LC-011

イーサネットを用いた大規模クラスタネットワーク構築法

鯉渕道紘[†]　大塚智宏[††]
工藤知宏[†††]　天野英晴[††]

A Methodology for Building Large-Scale Cluster Interconnects using Ethernet

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

1. 概　要

本稿では，イーサネットを用いて大規模なクラスタネットワークを構築するために VLAN リネーミングを提案する．VLAN リネーミングは，スイッチの既存の機能を制御することにより実現でき，1) 必要となる VLAN 数がスイッチの次数以内と少数，かつ 2) システムソフトウェアが VLAN 技術に対応していない場合にも様々な (デッドロックフリー) ルーティングを利用可能，という 2 点の特徴を持つ．32 台の PC を用いたクラスタシステムによる評価結果より，VLAN リネーミングは導入によるオーバヘッドがほとんどなく，ネットワーク資源を効率良く使った大規模クラスタネットワークの構築に適しているといえる．

2. はじめに

イーサネットは，管理の容易さ，高い耐故障性，安価なハードウェアコストなどの利点から，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，PC クラスタのネットワークとして幅広く採用されている．特に，GbE の普及，リンク集約化，ツイストペアケーブルを用いる 10GbE-T(802.3an-2006) の標準化などにより，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 SAN に迫る高性能 PC クラスタネットワークとして注目を集めている．

現状では，イーサネットはグリッド，HPC 分野においても，スパニングツリープロトコル (STP)，TCP/IP などの通信プロトコルスタックにより運用されることが多い．しかし，これらは本来 HPC 分野の技術ではないため，ネットワーク資源を効率的に利用することが難しい場合が多い．そのため，HPC 分野で用いる場合にはさらなる改良が必要となる．

そこで，リンク集約化 (IEEE203.3ad) 以外にも，STP を用いずに，同一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する方法が提案されてきた[1)]．これらは，リンク集約化と異なり，トーラスなどのループを含むトポロジを取ることができる特徴を持つ．STP を用いずに大規模クラスタシステムを構築する場合，ホストの追加，スイッチの故障，操作ミス等によるブロードキャストストームの発生を抑えるために MAC アドレスの管理が 1 つの課題となる．この点において，既存の方法の中で IEEE 802.1Q 標準のタグ VLAN 技術を応用する VLAN ルーティング法[1)] が有力である．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が，VLAN ルーティング法ではネットワークのスループット向上のために用いる．VLAN ルーティング法は図 1 のように，1 つのホストが複数の VLAN グループのメンバーになるようにしておき，各 VLAN ツリー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，MAC アドレスのブロードキャストストームを避けつつ，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．

† 国立情報学研究所/総合研究大学院大学
National Institute of Informatics/Sokendai
†† 慶應義塾大学大学院 理工学研究科
Graduate School of Science and Technology, Keio University
††† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

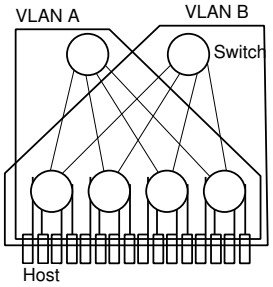

図 1 VLAN ルーティング法
Fig. 1 VLAN Routing Method

しかし，現状の VLAN 手法を用いて大規模クラスタネットワークを構築する場合，次の 2 点の問題が生じる．

- 多数の VLAN が必要となる．
- システムソフトウェアが VLAN 技術に対応していない場合[2)]，ルーティングアルゴリズムが限定される．

1 つ目の点については，安価な商用のレイヤ 2 ノンブロッキングスイッチは，数十個〜256 個程度の VLAN のサポートに留まるものが多い．しかし，既存の VLAN ルーティング法において必要となる VLAN 数は，ネットワークサイズに応じて増える．そのため VLAN 数が大規模クラスタネットワークを構成するための制限要因となる．

2 つ目の点は，並列計算機のネットワークと同様に，ループを含むイーサネットトポロジにおいても有効性が確認されているデッドロックフリールーティングを採用できない場合が生じ，その場合フレームの廃棄が頻繁に発生する[2)]．

そこで，本研究では，これら 2 点の問題を解決するために，各スイッチ内において，フレームを異なる VLAN ID に乗せかえながら転送する VLAN リネーミングを提案する．各スイッチが，そのスイッチ内の転送のみに有効な VLAN タグをフレームに割り当てることにより VLAN リネーミングはスイッチの次数 (使用ポート数) 以内の少数の VLAN 数で任意のサイズ，トポロジを構築することができるようになる．さらに，VLAN リネーミングは，ホストのシステムソフトウェアが VLAN 技術に対応していない場合にも使用でき，様々な (デッドロックフリー) ルーティングアルゴリズムを使うことができるようになる．VLAN リネーミングは多くの安価な商用 L2 イーサネットスイッチにおいてサポートされている機能を制御することにより実現できる点で高い実用性を持つ技術といえる．

3. VLAN リネーミング

VLAN リネーミングは，多くの商用スイッチが提供している VLAN の機能を次のように利用する．

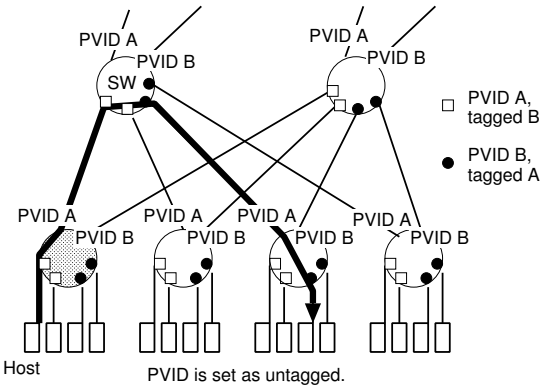

図 2 Fat ツリーにおける VLAN リネーミング
Fig. 2 VLAN Renaming on Fat Tree

- スイッチのすべての出力ポートにおいて，フレームの VLAN タグをすべて除去 (untagged) する．
- スイッチの入力ポートには常に VLAN タグなしフレームのみが到着し，それらは，PVID(ポートに対して与えた VLAN ID) 値にもとづいて，出力ポートに転送される．

そして，VLAN リネーミングはスイッチの各ポートの PVID を次の手順により割り当てることで実現する．

(1) 各入力ポート i に対して PVID p_i を割り当てる (図 3.(a))．
(2) 入力ポート i を通過するフレームが転送されうる出力ポートには p_i の通過を tagged として設定する (図 3.(b))．
(3) Step (2) において挙動が重複している PVID 群を 1 つにする (図 3.(c))．

Fat ツリートポロジに VLAN リネーミングを適用した例を図 2 に示す．この図においてスイッチの左側 2 ホストからのフレームは左側の上位スイッチ，右側 2 ホストからのフレームは右側の上位スイッチを用いて転送される．

ルーティングアルゴリズムにチャネル循環依存が存在しない (すなわちデッドロックフリー) 限り，VLAN リネーミングにおいてブロードキャスタトストームは発生しない．VLAN リネ ーミングにより，フレームは各スイッチ内において，異なる VLAN ID にのせかえられながら転送されることになる．また，VLAN リネーミングは，VLAN 毎に MAC アドレスをスイッチに静的に登録することにより様々な経路を実装することができる．

4. 評　　価

本章では，VLAN リネーミングの評価，ならびに，VLAN を用いずにスケーラブルな大規模クラスタネットワークを構築する既存の方法として，リンク集約化

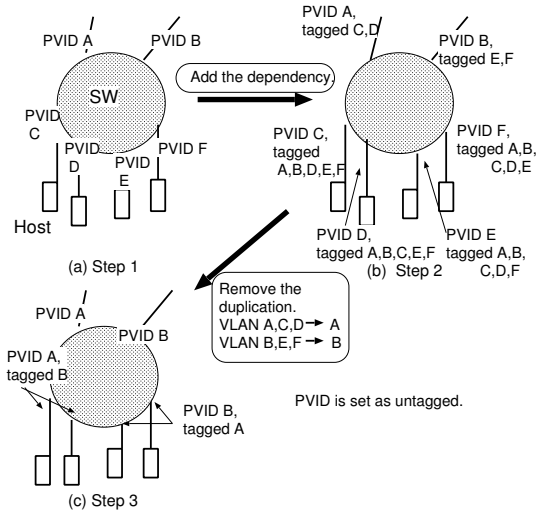

図 3 図 2 における VLAN 割り当て例
Fig. 3 VLAN Assignment on Figure 2

(LAG) との比較を行う．

4.1 VLAN リネーミングのオーバヘッド

スイッチにおける VLAN リネーミングのオーバヘッドを表 1 に示す．2 ホスト間の ping(ICMP メッセージ) を用いて，ノンブロッキング L2 スイッチであるデル PowerConnect5324, アライドテレシス GS916M，ネットギア GSM7212 におけるフレーム通過時間を GtrcNET-1[3] で各 300 回測定した．表 1 において，U-U は VLAN タグを用いない場合，T-T はホストにおいて VLAN タグを付加した場合，RENAME はホストにおいて VLAN タグを付加せず，スイッチ内のみ PVID の VLAN 処理を行うリネーミングの場合をそれぞれ示す．表 1 より，VLAN リネーミングの導入による遅延はほとんどないことが分かった．なお，ホストでは VLAN 処理を一切行わないため，ホストにおける VLAN リネーミング導入のオーバヘッドはない．

表 1 スイッチの通過遅延 (μsec)

	Min	Ave	Max
U-U (PC5324)	2.47	2.74	2.79
T-T (PC5324)	2.47	2.76	2.79
RENAME(PC5324)	2.47	2.75	2.79
U-U (GS916M)	2.44	3.30	3.72
T-T (GS916M)	2.40	3.29	3.71
RENAME (GS916M)	2.40	3.14	3.56
U-U (GSM7212)	2.47	2.77	2.79
T-T (GSM7212)	2.47	2.76	2.79
RENAME (GSM7212)	2.43	2.73	2.75

次に Tperf 1.5[4] を用いたバンド幅測定結果を表 2 に示す．VLAN リネーミングでは，イーサネットのフレームはリンク上において VLAN タグを含まない．つまり，従来の VLAN ルーティング法と異なり，VLAN リネーミングは VLAN 処理をスイッチ内で閉じることにより，VLAN によるバンド幅の低下が生じていないことが分かる．

表 2 バンド幅 (Mbps)

	PC5324	GS916M	GSM7212
U-U (UDP)	957.0	956.8	957.1
T-T (UDP)	954.4	954.5	954.1
RENAME(UDP)	957.0	956.8	957.1
U-U (TCP)	941.1	940.6	941.0
T-T (TCP)	936.9	938.0	938.0
RENAME(TCP)	941.1	941.0	941.0

4.2 VLAN 数

並列分散システムで採用されている典型的なトポロジにおいて，ホストで VLAN 割り当てを行う従来の場合と VLAN リネーミングで必要となる VLAN 数の比較を表 3 に示す．Fat ツリー (u,d,r) は，スイッチの上位リンク数，下位リンク数，レイヤ数を示す．また，メッシュ，トーラスでは経路を分散させることができる次元順ルーティングを想定した．トーラスは次元順ルーティングにおいてラップアラウンドチャネルによる循環依存を除くため，同一スイッチ間に 2 本のリンクを用いている．この表より，既存のホストにおいてフレームに VLAN を割り当てる場合と異なり，VLAN リネーミングはスイッチの次数 (使用ポート数) 以内というごく少数の VLAN 数でシステムを構築できていることが分かる．

表 3 VLAN 数の比較

	ホスト VLAN 割り当て	RENAME
Fat ツリー (u,d,r)	u^r	u
メッシュ(k-ary n-cube)	k^{n-1}	n
トーラス (k-ary n-cube)	$2k^{n-1}$	$3n$

4.3 PC クラスタを用いた評価

最後に表 4，図 4 に示した 32 台のホストにより構成されるシステムを用いて VLAN リネーミングを実装した．通信に TCP/IP を用いた場合，ならびに SCore システムソフトウェアを用いた場合における NAS 並列ベンチマーク (NPB)3.2[5] の評価結果を図 5 に示す．NPB は，すべてプロセス数 32，クラス B とし，32 プロセスで動作させることができるアプリケーションを選択した．コンパイルは gcc/g77 3.4.6 を用いてオプションは -O3 として行った．図 5 において縦軸はリンク集約化を用いた場合の実行時間に対する VLAN リネーミングを用いた場合の相対的な実行時間を示し，小さいほど性能が高いことを表す．VLAN リネーミングはトポロジをツリーに限定する必要はないが，本評価ではリンク集約化との比較のため，ツリートポロジを用いた．VLAN リネーミングにおいて，4 本のスイッチ間リンクは，4 個の VLAN タグにおいて識別され，フレームの出発地毎にリンクを

表 4 ホストの仕様

CPU	Intel Xeon 2.8GHz × 2 (SMP)
Memory	PC2-3200 DDR2 SDRAM 1Gbytes
Chipset	Intel E7520
PCI	64bit/133MHz PCI-X
NIC	Intel PRO/1000 MT Server Adapter
NIC Driver	Intel e1000 7.0.33
OS	Linux Kernel 2.6.9
MPI	MPICH-1.2.5
SCore	6.0.2.1

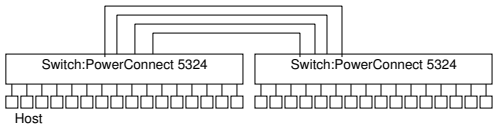

図 4 PC クラスタの構成
Fig. 4 Composition of the PC Cluster

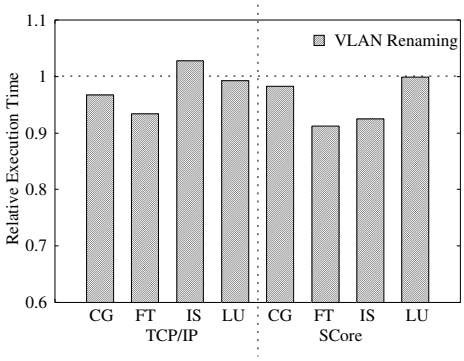

図 5 NPB の実行結果
Fig. 5 Execution Results of NPB

均等に使い分けるようにした．

図 5 より，VLAN リネーミングはリンク集約化を用いた場合に比べて実行時間が最大 9%向上した．この性能差は，表 5 に示した通り，VLAN リネーミングが 4 本のスイッチ間リンクの中でトラフィックをほぼ均一化していた一方，PowerConnect 5324 で実装されているリンク集約化アルゴリズムがスイッチ間リンクのトラフィックを均一に分散できなかったことによる．その他の要因としては，表 6 に示したようにリンク集約化によりスイッチの通過遅延が 2%増加する点が挙げられる．

表 5 4 本のスイッチ間リンクの通過トラフィック量の偏り (1 リンクのフレームの Octet 量)

(application)	Min ($\times 10^6$)	Max ($\times 10^6$)
LAG (IS)	73	111
RENAM(IS)	99	100
LAG (FT)	582	891
RENAM(FT)	786	788
LAG (CG)	159	478
RENAM(CG)	317	318
LAG (LU)	43	136
RENAM(LU)	86	90

表 6 2 スイッチ間の遅延 (μsec)

	Min	Ave	Max
LAG	5.54	5.58	5.60
LAG なし	4.90	5.46	5.51

5. ま と め

本稿では，イーサネットを用いて大規模クラスタを構築するために VLAN リネーミングを提案した．VLAN リネーミングは 1) 必要となる VLAN 数がスイッチの次数以内と少数，2) ホストのシステムソフトウェア，ドライバが VLAN 技術に対応している必要がなく，さらに様々な (デッドロックフリー) ルーティングアルゴリズムを利用可能，という利点を持つ．また，VLAN リネーミングは多くの安価な商用イーサネットスイッチにおいてサポートされている機能を制御することにより実現できる点で高い実用性を持つ技術といえる．

評価結果より，VLAN リネーミングのスイッチ通過遅延，バンド幅のオーバヘッドはほとんどない．また，既存のバンド幅向上技術であるリンク集約化は，スイッチが採用しているアルゴリズムによりクラスタトラフィックが分散されない場合があるが，VLAN リネーミングでは，明示的にリンクのトラフィック分散を設定することができる．スパニングツリープトロコル (STP) とリンク集約化を用いた場合と，VLAN リネーミングを含む VLAN 手法を利用した場合の他の特徴を表 7 に示す．

表 7 大規模クラスタの構築方法

	従来の方法 (LAG)	VLAN 使用
トポロジ	ツリー (STP)	任意
経路	ツリールーティング	デッドロックフリー ルーティング
経路数	1	複数
VLAN 数	1(無し)	次数個以内

以上より，VLAN リネーミングは，リンク集約化と比べてトポロジの柔軟性，スイッチの通過遅延等の性能の点で優れており，ネットワーク資源を効率良く使ったスケーラブルな大規模クラスタネットワークの構築に適しているといえる．

参 考 文 献

1) 工藤, 松田, 手塚, 児玉, 建部, 関口: 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(ACS 6), pp. 35–43 (2004).
2) 大塚, 鯉渕, 上樂, 工藤, 天野: 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(ACS 15), pp. 46–58 (2006).
3) GtrcNET-1: http://projects.gtrc.aist.go.jp/gnet/.
4) Tperf: http://www.am.ics.keio.ac.jp/~terry/tperf/.
5) NAS Par. B.: http://www.nas.nasa.gov/Software/NPB/.