「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22年度採択研究代表者

建部　修見

筑波大学システム情報系・准教授

ポストペタスケールデータインテンシブサイエンスのためのシステムソフトウェア

§1. 研究実施体制

(1)筑波大グループ

①研究代表者:建部　修見 （筑波大学システム情報系、准教授）

②研究参加者:田中　昌宏、平賀　弘平、Joel Tucci、木村　浩希、小林　賢司、三上　俊輔、大辻　弘貴、大西　健太

③研究項目

・分散ファイルシステム

・大規模データ処理実行基盤

(2)電通大グループ

①主たる共同研究者:大山　恵弘 （電気通信大学大学院情報理工学研究科、准教授）

②研究参加者:石黒　駿、村上　じゅん

③研究項目

・計算ノード OS

§2. 研究実施内容

・分散ファイルシステム

研究の狙いは、CPU コア数の増加に対し，アクセス性能がスケールアウトし，かつアクセス応答時間が長くならない分散ファイルシステムの設計を行うことである。

本年度は，メタデータの冗長管理の設計を行った。メタデータの冗長管理においては、冗長に保持するデータの同期方式によりどのような障害について耐障害性をもつかが決まる。まず、それらの関係を明らかにし、それぞれの同期方式について定量的に性能を計測した。ログの書き込みが大きなオーバヘッドとなり得るが、その書き込みをRAMディスクにすることにより、同期方式の違いによる性能差はほとんど見られなかった。研究成果は情報処理学会 HPC 研究会において発表した。また，遠隔ファイルアクセスの性能を向上させるための設計および評価を行った。性能向上のためには、通信遅延を隠蔽し、ネットワーク帯域を十分に活用する必要がある。まず、同期通信プロトコルについて、アクセスパターンを認識する手法、アクセスパターンに応じてバッファサイズを変更する手法を提案し、評価を行った。その結果、ほとんど最適なバッファサイズを選択することに成功した2．。さらに、その手法を非同期通信プロトコルについて拡張した。非同期通信プロトコルを用いることにより、同期通信プロトコルを上回る性能を達成した。研究成果は情報処理学会 HPC 研究会において発表した。

今後、CPU コア数の増加に対し、アクセス性能がスケールアウトし、アクセス応答時間が長くならないために、メタデータを分散管理するための方式について検討し、設計をすすめていく。

・計算ノード OS

研究の狙いは，分散ファイルシステムの性能を最大限に引き出すためのカーネルドライバおよびキャッシュ管理技術を構築することである。

現在までに，キャッシュ管理およびカーネルドライバのプロトタイプの大まかな設計を行った。第一に，ローカルストレージへのアクセスの高速化を実現するカーネル機構の設計と予備実験を行った。ユーザレベルファイルシステム構築を支援するシステム FUSE に存在する実行時間オーバヘッドを，キャッシュやカーネルドライバの利用により削減する手法を構築した。実験を通じて，理想的な条件において実現可能な性能の見積もりを得た。本手法をGfarmファイルシステムに導入することにより，シーケンシャル書き込みの性能が著しく向上することが確認できている。その後，現実的な科学技術計算アプリケーションを用いた，本手法の評価実験を進めている。同時に，Gfarm のメタデータサーバとの通信処理の一部をカーネル内で行うためのドライバを実装している。第二に，キャッシュの導入によるローカルストレージの活用についても研究を行った。ファイルのキャッシングを重複除外と組み合わせて，データ転送量やメモリ消費量を削減する手法を設計した。様々なアプリケーションを用いて，重複除外によるデータ削減量を測定する実験も行った。

今後に行う研究は以下の通りである。第一に，余剰コアを有効利用するためのノード内スケジューラの設計を進める。OS ノイズを削減し，アプリケーションへの性能面での影響を小さくする手法

を設計する。第二に，重複除外に視野を限定しない柔軟な形でキャッシュ機構の設計と実装を進める。その際には余剰コアの利用を検討する。第三に，メタデータサーバとの通信処理のためのカーネルドライバに関しては，より多くの通信処理をカーネル内で実行できるように，これまで未実装であったものについて実装を行う。

・大規模データ処理実行基盤

研究の狙いは、データインテンシブサイエンスのアプリケーションを効率的に実行するためのMPI-IO，大規模ワークフロー実行，MapReduce 処理などの実行環境の研究開発を行うことである。

本研究提案で研究開発する分散ファイルシステムは，全体としてのファイルアクセス性能はスケールアウトするが，ファイルアクセス性能が非均一となる。そのため，効率的に利用するためには，データアクセスについての局所性を利用し、データ移動を最小化することが重要となる。本年度は、データアクセスの局所化を行い、データ移動を最小化するためのプロセススケジューリングに関する研究を行った。大規模ワークフロー実行は、タスク間のデータ依存によりタスクグラフが構成される。タスクグラフの枝はデータの依存関係を表し、データ移動を最小化するためには、エッジカットを最小にするグラフ分割を考えることになる。ただし、並列実行を目的とする場合は、並列に実行できるタスクを分割する必要がある。このことにより、単純なグラフ分割問題に帰着することはできないことを示し、多制約タスク分割問題に帰着できることを示した。さらに、開発しているワークフローエンジンに、その多制約タスク分割を組み込み、性能評価を行い、データ転送量、ワークフロー実行時間を短縮させられることを確認した。

今後は、より大規模なワークフローについての実行を可能とするための階層的なワークフロー実行エンジンの設計を行う。

§3. 成果発表等

（3－1） 原著論文発表

●論文詳細情報

1. 木村浩希，建部修見，「MPI-IO/Gfarm：分散ファイルシステム Gfarm のための MPI-IO の実装と評価」，情報処理学会論文誌，No.52，Vol.12，pp.3239-3250，2011
2. 大辻弘貴，建部修見，「アクセスパターンと回線遅延を考慮した遠隔ファイルアクセスの最適化」，論文誌コンピューティングシステム(ACS)，情報処理学会，No.4，Vol.4，pp.122-134，2011
3. Shunsuke Mikami, Kazuki Ohta and Osamu Tatebe, "Using the Gfarm File System as a POSIX compatible storage platform for Hadoop MapReduce applications", Proceedings of 12th IEEE/ACM International Conference on Grid Computing (Grid

2011), pp.181-189, 2011 (DOI: 10.1109/Grid.2011.31)
4. Kenji Kobayashi, Shunsuke Mikami, Hiroki Kimura, Osamu Tatebe, "The Gfarm File System on Compute Clouds", Proceedings of 1st International Workshop on Data Intensive Computing in the Clouds (DataCloud 2011), 2011 (DOI: 10.1109/IPDPS.2011.255)
5. Hiroki Kimura, Osamu Tatebe, "MPI-IO/Gfarm: An Optimized Implementation of MPI-IO for the Gfarm File System", Proceedings of 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp.610-611, 2011 (DOI: 10.1109/CCGrid.2011.82)
6. Masahiro Tanaka, Osamu Tatebe, "Workflow Scheduling to Minimize Data Movement using Multi-constraint Graph Partitioning", Proceedings of IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2012 (accepted)

（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)