様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １１ 日現在

機関番号：１２１０２
研究種目：若手研究
研究期間：2022～2023
課題番号：２２Ｋ１７８９７

研究課題名（和文）次世代データサイエンスのための不揮発性メモリを用いたストレージシステムの研究

研究課題名（英文）Persistent memory-based storage systems for next-generation data science

研究代表者
平賀　弘平（HIRAGA, Kohei）

筑波大学・計算科学研究センター・研究員

研究者番号：２０９３７１２２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では，高性能計算を支えるスーパーコンピュータ用に，不揮発性メモリを活用したストレージシステム「PEANUTS」を設計・開発した．PEANUTSは，データサイエンスに不可欠な単一共有ファイルの大規模並列読み書きを主目的としており，従来のストレージシステムでは解決が困難だった大規模環境での性能問題に対処した．高速通信デバイスと不揮発性メモリの効率的な組み合わせにより，これらの課題を克服し，データ転送速度を飛躍的に向上させた．この進展は，大規模単一共有ファイルの利用がデータサイエンスにおける標準的なI/Oフレームワークとなる可能性を示唆している．

研究成果の学術的意義や社会的意義
学術的意義として，本研究は不揮発性メモリと高速通信デバイスの特性を活かし，
データサイエンスにおけるI/O課題を解決するストレージシステムの設計と最適化手法を明らかにした．これにより，高性能計算におけるストレージシステムの新たな標準を設定し，次世代の計算機設計への理解を深める貢献をしている．
社会的意義としては，データサイエンスアプリケーションのデータアクセスの高速化と効率化により，新たな科学的発見やビジネスインサイトの獲得が加速され，広範囲の分野にわたってイノベーションが促進されることが期待される．

研究成果の概要（英文）：In this research, we developed a storage system called "PEANUTS", which utilizes persistent memory for supercomputers supporting high-performance computing. PEANUTS primarily aims to enhance the parallel read and write operations of single shared files, crucial in data science, addressing performance issues in large-scale environments that conventional storage systems struggled to resolve. By efficiently combining high-speed communication devices with persistent memory, these challenges were overcome, and data transfer speeds were significantly improved. This advancement suggests the potential for the use of large single shared files to become a standard I/O framework in data science.

研究分野：高性能計算

キーワード：高性能計算　ストレージ　I/O　不揮発性メモリ　片側通信　RDMA　MPI-IO

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

蓄積されたデータから有益な知見の発見をするデータサイエンスでは，データを分析する学問的性質上ストレージ性能がボトルネックとなっている．

近年コモディティで利用可能になった新しいストレージデバイスに，不揮発性メモリがある．不揮発性メモリは，ハードディスクや SSD とは全くデバイス特性が異なる新しい高速ストレージデバイスであり，データサイエンスアプリケーションのストレージ性能を高速化するポテンシャルを持っており，その活用が期待される．

大規模なデータサイエンスでは，スーパーコンピュータ（スパコン）がよく利用されるが，スパコンの計算ノードに実装された不揮発性メモリを，データサイエンスアプリケーションから利用する方法や最適化手法は十分に明らかではない．特に，高性能計算分野で広く利用されている MPI-IO ライブラリや，MPI-IO を前提とする HDF5 や NetCDF を利用するアプリケーションのストレージ性能を高速化することが求められている．

２．研究の目的

本研究は，計算機科学の研究分野の見地から，スーパーコンピュータにおける不揮発性メモリのデバイス特性を生かした I/O 最適化手法を明らかにし，従来のハードディスクや SSD の利用を前提としていたストレージシステムの全体の設計を見直して，最適な設計を明らかにする．また，スーパーコンピュータの計算ノード上の不揮発性メモリをデータサイエンスアプリケーションから利用可能にして，ストレージ性能の向上を通じてデータサイエンス研究分野に貢献する．

目的達成のために，以下の３つの課題を設定した．

(1) 不揮発性メモリの特性を活かした I/O 高速化
これまでの研究により明らかになっている不揮発性メモリのデバイス特性を踏まえた上で，データサイエンスアプリケーションが要求するストレージシステムに適した最適化と設計を明らかにする．

(2) 計算ノード上の不揮発性メモリの有効活用
実際のスーパーコンピュータの計算機構成に即したストレージシステムの構成を明らかにする．

(3) データサイエンスアプリケーションとの統合，高速化
幅広いアプリケーションが対象となるように，データサイエンス分野で世界的に広く利用されている MPI-IO を対象とする．　　MPI-IO は，通常の POSIX I/O と違い，複数のプロセスから集団的にファイル I/O を行うための I/O ライブラリで，高性能計算分野におけるデファクトスタンダードである．

３．研究の方法

研究開始前の事前準備状況としては，不揮発性メモリの具体的な利用方法，I/O 最適化手法について調査を行った．既存のデータサイエンスアプリケーションがどのようなフレームワーク，ライブラリ，ストレージシステムを利用しているか，どのような I/O アクセスパターンなのかについて，調査を行った．

(1) 2022 年度計画：計算ノード上の不揮発性メモリを活用した書き込みの高速化

2022 年度は，それまでの調査を踏まえて，データサイエンスアプリケーションを高速化するためのストレージシステムを設計，構築する．2022 年度時点では MPI-IO ライブラリの不揮発性メモリ対応モジュールの構築を予定している．

(2) 2023 年度計画：不揮発性メモリ内のデータ参照による読み込み高速化

2023 年度は，不揮発性メモリ上のデータを直接参照することで，データ読み込みの高速化を行う．スーパーコンピュータで広く利用されている高速ネットワークである，InfiniBand の Remote Direct Memory Access 機能を用いて，リモートの計算機上の不揮発性メモリに読み書きする場合の性能評価を行い，期待する性能が得られるのか明らかにする．

(3)データサイエンスアプリケーションへの適用

2022 年度，2023 年度それぞれにおいて，提案ストレージシステムを用いてデータサイエンスアプリケーションが実際にどの程度高速化するか，評価を行い明らかにする．

４．研究成果

不揮発性メモリを搭載する新たなスパコン Pegasus が筑波大学で 2022 年末頃から稼働を開始し，利用可能となった一方で，Intel 社の Optane DC Persistent Memory 事業が 2022 年夏頃に中断となり，不揮発性メモリの今後のロードマップが不透明になった．今後は不揮発性メモリに変わり，Compute Express Link (CXL)がストレージデバイスの主流となり得る状況となった．CXL は，Persistent Memory と同様の API を利用可能で，シームレスな移行が期待できるため，本研究の成果や得られた知見は，CXL と共に今後公開されると考えられる新たなストレージデバ

イスへの適用を期待できる．

本研究では，(1) 計算ノード上の不揮発性メモリを MPI-IO を通じて利用可能にするストレージシステム「PEANUTS」の設計・開発・性能評価を主軸としつつ，(2)アドホックファイルシステムと既存のデータ管理ライブラリとの統合を通じた不揮発性メモリの活用についても研究を行った．

まず，Intel 社の最新かつ現時点で最後の不揮発性メモリとなった，第三世代 Intel Optane DCPMM を搭載するスーパーコンピュータ Pegasus における，不揮発性メモリの基礎的な性能評価を行い，デバイス特性を明らかにした．

次に，典型的な高性能計算アプリにおける，単一共有ファイルへの I/O 性能を評価するために，RDBench ベンチマークを開発した．

その後，(1)の研究成果として，PEANUTS (PErsistent memory And Network Unilateral Transfer System)を開発し，評価を行った．PEANUTS は MPI ランタイムとの統合により，MPI-IO を通じて計算ノード上の不揮発性メモリを透過的に利用可能とし，特にデータサイエンスアプリケーションで重要であるにもかかわらず従来の並列ファイルシステムが不得手としていた，単一共有ファイルへの並列 I/O の高速化を実現した．

Pegasus との統合評価により PEANUTS の有効性を検証した（図 2）．100 ノード，4800 プロセスからなる大規模環境での評価では，書き込み速度 2.47 TB/s，リモート読み取り速度 2.39 TB/s，ローカル読み取り速度 7.75 TB/s を記録し，ハードウェアの性能限界に近い結果を示した．これは既存研究の最先端システムと比較しても大幅な性能向上を達成している．さらに，PEANUTS の実アプリケーションへの適用評価も行い，データサイエンスアプリケーションを高速化できることを示した．

(2)の研究成果としては，既存のアドホックファイルシステムである CHFS の MPI-IO 統合システムを開発し評価を行った．従来の高性能計算アプリ以外のアプリケーションからのストレージ要求に対応するため，Apache Arrow，Tensor Store と CHFS の統合についても研究を行った．これらの研究成果については国際学会で発表を行い，オープンソースソフトウェアとしても公開を行っている．

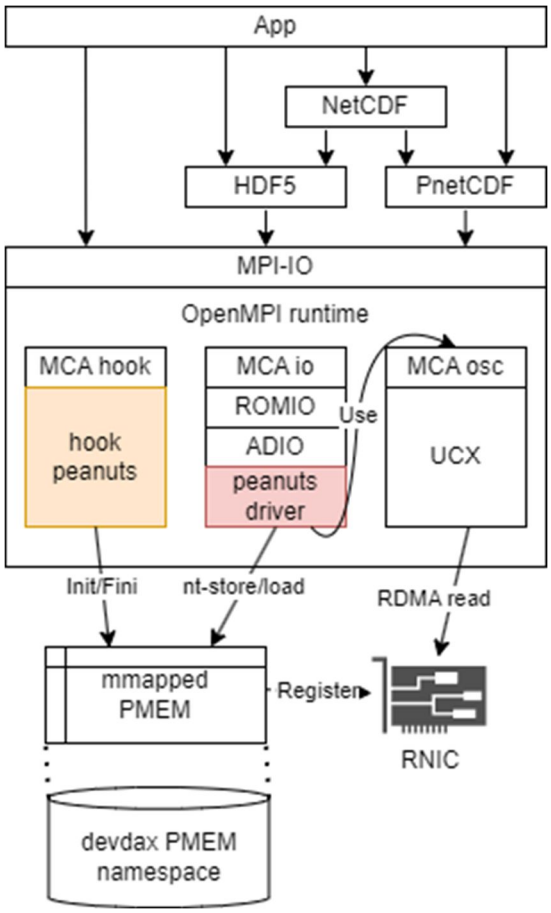

図 1 PEANUTS 概要図

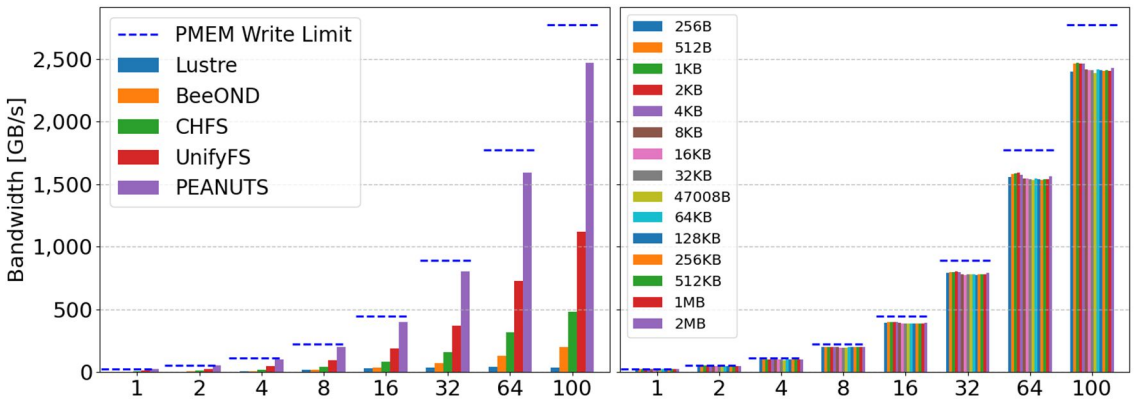

図 2　PEANUTS の単一共有ファイルへの並列書き込み性能と，既存の最先端システムとの比較．横軸はクライアントノード数を表す．左図はシステム間の比較．右図は PEANUTS のデータ転送サイズを変化させた場合の性能を表す．PEANUTS は既存システムと比較して大幅な性能向上を達成し，不揮発性メモリのハードウェア性能に近い性能が出ている

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Kohei Hiraga, Osamu Tatebe	４．巻 －
２．論文標題 PEANUTS: A Persistent Memory-Based Network Unilateral Transfer System for Enhanced MPI-IO Data Transfer	５．発行年 2024年
３．雑誌名 Euro-Par 2024: Parallel Processing - 30th International Conference on Parallel and Distributed Computing, Madrid, Spain, August 26-30, 2024, Proceedings	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Osamu Tatebe, Kohei Hiraga, Hiroki Ohtsuji	４．巻 1
２．論文標題 I/O-Aware Flushing for HPC Caching Filesystem	５．発行年 2023年
３．雑誌名 2023 IEEE International Conference on Cluster Computing Workshops (CLUSTER Workshops)	６．最初と最後の頁 11-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CLUSTERWorkshops61457.2023.00012	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計12件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 建部 修見, 平賀弘平, 前田 宗則, 藤田 典久, 小林 諒平, 額田 彰
２．発表標題 Pegasusビッグメモリスーパコンピュータの性能評価
３．学会等名 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 （HPC）研究発表会
４．発表年 2023年

１．発表者名 平賀弘平, 建部 修見
２．発表標題 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
３．学会等名 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 （HPC）研究発表会
４．発表年 2024年

１．発表者名
Kohei Hiraga, Osamu Tatebe

２．発表標題
PEANUTS: A Persistent Memory-Based Network Unilateral Transfer System for Enhanced MPI-IO Data Transfer

３．学会等名
Euro-Par 2024: 30th International European Conference on Parallel and Distributed Computing（国際学会）

４．発表年
2024年

１．発表者名
Sohei Koyama, Kohei Hiraga, Osamu Tatebe

２．発表標題
FINCHFS: Design of Ad-hoc File System for High-Performance Computing Workload

３．学会等名
22nd USENIX Conference on File and Storage Technologies (FAST 24)（国際学会）

４．発表年
2024年

１．発表者名
中野 将生, 平賀弘平, 建部 修見

２．発表標題
RustのUCXラッパーasync-ucxの性能評価

３．学会等名
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 （HPC）研究発表会

４．発表年
2023年

１．発表者名
小山 創平, 平賀弘平, 建部 修見

２．発表標題
FINCHFS: アドホック並列ファイルシステムの設計

３．学会等名
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 （HPC）研究発表会

４．発表年
2023年

１．発表者名
小山 創平，平賀弘平，建部 修見

２．発表標題
Apache Arrow CHFSによるビッグデータ処理のI/O高速化

３．学会等名
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 （HPC）研究発表会

４．発表年
2023年

１．発表者名
Osamu Tatebe, Kohei Hiraga, Hiroki Ohtsuji

２．発表標題
I/O-Aware Flushing for HPC Caching Filesystem

３．学会等名
4th Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO 24)（国際学会）

４．発表年
2024年

１．発表者名
Sohei Koyama, Kohei Hiraga, Osamu Tatebe

２．発表標題
Accelerating I/O in Distributed Data Processing Systems with Apache Arrow CHFS

３．学会等名
4th Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO 24)（国際学会）

４．発表年
2024年

１．発表者名
Sohei Koyama, Kohei Hiraga, Osamu Tatebe

２．発表標題
Fast Checkpointing of Large Language Models with TensorStore CHFS

３．学会等名
The International Conference for High Performance Computing Networking, Storage and Analysis (SC '23)（国際学会）

４．発表年
2023年

１．発表者名
巨畠 和樹，小山 創平，平賀 弘平，建部 修見

２．発表標題
HPC環境を想定した探索的データ解析におけるノードローカルストレージの利用の検討

３．学会等名
第１８５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 （HPC）研究発表会

４．発表年
2022年

１．発表者名
平賀 弘平，建部 修見

２．発表標題
MPI-IO/CHFS: ノードローカル不揮発性メモリを活用するアドホック分散ファイルシステムのためのMPI-IOの設計

３．学会等名
第１８５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 （HPC）研究発表会

４．発表年
2022年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

The core library of PEANUTS
https://github.com/tsukuba-hpcs/peanuts
PEANUTS integrated with OpenMPI
https://github.com/tsukuba-hpcs/ompi-peanuts
Artifact Evaluation Environment for PEANUTS
https://github.com/tsukuba-hpcs/peanuts-playground
K. Hiraga and O. Tatebe, “Artifact of the paper: PEANUTS: a persistent memory-based network unilateral transfer system for enhanced MPI-IO data transfer.” Zenodo, Jun. 2024. doi: 10.5281/zenodo.11558678.
本研究で開発したソフトウェアPEANUTSとそのOpenMPI統合，およびPEANUTSの論文中の評価を再現するための環境
rdbench
https://github.com/range3/rdbench
本研究で開発した典型的なHPCアプリによる単一共有ファイルへの書き込みを評価するための，リアルアプリケーションに基づくベンチマークソフトウェア

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関