情報処理学会論文誌　コンピューティングシステム　Vol. 1　No. 1　28–39 (June 2008)

2-D トーラスネットワークにおける動的通信予測による低遅延化

吉永　努[†1]　村上弘和[†1]　鯉渕道紘[†2]

本論文では，予測スイッチングを用いた並列計算機ネットワークの低遅延通信技術について述べる．特に，予測ミスによって発生するパケットへの対処方法を示すとともに，予測ミスパケットの発生を削減する手法を提案する．また，シミュレーション結果を示し，通信予測のヒット率と通信遅延の削減効果について考察する．2-D トーラスネットワークで次元順ルーティングを行う場合，ネットワーク負荷が低い場合には予測ヒット率に応じたルーティング計算時間の削減効果が得られた．また，予測器の精度を上げることにより，ネットワークが飽和する直前までの比較的高い通信負荷を与えても，パケット平均遅延の削減が可能であることを示す．

Latency Reduction Utilizing Dynamic Communication Prediction in 2-D Tori

Tsutomu Yoshinaga,[†1] Hirokazu Murakami[†1] and Michihiro Koibuchi[†2]

This paper describes a predictive switching technique for low-latency communication in parallel computer networks. We show a method to treat prediction misses as well as reducing the occurrence of mis-prediction packets. We also discuss relation between prediction hit rates and latency reduction. Our experiments for 2-D tori with dimension-order routing show that we can reduce latency for routing computation depending on the prediction hit rates. Not only zero-load latency but also average packet latency up to just before the network saturation load can be reduced when prediction accuracy is high.

†1 電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, University of Electro-Communications
†2 国立情報学研究所
National Institute of Informatics

1. はじめに

並列計算機ネットワークの重要な性能指標として，通信遅延時間とバンド幅があげられる．これまで低遅延・高バンド幅通信を実現するために，ルータ動作周波数の高速化，ネットワークトポロジやルーティングアルゴリズム上の改善，スイッチング方式やフロー制御に関する工夫等様々な技術が開発されてきた．特に，大規模なネットワークにおける通信距離の大きなメッセージの遅延を減らす目的で，ルータあたりのホップ遅延を短縮するための手法が提案されている．

投機的ルーティングはその 1 つであり，ルータ内の複数のパイプラインステージ，たとえば出力ポートの仮想チャネル割当てとクロスバスイッチの設定，を投機的に同一サイクルで実行する[16]．また，ルックアヘッドルーティングでは，1 ホップ先のルーティングを実行してルーティング結果をヘッダに持たせることで，ルーティング計算に続くルータ内パイプラインステージの開始サイクルを早め，所要サイクル数の削減を図る[4]．ただし，投機/ルックアヘッドルーティングではパケットヘッダの受信後にルーティング計算が必要であり，それらの遅延は 1 ホップごとに積算される．マッドポストマンスイッチングは，逐次転送されるパケットヘッダの受信とルーティング計算の遅延を削減するために，出力ポートを静的に決定してルーティング計算前に出力を開始する[7]．この方式は，2-D メッシュ上で次元順ルーティングを行う場合，メッセージが次元をターンするのはたかだか 1 回であり直進することが多いという性質を利用したものである．しかしながら，出力ポートを静的に直進すると決める方式では局所性のある通信パターンにうまく適合しない．

これらの問題を解決するため，本論文では動的通信予測スイッチングを提案する．動的予測スイッチングは，ルータに通信履歴を格納するメモリと予測器を付加し，次に入力するパケットの出力ポートを動的に予測する方式である[9),17),18)]．これにより，パケット受信前にルータ内の通信パス設定を行い，パケットを構成する phit（Physical Transfer Unit）を受信するとすぐに予測した通信パスに転送を行う．

本論文では，2-D トーラスネットワークを対象とし，動的通信予測の適用に必要な要素技術として予測ミスパケットを削減する手法，および予測ミスパケットを破棄する手法についても考察する．また，ネットワークシミュレータによる実験結果を示し，動的通信予測の有効性について議論する．

2. 動的予測スイッチング

2.1 2-D トーラスルータ

図 1 に，2-D トーラス用の動的予測スイッチングルータの構成を示す．本ルータは，予測器，ルーティングロジック，東西南北 4 つのネットワーク入出力ポートと PE（Processing Element）用の注入/排出ポート，およびクロスバスイッチで構成する．各入出力ポートは 1 つ以上の仮想チャネル（VC）を有し，それぞれ入力 VC，出力 VC と呼ぶことにする．また，入力ポートと注入ポートに対しては通信履歴を保存するためのメモリを設ける．このメモリは，先頭と末尾を示すポインタを持つ循環キュー構造とし，それぞれの入力/注入ポートがパケットの先頭 phit を転送するときに出力ポート番号をキューの末尾に追加記憶する．

図 2 (a) に示すように，ルータの非予測スイッチング・パイプラインは，入力 VC でのバッファリング（IB），ルーティング計算（RC），出力 VC 割当て（VA），クロスバスイッチ設定（SA），スイッチ転送（ST）の順で実行する．ただし，図 2 (a) はパケットヘッダが 3-phit で構成され，SA に投機処理を用いて VA と同一サイクルで実行する場合を示している．入力パケットは RC の結果に基づいて VA と SA を実行し，それらが成功すると ST によって phit 単位で転送されるものとする．VA または SA において，他の入力ポートからの要求との調停によって出力 VC 獲得またはクロスバスイッチ設定が行えなかった場合には，それらが完了するまで VA/SA を繰り返す．この状態をパケットのブロッキングと呼ぶ．

動的予測スイッチングでは，通信履歴の更新をトリガとして入力ポートまたは注入ポートが予測処理要求を行う．これを受けて，予測器が更新された通信履歴を参照し，次の出力ポートを予測する．そして，入力/ 注入ポートがパケット待ち状態になると，予測された出力ポートに利用可能な VC があるか調べ，VA を実行する．なお，VA は図 2 (b) に示すように IB ステージまでに完了すればよい．パケットの先頭 phit が入力ポートに到着すると，IB と同時に VA/SA を実行し，RC とそれに続く VA/SA の遅延なしに ST を実行する．この予測に基づく ST を PST（Predictive Switch Traversal）と呼ぶ．

PST は予測に基づくため，必ずしも正しい通信経路の確立を保証するものではない．そこで，パケットヘッダの IB を実行した後 PST と並行して通常の RC も実行する．RC が計算した出力ポート候補に予測処理で得られた出力ポートが含まれている場合，予測が成功したことになるので後続 phit に対しては IB/SA と PST ステージのみ実行する[⋆1]．そして，通信履歴に出力ポート番号を追加する．一方，RC によって予測が失敗したことが分かった場合には，予測によらない VA/SA，ST を実行する．これにより，入力 VC にバッファされたパケットを先頭 phit から正しい出力 VC に転送し直すことで通信を保証する．また，正しい出力ポートを通信履歴に追加する．通信履歴の更新が完了すると，現在のパケット転送とオーバラップして次の入力パケットの予測処理を要求することができる．したがって，

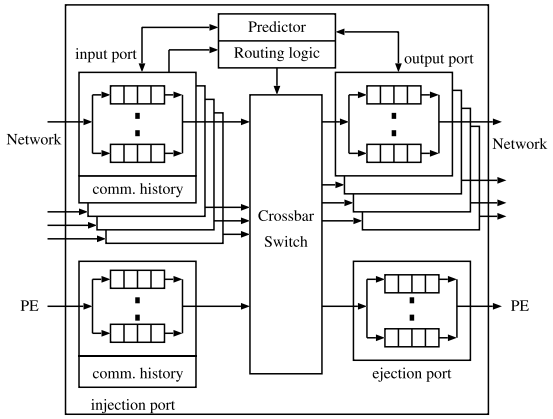


図 1　2-D トーラス用予測スイッチングルータ
Fig. 1　A predictive switching router for 2-D tori.

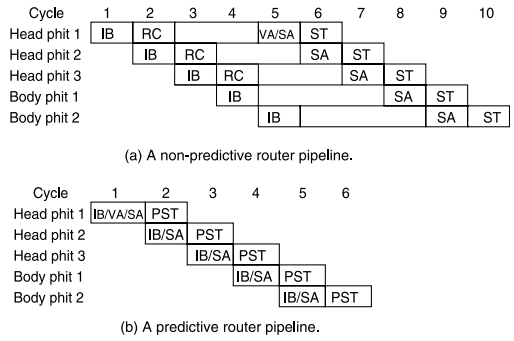


図 2　ルータのパイプライン：(a) 非予測スイッチング，(b) 予測スイッチング
Fig. 2　Router pipelines: (a) non-predictive switching, and (b) predictive switching.

⋆1 パケット全体に対してクロスバスイッチ設定を保持する場合は，phit ごとに SA を実行する必要はない．

予測処理はパケットの到着間隔内で完了すればよい．ただし，予測処理を行った入力ポートは実際にパケットが到着するまでは出力 VC を仮予約した状態となる．この状態で別の入力ポートに先着したパケットが仮予約された出力 VA を要求した場合には，仮予約をキャンセルして先着パケットの転送を優先する．すなわち，出力 VC の要求に 2 レベルの優先順位を設け，仮予約は低優先順位で VA を実行した状態を表す．そして，実際に入力したパケットヘッダの IB と同一サイクルで実行する VA によって出力 VC 獲得を確定する．また，RC の結果に基づく VA は高優先順位で実行する．図 2 (b) に示すサイクル 1 の VA/SA で出力 VC の獲得またはクロスバスイッチ設定に失敗したパケットは，ブロッキングされた状態となる．ブロッキングされたパケットは，非予測スイッチングパイプラインの RC が完了するか，または予測スイッチングによる出力 VC 獲得とクロスバスイッチ設定が完了するまで VA/SA を繰り返す．RC 完了まで出力 VC 獲得とクロスバスイッチ設定が完了しなかったパケットは，非予測スイッチングによって転送する．ここで，予測処理による出力 VC の仮予約時点では SA を実行していない．すなわち，クロスバスイッチ設定と入力 VC の出力調停は IB の前に処理していないことに注意されたい．

なお，ルーティングアルゴリズムによってはパケットの使用可能な VC に制約があることがある．その場合，仮予約した出力 VC を使用できるか否かは実際に入力したパケットヘッダの IB と同時に決定する．これにより，PST はルーティングアルゴリズム制約を満たす入出力 VC 割当てを守って実行する．また，予測処理は入出力ポート単位で行えばよく，通信履歴も仮想チャネルごとに区別する必要はない．

図 2 (a) のパイプラインでは，VA/SA ステージがクリティカルパスとなると考えられる[16]．ただし，パイプラインの定義から VA/SA は IB と並列実行可能である．予測スイッチングでは，2.3 節でパケット出力方向と予測値の比較を IB ステージに追加する．この追加処理は比較的簡単な論理演算ですむことから，本論文では IB が VA/SA の実行時間を超えないと仮定する．同様に，仮予約を前提とした 2 レベルの優先順位付き VA が投機的 SA の処理時間を超えないと仮定し，4 章では図 2 (a) と (b) は同一クロック周期で動作するとして評価する．

2.2　予測ミスに対するパケットの破棄

動的予測スイッチングは，ルータ間接続がビットシリアルリンクの場合やパケットヘッダが複数の phit に分割されて転送される場合，およびルーティング表参照等により RC の完了までに複数サイクルを必要とする場合の遅延短縮に効果が大きい．図 2 (a) と (b) の例では，予測スイッチングによるパケット 1 ホップあたりの遅延の削減は RC（3 サイクル）+VA/SA の最大 4 サイクルであるが，RC に必要なサイクル数によって遅延の削減量が変化する．RC が 1 サイクルで完了する場合，図 2 (a) の 2 サイクル目で予測ミスしたか否か確定する．図 2 (b) に示したように PST は 2 サイクル目から実行するが，送信 phit を隣接ルータで受信するための信号エッジを PST ステージの終わりの方で生成するとして，予測ミスした場合にヘッダ phit1 の受信信号エッジを隣接ルータに伝搬しないように制御できれば誤った経路へのパケット転送は発生しない．しかしながら，RC が 2 サイクル以上の場合に予測をミスすると誤った経路へパケットを転送してしまう．本論文ではこれを予測ミスパケットと呼ぶことにする．予測ミスパケットは不必要なネットワーク資源を消費するので，適正に破棄することが必要となる．また，パケットの最終 phit を識別する信号またはフラグを有する転送方式の場合には，予測ミスが判明した段階で最終 phit を転送し，予測ミスパケットのオーバヘッドを軽減することが望ましい．

X-Y 次元順ルーティング[⋆1]（以降 DOR と略す）や Duato のプロトコルに従う適応ルーティング[6]（以降 DPR と略す）のような最短経路ルーティングアルゴリズムでは，予測ミスパケットは最短経路範囲内にあるか否かを調べることにより検出できる．そこで，ネットワーク中に予測スイッチングを実行しない入力ポートを持つルータを配置し，通常の RC ステージに予測ミスパケットの検出機構を付加して破棄する方式を提案する．**図 3** に，5×5 の 25 ノードで構成するトーラスの例を示す．ここで，右端の列に位置するルータの X 次元入力ポートと上端の行に配置されたルータの Y 次元入力ポートは，予測スイッチングを行わないものとする．例としてノード 16 からノード 13 宛てのパケット通信を考える．DOR による通信経路は $16 \rightarrow 17 \rightarrow 18 \rightarrow 13$ であるが，いずれかのルータで予測ミスが発生しても右端または上端のルータに達すると最終 phit まで破棄される．また，予測スイッチングを実行するルータにおいても，予測器が与える出力ポートがブロッキングにより利用できない場合には，入力パケットに対して RC を実行する．したがって，パケットの宛先までの最短経路範囲外にあるいずれのルータにおいても，予測ミスパケットは検出/破棄されうる．たとえば，上記の例でルータ 17 が予測ミスによってパケットをルータ 12 方向に出力した場合，DOR に従って予測ミスパケットは Y 次元を直進するが，ルータ 7 または 2 で破棄することができる．ただし，予測ミスパケットが VC フロー制御[3] を用いる場合にトーラスサイクルを形成してデッドロックすることを防ぐこと，および予測ミスパケットのライブ

⋆1 まず X 次元方向に宛先ノードと X 座標が等しくなるまで転送した後，Y 次元方向に宛先まで転送する決定性ルーティングである．

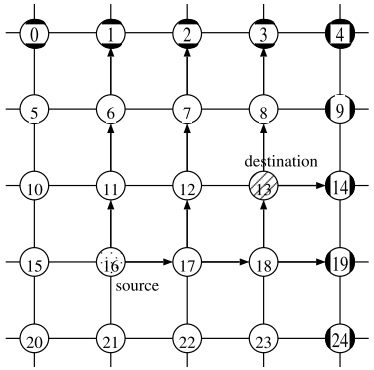


図 3　予測ミスパケットの破棄
Fig. 3　Discarding mis-prediction packets.

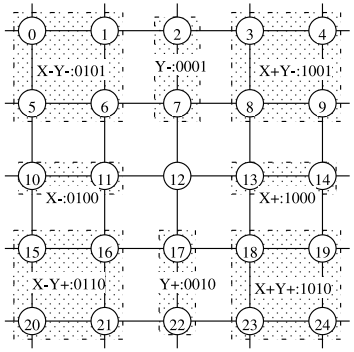


図 4　DOR に対するパケット出力方向の分類（ソース 12）
Fig. 4　Classification of output direction for packets in DOR from node 12.

ロックを防止するために予測スイッチングを実行しない入力ポートを持つルータを適切に配置することが必要であると考える．

予測スイッチングを行わないルータを図 3 の位置に加え，中央の行と列にも配置すれば予測ミスパケットの伝搬距離が短くなる．また，対角線ノードのみに配置すれば，予測スイッチングできる範囲が大きくなる．このように，予測ミスパケットが予測スイッチングを行わないノードで検出されるまでの平均距離を変えることで低遅延化と予測ミスのトレードオフをとることができる．

2.3　パケット出力方向による予測ミスの削減

DOR や DPR のような最短経路ルーティングでは，パケットの出力方向は静的に決定できる．たとえば DOR において，X 次元のパケットの出力方向が+方向であるときに予測スイッチングによって誤って 1 方向に出力することは避けるべきである．そこで，出力方向をパケットヘッダにエンコードしておき，最短経路と異なる方向へ予測ミスパケットが伝搬する機会を減少させる．このエンコード情報をヒントビットと呼ぶ．ヒントビットはルックアヘッドルーティングにおいて付加される 1 ホップ先のルーティング済み出力方向と類似したものであるが，予測スイッチングでは各ルータで更新されることはない．ヒントビットと予測値の比較は，IB ステージに追加する．これは，Path-Sensitive ルータ[11] や Guided Flit Queuing[12] をサポートするルータが，ルックアヘッドルーティングによってパケットヘッダに埋め込んだ VC 識別子を IB ステージの初めに抽出して使用するのと同様に，予測スイッチングではヒントビットを IB ステージで使用することを意味する．予測スイッチングでは，RC ステージの完了を待たずにパケット転送を開始することがルックアヘッドルーティングと異なる．ただし，静的に付加されたパケットの出力方向にルーティングし続けると宛先ノードを通り過ぎてしまう．したがって，静的な出力方向エンコードは明らかに誤った予測を排除する目的に使用するもので，単純にエンコードされた出力方向にルーティングするだけでは予測ミスパケットの発生を完全に排除することはできない．

以下に，DOR に対して具体的に説明する．

2.3.1　X-Y 次元順ルーティングへの適用

図 4 に，5 × 5 トーラスにおいて DOR を用いるとき，ノード 12 から出力するパケットに対する宛先ノードごとのヒントビットを示す．ヒントビットは，東西南北に対する $D_{x+}D_{x-}D_{y+}D_{y-}$ の 4 ビットで表し，パケットが進むべき方向のビットを 1 とする．たとえば，ノード 13 と 14 宛てのパケットは 1000 を持ち，ノード 16 宛てのパケットは 0110 を持つ．ヒントビットをパケットの先頭にエンコードしておき，入力/注入ポートにおいて予測器が出力する値とビット論理演算を行うことにより，PST ステージを実行するか否かを決定する．この機構を**図 5** に示す．

予測器は，東西南北の出力ポートと排出ポートを 5 ビット $P_{x+}P_{x-}P_{y+}P_{y-}P_c$ で表現した値を出力する．5 ビットのうち，予測した出力ポートに相当するいずれかの 1 ビットだけが 1 となる．ヒントビットと予測値をルーティングアルゴリズムに応じたデコーダでビット論理演算することで，PST ステージを実行するための PST_enable 信号を生成する．**表 1**

表 1　X-Y 次元順ルーティングに対する予測スイッチングの実行条件
Table 1　Conditions to validate predictive switching requests for the X-Y dimension order routing.

input port	output port				
	X+	X-	Y+	Y-	Ejection
Injection	$D_{x+} * P_{x+}$	$D_{x-} * P_{x-}$	$D_{y+} * P_{y+} * \overline{D_{x+}} * \overline{D_{x-}}$	$D_{y+} * P_{y+} * \overline{D_{x+}} * \overline{D_{x-}}$	
X+		P_{x-}	$D_{y+} * P_{y+}$	$D_{y-} * P_{y-}$	P_c
X-	P_{x+}		$D_{y+} * P_{y+}$	$D_{y-} * P_{y-}$	P_c
Y+				P_{y-}	P_c
Y-			P_{y+}		P_c

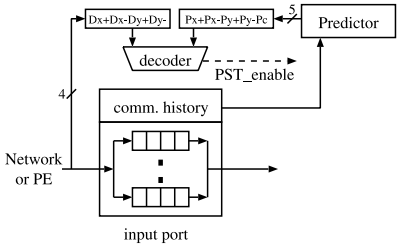


図 5　入力パケットに対する出力方向と予測値の比較
Fig. 5　Comparison between an output direction and a prediction value.

に X-Y DOR に対する入力ポートごとの各出力ポートに対する PST_enable が有効となる条件を示す．ここで，表中の空欄はルーティング制約によって禁止されたターンを表すものとする．

たとえば，注入ポートから出力ポート X+に予測スイッチングするための条件は $D_{x+} * P_{x+} = 1$ となり，簡単なビット AND 演算で高速に判定することができる．注入ポートから Y 方向の出力ポートに予測スイッチングするための条件が $D_{y+} * P_{y+} * \overline{D_{x+}} * \overline{D_{x-}}$ となっているのは，X-Y DOR のルーティング制約が Y 方向よりも X 方向を優先するためである．これらの条件により，注入ポートから予測ミスパケットが伝搬することは防止することが可能である．また，入力ポート X+における出力ポート X-に対する条件は $P_{x-} = 1$ のみでよい．これは，注入ポートにおいてすでに $D_{x-} = 1$ が保証されていることによる．排出ポートに対する条件は $P_c = 1$ であり，これは注入ポートを除くネットワークポートから入力されたパケットは予測値のみによって PST ステージを実行することを表している．

以上の条件が成立しない場合は，予測が失敗したことになるので，PST ステージを実行せずに RC ステージによるルーティングに従う．なお，本方式はデコーダの論理を変更することによって，Y-X DOR や DPR のような適応ルーティングに容易に対応可能である．

3. 予測アルゴリズム

本章では，4 章で評価に用いる動的予測アルゴリズム SPM（Sampled Pattern Matching）と静的予測アルゴリズム SS（Static Straight），およびそれら両者の中間的なコストで実装可能と考えられる LP（Latest Port）について説明する．ただし，各予測器の具体的なハードウェアコスト比較は今後の課題であり，本論文では詳しく議論しない．

3.1 パターンマッチ予測アルゴリズム SPM

SPM は，文献 8) で提案されたパターンマッチングに基づくユニバーサル予測アルゴリズムに，系列の長さ制限等の制約条件を付けたものである．我々は，過去の通信履歴から繰返しパターンを検索することによって，並列プログラムとルーティングアルゴリズムが持つ通信の規則性を抽出できると考え予測スイッチングに応用する．

具体的な予測手順を以下に示す．ここで，通信履歴は有限長の出力ポート番号列とし，$X_1^n = X_1, X_2, ..., X_n$ で表す．予測器は，通信履歴 X_1^n を参照して，次の出力ポート番号 X_{n+1} を予測するものとする．

(1)　X_1^n とその接尾系列（最近の通信履歴）X_j^n，$1 \leq j < n$ のパターンの一致を検索し，最長一致となる系列 $X_{j_{min}}^n$，$j_{min} = min\{j | X_j^n = X_k^{k+n-j}, 1 \leq k < n\}$ を求める．最長一致系列長は $D_n = n - j_{min} + 1$ となる．
(2)　最長一致系列の接尾系列（長さ $l = \lceil \alpha D_n \rceil$，$0 < \alpha \leq 1$）を探索系列 X_{n-l+1}^n とする．
(3)　探索系列 X_{n-l+1}^n と一致する系列の次の値の出現回数を調べる．
(4)　出現回数の最も大きい値が唯一に決まるときは，その値を予測値 X_{n+1} とする．唯一に決まらないときは，出現回数の最も大きい値の中でより最近に現れた値を予測値とする．

以下に具体例を示す．通信履歴を古い順に左から過去の出力ポート番号とし，右端の値?を予測するものとする．

00 0012 312 0012 233 0012 21 0012 ?

この例では，最長一致系列は 0012 となる．$\alpha = 1$ とすれば探索系列も 0012 であるから，それに一致する系列の次の値は下線部の 3，2，2 である．したがって，出現回数の多い 2 が予測値となる．

SPM は，繰返しパターンを含む系列（履歴）から次に出現する値を予測するという性質によりプロセッサの分岐予測においても高い予測精度を達成することが報告されている[13]．

3.2　静的直進予測アルゴリズム SS

SS は，入力パケットがつねに同一次元を直進すると予測する．すなわち，北の入力ポートに到着したパケットは南の出力ポートへ，東からは西へといった具合である．したがって，予測は過去の通信履歴に関係なく静的に出力ポートを決定する．この方式は，マッドポストマンスイッチングで用いられるもので，2-D メッシュ/トーラス上の次元順ルーティングで通信距離が長いときに有効と考えられる．なぜならば，パケットが X 次元から Y 次元に曲がる回数はたかだか 1 回であることによる．

本論文の実験では，PE からネットワークへの注入ポートに対する出力ポートの予測は，ランダムに行うものとする．

3.3　直前ポート予測アルゴリズム LP

LP は，入力パケットが 1 つ前のパケットと同一の出力ポートを選択すると予測する．したがって，通信履歴は入力ポートごとに 1 つ分で済み，予測処理も高速に実行可能である．単純な予測機構であるが，SS と同様に直進が多いルーティングアルゴリズムに対する有効性が期待できる．また，隣接通信パターンでは，ネットワークからの入力パケットがすべて自ノード宛てとなるため，つねに排出ポートが選択される．したがって，隣接通信を多く含むような空間的局所性の高い通信パターンに対しては SS よりも高い予測精度が期待できる．

4. 評　　価

4.1　シミュレーション

3 章に述べた各予測アルゴリズムについて，ネットワークシミュレータを使用した実験を行った．シミュレータは，文献 4) に紹介されている booksim をもとにして，予測スイッチングへの対応を図ったものを使用した．シミュレーション条件は，以下のとおりである．

ネットワーク： 32-ary 2-cube（4.6 節の LU の評価では，通信トレース作成の都合により 8-ary 2-cube を対象とする）．

通信パターン： ユニフォームランダムとビット列逆順（bit-reversal）の 2 つでは，全ノードに同一のパケット生成率（phit/cycle）を設定し，ノードごとにパケット生成タイミングのランダム性を持たせるためベルヌーイ・プロセスを使用[4]．NAS 並列ベンチマークの LU（サイズ W）における MPI メッセージトレースから作成する通信パターンでは，シミュレータの 1 サイクルをトレースデータの何サイクルに割り当てるかを変更して通信負荷を変化．

パケット長： 16-phit．ただし，予測ミスパケットは 4-phit 目を最終 phit に設定して短縮．

スイッチング： バーチャルカットスルー方式．

ルータ間ケーブル遅延： 2 サイクル．

ルーティング： X-Y 次元順ルーティング（DOR）．

VC： 物理チャネルあたり 2 本の VC（各 16-phit 分）を実装し，各次元とも dateline[⋆1]を超えないパケットは VC0 を使用し，dateline を超えると VC1 を使用．

パイプライン： 図 2 に示したように，予測スイッチングによらない場合は IB-RC（3 サイクル)-VA/SA-ST の 6 サイクルで実行．予測スイッチングの場合は，IB/VA/SA-PST の 2 サイクルで実行．

予測スイッチングをしないポートの配置： X，Y 各次元の入力ポートで予測スイッチングしないルータを，それぞれ $32/m$ 離れた m 個の行と列に配置（$m = 2$，3，4）．

SS の実験条件： 予測値取得の遅延は特に考慮せず，潜在的に全パケットが予測スイッチング可能と仮定．

LP の実験条件： 入力ポートごとの通信履歴長は，直前のパケットの出力ポートを記憶する 1×3 ビット．予測処理遅延は発生せず，全パケットについて通信履歴の値を参照可能．

SPM の実験条件： $\alpha = 1$ とし，入力ポートごとの通信履歴はパターンマッチによって十分な予測精度が得られるよう 512 パケットまでの出力ポート番号を格納[⋆2]．また，予測器はルータあたり 1 つとし，入力ポートまたは注入ポートが通信履歴を更新してから当該ポートに対する次パケットの出力ポートを予測するのに 4 サイクルかかると仮定[⋆3]．

⋆1 ラップアラウンドチャネルによって形成されるトーラスサイクルによるデッドロックを防止するため，各次元に dateline と呼ぶ VC の切替え位置を設定する．

⋆2 NAS 並列ベンチマークを用いた予備評価では，最長一致系列長が短いと高い予測精度を得られない場合があった[17]．

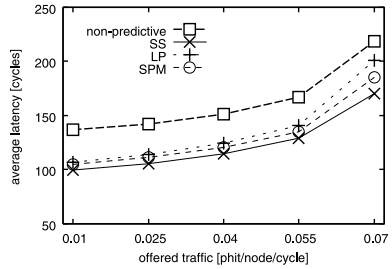
⋆3 予測遅延は具体的な SPM のハードウェア設計に基づいた値ではない．本論文は予測精度の影響を明らかにすることを主眼とし，実験において予測遅延が予測スイッチング実行の大きな障害とならない値を仮定した．

他の入力ポートの予測処理待ちの間に到着したパケットは予測スイッチングしない．

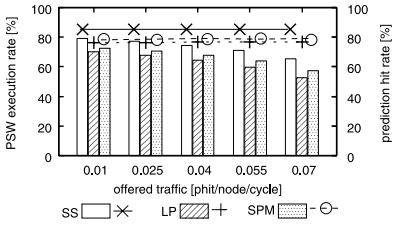
データ収集： シミュレーション開始から 1 万パケットを受信するまでをネットワークのウォームアップとして無視し，その後，約 12 万パケットを受信するまでの平均遅延，スループット，予測ヒット率，予測スイッチング実行率（受信パケットの全ホップ数に対して，予測スイッチングを実行したホップ数の割合）を測定[⋆1]．ただし，LU の評価ではウォームアップは行わない．

4.2　ユニフォームランダム通信に対する考察

図 6 に，ユニフォームランダム通信の実験結果を示す．ここで，$m = 2$（4.1 節の予測スイッチングをしないポートの配置を参照）とした．それぞれ，図 6 (a) はネットワークへのパケット注入負荷（横軸）を変化させたときのパケット平均遅延（縦軸）を表す．図 6 (b) は，予測スイッチング（図中 PSW と表記）実行率を棒グラフ（縦軸左）で，予測スイッチングを実行した受信パケットのホップに対する予測ヒット率を折線グラフ（縦軸右）で示す．



(a) Average latency



(b) Prediction

図 6　32-ary 2-cube 上のユニフォームランダム通信の結果
Fig. 6　Experimental results for uniform random traffic on a 32-ary 2-cube.

⋆1 本章に示すすべての実験において，12 万パケット受信時点でネットワークは十分に安定していることを確認した．

図 6 (a) から，予測スイッチングを行わない場合（凡例 non-predictive）に対する平均遅延の低下が確認できる．低遅延化効果は，予測ヒット率が高く，予測スイッチング実行率が大きいほど大きい．予測ヒット率は SS が最も高く約 85%であった．これは，DOR の性質上，パケットが 1 つの次元を直進する割合が高いためである．この傾向はネットワークサイズが大きくなるほど強くなるため，大きなネットワークほど予測スイッチングの効果が大きくなる[18]．LP と SPM は，それぞれ約 77，79%の予測ヒット率を示した．これらの動的通信予測ヒット率が SS よりも低い理由は，ユニフォームランダム通信パターンに通信の規則性がないことによる．

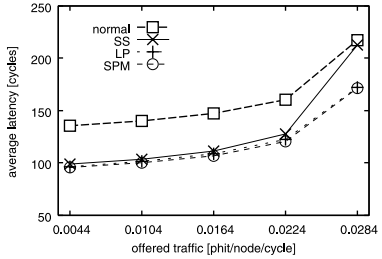
いずれの予測アルゴリズムにおいてもネットワーク負荷が高くなると予測スイッチング実行率は低くなる．これは，ブロッキングにより出力ポートが使用できない場合が増えることによる．SS の予測スイッチング実行率が高い理由は，ネットワークポートがつねに直進方向の出力ポートを予測するため他の入力ポートとの競合が少ないことがあげられる．また，図 6 (b) の SPM は予測遅延 4 サイクルの場合を示しているが，SPM の予測遅延を 8 サイクルに増やした場合，予測ヒット率はほとんど変わらないが予測スイッチング実行率は高通信負荷時に約 8%減少した．これは，通信負荷が高くなるにつれて予測処理がパケット到着に間に合わない場合が発生するためである．

結果として，ユニフォームランダム通信での平均パケット遅延は，予測ヒット率と予測スイッチング実行率の高い SS が最も小さくなった．低負荷時は，予測スイッチング実行率 × ホップ遅延の減少サイクル数 (1〜(RC+VA/SA) の範囲)× 平均ホップ数分の低遅延化効果となって現れている．なお，通信負荷が 0.07 を超えると，非予測スイッチングおよびいずれの予測アルゴリズムを用いた予測スイッチングでもネットワークが飽和する．ネットワークが飽和するまでの受信スループットは，ネットワークに注入するパケットの通信負荷値とほぼ同一の値で線形に上昇し，非予測スイッチングと予測スイッチングに大きな差はなかった．

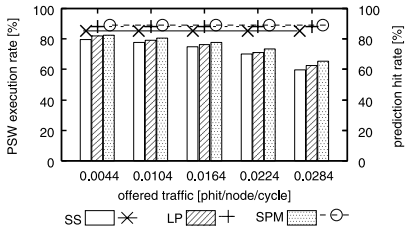
4.3　ビット列逆順通信に対する考察

図 7 に，ビット逆順通信の結果を示す（$m = 2$）．この通信パターンは，2 進数で表した送信元ノード番地（$b_{n-1}, b_{n-2}, ..., b_0$）から宛先（$b_0, b_1, ..., b_{n-1}$）に送信を繰り返すものである[4]．

ビット列逆順通信は通信の規則性を有するため，LP と SPM の予測ヒット率（それぞれ約 88，89%）がユニフォームランダム通信のときよりも高い．SS については，ユニフォー

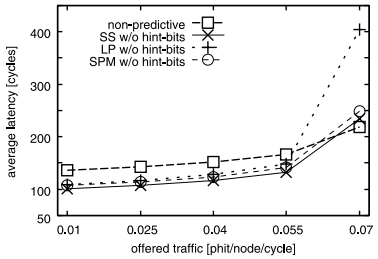


(a) Average latency

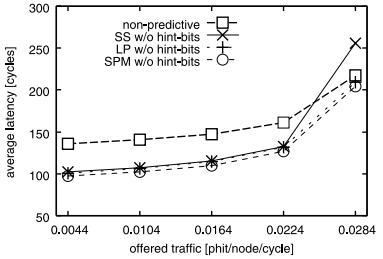


(b) Prediction

図 7　32-ary 2-cube 上のビット列逆順通信の結果
Fig. 7　Experimental results for bit-reversal traffic on a 32-ary 2-cube.



(a) Average latency for uniform random traffic



(b) Average latency for bit-reversal traffic

図 8　ヒントビット無効時の予測ミスパケットの影響
Fig. 8　Effect of the mis-prediction packets when the hint-bits are annulled.

ムランダム通信時と同等の予測ヒット率（約 85%）であった．予測スイッチング実行率も LP と SPM の方が SS よりも高いため，低遅延化効果も大きい．したがって，通信パターンに規則性が在ると，動的通信予測が SS のような静的予測よりも有効であることが分かる．

図 7 (a) で LP と SPM のグラフはほとんど重なっているが，これは低通信負荷時には両者の予測ヒット率と予測スイッチング実行率があまり変わらないこと，高負荷時にはビット列逆順通信パターン自体の輻輳が遅延増加の支配要因であることによる．なお，SPM の予測遅延を 8 サイクルに増やすと SPM の予測スイッチング実行率は高通信負荷時に約 5%減少した．SS の低遅延化効果は，ネットワーク負荷の上昇とともに小さくなっている．これは，SS によって発生する予測ミスパケットが本来のビット列逆順通信とネットワーク資源を競合する機会が増えることによる．予測ミスパケットのオーバヘッドについては，4.4 と 4.5 節で考察する．

ビット列逆順通信パターンは，通信負荷約 0.03 でネットワークが飽和する．飽和するまでの受信スループットは，非予測スイッチングと予測スイッチングで大きな差は見られなかった．

4.4　予測ミスパケット削減の効果

図 8 に，2.3 節で述べたヒントビットを無効にした場合（図中では w/o hint-bits と表記）の平均遅延を示す．図 8 (a) がユニフォームランダム通信，(b) がビット列逆順通信の結果であり，それぞれ non-predictive の値は図 6 と図 7 と同一である．

ユニフォームランダム通信では，いずれの予測アルゴリズムに対してもヒントビットを無効にすると高負化時の平均遅延が非予測スイッチングよりも大きくなった．その理由は，予測ミスパケットによってネットワークが飽和してしまうことによる．図 6 (a) と図 8 (a) を比較すると，図 6 (a) の高負化時に予測スイッチングによる低遅延化が小さい予測アルゴリズムほど図 8 (a) の高負化時に予測ミスパケットのオーバヘッドが大きいことが分かる．高負荷時には，予測ミスパケット以外の正しく経路選択されたパケットだけでもネットワーク資源が飽和状態近くまで使用されている．その状態で予測ミスパケットが正しく経路選択さ

れたパケットとネットワーク資源を競合する機会が増えるほど，予測ミスパケットのオーバヘッドが大きくなる．

ビット列逆順通信では，ユニフォームランダム通信に比べて予測ヒット率が高いため，予測ミスパケット数も少ない．そのため，予測ミスパケットによるオーバヘッドも相対的に小さい．ただし，高負化時に SS は非予測スイッチングよりも平均遅延が大きくなり，LP と SPM は予測ミスパケットのオーバヘッドが予測スイッチングによる低遅延化効果を相殺する結果となった．

以上の結果から，ヒントビットによる予測ミスパケットの削減はネットワークが高負荷時に有効であることが確認できた．なお，全受信パケットのうち，1 回以上ヒントビットによって予測ミスパケットの発生を抑制したパケット数の割合は，52〜76%であった．また，今回用いた実験条件（12 万パケット）に対するネットワークの受信スループットに関しては，ヒントビットによる予測ミスパケットの削減を行わないと，ユニフォームランダム通信で最大 7%，ビット列逆順通信で最大 9%低下した．

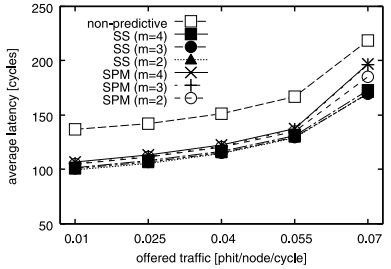
4.5　予測スイッチングをしないポートの配置に対する評価

図 9 に，予測スイッチングをしない入力ポートを持つルータの配置を変えた場合（$m = 2$, 3, 4）の平均遅延の変化を示す．図 9 (a) がユニフォームランダム通信，(b) がビット列逆順通信の結果である．なお，予測アルゴリズムは最も簡単な SS と最も複雑な SPM の場合を示した．$32/m$ の値が X と Y の各次元で予測ミスパケットを検出/破棄するルータ間の最大直進距離を表すので，m の値が大きい方が予測ミスパケットが伝搬する最大距離は短くなる．ただし，予測スイッチングできる距離も短くなる．

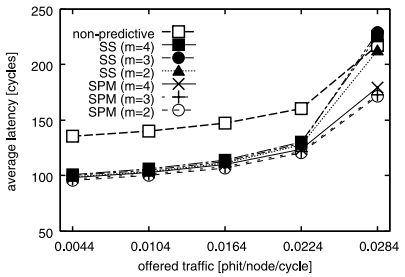
実験の結果，$m = 2$ の場合が最も低遅延となった．これは，予測スイッチングしない入力ポートを持つルータを増やすよりも，予測スイッチングの実行可能距離の長い方が効率的に機能したことを表す．ただし，SS と SPM のどちらも m の値による平均遅延への影響は比較的小さかった．この理由は，いずれのルータにおいても入力パケットがブロッキングされると RC を実行し，予測ミスパケットを検出/破棄できるからと考察される．同様な理由で，受信スループット値についても m の値による大きな影響は観測されなかった．また，前節の結果と比較すると，予測スイッチングをしない入力ポートを持つルータの配置を変えるよりも，ヒントビットによる予測ミスパケット削減の方が高負荷時に有効であることが分かる．

4.6　LU の通信パターンに対する考察

図 10 に，8-ary 2-cube に LU 通信パターン（サイズ W で発生した総メッセージ数約 215



(a) Average latency for uniform random traffic



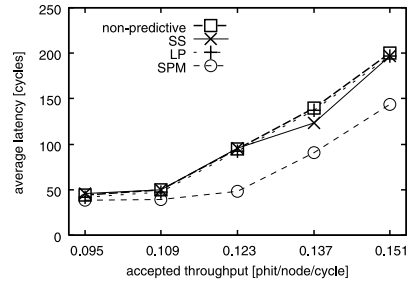
(b) Average latency for bit-reversal traffic

図 9　予測スイッチングをしないポートの配置に対する評価
Fig. 9　Experimental results for arrangement of non-predictive ports.

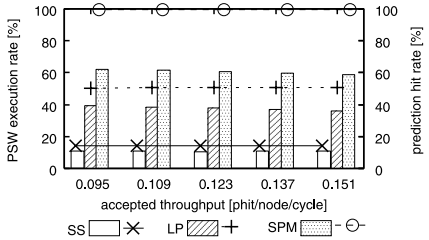
万の先頭 12 万メッセージ分）を与えた場合の実験結果を示す．この通信パターンは，すべて隣接ノード間通信で構成される．また，送信元ノードから 4 隣接ノードへの通信順序にも強い規則性が存在する[9]．

SPM は，通信の規則性を反映し 99%以上の高い予測ヒット率を示した．SPM の予測スイッチング実行率は，通信負荷の上昇につれて約 62%から 58%まで減少した．6 割前後のパケット転送（ホップ）について予測スイッチングが成功し，受信スループットが大きくなった場合に非予測スイッチング（図中 non-predictive）よりも平均遅延が小さくなった．今回の LU の実験条件では，SPM の予測遅延を 4 サイクルから 8 サイクルに変更しても，予測スイッチング実行率はほとんど低下しなかった．

LP は，約 50%の予測ヒット率であった．これは，隣接ノードへの通信パケットが送信元ノードでの注入ポートにおいて予測ミスとなり，宛先ノードでのネットワークポートにおい



(a) Average latency



(b) Prediction

図 10　8-ary 2-cube 上の LU の結果
Fig. 10　Experimental results for a traffic pattern in LU on a 8-ary 2-cube.

て予測がヒットすることによる．前者は，パケットの出力先が東西南北と変化することによる．後者は，ネットワークポートから入力されたパケットが，排出ポートへの転送を繰り返すことによる．LP の予測スイッチング実行率は，通信負荷の上昇にともない約 39%から 36%に漸減した．SS は注入ポートの予測がランダムに的中する以外，ネットワークポートの予測はすべてミスとなる．予測ヒット率は約 14%，予測スイッチング実行率も約 11%と小さい．受信スループット 0.137 のとき，SS の平均遅延が LP よりも小さい理由は，LP の方が予測ミスパケットのオーバヘッドが大きいことによる．結果的に，LP と SS は非予測スイッチングと比較して受信スループットに対する顕著な平均遅延の低下は観測されなかった．

なお，約 215 万メッセージの通信トレースの中ほどの 12 万メッセージ分，および最後の 12 万メッセージ分についても同様の実験を行った．LU の通信パターンは，実行経過にともないメッセージサイズが小さくなり，通信頻度が高くなる．本論文の実験では，通信トレースのメッセージサイズは無視して 1 メッセージを 16-phit の 1 パケットに置き換えた．したがって，図 10 (a) のグラフの横軸の値が後半の通信トレースを用いるほど小さな受信スループット値でネットワークが飽和する結果となった．受信スループット値以外の結果は，用いる通信トレース位置を変えても図 10 (a) と (b) とも同一の概形を示した．

以上の結果から，局所性/規則性を有する通信パターンに対しては，SPM のように通信履歴を活用した動的通信予測が有効であるといえる．

5. 関 連 研 究

通信の局所性を利用した予測スイッチングとして，エンドノード間の通信集合を静的，または動的に予測し，時分割多重方式（TDM: Time Division Multiplexing）でエンドノード間接続をスケジューリングする方式が提案されている[5]．この方式は，大域的な予測器および通信スケジューラを必要とするので，メッシュ/トーラスのような分散ルーティングを行う直接網に応用することは困難と考えられる．本論文で示した予測スイッチングは，2-D トーラスを構成する各ルータが個別にルータ内パスを確立することが TDM によるエンドノード間接続予測と異なる．

投機ルーティング[16] やルックアヘッドルーティング[4] は，ルーティング計算と出力 VC 割当て等のルータ内パイプラインステージを同時に実行することでルータあたりのホップ遅延を削減する．したがって，ルーティング計算の遅延を削減することは困難であり，パケットヘッダが複数の phit に分割されて転送される場合にはその逐次受信遅延が発生する．たとえば，IBM Blue Gene/L の 3-D トーラスネットワークでは 6 入力 6 出力のビットシリアルリンクを用い，入力ポートにおいて 8 バイトのパケットヘッダを 8 ステージの入力パイプラインで処理する[1]．ルーティングには，ヒントビットと呼ぶ 6 ビットの情報を使用するが，各ルータでパケットヘッダ内の宛先を参照して隣接ルータのためのルックアヘッドルーティングを行い，ヒントビットを正しく設定することが必要となる．

Express VC は，仮想的に非隣接ルータ間でバイパス経路を構成することにより中継ルータにおける所要パイプライン段数を削減する[14]．したがって，局所性を持つ通信パターンに対する低遅延化効果は有さない．また，Express VC を使用する場合とそうでない場合を高速に識別する必要があること，そのためにパケットヘッダが 1phit に納まるような広い物理チャネルを仮定していること，パケット転送に先立って Express VC 使用のための信号送信が必要となる等，ネットワークオンチップに特化した設計となっている．予測スイッチングは，局所的な通信パターンやパケットヘッダを複数 phit に分割して転送する細い物理

チャネルの場合にも低遅延化効果を有することが異なる．

Path-Sensitive ルータ[11] や Guided Flit Queuing[12] は，ルックアヘッドルーティングを行ってパケット出力方向に応じた入力 VC 番号をヘッダに埋め込む．また，DBBM は HoL ブロッキングを軽減する目的で宛先集合に応じたバッファ管理を行う[15]．本論文で述べたパケット出力方向のエンコードはこれらと異なり，必ずしも個別の入力 VC を指定するものではない．

マッドポストマンスイッチングは，パケットヘッダの受信完了を待たずにパケット出力を開始する[7]．それにより，ビットシリアルリンクを使用した場合のパケットヘッダの逐次受信遅延とルーティング計算遅延を削減する．本論文で提案する予測スイッチは，動的通信予測までを扱うことがマッドポストマンスイッチングと異なる．

6. おわりに

本論文では，通信予測を利用した低遅延通信技術を提案し，2-D トーラスネットワークに対する実現手法とその効果について考察した．ルーティングアルゴリズムとして，2-D トーラスに用いられることの多い次元順ルーティングを用いたシミュレーション実験から，次のことが分かった．

(1) DOR のように規則的なルーティングを用いる場合，通信距離がある程度長ければ，比較的簡単な SS や LP でも予測スイッチングによる低遅延化効果が得られる．
(2) 通信の規則性が強い通信パターンほど動的予測スイッチングの効果が得られやすく，ゼロ負荷からネットワークが飽和する直前の通信負荷までに対するパケット平均遅延を低下させる．
(3) 予測ミスパケットは，通信負荷が高い場合の通信遅延に悪影響を与える．この対策として，ヒントビットを用いた予測ミスパケットの削減が有効である．
(4) 隣接通信のような局所性の強い通信パターンに対して，通信履歴を利用した動的通信予測器は高い予測精度を示し，通信の低遅延化に効果的である．

今後の課題として，コスト性能比を考慮した通信予測器の検討，高次元ルータへの予測スイッチングの適用などがあげられる．

謝辞　本研究は，一部科学研究費補助金基盤研究（B）課題番号 17360178，基盤研究（C）課題番号 19500040 および NII 共同研究（提案型）予測機構を持つルータに関する研究の援助による．また，予測器に関して貴重なご意見をいただいた東工大・吉瀬謙二講師と山形大・岩田賢一准教授に感謝します．

参考文献

1) Adiga, N.R., et al.: Blue Gene/L Torus Interconnection Network, IBM Res. & Dev., Vol.49, No.2/3, pp.265–276 (2005).
2) Bailey, D., Harris, T., Saphir, W., Wijngaart, R., Woo, A. and Yarrow, M.: The NAS Parallel Benchmarks 2.0, Technical Report NAS-95-020, NASA Ames Research Center (1995).
3) Dally, W.J.: Virtual-Channel Flow Control, *ISCA90*, pp.60–68 (1990).
4) Dally, W.J. and Towles, B.: *Principles and Practices of Interconnection Networks*, p.550, Morgan Kaufman Publishers (2003).
5) Ding, Z., et al.: Switch design to enable predictive multiplexed switching in multiprocessor networks, *Proc. IPDPS05* (2005).
6) Duato, J.: A New Theory of Deadlock-Free Adaptive Routing in Wormhole Network, *IEEE Trans. Parallel and Distributed Systems*, Vol.4, No.12, pp.1320–1331 (1993).
7) Izu, C., Beivide, R. and Jesshope, C.: Mad-Postman: A look-ahead message propagation method for static bidimensional meshes, *Proc. 2nd Euromicro Workshop on Parallel and Distributed Processing*, pp.117–124 (1994).
8) Jacquet, P., Szpankowski, W. and Apostol, I.: A universal predictor based on pattern matching, *IEEE Trans. Info. Theory*, IT-48, 6, pp.1462–1472 (2002).
9) 鎌倉，吉永，鯉渕：2D トーラスネットワークにおける動的予測ルーティング，情報処理学会研究報告，2006-ARC-169, pp.97–102 (2006).
10) Kim, J. and Lilja, D.J.: Characterization of Communication Patterns in Message-Passing Parallel Scientific Application Programs, Technical Report, HPPC-97-10, University of Minnesota (1997).
11) Kim, J., Park, D., Theocharides, T., Vijaykrishnan, N. and Das, C.R.: A Low Latency Router Supporting Adaptivity for On-Chip Networks, *42nd DAC'05* (2005).
12) Kim, J., Nicopoulos, C., Park, D., Narayanan, V., Yousif, M.S. and Das, C.R.: A Gracefully Degrading and Energy-Efficient Modular Router Architecture for On-Chip Networks, *ISCA'06* (2006).
13) 吉瀬，岩田：分岐予測の精度と履歴情報との関係について，信学会大会論文集，A-1-25 (2005).
14) Kumar, A., Peh, L.-S., Kundu, P. and Jha, N.K.: Express Virtual Channels: Towards the Ideal Interconnection Fabric, *Proc. ISCA'07*, pp.150–161 (2007).
15) Nachiondo, T., Flich, J. and Duato, J.: Destination-Based HoL Blocking Elimination, *ICPADS'06* (2006).
16) Peh, L-S. and Dally, W.J.: A delay model and speculative architecture for pipelined routers, *Proc. HPCA*, pp.255–266 (2001).

17) Yoshinaga, T., Kamakura, S. and Koibuchi, M.: Predictive Switching in 2-D Torus Routers, *Proc. IWIA06*, pp.65–72 (2006).
18) 吉永，村上，鯉渕：2-D トーラスネットワークにおける動的通信予測の効果，SACSIS07 論文集，pp.219–226 (2007).

(平成 19 年 10 月 3 日受付)
(平成 20 年 2 月 13 日採録)

吉永　努（正会員）

昭和 61 年宇都宮大学工学部情報工学科卒業．昭和 63 年同大学大学院工学研究科修士課程修了．同年より宇都宮大学工学部助手．平成 9 年から翌年にかけて電子技術総合研究所・客員研究員．平成 12 年より電気通信大学大学院情報システム学研究科助教授．現在，同准教授．博士（工学）．並列計算機アーキテクチャ，クラスタ計算，ホームネットワーク等に興味を持つ．IEEE，電子情報通信学会各会員．

村上　弘和（学生会員）

平成 18 年電気通信大学電気通信学部卒業．平成 20 年同大学大学院情報システム学研究科博士前期課程修了．並列計算機ネットワーク，ルーティングアルゴリズム等に興味を持つ．現在，株式会社ブリヂストンに勤務．

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 年度より日本学術振興会特別研究員．現在，国立情報学研究所助教，総合研究大学院大学複合科学研究科情報学専攻助教（兼任）．相互結合網と並列処理に関する研究に従事．