RHiNET-2クラスタを用いたデッドロックフリー固定ルーティングの実機評価

鯉渕道紘[†,††] 渡邊幸之介[†] 大塚智宏[†]
上樂明也[†] 天野英晴[†]

高性能PCクラスタでは，パーソナルコンピュータ（PC）間をシステムエリアネットワーク（SAN）で接続する．SANにおけるデッドロックフリールーティングは近年多くの提案がなされ，シミュレーションによる評価がさかんに行われてきた．しかし，多くのシミュレーションではホスト内のパケット処理などを抽象化しているため，実機のPCクラスタにおける各デッドロックフリールーティングの性能差を正確に見積もることが難しい．そこで，本稿では64台のホストにより構成されたPCクラスタであるRHiNET-2クラスタを用いて各デッドロックフリー固定ルーティングの実機の評価を示す．評価結果より，RHiNET-2クラスタにおいてDLルーティング，構造化チャネル法はほぼ同等のバンド幅，バリア同期時間を示し，これらはUp*/Down*ルーティングに比べて最大51%のバンド幅向上を達成した．また，DLルーティングはUp*/Down*ルーティングに比べNAS Parallel Benchmarksの IS，CG，LUの実行時間を最大3.2%削減することが確認された．

Performance Evaluation of Deadlock-free Deterministic Routings on RHiNET-2 Cluster

Michihiro Koibuchi,[†,††] Konosuke Watanabe,[†]
Tomohiro Otsuka,[†] Akiya Jouraku[†] and Hideharu Amano[†]

System Area Networks (SANs) have been used to connect personal computers in high-performance PC clusters. A large number of deadlock-free routings for SANs have been proposed, and their evaluation on simulations have been widely done. Since hosts, network interfaces and switches used in simulation are often simplified for achieving enough simulation speed, simple simulation results may be hard to estimate the impact of them in real systems. In this paper, we implement deadlock-free routings on a real PC cluster called RHiNET-2, and evaluate their performance. Execution results show that DL routing and structured channel pools achieve almost the same bandwidth and latency of the barrier synchronization. Compared with up*/down* routing, they improve 51% of bandwidth. In addition to the fundamental evaluation, we appraise them using IS, CG, and LU from NAS Parallel Benchmarks (NPB), and DL routing achieves 3.2% improvement on its execution time compared with up*/down* routing.

1. はじめに

PCクラスタにおいてパーソナルコンピュータ（PC）間を接続するシステムエリアネットワーク（SAN）は，システムの性能向上の鍵となっている（Myrinet[1)]，InfiniBand[2)]）．SANは，専用の高速スイッチ群と大容量のpoint-to-pointリンクを用いて構成されるが，ローカルエリアネットワーク（LAN）と異なり，高速なダイレクトメモリ通信を行うためにバーチャルカットスルー方式（VCT方式）もしくはワームホール方式（WH方式）によりパケットを転送する．そのため，SANでは通常，デッドロックフリールーティングが用いられる．

しかし，SANは大規模な並列計算機の結合網[3)]と異なり，任意のスイッチトポロジをサポートしていることが多い．したがって，デッドロックフリーと経路保証を両立させるルーティングアルゴリズムを開発することは難しい．そのため，SANにおけるルーティングアルゴリズムは，1) スパニングツリーの持つ非循環性と連結性を利用するもの（Up*/Down*[4)]，Prefix[5)]，L-turn/R-turn[6)]），もしくは，2) 循環を除去するために仮想チャネルを使用するもの（構造化チャネル法

† 慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
†† バレンシア工科大学
Technical University of Valencia

(SBP)[7]，LASH[8]，DL[9])，などのシンプルな考え方に基づくものが提案されてきた．

これらのデッドロックフリールーティングの性能評価は，現状ではほとんどの場合，確率モデル[10]，エグゼキューションドリブン[11] などのシミュレーションにより行われている．これらのシミュレーションは，様々なルーティングの評価を容易に行うことができるが，スイッチやホスト内のパケット処理を簡素化してモデル化していることが多い．そのため，実機の SAN において各デッドロックフリールーティングの性能差を正確に見積もることが難しい．

そこで，本稿では，64 台のホストで構成される RHiNET-2 クラスタ[12),13)] に各デッドロックフリールーティングを実装し，その性能を評価する．RHiNET-2 クラスタのネットワーク RHiNET-2 は，1) ユーザレベルダイレクトメモリ通信をハードワイヤードで実現したネットワークインタフェース RHiNET-2/NI，2) 8 Gbps の光リンク，3) 64 Gbps カットスルースイッチ RHiNET-2/SW により構成されるネットワークである．RHiNET-2 は代表的な SAN である Myrinet[1]，InfiniBand[2] と同様に多くの固定ルーティングを実装することができる．

以後，2 章では既存の固定ルーティングについて述べ，3 章では RHiNET-2 クラスタについて述べる．そして，4 章において RHiNET-2 クラスタにおけるルーティングアルゴリズムの評価結果を示し，5 章でまとめを述べる．

2. 既存の固定ルーティング

InfiniBand，Myrinet および RHiNET-2 で採用されている固定ルーティングは，パケットの経路が出発地と目的地の組により一意に定まる．そのため，固定ルーティングは，1) パケット配送エラーの検出が容易であり，かつ，2) 目的地におけるパケットのソートなしにパケット配達の FIFO 性を保証することができる，という利点を持つ．

SAN における最も単純な固定ルーティングは Up*/Down*ルーティング[4]☆である．Up*/Down* ルーティングはトポロジ上へスパニングツリーのマッピングを行い，ツリー構造が持つ連結性および非循環性の特性を利用して経路とデッドロックフリーの両方を保証する．しかし，この方法は単純にツリー構造を利用したことにより，1) 非最短経路が発生する，2) トラフィックに偏りが生じやすい，という問題を持つ．ルーティングテーブルのサイズを削減することを目的にした Prefix ルーティング[5] も同様の問題をかかえている．

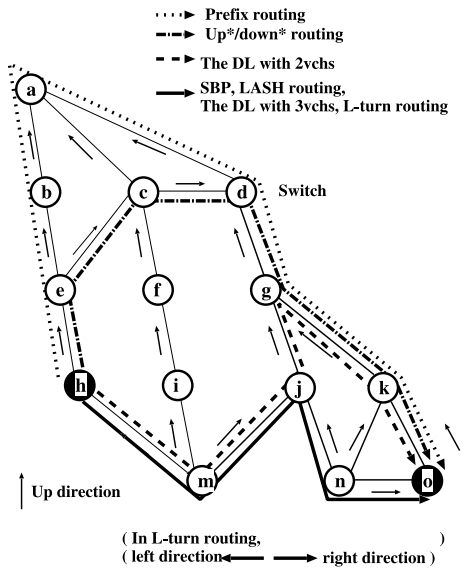


図 1 ルーティング例
Fig. 1 Routing examples.

たとえば，**図 1** に示したスパニングツリーを用いてスイッチ h からスイッチ o へパケットを転送する場合，Up*/Down*ルーティングではスイッチ c を通過する 6 ホップの経路をとる．また，Prefix ルーティングではスイッチ a を通過する 7 ホップの経路となる．これらは非最短経路であり，いずれもルート付近を通過してしまう．

そこで，この問題を解決するために，我々は Up*/Down*ルーティングで用いた 1 次元有向グラフを 2 次元有向グラフに拡張し，論理的に細かい経路制御を行うことでトラフィックの分散を行う L-turn/R-turn ルーティング[6] を提案した．

一方，仮想チャネルを利用して最短経路を保証するルーティング手法[7),8)] も提案されている．構造化チャネル法（SBP）は 1 ホップ進むごとに使用する仮想チャネル番号を増やすことで，最短経路とデッドロックフリーを両立させる．しかし，構造化チャネル法は結合網の直径よりも多い仮想チャネル数がスイッチに必要となる欠点がある．また，LASH ルーティング[8] は SAN を同一トポロジの仮想ネットワークの層に論理的に分割し，循環がおきないように各経路を仮想ネットワークに割り当てる．LASH ルーティングは構造化チャネル法に比べて使用する仮想チャネル数を削減しつつ，最短経路が保証できる利点を持つ．我々が

☆ 元々，適応型ルーティングとして提案されたが，各スイッチ間の経路を 1 つに選択することにより固定ルーティングとして実装することができる[14]．

表 1 固定ルーティングの比較
Table 1 Routing character.

	Up*/Down*, Prefix, L/R-turn	構造化チャネル法	LASH	DL
サイズ制限？	no	yes	yes	no
最短型?	no	yes	yes	no
仮想チャネル？	no	yes	yes	yes

提案している DL ルーティング[9] は LASH ルーティングと同様に仮想ネットワークを使用するが，SAN のトポロジ，仮想チャネル数による適用制限がない点が異なる．DL ルーティングは Up*/Down*ルーティング，L-turn/R-turn ルーティングに比べると経路長が短くなる．しかし，DL ルーティングは最短経路を必ずしも保証することができない．たとえば，図 1 において SAN が仮想チャネルを 2 本提供している場合，DL ルーティングはスイッチ m，g を経由する 5 ホップの経路となる．しかし，SAN が仮想チャネルを 3 本提供している場合，スイッチ n を経由する 4 ホップの最短経路をとることができる．

これらの固定ルーティングの特徴を**表 1** に示す．

3. RHiNET-2 クラスタ

本章では，評価に用いた RHiNET-2 クラスタについて述べる．

3.1 RHiNET-2 クラスタの構成

RHiNET-2 は新情報処理開発機構（RWCP），日立（株），慶應義塾大学により，分散配置されている PC を用いた並列分散環境の構築を目的として開発されたネットワークである．

16 スイッチ，64 台のホストで構成される RHiNET-2 クラスタを**図 2** に示す．各ホストの PCI バス（64 bit, 66 MHz）にはネットワークインタフェース RHiNET-2/NI が装着されている．また，スイッチ，およびホストは 8 Gbps の光リンク（2 m および 5 m）により相互接続されている．**表 2** にホストの仕様を示す．

3.1.1 ネットワークインタフェース RHiNET-2/NI

ネットワークインタフェース RHiNET-2/NI はネットワークコントローラチップ Martini[12]，256 MByte SDRAM，および光インタフェースを持ち，汎用の 64 bit/66 MHz PCI バスを持つ PC に装着する．コントローラチップ Martini はユーザレベルゼロコピー通信，アドレス変換機構，メモリ保護などをハードワイヤードロジックで実装した ASIC チップである．Martini は大きく分けて 2 種類の基本通信命令—リモート

図 2 RHiNET-2 クラスタ
Fig. 2 RHiNET-2 cluster.

表 2 ホストの仕様
Table 2 Specification of host.

CPU	Intel Pentium III 933 MHz × 2 (SMP)
Chipset	Serverworks ServerSet III HE-SL
Memory	PC133 SDRAM 1 GByte
PCI	64 bit/66 MHz
OS	RedHat Linux 7.2 (kernel 2.4.18)

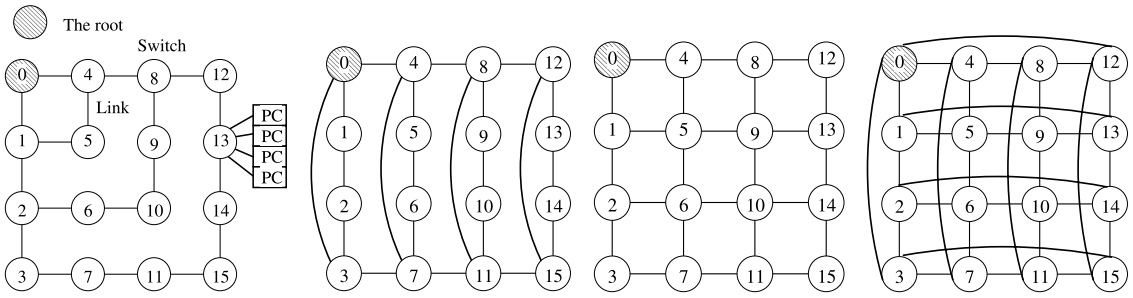
DMA 転送と PIO による転送—を提供する．前者は高バンド幅を実現するためのもので，PUSH（リモートライト）と PULL（リモートリード）の 2 種類の方式が用意されている．後者は，低レイテンシを実現することができるため，PCI バスを用いる場合小さいサイズのデータ転送に適している．パケットはデータ転送単位である 8 Byte のフリットに細分化して転送される．また，ヘッダとテイルは計 40 Byte（5 フリット）である．

3.1.2 スイッチ RHiNET-2/SW

スイッチ RHiNET-2/SW[15] は 8 個の入出力ポートを持ち，8 Gbps の光リンクでホストや他のスイッチと接続される．ただし，現在，より安定した環境を構築するために光リンクの周波数を 800 MHz から 600 MHz に落としている．そのため，現在はリンクの最大転送容量は 6 Gbps となっている．よって，RHiNET-2/SW は本来 64 Gbit/sec のスループットを持っているが，現在は 48 Gbit/sec のスループットで稼働していることになる．

また，RHiNET-2/SW は 16 本の仮想チャネルを提供し，Go & Stop フローコントロールを採用している．各仮想チャネルは 4 KByte のバッファを持っており，各リンクは最大 200 m のリンク長をサポートする．

RHiNET-2 クラスタにおいて，光リンクモジュールの bit-error-rate（BER）は 10^{-20} オーダときわめて低く，さらに，各フリットに ECC を付加することでエ



(a) トポロジ A (b) トポロジ B (c) メッシュ (d) トーラス

図 3 評価に用いたトポロジ
Fig. 3 Topologies considered in execution.

ラー検出，訂正を行っている．そのため，RHiNET-2 では信頼性のある通信がハードウェアレベルで保証されていると見なすことができる．RHiNET-2 クラスタの詳細および性能評価は文献 16) に示されている．

3.2 デッドロックフリールーティング

RHiNET-2 におけるルーティングは，それぞれのスイッチにおいてパケットのヘッダフリットに記述されている目的地をインデックスにしてルーティングテーブルから出力ポートを得るテーブルルックアップ（分散）方式の固定ルーティングである．また，出力仮想チャネル番号の増減は出力ポートと入力ポートの組をインデックスにしてルーティングテーブルから決定される．RHiNET-2/SW は 16 本の仮想チャネルを持つが，データ転送パケットが番号 0 から 7 までの仮想チャネルを使い，応答などのシステム制御パケットが番号 8 から 15 までの仮想チャネルを使う．しかし，両方のパケットとも各スイッチにおいて同一のルーティングテーブルを用いる．

RHiNET-2 クラスタではデッドロックフリールーティングとして元々，構造化チャネル法を改良した縮約構造化チャネル法[15] をサポートしている[7),10)]．縮約構造化チャネル法は 3 本以上の隣接スイッチへのリンクを持つスイッチにおいて，構造化チャネル法と同様にパケットの使用する仮想チャネル番号を 1 増加させる．一方，その他のスイッチを通過するパケットは仮想チャネル番号を切り換えない．

RHiNET-2/SW は縮約構造化チャネル法を任意のトポロジで実装するために，ルーティングテーブルにおいて，1) 目的地ごとのパケットの出力ポート（方向），2) 入力ポートと出力ポートの組ごとの仮想チャネル番号の切換えの有無を変更することができる．このため，RHiNET-2 クラスタでは，任意のトポロジにおいて様々なデッドロックフリールーティングを実装することが可能である．

3.3 システムソフトウェア

RHiNET-2 クラスタには，新情報処理開発機構で開発されたオープンソースのクラスタシステムソフトウェアである SCore[17] が移植されている．SCore では，低レベルのメッセージ通信機構である PM[18] を用いた MPI ライブラリ MPICH-SCore や分散共有メモリシステム SCASH などが利用可能である．また，RHiNET-2 クラスタでは基本通信処理へのソフトウェアの介入を極力減らすために，独自のソフトウェアレイヤを持っている[16]．RHiNET-2 クラスタのソフトウェアレイヤの実装についての詳細は文献 16) に述べられている．

4. 評　　価

4.1 条　　件

4.1.1 トポロジと固定ルーティング

図 3 に示した 4 つのスイッチのトポロジについて評価した．メッシュ，トーラスは多くのルーティングアルゴリズムが最短経路をとることができるのに対し，トポロジ A，B は若干の不規則性を持たせることで各ルーティングアルゴリズムのホップ数が異なるように意図して設計した．各スイッチの 4 つのポートはそれぞれ異なるホストに接続し，残りの 4 つのポートは隣接スイッチに接続する，もしくは使用しない．RHiNET-2 クラスタは 16 台のスイッチで構成されているため，64 ホストを持つ計算システムとなる．

デッドロックフリールーティングについてはスイッチ RHiNET-2/SW が多数の仮想チャネルを持つ利点を生かし，1) 仮想チャネルを必要としない Prefix ルーティング，Up*/Down*ルーティング，2) 多数の仮想チャネル☆を必要とする構造化チャネル法，3) その中

☆ 仮想チャネル数とは，以後データ転送用パケットが使用する本数のことを指す．なお，システム制御用パケットも同数の他の仮想チャネルを使用する．

間の性質を持つ DL ルーティング，の 3 つを次のように実装した．

Up*/Down*ルーティング，Prefix ルーティング
Autonet と同様にスパニングツリーはスイッチ 0 をルートとした幅優先探索（Breadth-first search）で構築した．Up*/Down*ルーティングは本来仮想チャネルを必要としないが，仮想チャネルの効果を調べるために，仮想チャネル数が 1 本，2 本，4 本の場合について各々評価した．そして，ホストごとに使用する仮想チャネル番号を固定し，ネットワーク内では仮想チャネルの切換えを行わないことで仮想チャネル間のトラフィックの分散を図った．

DL ルーティング
DL ルーティングは上記のスパニングツリーを用いた Up*/Down*ルーティングをすべての仮想ネットワークに適応することで実装した．この場合，図 3 の 4 つのトポロジではたかだか 2 つの仮想ネットワーク，つまり，2 本の仮想チャネルがあれば最短経路を保証することができる．そのため本実装では 2 本の仮想チャネルを用いて測定した．RHiNET-2 クラスタは 16 スイッチと小規模であるため，LASH ルーティングをたかだか 2 本の仮想チャネルで実装することができる．そのため，LASH ルーティングは DL ルーティングと同数の仮想チャネルを用いて同一の物理経路をとることになるため評価を省略した．

構造化チャネル法
ホストからパケットを注入するときに仮想チャネル番号 0 を使い，1 ホップ進むごとに使用するチャネル番号を 1 増加させていく．ただし，スイッチ内においてホストと接続しているポートを通過するパケットに対してはチャネルを増加させず，使用する仮想チャネル数をできるだけ抑えるように実装した．そのため，図 3 のトポロジ A，B およびメッシュでは各々6 本の仮想チャネルが必要となり，トーラスでは 4 本の仮想チャネルが必要となる．

これらのうち，Prefix ルーティングを除く 3 つのデッドロックフリールーティングは，元々適応型ルーティングであるため，同一ホスト間で複数経路が選択できる場合がある．そのため，これらを固定ルーティングとして実装するために経路を選択する必要がある[14]．本実装では，メッシュ，トポロジ A，B において経路を分散させるために，同一スイッチ間に複数経路が存在する場合，ホストごとに異なる経路を割り当てた．また，経路の分散が性能に与える影響を調査するために，各スイッチにおいて選択可能な最短経路の中から最も小さい出力ポート番号を通過する経路を選択する low port first[14] も評価した．

図 3 のトポロジ A，B において Up*/Down*ルーティングでは非最短経路が生じるが，DL ルーティング，構造化チャネル法では最短経路をとる．一方，メッシュではすべてのデッドロックフリールーティングにおいてパケットは最短経路をとることができる．なお，Prefix ルーティングはいずれのトポロジにおいても最短経路をとることができない．

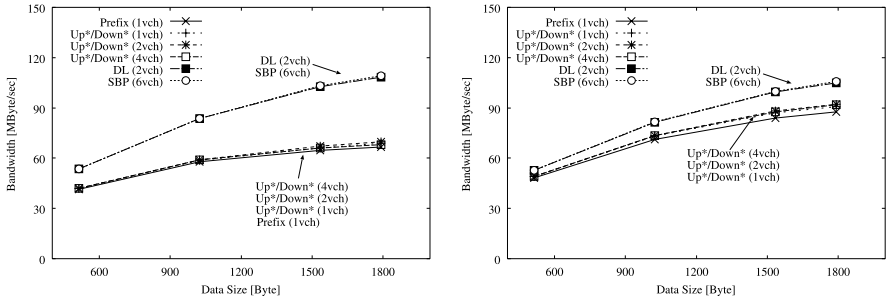
4.1.2 測定項目

平均バンド幅，64 ホストのバリア同期時間，および NAS Parallel Benchmarks（NPB）2.3[19),20)] の中から CG（Conjugate Gradient），IS（Integer Sort），LU（LU-decomposition）の実行時間を測定した．

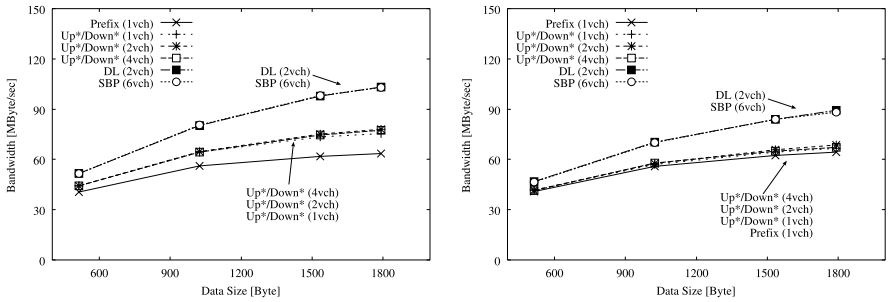
バンド幅 代表的なトラフィックパターン（bit-reversal，matrix transpose，butterfly，complement）[10] に従って，パケットをリモート DMA 転送で送信した場合のホスト間の平均バンド幅とした．ただし，各スイッチにおいて隣接する 4 ホストの中で，送信のみを行う 1 ホストと受信のみを行う 1 ホストの計 2 ホストのみを用いた．これにより，ホスト内のパケット処理を軽くしつつ，1 つの送信ホストあたりのネットワークへのパケット注入量を多くできる．バンド幅測定において発生するパケットはデータサイズが最大 1,792 Byte，1 フリットが 8 Byte であるため，ヘッダ，テイルフリットを含めて最大 229 フリットとなる．

バリア同期時間 全ホストのバリア同期時間の平均とした．データ転送には PIO を用いる．RHiNET-2 クラスタではユニキャストを基にしたマルチキャスト[21] によりバリア同期を行うことができる．ユニキャストを基にしたマルチキャストでは，目的地のホスト数を d とすると，$\lceil log_2(d+1) \rceil$ ステップが必要になり[21]，ホストへの訪問順が性能に影響を及ぼす．そこで，本評価では，MPI の実装でしばしば使用されているランダムに訪問リストを生成する手法[21] をマルチキャストに用いた．そして，1 条件につき 10 パターンの訪問リストを生成し，1 パターンにつき 100,000 回実行して平均をとった．また，バリア同期において発生するパケットの長さは本実装では 17 フリット（ヘッダ，テイル計 5 フリット，データ 1 フリット，残りはハードウェアパディング）となる．

NAS Parallel Benchmarks NPB 2.3 の中から，



(a) bit-reversal トラフィック (b) matrix transpose トラフィック



(c) butterfly トラフィック (d) complement トラフィック

図 4 トポロジ A における固定ルーティングのバンド幅
Fig. 4 Bandwidth of deterministic routings under Topology A.

CG，IS，LU について評価を行った．計算時間に対する通信時間の割合が大きくなるように問題サイズはクラス S とし，ホスト（プロセス）数として 16，32，64 の場合について各々評価した．ただし，LU ではその仕様からクラス S では測定できないため，クラス W で測定した．また，使用したホスト数が 16，32 の場合，1 スイッチあたりそれぞれ 1，2 ホストを用いることで，ネットワーク全体を使用するようにした．

4.2 実行結果

4.2.1 バンド幅

図 4，**図 5**，**図 6**，**図 7** に各デッドロックフリールーティングのバンド幅を示す．

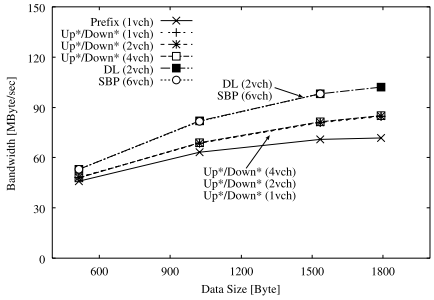
メッシュについては，比較のため e-cube ルーティング[22] についても評価を行った．e-cube ルーティングは，パケットを必要ホップ数 x 方向に転送した後，必要ホップ数 y 方向に転送することで最短経路とデッドロックフリーを保証する．これら 4 つの図において縦軸はバンド幅，横軸はパケットのデータサイズを示しており，SBP（6 vch）は 6 本の仮想チャネルを用いた構造化チャネル法を示している．図 4，5，6，7 より，DL ルーティングと構造化チャネル法は，Up*/Down* ルーティングに比べ最大 51%のバンド幅向上を達成した．一方，DL ルーティングと構造化チャネル法のバンド幅の差はほとんどない．これらの性能差はルーティングアルゴリズムの，1) パケットの平均ホップ数と，2) 経路の分散，の 2 つに起因すると考えられる．

そこで，前者について詳細を調べるために，バンド幅と 2 ホスト間の経由スイッチ数の関係を**図 8** に示す．

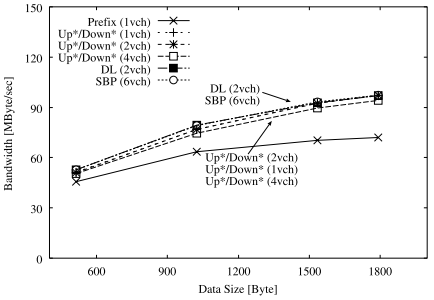
図 8 より，1 ホップ増えるごとにバンド幅が約 7.5 MByte 低下していることが分かる．これは，RHiNET-2 クラスタにおいてリモート DMA 転送を行う場合，各ホストは前のデータ転送の応答パケットを受け取った後，次のデータ転送を行うことが要因と考えられる．図 8 より，ルーティングアルゴリズムのバンド幅は，平均ホップ数に大きく影響されているといえる．

次に，後者の経路の分散について焦点をあてる．一般的に経路が分散することにより，パケットの衝突を抑えることができるため性能は向上すると考えられる．図 6 より，e-cube ルーティングは仮想チャネル 1 本のみを用いるにもかかわらず最も高い性能を示している．e-cube ルーティングはホスト間で均一なアクセスが発生する場合，最も経路が分散する手法の 1 つである．このことから，バンド幅については，ルーティングアルゴリズムにおけるパケットの平均ホップ数と経路の分散の両方が重要であると考えられる．

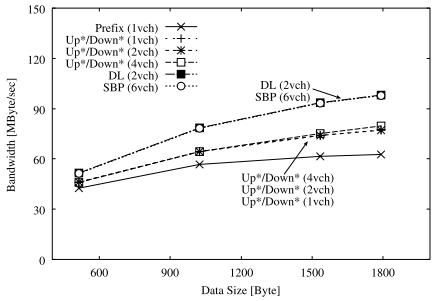
そこで，Up*/Down*ルーティング，構造化チャネ



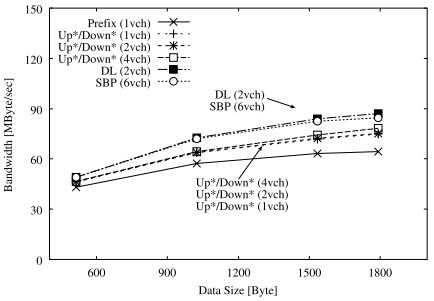
(a) bit-reversal トラフィック



(b) matrix transpose トラフィック

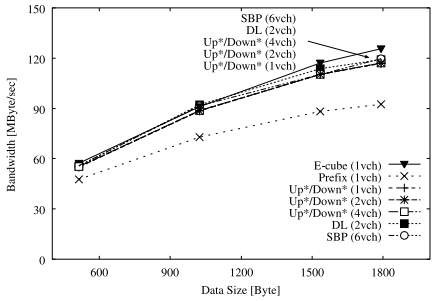


(c) butterfly トラフィック

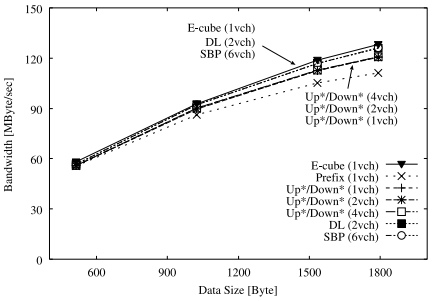


(d) complement トラフィック

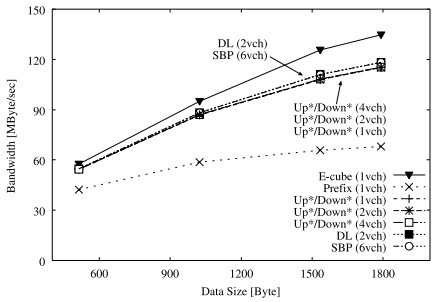
図 5 トポロジ B における固定ルーティングのバンド幅
Fig. 5 Bandwidth of deterministic routings under Topology B.



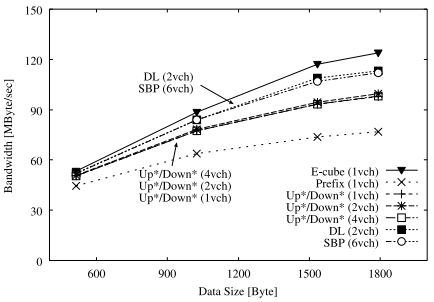
(a) bit-reversal トラフィック



(b) matrix transpose トラフィック

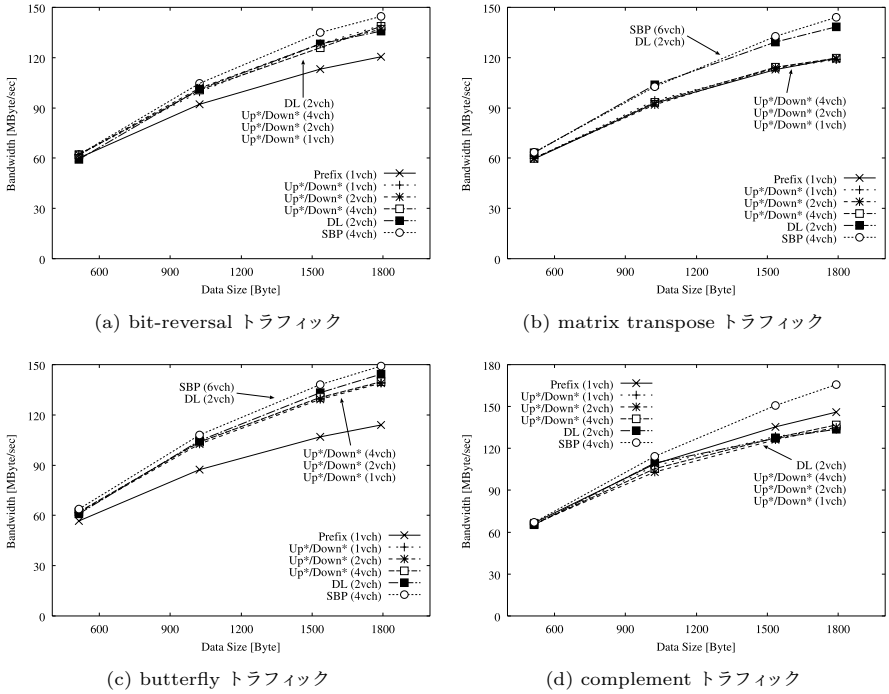


(c) butterfly トラフィック



(d) complement トラフィック

図 6 メッシュにおける固定ルーティングのバンド幅
Fig. 6 Bandwidth of deterministic routings under the mesh.



(a) bit-reversal トラフィック (b) matrix transpose トラフィック

(c) butterfly トラフィック (d) complement トラフィック

図 7 トーラスにおける固定ルーティングのバンド幅
Fig. 7 Bandwidth of deterministic routings under the torus.

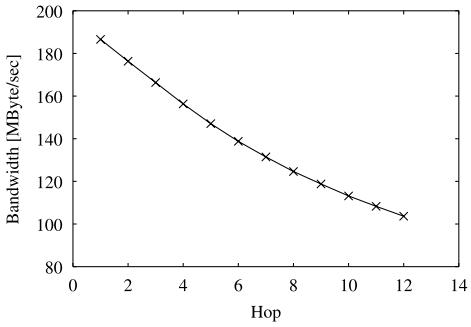


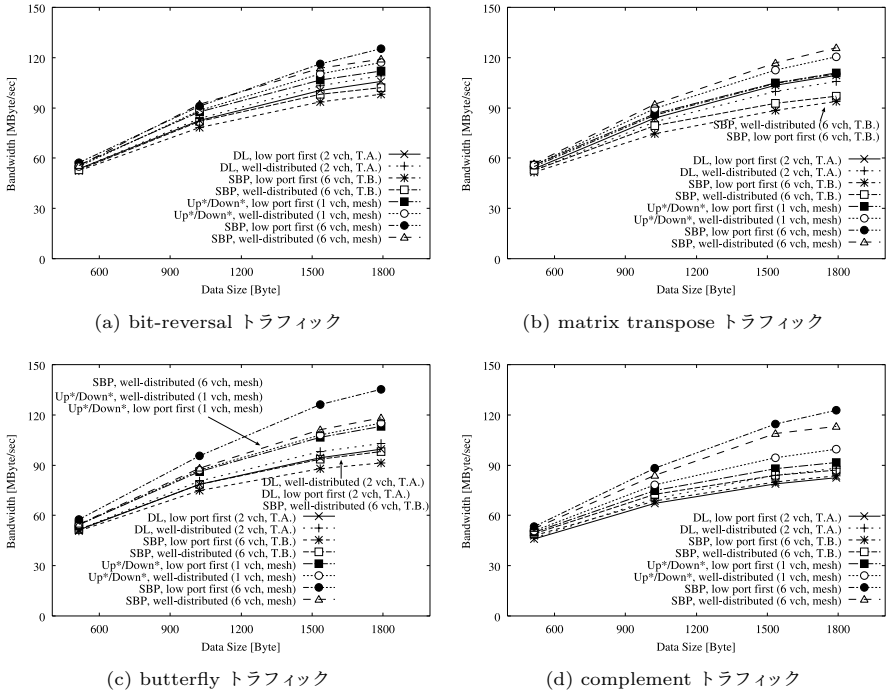
図 8 1,792 Byte データ転送における 2 ホスト間のバンド幅
Fig. 8 Bandwidth between two hosts under 1,792 Byte data.

ル法，DL ルーティングについて，各ホスト間の複数の経路候補の中から，最も小さい番号の出力ポートを使用する経路を使用する方法—low port first—との比較の結果を**図 9** に示す．図 9 において “DL, low port first (2 vch, T.A.)” はトポロジ A において low port first により経路を選択した DL ルーティングを表し，2 本の仮想チャネルを用いていることを示している．経路の分散が難しいトポロジ A，B においては選択可能な物理経路の多い構造化チャネル法，DL ルーティングをそれぞれ用いた．一方で，ノード間の物理経路数が多いメッシュにおいては，構造化チャネル法だけでなく，選択可能な経路数の少ない Up*/Down ルーティングを用いた．図 9 より，ルーティングアルゴリズムの経路の選択は同一ノード間のリンク数が多いトポロジ—メッシュ，トポロジ B，トポロジ A の順—ほど性能に影響し，最大 15%の性能差が生じることが分かる．また，メッシュの構造化チャネル法における経路選択法による性能差は Up*/Down*ルーティングの場合に比べて大きい傾向を示している．よって，同一ホスト間における選択可能な経路数が大きいほど，経路選択法による性能差が大きくなることが確認された．

図 9 において，メッシュにおける構造化チャネル法では low port first が同一スイッチ間の異なるホスト間ごとに異なる経路を使用した場合に比べて高い性能を示している．これは，メッシュなどの均一なトポロジにおいて，すべての最短経路を選択できる場合，e-cube ルーティングのように特定の方向へのパケット転送を優先させて経路を使用—low port first—することにより，全体として経路がより分散したものと考えられる．

一方，図 4，5，6，7 より，同一デッドロックフリールーティングにおける使用仮想チャネル数によるバンド幅の差がほとんどないことが分かる．そのため，使用する仮想チャネル数はバンド幅にほとんど影響しない，ということがいえる．

これらの結果より，バンド幅は，ルーティングアルゴ



(a) bit-reversal トラフィック

(b) matrix transpose トラフィック

(c) butterfly トラフィック

(d) complement トラフィック

図 9 固定ルーティングの経路選択方法のバンド幅
Fig. 9 Bandwidth of path selection method in deterministic routings.

リズムの，1) パケットの平均ホップ数と，2) 経路の分散，が重要であることが確認された．また，この点よりDL ルーティングと構造化チャネル法は Up*/Down* ルーティングより優れているといえる．また，仮想チャネルの本数が性能にあまり影響ないことから，RHiNET-2 クラスタではより少ない本数でデッドロックフリーと最短経路を両立できる DL ルーティングが構造化チャネルよりも優れているといえる．

4.2.2 バリア同期時間

表 3 に各デッドロックフリールーティングのバリア同期の実行時間を示す．また，表 3 の括弧内にトポロジ A において最もルーティングアルゴリズム間の性能差が大きかったホスト訪問順の実行時間を示し，ランダムにリストを生成した場合に生じうる性能差を示した．

ルーティングアルゴリズムのホップ数に焦点をあてると，表 3 より，1) パケットの平均ホップ数が最も大きい Prefix ルーティングがいずれのトポロジにおいても最もバリア同期時間が大きい，2) トポロジ A および B において DL ルーティングと構造化チャネル法はほぼ同じバリア同期時間であり，Up*/Down* ルーティングに比べて最大 26%のバリア同期時間向上を実現している，3) メッシュにおいては Up*/Down* ルーティング，DL ルーティング，構造化チャネル法はほぼ同じバリア同期時間である，ということが分かる．これは，パケットの平均ホップ数に違いがある場合にバリア同期時間に大きな差が出ることを示しており，平均ホップ数の削減がバリア同期時間の削減に効果的であるといえる．

表 3 バリア同期の実行時間（μsec）
Table 3 Latency of barrier synchronization (μsec).

	トポロジ A	トポロジ B	メッシュ
Prefix (1vch)	52.39 (60.54)	49.70	51.49
Up*/Down* (1vch)	50.49 (59.77)	48.65	45.78
Up*/Down* (2vch)	50.53 (59.87)	48.68	45.77
Up*/Down* (4vch)	50.55 (59.58)	48.53	45.78
構造化チャネル法	47.83 (44.58)	46.61	45.61
DL (2vch)	47.86 (44.60)	46.65	45.62

そこで，ホップ数がバリア同期に与える影響について調査した結果を**図 10** に示す．図 10 は 2 ホスト間の距離—経由スイッチ数—を変化させたときのバリア同期時間を示しており，1 ホップ増えるごとに約 0.7 μsec 増えていることが分かる．64 ホストのバリア同期の場合，収集と解放の手続きで計 12 ステップのユニキャストが必要になる．そのため，図 10 からルーティングアルゴリズムの平均ホップ数がバリア同期時間に大きく影響するといえる．したがって，DL ルーティングと構造化チャネル法が Up*/Down*ルーティングおよび Prefix ルーティングに比べてバリア同期を高速

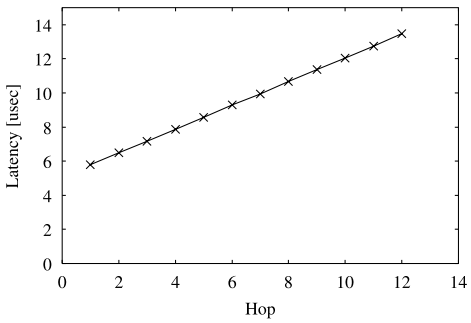


図 10 2 ホスト間のバリア同期時間
Fig. 10 Executiohn time of barrier synchronization on 2 hosts.

に行うことができるため優れているといえる．

次に，パケットの衝突に焦点をあて，バリア同期時間について検討する．本実装では，メッシュにおいて DL ルーティングと Up*/Down*ルーティングは同ホップ数であるが，DL ルーティングと異なり，Up*/Down*ルーティングでは経路に偏りが生じる．そのため，メッシュにおいて Up*/Down*ルーティングは DL ルーティングに比べて衝突が多く発生すると考えられる．しかし，表 3 より，メッシュにおいて，DL ルーティングと Up*/Down*ルーティングの性能差は 1 μsec 未満であり小さいといえる．本評価ではパケット長が直径に比べて大きいため，パケットの衝突は頻繁に起きていると考えられる．しかし，これらの結果よりバリア同期時間にはパケットの衝突はあまり影響しないといえる．

最後に仮想チャネルの影響について焦点をあてる．表 3 より Up*/Down*ルーティングにおいて仮想チャネル数の違いによるバリア同期時間の差はほとんど見られない．そのため，同一デッドロックフリールーティングにおいて仮想チャネル数の増加によるバリア同期時間の削減効果はほとんどないといえる．Slack buffer を用いたフロー制御ではパケットの先頭フリットが衝突した場合にも後続フリットをそのまま Slack buffer に格納することができる．そのため，RHiNET-2 クラスタにおいて仮想チャネル数の効果が少ない理由は実際には 1 つの物理チャネルにおいて仮想チャネルの切換えが起こりにくいためと考えられる．

4.2.3 NAS Parallel Benchmarks

NPB を用いたルーティングアルゴリズムの評価結果を**表 4**，**表 5**，**表 6**，**表 7** に示す．表 4，5，6，7 より，DL ルーティングは Up*/Down*ルーティングに比べ最大 3.2%の性能向上を達成していることが分かる．Prefix ルーティングと構造化チャネル法は未測定であるが，IS では MPI ライブラリの ALLTOALL 関数を使った全対全通信が多いため，図 4，5，6，7 に示した基礎的なバンド幅の評価結果と近い傾向が出ると予想される．具体的には，DL ルーティングと構造化チャネル法はほぼ同じ実行時間を示し，Prefix ルーティングは他のものに比べて劣ると考えられる．ベンチマークの振舞いを見るために，実行時間の内訳を**表 8** に示す．表 8 において，通信時間はスイッチング遅延，リンク遅延，メモリコピー，およびソフトウェアオーバヘッドを含む．表 8 より，通信時間がシステム性能を支配していることが分かる．しかし，通信時間のうち，ソフトウェアオーバヘッドが相対的に大きいと考えられ，ルーティングアルゴリズムの影響は限られている．また，通信時間が相対的に大きいため，ホスト数を 16，32，64 と増加させているにもか

表 4 トポロジ A における CG，LU，IS のベンチマークの実行時間（sec）
Table 4 Execution time of CG, LU and IS benchmarks under Topology A (sec).

	CG.S.16	LU.W.16	IS.S.16	IS.S.32	IS.S.64
U/D, 1vc	0.20600	7.78050	0.01886	0.02169	0.11798
U/D, 2vc	0.20633	7.81633	0.01883	0.02157	0.11798
U/D, 4vc	0.20700	7.78900	0.01863	0.02162	0.11797
DL, 2vc	0.20450	7.76525	0.01850	0.02091	0.11751

表 5 トポロジ B における CG，LU，IS のベンチマークの実行時間（sec）
Table 5 Execution time of CG, LU and IS benchmarks under Topology B (sec).

	CG.S.16	LU.W.16	IS.S.16	IS.S.32	IS.S.64
U/D, 1vc	0.20325	7.79975	0.01846	0.02087	0.11704
U/D, 2vc	0.20450	7.79450	0.01843	0.02077	0.11704
U/D, 4vc	0.20400	7.82300	0.01854	0.02090	0.11698
DL, 2vc	0.20400	7.77257	0.01838	0.02052	0.11681

表 6 メッシュにおける CG，LU，IS のベンチマークの実行時間（sec）
Table 6 Execution time of CG, LU and IS benchmarks under mesh (sec).

	CG.S.16	LU.W.16	IS.S.16	IS.S.32	IS.S.64
U/D, 1vc	0.20433	7.78100	0.01842	0.02039	0.11686
U/D, 2vc	0.20325	7.78433	0.01854	0.02043	0.11670
U/D, 4vc	0.20433	7.78300	0.01833	0.02029	0.11667
DL, 2vc	0.20325	7.76200	0.01836	0.02052	0.11673

表 7 トーラスにおける CG，LU，IS のベンチマークの実行時間（sec）
Table 7 Execution time of CG, LU and IS benchmarks under torus (sec).

	CG.S.16	LU.W.16	IS.S.16	IS.S.32	IS.S.64
U/D, 1vc	0.20367	7.77100	0.01837	0.02013	0.11580
U/D, 2vc	0.20367	7.75350	0.01837	0.02010	0.11631
U/D, 4vc	0.22467	7.71600	0.01835	0.02006	0.11629
DL, 2vc	0.20367	7.75225	0.01820	0.02008	0.11631

表 8 トポロジ A における DL ルーティングの CG，IS の実行時間の内訳（sec）
Table 8 Itemized statement of DL routing under CG and IS in Topology A (sec).

	comp.	Comm.	Wait
CG.S.16	0.02551	0.10675	0.07186
IS.S.64	0.00061	0.11387	0.00243

かわらず実行時間が改善されなかったと考えられる．

4.2.4 実行結果の比較，有効性

図 4，5，6，7 より，bit-reversal などのトラフィックにおけるルーティングアルゴリズムのバンド幅の差は最大 51%と大きい．一方で，表 3，4，5，6，7 より，バリア同期時間，NPB の実行時間におけるこれらの性能差は小さい．したがって，今回の規模の PC クラスタにおいては，ルーティングアルゴリズムの差がアプリケーション実行時の性能に必ずしも直結していないことが分かる．また，ルーティングアルゴリズムの差がアプリケーションの実行に影響を与えるのは，よりシステムのサイズが大きい場合であることが予想される．

RHiNET-2 は，光インタコネクトを用いて机上で利用中の PC を接続するための独自機能を持っている点に特徴がある．しかし，本評価においては，光インタコネクトは通常の PC クラスタと同程度の配線長のものを利用し，同一構造の PC を他のアプリケーションの負荷なしに稼働させている．この点で本クラスタの利用法は Myrinet-2000 などを用いて構成した一般的な PC クラスタと同じである．また，机上で利用中の PC を接続するために必要な RHiNET 特有の処理のオーバヘッドは，ネットワークインタフェース内の専用ハードウェアが実行することで隠蔽されるため，RHiNET-2 は Myrinet-2000 とほぼ同等の高バンド幅，低レイテンシを提供することが報告されている[16]．したがって，本稿における結果は，一般的な SAN における 1 つの実行結果として扱うことができる．

4.2.5 既存のシミュレーション研究との比較

これまで，確率モデルシミュレーションを用いたルーティングアルゴリズムの評価はさかんに行われてきた[9),10)]．それらの多くにおいて，SAN におけるデッドロックフリールーティングの性能差は，本稿で示した結果に比べて大きいと報告されている．たとえば，DL ルーティングの評価のために我々が行ったシミュレーション結果[9)] と本評価結果の比較を**表 9** に示す．表 9 において，パケットの平均ホップ数の差は，同一の経路選択方法を用いた DL ルーティングと Up*/Down*ルーティングの最大性能差を示し，経路分散は，DL ルーティングにおいて，静的な経路解析を行い経路を分散させた手法と low port first の最大性能差を示す．

表 9 ホップ数，経路分散が及ぼすスループットへの影響（%）
Table 9 Influence of hops and path distribution on throughput (%).

要因	シミュレーション	RHiNET-2 クラスタ
パケットのホップ数	67	51
経路分散	71	15

このシミュレーション条件[9)] は，16 スイッチのトポロジを採用し，スイッチング技術としてバーチャルカットスルー方式を用いており，仮想チャネルは 3 本，パケット長は 128 フリットである．よって，これは RHiNET-2 クラスタにおける本測定条件に近い構成といえる．

本評価により得られた結果が，これらの確率モデルシミュレーションとの結果と特に異なる点は次である：1) 経路選択方法によるバンド幅の差は最大 15%と小さい．2) 仮想チャネルの本数による性能への影響は小さい．両者より，実機ではルーティングアルゴリズムの経路長が，相対的により重要な性能向上の要因になっているといえる．ただし，これは，実機と確率モデルシミュレーションの測定項目の違いも影響していると考えられる．シミュレーションにおけるスループットは，各ホストにおける単位時間あたりの受信フリット数で表されることが多い[10)]．一方で，本測定を含む実機におけるバンド幅 B は，通常，送信データサイズを D，目的地ホストからの応答パケットを受信するまでの時間を T とすると

$$B = D/T$$

で測定される．したがって，経路長はシミュレーションにおけるスループットでは間接的な影響にとどまる一方，実機におけるバンド幅測定では大きく影響したと考えられる．また，これより，実機における経路分散の影響がシミュレーションに比べて小さいことと同様に，経路長の影響もシミュレーションに比べて小さいといえる．

一般的に，シミュレーションでは実行時間をおさえるために，ルーティングアルゴリズムに依存しない機能を簡略化している場合が多い．たとえば，ホストからのパケットの注入間隔は，ホスト処理の時間を考えていないことが多い[9),14)]．一方で，RHiNET-2 クラスタではパケットの注入間隔はホストにおける処理（e.g. メモリへの DMA 転送時間や応答パケットの生成時間）を含む．そのため，RHiNET-2 クラスタに

おけるルーティングアルゴリズムの影響はシミュレーションと比較して全体的に小さくなったと考えられる．そのため，bit-reversal トラフィックなどのトラフィックにおいても実機におけるトレースを収集，解析し，パケットの注入間隔を調整することでシミュレーションの精度を上げることは可能であると考えられる．

5. ま と め

16 スイッチ 64 ホストを用いた RHiNET-2 クラスタにおける 4 つのデッドロックフリールーティングのバリア同期時間およびバンド幅を測定した．また，NAS Parallel Benchmarks（NPB）2.3 を用いた実行時間の比較も行った．実行結果より，1) DL ルーティング，構造化チャネル法が Up*/Down*ルーティングに比べ，51%のバンド幅向上を達成した，2) DL ルーティングと構造化チャネル法はほぼ同じバンド幅，バリア同期時間であった，3) DL ルーティングは Up*/Down*ルーティングに比べ NPB 2.3 の実行時間を最大 3.2%削減した，ということが分かった．つまり，bit-reversal などのトラフィックにおけるルーティングアルゴリズムのバンド幅の差はと大きい一方で，バリア同期時間，NPB の実行時間におけるこれらの性能差は小さい．したがって，PC クラスタの実用においては，ルーティングアルゴリズムの性能がシステム性能に必ずしも直結しないといえる．

また，これらの結果より，ルーティングアルゴリズムはパケットの平均ホップ数の削減，および経路の分散が重要であり，仮想チャネル数は性能にあまり影響しないということが分かった．この点より，使用仮想チャネル数が少なく，経路を分散しつつ，平均ホップ数が小さい DL ルーティングは Up*/Down*ルーティング，Prefix ルーティング，構造化チャネル法に比べてバランスのとれた手法であるといえる．また，RHiNET-2 クラスタにおける評価結果は，一般的な確率モデルを用いたシミュレーション結果と比べてルーティングアルゴリズムと仮想チャネルの影響が小さい傾向にあることが分かった．

謝辞 本研究を行うにあたり RHiNET-2 クラスタについて貴重なご意見を下さった慶應義塾大学理工学部西宏章助手，河野賢一氏，北村聡氏に感謝いたします．

参 考 文 献

1) Myricom, Inc.. http://www.myri.com/
2) I.T. Association: InfiniBand architecture, Specification Volumen 1, Release 1.0.a (2001). available from the InfiniBand Trade Association, http://www.infinibandta.com
3) 天野英晴：並列コンピュータ，昭晃堂 (1996).
4) Schroeder, M.D. et al.: Autonet: a high-speed, self-configuring local area network using point-to-point links, *IEEE Journal on Selected Areas in Communications*, Vol.9, pp.1318–1335 (1991).
5) Wu, J. and Sheng, L.: Deadlock-Free Routing in Irregular Networks Using Prefix Routing, *Proc. Parallel and Distributed Computing Systems*, pp.424–430 (1999).
6) 上樂，鯉渕，天野：2 次元 Turn モデルに基づくイレギュラーネットワーク向けルーティングアルゴリズムの設計と評価，情報処理学会論文誌：コンピューティングシステム，Vol.44, No.SIG 11 (ACS 3), pp.157–168 (2003).
7) 堀江，石畑，池坂：並列計算機 AP1000 における相互結合網のルーチング方式，電子情報通信学会論文誌，Vol.J75-D-1, No.8, pp.600–606 (1992).
8) Skeie, T., Lysne, O. and Theiss, I.: Layered Shortest Path (LASH) Routing in Irregular System Area Networks, *Proc. International Parallel and Distributed Processing Symposium*, pp.162–169 (2002).
9) Koibuchi, M., Jouraku, A. and Amano, H.: Descending Layers Routing: A Deadlock-Free Deterministic Routing using Virtual Channels in System Area Networks with Irregular Topologies, *Proc. International Conference on Parallel Processing*, pp.527–536 (2003).
10) Duato, J., Yalamanchili, S. and Ni, L.: *Interconnection Networks: an engineering approach*, Morgan Kaufmann (2002).
11) Flich, J., Malumbres, M.P., Lopez, P. and Duato, J.: Performance Evaluation of Networks of Workstations with Hardware Shared Memory Model Using Execution-Driven Simulation, *Proc. International Conference on Parallel Processing*, pp.146–153 (1999).
12) Watanabe, K., Otsuka, T., Tsuchiya, J., Harada, H., Yamamoto, J., Nishi, H., Kudoh, T. and Amano, H.: Performance Evaluation of RHiNET-2/NI: A Network Interface for Distributed Parallel Computing Systems, *Proc. International Symposium on Cluster Computing and the Grid*, pp.318–325 (2003).
13) 鯉渕，渡邊，河野，上樂，天野：RHiNET-2 クラスタを用いたルーティングアルゴリズムの実機評価，電子情報通信学会技術研究報告 CPSY-2003-13，pp.43–48 (2003).
14) Koibuchi, M., Jouraku, A. and Amano, H.: The Impact of Path Selection Algorithm of Adaptive Routing for Implementing Determin-

istic Routing, *Proc. International Conference on Parallel and Distributed Processing Techniques and Applications*, pp.1431–1437 (2002).
15) Nishimura, S., Kudoh, T., Nishi, H., Yamamoto, J., Harasawa, K., Matsudaira, N., Akutsu, S., Tasho, K. and Amano, H.: High-speed network switch RHiNET-2/SW and its implementation with optical interconnections, *Hot Interconnect*, pp.31–38 (2000).
16) 大塚, 渡邊, 北村, 原田, 山本, 西, 工藤, 天野：分散並列処理用ネットワーク RHiNET-2 の性能評価, 先進的計算基盤システムシンポジウム SACSIS, pp.45–52 (2003).
17) Ishikawa, Y., Tezuka, H., Hori, A., Sumimoto, S., Takahashi, T., O'Carroll, F. and Harada, H.: RWC PC Cluster II and SCore Cluster System Software — High Performance Linux Cluster, *5th Annual Linux Expo*, pp.55–62 (1999).
18) Takahashi, T., Sumimoto, S., Hori, A., Harada, H. and Ishikawa, Y.: PM2: High Performance Communication Middleware for Heterogeneous Network Environment, *SC2000*, pp.52–53 (2000).
19) Bailey, D., Harris, T., Saphir, W., Wijngaart, R., Woo, A. and Yarrow, M.: The NAS Parallel Benchmarks 2.0, NAS Technical Report, NAS-95-020 (1995).
20) Bailey, D., Harris, T., Saphir, W., Wijngaart, R., Woo, A. and Yarrow, M.: New Implementations and Results for the NAS Parallel Benchmarks 2, *PP97* (1997).
21) Kesavan, R. and Panda, D.: Efficient Multicast on Irregular Switch-Based Cut-Through Networks with Up-Down Routing, *IEEE Trans. Parallel and Distributed Systems*, Vol.12, No.8, pp.808–828 (2001).
22) Dally, W.J. and Seitz, C.L.: Deadlock-Free Message Routing in Multiprocessor Interconnection Networks, *IEEE Trans. Comput.*, Vol.36, No.5, pp.547–553 (1987).

(平成 16 年 1 月 31 日受付)
(平成 16 年 5 月 9 日採録)

鯉渕 道紘

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）．現在，同大学理工学部情報工学科およびバレンシア工科大学コンピュータ工学科訪問研究員，平成 14 年度より日本学術振興会特別研究員．相互結合網に関する研究に従事．

渡邊幸之介

平成 15 年慶應義塾大学大学院理工学研究科開放環境科学専攻前期博士課程修了．現在，同後期博士課程に在学．平成 16 年度より日本学術振興会特別研究員．PC クラスタ向けネットワークインタフェースに関する研究に従事．

大塚 智宏

平成 15 年慶應義塾大学大学院理工学研究科開放環境科学専攻前期博士課程修了．現在，同後期博士課程に在学．PC クラスタのネットワーク，通信ミドルウェアの研究に従事．

上樂 明也

平成 12 年慶應義塾大学大学院修士課程修了．現在，慶應義塾大学大学院理工学研究科博士課程に在学中．相互結合網に関する研究に従事．

天野 英晴（正会員）

昭和 56 年慶應義塾大学理工学部電気工学科卒業．昭和 61 年同大学大学院理工学研究科電気工学専攻博士課程修了．現在，慶應義塾大学理工学部情報工学科教授．工学博士．計算機アーキテクチャの研究に従事．