マルチパスイーサネットにおける 省電力 On/Off リンクアクティベーション法

鯉渕道紘[†] 大塚智宏[††]
松谷宏紀[††] 天野英晴[††,†]

近年，PC クラスタでは，性能向上とともに消費電力の削減が強く求められている．本稿では PC クラスタのインターコネクトの主流であるイーサネットにおいてスイッチの消費電力を削減するために，トラフィック負荷に応じてリンクを On/Off するアクティベーション法を提案する．On/Off リンクアクティベーション法は，VLAN ルーティング法を応用することで任意のトポロジにおいてブロードキャストストームを避けつつ，ホスト間の経路を更新する．On/Off リンクアクティベーション法は，スイッチの既存の機能を制御することにより実現でき，ホスト側の設定は不要である．リンクの On/Off 操作，経路更新の総オーバヘッドは商用スイッチでは多くの場合，数秒であった．NAS パラレルベンチマークを用いた PC クラスタの評価結果より，On/Off アクティベーション法は性能を維持しつつ，スイッチの総消費電力を最大 37%削減できることがわかった．

An On/Off Link Activation Method for Power Saving in Multipath Ethernet

Michihiro Koibuchi,[†] Tomohiro Otsuka,[††] Hiroki Matsutani[††]
and Hideharu Amano[††,†]

Power saving is required for modern PC clusters as well as the performance improvement. To reduce the power consumption of Ethernet switches used in the clusters, this paper proposes an on/off link method that activates and deactivates the links based on the traffic information. The proposed method can renew a path without creating broadcast storms on various topologies that include loops using the VLAN routing method, and the proposed method has advantages in both simple host configuration and high portability. Evaluation results using NAS Parallel Benchmarks show that the proposed method reduces the power consumption of switches by up to 37% without performance degradation on a PC cluster.

1. はじめに

イーサネット (Ethernet) は，管理の容易さ，高い耐故障性，安価なハードウェアなどの利点から，ローカルエリアネットワーク (LAN) のみならず，広域ネットワークや PC クラスタのインターコネクトとしても幅広く採用されている．特に，Gigabit Ethernet (GbE) の普及，ツイストペアケーブルを用いる 10GBase-T の標準化 (IEEE 802.3an-2006) などにより，イーサネットは PC クラスタにおいて，Myrinet などの高価なシステムエリアネットワーク (SAN) に迫るインターコネクトとして主流になりつつある．例えば，TOP500 スーパーコンピュータのランキング[1] において上位 500 台の中で GbE を用いたクラスタシステムが 57%と過半数を占めている．

PC クラスタでは，ホスト内のプロセッサ等の各要素の省電力化は進んでいる．しかし，インターコネクトはリンクバンド幅が向上するにつれて，消費電力が増えるため，インターコネクトの省電力技術の研究開発は，今後の PC クラスタ構築において極めて重要となる．

イーサネットのスイッチの消費電力は，使用しているポート数に大きく依存し，ポート使用の有無は，スイッチのポート・シャットダウン操作により制御できる．また，フレーム転送の有無に関わらず，ポートが隣接デバイスと通信可能な場合 (On)，ほぼそのポートの最大電力を消費する．ポート・シャットダウンは，他の隣接デバイスとの通信を遮断させるために，多くのイーサネットスイッチに備わった制御機能であり，表 1 のように，スイッチの消費電力を削減することができる (測定方法は 5 章において述べる)．

表 1 GbE スイッチの消費電力 (W)

	ポート以外	ポート	合計 (ポートの割合)
PC5324	14.9	1.2	42.9(65%)
PC6248	56.8	2.1	155.2(63%)
C3750	84.5	1.8	127.7(34%)

表 1 において C3750 はシスコシステム WS-C3750G-24-TS-S(24 ポート) を表し，PC5324(24 ポート)，PC6248(48 ポート) はデル PowerConnect スイッチシリーズの型番を表す．“ポート以外” はすべてのポートをシャットダウンして未使用の場合 (Off)，“合計” はすべてのポートが On の場合，“ポート” はシャットダウン操作により削減されるポート当たりの消費電力を表す．本稿ではこのシャットダウン操作により削減される消費電力を “ポートの消費電力” として扱う．現在，GbE スイッチではポートあたり高々 2W 程度の消費電力であるが，今後の PC クラスタのインターコネクトの候補である 10GBase-T のリンクではその数倍の消費電力となる．

本稿では，イーサネットにおいてスイッチの消費電力を削減

† 国立情報学研究所/総合研究大学院大学/JST
National Institute of Informatics/SOKENDAI/JST
†† 慶應義塾大学大学院 理工学研究科
Graduate School of Science and Technology, Keio University

するために，ポートシャットダウン機能を利用する On/Off リンクアクティベーション法を提案する．提案手法は，高負荷時に全リンクのポートをアクティベーションする一方，低負荷時に使用率の低いリンクのポートをシャットダウンさせる．リンクが Off から On あるいは On から Off に遷移した場合，そのリンクを利用，あるいは以後使用しないために経路を更新する必要がある．そのため，頻繁に生じるスイッチの MAC アドレステーブル管理を高速かつ安定的に行う必要がある．On/Off リンクアクティベーション法ではスイッチで VLAN タグ付けを行う VLAN ルーティング法を応用することで，任意のトポロジにおいて MAC アドレステーブルの更新を安定的に実現する．

以下，2 章において関連研究を紹介し，3 章において On/Off リンクアクティベーション法を提案し， 4 章において On/Off リンク選択アルゴリズムを提案する．そして，5 章にて評価結果を示し，最後に 6 節でまとめを述べる．

2. 関 連 研 究

リンク制御に時間のかかるクラスタインターコネクトなどのチップ間通信と wakeup, sleep 状態への遷移が高速なチップ内ネットワークの各々において，トラフィック負荷に応じてリンクを On/Off する手法が提案されている[2]．しかし，イーサネットは (1) トポロジがツリー，(2) 通常，スイッチの MAC アドレステーブルが学習により設定される，という 2 つの大きな特徴があるため，この手法をイーサネットにそのまま実装することが難しい．

PC クラスタなどの閉じたネットワークにおいて，学習ではなく，静的にスイッチの MAC アドレステーブルを設定することでループ構造を含む任意のトポロジを採用する方法，任意のトポロジにおいて使用することができるルーティングアルゴリズムについても提案されている[3]．しかし，安定して経路を更新するためには一旦全スイッチポートをシャットダウンし，MAC アドレステーブルを更新する必要がある．しかし，この更新法では通信の大きな遮断時間が生じてしまう．さらに，この手法はそもそもブロードキャストストームが生じた場合の対処が難しい．

スイッチにおける MAC アドレスの管理の点で，IEEE 802.1Q 標準のタグ VLAN 技術を応用した VLAN ルーティング法[4][5] が経路の実装方法として有力である．さらに，三浦らにより，経路の動的な切り替えを行うために VLAN ルーティング法を利用する手法が提案されている[6][7]．この方法では，ホストにおいて更新した経路に対応した VLAN を選択することで，全スイッチポートを一旦シャットダウンすることなく経路を変更することができる．そのため，耐故障性の向上とアプリケーションにあわせた経路の設定を実現している．しかし，VLAN の設定を各ホストで行う必要があり，さらに，ドライバあるいは通信ライブラリがタグ VLAN に対応している必要がある．IEEE 802.1s MSTP (Multi STP) やシスコシステム PVST (Per VLAN Spanning Tree) を使うことも可能であるが，安価なスイッチではサポートされていないことが多い．

このように，イーサネットおけるトポロジ，ルーティングとその実装に関する研究は行われているが，商用のイーサネットスイッチを用いた省電力技術に関する研究は，我々の知る限りほとんどない．

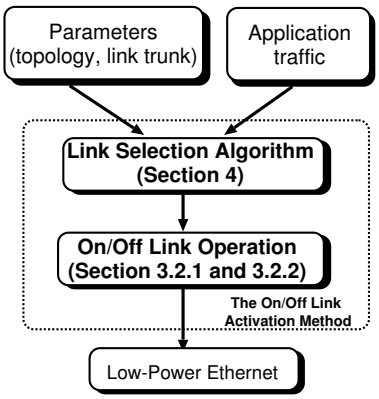


図 1 On/Off リンクアクティベーション法の概要

3. イーサネットにおける On/Off リンクアクティベーション法

イーサネットにおける On/Off リンクアクティベーション法の概要を図 1 に示す．本章では，提案手法においてもっとも重要であるリンクの On/Off 操作の実現方法を述べ，On/Off するリンクを定める On/Off リンク選択アルゴリズムについては次章で提案する．

3.1 準　　備

3.1.1 スイッチにおいてタグ付けを行う VLAN ルーティング法

ドライバ，ホストの更新なしに，隣接デバイスと通信可能な (On) リンク群を用いた経路の管理を行うために，本稿では我々が以前に提案を行った，スイッチにおいて VLAN タグ付けを行う VLAN ルーティング法を応用する[8]．

イーサネットのトポロジはツリーであるが，VLAN ルーティング法では実ネットワークの部分集合で構成されるツリートポロジの VLAN を組み合わせることで任意のトポロジを形成する．

スイッチにおいて VLAN タグ付けを行う VLAN ルーティング法において，ホストと接続されたスイッチポートは，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スイッチポートに対し，以下の 2 種類の設定を行う．

- スイッチポートの 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” メンバとする．

図 2 は，同志社大学 SuperNova クラスタのトポロジの例であり，4 個の VLAN を用いて次元順ルーティングを実装している．図中の円はスイッチを表しており各スイッチは 28〜29 台のホストが接続されている．図 2 において，ホスト 1〜28 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 #101 を付与され，すべての宛先について VLAN #101 内によって転送される．そして，宛先ホストに接続しているスイッチの出力ポートにおいて VLAN タグ #101 を除去する．一方，ホスト 29〜56 から送出されたフレームも同様の

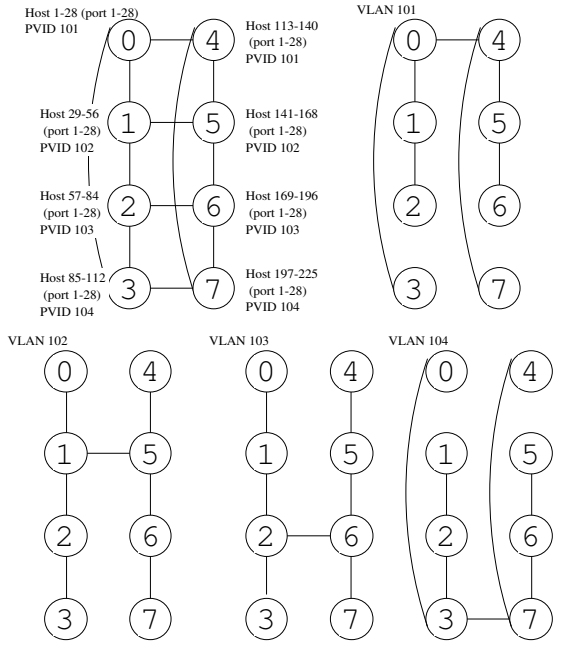


図 2 スイッチで VLAN タグ付けを行うルーティングの例

方法で VLAN #102 によって転送される．このようにして，ホスト側で VLAN がサポートされていなくても，各 VLAN におけるブロードキャストストームを避けつつ，全ホストの相互通信が可能になる．

3.1.2 スイッチにおける MAC アドレスの管理

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，スイッチが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 VLAN メンバとなっている全ポートからフレームを出力するため (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伴わずにフレームの交換が実現されるようになる．

しかし，MAC アドレス登録は，VLAN 毎に独立して行われる．VLAN ルーティング法を用いる場合，往路と復路で異なる VLAN を用いるため，そのままでは MAC アドレステーブルに宛先ホストが登録されない問題が生じる．

この問題は，静的に MAC アドレスをスイッチに登録することで解決できる．ただし，この場合は，MAC アドレスの自動学習を利用できないため，設計者のスイッチ設定の負担が大きい．そこで，この問題を解決する MAC アドレス学習法を次のように提案する．

(1) 各スイッチの MAC アドレスの保持期間である Aging Time を無限に設定する．無限に設定できない場合は最大値とし，その時間間隔で以下の手続きを繰り返し実行する．

(2) 全 VLAN に対応する仮想インタフェースを各ホストにおいて vconfig 等を使って作成する．例えば図 2 の場合，vlan 101〜104 を用いるため，各ホストにおいて eth0.101〜eth0.104 までを作成する．

(3) VLAN 毎に一意のネットワーク (IP) アドレスを与え，VLAN 毎に別々のセグメントに属するように各ホストの仮想インタフェースに IP アドレスを割り振る．

(4) 各仮想インタフェースごとに，ICMP または UDP メッセージを一度ブロードキャストする．

ステップ (4) では，各ホストにおいて，例えば各 VLAN セグメント内で全ホストに対して ping(ICMP echo req.) を送信することで実現することができる．これにより，各スイッチにおいて，各 VLAN のアドレステーブルに送信ホストの MAC アドレスが登録される．

MAC アドレス学習法において，ホストの仮想インタフェースはスイッチの MAC アドレスを学習するためのみに使用される．

3.2 On/Off リンクによる経路更新

スイッチの On/Off ポート操作により，一部のホスト間の経路が変更される．ここでは，その経路管理法について説明する．

ここで前提として，(1) On/Off 操作後に使用する経路群は，前節で述べた，スイッチでタグ付けを行う VLAN ルーティング[8] で実装し，(2) ルーティングアルゴリズムとして，任意のトポロジに適用することができる既存のルーティングアルゴリズム[3] を用いることとする．

3.2.1 リンクを On から Off にする場合

(1) リンクを Off にした後に使用する経路群を含む VLAN を作成し，その経路群が使用するスイッチ間ポートに対して “tagged メンバ” として，その VLAN を追加する．

(2) 新たに作成した VLAN ID における MAC アドレステーブルを静的に設計者が与える，あるいは，前節の MAC アドレス学習法により自動生成する．

(3) 更新される経路の出発スイッチにおいて，ホストへ接続しているポートの PVID をステップ (1) で作成した VLAN ID に更新する．

(4) 対象とするリンクを Off にする．

各ホスト間の通信は，ステップ (3) が完了するまでは，更新前の経路を用いて行うことになる．例えば，図 2 で示したトーラストポロジ，次元順ルーティングにおいて，スイッチ 1-5 間のリンクを Off にしたとする．この場合，このリンクを利用する経路は，スイッチ 1,5 のホストを起点とするため，スイッチ 1,5 のホストと接続しているポートの PVID を# 101 から# 102 にすることで Off リンクを迂回することができる．

3.2.2 リンクを Off から On にする場合

(1) リンクを On とした後に使用する経路群を含む VLAN を作成し，その経路群が使用するスイッチ間ポートに対して “tagged メンバ” として，その VLAN を追加する．

(2) 対象とするリンクのスイッチのポートを On にする．

(3) 新たに作成した VLAN ID における MAC アドレステーブルを静的に設計者が与える，あるいは前節の MAC アドレス学習法により自動生成する．

(4) 更新される経路の出発スイッチにおいて，ホストへ接続しているポートの VLAN ID を，ステップ (1) で作成した VLAN ID に更新する．

本節で述べたスイッチのポートの On/Off 操作は，現状では，PC クラスタのオペレータが手動で行う必要があるため，スイッチ数が増加した場合には，その分工数がかかることにな

る．ただし，SNMP 等で外部からスイッチのトラフィック量をモニタリングし，スイッチのリンクの On/Off のコマンドを外部から実行できる場合には，スイッチの外から自動制御可能である．

3.3 適用範囲および制限事項

現在，イーサネットスイッチは，GbE に限定したとしても数千円〜100 万円前後のものまで多岐に渡る．しかし，極めて安価なスイッチは VLAN 技術にすら対応していないものが多く，ごく一部の機能のみを提供している安価なイーサネットスイッチでは On/Off リンクアクティベーション法を利用できない場合がある．つまり，On/Off リンクアクティベーション法は IEEE 802.1Q に準拠した VLAN タグ操作をサポートしたスイッチでのみ適用可能である．

また，構築可能なシステムの規模は，使用するスイッチの仕様 (使用可能な VLAN 数，および登録可能な MAC アドレス数) により上限が定まる．VLAN 数は IEEE 802.1Q の規定では 4,094 ($2^{12}-2$) 個と有限であり，コストパフォーマンスの高い安価な商用スイッチでは数十〜数百個程度しかサポートしていないものも多い．ただし，スイッチでタグ付けを行う VLAN ルーティング法で必要となる VLAN 数は，Fat ツリーやメッシュなどの規則性の強いトポロジでは上記に比べて少ない[8]．そのため，スイッチのサポートしている VLAN 数がシステム規模を制限することは，ほとんどない．

一方，スイッチ内で静的または (学習により) 動的に登録されるホストの MAC アドレスは，同じアドレスであっても VLAN ごとに別々に登録されることになる．このため，登録可能なホストの MAC アドレス数 H は，使用する VLAN 数 V，スイッチに (静的に) 登録できる MAC アドレス数 M により以下となる．

$$H = \frac{M}{V} \tag{1}$$

4. On/Off リンク選択アルゴリズム

本章では，On/Off するリンクをトラフィック負荷に応じて選択するアルゴリズムを提案する．提案アルゴリズムでは，予期せぬトラフィックにより Off リンクが On になるまで待ち続けるフレームが生じることを避けるために，ネットワークの連結性を保証するツリー (スイッチ間リンクは 1 本) に属するリンクを常に On (ever-on) にする．

On/Off リンク選択アルゴリズムでは，サンプル調査，事前に対象とする並列アプリケーションを 1 度実行するなどにより各ホスト間の通信量が予測できるとする．また，既存のイーサネットスイッチでは，リンクの On/Off アクティベーションに秒単位の時間がかかるため，並列アプリケーション単位やそれ以上の大きな時間単位でリンクの On/Off を制御するものとする．

On/Off リンクアクティベーションアルゴリズムは，以下の手順により，各リンクの通信量を見積り，トラフィック量が一定量 Th に満たないチャネルで構成されるリンクを，前節で提案した On/Off 操作の実現方法により Off，あるいは On にする．なお，リンクは反対向きの単方向チャネル 2 本により構成されるものとする．

(1) 各チャネルのトラフィック負荷を見積る．
  (a) 各チャネルに，本アルゴリズム上で用いる論理的なカウンタを割り当て，全ての値を 0 に初期化する．
  (b) 通信が発生するホストペアの集合から未解析なペアを一つ取り出す．
  (c) そのホストペアの経路が利用する全てのチャネルのカウンタを，そのトラフィック量だけ増やす．ただし，リンク集約化 (スイッチ間に複数リンクを設置) を行っている場合，カウンタ値とは 1 本のチャネルあたりの平均値とする．
  (d) b〜c のステップを，全ての通信パスの解析が完了するまで繰り返す．
(2) カウンタの最大値を持つチャネルのリンクを起点として，幅優先探索でカウンタ値の大きいリンクを選択して，(ever-on となる) ツリーを構築する．
(3) チャネルのカウンタ値の最大値と Th の小さい方の値を Max とする．
(4) ツリーに属さない未解析なチャネルの中から，トラフィック量のもっとも少ないチャネルを通過するすべての経路を，そのチャネルを迂回する経路に更新する．なおリンク集約化を行っている場合は，経路は変更せずに，そのスイッチ間のチャネル数を 1 つ削減する．
(5) 各チャネルの経路更新後のカウンタ値を計算し，その中の最大値と Max を比較する．
  - Max の方が大きい場合，チャネルのカウンタを経路更新後の値に更新する．そして選択チャネルを Off にする．
  - Max の方が小さい場合，経路更新を破棄し，更新前の経路，チャネル集合に戻す．
(6) 4〜5 のステップを，全てのツリーに属さないチャネルの解析が完了するまで繰り返す．
(7) リンクを構成する 2 つのチャネルが Off の場合，そのリンクを Off にする．

5. 評　価

本章では，商用 GbE スイッチにおけるポートシャットダウン，速度減速による電力削減効果と，リンクの On/Off 制御時間，経路更新にかかるオーバヘッド，並列アプリケーションにおける提案手法の評価結果を明らかにする．

5.1 スイッチのポートあたりの消費電力

既存のスイッチにおけるポートシャットダウンによる電力削減を表 2 に示す．電力測定はシステムアートウェア ワットアワーメータ (SHW3A) を用いて行った．

表 2 において，Gb/Port，100M/Port はポートの通信速度をスイッチ外部から各々設定した場合に，ポートあたりに増加する消費電力を表している．PC5324, PC6224(24 ポート), PC6248(48 ポート) はデル PowerConnect スイッチシリーズの型番であり，SF-420 は Planex SF-0420G(24 ポート) である．また “ポート以外” とは，全ポートをシャットダウンした場合の消費電力を表している．表 2 より，リンク速度を落とすにしたがって，消費電力が劇的に削減されることが分かる．これらはトラフィック負荷によらず一定である．1 本のリンクを Off にする場合，両端のポートをシャットダウンすることになるため，消費電力の削減量は実際には 2 倍となる．L3 スイッチである PC6224, PC6248 は提供サービスが豊富かつ高度であ

表 2　GbE スイッチの消費電力 (W)

	ポート以外	Gb/ポート	100M/ポート
PC5324	14.9	1.2	0.9
PC6224	42.5	2.0	0.9
PC6248	56.8	2.1	1.3
SF-420	32.6	1.0	0.2

表 3　Overhead of On/Off Operation (Sec)

GbE SW	On/Off Operation	VLAN Modification
PC6224	3.4	0
PC6248	2.2	0
PC5324	4.0	0
SF-420G	12.0	0

るため，L2 スイッチである PC5324, SF-0420G に比べてポート，ポート以外の各要素の消費電力が大きいと考えられる．

5.2　On/Off リンク操作のオーバヘッド

スイッチの On/Off リンク操作のオーバヘッドを表 3 に示す．本測定では，オーバヘッドを，測定対象のスイッチを介して 2 ホストの ping (ICMP echo メッセージ) を 0.1 秒間隔で注入し，スイッチのポートに対して On/Off 操作を連続して行った場合の通信中断時間とした．ICMP echo メッセージのサイズはヘッダを含めて 64byte とした．よって，イーサネットフレームのデータサイズは IP ヘッダ 20byte を含めて 84byte である．評価結果より，リンクの On/Off 操作は数秒のオーバヘッドがかかり，そのほとんどがリンクの Off から On に状態が wakeup する場合にかかっていることが分かった．また，経路の更新は，ホストが接続しているスイッチポートの PVID を更新することにより実現されるが，そのオーバヘッドは表 3 のように，本測定では *zero* であった．

5.3　NAS パラレルベンチマークにおける消費電力

5.3.1　評価に用いた PC クラスタ

本評価では 128 台のホストを 8 台の 48 ポートの GbE スイッチ (Dell 社 PowerConnect6248) で接続した PC クラスタを用いた．スイッチ間リンク数は 6 本，ホストとスイッチの間は 1 本のリンクで接続されており，リンク集約化は出発地，目的地の IP アドレス，UDP/TCP のポート番号でリンク間のトラフィック分散を行った．図 2 のようにトポロジはトーラスであり，4 個の VLAN を用いて次元順ルーティングを VLAN ルーティング法により実装している．

本評価では任意のトポロジに適用でき，かつ，次元順ルーティングの経路を含むルーティングアルゴリズムである L-Turn ルーティング[3] を用いた．ただし，スイッチ間リンク数を 6 本としたため，本トラフィックパターンではスイッチ間リンク数の調整のみで，リンク選択アルゴリズムによる経路の更新はほとんど生じなかった．

表 4 に汎用のコンポーネントで構成されているホストの仕様を示す．並列ベンチマークは，MPICH 1.2.7.p1 を用いた IP パケットによりプロセス間通信を行った．

On/Off リンクアクティベーション法に用いたトレースは，各プロセス数において NAS パラレルベンチマーク 3.2 のアプリケーションが実行可能な最小のクラス (クラス W，A) について，MPE プロファイリングライブラリを用いて収集した．

128 プロセスを用いた場合は，1 つのスイッチにつき 16 台のホストの計 128 台のホストを用い，64 プロセスを用いた場合は，1 つのスイッチにつき 8 台，計 64 台のホストを用いた．

表 4　ホストの仕様

CPU	AMD Opteron 1.8GHz × 2
Chipset	AMD 8131+8111
Memory	PC2700 Registered ECC 2GB
OS	Debian GNU/Linux 4.0
Kernel	2.6.18-4-amd64
MPICH	1.2.7p1

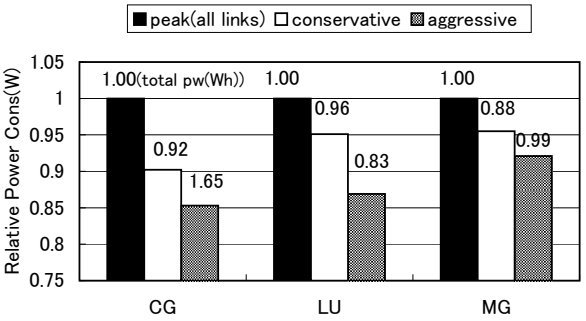


図 3　電力性能 (128 プロセス，クラス C, PC6248)

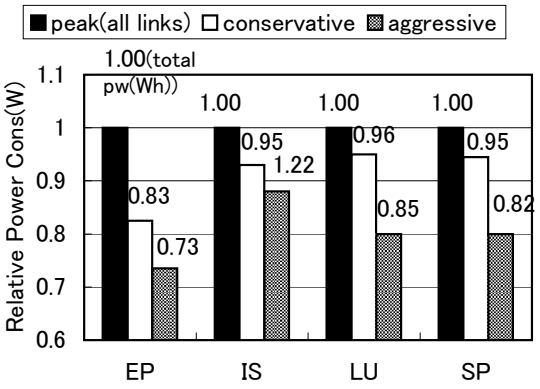


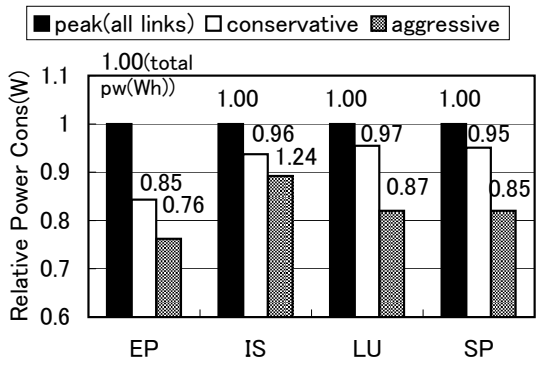
図 4　電力性能 (64 プロセス，クラス C, PC6248)

5.3.2　評価結果

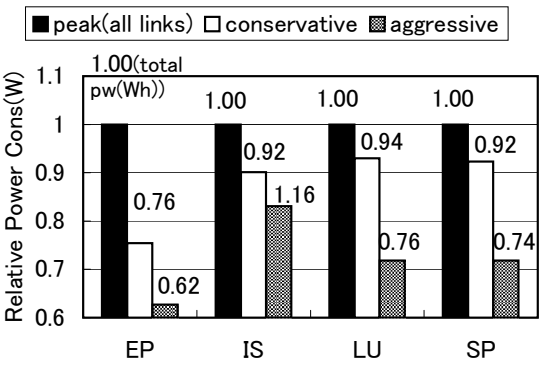
表 2 と Off リンク数からスイッチ (PowerConnect 6248) の総消費電力量を算出した結果を図 3 と 4 に示す．比較評価のため，64 プロセスの場合はスイッチ間リンク数を 5 本とした．

すべてのリンクを On にした場合 (“peak”) の性能値により正規化し，“conservative” はリンク選択アルゴリズムにおいて *trf=0* の場合，“aggressive” は性能よりもリンクの削減を重視して *trf* を設定した場合である．図 3，4 内の数値 (total pw(wh)) は，実行時間と消費電力から求めたネットワークの相対的な電力消費量である．ただし，ネットワークの性能低下によりアプリケーションの実行時間が増加した場合，ホスト群の電力消費量が大きくなる．そのため，リンク選択アルゴリズムの指標として (ネットワークの性能低下が生じない範囲での) スイッチの消費電力が最も重要となる．

評価結果より，総トラフィック量を見積る提案リンク選択法により，“conservative” の場合，64，128 プロセスの両方においてアプリケーションの性能低下は生じず，さらに “aggressive” の場合についても一部の場合には性能低下が生じないことが分



(a) SF-0420G



(b) PowerConnect 5324

図 5 電力性能 (64 Processes, NPB Class C, SF-0420G と PC5324)

かった．つまり，提案リンク選択法は性能低下をすることなく，スイッチの消費電力を最大 26%削減することができた．

64 プロセスの場合において，本評価に用いた PC6248 と同様に，他のノンブロッキングスイッチである PC5324，SF-0420G を用いた場合の電力性能見積りを表 2 から算出した結果を図 5 に示す．いずれのスイッチもノンブロッキングであり，フレームの通過遅延時間に大差がないためネットワークの最大性能はほぼ同じといえる．評価結果より，L3 スイッチである PC6248 に比べ，L2 スイッチを用いることにより，本提案手法の電力削減効果はより大きくなり，最大 37%であることが分かる．

6. ま と め

本稿ではイーサネットのレイヤ 2 スイッチの消費電力を削減するために，トラフィック負荷に応じてリンクを On/Off するアクティベーション法の提案，評価を行った．On/Off リンクアクティベーション法は，VLAN ルーティング法を応用することでブロードキャストストームを避けつつ，ホスト間の経路を更新する．On/Off リンクアクティベーション法は，スイッチの既存の機能を制御することにより実現でき，ホスト側の設定は不要である．

PC クラスタの使い方の 1 つに，パラメータサーベイのために，特定の科学技術演算等の並列プログラムを繰り返し実行することが挙げられる．また，インターネットにおいて盛んに行われているトラフィック解析と同様に，運用のために PC クラスタの通信履歴を利用できる場合がある．この 2 つの場合には，静的解析により，各ホスト間の通信量を予測することができるため，特に On/Off リンクアクティベーション法を適用しやすいといえる．

評価結果より，リンクの On/Off 操作，経路更新の総オーバヘッドは商用スイッチでは数秒であった．NAS パラレルベンチマークを用いた PC クラスタの評価結果り，On/Off アクティベーション法は性能低下なしに，スイッチの総消費電力を最大 37%削減できることがわかった．インターコネクトはリンクバンド幅が向上するにつれて，消費電力は増える一方であり，今後，PC クラスタのインターコネクトの候補である 10GBase-T リンクでは現状の数倍の消費電力となる可能性が高い．よって，On/Off リンクアクティベーション法の電力削減効果は，今後より大きくなると考えられる．

謝　　辞

同志社大学 SuperNova クラスタの使用をご快諾いただいた同大学廣安 知之教授、中尾 昌広氏、渡辺 崇文氏に感謝致します。本研究の一部は、科学技術振興機構「JST」の戦略的創造研究推進事業「CREST」の支援による．

参 考 文 献

1) Top 500 Supercomputer Sites: `http://www.top500.org/`.
2) Soteriou, V. and Peh, L.-S.: Exploring the Design Space of Self-Regulating Power-Aware On/Off Interconnection Networks, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 18, No. 3, pp. 393–408 (2007).
3) Jouraku, A., Koibuchi, M. and Amano, H.: An Effective Design of Deadlock-Free Routing Algorithms Based on 2-D Turn Model for Irregular Networks, *IEEE Transaction on Parallel and Distributed Systems*, Vol. 18, No. 3, pp. 320–333 (2007).
4) Sharma, S., Gopalan, K., Nanda, S. and Chiueh, T.: Viking: A Multi-Spanning-Tree Ethernet Architecture for Metropolitan Area and Cluster Networks, *Infocom*, pp. 2283–2294 (2004).
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) Okamoto, T., Miura, S., Boku, T., Sato, M. and Takahashi, D.: RI2N/UDP: High bandwidth and fault-tolerant network for PC-cluster based on multi-link Ethernet, *The Workshop on Communication Architecture for Clusters (CAC), IPDPS* (2006).
7) Miura, S., Boku, T., Okamoto, T. and Hanawa, T.: A dynamic routing control system for high-performance PC cluster with multi- path Ethernet connection , *IEEE International Symposium on Parallel and Distributed Processing (IPDPS)* (2008).
8) 大塚智宏, 鯉渕道紘, 工藤知宏, 天野英晴: 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 12(ACS 15), pp. 46–58 (2006).