ネットワークを用いた疎結合型並列ルータアーキテクチャ

地引昌弘[†] 近藤栄一[††]
勝又憲一[††] 太田昌孝[†††]

インターネット上の通信を支えるルータには，通信量の増大に対応する高速性に加えて，高品質な通信を提供するための高機能性も必要となる．ルータの高速化/高機能化は，専用の高速ハードウェアを用いてパケット転送や経路検索などの並列度を上げることにより実現できるが，この方式では設計上の制約から，利用環境に応じて必要なルータの各機能を自由に増減することが難しい．本論文では，ネットワーク規模や利用状況の変化に応じて，高速性/高機能性を柔軟に提供できるルータアーキテクチャを提案する．ここで提案するアーキテクチャは，次の特徴を備える．(1) ルータの基本機能を分割して互いに独立した要素ルータに分担させる．(2) 要素ルータ間の結合には，専用ハードウェアを介した密結合ではなく，LAN などの既存ネットワーク技術による疎結合を採用する．(3) ネットワークを介して疎結合した要素ルータ群を，外部から仮想的な 1 台のルータに見せる．本方式は疎結合型アーキテクチャを採用したことにより，必要に応じてルータ内部結合網を動的に拡張し，要素ルータを増減することで，ネットワーク規模や利用状況の変化に柔軟に対応できる．さらに本論文では，提案方式の実現に必要な分散経路制御機構およびこれらの実装/評価について述べる．

A Parallel Router Architecture Based on Sparse-combining with a Network

Masahiro Jibiki,[†] Eiichi Kondoh,[††] Kenichi Katsumata[††]
and Masataka Ohta[†††]

High performance/high functionality router is realizable by increasing the parallel degree of packet transmission and route search using exclusive hardwares. However, in this method, it is difficult to increase and decrease dynamically each required router function according to a network environment because of the restrictions on a design. This paper proposes the novel router architecture which can provide flexibly high performance/high functionality according to change of a network scale or an use situation. The architecture proposed in this paper has the following feature: (1) dividing the basic function of a router and assigning each function to the element routers, (2) adopting not the dense-combination by a exclusive hardware but the sparse-combination by the existing network technology as combination between the element routers, (3) constructing virtually one parallel router by sparse-combining plural element routers with the network. This method can be adapted to change of a network scale or an use situation flexibly by adjusting the size of the router internal network and the number of element routers based on the sparse joint architecture. Furthermore, this paper describes distributed routing mechanisms required for construction of the proposed architecture, and the implementation/estimations.

1. はじめに

インターネットが様々な通信メディアを融合し，知的活動に対する本質的な基盤となるに従い，インターネット上の通信を支えるルータには，通信量の増大に対応する高速性に加えて，高品質な通信を実現するための高機能性も必要となる．ルータの高速化/高機能化は，専用の高速ハードウェアを用いてパケット転送や経路検索などの並列度を上げることにより実現できる[1)~5)]．しかし，この方式では設計上の制約から，利用環境に応じて必要なルータの各機能を低コストで自由に増減することが難しい．

本論文では，ネットワーク規模や利用状況の変化に対し，高速性/高機能性を低コストで柔軟に提供でき

† NEC ネットワークス開発研究所
Development Laboratories, NEC Networks
†† NEC ネットワークス IP ネットワーク事業本部
IP Networks Operations Unit, NEC Networks
††† 東京工業大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

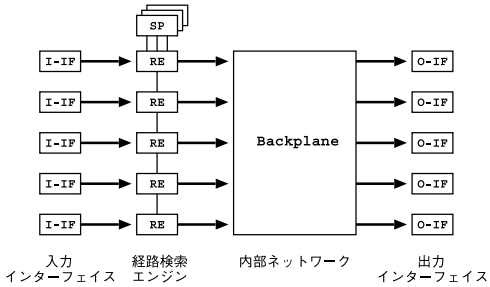

図 1 密結合型並列ルータアーキテクチャ
Fig. 1 Close coupling type parallel router architecture.

るルータアーキテクチャを提案する．ここで提案するアーキテクチャでは，ルータを互いに独立した次の 3 要素（以下，要素ルータ）の集合体として構築する．

FE： パケットを転送するフォワーディングエンジン
RE： 経路表を検索してパケットの転送先を決定するルーティングエンジン
SP： 経路制御プロトコルを用いて取得した経路情報から経路表を作成するシグナリングプロセッサ

提案方式では，LAN などの既存ネットワーク技術を用いてスイッチ（以下，要素スイッチ）を疎結合することにより動的に拡張可能なルータ内部結合網（Backplane）を構成し，複数の要素ルータから仮想的に 1 台の並列ルータを構築する．さらに本論文では，提案方式の構築に必要な分散経路制御機構およびこれらの実装と評価について述べる．本方式は，必要に応じて要素ルータ/スイッチを増減することにより，接続されるネットワークや回線数の変化，パケットおよびプロトコル処理の負荷変動に対して低コストで柔軟に対応できる．

2. ルータアーキテクチャとネットワーク設計

2.1 密結合型並列アーキテクチャの硬直性

ルータを高速化/高機能化するには，一般に専用の高速ハードウェアを用いてパケット転送を高速化し[4]，また経路検索の並列度を上げる方式が採用されている[5]．専用のハードウェア/内部結合網を用いた密結合型並列ルータのアーキテクチャを，**図 1** に示す．

図 1 のルータアーキテクチャでは，シグナリングプロセッサ（SP）が経路制御プロトコルを用いて経路情報を取得し，経路表を作成する．SP 以外の構成要素はハードウェア化され，これらを並列に接続することで高速なパケット転送を実現している[1)~3)]．また，ルータ内部結合網（Backplane）は，通常，N 本の入力/出力回線を互いに交差させ，交差点を半導体スイッチで接続したクロスバー方式で構成される（**図 2** 上）．回線収容数や交換容量，製作コストなどの理由から，図 2 下に示す複数のクロスバーを組み合わせた方式もある．ルータは，以下の手順でパケットを転送する．

(1) 入力インタフェース（I-IF）がパケットを受信
(2) パケットの転送先を決定するルーティングエンジン（RE）は，SP が作成した経路表を検索して出力インタフェース（O-IF）を決定
(3) Backplane を介して，RE で決定した O-IF へパケットを転送

一方，このアーキテクチャには，拡張性に関して以下の問題がある．

- ルータ内部結合網は専用ハードウェアにより構成されるため，内部結合網の規模を自由に拡張して並列度を上げることができない．ルータを構成する各要素（I-IF，RE，O-IF）の増設は，内部結合網の規模により制約を受ける[2),3)]．
- ネットワーク規模の変化に応じてルータの構成要素を増設するには，前もって将来の拡張を見越したより大きな内部結合網を用意する必要がある．しかし，専用ハードウェアにより構成される内部

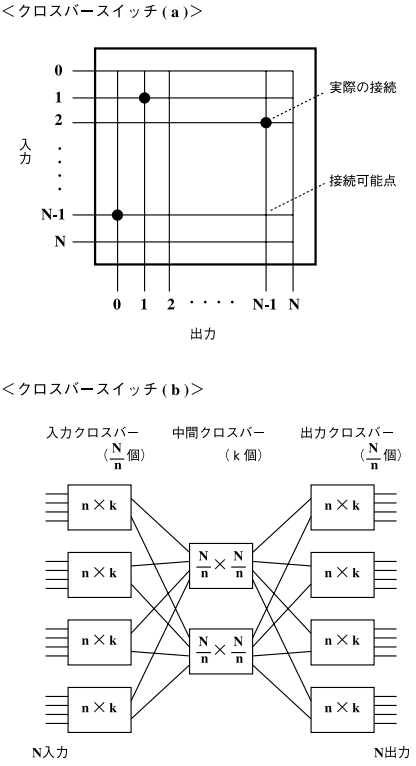

図 2 密結合型並列ルータにおける内部結合網の構成
Fig. 2 The structure of a backplane in the close coupling type parallel router.

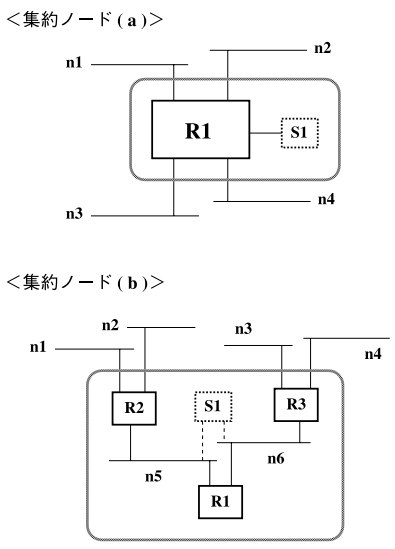

図 3 集約ノードのデザイン
Fig. 3 A design of the exchange node.

結合網は高価であり，低コストでネットワーク規模の変化に対応することができない．

ルータ内部結合網を専用ハードウェアにより構成する密結合型並列ルータアーキテクチャのこのような硬直性は，インターネットの設計/運用に対して，次節で述べるような制限を与えている．

2.2 ネットワーク設計に対する制限

インターネットは，複数のネットワークや回線を集約して管理する NOC☆/IX☆☆あるいはユーザからの回線を直接収容する局/AP☆☆☆などの集約ノードを基本単位として，これらを互いに接続することで構築されている[6]．集約ノードの管理では，増大する接続回線数や通信量に対して，ノード自体のパケット交換性能を落とさずにサービスを提供するためのコストをいかに軽減するかが問題となる[7]．

集約ノードの構築方法としては，

(a) 大規模/高性能なルータを単体で用いた放射状のトポロジーとして構築する方法（**図 3** 上）
(b) 安価な複数のルータを組み合わせたネットワークとして構築する方法（図 3 下）

の 2 種類が存在する[8]．

大型の密結合型並列ルータを利用する (a) の手法は，前節で述べたルータアーキテクチャの硬直性により，ネットワーク規模や利用状況の変化に対応することが難しい．したがって，集約ノードとしてパケット交換の性能が同じであれば，安価に集約ノードを構築できる (b) の手法が通常は用いられる．しかし，集約ノードを (b) の手法により構築する場合は，以下にあげる問題が存在する[9]．これらは，本来ルータの拡張により対応可能な問題を，ネットワークの再設計だけで対応する点に原因がある．

a. 複数のルータに対する管理のコスト

(a) の方式では（図 3 上），ルータ R1 上だけで n1〜n4 までの 4 種類の経路情報を管理できる．一方，(b) の方式では（図 3 下），n1〜n6 までの 6 種類の経路情報を，各ルータ（R1〜R3）ごとに管理する必要があり，管理コストが上昇する．管理コストの低減を目指したノード内への経路制御プロトコルの導入は，ノードの内部と外部で経路情報を交換するための設定を繁雑化させる．また，設定に誤りがあった場合は，他の地域に対するパケットの誤配信や喪失など深刻な影響を及ぼす[10]．

b. ネットワークの変化にともなう設計変更のコスト

新たなサービスの開始にともない集約ノード内へサーバを導入する場合，(a) の方式ではルータ R1 の直下にサーバが配置される（図 3 上のサーバ S1）．一方，(b) の方式では，たとえばルータ R2 に比べてルータ R3 からのアクセスが多い場合には，サーバをルータ R3 のリンク上に配置するなど，負荷分散を考慮してサーバの位置を決定する必要がある（図 3 下のサーバ S1）．ネットワークに変更がある場合も，サーバの導入と同様に設計変更のコストが問題となる．

c. 経路制御に関するソフトウェア処理のコスト

高品質なインターネット通信を提供する QoS 保証通信に必要なシグナリング処理は，一般に計算量が多く，その上フローごとに計算を行う必要がある[11]．特に局や AP などに配置され，ユーザからの回線を直接収容するルータは，シグナリング処理の回数が多くなり，負荷の増大に応じて高性能な CPU が必要となる．これに対し安価なルータでは，一般に複数の CPU を装備するといった強力なソフトウェア処理能力を備えておらず，また，経路制御に関連する処理は集約ノードの外部環境に依存するため，ネットワーク設計だけではソフトウェア処理のコスト増加に対応できない．

3. 疎結合型並列ルータの提案

3.1 疎結合型並列ルータアーキテクチャ

ネットワーク規模や利用状況の変化に低コストで効率良く対応できる高速/高機能ルータとして，我々は，以下の特徴を備える疎結合型並列ルータを提案する．

☆ Network Operation Center
☆☆ Internet Exchange
☆☆☆ Access Point

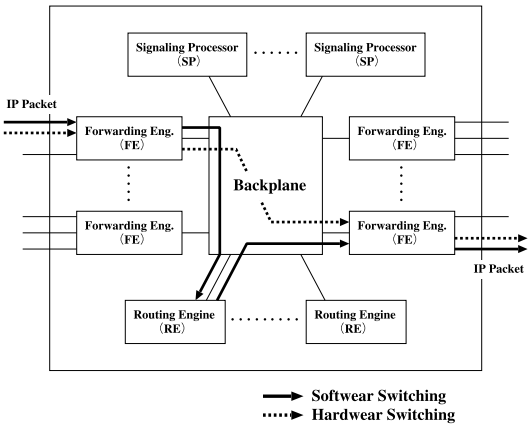

図 4 疎結合型並列ルータアーキテクチャ
Fig. 4 Loose coupling type parallel router architecture.

表 1 要素ルータの一覧
Table 1 The list of element routers.

要素ルータ	処理
SP☆	・経路制御プロトコルの処理 ・経路表の作成/管理
FE☆☆	・インタフェースの収容 ・ハードウェアスイッチング
RE☆☆☆	・経路検索

- ルータの基本機能を分割し，潜在的に高負荷となりうる部分を複数の要素ルータで分担させる．
- 専用ハードウェアではなく，LAN などの既存ネットワーク技術により要素スイッチ群を疎結合することで，動的に拡張可能なルータ内部結合網を構築する．
- ネットワークを介して疎結合した要素ルータ群/内部結合網を，外部から仮想的な 1 台のルータに見せる．

図 4 に疎結合型並列ルータのアーキテクチャを示す．これにより，負荷の高い要素ルータを必要に応じて増設することで，負荷への柔軟な対応ができる．疎結合型並列ルータを構成する各要素ルータおよびその処理分担を**表 1** に示す．疎結合型並列ルータ内のパケット転送には，以下の 2 種類がある．

- 要素ルータ FE 内に経路情報がキャッシュされていない場合（その宛先に対する最初のパケット転送の場合）は，要素ルータ RE にパケットを転送して経路検索を行う（図 4 の実線矢印）．
- 同一の宛先に対する以後の転送では，要素ルータ FE 内にキャッシュされている経路情報を利用したハードウェアスイッチングによる転送を行う（図 4 の点線矢印）．

要素ルータ RE による検索結果は，経路情報として要素ルータ FE へ通知され，FE 上にキャッシュされる．これにより，オーバヘッドとなる経路検索およびそれにともなう要素ルータ間の転送を最小限に抑えられる．並列ルータがアクセス網の近くへ配置される場合は，ホスト単位でフィルタリングなどを行う機会が多くなるため，要素ルータ FE の経路情報キャッシュは，ホスト経路で行う必要がある．一方，バックボーンに配置される場合は，キャッシュサイズを抑えるため，要素ルータ RE で管理する経路表のプレフィックス単位に行う必要がある．また，要素ルータ FE でキャッシュされる経路情報は以下の場合に削除されるため，経路情報の不一致などによる誤動作は生じない．

- キャッシュ上の経路に変更が生じた場合
- キャッシュ上の経路に対する通信が一定時間発生しなかった場合

ルータ内部結合網（Backplane）は，ネットワークを介した要素スイッチの疎結合により構築するが，$K \times K$ の要素スイッチを用いて N 個の要素ルータを結合するには，$\log_K N$ 段の要素スイッチを経由する必要がある．この場合，各段ごとに全要素ルータの通信容量と等しい通信容量を確保するには，最低でも各段で $\frac{N}{K}$ 個の要素スイッチを相互結合する必要がある☆4．この最低値を達成する結合方式としては，ハイパークロスバーやバンヤン，シャッフルエクスチェンジなど多少のバリエーションがあるものの，これらは等価であり，基本的には**図 5** のようなトポロジーで構成される[12]．

内部結合網で複数の要素スイッチを経由することによる遅延は，以下の理由から抑えることが可能である．

- 内部結合網の経路制御とルータ外部のネットワークに対する経路制御を分離することにより（後述），内部結合網で利用される経路情報数はたかだか $O(N)$ 個☆5に抑えられる．各要素スイッチでは全内部経路情報がキャッシュ上に格納され，ハードウェアスイッチングによりパケットをワイヤスピードで転送できる．
- 内部結合網における転送では，並列計算機の相互結合網で生じる同期処理/通信ループなどが存在せず，個々の通信は，互いに独立した $O(\log_K N)$

☆ Signaling Processor
☆☆ Forwarding Engine
☆☆☆ Routing Engine

☆4 全体では $\frac{N}{K}\log_K N$ 個の要素スイッチが疎結合される．
☆5 内部結合網で利用される経路情報数
≈ 外部にパケットを出力するインタフェース（IF）数
= (要素ルータ FE の個数)×(FE 1 個あたりの IF 数)

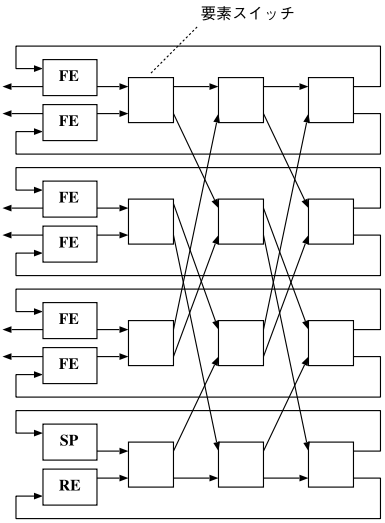

図 5 疎結合型並列ルータにおける内部結合網の構成
Fig. 5 The structure of a backplane in the loose coupling type parallel router.

ホップの転送で必ず終了する．

さらに，各要素スイッチでは，同じ宛先のパケットが同時に複数入力された場合に備えてバッファを必要とするが，これも以下の理由により問題となるほどの影響は発生しない．

- インターネットのプロトコルは，本来パケットの欠落に対処する機能を有しており，ベストエフォート通信においてバッファがあふれた場合は，単にそれを捨てることで対応できる．
- QoS 通信では，ベストエフォート通信用キューの他に QoS 通信用キューを 1 つ用意することで，特定の許容範囲に収まる通信品質を提供できる[13]．

以上より，疎結合型並列ルータの内部結合網には特別な要素スイッチを必要とせず，低価格な汎用スイッチを利用できる．また，内部結合網を要素スイッチの疎結合ネットワークとして構成するため，密結合型並列ルータとは異なりその規模は自由に変更が可能で（2.1 節），利用状況の変化に応じて無駄なく並列度を上げられる．したがって，並列ルータを低コストで提供できる．

各機能ごとに分割した要素ルータ群/内部結合網を結合し，外部に対して仮想的に 1 台のルータとして見せるには，以下の機能が必要となる．

(a) 分散経路制御 複数の要素ルータ SP 上で経路制御プロトコルの分散処理を行い，各経路情報群を統合して 1 組の外部経路表を作成する．これにより，複数の要素ルータ SP による経路情報の並列処理が可能となり，負荷への柔軟な対応が可能なアーキテクチャを実現できる．

(b) 内部経路制御 疎結合型並列ルータ内でのパケット転送を管理する内部経路制御と外部ネットワークの経路制御を分離し，それぞれ独立に管理する．これにより，要素ルータ RE が，外部ネットワークの変化に応じて出力先の要素ルータ FE を決定するだけで，パケットを転送できる．

(c) インタフェースの共有 要素ルータ SP が，経路制御プロトコルを受信した要素ルータ FE 上のインタフェース情報を共有する．これにより，要素ルータ SP は，各要素ルータ FE が収容する（SP では直接収容しない）インタフェースに関する経路情報を扱うことができる．

(a) については次節で述べる．(b) および (c) については，3.3 節で述べる．

3.2 分散経路制御機構

ルータ内で行われる処理のうち，経路情報の管理や経路表の作成などはソフトウェア処理に頼らざるをえず，一般に処理コストが高い．また，ルータを経路制御ドメインの境界上に配置して各サブネットごとに異なる経路制御を行う場合は（たとえば，VPN 網を構築するなど），多数の経路制御プロトコルを同時に扱う必要もある．さらに，QoS 保証通信のためのシグナリングでは，QoS 情報をもとにフローごとのグラフ（最適経路）を計算する必要がある．グラフ計算は転送処理/経路検索に比べてコストが高いことから，ネットワーク規模が拡大するにつれ，シグナリング処理が他の経路情報処理を圧迫してルータ全体の効率を低下させる[14]．

疎結合型並列ルータでは，経路情報処理の負荷に応じて要素ルータ SP を増設し，その処理を分散させて並列に実行することで，負荷の増大に対処できる．複数の要素ルータ SP を用いて経路情報の分散処理を行うために，経路情報処理を以下の 2 種類に分離した．

- 経路制御プロトコルごとの処理モジュール（プロトコルエンジン：PE）
- 各 PE が作成する最適経路表（ローカル経路表）を統合して 1 組のグローバル経路表を作成する経路表管理モジュール（Core）

PE は，負荷に応じて要素ルータ SP 上に分散配置される．Core は，要素ルータ SP の 1 つで稼働する（疎結合型並列ルータ内には 1 個しか存在しない）．Core の稼働する要素ルータ SP が経路情報を要素ルータ RE に配布する．分散経路制御機構のアーキテクチャを**図 6** に示す．

複数の要素ルータを仮想的に 1 台のルータとして動

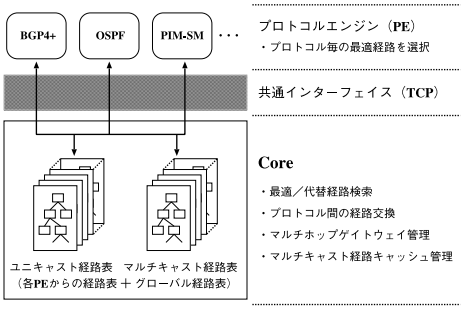

図 6 分散経路制御機構のアーキテクチャ
Fig. 6 Architecture of the distributed routing mechanism.

作させるには，分散経路制御機構のアーキテクチャに以下の性質を持たせる必要がある．

(a) 分散性 経路制御プロトコルの処理を行う PE と，各 PE から渡される経路情報を統合して経路表を作成する Core は，別々のプロセスとして定義される．両者を TCP で接続することで，通信にともなう信頼性の低下を引き起こさず複数の要素ルータ SP に PE などの機能を分散配置できる．

(b) 抽象性 Core 内には経路制御プロトコルに依存する処理が存在せず，これら固有の処理は各 PE 内にカプセル化される．Core は，経路情報を PE 単位で管理するが，経路制御プロトコルの種類は関知しない．

(c) 対称性 経路制御に関する情報は，PE → Core の向きだけに伝達されるのではなく，必要に応じて（特定のイベントに起因して）Core → PE の向きにも伝達される．

Core〜PE 間で TCP による通信を行うため，要素ルータ SP には制御用 IP アドレスが割り当てられる．外部ネットワークに要素ルータと同じ IP アドレスが存在した場合は，両者の通信を区別する必要が生じる．疎結合型並列ルータの内部転送では，各パケットに入力元あるいは出力先インタフェースの識別子を付加するが（3.3 節），この識別子として要素ルータ間の制御用通信を表す値を定義することで，両者を区別できる．

3.3 内部経路制御とインタフェースの共有

疎結合型並列ルータでは，要素ルータ SP 上の分散経路制御機構が，要素ルータ FE 上のインタフェースに対して外部経路表を作成するため，SP は各 FE のインタフェース情報を共有する必要がある．インタフェース情報の共有および外部経路表の作成は以下のように行う．

- 各要素ルータ FE が収容するインタフェースと 1 対 1 に対応する論理的なインタフェースを，要素ルータ SP に作成する．
- 要素ルータ SP 上の分散経路制御機構は，論理インタフェースに対して外部経路表を作成する．

疎結合型並列ルータにおけるパケットの転送制御（経路表の検索）は，内部/外部を問わず要素ルータ RE が行う．分散経路制御機構の Core が作成した外部経路表は，要素ルータ RE に配布される．RE では，パケットの内部転送先となる要素ルータを外部経路表に基づいて決定する（後述）．外部からパケットを受信した要素ルータ FE を IngressFE と表記し，外部にパケットを送信する FE を EgressFE と表記すると，内部パケット転送には以下の 3 種類が存在する．内部パケット転送 (a) および (b) は近隣ルータへパケットを転送する場合であり，(c) は主に経路制御プロトコルのパケットを要素ルータ SP へ転送する場合である．

(a) 要素ルータ RE の経路決定により IngressFE から EgressFE へ転送する場合☆
(b) ハードウェアスイッチングにより直接 IngressFE から EgressFE へ転送する場合☆☆
(c) IngressFE が収容するインタフェース宛のパケットを要素ルータ SP へ転送する場合

内部経路制御として，内部結合網を構成する各要素スイッチが全要素ルータに対するデータリンク層レベルの経路情報を保持することで☆☆☆，データリンク層のスイッチにより，各ホップ（要素スイッチ）では IP ヘッダを変更することなく高速な内部転送を実現できる．また，要素ルータ FE 上でパケットが実際に入出力するインタフェースは，対象となるインタフェースの識別子を内部転送パケットへ付加することにより管理できる．並列ルータ内のパケット転送におけるデータリンク層ヘッダの変化を**図 7** に示す．

パケットの内部転送を行う各要素ルータの動作を以下に述べる．

a. 要素ルータ FE （IngressFE）の動作

IngressFE では，経路情報がキャッシュに存在しない場合（ある宛先に対して初めてパケットを転送する場合）と，存在する場合（2 回目以降の転送の場合）とでパケットの転送先が異なる．最初のパケット転送は，要素ルータ RE を介した出力インタフェース（O-IF）の決定をともなう（図 7 の矢印 a）．一方，同じ宛先に対する 2 回目以降の転送では，IngressFE が持つ経路情報によりハードウェアスイッチングを行い，転送速度の向上をはかる（図 7 の矢印 b）．

b. 要素ルータ RE の動作

☆ IngressFE に経路情報のキャッシュが存在しない場合
☆☆ IngressFE に経路情報のキャッシュが存在する場合
☆☆☆ 経路情報の量については 3.1 節参照

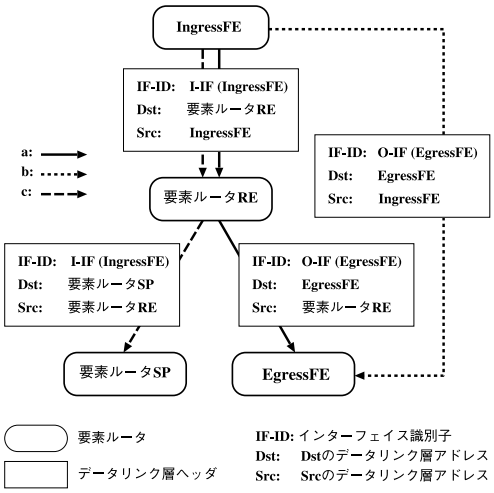

図 7 内部パケット転送とデータリンク層ヘッダ
Fig. 7 The internal packet forwarding and the data link header.

要素ルータ RE は，受け取ったパケットの IP ヘッダを解析し，その宛先が IngressFE のアドレスだった場合は要素ルータ SP へパケットを転送する（図 7 の矢印 c）．IngressFE 宛でなかった場合は，経路表から宛先に対応する出力インタフェース（O-IF）を決定し，それを収容する EgressFE へパケットを転送する（図 7 の矢印 a）．同時に，要素ルータ RE は，EgressFE の情報を経路情報として IngressFE に通知する．

c. 要素ルータ SP の動作

要素ルータ SP は，受け取ったパケットのインタフェース識別子を調べ（図 7 の IF-ID），IngressFE が実際にパケットを受け取ったインタフェース（I-IF）を識別する．要素ルータ SP の物理インタフェースは，内部結合網に接続するインタフェースだけであるが（図 4），SP のカーネルは，上で識別した IngressFE のインタフェースに対応する論理インタフェースを経由して上位層にパケットを渡す．経路制御プロトコルの処理では，パケットを受信したインタフェースの情報も利用する．要素ルータ SP 上の分散経路制御機構は，経路制御プロトコルのパケットを論理インタフェースから取得することで，要素ルータ FE 上の各インタフェースに対して経路表を作成できる．

4. 実　　装

疎結合型並列ルータの有効性および要素ルータ/スイッチの多段結合による性能の低下とハードウェアスイッチングの効果を検証するため，**表 2** の構成で並列ルータを実装した．

要素ルータ FE および RE には，NEC 製 ES100e/D4000 を利用した．要素ルータ SP には，Pentium III（500 MHz）CPU および 128 MByte のメモリを備えたパーソナルコンピュータを利用した．

表 2 疎結合型並列ルータの構成
Table 2 The component of the parallel router based on sparse-combining.

要素ルータ FE	要素ルータ RE	要素ルータ SP
2 台	2 台	1 台

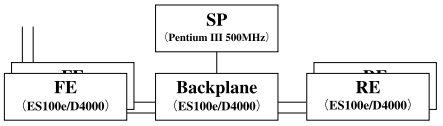

図 8 疎結合型並列ルータの実装
Fig. 8 The implementation of the parallel router based on sparse-combining.

ルータ内部結合網（Backplane）を構成する要素スイッチも，同様に ES100e/D4000 を利用した．ES100e/D4000 には 32 ポート分の交換能力があり，表 2 の構成における各要素ルータが 1 ポートだけを利用することで，Backplane を 1 台の ES100e/D4000 により構築できる．したがって，今回は最も単純な構成として，各要素ルータを 100 Base-T の Ethernet でスター状に接続した（**図 8**）．

また，分散経路制御機構については，PC-UNIX（FreeBSD/Linux）上で C 言語を用いて開発した．PE としては，BGP4+/OSPF/PIM-SM を既存のコードより移植し，その規模は合計で約 55 K 行となった．Core 部分は新規に開発し，約 14 K 行であった．これら以外に，各 OS 独自のコード☆をそれぞれの OS ごとに数 K 行ずつ（OS により異なる）作成した．

今回の実装は，以下の理由により，疎結合型並列ルータの一般的な実装と見なすことができる．

- Backplane は，全要素ルータの通信容量（最大 8 ポート）より大きい通信容量（32 ポート）を持つ．
- Backplane 内では，ハードウェアスイッチングによりパケットをワイヤスピードで転送するため，経由するホップ数の多寡は遅延全体に大きな影響を与えない．
- 全要素ルータは独立したルータで構成されており，それぞれが並列に動作できる．

5. 評　　価

本章では，2 章であげた問題に対する疎結合型並列

☆ インタフェース情報の取得やカーネルが保持する経路情報へのアクセスなど

ルータの有効性および性能について，4 章の実装をもとに評価を行う．一般にルータの性能を規定する項目として，経路情報の処理能力およびパケットの転送速度/転送容量があげられる．疎結合型並列ルータでは，このそれぞれについて次のオーバヘッドが想定される．

経路情報処理： 経路制御の分散化にともなう処理の転送によるオーバヘッド

パケット転送： 要素ルータ/スイッチ間の内部転送によるオーバヘッド

疎結合型並列ルータの有効性については次節で述べる．経路制御の効率については 5.2 節で，パケット転送の性能については 5.3 節で述べる．

5.1 有 効 性

集約ノードを複数のルータで構成する場合は（図 3 下），各ルータにおける経路情報の管理コストが高くなる（2.2 節 a）．一方，疎結合型並列ルータでは，利用する経路情報を外部/内部の 2 種類に分けて管理する．外部経路情報は，要素ルータ SP 上の分散経路制御機構で一括管理され，要素ルータ RE に配布される．内部経路情報は，外部経路情報から独立して各要素スイッチが管理し，その変化はルータの内部構成が変化した場合に限定される．以上より，経路情報の管理コストを大幅に低減できる．

また，複数のルータによる集約ノードは，ネットワークの設計/変更のコストも問題となる（2.2 節 b）．疎結合型並列ルータでは，要素ルータ FE がすべて平等に扱われ，内部結合網も全 FE の通信容量に等しい交換能力を備えている（3.1 節）．分散経路制御機構により，経路的にも外部からは 1 台のルータとして見える（3.2 節）．したがって，ネットワークの設計/変更に際しては単体ルータと等価に扱えるため，図 3 上と同様，余分なコストは発生しない．

経路情報の処理については（2.2 節 c），その処理負荷に応じた要素ルータ SP の増設により対応できる．処理を複数の要素ルータ SP へ振り分けるには，SP の 1 つをディスパッチャとして，そこから振り分ける方式（QoS 保証通信におけるシグナリング処理の負荷分散）や IngressFE から直接各 SP へ振り分ける方式（経路制御プロトコル処理の負荷分散）がある．IngressFE では自身宛のパケットを要素ルータ SP へ転送しているが，パケットの内部も解析することで，対応する SP へ転送先を振り分けられる．この場合，IngressFE で余分なパケット処理が発生するものの，経路制御プロトコルの通信は，遅延に対する許容範囲が大きいことおよびパケット数が多くないことから，問題にはならない．

5.2 経路情報の処理

疎結合型並列ルータでは，コストの高い経路情報の処理を，ネットワークで結合した要素ルータ SP 間で分散することにより処理能力の向上を目指すが（分散経路制御機構），処理の転送によるオーバヘッドを無視できない可能性もある．しかし，分散させる処理が以下のような性質を備える場合は，分散モデルとして単純なクライアントサーバ型を想定できる．

(a) 要求の転送元/先で処理の対象となるデータがすでに共有されており，必要なデータはすべてローカルにアクセスできる場合

(b) 要求の転送に際して通信量の上限を定めることが可能であり，要求がすべてその上限内に収まる場合

分散モデルが条件 (a) を満たす場合は，処理の途中で新たな通信によるオーバヘッドは生じない．また，条件 (b) を満たす場合は，処理の種類や処理を行う状況に応じて通信のオーバヘッドは大きく変化しない．つまり，これらを満たす分散モデルでは，処理の転送によるオーバヘッドを一定量以内と見積もることができる．以下では，分散経路制御機構におけるコストの高い処理として QoS 保証通信のためのシグナリング処理を取り上げ，ディスパッチャ方式による分散化（5.1 節）のオーバヘッドを見積もる．

分散経路制御機構は，要素ルータ SP 間で処理の対象となる経路情報を共有できるため，条件 (a) を満たす．単位時間あたりに発生するシグナリング要求数を λ，同じく単位時間あたりに処理可能な要求数を μ とすると，処理待ちによる平均遅延時間 (W_q) は待ち行列理論を用いて以下の式で与えられる．

$$W_q = \frac{\lambda}{(\mu - \lambda)\mu}$$

処理の負荷に応じて要素ルータを k 台増設し，シグナリング要求を機械的に各要素ルータへ分配する方式は，上記の W_q に対して単位時間あたりのシグナリング処理能力が k 倍になったモデルと考えることができる[15]．分散化により発生する通信は，要素ルータへ転送されるシグナリング要求と，その処理結果であるシグナリング経路☆の返送であり，データ量は各通信ごとにほぼ一定と考えることができるため条件 (b) を満たす．分散化にともなう通信遅延を W_c とすると，分散経路制御機構の平均遅延は以下の式で与えられる．

$$W_q' = \frac{\lambda}{(k\mu - \lambda)\mu} + W_c$$

☆ シグナリングの目的地までのパス情報

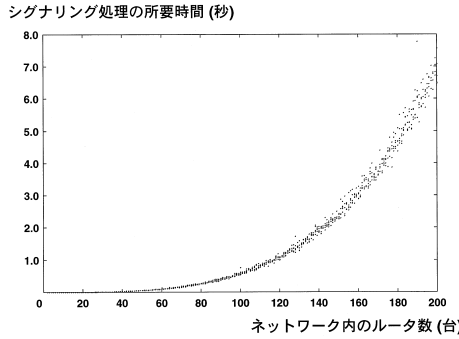

図 9 ネットワーク規模に対するシグナリング処理時間の分布
Fig. 9 The distribution of signaling time in contrast with a network scale.

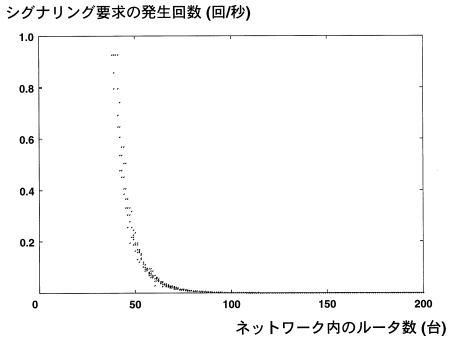

図 10 分散化による効率化が望めるシグナリング要求数の推移
Fig. 10 Transition of the number of signaling requests which can be expected the increase in efficiency by distributing.

W_q と W_q' を比較し，以下の条件が成り立つ場合は分散化による効果が得られるといえる．

$$W_q - W_q' > 0 \quad (1)$$

複数の要素ルータ SP 間で処理の転送によるオーバヘッドを測定するため，4 章の実装とは別に，要素ルータ SP とほぼ同性能のパーソナルコンピュータ（Pentium III（600 MHz）CPU/192 MByte メモリ）2 台を 100Base-T の Ethernet で接続した環境を用意した．1 台の PC が，シグナリング処理に要する時間を**図 9** に示す．これは，ルータを 2〜200 台まで変化させながら完全結合させた 199 種類のネットワークを疑似的に作成し，各ネットワークごとに経路情報を用いたグラフ計算の所要時間を計測したものである．また，分散化により発生する通信の所要時間 (W_c) を計測したところ往復 0.3 msec であった．計測は，要素ルータ SP に相当する 2 台の PC 間で行った．以下では，各ネットワーク規模に対して，式 (1) を満たす λ の条件を考える．

μ の値は，シグナリング処理に要する時間の逆数として図 9 より得られる．$\mu > \lambda$ および $k > 1$ のもとで式 (1) から得られる以下の式 (2) より，λ の条件を求めることができる．

$$W_c\lambda^2 + (1 - k - W_c\mu - W_ck\mu)\lambda + W_ck\mu^2 < 0 \quad (2)$$

式 (2) から得られる λ の範囲を $\alpha < \lambda < \beta$ とすると，$k > 1$ より以下の関係が成り立つ．

$$\mu < \frac{(k+1)W_c\mu + (k-1)}{2W_c} < \beta$$

したがって，式 (1) を満たす最終的な λ は，

$$\alpha < \lambda < \mu$$

として求められる．

2 台の要素ルータでシグナリングを処理する場合 ($k = 2$) に，ネットワーク規模の変化に対する λ の最小値 α の推移を**図 10** に示す．各ネットワーク規模に対し，1 秒間に発生するシグナリング要求数 λ が図 10 の曲線で示す値より大きい場合は，分散化による効率化が期待できる．たとえば，50 台のルータから構成されるネットワークでは，5 秒に 1 回以上（0.2 回/秒以上）の頻度でシグナリング要求が発生する場合に，シグナリング処理を分散化した方が効率が良いといえる．

5.3 パケットの転送

疎結合型並列ルータのアーキテクチャでは，パケットの転送に際して要素ルータ/スイッチ間の内部転送によるオーバヘッドが存在する．以下では，要素ルータ/スイッチの多段結合による性能低下とハードウェアスイッチングの効果を検証するため，4 章の実装を用いて転送能力および転送遅延を測定した．

ルータのパケット転送能力については，単位時間あたりのパケット転送数で評価できる．今回実装した疎結合型並列ルータのパケット転送能力は 1.98 Mpps であった．

次に，転送遅延の比較として，疎結合型並列ルータおよび単体の ES100e/D4000 へそれぞれ 2 台のパーソナルコンピュータを接続し，ping コマンドにより PC 間で 64 バイト（データ部分 56 バイト）のパケットを 20 個往復させ，その平均 RT 時間（往復時間）を計測した．並列ルータの内部転送には，RE 経由で転送される場合（最初のパケット）とハードウェアスイッチングにより転送される場合（2 個目以降）があり，計測にあたり両者を区別した．さらに，16,384 バイトのパケットをバースト的に発生させ，それぞれのスループットも計測した．その結果を**表 3** に示す．

RT 時間の計測に際し，疎結合型並列ルータの内部では，3 台の ES100e/D4000 間（IngressFE → Back-

表 3 転送遅延の比較
Table 3 The comparison of the forwarding delay.

機種	RT 時間	スループット
ES100e/D4000 単体	0.21 msec	92.36×10^6 bps
並列ルータ・RE 経由	10.54 msec	2.85×10^6 bps
並列ルータ・HW Switch	0.32 msec	92.23×10^6 bps

plane → EgressFE）でパケットが往復する．今回の実装は，内部結合網に 100Base-T の Ethrenet を用いており，Ethrenet の転送効率が 20〜30%であることを考えると[16]，64 バイトのパケットに対する転送遅延は 1 リンクあたり 20〜25 μsec となる．したがって，ハードウェアスイッチングによる転送において，ES100e/D4000 単体と比べ増加した転送遅延（0.1 msec）は，内部ホップ数の増加（4 リンク分）に起因するものにとどまっている．これより，たとえば内部結合網が 100 ホップ存在したとしても，Gbps クラスのリンクを用いることで転送遅延を μsec のオーダに抑えられる．

これに対し，要素ルータ RE 経由で転送される場合は，ハードウェアスイッチングによる転送と比較してその遅延に 30 倍以上の開きがある．要素ルータ RE では，パケットの出力インタフェースを決定しており，RE 経由での転送遅延はその大半が経路検索に費されている．今回の実装は，IngressFE がパケットの宛先をもとにハッシュ関数を引くことで，経路検索を依頼する要素ルータ RE を決めている．これは，複数の RE による完全な並列経路検索を目指したものである[17]．しかし同時に，ベースとなる ES100e/D4000 へはできる限り手を加えずに疎結合型並列ルータを構築することも目指した．これにより，IngressFE で依頼先の要素ルータ RE を決定する処理（ハッシュ処理）が，CPU の能力不足による影響を受けたと思われる．

一方，経路検索の頻度について考えると，バックボーンなどでは，1 つのインタフェースに対して同時に 10^3〜10^4 回のオーダで通信が発生する[18]．しかし，2 章で述べたユーザを直接収容する場合では，事実上，各インタフェースは 1 人のユーザに対してのみ割り当てられることから，経路情報キャッシュのサイズは同時通信数に比べて十分に大きくなる．したがって，各 FE では経路情報キャッシュが多くの場合に有効となり，ハードウェアスイッチングが選択されるため，転送遅延は単体ルータと比較しても実用範囲に収まると考えられる．

スループットについては，ES100e/D4000 単体と疎結合型並列ルータとでは 1%程度の違いしか見られなかった．以下では，より詳細な比較を行うため，パケットのサイズを 64〜16,384 バイトへと変化させながらバースト的に発生させ，ES100e/D4000 単体と疎結合型並列ルータとのスループットをそれぞれ計測した．**図 11** に，ES100e/D4000 単体のスループットを 100 とした場合の疎結合型並列ルータのスループット比を示す．図 11 は，スループット比が小さいほど疎結合型並列ルータのオーバヘッドが大きく，100%に近い値ほど ES100e/D4000 単体に近いスループットを発揮できていることを表す．

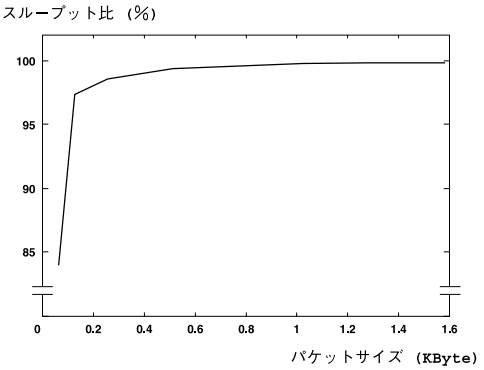

図 11 転送パケットサイズとスループットの比較
Fig. 11 The comparison of the forwarding packet size and the throughput.

疎結合型並列ルータでは，100 バイトより小さなパケットの転送において，単体ルータに比べてオーバヘッド（5〜15%の効率低下）が生じた．一般にルータは，小さなパケットを大量に転送する場合にスループットの低下が見られるが，疎結合型並列ルータでは，パケットの転送に際して内部で複数の要素ルータ/スイッチを経由することから，単体ルータに比べてより多くのオーバヘッドが生じていると思われる．しかし，IP 層では ATM と異なり，小サイズのセルなどによるフラグメント化を行わないことから，そのデータパケットの平均サイズは 100 バイトを超える[19]．したがって，スループットに関するオーバヘッドについても，単体ルータと比較して実用範囲に収まるといえる．

以上より，ルータ内部結合網を LAN などの既存ネットワーク技術で構成した場合でも，それにともなう性能の大きな低下は見られず，密結合型並列ルータに比べて拡張性を増した疎結合型並列ルータは，ネットワーク規模の変化に柔軟に対応可能である．

6. ま と め

本論文では，ネットワーク規模やパケット/プロトコル処理の負荷に対して柔軟に適応可能なルータアーキテクチャとして，以下の特徴を備える疎結合型並列

ルータを提案した．

- ルータの基本機能を分割して，互いに独立した要素ルータに分担させる．
- 内部結合網は，LAN などの既存ネットワーク技術を用いた要素スイッチの疎結合として構築する．
- ネットワークにより疎結合した要素ルータ群/内部結合網を，仮想的な 1 台のルータに見せる．

さらに本論文では，提案方式の実現に必要な分散経路制御機構およびこれらの実装と評価について報告した．

今回の開発では，疎結合型並列ルータの転送オーバヘッドを，要素ルータ/スイッチ間での内部転送に起因するものにとどめることができた．また，分散経路制御機構を利用することで，経路制御プロトコル処理を含む経路情報の管理に対する負荷分散を容易に行うことが可能となった．現在は，上記の成果より，小さなパケットに対する転送の効率化や経路検索の高速化，またマルチキャストパケットの処理などについて研究開発を進めている段階である．

謝辞 筑波大学大学院経営システム科学専攻の久野靖氏によるコメントは，本論文を改善するうえで大きな助けとなった．また，本研究は平成 10 年度通産省補正予算事業として，(財) 日本情報処理開発協会の委託研究開発プロジェクト「超高速・高性能次世代インターネット技術の研究開発」の支援を受けて実施された．ここに記して謝意を表す．

参 考 文 献

1) Development of High Speed Routers in USA and Japan, 情報処理学会 HQI 研究報告 (Oct. 1999). http://hsn-crl.koganei.wide.ad.jp/ipsj-hqi/prev_workshops/4.html
2) Avici Systems Inc.: The Avici Terabit Switch Router. http://www.avici.com/
3) Pluris Inc.: Teraplex 20 Product Overview. http://www.pluris.com/
4) Partridge, C., et al.: A 50-Gb/s IP Router, *IEEE/ACM Trans. Networking,* Vol.6, pp.237–248 (1998).
5) Lampson, B., Srinivasan, V. and Varghese, G.: IP Lookups Using Multiway and Multicolumn Search, *IEEE/ACM Trans. Networking,* Vol.7, pp.324–334 (1999).
6) 中川，江崎，永見：ラベルスイッチを用いた分散 IX の設計，情報処理学会 DSM 研究報告，1999-DSM-14, pp.85–90 (Jul. 1999).
7) 林：地域ネットワークの目的と新しい展開，情報処理，Vol.41, No.1, pp.3–7 (2000).
8) 北辻，小林，北村，加藤，小西：APAN 東京 XP の構築と運用について，情報処理学会 DSM 研究報告，2000-DSM-17, pp.37–42 (Mar. 2000).
9) 中川：地域 IX の現状と展望，情報処理，Vol.41, No.1, pp.8–13 (2000).
10) 中川：地域 IX における地域内経路制御の実現，情報処理学会 DSM 研究報告，1998-DSM-11, pp.37–42 (Sep. 1998).
11) Sola, M., Ohta, M. and Maeno, T.: Scalability of Internet Multicast Protocols, *Proc. ISOC INET'98*, Geneva, Switzerland (Jul. 1998). http://www.isoc.org/inet98/procee-dings/6d/6d_3.htm,
12) 黒川，相磯：並列処理の諸問題 結合方式，情報処理，Vol.27, No.9, pp.1005–1021 (1986).
13) 藤本，藤川，太田，池田：M/D/1/K 待ち行列モデルを用いた通信品質保証の要件に関する考察，第 58 回情報処理学会全国大会，pp.405–406 (Mar. 1999).
14) Ghosal, D., Lakshman, T.V. and Huang, Y.: Parallel Architectures for Processing High Speed Network Signaling Protocols, *IEEE/ACM Trans. Networking*, Vol.3, pp.716–728 (1995).
15) 森村，大前：応用待ち行列理論，日科技連出版社 (1975).
16) Tanenbaum, A.S.: Computer Networks 3rd Edition, Prentice-Hall, Inc. (1996).
17) Ohta, M., et al.: Hash Parallel and Label Parallel Routing for High Performance Multicast Router with Fine Grain QoS Control, *Proc. JSPS/IEEE IWS'99*, Osaka, Japan, pp.13–16 (Feb. 1999).
18) Jibiki, M., Terano, T. and Hashida, O.: Comprehensive Bottleneck Detection via Non-Linear Optimization Techniques, *Proc. JSPS/IEEE IWS'99*, Osaka, Japan, pp.100–107 (1999).
19) 地引：国際線 TCP コネクションの統計的解析，WIDE プロジェクト研究報告書 1996，pp.257–272 (1996).

(平成 13 年 5 月 9 日受付)
(平成 13 年 10 月 16 日採録)

地引 昌弘（正会員）

平成 2 年東京工業大学理学部情報科学科卒業．平成 4 年同大学院理工学研究科情報科学専攻修了．同年日本電気（株）入社．分散アーキテクチャ，ソフトウェア，通信システム，経路制御等の研究に従事．現在，同社ネットワークス開発研究所に勤務．

近藤 栄一（正会員）

平成元年明治大学工学部電子通信工学科卒業，同年日本電気（株）入社．以来，高速トランスポートプロトコル，ルータアーキテクチャの研究，IP マルチキャストルータの開発に従事．現在，同社 IP ソフトウェア技術本部主任．電子情報通信学会会員．

勝又 憲一（正会員）

昭和 56 年早稲田大学理工学部電子通信学科卒業，同年日本電気（株）入社．以来，パケット交換システム，IP ルータ製品の開発に従事．現在，同社 IP ソフトウェア技術本部技術マネージャー．電子情報通信学会会員．

太田 昌孝（正会員）

昭和 34 年生まれ．昭和 57 年東京大学理学部情報科学科卒業．昭和 59 年，同大学大学院理学系研究科情報科学専門課程修士課程修了．昭和 62 年，同研究科博士課程単位取得退学．同年より東京工業大学総合情報処理センター助手．平成 12 年より同大学大学院情報理工学研究科講師．平成 13 年よりモバイルインターネットサービス（株）CTO を兼務．博士（理学）．コンピュータグラフィックス，マルチメディア，UNIX，計算の高速化，国際化文字コード，ドメイン名システム，超高速インターネット，QoS 保証，マルチキャスト等の研究に従事．平成 7 年度山下記念研究賞，平成 9 年度第 12 回電気通信普及財団賞奨励賞．「いま，日本語が危ない」，「ドキュメント伊那 ADSL」他著書多数．