

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4205931号
(P4205931)

(45) 発行日 平成21年1月7日(2009.1.7)

(24) 登録日 平成20年10月24日(2008.10.24)

(51) Int.Cl. F I
H O 4 L 12/56 (2006.01) H O 4 L 12/56 1 O O Z

請求項の数 7 (全 17 頁)

(21) 出願番号	特願2002-333461 (P2002-333461)	(73) 特許権者	000005108
(22) 出願日	平成14年11月18日(2002.11.18)		株式会社日立製作所
(65) 公開番号	特開2003-218918 (P2003-218918A)		東京都千代田区丸の内一丁目6番6号
(43) 公開日	平成15年7月31日(2003.7.31)	(74) 代理人	110000350
審査請求日	平成17年11月16日(2005.11.16)		ポレール特許業務法人
(31) 優先権主張番号	10/052, 684	(74) 代理人	100068504
(32) 優先日	平成14年1月18日(2002.1.18)		弁理士 小川 勝男
(33) 優先権主張国	米国 (US)	(74) 代理人	100086656
			弁理士 田中 恭助
		(74) 代理人	100094352
			弁理士 佐々木 孝
		(72) 発明者	金武 達郎
			アメリカ合衆国 カリフォルニア州 94
			086 サニーベイル、ラニトス・アベニ
			ュー221

最終頁に続く

(54) 【発明の名称】 ラベル・スイッチ・ネットワークの仮想リンクを作成するためのシステムおよび装置

(57) 【特許請求の範囲】

【請求項1】

ラベル・スイッチ・ネットワークで多数のリンクを管理するシステムにおいて、
複数の入力仮想リンクと複数の出力仮想リンクとを含み、入力又は出力仮想リンク各々が複数の個別のリンクを有する複数の仮想リンクと、

各々のインGRESS・ノードがパケットを受信して付属のラベルを前記パケットにラベル付けするように構成された複数のインGRESS・ノードと、

各々のラベル・スイッチング・ノードが1つまたはそれ以上の入力仮想リンク経路で各々に関連したラベルを有する前記ラベル付けされたパケットを受信して前記受信したラベル付けされたパケットを1つまたはそれ以上の出力仮想リンク経路で各々に関連するラベルに基づいて転送するように構成され、各ラベル・スイッチング・ノードはさらに前記関連ラベルに関するラベル情報を管理する制御コンポーネントと前記受信したラベル付けされたパケットを前記ラベル情報に基づいて転送するように構成された転送コンポーネントを含む複数のラベル・スイッチング・ノードと、

各エGRESS・ノードが前記複数のラベル・スイッチング・ノードの一つから転送された前記ラベルのついたパケットを受信するように構成される複数のエGRESS・ノードとを含み、

1つまたはそれ以上のラベル・スイッチング・ノードがラベル・スイッチ経路に属し、仮想リンクを使用して前記ラベル・スイッチ経路に属する前記識別されたラベル・スイッチング・ノードを相互接続し、

10

20

さらに、前記ラベルのついたパケットの少なくとも一つは、初期シグナリング処理で設定される付属のラベルを変更することなしに、出力仮想リンク内の前記複数の個別リンクのいずれか一つへ転送されることが可能であることを特徴とすることを特徴とするシステム。

【請求項 2】

出力仮想リンク内の前記個別のリンクのどれを使って前記受信したパケットを転送するかの詳細が前記制御コンポーネントから隠蔽されることを特徴とする請求項 1 に記載のシステム。

【請求項 3】

ラベル・スイッチ・ネットワークにおいてラベル・スイッチ経路を設定するためのシステムであって、

複数のリンクと、

前記複数のリンク経路で互いに相互接続される複数のノードとを含み、

ラベル・スイッチ経路は前記ラベル・スイッチ・ネットワーク経路でパケットを送信するために識別され 1 つまたはそれ以上のノードを相互接続する 1 つまたはそれ以上のリンクから構成されることと、

前記ラベル・スイッチ経路内の前記 1 つまたはそれ以上のリンクは 1 つまたはそれ以上の仮想リンクを含むことと、

前記 1 つまたはそれ以上の仮想リンクの各々が 1 つまたはそれ以上の物理リンクから構成され、

さらに、前記ラベル・スイッチ経路の各ノードはこれに接続された 1 つまたはそれ以上のリンク経路で各々のラベルを有するパケットを受信して、これに接続された 1 つまたはそれ以上のリンクを経由して各々のラベルに基づいて前記受信したパケットを転送するように構成されることと、

前記ラベル・スイッチ経路の各ノードは前記ラベルに関するラベル情報を保持するように構成された制御コンポーネントと前記ラベル情報に基づいて前記受信したパケットの転送を実行するように構成された転送コンポーネントとを含むことと、

前記ラベル・スイッチ経路のノードについて、受信パケットを転送するために使用する前記リンクが仮想リンクであれば、前記受信パケットは初期シグナリング処理で設定されたラベルを変更することなく前記仮想リンク内の前記 1 つまたはそれ以上の物理リンクのいずれか一つへ転送されることが可能であることを特徴とするシステム。

【請求項 4】

パケットに付属する前記ラベルは、固定長のビット列、TDM フレームの時間スロット位置、光キャリアの波長を含む実体又はその組み合わせのいずれか一つから構成されることを特徴とする請求項 3 に記載のシステム。

【請求項 5】

マルチプロトコル・ラベル・スイッチ・ネットワークで使用するラベル・スイッチング・ルータであって、

複数の入力仮想リンクと複数の出力仮想リンクとを含み各々の入力又は出力仮想リンクは複数の個別リンクを有する複数の仮想リンクと、

1 つまたはそれ以上の前記入力仮想リンク経路で受信したパケットによって運ばれるラベルに関するラベル情報を保持するように構成された制御コンポーネントと、

前記複数の出力仮想リンクの 1 つまたはそれ以上を経由して前記ラベル情報に基づいて前記受信したパケットの転送を実行するように構成された転送コンポーネントと、

複数のエントリを有し、各エントリは入力仮想ポート番号と、入力ラベルと、出力ラベルと、出力仮想ポート番号とを有する、前記ラベル情報を格納するための少なくとも 1 個のラベル転送テーブルとを含み、

前記ラベル・スイッチング・ルータはパケットをルーティングするためのラベル・スイッチ経路の一部として認識され、

さらに、前記パケットの少なくとも一つは初期のシグナリング処理で設定される付属ラ

10

20

30

40

50

ベルを変更することなく出力仮想リンク内の前記複数の個別リンクのいずれか一つに転送されることが可能であることを特徴とするラベル・スイッチング・ルータ。

【請求項 6】

前記制御コンポーネントは前記受信したパケットを転送するために使用しようとする前記出力仮想リンク内の前記個別リンクがどれかについて関与しないことを特徴とする請求項 5 に記載のラベル・スイッチング・ルータ。

【請求項 7】

受信パケットは付属の入力ラベルを有し、前記付属する入力ラベルは転送等価クラスと付属出力ラベルに対応することと、

前記ラベル・スイッチング・ルータは前記付属の入力ラベルと前記受信パケットに付属する入力仮想ポート番号とをキーとして使用して、前記付属の出力ラベルと、前記受信パケットを転送しようとする出力仮想リンクに対応する出力仮想ポート番号とを前記少なくとも一つのラベル転送テーブル内で検索することと、

第 1 のハッシュ関数を使用して前記付属の出力ラベルについて演算してハッシュ値を取得し、前記ハッシュ値が前記受信パケットを転送しようとする前記出力仮想リンク内の前記複数の個別リンクの一つに対応する出力実用ポート番号を表わすことを特徴とする請求項 5 に記載のラベル・スイッチング・ルータ。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は一般にネットワーク内の 2 台の装置間の通信スパン内の複数リンクに関する。さらに詳しく説明すると、本発明はラベル・スイッチ・ネットワーク内の次ホップ装置用に複数リンク間で負荷バランスを採るための方法および装置に関する。

【0002】

【従来の技術】

複数リンク技術はネットワーク内の 2 台の通信装置またはノード間でバンド幅を拡大する簡単な方法として広く使用されている。たとえば、波長分割多重化ネットワークでのパラレル・リンクは互いに隔離された多数の波長を用いて設定される。複数リンクを物理的に別の媒体、たとえば数本のファイバーおよび/または光ファイバ内の数種類の波長、または SONET (同期光ネットワーク) / SDH (同期デジタル階層) 上の STM (同期転送モード) における VC (仮想コンテナ) など、なんらかの多重化技術を応用した共有媒体内のいくつかのスロットに設定できる。

【0003】

リンク・コスト・チューニングと呼ばれる初期のいくつかの方法は複数リンク間で負荷バランスを提供するために使用され、負荷によりリンクのコストを変更し、リンク・コストを用いて送信元と宛先の各対についてもっとも短いリンクまたは経路を見つけるルーティング・アルゴリズムにおいて使用されていた。これらの方法は一部に調節の粒状性が粗いため、またフィードバック処理における遅延に起因する固有の不安定性によるトラフィック発振のため、不十分であることがわかった。

【0004】

「等コスト・マルチパス」法 (ECMP) は多重リンク間で負荷バランスを実現するために使用されていたもう一つの方法である。ECMP 法では負荷に基づいた OSPF コストの動的調節を行なおうとはしないので、ECMP 法を信頼性のあるものにすることができる。等コストの経路が存在するようなトポロジーであれば、経路間でトラフィックを等分割するように試みる。トラフィック分割の次のような方法が使用されてきた。

【0005】

「パケットごとラウンドロビン・フォワーディング」法は経路上の遅延がほとんど等しい場合にのみ適用できる。遅延差はパケット・シリアル化時間に比べて十分に小さくなければならない。パケット・シリアル化時間の 3 倍以上の遅延差ではひどい TCP パフォーマンスの劣化が起こるが、これは順序が正しくないパケットの到着で TCP の高速再送を誘

10

20

30

40

50

発し、TCPを小さな時間枠に制限し、長い遅延路ではパフォーマンスを非常に低くしている。

【0006】

「発信元/宛先ハッシュ」法はIBM社製RT-PCベースのルータでT1-NSFNETと同程度のハッシュ関数たとえばCRC-16を発信元アドレスと宛先アドレスに対して適用する。閾値を設定するかモジュロ演算を実行するかのいずれかによって、ハッシュ空間を利用できる経路の間で均等に分割する。任意の発信元と宛先の間トラフィックは同一経路上にとどまる。

【0007】

【発明が解決しようとする課題】

ここで説明したこれらの方法はIPパケット転送メカニズムに基づいたものである。よって各パケットが転送される経路に沿って資源を確保するためのなんらかのシグナリング・メカニズムはサポートしていない。またトラフィック・エンジニアリングを実現するための明示的ルーティングもサポートしていない。そのため、なんらかのIPパケット転送メカニズムを使用することなしに負荷バランシングを実行することが可能な方法および装置を提供することが望まれる。

【0008】

さらに、多重リンクを使用することは一つまたはそれ以上の潜在的短所を発生させかねない。たとえば、リンク利用における効率の劣化は負荷バランシングなしのバランスのとれていないトラフィックで発生することがある。図1はこの状況を図示したものである。他のリンクたとえばリンク#2やリンク#3を使用していないとしても、過負荷状態の物理リンク#1では重大なパケット損失が発生しうる。負荷バランシングなしでは、多重リンク内のパケット損失を減少するためオーバーエンジニアリングが行なわれる。図2は従来の負荷バランシングを行なっている状態を示している。図2に図示してあるように、同じクラスに所属するパケットは多重リンクの内の特定の一つに転送されて、多重リンク全体にわたるトラフィックの均等な分配またはバランスを行なえるようにする。あるパケットが一つの物理リンクから別のリンクへ分配されるとき、これに付随するラベルも変更される。たとえば、物理リンク#1でラベルAのパケットが物理リンク#2へ再ルーティングされる場合、パケットは物理リンク#2で新しいラベルA'が付与される。

【0009】

多重リンクにわたる負荷バランシングにより、ある種のネットワークたとえばマルチプロトコル・ラベル・スイッチ(MPLS)ネットワークでは相当量の処理オーバーヘッドが発生する。従来のMPLSネットワークでは、各パケットにラベルを付与しパケットの転送はラベルに依存する。同じラベルを有するパケットは同一の方法で処理される、たとえばあらかじめセットアップしてあるラベル・スイッチ経路またはリンクに沿って全部が転送される。この種のネットワークでは、一つのリンクから別のリンクへ、たとえば渋滞のためパケットを再転送する場合、再転送されるパケットには新しいリンクに沿って伝送するための新しいラベルを付与する必要がある。これには新しいラベルの作成が必要であり、大量の処理オーバーヘッドが発生する。たとえば、新しいラベル・スイッチ経路をセットアップする必要があり、新しいラベル転送情報を新しいラベルスイッチ経路に沿ったすべてのノードに配布する必要が生じる。大量の平行多重リンクがあるMPLSネットワークでは、負荷バランシングは問題を生じやすい。したがって、ネットワークたとえばMPLSネットワークにおける多重リンクにわたって負荷バランシングを実行しつつ処理オーバーヘッドを満足できる水準に維持できるような方法および装置を提供することが望まれる。

【0010】

さらに、MPLSネットワークにおいて、到着パケットをラベル・スイッチ・ルータで受信する場合、到着パケットに付随するラベルを新しいラベルに変更するかマッピングし直してから次のラベル・スイッチ・ルータへ供給する。このラベル・マッピング処理は付随ラベル転送テーブルの関連ラベル転送情報を参照することで行なう。付随ラベル転送テ

10

20

30

40

50

ブルは多重リンク全部を経由してラベル・スイッチ・ルータへ転送されるエントリを全部含むのが普通である。全部の多重リンクについてのエントリ数は極めて大きくなることがある。その結果、関連のあるラベル転送情報をそれぞれのラベル・スイッチ・ルータで検索する時間が実質的に大きくなり、これによりラベル・マッピング処理に遅延が生じることで、パケットの転送レイテンシに影響を与えるようになる。したがって、ネットワーク、たとえばMPLSネットワークでの転送レイテンシを改善することができるような方法および装置を提供することが望まれる。

【0011】

【課題を解決するための手段】

本発明の代表的な一例によれば、ラベル・スイッチ・ネットワーク内の2台の装置間の一組のリンクから構成される仮想リンクがラベル・スイッチング・ルータの制御コンポーネントにより一つの実体として扱われる。仮想リンクへ伝送しようとする各パケットはこれに割り当てられるクラスに基づいてラベルが付与される。各パケットに付与されるラベルは、短く、固定長で、ローカルに有意な識別子であって特定のクラスを識別するために使用される。特定のラベルが対応するクラスに結合される。特定のラベルは同一クラスに属するパケットについて使用され、このようなパケットが仮想リンク内のリンクのいずれか一つへ転送することができる。仮想リンク内で一つのリンクから別のリンクへパケットを再ルーティングする場合、そのパケットに付随するラベルは変更されない。ラベル配布プロトコルを使用してイングレス・ノードからエグレス・ノードへ各クラスについての経路（またはラベル・スイッチ経路）を決定する。

【0012】

さらに、本発明の別の代表的な一例によれば、各クラスは出力ラベルを変更することなく出発仮想リンク内のリンクの一つにマッピングすることができる。このマッピングは選択したハッシュ関数を使用して行なわれる。選択したハッシュ関数を使用してラベルを整数にマッピングし、仮想リンク内の少数のリンクの組に分配するようにする。

【0013】

さらに、本発明のさらに代表的な一例によれば、別のラベルを有するパケットは仮想リンク内のリンクの間で配分される。同一ラベルを有するパケット（つまり同一クラスのパケット）は仮想リンク内で同じリンクへ転送される。仮想リンク内の各種リンクについて負荷バランスをとるためには、選択した時間に別の選択したハッシュ関数を使用して負荷条件によりパケットを配分することができる。

【0014】

最後に、本発明のさらに別の代表的な一例によれば、送出側と受信側ラベル・スイッチ・ルータで使用されるそれぞれのハッシュ関数が同期される。ハッシュ関数を同期することで、負荷転送テーブル内のエントリ数を減少し、これによって検索時間とメモリ記憶条件を減少する。

【0015】

本発明の代表的な実施例は多数のノードを有するラベル・スイッチ・ネットワークである。各ノードは、多数の入力仮想リンクと多数の出力仮想リンクとを含み各入力または出力仮想リンクが多数の独立したリンクを有する多数の仮想リンクと、入力仮想リンクの一つまたはそれ以上を経由してそれぞれのラベルを有するパケットを受信しまた出力仮想リンクの一つまたはそれ以上を経由してそれぞれのラベルに基づいて受信パケットを転送するように構成され、ラベルに関連したラベル情報を保持するように構成された制御コンポーネントとラベル情報に基づいて受信パケットの転送を実行するように構成された転送コンポーネントとをさらに含むラベル・スイッチング・ルータとを含む。

【0016】

ラベル・スイッチング・ルータはルーティング・パケットのラベル・スイッチ経路の一部として認識される。ラベル・スイッチ経路に関して、入力仮想リンクそれぞれの中の個別のリンクと出力仮想リンクそれぞれの中の個別のリンクがそれぞれ制御コンポーネントによって一つの実体として扱われる。

【 0 0 1 7 】

受信パケットを転送するのに出力仮想リンク内の個別のリンクの内のどれを使用することになるかについての詳細は制御コンポーネントから隠蔽される。言い換えれば、制御コンポーネントはパケットのそれぞれのラベルをつけ変えることなく、出力仮想リンク内の個別リンクのどれにでもパケットを転送する。

【 0 0 1 8 】

パケットに付随するラベルは、固定長のビット列、TDMフレーム内の時間スロット位置、光キャリアの波長を含む実体のいずれか一つまたはその組み合わせから構成される。さらに、各ラベルは多数の転送等価クラスの一つに割り当てられ、ラベルをラベル・スイッチング・ルータで使用して出力仮想リンクの個別リンクの一つへ対応するクラスに属するパケットを転送する。

10

【 0 0 1 9 】

各ノードはさらにハッシュ関数を使って出力仮想リンク経路で転送しようとするパケットのラベルを計算してハッシュ値を得るように構成された制御論理を含み、ハッシュ値はパケット転送に使用される出力仮想リンク内の個別リンクの一つを表わす。ハッシュ関数を使用してパケットのラベルを計算することで、ラベルを変更せずに出力仮想リンク内の個別のリンクの一つへパケットを転送することができるようになる。

【 0 0 2 0 】

制御論理はまた出力仮想リンク経路で転送しようとするパケットのそれぞれのラベルについて演算するのに使用しようとする多数のハッシュ関数のうちの一つを選択して対応するハッシュ値を得るようにも構成される。それぞれのラベルは一つまたはそれ以上の対応するクラスに対応する。対応するハッシュ値はパケットを転送するために使用しようとする出力仮想リンク内の個別のリンクの一つまたはそれ以上を表わす。それぞれのラベルについて演算するハッシュ関数の一つを選択することによりそれぞれのラベルを付与したパケットが出力仮想リンク内の個別のリンクの中から配分される。それぞれのラベルについて演算するために使用するハッシュ関数は負荷バランシング条件の一つまたはそれ以上に基づいて選択される。

20

【 0 0 2 1 】

各ノードはさらに多数のラベル転送テーブルを含み、各ラベル転送テーブルは多数のエントリを有し、各エントリは入力仮想ポート番号、入力ラベル、出力ラベル、出力仮想ポート番号を含む。入力仮想ポート番号は入力仮想リンクに関する識別情報を表わす。入力ラベルは受信パケットに関するラベル情報を表わす。出力ラベルは出力パケットに関するラベル情報を表わし、出力仮想ポート番号は出力仮想リンクに関する識別情報を表わす。

30

【 0 0 2 2 】

それぞれのラベル転送テーブルには入力仮想リンク内の個別リンクの一つが関連する。第1のラベル転送テーブルでは、そのエントリについての入力ラベルのハッシュ値が第1のラベル転送テーブルに付随する入力仮想リンク内の個別リンクに対応する入力実用ポート番号に等しければエントリがタグされる。第1のラベル転送テーブルを変更して、タグのついたエントリだけを含めるようにする。第1のラベル転送テーブルに付随する入力仮想リンク内の個別リンク経路でパケットを受信した場合、受信パケットに対応するエントリについて第1のラベル転送テーブル内のタグつきエントリだけを検索する。

40

【 0 0 2 3 】

本発明により多数の利点および/または利益が実現できる。たとえば、ラベル配布プロトコルを使用することにより、ラベル・スイッチ経路の決定は、最短経路を見つけるためのIPルーティング処理で使用されるIPパケット転送メカニズムやリンク・コストのどれからも独立して行なわれるので、これによりIPパケット転送メカニズムに付随する固有の不安定性やトラフィック発振を回避できる。ラベル配布プロトコルはトラフィック・エンジニアリングで使用する明示的ルーティングもサポートしているので、ネットワーク設計やトラフィック・フローの最適化が実現可能である。

【 0 0 2 4 】

50

さらに、仮想リンクは一つの実体として扱われるので、ラベル・スイッチ経路を設定・保持するために必要とされるシグナリングおよびメッセージングの総量を減少し、これにより大量の平行多重リンクから構成されるスケーラブル・ネットワークのサポートが容易になる。

【0025】

さらに、本発明によればラベル転送テーブルにおける受信パケットのラベル検索が容易になるので結果としてそれぞれのラベル・スイッチング・ルータで多重リンクへのラベルのマッピングし直しに必要な時間が短縮される。これによりフィードバック処理の遅延や発振の危険性を減少しデータパーストにたいして応答が高速の多重リンクの利用を向上する。

10

【0026】

図面や請求項を含む明細書の残りの部分を参照することで、本発明のその他の特徴や利点が明らかになる。本発明の更なる特徴や利点ならびに本発明の各種実施例の構造と動作は、添付の図面を参照して以下で詳細に説明し、図面において同じ参照番号は同一または機能的に類似した要素を表わすものとする。

【0027】

【発明の実施の形態】

一つまたはそれ以上の代表的実施例の形で本発明を説明する。本発明の代表的実施例によれば、制御の観点から、負荷バランシングを実現する特定の物理リンクの識別および選択は対象ではない。言い換えれば、多重リンクは負荷バランシング制御の観点から、集合的に仮想リンクとして（すなわち、擬似的な1本の太いパイプとして）扱われる。したがって、あるパケットについて初期のシグナリング処理で設定されるラベル情報は、一つの物理リンクから別のリンクへパケットが分配される場合に変更を受けない。図3は本発明の代表的実施例の動作を示す。図3に図示したように、物理リンク#1から物理リンク#2へラベルAをつけたパケットを再ルーティングする場合、パケットは同じラベル、ラベルAのままである。

20

【0028】

代表的実施例において、本発明は多重プロトコル・ラベル・スイッチ（MPLS）ネットワークへ適用される。しかし、本明細書で提供される開示に基づいて本発明をその他の同様な種類のネットワークへ適用できることは当業者には明らかであろう。図4はMPLSネットワーク30を示す略模式図である。図4に図示したように、MPLSネットワーク30は多数のラベル・エッジ・ルータ（LER）32、34と多数のラベル・スイッチング・ルータ（LSR）36を含む。LERはさらに2種類に分類される、すなわちイングレスLER32a、32bとエグレスLER34a、34bである。LER32、34は後述する同様の機能をそれぞれ実行する他のノードにより置換可能であり、LSR36はL3スイッチ、光クロスコネクタなどを含む他の種類のラベル・スイッチング・ノードで置き換えうることを理解されるべきである。

30

【0029】

イングレスLER32aはたとえば送信元（図示していない）から指定された宛先アドレスA1をもつIPパケットを受信する。送信元はIPパケットを送信することができるなんらかの装置、たとえばクライアント端末やネットワークのノードなどである。IPパケットをイングレスLER32aで受信した場合、イングレスLER32aはIPパケットを転送等価クラス（FEC）たとえばFEC A1に分類する。転送等価クラスは伝送について同じ条件を共有するパケットのグループを表現したものである、すなわち同じ転送等価クラス内の全部のパケットはMPLSネットワーク30を通る同じ経路をたどり各LSR36で同じ処理を受ける。IPパケットに適当なFECを付与したあと、イングレスLER32aはIPパケットに出力ラベル、たとえば1をつけ、これはこの場合FEC A1である所定のFECに対応するものである。IPパケットに関連するラベルは以下の実体のいずれか一つまたはその組み合わせから構成され、これにはたとえば固定長ビット列、TDMフレームの時間スロット位置および光キャリアの波長を含む。イングレスLER

40

50

3 2 a は新しくラベルをつけた I P パケットを L S R 3 6 a へ転送する。同様に、イングレス L E R 3 2 b は指定された宛先アドレス B 1 を有する I P パケットに対して同じ機能を実行する。

【 0 0 3 0 】

ラベルのついた I P パケットを受信すると、L S R 3 6 a はラベル転送テーブルを使用して I P パケットのラベルを 1 から 2 へ切り換える。ラベル 2 は同様に I P パケットに割り当てられた F E C である F E C A 1 に対応する、すなわち F E C A 1 に属するすべてのパケットにはラベル 2 が付与される。L S R 3 6 a で実行されるこのラベルスイッチ機能は M P L S ネットワーク 3 0 の特徴である。各 L S R 3 6 はそれ自身のラベル転送テーブルと特定 F E C についてそれ自身のラベルを保持する。ラベル転送テーブルは入力および出力のラベリングと F E C 情報とを内包する。この情報を用いて到着 I P パケットからの入力ラベルを、同じ F E C に所属する出力ラベルへマッピングする。出力ラベルのついた I P パケットはここで L S R により別の L S R へ転送される。したがって、各 I P パケットをカプセル化するために使用されるラベルはリンク内部の送出側 L S R から受信側 L S R へ同一のまま維持される。しかし、リンクごとにラベルを変更できることは注意すべきである。ラベルは、I P パケットを受信し受信側 L S R で検証した場合、受信側 L S R の観点からの入力ラベルとして特徴づけられ、またラベルは、I P パケットが送出側 L S R から次ホップ L S R へ送出された場合、送出側 L S R の観点からの出力ラベルとして特徴づけられる。本実施例では、L S R 3 6 a はそのラベル転送テーブルを使用することにより、L E R 3 2 から（特定の入力ポート経由で）受信した I P パケットのラベル 1 が F E C A 1 に対応することを判定し結果として I P パケットのラベルを 2 に切り換えるが、このラベルも F E C A 1 に対応するものである。新しいラベル 2 のついた I P パケットは L S R 3 6 a により次の L S R 3 6 b へ送信される。よって、各 L S R 3 6 で、I P パケットのラベルは関連するローカル・ラベル転送テーブルにより切り換えられるが、I P パケットの F E C は M P L S ネットワーク 3 0 内の伝送を通じて同じまま保持される。

【 0 0 3 1 】

エグレス L E R 3 4 a が I P パケットを受信した場合、ラベル 4（これは L S R 3 6 c によって割り当てられる）が除去され、I P パケットについて従来のレイヤー 3（ネットワーク層）参照を行なう。参照の結果に基づき、I P パケットは宛先アドレス A 1 へ転送される。

【 0 0 3 2 】

ラベル・スイッチド・パス（L S P）としても公知で、同じ F E C に所属する I P パケットが M P L S ネットワーク 3 0 内部で転送される経路は、初期のラベル結合手順でユニークに決定される。つまり、L S P はデータパケット伝送の開始より前に設定される。初期ラベル結合手順はラベルに F E C をマッピングまたは結合するために行ない、ラベル結合情報はあるプロトコル、たとえば、インターネット・エンジニアリング・タスクフォース（I E T F）で策定されるラベル配布プロトコル（L D P）などを用いて L S R 3 6 へ分配される。

【 0 0 3 3 】

図 4 に図示してあるリンクのどれか、たとえばイングレス L E R 3 2 a と L S R 3 6 a の間のリンク # 1 3 8 と L S R 3 6 a と L S R 3 6 d の間のリンク # 5 4 0 は、多数のリンクから構成されることがある。代表的実施例において、L S R 3 6 の入力リンク（群）（すなわち、パケット受信用のリンク）と出力リンク（群）（すなわち、パケット転送用のリンク）はそれぞれ多数のリンクから構成される。しかし、各 L S R 3 6 で、入力リンク（群）と出力リンク（群）がどちらも多数のリンクから構成されなくとも良いことには注意すべきである。たとえば、別の実施例において、入力リンクまたは出力リンクのどちらかが多重リンクで構成される。図 4 a から図 4 d は多重リンクの各種代表的実施例を示す。2 台の装置間の通信スパン内部の多重リンクは本明細書において集合的に「仮想リンク」と呼ぶ。図 5 a を参照すると、たとえば、仮想リンクはイーサ・フレームを使用し

10

20

30

40

50

てパケットを送信する多数のファイバで構成されるように図示してあり、図5 bでは仮想リンクはSONET/SDHフレームを使用してパケットを送信する多数のファイバから構成されるように図示してある。多数のフレームを使用すると波長分割多重化技術を用いてファイバーのうちの1本にパケットを送信することができる。図5 cにおいて、仮想リンクはイーサ・リンクとSONET/SDHリンクの組み合わせから構成されるように図示してある。また図5 dでは仮想リンクはSONET/SDHリンクのタイムスロットにより提供される多数の論理リンクから構成されるように図示してある。

【0034】

図6は本発明を実施するために使用できるLSR36のアーキテクチャを図示している。図6を参照すると、アーキテクチャには2つの独立したコンポーネントが含まれており、すなわち転送コンポーネント(データプレーンとも呼ばれる)50と制御コンポーネント(制御プレーンとも呼ばれる)52である。転送コンポーネント50は多数の到着仮想リンクからデータ・パケットを受信して当該パケットを多数の出発仮想リンクへ転送する。各仮想リンクは(到着側と出発側)多数の個別リンクまたは多重リンクを含む。一つの実装において、到着仮想リンク内部の各リンクは入力実用ポートに接続され、これがさらにネットワーク・カード(図示していない)へ接続される。同様に、出発仮想リンク内部の各リンクは出力実用ポートへ接続される。各入力実用ポートで、転送コンポーネント50はラベル転送テーブルまたはデータベースを使用してデータパケットが保持しているラベルに基づきその入力実用ポート経由で受信したデータパケットの転送を実行する。ラベル転送テーブルに関する詳細については以下で提供する。

【0035】

制御コンポーネント52はMPLSネットワーク30内部のLSPに沿って相互接続されているLSRのグループ間でラベル転送情報(バインディング(結合)とも呼ばれる)を作成、保持、分配する役割がある。ラベル転送情報はラベルが作成された時点で生成され制御コンポーネント52によって保持されるラベル転送テーブル内に格納される。各LSR36は自分のラベル転送情報をMPLSネットワーク30内の他のLSR36へ配布する。このラベル配布処理はまたラベル・バインディング交換とも呼ばれることがある。制御コンポーネント52はIPルーティング・テーブルも使用してラベル・バインディング交換を行ない、ここで隣接するMPLSノードがIPルーティングテーブル内に含まれている個別の発信元-宛先ペアについてのラベルを交換する。一つの例では、宛先ユニキャスト・ルーティングのラベル・バインディング交換はIETF策定LDPを使用して行なわれる。明示的なルーティングをサポートシトラフィック・エンジニアリングを実現するため、基本LDPへの拡張も策定されている。その他のプロトコル、たとえばMPLS-RSVP(資源予約プロトコル)やCR-LDP(制限ベースのラベル配布プロトコル)もラベル・バインディング交換を実行するために使用できる。

【0036】

図7は従来の方法で図6に図示したLSR36のアーキテクチャ内部でどのように多重リンクが処理されるかを示す。図7を参照すると、多重リンクの中の各リンクは別個の実体として扱われている。たとえば、3本のリンク60、62、64と各々のLSPは別個の独立した実体として扱われている。各々のLSPについて、ラベル配布処理やLSPを維持するためのその他の処理が必要である。

【0037】

本発明の代表的実施例によれば、制御コンポーネント52は多重リンク(すなわち仮想リンク)を単一の実体として扱う。言い換えれば、仮想リンクを構成する多くの物理および/または論理リンクの詳細は制御コンポーネント52から隠蔽される。図8は本発明の代表的実施例によりどのように多重リンクを扱うかを示す。図8に図示してあるように、多重リンクは単一の仮想リンク70として扱われる。多重リンクを単一の仮想リンクとして扱うことにより、仮想リンクを構成する物理および/または論理リンクの各々の詳細が必要ではなくなり単一のLSPが多重リンクの組全体について仮想的に設定されるため、シグナリングとメッセージングの総量が効果的に減少する。言い換えれば、制御コンポーネ

10

20

30

40

50

ント52の観点から、仮想リンク70内の多重リンクは単一のLSPとして見なされる、すなわちパケットを転送するために使用される仮想リンク70内の多重リンクのどれを用いるかが制御コンポーネント52から隠蔽され、またこれらの多重リンクは単一のLSPを設定するために使用されるのと等しい量のシグナリングおよびメッセージングの量で集会的に設定される。代表的実施例によれば、ラベル配布処理はFECに属するパケットの共通のラベルを結合し、ここで共通のラベルは同じFECに属するパケットを出発仮想リンク内の個別のリンクのいずれか一つへ転送するために使用できる。

【0038】

図9と図10は本発明によるLSR36の転送コンポーネント50の代表的動作を示す。図9と図10を参照すると、転送コンポーネント50は多数の入力・出力仮想ポートと、多数の入力・出力実用ポートと、多数のフロントエンドおよびバックエンド・ラベル転送モジュールおよびバックプレーン・スイッチを含む。仮想ポートは制御コンポーネント52により単一の実体と見なされる一組の実用ポートであると見なされる。仮想ポート内の実用ポートは各々が一つの実用ポートを他のポートから識別するために使用するユニークなポート番号を有する。

10

【0039】

到着仮想リンクは各々の入力仮想ポート（およびこれを構成する入力実用ポート）へ接続される。入力実用ポートはさらにフロントエンド・ラベル転送モジュールへ接続される。代表的には、各々の入力実用ポートはネットワーク・カード（図示していない）へ接続され、これに対応するフロントエンド・ラベル転送モジュールを含む。フロントエンド・ラベル転送モジュールはバックプレーン・スイッチへ接続される。バックプレーン・スイッチはバックエンド・ラベル転送モジュールへ接続され、これがさらに出力実用ポート（および出力仮想ポート）へ接続される。最後に、出力仮想ポートは各々の出発仮想リンクへ接続される。

20

【0040】

転送コンポーネント50は本発明による以下の代表的な方法で動作する。各々の入力仮想ポートを用いて対応する到着仮想リンクからパケットを受信する。つまり、到着仮想リンクたとえば仮想リンクAからのパケットは、到着仮想リンク内の個別のリンクに対応する多数の入力実用ポート（これは同じ入力仮想ポートに属する）によって受信される。入力実用ポートはこれに対応する多数のフロントエンド・ラベル転送モジュールへ接続される。各フロントエンド・ラベル転送モジュールは到着パケットの転送に使用すべき出力仮想ポートおよび実用ポートがどれなのかを決定する責任がある。各フロントエンド・ラベル転送モジュールは関連するラベル転送テーブルの適切なエントリを参照してこの決定を行なう。たとえば、パケットに付けられたラベルと仮想ポート番号を参照のキーとして使用できる。到着パケットを受信する入力仮想ポート番号は分かっている。よって、到着パケットについて入力仮想ポートの仮想ポート番号は容易に取得できる。参照の結果として、パケットを転送するために使用すべき出力仮想ポートの仮想ポート番号と次ホップLSRへの仮想リンクで使用すべき新しいラベルが取得される。前述したように、各々のフロントエンド・ラベル転送モジュールまたは入力実用ポートはラベル転送テーブルに関連するが、ラベル転送テーブルは多数の方法で構成できることに注意すべきである。たとえば、各々のLSR36は一つだけのラベル転送テーブルを含んでいたり、幾つかのフロントエンド・ラベル転送モジュール（これが仮想ポートを含む多数の入力実用ポートを処理する）が一つのラベル転送テーブルを共有することがある。本明細書に提供した開示に基づいて、ラベル転送テーブルを構成する他の方法が当業者には理解されよう。参照処理については以下でさらに詳しく説明する。出力仮想ポート内の出力実用ポート（とこれに対応する実用ポート番号）のどれを使うかを決定するメカニズムについては後述する。

30

40

【0041】

参照から得られた情報を使って、フロントエンド・ラベル転送モジュールはバックプレーン・スイッチに利用可能な出力仮想ポートの一つへパケットをスイッチさせる。図9に図示してあるように、たとえばバックプレーン・スイッチは出力仮想ポートの各々の中の出

50

力実用ポートの間でパケットをスイッチするために使用する。さらに詳しく説明すると、到着する、 α 、 β パケットは、入力仮想ポートA経由で仮想リンクAから受信される。関連する参照情報がフロントエンド・ラベル転送モジュールによって得られてから、フロントエンド・ラベル転送モジュールはバックプレーン・スイッチにパケットを分割させて2個の別々の仮想リンクへこれらを配送させる。および α パケット(新しいラベル α' と β がついている)は出力仮想ポートK経由で仮想リンクKへ配送され、 β パケット(新しいラベル β' がついている)は出力仮想ポートX経由で仮想リンクXへ配送される。

【0042】

さらに、バックエンド・ラベル転送モジュールもパケットのトラフィックをバランスさせるスイッチング機能を実行する。たとえば、図10に図示してあるように、バックエンド・ラベル転送モジュールは出力仮想ポート内の出力実用ポートの間でパケットを割り当てて出力仮想リンクへ配送すべきトラフィックをバランスさせるようにスイッチング動作する。詳細には後述するように、ハッシュ関数をパケットの出力ラベルとの関連で使用して仮想リンク内の各種リンクの間にパケットを割り当ててパケット・トラフィックをバランスさせる。

10

【0043】

図11は本発明によりパケットを転送するLSR36の各フロントエンド・ラベル転送モジュールで使用するラベル転送テーブルの好適実施例を示す。前述のように、好適実施例において、各々のフロントエンド・ラベル転送モジュールはそれ自身のラベル転送テーブルを保持する。図11に図示してあるように、ラベル転送テーブルは入力仮想ポート番号90のフィールド、入力ラベル92、出力ラベル94、出力仮想ポート番号96を含む。これらのフィールドにある情報は初期にたとえばラベルとLSPを作成する時点でLDPを使用して伝搬する。前述のように、ラベル転送テーブルはLSR36の制御コンポーネント52によって維持される。

20

【0044】

前述の参照処理をさらに以下で説明する。もう一度図9を参照すると、たとえば、 α とラベルのついたパケットを入力仮想ポートAで仮想リンクAから受信する。図11に図示したラベル転送テーブルを使用して、ラベル転送テーブルからAと α をキーとして参照を行なうと出力仮想ポート番号“K”と出力ラベル“ α' ”を次ホップLSRへの出発仮想リンクで使用すべきことが分かる。出力仮想ポート番号はパケット転送に使用すべき出力仮想ポート(とそれに対応する出発仮想リンク)を識別する。

30

【0045】

次に、識別された出力仮想ポート内でパケット転送に使用すべき出力実用ポートを決定する。あらかじめ選択しておいたハッシュ関数を使って出力ラベルを整数(又はハッシュ値)にマッピングして識別した出力仮想ポートで少ないリンクの組に対して配布を行なう。別のハッシュ関数を使えば別のトラフィック・パターンでの変動に対する負荷バランスに対応することができる。出力実用ポート番号は出力ラベルのハッシュ値を取ることにより得られる。図11に図示してあるように、たとえば、 $f(X)$ がラベルXのハッシュ値を表わすものとする、出力実用ポート番号が $f(\alpha')$ の出力実用ポートへパケットを転送する。

40

【0046】

LSR36の制御コンポーネント52は仮想リンク内のFEC各々にユニークなラベルを割り当てるので、同じ出力ラベルが同じFECに属するパケットに割り当てられる。したがって、出力ラベルのハッシュ値として定義された出力実用ポート番号も又同じFECに属する全てのパケットでユニークである。つまり、同じFECに属する全部のパケットは各々の出力仮想ポート内の同じ出力実用ポートを使用することになり、パケット順序が保存される。

【0047】

図12は本発明による到着パケット受信時のLSR36のフロントエンド・ラベル転送モジュール各々で使用するラベル転送テーブルの別の好適実施例を示す。後述するように、

50

代表的なラベル転送テーブルはもっと少ない数のエントリを含み、これによって検索時間とメモリ要求の量を減少する。

【0048】

本発明の代表的実施例によれば、受信側LSRによって使用されるハッシュ関数は送出側LSRで使用されるハッシュ関数と同期される。前述のように、ハッシュ関数は出力ラベルに演算を行なって対応する出力実用ポート番号を得るために使用される。送信側LSRから受信側LSRへ送出されたパケットについて、送信側LSRで付与した出力ラベルは受信側LSRで受信される入力ラベルと同一である。したがって、送信側と受信側のLSRでの各々のハッシュ関数を同期することにより、入力実用ポート番号が入力ラベルのハッシュ値として計算できるため、受信側LSRは入力実用ポートのどれを使用してパケットを受信するか通知を受けることができる。これは入力実用ポートに接続された各々のフロントエンド・ラベル転送モジュールに付随するラベル転送テーブルにおけるエントリ数を減少するために使用できる。

10

【0049】

ラベル転送テーブルのエントリ数は以下のようにすると減少できる。図12を参照すると、入力仮想ポート番号90、入力ラベル92、出力ラベル94、出力仮想ポート番号96といったフィールドの情報に加えて、フィールドの他の情報すなわち入力ラベルのハッシュ値98と入力実用ポート番号100も利用できる。入力ラベルのハッシュ値は入力ラベルについてハッシュ関数を演算することで得ることができ、入力実用ポート番号についてはフロントエンド・ラベル転送モジュールはこれに付属する入力実用ポートの認識に関する情報を有している。「タグ」フィールドの情報は以下の決定に基づいて伝播する：各エントリについて、入力ラベルのハッシュ値が入力実用ポート番号に等しい場合には、特定のエントリをマークする又はタグ付けする。たとえば、図12に図示したように、実用ポートが実用ポート番号[n]を有しこの実用ポートが仮想ポート番号「A」を有する仮想ポート内にあると仮定する。入力ラベル'のハッシュ値f(')が入力実用ポート番号「n」に等しい場合、入力ラベル'のエントリがタグされる。言い換えれば、入力実用ポート番号「n」に対応するハッシュ値を有するエントリだけがタグされる。前述の方法で配布したラベル転送テーブルは入力実用ポート番号「n」を有する入力実用ポート経由でパケットを受信するフロントエンド・ラベル転送モジュールとの関連で使用される。タグのついたエントリを用いることでより高速な参照が実現できる。このラベル転送テーブルを使用して参照又は検索を行ない関連するラベル転送情報を取得する場合、タグのついたエントリだけを検索することにより検索時間の総量を減少できる。他のエントリについての検索は、送信側LSRが番号を出力ラベルのハッシュ値であると決定した一つの出力実用ポートへ各パケットを転送するため、不要である。

20

30

【0050】

別の代表的実施例において、ラベル転送テーブルはタグのついたエントリだけを含む。タグのついたエントリだけを格納することにより、ラベル転送テーブルのサイズを小さくしておくことができ、ラベル転送テーブルを格納するのに必要なメモリ・ストレージ量も対応して減少する。MPLSネットワーク内には多数のラベル転送テーブルが存在しており、各フロントエンド・ラベル転送モジュールについて潜在的に一つのラベル転送テーブルがあるので、メモリ・ストレージについて有意な節約が本発明により実現できる。

40

【0051】

代表的実施例において、本発明はモジュール型又は内蔵型のいずれかの方法で制御論理の形でソフトウェアを使用して実現される。しかし、本発明はハードウェアを使用して又はハードウェアとソフトウェアの組み合わせを使用して実現することもできることは当業者には明らかなはずである。

【0052】

本明細書で説明した例および実施例は本発明の説明を目的としたものであってこれに鑑みた各種の変更又は変化が当業者には示唆されまた本出願の趣旨と範囲内又添付の請求項の範囲内に含まれるべきであることが理解されよう。本明細書で引用した全ての刊行物、特

50

許、特許出願はその全体として全ての目的で参照に含まれる。

【図面の簡単な説明】

【図 1】図 1 は負荷バランシングなしのバランスしていないトラフィック下でのリンク使用における効率の劣化を示す略模式図である。

【図 2】図 2 は負荷バランシングありの多重リンクに対する分散トラフィックを示す略模式図である。

【図 3】図 3 は本発明の代表的実施例の動作を示す略模式図である。

【図 4】図 4 はマルチプロトコル・ラベル・スイッチ・ネットワークを示す略模式図である。

【図 5】図 5 (a) から図 5 (d) は多重リンクの各種代表的実施例を示す略模式図である。 10

【図 6】図 6 は本発明の実施に使用することができるラベル・スイッチング・ルータのアーキテクチャを示す略模式図である。

【図 7】図 7 は図 5 (a) ~ (d) に図示したラベル・スイッチング・ルータの従来アーキテクチャ内でどのように多重リンクが扱われるかを示す略模式図である。

【図 8】図 8 は本発明の代表的実施例により多重リンクがどのように処理されるかを示す略模式図である。

【図 9】図 9 は本発明によるラベル・スイッチング・ルータの転送コンポーネントの代表的実施例を示す略模式図である。

【図 10】図 10 は本発明によるラベル・スイッチング・ルータの転送コンポーネントの別の代表的実施例を示す略模式図である。 20

【図 11】図 11 は本発明によるラベル・スイッチング・ルータによりパケットを転送するために使用されるラベル転送テーブルの代表的実施例を示す略図である。

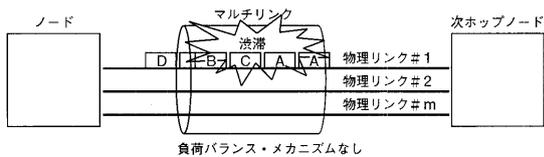
【図 12】図 12 は本発明による到着パケット受信時にラベル・スイッチング・ルータにより使用されるラベル転送テーブルの別の代表的実施例を示す略図である。

【符号の説明】

40 : リンク 5、C1, C2, C3, C4 : ラベル・スイッチング・ルータ (CORE)
 E1, E2, E3, E4 : ラベル・スイッチング・ルータ (エッジ)、A1, B1 : IP 宛先アドレス、
 1, 2, 3, 4, 1, 2, 3, 4 : ラベル、90 : 入力仮想ポート #、
 92 : 入力ラベル、94 : 出力ラベル、96 : 出力仮想ポート #。 30

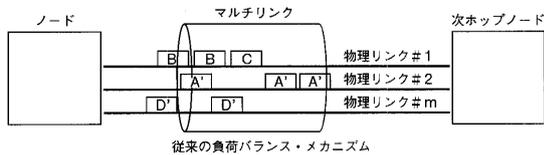
【図1】

図 1



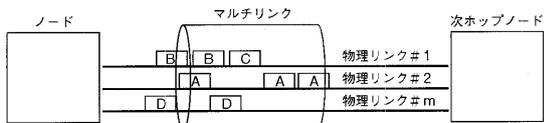
【図2】

図 2



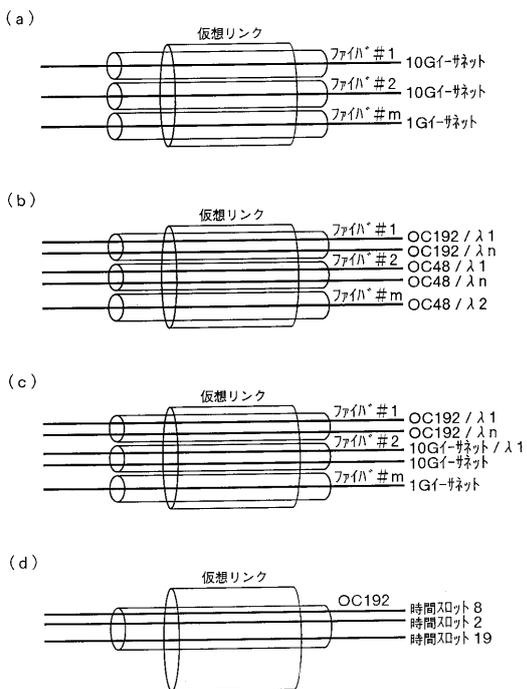
【図3】

図 3



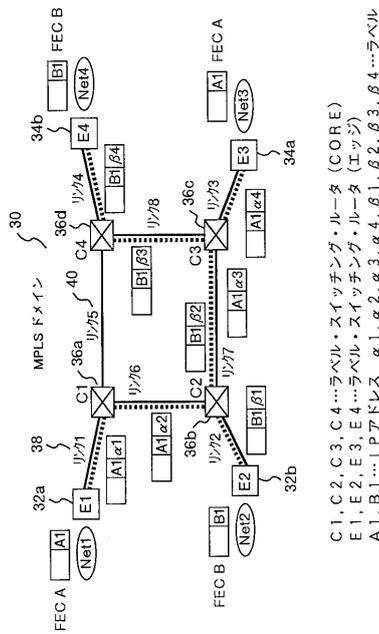
【図5】

図 5



【図4】

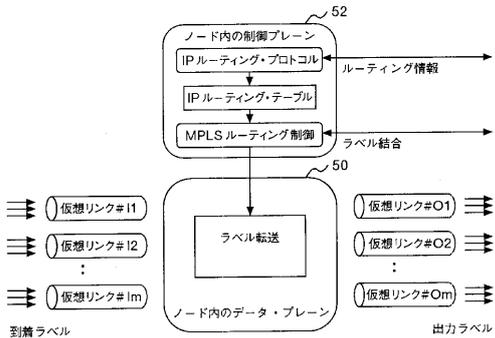
図 4



C1, C2, C3, C4...ラベル・スイッチング・ルータ (CORE)
 E1, E2, E3, E4...ラベル・スイッチング・ルータ (エッジ)
 A1, B1...IPアドレス α1, α2, α3, α4, β1, β2, β3, β4...ラベル

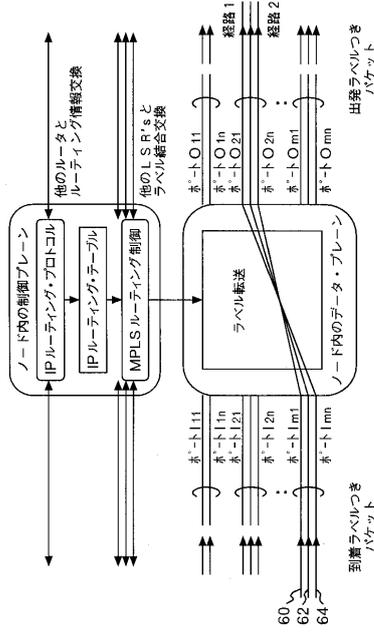
【図6】

図 6



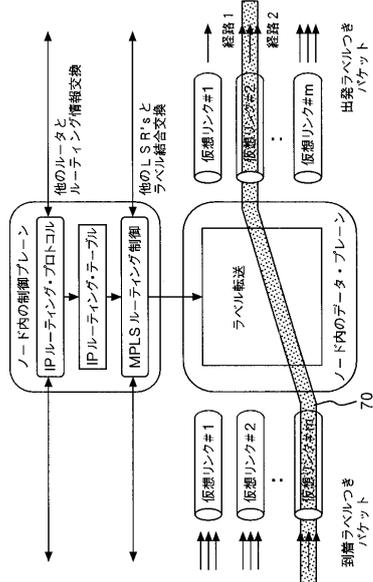
【 図 7 】

図 7



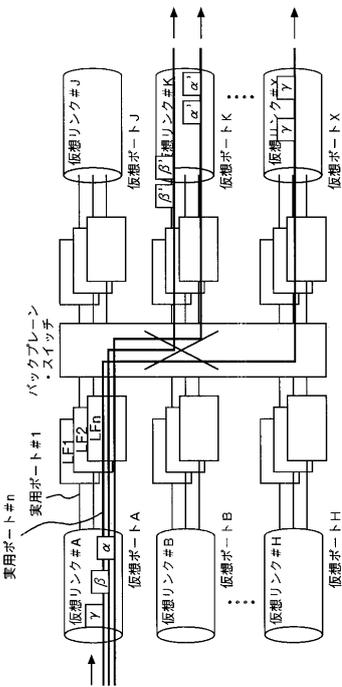
【 図 8 】

図 8



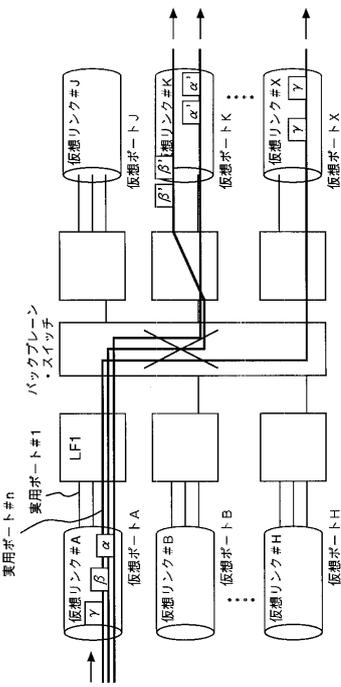
【 図 9 】

図 9



【 図 10 】

図 10



【図 1 1】

図 1 1

MPLS制御プレーンで維持されるラベル転送テーブル

	90	92	94	96	
入力実用ポート#	入力仮想ポート#	入力ラベル	出力ラベル	出力仮想ポート#	出力実用ポート#
n	A	α	α'	K	$f(\alpha')$
n	A	β	β'	K	$f(\beta')$
n	A	γ	γ'	X	$f(\gamma')$
2	A	δ	δ'	B	$f(\delta')$
1	A	ϵ	ϵ'	B	$f(\epsilon')$
2	A	ζ	ζ'	H	$f(\zeta')$
1	A	η	η'	J	$f(\eta')$
5	A	σ	σ'	K	$f(\sigma')$
n	A	κ	κ'	X	$f(\kappa')$

【図 1 2】

図 1 2

ハッシュを用いた高速参照メカニズム

MPLS制御プレーンで維持されるラベル転送テーブル

	98	100	90	92	94	96	
タグ	到着ラベルのハッシュ値	入力実用ポート#	入力仮想ポート#	入力ラベル	出力ラベル	出力仮想ポート#	出力実用ポート#
○	$f(\alpha')$	n	A	α'	α''	K	$f(\alpha'')$
○	$f(\beta')$	n	A	β'	β''	K	$f(\beta'')$
○	$f(\gamma')$	n	A	γ'	γ''	X	$f(\gamma'')$
○	$f(\delta')$	2	A	δ'	δ''	B	$f(\delta'')$
○	$f(\epsilon')$	1	A	ϵ'	ϵ''	B	$f(\epsilon'')$
○	$f(\zeta')$	2	A	ζ'	ζ''	H	$f(\zeta'')$
○	$f(\eta')$	1	A	η'	η''	J	$f(\eta'')$
○	$f(\sigma')$	5	A	σ'	σ''	K	$f(\sigma'')$
○	$f(\kappa')$	n	A	κ'	κ''	X	$f(\kappa'')$

フロントページの続き

審査官 齋藤 浩兵

- (56)参考文献 特開2001-144804(JP,A)
国際公開第99/056432(WO,A1)
特表2002-513244(JP,A)
特開2000-013439(JP,A)
特開平11-145990(JP,A)

- (58)調査した分野(Int.Cl., DB名)
H04L 12/56