スイッチでタグ付けを行うVLANルーティング法

大塚智宏[†] 鯉渕道紘[††]
工藤知宏[†††] 天野英晴[†]

コスト対性能比の高いPCクラスタでは，クラスタ内ネットワークとしてEthernetを採用している場合が多いが，通常，L2 Ethernetのトポロジはツリー構造に限られる．近年提案されたVLANルーティング法は，VLANによる複数の論理的な木構造を用いることで，Ethernetにおいて並列処理に適したさまざまなトポロジを利用できるようにする手法である．しかし，PCクラスタで使われる軽量通信ライブラリはEthernetフレームのVLANタグ付けに対応していない場合が多く，VLANルーティング法をそのまま適用できないという問題がある．本論文では，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をサポートしていない場合でもVLANルーティング法を利用できるようする手法を提案する．また，提案手法を効果的に利用するために，特にIEEE 802.3xフロー制御を使用する場合には，経路を設定する際にデッドロックフリールーティングが用いられるべきであることも明らかにする．NAS並列ベンチマークによる評価の結果，提案手法によるトポロジは，フラットなトポロジに近い性能を示すことが分かった．

Switch-tagged VLAN-based Routing for PC Clusters with Ethernet

Tomohiro Otsuka,[†] Michihiro Koibuchi,[††] Tomohiro Kudoh[†††]
and Hideharu Amano[†]

High performance-per-cost PC clusters usually employ Ethernet for intra-cluster networks, though the topology of L2 Ethernet is usually a simple tree. VLAN-based routing method presents various topologies suitable for parallel processing by using a combination of multiple logical trees in a L2 Ethernet network. However, since the communication library used in current PC clusters does not usually support VLAN technology, the original VLAN-based routing method cannot be applied to such PC clusters. In this paper, we propose switch-tagged VLAN-based routing method for PC clusters whose system software does not support VLAN tags. To efficiently use the proposed strategy, a deadlock-free routing should be applied to paths between hosts, especially when IEEE 802.3x link-level flow control is used. Evaluation results using NAS Parallel Benchmarks showed that performance of topologies supported by the proposed method is comparable with that of an ideal flat topology.

1. はじめに

Ethernetはその高いコストパフォーマンスにより，最近ではPCクラスタの内部ネットワークとしても利用されている．初期のベオウルフ型クラスタと違い，最近のEthernetを用いたクラスタでは，システムエリアネットワーク（SAN）[1),2)]で用いられるゼロコピー通信やワンコピー通信を実装したシステムソフトウェア[3),4)]を利用することができ，低遅延のノード間通信を実現している．また，高スループットのスイッチも容易に入手できるようになってきており，10 Gigabit Ethernet（10 GbE）の標準化など，CPUパワーの増加にともなってリンクバンド幅も急速に大きくなっている．これらの点から，EthernetはHPC向けPCクラスタのインタコネクトとしても有力な候補の1つとなっている．

しかし，現状では，Ethernetをインタコネクトに用いたPCクラスタのほとんどは，単純なツリー状トポロジを採用している．これは，L2 Ethernet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．ツリー状ネットワークにはトラフィックが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，リンク集

† 慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
†† 国立情報学研究所
National Institute of Informatics（NII）
††† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology（AIST）

約化だけではツリー状ネットワークの欠点を補いきれず，Ethernet を用いた PC クラスタは大規模化に向いていないとされてきた．PC クラスタの研究において，Ethernet のトポロジというものがこれまでにほとんど研究されてこなかったのもこのことが大きな原因と考えられる．

近年，工藤らによって提案された VLAN ルーティング法[5] は，IEEE 802.1Q 標準のタグ VLAN 技術を応用することで，Ethernet にお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手法である．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．しかし，以下のようにすることで，ホスト間に複数の経路を用意し，ネットワークのスループットを向上させるために利用することができる．すなわち，1 つのホストが複数の VLAN グループのメンバになるようにしておき，各 VLAN トポロジにそれぞれ異なるリンク集合を割り当てる．こうすると，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できるようになる．

各 VLAN トポロジは論理的にはツリー状ネットワークであるが，異なるリンク集合を持つ複数の VLAN を用いることで，Ethernet ネットワーク全体のトポロジはツリー状ネットワークから解放され，さまざまな物理トポロジを構築できるようになる．各経路はある 1 つの VLAN に属しているため，送信ホストは VLAN 番号（VLAN ID）を指定することで経路を選択しフレームを送信する．VLAN タグを格納した Ethernet フレームは，指定された VLAN 内で通常の L2 Ethernet の動作に従って転送される．

しかし，VLAN ルーティング法は，そのままでは現在一般に用いられている PC クラスタに対して適用するのは難しい．これは，PC クラスタのノード間通信に用いられる通信ライブラリが VLAN タグ付けに対応していない場合が多いからである．TCP/IP であれば VLAN に対応している場合がほとんどであるが，TCP はノード間通信に用いるには遅延が大きいためパフォーマンス上問題となる場合が多い．また，IP を使用する場合，VLAN ごとの仮想インタフェースにそれぞれ IP アドレスを割り振る必要があるが，MPI などの IP 上の通信ライブラリの実装では，ホストの指定に IP アドレス（またはそれに bind されたホスト名）を用いることが多く，同一のホストに複数の IP アドレス（ホスト名）がある状態を扱うのが難しいという大きな問題がある．これらの理由により，VLAN ルーティング法を利用した大規模なクラスタというのはまだ構築例がなく，小規模なクラスタによる手法の有効性の評価にとどまっているのが現状である．

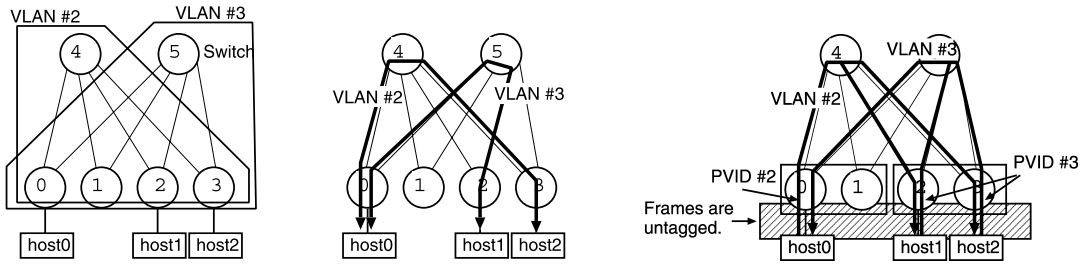
これらの問題に対し，三浦らは，用いる VLAN ID を MAC アドレスを基に決定する Linux 用デバイスドライバを開発し，クラスタシステムソフトウェアなどの上位レイヤに手を加えずに VLAN ルーティング法を利用した PC クラスタを実現している[6),7)]．

本論文では，VLAN ルーティング法を利用した PC クラスタを実現するもう 1 つの方法として，ホストではなくスイッチにおいて VLAN タグ付けを行う柔軟性の高い方法を提案する．ホストが VLAN タグを扱わないため，本提案手法は，OS などを含めたソフトウェア環境に依存することなく，たとえば VLAN をサポートしていないシステムソフトウェアや通信ライブラリを採用している PC クラスタにも適用できる．提案手法を用いることで，ホスト側のソフトウェアをいっさい変更することなく，VLAN ルーティング法によってサポートされるさまざまなトポロジを構築することができるようになる．また，提案手法を効率的に利用するために，特に IEEE 802.3x リンクレベルフロー制御を使用する場合においては，デッドロックフリーが保証されるような経路を設定すべきであることを明らかにする．

以下，まず 2 章において，提案手法の概要を説明する．3 章では，提案手法によって実現されるトポロジの例を示し，その性能要因について検討する．また，4 章では，VLAN ルーティング法を導入することで発生するデッドロックの問題について議論し，提案手法を効率的に利用するためにデッドロックフリールーティングが有効であることを示す．5 章では，実際に PC クラスタ上でさまざまなトポロジを構築して性能評価を行い，提案手法の有効性を検証した結果を示す．6 章で関連研究について述べ，最後に 7 章でまとめを述べる．

2. スイッチでタグ付けを行う VLAN ルーティング法

VLAN ルーティング法は，IEEE 802.1Q で標準化されているタグ VLAN を利用することで，Ethernet においてさまざまなトポロジを構築可能にする手法である．工藤らは，VLAN ルーティング法を提案した論文において，VLAN ルーティング法の動作原理を示し，少数のスイッチによる比較的単純なトポロジを TCP/IP ベースの通信ライブラリを用いて評価した結



(a) VLAN assignment (b) Host-tagged routing (c) Switch-tagged routing

図 1 Fat ツリーにおける VLAN ルーティングの例
Fig. 1 VLAN-based routing on a fat tree.

果を報告している[5]．この評価において，経路を選択するための VLAN 番号（VLAN ID）は，ID に関連付けられた仮想インタフェースの IP アドレスによって示される．

一方，現在の Ethernet を用いた高性能 PC クラスタは，TCP/IP をバイパスする軽量な通信ライブラリを利用している場合が多い．しかし，残念ながら，このようなシステムソフトウェアは通常 VLAN をサポートしていないため，VLAN ルーティング法を利用するにはライブラリに手を加える必要があるという問題点がある．

本章では，スイッチへのフレーム入力の際に VLAN タグを付加することで，VLAN をサポートしない通信ライブラリからでも VLAN ルーティング法を利用できるようにする手法を提案する．本手法は，IEEE 802.1Q 標準で定義された VLAN スイッチの機能のみを用いて実現するため，一般的な VLAN 対応の Ethernet スイッチ以外に特殊な機器やソフトウェアはいっさい必要なく，すべて既存のものを利用できる．

2.1 VLAN スイッチの動作

VLAN スイッチにおける VLAN タグ付けの動作は，以下のとおりである．まず，タグ付けされていないフレームがポートに入力された場合，そのポートのデフォルトの VLAN 番号（ポート VLAN ID，PVID）でタグ付けされる．タグ付けされたフレームが入力された場合は何も行わない．

一方，スイッチから出ていくフレームがタグ付けされているかどうかは，ポートごとの設定に依存する．ポートが “タグ付き” の VLAN メンバである場合，出力フレームはその VLAN 番号でタグ付けされている．ポートが “タグなし” の VLAN メンバである場合，出力フレームはタグ付けされていない．

2.2 提案手法によるルーティングの動作

提案手法では，ある 1 つのホストからフレームを送る場合の経路は，宛先によらずすべて単一の VLAN に属するように設定する．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ポートに対し，以下の 2 種類の設定を行う．

- ポートの 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．

提案手法の例を**図 1** (c) に示す．図において VLAN #2 はスイッチ 0，1，2，3，4，VLAN #3 はスイッチ 0，1，2，3，5 によって構成されており（図 1 (a)），スイッチ 0，1 のホスト側ポートには PVID #2 が，スイッチ 2，3 のホスト側ポートには PVID #3 が割り当てられている．ホスト 0 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 #2 を付与され，すべての宛先について VLAN #2 によってルーティングされる．そして，末端のスイッチ 1，2，3 の出力ポートにおいて VLAN タグ #2 を除去する．一方，ホスト 1 から送出されたフレームは同様の方法で VLAN #3 によってルーティングされる．

このようにすることで，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が相互に通信できるようになる．同じ経路集合によるパケット転送は，従来の VLAN ルーティング法[5] でも実現可能である（図 1 (b)）が．VLAN タグの処理の仕方が異なる．

フレームを転送する際の動作の詳細は，以下のとおりである．まず，送信側ホストは，通常の（タグ付けされていない）フレームを，IP アドレスや MAC アドレスで宛先ホストを指定することによって送出する．スイッチポートに入力される際に，フレームはそ

のポートに設定された PVID によってタグ付けされるため，以後はタグによって示される VLAN に属すると見なされる．これにより，ホストにおいてつねに単一の VLAN タグを付加してフレームを送出するのと同じ効果が得られる．スイッチ間の転送においては，フレームは通常の VLAN ルーティング法と同様，L2 Ethernet の動作に従って転送される．最後に，受信側ホストに接続されたスイッチポート（“タグなし” の VLAN メンバに設定されている）から出力される際に，フレームのタグは除去される．受信側ホストはタグなしのフレームを受け取ることになるため，システムソフトウェアでは通常どおりこれを処理するだけで済む．

2.3 提案手法の制限事項

提案手法では，その性質上，従来の VLAN ルーティング法に比べてとりうる経路集合に制限がある．すなわち，あるホストからの経路はすべて 1 つの VLAN に属していなければならない，ということである．たとえば，従来の VLAN ルーティング法による図 1 (b) に示される経路集合は，提案手法では実現できない．ホスト 0 からの経路が，VLAN #2 と#3 の 2 つの VLAN を使っているからである．

この制限により，不規則なトポロジにおいては，提案手法を用いてうまく経路をバランスさせるのは難しい．しかし，並列計算機で用いられているような規則的なトポロジであれば，うまく分散された経路を与える VLAN 集合を割り当てるのは比較的容易である．これは 3 章で議論する．

提案手法を用いる場合のもう 1 つの問題は，スイッチの MAC アドレス学習に関するものである．Ethernet 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，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，スイッチはフレームを VLAN メンバとなっている全ポートから出力するため（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ともなわずにフレームの交換が実現されるようになる．

しかし，VLAN ルーティング法では複数の VLAN を利用するため，ホスト A から B への経路と B から A への経路で異なる VLAN を使っている場合が当然考えられる．MAC アドレスの学習は VLAN ごとに独立して行われるため，このような場合，それぞれの経路の中間スイッチ群は，たとえ 2 つの経路がまったく同じスイッチの集合から構成されていても，宛先ホスト側の MAC アドレスの学習が不可能となってしまう．

この問題は，従来の VLAN ルーティング法の場合にもあてはまるが，提案手法の場合，各経路の往路と復路に同じ VLAN を割り当てようとすると，各ホストからの経路がすべて 1 つの VLAN に属していなければならないために，全体として 1 つの VLAN しか使用できないことになってしまう．幸い，最近の商用 Ethernet スイッチでは，DELL PowerConnect 5324 のような比較的安価なものであっても，MAC アドレステーブルの静的な設定ができるようになっているものが多い．このため，本提案手法では，各スイッチにおいて静的に MAC アドレステーブルに設定することを前提とする．

3. 提案手法における VLAN 割当て

本章では，スイッチでタグ付けを行う VLAN ルーティング法における VLAN 割当て法を示し，その性能要因を検討する．また，さまざまなトポロジにおける VLAN 割当ての例を示す．

3.1 VLAN 割当て方法

提案手法では，2.3 節で述べた制限事項を満たすすべての経路集合は，ネットワークが識別できる VLAN 数が十分にあれば実装することができる．

v_{max} 個の VLAN を識別☆できる n 台のスイッチで構成されるネットワークを仮定した場合，提案手法における VLAN 割当ては以下のような手続きとなる．

(1) ネットワーク内に，1 つのスイッチを起点とするすべての経路を含む論理ツリーを n 個作成する．
(2) もし，$n \leq v_{max}$ である場合，各 VLAN に異なる論理ツリーを割り当て，本手続きを終了する．
(3) もし，同一形状の論理ツリーが複数あれば，それらを 1 つの論理ツリーとする．そして，(2) に戻る．もし，同一形状の論理ツリーがなければ，その経路集合は，対象とするネットワークにおいては実装できないため，手続きを終了する．

☆ IEEE 802.1Q タグフィールドでは 4,094 個の VLAN を識別できるが，低価格のスイッチは，多くの場合，数百個の VLAN までしか識別できない．

3.2 VLAN ルーティングの設計

不規則なトポロジを含む一般のトポロジを仮定した場合の VLAN 割当て方法は前節で示したが，本節では，VLAN ルーティング法を用いた Ethernet ネットワークの性能要因について検討し，典型的なトポロジに提案手法を適用した場合の VLAN 割当て方法を示す．

3.2.1 性 能 要 因

3.2.1.1 経路のホップ数

Ethernet は SAN と異なり，スイッチング技術としてストアアンドフォワード方式を採用している場合が多い．ストアアンドフォワード方式は，カットスルー方式に比べて遅延が大きいため，Ethernet は SAN よりもスイッチング遅延がかなり大きくなる傾向にある．たとえば，典型的な SAN のスイッチである RHiNET-2 スイッチは，スイッチングに最小 270 ns しか必要としないのに対し[8]，Ethernet スイッチの遅延はその約 10 倍（2 μs から 4 μs 程度）である．

SAN や並列計算機の結合網では，経路の平均ホップ数がルーティングアルゴリズムの重要な性能要因として知られている．Ethernet の場合，ストアアンドフォワード方式でフレームを転送するため，各スイッチでの遅延がすべて加算されることになり，経路のホップ数は SAN の場合に比べてより重要な性能要因になると考えられる．そのため，本論文における VLAN ルーティング法は最短経路をとるように設計する．

3.2.1.2 経路の衝突

MPI などのメッセージパッシングモデルによる並列プログラミングでは，集団通信がしばしば用いられる．このため，QsNET など，SAN の中にはハードウェアによるマルチキャストがサポートされているものがあり，ブロードキャストやマルチキャスト操作において転送されるフレーム数を削減することができる[9]．

一方，Ethernet では，ユニキャストを繰り返すことでのみマルチキャストを実現できる．MPI では，MPI_Alltoall や，MPI_Reduce，MPI_Barrier などのさまざまな集団通信操作が用いられ，これらの性質がトポロジやルーティングアルゴリズムを設計するうえで重要な要素となる．ここでは，その中でも一度に大量の通信が発生する MPI_Alltoall（全対全通信）に着目する．Alltoall はすべてのプロセスが他のすべてのプロセスと通信を行うため，ユニキャストを基にした実装では最適化が難しい．そこで，本論文では，全対全通信において 1 つのチャネルに重なる経路数の最大値を抑えるように VLAN ルーティングを設計する．

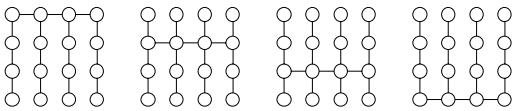


図 2 4 × 4 2 次元メッシュの DOR VLAN 集合
Fig. 2 The DOR VLANs in 4 × 4 2-D mesh.

3.2.2 VLAN ルーティングアルゴリズム

本項では，並列計算機などで広く採用されているトポロジ，すなわちメッシュ，トーラス，Fat Tree，Myrinet-Clos 網の各トポロジに提案手法を適用した場合の VLAN 割当て方法を示す．

3.2.2.1 トーラス，メッシュ

前節での議論に基づき，ここでは，トーラスおよびメッシュにおいて分散した経路集合を与える最短ルーティングとして，典型的なデッドロックフリー固定ルーティングである次元順ルーティング（DOR）[10] を用いる．DOR の経路に従う VLAN 割当て方法は文献 11) で明らかにされている．この手法では，ホストが VLAN タグ付けを行うことを前提にして，以下のように VLAN を割り当てる．

たとえば，$N \times N$ メッシュに対する DOR VLAN 集合は，$\left\{\mathrm{VL}(-, y) \,\middle|\, 0 \le y < N\right\}$ の N 個の VLAN で構成することができる（**図 2**）．図において，VLAN $\mathrm{VL}(-, y_0)$ は，全スイッチと y 方向への全リンク，y 座標 y_0 を持つスイッチ間リンクにより構成されるツリーを表す．スイッチ (x, y_S) に接続された送信元ホストは，すべての宛先に対して VLAN $\mathrm{VL}(-, y_S)$ を用いてフレームを送信する．一般に M 次元メッシュの場合，N を次元あたりのスイッチ数として，N^{M-1} 個の VLAN を必要とする．

この手法は，本論文の提案手法で利用可能なように容易に拡張できる．すなわち，スイッチ (x, y_S) において，ホストに接続されたポートすべてに VLAN $\mathrm{VL}(-, y_S)$ に対応する PVID を設定すればよい．

3.2.2.2 Fat Tree，Myrinet-Clos

Fat Tree および Myrinet-Clos 網（**図 3**）は，木構造を基本としたトポロジである．木はループ構造を含まないため，これらに対する最短ルーティングはすべてデッドロックフリーとなる．

これらのトポロジにおける VLAN 割当ては比較的単純である．たとえば，図 3 に示した Fat Tree $(2, 4, 2)$ では，**図 4** に示す 4 つの VLAN を用いることで，分散された経路集合を構築することができる．一般に，Fat Tree (u, d, m) において必要となる VLAN 数は u^m 個となる．ここで，u および d は各スイッチにお

ける上方向および下方向のリンク数，m は階層数である．

一方，図 3 に示した 4×4 の Myrinet-Clos 網では，**図 5** に示す 4 つの VLAN を用いることで，分散された経路集合を構築することができる．一般には，$n \times n$ の Myrinet-Clos 網において，n 個の VLAN が必要となる．

Fat Tree，Myrinet-Clos 網においては，あらゆる最短ルーティングでデッドロックが発生しない．このため，ホストに接続されたポートにどの PVID を割り当てるか，すなわち各ホストがどの VLAN を用いてパケットを送信するかについては，比較的選択の自由度が高い．特に，上記の Fat Tree $(2, 4, 2)$ の VLAN 割当て方法に対しては，どの VLAN を用いてもすべてのホスト間が最短ルーティングとなるため，経路の分散のみを考えて選択すればよい．本論文の評価（5 章）では，送信側のホスト ID を VLAN 数で割った余りによって使用する VLAN ID を選択する手法を用いている．

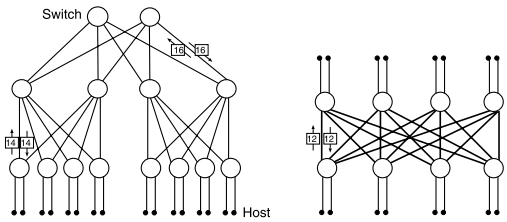


図 3 評価に用いた間接網
Fig. 3 Evaluated indirect networks.

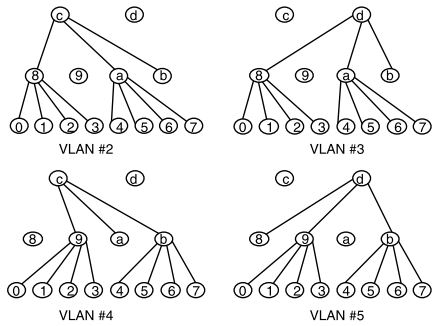


図 4 Fat Tree における VLAN 割当ての例
Fig. 4 An example of VLAN assignment for Fat Tree.

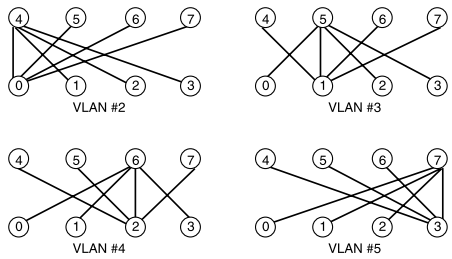


図 5 Myrinet-Clos 網における VLAN 割当ての例
Fig. 5 An example of VLAN assignment for Myrinet-Clos.

一方，Myrinet-Clos 網の場合，上記の割当て方法において最短ルーティングとなる VLAN ID の選択は一意に定まる．すなわち，図 5 において，スイッチ 0，4 に接続されたホストは VLAN #2 を用いてフレームを送信する．同様に，スイッチ 1，5，スイッチ 2，6，スイッチ 3，7 に接続されたホストはそれぞれ VLAN #3，#4，#5 を用いる．実際には，それぞれのホストが接続されたスイッチポートに対し，使用する VLAN ID を PVID として設定することになる．

3.2.2.3 各トポロジにおける VLAN ルーティング法の比較

各トポロジにおける前項で述べた性能要因，必要となる VLAN 数などを**表 1** に示す．

“#sw”，“#l”，“#VL” はそれぞれトポロジを構成するスイッチ数，リンク数，VLAN 数である．VLAN の割当ては，3.2.2.1 項および 3.2.2.2 項で述べた方法に従って行った．また，“AH” は経路の経由スイッチ数の平均値，“MH” はその最大値であり，“CP” 値は，各トポロジにおいて 1 つのチャネルに重なる経路数の最大値を表したものである．たとえば，各スイッチに 1 ホストのみを接続した 4 × 4 2 次元メッシュ上で次

表 1 さまざまなトポロジにおける VLAN ルーティングの諸元
Table 1 Evaluated topologies.

Topology	#sw	#l	#VL	AH	MH	CP
Mesh (4x2)	8	10	4	2.75	5	24
Torus (4x2)	8	12	4	2.50	3	32
Myri-Clos	8	16	4	2.25	3	12
M-Tree	16	15	(1)	4.81	10	64
Mesh (4x4)	16	24	4	3.50	7	16
Torus (4x4)	16	32	4	3.00	5	12
Fat-Tree	14	24	4	3.75	5	16

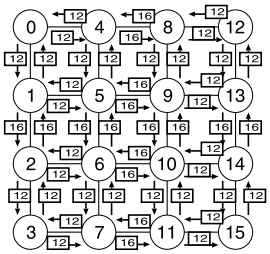


図 6 4 × 4 メッシュ上の各チャネルに重なる経路数
Fig. 6 The number of paths on a channel in the case of a mesh with DOR.

元順ルーティング（DOR）に従う経路を考えた場合，**図 6** に示すように，1 つのチャネルに最大で 16 の経路が重なることになる．

トポロジのうち，“M-Tree” は VLAN を用いない場合との比較のために導入した単純なツリー状トポロジであり，図 2 の VL$(-,0)$ と同じである．必要とするリンク数などの違いはあるが，これらのトポロジを，8 スイッチで構成されるものと 14 あるいは 16 スイッチで構成されるものとに分類することで，それぞれのグループ内でおおよそ公平に比較できるものと考えられる．

4. デッドロックフリールーティングの適用

本章では，VLAN ルーティング法を導入することで発生するデッドロックの問題について説明し，提案手法を効率的に適用するためにデッドロックフリールーティングが有効であることを示す．

4.1 Ethernet におけるデッドロック問題

並列計算機の相互結合網やシステムエリアネットワーク（SAN）では，メッシュなどのループ構造を含むトポロジが広く採用されている．VLAN ルーティング法を用いることで，Ethernet においてもループを含むトポロジを構築できるようになるが，同時に通常の Ethernet では考慮する必要のなかったデッドロックが問題となってくる．デッドロックが発生することで，ネットワークのスループットが劇的に低下する場合がある．

たとえば，**図 7** において，以下のフレーム転送が同時に発生すると，デッドロックが発生する．

- ホスト 0 が VLAN #2 を用いてホスト 3 へ
- ホスト 1 が VLAN #2 を用いてホスト 2 へ
- ホスト 2 が VLAN #3 を用いてホスト 1 へ
- ホスト 3 が VLAN #3 を用いてホスト 0 へ

一般に Ethernet では，スイッチやネットワークインタフェースのバッファが一杯になった場合，フレームは単に破棄される．通常，TCP などの上位層の End-to-End プロトコルが破棄フレームの再送処理を行うため，ネットワークのスループットは低下するものの，デッドロックは発生しない．ところが，SAN で行われているのと同様に，IEEE 802.3x リンクレベルフロー制御を用いた場合，フレームはほとんど破棄されなくなる．リンクレベルフロー制御は経路の衝突が起こった場合のバンド幅低下を抑制する効果があるが（次節で議論する），この場合，Ethernet においてもデッドロックが発生してしまう．

フレーム間のデッドロックを回避するためには，一般にデッドロックフリーの固定ルーティングアルゴリズムが有効である[9]．このようなアルゴリズムでは，デッドロックフリーを保証するために，チャネル依存グラフ（CDG）においてすべての循環依存を除去する，という操作を行う．

4.2 デッドロックの影響

VLAN ルーティング法におけるデッドロックの影響を評価するため，フレーム転送において，循環性の有無がバンド幅に与える影響を測定した．

デッドロックの問題が VLAN ルーティング法の実現方法によらない一般的な問題であることを示すために，ホストで VLAN ID を付与する従来の方法[5] と提案手法の両方を対象とした．

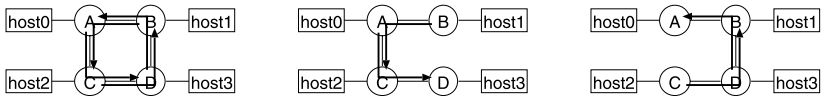
4.2.1 評価環境

表 2 に，評価に用いたクラスタの各ホストの仕様をまとめた．Ethernet スイッチには，DELL PowerConnect 5324（Gigabit Ethernet × 24 ポート，ノンブロッキング）を用いた．

測定プログラムには，各ホスト対間の TCP および UDP の転送バンド幅を測定する Tperf-1.4 [12] を用い，**図 8** に示す計 6 種類の転送パターンで測定を行った．また，送信プロセスと受信プロセスは同じスイッチに接続された異なるホストで起動した．

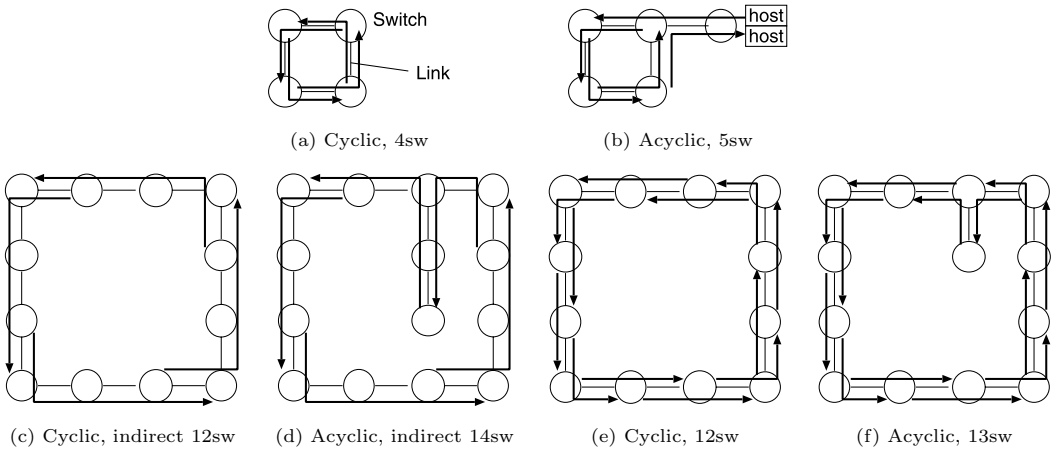
表 2 クラスタ内の各ホストの仕様
Table 2 Specifications of each host.

CPU	Intel Xeon 2.8 GHz × 2 (SMP)
Memory	PC2-3200 DDR2 SDRAM 1 Gbytes
Chipset	Intel E7520
PCI	64 bit/133 MHz PCI-X
NIC	Intel PRO/1000 MT Server Adapter
NIC Driver	Intel e1000 6.2.15
OS	Fedora Core 1 (kernel 2.4.21)



(a) 2 × 2 Mesh (b) VLAN #2 (c) VLAN #3

図 7 デッドロックを引き起こすフレーム転送
Fig. 7 Frame transfers with a deadlock.



(a) Cyclic, 4sw (b) Acyclic, 5sw

(c) Cyclic, indirect 12sw (d) Acyclic, indirect 14sw (e) Cyclic, 12sw (f) Acyclic, 13sw

図 8 循環および非循環のフレーム転送パターン
Fig. 8 Cyclic and acyclic frame transfers.

図 8 の (a)，(c)，(e) の各転送パターンでは，デッドロックを引き起こす経路間の循環構造が含まれているが，(b)，(d)，(f) のパターンでは循環が形成されていない．1 つのチャネルを使用する経路はどのパターンでも最大 2 であるため，循環性の有無以外に各パターン対間の条件に違いはない．

4.2.2 実験結果

ホストで VLAN ID を付与する従来の VLAN ルーティング法と提案手法の測定結果を**表 3**，**表 4** にそれぞれ示す．表において，バンド幅（単位 Mbps）の値は各経路ごとの測定結果の平均値であり，括弧内の数値はフレームの消失率である．また，リンクレベルフロー制御の影響を調べるため，4 種類のフロー制御の設定それぞれについて測定を行った．“FC None” はフロー制御を使用しない場合，“FC All” はすべてのリンクでフロー制御を使用した場合である．“FC Host” および “FC SW” はそれぞれ，ホスト–スイッチ間のリンクにのみフロー制御を用いた場合，スイッチ間リンクにのみフロー制御を用いた場合である．

表 3，表 4 から，循環を形成する転送パターン（(a)，(c)，(e)）では，ほとんどの場合において循環のないパターン（(b)，(d)，(f)）に比べてバンド幅が低下していることが分かる．特に，スイッチ間のリンクレベルフロー制御を使用している場合（“FC All” および “FC SW”）に，バンド幅が著しく低い値となっており，ほとんどゼロになっているものも多い．これは，いずれの VLAN ルーティング法においてもあてはまることから，VLAN ルーティング法を適用した場合の一般的な現象であるといえる．

この現象について解析するため，GtrcNET-1 [13] を用いて (a) のパターンと同様の転送実験を行った．GtrcNET-1 は，プログラマブルなネットワークテストベッドであり，Gigabit Ethernet のトラフィックをワイヤレートでモニタする機能を持つ．スイッチ間で転送されているフレームをキャプチャして解析したところ，フロー制御設定を “FC All” または “FC SW” にしたとき，循環を形成する各リンクにおいてフロー制御用の PAUSE フレームが大量に転送されており，通常のフレームがほとんど転送されていないことが判明した．PAUSE フレームは VLAN のトポロジとは無関係に転送されるため，これは PAUSE フレームの

表 3 各転送パターンにおけるバンド幅とフレーム消失率（従来の VLAN ルーティング法）
Table 3 Bandwidth and lost frame ratio with cyclic and acyclic transfers (Host-tagged).

	FC None	FC All	FC Host	FC SW
(a)/UDP	364.4	0.3	477.2	0.5
	(62%)	(28%)	(0.06%)	(99.9%)
(b)/UDP	394.3	317.0	505.8	230.3
	(59%)	(0.3%)	(7%)	(76%)
(c)/UDP	366.3	469.7	476.8	0.5
	(62%)	(0.04%)	(0.2%)	(99.9%)
(d)/UDP	437.6	475.4	507.4	247.7
	(54%)	(0.06%)	(7%)	(74%)
(e)/UDP	356.0	2.1	477.7	1.2
	(63%)	(9%)	(0.3%)	(99.9%)
(f)/UDP	361.3	204.9	494.3	106.4
	(62%)	(0.06%)	(4%)	(89%)
(a)/TCP	465.1	27.3	456.7	75.7
(b)/TCP	486.8	323.9	442.3	346.2
(c)/TCP	466.4	468.6	466.0	118.9
(d)/TCP	487.5	468.6	465.8	428.4
(e)/TCP	464.5	56.7	464.0	128.8
(f)/TCP	469.2	177.3	430.2	236.9

表 4 各転送パターンにおけるバンド幅とフレーム消失率（提案手法）
Table 4 Bandwidth and lost frame ratio with cyclic and acyclic transfers (Switch-tagged).

	FC None	FC All	FC Host	FC SW
(a)/UDP	345.2	209	477.2	0.3
	(64%)	(23%)	(0.07%)	(99.9%)
(b)/UDP	403.3	317.7	506.4	230.1
	(58%)	(0.02%)	(7%)	(76%)
(c)/UDP	339.0	464.9	477.2	0.3
	(65%)	(0.09%)	(0.1%)	(99.9%)
(d)/UDP	344.1	476.8	477.5	247.0
	(64%)	(0.02%)	(0.04%)	(74%)
(e)/UDP	349.1	1.9	479.4	19.2
	(64%)	(11%)	(1%)	(98%)
(f)/UDP	377.8	166.7	486.6	109.1
	(60%)	(0%)	(3%)	(89%)
(a)/TCP	445.0	0.8	454.6	90.4
(b)/TCP	465.4	317.6	440.9	345.0
(c)/TCP	462.2	469.3	466.4	401.4
(d)/TCP	443.2	464.1	469.7	427.9
(e)/TCP	465.6	47.8	447.0	153.9
(f)/TCP	472.4	158.3	416.6	227.5

循環によってデッドロックが発生し，他のフレームがまったく転送できなくなった状態であると考えられる．

さらに，表 3，表 4 から，スイッチ間のフロー制御を使用しない場合（“FC None” および “FC Host”）でも，非循環のパターンでは循環を含むパターンに比べて 2 割程度バンド幅が向上している．

これらの結果から，いずれかの VLAN ルーティング法を用いてループ構造を含むトポロジを構築する場合，デッドロックフリールーティングはリンクバンド幅を効率的に使うために非常に有効であることが分かる．特に，IEEE 802.3x リンクレベルフロー制御を使用する場合，デッドロックフリーであることが PAUSE フレーム間の循環を防ぐために必要となる．

5. 評　　価

本章では，提案手法におけるスイッチでのタグ付けがバンド幅，レイテンシに与える影響が小さいことを示す．さらに，従来 Ethernet ではほとんど採用されてこなかった，トーラスなどの並列計算機向けトポロジにおいて，提案手法の有効性を検証した結果を示す．

5.1 2 ホスト間の通信性能

評価には，4 章で用いたのと同じクラスタを用いた．クラスタには，SCore[3),14)] バージョン 5.8.2 が搭載されている．SCore はオープンソースの PC クラスタ向けシステムソフトウェアで，低レベル通信ライブラリ PM[4)] や MPICH-1.2.5[15)] をベースにした MPI ライブラリ MPICH-SCore を提供する．

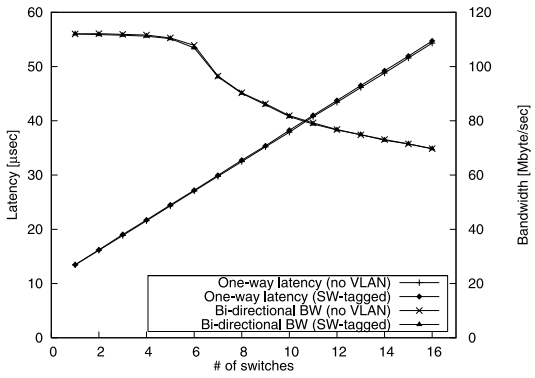


図 9 MPI 片道遅延と双方向バンド幅
Fig. 9 MPI one-way latency and bi-directional bandwidth.

まず，基礎評価として，Intel MPI Benchmarks（IMB）2.3[16)] を用いて，VLAN を用いない場合と，提案手法を用いてスイッチでタグ付けを行った場合の MPI レベルの遅延とバンド幅を測定した．**図 9** は，経由スイッチ数を変化させた際の，IMB PingPong テストによる MPI の片道遅延と，IMB PingPing テストによる双方向バンド幅の測定結果である．

図より，片道遅延は経由スイッチ数の増加にともないほぼ線形に増加していることが分かる．また，双方向バンド幅も経由スイッチ数が 6 を超えたあたりから大きく低下しており，遅延だけでなくバンド幅も経路のホップ数による影響を大きく受けることが分かる．このようにバンド幅が低下する理由としては，PM 通信ライブラリによる End-to-End の再送制御プロトコルによる影響が考えられる．宛先ホストからの Ack が返ってくるまでの時間が長い場合，Ack を待たずに一度に送信できるパケット数を使い切って待ち状態となっている時間が発生する．経由スイッチ数が多くなるほど Ack が返ってくるまでの時間が長くなるため，経由スイッチ数が 6 を超えるとこの状態が発生するようになるものと考えられる．このような再送制御は TCP などでも行われるため，信頼性を提供するプロトコルであれば同様の現象が起こるものと予想される．

一方，VLAN を用いない場合と，スイッチでタグ付けを行った場合の性能には差がないことから，スイッチでの VLAN タグ付けの処理オーバヘッドはほとんど性能に影響しないといえる．残念ながら，PM 通信ライブラリは現時点では VLAN タグ付けに対応していないため，MPICH-SCore を用いた本評価では，従来の VLAN ルーティング法を用いた場合との比較は不可能である．しかし，4 章において表 3 および表 4 に示した UDP および TCP のバンド幅測定結果より，

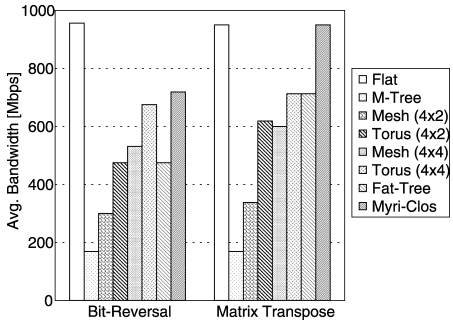


図 10 トラフィックパターンにおけるバンド幅
Fig. 10 Synthetic traffic pattern results.

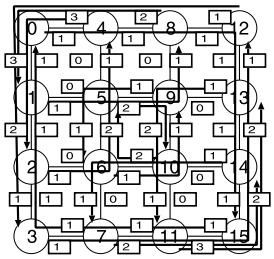


図 11 Bit-Reversal トラフィックにおける 4 × 4 メッシュ上の各チャネルに重なる経路数
Fig. 11 The number of paths on a channel in the case of a mesh with DOR in the case of bit-reversal traffic.

従来のホストでタグ付けを行う VLAN ルーティング法と，提案したスイッチでタグ付けを行う手法との間にはほとんど性能差はないといえる．

5.2 16 ホスト間の通信性能

次に，提案手法の評価として，表 1 にあげた各トポロジを構築し，通信性能を測定した．本論文で対象としているトーラスなどの並列計算機で採用されてきたトポロジは，これまで Ethernet ではあまり採用されてこなかった．そこで，これらのトポロジにおける提案手法の動作，有効性を確認した．

VLAN の割当て方法については，3 章に述べたとおりであり，必要となる VLAN 数は表 1 にまとめた．結合網の評価で用いられる典型的な通信パターンとして Bit-Reversal と Matrix Transpose [9] の 2 種類を使用し，ネットワーク全体のスループットの測定を行った．測定には Tperf-1.4 の UDP 転送を使用した．また，すべてのトポロジでホスト数は 16 とした．すなわち，Mesh（4×2）および Torus（4×2）ではスイッチあたり 2 台のホストを接続した．各ホストで送信プロセスと受信プロセスをそれぞれ起動し，UDP データグラムのサイズは最大の 1,470 byte とした．

図 10 は，各トポロジの測定結果において，すべての通信対のバンド幅の平均値をとったものである．比較のために 16 台のホストすべてを 1 つのスイッチに接続したフラットなトポロジ（Flat）でも測定を行った．このようなフラットトポロジはあくまで理想的なものであり，多数のホストを接続する大規模クラスタでは実現不可能なものであることに注意されたい．

図より，フラットトポロジと比較すると，ほとんどのトポロジでバンド幅が低下しているが，提案手法によるトポロジ（Mesh，Torus，Fat-Tree，Myrinet-Clos）は，単純なツリー状トポロジ（M-Tree）に比べて高いバンド幅を達成していることが分かる．特に，4 × 4 トーラスと Myrinet-Clos 網がどちらのトラフィックパターンの場合にも良い性能を示した．

これらのトポロジのバンド幅は，3.2.2.3 項で述べた経路の衝突により算出される値とほぼ一致する．たとえば，4 × 4 2 次元メッシュにおける Bit-reversal トラフィックでは，**図 11** に示すように，1 つの経路は最大で他の 2 つの経路とチャネルを共有することになる．すなわち，スイッチ 3 からスイッチ 12 への経路は，スイッチ 11 からスイッチ 15 へのチャネルを他の 2 つの経路と共有している．データグラムサイズが 1,470 byte の場合，UDP の理論転送性能は 958 Mbps であるため，この場合，1 つの経路あたりののバンド幅は 319 Mbps が上限となる．

このようにして概算したバンド幅は，たとえば 4×4 2 次元メッシュにおける Bit reversal トラフィックでは全経路の平均で 521 Mbps（測定値は 533.6 Mbps）となり，提案手法のスイッチでのタグ付けによるオーバヘッドはほとんどバンド幅に影響しないことが分かる．このことは，他のトポロジについても成立する．

なお，本測定では，経路の衝突がバンド幅に与える影響を明らかにするために UDP を用いて測定しているため，パケットのホップ数はほとんど性能に影響していない．しかし，TCP や，PM のようなクラスタ内通信向けのプロトコルでは，Ack などを用いて信頼性を提供する機構を備えているのが一般的である．前節で述べたように，このようなプロトコルでは，ホップ数が増加することによって遅延およびバンド幅が悪化する（図 9）ため，これらのプロトコルを用いて実際のアプリケーションを実行した場合は，ホップ数が性能に影響を与えると考えられる．

5.3 NAS 並列ベンチマーク性能

最後に，NAS 並列ベンチマーク 3.2 [17),18)] を用いて，提案手法によって実現される各トポロジにおいてアプリケーション実行性能の測定を行った．各ベンチマークの問題サイズはクラス B とし，実行プロセス数

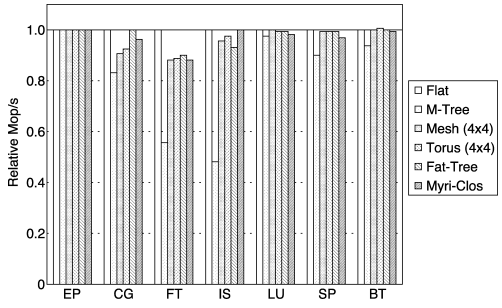


図 12 NAS 並列ベンチマーク性能
Fig. 12 NAS Parallel Benchmarks results.

はすべて 16 とした．コンパイルは gcc/g77 3.3.2 を用いてオプションを-O3 として行った．**図 12** は，各トポロジでのベンチマーク性能（Mop/s）を，比較のために用いるフラットトポロジの場合を 1 として正規化した相対性能を示したものである．

図より，提案手法によって構築した各トポロジは，ほとんどのアプリケーションにおいて理想的なフラットトポロジの 9 割以上の性能を達成している．前節でも述べたように，全ホストを 1 つのスイッチに接続するフラットトポロジはあくまで理想的なものであり，本評価環境のように比較的小規模なクラスタでは構築可能だが，数百・数千のホストを接続する大規模クラスタではほぼ実現不可能である．一方，提案手法では，8〜24 ポート程度の比較的小規模かつ低コストな Ethernet スイッチを用いて，3 章で述べたような，大規模な並列計算機や SAN で用いられてきたさまざまなトポロジを構築することが可能である．この点で，本手法は大規模なクラスタを構築する場合にも有効な手法であるといえる．

一方で，VLAN を用いない単純なツリー状トポロジである M-Tree は，EP を除くすべてのアプリケーションで提案手法を用いたトポロジに比べて性能が低くなっており，特に FT と IS では著しく性能が悪い．FT および IS は，MPI_Alltoall が頻繁に実行されるために高いバイセクションバンド幅を要求するアプリケーションであることが知られており，VLAN ルーティング法によるホスト間の経路の分散が非常に有効であることが分かる．

提案手法は，既存のホスト環境をいっさい変更することなく，低コストな Ethernet スイッチのみを用いて VLAN ルーティング法を利用できるようにする手法である．これらの評価結果より，提案手法を用いることで，VLAN ルーティング法による大規模かつ高性能な Ethernet クラスタを，低い導入コストで実現できる可能性が示されたといえる．

6. 関 連 研 究

三浦らの研究[6),7)] 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 アドレスを基にした 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．この点で柔軟性が高く，本研究とともに，VLAN ルーティング法を利用した PC クラスタの構築方法として有力であるいえる．また，両手法とも VLAN ID 制御のオーバヘッドが小さい点で優れている．

ただし，本研究と三浦らの研究とは，次の 2 つの点でその適用範囲が異なる．

- 本手法は，オペレーティングシステムやそのバージョンに依存しない．
- 本手法が適用可能な経路集合は，三浦らの手法に比べて限定される．

両者とも，変化が激しい HPC 分野を対象としているため，柔軟性は重要である．三浦らの手法も，Linux 用のデバイスドライバであるため高い柔軟性を持っている．しかし，本手法はさらに，たとえば最新バージョンのカーネルや，Microsoft の Windows Compute Cluster Server 2003 を用いるような PC クラスタにも，ホストの動作検証なしにそのまま適用することができる．この点で，提案手法は利用する際の柔軟性が高く，管理が簡単であるといえる．一方，提案手法は 2.3 節で述べたとおり，1 つのホストからの経路はすべて 1 つの VLAN に属していなければならず，適用可能な経路集合は限定される．さらに，三浦らの手法のようにスイッチにおける MAC アドレス学習は利用できず，静的に MAC アドレスを登録しなければならない．ただし，3 章で述べたとおり，提案手法は，規則性を持つ多くのトポロジにおいて典型的な最短ルーティングを実装することができる．

このほか，朴らの研究[19),20)] では，VLAN ルーティング法とは異なるが，大規模クラスタシステム用のネットワークとして Ethernet による 3 次元ハイパークロスバ網の実現を目指し，そのための専用通信ライブラリ PM/Ethernet-HXB を新たに開発している．VLAN ルーティング法によるハイパークロスバ網[5)] では節点において各次元方向のスイッチを直接接続するが，この方式では，ホスト上の PM ドライバがパケッ

トのルーティングを担当する点が異なる．

7. ま と め

本論文では，スイッチにおいて VLAN タグ付けを行うことで，ホスト側のシステムソフトウェアが VLAN をサポートしていない場合でも VLAN ルーティング法を利用できるようにする手法を提案した．また，VLAN ルーティング法によってサポートされるトポロジにおいて，デッドロックフリールーティングが有効であることを示し，特に IEEE 802.3x リンクレベルフロー制御を使用した場合にデッドロックフリーが必要であることを明らかにした．

提案手法を用いることで，ホスト上の既存のソフトウェアスタックに対する変更をいっさい行うことなく，低コストな Ethernet スイッチのみを用いて VLAN ルーティング法を利用でき，並列処理に適したさまざまなトポロジが構築可能になる．ベンチマークを用いた評価の結果，提案手法が設計どおり動作することを確認し，アプリケーションの実行において理想的なフラットトポロジに近い性能を示すことが分かった．これらのことから，提案手法を用いることで，VLAN ルーティング法による大規模かつ高性能な Ethernet クラスタを，低い導入コストで実現できる可能性を示した．

謝辞 本研究の一部は，国立情報学研究所共同研究「イーサネットを用いた並列分散処理に関する研究」による．

参 考 文 献

1) Boden, N.J., Cohen, D., Felderman, R.E., Kulawik, A.E., Seitz, C.L., Seizovic, J.N. and Su, W.: Myrinet: A Gigabit-per-Second Local Area Network, *IEEE Micro*, Vol.15, No.1, pp.29–36 (1995).
2) Association, I.T.: InfiniBand architecture, Specification Volume 1, Release 1.0.a, available from the InfiniBand Trade Association, http://www.infinibandta.com (2001).
3) 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, *Proc. 5th Annual Linux Expo*, pp.55–62 (1999).
4) Takahashi, T., Sumimoto, S., Hori, A., Harada, H. and Ishikawa, Y.: PM2: High Performance Communication Middleware for Heterogeneous Network Environment, *Proc. SC2000 High Performance Networking and Computing Conference*, pp.52–53 (2000).
5) 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).
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) Miura, S., Okamoto, T., Boku, T., Sato, M. and Takahashi, D.: Low-cost High-bandwidth Tree Network for PC Clusters based on Tagged-VLAN Technology, *Proc. 8th International Symposium on Parallel Architectures, Algorithms and Networks* (*I-SPAN 2005*), pp.84–93 (2005).
8) Koibuchi, M., Watanabe, K., Otsuka, T. and Amano, H.: Performance Evaluation of Deterministic Routings, Multicasts, and Topologies on RHiNET-2 Cluster, *IEEE Trans. Parallel and Distributed Systems*, Vol.16, No.8, pp.747–759 (2005).
9) Duato, J., Yalamanchili, S. and Ni, L.: *Interconnection Networks: an engineering approach*, Morgan Kaufmann (2002).
10) Dally, W.J. and Seitz, C.L.: Deadlock-Free Message Routing in Multiprocessor Interconnection Networks, *IEEE Trans. Comput.*, Vol.36, No.5, pp.547–553 (1987).
11) Otsuka, T., Koibuchi, M., Jouraku, A. and Amano, H.: VLAN-based Minimal Paths in PC Cluster with Ethernet on Mesh and Torus, *Proc. 2005 International Conference on Parallel Processing* (*ICPP-05*), pp.567–576 (2005).
12) Tperf. http://www.am.ics.keio.ac.jp/~terry/tperf/
13) Kodama, Y., Kudoh, T., Takano, R., Sato, H., Tatebe, O. and Sekiguchi, S.: GNET-1: Gigabit Ethernet Network Testbed, *Proc. 2004 IEEE International Conference on Cluster Computing* (*Cluster2004*) (2004).
14) PC Cluster Consortium. http://www.pccluster.org/
15) MPICH. http://www-unix.mcs.anl.gov/mpi/mpich/
16) Intel Cluster Toolkit. http://www.intel.com/cd/software/products/asmo-na/eng/cluster/clustertoolkit/
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) Saphir, W., Wijngaart, R., Woo, A. and

Yarrow, M.: New Implementations and Results for the NAS Parallel Benchmarks 2, *8th SIAM Conference on Parallel Processing for Scientific Computing* (1997).
19) 朴　泰祐，佐藤三久，宇川　彰：計算科学のための超並列クラスタ PACS-CS の概要，情報処理学会研究報告 2005-HPC-103，pp.133–138 (2005).
20) 住元真司，久門耕一，朴　泰祐，佐藤三久，宇川　彰：PACS-CS のための Ethernet を用いた高性能通信機構の設計，情報処理学会研究報告 2005-HPC-103，pp.139–144 (2005).

(平成 18 年 1 月 27 日受付)
(平成 18 年 6 月 23 日採録)

大塚　智宏（学生会員）
2003 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．2006 年度より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手．計算機アーキテクチャ，並列分散処理に関する研究に従事．

鯉渕　道紘（正会員）
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）．平成 14 年度から平成 16 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手．マルチプロセッサシステムの結合網に関する研究に従事．

工藤　知宏（正会員）
1991 年慶應義塾大学大学院理工学研究科修了．博士（工学）．東京工科大学情報工学科助手/講師/助教授を経て，1997〜2002 年技術研究組合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室長．光インターコネクションを用いた計算機間ネットワークの開発プロジェクトを指揮．2002 年 4 月より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クラスタ技術チーム長．クラスタ計算機およびグリッドのためのネットワークに関する研究に従事．

天野　英晴（正会員）
1986 年慶應義塾大学大学院理工学研究科電気工学専攻博士課程修了．博士（工学）．現在，慶應義塾大学理工学部情報工学科教授．計算機アーキテクチャの研究に従事．