並列 I/O とストレージシステムの研究

Research of Parallel I/O and Storage System

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 HPC、ビッグデータ AI を推進するにためには、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、並列ファイルシステムとの間に中間的なキャッシングファイルシステム、バーストバッファ等を活用する研究が進んでいる。本研究においては、計算ノードのローカルストレージを活用することによるストレージ性能の向上を目的とする。

２．研究成果の内容

筑波大学のスーパーコンピュータ Pegasus においては、並列ファイルシステムの性能が十分ではなく、性能を向上させるため計算ノードのローカルストレージを活用したストレージシステムの研究開発を実施した。

Pegasus ではローカルストレージとして不揮発性メモリと NVMe SSD を備えている。不揮発性メモリはバイト単位でアクセス可能であり、また RDMA により遠隔ノードからの直接アクセスも可能である。不揮発性メモリの書込み性能は 25 GiB/s 程であるのに対し読込性能は 80 GiB/s である。Pegasus のネットワーク性能は理論ピーク性能が 25 GiB/s であり、不揮発性メモリの性能を十二分に活用するためにはデータアクセスの局所性が重要となる。そのため、データアクセスの局所性を利用でき、かつ不揮発性メモリの性能を十二分に活用できるよう並列 1 次ストレージシステム PEANUTS の設計を行った。PEANUTS では、不揮発性メモリの全領域をメモリ登録し、全領域を RDMA 可能とし、その領域を並列 1 次ストレージとして利用する。また、ストレージサーバを設置せず、I/O ライブラリを設計、実装することにより並列プログラムが直接不揮発性メモリをアクセスするようにした。書込みにおいては、ログ形式でローカルの不揮発性メモリに書込み、同期とファイルクローズのタイミングで書込み領域の共有を行う。これにより、各プロセスはファイルの全領域のアクセスが可能となる。

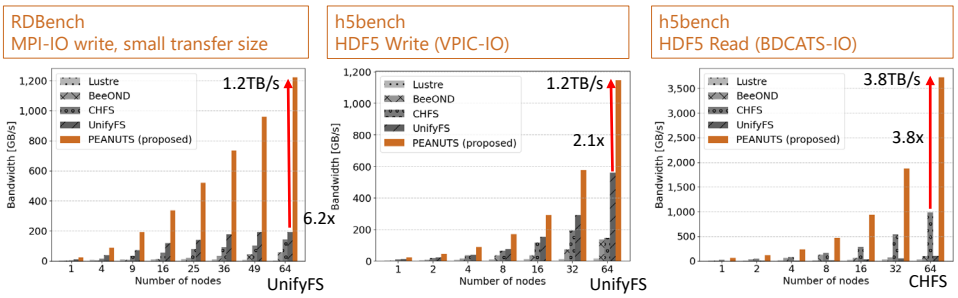

図 1 PEANUTS の性能評価

図 1 に RDBench および h5bench の Write と Read における最先端のシステムとの性能比較を示す。X 軸はノード数を示し、Y 軸は性能を示している。PEANUTS は全てのベンチマークにおいて高い性能を示している。本成果は国際会議 Euro-Par において発表した。

また、これまで研究開発していた CHFS の適用範囲を広げ、更に性能を向上させるため、FINCHFS の設計を行った。FINCHFS では一部制限はあるもののディレクトリ名やファイル名の変更をサポートし、またサーバ単体性能の向上とスケーラビリティの向上を図っている。

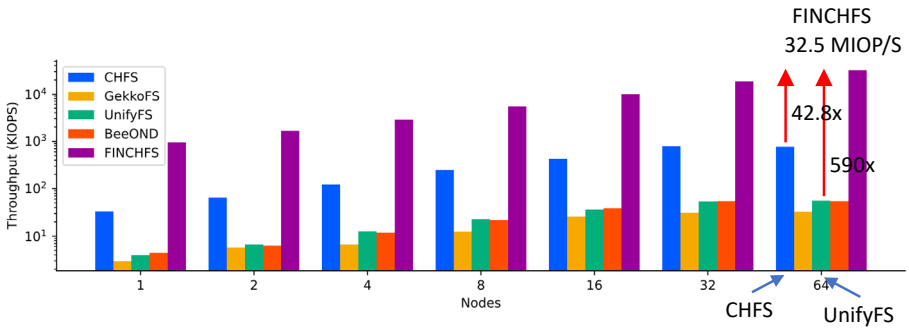

図 2 FINCHFS の性能評価

図 2 に MDtest Hard Write における最先端のシステムとの性能比較を示す。CHFS の 42.8 倍の性能向上を達成した。本成果は IEEE 国際会議 CLUSTER 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

学際共同利用プログラムにより Pegasus を利用できたことが研究推進につながった。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による評価を進め、実用化を図っていきたい。

５．成果発表

(1) 学術論文

1. Kohei Hiraga, Osamu Tatebe, "PEANUTS: A Persistent Memory-Based Network Unilateral Transfer System for Enhanced MPI-IO Data Transfer", Proceedings of 30th International European Conference on Parallel and Distributed Computing (Euro-Par), pp.439-453, 10.1007/978-3-031-69766-1_30, 2024
2. Sohei Koyama, Kohei Hiraga, Osamu Tatebe, "FINCHFS: Design of Ad-Hoc File System for I/O Heavy HPC Workloads", Proceedings of IEEE International Conference on Cluster Computing (CLUSTER), pp.440-450, 10.1109/CLUSTER59578.2024.00045, 2024

(2) 学会発表

1. Kohei Hiraga, Osamu Tatebe, "PEANUTS: A Persistent Memory-Based Network Unilateral Transfer System for Enhanced MPI-IO Data Transfer", 30th International European Conference on Parallel and Distributed Computing (Euro-Par), Madrid, Spain, Aug. 30, 2024
2. Sohei Koyama, Kohei Hiraga, Osamu Tatebe, "FINCHFS: Design of Ad-Hoc File System for I/O Heavy HPC Workloads", IEEE International Conference on Cluster Computing (CLUSTER), Kobe, Japan, Sept. 27, 2024
3. Hiroki Ohtsuji, Munenori Maeda, Reika Kinoshita, Masahiro Miwa, Osamu Tatebe, " Scalable RPC Layer Towards Millions of IOPS per Server", 9th International Parallel Data Systems Workshop (PDSW), Work-in-progress presentation, Atlanta, GA, Nov. 17, 2024
4. Haruka Miyauchi, Sohei Koyama (advisor), Osamu Tatebe (advisor), "Design of Reliable and Efficient Syscall Hooking Library for a Parallel File System", The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), Student Research Competition Undergraduate Poster, Atlanta, GA, Nov. 17-22, 2024
5. Ryosuke Maeda, Masaki Nakano, Osamu Tatebe, " Asynchronous scheduling of communication and I/O integrated with Rust", SupercomputingAsia 2025, Singapore, March 10-13, 2025
6. Osamu Tatebe, "National HPCI Shared Storage", Documenting and Classifying Research Data Storage Infrastructures, Singapore, Mar. 10, 2025
7. Osamu Tatebe, "Recent Trends in Ad-hoc HPC File Systems and Caching File Systems", 4th Workshop on Re-envisioning Extreme-Scale I/O for Emerging

Hybrid HPC Workloads (REX-IO), Keynote, Kobe, Japan, Sept. 24, 2024

8. Mingzhe Yu , Osamu Tatebe, "Distributed K-FAC Over Unstable Networks (unreferred)", 第 195 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, 徳島, Aug. 8, 2024
9. Mingzhe Yu , Osamu Tatebe, " Asynchronous Decentralized Distributed K-FAC: Enhancing Training Efficiency and Load Balancing in Heterogeneous Environments (unreferred)", 第 197 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, 沖縄, Dec. 16, 2024
10. 前田 椋祐, 中野 将生, 建部 修見, "分散ファイルシステムにおける通信イベントと I/O イベントの非同期スケジューリングを統合した非同期 I/O の性能評価", 第 197 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, 沖縄, Dec. 17, 2024

(3) その他

使用計算機	使用計算機に○	配分リソース※		
		当初配分	移行*	追加配分
Cygnus				
Pegasus	○	26,400		
Wisteria/BDEC-01	○	17,600		
	※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。 ＊バジェット移行を行った場合、「+2000」「-1000」のように記入			