

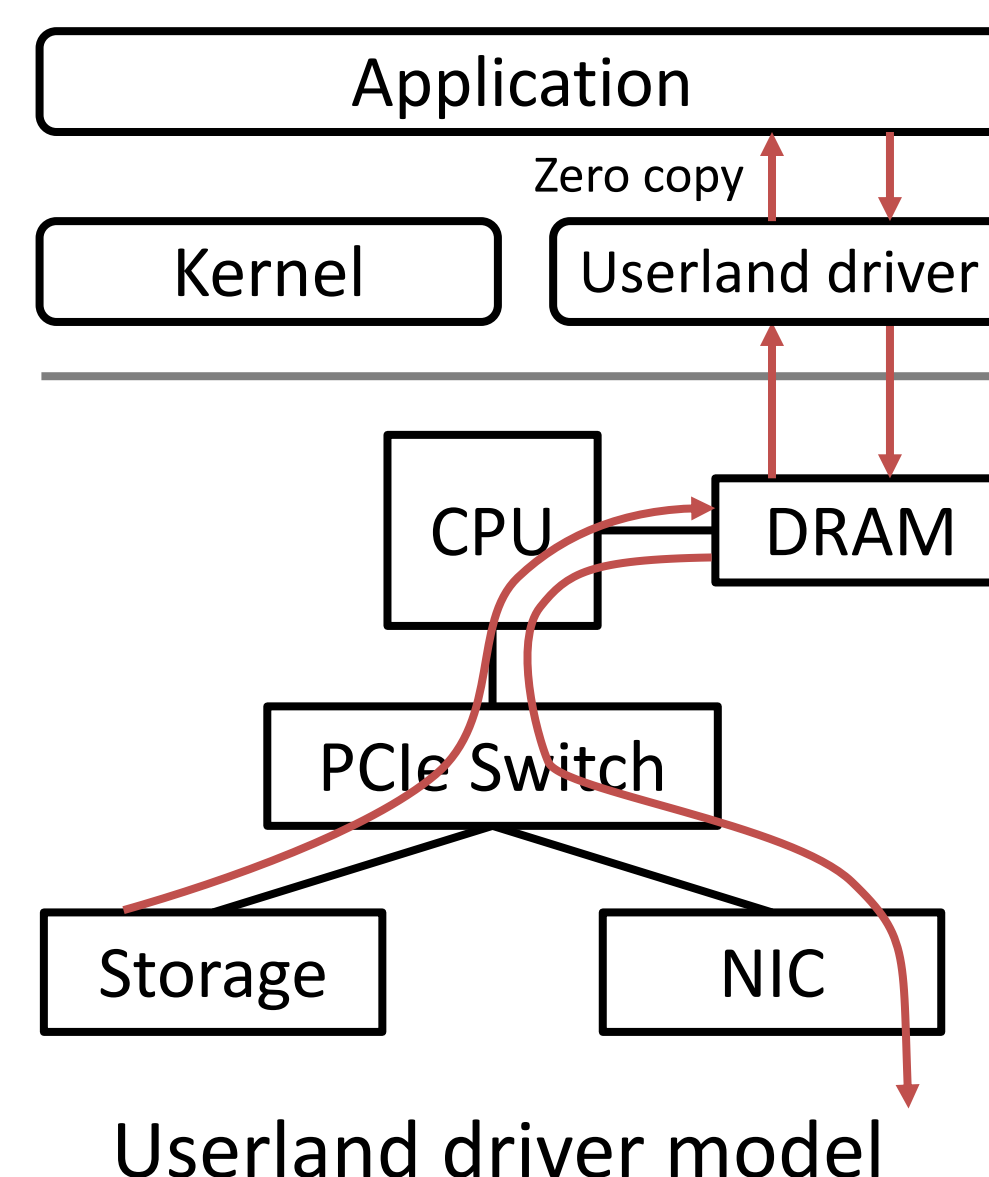
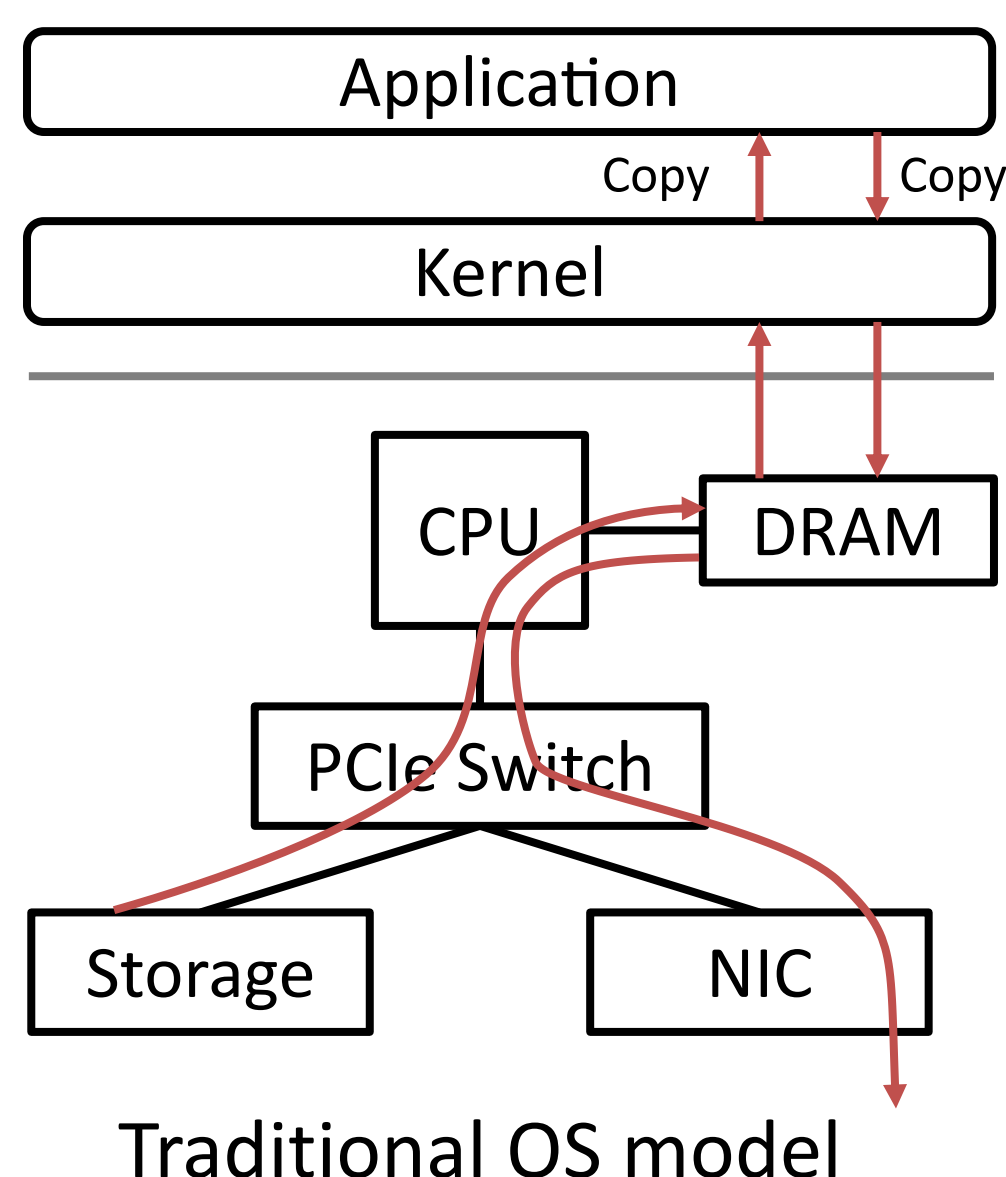
概要

本研究では、最新のNon-Volatile Memory Express (NVMe)の特徴を活かした、汎用マシンにおける新しい通信方式を提案する。本手法はNVMe のもつController Memory Buffer (CMB)を用いて、NVMe ストレージとNIC間でCPUやメインメモリを経由することなく直接PCIeバス越しにデータをやりとりする。これによって、汎用マシンのPCIeバスというハードの性能を限界まで使い切った、高速、大容量なデータ通信を実現する。本研究ではこの第一歩として、提案手法を用いた高速、大容量のトラフィックキャプチャおよびジェネレータを開発する。

現在のコンピュータにおけるPCIe通信とネットワーク高速化

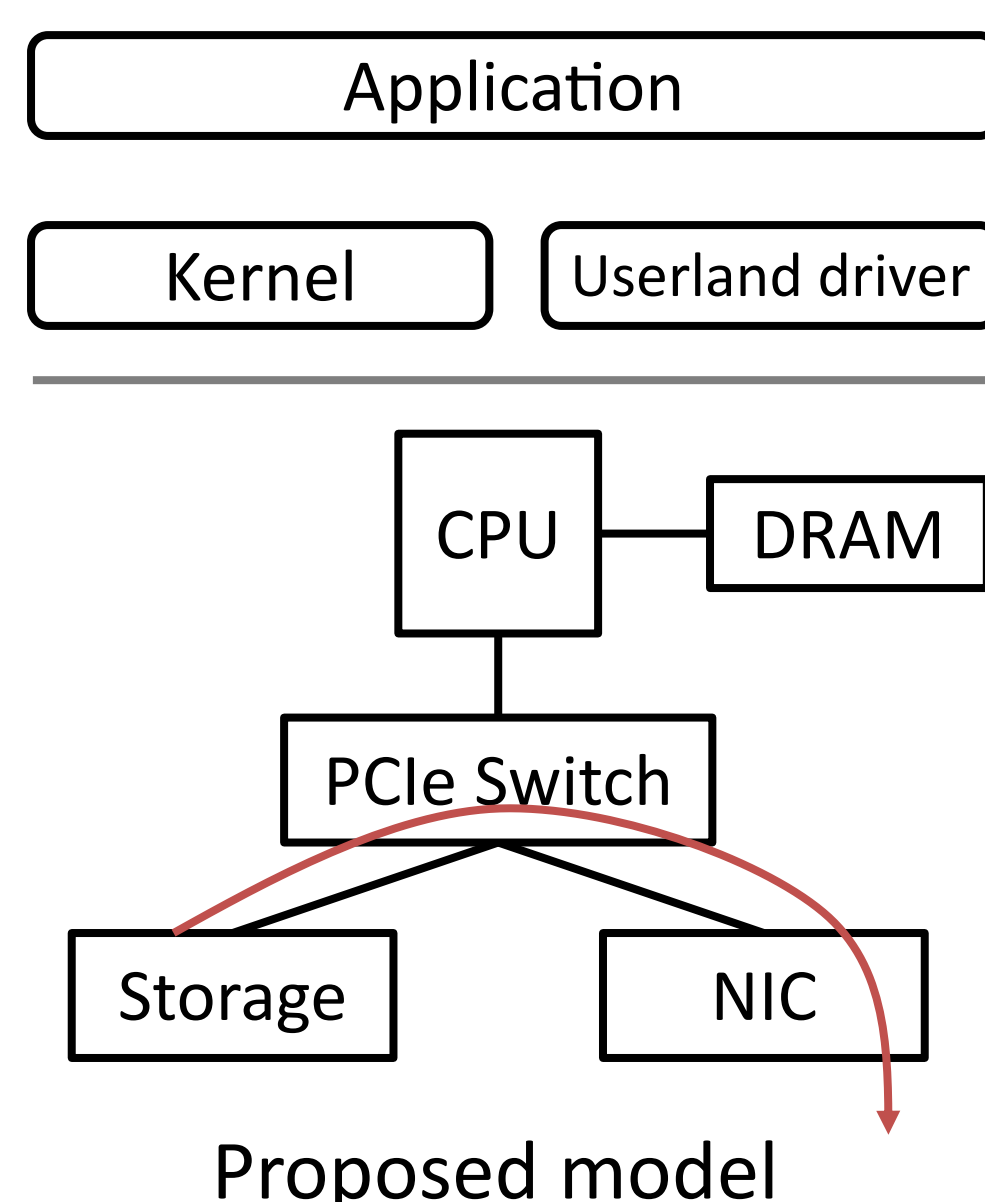
- DRAMへのDMAによるデータ転送
 - ストレージからの読み出し = DMA Write
 - NICへの送信 = DMA Read
 - 必ずDRAMを中継する
- 既存の高速化手法:
ユーザランドドライバを用いたコピーの削減
 - DPDK, netmap[1], uio_ring
 - SPDK, UNVME

一度の読み込みから書き出しに、データはDRAMを往復する必要がある



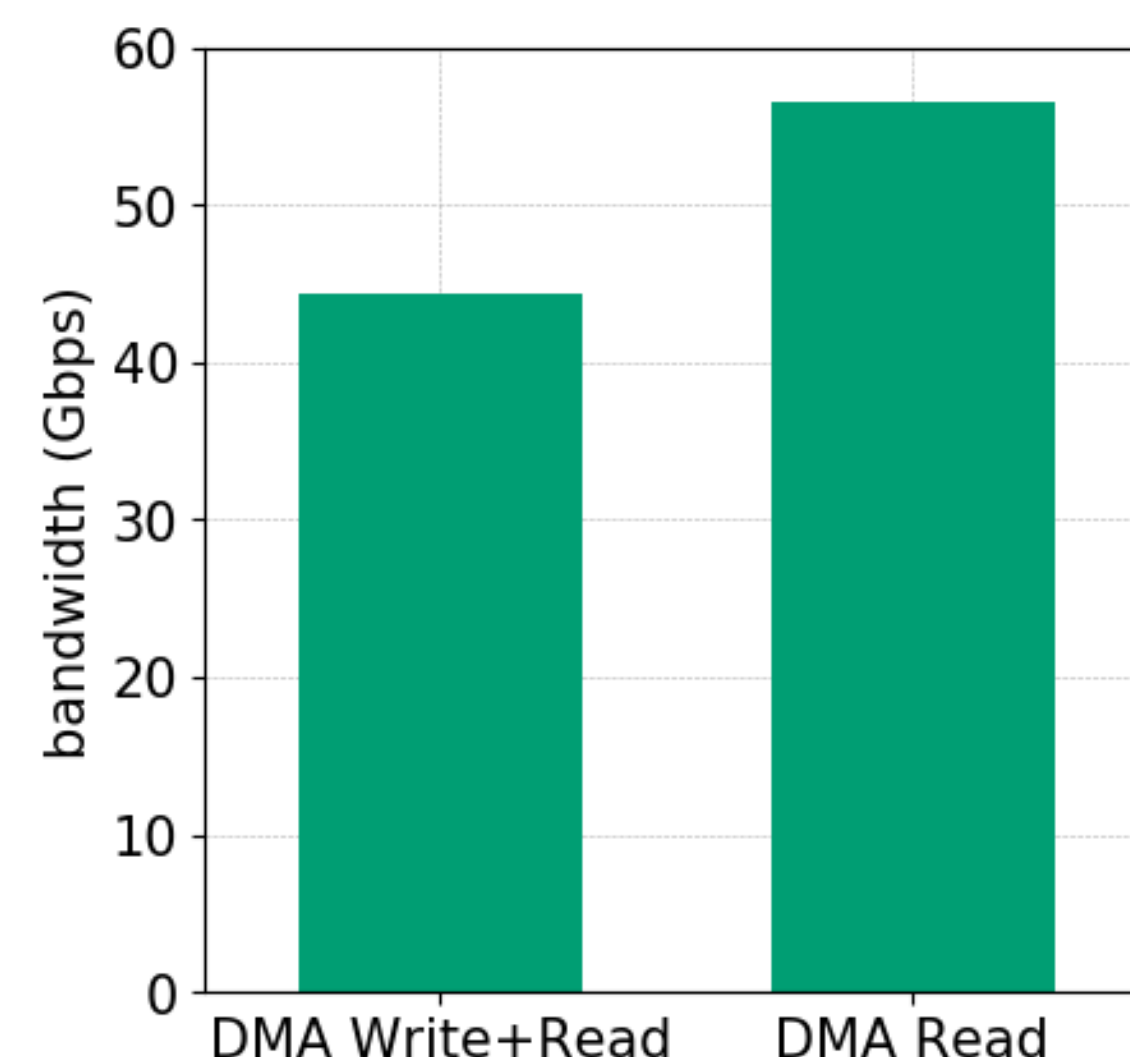
提案手法

- PCIeペリフェラル間の直接データ転送
 - CPUとDRAMを経由しない
=> DMAは一度だけ
 - PCIeバスの限界まで性能を発揮
 - ペリフェラル非依存
 - Not only RDMA
 - Not only GPU
- コンピュータ内のデータ通信と計算の分離
 - データ転送はPCIeバス上のみで行う
 - CPUは、データ転送先(address)や開始(doorbell)と終了(interrupt)のみをハンドル
- 必要な機器および実装
 - CMBサポートのNVMe
 - ペリフェラル上のメモリを管理するためのフレームワーク(実装中)



予備実験

- DMA Write + Read vs DMA Read
 - pcie-bench[2]で計測
 - 1M範囲の4kBをランダムに転送
 - DMA Readのみなら約1.2倍のスループット



[1] Luigi Rizzo, "netmap: A Novel Framework for Fast Packet I/O", USENIX ATC 2012

[2] Rolf Neugebauer, et al., "Understanding PCIe performance for end host networking", SIGCOMM'18