RHiNET-2クラスタを用いたシステムエリアネットワーク向けトポロジの実機評価

鯉渕道紘[†,††] 渡邊幸之介[†]
大塚智宏[†] 天野英晴[†]

高性能 PC クラスタで用いられているシステムエリアネットワーク（SAN）は，多くの場合，拡張性，耐故障性を重視して，様々なトポロジをとることができる．本稿では 64 台のホストにより構成される RHiNET-2 クラスタにおいてトポロジの性能を評価する．評価の結果，2 次元トーラスは 2 次元メッシュに対し最大 47%のバンド幅向上を示し，また，Myrinet-Clos 網は同数のホストを接続した Fat ツリー，2 次元トーラスに対し，最大 48%のバンド幅向上を示した．また，基礎的な評価に加えて NAS Parallel Benchmarks を用いたトポロジの評価も行った．その結果，トーラスが Myrinet-Clos 網に比べて最大 20%の性能向上を達成した．

Performance Evaluation of Topologies for System Area Networks on RHiNET-2 Cluster

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

System Area Network (SAN) usually accepts arbitrary topologies since connection flexibility and robustness are preferred in high-performance PC clusters. In this paper, we evaluate their performance on a real PC cluster with 64 hosts called RHiNET-2. Execution results show that two-dimensional torus improves up to 47% of bandwidth compared with two-dimensional mesh, and Myrinet-Clos network achieves up to 48% improvement on bandwidth compared with fat tree and two-dimensional torus with the same number of hosts. In addition to the fundamental evaluation, we appraise them using NAS Parallel Benchmarks (NPB), and two-dimensional torus achieves 20% improvement on their execution time compared with Myrinet-Clos network.

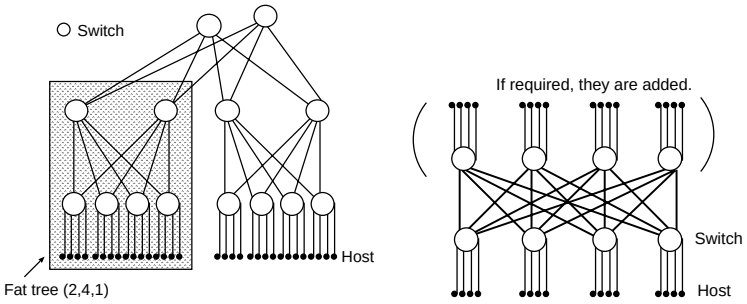
1. はじめに

高性能 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，システムの性能に影響を与える（Myrinet[1)]，QsNET[2)]，InfiniBand[3)]）．SAN はダイレクトメモリ通信を高速に行うために，従来の大規模並列計算機で用いられてきた相互結合網（T3E[4)]，Cavallino[5)]）と同様に高バンド幅，低レイテンシであることが求められる．SAN では，複数の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ため，パケットは複数のスイッチを経由して目的地に到達することになる．そのため，スイッチ群のトポロジが高バンド幅および低レイテンシを実現するための 1 つの鍵となる．

多くの SAN は，拡張性，耐故障性を重視しているため，様々なトポロジをとることができる．たとえば Myricom 社は Myrinet のトポロジとして Myrinet-Clos 網を推奨しているが，他のトポロジをとることも可能である[1)]．また，InfiniBand もトポロジに制限がない[3)]．そのため，トポロジが性能に与える影響について調査することが必要である．

これまでに並列計算機のトポロジについては，様々な提案，評価が行われてきた．しかし，SAN が任意のトポロジに適用できるルーティングアルゴリズム[6),7)]を用いているのに対し，並列計算機はトポロジに特化したルーティングアルゴリズムを用いることでルータの簡素化を図っている[8)]．そのため，同一トポロジをとる並列計算機の相互結合網と SAN では経路集合が異なる場合が多い．よって，SAN におけるトポロジ

† 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



(a) Fat ツリー (2,4,2) (b) Myrinet-Clos 網 (4×4)

図 1 間接網
Fig. 1 Indirect networks.

の性能を見積もるためには，並列計算機の相互結合網とは別に評価を行うことが望ましいと考えられる．また，並列計算機の相互結合網の研究では，ノード（プロセッシングユニット）間を直接リンクでつなぐ直接網とスイッチを介する間接網を，別個に議論する場合が多かった．しかし，直接網と間接網の違いはスイッチ（もしくはルータ）に接続しているノード数のみであり，SAN ではトポロジの許容範囲が大きいため両者の形式をとりうる．そのため，たとえば直接網の典型例であるトーラスと代表的な間接網である Myrinet-Clos 網や Fat ツリーとの比較が SAN のトポロジの評価には必要となる．

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[9] のトポロジを変化させ，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 SAN である．RHiNET-2 は代表的な SAN である Myrinet[1]，InfiniBand[3] と同様にトポロジに制限がないため，多くのトポロジを実装することができる．

以降，2 章では SAN で用いられる典型的なトポロジとその諸性質を述べ，3 章では本評価で用いた RHiNET-2 クラスタについて述べる．そして，4 章において RHiNET-2 クラスタによる評価結果を示し，5 章でまとめを述べる．

2. トポロジ

並列計算機や SAN で用いられてきた基本的なトポロジとしては，メッシュや k-ary n-cube（トーラス）があげられる[4),5)]．これらは対称性があるため，トラフィックの分散を比較的行いやすい利点がある．

表 1 各トポロジの直径と次数
Table 1 Diameter and degree of topologies.

トポロジ	直径	次数
2D Mesh (n×n)	2(n-1)	4
2D Torus (n×n)	n	4
Fat Tree (p,q,r)	2r	p+q
Myrinet-Clos (n×n)	2	n

一方，高速なバリア同期やデータ収集などの並列分散処理に適した階層網も提案されている．Fat ツリーはツリー構造を**図 1** (a) のように多重化したトポロジである．Fat ツリーはツリー構造が持つルート付近が混雑する問題を緩和しつつ，ツリーの階層構造を効果的に使う方法であり，QsNET[2] で採用されている．

Fat ツリーはツリーのルート方向へのリンク数 p，リーフ方向へのリンク数 q，および階層数 r の組 (p,q,r) により多少の柔軟性を持つ．その他の階層網としては超大規模並列計算機向けに Recursive Diagonal Torus（RDT），Shifted Recursive Torus（SRT）などが提案されている[8]．

また，Myricom 社はリンクのバンド幅を最大限に活かすために Myrinet のトポロジとして図 1 (b) に示した完全結合に近い Myrinet-Clos 網[1] を提唱している．図 1 (b) において，上層のスイッチにはホストが接続される場合とされない場合がある．

これらのトポロジは**表 1** に示した直径，次数により，大まかな特徴が分かる．

直径が小さいトポロジほど，一般的に，各パケットがネットワークに滞在する時間が短くなるためレイテンシが低く，結果的にスループットが高くなる．また，同数のポートを持つスイッチを用いた場合，次数の小さいトポロジほど多くの PC をスイッチに接続できる．そのため，同数のスイッチを用いた PC クラスタでは次数の小さいトポロジほど多くの PC を接続することができる．しかし，一般的に直径と次数はトレードオ

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)

フの関係にある．

表 1 のトポロジは，いずれも何らかの規則に従ってスイッチ間を接続する規則網であり，トラフィックの分散を比較的容易に行うことができる利点を持つ．しかし，物理的に不規則に配置された PC 群を結合する場合や規則的なトポロジにおいてリンク故障が起きた場合などでは，不規則なトポロジとして SAN を運用することもありうる．

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．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に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．図 2 においてスイッチ，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 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[9]，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．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 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 種類がある．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フリットに細分化して転送される．RHiNET-2 では 1 フリットは 8 Byte である．また，ヘッダとテイラは計 40 Byte（5 フリット）である．

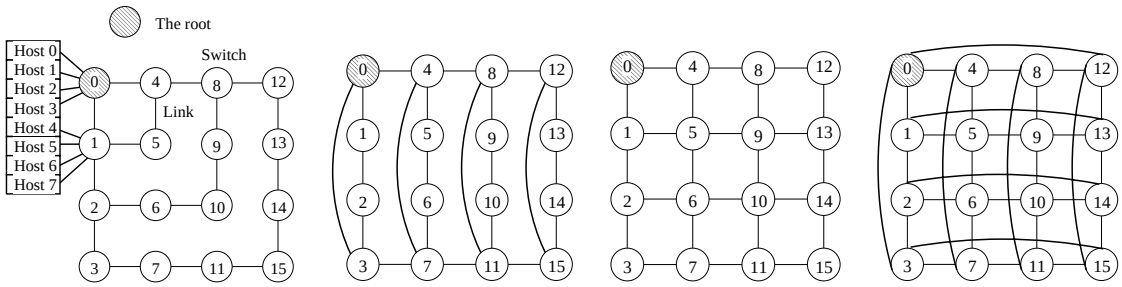
3.2 スイッチ RHiNET-2/SW

スイッチ RHiNET-2/SW[10] 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．現在，より安定した環境を構築するために光リンクの速度を 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 Gbps のスループットを持っているが，現在は 48 Gbps のスループットで稼働している．また，各ポートは 16 本の仮想チャネルを提供し，各仮想チャネルの持つ 4 KByte のバッファは，Go & Stop フローコントロールによりオーバーフローが起きないように制御する．

RHiNET-2/SW は，パケットの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分散方式の固定ルーティングを採用している[11]．そのため，ルーティングテーブルを変更することにより，RHiNET-2 クラスタは様々なトポロジ，ルーティングの組合せをとることができる．

3.3 システムソフトウェア

RHiNET-2 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[12] 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[13] 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 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．RHiNET-2 のソフトウェアレイヤの実装，評価についての詳細は文献 14) に述べられている．



(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.

4. 評　　価

4.1 条　　件

4.1.1 トポロジ

図 1，**図 3** に示した Myrinet-Clos 網，Fat ツリー，メッシュ，トーラスおよび若干の不規則性を持つトポロジ A，B の計 6 つのトポロジについて評価を行った．ホスト間の平均距離が大きく，経路が偏りやすいトポロジ A，B はこの 2 つの要因が性能に与える影響を調べる評価目的のために導入した．各スイッチのポートのうち，4 ポートは異なるホストに接続し，残りの 4 ポートは隣接スイッチに接続する，もしくは使用しない．ただし，図 1 (b) の Myrinet-Clos 網では，上層の 4 スイッチにホストを接続しない場合についての評価も行った．また，図 1 (a) の Fat ツリー (2,4,2) では最下層のスイッチにのみホストを接続し，中層のスイッチでは 6 ポートを隣接スイッチ間接続に使用した．同様に 6 スイッチを用いた Fat ツリー (2,4,1) も図 1 (a) の点線のように最下層のスイッチにのみホストを接続した．よって，図 3 に示した各トポロジでは 64 ホストの計算システム，図 1 の Fat ツリー，Myrinet-Clos 網では 16 もしくは 32 ホストの計算システムとなる．

本評価では，これらのトポロジを総ホスト数——16，32，64 ホスト——により，3 つに分類し，比較，検討を行う．

図 3 に示したトポロジにおいては特に記述していない場合，構造化チャネル法をデッドロックフリールーティングとして用いた．構造化チャネル法は，パケットが 1 ホップ進むごとに仮想チャネル番号を 1 増加させる最短型のルーティングである[15]．構造化チャネル法は多数の仮想チャネルを必要とする高性能なデッドロックフリールーティングの 1 つである．また，その他のトポロジではデッドロックが生じないため，仮想チャネル 1 本を用いた最短型ルーティングを用いた．構造化チャネル法は本来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，あらかじめ経路を 1 つ選択する必要がある[16]．

本実装では，メッシュ，トポロジ A，B，Fat ツリー，Myrinet-Clos 網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．一方，トーラスでは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—low port first—を選択した．これは，トーラスでは同一スイッチ間に最大 4 つの経路が存在するため，目的地のみをインデックスにする RHiNET-2/SW のテーブルフォーマット[11]ではメッシュなどに採用した経路分散を実装することが難しいことに起因する．

4.1.2 測定項目

基礎的な評価としてバリア同期時間，平均バンド幅，および，すべてのソフトウェアレイヤを含めた評価として NAS Parallel Benchmarks（NPB）2.3[17),18)] の実行時間を測定した．

バリア同期時間　全ホストによるバリア同期の平均時間とした．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[19)] 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[19)]，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[19)] を用いた．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サ

イズは本実装では 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．

トポロジを限定している並列計算機と異なり，SAN は多くの場合，トポロジに非依存のマルチキャストアルゴリズムが要求される．そのため，本稿では，様々な条件における基本性能を評価することに重点を置き，特定のトポロジや転送法に対する最適化は行わない．しかし，単純なランダムアルゴリズムを採用したことにより，性能に影響を与える要因について解析を行うことが可能である．

バンド幅 次のトラフィックパターン[20]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，各スイッチに接続されている 4 ホストの中で，送信のみを行うホストと受信のみを行うホストの計 2 ホストを測定に用いた．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 フリットは 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．

- bit reversal
- matrix transpose
- butterfly
- complement

NAS Parallel Benchmarks NPB 2.3[17),18)] の中から CG（Conjugate Gradient），IS（Integer Sort），LU（LU-decomposition）の実行時間を測定した．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，ホスト数は 16，32，64 とした．ただし，LU では仕様からクラス S では測定できないため，クラス W で測定した．また，64 ホストを接続したトポロジにおいて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．バリア同期の場合と同様に，様々な条件における基本性能を評価することに重点を置き，特定のトポロジや転送法に対する最適化は行わなず，ホスト番号順にプロセスを割り当てた．

4.2 バリア同期の測定結果

4.2.1 64 ホストのトポロジ場合

表 3 に 64 ホストによるバリア同期の実行時間を示す．表 3 において “Average Distance” はホスト間の経由スイッチ数の平均を示している．表 3 より，リンク数が多いトポロジの順――トーラス，メッシュ，トポロジ B，A の順――にバリア同期のレイテンシが小さく，その差は 8%である．これは，トポロジのリンク数が多いほど，1) パケットの平均ホップ数が小さくなり，かつ，2) 経路が分散されるためパケットの衝突が少なくなることに起因すると考えられる．

表 3 64 ホストのバリア同期時間（μsec）
Table 3 Execution time of barrier synchronization on 16 switches with 64 hosts (μsec).

Topology	Time	Average Distance
Topology A	47.83	4.1
Topology B	46.61	3.6
Mesh	45.61	3.5
Torus	44.09	3.0

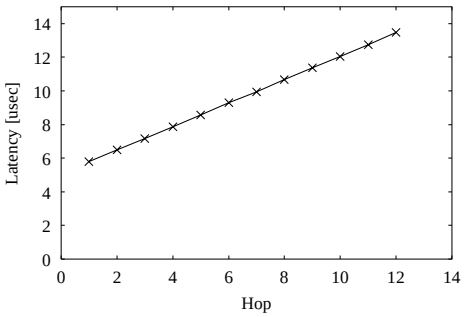


図 4 2 ホスト間のバリア同期時間
Fig. 4 Execution time of barrier synchronization on 2 hosts.

ここで，パケットのホップ数とコンテンションの与える影響について調べるために簡単な解析を行う．バリア同期時間におけるパケットのホップ数の影響を**図 4** に示す．図 4 は 2 ホスト間の経由スイッチ数を変化させた場合のバリア同期の実行時間を示している．

ここで，本稿ではバリア同期時間をパケット転送時間とホスト処理時間の 2 つに分ける．パケット転送時間を，各ステップにおいてボトルネックとなる最も到着の遅いパケットのヘッダが光リンクに注入された時間から，テイラが目的地ホストのネットワークインタフェースに到着するまでの時間の和と定義する．そして，その他のネットワークインタフェースでの処理を含めた時間をホスト処理時間と定義する．この定義により，パケット転送時間はパケットがスイッチもしくはリンクに滞在している時間を示す．よって，トポロジの選択はパケット転送時間のみに影響することになる．

パケット転送時間は，ネットワークインタフェース

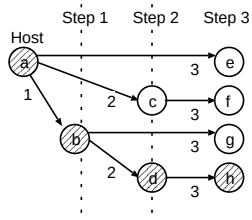


図 5 8 ホスト間のバリアの解放の例
Fig. 5 An example of barrier-release operation with 8 hosts.

からリンクへのパケット注入時間とパケットの経由スイッチとリンク遅延の和となる．パケットの注入時間 I はリンクのパケット転送容量が 4.8 Gbps☆であるため，

$$I = (136 \times 8/4.8/1000) = 0.226(\mu\,\mathrm{sec})$$

となる．一方で図 4 より，経由スイッチが 1 つ増えることに 0.697 μsec の遅延が発生していることが分かる．よって 2 ホストのバリア同期におけるパケット転送時間 T は，2 ホスト間のスイッチ数 s に対して

$$T(s) = 2I + 0.697s = 0.452 + 0.697s(\mu\,\mathrm{sec})$$

となり，ホスト処理時間の和は約 4.64 μsec となる．これより，64 ホストのバリア同期では 6 ステップ対が必要になるため，総ホスト処理時間は，約 27.85 μsec となる．

また，**図 5** に示したとおり，各ステップにおいて 1 つのホストが待つパケットはたかだか 1 つであり，ステップ内では異なるホストからのパケット転送は独立に行われる．さらに，ステップ数個のホスト間の依存関係は図 5 のホスト a，b，d，h 間のように 1 組発生するのみであり，これが各ステップのボトルネックになっていると見なすことができる．よって，ボトルネックになるパケットの平均ホップ数は，トポロジの平均距離と同じと考えることができる．

したがって，64 ホストのバリア同期時間に占めるパケット転送時間 T_{all} は，トポロジの平均距離を d とすると，ホスト注入時間，パケットのスイッチ，リンク移動時間とパケットの衝突による待ち時間 b の和

$$T_{all} = 6T(d) + b = 2.712 + 4.182d + b(\mu\,\mathrm{sec})$$

となる．なお，64 ホストのバリア同期においてはパケットのコンテンションが起こる点が，2 ホストのバリア同期と異なる．

パケット転送時間 T_{all} はバリア同期時間とホスト処理時間の差でも求められることから，その内訳を算出した結果を**表 4** に示す．表 4 において，inj，hops，cont はパケットをホストからリンクへの注入時間，スイッチとリンク移動遅延，パケットの衝突による待ち時間をそれぞれ表す．また，Host はホスト処理時間を示す．

☆ リンクのバンド幅は 6 Gbps であるが，そのうち 1.2 Gbps はスイッチインタフェースでパケットに付加される ECC の転送に使用される．

表 4 64 ホストのバリア同期時間の内訳（μsec）
Table 4 Itemized statement of barrier synchoronization time with 64 hosts (μsec).

Topology	Host	Pkt. Trans. Time		
		inj	hops	cont
Topology A	28	3	17	0
Topology B	28	3	15	1
Mesh	28	3	15	0
Torus	28	3	13	1

表 5 32 ホストのバリア同期時間（μsec）
Table 5 Execution time of barrier synchronization on 32 hosts (μsec).

Topology	Time	Average Distance
Ring (8)	36.51	3.0
Mesh (4×2)	35.00	2.8
Torus (4×2)	34.02	2.5
Myrinet-Clos (8)	33.60	2.3

表 4 より，いずれのトポロジにおいてもバリア同期時間に占めるパケット転送時間の割合は 40%程度であり，トポロジにより差が生じるパケットのスイッチとリンク移動時間，衝突の影響は 35%程度であることが分かる．特に，パケット転送時間に占めるスイッチとリンク移動時間は最大 86%と支配的であることが分かる．また，トーラスにおけるパケットの衝突による遅延はトポロジ A の場合に比べて大きい．トポロジ A はホスト間の距離の分散が大きいため，バリア同期の各ステップにおいて各ホストからパケットが注入される時間にばらつきが生じ，その結果，衝突による遅延が小さくなったと考えられる．

これらすべての結果より，性能の高いトポロジほどスイッチ，リンク移動時間が小さいため，バリア同期においてパケットホップ数の削減が最も重要であるといえる．

4.2.2 32 ホストのトポロジの場合

表 5 に 32 ホストによるバリア同期の実行時間を示す．表 5 では使用したホスト数は同じであるが，トポロジを構成するスイッチ数が異なるため，スイッチ数を括弧内に示した．表 3 と同様に，表 5 においてもホスト間の平均距離が短いトポロジほどバリア同期のレイテンシが小さいことが分かる．64 ホストの場合と同様にして，バリア同期時間の内訳を算出した結果を**表 6** に示す．

表 6 より，32 ホストの場合はパケット転送時間の割

表 6 32 ホストのバリア同期時間の内訳（μsec）
Table 6 Itemized statement of barrier synchoronization time with 32 hosts (μsec).

Topology	Host	Pkt. Trans. Time		
		inj	hops	cont
Ring (8)	23	2	10	1
Mesh (4×2)	23	2	10	0
Torus (4×2)	23	2	9	0
Myrinet-Clos (8)	23	2	8	0

表 7 16 ホストのバリア同期時間（μsec）
Table 7 Execution time of barrier synchronization on 16 hosts (μsec).

Topology	Time	Average Distance
Fat Tree (2,4,2)(14)	32.03	3.9
Fat Tree (2,4,1)(6)	27.68	2.5
Myrinet-Clos (8)	27.62	2.5
Ring (4)	26.33	2.0

合が約 30%と，64 ホストの場合に比べて小さくなっていることが分かる．これは，トポロジを構成するスイッチ数が減少したことにより，ホスト間の平均距離が小さくなったことが原因と考えられる．しかし，64 ホストの場合と同様にパケットのスイッチ，リンク移動時間が，衝突による待ち時間に比べて大きく性能に影響していることが分かる．

4.2.3 16 ホストのトポロジの場合

表 7 に 16 ホストによるバリア同期の実行時間を示す．表 7 では使用したホスト数は同じであるが，トポロジを構成するスイッチ数が異なるため，スイッチ数を括弧内に示した．また，表 7 において 14 スイッチの Fat ツリーでは 1 スイッチあたり 2 ホストを用いることで 16 ホストとした．

Fat ツリーにおいて，6 スイッチのものは 14 スイッチのものに比べて 16%もレイテンシを削減している．これはホスト間の平均距離が短いトポロジほどバリア同期のレイテンシが小さいことに起因すると考えられる．また，平均距離が等しいトポロジ——6 スイッチの Fat ツリーと Myrinet-Clos 網——では，リンク数が多い後者の方が若干レイテンシが小さいことが分かる．これは，リンク数が多いほど経路が分散できるためと考えられる．

4.3 バンド幅の測定結果

4.3.1 64 ホストのトポロジの場合

図 6 に 64 ホストのトポロジのトポロジのバンド幅を示す．図 6 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いる．図 6 より，トーラスはトポロジ A，メッシュに比べ最大 91%，47%のバンド幅向上を達成しており，トポロジ間の性能差が大きいことが分かる．これは，リンク数が多いほど，1) パケットの平均ホップ数が小さくなり，2) 経路が分散されることによりパケットの衝突が少なくなるためと考えられる．特に，ホスト間の平均距離が最も大きいトポロジ A が，トポロジ B に比べて高バンド幅であることから，後者の影響が大きいと考えられる．

一方で本評価ではリモート DMA 転送において，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，次のパケットを注入する．そのため，トポロジの平均パケットホップ数も性能に大きく影響したと考えられる．そこで，2 ホスト間のホップ数ごとに 1,792 Byte のパケットのバンド幅を測定した結果を**図 7** に示す．

図 7 より，バンド幅は 1 ホップ増えるごとに最大約 10 MByte 低下していることが分かる．本評価を含めた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

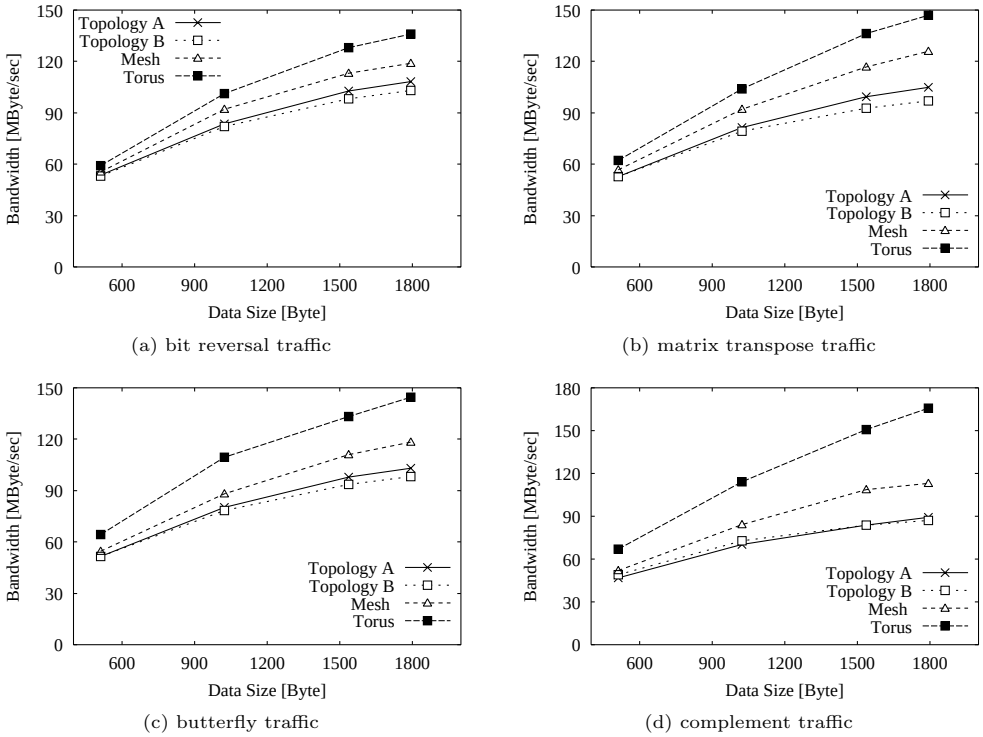
$$B = D/T$$

で測定される．RHiNET-2/SW において，チャネルバッファはパケットサイズに比べ十分に大きいため，パケットの衝突による待ち時間はパケット長によらず一定，つまり，データパケット，応答パケットとも同じと考えることができる．また，応答パケットは RHiNET-2 における最小パケット長—17 フリット，136 Byte—である．

これらより，2 ホスト間のバリア同期の場合と同様にして，1 パケットの転送時間をヘッダが出発地ホストからリンクに注入されてから，テイラが目的地ホストに到着するまでのパケット転送時間，ホスト処理時間，応答パケットの転送時間の項目別に算出した結果を**表 8** に示す．表 8 において s は 2 ホスト間の経由スイッチ数を表す．なお，1 つの通信ホスト組のバンド幅測定では，他のトラフィック負荷がないためパケットの衝突は起こらない．

表 8 より，バリア同期と異なりホップ数の影響は限られていることが分かる．一方で，パケットサイズが大きいため，パケットの注入時間が大きく影響していることが分かる．

次にパケットの衝突による性能の低下に焦点をあてる．トポロジ間の性能が大きい complement トラフィックにおける 1,792 Byte データのパケットの実測値と，パケットの衝突がないと仮定した場合のバンド幅を算出した結果を**表 9** に示す．パケットの衝突のない場合のバンド幅は，complement トラフィックにおける各通信ホスト間の距離を算出し，図 7 より得られたバンド幅の平均をとった．



(a) bit reversal traffic
(b) matrix transpose traffic
(c) butterfly traffic
(d) complement traffic

図 6 64 ホストのトポロジにおけるバンド幅
Fig. 6 Bandwidth of topologies on topologies with 64 hosts.

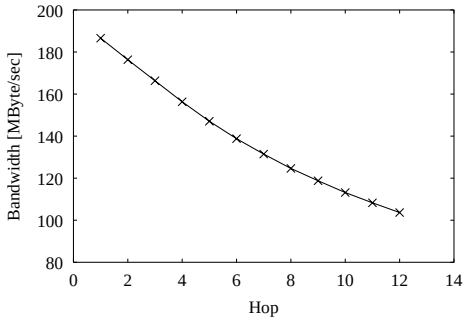


図 7 1,792 Byte データ転送における 2 ホスト間のバンド幅
Fig. 7 Bandwidth between two hosts under 1,792 Byte data.

表 8 1,792 Byte データのパケットの転送時間の内訳（μsec）
Table 8 Itemized statement of 1,792 Byte-data packet (μsec).

Total	Host	ACK Trans.		Pkt. Trans.	
		inj	hops	inj	hops
8.46+0.7s	5.19	0.23	0.35s	3.05	0.35s

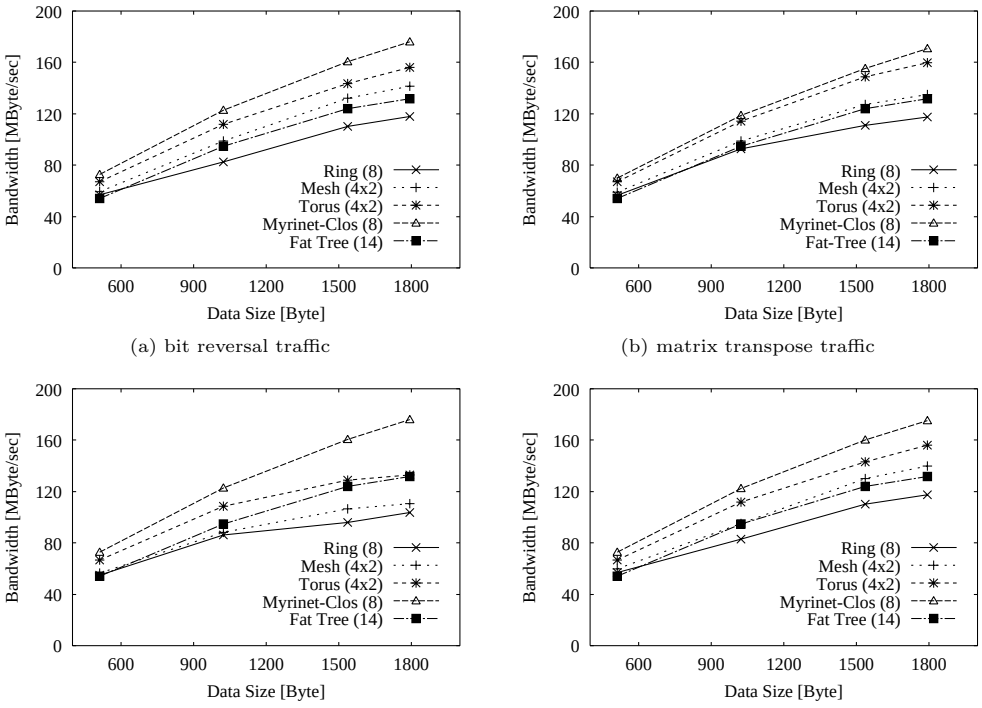
表 9 より，トーラスにおいて complement トラフィックの平均バンド幅は 3 ホップの場合の 2 ホスト間のバンド幅とほぼ等しいことが分かる．これは，トーラスにおける complement トラフィックでは，構造化チャネル法を用いた場合，すべてのパケットを 3 ホップで他の経路対のパケットとの衝突なしに転送することができるためである．また，表 9 より，ホスト間の平均距離がトポロジのバンド幅に与える影響が最大 17%にとどまるのに対し，パケットの衝突による性能低下が最大 41%と大きいことが分かる．

表 9 Complement トラフィックにおける 1,792 Byte データのパケットの衝突による影響（MByte/sec）
Table 9 The impact of contention on 1,792 Byte-data packet under complement traffic (MByte/sec).

Topology	Ideal	Results	Drop ratio
Topology A	137.74	89.40	35%
Topology B	148.02	87.11	41%
Mesh	148.02	113.03	24%
Torus	166.36	165.83	0%

4.3.2 32 ホストのトポロジの場合

図 8 に 32 ホストのトポロジのバンド幅を示す．図 8 ではトポロジを構成するスイッチ数が異なるため，スイッチ数を括弧内に示した．図 8 より，パケットの平均ホップ数が小さく，リンク数の多いトポロジである Myrinet-Clos 網が最も高い性能を示している．また，14 スイッチを用いた Fat ツリーは 8 スイッチを用いたトーラス，メッシュに比べてバンド幅が低い．これは，単にスイッチ数を増やしても，1) 平均ホップ数が増加し，2) ルート付近に経路が偏ってしまう場合には



(a) bit reversal traffic (b) matrix transpose traffic

(c) butterfly traffic (d) complement traffic

図 8 32 ホストのトポロジにおけるバンド幅
Fig. 8 Bandwidth of topologies on topologies with 32 hosts.

表 10 64 ホストのトポロジにおける CG，LU，IS のベンチマークの実行時間（sec）
Table 10 Execution time of CG, LU and IS benchmarks under topologies with 64 hosts (sec).

	CG.S.16	LU.W.16	IS.S.16	IS.S.32	IS.S.64
T. A	0.20450	7.76525	0.01850	0.02091	0.11751
T. B	0.20400	7.77257	0.01838	0.02052	0.11674
Mesh	0.20325	7.76200	0.01836	0.02052	0.11673
Torus	0.20367	7.75225	0.01820	0.02008	0.11631

表 11 32 ホストのトポロジにおける CG，LU，IS の実行時間（sec）
Table 11 Execution time of CG, LU and IS benchmarks under topologies with 32 hosts (sec).

	CG.S.16	LU.W.16	IS.S.16	IS.S.32
Ring (8)	0.16300	7.44067	0.01629	0.01736
Mesh (4×2)	0.16300	7.40233	0.01620	0.01735
Torus (4×2)	0.16233	7.43900	0.01611	0.01702
Myri.Clos (8)	0.20200	7.79100	0.01840	0.01994

必ずしも性能向上につながらないことを示している．

4.4 NAS Parallel Benchmarks

次に NPB 2.3 の CG，LU，IS の評価結果を**表 10**，**表 11** に示す．表 10，11 において，“IS.S.16” は IS を 16 ホストでクラス S において実行した場合の実行時間を示す．本評価では，構造化チャネル法と同一の物理経路を使用した DL ルーティングを用いた．文献 11) において，DL ルーティングと構造化チャネル法の性能差はほとんどないことが報告されている．

表 10 より，トーラスはトポロジ A に比べて，実行時間を最大 4.1%短縮していることが分かる．また，表 11 において，CG ではトーラスが Myrinet-Clos 網に比べて最大 20%の性能向上を達成している．これは，本実装における CG では近隣ホスト間の通信が多いことによると考えられる．つまり，Myrinet-Clos 網，Fat ツリーにおける連続した番号のホスト間通信は，トーラスの場合に比べて，遅延が大きく，これが性能に影響したと考えられる．

表 12 トポロジ A における CG，IS の実行時間の内訳（sec）
Table 12 Itemized statement of 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

ここで，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 12** に示す．表 12 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 12 より，通信時間が全体の性能を支配していることが明らかである．しかし，通信時間の中で，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，トポロジの影響は限られている．また，計算時間が相対的に小さいために，ホスト数を 16，32，64 と増加させているにもかかわらず，実行時間があまり変化しなかったと考えられる．

4.5 実行結果の比較，有効性

図 6，8 より，bit reversal などのトラフィックにおけるトポロジのバンド幅の差は最大 91%と大きい．一方で，表 3，5，7，10，11 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，本規模の PC クラスタの実用においては，トポロジ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，トポロジ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，光インターコネクトを用いて机上で利用中の PC を接続するを想定している．しかし，本評価においては，光インターコネクトは通常の 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，RHiNET-2 は，次のように高バンド幅，低レイテンシを実現するための SAN の機能を装備している．

- RHiNET-2/NI においてユーザレベルゼロコピー通信をハードワイヤードロジックで実現
- RHiNET-2/SW において 4 KByte のバッファを備えた 16 本の仮想チャネルと Go & Stop フローコントロールの採用により，バッファオーバフローを防ぎ効率的なパケット転送を実現
- デッドロックフリールーティングを用いることで，パケット間のデッドロックの検出，復旧にかかるオーバヘッドを除去

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．

4.6 既存のシミュレーション研究との比較

確率モデルシミュレーションを用いた並列計算機向けトポロジの評価はさかんに行われてきた．これらは，対象とする結合網のサイズ，1 スイッチに接続されているホスト数などが本評価条件と大きく異なるものが多いため，単純に比較することは難しい．一方で，我々が行った SAN のシミュレーション[7] は，条件—物理経路，トポロジ，ネットワークサイズ，ホスト数，バーチャルカットスルー方式—が本評価環境と似ている．シミュレーションにおけるスループットは，バンド幅の測定方法と異なり，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[20]．したがって，経路長はスループット測定において，間接的な影響にとどまるため，シミュレーションにおけるトポロジの性能差は本測定に比べて小さくなる傾向がある．しかし，本評価結果ではメッシュはトポロジ A に比べて最大 30%のバンド幅向上にとどまるのに対し，このシミュレーション結果[7] は最大 35%のスループット向上を達成したと報告している．よって，本測定におけるトポロジが性能に与える影響は，シミュレーションによる既存の研究報告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抑えるために，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[7),16]．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（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 クラスタにおけるトポロジの影響はシミュレーション結果と比べて小さくなったと考えられる．

5. ま と め

64 ホストを用いた RHiNET-2 クラスタにおける様々なトポロジのバリア同期時間，バンド幅，およびすべてのソフトウェアレイヤを含めた評価として NAS Parallel Benchmarks（NPB）2.3 から CG，LU，IS の実行時間を測定した．評価の結果，バリア同期時間はトポロジ間の性能差が 8%であり，バリア同期時間に占めるパケット転送時間は最大 40%であった．一方，2 次元トーラスは 2 次元メッシュに対し最大 47%のバンド幅向上を示し，また，Myrinet-Clos 網は同数のホストを接続した Fat ツリー，2 次元トーラスに対して最大 48%のバンド幅向上を示した．また，トポロジの平均距離がバンド幅に与える影響は 17%にとどまる一方，パケットの衝突による影響は 41%であった．NPB 2.3 の測定時間についてはトーラスが Myrinet-Clos 網に比べて最大 20%の性能向上を達成した．

謝辞 本研究を行うにあたり RHiNET-2 クラスタに関して貴重なご意見をくださった慶應義塾大学理工学部西宏章助手，河野賢一氏，上樂明也氏，北村聡氏に感謝いたします．

参 考 文 献

1) Myricom, Inc.. http://www.myri.com/

2) Petrini, F., Feng, W. and Hoisie, A.: The Quadrics network (QsNet): high-performance clustering technology, *Proc. Hot Interconnects*, pp.125–130 (2001).
3) I.T.Association: InfiniBand architecture. Specification Volumen 1, Release 1.0.a, available from the InfiniBand Trade Association (2001). http://www.infinibandta.com
4) Scott, S.L. and T.Horson, G.: The Cray T3E network: adaptive routing in a high performance 3D torus, *Proc. Hot Interconnects IV*, pp.147–156 (1996).
5) Carbonaro, J. and Verhoorn, F.: Cavallino: The teraflops router and NIC, *Proc. Hot Interconnects Symposium IV*, pp.157–160 (1996).
6) Rodeheffer, T. and Schroeder, M.: Automatic reconfiguration in Autonet, Technical Report SRC research report 77, DEC (1991).
7) Koibuchi, M., Jouraku, A., Watanabe, K. 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).
8) 天野英晴：並列コンピュータ，昭晃堂 (1996).
9) 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).
10) 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).
11) Koibuchi, M. and Watanabe, K., Kono, K., Jouraku, A. and Amano, A.: Performance Evaluation of Routing Algorithms in RHiNET-2 Cluster, *Proc. IEEE International Conference on Cluster Computing*, pp.395–402 (2003).
12) 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).
13) 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).
14) 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).
15) Merlin, M.P. and Schweitzer, J.P.: Deadlock Avoidance in Store-and-Forward Networks, *IEEE Trans. Comput.*, Vol.COM-28, No.3, pp.345–354 (1980).
16) Koibuchi, M., Jouraku, A. and Amano, H.: The Impact of Path Selection Algorithm of Adaptive Routing for Implementing Deterministic Routing, *Proc. International Conference on Parallel and Distributed Processing Techniques and Applications*, pp.1431–1437 (2002).
17) 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).
18) 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).
19) 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).
20) Duato, J., Yalamanchili, S. and Ni, L.: *Interconnection Networks: an engineering approach*, Morgan Kaufmann (2002).

(平成 16 年 1 月 29 日受付)
(平成 16 年 7 月 12 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．

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．