「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22年度採択研究代表者

H24 年度
実績報告

丸山　直也

独立行政法人理化学研究所計算科学研究機構・チームリーダー

高性能・高生産性アプリケーションフレームワークによるポストペタスケール高性能計算の実現

§1. 研究実施体制

(1)「丸山」グループ
① 研究代表者：丸山　直也　（理化学研究所計算科学研究機構、チームリーダー）
② 研究項目
・高い生産性と性能を両立する格子系流体向けアプリケーションフレームワーク

(2)「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
・格子系流体アプリケーションの大規模スーパーコンピュータにおける人手による参照実装

(3)「田浦」グループ
① 主たる共同研究者：田浦　健次朗　（東京大学大学院情報理工学研究科電子情報学専攻、准教授）
② 研究項目
・大域アドレス空間モデルと軽量マルチスレッドによるスケーラブルランタイム

(4)「泰岡」グループ
① 主たる共同研究者：泰岡　顕治　（慶應義塾大学理工学部機械工学科、教授）
② 研究項目
・分子動力学アプリケーションの大規模スーパーコンピュータにおける人手による参照実装

(5)「丸山」グループ
① 主たる共同研究者：丸山　直也　（東京工業大学学術国際情報センター、客員准教授）
② 研究項目
・高い生産性と性能を両立する格子系流体向けアプリケーションフレームワーク

§2. 研究実施内容

2.1 研究のねらい

本研究チームのねらいはアプリケーションドメイン特化型のソフトウェアスタックを構成し、それによって高い生産性と性能の両立をポストペタスケールスーパーコンピュータで達成することである。具体的には、アプリケーションドメインとして格子系流体シミュレーションおよび分子動力学法を対象とし、それぞれについてアプリケーションフレームワークをランタイムおよびプログラミング技術を応用することで構成する。これを達成するために本研究課題は以下の5つの研究項目から構成される。

【項目1】格子系流体アプリケーションの大規模スーパーコンピュータにおける人手による参照実装
【項目2】分子動力学アプリケーションの大規模スーパーコンピュータにおける人手による参照実装
【項目3】高い生産性と性能を両立する格子系流体向けアプリケーションフレームワーク
【項目4】大域アドレス空間モデルと軽量マルチスレッドによるスケーラブルランタイム
【項目5】高い生産性と性能を両立する分子動力学法向けアプリケーションフレームワーク

平成24年度はこれらの項目の内、項目1から項目4について以下をマイルストーンとして研究を実施した。

【項目1】格子系流体アプリケーションの大規模スーパーコンピュータにおける人手による参照実装（青木グループ）

- AMR の導入を前提とした tree 構造に対する Space Filling Curve による制御
- AMR を導入した構造格子での圧縮性流体計算の人手による GPU 実装
- 非構造格子流体計算における GPU 計算環境における初期検討

【項目2】分子動力学アプリケーションの大規模スーパーコンピュータにおける人手による参照実装（泰岡グループ）

- 負荷分散、通信と計算のオーバーラップ等について自動化アルゴリズムの開発

【項目3】高い生産性と性能を両立する格子系流体向けアプリケーションフレームワーク（丸山グループ）

- Physis フレームワークの拡張、最適化、耐故障性、自動チューニングによる性能と電力の最適化
- 均一構造格子差分法向けフレームワークのプロトタイプによる実アプリケーションの実装と評価
- 非構造格子計算向けフレームワークの初期設計

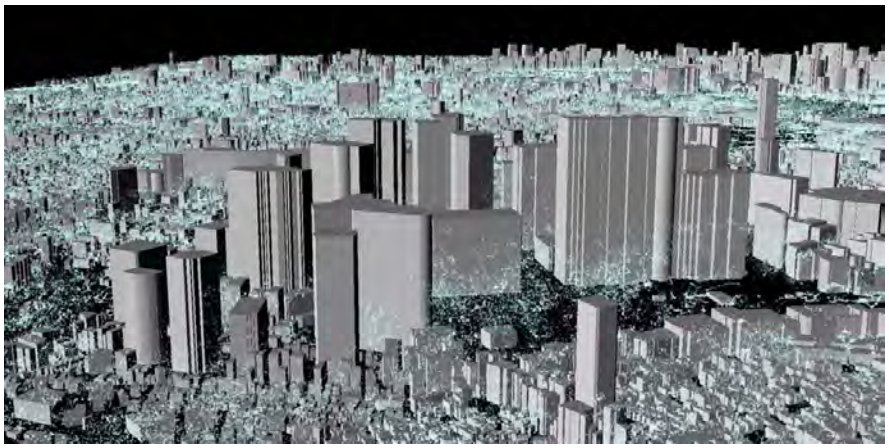
【項目4】大域アドレス空間モデルと軽量マルチスレッドによるスケーラブルランタイム（田浦グループ）

- InfiniBand 上で MPI, GasNet と比肩する性能を持つ、タスク並列をサポートする大域アドレス空間ライブラリの設計・実装を行い基本性能を評価する
- 分散メモリ上で負荷分散を行うマルチスレッドライブラリを設計・実装し、上記大域アドレス空間ライブラリと組み合わせて評価する。基本的なベンチマークのほか、密行列積による限界性能評価、FMM による記述性・性能トレードオフの評価を行う。
- MassiveThreads の Chapel 次期バージョンへの統合

2．2 これまでの研究の概要、進捗状況、および今後の見通し

【項目1】格子系流体アプリケーションの大規模スーパーコンピュータにおける人手による参照実装

2012 年度は、格子系流体アプリケーションの実アプリケーションとして格子ボルツマン法による都市気流の大規模計算を実行した。メモリ・アクセスが律速であるが、TSUBAME2.0 の 4000GPU を使って 0.6PFLOPS（単精度）というピーク性能に対して 15% の高い実行性能を達成し [13]、HPCS2013 では、最優秀論文賞を受賞した。単一 GPU に対して構造格子アプリケーションである格子ボルツマン法およびフェーズフィールド法への AMR 参照実装を行った [14]。また、木構造データのリーフに対して 2 種類の Space Filling Curve による制御を検討した。また、圧縮性流体計算の人手による Kepler GPU への実装では、Fermi コアと比較してコア当たりの L1 キャッシュ量が削減されたためヒット率が低下していることが分かり、チューニング指針を変更する必要があることが分かった。



今後は AMR 法による格子系アプリケーションを人手により複数 GPU 計算に対応させ、Space Filling Curve に基づいたノード間の動的負荷分散等を行う。

【項目2】分子動力学アプリケーションの大規模スーパーコンピュータにおける人手による参照実装

今年度までに数千の GPU を使って FMM(Fast Multipole Method)の計算が行えるコードを開発してきた [1]。分子動力学シミュレーションを行う場合、粒子の配置によって各ノードの計算負荷が変化する。特に FMM によって無限遠までのクーロン力を考慮すると、その負荷は複雑になるため動的負荷分散が必要となる。そこで今年度は、まずノード内動的負荷分散機能を実装した。具

体的にはQUARKやMassiveThreads、Intel TBBを使って実装しMassiveThreadsが性能的に優れていることが分かった。また、ノード間の動的負荷分散機能として、複数GPUを使った分子動力学シミュレーションのテストコードを作成してDS-CUDAにより動作を検証した。4GPUを使った場合に10%程度の性能向上が確認された。DS-CUDA[5),9)]はMPIを使わずに複数GPUを使うコードが書けるミドルウェアである。通信と計算のオーバーラップ機能に関しては、これまでの領域内部計算と袖領域通信のオーバーラップだけでなく、階層間のオーバーラップも初期実装した。今後は様々な機能をうまく共通化したモジュール等へと切り分けることが必要になってくる。

【項目3】高い生産性と性能を両立する格子系流体向けアプリケーションフレームワーク

均一構造格子差分法向けアプリケーションフレームワークPhysisの実装およびその性能評価を進めた。昨年度までに基本的なプロトタイプは完成しており、今年度は性能改善に注力した。単一GPUを対象としたコード生成では適切な最適化の組み合わせにより人出によるチューニングが施されたものと同等の性能を達成できている。また、単一GPU実行を対象とした自動チューニング機構を試作した。これは各種最適化パスの適用やそのパラメータについて網羅的に最適化組み合わせを探索するものであり、最適化の自動化を狙ったものである。来年度は同試作をベースに他実行環境への拡張等を実施する。また耐故障性の実現については要素技術の開発、評価について一定の成果を得ており、今後はフレームワークへの統合を進め[6),7),8)]、アプリケーション適用を通じた評価を行う。

また、高生産性と高性能の両立のための手法として関連技術である指示文によるアクセラレータプログラミングに関する評価も行い、フレームワークによるアプローチと比較したトレードオフを明らかにした[15)]。

【項目4】大域アドレス空間モデルと軽量マルチスレッドによるスケーラブルランタイム

[大域アドレス空間ライブラリ MassiveThreads/DM]

明示的なページのキャッシュ、マイグレーション，離散的な一括アクセスをサポートする大域アドレス空間ライブラリの実装を行い、TSUBAME 2.0(Infiniband ネットワーク)上で評価を行なった[33)]。C 言語レベルの軽量マルチスレッドをノード間で移送させる機能の実装も完了した。MassiveThreads に、応用に特化した負荷分散(ワークスチーリング)方法を実現するための API を設計、実装した[31)]。 また競合検知器および性能プロファイラを試作[32]するとともに、プロファイラの大量データを解析する基盤を設計・実装した[3), 11), 12)]。

[MassiveThreadsの他プラットフォームへの移植、他の言語処理系への組み込み]

MassiveThreads処理系を、京/FX10のCPUであるSPARC 64 fx[45)]およびXeon Phi上へ移植した。これによって将来的にMassiveThreads/DM処理系が京やFX10上で動作するための準備が整った。また，MassiveThreadsとGASNetを用いている言語処理系であるChapel処理系のFX10上への移植も完了し、初期的な評価を行った[4), 45)]。また,計画書で予定として述べたとおり、MassiveThreadsはChapel 1.6より、正式なリリースの一部として用いられることになった。

[MassiveThreads の実アプリケーションによる評価]

実アプリケーションとして、Fast Multipole Method の高速な実装である ExaFMM の分割統治法を用いた並列化を MassiveThreads を用いて行い、評価した [10),35)]。FMM は、直接計算の負荷が比較的小さい低精度な領域で台数効果を得ることが難しいが、本実装ではすべてのフェーズにおいて高い台数効果とタスク並列による記述性が検証できた。100 万粒子、3 桁精度の計算で、 SandyBridge 16 コアのノードで1ステップ 65ms の速度を達成した。

§3. 成果発表等

(3－1) 原著論文発表

●論文詳細情報

1. Rio Yokota, L.A. Barba, Tetsu Narumi, Kenji Yasuoka, "Petascale Turbulence Simulation Using a Highly Parallel Fast Multipole Method", Computer Physics Communications, Vol. 184, pp. 445-455, 2013, (DOI: 10.1016/j.cpc.2012.09.011)
2. Akihiro Nomura, Yutaka Ishikawa, Naoya Maruyama, Satoshi Matsuoka, "Design and Implementation of Portable and Efficient Non-blocking collective Communication", The 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2012), 2012, (DOI: 10.1109/CCGrid.2012.96)
3. 堀内美希、田浦健次朗, "広域分散ファイルシステムのための適応的な先読み手法", 先進的計算基盤システムシンポジウム(SACSIS)2012, 2012
4. Nan Dun, Kenjiro Taura, "An Empirical Performance Study of Chapel Programming Language", 26th IEEE International Parallel & Distributed Processing Symposium, 2012, (DOI: 10.1109/IPDPSW.2012.64)
5. Atsushi Kawai, Kenji Yasuoka, Kazuyuki Yoshikawa, Tetsu Narumi, "Distributed-Shared CUDA: Virtualization of Large-Scale GPU Systems for Programmability and Reliability", The Fourth International Conference on Future Computational Technologies and Applications, FUTURE CONPUTING 2012, 2012
6. Leonardo Bautista Gomez, Bogdan Nicolae, Naoya Maruyama, Franck Cappello, SATOSHI MATSUOKA, "Scalable Reed-Solomon-based Reliable Local Storage for HPC Applications in IaaS Clouds", International European Conference on Parallel and Distributed Computing, 2012, (DOI: 10.1007/978-3-642-32820-6_32)
7. Leonardo Bautista Gomez, Thomas Ropars, Franck Cappello, Naoya Maruyama, SATOSHI MATSUOKA, "Hierarchical Clustering Strategies for Fault Tolerance in Large Scale HPC Systems", 2012 IEEE International Conference on Cluster Computing, 2012(DOI: 10.1109/CLUSTER.2012.71)

8. Kento Sato, Naoya Maruyama, Kathryn Mohror, Adam Moody, Todd Gamblin, Bronis R. de Supinski, Satoshi Matsuoka, "Design and modeling of a non-blocking checkpointing system", Proceedings of the 2012 ACM/IEEE conference on Supercomputing (SC'12), 2012
9. Minoru Oikawa, Atsushi Kawai, Kentaro Nomura, Kazuyuki Yoshikawa, Kenji Yasuoka, Tetsu Narumi, "DS-CUDA: a Middleware to Use Many GPUs in the Cloud Environment", International Workshop on Sustainable HPC Cloud at SC12, 2012
10. Kenjiro Taura, Jun Nakashima, Rio Yokota, Naoya Maruyama, “A Task Parallelism Meets Fast Multipole Methods”, Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems at SC12, 2012
11. Ting Chen, Kenjiro Taura, "A Comparative Study of Data Processing Approaches for Text Processing Workflows", WORKS2012 at SC12, 2012
12. Miki Horiuchi, Kenjiro Taura, "Acceleration of Data-intensive Workflow Applications by Using File Access History", WORKS2012 at SC12, 2012
13. 小野寺直幸, 青木尊之, 下川辺隆史,小林宏充, "格子ボルツマン法による 1m 格子を用いた都市部 10km 四方の大規模 LES 気流シミュレーション", 情報処理学会ハイパフォーマンスコンピューティング研究会主催 HPCS シンポジウム 2013, 2013, (http://hpcs.hpcc.jp/program.html#session5)
14. 下川辺隆史、青木尊之、小野寺直幸, "ブロック AMR 法の GPU コンピューティング・フレームワーク", 情報処理学会ハイパフォーマンスコンピューティング研究会主催 HPCS シンポジウム 2013, 2013, (http://hpcs.hpcc.jp/)
15. Tetsuya Hoshino, Naoya Maruyama, Satoshi Matsuoka, Ryoji Takaki, "CUDA vs OpenACC: Performance Case Studies with Kernel Benchmarks and a Memory Bound CFD Application", Proceedings of the 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2013,(Accepted)