は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード (たとえば他の HCA) に接続されることができる。

#### 【0225】

ある実施形態に従うと、ターゲットグループ 2470 および 2471 のようなターゲットグループを、たとえばスイッチポートにおいて定義することができる。図に示すように、ターゲットグループ 2470 および 2471 は、それぞれスイッチ 2442 および 2443 のスイッチポートにおいて定義される。これらのターゲットグループ 2470 および 2471 は、たとえば、レート制限コンポーネント 2460 によってアクセス可能である、HCA に関連付けられるターゲットグループリポジトリ 2461 に記憶された、HCA オブジェクトとしての帯域幅割当を表すことができる。

#### 【0226】

ある実施形態に従うと、ターゲットグループ 2470 および 2471 は、特定の (および異なる) 帯域幅割当を表すことができる。これらの帯域幅割当は、ターゲットグループが表す、ファブリックにおける関連する経路内のさまざまな優先度についての予想される帯域幅使用を反映するために、特定の優先度 (「QoS クラス」) について割当に分割され得る。

#### 【0227】

ある実施形態に従うと、ターゲットグループ 2470 および 2471 は、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。すなわち、例えば、VM 2 2451 に対するデフォルト／元のエグレス制限が 1 つの閾値に設定されるが、VM 2 から送信されるパケットの宛先が、より低い帯域幅制限を設定するターゲットグループ 2470 を通過するであろう場合、VM 2 からのエグレス帯域幅は、VM 2 に課されたデフォルト／元のエグレス制限よりも低いレベルに制限され得る。HCA は、例えば VM 2 からのパケットのルーティングに関与するターゲットグループに応じて、そのようなスロットリング／エグレス帯域幅制限調整を担うことができる。

#### 【0228】

ある実施形態に従うと、ターゲットグループは、本質的に階層的であることもでき、それにより、本システムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループの階層 (帯域幅割当) を考慮することができる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ (最大レート) に関連付けられるべきであることを意味する。すなわち、例えば、ターゲットグループ 2470 がターゲットグループ 2471 よりも高い帯域幅制限を表し、パケットが 2 つのターゲットグループによって表される両方のスイッチ間リンクを介してアドレス指定される場合、ターゲットグループ 2471 の帯域幅制限は、制御的帯域幅制限ファクタである。

#### 【0229】

ある実施形態に従うと、ターゲットグループは、複数のフローによって共有することも

10

20

30

40

50

できる。たとえば、各フローに関連付けられるQoSおよびSLAに応じて、ターゲットグループによって表される帯域幅割当を分割することができる。例として、VM1およびVM2の両方が、例えば、10Gb/sの帯域幅割当を表すターゲットグループ2470を関与させるであろうフローを同時に送信し、それぞれのフローがそれに関連付けられる等しいQoSおよびSLAを有する場合、ターゲットグループ2470は、各フローについて5Gb/sの制限を表すであろう。ターゲットグループ帯域幅割当のこの共有または分割は、それぞれのフローに関連付けられるQoSおよびSLAに基づいて変更され得る。

#### 【0230】

ある実施形態に従うと、ターゲットグループリポジトリは、ターゲットグループ2470にクエリ2475を行い、たとえばターゲットグループの帯域幅割当を判断することができる。ターゲットグループの帯域幅割当を判断すると、ターゲットグループリポジトリは、ターゲットグループに関連付けられる割当値を記憶し得る。この割当は、次いで、レート制限コンポーネントによって以下のように使用され得る：a) ターゲットグループの帯域幅割当がVMの帯域幅割当のそれよりも低いかどうかをQoSまたはSLAに基づいて判定し、およびb) そのような判定で、ターゲットグループ2470を横断する経路に基づいてVMの帯域幅割当を更新2476する。

#### 【0231】

図25は、ある実施形態に係る、高性能コンピューティング環境において複数の共有帯域幅セグメントを組み合わせるためのシステムを示す。

#### 【0232】

ある実施形態に従うと、スイッチドネットワークまたはサブネット2500のような高性能コンピューティング環境内では、いくつかのエンドノード2501および2502は、リーフスイッチ2511および2512、スイッチ2521および2522、ならびにルートスイッチ2531および2532のようないくつかのスイッチを介して相互接続されるいくつかの仮想マシンVM1~VM4 2550~2553をサポートすることができる。

#### 【0233】

ある実施形態に従うと、ノード2501および2502の接続のための機能を提供するさまざまなホストチャネルアダプタ、ならびにサブネットに接続されるべき仮想マシンは図示されていない。そのような実施形態の議論はSR-IOVに関して上記されており、各仮想マシンはホストチャネルアダプタ上のハイパーバイザの仮想機能に関連付けられ得る。

#### 【0234】

ある実施形態に従うと、上記で論じたように、そのようなスイッチドファブリックに固有なのは、各エンドノードまたはVMが、そこに流入するトラフィックおよびそこから流出するトラフィックが遵守しなければならないそれ自体のエグレス／イングレス帯域幅制限を有してもよい一方で、サブネット内には、そこに流入するトラフィックについてのボトルネックを表すリンクまたはポートも存在し得るという概念である。したがって、VM1、VM2、VM3、またはVM4などのそのようなエンドノードにどのようなレートでトラフィックが流出入すべきかを判断するとき、レート制限コンポーネント2560および2561は、2550および2551などの様々なターゲットグループに問い合わせ、そのようなターゲットグループがトラフィックフローについてボトルネックを表すかどうかを判断することができる。そのような判断に応じて、レート制限コンポーネント2560および2561は、次いで、レート制限コンポーネントが制御するエンドポイントに、異なるまたは新たな帯域幅制限を設定することができる。

#### 【0235】

ある実施形態に従うと、さらに、ターゲットグループは、VM1からVM3へのトラフィックがターゲットグループ2550および2551の両方を利用する場合、レート制限2560は、VM1からVM3への帯域幅制限を決定する際に、そのようなターゲットグループの両方からの制限を考慮に入れることができるように、ネスト化された／階層的な

10

20

30

40

50

態様で問い合わせされることができる。

【0236】

図26は、ある実施形態に係る、高性能コンピューティング環境においてプライベートファブリックにおける輻輳制御のためにターゲットグループをサポートする方法のフローチャートである。

【0237】

ある実施形態に従うと、ステップ2610において、本方法は、1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することができ、第1のサブネットは複数のスイッチを含み、複数のスイッチは少なくともリーフスイッチを含み、複数のスイッチの各々は複数のスイッチポートを含み、第1のサブネットはさらに、複数のホストチャンネルアダプタを含み、ホストチャンネルアダプタの各々は、少なくとも1つのホストチャンネルアダプタポートを含み、複数のホストチャンネルアダプタは、複数のスイッチを介して相互接続され、第1のサブネットはさらに、複数の仮想マシンを含む複数のエンドノードを含む。

10

【0238】

ある実施形態に従うと、ステップ2620において、本方法は、複数のスイッチのうちの2つのスイッチ間のスイッチ間リンクまたは複数のスイッチのうちのいずれかのスイッチのポートのうちの少なくとも1つにおいて、ターゲットグループを定義することができ、ターゲットグループは、複数のスイッチのうちの2つのスイッチ間のスイッチ間リンクまたは複数のスイッチのうちのいずれかのスイッチのポートのうちの少なくとも1つにおける帯域幅制限を定義する。

20

【0239】

ある実施形態に従うと、ステップ2630において、本方法は、ホストチャンネルアダプタにおいて、ホストチャンネルアダプタのメモリに記憶されるターゲットグループリポジトリを提供することができる。

【0240】

ある実施形態に従うと、ステップ2640において、本方法は、定義されたターゲットグループをターゲットグループリポジトリに記録することができる。

【0241】

ターゲット固有の送信／RDMA書込およびRDMA読出帯域幅制限の組合せ（ORA200246-US-NP-2）

30

ある実施形態に従うと、ノード／VMは、ピアノード／VMによって開始される送信およびRDMA書込動作の結果とローカルノード／VM自体によって開始されるRDMA読出動作の結果との両方である着信データトラフィックについてのターゲットとすることができる。そのような状況では、これらのフローのすべてがレート制限に関して調整されない限り、ローカルノード／VMの最大または平均インGRESS帯域幅が必要な境界内にあることを保証することが問題になる。

【0242】

ある実施形態に従うと、本明細書に記載のシステムおよび方法は、ローカルメモリからのデータのフェッチおよび関連する遠隔ターゲットへのデータの送信を表すすべてのフローがすべて、同じ共有レート制限ならびに関連付けられるフロースケジューリングおよび調停に従うことを可能にする方法で、ターゲット固有のエGRESSレート制御を実現することができる。また、異なるフロータイプには、利用可能な帯域幅の、異なる優先度および／または異なるシェアが与えられてもよい。

40

【0243】

ある実施形態に従うと、「プロデューサ／送信元」ノードからのフローに対するターゲットグループ関連付けが、UD（信頼できないデータグラム）送信、RDMA書込、RDMA送信およびRDMA読出（すなわち、データを有するRDMA読出応答）を含むすべての発信データパケットの帯域幅規制を意味する限り、vHCAポートに対するすべてのインGRESS帯域幅の完全な制御がある。これは、ターゲットvHCAポートを所有するVMが複数のピアノードに対して「過剰な」量のRDMA読出要求を生成しているかどうか

50

とは無関係である。

【0244】

ある実施形態に従うと、ターゲットグループをフロー固有および「非請求」の両方のBECNシグナリングに結合することは、VHCAポートごとのイングレス帯域幅を、任意の数の遠隔ピアに対して動的にスロットリングすることができることを意味する。

【0245】

ある実施形態に従うと、「非請求BECN」メッセージを用いて、異なるステージ番号に対する純粋なCEフラグ立て／フラグ解除に加えて、特定のレート値を通信することもできる。このようにして、新たなピアからの初期着信パケット（例えば、通信管理（CM）パケット）が、着信パケットが来たHCA（すなわち、関連するファームウェア／ハイパー特権ソフトウェア）および現在の通信ピアの両方への1つ以上の「非請求BECN」メッセージの生成をトリガすることができるスキームを有することが可能である。

【0246】

ある実施形態に従うと、HCA上の両方のポートが同時に使用されるケース（すなわちアクティブ-アクティブスキーム）では、同時フローがいくつかのISLを共有しているか、またはさらには同じ宛先ポートをターゲットにしていることができる可能性がある場合、ローカルHCAポート間でターゲットグループを共有することは意味をなし得る。

【0247】

ある実施形態に従うと、HCAポート間でターゲットグループを共有する別の理由は、HCAローカルメモリ帯域幅が両方の（すべての）HCAポートについて全速リンク速度を維持できないかどうかである。この場合、ターゲットグループは、どのポートがソースHCAまたは宛先HCAのいずれに関与するかに関係なく、集約された総リンク帯域幅がローカルメモリ帯域幅を決して超えないように、設定することができる。

【0248】

ある実施形態に従うと、特定の宛先に向かう固定ルートの場合、任意の中間ターゲットグループは、典型的には、経路における特定のステージにおいて単一のISLのみを表すことになる。しかしながら、動的転送がアクティブである場合、ターゲットグループおよびECN処理の両方がこれを考慮しなければならない。スイッチのペア間の並列ISL間のトラフィック（例えば、単一のリーフスイッチから単一のスパインスイッチへのアップリンク）を均衡させるためにのみ動的転送判断が生じる場合、すべての処理は、原則的には、単一のISLのみが使用されている場合と非常に類似している。FECN通知は、関連するグループ内のすべてのポートの状態に基づいて行われると思われ、シグナリングは、ポートのいずれかからの輻輳指示に基づいてシグナリングされるという意味で「積極的」であり得るか、またはより保守的であり得、グループ内のすべてのポートについての共有出力キューのサイズに基づき得る。ターゲットグループ構成は、任意のパケットがその時点で最良の出力ポートを選択することを転送が可能にする限り、通常は、グループ内のすべてのリンクについての集約された帯域幅を表すであろう。しかしながら、フローごとに厳密なパケット順序維持の概念がある場合、いくつかのフローはある時点で同じISLを使用「しなければならない」場合があるため、帯域幅割当の評価はより複雑である。そのようなフロー順序スキームが、十分に定義されたヘッダフィールドに基づく場合、グループ内の各ポートを独立したターゲットグループとして表すことが最良である場合がある。この場合、送信元側HCAにおけるターゲットグループの選択は、RCQP接続またはアドレスハンドルに関連付けられることになるヘッダフィールドの評価を、スイッチがすべてのパケットについて実行時に実行するものと同じにすることができない。ならない。

【0249】

ある実施形態に従うと、デフォルトでは、新たな遠隔ターゲットについての初期ターゲットグループレートは、保守的に低く設定され得る。このようにして、ターゲットが関連レートを更新する機会を有するまで固有のスロットリングが存在する。したがって、すべてのそのようなレート制御は、関与するVM自体とは無関係であるが、VMは、イングレ

10

20

30

40

50

ストラフィックとエグレストラフィックとの両方について異なる遠隔ピアの割当を更新するようにハイパーバイザに要求することができるであろうが、これは、ローカル v H C A ポートと遠隔 v H C A ポートとの両方について定義された総制約内でのみ許可されるであろう。

#### 【0250】

図27は、一実施形態による、高性能コンピューティング環境においてターゲット固有の R D M A 書込帯域幅制限と R D M A 読出帯域幅制限とを組み合わせるためのシステムを示す。

#### 【0251】

より具体的には、ある実施形態に従うと、図27は、ハイパーバイザ2711を含むホストチャンネルアダプタ2701を示す。ハイパーバイザは、V F 2714~2716などのいくつかの仮想機能(V F)、および物理機能(P F)2713をホストするかまたはそれらに関連付けられることができる。ホストチャンネルアダプタは、ホストチャンネルアダプタをネットワーク2700などのネットワークに接続するために使用されるポート2702および2703などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、H C A 2701を、スイッチ、追加の別個の H C A などのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたは R o C E ネットワークなどのスイッチドネットワークを備えることができる。

#### 【0252】

ある実施形態に従うと、上述したように、仮想機能の各々は、V M 1 2750、V M 2 2751、および V M 3 2752 のような仮想マシン(V M)をホストすることができる。

#### 【0253】

ある実施形態に従うと、ホストチャンネルアダプタ2701は、さらに、ハイパーバイザを介して仮想スイッチ2712をサポートすることができる。これは、v S w i t c h アーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート(v P o r t)アーキテクチャをさらにサポートすることができる。

#### 【0254】

ある実施形態に従うと、ネットワーク2700は、図示のように、スイッチ2740, 2741, 2742, および2743など、相互接続され、たとえばリーフスイッチ2740および2741を介してホストチャンネルアダプタ2701に接続され得る、いくつかのスイッチを備え得る。

#### 【0255】

ある実施形態に従うと、スイッチ2740~2743は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード(たとえば他の H C A)に接続されることができる。

#### 【0256】

ある実施形態に従うと、ターゲットグループ2770および2771のようなターゲットグループは、リーフスイッチ2740とスイッチ2742との間およびリーフスイッチ2741とスイッチ2743との間の I S L のようなスイッチ間リンク(I S L)において定義することができる。これらのターゲットグループ2770および2771は、たとえば、レート制限コンポーネント2760によってアクセス可能である、H C A に関連付けられるターゲットグループリポジトリ2761に記憶された、H C A オブジェクトとしての帯域幅割当を表すことができる。

#### 【0257】

ある実施形態に従うと、ターゲットグループ2770および2771は、特定の(および異なる)帯域幅割当を表すことができる。これらの帯域幅割当は、ターゲットグループが表す、ファブリックにおける関連する経路内のさまざまな優先度についての予想される帯域幅使用を反映するために、特定の優先度(「Q O S クラス」)について割当に分割さ

10

20

30

40

50

れ得る。

#### 【0258】

ある実施形態に従うと、ターゲットグループ2770および2771は、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。すなわち、例えば、VM2 2751に対するデフォルト／元のエグレス制限が1つの閾値に設定されるが、VM2から送信されるパケットの宛先が、より低い帯域幅制限を設定するターゲットグループ2770を通過するであろう場合、VM2からのエグレス帯域幅は、VM2に課されたデフォルト／元のエグレス制限よりも低いレベルに制限され得る。HCAは、例えばVM2からのパケットのルーティン

10

#### 【0259】

ある実施形態に従うと、ターゲットグループは、本質的に階層的であることもでき、それにより、本システムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループの階層（帯域幅割当）を考慮することができる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ（最大レート）に関連付けられるべきであることを意味する。すなわち、例えば、ターゲットグループ2770がターゲットグループ2771よりも高い帯域幅制限を表し、パケットが2つのターゲットグループによって表される両方のスイッチ間リンクを介してアドレス指定される場合、ターゲットグループ2771の帯域幅制限は、制御的帯域幅制限ファクタである。

20

#### 【0260】

ある実施形態に従うと、ターゲットグループは、複数のフローによって共有することもできる。たとえば、各フローに関連付けられるQoSおよびSLAに応じて、ターゲットグループによって表される帯域幅割当を分割することができる。例として、VM1およびVM2の両方が、例えば、10Gb/sの帯域幅割当を表すターゲットグループ2770を関与させるであろうフローを同時に送信し、それぞれのフローがそれに関連付けられる等しいQoSおよびSLAを有する場合、ターゲットグループ2770は、各フローについて5Gb/sの制限を表すであろう。ターゲットグループ帯域幅割当のこの共有または分割は、それぞれのフローに関連付けられるQoSおよびSLAに基づいて変更され得る。

30

#### 【0261】

ある実施形態に従うと、帯域幅割当およびパフォーマンスの問題は、VM、たとえばVM1 2750が複数のソースからの過剰なイングレス帯域幅2790を受ける場合に生じ得る。これは、たとえば、VM1が1つ以上のRDMA書込動作と同時に1つ以上のRDMA読出応答を受け、VM1上のイングレス帯域幅が2つ以上のソース（例えば、接続されたVMからの1つのRDMA読出応答、および別の接続されたVMからの1つのRDMA書込要求）から来る状況において生じ得る。そのような状況では、たとえば、スイッチ間リンク上のターゲットグループ2770などのターゲットグループは、たとえば、クエリ2775を介して、典型的に許可されるであろうよりも低い帯域幅割当を反映するよう

40

#### 【0262】

ある実施形態に従うと、さらに、HCAのレート制限コンポーネント2760は、たとえばVM1についてのイングレス帯域幅制限を、VM1上でイングレス帯域幅を生成する役割を担うノードについてのエグレス帯域幅制限と協調させるために、他のピアHCAと交渉され得るVM固有のレート制限2762をさらに含み得る。これら他のHCA／ノードは、図には示されていない。

#### 【0263】

図28は、一実施形態による、高性能コンピューティング環境においてターゲット固有のRDMA書込帯域幅制限とRDMA読出帯域幅制限とを組み合わせるためのシステムを

50

示す。

【0264】

より具体的には、ある実施形態に従うと、図28は、ハイパーバイザ2811を含むホストチャンネルアダプタ2801を示す。ハイパーバイザは、VF2814~2816などのいくつかの仮想機能(VF)、および物理機能(PF)2813をホストするかまたはそれらに関連付けられることができる。ホストチャンネルアダプタは、ホストチャンネルアダプタをネットワーク2800などのネットワークに接続するために使用されるポート2802および2803などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA2801を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

10

【0265】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 2850、VM2 2851、およびVM3 2852のような仮想マシン(VM)をホストすることができる。

【0266】

ある実施形態に従うと、ホストチャンネルアダプタ2801は、さらに、ハイパーバイザを介して仮想スイッチ2812をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート(vPort)アーキテクチャをさらにサポートすることができる。

20

【0267】

ある実施形態に従うと、ネットワーク2800は、図示のように、スイッチ2840、2841、2842、および2843など、相互接続され、たとえばリーフスイッチ2840および2841を介してホストチャンネルアダプタ2801に接続され得る、いくつかのスイッチを備え得る。

【0268】

ある実施形態に従うと、スイッチ2840~2843は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード(たとえば他のHCA)に接続されることができる。

30

【0269】

ある実施形態に従うと、ターゲットグループ2870および2871のようなターゲットグループを、たとえばスイッチポートにおいて定義することができる。図に示すように、ターゲットグループ2870および2871は、それぞれスイッチ2842および2843のスイッチポートにおいて定義される。これらのターゲットグループ2870および2871は、たとえば、レート制限コンポーネント2860によってアクセス可能である、HCAに関連付けられるターゲットグループリポジトリ2861に記憶された、HCAオブジェクトとしての帯域幅割当を表すことができる。

【0270】

ある実施形態に従うと、ターゲットグループ2870および2871は、特定の(および異なる)帯域幅割当を表すことができる。これらの帯域幅割当は、ターゲットグループが表す、ファブリックにおける関連する経路内のさまざまな優先度についての予想される帯域幅使用を反映するために、特定の優先度(「QOSクラス」)について割当に分割され得る。

40

【0271】

ある実施形態に従うと、ターゲットグループ2870および2871は、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。すなわち、例えば、VM2 2851に対するデフォルト/元のエグレス制限が1つの閾値に設定されるが、VM2から送信されるパケット

50

の宛先が、より低い帯域幅制限を設定するターゲットグループ 2870 を通過するであろう場合、VM2 からのエグレス帯域幅は、VM2 に課されたデフォルト／元のエグレス制限よりも低いレベルに制限され得る。HCA は、例えば VM2 からのパケットのルーティングに参与するターゲットグループに応じて、そのようなスロットリング／エグレス帯域幅制限調整を担うことができる。

#### 【0272】

ある実施形態に従うと、ターゲットグループは、本質的に階層的であることもでき、それにより、本システムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループの階層（帯域幅割当）を考慮することができる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ（最大レート）に関連付けられるべきであることを意味する。すなわち、例えば、ターゲットグループ 2870 がターゲットグループ 2871 よりも高い帯域幅制限を表し、パケットが 2 つのターゲットグループによって表される両方のスイッチ間リンクを介してアドレス指定される場合、ターゲットグループ 2871 の帯域幅制限は、制御的帯域幅制限ファクタである。

#### 【0273】

ある実施形態に従うと、ターゲットグループは、複数のフローによって共有することもできる。たとえば、各フローに関連付けられる QoS および SLA に応じて、ターゲットグループによって表される帯域幅割当を分割することができる。例として、VM1 および VM2 の両方が、例えば、10Gb/s の帯域幅割当を表すターゲットグループ 2870 を関与させるであろうフローを同時に送信し、それぞれのフローがそれに関連付けられる等しい QoS および SLA を有する場合、ターゲットグループ 2870 は、各フローについて 5Gb/s の制限を表すであろう。ターゲットグループ帯域幅割当のこの共有または分割は、それぞれのフローに関連付けられる QoS および SLA に基づいて変更され得る。

#### 【0274】

ある実施形態に従うと、帯域幅割当およびパフォーマンスの問題は、VM、たとえば VM1 2850 が複数のソースからの過剰なイングレス帯域幅 2890 を受ける場合に生じ得る。これは、たとえば、VM1 が 1 つ以上の RDMA 書込動作と同時に 1 つ以上の RDMA 読出応答を受け、VM1 上のイングレス帯域幅が 2 つ以上のソース（例えば、接続された VM からの 1 つの RDMA 読出応答、および別の接続された VM からの 1 つの RDMA 書込要求）から来る状況において生じ得る。そのような状況では、たとえば、スイッチ間リンク上のターゲットグループ 2870 などのターゲットグループは、たとえば、クエリ 2875 を介して、典型的に許可されるであろうよりも低い帯域幅割当を反映するように更新され得る。

#### 【0275】

ある実施形態に従うと、さらに、HCA のレート制限コンポーネント 2860 は、たとえば VM1 についてのイングレス帯域幅制限を、VM1 上でイングレス帯域幅を生成する役割を担うノードについてのエグレス帯域幅制限と協調させるために、他のピア HCA と交渉され得る VM 固有のレート制限 2862 をさらに含み得る。これら他の HCA／ノードは、図には示されていない。

#### 【0276】

図 29 は、一実施形態による、高性能コンピューティング環境においてターゲット固有の RDMA 書込帯域幅制限と RDMA 読出帯域幅制限とを組み合わせるためのシステムを示す。

#### 【0277】

より具体的には、ある実施形態に従うと、図 29 は、ハイパーバイザ 2911 を含むホストチャネルアダプタ 2901 を示す。ハイパーバイザは、VF 2914～2916 などのいくつかの仮想機能（VF）、および物理機能（PF）2913 をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク 2900 などのネットワークに接続するために使用されるポート 29

10

20

30

40

50

02および2903などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA2901を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

#### 【0278】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 2950、VM2 2951、およびVM3 2952のような仮想マシン（VM）をホストすることができる。

#### 【0279】

ある実施形態に従うと、ホストチャネルアダプタ2901は、さらに、ハイパーバイザを介して仮想スイッチ2912をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（vPort）アーキテクチャをさらにサポートすることができる。

#### 【0280】

ある実施形態に従うと、ネットワーク2900は、図示のように、スイッチ2940、2941、2942、および2943など、相互接続され、たとえばリーフスイッチ2940および2941を介してホストチャネルアダプタ2901に接続され得る、いくつかのスイッチを備え得る。

#### 【0281】

ある実施形態に従うと、スイッチ2940～2943は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード（たとえば他のHCA）に接続されることができる。

#### 【0282】

ある実施形態に従うと、ターゲットグループ2971のようなターゲットグループは、リーフスイッチ2941とスイッチ2943との間のISLのようなスイッチ間リンク（ISL）に沿って定義することができる。他のターゲットグループは、例えば、スイッチポートにおいて定義することができる。図に示すように、ターゲットグループ2970は、スイッチ2952のスイッチポートにおいて定義される。これらのターゲットグループ2970および2971は、たとえば、レート制限コンポーネント2960によってアクセス可能である、HCAに関連付けられるターゲットグループリポジトリ2961に記憶された、HCAオブジェクトとしての帯域幅割当を表すことができる。

#### 【0283】

ある実施形態に従うと、ターゲットグループ2970および2971は、特定の（および異なる）帯域幅割当を表すことができる。これらの帯域幅割当は、ターゲットグループが表す、ファブリックにおける関連する経路内のさまざまな優先度についての予想される帯域幅使用を反映するために、特定の優先度（「QoSクラス」）について割当に分割され得る。

#### 【0284】

ある実施形態に従うと、ターゲットグループ2970および2971は、オブジェクトを特定の宛先アドレスから切り離し、本システムおよび方法は、ターゲットに加えて、ターゲット制限よりも制限的であり得る帯域幅制限を表してもよい中間の共有リンクまたはリンクのグループを表す能力を得る。すなわち、例えば、VM2 2951に対するデフォルト／元のエグレス制限が1つの閾値に設定されるが、VM2から送信されるパケットの宛先が、より低い帯域幅制限を設定するターゲットグループ2970を通過するであろう場合、VM2からのエグレス帯域幅は、VM2に課されたデフォルト／元のエグレス制限よりも低いレベルに制限され得る。HCAは、例えばVM2からのパケットのルーティングに関与するターゲットグループに応じて、そのようなスロットリング／エグレス帯域幅制限調整を担うことができる。

#### 【0285】

ある実施形態に従うと、ターゲットグループは、本質的に階層的であることもでき、それにより、本システムおよび方法は、異なるターゲットに向かう異なるステージでの帯域幅／リンク共有を反映するターゲットグループの階層（帯域幅割当）を考慮することができる。原則として、これは、特定のフローが、階層において最も制限されたレートを表すターゲットグループ（最大レート）に関連付けられるべきであることを意味する。すなわち、例えば、ターゲットグループ 2970 がターゲットグループ 2971 よりも高い帯域幅制限を表し、パケットが 2 つのターゲットグループによって表される両方のスイッチ間リンクを介してアドレス指定される場合、ターゲットグループ 2971 の帯域幅制限は、制御的帯域幅制限ファクタである。

#### 【0286】

ある実施形態に従うと、ターゲットグループは、複数のフローによって共有することもできる。たとえば、各フローに関連付けられる QoS および SLA に応じて、ターゲットグループによって表される帯域幅割当を分割することができる。例として、VM1 および VM2 の両方が、例えば、10Gb/s の帯域幅割当を表すターゲットグループ 2970 を関与させるであろうフローを同時に送信し、それぞれのフローがそれに関連付けられる等しい QoS および SLA を有する場合、ターゲットグループ 2970 は、各フローについて 5Gb/s の制限を表すであろう。ターゲットグループ帯域幅割当のこの共有または分割は、それぞれのフローに関連付けられる QoS および SLA に基づいて変更され得る。

#### 【0287】

ある実施形態に従うと、帯域幅割当およびパフォーマンスの問題は、VM、たとえば VM1 2950 が複数のソースからの過剰なイングレス帯域幅 2990 を受ける場合に生じ得る。これは、たとえば、VM1 が 1 つ以上の RDMA 書込動作と同時に 1 つ以上の RDMA 読出応答を受け、VM1 上のイングレス帯域幅が 2 つ以上のソース（例えば、接続された VM からの 1 つの RDMA 読出応答、および別の接続された VM からの 1 つの RDMA 書込要求）から来る状況において生じ得る。そのような状況では、たとえば、スイッチ間リンク上のターゲットグループ 2970 などのターゲットグループは、たとえば、クエリ 2975 を介して、典型的に許可されるであろうよりも低い帯域幅割当を反映するように更新され得る。

#### 【0288】

ある実施形態に従うと、さらに、HCA のレート制限コンポーネント 2960 は、たとえば VM1 についてのイングレス帯域幅制限を、VM1 上でイングレス帯域幅を生成する役割を担うノードについてのエグレス帯域幅制限と協調させるために、他のピア HCA と交渉され得る VM 固有のレート制限 2962 をさらに含み得る。これら他の HCA／ノードは、図には示されていない。

#### 【0289】

図 30 は、一実施形態による、高性能コンピューティング環境においてターゲット固有の RDMA 書込帯域幅制限と RDMA 読出帯域幅制限とを組み合わせるためのシステムを示す。

#### 【0290】

ある実施形態に従うと、スイッチドネットワークまたはサブネット 3000 のような高性能コンピューティング環境内では、いくつかのエンドノード 3001 および 3002 は、リーフスイッチ 3011 および 3012、スイッチ 3021 および 3022、ならびにルートスイッチ 3031 および 3032 のようないくつかのスイッチを介して相互接続されるいくつかの仮想マシン VM1～VM4 3050～3053 をサポートすることができる。

#### 【0291】

ある実施形態に従うと、ノード 3001 および 3002 の接続のための機能を提供するさまざまなホストチャネルアダプタ、ならびにサブネットに接続されるべき仮想マシンは図示されていない。そのような実施形態の議論は SR-IOV に関して上記されており、各仮想マシンはホストチャネルアダプタ上のハイパーバイザの仮想機能に関連付けられ得

10

20

30

40

50

る。

【0292】

ある実施形態に従うと、VM3 3052などのノードは、RDMA読出応答3050およびRDMA書込要求3051（入来時の帯域幅）を同時に処理しているときに、（たとえばレート制限3061から）帯域幅制限に入ることができる。

【0293】

ある実施形態に従うと、レート制限3060および3061は、たとえば、RDMA要求（すなわち、VM3によってVM4に送信され、RDMA読出を要求し、RDMA読出応答3050をもたらすメッセージ）およびRDMA書込動作（例えば、VM2からVM3へのRDMA書込）を調整することによって、インGRES帯域幅割当に違反しないことを保証するように構成されることができる。

10

【0294】

各個々のノードについて、システムおよび方法は、そのようなターゲットグループのチェーンを有して、フローが、ターゲットグループにおいて表されるファブリックの異なる部分におけるリンク帯域幅を共有するすべての他のフローと常に調整されるようにすることができる。

【0295】

図31は、一実施形態による、高性能コンピューティング環境においてターゲット固有のRDMA書込帯域幅制限とRDMA読出帯域幅制限とを組み合わせるための方法のフローチャートである。

20

【0296】

ある実施形態に従うと、ステップ3110において、本方法は、1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することができ、第1のサブネットは複数のスイッチを含み、複数のスイッチは少なくともリーフスイッチを含み、複数のスイッチの各々は複数のスイッチポートを含み、第1のサブネットはさらに、複数のホストチャンネルアダプタを含み、ホストチャンネルアダプタの各々は、少なくとも1つのホストチャンネルアダプタポートを含み、複数のホストチャンネルアダプタは、複数のスイッチを介して相互接続され、第1のサブネットはさらに、複数の仮想マシンを含む複数のエンドノードを含む。

【0297】

ある実施形態に従うと、ステップ3120において、本方法は、複数のスイッチのうちの2つのスイッチ間のスイッチ間リンクまたは複数のスイッチのうちのいずれかのスイッチのポートのうちの少なくとも1つにおいて、ターゲットグループを定義することができ、ターゲットグループは、複数のスイッチのうちの2つのスイッチ間のスイッチ間リンクまたは複数のスイッチのうちのいずれかのスイッチのポートのうちの少なくとも1つにおける帯域幅制限を定義する。

30

【0298】

ある実施形態に従うと、ステップ3130において、本方法は、ホストチャンネルアダプタにおいて、ホストチャンネルアダプタのメモリに記憶されるターゲットグループリポジトリを提供することができる。

【0299】

ある実施形態に従うと、ステップ3140において、本方法は、定義されたターゲットグループをターゲットグループリポジトリに記録することができる。

40

【0300】

ある実施形態によると、ステップ3150において、本方法は、ホストチャンネルアダプタのエンドノードにおいて、少なくとも2つの遠隔エンドノードからインGRES帯域幅を受信することができ、インGRES帯域幅は、エンドノードのインGRES帯域幅制限を超える。

【0301】

ある実施形態に従うと、3160において、少なくとも2つのソースからのインGRES帯域幅の受信に応答して、本方法は、ターゲットグループの帯域幅割当を更新することが

50

できる。

【0302】

イングレス帯域幅調停と輻輳フィードバックとを組み合わせる (ORA200246-US-NP-2)

ある実施形態に従うと、複数の送信側ノード／VMの各々および／またはすべてが単一の受信側ノード／VMに送信している場合、送信側間の公平性の最適なバランスを達成して輻輳を回避し、同時に、受信側ノード／VMが消費するイングレス帯域幅の使用を、関連するネットワークインターフェイスがイングレストラフィックに提供できる最大物理リンク帯域幅を（十分に）下回る最大限度未満に制限することは、単純ではない。さらに、異なる送信側が異なるSLAレベルに起因して異なる帯域幅割当を割り振られると想定される場合、方程式はさらに複雑になる。

10

【0303】

一実施形態によれば、本明細書のシステムおよび方法は、帯域幅割当の初期交渉、そのような帯域幅割当の動的調整（例えば、利用可能な帯域幅を共有する送信側ノードの数の変化、またはSLAの変化に適應するため）、および全体的な帯域幅割当が同じままであるにもかかわらず送信側が関連するエグレスデータレートを一時的に減速する必要があることを示すための動的輻輳フィードバックの両方を含むように、エンドツーエンド輻輳フィードバックのための旧式のスキームを拡張することができる。明示的な非請求メッセージおよびデータパケット内の「ピギーバック」情報の両方を用いて、ターゲットノードから送信側ノードに関連情報を伝達する。

【0304】

図32は、ある実施形態に係る、高性能コンピューティング環境においてイングレス帯域幅調停および輻輳フィードバックを組み合わせるためのシステムを示す。

20

【0305】

より具体的には、ある実施形態に従うと、図32は、ハイパーバイザ3211を含むホストチャネルアダプタ3201を示す。ハイパーバイザは、VF3214～3216などのいくつかの仮想機能（VF）、および物理機能（PF）3213をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク3200などのネットワークに接続するために使用されるポート3202および3203などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA3201を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

30

【0306】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 3250、VM2 3251、およびVM3 3252のような仮想マシン（VM）をホストすることができる。

【0307】

ある実施形態に従うと、ホストチャネルアダプタ3201は、さらに、ハイパーバイザを介して仮想スイッチ3212をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（vPort）アーキテクチャをさらにサポートすることができる。

40

【0308】

ある実施形態に従うと、ネットワーク3200は、図示のように、スイッチ3240、3241、3242、および3243など、相互接続され、たとえばリーフスイッチ3240および3241を介してホストチャネルアダプタ3201に接続され得る、いくつかのスイッチを備え得る。

【0309】

ある実施形態に従うと、スイッチ3240～3243は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード（たとえば他のHCA

50

）に接続されることができる。

#### 【0310】

ある実施形態に従うと、帯域幅割当およびパフォーマンスの問題は、VM、たとえばVM1 3250が複数のソースからの過剰なイングレス帯域幅3290を受ける場合に生じ得る。これは、たとえば、VM1が1つ以上のRDMA書込動作と同時に1つ以上のRDMA読出応答を受け、VM1上のイングレス帯域幅が2つ以上のソース（例えば、接続されたVMからの1つのRDMA読出応答、および別の接続されたVMからの1つのRDMA書込要求）から来る状況において生じ得る。

#### 【0311】

ある実施形態に従うと、さらに、HCAのレート制限コンポーネント3260は、たとえばVM1についてのイングレス帯域幅制限を、VM1上でイングレス帯域幅を生成する役割を担うノードについてのエグレス帯域幅制限と協調させるために、他のピアHCAと交渉され得るVM固有のレート制限3261をさらに含み得る。そのような初期交渉は、たとえば、利用可能な帯域幅を共有する送信側ノードの数の変化、またはSLAの変化に適応するように実行され得る。これら他のHCA／ノードは、図には示されていない。

#### 【0312】

ある実施形態に従うと、上記交渉は、たとえばイングレス帯域幅の結果として生成される明示的で非請求のフィードバックメッセージ3291に基づいて更新することができる。そのようなフィードバックメッセージ3291は、たとえば、VM1上でイングレス帯域幅3290を生成することを担う複数の遠隔ノードに送信され得る。そのようなフィードバックメッセージを受信すると、送信側ノード（VM1上のイングレス帯域幅を担う帯域幅の送信側）は、QoSおよびSLAを維持しようと試みながら、例えばVM1に接続するリンクなどを過負荷にしないように、送信側ノードに対するそれらの関連するエグレス帯域幅制限を更新することができる。

#### 【0313】

図33は、ある実施形態に係る、高性能コンピューティング環境においてイングレス帯域幅調停および輻輳フィードバックを組み合わせるためのシステムを示す。

#### 【0314】

より具体的には、ある実施形態に従うと、図33は、ハイパーバイザ3311を含むホストチャネルアダプタ3301を示す。ハイパーバイザは、VF3314～3316などのいくつかの仮想機能（VF）、および物理機能（PF）3313をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク3300などのネットワークに接続するために使用されるポート3302および3303などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA3301を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

#### 【0315】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 3350、VM2 3351、およびVM3 3352のような仮想マシン（VM）をホストすることができる。

#### 【0316】

ある実施形態に従うと、ホストチャネルアダプタ3301は、さらに、ハイパーバイザを介して仮想スイッチ3312をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート（vPort）アーキテクチャをさらにサポートすることができる。

#### 【0317】

ある実施形態に従うと、ネットワーク3300は、図示のように、スイッチ3340、3341、3342、および3343など、相互接続され、たとえばリーフスイッチ33

10

20

30

40

50

40および3341を介してホストチャネルアダプタ3301に接続され得る、いくつかのスイッチを備え得る。

【0318】

ある実施形態に従うと、スイッチ3340～3343は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード（たとえば他のHCA）に接続されることができる。

【0319】

ある実施形態に従うと、帯域幅割当およびパフォーマンスの問題は、VM、たとえばVM1 3350が複数のソースからの過剰なイングレス帯域幅3390を受ける場合に生じ得る。これは、たとえば、VM1が1つ以上のRDMA書込動作と同時に1つ以上のRDMA読出応答を受け、VM1上のイングレス帯域幅が2つ以上のソース（例えば、接続されたVMからの1つのRDMA読出応答、および別の接続されたVMからの1つのRDMA書込要求）から来る状況において生じ得る。

【0320】

ある実施形態に従うと、さらに、HCAのレート制限コンポーネント3360は、たとえばVM1についてのイングレス帯域幅制限を、VM1上でイングレス帯域幅を生成する役割を担うノードについてのエグレス帯域幅制限と協調させるために、他のピアHCAと交渉され得るVM固有のレート制限3361をさらに含み得る。そのような初期交渉は、たとえば、利用可能な帯域幅を共有する送信側ノードの数の変化、またはSLAの変化に適應するように実行され得る。これら他のHCA／ノードは、図には示されていない。

【0321】

ある実施形態に従うと、上記交渉は、たとえばイングレス帯域幅の結果として生成されるピギーバックメッセージ3391（エンドノード間で送信される通常のデータまたは他の通信パケットの上に存在するメッセージ）に基づいて更新することができる。そのようなピギーバックメッセージ3391は、たとえば、VM1上でイングレス帯域幅3390を生成することを担う複数の遠隔ノードに送信され得る。そのようなフィードバックメッセージを受信すると、送信側ノード（VM1上のイングレス帯域幅を担う帯域幅の送信側）は、QoSおよびSLAを維持しようと試みながら、例えばVM1に接続するリンクなどを過負荷にしないように、送信側ノードに対するそれらの関連するエグレス帯域幅制限を更新することができる。

【0322】

図34は、ある実施形態に係る、高性能コンピューティング環境においてイングレス帯域幅調停および輻輳フィードバックを組み合わせるためのシステムを示す。

【0323】

ある実施形態に従うと、スイッチドネットワークまたはサブネット3400のような高性能コンピューティング環境内では、いくつかのエンドノード3401および3402は、リーフスイッチ3411および3412、スイッチ3421および3422、ならびにルートスイッチ3431および3432のようないくつかのスイッチを介して相互接続されるいくつかの仮想マシンVM1～VM4 3450～3453をサポートすることができる。

【0324】

ある実施形態に従うと、ノード3401および3402の接続のための機能を提供するさまざまなホストチャネルアダプタ、ならびにサブネットに接続されるべき仮想マシンは図示されていない。そのような実施形態の議論はSR-IOVに関して上記されており、各仮想マシンはホストチャネルアダプタ上のハイパーバイザの仮想機能に関連付けられ得る。

【0325】

ある実施形態に従うと、VM3 3452などのノードは、たとえば3451および3452などの複数のRDMAイングレス帯域幅パケット（たとえば複数のRDMA書込）を受信すると、（たとえばレート制限3461から）イングレス帯域幅制限に入ることが

10

20

30

40

50

できる。これは、例えば、帯域幅制限を調整するために様々な送信側ノード間に通信が存在しない場合に生じ得る。

【0326】

一実施形態によれば、本明細書のシステムおよび方法は、帯域幅割当の初期交渉（すなわち、VM3、またはVM3に関連付けられる帯域幅制限が、イングレストラフィック内でVM3をターゲットとするすべての送信側ノードと交渉する）、そのような帯域幅割当の動的調整（例えば、利用可能な帯域幅を共有する送信側ノードの数の変化、またはSLAの変化に適応するため）、および全体的な帯域幅割当が同じままであるにもかかわらず送信側が関連するエグレスデータレートを一時的に減速する必要があることを示すための動的輻輳フィードバックの両方を含むように、エンドツーエンド輻輳フィードバックのためのスキームを拡張することができる。そのような動的輻輳フィードバックは、例えば、VM3にトラフィックを送信する際に利用するように更新された帯域幅制限について各送信側ノードに指示する、様々な送信側ノードへの返送メッセージ（例えば、フィードバックメッセージ3470）において生じ得る。そのようなフィードバックメッセージ3460は、明示的な非請求メッセージ、およびターゲットノード（すなわち、示される実施形態ではVM3）から送信側ノードに関連情報を伝達するためのデータパケット内の「ピギーバック」情報の形態をとることができる。

10

【0327】

図35は、一実施形態による、高性能コンピューティング環境においてイングレス帯域幅調停と輻輳フィードバックとを組み合わせる方法のフローチャートである。

20

【0328】

ある実施形態に従うと、ステップ3510において、本方法は、1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することができ、第1のサブネットは複数のスイッチを含み、複数のスイッチは少なくともリーフスイッチを含み、複数のスイッチの各々は複数のスイッチポートを含み、第1のサブネットはさらに、複数のホストチャネルアダプタを含み、ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、複数のホストチャネルアダプタは、複数のスイッチを介して相互接続され、第1のサブネットはさらに、複数の仮想マシンを含む複数のエンドノードを含む。

【0329】

ある実施形態に従うと、ステップ3520において、本方法は、ホストチャネルアダプタにおいて、ホストチャネルアダプタに取り付けられたエンドノードに関連付けられるエンドノードイングレス帯域幅割当を提供することができる。

30

【0330】

ある実施形態に従うと、ステップ3530において、本方法は、ホストチャネルアダプタに取り付けられたエンドノードと遠隔エンドノードとの間で帯域幅割当を交渉することができる。

【0331】

ある実施形態に従うと、ステップ3540において、本方法は、ホストチャネルアダプタに取り付けられたエンドノードにおいて、遠隔エンドノードからイングレス帯域幅を受信することができ、イングレス帯域幅はエンドノードのイングレス帯域幅制限を超える。

40

【0332】

ある実施形態に従うと、3550において、遠隔エンドノードからイングレス帯域幅を受信することに応答して、本方法は、ホストチャネルアダプタに取り付けられたエンドノードから遠隔エンドノードに応答メッセージを送信することができ、応答メッセージは、ホストチャネルアダプタに取り付けられたエンドノードのイングレス帯域幅割当を超えていることを示す。

【0333】

FECN（順方向明示的輻輳通知）シグナリングおよびBECN（逆方向明示的輻輳通知）シグナリングの両方における複数のCE（輻輳経験）フラグの使用（ORA200246-US-NP-4）

50

一実施形態によれば、従来、輻輳通知は、あるポイント（例えば、ネットワーク／ファブリックトポロジーを通る送信側からターゲットへの経路に沿ったいくつかのノード／スイッチペア間のなんらかのリンクセグメント）で輻輳に遭遇するデータパケットに基づいており、「輻輳」ステータスフラグ（C Eフラグとも呼ばれる）でマーキングされ、次いで、このステータスは、ターゲットから送信側に返送される応答パケットに反映される。

【0334】

ある実施形態に従うと、このスキームの問題は、送信側ノードが、同じリンクセグメントにおいて輻輳の対象となるフロー間を、それらが異なるターゲットを表すにもかかわらず、区別することを可能にしないことである。また、複数の経路が送信側ノードとターゲット側ノードとのペア間で利用可能であるとき、異なる代替経路についての輻輳についてのいかなる情報も、関連する経路を介する関連するターゲットについて何らかのフローがアクティブであることを必要とする。

【0335】

ある実施形態に従うと、本願明細書に記載のシステムおよび方法は、輻輳マーキングスキームを拡張して、同じパケットにおける複数のC Eフラグを容易にし、どのC Eフラグインデックスを更新すべきかを定義するステージ番号を表すようにスイッチポートを構成する。特定の送信側と特定のターゲットとの間で、順序付けられたスイッチポートのシーケンスを通る特定の経路は、一意のステージ番号の特定の順序付けられたリストを表し、それによってC Eフラグインデックス番号も表す。

【0336】

ある実施形態によれば、このようにして、複数のC Eフラグがセットされた輻輳フィードバックを受信する送信側ノードが、様々なC Eフラグを、関連する輻輳条件状態および関連付けられる動的レート低減を表すことになる異なる「ターゲットグループ」コンテキストにマッピングすることが可能である。さらに、異なるターゲットに対する異なるフローは、送信側ノードにおいて共有「ターゲットグループ」によって表される共有リンクセグメントに関連付けられる輻輳情報および動的レート低減状態を共有することになる。

【0337】

ある実施形態に従うと、輻輳が発生する場合、鍵となる問題は、輻輳フィードバックが、理想的には、輻輳フィードバックを受信するフローに関連付けられる階層におけるすべての関連するターゲットグループに関連付けられるべきであるということである。次いで、影響を受けたターゲットグループは、応じて、最大レートを動的に調整すべきである。したがって、各ターゲットグループのHW状態は、任意の現在の輻輳ステータスおよび関連する「スロットル情報」も含まなければならない。

【0338】

ある実施形態に従うと、ここでの重要な局面は、輻輳を検出するスイッチがトポロジーにおけるそのステージに対応するフラグをマーキングできるように、F E C Nシグナリングが複数の「輻輳経験」（C E）フラグを含む能力を有するべきであることである。一通常のファットツリーでは、各スイッチは、上方向に一意の（最大）ステージ番号を有し、下方向に別の一意の（最大）ステージ番号を有する。したがって、特定の経路を用いるフローは、次いで、完全なファブリックにおけるステージ番号のセット全体のすべてまたはサブセットのみを含むことになる特定のステージ番号のシーケンスに関連付けられることになる。しかしながら、その特定のフローについて、その経路に関連付けられる様々なステージ番号は、次いで、そのフローに関連付けられる1つ以上のターゲットグループにマッピングされ得る。このようにして、フローに対する受信されたB E C Nは、B E C N内の各C Eフラグ付きステージに関連付けられるターゲットグループが輻輳を示すように更新されることになることを意味することができ、これらのターゲットグループに対する動的最大レートは、その後、それに応じて調整することができる。

【0339】

ある実施形態に従うと、本来ファットツリートポロジーに適しているが、スイッチの「ステージ番号」概念は、そのような番号をスイッチに割り当てることが可能であるほとん

10

20

30

40

50

どすべてのトポロジを表すように一般化することができる。しかしながら、このような一般的な場合、ステージ番号は単に出力ポートの関数ではなく、各入出力ポート番号タプル関数の関数である。ステージ番号の必要量およびターゲットグループへの経路固有マッピングもまた、一般的な場合においては、より複雑である。したがって、この文脈では、推論は、ファットツリートポロジのみを仮定する。

#### 【0340】

ある実施形態に従うと、単一の packets における複数の CE フラグは、現在、標準プロトコルヘッダに関して、サポートされる特徴ではない。したがって、これは、標準ヘッダの拡張に基づいてサポートすることができるかもしれない、および／またはフローに追加の独立した FECN packets を挿入することによってサポートすることができるかもしれない。概念的には、フローにおける追加の packets の生成は、スイッチ内のカプセル化スキームの使用と同様であり、その影響は、より多くの「オーバーヘッドバイト」が下流に送信されなければならないので、ワイヤ速度で受信されている packets が同じワイヤ速度で転送され得ないことである。追加の packets を挿入することは、典型的には、カプセル化よりも多くのオーバーヘッドになるが、このオーバーヘッドが複数のデータ packets にわたって償却される限り（あらゆるデータ packets に対してそのような追加の通知を送信する必要はない）、このオーバーヘッドは許容可能である可能性が高くなることになる。

#### 【0341】

ある実施形態によれば、スイッチファームウェアがスイッチ内の輻輳条件を監視し、その結果、関連する送信側ノードに「非請求 BECN」を送信することができるスキームを有することも可能である。しかしながら、これは、スイッチファームウェアが、関連する送信側、ならびに、輻輳を経験する packets にどのようなアドレスが関与するかに関する動的情報も含み得る、ポートと優先度と関連する送信側との間のマッピングに関する、より多くの状態情報を有さなければならないことを意味する。

#### 【0342】

ある実施形態に従うと、RC QP の場合、「CE フラグからターゲットグループへの」マッピングは、典型的には QP コンテキストの一部であることになり、ACK／応答 packets において受信される任意の BECN 情報は、それによって、関連の QP コンテキストおよび関連付けられるターゲットグループについて、単純な態様で扱われることになる。しかしながら、「未請求の BECN」の場合（例えば、アプリケーションレベル応答／ACK のみを有するデータグラムトラフィックの結果として、または「輻輳警告」が複数の潜在的な送信側にブロードキャストされる結果として）、逆方向マッピングは、一少なくとも、HW によって自動的に処理される点においては単純ではない。したがって、よりよい手法は、FECN が、両方とも、接続された（RC）フローの場合に、自動 HW 生成 BECN に至り得るが、HW 自動 BECN 生成を伴う FECN イベントおよび HW 生成 BECN を伴わない FECN イベントの両方が、FECN を受信する HCA に関連付けられるファームウェアおよび／またはハイパー特権ソフトウェアによって処理され得るスキームを有することである。このようにして、観測された輻輳によって影響を受ける 1 つ以上の潜在的な送信側に送信される FW／SW 生成「未請求 BECN」があり得る。これらの「非請求 BECN」を受信する FW／SW は、次いで、受信された「BECN メッセージ」内のペイロードデータに基づいて、関連するローカルターゲットグループへのマッピングを実行することができ、次いで、ローカル HW をトリガして、RC 関連 BECN の完全 HW 制御処理において生ずるものと同様に、ターゲットグループ状態を更新することができる。

#### 【0343】

ある実施形態に従うと、BECN 通知なしの RC ACK／応答 packets、または CE フラグがセットされたステージ番号のサブセットが以前に記録された状態とは異なる（少ない）場合は、ローカル HCA 内の関連するターゲットグループの対応する更新につながってもよい。同様に、「非請求 BECN」が、応答側 HCA（すなわち、関連する sw／fw）によって送信されて、以前にシグナリングされた輻輳がもはや存在しないことを示

することができる。

【0344】

ある実施形態に従うと、上述したように、HWレベルまたはFW/SWレベルのいずれかにおいて動的輻輳フィードバックと組み合わせられたターゲットグループ概念は、HCAによって、ならびに個々のvHCAおよび物理HCAを共有するテナントによって生成されるエグレス帯域幅の柔軟な制御を提供する。

【0345】

ある実施形態に従うと、ターゲットグループは、VMレベルでの関連付けられる遠隔アドレスおよび経路情報とは完全に独立して識別されるので、ターゲットグループの使用と、VMからの通信がオーバーレイまたは他の仮想ネットワークスキームに基づく程度との間に依存性はない。唯一の要件は、HCAリソースを制御するハイパー特権ソフトウェアが関連するマッピングを定義できることである。また、VM/vHCAレベルにおいて、HCAによって実際のターゲットグループにマッピングされる「論理ターゲットグループID」を有するスキームを用いることが可能であろう。しかしながら、実際のターゲットグループIDをテナントから隠すことを除いて、これが有用であることは明らかではない。—基礎となる経路が変化した場合、どのようなターゲットグループが特定の宛先に関連付けられているかを更新する必要がある場合、これは他の宛先を伴わないことがある。したがって、一般的な場合、ターゲットグループの更新は、単に論理—物理ターゲットグループIDマッピングを更新するのではなく、すべての関与するQPおよびアドレスハンドルを更新することを伴う必要がある。

【0346】

ある実施形態に従うと、仮想化されたターゲットHCAについて、物理HCAポートではなく個々のvHCAポートを最終宛先ターゲットグループとして表すことが可能である。このようにして、遠隔ピアノードのためのターゲットグループ階層は、宛先物理HCAポートを表すターゲットグループと、vHCAポートに関して最終宛先を表す追加のターゲットグループとの両方を含むことができる。このようにして、本システムおよび方法は、物理HCAポートあたりの帯域幅および関連付けられる送信側ターゲットグループが、vHCAポートあたりのイングレス帯域幅割当の合計が物理HCAポート帯域幅（または関連付けられる帯域幅割当）未満に保たれる必要がないことを意味すると同時に、個々のvHCAポート（VF）のイングレス帯域幅を制限する能力を有する。

【0347】

ある実施形態に従うと、送信側HCA内では、ターゲットグループを用いて、異なるターゲットグループを異なるテナントに割り当てることによって、エグレス方向における物理HCAポートの共有を表すことが可能である。また、物理HCAポートのためのテナントレベルターゲットグループを共有する同じテナントからの複数のVMを容易にするために、異なるターゲットグループを異なるそのようなVMに割り当てるのが可能である。そのようなターゲットグループは、次いで、そのVMからのすべてのエグレス通信のための初期ターゲットグループとしてセットアップされることになる。

【0348】

図36は、ある実施形態に係る、高性能コンピューティング環境においてFECNおよびBECNの両方において複数のCEフラグを用いるためのシステムを示す。

【0349】

より具体的には、ある実施形態に従うと、図36は、ハイパーバイザ3611を含むホストチャネルアダプタ3601を示す。ハイパーバイザは、VF3614~3616などのいくつかの仮想機能（VF）、および物理機能（PF）3613をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク3600などのネットワークに接続するために使用されるポート3602および3603などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA3601を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークま

たは R o C E ネットワークなどのスイッチドネットワークを備えることができる。

【0350】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 3650、VM2 3651、および VM3 3652 のような仮想マシン (VM) をホストすることができる。

【0351】

ある実施形態に従うと、ホストチャネルアダプタ 3601 は、さらに、ハイパーバイザを介して仮想スイッチ 3612 をサポートすることができる。これは、v S w i t c h アーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート (v P o r t) アーキテクチャをさらにサポートすることができる。

10

【0352】

ある実施形態に従うと、ネットワーク 3600 は、図示のように、スイッチ 3640、3641、3642、および 3643 など、相互接続され、たとえばリーフスイッチ 3640 および 3641 を介してホストチャネルアダプタ 3601 に接続され得る、いくつかのスイッチを備え得る。

【0353】

ある実施形態に従うと、スイッチ 3640 ~ 3643 は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード (たとえば他の H C A) に接続されることができる。

20

【0354】

ある実施形態に従うと、イングレスパケット 3690 は、ネットワークを横断している間、その経路の任意のステージにおいて輻輳を経験し得、パケットは、ステージのいずれかにおいてそのような輻輳を検出すると、スイッチによってマーキングされることができる。パケットを、輻輳を経験したとマーキングすることに加えて、マーキングを行うスイッチは、パケットが輻輳を経験したステージをさらに示すことができる。宛先ノード、例えば VM1 3650 に到着すると、VM1 は、パケットが輻輳を経験したこと、およびどのステージでパケットが輻輳を経験したかを送信ノードに示すことができる明示的なフィールドバックメッセージ 3691 を介して応答パケットを (例えば自動的に) 送信することができる。

30

【0355】

ある実施形態に従うと、イングレスパケットは、パケットがどこで輻輳を経験したかを示すように更新されるビットフィールドを含むことができ、明示的フィールドバックメッセージは、このような輻輳を送信側ノードに通知する際にこのビットフィールドをミラーリング／表現することができる。

【0356】

ある実施形態に従うと、各スイッチポートは、サブネット全体におけるあるステージを表す。したがって、サブネットにおいて送信される各パケットは、最大数のステージを横断することができる。輻輳がどこ (複数の場所にあり得る) で検出されたかを識別するために、輻輳マーキング (例えば、C E フラグ) は、単なるバイナリフラグ (輻輳を経験) から複数のビットを含むビットフィールドに拡張される。ビットフィールドの各ビットは、次いで、各スイッチポートに割り当てられ得るステージ番号に関連付けられ得る。例えば、3 ステージからなるファットツリーでは、ステージの最大数は 3 であろう。システムが A から B への経路を有し、ルーティングが既知である場合、各エンドノードは、経路の任意の所与のステージでパケットがどのスイッチポートを介して横断したかを判断することができる。そうすることによって、各エンドノードは、ルーティングを、受信された輻輳メッセージと相関させることによって、パケットがどの別個のスイッチポートにおいて輻輳を経験したかを判断することができる。

40

【0357】

ある実施形態に従うと、システムは、どのステージで輻輳が検出されるかを示す、戻っ

50

てくる輻輳フィードバックを提供し得、次いで、エンドノードが共有リンクセグメントに起因して生じる輻輳を有する場合、輻輳制御は、異なるエンドポートではなく、そのセグメントに適用される。これは、輻輳に関して、より細かい粒度の情報を提供する。

【0358】

ある実施形態に従うと、このようなより細かい粒度を提供することにより、エンドノードは次いで、将来のパケットをルーティングする際に代替経路を用いることができる。または、例えば、あるエンドノードが、すべてが異なる宛先に行く複数のフローを有するが、輻輳が経路の共通のステージで検出される場合、再ルーティングがトリガされ得る。本システムおよび方法は、10個の異なる輻輳通知を有するのではなく、それに関連付けられるスロットリングに関して即時の反応を提供する。これは、より効率的な輻輳通知の処理である。

10

【0359】

図37は、ある実施形態による、高性能コンピューティング環境においてFECNおよびBECNの両方において複数のCEフラグを用いるためのシステムを示す。

【0360】

より具体的には、ある実施形態に従うと、図37は、ハイパーバイザ3711を含むホストチャネルアダプタ3701を示す。ハイパーバイザは、VF3714~3716などのいくつかの仮想機能(VF)、および物理機能(PF)3713をホストするかまたはそれらに関連付けられることができる。ホストチャネルアダプタは、ホストチャネルアダプタをネットワーク3700などのネットワークに接続するために使用されるポート3702および3703などのいくつかのポートをさらにサポートするかまたは備えることができる。ネットワークは、たとえば、HCA3701を、スイッチ、追加の別個のHCAなどのいくつかの他のノードに接続することができる、インフィニバンドネットワークまたはRoCEネットワークなどのスイッチドネットワークを備えることができる。

20

【0361】

ある実施形態に従うと、上述したように、仮想機能の各々は、VM1 3750、VM2 3751およびVM3 3752のような仮想マシン(VM)をホストすることができる。

【0362】

ある実施形態に従うと、ホストチャネルアダプタ3701は、さらに、ハイパーバイザを介して仮想スイッチ3712をサポートすることができる。これは、vSwitchアーキテクチャが実現される状況のためである。図示されていないが、本開示の実施形態は、上述のように、仮想ポート(vPort)アーキテクチャをさらにサポートすることができる。

30

【0363】

ある実施形態に従うと、ネットワーク3700は、図示のように、スイッチ3740, 3741, 3742, および3743など、相互接続され、たとえばリーフスイッチ3740および3741を介してホストチャネルアダプタ3701に接続され得る、いくつかのスイッチを備え得る。

【0364】

ある実施形態に従うと、スイッチ3740~3743は相互接続されることができ、さらに、図には示されていない他のスイッチおよび他のエンドノード(たとえば他のHCA)に接続されることができる。

40

【0365】

ある実施形態に従うと、イングレスパケット3790は、ネットワークを横断している間、その経路の任意のステージにおいて輻輳を経験し得、パケットは、ステージのいずれかにおいてそのような輻輳を検出すると、スイッチによってマーキングされることができる。パケットを、輻輳を経験したとマーキングすることに加えて、マーキングを行うスイッチは、パケットが輻輳を経験したステージをさらに示すことができる。宛先ノード、例えばVM1 3750に到着すると、VM1は、パケットが輻輳を経験したこと、およびどのステージでパケットが輻輳を経験したかを送信ノードに示すことができるピギーバック

50

クメッセージ（受信側ノードから送信側ノードに送信される別のメッセージ／パケット上にあるメッセージ）３７９１を介して応答パケットを（例えば自動的に）送信することができる。

#### 【０３６６】

ある実施形態に従うと、イングレスパケットは、パケットがどこで輻輳を経験したかを示すように更新されるビットフィールドを含むことができ、明示的フィードバックメッセージは、このような輻輳を送信側ノードに通知する際にこのビットフィールドをミラーリング／表現することができる。

#### 【０３６７】

ある実施形態に従うと、各スイッチポートは、サブネット全体におけるあるステージを表す。したがって、サブネットにおいて送信される各パケットは、最大数のステージを横断することができる。輻輳がどこ（複数の場所にあり得る）で検出されたかを識別するために、輻輳マーキング（例えば、ＣＥフラグ）は、単なるバイナリフラグ（輻輳を経験）から複数のビットを含むビットフィールドに拡張される。ビットフィールドの各ビットは、次いで、各スイッチポートに割り当てられ得るステージ番号に関連付けられ得る。例えば、３ステージからなるファットツリーでは、ステージの最大数は３であろう。システムがＡからＢへの経路を有し、ルーティングが既知である場合、各エンドノードは、経路の任意の所与のステージでパケットがどのスイッチポートを介して横断したかを判断することができる。そうすることによって、各エンドノードは、ルーティングを、受信された輻輳メッセージと相関させることによって、パケットがどの別個のスイッチポートにおいて輻輳を経験したかを判断することができる。

#### 【０３６８】

ある実施形態に従うと、システムは、どのステージで輻輳が検出されるかを示す、戻ってくる輻輳フィードバックを提供し得、次いで、エンドノードが共有リンクセグメントに起因して生じる輻輳を有する場合、輻輳制御は、異なるエンドポートではなく、そのセグメントに適用される。これは、輻輳に関して、より細かい粒度の情報を提供する。

#### 【０３６９】

ある実施形態に従うと、このようなより細かい粒度を提供することにより、エンドノードは次いで、将来のパケットをルーティングする際に代替経路を用いることができる。または、例えば、あるエンドノードが、すべてが異なる宛先に行く複数のフローを有するが、輻輳が経路の共通のステージで検出される場合、再ルーティングがトリガされ得る。本システムおよび方法は、１０個の異なる輻輳通知を有するのではなく、それに関連付けられるスロットリングに関して即時の反応を提供する。これは、より効率的な輻輳通知の処理である。

#### 【０３７０】

図３８は、ある実施形態に係る、高性能コンピューティング環境においてＦＥＣＮおよびＢＥＣＮの両方において複数のＣＥフラグを用いるためのシステムを示す。

#### 【０３７１】

ある実施形態に従うと、スイッチドネットワークまたはサブネット３８００のような高性能コンピューティング環境内では、いくつかのエンドノード３８０１および３８０２は、リーフスイッチ３８１１および３８１２、スイッチ３８２１および３８２２、ならびにルートスイッチ３８３１および３８３２のようないくつかのスイッチを介して相互接続されるいくつかの仮想マシンＶＭ１～ＶＭ４ ３８５０～３８５３をサポートすることができる。

#### 【０３７２】

ある実施形態に従うと、ノード３８０１および３８０２の接続のための機能を提供するさまざまなホストチャネルアダプタ、ならびにサブネットに接続されるべき仮想マシンは図示されていない。そのような実施形態の議論はＳＲ－ＩＯＶに関して上記されており、各仮想マシンはホストチャネルアダプタ上のハイパーバイザの仮想機能に関連付けられ得る。

10

20

30

40

50

## 【0373】

ある実施形態に従うと、ノード、VM3 3852からVM1 3850へ送られるパケット3851は、図に示されるように、いくつかのリンク、またはステージ1～ステージ6などのステージを介して、サブネット3800を横断することができる。パケット3851は、サブネットを横断している間、これらのステージのいずれかにおいて輻輳を経験し得、ステージのいずれかにおいてそのような輻輳を検出すると、スイッチによってマーキングされ得る。パケットを、輻輳を経験したとマーキングすることに加えて、マーキングを行うスイッチは、パケットが輻輳を経験したステージをさらに示すことができる。宛先ノードVM1に到達すると、VM1は、パケットが輻輳を経験したこと、およびどのステージでパケットが輻輳を経験したかをVM3 3852に示すことができるフィードバックメッセージ3870を介して応答パケットを（例えば、自動的に）送信することができる。

10

## 【0374】

ある実施形態に従うと、各スイッチポートは、サブネット全体におけるあるステージを表す。したがって、サブネットにおいて送信される各パケットは、最大数のステージを横断することができる。輻輳がどこ（複数の場所にあり得る）で検出されたかを識別するために、輻輳マーキング（例えば、CEフラグ）は、単なるバイナリフラグ（輻輳を経験）から複数のビットを含むビットフィールドに拡張される。ビットフィールドの各ビットは、次いで、各スイッチポートに割り当てられ得るステージ番号に関連付けられ得る。例えば、3ステージからなるファットツリーでは、ステージの最大数は3であろう。システムがAからBへの経路を有し、ルーティングが既知である場合、各エンドノードは、経路の任意の所与のステージでパケットがどのスイッチポートを介して横断したかを判断することができる。そうすることによって、各エンドノードは、ルーティングを、受信された輻輳メッセージと関連させることによって、パケットがどの別個のスイッチポートにおいて輻輳を経験したかを判断することができる。

20

## 【0375】

ある実施形態に従うと、システムは、どのステージで輻輳が検出されるかを示す、戻ってくる輻輳フィードバックを提供し得、次いで、エンドノードが共有リンクセグメントに起因して生じる輻輳を有する場合、輻輳制御は、異なるエンドポートではなく、そのセグメントに適用される。これは、輻輳に関して、より細かい粒度の情報を提供する。

30

## 【0376】

ある実施形態に従うと、このようなより細かい粒度を提供することにより、エンドノードは次いで、将来のパケットをルーティングする際に代替経路を用いることができる。または、例えば、あるエンドノードが、すべてが異なる宛先に行く複数のフローを有するが、輻輳が経路の共通のステージで検出される場合、再ルーティングがトリガされ得る。本システムおよび方法は、10個の異なる輻輳通知を有するのではなく、それに関連付けられるスロットリングに関して即時の反応を提供する。これは、より効率的な輻輳通知の処理である。

## 【0377】

図39は、ある実施形態に係る、高性能コンピューティング環境においてFECNおよびBECNの両方において複数のCEフラグを用いる方法のフローチャートである。

40

## 【0378】

ある実施形態に従うと、ステップ3910において、本方法は、1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することができ、第1のサブネットは複数のスイッチを含み、複数のスイッチは少なくともリーフスイッチを含み、複数のスイッチの各々は複数のスイッチポートを含み、第1のサブネットはさらに、複数のホストチャネルアダプタを含み、ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、複数のホストチャネルアダプタは、複数のスイッチを介して相互接続され、第1のサブネットはさらに、複数の仮想マシンを含む複数のエンドノードを含む。

## 【0379】

50

ある実施形態に従うと、ステップ 3920 において、本方法は、ホストチャネルアダプタに取り付けられたエンドノードにおいて、遠隔エンドノードからインGRES パケットを受信することができ、インGRES パケットは、エンドノードにおいて受信される前に第 1 のサブネットの少なくとも一部を横断し、インGRES パケットは、インGRES パケットが第 1 のサブネットの少なくとも一部の横断中に輻輳を経験したことを示すマーキングを含む。

#### 【0380】

ある実施形態に従うと、インGRES パケットを受信すると、ステップ 3930 において、方法は、エンドノードによって、ホストチャネルアダプタに取り付けられたエンドノードから遠隔エンドノードに応答メッセージを送信することができ、応答メッセージは、インGRES パケットが第 1 のサブネットの少なくとも一部の横断中に輻輳を経験したことを示し、応答メッセージはビットフィールドを含む。

#### 【0381】

#### プライベートファブリックなどのスイッチドファブリックにおける QOS および SLA

ある実施形態によると、クラウドならびに顧客および構内設置におけるより大きいクラウド内のプライベートネットワークファブリック（例えば、専用分散型アプライアンスまたは汎用高性能コンピューティングリソースを構築するために使用されるもの等のプライベートファブリック）は、VM ベースのワークロードを展開する能力を望み、ある固有の要件は、異なるタイプの通信フローに対してサービス品質（QOS）を定義および制御することが可能であることである。加えて、異なるテナントに属するワークロードは、そのようなワークロード間の干渉を最小限に抑え、異なる通信タイプに対する QOS 仮定を維持しながら、関連するサービスレベル合意（SLA）の境界内で実行されなければならない。

#### 【0382】

ある実施形態に従って、以下のセクションでは、関連する問題シナリオ、目標および潜在的な解決策について論じる。

#### 【0383】

ある実施形態に従うと、クラウド顧客（別名「テナント」）にファブリックリソースをプロビジョニングするための初期スキームは、テナントにラックの専用部分（たとえば割り当てラック）または 1 つ以上のフルラックを割り振ることができることである。この粒度は、割り振られたリソースが完全に動作可能である限り、各テナントが常に満たされる通信 SLA を有することを保証されることを意味する。これは、単一のラックが複数の部分に分割される場合にも当てはまり、なぜならば、粒度は常に HCA を伴う完全な物理サーバであるためである。単一のラック内の異なるそのようなサーバ間の接続性は、原則として、常に単一のフルクロスバースイッチを介することができる。この場合、同じテナントに属するサーバのセット間の通信トラフィックの結果として異なるテナントに属するフロー間の回線争奪または輻輳につながり得る方法で共有されるリソースは存在しない。

#### 【0384】

しかしながら、実施形態によれば、冗長スイッチが共有されるので、あるサーバ上のワークロードによって生成されるトラフィックが別のテナントに属するサーバをターゲットにすることは不可能であることは重大である。そのようなトラフィックは、テナント間のいかなる通信またはデータ漏洩／観察も容易にしないが、その結果は、他のテナントに属する通信フローに対する深刻な干渉または DOS（サービス不全）のような影響でさえあり得る。

#### 【0385】

ある実施形態に従うと、フルクロスバーリーフスイッチが、ローカルサーバ間のすべての通信がローカルスイッチのみを介して行われ得ることを、元来意味するという事実にもかかわらず、これが可能でないかもしれないか、または他の実際的な問題のために達成されないいくつかのケースがある。

・ある実施形態に従うと、たとえばホストバス（PCIE）生成が一度に 1 つのファブリ

10

20

30

40

50

ックリンクの帯域幅しか持続できない場合、任意の時点で1つのH C Aポートのみがデータトラフィックのために使用されていることは重要である。したがって、すべてのサーバがデータトラフィックのためにどのローカルスイッチを用いるかについて合意していない場合、いくつかのトラフィックは、ローカルリーフスイッチ間のスイッチ間リンク（I S L）を通過しなければならないことになる。

- ・ある実施形態に従うと、1つ以上のサーバがスイッチの1つへの接続性を失った場合、すべての通信は他方のスイッチを介して行われなければならないことになる。したがって、またもや、サーバのすべてのペアが同じ単一のスイッチの使用について合意できない場合、いくつかのデータトラフィックはI S Lを通過しなければならないことになる。

- ・ある実施形態に従うと、サーバが両方のH C Aポート（したがって両方のリーフスイッチ）を用いることができるが、接続が同じスイッチに接続するH C Aポートを介してのみ確立されることを実施することが不可能である場合、いくつかのデータトラフィックはI S Lを通過する場合がある。

○このスキームで終わる1つの理由は、ファブリックホストスタックにおけるソケット／ポート番号の欠如が、プロセスが着信接続を受け入れるために1つのソケットしか確立できないことを意味することである。その場合、このソケットは、一度に単一のH C Aポートにしか関連付けることができない。同じ単一のソケットが、発信接続を確立するときにも使用される限り、多数のプロセスはそれらの単一ソケットがローカルH C Aポート間で均等に分散されるにもかかわらず、システムは、I S Lを必要とする、ある数の接続に終わることになる。

#### 【0386】

一実施形態によれば、上記で概説したI S L使用／共有を課す特別なケースの単一ラックのシナリオに加えて、プロビジョニングの粒度が、各ラック内のリーフスイッチがスパインスイッチによって相互接続されるマルチラック構成に拡張されると、次いで、異なるテナントのための通信S L Aは、どのサーバがどのテナントに割り振られるか、および異なる通信フローが異なるスイッチースイッチリンク上にファブリックレベルルーティングスキームによってどのようにマッピングされるかに大きく依存するようになる。このシナリオにおける重要な問題は、2つの最適化局面が幾分矛盾することである。

- ・ある実施形態に従うと、一方では、可能な限り最良の性能を提供するために、異なる宛先ポートをターゲットとするすべての同時フローは、ファブリックを通る異なる経路（すなわち、異なるスイッチースイッチリンクーI S L）をできるだけ多く使用すべきである。

- ・実施形態によれば、他方で、異なるテナントに対して予測可能なQ O SおよびS L Aを提供するために、異なるテナントに属するフローが帯域幅について同じI S L上で同時に競合しないことが重要である。一般的な場合、これは、異なるテナントによってどの経路を使用できるかについての制限が存在しなければならないことを意味する。

#### 【0387】

しかしながら、ある実施形態に従うと、システムのサイズ、テナントの数、およびサーバが異なるテナントにどのようにプロビジョニングされたかに依存するいくつかの状況においては、異なるテナントに属するフローが同じI S L上で帯域幅について競合することになることを回避することが可能ではない場合がある。この状況では、問題に対処し、回線争奪の可能性を低減するために、ファブリックの観点から用いることができる、主に2つの手法がある。

- ・ある実施形態に従うと、同じI S L帯域幅について競合するにもかかわらず、異なるテナントからのフローが他のテナントとは無関係に順方向進行を有することを確実にするために、どのスイッチバッファリソースが異なるテナント（またはテナントのグループ）によって占有され得るかを制限する。

- ・ある実施形態に従うと、他のテナントを犠牲にして1つのテナントによって消費され得る最大帯域幅を制限する「許可制御」メカニズムを実現する。

#### 【0388】

ある実施形態に従うと、物理的なファブリック構成について、1つの問題は、二分帯域

10

20

30

40

50

幅が可能な限り高く、理想的には非ブロッキングであるかまたはオーバープロビジョニングさえされることである。しかしながら、非ブロッキング二分帯域幅であっても、1つ以上のテナントへのサーバの現在の割振りから、それら異なるテナントのための所望のSLAを達成することが困難であるシナリオがあり得る。そのような状況では、最良の手法は、概して、独立ISLおよび二分帯域幅の必要性を低減させるために、異なるテナントのためのサーバのうちの少なくともいくつかの再プロビジョニングを行うことであろう。

#### 【0389】

ある実施形態に従うと、いくつかのマルチラックシステムはブロッキングファットツリートポロジを有し、これの背後にある仮定は、関連する通信サーバがかなりの程度まで同じラック内に位置するようにワークロードがプロビジョニングされることになり、帯域幅利用のかなりの部分がローカルリーフスイッチ上のポート間のみにあることを意味する、ということである。また、従来のワークロードでは、データトラフィックの大部分は、1つの固定されたノードのセットから別の固定されたノードのセットまでである。しかしながら、不揮発性メモリならびにより新たな通信およびストレージミドルウェアを伴う次世代サーバでは、ある実施形態によると、異なるサーバが同時に複数の機能を提供し得るため、通信ワークロードは、さらにより要求が厳しくなり、予測可能性がより低くなることになる。

#### 【0390】

ある実施形態に従うと、ある目標は、物理サーバレベルとは対照的なVMレベルにあるテナントあたりのプロビジョニング粒度を提供することである。また、その目標は、同じ物理サーバ上で数十までのVMを展開することができ、同じ物理サーバ上の異なるセットのVMが異なるテナントに属し得、様々なテナントが、各々、異なる特性を有する複数のワークロードを表し得ることである。

#### 【0391】

実施形態によれば、加えて、現在のファブリック展開は、（例えば、ロックメッセージが大きなバルクデータ転送の後に「ストール」されることを防止するよう、）異なるフロータイプに対して基本的なQOS（トラフィック分離）を提供するために、異なるサービスのタイプ（TOS）関連付けを用いてきたが、異なるテナントに対して通信SLAも提供することが望ましい。これらのSLAは、たとえワークロードが他のテナントによって共有される物理インフラストラクチャ上でプロビジョニングされていたとしても、テナントが予想に従うワークロードスループットおよび応答時間を経験することを保証すると想定される。テナントに対する関連のSLAは、他のテナントに属するワークロードによる同時アクティビティとは無関係に満たされるものとする。

#### 【0392】

ある実施形態に従うと、ワークロードは、固定（最小および／または最大）セットの物理サーバ上の固定（最小）セットのCPUコア／スレッドおよび物理メモリをプロビジョニングされてもよいが、固定／保証されたネットワーキングリソースのプロビジョニングは、一般に、展開がサーバにおけるHCA／NICの共有を意味する限り、それほど簡単ではない。HCAの共有はまた、少なくともファブリックに対するインGRESリンクおよびエGRESリンクが異なるテナントによって共有されることを元来意味する。したがって、異なるCPUコア／スレッドは真に並列に動作することができるが、何らかの種類の帯域幅多重化または「時分割」を除いて、単一のファブリックリンクの容量を分割する方法はない。この基本帯域幅共有は、バッファ選択／割振りおよびファブリック内での帯域幅調停を実現するときに考慮される異なる「QOS ID」（例えば、サービスレベル、優先順位、DSCP、トラフィッククラスなど）の使用と組み合わせられてもされなくてもよい。

#### 【0393】

ある実施形態に従うと、全体的なサーバメモリ帯域幅は、いくつかのコアにおけるメモリ集約的なワークロードが他のコアについて遅延を課すことを防止するために、任意の個々のCPUコアの典型的なメモリ帯域幅ニーズと比較して非常に高くあるべきである。同

10

20

30

40

50

様に、理想的なケースでは、物理サーバのための利用可能なファブリック帯域幅は、サーバを共有する各テナントが、関連するワークロードが発生させる通信アクティビティに対して十分な帯域幅を有することを可能にするよう、充分大きくあるべきである。しかしながら、いくつかのワークロードが、すべて、大容量データ転送を実行しようとするとき、複数のテナントが、全リンク帯域幅を  $-100\text{Gb/s}$  以上であっても利用し得る可能性が非常に高い。このシナリオに対処するために、同じ物理サーバ上への複数のテナントのプロビジョニングは、利用可能な帯域幅の少なくとも所与の最小パーセンテージを各テナントが得るよう保証されることを確実にする方法で行われ得ることが、必要とされる。しかしながら、RDMAベースの通信では、テナントがエグレス方向にどのくらいの帯域幅を生成することができるかについての制限を実施する能力は、イングレス帯域幅が同じ方法で制限され得ることを意味しない。すなわち、複数の遠隔通信ピアは、すべて、各送信側が最大送信帯域幅によって制限されるにもかかわらず、受信側を完全に過負荷にする方法で、同じ宛先にデータを送信する場合がある。また、RDMA読出動作は、ほんのわずかなエグレス帯域幅を用いてローカルテナントから生じ得る。これは、バルクRDMA読出動作が複数の遠隔ピアのために生じる場合、壊滅的なイングレス帯域幅をもたらす潜在性があり得る。したがって、単一のサーバ上で単一のテナントによって使用される総ファブリック帯域幅を制限するためにエグレス帯域幅に最大制限を課することは充分ではない。

#### 【0394】

ある実施形態によると、本システムおよび方法は、テナントが、RDMA読出動作の使用とは無関係に、ならびにアクティブデータトラフィックを伴う遠隔ピアの数とは無関係に、および遠隔ピアの帯域幅制限とは無関係に、イングレスまたはエグレス方向のいずれかにおいて、関連リンク帯域幅のその相対的部分を決して超えないであろうことを確実にするであろう、テナントのための平均帯域幅制限を構成することができる。（これを達成する方法は、以下の「長期目標」のセクションで論じられる。）

ある実施形態に従うと、本システムおよび方法が通信帯域幅制限のすべての局面を実施することができない限り、テナントのための最高レベルの通信SLAは、テナントが物理サーバを他のテナントと共有できないこと、または潜在的に、テナントが物理HCAを他のテナントと共有しないことになること（つまり、複数の物理HCAを有するサーバの場合）を制限することによってしか、達成できない。物理HCAが、アクティブモードにおいて、両方のHCAポートについて全リンク帯域幅利用で動作できる場合、所与のテナントが通常の場合においてHCAポートの一方への排他的アクセスを与えられる制限を用いることも考えられる。それでも、HA制約により、完全なHCA（サーバごとのマルチHCAのケースにおける）または単一のHCAポートの障害は、所与のテナントについて予想される通信SLAをもはや保証しない再構成および共有を意味する場合がある。

#### 【0395】

ある実施形態に従うと、単一リンクについての全体的な帯域幅利用に対する制約に加えて、各テナントが異なる通信フロー間またはフロータイプ間でQOSを実現する能力は、それが、ファブリックレベルバッファリソースに対する厳しい輻輳競合、または他のテナントによる通信アクティビティに起因する調停を経験していないことに依存する。特に、これは、あるテナントが低レイテンシメッセージングを実現するために特定の「QOS ID」を用いている場合、そのテナントは、他のテナントがどのように「QOS ID」を用いているか、および／またはファブリック実現が「QOS ID」の使用をどのように実施しているか、および／または、これがバッファ割振りおよび／またはファブリック内のパケット帯域幅調停にどのようにマッピングするかによって、当該テナント自身を、別のテナントからの大容量データトラフィックと「競合している」、と見出すべきではないことを意味する。したがって、テナント通信SLAが、テナント内部QOS仮定が、同じファブリックリンクを共有する他のテナントが「うまく振る舞う」ことに依存せずには満たされ得ないことを意味する場合、これは、テナントが、他のテナントとのHCA（ま

10

20

30

40

50

たはH C Aポート) 共有なしにプロビジョニングされなければならないことを課す場合がある。

【0396】

ある実施形態に従うと、上述した基本帯域幅割り振りおよびQ O S問題の両方について、共有制約がファブリック内部リンクおよびサーバローカルH C Aポートリンクに適用される。したがって、所与のテナントに対する通信S L Aの性質および厳しさに応じて、テナントについてのV Mの展開は、物理サーバおよび／またはH C Aの共有ならびにファブリック内部I S Lの共有に制約を有し得る。I S L共有を回避するために、ルーティング制限と、プライベートファブリックトポロジ内でV Mが互いに対してプロビジョニングされ得る場所に対する制限との両方が適用され得る。

10

【0397】

トポロジー、ルーティングおよびブロッキングシナリオに関する考慮事項：

ある実施形態に従うと、上述のように、テナントが、他のテナントに属するV Mの動作に何ら依存することなく、通信するV Mのセットの間で期待される通信性能を達成できることを保証するために、他のテナントと共有されるH C A／H C Aポートまたは任意のファブリックI S Lは存在し得ない。したがって、提供される最高のS L Aクラスは、典型的には、これを、暗黙の実現として有するであろう。これは、原則として、クラウド内の多くの従来のシステムに対する現在のプロビジョニングモデルと同じスキームである。しかしながら、共有リーフスイッチでは、このS L Aは、他のテナントとのI S L共有がないことについての保証を必要とするであろう。また、テナントがフローと利用可能なファブリックリソースの利用との最良の可能な均衡を達成するためには、明示的な様式で「非ブロッキング性を最適化」できる必要があるだろう（すなわち、通信S Wインフラストラクチャは、異なるフローが同じリンク帯域幅に対して競合しない方法で通信が生ずることを保証する方法をテナントに与えなければならない）。これは、単一のリーフスイッチを介して生じ得る通信が実際にこの方法で実現されることを保証する方法を含むであろう。また、通信がI S Lを関与させなければならない場合、スループットを最大化するために、利用可能なI S Lにわたってトラフィックを均衡させることが可能であるべきである。

20

【0398】

ある実施形態に従うと、単一のH C Aポートからは、利用可能な最大帯域幅がファブリックにおけるすべてのリンクについて同じである限り、複数のI S Lにわたってトラフィックを均衡させようと試みることに意味はない。この観点から、利用可能なI S Lが送信側に対して非ブロッキングサブトポロジーを表す限り、送信H C Aポートごとに専用の「次ホップ」I S Lを用いることは意味をなすであろう。しかしながら、関連するI S Lが2つのリーフスイッチ間の接続性のみを表すのでなければ、送信側ポートごとに専用の次ホップI S Lを有するスキームは、実際には持続可能ではなく、なぜならば、ある時点で、通信が、異なるリーフスイッチに接続された複数の遠隔ピアH C Aポートとである場合、複数のI S Lが使用されなければならないからである。

30

【0399】

ある実施形態に従うと、非ブロッキングインフィニバンドファットツリートポロジでは、普及しているルーティングアルゴリズムは、「専用ダウン経路」を使用し、これは、非ブロッキングトポロジでは、ファットツリーの各層に同数のスイッチポートがあることを意味する。これは、各エンドポートが、1つのルートスイッチから、各中間スイッチ層を通して、関連するH C Aポートを接続するエグレスリーフスイッチポートまで、専用のポートチェーンを有することができることを意味する。したがって、単一のH C Aポートをターゲットとするすべてのトラフィックは、この専用ダウン経路を使用し、これらのリンク上において（下方向における）他の宛先ポートへのトラフィックはない。しかしながら、上方向では、各宛先への専用経路は存在し得ず、結果として、上方向におけるいくつかのリンクは、異なる宛先へのトラフィックによって共有されなければならないことになる。次のラウンドでは、これは、異なる宛先への異なるフローがすべて共有中間リンク上で全帯域幅を利用しようとするときに、輻輳につながり得る。同様に、複数の送信側が

40

50

同じ宛先に同時に送信している場合、これは、専用ダウン経路に輻輳を引き起こし得、これは、次いで、すぐに他の無関係のフローに広がり得る。

#### 【0400】

ある実施形態に従うと、単一の宛先ポートが単一のテナントに属する限り、専用ダウン経路における複数のテナント間の輻輳のリスクはない。しかしながら、専用ダウン経路を表すルートスイッチ（または中間スイッチ）に到達するために、異なるテナントが上方向において同じリンクを用いる必要があることは、依然として問題である。可能な限り多くの異なるルートスイッチを特定のテナントに専用化することによって、本システムおよび方法は、異なるテナントが上方向において経路を共有する必要性を低減させるであろう。しかしながら、単一のリーフスイッチからは、このスキームは、関連するルートスイッチに向かう利用可能なアップリンクの数を低減し得る。したがって、同じテナントに属するサーバ（またはむしろHCAポート）間で非ブロッキング二分帯域幅を維持するために、特定のリーフスイッチ上で（すなわち、単一のラック内で）単一のテナントに割り振られるサーバの数は、そのテナントによって使用されるルートスイッチに向かうアップリンクの数以下である必要があるであろう。他方では、単一のクロスバーを介して通信する能力を最大にするために、同じラック内の同じテナントに対してできるだけ多くのサーバを割り振ることは意味をなす。

#### 【0401】

一実施形態によれば、これは、元来、単一のリーフスイッチ内で保証された帯域幅を利用できることと、異なるラック内の通信ピアに向けて保証された帯域幅を利用できることとの間の矛盾を意味する。このジレンマに対処するために、最良の手法はおそらく、テナントVMは、それらがどのリーフスイッチ（すなわち、リーフスイッチペア）に直接接続されるかに基づいてグループ化され、次いで、そのようなグループ間の利用可能な帯域幅を定義する属性が存在する必要がある、スキームを用いることである。しかしながら、または、2つのそのようなグループの間（例えば、2つのラック内の同じテナントの間）の帯域幅を最大化できることと、複数の遠隔グループに対する帯域幅を保証できることとの間には、トレードオフがある。さらに、スイッチの2つの層（すなわち、単一のスパイン層によって相互接続されるリーフ層）のみの特別なケースでは、非ブロッキングトポロジは、N個のHCAポートが同じテナントに属するリーフスイッチとN個のスパインポートとの間にN個の専用アップリンクを有することが常に可能であることを意味する。したがって、これらのN個のスパインポートが、すべての関連する遠隔ピアポートについてすべての専用ダウン経路を「所有」するスパインを表す限り、構成は、そのテナントに対して非ブロッキングである。しかしながら、関連する遠隔ピアがN個を超えるスパインスイッチからの専用ダウン経路を表す場合、またはN個のアップリンクがすべての関連するスパインスイッチ間に分散されていない場合、本システムおよび方法は、他のテナントに対して回線争奪競合の可能性はある。

#### 【0402】

一実施形態によれば、単一のテナントのVMの間で、非ブロッキングまたはブロッキング接続性とは無関係に、同じリーフスイッチに接続された異なるソースからのフロー間の回線争奪の可能性が依然として存在する。—すなわち、宛先が同じスパインから専用ダウン経路を有し、ソースリーフスイッチからそのスパインへのアップリンクの数がそのような同時フローの数未満である場合、すべての送信側が全速リンク速度で動作する限り、アップリンク上で何らかの種類のブロッキング／輻輳を回避する方法はない。この場合、帯域幅を維持するための唯一の選択肢は、異なるスパインを介して宛先のうちの1つへの二次経路を用いることであろう。これは、次いで、別の専用ダウン経路との潜在的な競合を表すであろうと考えられ、なぜならば、標準的な非ブロッキングファットツリーはエンドポートごとに1つの専用ダウンリンクしか有し得ないからである。

#### 【0403】

ある実施形態に従うと、いくつかの従来のシステムの場合、リーフスイッチとスパインスイッチとの間に3のブロッキングファクタが存在し得る。したがって、ワークロードが

10

20

30

40

50

、通信トラフィックの3分の1以上がラック内部ではなくラック間にあることを意味する  
態様で分散されるマルチラックシナリオでは、結果として生じる二分帯域幅はブロッキング  
となる。例えば8ラックシステムにおけるノードの任意の対の間のトラフィックの均等  
な分散を伴う最も一般的なシナリオは、通信の7/8がラック間にあり、ブロッキング効  
果が実質的になることを意味する。

#### 【0404】

ある実施形態に従うと、オーバープロビジョニングのケーブルコストがシステムにおい  
て許容され得る場合（すなわち、固定されたスイッチユニットコストが与えられた場合）  
、追加のリンクを用いて、各リーフスイッチへの「バックアップ」ダウンリンクを提供す  
ること、および各リーフから各スパインへの予備のアップリンク容量を提供することの両  
方が可能である。—すなわち、両方の場合において、トラフィックの不均一な分布を表し  
、したがって、そもそも元来非ブロッキングであるトポロジーを利用することができない  
動的ワークロード分布に対する少なくともいくつかの潜在的な改善を提供する。

#### 【0405】

ある実施形態によると、より高い基数のフルクロスバースイッチは、各単一の「リーフ  
ドメイン」のサイズを増加させ、および所与のシステムサイズに必要とされるスパインス  
イッチの数を減少させる可能性も有する。例えば、128個のポートスイッチの場合、3  
2個のサーバを有する2つのフルラックを単一のフルクロスバリーフドメインに含める  
ことができ、それでも依然として非ブロッキングアップリンク接続性を提供し得る。同様  
に、16個のラック（512個のサーバ、1024個のHCAポート）間の非ブロッキング  
接続を提供するために、わずか8個のスパインが必要とされるであろう。したがって、  
依然としてわずか8つの、各リーフから各スパインへのリンクがある（すなわち、単一の  
、完全に接続されたネットワークの場合においてである）。1つのリーフ上のすべてのH  
CAポートが単一のスパインを介して単一の遠隔リーフに送信する極端なケースでは、こ  
れは依然として8のブロッキングファクタを意味する。他方、すべてのスパインの間で各  
リーフスイッチのための専用ダウン経路の均等な分布を考慮すると、そのような極端なシ  
ナリオの可能性は無視できるはずである。

#### 【0406】

一実施形態によれば、デュアル独立ネットワーク／レールの場合、冗長リーフスイッチ  
ペアにおける各リーフスイッチは、専用スパインを有する単一レールに属する。同じ8つ  
のスパインは、4つのスパインからなる2つのグループ（各レールに対して1つ）に分割  
され、したがって、この場合、レール内の各リーフは、4つのスパインのみに接続する必  
要があるだろう。したがって、この場合、最悪の場合のブロッキングファクタはわずか4  
であろう。他方、このシナリオでは、両方のレールにわたる負荷平衡を提供するために、  
各通信動作のためのレールの選択がさらに重要になる。

#### 【0407】

動的対静的パケットルート選択／転送＋マルチパス化：

ある実施形態に従うと、標準インフィニバンドは宛先アドレスごとに静的なルートを用  
いるが、イーサネットスイッチにおける動的ルート選択についてはいくつかの標準的で知  
的所有権下にあるスキームがある。インフィニバンドについては、「適応ルーティング」  
のための様々な知的所有権下にあるスキームもある（そのうちのいくつかは標準化され得  
る）。

#### 【0408】

ある実施形態に従うと、動的ルート選択の1つの利点は、ファブリック内において関連  
する二分帯域幅を最適に利用する確率が高くなり、それによって全体的なスループットも  
高くなることである。しかしながら、潜在的な欠点は、順序付けが妨害される可能性が  
あり、また、ファブリックの一領域における輻輳が他の領域に（すなわち、静的ルート選択  
が使用されていたなら回避され得たであろう態様で）より容易に広がる可能性があること  
である。

#### 【0409】

一実施形態によれば、「動的ルーティング」または「動的ルート選択」は、典型的には、スイッチ内およびスイッチ間で行われる転送判断に関して使用されるが、「マルチパス化」は、単一の宛先へのトラフィックが、送信側からの明示的なアドレス指定に基づいて、複数の経路にわたって拡散され得るときに使用される用語である。そのようなマルチパス化は、複数のローカルHCAポートにわたる単一のメッセージの送信を「ストライプ化」する（すなわち、完全なメッセージが、個々の転送動作を各々が表す複数のサブメッセージに分割される）ことを含み得、それは、同じ宛先への異なる転送が、ファブリックを通る異なる経路を動的に用いるためにセットアップされることを意味し得る。

#### 【0410】

ある実施形態に従うと、一般的な場合において、ローカルリーフドメイン外の宛先をターゲットとしているすべてのソースからのすべての転送が、（より）小さいチャンクに分割され、次いで、その宛先に向かうすべての可能な経路／ルートにわたって分散される場合、本システムは、利用可能な二分帯域幅の最適な利用を達成し、「リーフ間スループット」も最大化するであろう。それでも、これは、通信ワークロードもすべての可能な宛先にわたって均等に分散される限りのみ当てはまる。そうでない場合、その影響は、単一の宛先に向かういかなる輻輳も、すぐにすべての同時フローに影響を及ぼすことになる。

#### 【0411】

ある実施形態に従うと、輻輳が動的ルート選択およびマルチパス化に及ぼす意味は、単一の宛先へのトラフィックを制限して、単一の経路／ルートのみを、その経路／ルートが他のターゲットまたは任意の中間リンクにおける輻輳の犠牲とならない限り、用いるようにすることは意味をなす、ということである。専用ダウン経路を有する2層ファットツリートポロジでは、これは、エンドポートに関連しない唯一の考えられ得る輻輳が、同じスパインスイッチをターゲットとするアップリンク上に存在することになることを意味する。これは、特定のターゲットのために使用される個々のポートが動的に選択されるであろうことを除いて、同じスパインへのすべてのアップリンクを、同じ静的ルートを共有するポートのグループとして扱うことは意味をなすであろうことを意味する。代替として、個々のポートは、テナント関連付けに基づいて選択され得る。

#### 【0412】

ある実施形態に従うと、テナント関連付けを用いてそのようなグループ内でアップリンクポートを選択することは、固定された関連付けに基づき得るか、または異なるテナントがあるポートを用いるための「第1の優先度」を有するが他のポートを用いる能力も有するスキームに基づき得る。その場合、別のポートを用いる能力は、これが他のポートの「第1の優先度」トラフィックと競合しないことに依存するであろう。このようにして、テナントは、競合がない限り、すべての関連する二分帯域幅を用いることが可能であろうが、競合が存在する場合は、保証された最小帯域幅が存在するであろう。次いで、この最小保証帯域幅は、単一もしくはいくつかのリンクに対するすべての帯域幅、または1つ以上のリンクの帯域幅のパーセンテージを反映し得る。

#### 【0413】

ある実施形態によれば、原則として、同じ動的スキームが、下向きの経路において、スパインから特定のリーフへも使用され得る。一方では、これは、異なるエンドポートをターゲットとするフロー間でダウンリンクを共有することに起因する輻輳のリスクを増加させるであろうが、他方では、2つの異なるリーフスイッチに接続された2つのノードセット間で追加の代替経路を利用する方法を提供し得、しかし、依然として、異なるテナント間で輻輳が広がることを防止する方法を提供し得る。

#### 【0414】

ある実施形態に従うと、スパインからリーフへの異なる専用ダウン経路が既に特定のテナントを表すシナリオでは、これらのリンクが、関連するリーフスイッチ上において、別のスパインからの（一次）専用ダウン経路を有するエンドポートへの（同じテナントに属する）トラフィックのための「予備」として使用されることを可能にするスキームを有することは比較的単純であろう。

10

20

30

40

50

## 【0415】

ある実施形態に従うと、ある可能なモデルは、その場合、スイッチが、単一のスパインまたはリーフスイッチを接続する並列ISL間の動的ルート選択を処理することとなるであろうが、関連するターゲットへの（一次）専用ダウン経路を表さないスパインを介した明示的なマルチパス化を用いることについてのホストレベルの決定を有することとなるであろう。

## 【0416】

ペルテナントごとの帯域幅許可制御：

ある実施形態に従うと、単一のHCAが単一のテナントによってのみ使用される場合、本システムおよび方法は、HCAポートから生じ得る帯域幅を制限することができる。特に、このことは、遠隔リーフスイッチへ向かうトラフィックに関して、そのテナントについて、制限された二分帯域幅がある場合に当てはまる。

## 【0417】

ある実施形態に従うと、このような帯域幅制限の1つの局面は、制限された二分帯域幅の影響を受けるターゲットにのみ制限が適用されることを保証することである。原則として、これは、異なるのターゲットグループが特定の帯域幅割当（すなわち、厳密な最大レートおよび／またはある量の転送されたデータにわたる平均帯域幅のいずれか）に関連付けられるスキームを含むであろう。

## 【0418】

ある実施形態に従うと、このような制限は、定義上、HCAレベルで実現されなければならないであろう。また、そのような制限は、異なるテナントに属するVMが異なる仮想機能を介してHCAを共有している仮想化されたHCAシナリオに多かれ少なかれ直接マッピングするであろう。この場合、上で紹介された様々な「共有帯域幅割当グループ」は、1つ以上のVFのグループに関連付けられ、完全な物理HCAポートだけではない、という点で、追加の次元を必要とするであろう。

## 【0419】

ISL上のテナント単位帯域幅予約：

ある実施形態に従うと、上記のように、テナント（またはテナントのグループ）のために1つ以上のISLにわたってなんらかの保証された帯域幅を予約することは意味をなし得る。あるシナリオでは、どのテナントが完全なリンクを本当に使用することを許可されるかを制限することによって、そのリンクをテナントのために予約することができる。しかしながら、より柔軟でより細かい粒度のスキームを有するために、代替の手法は、スイッチ調停メカニズムを用いて、（いくつかの）イングレスポートが、1つ以上のエグレスポート上でどのような他のイングレスポートが帯域幅について競合しているかとは無関係に、同エグレスポートの帯域幅の最大X%を用いることを許可されることを保証することである。

## 【0420】

一実施形態によれば、このようにして、すべてのイングレスポートが、関連するエグレスポートの帯域幅の100%までを用いることができるが、ただし、これが、優先されるイングレスポートからのいかなるトラフィックとも競合しない限りにおいてのみである。

## 【0421】

ある実施形態に従うと、異なるテナントが異なるイングレスポート（たとえばHCAポートを接続するリーフスイッチポート）を「所有」するシナリオにおいて、このスキームは、1つ以上のスパインスイッチへのアップリンク帯域幅の割振りのための柔軟かつきめ細かいスキームを容易にするであろう。

## 【0422】

ある実施形態に従うと、スパインからリーフスイッチまでのダウンリンク経路において、そのようなスキームの有用性は、厳密な専用ダウン経路を有するスキームがどの程度使用されるかまたはされないかに依存するであろう。厳密な専用ダウン経路が使用され、ターゲットエンドポートが単一のテナントを表す場合、デフォルトでは、ダウンリンクを使

10

20

30

40

50

用しようとしている異なるテナント間に潜在的な競合はない。したがって、この場合、関連するダウンリンクへのアクセスは、通常、すべての関連するイングレスポートについて等しいアクセスを有するラウンドロビン調停スキームを用いるようにセットアップされるべきである。

#### 【0423】

ある実施形態に従うと、イングレスポートは異なるテナントに属するトラフィックを表すことができるので、1つのテナントに属するパケットが送信され得、関連するテナントが送信を許可されていないエグレスポート上で帯域幅を消費し得ることは、決して問題ではないはずである。この場合、仮定は、そのようなパケットが帯域幅を浪費することを防止するために、調停ポリシーではなく、厳密なアクセス制御（例えば、様々なポートについてのVLANベースの制限）が採用されることである。

10

#### 【0424】

ある実施形態に従うと、リーフスイッチでは、スパインからのダウンポートには、他のローカルエンドポートと比較して、さまざまなエンドポートに向かって、より多くの帯域幅が与えられるかもしれず、なぜならば、ダウンリンクが原則的には複数の送信側HCAポートを表すことができるのに対し、ローカルエンドポートは単一のHCAポートを表すのみであるからである。これが当てはまらない場合、いくつかの遠隔サーバが、ターゲットリーフスイッチへの単一のダウン経路を共有しているが、次のラウンドでは、リーフスイッチに直接接続されたN-1 HCAポートも同じローカルターゲットポートに送信しようとする場合に、そのリーフスイッチ上で単一の宛先に向かって帯域幅の1/Nを共有していることになるシナリオを有することが可能である。

20

#### 【0425】

ある実施形態に従うと、仮想化されたHCAが異なるテナントを表す場合、ファブリックISL内で（すなわち、さまざまなISLにわたって）帯域幅を予約する問題は、著しくより複雑になり得る。イングレス／アップリンク経路については、ある単純化された手法は、異なるテナント間で帯域幅調停を提供することはHCA次第であり、次いで、HCAポート上で何が送出されようと、それは、ポートレベル調停ポリシーに従ってイングレスリーフスイッチによって処理されることになる、ということである。したがって、この場合、リーフスイッチの観点からの変化はない。

#### 【0426】

一実施形態によれば、ダウンリンク経路（スパインからリーフへ、およびリーフイングレスからエンドポートへ）においては、調停判断は、パケットを転送しようとするポートだけでなく、様々な保留中のパケットがどのテナントに属するかにも依存し得るので、状況は異なる。1つの可能な解決策は、（またも）いくつかのISLを、特定のテナント（またはテナントのグループ）を表すようにのみ制限し、次いで、これをポートレベル調停スキームに反映することである。代替的に（または追加的に）、以下に概説されるように、異なるテナントを表すために異なる優先度またはQOS IDを用いることができる。最後に、調停論理の一部として使用されるVLAN IDまたはパーティションIDのような「テナントID」または任意の関連するアクセス制御ヘッダフィールドを有することは、調停に必要なレベルの粒度を容易にするであろう。しかしながら、これは、スイッチにおける、既に著しい「時間および空間」の複雑さを有する調停論理の複雑さを著しく増大させる可能性がある。また、そのようなスキームは、エンドツーエンドワイヤプロトコルにおいてすでに役割を有し得る情報の過負荷を伴うため、そのような余分な複雑性が、そのようなヘッダフィールド値に関する任意の既存の使用または仮定と競合しないことが重要である。

30

40

#### 【0427】

ISLおよびエンドポートリンクにわたる、異なる優先順位、フロータイプおよびQOS ID／クラス：

ある実施形態に従うと、異なるフロータイプが同じリンク上で同時に進行するためには、それらがスイッチおよびHCAにおいて同じパケットバッファに対して競合しないこと

50

が重要である。また、異なるフロータイプ間の相対的な優先度を区別するために、様々なスイッチエグレスポート上で次にどのパケットを送信するかを決定する調停論理は、どのようなパケットタイプキューがどのエグレスポート上で送出するものを有するかを考慮に入れなければならない。調停の結果は、すべてのアクティブなフローが、それらの相対的な優先度に従って、および、関連する下流ポートに関する関連するフロータイプに関するフロー制御条件（もしあれば）が、現在、任意のパケットを送信することをどの程度可能にするかに従って、順方向進行をなしていることであるべきである。

#### 【0428】

ある実施形態に従うと、原則として、たとえ異なるテナントが同じリンクを用いている場合であっても、異なるQOS IDを用いて、異なるテナントからのトラフィックフローを互いから独立させることができる。しかしながら、各ポートのためにサポートされ得るパケットキューおよび独立バッファプールの数は、典型的には、10未満に制限されるため、この手法のスケラビリティは、非常に制限される。また、単一のテナントが、異なるフロータイプを互いから独立させるために、異なるQOS IDを使用したい場合、スケラビリティはさらに低減される。

10

#### 【0429】

ある実施形態に従うと、上述したように、スイッチの単一ペア間で複数のISLを論理的に組み合わせることによって、本システムおよび方法は、いくつかのリンクをいくつかのテナントに制限し、次いで、異なるテナントが異なるQOS IDを異なるISL上で互いから独立して使用できることを保証することができる。しかしながら、ここでも、これは、他のテナントの独立性が100%保証される場合、任意の単一のテナントに利用可能な総帯域幅に制限を課す。

20

#### 【0430】

ある実施形態に従うと、理想的なケースでは、HCAイングレス（受信）パケット処理は、着信パケットがどのトランスポートレベル動作を表すかに関係なく、常に、関連するリンク速度よりも高いレートで生ずることができる。これは、その最後のリンク、（すなわち）HCAポートに接続するリーフスイッチ上のエグレスポート上で異なるフロータイプを制御するフローの必要がないことを意味する。しかしながら、リーフスイッチにおける異なるキューからの異なるパケットのスケジューリングは、依然として、優先度、公平性、および順方向進行に関する関連するポリシーを反映しなければならない。—例えば、1つの小さい高優先度パケットが、あるエンドポートをターゲットとすると同時に、N個のポートも、最大MTUサイズの「バルク転送パケット」を同じターゲットポートに送信しようと試みる場合、高優先度パケットは、他のポートのいずれよりも前にスケジューリングされるべきである。

30

#### 【0431】

ある実施形態に従うと、エグレス経路において、送信側HCAは、多くの異なる方法でパケットをスケジューリングおよびラベル付けすることができる。特に、VM+仮想HCAと物理ファブリックとの間の「ワイヤ内のポンプ」としてのオーバーレイプロトコルの使用は、テナント仮想HCAインスタンス間のエンドツーエンドプロトコルのいかなる局面も混乱させることなく、スイッチが関係し得るファブリック固有情報の符号化を可能にするであろう。

40

#### 【0432】

ある実施形態に従うと、スイッチは、現在のワイヤプロトコルが想定するよりも多くのバッファリングおよび内部キューイングを提供することができる。このようにして、異なるSLAを有する複数のテナントを表すトラフィックによってリンクが共有されることを考慮し、異なるフロータイプに対して異なるQOSクラスを用いる、バッファリング、キューイング、および調停ポリシーをセットアップすることが可能であろう。

#### 【0433】

ある実施形態に従うと、このようにして、異なる高優先度テナントは、スイッチ内に、より多くのプライベートパケットバッファ容量も有するかもしれない。

50

## 【0434】

無損失パケット転送対有損失パケット転送：

ある実施形態に従うと、高性能RDMAトラフィックは、個々のパケットがスイッチにおけるバッファ容量の欠如に起因して失われないことに大きく依存し、また、パケットが個々のRDMA接続ごとに正しい順序で到着することにも大きく依存する。原則として、潜在的な帯域幅が高いほど、これらの局面は、最適な性能を達成するのに、より重要である。

## 【0435】

ある実施形態に従うと、無損失動作は、明示的なフロー制御を必要とし、非常に高い帯域幅は、バッファ容量と、MTUサイズと、フロー制御更新頻度との間のトレードオフを意味する。

10

## 【0436】

実施形態によれば、無損失動作の欠点は、生成される総帯域幅が下流／受信容量よりも高い場合に、それが輻輳につながることである。輻輳は、次いで（おそらく）広がり、ファブリック内のどこかで同じバッファについて競合するすべてのフローを減速させることに終わることになる。

## 【0437】

ある実施形態に従うと、上述したように、独立したバッファプールに基づいてフロー分離を提供する能力は、スイッチ実現について、ポートの数、異なるQOSクラスの数、および（上記で紹介したように）潜在的には、異なるテナントの数にも依存する、主要なスケーラビリティの問題である。

20

## 【0438】

ある実施形態に従うと、代替手法は、真に無損失の動作（すなわち、保証されたバッファ容量に基づく無損失）を「プレミアムSLA」属性とし、それにより、この特徴を、そのようなプレミアムSLAを購入したテナントのみに制限することであり得る。

## 【0439】

ある実施形態に従うと、ここでの重要な問題は、利用可能なバッファ容量を「超過予約」できることであり、同じバッファを有損失フローおよび無損失フローの両方に用いることができるが、有損失フローに割り振られたバッファは、無損失フローからのパケットが到着し、同じプールからのバッファを用いる必要があるときはいつでも、強制排除されることができる。有損失フローが順方向に進むことを可能にするために、非常に最小限のセットのバッファが提供され得るが、それは、最適なバッファ割振りで達成され得るものよりも（はるかに）より低い帯域幅においてである。

30

## 【0440】

ある実施形態に従うと、（必要とされるときに）バッファを強制排除し、よりプレミアムなSLAタイプのフロークラスに与えなければならない前にバッファを占有できる最大時間の差に関して、異なるクラスのハイブリッド無損失／有損失フロークラスを導入することも可能である。—これは、リンクレベルクレジットを有するファブリック実現の文脈で最も良く機能するであろうが、潜在的に、xon／xoff型フロー制御（すなわち、RoCE／RDMAに使用されるイーサネット一時停止ベースのフロー制御スキーム）で機能するように適合させることもあり得る。

40

## 【0441】

厳密なパケット順序付け対（より）緩和されたパケット順序付け：

ある実施形態に従うと、ファブリック内での厳密な順序付けおよび無損失パケット転送により、HCA実現例は、信頼性のある接続およびRDMAを、トランスポートレベルにおいて最小状態オーバーヘッドで実現し得る。しかしながら、（ファブリック内の適応的／動的転送判断による）ルートの偶発的な変化による何らかの量の順序外パケット配信をより良好に許容するために、ならびにファブリック内の有損失または「ハイブリッド無損失／有損失」モード転送により失われたパケットに関連するオーバーヘッドおよび遅延を最小限に抑えるためにも、効率的なトランスポート実現形態は、多数の個々のパケット（

50

シーケンス番号) が順序を外れて到着しているとともに、より後のシーケンス番号を有する他のパケットが受け入れられ確認応答されている間に個々に再試行されることを可能にするために、十分な状態を保つことを必要とするだろう。

#### 【0442】

ある実施形態に従うと、ここでの重要なポイントは、紛失パケットまたは順序外パケットが現在のデフォルトのトランスポート実現で再試行を引き起こす場合に、平均帯域幅の長い遅延および損失を回避することである。また、配達されるパケットの列における後続のパケットが脱落することを回避することによって、本システムおよび方法は、そうでなければ他のフローによって消費され得た多くの帯域幅を消費し得たファブリック帯域幅の浪費も著しく低減している。

10

#### 【0443】

共有サービスおよび共有HCA:

ある実施形態に従うと、複数のテナントによって使用されるファブリック上の共有サービス(たとえばバックアップデバイス)は、そのサービスが特定のテナント(または制限されたテナントのグループ)に専用であり得るエンドポートを提供できなければ、いくつかのエンドポートリンクが異なるテナントによって共有されることになることを意味する。複数のテナントに属するVMが同じサーバおよび同じHCAポートを共有している場合、同様のシナリオが存在する。

#### 【0444】

ある実施形態に従うと、異なるテナントに対して、微調整されたサーバおよびHCAリソースを割り振ることができ、また、HCAからの発信データトラフィック帯域幅が、関連するSLAレベルに従って異なるテナント間で公平に分割されることを保証することもできる。

20

#### 【0445】

ある実施形態に従うと、ファブリック内において、異なるテナントに属するデータトラフィック間の相対的な重要性およびそれによる公平性を反映するパケットバッファ割り振りおよびキューイング優先度および調停ポリシーを設定することも可能であり得る。しかしながら、ファブリック内の非常に微調整されたバッファ割り振りおよび調停ポリシーであっても、粒度は、異なるテナントについての相対的な優先度および帯域幅割当が共有HCAポートに対するインGRESS帯域幅に関して正確に反映されることを保証するほど十分に精細ではないかもしれない。

30

#### 【0446】

一実施形態によれば、そのような精細な粒度の帯域幅割り振りを達成するために、1つ以上のテナントに属するいくつかの遠隔通信ピアの間で利用可能なインGRESS帯域幅を効果的に分割およびスケジューリングすることができる動的エンドツーエンドフロー制御スキームが必要とされている。

#### 【0447】

ある実施形態に従うと、そのようなスキームの目標は、任意の時点において、関連するアクティブな遠隔クライアントのセットが、利用可能なインGRESS帯域幅の、その公平な(必ずしも等しいとは限らない)シェアを利用できることであろう。また、この帯域幅利用は、エンドポートにおいて過剰な帯域幅を用いる試みに起因してファブリックに輻輳を生じさせることなく行われるべきである。(それでも、ファブリックレベルの輻輳は、依然として、ファブリックの残りの部分内の共有リンク上の過負荷に起因して生じ得る。)

40

ある実施形態に従うと、この目標を達成するためのハイレベルモデルは、受信側が関連する遠隔クライアントのセットのために利用可能な帯域幅を動的に割り振って更新することができる、ということであろう。各遠隔クライアントの現在の帯域幅値は、各クライアントに現在提供されているもの、および次に必要とされるものに基づいて計算される必要があるであろう。

#### 【0448】

ある実施形態に従うと、これは、単一のクライアントが現在すべての利用可能な帯域幅

50

を用いることを許可されており、別のクライアントもイングレス帯域幅を用いる必要がある場合、現在アクティブなクライアントに対して、新たな低減された最大帯域幅について伝える更新命令が配信されなければならない、新たなクライアントは、現在のクライアントに対する低減に対応する最大帯域幅を用いることができるという命令を配信されなければならないことを意味する。

#### 【0449】

一実施形態によれば、原則として、同じスキームが「任意の」数の同時クライアントに適用されるであろう。しかしながら、当然ながら、利用可能な帯域幅がいかなる時点においても決して「超過予約」されないことを保証できることと、利用可能な帯域幅が必要とされるときに常に完全に利用されることを保証することとの間には、巨大なトレードオフがある。

#### 【0450】

ある実施形態に従うと、この種のスキームによるさらなる課題は、それが動的輻輳制御とうまく相互運用することを保証し、複数のターゲットについて共有経路に係る輻輳が各送信側内で協調される態様で扱われることも保証することである。

#### 【0451】

##### 高可用性およびフェイルオーバー：

ある実施形態に従うと、性能に加えて、プライベートファブリックの鍵となる属性は、任意のクライアントアプリケーションについてサービスの喪失なく、任意の単一の障害点に続いて通信をフェイルオーバーする冗長性および能力であってもよい。さらに、「サービスの喪失」は二値条件（すなわち、サービスが存在するかまたは失われる）を表すが、いくつかの同等に重要であるが、より多くのスカラー属性は、フェイルオーバー中にどの程度まで電力不足時間があり、その場合、それがどれほど長い、である。別の重要な局面は、フェイルオーバー動作の完了中および完了後に、期待される性能がどの程度まで提供される（または再確立される）かである。

#### 【0452】

ある実施形態に従うと、単一のノード（サーバ）の観点からは、目標は、サーバ自体の外部のファブリック通信インフラストラクチャにおけるどのような単一の障害点（すなわち、単一のローカルHCAを含む）も、ノードが通信できなくなることを意味するべきではない、ということである。しかしながら、完全なファブリックの観点からは、1つ以上の構成要素の損失がどのようなレベルのファブリック全体のスループットおよび性能影響を意味するかについても疑問がある。例えば、2つのスパインスイッチのみで動作することができるトポロジサイズの場合。—その場合、二分帯域幅および輻輳のリスクの増加に関して、スパインのうちの1つがサービス停止している場合、リーフとリーフとの通信容量の50%が失われることは許容可能か？

ある実施形態に従うと、テナント毎SLAに対する別の問題は、プレミアムSLAを有するテナントが、障害およびその後のフェイルオーバー動作に続いて、残りの利用可能なリソースの、比例してより大きいシェアを得るという点において、そのようなテナントのためにファブリックリソースを予約および／または優先する能力がどの程度反映されるべきかということであり、—すなわち、かくして、障害の影響は、プレミアムSLAテナントについてはより少なくなるが、他のテナントについては、より多くの影響を犠牲にすることになる。

#### 【0453】

ある実施形態に従うと、冗長性に関して、そのようなテナントのための初期リソースプロビジョニングが、どのような単一の障害点も、関連する性能/QoS SLAが障害中または障害後のいずれにおいても満たされ得ないことを意味しないであろうことを確実にするであろうことはまた、「スーパープレミアムSLA属性」でもあり得る。しかしながら、そのようなオーバープロビジョニングに関する基本的な問題は、利用可能なリソースが常に最も最適な方法で活用されること、および、いかなる単一の障害点の結果としても非常にわずかな期間を超えて通信が決して停止されないことを確実にするために、極めて

10

20

30

40

50

高速のフェイルオーバー（およびフェイルバック／再均衡）が存在しなければならないことである。

【0454】

ある実施形態に従うと、そのような「スーパープレミアム」セットアップの例は、デュアルHCAベースのサーバを有するシステムであり得、両方のHCAはアクティブ-アクティブ方式で動作しており、両方のHCAポートはまた、代替経路が試される前の遅延が非常に短いAPM（自動経路マイグレーション）スキームを用いるアクティブ-アクティブ方式でも利用される。

【0455】

経路選択：

ある実施形態に従うと、2つのエンドポイント間に複数の可能な経路が存在する場合、関連するRDMA接続のための最良または「正しい」経路の選択は、理想的には、通信ワークロードが、関連付けられたSLAの制約内で最良の可能な性能を経験するように、かつシステムレベルファブリックリソースが最も最適な方法で活用されるように、自動的にあるべきである。

【0456】

ある実施形態に従うと、理想的な場合、これは、VM内のアプリケーション論理が、どのローカルHCAおよびどのローカルHCAポートがどの通信のために使用され得るかまたは使用されるべきかを扱う必要がないことを意味するであろう。これはまた、ポートレベルアドレス指定スキームではなくノードレベルを意味し、基礎をなすファブリックインフラストラクチャがアプリケーションに対して透過的に使用されることを意味する。

【0457】

ある実施形態に従うと、このようにして、関連するワークロードは、異なるシステムタイプまたはシステム構成の明示的な取り扱いを必要とすることなく、異なるインフラストラクチャ上でより容易に展開され得る。

【0458】

特徴：

ある実施形態に従うと、このカテゴリの特徴は、現行のファームウェアおよびソフトウェアを用いる既存のHCAおよび／またはスイッチハードウェアによってサポートされると想定される。

【0459】

ある実施形態に従うと、このカテゴリにおける主な目標は以下のとおりである：

- ・ローカル物理HCAインスタンスごとに単一のVMまたはテナントに属するVFによって生成される総エグレス帯域幅を制限する能力。
- ・ローカル物理HCAの単一のVMまたはテナントに属するVFが、利用可能なローカルリンク帯域幅の少なくとも最小パーセンテージを利用できることになることを保証する能力。

- ・VFがどのネットワーク（E n e t）優先度を使用できるかを制限する能力。

○これは、単一のVMが複数の優先度を用いるために（すなわち、優先度制限が有効にされると、単一のVFは単一の優先度を用いることしか許可され得ない限り）、複数のVFが割り振られなければならないことを意味し得る。

- ・どのISLが単一のテナントまたはテナントのグループに属するフローのグループによって使用され得るかを制限する能力。

【0460】

ある実施形態に従うと、物理HCAを他のテナントと共有しているテナントによるHCA使用を制御するために、「HCAリソース制限グループ」（本明細書では「HRLG」と呼ばれる）がそのテナントのために確立されることになる。HRLGは、HRLGによって生成することができる実際のデータレートを定義する最大帯域幅で設定することができ、また、他のテナント／HRLGとの回線争奪があるときにHRLGがHCA帯域幅の少なくとも指定されたパーセンテージを達成することを確実にする最小帯域幅シェアで設

10

20

30

40

50

定することもできる。他のHRLGとの回線争奪がない限り、HRLG内のVFは、指定されたレート（またはレート制限が定義されていない場合はリンク容量）まで永続的に用いることができる。

#### 【0461】

ある実施形態に従うと、HRLGは、HCAインスタンスがサポートし得る数までのVFを含み得る。HRLG内では、各VFが、HRLGに割り当てられた「割当」の公平なシェアを得るであろうことが予想される。各VFについて、関連するQPはまた、ローカルリンクへのアクセスの公平なシェアを、利用可能なHRLG割当の関数として、およびQRに対する任意の現在のフロー制御制限を得ることになる（すなわち、QPが、それ自体をスロットリングするようにそれに命ずる輻輳制御フィードバックを受信した場合、または関連する優先度で送信する「クレジット」が現在ない場合、QPは、ローカルリンクアクセスについて考慮されないことになる。）

10

ある実施形態に従うと、HRLG内では、VFが使うことができる優先度に対する制限を実施することが考えられる。この制限が、VFに対して許可される単一の優先度に関してしか定義され得ない限り、その意味は、（依然としていくつかの優先度のみに制限されているが）複数の優先度を用いると想定されるVMが、複数のVFを一要求される優先度ごとに1つ一使用しなければならないことになる、ということである。（注：複数のVFの使用は、異なる優先度を用いる複数のQP間のローカルメモリリソースの共有が、問題を表しそうであることを意味し、なぜならば、それは、どの優先度制限／実施ポリシーが定義されるかに応じて、異なるVFがVM内のULP／アプリケーションによって割り振られ、使用されなければならないことを意味するからである。）

20

ある実施形態に従うと、単一のHRLG内では、VF／QPが現在どの優先度を用いているかに依存して帯域幅割振りに差異はない。—それらはすべて、関連する割当を公平／均等な態様で共有する。したがって、異なる帯域幅割当を異なる優先度と関連付けるために、関連するHRLGが表す共有割当と関連付けられるべき優先度を用いるよう制限されるVFのみを含むことになる1つ以上の専用HRLGを定義することが必要とされる。このようにして、VMまたは同じ物理HCAを共有する複数のVMを有するテナントは、異なる優先度に対して異なる帯域幅割当を与えられ得る。

#### 【0462】

ある実施形態に従うと、現在のハードウェア優先度制限は、不正な優先度で送信されるのを試みられるデータが外部リンクに送信されるのを防止するが、ローカルメモリからの関連するデータのフェッチは防止しない。したがって、HCAがエグレス方向に維持することができるローカルメモリ帯域幅が利用可能な外部リンク帯域幅とほぼ同じである場合、依然として無駄になる全体的なHCAリンク帯域幅がある。しかしながら、関連するメモリ帯域幅が外部リンク帯域幅よりも（有意に）大きい場合、違法な優先度を用いる試みは、HCAパイプラインが最適な効率で動作する限り、外部リンク帯域幅をより少なく浪費することになる。それでも、外部リンク帯域幅に関してあまり節約がない限り、違法な優先度の使用を防止するための可能な代替スキームは、スイッチングレスポートにおいてACLルール実施を活用することであり得る。関連するテナントがいかなるスプーフィングの可能性も伴わずに効果的に識別され得る場合、これを用いて、同じVMに対して優先度ごとに個々のVFを割り振る必要なく、テナント／優先度関連付けを実施し得る。しかしながら、パケット／テナント関連付けが常に明確に定義され、送信VMからスプーフィングすることができないことを保証することと、およびVFがVM／テナントによって使用されるようにセットアップされているときはいつでも、関連するスイッチポートを動的に更新して関連する実施を実行すること両方は、トリビアルではない複雑さを表す。1つの可能なスキームは、VM／テナントに関連付けることができるスプーフィング不可なIDを表すためにVFポートごとのMACを用いることであろう。しかしながら、VxLANまたは他のオーバーレイプロトコルが使用されている場合、これは、—特に外部スイッチが使用されているオーバーレイスキームに関与している（はまたはそれを認識している）と想定されない限り—簡単ではない。

30

40

50

## 【0463】

ある実施形態に従うと、どのフローがどのISLを用いることができるかを制限するために、スイッチ転送論理は、関連するフローを識別し、それに応じて転送を設定するためのポリシーを有する必要がある。一例は、フローグループを表すためにVLAN IDを用いることである。異なるテナントがファブリック上の異なるVLAN IDにマッピングする場合、1つの可能なスキームは、どのVLAN IDが任意のLAGまたは他のポートグループ化における様々なポートに対して許可されるかに基づいて、スイッチがLAGタイプ均衡を動的に実現し得ることであろう。別のオプションは、宛先アドレスとVLAN IDとの組み合わせに基づくパケットの明示的な転送を含むであろう。

## 【0464】

ある実施形態に従うと、VxLANベースのオーバーレイが物理スイッチファブリックに対して透過的に用いられる場合、上記で概説したようにスイッチがVLAN IDをISLにマッピングすることを可能にするために、異なるオーバーレイを異なるVLAN IDにマッピングすることが可能であろう。

## 【0465】

ある実施形態に従うと、別の可能なスキームは、個々のエンドポイントアドレスの転送が、VLANメンバーシップまたは「テナント」関連付けの何らかの他の概念を考慮に入れるルーティングスキームに従ってセットアップされることである。しかしながら、同じエンドポイントアドレス値が異なるVLANにおいて許容される限り、VLAN IDは転送判断の一部である必要がある。

## 【0466】

ある実施形態に従うと、共有ISLまたは排他的ISLのいずれかへのテナントあたりのフローの配信は、ファブリック（ファットツリー）トポロジー内でグローバルに最適化された方法でトラフィックを配信するために、ホリスティックなルーティングスキームを必要とし得る。そのようなスキームの実現は、典型的には、スイッチのためのSDNタイプ管理インターフェースに依存するであろうが、ホリスティックなルーティングの実現は、トリビアルではないであろう。

## 【0467】

短期および中期SLAクラス：

ある実施形態に従うと、以下は、非ブロッキング2層ファットツリートポロジーが、単一のリーフスイッチの基数を超えるシステムサイズ（物理ノードカウント）のために使用されていると仮定する。また、物理サーバ上の単一のVMが（1つ以上のvHCA/VFを介して）すべてのファブリック帯域幅を用いることができると仮定される。したがって、物理サーバあたりのテナントあたりのVMの数は、HCA/ファブリックの観点からテナントレベルSLAファクタと見なされる必要があるパラメータではない。

## 【0468】

ある実施形態に従うと、一番上の層（たとえばプレミアムプラス）は、

- ・専用サーバを用いることができる。
- ・VM（またはHAポリシー）の数およびサイズが追加の距離を意味する場合を除いて、可能な限り同じリーフドメインに割り振られることができる。
- ・テナントが複数のリーフドメインを用いている場合（すなわち、同じリーフドメイン内でこのテナントに割り振られるサーバの数に対して）、平均して、ローカルリーフからの非ブロッキングアップリンク帯域幅を有することができるが、専用アップリンクまたはアップリンク帯域幅は有さないことになる。
- ・すべての「フローグループ」（すなわち、ファブリック内の異なるバッファプールおよび調停グループを表す優先度など）を用いることができる。

## 【0469】

ある実施形態に従うと、下位の層（たとえばプレミアム）は、

- ・専用サーバを用いることができるが、同じリーフドメインについての保証はない。
- ・平均して、少なくとも50%の非ブロッキングアップリンク帯域幅（すなわち、同じリ

10

20

30

40

50

ーフドメイン内でこのテナントに割り振られるサーバの数に対して)を有することができるが、専用アップリンクまたはアップリンク帯域幅は有さないことになる。

- ・すべての「フローグループ」を用いることができる。

#### 【0470】

ある実施形態に従うと、第3の層(たとえばエコノミープラス)は、

- ・共有サーバを用いてもよいが、専用の「フローグループ」を有することになる。

○これらのリソースは、ローカルHCAおよびスイッチポート専用であることになるが、ファブリック内で共有されることになる。

・ローカルサーバからのすべての利用可能なエグレス帯域幅を用いる能力を有することができるが、全エグレス帯域幅の少なくとも50%を有することが保証される。

- ・物理サーバ当たり1つのエコノミープラステナントのみ。

・平均して、このエコノミープラステナントによって使用されているサーバの数に対して少なくとも25%の非ブロッキングリーフアップリンク帯域幅を有することができる。

#### 【0471】

一実施形態によれば、第4の層(例えば、エコノミー)は、

- ・共有サーバを用いることができる

- ・専用優先順位なし

・サーバエグレス帯域幅の50%までを使用できるようにされることができ、これを3つまでの他のエコノミーテナントと共有する場合がある

・平均して、利用可能なリーフアップリンク帯域幅の25%までを同じリーフドメイン内の他のエコノミーテナントと共有することができる。

#### 【0472】

ある実施形態に従うと、最下層(たとえばスタンバイ)は、

- ・保証された帯域幅を伴わない予備容量を使用し得る

より長期の特徴:

ある実施形態に従うと、このセクションで議論される主要な特徴は以下の通りである:

・単一のVFが、サポートされる優先度のセット全体のうちの任意のサブセットを用いるように制限されることを可能にする方法で、VFごとに優先度制限を実施する能力。

・開始VFに対して許可されない優先度を用いてデータ転送が試みられるときはいつでも、浪費されるメモリ帯域幅またはリンク帯域幅をゼロにする。

・様々なHRLG内の各VFが、関連するHRLG総最小帯域幅および／または最大レートの、それらの公平なシェアを、ただし、様々な関連付けられる優先度ごとの割当によって定義される制約の下で得るように、複数のHRLGにわたる異なる個々の優先度に対するエグレスレートを制限する能力。

・ターゲットごとおよび共有経路／ルートごとの両方に基づいて送信側帯域幅制御および輻輳調整を実行し、これがVM／VF(すなわちvHCAポート)レベルおよびHCAポートレベルの両方で集約される能力。

・協働する遠隔送信側の受信側スロットリングに基づいて、VFに対して総平均送信およびRDMA書込イングレス帯域幅を制限する能力。

・協働する遠隔RDMA読出応答側に依存することなく、VFに対して平均RDMA読出イングレス帯域幅を制限する能力。

・協働する遠隔送信側の受信側スロットリングに基づいてVFに対して総平均イングレス帯域幅を制限するとき、送信およびRDMA書込に加えてRDMA読出を含む能力。

・テナントVMが異なるピアVMのグループに対して利用可能な二分帯域幅を観察する能力。

- ・ファブリック内のルーティング制御および調停ポリシーのためのSDN特徴。

#### 【0473】

ある実施形態に従うと、HCA VFコンテキストは、正当な優先度のリスト(IBTA IB vPortのための正当なSLのセットに類似している)を含むように拡張することができる。作業要求がVFにとって正当でない優先度を使用しようとしているときはい

10

20

30

40

50

つでも、その作業要求は、任意のローカルまたは遠隔データ転送が開始される前に失敗するべきである。加えて、任意の優先度が使用され得るという錯覚をアプリケーションに与えるために、優先度マッピングも使用され得る。しかしながら、パケットが送信される前に複数の優先度が同じ値にマッピングされ得るこの種のマッピングは、異なるフロータイプを異なる「QOSクラス」に関連付けるという点で、アプリケーションはそれ自体のQOSポリシーをもはや制御することはできない可能性がある、という欠点を有する。そのような制限されたマッピングは、SLA属性を表す（すなわち、より特権的なSLAは、マッピング後、より多くの実際の優先度を意味する）。しかしながら、アプリケーションが、ファブリックにおいて独立したフローを表すことにもなる態様で、どのフロータイプをどのQOSクラス（優先度）に関連付けるかを決定できることは、常に重要である。

10

【0474】

ターゲットグループおよび動的BW割当更新を介したインGRESS RDMA読出BW割当の実現：

ある実施形態に従うと、「プロデューサ／送信元」ノードからのフローに対するターゲットグループ関連付けが、UD送信、RDMA書込、RDMA送信およびRDMA読出（すなわち、データを有するRDMA読出応答）を含むすべての発信データパケットの帯域幅規制を意味する限り、vHCAポートに対するすべてのインGRESS帯域幅の完全な制御がある。これは、ターゲットvHCAポートを所有するVMが複数のピアノードに対して「過剰な」量のRDMA読出要求を生成しているかどうかとは無関係である。

【0475】

20

ある実施形態に従うと、上で論じたように、ターゲットグループをフロー固有および「非請求」の両方のBECONシグナリングに結合することは、vHCAポートごとのインGRESS帯域幅を、任意の数の遠隔ピアに対して動的にスロットリングすることができることを意味する。

【0476】

ある実施形態に従うと、上で概説した「非請求BECON」メッセージを用いて、異なるステージ番号に対する純粋なCEフラグ立て／フラグ解除に加えて、特定のレート値を通信することもできる。このようにして、新たなピアからの初期着信パケット（例えば、CMパケット）が、着信パケットが来たHCA（すなわち、関連するファームウェア／ハイパー特権ソフトウェア）および現在の通信ピアの両方への1つ以上の「非請求BECON」メッセージの生成をトリガすることができるスキームを有することが可能である。

30

【0477】

ある実施形態に従うと、HCA上の両方のポートが同時に使用されるケース（すなわちアクティブ-アクティブスキーム）では、同時フローがいくつかのISLを共有しているか、またはさらには同じ宛先ポートをターゲットにしていることができる可能性がある場合、ローカルHCAポート間でターゲットグループを共有することは意味をなし得る。

【0478】

ある実施形態に従うと、HCAポート間でターゲットグループを共有する別の理由は、HCAローカルメモリ帯域幅が両方の（すべての）HCAポートについて全速リンク速度を維持できないかどうかである。この場合、ターゲットグループは、どのポートがソースHCAまたは宛先HCAのいずれに関与するかに関係なく、集約された総リンク帯域幅がローカルメモリ帯域幅を決して超えないように、設定することができる。

40

【0479】

ある実施形態に従うと、特定の宛先に向かう固定ルートの場合、任意の中間ターゲットグループは、典型的には、経路における特定のステージにおいて単一のISLのみを表すことになる。しかしながら、動的転送がアクティブである場合、ターゲットグループおよびECN処理の両方がこれを考慮しなければならない。スイッチのペア間の並列ISL間のトラフィック（例えば、単一のリーフスイッチから単一のスパインスイッチへのアップリンク）を均衡させるためにのみ動的転送判断が生じる場合、すべての処理は、原則的には、単一のISLのみが使用されている場合と非常に類似している。FECN通知は、関

50

連するグループ内のすべてのポートの状態に基づいて行われると思われ、シグナリングは、ポートのいずれかからの輻輳指示に基づいてシグナリングされるという意味で「積極的」であり得るか、またはより保守的であり得、グループ内のすべてのポートについての共有出力キューのサイズに基づき得る。ターゲットグループ構成は、任意のパケットがその時点で最良の出力ポートを選択することを転送が可能にする限り、通常は、グループ内のすべてのリンクについての集約された帯域幅を表すであろう。しかしながら、フローごとに厳密なパケット順序維持の概念がある場合、いくつかのフローはある時点で同じ E S L を使用「しなければならない」場合があるため、帯域幅割当の評価はより複雑である。そのようなフロー順序スキームが、十分に定義されたヘッダフィールドに基づく場合、グループ内の各ポートを独立したターゲットグループとして表すことが最良である場合がある。この場合、送信元側 H C A におけるターゲットグループの選択は、R C Q P 接続またはアドレスハンドルに関連付けられることになるヘッダフィールドの評価を、スイッチがすべてのパケットについて実行時に実行するものと同じにすることができなければならない。

#### 【0480】

ある実施形態に従うと、デフォルトでは、新たな遠隔ターゲットについての初期ターゲットグループレートは、保守的に低く設定され得る。このようにして、ターゲットが関連レートを更新する機会を有するまで固有のスロットリングが存在する。したがって、すべてのそのようなレート制御は、関与する V M 自体とは無関係であるが、V M は、イングレストラフィックとエグレストラフィックとの両方について異なる遠隔ピアの割当を更新するようにハイパーバイザに要求することができるであろうが、これは、ローカル v H C A ポートと遠隔 v H C A ポートとの両方について定義された総制約内でのみ許可されるであろう。

#### 【0481】

##### ピアノード、経路およびターゲットグループの相関付け：

ある実施形態に従うと、V M が異なるピアノードおよび異なるグループのピアノードに関連付けられる帯域幅制限を識別できるようにするためには、どのターゲットグループが様々な通信ピア（および関連付けられるアドレス／経路情報）に関連付けられるかを問い合わせる方法が必要であろう。通信ピアのセットを様々なターゲットグループと相関させること、ならびに様々なターゲットグループが表すレート制限に基づいて、V M は、様々な通信ピアに対してどのような帯域幅を達成することができるかを追跡することができるであろう。これは、次いで、原則として、V M が、競合するターゲットグループを伴わない同時転送をできるだけ多く有することにより、最良の可能な帯域幅利用を経時的に達成する態様で、通信動作をスケジューリングすることを可能にするであろう。

#### 【0482】

##### H C A リソース制限グループとターゲットグループとの間の関係：

ある実施形態に従うと、H R L G 概念およびターゲットグループ概念は、それらが両方とも柔軟な態様で V M およびテナント間で定義ならびに共有されることができる帯域幅制限を表すという点で、いくつかの態様で重複している。しかしながら、H R L G の主な焦点は、ローカル H C A ／ H C A ポート容量の異なる部分を異なる V F （およびそれによって V M およびテナント）にどのように割り当てることができるかを定義することである一方、ターゲットグループ概念は、最終宛先および中間ファブリックトポロジの両方に関してローカル H C A の外部に存在する帯域幅制限およびフロー制御制約に焦点を当てている。

#### 【0483】

ある実施形態に従うと、このようにして、様々な V F がローカル H C A 容量のどのようなシェアを用い得るかを制御するための方法として H R L G を用いるが、許可された容量がどのようなファブリックまたは遠隔ターゲット制限または輻輳条件とも競合しない態様でのみ使用され得ることを保証することは意味がある。—これらの外部制約は、次いで、動的に制御され、関連するターゲットグループを介して反映される。

#### 【0484】

ある実施形態に従うと、実現化の点から、すべての関連するターゲットグループの状態は、どのローカルQPに対するどの保留中の作業要求がフロー制御状態にあることになるかを定義することになり、そこでは、それらは、任意の時点で、より多くのエグレスデータトラフィックを生成することを許可される。この状態は、QPが行うことが実際に送信すべき何かを有するかに関する状態とともに、次いで、どのVFが次に送信するための候補であるかに関して、VF/VHCAポートレベルで集約され得る。HCAポート上で次に送信するためにどのようなVFをスケジューリングするかについての決定は、HRLG階層における様々なHRLGの状態およびポリシー、「送信準備完了」VFのセット、およびどのVFがどのようなエグレストラフィックを生成したかに関する最近の履歴に基づくことになる。選択されたVFについて、VF固有の調停ポリシーは、データ転送のためにどのQPが選択されるかを定義することになる。

10

#### 【0485】

ある実施形態に従うと、保留中のデータ転送を有するQPのセットは、ローカル作業要求を有するQPと、関連する遠隔ピアからの保留中のRDMA読出要求を有するQPとの両方を含むので、上記のスケジューリングおよび調停は、すべての保留中のエグレスデータトラフィックに対処することになる。

#### 【0486】

ある実施形態に従うと、イングレストラフィック（着信RDMA読出応答を含む）は、遠隔ピアノードにおけるすべての関連するターゲットグループの現在の状態によって制御されることになる。—この（遠隔）状態は、輻輳状況に基づく動的フロー制御状態と、このHCA上のローカルVFに対するイングレス帯域幅割当の変化を反映する、このHCAからの明示的な更新との両方を含む。そのようなイングレス帯域幅割当は、HRLG階層によって反映されるポリシーに基づくことになる。このようにして、様々なVMは、イングレスとエグレスとの両方に対して、およびイングレスとエグレスとの両方に対する優先度ごとに基づいても、「微調整された」独立した帯域幅割当を有し得る。

20

#### 【0487】

##### SLAクラス：

ある実施形態に従うと、以下の提案は、非ブロッキング2層ファットツリートポロジーが、単一のリーフスイッチの基数を超えるシステムサイズ（物理ノードカウント）のために使用されている、と仮定する。また、物理サーバ上の単一のVMが（1つ以上のHCA/VFを介して）すべてのファブリック帯域幅を用いることができると仮定される。したがって、物理サーバあたりのテナントあたりのVM数は、SLA要因とみなす必要があるパラメータではない。

30

#### 【0488】

ある実施形態に従うと、1番上のレベルの層（たとえばプレミアムプラス）は、

- ・専用サーバしか使用できない。
- ・VM（またはHAポリシー）の数およびサイズが追加の距離を意味する場合を除いて、可能な限り同じリーフドメインに割り振られることができる。
- ・テナントが複数のリーフドメインを用いている場合、ローカルリーフからの非ブロッキングアップリンク帯域幅を常に有することができる。
- ・すべての「フローグループ」を用いることができる。

40

#### 【0489】

ある実施形態に従うと、下位レベルの層（たとえばプレミアム）を提供することができる

- ・専用サーバのみ使用することができ、同じリーフドメインについての保証はない。
- ・非ブロッキングアップリンク帯域幅の少なくとも50%（すなわち、同じリーフドメイン内でこのテナントに割り振られるサーバの数に対して）を保証されることができる。
- ・すべての「フローグループ」を用いることができる。

#### 【0490】

ある実施形態に従うと、第3レベルの層を提供することができる（たとえばエコノミープラス）

50

・共有サーバを用いてもよいが、4つの専用の「フローグループ」を有することになる（すなわち、ファブリック内の異なるバッファプールおよび調停グループを表す優先度など）。

○これらのリソースは、ローカルHCAおよびスイッチポート専用であることになるが、ファブリック内で共有されることになる。

・ローカルサーバからのすべての利用可能な帯域幅（エグレスおよびイングレス）を用いる能力を有することができるが、全帯域幅の少なくとも50%を有することが保証される。

・物理サーバ毎に1つのエコノミープラステナントに限定される。

・このエコノミープラステナントによって使用されているサーバの数に対して少なくとも25%の非ブロッキングリーフアップリンク帯域幅を保証されることができる。

【0491】

ある実施形態に従うと、第4の層を提供することができる（たとえばエコノミー）

・共有サーバを用いることしかできない

・専用の優先度を有さない

・サーバ帯域幅（エグレスおよびイングレス）の50%までを使用できるようにされることができるが、これを3つまでの他のエコノミーテナントと共有する場合がある

・利用可能なリーフアップリンク帯域幅の25%までを同じリーフドメイン内の他のエコノミーテナントと共有することができる。

【0492】

ある実施形態に従うと、保証された帯域幅を伴わない予備容量を用いることができる底層（たとえばスタンバイ）を提供することができる

本教示の様々な実施形態を説明してきたが、上記実施形態が限定ではなく例示として提示されていることが理解されるべきである。上記実施形態は、本教示の原理およびそれらの実際の適用例を説明するために選択され記載されたものである。上記実施形態は、新たな特徴および／もしくは改善された特徴を提供することによって、ならびに／または、リソース利用の低減、容量の増加、効率の向上および待ち時間の低下などの利点を提供することによって、システムおよび方法の性能を向上させるために本教示が利用されているシステムおよび方法を例示している。

【0493】

いくつかの実施形態においては、本教示の特徴は、全体的または部分的に、プロセッサ、メモリなどの記憶媒体、および他のコンピュータと通信するためのネットワークカードを含むコンピュータにおいて実現される。いくつかの実施形態においては、本教示の特徴は、コンピュータの1つ以上のクラスタがローカルエリアネットワーク（Local Area Network: LAN）、スイッチファブリックネットワーク（例えば、インフィニバンド）、またはワイドエリアネットワーク（Wide Area Network: WAN）などのネットワークによって接続されている分散コンピューティング環境において実現される。分散コンピューティング環境は、一箇所において全てのコンピュータを有していてもよく、または、WANによって接続されているさまざまな遠隔地理位置においてコンピュータのクラスタを有していてもよい。

【0494】

いくつかの実施形態においては、本教示の特徴は、全体的または部分的に、ウェブ技術を用いたセルフサービスの調整された態様でユーザに送達される共有型で融通性のあるリソースに基づいて、クラウド・コンピューティング・システムの一部またはサービスとしてクラウドにおいて実現される。（米国標準技術局（National Institute of Standards and Technology）によって定義される）クラウドの5つの特徴がある。すなわち、オン・デマンドのセルフサービス、広域ネットワークアクセス、リソースプール化、高速伸縮性、およびメジャーサービスである。例えば、本明細書に引用によって援用されている「クラウドコンピューティングのNIST定義（The NIST Definition of Cloud Computing）」（特殊出版（Special Publication）800～145（2011））を参照されたい。

クラウド展開モデルは、パブリック、プライベートおよびハイブリッドを含む。クラウドサービスモデルは、ソフトウェア・アズ・ア・サービス（Software as a Service：SaaS）、プラットフォーム・アズ・ア・サービス（Platform as a Service：PaaS）、データベース・アズ・ア・サービス（Database as a Service：DBaaS）およびインフラストラクチャ・アズ・ア・サービス（Infrastructure as a Service：IaaS）を含む。本明細書で使用するとき、クラウドは、セルフサービスの調整された態様で、共有される融通性のあるリソースをユーザに対して配信する、ハードウェア技術とソフトウェア技術とネットワーク技術とウェブ技術とを組合せたものである。特に指定がなければ、クラウドは、本明細書で使用するとき、パブリッククラウド、プライベートクラウドおよびハイブリッドクラウドの実施形態を包含しており、全てのクラウド展開モデルは、クラウドSaaS、クラウドDBaaS、クラウドPaaSおよびクラウドIaaSを含むもののこれらに限定されない。

【0495】

いくつかの実施形態においては、本教示発明の特徴が、ハードウェア、ソフトウェア、ファームウェアまたはそれらの組合せを用いて、またはそれらの組合せの助けを借りて実現される。いくつかの実施形態においては、本教示の特徴は、本教示の1つ以上の機能を実行するように構成されたかまたはプログラムされたプロセッサを用いて実現される。プロセッサは、いくつかの実施形態においては、シングルプロセッサもしくはマルチチッププロセッサ、デジタル信号プロセッサ（digital signal processor：DSP）、システム・オン・ア・チップ（system on a chip：SOC）、特定用途向け集積回路（application specific integrated circuit：ASIC）、フィールドプログラマブルゲートアレイ（field programmable gate array：FPGA）もしくは他のプログラマブルロジックデバイス、ステートマシン、離散的なゲートもしくはトランジスタ論理、離散的なハードウェアコンポーネント、または、本明細書に記載される機能を実行するように設計されたそれらのいずれかの組合せである。いくつかの実現例においては、本教示の特徴が、特定の機能に特化した回路類によって実現され得る。他の実現例においては、これらの特徴は、例えば、コンピュータ可読記憶媒体上に格納された命令を用いて特定の機能を実行するように構成されたプロセッサにおいて実現され得る。

【0496】

いくつかの実施形態においては、本教示の特徴は、処理システムおよび／またはネットワークシステムハードウェアを制御するために、かつ、プロセッサおよび／またはネットワークが本教示の特徴を利用する他のシステムと対話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組込まれている。このようなソフトウェアまたはファームウェアは、アプリケーションコード、デバイスドライバ、オペレーティングシステム、仮想マシン、ハイパーバイザ、アプリケーションプログラミングインターフェイス、プログラミング言語、および実行環境／コンテナを含み得るがこれらに限定されない。適切なソフトウェアコーディングは、ソフトウェア技術に精通した当業者にとって明らかになるように、熟練したプログラマであれば本開示の教示に基づいて容易に準備することができる。

【0497】

いくつかの実施形態においては、本手法は、本教示を実施するよう用いられ得る命令を担持するコンピュータ可読媒体などのコンピュータプログラムプロダクトを含む。いくつかの例では、コンピュータ可読媒体は、命令が格納されたことによって命令を担持する記憶媒体またはコンピュータ可読媒体であり、これらの命令を用いて、本教示の処理または機能のいずれかを実行するように、コンピュータなどのシステムをプログラムするか、または他の方法で構成することができる。記憶媒体またはコンピュータ可読媒体は、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、および磁気光ディスクを含む任意のタイプのディスク、ROM、RAM、EPROM、EEPROM、DRAM、VRAM、フラッシュメモリデバイス、磁気または光カード、ナノシステム（分子メモリICを含む）、ならびに、命令および／またはデータを格納するの

に適した任意のタイプの媒体または装置、を含み得るが、これらには限定されない。特定の実施形態においては、記憶媒体またはコンピュータ可読媒体は、非一時的な記憶媒体または非一時的なコンピュータ可読媒体である。コンピュータ可読媒体は、また、または代替的に、そのような命令を伝搬する搬送波または伝送信号などの過渡的媒体を含み得る。

【0498】

したがって、一観点から、高性能コンピューティング環境においてプライベートファブリックにおける帯域幅輻輳制御を提供するためのシステムおよび方法が説明されてきた。例示的な方法は、1つ以上のマイクロプロセッサにおいて、第1のサブネットを提供することができ、第1のサブネットは、複数のスイッチと、複数のホストチャネルアダプタを含み、ホストチャネルアダプタの各々は、少なくとも1つのホストチャネルアダプタポートを含み、複数のホストチャネルアダプタは、上記複数のスイッチを介して相互接続され、第1のサブネットはさらに、複数のエンドノードを含む。本方法は、ホストチャネルアダプタにおいて、ホストチャネルアダプタに取り付けられたエンドノードに関連付けられるエンドノードイングレス帯域幅割当を提供することができる。本方法は、ホストチャネルアダプタのエンドノードにおいて、イングレス帯域幅を受信することができ、イングレス帯域幅はエンドノードのイングレス帯域幅割当を超える。

【0499】

上述の記載は、網羅的となるように意図されたものではなく、または、本範囲を開示通りの形態に限定するように意図されたものではない。また、本発明の実施形態を特定の一連のトランザクションおよびステップを用いて説明したが、本範囲が上述の一連のトランザクションおよびステップに限定されないことは、当業者にとって明らかであろう。さらに、実施形態をハードウェアとソフトウェアとの特定の組合せを用いて説明したが、ハードウェアとソフトウェアとの他の組合せが本教示の範囲内にあることも認識すべきである。さらに、さまざまな実施形態で本発明の特徴の特定の組合せを記載したが、一実施形態の特徴が別の実施形態に組み込まれ得るというように、これらの特徴の異なる組合せが本教示の範囲内にあることは当業者にとって明らかであることを理解すべきである。さらに、本精神および範囲から逸脱することなく、形態、詳細、実施および用途のさまざまな追加、削減、削除、変形および他の変更がなされ得ることも、当業者にとっては明らかであろう。本発明は特許請求の範囲の適切な解釈によって規定されることが意図される。

10

20

30

40

50

【図面】  
【図 1】

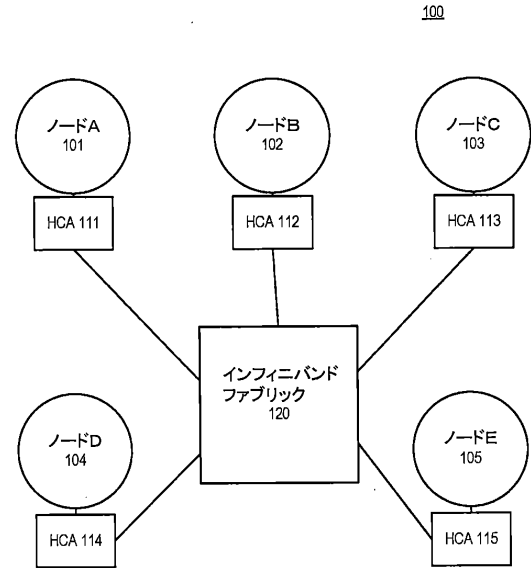


FIGURE 1

【図 2】

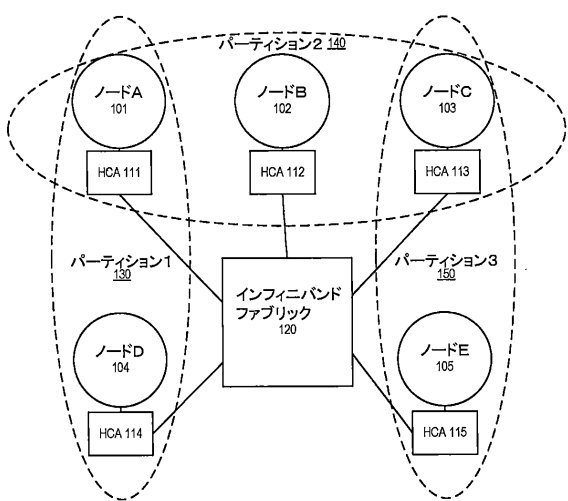


FIGURE 2

【図 3】

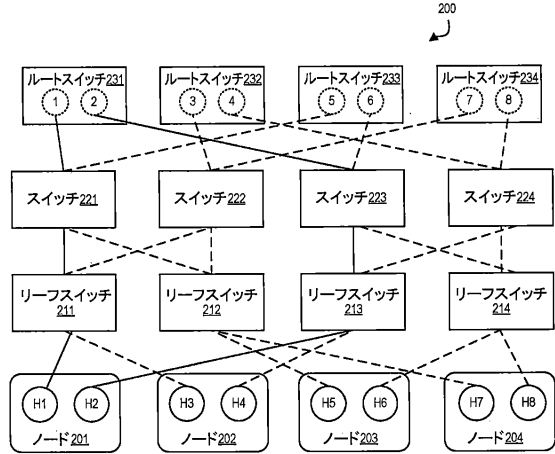


FIGURE 3

【図 4】

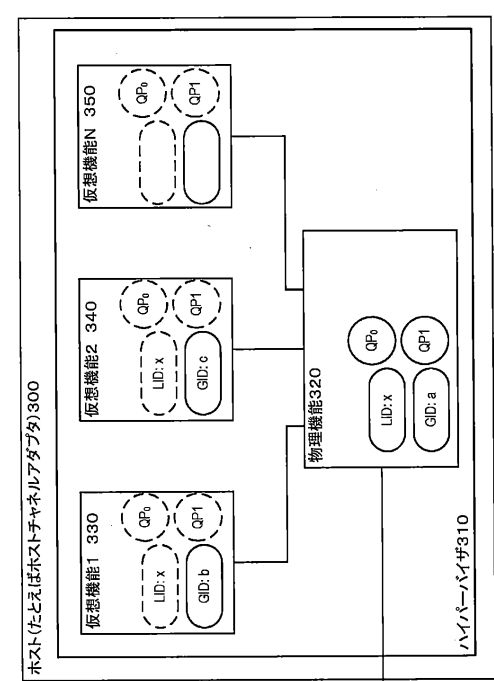


FIGURE 4

【図 5】

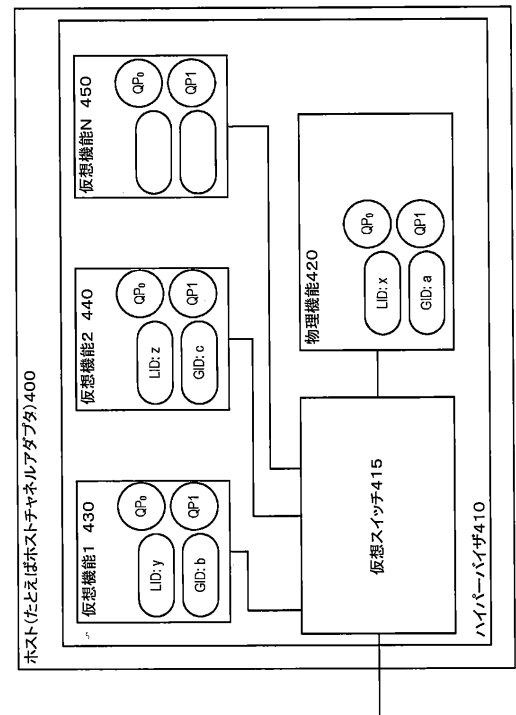


FIGURE 5

【図 6】

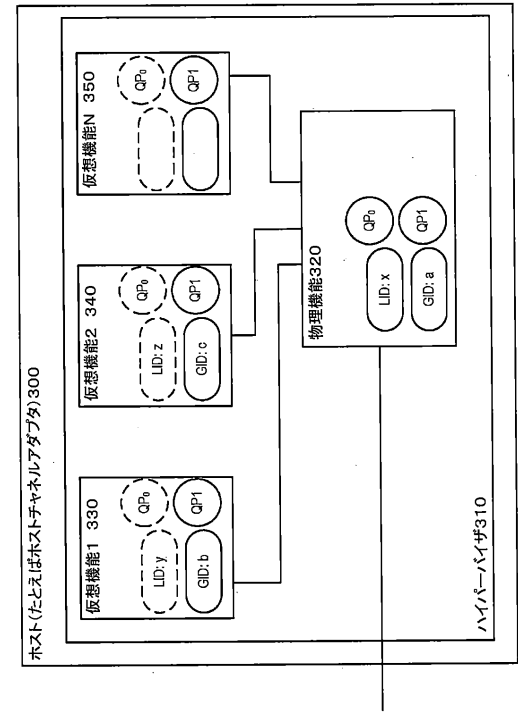


FIGURE 6

【図 7】

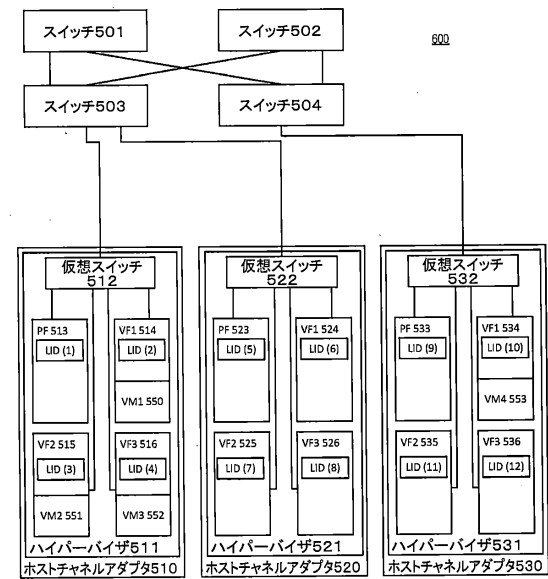


FIGURE 7

【図 8】

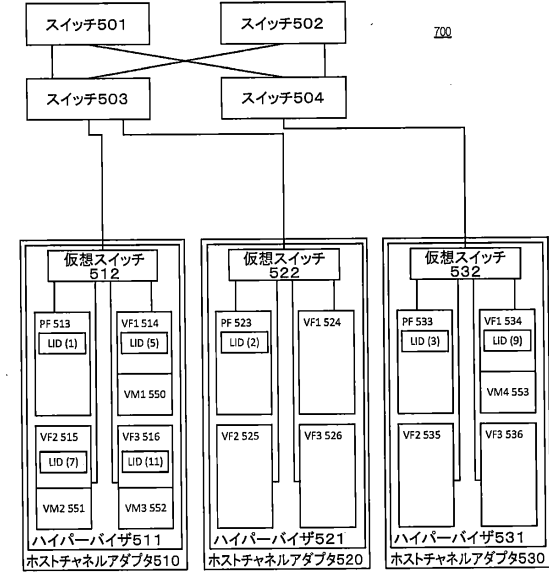


FIGURE 8

10

20

30

40

50

【図 9】

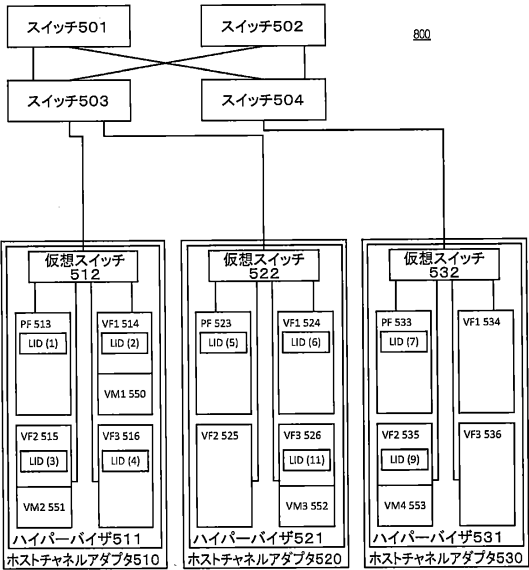


FIGURE 9

【図 10】

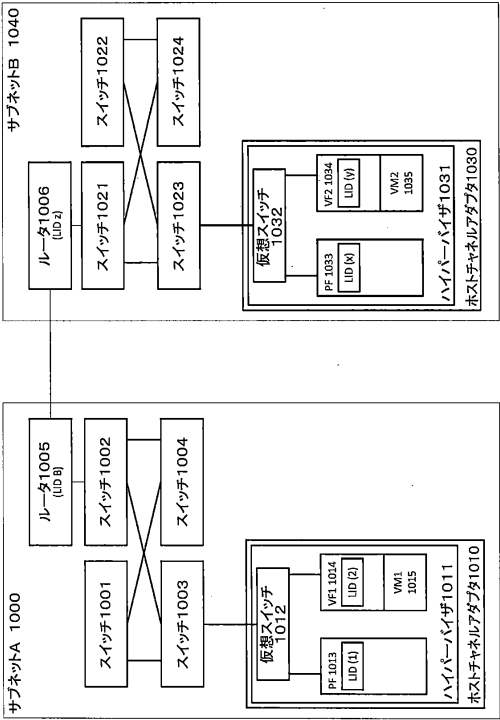


FIGURE 10

【図 11】

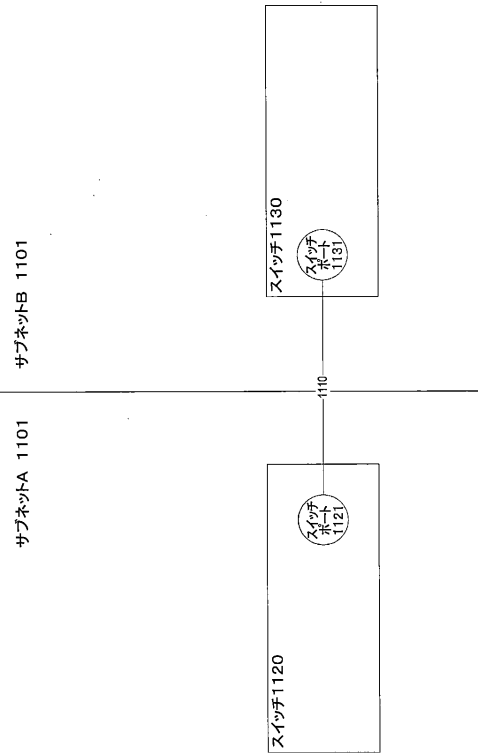


FIGURE 11

【図 12】

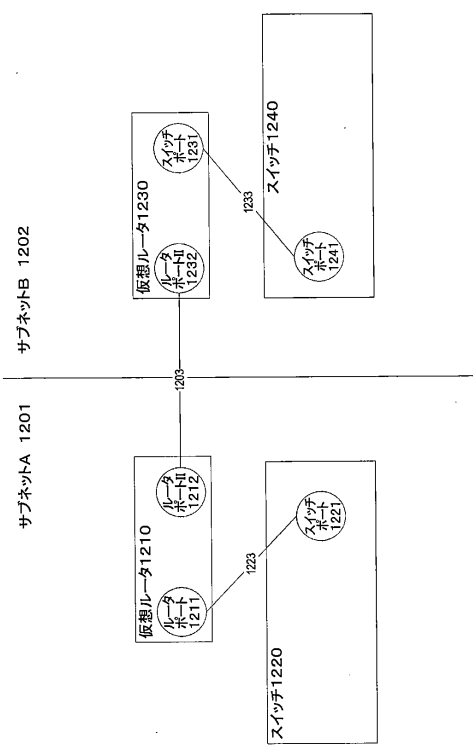


FIGURE 12

10

20

30

40

50

【図 13】

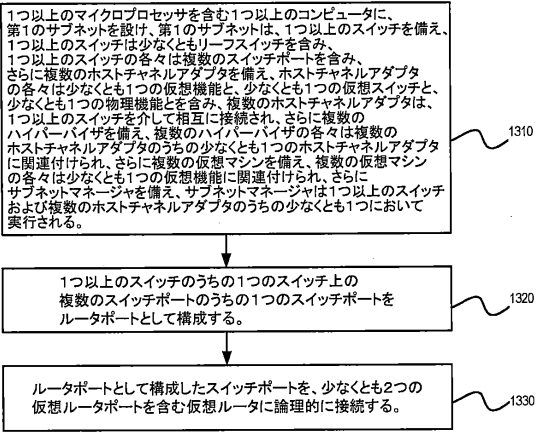


FIGURE 13

【図 14】

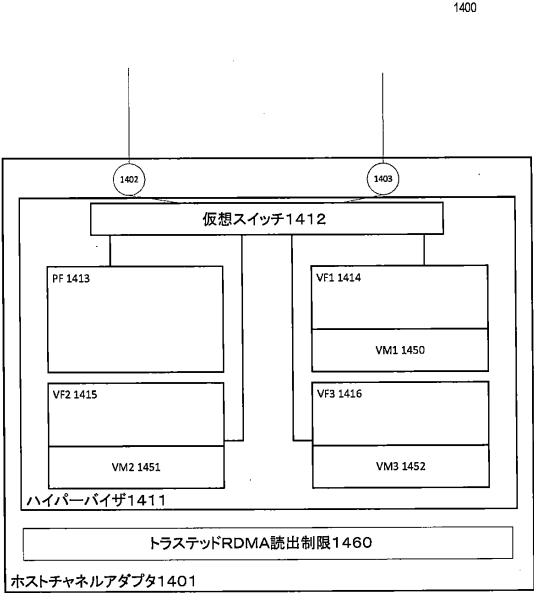


FIGURE 14

【図 15】

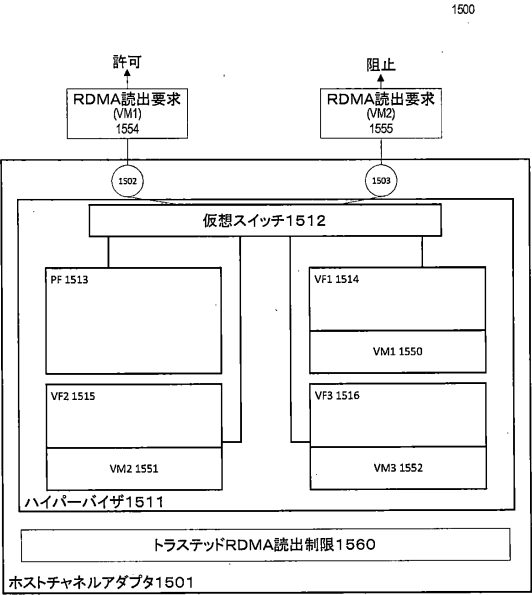


FIGURE 15

【図 16】

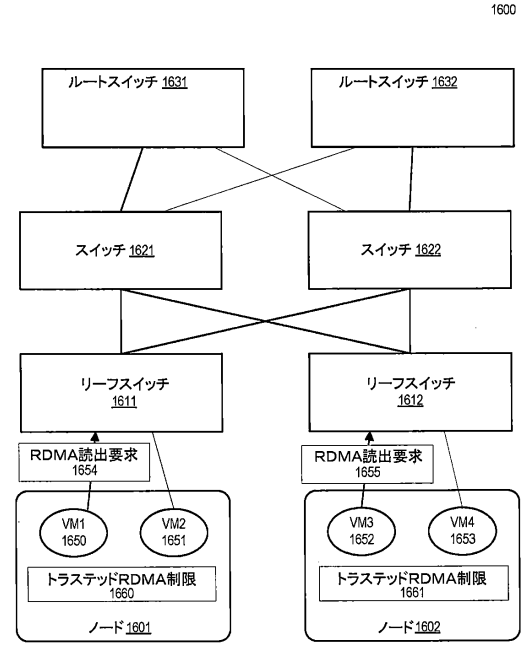


FIGURE 16

10

20

30

40

50

【図 17】

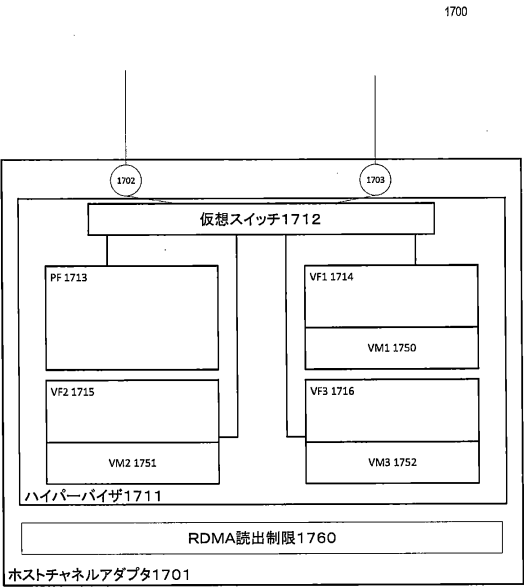


FIGURE 17

【図 18】

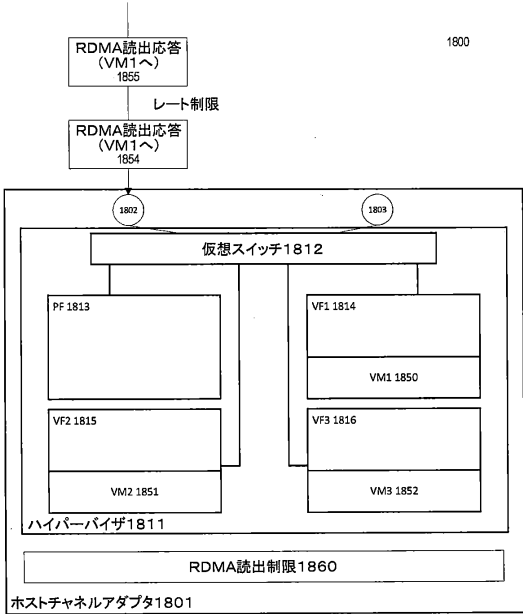


FIGURE 18

【図 19】

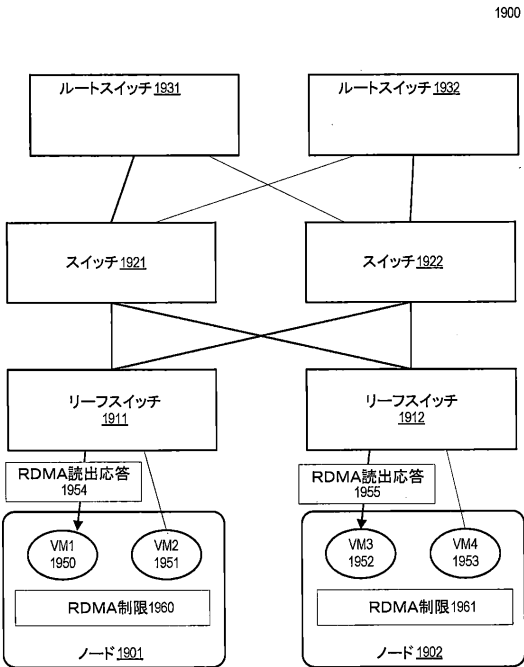


FIGURE 19

【図 20】

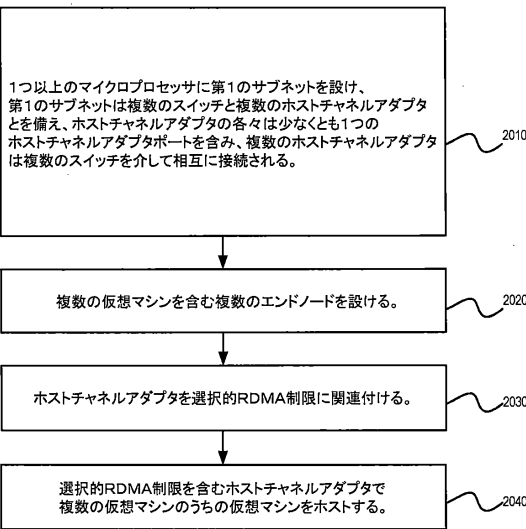
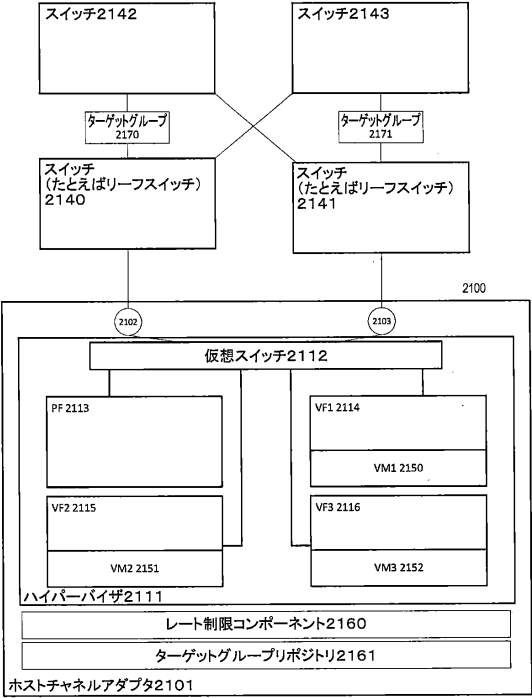
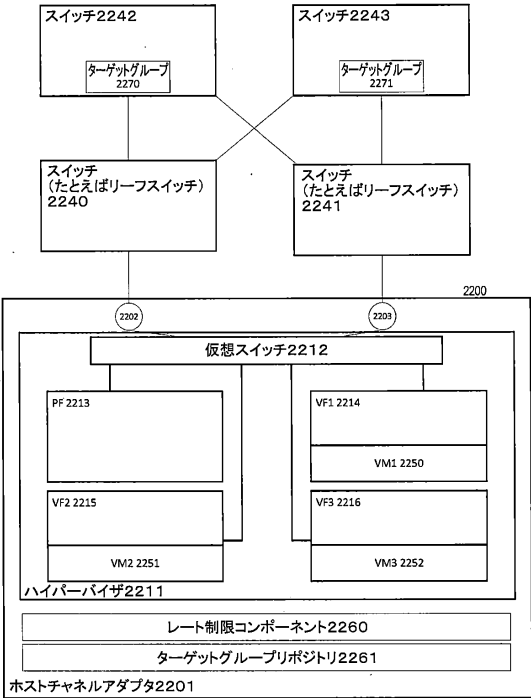


FIGURE 20

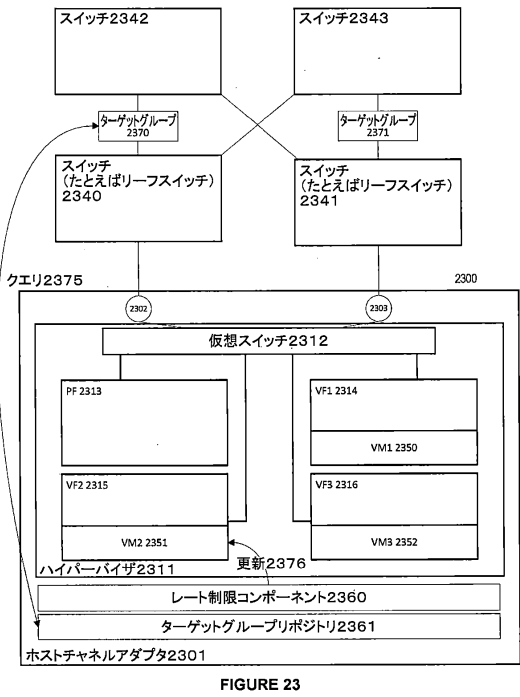
【図 2 1】



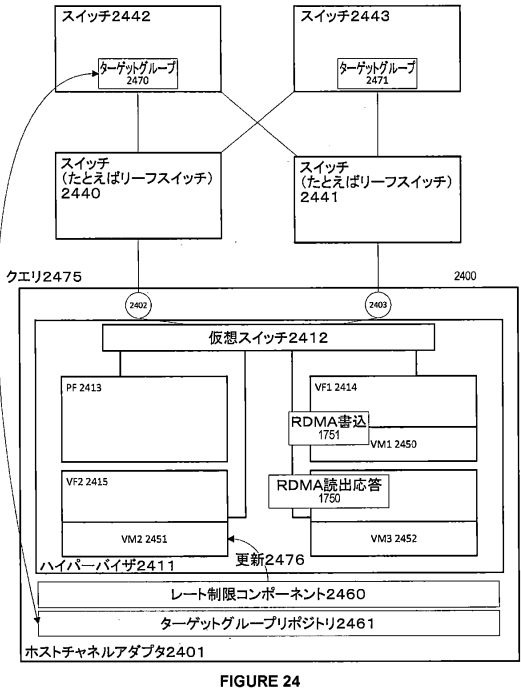
【図 2 2】



【図 2 3】



【図 2 4】



10

20

30

40

50

【図 25】

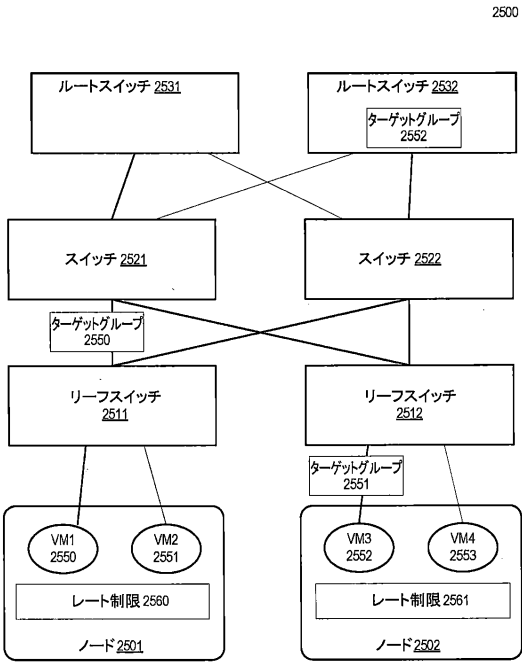


FIGURE 25

【図 26】

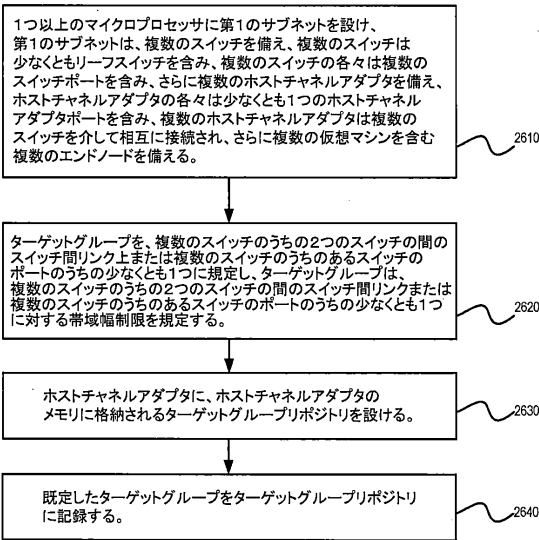


FIGURE 26

【図 27】

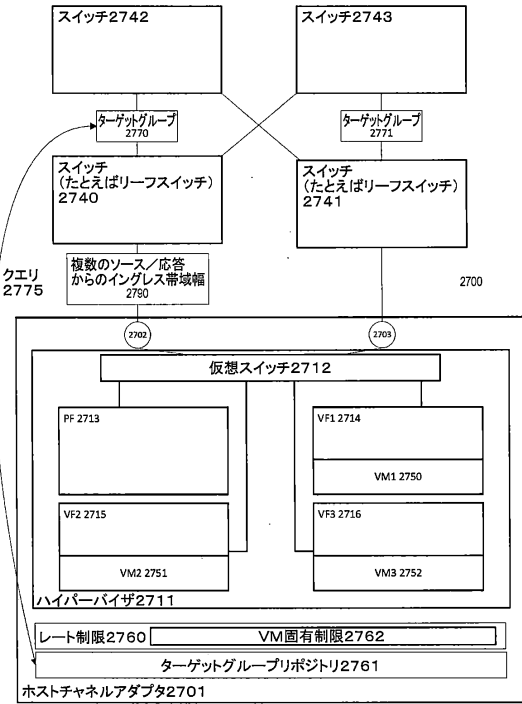


FIGURE 27

【図 28】

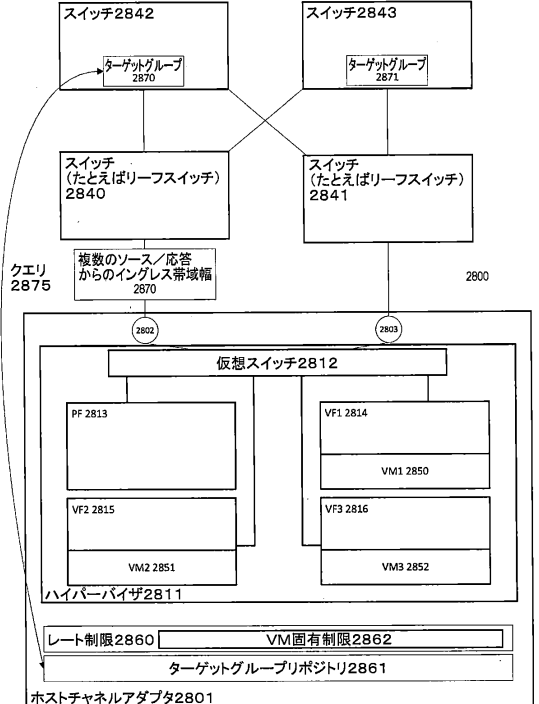


FIGURE 28

【図 29】

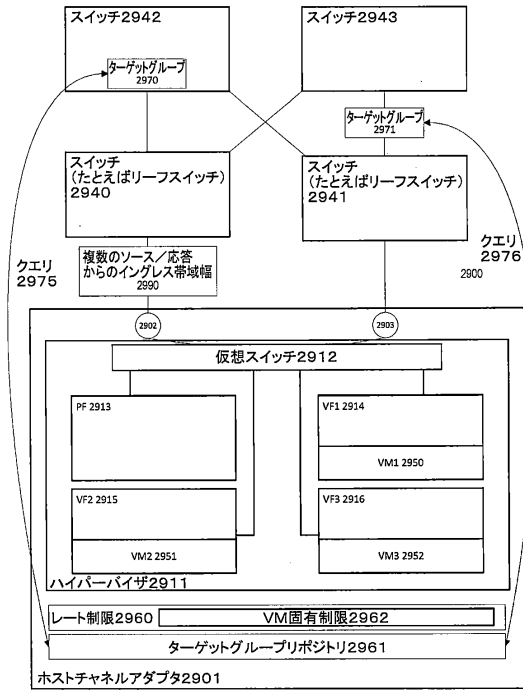


FIGURE 29

【図 30】

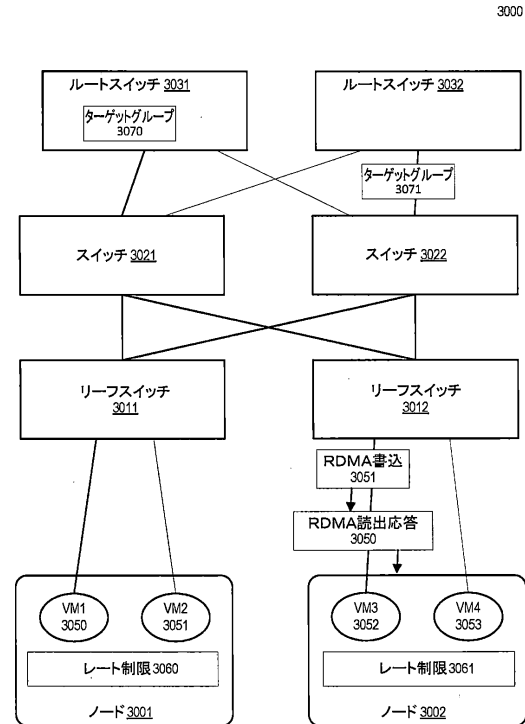


FIGURE 30

【図 31】

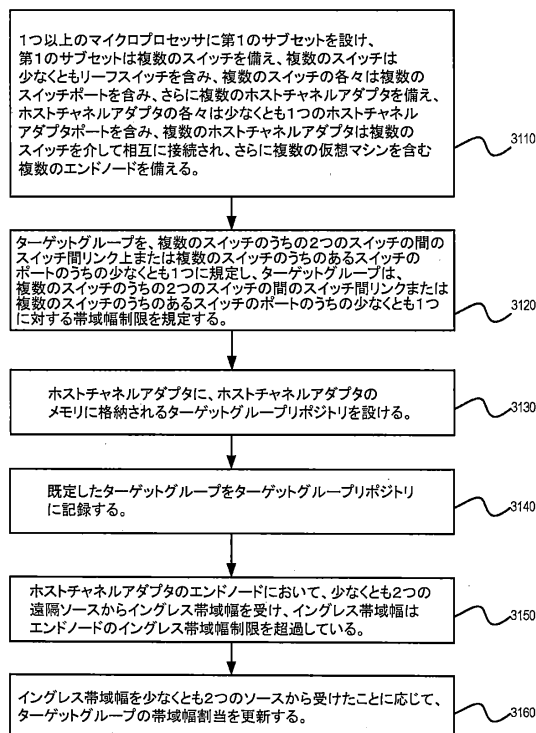


FIGURE 31

【図 32】

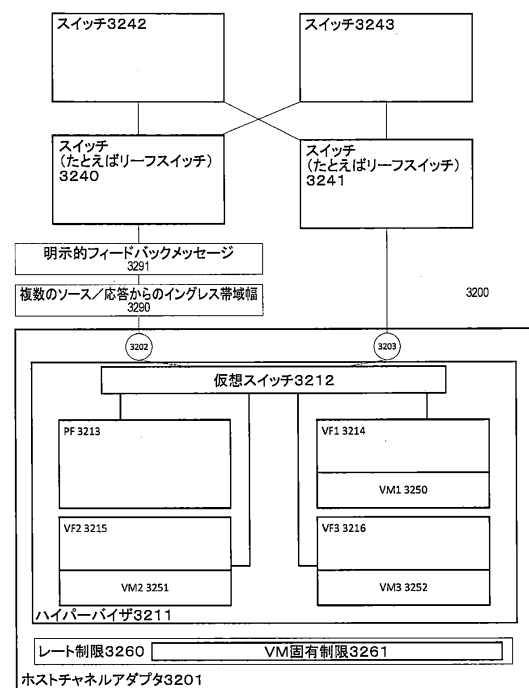


FIGURE 32

10

20

30

40

50

【図 3 3】

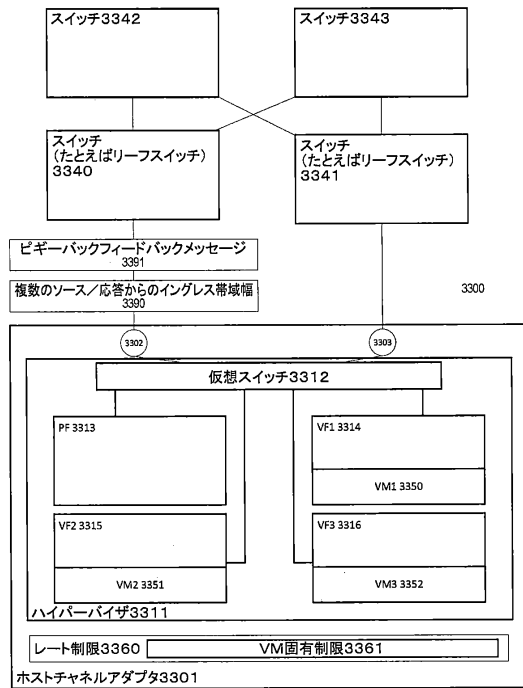


FIGURE 33

【図 3 4】

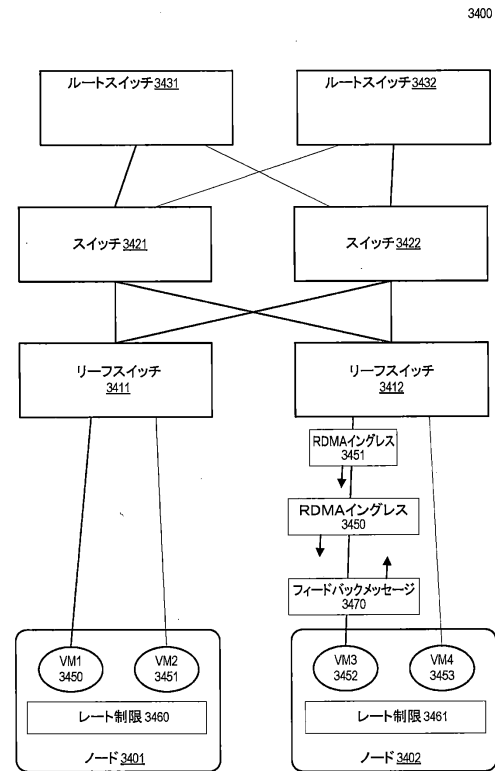


FIGURE 34

【図 3 5】

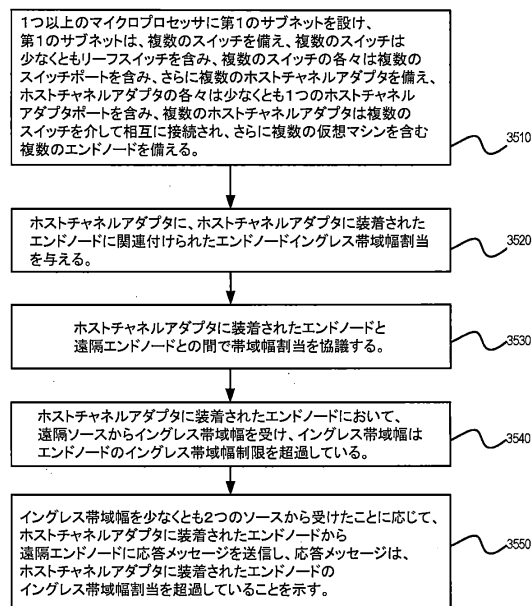


Figure 35

【図 3 6】

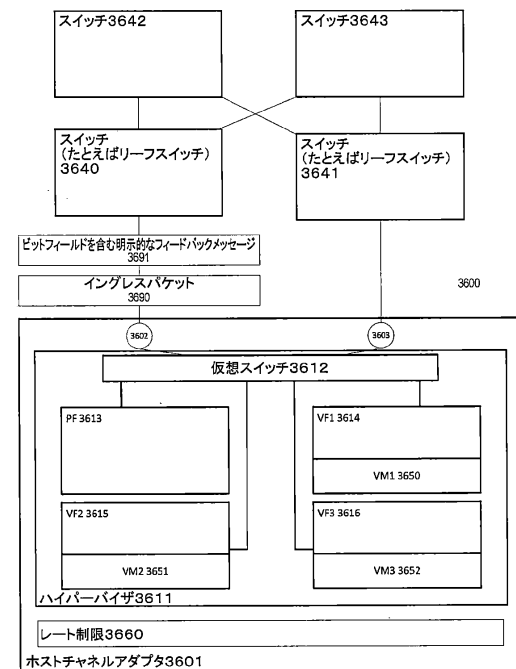


FIGURE 36

10

20

30

40

50

【圖 3 7】

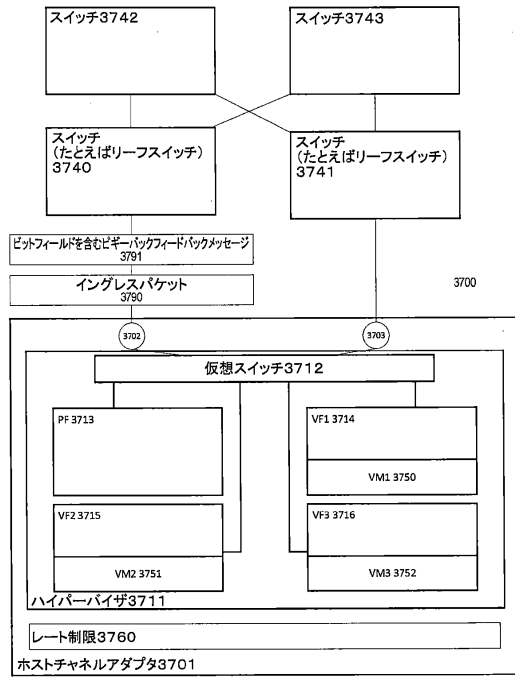


FIGURE 37

【図 38】

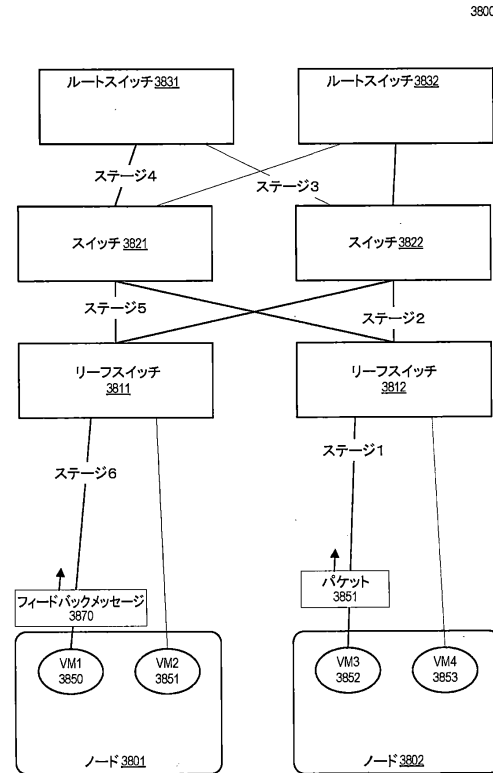


FIGURE 38

【圖 39】

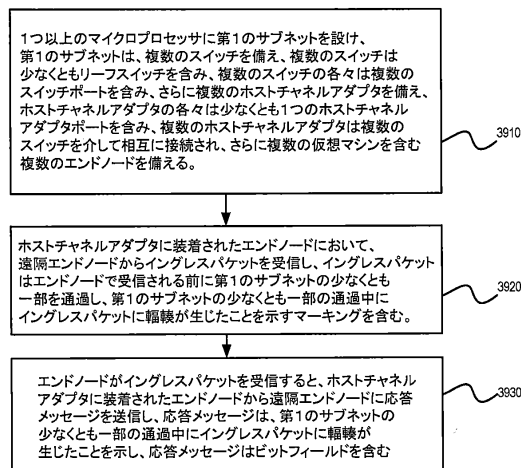


FIGURE 39

## フロントページの続き

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 16/872,038

(32)優先日 令和2年5月11日(2020.5.11)

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 16/872,039

(32)優先日 令和2年5月11日(2020.5.11)

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 16/872,043

(32)優先日 令和2年5月11日(2020.5.11)

(33)優先権主張国・地域又は機関

米国(US)

(72)発明者 シッダバスニ, アジョイ

アメリカ合衆国、94065 カリフォルニア州、レッドウッド・シティー、オラクル・パークウェイ、500、オラクル・インターナショナル・コーポレーション

(72)発明者 パント, アブニーシュ

アメリカ合衆国、94065 カリフォルニア州、レッドウッド・シティー、オラクル・パークウェイ、500、オラクル・インターナショナル・コーポレーション

(72)発明者 ジョラード, アマーナス

アメリカ合衆国、94065 カリフォルニア州、レッドウッド・シティー、オラクル・パークウェイ、500、オラクル・インターナショナル・コーポレーション

審査官 中川 幸洋

(56)参考文献 特表2018-537004(JP, A)

特開2000-041060(JP, A)

米国特許出願公開第2016/0094450(US, A1)

特表2019-503595(JP, A)

(58)調査した分野 (Int.Cl., DB名)

H04L 47/12

H04L 47/70