(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3603524号
(P3603524)
(45)発行日　平成16年12月22日(2004.12.22)　(24)登録日　平成16年10月8日(2004.10.8)

(51)Int.Cl.[7]　F I
H04L 12/56　H04L 11/20　102D

請求項の数 8　(全 18 頁)

(21)出願番号　特願平9-22402
(22)出願日　平成9年2月5日(1997.2.5)
(65)公開番号　特開平10-224400
(43)公開日　平成10年8月21日(1998.8.21)
審査請求日　平成13年10月5日(2001.10.5)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100075096
弁理士　作田　康夫
(72)発明者　マシエル・フレデリコ
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社　日立製作所　中央研究所内
(72)発明者　北井　克佳
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社　日立製作所　中央研究所内

最終頁に続く

(54)【発明の名称】ネットワーキング方法

(57)【特許請求の範囲】
【請求項１】
各々が複数のデータ処理装置を相互接続し、かつ前記データ処理装置のサブセットによって相互接続された、第一と第二のネットワークにおいて、
前記第一ネットワークと、前記第二ネットワークの前記データ処理装置間を流れるネットワーキングトラヒック量を取得する第１段階と、
前記第一と第二のネットワークを相互接続する前記データ処理装置の処理負荷の測定値を取得する第２の段階と、
前記ネットワークトラヒック量と前記処理負荷の測定値に応じてルートを選択する第３の段階であり、第二ネットワークの各データ処理装置に対して、第一と第二のネットワークを相互接続する前記データ処理装置の間の前記ネットワーキングトラヒックを分配して、前記第一と第二のネットワークを相互接続する前記データ処理装置間の前記ネットワーキングトラヒックの平衡を保つためのアルゴリズムを利用することによって、第二ネットワークの前記データ処理装置からおよび／または前記データ処理装置へ行き来するトラヒックのためのルータとして利用される、第一と第二のネットワークを接続する前記データ処理装置を選択する段階とを有し、
第一と第二のネットワークを相互接続する前記データ処理装置間のネットワーキングトラヒックの平衡を保つためのルートを選択することを特徴とするネットワーキング方法。
【請求項２】
アドレス・レゾリューション・プロトコル（ＡＲＰ）とプロクシＡＲＰが第一ネットワー

クにおいて利用され、かつ第一と第二のネットワークを相互接続する前記データ処理装置間における前記第一ネットワークから前記第二ネットワークへのネットワークトラヒックに対して動的トラヒック平衡を与える請求項１のネットワーキング方法において、
プロクシＡＲＰによって、第一と第二のネットワークをルータとして選択されて相互接続している前記データ処理装置の第二ネットワークにおけるネットワークアドレスを、前記データ処理装置の前記第一ネットワークに接続したインターフェースの物理アドレスに関連させる第４の段階、および
上述の段階を周期的に繰り返す第５の段階、
を有することを特徴とするネットワーキング方法。
【請求項３】
前記第３の段階では、前記第一と第二のネットワーク間のネットワークトラヒックの転送を阻止する誤動作が存在しないデータ処理装置にのみネットワーキングトラヒックを分配することによって信頼性を与えることを特徴とする請求項１記載のネットワーキング方法。
【請求項４】
請求項１に記載のネットワーキング方法において、
第二ネットワークの前記各データ処理装置のルーティング表における前記第一ネットワークへのルートを、第一と第二のネットワークを相互接続し、第二ネットワークの前記データ処理装置に対するルータとして動作するよう選択された前記データ処理装置へ変更する段階、および
上述の段階を周期的に繰り返す段階、
を更に有することを特徴とするネットワーキング方法。
【請求項５】
オープン・ショーテスト・パス・ファースト（ＯＳＰＦ）プロトコルが前記第二ネットワークにおいて利用されて、前記第一と第二のネットワークを相互接続するデータ処理装置間における前記第二ネットワークから前記第一ネットワークへのネットワークトラヒックに対して動的トラヒック平衡を与える請求項１の記載のネットワーキング方法において、
ＯＳＰＦプロトコルを利用してデータ処理装置の１つにおいて行われようとする第一拡張、および
ＯＳＰＦプロトコルを利用して他のデータ処理装置において行われようとする第二拡張、
を更に有することを特徴とするネットワーキング方法。
【請求項６】
ＯＳＰＦプロトコルを利用してデータ処理装置の１つにおいて前記第一の拡張が行われる場合、
請求項１のルートを選択するネットワーキング方法によってルートを選択する段階、
選択されたルートに応じてＯＳＰＦプロトコルを利用しないで、自己のルートおよびデータ処理装置におけるルートを変更するようセットする段階、
ＯＳＰＦプロトコルを利用して、選択されたルートを前記他のデータ処理装置に同報通信する段階、および
上述の段階を周期的に繰り返す段階、
を有することを特徴とする請求項５のネットワーキング方法。
【請求項７】
ＯＳＰＦプロトコルを利用して、他のデータ処理装置において第二拡張が行われようとする場合、
第一拡張のデータ処理装置によって同報通信されたルートを受信する段階と、
受信したルートに応じて自己のルートをセットする段階、および
上述の段階を周期的に繰り返す段階、
を有することを特徴とする請求項５のネットワーキング方法。
【請求項８】
前記第二ネットワークに接続した前記データ処理装置は並列処理装置のノードであり、前

記第一ネットワークは前記並列処理装置を他のデータ処理装置に接続することを特徴とする請求項１，２，３，４，５，６，７のいずれかに記載のネットワーキング方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明はデータネットワーク間の相互接続に関し、特に、例えば、複数のルータ間における、あるいは並列処理装置の複数のインターフェース間におけるトラヒック負荷平衡に関する。

【０００２】

【従来の技術】

コンピュータネットワークは通常、他のコンピュータネットワークに接続して、インターネットを形成する。所定の２つのネットワーク間の接続は１つ以上のデータ処理装置によって実行される（Ｄ．Ｅ．カマー，“ＴＣＰ／ＩＰによるインターネットワーキング”Ｖｏｌ．１，プレンティス　ホール，１９９１年参照）。

【０００３】

これらのデータ処理装置には、限定するわけではないが、ルータ，ゲートウエイおよびスイッチが含まれており、以後、本願明細書では、一般に、ルータと称することにする。

【０００４】

図１に与えられた例は、複数のデータ処理装置（３ａ，３ｂ，３ｃ，３ｄ，　３ｅ，４ａ，４ｂ）及び３つのネットワーク（１，２，５）を示している。データ処理装置３ａと３ｂはネットワーク１と２間のルータとして動作し、データ処理装置３ｃはネットワーク２と５間のルータとして動作し、そしてデータ処理装置４ｂはネットワーク１と５間のルータとして動作する。データ処理装置はネットワークインターフェース（３０ａ，３０ｂ，３０ｃ，３０ｄ，３０ｅ，３１ａ，３１ｂ，３１ｃ，４１ａ，４１ｂ，４２ｂ）によってネットワークに接続されている。これらのネットワークインターフェース（３０ａ，３０ｂ，３０ｃ，　３０ｄ，３０ｅ，３１ａ，３１ｂ，３１ｃ，４０ｂ，４１ａ，４１ｂ）は物理アドレスおよびネットワークアドレスによっても識別される。

【０００５】

種々のネットワークにおいて、ＡＲＰプロトコル（インターネット　エンジニヤリング　タスクフォース　ＲＦＣ８２６参照）が、物理アドレスおよびネットワークアドレスを相関させるのに利用される。物理アドレスおよびネットワークアドレスは、以後、図１におけるインターフェースの数字，ダッシュ、そしてそれぞれ、アドレスが物理アドレスあるいはネットワークアドレスであることを示す添字“Ｐ”あるいは“Ｎ”の表記法によって与えられる。

【０００６】

別々のネットワークにおける２つの所定のデータ処理装置が通信するためには、１つ以上のルータを利用して、一方のネットワークから他方へ通信を転送する。通常、通信の目標であるネットワーク（目標ネットワーク）と、このネットワークにパケットを転送するルータとの間の相関はデータ処理装置のルーティング表で示される（上記ＴＣＰ／ＩＰによるネットワーキング参照）。このルート表示方法は以後、“明示ルーティング表セットアップ”と称する。図１の３２ａ，３２ｂ，３２ｃ，３２ｄ，３２ｅはデータ処理装置３ａ，３ｂ，３ｃ，３ｄおよび３ｅのルーティング表を示す。図２に例示した明示ルーティングセットアップの例は、図１のネットワーク２に接続したデータ処理装置（３ａ，３ｂ，３ｃ，３ｄ，３ｅ）のネットワーク１へのルートに対して与えられている。各ルーティング表（３２ａ，３２ｂ，３２ｃ，３２ｄ，３２ｅ）は目標ネットワークを示すエントリ（３２１ａ，３２１ｂ，３２１ｃ，３２１ｄ，３２１ｅ），目標ネットワークの次のホップであるアドレスを示すエントリ（３２２ａ，３２２ｂ，　３２２ｃ，３２２ｄ，３２２ｅ）、および次のホップであるアドレスがデータ処理装置に直接接続したネットワークインターフェース（“インターフェース”）か、あるいはルータのアドレス（“ゲートウエイ”）かを表わすフラグ（３２３ａ

，３２３ｂ，３２３ｃ，３２３ｄ，３２３ｅ）を持っている。
【０００７】
データ処理装置３ａおよび３ｂのルーティング表は、ネットワーク１に対する各自のインターフェースのネットワークアドレス（３１ａ，３１ｂ）を次のホップ（３２２ａ，３２２ｂ）として、そして“インターフェース”をフラグ（３２３ａ，３２３ｂ）として表わす。データ処理装置３ｃ，３ｄおよび３ｅのルーティング表は、次のホップ（３２２ｃ，３２２ｄ，３２２ｅ）として３０ａ－Ｎをそしてフラグ（３２３ｃ，３２３ｄ，３２３ｅ）として“ゲートウエイ”を与えることによって、データ処理装置３ａがネットワーク１へのルータであることを表わしている。
【０００８】
その他の方式を利用して２つ以上のネットワーク間の相互接続を行うことができる。それらのうちの２つ、すなわちプロクシＡＲＰおよびＯＳＰＦプロトコルは以下で説明する。
【０００９】
プロクシＡＲＰ（ＲＦＣ１０２７参照）では、１つ以上のルータが代理として働き、他のネットワークにおけるネットワークアドレスに対するＡＲＰ要求に応答し、次いでこれらのアドレスに対してアドレスされた通信を受信し、そしてそれらをその宛先にルーティングし、従って２つ以上のネットワークを透過的にブリッジする（上記ＴＣＰ／ＩＰによるインターネットワーキング参照）。このために、物理アドレスおよびネットワークアドレス間の相関のセットアップが必要である。
【００１０】
図３はそのようなセットアップの例を示しており、ここでは、ネットワークアドレス（７１１）と物理アドレス（７１２）間の相関は、ＡＲＰキャッシュ　（７１）の特殊エントリとしてセットされる。この例では、ルータ３ａはネットワーク１からネットワーク２へ流れる通信の代理として動作する。“公用”フラグ（７１３）は、そのエントリはＡＲＰ問合わせに応答するために利用されるべきであることを示す。この例では、ネットワーク１におけるネットワークアドレス３０ｃ－Ｎ，３０ｄ－Ｎおよび３０ｅ－Ｎに対するいずれのＡＲＰ問合わせも、３１ａ－Ｐによって応答されるであろう。一例として、図１はデータ処理装置３ａのＡＲＰキャッシュで行われたこのセットアップを示すが、このセットアップは、ネットワーク１に接続したデータ処理装置（３ａ，３ｂ，４ａ，４ｂ）のいずれのＡＲＰキャッシュにおいても行うことができる。
【００１１】
上で行われたようなプロクシＡＲＰセットアップおよび明示ルーティング表セットアップは共に、各自のデータ処理装置の管理者によって行われる。従って、これらのセットアップは静的である、すなわち一旦行われると、それらは同じままであり、管理者による手動介入によってのみ変化する。従って、新規ルータがインストールされるやルータに誤動作がある場合、これらのセットアップは手動で変更されなければならない。
【００１２】
ネットワークの変化に応じてルーティング表を動的に変更する方法は、ＯＳＰＦプロトコル（ＲＦＣ１２４５，１２４６および１２４７参照）によって与えられ、この場合、ルータはそれら自体の間でルーティング情報を交換し、そしてこの情報に従ってそれらのルーティング表を変更する。
【００１３】
ＯＳＰＦの基本アルゴリズムは図４で示される。データ処理装置は、到達しうるネットワークおよびホップの数によって与えられたこれらのネットワークへの距離を含むメッセージを同報通信し（８２２）、そしてまた、他のデータ処理装置からこれらのメッセージを受信する（８２３）。ルートが変化した場合には（８２４）、各ルータはそれ自体からこれらのネットワークの各々への最短径路を計算し（８２５）、そしてこれらの径路に応じてそのルーティング表をセットする（８２６）。
【００１４】
図１のネットワークにおけるＯＳＰＦの利用例は図５で与えられている。この図では、全

データ処理装置（３ａ，３ｂ，３ｃ，３ｄ，３ｅ，４ａ，４ｂ）は ＯＳＰＦプロトコルを利用する交換データであり、従ってすべては、ＯＳＰＦ基本アルゴリズムを実行する制御アドオン（９１）を持っている。あるいはまた、ＯＳＰＦはネットワーク（１，２，５）のサブセットかデータ処理装置（３ａ，３ｂ，３ｃ，３ｄ，３ｅ，４ａ，４ｂ）のサブセットにおいてのみ、或いは両方のサブセットにおいて利用することができる。

【００１５】

相互接続ネットワークの特定の場合は並列処理装置を含むだけでなく、大規模並列処理装置およびワークステーションクラスタを含む。これらの並列処理装置は、ネットワークによって相互接続する複数のノードを含む。そのようなマシンの例は富士通社のＡＰ３０００，ＩＢＭ社のＲＳ／６０００ ＳＰおよびディジタルＣｏｒｐ．社のＴｒｕ－クラスタである。並列処理装置の場合は図１に示されている。この図では、第二ネットワーク２に接続したデータ処理装置は並列処理装置６のノードである。この構成では、並列処理装置は他のネットワークに対して複数のインターフェースを持って、信頼性およびネットワーキング性能を改善する。他のネットワークの処理装置（４ａ，４ｂ）は主に、並列処理装置６によって提供されるサービスにアクセスするクライアントである。

【００１６】

【発明が解決しようとする課題】

上記例では、ネットワーク１とデータ処理装置３ｃ，３ｄおよび３ｅとの間の全通信はデータ処理装置３ａのみを通過する。従って、３ａはこの通信の潜在的ボトルネックになる。このボトルネックは、ネットワーク２のデータ処理装置の一部をルータ３ａに、そして他のものをルータ３ｂに割り当てることによって、トラヒックは２つのルータを通過するので軽減させることができる。しかし、ネットワークトラヒックは時間と共に変化し、そしてある時間に多くのネットワークトラヒックが、ルータの１つに割り当てられたデータ処理にあるいはそれから行き来する場合、ボトルネックは存続する。従って、複数のルータ間でのトラヒックの均衡が有効であるためには、静的ではなく、動的であることが必要である。明示ルーティング表セットアップおよびプロクシＡＲＰセットアップは静的であり、従って静的トラヒック均衡を与えることができるが、動的平衡を与えることはできない。ＯＳＰＦプロトコル定義（ＲＦＣ１２４７）は、“宛先への幾つかの等価ルートがある場合、トラヒックはそれらの間で等しく分配される”、と明言しているが、トラヒック分配方法は述べておらず、これをルーティングアルゴリズムの実行にゆだねている。従って、ＯＳＰＦでは、ルートは動的に変化するにもかかわらず、ＯＳＰＦそれ自体では動的トラヒック平衡を与えることができない。

【００１７】

パーソナル コンピュータおよびワークステーションを備える、ユーザプログラムを実行するデータ処理装置は、ルータとしてしばしば利用される。ルーティングによって処理負荷を生じるので、そのようなデータ処理装置の処理性能は、ルーティングによって生じた処理負荷によって低減される。従って、ネットワーキング性能や処理能力を改善するために、ネットワーキングトラヒックの平衡を保つようルートを選択する場合、処理負荷を追加して考慮に入れなければならない。ネットワークトラヒックの場合におけるように、処理負荷は時間と共に変化し、従ってルートは処理負荷に応じて動的に変化すべきである。前述したように、明示ルーティング表セットアップおよびプロクシＡＲＰセットアップは静的であり、従ってその種の動的トラヒック平衡を与えることができない。ＯＳＰＦプロトコルは、目標ネットワークへの最短路を提供するルータを選択するが、この選択ではルータの処理負荷を考慮していない。従ってＯＳＰＦはルータの処理負荷に応じた動的トラヒック平衡を与えることができない。

【００１８】

信頼性を提供するために、ネットワークの故障の際にはルーティングが変更されることが望ましい。そのような特徴はＯＳＰＦによって提供され、それはプロクシＡＲＰにとっても望ましいのであるが、プロクシＡＲＰセットアップは静的であるので、その種の動的トラヒック変化を与えることができない。

【００１９】
【課題を解決するための手段】
２つのネットワーク間の複数のルータにおけるトラヒック平衡の問題は、両方のネットワークの各データ処理装置に対して、他のネットワークへの通信やこのデータ処理装置からの通信に利用しようとするルータを、これらのデータ処理装置に関連するネットワーキングトラヒックに応じて、選択することによって解決される。次いで、データ処理装置のルーティング表，物理アドレスとネットワークアドレス間のプロクシＡＲＰ連関、およびＯＳＰＦプロトコルで利用しようとするルートが、上述の選択によってセットされる。この手順を周期的に繰り返して、動的トラヒック平衡を与える。
【００２０】
トラヒック平衡において処理負荷を考慮するという問題は、適切であればルータの処理負荷を取得し、この処理負荷を等価量のネットワークトラヒックに換算し、そしてこのトラヒックを上述のルータを選択する手順に利用することによって、解決される。
【００２１】
プロクシＡＲＰの信頼性問題は、上述のトラヒック平衡手順を実行するのであるが、何らかの誤動作のためにルータとしてのその動作が阻止されているルータには、ネットワーキングトラヒックを分配しないようにすることによって解決される。さらに、ルータ故障あるいはルータ回復が検出される場合に、ルートもまた選択されるべきである。
【００２２】
【発明の実施の形態】
トラヒック平衡の全プロセスは、下記の順の三段階、すなわちルータを選択する段階と、これらのルータ間のトラヒックの分配を計算する段階と、そしてこの分配に応じてルートを変更する段階を含む。これらの三段階を順を追って説明し、そして第三段階の説明において幾つかの例を挙げる。以下の説明のすべては、以後“当該ネットワーク”と称する単一ネットワークに関する。その他のネットワークは以後、“目標ネットワーク”と称される。明示ルーティング表セットアップおよびＯＳＰＦは、当該ネットワークから目標ネットワークへ行く通信のためのルートをセットするのに利用され、そしてプロクシＡＲＰは目標ネットワークから当該ネットワークへ行く通信のためのルートをセットするのに利用される。従って、第二ネットワーク２に接続したデータ処理装置が並列処理装置６のノードである場合、明示ルーティング表セットアップおよびＯＳＰＦは、並列処理装置６からクライアント４ａ，４ｂへの通信のためのルートをセットするのに利用され、そしてプロクシＡＲＰは反対方向でのルートをセットするのに利用される。
【００２３】
信頼性を与えるためにルータの選択が行われる。これは、所定の目標ネットワークへのすべての可能なルータの中で、誤動作なく働いているルータを選択することによって行われる。このプロセスはＯＳＰＦで行われる。明示ルーティング表セットアップおよびプロクシＡＲＰの場合、下記の手順が利用される。先ず、ネットワークの管理者は当該ネットワークに対する大域ルーティング表を準備しなければならない。図６に示される大域ルーティング表の例７３は、当該ネットワークのネットワーク２に対してのものである。各目標ネットワーク（７３１）に対して、この表はこのネットワークに対するルータ（７３２）を示し、そして各ルータに対して、目標ネットワークとのインターフェースの物理アドレス　（７３３）、当該ネットワークにおけるネットワークアドレス（７３４）、このルータ後の目標ネットワークへの次のホップ（７３５）およびルートに対する優先度（７３６）を表わす。この優先度はネットワーク管理者によって、目標ネットワークまでのホップ数に応じて、そしてまた、ネットワーク処理能力に応じて、セットされ、従って最良のルートが優先的に選択されるであろう。
【００２４】
大域ルーティング表７３は、ルータを選択する図７に示されるアルゴリズムのための入力として利用される。第一段階（８１２）として、アルゴリズムは、メモリの大域ルーティング表（７３）への全エントリを読取る。次の段階（８１３）では、全ルータ（７３２）

の状態、すなわち、ルータ（７３２）はアクティブであるかどうか、そして表示されたネットワークインターフェース（７３３，７３４，７３５）はアクティブであるかどうかを、検証する。次の段階（８１４）では、アルゴリズムはメモリの表から、ルータあるいはインターフェースがアクティブでない全ルータエントリ（７３２，７３３，７３４，７３５，７３６）を除去する。次の段階（８１５）では、アルゴリズムは各ネットワーク（７３１）に対して、優先性（７３５）に対する最低値を持つルータエントリ（７３２，７３３，７３４，７３５，７３６）を選択し、他のルータエントリ（７３２，７３３，７３４，７３５，７３６）をメモリから削除する。この段階で、各ネットワーク（７３１）に対して、目標ネットワーク（７３１）に対するルータとして動作することになっているデータ処理装置が、欄７３２に現われる。

【００２５】

一旦、ルータが選択されると、ルートを選択することができる。最初に、ネットワークの管理者は当該ネットワークに対して、トラヒック平衡に関連するデータ処理装置のリストを準備しなければならない。当該ネットワークとしてのネットワーク２のためのこのリストの例は図８で与えられる。データ処理装置７２のリストは、トラヒック平衡が適用される全データ処理装置の識別子（７２１）および、この装置における処理とルーティング負荷を相関させる係数（７２２）を含んでいる。この係数は以下でより詳細に説明する。

【００２６】

トラヒック平衡を行うルートの計算（８４）は図９に示される。ｉをルータとして動作しないデータ処理装置に対するインデックスとし、ｊを目標ネットワーク（７３１）に対するインデックスとし、そしてｋをルータ（７３２）として動作するデータ処理装置に対するインデックスとする。また、トラヒック（ｉ，ｊ）をデータ処理装置ｉとネットワークｊ間のトラヒックを持つものとし、そして負荷（ｋ）をゼロに等しいか大きい数とする、この数は、このデータ処理装置がユーザプログラムを実行できる場合、データ処理装置の処理負荷ｋと共に増加し、あるいはそれができない場合は、ゼロである。アルゴリズムはまた、ルータｋに割り当てられたトラヒック（ｉ，ｊ）の量を表わすルート（ｋ）の値を利用する。アルゴリズムは基本的には、別々のルータｋにおけるルート（ｋ）の値の平衡を保つような方法で、ルート（ｋ）の間でトラヒック（ｉ，ｊ）を分配する。

【００２７】

ルートを選択するアルゴリズムは図９で与えられる。最初の段階（８４２）では、アルゴリズムは他のデータ処理装置から全トラヒック（ｉ，ｊ）および負荷（ｋ）を取得する。次の段階（８４３）では、アルゴリズムはルート（ｋ）の値をａ（ｋ）×負荷（ｋ）として初期化する、ただし、ａ（ｋ）（７２２）はデータ処理装置表（７２）から取られている。従って、ルータの処理負荷はネットワーキング負荷として計算に組入れられる。ａ（ｋ）定数は、ルータにおけるネットワークトラヒック平衡と処理負荷平衡間にトレードオフを与えている：ａ（ｋ）の低い方の全体値はネットワークトラヒック平衡に好都合であろうし、そして高い方の全体値は処理負荷平衡に好都合であろう。また、ａ（ｋ）の相対値はデータ処理装置の性能を表わし、ａ（ｋ）の低い方の値は高性能を意味する（すなわち、ネットワークトラヒックの同量をルーティングすることによって、より少ない処理負荷を生じる）。

【００２８】

次の段階（８４４）では、全ｋに対して、アルゴリズムは、ルータ（ｋ）＝ルータ（ｋ）＋トラヒック（ｉ，ｊ）を計算することによって、このルータによってのみ流れることができる全トラヒックを、ルータｋに割り当てる。この段階で利用されたトラヒック（ｉ，ｊ）の全値はゼロにされる。次の段階（８４５）では、アルゴリズムは、トラヒック（ｉ，ｊ）が残りのトラヒック（ｉ，ｊ）の最高値になるようにｉとｊを取得する。次の段階（８４６）では、アルゴリズムは、ネットワークｊのためのルータの中で、ルート（ｋ）の最小値を持つｋを見つける。この選択は、ネットワークｊとデータ処理装置ｉ間のネットワークトラヒックがルータｋを通過することになっていることを意味する。次の段階（８４７）では、トラヒックはこのルータに割り当てられる、すなわち、選択されたｉとｊ

の値に対してルータ（ｋ）＝ルータ（ｋ）＋トラヒック（ｉ，ｊ）。また、トラヒック（ｉ，ｊ）はゼロにされる。最終段階（８４８）では、アルゴリズムは、トラヒック（ｉ，ｊ）の非ゼロ値がまだあれば、段階（８４５）に戻る。

【００２９】

一旦、データ処理装置とその各自のルータ間の対応が得られれば、それに応じてそれらのルートがセットされる。ルートをセットするプロセスは、ＯＳＰＦが利用されているか、いないかによって異なる。ＯＳＰＦが利用されていない場合、トラヒック平衡の全プロセス（すなわち、ルータを選択し、これらのルータにおけるトラヒックの分配を計算し、そしてその分配に応じてルートを変更する３つの段階）は図１０に示される。ルータの選択（８１）およびトラヒック分配　（８４）の後、明示ルーティングセットアップ（８５２）およびプロクシＡＲＰセットアップが行われ、そしてプロセスは周期的に繰り返されて、動的負荷平衡を与える。

【００３０】

明示ルートセットアップ（８５２）のためのルートのセットアップは次の通りである。ルータとして利用しようとするデータ処理装置においてセットしようとする目標ネットワーク（７３１）のためのルートは、大域ルーティング表（７３）の次のホップの欄（７３５）から取られ、この列でそれはこのルータに対応する。この場合、フラグは“インターフェース”にセットされる。他のデータ処理装置のためのルートおよび、ルータとして動作できるが目標ネットワークに対するそれらのネットワークインターフェースが不動作であるデータ処理装置もまた、ネットワークの欄（７３４）から取られ、この列ではそれは、段階８４７でこのデータ処理装置に割り当てられたルータに対応する。この場合、フラグは“ゲートウエイ”にセットされる。不動作のデータ処理装置における、あるいは当該ネットワークに対するそのネットワークインターフェースがアクティブでないルートはセットされない。

【００３１】

プロクシＡＲＰのためのルートのセットアップ（８５３）は次の通りである。データ処理装置のインターフェースのネットワークアドレスは、このデータ処理装置に対して選択されたルータのインターフェースの物理アドレスに関連する。物理アドレスは大域ルーティング表（７３）の物理アドレスの欄（７３３）から取られる。プロクシＡＲＰ連関を持つデータ処理装置の故障の場合には、これらの連関は、目標ネットワーク（７３１）に接続した別のデータ処理装置にセットされるべきである。プロクシＡＲＰセットアップは、目標ネットワーク（７３１）が当該ネットワークに隣接している場合にのみ行われる。また、大域ルーティング表（７３）の物理アドレスエントリ（７３３）がブランクである場合には、プロクシＡＲＰを利用する相互接続は不可能であり、従って、プロクシＡＲＰセットアップは行われない。ＡＲＰ連関が変化する場合、それらを持つデータ処理装置は新規連関を持つＡＲＰ応答を同報通信することができる。この場合、目標ネットワーク（７３１）の他のデータ処理装置はこの応答を受信し、応答の連関をそれら自体のＡＲＰキャッシュの内容と比較し、そして変化が認められれば、それらのＡＲＰキャッシュの連関を応答内のそれに変更し、このようにして、トラヒックを新規物理アドレスに送信する。従って連関が変化した場合にＡＲＰ応答が同報通信されれば、新規連関はデータ処理装置に反映されて、ルータ間のそれらのネットワーキングトラヒックの動的トラヒック平衡を実現する。ＡＲＰ応答が同報通信されない場合、後に目標ネットワーク（７３１）にＡＲＰ問合わせを送信することで新規連関はデータ処理装置に反映される。

【００３２】

図１０のアルゴリズムは、ユーザプログラムを実行するいずれのデータ処理装置（３，４）においても制御アドオン走行によって、実行することができる。

【００３３】

図１１は本発明の好適な実施例を示しており、制御アドオン（９２）は、ＡＲＰキャッシュのセットアップを容易にするために、ＡＲＰキャッシュ（７１）を設けてあるのと同じデータ処理装置内に配置してある。あるいはまた、この制御アドオン（９２）は、データ

処理装置（３ａ，３ｂ，３ｃ，３ｄ，３ｅ，４ａ，４ｂ）のいずれにも位置することができる。並列処理装置（６）の場合には、制御アドオン（９２）のための好適な実施例は、並列処理装置のノード（３ａ，３ｂ，　３ｃ，３ｄ，３ｅ）の１つにあるが、それは並列処理装置（６）のオペレーティングシステムの特定特徴が、他のノードからデータを収集し、そして他のノードにおけるセットアップを実行するために、調査されることができるからである。このように、この制御アドオン（９２）はデータ処理装置（３ａ，３ｂ，３ｃ，３ｄ，３ｅ，４ａ，４ｂ）のいずれにも位置することができる。トラヒック（ｉ，ｊ）と負荷（ｋ）の値は、制御アドオン（９２）がシステムコールによって位置決めされているデータ処理装置から取得することができる。他のデータ処理装置のネットワーキングトラヒック，トラヒック（ｉ，ｊ）は、例えば、単一ネットワーク管理プロトコル（ＳＮＭＰ，ＲＦＣ１１５７参照）によって取得することができるし、そして他のデータ処理装置の処理負荷は、例えば、ｒｗｈｏプロトコルあるいはインターネット　シスタート　サービスによって取得することができる。ルートおよびプロクシＡＲＰ連関は、制御アドオン（９２）が走行しているデータ処理装置においては、システムコールによって、そしてその他のデータ処理装置においてはＳＮＭＰプロトコルを利用することによって、変更することができる。

【００３４】

図１および図６の例に適用された図１０のアルゴリズムの例は次の通りである。データ処理装置３ｃ，３ｄおよび３ｅのネットワーク１に関連するネットワーキングトラヒックは、その各々に対して１Ｍバイト／秒であり、そしてトラヒック（ｉ，ｊ）はＭバイト／秒で測定したネットワーキングトラヒックであると仮定する。また、データ処理装置３ａ，３ｂ，３ｃおよびそれらのインターフェースは誤動作なく動作し、それらはユーザプロセスを実行することができ、データ処理装置３ａと３ｃはユーザプロセスを実行せず、データ処理装置３ｂが１つのユーザプロセスを実行し、そして負荷（ｋ）はユーザプロセスの数によって測定されるものと仮定する。

【００３５】

ルータを選択するアルゴリズム（８１）は最初（８１２）に、大域ルーティング表（７３）を読取り、そして次の段階（８１３）で、データ処理装置３ａ，　３ｂおよび３ｃとそれらのインターフェースが誤動作なく動作していることを検証する。これらのデータ処理装置（３ａ，３ｂおよび３ｃ）のエントリは、次の段階（８１４）において表から除去されず、データ処理装置３ａと３ｂは、それらが優先性に対して最低値を持つので選択される（８１５）。従って、データ処理装置３ａと３ｂはネットワーク１と２間のルータとして選択される。

【００３６】

次いで、トラヒックの平衡を保つアルゴリズム（８４）では、最初に（８４２）、データ処理装置３ａと３ｂの処理負荷および、ネットワーク１とデータ処理装置３ｃ，３ｄおよび３ｅの各々との間のネットワーキングトラヒックが取得される。段階８４３では、ルート（３ａ）は０を割り当てられ、そしてルート（３ｂ）は１．５　を割り当てられる。段階８４４では、トラヒックはネットワーク１には割り当てられない。

【００３７】

次いで、アルゴリズムは段階８４５，８４６および８４７までを３回繰り返して、トラヒックをルータに割り当てる。データ処理装置３ｃ，３ｄおよび３ｅが段階８４５において、この順で選択されたと仮定する。従って、最初の繰り返しでは、データ処理装置３ｃがルータ３ａに割り当てられ、そしてルート（３ａ）は１に更新される。２回目の繰り返しでは、データ処理装置３ｄもまた、ルータ３ａに割り当てられ、そしてルート（３ａ）は２に更新される。３回目の繰り返しでは、データ処理装置３ｅは、ルータ３ｂに割り当てられ、ルート（３ｂ）は２．５　に更新される。従って、データ処理装置３ｃ，３ｄおよび３ｅのネットワーキングトラヒックは、データ処理装置３ａと３ｂ間で分配され、３ｂはより高い処理負荷を持っているので、より多いトラヒックが３ａに与えられている。データ処理装置３ｅへ行き来するルートは、このようにしてルータ３ｂに変更される。

【００３８】
明示ルーティング表セットアップ（８５２）が行われた後のデータ処理装置　３ｅのルーティング表は、図１２に示されている。プロクシＡＲＰセットアップ（８５３）が行われた後のＡＲＰキャッシュは、図１３に示されている。点線は変化したエントリを示す。
【００３９】
信頼性を説明する２つの例を以下に述べる。
【００４０】
第一の例は、ネットワーク２におけるデータ処理装置３ａ（３１ａ）のネットワークインターフェースは不調だが、動作はし続けている場合のためのものである。ルータを選択するアルゴリズム（８１）において、データ処理装置３ａの故障は段階８１３で検出され、そしてこのルータに対するエントリは段階８１４において削除される。データ処理装置３ｂに対するエントリは、それが優先性の最低値を持つので、段階８１５で選択される。従って、データ処理装置３ｂのみがルータとして利用される。次いで、トラヒック分配アルゴリズム（８４）において、データ処理装置３ａ，３ｃ，３ｄおよび３ｅのトラヒックはルータ３ｂに割り当てられる。明示ルーティング表セットアップ（８５２）が行われた後のデータ処理装置３ａ，３ｃ，３ｄおよび３ｅのルーティング表（３２ａ，３２ｃ，　３２ｄおよび３２ｅ）は図１４に示される。プロクシＡＲＰセットアップ（８５３）が行われた後のＡＲＰキャッシュ（７１）は図１５に示される。点線は変化したエントリを示す。この例では、第二ネットワークのデータ処理装置（３ａ，３ｂ，３ｃ，３ｄ，３ｅ）間の通信は、ネットワークインターフェースの故障によって影響を受けない。同様に、並列処理装置（６）の場合、それとクライアント　（４ａ，４ｂ）間の通信は、ルートが変化した後も続く。
【００４１】
信頼性に関するもう１つの例は、両データ処理装置３ａと３ｂが全体的に故障する場合である。この場合、ルータを選択するアルゴリズム（８１）において、データ処理装置３ａと３ｂにおける故障は、段階８１３において検出され、そしてこれらのルータに対するエントリは段階８１４において削除される。従って、データ処理装置３ｃのエントリのみが残り、従ってデータ処理装置３ｃがルータであると選択される。次いで、トラヒック分配アルゴリズム（８４）では、データ処理装置３ｄと３ｅのトラヒックがルータ３ｃに割り当てられる。明示ルーティング表セットアップ（８５２）が行われた後のデータ処理装置３ｃ，３ｄおよび３ｅのルーティング表（３２ｃ，３２ｄおよび３２ｅ）は図１６に示される。ＡＲＰキャッシュ（７１）内の全エントリは削除されたが、それは、ルータ３ｃに対応する物理アドレス（７３３）がセットされていないからである。この例では、当該ネットワークにおいてなお動作可能なデータ処理装置（３ｃ，３ｄ，　３ｅ）と目標ネットワークのデータ処理装置（４ａ，４ｂ）の間の通信は中断されるが、それはネットワーク２から１へのルートのみがセットされ、そしてネットワーク１から２への通信のためにルートがセットされていないからである。しかし、同じ方式がネットワーク１にも適用される場合には、ネットワーク１から２へのトラヒックは、データ処理装置４ｂを通って流れることができるようになり、従って、ネットワーク１と２間の通信を中断することはない。同じことが、並列処理装置（６）の場合にも適用される。
【００４２】
ＯＳＰＦが当該ネットワークで利用されている場合には、２つの主な相違がある。第一に、ルータはＯＳＰＦ処理によって選択される、従ってＯＳＰＦがルータを選択するアルゴリズム（８１）の代わりに利用される。第二に、ＯＳＰＦによるルータ交換ルートは分配するように動作するが、ルータ選択アルゴリズム　（８１）とトラヒック分配アルゴリズム（８４）は、種々のデータ処理装置によって選択されたルート間の不一致を回避し、そして、全データ処理装置がお互いの処理負荷およびネットワーキングトラヒックを問合わせる場合に起こるかもしれない処理負荷を回避するために、集中化するように実行されるべきである。
【００４３】

従って、ＯＳＰＦは次のように拡張されている。ＯＳＰＦによってデータを交換するデータ処理装置の１つ（以後、“マスター”と称する）は、ルートを選択するような集中化タスク、およびＯＳＰＦを実行する他のデータ処理装置にこれらのルートを同報通信する責任がある。これらの他のデータ処理装置（以後、“スレーブ”と称する）はそれに応じてそれらのルートをセットする。マスターおよびスレーブによって行われる手順のより詳細な説明は、以下の通りである。

【００４４】

マスターの基本アルゴリズムは図１７に示される。ＯＳＰＦ基本アルゴリズムは、追加段階（８３１）によって拡張され、周期的に実行されることになっている。この段階の詳細は図１８に示される。先ず、トラヒック平衡を計算するアルゴリズム（８４）が実行される。次いで、マスターデータ処理装置におけるルートがセットされる（８３１１）。次いで、選択されたルートはスレーブデータ処理装置に同報通信される（８３１２）。次いで、ＯＳＰＦによってデータを交換しないデータ処理装置におけるルートがセットされる。最後に、プロクシＡＲＰがＯＳＰＦと共に利用されている場合には、このプロクシＡＲＰ対応がセットされる（８５３）。スレーブの基本アルゴリズムは図１９に示される。ここでは、基本ＯＳＰＦアルゴリズムは、２つの追加段階（８３２と８３３）によって拡張されて、やはり周期的に実行されることになっている。これらの段階の第一において、スレーブデータ処理装置はマスターによって同報通信されたルートを受信し、そして次の段階（８３３）において、それに応じて、それ自体のルートをセットする。ＯＳＰＦプロトコルとは独立して、データの特別交換が行われることに注目すべきであり、従ってＯＳＰＦプロトコルの変更は必要でない。マスターおよびスレーブにおいて、ルータの状態の変化によって利用しようとするルータが変化する場合、トラヒック平衡計算の結果が計算され、そして同報通信されるまで、マスターおよびスレーブは利用できるものの中で任意にルートを選択すべきであり、そうでなければ、まだ有効であれば、前のルートを保持すべきである。このルートは、次いで、トラヒック分配アルゴリズム（８４）の結果に従って、変化する。

【００４５】

マスター（９３）およびスレーブ（９４）の制御アドオンの２つの好適な実施例を図２０と図２１に示す。図２０の実施例において、ＯＳＰＦは全ネットワーク（１，２，５）において利用されている。図２０では、マスター（９３）の制御アドオンがデータ処理装置３ａに位置決めされている；あるいはまた、それはＯＳＰＦによってデータを交換する他のいずれのデータ処理装置（３ａ，３ｂ，３ｃ，３ｄ，３ｅ，４ａ，４ｂ）にも位置決めすることができる。図２１の実施例では、ＯＳＰＦは１つ以上のネットワーク（図ではネットワーク２）において利用されており、そしてプロクシＡＲＰは、ＯＳＰＦが利用されているそれに隣接する１つ以上のネットワーク（図ではネットワーク１）から来るトラヒックのためのルートをセットするために利用されている。マスター（９３）の制御アドオンは、ＡＲＰキャッシュ（７１）のセットアップを容易にするために、ＡＲＰキャッシュ（７１）の同じデータ処理装置に位置決めされている。あるいはまた、マスターは、ＯＳＰＦが利用されているいずれのネットワークにでも位置することができる、そしてＡＲＰキャッシュ（７１）は目標ネットワークのいずれのデータ処理装置（３ａ，３ｂ，４ａ，４ｂ）にでも配置することができる。

【００４６】

明示ルーティング表セットアップおよびプロクシＡＲＰに対して与えられた３つの例は、最短路の選択（８２５）によってルータを選択するアルゴリズム　（８１）と同様に、同じルートがルート選択アルゴリズム（８４）によって選択されるであろう、そして当該ネットワークにおけるデータ処理装置のルーティング表にセットされたルート（３２ａ，３２ｂ，３２ｃ，３２ｄ，３２ｅ）およびＡＲＰキャッシュ（７１）にセットされた連関は同じになるであろう、という意味において、ＯＳＰＦの場合と同様に動作する。同じことはまた、並列処理装置（６）の場合にも適用される。

【００４７】

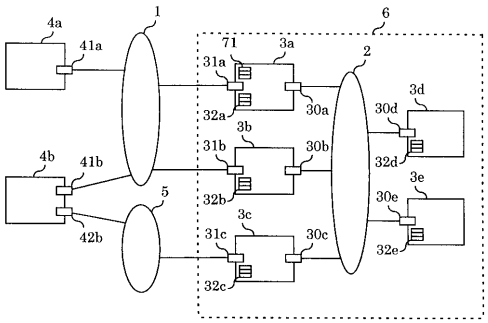
【発明の効果】
この発明は５つの利点をもたらす。第一に、この発明は、２つのネットワーク間の複数のルータの間でネットワークトラヒックを動的に平衡を保つことによって、良好なネットワーキング性能を提供する。第二に、この発明は、より高い処理負荷を持つデータ処理装置からルーティングトラヒックを取り、従ってデータ処理装置をより有効に利用し、かつジョブ処理能力を増加させている。第三に、この発明は、１つ以上のルータが故障した場合にプロクシＡＲＰのルーティングを透過的に変化させて、ネットワーキング信頼性を改善している。第四に、この発明は現存のネットワークプロトコルと互換性があり、従ってこれらのプロトコルを変更する必要はない。また、明示ルートセットアップおよびプロクシＡＲＰの場合に、データ処理装置のオペレーティングシステムに何の変化も必要としない。ＯＳＰＦプロトコルに利用する際に、等価複数通路の場合、ルートを選択するＯＳＰＦプロトコルを実行する制御アドオンの変化のみが必要とされる。第五に、並列処理装置の場合、この発明によれば、他のネットワークと効率よく通信するために、すべてのノードが外部ネットワークインターフェースを必要とするわけではないから並列処理装置のコストを低下させることができる。

【図面の簡単な説明】
【図１】本発明が実施されるネットワークの例を示すブロック図。
【図２】ルーティング表セットアップの説明図。
【図３】プロクシＡＲＰのためのＡＲＰキャッシュセットアップの説明図。
【図４】基本ＯＳＰＦアルゴリズムの説明図。
【図５】ＯＳＰＦのための制御アドオンの説明図。
【図６】大域ルーティング表の説明図。
【図７】ルータを選択するアルゴリズムの説明図。
【図８】処理負荷とネットワークトラヒックを相関させる表を示す図。
【図９】トラヒックを分配するアルゴリズムの説明図。
【図１０】ＯＳＰＦが利用されない場合にトラヒックの平衡を保つ手順の説明図。
【図１１】ＯＳＰＦが利用されない場合の実施例を示す図。
【図１２】トラヒックの平衡を保つためのルーティング表セットアップの説明図。
【図１３】トラヒックを分配するためのＡＲＰキャッシュセットアップの説明図。
【図１４】インターフェース故障の場合のルーティング表セットアップの説明図。
【図１５】インターフェース故障の場合のＡＲＰキャッシュセットアップの説明図。
【図１６】複数ルータ故障の場合のルーティング表セットアップの説明図。
【図１７】ＯＳＰＦマスター装置によって実行される手順の説明図。
【図１８】ＯＳＰＦマスター装置によって実行される手順の詳細の説明図。
【図１９】ＯＳＰＦスレーブ装置によって実行される手順の説明図。
【図２０】ＯＳＰＦのみが利用される場合の実施例の説明図。
【図２１】ＯＳＰＦがプロクシＡＲＰと共に利用される場合の実施例の説明図。

【符号の説明】
１，２，５…ネットワーク、３，４…データ処理装置、３０，３１，４１…ネットワークインターフェース、６…並列処理装置、３２…ルーティング表、７１…ＡＲＰキャッシュ、７２…処理負荷とネットワークトラヒックを相関させる表、７３…大域ルーティング表、９１…ＯＳＰＦの制御アドオン、９２…ＯＳＰＦなしの負荷平衡のための制御アドオン、９３…ＯＳＰＦのためのマスター制御アドオン、９４…ＯＳＰＦのためのスレーブ制御アドオン。

【図１】

図1



【図２】

図2

32a

Target Network	Next Hop	Flags
…	…	…
Network 1	31a-N	interface
…	…	…

321a 322b 323c

32b

Target Network	Next Hop	Flags
…	…	…
Network 1	31b-N	interface
…	…	…

321b 322b 323b

32c

Target Network	Next Hop	Flags
…	…	…
Network 1	30a-N	gateway
…	…	…

321c 322c 323c

32d

Target Network	Next Hop	Flags
…	…	…
Network 1	30a-N	gateway
…	…	…

321d 322d 323d

32e

Target Network	Next Hop	Flags
…	…	…
Network 1	30a-N	gateway
…	…	…

321e 322e 323e

【図３】

図3

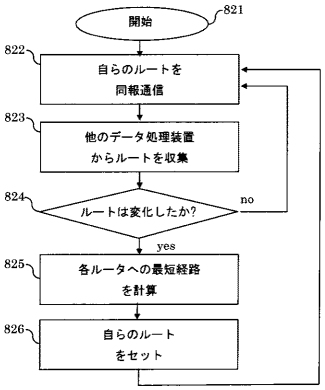
71

Network Address	Physical Address	Flags
…	…	…
30c-N	31a-P	public
30d-N	31a-P	public
30e-N	31a-P	public
…	…	…

711 712 713

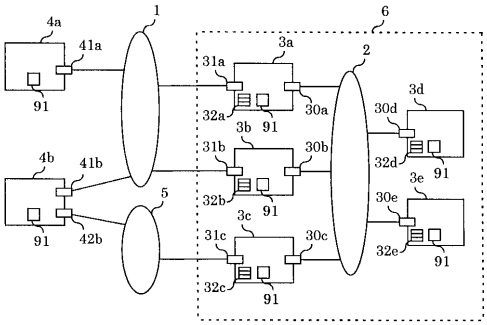
【図４】

図4



【図５】

図5



【図６】

図6

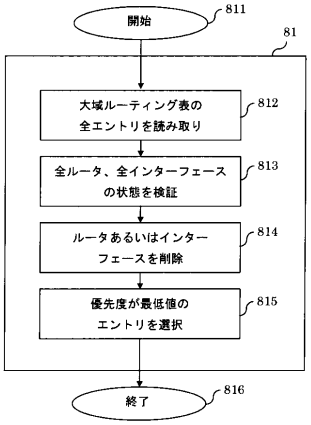
73

Target		Router		Next Hop	Priority
	ID	Physical	Network		
…	…	…	…	…	…
Network 1	3a	31a-P	30a-N	31a	1
	3b	31b-P	30b-N	31b	1
	3c	—	30c-N	42b	2
…	…	…	…	…	…

731 732 733 734 735 736

【図７】

図7



【図８】

図8

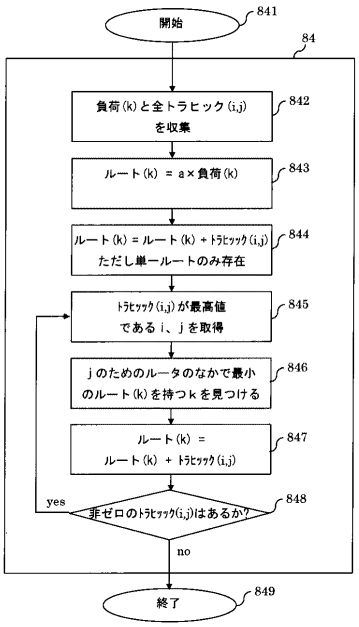
72

ID	a(k)
3a	1.5
3b	1.5
3c	1.5
3d	0
3e	0

721 722

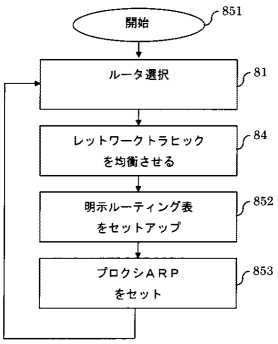
【図９】

図9



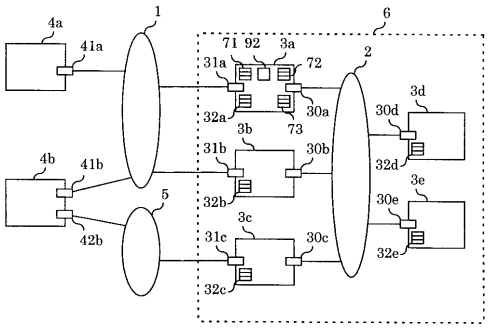
【図１０】

図１０



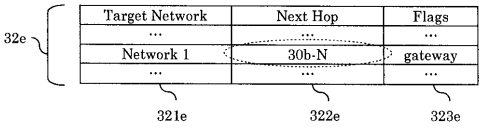
【図１１】

図１１



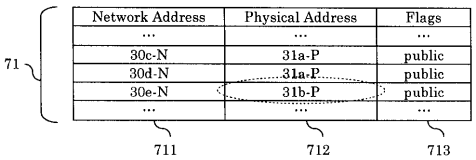
【図１２】

図１２



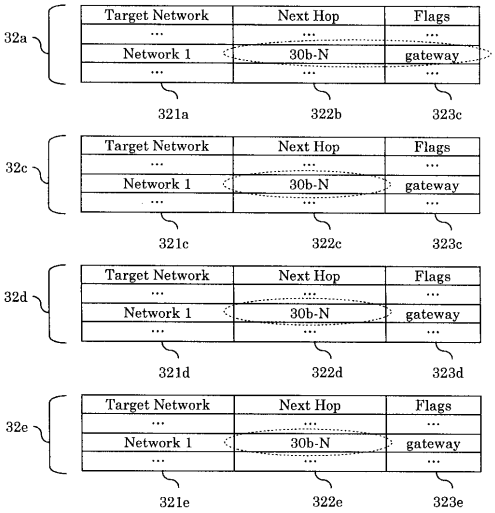
【図１３】

図１３



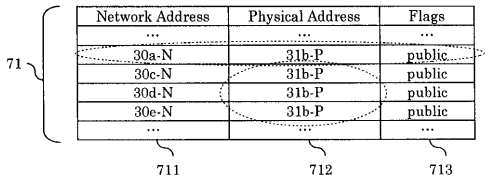
【図１４】

図１４



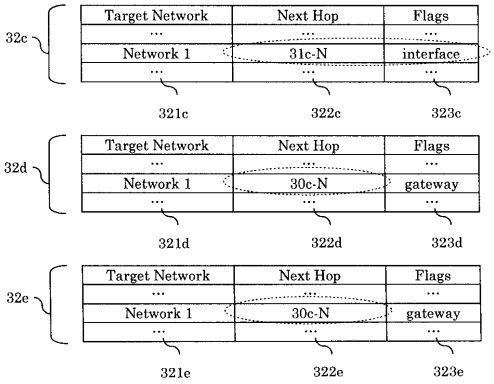
【図１５】

図１５



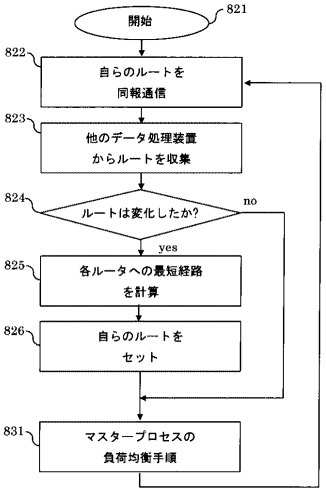
【図１６】

図１６



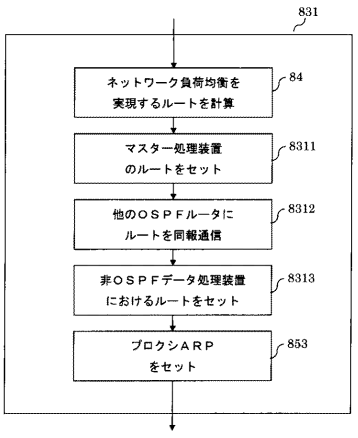
【図１７】

図１７



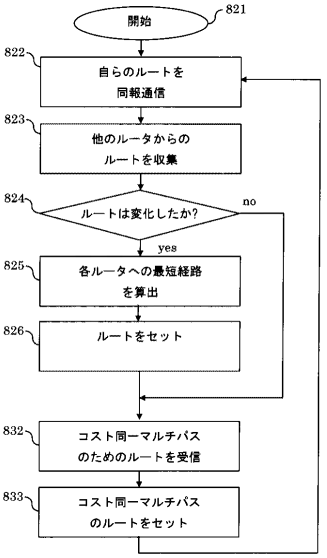
【図１８】

図１８



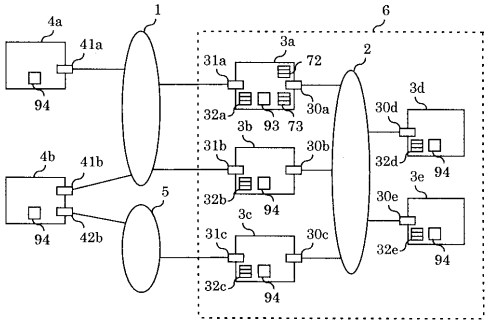
【図１９】

図１９



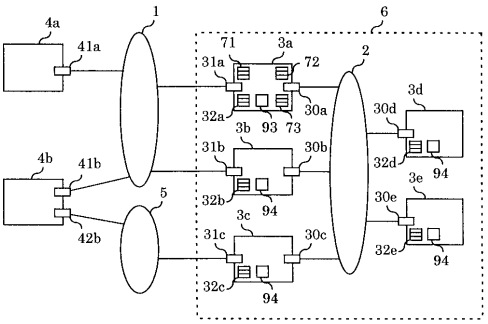
【図２０】

図２０



【図２１】

図２１



フロントページの続き

(72)発明者 樋口 達雄
東京都国分寺市東恋ケ窪一丁目２８０番地 株式会社 日立製作所 中央研究所内
(72)発明者 吉澤 聡
東京都国分寺市東恋ケ窪一丁目２８０番地 株式会社 日立製作所 中央研究所内
(72)発明者 村橋 英樹
東京都国分寺市東恋ケ窪一丁目２８０番地 株式会社 日立製作所 中央研究所内

審査官 石井 研一

(56)参考文献 特開平０３－０７０３３０（ＪＰ，Ａ）
特開平０５－１３０１４４（ＪＰ，Ａ）
特開平０９－２７５４０３（ＪＰ，Ａ）
特開平０７－２６４２３３（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.^{7}，ＤＢ名)
H04L 12/56