HPC/ビッグデータ/AI を推進するシステムソフトウェアの研究

Research of System Software for HPC/Big Data/AI

建部 修見
筑波大学計算科学研究センター

１．研究目的

大規模データ解析，ビッグデータによる AI，大規模 HPC アプリケーションなど大規模データを入出力するアプリケーションをスーパコンピュータで実行する需要が増している。このようなアプリケーションでは，演算性能を向上させたとしても I/O 性能，ストレージ性能がボトルネックとなってしまい，このことが大きな問題となっている。本プロジェクトでは，この性能ボトルネックを軽減し，I/O 性能，ストレージ性能をスケーラブルに向上させるためのスーパコンピュータのアーキテクチャ，システムソフトウェアの研究開発を行う。特に，計算ノードのストレージシステムを活用したアーキテクチャとシステムソフトウェアの研究開発を進める。

２．研究成果の内容

これまでストレージ性能のボトルネックを解消するため，計算ノードのローカルストレージを活用したアドホック分散ファイルシステム CHFS の設計を行ってきた。CHFS は，既存のシステムと比較して高い I/O バンド幅，メタデータ性能を示したが，並列ファイルシステムとは別のファイルシステムであるため，利用者が並列ファイルシステムとの間で必要なファイルのコピーを行う必要がある。この操作はステージングと呼ばれるが，しばしばファイルやディレクトリの指定誤りにより問題が起こる。この問題を解決するため，CHFS にキャッシングファイルシステムの機能を追加するための設計を行った。キャッシングファイルシステムは，並列ファイルシステムと同じ名前空間をもち，並列ファイルシステムとの間のファイルコピーはシステムが自動的に行う。しかしながら，これまでキャッシングファイルシステムではメタデータ性能が低くなる問題があった。メタデータは並列ファイルシステムで管理するため，純粋にオーバヘッドが増えるためである。この問題を解決するため，利用者にとって無理のない範囲で並列ファイルシステムとキャッシングファイルシステムの間の一貫性を緩めることを考える。ここで以下の仮定をおいた。

1. 入力ファイルは，ジョブ実行中は変更されない
2. ジョブによるファイル，ディレクトリ作成は必ず成功する
3. 既存ファイルを更新するときにはその前に読込む
4. 並列ファイルシステムへの更新はジョブ終了時までに行われる

これらの仮定はバッチキューイングシステムを利用する場合，暗黙に行われている仮定である。これらの仮定をおくと，並列ファイルシステムとの一貫性のチェックを大幅に削減することが可能となる。これらの仮定のもと，CHFS にキャッシングファイルシステムの機能を追加するため，メタデータにダーティフラグとキャッシュフラグを追加し，キャッシュ機能，フラッシュ機能の設計を行った。

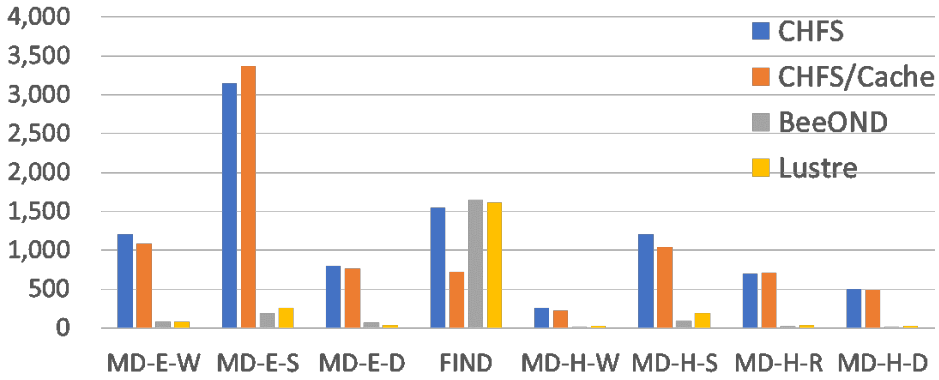

図 Cygnus 48 ノードにおけるメタデータ性能

図に Cygnus 48 ノードを用いた時のメタデータ性能の評価結果を示す。MD-E は各プロセスがそれぞれ異なるディレクトリに 0 バイトのファイルを作成，メタデータ取得，消去する操作を表し，グラフは 1 秒あたりの操作数を示す。MD-H は各プロセスが同一のディレクトリに 47,008 バイトのファイルを作成，メタデータ取得，参照，消去する操作である。FIND は作成したファイルのなかから特定のファイルを検索する操作である。評価対象は，キャッシュ機能を持たない CHFS，今回設計したキャッシュ機能を含む CHFS/Cache，キャッシュ機能を持たない BeeOND，並列ファイルシステム Lustre である。まず，CHFS と CHFS/Cache の比較では FIND を除きほとんどメタデータ性能が変わっていないことが分かった。これにより，当初の目的であるキャッシングファイルシステムにおけるメタデータ性能の低下が解決した。また，BeeOND との比較により，CHFS/Cache の性能が高いことが分かった。Lustre との比較により，CHFS/Cache を用いることにより Lustre の性能を大幅に改善可能であることが分かった。これらの成果は，国際ワークショップ ESSA において発表した。

３．学際共同利用プログラムが果たした役割と意義

学際共同研究プログラムにより，本研究の推進が可能となった。極めて大きな意義を持つ制度である。

４．今後の展望

今後，さまざまなアプリケーションベンチマークによる評価により研究開発を進め，Pegasus 等のスーパコンピュータに導入できるよう努めていきたい。

５．成果発表

(1) 学術論文

1. Osamu Tatebe, Hiroki Ohtsuji, “Caching Support for CHFS Node-local Persistent Memory File System”, Proceedings of 3rd Workshop on Extreme-Scale Storage and Analysis (ESSA 2022), pp.1103-1110, 2022
2. Sohei Koyama, Osamu Tatebe, “Scalable Data Parallel Distributed Training for Graph Neural Networks”, Proceedings of Workshop on AI for Datacenter Optimization (ADOPT'22), pp.699-707, 2022

(2) 学会発表

1. 巨畠 和樹，建部 修見，不揮発性メモリを用いた分散オブジェクトストレージの設計，研究報告ハイパフォーマンスコンピューティング（HPC），Vol. 2022-HPC-184，No. 3，pp. 1-10，2022 年 5 月
2. 巨畠 和樹，小山 創平，平賀 弘平，建部 修見，HPC 環境を想定した探索的データ解析におけるノードローカルストレージの利用の検討，研究報告ハイパフォーマンスコンピューティング（HPC），Vol. 2022-HPC-185，No. 19，pp. 1-8，2022 年 7 月
3. 建部 修見，CHFS アドホック並列分散ファイルシステムのアクセス性能の評価，研究報告ハイパフォーマンスコンピューティング（HPC），Vol. 2022-HPC-185，No. 31，pp. 1-6，2022 年 7 月
4. 平賀 弘平，建部 修見，MPI-IO/CHFS:ノードローカル不揮発性メモリを活用するアドホック分散ファイルシステムのための MPI-IO の設計，研究報告ハイパフォーマンスコンピューティング（HPC），Vol. 2022-HPC-185，No. 33，pp. 1-9，2022 年 7 月
5. 笠井 大暉，建部 修見，分散キャッシュファイルシステムの設計と実装，研究報告ハイパフォーマンスコンピューティング（HPC），Vol. 2022-HPC-186，No. 6，pp. 1-6，2022 年 9 月

使用計算機	使用計算機に ○	配分リソース※	
		当初配分	追加配分
Cygnus	○	14,000	