10

研究用MPLSシステムの開発と運用実験

宇夫陽次朗* 宇多 仁** 小柏伸夫***

インターネットはグローバルなコンピュータネットワークの社会基盤として確固たる位置を占めるようになり，利用者数および利用者層ともに大きく拡大した．しかし，その成功ゆえにネットワークが提供する機能に対する要求項目は増大し続けており，インターネットのアーキテクチャ自体を改変する要因となっている．

このような状況の中でネットワークの制御機構と転送機構をより明確に分離するパラダイムの有用性が認知されつつある．このパラダイムは現在のアーキテクチャではひとまとめとなっている転送およびその制御機構を分離することで，より柔軟なネットワーク制御の実現を目指している．

本稿ではその基本技術として注目されているラベルスイッチング技術について解説し，我々が開発および実装しているMPLS研究用実験環境AYAMEについて説明する．また2000年9月に開催されたWIDE合宿におけるMPLS実験ネットワークによって得られた経験を報告する．

インターネットの変遷と発展

インターネットが設計されてからすでに30年近い年月が経過した．その間にインターネットは広域に配備され，その利用者数は爆発的に増大した．この大きな成功の結果，インターネットはコンピュータネットワークの社会基盤となり，さらに多様な要求を実現する圧力にさらされるようになった．

現在のインターネットアーキテクチャは接続性を最も重視した設計になっている．それ以外のサービスに関してはあえて考慮しないことで限られたリソースを有効に利用し，より拡大しやすいネットワークを目指したからだ．そのため，接続性を提供するために最適かつ最低限の構成となっており，通信の根本となる部分の拡張性はあまり考慮されていない．

計算機技術および伝送技術の発展に伴い，利用者はそれ以上の何か，つまり通信品質の向上に目を向けるようになっている．このような既存アーキテクチャで想定していなかった機構をネットワーク上に導入するためには，必然的にインターネットのアーキテクチャ自体から見直す必要がある．

その流れの1つとして，ネットワークのレイヤリングを見直す動きがある．特に『転送／制御分離モデル』は今まで第3層によって提供されてきたパケット転送機能と経路制御に代表されるネットワーク制御機能を分離して扱うことで，より柔軟なネットワーク制御を狙う試みである．マルチプロトコルラベルスイッチング（MPLS）技術はこのパラダイムを実現するための基本技術として注目されている．第3層経路制御によって決定される経路以外を扱うためのトラフィックエンジニアリング用技術といった既存の制御フレームワークの制約を受けない技術として現在でも広く用いられるようになってきている．

MPLS技術はその有用性ゆえにか，技術開発と製品化の波が同時に出現したため，研究プラットフォームとなる実装が出現する前に，商用化のフェーズに移りつつある．しかし，MPLS技術はまだ開発途上であり，その可能性はまだまだ大きいと考えられる．そのため，我々は，ラベル転送技術一般の研究プラットフォームの必要性を強く感じ，研究向け実装の開発を進めている．

本稿ではMPLS技術一般に関する説明を行ったあと，MPLSの研究的側面に触れる．その上で，本連載で取り上げているWIDE合宿でのMPLSネットワーク運用実験から得られた経験を説明していく．

マルチプロトコルラベルスイッチング技術

ラベルスイッチングとは『ラベル』と呼ばれる識別子を利用してデータを転送する技術の総称で，過去には東芝のCSR（Cell-Switch Router），CiscoのTag Switching，IpsilonのIP Swithing，IBMのARISなどが提案されてきた．マルチプロトコルラベルスイッチング技術（以下MPLS技術）[3]はインターネットの標準化策定機関であるIETF（Internet Engineering Task Force）によって標準化が進められているラベルスイッチング技術である．

* 北陸先端科学技術大学院大学 情報科学研究科
yuo@jaist.ac.jp
** 北陸先端科学技術大学院大学 情報科学研究科
zin@jaist.ac.jp
*** 北陸先端科学技術大学院大学 情報科学研究科
n-ogashi@jaist.ac.jp

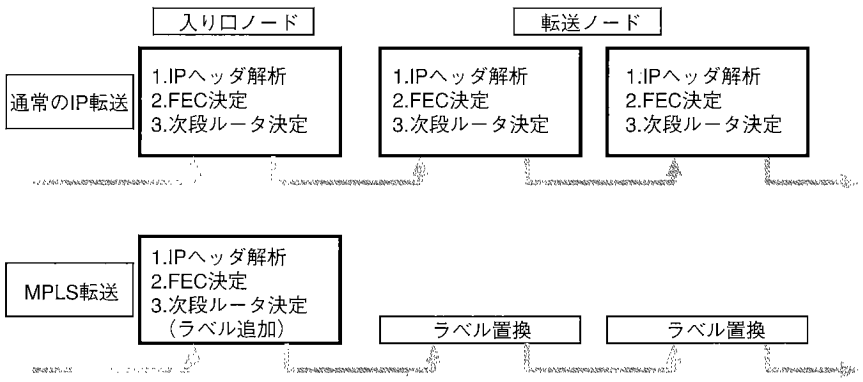

図-1　一般のL3転送とラベルスイッチングの比較

■MPLSネットワークの動作

一般に現在のインターネットはパケットは『第3層の転送処理』によって始点から終点に伝送される．この転送処理は各ルータで独立して行われている．各ルータではパケットが到達するたびに (1) パケットヘッダの解析，(2) 転送方法の確定，(3) 次転送先の決定と転送，といった一連の処理を実行し，パケットを次転送先に送り出している．パケットが最終点に到達するまでこの手続きが繰り返される (**図-1** (上))．

それに対して，MPLSでは第3層処理，すなわちIP層の処理を大部分で省略するようになっている．MPLS転送はルータ単体ではなく一連のMPLS対応ルータ群 (LSR: Label Switch Router) から構成されるMPLS-cloud (MPLS雲) が単位であり，MPLS雲の出入口LSRと途中LSRでの処理は異なる．入口LSRでは通常の第3層ルータと同様に上位層での情報を解析し，パケットをどのように転送すべきか決定する．それからパケットにラベルを付加して転送する．途中LSRはラベルだけを解釈して次々に転送を行い，MPLS雲の出口である出口LSRではラベル取り外してから隣接ネットワークへ転送する (**図-1** (下))．

■FECとMPLSシグナリング

入口LSRで解析された上位層情報 (たとえばIPヘッダ) はFEC (同一転送クラス: Forwarding Equivalent Class) と対応付けられる．FECは『同等の転送を行うパケット群を示すクラス』である．MPLS雲内ではMPLSのシグナリング機構 (ラベル配布プロトコル等) を用いてFECに関する情報を共有している．この機構を通じてFECはラベルと対応付けられており，途中LSRはラベルを処理するだけでそのパケットの転送に関する情報を得られる．このようにラベルを用いてFECを参照させる理由は主に2つある．処理の単純化と拡張性である．

- 単純化：
  ラベルの値を短い固定長としたことによって各LSRでの処理を簡略化できる (完全マッチングによる経路テーブルの検索)．長い値を扱うためにはハードウェアなどの回路量が増大し高速化のコストがかかるとともに，可変長のラベルだと処理が複雑となり (ベストマッチによる経路テーブルの検索) 高速化への技術的なハードルの1つとなる．
  一方，短いラベルだと小さな名前空間しか表現できないという問題点が発生するが，MPLSのアーキテクチャではこの問題をラベルをリンクローカルな識別子として扱い，ラベルの書き換え (スワッピング) を行うことで解決している．

- 拡張性：
  FECは抽象的な概念であり，その内容をすべて表現する情報をパケットごとに与えることは困難である．そのため，パケット流をラベルに写像し，ラベル情報を各パケットに付加し，このラベル情報を用いたパケットの転送を行う．このようにFECとFECへのポインタとしてのラベルを分離したことによって，MPLSはネットワーク制御に関する大きな拡張性と柔軟性を持つことになった．MPLS網内で何らかの処理を示すFECを共有することで，MPLS網内のパケット挙動 (取り扱いのポリシ) を任意に指定できる可能性が出たからである．

■制御と転送の分離

このようにMPLSは単純な転送処理と制御処理を分離するアーキテクチャを持つ (**図-2**)．既存のレイヤ構造では第3層以下は転送と制御が一体化されたブラックボックスとなっており，パケットはあらかじめ決められた形式 (経路制御および転送機構) でしか扱うことができなかった．しかし，MPLSでは第3層が持っているこれらの機能を分解して扱えるため，IP層での制御機構以外の制御機構を用いてネットワークを制御できる．これを『転送と制御の分離パラダイム』と呼び，最近ではMPLSの提供する性質の中で最も着目されている機能の1つである．

■ラベル配布プロトコル

複数のLSR間でFECとラベルのマッピングを行うプロトコルを一般にラベル配布プロトコルと呼ぶ．現時点では，第3層の経路制御情報をMPLS空間に写像するためのラベルマッピングを配布するLDP[4]，経路制約に基づいたラベルマッピングを配布するCR-LDP[5]などがある．他にも各種経路制御プロトコルで，配布する経路に対してラベルを相乗りして配布する拡張などが提案されている．ラベル配布プロトコルはMPLSの制御に大きな影響を与える部分であり，制御と配送の分離パラダイムにとっても重要な要素技術である．

MPLS研究／実験用環境

今まで説明したようにMPLSは，今までのネットワークで触れることができなかった部分に手を入れるための道具として現時点で最も利用しやすい技術である．現時点でも多くのベンダがMPLS対応をうたう製品を市場に投入しつつある．これらはMPLSが提供する利点を活用し実運用を可能としているという点で非常に評価されるべきだが，『MPLSという技術』自体を研究しようとした場合には役に立つとは言いがたい．研究を遂行するためには，製品として提供されている機構を組み合わせるだけではなく，オリジナルな要素の作成が強く望まれるからだ．

残念なことに，MPLS技術はその発端が商用サイドから発生したことや，立ち上がりが速かったことから，既存のネットワークコードのような研究者の自由になるコードがほとんど存在しないのが現状であった．そのため，北陸先端科学技術大学院大学ではMPLSの研究／実験用環境であるAYAME[1),2)]の実装を行っている．

■MPLS実験環境

AYAMEはラベルスイッチング技術を利用した研究を行うことを目的として開発が続けられているMPLS実験環境である．研究支援を主な目的とする実装は『改変および拡張が容易であるべきだ』との考えのもと，AYAMEでは徹底したモジュール構造を採用して拡張性を向上させている．また，BSD系UNIXであるNetBSD[6]をベースとしBSD (AS-IS) ライセンスを採用することで改変／再配布の自由を保証している．

AYAMEは大きく3つの部分に分かれる．主にカーネル部分を拡張して作成されているLSR実装，ユーザランド上で動作するラベル配布プロトコル実装，そしていくつかのサポートプログラム群である（**図-3**）．

MPLSは制御と転送を分離したアーキテクチャを持っていることはすでに説明した．そのためMPLSを対象とした研究は転送部分を対象としたものと制御部分を対象としたものに分類できる．AYAMEはこのどちらの研究のプラットフォームとしても利用できることを目指している．以下で，AYAMEのLSR実装およびラベル配布プロトコル実装について説明する．

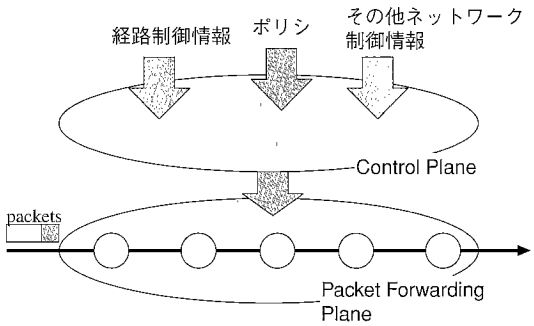

図-2　制御構造
(Control/forwarding separation)

LSR実装

AYAMEのLSR実装はBSDのネットワークスタックを拡張することで実現している．設計にあたって，最も考慮した部分はコードの見通しの良さである．MPLS LSRを実現するために必要な機能を分類し，それぞれの要素ごとにモジュール化してある．そのため必要な部分だけの変更や置換が容易な構造となっている．

特に既存のBSDのネットワークコードのセマンティクスの変更に関しては非常に多くの検討を行った．BSD系のコードをベースとする利点として，すでに多くの研究者が内容を把握しているという点が挙げられるが，MPLS拡張によって既存部分を大幅に変更してしまうとこの利点を損なうことになる．BSDのネットワークコードはもともと第2層と第3層の実装がIPのレイヤ構造に基づいた形式で分離されている．AYAMEではこの構造に基づいてMPLS機能を挿入することで見通しの良い拡張を実現した．

ラベル配布プロトコル実装

AYAMEでは，ラベル配布プロトコル実装としてLDP[4]およびCR-LDP[5]を提供している．MPLSネットワークおよびLSRの挙動制御の大部分はラベル配布プロトコルの制御下にあることを考慮すると，今後も実験や研究を目的とした新規ラベル配布プロトコルの提案や，既存のラベル配布プロトコルの拡張は重要である．

AYAMEで配布されているラベル配布実装は，ラベル配布のコアプロトコル部分とラベル配布機構間の通信のための要素を明確に分離した実装となっている．このため，コアプロトコル部分を拡張するだけで，容易に新しいラベル配布プロトコルを実現できるようになっている．また，研究を行ううえで単一のLSR上で複数のラベル配布プロトコルが動作するような場面があり得るが，この場合LSRで複数のラベル制御セマンティクスを扱う必要がある．そこでAYAMEのLSRにはラベル空間の分割およびそれらの間の整合性を維持する機構が含まれている．

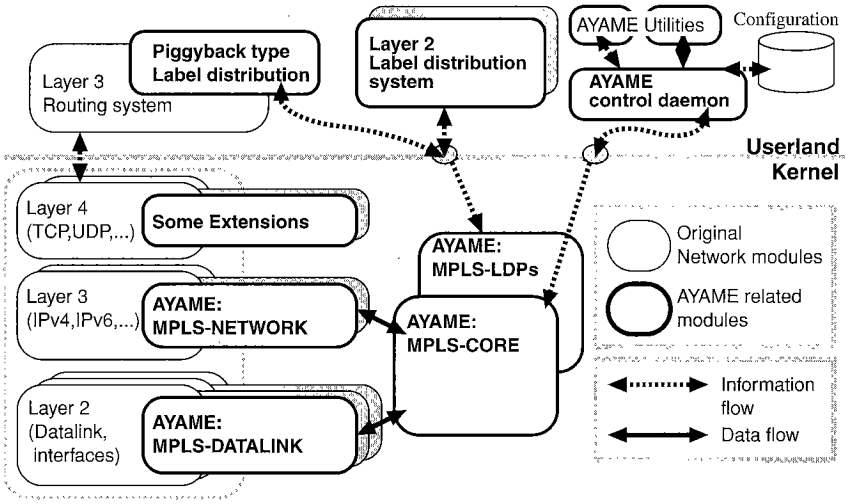

図-3　AYAMEの全体構造

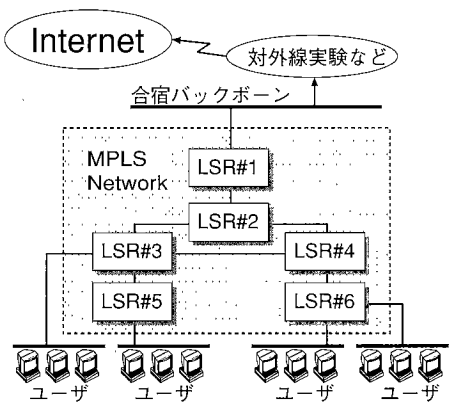

図-4　合宿ネットワーク（MPLS部分）

WIDE合宿ネットワークでの運用実験

本連載で何回か紹介しているように，WIDE Projectでは年2回（3月／9月）に4日間におよぶ合宿形式の研究会を開催している．研究会では，一時的なネットワークを構築するとともにその上でさまざまなネットワーク実験を行う．このネットワークを『合宿ネットワーク』と呼ぶ．2000年9月の研究会は長野県茅野市白樺湖で参加者250人程度規模で開催された．この研究会で，我々は先ほど紹介したMPLS実験環境AYAMEを用いたMPLSネットワークの動作実証実験を行った．

■合宿ネットワークの構成

今回の合宿ネットワークではネットワークのユーザセグメント部分をほぼ全部MPLSネットワークの下に配置した．このように多くのユーザからのトラフィックを扱うことで，実ネットワークに近い環境下でのMPLS実装の動作検証を行うことが目的である．

合宿ネットワークのMPLS部分のトポロジを**図-4**に示す．図から分かるように，MPLSネットワークと対外接続ネットワークの間にはNAT box（ネットワークアドレス変換装置）が配置されており，外部との接続部分でプライベートアドレスへ変換される構成になっている．MPLS網を構成するすべてのLSRはAYAMEであり，LSR間の接続はEthernet（100BT）を用いた．MPLS網の制御には，

- 第3層の経路制御プロトコルはOSPF
- ラベル配布プロトコルはLDP

を用いている．

AYAME環境を用いたMPLSネットワークは，このような実ネットワークに近いトラフィックの下でも『単体では』おおむね良好に動作することが確認された．しかし，他技術との組合せ，もしくは既存のネットワーク機器の制約などからいくつかの問題点も明らかになった．

■MPLSネットワーク運用から得られた経験

合宿ネットワークではMPLS技術の導入をきっかけにいくつかのネットワーク障害が発生した．MPLSではパ

ケットにラベルを付加するために最大MTU長が見かけ上減少してしまう．そのため潜在的に存在していた問題が表面化することになった．

問題を分類すると，

- 大きなパケットを扱えないデバイスの存在
- NAT実装の問題の表面化
- ICMPパケットフィルタリングと経路MTU探索

の3種類が確認された．これらの問題はそれぞれ個別ではあるが，合宿ネットワーク中では同時に出現したためトラブルシューティングが困難であった．それぞれの問題の概略を以下に示す．

大きなパケットを扱えないデバイスの存在

一般にイーサネットの最大MTUは歴史的に1,500byte程度と規定されていた．最近IEEE802系列の拡張によってこの値は拡張される傾向にあるが，古いデバイスなどでは1,500byteに固定されているものも散逸される．今回LSRで利用したネットワークインタフェースの一部にこのようなデバイスが含まれていた．この問題を回避するためにネットワークスタックの設定でMTUサイズを小さめに設定して運用した．

NAT実装の問題の表面化

前述の問題から，MPLS網では外部ネットワークで設定より小さなMTUで運用されていた．そのため，パケットがMPLS網への転送される際に断片化が発生した．今回の実験網ではNAT boxが断片化の境界点であったため，NAT実装の問題が表面化し接続性が維持できなくなった．

ICMPパケットフィルタリングと経路MTU探索

現状のインターネットでは多くのリンクMTUは1,500byte以上である．今回の合宿ネットワークで1,500byte未満のMTUでMPLSネットワーク部分を運用したことによって，一部の外部サイトとの接続ができないという問題点が発生した．本来はこのような問題は発生しないはずだが，最近のインターネットではサイトの設定によっては1,500byte未満のMTUを持つネットワークと通信できない場合があることが観測された．原因については『コラム：ファイアウォールと経路MTU探索』で説明する．

■非対称ラベル配布の運用

合宿ネットワークではMTUに関するさまざまな問題が発生したため，急きょこの問題を回避する目的で『非対称ラベル配布機構』を実装し運用した．これはLDPプロトコルをベースとしたラベル配布機構であるが，ピアリングした際にラベル情報を流す方法を制限し，LSRに対して単方向のラベル-FECマッピング情報しか与えない．この結果，同一のリンクであってもパケットの流れる方向によって，第3層転送とMPLS転送を切り替えて運用することが可能となった．このような状況はきわめて特殊な例ではあるが，AYAMEの実装の柔軟性を示す1つの例ともなった．

今後の課題と現在の取組み

MPLS技術を研究するためのプラットフォームの実装はおおむね終了している．今後はこのプラットフォームを用いてさまざまな研究を行っていく予定である．

■AYAMEの拡張

- 第2層インタフェース部分の拡張
  現在のAYAMEがサポートしているネットワークインタフェースはEthernetである．しかし，異なるMPLS実装との相互接続およびさらなる高速化を実現するために，PPPやATMリンクサポートが重要だと考えられる．
- ハードウェアスイッチの利用
  ATMのような第2層技術の一部にはハードウェアでパケットをスイッチングできる機器が存在する．AYAMEのソフトウェアによるパケットスイッチング機構とこれらの機構を協調的に動作させることでシステム全体の性能を向上させることを計画している．

■MPLS技術の拡張

MPLS技術は広く利用されるようになったとはいえ，まだまだ検討していかなければならない部分は多い．たとえば，マルチキャスト型配送の実現，IPv4以外の第3層プロトコルの利用，より多様な第2層技術の利用といった課題が存在する．

- マルチキャスト対応
  AYAMEプロジェクト[7]ではMPLSマルチキャストを重点課題として研究を進めている．MPLSマルチキャストは現在まだ標準化の途中であり，まだまだ多くの検討が必要だと考えられている．MPLSのアーキテクチャが転送と制御を分離するようになっているため，MPLS自体を研究対象とする場合にもそれらの分類が適用される．マルチキャスト型配送を実現するためには，1) LSRの転送機構を拡張，2) マルチキャスト網を構築するためのラベル配布機構の実現，の2点を考慮しなければならない．
- COPS over MPLS/Diffservの実現
  MPLSは汎用のデータトランスポートとしての側面もあることから，Policyフレームワークのような既存のIP制御層よりも複雑／高性能のシグナリングトランスポートの利用も広く議論されている．
  すでに連載中で説明されているようにWIDEプロジェクトではPolicyフレームワークにおけるQoSシグナリングに関するアクティビティ（Moon Bear Project）[8]が存在する．現在，これらの技術とMPLSを統合したネットワーク構築の共同研究を行っている．

コラム: ファイアウォールと経路MTU探索

ファイヤウォール技術はインターネットに接続する際に必須の技術となっている．パケットフィルタリングはその中でも広く用いられている技術であるが，その設定によって想定しないような副作用を引き起こす場合がある．ここでは，パケットフィルタリングと経路MTU探索に関係について述べる．

インターネットはさまざまな種類の物理リンクで構成されている．これらのリンクは特性に応じて最大転送単位（MTU）が規定されている．原則としてMTUを超えるパケットはそのままでは転送できない．この問題に対処するために以下のいずれかが用いられる．

- 経路MTU探索（PMTUD）：事前に経路上のMTUを調査してパケットの最大長を制限する．
- 断片化（フラグメント）：MTUが異なるリンク間に存在するノードがパケットを断片化する．

経路MTU探索ではパケットにはDF（断片化禁止フラグ）が付加され，経路中に断片化しなければ通過できないリンクが存在する場合にはそのパケットが破棄されICMPエラーを送信することになっている．しかし，適切な設定がされていないパケットフィルタ型ファイヤウォールでは，すべてのICMPパケットを破棄してしまうものがある．そのため，経路MTU探索のMTU調整メカニズムがうまく動作せずパケットの到達性が阻害されてしまう．

現在のインターネットでは大多数のリンクのMTU値が1,500バイト以上であるために，Ethernet（MTU: 1,500byte）に接続されたホストでは問題が表面化しにくく，自サイトがこのような問題を持っていることに気が付いていない場合がほとんどであろう．合宿での表面化を機会にいくつかの例を調査したところ，非常に多くのアクセスがある大手サイトでも潜在的に問題を抱えている場合が少なくなかった．最近では，インターネットへの接続に用いられるリンクが多様化しつつあるが，その中にはPPPoE（RFC2516）などのようなMTUが小さめなリンクもあり，今までは見過ごされやすかったこの問題が表面化する日もそう遠くはないだろう．

この問題を回避するためには，

- パケットフィルタでPMTUDにかかわるICMPパケットの通過を許可する．
- パケットフィルタの内側のホストでのPMTUDの利用を抑制し，途中ルータでのフラグメントを可能とする．

のいずれかが必要だ．もしファイアウォールの設定をした際に，単純にICMPを破棄するような設定にした覚えのある方は今一度見直してみるべきだろう．

- ラベル配布プロトコルの開発
  MPLSの転送部分はラベル配布プロトコルによって制御されている．したがって，現在とは異なる目的で転送部分を扱う場合には，それに応じたラベル配布プロトコルが必要となる場合が考えられる．そこで，ラベル配布プロトコルを容易に実現可能なツールキットを作成中である．

■MPLSを利用したネットワーク技術の開発

MPLSを用いることで既存のネットワークアーキテクチャでは実現が難しかった形式のネットワークを構築できると考えられている．そこで，MPLS技術を前提としたネットワークアーキテクチャという立場での次世代ネットワークに関する研究を行っている．

おわりに

転送機構を簡略化するとの目的で開発されたMPLSは，転送と制御の分離パラダイムを作り出すに至り，現在では次世代のネットワークを構築するうえでの要素技術とみなされるようになってきている．現在でも多くの分野でMPLSは利用されつつありそのための製品も少なくない．しかし，研究素材としてMPLSを扱うときに，そのためのプラットフォームが欠如していることは非常に問題であろう．本稿では『MPLS技術』およびその周辺を研究対象とするための実験プラットフォームの設計および実ネットワークへの適用について考察した．現時点ではMPLS環境を構築し終わったという状態であるが，今後さらに研究としてのMPLS技術を扱っていく予定である．

参考文献

1) Uo, Y., Uda, S., Ogashiwa, N., Ohta, S. and Shinoda, Y.: AYAME: A Design and Implementation of the CoScapable MPLS Layer for BSD Network Stack, INET2000 (2000).
2) 宇多　仁，宇夫陽次朗，篠田陽一：MPLS実装AYAMEにおけるパケット転送機構の設計および実装，DPS2000 (2000).
3) Eric, C.R., Arun, V. and Ross, C.: Multiprotocol Label Switching Architecture, RFC3031 (2001).
4) Loa, A., Paul, D., Nancy, F., Andre, F. and Bob, T.: LDP Specification, RFC3036 (2001).
5) Andersson, L., Fredette, A., Jamoussi, B., Callon, R., Doolan, P., Feldman, N., Gray, E., Halpern, J., Heinanen, J., Kilty, T.E., Malis, A.G., Girish, M., Sundell, K., Vaananen, P., Worster, T., Wu, L. and Dantu, R.: Constraint-Based LSP Setup using LDP, IETF Work in Progress (2000).
6) NetBSD Project, http://www.netbsd.org/.
7) AYAME Project, http://www.ayame.org/.
8) Moon Bear Project, http://www.moon-bear.net/.

（平成13年2月22日受付）