様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １１ 日現在

機関番号：３２６８５
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２６６５

研究課題名（和文）並列化コンパイラによる解析情報を活用した仮想環境の省電力化・処理分散配置最適化

研究課題名（英文）Task Distribution/Assignment and Energy Efficiency Optimization on Virtual Environments by Utilizing Parallelizing Compiler Analysis Information of User Applications

研究代表者

和田　康孝（WADA, Yasutaka）

明星大学・情報学部・准教授

研究者番号：４０４３４３１０

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：クラウドサービスの基盤となっている仮想環境の高効率化を目的として，特に，a）並列アプリケーション実行時の性能と消費電力をモデル化する手法の提案と自動化フレームワークへの実装，b）比較的大規模なサーバと小規模なデバイスが連携して深層学習処理を高効率化する実行方式の提案，c）複数の仮想マシンからの要求を整理し，適切かつ安全に実ハードウェアの消費電力を制御する方式の提案，を行なった．これらの成果・提案により，クラウドシステムをより簡単に高効率化することができる．

研究成果の学術的意義や社会的意義
我々が日常的に利用するクラウドサービスを支えるコンピュータシステムはその規模・数を日々増大させており，サービスの質を保ちつつ消費電力を削減することは急務である．本研究課題はクラウドシステム，引いては並列システムや並列アプリケーションの高効率化を目指し，システムとアプリケーションの両面からそれを実現しようとするものである．その成果を発展・活用させることで，より利便性が高く，かつ持続可能性の高い社会の実現に資することができると考えられる．

研究成果の概要（英文）：In order to improve the efficiency of a virtualized environment, which is the basis of today's cloud services, we have realized/proposed: a) a power-performance modeling strategy for parallel applications and its implementation on a software framework, b) a task sharing method between a small-size edge device and a large server for an efficient deep-learning processing (both of learning and inference), and c) a management method and interface for power-performance control requests from multiple virtual machines to control an actual hardware appropriately and safely. With these achievements and proposals, it is possible to make cloud systems more efficient easily.

研究分野：計算機工学

キーワード：低消費電力化　性能モデリング　アプリケーション解析　仮想環境

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

クラウドコンピューティング環境への需要・要求は高まる一方であるが，それゆえに消費電力が大きな課題となる．例えば，2011 年時点で Google 社の持つデータセンタが消費する電力は 2 億ワット以上と報告されている．さらに，IoT 技術の進展とともにリアルタイムデータ処理への要求が高まっており，これを実現するために，センサノード等のデータ発生箇所により近い位置に小～中規模のサーバを配置する Fog Computing あるいは Edge Computing という考え方が標準化・導入されつつある．これにより，ハードウェア量がさらに拡大する傾向にある．以上から，ユーザに十分な計算資源を提供しつつ，消費電力を低減することは急務である．

データセンタにおいて低消費電力化・省エネルギー化を行う方策としては，空調の高効率化など，現在は主にハード面での技術が用いられているが，上述のハードウェア量の増加を考えると，ソフトウェアとハードウェアの協調によってこの問題を解決する必要がある．また，クラウドコンピューティングの基盤となる仮想環境において DVFS（Dynamic Voltage/Frequency Scaling）や PG（Power Gating）を適用する手法や，仮想マシンの配置を最適化する手法はこれまでも多く提案されているが，外部から取得できる情報のみに基づいており，ユーザアプリケーションの並列性や構造などの詳細な特性は活用されてこなかった．そのため，実際には必要のない過剰な計算資源を仮想マシンに割り当ててしまい，十分な消費電力削減が行われないという問題が依然として残る．

２．研究の目的

クラウド環境におけるビッグデータ処理のリアルタイム性を向上させる取り組みがなされているが，そのために消費電力の増加がこれまで以上に大きな問題となる．研究代表者らはこれまで，逐次アプリケーションを自動的に並列化するコンパイラ技術を活用し，並列アプリケーションを実行する際の消費電力を削減する手法を実現してきた．一方，クラウド基盤において広く用いられる仮想化環境に対するこれまでの消費電力削減手法では，上述の通り，アプリケーションの並列性や構造などの詳細な情報についてはほとんど考慮されてこなかった．そこで本研究課題では，コンパイラ等のツールによって得られるアプリケーション内部の情報を活用し，クラウド基盤の根幹をなす仮想環境において消費電力のさらなる削減を目指す．

３．研究の方法

本研究課題の目的を達成するためには，a）アプリケーションの特性，特に実行性能と消費電力の関係を解析・モデル化し，DVFS や PG などの電力制御の仕組みを適切に利活用すること，b）クラウド基盤（サーバ）とエッジ（センサ等を搭載する組込みデバイス）の間で適切に処理分担を行うこと，c）実際にアプリケーションが動作する仮想マシンからの要求を適切に取り扱い，実ハードウェアを制御すること，が特に必要となってくる．そこで，本研究課題を進めるにあたっては，特に以下の３点について検討・研究開発を実施した．

（１）アプリケーションの構造や特性を解析し，消費電力制御の最適化に活用する手法

上記 a）を実現するため，並列アプリケーション実行時の消費電力・性能の変化の様子を，プロファイリング結果等をもとにモデル化・推定するフレームワークについて研究開発を行った．その結果を用いることで，要求された性能あるいは消費電力に合わせて適切に DVFS 等の電力制御を適用することが可能となる．

（２）クラウドとエッジによる処理分担に関する検討

深層学習やビッグデータ等の，近年需要が高く，かつ計算資源に対する要求の高いアプリケーションを対象として，クラウド基盤（サーバ）とエッジ（組込みデバイス）の間で適切に処理を分担・配置し，全体として効率よくサービスを提供する事例について検討・評価を行なった．

（３）仮想環境上で物理ハードウェアを適切に制御する仕組み・手法

クラウド基盤上では，実際のアプリケーションは仮想マシン上で動作し，物理ハードウェアを直接制御することはできない．また，同一の物理ハードウェアで動作する別の仮想マシンの状況を直接監視することも難しい．そのため，物理ハードウェアを直接制御するホスト OS とその上で動作する仮想マシンを連携させることで，仮想環境においても適切に物理ハードウェアを制御するための仕組み・インタフェースについて検討を行う．

４．研究成果

（１）アプリケーションの構造や特性を解析し，消費電力制御の最適化に活用する手法

アプリケーションの特性に応じた電力制御を行うため，並列アプリケーションのプロファイリングおよび電力制御のための API を埋め込むソフトウェアフレームワークを拡張し，より高精度に消費電力と実行性能のモデル化を行う手法を検討した[1]．提案手法では，アプリケーションの実行性能と消費電力の関係を考慮し，事前にプロファイリングを通じた情報取

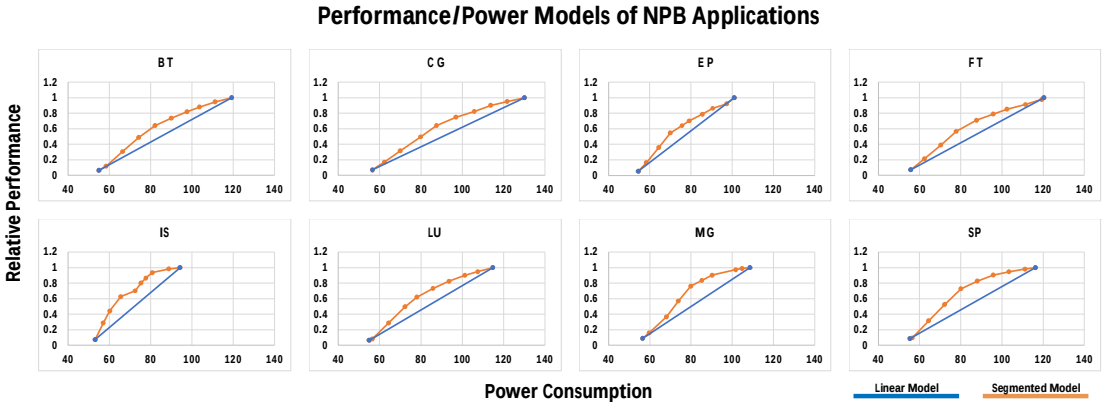

図１．アプリケーションごとの実行性能・消費電力の関係（[1]より引用）

得を行うことで，より精度高く電力・性能の関係をモデル化できる．また，フレームワークとして実装し簡易に一連の処理を行えるようにすることで，利便性を高めた．

その結果，図１に示すように，アプリケーション個別に実行性能と消費電力の関係をモデル化することが可能となった．図１では，消費電力を横軸，消費電力制御を行わない場合に対する実効性能を縦軸として，その間の関係を示している．

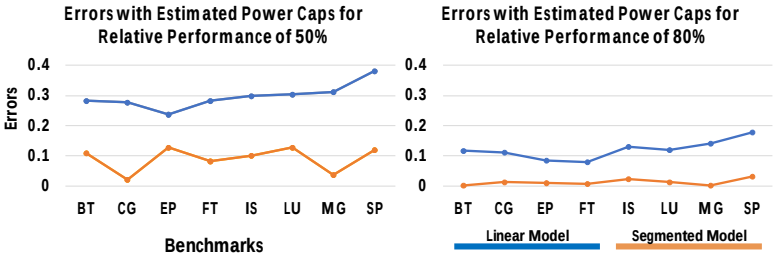

図２．単純な線形モデルと提案フレームワークの比較（[1]より引用）

また，図２では，電力制御を行わないときの 50[%]あるいは 80[%]の性能を保つように，図１のモデルをもとに電力制約をかけた際，実際の測定結果と消費電力にどの程度の差異があったかを示している．単純な線形の電力・性能モデルと比較して，アプリケーションの実行性能を本来の 50[%]に制限した際の消費電力の予測精度を，およそ 3 倍程度向上させることができた．この予測に基づいて電力制御を適用することで，アプリケーション実行性能の低下を抑えつつ，効果的にシステムの消費電力が削減できる．

（２）クラウドとエッジによる処理分担に関する検討

大規模なクラウド基盤のみに着目していては，システム全体の高効率化が難しいことから，近年需要の高まっている深層学習処理を対象に，クラウド（サーバ）とエッジ間で適切に処理分担し効率を向上させることができないか，という視点から検討を実施した．

図３に，検討・提案したクラウド・エッジ連携の仕組みを示す．エッジデバイス上では低電力な組込みシステム向け CPU や FPGA を用いて推論を実施し，センサから取り込んだデータをネットワークを介してサーバに送信する．サーバでは収集されたデータを用いて大規模な学習処理を実施する．学習結果は再度ネットワークを介してエッジデバイスに供給され，より高精度な推論を実施することができる．また，FPGA を利用する際には，そのための回路を配置配線・合成するような処理をサーバが担当することで，より効率よくシステム全体を利用することも考えられる[3]．

本研究課題では，FPGA を用いた画像認識を例題にプロトタイプを作成し，FPGA を備えたエッジデバイスによる推論と，サーバ上での学習処理を連携させ，継続的に推論精度を高めることが可能であることを確認した[2]．今後は，FPGA の動的再構成機能も利用し，さらなる効率化を目指す．

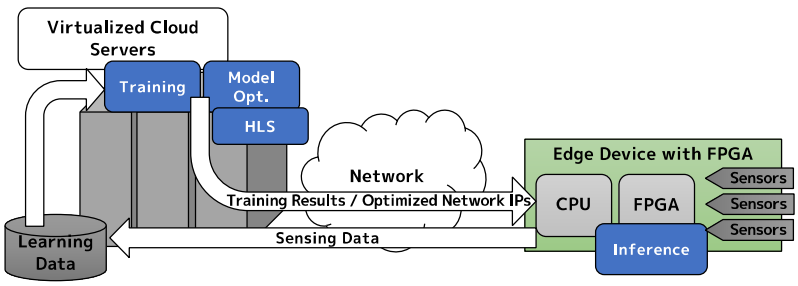

図３．エッジでの深層学習の推論処理を想定したエッジ・サーバ連携（[3]より引用）

（３）複数の仮想マシンが動作する状況を考慮した物理ハードウェアの制御

仮想環境上で動作する各仮想マシンから物理ハードウェアを直接制御する方式ではセキュ

リティが担保できない，各仮想マシンは互いに独立して動作するため互いの情報を直接やり取りできない，といった問題点を解決しつつ，仮想環境の低消費電力化を実現する仕組みについて研究開発を行った．

本研究では特に，クラウドシステムに広く用いられている OpenStack やその基盤となる Qemu, KVM 等を対象にして検討を進めた．本研究課題では，仮想マシンから直接物理ハードウェアを制御するのではなく，仮想マシンからの要求に応じて，ホスト OS が実ハードウェアの制御を行う手法を提案・検討した．提案手法では，図４に示すように，仮想マシン上で動作するアプリケーションからは仮想マシン上で動作するゲスト OS へ電力制御に関する要求を送出し，それを受け取ったゲスト OS はあるインタフェースを介してホスト OS にその要求を通知する．ホスト OS は複数のゲスト OS からの要求を取りまとめ，最終的にどのように実ハードウェアを制御するかを決定する．このような，仮想マシンから物理マシンのホスト OS に対して動作周波数制御の要求をそれぞれ通知し，それに従ってホスト OS が物理ハードウェアを制御する仕組みを提案し，その初期プロトタイピングを行った．これにより，複数の仮想マシンの状況に応じた簡単な制御を実現することができた．

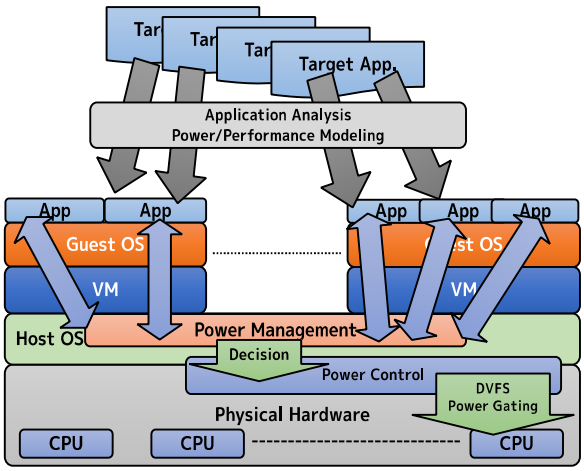

図４．仮想環境における電力制御（[3]より引用）

以上より，（１）並列アプリケーションの実行性能・消費電力の関係を適切にモデル化し，消費電力制御に活用する手法，（２）サーバ・エッジ間で処理を適切に分担することにより，アプリケーション全体を円滑に実行する仕組みの検討，（３）仮想マシン・物理マシン間の電力制御要求の授受を仲立ちするインタフェース・電力制御機構の提案，といった要素技術について成果を得ることができた．今後は，これらをさらに高度化・融合させ，より高効率なシステムの実現について研究開発を進める予定である．

参考文献

[1] Yuan He, Yasutaka Wada, Guanqin Pan, and Masaaki Kondo, "Simple DSL for Power-Performance Modeling with Segmented Linear Models", Proc. of the 48th International Conference on Parallel Processing (ICPP2019), Aug., 2019.

[2] 青戸 武蔵，和田 康孝，三ツ木 萌，"FPGA 上での CNN パラメータ動的更新手法の性能評価"，情報処理学会 第 82 回全国大会，7J-02，Mar.，2020.

[3] Yasutaka Wada, Ken'ya Onai, and Musashi Aoto, "Towards Energy-Efficient Neural Network Training on the Cloud for Effective Inference on IoT/Edge Devices", Proc. of Asia Pacific Conference on Robot IoT System Development and Platform 2019 (APRIS2019), Work-In-Progress Paper, Nov., 2019.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計13件（うち招待講演　4件／うち国際学会　8件）

１．発表者名 青戸武蔵，和田康孝，三ツ木萌
２．発表標題 FPGA上でのCNNパラメータ動的更新手法の性能評価
３．学会等名 情報処理学会　第82回全国大会
４．発表年 2020年

１．発表者名 Musashi Aoto, Moe Mitsugi, Takumi Momose, Yasutaka Wada
２．発表標題 Towards the Improvement of Training Efficiency and Image Recognition Accuracy for an FPGA Controlled Mini-Car by Offloading Neural Network Training
３．学会等名 2019 International Conference on Field-Programmable Technology (ICFPT)（国際学会）
４．発表年 2019年

１．発表者名 Yasutaka Wada, Musashi Aoto, Ken'ya Onai
２．発表標題 Towards Energy-Efficient Neural Network Training on the Cloud for Effective Inference on IoT/Edge Devices
３．学会等名 Asia Pacific Conference on Robot IoT System Development and Platform (APRIS2019)（国際学会）
４．発表年 2019年

１．発表者名 青戸武蔵，比留川翔哉，和田康孝，丸山一貴
２．発表標題 単機能なニューラルネットワークを複数用いた高速・高精度な画像認識のFPGAによる実現
３．学会等名 電子情報通信学会　リコンフィギャラブルシステム研究会（RECONF）
４．発表年 2019年

１．発表者名
Yuan He, Yasutaka Wada, Guanqin Pan, Masaaki Kondo

２．発表標題
Simple DSL for Power-Performance Modeling with Segmented Linear Models

３．学会等名
48th International Conference on Parallel Processing (ICPP2019)（国際学会）

４．発表年
2019年

１．発表者名
Musashi Aoto, Shoya Hirukawa, Kazutaka Maruyama, Yasutaka Wada

２．発表標題
An FPGA based Autonomous Driving Car Design using Multiple Simple Neural Networks for Decision Making

３．学会等名
The 10th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART 2019)（国際学会）

４．発表年
2019年

１．発表者名
Yasutaka Wada

２．発表標題
Software-Based Resource Management Techniques for Computer Systems of Various Scales

３．学会等名
International Conference on Soft Computing and Machine Learning (SCML2019)（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
和田康孝

２．発表標題
ソフトとハードの連携によるコンピューターシステムの高効率化に向けた取り組み

３．学会等名
令和元年度イノベーション多摩支援事業 「大学研究室見学会シリーズ～明星大学編～」（招待講演）

４．発表年
2020年

１．発表者名
和田康孝

２．発表標題
ソフトウェア・ハードウェア連携による深層学習処理の高効率化・高精度化・応用に向けた取り組み～エッジ・クラウド間連携を例にして～

３．学会等名
一般社団法人電子実装工学研究所（IMSI）公開講演会「電子実装工学の新潮流：ディープラーニング電子システム応用と常温接合パワーモジュール応用」（招待講演）

４．発表年
2019年

１．発表者名
Yasutaka Wada

２．発表標題
Parallel Processing and Parallelizing Compilation Technique for "Green Computing"

３．学会等名
The First International Colloquium of Mexican and Japanese Studies（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Musashi Aoto, Yousuke Numata, Yasutaka Wada

２．発表標題
Development of an FPGA controlled "Mini-Car" toward Autonomous Driving

３．学会等名
The 2018 International Conference on Field-Programmable Technology（国際学会）

４．発表年
2018年

１．発表者名
Yasutaka Wada, Daisuke Ogawa, Kanji Otsuka, Yoichi Sato

２．発表標題
Toward Software-Hardware Cooperative Systems for Energy Efficiency with Virtualization Platforms

３．学会等名
Exhibition, The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC18)（国際学会）

４．発表年
2018年

１．発表者名 樋口耀，和田康孝，山中脩也
２．発表標題 使用コア数の動的変更による仮想環境上でのIn-Situ Visualizationの高速化に関する検討
３．学会等名 情報処理学会第80回全国大会
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	和　遠 (HE Yuan)	the School of Information Science and Engineering, Shenyang University of Technology・Associate Professor	