「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」
平成 24 年度採択研究代表者

H28 年度
実績報告書

遠藤　敏夫

東京工業大学
准教授

ポストペタスケール時代のメモリ階層の深化に対応するソフトウェア技術

§1. 研究実施体制

(1)遠藤グループ(東京工業大学)
① 研究代表者:遠藤　敏夫 (国立大学法人東京工業大学学術国際情報センター、准教授)
② 研究項目
・メモリ階層対応ランタイムの研究開発とプログラミングモデル・アーキテクチャ統合
・メモリ階層対応ダイナミックコンパイレーション技術の研究開発

(2)緑川グループ(成蹊大学)
① 主たる共同研究者:緑川　博子 (成蹊大学　理工学部、助教)
② 研究項目
・大容量, 高性能を実現する多種多階層型メモリ構成技術と管理手法の研究

§2. 研究実施の概要

本 CREST チーム(通称「メモリ CREST」)の研究目的は、ポストペタ時代における科学技術計算の高性能化と大規模化の実現のために、今後ますます厳しくなるメモリウォール問題に対処するソフトウェア技術を確立することである。そのためにアプリケーションアルゴリズム・システムソフトウェア・メモリアーキテクチャの分野にまたがった研究開発を行っている。

特に本年度は、当チームのメモリ階層システムソフトウェアと高速 Flash デバイスの利用によるノードあたり数百 GB の問題規模の高性能計算の実現・性能評価や、コンパイルツールチェインの改良によるプログラマへのメモリ最適化支援の強化を中心に研究を推進した。また成果ソフトウェアの多くを、github サイトにて公開開始した。

メモリ階層対応ランタイムやソフトウェア・アーキテクチャ技術統合を担当する遠藤グループでは

遠藤らを中心に、システムソフトウェアについての下記の研究を推進した。メモリ階層活用ランタイムライブラリ HHRT において、多メモリ階層利用時の性能評価を詳細に行った。基本的に規模と速度の両立には成功しているものの、上位メモリ階層の 20 倍以上(100GB 以上)の問題規模において、解放できずにピンダウンされるメモリ領域の存在のために性能維持が阻害されることを確認し、解決手法の検討を行った。また HHRT モデルに基づいた PGAS ランタイムライブラリの開発および、メインメモリ遅延が(3D Xpoint などの採用により)現状より大幅に大きくなった場合の性能評価ツールの研究開発を行った。

また本グループでは佐藤らを中心に、メモリ階層の効率的利用・局所性の向上支援を目的とするコンパイラツールチェインおよびプロファイリング技術を中心とする研究開発に取り組んだ。コンパイラツールチェインに関しては、単一命令セット環境における実行時バイナリ変換に基づくコード変換機構の開発とその評価を平成 28 年度の前半に実施し、研究成果としての取りまとめを進めた。プロファイリング技術に関しては、フィードバック駆動型のチューニング機構へ応用することを念頭に実装を進め、キャッシュラインコンフリクトの原因となる部分をソースコードレベルで理解する仕組みとしてパッケージングを行った。本機構によりシステム性能向上に資する情報としてユーザが自身でソースコードを修正する際のヒントが提供でき、性能チューニングにおける生産性を大幅に改善することが可能となった。

多種多階層型メモリ構成技術として、主記憶サイズを超えるテラバイト級の大規模データを扱うステンシル計算を、最新 NVMe フラッシュデバイスを主記憶拡張として用い、DRAM のみを利用した場合（通常時）の性能（Mflops）の 80%～90%の性能で処理できるアルゴリズム（MultiMem-Stencil）を構築した。またこのアルゴリズムの性能を最も引き出す最適パラメタ（時間，空間ブロックサイズなど）を、用いる計算プラットフォームや問題サイズに合わせ、ステンシル計算実行時に Just-in-Time で抽出する自動チューニングシステム（Blk-Tune)を実現した。さらに、各ノードに SSD を備えたコンピュータクラスタ向けに、ノードに分散するメモリの総容量を超える大規模データを扱うためのステンシルアルゴリズムと、その最適パラメタ（ブロックサイズ）決定手法を提案し、実クラスタで効果を得た。遠隔メモリ利用技術として開発中の分散大容量メモリ（DLM）では、動的に生成・消滅するユーザスレッドに対応する遠隔ページフェッチ機能、ページ交換プロトコルの改良を行った。

以上のような研究開発(前年度までを含む)から得られた知見は新世代のスーパーコンピュータ技術の設計に活用される。遠藤がコアメンバーの一人として仕様策定に携わった東京工業大学 TSUBAME3.0 スパコンは、平成 29 年 1 月に開札が行われた。その結果、各計算ノードはローカルストレージとして、TB 級の容量と GB/s 級のアクセス速度を備える高速 Flash デバイスを搭載する予定であり、本チームの技術との統合により高性能計算からの利用可能なレベルとなる見込みである。

代表的論文

- ″Toshio Endo. ″Realizing Out-of-Core Stencil Computations using Multi-Tier Memory Hierarchy on GPGPU Clusters″. In Proceedings of IEEE Cluster Computing (CLUSTER2016), pp. 21-29, Taipei, Sep 2016. (DOI: DOI: 10.1109/CLUSTER.2016.61)
- Hiroko Midorikawa, Hideyuki Tan: ″Evaluation of Flash-based Out-of-core Stencil

Computation Algorithms for SSD-Equipped Clusters", The 22nd IEEE International Conference on Parallel and Distributed Systems ICPADS2016 , pp.1031-1040, 2016 (DOI: 10.1109/ICPADS.2016.013)