平成 20 年度
実績報告

「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

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. 研究実施の概要

平成 20 年度は引き続き超低消費電力の研究を、システム・アプリケーション・数理に基づいたチューニングの観点から推進した。システム面では GPU などのアクセラレータ、メモリ、ネットワークなどの構成要素の電力・性能モデルの精緻化およびそれらに基づく最適化手法を数多く提案した。これらの結果を活用するための、プログラミングシステムの研究も推進した。HPC アプリケーションとしては、高精度流体計算を GPU を用いて CPU の数 10 倍以上高速化することに成功している。GPU の消費電力は 170W 程度であり、CPU の場合の数 % 以下のエネルギーで計算できることが明らかになった。チューニングについては、事前情報が適切でない場合でさえも漸近的に最適な選択に近づくロバストな手法を提案した。この結果はグループ間で共有することにより、さらなる最適化を想定している。また引き続きグループ間で共用するためのプラットフォームの導入・保守を行っている。その一環として、本プロジェクトの成果を活用し、東京工業大学 TSUBAME スーパーコンピュータに大規模な GPU 群を導入し、GPU を積極的に用いるスパコンとしては世界最大のものとなり、高い注目を浴びた。

2. 研究実施内容（文中にある参照番号は 4.(1)に対応する）

「研究代表者・松岡」グループ（東京工業大学）

- GPU 上の疎行列ベクトル積、FFT などの数値計算カーネルのさらなる効率化の研究を推進した[A-3]。連立1次方程式の反復解法などで重要な疎行列ベクトル積の計算を高いメモリバンド幅を持つ GPU に最適な計算手法を提案した[A-10][A-12]。 CUDA GPU の各種キャッシュメモリを適切に使い分けることによりあらゆる行列で1GPU で 16 コアを超え、最大で約30倍の性能を実現し、電力効率の大きな向上を果たした。また複数 GPU を用いることでデバイスメモリより大きな行列サイズにも対応可能で、また使用する GPU 数に対してスケーラブルな性能向上が得られた。三次元 FFT 計算において、GPU 内部のメモリアクセスの最適化手

法を考案・実装した[A-7][A-8][A-11]。これによりGPU内での性能で約140GFlopsを達成した。これは従来最速であったNVIDIA社の純正ライブラリと比べ約3倍の性能向上である。このとき電力性能比においては、最新のQuad core CPUを用いる場合より約4倍優れることを示した。また三次元FFTにおいて問題サイズがGPUメモリより大きく、計算中にホストと(場合によっては複数の)GPU間の通信が必須となる場合のスケジューリング最適化にも取り組んだ。また引き続きCPUとGPUを併用するためのモデリングの精緻化を行った[A-3][A-6]。これらの過程において既存のAPIの改良すべき点を指摘し、GPUメーカーとの連携も行っている。

- GPUの耐故障性の研究を推進した。GPUは上述のようにCPUに比べて電力効率に優れているが、その信頼性・耐故障性は明らかではない。信頼性・耐故障性向上に向けて、GPUメモリの誤り検出をソフトウェアによって行う手法の設計・初期評価を行った。本手法では、GPGPUアプリケーション中にECCを計算、検査するコードを追加することで、ビットフリップなどの誤りを検出、訂正する。提案手法を行列積とN体問題に適用したところ、行列積で最大300%程度、N体問題で15%程度のオーバヘッドになることを確認し、N体問題のようにメモリアクセス頻度に対して計算量の多いアプリケーションでは比較的小さなオーバヘッドで実現可能であることを確認した。
- 省電力型アクセラレータによるHPCの大規模加速実験およびモデリングの研究を推進した[A-1]。大規模加速実験として、東京工業大学TSUBAMEシステム上で並列Linpackの実験を、これまでに導入したClearSpeedアクセラレータに加え、TESLA GPUを併用して行った。10000コア以上の汎用CPU、600枚以上のClearSpeed、600枚以上のTESLA GPU(後述)という大規模ヘテロ型システムにおいて、それぞれの計算資源にカーネル演算を適切に割り振ることにより77.48TFlopsの性能を達成した。この結果はTop500スパコンランキングにおいて世界29位にランクされた。ヘテロ型システムとしてはLANL RoadRunnerに次ぐ二位である。またTSUBAMEのような汎用CPUとアクセラレータが混在するHPCシステムを想定し、ジョブスケジューリングアルゴリズムの研究[A-9]を推進している。アルゴリズムでは各ジョブの加速特性の情報を知ることができるという仮定のもと、ジョブを適切なプロセッサに割り当てる。シミュレーションによりmake-spanおよびED積の評価を行い、ECT(earliest completion time)方式などの、より詳細な情報を仮定する方式と同等かそれ以上の性能を示した。提案方式はECTに比べ最大5%、単純な割り振り方法に比べ最大75%、ED積を改善した。
- 次世代低電力メモリを有効利用するシステムの研究を推進した[A-2]。電力コストが大きいDRAMの搭載容量を削減するために、これまでにメインメモリの一部をMRAMに置き換え、さらにスワップデバイスとしてFLASHメモリを使用するアーキテクチャを提案し、スワップを起こしてでもDRAM容量削減することによりエネルギー消費を抑制できることを示してきた。その上でアプリケーションの実行中のページング処理に関する消費電力を最小にする手法を提案し、実験中である。具体的にはさまざまなDRAM容量の場合のスワップ処理の合計コストを、一度のアプリケーション実行中に見つけるものである。類似手法と比較して、クリーンページとダーティーページの置換処理コストが異なることに注目することに特色がある。
- 東京工業大学学術国際情報センターでは本年度TSUBAMEスパコンにTESLA S1070 GPUを170台、680デバイス導入した。これは松岡代表が中心となり、これまでのCRESTにおける実験結果/研究成果[A-1][A-3][A-4][A-6][A-7][A-8][A-10]に基づくことにより可能となっ

たものである。この導入により TSUBAME の演算性能は 110TFlops→170TFlops(倍精度ピーク)および 170TFlops→870TFlops(単精度ピーク)と大幅に向上した。一方消費電力については、15%以下の上昇で抑えられている。これらを用いてすでに上記の Linpack 実験に成功しているが、今後さらに多様な HPC アプリケーションの電力最適化の研究をこれらの資源を用いて推進していく。

「主たる共同研究者①・須田」グループ（東京大学）

自動チューニング研究グループでは、低消費電力のための自動チューニング技術の研究開発を進めた。須田は、電力のように測定誤差が避けられないものを目的関数としたソフトウェア自動チューニングに適した数理基盤として、Bayes 統計を基礎とした数理手法を提案してきた。本年度は、ユーザから与えられた事前情報が適切でない場合でも、漸近的に最適な選択肢を確実に探し出すロバストな数理手法を定式化し、具体的な手法を提示した。黒田と Ren は、行列計算における電力消費の実態を調査した。このうち黒田は、消費電力を最小にするパラメタが計算システムに依存することを示した。また、Ren は消費電力モデルを構築し、実際の電力消費とよい一致が得られることを示した［B-1］。片桐は、自動チューニングプログラミングのための言語である ABCLibScript を低消費電力向けに拡張した。また、小谷は機械学習の手法に基づく自動チューニングのための実験計画法を開発した。このほか、超低消費電力スーパーコンピュータにおける高性能ソフトウェアに向けて、マルチコア超並列計算システムにおける線形解法や固有値計算の自動チューニング手法の実装、GPU を含む並列システムにおけるスケジューリング［B-2］や動的負荷分散［B-3］の手法の開発、数値計算アルゴリズムの研究開発を行った。

「主たる共同研究者②・青木」グループ（東京工業大学）

電力当たりのピーク性能が通常の CPU より 1 桁以上高い GPU を用い、HPC の主要アプリケーションである流体計算の高速化を試みた。産業応用などの利用が多い非圧縮性流体解析については、移流項に Cubic セミラグランジュ法を適用し、圧力の Poisson 方程式を Red & Black アルゴリズムとマルチグリッド法で解き、空力解析の問題としては典型的な円柱周りの流れを GPU を用いて CPU の 20～30 倍高速に計算することができた。また、圧縮性流体解析としては、レーリーテーラー不安定性の成長過程の計算に対し、CPU の数 10 倍の加速を達成した。数値計算手法としても最新の保存形 IDO 法を用いて Euler 方程式を高精度に解いていて、空間には格子点上の値と格子点間の積分値を従属変数として定義し、高次精度補間関数を構築している。また、浅水波方程式を解くことりよる高精度津波シミュレーションでは、実地形に対して遡上を含む津波の振る舞いを CPU の数 10 倍高速に計算することができた。流体計算は GPU を用いる場合でもメモリバンド幅が計算速度の律則となっていて、隣接格子点への複数回の参照に対し、GPU のオンチップの高速な共有メモリをキャッシュ・メモリ的に用いることによりアクセス回数を低減するアルゴリズムを開発した。材料科学の分野で重要な Cahn-Hilliard 方程式の 3 次元計算では 4 階微分の離散化が多数の隣接格子点データの参照を必要とするが、上記アルゴリズムの適用により CPU での計算と比較して約 160 倍の高速化を達成することができた。

これらのアプリケーションを GPU 上で計算しているときの消費電力をデジタルオシロスコープで詳細に計測し、GeForce GTX 280 では 160～200W 程度であり、待機電力から上昇分は 20W～

40W であることが分かった。学術レベルの流体計算が GPU では CPU の数 % 以下のエネルギーで計算できることが明らかになった。

「主たる共同研究者③・本多」グループ（電気通信大学）

本多(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その特徴である CPU から分離された演算コア、メモリに対応する並列プログラミングインタフェースの仕様を考察し、評価した。具体的には、SIMD 型アクセラレータ固有のプログラミングインタフェースが持つ性能を可能な限り維持しつつ使用することが可能なプログラミングインタフェースとして OpenMP プログラミング環境の処理系を提供することを目指し、NVIDIA 社が提供する GPGPU プログラミングインタフェースである CUDA に対応する OpenMP 処理系の実装を行った。その際、GPU 内の SIMD 型プロセッサで実行するタスクのスケジューリングと演算後の結果を集約するリダクション演算の最適化を施し、行列演算において CPU による並列実行に対して約 62 倍の高速化を達成することができた[D-2]。

また、SIMD 型アクセラレータの持つそれぞれ異なる SIMD 型命令を統一的に記述できる SIMD 共通記述方式を開発し[D-3]、高性能 SIMD 型アクセラレータに対する可搬性のあるプログラミング環境の検討を行った[D-1]。

「主たる共同研究者④・鯉渕」グループ（国立情報学研究所）

昨年度に提案したリンクの On/Off アクティベーション法を発展させ、PC クラスタのインターコネクトの主流であるイーサネットにおいて実装した。イーサネットにおいて On/Off リンクアクティベーション法は、VLAN ルーティング法を応用することで任意のトポロジにおいてブロードキャストストームを避けつつ、ホスト間の経路を更新する。On/Off リンクアクティベーション法は、スイッチの既存の機能を制御することにより実現でき、ホスト側の設定は不要である。従来のイーサネットではツリートポロジを採用するため性能面のスケーラビリティが不足していたが、提案した On/Off リンクアクティベーション法ではイーサネットにおいて任意のトポロジを可能とし、安定的にスイッチの MAC アドレステーブルの更新、管理を実現することができる。測定の結果、安価な商用スイッチにおいて、リンクの On/Off 操作、経路更新の総オーバヘッドは数秒であった。さらに、イーサネットのリンク(ポート)の消費電力がトラフィック量によらず一定である点に着目し、各商用スイッチの消費電力モデルを示した。そして、128 台のホストで構成される PC クラスタにおける NAS パラレルベンチマークを用いた評価結果より、On/Off アクティベーション法は性能を維持しつつ、スイッチの総消費電力を最大 37%削減できることがわかった[E-5、E-6、E-7]。

また、典型的な SIMD 型プロセッサなどのインターコネクトの通信性能を解析し、遅延、スループットに関するモデルを提案、検証した。さらにトラフィックパターンを事前に解析し、使用率の低いリンクを Off 状態にする軽量なルータ技術を示し、そのリンクを以後利用しないように経路を更新するルーティングのデッドロック除去技術、予測により通信遅延を抑える手法、電力増加を抑える超高信頼技術を提案、評価した[E-1、E-2、E-3、E-4]。

「主たる共同研究者⑤・日向寺」グループ（東海大学）

平成20年度、東海大グループでは、小規模データを用いた量子化学計算（フラグメント分子軌

道計算）による電力コストの測定を行った。測定実験は、16コア（2コア×8ノード）のPCクラスタ上で行い、量子化学計算ソフトABINIT－MP（RSS21版）を用いた。Gly3～10の小規模な入力データを構築し、1～16コアで計算した際の計算手法（MP2法とHartree－Fock法）の違いによる電力コストの差を測定した。

また、現実的な生体分子系により近付けるためGlyからなるポリペプチドの分子量を増やし、Gly8、Gly12、Gly16の安定構造を用いた電力測定も行った。その結果、フラグメント分子軌道計算の前半に行われるフラグメントモノマー計算では、各ノード間でほぼ均一に電力が消費されるのに対し、後半のフラグメントダイマー計算では各ノードで差異が見られる。

3. 研究実施体制

(1)「研究代表者・松岡」グループ
①研究分担グループ長：松岡　聡（東京工業大学　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)「主たる共同研究者①・須田」グループ
①研究分担グループ長：須田 礼仁（東京大学大学院　准教授）
②研究項目
超省電力 HPC システムに適したロバストな性能モデルや高性能と省電力の複合目的関数最適化などの数理の研究を行う。その成果は、自動チューニング数理基盤ライブラリおよび自動チューニングスクリプト言語 ABCLibScript の超省電力 HPC システム向けの拡張の形で実体化する。最終的には、ヘテロ複合アーキテクチャである超省電力 HPC システムに、柔軟かつロバストに適応するソフトウェアに必須である、自動チューニング基盤システムの完成を目指す。

(3)「主たる共同研究者②・青木」グループ
①研究分担グループ長：青木尊之（東京工業大学　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)「主たる共同研究者③・本多」グループ
①研究分担グループ長：本多　弘樹（電気通信大学大学院　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)「主たる共同研究者④・鯉渕」グループ
①研究分担グループ長：鯉渕　道紘（国立情報学研究所　助教）

②研究項目

省電力インターコネクトの研究開発

(6)「主たる共同研究者⑤・日向寺」グループ

①研究分担グループ長：合田（日向寺）　祥子（東海大学　講師）

②研究項目

巨大分子量子化学計算における超省電力 HPC システムの性能評価

4. 研究成果の発表等

(1) 論文発表（原著論文）

【松岡G】

1. A-1. Toshio Endo and Satoshi Matsuoka. Massive Supercomputing Coping with Heterogeneity of Modern Accelerators. In Proceedings of IEEE International Parallel & Distributed Processing Symposium (IPDPS 2008), 10pages, April 2008.
2. A-2. Yuto Hosogaya and Toshio Endo and Satoshi Matsuoka. Performance Evaluation of Parallel Applications on Next Generation Memory Architecture with Power-Aware Paging Method. In Proceedings of 4th IEEE Workshop on High-Performance, Power-Aware Computing (HPPAC08), in conjunction with IPDPS2008, 8pages, April 2008.
3. A-3. Yasuhiko Ogata, Toshio Endo, Naoya Maruyama, and Satoshi Matsuoka. "An Efficient, Model-Based CPU-GPU Heterogeneous FFT Library". In the 17th International Heterogeneity in Computing Workshop (HCW'08), in conjunction with IPDPS 2008, Miami, FL, USA, April 2008.
4. A-4. 額田彰,尾形泰彦,遠藤敏夫,松岡聡. CUDA 環境における高性能 3 次元 FFT. 先進的計算基盤システムシンポジウム SACSIS2008 論文集,pp. 81-88,2008 年 6 月.
5. A-5. 山崎翔平,丸山直也,松岡聡. モデルベース資源選択による効率的な仮想クラスタ構築 . 先進的計算基盤システムシンポジウム(SACSIS2008) 論文集, Vol. 2008, No.5 pp.325-332, 2008 年 6 月
6. A-6. 尾形 泰彦,遠藤 敏夫,丸山 直也,松岡 聡. 性能モデルに基づく CPU 及び GPU を併用する効率的な FFT ライブラリ. 情報処理学会論文誌コンピューティングシステム,Vol.1,No.1 (ACS 22), pp. 40-50,2008 年 6 月.
7. A-7. 額田彰,尾形泰彦,遠藤敏夫,松岡聡. CUDA 環境における高性能 3 次元 FFT. 情報処理学会論文誌コンピューティングシステム（ACS）,Vol. 1,No. 2,pp. 231-239,2008 年 8 月.
8. A-8. Akira Nukada, Yasuhiko Ogata, Toshio Endo and Satoshi Matsuoka. Bandwidth Intensive 3-D FFT kernel for GPUs using CUDA. In Proceedings of the ACM/IEEE conference on Supercomputing (SC'08), 11pages, Austin, November 2008.
9. A-9. Tomoaki Hamano, Toshio Endo, Satoshi Matsuoka. Power-Aware Dynamic Task Scheduling for Heterogeneous Accelerated Clusters. The Fifth Workshop on

High-Performance, Power-Aware Computing (HPPAC), in conjunction to IEEE IPDPS 2009, Rome, Italy, May 2009 (to appear).

10. A-10. A. Cevahir, A. Nukada and S. Matsuoka. Fast Conjugate Gradients with Multiple GPUs, Proc. The International Conference on Computational Science (ICCS 2009), 2009 (to appear).
11. A-11. 額田 彰,松岡 聡. CUDA GPU 向けの自動最適化 FFT ライブラリ. 先進的基盤システムシンポジウム SACSIS2009,採択済み,2009.
12. A-12. A. Cevahir, A. Nukada and S. Matsuoka. An Efficient Conjugate Gradient Solver on Double Precision Multi-GPU Systems. 先進的基盤システムシンポジウム SACSIS2009, 採択済み,2009.

【須田Ｇ】

13. B-1.D.-Q. Ren, R. Suda, "Modeling and Estimation for the Power Consumption of Matrix Computations on Multi-core CPU and GPU platform," Proceedings of IEEE IWHGA 2009, to appear.
14. B-2.R. Suda, "Divisible Load Scheduling with Improved Asymptotic Optimality", IEEE Cluster 2008, Poster + work-in-progress, 6 pages (CD-ROM), 2008.
15. B-3.D.-Q. Ren, D. D. Giannacopoulos, R. Suda, "An Optimized Dynamic Load Balancing Method for Parallel 3-D Mesh Refinement for Finite Element Electromagnetics with Tetrahedra", iWAPT2008 / IEEE Cluster 2008, 7 pages (CD-ROM).

【本多Ｇ】

16. D-1. Shoichi Hirasawa, Hiroki Honda, "Toward a Portable Programming Environment for Distributed High Performance Accelerators", In Proceedings of The First International Workshop on Software Technologies for Future Dependable Distributed Systems (STFSSD 2009), pp.189-194, Mar, 2009
17. D-2. 大島 聡史, 平澤 将一, 本多 弘樹: OMPCUDA: GPU 向け OpenMP の実装, HPCS2009 2009 年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, pp.131-138, Jan, 2009
18. D-3. Shoichi Hirasawa, Yu Nakanishi, Hiromasa Watanabe, Hiroki Honda: "Common Description Language of SIMD Instructions for Performance Portability", In Proceedings of The 2008 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'08), pp.52-58, Jul, 2008

【鯉渕Ｇ】

19. E-1. Michihiro Koibuchi, Hiroki Matsutani, Hideharu Amano, Timothy M. Pinkston, "Lightweight Fault-tolerant Mechanism for Network-on-chip", Proc. of the 2nd ACM/IEEE International Symposium on Networks-on-Chip (NOCS'08), pp.13-22, Apr 2008
20. E-2. 鯉渕 道紘, 吉永 努, 村上 弘和, 松谷 宏紀, 天野 英晴, "予測機構を持つルータを用いた低遅延チップ内ネットワークに関する研究", 先進的計算基盤システムシンポジウム SACSIS'08 論文集, pp.393--401, Jun 2008
21. E-3. 鯉渕 道紘,吉永 努,村上 弘和,松谷 宏紀,天野 英晴,``予測機構を持つルータを

用いた低遅延チップ内ネットワークに関する研究", 情報処理学会論文誌：コンピューティングシステム(ACS),vol.1 No.2, pp. 59-69, Aug 2008

22. E-4. Daihan Wang, Hiroki Matsutani, Michihiro Koibuchi, Hideharu Amano, ``A Link Removal Methodology for Network-on-Chip on Reconfigurable Systems", Proc. of the 18th International Conference on Field Programmable Logic and Applications (FPL'08), pp.269-274, Sep 2008
23. E-5. Takafumi Watanabe, Masahiro Nakao, Tomoyuki Hiroyasu, Tomohiro Otsuka, Michihiro Koibuchi ``The Impact of Topoloy and Link Aggregation on PC Cluster with Ethernet", (Work-in-progress presentation) IEEE International Conference on Cluster Computing (Cluster2008), pp.380-385, Sep. 2008
24. E-6. 大塚 智宏, 鯉渕 道紘, 工藤 知宏, 天野 英晴, ``VLAN イーサネットを用いた PC クラスタ向け大規模ネットワーク構築法", 情報処理学会論文誌：コンピューティングシステム(ACS), 情報処理学会論文誌：コンピューティングシステム(ACS), Vol.1 No.3, pp.96-107, Dec 2008
25. E-7. Michihiro Koibuchi, Tomohiro Otsuka, Hiroki Matsutani, Hideharu Amano, An On/Off Link Activation Method for Low-Power Ethernet in PC Clusters, Proc. of the 23rd IEEE International Parallel and Distributed Processing Symposium (IPDPS'09), May 2009. (to appear)

（2）特許出願

平成 20 年度 国内特許出願件数：0 件（CREST 研究期間累積件数：0 件）