並列分散型高速通信スイッチ COREswitch

3G－6　高橋 直久　村上 健一郎　丸山 充　八木 哲　小倉 毅　川野 哲生
NTT 光ネットワークシステム研究所

1　はじめに

これまでに超高速データ通信向きリンクレイヤプロトコルとしてMAPOS (Multiple Access Protocol over SONET/SDH) を提案し，機能検証を進めてきた [1, 2]. 本稿では，MAPOS に準拠して実現した高速通信スイッチ COREswitch の概要を述べる．COREswitch は，17 の双方向ポート（各ポートの転送速度 320MB/s/方向）のノンブロッキング内部結合系を備えたメモリ分散型の非均質マルチプロセッサシステムである．PC，ワークステーション，サーバ，専用線，他の COREswitch との間の最大 16 の SONET/SDH リンク (OC-3/STM-1 (156$Mbps$) または OC-12/STM-4 (622$Mbps$)) での MAPOS 通信が可能である．

2　COREswitch 実現上の課題

2.1　MAPOS 準拠スイッチの要件

アーキテクチャ設計の観点から，MAPOS 準拠スイッチが満たすべき条件を考察する．

- 速度のスケーラビリティ
  MAPOS は，SONET/SDH の仕様に従い 156$Mbps$, 622$Mbps$, 2.4$Gbps$,... と回線速度がスケールアップするので，スイッチは高速で多様な回線速度に対応する必要がある．
- IP(Internet Protocol) との親和性
  MAPOS は，最大 64$kbyte$ の長大フレームをサポートし，コネクション設定を要しない，IP と親和性の高いプロトコルである．コネクションレスのため制御機構の簡単化が期待できる反面，可変長フレームを効率良く処理できる機構を実現する必要がある．
- 自動構成制御
  MAPOS では，ネットワークのノードにアドレスを自動的に割り当てるプロトコルとスイッチの経路制御を自動的に取得するスイッチ間プロトコルを規定し，ノードアドレスや経路表などの設定を自動化している．このため，経路表の動的な更新と高速な検索処理，および，ソフトウェアによるプロトコル処理が必要である．
- 小規模セグメント
  MAPOS では，一つのネットワークに接続できるノードを最大 60 台程度（基本プロトコル（$RFC2171$) の場合），あるいは，最大 8 千台程度（拡張プロトコル ($RFC2175$) の場合）までに制限することにより，適切なセグメントの規模を保つようにしている．ノードアドレスのフィールドが小さい点を活かして，高速な検索処理を実現することが望まれる．

2.2　高速通信スイッチの要件

高速通信スイッチは，前節の要件に加えて，次のような点を考慮する必要がある．

- 輻輳制御
  特定の回線への出力が集中したり，あるいは，高速回線から低速回線へのデータが連続するなどの事態が発生した場合に，フレーム廃棄などにより輻輳制御する．
- HOL($Head\ of\ Line$) ブロッキングの回復制御
  ノンブロッキングの内部結合系を用いても，外部回線で輻輳が生じたときに転送待ちフレームが生じる．このフレームの後続フレームは，他の空き回線への転送の場合でも，先頭フレーム (HOL) がブロックしているため，転送要求を出せないで待ち続ける．これは，回線の輻輳が続く場合に深刻な問題になる．HOL の要求を後回しにしたり，廃棄するなどにより，後続フレームを長期間ブロックしないようにする．
- 高速低遅延転送制御
  スイッチを多段接続した場合でも実用上問題ないように，低遅延の内部転送機構を実現する．また，ユニキャストだけでなく，ブロードキャストとマルチキャスト通信の高速化が望まれる．これら 3 種類の通信が混在する場合にも効率的，かつ公平に処理する機構が必要である．

3　COREswitch の実現

3.1　構成

COREswitch は，図 1 に示すように，外部の通信回線対応のプロセッサ (CIF_{622},CIF_{156}. 以下総称して CIF)，システム全体の監視／制御用のプロセッサ (IFP)，アービトレーションモジュール (ABT)，クロスバスイッチ (XSW)，制御バス ($C-bus$) からなるメモリ分散型並列システムである．表 1 に，これまでに実現した構成要素の機能と諸元を示す．

3.2　特徴

COREswitch アーキテクチャの特徴を以下に示す．

- 高速ストリーム処理
  図 2 に示す，SONET/SDH 回線制御，HDLC フレーム制御,XSW, 送受信 FIFO (TxFIFO, RxFIFO) からなる単方向の簡明なデータパスを用いて，可変長フレームのストリームを高速処理する．RxFIFO，TxFIFO と回線の間は回線速度で動作させ，残りは XSW の速度で動作させることにより，多様な回線速度のストリームを高速で効率的に処理する．XSW では，32 ビット幅データと 4 ビット制御情報とを 80MHz で同期伝送して，ポート

COREswitch: A Parallel and Distributed High-speed Communication Switch.
Naohisa TAKAHASHI, Ken MURAKAMI, Mitsuru MARUYAMA, Satoru Yagi, Tsuyoshi OGURA and Tetsuo KAWANO
NTT Optical Networks Laboratories
3-9-11 Midori-cho Musashino-shi Tokyo 180 Japan

表 1: COREswitch の構成要素

略称	機能／諸元
	プロセッサモジュール
CIF_{622}	1 ポート $OC-12/STM-4(622Mbps)$MAPOS SONET/SDH 回線制御，HDLC フレーム制御 フレームのフォワーディング，キューイング
CIF_{156}	1 ポート $OC-3/STM-1(156Mbps)$MAPOS
IFP	MAPOS プロトコル処理，構成管理 保守インターフェース
	内部結合モジュール
XSW	17 ポート（双方向）ノンブロッキング転送 $320MB/s(2.56Gbps)$/方向 ユニキャスト／マルチキャスト
ABT	XSW アービトレーション
	制御系
$C-bus$	64 ビット制御バス 最大 40MB/s

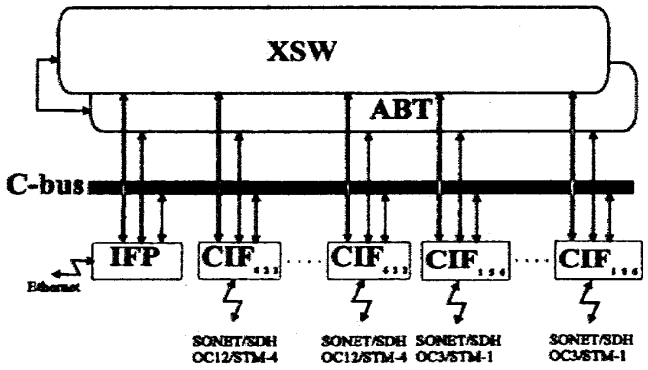

図 1: COREswitch の構成

当たり $320MB/s(2.56Gbps)$/方向 の高速低遅延転送を実現し，$2.4Gbps$ 回線まで対応可能にしている [3].

- 自動構成制御
  IFP のソフトウェアによる MAPOS プロトコル処理により，経路情報を更新／管理して，ネットワークの構成変化に自動的に対応可能にしている．また，回線の挿抜や CIF の活線挿抜を実現し，IFP が CIF の状態を監視して，自動構成制御に反映させている．これにより，回線の障害時や保守時にも他の回線での通信を継続して利用できるようにしている．

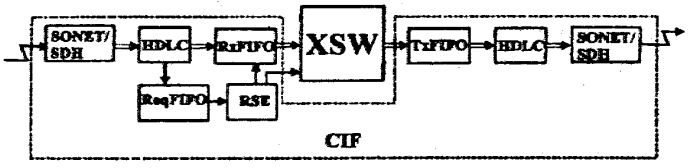

図 2: COREswitch におけるストリーム処理

- テーブル駆動型分岐制御
  CIF は，図 2の RSE(Route Search Engine) で経路表をキャッシングし，この表に従いフレームの転送／複製を制御する．この機構は，外部の出力回線への通常のフォワーディングだけでなく，機能分散型マルチプロセッサにおいて，受信フレームに対応したプロセッサへ処理を高速に分岐させる場合にも適用可能な汎用性の高い機構である．たとえば，他のフレームのフォワーディグ性能に影響を与えずに，IFP がフレームをタッピングしてソフトウェア処理を加えたり，CIF が他の CIF への転送と同時に複製フレームを IFP に送ってトラフィックに関する統計処理を依頼するなどが可能になる．

- ルックアヘッドキューイング
  図 2において，RxFIFO と TxFIFO は論理的には 2段のキューとして動作し，RxFIFO の HOL が TxFIFO にキューイングされる．このとき，RxFIFO は，TxFIFO の状態を監視し，TxFIFO へのキューイング開始のタイミングを遅延させたり，HOL を廃棄して次のフレームをキューイングする．これは，HOL ブロッキングを軽減する．

- 予約型アービトレーション
  固定長セルに比べて，可変長フレームに対するマルチキャストは，実現が難しい．XSW に対する転送要求のアービトレーションに予約機能を導入して，ユニキャストとマルチキャストのフレームが混在したストリームを高速かつ公平に転送する．

- コンパクトな実装
  前述の簡明なストリーム処理により，機能を単純化するとともに，HDLC フレーム制御，アービトレーション制御，経路表検索，XSW 制御のチップ化により，コンパクトな実装になっている．

4 おわりに

超高速データ通信向きリンクレイヤプロトコル MAPOS 準拠の高速通信スイッチ COREswitch の概要を述べた．現在，$622Mbps$ および $156Mbps$ の CIF が稼働しており，これらの回線について，COREswitch 間および PC／WS (S-bus または PCI) と COREswitch 間での MAPOS 通信が可能である．今後，実環境での評価とともに $2.4Gbps$ 対応の CIF の設計開発などを進める予定である．

謝辞
本研究をご支援下さる分散ネットワークシステム研究部山田茂樹部長，ならびに，御助言，御協力いただいた吉田敏明氏，小林正幸氏，佐島隆博氏に感謝します．

参考文献

[1] Murakami, K. and M. Maruyama, "MAPOS - Multiple Access Protocol over SONET/SDH, Version 1", RFC2171, June 1997

[2] 村上, 高橋, 丸山, 八木, 小倉, 川野, "超高速データ通信用プロトコル MAPOS の概要", 情処第 56 回大会 3G-05, Mar. 1998.

[3] 丸山, 高橋, 八木, 小倉, 川野, "COREswitch のハードウェアアーキテクチャ", 情処第 56 回大会 3G-07, Mar. 1998