ハイパフォーマンス
コンピューティング 83－8
（2000．10．6）

# 光インタコネクトを使った高速転送におけるフロー制御手法

上野龍一郎† 稲沢悟†† 西宏章†††
工藤知宏††† 天野英晴†

RHiNET-3/SWはフロア内に分散するPCを光インターコネクトで接続する並列システム用スイッチである．先に実装したRHiNET-2/SWよりも大規模かつ信頼性の高いネットワークを構築するため,creditベースのフロー制御によりRHiNET-2/SWの様なバッファサイズに起因する伝送距離の制限を取り除き，再送バッファを備えることでスイッチ間で信頼性を持たせるなどの高機能化を図った．本稿ではRHiNET-3/SWのフロー制御法をシミュレーションにより評価し，並列ネットワークへの影響等について考察する．

# Flow Control Method in high speed transfer using Optical Interconnect

RYUICHIRO UENO ,† SATORU INASAWA ,†† HIROAKI NISHI ,†††
TOMOHIRO KUDOH ††† and HIDEHARU AMANO †

RHiNET-3/SW is an advanced switching fabric for a parallel system which consisting PCs distributed one or several floors in a building connected by optical interconnect. Although previous switching fabrics RHiNET-2/SW achieved satisfactory performance under a certain condition, a credit based flow control is introduced to connect larger number of PCs distributed in wider area. For reliable communication, re-transmission buffer is also added. In this paper, transmission performance of RHiNET-3/SW is evaluated with computer simulation and compared with RHiNET-2/SW.

## 1. はじめに

クラスタによる並列分散処理は，安価なPS/WSクラスタを数十台から数百台でネットワークを構成し、大型計算機に匹敵する非常に高い性能を提供するシステムとして脚光を浴びている．

従来の高性能クラスタシステムは，計算機間の接続にMyrinet[1]などのSystem Area Network(SAN)を用いたものが多い.SANは基本的にパケットを廃棄することのない低遅延大容量の通信ネットワークで，多くの並列アプリケーションで要求される通信性能を満たしている．しかし従来のSANはリンク長やネットワークトポロジに制限があるため，計算機室内などに設置されたクラスタ専用のネットワークとして用いられてきた．

† 慶應義塾大学 理工学部
Dept. of ComputerScience,Keio University
†† 日立通信システム株式会社
Hitachi Communication Systems,Inc.
††† 新情報処理開発機構 並列分散システムアーキテクチャつくば研
RWCP Parallel and Distributed System Architecture Lab.

SANに匹敵する性能を持つネットワークで，フロア内やビル内に分散した計算機群を接続することができれば，日常の業務に用いている計算機の余剰性能を用いて高性能並列処理環境を実現したり，別々の計算機室に設置された様々な計算機群を接続して統合した処理環境を実現することができると考えられる．

フロア内やビル内の計算機群を接続するためには従来LANが用いられてきたが，LANには，遅延が大きいこと，大きなbi-section bandwidthを提供することが難しいこと，混雑時にはパケットを廃棄する可能性があること，パケットの到着順序が保証されないことなどの問題がある．

そこで我々は,LANとSANの両方の利点を追及した，LASN(Local AreaSystem Network)を提唱している[2]．LASNは，パケットを廃棄しない低遅延大容量ネットワークというSANの性質を保ちつつ，フロア内やビル内の計算機群を接続するのに十分なリンク長とトポロジの自由度を提供する．

近年の大容量ネットワークでは，光伝送を使った伝送が用いられるようになりつつある．光伝送を用いれば,10Gbpsクラスの伝送が比較的容易に実現できる．

しかしこのような大容量の伝送でパケットを廃棄しないためには, 伝送路上の伝搬遅延を考慮したフロー制御が必要になる. 伝送路長が長くなれば伝搬遅延も長くなるので,LASNがターゲットとする領域ではフロー制御手法が重要なポイントとなる.

我々は, これまでに RHiNET-1/SW[3],RHiNET-2/SW[2] と呼ぶLASN用スイッチを開発してきた. これらのスイッチは,Windowベースのフロー制御方式を用い, ビット誤り率がほぼ無視できるほど低い ($10^{-20}$) デバイスを用いた. しかし, このフロー制御方式は, 使用できるリンク長が制限されること, デバイスが高価であることなどの問題があった. そこで現在, クレジットベースのフロー制御方式を用い再送機構を持つスイッチ RHiNET-3/SW の開発を行なっている.

本報告では, これらのフロー制御手法について比較検討を行なう.

## 2. RHiNETの概念

RHiNETは,PCIバスに装着するネットワークインターフェース RHiNET/NI, ネットワークのスイッチとなる RHiNET/SW およびこれらを接続する光インタコネクトにより実装される.

### 2.1 RHiNET-2/SW

RHiNET-2/SW は LASN 特有の要求事項に対応するよう設計された, 高速インタコネクトスイッチで, SAN 同様の Asynchronous wohmhole routing[4] を利用した. また, テーブルルーティングと, 多数の仮想チャネルによる拡張構造化バッファ法[3] により, ループを含む自由なトポロジを実現した. さらに,ECC の付加により, ハードウェアでエラーレートを低くする事で上位プロトコル層における通信品質を保障している.

RHiNET-2/SW は, 光インターコネクトの転送速度に対応するため, 10bit×800MHz の光リンクからのデータをマルチプレクサにより 80bit×100MHz に変換してコア LSI においてフレーム処理をする (図 1). 各ポートは 16 の仮想チャネルをもち, 各仮想チャネルは 512 フレームのデータを保持するバッファを持つ.

### 2.2 RHiNET-2/SW のフロー制御

RHiNET-2/SW コア内において, それぞれの仮想チャネルはシステム LSI 内の Slack Buffer[1] (4KByte × 16 channel × 8 port) で実装し, フロー制御は単純な Slack Buffer based 制御を採用した. この方法は, 受信側のバッファに一定量のデータが溜ると, CA(Channel Announce) パケットの所定の bit を 0 にして送信側に対して送り, 転送の停止を要求する [1]. 送信側と受信側の間には, 長距離の光ケーブルによる遅延があるため,CA パケットが到着して送信が停止するまでの間に相当量のデータが送られてしまう. したがって, 受信側のバッファは, この間送られたデータを蓄えるだけの余裕を見て CA を発生してやる. 受信側のバッファ中のデータ量が一定以下になると, 再び CA パケットの所定の bit を 1 にして送信側に送り, データ転送の再開を促す.

このフロー制御法は, ハードウェアがシンプルに構成でき, 制御情報も幅をとらないため, 高速転送においてその効果を発揮した.

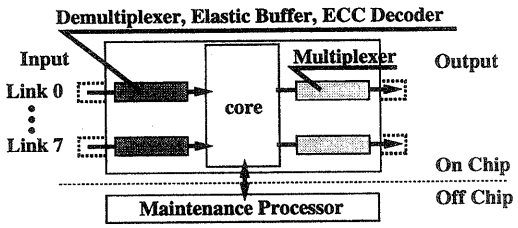


図 1 RHiNET-2/SW のブロック図

### 2.3 RHiNET-2/SW の問題点

RHiNET-2/SW は当初の RHiNET-2 の要求を満足する性能を得ることができたが, フロー制御方式は, 以下の問題点を持つことが明らかになった.

- RHiNET-2/SW では, 光伝送デバイスが非常にビット誤り率が低く, 現実の用途ではビット誤りは発生しないと仮定した. しかしこのようなデバイスは一般に高価である. 今後普及すると考えられる 10Gbit Ethernet などで用いることが検討されている光デバイスは,$10^{-12}$～$10^{-15}$ 程度のビット誤り率を持つと考えられ, これらのデバイスを用いて信頼性のあるネットワークを実現するには再送機構を持たせる必要がある.
- RHiNET-2/SW が現在用いている Slack Buffer based フロー制御はバッファ量により, 接続する距離が制限される.

## 3. RHiNET-3/SW

### 3.1 RHiNET-3/SW の概観

RHiNET-3/SW は 12.5G bit/s の並列光信号を 8 ポート入出力し全体として 100G bit/s の通信容量を実現する. スイッチ LSI 内では,1 ポートあたり 80bit × 125MHz の速さでのスイッチング処理を実現し, 光インターフェースと同等のパフォーマンスを発揮する. RHiNET-3/SW コアの構造を図 2に示す

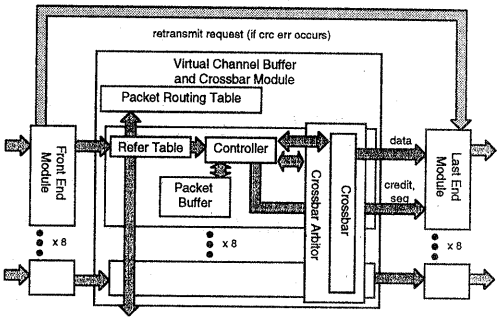


図 2　RHiNET-3 コア構成図

新たに生じた要求を満足するために，RHiNET-3/SW で新たに備えた機能は次の通りである．

- CRC チェックと再送機能
  送信側ポートはシーケンス番号，仮想チャネルのクレジット，および CRC の付与を行う．受信側ポートはこれを受け取り，CRC チェック，シーケンス番号のチェックを行い，エラーが検出されると，再送を行う．受信側ポート，内蔵するバッファに蓄えられた出力データのコピーを元に再送を行う．これにより，個々のリンク間での信頼性を確保する．
  再送制御用のバッファは各出力ポートあたり 128 フレーム分持っている近距離用ポートと 1024 フレーム持っている長距離用ポートがあり，それぞれ 250m,1000m までの光伝送長に対する再送をサポートする．
- クレジットベースのフロー制御機能
  クレジットベースのフロー制御は RHiNET-2/SW で採用していた Slack Buffer based フロー制御に比べてハードウェアが複雑化するが，ASIC の高密度化により実装が可能になった．オンチップメモリの有効活用，長距離リンクへの対応が可能になり，また仮想チャネル数が多量にとれるという利点もあり，なるべくバンド幅を失なわない形で採用した．
  具体的な方法は次節に述べる．
- ソースルーティング機能
  RHiNET-3/SW は RHiNET-2/SW 同様，テーブルによってルーティングを行い，縮約構造化チャネル法 [5] により，デッドロックを回避する．ただし，多数のノードを接続した場合のルーティングの自由度を増すため，仮想チャネル数を 16 から 64 に拡張した．しかし，このルーティング法はテーブル引きに一定の時間を要することから，低遅延の交信用に，Myrinet 同様のヘッダ首切りルーティングを導入した．
  RHiNET-3 のパケットヘッダにはソースノードから 8 つ目のスイッチまでのルーティング情報を記す事ができ，ルーティングを行う毎にこの情報を順に使っていく．この方法でテーブル引きのロスを回避する．
- パケットのカットスルー機能
  送信ポートがあるチャネルのパケットをクロスバーから入力されつつ出力ポートに送信している時，何らかの理由でクロスバーからの入力が途絶えた，もしくは，クレジットがなくなる等の理由で受信側に送信できなくなった場合，パケットをカットして，別のチャネルのデータの送信をする事ができる．その時中断のパケットを受信側に送信し，再開の時には再開用のビットをたてヘッダを付加する事により，独立したパケットしてルーティングを可能にする．この際，パケット同士の FIFO 性は保証する．

表 3.1に SW2 と SW3 のハードウェア性能の比較を示す．

| 項目 | SW2 | SW3(予定) |
|---|---|---|
| I/O 信号バンド幅 | 800Mbps | 1000Mbps |
| 内部動作周波数 | 100MHz | 125MHz |
| スループット | 64Gbps | 80Gbps |
| 仮想チャネル数 | 16 | 64 |
| バッファサイズ | 4KByte | 2KByte |

表 1　RHiNET-2/SW と RHiNET-3/SW の主なハードウェア性能比

### 3.2 RHiNET-3/SW Frame の機能

RHiNET-3 フレームフォーマットを図 3に示す．

光インターフェースの転送速度を損なわぬよう，LSI 内の処理単位は 1 フレーム 80bit，そのうち payload 部分を 64bit 確保し，残りの 16bit を Credit，再送 Sequence,CRC 等の制御ビットとした．16 ビットにおいてこれらの機能を実装するのは難しいので，2frame 単位でフレーム処理をする事にし，制御用に 32bit 確保している．

RHiNET-3 のフレームの機能の特徴は，クレジット情報と再送用シークエンス情報がフレームの同じビット上に記され，クレジットか再送シークエンスかを示すフラグで識別される事である．必要に合せて両情報ををかわるがわる出し，浮いた分をペイロードと CRC 情報に回す事により，信頼性を失わずに効率のよい転送を実現する．

送信側は,"Credit or Sequence" で Credit か Sequence かを判断する.0 であれば MicroFrame Info には受信側で受けとられた最後のシークエンス番号 (7 bit) が記してある.1 であれば, 仮想チャネル番号 (7 bit) と, そのチャネルのクレジットの増加量 (2bit) が記してある.

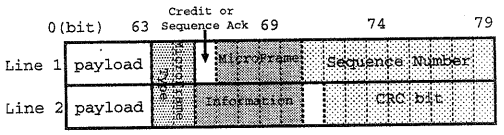


図 3 RHiNET-3 フレームフォーマット図

## 4. クレジットベースのフロー制御

各仮想チャネルは クレジットベースのフロー制御を行う.

リセット時, 各仮想チャネル送信バッファは受信側バッファの容量分のクレジットを保有している. 送信バッファ側はフレーム発信時に 1 クレジットを消費する. 受信側は受信したフレームをバッファに送信して新たに書きこむ. このバッファの読み出しによりクレジット数 (最大 4) を送り出す. 送信側はクレジット情報を受け取った後, クレジットを増加させる.

クレジットベースのフロー制御は Slack Buffer Based のフロー制御に比べ、リンク長の制限がないという利点がある. 最大バンド幅 × 往復分の転送遅延に相当するバッファ量が得られない場合,Slack Buffer Based のフロー制御では, 仮想チャネルバッファがあふれ, パケットがドロップされてしまう. これに対しクレジットフロー制御下ではバンド幅が低下するだけで, データは正常に受け渡される.

クレジットベースのフロー制御方法にも種類が複数存在するが, バンド幅の確保を優先するという方針のため, 制御情報にビットを多量に使う HIPPI-6400[6] の様な手法は困難である. 本スイッチでは, バッファが新たに空いた分のクレジットを送るという ATM の ATLAS-I Chip[7] に似た手法を用いた. この場合, パケットロスによるクレジットスターベーションの問題があるが, これはクレジット情報も再送バッファに格納するので問題ない.

## 5. シミュレーションによるフロー制御の性能評価

我々は RHiNET-3/SW の構築, 実装と並行してシミュレーションによるシステムの性能評価を行った.

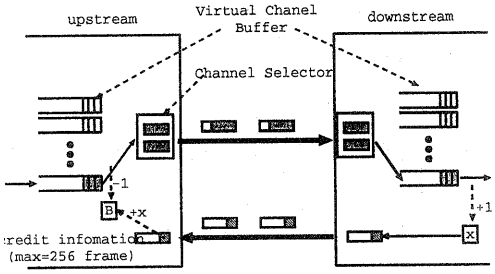


図 4 RHiNET-3 フロー制御図

ここではスイッチを含めた RHiNET-3 ネットワークの使用時において,RHiNET-3 スイッチの構成要素にかかる負荷を測定し, 従来の RHiNET-2/SW のシステムと比較する.

### 5.1 シミュレーションモデル

RHiNET のパケット転送方式に基づいて, スイッチのフリットレベルシミュレータを C++で構築した.

#### 5.1.1 時間 と フリットサイズ

全てのフリットは 2 フレームの集合で構成されている. フリットは 1clock で 1hop 移動する.

#### 5.1.2 ユ ニット

シミュレータの次の 3 つの構成要素から成る.

- ノード: PC のモデルであり, パケット発生器と受信器と 1 つのポートで構成され, 送信, 受信を繰り返す.
- スイッチ: 8×8 のポートを備えたスイッチでクロスバ部分で制御する
- コネクション: 光リンクにあたるコネクションには 30 フリット入る FIFO バッファを用意した. 計算では 60 メートル程度の光ファイバに相当する

#### 5.1.3 ポ ー ト

ポートはノード, スイッチの中で用い, それぞれのコネクションと接続され, 毎クロック, フリットを入出力する. 今回評価対象となる, 仮想チャネルバッファ制御の部分はここの中で記述されている.

1 ポートあたり 16 の仮想チャネルをもち, 各仮想チャネルは 128(256 フレーム分) の容量のバッファを持つ. それぞれの仮想チャネルはクレジットベースもしくは Slack Buffer based フロー制御で各フリットをコントロールする.

SW2 では CA を送りだす水準線を決めなければいけない.GO をバッファ容量 0 から 60, STOP を 68 に指定した.

#### 5.1.4 トポロジとトラフィック

- ネットワークのトポロジとそのサイズ

  スイッチで 4 × 4 の torus 網 を組みそれぞれに

ノードを一つ接続した torus 網モデルと, スイッチ 4 個を完全接続し, それぞれにノードを 5 つ接続した完全結合網モデル, この 2 つのトポロジをシミュレーションモデルに使用した.

- ルーティング
  各スイッチ毎にルーティングテーブルを用いる方法で,x 方向優先でルーティングを行なった.
- トラフィックソース通信方法としては, すべてのノードが定間隔に自分以外のノードに等確率で送信先ノードをランダムにしてメッセージを転送する, ランダム通信を使用した. また、torus 網においては, ホットスポット通信を追加した. この場合, すべてのノードが定間隔に自分以外のノードに等確率に特定のノードに送信する.

## 6. 測定とその評価

### 6.1 torus 網

ランダム転送の評価の図を図 5に示す.

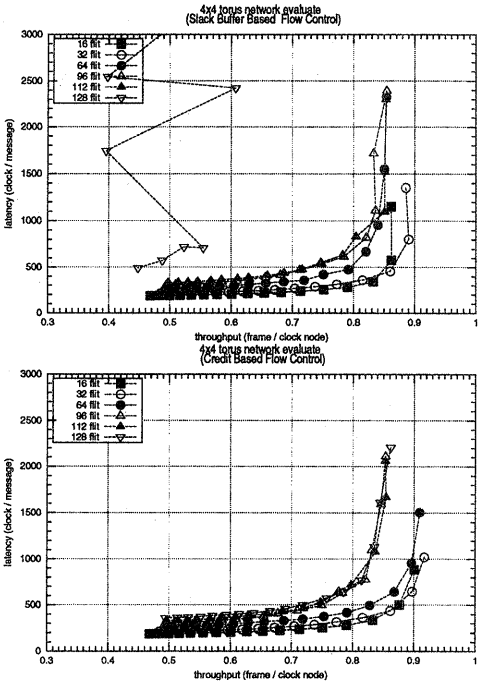


図 5 torus 網ランダム転送における RHiNET-2/SW(上) と RHiNET-3/SW(下) のフロー制御の評価

ホットスポットの評価の図を図 6に示す.

### 6.2 perfect 網

完全結合の評価の図を図 7に示す.

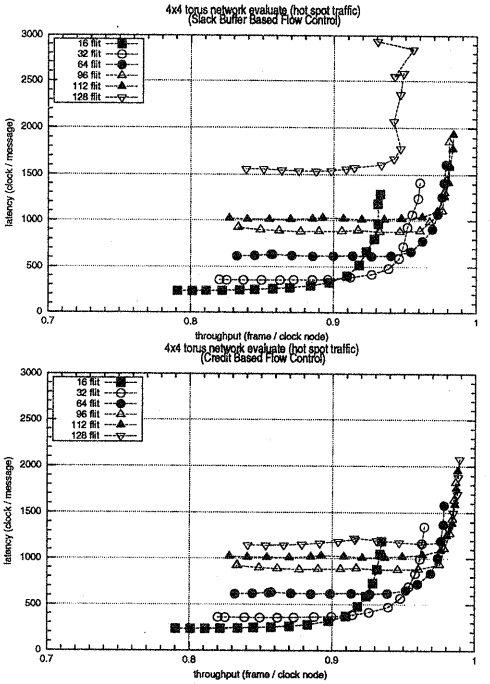


図 6 torus 網ホットスポットにおける RHiNET-2/SW(上) と RHiNET-3/SW(下) のフロー制御の評価

### 6.3 ランダム転送による評価

Slack Buffer based フロー制御ではメッセージサイズによって一部レイテンシとスループットに悪影響を及ぼすものがあった. これはパケットサイズが丁度 go-line より少し上に達していたため, 本来は次のパケットの一部が送れたにもかかわらず, 止っていたための弊害と思われる. この様な場合を避けるためにはアプリケーション側で, パケットサイズを制限する必要がある.

更にパケットサイズが仮想チャネルの容量に近づいた 128flit において, Slack Buffer based フロー制御の性能は落ち, ばらつきは多くなった. これは, パケットのサイズが仮想チャネルの容量に近かったため, パケットのトレイラに近い部分が最後に少しだけ残っている状態が非常にネットワークに悪影響をあたえているものと思われる. この状態は次に Go-line まで水準線が下がるまで送りだされないため, レイテンシのもとになる. 更に, 同じバッファ内にある次のパケットが送りだされない原因にもなったと思われる.

クレジット制御ではパケットサイズに関わりなく全てのパケットでコンスタントにスループットが出ていた.64flit 以下と 64flit 以上でスループットの飽和とレイテンシが大きく別れた. これは, パケットサイズが

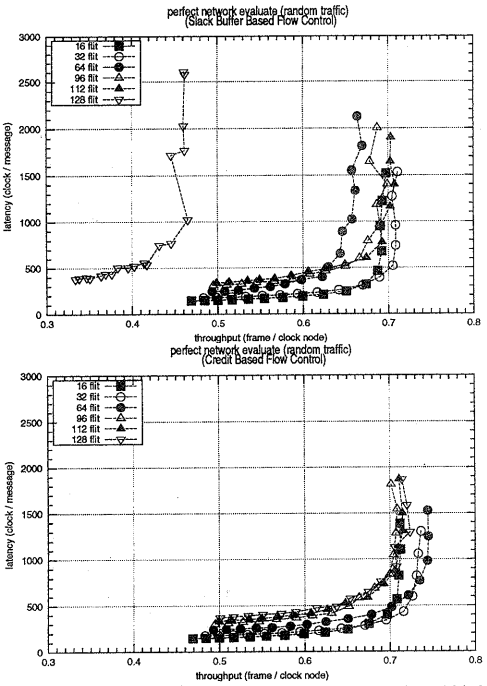


**図 7** 完全結合網ランダム転送における RHiNET-2/SW(上) と RHiNET-3/SW (下) のフロー制御の評価

バッファサイズの半分を超える事により, バッファの先のパケットの出力が中断される事により, 次のパケットのバッファへの読み込みが中断される状況ができ, 転送効率が悪くなったためと思われる.

### 6.4 ホットスポットによる評価

今回作ったホットスポットトラフィック内においては, 特定のリンクに負荷がかかる設計を作ってしまったため, フローコントロールによって転送性能は変わらなかった.

クレジットベースドのフロー制御においては, ランダム転送と違ってパケットサイズが 128 の所でスループットが飽和している. ここより, ホットスポットのスイッチにおいて, クレジットがなくなった事によってパケット転送を中断する事態はあまりなかったものと思われる.

つまりホットスポットではあってもスイッチのアービトレーションが効率よくなっている状況ではフローコントロールの差はあまりないものと思われる.

## 7. 結　　論

RHiNET-3/SW ではそれぞれ仮想チャネルが 256 フレーム入るバッファを保有している.256 フレームであると従来のフロー制御だと最大パケット長を伝送遅延の合計程度に抑えなければいけないが, クレジットフロー制御採用により最大パケット長は仮想チャネルバッファ容量を超える事も可能である.

本研究において,RHiNET-2/SW と RHiNET-3/SW のフローコントロールモデルを提案し,C++ においてネットワークシミュレーションモデル構築し,torus 網, 完全結合網を組んで確率モデルシミュレーションを実行した. その結果, クレジット制御に基く RHiNET-3/SW のシミュレーションでは, パケットの大きさに対する仮想チャネルのバッファ量に依存する事なく, 常に高効率でネットワーク転送が出来る事を確認した.

今回, シミュレータの動作速度とメモリの限界により小規模なシステムの評価に留まったが,RHiNET-3/SW の開発動機に基づけば, より大きなシステムを評価する必要があり, 今後の課題である.

## 8. 謝　　辞

本研究において貴重な意見を頂きました, 新情報処理開発機構の西村信治氏, 多昌廣治氏, 田辺昇氏に感謝します. また本スイッチの構築に協力いただいた, 日立製作所及び日立情報システムの皆様に感謝します。

## 参 考 文 献

1) Boden, N. J. et al.: Myrinet – A gigabit-per-second local-area network, *IEEE Micro*, Vol.15, No. 1, pp. 29–36 (1996).
2) 西宏章, 多昌廣治, 西村信治, 山本淳二, 工藤知宏, 天野英晴: LASN 用 8 Gbps/port 8x8 One-chip スイッチ:RHiNET-2/SW, 並列処理シンポジウム JSPP'2000, Vol. 2000-6, pp. 173–180 (2000).
3) 西宏章, 多昌廣治, 工藤知宏, 天野英晴: 仮想チャネルキャッシュを持つネットワークルータの構成と性能, 並列処理シンポジウム JSPP'99, Vol. 99-6, pp. 71–78 (1999).
4) 天野英晴: 情報系教科書シリーズ第 18 巻 並列コンピュータ, 株式会社昭晃堂, ISBN4-7856-2045-5 (1996).
5) 堀江健志, 石畑俊幸, 池坂宏明: 並列計算機 AP1000 における相互結合網のルーチング方式, 電子情報通信学会論文誌, Vol. J75-D-I, No. 8, pp. 600–606 (1992).
6) : http://www.hnf.org.
7) Katevenis, M., Serpanos, D. and Vatsolaki, P.: ATLAS I: A General-Purpose, Single-Chip ATM Switch with Credit-Based Flow Control, *IEEE Hot Interconnects IV Symposium*.