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

H25 年度
実績報告

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 と、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 25 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、前年度までのプロトタイプ実装を元に分散ファイルシステムのメタデータサーバの設計の改良、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とプロトタイプ実装、ノード間冗長符号書込の効率実装に関する設計とプロトタイプ実装を行った。メタデータサーバでは 15 サーバで 270,000 IOPS を達成したが、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、まず、分散ファイルシステム Gfarm のためのカーネルドライバの実装を進めた。具体的には、InfiniBand 高速通信機能や cooperative caching のための機能の追加など、超高性能計算環境上での利用において重要となる機能の追加を行った。次に、cooperative caching 機構の設計とプロトタイプ実装を進めた。どのファイルデータがどのノードにキャッシュされているかを管理する手法は複数あるが、それらを比較検討し、良好と考えられる手法の実装を進めた。特に、少ないノード間通信量でキャッシュ情報管理が可能な、統計的距離の概念を利用した手法を提案した。重複除外キャッシュ機構についても研究を進め、マルチコアと GPU を用いた並列化によりファイルシステムの処理を高速化する手法の提案、実装、評価を行った。最後に、現実的な科学技術計算アプリケーションに影響を与える OS ノイズについての調査研究をさらに進め、新しい実験結果や知見を得た。

・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、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開発と評価、バッチキューイングシステムにおけるジョブ記述言語の拡張とファイル位置を考慮したジョブスケジューリング手法、自動複製作成手法の開発と評価、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。