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 23 年度採択研究代表者

H25 年度
実績報告

南里 豪志

九州大学情報基盤研究開発センター
准教授

省メモリ技術と動的最適化技術によるスケーラブル通信ライブラリの開発

§1. 研究実施体制

(1)「インタフェース」グループ
① 研究代表者:南里 豪志 (九州大学情報基盤研究開発センター、准教授)
② 研究項目
・隣接通信インタフェースの実装
・非ブロッキング集団通信インタフェースの実装
・隣接・集団通信の動的最適化技術の開発
・スケーラブルな通信ライブラリの実装と公開

(2)「プロトコル」グループ
① 主たる共同研究者:住元 真司 (富士通株式会社次世代 TC 開発本部、シニアアーキテクト)
② 研究項目
・通信バッファを削減した通信モデルにもとづいた通信プロトコル

(3)「通信路制御」グループ
① 主たる共同研究者:柴村 英智 ((財)九州先端科学技術研究所次世代スーパーコンピュータ開発支援室、研究員)
② 研究項目
・パケット送信間隔動的最適化技術
・Exa FLOPS 環境のアプリケーション性能予測技術

(4)「アプリケーション」グループ

① 主たる共同研究者:高見　利也　（九州大学情報基盤研究開発センター、准教授）
② 研究項目
・非ブロッキング集団通信と遠隔 Atomic 通信を活用した OpenFMO の開発と評価
・隣接通信を活用した電磁流体プログラムの開発と評価
・既存アプリケーションの隣接通信、非ブロッキング集団通信による改良
・ExaFLOPS 環境に向けた高スケーラブルなアプリケーション作成技術の確立

§2. 研究実施の概要

本年度は、昨年度までの成果を踏まえ、本プロジェクトで開発する通信ライブラリ ACP (Advanced Communication Primitives) の構造の策定、ACP 基本層の設計・実装、通信効率化に向けた通信インタフェースのプロトタイプ実装、通信路制御技術の実機における検証、およびアプリケーションにおける通信効率化技術の検証を行った。

まず ACP については、片側通信のアトミック操作を用いる通信の有効性を確認する事を主目的に研究を実施し、省メモリ・低遅延通信プロトコルの確立を目指した。まず ACP スタック構造を検討し、基本的な通信を担う ACP 基本層、ストリーム転送の使用メモリを最適化できるチャネルインターフェース、大域的なデータ配置のアクセスを最適化できるグローバルデータ構造コレクションから構成した。さらに、ACP 基本層について、インターコネクトデバイス抽象化層の機能を定義することを目標に、グローバルメモリ管理機能、グローバルメモリ参照・Atomic 操作機能の詳細検討を行い、関数仕様を策定した。この仕様を UDP スタックと Tofu インターコネクト上に実装し、通信デバイスの機能や性質に依存せずに、片側通信による低遅延なアトミック操作が実現できることを確認した。

通信効率化に向けた通信インタフェースについては、隣接通信、集団通信のプロトタイプ実装を行った。隣接通信については、計算ノードへのプロセスの配置状況と使用可能なネットワークインタフェース数を考慮して通信順序を調整するアルゴリズムを開発した。一方、集団通信については、実行時の状況に応じた性能予測によるアルゴリズムの絞り込みと、実行中の実測を併用して適切なアルゴリズムを選択する技術を実装し、効果を確認した。また、通信と計算を並行して行うことによって通信時間を隠蔽するための非ブロッキング集団通信インタフェースについて、既存の手法と本プロジェクトで提案する手法で、それぞれ通信隠蔽の効果を検証した。さらに、プログラムの付加的な情報を通信ライブラリに提供することにより高度な通信効率化を可能とするヒントインタフェースを提案し、効果を確認した。

通信路制御技術については、実機におけるパケットペーシングの有効性を実証することを目的とし、既存の HPC システムによる検証実験を行った。パケットの送出間隔を制御できる富士通社製 PRIMEHPC FX10 を利用し、ランダムリング通信と全対全通信にパケットペーシングを適用した場合の通信性能を調査した。その結果、これまでのシミュレーション評価で認められてきたパケットペーシングの有効性をはじめ、メッセージ長やノード数に応じたペーシング効果の向上を確認した。

エクサスケール級システムにおけるアプリケーションの実行性能を詳細に評価するためには、ノード演算性能の推定に加え、通信衝突によって発生する通信レイテンシを含めた通信時間の推定が重要となる。そこで、衝突も含めたエクサスケール級の通信を模擬し、システムの仕様や通信パターンに則した実行時間を算出するインターコネクトシミュレータ NSIM を核としたエクサスケール級アプリケーションの性能推定環境の整備を行った。

アプリケーションについては、片側通信ライブラリ、あるいは、非ブロッキング通信を利用してアプリケーションを実装し、可能な限り大規模なシステムで性能測定を実施した。実際に大規模並列動作が可能なプログラムを利用して性能評価をするため、本プロジェクトでは、OpenFMO プログラム、および、電磁流体プログラムを開発している。OpenFMO プログラムでは、内部で利用する動的負荷分散機構を、ARMCI や OpenSHMEM など、既存の片側通信ライブラリや遠隔 Atomic 通信を利用した実装を行った。電磁流体プログラムに関しては、京、および、FX10 での大規模並列実行の結果を解析し、性能、および、スケーラビリティに関連する問題点の特定を実施した。これら以外のアプリケーションプログラムについても解析を実施し、パイプライン型のプログラムの通信部分(バケツリレー通信)では、データの細分化による性能向上を確認した。

§3. 成果発表等

(3－1) 原著論文発表

論文詳細情報(国内)

A-1 森江 善之, 南里 豪志, “多次元メッシュ/トーラスにおける通信衝突を考慮したタスク配置最適化技術”, 情報処理学会論文誌コンピューティングシステム, Vol. 6, No. 3, pp. 12-21, Sep. 2013.

論文詳細情報(国際)

D-1 K. Fukazawa, T. Nanri, and T. Umeda, “Performance evaluation of magnetohydrodynamics simulation for magnetosphere on K computer”, In: AsiaSim 2013, Communications in Computer and Information Science, Vol. 402, edited by G. Tan, G. K. Yeo, S. J. Turner, and Y. M. Teo, pp. 570-576, Springer-Verlag Berlin Heidelberg, 2013. (ISBN: 978-3-642-45036-5) (DOI: 10.1007/978-3-642-45037-2_61)

D-2 K. Fukazawa, T. Nanri and T. Umeda, “Performance Measurements of MHD Simulation for Planetary Magnetosphere on Peta-Scale Computer FX10”, Parallel Computing: Accelerating Computational Science and Engineering (CSE), Advances in Parallel Computing 25, pp.387-394, IOS Press, 2014. (DOI: 10.3233/978-1-61499-381-0-387)

D-3 T. Takami and D. Fukudome, “An efficient pipelined implementation of space-time parallel applications”, Parallel Computing: Accelerating Computational Science

and Engineering (CSE), Advances in Parallel Computing 25, pp. 273-281, IOS Press, 2014. (DOI: 10.3233/978-1-61499-381-0-273)