「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

H24 年度
実績報告

遠藤 敏夫

東京工業大学学術国際情報センター・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 (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

(2)佐藤グループ(北陸先端科学技術大学院大学)
① 主たる共同研究者:佐藤 幸紀 (北陸先端科学技術大学院大学情報社会基盤研究センター、助教)
② 研究項目
・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

(3)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. 研究実施内容

【概要】

初年度である平成 24 年度においては、各グループがシステムソフトウェアに関する研究・プロトタイプ開発等を推進した。それに並行して、各グループのソフトウェア成果、特に佐藤グループのメモリ局所性プロファイラをチーム内で共有し、議論・設計へのフィードバックを行った。本プロファイラは、本チームによるシステムソフトウェアが最適化のために活用するのに加え、アプリケーションの将来の HPC システム上での性能評価にも大いに有用であるため、「アプリケーション分野からみた将来の HPCI システムのあり方の調査研究」プロジェクトの理研-東工大チームとも連携し、メモリ部分の評価のために提供を行った。さらにアプリケーション分野については、よく知られたベンチマークプログラムに加え、CREST 藤澤克樹チーム・丸山直也チームらによる大規模アプリケーションを用いた評価を開始した。

【遠藤グループ】

遠藤グループにおいては、メモリ階層を有効利用するためのランタイムライブラリおよび局所性向上アルゴリズムの提案・評価を、TSUBAME2.0 スパコンおよび導入済の新世代 GPU および Flash 搭載サーバ上を用いて推進した。

提案しているランタイムライブラリである HHRT/MC(hybrid-hierarchical runtime with MPI/CUDA)のプロトタイプ設計および実装を推進した。本ライブラリでは物理 GPU 数よりも多数の仮想プロセスを起動し、その間でメッセージ通信が起こる。同一 GPU 上の仮想プロセス間、別ノード上の仮想プロセス間といった距離の差異が生まれるため、それぞれのケースにおいて効率的な通信機構のプロトタイプを開発した。7 点ステンシルなどのカーネル計算を用いランタイムの動作検証を行った。

局所性向上アルゴリズムの研究として、ステンシル演算における大規模問題サイズと高性能の両立の実現に取り組んだ[A-1]。GPU の高計算性能と広大なホストメモリの容量の双方を活用するのはステンシル計算では一般的には困難である。そのために、格子データの一部ずつについて、複数時間ステップ分の演算を進める手法である時間ブロッキングを導入した。この手法には冗長な計算やデータ移動の発生という課題があり、この緩和のための技法を提案した。問題サイズがデバイスメモリサイズを超える場合に、単純な手法では性能が 5%程度に下がってしまってしまうところを、70%以上の性能保持を達成し、大規模問題サイズと高性能の両立を実現した。今後は多数 GPU 版の実現・CREST 丸山直也チームのアプリケーションにおける実現・記述性の向上のための上記 HHRT ランタイム上の実装等に取り組む。

さらに CREST 藤澤克樹チームと連携し、半正定値計画問題ソルバーの TSUBAME2.0 上の大規模実行およびスケーラビリティ向上を行った。すでに約 4000GPU 上で 533TFlops の世界記録を達成し、その内容を Supercomputing'12 などで発表した[A-2]が、これには我々の GPU の高計算性能とホストメモリの大容量の両方を活用する技術が用いられている。これに加え、大規模

ツリーネットワークのボトルネック解析や通信と計算のオーバラップ部分のさらなる改良により 5～10%の性能向上を実現した。

【佐藤グループ】

佐藤グループでは、メモリ階層対応ダイナミックコンパイレーション技術を研究開発し、既存のキャッシュメモリやコンパイラでは吸収しきれないメモリ階層間の特性を考慮したチューニングを自動／半自動にて行うことを目指す。平成 24 年度は、メモリ局所性プロファイラの要素機能の研究開発、および、メモリ階層におけるメモリ性能シミュレータのプロトタイプの実装に取り組んだ。メモリ局所性プロファイラに関しては、実行バイナリコードからメモリ参照に関する情報を抽出し、アプリケーションのメモリ局所性情報として提供する上での主要な基本的機能の実装を行い、逐次実行ベンチマークプログラムを用いて評価を行った[B-1]。また、メモリ局所性プロファイラにより生成されるアプリケーションのメモリ局所性情報を遠藤グループにおける通信ランタイムの最適化や実アプリケーションプログラムにおけるメモリ参照特性を得る手段として、および、緑川グループによるノード内外のメモリ管理やデータ移動の最適化に活用するために必要な要件を検討し、チーム内の課題との連携を行うための基礎を固めた。検討を行った要件は次年度にて実装する予定である。

メモリ性能シミュレータに関しては、実行バイナリコードを入力として実行時に L1 キャッシュおよび L2 キャッシュの挙動をシミュレーション可能な実行駆動型キャッシュシュミレータのプロトタイプの実装を行った。また、既存のアクセラレータにおけるメモリ階層の詳細構造や特性に関する調査を実施し、調査の一環としてアクセラレータにおけるメモリ階層構造をアプリケーション特性に応じて再構成可能デバイスを用いてカスタマイズするという自由度が与えられた場合の性能を評価した[B-2]。次年度においては、より複雑なメモリ階層や各種メモリデバイスに対応するシミュレータとして拡張し、メモリ性能を見積った結果を統計的にモデル化しメモリモデルとして確立する予定である。

【緑川グループ】

緑川グループでは、平成24年度に購入したInfiniBand（FDR）クラスタを用い、遠隔メモリアクセス性能評価実験を行った[C-1]。マルチスレッドプログラムに対応する遠隔ページスワップ方式を構築し、世代の異なるクラスタ（Myri10GとInfiniBand（QDR/IPoIB）など）とも比較した。結果、Myri10Gに比べInfiniBand利用時の遠隔メモリアクセス性能がむしろ低下することがわかった。InfiniBand 向けフルマルチスレッド対応 MPI(MVAPICH2)の実際のラウンドトリップ通信性能を調査したところ、公表性能に比べ、平均で 3 倍程度の性能低下があり、毎回の転送時間にも 4 倍以上の大きなばらつきがあることがわかった。さらに詳細な調査が必要であるが、現状、フルマルチスレッド対応で高速通信媒体を利用できる MPI が限られているため、最適化手法や他の代替ネットワーク（40GbpsEthernet など）の比較性能実験も考えている。

この他に、繰り返し処理を含む応用を対象に遠隔メモリアクセス時の遠隔メモリサーバとのスラッシングを抑制するため、ループ内におけるアクセスデータセットサイズ（WS）を推定し、自動的に

ページサイズを調整する機構を構築し、初期評価を行った。また，バス接続型 Flash メモリに関する公開技術や、現在、世界的に活発化している不揮発性メモリへの各レベルでのアクセス API の標準化動向などを調査し、今後、どのレベルでの実装を行うかについても検討している。さらに、コンパイラやメモリアクセストレース情報をもとにデータ初期配置や事前移動などを行うため、佐藤グループとも議論し情報を共有した。

§3. 成果発表等

（3－1） 原著論文発表

●論文詳細情報

[A-1] Guanghao Jin, Toshio Endo, Satoshi Matsuoka. A Multi-level Optimization Method for Stencil Computation on the Domain that is Bigger than Memory Capacity of GPU . In Proceedings of The Third International Workshop on Accelerators and Hybrid Exascale Systems (AsHES), in conjunction with IEEE IPDPS 2013, Boston, May 2013 (accepted).

[A-2] Katsuki Fujisawa, Toshio Endo, Hitoshi Sato, Makoto Yamashita, Satoshi Matsuoka, Maho Nakata. High-Performance General Solver for Extremely Large-scale Semidefinite Programming Problems. In Proceedings of IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC12), Saltlake City, November 2012 (DOI: 10.1109/SC.2012.67).

[B-1] Yukinori Sato, Yasushi Inoguchi and Tadao Nakamura. Whole Program Data Dependence Profiling to Unveil Parallel Regions in the Dynamic Execution. In Proceedings of 2012 IEEE International Symposium on Workload Characterization (IISWC 2012). La Jolla, Nov. 2012. (DOI: 10.1109/IISWC.2012.6402902).

[B-2] Yukinori Sato, Yasushi Inoguchi and Tadao Nakamura. Evaluating Reconfigurable Dataflow Computing Using the Himeno Benchmark. In Proceedings of 2012 International Conference on ReConFigurable Computing and FPGAs (ReConFig2012). Cancun, Dec. 2012. (DOI: 10.1109/ReConFig.2012.6416746).

[C-1] Hiroko Midorikawa, Yuichiro Suzuki, Masatoshi Iwaida. User-level Remote Memory Paging for Multithreaded Applications, Procedding of 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2013), Delft, Netherlands, May 2013 (accepted)