(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2012-528552
(P2012-528552A)
(43)公表日　平成24年11月12日(2012.11.12)

(51)Int.Cl.		F I			テーマコード（参考）
H04L 12/56	(2006.01)	H04L	12/56	H	5B089
G06F 13/00	(2006.01)	G06F	13/00	357Z	5K030

審査請求 未請求　予備審査請求 未請求　（全 31 頁）

(21)出願番号　特願2012-513344 (P2012-513344)
(86)(22)出願日　平成22年5月28日 (2010.5.28)
(85)翻訳文提出日　平成24年1月30日 (2012.1.30)
(86)国際出願番号　PCT/US2010/036758
(87)国際公開番号　WO2010/138937
(87)国際公開日　平成22年12月2日 (2010.12.2)
(31)優先権主張番号　61/182,063
(32)優先日　平成21年5月28日 (2009.5.28)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　12/578,608
(32)優先日　平成21年10月14日 (2009.10.14)
(33)優先権主張国　米国 (US)

(71)出願人　500046438
マイクロソフト　コーポレーション
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
(74)代理人　110001243
特許業務法人　谷・阿部特許事務所
(72)発明者　アルバート　グリーンバーグ
アメリカ合衆国　９８０５２　ワシントン州　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　エルシーエー－インターナショナル　パテンツ内

最終頁に続く

(54)【発明の名称】アジャイルデータセンタネットワークアーキテクチャ

(57)【要約】
本特許出願は特にデータセンタで用いられ得るアジャイルネットワークアーキテクチャに関する。１つの実施例はレイヤ３のインフラストラクチャのマシンを接続する仮想レイヤ２のネットワークを提供する。

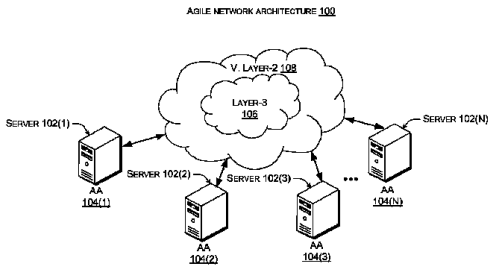


FIG. 1

【特許請求の範囲】
【請求項１】
アプリケーションアドレス（１０４）を複数のマシンに割り当て、ローカルアドレス（２０６）をレイヤ３のインフラストラクチャ（１０６）のコンポーネントに割り当てることによって前記マシンを接続する仮想レイヤ２のネットワーク（１０８）を提供するステップを備えることを特徴とする方法。
【請求項２】
個々のマシン間でターンアラウンドパスを早期に使用するステップを更に備えることを特徴とする請求項１記載の方法。
【請求項３】
前記マシンは、サーバ又は仮想マシンを備えることを特徴とする請求項１記載の方法。
【請求項４】
前記仮想レイヤ２のネットワークを提供するステップは、複数の仮想レイヤ２のネットワークを提供するステップを備えることを特徴とする請求項１記載の方法。
【請求項５】
レイヤ３のコンポーネントのローカルアドレスを有する第１のマシンと第２のマシンとの間でパケットを、前記第１のマシンと前記第２のマシンとの間の前記レイヤ３のインフラストラクチャの個々のパスに従ってカプセル化するステップを更に備えることを特徴とする請求項１記載の方法。
【請求項６】
前記マシンのうちの２つの間で前記レイヤ３のインフラストラクチャの個々のパスをランダムに選択するステップを更に備えることを特徴とする請求項１記載の方法。
【請求項７】
前記個々のパスを選択するためにバリアントロードバランシング（ＶＬＢ）を利用するステップを更に備えることを特徴とする請求項６記載の方法。
【請求項８】
周期的に又はネットワークイベントに応答して前記個々のパスを再選択するステップを更に備えることを特徴とする請求項６記載の方法。
【請求項９】
サーバ（３１６（１））であって、
コンピュータ可読命令を実行する少なくとも１つのプロセッサと、
前記少なくとも１つのプロセッサによって実行可能でかつ他のサーバ（３１６（Ｎ））への配信のためのパケットを受信し、中間スイッチ（３１０）を介して前記配信用の前記パケットをカプセル化するように構成されたアジャイルエージェント（３２０）と、
を備えることを特徴とするサーバ。
【請求項１０】
前記アジャイルエージェントは前記中間スイッチを複数の中間スイッチからランダムに選択するように構成されていることを特徴とする請求項９記載のサーバ。
【請求項１１】
前記アジャイルエージェントは前記配信用のパスを選択し、通信障害の指示を受信すると、複数の中間スイッチから新しい中間スイッチを含む新しいパスを再選択するように構成されていることを特徴とする請求項１０記載のサーバ。
【請求項１２】
前記サーバは複数の仮想マシンをサポートするように構成され、前記アジャイルエージェントは２つの仮想マシン間で前記パケットを配信するためのパスを選択するように構成されていることを特徴とする請求項１０記載のサーバ。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、アジャイルデータセンタネットワークアーキテクチャに関する。

【背景技術】
【０００２】
従来のデータセンタネットワークアーキテクチャはそれらのアジリティ（データセンタネットワークのいずれかのサーバを何らかのサービスに割り当てるためのそれらの能力）を弱体化可能ないくつかの設計上の欠点に悩まされている。第一に、従来のネットワークの構成は典型的に、実質的にツリー状であり、比較的高価な機器からなる。これはネットワークの他のどこかで予備能力が利用可能であるときでさえ、計算ホットスポットの混雑及びその発展の結果である。第二に、従来のデータセンタネットワークは１つのサービスにおけるトラフィック輻輳がその周囲の他のサービスに影響を与えることを防止することにほとんど効果がない。１つのサービスがトラフィック輻輳を経験すると、同一のネットワークのサブツリーを共有する全てのサービスが巻き添え被害（ダメージ）を被ることは普通である。第三に、従来のデータセンタネットワークのルーティング設計は、通常、サーバにトポロジ的に重要なインターネットプロトコル（ＩＰ）アドレスを割り当て、サーバをＶＬＡＮ(Ｖｉｒｔｕａｌ　Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ)のうちで分けることによりスケールを達成する。しかしながら、これはサーバがサービスの中で再割り当てされるとき大きな設定負担を作り出し、よってデータセンタのリソースを更に使ってしまう。更に、人間の関与がそれらの再設定で通常要求され、これによりこの処理の速度を制限してしまう。最後に、従来のデータセンタネットワークを設定する際の難しさ、そのようなネットワークで使用される機器のコスト等の他の事項は、これらのネットワークのアジリティを否定的に影響を与えることができる。
【発明の概要】
【０００３】
本特許出願は、特にデータセンタ内で使用され得るアジャイル(ａｇｉｌｅ)ネットワークアーキテクチャに関する。１つの実施例は、レイヤ３のインフラストラクチャ（インフラ）のサーバ等のマシンを接続する仮想レイヤ２のネットワークを提供する。
【０００４】
他の実施例は複数のスイッチを介して通信可能に接続された複数の計算装置を含んでいる。個々の計算装置をアプリケーションアドレスで関連付けすることができる。１つの計算装置をソースとして動作するように設定することができ、他の計算装置を送り先として動作するように設定することができる。パケットを送り先計算装置のアプリケーションアドレスに送信するようにソース計算装置を設定することができる。この実施例は、パケットを捕まえ、送り先計算装置と関連付けられたロケーションアドレスを識別し、そのロケーションアドレスにパケットを送信するために個々のスイッチを選択するように設定されるアジャイルエージェントを含むことができる。
【０００５】
上記した各実施例は紹介目的のために提供され、請求項に記載した構成の全てを包含及び／又は限定しない。
【図面の簡単な説明】
【０００６】
添付図面は本願で伝えられるコンセプトの実施例を示している。添付図面と関連した次の詳細な説明を参照することにより図示の実施例の特徴を更に容易に理解することができる。各図面の同類の参照番号は実現可能であれば同類の要素を示すために用いられている。更に、各参照番号の最も左の数字は図面と、参照番号が最初に紹介されている関連記載とを表す。
【０００７】
【図１】本コンセプトのいくつかの実施に応じたアジャイルネットワークアーキテクチャの例を示す図である。
【図２】本コンセプトのいくつかの実施に応じたアジャイルネットワークアーキテクチャの例を示す図である。
【図３】本コンセプトのいくつかの実施に応じたアジャイルネットワークアーキテクチャ

の例を示す図である。
【図４】本コンセプトのいくつかの実施に応じたアジャイルネットワークアーキテクチャの例を示す図である。
【図５】本コンセプトのいくつかの実施に応じたアジャイルネットワークアーキテクチャの例を示す図である。
【図６】本コンセプトのいくつかの実施に応じたアジャイルネットワークアーキテクチャの例を示す図である。
【図７】本コンセプトのいくつかの実施に応じたアジャイルネットワークデータセンタレイアウトの例を示す図である。
【図８】本コンセプトのいくつかの実施に応じたアジャイルネットワークデータセンタレイアウトの例を示す図である。
【図９】本コンセプトのいくつかの実施に応じたアジャイルネットワークデータセンタレイアウトの例を示す図である。
【図１０】本コンセプトのいくつかの実施に応じて達成可能なアジャイルネットワークのフローチャート図である。
【発明を実施するための形態】
【０００８】
（概要）
本特許出願は特にデータセンタ内で使用され得るアジャイルネットワークアーキテクチャに関する。クラウドサービスは潜在的に１０から何十万ものサーバを保有する大規模なデータセンタの構築を押し進めている。それらのデータセンタは多くの動的な数の異なるサービス（Ｗｅｂアプリケーション、電子メール、マップリデュースクラスタ等）を同時にサポートすることができる。クラウドサービスデータセンタの実施例はスケールアウト設計：必要に応じてサービス間で急いで再割り当てされ得るリソース（例えば、サーバ）の大きなプールを介して達成される信頼性及び性能に依存することができる。データネットワークのいずれかサーバに何らかのサービスを割り当てるための能力にはデータセンタネットワークのアジリティを考慮することができる。大きなコストと関連付けられ得るデータセンタの利益を効果的に引き出すためには、ネットワークアジリティは高価になる可能性がある。ネットワークアジリティなしではデータセンタサーバリソースは立ち往生となり、よって無駄なお金となる。
【０００９】
（第１のアジャイルネットワークアーキテクチャ例）
紹介目的のためにアジャイルネットワークアーキテクチャ１００を例示する図１及び図２が参照される。アジャイルネットワークアーキテクチャ１００はサーバ１０２（１）、１０２（２）、１０２（３）及び１０２（Ｎ）等の複数のサーバサイド計算装置を含むことができる。
【００１０】
用語サーバ及びマシンはデータを送信又は受信することができるいずれかの装置に言及していることを理解されるべきである。例えば、それらの用語は物理サーバ、（例えば、仮想化技術を用いて）サーバ上で動作する仮想マシン、１つのオペレーティングシステムを動作する計算装置、１以上のオペレーティングシステムを動作する計算装置、異なるオペレーティングシステム（例えば、マイクロソフトウインドウズ（登録商標）、Ｌｉｎｕｘ、ＦｒｅｅＢＳＤ）を動作する計算装置、サーバ以外の計算装置（例えば、ラップトップ、アドレサブルパワーサプライ(ａｄｄｒｅｓｓａｂｌｅ　ｐｏｗｅｒ　ｓｕｐｐｌｙ)）、又は計算装置の一部（例えば、ネットワーク接続ディスク、ネットワーク接続メモリ、記憶サブシステム、記憶エリアネットワーク（ＳＡＮ）、グラフィックス処理ユニット、計算アクセラレータ、量子計算装置）のいずれかに言及していることを理解されるべきである。
【００１１】
アジャイルネットワークアーキテクチャ１００はサーバの数に関連してスケラビリティ

（拡張性）を進展させることができる。スケラビリティを達成することができる１つの方法は、アプリケーションを利用してサーバ１０２（１）～１０２（Ｎ）に対するイーサネットのようなフラットアドレッシングを作り出すことによって達成され得る。イーサネットのレイヤ２セマンティクスは、フラットアドレッシングをサポートするネットワーク状態を達成することと関連付けることができ、フラットアドレッシングはサーバがＬＡＮ（ローカルエリアネットワーク）上に存在したように、いずれかのインターネットプロトコル（ＩＰ）アドレスをいずれかのネットワークポートに接続されたいずれかのサーバに割り当てることができる。

【００１２】

この場合において、アプリケーションアドレス（ＡＡ）１０４（１），１０４（２），１０４（３），１０４（Ｎ）を各サーバ１０２（１）、１０２（２）、１０２（３），１０２（Ｎ）に各々割り当てていることができる。サーバの観点からすれば、いずれかのサーバ関連付けされたアプリケーションアドレス１０４（１），１０４（２），１０４（３），１０４（Ｎ）を介していずれかの他のサーバと通信することができる。これは何らかの様式でアプリケーションアドレスを配置することができる場合に、サーバ１０２（１）、１０２（２）、１０２（３），１０２（Ｎ）を備えるＬＡＮに対して有効であるそれら全てを含む、レイヤ２の機能として考えることができる。しかしながら、以下に説明されるように、いくつかの実施例では、アジャイルネットワークアーキテクチャの根本のインフラは１０６で示されるようにレイヤ３であっても良い。よって、それらの実施例は、レイヤ３のインフラ１０６上に（又は利用して）仮想レイヤ２のネットワーク１０８を作り出すことができる。それらは同一のレイヤ３のインフラ１０６上に作り出した１以上の仮想レイヤ２のネットワーク１０８であっても良く、また各サーバはそれらの１以上の仮想レイヤ２のネットワーク１０８に属することができる。

【００１３】

図２はアジャイルネットワークアーキテクチャ１００にインターネット２０４を介して接続されている外部クライアント２０２を示している。アジャイルネットワークアーキテクチャ１００は、アプリケーションアドレス１０４（１）～１０４（Ｎ）の知識を有する外部クライアントなしで１以上のサーバ１０２（１）～１０２（Ｎ）に割り当てられているグローバル又はロケーションアドレス２０６と外部クライアントが通信することを可能にする。それらのコンセプトは図３～図５の説明において以下に詳細に説明される。

【００１４】

（第２のアジャイルネットワークアーキテクチャ例）

図３は上記のコンセプトを実施可能なアジャイルネットワークアーキテクチャ３００を例示している。この場合において、外部クライアント３０２はアジャイルシステム３０４とインターネット３０６及び／又は他のネットワークを介して通信することができる。この実施例では、アジャイルシステム３０４は、３０８で概して示された複数のルータ３０８（１）～３０８（Ｎ）、３１０で概して示された複数の中間スイッチ３１０（１），３１０（２）及び３１０（Ｎ）、３１２で概して示された複数の集合スイッチ３１２（１），３１２（２）及び３１２（Ｎ）、３１４で概して示された複数のトップオブラック（ＴＯＲ又はＴｏＲ）スイッチ３１４（１），３１４（２）及び３１４（Ｎ）、並びに３１６で概して示された複数のサーバ３１６（１），３１６（２），３１６（３），３１６（４），３１６（５）及び３１６（Ｎ）を含んでいる。図面のページの空間制限のため、６つのサーバ３１６（１）～３１６（Ｎ）だけがここでは示されているが、アジャイルシステム３０４は、何千、何万、何十万、又はそれ以上のサーバを提供することができる。なお、簡潔及び図面のページの空間制限のため、コンポーネント間の接続（すなわち、通信パス）は図３～図８に示されているとは限らない。

【００１５】

サーバ３１６（１）及び３１６（２）はサーバラック３１８（１）としてＴＯＲスイッチ３１４（１）と関連付けられている。同様に、サーバ３１６（３）及び３１６（４）はサーバラック３１８（２）としてＴＯＲスイッチ３１４（２）と関連付けられ、サーバ３

１６（５）及び３１６（Ｎ）はサーバラック３１８（Ｎ）としてＴＯＲスイッチ３１４（Ｎ）と関連付けられている。これは上記したように図面ページの空間制限のためであり、度々、サーバラックは１０以上のサーバを含む。更に、個々のサーバはアジャイルエージェントと関連付け可能である。例えば、サーバ３１６（１）はアジャイルエージェント３２０（１）と関連付けられる。同様の関係はサーバ３１６（２）～３１６（Ｎ）とアジャイルエージェント３２０（２）～３２０（Ｎ）との間に各々示される。

【００１６】

アジャイルエージェント３２０（１）～３２０（Ｎ）の機能は以下に詳細に説明される。短くは、アジャイルエージェントは個々のサーバ間での通信を容易にすることができる。特にこの例では、アジャイルエージェントはコンピュータ読み取り可能な命令としてサーバ上に保存された論理モジュールとして考えても良い。他の実施例は、サーバ群を提供するアジャイルエージェント３２０がスイッチ、例えば、ＴＯＲスイッチ又は中間スイッチに配置される構成を含んでも良い。スイッチに配置されたときには、アジャイルエージェントは、パケットがサーバ３１６から中間スイッチ３１０に向けてネットワークを逆送するようにパケットを処理する。そのような構成では、アジャイルエージェント３２０はパケット転送パス上のカスタムハードウエアと、その転送パス又はスイッチの制御プロセッサで実施するソフトウエア命令との組み合わせを用いて実施されても良い。

【００１７】

アジャイルシステム３０４は３つのディレクトリサービスモジュール３２２（１）～３２２（Ｎ）を更に含んでいる。ディレクトリサービスモジュールの図示の数はそのアジャイルシステムに必須ではなく、他の実施例は２、３以上のディレクトリサービスモジュール（及び／又は他の図示のコンポーネント）を使用することができる。ディレクトリサーバの機能は以下に詳細に説明される。短くは、ディレクトリサービスモジュールは、他の情報のうちの、アジャイルエージェント３２０（１）～３２０（Ｎ）（及び／又は他のコンポーネント）によって利用され得るアプリケーションアドレス対ロケーションアドレスマッピング（順方向又は逆方向のマッピングのいずれか一方又は両方）を含むことができ、これによりアジャイルシステム３０４内の通信を容易にする。この場合において、ディレクトリサービスモジュール３２２（１）～３２２（Ｎ）は特定のサーバ３１６（１）、３１６（３）及び３１６（５）と関連付けされる。他の設定では、そのディレクトリサービスモジュールは、データセンタ制御サーバ、スイッチ等の他のコンポーネントと共に、及び／又は専用の計算装置上で存在することができる。

【００１８】

アジャイルシステム３０４は２つの論理グループ化として考えても良い。第１の論理グループ化は３２６で示されたようにロケーション又はグローバルアドレスを担うリンクステートネットワークである。第２の論理グループ化は３２８で示されたようにアプリケーションアドレスを所有するサーバの代替可能なプールである。短くは、リンクステートネットワーク３２６のコンポーネントは、サーバ３２８のプールのどのサーバがどのアプリケーションアドレスを現在用いているかを追跡するために情報を交換することを必要としない。また、サーバの見地からは、サーバはサーバプール３２８内のいずれかの他のサーバと他のサーバのアプリケーションアドレスを介して通信することができる。この処理は、サーバに対して透過的であるような方式でアジャイルエージェント、ディレクトリサービス、及び／又は他のコンポーネントによって容易にされる。他の方法としては、その処理は、サーバ上の他のコンポーネントがその処理に気づくかもしれないけれども、サーバ上で動作しているアプリケーションに対して透過的であっても良い。

【００１９】

ルータ３０８、中間スイッチ３１０、集合スイッチ３１２、ＴＯＲスイッチ３１４及びサーバ３１６（１）～３１６（Ｎ）は、レイヤ３の技術を用いる等で通信可能に接続されても良い。個々のサーバの見地からは、他のサーバとの通信はレイヤ２の通信（例えば、仮想レイヤ２）として現れる。しかしながら、サーバラック３１８（１）のソースサーバ３１６（１）からサーバラック３１８（２）の送り先サーバ３１６（３）へのようなラッ

ク間通信はレイヤ３のインフラを介して実際に起きる。例えば、アジャイルエージェント３２０（１）は通信（例えば、サーバ３１６（３）のアプリケーションアドレスにアドレス指定されたパケット）を捕らえ、その伝送を容易にすることができる。

【００２０】

アジャイルエージェント３２０（１）はディレクトリサービスモジュール３２２（１）～３２２（Ｎ）のうちの１以上にアクセスして、サーバ３１６（３）で関連付けられたロケーションアドレスに対するアプリケーションアドレスのマッピングを得ることができる。例えば、そのマップ化したロケーションアドレスはＴＯＲスイッチに対してでも良い。アジャイルエージェントはロケーションアドレスを有するパケットをカプセル化することができる。そのアジャイルエージェントはそのカプセル化パケットを送信又は中継するように個々（又はセット）の集合及び／又は中間スイッチを選択することができる。この選択処理の機能は以下に更に詳細に説明される。ＴＯＲスイッチ３１４（２）でそのカプセル化パケットを受け取ると、そのＴＯＲスイッチはそのパケットをデカプセル化してそのパケットをサーバ３１６（３）に送信することができる。代替実施形態では、ロケーションアドレスはサーバ３１６（３）又はサーバ３１６（３）上で動作する仮想マシンと関連付けされても良く、パケットは送り先サーバそれ自身上でデカプセル化されても良い。それらの実施形態において、アプリケーションアドレスが他のホストがアプリケーションとの通信のために使用しているアドレスであるＬＡＮによって接続されているアプリケーションに対して示し続けるために、サーバ又は仮想マシンに割り当てられたロケーションアドレスはサーバ上で動作する他のアプリケーションから隠されても良い。

【００２１】

代替の実施形態において、パケットはレイヤ３／レイヤ２の境界を横切る際に他のコンポーネントによってカプセル化されても良い。例えば、デカプセル化を行うことができるコンポーネントの例はハイパーバイザ及び／又は仮想マシンモニタのルートパーティションを含んでも良い。

【００２２】

この設定はサーバがサーバプール３２８の多数の中に加えられることを可能にし、また、サーバの見地からは、他のサーバはそれらが同一のサブネットワーク上にあるように現れることができる。代替的に又は追加的に、リンクステートネットワーク３２６のコンポーネントはサーバアプリケーションアドレスを気にする必要がない。更に、サーバが追加又は削除された時等のアドレス情報が変化したときはいつでもディレクトリサーバは多数の異なる種類のコンポーネントを更新するよりもむしろ単に更新されても良い。

【００２３】

要約すれば、レイヤ２のセマンティックは、サーバがＬＡＮ上に存在したように、いずれかのネットワークポートに接続されたいずれかのサーバにいずれかのＩＰアドレスを割り当てることができるフラットアドレッシングをサポートするネットワーク状態を達成することと関連付けられても良い。また、リンクステートネットワーク３２６内のコンポーネント（すなわち、スイッチ）はそのリンクステートネットワーク内の他のコンポーネントに気づいても良いが、サーバプール３２８のコンポーネントに気づく必要はない。更に、ＴＯＲスイッチはそれら各々のラック内のサーバについて知っても良いが、他のラックのサーバについては知る必要がない。更に、アジャイルエージェントはサーバアプリケーションアドレス（ＡＡ）パケットを捕らえて、ＡＡの送り先計算装置と関連付けされたロケーションアドレス（ＬＡ）を識別することができる。そして、アジャイルエージェントはＬＡにパケットを送信するために個々のスイッチ（又はスイッチのセット）を選択することができる。この場合において、その個々のスイッチは利用可能なスイッチの１以上を含むことができる。

【００２４】

また、この設定はサービスに関係した他のサーバの機能を促進する。例えば、ディレクトリサービスモジュール３２２（１）～３２２（Ｎ）に含まれても良いようなデータセンタ管理ソフトウエアはいずれかのサービスにいずれかのサーバ３１６（１）～３１６（Ｎ

）を割り当て、そのサービスが期待するどのようなＩＰアドレスを有するサーバでも設定することができる。各サーバのネットワーク設定はＬＡＮを介して接続された場合と同一とすることができ、リンクローカルブロードキャストのような機能をサポートすることができる。サービス間の通信分離の目的は、サービス及び通信グループを定義する簡単で一貫したアプリケーションプログラムインターフェース（ＡＰＩ）を提供することと関連付けられても良い。この点において、ディレクトリサービスはサービス（例えば、顧客）と関連付けられたサーバのグループを定義することができる。１つのグループ内のサーバ間での完全接続を許可しても良く、異なるグループのどのサーバが通信することを許可されるべきかを管理するためにアクセスコントロールリスト（ＡＣＬ）のような方針を規定しても良い。

【００２５】

更に、上記の設定はトラフィック管理にそれ自身役立つ。説明の目的のために、第１の顧客がアジャイルシステム３０４のサーバによって実行されるべきサービスに対して比較的高いレートを支払い、それによってサービス契約の比較的高い品質を得ると仮定する。更に、第２の顧客が比較的低いレートを支払い、それによってサービス契約の比較的低い品質を受けると仮定する。そのような場合において、第１の顧客のためにトラフィックを運営するように中間スイッチ３１０（１）～３１０（Ｎ）の比較的高いパーセンテージ、又は全てを割り当てることができ、第２の顧客に対してより少ない数のスイッチを割り当てることができる。別の言い方をすれば、第１の顧客にスイッチの第１のサブセットを割り当てることができ、第２の顧客にスイッチの第２のサブセットを割り当てることができる。第１のセット及び第２のセットを互いに排他的に又は重複させても良い。例えば、いくつかの実施例では、個々のスイッチを特定顧客に対して専用に、又は複数の顧客に割り当てることができる。例えば、中間スイッチ３１０（１）を両方の顧客に割り当てることができ、中間スイッチ３１０（２）及び３１０（Ｎ）を排他的に第１の顧客に割り当てることができる。

【００２６】

要約すると、以下に詳細に説明されるように、アジャイルネットワークアーキテクチャ３００は次の１以上の目的、すなわちサーバ間の一様な高い能力、サービス間のパフォーマンス分離、イーサネットレイヤ２のセマンティック、及び／又はサービス間の通信分離に関連付けられても良い。サーバ間の一様な高い能力の目的は、ネットワークを流れるトラフィックの割合が、先ず、送信するサーバ及び受信するサーバのネットワークインターフェースカードの利用可能な能力による場合を除いて、制限されないネットワーク状態を達成することと関連付けられても良い。よって、開発者の見地からは、その目的を達成することによって、ネットワークトポロジは、サーバを１つのサービスに加えるときにはもはや第一の関心事ではないかもしれない。サービス間のパフォーマンス分離の目的は、各サービスが個別の物理スイッチによって接続されたかのように、１つのサービスのトラフィックが何らかの他のサービスによって運営されたトラフィックによって影響を受けないネットワーク状態を達成することと関連付けられても良い。イーサネットレイヤ２のセマンティックの目的は、サーバがＬＡＮ上に存在したように、ほとんどのＩＰアドレスが何らかのネットワークポートに接続されたいずれかのサーバに割り当てられ得るフラットアドレッシングをサポートするネットワーク状態を達成することと関連付けられても良い。よって、データセンタ管理ソフトウエアはいずれかのサーバを何らかのサービスに割り当て、サービスが期待するどんなＩＰアドレスを有するサーバでも設定することができる。

【００２７】

各サーバのネットワーク設定はＬＡＮを介して接続された場合と同一とすることができ、リンクローカルブロードキャストのような機能をサポートすることができる。サービス間の通信分離の目的は、サービス及び通信グループを定義する簡単で一貫したＡＰＩを提供することと関連付けられても良い。この点において、サーバのグループを定義するディレクトリシステム（すなわち、例えば、ディレクトリサービスモジュール３２２（１）～３２２（Ｎ）を介して）は提供されても良い。１つのグループ内のサーバ間での完全接続

を許可しても良く、異なるグループのどのサーバが通信することを許可されるべきかを管理するために方針を規定しても良い。

【００２８】

説明したアジャイルネットワークアーキテクチャを利用することによって、次のネットワークの特徴の１以上と関連付けられているデータセンタネットワークを備えることができる。（１）サービスインスタンスがネットワークのどこにでも配置されることを可能にするフラットアドレッシング、（２）ネットワークパスを越えて一様にトラフィックを分散するようにランダム化を使用するロードバランシング（例えば、バリアントロードバランシング（ＶＬＢ：Ｖａｌｉａｎｔ　Ｌｏａｄ　Ｂａｌａｎｃｉｎｇ））、（３）イーサネットレイヤ２のセマンティックを達成し、大きなサーバプールへスケーリングを行う新しいエンドシステムベースのアドレス解決サービス。

【００２９】

上記した目的を達成するために、少なくともいくつかの実施形態において、次のアジャイルネットワークアーキテクチャの設計原理の１以上を様々な実施例で用いることができる。

【００３０】

（多数のパスダイバシティを有するトポロジの利用）

「メッシィ（ｍｅｓｈｙ)」トポロジを利用することにより、サーバの個々のセット間の多数のパスを提供することができる。例えば、サーバラック３１８（１）の複数のサーバとサーバラック３１８（Ｎ）の複数のサーバとの間の通信はＴＯＲスイッチ３１４（１）から集合スイッチ３１２（１）～３１２（２）のいずれかを介して中間スイッチ３１０（１）～３１０（Ｎ）のいずれかに達することができる。中間スイッチから、その通信は集合スイッチ３１２（２）～３１２（Ｎ）のいずれかを介してＴＯＲスイッチ３１４（Ｎ）に達することができる。

【００３１】

この設定はいくつかの利益を得る結果にすることができる。例えば、多数のパスの存在は、明白なトラフィックエンジニアリング又はパラメータのチューニングの必要性なしにネットワークの輻輳状態の減少及び／又は解消を可能にする。更に、多数のパスは「スケールアウト(ｓｃａｌｅ－ｏｕｔ)」ネットワーク設計を可能にさせる。言い換えると、より低コストのスイッチを加えることにより更なる能力を加えることができる。逆に、従来の階層的なネットワーク設計は階層の高いレベルで１又は非常に僅かのリンクのトラフィックを集中させる。その結果、従来のネットワークはトラフィックの高い密度に対処するために高価な「ビッグアイロン(ｂｉｇ　ｉｒｏｎ：大型コンピュータ)」スイッチの購入を必要とする可能性がある。

【００３２】

更に、「メッシィ」トポロジを利用することにより、多数のパスはリンク又はスイッチが機能しなくなったとき正常な劣化を可能にさせる。例えば、所定のレイヤで「ｎ」のスイッチを有する説明したアジャイルデータセンタネットワークアーキテクチャに応じて実施されたアジャイルネットワークは、スイッチが機能しなくなったとき、その能力の５０％を失う可能性がある従来のネットワークと比較して、その能力の１／ｎだけ失う可能性がある。アジャイルデータネットワークアーキテクチャに応じて実施されたアジャイルネットワークは、完全な二分トポロジを潜在的に利用することができる。

【００３３】

（アドレス不安定さに対するランダム化）

データセンタはそれらの負荷、それらのトラフィック、及びそれらの故障パターンにおいて非常に大きい不安定さを持つ可能性がある。よって、リソースの複数の大きなプールを作り出すことができる。負荷をランダムにそれらに分散することができ、最悪の場合を平均的な場合に改善するために、最良の場合のいくつかの性能を、引き換えにすることができる。少なくともいくつかの実施形態において、広範囲のパスダイバシティと関連付けられたトポロジ（例えば、図３に明らかにされたように）が利用されても良い。ＶＬＢ技

術等のロードバランシング技術を用いてトポロジを越えてワークフローのルート決めすることができる。短くは、ＶＬＢ技術はデータ伝送を担うために用いられるパス又は複数のパスをランダムに選択することを含むことができ、ここでパスは一連のリンク及び／又はスイッチからなる。続いて、パスを再選択し、ここで再選択は初期のパスを構成するスイッチ又はリンクの１以上を変えることを必要とする。その再選択は、特定のバイト／パケット数を送信/受信した後に、及び／又は選択したパス、スイッチ又はリンクと関連付けられた伝送状況が生じたことに応答するように周期的に起きても良い。例えば、パケット遅延又は他の通信障害が検出されたならば、その選択処理が繰り返されても良い。この原理のアプリケーションを通して、一様な能力及びパフォーマンス分離目的を合致することができる。

【００３４】

特に、データセンタトラフィックマトリックスにおけるアドレスの不安定さ及び不確かさに対して、ネットワークパス間でフローをランダムにハッシュ（ｈａｓｈ）するためにロードバランシング技術（例えば、ＶＬＢ）を利用することができる。このアプローチに対する目的は、ホーストラフィックモデルにおいてのように、ネットワークの入出制限の影響下にある任意のトラフィック変動のための帯域幅保証を提供することである。短くは、ホースモデルは、所定のパスに亘るデータ伝送レートがそのパスの最も遅い又は最も規制された部分を越えることができないことを特定する。

【００３５】

フロー単位（フローの大部分のパケットがパスを再選択したときを除いてネットワークを介して同一のパスに続いていることを意味する）で、ＶＢＬのようなロードバランシング技術を用いることは、有利である可能性がある。それは、フローのパケットが並べ直しされる機会、又は送り先で判断された待ち時間を急に変更する体験、及び／又はフロー内のＭＴＵ(Ｍａｘｉｍｕｍ　Ｔｒａｎｓｍｉｓｓｉｏｎ　Ｕｎｉｔ)の違いのためにパスＭＴＵ発見プロトコルの混乱動作を減少することができるからである。トラフィックのいくつかの種類（例えば、パケットの並べ直しによって悪影響を与えない種類）及びいくつかの環境（例えば、全てのパスに亘って非常に均一な遅延を有する環境）は、パケット精度（潜在的な異なるパスがパケットのシーケンスにおけるパケット毎に使用されることを意味する）でＶＬＢのようなロードバランシングを用いることを好むかもしれない。フローの共通に受け入れられた定義のいくつかを用いることができる。例えば、ＩＰ５タプルフロー（５－ｔｕｐｌｅ　ｆｌｏｗ)、ＩＰ２タプルフロー（２－ｔｕｐｌｅ　ｆｌｏｗ)、又は２つのサブネット又はアドレスレンジ間のパケットのセット。

【００３６】

アジャイルデータセンタネットワークを提供するコンテキストにおいて、入出制限はサーバラインカード速度に対応することができる。ハイバイセクション帯域幅トポロジ（ｈｉｇｈ　ｂｉｓｅｃｔｉｏｎ　ｂａｎｄｗｉｄｔｈ　ｔｏｐｏｌｏｇｙ）（例えば、折り畳みＣｌｏｓトポロジ）との組み合わせにおいて、非干渉パケットスイッチ切替ネットワーク（非ブロッキング回路スイッチ切替ネットワークの対応）を作り出し、サーバ入出ポート速度を越える負荷を維持していない）トラフィックパターンのためのホットスポットフリーパフォーマンスを提供するためにロードバランシング技術を利用することができる。この点において、いくつかの実施例では、ＴＣＰ(Ｔｒａｎｓｍｉｓｓｉｏｎ　Ｃｏｎｔｒｏｌ　Ｐｒｏｔｏｃｏｌ)のエンドツーエンド輻輳制御機構はホースモデルを実行して過動作サーバポート速度を避けるために利用されても良い。この原理は図３に示された論理トポロジに導入することができ、それは３つの異なるスイッチ層、ＴＯＲ３１４、集合３１２及び中間３１０から構成することができる。１つのサーバから他のサーバへのフローは、ランダム中間スイッチを介し、複数のＴＯＲ及び集合スイッチを越えてランダムパスをとることができる。ＶＬＢ等のロードバランシング技術は、利用を円滑にして持続するトラフィック輻輳を解消するためにデータセンタの中間スイッチ構造のコンテキストに利用されても良い。

【００３７】

（ロケーションからの名称分離）

ロケーションから名称を分離することは、新しい機能を実施するために用いられ得る自由度を作り出すことができる。この原理はデータセンタネットワークにおいてアジリティを可能にし、アドレスとロケーションとの間を結びつけることが起き得るフラグメンテーションを減少することによって利用を改善するために利用することができる。この原理のアプリケーション及び以下に説明されるエンドシステムを包括する原理を通して、レイヤ２のセマンティクス目的を合致することができる。よって、ネットワークトポロジに関係なく、またそれらのアプリケーション又はネットワークスイッチを再設定することなくＩＰアドレスを割り当てることが開発者に可能になる。

【００３８】

ネットワークアジリティ（サーバ上のサービスをサポートすること、サーバプールの動的成長及び縮小、及び負荷マイグレーション）を拡張するために、ＡＡと称した名称及びＬＡと称したロケーションを分離するＩＰアドレッシングスキームを用いることができる。ディレクトリサービスモジュール３２２（１）～３２２（Ｎ）として示されたようなアジャイルディレクトリサービスは、拡張性及び信頼性ある方法でＡＡとＬＡとの間のマッピングを管理するために定義され得る。アジャイルディレクトリサービスは、個々のサーバ上でのネットワーキングスタックにおいて動作するシム層によって含まれても良い。図３に表された実施例では、このシム層をアジャイルエージェント３２０（１）～３２０（Ｎ）として示している。

【００３９】

（エンドシステムの包括）

データセンタサーバ上のオペレーティングシステムを含むソフトウエアは、通常、データセンタ内での利用のために広範囲に修正される。例えば、新しい又は修正したソフトウエアは仮想化のためのハイパーバイザ又は複数のサーバに亘ってデータを保存するためにブロブファイルシステムを作り出すことができる。スイッチ上のソフトウエアを変更するよりむしろ、このソフトウエアのプログラム可能性を利用することができる。更に、スイッチ又はサーバのハードウエアへの変更を避ける又は制限することができ、使われていないアプリケーションが修正されずに残すことができる。現在利用可能な低コストのスイッチＡＳＩＣ(ａｐｐｌｉｃａｔｉｏｎ ｓｐｅｃｉｆｉｃ ｉｎｔｅｇｒａｔｅｄ ｃｉｒｃｕｉｔ)の制限内で動作するようにサーバ上でソフトウエアを用いることによって、今日作り開発され得る設計を作り出すことができる。例えば、ＡＲＰ(Ａｄｄｒｅｓｓ Ｒｅｓｏｌｕｔｉｏｎ Ｐｒｏｔｏｃｏｌ)パケットのブロードキャストにより作り出される拡張性の問題は、スイッチ上でソフトウエア又はハンドウエアの変更を介してＡＲＰを制御する試みよりむしろ、サーバ上のＡＲＰリクエストを捕らえ、それらをディレクトリシステムに対してルックアップリクエストに変換することによって、減少及び／又は解消され得る。

【００４０】

図４はアジャイルエージェント３２０（１）を詳細に例示している。この場合において、アジャイルエージェント３２０（１）はユーザモード４０６及びカーネルモード４０８を含むサーバマシン４０２（１）上で動作する。そのサーバマシンはユーザモードでユーザモードエージェント４１０を含む。カーネルモードはＴＣＰコンポーネント４１２、ＩＰコンポーネント４１４、エンカプスレータ４１６、ＮＩＣ４１８及びルーティング情報キャッシュ４２０を含む。サーバマシンはディレクトリサービス３２２（１）を含む、及び／又はそれと通信することができる。ディレクトリサービスはサーバロールコンポーネント４２２、サーバヘルスコンポーネント４２４、及びネットワークヘルスコンポーネント４２６を含むことができる。アジャイルエージェント３２０（１）はユーザモードエージェント４１０、エンカプスレータ４１６、及びルーティング情報キャッシュ４２０を含んでも良い。エンカプスレータ４１６はＡＲＰを捕らえ、それをユーザモードエージェント４１０に送信することができる。ユーザモードエージェントはディレクトリサービス３２２（１）に問い合わせすることができる。カーネルモードコンポーネントにユーザモー

ドエージェントを含むように、又はルーティングテーブルルックアップの期間のようなＡＲＰ以外の機構を介して又はＩＰテーブル又はＩＰチェイン等の機構を介してディレクトリルックアップを含むように、それらのブロックの他の配置が可能であることが理解されるべきである。

【００４１】

図３のアジャイルネットワークアーキテクチャにおいて、エンドシステム制御は新しい機能を速く注入する機構を提供することができる。よって、アジャイルエージェントはロードバランシングに使用されるランダム化を制御することによってきめ細かいパス制御を提供することができる。加えて、名称及びロケーションの分離を実現するために、アジャイルエージェントはイーサネットのＡＲＰ機能をアジャイルディレクトリサービスへの問い合わせ（クエリ）で置き換えることができる。アジャイルディレクトリサービス自身は、スイッチよりむしろ、サーバ上に実現され得る。このアジャイルディレクトリサービスは、サーバ到達可能性、グループ化、アクセス制御、リソース配分（例えば、中間スイッチの能力）、分離（例えば、中間スイッチの非重複化）、及び動的成長及び縮小のきめ細かい制御を可能にする。

【００４２】

（ネットワーク技術の利用）

ネットワークスイッチにおけるロバスト実装を有する１以上のネットワーク技術を利用することは、アジャイルネットワークの設計を簡単にし、そのようなネットワークを開発するためにオペレータ意欲を増加させることができる。例えば、少なくともいくつかの実施形態において、リンクステートルーティングプロトコルは、サーバから障害を隠すためにネットワークスイッチ上に実装されても良く、また、アジャイルディレクトリサービスでの負荷を減少させる手助けに利用され得る。それらのプロトコルはアジャイルネットワークのためにトポロジ及びルートを維持するために利用されても良く、それはアジャイルディレクトリサービスとネットワーク制御プレーンとの間の結合を減少させることができる。スイッチ上のエニキャスト(ａｎｙｃａｓｔ)アドレスを定義するルーティング設計を通して、上記したアジャイルアーキテクチャはサーバからスイッチの障害を隠すためにＥＣＭＰ(Ｅｑｕａｌ　Ｃｏｓｔ　Ｍｕｌｔｉ－Ｐａｔｈ)を利用することができる。更に、これはディレクトリシステム上の負荷を減少させることができる。また、多数のパスの使用をサポートする他のルーティングプロトコルは適切である。

【００４３】

（仮想レイヤ２のネットワーク例に関する実施詳細）

（トポロジのスケールアウト）

従来のネットワークは、典型的に、ネットワークの最高レベルでトラフィックを少しのスイッチに集中させる。これはそれらの装置の能力に二等分の帯域幅を両方制限し、それらが機能しなくなったときネットワークに影響をかなり与える可能性がある。しかしながら、それらの問題を避けるために、トラフィックの不安定さに対処するためにランダム化を用いる原理によって運営されるアジャイルネットワークトポロジを利用することができる。この点について、ネットワーク装置をスケールアウトするアプローチをとることができる。これは、図３に示されたように、早送りに専門化され得る低複雑スイッチの比較的ブロードなネットワークに、結果的になる可能性がある。これは、中間スイッチ３１０（１）～３１０（Ｎ）と集合スイッチ３１２（１）～３１２（Ｎ）との間のリンクが完全な二分グラフを形成することができる折り畳みＣｌｏｓネットワークの例である。従来のトポロジにあるように、ＴＯＲは２つの集合スイッチに接続することができる。しかしながら、いずれかの２つの集合スイッチの間の多数のパスは、ｎの中間スイッチがあるならば、それらのいずれかの故障が１／ｎだけ－帯域幅の正常な低下と称され得る望ましいプロパティ、二分帯域幅を減少させることを意味する。更に、オーバサブスクリプション(ｏｖｅｒｓｕｂｓｃｒｉｐｔｉｏｎ)がないようにクロネットワーク等のネットワークを設計することができる。例えば、図３において、Ｄの数のインターフェースポートを有する集合スイッチ及び中間スイッチを用いることができる。それらのスイッチは、スイッチの

各層間の能力がリンク能力のＤ＊Ｄ／２倍であるように接続され得る。

【００４４】

ネットワークの上段又は「スパイン（ｓｐｉｎｅ)」で中間スイッチを介して戻ってくることによってネットワークがサーバラインカードで入出限度の影響下にある潜在的に全ての可能なトラフィックマトリックスに対する帯域幅保証を提供することができることにおいて、ロードバランシング（例えば、ＶＬＢ）に対してＣｌｏｓネットワークのようなネットワークは格別によく適合されても良い。ルーティングは単純で回復力（例えば、ランダムパスはランダム中間ノード及び取り払われたランダムパスを必要としても良い）を持っていて良い。

【００４５】

上記したアジャイルアーキテクチャは従来のネットワークアーキテクチャで達成され得るよりも大きなパス制御を提供することができる。特に、複数の中間ノードを区分することができ、異なる区分に専用化されたトラフィッククラスは高い帯域幅全体をいくつかのトラフィッククラスに割り当てる。輻輳指示は、ＩＥＥＥ(Ｉｎｓｔｉｔｕｔｅ ｏｆ Ｅｌｅｃｔｒｉｃａｌ ａｎｄ Ｅｌｅｃｔｒｏｎｉｃｓ Ｅｎｇｉｎｅｅｒｓ)８０２．１Ｑａｕ輻輳制御におけるように、ＥＣＮ(Ｅｘｐｌｉｃｉｔ Ｃｏｎｇｅｓｔｉｏｎ Ｎｏｔｉｆｉｃａｔｉｏｎ)又は類似の機構を介して送り元に信号で戻されても良い。よって、ＥＣＮ信号を集める送り元は、（上記のパスを再選択すると呼ばれた）ネットワークを介して代替パスを選択するために用いられるソースパケット内のフィールドを変えることによって応答することができる。

【００４６】

（アジャイルルーティング）

ロケーションから名称を分離する原理を実施するために、アジャイルネットワークは２つのＩＰアドレスファミリを用いることができる。図３はそのような分離を示している。ネットワークインフラは複数のＬＡに関して動作することができる。スイッチ及びインターフェース（３１０（１）～３１０（Ｎ），３１２（１）～３１２（Ｎ），及び３１４（１）～３１４（Ｎ））は割り当てられた複数のＬＡであることができる。スイッチはそれらのＬＡを担うリンクステートＩＰルーティングプロトコルを動作させることができる。

【００４７】

サーバ３１６（１）～３１６（Ｎ）上で動作するようなアプリケーションは複数のＬＡに気付かず、複数のＡＡに気付くことができる。この分離をいくつかの利益と関連付けることができる。第一に、パケットはＡＡに直接送信されるよりむしろ適当なＬＡに送られても良い（スイッチはそれらを送信するためにホスト毎にルーティングエントリを維持する必要がない）。これは複数のＡＡを複数のＬＡに変換するアジャイルディレクトリサービスが、どのサービスが通信のために許可されるべきかについての方針を実行することができることを意味している。第二に、低コストスイッチは度々、全てのＬＡルータを保持することができる小さなルーティングテーブル（例えば、１２Ｋエントリ）を有するが、ＡＡの数だけ負担が掛かる。このコンセプトは特に、スイッチが保持することができるルーティングエントリの数より多いネットワークが作られることを可能にすることにおいて価値がある可能性がある。第三に、いずれかのＡＡがトポロジとは関係なくいずれかのサーバに割り当てられ得るので分離はアジリティを可能にする。第四に、ＡＡから分離して複数のＬＡを割り当てる自由は、複数のＬＡがトポロジ的に重要な方法で要約され得るように割り当てられ、更に、スイッチが担う必要があるルーティング状態の量を制限することを意味している。一方、データセンタ内で動作しているサービス又はデータセンタのオペレータは、どんな方法においても、アプリケーションアドレスを割り当てる能力を妨げないことを望んでいる。

【００４８】

本発明の代替的な実施形態はＬＡ及びＡＡアドレスのための他の種類のデータを用いても良い。例えば、ＬＡアドレスはＩＰｖ４で可能であり、ＡＡアドレスはＩＰｖ６で可能であり、又は逆も同様である。又はＩＰｖ６はＡＡ及びＬＡの両方のアドレスのために使

用可能であり、又はＩＥＥＥ８０２．１ＭＡＣアドレスはＡＡアドレスとして使用可能である。一方、ＩＰアドレス（ｖ４又はｖ６）はＬＡアドレスのために用いられ、又は逆も同様である。また、アドレスは、例えば、ＩＰアドレスを有するＶＬＡＮタグ又はＶＲＦ識別子のように、異なる種類のアドレスと一緒に組み合わせることによって作り出されても良い。

【００４９】

次の記載は、下層のネットワーク構造を仮想化し、レイヤ２ＬＡＮ及びそれの上の何かのそれらのグループの他のサーバ３１６（１）～３１６（Ｎ）に接続され、ホストが比較的大きいデータセンタワイドレイヤ２ＬＡＮの一部であるアジャイルネットワークのサーバ３１６（１）～３１６（Ｎ）にイルージョンを作り出すために、トポロジ、ルーティング設計、アジャイルエージェント、及びアジャイルディレクトリサービスがどのように組み合わされるかを説明する。

【００５０】

（アドレス決定及びパケット転送）

少なくともいくつかの実施例では、サーバ３１６（１）～３１６（Ｎ）が信じることを可能にするためそれらは同一のサービスで他のサーバと１つの大きなＶＬＡＮを共有し、大きなイーサネットを悩ますブロードキャストＡＲＰスケーリング障害を解消し、以下に記載した解決策が提供される。なお、予め、次の解決策は下位互換があり、存在するデータセンタアプリケーションに対して透過であることが可能である。

【００５１】

（パケット転送）

ＡＡは通常、ネットワークのルーティングプロトコルに示されても良い。よって、パケットを受信するサーバのために、パケットのソースは先ず、そのパケットをカプセル化することができ、ホストのために外側のヘッダの送り先をＬＡに設定する。ＬＡアドレスを保持する装置に到達すると、パケットはデカプセル化され、送り先のサーバに配信される。１つの実施形態において、送り先のサーバのためのＬＡは、その送り先サーバが設置されている元のＴＯＲに割り当てられる。パケットがその送り先ＴＯＲに到着すると、通常のレイヤ２の配信規則によれば、そのＴＯＲスイッチがそのパケットをデカプセル化して内側のヘッダの送り先ＡＡに基づいてパケット配信することができる。代替的に、ＬＡは物理的送り先サーバ又はそのサーバ上で動作する仮想マシンと関連付けられても良い。

【００５２】

（アドレス決定）

複数のＡＡアドレスが複数のサーバと同一のＬＡＮに存在することを信じるようにその複数のサーバを設定することができ、それでアプリケーションが初めてＡＡにパケットを送信するときホスト上のカーネルネットワークスタックは送り先のＡＡのためのブロードキャストＡＲＰリクエストを生成することができる。ソースサーバのネットワーキングスタックで動作するアジャイルエージェントはそのＡＲＰリクエストを捕らえてそれをアジャイルディレクトリサービスへのユニキャスト問い合わせ（クエリ）に変換することができる。アジャイルディレクトリサービスがその問い合わせに答えるとき、アジャイルディレクトリサービスはパケットが送られるべきＬＡを提供することができる。また、アジャイルディレクトリサービスはパケットを中継するために用いられ得る中間スイッチ又は中間スイッチのセットを提供することができる。

【００５３】

（ディレクトリサービスによるインターサービスアクセス制御）

サーバは、ＡＡのためのパケットを送る必要がないＴＯＲのＬＡを得ることができないならば、パケットをＡＡに送信することができなくても良い。よって、アジャイルディレクトリサービス３２２（１）～３２２（Ｎ）は通信方針を実行することができる。アジャイルディレクトリサービスはルックアップリクエストを処理するとき、どのサーバがそのリクエストを作っているか、ソース及び送り先の両方が属するサービス、及びそれらのサービスの分離方針を知る。その方針が「拒否(ｄｅｎｙ)」であるならば、アジャイルディ

レクトリサービスはＬＡを提供することを単に拒否することができる。上記したアジャイルネットワークアーキテクチャの有利な点はインターサービス通信が許可されたとき、ＩＰゲートウエイに迂回されることなくパケットが送信サーバから受信サーバへ直接流れることができることである。これは、従来のアーキテクチャでの２つのＶＬＡＮの接続とは異なっている。

【００５４】

（インターネットとのインターラクション）

度々、データセンタによって処理されるトラフィックのほぼ２０％がインターネットへの又はインターネットからのトラフィックである。よって、データセンタネットワークがそれらの大きなボリュームを処理することができることは有利な点である。先ず、上記したアジャイルネットワークアーキテクチャが仮想レイヤ２ネットワークを実施するためにレイヤ３の構造を利用することは不思議に思えるかもしれないが、この１つの有利な点は、いくつかの従来提案されたネットワーク環境で要求されたようなヘッダの書き直しのためにゲートウエイサーバを経由することが強制されることなく、このアーキテクチャを有するアジャイルデータセンタネットワークを作り出すことが可能なスイッチの高速シリコンを横切って外部のトラフィックが直接流れることができることである。

【００５５】

インターネットから直接到達可能であることを必要とするサーバ（例えば、フロントエンドウェブサーバ）には２つのアドレス、ＬＡ及びＡＡを割り当てることができる。ＬＡはインターネットワーク通信のために用いられても良い。ＡＡはバックエンドサーバを用いたイントラデータセンタ通信のために用いられても良い。ＬＡは、ＢＧＰ(Ｂｏｒｄｅｒ　Ｇａｔｅｗａｙ　Ｐｒｏｔｏｃｏｌ)を介して伝えられかつ外部から到達可能であるプールから引き出されても良い。そのため、インターネットからのトラフィックはサーバに直接到達することができる。そのサーバから外部の送り先へのパケットはコアルータに向けてルート決めされ、ＥＣＭＰによって利用可能なリンク及びコアルータを介して拡がっても良い。

【００５６】

（ブロードキャストの処理）

上記したアジャイルネットワークアーキテクチャは下位互換性のアプリケーションに対してレイヤ２セマンティクスを提供することができる。これはブロードキャスト及びマルチメディアをサポートすることを含むことができる。アジャイルネットワークアーキテクチャのアプローチはＡＲＰ及びＤＨＣＰ(Ｄｙｎａｍｉｃ　Ｈｏｓｔ　Ｃｏｎｆｉｇｕｒａｔｉｏｎ　Ｐｒｏｔｏｃｏｌ)のようなブロードキャストの大部分の共通のソースを完全に除去することである。ＡＲＰはアジリティエージェント３２０のＡＲＰパケットを捕らえて上記したようにアジャイルディレクトリサービスから情報を調べた後、応答を与えることによって処理されても良く、ＤＨＣＰパケットは従来のＤＨＣＰリレーエージェント及びＤＨＣＰサーバに送られたユニキャストを用いてＴＯＲで捕らえられても良い。他のブロードキャストパケットを処理するために、セット内の他のホストによって送信されたブロードキャストパケットを受信可能であるべきホストの各セットはＩＰマルチキャストアドレスを割り当てられても良い。このアドレスはディレクトリサービスによって割り当てられても良く、アジリティエージェントはそのディレクトリシステムを問い合わせることによってそのアドレスを知ることができる。

【００５７】

ブロードキャストアドレスに送信されたパケットは代わりにそのサービスのマルチキャストアドレスに進行するように修正されても良い。アジャイルネットワークアーキテクチャのアジャイルエージェントは混乱を防止するために限界ブロードキャストトラフィックのレートを定めても良い。アジャイルエージェントは、サーバが最新の時間間隔（例えば、過去１秒及び過去６０秒）に亘って送信したブロードキャストパケットのレートの見積もりを維持し、サーバが各間隔の期間にブロードキャストパケットの設定数よりも多く送信することを防止することができる。許可されたパケット数を越えた分のパケットは次の

間隔まで引っ込められるか又は遅延させても良い。また、ネイティブＩＰマルチキャストがサポートされても良い。

【００５８】

複数のスイッチがレイヤ３のルータとして動作する実施形態の潜在的な有利な点は、マルチキャストグループに属する全てのマシン又はマシンに対してマルチキャストグループにアドレス指定したパケットの配信を実施することが特に容易であることである。ＰＩＭ－ＢＩＤＩＲ等の存在するＩＰマルチキャストルーティングプロトコルのいずれかは複数のスイッチに設定されても良い。これは、それらにマルチキャストグループに属する各ホスト又はマシンでエンドポイントを有するマルチキャスト分散ツリーを計算することをさせる。ホスト上のアジャイルエージェント、マシン、又はサーバは、通常、ＩＧＭＰをそのデフォルトゲートウエイに送信することによって適切なマルチキャストグループの一部であるとしてホスト、マシン、又はサーバを登録する。それ故、そのマルチキャストルーティングプロトコルはホスト、マシン、又はサーバをそのマルチキャストグループのための分散ツリーに加えることを処理する。レイヤ２で動作するスイッチはマルチキャストグループ毎にＶＬＡＮのような様々な機構を用いることができ、或いは、ネットワークを通して存在するパケットを、アジャイルエージェントのホスト、マシン又はサーバが受信すべきでないパケットを取り除く各ホスト、マシン、又はサーバ上のアジャイルエージェントで満たさせることができる。

【００５９】

（マルチパスルーティングでのランダム化）

上記したアジャイルネットワークアーキテクチャは、少なくともいくつかの実施形態では、２つの関係した機構、ＶＬＢ及びＥＣＭＰ(Ｅｑｕａｌ　Ｃｏｓｔ　Ｍｕｌｔｉｐａｔｈ)を用いて不安定さに対処するためにランダム化を用いる原理を活用／利用することができる。両方の目的は似ており、持続的な輻輳状態を減少又は防止するためにＶＬＢは複数の中間ノードに亘ってランダムにトラフィックを分散し、ＥＣＭＰは等コストパスを越えるトラフィックを発する。以下に詳細に説明されるように、ＶＬＢ及びＥＣＭＰは、各々がその他において制限を克服するために用いられる点で相補的である。両方の機構は、パケットの送り元がネットワークを越えるパスの選択に影響を与えるために使用可能な制御を提供することができる。アジャイルエージェントはそれらの制御が輻輳状態を避けるために利用されることを可能にする。

【００６０】

図５は、図３で紹介されたアジャイルネットワークアーキテクチャ３００のサブセットを示している。図５はサーバ通信に対するサーバの更なる詳細を提供する。この例はサーバ３１６（５）と通信するサーバ３１６（１）を含む。送信サーバ３１６（１）及び送り先サーバ３１６（５）は、ＶＬＡＮとして機能し１０．１２８／９のアプリケーションアドレスを有するサーバプール３２８内で動作することができる。中間スイッチ３１０（１）～３１０（Ｎ）はリンクステートネットワーク３２８内に存在する。

【００６１】

アジャイルネットワークアーキテクチャ３００は、ＶＬＢの利益がランダムに選択された中間ノードを外れるようにバケットを強制することによって達成されることを可能にする。この場合に、送り元のアジャイルエージェント３２０（１）は各パケットを中間スイッチ３１０（１）～３１０（Ｎ）に対してカプセル化することによりこれを実施することができる。その中間スイッチはそのパケットを送り先のＴＯＲ（この場合、３１４（１））に送る。ここで、パケットは先ず、中間スイッチのうちの、３１０（２）のような１つに配信され、そのスイッチによってデカプセル化され、ＴＯＲ３１４（１）のＬＡに配信され、再びデカプセル化され、そして、最後に送り先サーバ３１６（５）に送信されても良い。

【００６２】

アジャイルエージェント３２０（１）は活性状態の中間スイッチ３１０（１）～３１０（Ｎ）のアドレスを知っているならば、パケットを送信するときそれらの中からランダム

に選択することができる。しかしながら、これは、中間スイッチが機能しなくなったとき潜在的な何十万のアジャイルエージェントの更新を必要とする可能性がある。代わって、同一のＬＡアドレスが多数の中間スイッチに割り当てられても良い（この場合、ＬＡアドレス１０．０．０．５）。そのアジャイルディレクトリサービス（図３に示された）は、このエニキャストアドレスを１以上のルックアップ結果の一部としてアジャイルエージェント３２０（１）に戻すことができる。ＥＣＭＰはそのエニキャストアドレスにカプセル化されたパケットを活性状態の中間スイッチ３１０（１）～３１０（Ｎ）のうちの１つへ配信する処理を行うことができる。もしスイッチが機能しなくなったならば、ＥＣＭＰは再動作してアジャイルエージェントを通知する必要性を除去することができる。

【００６３】

しかしながら、ＥＣＭＰはスケーリング限界を有しても良い。従来のスイッチは、今日、１６ｗａｙのＥＣＭＰをサポートすることができ、また、２５６ｗａｙのＥＣＭＰスイッチが利用可能、又はまもなく利用可能である。ＥＣＭＰを使用するよりも利用可能なパスが存在することが生じたならば、ＶＬＢカプセル化は補償することができる。１つの解決法はいくつかのエニキャストアドレスを定義することであり、個々のエニキャストアドレスはＥＣＭＰが蓄積できるように多くの中間スイッチ３１０（１）～３１０（Ｎ）と関連付けされる。送り元は負荷を分散するためにエニキャストアドレスに亘ってハッシュを行うことができ、スイッチが機能しなくなると、個々のサーバが通知される必要性がないようにそのエニキャストアドレスがディレクトリシステムによって他のスイッチに再割り当てされても良い。説明のために、この態様はディレクトリシステムによって提供されたネットワーク制御機能として考慮されても良い。

【００６４】

上記したＶＬＢベースの気がつかないルーティングが、折り畳みＣｌｏｓネットワークトポロジについて純粋なＯＳＰＦ／ＥＣＭＰ機構を用いて実施されても良い。そのような設定は中間スイッチにおいてデカプセル化のサポートを必要としない。例えば、Ｎが各ＴＯＲ上のアップリンクの数であるならば、集合スイッチはセットにグループ化されても良い。いくつかの場合には、それらのセットの各々はＮのスイッチを正確に含むことができる。各ＴＯＲは１つのセット内にＮのスイッチ全てにアップリンクを有するか又は１つのセット内にスイッチのどれもアップリンクを有しないとすることができる。ＴＯＲのこの書き込みで、ＴＯＲ間のルート決定のためにＯＳＰＦ及び／又はＥＣＭＰのようなプロトコルが用いられるときでさえ、サーバの入／出制限の影響下にある任意のトラフィックのための帯域幅保証が保持し続けることを示すことができる。

【００６５】

ＴＯＲ間のルート決定のためのＯＳＰＦ又はＥＣＭＰの使用は、集合スイッチの同一のセット内における２つのＴＯＲ間のパケットのようないくつかのパケットが、中間スイッチを介して進行しないパスを利用するようにさせることができる。よって、ソースと送り先との間の最も短いパスに従い、同一の集合スイッチ又は複数の集合スイッチに接続された同一のＴＯＲ下、又は複数のＴＯＲ下のサーバ間のトラフィックの早期のターンアラウンドを可能にするようにそれらのパスを「アーリターンアラウンドパス(ｅａｒｌｙ　ｔｕｒｎａｒｏｕｎｄ　ｐａｔｈｓ)」と称することができる。それらのトラフィックフローはコア集合／中間ネットワークのいずれにも入ることを必要としない。

【００６６】

アーリターンアラウンドパスを用いることの潜在的な利益は、トラフィック（例えば、外部）の他のクラスのためのコアにおける能力を制限から解くことを含むことである。自由にされた能力は、存在するアプリケーションがクロス(ｃｒｏｓｓ)、例えば、ＴＯＲトラフィックを最小化するために書き込まれているとき、ほぼ「平均」の場合にあることができる。他の方法を見ると、これはコアが何らかの要因に基づいて設定されることを可能にし、サーバ間のトラフィックのために旨く動作することができる。また、アーリターンアラウンドパスの使用は装置の広範囲が中間スイッチとして使用されることを可能にし、それらのスイッチについて結果的に低コストにすることができる。

【００６７】
（輻輳状態に対する対処）
ＥＣＭＰ及びＶＬＢの両方を用いて、大きなフローが同一のリンク及び中間スイッチ各々にハッシュされる機会があっても良く、それは輻輳状態を生じさせるかもしれない。もし、これが起きるならば、送信アジャイルエージェントは、ＥＣＭＰがネクストホップ、すなわちパケットが通るべき次のスイッチを選択するために使用するフィールドの値を変更することによって、そのフローがアジャイルネットワークを介して利用するパスを変更することができる。この点において、アジャイルエージェントは、周期的に大きなフローをリハッシュする、又は激しい輻輳イベント（例えば、フルウインドウ損失）又はＥＣＮがＴＣＰによって検出されたとき、又は閾値のバイト／パケット数を送信／受信した後のような状況を簡単な機構で検出して処理することができる。
【００６８】
（ホスト情報の維持）
上記したアジリティネットワークアーキテクチャに応じて実施されるネットワークシステムは、データセンタ負荷のために設計されたスケーラブル（拡張可能な）、信頼がある、及び／又は高性能なストア又はディレクトリシステムを使用することができる。アジリティネットワークアーキテクチャに応じて実施されるネットワークは、一様な高能力、パフォーマンス分離、レイヤ２のセマンティック、及びサービス間の通信分離、それらの４つのプロパティの１以上を所有することができる。また、そのネットワークは正常な劣化を示すことができ、ここで、そのネットワークは故障後にどんな能力が残っていても使用することを継続することができる。よって、そのネットワークは故障にあっても信頼性／回復力がある。この点について、そのようなネットワークのディレクトリシステムは２つの潜在的なキー機能、（１）ＡＡ対ＬＡマッピングのためのルックアップ及び更新、及び（２）例えば、ライブ仮想マシンマイグレーション等の待ち時間検出動作をサポートすることができる反応型キャッシュ更新機構を提供することができる。
【００６９】
（要求の特徴付け）
ディレクトリシステムのためのルックアップ負荷は頻繁で集中的である。サーバは、ＡＡ対ＬＡマッピングのためのルックアップを生成する各フローで短時間の期間に何千又は何万の他のサーバまでと通信することができる。更新のために、負荷を故障及びサーバスタートアップイベントによって駆動することができる。多くの故障は一般に、サイズ的には小さく、大きな相関性がある故障はおそらくまれである。
【００７０】
（パフォーマンス要求）
負荷のバースト性特質は、多くの接続を素早く確立するようにルックアップが高い処理能力及び低応答時間を要求しても良いことを暗示する。ルックアップは初めてサーバと通信するために要求された時間を増加させるので、その応答時間はできだけ小さく保持されるべきである。例えば、１／１００秒が妥当な値である。しかしながら、更新のために、潜在的なキー要求は信頼性であって良く、応答時間はあまり気に掛けなくても良い。更に、更新は通常、前もって予定されるので、高い処理能力はバッチ処理の更新によって達成されても良い。
【００７１】
（一貫性の考慮）
従来のレイヤ２のネットワークにおいて、ＡＲＰはＡＲＰ中断のために最終的一貫性を提供することができる。加えて、ホストは余計なＡＲＰを発することによってその到着を伝えることができる。極端な例として、上記したアジリティネットワークアーキテクチャに応じて実施されたネットワークにおけるライブ仮想マシン（ＶＭ）マイグレーションを考慮する。ＶＭマイグレーションはステイルマッピング（ＡＡ対ＬＡ）の高速更新を利用することができる。ＶＭマイグレーションの潜在的な目的は、ロケーション変更を越えて進行している通信を保持することであることができる。それらの考慮は、信頼性ある更新

機構を提供することができる限り、ＡＡ対ＬＡマッピングの弱い又は最終的な一貫性が受け入れ可能であることを暗示している。

【００７２】

（アジャイルディレクトリシステム又はサービス設計）

パフォーマンスパラメータ及びルックアップの負荷パターンは、更新の各々からから異なっていても良い。よって、図６に示された２層になったアジャイルディレクトリサービスアーキテクチャ６００を考慮する。この場合において、アジャイルディレクトリサービスアーキテクチャ６００は、アジャイルエージェント６０２（１）～６０２（Ｎ）、ディレクトリサービスモジュール６０４（１）～６０４（Ｎ）、及びＲＳＭ(Ｒｅｐｌｉｃａｔｅｄ　Ｓｔａｔｅ　Ｍａｃｈｉｎｅ：複製ステートマシン)サーバ６０６（１）～６０６（Ｎ）を含んでいる。特に、この例では、個々のディレクトリサービスモジュールは専用のコンピュータ６０８（１）～６０８（Ｎ）各々上で実施される。他の実施例では、ディレクトリサービスモジュールは、他のシステム機能を実行するコンピュータ上で明白になっても良い。この実施例では、ディレクトリサービスモジュールの数は一般に全体のシステムサイズと比較して少ない。例えば、１つの実施例は１００Ｋサーバ（すなわち、図３のサーバ３１６（１）～３１６（Ｎ））に対してほぼ５０～１００のディレクトリサービスモジュールを用いることができる。この範囲は説明の目的で与えられ、重要ではない。

【００７３】

ディレクトリサービスモジュール６０４（１）～６０４（Ｎ）は、ＡＡ対ＬＡマッピングをキャッシュすることができる読み取り最適化された、複製ディレクトリサービスとして考慮され得る。そのディレクトリサービスモジュール６０４（１）～６０４（Ｎ）はアジャイルエージェント６０２（１）～６０２（Ｎ）、及び少量（例えば、ほぼ５～１０サーバ）の読み取り最適化された、ＲＳＭサーバ６０６（１）～６０６（Ｎ）と通信することができる。ＲＳＭサーバ６０６（１）～６０６（Ｎ）はＡＡ対ＬＡマッピングの強力な一貫して信頼性あるストア（保存部）を提供することができる。

【００７４】

ディレクトリサービスモジュール６０４（１）～６０４（Ｎ）は少ない待ち時間、高い処理能力、及び高ルックアップレートのための高い利用可能性を確かにすることができる。一方、ＲＳＭサーバ６０６（１）～６０６（Ｎ）は、少なくともいくつかの実施形態において、更新の少ないレートのために、Ｐａｘｏｓ合意アルゴリズム等を用いて強力な一貫性及び耐久性を確かにすることができる。

【００７５】

個々のディレクトリサービスモジュール６０４（１）～６０４（Ｎ）はＲＳＭサーバ６０６（１）～６０６（Ｎ）で保存されたＡＡ対ＬＡマッピングをキャッシュすることができ、キャッシュした状態を用いてアジャイルエージェント６０２（１）～６０２（Ｎ）からのルックアップに対して独立して応答することができる。強力な一貫性は必要条件でなくても良いので、ディレクトリサービスモジュールは定期的に（例えば、３０秒毎）ＲＳＭサーバとそのローカルマッピングをゆっくりと同期させることができる。高利用可能性及び少ない待ち時間を同時に達成するために、アジャイルエージェント６０２（１）～６０２（Ｎ）はランダムに選択されたディレクトリサービスモジュール６０４（１）～６０４（Ｎ）の数ｋ（例えば、２）に対してルックアップを送信することができる。多くの応答が受信されたならば、アジャイルエージェントは単に最も速い応答を選択してそれをそのキャッシュに保存する。

【００７６】

また、ディレクトリサービスモジュール６０４（１）～６０４（Ｎ）は、ネットワーク設定（プロビジョニング）システムからの更新を処理することができる。一貫性及び耐久性のために、更新は１つのランダム選択のディレクトリサービスモジュールに送信されても良く、ＲＳＭサーバ６０６（１）～６０６（Ｎ）にライトスルーされても良い。特に、更新について、ディレクトリサービスモジュールは先ず、ＲＳＭに更新を送ることができ

る。確実に、そのＲＳＭは個々のＲＳＭサーバに対してその更新を複製し、そして、ディレクトリサービスモジュールにアクナリッジメントをもって応答することができ、それはそのアクナリッジメントを元のクライアントに順に送り戻すことができる。

【００７７】

一貫性を拡張するための潜在的最適化として、ディレクトリサービスモジュール６０４（１）～６０４（Ｎ）は、少数の他のディレクトリサービスモジュールにアクナリッジメントを行った更新を任意に広めることができる。元のクライアントは中断（例えば２秒）内にアクナリッジメントを受信しなかったならば、そのクライアントは同一の更新を他のディレクトリサービスモジュールに送信して、信頼性及び／又は利用可能性と引き換えに応答時間を提供することができる。

【００７８】

また、ディレクトリサービスの他の実施形態は可能である。例えば、ＤＨＴ(Ｄｉｓｔｒｉｂｕｔｅｄ　Ｈａｓｈ　Ｔａｂｌｅ)がディレクトリサーバと、ＤＨＴにエントリーとして保存されたＡＡ／ＬＡマッピングとを用いて構成されても良い。また、パフォーマンスが前述した実施形態と同然でないか、又は強力な一貫性がなくて良いけれども、アクティブディレクトリ(Ａｃｔｉｖｅ　Ｄｉｒｅｃｔｏｒｙ)又はライトウエイトディレクトリシステム(Ｌｉｇｈｔｗｅｉｇｈｔ　Ｄｉｒｅｃｔｏｒｙ　Ｓｙｓｔｅｍ)等の他の存在するディレクトリシステムを用いても良い。

【００７９】

（最終的一貫性の確保）

ＡＡ対ＬＡマッピングはディレクトリサービスモジュール及びアジャイルエージェントのキャッシュでキャッシュされ得るので、アップデートは一貫性がないことに導く可能性がある。サーバ及びネットワークを浪費することなく一貫性がないことを解決するために、同時にスケラビリティ（拡張性）及びパフォーマンスの両方を確実にするために反応型キャッシュ更新機構を用いることができる。キャッシュ更新プロトコルはキー観察を利用することができる。ステイルホストマッピングは、そのマッピングがトラフィックを配信するために用いられるときだけに、修正される必要がある。特に、ステイルマッピングが用いられるとき、いくつかのパケットはステイルＬＡ（送り先サーバをもはや管理しないＴＯＲ又はサーバ）に到着することができる。そのＴＯＲ又はサーバはそのような配信不可能なパケットをディレクトリサービスモジュールに送り、例えば、ユニキャストを介して、ソースサーバのキャッシュにおいてステイルマッピングを選択的に修正するようにディレクトリサービスモジュールを動作させることができる。更新の他の実施形態において、ディレクトリサービスは影響を受けたサーバと通信するために許可されている全てのサーバグループに対する更新をマルチキャストによって行っても良い。

【００８０】

（更なる実施例）

（ロードバランシングの最適性）

上記したように、ＶＬＢのようなロードバランシング技術は不安定さに対処するためにランダム化を用いることができ、トラフィックパターン（最良の場合及び最悪の場合の両方を含む）を平均の場合に変えることによって最良の場合のトラフィックパターンのいくつかのパフォーマンスを潜在的に犠牲にする。このパフォーマンス損失は、より最適なトラフィックエンジニアリングシステム下にあるよりも高いいくつかのリンクの利用としてそれ自身を明らかにすることができる。しかしながら、活性状態にあるデータセンタ負荷についての評価は、更に複雑なトラフィックエンジニアリングスキームと比較したとき比較的小さい能力損失と関連づけられ得る、ＶＬＢのようなロードバランシング技術の簡潔さ及び普遍性を示している。

【００８１】

（レイアウト設定）

図７～図９は説明したアジャイルネットワークアーキテクチャに応じて実施されるデータセンタネットワークの３つの可能なレイアウト設定を示している。図７～図９において

、図面ページ上の空間制限のためにＴＯＲは関連したサーバなしに示されている。

【００８２】

図７はオープンフロアプランデータセンタレイアウト７００を示している。データセンタレイアウト７００はＴＯＲ７０２（１）～７０２（Ｎ）、集合スイッチ７０４（１）～７０４（Ｎ）、及び中間スイッチ７０６（１）～７０６（Ｎ）を含んでいる。図７において、ＴＯＲ７０２（１）～７０２（Ｎ）は、センタ「ネットワークケージ」７０８を取り囲むとして示され、（例えば、銅及び／又はファイバケーブル等を用いて）接続されても良い。集合スイッチ７０４（１）～７０４（Ｎ）及び中間スイッチ７０６（１）～７０６（Ｎ）各々はネットワークケージ７０８内側でごく接近してレイアウトされても良く、それらの中間接続のために銅ケーブルの使用を可能にする（銅ケーブルは低コストで極太であり、ファイバに対して低距離範囲用である）。ネットワークケージ７０８内側のケーブルの数を（例えば、約４分の１）、例えば、ＱＳＦＰ(Ｑｕａｄ　Ｓｍａｌｌ　Ｆｏｒｍ　Ｐｌｕｇｇａｂｌｅ)規格のような適切な規格を用いて１つのケーブルに１０Ｇリンクの数（例えば、４）と一緒に構築することによってそれらのコスト（例えば、約２分の１）と共に減らすことができる。

【００８３】

オープンフロアプランデータセンタレイアウト７００において、中間スイッチ７０６（１）～７０６（Ｎ）は、ネットワークケージ７０８内の中央に配置され、集合スイッチ７０４（１）～７０４（Ｎ）は、中間スイッチ７０６（１）～７０６（Ｎ）とＴＯＲスイッチ７０２（１）～７０２（Ｎ）（及び関連したサーバ）との間に介在されている。

【００８４】

オープンフロアプランデータセンタレイアウト７００は望んだように拡張可能である。例えば、追加のサーバラックは、サーバラックを作り出すためにＴＯＲ７０２（１）～７０２（Ｎ）を有するサーバの形で計算装置を関係付けることによって加えられても良い。そして、サーバラックはネットワークゲージ７０８の集合スイッチ７０４（１）～７０４（Ｎ）に接続されても良い。他のサーバラック及び／個々のサーバはオープンフロアプランデータセンタレイアウトによって提供されたサービスを妨げることなく削除されても良い。

【００８５】

図８はモジュラー化されたコンテナベースのレイアウト８００を示している。レイアウト８００はＴＯＲ８０２（１）～８０２（Ｎ）、集合スイッチ８０４（１）～８０４（Ｎ）、及び中間スイッチ８０６（１）～８０６（Ｎ）を含んでいる。この場合において、中間スイッチ８０６（１）～８０６（Ｎ）はそのレイアウトのデータセンタインフラ８０８に含まれる。集合スイッチ及びＴＯＲスイッチはそのデータセンタインフラに接続されているプラグ着脱可能なコンテナとして関連付けられても良い。例えば、集合スイッチ８０４（１）及び８０４（２）は、データセンタインフラ８０８に接続され得るプラグ着脱可能なコンテナ８１０（１）内でＴＯＲ８０２（１）及び８０２（２）と関連付けられても良い。同様に、集合スイッチ８０４（３）及び８０４（４）は、プラグ着脱可能なコンテナ８１０（２）内でＴＯＲ８０２（３）及び８０２（４）と関連付けられ、集合スイッチ８０４（５）及び８０４（Ｎ）は、プラグ着脱可能なコンテナ８１０（Ｎ）内でＴＯＲ８０２（５）及び８０２（Ｎ）と関連付けられる。

【００８６】

図７のように、図８において、サーバラックを作り出すためにＴＯＲと関連付けられているサーバは、図面ページの空間制限のために示されていない。更に、空間制限のために、２つの集合スイッチ及び２つの中間スイッチだけがプラグ着脱可能なコンテナ毎に示されている。勿論、他の実施例はそれらのコンポーネントのいずれか又は両方より多い又は少ないコンポーネントを用いることができる。また、他の実施例は本明細書に示された３より多い又は少ないプラグ着脱可能なコンテナを用いることができる。興味ある１つの特徴は、レイアウト８００が１のケーブルの束を各プラグ着脱可能なコンテナ８１０（１）～８１０（Ｎ）からデータセンタスパイン（すなわちデータセンタインフラ８０８）にも

たらすためにそれ自身を貸すことができることである。要約すると、データセンタインフラ８０８はレイアウト８００が個々のプラグ着脱可能なコンテナ８１０（１）～８１０（Ｎ）加えるか、又は除去することによってサイズとして拡張又は縮小することを可能にする。

【００８７】

図９は「インフラ減少」及び「コンテナ化」データセンタレイアウト９００を示している。レイアウトは多くのコンテナ９０８（１）～９０８（Ｎ）に配置されたＴＯＲ９０２（１）～９０２（Ｎ）、集合スイッチ９０４（１）～９０４（Ｎ）、及び中間スイッチ９０６（１）～９０６（Ｎ）を含んでいる。例えば、ＴＯＲ９０２（１）～９０２（２）、集合スイッチ９０４（１）～９０４（２）、及び中間スイッチ９０６（１）はコンテナ９０８（１）に配置されている。

【００８８】

コンテナ９０８（１）～９０８（Ｎ）は「インフラ減少」及び「コンテナ化」データセンタレイアウト９００の実現を可能にする。このレイアウト９００はコンテナ９０８（１）及び９０８（３）の個々のペア間でケーブル束９１０（１）を繋ぐことと関連付けられても良い。他のケーブル束９１０（２）はコンテナ９０８（２）及び９０８（Ｎ）の個々のペア間で繋ぐことができる。個々のケーブル束９１０（１），９１０（２）は、コンテナ９０８（１）内の集合スイッチ９０４（１），９０４（２）をコンテナ９０８（３）内の中間スイッチ９０６（３）と接続するリンク、また逆の場合も同じリンクを担うことができる。

【００８９】

要約すると、個々のコンテナ９０８（１）～９０８（Ｎ）は複数のスイッチを含むことができる。それらのスイッチは、相補型プラグ着脱可能なコンテナに配置されているＴＯＲ９０２（１）～９０２（Ｎ）、集合スイッチ９０４（１）～９０４（Ｎ）、及び中間スイッチ９０６（１）～９０６（Ｎ）を含むことができる。相補型プラグ着脱可能なコンテナのペアは、第１のプラグ着脱可能なコンテナの集合スイッチを第２のプラグ着脱可能なコンテナの中間スイッチに接続することによって、また逆の場合もケーブル束を介して同じによって結合されても良い。例えば、コンテナ９０８（１）はケーブル束９１０（１）を介してコンテナ９０８（３）に接続されても良い。特に、ケーブル束は、コンテナ９０８（１）の集合スイッチ９０４（１）及び９０４（２）をコンテナ９０８（３）の中間スイッチ９０６（３）と接続することができる。同様に、束９１０（１）はコンテナ９０８（３）の集合スイッチ９０４（５）及び９０４（６）をコンテナ９０８（１）の中間スイッチ９０６（１）と接続することができる。

【００９０】

少なくともいくつかの実施形態において、アジャイルネットワークアーキテクチャは次のコンポーネント（１）～（５）から構成することができる。（１）トポロジに一緒に接続されたスイッチのセット、（２）スイッチの１以上に各々接続されたサーバのセット、（３）サーバがパケットを他のサーバに送信することを望んだときリクエストが作られ、サーバ（又はサーバの代理アジャイルエージェント）がスイッチのトポロジを越えることができるように送信することを望むパケットを加えるか、又はカプセル化することにおいて使用する情報で応答するディレクトリシステム、（４）パケットがスイッチによって引っ込められてリンクに送られるほどいずれかのリンク上での利用の増大を減少又は防止するネットワークにおいて輻輳状態を制御する機構、（５）ディレクトリサービスと通信し、必要に応じてパケットをカプセル化し、アドレス指定し、又はデカプセル化し、そして、必要に応じて輻輳制御に加わるサーバ上のモジュール。

【００９１】

少なくとも１つの実施形態において、機能（１）～（５）を提供する各サーバについてアジャイルエージェントが存在することができる。（１）パケットを送り先に送り、システム内のこのサーバに登録する等のために利用されるカプセル化情報を取り出すためにアジャイルディレクトリサービスと通信する、（２）必要に応じて代替手段のセットの中か

ら（例えば、中間スイッチの中から）ランダム選択を行い、それらの選択をキャッシュする、（３）パケットをカプセル化／デカプセル化する、（４）ネットワークから輻輳表示を検出して応答する。よって、少なくともいくつかの実施形態において、それらの機能はネットワーク内の複数のサーバと複数のスイッチとの間で分散されても良い。例えば、デフォルトのルート決定はパケットを（中間スイッチ等の）スイッチのセット、及びパケットが横断する中間スイッチ上のパケット毎に実施された上記のリストの機能を管理するために用いられても良い。

【００９２】

少なくともいくつかの実施形態において、本明細書で説明されたアジャイルネットワークアーキテクチャを実施することは、ネットワーク内の各スイッチがネットワーク内の他のスイッチにパケットを送信することができるように、データセンタ内でスイッチのセットの中からネットワークを作り出すことを含むことできる。他のサーバと通信するためにサーバによって用いられたアドレスと同じ種類のアドレスをそれらのスイッチ又はそのネットワークがそれら自身の中からパケットを管理するために使用することが必要でない。例えば、ＭＡＣアドレス、ＩＰｖ４アドレス、及び／又はＩＰｖ６アドレスは全て適切であれば良い。

【００９３】

アジャイルネットワークの少なくとも１つの実施形態において、データセンタ内のスイッチのセットの中における１つの考慮は、それらの各々をＩＰアドレス、ＩＰｖ４又はＩＰｖ６のいずれかを用いて設定されること、及びそれらを１以上の標準レイヤ３ルーティングプロトコルを動作させるために一般的な例ＯＳＰＦ(Ｏｐｅｎ－Ｓｈｏｒｔｅｓｔ　Ｐａｔｈ　Ｆｉｒｓｔ)、ＩＳ－ＩＳ(Ｉｎｔｅｒｍｅｄｉａｔｅ　Ｓｙｓｔｅｍ－Ｉｎｔｅｒｍｅｄｉａｔｅ　Ｓｙｓｔｅｍ)又はＢＧＰ(Ｂｏｒｄｅｒ　Ｇａｔｅｗａｙ　Ｐｒｏｔｏｃｏｌ)を用いて設定することである。そのような実施形態の利益は、ディレクトリシステムがトポロジに対して反応してサーバに大部分の変更を知らせる必要がないように、スイッチ間でパケットを送るためにネットワークの能力を維持するそのルーティングプロトコルによって作り出されるネットワークの制御プレーンを用いて、ネットワークとディレクトリシステムとの間の接続が減少されることである。

【００９４】

代替的又は追加的に、ディレクトリシステムはネットワークのトポロジを監視し（例えば、スイッチ及びリンクの調子（ヘルス）を監視すること）、トポロジが変更するとサーバに提供するカプセル化情報を変更することができる。また、ディレクトリシステムは、前に応答を送信したサーバにそれらの応答がもはや有効でないことを通知しても良い。その代替に亘って第１の実施形態の潜在的な利点は、ディレクトリシステムがトポロジに対して反応してサーバに大部分の変更を知らせる必要がないように、スイッチ間でパケットを送るためにネットワークの能力を維持するそのルーティングプロトコルによって作り出されるネットワークの制御プレーンを用いて、ネットワークとディレクトリシステムとの間の接続が減少されることである。要約すれば、ネットワークパフォーマンスに関係した１以上のパラメータを監視することによってパケット配信遅延を減少又は避けることができる。そのパラメータは、特定のパスに亘る通信障害等のネットワークイベントを表しても良い。

【００９５】

１つの実施形態において、ネットワークの複数のスイッチは、ＬＡアドレスのサブネットから引き出されたＩＰｖ４のアドレスを用いて設定される。そのスイッチはＯＳＰＦルーティングプロトコルを動作させるように設定される。そのスイッチのアドレスはＯＳＰＦプロトコルによってスイッチの中で割り振られる。番号が付けられてないＯＳＰＦのインターフェース拡張はＯＳＰＦプロトコルによって分配される情報量を減らすために用いられても良い。各トップオブラック（ＴＯＲ）スイッチのサーバ向き部分はＶＬＡＮ(Ｖｉｒｔｕａｌ　Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ)の一部であるようにそのスイッチ上で設定される。ＡＡ空間を含むサブネットはサーバ向きのＶＬＡＮに割り当てられた

ようにスイッチに設定される。このＶＬＡＮのアドレスはＯＳＰＦには分配されず、ＶＬＡＮは通常、トランキングされない。サーバへ行くパケットはそのサーバが接続されたＴＯＲに対してカプセル化される。このＴＯＲはそれらカプセル化パケットを受信するとそのパケットをデカプセル化し、そして、サーバの送り先アドレスに基づいてサーバ向きのＶＬＡＮにそれらを送る。そのためそのサーバは通常のＬＡＮのように、パケットを受信する。

【００９６】

他の実施形態においては、ＴＯＲスイッチのサーバ向きのＶＬＡＮにＡＡサブネットを設定する代わりに、各ＴＯＲに特有のＬＡサブネットはそのサーバ向きのＶＬＡＮに割り当てられる。このＬＡサブネットはＯＳＰＦによって分配される。ＴＯＲに接続されたサーバは少なくとも２つのアドレスを用いて設定される。ＬＡサブネットから引き出されるＬＡアドレスはサーバ向きのＶＬＡＮに割り当てられ、それはＶＬＡＮの一部とＡＡアドレスである。サーバ行きのパケットはサーバ上に設定されているＬＡに対してカプセル化される。サーバ上のモジュールはカプセル化パケットを受信するとデカプセル化して、パケットに含まれたＡＡアドレスに基づいた送り先であるサーバ上の仮想マシン又は処理にローカル的に配信することができる。

【００９７】

他の実施形態において、ＴＯＲスイッチはレイヤ２のスイッチとして動作しても良く、一方、集合レイヤスイッチはレイヤ３のスイッチとして動作しても良い。この設計は、潜在的に安価なレイヤ２スイッチがＴＯＲスイッチとして使用されることを可能にしても良く（多くのＴＯＲスイッチが存在する）、一方、レイヤ３の機能は比較的少ない数の集合レイヤスイッチで実施されても良い。この設計において、デカプセル化機能はレイヤ２のスイッチ、レイヤ３のスイッチ、送り先サーバ、又は送り先仮想マシンで実行されても良い。

【００９８】

いくつかの実施形態において、追加のアドレスはスイッチ上に設定、又はＯＳＰＦのようなルーティングプロトコルを介して分配されても良い。それらのアドレスは通常、トポロジ的に重要である（すなわち、ＬＡ）。アドレスは通常、パケットをインフラサービス－すなわちサーバ、スイッチ、又は追加のサービスとして知られていることを提供するネットワーク装置に送るために用いられる。そのようなサービスの例はロードバランサ（それらはＦ５からのＢｉｇＩＰのようなハードウエアベース又はソフトウエアベースのロードバランサであって良い）、Ｓ－ＮＡＴ(Ｓｏｕｒｃｅ　Ｎｅｔｗｏｒｋ　Ａｄｄｒｅｓｓ　Ｔｒａｎｓｌａｔｏｒｓ)、ディレクトリシステムの一部であるサーバ、ＤＨＣＰサービスを提供するサーバ、又は（インターネット又は他のデータセンタ等の）他のネットワークへのゲートウエイを含む。

【００９９】

１つの実施形態において、各スイッチはＢＧＰプロトコルを用いてルートリフレクタクライアントとして設定されても良い。追加のアドレスはそのルートリフレクタ上でそれらを設定し、ＢＧＰがそれらをスイッチに分配することを許可にすることによってスイッチへ分配される。この実施形態は、追加のアドレスを加える又は除去することがそのスイッチのルーティングプロセッサを過負荷にするＯＳＰＦ再計算をさせないという利益を有する。

【０１００】

他の実施形態において、ネットワーク内の輻輳状態を制御する機構は複数のサーバそれら自身上で実施される。適切な機構はＴＣＰ(Ｔｒａｎｓｐｏｒｔ　Ｃｏｎｔｒｏｌ　Ｐｒｏｔｏｃｏｌ)のようなものであり、ここで送り先にサーバによって送られるトラフィックは、ネットワークが担うことができるらしいレートにサーバによって制限される。ＴＣＰのようなプロトコルの使用の改善は次に説明される。代替的な実施形態においては、スイッチ上のＱｏＳ(Ｑｕａｌｉｔｙ　ｏｆ　Ｓｅｒｖｉｃｅ)機構が輻輳状態制御のために用いられても良い。そのような機構の例はＷＦＱ(Ｗｅｉｇｈｔｅｄ　Ｆａｉｒ　Ｑｕ

ｅｕｉｎｇ)及びその派生物、ＲＥＤ(Ｒａｎｄｏｍ　Ｅａｒｌｙ　Ｄｅｔｅｃｔｉｏｎ)、ＲＳＶＰ、ＸＣＰ(ｅＸｐｌｉｃｉｔ　Ｃｏｎｔｒｏｌ　Ｐｒｏｔｏｃｏｌ)、及びＲＣＰ(Ｒａｔｅ　Ｃｏｎｔｒｏｌ　Ｐｒｏｔｏｃｏｌ)を含む。

【０１０１】

少なくとも１つの実施形態において、サーバ上のモジュールはアジャイルネットワークから受信されているパケットを観察し、受信したパケットから得た又は推測した情報に基づいてパケットの送信又はパケットのカプセル化を変更する。アジャイルエージェントは（１）パケットが送信されるレートを減らすためにパケットの送信を変更する、又は（２）ネットワークを通して異なるパスを利用するようにパケットのカプセル化を変更する、ことによってネットワークにおける輻輳状態を減らすことができ、それは、パケットのカプセル化及びアドレス指定を先ず選択するとき作った可能な選択肢の中からのランダム選択のいずれか又は全てを作り変えることによって達成されても良い。

【０１０２】

アジャイルエージェントがすることができる観察及びその反応の例は、次の（１）～（５）を含む。（１）アジャイルエージェントはＴＣＰパケットのフルウインドウの損失を検出するならば、パケットがネットワークを通して利用するパスを再ランダム化する。これは、特に、パケットによって利用されたパスの変更が並べ直されたパケットを送り先によって受信されることが起きないようにフローを前に送られた全てのパケットがネットワークから抜け出たと思われると同時に異なる（望ましくは輻輳してない）パスにフローを置くようにすると都合が良い。（２）アジャイルエージェントはパケットによって利用されるパスを周期的に再ランダム化することができる。（３）アジャイルエージェントは１つのフローによって達成されている実効レートを計算し、そのレートが期待閾値より低いならば再ランダム化することができる。（４）アジャイルエージェントはＥＣＮ(Ｅｘｐｌｉｃｉｔ　Ｃｏｎｇｅｓｔｉｏｎ　Ｎｏｔｉｆｉｃａｔｉｏｎ)マークの受信パケットを監視してそのレートを減らすか又はその送り先へのパケットのパスを再ランダム化することができる。（５）スイッチは輻輳状態に入った又は入ろうとしているリンクを検出するようにロジックを実行し（例えば、ＩＥＥＥ　ＱＣＮ及び８０２．１ａｕにおけるように）、通知を上流スイッチ及び／又はサーバに送ることができる。それらの指示を受信するアジャイルエージェントはそれらのパケットのレートを減らすか又はパケットのパスを再ランダム化することができる。

【０１０３】

説明した複数の実施形態の１つの有利な点は、ＶＭ（仮想マシン）が１つのサーバから他のサーバへ再配置されても良く、一方、同一のＩＰアドレスの使用を保持しても良いので、それらが仮想マシン（ＶＭ）のライブマイグレーションを可能にすることである。ディレクトリシステムは、移動の間に再配置されるサーバのＶＭのＩＰアドレスに向かうパケットを管理するために単に更新されても良い。ロケーションにおける物理的な変更は行われている通信を妨げる必要はない。

【０１０４】

少なくとも１つの実施形態において、ネットワークの能力の一部は、好ましいサービスが多数又は少数のパス、又はサービスの他のセットによって使用されるパスから分離されたパスのセットに広がったそれらのパケットを有するように、分割割合の一様な計算によってネットワークに亘って動作するサービスのセットに予約され又は優先的に割り当てられても良い。プリファレンス又はＱｏＳの多数のクラスはこの同じ技術を用いて作り出されても良い。

【０１０５】

図１０は、本コンセプトの少なくともいくつかの実施例と一致するアジャイルネットワーキング技術又は方法１０００のフローチャートを示している。方法１０００が説明される順番は、限定として構成されるつもりではなく、その方法又は代替方法を実施するために様々な説明のブロックを組み合わせても良い。更に、方法は、計算装置がその方法を実施することができるように、適切なハードウエア、ソフトウエア、ファームウエア、又は

それらの組み合わせで実施されても良い。１つの場合において、その方法は、計算装置のプロセッサによる実行が計算装置にその方法を行わせるように、命令のセットとしてコンピュータ可読記憶媒体に保存される。他の場合において、その方法はＡＳＩＣによる実行のためにＡＳＩＣのコンピュータ可読記憶媒体に保存される。

【０１０６】

１００２では、その方法はパケットを送り先に送るために利用されるカプセル化情報を得る。

【０１０７】

１００４では、その方法は複数のスイッチ等の利用可能なハードウエアを通してパスを選択する。

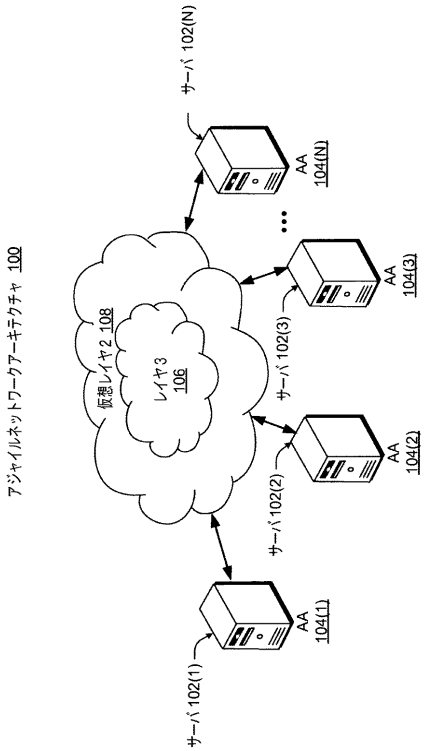
【０１０８】

１００６では、その方法はパスを介した配信のためにパケットをカプセル化する。

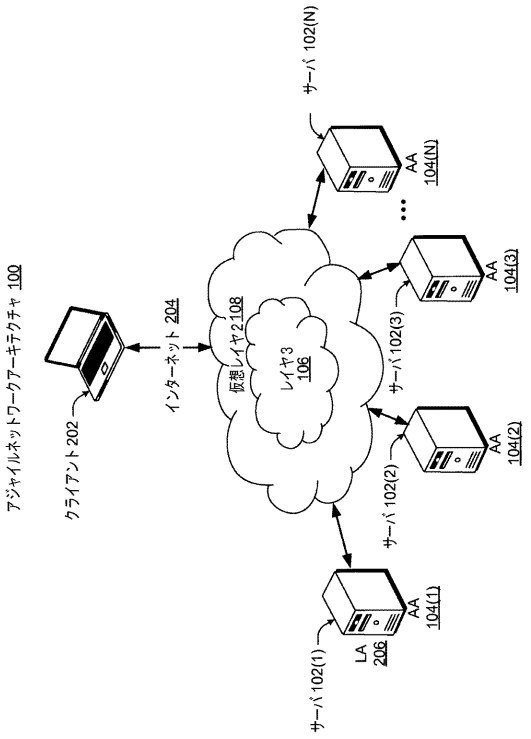
【０１０９】

１００８では、その方法は輻輳状態の表示に対して監視する。例えば、その方法はネットワークパフォーマンスに関係したパラメータを監視することができる。例えば、ＴＣＰは、パケット伝送レートに関係した更新、及び／又は輻輳状態に関係するネットワークパラメータとして動作可能なネットワークコンポーネント上の負荷を提供することができる。その方法は輻輳状態が検出されたときパスを再選択するか又は他の動作をとることができる。

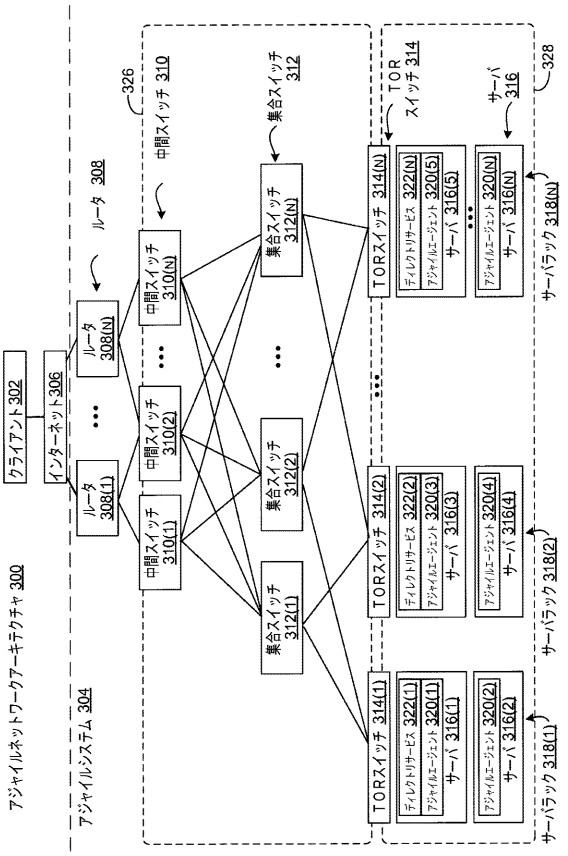
【図１】



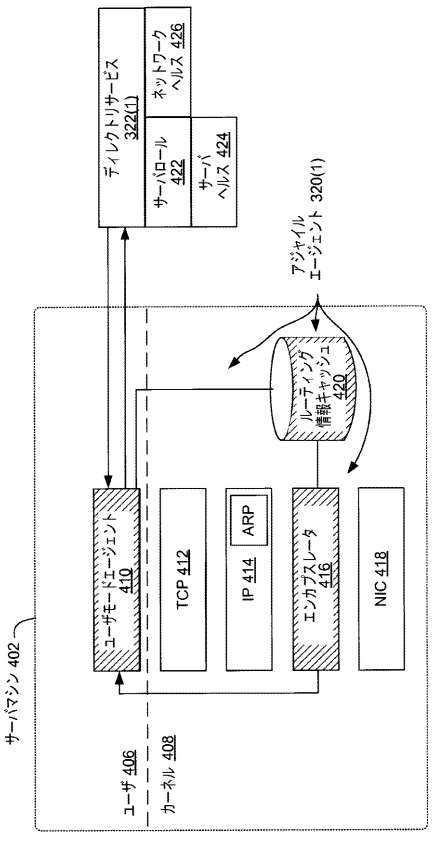
【図２】



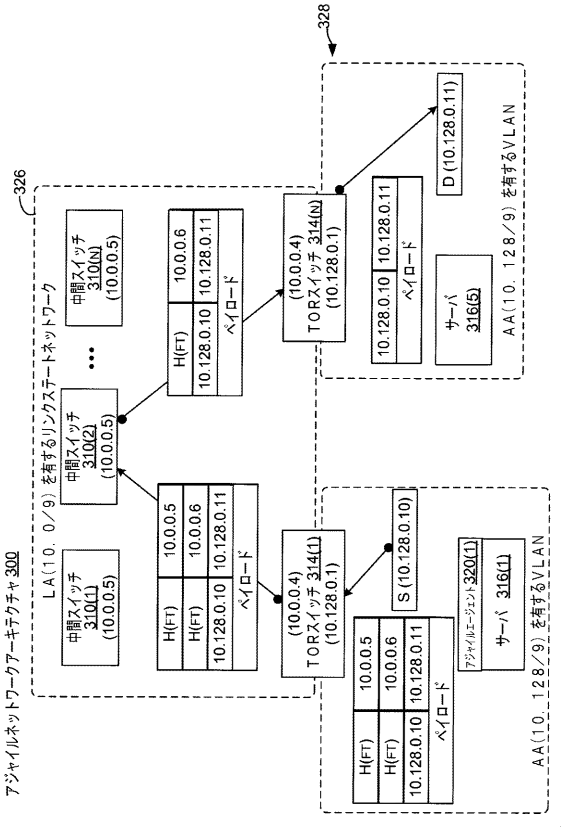
【図３】


アジャイルネットワークアーキテクチャ 300
アジャイルシステム 304
クライアント 302
インターネット 306
ルータ 308
ルータ 308(1)
ルータ 308(N)
中間スイッチ 310
中間スイッチ 310(1)
中間スイッチ 310(2)
中間スイッチ 310(N)
326
集合スイッチ 312
集合スイッチ 312(1)
集合スイッチ 312(2)
集合スイッチ 312(N)
TOR スイッチ 314
TORスイッチ 314(1)
TORスイッチ 314(2)
TORスイッチ 314(N)
328
サーバ 316
サーバラック 318(1)
サーバラック 318(2)
サーバラック 318(N)
ディレクトリサービス 322(1)
アジャイルエージェント 320(1)
サーバ 316(1)
ディレクトリサービス 322(2)
アジャイルエージェント 320(3)
サーバ 316(3)
ディレクトリサービス 322(N)
アジャイルエージェント 320(5)
サーバ 316(5)
アジャイルエージェント 320(2)
サーバ 316(2)
アジャイルエージェント 320(4)
サーバ 316(4)
アジャイルエージェント 320(N)
サーバ 316(N)

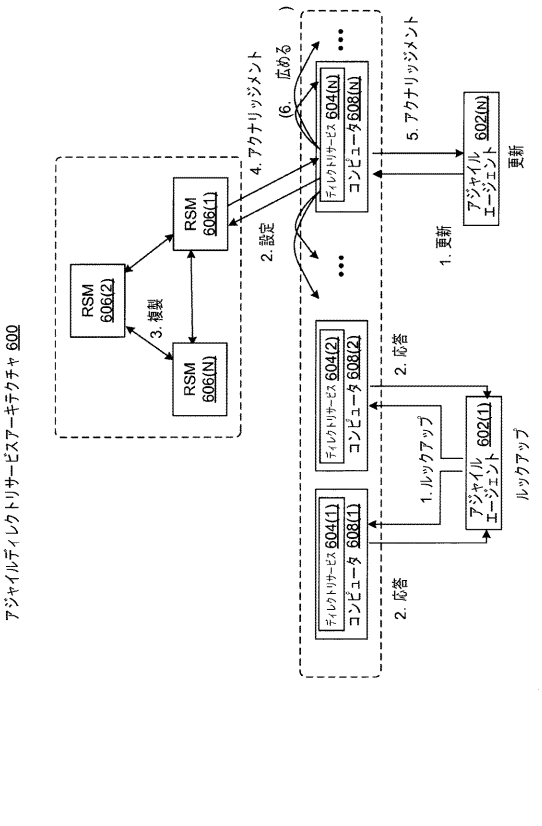
【図４】


サーバマシン 402
ユーザ 406
カーネル 408
ユーザモードエージェント 410
TCP 412
IP 414
ARP
エンカプスレータ 416
NIC 418
ルーティング情報キャッシュ 420
アジャイルエージェント 320(1)
ディレクトリサービス 322(1)
サーバロール 422
サーバヘルス 424
ネットワークヘルス 426

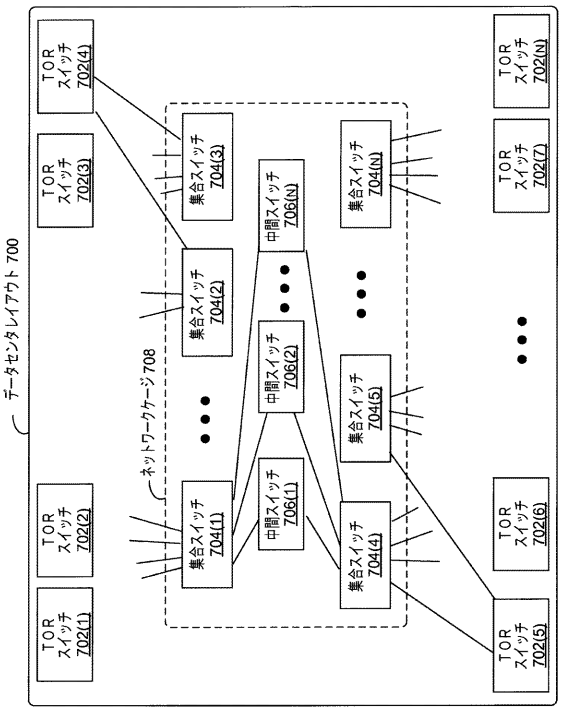
【図５】


アジャイルネットワークアーキテクチャ 300
LA(10. 0/9)を有するリンクステートネットワーク
326
328
中間スイッチ 310(1) (10.0.0.5)
中間スイッチ 310(2) (10.0.0.5)
中間スイッチ 310(N) (10.0.0.5)
(10.0.0.4) TORスイッチ 314(1) (10.128.0.1)
(10.0.0.4) TORスイッチ 314(N) (10.128.0.1)
H(FT) 10.0.0.5
H(FT) 10.0.0.6
10.128.0.10 10.128.0.11
ペイロード
S (10.128.0.10)
D (10.128.0.11)
サーバ 316(1)
サーバ 316(5)
アジャイルエージェント 320(1)
AA(10. 128/9)を有するVLAN

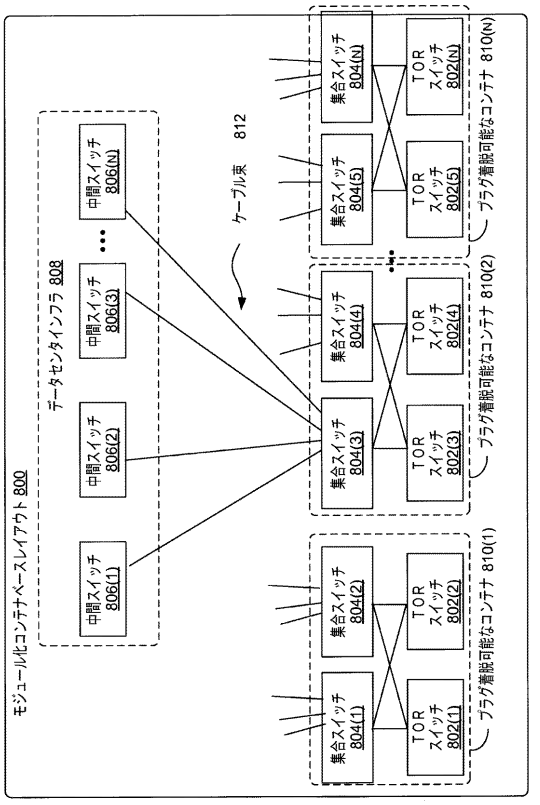
【図６】


アジャイルディレクトリサービスアーキテクチャ 600
RSM 606(1)
RSM 606(2)
RSM 606(N)
3. 複製
2. 設定
4. アクナリッジメント
(6. 広める)
5. アクナリッジメント
1. 更新
2. 応答
1. ルックアップ
ディレクトリサービス 604(1)
コンピュータ 608(1)
ディレクトリサービス 604(2)
コンピュータ 608(2)
ディレクトリサービス 604(N)
コンピュータ 608(N)
アジャイルエージェント 602(1)
アジャイルエージェント 602(N)
ルックアップ
更新

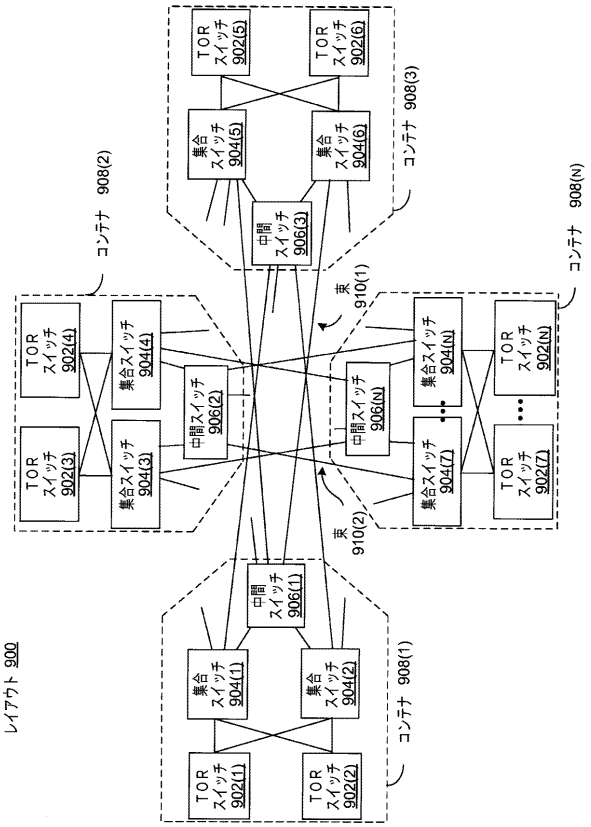
【図７】


データセンタレイアウト 700
ネットワークケージ 708
TOR スイッチ 702(1)
TOR スイッチ 702(2)
TOR スイッチ 702(3)
TOR スイッチ 702(4)
TOR スイッチ 702(5)
TOR スイッチ 702(6)
TOR スイッチ 702(7)
TOR スイッチ 702(N)
集合スイッチ 704(1)
集合スイッチ 704(2)
集合スイッチ 704(3)
集合スイッチ 704(4)
集合スイッチ 704(5)
集合スイッチ 704(N)
中間スイッチ 706(1)
中間スイッチ 706(2)
中間スイッチ 706(N)

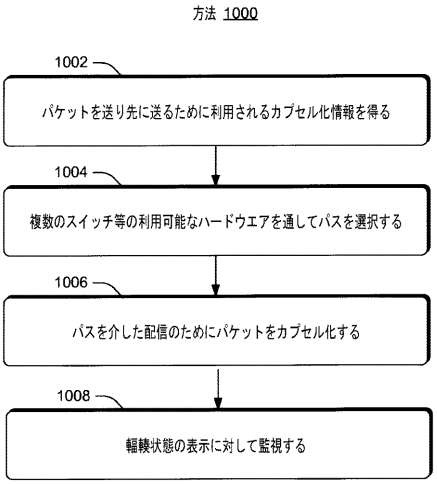
【図８】


モジュール化コンテナベースレイアウト 800
データセンタインフラ 808
ケーブル束 812
中間スイッチ 806(1)
中間スイッチ 806(2)
中間スイッチ 806(3)
中間スイッチ 806(N)
集合スイッチ 804(1)
集合スイッチ 804(2)
集合スイッチ 804(3)
集合スイッチ 804(4)
集合スイッチ 804(5)
集合スイッチ 804(N)
TOR スイッチ 802(1)
TOR スイッチ 802(2)
TOR スイッチ 802(3)
TOR スイッチ 802(4)
TOR スイッチ 802(5)
TOR スイッチ 802(N)
プラグ着脱可能なコンテナ 810(1)
プラグ着脱可能なコンテナ 810(2)
プラグ着脱可能なコンテナ 810(N)

【図９】


レイアウト 900
コンテナ 908(1)
コンテナ 908(2)
コンテナ 908(3)
コンテナ 908(N)
束 910(1)
束 910(2)
TOR スイッチ 902(1)
TOR スイッチ 902(2)
TOR スイッチ 902(3)
TOR スイッチ 902(4)
TOR スイッチ 902(5)
TOR スイッチ 902(6)
TOR スイッチ 902(7)
TOR スイッチ 902(N)
集合スイッチ 904(1)
集合スイッチ 904(2)
集合スイッチ 904(3)
集合スイッチ 904(4)
集合スイッチ 904(5)
集合スイッチ 904(6)
集合スイッチ 904(7)
集合スイッチ 904(N)
中間スイッチ 906(1)
中間スイッチ 906(2)
中間スイッチ 906(3)
中間スイッチ 906(N)

【図１０】


方法 1000
1002
パケットを送り先に送るために利用されるカプセル化情報を得る
1004
複数のスイッチ等の利用可能なハードウエアを通してパスを選択する
1006
パスを介した配信のためにパケットをカプセル化する
1008
輻輳状態の表示に対して監視する

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2010/036758

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/06(2006.01)i, H04L 12/56(2006.01)i, H04L 12/28(2006.01)i

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 29/06; G06F 15/16

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords:virtual routing, encapsulate, layer-3, layer2

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	GIHWAN CHO, "An Efficient Location and Routing Scheme for Mobile Computing Environments", IEEE JSAC. JUNE 1995, VOL. 13. NO. 5. See page 868, left column, lines 11-20; page 871, right column, lines 1-24.	9-10
Y	JESSE M. GORDON, "Hypercube Message Routing in the Presence of Faults", ACM, 1988 See abstract; page 319, lines 20-30.	9-10
A	US 2009-0063706 A1 (GOLDMAN JOEL et al.) 05 March 2009 See whole documents.	1-8
A	BRAIN VETTER, "An Experimental Study of Insider Attacks for OSPF Routing Protocol", IEEE International Conference on Network Protocols, 1997. See whole documents.	9-12

☐ 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 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 DECEMBER 2010 (29.12.2010)	**03 JANUARY 2011 (03.01.2011)**
Name and mailing address of the ISA/KR Korean Intellectual Property Office Government Complex-Daejeon, 139 Seonsa-ro, Seo-gu, Daejeon 302-701, Republic of Korea Facsimile No. 82-42-472-7140	Authorized officer Lee Seoung Young Telephone No. 82-42-481-8591

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2010/036758

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2009-0063706 A1	05.03.2009	None	

フロントページの続き

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

（特許庁注：以下のものは登録商標）
１．Ｌｉｎｕｘ
２．イーサネット
３．ＦｒｅｅＢＳＤ

(72)発明者 パランタップ ラヒリ
アメリカ合衆国 ９８０５２ ワシントン州 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 エルシーエー－インターナショナル パテンツ内
(72)発明者 デイビッド エー．マルツ
アメリカ合衆国 ９８０５２ ワシントン州 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 エルシーエー－インターナショナル パテンツ内
(72)発明者 パルヴィーン ケー．パテル
アメリカ合衆国 ９８０５２ ワシントン州 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 エルシーエー－インターナショナル パテンツ内
(72)発明者 スディプタ センングプタ
アメリカ合衆国 ９８０５２ ワシントン州 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 エルシーエー－インターナショナル パテンツ内
(72)発明者 ナヴェンデュ ジェイン
アメリカ合衆国 ９８０５２ ワシントン州 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 エルシーエー－インターナショナル パテンツ内
(72)発明者 チャンホン キム
アメリカ合衆国 ９８０５２ ワシントン州 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 エルシーエー－インターナショナル パテンツ内

Ｆターム(参考) 5B089 GA11 HB19 KB03 KC23 MA03
5K030 GA19 HA08 HC13 HD03 HD06 HD09 KA01 KA07