HPC とビッグデータ 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 and 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

HPC、ビッグデータ AI を推進するにあたり、ストレージ性能は重要な要素となっている。ストレージ性能を向上させるための方法として、計算ノードのローカルストレージを活用する研究が進んでいる。筑波大学計算科学研究センターでは Pegasus システムを導入し、HPC、ビッグデータ AI の推進を行うが、ストレージ性能を十二分に活用するためにはこのローカルストレージを活用する研究の推進と実用アプリでの利活用が求められる。本研究においては、これまで研究開発してきた Pegasus システムの計算ノードに搭載された不揮発性メモリを用いた一時的な並列ファイルシステム CHFS をキャッシングファイルシステムに拡張し、これまで未解決の問題であった並列ファイルシステムへのフラッシュ時の性能低下問題について解決を図る。これにより、HPC とビッグデータ AI の推進を図る。

２．研究成果の内容

スーパーコンピュータにおける並列ファイルシステムの性能は十分ではなく、その性能不足を補うため計算ノードの不揮発性メモリを活用した並列キャッシングシステムの研究開発を行った。並列キャッシングシステムでは、書込んだデータを並列ファイルシステムに書き戻す（フラッシュする）必要があるが、この処理を行っている間、並列キャッシングシステムのアクセス性能が低下する問題があった。この問題の原因は、フラッシュ処理と並列キャッシングシステムのアクセスにおいて計算ノード上のストレージアクセスの競合と、ネットワークアクセスの競合が起こるためである。

この問題を解決するため、I/O-aware フラッシング方式の提案を行った。I/O-aware フラッシング方式では、各キャッシングシステムのサーバにおいてキャッシングシステムのアクセス状況のモニタリングを行う。アクセスがなくなったら一定時間待ち、まだアクセスがないようであればフラッシュする。フラッシュ中に、アクセスが再開したらフラッシュを中断する。この方式は、各サーバがそれぞれフラッシュの判断を行うため、中央サーバが不要で、サーバ数が増えてもオーバヘッドが増えることはない。また、典型的な HPC アプリケーションにおいては、計算フェーズと I/O フェーズが交互に現れることから、この方式により I/O を行わない計算フェーズにフラッシュ処理が行われることが期待される。この方式のための実装方式を示し、研究

開発を進めている CHFS/Cache に実装した。

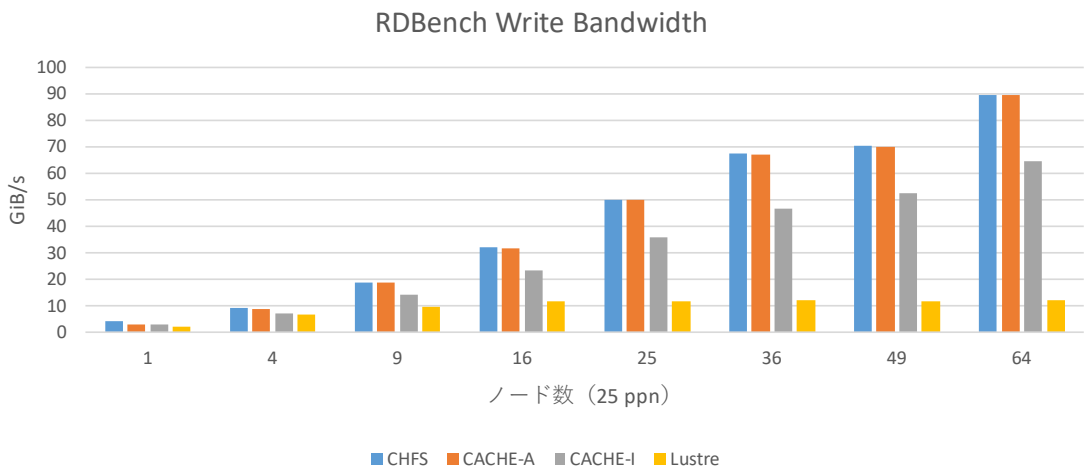

図 1 RDBench における書込み性能

図 1 に RDBench ベンチマークを用いた書込み性能を示す。RDBench は拡散反応系のベンチマークプログラムであり、反復毎にスナップショットファイルを単一共有ファイルに書き出す。実験は Pegasus を用い、各ノード 25 プロセスのウィークスケーリングの評価を実施した。Lustre は、直接並列ファイルシステムに書込んだ時の性能であるが、16 ノード以降は性能が向上していない。CHFS はフラッシュを行わない場合、CACHE-A は提案手法の I/O-aware フラッシング方式を用いた場合であるが、これらはノード数を増やすと性能が向上し、また CHFS と CACHE-A はほぼ同様の性能となっている。これにより I/O-aware フラッシング方式ではフラッシュにおける競合が回避できていることがわかる。一方、よく利用されるフラッシュ方式である書込んだ直後にフラッシュする方式の CACHE-I では競合により 30%ほどの性能低下となっている。これらの成果は国際ワークショップ REX-IO 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 等のスーパー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, Kohei Hiraga, Hiroki Ohtsuji, "I/O-Aware Flushing for HPC Caching Filesystem", Proceedings of 3rd Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO), pp.11-17, 10.1109/CLUSTERWorkshops61457.2023.00012, 2023
2. Sohei Koyama, Kohei Hiraga, Osamu Tatebe, "Accelerating I/O in Distributed Data Processing Systems with Apache Arrow CHFS", Proceedings of 3rd Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO), pp.1-4, 10.1109/CLUSTERWorkshops61457.2023.00009, 2023

(2) 学会発表

1. Osamu Tatebe, Kohei Hiraga, Hiroki Ohtsuji, "I/O-Aware Flushing for HPC Caching Filesystem", 3rd Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO), Santa Fe, NM, Oct. 31, 2023
2. Sohei Koyama, Kohei Hiraga, Osamu Tatebe, "Accelerating I/O in Distributed Data Processing Systems with Apache Arrow CHFS", 3rd Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO), Santa Fe, NM, Oct. 31, 2023
3. Sohei Koyama, (Advisor) Kohei Hiraga, Osamu Tatebe, Fast Checkpointing of Large Language Models with TensorStore CHFS, The International Conference for High Performance Computing, Networking, Storage, and Analysis, ACM Student Research Competition Posters, Denver, CO, Nov. 14-16, 2023
4. Sohei Koyama, Kohei Hiraga, Osamu Tatebe, FINCHFS: Design of Ad-hoc File System for High-Performance Computing Workload, 22nd USENIX Conference on File and Storage Technologies (FAST), Work-in-Progress Report and Poster, Santa Clara, CA, Feb. 27-28, 2024
5. 小山 創平, 平賀 弘平, 建部 修見, Apache Arrow CHFS によるビッグデータ処理の I/O 高速化, 第 190 回情報処理学会ハイパフォーマンスコンピューティング研究発表会, 2023
6. 建部 修見, 平賀 弘平, 前田 宗則, 藤田 典久, 小林 諒平, 額田 彰, Pegasus ビッグメモリスーパコンピューターの性能評価, 第 190 回情報処理学会ハイパフォーマンスコンピューティング研究発表会, 2023
7. 小山 創平, 平賀 弘平, 建部 修見, FINCHFS: アドホック並列ファイルシステムの設計, 第 192 回情報処理学会ハイパフォーマンスコンピューティング研究発表

会, 2023

8. 杉原 航平, 建部 修見, スパースセグメントを活用した局所性志向バーストバッファにおけるフラッシュ手法の検討, 第 192 回情報処理学会ハイパフォーマンスコンピューティング研究発表会, 2023
9. 丸山 泰史, 建部 修見, NVMe SSD 環境における CHFS の設計最適化の検討, 第 192 回情報処理学会ハイパフォーマンスコンピューティング研究発表会, 2023
10. 中野 将生, 建部 修見, Rust の UCX ラッパーasync-ucx の性能評価, 第 192 回情報処理学会ハイパフォーマンスコンピューティング研究発表会, 2023
11. 平賀 弘平, 建部 修見, PMEMBB: 不揮発性メモリと MPI 片側通信を用いた MPI-IO バーストバッファの設計, 第 193 回情報処理学会ハイパフォーマンスコンピューティング研究発表会, 2024
12. 木下 嵩裕, 建部 修見, CA VOL: HDF5 におけるコンテキストによる I/O 最適化, 第 193 回情報処理学会ハイパフォーマンスコンピューティング研究発表会, 2024

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			
Pegasus	○	20,000	40,000
Wisteria/BDEC-01			
※配分リソースについてはノード時間積をご記入ください。			