社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－HPC－96 (14)
2003／10／16

VLANを用いた複数パスを持つL2 Ethernetネットワーク

工藤知宏† 手塚宏史† 児玉祐悦†
建部修見† 松田元彦† 関口智嗣†

本稿では、複数パスを持つLayer2 Ethernetネットワークを実現するVLANルーティング法を提案する。VLANルーティング法により、従来Myrinetなどの System Area Networkで広く用いられてきたFat Treeのようなクラスタに適したトポロジを、安価なEthernetを用いて構成することができる。

また、VLANルーティング法を用いたクラスタ向きネットワークトポロジをいくつか提案するとともに、これを用いた小規模なFat Treeネットワークの通信性能評価とNAS並列ベンチマークの性能評価により有効性を示す。

VLAN-based multi-path L2 Ethenet network

Tomohiro Kudoh ,† Hiroshi Tezuka ,† Yuetsu Kodama ,†
Osamu Tatebe ,† Motohiko Matsuda † and Satoshi Sekiguchi†

In this report, VLAN-based routing, which realizes multi-path Layer 2 Ethernet network is proposed. By using VLAN-based routing, those topologies such as Fat Tree which have been widely used in System Area Networks like Myrinet can be formed using Ethernet.

We also show several topologies which use VLAN-based routing and are suitable for high performance clusters. Communication bandwidth measurement result and performance of NAS parallel benchmark executed on a small scale cluster with the VLAN-based fat tree network is shown to demonstrate the effectiveness of this method.

1. はじめに

Gigabit Ethernetはバンド幅あたりの価格が安価であり、クラスタ内ネットワーク用のインタコネクトとして非常に魅力的である。しかし、ポート数が30程度以上のスイッチは高価であるため、安価にネットワークを構成するには小規模なスイッチを多数用いる必要がある。

従来、高性能なHPC向けクラスタでは、Myrinet[1]やQsNET[2]のようなSystem Area Network(SAN)が用いられてきた。このようなクラスタ向きインタコネクトではFat Treeなどの上位スイッチがボトルネックとならないトポロジが用いられる。このようなトポロジでは、クラスタのノード間に複数のパスが存在し、高いバイセクションバンド幅などのHPCクラスタ向けネットワークに必要な性質を備えている。

ところが、L2 Ethernetのネットワークトポロジは基本的に単純な木構造に限られる。このため、ノード間に複数のパスが存在するネットワークは、L2 Ethernetでは構成できない。上位リンクに下位よりも高いバンド幅を持つリンクを用いればこの問題は解決できるが、Gigabit Ethernetの上位の規格である10Gigabit Ethernetは高価であり、また、上位のスイッチに高いスイッチング能力が要求されることになり、システム全体のコストが高くなってしまう。

本稿では、VLANを用いることによりL2 Ethernetネットワーク上に複数パスを実現するVLANルーティング法(VLAN-based routing)を提案する。

本手法を用いれば、Fat Treeなどの複数パスを持つトポロジのネットワークを構築できる。ノード間の複数のパスは異なるVLANに属する。各ノードはVLANごとに異なる複数のIPアドレスを持ち、送信側のノードは、宛先のIPアドレスを指定することにより、どのVLANを介して通信するかを明示的に指定できる。HPCクラスタでは、実行する問題の性質が既知であることが多く、パスの選択をアプリケーションプログラムやミドルウェアにより明示的に行えることにより適切な通信負荷の分散ができ、システム全体の性能を向上させることができる。

本稿では、Fat Tree、完全グラフ、Hyper Crossbarを構成する例を示す。また、小規模なFat Treeネットワークを用いたクラスタにおいて通信性能とNAS並列ベンチマークの性能を評価した結果を示す。

† 産業技術総合研究所グリッド研究センター
Grid Technology Research Center, AIST

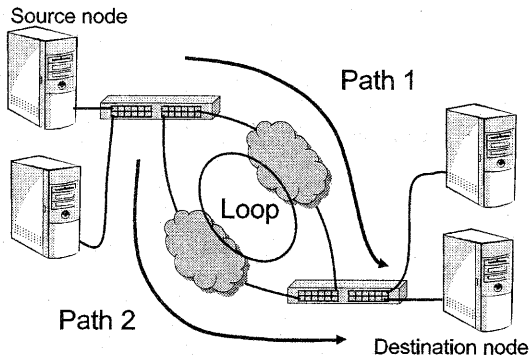


図 1　スイッチ間に複数パスを持つネットワーク

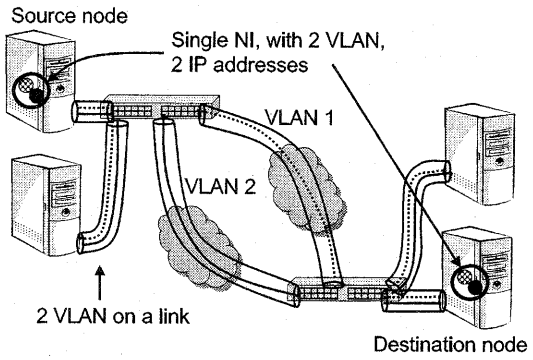


図 2　VLAN ルーティング法による複数パスの実現

2. VLAN ルーティング法

図 1 に示すように、複数のノードが接続されたスイッチ間を複数のパスで接続し、ノード毎にパスを選択すれば、それぞれのスイッチに接続されたノード群の間の通信バンド幅は、パスが一つしかない場合と比べてパス数倍になる。しかし、このようなネットワークにはループが存在する。Layer2 Ethernet では、ネットワーク中にループが存在すると broadcast storm が発生し、また MAC アドレステーブルが不安定になってしまう。このような現象を避けるために、多くのスイッチが STP(Spanning Tree Protocol) をサポートしている。Spanning Tree Protocol では、ネットワーク上のトポロジを検出するための専用のメッセージをスイッチが交換し、ループが検出されたら、ループ上のあるリンクを使用しないようにすることにより実際にループを使用することを回避する。従って、スイッチ間の複数パスを持つように配線しても、実際には複数パスを利用することはできない。

本稿で提案する VLAN ルーティング法では、VLAN を用いてこの問題を解決する。

VLAN 技術を用いると、物理的なネットワーク上に仮想的に複数のネットワーク (VLAN) を構築することができる。異なる VLAN 間ではイーサネットフレームは伝搬しない。従って、物理的なネットワークがループを含んでいても、それぞれの VLAN が木構造のトポロジを持っていれば broadcast storm は発生しない。IEEE803.1Q による VLAN の規定では、イーサネットフレームには Tag 付のものと Tag 無のものがある。

スイッチのポートに入力されたイーサネットフレームは、

- Tag 付であれば、その Tag によって示される VLAN に属するものとみなされる。その VLAN id がそのポートに登録されていなければ、フレームを破棄するように設定することもできる。
- Tag 無であれば、あらかじめそのポートに設定された PVID(Port VLAN id) の VLAN に属するものとみなされる。

一方、スイッチの出力ポートには、VLAN id 毎に Tag 付と Tag 無の設定をすることができ、

- そのポートに Tag 無で設定されている VLAN id のフレームは、Tag の無い通常のフレームとして出力される。
- そのポートに Tag 付で設定されている VLAN id のフレームは、Tag 付で出力される。
- そのポートに Tag 付、Tag 無のいずれも設定されていない VLAN id のフレームは、そのポートからは出力されない。

という動作をする。Tag 付で受け取ったフレームを Tag 無のポートに出力する際にはフレームの Tag フィールドは削除される。逆に Tag 無で受け取ったフレームにはそのポートの PVID が割り当てられ、Tag 付のポートに出力する際には Tag フィールドがスイッチによって付加される。特に VLAN 設定をしていないスイッチは、すべてのポートが同一 id の VLAN に Tag 無フレームを送受するように設定されているととらえることができる。

VLAN ルーティング法では、図 2 に示すように、ノードやスイッチ間の複数のパスに異なる VLAN を割り当てる。同一リンクに単一の VLAN しか割り当てられない場合には、Tag 無フレームを使うことができる。同一リンクに複数の VLAN が割り当てられる場合には、識別のために Tag 付フレームを用いなくてはならない。それぞれのリンクは、そのリンクに割り当てられた VLAN に属するフレーム以外は通過できない。各 VLAN はループを含まないため、broadcast storm が発生することはない。各ノードのネットワークインタフェース (NIC) には、それぞれの VLAN に属する仮想インタフェースがあり、VLAN 毎に IP アドレスを持つ。各ノードは仮想インタフェースを持ついずれの VLAN のフレームも送受できる。

送信側のノードは、宛先 IP アドレス毎にどの仮想インタフェースにパケットを送るかをあらかじめ IP 経路テーブルに設定しておく。実際には VLANid 毎

にネットワークを構成し、ネットワーク毎に経路テーブルに登録すればVLAN数の登録ですむ。送信時には、宛先のIPアドレスを指定することにより、どのVLANを介して通信するかを明示的に指定できる。図2で、送信側ノードは、宛先ノードのVLAN1に属するIPアドレスに送信すればVLAN1に属する上側のパスでパケットを送ることができ、VLAN2に属するIPアドレスに送信すれば下側のパスで送ることができる。

通常VLANは、単一の物理ネットワークを用いて、複数の仮想ネットワークを構成し、仮想ネットワーク間での通信ができないようにするために用いられる。これに対して、VLANルーティング法は複数パスを構成するためにVLANを用いており、基本的にいずれのノードも他のすべてのノードとLayer2で通信することができる。

3. VLANルーティング法を用いたネットワークトポロジ

VLANルーティング法ではループを含む様々なトポロジを実現できる。SANや超並列計算機で用いられてきたトポロジがHPCクラスタ向けにも有効であると考えられる。本節では、VLANルーティング法を用いたトポロジの例をいくつか提案、考察する。

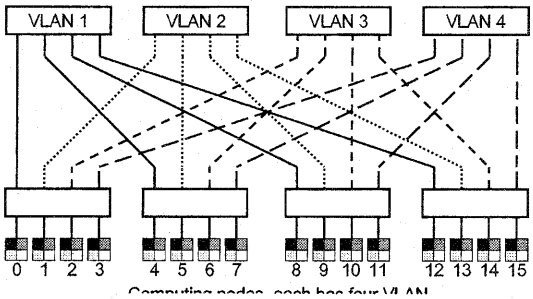
3.1 VLAN-based Fat Tree(VBFT)

VBFTは、VLANルーティング法を用いたFat Treeである。VBFTの例を図3に示す。図の下の部分にある各ノードは、図4に示すように、単一のNIC上に4つのVLANの仮想インタフェースを持つ。ノードと下部のスイッチの間はTag付フレームが交換される。上部と下部のスイッチ間では、各リンクには1種類のVLANのフレームしか流れない。また上部の各スイッチは単一のVLANに属する。このため、下部のスイッチの上部スイッチと接続されるポートの設定をTag無のポート単位のVLANとし、上部のスイッチにはVLANをサポートしていないスイッチを用いることもできる。

VBFTはFat Treeであり、この図の例ではネットワークを2つに分割した際にフルバイセクションバンド幅を持っている。これを拡張し下部に24ポートのVLANをサポートするスイッチ12台を用い、上部に12ポートのスイッチ12台を用いれば、144台のノードを接続することができる。

3.2 完全グラフネットワーク

スイッチ間を完全グラフネットワークで接続する場合、ネットワーク上にVLANを設定する様々な方法が考えられる。図5にその一例を示す。この図では丸印がスイッチを表しており、n個 (図では6) のスイッチが (a) に示す物理リンクで接続されている。図には示していないが、各スイッチに複数のノードが接続され



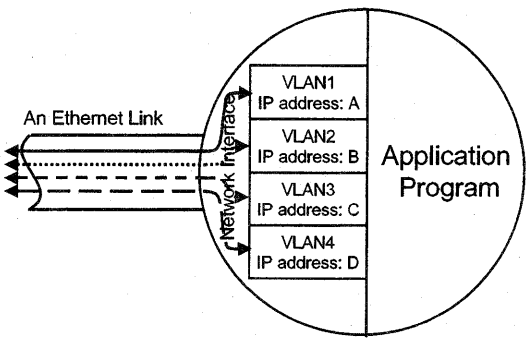


図4 ノード上のVLAN仮想インタフェース

ることを想定している。この物理リンク上に (b) に示すようにn種類のVLANを構成する。個々のVLANは (c) に示すように各スイッチから他の$n-1$個のスイッチにつながる放射状のトポロジを持つ。このように設定すると、(d) に示すように、2つのスイッチ間にはVLANの選択により$n-1$種類の経路が存在することになる。従って、スイッチx,y間を直結するリンクのみを用いる場合と比べ$n-1$倍のスイッチ間通信バンド幅を利用できることになる。

この例ではスイッチ間に最大で1つのスイッチを経由するようにVLANを設定しているが、さらにVLANの数を増やして2つ以上のスイッチを経由するVLANを用意することも可能である。

3.3 メッシュ状ネットワークとHyper Crossbar

図6にメッシュ状のネットワークにVLANルーティング法を適用した例を示す。$n \times n$のメッシュでは$2n$個のVLANを用いる。図の例では3×3のメッシュであるから6個のVLANが用いられる。図が煩雑になるため、(a) と (b) に分けて3つずつのVLANのトポロジを示している。このうち一つのVLANのトポロジを示すと (c) のようになる。適切にVLANを選択することで、xyルーティングやその迂回ルーティングなどを行うことができる。

メッシュ状のネットワークでは、格子点をスイッチとし、スイッチ間をリンクで接続することもできるし、

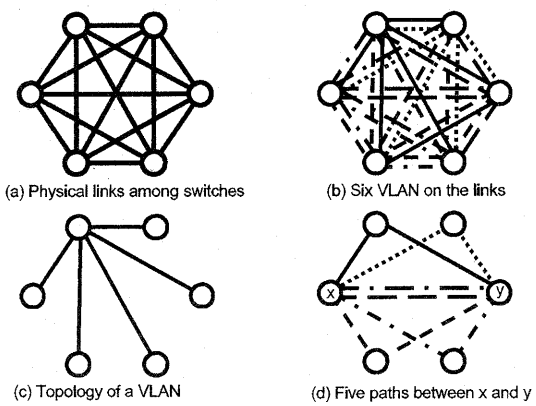


図 5　VLAN を用いた完全グラフネットワーク

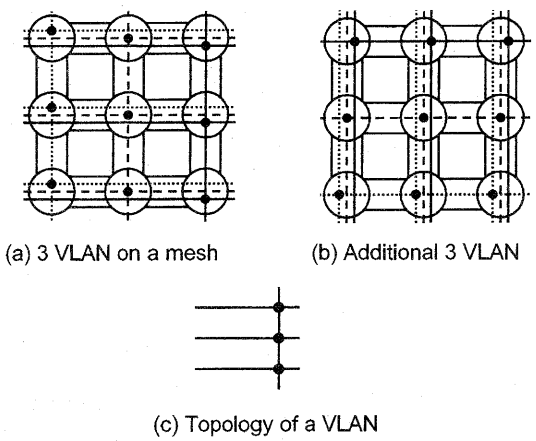


図 6　VLAN を用いた Mesh トポロジ

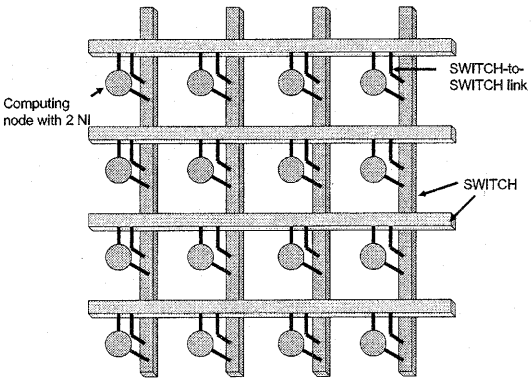


図 7　VLAN を用いた 2 次元 Hyper Crossbar

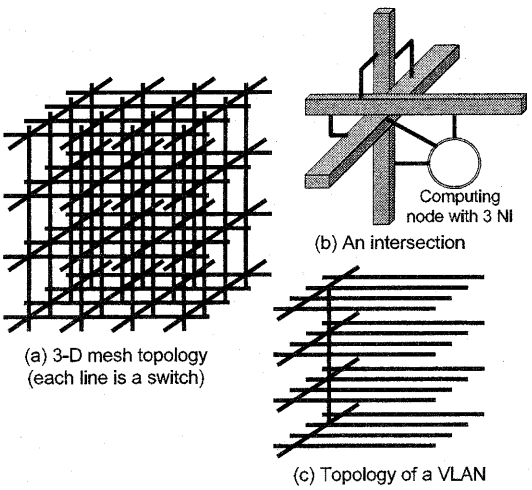


図 8　VLAN を用いた 3 次元 Hyper Crossbar

X 方向、Y 方向にそれぞれスイッチを置き、スイッチ間をリンクで接続することもできる。後者はハイパークロスバを構成することになる。図 7 に 2 次元ハイパークロスバの例を示す。ノードの接続には様々な方法が考えられるが、図では各格子点にノードを置き、X, Y 両方のスイッチに接続する例を示している。この場合、各ノードは 2 つのネットワークインタフェースを持つことになる。どちらのネットワークインタフェースからもいずれの VLAN を用いたルーティングも可能である。

図 8 に 3 次元のハイパークロスバの例を示す。(a) における X, Y, Z 各方向の線がスイッチになる。各格子点の接続は (b) のようになり、ある VLAN のトポロジは (c) のようになる。この方式で $n \times n \times n$ のハイパークロスバを構成するには、$3n^2$ 個のスイッチと同数の VLAN が必要で、n^3 のノードを接続することができる。

比較的安価に入手できる 24 ポートのスイッチで構成することを考えると、X, Y, Z のクロスバを構成する各スイッチは、格子点毎に 3 つのポートが必要だから、$8 \times 8 \times 8$ のハイパークロスバを構成することができることになる。512 ノードからなるネットワークとなり、192 個のスイッチで 192 の VLAN を用い、スイッチ間およびノードを接続するリンクの総数は 3,072 本となる。

4. 評　　価

VLAN ルーティング法の有効性を確認するために小規模なクラスタシステムを構築し、通信性能と並列ベンチマークの実行性能を評価した。本節で示す評価は、すべて以下の環境で行った。評価で用いたリンクはすべて Gigabit Ethernet である。

```
Node PC:
 Processor: Pentium4 2.4C (2.4GHz)
 Motherboard: Intel D865GLC
 Memory: 512MB DDR400
 NIC: on board Intel 82547EI(CSA interface)
 OS: RedHat 9
```

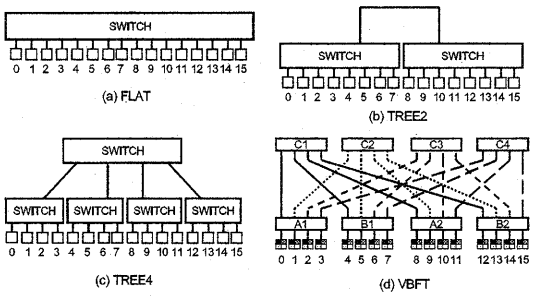


図 9 評価に用いたネットワーク

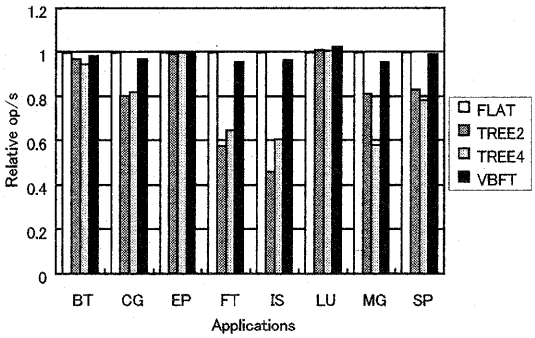


図 10 NAS 並列ベンチマークの性能

設定		通信性能 (Mbps)
U-s-U	VLAN 無で単一スイッチでノードを接続	941.0
T-s-T	Tag 付 VLAN で単一スイッチでノードを接続	938.8
T-s-U-s-T	Tag 付 VLAN のノードが繋がったスイッチ間を Tag 無で接続	938.8

表 1 2 ノード間の単方向通信性能

構造	平均ノード間バンド幅 (Mbps)	バイセクションバンド幅 (Mbps)
単純木構造	235.375	1883
Fat Tree	938.25	7506

表 2 バイセクションバンド幅

```
NIC driver: Intel e1000 5.2.16
Network switch:
 Dell PowerConnect 5224
 (Gigabit Ethernet 24-port, non-blocking)
```

評価には、図 9(d) に示す、VLAN ルーティング法による Fat tree(VBFT) ネットワークで、16 台のノードを接続したクラスタを用いた。但し、実際にはスイッチ A1 と A2、B1 と B2、C1〜C4 はそれぞれ同一のスイッチを VLAN により分割して用いている。各ノードの VLAN 仮想インタフェースは、RedHat9 標準の `vconfig` コマンドにより設定した。

並列ベンチマークの実行性能の比較対照として図 9(a) に示すように単一のスイッチに 16 台のノードを接続した構成 (FLAT)、(b) に示す 8 台のノードが接続された 2 台のスイッチを 1 本のリンクで接続した構成 (TREE2), (c) に示す 4 台のノードが接続されたスイッチをそれぞれ上位のスイッチに 1 本のリンクで接続した構成 (TREE4) についても評価した。(こちらも実際には同一スイッチを VLAN により分割している)

4.1 基本通信性能

まず、スイッチを介してノード間で通信した際の単方向通信性能を、iperf1.7.0 を用いて測定した結果を表 1 に示す。ソケットバッファサイズは 128KB とした。

VLAN を用いない Gigabit Ethernet で TCP 通信を行った場合の理論上のデータ（ペイロード）転送性能は、ACK を考慮せずに最大 949.3Mbps である。VLAN なしの結果はこれにかなり近く、理論上の限界に近い片方向通信性能が得られている。Tag 付のイーサネットフレームは通常のフレームよりも 4Byte 大きくなるため、通常の MTU サイズのフレームでは実効データ転送レートが 2.4Mbps 程度低下する計算になる。結果はこの理論値とほぼ一致している。また、スイッチにおける Tag 付フレームと Tag 無フレームの変換にはほとんどオーバヘッドがないことがわかる。

次に、図 9(d) の VBFT において、左側の 8 ノードから右側の 8 ノードに同時に iperf による片方向通信を行い、バイセクションバンド幅を測定した。すべての通信がスイッチ C1 を経由する (C2〜C4 は使用しない：単純木構造) 場合と、C1〜C4 を全て用いる (Fat Tree) 場合について計測を行った。C1〜C4 を用いる場合には、ノード 0 はスイッチ C1 経由でノード 8 に、ノード 1 はスイッチ C2 経由でノード 9 に、というように 8 離れたノードに、ノード id を 4 で除した余りに従って上位スイッチを選択して送信した。この結果を、表 2 に示す。C1〜C4 を用いた Fat Tree では、木構造の 4 倍のバンド幅となり、フルバイセクションバンド幅が得られていることがわかる。

5. 並列ベンチマークの実行性能

NAS 並列ベンチマークを実行し、性能を測定した。NAS 並列ベンチマークは ver.2.3、問題サイズは ClassB である。MPI ライブラリは LAM/MPI 7.0 を使用した。コンパイラには gcc-3.2.2 を使用し、すべての最適化は-O3 とした。

VBFT では各ノードから宛先ノード番号に従ってサイクリックにスイッチを使用するように設定した。送信元ノードと宛先ノードの組によってパスが異なるため、宛先ノードの見かけの IP アドレスが送信元ノード毎に異なることになる。このため、`/etc/hosts` に記述されるホスト名と IP アドレスの対応表を各ノー

ド毎に用意した。MPIプログラムの実行にあたって、ホスト名とIPアドレスの対応表以外に特別な設定は行なっていない。

図9の各構成での性能を比較した結果を図10に示す。Mops値を測定した結果をFLATにおける性能により正規化している。TREE2, TREE4では、問題によってFLATに対して性能が大きく低下しているが、VBFTではほぼFLATと同等の性能が得られていることがわかる。ベンチマークのうちFTとISは全体全通信を行なっており、バンド幅を必要とするベンチマークであることが知られている。図10からもこの2つでVBFTの効果が大きいことが読みとれる。

6. 関連研究

松岡[3]は、Ethernetで、大規模なスイッチを用いることなくクラスタ向けネットワークを構築することを提案し、ノードがパケットをいったん受け取りバケツリレー式にパケットを伝達することにより、複数のパスを分散して使えるとしている。しかし、この提案では、ループがある場合に発生するbroadcast storm等の問題にどのように対処するかは示されていない。

一方、森川[4]は、スイッチのルーティングテーブルを静的に設定することによりL2 Ethernetに複数の経路を設定できるとしている。しかし、これには静的にテーブルを設定する機能を持つスイッチが必要である。このような機能は標準規格には定められておらず、一般に用いることはできないと考えられる。

これに対して、VLANルーティング法では、規格により定められたVLAN機能を持つ一般的なスイッチを用いて、broadcast stormなどの問題が起きない複数パスを持つネットワークを構築できる。通常VLANは、単一のネットワーク上に複数の仮想ネットワークを構築し、それぞれのノードが参加できる仮想ネットワークを制限するために用いられるのに対し、本方式通信パスを選択するためにVLANを用いており、基本的に全ノードがすべてまたは複数のVLANに参加する。

7. おわりに

本稿では、VLANを用いることにより複数パスを持つL2 Ethernetネットワークを構築する　VLANルーティング法を提案し、これを用いたネットワークトポロジの例を示すと共に、小規模なクラスタで性能を評価して有効性を示した。VLANルーティング法を用いれば、本稿で示したもの以外にも様々なトポロジのネットワークを構築できると考えられる。IEEE803.1Qでは、Tagフィールドで最大$4094(2^12-2)$個のVLANを指定できる。本稿の評価で用いたDell社のPowerConnect 5224は最大255個のVLANを使用でき、かなり大規模なネットワークを構築できる。

VLANルーティング法では、VLANを用いることにより、物理的にはループを含むトポロジを、broadcast stormを起こすことなく利用できる。しかし、誤って単一VLANでループを構成してしまうと、broadcast stormが発生してしまう。これを防ぐためにはVLAN毎にSTPを適用することが考えられる。このような機能はSpanning Forestと呼ばれ、一部のスイッチには実装されている。本稿の評価で用いた比較的安価なスイッチでは、Spanning Forestは実装されておらず、STPはVLANの構成に関係なく物理リンクのトポロジに対して働く。従って、物理リンク上にループがある状態でVLANルーティング法を用いるには、STP機能を停止した上で使用する必要がある。さらに、STP機能を停止すると、スイッチは他から受け取ったSTP構成メッセージを、VLANの設定に関係なくなくすべてのポートに伝える。このため、STP機能が動作しているスイッチを接続すると、構成メッセージがループを循環し増殖してしまう。安価なスイッチでVLANルーティング法によるネットワークを構築する場合、このような点に注意する必要がある。

VLANルーティング法では、同一ノードが複数のIPアドレスを持ち、パスは宛先IPアドレスにより選択される。しかし、既存のMPIはノードとIPアドレスが一意に対応する場合しか想定していないため、VLANルーティング法を用いたネットワーク上で実行するためにはIPアドレスとホスト名の対応をノード毎に変更する必要があった。MPIなどのアプリケーションが指定する宛先IPアドレスと通信時に実際に使用するIPアドレスの変換を行うレイヤーを導入することによりこのような問題を解決することができると考えられる。また、このレイヤーを用いて、故障時に代替パスに切り替えることにより高い信頼性を実現することを検討していく予定である。

謝辞　本研究の一部は、新エネルギー・産業技術総合開発機構基盤技術研究促進事業（民間基盤技術研究支援制度）の一環として委託を受け実施している「大規模・高信頼サーバの研究」の成果である。

参考文献

1) Myricom, Inc. *http://www.myri.com/*.
2) Fabrizio Petrini, Wuchun Feng, Adolfy Hoisie, Salvador Coll, and Eitan Frachtenberg. The quadrics network (qsnet): High-performance clustering technology. In *Hot Interconnects 9*, August 2001.
3) Satoshi Matsuoka. You don't really need big fat switches anymore-almost. 情報処理学会研究報告 ARC-154-27, pp. 157–162, 2003.
4) 森川誠一. グリッドコンピューティングに要求される通信技術. In *NETWORLD+INTEROP 2003 TOKYO Conference Notes*, pp. 75–87, 2003.