(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2019-514294
(P2019-514294A)
(43)公表日　令和1年5月30日(2019.5.30)

(51)Int.Cl.			F I			テーマコード（参考）
H04L	12/803	(2013.01)	H04L	12/803		5K030
H04L	12/70	(2013.01)	H04L	12/70	D	

審査請求 未請求　予備審査請求 未請求　（全 20 頁）

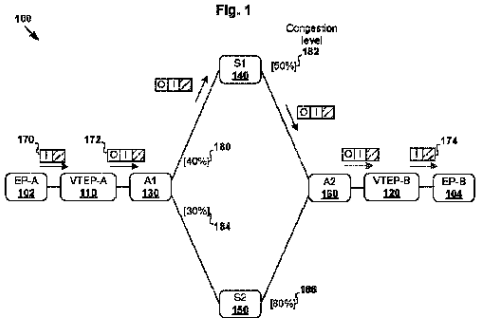
(21)出願番号	特願2018-553946 (P2018-553946)
(86)(22)出願日	平成29年4月12日 (2017.4.12)
(85)翻訳文提出日	平成30年12月12日 (2018.12.12)
(86)国際出願番号	PCT/US2017/027190
(87)国際公開番号	WO2017/180731
(87)国際公開日	平成29年10月19日 (2017.10.19)
(31)優先権主張番号	62/321,730
(32)優先日	平成28年4月12日 (2016.4.12)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	15/485,089
(32)優先日	平成29年4月11日 (2017.4.11)
(33)優先権主張国	米国 (US)

(71)出願人　511235548
ニシラ，　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９４３０４，　パロアルト，　ヒルビュー　アベニュー　３４０１
(74)代理人　100105957
弁理士　恩田　誠
(74)代理人　100068755
弁理士　恩田　博宣
(74)代理人　100142907
弁理士　本田　淳
(72)発明者　ヒラ、ムケーシュ
アメリカ合衆国　９４３０４　カリフォルニア州　パロ　アルト　ヒルビュー　アベニュー　３４０１

最終頁に続く

(54)【発明の名称】輻輳を考慮したロードバランシングのための仮想トンネルエンドポイント

(57)【要約】
例示的な方法は、送信元仮想トンネルエンドポイント（ＶＴＥＰ）のために提供されて、データセンタネットワークにおいて輻輳を考慮したロードバランシングを実行する。方法は、送信元ＶＴＥＰが、送信元ＶＴＥＰを送信先ＶＴＥＰに接続する対応する複数の中間スイッチによって提供される複数の経路に関連付けられた輻輳状態情報を学習すること、を備える。また方法は、送信元ＶＴＥＰが、送信元エンドポイントによって送信され且つ送信先ＶＴＥＰに関連付けられた送信先エンドポイント宛ての複数の第２のパケットを受信すること、輻輳状態情報に基づいて複数の経路から特定の経路を選択すること、を備える。さらに方法は、送信元ＶＴＥＰが、特定の経路に関連付けられたタプルセットを含むヘッダ情報を有する前記複数の第２のパケットの各々をカプセル化することによってカプセル化された複数の第２のパケットを生成すること、カプセル化された複数の第２のパケットを送信先エンドポイントに送信すること、を備える。



Congestion state information at VTEP-A 190

Outer source port number (source_PN) 192	Path (path_ID) 194	Congestion flag (congestion_flag) 196	Path weight (weight) 198
SP1	P1 (via S1)	false	w1
SP2	P2 (via S2)	true	w2

【特許請求の範囲】
【請求項１】
送信元仮想トンネルエンドポイント（ＶＴＥＰ）が、該送信元仮想トンネルエンドポイント、送信先仮想トンネルエンドポイント、送信元エンドポイント、送信先エンドポイント、複数の中間スイッチを含むデータセンタネットワークにおいて輻輳を考慮したロードバランシングを実行する方法であって、
前記送信先仮想トンネルエンドポイントからの複数の第１のパケットに基づいて、前記送信元仮想トンネルエンドポイントが、前記送信元仮想トンネルエンドポイントを前記送信先仮想トンネルエンドポイントに接続する対応する複数の中間スイッチによって提供される複数の経路に関連付けられた輻輳状態情報を学習すること、
前記送信元仮想トンネルエンドポイントが、前記送信元エンドポイントによって送信され且つ前記送信先仮想トンネルエンドポイントに関連付けられた送信先エンドポイント宛ての複数の第２のパケットを受信すること、
前記送信元仮想トンネルエンドポイントが、前記輻輳状態情報に基づいて前記複数の経路から特定の経路を選択すること、
前記送信元仮想トンネルエンドポイントが、前記特定の経路に関連付けられたタプルセットを含むヘッダ情報を有する前記複数の第２のパケットの各々をカプセル化することによってカプセル化された複数の第２のパケットを生成すること、
前記カプセル化された複数の第２のパケットが前記タプルセットに基づいた前記特定の経路を介して転送されるように、前記送信元仮想トンネルエンドポイントが、前記カプセル化された複数の第２のパケットを前記送信先エンドポイントに送信すること、を備える方法。
【請求項２】
前記輻輳状態情報を学習することは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットに基づいて前記特定の経路に関連付けられた輻輳フラグを決定することを含み、
前記特定の第１のパケットは、
前記特定の経路が輻輳しているかどうかを示すための前記複数の中間スイッチのうちの少なくとも１つからの輻輳通知を含む、請求項１に記載の方法。
【請求項３】
前記輻輳状態情報を学習することは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットの受信時刻、および前記特定の第１のパケットをトリガするための前記送信元仮想トンネルエンドポイントによって送信されたプローブパケットの送信時刻に基づいて、前記特定の経路に関連付けられたラウンドトリップタイムを決定することを含む、請求項１に記載の方法。
【請求項４】
前記輻輳状態情報を学習することは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットに基づいて、選択の可能性を増加または低下させるために前記特定の経路に関連付けられた重み付けを決定すること、を含む、請求項１に記載の方法。
【請求項５】
前記輻輳状態情報を学習することは、
複数の外部送信元ポート番号と、対応する複数の経路との間のマッピングを学習することを含む、請求項１に記載の方法。
【請求項６】
前記複数の第２のパケットの各々をカプセル化することは、
前記複数の外部送信元ポート番号から、前記特定の経路に関連付けられた特定の外部送信元ポート番号を決定すること、
前記特定の送信元ポート番号を含むため、前記複数の第２のパケットの各々において前記ヘッダ情報のタプルセットを構成すること、を含む、請求項５に記載の方法。

【請求項７】
前記送信先エンドポイント宛ての前記現在の第２のパケットと以前の第２のパケットとの間のパケット間ギャップに基づいて、現在の第２のパケットである前記複数の第２のパケットのそれぞれをフローレットに割り当てること、
前記特定の経路に関連付けられた特定の外部送信元ポート番号に前記フローレットを関連付けるフローレット情報を記憶すること、をさらに備える請求項６に記載の方法。
【請求項８】
１組の命令を含む非一時的なコンピュータ可読記憶媒体であって、
前記１組の命令は、ホストのプロセッサによって実行されると、
前記プロセッサに送信元仮想トンネルエンドポイント（ＶＴＥＰ）を実行させて、前記送信元仮想トンネルエンドポイント、送信先仮想トンネルエンドポイント、送信元エンドポイント、送信先エンドポイント、複数の中間スイッチを含むデータセンタネットワークにおいて輻輳を考慮したロードバランシングの方法を実行し、前記ホストは、前記送信元仮想トンネルエンドポイントをサポートし、
前記方法は、
前記送信先仮想トンネルエンドポイントからの複数の第１のパケットに基づいて、前記送信元仮想トンネルエンドポイントが、前記送信元仮想トンネルエンドポイントを前記送信先仮想トンネルエンドポイントに接続する対応する複数の中間スイッチによって提供される複数の経路に関連付けられた輻輳状態情報を学習すること、
前記送信元仮想トンネルエンドポイントが、前記送信元エンドポイントによって送信され且つ前記送信先仮想トンネルエンドポイントに関連付けられた送信先エンドポイント宛ての複数の第２のパケットを受信すること、
前記送信元仮想トンネルエンドポイントが、前記輻輳状態情報に基づいて前記複数の経路から特定の経路を選択すること、
前記送信元仮想トンネルエンドポイントが、前記特定の経路に関連付けられたタプルセットを含むヘッダ情報を有する前記複数の第２のパケットの各々をカプセル化することによってカプセル化された複数の第２のパケットを生成すること、
前記カプセル化された複数の第２のパケットが前記タプルセットに基づいた前記特定の経路を介して転送されるように、前記送信元仮想トンネルエンドポイントが、前記カプセル化された複数の第２のパケットを前記送信先エンドポイントに送信すること、を備える、非一時的なコンピュータ可読記憶媒体。
【請求項９】
前記輻輳状態情報を学習することは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットに基づいて前記特定の経路に関連付けられた輻輳フラグを決定することを含み、
前記特定の第１のパケットは、
前記特定の経路が輻輳しているかどうかを示すための前記複数の中間スイッチのうちの少なくとも１つからの輻輳通知を含む、請求項８に記載の非一時的なコンピュータ可読記憶媒体。
【請求項１０】
前記輻輳状態情報を学習することは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットの受信時刻、および前記特定の第１のパケットをトリガするための前記送信元仮想トンネルエンドポイントによって送信されたプローブパケットの送信時刻に基づいて、前記特定の経路に関連付けられたラウンドトリップタイムを決定することを含む、請求項８に記載の非一時的なコンピュータ可読記憶媒体。
【請求項１１】
前記輻輳状態情報を学習することは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットに基づいて、選択の可能性を増加または低下させるために前記特定の経路に関連付けられた重み付けを決定

することを含む、請求項８に記載の非一時的なコンピュータ可読記憶媒体。
【請求項１２】
前記輻輳状態情報を学習することは、
複数の外部送信元ポート番号と、対応する複数の経路との間のマッピングを学習することを含む、請求項８に記載の非一時的なコンピュータ可読記憶媒体。
【請求項１３】
前記複数の第２のパケットの各々をカプセル化することは、
前記複数の外部送信元ポート番号から、前記特定の経路に関連付けられた特定の外部送信元ポート番号を決定すること、
前記特定の送信元ポート番号を含むため、前記複数の第２のパケットの各々において前記ヘッダ情報のタプルセットを構成すること、を含む、請求項１２に記載の非一時的なコンピュータ可読記憶媒体。
【請求項１４】
前記送信先エンドポイント宛ての前記現在の第２のパケットと以前の第２のパケットとの間のパケット間ギャップに基づいて、現在の第２のパケットである前記複数の第２のパケットのそれぞれをフローレットに割り当てること、
前記特定の経路に関連付けられた特定の外部送信元ポート番号に前記フローレットを関連付けるフローレット情報を記憶すること、をさらに備える請求項１３に記載の非一時的なコンピュータ可読記憶媒体。
【請求項１５】
送信元仮想トンネルエンドポイント（ＶＴＥＰ）を実装して、データセンタネットワークにおいて輻輳を考慮したロードバランシングを実行するように構成されたスイッチであって、
送信先仮想トンネルエンドポイントから複数の第１のパケットを受信するための１つまたは複数の第１のポートであって、前記送信元仮想トンネルエンドポイントを前記送信先仮想トンネルエンドポイントに接続する対応する複数の中間スイッチによって提供される複数の経路を介して前記送信元仮想トンネルエンドポイントに接続される前記１つまたは複数の第１のポートと、
送信元エンドポイントによって送信され且つ前記送信先仮想トンネルエンドポイントに関連付けられた送信先エンドポイント宛ての複数の第２のパケットを受信するための第２のポートと、
スイッチロジックであって、
前記複数の第２のパケットを受信することに応答して、前記複数の第１のパケットに基づいて前記複数の経路に関連付けられた輻輳状態情報を学習すること、
前記輻輳状態情報に基づいて前記複数の経路から特定の経路を選択すること、
前記特定の経路に関連付けられたタプルセットを含むヘッダ情報を有する前記複数の第２のパケットの各々をカプセル化することによってカプセル化された複数の第２のパケットを生成すること、
前記カプセル化された複数の第２のパケットが前記タプルセットに基づいた前記特定の経路を介して転送されるように、前記カプセル化された複数の第２のパケットを前記送信先エンドポイントに送信すること、を実行するように構成された前記スイッチロジックと、を備えるスイッチ。
【請求項１６】
前記スイッチロジックは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットに基づいて前記特定の経路に関連付けられた輻輳フラグを決定することによって、前記輻輳状態情報を学習するように構成され、
前記特定の第１のパケットは、
前記特定の経路が輻輳しているかどうかを示すための前記複数の中間スイッチのうちの少なくとも１つからの輻輳通知を含む、請求項１１に記載のスイッチ。

【請求項１７】
前記スイッチロジックは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットの受信時刻、および前記特定の第１のパケットをトリガするための前記送信元仮想トンネルエンドポイントによって送信されたプローブパケットの送信時刻に基づいて、前記特定の経路に関連付けられたラウンドトリップタイムを決定することによって、前記輻輳状態情報を学習するように構成される、請求項１１に記載のスイッチ。
【請求項１８】
前記スイッチロジックは、
前記送信先仮想トンネルエンドポイントからの特定の第１のパケットに基づいて、選択の可能性を増加または低下させるために前記特定の経路に関連付けられた重み付けを決定することによって、前記輻輳状態情報を学習するように構成される、請求項１１に記載のスイッチ。
【請求項１９】
前記スイッチロジックは、
複数の外部送信元ポート番号と、対応する複数の経路との間のマッピングを学習することによって前記輻輳状態情報を学習するように構成される、請求項１１に記載のスイッチ。
【請求項２０】
前記スイッチロジックは、
前記複数の外部送信元ポート番号から、前記特定の経路に関連付けられた特定の外部送信元ポート番号を決定すること、および
前記特定の送信元ポート番号を含むため、前記複数の第２のパケットの各々において前記ヘッダ情報のタプルセットを構成することによって、前記輻輳状態情報を学習するように構成される、請求項１９に記載のスイッチ。
【請求項２１】
前記スイッチロジックは、
前記現在の第２のパケットと前記送信先エンドポイント宛ての以前の第２のパケットとの間のパケット間ギャップに基づいて、現在の第２のパケットである前記第２のパケットのそれぞれをフローレットに割り当てること、
前記特定の経路に関連付けられた特定の外部送信元ポート番号に前記フローレットを関連付けるフローレット情報を記憶すること、をさらに実行するように構成される、請求項２０に記載のスイッチ。
【発明の詳細な説明】
【背景技術】
【０００１】
本明細書で別段の説明がない限り、このセクションに記載されたアプローチは、このセクションに含めることによって先行技術であると認められるものではない。
複数のデータセンタネットワークは、概して、多数のマルチ経路によって特徴付けられるマルチルートトポロジを採用する。例えば、複数の物理サーバは、パケット転送の代替経路を提供する多数のスイッチを用いて相互に接続される。物理サーバが別の物理サーバに送信するデータを有する場合、複数の経路のうちの１つが選択されて、データが複数のパケットのフローとして送信される。実際には、トラフィックは、異なる経路間で均等に分配されないことがあり、１つの経路の過剰利用と別の経路の利用不足の原因となる。ロードバランシングは、トラフィックをできるだけ均等に分散させて輻輳を減らし且つネットワークパフォーマンスを向上させるために重要である。
【図面の簡単な説明】
【０００２】
【図１】輻輳を考慮したロードバランシングが実行されるデータセンタネットワークの一例を示す概略図。

【図２】送信元仮想トンネルエンドポイント（ＶＴＥＰ）がデータセンタネットワーク内で輻輳を考慮したロードバランシングを実行するための例示的なプロセスのフローチャート。
【図３】送信元ＶＴＥＰがデータセンタネットワーク内の輻輳状態情報を学習するための第１の例示的なプロセスのフローチャート。
【図４Ａ】送信元ＶＴＥＰがデータセンタネットワーク内の輻輳状態情報を学習するための第２の例示的なプロセスのフローチャート。
【図４Ｂ】図４Ａの例に従って学習された輻輳状態情報の一例を示す概略図。
【図５】送信元ＶＴＥＰがデータセンタネットワークにおいてデータパケット処理を実行するための例示的なプロセスのフローチャート。
【発明を実施するための形態】
【０００３】
以下の詳細な説明では、本明細書の一部を形成する添付の図面を参照する。図面において、類似の記号は、説明が別途指示しない限り、典型的には同様の構成要素を特定する。詳細な説明、図面、および特許請求の範囲に記載された例示的な実施形態は、限定を意味するものではない。本明細書に提示される主題の趣旨または範囲から逸脱することなく、他の実施形態が利用されてもよく、他の変更が行われてもよい。本明細書に全体的に記載され、図面に示された本開示の態様は、多種多様な異なる構成で配置され、置換され、結合され、構成され、これらのすべてが本明細書において明示的に企図されることは容易に理解されるだろう。
【０００４】
複数のデータセンタネットワークにおけるロードバランシング（load balancing）の課題は、輻輳を考慮したロードバランシングが実行される例示的なデータセンタネットワーク１００を示す概略図である図１を参照してより詳細に説明される。データセンタネットワーク１００は、所望の実施形態に従って、示されるコンポーネントに対する追加および／または代替のコンポーネントを含むことができることを理解されたい。
【０００５】
図１の例では、データセンタネットワーク１００は、「ＶＴＥＰ－Ａ」１１０および「ＶＴＥＰ－Ｂ」１２０等の仮想トンネルエンドポイント（virtual tunnel endpoints：VTEPs）によって提供される複数の経路を介して接続される第１のエンドポイント１０２（「ＥＰ－Ａ」参照）および第２のエンドポイント１０４（「ＥＰ－Ｂ」参照）を含む。「ＶＴＥＰ－Ａ」１１０は、「Ａ１」１３０、「Ｓ１」１４０、「Ｓ２」１５０および「Ａ２」１６０などの複数の中間スイッチによって提供される複数の経路を介して「ＶＴＥＰ－Ｂ」１２０に接続される。送信元「ＥＰ－Ａ」１０２から送信先「ＥＰ－Ｂ」１０４に複数のデータパケットを転送する場合、複数のデータパケットは、「Ａ１」１３０、「Ｓ１」１４０および「Ａ２」１６０を経由する第１の経路、および「Ａ１」１３０、「Ｓ２」１５０、および「Ａ２」１６０を経由する第２の経路のうちの一つで転送され得る。
【０００６】
実際には、用語「仮想トンネルエンドポイント」（例えば、「ＶＴＥＰ－Ａ」１１０および「ＶＴＥＰ－Ｂ」１２０）は、複数のパケット転送サービス、複数のロードバランシングサービス、複数のゲートウェイサービス等を複数のエンドポイント（例えば、「ＥＰ－Ａ」１０２および「ＥＰ－Ｂ」１０４）に提供するように構成される任意の好適な複数のネットワーク要素に言及する。ＶＴＥＰ１１０／１２０は、１つまたは複数の物理エンティティまたは仮想エンティティによって具体化されてもよい。例えば、ＶＴＥＰ１１０／１２０は、物理コンピューティングデバイス（例えば、エッジデバイス、物理サーバなど）によってサポートされる（supported）ハイパーバイザ（例えば、ハイパーバイザの仮想スイッチ）によって具体化されてもよい。ＶＴＥＰ１１０／１２０およびそれに関連するエンドポイント１０２／１０４は、同じ物理コンピューティングデバイス上に、または異なるコンピューティングデバイス上に存在してもよい。例えば、「ＥＰ－Ａ」１０２は仮想マシンであり、「ＶＴＥＰ－Ａ」１１０は同じ物理サーバによってサポートされる

仮想スイッチである。別の例では、「ＥＰ－Ａ」１０２は、第１の物理サーバによってサポートされる仮想マシンであり、「ＶＴＥＰ－Ａ」１１０は、第１の物理サーバに接続される第２の物理サーバまたは物理的トップ・オブ・ラック（top-of-rack : ToR）スイッチによってサポートされる。

【０００７】

用語「エンドポイント」（例えば、「ＥＰ－Ａ」１０２および「ＥＰ－Ｂ」１０４）は、一般的に、双方向性プロセス間通信フローの送信元ノード（「送信元エンドポイント」）または終端ノード（「送信先エンドポイント」）に言及する。実際には、エンドポイントは、物理コンピューティングデバイス（例えば、物理サーバ、物理ホスト）、物理コンピューティングデバイスによってサポートされる仮想化コンピューティングインスタンスなどであってもよい。仮想化されたコンピューティングインスタンスは、ワークロード、仮想マシン、アドレス指定可能なデータコンピューティングノード、隔離されたユーザ空間インスタンスなどを表してもよい。実際には、任意の適切な技術が、ハードウェア仮想化を含むが限定されない隔離された複数のユーザ空間インスタンスを提供するように用いられる。他の仮想化されたコンピューティングインスタンスは、（例えば、ハイパーバイザまたはＤｏｃｋｅｒなどの別個のオペレーティングシステムを必要とせずにホストオペレーティングシステムの上で動作するか、またはオペレーティングシステムレベルの仮想化として具体される）複数のコンテナ（container）、複数の仮想プライベートサーバなどを含み得る。複数の仮想マシンは、物理コンピューティングシステムのハードウェアおよびソフトウェアコンポーネントの仮想等化物を含む完全な計算環境であってもよい。用語「ハイパーバイザ」は、一般的に、Ｄｏｃｋｅｒなどの複数の名前空間コンテナをサポートするシステムレベルのソフトウェアを含む複数の仮想化されたコンピューティングインスタンスの実行をサポートするソフトウェア層またはコンポーネントに言及する。

【０００８】

用語「スイッチ」（例えば、「Ａ１」１３０、「Ｓ１」１４０、「Ｓ２」１５０および「Ａ２」１６０）は、概して複数のパケットを受信および転送するように構成された任意の適切なネットワーク要素に言及してもよく、レイヤ３ルータ、レイヤ２スイッチ、ゲートウェイ、ブリッジなどでもよい。ネットワークトポロジに応じて、スイッチはＴｏＲスイッチ、アグリゲートスイッチ（aggregate switch）、スパインスイッチ（spine switch）などでもよい。簡略化のために、図１には２つの代替的な経路が示されているが、経路の数は、相互接続されたスイッチの数およびマルチルートトポロジ（例えば、リーフスパイントポロジ（leaf-spine topology）、ファットツリートポロジ（fat-tree topology）など）等のデータセンタネットワーク１００に依存する。さらに、「ＶＴＥＰ－Ａ」１１０と「ＶＴＥＰ－Ｂ」１２０を接続する追加のスイッチが図１に示すものよりも存在してもよい。

【０００９】

用語「レイヤ２（layer-2）」は、一般的に、データリンクレイヤ（例えば、メディアアクセス制御（Media Access Control : MAC）またはイーサネットレイヤ）に言及し、用語「レイヤ３」は、ネットワークレイヤ（例えば、インターネットプロトコル（Internet Protocol : IP）レイヤ）に言及し、「レイヤ４」は、オープンシステム相互接続（Open System Interconnection : OSI）モデルにおけるトランスポート層（例えば、伝送制御プロトコル（Transmission Control Protocol : TCP）層）に言及するが、本明細書で説明される概念は、他のネットワークモデルにも適用可能であり得る。用語「パケット」は、一般的に、一緒に転送される１群のビットに言及し、「フレーム」、「メッセージ」、「セグメント」などの別の形態であってもよい。

【００１０】

「ＶＴＥＰ－Ａ」１１０と「ＶＴＥＰ－Ｂ」１２０との間の接続を提供するために、「トンネル」（簡略化のために図示せず）が、任意の適切なプロトコル（例えば、ＧＥＮＥＶＥ（Generic Network Virtualization Encapsulation）、ＳＴＴ（Stateless Transport Tunneling）またはＶＸＬＡＮ（Virtual eXtension Local Area Network））を用いて

複数のＶＴＥＰ間で確立される。用語「トンネル」は、一般的に、一対のＶＴＥＰ間のエンドツーエンド且つ双方向通信の経路に言及する。この場合、「ＥＰ－Ａ」１０２から複数のデータパケット（図１の１７０参照）を転送する前に、「ＶＴＥＰ－Ａ」１１０はカプセル化を実行してカプセル化された複数のパケットを生成する（図１の１７２参照）。

【００１１】

より詳細には、各データパケット１７０は、（図１では「Ｉ」の符号が付されている）「内部ヘッダ情報（inner header information）」およびペイロードとしてのアプリケーションデータを含む。カプセル化された後、カプセル化された各パケット１７２は、（図１では「Ｏ」の符号が付されている）外部ヘッダ情報（outer header information）およびペイロードとしてのデータパケット１７０を含む。（「外部トンネルヘッダ」としても知られている）「外部ヘッダ情報」は、外部レイヤ２ヘッダ、外部レイヤ３ヘッダ、外部レイヤ４ヘッダ等を含み得る。カプセル化が実行されて、（例えば、１３０～１６０によって形成される）ファブリック・オーバーレイ（fabric overlay）が外部トンネルヘッダに基づいて一対のＶＴＥＰ間のパケット転送を実行するだけでよい。

【００１２】

実際には、トラフィック負荷は、データセンタネットワーク１００内の異なる複数の経路間で不均一に広がり、輻輳および性能低下を引き起こす可能性がある。従来、（例えば、ホップ数が等しい）等価コストで複数の経路にわたって均一にトラフィックを分散させるために、データプレーン・ロードバランシングメカニズム（data plane load balancing mechanism）として、等価コストマルチパスルーティング（equal cost multipath routing：ECMP）が一般的に用いられている。ＥＣＭＰスイッチは、単純なハッシュベースのロードバランシングスキーム（hash-based load balancing scheme）を用いて、新しい各トラフィックのフローを利用可能な複数の経路のうちの１つにランダムに割り当てる。ＥＣＭＰは、通常、ロードバランシングスキームを更新するために、柔軟性に欠けるカスタムシリコン（例えば、特定用途向け集積回路（ＡＳＩＣ））に実装される。さらに、ＥＣＭＰは輻輳に依存せず、性能低下の原因となる複数の経路のオーバーサブスクリプション（oversubscription）を防止しない。

【００１３】

例えば、図１では、「ＥＰ－Ａ」１０２から「ＥＰ－Ｂ」１０４へ転送する複数のパケットのキュー占有レベル（queue occupancy levels）（複数の括弧内の１８０～１８６を参照）を用いて示されるように、異なる対のスイッチを接続する複数のリンクは異なる輻輳レベルを有する。「Ｓ１」１４０を介した第１の経路に沿ったキュー占有レベルは４０％（１８０参照）および５０％（１８２参照）である。「Ｓ２」１５０を介した第２の経路に沿ったキュー占有レベルは３０％（１８４参照）および８０％（１８６参照）である。ＥＣＭＰは異なる輻輳レベルを考慮しないため、長時間実行されるフローは、８０％（１８６参照）のキュー占有レベルを有する輻輳が発生する「Ｓ２」１５０を介した第２の経路に割り当てられる可能性がある。

【００１４】

従来、制御プレーン・ロードバランシングメカニズムが、ＥＣＭＰの欠点に対処するために用いられている。この場合、中央制御装置がデータセンタネットワーク１００に配置されて、「Ａ１」１３０、「Ｓ１」１４０、「Ｓ２」１５０および「Ａ２」１６０から統計を収集し且つ転送ルールを「Ａ１」１３０、「Ｓ１」１４０、「Ｓ２」１５０および「Ａ２」１６０にプッシュして（push）、制御プレーン・ロードバランシングを実現する。しかしながら、中央制御装置が必要とされるので、制御プレーンメカニズムは、制御ループの待ち時間（latency）が長いために比較的遅く且つ高い揮発性のトラフィック（volatile traffic）を扱うことができない。

【００１５】

従来、ホストベース（host-based）のアプローチが、ＥＣＭＰの欠点に対処するために用いられている。例えば、マルチパスＴＣＰ（multipath TCP：MPTCP）と呼称される修正されたバージョンの伝送制御プロトコル（ＴＣＰ）が、複数のエンドポイント間に複数

のサブフローを確立して、異なる複数の経路でトラフィックを分割するように用いられる。しかしながら、ホストベースのアプローチは、通常、ＭＰＴＣＰの場合に「ＥＰ－Ａ」１０２および「ＥＰ－Ｂ」１０４のＴＣＰ／ＩＰスタックを変更するなど、すべてのエンドポイントへの変更を必要とする。そのような変更は、特に、「ＥＰ－Ａ」１０２および「ＥＰ－Ｂ」１０４が異なるオペレーティングシステムを実行している場合、または異なるエンティティによって制御されている場合には、特に困難である（場合によっては不可能である）。

【００１６】

輻輳を考慮したロードバランシング（Congestion-aware load balancing）

本開示の例によれば、輻輳を考慮したロードバランシングのアプローチは、関連する「ＥＰ－Ａ」１０２に気付かれない方法で「ＶＴＥＰ－Ａ」１１０によって実施され得る。上述した従来のアプローチとは異なり、本開示の例は、ＭＰＴＣＰを実施するための「ＥＰ－Ａ」１０２の変更、または新しいロードバランシングスキームを実施するための中間スイッチ１３０～１６０の変更を必要とせずに具体化され得る。さらに、制御プレーンロードバランシングメカニズムとは異なり、中央制御装置を設けて輻輳監視を実行し且つ中間スイッチ１３０～１６０に転送ルールをプッシュする必要がない。

【００１７】

より詳細には、図２は、送信元ＶＴＥＰ１１０がデータセンタネットワーク１００内で輻輳を考慮したロードバランシングを実行するための例示的なプロセス２００のフローチャートである。例示的なプロセス２００は、２０５～２４０などの１つまたは複数のブロックによって例示される１つ以上の工程、機能、動作を含むことができる。様々なブロックは、より少数のブロックと組み合わされ、追加のブロックに分割され、および／または所望の実施形態に応じて除去され得る。

【００１８】

以下、「ＶＴＥＰ－Ａ」１１０が送信元ＶＴＥＰの例として用いられ、「ＶＴＥＰ－Ｂ」１２０が送信先ＶＴＥＰの例として用いられ、「Ｓ１」１４０および「Ｓ２」１５０が中間スイッチの例として用いられ、「ＥＰ－Ａ」１０２が「送信元エンドポイント」として用いられ、「ＥＰ－Ｂ」１０４が「送信先エンドポイント」として用いられる。キュー占有レベルは図１のデータセンタネットワーク１００における輻輳を示す一例として用いるが、リンク利用レベル、ラウンドトリップタイム（round trip time：RTT）などの輻輳の他の適切なインジケータが用いられてもよい。

【００１９】

図２の２０５において、「ＶＴＥＰ－Ａ」１１０は、「ＶＴＥＰ－Ａ」１１０と「ＶＴＥＰ－Ｂ」１２０とを接続する中間スイッチ１３０～１６０によって提供される複数の経路に関連する輻輳状態情報（図１の１９０参照）を学習する（learn）。輻輳状態情報は、「ＶＴＥＰ－Ｂ」１２０からの送信先から送信元への（destination-to-source）フィードバック情報（図１の１８８参照）を示す第１のパケットに基づいて学習され得る。図２の２１０において、「ＶＴＥＰ－Ａ」１１０は、「ＥＰ－Ａ」１０２によって送信され且つ「ＥＰ－Ｂ」１０４宛ての複数の第２の（データ）パケット１７０を受信する。例えば、複数の第２のパケット１７０は、「ＥＰ－Ａ」１０２上で動作するアプリケーションから「ＥＰ－Ｂ」１０４上で動作する別のアプリケーションまでのアプリケーションデータを含み得る。各第２のパケット１７０は、概して、「ＥＰ－Ａ」１０２と「ＥＰ－Ｂ」１０４との間のプロセス間通信に関連する（図１では、「Ｉ」の符号が付された）内部ヘッダ情報を含む。

【００２０】

図２の２２０において、「ＶＴＥＰ－Ａ」１１０は、複数の経路から特定の経路（「選択経路」とも呼称される）を選択する。例えば、図１では、「ＶＴＥＰ－Ａ」１１０は、複数の経路に関連する輻輳状態情報に基づいて、「Ａ１」１３０、「Ｓ１」１４０および「Ａ２」１６０を介した第１の経路を選択し得る。図２の２３０において、「ＶＴＥＰ－Ａ」１１０は、２２０で選択された経路に関連するタプルセット（a set of tuples）を

含む（外部）ヘッダ情報を有する各第２のパケット１７０をカプセル化することによってカプセル化された複数の第２のパケット１７２を生成する。図２の２４０において、「ＶＴＥＰ－Ａ」１１０は、カプセル化された第２のパケット１７２を送信先「ＥＰ－Ｂ」１０４に送信して、カプセル化された第２のパケット１７２は、タプルセットに基づいて選択された経路を介して転送される。

【００２１】

図３を用いてさらに説明されるように、「ＶＴＥＰ－Ａ」１１０は、異なる外部送信元ポート番号（ｓｏｕｒｃｅ＿ＰＮ１９２を参照）を対応する経路（ｐａｔｈ＿ＩＤ１９４）および輻輳を示すフラグ（ｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ１９６を参照）に関連付ける輻輳状態情報（図１の１９０参照）に依存し得る。例えば、パケット転送を実行する前に、「ＶＴＥＰ－Ａ」１１０は、ｓｏｕｒｃｅ＿ＰＮ１９２とｐａｔｈ＿ＩＤ１９４との間のマッピングを学習するために、経路学習（path learning）を実行し得る。「Ｓ１」１４０を介した第１の経路については、ｓｏｕｒｃｅ＿ＰＮ＝ＳＰ１およびｐａｔｈ＿ＩＤ＝Ｐ１である。「Ｓ２」１５０を介した第２の経路については、ｓｏｕｒｃｅ＿ＰＮ＝ＳＰ２およびｐａｔｈ＿ＩＤ＝Ｐ２である。

【００２２】

図３を用いてさらに説明されるように、「ＶＴＥＰ－Ａ」１１０は、ｓｏｕｒｃｅ＿ＰＮ１９２とｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ１９６の異なる対の間のマッピングを学習してもよい。一例では、「ＶＴＥＰ－Ａ」１１０は、データセンタネットワーク１００内の輻輳状態情報のエンドツーエンド通知を可能にする明示的輻輳通知（Explicit Congestion Notification：ECN）などの既存の機能中間スイッチ（existing capabilities intermediate switches）１３０～１６０に依存し得る。この場合、パケットをドロップする代わりに、中間スイッチ１３０～１６０は、特定の経路に関連する現在のまたは目下の輻輳を「ＶＴＥＰ－Ｂ」１２０に通知するために、輻輳通知のフォームとしてパケットマーキングを受信する。その後、「ＶＴＥＰ－Ｂ」１２０は、「ＶＴＥＰ－Ａ」１１０（図１の１８８参照）に輻輳通知を報知する。ＥＣＮの他に、他の適切なアプローチが、輻輳状態情報を学習するために用いられてもよい。例えば、図４Ａおよび図４Ｂを用いてさらに説明されるように、「ＶＴＥＰ－Ａ」１１０は、「ＶＴＥＰ－Ａ」１１０と「ＶＴＥＰ－Ｂ」１２０との間のＲＴＴを測定することができる。

【００２３】

図１の例では、カプセル化された各第２のパケット１７２は、選択された経路に関連するタプルセットを含む外部ヘッダ情報を含む。具体的には、カプセル化された各第２のパケット１７２は、外部レイヤ２ヘッダ、外部レイヤ３ヘッダ、および外部レイヤ４ヘッダなどの（「Ｏ」の符号が付けされている）外部ヘッダ情報を含む。タプルセットは、送信元ポート番号、送信先ポート番号、送信元ＩＰアドレス、送信先ＩＰアドレス、およびプロトコルを含み得る。「Ｓ１」１４０を介した第１の経路（すなわち、ｐａｔｈ＿ＩＤ＝Ｐ１）について、外部レイヤ４ヘッダは、ｓｏｕｒｃｅ＿ＰＮ＝ＳＰ１の値を有する送信元ポート番号を含み得る。送信元ポート番号は、選択された経路に沿って「Ａ１」１３０、「Ｓ１」１４０および「Ａ２」１６０を介してカプセル化された複数のパケット１７２が転送されるように設定される。送信先「ＶＴＥＰ－Ｂ」１２０において、カプセル化の解除が実行されて外部ヘッダ情報を除去し、データパケット１７４が「ＥＰ－Ｂ」１０４に送信される。

【００２４】

例示的なプロセス２００を用いて、「ＶＴＥＰ－Ａ」１１０は、異なる複数の経路に関連付けられた輻輳状態情報１９０を考慮して、データセンタネットワーク１００内の異なる複数の経路上に仮想ネットワークトラフィックを分配することができる。「ＶＴＥＰ－Ａ」１１０は、異なる経路上にカプセル化された第２のパケット１７２を送信したいときはいつでも、異なる外部送信元ポート番号を選択することができる。「ＶＴＥＰ－Ａ」１１０と「ＶＴＥＰ－Ｂ」１２０とを接続する中間スイッチ１３０～１６０は外部ヘッダ情報に基づいてロードバランシングを実行するので、外部レイヤ４ヘッダ内の外部送信元ポ

ート番号はエントロピーとして用いられ、データセンタネットワーク１００内の複数の経路（例えば、等コストパス（equal-cost paths））を利用する。

【００２５】

輻輳状態情報

図３は、第１の例によるデータセンタネットワーク１００内の輻輳状態情報１９０を学習するための送信元ＶＴＥＰ１１０の例示的な第１のプロセス３００のフローチャートである。例示的なプロセス３００は、３１０～３６５などの１つまたは複数のブロックによって示される１つまたは複数の工程、機能、または動作を含み得る。様々なブロックは、より少数のブロックに組み合わされ、追加のブロックに分割され、および／または所望の実施形態に応じて除去され得る。図１の例を用いて、例示的なプロセス３００は、「ＶＴＥＰ－Ａ」１１０によって実行され得る。

【００２６】

図３の３１０において、「ＶＴＥＰ－Ａ」１１０は、ｓｏｕｒｃｅ＿ＰＮ１９２およびｐａｔｈ＿ＩＤ１９４の複数の対の間のマッピングまたは関連付けを学習する。例えば、図１において、ｓｏｕｒｃｅ＿ＰＮ１９２の異なる値は、データセンタネットワーク１００内の異なる重なり合わない複数のＥＣＭＰ経路を導く可能性がある。マッピングは、「ＶＴＥＰ－Ａ」１１０がパケット転送のために選択することができる異なる複数の経路の先験的知識を示す。ＥＣＭＰハッシング（ECMP hashing）がｓｏｕｒｃｅ＿ＰＮ１９２の特定の値を含むタプルセットに適用される場合、結果として関連する経路がｐａｔｈ＿ＩＤ１９４によって識別される。

【００２７】

実際には、「ＶＴＥＰ－Ａ」１１０は、各経路で検出されたすべてのインタフェースＩＰに関する「トレースルート（traceroute）」スタイル情報を収集するために、（例えば、Paris tracerouteの後にモデル化される）バックグラウンドデーモンを実施して、データセンタネットワーク１００内の他のすべてのＶＴＥＰに定期的なプローブパケットを送信する。「ＶＴＥＰ－Ａ」１１０は、各プローブパケットの外部ヘッダ情報内の外部送信元ポート番号を変更して（rotate）、各ポート番号に対する経路トレースを収集することができる。次に、「ＶＴＥＰ－Ａ」１１０は、対応するプローブ経路トレースがこれまで収集されたトレースと異なるたびに、輻輳状態情報１９０を更新してｓｏｕｒｃｅ＿ＰＮ１９２を追加または更新する。

【００２８】

図３のブロック３１５～３６０は、ｐａｔｈ＿ＩＤ１９４によって識別される経路と、その経路に沿った輻輳を示すｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ１９６とを関連付ける輻輳状態情報１９０を学習する「ＶＴＥＰ－Ａ」１１０に関する。特に、図３の３１５および３２０において、「ＶＴＥＰ－Ａ」１１０は、ｓｏｕｒｃｅ＿ＰＮ１９２の特定の値を含む外部トンネルヘッダを有する複数のパケットをカプセル化し、カプセル化された複数のパケットを送信する。

【００２９】

図３の例では、「ＶＴＥＰ－Ａ」１１０は、現在のまたは目下の輻輳などを示すためのＥＣＮマーキングなどの既存のスイッチの輻輳通知機能に依存することができる。ＥＣＮの詳細情報は、コメント番号３１６８のインターネットエンジニアリングタスクフォース（Internet Engineering Task Force：IETF）で見つけられ、「ＩＰへの明示的輻輳通知（ＥＣＮ）の追加」と題されており、これは参照によりその全体が本明細書に組み込まれる。例としてＥＣＮが説明されているが、任意の他の適切なパケットマーキングのアプローチを使用できることを理解されたい。

【００３０】

図３の３２５，３３０，３３５及び３４０において、「ＶＴＥＰ－Ａ」１１０からカプセル化された複数のパケットを受信することに応答して、スイッチ（例えば、「Ｓ１」１４０、「Ｓ２」１５０）は、ＥＣＮマーキングを実行して、カプセル化された複数のパケットを転送する前にこのスイッチにおける輻輳にフラグを立てる。この場合、スイッチは

、輻輳通知のフォームとしてカプセル化された複数のパケットのヘッダ情報（例えば、ＴＣＰヘッダの予約されたフィールド）を変更することができるＥＣＮ対応スイッチ（ECN-enabled switch）として知られている。

【００３１】

例えば、図１において、「Ｓ２」１５０におけるキュー占有レベル（例えば、８０％）が所定の閾値（例えば、T_Q＝６０％）を超えるとき、「Ｓ２」１５０は、「ＶＴＥＰ－Ａ」１１０から受信された複数のパケットをマークキングして、「ＶＴＥＰ－Ｂ」１２０に輻輳を通知して、第２の経路が輻輳した経路であることを示す。一方、関連するキュー占有レベル（例えば、５０％）が閾値未満であるため、「Ｓ１」１４０が任意のパケットマーキングを実行する必要はなく、それによって第１の経路が非輻輳の経路であることを示す。

【００３２】

図３の３４５および３５０において、送信先「ＶＴＥＰ－Ｂ」１２０は、カプセル化されたパケットを受信し、外部ヘッダ情報内の外部送信元ポート番号と送信元「ＶＴＥＰ－Ａ」１１０（図１の１８８も参照）への任意の輻輳通知（例えば、ＥＣＮマーキング）との間のマッピングを通知する。図３の３５５および３６０において、「ＶＴＥＰ－Ａ」１１０は、ｓｏｕｒｃｅ＿ＰＮ１９２と、関連するｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ１９６（すなわち、フラグ情報）との間のマッピングを更新する。例えば、図１において、「ＶＴＥＰ－Ａ」１１０は、「Ｓ１」１４０を介した第１のパスのｓｏｕｒｃｅ＿ＰＮ＝ＳＰ１についてはｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ＝ｆａｌｓｅであることを判定し、「Ｓ２」１５０を介した第２のパスのｓｏｕｒｃｅ＿ＰＮ＝ＳＰ１についてはｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ＝ｔｒｕｅであると判定する。

【００３３】

ｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇは、関連する経路の選択の可能性に影響を与える重み付けアルゴリズムにおいて使用されてもよい。ロードバランシングプロセスの開始時に、「ＶＴＥＰ－Ａ」１１０は、図３の３１０で経路学習機能によって発見されたすべての等コスト経路について等しい重み付けで開始することができる。続いて、図３の３６５で、輻輳フラグに基づいて経路の重み付け（図１の重み付け１９８参照）が調整される。例えば、ｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ＝ｆａｌｓｅ（すなわち、クリア）であるとの判定に応答して、「Ｓ１」１４０を介した第１の経路に関連するｗｅｉｇｈｔ＝ｗ１は、選択の可能性を増大するために増大される。一方、ｃｏｎｇｅｓｔｉｏｎ＿ｆｌａｇ＝ｔｒｕｅ（すなわち、セット）であるとの判定に応答して、「Ｓ２」１５０を介した第２の経路に関連するｗｅｉｇｈｔ＝ｗ２は、選択の可能性を低減するために低減される。

【００３４】

輻輳を示すために使用され得る別のメトリックは、各経路について測定され且つアクティブに追跡され得るＲＴＴである。例は、送信元ＶＴＥＰ１１０がデータセンタネットワーク１００内の輻輳状態情報を学習する第２の例示的なプロセス４００のフローチャートである図４Ａを用いて説明される。プロセス４００の例は、４１０～４５５などのより多くのブロックにより例示される１つ以上の工程、機能、または動作を含み得る。様々なブロックは、より少数のブロックに組み合わされ、複数の追加のブロックに分割され、および／または所望の実施形態に応じて除去され得る。

【００３５】

図４Ａの４１０，４１５および４２０において、「ＶＴＥＰ－Ａ」１１０は、関連するスイッチ１４０／１５０を介した各経路上に複数のプローブパケットを周期的に送信し、各プローブパケットはｓｏｕｒｃｅ＿ＰＮおよび送信（Ｔｘ）タイムスタンプを識別する。図４Ａの４２５，４３０および４３５において、「ＶＴＥＰ－Ｂ」１２０は、スイッチ１４０／１５０を介してｓｏｕｒｃｅ＿ＰＮおよびＴｘタイムスタンプも識別する確認応答（ＡＣＫ）パケットを送信することによって、各プローブパケットに応答する。図４Ａの４４０および４４５で、特定の受信時刻にＡＣＫパケットを受信したことに応答して、「ＶＴＥＰ－Ａ」１１０は、受信時刻とＴｘタイムスタンプとの間の差に基づいてＲＴＴ

を決定することができる。図４Ａの４５０および４５５において、「ＶＴＥＰ－Ａ」１１０は、ＲＴＴをｓｏｕｒｃｅ＿ＰＮに関連付け、経路に関連する重み付けを調整する。

【００３６】

図４Ｂは、図４Ａの例を用いて学習された例示的な輻輳状態情報４６０を示す概略図である。例えば、「Ｓ１」１４０を介した第１の経路は、ｓｏｕｒｃｅ＿ＰＮ＝ＳＰ１（４６２参照）、ｐａｔｈ＿ＩＤ＝Ｐ１（４６４参照）、ＲＴＴ＝Ｒ１（４６６参照）およびｗｅｉｇｈｔ＝ｗ１（４６８参照）に関連付けられる。「Ｓ２」１５０を介した第２の経路は、ｓｏｕｒｃｅ＿ＰＮ＝ＳＰ２、ｐａｔｈ＿ＩＤ＝Ｐ２、ＲＴＴ＝Ｒ２およびｗｅｉｇｈｔ＝ｗ２に関連付けられる。図１の例では、第２の経路が、第１の経路よりも高い輻輳レベルを有し、Ｒ２はＲ１よりも大きくすべきである。この場合、ｗ２はｗ１よりも小さく調整されて、第２経路の選択の可能性が低減される。実際には、タイムスタンプと肯定応答（acknowledgement）が行われることが近いほど、トランスミッタとレシーバのソフトウェアスタックによって導入されるレイテンシを含まないため、ＲＴＴは経路の実際のネットワークレイテンシをより正確に反映する。

【００３７】

データパケット処理

本開示の例によれば、ロードバランシングは、ＴＣＰなどのトランスポート層プロトコルに関連するパケット並べ替え問題を回避または改善するために、複数のフローレットの粒度（granularity of flowlets）で実行され得る。これは、複数のパケットのフローを「フローレット（flowlet）」と呼称される複数のより小さいグループに分割することによって達成され得る。本明細書で使用される用語「フローレット」は、概して、フロー内のパケットのグループまたはバースト（burst）に言及する。

【００３８】

図５は、仮想トンネルエンドポイント１１０がデータセンタネットワーク１００においてデータパケット処理を実行する例示的なプロセス５００のフローチャートである。例示的なプロセス５００は、５１０～５６０等の１つまたは複数のブロックによって示される１つまたは複数の工程、機能、動作を含み得る。種々のブロックは、より少数のブロックに組み合わされ、追加のブロックに分割され、および／または所望の実施形態に応じて除去され得る。

【００３９】

５１０および５１５において、「ＥＰ－Ａ」１０２からのデータパケット１７０を受信したことに応答して、「ＶＴＥＰ－Ａ」１１０は、データパケット１７０が新しいフローレットまたは現在のフローレットに属するかどうかを判定する。例えば、同じフロー内の２つの連続するパケットの到着の時間間隔（すなわち、パケット間ギャップ（inter-packet））が所定の閾値（例えば、$T_{flowlet}$秒；５１５参照）を超えるたびに、新しいフローレットが検出され得る。しきい値を超えないすべての後続パケットは、同じフローレットの一部であるとみなされる。２つのシナリオが以下で説明される。

【００４０】

（ａ）新しいフローレットが検出されたとき（すなわち、パケット間ギャップ＞$T_{flowlet}$またはフローの第１パケットが検出されたとき）、「ＶＴＥＰ－Ａ」１１０は、図５の５２０においてデータパケット１７０にｆｌｏｗｌｅｔ＿ＩＤ（例えば、「Ｆ１」）を割り当てる。図５の５２５において、その時点の輻輳状態情報１９０／４６０に基づいて、新しいフローレット（５２５参照）に対して、ｐａｔｈ＿ＩＤ（例えば、「Ｐ１」）によって識別される経路が選択される。図５の５３０において、選択された経路に関連するｓｏｕｒｃｅ＿ＰＮが決定され、選択された経路に関連するｆｌｏｗｌｅｔ＿ＩＤとｓｏｕｒｃｅ＿ＰＮ（例えば、「ＳＰ１」）との間の関連性が記憶される。「ＶＴＥＰ－Ａ」１１０は、フローレットの最新のデータパケット１７０が受信された時刻を記録するためのｆｌｏｗｌｅｔ＿ｔｉｍｅも記憶する。

【００４１】

（ｂ）既存のフローレットのデータパケット１７０が検出されたとき（すなわち、パケ

ット間ギャップ＝現在の時刻－ｆｌｏｗｌｅｔ＿ｔｉｍｅ≦Ｔｆｌｏｗｌｅｔ）、「ＶＴＥＰ－Ａ」１１０は、図５の５５０および５５５において、現在のｆｌｏｗｌｅｔ＿ＩＤ（たとえば「Ｆ１」）および関連するｓｏｕｒｃｅ＿ＰＮ（たとえば「ＳＰ１」）を取得する。同様に、図５の５６０において、「ＶＴＥＰ－Ａ」１１０は、フローレットの最新のデータパケット１７０が受信された時刻を記録するためのｆｌｏｗｌｅｔ＿ｔｉｍｅも記憶する。

【００４２】

（ａ）および（ｂ）の両方の場合、例示的なプロセス５００は、図５の５３５および５４０に続き、「ＶＴＥＰ－Ａ」１１０は、ｆｌｏｗｌｅｔ＿ＩＤ（例えば、「Ｆ１」）に関連付けられたｓｏｕｒｃｅ＿ＰＮ（例えば「ＳＰ１」）の値を有する外部送信元ポート番号を含むように構成された外部トンネルヘッダを有するデータパケット１７０をカプセル化し、カプセル化されたデータパケット１７２を送信する。外部トンネルヘッダはまた、「ＶＴＥＰ－Ａ」１１０に関連する送信元ＩＰアドレスと、「ＶＴＥＰ－Ｂ」１２０に関連する送信先ＩＰアドレスを含む。送信先では、パケットが「ＥＰ－Ｂ」１０４に転送される前に、外部トンネルヘッダが「ＶＴＥＰ－Ｂ」１２０によって除去される。

【００４３】

図５の５１５でパケット間ギャップを超過したとき、ｆｌｏｗｌｅｔ＿ＩＤ＝「Ｆ２」などの別のフローレットに属する後続のデータパケット１７０に対して上記事項を繰り返すことができることを理解されたい。この場合、輻輳状態情報１９０／４６０が変更され、「Ｓ１」１４０を介した第１の経路の代わりに「Ｓ２」１５０を介した第２の経路（すなわち、ｐａｔｈ＿ＩＤ＝Ｐ２）が選択され得る。この場合、関連するｓｏｕｒｃｅ＿ＰＮ＝ＳＰ２は、カプセル化されたデータパケット１７２が第２の経路に沿って「Ａ１」１３０、「Ｓ２」１５０および「Ａ２」１６０によって「ＶＴＥＰ－Ｂ」１２０に転送されるような外部ヘッダ情報に含まれる。

【００４４】

本開示の例を用いて、「ＶＴＥＰ－Ａ」１１０は、新しいフローレットのｆｌｏｗｌｅｔ＿ＩＤを記憶し、同じフローレットの複数のパケットのために関連するｓｏｕｒｃｅ＿ＰＮを再利用する。実際には、閾値Ｔｆｌｏｗｌｅｔは、ネットワークにおいて（例えば、ＲＴＴのオーダーにおいて）推定ＲＴＴに基づいて設定されてもよい。複数のフローレット間のパケット間ギャップが大きいことは、（技術的には同じフローの一部である）複数のフローレットが異なる経路を用いる場合、パケットの並べ替えを最小限に抑えることを保証する。さらに、「ＶＴＥＰ－Ａ」１１０は、選択された経路を用いるようにするために、新しい各フローレットに対して最小の輻輳を有する外部送信元ポート番号を割り当てることができる。これにより、すべてのアクティブな経路のキューが常に低く保たれるため、ワークロードのスループットおよびレイテンシが向上する。

【００４５】

本開示の複数の例は、（例えば、「ＥＰ－Ａ」１０２および「ＥＰ－Ｂ」１０４として動作する）複数のゲスト仮想マシンに気付かれない方法でネットワークフローを複数の非輻輳の経路に分割するように「VTEP-A」110を構成することによって、エンドユーザの複数のゲスト仮想マシンに気付かれないように実施される。一例では、（例えば、「ＶＴＥＰ－Ａ」１１０として動作する）送信元仮想スイッチと、（例えば、「ＶＴＥＰ－Ｂ１２０」として動作する）送信先仮想スイッチの両方が用いられ得る。この場合、フローは任意に複数のフローレットに分割され、複数のフローレット間のアイドルギャップ（idle gap）に左右されない。

【００４６】

例えば、各ＴＣＰセグメンテーションオフロード（TCP segmentation offload : TSO）セグメントは、フローレットとして扱うことができる。用いられる複数のフローが異なるために送信先仮想スイッチで複数のフローレットが順番に関係なく（out of order）到着した場合、送信先仮想スイッチは、それらを送信先エンドポイント（例えば、ゲスト仮想マシン）に配信する前に、複数のフローレットを並べ替えることができる。これにより、

送信先仮想スイッチは、送信先エンドポイントの送信先プロトコルスタック（例えば、ＴＣＰ／ＩＰスタック）から順番に関係なく到達したことを隠して、送信先プロトコルスタックが、送信元エンドポイントでの送信元プロトコルスタックの遅延を防止することができる。

【００４７】

本開示の例は、ハイパーバイザの仮想スイッチで具体化され、トラフィックの第１のエントリポイントからインテリジェントな経路選択に導かれる。たとえば、仮想スイッチ以降の複数のＥＣＭＰ経路が存在する可能性があり、各経路は、異なる物理ＮＩＣを用いる。この場合、本開示の例は、プロセッサを含むホストと、一組の命令を記憶する非一時的なコンピュータ可読記憶媒体とを含むホストを用いて具体化され得る。プロセッサによる実行に応答して、１組の命令は、図１～図５の例にしたがって、プロセッサに、データセンタネットワーク１００における輻輳認識ロードバランシングを実行するために、送信元ＶＴＥＰ（例えば、ホストによってサポートされる仮想スイッチにおける「ＶＴＥＰ－Ａ」１１０）を具体化させる。

【００４８】

一例では、ホストによって実装された送信元ＶＴＥＰは、送信先ＶＴＥＰからの複数の第１のパケットに基づいて、送信元ＶＴＥＰを送信先ＶＴＥＰに接続する対応する複数の中間スイッチによって提供される複数の経路に関連する輻輳状態情報を学習することができる。また、送信元エンドポイントによって送信され、送信先ＶＴＥＰに関連付けられた送信先エンドポイント宛ての複数の第２のパケットを受信することに応答して、送信元ＶＴＥＰは、輻輳状態情報に基づいて複数の経路から特定の経路を選択し、複数の第２のパケットの各々を、前記特定の経路に関連する１組のタプルを含むヘッダ情報でカプセル化することによってカプセル化された複数の第２のパケットを生成し、カプセル化された複数の第２のパケットが１組のタプルに基づいて特定の経路を介して転送されるように、カプセル化された複数の第２のパケットを送信先エンドポイントに送信する。

【００４９】

すべての物理ＮＩＣが同じレイヤ３に接続するシナリオでは、ネクストホップ（next-hop）および経路ダイバーシティ（path diversity）が第１ホップスイッチ（例えば、ＴｏＲスイッチ）を超えて開始する場合、本開示の例は、仮想スイッチソフトウェアよりも早いスピードでＮＩＣドライバ／ハードウェアまたは第１ホップスイッチにおいて具体化され得る。高度なスイッチアーキテクチャを必要とする従来のアプローチと比較して、本開示の例は、エッジハイパーバイザ（edge hypervisor）において（例えば、完全にソフトウェアで）実行され、複数の送信元と複数の送信先との間の任意の数のホップにスケーリングされ得る。この場合、図１から図５の例に従ってデータセンタネットワーク１００内で輻輳認識ロードバランシングを実行する送信元ＶＴＥＰ（例えば、「ＶＴＥＰ－Ａ」１１０）を具体化するために、例示的な（第１ホップ）スイッチが用いられる。スイッチは、ハードウェアロジック（例えば、ハードウェア回路）、プログラマブルロジック、またはそれらの組み合わせなど、任意の適切なスイッチロジックを含み得る。

【００５０】

一例では、スイッチは、１つまたは複数の第１のポート、１つまたは複数の第２のポート、およびスイッチロジックを含み得る。１つまたは複数の第１のポートは、送信元ＶＴＥＰを送信先ＶＴＥＰに接続する対応する複数の中間スイッチによって提供される複数の経路を介して送信元ＶＴＥＰに接続された送信先ＶＴＥＰから複数の第１パケットを受信するように用いられてもよい。第２のポートは、送信元エンドポイントによって送信され、送信先ＶＴＥＰに関連付けられた送信先エンドポイント宛である第２のパケットを受信するように用いられてもよい。スイッチロジックは、複数の第１のパケットに基づいて複数の経路に関連付けられた輻輳状態情報を学習するように構成され得る。複数の第２のパケットを受信したことに応答して、スイッチロジックは、輻輳状態情報に基づいて複数の経路から特定の経路を選択し、特定の経路に関連付けられたタプルセットを含むヘッダ情報を有する複数の第２のパケットの各々をカプセル化することによってカプセル化された

複数の第２のパケットを生成し、カプセル化された複数の第２のパケットがタプルセットに基づいて特定の経路を介して転送されるように、カプセル化された複数の第２のパケットを送信先エンドポイントに送信するように構成され得る。

【００５１】

上記で紹介した技術は、専用ハードワイヤード回路、プログラマブル回路と組み合わせたソフトウェアおよび／またはファームウェア、またはそれらの組み合わせで具体化され得る。専用ハードワイヤード回路は、例えば、１つまたは複数の特定用途向け集積回路（ＡＳＩＣ）、プログラマブルロジックデバイス（ＰＬＤ）、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、プログラマブルスイッチアーキテクチャなどの形態であってもよい。用語「プロセッサ」は、処理ユニット、ＡＳＩＣ、論理ユニット、またはプログラマブルゲートアレイなどを含むように広く解釈されるべきである。

【００５２】

前述の詳細な説明は、ブロック図、フローチャート、および／または実施例の使用を介して、デバイスおよび／またはプロセスの様々な実施形態を示している。そのようなブロック図、フローチャート、および／または実施例が１つまたは複数の機能および／または動作を含む限り、そのようなブロック図、フローチャート、または実施例における各機能および／または動作が、ハードウェア、ソフトウェア、ファームウェア、またはそれらの任意の組み合わせの広い範囲によって個別におよび／または集合的に具体化可能であることは、当業者によって理解されよう。

【００５３】

当業者であれば、本明細書に開示された実施形態のいくつかの態様は、全体的または部分的に、集積回路内で、１つまたは複数のコンピュータ上で実行される１つまたは複数のコンピュータプログラムとして（例えば、１つまたは複数のコンピュータシステム上で実行される１つまたは複数のプログラムとして）、１つまたは複数のプロセッサ上で実行される１つまたは複数のプログラムとして（例えば、１つまたは複数のマイクロプロセッサ上で実行される１つまたは複数のプログラムとして）、ファームウェアとして、またはそれらの仮想的な任意の組み合わせとして、透過的に具体化され得ること、および回路の構成および／またはソフトウェアおよび／またはファームウェア用のコードの記述はこの開示に照らして当業者の技術範囲内であることを当業者によって認識されよう。

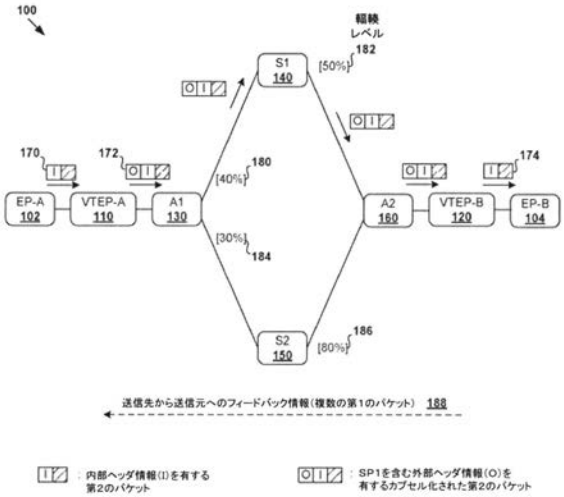
【００５４】

明細書に紹介されたソフトウェアおよび／または技術を具体化することは、非一時的なコンピュータ可読記憶媒体に記憶されてもよく、１つまたは複数の汎用または専用のプログラム可能なマイクロプロセッサによって実行されてもよい。「コンピュータ可読記憶媒体」は、本明細書で使用されるように、機械（例えば、コンピュータ、ネットワーク装置、パーソナルデジタルアシスタント（ＰＤＡ）、モバイルデバイス、製造ツール、１つまたは複数のプロセッサのセットを有する任意のデバイスなど）を含み得る。コンピュータ可読記憶媒体は、記録可能／記録不可能媒体（例えば、読み出し専用メモリ（ＲＯＭ）、ランダムアクセスメモリ（ＲＡＭ）、磁気ディスクまたは光記憶媒体、フラッシュメモリデバイスなど）を含み得る。

【００５５】

図面は、例示にすぎず、図面に示される要素または手順は、本開示を実施するために必ずしも必須ではない。当業者であれば、実施例中の要素は、記載された実施例の装置に配置されるか、あるいは、実施例の装置とは異なる１つ以上の装置に代替的に配置されることを理解するであろう。記載された実施例における要素は、１つのモジュールに組み合わされるか、またはさらに複数のサブ要素に分割され得る。

【図１】

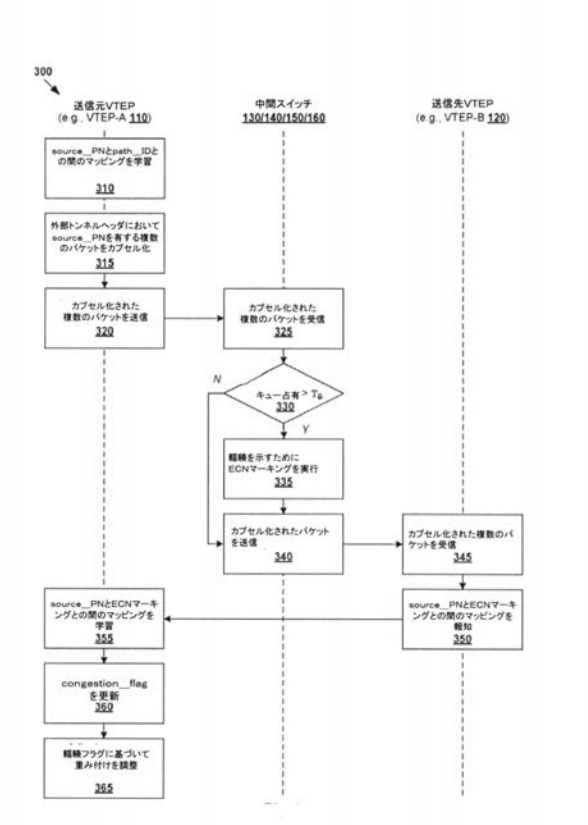


VTEP－Aにおける輻輳状態情報 190

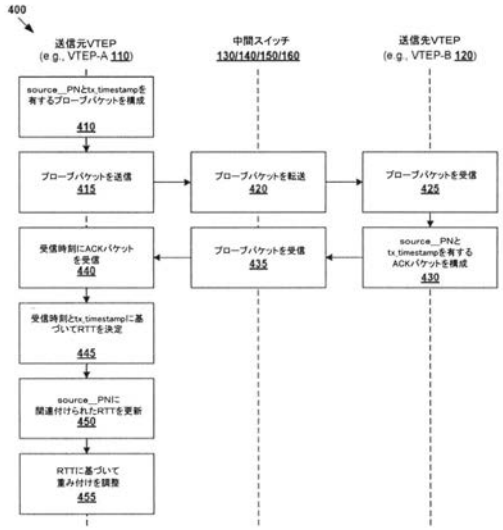
外部送信元ポート番号（source_PN） 192	経路（path_ID） 194	輻輳フラグ（congestion_flag） 196	経路の重み付け（weight） 198
SP1	P1 (via S1)	false	w1
SP2	P2 (via S2)	true	w2

【図２】

【図３】



【図４Ａ】

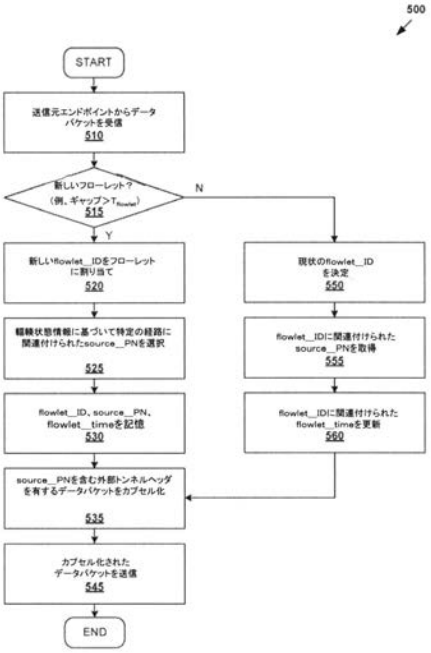


【図４Ｂ】

VTEP－Aにおける輻輳状態情報 460

外部送信元ポート番号（source_PN） 462	経路（path_ID） 464	ラウンドトリップタイム（RTT） 466	経路の重み付け（weight） 468
SP1	P1 (via S1)	R1	w1
SP2	P2 (via S2)	R2	w2

【図５】


500
START
送信元エンドポイントからデータパケットを受信
510
新しいフローレット？
（例、ギャップ>T_flowlet）
515
N
Y
新しいflowlet_IDをフローレットに割り当て
520
現状のflowlet_IDを決定
550
経験状態情報に基づいて特定の経路に関連付けられたsource_PNを選択
525
flowlet_IDに関連付けられたsource_PNを取得
555
flowlet_ID、source_PN、flowlet_timeを記憶
530
flowlet_IDに関連付けられたflowlet_timeを更新
560
source_PNを含む外部トンネルヘッダを有するデータパケットをカプセル化
535
カプセル化されたデータパケットを送信
545
END

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2017/027190

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L12/715 H04L12/729 H04L12/803
ADD.

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

B. FIELDS SEARCHED

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MOHAMMAD ALIZADEH ET AL: "CONGA", SIGCOMM, ACM, 2 PENN PLAZA, SUITE 701 NEW YORK NY 10121-0701 USA, 17 August 2014 (2014-08-17), pages 503-514, XP058053862, DOI: 10.1145/2619239.2626316 ISBN: 978-1-4503-2836-4 the whole document -----	1-21

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 May 2017	09/06/2017

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

1

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

フロントページの続き

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

(72)発明者 カッタ、ナガ
アメリカ合衆国 ９４１０９ カリフォルニア州 サンフランシスコ ポスト ストリート ７３７ アパートメント ７２１
(72)発明者 ケスラシー、アイザック
アメリカ合衆国 ９４３０４ カリフォルニア州 パロ アルト ヒルビュー アベニュー ３４０１
(72)発明者 ガグ、アディティ
アメリカ合衆国 ９４３０４ カリフォルニア州 パロ アルト ヒルビュー アベニュー ３４０１

Ｆターム(参考) 5K030 GA13 HC13 KA07 LB06 LC11 MA07 MB02 MC07