情報処理学会論文誌　コンピューティングシステム　Vol. 3　No. 3　88–99 (Sep. 2010)

単フリット・単サイクルルータを用いたNoC向け非最短完全適応型ルーティング

西川由理[†1]　鯉渕道紘[†2,†3]
松谷宏紀[†4]　天野英晴[†1]

チップマルチプロセッサ（CMP）のチップ内ネットワークではコヒーレンス転送などで生じる 1 flit におさまる小さいメッセージを扱う場合が多い．そこで，本稿では，1-flit パケット転送に適した非最短完全適応型ルーティング機構である Semi-deflection ルーティングを提案する．Semi-deflection ルーティングは一部のルータ間のパケット転送をノンブロッキングで行うことで仮想チャネルを用いずにパケット間のデッドロックフリーを実現する．評価結果より，Semi-deflection ルーティングをサポートするルータは，典型的な適応型ルーティングである North-last ターンモデルをサポートするルータとほぼ同等のハードウェア量で実現でき，8×8 2 次元 Mesh トポロジにおけるスループットは最大 3.32 倍の向上を達成した．

A Non-minimal Fully Adaptive Routing Using Single-flit Single-cycle Routers for NoCs

Yuri Nishikawa,[†1] Michihiro Koibuchi,[†2,†3]
Hiroki Matsutani[†4] and Hideharu Amano[†1]

A Network-on-Chip of chip multiprocessors (CMP) usually forwards a message that can be packed into a single flit, generated by coherence transfer. In this paper, we propose a non-minimal fully adaptive routing mechanism for 1-flit packet transfers called “Semi-deflection”. Semi-deflection routing guarantees deadlock-free packet transfer without use of virtual channels by allowing non-blocking transfer between specific pairs of routers. Evaluation results show that a router that supports Semi-deflection routing is almost equal to the hardware amount compared with that of north-last turn model, which is a typical adaptive routing. As the result of throughput evaluation, the Semi-deflection routing provided 3.32 times higher throughput.

1. はじめに

半導体技術の進歩によって単一チップ上にプロセッサやメモリ，I/O など複数の設計モジュールをタイル状に実装できるようになり，このようなタイルどうしの結合にパケット構造を用いたチップ内ネットワーク（Network-on-Chip: NoC）が用いられるようになった[1)]．

多くの既存のチップ内ネットワークはプロセッサ内演算データ転送やチップマルチプロセッサ（CMP）におけるコヒーレンス転送を主に行うため，パケット長が短いという特徴を持つ．たとえば，TRIPS では，On-Chip Network（OCN: 138 ビット幅）におけるトラフィックはメモリ転送が多く，キャッシュの line size である 64-Byte 転送は 5-flit パケットに，データ転送要求などの細かい通信は 1-flit パケットとして転送する．さらに，Operand Network（OPN: 142 ビット幅）には演算データが流れるが，90%のパケットは 99-bit 以下であり，1-flit パケットに収まる[2),3)]．

NoC は end-to-end でパケットの消失が生じない lossless ネットワークであるため，ルーティングはパケット間の循環依存により生じるデッドロックの除去が 1 つの課題となる．ルーティングにおけるデッドロックの除去は 1-flit パケットを用いる場合，各フリットが独立してブロックされているフリットを迂回できるため複数フリットで構成されるパケットの場合と比べて容易である．

それにもかかわらず，既存の NoC では並列計算機で採用されてきた可変長パケット構造を対象としたデッドロックフリールーティングアルゴリズムが現在も使われ続けている．たとえば，NoC における典型的なトポロジである 2 次元 Mesh では，パケットを x 方向に必要ホップ数移動した後，y 方向に移動する最短型固定ルーティングである次元順ルーティングが頻繁に用いられている．また，複数経路の中から使用する経路を動的に選択することができる非最短適応型ルーティングについても，Turn モデル[4)] も利用されている[5)]．

一般的に，1 つの目的地に対し，選択可能な経路数が多い自由度の高いルーティングアル

†1 慶應義塾大学大学院理工学研究科
Graduate School of Science and Technology, Keio University
†2 国立情報学研究所
National Institute of Informatics
†3 総合研究大学院大学
The Graduate University for Advanced Studies
†4 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

ゴリズムは性能が高い傾向にある．最も選択可能な経路数が多いルーティングは，非最短完全適応型ルーティングアルゴリズムである．非最短完全適応型ルーティングとは，目的地へのあらゆる経路（物理チャネル，ルータ群）を動的に選択可能なルーティングアルゴリズムであり，経路が静的に 1 つに定まる固定ルーティングに比べてパケット間のデッドロックの除去が複雑となる．

そのため，非最短完全適応型ルーティングアルゴリズムは，通常，仮想チャネルを用いることでデッドロックフリーを実現する．たとえば，最短完全適応型ルーティングである Duato's protocol では，Mesh トポロジにおいて 2 本，Torus トポロジにおいて 3 本の仮想チャネルの追加が必要である．しかし，チップ内ネットワークで採用されているルータは軽量であるため，チャネルバッファの大きさがルータのハードウェア量に対して支配的となる．仮想チャネルは，1 つのポートあたり仮想チャネル数分の独立したチャネルバッファが必要となるため，現状では自由度の高い適応型ルーティングの実装コストはきわめて大きい．

そこで，本稿では，1-flit パケット構造を利用することで，仮想チャネルを使わずにデッドロックフリーを実現する非最短完全適応型ルーティング機構である Semi-deflection ルーティングを提案する．Semi-deflection ルーティングではデッドロックフリーパケット転送を実現するために，一部のルータ間のパケット転送を，他のパケットにブロックされることなく行う．このパケット転送を実現するには，一部のルータの入力ポートに到着したパケットを，必ず 1 つの出力ポートにノンブロッキングで転送する必要がある．そのため目的地に最短で近づく出力ポートが選択できるとは限らない．これは deflection routing（ホットポテト）[6),7)]，カオスルータ[8)] と同様のアプローチであり，これらは既存の研究によりライブロックを生じることなく目的地に到達できることが示されている．

また，転送データが 1-flit に収まらない場合には，転送データを複数の 1-flit パケットに分割し，独立に転送することで，Semi-deflection ルーティングを利用することが可能である．

NoC のルータは，パケット転送を複数のステージに分割したパイプライン転送を行うが，最近の NoC では，依存関係のない複数ステージの並列実行，投機実行（speculation），あるいは動作周波数の低いチップにおける複数ステージの cascading による統合（pipeline integration）により，転送遅延の小さい単サイクルあるいは 2-cycle ルータが開発されている．そのため，ルータにおけるパケットのルーティング処理，出力ポートの設定などのパケットごとに 1 度だけ必要となる処理遅延の影響が小さくなってきており，1-flit パケット構造による可変長パケット構造と比べた場合のルータ内処理遅延の増大はほとんど生じない．

なお，適応型ルーティングの問題として，並列アプリケーションの実装によっては必要となる In-order パケット配送が保証されていない点があげられる．この問題に対して並列計算機の結合網では，ネットワークインタフェースにおいて少量のバッファでパケットソートを実現する仕組みなどが検討されてきた[9),10)]．チップ内ネットワークにおいても，適応型ルーティングのパケットの配送順に関する対策が同様に必要となる場合が生じるが，本稿では適応型ルーティングによるチップ内ネットワークの性能向上を論点とする．

以下，2 章において Semi-deflection ルーティングについて提案する．3 章において評価結果を示し，4 章で関連研究を述べる．最後に 5 章でまとめと今後の課題を述べる．

2. 非最短完全適応型ルーティング機構

本章では，非最短完全適応型ルーティング機構である Semi-deflection ルーティングを提案する．これは Turn モデルを基に，1-flit 固定長パケット，ルータの出力方向選択機構とアービタの工夫によりデッドロックフリーなパケット転送を実現する．

2.1 Turn モデル

Turn モデルは，チャネルの循環構造を解析し，トポロジ内におけるすべての循環依存を除去することで，デッドロックフリーを実現する適応型ルーティングである．代表的なものとして North-last turn モデル，West-first turn モデルがあげられる[4)]．

Turn モデルの例として，2 次元メッシュトポロジにおける North-last turn モデルルーティングを**図 1** に示す．North-last turn モデルルーティングでは，ルータにおける入出力チャネルの方向が東西南北と 4 種類あり，パケットはルータにおいて，計 8 種類のターンの組合せにより循環構造が生じることに着目する．そして，この循環構造を除去するために，そのうちの 2 種類のターン（南から西，南から東）を禁止する．よって，North-last turn モデルルーティングにおいてパケットは 0 ホップ以上の北方向への移動を最後に行うことで目的地に到着することになる．

Turn モデルは目的地までの非最短経路を許すが，循環除去のために禁止したルータにおけるパケットのターン（以後，禁止ターンと呼ぶ）が生じるため，一部の経路を選択することができない．

2.2 1-flit パケット構造を用いたデッドロック除去

本節では，デッドロックフリーを保証しつつ，Turn モデルのうち North-last turn モデル，West-first turn モデルにおける禁止ターンを許可する拡張を行う．

ここで，この拡張の鍵となる用語「ノンブロッキング」を以下のとおり定義する．

定義 1　ルータにおけるクロスバ設定，アービタなどのすべての処理において，ある入出

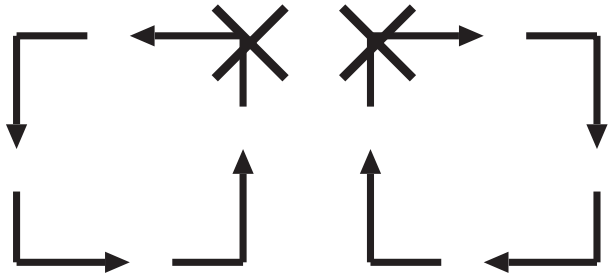


図 1　North-last turn モデルの禁止ターン
Fig. 1　Forbidden turn of North-last turn model.

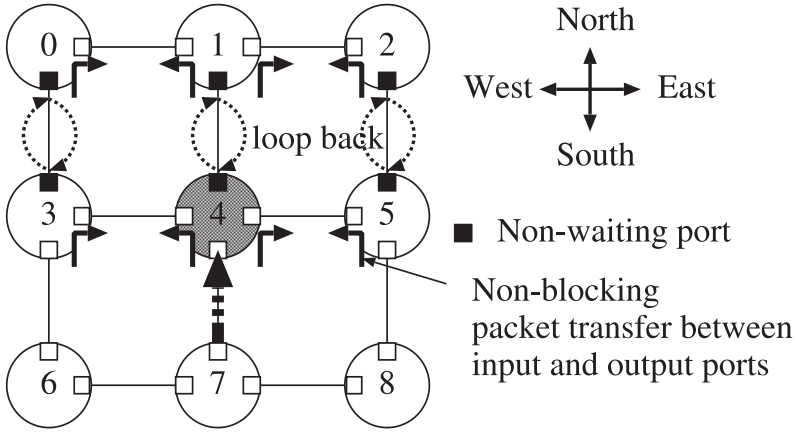


図 2　非最短完全適応型ルーティングの例
Fig. 2　Example of a non-minimal fully-adaptive routing.

力ポート間のパケット転送が，他の入出力ポート間のパケット転送により妨害されずに実行された場合，そのパケット転送をノンブロッキングで行われたと呼ぶ．

禁止ターンを許可した場合でも，禁止ターンに沿って転送される 1-flit パケットがつねにノンブロッキングで，次のルータへ転送することを保証できた場合，デッドロックは生じない．つまり，Turn モデルが禁止するターンをとろうとするパケット転送は，必ずノンブロッキングで行うことで，完全デッドロックフリールーティングを実現することができる．**図 2** に，North-last turn モデルを基にした，ノンブロッキングでパケット転送する必要がある入出力ポート対（図中：Non-blocking packet transfer between input and output ports）を示す．

このノンブロッキングは，1-flit パケットを用いる場合，各フリットが独立して他のフリットにより使用されているバッファを迂回できるため，実現可能である．具体的には，各パケットは，各ルータにおいて入力チャネルバッファが一杯，かつ，禁止ターンとなる出力チャネルが使用中の場合，パケットはその出力ポートのみを待つことはできないという制約を課す．つまり，禁止ターンとなる出力チャネルが使用中で，その他の出力チャネルが空いていた場合，その他の出力ポートへ転送しなければならない．これは，パケットがつねに動き続けることが要求される deflection ルーティング[6] と同様の考え方である．

たとえば，図 2 の灰色のルータに南方向からパケットが到着したと仮定する．西，南，および東方向へのパケット転送はすべて禁止ターンであるため，東西南方向への出力ポートがすべて使用されている場合，目的地がどこであれ北方向へ転送されることになる．

ただし，北方向の出力ポートが使用中である限り，南方向からのパケットは，全方向の出力ポートを含めて待つことができる．これは禁止ターン以外の方向へのターンが循環構造を防ぐ逃げ道となるためである．この逃げ道は，Duato's protocol における転送制限のない適応型パスからデッドロックを除去するデッドロックフリーな経路と同様の考え方であるが，Semi-deflection ルーティングの場合は，仮想チャネルを必要としない．

図 2 において，ルータ 1，2，3 の黒色の入力ポートでは，すべての出力ポートに対するパケット転送が禁止ターンをとることになる．そのため，この入力ポートではすべてのパケットはつねにノンブロッキングで転送されることが要求される．この要求実現のために，以下，Semi-deflection ルーティングに関する定義，定理，証明を行うが，これらの議論は，各ルータは入力に最低 1 flit 分のバッファを持ち，かつ，パケットの出力先を決定し，パケットを転送する操作を同期して行うことを前提とする．

定義 2　ルータの入力ポートのうち，以下の条件のいずれかを満たすものを "non-waiting ポート" と呼ぶ．(a) 出力ポートのすべてが，採用した Turn モデルの禁止ターンとなっている．(b) 隣接ルータの (a) を満たす入力ポートと対になっている．

定理 1　2 次元メッシュ ($N \times N$) の Turn モデル：North-last，West-first ではそれぞれのルータに対して non-waiting ポートはたかだか 1 つである．

証明 1　North-last turn モデルルーティングでは，ターンに制限が生じるのは，南方向から入力されたパケットのみである．したがって，定義 2 (a) の条件で non-waiting ポートとなる可能性があるのは，南方向からの入力ポート 1 つである．定義 2 (b) を満たす入力

ポートは，(a) に対して 1 つのみであるため，自動的に満足される．West-first turn モデルルーティングでは，ターンに制限が生じるのは，$x = N - 1$ のルータのみである．ここで $y \neq 0$，$y \neq N - 1$ の場合は y 方向のどちらかに禁止ターンではない出力が存在する．定義 2 (a) の条件が成立するのは $x = N - 1$，$y = 0$ および $x = N - 1$，$y = N - 1$ である．前者の場合，条件が成立するのは，北方向からの入力ポート 1 つであり，後者の場合は南方向からの入力ポート 1 つである．定義 2 (b) については，(a) に対して 1 つのみであるため，自動的に満足される． ∎

定義 3　ある入力ポートから，その入力ポートが接続されている送り元のルータに対して，パケットを転送することをループバック転送と呼ぶ．

定理 2　あるルータにおいて non-waiting ポートがただ 1 つである場合，ループバック転送を許し，non-waiting ポートに対してアービトレーション上の優先権を与えれた場合，non-waiting ポートには同一パケットが一定時間以上とどまらない．

証明 2　転送相手先の入力ポートのバッファに空きがある場合，non-waiting ポート内のパケットは空いたポートのいずれかに転送される．定義 2 により，ループバック転送の相手先ルータには必ず non-wainting ポートが存在する．転送相手先の入力ポートに空きがない場合，相手先ルータの non-waiting ポートに対してパケットをループバック転送する．相手先ルータの non-waiting ポート内に別のパケットが存在する場合でも，non-waiting ポートに優先度を与えた場合，互いのバッファに対してパケットが転送され，パケットの交換が起きる．以上の転送および交換は，ルータがパケットを処理するのに要する一定時間でつねに起きるため，non-waiting ポートに同一パケットはとどまることはない． ∎

図 2 における non-waiting ポートは，黒色のポート群となる．ループバック転送によるライブロック除去については次節で述べる．

2.3　ルータ機構

デッドロックとライブロックを防ぐためには，ルータの出力選択機構（output selection function: OSF）とアービタの設定を適切に行うことが必要となる．

2.3.1　出力選択機構

適応型ルーティングは，適応型アルゴリズムと出力選択機構の 2 段階に分けられる．適応型アルゴリズムはデッドロックフリーな出力チャネルの候補の集合を求め，OSF によってその候補の中から実際のパケットの出力チャネルが決定される．Semi-deflection ルーティングのように（非最短型）完全適応型ルーティングの場合，すべての出力チャネルがパケットの転送先の候補となるため，出力選択機構がパケットのとる経路を決定することになる．

定義 4　Semi-deflection OSF は，OSF の優先順序を高い方より以下の順に設定する．

(1)　目的地に近づく出力ポート
(2)　ループバック転送を除く目的地から遠のく出力ポート
(3)　ループバック転送を生じる出力ポート（non-waiting ポートのみ）

Semi-deflection OSF を用いることで，non-waiting ポートにおいてループバックとなる出力ポートの優先度が最も低くなりライブロックを抑えることができる．なお，パケットをより分散させるため，同一優先度の複数のポートはランダムに選択させるなどが可能である．

2.3.2　アービトレーション

ルータ内のアービタは，入力ポートから要求のあった複数のフリットに対して，各出力ポートにおける優先順位制御を行う．この制御により，ルータにパケットが入力された時系列順に，前述の OSF によって与えられた優先度に基づいて，以下の順で出力ポートが割り当てられることになる．

定義 5　Semi-deflection アービトレーションは，以下の順にパケットを選択する．

(1)　non-waiting ポートのバッファ中のパケット．
(2)　禁止ターン方向の出力ポートに OSF で最高優先度が与えられているパケット．
(3)　その他のパケット．

(1) により，non-waiting ポート間をパケットが往復しつづけるライブロックを防ぎつつ，（OSF で他の優先度が高い出力ポートが使用中の場合）自身のポートへのループバック転送をノンブロッキングで行うことができる．(2) について，禁止ターン方向の出力ポートが別のパケットによってロックされている場合は，次の優先度に従って別の空いているポートが割り当てられる．これにより禁止ターン方向のノンブロッキング転送を実現することができる．

2.4　Semi-deflection ルーティング機構

前節での検討に基づき，Semi-deflection ルーティングを定義する．

定義 6　以下に定めたチップ内ネットワークにおけるデータ転送を Semi-deflection ルーティング機構と呼ぶ．

- パケットサイズを 1 flit とし，完全適応型ルーティングを用いる．
- North-last あるいは West-first turn モデルルーティングに基づいて禁止ターン群とその non-waiting ポートを設定する．
- non-waiting ポートに限り，入力ポートに接続されたルータに対して，ループバック転送を許可する．

- 各ルータは，Semi-deflection OSF，Semi-deflection アービトレーションに基づきパケット転送を行う．

この Semi-deflection ルーティング機構を備えたルータによるルーティングを Semi-deflection ルーティングと呼ぶ．

定理 3　Semi-deflection ルーティングはデッドロックフリーである．

証明 3　(i) 禁止ターン以外のターンに沿った経路でパケット転送を行った場合，循環経路は存在せず，デッドロックは生じない．(ii) 禁止ターンを用いた場合でも，そのときに相手先ルータの入力ポートのバッファのいずれかが空いていれば，パケットはブロックされず，1 flit パケット転送においては，デッドロックは生じない．(iii) すべての相手先ルータの入力ポートのバッファにパケットが存在する場合でも，相手先ルータが禁止ターン以外のターンに沿ったルータであれば，その経路上はデッドロックは生じない．したがって (i) より必ず入力ポートのバッファには空きが生じ，パケットの転送が可能になる．以上により，1 flit パケットの完全適応型ルーティングを用いた場合，デッドロックを生じる可能性があるのは，すべての相手先のルータが禁止ターンの経路上にあり，その入力ポートのバッファにパケットが存在する場合である．定義 2 により上記の条件にあてはまる入力ポートは non-waiting ポートであり，定理 1 および定義 5 により non-waiting ポートには一定時間以上パケットは滞在しないことが保証される．よって，Semi-deflection ルーティングはデッドロックフリーである． ∎

定理 4　Semi-deflection ルーティングはライブロックフリーである．

証明 4　定義 4 および定義 5 により，Semi-deflection ルーティングにおいては目的地へ向かう出力ポートを優先的に選択する．また，Semi-deflection ルーティングは，1 flit パケットに基づく完全適応型ルーティングである．以上により，カオスルータのライブロックフリーの条件[8] が満足される．よって，Semi-deflection ルーティングはライブロックフリーである． ∎

3. 評　　価

非最短完全適応型 Semi-deflection ルーティングについて，まず Verilog で実装を行ったルータのハードウェア面積の結果を示す．次にフリットレベルシミュレーションにより，1-flit 構造を適用したときのスループットとレイテンシを，従来手法とあわせて評価する．最後に，メッセージがすべて 1-flit に収まるとは限らないため，メッセージが大きい場合に，1 つの可変長のパケット構造（wormhole 方式）を利用して転送した場合と，複数の 1-flit パケットに分割する場合の比較を行う．

表 1　8 × 8 Mesh に対応するルータ 1 個の面積
Table 1　Area of a router for 8 × 8 Mesh network.

ルータ	ルーティング手法	ゲート数
1-cycle 1-flit router	ecube	35,026
1-cycle 1-flit router	North-last	34,986
1-cycle 1-flit router	Semi-deflection	35,600

一般的に，適応型ルーティングは選択可能な経路数が多いため，ネットワークの性能はプロセッサコアからの注入制限（スロットリング）に大きく影響を受ける．よって，この性能チューニングについてもその効果を示す．

3.1　ハードウェア量

本評価において，軽量な wormhole 型オンチップ 1-cycle ルータ[11] を基に拡張することで，本ルーティング方式をサポートするルータと既存のルーティングアルゴリズムをサポートするルータのハードウェアを実装した．パケット構造は 1-flit だが，フロー制御上，アービトレーションの結果出力チャネルに転送できないパケットが入力ポートにとどまることになるため，入力チャネルは 2-flit，出力チャネルは 1-flit 分のバッファを持つ．また前章で述べたように，仮想チャネルは用いていない．Synopsis 社の Design Compiler 2007.12-SP3 を使い，45nm CMOS プロセスライブラリを用いて論理合成，SoC Encounter を用いて配置配線を行った．

パケットフォーマットは，1 章で述べた TRIPS の OCN ネットワークのビット幅を参考に，140 ビットに設定した．この内訳は，制御情報として目的地（6 ビット），パケット長（3 ビット），順序番号情報（3 ビット）とし，他をすべてペイロード（128 ビット）とした．

次元順（ecube）ルーティング，North-last ルーティング，および Semi-deflection ルーティングに適用される，8 × 8 Mesh トポロジに対応したルータ 1 個の面積を**表 1** に示す．表より，次元順（ecube）ルータに比べ，適応型（North-last）ルータは出力ポートを割り当てる OSF およびアービタがやや複雑になるが，きわめてオーバヘッドが小さく，ほぼ同じハードウェア量となっている．さらに，ecube ルータ，適応型ルータに比べ，Semi-deflection ルータは 1.02%ときわめて小さいハードウェア量の増加で実現できることが分かった．

3.2　スループット

3.2.1　シミュレーション条件

本節では，C++で記述されたフリットレベルネットワークシミュレータ irr_sim[12] を用

いて，16，64，256 コアのチップ内ネットワークのスループットと遅延の評価を行う．ここで遅延とは，出発地のコアがパケットを生成してから目的地のコアが受信するまでのサイクル数であり，ネットワークへの注入時間および滞在時間の合計である．また accepted traffic は各コアが 1 サイクルあたりに受信する平均フリット数とした．また，本評価における最大スループットは，文献 11) と同様に，通信遅延が 1,000 サイクルを超えない accepted traffic の最大値として定める．

トポロジは 4×4，8×8，16×16 の 2 次元 Mesh とした．またトラフィックパターンには，uniform トラフィック，および以下の合成トラフィックパターンを用いた．

- *matrix transpose*
  行列転置の際のトラフィックパターン．アレイサイズが k のとき，ノード (x, y) がノード $(k-y-1, k-x-1)$ にパケットを転送する．ただし対角線上（$x+y=k-1$ が成り立つ）のノードはノード $(k-x-1, k-y-1)$ に転送する．
- *bit reversal traffic*
  送信元ノードを $(a_0, a_1, \ldots, a_{n-1})$ と表記するとき，それぞれが宛先ノード $(a_{n-1}, \ldots, a_1, a_0)$ にパケットを転送する．

また上記の合成トラフィックに加え，本稿では，実アプリケーションとして NAS Parallel Benchmark（NPB）[13] プログラムから得られた通信パターンを用いる．NPB は並列計算機向けのベンチマークであるが，最近では，MPI ライブラリを用いて並列処理を行う CMP が登場している[14]．今回は NPB から次の並列プログラムを用いる．

- Conjugate Gradient（CG）
- Multi-Grid solver（MG）

プログラムのクラスは “W” とし，計算コアの数は 64 とした．

3.2.2　注入制限による性能チューニング

図 3 に示すように，Semi-deflection ルーティングを用いた場合，トラフィック量が一定量を超過すると，いずれのトラフィックパターンでもスループットが急速に下落する．一方，ecube，North-last ルーティングでトラフィック量が増えたときは，緩やかに飽和状態に達するものの，急速に性能が悪化することがない．なお，同様の性能悪化の現象は，完全適応型 wormhole 方式において，ネットワークが高負荷になった場合にも見られる[15]．よって特に Semi-deflection ルーティングにおいては，その弱点であるトラフィック負荷が一定量を超えてネットワーク性能が悪化する前に，パケットの注入を中断する機構が必要となる．

図 3 より，最大スループットが得られるネットワーク全体の負荷はトラフィックパターンによって異なる．また，局所的なトラフィックにおいては，ネットワーク全体の負荷値でローカルノードの制御をすることは適切ではない．このため，各ノードが独自に制限を切り替えられることで良い性能が得られると考えられており，wormhole ルータでは，負荷を動的に制限するにあたって，仮想チャネルの利用率を監視する手法が複数提案されている[15),16]．特に文献 15) では，同時に利用可能な仮想チャネル数の上限を動的に制御するにあたり，ホスト側のキュー長をパラメータとする方法を提案している．一般にネットワーク全体の負荷を各ノードが把握するのが困難であり，ローカルホストのキュー長が負荷の指標として使用されている．この手法の有効性も同文献に示されている．なお Semi-deflection ルーティングは仮想チャネルを持たないため，これらに相当する注入制限手法を決定する．

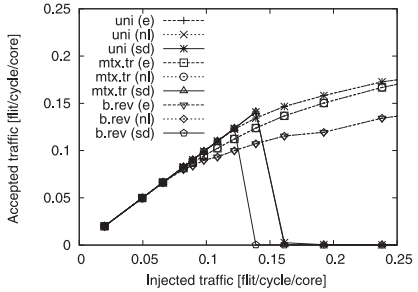


図 3　注入制限を行わない 8×8 Mesh におけるネットワーク負荷と性能の関係．“uni”，“mtx.tr”，“b.rev” はそれぞれ uniform，matrix transpose，bit reversal トラフィックパターンの略．(e)，(nl)，(sd) はそれぞれ ecube，North-last，Semi-deflection ルーティングの略である

Fig. 3　Relationship between network load and performance for 8×8 Mesh network without injection limitation. “uni”, “mtx.tr”, “b.rev” represent uniform, matrix transpose and bit reversal traffic, respectively. (e), (nl), (sd) stand for ecube, North-last, Semi-deflection routing, respectively.

まずルータのローカルホストで，ルータの入力ポートの利用数，およびホスト側のキューに入力されているパケットの数を監視する．そして，ネットワークの負荷に対するこれらの値を求める．前者の平均値を**図 4** に示す．これより，同時に利用できる入力ポート数を 2 から 3 の間で調整するのが適当であると分かる．そこで，この結果を基に，ホストからのパケットの注入制限を，ホストのキュー長と隣接ルータの（上限）使用入力ポート数（2 あるいは 3）により決定する性能チューニングを行う．各トラフィックパターンについて，図 3 でスループットが悪化した直後のネットワーク負荷におけるホストのキュー長のスレッショルド値とレイテンシの関係に関するシミュレーション結果を**図 5** に示す．これより，uniform，

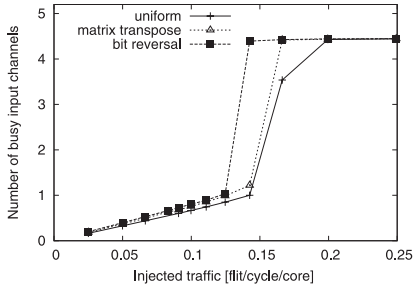


図 4　注入制限を行わない 8 × 8 Mesh におけるネットワーク負荷と利用中の入力ポート数の関係
Fig. 4　Relationship between network load and number of occupied input ports for 8 × 8 Mesh network without injection limitation.

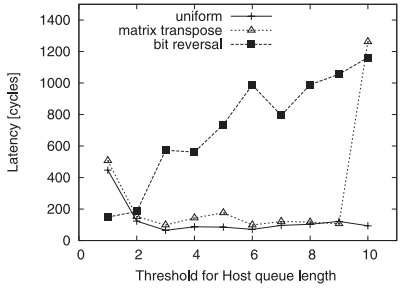


図 5　8 × 8 Mesh におけるホストキューのスレッショルド値
Fig. 5　Threshold value of host queue length for 8 × 8 Mesh network.

matrix-transpose についてはホストキュー長が 3 以上のとき，bit-reversal では 1 以上のときに，入力ポート数の上限を 2 に定めることでレイテンシが短くなることが分かる．今回はこのように，トラフィックごとにレイテンシが最も短くなるホストキュー長をスレッショルド値として定め，評価を行った．スレッショルド値を実行時に動的に制限する機構は今後の課題とする．

なお，この制限機構は，ホスト側でルータの各入力ポートの利用率，および自身のキュー長を監視することで実現可能となる．ここで，表 1 にハードウェア量を示した Semi-deflection ルータは，wormhole ルータを元に拡張を行っているため，各入力ポートがその他のすべての入力ポートの利用状況を示す信号線をルータ外部に出力している．Wormhole ルータをこのように実装した理由は，あるノードの入力バッファの利用状況が，隣接ノードの転送要求となるためである．よって，ホスト側は各入力ポートの出力信号線をルータからの入力として受け取るのみでよく，3.1 節で設計した Semi-deflection ルータのハードウェア量を増やすことなく注入制限機構を実現することができる．

3.2.3　完全適応ルーティングと既存のルーティングアルゴリズム

前項で述べた条件における Semi-deflection ルーティングのスループットを**図 6** に示す．それぞれ，次元順ルーティング（ecube），および Turn モデルのうち North-last を用いた適応型ルーティングアルゴリズムにおいて 1-flit パケットを転送する場合とスループットを比較した．なお，すべてのルーティングアルゴリズムにおいて，入力バッファサイズ，およびパケットサイズは 1-flit とした．なお，ecube，North-last ルーティングは，図 3 よりネットワーク負荷が増加しても性能が悪化しないが，評価の公平性のため，注入制限をした場合としない場合の両方の結果を示す．Semi-deflection ルーティングについては，注入制限を行った結果を示す．

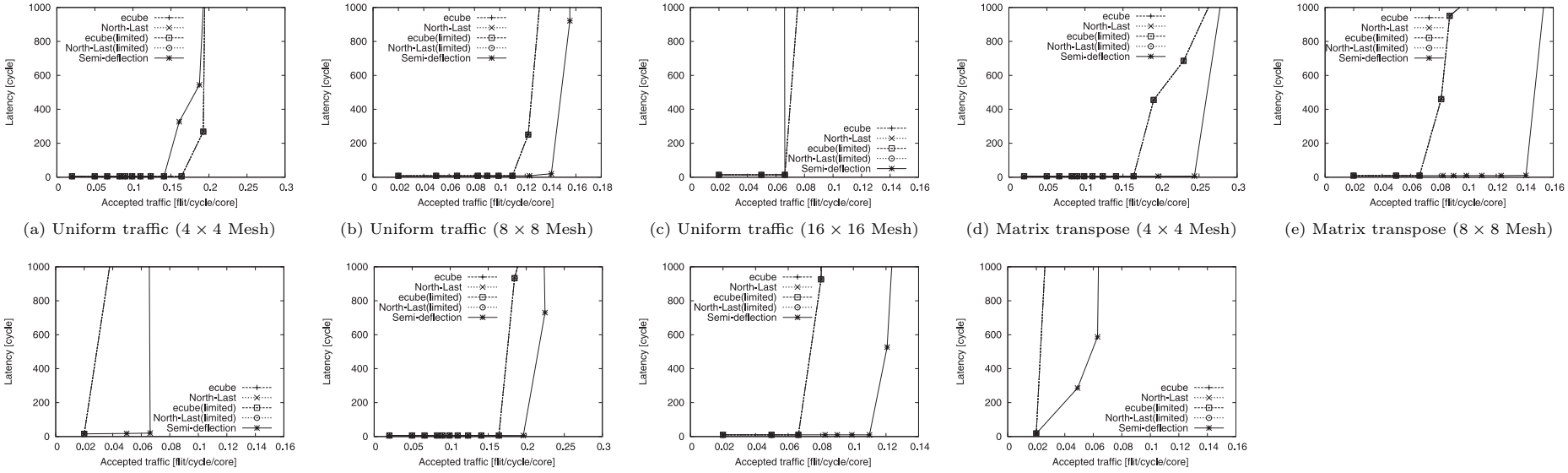
結果より，各アレイサイズおよびトラフィックパターンにおいて，Semi-deflection ルーティングは uniform トラフィックを除き，比較対象とほぼ同等かそれ以上の最大スループットを得ることができた．またその最大スループットは，uniform トラフィックおよび 16 × 16 の bit reversal トラフィックを除き，きわめて低レイテンシで実現されている．

uniform トラフィックは，一般に適応型ルーティングアルゴリズムの効果が得にくいパターンとして知られており，図 6 (a) から図 6 (c) では Semi-deflection ルーティングの性能が劣る，または大きな有効性を示すことができていない．一方，図 6 (d) から図 6 (i) に示すように，局所的なトラフィックが生じる matrix transpose や bit reversal において，Semi-deflection ルーティングがいずれも従来手法より高い最大スループットを示している．さらにアレイサイズが大きくなるほど本手法が有利となり，図 6 (i) に示すように，16 × 16 Mesh トポロジにおいて matrix transpose traffic を適用したとき，最大 3.32 倍のスループットが得られた．

また，図 6 (a) から図 6 (i) のすべての結果より，まず ecube および North-last ルーティングに対する注入制限の効果が低いことが分かる．これは図 3 で示したように，ネットワーク負荷が上がっても注入制限をすることなく性能が維持できているためである．さらに，Semi-deflection ルーティングについても性能が維持できており，3.2.2 項で示した注入制限手法が有効であったといえる．

また，各トラフィックパターンにおける平均ホップ数を**表 2** に示す．ここで

- 低負荷：トラフィック量がきわめて少ない（50 サイクルごとに，すべてのホストがパ



(a) Uniform traffic (4 × 4 Mesh)　(b) Uniform traffic (8 × 8 Mesh)　(c) Uniform traffic (16 × 16 Mesh)　(d) Matrix transpose (4 × 4 Mesh)　(e) Matrix transpose (8 × 8 Mesh)

(f) Matrix transpose (16 × 16 Mesh)　(g) Bit reversal (4 × 4 Mesh)　(h) Bit reversal (8 × 8 Mesh)　(i) Bit reversal (16 × 16 Mesh)

図 6　合成トラフィックを用いて測定した，既存の Turn モデルによる適応型ルーティングと比較した Semi-deflection ルーティングのスループット
Fig. 6　Synthetic traffic throughput of Semi-deflection routing compared with existing Turn-model-based adaptive routing.

表 2　8 × 8 Mesh における Semi-deflection ルーティングの平均ホップ数
Table 2　Average hop count of Semi-deflection routing for 8 × 8 Mesh network.

	低負荷	中負荷	超高負荷	最短経路のみ
Uniform	5.57	5.91	7.93	5.29
Matrix transpose	6.41	7.47	10.02	5.06
Bit reversal	6.97	7.29	11.31	5.78

ケットをネットワーク中に投入した）場合

- 中負荷：低負荷状態の平均ホップ数を四捨五入してそのサイクルごとにすべてのホストがパケットを投入した場合
- 超高負荷：1 サイクルごとにすべてのホストがパケットを投入した場合

とした．

トラフィック負荷が高いほど，パケットのブロックが生じて最短経路をとることができない確率が上がり，平均ホップ数が増加していることが分かる．また，逆に考えると，注入制限によるネットワーク内のトラフィック負荷の低減がホップ数の増加を抑え，Semi-deflection ルーティングの性能向上につながっているといえる．

次に，NAS Parallel Benchmark から CG，MG の結果を**図 7** に示す．ecube，North-last，Semi-deflection ルーティングにおいて，すべて 1-flit 構造，および 8 × 8 Mesh トポロジを想定している．また，横軸は通信トラフィックの受信レートを表す．横軸の数値が大きくなるに従って，ネットワークの転送能力がコアの計算能力に対して大きくなり，通信が頻繁に発生する場合に対応する．

図 7 (a) より，CG においては，ecube と North-last ルーティングの 1.2 倍の最大スループットが得られた．CG のトラフィックパターンを解析した結果，8 × 8 Mesh トポロジの対角線に対して線対称であるノード間にトラフィックが集中しており，合成トラフィックの matrix transpose ときわめてパターンが類似している．このため，良い性能が得られたと考

えられる．一方図 7 (b) より，MG は従来手法とほぼ同等の性能となった．このベンチマークでは，8×8 Mesh トポロジにおいて同じ x 軸上，y 軸上にある宛先ノードとの通信が頻発している．したがって ecube ルーティングに適したパターンとなっており，Semi-deflection ルーティングで高い性能が得られなかったと考えられる．

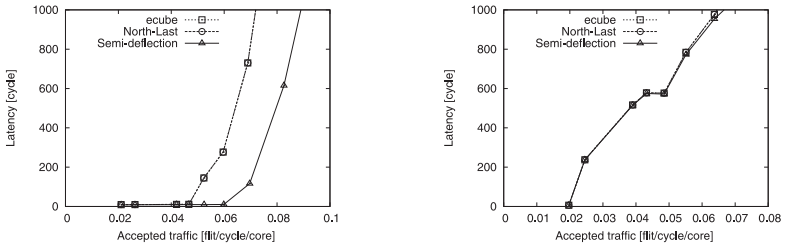
3.2.4 Semi-deflection ルーティングと可変長パケットの wormhole 方式における性能差

Semi-deflection ルーティングで，1-flit 分を超えるメッセージ長を転送した際のスループット性能について検討する．そのため，メッセージ長が 4-flit と 8-flit の場合において，以下の比較を行う．

- 1-flit パケットの場合，メッセージを複数パケットに分割し，1-flit 分の入力バッファを持つルータにて転送
- 可変長パケットの場合，メッセージを 1 つのパケット（複数フリットで構成）にして，wormhole ルータで転送

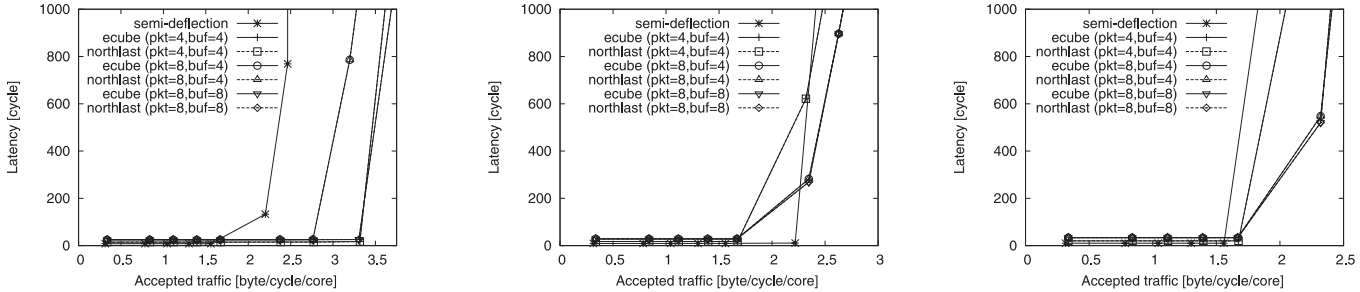
ここで比較対象の wormhole ルータは，メッセージ長が 4-flit の場合は 4-flit 分，8-flit 長の場合は 4-flit および 8-flit 分の入力バッファを持つものを想定して評価した．いずれのルータも仮想チャネルは持たない．前述のフリットレベルシミュレータ上で，uniform トラフィックパターンを利用し，8×8 Mesh トポロジにおける性能を測定した．その結果を**図 8** に示す．なお「最大スループット」の定義は前節と同様だが，データのペイロード部分の転送効率を示すため，横軸の単位を [Byte/cycle/nodes] とした．具体的には 1-flit パケットでは，制御情報として，目的地（6 ビット）パケット長（3 ビット），順序番号情報（3 ビット）と仮定し，他をすべてペイロードとして算出した．一方，wormhole 方式におけるパケットでは，各フリットにフリット識別情報（2 ビット：ヘッダ，テイル，ボディの区別），加えてヘッダフリットが制御情報として目的地（6 ビット），パケット長（3 ビット）を持つと仮定し，他をすべてペイロードとして算出した．

結果より，uniform トラフィックにおける Semi-deflection ルーティングの最大スループットは，8-flit 長の wormhole ルータ（入力バッファ量 8-flit）における次元順ルーティングお



(a) NAS Parallel Benchmark CG (Class W)　(b) NAS Parallel Benchmark MG (Class W)

図 7　NAS Parallel Benchmarks のトラフィックパターンを用いて測定した，既存の Turn モデルによる適応型ルーティングと比較した Semi-deflection ルーティングのスループット
Fig. 7　Throughput of Semi-deflection routing compared with existing Turn-model-based adaptive routing using traffic patterns of NAS Parallel Benchmarks.



(a) Uniform traffic (versus 4-flit and 8-flit WH)　(b) Matrix transpose (versus 4-flit and 8-flit WH)　(c) Bit reversal (versus 4-flit and 8-flit WH)

図 8　可変長パケットを Semi-deflection ルーティングと wormhole ルーティングで転送したときのスループットの比較
Fig. 8　Throughput comparison for transferring variable-length packets with Semi-deflection routing and adaptive wormhole routing.

よび North-last ルーティングに対し，75%程度にとどまる．一方 matrix transpose では，8-flit 長の wormhole ルータを用いた North-last ルーティングに対し，84%，bit reversal トラフィックでは，93%が最大スループットだが，低レイテンシでそれを提供することができることが分かった．

4. 関 連 研 究

本章では，チップ内ネットワークにおける既存のルーティングアルゴリズムについて説明する．

チップ内ネットワークでは並列計算機で採用されてきたデッドロックフリールーティングアルゴリズムがそのまま使われてきた．チップ内ネットワークにおける典型的なトポロジである 2 次元 Mesh では，パケットを x 方向に必要ホップ数移動した後，y 方向に移動する最短型固定ルーティングである次元順ルーティングが頻繁に用いられている．

固定ルーティングは，目的地までの経路が静的に 1 つに限定される．単位時間あたりのスループットを向上させるために，目的地までの経路を動的に切り替えることが可能な適応型ルーティングも様々な提案がなされている．2 章で述べた Turn モデルは，チャネルの循環構造を解析し，トポロジ内におけるすべての循環依存を除去することでデッドロックフリーを実現する．また，完全適応型ルーティングである Duato's protocol は，Mesh や Torus トポロジにおいて，さらに仮想チャネルを 1 本追加することにより，任意の最短経路を選択することができる．

本稿で Semi-deflection ルーティングアルゴリズムは，これらの適応型ルーティングと比べて，

- 仮想チャネルが不要であるため軽量なルータが実現可能である，
- 経路選択の自由度が最も大きい非最短完全適応型ルーティングである，

という点が特徴である．

また，デッドロックフリールーティングだけでなく，パケットのデッドロックが発生した場合にその中の 1 つのパケットを除去するなどの対応を行うデッドロックリカバリー方式も提案されている．デッドロックが発生しない退避経路を準備する progressive 方式と，一部のパケットを廃棄再送する regressive 方式の 2 つがあるが，前者は退避経路のための仮想チャネルが必要になり，後者は廃棄再送のための制御機構が必要となる．なお，Semi-deflection ルーティングアルゴリズムは，デッドロックフリー方式であるためパケットの廃棄は生じない．

さらに積極的にパケットのデッドロック対処をする方法も提案されている．Deflection ルーティングは，各ルータにおいて，毎サイクル，到着するフリットと同数あるいはそれ以上のフリットが出力ポートから転送することでデッドロックフリーを保証する．各パケットは目的地へ向かう出力ポートが選択できるとは限らないため，ホップ数が増加するが，ルータ内におけるパケットの衝突が生じないためデッドロックは生じない．また，カオスルータの研究により，このアプローチはライブロックフリーであることが証明されている．Deflection ルーティングの欠点としては，wormhole 方式には適用できない点，ルータに 1 つのパケットを格納する専用のバッファが追加で必要となる点があげられる[17]．本稿で提案したルーティングアルゴリズムは，Deflection ルーティングと以下の点で異なる．

- Semi-deflection ルーティングにおいては，一部のルータ，かつその一部のポートにおいてのみ，ノンブロッキング転送を行う．その他のルータは，目的地へ向かう出力ポートが使用中の場合，空くまで，入力ポートにある多くのパケットは待つことができる．そのため，パケットのホップ数を一定に保つことができる．
- deflection において必要な専用バッファは追加不要である．

5. まとめと今後の課題

CMP のチップ内ネットワークではコヒーレンス転送などで生じる 1 フリットにおさまる小さいメッセージを扱う場合が多い．

そこで，本稿では，1 フリットパケット転送に適した非最短完全適応型ルーティング機構である Semi-deflection ルーティングを提案した．Semi-deflection ルーティングは一部のルータ間のパケット転送をノンブロッキングで行うことで仮想チャネルを用いずにパケット間のデッドロックフリーを実現する．

評価結果より，Semi-deflection ルーティングをサポートするルータは典型的な適応型ルーティングである North-last ターンモデルをサポートするルータとほぼ同等のハードウェア量で実現でき，8×8 2 次元 Mesh トポロジにおけるスループットは最大 3.32 倍の向上を達成した．また，単サイクルルータで構成されたチップ内ネットワークにおいて，長いメッセージを 1 フリット長の複数パケットに分割して独立して転送する場合のスループット低下はほとんどないことが分かった．

今後は，2 次元 Torus など他のトポロジへの適用や，non-waiting ポートの影響を最小限におさえるルーティングポリシの設計とその性能比較などを行う予定である．

謝辞　本研究の一部は科学研究費（22700061），JST CREST の助成を受けたものである．

参 考 文 献

1) Dally, W.J. and Towles, B.: *Principles and Practices of Interconnection Networks*, Morgan Kaufmann (2003).
2) Burger, D., Keckler, S.W., McKinley, K., Dahlin, M., John, L., Lin, C., Moore, C., Burrill, J., McDonald, R., Yoder, W. and the TRIPS Team: Scaling to the End of Silicon with EDGE Architectures, *IEEE Computer*, Vol.37, No.7, pp.44–55 (2004).
3) Gratz, P., Kim, C., Sankaralingam, K., Hanson, H., Shivakumar, P., Keckler, S.W. and Burger, D.: On-Chip Interconnection Networks of the TRIPS Chip, *IEEE Micro*, Vol.27, pp.41–50 (2007).
4) Glass, C.J. and Ni, L.M.: The Turn Model for Adaptive Routing, *Proc. International Symposium on Computer Architecture*, pp.278–287 (1992).
5) Ilitzky, D.A., Hoffman, J.D., Chun, A. and Esparza, B.P.: Architecture of the Scalable Communications Core's Network on Chip, *IEEE Micro*, Vol.27, No.5, pp.62–74 (2007).
6) Nilsson, E.: Design and Implementation of a Hot-Potato Switch in Network On Chip, Master's thesis, Laboratory of Electronics and Computer Systems, Royal Institute of Technology (KTH) (2002).
7) Moscibroda, T. and Mutlu, O.: A Case for Bufferless Routing in On-Chip Networks, *Proc. International Symposium on Computer Architecture* (*ISCA'09*) (2009).
8) Konstantinidou, S. and Snyder, L.: The Chaos Router, *IEEE Trans. Comput.*, Vol.43, No.12, pp.1386–1397 (1994).
9) Martinez, J.C., Flich, J., Robles, A., Lopez, P., Duato, J. and Koibuchi, M.: In-Order Packet Delivery in Interconnection Networks using Adaptive Routing, *Proc. IEEE International Parallel and Distributed Processing Symposium*, p.101a (2005).
10) Koibuchi, M., Martinez, J.C., Flich, J., Robles, A., Lopez, P. and Duato, J.: Enforcing In-Order Packet Delivery in System Area Networks with Adaptive Routing, *Journal of Parallel and Distributed Computing*, Vol.65, pp.1223–1236 (2005).
11) 枚田優人，松谷宏紀，鯉渕道紘，天野英晴：パイプラインステージ統合による省電力・可変パイプラインルータに関する研究，情報処理学会論文誌コンピューティングシステム，Vol.2, No.3, pp.71–82 (2009).
12) Matsutani, H., Koibuchi, M., Wang, D. and Amano, H.: Adding Slow-Silent Virtual Channels for Low-Power On-Chip Networks, *Proc. International Symposium on Networks-on-Chip* (*NOCS'08*), pp.23–32 (2008).
13) Bailey, D., Harris, T., Saphir, W., van der Wijngaart, R., Woo, A. and Yarrow, M.: The NAS Parallel Benchmarks 2.0, NAS Technical Reports NAS-95-020 (1995).
14) Howard, J., et al.: A 48-Core IA-32 Message-Passing Processor with DVFS in 45nm CMOS, *Proc. International Solid-State Circuits Conference* (*ISSCC'10*), pp.108–109 (2010).
15) Baydal, E., Lopez, P. and Duato, J.: A Simple and Efficient Mechanism to Prevent Saturation in Wormhole Networks, *Parallel and Distributed Processing Symposium, International*, Vol.0, p.617 (2000).
16) Lopez, P., Martinez, J. and Duato, J.: DRIL: dynamically reduced message injection limitation mechanism for wormhole networks, *Proc. International Conference on Parallel Processing*, pp.535–542 (1998).
17) Duato, J., Yalamanchili, S. and Ni, L.: *Interconnection Networks: an engineering approach*, Morgan Kaufmann (2002).

(平成 22 年 1 月 26 日受付)
(平成 22 年 5 月 19 日採録)

西川　由理（学生会員）

平成 18 年慶應義塾大学理工学部情報工学科卒業．平成 20 年同大学大学院理工学研究科開放環境科学専攻修士課程修了．現在，同大学院理工学研究科開放環境科学専攻博士課程在籍中．平成 20 年度より日本学術振興会特別研究員 DC1．ハイパフォーマンスコンピューティングとインタコネクトに関する研究に従事．

鯉渕　道紘（正会員）

平成 12 年慶應義塾大学理工学部情報工学科卒業．平成 15 年同大学大学院理工学研究科開放環境科学専攻博士課程修了．博士（工学）．平成 14 年度より 16 年度まで日本学術振興会特別研究員．現在，国立情報学研究所准教授，総合研究大学院大学複合科学研究科情報学専攻准教授（兼任）．ハイパフォーマンスコンピューティングとインターコネクトに関する研究に従事．IEEE Computer Society Japan Chapter Young Author Award 2007，平成 19 年度情報処理学会論文賞受賞．IEEE，電子情報通信学会各会員．

松谷　宏紀（正会員）

平成 16 年慶應義塾大学環境情報学部卒業．平成 20 年同大学大学院理工学研究科開放環境科学専攻博士課程修了．博士（工学）．現在，東京大学大学院情報理工学系研究科特別研究員．平成 21 年度より日本学術振興会特別研究員 SPD．計算機アーキテクチャ，オンチップネットワークの研究に従事．

天野　英晴（正会員）

昭和 56 年慶應義塾大学工学部電気工学科卒業．昭和 61 年同大学大学院工学研究科電気工学専攻博士課程修了．工学博士．現在，慶應義塾大学理工学部情報工学科教授．計算機アーキテクチャの研究に従事．