マルチレイヤスイッチ MultiFlow 1000

Multilayer Switch MultiFlow 1000

UDC 654.022 : 681.39

小畑時男	Tokio Obata	情報通信事業本部　情報システム事業部　技術部
川藤光裕	Mitsuhiro Kawafuji	情報通信事業本部　情報システム事業部　技術部
木村幸泰	Yukiyasu Kimura	情報通信事業本部　情報システム事業部　技術部
東山　満	Mitsuru Higashiyama	情報通信事業本部　情報システム事業部　技術部
関水宗樹	Muneki Sekimizu	情報通信事業本部　情報システム事業部　技術部

1 まえがき

インターネット／イントラネットの普及とパーソナルコンピュータ（PC）の急速な高性能化，また，アプリケーションのマルチメディア化に伴い，PCなどを接続するLANの高速化が求められるようになった。この要求を満たすため，従来の10 Mbit/sの帯域を共有するシェアード方式から，通信する機器同士だけを論理的に接続して複数の10 Mbit/sの帯域を同時に利用できるようにしたスイッチング方式が広まってきている。さらに，帯域そのものを100 Mbit/sに広げた100VG-AnyLANや100Base-TXなどの方式も使われるようになった。

また，スイッチング方式によりVLAN（Virtual LAN: 仮想LAN）という考え方が生まれた。これによると，PCなどの物理的な設置場所に制約を受けることなくグループ化することが可能になる。すなわち，設置場所にとらわれることなく，用途などにより柔軟にLANを構成したり，PCなどのネットワーク機器の移動にも柔軟に対応できる。VLANによって複数のLANを構築すると，そのLAN間の通信を行なう手段が必要になる。もちろん，このLAN間の通信手段にも高速性が求められる。

今回，我々は前記の要求に応えるLAN機器MultiFlow 1000を開発した。本装置は，インタフェースが100 Mbit/sに対応したスイッチング方式のハブであり，VLAN間通信のためにハードウェアによる独自の高速ルーティング機構を有している。

2 基本構成

コンピュータネットワークは，OSI参照モデルとして知られる7層に論理的に分割できる（図1）。これらの層のうち，通常のスイッチングハブでは第2層を扱ってスイッチングを行なう。本装置は，この第2層だけではなく，上位の第3層も扱うマルチレイヤスイッチングハブである。第3層のスイッチング処理は，従来のルータのそれと内容は同じである。なお，インターネット／イントラネットで使われているTCP/IPではIPが第3層に相当する。本装置では，第2層と第3層を高速に処理するために，それぞれ専用のLSIを開発し，採用している。第2層を処理するLSIは2つで構成され，それぞれMACSW（Media Access Control layer SWitch）とASE（Address Search Engine）と呼び，第3層を処理するLSIをRE（Routing Engine）と呼んでいる。詳細は後述する。

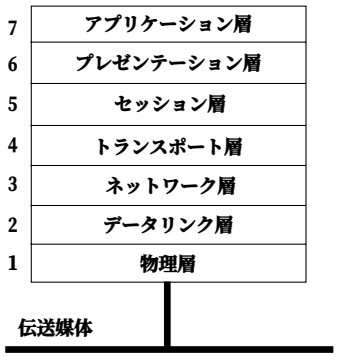

図1　OSI参照モデル
OSI reference model

本装置は，基本筐体内にシステムの制御部を有し，その他のインタフェース部や第3層スイッチ部はプラグインのモジュールとして用意する。インタフェースとしては100 Mbit/sの100Base-TXモジュール，100Base-FXモジュール，100VG-AnyLANモジュールがあり，100Base-TX モジュールは10

Mbit/sの10Base-Tにも対応している。また，各インタフェースモジュールはモジュール1つあたりインタフェースを4つ有している（100Base-FXは2つ）。さらに，第3層スイッチモジュールを含めて，すべてのモジュールは活線挿抜に対応している。したがって，モジュールの追加や交換時に装置の電源を切る必要がなく， LANを停止させることなく装置のメンテナンスをすることが可能である。

図2　装置外観
External View of Multi Flow 1000

これらの機構を内蔵した本装置の外観を図2に示す。

3 第2層スイッチ

第2層でのスイッチングは専用のLSIを開発し，高速処理とVLAN対応を実現している。

3.1 設計方針

第2層スイッチング部は以下の方針で設計を行った。

1. スイッチング方式は，異常フレームの廃棄が可能でセキュリティ上有利なストアアンドフォワードとする。
2. スイッチングを高速に行うための2種類の専用LSIを開発する。目標性能は，

   バス性能: 1.44 Gbit/s以上
   フォワーディング性能: 1.67 M packet/s以上
   フィルタリング性能: 各インタフェースのワイヤスピード

   とする。
3. VLAN方式はIEEE802.1Q標準に準拠する。したがって，設定はポート単位で行なう。

3.2 設計の要点

各インタフェースモジュールで受信したパケットから送信元アドレス，宛先アドレスの抽出，アドレスデータベースへの問い合わせ，装置内部のデータ転送バスへの送信，データ転送バスからの受信などの機能をもつ専用LSI(MACSW)と，アドレスデータベースの登録・検索を高速に行う専用LSI(ASE)を開発した。

MACSWとASEを組み合わせて使用することにより以下の機能を実現している。

- ダイナミックフィルタリング
- スタティックフィルタリング
- VLAN
- 統計情報の収集，カウント
- カスケード接続

3.2.1 ダイナミックフィルタリング

転送パケット中の送信元アドレスとポート番号の組合わせを登録し，アドレスデータベースを作成することにより，自ポート内に接続された端末宛のフレームを，他のポートに送らないようにして，各ポートに接続されているネットワークの負荷を軽減する。

3.2.2 スタティックフィルタリング

送信先アドレスと宛先ポート（複数指定可能）の組合わせを予めアドレスデータベースに登録しておき，転送パケット中の宛先アドレスがこの中にあれば，指定されたポートに無条件で転送する。

3.2.3 VLAN

本装置に接続されたネットワーク同士を，ポート単位で論理的なグループに分け，その中でのみ通信が行えるようにしたもので，これによりブロードキャストドメインを分割することが可能となる。ネットワーク全体のトラフィックの低減によりトータルパフォーマンスを向上させるとともに，グループ間でのアクセス制限により，セキュリティを高めることができる。

3.2.4 統計情報の収集，カウント

ネットワーク全体の管理を行うため，各ポートから入ってくるパケットをチェックし，データ長，パケット数，エラー情報などのデータを収集，カウントする。

3.2.5 カスケード接続

本装置を複数台カスケード接続するとき，そのポートを流れるフレームにVLAN情報を付加することによりVLANを共有することが可能となる。

3.3 LSIの実装

MACSW LSIはシステム制御部と第3層スイッチモジュール，およびモジュール，および，すべてのインタフェースモジュールに実装される。インタフェースモジュールではインタフェースごとにMACSW LSIを実装する。すべてのMACSW

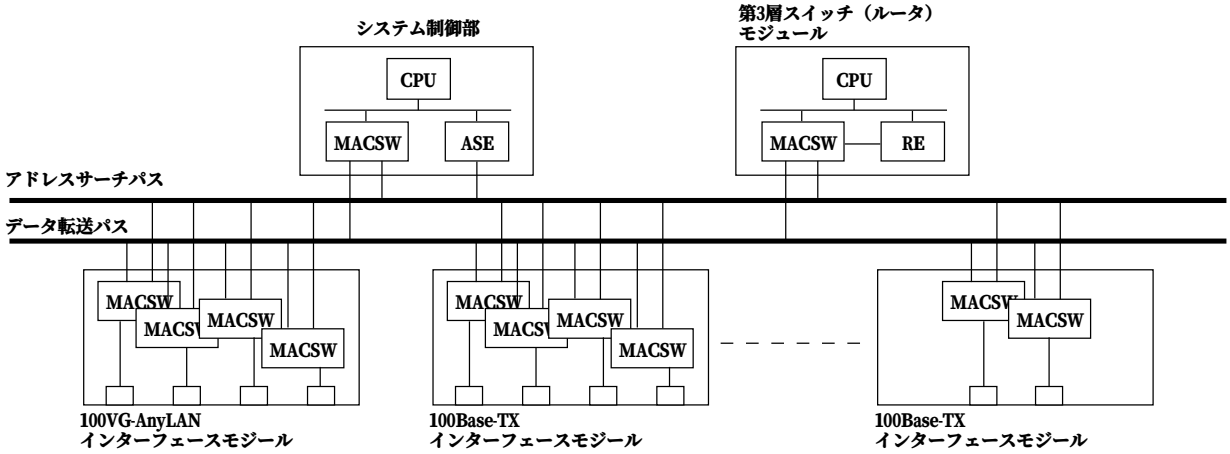

図3　内部構成（概略）
Block diagram（over view）

LSIは装置内バスで接続される。

ASE LSIはシステム制御部に実装し，装置内すべてのMACSW LSIがアドレスデータベース検索のためにアドレスサーチバスを通じてこのASE LSIをアクセスする。

内部構成の概略を図3に示す。

4 第3層スイッチ

前述のように，スイッチングハブを用いてVLANを構成した場合，そのVLAN間の通信手段が必要になる。本装置では，VLAN間通信手段を第3層スイッチモジュール（ルータモジュール）として内蔵できる。

4.1 設計方針

VLANという概念が生まれた当初は，VLAN間通信を行なうために従来のルータを用いる方法が取られた。しかし，この方法では，VLAN間通信の速度がルータのインタフェースの速度，すなわち，10 Mbit/sや100 Mbit/sといった速度の制限や，ルータを接続するためにインタフェースのポートを占有されるという制限を受ける。また，従来のルータでは，その処理内容の複雑さから処理はソフトウェアで行なわれていたため，この面からも速度が大きく制限されてしまう。

本装置は高速スイッチングを目指した装置であり，VLAN間通信も当然高速でなければならない。したがって，次のような方針で設計を行なった。

1. インタフェース速度による制限を受けないように，装置本体の高速内部バスに直結する内蔵モジュールとする。
2. 処理そのものを高速にするためにハードウェアで処理を行なう。処理目標時間は2μsとする（最小パケット）。
3. 業界標準に準拠し，特殊なプロトコルなどは用いない。

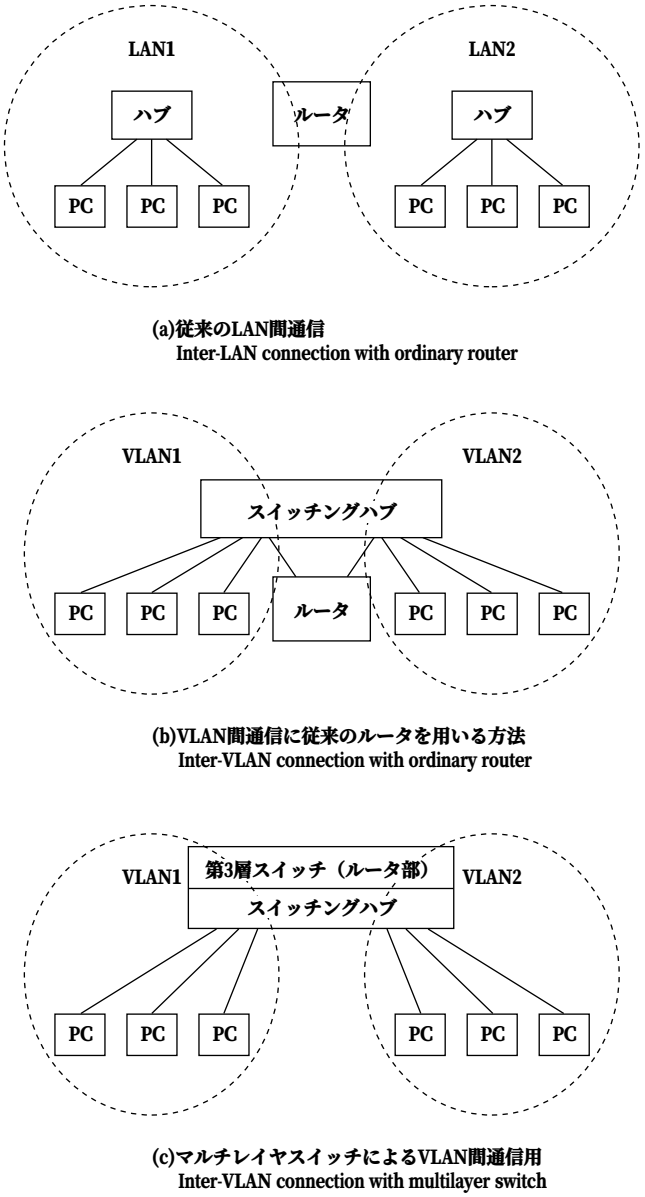

図4　LAN間通信の形態
Inter-LAN communication forms

LAN間通信の形態を図4（a）～（c）に示す。本装置は，上述のように（c）の第3層スイッチを装置内に内蔵した形態である。

このように，VLAN間通信を行なう第3層スイッチを装置に内蔵することによって，高価なルータを別途用意する必要がないということも大きな利点である。

4.2 設計の要点

4.2.1 処理対象の絞り込み

ネットワークの第3層に当たるプロトコルには，IP（TCP/IP），IPX（NetWare），AppleTalkなど複数のものがある。しかし，インターネット／イントラネットが普及した現在，LANを流れるトラフィックのほとんどはIPである。そこで，ハードウェアによる高速処理はIPをターゲットとし，IPX，AppleTalkはソフトウェアで処理する。

しかしながら，IPだけに限っても従来のルータで行なっていた処理は非常に複雑であり，すべての機能をそのままハードウェア化することは困難である。実際のトラフィックの内訳を考慮し，一般的なものをハードウェア処理，例外的なものを従来通りソフトウェアで処理するという方法が現実的な実装であろう。本装置では，次の条件をすべて満たすパケットをハードウェアで処理を行なう対象として絞り込んだ。

- EthernetⅡフレームフォーマット
- ユニキャスト
- IPデータグラム
- IP version 4
- IP optionsフィールドがない（IP IHLが5である）
- IP TTL（Time To Live）が1より大きい
- IP Header Checksumによるチェックの結果が正常である
- Destination IP Addressがルーティングテーブルに存在する
- Destination MAC AddressがARPテーブルに存在する

ファイル転送やWWWのデータなどのように転送するデータが多く，かつ，連続して通信が行なわれる可能性が非常に高いパケットはこれらの条件を満たす。逆に，これらの条件を満たさないパケットは転送データ量が少なかったり，例外的なものである。つまり，これらの条件を満たすパケットを高速に処理することによって実効的に非常に高い性能を得ることができる。

なお，これらの条件はパケットのEthernetヘッダとIPヘッダを参照することで判定できる（図5，図6参照）。

4.2.2 E-ARPテーブルの開発

DSTMAC ADRS	SRCMAC ADRS	TYPE	IP HEADER	DATA	FCS

DST: Destination
SRC: Source
ADRS: Adress
FCS: Frame Check Sequence

図5 Ethernet Ⅱフレームフォーマット
Frame format of EthernetⅡ

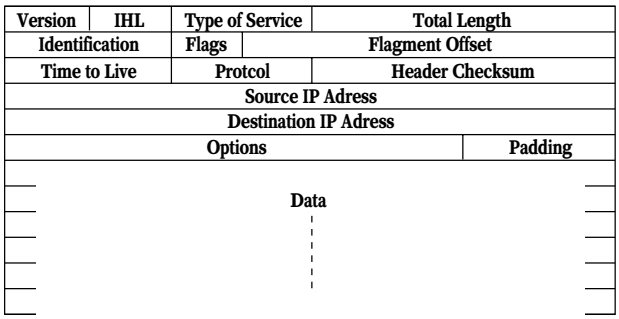

図6 IPヘッダ
IP header

ソフトウェアによる従来のIPルーティング処理は，ルーティングテーブルとARP（Address Resolution Protocol）テーブルという2つのテーブルを使って行なわれる。例えば，図7のLANの場合，Router 1のルーティングテーブルとARPテーブルはそれぞれ表1，表2のようになる。実際のルーティング処理は，これらの表を用いて次のような手順で行なわれる。

1. ルーティングするIPデータグラムから，ディスティネーションIPアドレス（DST IP ADRS）を取り出す。
2. ルーティングテーブルからDST IP ADRSにマッチするエントリを検索する。
   (a) ルーティングテーブルのあるエントリのNetmaskとDST IP ADRSとの論理積を求める。
   (b) その結果とそのエントリのDestinationを比較する。
   (c) 上記を各エントリについて行い，一致する物を求める。複数一致するものが得られた場合には，それらの中でMetricの最も小さいものを採用する（ベストマッチ）。
3. ARPテーブルからMACアドレス(DST MAC ADRS)を求める。この際，ルーティングテーブルの検索の結果，ゲートウェイが存在すればそのMACアドレスを，存在しなければDST IP ADRS自身のMACアドレスを求める。
4. ルーティングするIPデータグラムのIPヘッダの再構成（TTLの減算とChecksumの再計算）を行う。
5. DST MAC ADRSを用いてEthernetフレームを再構築する。

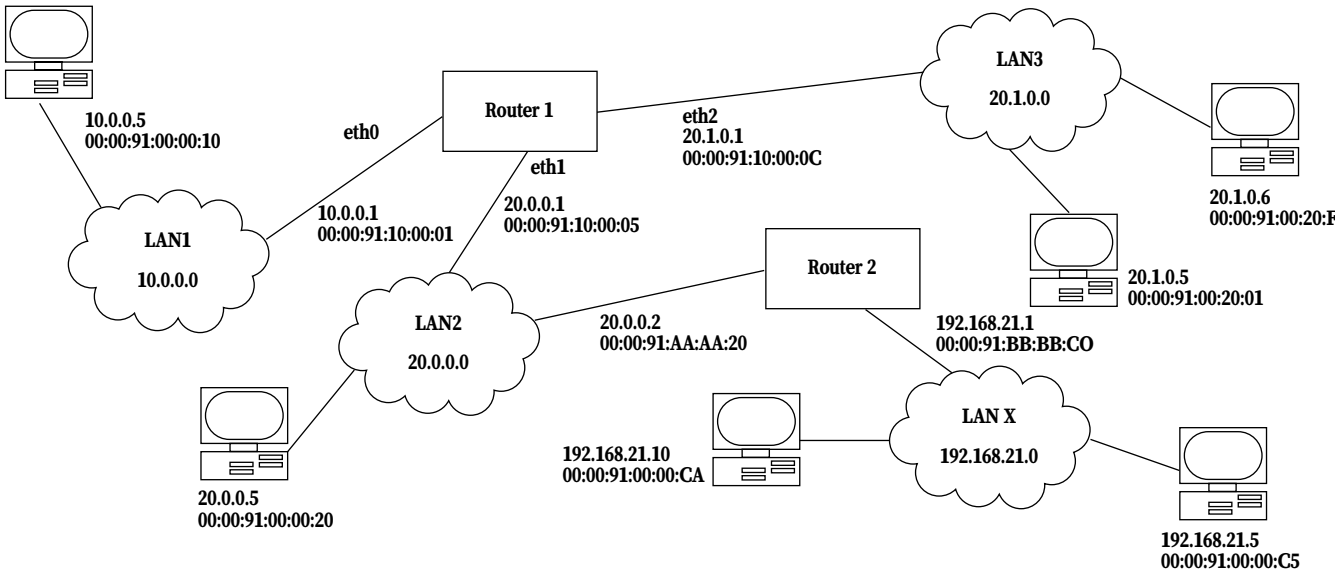

図7　従来のルータを用いたLANの例
Example of LAN with ordinary router

表1　ルーティングテーブルの例
Example of routing table

Destination	Gateway	Netmask	Metric	Interface
10. 0. 0.0	0.0.0.0	255. 0. 0.0	1	eth0
20. 0. 0.0	0.0.0.0	255.255. 0.0	1	eth1
20. 1. 0.0	0.0.0.0	255.255. 0.0	1	eth2
192.168.21.0	20.0.0.2	255.255.255.0	2	eth1

表2　ARPテーブルの例
Example of ARP table

IP Address	MAC Address
10.0.0.5	00:00:91:0 0:0 0:10
20.0.0.2	00:00:91:AA:AA:20
20.0.0.5	00:00:91:0 0:0 0:20
20.1.0.5	00:00:91:0 0:2 0:01
20.1.0.6	00:00:91:0 0:2 0:F9

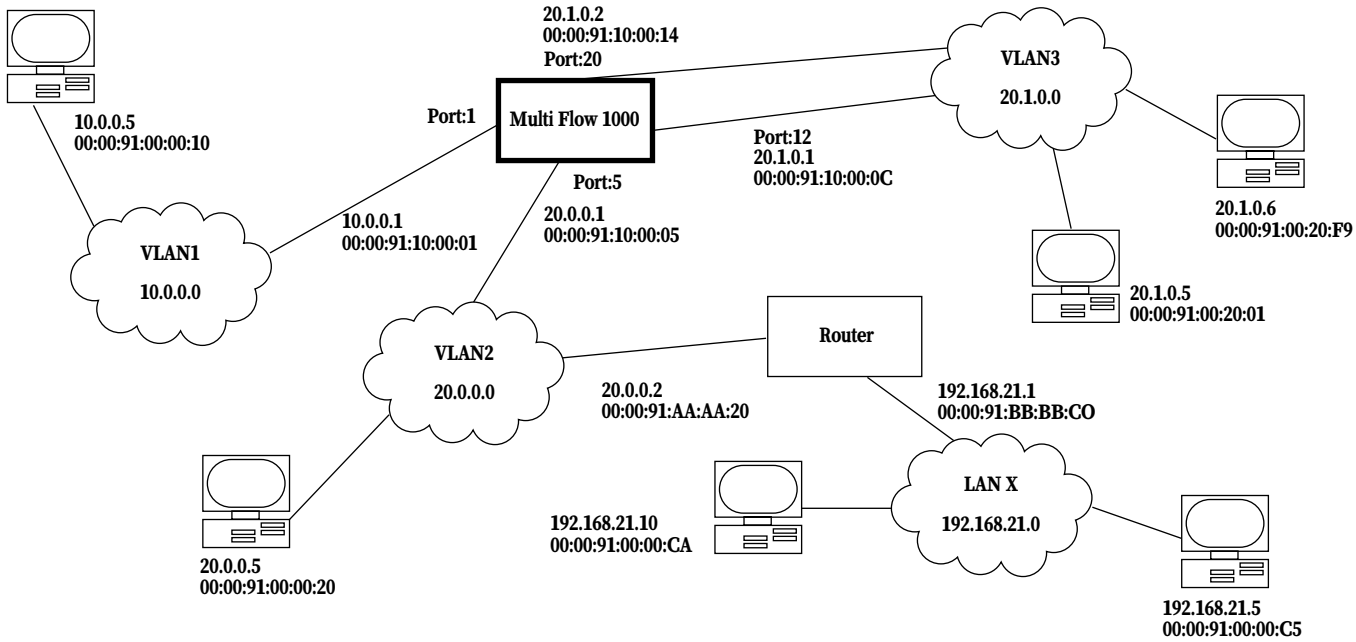

図8　MultiFlow 1000を用いたLANの例
Example of LAN with MultiFlow 1000

表3　E-ARPテーブルの例
Example of E-ARP table

IP Address	MAC Address	VLAN ID
10.0.0. 5	00:00:91:00:00:10	1
20.0.0. 2	00:00:91:AA:AA:20	2
20.0.0. 5	00:00:91:00:00:20	2
20.1.0. 5	00:00:91:00:20:01	4
20.1.0. 6	00:00:91:00:20:F9	4
192.168.21. 1	00:00:91:AA:AA:20	2
192.168.21. 5	00:00:91:AA:AA:20	2
192.168.21.10	00:00:91:AA:AA:20	2

6. ARPテーブルから得られたInterfaceに出力する。

このようにテーブルを2回ひいたり，ネットマスクやメトリック値を考慮して最も良い条件の経路（ベストマッチ）を探す必要があるため処理が非常に繁雑で時間がかかる。ハードウェアで実現する場合，これらの処理の複雑さが問題になる。

そこで，本装置では，ルーティングテーブルやARPテーブルを用いない新たな手法を開発した。これは，上述のルーティング処理の結果だけを予め表にしておき，複雑な手順なしにルーティング結果が得られるようにしたものである。この表をE-ARP（Enhanced ARP）テーブルと呼ぶ。図8は本装置を用いたネットワークの例であり，このネットワークにおけるE-ARPテーブルを表3に示す。なお，VLAN番号はユーザが各LANに便宜的に付ける番号であり，一定の範囲内で任意に選ぶことができる。また，本装置を含めて，一般的にスイッチングハブでは複数のポートを同じVLANに属するようにすることが可能である。

図8のLAN xに属するPCのように，他のルータを介して接続されているLANに属している装置のMACアドレスは知ることはできない。しかし，他のルータを介して接続している場合にはそのルータに次の処理を任せることになるので，ルーティング処理で必要なMACアドレスはそのルータのMACアドレスである。したがって，E-ARPテーブルでは，他のルータを介して接続されているLANに属している装置については，中継するルータのMACアドレスを登録している。

4.2.3　E-ARPテーブルの管理

ハードウェアで処理できないパケットを従来通りソフトウェアで処理するということは前に述べた通りである。本装置では，このソフトウェアでE-ARPテーブルの作成・管理も行なっている。

E-ARPエントリを作成する処理過程を以下のようになる（図9参照）。

1. DST IP ADRSがE-ARPテーブルに未登録であるIPパケットがソフトウェア処理に回される。
2. ソフトウェアはルーティングテーブルを参照し，DST IP ADRSが存在するVLAN決定する。
3. ARPテーブルを参照して，DST IP ADRSを持つホストのMACアドレスを求める。ARPテーブルにMACアドレスが登録されていなければ，アドレス解決プロトコルを用いてMACアドレスを取得し,ARPテーブルへの登録を行う。DST IP ADRSがVLANに接続された別のルータを介して接続されている場合は，その別ルータの物理アドレスを求める。
4. MACアドレスの付け換え，TTLの減算などを行った後，フォワーディングする。
5. 上記1.から3.の処理で取得したDST IP ADRS，VLAN ID，MAC ADRSをE-ARPテーブルに登録する。

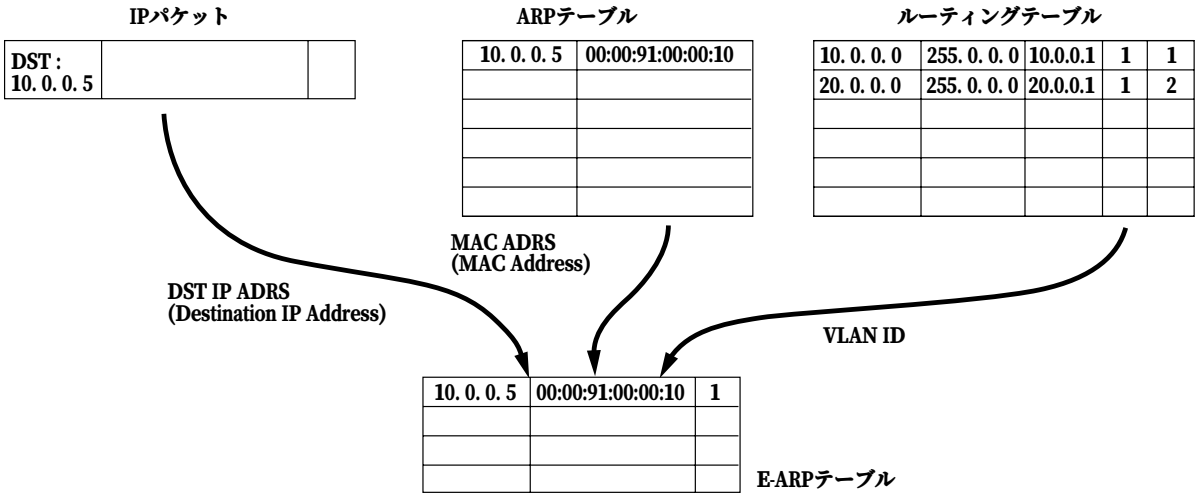

図9　E-ARPテーブルの概要
Overview of E-ARP table

このように，E-ARPテーブルに未登録のDST IP ADRSを持つIPパケットは，最初のものはソフトウェアによりフォワーディング処理される。しかし，フォワーディング処理と同時にE-ARPの登録を行うため，2個目以降のパケットはソフトウェアを介さず，ハードウェアで直接フォワーディング(ショートカット)できるようになる。

E-ARPテーブルは標準で2728エントリの大きさを持っている。この大きさは，VLANをまたがる通信を行うホスト数を考えると一般的なLANでは十分な大きさであるが，大規模ネットワークやインターネットプロバイダに直接接続される場合には十分とは言えない。そこで一定時間参照されなかったE-ARPエントリは，E-ARPエントリから削除するエージング処理を行っている。

エージングの方式について説明する。任意のE-ARPエントリに対してそのDST IP ADRSを持つIPパケットをフォワーディングしたかどうかをハードウェアで監視する。これを全エントリに対して実施し，一定時間内にそれぞれのエントリが参照されたか否かをソフトウェアに通知する。この時間間隔はソフトウェアからハードウェアに対して与えており，標準で3分としている。ソフトウェアは「参照されなかった」と通知されたエントリをE-ARPテーブルから削除する。

エージングによりE-ARPテーブルには常に新しく発生したトラフィックに必要な情報のみが残り，それ以外は削除される。つまり，E-ARPテーブルはハーティング情報のキャッシュとして存在している。この結果MultiFlow 1000はE-ARPテーブルのエントリサイズに関係なく，2728ホスト以上の大規模ネットワークにも適用できる。

4.2.4　ルータモジュールの構成

これまで述べてきた第3層スイッチング機構を含むルータモジュールのハードウェア構成を図10に示す。

E-ARPテーブルはCAM（Content Addressable Memory）で構成されており，非常に高速な検索ができる。

5 むすび

本装置は大量のトラフィックが発生するマルチメディア環境でのインターネット／イントラネットの構築に応えられる製品である。しかし，インターネット／イントラネットはますます使われるようになり，トラフィックの増加は指数関数的な勢いで伸びている。今後はさらに高速な装置が求められることは明らかである。

今後は，より速いバス，あるいはそれに代わる手法を開発すると共に，現在の第2層スイッチLSIと第3層スイッチLSIを統合し，第3層スイッチ機構を分散して高速化を図るなどによって，より高速にという要求に応えていきたい。

参考文献

1) 岩崎，小畑，松岡，東山，関水：“CSMA/CD, Demand Priorityの方式と性能比較”，アンリツテクニカル，72号，pp. 87-91 (1996-9)
2) D. Comer著，村井・楠本訳：“TCP/IPによるネットワーク構築”，共立出版
3) W. R. Stevens著，井上・橘訳：“詳解TCP/IP”，ソフトバンク
4) S. A. Thomas著，塚本・春本訳：“次世代TCP/IP技術解説”，日経BP
5) IEEE802.1Q：“Draft Standard P802.1Q/D8 IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks”

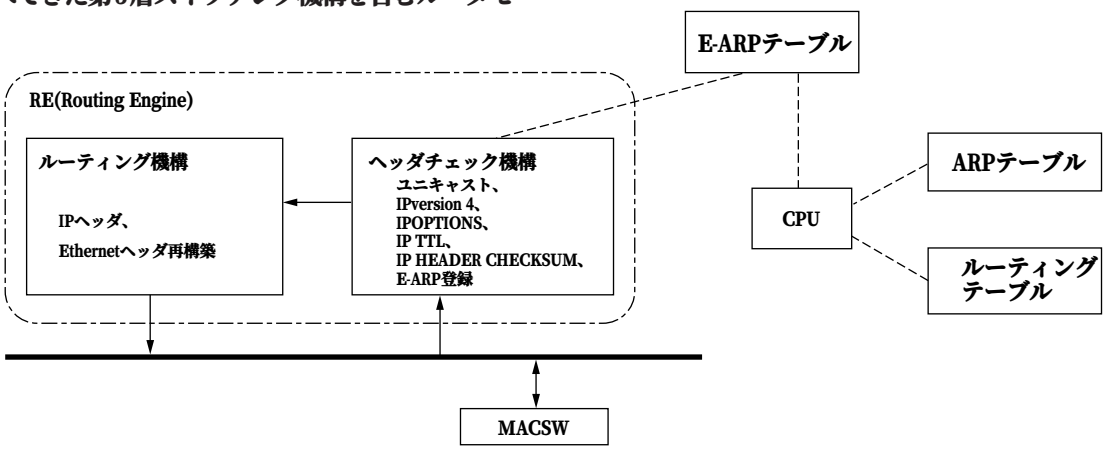

図10　ルータモジュールのハードウェア構成
Block diagram of router module