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．また同時に MegaProto は，プロジェクトで開発中の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を実証するためのプラットフォームとしても機能する．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，1 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 32.0 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，16 Gbps である．また，消費電力は待機時で 150 W，最大計算負荷を課した条件でも 300〜320 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ており，コモディティ技術により高密度・低消費電力・高性能が同時に達成できることが実証された．

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. It also serves as a platform to implement and evaluate our new technologies such as power conscious compilation, highly reliable and high performance networking, highly dependable cluster management, and multi-level scalable parallel programming. 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 32.0 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 idles at less than 150 W and consumes 300-320 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. This demonstrates that higher performance can be achieved with low-power, densely populated architectures with commodity components.

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

1. はじめに

我々は，科学技術振興機構・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．なか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唯一の方法であると主張している．

この主張を裏付ける 1 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．

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，1 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 32.0 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，16 Gbps である．また，消費電力は待機時で 150 W，最大計算負荷を課した条件でも 300〜320 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ており，コモディティ技術により高密度・低消費電力・高性能が同時に達成できることが実証された．

以下本論文では，2 章でプロジェクトの概要を述べた後，3 章で MegaProto の設計方針を，また 4 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．5 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，6 章で関連研究について述べた後，7 章でまとめを行う．

2. プロジェクトの概要

PetaFlops 級の計算能力を得るためにはきわめて大規模な並列システムの構築が不可欠であるが，従来の MPP やクラスタ計算機技術の延長でのプロセッサ数増加は，設置面積，消費電力，メンテナンス，ソフトウェア開発の面で限界に来ている．たとえば，ASCI プロジェクトの MPP や地球シミュレータは，数千〜1 万プロセッサですでに小スタジアムほどの大きさを占め，電力も 10 メガワット以上を消費する．

一方，より一般的な計算機技術分野において，メガスケールコンピューティングを実現する技術の別のコンテクストでの研究開発が進みつつ，または注目されつつある．これらは従来のようにハイエンドではなく，むしろ汎用的なコモディティ技術の基盤となるもの，あるいはそれをベースとするものである．我々の主張は，このような技術をベースとするアプローチ，すなわち単純に高性能や高機能を目指した従来型の高性能システムの研究開発とは根本的に異なったアプローチで，はじめてメガスケールの高性能計算を達成できるというものである．

メガスケール研究プロジェクトの目的は以下に示す，(1) ハードウェア/ソフトウェア協調による低電力化技術と，(2) 大規模並列タスクの実行モデル構築・利用技術を柱として，種々のコモディティ技術を活用したメガスケールコンピューティングの基盤技術を確立することにある．すなわち，この 2 つの技術を中核としてプロセッサ，コンパイラ，ネットワーク，クラスタ構築，およびプログラミングに関する研究を行い，それらにより 100 万プロセッサ級の汎用メガスケールコンピューティングが実現できることを示すことと，そのプロトタイプとして低電力・高密度大規模クラスタ MegaProto を構築して技術の有効性を実証することが，本プロジェクトの目的である．

2.1 ハードウェア/ソフトウェア協調による低電力化技術

現実的な設置規模でメガスケールのシステムを構築するためには高密度実装が不可欠であるが，そのためにはまずプロセッサの消費電力を極力削減する必要が

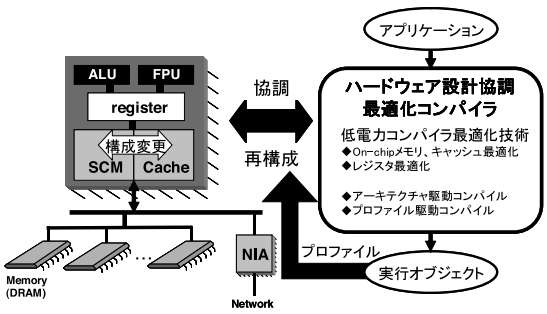


図 1 ハードウェア/ソフトウェア協調低電力化
Fig. 1 Hardware/software cooperative power-aware optimization scheme.

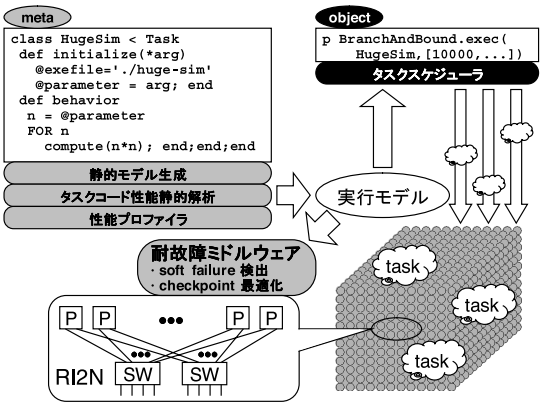


図 2 実行モデルによる実行・環境管理
Fig. 2 Workload modeling and dependability management.

ある．そこで我々は，ハードウェアとソフトウェアの協調によりデータ転送を中心とする最適化を行い，低消費電力と高性能の両立を目指した研究を行っている．

この研究の鍵となる技術は，SCIMA（Software Controlled Integrated Memory Architecture）[19] と呼ぶ，ソフトウェアから可視かつ構成の変更が可能な高速メモリ階層アーキテクチャと，それを利用した性能と消費電力の両面での最適化コンパイル技術である．SCIMA は**図 1** に示すように，通常のキャッシュとの境界が可変である高速メモリ SCM を中心に構成され，配列などのデータは再利用性，アクセスの規則性，容量に応じて SCM あるいは通常のキャッシュ可能な空間に割り付けられる．この割付けをコンパイラが最適化することにより[7),15)]，プロセッサチップと主記憶の間のデータ転送の回数や量を大幅に削減することができ，さらにオンチップメモリのアクセスによる消費電力も削減できる．この結果，実行時間と消費エネルギーの両面で，大きな削減効果が達成される[6),16)]．また同じ発想に基づくメモリアクセスの最適化はキャッシュのみを持つプロセッサにも適用可能であり，特に低電力プロセッサで高い効果が得られることが明らかになっている[9)]．

2.2 大規模並列タスクの実行モデル構築・利用技術

メガスケールのシステムは膨大な計算資源を持つため，ある意味で超大規模の広域分散計算環境に相通じる性格を持っている．すなわち，大きな粒度の並列タスクを単位としたプログラミングと，その実行と環境の管理の大規模な分散化は必然である．しかしその一方で，現実的な設置規模に収められた単一あるいは少数の計算環境の集合体であることを生かし，システム全体を統一的に管理・運用する機構を持つことが求められる．我々は，この分散と統一という背反する課題を解決する鍵が，並列タスクの実行挙動をあらかじめ把握することにあるとの考察に基づき，タスク実行モデルの構築とモデルを利用した実行および環境の管理技術を研究している（**図 2**）．

実行モデルの構築のために，我々は並列タスクの挙動情報を記述可能なタスク並列スクリプト言語 MegaScript を設計した[22)]．この言語ではコンパイラの解析情報からは決定困難なタスク挙動に関する量的情報を与えることができ，静的あるいは動的な実行モデルを高精度に構築することができる．生成されたモデルは，MegaScript で記述された並列タスク実行のスケジューリングのために用いられ，タスク粒度の調整や最適な配置が行われる．

また，実行モデルを利用したシステムの信頼性向上に関する研究も行っている．大規模システムにおける脆弱性の主要因であるネットワークの耐故障性のために，我々はノード間リンクを多重化して高バンド幅と高信頼性を同時に実現する RI2N（Redundant Interconnection with Inexpensive Network）を提案している[18)]．さらにシステムレベルでの耐故障性，特に Soft Failure の検出やチェックポイントの生成・回復戦略の最適化を，実行モデルと実際の挙動との比較などモデルを活用して行う方式を研究している．またモデルを動的に精緻化するための動的なプロファイリング技術についても研究している[24),25)]．

3. 設 計 方 針

3.1 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

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 [2]	2.20	4.40	76.8	57.3
Xeon [12]	3.00	6.00†	85.0	70.6
Pentium 4 [12]	3.20	6.40†	82.0	78.0
Mob. Pentium 4 [12]	3.06	6.12†	70.0	87.4
Mob. Pentium III-M [12]	1.00	1.00	10.5	95.2
TM5800 [27]	0.93	0.93	7.5	124.0
Mob. Celeron [12]	2.40	4.80†	35.0	137.1
Mob. Pentium 4-M [12]	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 [13] (*5)	2.80	5.60†	111.0	50.5
Pentium 4 [13]	3.80	7.60†	115.0	66.1
Mob. Pentium 4 [13]	3.33	6.66†	88.0	75.7
Celeron M [13]	1.50	1.50	24.5	61.2
	0.90	0.90	7.0	128.6
Pentium M [13]	2.10	2.20	21.0	100.0
	1.10	1.10	5.0	220.0
TM8820 [28] (*6)	1.00	2.00†	3.0	666.7

(*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) 3.0〜3.6 GHz 品もリリースされているが，それらの TDP は 13) には記載されていない． (*6) TDP は概算値．

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 のさらに 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．

100 MFlops/W という値自体は，現時点での高性能計算システム，たとえば TOP500 [26] にランクされているほとんどのシステムの性能電力比よりも 1 桁程度高く，きわめて挑戦的な目標のようにも見える．しかし TOP500 リストには，この値をすでにクリアし約 200 MFlops/W という高い値を達成した BlueGene/L [1] という重要な例外が存在する．BlueGene/L はコモディティ技術によるシステムではなく，その高い性能電力比は専用に開発されたチップに負うものであるが，低電力を重要な設計ポイントとしたシステムが TOP500 の第 1 位を占めたことは，我々の主張の正しさを裏付ける 1 つの証拠と考えることができる．

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いわざるをえない．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，TM8820 が上記の目標値 400 W/Flops の 1.7 倍もの数値を達成しているほか，Pentium M も目標値の 1/2 を超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 TM8820 が 667 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．

またこの表が示す興味深い事実として，モバイルプロセッサの性能電力比が (a) の 130 nm 世代から (b) の 90 nm 世代への移行により顕著な改善を示しているのに対し（Intel の場合は約 2 倍，TransMeta では約 5 倍☆），ハイエンドプロセッサの値はほとんど改善

されていない（若干悪化している）ことがあげられる．またこの事実からも，近い将来に Peta-Flops を実現する道はモバイルプロセッサなどの低電力プロセッサの利用に限られることが強く示唆されている．

3.2 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 × 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．

3.3 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[18]．RI2N は，通信メッセージを数 KB の chunk に分割し，それらを複数の Ethernet リンクに動的に分配することにより，ほぼリンク数に相当するだけのバンド幅向上を達成する技術である．また送信可能なリンクを動的に選択して chunk を分配することにより，輻輳したパスを自然に回避できるだけでなく，故障したリンクの回避も自動的に行うことができる．この故障回避機能と chunk の再送制御機能を組み合わせることで，信頼性の面でも優れたネットワークを構築することができる．

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．なおこの構成は GbE を 2 ポート持つ 1U サーバに一見類似しているが，クラスタユニット内の全プロセッサを接続するためのスイッチがユニット内に搭載されていることが本質的に異なっている．また 2 つのポートは別々のスイッチに接続され，ユニット内に完全に独立した 2 系統のネットワークが構成されるため，高い信頼性を確保することができる．

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，実装スペース，消費電力はかなり大きい．たとえば 10 Gbps Ethernet では，4 ポートスイッチの実装スペースが 24 ポートスイッチ 2 個分と同じ程度であり，消費電力も 4 倍程度を要している．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[17]，上記のようにスペース比や電力比でも優れているため，この方法を選択することとした．

この結果 1 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．このような構成は，たとえば Dell の PowerConnect 5324 に用いられているスイッチを用いて実現することができ，スイッチに要するコストを十分小さいものとすることができる．また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．

4. クラスタユニットの設計

前章で述べた設計方針により，MegaProto のシステム構成は**図 3** に示すものとなる．本章では，Mega-

☆ TM8820 と TM5800 の比，あるいは Pentium M（1.1 GHz）と Mobile Pentium 4-M（1.2 Hz，TDP = 20 W[13]）の比．

☆☆ 技術の動向から考えて，おそらく将来についても．

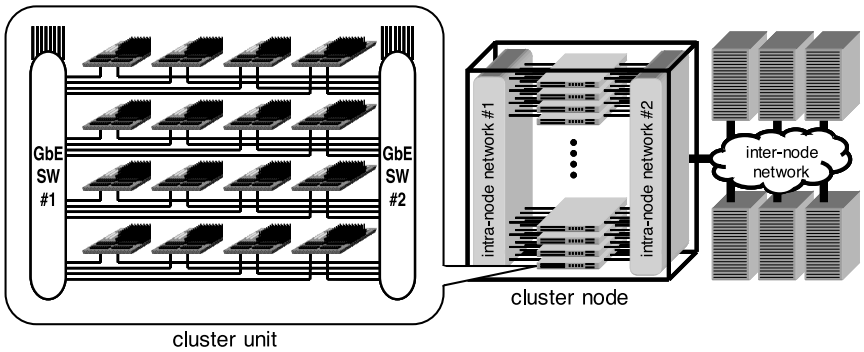


図 3 システム構成
Fig. 3 System configuration.

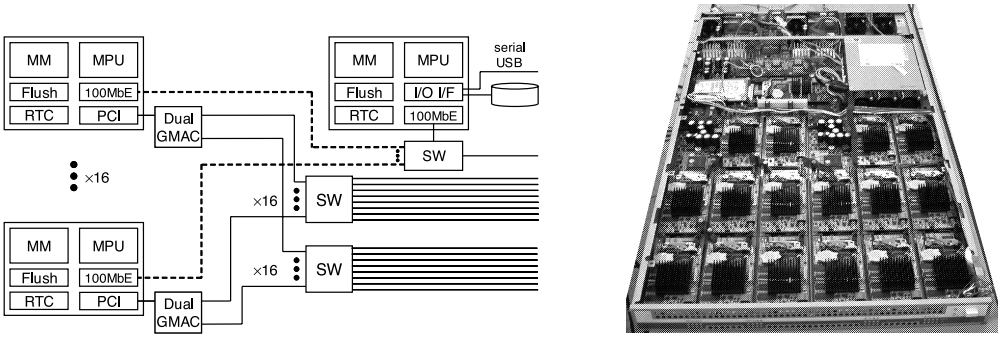


図 4 クラスタユニット
Fig. 4 Cluster unit.

Proto の構成単位であるクラスタユニットの設計について詳細に議論する．

クラスタユニットの構成と外観写真を，**図 4** に示す．クラスタユニットは 432 mm（W）× 756 mm（D）× 44 mm（H）のシャーシに実装され，その約半分（写真では前方）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，TM8820 を用いた第 2 バージョンでは 320 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．

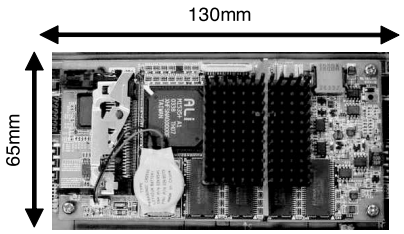
クラスタユニットは大規模クラスタである MegaProto の構成要素であるが，それ自体も小規模なクラスタであると見なすことができる．すなわち各プロセッサはディスクレスの PC として機能し，Linux のソフトウェアスタックや MPI などのミドルウェアが通常のクラスタノードと同様に実装される．またプロセッサ間通信はクラスタユニットの内外で論理的に等価であり，少なくとも論理的には内外の通信を区別する必要はない．

4.1 プロセッサカード

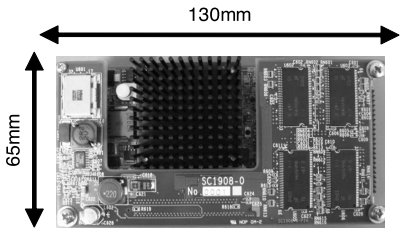
プロセッサカードは低電力プロセッサを中心に構成され，**図 5** に示す千円札よりもやや小さい 65 mm × 130 mm のカード上に，主記憶，フラッシュメモリ，I/O インタフェースなどの周辺回路も実装されている．前章で述べたように，モバイルプロセッサ技術の最新の技術を活用するため，クラスタユニットは共通のマザーボードと交換可能なプロセッサカードから構成されている．

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 TM8820 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 14.9 GFlops であるのに対し，第 2 バージョンでは 1.0 GHz の TM8820 によって 32.0 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 1.34 TFlops となり，きわ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



(a) 第 1 バージョン



(b) 第 2 バージョン

図 5 プロセッサカード
Fig. 5 Processor card.

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

	1st version	2nd version
MPU	TM5800 (0.93 GHz)	TM8820 (1.0 GHz)
Caches	L1 = 64 KB(I) +64 KB(D)	L1 = 128 KB(I) +64 KB(D)
	L2 = 512 KB(D)	L2 = 1 MB(D)
Memory	256 MB SDR-133	512 MB DDR-266
Flush	512 KB	1 MB
I/O Bus	PCI (32 bit, 33 MHz)	PCI-X (64 bit, 66 MHz)

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．すなわち TM8820 は TM5800 よりも低電力であるため，プロセッサカードに割り当てた 10 W の消費電力枠の中から周辺部分により多くの電力を割くことが可能になった．周辺部分改良の第 1 のポイントは，メモリ容量を 256 MB から 512 MB に増やしつつ，SDR-133 を DDR-266 に置き換えてメモリバンド幅を向上させたことである．また第 1 バージョンでは，電力とプロセッサ能力の不足から I/O バスを 32 ビット/33 MHz の PCI とせざるをえず，2 つの GbE リンクを十分に活用することができなかったが，この問題は 64 ビット/66 MHz の PCI-X を第 2 バージョンで採用することで解消している．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．

4.2 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．1 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．

このほか，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．

4.3 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．

4.4 考　　察

本節では MegaProto が，我々の研究の究極の目標である Peta-Flops 級の性能を持つメガスケール計算システムの実現に向けて，どのように位置付けられるものであるかを議論する．

すでに述べたように，我々の設計目標値であったラックあたりのピーク性能 1 TFlops および消費電力 10 kW を単純に外挿すると，ピーク 1 Peta-Flops は 1,000 ラック，10 MW のシステムを構築すれば達成可能ということになる．ここで 19 インチラックの占

☆ 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．

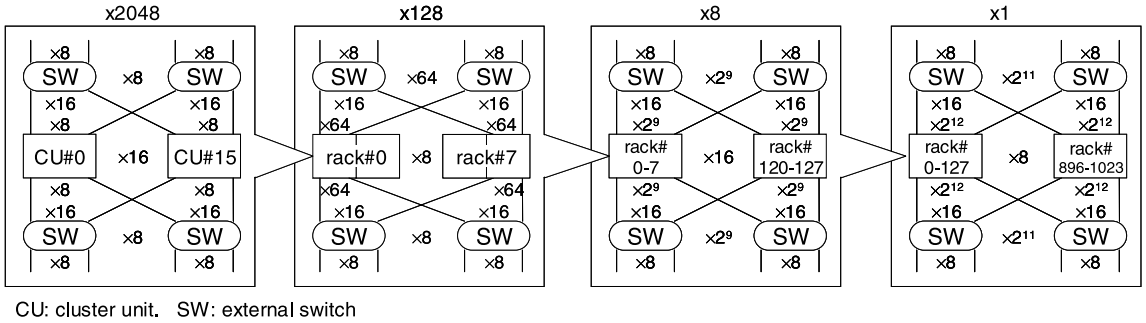


図 6 MegaProto クラスタユニットによる Peta-Flops システムの実現イメージ
Fig. 6 Hypothetical configuration of Peta-Flops system with MegaProto cluster unit.

有床面積を 0.54 m^2（= 600 mm × 900 mm）とすると，1,000 ラックの占有面積は 540 m^2 となり，地球シミュレータのプロセッサノード筐体の占有面積 448 m^2（= 1,000 mm × 1,400 mm × 320）や，ネットワーク筐体を含む面積 557 m^2 にほぼ相当する値となる[8]．一方消費電力は，地球シミュレータが 5.5〜6.0 MW であるのでその 2 倍弱となるが，必ずしも夢想的な数字ではないといえる．

一方 MegaProto の第 2 バージョンは，クラスタユニットのレベルでは 100 MFlops/W を達成しており，ユニット間接続のためのスイッチの設置容積・電力を加味しても，ラックあたり 1 TFlops のピーク性能を得ることは十分可能であると考えられる．たとえば**図 6** に示すような，2 系統の 16 進 8 多重 fat-tree というかなり大規模なネットワークを仮定すると，1 Peta-Flops の達成に必要な 512 K 個のプロセッサ（32 K ユニット）の接続は，以下のように実現することができる．

- ユニット内と同じ 24 ポートのコモディティスイッチを使用．実装密度は 1U あたり 4 スイッチ☆，また消費電力はスイッチあたり 80 W と仮定☆☆．
- 必要なスイッチ個数は，系統あたり下位レベルから順に 16 K，8 K，4 K，2 K であるので，2 系統で総数は 60 K 個．
- 1 ラックに 32 ユニット（512 プロセッサ）とそれらを結合する最下位レベルのスイッチ 32 個（8U）を搭載（合計 40U）．ラックあたりの消費電力は 12.8 KW．システム全体では 1024 ラックからなり，消費電力は 13.1 MW．
- 上位 3 レベルのスイッチ（総計 28 K 個）をラックあたり 128 個（32U）ずつ実装．ラックあたりの消費電力は 10.2 KW．システム全体では 224 ラックからなり，消費電力は 2.3 MW．

上記を総計すると，システムは 1248 ラックで構成され，ラックの占有面積は総計 674 m^2，消費電力総計は 15.4 MW となる．これらの値を地球シミュレータと比較すると，占有面積では 1.2 倍，消費電力では約 3 倍となる．

この値もやはり夢想的なものではないが，現実的にはきわめて実現困難であり，MegaProto の延長線上に Peta-Flops システムを描くためには，解決しなければならない課題がいくつか存在する．その第 1 のポイントは消費電力であり，特に上記の試算値の中でスイッチの消費電力が，クラスタユニット内のものを含めると約 2/3 を占めることが重大な問題である．この要因の 1 つとして，従来のネットワーク技術が低消費電力化をさほど重視していなかったことがあげられるが，情報機器全般における低電力化の技術トレンドに沿って改善されるものと期待できる．また後述のように，スイッチの消費電力は通信の有無や転送量によらずほぼ一定であるという問題もあるが，通信パターンの解析・予測などにより動的にスイッチの電源供給を制御することで，システムレベルでの解決が可能であると考えている．このほか，近い将来に予想される 10 Gbps Ethernet のコモディティ化を利用すれば，クラスタユニット間の接続を 10 G に置換してスイッチ数を大幅に削減することが可能である．

第 2 のポイントは，プロセッサの性能・実装密度・電力効率の改善である．前述のように，130 nm 世代から 90 nm 世代への移行による性能電力比の改善はモバイルプロセッサにおいて顕著であり，この傾向が続けばハイエンド/ミドルレンジに対するモバイルプロセッサの優位性は揺るがないものと考えられる．実際，ITRS による半導体ロードマップ[14] によれば，ハイエンド/ミドルレンジのプロセッサの消費電力増が年

☆ MegaProto の実装と同程度の技術で十分可能である．
☆☆ DELL の PowerConnect 5324 の値．

率 5%強であるのに対し☆，モバイルプロセッサでは 2.6%と見積もられている．したがって，クロック速度や素子密度の向上率が従来と同様にクラスによらず一定とすれば，モバイルプロセッサの優位性はさらに拡大することになる．また消費電力抑制に最も効果的な電源電圧低下も，ハイエンドでの年率 3.3%に対してモバイルでは 4.8%と大きい．さらに携帯・組み込み用途のプロセッサが機能・性能面で向上し，高性能計算への応用が可能になれば，性能電力比が劇的に改善することも予想できる．このほか，実装密度の面で最も効果的なマルチコア技術が，モバイルプロセッサなどに適用されるかどうかは不透明であるが，メガスケール計算では単純な複数コアの集積で十分であり，純技術的にはごく近い将来に実用可能であると考えられる．

第 3 のポイントは，コモディティ部品の大量集積によって構築されたシステムが十分な信頼性を保ちうるかという問題である．この点について我々は，プロセッサ多重化など信頼性向上のみを目的としたハードウェア投資よりも，ソフトウェア技術によって現実的に得られる信頼性に対処する方法が優れていると考え，プロジェクトの重要な項目として研究を進めている．たとえば MegaProto のネットワークである RI2N は，ソフトウェアの管理下で状況に応じて高バンド幅と耐故障性を切り替える技術である．またシステムレベルの故障検出・回復の機構など[24),25)]，ソフトウェアによる耐故障性技術を MegaProto に実装する予定であり，故障が避けられないという前提でのメガスケール計算の実現性を検証することとしている．

5. 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 5.1 節において，重要な評価項目である消費電力の測定環境について述べる．次に 5.2 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[23)] 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 5.3 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．

☆ 2004 年版ではハイエンドプロセッサの消費電力が 200 W 程度で頭打ちになるという予測に修正されているが，これはチップ技術の進歩を反映したものではなく，冷却技術の限界を示したものであると解説されている．

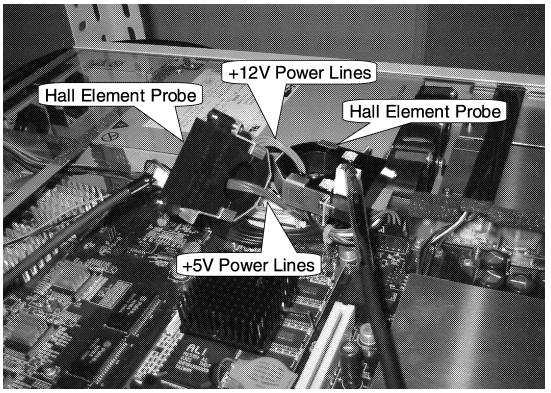


図 7 ホール素子プローブによる電力測定
Fig. 7 Power measurement using hall element probes.

5.1 消費電力測定

MegaProto の評価項目として，消費電力は実行速度と同等の重要性を持っている．特に，プログラム実行中のどの時点で，どのハードウェアが，どの程度の電力を消費するかを知ることは，将来の設計改良や電力・性能最適化コンパイルにとって重要な意味を持つ．したがって，時間的および空間的に高い解像度で消費電力を測定することが必要となる．

この詳細な消費電力測定のために，我々はホール素子を用いた電力測定器を用意した[11)]．この測定器には環状のホール素子プローブが備えられており，**図 7** に示すように電源ラインをプローブ環に通すことで，電気的に接続することなくその電流を擾乱なしに測定することができる．また測定した電流値は，A/D 変換器を経由して 10 μs という高いサンプリングレートで PC などの解析システムに送信することができる．

この測定器を用いて，MegaProto の電源の 100 V AC 入力と，5 V および 12 V の DC 出力の電源電流を測定した．5 V 電源はプロセッサを駆動する電源であり，12 V 電源はネットワーク系を主とする他のデバイスに供給されている☆☆．

5.2 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

☆☆ このほかにプロセッサの周辺回路を駆動する 5 V DC 電源があるが，測定困難な実装となっていることと，供給される電力が他の DC 電源に比べて無視できる程度あるため，測定は行っていない．

☆☆☆ 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超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	52.3	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.

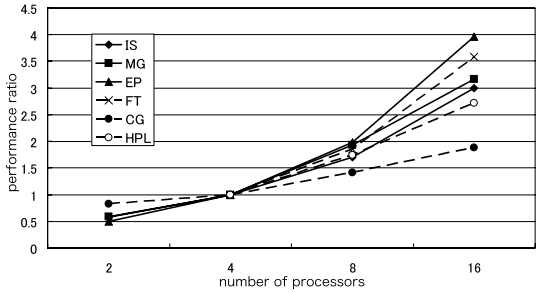


図 8 NPB と HPL の台数効果
Fig. 8 Speedup of NPB and HPL.

た台数効果を**図 8** に示す.

NPB の結果は，多くの高性能クラスタと類似した傾向となっており，クラスタユニットが小規模なクラスタとしても妥当な構成となってい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いえない．この主な理由は，プロセッサあたり 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 9 は，もう 1 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．

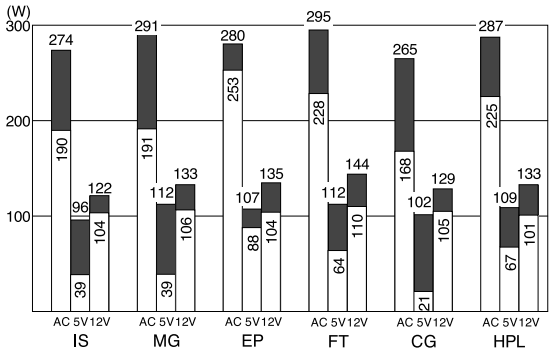


図 9 消費電力の最大値と平均値
Fig. 9 Peak and average power consumption.

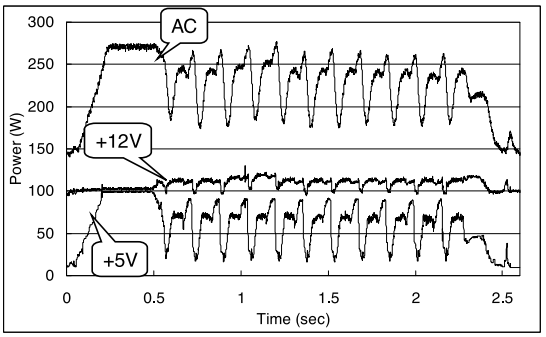


図 10 IS の電力プロファイル
Fig. 10 Power profile of IS.

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，つね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 10** 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

☆ 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.

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あげられる．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，通信の有無にかかわらず「つねに」リンクを駆動し続けることにある☆1．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まった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る．現時点ではこの問題への有効な解は見出せていないが，RI2N の機構を活用した電力削減が可能ではないかと考えている．たとえばプログラム中の通信挙動を実行モデルなどを用いて予測し，要求バンド幅が小さいと判断した場合に複数リンクの一部の電源をオフにすることができれば，性能劣化をともなわない大幅な消費電力削減が可能となる☆2．

5.3 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 [12] 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 11 はデュアル Xeon 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 30 %，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（あるいはそれがゆえに）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．

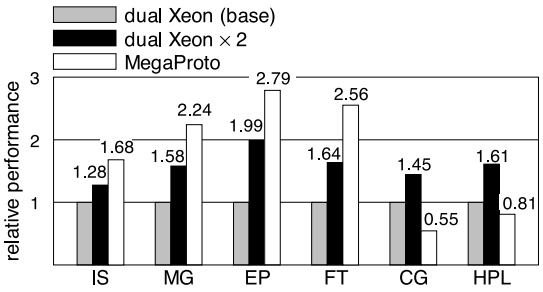


図 11 デュアル Xeon サーバとの性能比較
Fig. 11 Performance relative to dual-Xeon server.

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 にとどまり，HPL でもわず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 TM8820 のピーク演算性能は 130 nm 世代である TM5800 の約 2.1 倍（2.0 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．

6. 関 連 研 究

低電力プロセッサからなるクラスタを高性能計算システムとして用いる発想は，まず最初に Green Destiny により具現化された[29]．Green Destiny の構成要素は，667 MHz の TM5600☆3 と 640 MB のメモリ，および 10 GB のハードディスクを搭載した，高さ 3U のブレードである．このブレードは，19 インチラックに横方向に 24 枚搭載可能であるので，1U あたりの実装密度は 8 プロセッサ，すなわち MegaProto の 1/2 である．ブレードのピーク性能電力比は 35.6 MFlops/W であり，MegaProto 第 1 バージョンの約 2/3，第 2 バージョンの約 1/3 である．また MegaProto と大きく異なる点は，ネットワークがプロセッサあたり 1 ポートの 100 Mbps Fast Ethernet であるため☆4，バンド

☆1 MegaProto のデバイスだけではなく，一般的な性質である．
☆2 MegaProto には，部分的な電源オフ機能が備えられていないので，このアイデアの実現・実証には新たな設計・開発が必要である．
☆3 MegaProto 第 1 バージョンの TM5800 は，このプロセッサの後継版である．
☆4 管理用および代替用のポートもあるが，計算には使われていない．

表 4 他機種との比較
Table 4 Comparison with other systems.

システム	Green Destiny	BlueGene/L	Altix 3000	Earth Sim.	MegaProt(v.1)	MegaProto(v.2)
プロセッサ	TM5600	custom (PowerPC 440)	Itanium 2	custom (vector)	TM5800	TM8820
周波数 [GHz]	0.667	0.7	1.5	0.5	0.93	1.0
プロセッサ数	216	2048	64	16	512	512
消費電力 [kW]	5.2	28.1	12.2	16.0	12.0	12.8
設置面積 [m^2]	0.56	0.84	1.02	1.40	0.54	0.54
ラック容積 [m^3]	1.13	1.64	1.95	2.80	1.08	1.08
ピーク性能						
/proc (GFlops)	0.667	2.8	6.0	8.0	0.93	2.0
/rack (TFlops)	0.14	5.73	0.38	0.13	0.48	1.02
/W (MFlops/W)	27.7	204.0	31.5	8.0	39.7	80.0
/m^2 (TFlops/m^2)	0.26	6.85	0.37	0.09	0.88	1.90
/m^3 (TFlops/m^3)	0.12	3.50	0.20	0.05	0.44	0.95
メモリ容量						
/proc (MB)	640	128	1900	2048	256	512
/rack (GB)	133	256	119	32	128	256
/Flops (B/Flops)	0.94	0.04	0.31	0.25	0.27	0.25

幅がプロセッサあたり 1/20，ピーク性能比でも約 1/7 と，非常に小さいことである．Green Destiny の一般的なベンチマークによる性能は明らかになっていないが，5 章で述べたように，MegaProto の第 1 バージョンにおいても問題によってはネットワーク性能に起因したスケーラビリティ不足が生じており☆，Green Destiny のスケーラビリティには深刻な問題があると考えられる．このように開発時期の差や設計方針の違いにより，Green Destiny と MegaProto の性能差は大きなものとなっているが，N 体問題による ASCI Q との比較では性能電力比で約 3 倍，性能設置面積比では約 14 倍の値を達成するなど，低電力・高密度・高性能計算の可能性を示した意義は大きい．

また高性能計算における低電力化の重要性を強く印象付けた出来事として，BlueGene/L [1] が 2004 年 11 月期の TOP500 [26] の第 1 位を占めたことがあげられる．BlueGene/L は，5.6 GFlops の専用デュアルプロセッサチップを最小単位とし，それを 2 個搭載したドーターボード 16 枚からなるマザーボード（32 チップ，64 プロセッサ，180 GFlops）が，1 筐体に 32 枚搭載される．したがって筐体あたり 2048 プロセッサで 5.7 TFlops という高いピーク性能を達成しつつ，消費電力は約 28 kW に抑えられており，性能電力比は 204 MFlops という非常に優れた値となっている．またリンクあたり 1.4 Gbps の 3 次元トーラスネットワークと，リンクあたり 2.8 Gbps のツリーネットワークを備えている．TOP500 の第 1 位となった 16 筐体（32 K プロセッサ）のシステムは Linpack 性能で 70.7 TFlops を達成しており，ピーク性能との比は 77 %という比較的高い値になっている．BlueGene/L の絶対的な高性能と高い性能電力比は専用チップに負う部分が大きく，コモディティ技術のみで構成された MegaProto との対比は困難であるが，高性能計算分野での低電力化の重要性を強くアピールしたシステムとして注目される．

ここで上記の 2 システムと，2004 年 11 月期の TOP500 第 2 位である NASA Ames Resarch Center の SGI Altix 3000 [21]，第 3 位の地球シミュレータ [8] について，1 ラックの諸元を MegaProto 第 1 および第 2 バージョンと対比したものを**表 4** に示す．なお MegaProto については，4.4 節で述べた 32 クラスタユニットと 32 個のユニット間接続スイッチからなる構成を仮定する．前述のように専用チップで構成された BlueGene/L は，ピーク性能やその電力比，面積比，容積比などにおいて非常に優れた値を示しており，いずれの点でも第 2 位の Altix 3000 を 1 桁あるいはそれ以上凌駕している．ただしメモリ容量が性能に比較して小さく，容量性能比が他のシステムに比べて 1 桁程度小さいことが総合的なシステム性能に影響する可能性がある．

一方 MegaProto の第 2 バージョンは，BlueGene/L には及ばないものの Altix 3000 を大幅に上回る性能指標を示している．すなわちピーク性能と性能電力比は約 2.5 倍，性能面積比と容積比は約 5 倍であり，コモディティ技術をベースとしたクラスタとして非常に優れたシステムであるといえる．またラックあたりのメモリ容量は表に示したものの中で最大であり（Blue-

☆ 前述のように第 1 バージョンでは 1 系統のネットワークのみを使用しているが，バンド幅のピーク性能比はやはり Green Destiny の 7 倍程度である．

Gene/L と同容量），容量性能比も Altix 3000 や地球シミュレータとほぼ同じ値である．地球シミュレータで実証されているように，大規模システムの性能を発揮するには大規模な問題に対応できることが必須であるが，この点でも MegaProto はバランスの良い構成であるといえる．

また，ビジネス計算の分野においても，低電力プロセッサを用いたクラスタやサーバが研究・開発されている．たとえば IBM Austin 研究所による Super Dense Server では，300 MHz と 500 MHz に切替え可能な Ultra Low Voltage Pentium III を用いて，ノードあたり 12 W の低消費電力と，8U で 36 ノード（1U あたり 4.5 ノード）の実装密度を達成している[4]．また Orion Multisystem は，MegaProto の第 2 バージョンと同一コアのプロセッサ TM8800 を 12 個搭載した “Desktop Cluster” や，そのマザーボード 6 枚からなる 96 プロセッサの “Deskside Cluster” を開発し，その低消費電力と高密度実装をセールスポイントとしている[3]．このほか，128 台の PowerPC 750CXe（600 MFlops, 128 MB）を 100 Mbps Ethernet で結合した Argus[5] では，性能電力比で MegaProto 第 1 バージョンにほぼ匹敵する 38.4 MFlops/W を，また性能容積比では第 2 バージョンにほぼ匹敵する 0.72 TFlops/m^3 を，それぞれ達成している．

7. ま と め

本論文では，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，2.0 GFlops の TM8820 を用いた第 2 バージョンとがあり，どちらも 300〜320 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 100 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いっそう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，冷却や電源供給がつね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．

MegaProto の第 1 バージョンは，ハードウェア開発が 2004 年 3 月に完了し，現在 2 クラスタユニット（32 プロセッサ）を対象として種々の性能評価作業を行っている．また第 2 バージョンの設計はすでに完了しており，2005 年の第 2 四半期に 20 ユニット（320 プロセッサ）のシステムが完成する予定である．このように大規模で，高性能（ピーク 640 GFlops）かつ低電力（6.4 kW）のシステムはほとんど例がなく，我々の基本主張である低電力・高性能計算の有効性を実証できるものと強く期待している．

今後の緊急の課題は，第 2 バージョンのシステム構築と性能評価である．特に第 1 バージョンの評価で明らかになった，メモリ容量・性能やネットワーク性能に起因するスケーラビリティの不足が，これらを増強した第 2 バージョンでどのように改善されるかが重要なポイントである．また本論文では報告することができなかった複数クラスタユニットを用いた性能評価や，プロジェクトで開発している各種ソフトウェアの実装と評価も，短期的な課題としてあげられる．

一方中長期的な課題は，究極的な目標である PetaFlops 級のメガスケール計算に向けて，より進化した高性能・低電力・高密度システムの構成を検討することである．特に 4.4 節や 5.2 節で議論したネットワークの消費電力削減は重要な課題であり，コモディティネットワーク技術の動向や各種の低電力化技術をふまえつつ，高性能計算に適合した省電力型ネットワークを提案することを目指している．

謝辞 技術的な支援をいただいた日本 IBM 社の技術スタッフに謝意を表する．本研究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．

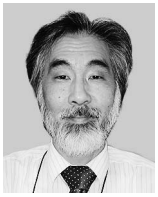
参 考 文 献

1) Adiga, N.R., et al.: An Overview of the BlueGene/L Supercomputer, *Proc. Supercomputing 2002* (2002).
2) Advanced Micro Devices, Inc.: *AMD Athlon XP Processor Model 10 Data Sheet* (2003).

3) Costa, D.: Orion Puts a Cluster on Your Desktop (2004). http://www.workstationplanet.com/features/article.php/3437011
4) Felter, W.M., et al.: On the Performance and Use of Dense Servers, *IBM J. R & D*, Vol.47, No.5/6, pp.671–688 (2003).
5) Feng, X., Ge, R. and Cameron, K.W.: Supercomputing in 1/10 Cubic Meter, *Proc. Intl. Conf. Parallel and Distributed Computing and Networks* (2005).
6) 藤田元信，田中慎一，近藤正章，中村 宏：ソフトウェア制御オンチップメモリにおけるスタティック消費電力削減手法，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.SIG11 (ACS7), pp.219–228 (2004).
7) 藤田元信，近藤正章，中村 宏：ソフトウェア制御オンチップメモリ向け自動最適化コンパイラの提案，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.SIG1(ACS4), pp.77–87 (2004).
8) 幅田伸一，横川三津夫，北脇重宗：地球シミュレータのハードウェア，情報処理，Vol.45, No.2, pp.116–121 (2004).
9) 堀田義彦，佐藤三久，朴 泰祐，高橋大介，中島佳宏，高橋睦史，中村 宏：プロセッサの消費電力測定と低消費電力プロセッサによるクラスタの検討，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.SIG11 (ACS7), pp.207–218 (2004).
10) Hotta, Y., et al.: MegaProto: A Prototype of the Ultra Low-Power Mega-Scale System, *Proc. COOL Chips VII* (2004).
11) Hotta, Y., Sato, M., Boku, T., Takahashi, D. and Takahashi, C.: Measurement and Characterization of Power Consumption of Microprocessors for Power-Aware Cluster, *Proc. COOL Chips VII* (2004).
12) Intel Corp.: Datasheets of the following Intel processors on 0.13 micron process: Xeon (298642-006), Pentium 4 (298643-010), Mobile Pentium 4 (253028-001), Mobile Pentium III-M (298340-006), Mobile Celeron (251308-005) and Mobile Pentium 4-M (250686-007) (2003).
13) Intel Corp.: Datasheets of the following Intel processors on 90 nm process: Xeon (302355-001), Pentium 4 (303128-004), Mobile Pentium 4 (302424-002), Celeron M (300302-003) and Pentium M (302189-004) (2004).
14) International Technology Roadmap for Semiconductors: Executive Summary (2003 Edition) (2003). http://public.itrs.net/Files/2003ITRS/Home2003.htm
15) 近藤正章，中村 宏，朴 泰祐：SCIMA における性能最適化手法の検討，情報処理学会論文誌：ハイパフォーマンスコンピューティングシステム，Vol.42, No.SIG12 (HPS4), pp.37–48 (2001).
16) Kondo, M. and Nakamura, H.: Reducing Memory System Energy by Software-Controlled On-Chip Memory, *IEICE Trans.*, Vol.E86-C, No.4, pp.550–588 (2003).
17) Matsuoka, S.: You Don't Really Need Big Fat Switches Anymore—Almost, *IPSJ SIG Notes*, 2003-ARC-154, pp.157–162 (2003).
18) Miura, S., Boku, T., Sato, M. and Takahashi, D.: RI2N—Interconnection Network System for Clusters with Wide-Bandwidth and Fault-Tolerancy Based on Multiple Links, *Proc. Intl. Symp. High Performance Computing 2003*, pp.342–351 (2003).
19) 中村 宏，近藤正章，大河原英喜，朴 泰祐：ハイパフォーマンスコンピューティング向けアーキテクチャSCIMA，情報処理学会論文誌：ハイパフォーマンスコンピューティングシステム，Vol.41, No.SIG5 (HPS1), pp.15–27 (2000).
20) Nakashima, H., et al.: MegaProto: A Low-Power and Compact Cluster for High-Performance Computing, *Proc. WS High-Performance Power-Aware Computing* (included in *Proc. IPDPS 2005*) (2005).
21) National Aeronautics and Space Administration: NAS Computing Resources—Columbia Supercomputer (2004). http://www.nas.nasa.gov/Users/Documentation/Altix/hardware.html
22) 大塚保紀，深野佑公，西里一史，大野和彦，中島浩：タスク並列スクリプト言語 MegaScript の構想，先端的計算基盤システムシンポジウム SACSIS 2003，pp.73–76 (2003).
23) Petitet, A., Whaley, R.C., Dongarra, J. and Cleary, A.: HPL—A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers (2004). http://www.netlib.org/benchmark/hpl/
24) Sakae, Y., Matsuoka, S., Sato, M. and Harada, H.: Towards Dynamic Load Balancing Using Page Migration and Loop Re-partitioning on Omni/SCASH, *Proc. 4th European WS OpenMP* (2002).
25) 高宮安仁，松岡 聡：ユーザー透過な耐故障性を実現する MPI へ向けて，並列処理シンポジウム JSPP2002，pp.217–224 (2002).
26) TOP500 Team: TOP500 list for November 2004 (2004). http://www.top500.org/lists/2004/11/
27) Transmeta Corp.: *Crusoe Processor Product Brief—Model TM5800* (2003).
28) Transmeta Corp.: *Transmeta Efficeon TM8800*

Processor—Product Sheet (2004).
29) Warren, M., Weigle, E. and Feng, W.: High-Density Computing: A 240-Node Beowulf in One Cubic Meter, *Proc. Supercomputing 2002* (2002).

(平成 17 年 1 月 24 日受付)
(平成 17 年 4 月 26 日採録)



中島　浩（正会員）

1981 年京都大学大学院工学研究科情報工学専攻修士課程修了．同年三菱電機（株）入社．推論マシンの研究開発に従事．1992 年京都大学工学部助教授．1997 年豊橋技術科学大学教授．並列計算機のアーキテクチャ等並列処理に関する研究に従事．工学博士．1988 年元岡賞，1993 年坂井記念特別賞受賞．情報処理学会計算機アーキテクチャ研究会主査，同論文誌コンピューティングシステム編集委員長等を歴任．IEEE-CS，ACM，ALP，TUG 各会員．

中村　宏（正会員）

1990 年東京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 年より東京大学先端科学技術研究センター助教授．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．高性能・低消費電力プロセッサのアーキテクチャ，ハイパフォーマンス・ディペンダブルコンピューティング等の研究に従事．情報処理学会より論文賞（1993 年度），山下記念研究賞（1994 年度），坂井記念特別賞（2001 年度），各受賞．IEICE，IEEE，ACM 各会員．

佐藤　三久（正会員）

1982 年東京大学理学部情報科学科卒業．1986 年同大学大学院理学系研究科博士課程中退．同年新技術事業団後藤磁束量子情報プロジェクトに参加．1991 年通産省電子技術総合研究所入所．1996 年新情報処理開発機構並列分散システムパフォーマンス研究室室長．2001 年より，筑波大学システム情報工学研究科教授．同大学計算科学研究センター勤務．理学博士．並列処理アーキテクチャ，言語およびコンパイラ，計算機性能評価技術，グリッドコンピューティング等の研究に従事．IEEE，日本応用数理学会会員．

朴　泰祐（正会員）

1990 年慶應義塾大学大学院理工学研究科電気工学専攻後期博士課程修了，工学博士．1988 年慶應義塾大学理工学部物理学科助手．1992 年筑波大学電子・情報工学系講師，1995 年同助教授．2004 年同大学大学院システム情報工学研究科助教授，2005 年同教授，現在に至る．同大学計算科学研究センター勤務．超並列計算機アーキテクチャ，ハイパフォーマンスコンピューティング，クラスタコンピューティング，グリッドに関する研究に従事．2002 年度および 2003 年度情報処理学会論文賞受賞．日本応用数理学会，IEEE-CS 各会員．

松岡　聡（正会員）

1986 年東京大学理学部情報科学科卒業．2001 年東京工業大学学術国際情報センター教授，2002 年国立情報学研究所客員教授併任．博士（理学）（東京大学）．高性能システム，並列処理，グリッド計算，クラスタ計算機等．1996 年度情報処理学会論文賞，1999 年情報処理学会坂井記念賞受賞．ACM OOPSLA'2002，IEEE CCGrid2003 を含む種々の国際会議のプログラム・大会委員長を歴任．2003 年よりグリッドの国家プロジェクトの NAREGI プロジェクトにも従事．グリッド国際標準化団体 Global Grid Forum の Area Director．

高橋 大介（正会員）

1997 年東京大学大学院理学系研究科情報科学専攻博士課程中退．同年同大学大型計算機センター助手．1999 年同大学情報基盤センター助手．2000 年埼玉大学大学院理工学研究科助手．2001 年筑波大学電子・情報工学系講師．2004 年同大学大学院システム情報工学研究科講師．博士（理学）．並列数値計算アルゴリズムに関する研究に従事．1998 年度情報処理学会山下記念研究賞，1998 年度および 2003 年度情報処理学会論文賞各受賞．日本応用数理学会，ACM，IEEE，SIAM 各会員．

堀田 義彦（学生会員）

2003 年 3 月筑波大学第三学群情報学類卒業．現在，同大学大学院システム情報工学研究科在学中．HPC クラスタの低消費電力化，組み込み CMP 向け OpenMP による消費電力最適化の研究に従事．