MPLS 実装 AYAME における パケット転送機構の設計および実装

宇多 仁　　宇夫 陽次朗　　篠田 陽一

北陸先端科学技術大学院大学 情報科学研究科

MPLS は 次世代インターネットにおけるパケット配送機構の高機能化の実現が可能な基礎技術の 1 つとして盛んに研究されている。我々は、MPLS に関する研究開発において評価実装等を行うための基礎として用いることが可能な MPLS 実装として、AYAME を設計・実装した。

本稿では、特に AYAME におけるパケット転送機能の設計・実装に焦点を当て、汎用性と拡張性を持ったパケット転送機構の設計と、NetBSD 上に実装する際に採用した実装手法について述べる。

Design and Inplementation of Packet Forwarding Engine in MPLS Environment "AYAME"

Satoshi Uda　　Yojiro Uo　　Yoichi Shinoda

Japan Advanced Institute of Science and Technology, Hokuriku

The MPLS is a hot technology, because this is one of the base technologies for the Internet-ng, which aspire to high speed and high functional packet forwarding. We proposed experimental implementation of Multi Protocol Label Switching (MPLS) base a Label Switching Router (LSR), named "AYAME". In this paper, we especially focus on flexible and extensible design of the AYAME's packet forwarding engine, among with techniques utilized in AYAME's NetBSD based implementaion.

1 はじめに

インターネットの利用者層および利用用途の広がりに伴い、インターネットのパケット配送機構に対する高速化・高機能化への要望が高まっている。現在用いられているインターネットのパケット配送機構は、最善努力型を基本として設計された。しかし、既に、最善努力型の配送機構のみでは利用者の要求を満たすことが困難となりつつある。

そこで、利用者の要求を満たすために、高速・高機能配送機構を実現する次世代インターネット技術の開発が盛んに行われている。この高速・高機能配送機構を実現する基礎技術の 1 つとして MPLS (Multi Protocol Label Swintching) [4] [5] が提案されている。また、高機能な配送機構を MPLS を用いて実現する手法が、盛んに研究されている。

我々は、MPLS を用いた高機能配送機構等に関する研究を行う上で、その評価実装を行うための汎用的基盤実装の必要に迫られ MPLS 実装 AYAME の設計・実装を行った [1]。その設計・実装においては、今後の MPLS に関する研究開発活動の中で新機能の追加等に柔軟に対応できるよう汎用性を重視し、特定の用途やプロトコル等に特化しない事に留意した。

MPLS の実現には、実際のパケット転送処理を行うパケット転送機構と、パケット転送機構を制御するための機構が必要である。本論文では、特に、AYAME の実装におけるパケット転送機構に

焦点を当て、その設計と NetBSD[1] ネットワークスタックへの組み込み手法について述べている。

2 MPLS

2.1 MPLS の概要

MPLS はネットワークプロトコルに依存しないラベルスイッチング技術であり、そのラベルとして積み上げ可能なラベルスタック構造を採用している。

MPLS 網内では、パケットにはラベルが付けられた状態で配送され、中継する MPLS 対応ルータ (LSR: Label Switched Router) 毎に、ラベルの参照/書き換えが行われた上で次ホップへ転送される。パケットに付けられるラベルは、そのパケットを受け取る LSR が行うべき転送挙動に対応するものである。あるルータで同一の転送挙動をすべきパケットの特徴・種類の集合を**同一転送クラス** (FEC: Forwarding Equivalence Class) と呼ぶ。パケットに付加されるラベルについて FEC 概念を用いて表現すると、「パケットを受け取る LSR の FEC に対応したラベルが付けられている」と言える。

MPLS は、次世代インターネットにおけるネットワークの高機能化・高速化に対応するための技術として設計されたため、インターネット技術との親和性に優れている。また、次のような特徴も持っている。

- 高速なパケット転送が可能
  LSR のパケット転送の際の各パケットに対する処理内容が単純なため。
- 扱うことのできる FEC 集合への制限がない
  パケット転送機構が特定の FEC 集合に依存していないため。

2.2 MPLS 網の動作概要

MPLS を用いてパケットを配送するには、各 LSR は次ホップ LSR における、1) FEC 情報、2) FEC / ラベル対応情報を知る必要がある。これは、先にも述べた通り、MPLS を用いてパケットを送信する際は、次ホップ LSR の転送挙動に対応したラベルを付加する必要があるためである。この情報交換を実現するためのプロトコル群を、ラベル配布プロトコルと呼ぶ。

[1][NetBSD] *http://www.netbsd.org/*

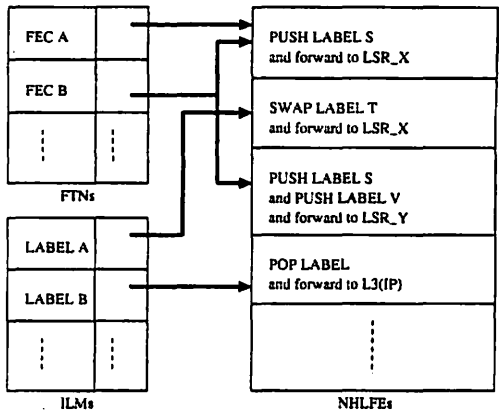


図 1: NHLFE / ILM / FTN

ここで注意すべき点は、ラベル配布プロトコルは経路制御を行うためのものではないという点である。MPLS には MPLS 網を制御するような経路制御機構は存在しない。LSR における各 FEC に対する転送挙動は、その FEC を扱うネットワーク層等の既存の経路制御機構によって決定される。LSR では、各種経路制御機構がもたらす FEC とそれに対する転送挙動を、FEC に対するラベルとそのラベルに対する操作に置き換えて処理する。

ラベルの付いたパケットの配送に関わる各 LSR の転送挙動は非常に単純である。ラベル付きパケットを受信した LSR は、そのラベルからデータベースを参照することにより転送挙動を決定する。LSR における転送挙動は、ラベルスタックに対する n 回の PUSH (ラベル追加) / POP (ラベル除去) / SWAP (ラベル置換) 操作と、スタック操作後の適当な次ホップ LSR への送信である。

MPLS アーキテクチャにおいて、転送挙動を決定するためのデータベースは図 1 のような構造であり、それぞれの表には以下の情報が記載される。

- **NHLFE (Next Hop Label Forwarding Entry)**
  パケットに対する転送挙動
- **ILM (Incoming Label Map)**
  あるラベルの付いた入力パケットに対し用いるべき NHLFE 集合
- **FTN (FEC-to-NHLFE Map)**
  ある FEC に属するパケットに対し用いるべき NHLFE 集合

ここで述べた MPLS の動作概要を各 LSR での処理の観点から見た場合、一般的 IP ルータの持つ機構と比べて、大きく次の 2 つの拡張がなされていると言える。

- パケット転送機構のラベルへの対応
  NHLFE / ILM / FTN に基づき、ラベルを用いたパケット転送動作を行う。
- FEC/ラベル管理機構
  ラベルの割り当て/管理、ラベル配布プロトコルを用いたラベル情報等の交換を行い、各種経路制御機構の情報を基に NHLFE / ILM / FTN の生成・管理を行う。

3 MPLS 実装 AYAME

我々は、研究・開発用の MPLS スタックとして、"AYAME" の設計・実装を行った。

AYAME は、NetBSD に対する拡張として実装した。NetBSD は、4.4BSD の流れをくみ自由に利用可能で再配布可能な、UNIX-like なオペレーティングシステムである。我々は、AYAME の利用/配布が、基礎とするコードの利用条件/配布条件に束縛されることは研究開発目的に利用する上で好ましくないと考え、自由に利用可能で再配布可能な NetBSD を選択した。

AYAME の設計・実装においては、以下の点について特に留意した。

- 拡張性の高い FEC 決定手法
  通常の IP ルータにおいては、各パケットに対する転送方針はネットワーク層の経路表のみで決定される。これはつまり、パケットの属する FEC は経路表のみから決定することが可能であると言える。しかし、この先のインターネットの高機能化を考慮すると、常にパケットの属する FEC が経路表のみから決定されるとは限らなくなると考え、より拡張性の高い FEC 決定手法を検討した。
- モジュール構造を用いた実装
  パケット転送機構において、特定のデータリンク層やネットワーク層に依存する部分と汎用部との切り分けを行った。これにより、現在対応していない新たな層への対応や新機能の追加を容易とした。
- 複数のラベル配布プロトコルの使用が可能
  同一の FEC 群を取り扱う複数のラベル配布プロトコルの共存が可能な FEC/ラベル管理機構について検討を行った。
- 既存スタックへの最小限の影響
  既存の NetBSD のネットワークスタックへの変更点が最小限に押さえられるよう設計実装した。これは、既存スタックへの変更は MPLS 拡張以外の新たな機能追加との親和性を低くし、また、MPLS 拡張自体の汎用性を低減する恐れが強いためである。

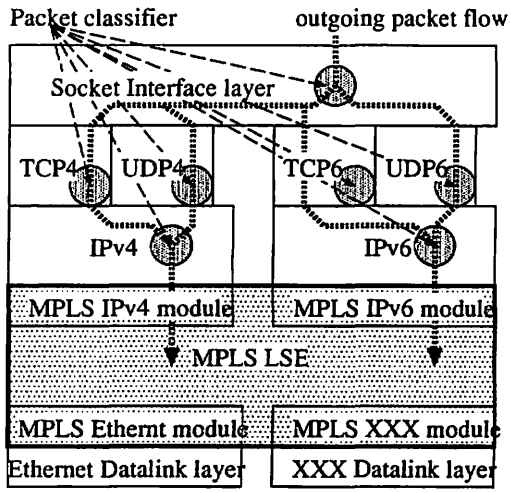


図 2: LSE と FEC 決定機構

先にも、LSR を実装する上で拡張が必要となる機能は、パケット転送機構と FEC/ラベル管理機構であると述べたが、本論文では特にパケット転送機構に着目している。AYAME のパケット転送機構について見ると、以下の 2 つの機構で構成される (図 2)。

- LSE (Label Switching Engine)
  ラベルの付いたパケットに対し、必要な処理を行ったうえで次ホップ LSR への転送を行う機構。
- FEC 決定機構
  MPLS ネットワークの入口点等において、ラベルの付いていないパケットに対しそのパケットが属する FEC を決定する機構。この決定に基づいて、パケットにラベルが付けられる。

本章では、以下に、AYAME におけるパケット転送機構の実装手法について述べる。

3.1 LSE の実現

AYAME における LSE は、NetBSD のネットワーク層スタックの 1 つとして実装した。その際、既存のネットワーク層/データリンク層スタックと LSE スタック間の API として、ネットワーク層-データリンク層間 API を用いている (図 3)。これによ

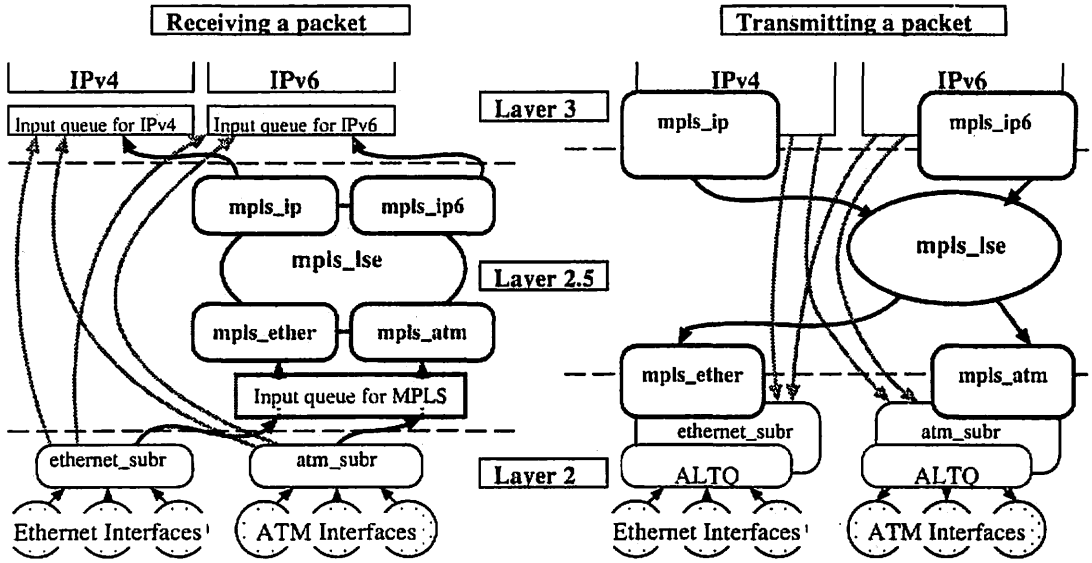


図 3: LSE モジュール構成

り、データリンク層およびネットワーク層の既存スタックへの変更が最小限に押さえられた。

また、LSE 内部構造は、複数のデータリンク層/ネットワーク層依存部分と汎用部を分離しモジュール化している。

汎用部 (*mpls_lse*) は、ラベルの付いたパケットに対し、ラベルから処理内容を検索し、pop/push/swap の操作を行った後、適当なインタフェイスに出力するのみの単純なものである。複数回の pop/push/swap 動作は、再度 *mpls_lse* を通過することにより実現している。

データリンク層依存モジュール (*mpls_ether* など) は、ラベルのエンコーディング方法等データリンクに依存した処理を行う。

ネットワーク層依存モジュール (*mpls_ip* など) は、ネットワーク層より MPLS 層に送られてくるラベルの付いていないパケットについて、そのパケットの属する FEC に対応するラベルを付加した上で LSE に渡す等、特定のネットワーク層に依存した処理を行う。

このような構成のため、*mpls_lse* は、データリンク層からのパケットもネットワーク層からのパケットも同一視して扱うことができ、単純化されている。また、処理方針を記述する表は、ILM と NHLFE を組合わせた形で実現した。

3.2 FEC 決定機構の実現

先にも述べた通り、現状では、一般的にパケットの属する FEC はネットワーク層の情報で決まる。しかし、今後のインターネットの高機能化に伴い、パケットの転送により細かい粒度での制御が必要となると考えられる。ネットワークスタック内でパケットの持つ情報量は上位層ほど多く、下位層ほど少ない。例えば、TCP/IP においては、IP 層では { 始点アドレス, 宛先アドレス, 上位プロトコル種別 } しか知ることができないが、TCP 層では、これに加え「サービス種別 (ポート番号)」を知ることができる。

そこで、AYAME では、パケットの属する FEC の決定戦略として、パケットが上位層から下位層に流れる過程において、各層でそのパケットが属する FEC が決定できる場合はその FEC を採用し、決定できない場合はその決定を下位層に委ねるとした (図 4)。また、IP パケットの場合は上位層で FEC が決定できなかった際は、既存の IP ルータと同じく IP 層で経路表を検索した際に決定することとなる。

3.2.1 mbuf 拡張

この FEC 決定戦略を NetBSD のネットワークスタック上で実現するためには、ネットワークスタック間でパケットの受け渡しを行う際に、FEC に関する情報とともに受け渡す必要がある。

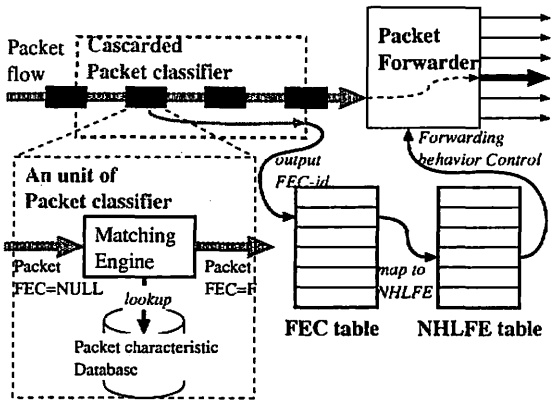


図 4: FEC 決定戦略

FEC に関する情報の受け渡し手段としては、以下の 2 手法が考えられる。

- スタックを実現している関数間で関数呼び出しの際の引数として受け渡す。
- パケットを格納している構造 (mbuf) の持つ情報の 1 つとして受け渡す。

前者は、変更の影響範囲が FEC 決定機構に関連する関数のみに限定される利点が挙げられるが、関連する多くの関数間の API を変更する必要が生じる。後者は、mbuf の構造を拡張するため mbuf を利用しているネットワークスタック全域への影響が懸念されるが、関数間の API を拡張する必要がなくコード上の変更は小さなもので済ますことができる。

AYAME では、今後 IP のみに限らず、さまざまなスタックを MPLS に対応させることを考え、後者 (mbuf の拡張) を選択し実装した。

具体的な mbuf の拡張は、パケットが格納される mbuf 群に付けられるパケット情報 (パケットの全長等) が格納されている *pkthdr* 構造体を拡張し、FEC 情報を格納する事とした。これに対応するため、mbuf の確保/分割等に関わる部分に若干の変更を加えた。

3.2.2 経路表の拡張

ここで述べた FEC 決定戦略においても、上位層でパケットの属する FEC が決定できない場合、最終的にはネットワーク層の情報を用いて属する FEC を決定する。また、NetBSD においては、ほとんどのネットワーク層はパケットの取り扱い方針を経路表に格納している。

そこで AYAME では、経路表の各エントリに対し、FEC 情報を格納できるよう拡張した。経路表に FEC 情報を格納する手法としては、以下の 2 手法が考えられる。

- 次ホップを表す項目に、FEC 情報を格納し MPLS スタックを用いて転送を行うことを指示する。
- 新たに FEC 情報を格納する項目を追加し、ここに FEC 情報を格納する。

ここで、次ホップを表す項目は、既存の各種経路制御機構が直接読み書きする。そのため、前者の手法を採るためには、既存の経路制御機構への変更が必要となる。既存の経路制御機構は、既に各種さまざまなものが提案・実装されており、これらをそのままの形で利用し続けることが困難となる前者の実装は好ましくない。そこで、AYAME では後者の手法を選択し実装した。

具体的には、FEC 情報を格納するために経路表のエントリ内の項目を追加し、この情報を操作するために経路表操作 API の拡張を行った。

4 評価

我々の行った AYAME 実装は、パケット転送機構について考察すると、汎用性/拡張性の高いものとなった。現在の実装では、ネットワーク層として IPv4 および IPv6 のみにしか対応していないが、他のネットワーク層をただ単純に MPLS 対応される程度であれば、対応すべき新たなネットワーク層自体への非常に小さな変更とこれに対応した LSE のネットワーク層依存モジュールの実装のみで可能である。これは、設計時から、特定のプロトコル等への依存を極力避け、依存が不可欠な場合はその部分を汎用部と明確に切り分けることを着実に行った成果である。また、現在の実装では、データリンク層として Ethernet のみしか対応していないが、これも、一部の特殊なデータリンクを除いては小さな労力で対応が可能である。ここで、挙げた一部の特殊なデータリンクとは ATM およびフレームリレーを指している。ATM 等のデータリンクでは、パケットに対するラベルの付加手法として、各パケットのヘッダに直接付加せず、データリンクの持つ機能を利用することでラベルの情報を次ホップに伝える。このような場合、データリンク層コードの機能拡張とその拡張機能を用いるデータリンク層依存モジュールが必要となる。し

かし、この場合も、LSE の汎用部自体は我々が実装したものを用いることができると考えている。

公開されている MPLS 実装としては、1) NIST Switch [2]、2) MPLS for Linux (Univ. of Wisconsin) [3] など数種の実装が存在する。しかし、1) は RSVP を MPLS 上で実現する事を目的とし、2) は LDP の実験を行う事を目的とし、特定用途向け実装との色合いが濃い。我々の AYAME は、設計段階より汎用性と拡張性を強く意識しており、その成果が先にも述べた高い汎用性/拡張性に繋がっている。特に、AYAME の FEC 決定機構はネットワーク層以外の情報を用いた経路制御への道を開いており、既存の IP 経路制御 (宛先アドレスに基づく経路制御) に縛られない新たな経路制御構造の導入も可能としている。

実際にネットワークを運用する場合には、現状の AYAME 実装には問題がある。現状の実装はコードの汎用性等と実装の容易さ優先して行われたため、パケット転送能力に関する最適化ををほとんど行っておらず、MPLS の特徴の 1 つである「高速なパケット転送」が十分には生かされていない。しかし、現状の実装においても、既存ネットワーク層を用いたパケット転送と同等以上の性能は維持されており、さらに、各モジュール毎のコードの最適化を行うことによりさらなるパケット転送能力の増強を見込むことができると考えられる。また AYAME 実装の動作検証として、WIDE 9 月研究会 (平成 12 年 9 月 11 日～14 日/参加者約 250 人) において AYAME 実装のルータ 7 台を持ち込み、ユーザセグメント (IPv4) のバックボーンとして大きな障害もほとんどなく約 4 日稼働した実績が挙げられる。

さらに、我々は、MPLS の特徴として挙げれられる「扱うことが可能な FEC 集合に制限がない」点を用いて、より粒度の細かい FEC を扱う経路制御機構を用いた研究・提案として次のようなものなどを手がけている [2][3]。

- マルチキャスト配送のための高機能配送機構の実現手法
- QoS 対応配送機構の実現手法

これらの設計・実装において、本論文で述べた AYAME のパケット転送機能は、その用途での利用に十分耐えられるものである。

[2] [NIST Switch] *http://www.antd.nist.gov/itg/nistswitch/*
[3] [MPLS for Linux] *http://nero.doit.wisc.edu/mpls-linux/*

5 まとめ

我々は、MPLS に対応した LSR 実装 AYAME を提案、設計、実装した。本論文では、AYAME の実装の中で、特に、パケット転送機構の実装、その汎用性を実現するために我々が用いた実装手法について述べた。

AYAME のパケット転送機構における、今後の課題としては、その高速化の実現が挙げられる。先に挙げた、モジュール単位でのコードの最適化に加え、各種データベースのキャッシュ機構の整備、パケット処理用ハードウェアとの連携等を検討・実装することによりパケット転送機構を高速化することを検討している。また、最適化後のパケット転送能力等に関する評価も行いたいと考えている。

また、我々は、インターネットにおけるパケット配送機構の高機能化に用いることができる技術の 1 つとして MPLS に注目しており、その評価実装に用いる基盤実装として AYAME を活用していく予定である。

我々は、AYAME は MPLS を用いた研究開発において、その評価実装等を行う際の基礎実装とできるものであると考えている。AYAME は、自由な利用、再配布が可能となるよう実装されており、今後の AYAME 実装の配布等についても検討している。

参考文献

[1] Yojiro Uo, Satoshi Uda, Nobuo Ogashiwa, Satoshi Ohta, Yoichi Shinoda: "AYAME: A design and implementation of the CoS capable MPLS layer for BSD Network stack". Proceedings of INET2000, Internet Society. July 2000.

[2] 小柏伸夫, 宇夫陽次朗, 篠田陽一: "多層経路制御層型マルチキャスト配送の MPLS による実現". WIT 2000, 日本ソフトウェア科学会.

[3] 小柏伸夫, 宇夫陽次朗, 宇多仁, 篠田陽一: "多層経路層型マルチキャスト配送機構の設計と実装に関する考察". DPS 2000, 情報処理学会. (投稿中).

[4] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, A. Viswanathan: "A Framework for Multiprotocol Label Switching". Internet-Draft, Network Working Group, IETF. work in progress, September 1999. (draft-ietf-mpls-framework-05)

[5] Eric C. Rosen, Arun Viswanathan, Ross Callon: "Multiprotocol Label Switching Architecture". Internet-Draft, Network Working Group, IETF. work in progress, August 1999. (draft-ietf-mpls-arch-06)