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] 中尾昌広[†2]
大塚智宏[†3] 鯉渕道紘[†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わかった．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]
Masahiro Nakao,[†2] Tomohiro Otsuka[†3]
and Michihiro 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 existing 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）のみならず，広域ネットワークや PC クラスタのインタコネクト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ト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

†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

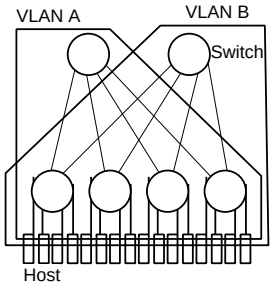

図 1 VLAN ルーティング法
Fig. 1 VLAN routing method

する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切れなくなる．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バ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 グループのメンバー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関わらず，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) システムソフトウェアが 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 個と極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 節で関連研究を述べ，3 節において VLAN ルーティング法の実装について述べ，4 節にて，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 節でまとめを述べる．

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，ループ構造を含むさまざま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 ルーティング法の利用を実現している．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ヘッドはレイヤー 2 スイッチング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 等が提唱し，現在 IEEE 等で 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用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．

- 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．

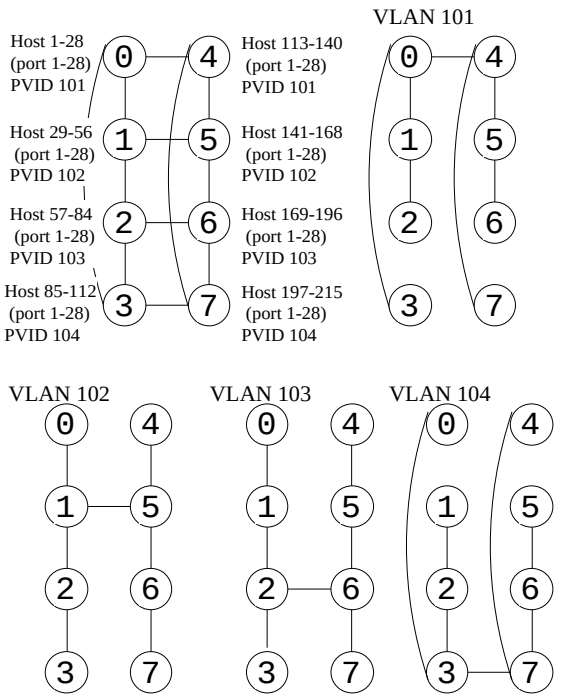

図 2 スイッチで VLAN タグ付けを行うルーティングの例
Fig. 2 Example of switch-tagged VLAN routing

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，さまざま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（これをフラッディングと呼ぶ），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，以後はフラッディングを伴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．例えば，本評価に用いた Dell 社 PowerConnect 6248 は高々 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 等を使って作成する．例えば図 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 毎に一意のネットワーク（IP）アドレスを与え，VLAN 毎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毎に，ICMP または UDP メッセージを一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，例えば各 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 などで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≒ 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 日に一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．

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

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

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）の各トポロジについて，スイッチ間のリンク数を 1〜8 本に変化させて評価を行った．いずれのトポロジも，規則性が強くあるため高々数個の VLAN で実装している．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，Tperf[12] を用いた測定結果から，ポートあたり 957Mbps(UDP), 939Mbps(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）[13] 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 ではシステムの特性にあったパラメータを設定す

表 2　主な HPL のパラメータ
Table 2　The main parameters of HPL

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

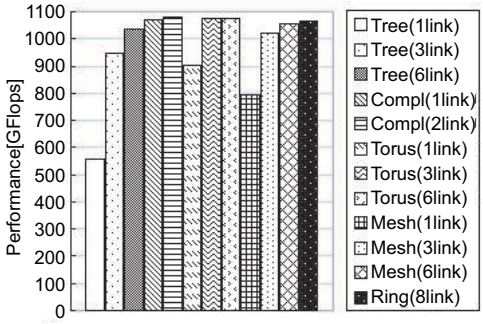

図 3　SuperNova における HPL 結果
Fig. 3　HPL results of SuperNova

ることが可能である．今回計測に利用した HPL の主要なパラメータを**表 2** 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．しかし，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まった解の精度確認時にエラーが起きていたため，これらの計測値は参考値である．

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性能向上が可能であることが分かる．しかし，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）に勝っていると言える．例えば，Compl（1link）の場合，リンク 1 本あたりの性能は 38.2GFlops であり，Tree（6link）より 13.6GFlops 高いことが分かる（**表 3**）．

表 3　スイッチ間リンク 1 本あたりの性能の比較
Table 3　Comparison of a performance per link between switches

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

表 4　Force 10 E1200 スイッチ使用時との比較
Table 4　Comparison wich using FORCE10 E1200 switch

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

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[14] との比較を**表 4** 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，1.44Tbps 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 4 より，本実験で得られた最高計測値（1081GFlops）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ーラス，完全結合などのトポロジを構成することは，費用対効果が極めて高いといえる．

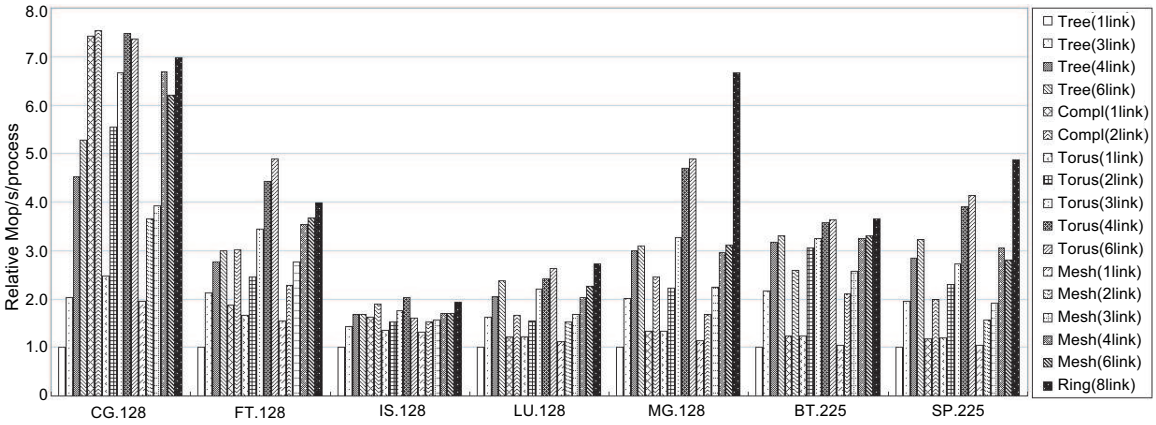

図 4 SuperNova における NAS Parallel Benchmarks 結果
Fig. 4 NAS Parallel Benchmarks results of SuperNova

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．コンパイルは 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．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）に変えることによってさらに 227.7% 性能向上を達成している．これは，リンク集約化だけではなく，トポロジの変更が性能向上に効果的であると言える．各アプリケーションにおいて，Tree（6link）の結果を基準に比較を行うと，CG 法では，Compl（2link）が最大 43.2%の性能向上を達成，FT 法では，Torus（6link）が最大 63.6%の性能向上を達成，MG 法では，Ring（8link）が最大 115.6%の性能向上を達成，SP 法では，Ring（8link）が最大 50.6%の性能向上を達成するなど，これらのアプリケーションにおいてトポロジの変更が性能向上に影響を及ぼしていることが見てとれる．以上の結果より，完全結合，トーラス，リングトポロジが極めて有効であることが分かる．

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

表 5 Misc におけるホストの仕様
Table 5 Specifications of each host of Misc

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

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 5** 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，SuperNova クラスタで性能が高かった完全結合，ツリートポロジの比較に焦点をあてる．単純ツリー，完全結合（次元順ルーティング）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．並列ベンチマークは Open MPI1.3[15] を用いた IP によりプロセス間通信を行った．

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 2 に示した通りである．

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[16]．GotoBLAS ではユーザが任意に指定したスレッド数で行列演算をノード内で並列化できるため[17]，1 ホストに対する MPI 実行プロセス数と GotoBLAS における行列演算の並列スレッド数を変化させて予備評価を行った．評価の結果，1 ホストに対して MPI の実行プロセス数が 8，GotoBLAS の並列スレッド数が 1 である場合に最も高い性能が得られたため，HPL の実行プロセス数は 1 ホストに対して 8MPI プロセス（1 コアあたり 1MPI プロセス）として評価を行った．

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時に比べて 407.0GFlops 高い性能が得られることを確認した．

図 6 より，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 本の場合で比較すると，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．

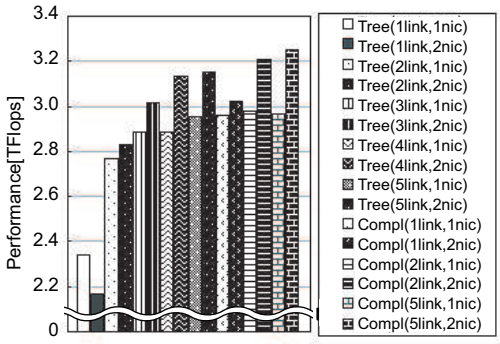

図 6 Misc クラスタにおける HPL 結果
Fig. 6 HPL results of Misc Cluster

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，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て**図 7** に示す．

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 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，全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，全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，全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．

トポロジ間の比較では，CG 法，MG 法において，Tree（5link）の結果に比べて Compl（5link）が 688.1% 〜909.2% と劇的な性能向上を達成したことから，128 プロセス実行でのこれらのアプリケーションにおいては，トポロジの変更が極めて効果的であることが分かった．

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

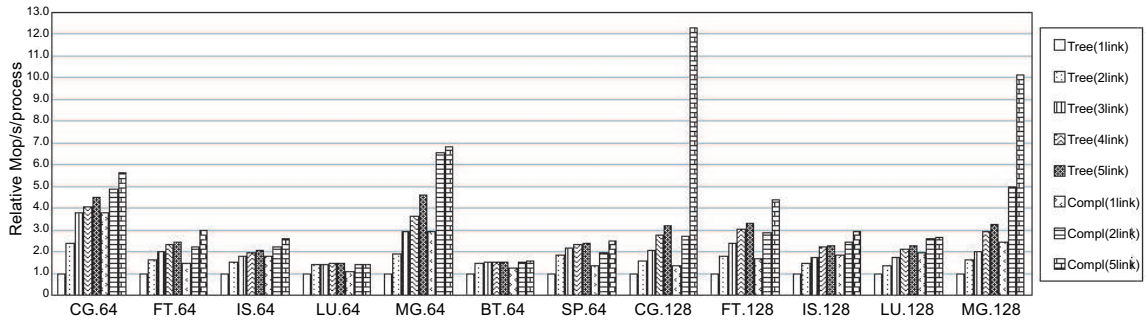

図 7 Misc クラスタにおける NAS Parallel Benchmarks 結果
Fig. 7 NAS Parallel Benchmarks results of Misc Cluster

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わかった．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．

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, 鯉渕道紘, 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, 鯉渕道紘, 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. of the 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 Comput ing (Cluster2008)* (2008).
12) Tperf: `http://www.am.ics.keio.ac.jp/~terry/tperf/`.
13) HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers: `http://www.netlib.org/benchmark/hpl/`.
14) 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).
15) OpenMPI: `http://www.open-mpi.org/`.
16) 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).
17) Goto,K.: `http://www.tacc.utexas.edu/resources/software`.