バーストバッファの研究

Research of burst buffer

建部修見
筑波大学計算科学研究センター

１．研究目的

CPU 性能とストレージ性能のギャップを埋めるため、バーストバッファの研究開発を行う。バーストバッファはストレージアクセスに対するキャッシュとして動作し、アプリケーションによる一時的でバースト的なストレージアクセスを吸収し、性能ギャップを埋めるものである。Cygnus では、各計算ノードに NVMe SSD が搭載されている。この NVMe SSD を活用して分散バーストバッファの研究開発を行う。これまで分散バーストバッファとして Gfarm/BB の研究開発を行ってきたが、アプリケーションにより I/O 要求が異なるため、最適化の方法も様々なものとなる。本プロジェクトでは、それぞれのアプリケーションの I/O 要求に応じた最適化の方法を研究開発する。

２．研究成果の内容

並列ファイルシステムに対するアクセス性能はアクセスパターンにより変わり、特によく利用される単一共有ファイル（single shared file）パターンでは性能が低下することが知られている。単一共有ファイルパターンとは、並列プロセスが単一のファイルにアクセスするパターンである。本研究では、この単一共有ファイルパターンにおいて、ノードローカルストレージを活用してストレージ性能を向上するために MPI-IO の研究を行った。MPI-IO は HPC アプリケーションにおいて標準的に利用される並列 I/O インターフェースであり、MPI-IO ライブラリにおいてノードローカルストレージを用いた最適化を行うことにより、多くの並列アプリケーションにおいて並列 I/O 性能の向上を行うことができる。

ノードローカルストレージは、計算ノードのローカルストレージである。ジョブが投入され、計算ノードが割当てられ、ジョブの実行が開始してから終了するまでしか利用することはできない。これまで、単一共有ファイルパターンにおいてローカルストレージを活用する研究は BurstFS, BeeOND, Gfarm/BB, GekkoFS, SymphonyFS などアドホック分散ファイルシステムを用いるものがあるが、これまで読込、書込で十分な性能を引き出しているものはない。本研究は、この問題に対し MPI-IO 層での解決を図り、NS-MPI の提案を行うものである。NS-MPI ではアドホック分散ファイルシステムの Gfarm/BB を用い、NS-MPI においてファイル分割を行うことで、単一

共有ファイルへのアクセスに対し、プロセス毎ファイルと同様の高いアクセス性能を実現することを目指す。これを実現するために、sparse segments という新しいファイルフォーマットを提案した。その新しいフォーマットを用いて NS-MPI の設計、実装を行った。

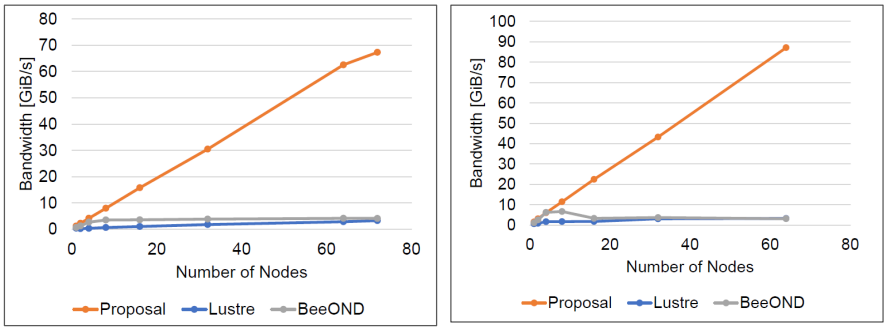

図 1 LES-IO と VPIC-IO における提案手法（NS-MPI）と Lustre、BeeOND との性能比較

図 1 にアプリケーションベンチマークである LES-IO と VPIC-IO を用いた性能評価を示す。提案手法と同様にノードローカルストレージを用いる BeeOND では計算ノード数を増やしたときに性能が向上していないが、提案手法では性能がスケーラブルに向上していることが分かる。

本成果は、IEEE International Workshop on High-Performance Storage および、6th IEEE International Conference on Data Science and Systems (DSS)において発表を行った。

３．学際共同利用が果たした役割と意義

学際共同利用により、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、研究開発した MPI-IO を実用に供するための準備を進めていきたい。また、本研究ではノードローカルストレージとして NVMe SSD を対象としたが、今後不揮発性メモリを搭載することも考えられ、不揮発性メモリに対してさらに研究を進めていきたい。

５．成果発表

(1) 学術論文

- 町田 健太，建部 修見，「大規模メタゲノムデータに対する分散並列相同性検

索システム GHOSTZ PW/GF の提案」，情報処理学会論文誌コンピューティングシステム(ACS)，Vol.13, No.2, pp.13-27, 2020 年 9 月

- Kohei Sugihara, Osamu Tatebe, "Design of Direct Read from Sparse Segments in MPI-IO", Proceedings of the 6th IEEE International Conference on Data Science and Systems (DSS), pp.1308-1315, 2020

(2) 学会発表

- Kohei Sugihara, Osamu Tatebe, "Design of Direct Read from Sparse Segments in MPI-IO", 6th IEEE International Conference on Data Science and Systems (DSS), Yanuca Island, Cuvu, Fiji (Virtual conference), 2020/12/14-16

(3) その他

なし

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	○	31,500	0
Oakforest-PACS	○	240,000	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。			