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

演算性能の向上に対しデータ入出力性能が向上しておらず、データ入出力は性能ボトルネックとなっている。本研究ではそのボトルネックを解消すべく、計算ノードのローカルストレージをキャッシュとして利用する研究を進める。これまでは利用障壁が大きいこと、必ず性能が向上するわけではないこと、導入・保守のサポートがないことなどいくつか問題があり普及が進んでいなかった。本研究では、一般のユーザが普通に並列ファイルシステムを利用するようにキャッシングファイルシステムを利用することができ、また性能を大きく向上させることを目的とする。

２．研究成果の内容

アプリケーションのデータ入出力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利用したストレージシステムの研究開発を行った。

これまで研究を続けているキャッシングファイルシステム CHFS/Cache については、POSIX インターフェースに対する対応、相対パスの対応等を進めた。これにより大多数の POSIX インターフェースについてシステムコールのフックが可能となった。既存のアプリケーションでは、カレントディレクトリからの相対パスでファイル等を指定していることが多いため、相対パスの対応は重要であった。また、絶対パスの指定において、これまではキャッシングファイルシステム用のマウントポイント以下でキャッシュしていたが、その場合、アプリケーションで指定するパス名の変更が必要であり、全てのアプリケーションで対応できるわけではなかった。この問題を解決するため、/（ルート）ディレクトリをキャッシングファイルシステムのマウントポイントとして、キャッシュするディレクトリを指定可能とした。これにより、アプリケーションの変更をすることなくキャッシュしたいディレクトリだけをキャッシュすることが可能となった。これらの設計により大きく利用障壁を下げられたと考えている。CHFS/Cache を全利用者に利用可能とするため、Pegasus に導入し、Pegasus の利用の手引きにおいてその利用方法の追加を行った。

また、MPI-IO に特化したデータ出力の高速化のための研究を行った。書込みをローカルストレージに行ったからといって高速に書き出せるわけではない。高速に書き出すためには書き出すサイズを大きくする必要がある。連続領域への書込みであれば

バッファリングなどで書込み単位を大きくすることは可能であるが、多次元配列の書き出しなどの場合、メモリ上では連続であっても書込み先が連続ではないため書込み単位が小さくなってしまう。そのため、この問題を解決するバーストバッファの設計を行った。メモリ上では連続であることに着目し、メモリ上のレイアウトを持つ MPI ファイルビューでローカルストレージに書込みを行い、その後元のファイルビューで非同期に並列ファイルシステムに書き込む。アプリケーションはローカルストレージへの書込みしか待つ必要はないため、アプリケーション上での書込み性能は向上する。並列ファイルシステムへの書込みは速くはないが、アプリケーションの実行とオーバラップするため隠蔽される。

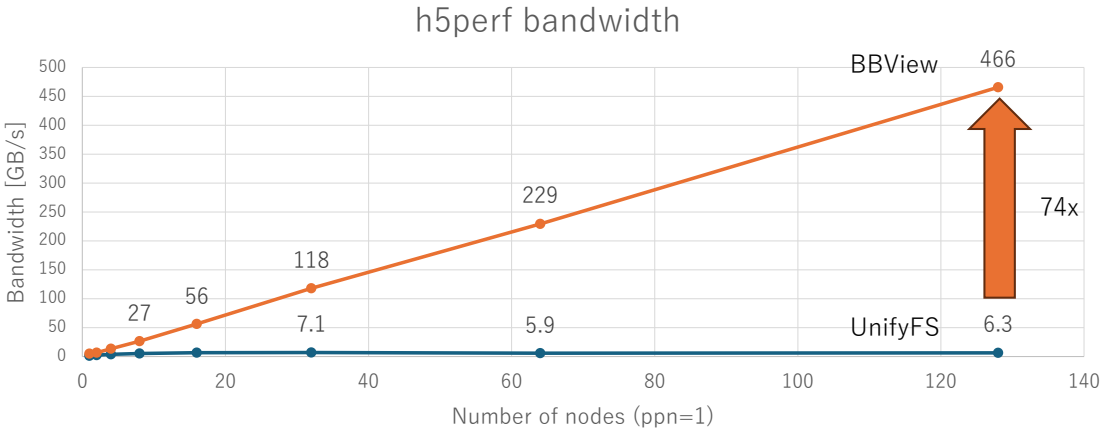

図 1 h5perf によるバンド幅の評価

図 1 に既存の最高性能のシステムである UnifyFS との性能比較を示す。UnifyFS もローカルストレージへ書込みを行っているが、提案手法である BBView では書込み単位を大きくしているため、128 ノードでは 74 倍の性能差となっている。なお、BBView は MPI のモジュールで実装されており、アプリケーションの修正は不要である。本成果については国際ワークショップ REX-IO で発表した。今後はこれらの成果についても一般ユーザが利用可能としていくことを考えている。

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. Mingzhe Yu, Osamu Tatebe, "AD-KFAC: Asynchronous Decentralized Distributed K-FAC with Dynamic Load Balancing and Fault Tolerance", Proceedings of 2025 10th International Conference on Machine Learning Technologies (ICMLT), pp.373-382, 10.1109/ICMLT65785.2025.11193392, 2025
2. Sohei Koyama, Osamu Tatebe, "BBView: A View-Aware Burst-Buffer Mechanism for MPI-IO", Proceedings of 4th Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO), pp.1-9, 10.1109/CLUSTERWorkshops65972.2025.11164194, 2025
3. Mingzhe Yu, Osamu Tatebe, "Block-Diagonal K-FAC: A Trade-off Between Curvature Information and Resource Efficiency", Proceedings of 17th International OPT Workshop on Optimization for Machine Learning (OPT), 10 pages, 2025
4. Norihisa Fujita, Keita Ito, Kohji Yoshikawa, Kohei Hiraga, Osamu Tatebe, Akira Nukada, Taisuke Boku, "Large-Scale Vlasov Simulations for Astrophysics using Non-volatile Memory as Large Memory", Proceedings of International Workshop on Intel eXtreme Performance Users Group (IXPUG SCA/HPCAsia 2026 workshop), pp.339-347, 10.1145/3784828.3785352, 2026

(2) 学会発表

1. Mingzhe Yu, Osamu Tatebe, "AD-KFAC: Asynchronous Decentralized Distributed K-FAC with Dynamic Load Balancing and Fault Tolerance", 2025 10th International Conference on Machine Learning Technologies (ICMLT), Helsinki, May 23-25, 2025
2. 前田 宗則, 平賀 弘平. 大辻 弘貴, 建部 修見, 共有メモリアーキテクチャにおける高性能 RPC の方式検討, 第 200 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, 高松, 8/4-6, 2025
3. 前田 椋祐, 中野 将生, 建部 修見, Pluvio: アドホックファイルシステムのための zero-copy I/O 非同期ランタイムの設計, 第 200 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, 高松, 8/4-6, 2025
4. 小山 創平. 建部 修見, BBView: View を意識した MPI-IO 対応バーストバッファの設計, 第 200 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, 高松, 8/4-6, 2025
5. 中野 将生, 建部 修見, 大規模高速ストレージアーキテクチャの実現に向けた非同期 RPC 基盤の設計, 第 200 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, 高松, 8/4-6, 2025

6. Sohei Koyama, Osamu Tatebe, "BBView: A View-Aware Burst-Buffer Mechanism for MPI-IO", 4th Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO), Edinburgh, Sep. 2, 2025
7. Shingo Hattori, Osamu Tatebe (advisor), "SRAP: Sender-Side Receiver-Aware Port Selection for High-Speed Multi-Flow TCP", The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), Student Research Competition Graduate, Poster, St. Louis, Nov. 16-21, 2025
8. Mingzhe Yu, Osamu Tatebe, "Block-Diagonal K-FAC: A Trade-off Between Curvature Information and Resource Efficiency", 17th International OPT Workshop on Optimization for Machine Learning (OPT), Poster, San Diego, Dec. 6, 2025
9. 小木 勇輝, 建部 修見, 第 3 世代 Optane メモリによる大規模言語モデル KV キャッシュの性能評価, 第 202 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, 12/15-16, 2025
10. 服部 真吾, 建部 修見, ネットワーク転送ベンチマーク向けの仮想ファイルシステムの提案, 第 202 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, 12/15-16, 2025
11. Norihisa Fujita, Keita Ito, Kohji Yoshikawa, Kohei Hiraga, Osamu Tatebe, Akira Nukada, Taisuke Boku, "Large-Scale Vlasov Simulations for Astrophysics using Non-volatile Memory as Large Memory", International Workshop on Intel eXtreme Performance Users Group (IXPUG SCA/HPCAsia 2026 workshop), Osaka, Jan. 26, 2026
12. 藤田 典久, 吉川 耕司, 辻 美和子, 朴 泰祐, 建部 修見, AMD MI300A APU における共有メモリシステムの性能評価, 第 203 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, 札幌, 3/16-18, 2026
13. 中野 将生, 前田 宗則, 建部 修見, LocustaRPC: 次世代リーダーシップマシンのためのスケーラブルな RPC 基盤, 第 203 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, 札幌, 3/16-18, 2026
14. Alexander Klassen, Osamu Tatebe, Tracing the RPC Lifecycle for Performance Analysis in Margo-Based HPC Data Services (unreferred), 第 203 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, 札幌, 3/16-18, 2026

使用計算機	使用計算機に○	配分リソース※		
		当初配分	移行*	一般利用による追加
Pegasus	○	17,600		15,000
Miyabi-G	○	49,500	-720	0
Miyabi-C	○	0	+900	0
	※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			