

PC クラスタにおける VLAN イーサネットのトポロジの評価

廣 安 知 之^{†1} 渡 辺 崇 文^{†2,*1} 中 尾 昌 広^{†2}
大 塚 智 宏^{†3,*2} 鯉 淵 道 紘^{†4}

イーサネットにおいて、VLAN ルーティング法を用いることで、様々なトポロジを採用することができるようになった。しかし、これまで大規模 PC クラスタにおけるイーサネットのトポロジの性能評価はほとんど行われていない。そこで、本稿では 450 コア 225 台のホストで構成される SuperNova クラスタ、および、528 コア 66 台のホストで構成される Misc クラスタにおいて、トポロジが性能に与える影響について調べた。既存の PC クラスタにおいて最小限のシステム更新で VLAN ルーティング法を実装するために、(1) スイッチにおいてフレームに VLAN タグを付与し、(2) MAC アドレステーブルの学習により各スイッチの経路管理を行う汎用性の高い方法を提案し、実装した。評価結果より、SuperNova クラスタにおいて 8 台の安価な 48 ポートスイッチをトラス、完全結合トポロジで接続した場合の High-Performance LINPACK benchmark (HPL) の性能は、336 ポートの高価なノンブロッキングスイッチを 1 台使用した場合とほぼ同等であることが分かった。さらに、Misc クラスタにおいて完全結合トポロジにおける NAS Parallel Benchmarks の性能は、リンク集約化を行ったツリートポロジに比べて最大 909.2%向上することが分かった。

Performance Evaluation of VLAN-Ethernet Topologies on PC Clusters

TOMOYUKI HIROYASU,^{†1} TAKAFUMI WATANABE,^{†2,*1}
MASAHIRO NAKAO,^{†2} TOMOHIRO OTSUKA^{†3,*2}
and MICHHIRO KOIBUCHI^{†4}

VLAN routing method allows us to employ various topologies in Ethernet. However, their evaluation on real large-scale PC clusters was rarely done. In this paper, we investigate the impact of topology on the performance of PC clusters called SuperNova that consists of 225 450-core hosts and called Misc that consists of 66 528-core hosts. To minimize the system modification of ex-

isting PC clusters, we implement it by (1) adding the VLAN tag to frames at switches, and (2) managing routing paths by address self-learning of switches. Evaluation results show that the performance of torus and completely connected topologies with eight 48-port switches is comparable to those of an ideal 1-switch (full crossbar) network in High-Performance LINPACK Benchmark (HPL) on SuperNova cluster. In Misc cluster, the completely connected topology achieves up to 909.2% improvement on their execution time of NAS Parallel Benchmarks compared with that of tree topology with link aggregation.

1. はじめに

イーサネット (Ethernet) は、管理の容易さ、高い耐故障性、安価なハードウェアなどの利点から、ローカルエリアネットワーク (LAN) だけでなく、広域ネットワークや PC クラスタのインタコネクタとしても幅広く採用されている。特に、ギガビットイーサネット (Gigabit Ethernet) の普及、ツイストペアケーブルを用いる 10GBASE-T の標準化 (IEEE 802.3an-2006) などにより、イーサネットはハイパフォーマンスコンピューティング (HPC) 分野において、Myrinet などの高価なシステムエリアネットワーク (SAN) に迫るインタコネクタとして主流になりつつある。

しかし、イーサネットを用いた PC クラスタの多くは単純なツリートポロジを採用している。これは、基本的にイーサネットがループ構造を含むトポロジを許していないためである。ツリートポロジにはトラフィックがツリーのルート付近に偏りやすいという欠点があるため、リンク集約化 (IEEE 802.3ad) などによってルート付近のリンクを強化するのが一般的である。しかし、クラスタが大規模になると、1 つの集約化されたリンクを構成する

†1 同志社大学生命医科学部

Department of Life and Medical Sciences, Doshisha University

†2 同志社大学大学院

Graduate School of Engineering, Doshisha University

†3 慶應義塾大学大学院理工学研究科

Graduate School of Science and Technology, Keio University

†4 国立情報学研究所/総合研究大学院大学/JST

National Institute of Informatics/SOKENDAI/JST

*1 現在、株式会社インターネットイニシアティブ

Presently with Internet Initiative Japan Inc.

*2 現在、慶應義塾インフォメーションテクノロジーセンター

Presently with Information Technology Center, Keio University

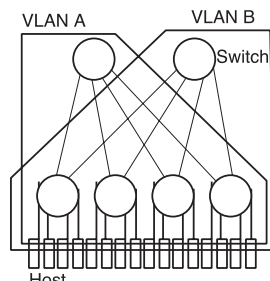


図 1 VLAN ルーティング法
Fig.1 VLAN routing method.

ポート数は限られている場合が多いためツリートポロジの欠点を補いきれなくなる。また、リンク集約化のためにスイッチのポートを多数占有してしまうため、限られたホストの数しかスイッチに接続できないといった制約が生じる。これらのことから、ユーザやアプリケーションの要求に応じたトポロジ・ルーティングを採用している SAN や並列計算機の相互結合網に比べて、イーサネットを用いた PC クラスタは大規模化には向かないとされてきた。

リンク集約化以外にも、スイッチ間に複数リンクを接続することでパンド幅を向上させる方法として、IEEE 802.1Q 標準のタグ VLAN 技術を応用した VLAN ルーティング法^{1),2)}が提案されている。

VLAN 技術は本来、同じ物理ネットワークに接続されたホストの集合を、複数の論理的なグループに分割するために用いられるが、VLAN ルーティング法ではこれをネットワークのスループット向上のために用いる。図 1 のように、各ホストが複数の VLAN グループのメンバになるようにしておき、各 VLAN にそれぞれ異なるリンク集合を割り当てる。ここで、各 VLAN ネットワークのトポロジはツリー構造となっているため、ブロードキャストストームは発生しない。上記の方法により、すべてのホストがどの VLAN を用いても互いに通信でき、VLAN を選択することで複数の経路を切り替えて使うことが可能になる。

しかし、TOP500 スーパーコンピュータのランキング³⁾において上位 500 台の中でギガビットイーサネットを用いたシステムが 56%と過半数になっているにもかかわらず、これらがトーラストポロジなどのループを含むトポロジを採用した報告はほとんどない。

これは、現時点において運用されている PC クラスタのホスト、システムソフトウェアが VLAN 技術に対応していない点が一因と考えられる。さらに、最近の高性能 PC クラスタにおいて、ループを含むトポロジを採用することにより、ツリートポロジに比べてどの程度

性能が向上するのか、定量的な評価結果がないことも大きな原因と考えられる。

そこで、本稿では、(1) MPI 通信ライブラリが VLAN 技術に対応しておらず、(2) 静的な MAC アドレスのエントリ数が 100 個ときわめて少ないスイッチを用いた既存の大規模 PC クラスタにおいて、システムの更新をできるだけ抑えるように VLAN ルーティング法を実装した。そして、マルチコアプロセッサを用いた Misc クラスタと従来のシングルコアプロセッサを用いた SuperNova クラスタの 2 種類のホスト構成の PC クラスタにおけるトポロジが与えるシステム性能について明らかにする。

以下、2 章で関連研究を述べ、3 章において VLAN ルーティング法の実装について述べ、4 章で、2 種類の PC クラスタの概要、ならびに評価結果を示す。最後に 5 章でまとめを述べる。

2. 関連研究

イーサネットにおいて VLAN を用いてホスト間に複数の経路を設定し、ループ構造を含む様々なトポロジを利用できるようにするルーティング技術は国内外でほぼ同時期に提案された^{1),2)}。工藤らが提案した VLAN ルーティング法¹⁾では、図 1 のようにループを含まない各リンク集合にそれぞれ異なる VLAN を割り当てることで、ブロードキャストストームを避けつつ同一スイッチ間に複数経路を実現する。ループを含むトポロジにおけるルーティングアルゴリズムは、リンク間の循環依存を除去する必要があるためデッドロックフリールーティングが必要となる^{4),5)}。

ループ構造を防ぐために、IEEE 802.1D STP (Spanning Tree Protocol), 802.1D-2004 RSTP (Rapid STP) があるが、これらは VLAN 処理とは独立に行われるため併用することはできない。802.1Q-2003 MSTP (Multiple STP) と Cisco Systems' PVST (Per VLAN Spanning Tree) は VLAN を扱うことができるため VLAN ルーティング法に有益だが、すべての安価なスイッチにおいて採用されているわけではない。

VLAN 技術を利用して PC クラスタのインタコネクトを構築する手法は国内を中心に活発に議論され、三浦らの研究⁶⁾では、MAC アドレスから VLAN ID を決定しタグ付けを行うための Linux 用デバイスドライバを開発し、TCP/IP を用いた VLAN ルーティング法の実現している。この手法では、MAC アドレスに基づいた VLAN ID の制御とすることで、送信先に応じた VLAN の選択をドライバに任せることができるようになるため、上位レイヤのソフトウェア環境に手を加えることなく VLAN ルーティング法を実現できる。

これに対し我々は、様々なトポロジにおける VLAN の割当て方法や、スイッチにおいて

VLAN タグ付けを行うことで、システムソフトウェアが VLAN 技術をサポートしていない場合にも VLAN ルーティング法を利用できるようにする手法^{7),8)}を提案し、32 台のホストで構成される PC クラスタにおいて評価を行った。

VLAN 技術を用いずに、静的にホストの MAC アドレスを登録することでルーティングを行う方法も検討されているが、ブロードキャストストームが発生した場合の対処、ならびに各スイッチから宛先への出力ポートが入力ポートによらずに定まるため利用可能なルーティングアルゴリズムが限定される。ループ構造を扱うことができる Transparent ブリッジ⁹⁾、複数経路を扱う spanning tree alternate routing (STAR)¹⁰⁾なども提案されているが、VLAN ルーティング法は既存のイーサネットの機能により実現できる点で異なる。

また、レイヤ 3 ルーティングを用いることにより、VLAN ルーティング法と同等の並列計算向けトポロジを構築することができる。しかし、(1) レイヤ 3 ルーティングをサポートするスイッチは高価である、(2) レイヤ 3 ルーティングのオーバーヘッドはレイヤ 2 スwitching に比べて大きい場合が多い、(3) レイヤ 3 ルーティングのスイッチ設定はトポロジによっては設定が煩雑になる、などの問題がある。

IBM や Cisco Systems などが提唱し、現在 IEEE などで 10 ギガビット・イーサネットの拡張仕様として標準化作業が進められている次世代イーサネットのアーキテクチャである Data Center Ethernet (DCE) は、マルチパスルーティングにより複数の最短パスを提供可能などの点からクラスタのノード間通信インターコネクトとしても将来利用される可能性がある。ただし、VLAN ルーティング法とは、レイヤ 3 ルーティングと同様にスイッチのコストの点、対象システムが異なる点で、現時点では詳細な比較は難しい。

3. VLAN ルーティング法の実装

最小限のシステム更新で VLAN ルーティング法を既存の一般的な PC クラスタに実装するために、本章では、(1) 各スイッチは、ホストから注入される VLAN タグのないフレームに VLAN タグを挿入し⁷⁾、(2) MAC アドレスの学習により各スイッチの経路管理を行う汎用性の高い方法を提案し、実装する。

3.1 スイッチにおける VLAN タグ付け

ここでは、文献 7) の方法に従って、ホストと接続されたスイッチポートでは、ホストからの入力フレームに VLAN タグを付加し、ホストへ出力するフレームから VLAN タグを除去する。これを行うため、ホストと接続された各スイッチポートに対し、以下の 2 種類の設定を行う。

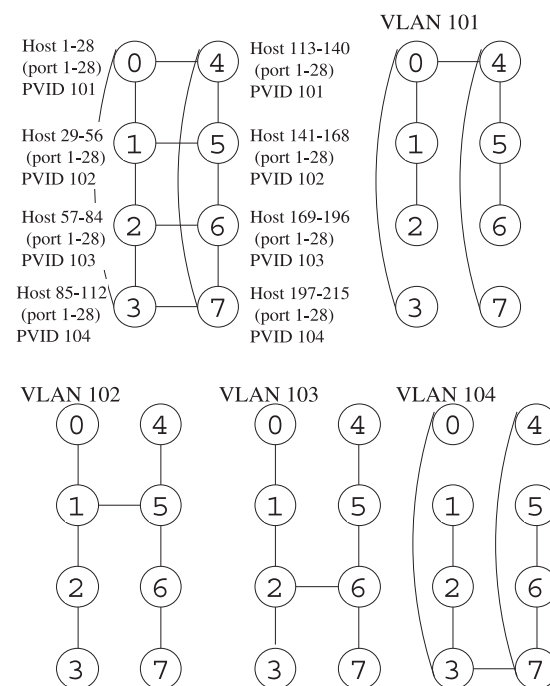


図 2 スイッチで VLAN タグ付けを行うルーティングの例
Fig. 2 Example of switch-tagged VLAN routing.

- Port VLAN id (PVID) として、接続されたホストがフレームを送信する際の経路として使う VLAN の ID をあらかじめ設定する。
- 各リモートホストから送られてくるフレームのタグを除去するため、ポートをネットワーク全体で使われる全 VLAN の“タグなし”メンバとしておく。

この例を使用した PC クラスタの構成図を図 2 に示す。図中の円はスイッチを表している。図 2 において、ホスト 1~28 から送出されたフレームは、スイッチ 0 の入力ポートにおいて VLAN タグ #101 を付与され、すべての宛先について VLAN #101 内によってルーティングされる。そして、宛先ホストに接続しているスイッチの出力ポートにおいて VLAN タグ #101 を除去する。一方、ホスト 29~56 から送出されたフレームも同様の方法で VLAN #102 によってルーティングされる。

上記の方法により、ホスト側で VLAN がサポートされていなくても、様々なトポロジにおいて全ホストの相互通信が可能になる。

3.2 スイッチにおける MAC アドレスの管理

スイッチは通常、以下のように MAC アドレスを学習する。まず、スイッチがフレームを受信した際、スイッチはその送信元 MAC アドレスを参照し、入力されたポート番号とともに MAC アドレステーブルに登録する。次に、宛先 MAC アドレスを参照し、テーブルを引いてそのアドレスのエントリがあるかどうかを調べる。エントリが見つからなかった場合、スイッチは VLAN メンバとなっている全ポートからフレームを出力するため（これをフラディングと呼ぶ）、最終的にフレームは宛先ホストへ到達する。この宛先 MAC アドレスのエントリは、宛先ホストからの返信フレームを受信した際に登録されるため、以後はフラディングをとまわずにフレームの交換が実現されるようになる。

しかし、本手法では往復の経路で使用する VLAN が異なるため、各スイッチにおける MAC アドレステーブルの管理が 1 つの課題となる。

この課題は、静的に MAC アドレスをスイッチに登録することで解決することができる。しかし、スイッチの多くは静的に登録できる MAC アドレス数が限られているため、大規模 PC クラスタには適用できない場合がある。たとえば、本評価に用いた Dell 社 PowerConnect 6248 はたかだか 100 個の MAC アドレスのみ静的に登録可能であり、文献 7) の方法を実装することができない。しかし、MAC アドレスの学習を用いることで最大 8,000 個の MAC アドレスのエントリを持つことが可能である。

そこで、本実装ではスイッチにおいてタグ付けを行う VLAN ルーティング法において次のように MAC アドレスの学習を実現した¹¹⁾。

- (1) 全 VLAN に対応する仮想インタフェースを各ホストにおいて vconfig などを使って作成する。たとえば図 2 の場合、VLAN 101 ~ 104 を用いるため、各ホストにおいて eth0.101 ~ eth0.104 までを作成する。
- (2) VLAN ごとに一意のネットワーク (IP) アドレスを与え、VLAN ごとに別々のセグメントに属するように各ホストの仮想インタフェースに IP アドレスを割り振る。
- (3) 仮想インタフェースごとに、ICMP または UDP メッセージを 1 度ブロードキャストする。

ステップ (3) では、各ホストにおいて、たとえば各 VLAN セグメント内で全ホストに対して ping (ICMP echo req.) を送信することで実現することもできる。これにより、各スイッチにおいて、各 VLAN のアドレステーブルに送信ホストの MAC アドレスが登録さ

れる。

本 MAC アドレス登録方式はスイッチの MAC アドレス学習のみが目的であり、MPI など発生する並列計算の通信レイヤ、通信経路には影響を与えない。

なお、PC クラスタはホストの追加、削除は一般的に、LAN 環境に比べて頻繁ではない。そのため、スイッチにおいて学習した MAC アドレステーブルの保持時間を定める Aging Time を大きくすることが現実的である。本評価で使用した PowerConnect 6248 では最大値である 100000 秒 \approx 11.6 日とした。そのため上記の MAC アドレスの学習手続きを 11.6 日に 1 度行うことでスイッチの MAC アドレステーブルを維持することができる。なお、商用のギガビットイーサネットスイッチによっては保持期間を無限に設定することができる。

4. 評価

本章では VLAN ルーティング法により実現された様々なトポロジの評価結果を示す。本実験では、従来のシングルコアプロセッサを用いた SuperNova クラスタの一部と、マルチコアプロセッサを用いた Misc クラスタの 2 つを用いた。これらは、現在、同志社大学に設置、運用されている。

4.1 SuperNova クラスタ

4.1.1 システム構成

SuperNova クラスタは 1 台の Force10 E1200 スイッチを用いて 256 台のホストを接続することで、2003 年の TOP500 ランキング³⁾ において 93 位となった大規模計算システムであり、今回の実験では、225 台のホストと 8 台の 48 ポートのギガビットイーサネットスイッチ (Dell 社 PowerConnect6248) で構成した。

表 1 にホストの仕様を示す。SuperNova クラスタでは、図 2 に示した VLAN #101 の単純ツリー、図 2 に示した 4×2 トーラス (次元順ルーティング) (3-bit hypercube)、完全結合 (次元順ルーティング)、 8×1 リング、 4×2 メッシュ (次元順ルーティング) の

表 1 SuperNova におけるホストの仕様
Table 1 Specifications of each host of SuperNova.

CPU	AMD Opteron 1.8 GHz \times 2
Memory	DDR 333 MHz 2 GB
NIC	Broadcom BCM95704A7 1000BaseT
NIC driver	Broadcom Tigon3
OS	Debian GNU/Linux 4.0
Kernel	2.6.18-4-amd64

表 2 SuperNova クラスタのトポロジの特徴
Table 2 Statics of topologies on SuperNova cluster.

	直径	Bi. BW (Gbps)	VLAN 数
Tree (n link)	6	n	1
Compl (n link)	2	$16n$	8
Mesh (n link)	5	$2n$	4
Torus (n link)	4	$4n$	4

各トポロジについて、スイッチ間のリンク数を 1 本からスイッチの設定の上限である 8 本までに変化させて評価を行った。これらの典型的なトポロジにおける VLAN の割当ては文献 12) に述べた直感的方法を用いた。また、スイッチ間リンク数を増やす場合は単に Link Aggregation を使った。そのため、スイッチ間リンク数が増加した場合も使用する VLAN 数は一定である。なお、Tree の場合は 1 つの VLAN のみを用いた VLAN ルーティング法を使用した。VLAN 操作の遅延、スループット面でのオーバーヘッドはほとんどない旨、文献 12) において確認している。これらのトポロジの静的な特徴を表 2 にまとめる。直径はホスト間通信における最大経路スイッチ数、“Compl (n link)” はスイッチ間リンク数が n 本である完全結合トポロジ、“Bi. BW” はバイセクションバンド幅を表す。

また、各並列ベンチマークは、MPICH1.2.7p1 を用いた IP パケットによりプロセス間通信を行い、ホスト-スイッチ間のリンク数は 1 本である。

すべてのトポロジにおいて各スイッチは IEEE 802.3x リンクレベルフロー制御を用いており、いずれのトポロジにおいてもリンク集約化は送信元、宛先ホストの IP アドレス、UDP/TCP のポート番号でリンク間のトラフィック分散を行った。

なお、本評価で用いた Dell 社 PowerConnect 6248 ギガビットイーサネットスイッチは、ノンブロッキングであり、Tperf¹³⁾ を用いた測定結果から、ポートあたり 939 Mbps (TCP) のバンド幅を達成することを確認しており、既存の他の商用ギガビットイーサネットスイッチの性能⁸⁾ と比べて遜色ないことを確認している。また、このスイッチはレイヤ 3 の機能も含んでいるが前章で述べたレイヤ 2 の機能のみを用いて実装している。

4.1.2 HPL の評価結果

High-Performance LINPACK Benchmark (HPL)¹⁴⁾ は、分散メモリ型並列計算機用のベンチマークソフトウェアであり、ガウス消去法を用いた密行列連立一次方程式を解き、その速度を Flops 値で評価する。

HPL ではシステムの特성에あったパラメータを設定することが可能である。今回計測に

表 3 主な HPL のパラメータ
Table 3 The main parameters of HPL.

	SuperNova	Misc
N	180000	234960
NB	240	220
P, Q	18, 25	6, 88
BCAST	1ring Modified	1ring

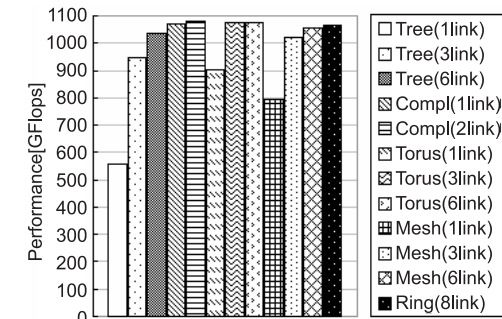


図 3 SuperNova における HPL の結果
Fig. 3 HPL results of SuperNova.

利用した HPL の主要なパラメータを表 3 に示す。

HPL 性能評価のプログラムとしては HPL1.0a を用い、数値演算ライブラリには GotoBLAS1.22 を用いた。HPL1.0a, MPICH1.2.7p1 のコンパイルは pgcc/pgf 7.1 を用い、最適化オプションは -fastsse -tp k8-64 とした。GotoBLAS1.22 のコンパイルには gcc4.2.2/pgf7.1 を用いた。

HPL の評価結果を図 3 に示す。Tree は図 2 に示した VLAN #101 の単純ツリー、Compl は完全結合、Torus は 4×2 トーラス、Mesh は 4×2 メッシュ、Ring は 8×1 リングの各トポロジを示し () 内はスイッチ間リンク数である。HPL では数値解の精度が要求される。SuperNova クラスタにおける計測では、Tree (1link), Compl (2link), Mesh (1link), Mesh (6link), Ring (8link) において並列計算により求めた解の精度確認時にエラーが生じた。ただし、HPL のパラメータであるブロックサイズなどを変化させることで Tree (1link) ではエラーが生じないことを確認している。しかし、この場合は性能が低下するため、トポロジ間の比較に重点を置き、Torus (6link) で実行に成功したパラメータにおいて

表 4 スイッチ間リンク 1 本あたりの性能の比較

Table 4 Comparison of performance per link between switches.

	Tree 6link	Compl 1link
Links between switches	42	28
Performance (GFlops)	24.6	38.2

比較を行った。

図 3 より、提案手法を実装した Compl (2link) において Tree (1link) に比べて最大 93.5%性能向上することが分かった。Tree (6link) の結果と Compl (2link) の実行性能を比較すると、その差は 2.9%であり、HPL においては単純ツリートポロジにてリンク集約化技術を用いれば十分な性能向上が可能であることが分かる。Compl (2link) は Tree (6link) に比べてパイセクションバンド幅が約 5 倍であるため、HPL に関してはパイセクションバンド幅が性能に直結していないことが分かる。

ただし、スイッチ間リンク総数を考慮してトポロジ間の性能を比較した場合、Tree (6link) がスイッチ間リンク総数 42 本で 63.8%の実行効率を達成しているのに対し、Compl (1link) は同 28 本で 66.0%の実行効率を達成、Torus (3link) は同 36 本で 66.3%の実行効率を達成、Mesh (3link) は同 30 本で 63.1%の実行効率を達成できている。これらのトポロジはリンク 1 本あたりの性能で Tree (6link) を上回っており、費用対効果の面で Tree (6link) に勝っているといえる。たとえば、Compl (1link) の場合、リンク 1 本あたりの性能は 38.2 GFlops であり、Tree (6link) より 13.6 GFlops 高いことが分かる (表 4)。

2003 年、9 月に Force10 Networks 社の E1200 を用いて計測した際に得られた結果¹⁵⁾ との比較を表 5 に示す。ただし、当時の結果と今回の結果では、HPL のパラメータだけでなく用いたコンパイラが異なるため、参考としての比較とする。E1200 は、1.44 Tbps のバックプレーンを持ち、最高 336 ホスト間のノンブロッキング通信が可能な超高性能スイッチである。表 5 より、本実験で得られた最高計測値 (1081 GFlops) は、256 台のホストを 1 台のスイッチに接続したフラットなトポロジに匹敵する値である。さらに、今回は 2003 年の計測時に比べて 62 個少ない CPU を用いて、E1200 を用いた場合より 3.3%高い 66.7%の実行効率 (Rmax/Rpeak) を得ることができた。2008 年 11 月に発表された TOP500 スーパーコンピュータのランキング³⁾ では、ギガビットイーサネットを用いた PC クラスタの実行効率は、最高で 63.0%であり、TOP500 にランクインしているギガビットイーサネットを用いた PC クラスタのうち約 9 割のシステムの実行効率は 55.0%以下である。このように、ギガビットイーサネットを用いた大規模 PC クラスタにおいて実行効率が 60.0%を超

表 5 Force 10 E1200 スイッチ使用時との比較

Table 5 Comparison to the case using FORCE10 E1200 switch.

	Powerconnect6248	E1200
Number of Processors	450	512
Cable	CAT6E	CAT5E
Number of Switches	8	1
Rmax (TFlops)	1.081	1.169
Rpeak (TFlops)	1.620	1.843
Rmax/Rpeak (%)	66.7	63.4
Nmax	180000	220000
Cost (\$)	16,000	400,000

える結果を得るのは困難であるが、今回それを大きく上回る 66.7%という結果を得ることができた。

また、スイッチの費用対効果を比較した場合、E1200 の 25 分の 1 の費用で同等の性能を出せていることから、小規模なスイッチを多数用いてトラス、完全結合などのトポロジを構成することは、費用対効果がきわめて高いといえる。

4.1.3 NPB の評価結果

次に NAS Parallel Benchmarks 3.2 を用いて、各トポロジにおける各アプリケーションの実行性能を測定し、トポロジ間の性能を比較する。

各ベンチマークの問題サイズはクラス C とし、各アプリケーションの実行プロセス数は、計算を実行できるホスト数 225 内の最大値 128 あるいは 225 とした。アプリケーションは、CG 法、FT 法、IS 法、LU 法、MG 法、BT 法、SP 法を使用した。コンパイラは gcc 3.3.6/g77 3.4.6 を用いてオプションを-O3 として行った。各トポロジでのベンチマーク性能 (Mop/s/process) を測定した結果を図 4 に示す。図 4 では、Tree (1link) における性能値により正規化している。トポロジの表記は図 3 と同様である。なお、アプリケーションの表記である CG.128 は、CG 法における実行プロセス数が 128 であることを示す。各アプリケーションにおいて、提案手法を実装したトポロジを用いることで Tree (1link) に比べて最大 654.4%高い性能値を計測できたことが分かる。

1 つのスイッチに 28~29 台のホストが接続されている本環境では評価に用いたすべてのトポロジにおいて、ホスト間通信がノンブロッキングとはならない。そのため、FT 法、IS 法は全対全通信が生じ、通信量が多いため、すべてのトポロジにおいて多数のパケットの衝突が発生していると考えられる。そのため、FT 法、IS 法において Torus (6link) が Tree (6link) に比べて各々 63.6、20.7%の性能向上を達成した。

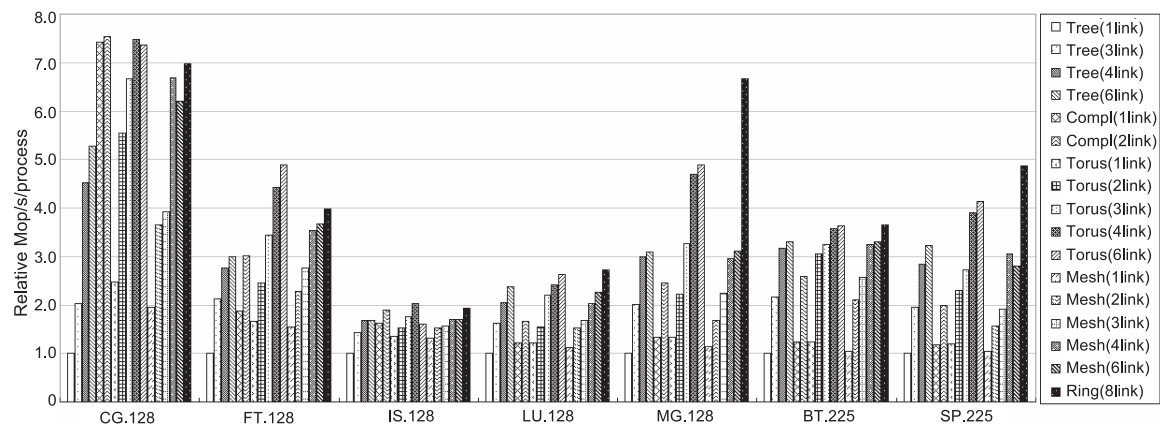


図 4 SuperNova における NAS Parallel Benchmarks の結果
Fig. 4 NAS Parallel Benchmarks results of SuperNova.

さらに, CG 法では, Tree (1link) の結果に比べて Tree (6link) の結果では 426.7% の性能向上を達成しているが, トポロジを Compl (2link) に変更することによってさらに 227.7% 性能向上を達成した. また, MG 法では, Ring (8link) が Tree (6link) に比べて最大 115.6% の性能向上を達成した. MG 法, CG 法では単純にバイセクションバンド幅の改善が性能向上に直結するわけではないが, 通信パターンにあわせてルーティングを最適化することで大きく性能が向上することが報告されている¹⁶⁾. 本評価環境においても, 単純なツリー構成において偏りのある経路を用いた場合に比べて, 完全結合, あるいは, トーラスでは分散された経路を用いることができるため, MG 法, CG 法ではきわめて大きなトポロジ間の性能差につながったと考えられる. つまり, 本評価では, アプリケーションが使用する経路群の偏りがトポロジに与える影響がきわめて大きいといえる.

よって, SuperNova クラスタでは, 今回すべてのアプリケーションにおいて, VLAN を用いてチューニングを行ったトポロジが高い性能値を得られており, VLAN ルーティング法の有効性が示されたといえる.

4.2 Misc クラスタにおける評価

4.2.1 システム構成

Misc クラスタは 2008 年に同志社大学に導入された PC クラスタであり, 528 コア 66 台のホストで構成される (図 5). SuperNova クラスタと同様に PowerConnect6248 を使用



図 5 Misc クラスタの概観
Fig. 5 Overview of Misc Cluster.

した. ただし, 各ホストは MPI 通信のために, 2 本のギガビットイーサネットを用いることができる. 表 6 にホストの仕様を示す. 各スイッチには 11 台のホストが接続されており, 計 6 台のスイッチを用いている.

ここでは, ツリートポロジと SuperNova クラスタで性能が高かった完全結合の比較に焦点をあてる. ツリー, 完全結合の各トポロジについて, スイッチ間のリンク数を 1~5 本に変化させ, さらにホスト-スイッチ間のリンク数が 1 本, 2 本の場合について評価を行った.

表 6 Misc クラスタにおけるホストの仕様
Table 6 Specifications of each host of Misc.

CPU	Quad-Core AMD Opteron 2.3 GHz × 2
Memory	DDR2 667 MHz 8 GB
NIC	Broadcom BCM95721 1000BaseT × 2
NIC driver	Broadcom Tigon3
OS	CentOS 4.6
Kernel	2.6.9-67.0.15.ELsmp

表 7 Misc クラスタのトポロジの特徴
Table 7 Statics of topologies on Misc Cluster.

	直径	Bi. BW (Gbps)	VLAN 数
Tree (<i>n</i> link)	3	3 <i>n</i>	1
Compl (<i>n</i> link)	2	9 <i>n</i>	6

なお、ツリートポロジは1つのスイッチに他の5つのスイッチが接続される構成とした。並列ベンチマークは Open MPI1.3¹⁷⁾ を用いた IP によりプロセス間通信を行った。

これらのトポロジの静的な特徴を表7にまとめる。“Bi. BW”はバイセクションバンド幅を表す。

Misc クラスタにおいてホスト-スイッチ間のリンク数が2本の場合、目的地のIPアドレス、UDP/TCPのポート番号でリンク間のトラフィック分散を行った。

4.2.2 HPL の評価結果

Misc クラスタを用いて評価を行った結果を図6に示す。()内はスイッチ間リンク数およびホスト-スイッチ間リンク数である。HPL性能評価のプログラムとしてはHPL2.0を用い、数値演算ライブラリにはGotoBLAS1.26を用いた。コンパイルにはgcc 3.4.6/g77 3.4.6を用いた。計測に利用したHPLの主要なパラメータは表3に示したとおりである。

Misc クラスタは NUMA アーキテクチャを採用したマルチコアプロセッサを用いた PC クラスタであり、HPL 実行時の MPI プロセス数およびスレッド数によって性能が変化することが報告されている¹⁸⁾。本評価では MPI プロセス数に関しては、以下を比べた結果、前者の方が性能が高かったため、前者を採用した。

- 1 コアに対して 1 MPI プロセス (MPI processnum=8)
- 1 ホストに対して 1MPI プロセス (MPI processnum=1)

なお、これはマルチコア Quad-Core AMD Opteron を用いた大規模 PC クラスタにおける評価結果と同様の傾向である¹⁸⁾。

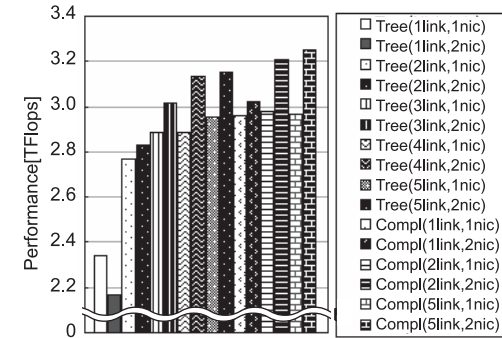


図 6 Misc クラスタにおける HPL の結果
Fig. 6 HPL results of Misc Cluster.

なお、HPL 実行時には、MPI プロセスのリモートメモリ参照のオーバーヘッドを最小限に抑えるため、Open MPI1.3 のプロセッサアフィニティ機能を利用し、連続した4つのランク番号の MPI プロセスを同一 CPU 内の4コアにバインドさせて評価を行った。本設定を行った場合、設定を行わなかったときに比べて407.0 GFlops 高い性能が得られることを確認した。

図6より、直径が小さく、バイセクションバンド幅が大きいトポロジである Compl(5link, 2nic)は Tree(1link, 1nic)に比べて最大38.7%の性能向上を達成できたことが分かる。SuperNova クラスタの HPL 評価と同様、リンク集約化が性能向上に効果的であり、単純ツリートポロジ内で、ホスト-スイッチ間リンク本数が1本の場合と比較すると、Tree(1link, 1nic)の実行効率が48.2%であるのに対して、Tree(5link, 1nic)では、60.8%の実行効率を達成できた。

また、ホスト-スイッチ間リンク本数が1本と2本の場合で比較した結果、Tree(1link)以外の各トポロジにおいて、2本の方が1本の場合よりも2.2%から9.6%ほど高い実行性能を得ることができた。

トポロジ間の比較では、Tree(5link, 2nic)と Compl(5link, 2nic)で得られた最大実行効率の差は2.1%であり、リンク集約化とホスト-スイッチ間のリンク数が大きく影響していることといえる。

なお、Misc クラスタでの結果では、どのトポロジであってもエラーが生じないことを確認している。

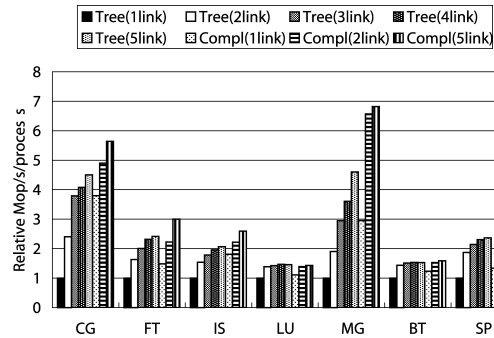


図 7 Misc クラスタにおける NAS Parallel Benchmarks の結果 (64 プロセス)
Fig. 7 NAS Parallel Benchmarks results of Misc Cluster (64 processes).

4.2.3 NPB の評価結果

各ベンチマークの問題サイズはクラス C とし、各アプリケーションの実行プロセス数は、64 あるいは 128 とした。用いたアプリケーションは SuperNova クラスタの評価時と同様であり、コンパイラは Misc の HPL 評価時と同様である。ただし、ホスト-スイッチ間のリンク数は 1 本とした。各トポロジでのベンチマーク性能 (Mop/s/process) を Tree (1link) における性能値により正規化した結果を図 7, 図 8 に示す。なお、128 プロセス実行時ではスイッチ間リンク数の影響が大きかったため、この詳細を見るために図 8 では Compl (3link), Compl (4link) の結果を含めた。

各アプリケーションにおいて、提案手法を実装したトポロジを用いることで Tree (1link) に比べて最大 1131.3% 高い性能値を計測できたことが分かる。

64 プロセス実行では、すべてのトポロジにおいてリンク集約化が性能向上に効果的であり、Tree (1link) に対して、Tree (5link) では、CG 法、FT 法、IS 法、MG 法、SP 法では 100% 以上の性能向上を達成できており、特に MG 法では 359.9% の性能向上を達成した。トポロジ間の比較では、CG 法、FT 法、IS 法、MG 法において Tree (5link) に対して Compl (5link) が 52.4% ~ 222.1% 高い性能向上を達成していることから、これらのアプリケーションではトポロジの変更が効果的であることが分かる。128 プロセス実行においても、すべてのトポロジにおいてリンク集約化が性能向上に効果的であり、Tree (5link) では、すべてのアプリケーションにおいて、100% 以上の性能向上を達成でき、特に FT 法では 231.9% の性能向上を達成した。

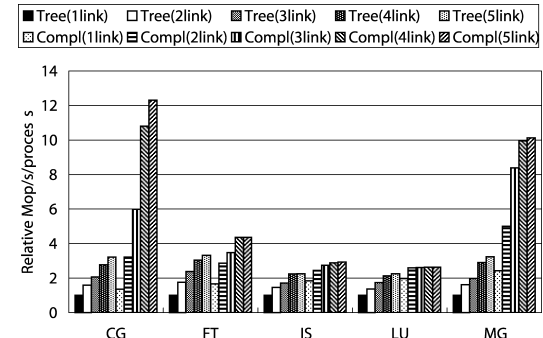


図 8 Misc クラスタにおける NAS Parallel Benchmarks の結果 (128 プロセス)
Fig. 8 NAS Parallel Benchmarks results of Misc Cluster (128 processes).

CG 法に関しては、128 プロセス実行時の Compl トポロジにおいて、スイッチ間リンク数が特に大きく性能に影響を与えていることが分かる。4.1.3 項に述べたとおり、CG 法では経路の分散が大きく影響を与えるため、スイッチ間リンク数が少ない場合は、経路が集中するリンクが性能のボトルネックとなることが考えられる。そのため、Compl (1link) は Tree (3link) に比べて性能が劣っているが、Compl (2link) 以上では劇的に性能が向上していると考えられる。

これらより、NAS Parallel Benchmarks では、各アプリケーションにおいてネットワークトポロジのチューニングが必要であり、用いるトポロジによっては単純木構造の場合に比べて非常に高い性能値が得られることが分かる。

4.3 2つのPCクラスタにおけるトポロジ評価の傾向

本章では従来のシングルコアプロセッサを用いた SuperNova クラスタの一部と、マルチコアプロセッサを用いた Misc クラスタの両方におけるトポロジの評価結果を示した。いずれのトポロジも同一のスイッチ、リンクを用いて構成されている。

スイッチの台数が限られている本環境では 1 つのスイッチに 11 ~ 29 台のホストが接続されている。そのため評価に用いたすべてのトポロジにおいて、ホスト間通信がノンブロッキングとはならない点が 1 つの特徴である。両クラスタにおいて以下の傾向が見られた。

- HPL に関しては、完全結合などのトポロジによる性能向上よりも、リンク集約化による性能向上が大きかった。
- NAS Parallel Benchmarks に関しては多くの場合、トラスなどの経路を分散するこ

とができるトポロジによる性能向上が大きかった。

以上より、アプリケーションにより適したトポロジは異なるが、VLAN ルーティング法によりツリー以外のトポロジ（完全結合，トーラス）を選択することにより性能が向上する場合が多いことが分かった。

5. ま と め

イーサネットにおいて、VLAN ルーティング法を用いることで、様々なトポロジを採用することができるようになった。しかし、これまで大規模 PC クラスタにおけるイーサネットのトポロジの性能評価はほとんど行われていない。そこで、本稿では 450 コア 225 台のホストで構成される SuperNova クラスタ、および、528 コア 66 台のホストで構成される Misc クラスタにおいて、トポロジが性能に与える影響について調べた。既存の PC クラスタにおいて最小限のシステム更新で VLAN ルーティング法を実装するために、(1) スイッチにおいてフレームに VLAN タグを付与し、(2) MAC アドレステーブルの学習により各スイッチの経路管理を行う汎用性の高い方法を提案し、実装した。

評価結果より、SuperNova クラスタにおいて 8 台の安価な 48 ポートスイッチをトーラス、完全結合トポロジで接続した場合の High-Performance LINPACK benchmark (HPL) の性能は、336 ポートの高価なノンブロッキングスイッチを 1 台使用した場合とほぼ同等であることが分かった。さらに、Misc クラスタにおいて完全結合トポロジにおける NAS Parallel Benchmarks の性能は、リンク集約化を行ったツリートポロジに比べて最大 909.2%向上することが分かった。

今後は、大規模クラスタにおいて安定的に運用するツールの検討などを行う予定である。謝辞 本研究の一部は、科学技術振興機構「JST」の戦略的創造研究推進事業「CREST」の支援による。

参 考 文 献

- 1) 工藤知宏, 松田元彦, 手塚宏史, 児玉祐悦, 建部修見, 関口智嗣: VLAN を用いた複数パスを持つクラスタ向き L2 Ethernet ネットワーク, 情報処理学会論文誌: コンピューティングシステム, Vol.45, No.SIG 6(ACS 6), pp.35-43 (2004).
- 2) Sharma, S., Gopalan, K., Nanda, S. and cker Chiueh, T.: Viking: A Multi-Spanning-Tree Ethernet Architecture for Metropolitan Area and Cluster Networks, *Infocom*, pp.2283-2294 (2004).
- 3) Top 500 Supercomputer Sites. <http://www.top500.org/>

- 4) Pellegrini, F.D., Starobinski, D., Karpovsky, M.G. and Levitin, L.B.: Scalable Cycle-Breaking Algorithms for Gigabit Ethernet Backbones, *Infocom*, pp.2175-2184 (2004).
- 5) Reinemo, S.-A. and Skeie, T.: Effective Shortest Path Routing for Gigabit Ethernet, *IEEE International Conference on Communications (ICC)*, pp.6419-6424 (2007).
- 6) 三浦信一, 岡本高幸, 朴 泰祐, 佐藤三久, 高橋大介: tagged-VLAN に基づく PC クラスタ向け高バンド幅 ツリーネットワークの開発, 情報処理学会研究報告 2005-HPC-104, pp.13-18 (2005).
- 7) 大塚智宏, 鯉淵道紘, 工藤知宏, 天野英晴: スイッチでタグ付けを行う VLAN ルーティング法, 情報処理学会論文誌: コンピューティングシステム, Vol.47, No.SIG 12(ACS 15), pp.46-58 (2006).
- 8) 大塚智宏, 鯉淵道紘, 工藤知宏, 天野英晴: VLAN イーサネットを用いた PC クラスタ向け大規模ネットワーク構築法, 情報処理学会論文誌: コンピューティングシステム, Vol.1, No.3, pp.96-107 (2008).
- 9) Garcia, R., Duato, J. and Serrano, J.J.: A New Transparent Bridge Protocol for LAN Internetworking using Topologies with Active Loops, *Proc. International Conference on Parallel Processing (ICPP)*, pp.295-303 (1998).
- 10) Lui, K., Lee, W. and Nahrstedt, K.: STAR: A transparent spanning tree bridge protocol with alternate routing, *ACM SIGCOMM Computer Communication Review*, Vol.32, No.3, pp.33-46 (2002).
- 11) Watanabe, T., Nakao, M., Hiroyasu, T., Otsuka, T. and Koibuchi, M.: The Impact of Topology and Link Aggregation on PC Cluster with Ethernet, *Poster (Work-in-progress presentation), IEEE International Conference on Cluster Computing (Cluster2008)* (2008).
- 12) 大塚智宏: VLAN イーサネットを用いた大規模クラスタネットワークの構築に関する研究, 慶應義塾大学大学院博士論文 (2009).
- 13) Tperf. <http://www.am.ics.keio.ac.jp/~terry/tperf/>
- 14) HPL — A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers. <http://www.netlib.org/benchmark/hpl/>
- 15) 廣安知之, 三木光範, 荒久田博士: テラフロップスクラスタの構築と Benchmark による性能評価, 同志社大学理工学研究報告, Vol.45, No.4, pp.187-198 (2005).
- 16) 三浦信一, 岡本高幸, 朴 泰祐, 佐藤三久, 高橋大介: VFREC-Net: ドライバ制御による tagged-VLAN を用いた PC クラスタ向け マルチパスネットワーク, 情報処理学会論文誌: コンピューティングシステム, Vol.47, No.SIG 12(ACS 15), pp.35-45 (2006).
- 17) OpenMPI. <http://www.open-mpi.org/>
- 18) 高橋大介, 後藤和茂, 朴 泰祐, 建部修見, 佐藤三久, 三上和徳: T2K 筑波システムにおける Linpack 性能評価, 情報処理学会研究報告, Vol.2008, No.74, pp.55-60 (2008).

19) Goto, K. <http://www.tacc.utexas.edu/resources/software>

(平成 21 年 1 月 27 日受付)

(平成 21 年 5 月 25 日採録)



廣安 知之 (正会員)

1997 年早稲田大学理工学研究科後期博士課程修了。早稲田大学工学部助手、同志社大学工学部助手、知識工学科専任講師、インテリジェント情報工学科准教授を経て 2008 年から生命医科学部教授。進化的計算、最適設計、並列処理、設計工学、医療画像工学等の研究に従事。IEEE、電子情報通信学会、計測自動制御学会、日本機械学会、超並列計算研究会、日本計算工学会各会員。



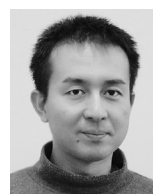
渡辺 崇文

2007 年同志社大学工学部知識工学科卒業。2009 年同大学大学院工学研究科知識工学専攻博士前期課程修了。修士(工学)。同年(株)インターネットイニシアティブ入社。在学中、PC クラスタのネットワークに関する研究に従事。



中尾 昌広 (学生会員)

2005 年同志社大学大学院工学研究科博士前期課程修了。同年 NTT アドバンステクノロジー株式会社入社。2007 年同志社大学大学院工学研究科博士後期課程入学。ハイパフォーマンスコンピューティング、バイオインフォマティクス、進化的計算に興味を持つ。人工知能学会学生会員、超並列計算研究会会員。



大塚 智宏 (正会員)

2001 年慶應義塾大学理工学部情報工学科卒業。2009 年同大学大学院理工学研究科開放環境科学専攻後期博士課程修了。博士(工学)。2006 年度から 2008 年度まで慶應義塾インフォメーションテクノロジーセンター本部助教。現在、同センターに勤務。在学中は並列分散処理システムおよびインターコネクトに関する研究に従事。



鯉淵 道紘 (正会員)

2000 年慶應義塾大学理工学部情報工学科卒業。2003 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)。2002 年度より 2004 年度まで日本学術振興会特別研究員。現在、国立情報学研究所助教、総合研究大学院大学複合科学研究科情報学専攻助教(兼任)。ハイパフォーマンスコンピューティングとインターコネクトに関する研究に従事。

IEEE Computer Society Japan Chapter Young Author Award 2007, 2007 年度情報処理学会論文賞受賞。IEEE, 電子情報通信学会各会員。