「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

H24 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

(1) インタフェースグループ

通信衝突を低減する隣接通信、集団通信アルゴリズム性能予測、および非ブロッキング集団通信推進機構について、研究開発を行った。

まず隣接通信については、通信パターンをファットツリーにおいて実行した場合に発生する通信衝突を削減することで隣接通信の性能を向上させるタスク配置最適化技術に関する研究を実施した[7]。この技術では、隣接通信の通信パターンをファットツリーにマッピングする際に、律速点となるリンクでの通信量がもっとも低くなるタスク配置を探索、適用することで通信性能を向上させる。この時の隣接通信は、各通信に順序関係のないものを考えており、同時に実行されるため、通信量がもっとも多いリンクで性能が決まるからである。また、本技術のメッシュ/トーラスへの拡張をめざし、メッシュ/トーラスでの通信衝突削減のためのタスク配置最適化技術に関する研究を実施した[11]。この研究では、メッシュ/トーラスのルーティングを考慮したタスク配置最適化を行えるようにした。また、通信タイミングを考慮することで、各通信に順序関係がある通信パターンにも適用できるタスク配置最適化技術の開発を行った。CG 法の通信パターンを用いてネットワークトポロジがメッシュ/トーラスである九州大学の FX10 上で評価実験を行い、本技術の有効性を示した。この研究成果を適用することで通信に順序関係のある隣接通信においても性能向上が可能となる。

次に集団通信アルゴリズム性能予測については、ファットツリー上のプロセスの配置に応じた性能予測技術と、それに基づいたアルゴリズム選択技術について研究開発した[T. Nanri, et al, "Efficient Runtime Algorithm Selection of Collective Communication with Topology-Based Performance Models", PDPTA'12]。さらに、メッシュ/トーラス上のプロセス配置に応じた性能予測技術についても研究開発した[H. Sugiyama, et al, "Performance Prediction Technology for Collective Communication Algorithm on Multi-Dimensional Mesh/Torus", International workshop on HPC, Krylov Subspace method and its application]。これは、メッシュ/トーラス上の衝突の状況をアルゴリズム内の各フェーズに対して解析し、この衝突の影響を考慮してアルゴリズムの所要時間を見積もるものである。衝突を考慮することにより、予測精度を大幅に向上できることを示した。

最後に非ブロッキング集団通信推進機構については、非ブロッキング通信の推進専用にスレッドを用意し、要求された集団通信を順に処理する機構を実装し、既存の機構である LibNBCと比較した[T. Okuma, et al, "Evaluation of Implementation Methods for Non-Blocking Collective Communications in Overlapping Communication and Computation", International workshop on HPC, Krylov Subspace method and its application]。その結果、提案手法は同時に発行される非ブロッキング集団通信数が 1 個の場合に効率が良いのに対し、LibNBC のように集団通信を個々の一対一通信に分けて並行して推進させる手法は、複数の非ブロッキング集団通信が同時に発行される場合に高速であることが分かった。

(2) プロトコルグループ

ポストペタスケールのスーパーコンピュータで想定される数千万から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。本年度は「メッセージパッシング通信プロトコル低遅延・省メモリ化技術の研究開発」として以下を行った。

通信ライブラリのメモリ使用量を定量的に評価する方法の確立として、測定ツールの提案および実装を行った[秋元秀行, 他, “DMATP-MPI: MPI 向け動的メモリ割当分析ツール”, 第 138 回 HPC 研究会, 秋元秀行, 他, “資源使用量集計プログラム、資源使用量集計方法及び資源使用量集計装置”, 特許出願番号 2013-016970]。本測定ツールではアプリケーションやライブラリに変更を加えることなく、動的メモリ使用量をライブラリやその内部関数毎に測定・集計可能である。さらに測定ツールを使用し既存の MPI ライブラリ(Open MPI および MVAPICH MPI の一部)の主要な関数について、プロセス数に対するメモリ使用量の変化を調査した[住元真司, 他, “DMATP-MPI を用いた MPI ライブラリの関数別メモリ使用量評価”, 第 138 回 HPC 研究会]。その結果 MPI_Init 関数のプロセス数に対するメモリ使用量の増加と Unexpected Message のメモリ開放処理に大きな問題があることが判明し、対策が必要との結論を得た。他にもメモリ使用量に関する対策が必要な部分を抽出し対策を進めている。今後、開発ライブラリのメモリ使用量観点の評価は本測定ツールを使用する。

一方、遠隔 Atomic 通信を用いた通信バッファの削減方式や通信データの局所的な集中に対する性能改善の方式検討を進めた[住元真司, 他, “遠隔 Atomic 通信を用いた省メモリ性実現のための方式検討”, 第 138 回 HPC 研究会]。通信バッファのメモリ削減については、「受信バッファ獲得方式」、「送信メッセージ数制限方式」、「遠隔 Atomic 通信を用いた受信バッファ固定方式」、および「遠隔 Atomic 通信を用いた受信バッファ数制限方式」をそれぞれ検討した。また、通信データの局所集中に対する性能改善方式として、サーバーへのアクセス数を一定以下に抑える機能をインターコネクトの構成に合わせて分散配置する方式等を提案・検討した。

複数プロセスのメモリに配置したグローバルデータ構造を主たるデータ格納先として利用する場合のメモリ管理手法として非同期グローバルヒープの提案と初期検討を行った[安島雄一郎, 他, “非同期グローバルヒープの提案と初期検討”, 第 138 回 HPC 研究会]。本提案の特徴はグローバルヒープとして使用するメモリを非同期に、ローカル・リモートノードから動的に変更できる点にあり、その制御方式の提案を行った。初期評価としてグローバルヒープ操作関数の性能評価を行い、インターコネクトの通信順序保証や CPU/インターコネクト間の Atomic 操作の不可分性の有無による性能評価・比較を行った。

(3) パケット制御グループ

隣接・集団通信におけるパケットペーシングの評価とモデル化を実施した。隣接通信は近隣ノード群との 1 対多あるいは多対多通信であり、科学技術計算アプリケーションに多く利用されているものの、多くの HPC 向け通信ライブラリでは規定されていない。そこで、実践的なアプリケーションから利用頻度が高く、効果的な隣接通信パターンの調査を行うとともに、本研究で開発する通

信ライブラリへの実装方針ならびに汎用性のあるユーザ API について検討した。具体的には、隣接通信を行う領域をビットマップパターンで表現し専用の構造体で管理する。そして、構造体の情報とネットワークのトポロジから通信の衝突具合を算出し、通信スループットを最大化させるパケット間ギャップを導くものである。

また、パケットペーシング適用時の隣接通信、ならびに集団通信のモデル化（通信実行時間の定式化）を行った。これは、通信衝突の無い理想的な通信を対象に、様々な通信遅延要素を積算する基本モデル式を定め、通信衝突の発生確率およびパケットペーシングを反映したモデル式を導くことで、異なるアプリケーションやネットワークのトポロジの衝突確率に応じた通信実行時間を求める。

一方、この通信モデル式から、負荷不均衡（ロードインバランス）や通信不均衡（コミュニケーションインバランス）といったインバランスを考慮すると、パケットペーシングの効果が抑制される場合があることがわかった。そこで、パケットペーシングを用いた集団通信に対するインバランスの影響を調査した。具体的には、ロードインバランスやネットワークインバランスに起因する通信開始時刻のインバランスが、パケットペーシングを用いた集団通信の実行に与える影響をシミュレーションによって評価した。3 次元トーラス網ならびに 2 次元トーラス網を対象に、インバランスの付加や MOD ペーシングと呼ぶ最適ペーシング手法の一つを適用した様々な集団通信の実行性能について、インターコネクトシミュレータ NSIM を用いて測定した。その結果、集団通信のアルゴリズムによってインバランスの感受性が異なることがわかった。また、集団通信に対するペーシングの有効性を確認するとともに、メッセージサイズやノード数の増加に応じて実行時間の高速化率も向上することが確認された。さらに、ペーシングを適用した集団通信にインバランスが及ぼす影響を評価した結果、通信アルゴリズムによっては、わずかなインバランスが加わることで実行時間が大幅に増加し，ペーシングの効果を損なう場合があることが明らかになった。このように、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、ロードインバランスやネットワークインバランスが与える影響をシミュレーションによって明らかにした[柴村英智，他，“パケットペーシングを用いた集団通信にするロード／ネットワークインバランスの影響”，第 135 回 HPC 研究会]。来年度は、これまで蓄積してきた様々なパケットペーシング技術について、実機での検証を行う予定である。

以上の研究成果から、来年度以降に開発する、ネットワーク混雑状況に応じた最適なパケット送信間隔の制御技術に向けた指針を得た。

また、大規模並列計算環境の性能予測に関する調査として、重力多体系の時間発展計算を実施するプログラムのスケーラビリティを、Fujitsu の FX10 や InfiniBand クラスタでの大規模な並列実行による実測値に基づいて評価し、性能予測モデルを構築するための基礎データを収集した。

(4) アプリケーショングループ

ポストペタスケールでの実行に向けて開発中のプログラム(OpenFMO、および、電磁流体コード)に関して、通信性能評価のためのモデル構築、および、入手可能な情報の範囲内で並列アプ

リケーションに対するスケーラビリティ調査を実施した。OpenFMO について、京コンピュータに向けて高性能化を進め[5]、マルチノードでの実行時の性能予測を行うためのモデルを構築した。電磁流体コードでは、様々な計算機での性能評価と性能向上のためのチューニングを実施し[1, 2]、隣接ノードとの袖領域通信時間と計算時間の実測値から全体の性能予測を可能にするモデルを構築した。このモデルに関して、現在実測が可能なノード数の範囲内で検証した結果、かなり信頼性が高い性能予測が可能であることが明らかになった。また、本研究で評価対象とするアプリケーションプログラムの多様性を確保するために、これら以外のプログラムとして、依存関係のために並列化できない計算を並列実行するためのアルゴリズム(時間領域並列化)を実装し、スケーラビリティの実測を行った[4]。このほか、公開されている情報の範囲で、様々なプログラムのスケーラビリティについて調査を行った。

§3. 成果発表等

(3－1) 原著論文発表

●論文詳細情報

[1] T. Umeda and K. Fukazawa, “Performance measurement of parallel Vlasov code for space plasma on various scalar-type supercomputer systems”, In: Algorithms and Architectures for Parallel Processing, Lecture Notes in Computer Science, Vol.7439, edited by Y. Xiang, I. Stojmenovic, B.O. Apduhan, G. Wang, K. Nakano, and A. Zomaya, pp.233-240, 2012. (ISBN: 978-3-642-33077-3)
(DOI:10.1007/978-3-642-33078-0_17)

[2] T. Umeda, K. Fukazawa, Y. Nariyuki, and T. Ogino, “A Scalable Full Electro-Magnetic Vlasov Solver for Cross-scale Coupling in Space Plasma”, IEEE Transactions on Plasma Science, Vol. 40, No. 5, pp.1421-1429, 2012.
(DOI:10.1109/TPS.2012.2188141)

[3] 松本幸, 安達知也, 住元真司, 南里豪志, 曽我武史, 宇野篤也, 黒川原佳, 庄司文由, 横川三津夫,「MPI Allreduce の「京」上での実装と評価」, 情報処理学会論文誌コンピューティングシステム(ACS), Vol. 5, No. 5, pp.152-162, 2012.

[4] T. Takami and A. Nishida, “Parareal Acceleration of Matrix Multiplication”, Advances in Parallel Computing, vol. 22, pp.437-444, 2012. (DOI:10.3233/978-1-61499-041-3-437)

[5] 稲富雄一, 眞木淳, 本田宏明, 高見利也, 小林泰三, 青柳睦, 南一生, "京コンピュータでの効率的な動作を目指した並列 FMO プログラム OpenFMO の高性能化", Journal of Computer Chemistry, Japan (in press).

[6] 松本幸, 安達知也, 住元真司, 曽我武史, 南里豪志, 宇野篤也, 黒川原佳, 庄司文由, 横川三津夫, "MPI Allreduce の「京」上での実装と評価", 先進的計算基盤システムシンポジウム（SACSIS2012）, May. 2012.

[7] Yoshiyuki Morie, and Nanri Takeshi, "Task Allocation Optimization for Neighboring Communication on Fat Tree", in Proceedings of 14th IEEE International Conference on High Performance Computing and Communication , pp.1219–1225, Liverpool, United Kingdom, Jun. 2012. (DOI:10.1109/HPCC.2012.179)

[8] K. Fukazawa, T. Nanri, "Performance of Large Scale MHD Simulation of Global Planetary Magnetosphere with Massively Parallel Scalar Type Supercomputer Including Post Processing", in Proceedings of 14th IEEE International Conference on High Performance Computing and Communication, pp. 976-982, Liverpool, United Kingdom, Jun. 2012. (DOI:10.1109/HPCC.2012.142)

[9] S. Fujino, T. Nanri, K. Kusaba, "Balancing Communication and Execution Technique for Parallelized Sparse Matrix-Vector Multiplication", 4th International Conference on Future Computational Technologies and Applications, Jul. 2012.

[10] K. Fukazawa and T. Nanri, "Effective Performance of Large-Scale MHD Simulation for Planetary Magnetosphere with Massively Parallel Computer", Proc. JSST, pp.243-247, 2012.

[11] 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", 2013 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, No. 2013, pp.95-103, Jan. 2013.

（3－2） 知財出願

① 平成 24 年度特許出願件数（国内 1 件）

② CREST 研究期間累積件数（国内 1 件）