「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

H24 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(筑波大)
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(電通大)
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、メタデータの分散管理の設計を行った。目標はスケールアウト性を確保するため、シェアードナッシング型の分散キーバリューストアにおける実装が可能な設計である。ファイルシステムのメタデータで分散管理が難しいところはディレクトリ構造の管理であり、問題は同一ディレクトリ内のエントリの並列追加、ディレクトリの移動、消去などのファイルシステムのディレクトリ操作を、いかに一貫性を保ちつつ行うことができるかである。この問題に対し、ダイナミックソフトウェアトランザクションメモリ（DSTM）に基づくノンブロッキング分散トランザクションをキーバリューストアで設計することにより解決を図った。初期性能評価においてメタデータサーバ数を増加させることにより IOPS が増加することが示された。

NVRAM 等、次期ストレージデバイスにおける効率的なファイルシステムを目標として、ログ構造化ファイルシステムをベースに設計を行った。データインテンシブコンピューティングにおいて重要な操作となる並列書き込みを効率的に行うため、書き込み、更新操作は連続アクセスとなるように設計を行った。初期性能評価においては、物理的な性能に迫る性能を示し、現在利用されているファイルシステムに比べ高い性能を示した。

また、ストレージノードをまたがった冗長符号格納方式、遠隔リモートアクセスに基づく高速なファイルアクセスについても研究を行った。

今後は、分散メタデータサーバの網羅的な性能評価、次期ストレージデバイスにおけるファイルシステムの性能評価、冗長符号による書き込みの効率化をすすめていく。

・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.

本年度は，以下の研究を行った. まず，現実的なアプリケーションを用いて OS ノイズによる性能の変化を測定し，ページ管理に関する OS ノイズに起因する性能低下を抑える手法を提案した. この手法により，OS ノイズがアプリケーションの性能に与える影響が小さくなることを確認した. また，分散ファイルシステムのファイルデータを計算ノード OS にキャッシュする機構については，設計と実装を進めた. 多くの余剰コアがある環境を想定し，キャッシュ機構の処理を並列化した. ファイルデータのキャッシュをストレージにではなくメモリに配置する手法についても予備実験を行った. さらに，分散ファイルシステムのためのカーネルドライバについては，実装を進め，ファイルの読み書きなどのための通信もカーネルレベルで処理できるようにしている. コンテキストスイッチやメモリコピーの回数を減らして高速化を図るカーネルドライバの研究については，国際会議 SC12 のワークショップにて発表した[2].

今後は，OS ノイズを削減する方法についてさらに調査，実験を行い，OS ノイズを効果的に削減するための方法を構築する．また，今までに開発してきたカーネルドライバを用いて性能評価を行う．キャッシュ機構に関しては，計算ノード OS のローカルメモリを有効に活用する方式の構築を目指す．

・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、昨年度に引き続き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すすめた。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ず、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[1]。また、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い、初期性能評価を行った。

今後は、大規模アプリケーション実行を可能とするためのプロトタイプ実装をすすめ，性能評価を行う。また，より効率的に実行するための，データ移動と計算処理のオーバラップを考慮したスケジューリングなどの研究をすすめ性能評価を行う。

§3. 成果発表等

（3－1） 原著論文発表

●論文詳細情報

1. Masahiro Tanaka and 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), pp.65-72, 2012 (DOI: 10.1109/CCGrid.2012.134)
2. Shun Ishiguro, Jun Murakami, Yoshihiro Oyama and Osamu Tatebe, “Optimizing

Local File Accesses for FUSE-Based Distributed Storage", Proceedings of the International Workshop on Data-Intensive Scalable Computing Systems (DISCS), 2012 (DOI: to be assigned)