

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5632923号
(P5632923)

(45) 発行日 平成26年11月26日 (2014.11.26)

(24) 登録日 平成26年10月17日 (2014.10.17)

(51) Int. Cl.	F I
H04 L 12/44 (2006.01)	H04 L 12/44
G06 F 13/00 (2006.01)	G06 F 13/00 353C

請求項の数 20 (全 13 頁)

(21) 出願番号	特願2012-533293 (P2012-533293)	(73) 特許権者	500046438
(86) (22) 出願日	平成22年10月6日 (2010.10.6)		マイクロソフト コーポレーション
(65) 公表番号	特表2013-507833 (P2013-507833A)		アメリカ合衆国 ワシントン州 9805
(43) 公表日	平成25年3月4日 (2013.3.4)		2-6399 レッドモンド ワン マイ
(86) 国際出願番号	PCT/US2010/051698		クロソフト ウェイ
(87) 国際公開番号	W02011/044288	(74) 代理人	100140109
(87) 国際公開日	平成23年4月14日 (2011.4.14)		弁理士 小野 新次郎
審査請求日	平成25年8月20日 (2013.8.20)	(74) 代理人	100075270
(31) 優先権主張番号	61/250,013		弁理士 小林 泰
(32) 優先日	平成21年10月9日 (2009.10.9)	(74) 代理人	100096013
(33) 優先権主張国	米国 (US)		弁理士 富田 博行
(31) 優先権主張番号	12/723,697	(74) 代理人	100092967
(32) 優先日	平成22年3月15日 (2010.3.15)		弁理士 星野 修
(33) 優先権主張国	米国 (US)	(74) 代理人	100147991
			弁理士 鳥居 健一

最終頁に続く

(54) 【発明の名称】 データセンターにおけるファイウェイ

(57) 【特許請求の範囲】

【請求項 1】

コンピュータネットワーク環境において、

サーバーの第1のラックに結合された1つのファイウェイ機構を含む、互いに通信するように動的にプロビジョニングされるファイウェイ機構の組を備え、前記ファイウェイ機構は、オーバーサブスクライブされたベースネットワークに追加のネットワーク通信容量を提供するためにサーバーの第2のラックに結合される別のファイウェイ機構と通信するように構成され、

その間でプロビジョニングされたファイウェイをまだ有していない前記第1のラックと前記第2のラックとの間のネットワークトラフィックが決定され、

前記ネットワークトラフィックに基づいて、前記第1のラックと前記第2のラックとの間でファイウェイが動的にプロビジョニングされる、システム。

【請求項 2】

追加のネットワーク通信容量の必要性を検出するように構成されるコントローラーをさらに備え、前記ファイウェイ機構が、前記コントローラーによって互いに通信するために動的にプロビジョニングされる請求項1に記載のシステム。

【請求項 3】

前記オーバーサブスクライブされたベースネットワークはツリー状のトポロジーで構成される請求項1に記載のシステム。

【請求項 4】

10

20

前記オーバーサブスクライブされたベースネットワークはメッシュトポロジーで構成される請求項 1 に記載のシステム。

【請求項 5】

前記フライウェイ機構は、前記ネットワークの物理的なレイアウトに少なくとも部分的に基づく請求項 1 に記載のシステム。

【請求項 6】

前記フライウェイ機構が物理的なワイヤを介して他のフライウェイ機構と通信する請求項 1 に記載のシステム。

【請求項 7】

前記フライウェイ機構が無線技術を介して他のフライウェイ機構と通信する請求項 1 に記載のシステム。

【請求項 8】

前記無線技術は、60GHz 周波数帯技術を含む請求項 7 に記載のシステム。

【請求項 9】

前記無線技術は光リンクを含む請求項 7 に記載のシステム。

【請求項 10】

前記無線技術は 802.11 ベースの Wi-Fi 技術を含む請求項 7 に記載のシステム。

【請求項 11】

前記フライウェイ機構及び少なくとも 1 つの他のフライウェイ機構によって確立される 1 つ又は複数のフライウェイを用いてネットワークパスが構築される請求項 1 に記載のシステム。

【請求項 12】

コンピューターによる実行に応答して、前記コンピューターに、
ネットワークマシンのペア間のネットワークトラフィックを決定するステップと、
前記ネットワークトラフィックに基づいて、選択されたマシンのペアを選択するステップであって、前記選択されたマシンのペアは、その選択されたマシンのペアの間にプロビジョニングされたフライウェイをまだ有していない、ステップと、
その選択されたマシンのペアの間にフライウェイを動的にプロビジョニングするステップと
を実行させる、コンピューター実行可能な命令を有する、1 つ又は複数のコンピューター記憶媒体。

【請求項 13】

フライウェイを切断するステップを含むコンピューター実行可能命令をさらに有する請求項 12 に記載の 1 つ又は複数のコンピューター記憶媒体。

【請求項 14】

マルチホップ通信のための少なくとも 1 つの媒介マシンを使用することを含む、マシン間の別のフライウェイをプロビジョニングするステップを含む、コンピューター実行可能命令をさらに有する請求項 12 に記載の 1 つ又は複数のコンピューター記憶媒体。

【請求項 15】

前記ネットワークトラフィックを決定するステップは、サーバーマシンと通信するステップもしくはスイッチをポーリングするステップを含むか、又はサーバーマシンと通信するステップ及びスイッチをポーリングするステップの両方を含む請求項 14 に記載の 1 つ又は複数のコンピューター記憶媒体。

【請求項 16】

前記フライウェイをプロビジョニングするステップは、その間で前記フライウェイがプロビジョニングされるマシンを選択する配置アルゴリズムを実行するステップを含む請求項 14 に記載の 1 つ又は複数のコンピューター記憶媒体。

【請求項 17】

コンピューティング環境において、少なくとも 1 つのプロセッサ上で実行される方法

10

20

30

40

50

であって、

ネットワークマシンのペア間のネットワークトラフィックを決定するステップと、

前記ネットワークトラフィックに基づいて、選択されたマシンのペアを選択するステップであって、前記選択されたマシンのペアは、その選択されたマシンのペアの間でプロビジョニングされたフライウェイをまだ有していない、ステップと、

その選択されたマシンのペアの間でフライウェイを動的にプロビジョニングするステップと

を含む方法。

【請求項 18】

マシンの別のペアの間でフライウェイを切断するステップをさらに含む請求項 17 に記載の方法。

【請求項 19】

選択されたマシンのペアを選択する前記ステップは、最適なルーティングを計算するために配置アルゴリズムを実行するステップを含む請求項 17 に記載の方法。

【請求項 20】

単一のフライウェイを介して通信することができない 2 つのマシン間でマルチホップネットワーク通信を可能にするために、少なくとも 2 つのフライウェイをプロビジョニングするステップをさらに含む請求項 17 に記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、データセンターにおけるフライウェイに関する。

【背景技術】

【0002】

[0001]大規模ネットワークのデータセンターは、規模の経済性、大規模なリソースプール、簡略化された IT 管理、及び大規模なデータマイニングジョブを実行する能力を提供する。大規模なデータセンターを構築するときにネットワークのコストを抑えることは重要な考慮事項である。ネットワークコストは主要な経費の 1 つである。知られているように、サーバークラスター内のサーバーの任意のペア間に回線速度の通信帯域幅を提供することに伴うコストは、一般的に、サーバークラスターのサイズに対して超直線的に増加する。

【発明の概要】

【発明が解決しようとする課題】

【0003】

[0002]生産データセンターネットワークは、必要な容量を提供するために、高帯域幅のリンクやハイエンドのネットワークスイッチを使用するが、なおオーバーサブスクライブされている (over-subscribed) (時折容量が不足する) ため、散発的な性能の問題に悩まされている。オーバーサブスクリプションは、一般的に、技術上の制約、高価な「超高速大型コンピューター (big-iron)」スイッチを必要とするこれらのネットワークのトポロジー (例えば、ツリー状)、及び、コストを低く保つためのネットワーク管理者への圧力の組み合わせの結果である。他のネットワークトポロジーは同様の問題を抱えている。

【課題を解決するための手段】

【0004】

[0003]この概要は、詳細な説明において以下にさらに記載される代表的な概念のうちの選択されたものを簡略化した形式で紹介するために提供される。この概要は、特許請求の主題の重要な特徴又は不可欠な特徴を特定するようには意図されておらず、いかなるようにも特許請求の主題の範囲を限定するように使用されるようには意図されない。

【0005】

[0004]簡潔に言えば、本明細書に記載の主題の様々な側面は、追加のネットワーク通信

10

20

30

40

50

容量がフライウェイ (flyway) と呼ばれる通信リンクを使用してオーバーサブスクライブされたベースネットワークに提供される技術を対象とする。一般に、各フライウェイは、必要に応じて動的に設定し、必要に応じて取り壊すことのできるネットワークマシン間の追加のネットワーク通信経路を備える。

【0006】

[0005] 1つの実施例において、ネットワーク内の各サーバーラックは、他のフライウェイ機構と通信することができる、関連するフライウェイ機構を有する。フライウェイ機構は有線技術又は無線技術に基づいてもよい。コントローラーは、2つのラックの間に追加のネットワーク通信容量が必要であることを検出すると、フライウェイ機構を互いに通信させることによって、フライウェイを動的にセットアップ (プロビジョニング) する。一旦セットアップされると、マシン間の追加の通信リンクは、それらのマシン間のネットワークトラフィックの少なくともいくつかを運ぶのに使用される。

10

【0007】

[0006]他の利点は、図面と併せて以下の詳細な説明から明らかになる。

[0007]本発明は実施例によって説明され、添付の図面に限定されず、同様の参照符号は同様の要素を示す。

【図面の簡単な説明】

【0008】

【図1】[0008]フライウェイが確立され得るフライウェイ機構を組み込んだ例示的なデータセンターを示すブロック図である。

20

【図2】[0009]ネットワークマシン間で設定されるフライウェイを表現したものである。

【図3】[0010]サーバーラック及びラックのセットの中央付近のラック上に取り付けられた例示的な60GHz装置の範囲を示す部分的なデータセンターの表現である。

【図4】[0011]ネットワークマシンのペア間でフライウェイをプロビジョニングするための例示的なステップを示すフロー図である。

【発明を実施するための形態】

【0009】

[0012]本明細書に記載された技術の様々な態様は、一般的に、既存のネットワークの利点を保持するが、実質的にその性能を向上させる、ネットワーク設計のためのものである。この目的のために、ベースネットワーク (ツリー状又はメッシュ状を含む) が平均的なケースのためにプロビジョニングされている (したがってオーバーサブスクライブされている) ハイブリッドアーキテクチャーが記載されるが、ネットワークで必要とされる場所及び時において追加の容量を提供するためにオンデマンドベースで追加される、余分なリンクを含めることによって、ホットスポットを取り扱う。本明細書中で使用されるとき、このようなリンクはフライウェイと呼ばれる。将来のネットワーク設計 (例えば、クロ・ネットワーク (Clos networks)) もまた、本明細書に記載の技術によって強化され得ることに留意されたい。理解されるように、強化された設計は、ネットワーク機器の量やコストの大きな増加なくしてサーバー数を拡大するときに、ネットワークがその性能を維持することを可能にする。

30

【0010】

[0013]本明細書に記載された例のいずれもが限定的でない例であることが理解されるべきである。このように、本発明は、本明細書に記載の任意の特定の実施例、態様、概念、構造、機能、又は例に限定されない。むしろ、本明細書に記載される実施例、態様、概念、構造、機能又は例の任意のものは限定的でなく、本発明は、一般にコンピューティングネットワーク及びコンピューターネットワークにおいて利益及び利点を提供する様々な方法で可以使用することができる。

40

【0011】

[0014]図1は、ツリー状のトポロジーに基づく生産ネットワークを示す。複数のラック $102_1 - 102_n$ は、各々が、ラックスイッチ $104_1 - 104_n$ の最上部を介して通信するサーバーを有する。典型的なネットワークは、ますます強力なリンク及びツリーを

50

上るスイッチとともに、ラックあたり20から40のサーバーを有する。フライウェイがツリー状のトポロジに限定されず、closネットワークや他の形式のメッシュトポロジ、及びFatTreeトポロジを含む任意のトポロジで 사용할 ことができることに留意されたい。

【0012】

[0015]図1に表されるように、ラックスイッチ $104_1 - 104_n$ のそれぞれの最上部は、1つ又は複数の集約スイッチ $106_1 - 106_k$ を介して互いに結合される。このように、各サーバーは異なるラック内のサーバーを含む任意の他のサーバーと通信することができる。この例では、高レベルのアグリゲーションスイッチ108は、ラックレベルの集約スイッチ $106_1 - 106_k$ を結合し、集約スイッチの結合について1つ又は複数の追加のレベルが存在し得ることになる。

10

【0013】

[0016]アプリケーションの要求は、一般的に、オーバーサブスクライブされたネットワークによって満たすことができるが、時折、ネットワークは「ホットスポット」を処理するのに十分な容量を有していない。本明細書に記載の技術は、必要に応じて追加のデータトラフィックを処理するためにフライウェイを使用して追加の容量を提供する。

【0014】

[0017]図1及び2に示されるように、フライウェイ（図2の曲線の矢印）、フライウェイコントローラー112によって制御されるフライウェイ機構 $110_1 - 110_n$ を介して実施される。フライウェイは、必要に応じてフライウェイコントローラー112によって動的に設定され、必要とされない場合（又は他の場所で必要とされる場合）に停止され得る。図2は、1つのフライウェイ220がラック 102_1 及び 102_n 並びにそれぞれのラック最上部スイッチ 104_1 及び 104_n をリンクするためにどのように使用され、その一方、別のフライウェイ222がラック 102_2 及び 102_m 並びにそれぞれのラック最上部スイッチ 104_2 及び 104_n をどのようにリンクするかを示す。ラック／ラック最上部スイッチが、としてラック 102_1 及び 102_2 の間のフライウェイ221によって表されるように、任意の時点で複数のフライウェイを有し得ることに留意されたい。ラックごとに単一のフライウェイ機構が示されるが、おそらくは異なる通信技術（例えば、無線及び光）を使用して、ラックごとに複数のフライウェイ機構（又は単一のフライウェイ機構内の複数の装置）があってもよいことが理解される。

20

30

【0015】

[0018]データセンターネットワークからのトレースの分析は、いつでも、ごくわずかのラック最上部スイッチが「ホット」であること、つまり、それらが大量のトラフィックを送信及び／又は受信していることを示す。さらに、ホットである場合、ラック最上部スイッチは、通常、それらのデータの多くをごく少数のラック最上部スイッチと交換する。これは、ラック最上部スイッチのほんの一部が後れをとり、ネットワーク全体を妨げる、偏ったボトルネックにつながる。本明細書に記載されたフライウェイは、これらのいくつかのラック最上部スイッチへ追加の容量を提供し、したがって全体的な性能をおおきく向上させる。実際、比較的低い帯域幅の少数のフライウェイは、オーバーサブスクライブされたデータセンターネットワークの性能を大幅に向上させる。

40

【0016】

[0019]フライウェイは、比較的少額の追加コストでネットワークに追加することができる。これは、無線リンク（例えば、60GHz、光リンク及び／又は802.11n）の使用及び／又はランダムな方法で容量を追加するための商品スイッチ（commodity switches）の使用によって達成することができる。一般に、任意のフライウェイ機構は、結合要件（例えば、無線の範囲内で、光などのために視線を有する）を満たす限り、他のフライウェイ機構にリンクすることができる。

【0017】

[0020]したがって、フライウェイは、フライウェイ機構（例えば、適切な無線装置）、及び／又はラック最上部スイッチのサブセット（部分集合、subset）を相互接続する商品

50

スイッチ間のオンデマンド設定されている無線リンクを介してなど、さまざまな方法で実施されてもよい。後述するように、60GHz無線技術は、短距離（1～10メートル）、高帯域幅（1Gbps）の無線リンクをサポートするので、フライウェイを作成するための1つの実施例である。さらに、60GHzの高容量及び限定された干渉範囲が利点をもたらす。

【0018】

[0021]フライウェイで強化された（flyway-enhanced）オーバーサブスクリプされたネットワークの性能は、オーバーサブスクリプされていないネットワークの性能に近くか又はそれと同等ですらある。最大の利益を達成する1つの方法は適切な位置にフライウェイを配置することである。ネットワークトラフィックの需要は一般に短い時間スケールで予測可能／決定可能であって、フライウェイのプロビジョニングが需要の変化に追いつくことを可能にすることに留意されたい。本明細書に記載されるように、中央フライウェイコントローラー112は、需要データを収集し、動的な方法でフライウェイを適応させ、経路を交通路に切り替える。

【0019】

[0022]無線フライウェイは、必要に応じてリンクを形成することができ、したがって、利用可能な容量が中央フライウェイコントローラー112によって決定されることを要求するいずれのラック最上部スイッチペアに対しても、利用可能な容量を分配するために使用することができる。有線フライウェイは、同等の利点を提供し、安価なスイッチがラック最上部スイッチのサブセットに接続されている場合、これらのスイッチにおける限定されたバックプレーンの帯域幅は、それを必要とする接続されている多くのラック最上部スイッチペアのいずれかの間で分割することができる。（以下に記載するように、マルチホップフライウェイは、間接的な利益を可能にする1つの代替例であるが）有線フライウェイは、ラック最上部スイッチペアがフライウェイスイッチのうちの1つを介して偶然接続される場合にのみ、直接的にフライウェイの恩恵を受けることができると言う点で、いくらかより限定的である。しかしながら、有線フライウェイは、（例えば、NICが10Gbpsまで上がり、リンクが40Gbpsまで上がるように）有線の速度に追いつく可能性がより高い。いずれにせよ、フライウェイは、すべてのラックにわたって同じ帯域幅を拡散させることの代わりに対して利点をもたらす、それを必要としないリンク上でそれができるだけ使用されないようにする。さらに、フライウェイは、今日では、既存の生産データセンターのトポロジーにおいて展開することができる。

【0020】

[0023]ホットなラック最上部スイッチと、それらがトラフィックを交換する、いくつかの他のラック最上部スイッチとの間に追加の容量を提供するべくフライウェイを追加するために、最大の高速化を提供するフライウェイ機構のペアが選択され得る。各フライウェイはまた、十分な容量を必要とする。最も混雑しているラック最上部スイッチを備えたラックと、スイッチが最大のデータを交換する別のラックととの間にフライウェイを確立することは、容易である。しかし、同じラック又は別の場所に別のフライウェイを確立すべきかどうかを考慮する必要があるなど、その後の選択肢はあまり明確ではない。

【0021】

[0024]一般に、あまりにも薄くフライウェイを広げたり、又は数個のラック最上部スイッチにそれらを集中したりすることは、特にうまく機能するということはない。たとえば、上位50のラック最上部スイッチとその最大の対応するものとの間に、それぞれ、フライウェイを配置することは、ホットなラック最上部スイッチの完了時間を十分に短縮することにはならない。逆に、上位5つのラック最上部スイッチとそれらの上位10個の対応するものの各々との間にフライウェイを配置すると、上位5つにおける混雑が排除され、6番目のラック最上部スイッチはボトルネックとして終わることになる。一般的には、より多くのラック最上部スイッチを支援することと、ホットなラック最上部スイッチのいずれもにおいて混雑を十分に減らすこととの間の適切なバランスを達成することによって、最大の高速化が得られる。フライウェイの配置のための1つの適当なアルゴリズムが以下に記載される。

【0022】

[0025]各フライウェイがどのくらいの容量を必要としているかに関して、上位10のラック最上部スイッチと、それらが最大のデータを交換する5つの他のラック最上部スイッチの各々との間にフライウェイを追加する例（すなわち、全部で50のフライウェイ）を考慮する。ほとんどのフライウェイは、10パーセント未満のラック最上部スイッチのアップリンク帯域幅が有用であることを必要とする。1つの理由は、ラック最上部スイッチのアップリンクがは他のすべてのラック最上部スイッチにトラフィックを運ぶ一方、フライウェイは1つの他のラック最上部スイッチにトラフィックを運ぶだけでよいということである。

【0023】

[0026]フライウェイの有用性は一般に疎な（sparse）需要行列の原因となるアプリケーションの特性に起因する。フライウェイは、すべての要求に対しては利点を提供しないかもしれないが、実用的なアプリケーションの大規模なセットはフライウェイから恩恵を受ける。たとえば、ウェブサービスをサポートするデータセンターにおいて、要求トラフィックは、サーバー間で負荷分散され、その各々は、おそらくはいくつかの他のサーバーに応答の一部（例えば、広告）を生成するよう要請することによって、応答ページを組み立てる。データマイニングジョブにおけるマップ低減のうちの低減の一部はおそらく最悪のシナリオであり、各低減部（レデューサー、reducer）はすべてのマッパーからデータを引き出す。当該ジョブはすべてのレデューサーが完了するまで邪魔される。その場合でも、すべてのラック最上部スイッチが同時に混雑するほど多くのマッパー及びリデューサーを有することは稀である。

【0024】

[0027]60GHzの無線通信などに目を向けると、ミリ波無線通信が急速に進んできており、データセンター内にフライウェイを構築するのに適した技術である。60GHzの周波数帯は、許可されていない装置による使用のため、2001年にFCCによって確保された7GHz幅の帯域（57-64GHz）である。ISOやIEEEによって指定されたチャンネル帯域幅は、2.160GHzである。60GHz帯における十分に大きな帯域幅は、より大容量のリンクを容易にする。例えば、わずか1bps/Hzを実現する単純な符号化方式は、7Gbpsの公称帯域幅のリンクを構築することを可能にする。より良い変調技術及び符号化技術が実施されると、60GHzの装置はスペクトル効率がよくなる可能性が高く、7つの直交する1Gbpsのチャンネルは共通となり、1Gbpsのリンク速度はフライウェイを構築するのに十分である。4-5Gbpsのリンク速度は60GHz帯で複数のチャンネルを使用して、すでに利用可能であることに留意されたい。

【0025】

[0028]容量に加え、60GHz帯の通信の他の特徴は、フライウェイをインスタンス化するために特に魅力的である。1つには、60GHz帯の装置は、比較的短い範囲（5-10メートル）を有する。無線通信路では、損失は周波数の二乗に正比例し、したがって、2.4及び5GHzのWi-Fi信号と比較した場合、60GHz帯の信号は、距離によってより迅速に減衰する。もう1つには、高い指向性の60GHz帯の通信を実現するのは容易である。信号波長は周波数に反比例するため、60GHz帯の信号は非常に短い波長（5mm）を有し、これにより、RFチップの設計者は、トランシーバーチップにアンテナアレイを組み込むことができる。適切なフェージングを有する多くのアンテナ及びトランシーバーを単一のチップに統合する能力は、ビーム形成を可能にし、したがって指向性の通信を可能にする。さらに、フェーズドアレイアンテナは、信号品質を向上させ、短い距離においてビットエラーレートの低いよりクリーンな信号につながる。例えば、1平方インチのアンテナは、60GHzで25dBiの信号向上をもたらす。チップ上のアンテナの統合の副次的な利点は、これによってワイヤーがチップとの間で信号を伝送する必要性が回避され、二桁分までパッケージングのコストを削減されることである。

【0026】

[0029]短距離で大容量の指向性リンクは、物理的な空間が高品質のリソースであるため

にサーバーラックが通常密にパックされ、移動性及びサーバーラックあたり約10ワットの電力消費が問題でないような、データセンターにおいて有益である。指向性は、ネットワーク設計者が、空間の再利用により、全体のスペクトル効率を向上させることを可能にする。たとえば、4つのラック最上部スイッチ間の2組の通信は、指向性及び距離のために、同時に発生することができる。

【0027】

[0030]典型的なデータセンターのサーバーでは、ラックは約24インチ幅であり、図3に示すように、密度の高い、規則的なパターンで配置されている。図3において、各ボックスは、例えば、10個の行に配置された24x48インチのラックを表す。円は中央のラックの上にマウントされた60 GHz帯装置の10mの範囲を表し、約70の他のラックが含まれている。指向性アンテナを使用して、ラックの上部に取り付けられた小さな(2-3立方インチ)60 GHzのネットワークカードは、必要に応じて、いくつかのラックの間に複数の非干渉ギガビットフライウェイをインスタンス化できる。約60度のビーム幅を持つ電子可動フェーズドアレイアンテナは、ミリ秒の待ち時間で誘導することができる。

10

【0028】

[0031]指向性が結果をもたらすことに留意されたい。たとえば、発見はもはや容易ではなく、無指向性の通信にとってよく機能する一般の媒体アクセス制御プロトコルは修正する必要がある。それにもかかわらず、このことは、研究コミュニティではよく理解されている問題であり、多くの解決策が存在している。IEEE及びISO 60 GHzワーキンググループは、指向性通信及び空間再利用のためのサポートを組み込むことに積極的に取り組んでいる。彼らの主なシナリオは家庭内のマルチメディア通信であるが、議論されている解決策は、データセンターのネットワークに適用するように修正することができる。具体的には、ネットワークが構成される方法及びネットワークの規模のために、制御チャネルへと変えることができるすべての発見可能なラック最上部スイッチの間で、しばしば有線で、バックチャネル接続が存在するので、問題を解決することができる。制御チャネルは、次いで、ラック最上部スイッチを発見し、60 GHz帯のビームを調整し、干渉軽減のためにチャネルを割り当て、衝突のないメディアアクセスを解決するために使用することができる。

20

【0029】

[0032] 60 GHzの別の問題は、これらの高周波数において、信号が物体によって吸収され、視線のない(non-line-of-sight、NLOS)通信が可能でないということである。しかし、データセンターでは、装置は、ラックの上に取り付けて人間のオペレーターの邪魔にならないようにすることができ、それによって、視線が遮られないようにすることができる。さらに、最近の技術の進歩はこの制限を克服し始めている。

30

【0030】

[0033]上述したように、スペクトル効率を改善することと組み合わせた大量のスペクトルは、多くのフライウェイが構築されて独立して動作されることを可能にする(10 bits/Hzのスペクトル効率で、70のフライウェイまで、各々が1 Gbpsで動作するように構築することができる)。このことは、指向性に起因する低干渉性の特性及び経路損失に起因する小さな範囲の特性と組み合わせて、データセンター内に存在する数千のサーバーラックにわたる大幅な周波数再利用を可能にする。

40

【0031】

[0034]さらに、他の高帯域幅の無線技術とは異なり、60 GHz帯は、世界全体で利用可能である。したがって、チップセット及び装置は世界中で使用可能であって、規模の経済性を提供し、コストを押し下げる。

【0032】

[0035]フライウェイを備えたネットワークを実施することに目を向けると、ラック最上部スイッチのペア間にフライウェイを形成するために、対応するラックの上に配置された1つ又は複数の装置は無線リンクを作成する。技術の選択は、利用可能な帯域幅、空間の再利用に利用可能なチャネル数、フライウェイの干渉パターン及び範囲に影響を与える。アンテナ技術は、フライウェイを設定し取り壊すために必要な時間を決定づける。ラック

50

最上部スイッチごとにいくつかの無線装置を追加することはコストをわずかに増加させることに留意されたい。

【0033】

[0036] ラック最上部スイッチのランダムなサブセットを相互接続するラック最上部スイッチを現在のものとするなど、追加のスイッチを使用することによって、有線フライウェイを構築することができ、例えば、商用スイッチは、各々が1 Gbpsのリンクを備えた20のラック最上部スイッチを相互接続してもよい。リンクを短く保つために、フライウェイスイッチは、データセンター内の互いに近接しているラックを接続してもよい。有線フライウェイを配置する場合、スペクトル割り当てや干渉は問題ではない。しかし、それらのランダムな構造は有線フライウェイを制約する；例えば、多くのトラフィックを交換し余剰容量の恩恵を受けることができるラック最上部スイッチペアは、有線フライウェイなしに終わるかもしれない。

10

【0034】

[0037] いずれにせよ、有線又は無線のフライウェイの構築は、すべてのラック最上部スイッチのペアにわたって均一に同じ量の帯域幅を分割することに比べて改善される。需要行列が疎であるときに起こるように帯域幅を均一に拡散してその多くを浪費するよりも、フライウェイは、余剰容量から最も恩恵を受けることができる需要行列の部分を対象とするように、予備の帯域幅を使用する方法を提供する。

【0035】

[0038] 中央フライウェイコントローラー112（図1）は、ラックスイッチのペア間の需要の推定値を収集する。たとえば、情報はエンドサーバー自体における軽量の機器から、又はスイッチにおけるSNMPカウンタをポーリングすることによって、収集することができる。これらの推定値を使用して、コントローラー112は、周期的に又はその他の方法で、利用可能なフライウェイを配置する配置アルゴリズム（後述）を実行することができる。

20

【0036】

[0039] フライウェイベースのネットワークのトポロジーはしたがって動的であり、マルチパスルーティングを必要とする。同様の問題に対する解決策が知られている。コントローラー112は、ラック最上部スイッチのペア間でどれだけのトラフィックがベースネットワークとともに進み、又はその代わりに、存在する場合に、送信側のラック最上部スイッチから受信側のラック最上部スイッチへのフライウェイをとるかを決定する。ラック最上部スイッチは、異なるMPLS（マルチプロトコルラベルスイッチング）ラベル切り換え経路上に異なるフローを割り当てることにより、この比率に従ってトラフィックを分割する。もしあれば、ごくわずかのフライウェイが各ラック最上部スイッチにおいて利用可能であることに留意されたい。したがって、各スイッチにおいて必要なLSP（ラベル切り替え経路、label switched paths）の数は小さく、ベース及び1ホップ長（one hop long）のフライウェイにわたってトラフィックを分割する問題は、他のトラフィック分割技術よりもはるかに簡単である。

30

【0037】

[0040] 最適なフライウェイを作成する問題は、最適化問題と見なすことができる。ラック最上部スイッチ i 、 j の間の需要 D_{ij} を仮定し、 C_l がリンク l の容量である場合、最適なルーティングは、最大完了時間

40

【0038】

【数 1】

$$\begin{aligned}
 & \text{such that} \quad \min \max_{ij} \frac{D_{ij}}{r_{ij}} \quad (1) \\
 & \sum_{l \in \text{incoming}} r_{ij}^l - \sum_{l \in \text{outgoing}} r_{ij}^l = \begin{cases} D_{ij} & \text{at ToR } j \\ -D_{ij} & \text{at ToR } i \\ 0 & \text{at all other ToRs} \end{cases} \\
 & \sum_{ij} r_{ij}^l \leq C^l \quad \forall \text{ links } l
 \end{aligned}$$

10

を最小化するものであり、ここで r_{ij} は、ラック最上部スイッチのペア i, j について達成されるレートであり、 r_{ij}^l はリンク l 上のそのペアのトラフィックの一部である。

【0039】

[0041] 最適／ほぼ最適なフライウェイウェイの配置を計算することは、トポロジーを適切に変更すること及び上記の最適化問題を再度解くことを含む。たとえば、すべての可能性のあるフライウェイが、一定数以上が同時にアクティブになることができない、又はフライウェイのどれかが一定量より大きい容量を持つことができないという制約をもって、追加することができる。技術的な制約のために引き起こされる、上記の最適化問題のすべての変形が、扱いやすいとは限らない。代わりに、上記の最適化問題を解いて完了時間をもっとも低減するフライウェイを見つけることにより、一度に1つのフライウェイを追加する、欲張りな手順を使用することができる。他の手順がより良い近似であってもよい。

20

【0040】

[0042] 1つの例は、一般に、図4の例示的なフロー図に表され、考慮されているペアについてネットワークトラフィックデータを収集することを表すステップ402において始まる。一般的に、トラフィックデータに基づいて、2つの選択されたマシンのペア（ラック最上部スイッチ）の間で必要とされる最適なフライウェイ（例えば、スイッチがホットであるが、フライウェイがまだプロビジョニングされていない）が、ステップ404で選択される。

30

【0041】

[0043] フライウェイはまた、ラックの残りの部分への一次接続（primary connections）が失敗したラックへの接続を提供するために使用され得ることに留意されたい。このような障害補償フライウェイは、例えば、ステップ404で必要とされる最適なフライウェイであると考慮されてもよく、又は、容量に基づいて他を選択する前など、別個に考慮されてもよい。

【0042】

[0044] ステップ406は、最適なフライウェイをプロビジョニングすることができるかどうかを評価することを表す。例えば、フライウェイの合計数に制約がある場合、この時点で利用可能な別のフライウェイがないかもしれない。しかし、利用可能なフライウェイが存在しないが、最適なプロビジョニングされていないフライウェイが現在プロビジョニングされている別のフライウェイよりも必要とされている場合、ステップ408は、フライウェイが利用可能になるように、その他のフライウェイ（又は、以下に記載するようにマルチホップシナリオが使用される場合）を取り壊してもよい。例えば、最も必要とされないフライウェイが取り壊されてもよい。

40

【0043】

[0045] しかし、フライウェイを取り壊すことが適切でないかもしれないことに留意されたい。たとえば、すでにプロビジョニングされたフライウェイのすべてがさらに多くのトラフィックを処理しており、したがって、動作が行われない（例えば、ステップ408からの破線）ステップ404における選択されたペアのフライウェイより必要とされるかも

50

しれない。

【0044】

[0046]ステップ410は（適切な場合に）最適なフライウェイをプロビジョニングする。一旦プロビジョニングされると、フライウェイは、ステップ412によって表されるように、選択したペア間で、ネットワークトラフィックの少なくとも一部を処理するために使用することができる。次いで処理が繰り返される。図4が1つの例にすぎず、代替的なものとして、まだプロビジョニングされていない上位Nまでの最適なフライがステップ402においてペアの各ポーリングの間に追加することができる。

【0045】

[0047]フライウェイとともに、ネットワークの物理トポロジーは、技術に起因する制約と同様に、考慮する必要がある。無線フライウェイが範囲（距離）によって制約される一方、ラック最上部スイッチのランダムなサブセットを相互接続することにより構成されている有線フライウェイは、単にこれらのランダムなサブセットの間に余剰容量を提供することができる。60 GHzのフライウェイは、10メートルの距離に及び、生産データセンターからデータセンターのレイアウト（例えば、図3）を使用する。有線フライウェイについて、24ポートの1 Gbpsスイッチが、フライウェイチャネルに使用されるポートの一部とともに使用することができる。このような制約はフライウェイの利点に影響を与えるが、利益は依然として大きい。

【0046】

[0048]無線フライウェイと同じ利点を得るためにはより多くの有線フライウェイを追加する必要があることに留意されたい。たとえば、50個の24ポートスイッチが追加される場合、1200の二重リンクがネットワークに追加される。無線フライウェイは、50の半二重リンクのみによって、同等の性能を提供する。これが、無線フライウェイがターゲットを絞った形で追加される理由である。それらは、追加の容量を必要とするラック最上部スイッチのペアをスピードアップするのに役立つ。選択されたもののなかに偶然ある場合にのみ、有線フライウェイは、ランダムに追加され、ペアの利益になる。

【0047】

[0049]マルチホップフライウェイは実現可能である。図3の円内のマシンが円の外側のマシンと通信することを望むことを考える。しかし、別のマシンに十分近い円の内側のマシンからの第2のホップは、通信を達成するために使用することができる。

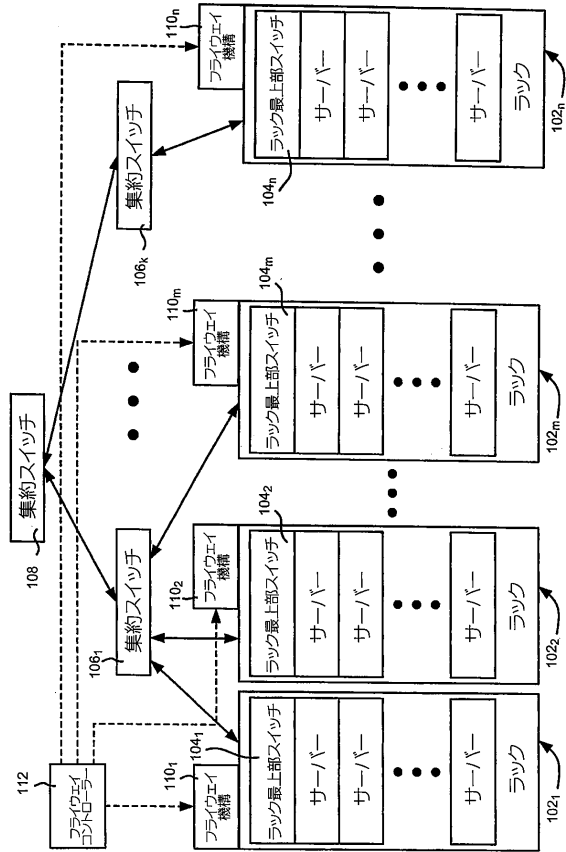
【0048】

[0050]さらに、コントローラー112（図1）は、2つのマシン（例えば、A及びZ）が単一のホップ範囲の外にあるかどうかを考慮することができる。そうである場合、近接しているがラック最上部スイッチを介して通信しており、その通信がフライウェイに切り換えられる、2つの他のマシン（例えば、C及びM）の間にフライウェイを構築することができる。新たに構築されたC及びMについてのフライウェイによって提供される帯域幅の節約は、次いで、マシンA及びZがラック最上部スイッチを使用して通信することを可能にするラック最上部スイッチを解放するために使用することができる。

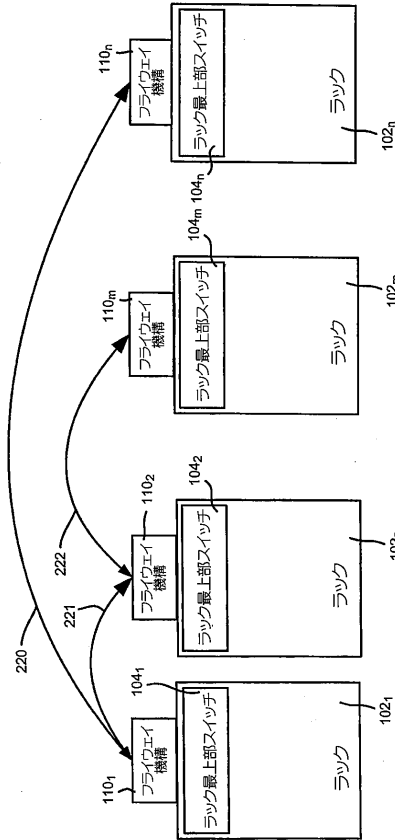
【0049】

[0051]本発明は、様々な修正及び代替構造の影響を受けやすいが、特定の図示された実施例が図面に示されており、詳細に上述されている。しかし、本発明を開示された特定の形態に限定する意図はなく、反対に、本発明の趣旨及び範囲内にあるすべての修正、代替構造及び均等物をカバーするよう意図されることが理解されるべきである。

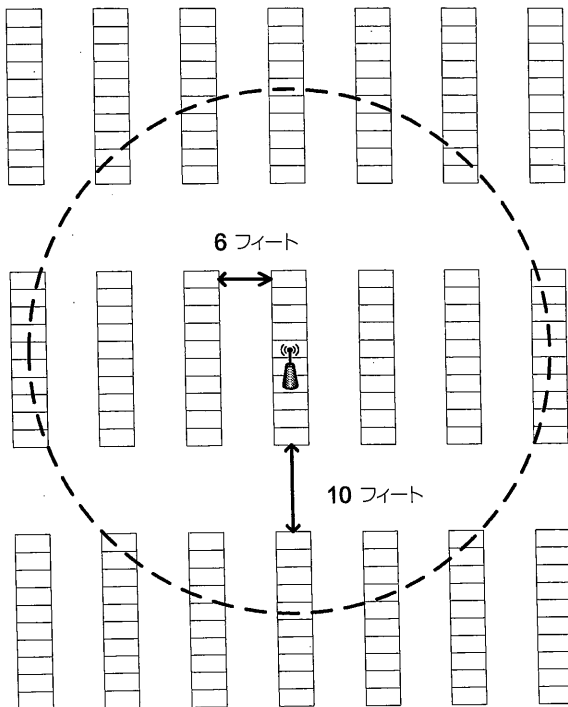
【図 1】



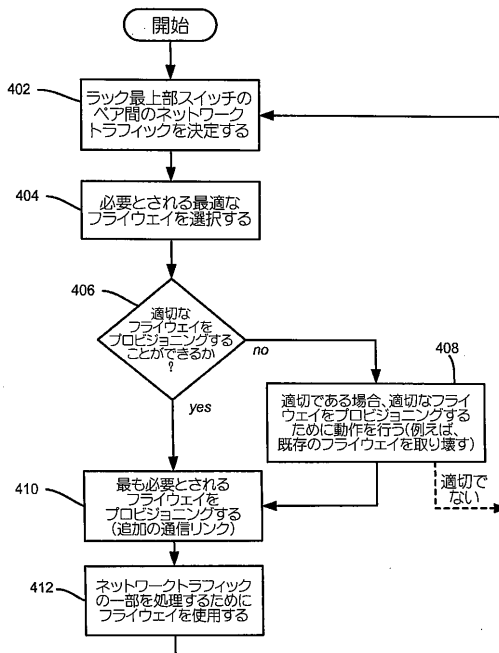
【図 2】



【図 3】



【図 4】



フロントページの続き

- (72)発明者 パディエ, ジテンドラ・ディー
アメリカ合衆国ワシントン州98052-6399, レッドモンド, ワン・マイクロソフト・ウェイ, マイクロソフト コーポレーション, エルシーエイインターナショナル・パテント
- (72)発明者 カンデュラ, スリカンス
アメリカ合衆国ワシントン州98052-6399, レッドモンド, ワン・マイクロソフト・ウェイ, マイクロソフト コーポレーション, エルシーエイインターナショナル・パテント
- (72)発明者 バール, パラムビール
アメリカ合衆国ワシントン州98052-6399, レッドモンド, ワン・マイクロソフト・ウェイ, マイクロソフト コーポレーション, エルシーエイインターナショナル・パテント

審査官 大石 博見

- (56)参考文献 特開2004-228828 (JP, A)
特開2000-232483 (JP, A)

- (58)調査した分野(Int.Cl., DB名)
H04L 12/44
G06F 13/00