マルチメディア通信と分散処理　94－16
分散システム／インターネット運用技術　14－16
（１９９９．７．１６）

ラベルスイッチを用いた分散IXの設計

中川 郁夫
インテック・システム研究所

江崎 浩
東京大学
情報基盤センター

永見 健一
東芝 コンピュータ&
ネットワーク開発センター

概要

インターネットでは、複数のネットワーク間の相互接続を効率的に実現するため、インターネット・エクスチェンジ (Internet eXchange: IX) が構築されることが多くなってきている。インターネット・エクスチェンジの構築手法、および運用手法は多種多様であり、それぞれ特徴をいかした相互接続を実現しているが、同時に、いくつかの課題も残されている。
本稿では、ラベルスイッチ技術を応用したインターネット・エクスチェンジの設計について提案する。本稿で提案するインターネット・エクスチェンジでは、ポリシに基づいた経路制御を行うためにラベルを利用し、柔軟な経路制御を実現にするとともに、分散環境においても、効率的に相互接続環境を構築することを可能にする。また、ルートサーバ機能などを組み合わせることによりその管理・運用コストを低減させる仕組みについても言及する。

A Design of Distributed IX using Label Switching Technogoly

Ikuo Nakagawa
INTEC Systems Laboratory

Hiroshi Esaki
Information Technology Center
The University of Tokyo

Kenichi Nagami
Computer & Network
Development Center, Toshiba

Abstract

Recently, many Internet eXchanges are build and operated to interconnect multiple networks. While there are several architectures and technologies to implement an Internet eXchange, many problems are still unsolved.
In this paper, we propose a new architecture to implement an Internet eXchange using 'Label Swiching technology'. This architecture is one of efficient ways to implement flexible policy based routing, even in distributed area. We also suggest to implement Route Server mechanism to our architecture to reduce management or operational cost.

1　はじめに

インターネットでは、複数のネットワークを相互に接続する手段のひとつとしてインターネット・エクスチェンジ (Internet eXchange: IX)[1] が構築されるケースが多くなってきている。インターネット・エクスチェンジはひとつのネットワーク上で複数のネットワーク組織が相互に接続し、効率的にトラフィック交換を行なうことを目的としている。

一方、インターネット・エクスチェンジ技術にはまだ未解決の問題も多く残されている。例えば、いくつかのインターネット・エクスチェンジでは、ネットワーク間の相互接続に Ethernet や FDDI などの LAN 技術を利用しているが、これらの技術では通信の帯域が限界になりつつあること、あるいは、物理的に離れた場所での接続ができないことなどが問題になる。また、最近ではインターネット・エクスチェンジの相互接続に ATM 技術を利用することも多くなってきているが、この場合、個別に PVC の設定を行なう必要があるため、相互接続組

織が増えると、運用コストが膨大になる、などの問題が残されている。

本稿では、インターネット・エクスチェンジにおいて、ラベルスイッチ (Label Switching)[12] の技術を応用して、分散された環境でも効率的にトラフィックの交換が可能なインターネット・エクスチェンジの設計を行う。

ラベルスイッチでは、特に ATM などのスイッチ網において、カットスルー (Cut Through) 技術を用いることにより高速なパケット転送を可能にするとともに、トラフィック・エンジニアリングの技術によって、ポリシに基づいたトラフィック制御が可能である。また ATM などの PVC の設定のほとんどを自動化することもできるため、運用コストを抑えたまま、分散された環境での高速なトラフィック交換も可能になる。

本稿では、この技術を応用し、インターネット・エクスチェンジにおいて、広域分散環境で複数のネットワークを相互に接続するための手法を提案する。本手法では、インターネット・エクスチェンジで必要とされるポリシの制御をラベルスイッチにおけるラベルを用いて行い、同時に、カットスルーを用いた効率的なトラフィック交換を実現する。また、ルートサーバの技術を用いることにより、経路制御における管理・運用を容易にし、インターネット・エクスチェンジ全体の運用コストを最小限に抑えながら効率的なトラフィック交換を可能にする。

2 インターネット・エクスチェンジ

インターネットは、数万ともいわれるネットワークが相互に接続された「ネットワークのネットワーク」である。これらの相互接続において、複数のネットワーク間を効率的に相互接続するための技術としてインターネット・エクスチェンジ (Internet eXchange: IX) [1] の技術がある。図 1 は複数のネットワーク間で相互接続を行う場合の典型的な接続構成を示しているが、このように、個別に相互接続を行う場合には、各ネットワークは相互接続を行うネットワーク数に比例した相互接続コストを必要とする。

一方、インターネット・エクスチェンジ技術を用いる場合、図 2 に示されるように、ひとつのネットワークに複数のネットワークが接続し、集中的にトラフィックを交換することを可能にするため、各ネットワーク組織は相互接続のための回線コストや運用コストを抑えることが可能でる。

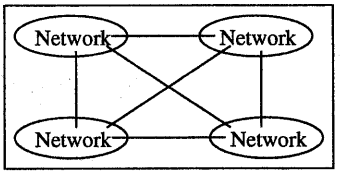

図 1: IX のない場合の相互接続

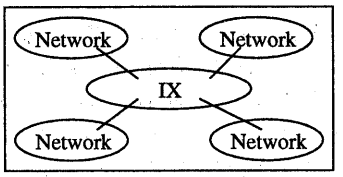

図 2: IX による相互接続

米国では CIX (Commertial Internet eXchange)[2], MAE(Metroporitan Area Network)[4], Chicago NAP (Network Access Point)[6] などに代表される、多くのインターネット・エクスチェンジが運用されている。また、国内では WIDE Project により実験・研究目的で構築・運用されている NSPIXP (Network Service Provider Internet eXchange Point)[8] や商用サービスの JPIX (JaPan Internet eXchange)[9], MEX (Media EXchange)[10] などが有名である。

2.1 インターネット・エクスチェンジにおけるポリシ

インターネット・エクスチェンジでは多数のネットワークが相互に接続されるが、経路情報の交換、すなわち、実際のトラフィックの交換は各ネットワークの運用ポリシに依存することが一般的になってきている。そのため、多くのインターネット・エクスチェンジでは全ネットワーク間で共通のポリシを強要されるマルチラテラル (Multi-Lateral) による相互接続協定よりも、個別にポリシを決定できるバイラテラル (Bi-Lateral) による相互接続協定を採用していることが多い。

2.2 Layer 3 のインターネット・エクスチェンジ

インターネット・エクスチェンジを構築する最も単純な方法は、IP データグラムを転送可能なルータ群で構成する (Layer 3 IX) 方法である。しかし、インターネット・エクスチェンジをルータ群で構成した場合、ルータのデータグラムの転送能力が全体のトラフィック交換性能に与える影響が大きく、性能上の問題が指摘されている。

また、前述の通り、近年ではインターネット・エクスチェンジにおいてバイラテラルによる相互接続協定を採用しているケースが増えてきているため、インターネット・エクスチェンジのネットワーク構成は、各ネットワークの運用ポリシを反映させることができなくてはならない。

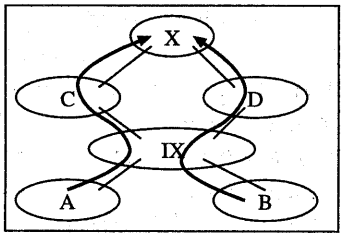

図 3: IX におけるポリシの反映

図 3 はインターネット・エクスチェンジの論理的な接続例を示している。例えば、インターネット・エクスチェンジに接続されるネットワーク A, B から X へデータグラムを送る場合にそれぞれ次のようなポリシが適用される場合、インターネット・エクスチェンジ内では、同じ宛先に対して異なる経路を選択する必要がある。

- A から X へは C を経由して送信する
- B から X へは D を経由して送信する

しかし、IP (Internet Protocol) においてルータがデータグラムの転送を行なう場合、その処理はルータ毎の Hop-by-Hop で宛先アドレスによって逐次処理されるため、上図において、インターネット・エクスチェンジがルータ群で構成されている場合 (Layer 3 IX) は、送信元が A, B のどちらかによって途中の経路を選択するということはできない。

このため、近年のインターネット・エクスチェンジではデータリンク層による接続を行なう Layer 2 IX を構築し、各ネットワーク間のトラフィック交換はデータリンク層のみで実現することにより、個別のポリシを反映させるケースがほとんどである。

2.3 LAN 技術を応用したインターネット・エクスチェンジ

これまでは、インターネット・エクスチェンジをデータリンク層 (Layer 2) で実現する手段として LAN 技術を利用するケースが多かった。例えば NSPIXP-2 や JPIX では、FDDI スイッチを用いて、ひとつのビル内で多数のネットワークのルータを相互に接続しており、NSPIXP-3 では Ethernet スイッチを用いて相互接続を行っている。

しかし、Ethernet や FDDI などの LAN 技術を用いた相互接続を行う場合、物理的に離れた場所での接続はできないため、各ネットワーク組織は、相互接続を行う場所へルータを持ち込み、物理的には一箇所、もしくは非常に近い場所での接続を強制される。

また、インターネットのバックボーン速度が LAN における通信速度を越えた現在においては、LAN 技術の通信速度はインターネット・エクスチェンジでのトラフィック交換には不適切だとの指摘もある。一部では Gigabit Ethernet などの利用実験を行っているケースもあるが、将来的には LAN 技術では帯域不足などの問題がより深刻になると思われる。

2.4 ATM によるインターネット・エクスチェンジ

近年では Chicago NAP や MAE-ATM[5] のように ATM(Asynchronous Transfer Mode) による分散型のインターネット・エクスチェンジが構築されるケースも増えてきている。

ATM を用いた場合、Ethernet や FDDI などの LAN 技術を用いる場合とは異なり、容易に広域分散環境においてインターネット・エクスチェンジを構成することが可能であり、データ転送速度も Ethernet や FDDI に比較して容易に高速化が可能であることなどから、拡張性の高いインターネット・エクスチェンジが構築できると期待されている。

ただし、ATM 網上でインターネット・エクスチェンジを構成する場合、相互接続を行なうネットワーク間で

PVC (permanent vertual circuit) を設定し、個別に経路情報の交換を行なう方法が一般的である。

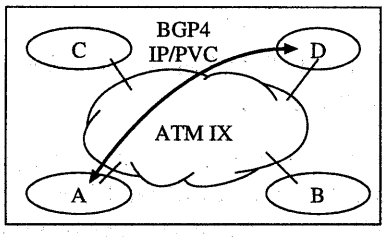

図 4: ATM による IX

この場合、各ネットワーク間の接続はデータリンク層で個別に確立・運用されるため、ネットワーク間で個別に経路制御を行なうことにより、それぞれのポリシを反映させることは容易であるが、各ネットワークでは相互接続ネットワーク数に比例した PVC の設定が必要であり、大規模なインターネット・エクスチェンジでは、管理・運用コストも膨大になる。

2.5 インターネット・エクスチェンジにおける経路制御とルートサーバ

多くのインターネット・エクスチェンジでは接続される個々のネットワーク間で個別のポリシを反映させるため、BGP4 (Border Gateway Protocol version 4)[11] による経路制御を行っている。この場合、ひとつのネットワークが他のネットワークとトラフィックの交換を行なうためには、それらのすべてのネットワークと BGP4 による経路情報の交換を行なう必要がある。接続組織の数を N とした場合、ひとつのネットワーク当たりの運用コストは $O(N)$ 、また、インターネット・エクスチェンジ全体での運用コストは $O(N^2)$ に相当することになり、運用上の負荷は膨大なものになる。また、接続組織数が多くなってきた場合、BGP4 を処理する各ルータの負荷も増大することになる。

一部のインターネット・エクスチェンジはこうした運用コストやルータの負荷の問題を解決するため、ルートサーバ (Router Server)[15] を導入している。ルートサーバは、Ethernet や FDDI などの共有型データリンク層のネットワークで構成されるインターネット・エクスチェンジ上に設置され、そこへ接続されるネットワークとの経路制御を一括して行なう。この際、経路情報データベース (Routing Regisotry[14]) を参照することにより各ネットワークの運用ポリシを反映し、運用負荷、あるいは各ルータの経路情報の処理負荷を最小限に抑えながら経路制御を行うことが可能になる。

ただし、ATM によるインターネット・エクスチェンジでは各ネットワーク間は PVC による個別の接続を行っており、全体で共通したネットワークが存在しないため、ルートサーバを用いて統括的な経路制御を行うことは難しい。

3 ラベルスイッチ技術を用いたインターネット・エクスチェンジ

本節ではラベルスイッチ技術を応用した、分散型のインターネット・エクスチェンジの構築手法を提案する。

本稿で提案する手法では、各ネットワークのポリシをラベルに対応付けることによりデータグラムの転送にポリシを反映させる。また、ラベルスイッチ技術により相互接続の設定を自動化し、ATM 上で構築されたインターネット・エクスチェンジであっても、その運用コストを最小限に抑えることが可能になる。

なお、本稿では、特に ATM によるインターネット・エクスチェンジ上でラベルスイッチ技術を応用する手法について述べるが、ラベルスイッチ技術は ATM 以外のデータリンク層技術でも利用が可能であり、将来的には POS や WDM などの次世代データリンク技術へも容易に応用が可能である。

3.1 ラベルスイッチを用いた IX モデル

図 5 はラベルスイッチを用いたインターネット・エクスチェンジの構成図を示している。Edge は各ネットワーク組織の管理下におかれ、インターネット・エクスチェンジ内の LSR (Label Switching Router) Core と接続される。

Edge-Core 間の接続に ATM 網を利用する場合、この間には PVC による直接接続を行い、LDP(Label Distribution Protocol)[13] 等によりラベルに関する情報を交換することになる。

なお、各ネットワーク間のトラフィック交換、すなわち Edge-Edge 間の通信においては、ラベルスイッチにおけるカットスルー (Cut Through) 技術によって高速

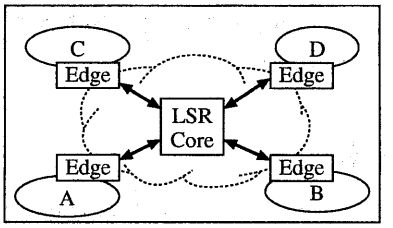

図 5: ラベルスイッチを用いた IX

なトラフィック交換を行う。

3.2 経路情報の交換

ATM PVC によるインターネット・エクスチェンジでは、各ネットワーク間の相互接続は個別に行われるため、それぞれにどのようなアドレスが付与され、また、各リンクが現在利用可能かどうかをインターネット・エクスチェンジ側で把握することは難しい。

一方、本稿で提案されるインターネット・エクスチェンジでは各ネットワークの Edge ルータはインターネット・エクスチェンジ内の LSR Core と IP 層で接続されるため、各 Edge のアドレスや、リンクの接続性情報などを LSR Core で一括して管理が可能である。本モデルでは、これらの特徴を活かして、LSR Core にルートサーバ機能を実装し、Edge-Core 間では BGP4 および LDP による制御を行う。

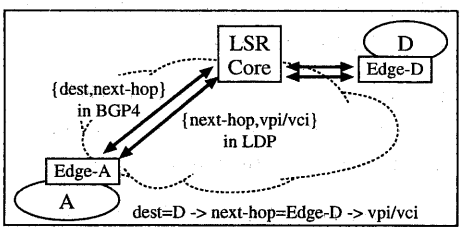

図 6: ラベルスイッチを用いたデータグラムの転送

BGP4 による経路情報には宛先アドレスと、それに対応する属性が含まれる。特に属性には Next-Hop アドレスが定義されており、Edge ルータは、この Next-Hop アドレスを用いて、その宛先へ向けたデータグラムを転送する、対向の Edge を決定する。

LDP によるラベルの対応情報は Next-Hop アドレスと、それに対応するラベル (ATM の場合 VPI/VCI) の対で表現される。実際にデータグラムを転送する場合には、宛先アドレスから Next-Hop を検索し、さらに Next-Hop に対応するラベルを用いてインターネット・エクスチェンジ上でカットスルーを行なうことになる。

本稿で提案するインターネット・エクスチェンジでは、このように、ラベルスイッチ技術により BGP4 で取得した Next-Hop 情報をラベルに対応付けることにより、トラフィック交換におけるポリシの反映を可能にしている。

3.3 ラベルスイッチを用いたインターネット・エクスチェンジの特徴

本稿では、インターネット・エクスチェンジにおいてラベルスイッチを適用する手法について提案したが、本手法は次のような特徴を持つ。

1. 各 Edge において必要な設定は Core との間でのラベルスイッチの設定、および BGP4 の設定のみであり、管理・運用コストを低減させることができる。
2. 実トラフィックはラベルスイッチ技術により Edge-Edge 間で直接転送するため、カットスルーが可能であり、高速なトラフィック交換が可能である。
3. ATM, POS, WDM などの技術に対応可能であり、広域分散環境での相互接続、あるいは超高速化時代への対応も可能である。

3.4 既存のインターネット・エクスチェンジ技術との比較

本節では、ラベルスイッチ技術を応用したインターネット・エクスチェンジの利点を既存のインターネット・エクスチェンジ技術と比較する。

1. Layer 3 IX との比較。
   Layer 3 IX では、データグラムの転送はインターネット・エクスチェンジ内のルータに依存し、ポリシを反映させることができないため、ネットワーク毎にポリシが異なる場合などには適用できない。本稿で提案するインターネット・エクスチェンジではポリシをラベルに対応させることにより、基本的な接続や設定は Layer 3 と同等でありながら、各ネットワークのポリシを反映させたデータグラムの転送

が可能である。
また、ラベルスイッチを用いたインターネット・エクスチェンジでは、実トラフィックはカットスルー技術により極めて高速なパケット転送が可能である。

2. Ethernet, FDDI による IX との比較。
LAN 技術を用いた場合、分散環境でインターネット・エクスチェンジを構成することが難しい。そのため、各ネットワークから接続拠点にルータを持ち込む必要があり、各ネットワークの運用負荷はさらに増大する。本稿によるインターネット・エクスチェンジは広域分散環境での相互接続を前提としており、各ネットワークの Edge ルータはそれぞれのネットワーク内に設置することも可能である。
また、LAN 技術を用いた場合、Ethernet で 10Mbps～1Gbps、FDDI で 100Mbps 程度が通信速度の上限になるが、ラベルスイッチ技術を応用する場合、ATM で 155Mbps～2.4Gbps、POS で 10Gbps 程度など、次世代の超高速通信技術でも容易に応用が可能である。

3. 既存の ATM IX との比較。
ATM PVC によるインターネット・エクスチェンジでは PVC の設定は接続する相手ごとに設定する必要があるため、その設定、管理、運用コストは、各ネットワーク当たり $O(N)$ になる。本稿におけるインターネット・エクスチェンジの場合、PVC の設定はラベルスイッチにより、経路制御の設定はルートサーバ機能により自動化されるため、個別の設定は 実質的にこれらのコストは $O(1)$ で抑えられる。

4 まとめ

本稿ではラベルスイッチ技術を応用したインターネット・エクスチェンジの構築手法を提案した。本提案によるインターネット・エクスチェンジは次のような特徴を持つ。

- 各ネットワーク組織の管理・運用コストが最小限に抑えられる。
- 広域分散環境において相互接続が可能である。
- カットスルーによるデータグラムの転送により、高速なトラフィック交換が可能である。
- ATM、POS や WDM などの技術へも応用可能である。

今後は、実証実験を通して性能評価等を行なうとともに、複数ルータの実装に関する相互接続試験などを行なうことを予定している。

参考文献

[1] Bill Manning: "Exchange Point Information", http://www.ep.net/

[2] CIX: "Commertial Internet eXchange", http://www.cix.org/

[3] Digital: "Digital Internet eXchange", http://www.ix.digital.com/

[4] "MCI WorldCom MAE Information", http://www.mae.net/

[5] "MAE-ATM Information", http://www.mae.net/atm/

[6] "Chicago NAP", http://nap.aads.net/main.html

[7] "Welcome to Star Tap", http://www.startap.net/

[8] WIDE Project, "WIDE/NSPIXP Home Page", http://xroads.sfc.wide.ad.jp/NSPIXP/

[9] "JaPan Internet eXchange", http://www.jpix.ad.jp/

[10] "Media EXchange", http://www.mex.ad.jp/

[11] Y. Rekhter, T. Li: "A Border Gateway Protocol 4", RFC1771, Mar. 1995

[12] E. Rosen, A. Viswanathan, R. Callon: "Multiprotocol Label Switching Architecture", IETF Internet-Draft, April, 1999.

[13] Paul Doolan, Nancy Feldman, Andre Fredette, Bob Thomas: "LDP Specification", IETF Internet-Draft, June, 1999.

[14] D. Estrin, J. Postel, Y. Rekhter: "Routing Arbiter Architecture", June 1994

[15] R. Govindan, C. Alaettinoglu, K. Varadhan, D. Estrin: "Route Servers for Inter-Domain Routing"