

# 実HPC環境におけるEEEの電力/性能評価

三輪 忍<sup>1,a)</sup> 會田 翔<sup>1</sup> 安島 雄一郎<sup>2</sup> 清水 俊幸<sup>2</sup> 安里 彰<sup>2</sup> 中村 宏<sup>1</sup>

受付日 2014年4月4日, 採録日 2014年9月21日

**概要:** 近年のHPCシステムではその規模が供給電力によって制限されることが多い。今後、HPCシステムの処理能力をさらに向上させてエクサフロップスを実現するためには、システムの電力性能を高めることが必要不可欠である。本稿ではインタコネクション・ネットワークの省電力化手法として、最新のイーサネットを採用されている技術であるEnergy Efficient Ethernet (EEE)に着目する。EEEは将来のHPCシステムにおいて採用が期待されている技術の1つであるが、それをHPCシステムに採用した場合にシステムの性能と電力にどのような影響を与えるかは明らかでない。そこで我々は、10GBASE-Tのネットワークによって構成された実機を用いてEEEの効果の詳細に評価した。本稿ではその結果を報告する。

**キーワード:** HPC, インタコネクション・ネットワーク, 電力, EEE

## Power/Performance Evaluation of EEE in Real HPC Environment

SHINOBU MIWA<sup>1,a)</sup> SHO AITA<sup>1</sup> YUICHIRO AJIMA<sup>2</sup>  
TOSHIYUKI SHIMIZU<sup>2</sup> AKIRA ASATO<sup>2</sup> HIROSHI NAKAMURA<sup>1</sup>

Received: April 4, 2014, Accepted: September 21, 2014

**Abstract:** Since modern HPC systems are often constrained by their power supply, improvement of performance per watt is indispensable to the future HPC systems like an Exa-scale system. This paper focuses on Energy Efficient Ethernet (EEE), which is adopted in state-of-the-art Ethernet for saving network power. EEE is expected to be adopted in the future HPC systems, but its impact on power/performance of HPC systems is still unclear. This paper reports the result of our through evaluation of EEE with a 10GBASE-T system.

**Keywords:** HPC, interconnection networks, power, EEE

### 1. はじめに

近年のスーパーコンピュータは、供給可能な電力によってシステムの規模が制限されるようになってきている。たとえば現在世界最速のコンピュータである Tianhe-2 は、33.9 PFLOPS の性能を達成するために 17.6 MW の電力を消費している [24]。1 つの計算機センタに供給可能な電力は現実的には 20~30 MW といわれており [26]、上記の数値は、現時点で世界最速のシステムがすでに供給電力による制約を受けつつあることを示唆している。今後さらにシ

ステムの性能を向上させてエクサフロップス級のシステムを実現するためには、上記の電力制約下での大幅な性能向上が求められる。そのためには、現在のシステムの電力性能を大幅に改善する必要がある。

我々はスーパーコンピュータのインタコネクション・ネットワークに着目する。近年のスーパーコンピュータでは高速、広帯域幅かつ冗長度の高いネットワークが採用されており、その消費電力が無視できないレベルにまで達している。たとえば、「京」コンピュータでは、ネットワークの制御を行う専用 LSI であるインタコネクト・コントローラは、CPU の約半分の電力を消費している [4]。将来の HPC システムではネットワークの消費電力がシステム全体のその 33% に達するとの予測もある [15]。CPU やメモリなどと比べて、ネットワークはこれまであまり省電力化の必

<sup>1</sup> 東京大学  
The University of Tokyo, Bunkyo, Tokyo 113-8656, Japan

<sup>2</sup> 富士通株式会社  
Fujitsu Limited, Kawasaki, Kanagawa 211-8588, Japan

<sup>a)</sup> miwa@hal.ipc.i.u-tokyo.ac.jp

要性が認識されてこなかったが、今後はネットワークにおいても省電力化を図ることが重要である。

ネットワーク機器においては、ネットワーク・リンクと機器とを接続する部分に位置する PHY と呼ばれる回路が多くの電力を消費している。PHY は、リンクの接続状態を確認するため、送信データが何もないときでもつねにアクティベートされ、パケットのやりとりを行っている。そのため、アプリケーションが通信を行っていない状態であっても、非常に多くの電力を消費している。

この問題を解決するため、イーサネットにおいては、Energy Efficient Ethernet (以下 EEE とする) という技術が最近になって用いられるようになってきた [14]。EEE は上述の無駄なパケットのやりとりを抑制することで、PHY の消費電力を削減する手法である。EEE は 2010 年に IEEE802.3az として標準化され、その結果、各ネットワーク機器メーカーは徐々に EEE への対応を進めている [10], [13], [19], [25]。EEE は今はまだイーサネット限定の標準であるが、バックプレーン用の 100 Gbps 以下のメタル配線 [7], [12] やファイバ・チャネル [11] など、他のネットワーク規格においても標準化の動きが進んでいる。そのため、EEE のような PHY の省電力化技術が将来のスーパーコンピュータのネットワークに採用される可能性は高いといえる。

ところが、EEE を HPC システムに採用した場合のシステム性能や消費電力を詳細に評価したケースはほとんどない。Saravanan らは、我々と同様、EEE を採用した HPC システムの電力/性能評価を行っている [22]。ただし、彼らの評価はシミュレーションによって行われており、実システムで EEE を採用した場合に、電力と性能にどのような影響があるかはまだ明らかでない。実機を用いて EEE の電力や性能を評価したケースもあるが、ピンポン通信などの簡単なプログラムを実行したときの評価にとどまっている [20], [21]。並列アプリケーションを実機上で実行し、評価を行ったわけではない。

そこで本稿では、並列アプリケーションを実行した際に EEE がシステムの性能と電力に与える影響を実機を用いて評価する。EEE による性能への影響が軽微で、かつ、大幅に電力を削減できるのであれば、HPC システム開発者にとって EEE を採用する大きなモチベーションとなる。我々は過去に EEE 対応のギガビット・イーサネットのシステム上でいくつかの並列アプリケーションを用いてシステムの性能を測定したことがあるが [17]、電力そのものの評価、および、10 ギガビット・イーサネットのシステムを用いた評価は今回が初めてである。スーパーコンピュータでは 1 リンク数十ギガバイトにも達する広帯域幅のネットワークが用いられることが多いことから、10 ギガビット・イーサネットを用いた今回のシステムの方が、HPC システムをより正確にスケールダウンしたものと見える。

本稿の構成は以下のとおりである。まず次章で EEE について詳しく説明する。続く 3 章では実験方法を説明し、結果は 4 章で示す。今回実験を行ったのは最大 16 ノードのシステムであるため、5 章では大規模環境において EEE を使用した場合について考察する。6 章では EEE が有効に機能する範囲を示し、7 章では Saravanan らの実験で得られた知見との相違点を述べ、最後に 8 章で本稿をまとめる。

## 2. EEE

前述のように、ネットワーク機器内に存在するモジュールの中でも特に、PHY は多くの電力を消費している。これは、リンクの接続状態を確認するため、リンクの両端に位置する 2 つの PHY はつねにアクティベートされ、お互いにパケットをやりとりしているからである。送信データが何もないときは、PHY は IDLE コードと呼ばれる特殊なパケットを生成し、それを定期的に送信することによって、そのリンクが通信可能な状態にあることを確認する。

PHY の消費電力を抑えるため、IEEE 802.3az タスク・フォースによって EEE は開発された。図 1 に EEE のコンセプトを示す。EEE の基本的なアイデアは上述の IDLE コードの送受信回数を減らすことである。送信データが何もないときに、EEE 対応の PHY は、IDLE コードの代わりに LPI (Low Power Idle) コードを送信する。受信側の PHY も EEE に対応していた場合は、受け取った LPI コードに対して応答を返す。これにより、2 つの PHY は LPI モードと呼ばれる省電力モードへと移行する。LPI モード中は、PHY 内の多くのハードウェアがシャットダウンされた状態となる。このモード移行にはある程度の時間 ( $T_s$ ) を要し、その間の電力はアクティブ状態とほぼ同じである。

LPI モード中の PHY にパケットがやってきたときは、PHY はウェイクアップを開始する。すべてのハードウェアの電源が復帰し、もう一方の PHY とパケットのやりとりを行うことでお互いに通信の準備が整っているかを確認する。このウェイクアップ処理もまた、アクティブ状態とほぼ同等の電力を消費する。ウェイクアップが ( $T_w$  の時間をかけて) 完了すると、PHY はパケットの送信を開始する。すなわち、LPI モード中の PHY に到着したパケットは、ウェイクアップ時間 ( $T_w$ ) 分の性能ペナルティを被る。

LPI モード中にリンクの接続状態を確認するため、リフレッシュと呼ばれる処理が行われる。LPI モードに移行し

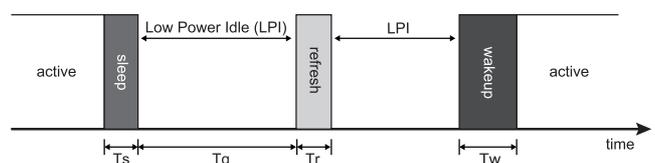


図 1 Energy Efficient Ethernet (EEE)

Fig. 1 Energy Efficient Ethernet (EEE).

表 1 EEE のモード移行時間 [21]

Table 1 Mode transition time of EEE [21].

Protocol	Min $T_s$ ( $\mu\text{sec}$ )	Min $T_w$ ( $\mu\text{sec}$ )
100BASE-TX	200	30
1000BASE-T	182	16
10GBASE-T	2.88	4.48

た PHY は、一定時間 ( $T_q$ ) 経過後に起動され、リンク接続状態を確認するためのパケットのやりとりを行う。このために、スリープ処理中に PHY は内部のタイマに起動時刻をセットしておく。パケットのやりとりが終わり接続状態が確認されると、リンクの両端の PHY は再びタイマをセットし、LPI モードへと移行する。

IEEE 802.3 task force が想定しているモード移行時間 ( $T_s$ ,  $T_w$ ) の最小値を表 1 に示す [21]。この値はモード移行に最低限必要な時間であり、後述するように、実際のモード移行時間とは異なることに注意されたい。スリープ処理およびウェイクアップ処理に要する時間は、表に示すように、10GBASE-T プロトコルの場合は数マイクロ秒である。このような高速なモード遷移により、アプリケーションの実行中に生じる短いアイドル期間であっても、EEE 対応の PHY は省電力モードへと移行することができる。

EEE は、LPI モードに移行するタイミングなど、PHY の電力管理方針を規定したものではないことに注意されたい。EEE は MAC 層のプロトコルであり、電力管理の方針は各ネットワーク機器メーカーが独自に定めることができる。そのため、現在市販されている EEE 対応の機器の電力管理方式は機器によって異なる。我々の調査によれば、多くの EEE 対応機器は、IT 機器の電力制御方式として広く用いられている、タイムアウト制御によるスリープ [16], [23], および、オンデマンドによるウェイクアップを行っている。

前述のように、EEE は 2010 年に IEEE 802.3az として標準化された。その後、この規格に従うネットワーク機器が徐々に増え始めている [10], [13], [19], [25]。しかし、EEE は現時点ではイーサネットの標準であるため、InfiniBand などの他のネットワークではまだ利用できない。たとえば IEEE P802.3bj タスク・フォースでは EEE に関する議論が続けられているが、100 Gbps 以下のメタル配線において EEE が採用されるにはあと数年はかかるだろう [7], [12]。そのため、HPC 分野では、EEE は主要な技術としてまだ広く認知されていない。

### 3. 実験方法

EEE 対応のネットワーク機器を用いてシステムを構築し、その基本電力/性能および並列アプリケーション実行時の電力/性能について評価を行った。本章では構築したシステムと行った実験内容について説明する。

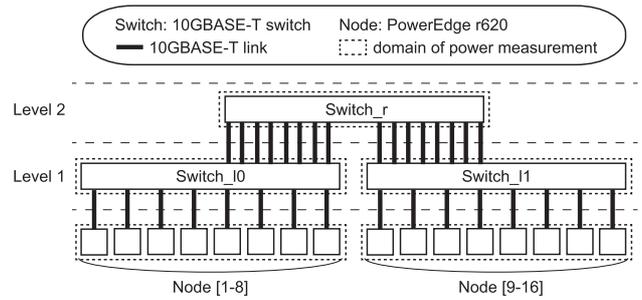


図 2 評価システム

Fig. 2 Evaluation system.

表 2 ノード (PowerEdge r620) の構成

Table 2 Configuration of a node (PowerEdge r620).

Device	Remarks
CPU	Xeon E5-2630L × 1 Clock frequency: 2.0 GHz # of core: 6 TDP: 60 W
Memory	32 GB 4 GB 1R × 4 DDR3-1333 LV-RDIMM × 8
HDD	292 GB
NIC	Broadcom 57810S DP 10Gb BASE-T adapter
OS	Scientific Linux 6.5

### 3.1 実験環境

今回の評価に用いたシステムを図 2 に示す。後述するように、各実験は図のシステムの一部または全部を使用して行う。図に示すように、評価システムは 16 ノードを有し、2 階層の fat-tree 構造をしている。2 台の 1 階層目のスイッチ (“switch\_l0” と “switch\_l1”) にはそれぞれ 8 ノードが接続されており、接続には 10GBASE-T のリンクが用いられている。1 階層目のスイッチと 2 階層目のスイッチ (“switch\_r”) 間はそれぞれ 8 本の 10GBASE-T のリンクを用いて接続する。なお、図のネットワークは研究室内の他のネットワークとは切り離れた状態で実験を行った。

ノードの構成を表 2 にまとめる。各ノードは 1 つの CPU を有している。CPU は Xeon E5-2630L、コア数は 6 コアである。クロック周波数は 2.0 GHz、TDP (Thermal Design Power) は 60 W である。この TDP 値は「京」コンピュータで用いられている SPARC64VIIIfx のそれとほぼ同じである [18]。主記憶は 4 GB の 1 ランクの DDR3-1333 LV-RDIMM 8 枚からなり、計 32 GB の容量となっている。NIC は、EEE 対応の 10GBASE-T ネットワーク・アダプタである、Broadcom 社の 57810S DP 10GBASE-T アダプタを使用した。リンクには Cat. 7 ケーブルを使用する。したがって、評価システムにおける 1 ポートあたりのネットワーク帯域幅は 10 Gbps となり、これは Tofu ネットワークのおよそ 5 分の 1 に相当する [1]。

前章で述べたように、EEE は PHY の電力管理方針を規定したものではないため、同じ EEE 対応のネットワーク

表 3 各スイッチの諸元

Table 3 Specification of each switch.

	PowerConnect	M7100
プロトコル	10GBASE-T	10GBASE-T
ポート数	24	24
スイッチ・ファブリック	640 Gbps	480 Gbps
転送レート	480 Mpps	357.1 Mpps
システム・メモリ	2 GB	256 MB
フラッシュ・メモリ	256 MB	128 MB
パケット・バッファ・メモリ	9 MB	16 MB
EEE 対応	yes	yes
ウェイクアップ時間	17 us	17 us
タイムアウト時間	600 us	不明

機器であっても、ネットワーク性能や電力削減効果は機器によって異なると考えられる。そのため、本稿では異なる機器を実験に使用し、その性能や電力削減効果を比較する。具体的には、図 2 のネットワーク・スイッチには次のいずれか一方を使用する。

**PowerConnect** Dell 社製 PowerConnect 8132

**M7100** Netgear 社製 M7100-24x

各スイッチの諸元を表 3 にまとめる。PowerConnect, M7100 いずれも 24 個の 10GBASE-T のポートを有する。これらのポートのうち、最大 16 個のポートを実験に使用した。EEE の機能は、いずれのスイッチもポート単位で有効化/無効化できる。なお、リンク・レベルのフロー制御は有効にして実験を行った。

PowerConnect に関しては、送信側 (Tx) の EEE のタイムアウト時間、および、ウェイクアップ時間を管理コンソールを用いて設定/確認することができる。そのため、管理コンソールを通してこれらの値を変更し、これらの値が電力と性能に与える影響についても評価を行った。なお、これらのデフォルト値は、ウェイクアップ時間が 17 マイクロ秒 (表 1 の最小値とは異なることに注意されたい)、タイムアウト時間が 600 マイクロ秒である。一方、M7100 については管理コンソールから Tx のウェイクアップ時間の確認のみを行うことができ、その設定変更ならびにタイムアウト時間の設定変更/確認を行うことはできなかった。M7100 のウェイクアップ時間は、PowerConnect のデフォルト値と同様、17 マイクロ秒であった。

スイッチ単体の電力評価には WattsUp? .Net [27] という電力計を使用した。WattsUp? .Net は電子機器の 1 秒ごとの消費電力をミリワット単位で計測できる。この電力計を AC プラグとコンセントの間に挿入し、スイッチの消費電力を測定した。

システム全系の電力評価には横河電機の WT1800 を使用した。WT1800 は 50 ミリ秒間隔と非常に高解像度で電子機器の電流を計測することが可能である。この電力計を用いて、図 3 に示すように、機器の AC 電源部分の流れ

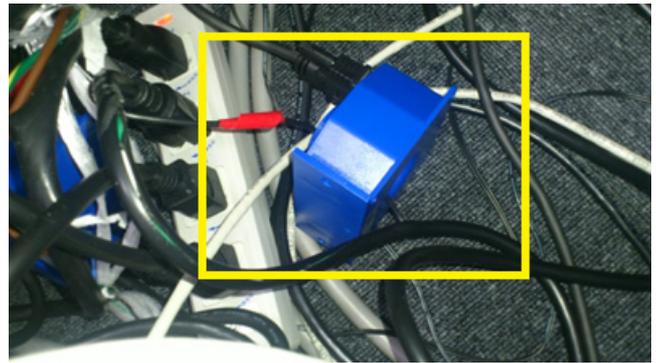


図 3 WT1800 を用いた電力測定の様子 (中央の四角で囲んだ物体が電流増幅器であり、そこに AC 電源ケーブルを通すことでケーブルを流れる電流を測定する)

Fig. 3 Power measurement with WT1800 (the device within the box is a current amplifier).

る電流を測定した。WT1800 には 6 本の測定チャンネルがあり、最大で 6 カ所までの電流を同時計測できる。本実験では、3 本のチャンネルを用いてそれぞれ 3 つのスイッチの AC 電源部分の電流計測を行い、残りの 2 本を用いてノード全体の電流計測を行った (図 2)。ノードの電流計測は、2 つの電源タップにノード 8 台ずつを接続し、2 つの電源タップを流れる電流をそれぞれ 2 本のチャンネルを用いて計測した。このようにして測定した各箇所の電流に 100 V の電源電圧を乗じることで各機器の消費電力を求めた。

### 3.2 基本電力/性能の評価実験

まずは簡単な実験を通して、EEE による省電力モード時の電力削減量および性能ペナルティを評価した。以下それぞれの実験内容について詳しく述べる。

#### 3.2.1 電力削減量の評価実験

省電力モード時の電力削減量を調べるため、ベンチマーク・プログラムを何も実行していない状態における電力測定を行った。実験には fat-tree の葉の部分に相当するシステム、すなわち、スイッチ 1 台とノード 8 台を用いた。スイッチとノードとを結ぶリンクの数を 1 本から 8 本まで変化させた際のスイッチの電力を測定した。なお、本実験に無関係なリンクは、ケーブルを物理的に抜いておくことでそれらの影響を排除する。EEE の機能を有効にした場合と無効にした場合それぞれの平均消費電力を測定し、その結果を比較する。また、スイッチの種類を変えて同様の実験を行い、スイッチによる電力削減量の違いを評価する。

#### 3.2.2 性能ペナルティの評価実験

前章で述べた性能ペナルティの影響を調べるため、パケットの送信間隔を変化させた際のピンポン通信の応答時間を評価した。実験には同一のスイッチに接続された 2 つのノードを使用した。実験に無関係なノードからのパケットの影響を排除するため、これらのノード以外のケーブルを物理的に抜き、実験を行った。パケットの送信間隔が

PHYのタイムアウト時間を超えるとPHYは省電力モードへと移行するため、次のパケットの送信時にはPHYのウェイクアップをしなければならず、その応答時間が悪化する。パケットのサイズは64B (ICMPヘッダを含む)とし、10,000個のパケットを一定間隔で送信し、その平均応答時間を求めた。ICMPパケットの生成にはBit-Twist [8]を使用し、応答パケットの観測にはWireshark [28]を使用した。パケットの送信間隔は0.1~2.0ミリ秒の範囲で変化させた。パケットが通過するリンクは2本存在するが、両方ともEEEを無効にした場合、1本のみEEEを有効にした場合、2本とも有効にした場合の計3通りについて測定を行った。

### 3.3 並列アプリケーションを用いた実験

続いて並列アプリケーションを実行した際の電力削減量、および、性能への影響を評価した。まずは図2のシステムの一部を用いて、スイッチの種類の違いによる電力削減量と性能を評価した。次いで今度は全系を使用し、実アプリケーションを実行した場合の電力削減量と性能を評価した。以下それぞれの実験内容を詳しく述べる。

#### 3.3.1 スイッチの比較実験

実験にはfat-treeの葉の一部、すなわち、スイッチ1台とそれに接続されたノード4台とを使用する。実験に無関係なノードのケーブルはすべて抜いておく。ノード4台、スイッチ1台の構成としたのは、1つには、ベンチマーク・プログラムの一部が $N$ の2乗 ( $N$ は正数) 並列にしか対応していないことによる。もう1つは、10GBASE-Tの24ポートのスイッチは高価であり、当研究室ではM7100を1台しか所有していないことによる。

実験にはNas Parallel Benchmark (NPB) 3.3 [6]を使用する。入力はクラスB, Cの2種類を用いた。各ノードに4MPIプロセス、計16MPI並列で実行した。コンパイルにはmpich2-1.4.1p1とgcc-4.8.0を使用する。コンパイル・オプションは“-O2 -funroll-loops”とした。なお、ISとMGの実行時間は1秒前後と非常に短く、本実験で使用した電力計ではプログラム実行中の消費電力の計測ができなかった。そのため、これらのプログラムの実験結果は次章で示すグラフからすべて省いてある。実験は各プログラムにつき5回行った。

実際の大型計算機センタでは、ジョブの実行時間の再現性を保証するため、Turbo BoostとHyper Threadingは無効化して運用されることが多い。そこで本実験でもこれらの機能は無効化して評価を行った。

#### 3.3.2 実アプリを使用した全系の評価実験

図2のシステムすべてを使用し、実アプリケーションを実行したときの平均電力と性能を評価した。3台のスイッチにはいずれもPowerConnectを使用する。

実験にはALPS/looper [3]を使用する。ALPS/looper

は、ALPS (Algorithm and Libraries for Physics Simulation) プロジェクト [2] の一環として開発されている、マルチクラスタ量子モンテカルロ・シミュレーションを行うプログラムである。ソース・コードはMPI/OpenMPのハイブリッド並列で記述されており、オープン・ソースで提供されている。ALPS/looperは「京」コンピュータやT2Kオープン・スパコンなどでの稼働実績を持ち、文部科学省が推進する「レイテンシコアの高度化・高効率化による将来のHPCIシステムに関する調査研究」においても、ポストペタスケール・システム開発の際に想定するアプリケーションの1つに選ばれている [30], [31], [32]。ALPS/looperは、新機能を持った強相関材料や磁性材料の物性予測や物性解明に利用されることが期待されている。

アプリケーションのコンパイルにはmpich2-1.4.1p1とgcc-4.8.0を使用した。コンパイル・オプションは“fopenmp”を使用した。プログラムの入力はデフォルトのもの ( $l = 524,288$ ,  $t = 0.00083$ ,  $n = 16$ ,  $\mu = 16$ ) を使用し、16MPI並列で実行した。実験の際はTurbo Boost, Hyper Threadingともに無効化し、測定はEEEを有効にした場合、無効にした場合それぞれ3回行った。

## 4. 実験結果

前章で述べた実験を行い、EEEによる電力削減量およびシステム性能への影響を評価した。本章ではその結果について詳しく述べる。

### 4.1 基本電力/性能の評価結果

まずは3.2.1項、および、3.2.2項で説明した、基本電力/性能に関する評価実験の結果を示す。

#### 4.1.1 電力削減量の評価結果

図4および図5にリンク数を変更したときの各スイッチの平均消費電力を示す。グラフの横軸はリンク数を表し、縦軸は平均消費電力を表している。左の棒グラフはEEEを無効化したときの電力を、図4の中央の棒グラフ

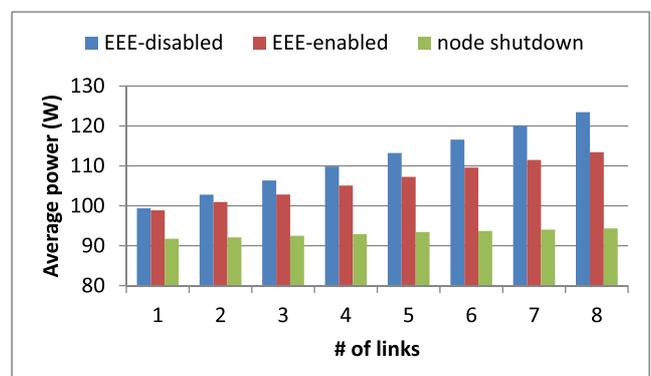


図4 リンク数を変えたときのPowerConnectの平均消費電力  
Fig. 4 Average power consumption of PowerConnect for various link count.

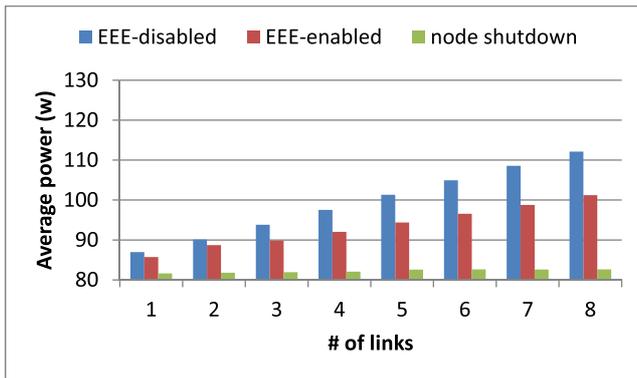


図 5 リンク数を変えたときの M7100 の平均消費電力

Fig. 5 Average power consumption of M7100 for various link count.

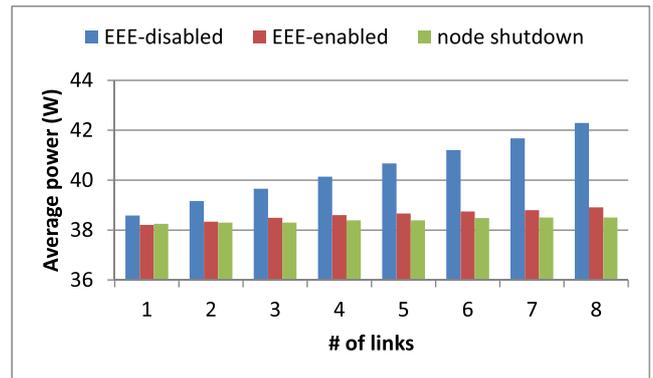


図 6 リンク数を変えたときの PowerConnect 5548 (1000BASE-T) の平均消費電力

Fig. 6 Average power consumption of PowerConnect 5548 (1000BASE-T) for various link count.

は EEE を有効化したときの電力を表す。一番右の棒グラフは、EEE を有効にし、かつ、ノードをシャットダウンした状態のスイッチの電力である。

グラフより、バックプレーンが消費する電力は異なるものの、1ポートが消費する電力、および、EEEによって削減される電力は、スイッチによる差がほとんどない。PowerConnect のバックプレーンの平均消費電力は約 96 W、M7100 のそれは約 83 W と 13 W の差がある。それに対しポートあたりの平均消費電力は、PowerConnect の場合は 3.44 W (8 リンク使用時)、M7100 の場合は 3.59 W (8 リンク使用時) とほぼ同じである。また、この状態で EEE を有効にすると、PowerConnect においては平均で 1.36 W、M7100 においては 1.37 W の電力削減が観測された。したがって、PowerConnect は EEE によりポートの電力を 39.8%、M7100 は 38.2%削減していることになる。これらのスイッチにおける PHY の電力削減量が限定的であるのは、省電力モード中であっても、PLL (Phase Locked Loop) の一部やパケットを格納するバッファなどの回路が稼働しているためと考えられる。

なお、どちらのスイッチも、ノードをシャットダウンした状態では、リンク数の違いによる消費電力の差はほとんど見られなかった。これは、ノードがシャットダウンされた状態では上記の回路もすべて電源遮断されるためと考えられる。

また、ノードをシャットダウンした状態における消費電力は、図 4 および図 5 から補間される、リンク数がゼロのときの消費電力よりも若干小さい。これは、スイッチにノードが何も接続されていない状態、あるいは、スイッチに接続されたすべてのノードがシャットダウンされた状態では、バックプレーンの回路の一部が省電力モードに移行するためと考えられる。実際、リンクが 1 本のみ挿さった状態からこのリンクを抜いてみると、スイッチの消費電力はいったんリンク 1 本分の消費電力 (1.36 W 程度) が下がった後に安定し、しばらく経つとそこからさらに数 W

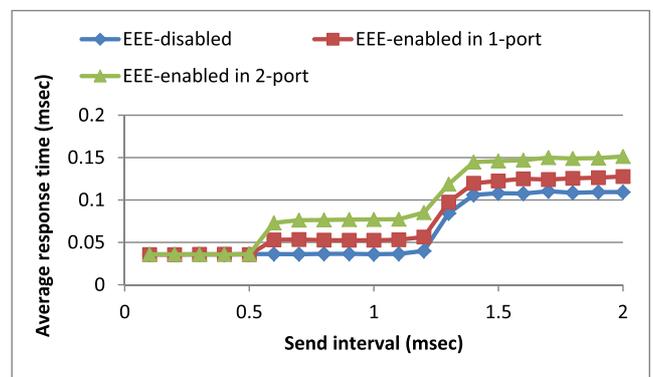


図 7 PowerConnect におけるパケット送信間隔に対する応答時間

Fig. 7 Average response time on PowerConnect for various send intervals.

低下する現象が見られた。

参考までに、EEE 対応の 1000BASE-T のスイッチにおいて同様の実験を行ったときの結果を図 6 に示す。使用したスイッチは、文献 [17] でも使用した、Dell 社の PowerConnect 5548 である。グラフより、1000BASE-T の場合は、ポートあたりの消費電力は 0.53 W (8 リンク使用時) であり、そのうちの 0.42 W の電力を EEE によって削減できた。すなわち、ポートの電力の 79.2% を EEE によって削減できる。このように、電力削減量で見れば 10GBASE-T の PHY が 1000BASE-T の PHY より大きい、電力削減割合で見れば後者の方が大きい。つまり、現在の 10GBASE-T の PHY は電力削減の余地をまだ残しているといえ、各ネットワーク機器メーカーにはこの部分の電力を削減することが期待される。

#### 4.1.2 性能ペナルティの評価結果

PowerConnect においてパケットの送信間隔を変えたときの応答時間を図 7 に示す。図の横軸はパケットの送信間隔、縦軸は 10,000 回のパケット送信を繰り返したときの平均応答時間である。3 本の折れ線グラフは、菱型のマークのグラフがリンク 2 本とも EEE を無効化した場合、四角

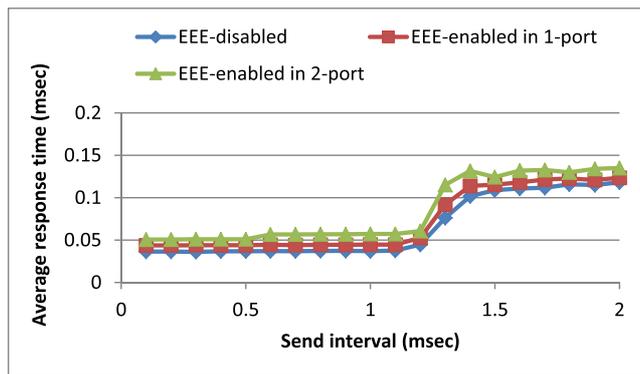


図 8 M7100 におけるパケット送信間隔に対する応答時間  
 Fig. 8 Average response time on M7100 for various send intervals.

形のマーカーのグラフが片方のリンクのみ EEE を有効にした場合、三角形のマーカーのグラフが 2 本とも EEE を有効にした場合を表す。

図 7 より、PowerConnect において EEE を用いた場合、パケットの送信間隔が 600 マイクロ秒を超えると応答時間が急激に悪化することが分かる。1 本のリンクの EEE を有効にした場合は 24 マイクロ秒（送信間隔 2 ミリ秒時）、2 本とも有効にした場合はさらに 18 マイクロ秒（同送信間隔時）の遅延の増加が見られた。これは表 3 に示した PowerConnect の PHY のウェイクアップ時間（17 マイクロ秒）とほぼ一致している。すなわち、送信間隔がタイムアウト時間である 600 マイクロ秒を超えたときは、パケット送信のたびに PHY のウェイクアップが発生していたことを表している。

応答時間の悪化率は、1 本のリンクの EEE を有効にした場合で 46.3%（送信間隔 0.6 ミリ秒時）、2 本とも有効にした場合で 201.4%（同送信間隔時）であった。このように、パケットのサイズと送信間隔によっては、EEE による性能ペナルティは無視できないものがある。

なお、実験に用いたシステムでは、EEE を有効にした場合でも無効にした場合でも、パケットの送信間隔が 1.3 ミリ秒を超えると応答時間が急激に悪化するという現象が見られた。ただし、この応答時間の急激な悪化は、後述するように、M7100 でも同様の現象が観測されたことから、NIC が原因である可能性が高いと考えている。同様の現象が別の NIC でも観測されるかは、今後詳しく調査する。

図 8 に M7100 において同様の実験を行ったときの結果を示す。グラフの見方は図 7 と同様である。M7100 の場合は、PowerConnect と異なり、本実験の測定範囲においてはタイムアウトの閾値らしきものは観測されなかった。いずれの送信間隔においても、1 本のリンクの EEE を有効にした場合は 5~16 マイクロ秒の、2 本とも有効にした場合は 7~23 マイクロ秒の遅延の増加が見られた。

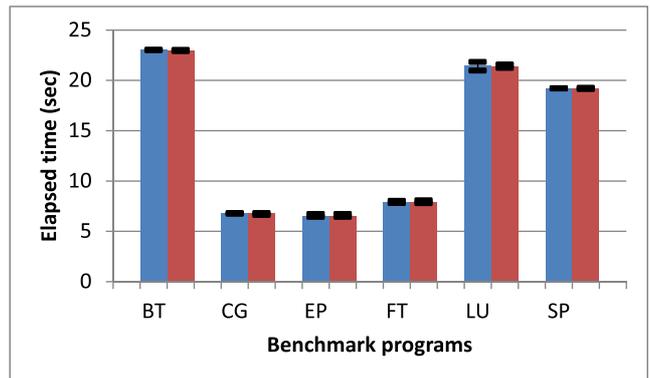


図 9 PowerConnect における NPB の実行時間（クラス B）  
 Fig. 9 Elapsed time of NPB (Class B) on PowerConnect.

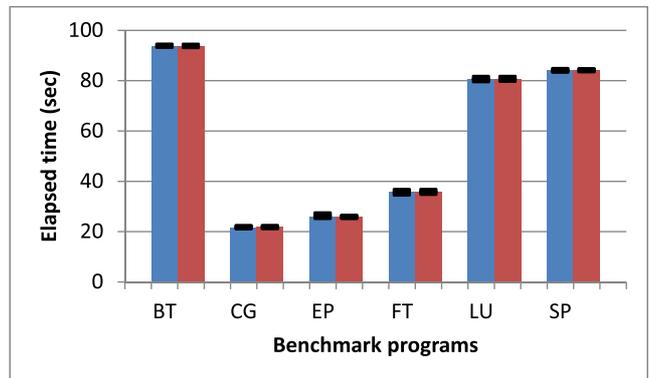


図 10 PowerConnect における NPB の実行時間（クラス C）  
 Fig. 10 Elapsed time of NPB (Class C) on PowerConnect.

#### 4.2 並列アプリケーションを用いた実験結果

続いて 3.3.1 項、および、3.3.2 項で説明した、並列アプリケーションを用いた評価実験の結果を示す。

##### 4.2.1 スイッチの比較実験結果

PowerConnect を使用したシステムにおける NPB の実行時間を図 9 および図 10 に示す。グラフの横軸はベンチマーク・プログラムを表し、縦軸は実行時間を表している。プログラムごとの 2 本の棒グラフは、左が全ポートにおいて EEE を無効化したときの実行時間を、右が全ポートにおいて EEE を有効にしたときの実行時間を表す。棒グラフは 5 回実行したときの平均値を表しており、箱ひげはそのときの最大値と最小値である。図 9 はクラス B、図 10 はクラス C に対する結果である。なお、スイッチの Tx のタイムアウト時間とウェイクアップ時間はデフォルト値とした。

グラフより、いずれのプログラム/入力においても、EEE を用いたときの性能ペナルティの影響は見られない。性能の悪化率が最も大きかったのはクラス C の FT を実行したときであるが、率にしてわずか 0.386% である。実行時間の増分としてはわずか 0.138 秒（EEE 無効時の実行時間が 35.726 秒に対して EEE 有効時の実行時間は 35.864 秒）にすぎなかった。

M7100 における NPB の実行時間を図 11 および図 12

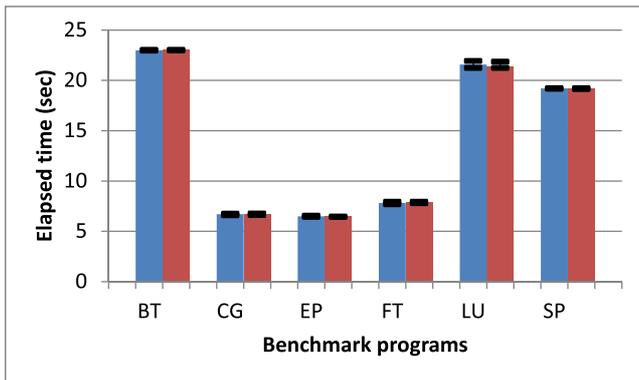


図 11 M7100 における NPB の実行時間 (クラス B)  
 Fig. 11 Elapsed time of NPB (Class B) on M7100.

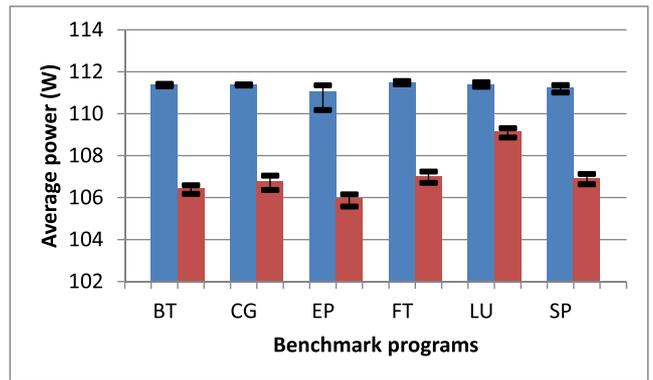


図 14 NPB 実行時の PowerConnect の平均消費電力 (クラス C)  
 Fig. 14 Average power consumption of PowerConnect for NPB (Class C).

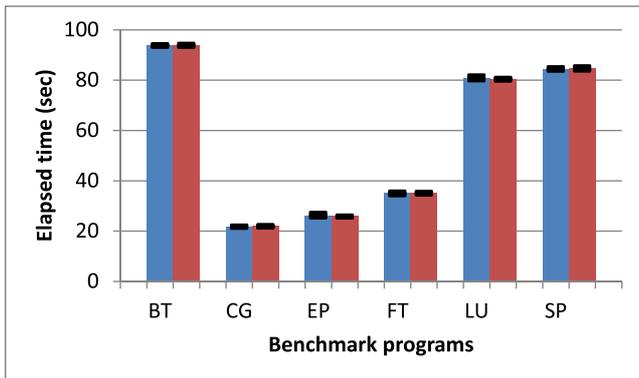


図 12 M7100 における NPB の実行時間 (クラス C)  
 Fig. 12 Elapsed time of NPB (Class C) on M7100.

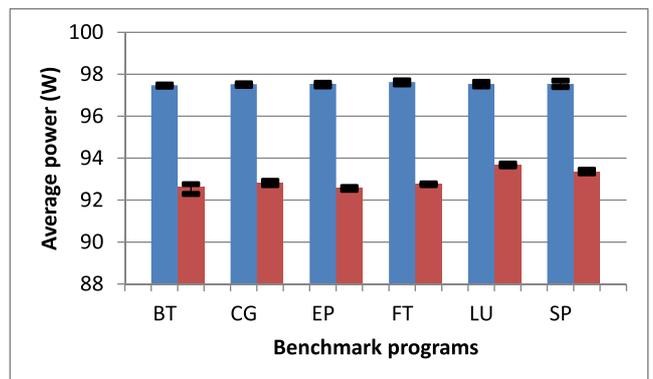


図 15 NPB 実行時の M7100 の平均消費電力 (クラス B)  
 Fig. 15 Average power consumption of M7100 for NPB (Class B).

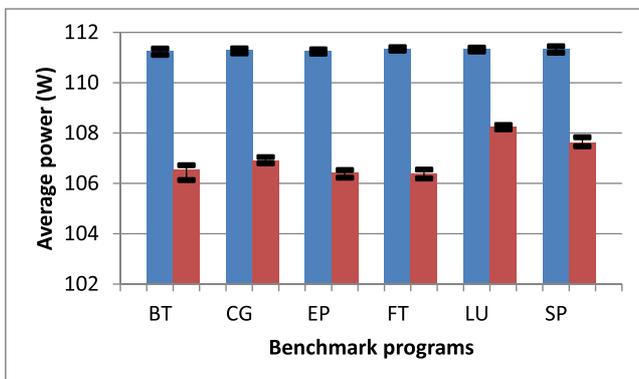


図 13 NPB 実行時の PowerConnect の平均消費電力 (クラス B)  
 Fig. 13 Average power consumption of PowerConnect for NPB (Class B).

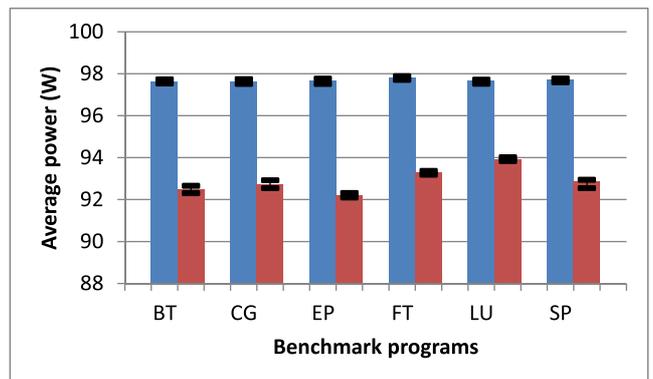


図 16 NPB 実行時の M7100 の平均消費電力 (クラス C)  
 Fig. 16 Average power consumption of M7100 for NPB (Class C).

に示す。グラフの見方は先の 2 枚のグラフと同様である。PowerConnect と同様、M7100 においても EEE の性能ペナルティの影響はほとんど見られない。最悪の場合 (クラス C の BT) でも、EEE を有効にしたときの性能劣化は 0.102 秒 (EEE 無効時の実行時間が 93.744 秒に対して EEE 有効時の実行時間は 93.846 秒) にすぎなかった (率に直すと 0.368%)。

これまで述べてきたように、EEE を HPC システムに用いた場合には、プログラムの実行中に PHY のウェイク

アップが発生することで性能が低下してしまうことが懸念された。しかし、今回の実験により、実用上はそのような性能低下はほとんど発生しないことが分かった。

図 13, 図 14, 図 15, 図 16 にかけて、NPB を実行した際の各スイッチの平均消費電力を示す。グラフの見方は、縦軸が消費電力に変わった点を除き、これまでと同様である。

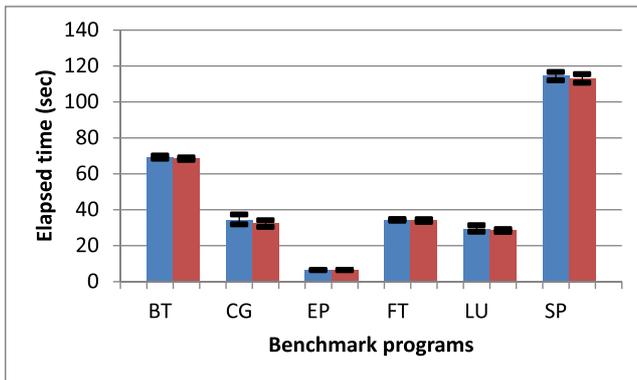


図 17 PowerConnect 5548 (1000BASE-T) における NPB の実行時間 (クラス B)

Fig. 17 Elapsed time of NPB (Class B) on PowerConnect 5548 (1000BASE-T).

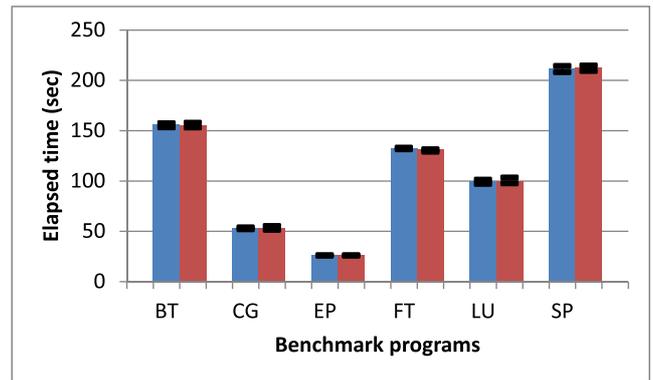


図 18 PowerConnect 5548 (1000BASE-T) における NPB の実行時間 (クラス C)

Fig. 18 Elapsed time of NPB (Class C) on PowerConnect 5548 (1000BASE-T).

グラフより、いずれのプログラム/入力においても、EEE を用いることでスイッチの消費電力を削減できていることが分かる。ただしその削減量はスイッチ、あるいは、プログラム/入力によって異なる。特に EP のような通信をほとんど行わないプログラムでは削減量が大きく、M7100 でクラス B の EP を実行したときは平均で 4.97 W、クラス C の EP を実行したときは平均で 5.44 W の電力が削減されていた。また、全プログラムの平均では、M7100 の場合は 4.67 W、PowerConnect の場合は 4.28 W の消費電力を削減していた。ポートあたりの電力削減量に直すと、M7100 の場合は平均 1.17 W、PowerConnect の場合は平均 1.07 W である。これは 4.1.1 項で示した電力削減量にかなり近い値であり、各 PHY はほとんどの期間省電力モードに滞在していたと考えられる。

図 14 の EEE 無効化時の EP の消費電力のみ、5 回測定時の分散が大きい。これは、5 回測定したうちの 1 回の実験において、Watts Up? .Net がサンプリングした 1 回分の観測値 (1 秒分の平均消費電力) が著しく (20 W 程度) 低かったことによる。本実験では Watts Up? .Net と測定用のホスト計算機とを USB ケーブルで接続し、Watts Up? .Net が測定した値を USB 経由でホスト計算機側で読み出しているが、この方法ではごく稀に電力を正しく計測できないことがある。その他のサンプリング値は他の実験時とほぼ同じ値 (11 W 前後) を示していることから、上記の異常な電力値は Watts UP? .Net のエラーが原因である。

スイッチを 4.1.1 項でも使用した 1000BASE-T のスイッチ (PowerConnect 5548) に交換し、NPB を実行したときの性能とスイッチの消費電力を測定した。結果を図 17、図 18、図 19、図 20 に示す。ただし、表 2 に示した 10GBASE-T の NIC は 1000BASE-T 接続時の EEE をサポートしていなかったため、別の EEE 対応の 1000BASE-T の NIC を用いて実験を行った。

まず、図 17 や図 18 を図 9 や図 10 と比べると、

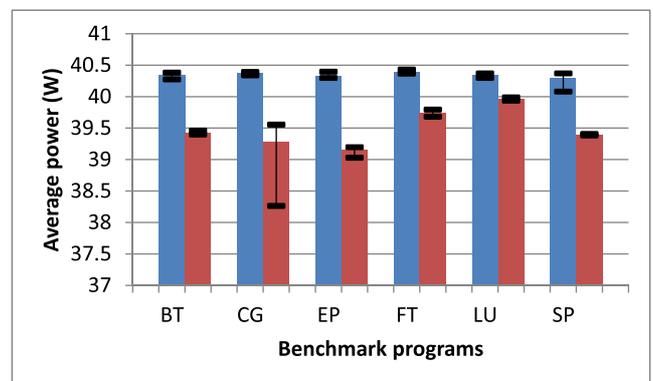


図 19 NPB 実行時の PowerConnect 5548 (1000BASE-T) の平均消費電力 (クラス B)

Fig. 19 Average power consumption of PowerConnect 5548 (1000BASE-T) for NPB (Class B).

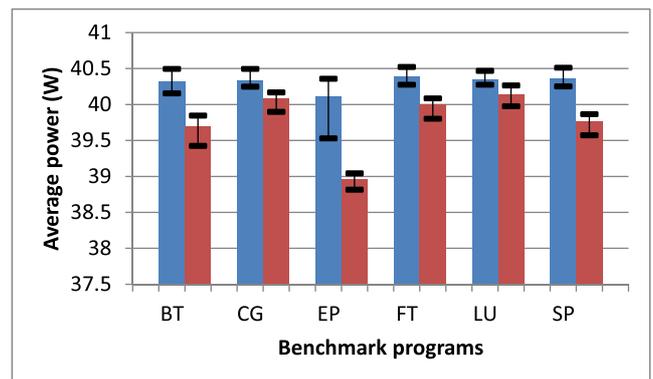


図 20 NPB 実行時の PowerConnect 5548 (1000BASE-T) の平均消費電力 (クラス C)

Fig. 20 Average power consumption of PowerConnect 5548 (1000BASE-T) for NPB (Class C).

10GBASE-T のネットワークの代わりに 1000BASE-T のネットワークを用いることで、EP を除くすべてのプログラムで実行時間が大幅に増加していることが分かる。特に、クラス B の SP の実行時に EEE を無効化した場合は、PowerConnect 8132 (10GBASE-T) では実行時間が 19.2

秒だったのに対し、PowerConnect 5548 (1000BASE-T) では実行時間が 114.6 秒であった。すなわち、1000BASE-T のネットワークを使用したときの方が 5.97 倍も遅い。このように、たとえ 4 ノードで NPB を実行する場合であっても、10GBASE-T のネットワークを利用する価値はある。

次に、図 17 および図 18 において EEE 有効時と無効時の性能を比較すると、10GBASE-T と同様、EEE による性能劣化はほとんどないことが分かる。最も性能が低下する場合（クラス C の SP）であっても、実行時間の増加率はわずか 0.422%（EEE 無効化時には 211.938 秒だったものが有効時には 212.832 秒）にすぎない。表 1 によれば 1000BASE-T のネットワークの方が 10GBASE-T のネットワークよりもウェイクアップ時間が 3~4 倍大きいですが、それでも EEE を用いることによる性能への悪影響はほとんど見られなかった。

図 19 および図 20 は、同スイッチを用いて NPB を実行したときのスイッチの平均消費電力である。グラフより、10GBASE-T のスイッチよりは消費電力削減量は低いものの、1000BASE-T のスイッチにおいても EEE を用いることで消費電力が削減できることが分かる。EEE による消費電力の削減量は平均で 0.174 W/ポートであった。なお、図 19 と図 20 中のいくつかのグラフ（たとえばクラス B の CG を EEE 有効にして実行した場合）は 5 回実行時の分散が大きいですが、これは前述のように Watts Up? .NET のエラーによるものである。

当初の予想に反し、実 HPC 環境では EEE の性能ペナルティが問題となるケースはほとんど観測できなかった。これは、以下で述べるように、リンクの使用間隔がタイムアウト時間を超えるケースが少ないためである。

図 21 は、PowerConnect を使用してクラス C の NPB を実行したときの、リンクのアイドル時間の分布の累積である。グラフの横軸はアイドル時間を表しており、縦軸はその時間以上のアイドル時間の出現回数を累積した値を表

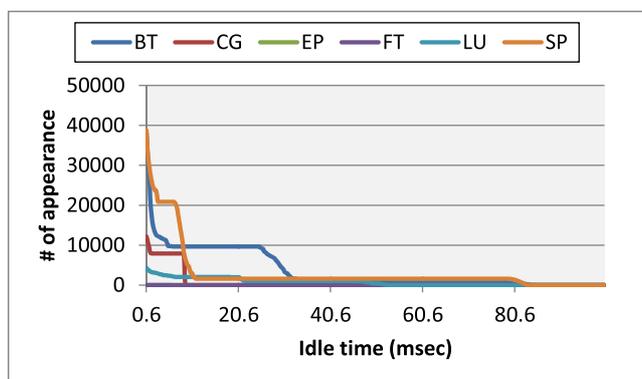


図 21 PowerConnect で NPB (クラス C) を実行したときのリンク・アイドル時間の累積分布

Fig. 21 Cumulative distribution of link idle time for NPB (Class C) on PowerConnect.

している。各リンクのアイドル時間は、MPE を用いてアプリケーション実行時の MPI 通信イベント・ログを取得し、それを次のようにして解析することで求めた。あるノードが接続されたリンクは、そのノードで発生した直前の通信イベントの終了から次の通信イベントの開始までアイドル状態であったと見なす。そして、それ以外の時間は、そのリンクはアクティブ状態であったと見なすものとする。

この仮定に基づくリンク・アイドル時間の見積りは概算値であることに注意されたい。なぜなら、通信イベントの間中ずっとリンクが使用され続けるわけではないからである。さらに、通信イベントの中にはノード内の他のプロセスとの通信イベントも含まれており、通信イベント中であってもリンクが使用されないこともある。そのため、実際のリンク・アイドル時間はこの解析結果よりも長い。したがって、以下の議論における省電力モードへの遷移回数は実際のそれよりも少ない見積り値となっているが、通信イベント中のリンクが使用されていない時間は通信そのものに要する時間と比べて通常は十分短いと考えられるため、両者の値の差はそれほど大きくはないと考えている。

グラフより、リンクが 600 マイクロ秒を超えてアイドル状態となるケースは、最も多い SP でも 38,770 回にすぎない。タイムアウト時間が 600 マイクロ秒のときはこれらのケースで省電力モードへ移行することから、ウェイクアップ時間を 17 マイクロ秒とすると、省電力モードからの復帰による総ペナルティ時間はたかだか 0.659 秒である。一方、図 10 より、クラス C の SP の実行時間は 84.1 秒であるから、実行時間に対する総ペナルティ時間の割合はわずか 0.783% である。

実際の性能ペナルティはこれよりも少ないことに注意されたい。なぜなら、複数リンクのウェイクアップがオーバーラップする、リンクのウェイクアップと CPU による計算とがオーバーラップする、などの理由により、一部の性能ペナルティは隠蔽されるからである。

このように、単位時間あたりの省電力モードへの遷移回数が十分少ないことから、ウェイクアップ時間が性能に与える影響はほとんどなかったものと考えられる。

省電力モードへの遷移回数が少ないにもかかわらず EEE により十分な電力を削減できるのは、HPC アプリケーションが一般には十分な長さの計算フェーズを有しており、その間のリンクの消費電力を削減できるからである。

図 22 は、PowerConnect を使用してクラス C の NPB を実行したときの、リンクのアイドル時間（ミリ秒単位）の CDF (Cumulative Distribution Function) である。グラフの横軸はアイドル時間を表しており、縦軸はそのアイドル時間が「全実行時間 × 総リンク数」に占める割合の積算値である。たとえば、CG はアイドル時間が 20 ミリ秒のときの積算値が 20.9% であるが、これはリンクが 20 ミリ秒以下のアイドル状態となるケースが「全実行時間 × 総リ

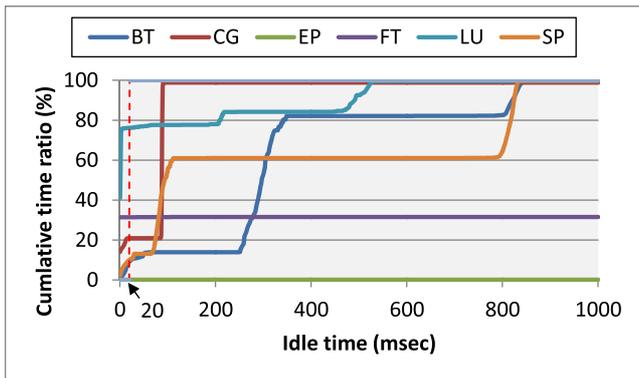


図 22 PowerConnect で NPB (クラス C) を実行したときのリンク・アイドル率

Fig. 22 Link idle ratio for NPB (Class C) on PowerConnect.

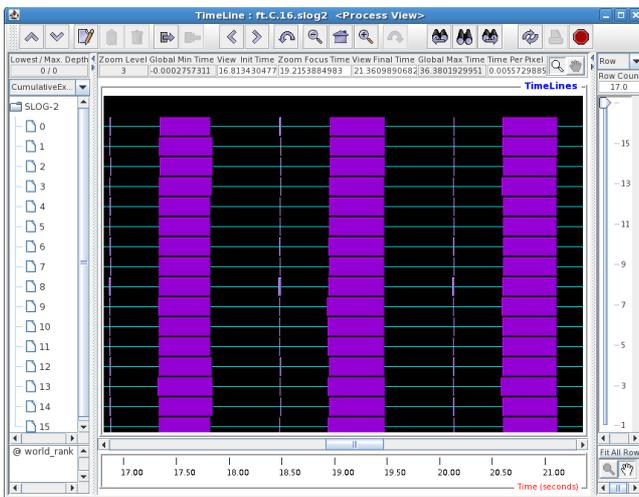


図 23 PowerConnect で FT (クラス C) を実行したときの通信の様子

Fig. 23 Communication pattern for FT (Class C) on PowerConnect.

「リンク数」の 20.9%を占めていたことを表している。なお、MPE による通信イベントの時間トレースはミリ秒単位の解像度のため、グラフ中では 1 ミリ秒未満の通信間隔はアイドル時間を 0 としてカウントしてある。

グラフより、多くのプログラムでリンクが 20 ミリ秒以上アイドル状態となるケースが多くを占めていることが分かる。リンクが 20 ミリ秒以上アイドル状態となるケースは、SP の場合は「全実行時間 × 総リンク数」の 89.8%、BT の場合は 90.7%であった。このように、NPB の実行中は、リンクが長時間アイドル状態となるケースが実行時間の多くを占めている。この間にリンクにおいて消費される電力の大半を削減できるため、EEE による消費電力削減効果は高い。

図 23 に FT の通信パターンを示す。表示には Jumpshot [9] を用いた。図より、FT は計算フェーズ (各プロセスの黒色の部分) と通信フェーズ (紫色の部分) を交互に繰り返しており、計算フェーズの時間は 500 ミリ秒を超え

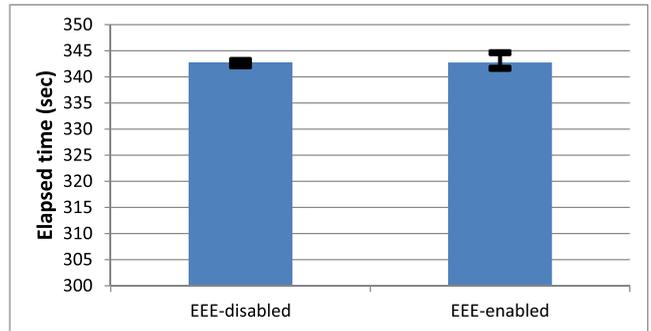


図 24 ALPS/looper の実行時間

Fig. 24 Elapsed time of ALPS/looper.

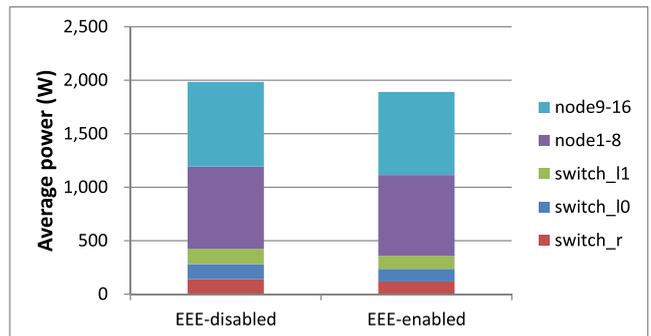


図 25 ALPS/looper を実行したときの平均消費電力の内訳

Fig. 25 Power breakdown of the evaluation system for ALPS/looper.

ていることが分かる。このような計算フェーズにおいてリンクを省電力モードにできるため、EEE は効果がある。

#### 4.2.2 実アプリを使用した全系の評価実験結果

4.1.2 項で述べたように、PowerConnect のポートあたりの電力削減量は最大で 1.35 W 程度である。したがって、ALPS 実行中のほとんどの期間においてほとんどの PHY が省電力モードで動作していたことになる。

図 2 の全系を使用し、ALPS/looper を実行したときの実行時間を図 24 に示す。グラフの横軸は EEE の使用の有無を表し、縦軸は実行時間を表している。棒グラフは 3 回の実行における平均実行時間を表しており、箱ひげの上端は 3 回の実行における最長実行時間を、下端は最短実行時間を表す。

グラフより、前項の実験結果と同様、全系を使用して ALPS/looper を実行した場合もまた EEE による性能低下は見られない。EEE 無効化時の平均実行時間は 342.79 秒だったのに対し、EEE 有効化時の平均実行時間は 342.76 秒であった。

ALPS/looper を実行した際の評価システムの平均消費電力の内訳を図 25 に示す。各棒グラフは 5 色に塗り分けられており、下から順に、switch\_r, switch\_l0, switch\_l1, ノード 1~8, ノード 9~16 の平均消費電力を表している。なお、グラフの値は 3 回の実行における平均値である。

グラフより、EEE を使用することによってシステム全

体で平均 93.7 W の消費電力を削減できた。この値はシステム全体の平均消費電力の 4.72% に相当する。このことは、図 24 に示したように実行時間の増加はほぼないことから、EEE を使用することでシステムの消費エネルギーが 4.72% 削減できることを意味する。

上記の電力削減量の大部分は 3 つのスイッチの消費電力削減によるものである。3 つのスイッチにおいては、1 台あたり 21.3~21.9 W の平均消費電力の削減効果があった。本実験では各スイッチは 16 ポートを使用していることから、ポートあたりの電力削減量は 1.33~1.37 W である。

このように、EEE は並列プログラムの性能を落とすことなく PHY の消費電力を削減できる。

### 5. スケーラビリティに関する考察

本稿では、最大 16 ノード、2 階層の fat-tree というシステム構成で、並列プログラム実行時に EEE が及ぼす影響を評価した。評価システムがこのように小規模であるのは、多ポートの 10 ギガビット・イーサネット・スイッチは高額であり、それを多数揃えた大規模環境での実験が現実的には難しいことによる。本章では、FX-10 上で NPB を実行したときの通信パターンの分析結果を交えながら、大規模環境で EEE を使用した場合についての考察を行う。

これまで見てきたように、EEE が本実験環境で効果があったのは、1) 単位時間あたりの省電力モードへの遷移回数が十分に少なく、かつ、2) 計算フェーズのような長時間リンクがアイドル状態となるケースが実行時間の多くを占めているからである。大規模環境でもこれらが成り立つのであれば、EEE は有効に機能すると考えられる。

図 26 は、64 ノードの FX-10 上でクラス D の NPB を実行したときのリンク・アイドル時間の累積分布である。グラフより、リンクが 20 ミリ秒以上アイドル状態となったケースの総数は、最も多い BT でも 817,131 回であった。Tofu の PHY のウェイクアップ時間が 10GBASE-T のそ

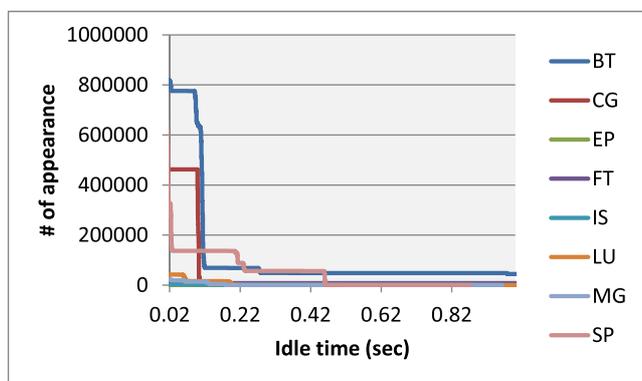


図 26 FX-10 64 ノード上で NPB (クラス D) を実行したときのリンク・アイドル時間の累積分布  
Fig. 26 Cumulative distribution of link idle time for NPB (Class D) on 64-node FX-10.

れと同じだと仮定すると、タイムアウト時間が 20 ミリ秒のときの総ペナルティ時間は 13.9 秒である。一方、クラス D の BT の FX-10 上での実行時間は 741.2 秒であったことから、実行時間に対する総ペナルティ時間の割合は 1.88% である。

図 27 にリンク・アイドル時間の CDF を示す [29]。リンクが 20 ミリ秒以上アイドル状態となるケースが「全実行時間 × 総リンク数」に占める割合は、いずれのプログラムにおいても多い。BT ではその割合は 97.0% に達する。したがって、タイムアウト時間が 20 ミリ秒であっても十分な省電力効果が期待できる。

また、後述するように、タイムアウト時間を適切に選べば 64 ノードの環境でも EEE が有効であることを示したシミュレーション結果もある [22]。このように、タイムアウト時間の再調整は必要かもしれないが、大規模環境でも EEE は有効と考えられる。

### 6. EEE が有効な範囲

HPC システムのインタコネクには、光インタフェースや InfiniBand など、イーサネットよりも高速なネットワークが使用されることもある [22]。これらのインタコネクが EEE と同様の仕組みを採用した場合、PHY のウェイクアップ時間は本実験で使用した 10GBASE-T のそれとは異なると考えられる。ウェイクアップ時間が短くなるのであればよいが、長くなるようであれば EEE による性能劣化が無視できなくなる可能性がある。EEE による性能ペナルティを軽減するためにタイムアウト時間をもっと長くすることも考えられるが、その場合は EEE による消費電力削減量が減少してしまう。

3 章で述べたように、PowerConnect は Tx のタイムアウト時間とウェイクアップ時間を管理コンソールから変更できる。そこで、ウェイクアップ時間とタイムアウト時間を変化させながら性能と消費電力を測定し、どの程度のウェ

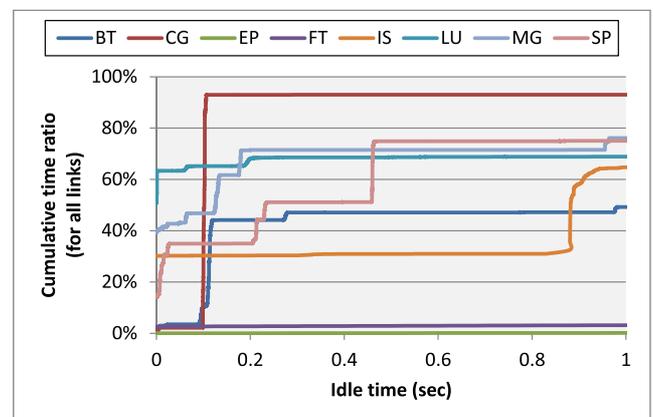


図 27 FX-10 64 ノード上で NPB (クラス D) を実行したときのリンク・アイドル率  
Fig. 27 Link idle ratio for NPB (Class D) on 64-node FX-10.

イクアップ時間とタイムアウト時間まで許容できるかを明らかにする。

本実験の目的は、10GBASE-T 以外のネットワークが EEE のような規格を制定する場合の、PHY のウェイクアップ時間とタイムアウト時間の目標値を与えることにある。これらの PHY ではまだ EEE のような規格が存在しないため、そのウェイクアップ時間がどの程度になるかは定かでない。他のネットワークが EEE を採用した場合の性能/電力への影響を正確に見積もることは困難であることから、代わりに 10GBASE-T のネットワークにおける PHY のウェイクアップ時間とタイムアウト時間の許容値を求めることで、他のネットワークが EEE のような規格を定める際の指針を与えることにした。

図 28 は、タイムアウト時間を 600 マイクロ秒、1,600 マイクロ秒、16,000 マイクロ秒、160,000 マイクロ秒と変化させたときの各プログラムの実行時間である。横軸はタイムアウト時間、縦軸は実行時間である。ただし、各プログラムの実行時間は、600 マイクロ秒のときのそれで正規化してある。実験には図 10 などと同様のシステムを使用した。プログラムはクラス C の NPB とした。なお、ウェイクアップ時間はデフォルトの値 (17 マイクロ秒) とした。

グラフより、測定した範囲内では、FT を除いてタイムアウト時間の違いによる性能への影響はほとんど見られなかった。FT 以外のいずれのプログラムも、実行時間の差は 0.5% 以内に収まっている。一方、FT に関しては、タイムアウト時間を 600 マイクロ秒から 1,600 マイクロ秒にすることで性能が 2% 程向上した。これは、図 22 に示したように、FT はリンクを 1 ミリ秒よりも短い間隔で使用することが多いためである。タイムアウト時間を 1,600 マイクロ秒にすることで省電力モードに遷移する回数が減ったため、実行時間が短縮されたものと考えられる。

図 29 は、タイムアウト時間に対するスイッチの消費電力削減量である。各消費電力削減量は、タイムアウト時間

が 600 マイクロ秒のときのそれで正規化してある。

グラフより、タイムアウト時間を 1,600 マイクロ秒にまで長くしても、スイッチの消費電力削減量はほとんど変わらない。1,600 マイクロ秒のときの消費電力削減量は、600 マイクロ秒のときのそれに比べて、最悪 (SP) でも 3.1% 減少するだけである。そのため、1,600 マイクロ秒程度のタイムアウト時間は許容範囲内といえる。これは、PowerConnect のデフォルトのタイムアウト時間の 2.67 倍の値である。

一方、1,600 マイクロ秒よりもタイムアウト時間が長くなるにつれて、スイッチの消費電力削減量は徐々に減少する。スイッチの消費電力削減量は、タイムアウト時間が 16,000 マイクロ秒のときで最大 41.8% (CG), 160,000 マイクロ秒のときで最大 56.5% (SP) 低下した。このことから、タイムアウト時間が 16,000 マイクロ秒を超えると、PHY の消費電力を削減する機会を大幅に奪ってしまうことが分かる。

続いて、タイムアウト時間をデフォルトの値 (600 マイクロ秒) にした状態で、ウェイクアップ時間を 17 マイクロ秒、170 マイクロ秒、850 マイクロ秒、1,700 マイクロ秒

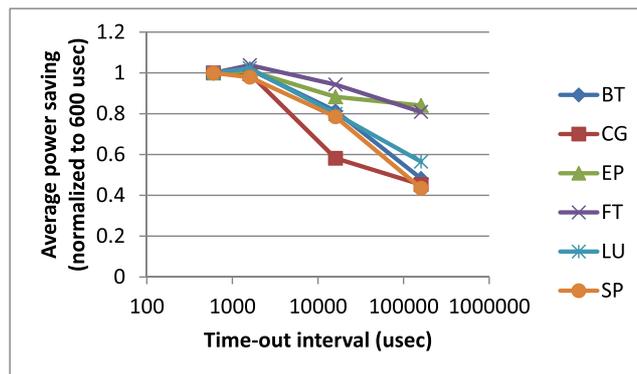


図 29 タイムアウト時間に対する消費電力削減量のセンシティブティ (PowerConnect, クラス C)

Fig. 29 Power-saving sensitivity for time-out intervals (PowerConnect, Class C).

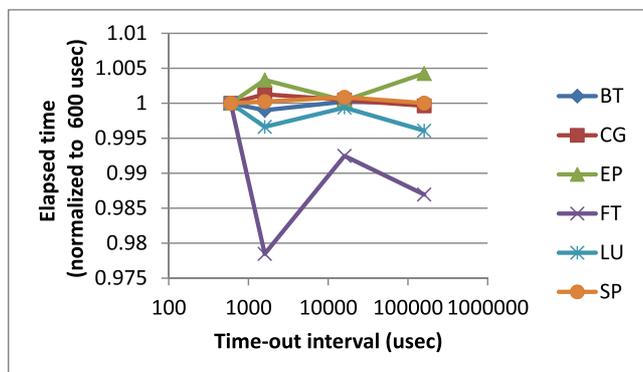


図 28 タイムアウト時間に対する性能のセンシティブティ (PowerConnect, クラス C)

Fig. 28 Performance sensitivity for time-out intervals (PowerConnect, Class C).

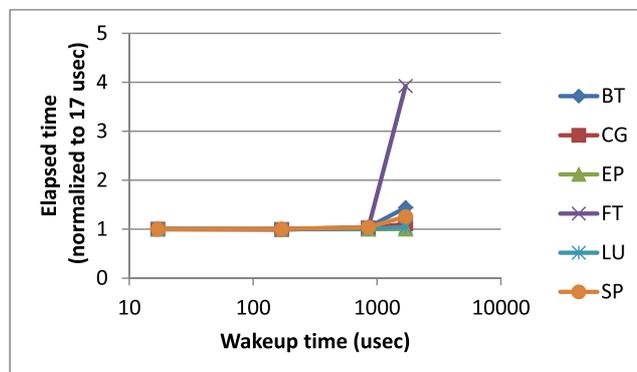


図 30 ウェイクアップ時間に対する性能のセンシティブティ (PowerConnect, クラス C)

Fig. 30 Performance sensitivity for wakeup time (PowerConnect, Class C).

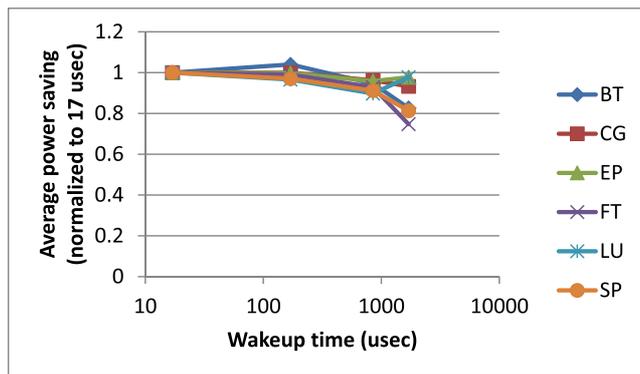


図 31 ウェイクアップ時間に対する消費電力削減量のセンシティブティ (PowerConnect, クラス C)

Fig. 31 Power-saving sensitivity for wakeup time (PowerConnect, Class C).

と変えたときの各プログラムの実行時間を図 30 に示す。グラフの横軸はウェイクアップ時間、縦軸は実行時間である。ただし、各実行時間はウェイクアップ時間が 17 マイクロ秒のときのそれで正規化してある。

グラフより、ウェイクアップ時間が 10 倍 (170 マイクロ秒) になっても性能はほとんど変わらない。ウェイクアップ時間が 17 マイクロ秒のときとの性能差は、いずれのプログラムも 1%以内であった。また、ウェイクアップ時間を 50 倍 (850 マイクロ秒) にした場合でも、ウェイクアップ時間が 17 マイクロ秒のときに対する性能低下は最大で 4.1% (SP) であり、十分許容の範囲内といえる。

一方、ウェイクアップ時間を 100 倍 (1,700 マイクロ秒) にした場合は、ほとんどのプログラムで大幅に性能が低下した。ウェイクアップ時間が 17 マイクロ秒のときに対する実行時間の増加率は、最大で 3.92 倍 (FT) であった。

図 31 はウェイクアップ時間を変えたときのスイッチの消費電力削減量である。グラフより、スイッチの消費電力削減量は、ウェイクアップ時間を増やしても 850 マイクロ秒まではあまり変わらないことが分かる。

図 32 および図 33 は、ウェイクアップ時間を 850 マイクロ秒にした状態でタイムアウト時間を変化させたときの性能と消費電力削減量のグラフである。どちらのグラフも、ウェイクアップ時間が 17 マイクロ秒、かつ、タイムアウト時間が 600 マイクロ秒のときの性能/消費電力削減量で正規化してある。

図 32 より、タイムアウト時間を 600 マイクロ秒から 1,600 マイクロ秒にすることで、いくつかのプログラムは EEE による性能低下を抑制できている。たとえば、BT では、タイムアウト時間が 600 マイクロ秒のときは 3.78%性能が悪化していたのに対し、タイムアウト時間を 1,600 マイクロ秒にすることで 0.80%の性能低下に抑えることができた。一方、スイッチの消費電力削減量は、タイムアウト時間が 600 マイクロ秒のときと 1,600 マイクロ秒のときと

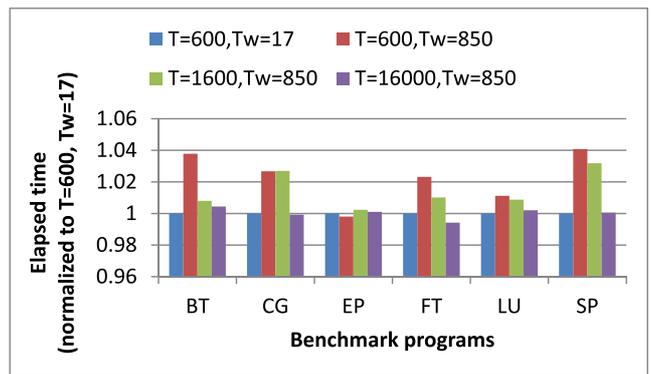


図 32 ウェイクアップ時間を 850 マイクロ秒にしてタイムアウト時間を変化させたときの性能 (PowerConnect, クラス C)

Fig. 32 Performance under the fixed wakeup time (850 microseconds) for various time-out intervals (PowerConnect, Class C).

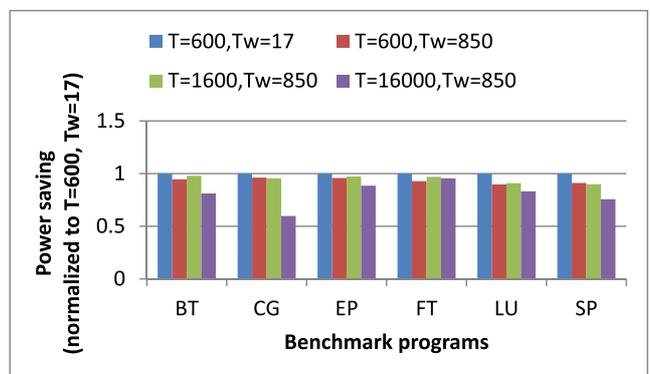


図 33 ウェイクアップ時間を 850 マイクロ秒にしてタイムアウト時間を変化させたときの消費電力削減量 (PowerConnect, クラス C)

Fig. 33 Power-saving under the fixed wakeup time (850 microseconds) for various time-out intervals (PowerConnect, Class C).

でほとんど差がない (図 33)。そのため、ウェイクアップ時間が 850 マイクロ秒 (デフォルト値の 50 倍)、タイムアウト時間が 1,600 マイクロ秒 (デフォルト値の 2.67 倍) までは許容の範囲内といえる。

InfiniBand などの高速ネットワークと 10GBASE-T のネットワークとでは、通信レイテンシが 1 桁以上異なる。そのため、10GBASE-T のネットワークを用いて求めたウェイクアップ時間とタイムアウト時間の許容値が、そのまま他の高速ネットワークにもあてはまるわけではない。通信レイテンシが小さくなるということは、CPU の通信待ち時間が減少するということである。これによりリンクのアイドル時間は短縮される方向に働くため、同じタイムアウト時間のもとでは、10GBASE-T のネットワークよりも高速ネットワークの方が省電力モードに遷移する機会が少なくなる。したがって、高速ネットワークにおいて 10GBASE-T と同様の制御を行った場合は、上述の見積り結果よりも性能ペナルティ/消費電力削減量ともに減少す

ることになることに注意されたい。

## 7. Saravanan らの実験との相違点

本稿の冒頭で述べたように、Saravanan らによって EEE を HPC 環境で使用した場合のシミュレーションが行われている [22]。彼らは、MareNostrum 上で取得したアプリケーションのトレース (MPI 通信イベントの詳細情報とイベント間で消費された CPU 時間) と、大規模クラスター・シミュレータである Dimemas [5] を用いて、HPC 環境において EEE を用いた場合の電力と性能を評価した。その結果、EEE によって 25% の性能低下が見られたことを報告している。また、彼らは Power-Down Threshold (後述するように、これはタイムアウト時間と本質的に同じである) と呼ぶ方法により、この性能オーバーヘッドを 2% にまで軽減できることを報告している。

このように、彼らは我々とは異なり、EEE による性能ペナルティを問題視しているが、これは彼らが実際の EEE 対応機器の性質をふまえずに議論を展開しているためである。彼らは、IEEE802.3az の仕様から、EEE 対応のポートが通信を終えるとただちに省電力モードへ移行すると仮定し、評価を行っている。パケットの送受信のたびに省電力モードへと移行する制御を行った結果として、大幅な性能低下が見られたと報告している。

しかし、これまで見てきたように、商用の EEE 対応機器はそのような制御を行っていない。最後の通信から一定時間経過した後に省電力モードへ移行するというタイムアウト制御を行っている。これは、たとえインターネット用途であっても、パケット送受信のたびに省電力モードへと移行していたのでは、PHY のウェイクアップによる性能低下が無視できなかつたためと考えられる。十分なタイムアウト時間を設けることによって、ポートが頻繁に使用されているときに省電力モードへと移行するのを防ぎつつ、ポートが長時間使用されないときに省電力モードへ移行することを可能にしている。

Saravanan らが文献 [22] で提案した Power-Down Threshold とは、このタイムアウト時間のことである。各リンクがパケットの転送を終えた直後に省電力モードへ移行するのではなく、一定の猶予期間を設けることで、同文献では性能低下を抑えつつポートの消費電力を大幅に削減できることを示した。しかし、上で述べたように、商用の EEE 対応機器はすでにタイムアウト制御を行っている。その結果、Saravanan らの実験において Power-Down Threshold を適用した場合とほぼ同様の結論が本実験では得られた。

本実験を通して初めて得られた知見として、EEE によるポートの消費電力削減量は、Saravanan らの想定よりもかなり小さいことがあげられる。4 章で述べたように、EEE によるポートの消費電力削減率は、10GBASE-T の場合で

およそ 4 割、1000BASE-T の場合で約 75% である。残りの部分の消費電力が本質的に削減困難であるか否かは不明であるが、この削減率のまま EEE を HPC 環境に応用しても、システム全体の消費電力を考慮すれば、EEE による省電力効果が小さく見えてしまう恐れがある。今後はこの削減率を改善していくことが望まれる。

また、本実験により、バックプレーンが消費する電力が予想外に大きいことも分かった。本実験で使用したスイッチはバックプレーンだけで 100 W 近い電力を消費していた。4 章で述べたように、ポートの消費電力は 3 W 強 (EEE を使用しない場合) であり、24 ポートのスイッチすべてを使用したとしても、リンクの総消費電力は 72 W 強である。今後は、ポートの消費電力だけでなく、バックプレーンの消費電力を削減する方法も考えていく必要があるだろう。たとえば、スイッチのスループットがそれほど必要ないときにバックプレーンに位置する CPU やメモリを DVFS することが考えられる。

## 8. まとめ

本稿では、将来のスーパーコンピュータにおける有力な省電力技術の 1 つである EEE に着目し、それが電力と性能に与える影響を実システムを用いて評価した。評価の結果、EEE は、10GBASE-T を使用する実環境においても、性能をまったくといっていいほど低下させることなく、並列アプリケーション実行中のネットワークの消費電力を削減できることが分かった。

謝辞 本研究の一部は文部科学省「将来の HPCI システムのあり方の調査研究」事業における研究課題「レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究」(アーキテクチャ評価およびコンパイラ技術と省電力機構)、および、JST CREST における研究領域「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」の研究課題「ポストペタスケールシステムのための電力マネジメントフレームワークの開発」による。

## 参考文献

- [1] Ajima, Y., Sumimoto, S. and Shimizu, T.: TOFU: A 6D Mesh/Torus Interconnect for Exascale Computers, *IEEE Computer*, Vol.42, No.11, pp.36-40 (2009).
- [2] ALPS project, available from ([http://alps.comp-phys.org/mediawiki/index.php/Main\\_Page](http://alps.comp-phys.org/mediawiki/index.php/Main_Page)).
- [3] ALPS project, available from (<http://exa.phys.s.u-tokyo.ac.jp/ja/projects/alps-looper>).
- [4] Asato, A.: Extracting the Performance of Multi-core Processor in Supercomputer, *The 12th International Forum on Embedded MPPSoC and Multicore (keynote)* (2012).
- [5] Badia, R.M., Labarta, J., Giménez, J. and Ascalé, F.: DIMEMAS: Predicting MPI applications behavior in Grid environments, *Workshop on Grid Applications and*

*Programming Tools (GGF8)* (2003).

[6] Bailey, D.H.: The NAS Parallel Benchmarks, *RNR-94-007* (1994).

[7] Bannet, M.J.: Energy-Efficient Ethernet for 100G Backplane and Copper, *IEEE P802.3bj task force* (2011).

[8] Bit-Twist, available from <http://bittwist.sourceforge.net/>.

[9] Chan, A., Gropp, W. and Lusk, E.: An Efficient Format for Nearly Constant-Time Access to Arbitrary Time Intervals in Large Trace Files, *Scientific Programming*, Vol.16, No.2-3, pp.155-165 (2008).

[10] D-Link: DGS-1100 EasySmart Switches 16/24 Port Gigabit Switches (Datasheet) (2011).

[11] Fibre Channel Industry Association (2010).

[12] Gustlin, M.: 100Gb/s Ethernet and EEE, *IEEE P802.3bj task force* (2011).

[13] Hewlett-Packard: HP E8200 z1 v2 Switch Series (Technical Specifications) (2011).

[14] IEEE: IEEE Std 802.3az-2010 (IEEE Standards) (2010).

[15] Kogge, P.M.: Architectural Challenges at the Exascale Frontier, *Simulating the Future: Using One Million Cores and Beyond (invited talk)* (2008).

[16] Li, K., Kumpf, R., Horton, P. and Anderson, T.: A Quantitative Analysis of Disk Drive Power Management in Portable Computers, *Proc. 1994 Winter USENIX Conference*, pp.279-291 (1994).

[17] Miwa, S., Aita, S. and Nakamura, H.: Performance Estimation for High Performance Computing Systems with Energy Efficient Ethernet Technology, *Proc. International Conference on Energy-Aware High Performance Computing* (2013).

[18] Okano, H., Kawabe, Y., Kan, R., Yoshida, T., Yamazaki, I., Sakurai, H., Hondou, M., Matsui, N., Yamashita, H., Nakada, T., Maruyama, T. and Asakawa, T.: Fine Grained Power Analysis and Low-Power Techniques of a 128GFLOPS/58W SPARC64 VIIIfx Processor for Petascale Computing, *Proc. 2010 IEEE Symposium on VLSI Circuits*, pp.167-168 (2010).

[19] One, L.: GEU-0820 8-Port Gigabit Switch (Datasheet) (2011).

[20] Reviriego, P., Christensen, K., Rabanillo, J. and Maestro, J.A.: An Initial Evaluation of Energy Efficient Ethernet, *IEEE Communication Letters*, Vol.15, No.5, pp.578-580 (2012).

[21] Reviriego, P., Sivaraman, V., Zhao, Z., Maestro, J.A., Vishwanath, A., Sánchez-Macian, A. and Russell, C.: An Energy Consumption Model for Energy Efficient Ethernet Switches, *Proc. 2012 International Conference on High Performance Computing and Simulation*, pp.98-104 (2012).

[22] Saravanan, K.P., Carpenter, P.M. and Ramirez, A.: Power/Performance Evaluation of Energy Efficient Ethernet (EEE) for High Performance Computing, *Proc. 2013 IEEE International Symposium on Performance Analysis of Systems and Software*, pp.205-214 (2013).

[23] Shye, A., Scholbrock, B. and Memik, G.: Into the Wild: Studying Real User Activity Patterns to Guide Power Optimizations for Mobile Architectures, *Proc. 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp.168-178 (2009).

[24] TOP500, available from <http://www.top500.org/>.

[25] Trendnet: 8-Port Gigabit GREENnet Switch (Datasheet) (2011).

[26] U. S. Department of Energy: *Final Minutes Advanced Scientific Computing Advisory Committee* (2012).

[27] Watts up?, available from <https://www.wattsupmeters.com/secure/products.php?pn=0&wai=0&more=2>.

[28] Wireshark, available from <http://www.wireshark.org/>.

[29] 三輪 忍, 會田 翔, 安島雄一郎, 清水俊幸, 安里 彰, 中村 宏: FX10におけるインタコネク・コントローラの省電力化手法の初期検討, 技術報告 5, 情報処理学会研究報告 2012-HPC-137 (2012).

[30] 片桐孝洋, 大島聡史, 中島研吾, 米村 崇, 熊洞宏樹, 樋口清隆, 橋本昌人, 高山恒一, 藤堂眞治, 岩田潤一, 内田和之, 佐藤正樹, 羽角博康, 黒木聖夫: レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究のためのアプリケーションと性能評価, 技術報告 2, 情報処理学会研究報告 2012-HPC-137 (2012).

[31] 片桐孝洋, 大島聡史, 中島研吾, 米村 崇, 熊洞宏樹, 樋口清隆, 橋本昌人, 高山恒一, 藤堂眞治, 岩田潤一, 内田和之, 佐藤正樹, 羽角博康, 黒木聖夫: レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究のためのアプリケーション最適化と異機種計算機環境での性能評価, 技術報告 4, 情報処理学会研究報告 2012-HPC-139 (2013).

[32] 片桐孝洋, 大島聡史, 中島研吾, 米村 崇, 熊洞宏樹, 樋口清隆, 橋本昌人, 高山恒一, 藤堂眞治, 岩田潤一, 内田和之, 佐藤正樹, 羽角博康, 黒木聖夫, 安達 斉, 江口義之: レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究のアプリケーションの異機種環境での評価—メニーコア環境を中心に, 技術報告 26, 情報処理学会研究報告 2012-HPC-143 (2014).



三輪 忍 (正会員)

1977年生。2005年京都大学大学院情報学研究所通信情報システム専攻博士後期課程単位認定退学。情報学博士。京都大学大学院法学研究科助手、東京農工大学工学府特任助教を経て、現在、東京大学大学院情報理工学系研究科助教。コンピュータ・アーキテクチャ、高性能計算機システム、組み込みシステム等の研究に従事。組み込みシステムシンポジウム 2010 優秀論文賞受賞。電子情報通信学会、IEEE 各会員。



會田 翔 (学生会員)

2012年東京大学工学部計数工学科卒業。2014年同大学大学院情報理工学系研究科修士課程修了。現在、(株)バンダイナムコゲームスに勤務。高性能計算機システムの研究に従事。



安島 雄一郎 (正会員)

1997年東京大学工学部電気工学科卒業。2002年同大学大学院工学系研究科情報工学専攻博士課程修了。博士(工学)。同年株式会社富士通研究所に入社。計算機アーキテクチャ、インタコネクトアーキテクチャの研究開発に

従事。2007年より富士通株式会社次世代テクニカルコンピューティング開発本部所属。スーパーコンピュータ「京」のTofuインタコネクトを開発。IEEE, ACM各会員。



清水 俊幸 (正会員)

1964年生。1988年東京工業大学大学院理工学研究科情報工学専攻修士修了。同年(株)富士通研究所入社。並列計算機アーキテクチャの研究に従事。現在、スーパーコンピュータの研究・開発に従事。電子情報通信学会

会員。



安里 彰 (正会員)

1983年東京大学理学部卒業、同年(株)富士通研究所入社。主として計算機アーキテクチャの研究開発に従事。現在、富士通(株)にてスパコン用CPUの開発を担当。2011年より電子情報通信学会コンピュータシステム研究会

副委員長。



中村 宏 (正会員)

1985年東京大学工学部電子工学科卒業。1990年同大学大学院工学系研究科電気工学専攻博士課程修了。工学博士。同年筑波大学電子・情報工学系助手。同講師、同助教授を経て、1996年東京大学先端科学技術研究センター

助教授。2010年東京大学大学院情報理工学系研究科教授。2014年より東京大学情報基盤センター長を兼務。この間1996~1997年カリフォルニア大学アーバイン校客員助教授。ハイパフォーマンスコンピューティング、省電力コンピューティング、高性能・低消費電力VLSIシステムの研究に従事。情報処理学会より論文賞(平成5年度,平成24年度),山下記念研究賞(平成6年度),坂井記念特別賞(平成13年度)各受賞。IEEE, ACM senior member。