社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ARC－169 (17)
2006／8／1

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

鎌倉 正司郎† 吉永 努† 鯉渕 道紘‡

†電気通信大学 大学院情報システム学研究科 ‡国立情報学研究所

並列計算機ネットワークにおいては，高スループット・低遅延通信を実現することが重要である．我々は，並列アプリケーションに見られる通信の規則性に着目し，その特性を利用して低遅延通信を実現するための動的予測ルーティングを提案する．本手法は，過去の通信履歴を基に次パケットの出力ポートを動的に予測する．予測が成立すれば，パケットヘッダのデコードからルータ内のスイッチ設定までに要する時間を短縮することができる．本論文では，動的予測機構を支援するルータのアーキテクチャと動的予測機構の有効性について考察する．NAS 並列ベンチマークプログラムから得られた通信トレースを用いた動的予測のヒット率を予備評価したところ，70～90%程度の予測精度を得られることがわかった．

Dynamic Predictive Routing For 2-D Torus Networks

Shojiro KAMAKURA† Tsutomu YOSHINAGA† Michihiro KOIBUCHI‡

†Graduate School of Information Systems, University of Electro-Communications
‡National Institute of Informatics

In parallel computer networks, it is important to provide high-bandwidth and low-latency communication performance. We propose a dynamic predictive routing scheme to archive low-latency communication by utilizing regularity of communication patterns in parallel applications. Our scheme dynamically predicts an output port for a next incoming packet based on the communication history. Then, the predicted output port and an internal router switch are allocated speculatively before the actual routing computation. Hence the packet hop latency is reduced when the prediction hits. In this paper, we consider a router architecture which supports the dynamic prediction as well as effectiveness of dynamic and static predictors. Our preliminary evaluation for the dynamic prediction model shows approximately 70 to 90% hit ratios for NAS parallel benchmark programs.

1. はじめに

並列計算機ネットワークの性能面で重要な要素に，スループットと通信遅延時間がある．近年，1 万ノード以上から構成される超並列計算機が存在し，そのようなシステムのネットワークは大規模なものとなる．アプリケーションの実行性能は，通信距離が大きくなると通信遅延の影響を受けやすい．また，ネットワークの構成要素となるルータ内のパイプライン段数が増加することも遅延の増加につながる．メッシュやトーラスのような直接網において低遅延通信を実現するためには，ルータごとにかかる通信処理時間をできるだけ短縮する必要がある．ルータあたりの通信遅延時間を短縮する関連研究としては，マッドポストマンスイッチング[1]や投機ルーティング[2]が挙げられる．マッドポストマンスイッチングは，シリアル通信においてパケットヘッダ全体の到着とデコードを待たずに，パケットが入力ポートと同次元を直進すると仮定して通信するモデルである．この手法は，通信距離が大きく，且つ次元順ルーティングのようにパケットが直進する頻度が高い場合に有効となる．投機ルーティングは，ルータ内の複数のパイプラインステージ(例えば仮想チャネルの割付とクロスバスイッチの設定)を同一サイクルに実行することで，ルータあたりのホップ遅延を小さくする．ただし，これらの先行研究ではパケットの出力ポートを動的に予測することは

行っていない．

並列アルゴリズムの通信パターンには時間的，空間的な局所性が存在することが多い[3]．例えば，NAS 並列ベンチマーク[5]は比較的高い空間的局所性を示すことが報告されている．これらの通信の規則性が直接網におけるルータで利用できれば，通信の低遅延化に活用できると考えられる．そこで，我々は通信特性に基づいた動的予測ルーティングを提案する[4]．対象とするネットワークは2D トーラスとする．通常，ルータは入力されたパケットに対してヘッダのデコードと出力候補となるポートの計算(Routing Computation; RC)，出力ポートの調停と仮想チャネルの割り当て(Virtual Channel Allocation; VA)，ルータ内クロスバスイッチの設定(Switch Allocation; SA)を順に行った後，データ転送(Switch Traversal; ST)を行う．我々の提案する動的予測ルーティングでは，ルータの入力ポートにおいて過去の出力履歴から次にどの出力ポートにパケットを出力すべきかの予測を前もって行う．パケット入力前に予測を行うことで，ヘッダのデコードからクロスバスイッチ設定までに生じる遅延を短縮することが可能となる．つまり予測が成立した場合はデータ転送の遅延しかかからない．

本論文の構成は以下の通りである．まず2 章で，我々の提案する動的予測機構を支援するルータアーキテクチャについて説明する．次に3 章ではアプリケーションプログラムでの予測性能を評価するため NAS 並列ベンチマークから得られた通信パターンを用いて，通信の局所性や規則性の有無について論じる．4 章では3 章で述べた局所性や規則性を利用する動的予測機構を検討する．5 章では動的予測機構及びその他の予測機構の予測精度を比較，考察し，動的予測機構の有効性について論じる．最後に6 章で本論文をまとめる．

2. ルータアーキテクチャ

図 1 に我々の提案する動的予測機構を支援するルータアーキテクチャを示す．これは，前章で述べた既存のルータに，予測機構，及び各入出力ポートの履歴保持機構を追加した構造を取る．2D トーラスネットワークにおいてルータは各 4 個の入出力ポート(Input / Output Port)を持つ．そして各ポートは複数本の仮想チャネルバッファを装備する．またクロスバスイッチサイズはProcessing Element(PE) I/F からの接続を含めて 5x5 である．このルータで処理されるメッセージの流れを以下に説明する．

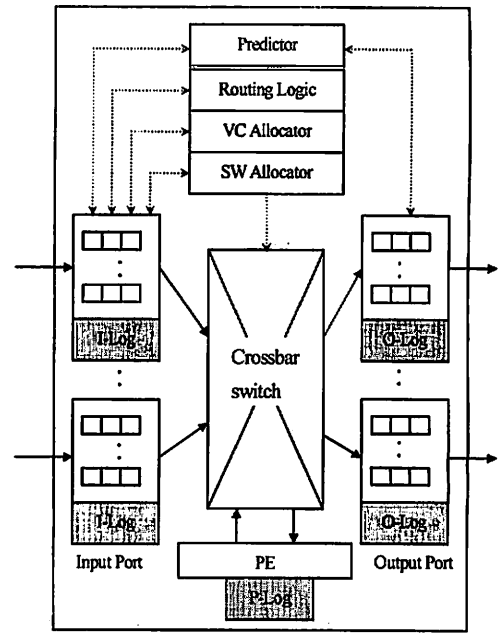

図 1 ルータアーキテクチャ

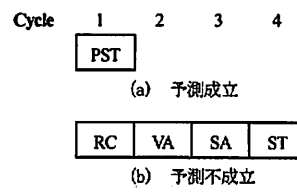

図 2 パイプラインステージ

(1) **Prediction** － 各入力ポート，PE I/F における次パケットの出力ポートを予測する．この予測はパケットがルータに入力される前に実行する．予測方法は各入力ポート，PE I/F が持つ過去の出力履歴(I-Log, P-Log)から動的予測機構(Predictor)を用いて次の予測出力ポートを決定する．I-Log，P-Log にはその入力ポート，PE I/F を使用したパケットの出力ポートが記録されている．出力ポートの予測完了後，予測した出力ポートに対して VA， SA ステージを投機的に実行する．

(2) **Input Buffering** － 入力パケットをバッファ(仮想チャネル)に一時保存する．

(3) **Predictive Switch Traversal (PST) / Routing Computation (RC)** － 投機的に VA，SA ステージが完了しているため，すぐに予測した出力ポートへ

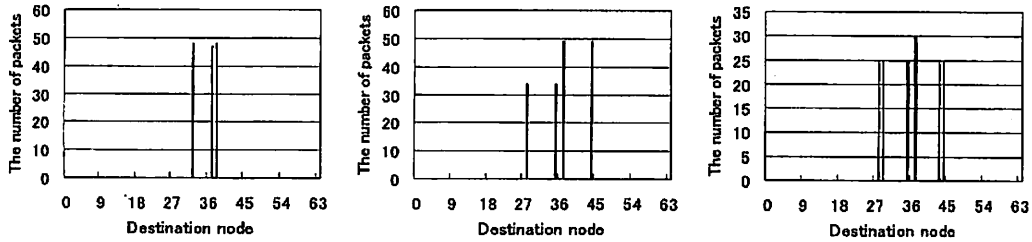

図3 36番ノードの宛先ノード

のパケット転送を開始する．但し，予測不成立の場合の保証としてバッファに残存するパケットヘッダを利用して正しい出力ポートを決定するRCステージを実行する．

(4) **Virtual Channel Allocation (VA)** ― 出力仮想チャネルに対してパケットの割り当て，調停を行う．

(5) **Switch Allocation (SA)** ― クロスバスイッチ設定を行う．

(6) **Switch Traversal (ST)** ― 出力ポートへパケットを転送する．

図2(a) に示す通り，パケットヘッダ到着後，予測が成立した場合に必要なステージはPSTだけとなる．一方，**図2(b)** に示すように予測不成立の場合は，通常の4ステージでルーティングが行われる．PSTとRCステージは同時に実行されるが，予測不成立の場合はPSTステージで転送開始したデータを破棄しなければならない．並列チャネル転送を行う多くのルータは，隣接ルータでデータをラッチするための信号をデータと共に送出するが，予測不成立の場合はこのラッチ信号をアクティブにしなければよい．シリアルチャネル転送を行うルータでは，予測の不成立は誤った隣接ルータへのデータ伝播(デッドフリット)を発生する．デッドフリットの処理には，予測ルーティングを行わないルータによる破棄や制御パケットによる破棄が必要になるが，本論文では扱わないこととする．

以上の仮定から，予測が成立すれば通常のルータ内パイプライン処理より3サイクル分だけパケットホップにかかる時間を短縮することが可能となる．

この予測処理は，各入力ポート及びPE I/Fそれぞれが独立して行う．そのため，複数の入力ポートが同じ出力ポートを予測する競合が起こる可能性がある．この解決策の1つとして，出力ポート側からの予測が考えられる．入力ポート同様に出力ポートにも過去の履歴(O-Log)を持たせる．そして出力ポート側でも次にどの入力ポートからパケットが転送されるのかを予測する．入出力ポートのマッチングを考慮したクロスバスイッチ設定を行うことで，上記のような予測ポートの競合を防ぐことができる．

3. 通信の局所性，規則性

本章ではNAS並列ベンチマーク(NPB)から得られた通信トレースを例に取り，通信の局所性及び規則性について考察する．以降，NPBのParallel Kernel BenchmarksからConjugate Gradient(CG)，Multi-Grid solver(MG)，Parallel CFD Application BenchmarksからLower-Upper diagonal(LU)，Scalar Pentadiagonal(SP)，Block Tridiagonal solver(BT)の計5つのベンチマークプログラムを使用する(プログラムサイズはW)．各プログラムともメッセージ数を1万個として解析を行う．ノード数は64個で，各ノードにノード番号として0～63番を割り当てる．

図3 はそれぞれCG，LU，BTで36番ノードがパケットを送出する宛先ノード番号(横軸)とそのノードに送られるパケット数(縦軸)を示したものである．例えば，**図3(a)** は32，37，38番の3ノードを宛先として，それぞれ約50個のパケットを送出することを示している．他のプログラムにおいても，特定の送受信ノードペアが通信することが確認できる．つまり，NAS並列ベンチマークから得られた通信には宛先ノードの空間的局所性が存在する．

次に通信の規則性について議論する．CGにおける36番ノードに着目する．36番ノードがパケットを送出する宛先ノードは32，37，38番ノードのみで，32→38→37→37→38→32→32→38→37 という繰り返しパターンでパケット転送を行う．同様にLUにおける36番ノードがパケットを送る宛先ノードは28，35，37，44番ノードで，37→35→44→28を2回繰り返した後，37→44を31回繰り返し，35→28を31回繰り返すというパターンで通信を行う．BTにおける36番ノードでは37→35→44→28→43→29の後に37を7回，35を7回，44を7回，28を7回，43を7回，29

を7回繰り返すというパターンでパケットを転送する．このようにいずれのベンチマークにおいても特定の送信元ノードに着目すると，ある規則的なパターンで宛先ノードへパケットを転送していることが分かる．

このような通信の局所性や規則性の存在は，我々の提案する動的予測ルーティングに対して有望といえる．

4. 予測機構

本章では我々の提案する動的予測機構を説明する．また提案手法との比較を行うための予測機構として直前ポートマッチング予測機構，静的直進予測機構についても述べる．

4.1. 動的予測機構

前章で通信の空間的局所性とノードごとの通信パターン，規則性を確認した．そこで我々は動的予測機構として Sampled Pattern Matching (SPM) [6]を使用する．SPM のアルゴリズムは以下の通りである．過去の出力ポートの履歴を $X_1^n = X_1, X_2, \cdots, X_n$ とし，次の予測すべき出力ポートを X_{n+1} と仮定する．まず，出力履歴 X_1^n から $X_i^n\ (1 \le i \le n)$ と一致するパターンを見つけ出す．さらに $X_{i\min}^n$，$i\min = \min\{ i \mid X_i^n = X_j^{j+n-i}, 1 \le j \le n \}$ (最長一致パターン)を求める．この時の X_j^{j+n-i} の次の出力ポート(予測出力ポート候補)の頻度を計算する．そして頻度が最も高い出力ポート $X_{j+n-i+1}$ を次の予測出力ポート X_{n+1} とする．但し，X_1^n が存在しない時は入力ポートに対する直進先ポートを，一致するパターンが存在しない時は前回の出力ポート X_n を X_{n+1} とする．以上のアルゴリズムを具体的な例で説明する．

DDDACDABDAC AABCD ABDD AABCD

この履歴は X_1^{25} で，左から順に X_1，$X_2, \cdots$ である．この例では一致パターンは"D"，"CD"，"BCD"，"ABCD"，"AABCD" であるので，最長一致パターンは "AABCD" となる．よってSPMでは，その次の出力ポート "A" を予測出力ポート X_{26} とする．このように SPM は過去の出力ポートのパターンに規則性が存在すれば効果的な予測手法といえる．

さらに我々は SPM の予測効果を上げるために改良した SPM-1 を考案した．SPM-1 は予測出力ポート候補が複数ある場合，1つの候補の頻度が極めて高い時のみ予測を行う．具体的には全ての予測出力ポート候補の頻度の平均値を求め，その平均値を超える予測出力ポート候補の頻度が1つである時のみ予測を実行する．

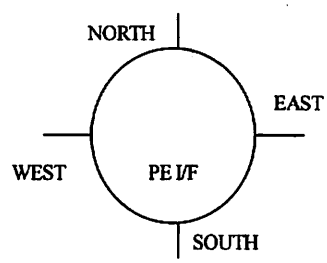

図4 ノードにおけるポート

SPM-1 は SPM より予測精度が高いものの，予測を行う条件を絞るため予測実行回数が減少する．よって SPM-1 には予測精度と予測回数とのトレードオフが存在する．

4.2. 直前ポートマッチング予測機構

直前ポートマッチング予測機構 (Latest port Matching predictor ; LM) は直前の出力ポートを予測出力ポートとする．前節で示した履歴を例に取ると，LM は直前の出力ポート "D" を予測出力ポートとする．LM の利点は入力ポートごとに保持する出力履歴を最低限(直前のポートのみ)に抑えることができる．また，空間的局所性が非常に高い通信では前回と同じ出力ポートを使用する可能性が高いと考えられ，その点で LM は有効である．

4.3. 静的直進予測機構

過去の履歴に応じて予測を行う動的予測に対して，常に同じ予測を行う単純な予測手法を静的予測と呼ぶ．各ノードで図4のように NORTH, EAST, SOUTH, WEST, PE I/F からメッセージの入力(注入)がある．静的直進予測機構 (Static Straight predictor ; SS) では入力ポートに対して直進先のポートを常に予測出力ポートとする．例えば，NORTH ポートに入力されたパケットは SOUTH ポートに出力されると予測し，WEST ポートに入力されたパケットは EAST ポートに出力されると予測する．PE I/F からネットワークに注入されるパケットに関しては予測を行わない．SS は過去の出力履歴を保持する必要がなく低コストで実装でき，予測実行時間も短いという利点がある．

5. 実 験

5.1. 条 件

ネットワークは8x8サイズ(計64ノード)の2D トーラスネットワークを用いる．ルーティングアルゴリズムは X-Y 次元順ルーティング (DOR)，完全適応ルーティング (Duato) を用いる．入力ポートごとに DOR では2本，Duato

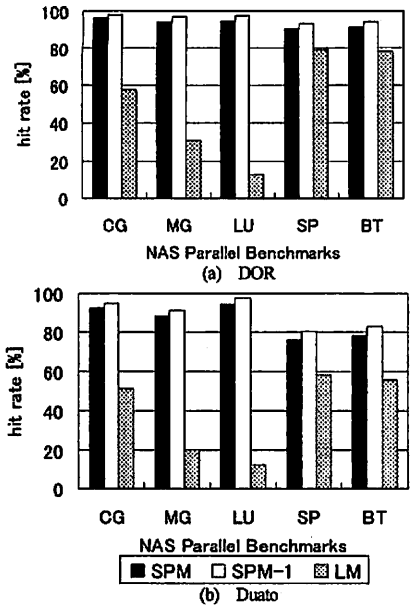

図5 PE I/F の平均予測精度

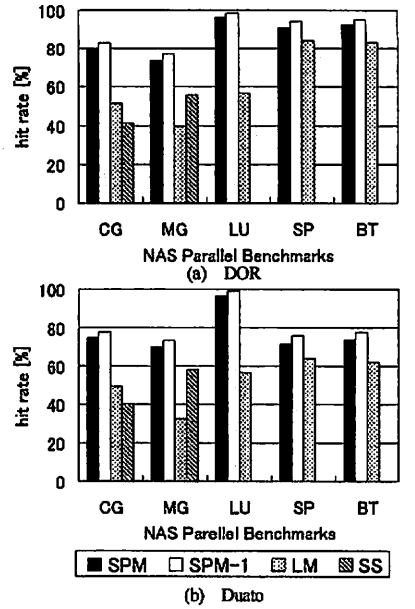

図6 全入力ポートの平均予測精度

では3本の仮想チャネルを使用する．実際のアプリケーションプログラムでの予測効果を示すためにNPB(CG, MG, LU, SP, BT)から得られた通信パターンを使用する．予測機構は4章で説明したSPM, SPM-1, LM, SSを用いる．パケットサイズは32フリット固定とし，これはいずれの予測機構においても予測を実行するのに十分な長さとする．つまりパケットが入力される前までに必ず予測は完了する．またここでは予測の競合は考慮しない．パケット転送方式はVirtual Cut-Through方式を採用する．

次節で論じる予測精度はNPBを用いてDOR, Duatoそれぞれで評価する．ネットワークへの注入パケット数を約1万2千，通信負荷を約0.2 [flit/node/cycle]として通信シミュレーションを行い，全てのPE I/F(64ポート)及び全入力ポート(5x64ポート)での平均予測精度で評価する．

5.2. 予測精度の評価

図5, 6にNPBにおける各予測機構の平均予測精度を示す．縦軸は予測精度(hit rate [%])，横軸はNPBプログラム(CG, MG, LU, SP, BT)を示しており，図5がPE I/Fにおける平均予測精度，図6がPE I/Fを含む全入力ポートにおける平均予測精度である．両図ともに(a)はDOR, (b)はDuatoでの予測精度を示す．

まずPE I/Fにおける予測精度について議論する．3章で確認したように送信元ノードは特定の複数ノードを宛先ノードとして決まった順序でパケットをネットワークに注入する．すなわちPE I/Fでの予測は通信パターンの有無に非常に影響を受ける．図5でDORとDuatoを比較すると，ほとんどのベンチマークプログラムでDuatoの予測精度が低い．これはDuatoのルーティング選択の自由度が高く，パケット注入を行うPE I/Fにおいてでさえもルーティング経路が適応的になりやすく，通信のパターンが崩れやすいことが原因である．DORのSPMはどのプログラムでも90%を超える予測精度を示す．過去の出力履歴から予測を行うSPMはDORにおいて非常に有効である．さらにSPM-1は予測が当たりやすい限られた条件でのみ予測を実行することで非常に高い予測精度を示している．一方，LMの予測精度はSPMに劣る．特にCG, MG, LUで低い予測精度を示す．これらのベンチマークは宛先ノードの繰り返しパターンは存在するが，連続して同じ宛先ノードにメッセージを送ることが少ない．そのため，前回と同じ出力ポートを予測するLMでは良い性能が期待できない．SP, BTは連続して同じ宛先ノードにメッセージを送ることが多いため，80%程度と他のベンチマークに比べ予測精度は高い．Duatoに関しても同様な傾向が見られる．

次に全入力ポートにおける平均予測精度について述べる．図6ではSPM, SPM-1, LMに加えてSSの評価も行う．DORのSPM, SPM-1は一番低いMGでは約70%，LUでは95%以上と高い予測精度を示す．いずれのベンチマークにおいてもSPM-1, SPMが最も高い予測精度に達する．

NPB	Avg. hop	
CG	2.53	1, 2, 4, 6, 8-hop を含む通信
MG	2.66	1, 2, 3, 4, 5-hop を含む通信
LU	1	隣接ノード(1-hop)通信のみ
SP	1.30	1, 2-hop を含む通信
BT	1.31	

表1 NPBによる通信ホップ数

NPB	DOR [%]	Duato [%]
CG	88	88
MG	87	86
LU	96	96
SP	93	86
BT	94	87

表2 SPMに対するSPM-1の予測実行割合

CG, MGに関してSPMは他のベンチマークに比べて低い. 表1にNPBによる平均ホップ数(Avg. hop)とその特徴を示す. CG, MGともに他のベンチマークより平均ホップ数が長く, メッセージによって様々なホップ数が存在する通信である. これはパケットのポート競合が他のベンチマークと比較して多いことを意味する. この競合の存在がポートでのSPM予測を困難にさせる原因だと考えられる. 反対に全ノードが隣接ノードとしか通信を行わないLUでは他パケットとの競合がないためPE I/Fでの予測同等の非常に高い予測精度を示す. ホップ数の短いLU, SP, BTではSSは全く機能しない. DuatoのSPM, SPM-1はDORと比べてLU以外のベンチマークで予測精度が低い. これはPE I/Fでの予測と同様にDuatoのルーティング自由度のためである. LUは1ホップ通信のみで, DORでもDuatoでもルーティング経路に差異がないため同等の予測精度を発揮する.

図5, 6よりSPMは効果的な予測機構であることが確認できた. さらにSPM-1は予測精度の面では非常に高い効果が期待できる. ただここで考えなければならないのはSPM-1の予測精度と予測実行数とのトレードオフについてである. SPM-1は1つの予測出力ポート候補の頻度が高い時のみ予測を実行するため, SPMに対して予測実行の回数が制限される. 表2にSPMに対するSPM-1の予測実行回数の割合を示す. 図6と表2の関係から分かるようにSPM-1で予測が成立する総数はSPMよりも少ない. これは今回のようなノード内で予測不成立のメッセージ破棄が実行できる状況ではSPMの方が効果的であることを意味している. しかし, デッドフリットが次ノードに出力されてしまう状況を考える場合, より予測精度の高い機構が望まれるためSPM-1の優位性は高い.

6. まとめと今後の課題

本論文では, 2Dトーラスネットワークにおいて入力ポートごとに次のパケットのための出力ポートを動的に予測する手法を提案した. この手法の優位性は通信に規則性及び局所性が存在することを前提としている点にある. また予測が成立すればルータあたり3サイクル分の遅延抑制が実現でき, データ転送のみの最小遅延時間でルーティングを行うことが可能である. 予測不成立の場合でも通信性能におけるデメリットはない. 他の予測機構と比較しても非常に効果的である.

今後の課題として, ポートの履歴長や予測実行時間によるハードウェア制約を考慮したデザインの検討などが挙げられる.

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, 18年度NII共同研究(提案型)の援助による. また本論文の予測機構について多大なる御指導, 御助言を頂きました東工大・吉瀬謙二博士, 電通大・岩田賢一博士に深く感謝し, ここに厚く御礼申し上げます.

参考文献

[1] C.Izu, R.Beivide and C.Jesshope : "Mad-Postman: a lookahead message propagation method for static bidimensional meshes", Proc. 2nd Euromicro Workshop on Parallel and Distributed Processing, pp.117-124 (1994).

[2] Li-Shiuan Peh and W.J.Dally. : "A delay model and speculative architecture for pipelined routers", Proc. of the HPCA, pp.255-266 (2001).

[3] Z.Ding, R.Hoare, A.Jones, D.Li, S.Shao, S.Tung, J.Zheng and R.Melhem : "Switch Design to Enable Predictive Multiplexed Switching in Multiprocessor Networks", Proc. of the IPDPS, p.100a (2005).

[4] 鎌倉正司郎, 西村康彦, 吉永努, 鯉渕道紘 : "2Dトーラスネットワークにおける通信方向予測ルーティング", 情報処理学会第68回全国大会論文集(1), pp.127-128 (2006).

[5] D.Bailey, T.Harris, W.Saphir, R.Wijngaart, A.Woo and M.Yarrow : The NAS Parallel Benchmarks 2.0, *NAS Technical Reports, NAS-95-020* (1995).

[6] 吉瀬謙二, 岩田賢一 : "分岐予測の精度と履歴情報との関係について", 電子情報通信学会2005年基礎・境界ソサイエティ大会講演論文集, A-1-25, p.25 (2005).