(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6117911号
(P6117911)
(45)発行日　平成29年4月19日(2017.4.19)　(24)登録日　平成29年3月31日(2017.3.31)

(51)Int.Cl.　　F I
H04L 12/733 (2013.01)　　H04L 12/733
H04L 12/741 (2013.01)　　H04L 12/741

請求項の数 27　(全 27 頁)

(21)出願番号　特願2015-507625(P2015-507625)
(86)(22)出願日　平成25年4月11日(2013.4.11)
(65)公表番号　特表2015-518345(P2015-518345A)
(43)公表日　平成27年6月25日(2015.6.25)
(86)国際出願番号　PCT/IB2013/052862
(87)国際公開番号　WO2013/160786
(87)国際公開日　平成25年10月31日(2013.10.31)
審査請求日　平成28年3月11日(2016.3.11)
(31)優先権主張番号　13/458,675
(32)優先日　平成24年4月27日(2012.4.27)
(33)優先権主張国　米国(US)

(73)特許権者　598036300
テレフオンアクチーボラゲット　エルエム　エリクソン（パブル）
スウェーデン国　ストックホルム　エス－１６４　８３
(74)代理人　100095957
弁理士　亀谷　美明
(74)代理人　100096389
弁理士　金本　哲男
(74)代理人　100101557
弁理士　萩原　康司
(74)代理人　100128587
弁理士　松本　一騎

最終頁に続く

(54)【発明の名称】８０２．１ＡＱのための３段折り返しＣＬＯＳの最適化

(57)【特許請求の範囲】
【請求項１】
３段折り返しＣＬＯＳネットワーク内の、計算上の複雑さ、ネットワーク運営、マルチキャストアドレッシング、及び当該ネットワークにおける障害のケースでの負荷再分配、についての改善された効率性を伴うイーサネットルーティングのための方法であって、
前記ネットワークは、各々がネットワークエレメントを表す複数のノードを含み、前記ノードはルートノードとエッジノードとを含み、各エッジノードは、ユーザに面する入力及び出力ポートのセットを含み、各ルートノードは、エッジノードを相互接続するためのポートのセットを含み、前記方法は、
前記ネットワーク内の各ノードにより、前記ルートノードから発するスパニングツリーについての転送ステートを計算し及びインストールするステップと、前記スパニングツリーにおけるデータ転送は、前記ネットワークに障害が無い場合にはマルチキャストパスについて任意ソースのアドレスを利用することと、
前記エッジノードのうちの所与の１つを、前記ルートノードのうちの所与の１つへの障害を起こした接続を有するものとして識別するステップと、
前記ネットワーク内の各ノード又はシステム管理機能により、前記所与のエッジノードへの接続性のためのプロトタイプとして、前記所与のエッジノードから発する最短パス優先（ＳＰＦ）ツリーを、前記所与のルートノードから発し前記障害を起こした接続を使用するスパニングツリーについてのＢ－ＶＩＤ（Backbone Virtual Local Area Network identifiers）のセットの各々について構築するステップと、前記ＳＰＦツリーは、前

記Ｂ－ＶＩＤについての前記所与のエッジノードへのユニキャスト及びマルチキャストの接続性のためのプロトタイプとして供されることと、
前記Ｂ－ＶＩＤの前記セットに関連付けられるサービスを共通して有するように識別される各ノードペアについての、前記所与のエッジノードと他のエッジノードとの間のペアごとの接続性のために、フィルタリングデータベースへデータ投入するステップと、
前記フィルタリングデータベース内にユニキャストステートをインストールするステップと、
トラフィックが前記所与のエッジノードへと向かうのか前記所与のエッジノードから来るのかに依存して選択されるマルチキャストアドレッシングのハイブリッドを用いて、前記フィルタリングデータベース内にマルチキャストステートをインストールするステップと、
前記フィルタリングデータベースに従って、前記ネットワークにおいてユニキャスト及びマルチキャストデータを転送するステップと、
を含む方法。
【請求項２】
データフレームを転送する前記ステップは、
前記所与のエッジノードから前記他のエッジノードへ、同じサービス識別子を有するグループ内の前記エッジノードの全てに到達する、８０２．１ａｈマルチキャストＭＡＣ（Medium Access Control）アドレッシングを用いて、データフレームの第１のセットを転送するステップと、
前記他のエッジノードの各々から前記所与のエッジノードへ、ユニキャスト又は８０２．１ａｑマルチキャストＭＡＣアドレッシングを用いて、データフレームの第２のセットを転送するステップと、前記ユニキャスト及び前記８０２．１ａｑマルチキャストＭＡＣアドレッシングは、前記グループ内の前記エッジノードの全てではなくサブセットへの通信を可能とすることと、
をさらに含む、請求項１の方法。
【請求項３】
前記スパニングツリーの各々について、当該スパニングツリーに関連付けられるマスク値によって前記ノードのシステムＩＤが変換された際に、前記ネットワーク内の前記ノードの中から、最小のシステムＩＤを有するルートノードを選択するステップ、
をさらに含む、請求項１の方法。
【請求項４】
各ノードにより、障害を起こしたルートノードの作業負荷が他の複数のルートノードにわたって分配されるように、スパニングツリールート選択の期間中のタイブレーキングのための複数のシステムＩＤを受信すること、
をさらに含む、請求項３の方法。
【請求項５】
計算し及びインストールする前記ステップは、
前記ネットワーク内の所与のＢ－ＶＩＤについてスパニングツリーを形成するために、ダイクストラのアルゴリズムを用いて、各ルートノードについて、前記エッジノードへのそのパスを計算するステップと、
前記スパニングツリーから、前記エッジノードまでに１つよりも多くのホップを有する前記パスを除外するステップと、
前記スパニングツリー内の３ホップパスを、前記３ホップパスの障害の終点にあたるエッジノードと前記ルートノードとの間のリンク障害を有するものとして識別するステップと、
をさらに含む、請求項１の方法。
【請求項６】
同じサービス識別子に関連付けられるエッジ－エッジノードペアを識別するステップと、

識別した前記エッジ－エッジノードペアに基づいて、前記ネットワーク内の各ノードについて前記フィルタリングデータベースを構築するステップと、
をさらに含む、請求項１の方法。
【請求項７】
前記フィルタリングデータベースを構築する前記ステップは、
各エッジノードにより、Ｉ－ＳＩＤマルチキャストアドレスがスパニングツリーのルートと前記エッジノードのノードとしてのユニキャストアドレスとを指し示すように、前記フィルタリングデータベースにデータ投入するステップと、
各ルートノードにより、前記フィルタリングデータベース内の、識別した前記エッジ－エッジノードペアへのマルチキャストエントリ及び前記ルートノードのノードとしてのユニキャストアドレスへのマルチキャストエントリ、を相互接続するように前記フィルタリングデータベースにデータ投入するステップと、
をさらに含む、請求項６の方法。
【請求項８】
前記スパニングツリーのサブセットを、前記ネットワーク内の他のスパニングツリーに対してプライオリティを有するものとして識別するステップと、
前記サブセットの外側での障害を前記サブセットの中へと拡散させることなく、前記サブセットの内側の障害を前記サブセットの外側へと拡散させるステップと、
をさらに含む、請求項１の方法。
【請求項９】
前記ネットワーク内の前記障害は、障害を起こしたルートノード、自ノードに接続している１つよりも多くのリンクが障害を起こした部分的に断絶されたルートノード、単一のリンクが生き残っている事実上断絶されたルートノード、自ノードに接続している１つよりも多くのリンクが障害を起こした部分的に断絶されたエッジノード、又は、異なるルートノード及び異なるエッジノードへ各々接続している障害を起こした複数のリンク、を含む、請求項１の方法。
【請求項１０】
前記３段折り返しＣＬＯＳネットワークは、データセンタ内のネットワークを表す、請求項１の方法。
【請求項１１】
３段折り返しＣＬＯＳネットワークのエッジノードとして機能するネットワークエレメントであって、計算上の複雑さ、ネットワーク運営、マルチキャストアドレッシング、及び当該ネットワークにおける障害のケースでの負荷再分配、についての改善された効率性を伴うイーサネットルーティングを使用し、前記エッジノードは、
ユーザに面する入力及び出力ポートの第１のセットと、
複数のルートノードへ連結される入力及び出力ポートの第２のセットと、
フィルタリングデータベースを記憶するメモリと、
前記メモリ、ユーザに面する入力及び出力ポートの前記第１のセット、並びに入力及び出力ポートの前記第２のセット、へ連結されるネットワークプロセッサと、
を備え、
前記ネットワークプロセッサは、
前記ルートノードから発するスパニングツリーについての転送ステートを計算し及びインストールし、
前記所与のエッジノードへの接続性のためのプロトタイプとして、前記所与のエッジノードから発する最短パス優先（ＳＰＦ）ツリーを、前記所与のルートノードから発し前記障害を起こした接続を使用するスパニングツリーについてのＢ－ＶＩＤ（Backbone Virtual Local Area Network identifiers）のセットの各々について構築し、前記ＳＰＦツリーは、前記Ｂ－ＶＩＤについての前記所与のエッジノードへのユニキャスト及びマルチキャストの接続性のためのプロトタイプとして供されるものであり、
前記Ｂ－ＶＩＤのセットに関連付けられるサービスを共通して有するように識別される

各ノードペアについての、前記所与のエッジノードと他のエッジノードとの間のペアごとの接続性のために、フィルタリングデータベースへデータ投入し、
前記フィルタリングデータベース内にユニキャストステートをインストールし、
トラフィックが前記所与のエッジノードへと向かうのか前記所与のエッジノードから来るのかに依存して選択されるマルチキャストアドレッシングのハイブリッドを用いて、前記フィルタリングデータベース内にマルチキャストステートをインストールし、
前記フィルタリングデータベースに従って、前記ネットワークにおいてユニキャスト及びマルチキャストデータを転送する、
ように構成される、ネットワークエレメント。
【請求項１２】
前記ネットワークエレメントは、管理システムへ連結され、前記管理システムが前記ネットワーク内の各ノードの代わりに各Ｂ－ＶＩＤについて前記ＳＰＦを構築する、請求項１１のネットワークエレメント。
【請求項１３】
前記ネットワークプロセッサは、
前記所与のエッジノードから前記他のエッジノードへ、同じサービス識別子を有するグループ内の前記エッジノードの全てに到達する、８０２．１ａｈマルチキャストＭＡＣ（Medium Access Control）アドレッシングを用いて、データフレームの第１のセットを転送し、
前記他のエッジノードの各々から前記所与のエッジノードへ、ユニキャスト又は８０２．１ａｑマルチキャストＭＡＣアドレッシングを用いて、データフレームの第２のセットを転送する、
ようにさらに構成され、
前記ユニキャスト及び前記８０２．１ａｑマルチキャストＭＡＣアドレッシングは、前記グループ内の前記エッジノードの全てではなくサブセットへの通信を可能とする、
請求項１１のネットワークエレメント。
【請求項１４】
前記ネットワークプロセッサは、前記スパニングツリーの各々について、当該スパニングツリーに関連付けられるマスク値によって前記ルートノードのシステムＩＤが変換された際に、前記ルートノードの中から、最小のシステムＩＤを有するルートノードを選択する、
ようにさらに構成される、請求項１１のネットワークエレメント。
【請求項１５】
各ノードは、障害を起こしたルートノードの作業負荷が他の複数のルートノードにわたって分配されるように、スパニングツリールート選択の期間中のタイブレーキングのための複数のシステムＩＤを受信する、請求項１４のネットワークエレメント。
【請求項１６】
前記ネットワークプロセッサは
前記ネットワーク内の所与のＢ－ＶＩＤについてスパニングツリーを形成するために、ダイクストラのアルゴリズムを用いて、各ルートノードについて、前記エッジノードへのそのパスを計算し、
前記スパニングツリーから、前記エッジノードまでに１つよりも多くのホップを有する前記パスを除外し、
前記スパニングツリー内の３ホップパスを、前記ルートノードとの間の隣接関係の上でリンク障害を有するノードへ到達するものとして識別する、
ようにさらに構成される、請求項１１のネットワークエレメント。
【請求項１７】
前記ネットワーク内の前記障害は、障害を起こしたルートノード、自ノードに接続している１つよりも多くのリンクが障害を起こした部分的に断絶されたルートノード、単一のリンクが生き残っている事実上断絶されたルートノード、自ノードに接続している１つよ

りも多くのリンクが障害を起こした部分的に断絶されたエッジノード、又は、異なるルートノード及び異なるエッジノードへ各々接続している障害を起こした複数のリンク、を含む、請求項１１のネットワークエレメント。
【請求項１８】
各ルートノードは、バックボーンコアブリッジ（ＢＣＢ）であり、各エッジノードは、バックボーンエッジブリッジ（ＢＥＢ）である、請求項１１のネットワークエレメント。
【請求項１９】
前記３段折り返しＣＬＯＳネットワークは、データセンタ内のネットワークを表す、請求項１１のネットワークエレメント。
【請求項２０】
３段折り返しＣＬＯＳネットワークのシステムであって、計算上の複雑さ、ネットワーク運営、マルチキャストアドレッシング、及び当該ネットワークにおける障害のケースでの負荷再分配、についての改善された効率性を伴うイーサネットルーティングを使用し、前記システムは、
ユーザに面する入力及び出力ポートのセットを各々が含む複数のエッジノードと、
前記複数のエッジノードを相互接続するためのポートのセットを各々が含む複数のルートノードと、
を含み、
前記ルートノード及びエッジノードの各々は、
フィルタリングデータベースを記憶するメモリと、
前記メモリへ連結されるネットワークプロセッサと、
を備え、
前記ネットワークプロセッサは、
前記ルートノードから発するスパニングツリーについての転送ステートを計算し及びインストールし、
前記所与のエッジノードへの接続性のためのプロトタイプとして、前記所与のエッジノードから発する最短パス優先（ＳＰＦ）ツリーを、前記所与のルートノードから発し前記障害を起こした接続を使用するスパニングツリーについてのＢ－ＶＩＤ（Backbone Virtual Local Area Network identifiers）のセットの各々について構築し、前記ＳＰＦツリーは、前記Ｂ－ＶＩＤについての前記所与のエッジノードへのユニキャスト及びマルチキャストの接続性のためのプロトタイプとして供されるものであり、
前記Ｂ－ＶＩＤのセットに関連付けられるサービスを共通して有するように識別される各ノードペアについての、前記所与のエッジノードと他のエッジノードとの間のペアごとの接続性のために、フィルタリングデータベースへデータ投入し、
前記フィルタリングデータベース内にユニキャストステートをインストールし、
トラフィックが前記所与のエッジノードへと向かうのか前記所与のエッジノードから来るのかに依存して選択されるマルチキャストアドレッシングのハイブリッドを用いて、前記フィルタリングデータベース内にマルチキャストステートをインストールし、
前記フィルタリングデータベースに従って、前記ネットワークにおいてユニキャスト及びマルチキャストデータを転送する、
ように構成される、システム。
【請求項２１】
前記ネットワーク内の各ノードの代わりに各Ｂ－ＶＩＤについて前記ＳＰＦを構築する管理システム、
をさらに含む、請求項２０のシステム。
【請求項２２】
前記ネットワークプロセッサは、
前記所与のエッジノードから前記他のエッジノードへ、同じサービス識別子を有するグループ内の前記エッジノードの全てに到達する、８０２．１ａｈマルチキャストＭＡＣ（Medium Access Control）アドレッシングを用いて、データフレームの第１のセットを

転送し、
前記他のエッジノードの各々から前記所与のエッジノードへ、ユニキャスト又は８０２．１ａｑマルチキャストＭＡＣアドレッシングを用いて、データフレームの第２のセットを転送する、
ようにさらに構成され、
前記ユニキャスト及び前記８０２．１ａｑマルチキャストＭＡＣアドレッシングは、前記グループ内の前記エッジノードの全てではなくサブセットへの通信を可能とする、
請求項２０のシステム。
【請求項２３】
前記ネットワークプロセッサは、前記スパニングツリーの各々について、当該スパニングツリーに関連付けられるマスク値によって前記ルートノードのシステムＩＤが変換された際に、前記ルートノードの中から、最小のシステムＩＤを有するルートノードを選択する、
ようにさらに構成される、請求項２０のシステム。
【請求項２４】
障害を起こしたルートノードの作業負荷が他の複数のルートノードにわたって分配されるように、スパニングツリールート選択の期間中のタイブレーキングのために使用される複数のシステムＩＤで、各ノードを構成する管理システム、
をさらに含む、請求項２３のシステム。
【請求項２５】
前記ネットワーク内の前記障害は、障害を起こしたルートノード、自ノードに接続している１つよりも多くのリンクが障害を起こした部分的に断絶されたルートノード、単一のリンクが生き残っている事実上断絶されたルートノード、自ノードに接続している１つよりも多くのリンクが障害を起こした部分的に断絶されたエッジノード、又は、異なるルートノード及び異なるエッジノードへ各々接続している障害を起こした複数のリンク、を含む、請求項２０のシステム。
【請求項２６】
各ルートノードは、バックボーンコアブリッジ（ＢＣＢ）であり、各エッジノードは、バックボーンエッジブリッジ（ＢＥＢ）である、請求項２０のシステム。
【請求項２７】
前記３段折り返しＣＬＯＳネットワークは、データセンタ内のネットワークを表す、請求項２０のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は、コンピュータネットワーキングの分野に関し、より具体的には、３段折り返し（three stage folded）ＣＬＯＳネットワークにおけるイーサネットルーティングの最適化に関する。
【背景技術】
【０００２】
ＩＥＥＥ８０２．１ａｑ標準（これ以降、８０２．１ａｑともいう）は、２０１２年に公開され、イーサネットのためのルーティングソリューションを定義している。８０２．１ａｑは、最短パスブリッジングあるいはＳＰＢとしても知られている。８０２．１ａｑは、ネイティブなイーサネット基盤上での論理的なイーサネットネットワークの生成を可能とする。８０２．１ａｑは、トポロジー及びネットワーク内のノードの論理的なネットワークメンバシップの双方を広告するために、リンクステートプロトコルを採用する。データパケットは、８０２．１ａｑを実装するネットワークのエッジノードにおいて、ｍａｃ－ｉｎ－ｍａｃの８０２．１ａｈフレーム又はタグ付き８０２．１Ｑ／ｐ８０２．１ａｄフレームのいずれかでカプセル化され、論理ネットワークの他のメンバにのみ移送される。ユニキャスト及びマルチキャストもまた、８０２．１ａｑによりサポートされる。全

てのそうしたルーティングは、対称的な最短パスを介して行われる。多くの等コストな最短パスがサポートされる。ネットワークにおける８０２．１ａｑの実装は、プロバイダネットワーク、企業ネットワーク及びクラウドネットワークを含む多様なタイプのネットワークの生成及び構成を簡素化する。構成は比較的簡素化され、エラーの可能性、特に人為的な構成エラーの可能性が減少する。

【発明の概要】

【０００３】

方法の一実施形態が、計算上の複雑さ、ネットワーク運営、マルチキャストアドレッシング、及び障害時の負荷再分配、についての改善された効率性を伴うイーサネットルーティングのために使用される３段折り返しＣＬＯＳネットワークにおいて実装される。上記ネットワークは、ユーザに面する（user-facing）入力及び出力ポートを有するエッジノードの配列に連結されるルートノードの配列を含む。上記ネットワーク内の各ノードは、上記ルートノードから発するスパニングツリーについての転送ステートを計算し及びインストールする。上記スパニングツリーにおけるデータ転送は、上記ネットワークに障害が無い場合には任意ソース（any source）のマルチキャストパスアドレスをマルチキャストパスのために利用する。これは、サービスに参加するノードの数とマルチキャストグループの数との積という組合せ積（combinatorial product）ではなく、マルチキャストグループの数に線型的に比例してスケーリングするためである。所与のルートノードへの障害を起こした接続を有するものとして所与のエッジノードが識別されると、上記ネットワーク内の各ノードは、上記所与のルートノードから発し上記障害を起こした接続を使用するスパニングツリーのための各Ｂ－ＶＩＤについての、上記所与のエッジノードから発する最短パス優先（ＳＰＦ）ツリーを計算し、上記ＳＰＦツリーは、上記Ｂ－ＶＩＤについての上記所与のエッジノードへのユニキャスト及びマルチキャストの接続性のためのプロトタイプとして供される。上記所与のエッジノードと他のエッジノードとの間のペアごとの接続性のために、各ノード内のフィルタリングデータベースへデータが投入され（populated）、それらは移動される（displaced）Ｂ－ＶＩＤに関連付けられるサービスを共通して有する。そして、ノードは、上記フィルタリングデータベース内にユニキャストステートをインストールし、トラフィックが上記所与のエッジノードへと向かうのか上記所与のエッジノードから来るのかに依存して選択されるマルチキャストアドレッシングのハイブリッドを用いて、上記フィルタリングデータベース内にマルチキャストステートをもインストールする。そして、各ノードは、自身の転送データベースを用いて、上記ネットワークにおいてユニキャスト及びマルチキャストデータを転送する。

【０００４】

上述した実施形態は、分散型のルーティングシステムであり、各ノードがフィルタリングデータベース内の自身の転送テーブルを計算する。代替的な実施形態において、転送テーブルを計算するシステム管理機能を、集中型のコントローラが実行する。そして、転送テーブルは、各ノードへ、当該ノードがデータ転送を実行するためにダウンロードされる。

【０００５】

エッジノードとして機能するネットワークエレメントの一実施形態が、マルチキャストアドレッシング及び障害時の負荷再分配、についての改善された効率性を伴うイーサネットルーティングのために使用される３段折り返しＣＬＯＳネットワークにおいて実装される。上記エッジノードは、ユーザに面する入力及び出力ポートの第１のセットと、複数のルートノードへ連結される入力及び出力ポートの第２のセットと、フィルタリングデータベースを記憶するメモリと、ネットワークプロセッサと、を備える。上記ネットワークプロセッサは、上記ルートノードから発するスパニングツリーについての転送ステートを計算し及びインストールするように構成され、上記スパニングツリーにおけるデータ転送は、上記ネットワークに障害が無い場合には任意ソースのマルチキャストパスアドレスをマルチキャストパスのために利用し、上記ネットワークプロセッサは、所与のルートノードへの障害を起こした接続を有するものとして所与のエッジノードを識別し、上記所与のエ

ッジノードから発するＳＰＦツリーを、上記所与のルートノードから発し上記障害を起こした接続を使用するスパニングツリーのための各Ｂ－ＶＩＤについて構築するように構成され、上記ＳＰＦツリーは、上記所与のエッジノードへのユニキャスト及びマルチキャストの接続性のためのプロトタイプとして供されるものであり、上記ネットワークプロセッサは、移動される上記Ｂ－ＶＩＤに関連付けられるサービスを共通して有する上記所与のエッジノードと他のエッジノードとの間のペアごとの接続性のために、フィルタリングデータベースへデータ投入し、上記フィルタリングデータベース内にユニキャストステートをインストールし、トラフィックが上記所与のエッジノードへと向かうのか上記所与のエッジノードから来るのかに依存して選択されるマルチキャストアドレッシングのハイブリッドを用いて、上記フィルタリングデータベース内にマルチキャストステートをインストールする、ように構成される。そして、ネットワーク内のユニキャスト及びマルチキャストデータを転送するために、転送データベースが使用される。

【０００６】

システムの一実施形態が、マルチキャストアドレッシング及び障害時の負荷再分配、についての改善された効率性を伴うイーサネットルーティングのために使用される３段折り返しＣＬＯＳネットワークにおいて実装される。上記システムは、ユーザに面する入力及び出力ポートのセットを有するエッジノード群へ連結されるルートノード群を含む。上記ルートノード及び上記エッジノードの各々は、フィルタリングデータベースを記憶するメモリと、ネットワークプロセッサと、を備える。上記ネットワークプロセッサは、上記ネットワークプロセッサは、上記ルートノードから発するスパニングツリーについての転送ステートを計算し及びインストールするように構成され、上記スパニングツリーにおけるデータ転送は、上記ネットワークに障害が無い場合には任意ソースのマルチキャストパスアドレスをマルチキャストパスのために利用し、上記ネットワークプロセッサは、所与のルートノードへの障害を起こした接続を有するものとして所与のエッジノードを識別し、上記所与のエッジノードから発するＳＰＦツリーを、上記所与のルートノードから発し上記障害を起こした接続を使用するスパニングツリーのための各Ｂ－ＶＩＤについて構築するように構成され、上記ＳＰＦツリーは、上記Ｂ－ＶＩＤについての上記所与のエッジノードへのユニキャスト及びマルチキャストの接続性のためのプロトタイプとして供されるものであり、上記ネットワークプロセッサは、移動される上記Ｂ－ＶＩＤに関連付けられるサービスを共通して有する上記所与のエッジノードと他のエッジノードとの間のペアごとの接続性のために、フィルタリングデータベースへデータ投入し、上記フィルタリングデータベース内にユニキャストステートをインストールし、トラフィックが上記所与のエッジノードへと向かうのか上記所与のエッジノードから来るのかに依存して選択されるマルチキャストアドレッシングのハイブリッドを用いて、上記フィルタリングデータベース内にマルチキャストステートをインストールする、ように構成される。そして、ネットワーク内のユニキャスト及びマルチキャストデータを転送するために、転送データベースが使用される。

【図面の簡単な説明】

【０００７】

本発明は、限定ではなく例示の手法で添付図面の図において示され、図中で類似の参照符号は同様のエレメントを指し示す。なお、本開示における「一」実施形態又は「１つの」実施形態についての異なる言及は、必ずしも同じ実施形態を指すものではなく、それらの言及は、少なくとも１つの実施形態を意味することに留意すべきである。さらに、一実施形態に関連して特定の特徴、構造、又は特性が説明される場合、他の実施形態に関連してそのような特徴、構造、又は特性を作用させることは、明示されているか否かに係わらず、当業者の知識の範囲内であることを提示しておく。

【０００８】

【図１】３段折り返しＣＬＯＳネットワークの一例を示している。

【図２】ルートノードに割り当てられるシステムＩＤセットの一例を示している。

【図３】スパニングツリールート選択のために使用されるスプリットタイブレーキングの

仕組みの一例を示している。
【図４】３段折り返しＣＬＯＳネットワークにおいてスパニングツリーのための転送ステージを計算し及びインストールする方法の一例を示している。
【図５】一実施形態に係るルートノードに障害が起きるシナリオを示している。
【図６】一実施形態に係るリンクに障害が起きるシナリオを示している。
【図７】一実施形態に係る２つのリンクに障害が起きる第１のシナリオを示している。
【図８】一実施形態に係る２つのリンクに障害が起きる第２のシナリオを示している。
【図９】一実施形態に係る２つのリンクに障害が起きる第３のシナリオを示している。
【図１０】一実施形態に係る３つのリンクに障害が起きる第１のシナリオを示している。
【図１１】一実施形態に係る３つのリンクに障害が起きる第２のシナリオを示している。
【図１２】３段折り返しＣＬＯＳネットワークにおいて障害が発生した場合のデータフレーム転送のための方法の一実施形態を示している。
【図１３】一実施形態に係る管理システムに連結されるネットワークエレメントを示すブロック図である。
【発明を実施するための形態】
【０００９】
以下の説明において、多くの具体的な詳細が説明される。しかしながら、本発明の実施形態が、それらの具体的な詳細がなくとも実践され得ることが理解されるべきである。他の例では、本説明の理解を曖昧にすることのないように、周知の回路、構造、及び技法については詳細を示していない。しかしながら、そのような具体的な詳細がなくとも本発明が実践され得ることが、当業者によって理解されるであろう。ここに含まれる説明によって、当業者は、過度の実験をすることなく、適切な機能性を実装することができるであろう。
【００１０】
ＩＥＥＥ８０２．１ａｑ標準では、ネットワーク上のイーサネットフレームの転送を制御するために、リンクステートプロトコルが利用される。１つのリンクステートプロトコルであるＩＳ－ＩＳ（Intermediate System to Intermediate System）が、ネットワークのトポロジー及び論理的なネットワークメンバシップの双方を広告するために、８０２．１ａｑネットワークにおいて使用される。
【００１１】
８０２．１ａｑは、２つの動作モードを有する。ＶＬＡＮ（Virtual Local Area Network）ベースのネットワークのための第１のモードを、最短パスブリッジングＶＩＤ（ＳＰＢＶ）という。ＭＡＣベースのネットワークのための第２のモードを、最短パスブリッジングＭＡＣ（ＳＰＢＭ）という。ＳＰＢＶネットワーク及びＳＰＢＭネットワークの双方は、データプレーンにおいて同時に等コスト転送ツリーのセット（ＥＣＴセット）を１つよりも多くサポートすることができる。ＥＣＴセットは、通常、複数の最短パスＶＬＡＮ識別子（ＳＰＶＩＤ）に関連付けられ、ＳＰＢＶのためにＳＰＶＩＤセットを形成し、及び、ＳＰＢＭのためにバックボーンＶＬＡＮ　ＩＤ（Ｂ－ＶＩＤ）と１対１で関連付けられる。
【００１２】
８０２．１ａｑのＭＡＣモードによれば、プロバイダネットワーク内のネットワークエレメントは、同じ宛て先アドレスを宛て先としつつも異なるＢ－ＶＩＤにマッピングされた異なるフレームがネットワークを通じて異なるパス（“マルチパスインスタンス”という）上で転送され得るように、Ｂ－ＶＩＤによって分離されるマルチパス転送トラフィックを実行するように構成される。サービスに関連付けられる顧客のデータフレームは、８０２．１ａｑに従い、別個のサービス識別子（Ｉ－ＳＩＤ）及びＢ－ＶＩＤを有するヘッダと共にカプセル化される。この分離は、ネットワークトポロジーから独立してサービスがスケーリングされることを可能とする。よって、Ｂ－ＶＩＤは、専らマルチパスインスタンスの識別子として使用されることができる。Ｉ－ＳＩＤは、Ｂ－ＶＩＤにより識別されるマルチパスインスタンスによって提供されるべき特定のサービスを識別する。８０２

．１ａｑにおけるマルチパスインスタンスの実際のルーティングは、各ノードのシステムＩＤに基づくタイブレーキングによって決定される。

【００１３】

ＷＡＮ（Wide Area Network）又はクラウドコンピューティングデータセンタといったデータセンタ内で、８０２．１ａｑをイーサネットルーティングのために使用することができる。データセンタ内のネットワークは、典型的には、ＣＬＯＳネットワークのトポロジーのように、非常に整然とした構造を有する。ＣＬＯＳネットワークは、スイッチングノードの配列を含む。ＣＬＯＳネットワークの一例は、入口ステージ、中間ステージ及び出口ステージを有する３段折り返しＣＬＯＳであり、当該ネットワークは、入口ステージが出口ステージにマージされるように、ノードの中間ステージをまたいで途中で折り返される。入口ステージ内のノードに入る各データフレームは、宛て先の出口ステージのノードへ到達するための中間ステージの利用可能なノードのいずれかを通じてルーティングされ得る。

【００１４】

障害の起きたノード又はリンクを有するＣＬＯＳネットワークにおける接続性を維持するために、いくつかの技法が開発されてきた。８０２．１ａｑによれば、ノード又はリンクの障害は、１つ以上の周囲ノードにより観測され、ルーティングシステムによってネットワークにわたって広告される。ネットワーク内の各ノードは、障害により影響を受けるトラフィックについて新たなパスを再計算し、新たなパスを用いて転送が自動的に継続されるであろう。

【００１５】

任意のトポロジーへの、非最適な転送パスをもたらすスパニングツリーの適用とは異なり、障害の無いＣＬＯＳネットワークでは、第２のティアノード群（即ち、中間ステージ内のノード群）から発する（rooted）複数のスパニングツリーにより提供される接続性が、最短パスツリーに基づくものと同一である。これは、ルーティング問題の計算上の複雑さの観点において簡易的であるだけでなく、よりスケーラブルなマルチキャストアドレッシングフォーマットを使用することをも可能とする。しかしながら、第２のティアノードが不完全な接続性を有するようにリンクに障害が起きた場合のトラフィックの分散は、課題（problematic）であり、最短パスツリーの動作への立ち戻りをもたらし、あまりスケーラブルでないマルチキャストアドレッシングの使用を要し得る。

【００１６】

さらに、８０２．１ａｑによれば、障害が存在する場合、移動を強いられるトラフィックはまとまってフェイルオーバパスへとシフトされる。フェイルオーバパスが今やトラフィックの大幅な増加を処理することから、トラフィックのまとまった再分配がネットワークの安定性を減少させるかもしれず、それにより性能が大幅に劣化し得る。さらに、フェイルオーバパスへのこのまとまったトラフィックのシフトは、フェイルオーバパス内のリンク及びノードを、それらエレメントを実質的に障害に陥らせるほど圧倒し得る。基本的な仕様８０２．１ａｑの振る舞いを修正して、障害シナリオにおけるマルチパス選択との共通性のある仕組みであって、障害シナリオでのネットワークキャパシティの劣化の公平な分配のために提供される分散的なスパニングツリーのルート選択（root election）の仕組みを提供することにより、マルチキャストアドレスのスケーラビリティを最大化し、計算上の複雑さを最小化し、及びネットワーク内のマルチパス化設計を簡易にするマルチパス化技法を適用することが望ましいであろう。

【００１７】

ここで説明される実施形態は、３段折り返しＣＬＯＳネットワークといった、フラット化されたスイッチング階層を有するネットワークトポロジーを利用する。３段折り返しＣＬＯＳネットワークは、入口ステージが出口ステージにマージされるように、ノードの中間ステージをまたいで途中で折り返される３ステージ型のＣＬＯＳネットワークである。マージされる入口／出口ステージはエッジノード（第１ティアノードともいう）の配列を含み、中間ステージはルートノード（第２ティアノードともいう）の配列を含む。ルート

ノードへ接続される１つ以上のリンクに障害が起きた場合、不必要に長い転送パスを生じさせてまで全ての接続性のためにそのルート（root）を使用し続けたり、又は“全てのペア”の計算に立ち返ったりする代わりに、特別なソリューションを前提とすることが可能であり、Ｂ－ＶＩＤについての障害の起きたリンク上で上記ルートと通常直接的な隣接関係を有する各エッジノードについて、最短パス優先（ＳＰＦ）ツリーが生成され、当該ＳＰＦツリーはエッジノードから発する。障害の起きたルートノード以外の全てのルートノードについてスパニングツリーが生成されるといった“全てのペア”の計算は存在せず、代わりに、ルートノード又はルートノードへ接続する１つ以上のリンクの障害について、Ｂ－ＶＩＤごとに１つのＳＰＦツリーのみが生成される。よって、ルートノードに関わる障害のケースでの計算上の負荷が、大幅に削減される。

【００１８】

ここで開示される実施形態は、各ＥＴＣセットについてデータフレームが横断するスパニングツリーのルートを選択するように、スプリットタイブレーキングの仕組みと共に増強される、８０２．１ａｑのタイブレーキングパス選択技法の組合せのバリエーションを利用する。１つの実施形態において、スパニングツリーのためのルートは、Ｂ－ＶＩＤに関連付けられるＸＯＲマスク値でのマスキング後の最小のシステムＩＤを伴うルートノードである。ネットワークは、ＣＬＯＳ内の第２のティアノードが一貫して、エッジノードに割り当てられるシステムＩＤと比較した場合にタイブレーキングにおける最小のシステムＩＤを有することを保証するように運営されてきた。ＣＬＯＳネットワーク内の各ノードは、タイブレーキングのために使用される複数のシステムＩＤセット内の複数のシステムＩＤを有する。それらシステムＩＤセット及びマスク値は、スパニングツリーの複数のルートの分散的かつ独立的な選択を可能とし、データセンタ内の負荷分散への単純なアプローチを提供する。スパニングツリーのルートの次善（second best）の選択（例えば、ＸＯＲマスキング後の２番目に小さいシステムＩＤを伴うルートノード）は、各システムＩＤセットにおける異なる第２のティアノードとなる。よって、第２のティアノードについて障害が起きた場合、当該ノードから発するスパニングツリーのセットを、１つよりも多くの他の第２のティアノードにわたって分配することができる。

【００１９】

折り返しＣＬＯＳネットワークについて、ルートノードから発するスパニングツリーは、エッジノードから発する等コストツリーと同じ接続性を提供する。しかしながら、等コストツリーに対して、スパニングツリーを使用するいくつかの利点が存在する。スパニングツリーの使用は、“任意ソース（any source）”というマルチキャストアドレッシングの使用を許容し、それによりネットワーク内のマルチキャストステートが大幅に削減される。対比において、“ソース固有（source specific）”マルチキャストアドレッシングを使用する等コストツリーは、所与のマルチキャストグループについてのマルチキャストソースノードの数をＳとすると、オーダ（Ｓ）だけ多くのマルチキャストアドレスをスイッチフィルタリングデータベース内にもたらす。

【００２０】

さらに、ここで開示される実施形態は、マルチキャストの接続性を構築する中で、３つの形式のバックボーン宛て先ＭＡＣアドレスを利用し、その３つは、８０２．１ａｑマルチキャストＭＡＣアドレッシング（“ソース固有”マルチキャストアドレッシングあるいは（Ｓ，Ｇ）アドレッシングともいう）、８０２．１ａｈＭＡＣアドレッシング（“任意ソース”マルチキャストアドレッシングあるいは（＊，Ｇ）アドレッシングともいう）、及び既存のバックボーンユニキャストトンネリングの再利用、を含む。本実施形態では、ルートノードの部分的な断絶（partial severing）を扱うために、“スプリットホライズン”ツリーアプローチが採られる。障害を起こした隣接関係（failed adjacency）にアタッチしているノードは、そうではないノードとは異なる形式のアドレッシングを使用する。障害を起こした隣接関係にアタッチしているノードは、（＊，Ｇ）アドレッシングを使用し得る一方、ＥＣＴセット内の障害を起こした隣接関係に直接的にはアタッチしていないノードは、ＢＵＭ（broadcast/unknown/multicast）データフレームを、（＊，Ｇ

）アドレッシング及び（Ｓ，Ｇ）（あるいはユニキャスト）アドレッシングのハイブリッドを用いてバイキャスト（bi-cast）する。

【００２１】

ここで説明されるハイブリッドアドレッシングのスタイルは、ノード内に維持されるマルチキャストステートを削減することにより、ネットワークのスケーラビリティを改善する。８０２．１ａｈによれば、（＊，Ｇ）マルチキャストのためのバックボーン宛て先ＭＡＣアドレスは、固定的なＯＵＩ（Organizationally Unique Identifier）（Ｉ－ＳＩＤについての全てのソースのセットである＊を表す）と、Ｉ－ＳＩＤなどのサービス識別子（Ｇを表す）との連接（concatenation）として符号化される。８０２．１ａｑによれば、（Ｓ，Ｇ）マルチキャストのためのバックボーン宛て先ＭＡＣアドレスは、マルチキャストツリールート（Ｓを表す）とＩ－ＳＩＤなどのサービス識別子（Ｇを表す）との連接として符号化される。単一の受信機が存在するユニキャストについて、バックボーン宛て先ＭＡＣアドレスは、例えば４６ビットの値などの、固定長のビット値である。いかなる個別のソースも自身のトラフィックに行き当たらないというスプリットホライズンに依存して、（＊，Ｇ）マルチキャストツリーは、任意のソースノードにより、Ｇ内の全ての受信機へ到達するために使用されることができる。（Ｓ，Ｇ）マルチキャストツリーは、ツリーが全ての“Ｓ”について同じようにルーティングされる（co-routed）わけではないことによるＦＤＢコンフリクトに起因してスプリットホライズンが不可能な最短パスツリーのために必要とされる。（Ｓ，Ｇ）マルチキャストツリーは、ツリーが各“Ｓ”について個別化（personalize）されることから、単一のソースによってＧ内の受信機のセット又はサブセットへ到達するために使用されることができる。従って、（＊，Ｇ）ツリーと等価なものを構築するために複数の（Ｓ，Ｇ）ツリーを要し、結果として、それら複数の（Ｓ，Ｇ）ツリーは、１つの（＊，Ｇ）マルチキャストツリーと同じことを実質的に行うために、ルートノード内でより多くのステート（state）を要する。ユニキャストは、ノードグループ（Ｇ）内に単一の受信機が存在する場合に使用される。ユニキャストの転送パスはサービス固有のマルチキャストツリーからは独立して存在するため、ユニキャストは、ステートを追加しない。従って、（＊，Ｇ）が使用されず又は使用不能である場合、（Ｓ，Ｇ）に対しユニキャストを使用することが有益である。

【００２２】

ここでは特定のバージョンの標準が説明されるが、本発明の実施形態は現時点のバージョンの標準に基づく実装には限定されず、将来のバージョンの標準が開発されればそれらと共に作動するように適合され得る。同様に、本発明の実施形態は、ここで説明される特定のプロトコルの１つとの関係で動作する実装には限定されず、イーサネットマルチエリアルーティングネットワークにおいて他のプロトコルもまた使用され得る。

【００２３】

図１は、３段折り返しＣｌｏｓネットワーク１０の一例としてのネットワークトポロジーの図である。折り返される入口／出口ステージ（図示した下のステージ）内のノードをエッジノード１２といい、中間ステージ（図示した上のステージ）内のノードをルートノード１１という。ルートノード１１及びエッジノード１２を、総称として“ノード”という。各ノード１１、１２は、自身の入力ポートの全てを自身の出力ポートへと相互接続するスイッチングエレメントである。当該ネットワークは、フレームにより横断されるノードの最大の数が３つ（即ち、エッジからルート、そしてエッジへ）であることから３段（three stage）である。各ルートノード１１は、エッジノード１２の全てへ接続される。各ルートノード１１は、エッジノード１２との間の送受信のための複数の入力／出力ポートを含み、各エッジノード１２もまた、ルートノード１１との間の送受信のための複数の入力／出力ポートを含む。追加的に、各エッジノード１２は、ネットワーク１０の外部との間でトラフィックを送受信するための、複数のユーザに面する（user-facing）入力／出力ポート１３をも含む。

【００２４】

図１の例としてのネットワークは、キャパシティが同一であるノードから作られ、８つ

のエッジノード１２を相互接続する４つのルートノード１１を含む。別の実施形態では、異なる数のルートノード１１及びエッジノード１２が含まれてもよい。１つの実施形態において、ネットワーク１０は、データセンタ内にある。１つの実施形態において、各ルートノード１１は、バックボーンコアブリッジ（ＢＣＢ）であり、各エッジノード１２は、バックボーンエッジブリッジ（ＢＥＢ）である。

【００２５】

一般には、３段折り返しＣＬＯＳネットワークのノンブロッキング特性は、そのルートノード及びエッジノード内のポートの数により左右され得る。例えば、ノンブロッキングなＣＬＯＳネットワークにおけるルートノード１１の最大数は、ノード別ポート数を２で割った商であり、エッジノード１２内のユーザに面するポート１３の数は、ルートノードの数とノード別ポート数との積である。

【００２６】

簡明さのために、以下の説明は、基礎となるネットワークとして、破線のボックス１５内に示したようなネットワーク１０の一部を使用する。理解されることとして、本技術は、様々な数のルートノード及びエッジノードを伴う３段折り返しＣＬＯＳネットワークに適用可能である。図２の実施形態に示したように、ルートノード２１がラベルＡ、Ｂ、Ｃ及びＤを付与されており、エッジノード２２がラベルｗ、ｘ、ｙ及びｚを付与されている。１つの実施形態において、ルートノード２１には、複数のシステムＩＤセット（例えば、セット１、セット２及びセット３）が割り当てられ、各システムＩＤセットはルートノード２１ごとに区別される（distinct）システムＩＤを含む。異なるシステムＩＤセット内の同じルートノード２１についてのシステムＩＤは、同じであっても異なってもよく、例えば、ノードＡは３つのシステムＩＤセットの中でそれぞれシステムＩＤ００、００、００を有し、ノードＢは３つのシステムＩＤセットの中でそれぞれ０１、０２、０２を有する。

【００２７】

複数のシステムＩＤセットの使用は、Ｂ－ＶＩＤについてスパニングツリーのルートノードを選択する際のスプリットタイブレーキングを可能とする。図３を参照しながら詳細に説明されるスプリットタイブレーキングの仕組みによれば、Ｂ－ＶＩＤ１、５、９（ｖ１／５／９として示されている）についてのスパニングツリーは、ノードＡから発する（rooted）。図２には示していないが、ルートノードＢはＢ－ＶＩＤ２、６、１０についてのスパニングツリーのルートであり、ルートノードＣはＢ－ＶＩＤ３、７、１１についてのスパニングツリーのルートであり、ルートノードＤはＢ－ＶＩＤ４、８、１２についてのスパニングツリーのルートである。よって、ヘッダ内でＢ－ＶＩＤ１、５又は９を搬送するデータフレームは、エッジノードのいずれからもネットワークへ進入することができ、ルートノードＡを通過して、他のエッジノードのいずれかを通じて退出する。ノードＡに障害が起き又はラベリングされたリンクのいずれかに障害が起きた場合、影響を受けるトラフィックを、他のルートノードを介する他のリンクへと再分配することができる。障害の無い（fault free）シナリオでは、エッジノードｗ、ｘ、ｙ及びｚの各々が、他のエッジノードへと到達するために、８０２．１ａｈの“任意ソース”マルチキャストＭＡＣ（＊，Ｇ）アドレッシングを用いて、データフレームを転送する。なお、顧客レイヤのトラフィックは、それがバックボーンレイヤでどのように転送されたかとは別に、顧客レイヤではブロードキャスト、ユニキャスト又はマルチキャストのいずれかである。

【００２８】

図３は、スパニングツリールート選択のために使用されるスプリットタイブレーキングの仕組みの一実施形態を示している。スプリットタイブレーキングの仕組みを説明する前に、８０２．１ａｑにおいて定義されているタイブレーキングの仕組みを説明することが有益である。８０２．１ａｑは、通常、ネットワーク内のトラフィックの各送信元（source）から発する、対称合同的（symmetrically congruent）な最短パスツリーのフルメッシュを生成する。１つのそうしたフルメッシュは、等コストツリー（ＥＣＴ）セットとして知られている。ＥＣＴセットは、通常、Ｂ－ＶＩＤに関連付けられる。ＥＣＴセットの

生成の一部としてのパス計算が１つよりも多くの等コストパスからの選択の必要性に至った場合、８０２．１ａｑは、ノードＩＤの辞書順の並び替えを用いて各等コストパスについての一意なパスＩＤを構築し、パスＩＤのセットをソートし、最も小さい値を選択する。追加的に、８０２．１ａｑは、各ＥＣＴセットに関連付けられるセットの値のノードＩＤとのＸＯＲ演算を介して複数のＥＣＴセットを生成し、各パスＩＤにおけるノードＩＤの辞書順の並び替えを改訂し、パスＩＤのランク付けを再度行い、最も小さい値を再選択するための手段を仕様化している。

【００２９】

ここで説明されるスプリットタイブレーキングの仕組みは、８０２．１ａｑに対する改善である。スプリットタイブレーキングの仕組みは、障害の無いシナリオのみならず１つ以上のルートノードに障害が起きた場合における、ＣＬＯＳネットワーク内のルートノードのセットにわたるトラフィックのより均等な分配を可能とする。スプリットタイブレーキングの仕組みは、次のように、いくつかの設計エレメントを使用する。（１）インスタンス化されるスパニングツリーの数は、ルートノードの数の何らかの倍数となるように選択される。その倍率は、障害の期間中にスプリットタイブレーカの性質を活用する（leverage）ことが望ましい場合、１よりも大きい。（２）ルートノードＩＤ（即ち、システムＩＤ）は、障害の無いＣＬＯＳネットワークにおいて各ルートノードから等しい数のスパニングツリーが発するように設計される。（３）エッジノードＩＤは、各エッジノードが決してスパニングツリーのルートにならないように設計される。これは、単純に、エッジノードＩＤの上位ビットに非ゼロの値を使用し、一方でルートノードは上位ビットにゼロを有し、それによりエッジノードＩＤのセットの最小値がルートノードＩＤのセットの最大値よりも大きくなるようにすることで、達成され得る。（４）（１）において倍率が１よりも大きい場合、スプリットタイブレーカ値は、障害の起きたルートノードのスパニングツリーのルート（root）が１つよりも多くの他のルートノードにわたって分配されるように設計される。

【００３０】

図３の実施形態において、システムＩＤセット２３内のシステムＩＤは、バイナリで示されている。加えて、ネットワーク内の各Ｂ－ＶＩＤには、例えばマスクセット３３に示した例のように、マスク値及びシステムＩＤセット番号が割当てられている。システムＩＤセット１に属するＢ－ＶＩＤ１～４の各々についてルートノードを決定するために、Ｂ－ＶＩＤ１～４という対応するマスク値を用いて、システムＩＤセット１について変換が行われる。１つの実施形態において、（所与のマスクを用いた）変換後の最小のシステムＩＤを有するルートノードが、当該所与のマスクに関連付けられるＢ－ＶＩＤについてのスパニングツリーのルートである。１つの実施形態において、上記変換はＸＯＲ演算である。例えば、マスク値００００を伴うＢ－ＶＩＤ１に関し、マスク値００００及びセット１：００００、０００１、００１０、００１１内のシステムＩＤについてＸＯＲが実行される。（変換されたシステムＩＤである）ＸＯＲの結果は、００００、０００１、００１０、００１１である。１つの実施形態において、最小のＸＯＲ値をもたらしたルートノードが、対応するＢ－ＶＩＤについてのスパニングツリーのルートとして選択される。よって、Ｂ－ＶＩＤ１について、４つの中でＸＯＲの結果として００００が最小であるため、ルートノードはノードＡである。

【００３１】

ノードＡの障害のケースでは、ノードＡを通過するトラフィックは、上で計算されたＸＯＲの結果に従って、他のルートノードへとリルートされ得る。例えば、次に小さいＸＯＲ値をもたらすルートノードを新たな通過ルートノードとして選択することができ、即ちノードＢである。従って、上で計算したＸＯＲの結果がＢ－ＶＩＤ１についてのフェイルオーバ順序を左右する。

【００３２】

同様に、Ｂ－ＶＩＤ５～８及びＢ－ＶＩＤ９～１２についてのルート選択を、同じスプリットタイブレーキングの仕組みで実行することができる。Ｂ－ＶＩＤ１～４、Ｂ－ＶＩ

Ｄ５～８及びＢ－ＶＩＤ９～１２は異なるセット内にあるため、各セット内のマスク値及びシステムＩＤは、他のセットには非依存で構成され得る。１つの実施形態において、管理システムは、各ルートノードをシステムＩＤ及びその対応するシステムＩＤセットと共に構成し得る。管理システムは、ネットワーク内で使用されるＢ－ＶＩＤにマスク値をも割り当て得る。システムセットＩＤ番号（例えば、セット１、セット２又はセット３）及び各Ｂ－ＶＩＤについてのマスク値は、拡張ＩＳ－ＩＳ（augmented Intermediate System to Intermediate System）ｈｅｌｌｏ手続を介してノード間で交換され、又は管理システムにより構成データとして各ノードへロードされ得る。

【００３３】

図４は、３段折り返しＣＬＯＳネットワーク内でスパニングツリーのための転送ステートを計算し及びインストールするための方法４００の一実施形態を示すフロー図である。１つの実施形態において、拡張ＩＳ－ＩＳのｈｅｌｌｏ手続を用いて、ノード間でタイブレーキング情報が交換される（ブロック４１０）。拡張ＩＳ－ＩＳ手続を用いて、例えば各Ｂ－ＶＩＤについて、パス生成アルゴリズム（例えば、スパニングツリー又は等コストツリー）、ルート選択のためのシステムＩＤセット、及び当該Ｂ－ＶＩＤについてのマスク値を交換することができる。例えば、ノードに関連付けられるＩＳ－ＩＳスピーカは、以下を広告し得る：Ｂ－ＶＩＤ１について、スパニングツリーを使用し、ルート選択のためにシステムＩＤセット１を使用し、マスク値＝００である；Ｂ－ＶＩＤ２について、スパニングツリーを使用し、ルート選択のためにシステムＩＤセット１を使用し、マスク値＝０１である、など。代替的な実施形態において、タイブレーキング情報は、管理システムにより各ノードへ構成され得る。タイブレーキング情報を取得した後、スパニングツリーのルートノードが、当該タイブレーキング情報に基づいて各Ｂ－ＶＩＤについて選択される（ブロック４２０）。図３において説明した通り、所与のＢ－ＶＩＤにより識別されるスパニングツリーについてのルートとして、全てのルートノードの中で最小のシステムＩＤを有するルートノードが選択されるように、システムＩＤの変換によってルート選択が行われる。ルート選択の後、各ルートノードについて、エッジノード群までのそのパスが識別される（ブロック４３０）。１つの実施形態において、ダイクストラのアルゴリズムを用いてネットワーク内の各ノードによりパスが計算される。その代わりに、異なるアルゴリズムが使用されてもよい。

【００３４】

ダイクストラのアルゴリズムが使用される実施形態において、計算の結果を用いて、リンク障害の影響を迂回するように転送を修正することができる。計算はネットワーク内の各ノードにより分散的なやり方で行われ、計算の結果は、ノードにより、ネットワーク内のそれらの位置に依存することなく使用され得る。よって、この情報を使用するためにノードは障害の起きたリンクに隣接していなくてよく、なぜならノードはグローバルのトポロジーの知識を取得し、その情報のローカルな個別化を計算するためである。ダイクストラのアルゴリズムは、プロトタイプツリーを生成し、プロトタイプツリーにおいて、１ホップの各パスはルートノードからエッジノードへのパスであり、それらはツリー内で維持され得る。１つより多くのホップを有するパスは、ツリーから除外（pruned）され得る（ブロック４４０）。例えば、ルートからエッジへ、さらにルートへ、という２ホップのパスが存在するかもしれない。それらパスは、プロトタイプツリーから除外され得る。リンク障害の標識として、３ホップパスが識別される（ブロック４５０）。３ホップパスの始点はスパニングツリーのルートであり、終点は当該ルートにより直接的に到達可能でないエッジノードである。よって、３ホップパスは、エッジから出発するトラフィックがスパニングツリーのルートを通過できず当該エッジノードへ至るために他のルート（root）を通過しなければならないことを示す。よって、それら３ホップパスをも検討から除外することができ、そのＢ－ＶＩＤ及び障害の起きたリンクに連結しているエッジノードについて、新たなルートが必要である。３ホップパスによって到達されるノードのリストは、ルートにもはや直接的にアタッチしていないエッジノードのセットとして、将来の計算のために別個に維持される。さらに、４ホップ以上のパスは、ネットワークに病的に（pathol

ogically）障害が起きていることを意味する。なお、上述したパスの除外及び障害検出は、例に過ぎず、最適化が存在してもよい。

【００３５】

スパニングツリーの生成の後、Ｂ－ＶＩＤに関連付けられる共通のサービス識別子を有する当該Ｂ－ＶＩＤについてのエッジノードの各ペアが識別される（ブロック４６０）。エッジ－エッジノードペアが識別されると、そのそれぞれのフィルタリングデータベース（ＦＤＢ）が、当該ノードペアの間でユニキャスト及びマルチキャストのデータフレームを転送するための転送エントリ（転送ステートともいう）を含むように構築される（ブロック４７０）。それらデータフレームは、ネットワークにより提供される対応するサービスを識別するために、ヘッダ内にＩ－ＳＩＤを含むことになる。エッジノードについては、スパニングツリーのルートへのＩ－ＳＩＤマルチキャストアドレスと共に、当該エッジノードのノードとしての（nodal）ユニキャストＢ－ＭＡＣ（Backbone MAC）アドレスへのＩ－ＳＩＤマルチキャストアドレスをポインティングすることにより、ＦＤＢエントリを生成することができる。ルートノードについては、エッジノードのペアへのマルチキャストエントリに加えて当該ルートノードのノードとしてのユニキャストＢ－ＭＡＣへのマルチキャストエントリを相互接続することにより、ＦＤＢエントリを生成することができる。ブロック４３０～ブロック４７０の動作は、スパニングツリーの各ルートについて繰り返される。

【００３６】

図５は、図２のネットワークの一実施形態を示しており、１つのルートノード（例えば、ノードＡ）が障害を起こす（“Ｆ”により示されている）。図３において説明したスプリットタイブレーキングの仕組みを用いることにより、ノードＡ上のデータ転送の作業負荷（workload）を、Ｂ－ＶＩＤ１、５及び９について２番目に小さいＸＯＲ結果（バイナリで０００１）を有する他のノードへとシフトすることができる。従って、ノードＢ、Ｃ及びＤが、Ｂ－ＶＩＤ１、５及び９についてそれぞれスパニングツリーのルートになる。作業負荷のシフトは、エッジノードにより使用されるマルチキャストアドレッシングを変化させない。よって、エッジノードｗ、ｘ、ｙ及びｚは、８０２．１ａｈの“任意ソース”マルチキャストＭＡＣ（＊，Ｇ）アドレッシング（簡明さのために（＊，Ｇ）アドレッシングともいう）を用いて、他のエッジノードへのマルチキャストデータフレームの転送を継続する。

【００３７】

図６～図１１は、１つ以上のリンクに障害が起きる、一例としての３段折り返しＣＬＯＳネットワークにおけるいくつかのシナリオを示している。図４のブロック４５０において説明したように、スパニングツリーのルートからエッジノードへのパスを計算するためにダイクストラのアルゴリズムが使用されるスパニングツリー生成の際に、リンク障害が検出され得る。図４のブロック４５０に示したように、スパニングツリー内の３ホップパスは、リンク障害及び影響を受けるノードを示している。理解されることとして、３段折り返しＣＬＯＳネットワークにおいてノード障害又はリンク障害の帰結を検出するために、他の方法が使用されてもよい。

【００３８】

図６は、ノードＡとノードＷとの間のリンクに障害が発生するシナリオを示している。ノードＡについてスパニングツリーが計算されると、ノードＷは、３ホップパスによりサービスされるものとして現れる。リンク障害を識別した後、障害の起きたリンクを使用する各Ｂ－ＶＩＤについてノードｗから発する最短パス優先（ＳＰＦ）ツリーが構築される。ＩＳ－ＩＳなどのリンクステートルーティングプロトコルを介して、各ノードは、ネットワーク内のトポロジー情報を学習し、当該情報をＳＰＦツリーを計算するために使用する。１つの実施形態において、各ノードにより、ダイクストラのアルゴリズムを用いてＳＰＦツリーを計算することができる。これらノードは、トポロジー情報に基づいてネットワークの同じビューを構築する。

【００３９】

１つの実施形態において、ＳＰＦツリーの構築は、どのルートノードを通過すべきかを決定するための、前述したスプリットタイブレーキングの仕組みの使用を含む。なお、これらＳＰＦツリーのルートはネットワークのエッジノードであり、ネットワークのルートノードはＳＰＦツリーの通過ノード（transit　nodes）になる。ＳＰＦツリーが構築された後、エッジノードは、Ｂ－ＶＩＤについてエッジ－エッジノードペアにより共有される、Ｉ－ＳＩＤのインターセクション（即ち、共通のＩ－ＳＩＤ）を識別する。識別されるエッジノードペアに基づいて、各ノードは、後続のデータ転送のために自身のＦＤＢにデータ投入を行う（populate）。

【００４０】

図６の例において、Ｂ－ＶＩＤ１、５及び９について作られるＳＰＦが、それぞれｖ１、ｖ５及びｖ９でラベリングされたリンクとして示されている。ラベルｖ１／５／９を伴うノードｘ、ｙ及びｚへ接続するリンクは、障害が無く、従って、ノードｘ、ｙ及びｚにより、それらの間でデータを転送するために、（＊，Ｇ）アドレッシングを用いて継続的に使用されることができる。一方、ノードｗと通信するためには、ノードｘ、ｙ及びｚは、ｖ１、ｖ５及びｖ９でラベリングされたＳＰＦツリーのリンクを使用する必要がある。よって、１つの実施形態において、エッジノードによりハイブリッドマルチキャストアドレッシングが使用される：ノードｘ、ｙ及びｚは、ノードｗへマルチキャストデータを転送するために、ユニキャスト又は８０２．１ａｑの“ソース固有”マルチキャストＭＡＣ（Ｓ，Ｇ）アドレッシング（簡明さのために（Ｓ，Ｇ）アドレッシングともいう）を使用し、ｘ、ｙ及びｚの間で使用される接続性が拡張され、一方で、ノードｗは、ノードｘ、ｙ及びｚへマルチキャストデータを転送するために、８０２．１ａｈマルチキャストＭＡＣ（＊，Ｇ）アドレッシングを使用する。これは、ノードｗからのマルチキャストは、全てのピアへ到達する必要があり、ｘ、ｙ及びｚからのマルチキャストは、それらが互いへとどのようにマルチキャストを行うかとは別に、ノードｗへ到達するために特有の扱いを必要とするからである。。

【００４１】

同じ原理が、図７～図１１に示した以下の例の各々に適用され、即ち、エッジノードが３ホップパスによってのみスパニングツリーから到達可能である場合、そのノードからのＳＰＦツリーが計算され、（＊，Ｇ）アドレッシングを用いてそのノードからマルチキャストデータが転送され、ユニキャスト又は（Ｓ，Ｇ）アドレッシングを用いてそのノードへマルチキャストデーが転送される。

【００４２】

図７は、ノードｗへ接続している２つのリンクに障害が発生する他のシナリオを示している。このシナリオにおいて、（ノードｗから発する）Ｂ－ＶＩＤ１についてのＳＰＦツリーはもはやノードＡ及びＢを通過しないため、ノードＣが通過ノードとして（例えば、タイブレーキングにより）使用される。Ｂ－ＶＩＤ５及びＢ－ＶＩＤ９についてのＳＰＦツリーは、以前としてノードＣ及びＤをそれぞれ通過することができる。ｖ１／５／９に関するアドレシング方式は以前として（＊，Ｇ）アドレッシングである。しかしながら、ノードｗと通信するために、ノードｘ、ｙ及びｚは、ｖ１／５（ｖ１及びｖ５を表す）並びにｖ９でラベリングされたＳＰＦツリーのリンクを使用する必要がある。よって、１つの実施形態において、ノードｘ、ｙ及びｚは、ノードｗへマルチキャストデータを転送するためにユニキャスト又は（Ｓ，Ｇ）アドレッシングを使用し、一方で、ノードｗは、ノードｘ、ｙ及びｚへマルチキャストデータを転送するために（＊，Ｇ）アドレッシングを使用する。

【００４３】

図８は、ノードＡへ接続している２つのリンクに障害が発生する他のシナリオを示している。このシナリオにおいて、ノードｙ及びｚは、それらの間で依然として、ｖ１／５／９リンクを介し、８０２．１ａｈマルチキャストＭＡＣ（＊，Ｇ）アドレッシングを用いて通信することができる。一方、ノードｘ及びｗと通信するために、ノードｙ及びｚは、ｖ１、ｖ５及びｖ９でラベリングされたＳＰＦツリーのリンクを使用する必要がある。よ

って、１つの実施形態において、共通のルートへの障害の起きた隣接関係を有する複数のノードが存在することから、ノードｙ及びｚは、ノードｘ及びｗへマルチキャストデータフレームを転送するために（Ｓ，Ｇ）アドレッシングを使用し、一方で、ノードｗ及びｘは、互いに並びにノードｙ及びｚへマルチキャストを行うために８０２．１ａｈマルチキャストＭＡＣ（＊，Ｇ）アドレッシングを使用する。

【００４４】

図９は、ノードＡ及びＷの間の第１のリンクに障害が起き、並びにノードＢ及びｘの間の第２のリンクにも障害が起きる他のシナリオを示している。ノードＡについてスパニングツリーが計算されると、ノードＷは、３ホップパスによりサービスされるものとして現れる。Ｂ－ＶＩＤ１についてのＳＰＦツリーはノードｗがノードｘと通信するための（ノードＢとノードｘとの間の）有効なパスをもはや有しないため、この障害の起きたパスをパスｗ－Ｃ－ｘによって置換えることができ、これは１つよりも多くのルートを通過するノードｗからのＳＰＦツリーの計算の結果として判定される。ラベルｖ１／５／９を伴うノードｘ、ｙ及びｚへのリンクは、障害が無く、従って、ノードｘ、ｙ及びｚにより、それらの間で（＊，Ｇ）アドレッシングを用いてデータを転送するために継続的に使用されることができる。一方、ノードｗと通信するために、ノードｘ、ｙ及びｚは、ｖ１、ｖ５及びｖ９でラベリングされたＳＰＦツリーのリンクを使用する必要がある。よって、１つの実施形態において、ノードｘ、ｙ及びｚは、ノードｗへマルチキャストデータを転送するためにユニキャスト又は（Ｓ，Ｇ）アドレッシングを使用し、一方で、ノードｗは、ノードｘ、ｙ及びｚへマルチキャストデータを転送するために（＊，Ｇ）アドレッシングを使用する。

【００４５】

図１０は、ノードｗへ接続する４つのリンクのうちの３つに障害が起きる他のシナリオを示している。このシナリオにおいて、ラベルｖ１／５／９を伴うノードｘ、ｙ及びｚへのリンクは、障害が無く、従って、ノードｘ、ｙ及びｚにより、それらの間で８０２．１ａｈマルチキャストＭＡＣ（＊，Ｇ）アドレッシングを用いてデータを転送するために継続的に使用されることができる。一方、ノードｗとの通信は、ノードｗとノードＤとの間を接続する唯一の有効なリンクを介する。よって、（ｖ１／５／９でラベリングされた）Ｂ－ＶＩＤ１、５及び９についてのノードｗから発する３つのＳＰＦツリーは、全てノードＤを通過する。ノードｗと通信するために、ノードｘ、ｙ及びｚは、ｖ１／５／９でラベリングされたＳＰＦツリーのリンクを使用する必要がある。よって、１つの実施形態において、ノードｘ、ｙ及びｚは、ノードｗへマルチキャストデータを転送するためにユニキャスト又は（Ｓ，Ｇ）アドレッシングを使用し、一方で、ノードｗは、ノードｘ、ｙ及びｚへマルチキャストデータを転送するために（＊，Ｇ）アドレッシングを使用する。

【００４６】

図１１は、ノードＡへ接続する４つのリンクのうちの３つに障害が起きる他のシナリオを示している。このシナリオにおいて、ノードｗとの通信は、ノードｗへ接続する唯一の有効なリンク（図中のｖ１／５／９でラベリングされた左端のリンク）を介する。Ａ－ｚの間のリンクは、ノードｚが他のいずれのエッジノードへ到達するためにもこのリンクを使用できないことから、使用されない。よって、ルートノードＡは、有効なデータ転送のためにエッジノードのいずれにもサービスできないことから、ネットワークから“事実上”断絶されている。このシナリオは、４つのマルチキャストアドレスを使用する：ノードｗ、ｘ及びｙは、８０２．１ａｑマルチキャストＭＡＣ（Ｓ，Ｇ）アドレッシングを使用し、ノードｚは、８０２．１ａｈマルチキャストＭＡＣ（＊，Ｇ）アドレッシングを使用する。加えて、ノードＡから発する１つのスパニングツリー、並びに、ノードｗ、ｘ及びｙからそれぞれ発する３つのスプリットホライズン計算が存在する。

【００４７】

図１２は、３段折り返しＣＬＯＳネットワークにおいて障害が発生した場合のデータフレーム転送のための方法１２００を示すフロー図である。１つの実施形態において、方法１２００は、ルートノードから発するスパニングツリーについての転送ステートを、各ノ

ードが計算し及びインストールすることから開始され（ブロック１２１０）、ここで、スパニングツリーにおけるデータ転送は、ネットワークに障害が無い場合にはマルチキャストパスについて８０２．１ａｈのマルチキャストアドレスを利用する。（例えば、図４にて上述したようにダイクストラのアルゴリズムによって）所与のルートノードへの障害を起こした接続を所与のエッジノードが有することが検出された場合（ブロック１２２０）、当該所与のエッジノードから発するＳＰＦツリーが、上記所与のエッジノードから発し障害の起きた接続を使用するＢ－ＶＩＤのセットの各々について構築される（ブロック１２３０）。上記所与のエッジノードと他のエッジノードとにより共有される共通のＩ－ＳＩＤが識別され、この情報を使用して、Ｂ－ＶＩＤのセットに関連付けられるサービスを共通して有するように識別されるエッジノードペアについて、上記所与のエッジノードと他のエッジノードとの間のペアごとの接続性のために、各ノード内のＦＤＢへデータが投入される（ブロック１２４０）。ＦＤＢ内にユニキャストステートがインストールされる（ブロック１２５０）。トラフィックが所与のエッジノードへと向かうのか所与のエッジノードから来るのかに依存して選択されるマルチキャストアドレッシングのハイブリッドを用いて、ＦＤＢ内にマルチキャストステートもまたインストールされる（ブロック１２６０）。そして、ノードは、ＦＤＢに従って、ユニキャスト及びマルチキャストデータを転送する（ブロック１２７０）。データは、ハイブリッドマルチキャストアドレッシングに従って転送され、それと共に、上記所与のエッジノードは８０２．１ａｈマルチキャストＭＡＣアドレッシングを用いてＳＰＦツリーを介して他のエッジノードへデータフレームを転送し、他のエッジノードはユニキャスト又は８０２．１ａＱマルチキャストＭＡＣアドレッシングを用いてＳＰＦツリーを介して上記所与のエッジノードへデータフレームを転送する。

【００４８】

図２～図１２を参照しながら上で説明した方法４００及び１２００は、さらに最適化されてもよい。図２の例において、Ｂ－ＶＩＤ１～１２について１２個のスパニングツリーが生成される（３つのスパニングツリーのみが示されている）。実際には４つのルートのみが存在するため、４つのスパニングツリーの計算の結果が再利用されてもよい。同様に、２つのリンクに障害が起きる最初のシナリオ（図７）及び３つのリンクに障害が起きる最初のシナリオ（図１０）において、ノードｗからの計算が再利用されてもよい。

【００４９】

より小規模なネットワークについて、ある程度の事前の計算と、別個のシステムとしての又はノードに統合されるより簡易な内部システム管理機能とのルーティングシステムの置換えと、を計画することも可能であり、例えば、８ポートのスイッチから作られるＣＬＯＳネットワークが３２個のリンクを有する。全ての単一のリンク及びノードの障害シナリオの疎なテーブルを前もって構築することもでき、当該テーブルは、Ｂ－ＶＩＤに対し、どのルート及びどのアドレッシングを使用すべきかをマッピングする。例えば、テーブル内には（２３２のあり得るネットワークステートに対して）約４０個のエントリがある。エッジノードについては、正確なルートノードを転送エントリで指し示すことによって、ＦＤＢエントリを生成することができる。ルートノードについては、各エッジペアについての対象のＩ－ＳＩＤのインターセクションを判定し、それに応じてルートＦＤＢにデータ投入することにより、ＦＤＢエントリを生成することができる。

【００５０】

上の説明は、移される負荷が生き残るノードにわたって可能な限り均等に共有されるように、障害時の負荷を拡散することを扱っている。しかしながら、全ての顧客がノンブロッキングなサービスからブロッキングなサービスを有するように遷移することになるため、均等な負荷の分配が常に望ましいわけではないかもしれない。１つの実施形態において、障害シナリオの下でのノンブロッキング性を保全するために、（スパニングツリーについてのそれぞれのＢ－ＶＩＤにより識別される）ＶＬＡＮのサブセットへプライオリティが付与され得る。プライオリティの与えられたスパニングツリーのセットの外側での障害は、それらセットの内側には拡散されない。それらセットの内側の障害は、それらセット

の外側へと拡散される。このようにして、顧客のサブセットに、障害に直接的に影響されない場合にはいつでも、ノンブロッキングな振る舞いを保証することができる。

【００５１】

さらに、その時点の（current）トラフィックパターンの知識もまた活用されてよい。例えば十分に利用されているネットワークのような、データセンタ内では、トラフィックは均等に分散しているものと想定される。アルゴリズムを、利用度の低い（under-utilized）ネットワークを活用するように開発することができる。例えば、利用度の低いツリーは、障害時にシフトされるトラフィックのより大きなシェアを受け取ることができる。従って、ネットワークは、真にノンブロッキングではなく、提供されるその時点の負荷に基づいてノンブロッキングであるだけである。

【００５２】

さらに、バックアップノード／パスの構成の変更を、ヒットレスに（即ち、データのロスなく）提供することができる。所与のシステムＩＤセット内の障害時の第２の又は第３の候補としてのノードの位置を、当該システムＩＤセット内でのそのシステムＩＤの変更を介して修正することができる。管理システムは、どのようにトラフィックがシフトされるかを修正するために、それらシステムＩＤを構成し及び調整することができる。その時点の転送パターンが最小のＸＯＲタイブレーカ値に基づくものと仮定すると、次の最良のものの序列を、転送パターンに影響を与えることなくサービス中に修正することができる。

【００５３】

図１３は、本発明の一実施形態を実装するために使用され得る３段折り返しＣＬＯＳネットワークエレメントの一例を示している。ネットワークエレメント３１０は、上述した３段折り返しＣＬＯＳネットワーク内の任意のノード（エッジノード又はルートノード）であってよい。

【００５４】

図１３に示したように、ネットワークエレメント３１０は、スイッチングファブリック３３０、複数のデータカード３３５、受信（Ｒｘ）インタフェース３４０、送信（Ｔｘ）インタフェース３５０及びＩ／Ｏポート３５５を含むデータプレーンを含む。Ｒｘ及びＴｘインタフェース３４０及び３５０は、Ｉ／Ｏポート３５５を通じて、ネットワーク内のリンクとインタフェースする。ネットワークエレメントがエッジノードである場合、Ｉ／Ｏポート３５５は、ネットワークの外部との間の通信を提供するための、ユーザに面する複数のポートをも含む。データカード３３５は、インタフェース３４０及び３５０上で受信されるデータについての機能を実行し、スイッチングファブリック３３０はデータカード及びＩ／Ｏカードの間でデータをスイッチングする。

【００５５】

ネットワークエレメント３１０は、制御プレーンをも含み、制御プレーンは、データトラフィックの経路制御、転送及び処理をハンドリングするように構成される制御ロジックを収容する１つ以上のネットワークプロセッサ３１５を含む。ネットワークプロセッサ３１５は、スパニングツリールート選択のためのスプリットタイブレーカを実行し、スパニングツリーについての転送ステートを計算し及びインストールし、リンク障害の発生時にＳＰＦツリーを計算し、並びにデータ転送のためにＦＤＢ３２６にデータ投入を行うようにも構成される。制御ロジック内に、他の処理もまた実装されてよい。

【００５６】

ネットワークエレメント３１０は、ＦＤＢ３２６及びトポロジーデータベース３２２を記憶するメモリ３２０をも含む。トポロジーデータベース３２２は、ネットワークのリンクステートを含む、ネットワークモデル又は類似のネットワークトポロジーの表現を記憶する。ＦＤＢ３２６は、１つ以上の転送テーブル内にネットワークエレメント３１０の転送ステートを記憶させ、それらはネットワークエレメント３１０へのインカミングのトラフィックをどこへ転送すべきかを示す。

【００５７】

１つの実施形態において、ネットワークエレメント３１０は、管理システム３８０へ連結され得る。１つの実施形態において、管理システム３８０は、メモリ３７０へ連結される１つ以上のプロセッサ３６０を含む。プロセッサ３６０は、システムＩＤ及びネットワークエレメント３１０の動作を構成するためのロジックを含み、当該動作は、システムＩＤの更新及びそれによるネットワーク内の作業の分散と、スパニングツリーのサブセットへの、ネットワークのノンブロッキング特性を少なくともそれらスパニングツリーについて維持するようなプライオリティの割り当てと、を含む。１つの実施形態において、管理システム３８０は、各ノードについて転送テーブルを計算し、そして転送テーブルをノードへダウンロードするという、システム管理機能を実行してもよい。システム管理機能は、（破線で示されているように）オプションであり、代替的な実施形態において、分散型のルーティングシステムは、各ノードが自身の転送テーブルを計算するように、上記計算を実行してもよい。

【００５８】

ここで説明した実施形態の利点の１つは、ネットワーク内に合計でＮ個のノードのあるネットワークについて、スパニングツリーのルート選択に適合されたスプリットタイブレーカの使用が、ルートの障害時に計算の負荷が大幅に削減されることを意味し、それにより、計算の複雑さは、（生き残るルートの数）×Ｏ（ＮｌｎＮ）となる。これは、複雑さがＯ（$Ｎ^{2}$ ｌｎＮ）である８０２．１ａｑに対して、大幅な改善である。ここで説明した実施形態について、計算上の複雑さは、障害の無いシナリオ及び障害シナリオの双方で削減される。

【００５９】

リンク障害のシナリオを解決するためのスプリットホライズン型のルーテッドツリーの使用もまた、計算上の複雑さを低減し、それにより、複雑さは、（ルートの数）×Ｏ（ＮｌｎＮ）＋（障害リンクに隣接するエッジの数）×Ｏ（ＮｌｎＮ）となる。対比として、８０２．１ａｑについての名目上の複雑さは、Ｏ（$Ｎ^{2}$ ｌｎＮ）である。ここで説明した実施形態についてあらためて言うと、計算上の複雑さは、障害の無いシナリオ及び障害シナリオの双方で削減される。

【００６０】

さらに、スプリットホライズン型のルーテッドツリーの使用は、障害リンクに隣接する他のエッジノード（例えば、図６～図１１の例におけるノードｗ）への送信を行うエッジノードだけが、それらの通常のスパニングツリー（＊，Ｇ）と、当該他のエッジノード（例えば、ノードｗ）へのユニキャストパスと、へバイキャストする必要があることを意味し、一方で、当該他のエッジノード自体（例えば、ノードｗ）は、（＊，Ｇ）アドレッシングを使用し続けることができる。

【００６１】

ルート選択のためのノードとしてのタイブレーカ値の使用は、ルートが他の手段で明示的に識別される必要の無いことを意味し、よって構成ミスをすることがより少ない。

【００６２】

上述の機能は、コンピュータ読取可能なメモリに記憶されてネットワークエレメントに関連付けられたコンピュータプラットフォーム上の１つ以上のプロセッサ上で実行される、プログラム命令のセットとして実現されてもよい。しかしながら、ここで説明した全てのロジックを、複数の個別の構成要素、特定用途向け集積回路（ＡＳＩＣ）等の集積回路、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）若しくはマイクロプロセッサ等のプログラマブルロジックデバイスと一緒に使用されるプログラマブルロジック、ステートマシーン、又は、それらのいずれの組み合わせをも含む他のいずれの装置を使用しても具現化することができることは、当業者には明らかであろう。プログラマブルロジックは、読取専用メモリチップ、コンピュータメモリ、ディスク、又は他の記憶媒体等の有形の媒体に、一時的に又は永久的に固定されることができる。また、プログラマブルロジックは、搬送波で具現化されるコンピュータデータ信号内に固定されることもでき、それにより、プログラマブルロジックを、コンピュータバス又は通信ネットワーク等のインターフェ

イス上で送信することが可能になる。このようなすべての実施形態は、本発明の範囲に含まれることが意図される。

【００６３】

図４及び図１２のフロー図の動作を、図１、図２及び図１３の例示的な実施形態を参照しながら説明した。しかしながら、図４及び図１２のフロー図に示す動作を、図１、図２及び図１３を参照して議論した実施形態以外の本発明の実施形態によっても実行することができること、また、図１、図２及び図１３を参照して議論した実施形態が、図４及び図１２の図を参照して議論した動作とは異なる動作を実行することができること、が理解されるべきである。図４及び図１２の図は、本発明の特定の実施形態によって実行される動作の具体的な順序を示しているが、この順序は例示であること（例えば、代替の実施形態では、動作が異なる順序で行われ、何らかの動作が組み合わされ、何らかの動作が重複してもよいこと、など）が理解されるべきである。

【００６４】

本発明の多様な実施形態は、ソフトウェア、ファームウェア、及び／又はハードウェアの多様な組み合わせを使用して実装されてよい。従って、図示した技法は、１つ以上の電子デバイス（例えば、エンドステーション、ネットワークエレメント）に記憶されかつそこで実行されるコード及びデータを使用して実装されることができる。そのような電子デバイスは、非一時的なコンピュータ読取可能な記憶媒体（例えば、磁気ディスク、光ディスク、ランダムアクセスメモリ、読出し専用メモリ、フラッシュメモリデバイス、相変化メモリ）及び一時的なコンピュータ読取可能な伝送媒体（例えば、電気的、光学的、音響的、若しくは他の形式の伝搬信号－搬送波、赤外信号、デジタル信号等）、といったコンピュータ読取可能な媒体を使用して、コード及びデータを記憶し並びに（内部で及び／又はネットワークを通じて他の電子デバイスとの間で）通信する。加えて、そのような電子デバイスは、典型的に、１つ以上の記憶デバイス（非一時的な機械読取可能な記憶媒体）、ユーザ入出力デバイス（例えば、キーボード、タッチスクリーン、及び／又はディスプレイ）、並びにネットワーク接続等の、１つ以上の他のコンポーネントに連結された１つ以上のプロセッサのセットを含む。そのプロセッサのセットと他のコンポーネントとの連結は、典型的に、１つ以上のバス及びブリッジ（バスコントローラとも呼ばれる）を通じて行われる。よって、所与の電子デバイスの記憶デバイスは、典型的に、その電子デバイスの１つ以上のプロセッサのセット上での実行のためのコード及び／又はデータを記憶する。

【００６５】

ここで使用されるところでは、ネットワークエレメント（例えば、ルータ、スイッチ、ブリッジ、コントローラ）とは、ネットワーク上の他の機器（例えば、他のネットワークエレメント、エンドステーション）を通信可能に相互に接続する、ハードウェア及びソフトウェアを含む、１つのネットワーキング機器のことである。いくつかのネットワークエレメントは、“マルチサービスのネットワークエレメント”であり、これらは、複数のネットワーキング機能（例えば、ルーティング、ブリッジング、スイッチング、レイヤ２統合、セッションボーダ制御、サービス品質（Quality of Service）、及び／又は加入者管理）についてのサポートを提供し、及び／又は、複数のアプリケーションサービス（例えば、データ、音声、及びビデオ）についてのサポートを提供する。加入者エンドステーション（例えば、サーバ、ワークステーション、ラップトップ、ネットブック、パームトップ、携帯電話、スマートフォン、マルチメディアフォン、ＶＯＩＰ（Voice Over Internet Protocol）フォン、ユーザ機器、端末、携帯型メディアプレーヤ、ＧＰＳユニット、ゲームシステム、セットトップボックス）は、インターネットを通じて提供されるコンテンツ／サービス、及び／又は、インターネットにオーバレイされる（例えば、インターネット経由でトンネリングされる）仮想プライベートネットワーク（ＶＰＮ）上で提供されるコンテンツ／サービスにアクセスする。それらのコンテンツ及び／又はサービスは、典型的に、サービスプロバイダ又はコンテンツプロバイダに属する１つ以上のエンドステーション（例えば、サーバエンドステーション）により、又はピアツーピアサービスに

参加するエンドステーションにより提供され、例えば、パブリックウェブページ（例えば、無料コンテンツ、ストアフロント、検索サービス）、プライベートウェブページ（例えば、ｅメールサービスを提供する、ユーザ名／パスワードによりアクセスされるウェブページ）、及び／又は、ＶＰＮ上の企業ネットワークを含み得る。典型的に、加入者エンドステーションは、（例えば、アクセスネットワークに（有線又は無線で）連結された顧客構内機器（customer premise equipment）を通じて）エッジネットワークエレメントに連結され、それらエッジネットワークエレメントは、（例えば、１つ以上のコアネットワークエレメントを通じて）他のエッジネットワークエレメントに連結され、それら他のエッジネットワークエレメントは、他のエンドステーション（例えば、サーバエンドステーション）に連結される。

【００６６】

本発明をいくつかの実施形態に関して説明してきたが、本発明が上述の実施形態に限定されるものではなく、添付の特許請求の範囲の思想および範囲内で修正及び変形と共に実践されることができることを、当業者は認識するであろう。よって、本説明は、限定ではなく例示と捉えられるべきである。

【図１】

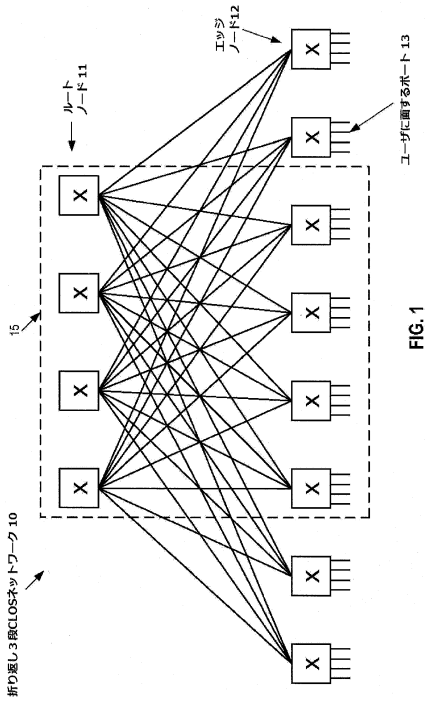


FIG. 1

【図２】

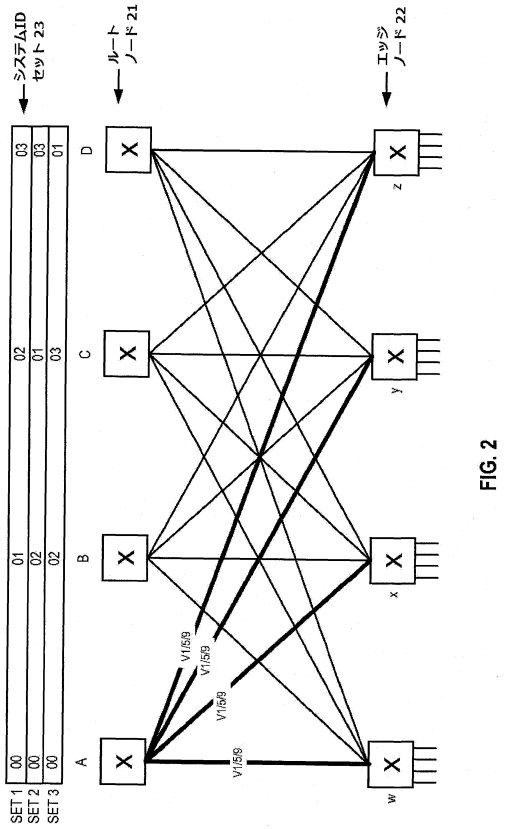
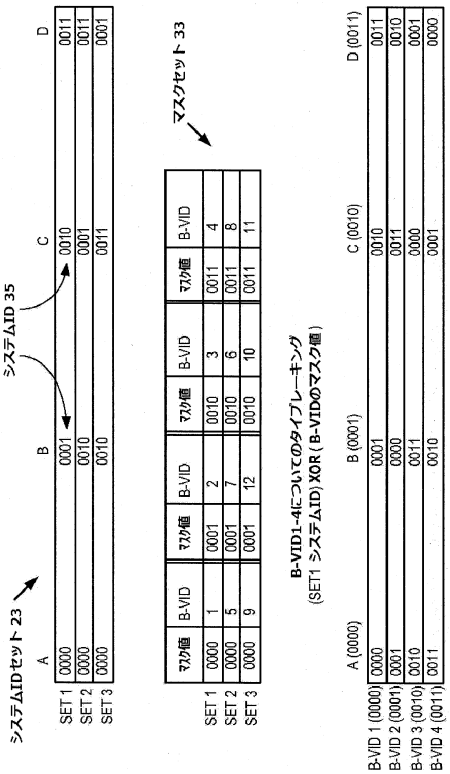


FIG. 2

【図３】



【図４】

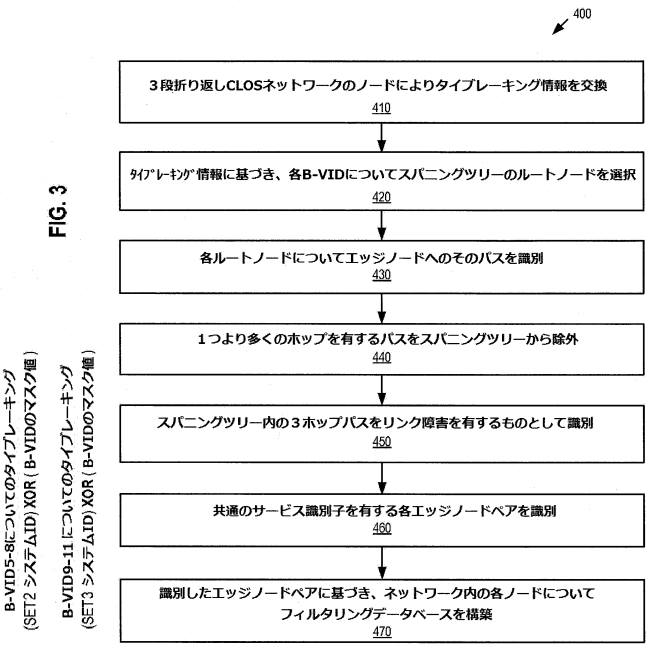


FIG. 3

B-VID5-8についてのタイプレーキング
(SET2 システムID) XOR (B-VIDのマスク値)

B-VID9-11についてのタイプレーキング
(SET3 システムID) XOR (B-VIDのマスク値)

FIG. 4

【図５】

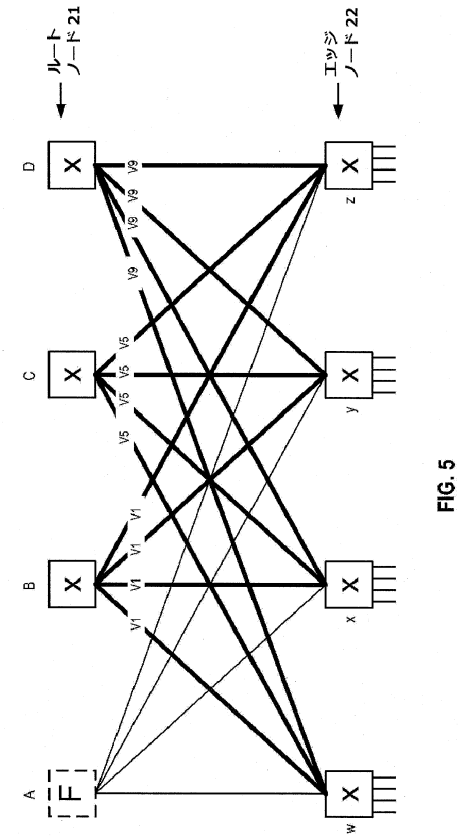


FIG. 5

【図６】

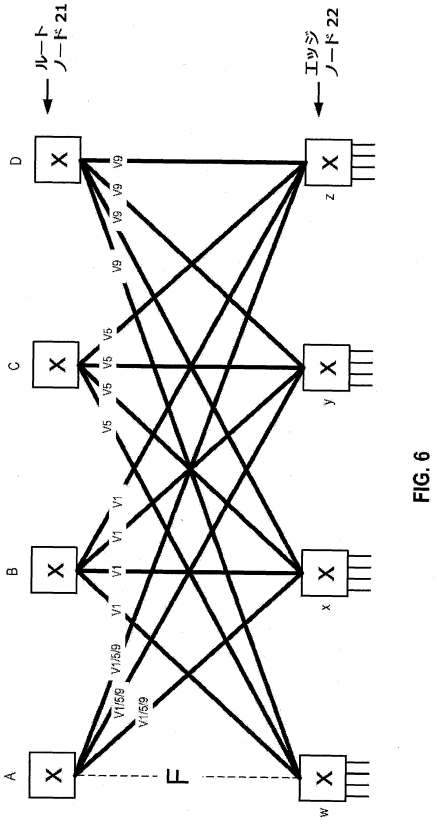


FIG. 6

【図７】

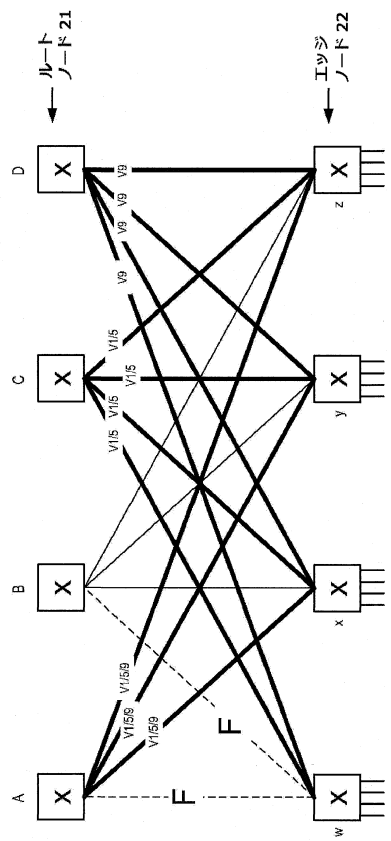


FIG. 7

【図８】

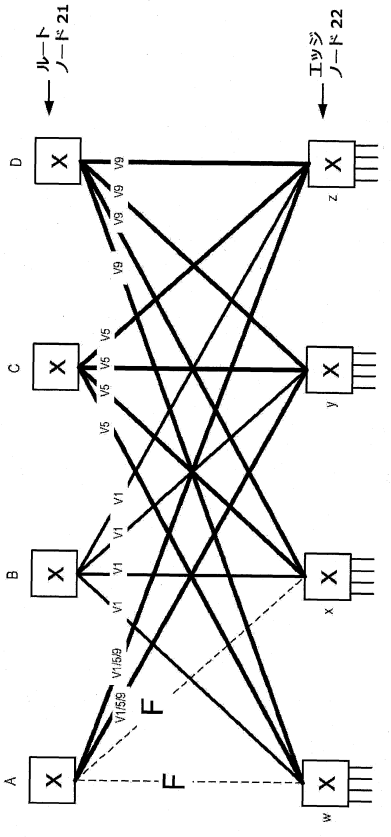


FIG. 8

【図９】

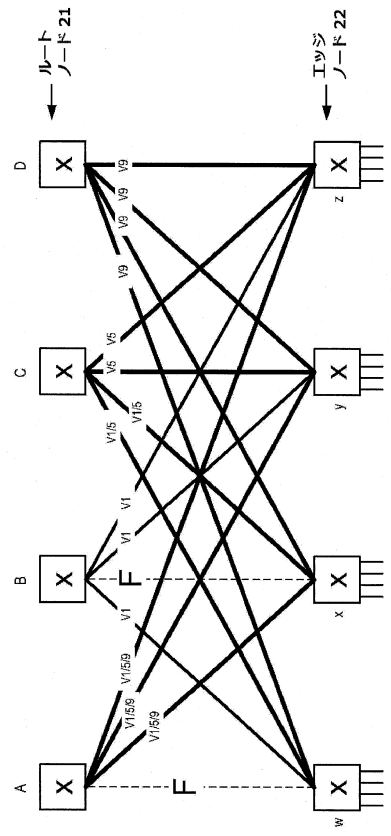


FIG. 9

【図１０】

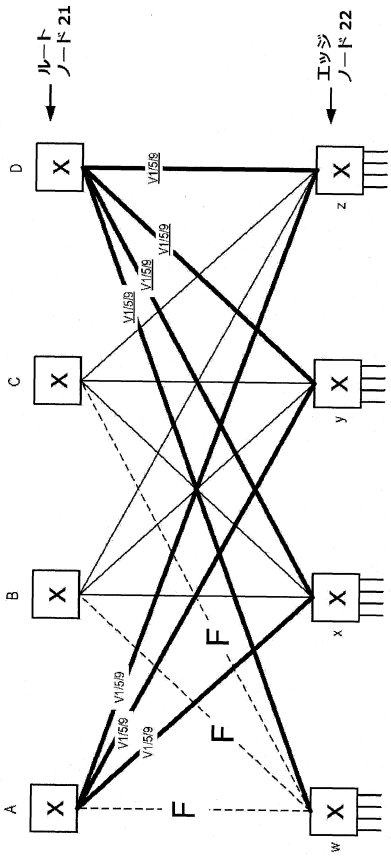


FIG. 10

【図１１】

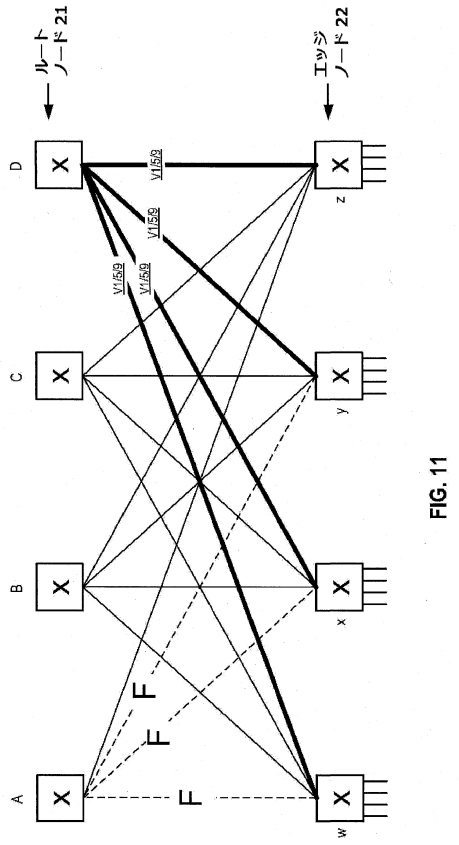


FIG. 11

【図１２】

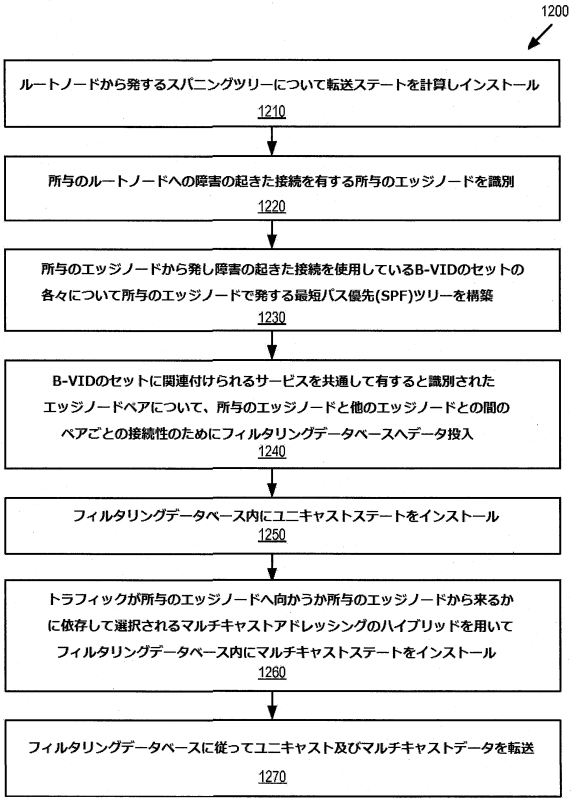


FIG. 12

【図１３】

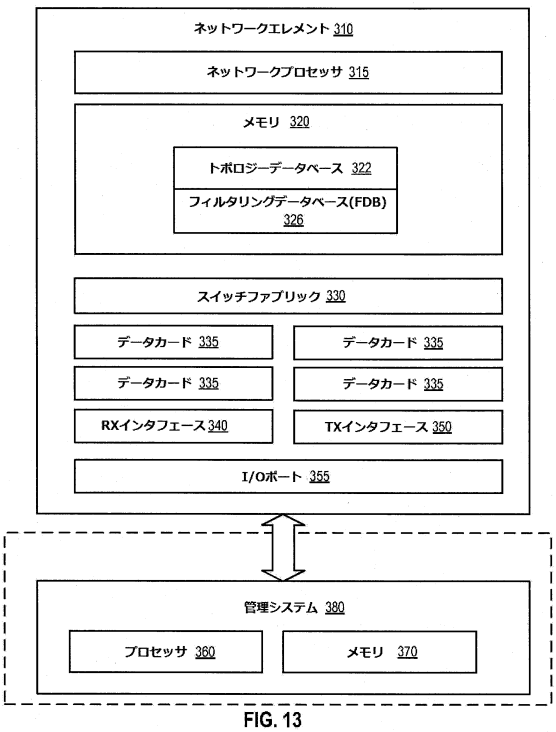


FIG. 13

フロントページの続き

(72)発明者　アラン、デイビッド　イアン
アメリカ合衆国　カリフォルニア州　９５１１２　サンノゼ　３０６　エス　フィフティーンス
ストリート

審査官　速水　雄太

(56)参考文献　特表２００８－５４６３３２（ＪＰ，Ａ）
米国特許出願公開第２００９／０２１３８６６（ＵＳ，Ａ１）
米国特許出願公開第２００５／０１１１３５６（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｌ　１２／７３３
Ｈ０４Ｌ　１２／７４１