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

2005－ARC－162 (21)
2005－HPC－101 (21)
2005／3／8

高性能計算のための低電力・高密度クラスタ MegaProto

中島　浩[†1]　中村　宏[†2]　佐藤三久[†3]
朴　泰祐[†3]　松岡　聡[†4]
高橋大介[†3]　堀田義彦[†3]

現在進行中の研究プロジェクト「低電力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」において，我々は百万プロセッサ規模の並列システムは低電力コモディティプロセッサの高密度実装によってのみ実現可能であると主張し，それを実証するためのプロトタイプ *MegaProto* を開発している．MegaProto は 19 インチラックに搭載可能な 1U サイズのクラスタユニットを単位として構成され，一つのユニットには 16 個の低電力プロセッサと，それらを結合するプロセッサあたり 2 Gbps の高バンド幅ネットワークが搭載される．ユニットあたりのピーク性能は第 1 バージョンで 14.4 GFlops，第 2 バージョンで 38.4 GFlops であり，ユニット内およびユニット間のネットワークバンド幅はそれぞれ 32 Gbps，8 Gbps である．また，消費電力は最大 300 ～ 330 W と小さく，従来型の 1U サーバ，たとえばハイエンドのデュアルプロセッササーバと同等以下である．一方 NPB による性能評価の結果，第 1 バージョンにおいても 4 つのベンチマークでデュアルプロセッササーバを大きく凌駕し，最大 2.8 倍の高い性能を発揮することが明らかになった．

MegaProto: A Low-Power and Compact Cluster for High-Performance Computing

Hiroshi Nakashima,[†1] Hiroshi Nakamura,[†2] Mitsuhisa Sato,[†3]
Taisuke Boku,[†3] Satoshi Matsuoka,[†4] Daisuke Takahashi[†3]
and Yoshihiko Hotta[†3]

MegaProto is a proof-of-concept prototype for our project "Mega-Scale Computing Based on Low-Power Technology and Workload Modeling", implementing our key idea that a million-scale parallel system should be built with densely mounted low-power commodity processors. The building block of the MegaProto is a 1U-high 19 inch-rack mountable motherboard unit on which 16 low-power, one-dollar note-sized, commodity PC-architecture daughterboards are mounted with a high bandwidth, 2 Gbps per processor network based on Gigabit Ethernet. The peak performance of each unit is 14.4 GFlops for the first version and will improve to 38.4 GFlops in the second version through a processor/daughterboard upgrade. The intra- and inter-unit network bandwidths are 32 Gbps and 16 Gbps respectively. As for power consumption, the entire unit consumes 300-330 W maximum under extreme computational stress; this is comparable to or better than conventional 1U servers comprised of dual high-performance, power hungry processors, while benchmarks exhibit up to 279 % superior performance for some NPB programs.

1. はじめに

我々は，科学技術振興事業団・戦略的創造研究推進事業の研究プロジェクトとして，「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」を実施している．このプロジェクトの目的は，Peta-Flops クラスの計算能力を有する百万プロセッサ級のメガスケール計算システム構築のための基盤技術の開発であり，その実現性，信頼性およびプログラム容易性に重点をおいた研究開発を行っている．中でもメガスケール計算システムの実現性の鍵は，現実的な設置面積・容積と消費電力の制約下で，いかに多数の計算資源を実装して高い性能を得るかにある．したがって我々は，高性能・高電力のプロセッサを用いる従来型の MPP やクラスタではなく，低電力プロセッサを高密度に実装するアプローチこそがメガスケール計算を実現する

†1 豊橋技術科学大学
Toyohashi University of Technology
†2 東京大学
University of Tokyo
†3 筑波大学
University of Tsukuba
†4 東京工業大学
Tokyo Institute of Technology

唯一の方法であると主張している．

この主張を裏付けるひとつの方法は，現時点で利用可能なコモディティ技術を用いて高密度・低消費電力・高性能のシステムを構築し，その延長線上に我々が目指すメガスケール計算システムが存在することを実証することである．そこで我々は，多数の低電力プロセッサを高密度に実装し，それらを高信頼・高バンド幅のネットワークで結合したプロトタイプシステム *MegaProto* を開発している．また MegaProto は，プロジェクトで研究・開発中の様々な技術，すなわち低電力化コンパイル技術，高信頼・高性能ネットワーク技術，高信頼クラスタ構築技術，多重並列プログラミング技術などのの実証プラットフォームとしても利用される．

以下本報告では，第 2 章で MegaProto の設計方針を，また第 3 章でその構成単位であるクラスタユニットの設計について述べる．また第 4 章では NPB および HPL による性能と消費電力の評価結果を示し，第 5 章でまとめを行う．

2. 設 計 方 針

2.1 性能／電力比

前述のように MegaProto の開発目的は，現時点で利用可能な技術を用いた高密度・低消費電力のシステム構築であり，そのためには低電力プロセッサの使用が不可欠である．しかし単に低電力というだけでは不十分であり，たとえば浮動小数点演算機構を持たない携帯機器用のプロセッサなどでは，我々が目指す Peta-Flops 計算への方向性と大きく乖離したものとなってしまう．そこで MegaProto の仕様設計に際して，まず以下に示す大まかな性能目標を定め，その値に近い性能を達成できる構成が可能かどうかを検討することとした．

- 19 インチの 42U ラックに搭載されるシステムの性能目標を，ピーク性能 ＝ 1 TFlops，消費電力 ＝ 10 kW と設定する．
- システムの消費電力の 1/2 はネットワークなど計算以外の部分で消費され，残りの 1/2 の更に 1/2 はメモリなどプロセッサ周辺のデバイスで消費されると仮定する．したがってプロセッサの消費電力は，システム全体の 1/4 となる．

第 1 の性能目標から性能電力比を求めると 100 MFlops/W となり，これを単純に外挿したピーク 1 Peta-Flops を達成するための規模と消費電力は 1,000 ラック，10 MW となる．この値は，実現困難ではあるものの夢想的な数字ではない．また性能電力比が将来的に 5 ～ 10 倍程度改善されると仮定すれば 1 Peta-Flops の達成は一気に現実的になるが，後述するようにこの仮定もやはり夢想的なものではない．

一方，第 2 の仮定を加味したプロセッサ単体の性能電力比は 400 MFlops/W となる．この値に対し，**表 1**(a) に示す 2003 年（すなわち MegaProto 設計開始時点）でのプロセッサの性能電力比は，4 ～ 6 GFlops の高性能プロセッサでは約 1/5 ～ 1/7 と大きく下回っており，かつピーク性能が浮動小数点命令の 2 命令同時実行によるもの（Intel のプロセッサでは SSE2 による）であることを考慮すると，表記した数値以上に乖離していると言わざるを得ない．一方，1 Gflops 程度のモバイルプロセッサでも約 1/3 ～ 1/4 の値しか得られなかったが，近い将来の改善を期待しつつ，このグループ中で絶対的な消費電力が最小で，かつピーク性能がクロックあたり 1 浮動小数点命令の実行で得られる TM5800 を選択した．

なお，この選択によって 1 TFlops/10 kW のシステムを断念したわけではなく，プロセッサを容易に交換できる設計を行い，上記のように短期間での性能改善を期待した．この期待の妥当性は，表 1(b) に示す 2004 年の調査によって裏付けられている．すなわち，TM8800 が上記の目標値 400 W/Flops の 1.5 倍もの数値を達成しているほか，Pentium M も目標値の 1/2 を越える値を達成している．特に我々が選択した TM5800 の後継機である TM8800 が 600 W/Flops という優れた値を示していることで，設計の妥当性が立証された．

2.2 プロセッサの実装

前節で定めた性能電力比から，消費電力 5 W のプロセッサを用いれば，2.5 kW ＝ 10 kW ÷ 4 の消費電力制約のもとで，ラックあたり 500 プロセッサのシステムを構築できることが導かれる．これを 1U あたりのプロセッサ数に換算すると 500 ÷ 42 ≈ 12 となり，現在の実装技術で十分達成可能な値となる．一方 1U サーバと同程度のマザーボード上に diskless のプロセッサノードを何ノード配置できるかを検討した結果，16 ノード（あるいはそれ以上）の実装は十分可能であるという結論に達した．

ここでシステム全体のネットワークの構造が，マザーボード内の結合とマザーボード間の結合の（少なくとも）2 階層となることと，マザーボード間の結線・接続コストが大きいことを考えると，マザーボード上にできるだけ多数のプロセッサを配置することが得策であることは明らかである．またブレードサーバーのように比較的少数のプロセッサからなるボードを多数搭載する構成は，ネットワーク階層の増加や最下層のプロセッサ数減少をもたらすため得策ではないと判断した．

これらの事項を総合的に検討した結果，1U マザーボードを「クラスタユニット」とし，1 ユニットに 16 プロセッサを配置して，ラックあたり $16 \times 42 = 672$ プロセッサの構成とすることとした．この結果ラックあたりの消費電力が目標値よりも 35 %程度上回ることとなるが，許容できる範囲であると判断した．

表 1 主なプロセッサの性能電力比
Table 1 Performance/power ratio of modern microprocessors

(a) 2003 年 9 月の調査

機種名	周波数 (*1)	性能 (*2)	TDP (*3)	性能電力比 (*4)
Athlon XP Model 10[1)]	2.20	4.40	76.8	57.3
Xeon[3)]	3.00	6.00†	85.0	70.6
Pentium 4[3)]	3.20	6.40†	82.0	78.0
Mob. Pentium 4[3)]	3.06	6.12†	70.0	87.4
Mob. Pentium III-M[3)]	1.00	1.00	10.5	95.2
TM5800[6)]	0.93	0.93	7.5	124.0
Mob. Celeron[3)]	2.40	4.80†	35.0	137.1
Mob. Pentium 4-M[3)]	2.60	5.20†	35.0	148.6

(b) 2004 年 11 月の調査

機種名	周波数 (*1)	性能 (*2)	TDP (*3)	性能電力比 (*4)
Xeon[4)] (*5)	2.80	5.60†	111.0	50.5
Pentium 4[4)]	3.80	7.60†	115.0	66.1
Mob. Pentium 4[4)]	3.33	6.66†	88.0	75.7
Celeron M[4)]	1.50	1.50	24.5	61.2
	0.90	0.90	7.0	128.6
Pentium M[4)]	2.10	2.20	21.0	100.0
	1.10	1.10	5.0	220.0
TM8800[6)] (*6)	1.20	2.40†	4.0	600.0

(*1) 動作周波数 (GHz). (*2) ピーク性能 (GFlops). † を付した数値は SSE2 による性能.
(*3) Thermal design power (W). (*4) ピーク性能／電力比 (MFlops/W).
(*5) 3.0∼3.6GHz 品もリリースされているが，それらの TDP は 4) には記載されていない. (*6) TDP は概算値.

2.3 ネットワーク

プロセッサの選定や実装と同様に重要な設計ポイントであるネットワークについては，複数のコモディティネットワークを束ねた構成が性能（バンド幅）と信頼性の両面で最適な解であることを，我々は RI2N (Redundant Interconnection with Inexpensive Network) の研究を通じてすでに示している[5)]．そこで MegaProto では，プロセッサあたり 2 ポートのコモディティネットワーク，すなわち 2 系統の Gigabit-Ethernet (GbE) を持つ構成とした．

アップリンクについては，GbE を複数用意してバンド幅を確保する方法と，Infiniband や 10 Gbps Ethernet などの高バンド幅リンクとする方法が考えられる．後者はクラスタユニット間の結線の面で魅力的ではあるが，現時点でのクラスタユニット内外のネットワーク部品・機器のコストは小さくなく，価格性能比の面で問題が大きいと判断した．一方前者はクラスタユニット間の結合に多数の結線やスイッチを必要とするが，低価格の小ポート数スイッチを多数用いる構成は価格性能比の面で優れていることが実証されており，この方法を選択することとした．

この結果一つの系統について，クラスタユニット上の GbE スイッチのポート数はプロセッサ数とアップリンク数の和となるが，現時点☆で価格性能比に優れたスイッチのポート数の上限が 24 であることから，アップリンクのポート数を 8 と設定した．この結果，系統あたりのバンド幅は，クラスタユニット内部で 16 Gbps，またユニット間では 8 Gbps となり，2 系統を合算するとそれぞれ 32 Gbps/16 Gbps という，十分大きな値を確保することができる．

3. クラスタユニットの設計

クラスタユニットの構成を**図 1** に示す．クラスタユニットは 432mm(W) × 756 mm(D) × 44 mm(H) の 1U シャシーに実装され，その約半分に 16 個のプロセッサカードが，また残りの半分に 2 系統の 24 ポート GbE スイッチを中心とするネットワーク，管理用プロセッサ，および電源が搭載されている．クラスタユニットの最大消費電力は，TM5800 を用いた第 1 バージョンで 300 W，TM8800 を用いた第 2 バージョンでは 330 W であり，いずれも小型の低速ファン 4 個による空冷で十分に冷却可能な値となっている．

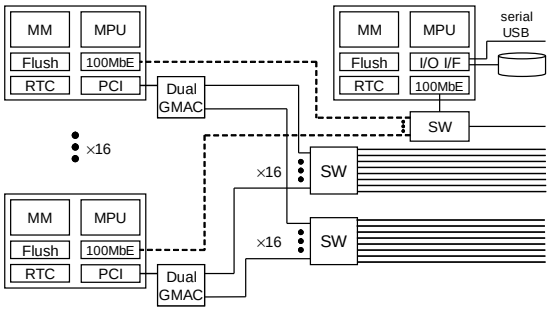

図 1 クラスタユニット
Fig. 1 Cluster unit

3.1 プロセッサカード

プロセッサカードは低電力プロセッサを中心に構成され，千円札よりもやや小さい 65 mm × 130 mm のカード上に，主記憶，フラッシュメモリ，I/O インタフェースなどの周辺回路も実装されている．前章で述べたように，モバイルプロセッサの最新技術を活用するため，クラスタユニットは共通のマザーボードと交換可能なプロセッサカードから構成されている．

表 2 は第 1 および第 2 バージョンのプロセッサカードの仕様を示したものである．最も重要なポイントは TM5800 から TM8800 への置換であり，第 1 バージョンでは 0.93 GHz の TM5800 を用いているためクラスタユニットあたりのピーク性能はで 14.9 GFlops であ

☆ 技術の動向から考えて，おそらく将来についても．

表 2 プロセッサカードの仕様
Table 2 Processor card specification

	1st version	2nd version
MPU	TM5800 (0.93GHz)	TM8800 (1.2GHz)
Caches	L1=64KB(I)+64KB(D)	L1=128KB(I)+64KB(D)
	L2=512KB(D)	L2=1MB(D)
Memory	256 MB SDR (133 MHz)	512 MB DDR (266 MHz)
Flush	512 KB	1 MB
I/O Bus	PCI (32 bit, 33 MHz)	PCI-X (64 bit, 66 MHz)

るのに対し，第 2 バージョンでは 1.2 GHz の TM8800 によって 38.4 GFlops の性能が得られる．後者の値から 42U ラックでのピーク性能を求めると 1.6 TFlops となり，極めて高性能かつコンパクトなクラスタシステムを構築することができる．

またプロセッサの置換だけではなく，これに付随するいくつかの改良設計も性能面で大きな意味を持っている．まず，メモリ容量を 256 MB から 512 MB に増やしつつ，133 MHz の SDR を 266 MHz の DDR に置き換えてメモリバンド幅を向上させた．また I/O バスを 32 ビット/33 MHz の PCI から 64 ビット/66 MHz の PCI-X に置き換え，2 本の GbE リンクとのデータ送受をカバーできる性能とした．このように，プロセッサの能力向上に応じて周辺の性能も向上させているため，2 つのバージョンの計算／メモリアクセス／通信の性能バランスはともに優れたものとなっている．

3.2 ネットワーク

データネットワークである RI2N は独立した 2 系統からなり，各々が 24 ポートの Layer-2 GbE スイッチを中心に構成される．一つのスイッチについて，16 ポートはプロセッサのネットワークインタフェースである 2 ポートの GMAC チップに接続され，PCI-X（第 1 バージョンでは PCI）バスを経由してプロセッサと接続されている．残りの 8 ポートには，クラスタユニット外への 1000Base-T アップリンクが接続される．スイッチ速度は 20 Gbps であり，ほぼ wire speed でのスイッチングが実現できる設計とした．前述のようにクラスタユニットには 2 系統の GbE ネットワークが搭載されるため，クラスタユニット内の総バンド幅は 32 Gbps，ユニット間のバンド幅は 16 Gbps となる．

この他，後述する管理プロセッサとの通信用に 100Base-TX のネットワークを用意し，クラスタユニット内の全プロセッサノードと管理プロセッサを接続することとし，そのアップリンクとして GbE (1000Base-T) のリンク 2 本が用意されている．

3.3 管理プロセッサ

管理プロセッサは，クラスタユニットの IPL，異常検出，およびネットワークの設定管理を行うために用意され，通常の計算処理には参加しない．したがって基本的にはプロセッサノードと同一の構成ではあるが☆，プロセッサノードとの通信は管理ネットワーク経由でのみ行い，データ転送用の 2 系統 GbE ネットワークの通信に悪影響を与えない構成とした．また I/O として，60 GB のハードディスク，USB およびシリアルポートが各々1 ポート備えられている．

OS (Linux) のブートイメージはこのハードディスクに格納されており，管理プロセッサと管理ネットワークを経由して個々のプロセッサにロードされる．一方メインファイルシステムはクラスタユニットの外部に置くことができ，データネットワークあるいは管理ネットワークを経由した NFS によってアクセスされる．

4. 性能評価

本章では MegaProto の第 1 バージョン（TM5800 バージョン）に関する，予備的な性能評価について述べる．まず 4.1 節では，5 つの NPB3.1 カーネルベンチマーク（IS, MG, EP, FT, CG）と HPL 1.0a を用いた性能と消費電力の評価結果を示す．これらのベンチマークは LAM-MPI 7.7.1 を用いてプログラムし，gcc/g77 3.3.2 によりコンパイルしたコードを Linux 2.4.22mmpu の管理下で実行した．また消費電力の測定にはホール素子を用いた測定器を使用し[2)]，100 V AC 入力と，5 V（プロセッサ電源）および 12 V の DC 出力の電源電流を測定した．

また 4.2 節では，Xeon のデュアルプロセッサ構成の 1U サーバとの性能比較を行うが，このサーバのソフトウェア構成は MegaProto とほぼ同じであり，LAM-MPI 6.5.6, gcc/g77 3.4.3, および Linux 2.4.20-20.7smp を用いた．

4.1 実行速度と消費電力

表 3 に，2 ～ 16 プロセッサでの NPB と HPL の実行速度を示す☆☆．また 4 プロセッサ性能を基準とした台数効果を**図 2** に示す．

NPB の結果は，多くの高性能クラスタと類似したものとなっており，MegaProto の設計がクラスタとして妥当なものであることを示している．すなわち EP, FT, MG は良好な台数効果を示し，それらには劣るものの IS でも高い並列性能が発揮されている．また CG の台数効果は相対的に小さなものとなっているが，これはプログラムのスケーラビリティが低いためである．一方 HPL については，16 プロセッサで 5.62 GFlops の性能が達成されているが，この値がピーク性能の 38 %であることを勘案すると，必ずしも満足できる性能であるとは言えない．この主な理由は，プロセッサ

☆ 第 2 バージョンにおいても，管理プロセッサには TM5800 を使用している．

☆☆ 単一プロセッサでの性能も測定したが，ほとんどのベンチマークで必要とする記憶域が主記憶容量を越えるため頻繁にスワップが発生し，複数プロセッサでの性能と意味ある比較が困難な低い性能であったため省略した．

表 3 NPB と HPL の性能値
Table 3 Performance of NPB and HPL

# of proc.	NPB (class A)[Mop/s] IS	MG	EP	FT	CG	HPL [GFlops]
2	10.1	153.1	5.0	(*1)	95.6	(*1)
4	17.4	262.6	10.0	257.9	115.7	2.07
8	29.6	507.9	19.9	476.4	163.4	3.61
16	43.8	831.6	39.8	923.9	217.5	5.62

(*1) メモリ容量不足のため測定不能.

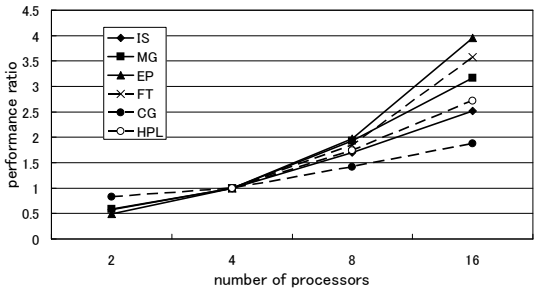

図 2 NPB と HPL の台数効果
Fig. 2 Speedup of NPB and HPL

あたり 256 MB というやや小さめの主記憶容量にある．すなわち問題サイズを十分大きなものとできないため，プロセッサの計算性能を完全に発揮することができていない．この問題は第 2 バージョンで主記憶容量を倍増することにより大幅に改善されるため，第 2 バージョンの実効／ピーク性能比は大きく向上することが期待できる．

図 3 は，もう一つの重要な評価項目である消費電力の測定結果を示したものである．図には AC および DC（5 V および 12 V）の各電源について，それぞれの最大消費電力（暗色の棒グラフ）と平均消費電力（白色）が示されている☆．いずれのベンチマークにおいても，全実行期間の中に計算集約的な部分は多かれ少なかれ存在するので，最大消費電力はベンチマーク間で差はなく，かつ最大設計値にかなり近い値となっている．一方プロセッサ電源である 5 V の平均消費電力は，CG の 21 W から EP の 88 W までの範囲に大きく広がっており，AC の平均電力もそれに対応する形で変化している．

この 5 V 電力の変動は，TM5800 の DVS である LongRun に負うものである．CG のようにスケーラビリティが小さく通信の占める割合が大きなプログラムでは，計算集約的な実行区間がほとんどなく，多くの部分で通信待ち状態になっている．LongRun はこの通信待ち状態を検出し，クロック速度と電源電圧を

☆ DC 電力の和が AC 電力に一致しないのは，約 20 %の変換損失が主要な理由であり，測定していないプロセッサ周辺回路の電力も若干影響している．

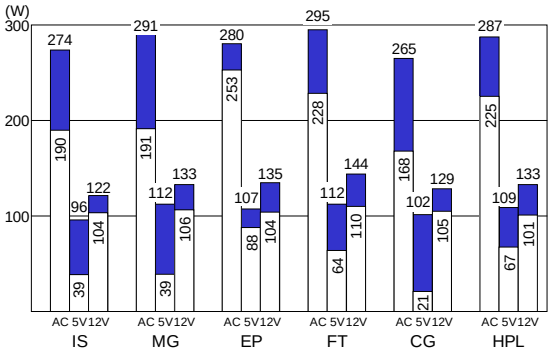

図 3 消費電力の最大値と平均値
Fig. 3 Peak and average power consumption

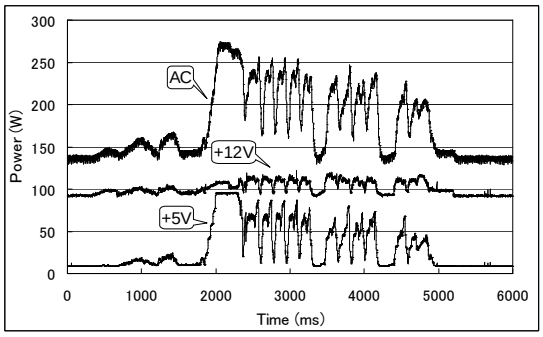

図 4 IS の電力プロファイル
Fig. 4 Power profile of IS

自動的に低下させるため，CG の消費電力は非常に小さな値に抑えられている．一方 EP は全実行期間がほぼ計算集約的であるため，常に最大値に近い電力が消費される．他のベンチマークはこの両者の中間的な挙動を示し，たとえば IS では**図 4** に示すように高電力の計算区間と低電力の通信区間の繰り返しが観測される．これらの結果から，LongRun がプログラム実行中の挙動変化を的確に検出してクロックと電源電圧を制御し，総体的な消費エネルギーを効果的に低く抑えていることが明らかになった．

また 2 つの図が示す別の興味深い事実として，12 V 電源の消費電力がほとんど変動しないことが挙げられる．この電源は主としてネットワーク系を駆動するものであるので，通信頻度や通信量によらずかなり多くの電力が定常的に消費されるのは奇異に思える．この原因は，MegaProto で採用したネットワーク系デバイスが，通信の有無に関わらず「常に」リンクを駆動し続けることにある☆☆．この消費電力に対する無頓着さは，研究室やオフィスのネットワーク用途では全く問題にならないと考えられるが，低電力・高密度クラスタにとってはきわめて重大な障害となり，その解決

☆☆ MegaProto のデバイスだけではなく，一般的な性質である．

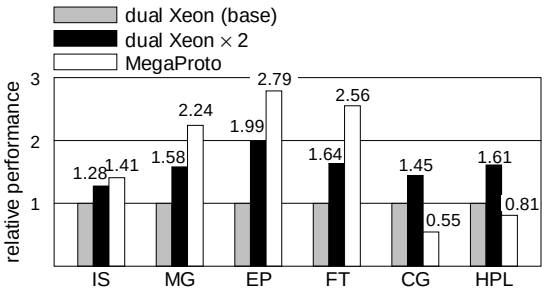

図 5 デュアル Xeon サーバとの性能比較
Fig. 5 Performance relative to dual-Xeon server

は今後の重要な課題である．

4.2 デュアル Xeon サーバとの性能比較

前節では MegaProto の性能と消費電力について，その絶対値に基づく議論を行ったが，本節では従来型のハイエンドサーバおよびミニクラスタとの比較評価を行う．比較対象は，3.06 GHz の Xeon 2 台と 1 GB の DDR 共有メモリから構成された，1U サイズのデュアル Xeon サーバ (Appro 1124Xi) である．2 台の Xeon の TDP とピーク性能はそれぞれ 170 W[3] および 12.2 GFlops であり，またサーバ全体の消費電力は約 400 W である．これらの値はいずれも MegaProto とほぼ同等であるため，同じ 1U サイズのシステムとして妥当な比較対象であるといえる．

図 5 はサーバの性能を 1 として正規化した相対性能を示したものであり，MegaProto の優位性を立証する結果となっている．すなわち MegaProto は IS, MG, EP および FT でデュアル Xeon サーバの性能を大きく凌駕し，最大 2.8 倍 (EP) の性能を発揮している．またこれらのベンチマークでは，2 つのデュアル Xeon サーバを GbE で結合したミニクラスタ（すなわち $2 \times 2 = 4$ プロセッサの小規模 SMP クラスタ）との比較においても，IS で 10 %，それ以外では 40 %も上回るという好結果が得られた．これらの結果は，多数の低電力プロセッサからなるクラスタは，ハイエンドであっても（あるいはそれが故に）プロセッサ数が小さなクラスタに優るという，我々の主張を明確に実証している．

しかし MegaProto の優位性は完全なものではなく，CG の性能はデュアル Xeon サーバの約 1/2 に留まり，HPL でも僅かに及ばない結果となっている．前節で述べたように，これらのベンチマークの性能が不十分である理由はメモリ容量が小さいことであり，第 2 バージョンではこの問題が解消される．また 90 nm 世代である TM8800 のピーク演算性能は 130 nm 世代である TM5800 の約 2.5 倍（2.4 GFlops/0.93 GFlops）であり，かつネットワークを駆動する I/O バスの性能も大幅に改善される．一方 Xeon では 130 nm 世代と 90 nm 世代の演算性能比は小さく，これらを総合すると 90 nm 世代での MegaProto の全面的優位性は，ほぼ明らかであると考えられる．

5. ま と め

本報告では，我々が開発中の低電力・高密度の高性能計算向けクラスタ MegaProto の設計について述べた．MegaProto は，16 個の低電力プロセッサとプロセッサあたり 2 本の GbE からなるネットワークを搭載した 1U サイズのクラスタユニットを単位として構成され，各プロセッサはディスクレスの完全な Linux PC として稼動する．MegaProto のクラスタユニットには，0.93 GFlops の TM5800 を用いた第 1 バージョンと，2.4 GFlops の TM8800 を用いた第 2 バージョンとがあり，どちらも 300 ～ 330 W という低消費電力を特徴としている．クラスタユニットのマザーボードは 2 つのバージョンで共通であるが，第 2 バージョンでのプロセッサやメモリの性能向上に対応可能なように，十分な大きさの I/O およびネットワークバンド幅が確保されている．

第 1 バージョンに関する性能評価の結果，4 種の NPB カーネルベンチマークにおいて，従来型の 1U サーバを大きく凌駕する性能を発揮することが明らかになった．第 2 バージョンではシステムの性能電力比が 116.4 MFlops/W に改善されるため，優位性がさらに向上し，低電力・高密度のクラスタは従来型のハイエンドクラスタに優るという事実が一層明らかになると予想される．大規模高性能計算においては，冷却や電源供給が常に大きな問題となってきたが，MegaProto によって 1 筐体あたり 1 TFlops を 10 kW 未満で達成することが可能になり，高性能計算システムの新たな局面を切り開くことができる．

参 考 文 献

1) Advanced Micro Devices, Inc.: *AMD Athlon XP Processor Model 10 Data Sheet* (2003).
2) Hotta, Y. et al.: Measurement and Characterization of Power Consumption of Microprocessors for Power-Aware Cluster, *COOL Chips VII* (2004).
3) Intel Corp.: Datasheets of the Intel processors on 0.13 micron process, `http://www.intel.com/` (2003).
4) Intel Corp.: Datasheets of the Intel processors on 90nm process, `http://www.intel.com/` (2004).
5) Miura, S. et al.: RI2N—Interconnection Network System for Clusters with Wide-Bandwidth and Fault-Tolerancy Based on Multiple Links, *ISHPC 2004*, pp.342–351 (2003).
6) Transmeta Corp.: Product Sheets of TM5800 and TM8800, `http://www.transmeta.com/` (2003-2004).