「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 22 年度採択研究代表者

H26 年度
実績報告書

建部 修見

筑波大学大学院システム情報系
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 研究項目
・計算ノード OS

§2. 研究実施の概要

次世代 DNA シーケンサ、放射光、加速器などの実験装置、スパコンによる数値シミュレーションなどによりペタバイト規模を超える大規模データが生成されるようになった。この大規模データを解析することによる科学的手法はデータインテンシブサイエンスと呼ばれる。本研究課題は、大規模データ解析のためのシステムソフトウェアとして、ペタバイト規模のデータの高速アクセスを実現する分散ファイルシステム、計算ノード OS、大規模データ処理実行基盤の研究項目を実施している。平成 26 年度の研究概要を以下にまとめる。

・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を設計、実装することである。本年度は、主にオブジェクトストレージ、キーバリューストアの次世代デバイス向け設計に基づくプロトタイプ実装による高度化、ノード間冗長符号書込の効率的な実装に関する設計とプロトタイプ実装による評価を行った。ノード間冗長符号書込については、ストレージ側で冗長符号処理とデータ転送を行うアクティブストレージ機構の提案を行った。プロトタイプ実装により、クライアント側のオーバヘッドを増加させることなくノード間 RAID-4 による書込が実現した。今後さらなる改良を続けていく予定である。

・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。本年度は、計算ノード OS が提供する各機構および全体システムの実装、性能評価、高性能化、高信頼化をさらに進めた。具体的には、キャッシュ管理機構とカーネルドライバを連携させ、性能測定実験を行った。そのキャッシュ管理機構はクライアント間でキャッシュを共有する cooperative caching 機構であり、計算ノードのメモリ資源を有効に利用することを可能にする。また、そのカーネルドライバは InfiniBand RDMA を用いて高速にファイルデータを転送することができる。そのファイルデータ転送機構についても高性能化、高信頼化、実験による性能評価を進めた。次に、OS ノイズ（OS ジッタ）に関する実験をさらに進め、様々なパラメタを与えたときの科学技術計算アプリケーションの性能を測定した[1]。また、分散ファイルシステムへの重複除外機構の導入についても研究を行った。クライアントノード上キャッシュに対する重複除外の適用についての研究を論文にまとめるとともに、I/O ノードが管理するファイルデータに対する重複除外機構を設計、実装した。

・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、MPI-IO、大規模ワークフロー実行、MapReduce 処理、バッチキューイングシステム、データベース管理システムなどの実行環境を設計、実装することである。本年度は、昨年度からすすめているワークフロー実行における効率的なノード内スケジューリング手法の高度化と詳細な解析をすすめた[2]。バッチキューイングシステムについては、I/O インテンシブなジョブと CPU インテンシブなジ

ョブが混ざったジョブについてファイル位置を考慮したジョブスケジューリング手法の設計を行った。データベース管理システムにおける大規模中間データを効率的に処理するための演算スケジューリングの開発と評価を行った。

【代表的な原著論文】

[1] Yoshihiro Oyama, Shun Ishiguro, Jun Murakami, Shin Sasaki, Ryo Matsumiya, Osamu Tatebe, “Reduction of Operating System Jitter Caused by Page Reclaim”, In Proceedings of the 4th International Workshop on Runtime and Operating Systems for Supercomputers (ROSS 2014), Munich, Germany, June 10, 2014. (DOI: 10.1145/2612262.2612270)

[2] Masahiro Tanaka, Osamu Tatebe, “Disk Cache-Aware Task Scheduling For Data-Intensive and Many-Task Workflow”, Proceedings of IEEE International Conference on Cluster Computing (Cluster), pp.167-175, 2014 (DOI: 10.1109/CLUSTER.2014.6968774)