社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－HPC－108 (12)
2006／10／6

MegaProto/Eにおける電力性能評価および電力性能最適化の検討

今田貴之[†] 佐藤三久[†] 堀田義彦[†]
木村英明[†] 朴泰祐[†]
高橋大介[†] 三浦信一[†]

我々は現実的な設置面積・容量と消費電力の条件下でより高い計算能力を得ることに焦点をあて，高性能でありながらも低消費電力・高密度実装を実現を目指す実証システムのプロトタイプとしてMegaProtoを開発している．先に開発したMegaProto/Cでは，プロセッサとしてCrusoeを用いていたが，MegaProto/Eは，Efficeon(TM8820)を用いて，それにメモリ，ネットワークに改良を加えたシステムである．本論文では，MegaProto/Eの基本性能と電力性能の評価を行い，MegaProto/Cとの比較評価を行った．その結果，MegaProto/EはMegaProto/Cと比較して最大約2倍の性能，および最大1.8倍の電力性能比であることが分かった．また，DVFS（Dynamic Voltage and Frequency Scaling）機能を用いて，電力性能最適化を試みたが，Efficeonでは電圧の削減幅が小さいなどの理由により，改善に至らなかった。

Power Performance Evaluation and Power Performance Optimization on MegaProto/E

TAKAYUKI IMADA,[†] MITSUHISA SATO,[†] YOSHIHIKO HOTTA,[†]
HIDEAKI KIMURA,[†] TAISUKE BOKU,[†] DAISUKE TAKAHASHI[†]
and SHIN'ICHI MIURA[†]

We have been developing MegaProto as a prototype to realize a concept of high-performance computing systems with low-power and power-aware technology and highly compact packaging. MegaProto/E is a successor of MegaProto/C which was a first prototype with Transmeta Crusoe. It uses Efficeon (TM8820) as a processor, and has improved memory system and networks. In this paper, we report the performance evaluation of MegaProto/E with respect to its performance and power with comparison with MegaProto/C. We found that MegaProto/E achieve almost twice performance and 1.8 times power-performance ratio comparing to MegaProto/C. In addition, we have applied the optimization of the power performance by controlling of DVFS(Dynamic Voltage and Frequency Scaling) to MegaProto/E. Unfortunately, we found no improvement by DVFS optimization because of the small difference of the possible voltage levels.

1. はじめに

近年，プロセッサの消費電力が上昇しており，高性能なプロセッサを用いるスパコンやPCクラスタシステムなどの大規模な高性能並列計算機の消費電力が増大している．Peta-Flopsコンピューティングを実現するような並列計算機システムを構築するためには数百万プロセッサを必要とし，計算機全体の消費電力は非現実的なものとなり運用は困難であると考えられる．また，プロセッサの消費電力が高いことによって高密度実装が困難となり，巨大な並列計算機を設置するための面積が増大してしまうという問題も生じる．

我々は現実的な設置面積・容量と消費電力の条件下でより高い計算能力を得ることに焦点をあて，高性能でありながらも低消費電力・高密度実装を実現を目指すための実証システムのプロトタイプとしてMegaProto[1)]を開発している．先行研究においてTransmeta社のCrusoeを用いてMegaProtoのプロトタイプの第1バージョンであるMegaProto/C[1)]では性能に加えて低消費電力高性能クラスタとしての指標である電力性能について評価を行った．その後，Crusoeの次期バージョンであるEfficeonを用いた第2バージョンとなるMegaProto/E[2)]ではまだ電力性能の評価が行われていない．MegaProto/Eではプロセッサ等の変更が行われているため，MegaProto/Cとは異なる電力特性を示すことが十分に考えられる．

† 筑波大学大学院システム情報工学研究科
University of Tsukuba, Graduate School of Systems and Information Engineering

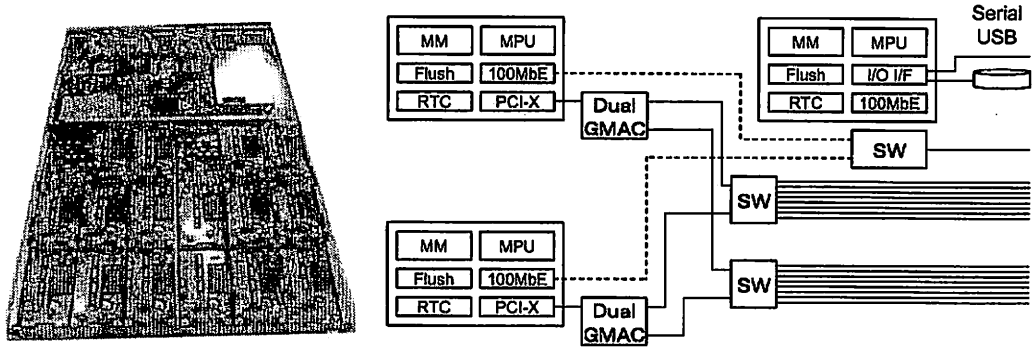

図 1 MegaProto/E クラスタユニット

また，現在多くのコモディティベースのプロセッサでは DVFS（Dynamic Voltage and Frequency Scaling）と呼ばれる機構が実装され，その利用範囲は PC を超えてサーバにまで広がっている．DVFS はプロセッサの周波数・電圧を変更する機能であり，これを用いて性能を低下させることなく消費電力を削減する手法が OS やハードウェアに実装されている．しかし，それらにおける実装は通信やメモリによる CPU のストールによって起こるアイドル状態を十分に活かしきれていない．そのため以前より DVFS 機能を用いて消費電力を削減する様々な手法がハイエンドサーバに対して研究されており，MegaProto のような低消費電力プロセッサを高密度実装したシステムにおいても有効ではないかと考えられる．我々はすでに電力プロファイルを用いた DVFS 制御による電力性能最適化手法[6),7)] を研究開発している．

そこで本論文では MegaProto/C と MegaProto/E の電力性能の違いを明らかにし，MegaProto/E 上でさらに消費電力を削減する方法について検討する．本稿では以下の 2 点に焦点を当てる．

- MegaProto/C で行われた以前の評価を踏まえ，MegaProto/E について評価を行う．
- MegaProto/E における DVFS 制御を用いた電力性能最適化の評価を行う．

本論文では，まず 2 章で MegaProto/E の概要を述べ，3 章で性能および電力性能評価について述べる．4 章では DVFS 制御による電力性能最適化について述べる．そして 5 章で考察を述べ，最後にまとめと今後について述べる．

2. MegaProto/E の概要

2.1 MegaProto

我々は低電力なコモディティ技術を用いることによってメガスケールコンピューティングを実現するためのプロトタイプシステムとして MegaProto を提案している．高性能・高消費電力プロセッサを単純に使用して並列計算機を構築する従来のアプローチと異なり，MegaProto は低消費電力プロセッサを高密度に実装するというアプローチに基づいて設計されている．1TFlop のピーク性能，10kW の消費電力を 19 インチ 42U ラックサイズで実現すること目標とし，プロトタイプの第 1 バージョンとして Crusoe を用いた MegaProto/C，さらにシステムを増強した第 2 バージョンとして Efficeon を用いた MegaProto/E を設計・開発した．

表 1 計算プロセッサカードの仕様

	MegaProto/C	MegaProto/E
CPU	TM5800(933MHz)	TM8820(1GHz)
TDP(Peak)	7.5W	3W
Peak Perf.	0.93 GFlops	2.0 GFlops
L1Cache	64KB(I) + 64KB(D)	128KB(I) + 128KB(D)
L2Cache	512KB	1024KB
Memory	256MB SDR(133MHz)	512MB DDR(266MHz)
I/O Bus	PCI(32bit, 33MHz)	PCI-X(64bit, 66MHz)

2.2 仕　　様

MegaProto/E の基本的な仕様は前バージョンの MegaProto/C と同じである．MegaProto/C の仕様は先行研究[1)] にて詳細に述べられているので，ここでは MegaProto/E のクラスタユニットについて説明し，特に MegaProto/C から MegaProto/E への変更点について述べる．

図 1 に MegaProto/E クラスタユニットの概観とクラスタユニット内部のネットワークを示す．MegaProto/E のクラスタユニットは MegaProto/C と同様 16 基の計算プロセッサカードと 1 基の管理プロセッサカードが 19 インチ 1U サイズのケース内に実装されている．各計算プロセッサカードが Gigabit Ethernet（以下 GbE）でつながれており，1 クラスタユニットが 16 ノードのクラスタシステムとして動作する．**表 1** に MegaProto/C および MegaProto/E の計算プロセッサカードの仕様を示す．先行研究において MegaProto/C では設計から性能評価までの段階を通して，クラスタシステムにおいて以下の問題点が生じた．

表 2 周波数および電圧

Processor	Frequency	Voltage
TM8820	1GHz	850mV
	800MHz	800mV
	700MHz	800mV
TM5800	933MHz	1350mV
	800MHz	1250mV
	667MHz	1200mV
	533MHz	1100mV
	300MHz	900mV

表 3 Xeon サーバのスペック

	Xeon server
CPU	Xeon 3.2GHz × 2
L2Cache	512KB
Memory	DDR1GB(266MHz)
Network	Gigabit Ethernet

表 4 各システムの環境

	MegaProto/E	MegaProto/C	Xeon
Linux kernel	2.6.16	2.4.22	2.4.20-20
gcc/g77	4.0.2	3.4.3	3.4.3
LAM-MPI	7.1.2	7.1.1	6.5.6

- 計算プロセッサカードに搭載されているメモリ容量が小さく，プロセッサの性能を完全に活かしきれなかった．
- 電力とプロセッサ能力の不足から，I/O バスを 32bit/33MHz の PCI にせざるを得なかった．

そこで，我々はこれらの問題点を解決するために第 2 バージョンの MegaProto/E では以下の改良を計算プロセッサカードに行った．

(1) 計算プロセッサを TM5800[3] から TM8820[4] に変更．
(2) メモリ容量を倍増するとともに DDR タイプに変更．
(3) I/O バスを PCI から PCI-X に変更．

計算プロセッサカードの改良により，MegaProto/E では以下の性能向上が期待できる．

- プロセッサあたりの計算性能が約 2 倍になると同時にピーク時の消費電力が半分以下になった．つまりプロセッサのピーク時における電力性能比が 4 倍以上になる．
- メモリ容量とバンド幅が 2 倍になり，プロセッサとメモリのバランスが取れるようになる．
- I/O のバス幅が 4 倍になり，バス幅の不足による GbE のボトルネックが解消される．

表 2 に TM8820 および TM5800 で利用できる周波数と電圧の組合せ（以下 Gear）を示す．TM8820 の最高周波数時の電圧は TM5800 の最低周波数時の電圧よりも低いものとなっている．さらに TM8820 では浮動小数点演算の 2 命令同時発行が可能となり，4 倍以上電力性能比を実現している．計算プロセッサカードの電力容量制限は MegaProto/C および MegaProto/E

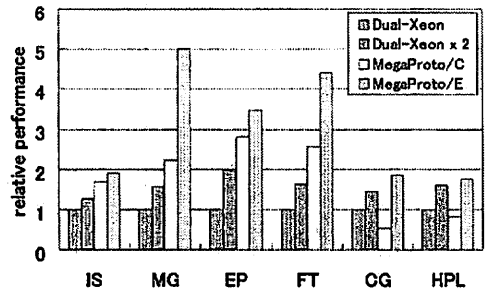

図 2 Dual Xeon サーバとの相対性能比較

ともに 10W であるが，制限を越えることなく 3 つのコンポーネントの性能向上を同時に行うことができた．これは MegaProto/E ではメモリと I/O バスの変更によりこれらの消費電力が MegaProto/C に比べて高くなっているにもかかわらず，プロセッサの変更によって消費電力が下がっているためである．1 クラスタシステム全体の最大負荷時における消費電力は 320W である．

3. MegaProto/E の評価

3.1 評価環境

MegaProto/E の性能評価には NAS Prallel Benchmark（以下 NPB）より IS, MG, EP, FT, CG カーネルベンチマーク（全てクラス A），および High Performance Linpack（以下 HPL）のサイズ N=10000 を使用する．評価には MegaProto/E の 1 クラスタユニット（16 基の計算プロセッサ）を使用する．

また，電力性能測定には我々が以前に提案した PowerWatch[7] を使用し，クラスタユニット全体に供給される DC5V および DC12V の電力について測定を行う．

3.2 計算性能の評価

3.2.1 性能比較の対象

MegaProto/E における各アプリケーションの性能について，MegaProto/C およびハイエンドサーバの代表である Xeon プロセッサを搭載したサーバ (以下 Xeon サーバ) との比較を行う．表 3 に Xeon サーバのスペックを示す．このサーバはハイエンドサーバを想定し，MegaProto/E と同じく 1U サイズで同じ出力容量の電源を備えている．性能の比較対象は MegaProto/C クラスタユニット 1 台，Xeon サーバ 1 台，そして Xeon サーバ 2 台とする．表 4 に使用されているシステムソフトウェアを示す．MegaProto/C と Xeon サーバはなるべくソフトウェア環境を揃えており，MegaProto/E は前記 2 システムよりも新しいバージョンのソフトウェアを使用している．

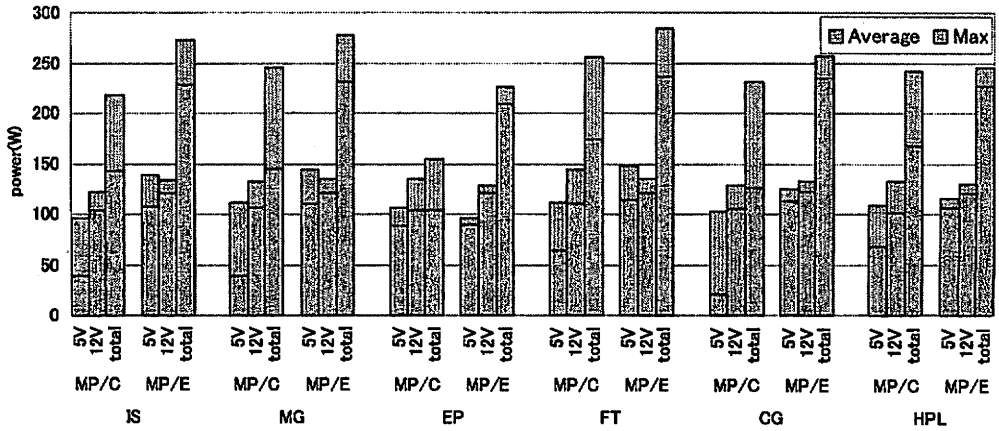

図 3　平均消費電力と最大消費電力

3.2.2 性能評価

図 2 に Xeon サーバ 1 台の性能を 1 としたときの各システムの相対性能を示す．図 2 より，MegaProto/E の性能はすべてのベンチマークにおいて他のシステムよりも高くなっている．MegaProto/E の計算性能において特徴的な点は以下の 2 つである．

(1) CG や HPL といった MegaProto/C で Xeon サーバ 1 台より性能が低かったベンチマークについても MegaProto/E 上では性能が向上し，その性能は Xeon サーバ 2 台よりも高いものとなった．

(2) MegaProto/C と比較して IS の性能があまり向上しなかった．

(1) については，CG はメモリ容量の増加および I/O 周りの性能向上，HPL は SSE2 機能の利用による浮動小数点命令の 2 命令同時発行が要因であると考えられる．以前の報告[1] において MegaProto/C における CG の性能が LongRun[5] の影響で低くなったと報告されているものの，HPL については純粋な性能向上であると考えられる．(2) については，16 ノード上で IS を実行したときに性能が場合によって不安定であった．調査をした結果，全対全通信においてパケット・ドロップによると思われる不具合が不定期に発生していた☆．

3.3 電力性能の評価

先行研究[1] において MegaProto/C の電力性能について触れているが，本評価実験ではさらに詳細な項目について検討を行い，MegaProto/E の電力性能について比較評価を行う．

図 3 に MegaProto/C および MegaProto/E の各ベンチマークにおける平均消費電力と最大消費電力を示す．両システムとも LongRun による DVFS 制御が行われているが，プロセッサの変更により以下の 2 つの特性が現れている．

(1) MegaProto/E における 5V の平均消費電力は MegaProto/C の場合と比較して全てのベンチマークで高くなった．

(2) MegaProto/E ではプロセッサに供給されている 5V の最高消費電力と平均消費電力の差が MegaProto/C と比較して全てのベンチマークにおいて小さくなった．

(1) については計算プロセッサカードの改良により消費電力が向上したメモリや I/O バスが定常的に 5V の電源供給を受けており，それらの消費電力が上乗せされているためでると考えられる．また，(2) については MegaProto/E で用いられている TM8820 で LongRun が使用できる Gear の幅が狭いことによるものであると考えられる．TM8820 では Gear の選択肢が 3 通りしかなく，Gear の上限値と下限値の差が TM5800 と比べて小さい．つまり LongRun による消費電力の削減が MegaProto/C に比べて効果的に行われなかったと考えられる．

また，**図 4** に各ベンチマークを実行したときの電力性能比について，MegaProto/C を 1 としたときの相対比を示す．MegaProto/E の電力性能比は MegaProto/C の場合に比べて最低 0.7 倍 (IS)，最高 1.8 倍 (CG) となっている．唯一 IS において電力性能比が MegaProto/C に対して低くなったのは，MegaProto/C に対する性能があまり高くなく，さらに LongRun の効果によって MegaProto/C の平均消費電力が低かったためである．その他のアプリケーションにおいてはシステム全体の消費電力が上がっているため，計算性能比ほど電力性能比が上がっていない．

☆ 使用した MPI ライブラリによる影響ではないかと考えられる．

4. DVFS 制御による電力性能の最適化

DVFS 制御のスケジューリングによるサーバやクラスタシステムの電力性能最適化は様々な手法が研究されている．そこで，以前我々が提案した DVFS 制御のスケジューリング手法[7] を用いて MegaProto/E において電力性能の最適化が行えるかどうか評価を行う．

4.1 最適化手法

電力性能最適化の評価には我々が以前提案した全体電力プロファイルを用いた DVFS 制御による電力性能最適化手法[6] を使用した．電力性能最適化の流れは以下のようになる．

(1) アプリケーションにおける各領域の実行時間プロファイルとシステム全体の消費電力プロファイルを各周波数ごとに取得する．

(2) 得られた 2 種類のプロファイルより，各領域での消費電力を推測し，各ノードにおける消費電力プロファイルの推測を行う．

(3) 周波数選択アルゴリズムにより各領域における最適な周波数を得る．アルゴリズムの詳細については論文[7] を参照のこと．

同手法ではシステム全体の電力プロファイルだけで最適化を行うことが可能であり，個々のノードで電力プロファイルを取ることが不可能な MegaProto/E のような高密度実装クラスタシステムにおいても電力性能最適化を行うことができる．

4.2 最適化の評価

電力性能最適化の評価には NPB より IS, MG, FT, CG (全てクラス A) を使用した．また，各ベンチマークにおける領域の区切り方は先行研究[7] と同様であり，アプリケーション内の通信部分と計算部分を区別し，関数を単位として区切っている．最適化の指標には PC クラスタの電力性能として広く利用されている EDP(Energy Delay Product) を使用し，アプリケーション実行のエネルギー量に加えて高性能計算で重要となる要素である実行時間を考慮した最適化を行った．

MegaProto/E において上記手法を用いて DVFS 制御による電力性能最適化を行った結果，どのベンチマークにおいてもすべての領域で最高周波数を選択するという結果となった．図 5 に最低周波数と最高周波数で FT を実行したときのシステム全体消費電力の時間変化を示す．FT では最高周波数と最低周波数の場合を比較すると，通信時のプロセッサ消費電力の増加が 10W 程度であり，システム全体消費電力に対する割合は 5%にも満たない．一方，最高周波数と最低周波数の場合の時間の変化幅が 1.2 秒であり，これは約 18%の実行時間短縮になる．EDP を評価指標として周波数選択を行うとき，個々の領域においても実行時間短縮による EDP 値の減少が周波数選択に優先されたと考えられる．

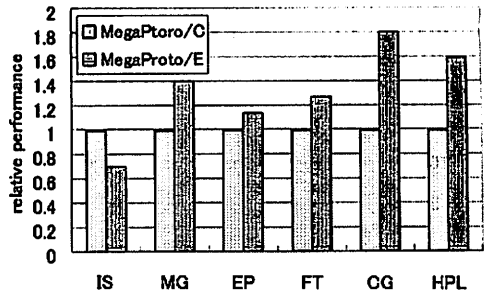

図 4 電力性能比の相対比較

他のアプリケーションでも同様の現象が見られ，FT と同じ理由で全ての領域において最高周波数を選択したと考えられる．

5. 考　　察

5.1 DVFS 制御による電力性能最適化

MegaProto/E において DVFS 制御による電力性能最適化を行った結果，どのベンチマークにおいてもすべての領域で最高周波数で実行するのが最適であると判断された．これには以下の 2 つが影響しているものと考えられる．

(1) **クラスタユニットのプロセッサ以外の消費電力**
MegaProto/E の 1 クラスタユニット全体の平均消費電力は，計算プロセッサが使用する DC5V よりもそれ以外で使われる DC12V の比率が高くなっていた．また，DC5V はプロセッサ以外にもメモリや PCI-X バスなど定常的に電力を消費するコンポーネントに供給されている．そのため，MegaProto/E ではプロセッサの消費電力削減による効果がシステム全体において相対的に小さくなってしまっている．

(2) **計算プロセッサのピーク電力の低さ**
計算プロセッサに用いた TM8820 は最高周波数時における TDP が 3W である．したがってこの上限値で DVFS 制御が行われるとき，削減できる消費電力がシステム全体の消費電力に対して相対的に小さかったと考えられる．つまり周波数と電圧を低下させることによる消費電力の減少よりも，同時に生じる実行時間の増加が EDP に影響して EDP を増加させてしまったと考えられる．

5.2 さらなる電力性能の向上の検討

MegaProto/E は MegaProto/C よりも低消費電力のプロセッサの使用，そしてその他のコンポーネントに消費される定常的な消費電力が相対的に増加していることが大きな特性であると考えられる．この状態か

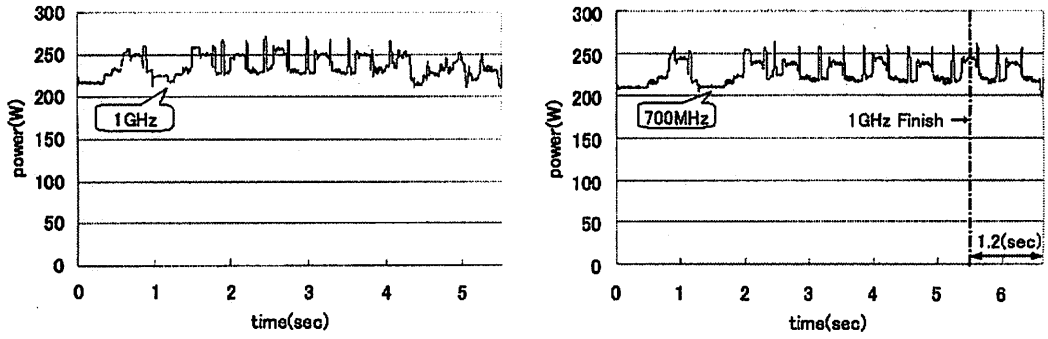

図 5 FT における消費電力の時間変化

らさらに消費電力の削減を行うとすれば，プログラム実行中において動的に使用プロセッサ数を変化させるなどの手法が必要になると考えられる．MegaProto/E では計算プロセッサカードに搭載されているプロセッサ以外のコンポーネントも定常的に電力を消費しているため，プログラム実行中に動的に使用するノード数を変化させれば，大きく消費電力を削減できるのではないかと考えられる．

6. おわりに

本論文では MegaProto/E の電力性能評価および電力プロファイル情報を用いた DVFS 制御による電力性能最適化の検討を行った．評価実験によって MegaProto/E は同サイズのハイエンドサーバより最大 5 倍，前バージョンの MegaProto/C よりも最大 2 倍の性能を発揮し，さらに電力性能が MegaProto/C の最大 1.8 倍になることが分かった．また，電力プロファイル情報を利用した DVFS 制御による電力最適化では最適化を行ったが，低電力プロセッサを用いた MegaProto/E にはあまり有効ではないことが分かった．

今後の課題として，MegaProto/E のような低消費電力プロセッサを高密度実装した大規模クラスタにおいてもう一度電力性能最適化について検討する必要がある．今回の評価実験では行えなかったネットワーク周りの詳細な電力評価を行い，ネットワークも含めたクラスタシステム内のコンポーネントを動的に利用することが実現できればより効果的に消費電力を削減することが期待できる．

謝辞 本研究の一部は科学技術振興機構・戦略的創造研究推進事業（CREST）の研究プロジェクト「低電力化とモデリング技術によるメガスケールコンピューティング」による．

参 考 文 献

1) H. Nakashima, et al.: MegaProto: 1TFlops/10 kW Rack Is Feasible Even with Only Commodity Technology, in Proc. Supercomputing Conference 05 (2005)

2) T. Boku, et al.: MegaProto/E: Power-Aware High-Performance Cluster with Commodity Technology, in Proc. High performance Power-Aware Computing Workshop(HPPAC) in IPDPS'06 (2006)

3) Transmeta Corp.: Crusoe TM5800 Processor Product Sheet (2003)

4) Transmeta Corp.: Efficeon TM8820 Processor Product Sheet (2005)

5) Transmeta Corp.: Crusoe LongRun Power Management White Paper (2001)

6) 堀田 義彦，佐藤 三久，木村 英明，朴 泰祐，高橋 大介: PC クラスタにおける全体電力プロファイルを用いた電力性能最適化，情報処理学会研究報告 2006-ARC-169, pp.1-6, (2006)

7) Y. Hotta, et al.: Profile-based Optimization of Power Performance on by using Dynamic Voltage Scalng on a PC cluster, High Performance Power-Aware Computing Workshop(HPPAC) in IPDPS'06 (2006)