2024 年度　利用報告書

課題名（タイトル）:

Research and Development of system software for high performance big data applications

利用者氏名:

○佐藤 賢斗(1), Amarjit Singh(1), 吉田 幸平(1), 浅井 陽介(1), Xin Huang(1), Shiman Meng(1), Weiping Zhang(1), Guoyuan Jia(1), Hui Li(1)

理研における所属研究室名:

(1) 計算科学研究センター 高性能ビッグデータ研究チーム

1.　本課題の研究の背景、目的、関係する課題との関係

当研究室は、『機械学習、深層学習および大規模ビッグデータ処理(AI 技術)』の高速化・スケール化のためのシステムソフトウェアの研究開発（HPC for AI）、さらにそれらの AI 技術を用いた高性能科学技術計算や高性能計算機の高速化・スケール化（AI for HPC）の研究・開発を行っています。これらの目的を達成するために、具体的に次のような研究開発を行っています。(1)大規模並列 I/O などのビッグデータのスケール化・高速化、(2)チェックポイント等の高信頼化技術のスケール化・高速化、(3)メモリ・ストレージ階層の深化に対応する超並列アルゴリズムやプログラミング、(4)マルチペタバイトデータのテラビット級ネットワークにおける高速転送、(5)ビッグデータ、機械学習、HPC のソフトウェアスタックの統合、及びそのスケール化・高速化、(6)超大規模ビッグデータの視覚化や対話型操作。

2.　具体的な利用内容、計算方法

今年度は先に挙げた(2)及び(3)に関連する、「マルチスレッドプログラムの効率的な記録と再生のための分散順序記録技術」の研究において HOKUSAI を利用しました。

大規模並列計算においては OpenMP などの API を利用したプログラミングおよび実行が行われますが、実行結果の検証や、デバッグを行うときに、一度実行したプログラムの動作を適切に再現することは容易ではありません。プログラムはマルチプロセス、マルチスレッドで動作し、共有メモリやその他のリソースにアクセスしますが、そのアクセス順序は、実行するタイミングや他のプログラムの動作の影響を受けるので、その時々で順序が入れ替わってしまうからです。

適切な順序で再実行するには、最初の実行時に適切な記録を行う必要がありますが、記録の際のオーバヘッドが大きくならないことが重要です。

これについて、分散クロック(DC)と分散エポック(DE)を記録する技術を新たに提案しました。DC は、スレッド ID の代わりに論理クロックを記録します。DC をベースに、再生効率を向上させる手法をさらに最適化したものが DE です。DE はプログラムの詳細な分析と特定の並列条件を利用することにより、シリアル化と同期の要件を効果的に軽減します。そしてプログラムの最終結果に影響を与えることなく、リソースへのアクセス順序を適切に再現できます。

この技術の有効性と実用性を検証するため、OpenMP 記録再生ツール ReOMP を実装しました。ReOMP は既存のマルチスレッド記録再生ツールとの互換性が高く、独立して実行することも、既存の記録再生ツールと補完的に実行することもできます。

性能検証として、HOKUSAI BigWaterfall2 上で並列計算プログラムを実行し、他の手法との比較を行いました。

3.　結果

マルチスレッドを用いる 5 つのアプリケーションを評価しました。私たちの手法は、従来の代表的なアプローチである ST recording 法よりも、2～5 倍効率的なことが確認できました。

4.　まとめ

「マルチスレッドプログラムの効率的な記録と再生のための分散順序記録技術」の研究において HOKUSAI を利用しました。分散クロック(DC)と分散エポック(DE)を記録する技術を提案し、HOKUSAI BigWaterfall2 上で動作検証を行って、他の手法よりも優れていることを確認しました。

5.　今後の計画・展望

並列タスクスケジューリングなど、より非決定的な OpenMP スケジューリングを調査します。 また、分散型記録再生技術を使用した先進的なチェックポイント／リスタート法の可能性を探ります。

2024 年度 利用研究成果リスト

【会議の予稿集】

Xiang Fu, Shiman Meng, Luanzheng Guo, Kento Sato, Dong H. Ahn, Ignacio Laguna, Gregory L. Lee, Martin Schulz, "Distributed Order Recording Techniques for Efficient Record-and-Replay of Multi - Threaded Programs", 2024 IEEE International Conference on Cluster Computing (CLUSTER), pp.27-38