Vol.2012-EVA-38 No.2
2012/8/2

分散データストアシステムに対するEnergyCapping制御

菅 真樹[1,2,a)]　小林 大[1,b)]　長谷部賀洋[1,c)]

概要：IT システムの利用電力に制約を掛ける EnergyCapping 制御の必要性が高まっている．本稿では，分散データストアに対して，構成ノードの停止制御のみで制御を実現する NodeStop 制御と，構成ノードのリソース利用に制約とノード停止制御の両方を用いる NodeCapping 制御の 2 つの手法を提案する．省電力機能を備えた分散 KVS プロトタイプを用いて性能評価の結果，両手法により制約電力値以下で分散データストアが運用できることを確認し，消費電力およびベンチマーク性能について確認した．また，NodeCapping 制御は制約電力動作への遷移時間に優れ，NodeStop 制御は性能効率に優れることを確認した．

キーワード：クラウドコンピューティング，分散システム，ストレージ，省電力化，電力制約

Energy capping control for distributed data store system

Masaki KAN[1,2,a)]　Dai KOBAYASHI[1,b)]　Yoshihiro HASEBE[1,c)]

Abstract: Energy capping control, which restricts power consumption on the IT system, is desired in the real world. This paper proposes 2 control methods to achieve energy capping control, the NodeStop control which use only node's power down and the NodeCapping control which combines node power down and node's power capping. The evaluation results on our distributed key value store for energy-saving show the operability of 2 control methods and the performance of benchmark and energy consumption. Moreover, the results also show the advantage of NodeCapping control is transition duration to energy capping state, and the advantage of NodeStop control is the storage performance.

Keywords: Cloud Computing, distributed system, storage, energy-saving, energy-capping

1. はじめに

東日本大震災は，我が国における電力環境に非常に大きな影響を与えており，2011 年夏には，多くの電力利用者は使用制限を求められてきた．このような厳しい状況にもかかわらず，データセンタは安定的な経済活動・社会生活に不可欠な需要設備として，使用制限の緩和対象となっている．しかし，近年のデータセンタの電力利用量は電力利用量全体の 1.1% - 1.5%に達しており [1]，データセンタについても電力利用の効率化が重要視されてきている．そして，データセンタの中で動作する装置の電力利用の観点では，利用負荷に応じた電力でシステムを動作させる Energy Propotional Computing という考え方が提唱されてきた [2]．電力需給逼迫による電力利用制約を考慮すると，このような考え方に加え，利用可能な電力が制限された条件で IT システムを動作させることが必要になる．

本稿では，このような指定電力値以下で IT システムを動作させるための制御を EnergyCapping 制御と呼ぶ．データセンタ内で電力を利用する IT システムの装置として，主に計算機サーバ，スイッチなどのネットワーク装置，ストレージ装置などがある．本稿では，このうちストレージ装置，特に近年増加している複数の計算機によってストレー

[1] NEC クラウドシステム研究所
Cloud System Research Laboratories, NEC Coporation
[2] 東京工業大学 情報理工学研究科
Graduate School of Infomation Science and Engineering, Tokyo Institute of Technology
[a)] kan@bq.jp.nec.com
[b)] daik@ay.jp.nec.com
[c)] y-hasebe@ah.jp.nec.com

ジ機能を提供する分散データストアを中心に議論する．

IT システムに対する EnergyCapping 制御のシナリオは次の 2 種類が考えられる．1 つは，システムの運用コストとして電力コストを一定以下に抑えたい場合や，事前に設定された計画に基づいて利用電力を抑えたい場合である．このようなシナリオでは，限られた電力内で可能な限り性能効率に優れた制御が求められる．もう 1 つは，災害などの非常時や夏季の需要増などで，電力供給量が急遽逼迫し，政府などの要望によって IT システムの利用電力を急遽一定以下に抑えなければならないという場合である．このようなシナリオでは，可能な限り高速に電力制約された状態に移行できるような制御が求められる．分散データストアに対する EnergyCapping 制御は，従来の省電力化手法を応用することで実現可能と予想されるが，その具体的な手法や性能等への影響については明らかではない．

近年，分散データストアに対する様々な省電力化手法が提案されている [3]．従来のディスクアレイによるストレージシステムは，CPU やメモリの搭載量に対し接続される HDD の数が多く，したがって HDD の利用電力がシステムで支配的であったため，一部の HDD を省電力モードに移行するアプローチが多く用いられてきた．

一方，複数の計算機を用いて実現される分散データストアにおいては，計算機ノード全体を省電力モードに移行する必要がある．それは，計算機において記憶装置に依存する消費電力はわずかなためである．計算機ノードの消費電力は計算機の負荷によって増減する変動部分と，負荷に寄らず一定の量を消費する定常消費部分がある．例えば，DB サーバでは定常消費が 54 %を占める [4] ことが示されている．この定常消費部分の電力を削減するため，システム全体の負荷量をある特定のノード群に割り当て，その他のノードの電力消費を停止する方法がある [5]. また，負荷変動部分の電力についても，計算機ノード自体に搭載される PowerCapping 機能や，DVS(Dynamic Voltage Scaling) 機構により CPU の動作周波数を調整することで利用電力を調整する事が可能である．

本稿では，分散データストア上での EnergyCapping 制御の実現手法について議論する．まず，EnergyCapping 制御を実現するためにデータストアが備えるべき機能と，我々の開発中の分散データストア DKVS に基づく実装について述べる．続いて，2 種類のアプローチに基づく EnergyCapping 制御手法を提案した上で，両者を実機上で比較評価を行う．

本稿における我々の貢献は次の通りである：

- EnergyCapping 制御を実現するため，ノード電源停止のみを活用する NodeStop 制御と，ノード電源停止と CPU 制御を組み合わせた NodeCapping 制御の 2 種類の手法を提案した .
- 電力効率に優れたサーバにおける NodeCapping 制御手法のサーバ利用効率性について述べ，論理検証を行った．
- 開発中の DKVS プロトタイプの省電力機能を実機上で評価し，高い省電力スケーラビリティ性能を示すこと，レスポンス性能の悪化と引き替えに，平均電力を約 34%抑えることが可能であることを示した．
- また，2 種類の制御手法の両方において指定電力以下の動作が可能なこと，実験に用いたサーバの電力効率においては NodeStop 制御の性能が有利であるが，NodeCapping 制御は EnergyCapping 制御状態に移行するまでの時間の観点で優れることを示した．

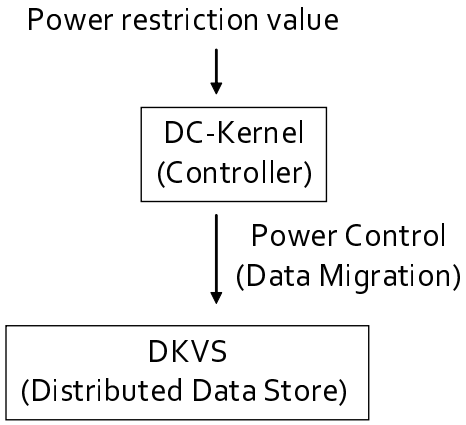

図 1　概略システムアーキテクチャ
Fig. 1　System Architecture

2. システムアーキテクチャ

本稿では，分散データストアに対する EnergyCapping 制御方式について議論する．まず，分散データストアに対する EnergyCapping 制御を実現するシステムアーキテクチャを図 **1** に示す．

本システムアーキテクチャは，電力制約制御の対象となる分散データストアと，制御命令を発行するコントローラからなる．分散データストアは，省電力機能として，構成ノードの電力制御機能を備える．コントローラは，システム管理者などから制約電力値を指定され，分散データストアがその制約電力値内で収まるように動作するよう，分散データストアにどのような電力制御を行うか（ノードの稼働停止台数など）を決定する機能を持つ．また，分散データストアの構成ノードに対して電力制御要求を行う．さらに，必要あれば，構成ノードの電源制御にデータアクセス処理に問題が起きないようにするための，データマイグレーション要求も共に行う．

本稿では，電源制御可能な分散データストアとして，我々が開発中であるインメモリ分散データストアである DKVS を用いる．2.1 においてその概要を述べるが，詳細は [6] を参照されたい．また，本稿で我々が開発したコントローラを DC-Kernel と呼ぶ．2.2 でその概要を述べる．

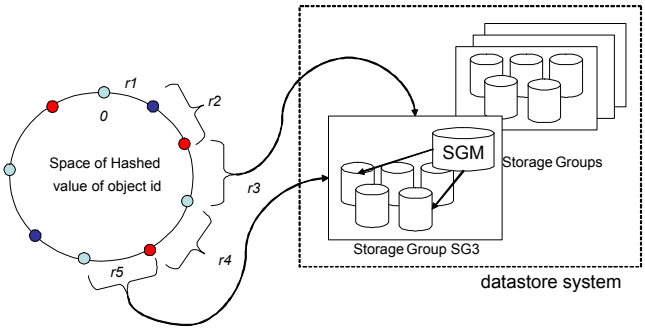

図 2　DKVS の 2 段階データ配置方式
Fig. 2　Hierarchical Data Location Management on DKVS

2.1 DKVS:Distributed Key Value Store

DKVS は，我々が開発中の分散 Key-Value ストアで省電力機能を持つ．DKVS ではデータは複数のプロパティを合わせたオブジェクトと呼ぶデータ単位に対しシステム内で一意の Key 値を付与し格納する．オブジェクトの複数の複製が複数の計算機の主記憶に格納される．また，オブジェクトレベルの一貫性があり，同じオブジェクトへのアクセスはトランザクショナルに行うことが出来る．

DKVS は，スケーラビリティとデータ配置の柔軟性を備えるための 2 段階データ配置方式と，構成ノードの一部を ACPI(Advanced Configuration and Power Interface) で規定された省電力モードに移行できる機能を備えることを特徴とする．**図 2** にこの 2 段階のデータ配置管理の概要を示す．DKVS では，システム構成ノード群を複数のストレージグループ (SG) に分割する．各 SG は 1 台の SG マスターノード (SGM) をもち，SG 内のオブジェクト配置情報を集中管理することで，SG 内での自由なオブジェクト配置を実現する．また，SGM の死活監視や SGM 障害時の SGM 再選出，状態情報の管理などを行う代表ノード (RootMaster) が存在する．RootMaster は別のノードに用意してもよいし，SGM のいずれかが担っても良い．

このように 2 段階管理し，SG 内の自由度を増すことで，オブジェクトの複製を適切に配置し，システム低負荷時にはより多くのノードの電源を停止することが出来る．このようなデータ配置の柔軟性により，各オブジェクトの複製のうち最低 1 つを生かしたまま，その他のオブジェクトを格納するノードを低電力モードへ移行することで，システム全体の消費電力を削減することができる．

2.2 DC-Kernel: Controller of Distributed Data Store

DC-Kernel は，我々が開発中の分散データストアシステムのコントローラである．**図 3** に構成と EnergyCapping 制御の命令フローを示す．

DC-Kernel による EnergyCapping 制御は次のように行われる．まず，システム管理者などが期間や制約電力値を

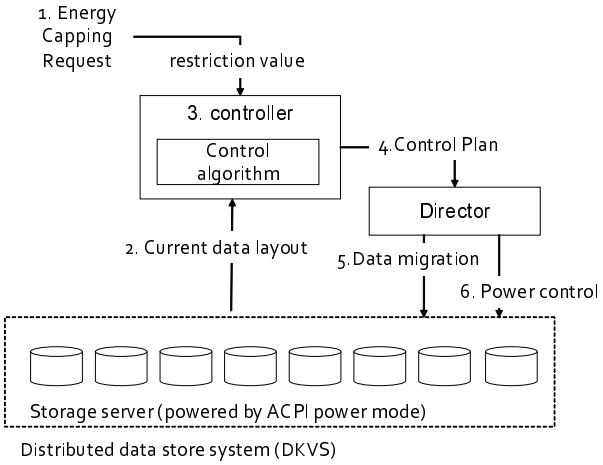

図 3　DC-Kernel のアーキテクチャと処理フロー
Fig. 3　The DC-Kernel's architecuture and power control flow

含む電力制約要求を DC-Kernel に要求する (1). これは，電力供給不足に基づく政府などからの要求や，企業の電力コスト削減のための節電要求に基づくことを想定している．

DC-Kernel はこの制約電力値に基づいて，分散データストアのデータ配置状態を取得 (2) した上で，あらかじめ登録された制御アルゴリズムに基づいて制御プランの算出を行う (3)．制御プランとして，電源制御の対象となるノードの選択，電源制御内容の決定，必要なデータマイグレーションの内容が決定される．具体的な EnergyCapping 制御アルゴリズムの内容については，3 で述べる．

算出されたプランを Director に渡し (4)，Director は制御プランに基づいて，データマイグレーション要求を分散データストアに対して行い (5), そのマイグレーション処理が完了したあとに，分散データストアに対して電力制御要求を行う (6).

3. EnergyCapping 制御手法

指定された電力値 P_{cap} 以内でシステムを動作させる EnergyCapping 制御は，電力をどのような内訳で利用するかによって幾つかの手法が考えられるが，本稿では 2 種類の EnergyCapping 制御手法を提案する．図 3 で示したように，コントローラは制御プランとして分散データストアの構成ノードに対する電力制御内容および，マイグレーション内容を決定する．電力制御内容とは，分散データストアのノードを何台稼働させ，何台停止させるか，稼働ノードにはどれだけの電力利用を許容させるか，ということである．

一部の構成ノードを停止可能なシステムの Total 電力 $P(a, P_{active})$ は，システムの総ノード台数を n, 稼働ノード台数を a, 停止状態のノード電力を P_{stop}, ノード 1 台あたりの利用電力を P_{active}，それ以外の電力消費量を α としたときに，数式 1 で示すことが出来る．P_{active} は実際にはノード毎に異なる電力となるが，簡略化のため全てのノー

ドでほぼ同じ電力利用と仮定して議論する．これは，本稿の主な対象である分散データストアは，適切にオブジェクトが分散配置されている条件のもとでは同程度の負荷であるためである．また，分散データストアを実現するソフトウェアで制御対象とできないプロセスについては α に反映するものとする．DKVS では SGM および RM のプロセスが利用する電力がこれに該当する．

$$P(a, P_{active}) = (n-a)P_{stop} + a * P_{active} + \alpha \qquad (1)$$

$P(a, P_{active}) \leq P_{cap}$ とすることが EnergyCapping 制御の目的である．そのための制御可能なパラメータは，a と P_{active} である．本稿ではこのパラメータの決定方法として，NodeStop 制御と NodeCapping 制御 2 種類の手法を提案する．3.1 において，a だけの制御で EnergyCapping 制御を実現する NodeStop 制御について述べる．3.2 では，a と P_{active} の両方を用いて制御する手法である NodeCapping 制御について述べる．

3.1 NodeStop 制御

NodeStop 制御は，システムのノード電源停止のみを活用して EnergyCapping 制御を実現する方法である．

NodeStop 制御では，制御プラン算出処理において，稼働ノード台数 a を決定する．この a は，式 1 を用いて算出することが出来る．P_{active} をシステムの構成ノードピーク電力 P_{peak} と見なした上で，$P \leq P_{cap}$ となる最大の a を算出する．残りの $n-a$ 台が電源停止されるノード台数となる．

P_{peak} は，サーバの CPU の利用率によって決定されることが [7] などの研究によって報告されている．従って，CPU を 100%使い切れるアプリケーションであれば P_{peak} はサーバの最大利用電力に近い値になる．一方，他のリソースがボトルネックになって CPU を使い切れないアプリケーションであれば測定に基づいて決める必要がある．

a 決定後は，DC-Kernel は $n-a$ 台の DKVS ノードを停止予定ノードとしてランダムに選択し，停止予定ノード上のオブジェクトを制御後にも稼働し続けるノードにマイグレーションする．マイグレーション先は，同一オブジェクトの複製が同一ノード上に配置されないように決定する必要がある．マイグレーションが完了した後に，停止予定ノードに対して ACPI を介して電源制御要求を行う．稼働ノードには利用電力の制約を掛けない．

3.2 NodeCapping 制御

NodeCapping 制御は，システムの構成ノードの電源停止と稼働ノードの最大利用電力への制約の 2 つを組み合わせた手法である．

3.2.1 制御プラン計算

NodeCapping 制御の制御プラン計算では，$P \leq P_{cap}$ を満たす P_{active} と a の組み合わせを決める．

この組み合わせについても，式 1 を用いて算出することが出来る．なお，条件として $P_{idle} \leq P_{active} \leq P_{peak}$ を満たす必要がある．P_{idle} はサーバのアイドル時の消費電力である．P_{idle} と P_{peak} の値によっては複数の P_{active} と a の組み合わせが条件を満たすことが出来る．この組み合わせから，任意のポリシーに基づいて組み合わせを選択する．例えば，P_{active} がある閾値以上の条件で a が最も多いものを選ぶ，といったポリシーが考えられる．また，a もシステムの最小稼働台数以上である必要がある．例えば，3 つ以上のサーバに複製を保持することで信頼性を担保している分散データストアでは，a は 3 以上でなければならない．

3.2.2 NodeCapping 制御の実装

a と P_{active} が決定された後に，NodeCapping 制御では，NodeStop 制御と同様に $n-a$ 台の DKVS ノードに対して電源制御要求を行う．これに加えて，a 台の稼働サーバの最大利用電力を P_{active} 以下に抑える電力制御を行う必要がある．この制御方法は，以下の 2 種類の実現方法が考えられる．

- BMC(Base Management Controller) ベース
- ソフトウェアベース

まず，BMC ベースの手法について述べる．最近の業務用サーバでは，サーバ単体の PowerCapping 制御を行う技術が搭載されている．[*1] この方法では，Intel Node Manager と連携し，サーバプロセッサのパフォーマンスを制限することで最大電力値を調整することが可能である．このような方法が利用できるサーバの場合には，P_{active} の値をサーバの利用電力の上限値として規定すればよい．

ソフトウェアベースの手法では，上記のようなサーバ単位の最大電力値制御をソフトウェアで CPU 利用率を制限することで実現する．例えば，LimitCPU[*2]のようなソフトウェアによって CPU 利用率を制限することが出来る．

ソフトウェアベースの手法では，制約すべき電力 P_{active} に対して制約 CPU 利用率を算出する必要がある．[7] で CPU 利用率あたりの消費電力について経験則に基づく推定モデル式 (下記式 2) が提案されている．u が CPU 利用率，r は calibration parameter である（本稿および [7] では 1.4）．この式を用いて P_{active} を満たすための CPU 利用率 u を決める事が可能である．

$$P_{active} = P_{idle} + (P_{busy} - P_{idle})(2u - u^r) \qquad (2)$$

3.3 NodeCapping 制御の優位性

本節では，利用可能電力に対して利用できるサーバリソース量という観点で，NodeCapping 制御の NodeStop 制

[*1] http://support.express.nec.co.jp/tech/PowerCapping_WhitePaper/PowerCapping_WhitePaper_rev1.5.pdf
[*2] http://limitcpu.sourceforge.net/

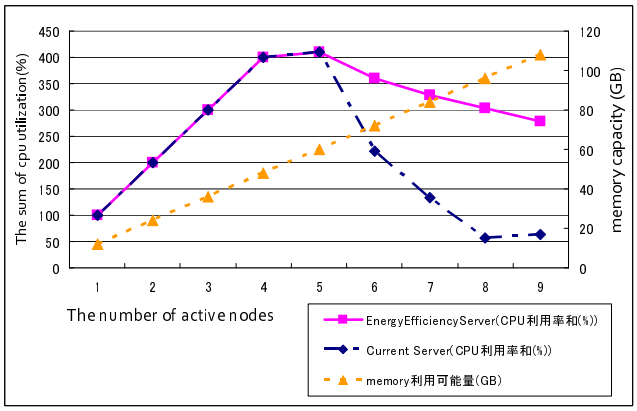

図 4　アイドル電力比率が異なるサーバにおける NodeCapping 制御のサーバリソース利用可能量

Fig. 4　The capacities of resource utilization by NodeCapping control in 2 different energy-efficiecy servers

御に対する優位性について議論する．また，サーバの idle 電力比率が NodeCapping 制御の利用サーバリソース量に与える影響についても述べる．

NodeCapping 制御は，NodeStop 制御よりも，利用可能電力に対してより多くのサーバリソース量を利用可能なことがある．NodeStop 制御よりも稼働ノード台数が多いために，メモリ容量や CPU コア数，CPU とメモリ間の帯域などは多く利用可能であるためである．しかし，CPU 利用量は一定以下に抑えられてしまう．[2] に述べられているような，Energy Efficiency の低い（すなわち idle 電力がピーク電力に比して高い）サーバについては，idle 電力に相当する電力分の削減を CPU 利用量の減少によって補う必要があるため，その優位性は小さい．しかし，近年のサーバのような idle 電力の割合が低いサーバ（[2] では約 10%のものが例示されている）であれば，NodeCapping 制御でその分多くのリソースを利用可能である．

図 4 は，900w の EnergyCapping 制御を行った場合に，稼働ノード数に応じた利用可能なサーバリソース量が，idle 電力比率が異なる 2 種類サーバにおいて異なる事を示すグラフである．CurrentServer は $P_{idle} = 100(w), P_{busy} = 180(w)$ としたサーバ (4 において本稿の実験で利用したサーバと同一特性)，EnergyEfficiencyServer は仮に $P_{idle} = 38(w), P_{busy} = 180(w)$ としたサーバ（電源効率の良いサーバ特性）である．このときに，a の台数を変動させて EnergyCapping 制御を行った際に，数式 2 に基づいて利用可能 CPU 利用率を算出し，稼働ノードで合算することで CPU 利用率和を算出した．また，1 ノードあたりのメモリ搭載容量は 12GB とした (同様に 4 において本稿の実験で利用したサーバと同一特性)．

a が 4 までの場合，稼働ノードは全て CPU を 100%利用可能であるため，2 種類のサーバにおいて両方とも CPU 利用率和は同一である．なお，$a = 4$ が NodeCapping 制御の制御条件となる．$a = 5$ のケースにおいてわずかに $a = 4$ より CPU 利用率を多く利用でき，以降は a が増加するにつれて CPU 利用率和が減少していく．このとき CurrentSever の減少度合いは，EnergyEfficiencyServer と比較して大きい．これは idle 電力が大きいためその電力分をカバーするために，1 サーバあたりで利用できる CPU 利用率が大きく下がってしまうためである．一方，当然ながら両方のサーバ特性において，利用可能なメモリ容量は a が増加するに辺り比例して増加する．

つまり，idle 電力の低いサーバを利用できれば，NodeCapping 制御は NodeStop 制御と比較して，CPU 利用率和の観点では価値が小さいが，メモリは多く利用できるために，CPU よりもメモリ利用の影響が大きいアプリケーションに対しては高い価値が得られる．一方，idle 電力が高いサーバでは，稼働台数が多くなるにつれて CPU 利用率に対する制約が厳しくなるため，a を大きくした NodeCapping 制御よりも，NodeStop 制御が有用である．

4. 評価実験

本節では，NodeStop 制御および NodeCapping 制御により，制約電力以下でのシステム動作について，実機上に構成したプロトタイプ・システムにより確認する．また，それぞれの制御手法の，DKVS 上のワークロードに対する性能および消費電力に対する影響について比較する．

4.1 評価システム

本実験の DKVS は合計 9 台のノードによって構成され，システム運用管理のみを行う RM ノードを 1 台，8 台のストレージノード（1SG 構成）からなる．また，1 つのオブジェクトを 3 つのストレージノードに保持する．DKVS ノードの一部停止は ACPI-S4 で規定されるハイバネーションを利用する．

DKVS を動作させたハードウェアは 2 つの Intel Xeon E5504 プロセッサ，12GByte の DDR2-800 メモリを搭載した IA サーバ 9 台である．これとは別にベンチマーク負荷を掛けるためのサーバを数台利用している．電力測定は，Raritan 社のインテリジェント電源タップ Dominion PX を用い，3 秒おきの電力を各サーバ毎に測定し，DKVS 構成ノード 9 台の合算として示している．

予備評価として，本実験で用いたサーバの CPU 利用率と利用電力の関係を測定した．測定した実測値と，数式 2 による推定結果のグラフを**図 5** に示す．$P_{busy} = 180(w), P_{idle} = 100(w), P_{stop} = 20(w)$ の 3 つのパラメータを指定することで，本実験のサーバも数式 2 とほぼ同様の利用電力特性を示すことがわかった．

以降の実験の評価ワークロードには，Yahoo! Cloud Serving Benchmark(YCSB) [8] を用いた．ワークロードとしては，1 オブジェクト 1KB のオブジェクトを 1,000,000

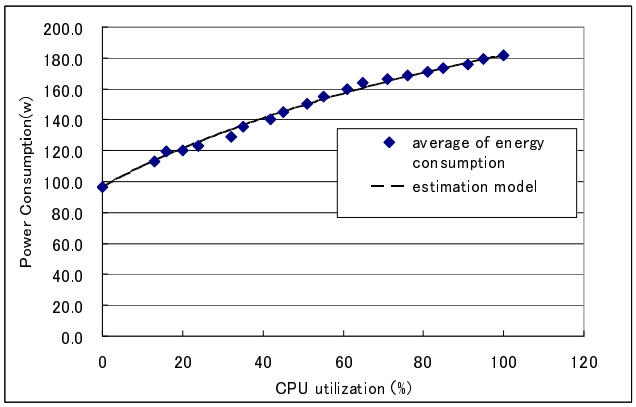

図 5　実験サーバの CPU 利用率と利用電力の関係

Fig. 5　Server power usage at varying cpu utilization from idle to peak

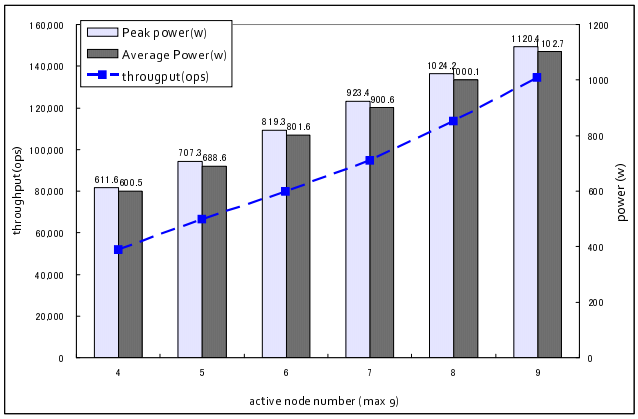

図 6　DKVS の稼働ノード数に応じたスループットおよび利用電力

Fig. 6　The throuput and power consumption on DKVS

個保持させ，Update 5%, Read 95%の比率で 8,000,000 回のアクセスを行った．アクセス先のオブジェクト分布は uniform で一律としている．また，DKVS はクライアントにオブジェクトの配置情報の一部をキャッシュに載せるまでは性能が安定しないため，測定前にある程度を負荷を掛け，充分キャッシュに乗っている状態の性能を示している．

4.2　DKVS の電力対性能比の評価

予備実験として，**図 6** に DKVS の基本的な省電力性能を示す．DKVS の停止ノード台数を変動させた際の，停止制御完了後の YCSB 性能と DKVS ノード 9 台 (1RM ノード，8 ストレージノード) の消費電力である．我々が開発した省電力データストアである DKVS は，稼働ノード数に比例したスループット性能および，消費電力特性を提供できている事がわかる．

また，YCSB の負荷量を 45,000 ops に抑えた際の応答性能について，**図 7** に示す．スループットは全てのケースにおいて約 45,000ops を達成したものの，稼働ノード数が低下するにつれてレスポンス性能の劣化と引き替えに消費電力を削減できる事がわかる．

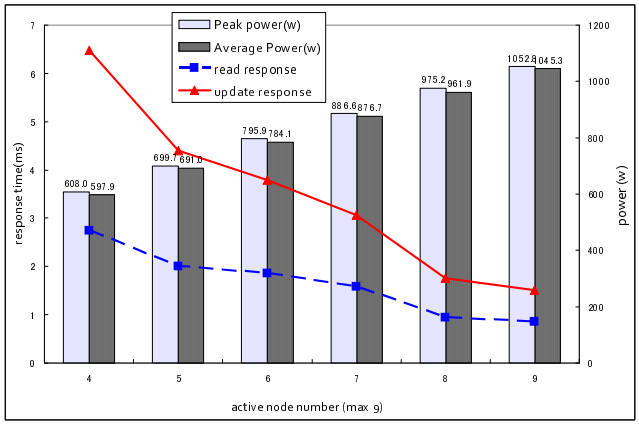

図 7　DKVS の稼働ノード数に応じたレスポンス性能および利用電力（低負荷時）

Fig. 7　The response time and power consumption on DKVS with low offload

予備実験より，例えば 4 台のノードを停止させる（稼働ノード 5 台）によって，スループット及びレスポンス性能が 2 倍程度に悪化するものの，平均電力を約 34%削減する事が出来る．レスポンス性能の悪化を許容出来る条件では十分に省電力データストアとして利用可能であることがわかる．

4.3　NodeStop 制御と NodeCapping 制御のワークロード性能および消費電力の比較

次に，本稿で提案したソフトウェアベースでの NodeCapping 制御と，NodeStop 制御について評価を行う．ここでは，800(w) に利用電力を制御するケースを想定し，NodeCapping 制御および NodeStop 制御によって EnergyCapping 制御を行う．

まず，EnergyCapping 制御のためのパラメータ P_{peak}, α を指定するため，4.2 で DKVS の限界負荷を掛けた際の DKVS ノード電力を測定した．その結果，現在の DKVS プロトタイプでは CPU を 100%使い切ることが出来ず，$P_{peak} = 140(w)$ が限界であることがわかった．

また，RootMaster は停止対象とせず，また DKVS 性能にもほとんど寄与しないため RootMaster の電力は α で加算することとした．つまり，数式 1 を数式 3 のように変形した．なお，RootMaster ノードの利用電力は約 100-120(w) であるため，α は 120 とした．

$$P(a, P_{active}) = (9-a)20 + (a-1)P_{active} + 120 \quad (3)$$

$$20 \leq P_{active} \leq 140 \quad (4)$$

この式に基づいて 800(w) の電力制限を掛ける場合，NodeStop 制御では $a = 5$，NodeCapping 制御では $a = 6, P_{active} = 122.8(w)$ となる．$P_{active} = 122.8(w)$ から数式 2 に基づいて，利用可能な最大の CPU 利用率を算出すると，19%となるため，LimitCPU を用いて 19%制約を

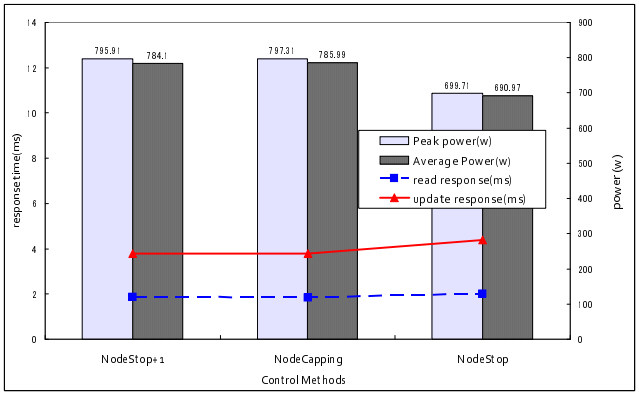

図 8　低負荷時の EnergyCapping 制御時の性能
Fig. 8　The performances on energy capping controls with low load

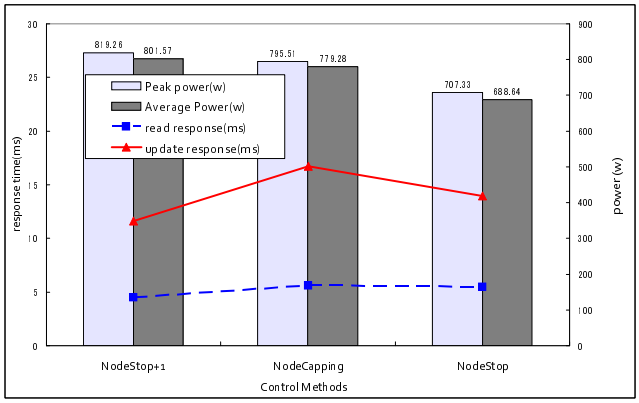

図 9　高負荷時の EnergyCapping 制御時のレスポンスタイム性能
Fig. 9　The response time and power consumption on energy capping controls with high load

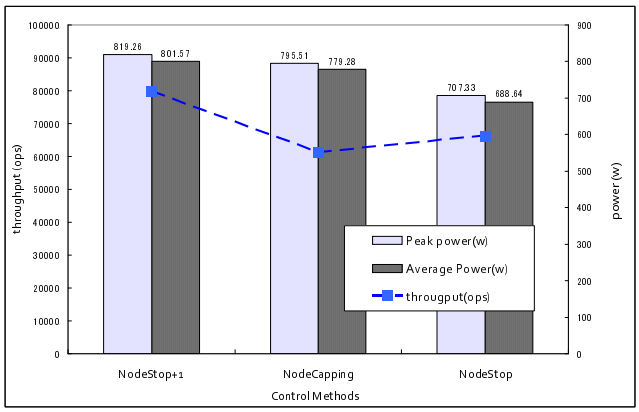

図 10　高負荷時の EnergyCapping 制御時のスループット性能
Fig. 10　The throughput and power consumption on energy capping controls with high load

DKVS サーバノードに対して掛ける．なお，$a = 7$ の場合，$P_{active} = 103.3(w)$ となるが，103.3(w) のケースは CPU 利用率として 2%しか利用できない．LimitCPU ではこのレベルの CPU 利用率制約は正しく動作しないため，実験を省略した（非常に低い性能を示すと予想される）．

まず，$a = 5$ の環境でも充分捌くことが可能なスループット負荷 (45,000 ops) におけるレスポンス性能について，**図 8** に示す．NodeStop 制御の測定値は，4.2 節の結果を用いている．スループット性能はいずれのケースにおいても約 45,000ops を示したのでグラフを省略した．また，NodeCapping 制御と NodeStop 制御の他に，比較として $a = 6$ の NodeStop 制御，すなわち NodeCapping 制御で CPU 利用率制約を掛けない時の性能値を NodeStop+1 として示した．

図 8 を参照すると，NodeCapping 制御および NodeStop 制御の両方で 800(w) 以下の動作が実現できることがわかる．また，NodeCapping 制御は NodeStop 制御よりもレスポンス性能がやや優れる結果を得た．NodeStop+1 の場合にもほぼ 800(w) 以下で動作することから，これは 45,000ops という負荷では各ノードの CPU 利用率が制約レベルの 19%に達しないものと予想される．従って，充分低い負荷であれば NodeCapping 制御は NodeStop 制御と同等以上の性能を示すことが可能である．なお，このように NodeStop+1 制御でも低負荷状態で安定していれば，800(w) 制約動作が可能であるが，アプリケーション負荷が増加した場合には 800(w) 以上の電力を利用してしまう．

次に，YCSB で無制限に負荷を掛けた場合の性能について**図 9**, **図 10** に示す．YCSB で可能な限りの負荷を掛けても両方の制御手法により，指定電力以下に制約動作が実現できていることがわかる．NodeCapping 制御では，CPU 利用率制約が厳しいため，スループットおよびレスポンス性能の悪化が大きく，NodeStop 制御の方が優れた性能を示している．従って，一定以上の負荷では NodeCapping 制御よりも NodeStop 制御の方が性能効率に優れると考えられる．これは，DKVS では全てのオブジェクトをメモリ上に保持しているため，アクセス性能に現れる CPU 利用率の影響が大きいためである．CPU 利用率に制約を掛けられている分，処理待ちの時間が発生してレスポンスタイム性能が悪化する．

4.4　各制御の遷移時間の評価

NodeStop 制御と比較して NodeCapping 制御の優れる点は，停止ノードが少ないために，電源制御に必要なデータマイグレーション量が少なく，制約電力以下に移行するまでの時間が短縮できることにある．また，追加でノード停止を行う必要が無い条件であれば，瞬時に電力制約状態に移行できるため，更に時間を短縮できる．ノード停止時間は，ACPI-S4(ハイバネーション) の場合，サーバのメモリ状態を HDD に書き出して停止処理を行う時間に依存する（本実験では約 90 秒掛かる）．

前節までの実験条件と同等条件において，NodeCapping 制御によりノード停止が必要無い制御として，6 台稼働 3 台停止の状態から 800(w) 制約を行う場合と，NodeStop 制

表 1　NodeCapping 制御と NodeStop 制御間の比較
Table 1　Comaprison of NodeCapping control and NodeStop control

制御方式	マイグレーション量 (objects)	移行時間 (sec)
NodeStop	59,982	222
NodeCapping	0	3.2

御で更に 1 台停止する場合の，マイグレーション量と，電力制約状態に移行するまでの時間を**表 1** に示す．EnergyCapping 状態に移行する時間を著しく短縮できることがわかる．

5. 関連研究

データセンターにおけるシステム負荷と消費電力の関係について，現在非線形であるものをより線形に近づけること (energy proportionality) の重要性については，Barroso らによって言及されており [2]，CPU の電力制御が中心に議論されている．本稿で提案した NodeCapping 制御手法はこの線形性をより高める手法の 1 つである．

本稿で対象とする分散ストレージシステムに対しても，様々な省電力化手法が行われている [3]．本稿で述べたような，計算機ノードを停止させることにより電力効率性を向上させる研究が数多く行われている [6], [9], [10]．これらの研究は，データの配置と停止方法にフォーカスしており，指定された電力以下での制御と，CPU 電力制御の組み合わせに着目した我々の研究とは求める特性が異なる．

また，分散データストアの性能 SLO を保証するためにデータ配置を制御するアプローチとフレームワークが [11] に提案されている．我々の開発したコントローラもこれと類似しているが，性能だけでなく電力制御の観点に着目している点で異なる．

6. おわりに

我々 NEC は，人と地球にやさしい情報社会の実現をグループビジョンとして掲げて活動しており，限られた電力内で IT システムを動作させる EnergyCapping 制御はその重要な要素のうちの 1 つである．本稿では，分散データストアに対する 2 種類の EnergyCapping 制御手法を提案し，比較評価を行った．その結果，双方の制御手法で目的の制約電力以下での分散データストアの動作が可能であることを確認した．また，idle 電力性能が高い物理サーバを用いた評価実験では，ノード停止制御のみを利用する NodeStop 制御が性能効率に優れ，ノード電力制約機能を活用する NodeCapping 制御は電力制約状態への遷移時間に優れる事を示した．さらに，NodeCapping 制御はアプリケーション負荷が充分低ければ，NodeStop 制御と同等の性能を発揮できることを示した．

今後の課題として，ソフトウェアベースのプロセスレベルの CPU 利用率制御ではなく，BMC を用いた PowerCapping 機能や CPU の DVS 機能を用いて NodeCapping 機能を実現することや，idle 電力性能が低い消費電力特性の優れた物理サーバを用いた評価がある．

謝辞　本研究の一部は, 独立行政法人新エネルギー・産業技術総合開発機構 (NEDO) の委託事業による成果である．本研究に向けてプログラム実装などについて多くの協力を頂いた NEC ソフトウェア東北株式会社の皆様に感謝する．

参考文献

[1] Koomey, J.: *Growth in data center electricity use 2005 to 2010*, Analytics Press, Oakland, CA (2011).
[2] Barroso, L. A. and Hölzle, U.: The Case for Energy-Proportional Computing, *Computer*, Vol. 40, No. 12, pp. 33–37 (2007).
[3] 長谷部賀洋，小林　大，菅　真樹：クラウドを支えるデータストレージ技術：5．クラウド時代を支えるグリーンなデータセンタのストレージ技術動向，情報処理，Vol. 52, No. 6, pp. 693–699 (2011).
[4] Tsirogiannis, D., Harizopoulos, S. and Shah, M. A.: Analyzing the energy efficiency of a database server, *Proc.of the 2010 international conference on Management of data (SIGMOD2010)*, pp. 231–242 (2010).
[5] 広渕崇宏，小川宏高，中田秀基，伊藤　智，関口智嗣：仮想クラスタ遠隔ライブマイグレーションにおけるストレージアクセス最適化機構，情報処理学会研究報告 HPC，Vol. 2008, No. 74, pp. 19–24 (2008-07-29).
[6] 小林　大，菅　真樹，大野善之，鳥居隆史：構成ノード電源停止によるシステム省電力化のためのインメモリ分散データストア設計，第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM 2010)，pp. C10–3 (2011).
[7] Fan, X., Weber, W.-D. and Barroso, L. A.: Power provisioning for a warehouse-sized computer, *Proceedings of the 34th annual international symposium on Computer architecture*, ISCA '07, No. 11, New York, NY, USA, ACM, pp. 13–23 (2007).
[8] Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R. and Sears, R.: Benchmarking cloud serving systems with YCSB, *Proceedings of the 1st ACM symposium on Cloud computing*, SoCC '10, No. 12, New York, NY, USA, ACM, pp. 143–154 (2010).
[9] Amur, H., Cipar, J., Gupta, V., Ganger, G. R., Kozuch, M. A. and Schwan, K.: Robust and flexible power-proportional storage, *Proceedings of the 1st ACM symposium on Cloud computing*, SoCC '10, No. 12, New York, NY, USA, ACM, pp. 217–228 (2010).
[10] Thereska, E., Donnelly, A. and Narayanan, D.: Sierra: practical power-proportionality for data center storage, *Proceedings of the sixth conference on Computer systems*, EuroSys '11, No. 14, New York, NY, USA, ACM, pp. 169–182 (2011).
[11] Trushkowsky, B., Bodík, P., Fox, A., Franklin, M. J., Jordan, M. I. and Patterson, D. A.: The SCADS director: scaling a distributed storage system under stringent performance requirements, *Proceedings of the 9th USENIX conference on File and stroage technologies*, Berkeley, CA, USA, USENIX Association, pp. 12–12 (2011).