「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

H26 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 (財団法人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、本プロジェクトで開発する通信ライブラリ ACP(Advanced Communication Primitives)について、昨年度設計・実装した基本層の上に構築する ACP Middle Layer の各インタフェースの設計と実装を行った。さらに、これらをパッケージとして整備し、プロジェクトの Web サイト (http://ace-project.kyushu-u.ac.jp/index.html)上で公開を開始した(図 1)。また、ACP のベースとなっている RDMA(Remote Direct Memory Access)通信に対応したインターコネクトシミュレータ NSIM-ACE を実装した。

図 1　ACE プロジェクト Web サイト

公開した ACP ライブラリの構成を図 2 に示す。基本層は、ネットワークの違いを隠蔽する低レベルインタフェースであり、RDMA 通信をベースとした設計により、省メモリと低オーバヘッドを両立させている。一方、Middle Layer では、Message Passing や分散データ構造等の高機能インタフェースを提供する。これらのインタフェースは、必要に応じてメモリの確保と解放を行うため、メモリ消費を最小限に抑えつつ、効率的な並列プログラミングを可能としている。

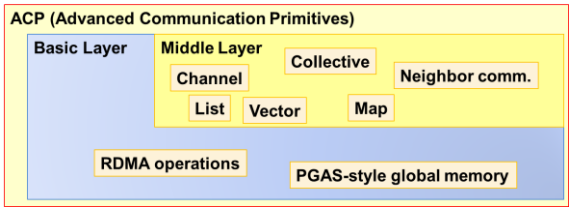

図 2　ACP ライブラリの構成

このうち分散データ構造に関しては、「PGAS 言語向け分散共有配列ライブラリの低遅延・省メモリ化技術の研究開発」を目標に、ACP 基本層上に非同期グローバルメモリアロケータおよび分散データ構造を持つデータライブラリを初期実装した。非同期グローバルメモリアロケータは、メモリを提供する側のプロセッサの介在なしにメモリ領域の確保を行うことができる。分散データ構造関数は C++の STL コンテナの vector と list に相当するデータ構造を実装した。データライブラリの初期実装は ACP 基本層の UDP 版および Tofu 版上で性能評価済みであり、Tofu インターコネクトのような低レイテンシな通信デバイス上ではリモート側データへのアクセスレイテンシがローカル側レイテンシの 2 倍から 4 倍に収まることが確かめられた。また、ACP 基本層を利用した、Ruby や

Python などのスクリプト言語ベースの PGAS 言語を試験実装した。Ruby や Python の高い生産性により、ACP の特徴であるリモート間コピーを損なうことなく PGAS 機能を容易に提供可能であることを確認した。

通信効率化に向けた通信インタフェースについては、アプリケーショングループが設計したチャネルインタフェースについて、ACP 基本層上にプロトタイプを実装した。このプロトタイプにより、チャネルインタフェースのように通信路の確立と解放を明示することで、通信に使用するメモリ量を最小限に抑制できることを確認した。一方、性能面では、特に Rendezvous プロトコルによる通信時のメモリ登録に伴うコストが問題となることが判明したため、現在改良を進めている。また、通信最適化技術としては、昨年度開発した隣接通信最適化技術の改良、および通信路の確立と解放を自動的に行う技術の研究開発を進めている。

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、昨年度に引き続き、既存の HPC システムにおける評価実験を行った。富士通社製 PRIMEHPC FX10 を利用し、ランダムリング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。昨年度よりも大規模なノード数（768 ノード）による評価実験を行い、実機におけるパケットペーシングの効果を実証するとともに、ノード数が増加した場合におけるペーシング効果の向上について確認した。

また、これまでに整備を行ってきた NSIM を拡張し、ユーザに使いやすい形で RDMA をサポートした大規模インターコネクトシミュレータ NSIM-ACE を実装した。基本的な RDMA 通信について NSIM-ACE によるシミュレーションと実機との比較評価を行い、今後の ACP アプリケーションの性能推定に向けて良好な精度を達成していることを確認した。

アプリケーションへの適用に関しては、これまでの研究で非ブロッキング通信を利用した実装方法に変更し、さらに高並列環境での性能検証を行ってきたが、今年度は、ポストペタ環境での性能予測を実施した。また、ACP 基本層を利用した実装を行うための準備と、ACP Middle Layer のインタフェースの検討を実施した。OpenFMO については、データ共有機構、及び、遠隔 Atomic 通信の利用により、プログラムコードの簡素化を行い、これまで以上にチューニングが容易となるような実装に変更した。MHD プログラムについては、これまで大規模並列環境で高性能化に取り組んだ成果を利用して、より高並列での性能予測を実施するとともに、領域分割により生ずる袖領域通信を ACP Middle Layer として効率よく実装するためのインタフェースを検討した。これら以外のアプリケーションに関しても、抽出したパターンについて ACP Middle Layer として実装するためのインタフェースを検討した。

【代表的な原著論文】

[1] 森江 善之、南里 豪志、″直接網において複数の通信デバイスを有効に使用する隣接通信アルゴリズムの提案″、ハイパフォーマンスコンピューティングと計算科学シンポジウム論文集、to appear(2015)

[2] T. Takami and D. Fukudome, "An identity parareal method for temporal parallel computations", Lecture Notes in Computer Science Vol. 8384, edited by R. Wyrzykowski, J. Dongarra, K. Karczewski, and J. Wasniewski, pp. 67-75 (2014)

[3] 本田 宏明, 稲富 雄一, 森江 善之, 南里 豪志, 高見 利也, ″分子軌道法に向けた RDMA に基づく通信ミドルウエアの開発″, Journal of Computer Chemistry, Japan, Vol.13, No.6, pp.335-336, 2014