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　５ 年　 ５ 月 ２９ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１９７８１

研究課題名（和文）P2P DMAを用いた高速ネットワークI/Oの研究

研究課題名（英文）Research on High-speed Packet I/O with P2P DMA

研究代表者
中村　遼（Nakamura, Ryo）

東京大学・情報基盤センター・准教授

研究者番号：９０８０４７８２

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では、コンピュータに接続されたデバイス同士が、CPUを介さずに直接データをやりとりするPeer-to-Peer Direct Memory Access (P2P DMA)を用いた高速なネットワークI/Oに関する研究を行なった。研究の中で、P2P DMAに関する計測ツールを実装し、P2P DMAを利用する利点を明らかにした上で、Network Interface Card同士が直接データをやりとりする新しい高速ソフトウェアルータを開発した。

研究成果の学術的意義や社会的意義
本研究は、コンピュータに搭載された複数のデバイスが連携する環境を想定し、デバイス間で直接データをやりとりするP2P DMAについて、その利点と応用についての研究を行なった。研究の結果、P2P DMAの有用なケースを明らかにし、応用事例としてCPUを利用するよりもさらに高速なソフトウェアルータの実現に至った。本成果は、計算での利用は当たり前となったGPUやアクセラレータをはじめとする周辺デバイスをより効率的に、幅広い用途へと利用していくための一助となるものである。

研究成果の概要（英文）：In this research, we have studied high-speed network I/O based on Peer-to-Peer Direct Memory Access (P2P DMA), which allows devices connected to a machine to exchange data directly without the CPU. During the research, we implemented a measurement tool for P2P DMA that clarified the advantages of using P2P DMA, and developed a new high-speed software router that leverages P2P DMA between Network Interface Cards.

研究分野：ネットワークシステム

キーワード：P2P DMA　Pcieデバイス　高速パケット転送　ネットワークI/O

１．研究開始当初の背景

コンピュータがシステムを構成する際、必ず周辺デバイスと協調して動作する。例えば Web サーバは、ストレージからファイルを読み出し、Network Interface Card (NIC)を介してクライアントへとファイルを送信する。ストレージや NIC に限らず、ムーアの法則で予想された CPU の性能向上が限界に近づく現在、GPU を始めとして、FPGA や Smart NIC、Tensor Processing Unit (TPU)など、PCIe で接続される周辺デバイスを用いて計算機の性能を向上する手法が広く研究されている。こうしたホストとデバイス間のデータ転送は、Direct Memory Access (DMA)と呼ばれる、メインメモリを介した通信で実現される。図 1 に示すように、例えばストレージ(NVMe SSD)は読み出すデータをメモリに書き込み(DMA Write)、NIC はメモリ上のデータを読み込んで(DMA Read)ネットワークへ送信する。

DMA はこのように、CPU がデバイスとデータをやりとりするために設計された通信方式である。そのため 2 つ以上のデバイス間でデータをやりとりするとき、下記に挙げる 2 つの問題が存在する。

- 通信がメインメモリを経由するために実効帯域上のオーバーヘッドが大きくなる。
- 通信がメインメモリを経由するために遅延が増加する。

DMA による通信は、PCIe リンク上では Transaction Layer Packet (TLP)と呼ばれるパケット通信によって実現される。そして、このパケットのヘッダが PCIe リンク上の実帯域におけるオーバーヘッドとなる。図 2 は、PCIe Gen 3 x8 リンク上で DMA を行なった際の実効帯域の理論値を示している。この図が示すように、2 つのデバイス間でメインメモリを経由してデータをやりとりする場合、メインメモリを経由した双方向の DMA (図 1 中 DMA Read/Write)が必要となり、単一方向の DMA (図 1 中 P2P DMA)に対して TLP ヘッダ分のオーバーヘッドが大きくなる。このオーバーヘッドに起因する実効帯域の減少は最大で 12Gbps にもなる。またメインメモリへ、そしてメインメモリから、デバイスへデータを転送させるための CPU 側の処理も必要になる。

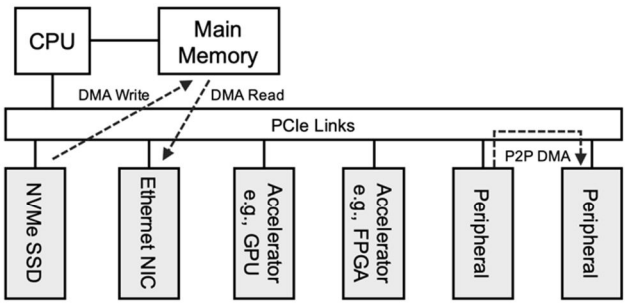


図1: 一般的なDMAとP2P DMA

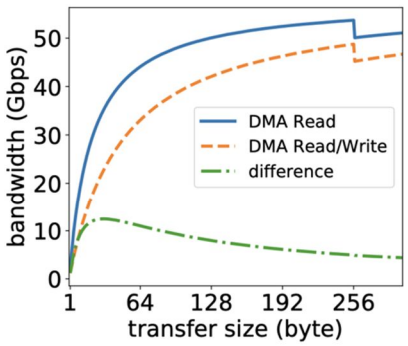


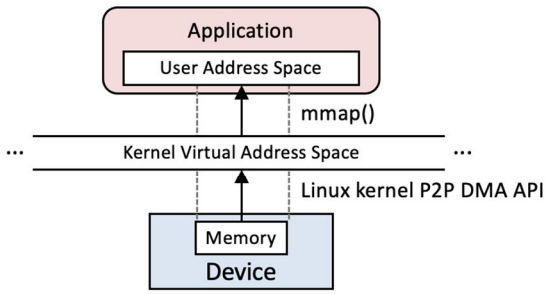
図2: 通常のDMAによるデバイス間通信とP2P DMAによる性能理論値の比較

２．研究の目的

本研究では、上記の問題を解決するため、Peer-to-Peer DMA (P2P DMA)を用いた新しいネットワーク I/O を研究開発する。P2P DMA とは図 1 中の右側に示すように、メインメモリを経由せずデバイス間で直接データを転送する DMA である。P2P DMA の通信には CPU が関与せず、データの移動にメインメモリを介す必要もないため、双方向の DMA によるオーバーヘッドとや CPU による調停処理やメインメモリの遅延も回避し、ハードウェアの性能を限界まで活かした高速な通信を実現できる可能性がある。

３．研究の方法

P2P DMA をネットワーク I/O に応用するにあたって、まずは P2P DMA を様々なデバイスで利用するためのライブラリとして libpop (Library for Peer-to-Peer DMA)開発し、libpop を用いた計測実験を通じて P2P DMA が有効なケースの確認を行なった。libpop は、図 3 に示すように P2P DMA の宛先となるデバイス上のメモリをオペーレーティングシステム(OS)の仮想メモリ空間にマッピングし、そのメモリを確保、解放するためのユーザランド API を提供する。libpop をデバイスごとのドライバやフレームワークに統合することで、元々P2P DMA に対応していないデバイスでも P2P DMA を利用することができるようになる。本研究では libpop を、Ethernet NIC 用のフレームワークである netmap、NVMe SSD 用のフレームワークである UNVMe、そしてベンチマークデバイスである pcie-bench の 3 つに統合し、今まで主に GPU と RDMA でしか利用されてこなかった P2P DMA を、Ethernet NIC と NVMe デバイスに適用し、その有効性を確認した。




```
pop_mem_t *pmem;
pop_buf_t *pbuf;
void *data;

pmem = pop_mem_init("1a:00.0", 4096);
pbuf = pop_buf_alloc(pmem, 2048);
data = pop_buf_data(pbuf);
```

図3: libpopのメモリマップの概要と、APIの使用例

次に P2P DMA を応用したアプリケーションとして、Ethernet NIC 間の P2P DMA を利用したソフトウェアルータの開発を行なった。libpop を用いた実験によって、汎用の x86 サーバで、デバイスの種別によらず P2P DMA が可能なことを確認できた。そこでコンピュータに接続された複数の NIC 同士が直接パケットをやりとりする、P2P DMA による高速ネットワーク I/O を根幹に据えたソフトウェアルータの開発を行なった。本ソフトウェアルータのアーキテクチャを図 4 に示す。従来のソフトウェアルータは、図 4 左に示す様に、CPU 上で動作するソフトウェアが、NIC からパケットをメインメモリに受信し、送信先の NIC がメインメモリからパケットを受け取って送信する。本研究では、この CPU を中心にパケットをやりとりする従来のアーキテクチャを CPU-driven アーキテクチャと呼ぶ。一方図 4 右に示す提案手法である P2PNIC アーキテクチャでは、NIC 同士が PCIe スイッチ越しに直接パケットを転送する。CPU-driven アーキテクチャと比較して、P2PNIC アーキテクチャではパケットが CPU もメインメモリも通らないため、PCIe のリンク速度がスループットの理論値となり、またパケット転送遅延も短くなることが期待される。

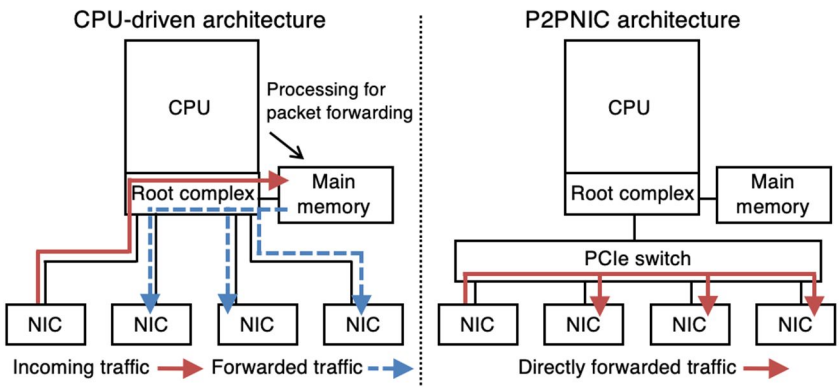


図4: 既存のソフトウェアルータのアーキテクチャ(CPU-driven architecture)と提案手法のアーキテクチャ(P2PNIC architecture)の比較

４．研究成果

libpop を用いた計測実験の成果のひとつとして、メインメモリへの書き込み負荷がデバイスからの DMA Write のスループットを最大 70%減少させることを明らかにした。図 5 は、ParaDNN を用いた機械学習モデルの訓練、メインメモリからの読み込み、メインメモリへの書き込み、という異なる負荷をかけながら、libpop を統合した pcie-bench を用いて、メインメモリへの DMA Write と、他のデバイスへの P2P DMA Write のスループットを計測した結果である。本グラフの示す通り、機械学習モデルの訓練は断続的にメインメモリへの DMA Write のスループットを低下させることがわかる。またとくに、メインメモリへの書き込みが DMA Write によるメインメモリの書き込みと重なると、DMA Write のスループットが最大で 70%ほど低くなることがわかった。一方で、P2P DMA Write はすべてのパターンで性能の低下は見られなかった。このように、メインメモリ負荷受けないという P2P DMA の利点を計測実験を通じて明らかにした。

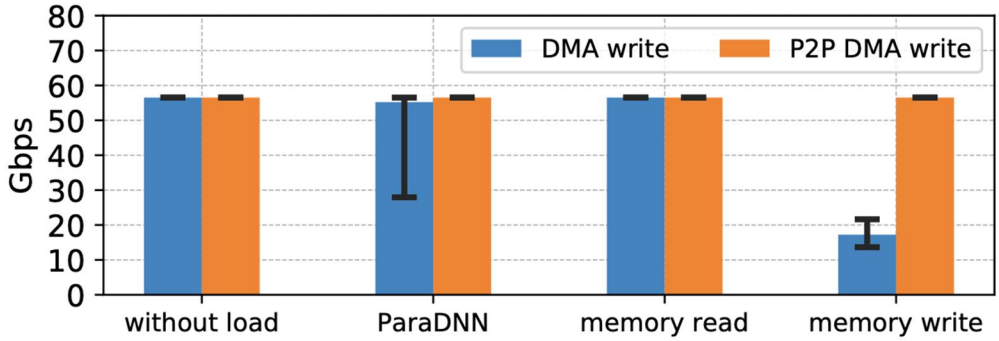


図5: pcie-benchで計測したDMA WriteとP2P DMA Writeのスループット

また図 5 の結果のほか、本研究の成果論文の中で、メインメモリの負荷がネットワークの送信性能には影響しないが受信性能には影響すること、パケット受信時のディスクリプタライトバックがその原因となっていること、ディスクリプタライトバックの影響はパケットのみを P2P DMA で転送するだけでは回避できないことなどが示されている。

P2P DMA によるネットワーク I/O を根幹に置いた P2PNIC アーキテクチャについて、Netronome の Smart NIC を用いて P2P DMA を行う新しい NIC を実装し、評価を行なった。図 6 は、実装したソフトウェアルータを用いて、 DMA Write による NIC 間パケット転送(P2PNIC-Wr)、DMA Read による NIC 間パケット転送(P2PNIC-Rd)、比較のために実装したメインメモリを経由する NIC 間パケット転送(P2PNIC-Bn)、そして既存の高速パケット I/O フレームワークである DPDK を用いた CPU-driven なパケット転送(L3FWD)のパケット転送性能を計測し、プロットしたものである。なお図 6 左は 1 ポートでトラフィックを受信し 1 ポートへ送信、図 6 右は 2 ポートから受信し 2 ポートへ送信した際の結果である。NIC 主導でパケットを転送する P2PNIC は、どの転送方式でも既存の CPU-driven アーキテクチャの L3FWD よりも高いスループットを達成した。とくに P2PNIC-Wr は PCIe の DMA Write が Posted Transaction であるという点もあり、高いスループットを実現している。

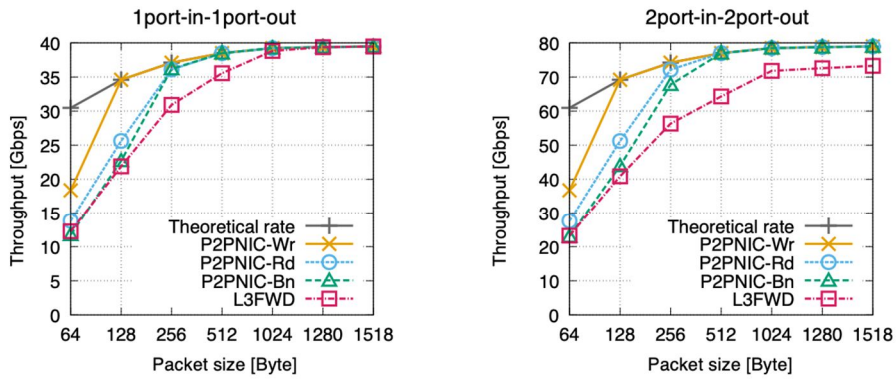


図6: Agillo CX 40Gbps NICで実装した提案手法によるソフトウェアルータのパケット転送性能

また本研究では、P2PNIC アーキテクチャをさらに推し進めた Pktpit アーキテクチャの開発も行なった。初期の P2PNIC アーキテクチャでは、NIC がパケットの宛先 IP アドレスに基づく送信ポートの決定を行い、当該ポートを持つ NIC へとパケットを P2P DMA で転送していた。一方、Ethernet NIC に搭載される高速なメモリのサイズはそれほど大きくはなく、例えばインターネットの全経路を NIC のメモリに展開することはできない。そこで Pktpit アーキテクチャでは、NIC は受信したパケットの宛先 IP アドレスのみを CPU へ送り、大量の経路から転送先を決定する処理を CPU で行い、パケットのデータは P2P DMA で送信ポートを持つ NIC へ送る。Pktpit も P2PNIC と同様に実装、評価され、CPU-driven な既存のソフトウェアルータ実装よりも高いスループット、低いパケット転送遅延を実現した。以上のように、本研究を通じて、P2P DMA の基本的な評価から、ソフトウェアルータをアプリケーションとして、データの転送はデバイスが主導し、またメモリと計算が必要な処理のみ CPU へオフロードする、デバイス主導のネットワーク I/O のひとつの形が実現された。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4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 Ueno Yukito、Nakamura Ryo、Kuga Yohei、Esaki Hiroshi	４．巻 0
２．論文標題 P2PNIC: High-Speed Packet Forwarding by Direct Communication between NICs	５．発行年 2021年
３．雑誌名 IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)	６．最初と最後の頁 1~6
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/INFOCOMWKSHPS51825.2021.9484641	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 Ueno Yukito、Nakamura Ryo、Kuga Yohei、Esaki Hiroshi	４．巻 30
２．論文標題 A NIC-driven Architecture for High-speed IP Packet Forwarding on General-purpose Servers	５．発行年 2022年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 226～237
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.30.226	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 Nakamura Ryo、Kuga Yohei、Akashi Kunio	４．巻 0
２．論文標題 How beneficial is peer-to-peer DMA?	５．発行年 2020年
３．雑誌名 APSys '20: Proceedings of the 11th ACM SIGOPS Asia-Pacific Workshop on Systems	６．最初と最後の頁 25 32
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3409963.3410491	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 Ueno Yukito、Nakamura Ryo、Kuga Yohei、Esaki Hiroshi	４．巻 0
２．論文標題 Pktpit: separating routing and packet transfer for fast and scalable software routers	５．発行年 2022年
３．雑誌名 SAC'22: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing	６．最初と最後の頁 943 951
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3477314.3507025	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0件

〔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

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