Recover-x 適応ルーティング

吉永 努[†] 林 匡哉[†] 堀田真貴[†]
中村さゆり[†] 大津金光[†] 馬場敬信[†]

適応ルーティングとデッドロック回復に使用するバーチャルチャネル数のバランスを考慮したRecover-xルーティングを提案する．Recover-xルータは，デッドロック回復の対象メッセージを通信経路に応じて効率的に分類することで，比較的少ないバーチャルチャネル数で構成可能である．また，並列，かつ，オーバヘッドの少ないデッドロック回復をサポートする．本論文では，Recover-xと他のいくつかのルータをハードウェア記述言語により設計し，その論理合成とシミュレーション結果に基づいた評価を行う．その結果，Recover-xは，他の適応ルータに比べて高速動作が可能であり，低レイテンシ，高バンド幅の通信性能を達成することを示す．

Recover-x Adaptive Routing

Tsutomu Yoshinaga,[†] Masaya Hayashi,[†] Maki Horita,[†]
Sayuri Nakamura,[†] Kanemitsu Ootsu[†] and Takanobu Baba[†]

A new fully adaptive routing, Recover-x, is proposed. Recover-x balances the cost and bandwidth between adaptive and non-adaptive virtual channels. The router is organized with a moderate number of virtual channels by effectively classifying deadlock recovery messages. It supports concurrent and low-overhead deadlock recovery. In this paper, we compare its hardware cost and performance with other routers based on HDL designs. Experimental results show that the Recover-x router attains a fast operating speed and low-latency, with high-bandwidth communication performance.

1. はじめに

並列計算機ネットワークにおける基本的，かつ重要な設計項目としてデッドロック回避技術があげられる[11]．Dallyらは，チャネル依存グラフにサイクルが存在しなければデッドロックが発生しないことを示した[5]．Dimension-orderルーティングはその典型であり，多くの並列計算機に採用されている．しかし，Dimension-orderルーティングにおけるチャネル利用順序に関する制約は，ホットスポットを形成する場合などの通信性能を損なう原因ともなる．

この問題への1つの対処法として，適応ルーティングが研究されている．Duatoは，サイクリックなチャネル依存関係が存在しても，メッセージがサイクルから退避できるパスを用意すればデッドロックを防止できることを証明した[6]．この考え方に基づく適応ルーティングとして，∗-channel[3]やDISHA[1]が提案されている．これらの適応ルータは，バーチャルチャネル（VC）やデッドロックバッファ（DB）と呼ばれるハードウェア資源をデッドロックからの退避パス用に装備する．ここで設計者が考慮すべき項目として，適応ルーティングとデッドロック回避に利用するハードウェア資源のバランスをどうするかという問題がある．∗-channelは，全ネットワークポートに退避パス用のVCを装備するため，適応ルーティングの自由度が限定される．一方，DISHAは退避パス用のDBを最少限とすることで適応ルーティングの自由度を大きくできるが，デッドロック回復のオーバヘッドが大きくなる．一般に，VC数はルータの動作速度やハードウェア量に影響を与えるため，適度なVC数で効率的なルーティングを実現するアルゴリズムが必要となる[4]．

本論文では，実装に必要なハードウェア量が比較的小さく，並列デッドロック回復が可能であり，かつ，デッドロック回復時のオーバヘッドが少ない完全適応ルーティングRecover-xを提案する．また，ハードウェア記述言語（HDL）によるLSI設計手法を活用

† 宇都宮大学工学部
Faculty of Engineering, Utsunomiya University

した並列計算機ルータのコストと性能に関する評価を示す．それにより，Recover-x ルータが他の適応ルータに比べて高速動作可能であり，良好な通信性能を達成することを示す．

2. Recover-x ルーティング

2.1 基本アルゴリズム

Recover-x ルーティングの着目点は，デッドロックサイクルから退避するメッセージの候補を必要な集合に限定し，それに見合う退避パスを装備することにある．全メッセージを退避候補とする必要がなければ，∗-channel のように全ポートに退避パス用の VC を装備したり，DISHA のように DB を全ポート共有の資源にする必要はない．また退避パスを DB のように複数ポート共有の資源にすると，デッドロック回復時にポート間の調停が必要となり，オーバヘッドが大きくなる．

そこで，必要なネットワークポートのみに退避用 VC を持たせ，適応ルーティングに利用可能な VC 数を十分に確保しつつ，退避用 VC の使用コストを小さく抑える．退避用 VC では，デッドロックを防止するために非適応ルーティングを行う．そこで，適応ルーティングに使用する VC を適応 VC，デッドロックサイクルからの退避用 VC を非適応 VC と呼ぶ．∗-channel では，非適応 VC に退避されたメッセージは再び適応 VC に戻ることができる．しかし，我々の予備評価によると，VC 切替えの組合せ数はルーティングロジックの複雑さに影響し，ルータの動作速度を低下させる[8]．そこで，Recover-x では非適応 VC から適応 VC へのメッセージ切替えは行わない．ただし，デッドロック回復を効率的に行うため，並列デッドロック回復はサポートする[2]．

2.2 2 次元トーラス用アルゴリズム

2 次元トーラス用 Recover-x ルータでは，一方の次元のネットワークポートに少なくとも 1 本の適応 VC と 2 本の非適応 VC を持ち，もう一方の次元のポートには 2 本以上の適応 VC を装備する．仮に，前者を X 次元，後者を Y 次元とする．この場合，非適応 VC への退避候補は，Y 次元のルーティングを完了して X 次元のみに進む必要のあるメッセージと，はじめから Y 次元方向に進む必要がないメッセージに限定できる．このことは，次節で証明する．全メッセージは，はじめ各ポートの適応 VC を用いて最短経路で適応ルーティングするが，退避候補メッセージがあらかじめ設定した時間以上ブロッキングされた場合，非適応 VC に退避する．非適応 VC 中のメッセージはあて先ノードまで非適応 VC を通してルーティングする．これをデッドロック回復と呼ぶ．

本方式では，デッドロック回復メッセージを X 次元の 2 本の非適応 VC にトーラスサイクル[12]によってデッドロックしないように割り当てることで，並列デッドロック回復が可能となる．また，X 次元方向へ直進するメッセージのみをデッドロック回復し，Y 次元ポートは適応 VC のみで構成するため，ルーティング自由度を比較的高くできる．

すべてのメッセージについて適応ルーティング可能な方式を完全適応ルーティングと呼び，その中で VC に関しても自由に選択できる方式を真の完全適応（true fully adaptive）ルーティングと呼ぶことがある．Recover-x は，完全適応であるが，非適応 VC とトーラスサイクル防止のための VC 割付け制約によって真の完全適応とはならない．ただし，前節に述べたように VC 切替えの組合せ数はルータの動作速度に影響するため，必ずしも不利とは限らない．

2.2.1 デッドロックフリーの証明

Recover-x ルーティングが，デッドロックフリーであることを証明する．

定理 1： Y 次元方向に進もうとするメッセージだけからなるサイクル（Y 次元トーラスサイクル）は，Y 次元ポートの 2 本の適応 VC を使用すればデッドロックしない．

証明： VC を切り替えるための基準線を Y 次元に設け，それを dateline と呼ぶ．最短経路ルーティングでは，メッセージが dateline を最大 1 回しか通過しない．したがって，**図 1** 左に示すように (1) dateline を越えるメッセージと越えないメッセージで使用する VC を区別する方法，または (2) メッセージが dateline を越えるときに VC を変更する方法により，2 本の VC でトーラスサイクルのデッドロックを防止できる[7]．また，トーラスサイクルが，図 1 右に示すように X 方

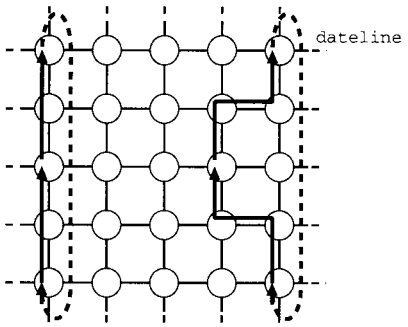


図 1 トーラスサイクル
Fig. 1 Torus cycles.

向に広がりを持っていたとしても，(1) または (2) の方法により Y 次元の 2 本の VC を使い分ければ，Y 次元トーラスサイクルとしてデッドロックすることはない．

定理 2：X 次元のトーラスサイクルは，X 次元ポートに 2 本の非適応 VC と 1 本の適応 VC があればデッドロックしない．

証明：2 本の非適応 VC は，Y 次元と同様に dateline によって使い分ける．適応 VC でトーラスサイクルが発生した場合，サイクルを構成するメッセージのタイムアウトを検出して非適応 VC へ退避する．これにより，サイクルは解消される．

定理 3：非適応 VC への退避候補メッセージを含まないサイクルはデッドロックしない．

証明：非適応 VC への退避候補メッセージを含まないサイクルがデッドロックしたと仮定する．定理 1 と 2 より，このサイクルはトーラスサイクルでないから，必ず X と Y の両方向へ進もうとするメッセージ，すなわち X と Y 両次元のルーティングが完了していないメッセージを含む．これらのメッセージの Y 方向がブロックされていることは，その方向に別のデッドロックサイクルが存在することを意味する．このようなデッドロックサイクルの連鎖に退避候補メッセージが含まれない状態は，その連鎖の一部がトーラスサイクルを形成してデッドロックしていることにほかならない．なぜならば，非適応 VC への退避候補にならないメッセージ群の Y 次元方向のサイクルを追っていけば，Y 次元方向に進もうとするメッセージからなるサイクル，すなわち Y 次元のトーラスサイクルがデッドロックしていることになる．しかし，これは定理 1 に矛盾する．よって仮定は誤りであり，非適応 VC への退避候補メッセージを含まないサイクルはデッドロックしない．したがって，デッドロックするサイクルには必ず退避候補メッセージが含まれる．

例として，**図 2** に 2 つのメッセージサイクルの連鎖の様子を示す．ここで，メッセージ M は右上方向にルーティングする必要があるが，右，上双方ともブロックされている．上方向をブロックするメッセージ N を含むサイクルに非適応 VC への退避候補メッセージが存在すれば，そのサイクルはデッドロックしないから，有限時間内に N のブロックは解消され，M も上方向に進むことができる．一方，N を含むサイクルに非適応 VC への退避候補メッセージが存在しない場合は，そのサイクル中で上方向へ進もうとするメッセージがブロックされていることになる．このように上（Y 次元）方向のサイクルを追っていくと，Y 次元のトーラスサイクルに帰着する．なぜならば，上方向へ進む必要がなく右（X 次元）方向へしかルーティングされないメッセージは，非適応 VC への退避候補となるからである．しかしトーラスサイクルはデッドロックしないことが保証されているので，結局 N を含むサイクルは解消され，M のブロックも解消される．

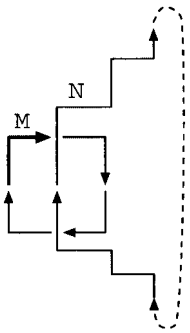


図 2 メッセージサイクルの連鎖
Fig. 2 Multiple cycles of messages.

表 1 ルータのバーチャルチャネル数
Table 1 The number of virtual channels for each router.

Router	X	Y	PE I/F	DB	計
Dimension-order	4	4	2	0	18
∗-channel	4	4	2	0	18
DISHA	4	4	1	1	18
Recover-x	4	4	2	0	18

計 = 2·X + 2·Y + PE I/F + DB

2.3 3 次元トーラスへの応用

Recover-x を 3 次元トーラスへ応用する場合について考察する．ここで，3 次元トーラスを X-Y，Y-Z，Z-X 平面に分割して考える．X，Y，Z の各次元のポートのうち，2 つの次元について非適応 VC があれば，デッドロック回復が可能である．たとえば，X と Z ポートに非適応 VC を持つ場合，X-Y 平面，Y-Z 平面のデッドロックは，それぞれ X 方向と Z 方向へ進むメッセージを非適応 VC へ退避することでデッドロック回復すればよい．また，Z-X 平面は，少なくとも X または Z 次元へ進むメッセージのどちらかをデッドロック回復対象にすれば問題ない．

3. ルータの構成

本論文では，ネットワークトポロジーとして 2 次元トーラスを取り上げ，**表 1** に示す 4 つのルータについてアルゴリズム横断的にコスト/パフォーマンスを比較する．

3.1 バーチャルチャネル

表 1 中の X と Y は各次元のネットワークポートを，

PE I/F は PE インタフェースを表し，数字は各ルータにおけるポートごとの VC 数を示している．なお，ポート X と Y は，各ノードに 2 つずつあるので，VC 数の合計は，2·X + 2·Y + PE I/F + DB となる．

これらのルータにおける VC 数の最少構成はそれぞれ異なるが，各ネットワークポートあたり 3 本以上の VC を装備すればすべてのルータが構成可能である．本実験では，実用規模のルータにおける公平な比較として，ネットワークポートの VC 数を 4 本に統一する[14]．DISHA では，メッセージがネットワークに過剰に出力されると，デッドロックが発生し性能が低下する[10]．そこで，メッセージの出力を抑制するため，ネットワークへの出口となる PE I/F の VC 数を 1 とする．DISHA 以外のルータでは，DB が必要ないので PE I/F の VC 数を 2 とし，全ルータの VC 数を 18 に統一する．なお，トーラスサイクルに対しては，メッセージが dateline を越えるか越えないかによって使用する VC を区別する方法を採用する．

以下に，各ルータにおけるメッセージに対する VC 切替えの制約条件を説明する．

(1) Dimension-order

2 次元トーラス用 Dimension-order ルータ（X-Y 順）で許容される VC 切替えを**図 3** に示す．図中の PE I/F は送信ノードを表し，図中央の X と Y はそれぞれ隣接ノードの X，または Y 次元の入力ポート（以降，それぞれ X，Y ポート）を表す．

X ポートと Y ポートの 4 つの VC は，VC0，2 と VC1，3 の 2 組に分け，それぞれ各次元において dateline を越えるメッセージと越えないメッセージが使用する．図右側の X から Y へのチャネル依存は，中継ノードの X ポートから出力されるメッセージが次ノードの Y ポートに格納される場合に可能な VC 切替えの様子を表している．ここで，X ポートから X ポートなど次元が変化しないメッセージは，VC 切替えを行わないため省略してある．また，X-Y 順ルーティングでは，Y ポートから X ポートへの切替えは発生しない．

(2) ∗-channel

図 4 に ∗-channel で可能な VC 切替えを示す．影なしと影付きの VC は，それぞれ適応 VC と非適応 VC を表す．適応 VC 中のメッセージは，VC0 と VC1 を利用して適応ルーティングする．適応 VC がどれも使用できない場合，メッセージは非適応 VC に退避する．VC2 と VC3 は，トーラスサイクルを解消するように使い分ける．非適応 VC は Dimension-order ルーティングを行うためデッドロックを起こさない．また，一度非適応 VC に入ったメッセージでも，適応 VC が使用可能になった場合，適応 VC に戻って適応ルーティングを再開することができる．

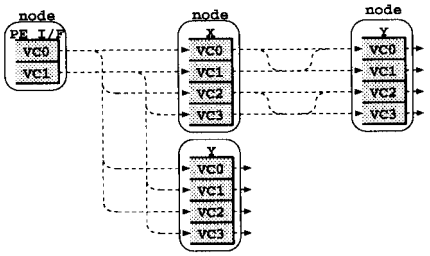


図 3 Dimension-order の VC 切替え
Fig. 3 VC assignment for a dimension-order router.

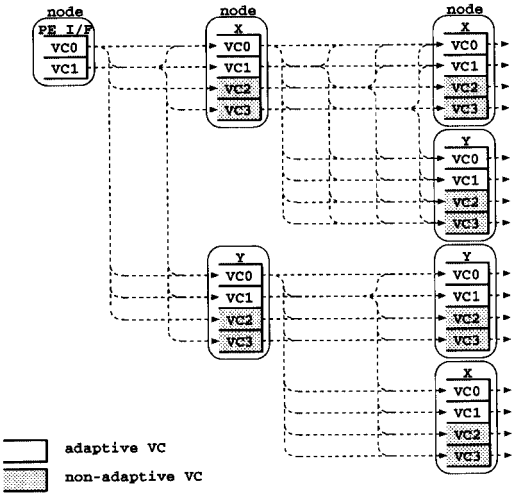


図 4 ∗-channel の VC 切替え
Fig. 4 VC assignment for a ∗-channel router.

このように，非適応 VC と適応 VC 間の切替えを多くサポートするため，ハードウェア構成は複雑になる．

(3) DISHA

DISHA は，トーラス用ルータを単一の DB によって構成可能である[1]．ただし，その場合デッドロック回復は逐次的となり，DB を使用するメッセージはネットワーク中で唯一に限定される．また，その調停方法としてはトークンをネットワーク中に巡回させる方法が提案されている．

図 5 に示すとおり，各ポートの VC はすべて適応 VC として使用することができ，VC 切替えにも制約がない．これは，真の完全適応（true fully adaptive）ルーティングと呼ばれる．メッセージの転送候補となるすべての適応 VC を使用できず，デッドロックを検出したノードがトークンを獲得したときにデッドロック回復を行う．一度 DB に入ったメッセージは，次元

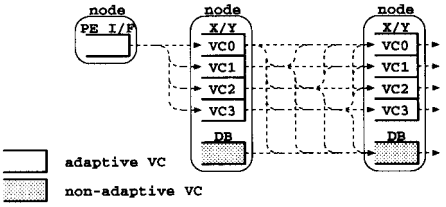


図 5 DISHA の VC 切替え
Fig. 5 VC assignment for a DISHA router.

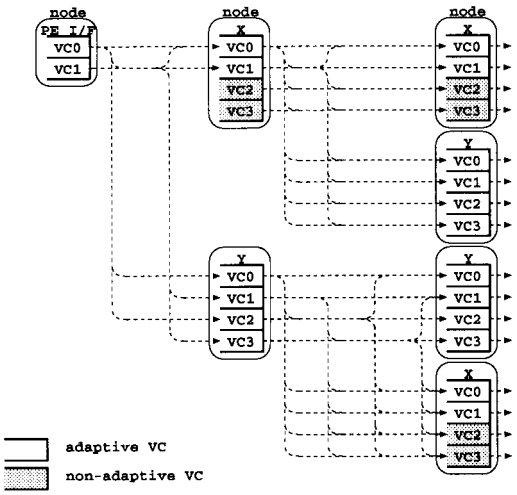


図 6 Recover-x の VC 切替え
Fig. 6 VC assignment for a Recover-x router.

順ルーティングによりあて先ノードまで配送する．

(4) Recover-x

図 6 に Recover-x の VC 切替えを示す．X ポートには適応 VC と非適応 VC を 2 本ずつ持ち，Y ポートはすべて適応 VC で構成する．∗-channel と違い，非適応 VC から適応 VC へのメッセージの移動を許さないため，PE I/F からは適応 VC にしかメッセージを出力しない．

X ポートの適応 VC からは，メッセージをすべての VC へ切り替えて転送することができる．また，X ポートの非適応 VC に退避するメッセージは Y アドレスのルーティングが完了したメッセージのみであるから，一度非適応 VC に入ったメッセージは，あて先ノードまで非適応 VC を直進すればよい．

一方，Y ポートの VC は，Dimension-order ルータと同様に dateline によって Y 次元のトーラスサイクルを解消するように 2 組に分ける．適応ルーティングするメッセージの FIFO 性は保証されないため，Y ポートから Y ポートへ直進するメッセージは，それら各組内の VC 間では自動切替えが可能である．FIFO 性を保証する場合には，Dimension-selective ルーティングをサポートすればよい[13]．

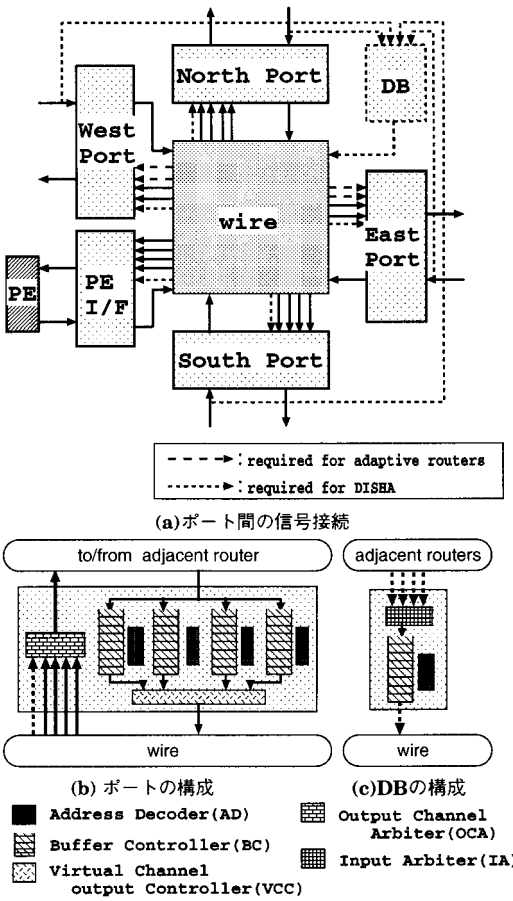


図 7 ルータのハードウェア構成
Fig. 7 Hardware organization of the routers.

3.2 ハードウェア構成

図 7 にルータのブロック図とポート，および DB の内部構成を示す．すべてのルータは，4 つのポートと PE I/F を持つ．さらに DISHA は，全ポートと結線された DB を装備する．なお，図 7 (a) では簡単化のために wire をブロックのように表しているが，実際にはこの部分は結線のみである．

ここに示すルータは，各ポートが独立して経路選択と出力チャネルの調停を行う．したがって，出力ポートが衝突しない限り，複数メッセージの経路選択，入出力ポートの接続，データ転送は，すべて並列動作可能である．各モジュールの機能を以下に説明する．

Buffer Controller（BC）： ネットワークからの入力メッセージをバッファリングし，フロー制御と経路選択要求を行う．

Address Decoder（AD）： 受信したメッセージ

表 2 論理合成結果
Table 2 Synthesis results.

Router	Dimension-order	∗-channel	DISHA	Recover-x
最大クロック速度（MHz）	156.2 (1.00)	114.9 (0.74)	100.0 (0.64)	133.3 (0.85)
セル領域（Kgates）	90.1 (1.00)	96.4 (1.07)	105.3 (1.17)	94.4 (1.05)
配線領域（Kgates）	51.6 (1.00)	60.1 (1.16)	69.8 (1.35)	58.1 (1.13)
総面積（Kgates）	141.7 (1.00)	156.5 (1.10)	175.1 (1.24)	152.5 (1.08)

括弧内の数字は Dimension-order を 1 とした場合の比

のアドレスをデコードし，出力ポートに出力要求を行う．デッドロック回復ルータでは，メッセージのブロック時間をカウントすることでデッドロックを検出する．

Output Channel Arbiter（OCA）： 物理チャネルの使用状況と隣接ノードの受信バッファの状態をもとに，各入力ポートからの出力要求を調停する．

Virtual Channel output Controller（VCC）： OCA からの出力許可に基づいて VC の出力を制御する．

Input Arbiter（IA）： デッドロック回復メッセージの入力を調停する．

4. ハードウェアコスト

4.1 論理合成条件

3 章で述べたルータを Verilog-HDL によって設計し，論理合成を行った．**表 2** に合成結果を示す．なお，論理合成条件は以下のとおりである．

シンセサイザ： Synopsys HDL Compiler Ver.1999.05

ライブラリ： LSI Logic 0.6 μm Gate Array

動作条件： 民生用最悪条件

マッピング最適化： Medium effort

配線負荷： セル面積による自動選択

ここで，動作条件とは，テクノロジライブラリで指定される製造プロセス，動作温度，供給電圧，配置配線の相互接続モデルなどのことであり，今回はライブラリに民生用最悪条件として定義されたものを使用した．また，マッピング最適化のレベルは，合成時間を考慮してデフォルトの Medium effort に統一した．そのほか，合成時にはすべてのサブ回路の境界最適化を指定した．

最大クロック速度は，論理合成結果がタイミング条件を満たす範囲で最速の値を表す．また，括弧内の数字は，Dimension-order を 1 とした場合の比を表す．なお，ルーティングロジックはクロックの立ち上がりエッジ駆動とし，メッセージヘッダの 1 ホップあたり 3 クロックを要する．ここで，回路のクリティカルパスを 2 ステップに分割することにより，ある程度動作速度の向上も可能であるが，我々の試みた範囲では上記のステップ数で動作する場合の 1 ホップ時間が最短であった．

面積は，最大クロック速度を制約条件としたときのゲート数を表す．また，Verilog-HDL ソースプログラム中では，適宜シンセサイザへのディレクティブを指定し，クリティカルパスが短くなるように配慮した．

4.2 動作速度

非適応ルータ Dimension-order とそれ以外の適応ルータでは，最大クロック速度に 15〜36%の差がある．この原因としては，次の 2 点があげられる．

第 1 に適応ルータの経路選択ロジックの複雑化にともない，AD のクリティカルパスが伸びる．第 2 にルーティングの自由度の増加にともない OCA が複雑化する．OCA の複雑さは，入出力チャネルの組合せ数と関係がある．Dimension-order では，Y ポートの OCA が各ポートから入力する VC と出力する隣接ノードにおける VC の組合せ数は 24 通り存在する．同様に，∗-channel では 48 通り，DISHA では 52 通り，Recover-x では X ポートの OCA で 46 通りが最大となる．各ルータの最大クロック速度は，この関係をよく反映しており，Dimension-order が最高速で DISHA が最も遅い．また，適応ルータ内では Recover-x の動作クロック速度が最高となっている．

なお，各ルータにおいてポートの VC 数を 3 に統一した場合にも，同様の傾向が確認できる[9]．

4.3 回路面積

適応ルータの総面積は，Dimension-order に対して 8〜24%増加する．これは，適応ルーティングのために，X ポートから Y ポートへの結線が追加されるとともに，AD や OCA のロジックが複雑化するためである．

適応ルータ間で面積を比較すると，∗-channel と Recover-x にはあまり大きな差が見られない．これはルータの面積に最も影響を与えるバッファ容量を統一しているためである．DISHA は，他と比較してや

や面積増加が顕著である．特に，配線領域の増加率が大きい．これは，VC 切替えに制約のない真の完全適応ルーティングをサポートするためである．予備評価では，DISHA においても VC 切替え数を制限すれば，面積の減少と動作速度の向上に寄与することを確認している．ただし，その場合ルーティング自由度を失うため，必ずしも通信性能が向上するとは限らない．

5. 通信性能

4 章で論理合成結果を示したルータに対して，典型的な通信パターンに対するメッセージの通信性能を示す．

5.1 シミュレーション条件

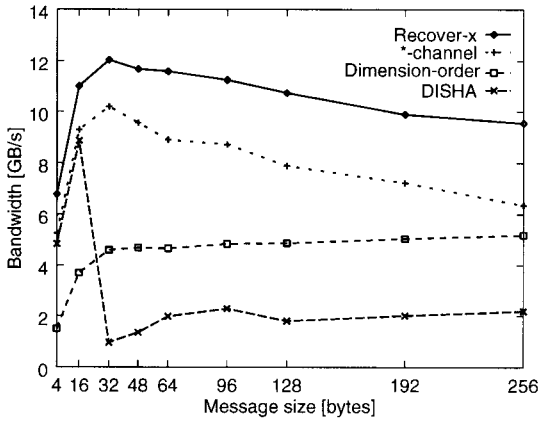
ネットワークサイズは 10×10 の 100 ノードとし，シミュレータには，Cadence 社の Verilog-XL を使用した．通信パターンには，偏りのある Hot-spot 通信とユニフォームな All-to-all 通信を用いる．

- **Hot-spot 通信：**全ノードが 100 個のメッセージを連続して送信するが，そのうちの 25%は，$0 \leq j \leq 9$ であるノードアドレス $(4, j)$ の 10 個のノードに送信する．残りのメッセージは自分以外の任意のノードに等確率で送信する．
- **All-to-all 通信：**全ノードが自ノード以外の 99 ノードに 1 つずつ（ノード n は $n+1 \rightarrow n+2 \rightarrow \cdots \rightarrow 99 \rightarrow 0 \rightarrow \cdots \rightarrow n-1$ の順）連続してメッセージを送信する．

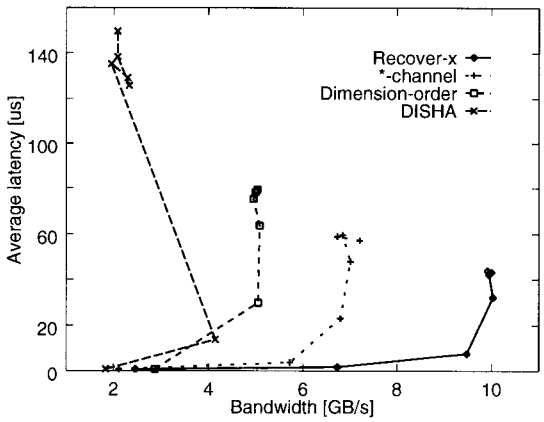
これらの通信パターンでは約 1 万メッセージを送信するが，ネットワークがある程度混雑した状態での評価を行うため，到着順で 2000〜7000 メッセージを評価対象とする．

各ルータの動作速度は表 2 に示した最大クロック速度とし，ルータ間の伝送遅延は 1 クロック時間以内と仮定する．我々の設計したルータは，メッセージヘッダの 1 ホップに 3 クロックを要するので，ノード間遅延と合計すると 1 ホップ時間は 4 クロックとなる．あて先ノードに到着したメッセージは，随時 PE に取り込まれるものとする．また，PE ではメッセージの送信と受信を並列処理可能とする．

デッドロック回復ルータが非適応 VC にメッセージを退避するまでのブロック時間は，通信性能に影響を与える．これに対しては，種々実験した中で最良の通信性能を示したクロックサイクル時間を使用する．この値は，DISHA では 256，Recover-x では 4 クロックであった．DISHA のブロック時間が長い理由は，デッドロック回復のオーバヘッドが大きいので，デッドロックを誤りなく検出しなければならないためである．



(a) バンド幅



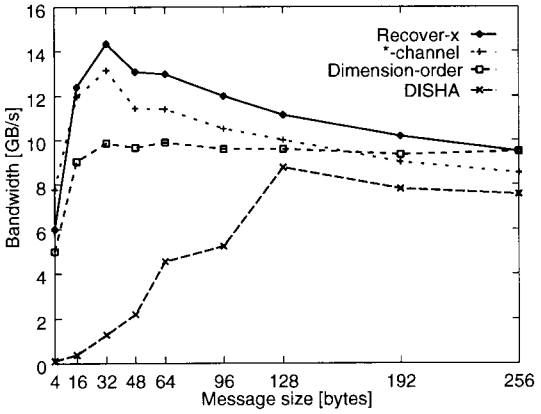
(b) 平均レイテンシ

図 8 Hot-spot 通信
Fig. 8 Hot-spot traffic.

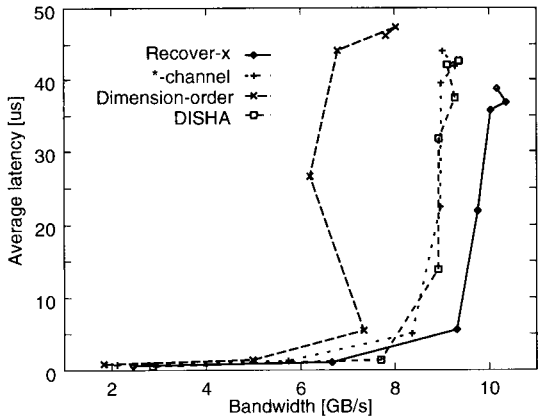
図 8 と**図 9** に，ネットワーク全体のバンド幅とメッセージの平均レイテンシを示す．グラフの横軸は，それぞれメッセージサイズとネットワーク負荷を表すバンド幅である．平均レイテンシの計測では，メッセージの送信間隔を変化させてネットワーク負荷を求めた．ここに示すレイテンシのグラフは，メッセージサイズを 192 バイト（48 フリット）とした場合の結果を表す．なお，他のメッセージサイズにおけるレイテンシ評価も行ったが，一般にそのサイズで高バンド幅なルータが低レイテンシとなり，ネットワークの飽和点が高バンド幅を示した．

5.2 Hot-spot 通信

図 8 (a) より，Recover-x と $*$-channel が高バンド幅を達成していることが分かる．これらのルータでは，比較的小さい 32 バイト程度のメッセージサイズにおけるバンド幅が高い．この理由は，VC 占有時間の関係で，適度に小さいメッセージに対する適応ルーティングの効率が良いためである．Recover-x が $*$-channel



(a) バンド幅



(b) 平均レイテンシ

図 9 All-to-all 通信
Fig. 9 All-to-all traffic.

よりも高バンド幅であることは，適応 VC 数の違いで前者が後者よりも適応ルーティングの自由度が高いことで説明できる．

Dimension-order は，メッセージサイズが大きくなるとバンド幅も高くなるが，ルーティングの自由度がないため 32 バイト以上では頭打ちになる．

DISHA は，16 バイト以下のメッセージでは動作速度が遅いことを考慮しても比較的高バンド幅であるが，それ以上のメッセージサイズでは急激にバンド幅が低下している．これは，デッドロック回復のオーバヘッドによるものである．したがって，デッドロック発生率を低下させる機構を設ける必要が必須といえる[10]．

図 8 (b) に，メッセージの平均レイテンシを示す．仮に，Dimension-order において 192 バイトのメッセージが無衝突で 10 ホップするとすれば，レイテンシは 0.72 μs となる．このグラフから，図 8 (a) の横軸 192 バイト時にバンド幅が高い順にネットワークの飽和点が高くなり，低レイテンシであることが分かる．たとえば，Dimension-order は，ネットワーク全体のバンド幅が 5 GB/s 付近で飽和し，それ以上の負荷をかけようとするとメッセージの平均レイテンシが急上昇する．これに対して Recover-x は，9 GB/s を超えるところまで低レイテンシでメッセージを通信できる．また，DISHA において平均レイテンシが単調増加でない理由は，メッセージの出力頻度によってデッドロック回復の発生回数が異なるためである．

5.3 All-to-all 通信

All-to-all 通信では，ネットワーク全体にメッセージが均等に分散する．そのため，適応ルーティングにより，メッセージの代替経路を探しても Hot-spot 通信に比べて利用可能な VC が見つかりにくい．したがって，図 9 (a) のように Recover-x，$*$-channel と Dimension-order の性能差が小さくなっている．特に，大きなメッセージサイズでは，動作速度の高い Dimension-order が適応ルーティングの効率が下がった $*$-channel を逆転し，Recover-x と同程度のバンド幅を示している．

DISHA は，メッセージサイズが小さいときのバンド幅が低い値を示している．これは，DISHA の VC がすべて適応 VC であり，デッドロック回復に備えて各適応 VC の FIFO には唯一のメッセージしか格納できないことによる[6]．

図 9 (b) を見ると，図 8 (b) と違い，どのルータもネットワークの負荷が約 6 GB/s までは低レイテンシで通信できることが分かる．また，ネットワークの飽和点は接近しているものの，やはり図 9 (a) のメッセージサイズ 192 バイトのときのバンド幅の順に飽和点が高くなっている．ただし，ネットワークが飽和した後の平均レイテンシは，どのルータも急上昇している．

以上のことから，All-to-all のようにユニフォームな通信パターンよりも Hot-spot のように偏りがある通信パターンでの適応ルーティングのメリットが大きいことが分かる．

6. ま と め

本論文では，適応 VC と非適応 VC の実装コストと性能バランスを考慮した完全適応ルーティングアルゴリズム Recover-x を提案した．また，2 次元トーラス用ルータを HDL により設計し，論理合成と RTL シミュレーションにより，他の代表的なルータとコスト/パフォーマンスを比較した．

Recover-x は，比較的少ないハードウェア量で実装でき，他の適応ルータに比べて高速動作が可能である．

また，並列，かつ，オーバヘッドの少ないデッドロック回復をサポートすることにより，高バンド幅，低レイテンシ通信を達成することを示した．

今回は，RTL シミュレータによる実験結果に基づいてその通信性能を議論した．この実験方法は，緻密な評価が行えるというメリットが大きいが，実行時間が長いという不都合ももたなう．したがって，ネットワークサイズやメッセージ数に対する制約も大きい．このため，現在，より高レベルなシミュレータを活用した種々の実験を行っている．ハードウェア設計に基づいて動作速度などを考慮するとともに，より実用的な条件（通信パターンやネットワークへの通信負荷）を使用した評価が今後の課題である．

謝辞 本研究において貴重なご意見をいただきました筑波大学の山口喜教氏に深く感謝いたします．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．

本研究は，一部文部省科学研究費基盤研究（B）課題番号 10558039，奨励研究（A）課題番号 11780190，実吉奨学会の援助による．

参 考 文 献

1) Anjan, K.V. and Pinkston, T.M.: An Efficient, Fully Adaptive Deadlock Recovery Scheme: DISHA, *Proc. 22nd ISCA*, pp.201–210 (1995).
2) Anjan K.V., Pinkston, T.M. and Duato, J.: Generalized Theory for Deadlock-Free Adaptive Wormhole Routing and its Application to *Disha* Concurrent, *Proc. IPPS*, pp.815–821 (1996).
3) Berman, P.E., Gravano, L., Pifarré, G.D. and Sanz, J.L.C.: Adaptive Deadlock and Livelock Free Routing with all Minimal Paths in Torus Networks, *Proc. SPAA* (1992).
4) Chien, A.A.: A Cost and Speed Model for k-ary n-Cube Wormhole Routers, *IEEE Trans. Parallel and Distributed System*, Vol.9, No.2, pp.150–155 (1998).
5) 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).
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) Draper, J.: The Red Rover Algorithm for Deadlock-free Routing on Bidirectional Rings, *Proc. PDPTA '96*, pp.345–354 (1996).
8) 林 匡哉，堀田真貴，吉永 努，大津金光，馬場敬信：適応ルータの効率的なデッドロックリカバリ方式の提案，JSPP '99 論文集，pp.55–62 (1999).
9) 堀田真貴，林 匡哉，中村さゆり，吉永 努，大津金光，馬場敬信：RTL 設計による並列計算機ルータの評価，情報処理学会 ARC 研究報告，Vol.99, No.67, pp.67–72 (1999).
10) López, P., Martínez, J.M. and Duato, J.: DRIL: Daynamically Reduced Message Injection Limitation Mechanism for Wormhole Networks, *Proc. ICPP'98* (1998).
11) Pinkston, T.M. and Warnakulasuriya, S.: On Deadlocks in Interconnection Networks, *Proc. 24th ISCA*, pp.38–49 (June 1997).
12) 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).
13) 吉永 努，林 匡哉，堀田真貴，山口喜教，大津金光，馬場敬信：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，情報処理学会論文誌，Vol.50, No.5, pp.1958–1967 (1999).
14) Vaidya, A.S., Sivasubramaniam, A. and Das, C.R.: LAPSES: A Recipe for High Performance Adaptive Router Design, *Proc.HPCA-5* (1999).

(平成 11 年 8 月 30 日受付)
(平成 12 年 2 月 4 日採録)

吉永 努（正会員）

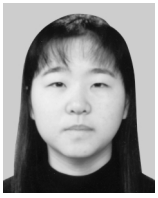
1986 年宇都宮大学工学部情報工学科卒業．1988 年同大学大学院修士課程修了．同年より宇都宮大学工学部助手．博士（工学）．1997 年から翌年にかけて電子技術総合研究所・客員研究員．並列計算機アーキテクチャ，リコンフィギュラブル・コンピューティング等に興味を持つ．電子情報通信学会会員．

林 匡哉（学生会員）

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（学生会員）

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．



中村さゆり

2000 年宇都宮大学工学部情報工学科卒業．HDL によるハードウェア設計に興味を持つ．

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．