「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

H23 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. 研究実施内容

(文中に番号がある場合は(3－1)に対応する)

(1)通信衝突の影響予測技術開発および集団通信アルゴリズム動的選択技術の基礎評価(インタフェースグループ)

通信衝突の影響予測技術として、隣接通信の通信衝突による性能悪化を改善するための方法に関する研究を実施した。隣接通信は格子状にタスクを並べた際に、隣接するタスク間との通信を行うため、メッシュ/トーラスで隣接通信を行うのであれば、すべての通信が 1 ホップとなり、通信衝突は発生しない。しかしながら、ファットツリーなどのツリー状のネットワークトポロジにおいて隣接通信を実行する際には通信衝突を発生させる可能性がある。このため、本年度は、隣接通信をファットツリーで実行した場合に通信衝突を削減することを目標とした。ここでの通信衝突は、タスク配置に依存しているため、タスク配置最適化により通信衝突を削減する手法の提案を行い、その評価を行った[森江善之、他、"隣接通信に対する通信衝突を考慮した通信性能向上のためのタスク配置最適化の評価" HPCS2012、ポスター発表]。この手法では、各リンクにおける通信量を調べて、隣接通信のボトルネックを割り出すことで通信性能を見積もることを行い、この見積もりにおいて最も通信性能が良くなるタスク配置を探索することで、隣接通信の通信性能の向上を行っている。この研究において、ファットツリーにおいて通信衝突を発生させる隣接通信を実行するときの通信性能の見積もりを行う方法を確立した。

一方、集団通信アルゴリズム動的選択技術として、ランク配置による通信性能の変動を考慮したアルゴリズム選択技術を提案し、基礎評価を行った。具体的には、実行前に取得しておいたネットワークのトポロジ、ルーティング情報と、実行時に得られるランク配置の情報から、通信時の平均的なバンド幅を計算し、その値を用いて各アルゴリズムの性能を予測する。この予測結果を基に、他のアルゴリズムよりも明らかに遅いと予測されたものを選択肢から除外し、残ったアルゴリズムについて実際に実行中に一つずつ試して最速のアルゴリズムを選択する。実験の結果、提案する手法により、オーバヘッドを低く保ちながら高い精度でアルゴリズム選択が行えることを確認した。[南里豪志、他、"ランク配置に応じた集団通信アルゴリズム動的選択技術の提案", 第 133 回ハイパフォーマンスコンピューティング研究会]

(2)基礎的な片側通信ライブラリの確立(プロトコルグループ)

Exa スケールのスーパーコンピュータで想定される数千万～数億プロセスに耐えうる、省メモリな通信レイヤを開発することがプロトコルグループの研究目標である。これに対して、我々は遠隔 Atomic 操作と片側通信を用いたグローバルデータの利用に着目している。ノードごとに分散したメモリの一部を、システム全体で共有するグローバルデータとすることで、ノードあたりのメモリ使用量を増加させることなく、システム規模の拡張を実現することが可能になると考えている。

本年度は、まず既存技術の確認として、オープンソースの MPI ライブラリについて、省メモリ化の

ための既存研究の調査と、実際に公開されているライブラリでの評価を行った[三浦健一,他. ”エクサスケールコンピューティングに向けた省メモリ通信ライブラリの検討”, 第 133 回ハイパフォーマンスコンピューティング研究会]。調査の結果、数千並列程度では問題のないメモリ消費量であったが、我々の想定している数千万プロセスの環境ではプロセスごとに数十 GB のメモリを消費し、そのままではアプリケーションの動作が不可能なレベルであることを確認した。
また、片側通信と遠隔 Atomic 操作を用いたグローバルデータ構造の効率的な操作方法を検討し、実際のハードウェア上で動作させた場合の性能の予測を行った[安島雄一郎, 他, “片側通信による, グローバルデータ構造の効率的な操作方法の検討”, 第 133 回ハイパフォーマンスコンピューティング研究会][秋元秀行, 他, “InfiniBand Atomic Operation の性能評価”, 第 133 回ハイパフォーマンスコンピューティング研究会]。その結果、遠隔 Atomic 操作を用いることで、Put, Get の単純な片側通信のみの場合に比べて、100 万プロセス規模で最大 1000 倍程度の性能改善が得られることを明らかにした。また、遠隔 Atomic 操作をグローバルデータ構造の操作に用いる場合、ハードウェアによる順序保証サポートが性能に大きく寄与することを明らかにした。

(3)パケット制御グループ
本年度は、研究計画に従って、既存の集団通信アルゴリズムを対象に通信衝突を緩和するパケットペーシングの効果について基礎的な評価を実施した。
まず、集団通信の実行時間は、通信衝突の発生によって理想実行時間よりも増加し、集団通信アルゴリズム、プロセス数、インターコネクトのトポロジ、通信サイズなどによっても大きく異なる。そこで、様々な実行環境・状況を対象に、パケットの送出間隔(以下、パケット間ギャップ)を制御するパケットペーシングを適用した集団通信の基本的な性能について調査した。具体的には、2 次元トーラス網、ならびに 3 次元トーラス網において、集団通信アルゴリズム、ノード数、メッセージサイズ、パケット間ギャップ値をそれぞれ変化させ、集団通信に要する時間をはじめとする各種統計値を、インターコネクトシミュレータ NSIM を用いて調査した。その結果、一般的な集団通信にパケットペーシングを適用した場合の有効性を確認するとともに、アルゴリズム、トポロジ、ノード数、メッセージサイズに応じて、集団通信を最適に実行するパケットの送出間隔が異なることが分かった。また、通信ホップ数に応じたパケット間ギャップを設定することで、アルゴリズムによっては大幅な実行時間の短縮が達成されることを確認した。
次に、平成 24 年度において実施するパケットペーシングのモデル化への指針を得るために、一部の集団通信アルゴリズムについて実行時間を最小化させるパケット間ギャップ値の導出手法について検討した。具体的には、同時期に通信リンクを流れようとするメッセージの数、すなわちリンクあたりのメッセージ重複数に着目し、通信ステップ毎に最適なパケット間ギャップのモデル式を設計するとともに、NSIM によるシミュレーション評価を行った。その結果、モデル式に基づくパケットペーシングを行うことで、8K ノードの 2 次元、3 次元トーラス網において、Pairwise exchange アルゴリズムでは約 2 倍速くなることがわかった。また、ノード数が大きくなるにつれて、パケットペーシングによる速度向上率も増加することが確認された。[吉田匡兵, 柴村英智, 井上弘士, 村上

和彰, 全対全通信向けパケットペーシングにおける送信間隔の導出手法, 情報処理学会第 74 回全国大会]
以上のことから、ポストペタスケール級システムでは、集団通信を高速化するパケットペーシングの大きな効果が期待できる。

(4)アプリケーショングループ

大規模超並列計算機での性能測定、および、ポストペタスケールでの性能推定を目的として、テストコードを開発するとともに現状利用できる並列計算機を利用した性能測定と解析を実施した。現在の超並列計算機で実施されているさまざまな科学計算をポストペタスケールの高並列環境で実行するためには、高性能な通信ライブラリを有効に利用することとアプリケーション自体の効率化を行うことが必要となる。そこで今年度は、領域分割計算での効率的な隣接通信の実装を目的として MHD コードを、並列化されたタスクの動的均等化の評価のために FMO コードを、解析対象とした。

MHDコードに関しては、ベクトル機(SXシリーズ)、スカラ機(X86系、SPARC系、POWER系)の計算機システムにおいて性能評価を行い、現状での計算効率を調べた[2]。特にスカラ機では5000 並列以上のシステムを利用し、コードのスケーラビリティを調べ、通信における同期が効率低下の原因になることを明らかにした[3]。

二段階に並列化されたフラグメント分子軌道(FMO)計算プログラムの、細粒度並列部分である電子状態計算に関して負荷分散による効果の実測を行った。この計算では、計算量の不均一な小規模タスクが多数存在するが、超並列実行時には静的な負荷均等化では限界があることが分かっているため、複数ノードにわたる計算機間で動的な均等化を実施するための大域カウンタの導入と、その効果の検証を行った[稲富 雄一、他、“並列 FMO プログラム OpenFMO の性能最適化” HPCS2012、ポスター発表]。また、負荷均等化に関して OS ジッタなどの影響を調査するために、ハードウェアカウンタを利用した測定手法の予備調査を実施した。

§3. 成果発表等

(3－1) 原著論文発表

●論文詳細情報

[1] Fukazawa, K., T. Ogino, and R. J. Walker, “A magnetohydrodynamic simulation study of Kronian field-aligned currents and auroras”, J. Geophys. Res., 117, A02214, 2012 (doi: 10.1029/2011JA016945).

[2] Fukazawa, K., T. Umeda, "Performance measurement of magnetohydrodynamic code for space plasma on the typical scalar type supercomputer systems with the large

number of cores", International Journal of High Performance Computing, 2012 (doi: 10.1177/1094342011434813).
[3] 深沢圭一郎、梅田隆行、南里 豪志、“超並列惑星磁気圏電磁流体シミュレーションに向けた隣接通信の効率化”、2012 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, 101-106, 2012.
[4] Umeda, T., K. Fukazawa, Y. Nariyuki, and T. Ogino, “A scalable full electromagnetic Vlasov solver for cross-scale coupling in space plasma”, IEEE Transactions on Plasma Science, 2012 (in press).

（3－2） 知財出願

① 平成 23 年度特許出願件数（国内 0 件）

② CREST 研究期間累積件数（国内 0 件）