「情報システムの超低消費電力化を目指した技術革新と統合化技術」
平成 19 年度採択研究代表者

H23 年度
実績報告

松岡　聡

東京工業大学　学術国際情報センター・教授

ULP-HPC: 次世代テクノロジのモデル化・最適化による超低消費電力ハイパフォーマンスコンピューティング

§1. 研究実施体制

(1)「東工大・松岡」グループ

① 研究代表者:松岡　聡 （東京工業大学　学術国際情報センター、教授）

② 研究項目

次世代 HPC システムにて超省電力・高性能を達成するハードウェア・ソフトウェア統合システムの研究開発

(2)「東大」グループ

① 主たる共同研究者:須田　礼二 （東京大学大学院　情報理工学系研究科、教授）

② 研究項目

超省電力 HPC システムのための自動チューニング技術の研究開発

(3)「東工大・青木」グループ

① 主たる共同研究者:青木　尊之 （東京工業大学　学術国際情報センター、教授）

② 研究項目

超省電力型の HPC アプリケーション及びアルゴリズムの研究開発

(4)「電通大」グループ

① 主たる共同研究者:本多　弘樹 （電気通信大学大学院　情報システム学研究科、教授）

② 研究項目

超省電力化 SIMD アクセラレータのための汎用プログラミング環境

(5)「NII」グループ

① 主たる共同研究者:鯉渕　道紘　(国立情報学研究所　アーキテクチャ科学研究系、准教授)

② 研究項目

省電力インターコネクトの研究開発

§2. 研究実施内容

(文中に番号がある場合は(3－1)に対応する)

[概要]

本年度も超低消費電力高性能計算の研究を、次世代 HW システムの利用技術(松岡 G・本多 G・鯉渕 G)、自動チューニング数理基盤およびソフトウェア(須田 G)、超省電力大規模アプリケーション(青木 G)の体制で推進した。特筆すべきこととしては以下のトピックが挙げられる。青木 G を中心に東工大 TSUBAME2.0 スパコンの 4000GPU クラスのスケーラビリティをデンドライト凝固シミュレーションにおいて実証し、電力測定も行った。その業績に対して高性能計算分野の最も著名な賞である Gordon Bell 賞(Special Achievement Award)が与えられた。また本年度は東日本大震災に伴う電力危機が社会問題となったが、その要請に応えるために TSUBAME2.0 を運用する東工大 GSIC と共同でピークシフト運用を実現し、昼間に 20%以上の電力削減を達成した。

【研究代表者・松岡 G】

松岡グループでは主に大規模計算機システムにおけるアクセラレータ利用技術を含むソフトウェアによる超低消費電力化を担当している。本年度は東工大 TSUBAME2.0 スパコンにおけるピークシフト運用と電力可視化の実現および、多数アクセラレータにおける数値演算アルゴリズムの改良等数多くの成果をあげた。

- 東日本大震災に伴う原発事故の影響で、全国的な電力供給危機が発生した。その情勢に対応するため、TSUBAME2.0 はすでに運用スパコン省電力世界一となっていた[A-5, A-9, A-37, A-39 など]が、さらに消費電力を削減する必要が生じた、ULP-HPC 松岡グループと GSIC の共同で、ピークシフト運用の検討および実現を行った[A-49]。稼働率とスケジュール手法の検討を行い、昼間はシステムの 70%、夜間は 100%運用とする自動化ツールの共同開発を行った。この運用は四月中旬から九月下旬にかけて行われ、ほぼ全期間において昼間電力を今回の目標値の 787kW 以下に抑えることに成功した。
- 上記に関連して、システム全体電力をリアルタイムに把握するため、分電盤の電力を分のオーダーでグラフィカルに確認可能なソフトウェアツールを開発し、web 公開した[A-49]。これは運用の意思決定に使われたのに加え、大規模 GPU ジョブの電力測定にも用いられた。4000GPU を用いたデンドライト凝固シミュレーション[C-2](青木 G, Gordon Bell 賞 Special Achievement award 受賞)、冠動脈血流シミュレーション[A-2](Gordon Bell 賞奨励賞受賞)などの測定により、実アプリケーションにおける優れた電力性能比を示した。

- アクセラレータ上の高性能省電力ソフトウェアの研究を以下のように行った。まず FFT に関しては複数 GPU への対応を行った。そもそも FFT においては通信量が大きいために複数 GPU 化は不利であるが、3D-RISM 等の実アプリに組み込む際に必要が生じうる。ソフトウェアパイプライニング手法などにより、4 NVIDIA GPU を搭載するシングルノードでは約 1.8 倍、32 ノードシステムでは約 5 倍の性能向上率(strong scaling)を実現した[A-1]。さらに FFT ライブラリの適用アーキテクチャを広げるために OpenCL 版も開発した。AMD Radeon GPU を用いた性能評価では、メモリバンド幅の差異などのために NVIDIA GPU より高性能という結果が得られた[A-59]。さらに、GPU 上における大規模ステンシル演算[A-3,A-58]、有限要素法[A-55]、Fast multipole method[A-50, A-62]、グラフ解析[A-51,A-63,A-64]の実装と性能解析を行い、良好なスケーラビリティを得た。このような複数 GPU を用いたアプリケーションの高信頼化技術についても、現状の CUDA システムの問題点の指摘および、スケーラビリティの検証を行った[A-4, A-8]。

【須田 G】

東大グループでは、高性能・低消費電力のための自動チューニング技術の研究を展開している。我々はこの研究課題を次の 5 つのテーマに分けて進めている。すなわち、(1) 計算システムの消費電力モデル, (2) GPU システムにおける高性能化・低消費電力化技術開発, (3) 低消費電力化自動チューニング数理基盤, (4) 低消費電力化自動チューニングプログラミングシステム, (5) 超低消費電力数値ライブラリ、である。

(1) 計算システムの消費電力モデルにおいては, GPU・マルチコア結合システムにおける性能と電力のモデル化の研究を一層進めた. 特に GPU の性能モデル・電力モデルにおいては, GPU の命令ごとの性能・電力をもとに, GPU コアの特性をパラメタ化して, 高精度で推定する手法を開発した(原著論文 B-2). (2) GPU 高性能化については, 引き続き大規模木探索で成果を挙げ, 世界最大サイズの探索を実施した. (3) 低消費電力化自動チューニング数理では, これまでの知見をライブラリ ATMathCoreLib としてまとめて公開した. また変動する条件に対する自動チューニング数理手法を開発した. (4) 低消費電力化自動チューニングプログラミングシステムでは, 電通大グループと協力して HxABCLibScript を開発した. また C から指示行なしに CUDA に変換する APTCC の開発を進めている(原著論文 B-3). (5) 超低消費電力数値ライブラリでは, 電力性能測定汎用 API を開発中である.

【本多 G】

電通大グループにおいては、省電力化に有効な SIMD 型アクセラレータの有効活用を目指し、特に GPU に対してその特徴である CPU から分離された演算コアとメモリに対応可能な並列プログラミングインタフェースの仕様を考察し、評価を行っている。この際に、実行性能のみ考慮されていた既存のシステムに加えてアプリケーションの消費エネルギーを最適化可能とするシステムの実行モデルの考案および実装を行なっている。

具体的には、NVIDIA 社が提供する GPU コンピューティングフレームワークである CUDA に対応するOpenMP処理系であるOMPCUDAの開発を行っている。実行時に判明する計算データサイズが小さい場合において GPU を使用することは消費エネルギーおよび実行性能を悪化させるため、動的にCPU使用とGPU使用を自動的に判断するチューニング機構を追加し、問題サイズに応じてオーバヘッドを最小限に抑える実行を可能とした。また、GPU 内の並列度だけではなく複数のGPUにまたがったGPU外の並列度を活用した実行を可能とする拡張を行ない、使用するGPUの数に比例して回数が多くなるGPU呼び出しやメモリ確保回数を最小化させる最適化を進めた[D-2] 。

また、東大グループと連携してアプリケーションからシステムまで統合的にチューニング可能とする省電力チューニングフレームワークの開発を行っている。具体的には省電力チューニングフレームワークにおいてユーザが指定するべきポリシーを用いて、CPU および GPU でデータを分担実行する際に最大の実行性能を達成するための実行、最小の消費エネルギーを達成するための実行を選択するチューニングが可能となることを明らかにした[D-1]。

【鯉渕 G】

NII グループでは，並列アプリケーション毎にインターコネクトの電力最適化を行う On/Off リンクアクティベーション法を、スイッチの各ポートのリンク速度を独立に変更する拡張，安定化を行い、イーサネット上においてHPL, NAS パラレルベンチマークなどを用いた評価を行った．その結果，評価を行った PC クラスタでは1%の性能低下でネットワーク部（スイッチとリンク）の消費電力を23%削減することができた[E-1].

さらに On/Off リンクアクティベーション法の適用可能なリンク数を最大にするためのデッドロックフリールーティングについても定量的な比較を行った．具体的には任意のトポロジに適用することができるデッドロックフリールーティングについて評価を行い，我々が以前提案した L-turn ルーティングは必要とするネットワーク資源と性能のバランスが優れていることが分かった[E-2].

また，近年，HPC システムの巨大化と並列アプリケーションの並列数の増大により，通信遅延を削減しつつ，省電力技術を適用することが要求されることが多い．そのためこの観点で優れたトポロジとして，既存のトポロジにランダムにショートカットリンクを加える新しいHPCシステム用のトポロジを提案，探求した．その結果，同一次数のトポロジと比べて，ホップ数の低下を実現することにより通信遅延が最大35%削減することができた[E-3][E-4].

以上のように，ULP-HPC の要素利用技術として省電力インターコネクト技術の推進のために理論的な側面と実評価による有効性と feasibility に関する成果を得ることができた．

[今後の見通し]

最終年度となる H24 年度には統合システムとして、大規模スパコンのパワーキャッピングを目的としてフィードバック制御システムの研究開発を行う予定である。本年度に行ったピークシフト運用はすでにそれを部分的に実現したものであるものの、その制御は、緊急に運用を開始しなければならなかったためではあるが、ON/OFF の単純な二値制御であり、また削減ノード数も経験則に基づく固定数であるという制限があった。その解決のためには、今年度の節電運用経験、前年度

までの性能・電力モデリングおよび須田・本多 G の自動チューニングソフトウェアの統合を予定し、青木 G の超大規模 GPU アプリケーションによる実証実験を行う。

§3. 成果発表等

（3－1） 原著論文発表

【松岡 G】

●論文詳細情報

[A-1] Akira Nukada, Yutaka Maruyama, Satoshi Matsuoka. “High Performance 3-D FFT using multiple CUDA GPUs”, In Proceedings of the Fifth Workshop on General Purpose Processing using Graphics Processing Units (GPGPU-5), London, UK, 7 pages, ACM Press, Mar. 2012.

[A-2] Massimo Bernaschi, Mauro Bisson, Toshio Endo, Massimiliano Fatica, Satoshi Matsuoka, Simone Melchionna, Sauro Succi, "Petaflop Biofluidics Simulations On A Two Million-Core System", In Proceedings of ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC11), Gordon Bell Paper, ACM Press, Nov. 2011.

[A-3] Naoya Maruyama, Tatsuo Nomura, Kento Sato, Satoshi Matsuoka, Physis: An Implicitly Parallel Programming Model for Stencil Computations on Large-Scale GPU-accelerated Supercomputers, Proceedings of the 2010 ACM/IEEE conference on Supercomputing (SC'11), 2011/11/15, Seattle, WA, USA

[A-4] Leonardo Bautista, Naoya Maruyama, Dimitri Komatitsch, Tsuboi Seiji, Franck Cappello, Satoshi Matsuoka, Nakamura Takeshi.FTI: High performance Fault Tolerance Interface for hybrid systems..In International Conference for High Performance Computing, Networking, Storage and Analysis (SC).Page 1-12.Nov. 2011.

[A-5] 遠藤 敏夫, 額田 彰, 松岡 聡. **スーパーコンピュータ** TSUBAME 2.0 における Linpack 性能 1 ペタフロップス超の達成. 情報処理学会論文誌コンピューティングシステム, Vol. 4, No.4 (ACS 35), pp.169—179, 2011 年 10 月.

[A-6] 滝澤 真一朗,松岡 聡,友石 正彦,佐藤 仁,東田 学.Point-of-Presence 連携による e-サイエンス分散環境..In インターネットカンファレンス 2011.Oct. 2011.

[A-7] Shuntaro Yamazaki, Akira Nukada, Masaaki Mochimaru, “Hamming Color Code for Dense and Robust One-shot 3D Scanning”, In Proc. of the 2011 British Machine Vision Conference, Dundee, Scotland, Springer, Aug. 2011.

[A-8] Akira Nukada, Hiroyuki Takizawa, Satoshi Matsuoka. NVCR: A Transparent Checkpoint-Restart Library for NVIDIA CUDA. Proceedings of the 20th International Heterogeneity in Computing Workshop (HCW 2011), in conjunction with IEEE IPDPS 2011.The IEEE Press. In The 20th International Heterogeneity in Computing Workshop (HCW 2011), in conjunction with IEEE IPDPS 2011.page 1--10.May. 2011.

[A-9] 遠藤 敏夫,額田 彰,松岡 聡.スーパーコンピュータ TSUBAME 2.0 における Linpack 性能 1 ペタフロップス超の達成.情報処理学会 SACSIS2011 論文集.情報処理学会.In 先進的計算基盤システムシンポジウム(SACSIS2011).pp. 1-8.May. 2011.

[A-10] Sumeth Lerthirunwong, Hitoshi Sato, Satoshi Matsuoka. "Multi-ring Structured Overlay Network for the Inter-cloud Computing Environment", In Proceedings of the 1st International Conference on Cloud Computing and Services Science (CLOSER 2011), pp. 5--14, Noordwijkerhout, Netherlands, 7-9 May, 2011.

[A-11] Mohamed Amin JABRI and Satoshi MATSUOKA. "Dealing with Grid-Computing Authorization using Identity-Based Certificateless Proxy Signature", In Proceedings of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2011), pp. 544--553, Newport Beach, CA, May 2011.doi=10.1109/CCGrid.2011.12

【須田 G】

●論文詳細情報

[B-1] Da Qi Ren, Reiji Suda, "Global Optimization Model on Power Efficiency of GPU and Multicore Processing Element for SIMD Computing with CUDA", International Conference on Energy-Aware High Performance Computing / Computer Science - Research and Development, Springer, 2011 for online first, DOI:10.1007/s00450-011-0197-6.

[B-2] Cheng Luo, Reiji Suda, A performance and energy consumption analytical model for GPU, International Conference on Cloud and Green Computing (CGC2011), Sydney, Australia, Dec 2011, 8 pages DOI: 10.1109/DASC.2011.117

[B-3] Takehiko Nawata and Reiji Suda ``APTCC: Auto Parallelizing Translator From C To CUDA", Proceedings of the International Conference on Computational Science, Procedia Computer Science, Volume 4, pp 352-361, 2011, doi:10.1016/j.procs.2011.04.037

[B-4] Da Qi Ren, Reiji Suda, “Experimental Estimation and Analysis of the Performance and Power Efficiency of CUDA Processing Element in SIMD Computation”, Proceeding of the 10th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2011), pp.405-408, Sanya, China, May 16-18,

2011. DOI: 10.1109/ICIS.2011.74

【青木 G】

●論文詳細情報

[C-1] S. Saito, K. Ohno, M. Sekijima, T. Suzuki, and H. Sakuraba,"Database of the clinical phenotypes, genotypes, and mutant arylsulfatase B structures in mucopolysaccharidosis type VI", J Hum Genet, Feb.2012, doi:10.1038/jhg.2012.6

[C-2] T. Shinozaki, T. Iwaki, S. Du, M. Sekijima, S. Furui, "Distance-based Factor Graph Linearization and Sampled Max-sum Algorithm for Efficient 3D Potential Decoding of Macromolecules", IPSJ Transactions on Bioinformatics.

[C-3] T. Shimokawabe, T. Aoki, T. Takaki, A. Yamanaka, A. Nukada, T. Endo, N., Maruyama, S. Matsuoka: Peta-scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer, in Proceedings of the 2011 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, SC'11, IEEE Computer Society, Seattle, WA, USA, Nov. 15, 2011, SC'11 Technical Papers

(科学技術計算としてステンシル計算としては世界で初めて TSUBAME2.0 上で 2 ペタフロップスというペタフロップス超の性能を達成し、ACM ゴードンベル賞を受賞、同時に 1.5GFlops/W と、Linpack を超える性能電力比を実アプリケーションで示した。)

[C-4] T. Udagawa and M. Sekijima, "The power efficiency of GPUs in multi nodes environment with molecular dynamics", In Proceedings of the 2011 International Conference on Parallel Processing Workshops, pp.399-405, 2011.

[C-5] Wang Xian and Aoki Takayuki: Multi-GPU performance of incompressible flow computation by lattice Boltzmann method on GPU cluster, Parallel Computing, pp.521-535, September 2011, doi:10.1016/j.parco.2011.02.007

[C-6] T. Miki, X. Wang, T. Aoki, Y. Imai, T. Ishikawa, K. Takase and T. Yamaguchi: Patient-specific modelling of pulmonary airflow using GPU cluster for the application in medical practice, Computer Methods in Biomechanics and Biomedical Engineering, DOI:10.1080/10255842.2011.560842, online: 02 Aug 2011

[C-7] Naoyuki Onodera, Takayuki Aoki, Hiromichi Kobayashi: Large-eddy simulation of turbulent channel flows with conservative IDO scheme, Journal of Computational Physics, Volume 230, Issue 14, 20 June 2011, Pages 5787-5805 (2011)

[C-8] 丹愛彦, 青木尊之, 井上景介, 吉谷清文: 回転体に駆動される気液二相流の数値計算, 日本機械学会論文集 B 編, Vol.77, No.781, 1699-1714, (2011)

【鯉渕 G】

●論文詳細情報

[E-1] Michihiro Koibuchi, Takafumi Watanabe, Atsushi Minamihata, Masahiro Nakao, Tomoyuki Hiroyasu, Hiroki Matsutani, Hideharu Amano, Performance Evaluation of Power-aware Multi-tree Ethernet for HPC Interconnects, The Second International Conference on Networking and Computing, pp.50-57, Nov, 2011

[E-2] J. Flich, T. Skeie, A.Mejia, O. Lysne, P. Lopez, A. Robles, J. Duato, M. Koibuchi, T. Rokicki, and J. C. Sancho, ``A Survey and Evaluation of Topology Agnostic Routing Algorithms'', IEEE Transactions on Parallel and Distributed Systems, Vol.23, No.3, pp.405-425, Mar. 2012 (DOI: 10.1109/TPDS.2011.190)

[E-3] 鯉渕 道紘, 松谷 宏紀, 天野 英晴, D. Frank Hsu Casanova Henri, 高性能計算機インターコネクトにおけるランダムショートカットトポロジ, ハイパフォーマンスコンピューティングと計算科学シンポジウム(HPCS)Jan 2012, pp.85-92, Jan 2012

[E-4] Michihiro Koibuchi, Hiroki Matsutani, Hideharu Amano, D. Frank Hsu, Henri Casanova, ``A Case for Random Shortcut Topologies for HPC Interconnects'', The 39th International Symposium on Computer Architecture (ISCA), June 2012 (accepted with shepherd)