科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　３０　日現在

機関番号：　３２６１２
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５０００５０

研究課題名（和文）クラウド環境におけるメニーコア向け省電力仮想マシン・スケジューラ

研究課題名（英文）Power-Saving Virtual Machine Scheduling for Many-Core CPUs in Cloud Environments

研究代表者
河野　健二（Kono, Kenji）

慶應義塾大学・理工学部・准教授

研究者番号：９０３０１１１８

交付決定額（研究期間全体）：（直接経費）　　3,900,000円、（間接経費）　　1,170,000円

研究成果の概要（和文）：クラウド環境における電力コストは，その運用コストの大きな部分を占めている．クラウド環境を実現するデータセンタでは仮想マシンによる資源管理が一般的になっている．本研究では仮想マシンモニタの階層においてデータセンタの省電力化を実現するためのさまざまな機構の開発を行った．ライブマイグレーションによって仮想マシンの集約を促進し，ネットワーク機器までを含めた省電力化の達成，地域毎の電力価格変動に応じたデータセンタの自動選択機構などを達成した．また，メニーコア・プロセッサのひとつである GPU の仮想化を達成し，メニーコア・プロセッサ向けの仮想環境における省電力化のための先鞭をつけた．

研究成果の概要（英文）：The cost of electric power is becoming dominant in modern cloud environments. In the modern cloud environments, it is norm to use virtual machines to manage their physical resources. In this research, several techniques for saving electricity in data centers have been established. They include the technique to promote the virtual machine consolidation for power saving by means of live migration. It also reduces the power consumption by network switches. Another technique that allows us to select cost-effective data centers in terms of electricity prices. In addition to those techniques, the virtualization of GPU, an example of many-core processors has been achieved. It opens a way to power saving of many-core processors.

研究分野：　情報学

科研費の分科・細目：　計算基盤・ソフトウェア

キーワード：　クラウド環境　省電力　仮想化　マイグレーション

１．研究開始当初の背景

データセンタに集約されたコンピュータ資源を公開し，一般に利用できるようにしたクラウド・コンピューティング環境が注目を集めている．クラウド環境は“pay-as-you-go” とも言われており，その利用者は，コンピュータ資源を利用した分に対してのみ支払いを行えばよい．1 台のコンピュータを 1,000 時間利用するのも，1,000 台のコンピュータを 1 時間利用するのも，同じだけの課金となる．

クラウド環境はその形態によっていくつかに分類されている．本研究では IaaS (Infrastructure as a Service) 型のクラウド環境を対象とする．IaaS 型のクラウドは，オペレーティングシステムを含むすべてのソフトウェア・スタックを利用者が自由に選択・インストール可能となっている．そのため，提供したいサービスに適合した最適なソフトウェア・スタックを用いることができるという利点がある．クラウド環境のひとつである Amazon EC2 では IaaS 型のクラウド環境を提供している．

Amazon EC2 に代表される IaaS 型のクラウド環境では，仮想マシンモニタを用いて仮想マシンを提供している．クラウド環境の利用者はその仮想マシン上にオペレーティングシステムなどのソフトウェア・スタックをインストールして利用する．そのためクラウド環境のほうでは，利用されるオペレーティングシステムをあらかじめ特定することができない．

クラウド環境における重要な問題のひとつは，その消費電力の削減である．大規模なデータセンタの消費電力は，その運用コストのかなりの部分を占めており，運用コスト削減という視点からも，自然環境保護という視点からも重要な研究テーマとなっている．

一般に電力制御のための機能はオペレーティングシステムの階層で実現されている．アプリケーションとオペレーティングシステムが連携して電力削減を行う場合であっても，実際にハードウェアの機能を使って電力削減の処理を行うのはオペレーティングシステムである．

しかし，IaaS 型のクラウドでは，特定のオペレーティングシステムの利用を強制することができないため，オペレーティングシステムによる省電力対策を期待することができない．また，一台のコンピュータ上で複数の仮想マシンが動作するため，個々の仮想マシン内で動作するゲスト・オペレーティングシステムが電力制御を行っていたとしても，オペレーティングシステムからは物理 CPU の状態が隠蔽されているため，その効果が十分に発揮できるとは限らない．したがって，仮想マシンモニタの階層において，電力制御のための機能を支援する必要がある．

２．研究の目的

本研究の目的は，IaaS 型のクラウド環境のための省電力化技術を確立することである．多数のコアをひとつのチップに集積したマルチコア・メニーコアのプロセッサに対して効果的な，仮想マシンモニタ階層での省電力化技術は十分に確立されているとは言えず，マルチコア・メニーコアのプロセッサを想定した仮想マシンモニタ階層での省電力化技術を確立することを目的とする．

研究の背景においても述べたとおり，IaaS 型のクラウド環境ではオペレーティングシステムの階層で十分な省電力対策が組み込まれていることは保証できず，仮に，そのような仕組みが組み込まれていたとしても，各仮想マシン内で動作するゲスト・オペレーティングシステムは物理的な CPU の状態を直接コントロールすることはできないため，仮想マシンモニタの階層における省電力化は必須である．

（１）ゲスト・オペレーティングシステムが特に省電力のための仕組みを備えていない場合，もっとも効果の高い省電力手法は，複数の仮想マシンをできる限り少ない台数の物理マシンに集約し，稼働する物理マシンの台数を減らすことである．

さらに，クラウド環境におけるネットワーク・トポロジの特性を活用することで，物理マシンだけではなくネットワーク・スイッチ等のネットワーク機器を含めた省電力化が可能となる．特に，仮想マシンの持つライブ・マイグレーションの機能を用いると，ネットワーク機器を含めた消費電力量を最適化するように仮想マシンを配置することが理論上，可能となる．本研究の目的のひとつは，そのような仮想マシンの再配置を行う効率的な手法を確立することである．

（２）また，近年，クラウド環境においてもマルチコア・メニーコアと呼ばれるプロセッサが計算資源として提供されるようになりつつある．本研究ではメニーコアのプロセッサの代表として，GPGPU (汎目的計算のためのグラフィック・プロセッシング・ユニット）を対象とする．GPGPU は価格に対する性能比，消費電力量に対する性能比がともに突出してすぐれており，ハイパフォーマンス・コンピューティングだけではなく，さまざまな分野において活発に活用されるようになっている．

現状では GPGPU はその仮想化が困難であるため，クラウド環境における資源の共有には適していないとされている．本研究では，GPGPU の仮想化技術を確立し，そのスケジューリング方法を仮想マシンモニタの階層で制御可能にすることによって，メニーコア・プロセッサの代表である GPGPU に対しても省電力化のための技術的なプラットフォームを提供することを目指す．

３．研究の方法

（１）クラウド環境におけるライブ・マイグレーションを利用した省電力化のための仕組みに関しては，クラウド環境を模したシミュレーション環境を構築し，シミュレーションによるアルゴリズムの検討・改善・開発を行う．

このようなシミュレーションを行うためには，実際のデータセンタにおけるワークロードをできる限り忠実に模倣する必要がある．しかし，このようなワークロードは一般に公開されていることは稀であり，入手は極めて困難である．本研究では文献調査により米国 Google や Amazon などのデータセンタにおけるネットワーク解析結果から，想定されるワークロードを抽出するという方法をとる．

シミュレータを用いることにより，さまざまな規模のデータセンタに対して，提案手法がどのように有効に機能するのか，あるいは，どのような場合に有効に機能しにくいのかといった網羅的な結果を得ることができる．

（２）メニーコア向けの仮想化環境における省電力化技術の確立のためには，まず，GPU のハードウェア上の詳細を入手する必要がある．しかし，NVIDIA 社などの GPU はハードウェア上の仕様が公開されていない．そこで，オープンソースの NVIDIA 向け GPU デバイスドライバである Nouveau の開発チームからの協力を仰ぎ，ハードウェア上の仕様をある程度，入手するということから始めた．

現状の GPU は，ホスト CPU からはデバイスのひとつとして認識されるため，既存の仮想化技術のうち，デバイスの仮想化に用いる技術の援用ができる部分と，そうではない部分との切り分けを明確に行った．また，GPGPU は通常のデバイスとは異なり，内部にアドレス変換のためのページテーブルを持つなど，その仕組みがはるかに複雑である．そうした部分の仮想化においても，既存の仮想化技術が利用できる部分とそうでない部分との切り分けを明確に行った．

特に，ページテーブル周りについては，従来の仮想化技術で用いられているシャドー・ページテーブル（shadow page table）という技術を使い，さらにシャドー・ページテーブルの設定を巧妙に行うことで，仮想マシン間のアイソレーションを保証するというアプローチを取った．

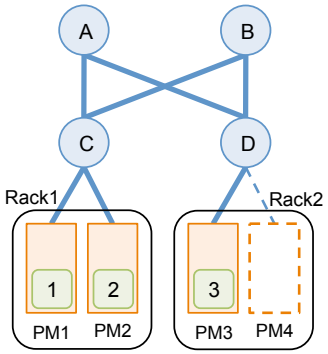
また，GPU 時間を仮想化するためには，仮想マシンモニタのレイヤにおいて， GPU に対するコマンドの発行をスケジューリングする必要がある．コマンドのスケジューリングを仮想マシンモニタが行うことによって，電力需要に応じて計算を遅延させるような処理が可能となり，システム全体での消費電力削減効果が期待できる．

４．研究成果

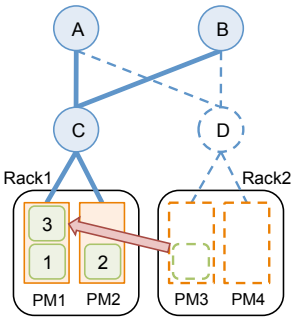
（１）仮想マシンモニタの持つライブ・マイグレーションという機能を使って，仮想マシンをできる限り少ない台数の物理マシン上に集約することによって消費電力量を削減するという手法はよく知られている．現在のデータセンタでは，ネットワーク機器による電力消費量も無視できないものとなっているという点に着目し，物理マシンだけではなくできるだけ多くのネットワーク機器も停止できるような形で仮想マシンをマイグレーションするという方式を確立した．

さらに，この手法ではネットワーク・スイッチの空きポートを利用し，ラック内に納められたいくつかの物理マシンを上位のネットワーク・スイッチに直接接続するという構成をとるようにする．このような構成をとることによって，あるラック内で動作するすべての仮想マシンを，上位のネットワーク・スイッチに直接接続された物理マシンに集約することができるようになる．そのような集約をおこなうと，下位のネットワーク・スイッチを停止できるというメリットがある．

また，このような構成をとることによって，ネットワーク機器の冗長性をそこなわずにネットワーク機器の停止が可能となるというメリットがある．下図に示した初期状態を考えよう．

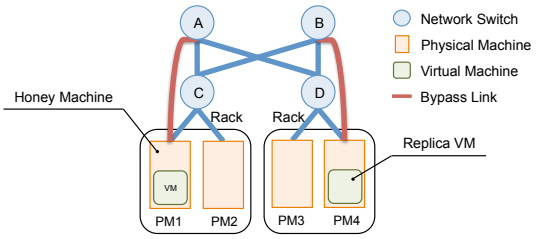


この状態で仮想マシン 3 を Rack1 にある物理マシンに移送を行うと下の図のようになり，ネットワークスイッチ D を停止できる．

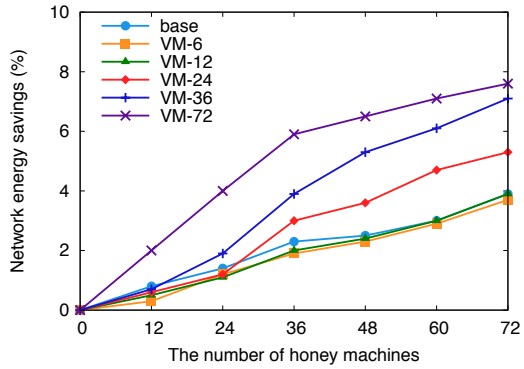


提案方式ではネットワーク・スイッチ A，B の空きポートを利用し，その空きポートからそれぞれ PM1，PM3 へのバイパスとなるネットワークを敷設する．このようにすることで，同じ状況においてもさらにネットワーク機

器の停止を行うことができる．その様子をしたの図に示す．この図をみると，赤色で示されたバイパスのネットワークを敷設することによって，ネットワークスイッチ D だけではなく，ネットワークスイッチ C も停止することができるようになっている．

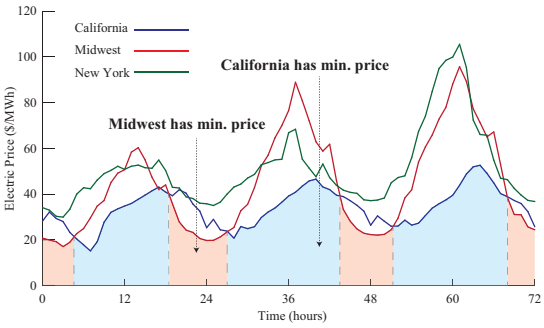


このような手法を用いて，シミュレーションによる電力消費効果の計測を行った．その結果を下図に示す．



このグラフに示すように，提案方式の効果が最も大きい場合，約 8% ほどの消費電力削減効果が得られていることがわかる．

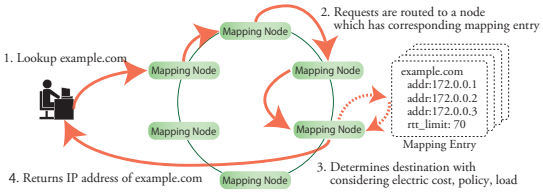
このようなクラウドを構成するデータセンタ内における省電力化だけではなく，複数のデータセンタから構成されるクラウド環境の場合，電力使用量が同じであっても，利用するデータセンタによって電力価格が異なるというケースがある．これは地域によって余剰電力の総量が異なり，余剰電力の多い地域では電力価格が安くなるという傾向にあるためである．したがって，電力価格の安い地域に設置されたデータセンタを積極的に利用することで，電力コストを削減できるという効果だけではなく，余剰電力を有効に活用することができるという環境上のメリットも多い．電力の自由化が行われている米国においてこの傾向が強い．その様子を下図に示す．



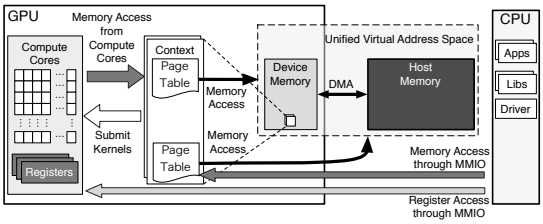
我が国においても電力の自由化や再利用可能エネルギーの利用が活発化しつつあり，近い将来により地域ごとの電力価格変動が生じると予測される．

上記のデータセンタ内省電力化機構に加え，このような地域毎の電力価格変動に合わせて，サービスを稼働させるデータセンタを選択することで，一層の電力コストの削減，および余剰電力の有効活用が可能となる．

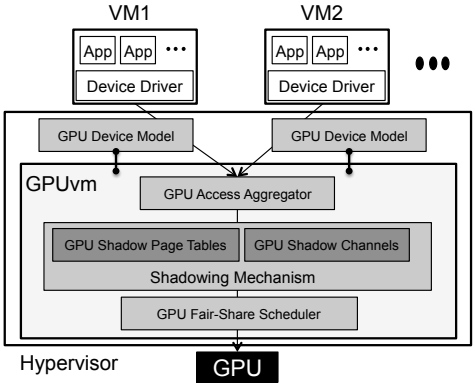
下の図に示したようなマッピング機能を介在させることにより，電力価格の変動に応じて適切なデータセンタの選択が可能となるようにした．



（2）メニーコア向けの省電力基盤の提供を行うため，GPU の仮想化を行った．GPU の仮想化を実現することによって，GPU に対するコマンドのスケジューリングを仮想マシンモニタの階層で行うことが可能となり，複数の仮想マシンからの要求を調停し，電力要求に合わせて GPU の使用量を粗粒度に調整することができるようになる．



上の図は GPU のハードウェア上の構成を外観したものである．ホスト CPU が GPU とやりとりする MMIO (memory-mapped I/O) の仮想化，GPU チャネルの仮想化，GPU ページテーブルの仮想化などを行い，GPU の仮想化を達成した．このアーキテクチャを下図に示す．



GPU Access Aggregator というモジュールによって GPU に対するコマンドリクエストがスケジューリングされるため，その部分に省電力化のための GPU スケジューリング機構を組み込むことが可能な仕組みとなっている．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　4　件）

① Hiroshi Yamada, Shuntaro Tonosaki, Kenji Kono, Efficient Update Activation for Virtual Machines in IaaS Cloud Computing Environments, IEICE Transactions on Information & Systems, 査読有, Vol. E97-D, No. 3, 2014, pp. 469-479. DOI: 10.1587/transinf.E97.D.469

② Hiroki Shirayanagi, Hiroshi Yamada, Kenji Kono, Honeyguide: A VM Migration-Aware Network Topology for Saving Energy Consumption in Data Center Networks, IEICE Transactions on Information & Systems, 査読有, Vol. E96-D, No. 9, 2013, pp. 2055-2064. DOI: 10.1587/transinf.E96.D.2055

③ Takumi Sakamoto, Hiroshi Yamada, Hikaru Horie, Kenji Kono, Energy-Price-Driven Request for Dispatching for Cloud Data Centers, IEEE International Conference on Cloud Computing, 査読有, 2012, pp. 974-976.

〔学会発表〕（計　6　件）

① Yusuke Suzuki, Shinpei Kato, Hiroshi Yamada, Kenji Kono, GPUvm: Why Not Virtualizing GPUs at Hypervisor?, USENIX Annual Technical Conference (USENIX ATC), 査読有, 2014. To appear.

② 白柳 広樹，山田浩史，河野健二，ネットワーク機器の省電力化のための仮想マシン移送を考慮したトポロジ，情報処理学会システムソフトウェアとオペレーティングシステム，査読無，2012.

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者
河野 健二（KONO, Kenji）
慶應義塾大学・理工学部・准教授
研究者番号：90301118

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：