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　４ 年　 ６ 月　 ９ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１９９１

研究課題名（和文）NVDIMM上の通信バッファによるスケーラブルな非同期通信レイヤの開発

研究課題名（英文）Development of Scalable Non-Blocking Communication Layer with NVDIMM as a Buffer

研究代表者
南里　豪志（NANRI, Takeshi）

九州大学・情報基盤研究開発センター・准教授

研究者番号：７０２８４５７８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、大規模並列計算環境において長メッセージの通信を隠蔽可能とするスケーラブルな非同期通信レイヤHyBuf-MPを実装し、その実用性を検証した。この通信レイヤの特徴は、通信隠蔽に必要なEagerプロトコルのバッファ領域として、DRAMだけでなくNVDIMMも利用可能とすることで、通信用のDRAMの消費を最小限に抑えつつ、長メッセージの通信隠蔽を実現できる点である。本研究の成果として、研究計画時の予想の通り、InfiniBandの通信遅延とNVDIMMのアクセス遅延が同程度であることから、これらを組み合わせた通信バッファが、DRAMの節約と通信隠蔽に効果的であることが確認できた。

研究成果の学術的意義や社会的意義
NVDIMMは、高速性と大容量を兼ね備えた新しい記憶デバイスであり、これを応用した新しいソフトウェア技術の開発が求められている。本研究では、従来DRAM上に実装していた通信バッファをこのNVDIMM上に構築し、十分実用的な性能が得られることを確認した。これにより、通信性能を落とすことなく、大規模並列計算で問題になっていた通信バッファによるDRAMの消費を大幅に削減できる。そのため、本研究の成果は、今後の計算科学やデータ科学におけるプログラムのスケーラビリティ向上に大きく貢献するものである。

研究成果の概要（英文）：In this study, HyBuf-MP, a scalable asynchronous communication layer has been implemented to enable long message communication overlapping in a massively parallel computing environment. The feature of this communication layer is that not only DRAM but also NVDIMMs can be used as buffer areas for the Eager protocol, which is necessary for communication overlapping, thereby minimizing the consumption of DRAM for communication and achieving communication overlapping of long messages.As a result of this research, since the communication delay of InfiniBand and the acceess delay of NVDIMM are comparable, it was confirmed that the communication buffer combining these two devices is effective in saving DRAM and hiding communication, as expected at the time of the research plan.

研究分野：高性能計算

キーワード：NVDIMM　通信　大規模並列計算

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景
近年、低コスト、省電力の主記憶装置として、DIMM スロットに装着可能な不揮発性メモリ装置 NVDIMM の開発が進んでいる。特に 2018 年には、揮発性メモリと不揮発性メモリの長所を合わせ持つ NVDIMM 規格 NVDIMM-P が公開された。そのため、高速小容量の DRAM と、低速大容量の NVDIMM によるハイブリッド構成の主記憶層は、アクセス速度と大容量化の両立を実現する将来の主記憶構成として期待されている。しかし、NVDIMM のアクセス遅延時間は DRAM の数百～数千倍であり、この領域を通常の演算に利用するのは非現実的である。そこで、NVDIMM の特性を活かした新しいソフトウェア技術の開発が求められていた。

２．研究の目的
本研究では、このハイブリッド構成の主記憶層の特性を活かした、ハイブリッドバッファ領域による通信ライブラリを提案する。通信ライブラリは、プロセス間のメッセージパッシングのために、内部でバッファ領域を確保する。このバッファ領域が、近年の計算機の大規模化に伴って増大し、アプリケーション用のメモリ領域を圧迫しつつあることが指摘されている。そこで、多くの容量が必要な全対全通信用のバッファ領域を、安価で低消費電力な NVDIMM の領域に配置することにより、DRAM のほとんどの領域をアプリケーション用に確保することが出来る。また、ネットワークの遅延時間が 1 μ秒程度であることから、NVDIMM のアクセス遅延時間は、通信用のバッファ領域としては十分実用的である。さらに、プログラム中で性能への影響が大きい通信のために、DRAM の一部の領域に高速バッファ領域を確保し、実行時に適切に使用バッファを選択することで、プログラムの性能を維持しつつ計算機の大規模化への対応が可能となるため、スケーラブルな通信ライブラリを実現できる。しかしながら、従来の通信ライブラリは単一の主記憶層を対象としており、このようなハイブリッド構成の主記憶層を対象としたものは存在しない。

３．研究の方法
まず、通信ライブラリの基盤となる、二つのバッファ領域を持つメッセージパッシング機構を実装する。このメッセージパッシング機構は、南里らが既に RDMA インタフェース上に実装済みのものをベースに構築する。片方のバッファ領域は、通常のメモリ領域として確保する。もう一方は、将来の NVDIMM 領域の仕様を想定し、libpmem ライブラリを用いて確保した領域を使用する。次に、実行中に任意のプロセス間で使用バッファを切り替えるプロトコルを設計し、実装する。切り替えの判断には受信側の高速バッファの空き状況の情報が必要であるため、このプロトコルは受信側が送信側に切り替え要求メッセージを出す形で設計する。さらに、プログラムの実行状況に応じて重要な通信相手に優先的に高速バッファを割り当てるバッファ選択機構を実装する。
完成した通信ライブラリは、NAS Parallel Benchmark により性能評価するとともに、オープンソースのライブラリとして GitHub で公開する。初年度、第二年度は、libpmem と numactl を用いて、NUMA 上の遠いメモリと近いメモリによる仮想的なハイブリッド主記憶層を構築して使用する。そのため、クラスタの各ノードは CPU を 2 ソケット搭載したものとする。第三年度には、その時点で導入可能な NVDIMM を各ノードの DIMM スロットに追加し、ハイブリッドな主記憶を構成する。今のところ、NVDIMM のメモリモジュールとしては、2018 年以降に発売予定となっている Intel 社の DIMM 版 Optane の購入を予定している。もし、本研究期間内での NVDIMM の購入が困難である場合、予算をノード増設に使用し、より大規模での実験を行う。HyBuf-MP の設計、実装および評価は、研究代表者の南里が担当する。一方、連携研究者の大江は、階層型ストレージシステムでの知見を活かし、不揮発性メモリの効率的な操作手段や、バッファ切り替えアルゴリズムの開発を支援する。

４．研究成果
本研究で構築する通信ライブラリは、一対一のメッセージキューイングシステムで構成される。入力側プロセスが enqueue コマンドでキューに投入したメッセージを、出力側プロセスが dequeue コマンドで取り出すことで、メッセージの転送が完了する。また、入力側プロセスは、flush コマンドにより、投入したメッセージが取り出されたことを確認することができる。
このキューでは、データ転送だけでなく、カウンタなどの管理情報の操作もすべて RDMA で実装する。これにより入力側は、キューが空いていれば、出力側のプロセスの状況によらずに enqueue 操作を完了することができる。RDMA を用いたメッセージキューイングシステムの一般的な実装として、キュー領域を出力側のプロセスのメモリ領域に配置する Push 型と、入力側のプロセスのメモリ領域に配置する Pull 型があげられる。多くのメッセージキューイングシステムでは Push 型が選択されている。これは、Push 型が enqueue 時にデータを出力側のメモリに RDMA write するため、dequeue 側で並行して計算を進めることで、通信時間の隠ぺいが期待できるためである。一方 Pull 型は、dequeue 時に入力側のメモリから RDMA read するため、通信時間を

隠蔽できない。しかし、Pmem では、RDMA write とローカルメモリへの書き込みが競合した場合の性能劣化の問題が報告されている。そのため、Pmem を用いたメッセージキューイングシステムの実装では、Push 型と Pull 型の優劣は、この書き込み競合による性能劣化の程度に依存する。そこで今回は、Push 型と Pull 型の両方を実装し、性能を比較した。このうち Pull 型による実装の概要を図１に示す。

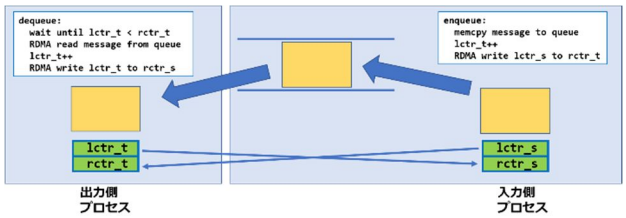

図１ Pull 型によるキューイングシステムの実装

図２および図３に、メッセージキューイングシステムの入力側プロセスと出力側プロセスの所要時間をそれぞれ示す。メッセージサイズが小さい場合、大きい場合とも、DRAM と Pmem の性能差は軽微である。また、メッセージサイズが小さい場合は当初の予想通り Push 型の方が高速である。一方、メッセージサイズ大きい場合は Pull 型の方が高速となる。これは、Push 型で dequeue 時に発生するキュー領域からのメモリコピーに要する時間が影響していると考えられる。

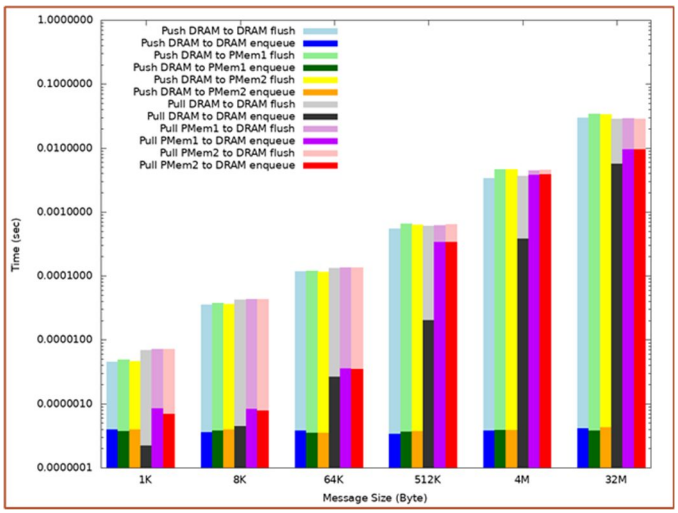

図２入力側プロセスの所要時間

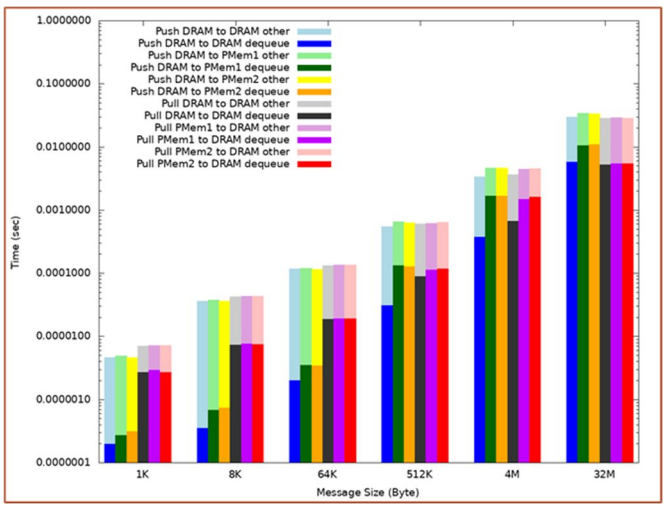

図３出力側プロセスの所要時間

これらの結果から、NVDIMM をメッセージキューとして使用する Eager Protocol の通信が、DRAM を使用する場合と同程度の性能であることが確認できた。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 南里 豪志
２．発表標題 DIMMスロット装着型不揮発性メモリ上のRDMAによるメッセージキューイングシステムの試作
３．学会等名 大学ICT推進協議会2020年度年次大会,
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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
連携研究者	大江　和一 (Oe Kazuichi) (80417451)	株式会社富士通研究所・その他部局等・研究員 (92707)	

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