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

工藤知宏† 松田元彦† 手塚宏史†
児玉祐悦† 建部修見† 関口智嗣†

本論文では，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の例をいくつか示すとともに，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VLANルーティング法が設計どおり動作することを示す．

VLAN-based Multi-path L2 Ethernet Network for Clusters

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

In this paper, 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 introduce 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 cluster with the VLAN-based Fat Tree network are 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安価な小規模スイッチを多数用いる必要がある．

HPCクラスタ向けネットワークには，高いバイセクションバンド幅が求められる．同一バンド幅のリンクを用いてネットワークを構成する場合，高いバイセクションバンド幅を実現するには，スイッチ間に複数のパスが必要である．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，複数のスイッチを用いる場合，Fat Treeなどの複数のパスを持つトポロジが用いられる．

Gigabit Ethernetを用いてクラスタ向けネットワークを構築する場合も，高いバイセクションバンド幅を実現するにはスイッチ間に複数パスが必要である．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．L3ルーティングを用いれば複数パスは実現可能であるが，L3機能を持つスイッチは高価であるし，トポロジによってはルーティングの設定が大変煩雑になる．また，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バイセクションバンド幅を高くすることができるが，Gigabit Ethernetの上位の規格である10 Gigabit Ethernetポートを持つスイッチは高価であり，システム全体のコストが高くなってしまう．

本論文では，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．VLANルーティング法では，ノード間あるいはスイッ

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

チ間に異なるVLANに属する複数のパスを用意し，どのVLAN上で通信を行うかにより，パスを決定する．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．

以下，まずVLANルーティング法を提案し，SANで一般的に用いられるソースルーティングとの違いを明らかにする．次に，VLANルーティング法をソフトウェアからどのように利用するかについて議論する．さらに，VLANルーティング法を用いたネットワークの一例として，Fat Tree，完全グラフ，Hyper Crossbarを紹介し，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し，VLANルーティング法が設計どおりに動作することを示す．

2. VLANルーティング法

2.1 VLANを用いた複数パスの実現

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，ノードごと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1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．STPはループを検出してループ上のあるリンクを使用しないようにするため，複数パスを利用することはできない．

VLANを用いると，この問題を解決することができる．VLANは，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．異なるVLAN間ではイーサネットフレームは伝搬しない．したがって，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．

この性質を用いて，**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を割り当てれば，いずれのVLANにデータを送るかにより，複数のパスを使い分けることができる．

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，それぞれの仮想ネットワークを独立したネットワークとして利用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，

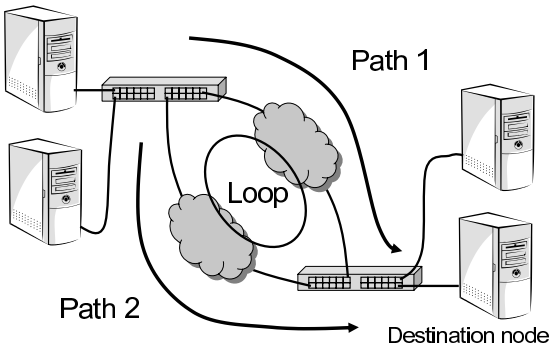

図1 スイッチ間に複数パスを持つネットワーク
Fig. 1 Multiple paths between switches.

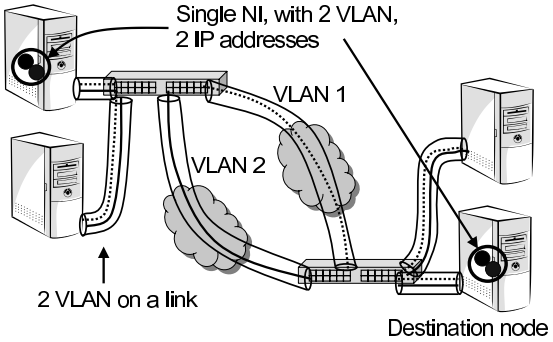

図2 VLANルーティング法による複数パスの実現
Fig. 2 VLAN-based routing and multiple paths.

基本的にいずれのノードも他のすべてのノードとL2で通信することができる．

2.2 VLANとスイッチの動作

VLANを実現するためには，イーサネットフレームにはTagと呼ばれるフィールドが付加される．IEEE803.1QによるVLANの規定では，イーサネットフレームにはTag付きのものとTagなしのものがある．Tagなしのフレームは通常のイーサネットフレームである．Tag付きのフレームでは，Tagフィールドに VLANの番号（VLAN id）が入る．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なし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付きフレームを用いなくてはならない．

VLANをサポートす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付きであれば，そのTagによって示されるVLANに属するものと見なされる．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なしであれば，あらかじめそのポートに設定されたPVID（Port VLAN id）のVLANに属

するものと見なされる．

一方，スイッチの出力ポートには，VLAN idごとにTag付きとTagなしの設定をすることができ，

- そのポートにTagなしで設定されているVLAN idのフレームは，Tagのない通常のフレームとして出力される．
- そのポートにTag付きで設定されているVLAN idのフレームは，Tag付きで出力される．
- そのポートにTag付き，Tagなしのいずれで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．

Tag付きで受け取ったフレームをTagなし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なし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付き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なし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．このため，VLANをサポートしているスイッチでは，VLANを用いた場合と用いない場合で帯域や遅延に大きな違いは生じない．

2.3 ノードの設定とVLANルーティング法の利用

RedHat9では，たとえばeth1というネットワークインタフェース（NI）に対し，

```
% vconfig add eth1 2
% vconfig add eth1 3
```

というコマンドを実行して，VLAN idが2および3の仮想インタフェース`eth1.2`，`eth1.3`を作成することができる．これらの仮想インタフェースのMACアドレスはすべてeth1のMACアドレスに等しい．また，`ifconfig`コマンドによりIPアドレスをそれぞれの仮想インタフェースに割り当てることができる．

たとえば，単一のネットワークインタフェース上に4つのVLANの仮想インタフェースを持つと**図3**のような構成となる．

各ノードは，そのノードがデータを送受する必要があるVLANすべてについて仮想インタフェースを持つ．そのノードが送受に関与しないVLANの仮想インタフェースは持たなくてよい．送信側ノードは，どの仮想インタフェースから送信するかによりVLANを選択する．

PM/Ethernet[3]で用いられるようなL2でデータを送受する方式でVLANルーティング法を用いるには，送信する際に用いる仮想インタフェースによりVLANを選択し，あて先ノードはMACアドレスにより指定すればよい．現在のPM/Ethernetの実装では，あて先に従って仮想インタフェースを指定する機能がないため，若干の拡張が必要である．

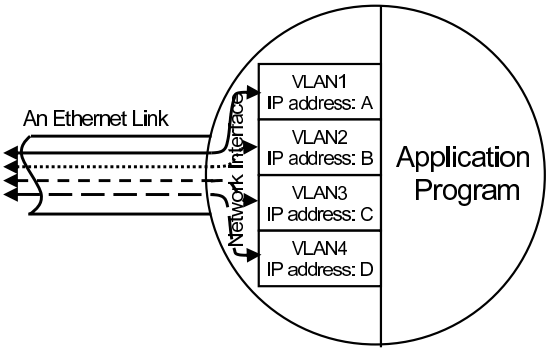

図3 ノード上のVLAN仮想インタフェース
Fig. 3 VLAN virtual interfaces on a node.

一方，IPを用いた通信では，仮想インタフェースはそれぞれ異なるIPアドレスを持つため，IPアドレスのみでVLANとあて先ノードの指定を行うことができる．送信側のノードでは，宛先IPアドレスごと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VLAN idごとにネットワークを構成し，ネットワークごとに経路テーブルに登録すれば，経路テーブルへの登録はVLAN数分のエントリですむ．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指定することになる．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．

2.4 VLANルーティングの振舞い

VLANの中では，イーサネットフレームは通常のEthernetの仕組みで伝搬される．Ethernetでは，イーサネットフレーム中のあて先MACアドレスによりあて先が指定される．ネットワークの中間にスイッチ（ブリッジデバイス）があれば，スイッチはMACアドレステーブルにより，どのポートに受け取ったフレームを送り出せばよいかを知る．MACアドレステーブルにフレームのあて先MACアドレスが登録されていない場合には，当該VLAN idが登録されているすべてのポートにフレームを送る（これをフラッディングと呼ぶ）．この機能により，送り出されたフレームは，あて先MACアドレスを持つデバイスがそのVLAN上にあれば，そのデバイスに確実に送られる．一方，スイッチは，受け取ったイーサネットフレームの送り

元 MAC アドレスを，入力ポート番号とともに MAC アドレステーブルに登録する．このため，いったんフラッディングが起きた後は，フラッディングの原因になったフレームを受け取ったデバイスが返答のためにフレームを送出すれば，そのデバイスの MAC アドレスがテーブルに登録されて，次回以降フラッディングは起こらず，送信元とあて先の間のパス上でのみフレームが受け渡される．

Myrinet などの SAN では，送信元が通信パスを指定するソースルーティングが用いられる．ソースルーティングでは，送信側が任意のパスを選択することが可能である．これに対して VLAN ルーティング法では，VLAN により送信されたフレームが通ることができる部分ネットワークが決められ，その中ではスイッチの学習機能によりルーティングされることになる．したがって，あるパスでデータを送るには，そのパス全体を含む VLAN を設定すればよい．しかし，利用できる VLAN の数には限りがある（たとえば本論文の評価で用いた Dell 社 PowerConnect5224 では最大 255）ため，全体として少ない VLAN 数で効率良くトラフィックを分散できるような構成が望ましい．

SAN 同様，VLAN ルーティング法においても同一の物理的なネットワークトポロジ上で複数のルーティングがありうる．上記の制約に従って，構築するクラスタシステムに適したトポロジとルーティングを選択する必要がある．

3. VLAN ルーティング法を用いたネットワークトポロジの例

VLAN ルーティング法では様々なトポロジとそのうえでのルーティングを実現できる．本章では，VLAN ルーティング法を用いたトポロジの例をいくつかあげる．ここであげるトポロジはごく一例であり，様々なトポロジが実現可能である．

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 4** に示す．下部のスイッチに $2n$ ポートのスイッチを用いると，n 個の下部スイッチと，n 個の n ポートの上部スイッチを用いて，n^2 のノードを接続する Fat Tree を構成できる．全体で n 種類の VLAN を用いることになる．静的に負荷を分散する場合には，送信元と宛先との id などを基に n の剰余により使用する VLAN を選択する方式が考えられる．

図 4 で，ノードと下部のスイッチの間は Tag 付きフレームが交換される．上部と下部のスイッチ間では，

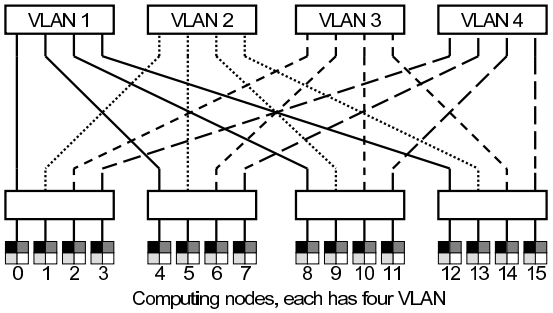

図 4 VLAN を用いた Fat Tree トポロジ
Fig. 4 VLAN-based Fat Tree.

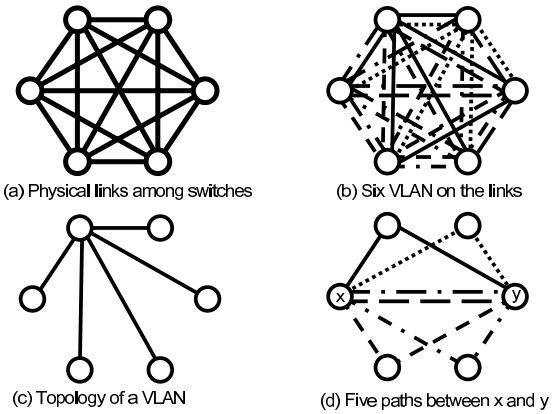

図 5 VLAN を用いた完全グラフネットワーク
Fig. 5 Complete graph network with VLAN-based routing.

各リンクには単一の 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 なしのポート単位の VLAN とすれば，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．たとえば下部に 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ることを想定している．この物理リンク上に (b) に示

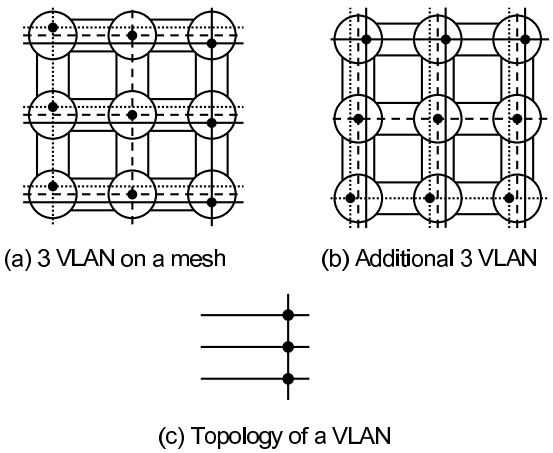

図 6 VLANを用いたMeshトポロジ
Fig. 6 VLAN-based mesh.

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(直結リンク上では2つのVLANのいずれを用いることもできる)．したがって，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．これは，通信するノードの組合せにより異なるVLANを用いることにより実現できる．

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1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，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．

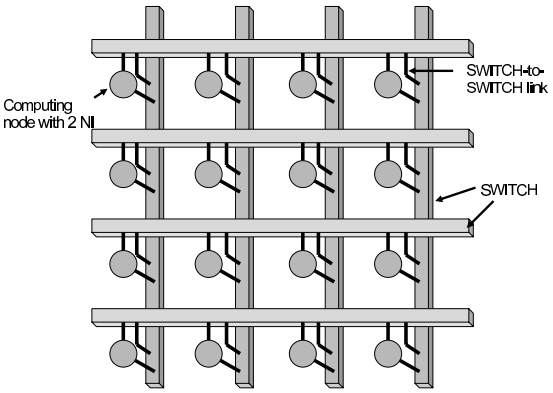

図 7 VLANを用いた2次元Hyper Crossbar
Fig. 7 VLAN-based 2D Hyper Crossbar.

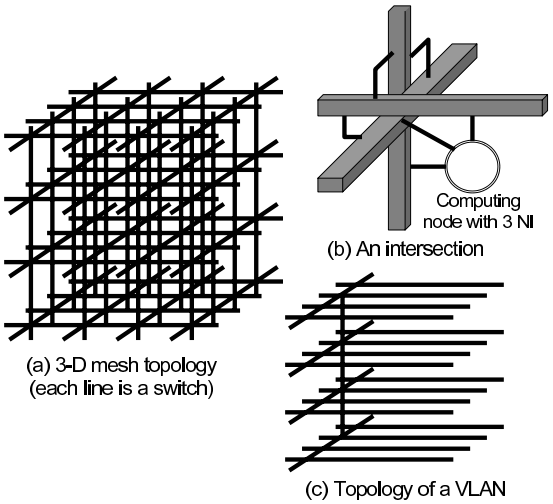

図 8 VLANを用いた3次元Hyper Crossbar
Fig. 8 VLAN-based 3D Hyper Crossbar.

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) のようになる．

このトポロジ上でルーティングを実現するVLANの構成の一例を考える．この例で，スイッチは x，y，z の直行する3軸のいずれかに平行におかれていると考えると，各スイッチには，z 軸に平行なものには $(x, y, -)$，x 軸に平行なものには $(-, y, z)$，y 軸に平行なものには $(x, -, z)$ という番号をつけることができる．$n \times n \times n$ のハイパークロスバでは x，y，z はそれぞれ 1〜n の範囲のいずれかの整数になる．

ここで，z 軸上のスイッチ $(x1, y1, -)$ を中心とし

た VLAN を以下のスイッチ上に設定する．ここで，$*$ は 1〜n の範囲のすべての整数が入る．

$$(x1, y1, -), (x1, -, *), (-, *, *)$$

これをすべての z 軸上のスイッチについて持つ．同様にスイッチ $(-, y2, z1)$ を中心とした VLAN を

$$(-, y2, z1), (*, y2, -), (*, -, *)$$

に，スイッチ $(x2, -, z2)$ を中心とした VLAN を

$$(x2, -, z2), (-, *, z2), (*, *, -)$$

に設定する．これらの VLAN のうちの 1 つのトポロジを例として示すと (c) のようになる．各スイッチごとに VLAN を持つので，全体では $3n^2$ 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，格子点ごとに 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．全体として 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

本章では，VLAN を用いた場合のスイッチ性能の評価結果を示すとともに，小規模なクラスタシステムを用いた性能評価により，VLAN ルーティング法が設計どおりに機能することを示す．本章で示す評価は，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
- NI: on board Intel 82547EI(CSA interface)
- OS: RedHat 9
- NI driver: Intel e1000 5.2.16

Network switch:
- Dell PowerConnect 5224
- (Gigabit Ethernet 24-port, non-blocking)

4.1 スイッチの性能

すでに述べたように，VLAN をサポートしているスイッチでは，VLAN を特に用いない状態では，実際には全ポートで同一 VLAN id の Tag なしのフレームを入出力しており，内部では VLAN id を付加して処理を行っている．このため，VLAN を用いてもスイッチ性能の低下はほとんどないと考えられる．

表 1 2 ノード間の単方向通信バンド幅
Table 1 Uni-directional bandwidth between two nodes.

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

表 2 スイッチのレイテンシ
Table 2 Latency of a switch.

	min. (μs)	avg. (mus)	max. (μs)
U-sw-U	2.46	3.60	4.06
T-sw-T	2.30	3.69	4.13
T-sw-U	2.40	3.79	4.22
U-sw-T	2.53	3.62	4.06

まず，スイッチを介して 2 ノード間で通信した際の単方向通信バンド幅を，iperf 1.7.0 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 128 KB 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 941.5 Mbps である．VLAN なしの結果はほぼこれに等しく，理論上の限界に近い片方向通信性能が得られている．Tag 付きのイーサネットフレームは通常のフレームよりも 4 Byte 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 2.4 Mbps 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 付きフレームと Tag なし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分かる．

次に，スイッチのレイテンシを測定した結果を**表 2** に示す．U は Tag なし，T は Tag 付きのフレームを意味し，`T-sw-U` はスイッチに Tag 付きのフレームを入力し Tag なしで出力したときのレイテンシを指す．それぞれ，ping（ICMP メッセージ）を送った場合のレイテンシを GNET-1 [4] を用いて 1,000 回測定した場合の最小，平均，最大値である．GNET-1 では 32 ns の精度でレイテンシを測定することができる．この結果から，遅延にはかなりばらつきがあるが，VLAN 機能のあるスイッチで特に VLAN を用いない場合（`U-sw-U`）と VLAN を用いた場合ではレイテンシに大きな差は

☆ TCP オプションヘッダにタイムスタンプを含む．

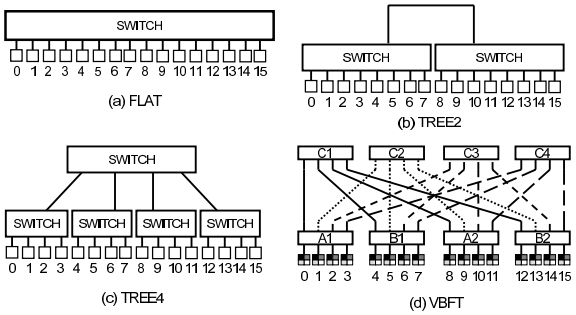

図 9 評価に用いたネットワーク
Fig. 9 Evaluated networks.

ないことが分かる．

4.2 小規模クラスタの性能評価

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．ただし，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象として図 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）．

4.2.1 基本通信性能

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 をすべて用いる（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 3** 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分かる．

4.2.2 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．

表 3 バイセクションバンド幅
Table 3 Bi-section bandwidth.

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	235.4	1,883
Fat Tree	938.3	7,506

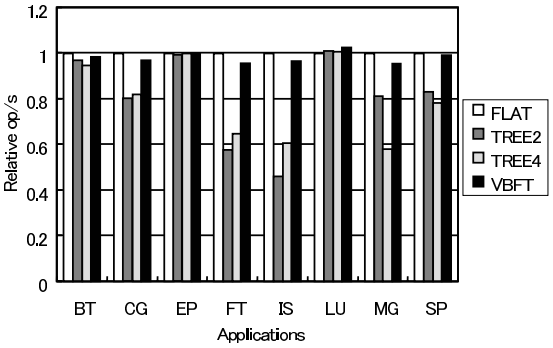

図 10 NAS 並列ベンチマークの性能
Fig. 10 NAS parallel benchmark results.

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ごと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ごとに用意した．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分かる．ベンチマークのうち FT と IS は全対全通信を行っており，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取れる．

5. 関連研究

Matsuoka[5]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 など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[6]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，本方式では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複数の VLAN に参加する．

L3 ルーティングを用いれば，VLAN ルーティング法を用いた L2 ネットワークと同様に，複数パスを持つネットワークを構築することができる．Cisco systems 社や Force10 networks 社では，L3 ルーティングを用いた，上位に複数のスイッチが存在する Fat Tree 状のネットワークを，クラスタ用として提案している．これらの L3 ネットワークでは，パスは送信ノードと受信ノードの組合せにより設定され，実行時に送信側ノードがパスを明示的に選択することは想定されていない．ただし，VLAN ルーティング法で用いているのと同様に仮想ネットワークインタフェースにより単一ネットワークインタフェースに複数の IP アドレスを持たせれば，明示的な選択は可能である．

しかし，L3 ルーティングを用いた複数パスネットワークには，

(1) L3 ルーティングをサポートするスイッチは，L2 スイッチングのみをサポートするスイッチと比べ高価である，
(2) L3 ルーティングは，L2 スイッチングよりもオーバヘッドが大きいことがある，
(3) L3 ルーティングでハイパークロスバのような複雑なトポロジを構成するのは大変煩雑である，

といった問題点がある．このうち (3) については，L3 ルーティングでは，スイッチの各ポートごとに，そのポートにルーティングするノードの IP アドレスを登録しなくてはならない．すなわち，各ノードのネットワークインタフェースに複数の IP アドレスを割り付けたうえで，各アドレスあてのトラフィックが通過する可能性がある全スイッチの全出力ポートにそのアドレスを登録をしなくてはならない．これに対して VLAN ルーティング法では，スイッチに VLAN の設定を行えば，L2 スイッチの機能によりルーティングテーブルはスイッチが自動的に学習する．すなわち，各スイッチは，ポートごとに受け取ったフレームのソース MAC アドレスを記録し，その MAC アドレスあてのフレームはそのポートにルーティングするようこのため，個々の IP アドレスごとのルーティングを設定する必要がなく，L3 ルーティングを用いる場合と比べて設定が簡単である．

6. おわりに

本論文では，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ともに，小規模なクラスタで性能を評価して有効性を示した．VLAN ルーティング法では，ソースルーティングとくらべると VLAN 数による制約があるものの，様々なトポロジ上で様々なルーティングが可能である．本論文ではその一例を示したにすぎない．どのようなトポロジ，VLAN 構成が良いかは，目的などによっても異なり，今後の研究課題である．なお，IEEE803.1Q では，Tag フィールドで最大 4,094（$2^{12}-2$）個の VLAN を指定できる．実際のスイッチでは，たとえば本論文の評価で用いた 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．また，重なりのない VLAN には同一の id を割り当てることができる．この性質を用いて各 VLAN のトポロジを選定すれば，使用する VLAN id 数を減らすことができる．

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 ごとに STP を適用する一般に Spanning Forest と呼ばれる機能が必要である．Cisco systems 社は，PVST（Per VLAN Spanning Tree）という独自機能により Spanning Forest をサポートしている．しかし，本論文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．したがって，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

うえで使用する必要がある．さらに，STP機能を停止すると，スイッチはほかから受け取った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．Spanning Forestをサポートしていないスイッチでで 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ルーティング法で実現する複数パスは，SANや超並列計算機用結合網と同じく，通信するノードの組によって静的に利用するパスを設定する方法，実行時に動的にパスを選択する方法，同一通信（コネクション）のデータを分割して複数パスで通信する方法などの利用法が考えられる．

VLANルーティング法により実現できるトポロジ，ルーティングは多種多様である．2章で示したように，VLANルーティング法ではVLANによりネットワーク上に利用可能なパスを設定する形でルーティングが決まること，利用できるVLANの数に制限があることにより，ソースルーティングなどを用いるSANとくらべて実現可能なトポロジ，ルーティングに制約がある．どのようなトポロジ，ルーティングが良いかは，利用できる機材の制約や，システムが対象とする問題の性質などより変わり，今後の研究課題である．

謝辞 性能評価にご協力いただいた（株）シナジェテックの清水敏行氏，産業技術総合研究所の岡崎史裕氏に感謝します．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) Petrini, F., Feng, W., Hoisie, A., Coll, S. and Frachtenberg, E.: The quadrics network (qsnet): High-performance clustering technology, *Hot Interconnects 9* (Aug. 2001).
3) 住元真司，堀 敦史，手塚宏史，原田 浩，高橋俊行，石川 裕：既存OSの枠組みを用いたクラスタ向け高速通信機構の実現，情報処理学会論文誌，Vol.41, No.6, 情報処理学会 (2000).
4) 児玉祐悦，工藤知宏，佐藤博之，関口智嗣：ハードウェアネットワークエミュレータを用いたTCP/IP通信の評価，情報処理学会研究報告 HPC-95-9，pp.47–52 (2003).
5) Matsuoka, S.: 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).
6) 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，*NETWORLD+INTEROP 2003 TOKYO Conference Notes*, pp.75–87 (2003).

(平成15年10月 9 日受付)
(平成16年 1 月29日採録)

工藤 知宏（正会員）

1991年慶應義塾大学大学院理工学研究科博士課程単位取得退学．東京工科大学助手，講師，助教授を経て，1997年より新情報処理開発機構並列分散システムアーキテクチャつくば研究室長，2002年より産業技術総合研究所グリッド研究センタークラスタ技術チーム長．博士（工学）．並列処理，通信アーキテクチャに関する研究に従事．電子情報通信学会，IEEE CS各会員．

松田 元彦（正会員）

1988年京都大学理学部卒業．同年住友金属工業（株）入社．1995年から1999年まで技術研究組合新情報処理開発機構に出向．2003年より独立行政法人産業技術総合研究所．現在同研究所グリッド研究センター主任研究員．工学博士．並列計算システム，クラスタシステムおよびグリッド環境での高性能計算に関する研究に従事．

手塚 宏史

1957年生．1985年ソニー株式会社入社．UNIXワークステーションの開発に従事．1989年ソニーコンピュータサイエンス研究所勤務．1993年北陸先端科学技術大学院大学研究生．1995年新情報開発機構研究員．ワークステーション/PCクラスタの開発に従事．2001年株式会社オムニサイソフトウエア入社．2003年より産業技術総合研究所グリッド研究センター勤務．

児玉　祐悦（正会員）

1962 年生．1986 年東京大学工学部計数工学科卒業．1988 年同大学大学院情報工学専門課程修士課程修了．同年通産省電子技術総合研究所入所．2001 年独立行政法人産業技術総合研究所に改組．現在，同研究所グリッド研究センター主任研究員．データ駆動やマルチスレッド等の並列計算機システムの研究に従事．特にプロセッサアーキテクチャ，並列性制御，動的負荷分散，並列探索問題等に興味あり．博士（工学）．情報処理学会奨励賞，情報処理学会論文賞（1990 年度），市村学術賞（1995 年）等受賞．電子情報通信学会，IEEE CS 各会員．

建部　修見（正会員）

1969 年生．1992 年東京大学理学部情報科学科卒業．1997 年同大学大学院理学系研究科情報科学専攻博士課程修了．同年電子技術総合研究所入所．理学博士．組織変更により現在独立行政法人産業技術総合研究所グリッド研究センターに所属．グリッドコンピューティング，並列数値アルゴリズム，並列計算機システムの研究に従事．日本応用数理学会，ACM 各会員．

関口　智嗣（正会員）

1959 年生．1982 年東京大学理学部情報科学科卒業．1984 年筑波大学大学院理工学研究科修了．同年電子技術総合研究所入所．以来，データ駆動型スーパーコンピュータ SIGMA-1 の開発等の研究に従事．2001 年独立行政法人産業技術総合研究所に改組．2002 年 1 月より同所グリッド研究センターセンター長．並列数値アルゴリズム，計算機性能評価技術，グリッドコンピューティングに興味を持つ．市村賞，情報処理学会論文賞受賞．グリッド協議会会長．日本応用数理学会，ソフトウェア科学会，SIAM，IEEE，つくばサイエンスアカデミー各会員．