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

平成　３０　年　　６　月　１８　日現在

機関番号：　８２６２６
研究種目：　挑戦的萌芽研究
研究期間：　2014～2017
課題番号：　２６５４００５０

研究課題名（和文）エクストリームコンピューティング向けた不揮発性メモリによるプログラム構成法

研究課題名（英文）Programming Model　for Non-Volatile Memory toward Extreme Computing

研究代表者
佐藤　仁（Sato, Hitoshi）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員

研究者番号：００５５０６３３

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：不揮発性メモリが登場し、スーパーコンピュータやクラウドへの搭載が進んでいるものの、その利用は従来のファイルと同様であり、不揮発性メモリデバイス本来の性能や可能性を活かしきれていない。そこで、不揮発性メモリをDRAMの拡張領域として扱うためのソフトウェア構成法を検討し、GPU向けMapReduceをはじめ、Sort,PrefixSum, Unique, SetIntersectionなどのビッグデータカーネルのOut-of-core化を進め、メモリ容量を超える大規模なデータセットに対しても高速な処理ができることを示した。

研究成果の概要（英文）：Emerging NVM (Non-Volitle Memory) devices such as Flash, which have positive aspects of inexpensive cost, high-energy-efficiency, and huge capacity compared with conventional DRAM devices, as well, as negative aspects of low throughput and latency, are widely employed to existing supercomputers and clouds. However, efficient implementation techniques and its productivity to overcome deepening memory hierarchy are open problems, although these NVMs will greatly expand the possibility of processing extremely large-scale datasets that exceed the DRAM capacity of the nodes. In order to address the issues, we investigated the programming model for NVM toward extreme data-intensive computing. Based on our GPU-based MapReduce implementation, we enhanced out-of-core features of the implementation, including various Big Data Kernels such as Sort, PrefixSum, Unique, SetIntersection, and demonstrated efficient performance to datasets that exceed the DRAM capacity of the nodes.

研究分野：　高性能計算

キーワード：　不揮発性メモリ　GPGPU　高性能計算　ビッグデータ

１．研究開始当初の背景

近年、フラッシュデバイスに代表される不揮発性メモリデバイスが登場し、モバイル PC だけでなく、スーパーコンピュータやクラウドデータセンターなどにも搭載されはじめている。例えば、東京工業大学の学術国際情報センターでは、2010 年 11 月よりスーパーコンピュータ「TSUBAME2.0」の運用を行っているが、実際に、各計算機ノード上に SSD が搭載され、総計 190TB、300GB/s の超高速な I/O を実現している。このような不揮発性メモリは、DRAM と比較するとアクセスレイテンシやスループットなどの性能面では劣るものの、容量あたりの価格や消費電力の点で優れており、今後は、従来の DRAM を補完していくものと考えられる。

一方、このような不揮発性メモリデバイス上のデータ管理は、従来手法では、ファイルの延長として扱われ、不揮発性メモリデバイス上にファイルシステムを構成しその上のファイルの操作として扱われてきた。しかし、近年、FusinoIO 社を中心としたオープンソースの OpenNVM プロジェクトや米国の SNIA(Storage Networking Industry Association) の NVM Programming Technical Work Group などでは、不揮発性メモリデバイスに対してフラッシュメモリを直接操作する API を定義し SDK を提供することで、不揮発性メモリに対するアクセスの更なる最適化による I/O 性能向上や、ファイルの形式を取らずに DRAM 上のデータを永続化できるという利点を活用した従来とは異なる不揮発性メモリの利用法、などの模索が盛んに行われている。一方で、このような不揮発性メモリに対するプログラミングの事例は、非常に新しい試みであるため、現状ではまだまだ事例が少なく、どのような I/O 性能の最適化が可能であるか、デバイスに対してどのようなインターフェースを定義すべきか、また、不揮発性メモリの性能を活かすためにどのようなライブラリを実現すべきか、など実際のソフトウェアの実装事例を基盤に解決すべき本質的研究課題が非常に多い。特に、数千～数万ノード、数万～数百万プロセスを必要とするアプリケーションから不揮発性メモリをどのように活用していくか、という点は未だ明らかでない。

２．研究の目的

ソフトウェアからの不揮発性メモリデバイス利用に関する要素技術の研究を推進し、将来のスーパーコンピュータやクラウドデータセンターへの適用可能な基盤技術のシーズが何であるかを明らかにすることを目的として、これまで我々が将来のエクサスケールスーパーコンピュータ想定し開発を進めてきた、GPU 向け MapReduce 処理フレームワークを基盤にし、1) 不揮発性メモリを DRAM の拡張領域として扱うためのデータ管理手法、性能最適化手法の検討、2)不揮発性メモリから GPU アクセラレータへの直接転送による性能最適化の検討、及び、GPU/CPU 上のメモリを超えるデータをホスト上の DRAM へオフロードするための手法の検討、3) MapReduce 処理の際の実行プロセスの永続化によるプロセスマイグレーション手法の検討、を行う。

３．研究の方法

3 カ年計画で、エクストリームコンピューティグに向けたソフトウェアからの不揮発性メモリデバイス利用に関する要素技術の確立を目指す。研究推進ためのソフトウェア基盤として、我々がこれまで開発を進めてきた GPU 向け MapReduce 処理フレームワークである “HAMAR”を利用する。その上で、初年度では、不揮発性メモリを DRAM の拡張領域として扱うための拡張方式の検討を行い、要素技術の基盤整備を行う。次年度では、不揮発性メモリと GPU アクセラレータを協調利用するための手法の検討を行う。最終年度は、これまでの研究成果を統合し、その上で、プロセスマイグレーションによるデータアクセスの高速化の検討を行う。研究成果は、単に学術会議での報告にとどまらず、オープンソースソフトウェアとして公開することを目指す。

４．研究成果

ソフトウェアからの不揮発性メモリ利用に関する要素技術として、不揮発性メモリを DRAM の拡張領域として扱うための技術を推進した。

具体的には、GPU アクセラレータと不揮発性メモリデバイスを搭載したスーパーコンピュータ向けの MapReduce 処理系 “HAMAR”を Out-of-core 処理へ拡張し、GPU に搭載されているデバイスの容量を超えるデータセットに対しても高速に処理が行えることを示した。我々が開発を進めている GPU 向け MapReduce 処理系を基盤に GIM-V(Generalized Iterative Matrix-Vector multiplication)アルゴリズムの実装と最適化を進め、TSUBAME2.5 の 1024 ノード（12288CPU コア, 3072 台の GPU）を用いて大規模な実証実験を行った結果、デバイスメモリの容量を超えるグラフデータ(171.8 億頂点、2749 億辺からなる大規模グラフ)を処理する際に 1 ノードあたり 3GPU を使用した場合、2.8 Giga Edges/sec(1 秒あたりに処理した辺数、47.7GB/sec)の性能になり、CPU 上での実行に対して 2.10 倍の高速化を確認した。また、ウィークスケーリングの性能を計測した結果、1024 ノード(3072 台の GPU)を使用した場合に 1 ノード(3 台の GPU)を使用した場合に対して、686 倍の性能向上を示し、良好なスケーラビリティを確認した。

また、MapReduce 処理系で性能律速になっていた Sort, PrefixSum, Unique, SetIntersection などの処理をビッグデータ

処理カーネルとして汎用化し Out-of-core 処理の実装を進めた。特に、GPU のデバイスメモリの容量を超える規模のデータに対しても高速処理が可能な大規模分散ソートの開発を進め、TSUBAME2.5 の 1024 ノードのうちの 2048 台の GPU を用いて、4TB の 64bit 整数をソートした結果、0.25TB/s のスループットが得られた。これは、CPU1 スレッドのみの実装と比べると 3.61 倍、CPU6 スレッド並列のものと比べると 1.40 倍の性能となっている。

さらに、大規模分散ソート中のローカルソートとして、GPU アクセラレータと不揮発性メモリを考慮した外部ソート xtr2sort(extreme external sort)を提案した。GPU の高い演算性能とメモリバンド幅を活かし、不揮発性メモリ、ホストメモリ、デバイスメモリ間のデータ移動に伴う遅延を隠蔽するために、不揮発性メモリ上のソートの対象となるレコードをデバイスメモリの収まるサイズへチャンク分割し、チャンク毎にパイプラインで不揮発性メモリへの I/O 操作、CPU-GPU 間のメモリ転送、GPU 上でのソート処理を非同期に行うことで、デバイスメモリやホストメモリの容量を超えたサイズのレコードに対しても高速に行う。提案手法を 2-way の Intel Xeon E5-2699 v3 2.30GHz(18 コア),NVIDIA Tesla K40 を搭載した 1 台のサーバ上で評価した結果、Linux Asynchronous I/O(libaio)を用いたノンブロッキング I/O による提案手法の実装において、CPU 上で実行可能なレコード数の 4 倍、GPU 上で実行可能なレコード数の 64 倍となる 25.6×10^9 の int64_t 型の整数値からなるレコードに対し、78,121,548 records/sec で動作し、2 ソケット 72 スレッドで動作させた CPU 版のノンブロッキング I/O による Out-of-core ソートとして 2.16 倍の性能を示すことを確認した。これらから、GPU アクセラレータを用いた Out-of-core な処理に向けて、不揮発性メモリを組み合わせ I/O のチャンク化と遅延隠蔽を行うことが良好であることを確認した。

これらの知見を基に、AI/ビッグデータ処理などへの応用を見越して，次世代の不揮発性メモリの主流と考えられている 3D XPoint メモリを対象に I/O ワークロードの詳細な解析を進め、有効性の確認を行った。ストレージ I/O (fio)、ストレージ I/O の遅延隠蔽(libaio)、メモリバンド幅 (STREAM)、演算性能 (GEMM)、ビッグデータ処理性能(Graph500) など不揮発性メモリのプログラム構成法の要素技術をベンチマークツール化し、これらを用いて、AI / ビッグデータ処理を模したワークロードを実行して性能評価を行い、DRAM メモリを超える規模のデータセットに対しても性能低下を抑えて透過的なメモリアクセスを提供できることを確認した。また、不揮発性メモリへのデータの永続化を前提としたプロセスマイグレーションに関して、Singularity などの HPC コンテナを対象とし要素技術の検討を行った。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 8 件）

① Hitoshi Sato, Ryo Mizote, Satoshi Matsuoka, Hirotaka Ogawa, “I/O Chunking and Latency Hiding Approach for Out-of-core Sorting Acceleration using GPU and Flash NVM”, 2016 IEEE International Conference on Big Data, 査読有, 398-403, 2016. DOI: 10.1109/BigData.2016.7840629

② Katsuki Fujisawa, Toyotaro Suzumura, Hitoshi Sato, Koji Ueno, Yuichiro Yasui, Keita Iwabuchi, Toshio Endo, “Advanced Computing and Optimization Infrastructure for Extremely Large-scale Graphs on Post Peta-scale Supercomputers”, Optimization in the Real World; Toward Solving Real-World Optimization Problems, 査読有, Vol.13, 1-13, 2016. DOI: 10.1007/978-4-431-55420-2

③ Hideyuki Shamoto, Koichi Shirahata, Aleksandr Drozd, Hitoshi Sato, Satoshi Matsuoka, “GPU-Accelerated Large-scale Distributed Sorting Coping with Device Memory Capacity”, IEEE Transactions on Big Data, 査読有, Vol.2, Issue 1, 57-69, 2016. DOI: 10.1109/TBDATA.2015.2511001

④ Keita Iwabuchi, Hitoshi Sato, Ryo Mizote, Yuichiro Yasui, Katsuki Fujisawa, Satoshi Matsuoka, “Hybrid BFS Approach Using Semi-External Memory”, 2014 IEEE International Parallel & Distributed Processing Symposium Workshops, 査読有, 1698-1707, 2014. DOI: 10.1109/IPDPSW.2014.189

⑤ Koichi Shirahata, Hitoshi Sato, Satoshi Matsuoka, “Out-of-core GPU Memory Management for MapReduce-based Large-scale Graph Processing”, 2014 IEEE Conference on Cluster Computing, 査読有, 221-229, 2014. DOI: 10.1109/CLUSTER.2014.6968748

⑥ Hideyuki Shamoto, Koichi Shirahata, Aleksandr Drozd, Hitoshi Sato, Satoshi Matsuoka, “Large-scale Distributed Sorting for GPU-based Heterogeneous Supercomputers”, 2014 IEEE Internatinal Conference on Big Data, 査読有, 510-518, 2014. DOI: 10.1109/BigData.2014.7004268

⑦ Keita Iwabuchi, Hitoshi Sato, Yuichiro Yasui, Katsuki Fujisawa, Satoshi Matsuoka, “NVM-based Hybrid BFS with Memory Efficient Data Structure”,

2014 IEEE International Conference on Big Data, 査読有, 529-538, 2014. DOI: 10.1109/BigData.2014.7004270

〔学会発表〕(計 16 件)

① 佐藤仁, 溝手竜, 小川宏高, “不揮発性メモリ 3D XPoint の AI/ビッグデータ処理への適用に向けた初期評価, 2018.
② Hitoshi Sato, “Building Software Ecosystems for AI Cloud using Singularity HPC Container”, 5th ADAC Workshop, 2018.
③ 佐藤仁, 小川宏高, “AI クラウドでの Linux コンテナ利用に向けた性能評価”, 第 162 回ハイパフォーマンスコンピューティング研究発表会, 2017.
④ 社本秀之, 佐藤仁, 松岡聡, “GPU アクセラレータと不揮発性メモリを考慮した大規模分散ソート”, 情報処理学会第 154 回ハイパフォーマンスコンピューティング研究発表会, 2016.
⑤ 佐藤仁, 溝手竜, 松岡聡, 小川宏高, “I/O 分割による遅延隠蔽を用いた Out-of-core な GPU Set Intersection の性能評価”, 情報処理学会第 155 回ハイパフォーマンスコンピューティング研究発表会, 2016.
⑥ 佐藤仁, 溝手竜, 松岡聡, “GPU アクセラレータと不揮発性メモリを考慮した外部ソート”, 情報処理学会第 150 回ハイパフォーマンスコンピューティング研究発表会, 2015.
⑦ Hitoshi Sato, Ryo Mizote, Satoshi Matsuoka, “Out-of-core Sorting Acceleration using GPU and Flash NVM”, The International Conference for High Performance Computing, Networking, Storage, and Analysis, 2015.
⑧ Satoshi Matsuoka, Hitoshi Sato, “Abstractions for Convergence of Big Data and HPC in Deep Memory Hierarchy Machines”, Workshop on Programming Abstractions for Data Locality,2014.
⑨ 佐藤仁, “不揮発性メモリを考慮した大規模グラフの高速処理”, メモリープラスワークショップ, 2014.
⑩ Hitoshi Sato, “Extreme Big Data(EBD) Next Generation Big Data Infrastructure Technologies Towards Yottabyte/Year”, 2014 ATIP Workshop: Japan Research Toward Next-Generation Extreme Computing in conjunction with SC14, 2014.

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況(計 0 件)
名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況(計 0 件)
名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
① 佐藤仁, “分散深層学習と I/O”, Gfarm Workshop 2018, 2018.
② 佐藤仁, “AI クラウドのソフトウェアエコシステム構築に向けた産総研の取り組み”, GTC Japan 2017, 2017.
③ Hitoshi Sato, Shuichi Ihara, Satoshi Matsuoka, “Reliability of NVM devices for I/O Acceleration on Supercomputing Systems”, Lustre User Group 2015,2015.
④ Hitoshi Sato, “Big Data Processing on GPU-based Supercomputers”, GPU Technology Conference GPU COE Achievement Award,2015.
⑤ 佐藤仁, “TSUBAME2 における GPU を用いた大規模グラフ処理“, GPU Technology Conference Japan 2014, 2014.
⑥ 佐藤仁, “Extreme Big Data: Convergence of Extreme Computing and Big Data Technologies”, Japan Lustre User Group 2014, 2014.

６．研究組織

(1)研究代表者
佐藤 仁 (SATO, Hitoshi)
産業技術総合研究所・情報・人間工学領域・主任研究員
研究者番号：00550633

(2)研究分担者
( )
研究者番号：

(3)連携研究者
( )
研究者番号：

(4)研究協力者
( )