■5 群（通信・放送） － 4 編（ノード技術）

5 章 IP データ系システム

（執筆者：矢崎武己）［2009 年 5 月 受領］

■概要■

インターネットの急速な普及に伴い，そのトラヒック量も急速に増加している．インターネットのバックボーンに位置する IX（Internet eXchange）では，年率 30～40 %でトラヒック量が増加しているとのデータもある．このため，ルータ，L2 スイッチなどの IP データ系システムの高速化が求められ，高速ルータ，L2 スイッチにおいては，パケットの転送動作をハードウェアで実現し，高速化することが一般的となっている．また，このハードウェア化に加え，ルータ，L2 スイッチを高速化するうえでボトルネックとなる，ルーティングテーブルや FDB（Filtering Data Base）の検索処理，フローの識別処理，バックプレーンに位置するスイッチなどの高速化技術が開発されている．これらの技術開発により，ルータ及び L2 スイッチは急速に高速化し，数 Tbit/s の交換容量を備える製品が登場している．

インターネットに代表される IP ネットワークの普及の一つの要因は，複数のユーザがネットワークの帯域資源を共有することによる利用コストの低廉性である．従来，専用線や電話網といった専用ネットワークで実現してきたサービスを，この低廉性を活かして IP ネットワークにより安価に実現する要求が高まっている．そのため，専用のネットワークで実現してきた通信品質及び信頼性・可用性を IP ネットワークにて確保するルータ，L2 スイッチの技術が開発されている．通信品質を確保する QoS 制御技術としては，ポリシング，シェーピング，バッファ制御技術などが，高信頼化・高可用化技術としては，L2 スイッチのリング技術，リンクアグリゲーションなどがある．

【本章の構成】

本章では，ルータの基本動作，高速化技術，QoS 制御技術（5-1 節），及び，L2 スイッチの基本動作，高速化技術，高信頼化・高可用化技術（5-2 節）に関して，製品化の状況などを交えつつ解説する．

■5群 － 4編 － 5章

5-1 IPルータシステム

（執筆者：矢崎武己）［2009年5月 受領］

5-1-1 ルータの高速化

世界初の商用IPルータシステム（以下，ルータ）は，1986年1月に発売され，その後，1986年3月に発売されたルータがベストセラーとなった．専用の筐体に収まる専用機器であったが，アーキテクチャは汎用のコンピュータと大きな違いはなく，プロセッサと回線を収容するインタフェースカードを1本のバスでつなげた構成であった（**図5･1**）[1)]．

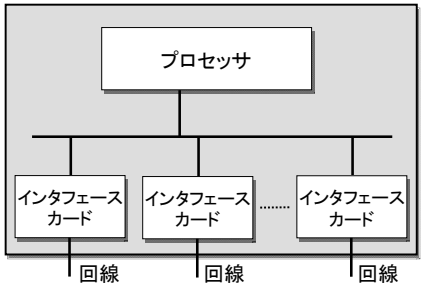


図5･1　プロセッサベースのルータアーキテクチャ

このルータにおいては，パケットの転送は1本のバスを通して行われるため，このバスが性能のボトルネックとなりやすく，性能は10 kパケット/s程度であった．全パケット長がイーサネットを用いた際の最大パケット長である1500バイトとすると120 Mbit/s程度の性能である．

1993年には，パケットの転送動作をハードウェアで高速化するハードウェアルータが登場した．それ以来，インターネットトラヒックの急増及び半導体技術の進展を背景に，ルータは急速に高速化していった．最近の10年間では，およそ年率30 %で高速化しており，現在では，数Tbit/sのルータ製品が登場している（**図5･2**）．次節以降では，ルータのパケット転送動作，パケット転送動作を高速化する技術，通信品質の確保や通信優先付けを行うQoS制御機能について説明する．

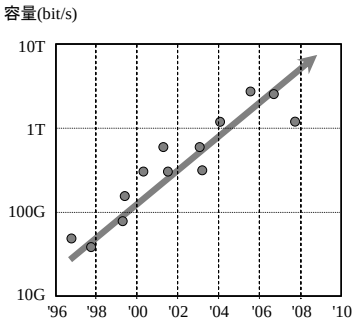


図5･2　ルータ容量の年次推移

5-1-2 ルータのパケット転送動作

ルータの基本的なパケット転送動作を**図 5･3** のルータの論理構成図を用いて説明する．ルータはパケットが入力する複数の入力回線とパケットを出力する出力回線を備える．また，パケットをスイッチングするスイッチ，パケット処理部，及びパケットの帯域や順序を制御して出力回線に出力する制御部（本章では出力制御部と呼ぶ）を備える．

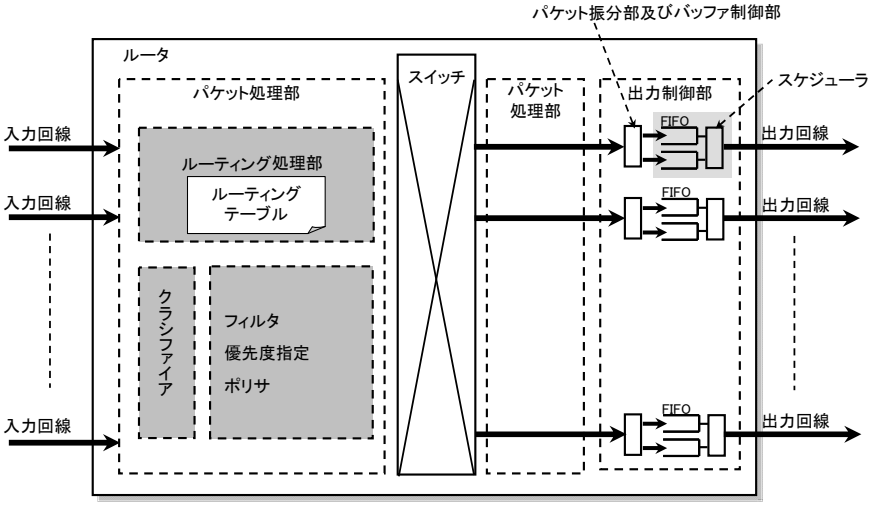


図 5･3　ルータの論理構成図

パケットが入力回線より入力すると，パケット処理部はパケットに付与されたヘッダ内の宛先 IP アドレスに基づき，ルーティングテーブルの検索処理を実施し，パケットを出力する出力回線や次ホップの IP アドレスなどの判定を行う（本章では本処理を行う機能部をルーティング処理部と呼ぶ）．なお，このルーティングテーブルは，ルータがルーティングプロトコルを用いて経路（ルート）を他のルータと互いに通知し合うことで作成される．

本処理と並行して，クラシファイアが IP ヘッダ，TCP/UDP ヘッダに格納される情報などに基づき，パケットが属するフローを識別する（本章ではフロー識別と呼ぶ）．更に，識別したフローごとに，(1)パケットを廃棄するフィルタ，(2)パケットの優先度指定，(3)帯域の監視を実施するポリシング（ポリシングを実施する機能部をポリサと呼ぶ）が実施される．

スイッチは，ルーティング処理部で判定された出力回線に基づいてパケットをスイッチングし，出力制御部の FIFO（First In First Out）に送信する．出力制御部の前段に配置されたパケット処理部が，フィルタ，優先度指定，ポリシングを再度実施する場合もある．出力制御部は，出力回線ごとに備えた複数の FIFO にパケット処理部で判定された優先度に従いパケットを蓄積し（図 5･3 ではパケット振分部と記載），スケジューラにより FIFO からパケットを読み出して出力回線に出力する．また，FIFO の前段に配置された機能部（図 5･3 ではバッファ制御部と記載）が，FIFO の蓄積状況をモニタし，パケットの蓄積・廃棄を判定する．

5-1-3 ハードウェアルータアーキテクチャ

前節にて説明したパケット転送動作を高速化するハードウェアルータの代表的なアーキテクチャとして，分散型アーキテクチャ [1), 2)] と集中型アーキテクチャが存在する．分散型アーキテクチャのブロック図の一例を**図 5･4** に示す．本アーキテクチャのルータはインタフェースカードと，複数のインタフェースカードを結合するスイッチカード，装置の制御やルーティングプロトコルの処理を実施する制御部より構成される．

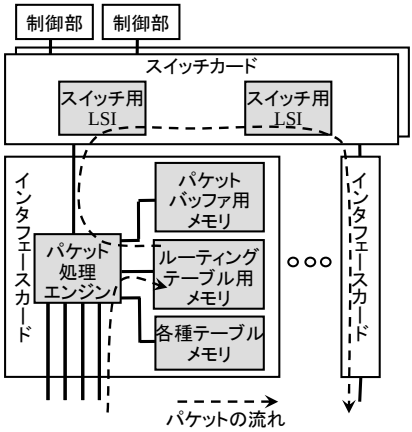


図 5･4　分散型アーキテクチャによるルータ

インタフェースカードは，ルーティング処理，フロー識別及びフィルタ，優先度指定，ポリシングを実施するパケット処理エンジンと，ルーティングテーブル用メモリ，フィルタ/優先度指定/ポリシングを行うための各種テーブルメモリと，パケットを蓄積するパケットバッファ用メモリより構成される．

パケット処理エンジンやスイッチ用 LSI は高速な ASIC（Application Specific Integrated Circuit），FPGA（Field Programmable Gate Array）及び NP（Network Processor）などにより実現される．また，パケットバッファ用メモリ，ルーティングテーブル用メモリ，各種テーブルメモリは SRAM（Static Random Access Memory），DRAM（Dynamic Random Access Memory）や後述する TCAM（Ternary Contents Addressable Memory）などのメモリにより実現される．

制御部はルーティングプロトコルを処理し，経路を他のルータに通知するとともに，IP アドレスに対応した宛先などを管理するルーティングテーブルを作成し，転送動作に必要な情報をインタフェースカードのルーティングテーブル用メモリに格納する．このインタフェースカード上のテーブルを FIB（Forwarding Information Base）と呼ぶこともある．分散型アーキテクチャのハードウェアルータにおいては，高速なパケット処理エンジンやスイッチ用 LSI が，ルーティングテーブル用メモリや各種テーブルメモリに格納された情報に従い，パケット処理，パケット転送を実施する．複雑なルーティングプロトコルの処理が制御部にて実施され，比較的単純なパケット転送処理がハードウェアにて高速実行されるアーキテクチャである．本アーキテクチャでは，パケット転送のボトルネックとなるバスの代わりにハードウェアのスイッチ用 LSI がパケットをスイッチングすることで，バスのボトルネックが解消

される．

分散型アーキテクチャを採用したルータは，スイッチの容量が許容する限り，インタフェースカードの枚数を増加することで容量を拡大することが可能であり，容量スケーラビリティが高い．一方で，パケットがパケット処理エンジンを 2 回通過し，パケットバッファ用のメモリの書き込み／読み出しが 2 回発生する．更に，ソフトウェアルータでは必要ないスイッチ LSI が必要となるため，本アーキテクチャを採用したルータは，高価となりやすい．

図 5･5 にこれらの問題に対応した集中型アーキテクチャを示す．本アーキテクチャでは，パケット処理カードに実装されたパケット処理エンジンやメモリを複数のインタフェースカードが共有する．分散型アーキテクチャのスイッチ LSI は存在せず，パケット処理エンジンとパケットバッファメモリがスイッチの機能を兼ね備え，宛先に対応するインタフェースカードにパケットを転送する．

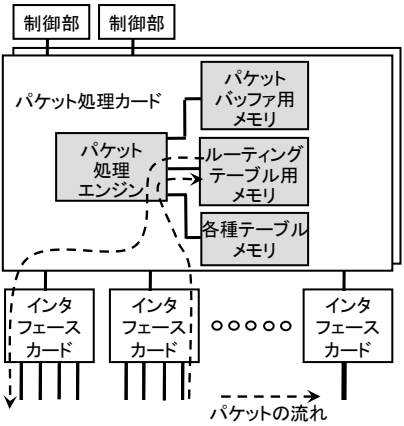


図 5･5　集中型アーキテクチャによるルータ

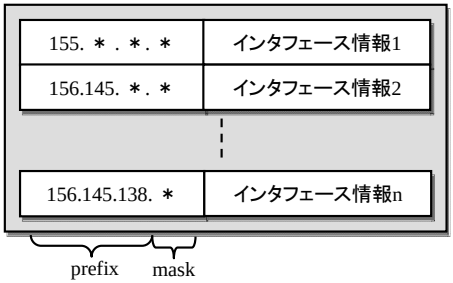
パケット処理用エンジンの処理は 1 回であり，パケットバッファ用メモリの書き込み／読み出しは 1 回となる．更に，スイッチカードが不要なため，本アーキテクチャのルータは分散型アーキテクチャのルータと比べ安価となる．一方で，パケット処理エンジンの性能以上にルータの容量を拡大することが難しく，ベンダや製品化時期によりばらつきはあるが，概ね 100 Gbps～数 100 Gbps 以上のルータでは分散型アーキテクチャが採用される．

5-1-4 ルーティングテーブルの検索技術

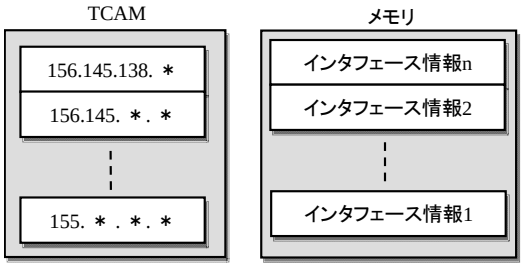
ルーティングテーブルには，IP アドレスの条件とその条件に対応する次ホップ IP アドレスや出力回線などのインタフェース情報が記載される．ルーティングテーブル検索では，パケットが入力した際にその宛先 IP アドレスに一致する IP アドレスの条件が検索され，一致する条件に対応するインタフェース情報が，そのパケットのインタフェース情報と判定される．

ハードウェアルータにおいては，ルーティングテーブル用メモリに，ルーティングテーブルが構成される．**図 5･6**(a) にこのルーティングテーブルの論理構成の例を示す．IP アドレ

スの条件は，上位ビットの prefix と下位ビットの mask から構成される．prefix 部分が IP アドレスの値を指定する情報であり，mask 部分は IP アドレスの値を指定しない．複数の IP アドレス条件に一致する場合があるが，prefix のビット長（prefix 長）が最も長い条件に対応するインタフェース情報が採用される．この検索を最長一致検索（Longest Prefix Match）と呼ぶ．最も単純な検索方式としてリニアサーチ（Linear Search）がある．本方式では，prefix 長の長い条件順に，その条件と対応するインタフェース情報を，メモリアドレスの小さなエントリより順に配置しておく．検索時には，メモリアドレスの小さなエントリ内の条件から順にパケットの宛先 IP アドレスと比較し，最初に一致した条件に対応するインタフェース情報を判定する．インターネットでは，経路が多い場合には数 100 k 程度となる場合もあり [3]，IP アドレスの条件としても数 100 k の条件が設定される．このため，リニアサーチは高速なルーティングテーブルの検索が難しく，以下に示す高速検索方式が開発されてきた．



(a)　ルーティングテーブルの論理構成



(b)　TCAM のよるルーティングテーブル検索の実現例

図 5・6

(1)　パトリシアツリー（Patricia Tree）

本方式は，IP アドレスの条件を二分木のパトリシアツリーに展開する方式である [5]．リニアサーチと異なり，設定される IP アドレスの条件の数 : N に検索性能が依存しない特徴がある．**図 5･7** に IP アドレスが 3 ビットの場合のルーティングテーブルと，対応するパトリシアツリーを記載する．IP アドレスの条件に記載されている ＊ は mask を表している．

ルーティングテーブル

IPアドレス	インタフェース情報
0 0 0	A
0 0 1	B
1 * *	C
1 0 *	D
1 1 0	E

パトリシアツリー

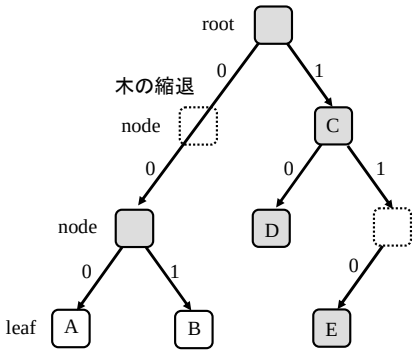


図 5・7　ルーティングテーブルと対応するパトリシアツリー

パトリシアツリーにおいては，IP アドレスの条件の各ビット（0 or 1）は枝として表現される．条件の上位ビットから順に根（root）から枝が接がれ，枝の先には条件に対応するインタフェース情報が記載された葉（leaf）が形成される．例えば，図 5･7 の場合，root より 1，1，0 の枝を辿った先には IP アドレス 110 に対応するインタフェース情報 E が格納される leaf がある．mask が存在する場合には，分岐点である節（node）にインタフェース情報が記載される．例えば，1** に対応するインタフェース情報 C は，root から分岐した 1 の枝の先にある node に記載される．また，分岐がない node は省略される．例えば，root から分岐した 0 の枝の先には 1 の枝が存在しないため，0 の枝の先の node は省略される．

検索時にはパケットの宛先 IP アドレスの上位ビットから，対応する枝が辿られ，最終的に到達した leaf に記載されたインタフェース情報が，パケットのインタフェース情報と判定される．対応する枝がない場合には，前の node に記載されるインタフェース情報が最終的なインタフェース情報と判定される．例えば，IP アドレスが 111 の場合には，3 ビット目に対応する 1 の枝がないため，C がインタフェース情報と判定される．

本方式の性能は，枝を辿る回数で決まり，IP アドレスの全ビットに対応する枝をすべて辿った場合に最悪となる．性能は経路数 N に依存しないため，最悪性能を見積もることが可能となる．一方で，IP アドレスのビット数に比例して性能が劣化するため，IPv6 アドレス（ビット長 128 ビット）のルーティングテーブルを検索する性能が大きく劣化することが課題となる．

(2)　TCAM (Ternary Contents Addressable Memory)

近年，TCAM と呼ばれる半導体メモリを使用してルーティングテーブルを高速に検索する方式 [6] がハイエンドルータに適用されている．TCAM は条件を記載するエントリを複数備え，検索キーが入力されると検索キーとエントリ内の条件を高速に一致比較し，一致したエントリのアドレスを出力する．

TCAM を用いた際のルーティングテーブルの実現例を図 5･6(b) に示す．TCAM のエント

リには，IP アドレスの条件が prefix 長が長い順に格納され，各条件に対応したインタフェース情報が通常のメモリに格納される．検索時には，パケットの宛先 IP アドレスを検索キーとして TCAM が検索され，一致したエントリのうち最も小さなアドレスに対応したメモリ内のインタフェース情報が，パケットのインタフェース情報として読み出される．

現状の TCAM に IPv4 の 32 ビットの IP アドレスの条件を格納する場合，一つの CAM により 500 k 以上の条件を格納し，毎秒 200 M 回以上の検索を実施することが可能である [7]．毎秒 200 M 回の検索は，最低でも 100 Gbit/s（= 64 バイト×8 ビット/バイト×200 M/s：64 バイトは回線がイーサネットを用いた際の最低パケット長）で入力されるパケットの検索に相当する．また，現状の経路数は 300 k 程度であるため [3]，一つの CAM によりルーティングテーブルが実現可能で，メモリ容量といった観点でも不足ない．しかしながら，TCAM は価格が通常のメモリに比べ高価であり，2005 年時点で DDR SDRAM と比べてビット辺りの単価が 30 倍とのデータもある [8), 9)]．また，消費電力もビット辺り 100～150 倍の電力となる [8), 9)]．このため，TCAM はハイエンドルータなどの高速性が必要なルータに適用される傾向がある．

これまでに説明した高速化方式に加え，様々な検索方式が提案されている．例えば，同一の宛先 IP アドレスのパケットが比較的短期間に集中して発生する時間的局所性に着目し，宛先 IP アドレスとインタフェース情報をキャッシュに格納する方法 [4] などである．最適な検索方式がルータに要求される経路数，性能などに応じて選択されることとなる．

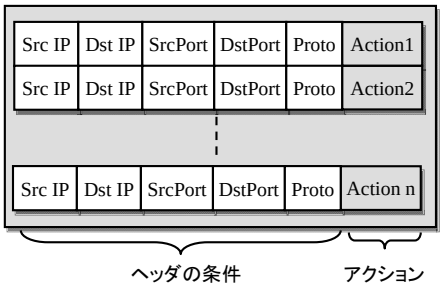
5-1-5 フロー識別技術

フローごとにフィルタ，優先度指定，ポリシングなどのパケットの処理を実施するために，クラシファイアはパケットのヘッダ内の情報などに基づいて，そのパケットが属するフローを判定する．この判定をフロー識別と呼ぶ．

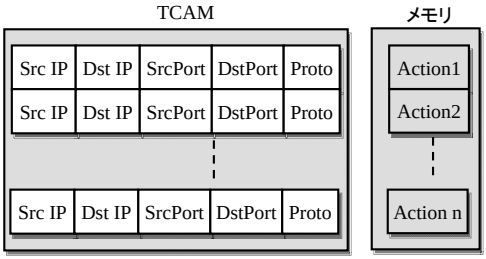
図 5･8(a) にクラシファイアとフロー識別結果に基づくパケットの処理を実施するためのテーブルの論理構成を示す．本テーブルはヘッダの条件と対応するアクションが格納される．このアクションとしては，フィルタであれば通過または廃棄が，優先度指定であれば優先度が，ポリシングであれば帯域を監視するための情報が記載される．一つのパケットに関して，複数の条件に一致する場合があるため，ルータの管理者により条件の優先度が付与される．

代表的なヘッダ条件は，5-tuple と呼ばれる IP ヘッダ内の送信元 IP アドレス（Src IP），宛先 IP アドレス（Dst IP），プロトコル（Proto），TCP/UDP ヘッダ内の送信元ポート番号（Src Port），宛先ポート番号（Dst Port）である．送信元 IP アドレス，宛先 IP アドレスには主に mask が，送信元ポート番号，宛先ポート番号には値の範囲が指定可能であるのが一般的である．

5-tuple によるフロー識別を行う場合，ヘッダの条件はアドレスのビット長が小さな IPv4 であっても 104 ビットにも達する．更に，数 k もの条件が設定される場合もあり [9]，フロー識別の高速化が課題となる．更に，ルータによっては，データリンクヘッダ内の送信元／宛先 MAC アドレスや，IPv6 の IP アドレス（128 ビット）が設定可能な場合もあり，ヘッダの条件の情報が更に大きくなる場合もある．このため，フロー識別はルーティングテーブルの検索とともに，ルータの高速化を実現するうえでのボトルネックとなる．



(a) クラシファイアの論理的なテーブル構成



(b) クラシファイアの実装例

図 5・8

フロー識別の方式は 4 方式に大きく分けることができる [9]．第 1 の方式はすべての条件を精査する Exhaustive Search である．順々に条件を精査していくリニアサーチ（Linear Search）や TCAM などが代表である．第 2 の方式は Decision Tree で，5-1-4 項に記載したパトリシアツリーと同様に，条件が展開されたツリーを順に辿っていく方式である．本方式に分類される方式として Grid-of-Tries [10] がある．第 3 の方式は，条件を Src IP，Dst IP といった各フィールドに分解し，フィールドごとの一致判定と，その結果より最終的な結果を判定する Decomposition である．最後の方式は Tuple Space であり，mask 位置が同一の条件ごとに一致検索（Exact Match）を行う方式である．

これらの検索方式の中でも TCAM を用いた方式は高速であり，ハイエンドのルータに適している．図 5･8(b) に TCAM によるフロー識別の実現例を示した．ルーティングテーブルの検索と同様にヘッダの条件を記載する TCAM と，その条件に対応するアクションを記載したメモリより構成される．

現在の TCAM は，500 ビット以上の条件を設定可能であり，IPv6 の IP アドレスにも対応可能である．また，IPv4 の 5-tuple：104 ビット程度であれば，毎秒 200 M 回以上の検索を実現可能であり [7]，高速性，検索のビット幅といった観点で問題ない．しかし，前述のとおり，TCAM は価格が通常のメモリに比べ高価であり，消費電力も大きい．更に，TCAM を用いた検索では，*N* ビットの条件が範囲指定されると，条件が 2N-1 の TCAM エントリに分割されて設定されるため [9]，メモリの利用効率が悪い．

5-1-6 QoS 制御技術

ルータの出力回線にその帯域以上のパケットが集中すると，すべてのパケットを出力回線より送信することができず，パケットの通信遅延や廃棄が発生するなど，通信品質（QoS）が劣化する．ルータの QoS 制御技術は，帯域，遅延時間，遅延ゆらぎ，廃棄率などの保証や，通信の優先度に応じた優先付け（例：優先クラスと非優先クラスの実現など）を実現するための技術である．

(1) ポリサ

ポリサは，クラシファイアにて識別されたフローごとにパケットの帯域を監視し，予め設定した帯域を超過した場合，パケットを廃棄したり，パケットの優先度を下げたりする機能である．このポリサの監視動作をポリシングと呼ぶ．ポリサは固定長のパケットである 53 バイトのセルを通信する ATM（Asynchronous Transfer Mode）により確立された．遵守／違反の判定アルゴリズムとしては，Continuous-state Leaky Bucket アルゴリズム（以下，リーキーバケットアルゴリズム）が用いられる [11]．なお，GCRA（Generic Cell Rate Algorithm），Virtual Scheduling Algorithm が用いられる場合もあるが，これらはリーキーバケットアルゴリズムと等価なアルゴリズムである．

リーキーバケットアルゴリズムは，穴の開いた深さをもった漏れバケツのモデルで説明される．予め設定した帯域でバケツから水が漏れ，セル到着時には 53 バイトに相当する水が蓄積される．バケツが溢れていない場合には，そのセルは遵守セル（Conforming Cell）と判定され，溢れている場合には違反セル（Non Conforming Cell）と判定される．

バケツの深さが L，単位時間当たりにバケツから漏れる水の量が 1，53 バイトに相当する水の量を I とする．k 番目のセルが到着するとその到着時刻 $t_a(k)$ と，直前の遵守セルが到着した時刻 LCT との差分（漏れた水の量）を計算し，蓄積水量 X より減算してその値を X' とする．もし，X' が負の値となった場合には，バケツが空であることを表す 0 に修正する．X' が正の値である場合には，バケツの深さ L と比較し，$L > X'$ の場合にはバケツから水が溢れるため，到着セルを違反セルと判定し，$L \leqq X'$ の場合には遵守セルと判定し，53 バイトに対応する水の量 I を X' に加算し，蓄積水量 X とする．

このアルゴリズムの特徴は L を備え，セルの到着時間の揺らぎを許容することにある．例えば，L が 0 であれば，セルの到着間隔が設定帯域より決まる時間より少しでも短くなると，即座に違反セルと判定する．一方，$L > 0$ とすると，セルの到着時刻がゆらいで早く到着したとしても，即座には違反セルと判定せずに遵守セルと判定する．しかし，帯域が長期にわたって設定した帯域を上回る場合には，蓄積水量 X が徐々に大きくなり，帯域超過を検出する．すなわち，リーキーバケットアルゴリズムはゆらぎによる短期の帯域の超過は許容しつつ，長期に渡る帯域超過を検出可能なアルゴリズムである．

多くのルータにおいては，リーキーバケットアルゴリズムを可変長のパケットに拡張したアルゴリズムが用いられる．その際，I の値をパケット長に応じて変化させればよい．

他の遵守／違反を判定する他のアルゴリズムとしては Jumping window や Sliding window がある．これらのアルゴリズムは，ある一定の時間（window）に入力する上限のセル数やバイト数を備え，この上限値を超えて入力するセルやパケットを違反セル，違反パケットと判定する．遅延時間のゆらぎを考慮すると，一定時間に入る上限のセル数，バイト数は，帯域か

ら決まるセル数，バイト数に一定量だけ余分なセル数，バイト数を加えた値となる．例えば，0.1ms の短期ゆらぎがあるパケットを 1 ms の window により帯域を監視する場合，このゆらぎを許容するのであれば，上限のバイト数は 1.1 ms の間に設定帯域で入力するバイト数となる．この際，遵守パケットと判定する帯域は最大で設定帯域の 1.1 倍となってしまう．この様に window による方式では，遅延時間のゆらぎを許容するのであれば，長期にわたる帯域超過をある程度許容する必要がある．しかし，リーキーバケットアルゴリズムと比較して実装が単純であり，精度の悪い帯域制限で十分なルータには本アルゴリズムが適している．

(2) スケジューラ

スケジューラは，出力回線ごとに備える複数の FIFO から読み出すパケットの送信順序を制御したり，FIFO から読み出すパケットの帯域を制御（スケジューリング）する機能部である．前者の制御を優先制御と呼び，後者の帯域制御をシェーピングと呼ぶ．シェーピングであれば出力回線ごとの FIFO は一つの場合もあるが，優先制御を実現する際には図 5･3 の出力制御部は複数の FIFO を備え，パケット振分部が優先度に応じた FIFO への蓄積を実施することとなる．

優先制御としては，PQ（Priority Queuing），CQ（Custom Queuing），WFQ（Weighted Fair Queuing）[12] などがある．PQ では，FIFO に絶対的な優先度が付与され，スケジューラは優先度の高い FIFO にパケットが存在する場合には，その FIFO より優先的にパケットを読み出して送信し，優先度の高い FIFO にパケットがない場合には，優先度の低い FIFO よりパケットを送信する．

PQ では，優先度の高いパケットの帯域が大きくなると，優先度が低いパケットの帯域が枯渇する．優先度の低い FIFO に対して一定帯域のパケット送信を確保可能とする優先制御の一つが CQ である．CQ では，それぞれの FIFO には送信するパケットのバイト数が設定される．スケジューラは設定されたバイト数分のパケットを FIFO より送信し，送信が完了すると別の FIFO の送信を開始する．このパケット送信を繰り返し，すべての FIFO からの送信が完了すると，再び最初の FIFO に戻り，同様のパケットの送信を実施する．CQ では，各々の FIFO に設定されるバイト数の変更により，各 FIFO から送信するパケットの送信帯域の比率が調整される．

送信帯域の比率を制御する他の優先制御として WFQ がある [12]．CQ が優先度ごとの FIFO からの送信を制御するのに対し，WFQ はフローごとの FIFO を備え，この FIFO からの送信を制御する．WFQ では，各 FIFO（あるいはフロー）に重み（weight）が割り当てられ，スケジューラはこの重みに比例した帯域でパケットを FIFO から出力する．

一方，シェーピングは設定帯域にパケットを制限して送信することをいい，シェーピングを実施する機能部をシェーパという．シェーピングとポリシングは，ともにパケットの帯域を設定値以下に抑える機能であるが，出力されるパケットの帯域は異なっている．

図 5･9 にポリシングとシェーピング前のパケットの帯域（メッシュ部分）と，ポリシングとシェーピングを適用した後のパケットの帯域（実線）の例を示す．ポリシングにおいては，設定帯域を越えた場合にはパケットが即座に廃棄されて帯域が設定帯域以下となるのに対し，シェーピングでは超過したパケットは FIFO に蓄積され，出力する帯域に余裕が有る際に出力される．ポリシングでは，パケットが廃棄されるため，トランスポート層のプロトコルと

して TCP（Transmission Control Protocol）が使用されている場合，TCP コネクションがスロースタート状態となり，ポリサに到着するパケットの帯域が設定帯域よりも極端に小さくなることがある．一方で，シェーピングでは，FIFO が枯渇しない程度の帯域超過であれば，パケットの廃棄が発生せず，スループットの極端な低下が抑止される．しかし，ルータに実装する際には，シェーパは FIFO を実現するメモリや帯域制御を実現する回路を備える必要があり，ポリサに比較して高価となる．

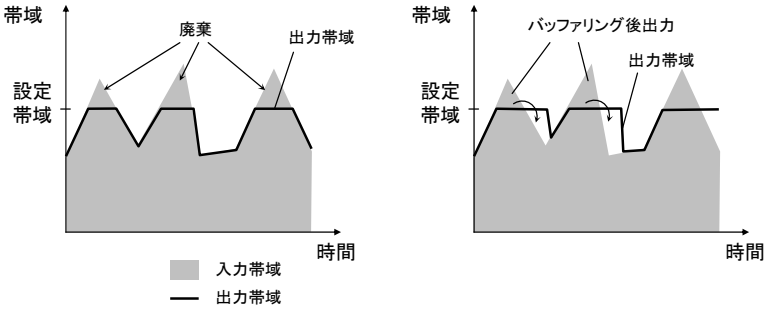


図 5・9　ポリシング（左図）とシェーピング（右図）の効果

シェーピングにおいて，帯域を制御するアルゴリズムとしては，トークンバケットアルゴリズムが一般的である．本アルゴリズムでは，トークンを蓄積するバケツを備え，一定時間に設定帯域に相当する一定量のトークンが蓄積される．スケジューラは，FIFO の先頭にあるパケットのバイト数分のトークンがバケツに蓄積されている場合には，そのパケットを出力し，相当するトークンを減算し，トークンが不足している場合には，蓄積されるまで待機する．

これまで，設定帯域が最大帯域となるシェーピングを説明してきたが，設定帯域が最低帯域であり，設定帯域でのパケット出力を保証しつつ，帯域に余裕が有る場合にはそれ以上の帯域でパケットを送信するシェーピングや，最低帯域と最大帯域を制御するシェーピングなどがある．

また，シェーピングと優先制御との組合せを実現するスケジューラも存在する．例えば，四つの FIFO を備え，最も優先度の高い FIFO より最大帯域を満たす範囲で優先的にパケットを出力し，残りの帯域に関しては CQ の出力アルゴリズムに従って出力するスケジューラなどである．遅延時間の制約が厳しい VoIP（Voice over IP）のパケットを最優先と，E-mail，Web 閲覧用のパケットなどのパケットを低優先に割り当てることで，VoIP のパケットの低遅延通信と，その他のパケットへの帯域割当が実現される．

(3)　バッファ制御（キューマネージメント）

出力回線へのパケットの集中が短期的であれば，出力制御部の FIFO にパケットが蓄積され，廃棄なしにパケットが出力回線より出力される．しかしながら，トラヒックの集中が長期間続くと FIFO が枯渇し，パケットが廃棄される．バッファ制御は，フロー間の優先度付けや回線帯域の利用効率を向上するために，FIFO が枯渇する前に前もってパケットを廃棄する制御である．

最も単純なバッファ制御は，FIFO に閾値までパケットが蓄積された際に，すべてのパケットを廃棄する Tail Drop である．優先度に応じて同一 FIFO 内に向かうパケットの閾値を変更することで，フローごとのパケット廃棄率が制御可能となる．

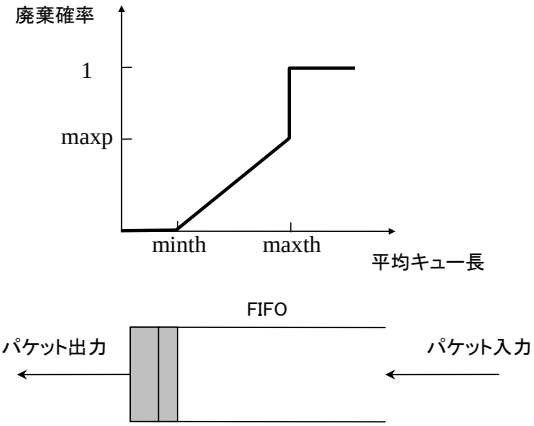


図 5・10　RED における平均キュー長と廃棄確率の関係

Tail Drop では，閾値に達するとバースト的なパケットの廃棄が発生するが，この際，TCP コネクションが一斉に送信帯域を減少させるスロースタートとなる Global Synchronization が発生し，帯域の利用効率が悪化してしまう問題がある．この Global Synchronization を回避するバッファ制御が RED（Random Early Detection）[13] である．RED では，**図 5･10** に示すように FIFO が溢れる前に予めパケットが確率的に廃棄される．平均キュー長が minth 以下であれば，パケットの廃棄確率は 0 であるが，minth 以上となると廃棄確率が正となり，キュー長の増加と供に廃棄確率は増加する．maxth で廃棄確率が maxp となり，それ以上では廃棄確率は 1 となる．RED は，パケットの確率的な廃棄により，すべての TCP コネクションがスロースタートに陥ることを防止し，Global Synchronization を抑止し，回線帯域の利用効率を向上する．

現在のルータには，RED の廃棄確率のプロファイルを複数備えて優先度付けを実現する WRED（Weighted RED）が実装されているのが一般的である．

5-1-7 高速スイッチ技術（高速スイッチング技術）

スイッチは，複数の入力ポートと出力ポートを備え，宛先に従いパケットを出力ポートにスイッチするルータの機能部である．分散型アーキテクチャによるルータでは，図 5･4 に示す様にスイッチカードとして独立して実装されるのが一般的である．一方，集中型アーキテクチャによるルータでは，パケット処理エンジンとパケットバッファ用メモリが，スイッチを実現している．スイッチはルータの高速化に伴い，その高速化が求められており，高速スイッチ技術も進化してきた．以下では，代表的なスイッチの特徴を説明し，そのスイッチを組合せて高速化する手法について説明する．

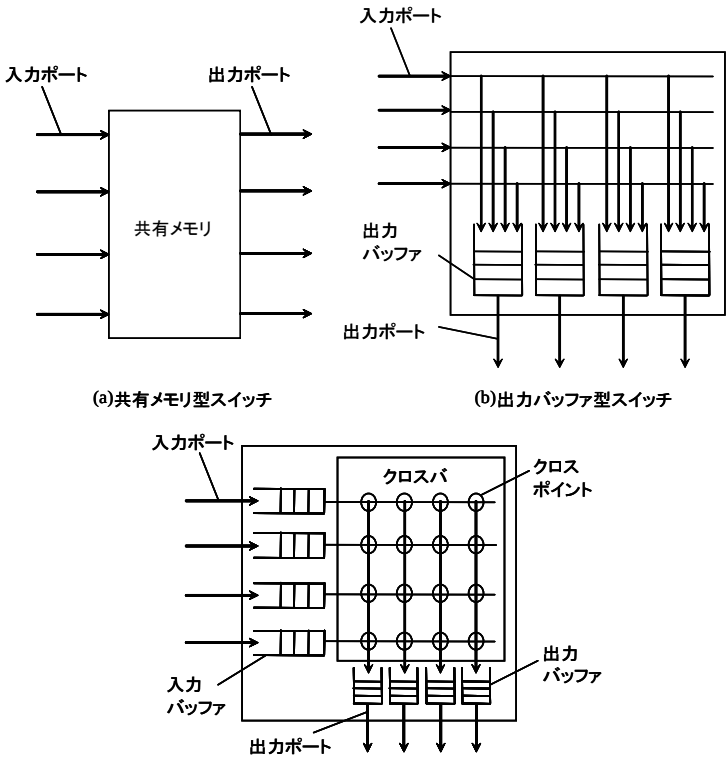


図 5・11　代表的な要素スイッチの構成

表 5・1　代表的な要素スイッチの特徴

スイッチ方式	共有メモリ型	出力バッファ型	入出力バッファ型 クロスバスイッチ
方式概要	全入出力ポートが，メモリを共有	出力ポートごとメモリを共有	バッファ間のクロスバ，もしくは，出力数個のセレクタで交換
メモリ スループット	(N+M) R	(N+1) R	2 R
メモリ数	1	M	N+M
総メモリ スループット	(N+M) R	(N+1) MR	2NR（入力） 2MR（出力）
特　長	メモリを有効利用可能	共有メモリ型よりメモリスループットが低くてよい	容量に対するスケーラビリティが高い

代表的なスイッチ（本章では要素スイッチと呼ぶ）の構成と特徴を**図 5･11** 及び**表 5･1** に示す．共有メモリ型スイッチは，全入出力ポートで共有メモリを共有し，入出力ポートからのパケットを時分割してメモリに read/write することでパケットを交換する．全入出力ポート

共通のメモリを保持するため，メモリ容量を有効利用できる．しかし，スイッチを大容量化するには，このメモリの高速化が要求される．例えば，N 入力，M 出力で入出力ポートの帯域を R とすると，メモリには (N+M) R ものスループットが要求される．そのため，高速化（R > 10 Gbit/s），多ポート化（8 ポート以上）への対応が難しい．

一方，出力バッファ型スイッチは，出力ポートごとにバッファメモリを共有し，全入力の書込み，1 出力ポートへの読出しを，時分割で行うスイッチ方式である．メモリが出力ポートごとに分散しているため，メモリの必要なスループットは共有メモリ型より小さく(N+1) R となり，共有メモリ型と比べれば大容量化を達成しやすいが，共有メモリ型スイッチと同様に，高速なスイッチへの適用が難しい．

高速なスイッチを実現する方式として開発された入出力バッファ型クロスバスイッチは，全入力／出力ポートを接続するクロスポイントによりクロスバを構成し，データを交換する．複数の入力ポートからのパケットが同一の宛先を目指す場合やクロスバの出力先ポートが出力できない場合（例えば，スイッチの後段にあるパケット処理部の処理が終了していない場合など）に，パケットを待機させるバッファをクロスバの前段と後段に備える（それぞれを入力バッファ，出力バッファと呼ぶ）．この方式は，バッファメモリへの要求スループットが 2 R であるので，共有メモリ型，出力バッファ型スイッチに比べて容量のスケーラビリティが高く，多くの高速ルータに採用されてきた．

最近では，さらなる高速回線（40 Gbit/s 以上の回線）の収容や，高速回線を多回線収容したルータへの要求の高まりにより，前述の要素スイッチを複数組み合わせることで更なる大容量化を実現するスイッチが登場している．**図 5･12** にそのスイッチ方式を示す．

集中型多段結合網 [14] は，入力ポート及び出力ポートを収容する交換部（要素スイッチ）を並列に配列し，そのスイッチ群を多段に構成することで，入出力ポートの数を増加させる．ポート数の増加に伴い，ルータとして多くの入出力回線のサポートが可能となる．分散型単段網 [14] は，並列に並べた交換部と，その交換部にパケットを振り分ける（または，パケットを固定長のセルに分割した後にセルを振り分ける）分散部，各出口に整列部を備える．同一の出力ポートに出力されるパケット群（または，セル群）は異なる交換部を経由して目的の出口（整列部）に転送されることがあるため，各交換部のスイッチング時間が異なると，パケットやセルの順序が入れ替わる場合がある．そのため，交換部の後段に配した整列部によってパケット群（または，セル群）の順序を元通りに整列したのち，出力ポートから出力する．

この方式は，入力ポートからのパケットを複数の要素スイッチで並列処理するため，入力ポートの高速化を実現できる．入力ポートの帯域がルータの入出力回線の最大帯域を決めるため，分散型単段網のスイッチは，集中型多段結合網より高速な入出力回線のサポートが可能である．分散型多段結合網 [14] は，集中型多段結合網と分散型単段網を組み合わせた方式である．分散型単段網に使用される交換部（要素スイッチ）に代わり集中型多段結合網によるスイッチを使用することで，高速回線の収容と多くの入出力ポートのサポートを実現する．

これらのスイッチ方式は用途に応じて使い分けられることとなる．例えば，高速回線（40 Gbit/s 以上の回線）を少数収容（例えば，4 回線）するルータには分散型単段網が，高速回線の回線を多数収容（例えば，128 回線）するルータには分散型多段結合網が適用される [15]．

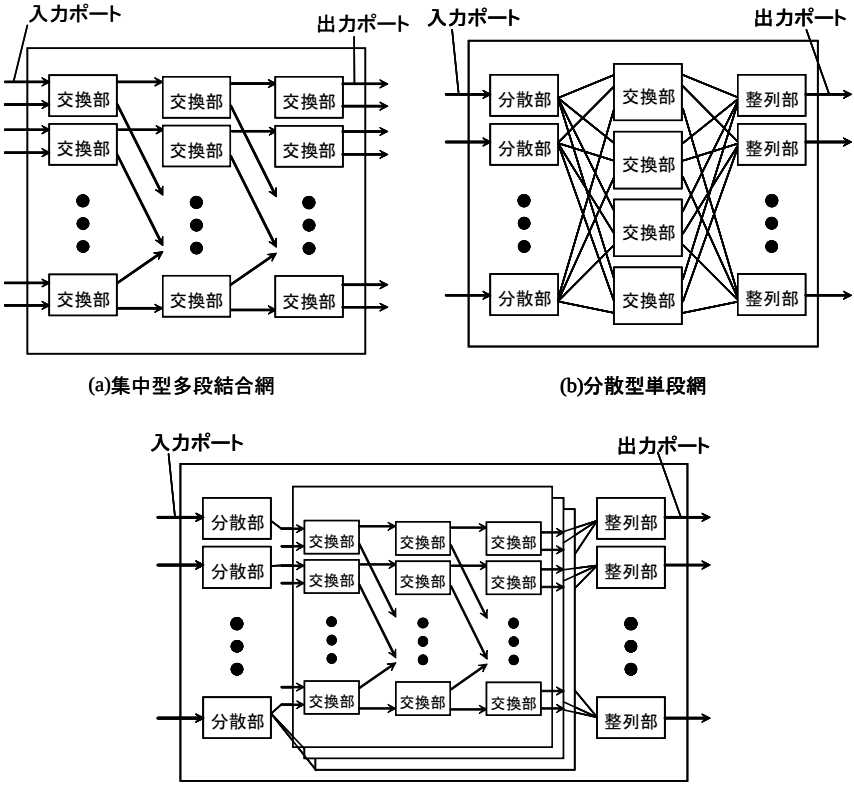


図 5・12　要素スイッチを組み合わせたスイッチ

■参考文献

1) “発掘！ルーター開発物語,” 日経 NETWORK, 2005 年 1 月号, 日経 BP 社.
2) T. Yazaki, T. Kanetake, S. Akahane, Y. Sakata, K. Sugai, and H. Yano, “High-Speed IPv6 Router/Switch Architecture,” Proceedings of SAINT2004, Jan. 2004.
3) http://bgp.potaroo.net/as2.0/bgp-active.html
4) Tzi-cker Chiueh and Prashant Pradhan, “High-Performance IP Routing Table Lookup Using CPU Caching,” Proceedings of INFOCOM99, pp.1421-1428, 1999.
5) D. R. Morrison, “Patricia-practical algorithm to retrieve information coded in alphanumeric,” Journal of ACM, 15(4):514-534, Jan. 1968.
6) A. McAuley and P. Francis, “Fast routing table lookupusing CAMs,” Proceedings of IEEE INFOCOM '93, vol.3, pp.1382-1391, Mar. 1993.
7) http://www.idt.com/products/getDoc.cfm?docID=18458762
8) 阿多信吾, 黄恵聖, 山本耕次, 井上一成, 村田正幸, “低コスト・低消費電力 TCAM における効率的なルーティングテーブル管理法,” 信学技報, vol.107, no.443, NS2007-120, pp.7-12, Jan. 2008.

9) David E. Taylor, “Survey and Taxonomy of Packet Classification Techniques,” ACM Computing Surveys, vol.37, no.3, Sep. 2005.
10) V. Srinivasan, G. Varghese,S. Suri, and M. Waldvogel, “Fast and scalable layer four switching,” Proceedings of ACM SIGCOMM '98, Sep. 1998.
11) The ATM Forum Technical Committee, Traffic Management Specification Version 4.1 AF-TM-0121.000, Mar. 1999.
12) S. Golestani, “A Self-Clocked Fair Queueing Scheme for Broadband Applications,” Proc. of INFOCOM94, pp.636-646, 1994.
13) S. Floyd and V. Jacobson, “Random Early Detection Gateways for Congestion Avoidance,” IEEE/ACM Transactions on Networking, 1(4):397-413, Aug. 1993.
14) H. J. Chao, “Next generation routers,” invited paper, IEEE Proceeding, vol.90, no.9, pp.1518-1558, Sep. 2002.
15) http://www.cisco.com/en/US/prod/collateral/routers/ps5763/prod_brochure0900aecd800f8118.pdf

■5 群 － 4 編 － 5 章

5-2 L2 スイッチシステム

（執筆者：矢崎武己）［2009 年 5 月 受領］

5-2-1 L2 スイッチの基本動作

L2 スイッチシステム（以下，L2 スイッチ）はデータリンク層のアドレスによりフレームを転送（スイッチング）するネットワーク機器であり，1990 年に開発された．L2 スイッチの論理構成は，5-1-2 項の図 5･3 に示すルータのルーティング処理部が，**図 5･13** に示す FDB（Filtering Data Base）を備えたフォワーディング処理部となる．FDB は MAC アドレス，VLAN（Virtual LAN）ID に対応して回線番号とエージング時間が記載されたエントリを複数備えたテーブルである．

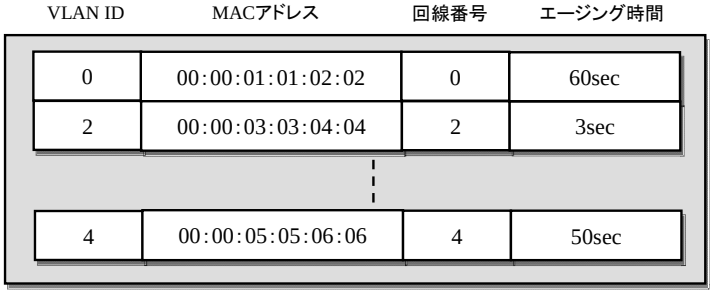


図 5・13　FDB の論理構成

フレームが入力された際のフォワーディング処理部の主な処理は，(1)アドレス学習のための FDB の検索と(2)フレーム転送のための FDB の検索の二つである．フレームが入力されると，まず，(1)アドレス学習のための検索を実施する．本検索では，フォワーディング処理部はヘッダ内の送信元 MAC アドレスと VLAN ID とを，FDB 内の値と一致比較する．一致したエントリが存在する場合，このフレームが入力した入力回線の番号を FDB 内の回線番号に上書きし，エージング時間を更新する．一方，一致するエントリが存在しない場合には，入力したフレームの送信元 MAC アドレス，VLAN ID，入力回線の番号，更新したエージング時間の値を FDB に新規に登録する．この学習動作により，MAC アドレスと VLAN ID に対応する回線番号を学習する．また，エントリ内のエージング時間は，MAC アドレスが学習されてから一定時間の間に対応するフレームが入力されない場合に，そのエントリを削除するエージングに使用される．

(2)フレーム転送のための検索においては，フォワーディング処理部はヘッダ内の宛先 MAC アドレスと VLAN ID とを，FDB 内の値と一致比較する．一致する MAC アドレスと VLAN ID が存在する場合には，対応する回線番号をフレーム転送すべき出力回線の番号と判定する．一致する MAC アドレスと VLAN ID が存在しない場合には，そのフレームが入力した回線以外のすべての回線をフレーム転送する出力回線と判定する．

なお，本章では，フレームのヘッダに VLAN ID が格納されるタグ VLAN[1)] を使用した場合を前提としたが，ポート VLAN[1)]，プライベート VLAN などが使用された場合にも，L2 スイ

ッチ内で VLAN ID に相当する情報が判定されるため，FDB の検索は同様に実施される．

5-2-2 FDB 検索技術

FDB の検索は，ルーティングテーブルの検索と似ているが，ルーティングテーブルの検索が，mask が設定される Longest Prefix Match であるのに対し，FDB 検索は条件に mask が設定されない Exact Match である．Exact Match を実現する検索方式として一般的な 3 方式を説明する．以下では簡単のため，VLAN を使用せず，VLAN ID を除く MAC アドレスを用いた場合について説明するが，VLAN ID を用いた場合も同様の検索が可能である．

(1) ハッシュ（Hash）

本方式を，**図 5･14** を用いて説明する．本方式の FDB は複数のテーブル領域に分割され，MAC アドレスはそのハッシュ値から決まるテーブル領域に格納される．アドレス学習のための検索時には，送信元 MAC アドレスのハッシュ値が計算され，対応するテーブル領域の MAC アドレのみが一致比較され，フレーム転送のための検索時には，宛先 MAC アドレスにて同様の処理が実施される．*N* のテーブル領域に分割されている場合，検索すべき MAC アドレスは平均して 1/*N* となり，高速化が実現される．

一方で，ハッシュ値が完全に分散せず，特定のテーブル領域に格納すべき MAC アドレスが偏ってしまうことが考えられる．特定のテーブル領域がすべて使用され，他のテーブル領域に空きがあるにもかかわらず MAC アドレスを登録できない状況が発生するため，テーブルを実現するメモリの使用効率の劣化が発生する．この問題を解決する方法として，オーバーフローしたテーブル領域を確保する方式，ハッシュ関数を動的に変更する方式などがある [2].

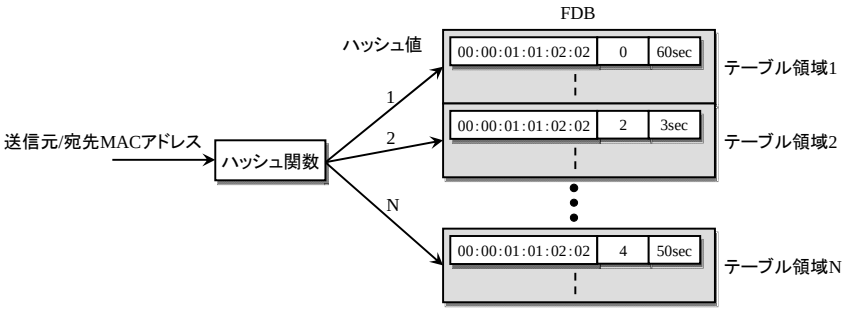


図 5・14　FDB の高速化手法（ハッシュ）

(2) バイナリーサーチ（Binary Search）

本方式では，FDB 内の MAC アドレスは降順あるは昇順にソートされ，その真ん中の MAC アドレスとフレームの宛先/送信元 MAC アドレスとの大小が比較される．その大小に応じて，検索対象が半分に限定され，限定された FDB 内の MAC アドレスで同様の処理が繰り返される．

N 個の MAC アドレスの条件が FDB に設定される場合，$\log_2 N$ のオーダの一致比較をすればよく，本方式は *N* が大きくなると *N* に比例した性能となるリニアサーチに比べ高速である．

また，ハッシュを使った方式と異なり，メモリ効率が劣化することもない．一方で，MAC アドレスを追加するためには，ソートを再度実施する必要があり，ソート時間中は検索が停止するといった問題が発生する．この問題を解決するために，予備用の FDB を備え，追加削除時には予備用の FDB のソートを実行し，ソート後にこの FDB を検索用の FDB に切り替える方式がある．しかし，本方式では，FDB が二つ必要であり，ハッシュの場合と同様にメモリの使用効率が劣化する課題がある．

(3) CAM (Contents Addressable Memory)

本方式は，ルーティングテーブルの検索にて記載した方法と同様の手法である．ただし，FDB には mask は設定されないため，mask を設定可能な TCAM ではなく，通常の CAM で十分である．mask 設定用のメモリ領域が不要であるため，メモリ効率は 2 倍となる．

5-2-3 高可用化技術

L2 ネットワークを高可用化する技術として，ネットワークの経路を冗長化する IEEE 802.1D に規定されるスパニングツリープロトコルがある．しかしながら，障害発生時に最大で 50 秒もの間，通信断が発生する場合がある．また，学習する MAC アドレスの数が多くなると動作が不安定となるといった課題も指摘されており [3]，これらの課題を解決する高信頼化技術が開発されている．以下では，特に L2 スイッチの高可用化技術としてリンクアグリゲーションとリング技術を説明する．

(1) リンクアグリゲーション

リンクアグリゲーションは，複数の回線を仮想的な一つのリンク（集約リンク）とする機能である（**図 5･15**）．ベンダ独自の方式が実装されてきたが，IEEE 802.3ad として標準化された．複数の物理回線からなるリンクアグリゲーショングループ（Link Aggregation Group：LAG）が構成され，分配部がこのグループを構成する回線にフレームを振り分けて出力する．リンクアグリゲーショングループの構成は，LACP（Link Aggregation Control Protocol）により自動に，あるいは，管理者により手動で行われる．

本機能の利点は主に 2 点である．1 点目は，高可用化である．グループを構成する回線が障害となった場合，1 回線の通信が可能であれば，通信断が回避される．

L2 スイッチにルーティング機能を備えた L3 スイッチのハイエンド製品は，5-1-3 項の図 5･4，図 5･5 に記載した構成を採用しているケースが多いが，この様な製品においては，リンクアグリゲーショングループを構成する回線が複数のインタフェースカードに跨ることが可能である．そのような設定をしておけば，インタフェースカードが最低でも一つ通信可能であれば，通信断が発生しない．

二点目の利点は回線帯域の柔軟な増加である．イーサネットであれば，10 Mbit/s → 100 Mbit/s → 1 Gbit/s → 10 Gbit/s と，回線帯域を増加させる際には 10 倍の回線に入れ替える必要がある．しかし，本技術を用いることで各回線の n 倍（n = 1, 2 …）の帯域を備える回線を実現可能である．しかし，分配部が各回線に均等にフレームを分散しないと，フレームの偏りが発生し，特定の回線にはフレームが送信されず，回線帯域を効率良く活用できないといった問題がある．回線へのパケット振分けは，宛先／送信元 MAC アドレス，宛先／

送信元 IP アドレス，宛先／送信元ポート番号などのヘッダフィールドに基づく振分が可能であるが，特定の回線にフレーム出力が集中しないフィールドを用いて振分けを実施することが回線帯域の有効活用の観点から肝要となる．

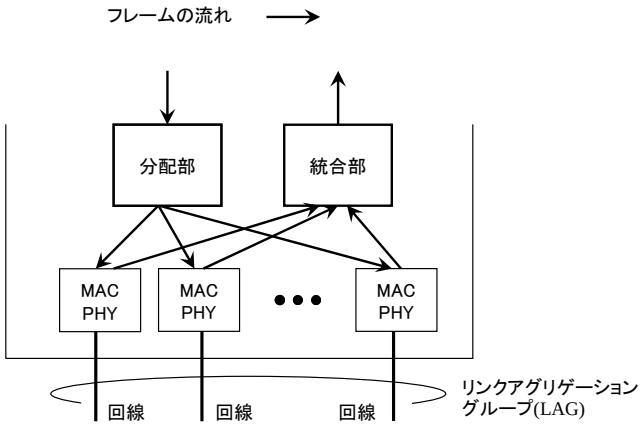


図 5・15　リンクアグリゲーションを実現する機能部のブロック図

(2)　リング技術

リング技術は，ノードをリング状に構成したリングトポロジーのネットワークにて適用される高可用化技術である．ネットワーク機器ベンダにより独自に開発されてきたが，一部の技術が標準化されている [4]．リングトポロジーのネットワークにおいては，ノード数と同数のケーブルで冗長経路を構成可能であり，他のトポロジー（例えば，スタートポロジー）と比べて，ケーブルコストや回線料金を安価に抑えつつ，ネットワークの高信頼化が実現される．例えば，**図 5･16** に記載したネットワークにおいては，五つのノードと五つの回線（ケーブル）により，N1 から N2 に至る経路：P1 及び P2 の経路が設定可能である．

本技術を適用したネットワークは，1 台のマスターノードと複数のトランジットノードから構成される．マスターノードは，ヘルスチェック用の制御フレームを定期的に送信し，一方の回線の通信をブロックする．図 5･16 では，マスターノードの回線 2 が論理的にブロックされ，P1 の経路はフレームの通信に使用されない状態となっている．

ヘルスチェック用の制御フレームが一定時間内にマスターノードに到達している際には，ネットワークが正常と判断される．一方，フレームが到達しない場合には障害発生と判断され，逆周りのフレーム転送が実施されるように，リングを構成するノードの FDB がクリア（flush）される．また，マスターノードのブロックしていた回線が通信可能となり，迂回経路によるフレームの転送が実施される．例えば，図 5･16 のネットワークにおいて，P2 の経路によりフレームが転送されている場合に，トランジットノード 2 と 3 間の障害が発生すると，FDB のクリアによりトランジットノード 1 がマスターノード側にもフレームを送信し，マスターノードが受信したフレームをトランジットノード 4 に転送することで，迂回経路 P1 によるフレーム転送が実現される．

リング技術を適用したネットワークにおいては，比較的単純なネットワーク構成による冗長経路の切り替えが，単純なプロトコルにより実施され，安定した動作と高速な経路切り替えが実現される．現在では，切り替え時間が 1 秒以下の製品もある．

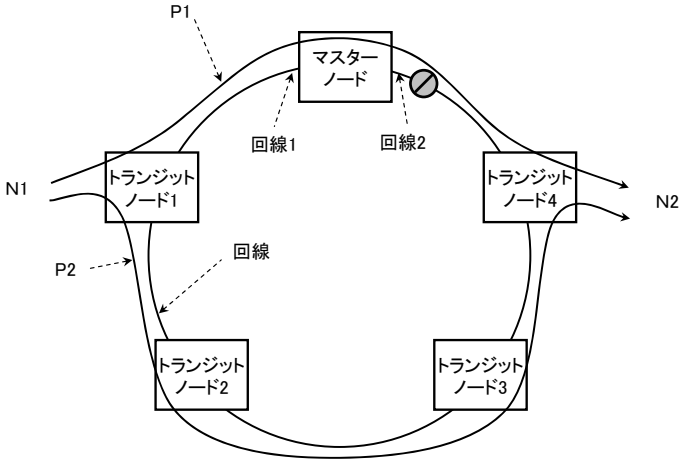


図 5・16　リングトポロジーのネットワーク

■参考文献

1) IEEE 802.1Q - Virtual LANs, Nov. 2006.
2) Rich Seifert 著, 間宮あきら訳, “LAN スイッチング徹底解説,” 日経 BP 社.
3) “スイッチに 10 倍詳しくなる！,” ネットワークワールド 2005 年 11 月号, アイ・ディ・ジー・ジャパン.
4) IETF RFC3619, “Extreme Networks' Ethernet Automatic Protection Switching (EAPS) Version 1,” Oct. 2003.