ストレージシステムの研究

Research of 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

大規模データ解析、ビッグデータによる AI などをスーパコンピュータで実施する場合、ストレージ性能がボトルネックとなる。この性能ボトルネックを軽減するため計算ノードのストレージシステムを活用するストレージシステムの研究を行う。

２．研究成果の内容

スーパコンピュータの計算ノードのストレージシステムを活用したアドホック並列分散ファイルシステム CHFS の設計を行った。アドホック並列分散ファイルシステムとは、計算ノードが割り当てられている間に計算ノードのローカルストレージを用いて一時的に構成される並列分散ファイルシステムである。並列ファイルシステムにおけるストレージ性能のボトルネックの解消を目的としている。

計算ノードのストレージシステムには NVMe SSD に加え不揮発性メモリも用いられる。不揮発性メモリは NVMe SSD と異なり、バイト単位でのアクセスが可能であり、従来のブロックデバイスを想定したファイルシステム等を用いたアクセスでは性能を活用することができない。本研究においては、不揮発性データ構造を用いたデータ管理を用い、具体的にはインメモリ不揮発性キーバリューストアを活用したストレージシステムの設計を行った。複数ノードの不揮発性メモリを用いて分散キーバリューストアを構成する。通信については、スーパコンピュータの高速ネットワークを活用するため RDMA を用いる。分散キーバリューストアについては、ノードの増減時にデータ転送を最小化するためにコンシステントハッシングを用いる。

分散キーバリューストア上に並列分散ファイルシステムを構成するための設計を行った。設計にあたり、ノード数に対するスケーラビリティを阻害する要因となる集中データ構造や逐次処理を避けた。HPC においては大規模ファイルの並列アクセスにおける性能向上が必須であるため、ファイルはチャンクで分割し分散させる。既存研究では GekkoFS が最も近いが、GekkoFS は NVMe SSD を想定し、ファイルデータはファイルシステムで管理し、メタデータは RocksDB で管理している。一方、CHFS では不揮発性メモリの性能を活用するためインメモリ不揮発性キーバリューストアをベースとした設計となっている。また、GekkoFS ではデータ分散はモジュラハッシングを用い、ノードの増減については考慮していない。

CHFS を Mochi-Margo を用いて実装した。Mochi-Margo は HPC 用の RPC ライブラリであり、InfiniBand 等の高速ネットワークを活用することが可能である。

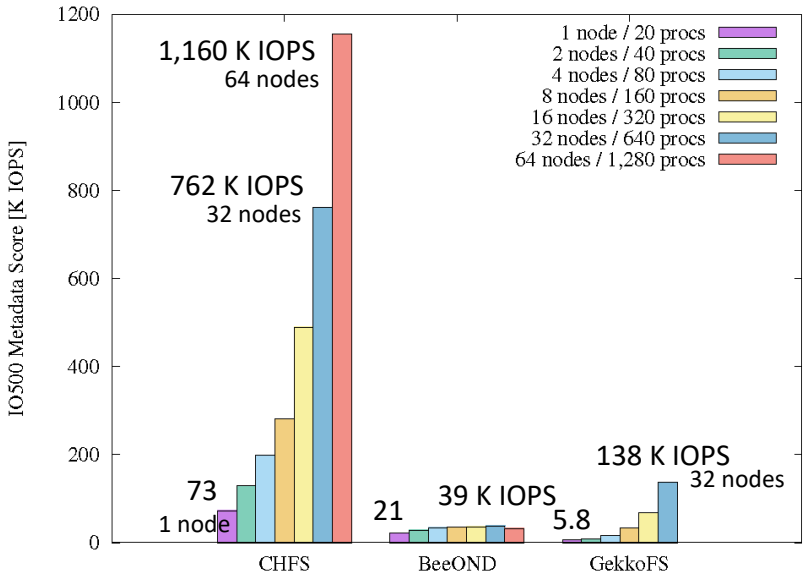

図 1 Cygnus における IO500 メタデータベンチマークの性能

図 1 に Cygnus を用いた IO500 メタデータベンチマークの CHFS の性能を示す。ノード数を 1 ノードから 64 ノードまで増やしていきメタデータ性能を計測した。IO500 メタデータベンチマークは 8 種類のメタデータアクセスパターンの性能の幾何平均である。ノード数を増加させるとメタデータ性能が向上していることがわかる。また既存システムである BeeOND と GekkoFS との性能の比較を行った。BeeOND ではノード数を増やしてもメタデータ性能はあまり向上しない。GekkoFS の性能は向上するが CHFS と比較すると性能差は 5.5 倍であった。なお、本研究成果は ACM 国際会議 HPC Asia 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

CHFS を HPC アプリケーションで利用し性能評価を行っていきたい。また、キャッシュファイルシステムとして利用するための研究開発を行い、並列ファイルシステムと CHFS 間のデータ移動をユーザが行わなくてもいいようにいていきたい。

５．成果発表

(1) 学術論文

- Osamu Tatebe, Kazuki Obata, Kohei Hiraga, Hiroki Ohtsuji, "CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory", Proceedings of the ACM International Conference on High Performance Computing in Asia-Pacific Region (HPC Asia), pp.115-124, 2022

(2) 学会発表

- Osamu Tatebe, Kazuki Obata, Kohei Hiraga, Hiroki Ohtsuji, "CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory", ACM International Conference on High Performance Computing in Asia-Pacific Region (HPC Asia), Virtual Event, 2022/1/12-14

(3) その他

- 平賀 弘平, 建部 修見, 計算ノード上の不揮発性メモリを用いた MPI-IO バーストバッファの設計, 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-183, No. 24, 9 pages, 2022
- 建部 修見, 計算ノードの不揮発性メモリを用いたキャッシュファイルシステムの設計, 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-183, No. 8, 9 pages, 2022

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	○	12,960	0
Oakforest-PACS			
※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。			