

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 淵 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ー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)

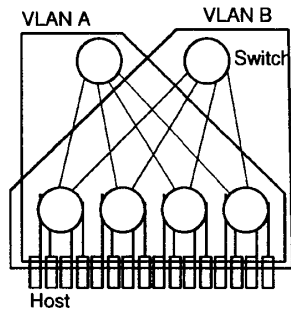


図1 VLAN ルーティング法
Fig. 1 VLAN Routing Method

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に乘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。

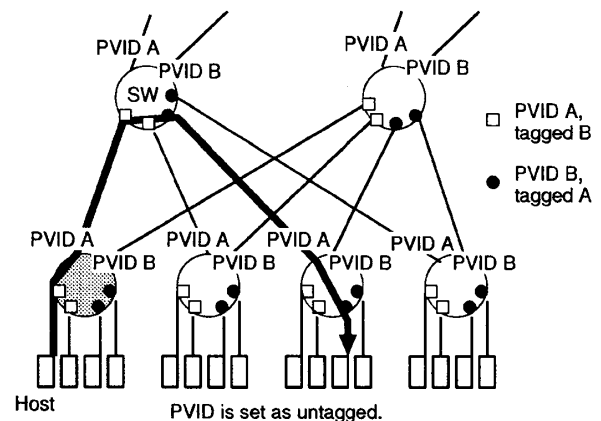


図2 Fat ツリーにおける VLAN リネーミング
Fig. 2 VLAN Renaming on Fat Tree

- 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ン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

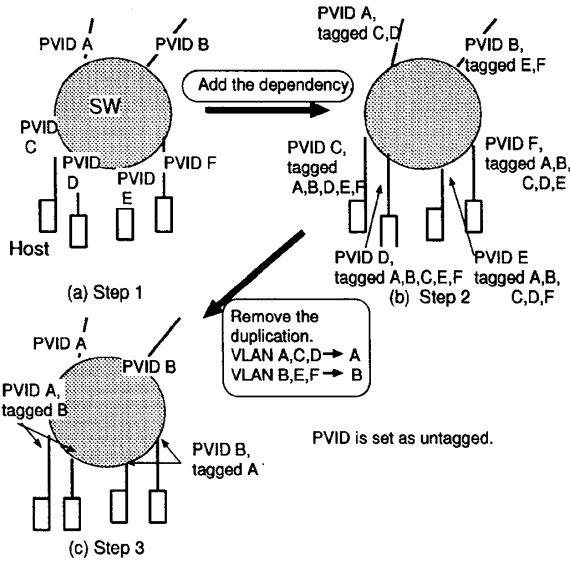


図3 図2におけるVLAN割り当て例
Fig. 3 VLAN Assignment on Figure 2

(LAG) との比較を行う。

4.1 VLAN リネーミングのオーバーヘッド

スイッチにおけるVLANリネーミングのオーバーヘッドを表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-1³⁾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ーヘッドはない。

表1 スwitchの通過遅延 (usec)

	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 1.5⁴⁾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

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

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。

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ン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)3.2⁵⁾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

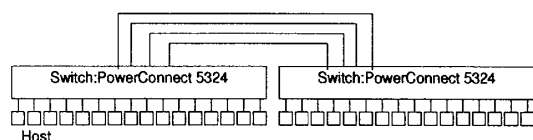


図4 PCクラスタの構成

Fig. 4 Composition of the PC Cluster

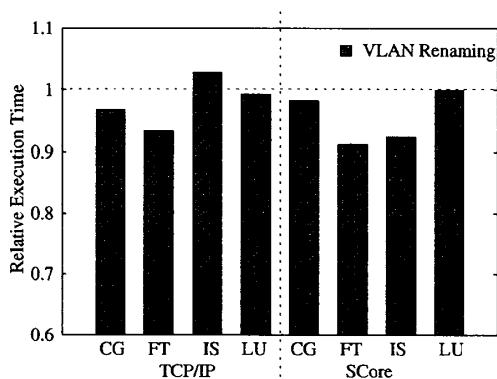


図5 NPBの実行結果

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、バンド幅のオーバー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。スパンニングツリープロトコル(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, 鯉淵, 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/>.