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

2006－ARC－167 (27)
2006－HPC－105 (27)
2006／2／28

電力制約下での高性能計算機クラスタ構成手法

池田佳路† 近藤正章† 中村宏†

計算機クラスタはその利点であるコストパフォーマンスを向上するには高性能なプロセッサを高密度に実装することが不可欠である．従来の計算機クラスタ構成手法では，最大消費電力が，冷却限界に対する許容電力 (TDP:Thermal Design Power) を超えない HW 実装で設計されてきた．しかし実際には TDP を消費することなく，性能の余裕を残した状態で稼動している時間が多い．本研究では，TDP を保証しない構成で実装されたクラスタに対し，動作周波数を変化させることにより，TDP 内で動作する最も性能の高い構成を選択し，従来の構成手法によるクラスタよりも高い性能を実現する計算機クラスタ実装方法を提案し，その初期検討を行った．

A High Performance Cluster System Design with Power Constraint

Yoshimichi IKEDA,† Masaaki Kondo† and Hiroshi Nakamura†

Compact and dense packaging is indispensable for cluster systems to achieve high performance/cost ratio. Packaging of cluster systems, so far, is designed to satisfy the restriction that that is, peak power consumption does not exceed a given TDP (Thermal Design Power) derived from cooling limitation. However, practical power consumption seldom reach TDP, and thus cluster systems operates with allowance for power consumption most of the time. Therefore, in this paper, we propose a new implementation scheme of cluster systems. In the proposed scheme, although theoretical peak power consumption exceeds TDP, practical power consumption is still below TDP by adjusting supply voltage and clock frequency of each node, and thus effective performance gets higher than conventional implementation.

1. はじめに

近年，大規模科学技術計算分野において重要な位置を占めるのが計算機クラスタである．計算機クラスタは，複数台の汎用計算機をネットワークによって接続し，並列処理を行うことで，スーパーコンピュータ並の性能を実現するシステムである．スーパーコンピュータは専用のハードウェアを開発することで高性能を得られる反面，高コストである．一方，計算機クラスタは比較的安価な汎用品を用いることで，高いコストパフォーマンスを実現することができるため，現在では広く用いられている．TOP500[1] によれば，世界最高性能 500 台の計算機システムの内，360 台が計算機クラスタにより構成されており，この点からも計算機クラスタの重要性が伺える．

並列処理を行う計算機クラスタシステムは，ノード数を増加させることで性能を向上させることができる．しかし，設置面積や電源などの制限から，必然的にノード数の上限は決定される．そこで，高性能ではあるが消費電力の高いプロセッサを搭載するノードを複数用いるよりも，プロセッサあたりの性能はそれほど高くないが省電力なプロセッサを多数用い，高密度に実装することで台数効果により性能向上を図る方が，高性能なシステムを構築できる場合が多い．この点に着目し，“Green Destiny”[2] や “Mega Proto”[3] は，省電力プロセッサを高密度に実装したクラスタシステムを構築し，既存のクラスタシステムと比較して，設置面積 (体積) あたりの性能や電力あたりの性能が優れたクラスタシステムであることを示している．また，IBM で開発された BluGene/L[4] は，PowerPC をベースとしたプロセッサを超高密度に実装し，非常に高い性能を省スペースで実現している．

このように，低消費電力プロセッサの高密度実装により，面積や消費電力効率の良いクラスタシステムを構築できることが示されているが，高密度実装にも限界がある．計算機システムを構築する上では，単位体積当たりの冷却能力には限界が存在するため，その冷却能力を越えるような実装は不可能なためである．例えば，通常の冷却システムのもとでは，19 インチラックに収納する 1U サイズの筐体では 300W 程度の消費電力が限界と言われている．したがって，クラスタシステムを設計する際には，冷却システムが許容する発熱量に対応した許容消費電力を上限とし，それを越え

† 東京大学先端科学技術研究センター
Research Center for Advanced Science and Technology (RCAST), The University of Tokyo

ないように設計する必要がある．

ここで，プロセッサを含む半導体チップには，放熱面での消費電力の最大値を定めた熱設計消費電力 (*TDP: Thermal Design Power*) が存在し，一般的にはそのチップのピークの消費電力に相当する．半導体チップで構成される計算機システムは，このTDPを基に冷却システムが設計されており，したがって，システムはピークの電力消費にも耐えられるように設計されていることになる．しかし，実際に製造されたプロセッサチップが，このピーク電力を消費することはまれである．また消費電力は実行するアプリケーションに応じて大きく異なり，例えばキャッシュミスが頻発する場合などは，非常に電力消費が少なくなる場合もある[5]．そのため，TDPをもとに設計されているシステムは，通常はその冷却能力を最大限に発揮しているわけではない．

そこで本稿では，TDPで考えた場合には冷却能力を越えてしまうような数のプロセッサを搭載する構成を利用し，アプリケーションの特性にあわせて消費電力制御を行うことで，冷却能力を最大限に活用して性能を最適化する「電力制約下での高性能計算機クラスタ構成手法」を提案する．本提案手法は，システムの電力を動的に監視し，冷却能力により決定された許容消費電力を越えそうな場合には，動的電源電圧変更手法 (DVS: Dynamic Voltage Scaling) により消費電力を制御しつつ，実行させるものである．本稿では，提案手法を適用したプロトタイプのクラスタシステムを構築し，従来のクラスタ構成に比べ高性能が達成できることを示す．

本稿の構成は以下の通りである．次章では，提案するクラスタ構成手法について述べ，3章でその有効性に関して議論する．4章では電源管理のアルゴリズムを述べ，5章でその評価結果を示す．6章でまとめと今後の課題について述べる．

2. 実効電力に基づくクラスタ構成手法

2.1 実効消費電力

多くの計算機システムにおいて，最も電力を消費する部分はプロセッサチップである．プロセッサの消費電力は，プログラムの性質に依存して異なる．例えば，並列処理プログラムにおいて，通信と演算の比率を対象に考えると，通信が多い場合には，プロセッサはストールしている時間が長くなり，実効消費電力が低下する．逆に通信頻度が少ない場合では，プロセッサは演算処理をずっと行うことができ，結果として実効消費電力が増大することが多い．実際に，Intel Pentium Mを搭載したPCを8台用いたクラスタシステムにおいて，ベクトル積を並列に計算するプログラムにおいて，演算と通信の割合を変化させた場合の，消費電力を図1に示す．なお，実験環境の詳細については，3.1節で述べる．

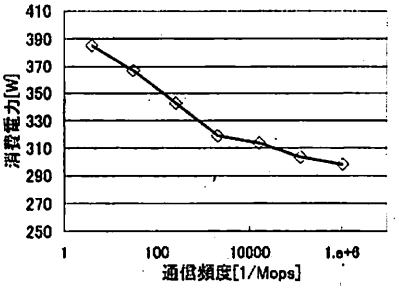

図1 通信頻度に対する消費電力の違い

図より，通信頻度に依存して，実行時の消費電力が大きく変化していることがわかる．通信頻度が小さいアプリケーションでは8台で400W程度であり，一台あたり50Wと比較的高い消費電力となっている．一方通信頻度が高く，通信にかかる時間が多い場合では，消費電力が300W程度にまで低下し，1台あたりでは37W程度まで低下する．

上記の結果より，TDPをもとに設計されているシステムにおいては，システムの許容電力に対して，余裕を残した電力しか消費しない場合が多いことがわかる．

2.2 実効電力に基づくクラスタ構成

従来の計算機クラスタを構築する場合は，ピークの消費電力が許容消費電力を超えない範囲でのハードウェアを用いて構成されている．これに対し，本稿で提案する手法では，全プロセッサがピークで動作した場合の総消費電力が許容電力を越えるようにハードウェアを構成し，実効消費電力を許容電力内に抑えることで，冷却能力を最大限に活用し，さらなる高性能の達成を目指す．

例として，前節で用いたPentium Mを搭載するPCを複数台用いて計算機クラスタシステムを実装する場合を考える．種々の実験の結果，本PCシステムの消費電力は60Wを越えることはなく，TDPはおよそ60W程度であると考えられる．許容電力が300Wであるユニットを仮定し，そのユニットにPCを複数台実装しようとする場合，TDPが60WのPCであれば，5台を実装するのが限度となる．しかし，実効電力に基づくクラスタ構成では，当該PCを例えば8台実装する．当然，各PCがピークに近い電力を消費すると許容電力を越えてしまうが，前節の図1の結果より，アプリケーションによっては8台がすべて動作した場合でも，実行消費電力が許容電力内に収まる場合が存在する．この場合，8台で処理を行うことで，5台で実行する場合に比べ高性能が期待できる．

逆に，8台すべて動作した場合に許容電力を越えてしまう可能性があるアプリケーションも存在するため，実効電力を管理する必要がある．実効電力が許容電力

表 1 評価環境

M/B	Commell LV673 - i915GM + ICH6M chip-set - Gb Ethernet x 2 12V CPU 電源駆動
Processor	Pentium M 760 (Max 2GHz, FSB533MHz)
Memory	DDR2-SDRAM 1GB

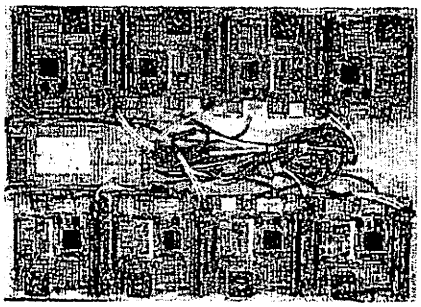

図 2 実験環境

を越える場合は，(1) ノード数を縮小して実行する，または (2)DVS により，周波数／電源電圧を低下させて実行する，のどちらか，あるいは両者を組み合わせにより，許容電力内で最高の性能が出せる条件で実行するよう制御し，効率的に実行する．

3. 実効電力に基づく構成手法の有効性

3.1 評価環境

提案する実効電力に基づくクラスタ構成手法の性能を評価するため，実際にクラスタシステムを構築し評価を行う．構築したクラスタシステムの各ノードの仕様を表 1 に示す．Pentium M 760 プロセッサを用いた PC を用い，これを 8 台 Gb Ethernet で接続しクラスタシステムを構築する．各ノードはディスクを持たず，外部の NFS サーバを用いたディスクレスシステムである．OS は Linux kernel-2.6.11 を用い，cpufreq によりソフトウェア上から周波数・電源電圧が制御可能である．表 2 に，設定可能な周波数および電圧のセットを示す．

各ノードのマザーボードのサイズは，17cm 四方であり，8 台を 2 列に並べて配置すると，ほぼ 19 インチラックの 1U のサイズになる (図 2)．ディスクを接続しない場合，マザーボードやメモリを含めたノードあたりのピーク消費電力は約 60W 程度であり，8 台分を合計すると 480W となり，実際に通常の 1U で許容される 300W 程度の消費電力を越えるシステムとなる．

ベンチマークプログラムとしては，High Perfromance Linpack Benchmark (HPL)，NAS Parallel Benchmark(NPB) の中から EP，姫野ベンチ (Himeno-Bench) のプログラムを用いた．消費電力の

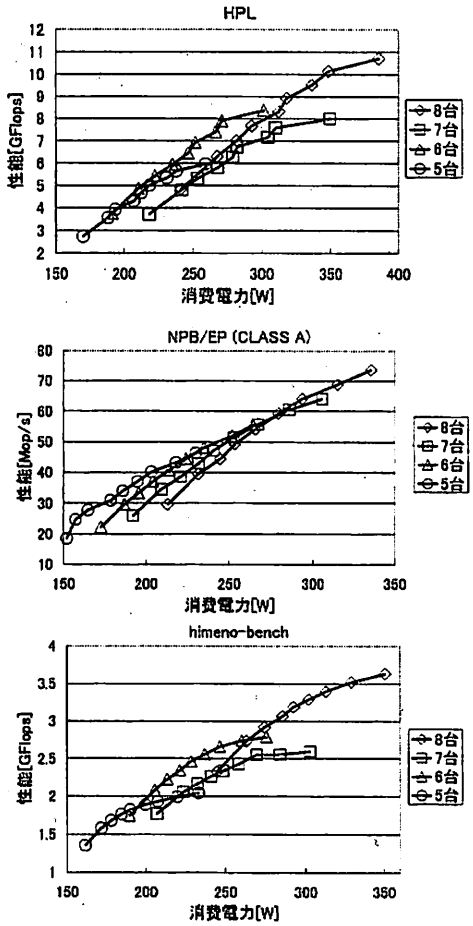

図 3 各ベンチマークにおける消費電力に対する性能

測定には，シナジェティック社製 CT-30000 を用いた (詳細については文献[5] を参照)．

3.2 評価結果

実際に提案するクラスタ構成手法を用いた場合に，どの程度性能向上の可能性があるかを調べるために，HPL，NPB の EP，Himeno-Bench において，ノード数を 4~8 台，周波数を表 2 の全通りで変えつつベンチマーク実行時の性能と電力を測定した．なお，処理に用いるノード数，クロック周波数は，アプリケーションの開始から終了まで固定とした．

結果を図 3 に示す．図は，あるノード台数，およびある周波数で実行した場合の消費電力／性能の関係をプロットしたのもである．なお，同じ台数の場合が線で結ばれている．図 3 より，許容電力を 300W と仮定すると，従来型の構成手法ではどのアプリケーションにおいても 300W を越えないことが保証される，ノード数が 5 台の構成を採用せざるを得ない．従って最高

表 2 Pentium M 760 プロセッサの周波数と電源電圧の関係

Clock (GHz)	2.00	1.86	1.73	1.60	1.46	1.33	1.20	1.06	0.80
Core Vdd (V)	1.356	1.308	1.260	1.228	1.196	1.164	1.132	1.084	0.988

表 3 許容電力を 300W とした場合の性能

ベンチマーク	HPL	NPB/EP	himeno
提案手法	8.37[GFlops]	64.2[Mop/s]	3.28[GFlops]
従来手法	5.98	46.4	2.04
性能比	1.40	1.38	1.60

性能はノード数 5 台，周波数 2.0GHz の場合となる．一方，提案手法では，消費電力 300W 以下で最も性能の高いノード台数および，周波数／電源電圧の組で各プログラムを実行する．図 3 を見ると，HPL では 6 台の 2GHz の場合，NPB-EP では 8 台の 1.7 GHz の場合，Himeno-Bench では 8 台の 1.6GHz で実行する場合が，許容電力内で最も性能の高い構成となっている．従来型，および提案するクラスタ構成手法での最高性能を，表 3 に示す．

表 3 より，提案手法により，従来型に比べ 1.38 倍から 1.6 倍の高い性能が得られる可能性がある．このことから，実効電力に基づくクラスタ構成手法は，クラスタの高性能化において非常に有効な手法であると考えられる．

4. 動的電力モニタリングによる実効電力制御

4.1 概　　要

2.2 節において，実効消費電力に基づくクラスタ構成手法を述べ，従来型のクラスタに比べ高い性能を得られる可能性があることを 3.2 節の評価で示した．ただし，本手法は，アプリケーションに応じて最適なノード数／周波数を決定しなければならないため，プロファイリングなどにより，あらかじめプログラム実行時の消費電力を予測した上で最適な構成を選択し，実行する必要がある．

しかし，並列処理を行うアプリケーションでは，問題の分割などの効率から，ノード数の選択肢にはあまり自由度がないことが多い．また，実行時に動的にノード数を変更し，実効消費電力を制御することは一般的には難しいと考えられる．

そこで本稿では，ノード数はシステムに搭載された最大の構成を固定的に選択し，実効消費電力に応じて周波数／電源電圧のみを動的に変更し，許容電力内で最も高い周波数で動作させることで高性能化を図る手法について検討する．ノード数を変更しないため，最適化の余地が減るが，プロファイリングなどを行う必要がなく，実効消費電力に基づくクラスタ構成手法の運用方法として有用であると考えられる．

4.2 電力管理手法

通常のクラスタシステムの場合，1U などのあるユニット単位で電源が搭載され，また許容される電力などもそのユニット単位で定められるのが普通である．そこで，電源より定期的に消費電力の値を取得し，その値に応じて該当ユニット内のプロセッサの周波数／電源電圧を制御することで，動的に実効消費電力を許容電力内に納める．

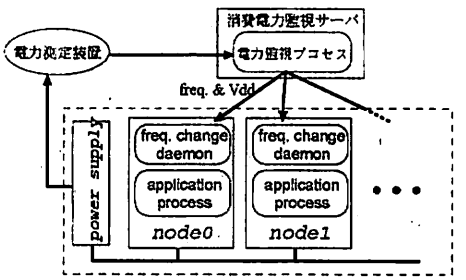

図 4 電力管理システム

図 4 に，電源管理システムの概要を示す．ユニットの電源の消費電力を測定しつつ，電力値を外部の消費電力管理サーバに送り，消費電力管理プロセスが，現在の電力と許容電力の値を比較し，ユニットの消費電力が許容電力を上回った場合は，周波数の変更命令をユニット内の各ノードに対し送信する．各ユニットには，自身の周波数／電源電圧を変更するデーモンを起動しておくことで，消費電力管理サーバからの指示により実際に変更を行う．

なお，消費電力が許容電力から外れた場合に，電力制御のための指示を送信してから実際に周波数が変更されるまで，ある程度の時間がかかる．したがって，瞬間的には許容電力以上の電力で動作してしまうが，これは熱に対する影響が小さいため無視できるものと考える．ここで，本稿では電源に関しては許容消費電力に対し，ある程度の余裕を持った実装を行っていることを仮定し，許容される消費電力を決定する要因は熱密度であることを前提にしている．高密度実装のクラスタを構築する上での制限は，熱密度であることから，この仮定は妥当であると考えられる．

4.3 電力制御アルゴリズム

消費電力管理プロセスが，ユニットの消費電力を監視しつつ，どのように消費電力を周波数および電源電圧を変更するかの戦略が重要となる．そこで，できる限り許容消費電力の範囲内でできるだけ高い周波数を選択することを目的に，以下の電力制御アルゴリズムを用いる．

(1) 消費電力の上限の閾値 (許容電力に相当) と下限の閾値を与える
(2) 各ノードの周波数を最も低い周波数に設定し，

```
freq = Freq_min;
set_freq_all_nodes(freq);

(invock application)

while(application_is_running) {
   W = get_power();
   if (W >= W_max_threshold) {
      freq--;
      set_freq_all_nodes(freq);
   }
   else if (W =< W_min_threshold) {
      freq++;
      set_freq_all_nodes(freq);
   }
}
```

図 5 周波数制御アルゴリズム

アプリケーションの処理を開始する

(3) 実行中のユニットの消費電力を測定／監視する

(4) 与えられた上限の閾値を上回る消費電力を観測した場合は，全ノードの周波数／電圧を 1 段階下げる

(5) 逆に与えられた下限の閾値を下回る消費電力を観測した場合は，全ノードの周波数／電圧を 1 段階上げる

(6) (3) にもどる

上記のアルゴリズムをまとめたものを，図 5 に示す．

5. 動的実効電力制御手法の評価

本節では，4.2 節の電力管理システム，および 4.3 節のアルゴリズムを実際のクラスタシステムに実装し，アプリケーション実行中に動的に周波数／電源電圧を変更することで，提案手法の有効性を示す．

5.1 評価環境および評価条件

評価環境は 3.1 節で述べたクラスタシステムに対し，新たに電力管理サーバを追加し評価を行う．また，ベンチマークプログラムは HPL，Himeno-Bench，および NPB 中のカーネルベンチマーク全てを用いる．

また，アルゴリズム中の閾値は上限 350W,300W,250W とし，下限は各上限に対して 10W の幅を取り，それぞれ 340W,290W,240W として評価を行った．

従来手法と比較を行うため，アプリケーション毎に，8 ノード使用時で周波数を固定して全周波数で実行し，実行電力が許容電力を超えない範囲で最高の周波数の場合を選択し，比較対象として用いる．

なお，今回実装した電力管理システムにおいて，電力監視ノードが消費電力の閾値からの逸脱を観測してから，8 ノード全ての周波数が変更されるまでの時間を計測したところ約 16ms であった．そこで，消費電力監視サーバが対象ユニットの電力を監視する時間間隔を 20ms に設定して評価を行う．

5.2 評価結果

提案する実効電力に基づくクラスタ構成手法に電力管理手法を適用して評価した場合の結果を図 6 に示す．図は，比較対象として選択されたものに対する相対性能比を表している．

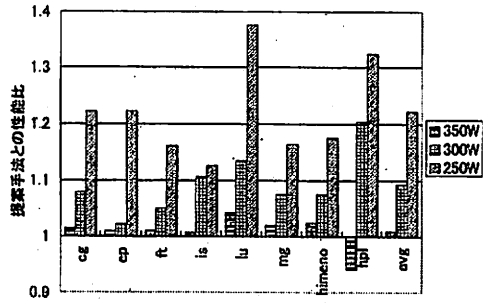

図 6 従来手法と提案手法の性能比

表 4 比較対象として選択された周波数 (GHz)

閾値	cg	ep	ft	is
350W	1.86	1.86	1.86	1.86
300W	1.33	1.73	1.33	1.60
250W	0.80	1.06	0.80	1.06
閾値	lu	mg	himeno	hpl
350W	1.86	1.73	1.86	1.86
300W	1.46	1.33	1.46	1.20
250W	0.80	0.80	0.80	0.80

結果より，提案する手法により，上限の閾値 (許容消費電力)350W の HPL の場合を除いた全ての場合で性能が向上しているのがわかる．また，許容消費電力が低い場合のほうが，提案手法の性能向上率が高いことがわかる．これは，従来型のクラスタ構成では，多くの時間は低い実効消費電力で動作しているにも関わらず，瞬間的に電力がピークになような動作をした場合，ピーク時の消費電力が許容消費電力を上回らないように設計する必要があるため，定常時でも低い周波数で動作せざるを得ないためである．一方，提案手法では，動的に周波数／電源電圧を制御するため，効率的な実行ができた結果である．平均で見ると，許容消費電力 350W の場合で 0.8%，300W の場合で 9.3%，250W の場合で 22.1%の性能向上を達成している．

次に，それぞれのアプリケーション実行時，各周波数が処理に用いられた時間の割合を図 7 に示す．また，比較対象として選択され，固定して動作した周波数の値を表 4 に示す．

これらの結果をみると，許容消費電力 350W においては，比較対象で選択される固定周波数が 1.86GHz と高く，提案手法は 2GHz でほとんどの時間動作しているものの，そのアドバンテージは小さい．一方，許容消費電力 300W，250W の場合では，比較対象で選択される固定周波数よりも，提案手法では高い周波数で動作する時間が長いため，許容消費電力 350W の

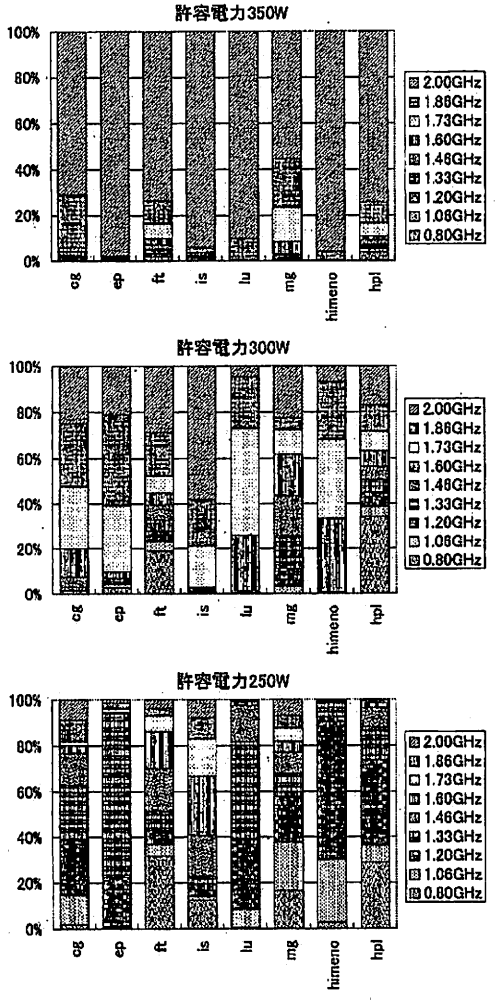

図 7　各ベンチマークにおける消費電力に対する性能

場合に比べて提案手法の性能向上率が高い．しかし，FT，HPL のベンチマークでは，比較対象で選択された周波数より低い周波数で動作している時間が比較的長い．これは，FT,HPL を処理する際に，消費電力が頻繁に閾値を越えてしまうため，周波数変更プロセスによる CPU 負荷が増加し，消費電力の増大を招いているためと考えられる．

6. まとめと今後の課題

本稿では，TDP で考えた場合には冷却能力を越えてしまうような数のプロセッサを搭載する構成を利用し，アプリケーションの特性にあわせて消費電力制御を行うことで，冷却能力を最大限に活用して性能を最適化する「電力制約下での高性能計算機クラスタ構成手法」を提案した．また，システムの電力を動的に監視し，冷却能力により決定された許容消費電力を越える場合には，DVS 手法により消費電力を制御しつつ実行させる手法も提案した．また，提案手法を適用したプロトタイプのクラスタシステムを構築し，従来のクラスタ構成と比較評価を行った．

評価の結果，提案するクラスタ構成手法を用いることで，従来型のクラスタシステムに比べ高い性能を達成できることがわかった．また，動的実効電力制御を実装し，実際に許容電力付近で動作を行うことで，従来型システムより高い性能を達成した．

しかし，現在の動的実効電力制御では，消費電力推移の急峻なアプリケーションにおいては，頻繁に周波数／電圧の変更が行われるため，性能が低下する．これを解消するため，閾値幅を大きくする，周波数変更の幅を閾値からの外れの度合いに対応して動的に調整する，また，ある程度の時間閾値の上限を越えなければ周波数を上げる，などのアルゴリズムの改良を行う必要がある．

また，アプリケーション毎に最適なノード数，周波数を設定するため，動的ノード数変更，プロファイリング等による電力予測などを行う必要がある．これら手法についても今後確立する必要がある．

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」，および文部科学省科学研究費補助金 (若手研究 (B) 17700049)，東レ科学振興会科学研究助成の支援によって行われた．

参 考 文 献

1) TOP500 team. "Top500 list for November 2005", http://www.top500.org/lists/2005/11
2) M. Warren et al. "High Density Computing: A 240-Node Beowulf in One Cubic Meter", In *Proc. Supercomputing 2002*, Nov. 2002.
3) H. Nakashima et al. "MegaProto:1TFlops/10kW Rack Is Feasible Even with Only Commodity Technology," In *Proc. Supercomputing 2005*,Nov.2005.
4) IBM and Lawrence Livermore National Laboratory, "An Overview of the BlueGene/L Supercomputer", In *Proc. Supercomputing 2002*, Nov.2002.
5) 堀田義彦 他, "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, Vol.45, No.SIG11(ACS7), pp.207–218, 2004.
6) R. Ge et al. " Performance-constrained Distributed DVS Scheduling for Scientific Applications on Power-aware Clusters," In *Proc. Supercomputing 2005*,Nov.2005.