リンクレイヤプロトコルMAPOS準拠の高速スイッチCOREswitch

小倉　毅[†]　川野　哲生[†]　清水　健司[†]　丸山　充[†]
高橋　直久[††]

COREswitch : A High-Speed Link-Layer Switch Compliant with MAPOS

Tsuyoshi OGURA[†], Tetsuo KAWANO[†], Kenji SHIMIZU[†], Mitsuru MARUYAMA[†], and Naohisa TAKAHASHI[††]

あらまし　本論文では，筆者らが提案するリンクレイヤプロトコル MAPOS に準拠した高速スイッチ COREswitch について述べる．COREswitch は，OC-3c (155 Mbps)，OC-12c (622 Mbps)，OC-48c (2.4 Gbps) の 3 種類の回線を最大 16 回線まで混在して収容でき，87.04 Gbps の内部転送容量をもつ．高速ストリーム処理向きのデータパスや高効率の可変長フレーム転送アービタ等により，ハードウェアの簡略化による開発期間の短縮と高速性を両立した．想定する主なアプリケーションとして，広帯域映像 IP 転送システム及びインターネットバックボーンスイッチを取り上げ，それぞれの性能評価を行った．その結果，広帯域映像 IP 転送システムとして約 1.5 Gbps の映像業界向け非圧縮 HDTV 映像をスイッチングするのに十分な性能をもつことを明らかにした．また，インターネットバックボーンスイッチの構成の一例として，14 本の OC-12c 回線，2 本の OC-48c 回線の構成を用いて，同時競合転送試験を行った結果，フレームサイズが 512 byte 以上であればワイヤレートのスイッチングが達成できることを明らかにした．

キーワード　インターネット，リンクレイヤプロトコル，MAPOS，スイッチ，広帯域映像 IP 転送システム

1. まえがき

インターネットにおいては，近年のアクセス系におけるブロードバンドの普及や映像配信サービスなどへの需要増加に象徴されるように，高速化への要求が高まる一方である．このようななか，現時点で IP のデータ通信における 1 回線当りの通信速度は数百 Mbps〜数 Gbps が技術的，経済的制約のもとでの解となっており，このクラスをターゲットとするいくつかの IP パケット転送のためのリンクレイヤプロトコルが開発，利用されている．例えば，ATM，ギガビットイーサネット，10 ギガビットイーサネット，PPP-over-SONET/SDH [1]（以下，PPP と略記）などが代表的である．

しかし，ATM については，Cell Tax と呼ばれる IP パケットの細分化によるペイロードの減少，セル欠落によるパケット（フレーム）転送効率低下の問題に対処するためのアルゴリズムの複雑さ，コネクション設定のオーバヘッド，異なる装置間での相互接続性，などの問題がある．ギガビットイーサネットについては，WAN 接続時における保守，監視機能の不足が問題となり，キャリヤクラスの WAN 接続品質を提供するためには，SONET/SDH への再マッピングが必要となる．また，速度の改定のたびに仕様策定作業が必要である．PPP については，HDLC (High-level Data Link Control) フレームを単位として転送するためオーバヘッドは低いが，point-to-point の接続形態に限定される．

筆者らは，既存のプロトコルがもつこれらの問題を解決するため，リンクレイヤプロトコル MAPOS（メイポス：Multiple Access Protocol over SONET/SDH）[2], [3] を提案している．MAPOS は PPP と同じく，専用線の標準である SONET/SDH [4], [5] 上で HDLC 互換のフレームを転送する．SONET/SDH による速度，距離のスケーラビリティを備えながら，ATM に比べて低オーバヘッドな通信環境を提供する．更に

† 日本電信電話株式会社 NTT 未来ねっと研究所，武蔵野市
NTT Network Innovation Laboratories, NTT Corporation, 3–9–11, Midori-cho, Musashino-shi, 180–8585 Japan
†† 名古屋工業大学電気情報工学科，名古屋市
Electrical & Computer Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, 466–8555 Japan

MAPOS ではスイッチング機能（MAPOS スイッチ）の導入により多対多の接続が可能で，PPP に比べより柔軟な接続形態が構成できる．仕様の詳細は，IETF (Internet Engineering Task Force) の標準化文書である RFC (Request for Comments) 2171〜2176 として公開している．

本論文では，筆者らが開発を行った MAPOS 準拠のリンクレイヤスイッチ COREswitch（コアスイッチ）について述べる．COREswitch は，MAPOS に完全準拠したスイッチプロダクトで，OC-48c (2,488.32 Gbps)，OC-12c (622.08 Mbps)，OC-3c (155.52 Mbps) の 3 種類の回線を最大 16 回線まで混在して収容できる．回線ボードを相互接続するバックプレーンは，双方向 87.04 Gbps の転送容量をもつ．MAPOS の可変長フレームを内部で固定長のセルに分割せずそのままスイッチングすることで回線インタフェースのハードウェアを簡略化し，更に，高速ストリーム処理向きのデータパスや高効率の可変長フレーム転送アービタを実現することで，開発期間の短縮と高速性を両立した．

以下，**2.** では MAPOS の概要を説明し，**3.** では筆者らが開発した COREswitch のアーキテクチャの詳細を述べる．**4.** では COREswitch のアプリケーション例として，広帯域映像 IP 転送システム，及びインターネットバックボーンスイッチへの適用例を紹介し，**5.** ではそれらのアプリケーションへの適用性の観点から行った性能評価の結果を述べる．そして，**6.** で関連研究との比較検討を行い，最後に **7.** でまとめる．

2. MAPOS の概要

MAPOS は，前述の従来プロトコルの問題点を解決するために，筆者らが提案しているプロトコルである．図 1 に MAPOS ネットワークの基本的な構成を示す．MAPOS スイッチは全二重アクセス可能な複数の SONET/SDH ポートをもち，複数のノード（ホストや IP ルータ）がスイッチを経由して接続される．

MAPOS は，PPP と同じく，SONET/SDH 上で HDLC 互換のフレームを転送する．図 2 に，PPP 及び MAPOS のフレームフォーマットを示す．MAPOS では，PPP で固定値 (0xff) とされている Address フィールドにフレームの送信先を示すあて先アドレスを挿入し，MAPOS スイッチでこの値を用いたフレームスイッチを行うことにより多対多の通信を実現する（送信元アドレスを格納するフィールドはない）．スイッチどうしが多段接続されているときは各スイッチ

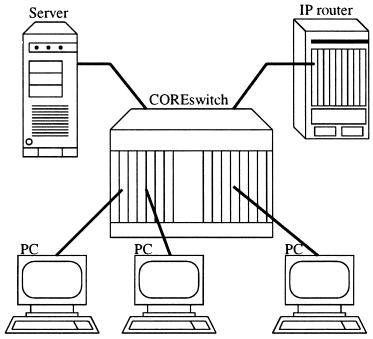

図 1　MAPOS ネットワーク
Fig. 1　MAPOS network.

Flag	Address	Control	Protocol	Information	FCS
01111110	11111111	00000011	8/16 bits		16/32 bits

(a) PPP-over-SONET/SDH

Flag	Destination Address	Control	Protocol	Information	FCS
01111110	8 bits	00000011	16 bits		16/32 bits

(b) MAPOS Version 1

Flag	Destination Address	Protocol	Information	FCS
01111110	16 bits	16 bits		16/32 bits

(c) MAPOS 16

図 2　PPP 及び MAPOS のフレームフォーマット
Fig. 2　Frame formats of PPP and MAPOS.

が中継動作を行い，最終的にあて先のノードへフレームが転送される．なお，MAPOS には，HDLC 互換フレームの 8 ビットの Address フィールドだけを使用する MAPOS Version 1 [2]，及び Address フィールドと Control フィールドの計 16 ビットをアドレスとして使用する MAPOS 16 [3] の二つのモードがある．MAPOS は以下のような特長をもつ．

- シンプルさと転送効率の高さ

コネクションレスのため，コネクション管理のオーバヘッドがなく，IPv4 や IPv6 との親和性 [6] も高い(注1)．更に，フレームを転送単位とするため，ATM より帯域利用率が高い．また，最大 64 kbyte の長大フレームをサポートし，IP パケットを分割せずに転送できるので，フラグメントに伴うホスト側でのヘッダ処理のオーバヘッドが少ない．

(注1)：MAPOS 上での IPv6 パケットの転送方法については，2002 年 11 月現在，インターネットドラフトとして公開中．

- シームレス性

LAN から WAN に至るまで SONET/SDH による継ぎ目のないネットワークが構成できる．

- 速度と距離のスケーラビリティ

SONET/SDH の速度体系がそのまま利用でき，スケーラビリティに優れる．また，SONET/SDH 伝送装置の利用によりセグメントの延長が可能である．

- PPP フレーム，SONET/SDH との互換性

PPP フレームとの互換性により，PPP のハードウェアの多くがそのまま利用可能であり，また，ソフトウェアも若干の修正で利用できるので，開発コストや期間が節約できる．また，SONET/SDH の部品もそのまま利用できる．

- プラグアンドプレイ

MAPOS のプロトコル群の一部である NSP (Node Switch Protocol) [7] によるアドレス自動設定や，SSP (Switch Switch Protocol) [8] による最適経路設定機能により，ユーザの負荷の低減や経路設定ミスによる障害の防止を図っている．

3. COREswitch のアーキテクチャ

3.1 特　　徴

COREswitch は，MAPOS の提案当初，プロトコル検証用のリファレンスマシンとして開発を始めた．一般の高速 IP ルータやスイッチでは，データ転送パスの使用効率の向上やデータ転送のスケジューリングの簡易化のために，転送データを固定長に分割してスイッチングする方式が多くみられるが，COREswitch では，MAPOS の可変長フレームをそのままスイッチングする．これにより，高速に動作する回線インタフェースにフレーム分割や再構成のための複雑なハードウェアをもたせず，この部分のハードウェアを簡略化している．そして，更に以下の方式を実現することにより，開発期間の短縮と高速性の両立を達成した．

（1） 高速ストリーム処理向きデータパス

SONET/SDH 回線制御，HDLC フレーム制御，クロスバスイッチ，単一の送受信 FIFO からなる簡潔なデータパス上に，異速度回線の効率的な収容，タグ制御情報の付加によるハンドシェイクの削減，FIFO 状態の先行監視による HOL (Head of Line) ブロッキング抑制機能等を実現し，可変長フレームを高速処理する．

（2） 高効率な可変長フレーム転送アービタ

クロスバスイッチへの転送競合に対し，スロット間の並列処理，フレーム転送要求のパイプライン処理，ユニキャスト/マルチキャスト転送の統合的扱い等により，可変長フレーム転送における複雑なアービトレーションを効率的に行う．

3.2 システム構成

COREswitch の外観を図 3 に，構成を図 4 に示す．複数の回線インタフェースカード (CIF: Cut-through Interface)，及びシステム全体の監視/制御用プロセッサ (IFP: Interface Processor) がバックプレーン (BP) で接続されている．BP 上では CIF，IFP 間をデータ転送用のクロスバスイッチ (XSW) と制御用のバス (Control Bus) で接続している．また，CIF や IFP からの XSW の使用要求を調停するためのアービトレーションモジュール (ABT) が搭載されている．回線速度は，現在，OC-3c，OC-12c 及び OC-48c の 3 種類に対応しており，OC-48c 及び OC-12c/OC-3c 切換型の 2 種類の CIF が利用可能である．COREswitch の主要諸元を表 1 に示す．

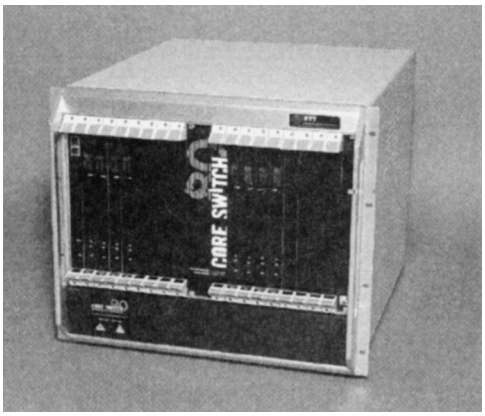

図 3　COREswitch の外観
Fig. 3　COREswitch.

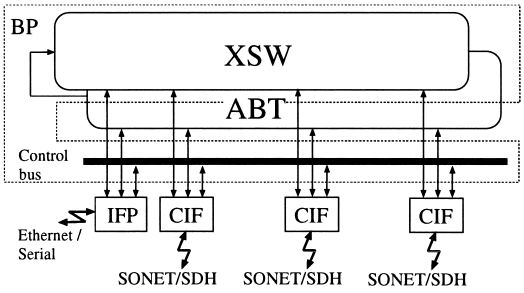

図 4　COREswitch の構成
Fig. 4　Block diagram of the COREswitch.

表 1　COREswitch の主要諸元
Table 1　Specifications of the COREswitch.

項目	機能/諸元
システムサイズ	430 mm×386 mm×500 mm
スロット数	17 スロット
回線ボードサイズ	233 mm×160 mm
BP	クロスバスイッチ (XSW)/制御用バス
・XSW	17×17 1 段 36 bit 幅パラレル伝送
・制御用バス	64 bit 非同期バス，最大 40 Mbyte/s
CIF	SONET/SDH，リンク制御
	アドレス検索，転送制御
・接続回線	OC-48c，OC-12c/OC-3c
・入出力 FIFO	各 512 kbyte
・回線制御	汎用 SONET/SDH LSI (OC-12c/
	OC-3c)，ASIC (OC-48c)
・HDLC 制御	FPGA (OC-12c/OC-3c),
	ASIC (OC-48c)
・検索転送制御	FPGA
・スロット実装	活線挿抜対応
IFP	NSP，SSP プロトコル制御,
	CLI (Command Line Interface),
	システム管理制御，L3 フォワード機能
・プロセッサ	Intel 960HD 33/66 MHz
・メモリ	RAM 128 Mbyte，ROM 16 Mbyte
・検索転送制御	FPGA
・インタフェース	シリアル/Ethernet
ABT	クロスバスイッチ制御
・主要ロジック	FPGA
・接続要求	17 スロット対応
・要求信号	各スロット 4 bit 幅
EMC	VCCI クラス A 準拠
電源	AC 100 V

3.3　高速ストリーム処理向きデータパス

スイッチ内では，図 5 に示すパスに沿って回線受信側 CIF (Ingress CIF) から回線送信側 CIF (Egress CIF) へフレームが転送される．実際には，各 CIF は図 6 に示すように送受信各方向のパスを備えているが，図 5 では省略して単方向のみを示している．Ingress CIF で受信したフレームを Egress CIF へ転送するまでの手順は次のとおりである．

（1） SONET/SDH 処理部で SONET/SDH のオーバヘッド処理と HDLC フレームの抽出を行い，HDLC 処理部にわたす．

（2） HDLC 処理部は，フレーム内のあて先アドレス値を抽出し，フレームデータとは別に ReqFIFO (Request FIFO) に格納する．また，残りのフレームデータを SONET/SDH 処理部から受信しながら Rx-FIFO へ書き込んでいき，そのバイト数をカウントしていく．

（3） ReqFIFO にあて先アドレス値が書き込まれると，RSE (Route Search Engine) は，その値を読出しアドレスとしてルートテーブルが格納されたサーチ RAM にリードアクセスし，フレームの転送先となる Egress CIF の番号等の情報を得る(注2)．

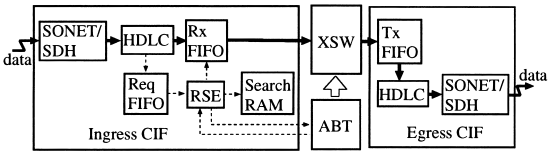

図 5　内部データ転送パス
Fig. 5　Data transmission path.

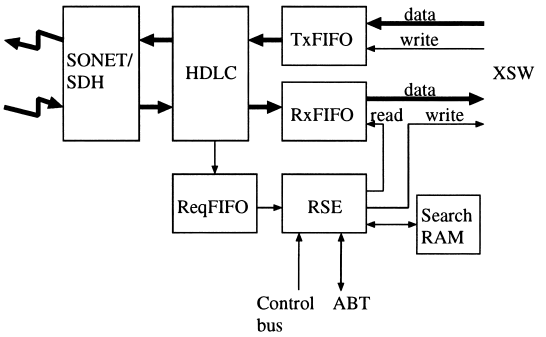

図 6　CIF のブロック構成
Fig. 6　Block diagram of the CIF.

（4） HDLC 処理部はフレームの全データの Rx-FIFO 内への受信が完了すると，バイトカウントによって得られたフレーム長を ReqFIFO に書き込む．

（5） フレーム長の ReqFIFO への書込みを検出した RSE は，XSW 設定要求を ABT へ送る．ABT から XSW 設定完了通知を受け取ると，RSE は，RxFIFO の読出し信号，及び Egress CIF の TxFIFO への書込み信号を発行する．

以上の処理により，RxFIFO からの出力であるフレームデータと書込み信号が XSW を介して Egress CIF へ転送される．RSE はフレーム長情報をもとに，フレームデータがすべて転送されるまで信号の発行を繰り返す．Egress CIF 側では，フレームの全データの TxFIFO への受信を待たずに外部回線へ出力するカットスルー動作を行う．

XSW は 36 bit 幅 17×17 の 1 段の構成で，16 枚の CIF と 1 枚の IFP 間をスイッチングする．各 CIF 及び

(注2)：各スイッチは経路交換プロトコル SSP により，他のスイッチの番号とそこへのネクストホップへ接続された自身の CIF 番号を知ることができる．一般に MAPOS アドレスはスイッチ番号と CIF 番号の組合せで構成されているため，この情報をもとにサーチ RAM 内に MAPOS アドレスに対する転送先情報を格納することができる．

IFP と XSW の間は，送受信それぞれにおいて，32 bit 幅のデータと 4 bit の制御情報とを BP 上の 80 MHz クロックに同期して伝送する．したがって，1 スロット当り 2.56 Gbps の送受信が同時に行える．また，ユニキャストのほかに，マルチキャスト，ブロードキャスト，ループバック転送も可能である．

スイッチ内のデータパスは送受信各方向について単一の FIFO からなるシンプルなものであるが，以下の機能により高速なストリーム処理を実現している．

- 異速度回線の効率的収容

Ingress CIF では，フレーム全体のデータが RxFIFO にそろってから XSW への転送を行う store-and-forward 方式を採用している．XSW 上では回線速度に関係なく 2.56 Gbps で転送が行われるので，低速度の CIF からの転送により XSW が長時間占有されることがなく，3 種類の速度の回線を効率良く収容できる．

- タグ情報によるハンドシェイクの削減

XSW 転送において，Egress CIF の TxFIFO への書込み信号，及びフレーム終了，バイトアライメント，エラーの有無を示す 4 ビットのタグ情報がフレームデータと一緒に転送される．データの流れる方向が 1 方向のみになるので，CIF 間のハンドシェイクが必要なく，オーバヘッドを削減できる．

- HOL (Head of Line) ブロッキングの抑制

RxFIFO と TxFIFO は論理的には 2 段のキューとして動作し，ABT による XSW の調停が完了し，TxFIFO に十分な空きがあることを確認してから，RxFIFO の先頭フレームが TxFIFO にキューイングされる．このとき Ingress CIF 側の RSE は，TxFIFO の空き待ち時間と ABT からの転送許可待ち時間を個別に監視し，どちらかが上限値を超えた場合にはこの先頭フレームを積極的に廃棄する(注3)．この機能は，HOL ブロッキングによるシステムのトータルのフォワーディング性能の低下を軽減する．

3.4 高効率な可変長フレーム転送アービタ

各 CIF や IFP から XSW へのデータ転送要求発生時には，他の CIF からのデータ転送要求と転送先の XSW ポートが競合することがある．ABT では，この競合の調停（アービトレーション）を行う．

COREswitch では，CIF のハードウェアの簡略化のため，回線から受信した可変長フレームを固定長セルに分割せず，可変長のまま内部転送する．このため，各 CIF (IFP) から XSW への転送要求の発生/終了のタイミングは全く任意であり，ABT には固定長セルの場合より複雑な処理が要求される．本 ABT では，(1)各 CIF (IFP) からの転送要求の発生/終了時の処理の並列化，(2)転送要求の受信から転送許可までの処理のオーバヘッドをいんぺいするためのパイプライン処理，(3)各 CIF (IFP) からの転送要求の有無の時系列サンプリングによるユニキャスト/マルチキャスト処理の統合などにより，転送要求のアービトレーションの高速化を実現した．これらすべての機能を ABT カード上の 1 チップの FPGA（60 万ゲート相当）内に実装している．詳細は文献 [9] を参照されたい．

3.5 自動構成制御のためのプロトコル処理

IFP では，制御用バスを介したシステムデバイス制御，CLI (Command Line Interface) の提供，IP アドレスを付与した CIF 間での IP パケット転送の実現などのほか，MAPOS の自動構成制御機能を実現する NSP (Node Switch Protocol) や SSP (Switch Switch Protocol) の処理を行う．

NSP は，アドレス自動割当てのためのプロトコルで，PC や IP ルータなどのノードが COREswitch に接続されると，ノードからの MAPOS アドレス要求が IFP へ転送され，IFP からノードへ MAPOS アドレスがセグメント内で重複なく割り当てられる．SSP は RIP などと同様の Distance Vector 型の経路情報交換プロトコルで，MAPOS スイッチ間で自動的にフレーム転送の経路情報を交換し経路表を作成する．

NSP については，ノードが受信する必要のあるマルチキャストフレームのあて先アドレスのリストを MAPOS アドレス要求と一緒に送信し，MAPOS スイッチがそれに基づいてマルチキャストアドレスと転送先 CIF の対応表を作成し，ノードへの不要なマルチキャストフレーム転送を抑制する機能拡張も行った(注4)．従来の IGMP snooping [10] や Cisco CGMP [10] などと同様の機能を簡単な方式で実現している．

4. COREswitch のアプリケーション

COREswitch は，既に動作検証まで終了したプロダクトであり，いくつかのシステムへの適用事例も存在する．ここでは，広帯域映像 IP 転送システム及びインターネットバックボーンスイッチへの適用例を紹

(注3)：TxFIFO への書込み信号を発行しない状態で RxFIFO のデータを読み出すことによりフレームを廃棄する．
(注4)：2002 年 11 月現在，NSP+としてインターネットドラフトで公開中．

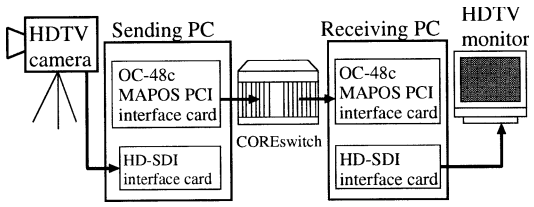

図 7　非圧縮 HDTV 転送システム
Fig. 7　Uncompressed HDTV transmission system.

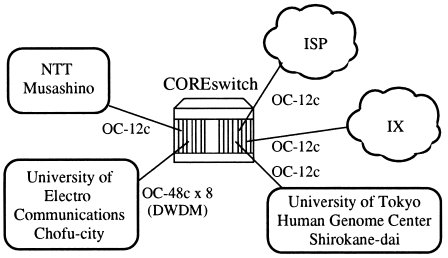

図 8　MAPOS 実験網
Fig. 8　MAPOS experimental network.

介する．

4.1　広帯域映像 IP 転送システム

筆者らは，MAPOS の高速性を生かしたアプリケーションとして，広帯域映像 IP 転送システム [11] の開発を行い，適用実験を進めている．本システムでは様々な種類の映像ストリームを IP でリアルタイム転送することが可能であり，なかでも，現在一般に普及しているギガビットイーサネットでは扱うことができない，約 1.5 Gbps の映像業界向け非圧縮 HDTV 映像ストリームの IP 転送が最も特徴的である．

図 7 は非圧縮 HDTV 映像を IP 転送する場合の基本的なシステム構成で，エンドシステム (PC) を OC-48c の MAPOS 回線で接続している．各 PC に筆者らが開発した OC-48c MAPOS PCI インタフェースカード [12] を搭載し，COREswitch を経由し相互接続している．1 対 1 通信だけでなくマルチキャスト通信も可能である．

HDTV カメラで撮影した映像データは送信 PC にて分割，IP パケット化され，MAPOS ネットワークを経由して受信 PC へ送られる．受信 PC では，受信した IP パケットデータをもとに HDTV 映像データを再構成し出力する．HDTV カメラ及び出力モニタと PC の間は，業務用映像機器用の HD-SDI 規格のインタフェースで，この部分を流れる非圧縮 HDTV 信号は 74.25 MHz サンプリング，10 ビット量子化の輝度信号と，37.125 MHz サンプリング，10 ビット量子化の 2 種の色差信号からなる 1.485 Gbps のディジタル信号である．この HD-SDI 信号の PC への入出力には市販の HD-SDI カードを用いている．

本システムは，市販の汎用 PC，市販の HD-SDI カード，及び MAPOS PCI インタフェースカードを用いて構成されており，PC 間の通信には MAPOS の特徴である 64 kbyte の長大フレームを用いている．長大フレームを用いた結果，エンドノードにおけるプロトコル処理の負荷を大幅に軽減することができ，汎用 PC でも十分な性能が得られることを実証できた [11].

4.2　インターネットバックボーンスイッチ

MAPOS の実フィールドへの適用性を検証するため，1999 年 3 月に NTT 武蔵野研究所を含む都内 5 地点を結ぶ MAPOS 実験網を構築し，今日まで運用を続けている．

本実験網は図 8 に示すように，各拠点がバックボーンの COREswitch を介したスター型の MAPOS ネットワークで接続された構成となっている．このバックボーンスイッチは NTT 武蔵野研究所内に設置されている．バックボーンスイッチと各拠点間は伝送装置を用いず，光ファイバ及び光アンプを用いた簡易な伝送路で接続されている．伝送距離の最長部分は 50.7 km である．実験開始時期の関係から多くの区間で OC-12c 回線を用いているが，順次 OC-48c 回線に置き換えて運用している．各拠点では，バックボーンスイッチとの接続点において MAPOS 対応の IP ルータを使用しており，各拠点の IP ルータ相互間で BGP (Border Gateway Protocol) を介した IP の経路交換を行っている．すなわち，本実験網は，独立したネットワーク運営主体である AS (Autonomous System) を BGP によって相互接続した，今日のインターネットバックボーンと全く同様の形態となっている．

運用開始から現在まで 3 年半の間，MAPOS のプロトコルや COREswitch に起因するトラブルはなく安定した運用を続けており，これらが実フィールドへの高い適用性をもつことがわかった．

MAPOS の実フィールドへの適用例としてこのほかに，「つくば WAN」におけるギガビットイーサネット LAN 間接続の例があり，筆者らが提案する GbE-MAPOS 変換装置 [13]〜[15] が用いられている．

5.　性 能 評 価

4. で述べた二つの代表的なアプリケーション，すな

わち広帯域映像 IP 転送システム，及びインターネットバックボーンスイッチへの適用を想定し，COREswitch の性能評価を行った．

5.1 広帯域映像 IP 転送用スイッチとしての評価

映像の種類として約 1.5 Gbps の非圧縮 HDTV 映像を転送する場合を想定し，スループット，及びフレームフォワード遅延の測定と評価を行った．

5.1.1 スループット

非圧縮 HDTV 転送においてはデータ帯域が一定であるため，あらかじめ必要な帯域を見積もることが可能である．非圧縮 HDTV 映像は約 1.5 Gbps の帯域を必要とすることから，今回は OC-48c 回線を用いて 1 回線当り一つの非圧縮 HDTV 映像を転送する場合を想定し，トラヒック間で回線競合が発生しない条件下で，以下の各場合のスループットを測定した．

（1） あて先固定ユニキャスト転送
（2） あて先変動ユニキャスト転送
（3） あて先固定マルチキャスト転送
（4） あて先変動マルチキャスト転送

具体的には,(1)では表 2 に示す八つのポート(注5)のペア間での全二重通信を行い,(2)では表 3 に示すポート 1〜ポート 8 までの各ポートが表に示すように 2 箇所のポートへ交互にフレーム転送を行い,(3)ではポート 1 が，ポート 2〜ポート 8 の 7 箇所へのマルチキャストを継続して行い,(4)では，ポート 1 から他の奇数番及び偶数番の全ポートへのマルチキャストを交互に行った．これらすべてにおいて，あて先競合は発生しない．

表 2 全二重通信時のポート割当て
Table 2 Port assignments for case (1).

port1 <-> port2
port3 <-> port4
port5 <-> port6
port7 <-> port8
port9 <-> port10
port11 <-> port12
port13 <-> port14
port15 <-> port16

表 3 あて先変動ユニキャスト時のポート割当て
Table 3 Port assignments for case (2).

port1 -> port2, port3
port2 -> port4, port5
port3 -> port6, port7
port4 -> port8, port9
port5 -> port10, port11
port6 -> port12, port13
port7 -> port14, port15
port8 -> port16, port1

（i） 測定系：測定系を図 9 に示す．MAPOS 対応の市販測定器である Anritsu 社製 MD1230A（以下，TG と略記）のトラヒックジェネレート機能を用いてトラヒックを生成し，それを光スプリッタで 16 分岐し，COREswitch の 16 個の OC-48c ポートへ入力する．上記トラヒックは，ペイロード内のデータがすべて 0 の IP パケットを収容した MAPOS フレームからなる．

（ii） 測定方法：(2),(4)では，TG より 2 種類のあて先の MAPOS フレームを交互に送信し,(1)では固定のユニキャストアドレス,(3)では固定のマルチキャストアドレスをあて先にもつ MAPOS フレームを連続送信する．各ポートのルートテーブルを操作して所望のポートへフレームが転送されるようにする．COREswitch 上でフレーム欠損の有無を観測しながら TG でフレーム間ギャップを調整し，60 秒間フレーム欠損がない状態を維持できた最大の送信トラヒックのビットレートを測定しスループットとした．この測定を，TG からの送信トラヒックのフレームサイズが表 4 に示す各値の場合について行った．なお，表中の値は，4 byte の MAPOS ヘッダと 4 byte の FCS (Frame Check Sequence) を含む値である．

（iii） 結果と考察：(1),(2)の測定結果を図 10 に,(3),(4)の測定結果を図 11 に示す(注6)．同じユニキャスト通信での(1),(2)のスループットの違いは，ABT

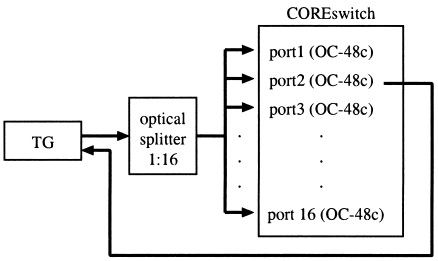

図 9 スループット測定系
Fig. 9 Throughput measurement system.

(注5)：1 枚の CIF は 1 本の全二重の回線をもち，この一つの回線をポートと呼ぶことにする．任意のポート間のデータ転送はバックプレーンのクロスバスイッチを経由して行われる．
(注6)：異なる種類のフレームを送信する場合の TG のフレーム間ギャップの下限の制約により,(2),(4)においてフレームサイズがそれぞれ 4,096，8,192 byte 以上の範囲では，フレーム欠損を起こさない限界近くの高いビットレートでトラヒックを送信することができなかった．このため，図にはこの下限のフレーム間ギャップを用いたときの送信トラヒックのビットレートをそのままプロットした．

表 4　測定に用いたフレームサイズ［単位：byte］
Table 4　Frame sizes used for the measurements.

54, 64, 128, 256, 512,
1,024, 1,280, 1,518, 2,048, 4,096,
4,472, 8,192, 16,384, 32,768, 65,288

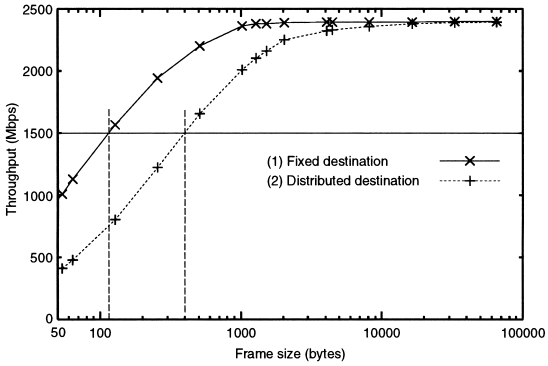

図 10　ユニキャスト時のスループット
Fig. 10　Throughput in unicasting.

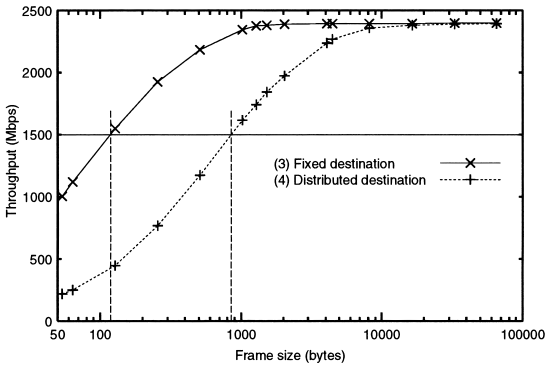

図 11　マルチキャスト時のスループット
Fig. 11　Throughput in multicasting.

におけるフレームの連続転送機能 [9] が稼働しているか否かの違いである.(1)の場合，各ポートに対してあて先ポートが同じフレームが連続して入力されるため，入力ポート上の RSE は，そのあて先ポートに対する ABT からの 1 回の転送許可ごとに，同じあて先ポートのフレームを連続して転送する．一方(2)の場合，あて先ポートの異なる 2 種類のフレームを交互に受信するため，入力ポート上の RSE は，1 フレームごとに ABT に対して転送要求を発行し転送許可を待つ．この処理の違いがスループットの違いの原因である.(3),(4)のスループットの違いも同じ理由による.

同じあて先変動がある場合での(2),(4)のスループットの違いは，入力ポート上の RSE からの転送要求に対し，ABT が入力ポートと出力ポート間のクロスバスイッチ接続を新たに行う必要があるかどうかの判断(注7)にかかる時間が，あて先ポートの数に依存することによる.(1),(3)においても同様の違いはあるが，連続転送機能により ABT が上記の判断を行う回数が少なくなるため，スループットへの影響が小さくなっている.

図 10，図 11 より,(1)〜(4)のそれぞれにおいてフレームサイズが 128，512，128，1,024 byte 以上であれば 1.5 Gbps のスループットが得られることがわかる．これらの値は，エンドシステムの PC 側で想定する 64 kbyte のフレームサイズ [11] に対して十分小さい値であり，COREswitch は非圧縮 HDTV 映像転送用スイッチとして十分なスループット性能をもつといえる.

5. 1. 2　フレームフォワード遅延

5. 1. 1 の(1)〜(4)について，スループットに加えフレームフォワード遅延の測定を行った．図 9 の系において，ポート 2 からの出力を TG へ入力し，TG の遅延測定機能を使用した．TG からのフレーム送信時にタイムスタンプを付加し，測定対象装置を経由して再度そのフレームを受信したときのタイムスタンプと比較して遅延を測定している．60 秒間の計測における遅延の平均値を測定値とした.

図 12 に,(1),(2)のユニキャスト通信時の遅延測定結果を示す.(3),(4)の場合については,(1),(2)の場合とほぼ同様の結果が得られたため省略する．測定は表 4 のフレームサイズについて行った．フレームサイズが最小の 54 byte では，(1) で 2.490 μs,(2) で 2.633 μs であり，最大の 65,288 byte では,(1)で 220.531 μs,(2)で 220.515 μs であった．COREswitch では，入力ポートにおいて store-and-forward 方式に基づくフレーム転送を行うため，フレーム長に応じて遅延時間も大きくなっている．フレームサイズが 512 byte までの範囲では前述の RSE/ABT 間の処理オーバヘッドの違いが(1)と(2)の測定値の違いに現れているが，1,024 byte 以上の範囲では入力ポートにおける RxFIFO へのフレームバッファリング時間が

(注7)：一度確立したクロスバスイッチの接続は，他のポートから同じあて先ポートへの転送要求がない限り解放しない．ABT はフレームの転送先が既に接続されているあて先ポートの集合のなかに含まれるかどうかを判断し，含まれる場合は直ちに転送許可を発行する．このとき，フレームが実際のあて先以外のポートへも転送されるのを防ぐため，TxFIFO への書込み信号を用いた制御を行っている.

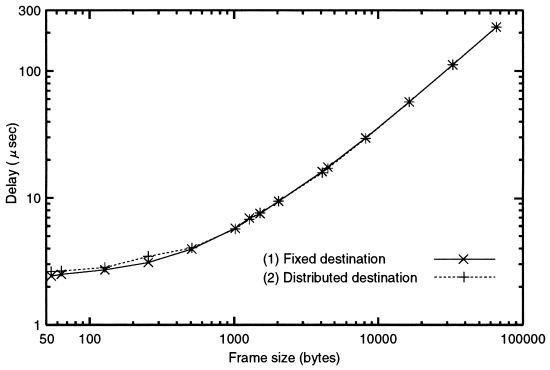

図 12　ユニキャスト時の遅延
Fig. 12　Forward delay in unicasting.

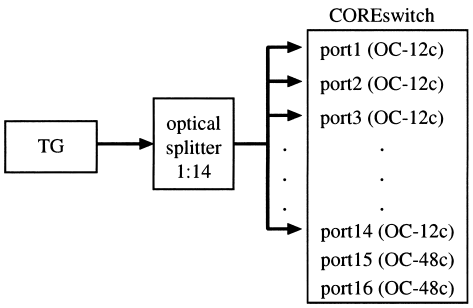

図 13　スループット測定系
Fig. 13　Throughput measurement system.

支配的になるため，この違いによる影響は見られない．筆者らが用いている 1080i 規格の HDTV システムでは，1 秒当りの映像フレーム数は 30 であり，1 映像フレームの表示時間は約 33 ms である．COREswitch のフレームフォワード遅延は，これに比べて十分小さい値であり，1 映像フレーム未満の転送遅延が要求される HDTV 映像転送システムへの適用も十分可能である．

5.2　バックボーンスイッチとしての評価

次に，インターネットバックボーンスイッチとしての使用を想定し，あて先競合が生じるトラヒックを入力した場合のスループットを測定した．14 本の OC-12c 回線と 2 本のアップリンク用 OC-48c 回線からなる構成を想定し，測定環境の制約の関係から，OC-12c 回線側にトラヒックを入力し，OC-12c 回線間のスイッチングにおけるスループットを測定した．

（i）　測定系　測定系を図 13 に示す．OC-12c 回線用の MAPOS 対応市販測定機である RADCOM 社製 Tetra2（以下，TG と略記）の出力を 14 分岐し，OC-12c のポート 1〜14 へ入力する．

（ii）　測定方法　TG から，ポート 2〜16 をあて先とする 15 通りのフレームを順番に送信するトラヒックを生成し，COREswitch のすべての OC-12c 回線へ入力する．各 CIF が自分のポートをあて先とするフレームを入力した際，これを廃棄しないようにするため，自分のポートをあて先とするフレームをポート 1 へフォワードするようにルートテーブルを設定する．

COREswitch 上でフレーム欠損の有無を観測しながら，TG から送信するトラヒックの回線速度に対する割合を 1%刻みで制御し，60 秒間フレーム欠損がない状態を維持できた最大の値を測定した．この測定を，TG から送信するトラヒックのフレームサイズが 256, 512，1,024 byte の三つの場合について行った．

（iii）　結果　フレームサイズが 256 byte の場合，OC-12c の回線速度に対し 89%のトラヒックの入力までフレーム欠損がない状態を維持することができた．また，フレームサイズが 512，1,024 byte の場合については，回線速度に等しい入力までフレーム欠損なしの状態を維持することができた．

本測定では，すべてのポートに対して同じあて先のフレームを同時に入力することを繰り返しており，あて先競合が非常に多い場合であると考えられる．このようなトラヒックの場合においても，フレームサイズが 512 byte 以上の場合 OC-12c 回線間でワイヤレートのスイッチングが行え，バックボーンスイッチとして十分な性能を有することが確認できた．

6.　関 連 研 究

高速 IP ルータやスイッチにおいては，文献 [16] のようにデータ転送パスの使用率の向上やスケジューリングの簡易化の観点から，回線上のデータ転送単位が可変長であっても，それらを固定長のセルに分割してスイッチングし，また再構成して出力するものが多い．これに対し，COREswitch では，可変長フレームをそのままスイッチングすることで高速に動作する回線インタフェースのハードウェアを簡略化し，更に，高速ストリーム処理向きデータパスや高効率な可変長フレーム転送アービタを実現することにより，開発期間の短縮と高速性の両立を達成している点が特徴である．

主なアプリケーションとして想定している広帯域映像の IP 転送については，以下のような報告例がある．

文献 [17] は，ネットワークアダプタ上にプロセッサ

を搭載し，プロトコル処理をホストプロセッサから独立させることで高速化を図るアーキテクチャの提案である．文献 [17] に報告のある OC-48c の例だけでなく，ギガビットイーサネットなどにも対応している．非圧縮 HDTV より帯域の狭い DV (Digital Video) を，IPsec を用いて実時間で暗号化している [18] のが特徴的である．

文献 [19]〜[21] は，非圧縮 HDTV の IP 転送に関する報告である．文献 [19] では，非圧縮 HDTV/SDTV のビデオ信号を 1,500 バイトの IP パケットに収容して SDH 上で長距離伝送する装置が紹介されている．ビデオ信号の入出力インタフェースと双方向の通信インタフェースをもち，この装置だけで HDTV/SDTV 機器間を結ぶ IP 伝送網の構築が可能である．しかし，完全な専用ハードウェアであり，様々なプロトコルが実装可能であるという汎用性の点では PC を用いた筆者らのシステムが優れている．

文献 [20] は，様々な種類の映像フォーマットを扱っている．文献 [20] には述べられていないが，非圧縮 HDTV の転送には高速 LAN 技術である HIPPI (High Performance Parallel Interface) の後継である GSN (Gigabyte System Network) を用いており，筆者らの SONET/SDH と比べて距離の制約がある．また，エンドノードに PC ではなく高価な高性能ワークステーションを用いている．

文献 [21] は，エンドノードに汎用 PC，ネットワークにギガビットイーサネットや OC-48c POS を用いている．汎用性の高い安価な構成要素を用いている点で筆者らのシステムに近い．しかし，文献 [21] の報告では非圧縮 HDTV を完全な形で転送するだけの性能は得られていない．また，報告のなかで筆者らのシステムを参照しており，筆者らのシステムの実装が先駆けている．

これらの例と比較して，**4.1** で述べた筆者らのシステムは，汎用性が高く安価な構成要素を用いながら，非圧縮 HDTV の転送に十分な性能を実現している点が優れている．これは，MAPOS の長大フレームによってエンドノードのプロトコル処理の負荷を大幅に軽減できたためであり，MAPOS の優位性を示すものである．

7. むすび

本論文では，筆者らが開発した MAPOS 準拠の高速スイッチ COREswitch について述べた．COREswitch は MAPOS 完全準拠のスイッチプロダクトで，1 回線当り OC-48c までの速度をサポートし，内部転送容量 87.04 Gbps のクロスバスイッチによりリンクレイヤでのスイッチングを行う．高速ストリーム処理向きのデータパスや，高効率の可変長フレーム転送アービタ等の実現により，ハードウェアの簡略化による開発期間の短縮と高速性を両立している．

想定する主なアプリケーションとして，広帯域映像 IP 転送システム及びインターネットバックボーンスイッチを取り上げ，それぞれの性能評価を行った結果，ギガビットイーサネットでは扱えない約 1.5 Gbps の映像業界向け非圧縮 HDTV 映像を扱うのに十分な性能をもつことを明らかにした．また，インターネットバックボーンスイッチの構成の一例として，14 本の OC-12c 回線，2 本の OC-48c 回線の構成を用いて，同時競合転送試験を行った結果，フレームサイズが 512 byte 以上であればワイヤレートのスイッチングが達成できることを明らかにした．

今後は，広帯域映像 IP 転送システムへの適用実験を進め，得られた知見をもとに，広域映像配信向き次世代高速スイッチの研究へと発展させていく予定である．

謝辞 COREswitch のハードウェア実装に御協力頂いた中央システム技研（株）小林正之氏，システムソフトウェア実装に御協力頂いた（有）ベルクマイクロシステムズ吉田敏明氏，日ごろから有益な御助言を頂いているサン・マイクロシステムズ（株）佐島隆博氏に感謝致します．また，MAPOS 実験網の構築及び運営に御協力頂いた，電気通信大学大学院情報システム学研究科伊藤秀一教授，同大学総合情報処理センター土屋英亮助教授，同大学電気通信学部情報工学科竹内郁雄教授，東京大学医科学研究所ヒトゲノム解析センター高木利久教授をはじめとする MAPOS 実験網関係者の皆様に感謝致します．

文　献

[1] A.G. Malis and W.A. Simpson, “PPP over SONET/SDH,” RFC-2615, June 1999.
[2] K. Murakami and M. Maruyama, “MAPOS - Multiple Access Protocol over SONET/SDH, Version 1,” RFC-2171, June 1997.
[3] K. Murakami and M. Maruyama, “MAPOS 16 - Multiple Access Protocol over SONET/SDH with 16 Bit Addressing,” RFC-2175, June 1997.
[4] “Synchronous Optical Network (SONET) - Basic Description Including Multiplex Structure, Rates and Formats,” ANSI T1.105-1995.
[5] “Network Node Interface for the Synchronous Digital

Hierarchy (SDH),” ITU-T Recommendation G.707, Oct. 2000.
[6] K. Murakami and M. Maruyama, “IPv4 over MAPOS Version 1,” RFC-2176, June 1997.
[7] K. Murakami and M. Maruyama, “A MAPOS version 1 Extension - Node Switch Protocol,” RFC-2173, June 1997.
[8] K. Murakami and M. Maruyama, “A MAPOS version 1 Extension - Switch-Switch Protocol,” RFC-2174, June 1997.
[9] T. Ogura, S. Yagi, T. Kawano, M. Maruyama, and N. Takahashi, “Crossbar Arbiter Architecture for High-Speed MAPOS Switch,” IEICE Trans. Inf. & Syst., vol.E83-D, no.5, pp.1028–1038, May 2000.
[10] “Multicast in a Campus Network: CGMP and IGMP Snooping,” http://www.cisco.com/warp/public/473/22.pdf
[11] 川野哲生，小倉　毅，清水健司，丸山　充，小柳恵一，“非圧縮 HDTV over IP システムにおける高速プロトコル処理技術,” 信学技報，NS2002-51, pp.47–50, June 2002.
[12] 清水健司，川野哲生，小倉　毅，丸山　充，“MAPOS 対応 OC-48c PCI カードの実現と性能評価,” 信学技報，NS2002-55, pp.9–12, June 2002.
[13] O. Okamoto, M. Maruyama, and T. Sajima, “Forwarding Media Access Control (MAC) Frames over Multiple Access Protocol over Synchronous Optical Network/Synchronous Digital Hierarchy (MAPOS),” RFC-3422, Nov. 2002.
[14] 岡本　治，原田啓司，丸山　充，“MAPOS を用いた LAN 間接続方式の検討,” 信学技報，NS2001-17, pp.37–42, April 2001.
[15] 原田啓司，岡本　治，“MAPOS 技術を用いた GbE-SONET/SDH 変換装置の開発,” NTT 技術ジャーナル，vol.14, no.3, pp.72–74, March 2002.
[16] N. McKeown, “Fast Switched Backplane for a Gigabit Switched Router,” http://www.cnaf.infn.it/ferrari/tfngn/doc/fasts_wp.pdf
[17] 小林伸治，的場宏純，都筑俊秀，陣崎　明，“Comet による OC48c クラスタの性能評価,” 情処学 HPC 研報，no.085-027, pp.157–162, March 2001.
[18] “Parallel and Distributed Systems Fujitsu Laboratory, RWCP,” http://www.comet-can.jp/PDSflab/#NP
[19] 栗林洋志，“デジタル放送用非圧縮映像伝送ネットワークの検討,” 映情学誌，vol.56, no.12, pp.1947–1950, Dec. 2002.
[20] 勝本道哲，原田雅博，“超高品質・映像音響技術の構築,” 信学技報，CQ2002-70, pp.31–36, July 2002.
[21] C. Perkins, L. Gharai, T. Lehman, and A. Mankin, “Experiments with Delivery of HDTV over IP Networks,” Proc. of the 12th International Packet Video Workshop, Pittsburgh, April 2002.

（平成 14 年 11 月 29 日受付，15 年 3 月 14 日再受付）

小倉　毅

平 4 神戸大・工・システム卒．平 6 同大大学院工学研究科システム工学専攻修士課程了．同年日本電信電話（株）入社．現在，NTT 未来ねっと研究所にてネットワークプロトコル，並列処理アーキテクチャ等の研究に従事．情報処理学会会員．

川野　哲生　（正員）

平 3 熊本大・工・電気情報卒．平 5 九大大学院総合理工学研究科情報システム学専攻修士課程了．平 8 同博士後期課程了．同年日本電信電話（株）入社．博士（工学）．現在，NTT 未来ねっと研究所にて超高速 IP 通信向けネットワークプロトコル及び SW/HW アーキテクチャの研究に従事．情報処理学会会員．

清水　健司

平 10 上智大・理工・電気電子卒．平 12 同大大学院理工学研究科電気電子工学専攻修士課程了．同年日本電信電話（株）入社．現在，NTT 未来ねっと研究所にてネットワークプロトコル，広帯域コンテンツ配信ネットワーク，並列処理アーキテクチャ等の研究に従事．

丸山　充　（正員）

昭 60 電通大大学院修士課程応用電子工学専攻了．同年日本電信電話（株）入社．主として，高精細画像情報提供システムの研究開発，ビデオ・オン・デマンドシステムの研究開発に従事．現在，超高速コンピュータネットワークと実時間並列分散アーキテクチャの研究に従事．工博．情報処理学会，日本ソフトウェア科学会，IEEE，ACM 各会員．

高橋　直久　（正員）

昭 49 電通大・応用電子卒．昭 51 同大大学院修士課程了．同年日本電信電話公社（現，NTT）武蔵野電気通信研究所入所．平 13・4 月より名工大電気情報工学科教授．この間，並列計算システム，ソフトウェア工学，ネットワークコンピューティングなどの研究に従事．工博（東工大）．情報処理学会，日本ソフトウェア科学会，ACM 各会員．