適応ルータの出力チャネル選択における優先次元指定の効果

吉永 努[†] 林 匡哉[†] 堀田真貴[†]
山口喜教[††] 大津金光[†] 馬場敬信[†]

本論文では，メッセージごとに適応ルーティング時に優先する出力チャネルやFIFO性保証のための適応ルーティングの禁止を指定可能なルーティングを提案する．また，そのハードウェアコストと性能をハードウェア記述言語により設計したルータの論理合成とシミュレーションによって評価する．その結果，優先次元指定ルーティングは，(1)バーチャルチャネルの追加と比較して，小さなコストでサポートできる，(2)ホットスポットを形成する通信パターンのスループットを大きく改善する，(3)適応性が制限可能であることは，FIFO性の保証のみでなく，ユニフォーム通信におけるネットワークの負荷を均等に保つことに対しても有効活用できる，などを明らかにした．

Prior-dimension Specification on Output Channel Selection for Adaptive Routers

Tsutomu Yoshinaga,[†] Masaya Hayashi,[†] Maki Horita,[†]
Yoshinori Yamaguchi,[††] Kanemitsu Ootsu[†]
and Takanobu Baba[†]

We propose a new adaptive routing method which is capable of selecting, based on a prioritizing system, a particular dimension to output each message. We have compared its hardware cost and performance based on HDL designs. The results of HDL synthesis and simulation lead to the following conclusions: (1) The dimension-selective routing can be supported inexpensively compared with adding virtual channels; (2) Adaptive routers which support communication scheduling are effective in improving network performance; (3) The ability to restrict adaptive routing is useful not only in maintaining in-order message delivery but also balancing the overall network load for uniform communication traffic.

1. はじめに

高並列計算機の通信スループットを制限する要因として，従来は，プロセッサ側で通信処理を行うためのソフトウェア的なオーバヘッドが問題にされてきた．しかし，(a)要素プロセッサとして使用されるマイクロプロセッサの高性能化，(b)ソフトウェア実装技術の発達，(c)ネットワーク・インタフェースの改良，などにより，ソフトウェア・オーバヘッドの要因がある程度解消されつつある．これにともなって，プロセッサ間のネットワークに関して，相対的にルータを介したネットワーク内での情報転送のスループット自体が問題となる場合が生じている[6]．

ネットワーク上でのメッセージ衝突によるスループットの低下を抑える技術として，適応ルーティングが提案されている．しかし，実際の高並列計算機で適応ルーティングを採用したものは比較的少ない．これには，いくつかの理由があげられる．

(1) 適応ルーティングのために専用のバーチャルチャネル[9]やデッドロックバッファ[3]が必要となり，ハードウェアが高価になる．

(2) 経路選択の柔軟性により，1ホップ時間が長くなる．

(3) アプリケーションに必要なメッセージのFIFO性が失われる．

(1)に関しては，ターンモデル（Turn model）[13]のように必ずしも適応ルーティング専用のバーチャルチャネル（VC: Virtual Channel）やバッファを必要としないアルゴリズムが存在する．ターンモデルは，チャネル依存グラフ[7]からサイクルをなくすという点で次元順（dimension-order）ルーティングと共通であり，次元順ルーティングを採用してきた多くのネッ

† 宇都宮大学工学部
Faculty of Engineering, Utsunomiya University
†† 電子技術総合研究所
Electrotechnical Laboratory

トワークトポロジーに適用可能である．では，その場合ハードウェア量と性能はどのような影響を受けるのか．本論文では，ハードウェア記述言語で設計した実際のルータ回路に基づいて，より現実的なハードウェア量と動作速度を明らかにする．

メッセージのFIFO性に関しては，従来，適応ルーティングと相反するように議論されることが多かった．しかし，FIFO性を保証する必要のあるメッセージのみ非適応ルーティングをし，他のメッセージについては適応ルーティング可能であれば，両者の利点を享受できる．そこで，メッセージごとに適応ルーティングするか，しないか，また，適応ルーティングするときに優先する出力チャネルを指定可能なルーティングを提案する．また，そのハードウェアコストと通信性能について議論する．

2. ルータの基本構成

ここでは，本研究で使用するルータの基本構成についてまとめる．

2.1 ネットワークトポロジー

高並列計算機には，2次元，または3次元のメッシュやトーラスを採用しているものが多い．最近では，ASCI Red[4]，T3E[21]，AP3000[22]などがこれにあたる．ルーティングアルゴリズムは，k-ary n-cube（n次元kトーラス）ネットワーク上で検討されることが多いが，実際には2次元や3次元の実装となっていることが分かる．トーラスのラップアラウンド・チャネルによって発生するチャネル依存グラフのサイクルは，一般にVCを導入することによって解消する．このことは，ターンモデルについても同様である．

次元数によって影響を受けるのは，ポート数やチャネル幅ばかりでなく，ポート間結線，またはルータ内のクロスバスイッチ，調停回路などがある．適応ルーティングを採用した場合には，ポート数が多いほど，また，ポートあたりのVC数が多いほど経路選択に自由度が大きくなり，ハードウェアが複雑になる．本論文では，上記すべてのネットワークの基本となる2次元メッシュについて検討する．

2.2 フロー制御

近年の高並列計算機では，ワームホール[7]方式やバーチャルカットスルー方式[16]が使用されている．ワームホール方式は，フロー制御に必要な最小限のバッファ容量で大きなメッセージを扱えることに特徴がある．バーチャルカットスルー方式は，ブロックされたメッセージを単一ルータ内のバッファに吸収できることに特徴があるが，チャネルあたりのバッファを大きくするか，メッセージを小さな単位に分割して送受信する必要がある．最近の研究によると，多くのアプリケーションで数十から数百バイトのメッセージが必要であると指摘するものがある[19]．この場合には，ワームホール方式の方がコスト/パフォーマンスが良い．

したがって，本研究ではワームホール方式を基本とする．ただし，適応ルーティングの特長であるネットワーク全体の高スループット性を生かすため，メッセージサイズはアプリケーション側から制御可能な形態とする．すなわち，ルータとしては任意長のメッセージを扱える構成とし，メッセージの終端は最終フリット（flit: flow control unit）に同期した信号によって判断する．プロセッサは，大きなメッセージを1つのパケットで送信することも，複数に分割して送信することもできる．単に，最終フリットにタグビットをセットすればよい．

2.3 適応ルーティング

適応ルーティングとして，ターンモデルを基本とした複数のアルゴリズムについて考察する．ターンモデルには，チャネル依存グラフからサイクルをなくすために，どのターンを禁止するかによって，いくつかの組合せが存在する．初めに，最も簡単な構成として，2次元メッシュの方向を東西南北としたときに，北進するメッセージのターンを禁止するNorth-lastアルゴリズムを取りあげる．North-lastルータは，VCを使用せずに実現できるが，メッセージのあて先ノードが送信ノードより北にあると適応ルーティングできない部分的適応型（partially adaptive）となる．

次に，X次元の物理チャネルあたり2本のVCを用いて，あて先が北の場合North-lastに，南の場合South-lastに割り当てる．これにより完全適応（fully adaptive）ルーティングが可能となる．このルータをDouble-xと呼ぶ[14]．Double-xは，Y次元にVCを持たないため，適応性を制限した次元順ルータとして動作する場合にY次元のチャネルがボトルネックとなるおそれがある．そこで，Y次元にもVCを2本使用し，Double-x完全適応ルータとVC 2本の次元順ルータの双方の動作が可能なルータについても考察する．これをDouble-xyルータと呼ぶ．

なお，ターンモデルでは，最短経路ルーティングと迂回を許す非最短経路ルーティングの双方を構成可能であるが，本論文で扱うルータは，ライブロック制御の不要な最短経路ルータとする．

2.4 優先次元指定ルーティング

適応ルーティングで出力可能チャネルが複数あった場合，どのチャネルを選択するかについては，静的に

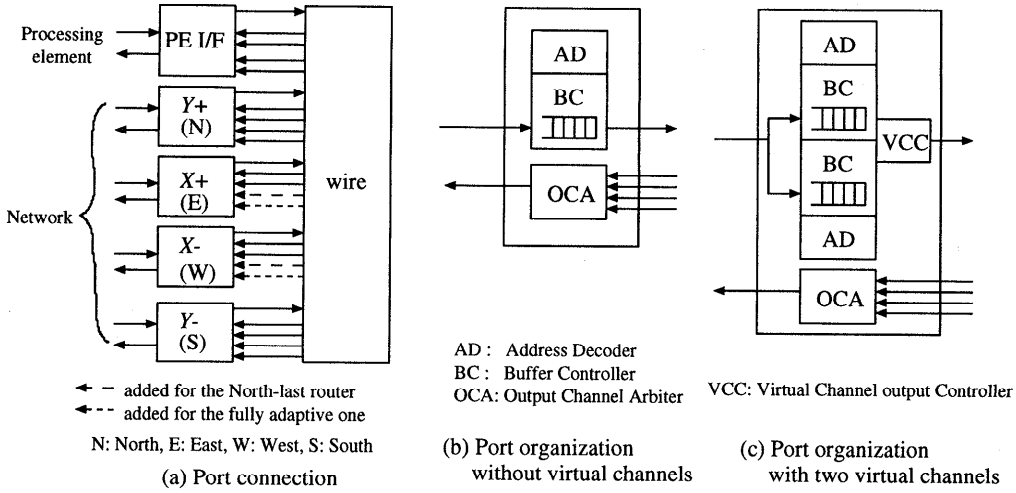


図1 ルータの構成
Fig. 1 A router organization.

優先順位を設ける方式と動的に優先順位を変更する方式，ランダムに選択する方式などが考えられる[2]．従来，経路選択はルータの実装に頼るのみで，アプリケーション側から静的な情報を与えることはあまり検討されていない．しかし，安易に適応ルーティングを実装しても，十分な効果が得られないことがある[24]．そこで，我々はメッセージごとに優先する出力チャネルの方向を指定可能な優先次元指定（dimension-selective）ルーティングを提案する．優先次元指定ルーティングでは，適応ルーティングしてよいメッセージに対して，優先的に選択する出力チャネルの方向を指定できる．たとえば，2次元メッシュにおいてX次元優先を指定すれば，そのメッセージは，適応ルーティングしながらできるだけX方向へ進もうとする．これにより，アプリケーションの通信パターンが既知であれば，より空いている方向へ一部のメッセージをルーティングすることで，ホットスポットを回避することが期待できる．また，適応ルーティングを許さないX-Y次元順のルーティングを指定すれば，FIFO性の保証が可能である．

2.5 ハードウェア構成

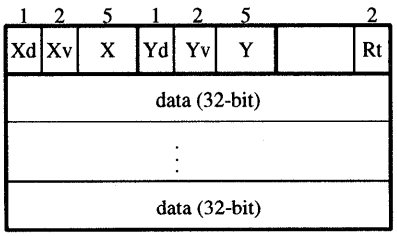
図1(a)に2次元メッシュ・ルータの構成を示す．ここでは，2次元をX，Yで表し，各次元の方向を +，− で表現している．2次元の各方向に相当する4つのネットワーク・ポート（$Y+, X+, Y-, X-$）とプロセッサ側のインタフェース・ポート（PE I/F），およびそれらの結線（wire）からなる．ネットワーク・ポートには，東西南北の方向を () 内に示した．また，結線部分をブロックのように図示したのは図を簡略化するためであり，この部分は純粋に結線のみである．ポートとwire間の実線は，次元（X-Y）順ルーティングで必要な結線を示している．North-lastアルゴリズムのみ，またはNorth-lastとSouth-lastの両方をサポートするためには，それぞれ図中に破線で示した結線が必要となる．

図1(b)は，VCを持たない場合のポートの構成を，また(c)は物理チャネルあたり2本のVCを持つ場合の構成を示す．各ブロックの機能と構成は以下のとおりである．

Buffer Controller（BC）： メッセージのバッファリングを制御する．大規模，かつ，高速なネットワークに対応するため，ネットワークから入力するメッセージはルータ内のクロックと非同期にFIFOに受信する．受信クロックは，メッセージを構成する各フリットとともにネットワーク経由で受信する．

Address Decoder（AD）： メッセージのアドレスをデコードして，出力候補となるポートに出力要求を行う．適応ルーティングの場合は，すべての出力候補に同時に要求を出し，出力許可を返したポートから1つを選択する．出力ポートの選択方針としては，静的，または動的な優先度を使用する．

Output Channel Arbiter（OCA）： ADからのメッセージ出力要求を調停し，物理チャネル，および隣接ノードの接続するポート内の受信バッファ（VC）の利用可能状態に応じた出力制御を行う．各ポートからのデータを物理チャネルに出



Xd, Yd: X and Y directions
Xv, Yv: X and Y virtual channel select
X, Y: X and Y absolute addresses
Rt: Additional routing information

図2 メッセージの構造
Fig. 2 Structure of a message.

力するセレクタの大きさは，経路選択アルゴリズムによって，図1(a)に実線と破線で示した数に依存する．

Virtual Channel output Controller（VCC）： 図1(c)において，OCAからの出力許可信号を受けて，バーチャルチャネルの出力を制御する．ルータ内の結線数とOCA内のセレクタを節約するために，同一ポート内のVCからの出力は，VCCでマルチプレックスする[8]．

2.6 メッセージの構造

図2にメッセージの構造を示す．1フリットは32ビットとし，ヘッダ・フリットに経路選択に必要な情報を保持する．2フリット目以降がデータである．メッセージサイズは，1フリット（データサイズ0バイト）以上の任意サイズをサポートする．メッセージ・ヘッダの構成と設計方針を以下に示す．

宛先ノードアドレス（X, Y）： X, Y各次元ごとに，絶対アドレスを使用する．絶対アドレス形式は，ホップカウント・ベースの相対アドレスに対して，デクリメントなどの書き換えが不要な分，高速に実装できる．

方向ビット（X_d, Y_d）： 次元ごとに1ビットで方向を指示することにより，デコード時の大小比較や表検索を不要にする．

VC選択（X_v, Y_v）： 調停回路を簡単化するため，静的なVC選択を用いる．必要ビット数はVC数に依存する．

経路選択補助（Rt）： 適応ルーティングをするか，しないか，また，適応ルーティングする場合に経路の優先次元を指定することを可能にする．

3. 基本性能

2.3節で述べた構成のルータをVerilog-HDLによって記述した．すべてのルータにおいて，1ホップに要する手数は次の4ステップである．

(1) 非同期に受信したメッセージ（ヘッダ）をルータ内部のクロックで検知する．
(2) アドレスをデコードして，出力ポートに要求を出す．
(3) 出力ポートが要求を調停して出力許可を出す．
(4) 入力ポートから出力ポートにデータを転送する．

VCが増えると調停回路も複雑化するが，VC間で静的な優先順位を仮定すれば，少数のVCは上記の手順を増やさずに実装できる．また，ステップ(3)と(4)の動作はオーバラップ可能である．したがって，1ホップあたり単相クロックならば3クロック，2相クロックならば1.5クロックで動作する．この手数は，最近の高速ルータとほぼ等しい[12),21)]．なお，ルータの動作速度に関しては，クリティカルパスを2ステップに分割してある程度高速化することも可能である．しかし，我々の試みた範囲では，上記の最小クロック数で動作するHDL記述の1ホップ時間が最短であった．

表1 論理合成環境
Table 1 Synthesis tool and conditions.

シンセサイザ	Synopsys HDL Compiler Version 1998.02
動作条件	民生用最悪条件
配線負荷	セル面積による自動選択
マッピング最適化	Medium effort
ライブラリ	LSI LOGIC 0.6-micron array-based gate array

3.1 ハードウェアコスト

ハードウェアコストの見積りは，論理合成ツールとターゲット・ライブラリに大きく依存するため，クロック制約条件を除き，同一の条件で合成した結果について考察する．使用したシンセサイザとその条件を**表1**に示す．なお，今回はルータ内をクロックの立ち上がりと立ち下がりの両エッジで動作するように設計した．したがって，回路のクリティカルパスが半クロック時間以下となるものがタイミング条件を満たすものとして，動作クロック速度を求めた．また，Verilog-HDLソースプログラムについては，適宜シンセサイザへのディレクティブを記述するなどして，タイミングマージンが大きくなるように配慮した．

3.1.1 次元順ルータ

表2に，次元順ルータのVC数とVC割当て方法を変えたときのハードウェアコストを示す．VC割当てについては，送信ノードにおいて静的に割り当てたVCを中継ノードで変更しないものと，各ルータで動的に空きVCを割り当てるものについて調べた．なお，

表2 次元順ルータのハードウェアコスト
Table 2 Synthesis results for the dimension-order routers.

FIFO/ポート VC 割当て	4 x 1 静的	4 x 2 静的	4 x 2 動的	4 x 4 静的
最大動作周波数（MHz）	98.2	72.4	69.4	68.9
クリティカルパス（ns）	5.09	6.90	7.20	7.25
セル面積（K gates）	13.1	28.6	28.8	53.8
配線領域（K gates）	7.5	16.4	16.5	32.1
総面積（K gates）	20.6	45.0	45.3	85.9
信号線数	280	296	296	320

各ポートの物理チャネルは幅32ビット×2（入出力），FIFOの深さは4フリット分（32×4ビット）とした．

表中のクリティカルパスと面積（ゲート数換算値）は，タイミング条件を満たす最大動作周波数をクロック条件に指定したときの値を示している．VCを導入すると，最大動作周波数が小さくなることが分かる．これは，VCC（図1(c)参照）が増えること，出力チャネルの調停ロジックが複雑化すること，などによる．ただし，VCなしからVC 2本にした場合に比べ，VC 2本から4本に増やした場合の速度低下は小さい．また，動的なVC割当ても動作速度に影響するが，その影響もVCの追加と比較すれば小さい．

セル面積は，バッファ容量に比例して増加する．ターゲットに指定したゲートアレイでは，配線領域にセル面積の半数強が必要であることが分かる．紙面の関係でセル面積の内訳は示していないが，バッファ（FIFO）の面積が支配的である．予備評価では，セル面積，配線領域ともにバッファ容量（深さ，または幅）にほぼ比例した値を示した．

信号線数は，ルータどうしの結線に必要なピン数を表す．VC数に従い，その選択信号とフロー制御信号が増加している．

3.1.2 適応ルータ

表3に，以下に示す適応ルータのハードウェアコストを示す．

(1) North-last：全メッセージをNorth-lastアルゴリズムで経路選択する．
(2) North-last/dimension-selective：(1)で優先次元指定ルーティングを指定できる．
(3) Double-x：X次元のポートのみにVCを2本使用し，全メッセージをあて先によってNorth-lastかSouth-lastに区別して経路選択する．
(4) Double-x/dimension-selective：(3)で優先次元指定ルーティングを指定できる．
(5) Double-xy：Y次元ポートにもVCを2本使用し，Double-xとVC 2本の次元順ルータの両方をエミュレートできる．
(6) Double-xy/dimension-selective：(5)で優先次元指定ルーティングを指定できる．

dimension-selectiveオプションが指定可能なルータについては，FIFO性を保証する次元順ルーティングと，出力可能チャネルが複数あるときの優先次元を指定できる．特に指定がなければ，X次元を優先する．

3.1.3 次元順ルータとの比較

表3の結果を表2に示した次元順ルータと比較すると，VCなしのとき，North-lastルータのクリティカルパスは0.31 ns伸び，最大動作周波数は5.7 MHz遅くなっている．また，総面積は約5.5%増加する．しかし，これらの影響は，次元順ルータにVCを追加する場合に比べると小さいといえる．

Double-xルータは，VC 2本の次元順ルータとほぼ同一のクロック周波数で動作する．Double-xルータは，Y次元ポート（南北）にVCを持たないが，VCを有するX次元（東西）ポートのロジックがクリティカルパスとなるためである．また，Double-xルータの面積は，バッファ容量を反映して，VCなしとVC 2本の次元順ルータの中間的な値となっている．したがって，ハードウェアコストとしては，完全適応ルーティングよりもVCの影響が大きいことが分かる．

なお，適応ルータのチップ面積は，Y次元ポートからX次元ポートへの結線，X次元ポート内のスイッチの増大などにより増加する（図1(a)参照）．また，回路のクリティカルパスは，経路選択ロジックの複雑さばかりでなく，回路面積が増大することによる配線負荷の増加によっても長くなる．

3.1.4 適応ルータ内での比較

表3から，優先次元指定ルーティングの追加（表中DS）は，最大動作周波数と回路面積に大きな影響を与えないことが分かる．実際，Verilog-HDLによる記述では，アドレスデコーダ部分にRtビットの条件を追加するだけでよい．

部分適応型のNorth-lastと完全適応型のDouble-xを比較すると，クリティカルパスの伸びにともなう最大動作周波数の違いは大きいが，Double-xとDouble-xyの最大動作周波数の違いは比較的小さい．これもVCの追加による影響が大きいためである．また，回路面積と信号線数もVC数に応じて増加している．ただし，表3に示した数万ゲート規模の回路面積の違いは，近年の集積回路技術の発展によって大きな制約とならない場合が多い．また，信号線の増加もVC数n本の片方向ポートに対して，n本のフロー制御信号と$\log n$本のVC選択信号の増加にとどまる．したがっ

表3 適応ルータのハードウェアコスト
Table 3 Synthesis results for the adaptive routers.

経路選択	NL	NL/DS	DX	DX/DS	DXY	DXY/DS
FIFO/X 次元	4 x 1	4 x 1	4 x 2	4 x 2	4 x 2	4 x 2
FIFO/Y 次元	4 x 1	4 x 1	4 x 1	4 x 1	4 x 2	4 x 2
FIFO/PE I/F	4 x 1	4 x 1	4 x 2	4 x 2	4 x 2	4 x 2
最大動作周波数（MHz）	92.5	92.5	70.9	68.4	68.0	68.0
クリティカルパス（ns）	5.40	5.40	7.05	7.30	7.35	7.35
セル面積（K gates）	13.7	13.8	23.7	23.8	29.8	29.9
配線領域（K gates）	8.0	8.1	14.3	14.3	18.0	18.1
総面積（K gates）	21.7	21.9	38.0	38.1	47.8	48.0
信号線数	280	280	288	288	296	296

NL: North-last DS: Dimension-selective DX: Double-x DXY: Double-xy

て，ターゲットチップのゲート数やピン数が許容する範囲であれば，Double-xy 適応ルータは，小さな動作速度のオーバヘッドで次元順ルーティングを積極活用することが可能である．

4. シミュレーション

以上に述べたルータが，衝突をともなう通信に対してどのような性能を示すか，Verilog-HDL RTL（Register Transfer Level）記述をシミュレートした．シミュレータには，Cadence 社の Verilog-XL を使用した．実験結果は，すべてのルータが論理合成において，タイミング条件を満たす 66 MHz のクロックの場合と，各ルータごとに表 2，3 に示した最大動作周波数を用いた場合のネットワークのバンド幅で示す．

シミュレーション条件として，ルータ間のデータ転送遅延は，各ルータの動作速度の半クロック（表 2，3 中のクリティカルパスに相当）以内と仮定した．したがって，3 章に述べた 2 相クロック動作時のルータ内遅延 1.5 クロックと合わせて，1 ホップにメッセージヘッダは 2 クロックを要する．また，VC を有する次元順ルータにおける静的な VC 割当てについては，転送距離や送信順に対する複数の割当て法を試みた中で最良の結果を示す．

なお，本論文では Matrix-transpose 通信と All-to-all 通信の 2 つの結果を示すが，他の通信パターンのシミュレーション結果については文献 15) を参照されたい．

4.1 Matrix-transpose 通信

ここでは，5×5 の 25 ノード 2 次元メッシュにおいて，ノードアドレス (i,j) と (j,i) がメッセージをピンポンする．シミュレーションでは，$i \neq j$ の 20 ノードが同時にメッセージの送信を開始した時点から，20 ノードすべてが 4 個のメッセージの受信が完了するまでの時間を計測し，バンド幅を求めた．各ルータに対して，メッセージ長を変えたときの結果を**図 3** に示す．

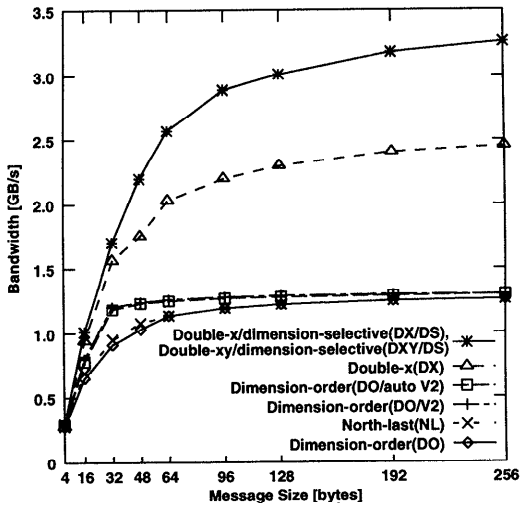
図 3 (a) 66 MHz 動作時は，優先次元指定付きの完全適応型（DX/DS，DXY/DS）の結果が最も良い．これは，実行全体を通してのクリティカルパスとなるノード対 (4,0)–(0,4) 間のメッセージを Y 次元優先に指定して，他のメッセージとの衝突を回避できるからである．すると (3,0)–(0,3) 間のメッセージがクリティカルパスとなるが，さらに，(1,0)–(0,1) 間のメッセージを Y 次元優先にして (3,0)–(0,3) 間のメッセージの経路から外すことで，無衝突時と同じピーク性能を達成する．優先次元指定のない完全適応型 DX は，次元順 DO や部分適応型 NL の性能を大きく改善するが，通信パターンに対して最適なスケジューリングを行う優先次元指定付きの完全適応型（DX/DS，DXY/DS）には及ばない．また，部分適応型 NL は，次元順 DO の結果をほとんど改善していない．これは，クリティカルパスになるノード (0,4) から (4,0) へのメッセージを適応ルーティングできずに，次元順に送るためである．

VC 付きの次元順ルータについてみると，VC 割付けに関して静的な DO/V2 と動的切替え可能な DO/auto V2 に大きな差はみられない．これは，この通信パターンでは，VC の動的切替えがほとんど行われないためである．

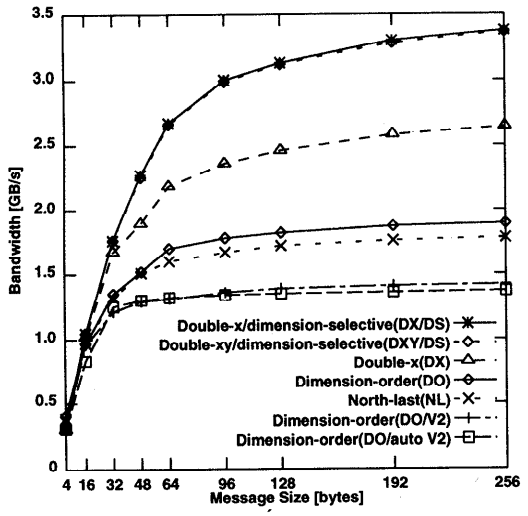
図 3 (b) 最大動作周波数のときは，動作速度の高い VC を持たないルータ（DO，NL）と，完全適応型（DX，DX/DS，DXY/DS）との性能差は小さくなっている．しかし，これらのルータ間で性能が逆転するまでには至っていない．なお，メッセージのピンポン回数を増やした場合にもバンド幅の性能順位に変動はないが，その差が大きくなる傾向を示す．これは，ホットスポットに弱いルータの性能低下が大きくなるためである．

4.2 All-to-all 通信

All-to-all 通信では，5×5 の 25 ノード 2 次元メッシュにおいて，全ノードが自分以外のノードに 1 つず

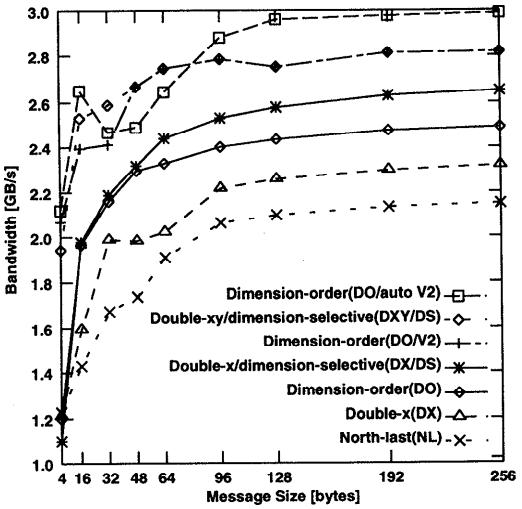


(a) 66 MHz 動作のとき
(DX/DS と DXY/DS のバンド幅は同値となる)

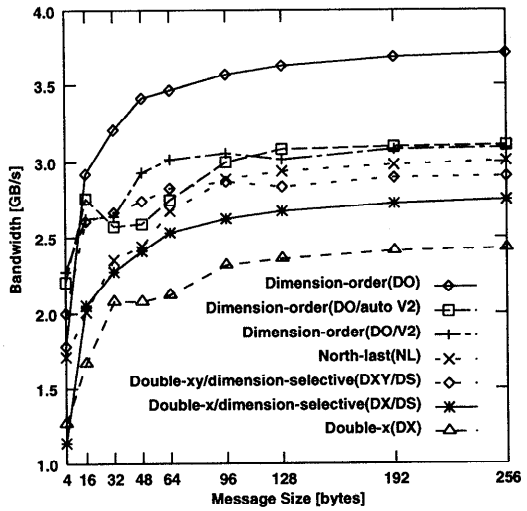


(b) 最大動作周波数のとき

図 3 Matrix-transpose 通信のバンド幅
Fig. 3 Bandwidth for the matrix-transpose traffic.



(a) 66 MHz 動作のとき



(b) 最大動作周波数のとき

図 4 All-to-all 通信のバンド幅
Fig. 4 Bandwidth for the all-to-all traffic.

つ連続的にメッセージを送信する．シミュレーションでは，25 ノードすべてが同時にメッセージの送信を開始した時点から，全ノードが 24 個のメッセージすべてを受信するまでの時間を計測し，バンド幅を求めた．結果を図 4 に示す．

All-to-all 通信では，一様にネットワーク全体が混雑するため，適応ルーティングを行うと一様であったネットワークの負荷に偏りを生じ，スループットが低下する．したがって，図 4 (a) 66 MHz 動作時には，適応ルーティングを行う NL や DX の結果が最も悪い．

一方，VC を有する次元順ルータ DO/auto V2 と DO/V2 は良好な結果を示す．このことは，4.1 節の結果と対照的であり，VC の効果も通信パターンと各メッセージへの VC の割当て方法に依存することを表している．

優先次元指定ルーティング DXY/DS では，全メッセージに次元順ルーティングを指定することにより，DO/V2 と同一のサイクル数で処理を完了する．しかし，DX/DS は Y 次元に VC を持たないため，DO/V2 と DO の中間的な性能となっている．

図 4 (b) 最大動作周波数のときは，ルーティング・アルゴリズムよりも動作周波数による影響を大きく受けている．そのため，高速に動作可能な VC なしの次

元順ルータDOが最も良い結果を示す．このことは，適応ルーティングやVCによる小さな必要サイクル数の減少は，より簡単なハードウェア構成を持つルータの動作速度の向上によって相殺されてしまうことを示している．

以上の結果から，ルータの設計においては，目標とする動作速度を満たしたうえで，適応ルーティングやVCによるスループットの向上について考えるべきであるといえる．また，アプリケーションの通信パターンがある程度既知であれば，最適化コンパイラやユーザ指定によって通信も最適化できることが分かる．

5. 関連研究

村上ら[20]は，実装を想定した正確な性能評価を行うために，並列計算機ネットワーク用ルータ・チップの自動設計システムの開発を行っている．

Chien[5]は，ルータのハードウェア量を機能別にモデル化し，特定のテクノロジのゲート遅延を基に各種ルータの動作速度を推定した．Chienの遅延モデルを用いた適応ルータと非適応ルータの性能比較も報告されている[10),17]．しかし，Chienの示したルータチップの遅延時間は，最新のルータと比較しても非常に小さい．また，文献10), 17)では，ルータ内の遅延よりもルータ間のデータ転送時間が長い場合，ルータのクロック間隔をルータ間遅延時間としている．しかし，近年のルータは，ルータ間のデータ転送速度と独立なチップ内クロックを採用するものが多い[23]．

我々はHDLを用いて設計したルータの論理合成結果から，より現実的なルータの動作速度を求め，より実際的なシミュレーション条件を用いて評価した．また，本論文では，適応ルーティングにおける経路選択の優先順位やFIFO性保証オプションを追加する優先次元指定ルーティングを提案し，3.1.4項と4章でそのハードウェアコストと通信性能を評価した．メッセージごとに配送経路を指定可能なルーティングとしては，iWarpのストリートサイン方式がある[1]．iWarpでは，メッセージヘッダに送信方向とターンするノード番号とあて先を指定する．これに対して，優先次元指定ルーティングでは，適応ルーティング時に出力可能チャネルが複数あった場合，どちらを優先するかをフラグによって指定する．

適応ルータにおいて，FIFO性保証のための非適応ルーティングをサポートするものにTriplexルータ[11]やハイブリッドルータ[18]がある．ただし，これらでは，適応ルーティング時の優先次元の指定については議論していない．また，本論文で示したDouble-xyルータは，FIFO性保証のためばかりでなく，ユニフォーム転送時のネットワークの負荷バランスによっては，適応ルータの一機能として次元順ルーティングを積極的に活用すべきであるという観点を示した．

6. まとめ

本論文では，優先次元指定ルーティングを提案した．また，優先次元指定ルータを含むいくつかの並列計算機ルータをHDL設計し，それらのハードウェアコストと通信性能を比較した．その結果，次のようなことが分かった．

(1) 適応ルーティングのためよりも，バーチャルチャネル（VC）がハードウェア量と動作速度に大きく影響する．これに対して，優先次元指定ルーティングは，ベースとする適応ルータに小さなコストで追加できる．

(2) 優先次元指定ルーティングは，定型的な通信パターンを持つアプリケーションに対して通信スケジューリングの可能性を開く．特に，ホットスポットを形成する通信パターンのスループットを大きく改善できる．

(3) ネットワークの通信性能は，ルータの動作周波数とルーティングアルゴリズムの通信パターンへの適応性の双方の影響を受ける．優先次元指定ルーティングは，ベースとする適応ルータの動作速度をほとんど落とすことなく，通信パターンへの適応性を増す．

現在，我々は，優先次元指定ルーティングをDuatoの適応ルーティング[9]やDISHAタイプのデッドロックリカバリールーティングアルゴリズム[3]と組み合わせた場合のHDLによる評価を行っている．今後の課題として，2次元メッシュ以外のネットワークトポロジーやアルゴリズム横断的なルータの評価があげられる．また，将来的には，プロセッサモデルを実験環境に取り込み，実アプリケーションによる通信スケジューリングの効果についても明らかにする必要がある．

謝辞 本研究に関してご指導いただいた電子技術総合研究所の大蒔和仁情報アーキテクチャ部長，並列アーキテクチャラボ，実時間システムラボの皆様に感謝いたします．また，本研究の一部は東京大学大規模集積システム設計教育研究センターより提供していただいたCADツールを使用しています．深く感謝いたします．

本研究は，一部文部省科学研究費基盤研究（C）課題番号09680324，基盤研究（B）課題番号10558039，奨励研究（A）課題番号09780237の援助による．

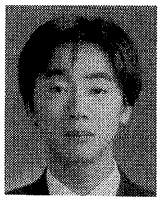
参考文献

1) Borkar, S., et al.: Supporting Systolic and Memory Communication in iWarp, *Proc. 17th ISCA*, pp.70–81 (1990).
2) Bolding, K., Fulgham, M. and Snyder, L.: The Case for Caotic Adaptive Routing, Technical Report, CSE-94-02-04, University of Washington (1994).
3) Anjan, K.V. and Pinkston, T.M.: DISHA: A Deadlock Recovery Scheme for Fully Adaptive Routing, *Proc. 22nd ISCA*, pp.201–210 (1995).
4) Carbonaro, J. and Verhoorn, F.: Cavallino: The Teraflops Router and NIC, *Proc. Hot Interconnects IV*, pp.157–160 (1996).
5) Chien, A.A.: A Cost and Speed Model for k-ary n-cube Wormhole Routers, *Proc. Hot Interconnects* (1993).
6) Dai, D. and Panda, D.K.: How Much Network Contention Affect Distributed Shared Memory Performance, *Proc. ICPP '97*, pp.454–461 (1997).
7) Dally, W.J. and Seiz, C.L.: Deadlock-Free Message Routing in Multiprocessor Interconnection Network, *IEEE Trans. Comut.*, Vol.C-36, No.5, pp.547–533 (1987).
8) Dally, W.J.: Virtual-Channel Flow Control, *Proc. 17th ISCA*, pp.60–68 (1990).
9) 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).
10) Duato, J. and López, P.: Performance Evaluation of Adaptive Routing Algorithms for k-ary n-cubes, *Proc. PCRCW*, Lecture Notes in Computer Science, Vol.853, pp.45–59 (1994).
11) Fulgham, M.L. and Snyder, L.: Triplex Router: A Versatile Torus Routing Algorithm, Techinical Report, UW-CSE-96-01-11, University of Washington (1996).
12) Galles, M.: SPIDER: A High-Speed Network Interconnect, *IEEE Micro*, Vol.17, No.1, pp.34–39 (1997).
13) Glass, C.J. and Ni, L.M.: The Turn Model for Adaptive Routing, *Proc. 19th ISCA*, pp.278–287 (1992).
14) Glass, C.J. and Ni, L.M.: Maximally Fully Adaptive Routing in 2D Meshes, *Proc. 1992 ICPP*, pp.I-101–I-104 (1992).
15) 林 匡哉，堀田真貴，大津金光，吉永 努，馬場敬信：HDL 設計に基づく並列計算機ルータの評価，情報処理学会研究報告，98-ARC-130-13, pp.79–84 (1998).
16) Kermani, P. and Kleinrock, L.: Virtual Cut-Through: A New Computer Communication Switching Technique, *Computer Networks*, Vol.3, No.4, pp.267–286 (1979).
17) Millar, D.R. and Najjar, W.A: Empirical Evaluation of Deterministic and Adaptive Routing with Constant-Area Routers, *Proc. PACT'97*, pp.64–75 (1997).
18) Millar, D.R. and Najjar, W.A: Preliminary Evaluation of a Hybrid Deterministic/Adaptive Router, *Proc. PCRCW '97*, Lecture Notes in Computer Science, Vol.1417, pp.21–32 (1997).
19) Mukherjee, S.S. and Hill, M.D.: The Impact of Data Transfer and Buffering Alternatives on Network Interface Design, *Proc. 4th HPCA* (1998).
20) 村上祥基，朴 泰祐：並列計算機ネットワーク用ルータ・チップの自動設計システム，情報処理学会研究報告，97-ARC-127-1, pp.1–8 (1997).
21) Scott, S.L. and Thorson, G.M.: The T3E Network: Adaptive Routing in a High Performance 3D Torus, *Hot Interconnects IV*, pp.147–156 (1996).
22) Shiraki, O., Nagatsuka, M., Horie, T., Koyanagi, Y., Shimizu, T. and Ishihata, H.: AP-Net: Advanced High-Performance Network for Scalable Parallel Server, *Proc. Hot Interconnects IV*, pp.19–28 (1996).
23) Stunkel, C.B.: Challenges in the Design of Contemporary Routers, *Proc. PCRCW '97*, Lecture Notes in Computer Science, Vol.1417, pp.21–32 (1997).
24) 吉永 努，山口喜教：適応ルータのコスト/パフォーマンス，並列処理シンポジウム JSPP'98 論文集，pp.55–62 (1998).

(平成 10 年 8 月 31 日受付)
(平成 11 年 2 月 8 日採録)

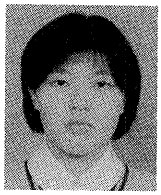
吉永 努（正会員）

1986 年宇都宮大学工学部情報工学科卒業．1988 年同大学大学院修士課程修了．同年より宇都宮大学工学部助手．博士（工学）．1997 年から翌年にかけて電子技術総合研究所客員研究員．並列計算機アーキテクチャ，リコンフィギュラブル・コンピューティング等に興味を持つ．電子情報通信学会会員．



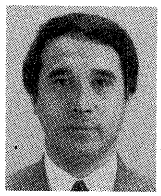
林 匡哉

1998年宇都宮大学工学部情報工学科卒業．現在同大学大学院博士前期課程在学中．並列計算機アーキテクチャ，特に，ルーティングアルゴリズムに興味を持つ．



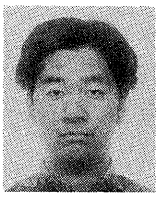
堀田 真貴

1999年宇都宮大学工学部情報工学科卒業．現在同大学大学院博士前期課程在学中．ハードウェア設計，特に，並列計算機ルータに興味を持つ．



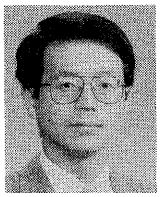
山口 喜教（正会員）

1972年東京大学工学部電子工学科卒業．同年通商産業省工業技術院電子技術総合研究所入所，計算機方式研究室長等を経て，1999年筑波大学電子・情報工学系教授（電子技術総合研究所併任），工学博士．高級言語計算機，並列計算機アーキテクチャ等の研究に従事．1991年情報処理学会論文賞，1995年市村学術賞受賞．著書「データ駆動型並列計算機」（共著，オーム社）．IEEE Computer Society，ACM，電子情報通信学会各会員．



大津 金光（正会員）

1993年東京大学理学部情報科学科卒業．1995年同大学大学院修士課程修了．1997年同大学大学院博士課程退学，同年より宇都宮大学工学部助手となり現在に至る．理学修士．高性能計算機システム，特にマイクロプロセッサアーキテクチャに興味を持つ．



馬場 敬信（正会員）

1970年京都大学工学部数理工学科卒業．1975年同大学大学院博士課程単位取得退学．同年より電気通信大学助手，講師を経て，現在宇都宮大学工学部教授．工学博士．1982年より1年間メリーランド大学客員教授．計算機アーキテクチャ，並列処理等の研究に従事．電子情報通信学会，IEEE各会員．1992年情報処理学会Best Author賞．著書「Microprogrammable Parallel Computer」（MIT Press），「コンピュータアーキテクチャ」（オーム社）等．