計算機アーキテクチャ 134−12
（1999. 8. 3）

RTL 設計による並列計算機ルータの評価

堀田真貴　林　匡哉　中村さゆり
吉永　努　大津金光　馬場敬信

宇都宮大学 工学部 情報工学科

本稿では、我々の提案する Recover-x 適応ルータと他のいくつかのルータについて、各ネットワークポートのバーチャルチャネル (VC) 数を 3 本と 4 本にした場合のハードウェアコストと性能を考察する。その結果、Recover-x ルータが偏りのある通信パターンに対して高バンド幅であることを示す。また、VC の増加はハードウェア量と動作速度に影響するものの、ルーティング自由度の向上がユニフォームな通信パターンに対するバンド幅の向上に貢献することを示す。

An Evaluation of Routers for Parallel Computers based on RTL Design

Maki HORITA, Masaya HAYASHI, Sayuri NAKAMURA,
Tsutomu YOSHINAGA, Kanemitsu OOTSU
and Takanobu BABA

Department of Information Science, Faculty of Engineering,
Utsunomiya University

In this paper, we have considered the hardware cost and performance for the proposed router "Recover-x" and several other routers. In our evaluation, we have designed those routers, which have three or four virtual channels per network port, using Verilog-HDL. The results lead to the following conclusions: the Recover-x router attains high-bandwidth for a non-uniform communication pattern; the number of VCs affects the hardware cost and the operation speed, but the improvement of routing flexibility contributes the network bandwidth, especially for a uniform communication pattern.

1. はじめに

並列計算機におけるネットワークの性能は、システム全体の性能に大きく影響する。我々は今までにネットワーク通信に影響を与えるルータのコスト・パフォーマンスについて様々な実験を行なってきた[6),7)]。また、並列デッドロック回復ルーティングを行なう Recover-x を提案し、その有効性を示してきた[3)]。

これまでの実験では、比較的少数のバーチャルチャネル (VC) 構成を用いていた。しかし、近年の集積回路技術の発展は、より多くの VC を実装可能とし、実際に 4 本程度の VC を持つルータが多い[5)]。そこで、本稿では、2 次元トーラスネットワーク用の Recover-x ルータと他のいくつかのルータについて、VC 数を 3 本と 4 本に統一した場合の RTL 設計を行ない、ハードウェアコストと性能を考察する。その結果、VC の増加はハードウェア量の増加と動作速度の低減に影響するものの、ルーティング自由度の向上が、特にユニフォーム系の通信パターンに対するバンド幅に貢献することを示す。

以降、第 2 章では、本稿で用いるルーティングアルゴリズムと VC 構成、第 3 章では、それを実装するための設計仕様を示す。第 4 章では、各ルータの論理合成結果を示し、ハードウェアコストと動作速度について考察する。第 5 章では、各ルータの RTL シミュレーション結果から性能を評価する。最後に、第 6 章で、本稿のまとめと今後の予定について述べる。

2. ルーティングとバーチャルチャネル構成

本研究で用いたルーティングアルゴリズムについて、VC の使用法を中心に説明する。図 1〜 5は、各ルータにおいて、メッセージを送信するノードの PE インタフェース (PE I/F) から 2 ノード分の可能な VC 切り替えの様子を示している。なお、VC 3 本のときについては文献 3) を参照されたい。

2.1 Dimension-order

2 次元の各次元を X、Y 次元としたとき、はじめに X 次元に沿って宛先ノードと等しい X 座標まで転送する。その後、Y 次元に沿って宛先ノードまで転送する。つまり、宛先までの経路が 1 通りしかない非適応ルーティングである。

VC 4 本は、図 1に示すように 2 本ずつで使い分け、デッドロックの発生を防ぐ。VC 0 と VC 2 はトーラスのラップアラウンドチャネルを越えないメッセージ、VC 1 と VC 3 はトーラスのラップアラウンドチャネルを越えるメッセージが使用する。なお、VC 3 本の Dimension-order ルータは VC バランスを考慮して VC の自動切り替えをサポートするが、VC 4 本では FIFO 性の保証を重視して VC の自動切り替えは行なわない。

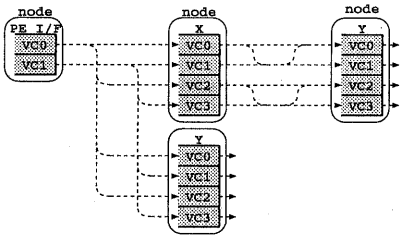

図1 Dimension-order の VC 切り替え

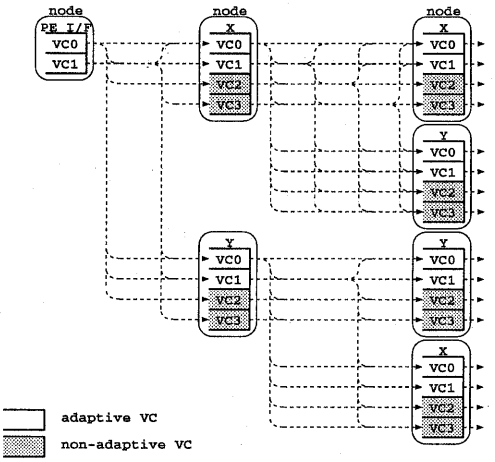

図2 ∗-channel の VC 切り替え

2.2 ∗-channel[2)]

各ネットワークポートに、適応 VC と非適応 VC を設ける。メッセージは、適応 VC を使用して完全適応ルーティングを行なうが、ブロックされると非適応 VC へエスケープする。非適応 VC に移動したメッセージは、Dimension-order ルーティングを行なうが、適応 VC が使用可能なノードでは、再び適応 VC へ戻ることができる。

図 2に VC 切り替えの様子を示す。VC 3 本版から追加する VC は、適応 VC として用いる。このため、VC 切り替え数とルーティングの自由度が増大する。

2.3 DISHA[1)]

どのネットワークポートも全て適応 VC で構成し、デッドロック時の回復パス用にデッドロックバッファ(DB)を用意する。適応 VC 数が多く、メッセージはそれらを自由に切り替えながら転送できるので、ルーティングの自由度は高い。ただし、デッドロック回復のオーバヘッドも大きい。

DISHA は、図 3に示すように VC 切り替えに制限がない真の完全適応 (true fully adaptive) ルーティングをサポートする。本稿では、これを DISHA-true と呼ぶ。ただし、VC 切り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する[3)]。そこで、図 4のように VC 切り替えに制限を与えたもの (DISHA-seq と呼ぶ) についても合わせて考察する。

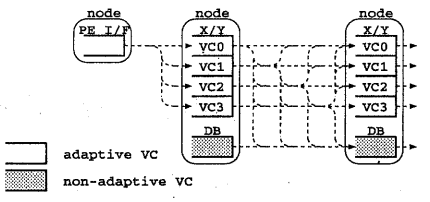

図3 DISHA-true の VC 切り替え

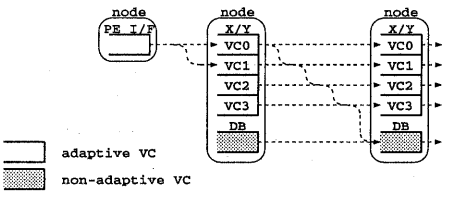

図4 DISHA-seq の VC 切り替え

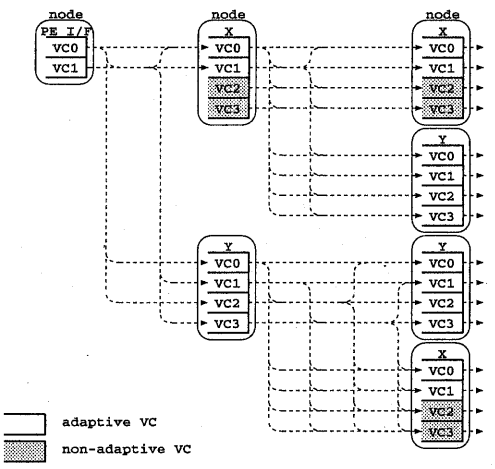

図5 Recover-x の VC 切り替え

2.4 Recover-x[3)]

X 次元ポートに適応 VC とエスケープチャネルとなる非適応 VC、Y 次元には適応 VC のみを持つ。メッセージは、各ポートの適応 VC を用いてルーティングを開始し、デッドロックが発生した場合には、Y 次元のルーティングを完了し、X 次元に沿って進もうとするメッセージの一部を非適応 VC にエスケープすることでデッドロックから回復する。並列デッドロック回復が可能であるため、デッドロックによる性能低下を小さく抑えることができる。

今回の設計では、図 5のように、VC 3 本から追加する VC は、X 次元、Y 次元共に、適応 VC として実装する。これは、VC 4 本の割り当て方法の中で、最も自由に経路選択を行なうことができる割り当てである。

3. ルータの設計仕様

2章で述べたルーティングを行なうルータを実現するための設計仕様を述べる。

3.1 ネットワークとフロー制御

本研究で用いたネットワークトポロジーは、実際の高並列計算機でも広く使用されている2次元トーラスを使用し、フロー制御方式は、パケットサイズに制限がなく、比較的小容量なバッファで実現可能なワームホール方式を用いる。また、全てのルータが最短経路ルーティングを行なう。

3.2 ハードウェア構成

図6(a)に、設計したルータのハードウェア構成を示す。どのルータも、4つのネットワークポート (North, East, West, South Port) と PE I/F を持つ。また、DISHA では全ポートの入力チャネルから結線される DB を追加する。VC数が4の場合のポートの構成を図6(b)に、DB の構成を図6(c)に示す。各ブロックの機能は以下の通りである。

Buffer Controller(BC)： ネットワークからの入力メッセージを受信し、そのバッファリングを制御する。

Address Decoder(AD)： メッセージのアドレスをデコードし、出力候補のポートに出力要求を行ない、出力ポートを選択する。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してデッドロックの検出を行なう。

Output Channel Arbiter(OCA)： 物理チャネルの使用状況と隣接ノードのバッファ状態を基にメッセージの出力要求を調停する。

Virtual Channel output Controller(VCC)： OCA からの出力許可により VC の出力を制御する。

Input Arbiter(IA)： DB において、BC の使用状況とトークン獲得の有無によって DB への入力要求を調停する[1]。

公平な比較をするため、各ルータの全VC数を統一する。東西南北ポートの VC 数は3または4本である。また、PE I/F の VC 数は Dimension-order、*-channel、Recover-x では2本であり、DISHA-true、DISHA-seq では DB があるため1本とする。各 VC は、効率的にフロー制御できるよう8フリット分の FIFO を持つ。

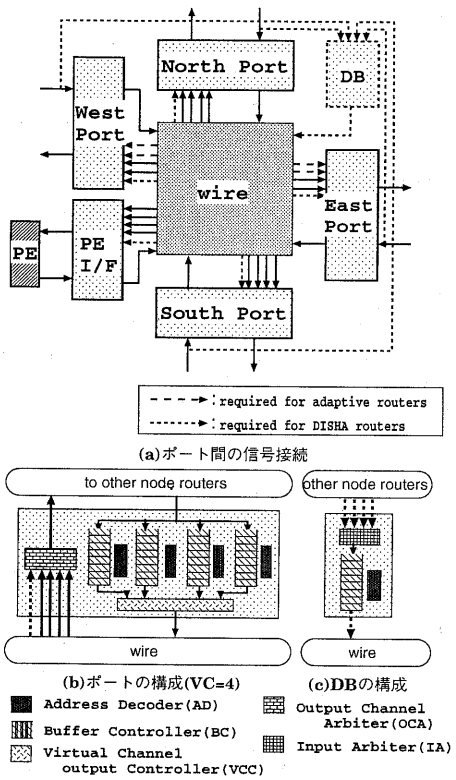

図6 ルータのハードウェア構成

4. 論理合成

以上の各ルータを Verilog-HDL で設計した。ここでは、それらの論理合成結果を示し、ハードウェアコストについて考察する。

4.1 論理合成条件

論理合成は、クロック制約条件を除き、どのルータも同一な条件のもとで行なった。本稿で用いた論理合成条件を以下に示す。

シンセサイザ:
Synopsys HDL Compiler Version 1999.05
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
ターゲットライブラリ:
LSI Logic 0.6μm Array-Based Gate Array

表中の最大動作速度は、各ルータをクロックの立上りエッジ駆動とした場合に、論理合成結果がタイミング条件を満たす中で最も高速であった場合の値を示す。また、面積は最大動作速度時のゲート数を表す。なお、Verilog-HDL ソースプログラムでは、シンセサイザへのディレクティブを適宜指定し、クリティカルパスが短くなるよう配慮した。

4.2 論理合成結果

表1に、ルータの論理合成結果を示す。

各ルータにおいて、VC数を3本から4本にすると動作速度が低下している。これは、OCA のマルチプレックス数がクリティカルパスに影響を与えているためである。次に、アルゴリズム横断的に同数の VC 構成の場合を比較すると、ハードウェア構成が簡単な Dimension-order が最も高速で、VC 切り替えの組合せ数が多い適応ルータほど最大動作速度が遅くなる。例えば、DISHA-true の最大動作速度は、DISHA-seq に比べて VC 3本、4本共に約 15% 低下することがわかる。結論として、最大動作速度は OCA のマルチプレックス数と VC 切り替えの組合せ数に依存するといえる。したがって、DISHA に比べて DB 分のマルチプレックス数が少なく、*-channel に比べて VC 切り替えの組合せ数の少ない Recover-x が適応ルータの中では最速となる。

表 1 論理合成結果

ルータ	(1)		(2)		(3)		(4)		(5)	
VCs/port	3	4	3	4	3	4	3	4	3	4
最大動作速度 (MHz)	161.2	156.2	120.4	114.9	108.1	100.0	123.4	116.2	142.8	133.3
クリティカルパス (ns)	6.20	6.40	8.30	8.70	9.25	10.00	8.10	8.60	7.00	7.50
セル面積 (Kgates)	70.8	90.1	72.1	96.4	79.6	105.3	74.6	94.0	75.6	94.4
配線領域 (Kgates)	40.0	51.6	43.1	60.1	50.3	69.8	43.7	56.5	43.0	58.1
総面積 (Kgates)	110.8	141.7	115.2	156.5	129.9	175.1	118.3	150.5	118.6	152.5

(1) Dimension-order (2) *-channel (3) DISHA-true (4) DISHA-seq (5) Recover-x

各ルータの面積は、バッファ容量、ポート間の結線とVC切り替えの組合せ数、デッドロック検出回路の有無などに影響を受ける。非適応ルータである Dimension-order は、適応ルータに比べてポート間結線も少なく、経路選択ロジックも簡単であるから各 VC 構成において最も面積が小さい。逆に、DB を持ち、VC 切り替えの組合せ数の多い DISHA-true の面積が最大となっている。

5. シミュレーション

ここでは、各ルータの RTL シミュレーション結果を示し、それぞれの通信性能について考察する。

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 ノードであり、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**：すべてのノードが 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**：ノード n が $n+1 \rightarrow n+2 \rightarrow \cdots \rightarrow 99 \rightarrow 0 \rightarrow \cdots \rightarrow n-1$ の順に連続してメッセージを送信する。

今回の評価は、ネットワークの定常状態におけるバンド幅で議論する。これは、実際のネットワークにおいて、ネットワークが空の状態からアプリケーションを開始しているとは限らないからである。このため、シミュレーション開始から終了までの間の一定間隔を評価対象時間とし、他をウォームアップ時間として取り除くこととする。

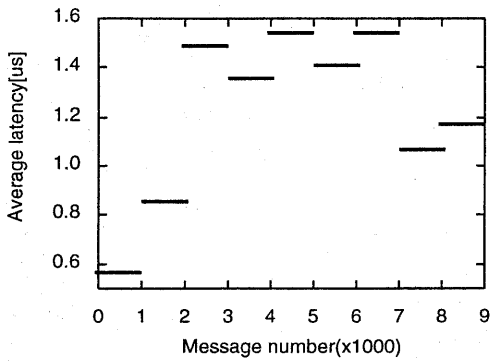

図 7 1000 メッセージ毎の平均レイテンシ

はじめに、ウォームアップ時間を設定するための予備実験について説明する。我々は、種々の通信パターンについてネットワークが空の状態から開始したときの、1000 個毎のメッセージの平均レイテンシを測定した。図 7に、メッセージサイズ 32 バイトに対する Dimension-order(VC 3 本版) の All-to-all 通信の結果を示す。これより、最初の 2000 メッセージは、ネットワークが空いているためにレイテンシが短く、そこから 7000 メッセージまでは、同程度のレイテンシを示すことがわかる。また、7000 メッセージ以降では、全転送を完了したノードが出てくるので、ネットワークの混雑状態が緩和され、再びレイテンシが短くなる。したがって、2000 ～ 7000 メッセージ間を定常状態と定義する。なお、他のメッセージサイズについても試したが、類似した結果が得られた。

ルータ間の伝送遅延は、各ルータの動作速度の 1 クロック時間以内と仮定した。全ルータが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。宛先 PE 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 I/F は送信と受信を並列処理することが可能である。

デッドロック回復ルータがエスケープチャネルを使用するまでのメッセージのブロック時間は、DISHA-seq では 64、Recover-x では 4 クロックとした。これは、種々の実験において最良の結果を示した値である。DISHA-seq のブロック時間が長い理由は、Recover-x よりもエスケープパスが少なく、デッドロック回復のオーバヘッドが大きいためである。

5.2 シミュレーション結果

各通信パターンにおいて、各ルータのクロックを 100MHz で一定にした場合と、表 1に示す最大動作速度の場合のネットワーク全体のバンド幅を示して考察する。なお、DISHA-seq と DISHA-true は、ルータの動作速度とルーティングの自由度が相殺して近似した結果を示したため、グラフの見やすさを考慮して DISHA-seq の結果のみを示す。

5.2.1 Hot-spot 通信

図 8に各ルータが 100MHz で動作する時のバンド幅を、図 9に最大動作速度時のバンド幅を示す。なお、それぞれの図において、VC 3 本版の結果を (a) に、VC 4 本版の結果を (b) に示す。

これらのグラフから、Recover-x が良好なバンド幅を

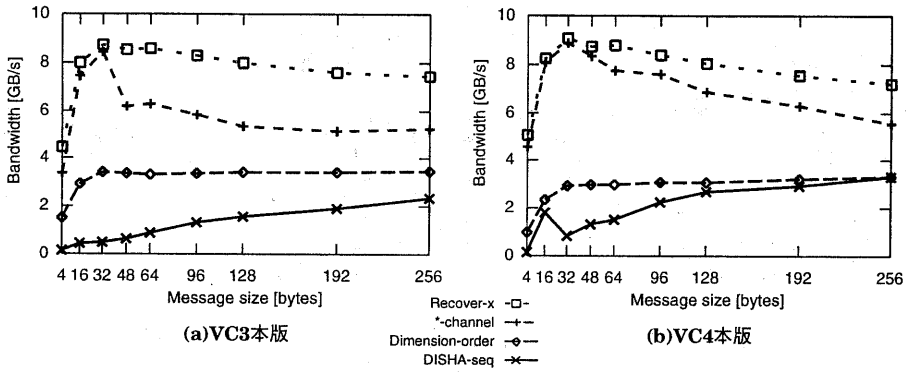

図8 Hot-spot 通信 (100MHz 動作時)

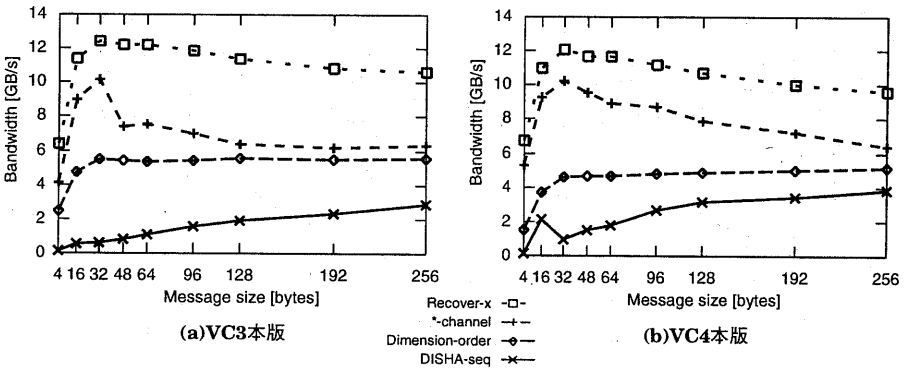

図9 Hot-spot 通信 (最高動作速度時)

示すことがわかる。ただし、Recover-xの4本目のVCはバンド幅の向上に貢献していないことも確認できる。*-channelは、VC数を3本から4本にした場合の効果が大きいが、動作速度が遅いことからRecover-xとの性能差が大きくなる。DISHA-seqは、逐次的なデッドロック回復のオーバヘッドのために、Dimension-orderに劣る結果となっている。Dimension-orderは、VC4本よりも3本の方がやや高バンド幅を示している。これは、VC4本版が静的なVC割り当てであるのに対して、VC3本版では動的な空きVCへの自動切り替えをサポートしたためである。ただし、動的なVC切り替えをサポートすると、非適応ルーティングであってもメッセージのFIFO性の保証が困難になる欠点がある。

5.2.2 All-to-all 通信

図10、11に、それぞれ100MHz動作時と最大動作速度時のバンド幅を示す。

この結果より、各適応ルータについてVC数の増加による効果が確認できる。これは、All-to-all通信が、ネットワーク全体にメッセージを分布させるユニフォーム通信であることによる。すなわち、Hot-spot通信ではメッセージが転送途中で適応経路を選択しても、最終的に輻輳領域が全体のバンド幅を決定してしまう。これに対して、All-to-all通信ではVC数の増加によるルーティングの自由度がバンド幅の向上に貢献していることによる。

VC3本の場合、図10(a)ではDISHA-seq以外のルータのバンド幅は同程度であるが、図11(a)では最大動作速度が勝るDimension-orderのバンド幅が最も高い。一方、VC4本の場合、図10(b)では*-channelとRecover-xが良好なバンド幅を示し、最大動作速度を考慮した図11(b)においても、よい結果を得た。特に、より適応経路を見つけやすい小さなメッセージに対して、これらの適応ルータはDimension-orderよりも高バンド幅である。DISHA-seqは、VC数の増加による効果は比較的大きいが、デッドロック回復のオーバヘッドが依然として大きい結果となった。

一般に、ユニフォーム通信ではDimension-orderルーティングが良好な通信性能を示すことが多いが、適度なルーティング自由度を持たせることにより、適応ルータの性能がDimension-orderよりもよいことがわかる。なお、今回の実験でDISHA-seqのバンド幅が低かった原因は、デッドロックが多発したためであるが、これについてはインジェクション制御などの手法が提案されている[4]。

6. ま と め

本稿では、各ネットワークポートのVC数が3本と4本の場合のDimension-order、*-channel、DISHA、

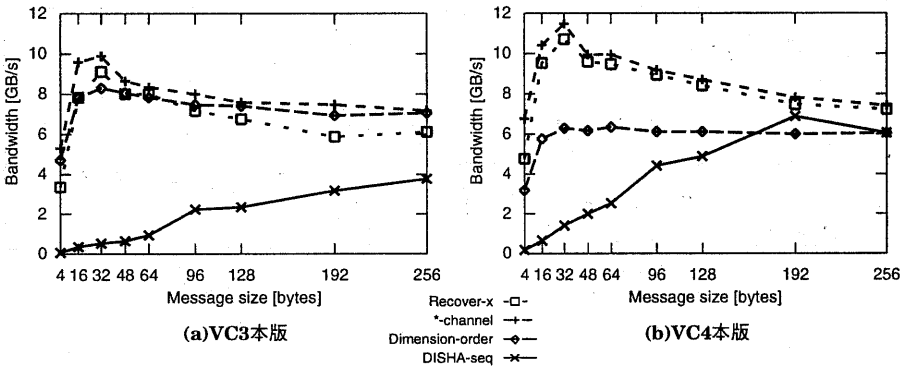

図10 All-to-all 通信 (100MHz 動作時)

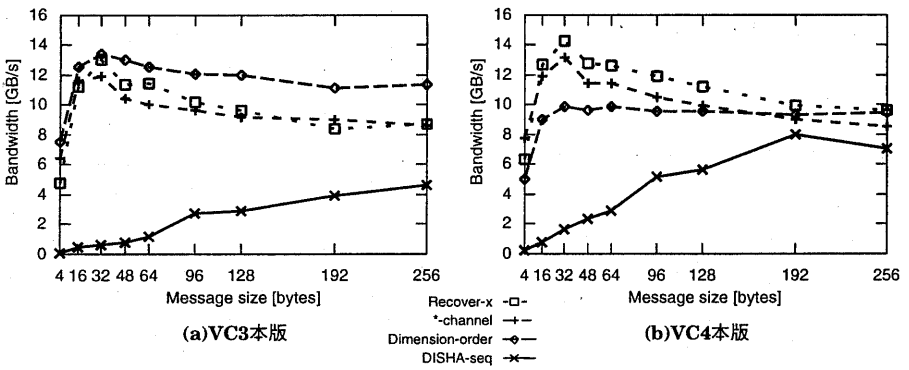

図11 All-to-all 通信 (最高動作速度時)

Recover-x の各ルータを Verilog-HDL で設計し、論理合成、RTL シミュレーションに基づくコスト・パフォーマンスを評価した。その結果、VC 数を 1 本増加することによる動作速度の低下は 3 ～ 8%程度であり、追加した VC を効率的に用いると、その動作速度を考慮しても VC 4 本の方が性能がよい場合が多いことを確認した。また、提案する Recover-x ルータの有効性を示した。

今後の課題として、ルーティングアルゴリズムに対して冗長な VC を持つ場合の効果的な VC 使用法に対する検討が挙げられる。また、より現実的な通信パターンに対するバンド幅やレイテンシの評価を行なう必要がある。

謝辞 本研究において貴重な御意見を頂きました筑波大学の山口喜教氏、宇都宮大学馬場研究室の諸氏に深く感謝致します。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頂いた 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 の援助による。

参 考 文 献

1) K.V. Anjan and T.M. Pinkston: "An Efficient, Fully Adaptive Deadlock Recovery Scheme: DISHA", *Proc. 22nd ISCA*, pp.201-210(1995).

2) P.E. Berman, L. Gravano, G.D. Pifarré and J.L.C. Sanz: "Adaptive Deadlock and Livelock Free Routing with all Minimal Paths in Torus Networks", *Proc. SPAA*(1992).

3) 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並列デッドロックリカバリ方式の提案", 並列処理シンポジウム JSPP'99 論文集, pp.55-62 (1999).

4) F. Petrini, J. Duato, P. López and J.M. Martínez: "LIFE: a Limited Injection, Fully adaptivE, Recovery-Based Routing Algorithm",*Proc. HiPC'97*(1997).

5) A.S. Vaidya, A. Sivasubramaniam and C.R. Das: "LAPSES: A Recipe for High Performance Adaptive Router Design", *Proc. HPCA-5'99*(1999).

6) T. Yoshinaga, M. Hayashi, M. Horita, Y. Yamaguchi, K. Ootsu, and T. Baba: "A Cost and Performance Comparison for Wormhole Routers based on HDL Designs", *Proc. ICPADS'98*, pp.375-382(1998).

7) 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.40, no.5, pp.1958-1967 (1999).