「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」

H23 年度
実績報告

平成23年度採択研究代表者

藤澤克樹

中央大学理工学部経営システム工学科・准教授

ポストペタスケールシステムにおける
超大規模グラフ最適化基盤

§1. 研究実施体制

(1)「大規模最適化」グループ

①研究代表者:藤澤克樹 （中央大学理工学部、准教授）

②研究項目

・超大規模データを伴う最適化問題に対する高速計算システムの構築と評価

・高速グラフ探索アルゴリズム開発

・超並列スレッドを用いた数理計画問題に対する高性能ソルバーの開発

・実グラフデータを用いた実証実験

(2)「大規模グラフ処理系」グループ

①主たる共同研究者:鈴村豊太郎 （東京工業大学情報理工学研究科、客員准教授）

②研究項目

・リアルタイム大規模グラフストリーム処理系及びグラフ最適化ライブラリの開発

・X10 言語上の大規模グラフ処理ライブラリの設計・開発

・大規模グラフストリーム処理系設計・開発

・高速グラフ探索アルゴリズム及び数理計画問題の高性能ソルバーの X10 による実装

・実グラフデータを用いた実証実験及び性能最適化

(3)「大規模グラフストア」グループ

①主たる共同研究者:佐藤仁 （東京工業大学・学術国際情報センター、特任助教）

②研究項目

・大規模グラフ処理向けデータストアの開発

・大規模グラフストア設計、プロトタイプ実装
・グラフ I/O ライブラリ開発（X10 ベース）、グラフ I/O 最適化アルゴリズム開発
・性能最適化・安定化
・他コンポーネントとの統合

（4）「対話型閲覧システムグループ」グループ
①主たる共同研究者：主たる共同研究者：脇田建　（東京工業大学情報理工学研究科、准教授）
②研究項目
・超大規模グラフ向けの対話型閲覧システムの開発
・ユーザインターフェース開発（フィルタリングなし）
・高次元レイアウトシステムｰﾌﾟ
・グラフクラスタリング
・ユーザインターフェース開発（フィルタリングあり）
・信頼性解析システム
・スケーラブルなグラフ処理ライブラリ

§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)

○超大規模ネットワークにおける高速探索技術の開発

超大規模ネットワーク(10 億頂点以上)規模での高速なグラフ探索(最短路、幅優先探索)ソフトウェアの開発と評価。またこれらのソフトウェアを用いたグラフ解析ソフトウェア(中心性の計算やクラスタリング)のプロトタイプの実装を行った。例えば全米道路ネットワークに対する全対全最短路問題に対しては、これまでの方法だと Δ-stepping algorithm で 9.1 年、また Dijkstra's algorithm with Multi-Level Buckets で 4.9 年の年月を要するところを、我々の実装では MSLC-algorithm の高速化を行い、厳密な distance-table を 7.75 日で求めることに成功した(使用環境:4-way Opteron 6174 2.2 GHz (12 cores x 4)、 256 GB、 GCC-4.6.0)。

○最適化ソフトウェア(数理計画問題)の開発

グラフ最適化問題に関するポストペタスパコンでの基盤技術の確立を目指すため、その重要な構成要素として半正定値計画問題(SDP)や混合整数計画問題(MIP)に対するソルバーの開発を推進し、メニーコア CPU におけるマルチスレッドでのアルゴリズム動作の高速化等を行った。特に SDP は組合せ最適化、 システムと制御、 データ科学、 金融工学、 量子化学など非常に幅広い応用を持ち、現在最適化の研究分野で最も注目されている最適化問題の一つとなっている。また今後のエネルギー供給計画(スマートグリッド等)では非線形の複雑な最適化問題を扱う必要があり、これらの問題に対して強力な緩和値を算出できる SDP の高速計算技術の確立が急務とされている。SDP に対しては高速かつ安定した反復解法である内点法アルゴリズムが存在しているが、巨大な線形方程式系の計算が大きなボトルネックとなっている。申請者のグループでは内点法アルゴリズムを記述したソフトウェアの開発・評価・公開を 15 年以上行っており、疎性の追求、計算量やデータ移動量などによる計算方法の自動選択などの技術を他に先駆けて実現し、大規模な並列計算等によって上記のボトルネックの高速化と世界最大規模の SDP を高速に解くことに成功した。具体的には TSUBAME 2.0 の 410 ノード、 820CPU、 4920 コアを用いて世界最大規模の SDP (n = 486,600、 m = 379,350; n は行列の大きさ, m は制約条件の数)を解いた(2012 年 2 月 12 日)。

○グラフ分割による大規模2部グラフのリアルタイム解析

近年、グラフ構造に対するマイニング技術が注目を集めている。従来の大規模グラフ処理の研究はデータを蓄積して実行するバッチ処理が中心となっているが、リアルタイムな処理が要求される分野も数多く存在し、データを蓄積せずに逐次に実行するデータストリーム処理の併用が求められる。本研究では、バッチ処理を定期実行しつつデータストリーム処理を行うシステムを実装し、大規模 2部グラフとして Wikipedia の編集履歴を用いて評価した。大規模グラフをリアルタイムで処理するために、ソーシャルネットワークに代表されるグラフの持つ性質であるコミュニティ構造を利用して、いくつかのコミュニティにグラフを分割して処理を行った。結果、4 ノード 48 コ アを用いて、

頂点数が 207,329 と1,847,166、エッジ数が 22,034,825 の 2 部グラフ を処理し、グラフ分割数に対してデータストリーム処理で 2乗、バッチ処理で 3 乗の 高速化を 30 分割において実現した。その結果、実データの到着レートの約 32 倍の速度でデータストリーム処理することができた。

○スケーラブルなグラフ探索アルゴリズムに研究開発

Graph 500とは、スーパーコンピュータのグラフ処理性能を測定する新しいベンチマークである。スパコンのベンチマークでは、数値計算性能を測るLinpackによるTop500が有名だが、近年、大規模グラフ処理が、重要性を増しており、Graph500ベンチマークが広がりを見せている。Graph500ベンチマークはリファレンス実装が公開されているが、リファレンス実装は分散メモリ環境で大規模にスケールさせることができない。そこで、大規模にスケール可能な隣接行列の2次元分割に注目した。本グループでは、2次元分割をベースとしたアルゴリズムに様々な最適化を施し、1366ノード、16392コアを使用し、頂点数2^36(687億)、枝数2^40(1.1兆)のグラフのBFS(幅優先探索)を10.955秒で計算することに成功した。これは100.366GE/sという速度であり、これによりSC2011で発表された最新のGraph500リストでTSUBAME2.0は3位となった。また、ベンチマークスコアを出すだけでなく、Infiniband Fat-Tree接続ネットワークによる大規模分散環境における、最適化した実装とリファレンス実装の性能特性の解析も行った。

○複数 GPU を用いた汎用グラフ処理モデル GIM-V 実装におけるデータ配置の最適化

ペタバイト級の大規模グラフ処理手法として、GIM-V と呼ばれる MapReduce プログラミングモデルに基づいた行列ベクトル積の処理モデルがある。一方、HPC の分野では、エクサスケールのスーパーコンピュータの実現に向けて、スーパーコンピュータへの GPU アクセラレータの搭載が進んでいる。しかし、大規模グラフ処理に対する複数 GPU の使用による高速化、特に、複数 GPU を利用したときの GPU デバイスへの効率的なデータ割り振り手法は明らかではない。我々は、複数 GPU を用い、MapReduce 処理における通信量削減手法としてグラフ分割の効果について調査した結果、GPU の使用により Map 処理が 7.17 倍の高速化を示した。一方で、Sort、Reduce 処理については高速化を示さず、性能改善の余地があることを示した。特に、グラフ分割によりデータ転送を 54%削減したものの、GPU 毎の負荷が不均衡になる現象が発生することを示した。

○PGAS 言語 X10 を用いた Storage Class Memory 考慮した非同期マルチスレッドグラフ探索の設計と実装

SC10 において Pearce らにより提案された Storage Class Memory 考慮した非同期マルチスレッドグラフ探索(MAGT)アルゴリズムがある。我々は、PGAS 言語における並列 I/O インターフェースの設計指針や、既存の PGAS 言語での実装の性能、生産性、問題点を明らかにするために、MAGT アルゴリズムを、PGAS 言語 X10 を用いて実装した。プロトタイプを TSUBAME2.0 上の 1 ノードで実行した結果、1 スレッドの実行と比較して最大 9 倍程度の性能向上を確認した。また、同じ MAGT アルゴリズムの C++実装と比較した結果、X10 のプロトタイプコードでは 10%程度の性能低下がみられ

たが、コード行数においては 40%程度の記述量の削減を確認した。

○細粒度 I/O を考慮したオンデマンド階層型データストアの TSUBAME2.0 への適用

従来のスーパーコンピュータ上では I/O 性能向上のために Lustre や GPFS などの並列ファイルシステムが広く用いられている。しかし、スーパーコンピュータ上のファイルシステムは計算ノード間で共有されるため、複数のユーザの複数のアプリケーションによりメタデータサーバ、及び、I/O サーバ上での I/O 競合が発生し、I/O スループットや IOPS 性能の低下が問題になる。特に、大規模グラフ処理で発生する I/O はランダムアクセスが多く、それに伴って細粒度な I/O が増加するため、この I/O 競合の問題が顕著である。我々は、TSUBAME2.0 上の PBS Pro ジョブスケジューラと連携して Gfarm ファイルシステムをオンデマンドで構築することで計算ノードに搭載された SSD を集約し、オンデマンドに単一のストレージとして扱うためのツールのプロトタイプを作成し、手法の妥当性を検討した。

○大規模グラフの可視化に関する研究

大規模グラフの可視化については，一般的なグラフのレイアウトアルゴリズムが計算量の複雑さの面で，それらを単純に応用することでは大規模化は難しい．また，表示されたグラフ自体が複雑なため，それを扱うための直感的で効率的なユーザインタフェイスの提供が欠かせない．前者の問題については，高速かつグラフの特性を生かしたグラフの分割手法，すなわちグラフクラスタリングの研究を実施した．後者の問題については，最適化手法を応用した効率的なグラフレイアウトアルゴリズムとインタラクティブなグラフ操作手法についての理論的な研究を行った．

§3. 成果発表等

（3－1） 原著論文発表

●論文詳細情報

1. Makoto Yamashita, Katsuki Fujisawa, Mituhiro Fukuda, Kazuhide Nakata, and Maho Nakata, “Parallel solver for semidefinite programming problem having sparse Schur complement matrix”, ACM Transactions on Mathematical Software, 2012 年(査読付き、採択済)

2. Yuji Shinano, Tobias Achterberg, Timo Berthold, Stefan Heinz, Thoresten Koch,” ParaSCIP: A Parallel Extenstion of SCIP”, Competence in High Performance Computing, edited by C.Bischof, H.-G.Hegering, W.Nagel, G.Wittum, Springer Germany, 2012 年(査読付き、採択済)

3. 西井俊介、鈴村豊太郎、”データストリーム処理によるリアルタイム性を考慮した大規模グラフ処

理基盤とスペクトラルクラスタリングへの応用”、情報処理学会 SACSIS 2012 先進的計算基盤システムシンポジウム、 2012 年 5 月(査読付き、採択済)

4. Koji Ueno and Toyotaro Suzumura, “2D Partitioning Based Graph Search for the Graph500 Benchmark”, IPDPS PerLearning Workshop 2012, 2012 年 5 月(査読付き、採択済)

5. Hiroya Matsuura and Toyotaro Suzumura, “A Highly EfficientConsolidated Platform for Stream Computing and Hadoop “, IPDPS HIPDC Workshop 2012, 2012 年 4 月, (査読付き、採択済)

6. Tamas Fleiner and Naoyuki Kamiyama, “A Matroid Approach to Stable Matchings with Lower Quoats”, In Proc. 23rd Annual ACM/SIAM Symposium on Discrete Mathematics (SODA2012), pp.135-142, 2012 年(査読付き)

(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)