基于NVM和RDMA的分布式存储系统综述

汇报人：刘旭冉

目　　录

问　　题　　背　　景

NVM

非易失性存储器，或PM，持久内存

接近内存的访问速度

字节寻址的访问方式

磁盘一样提供持久性存储

问　　题　　背　　景

RDMA

远程直接内存访问(remote direct memory access，RDMA)技术正在大数据领域被越来越广泛地应用

支持在对方主机CPU不参与的情况下远程读写异地内存

提供高带宽、高吞吐和低延迟的数据传输特性

RDMA原语

双向原语

Send和Recv

在发送消息之前，接收方需提前调用Recv原语，用于指定接收消息的存放地址

单向原语

Read，Write以及相应的变种

能在远端CPU不介入的情况下直接读取或更新远端内存

RDMA网络数据收发过程

以远程写操作为例

本地CPU直接以MMIO方式向网卡发远程写命令，并传递相应参数

本地网卡收到命令之后，根据参数本地数据块基地址将数据块从主存以DMA Read的方式读取到网卡缓存，并发送到远端

远端网卡接收到数据块之后，以DMA Write的方式直接将数据写入内存对应地址

RDMA的优势

零拷贝

数据能够被直接发送到缓冲区或者能够直接从缓冲区里接收，而不需要被复制到网络层。

内核旁路

应用程序可以直接在用户态执行数据传输，不需要在内核态与用户态之间做上下文切换。

高带宽、低延迟

基于RDMA的分布式存储统将为满足大数据高时效处理和存储带来新的机遇

简单替换现有系统中的网络和存储模块？

直接用RDMA替换网络模块

软件逻辑冗余，多层数据拷贝将明显降低系统整体性能

RDMA内核旁路等特性未被充分利用

直接用NVM替换存储介质

替换内存（memory）：不保证数据可靠性

替换磁盘（storage）：软件开销过大，不能充分利用NVM性能

相关论文介绍

题目	出处
RDMP-KV: Designing Remote Direct Memory Persistence based Key-Value Stores with PMEM	SC 2020
Disaggregating Persistent Memory and Controlling Them Remotely: An Exploration of Passive Disaggregated Key-Value Stores	USENIX ATC 2020
TH-DPMS: Design and Implementation of an RDMA-enabled Distributed Persistent Memory Storage System	TOS 2020

RDMP-KV: Designing Remote Direct Memory Persistence based Key-Value Stores with PMEM

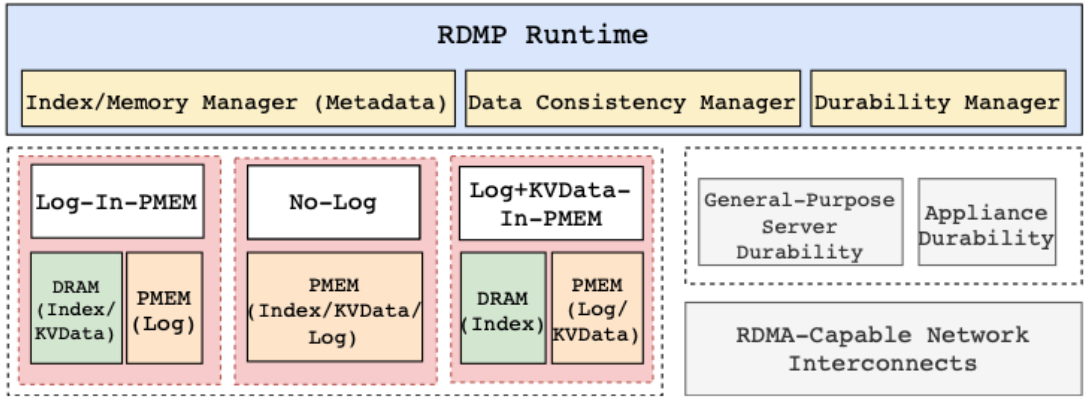


Fig. 3. Overview of RDMP-KV Runtime Architecture

动机

避免在RDMA communication buffers和PMEM之间拷贝数据(DRAM-to-PMEM staging)

贡献

提出一种针对带NVM的KV存储系统的RDMA通信架构Remote Direct Memory Persistence based Key-Value stores (RDMP-KV)，使客户端可以直接访问和持久化地更新服务端PMEM里存储的KV对。

RDMP-KV: Designing Remote Direct Memory Persistence based Key-Value Stores with PMEM

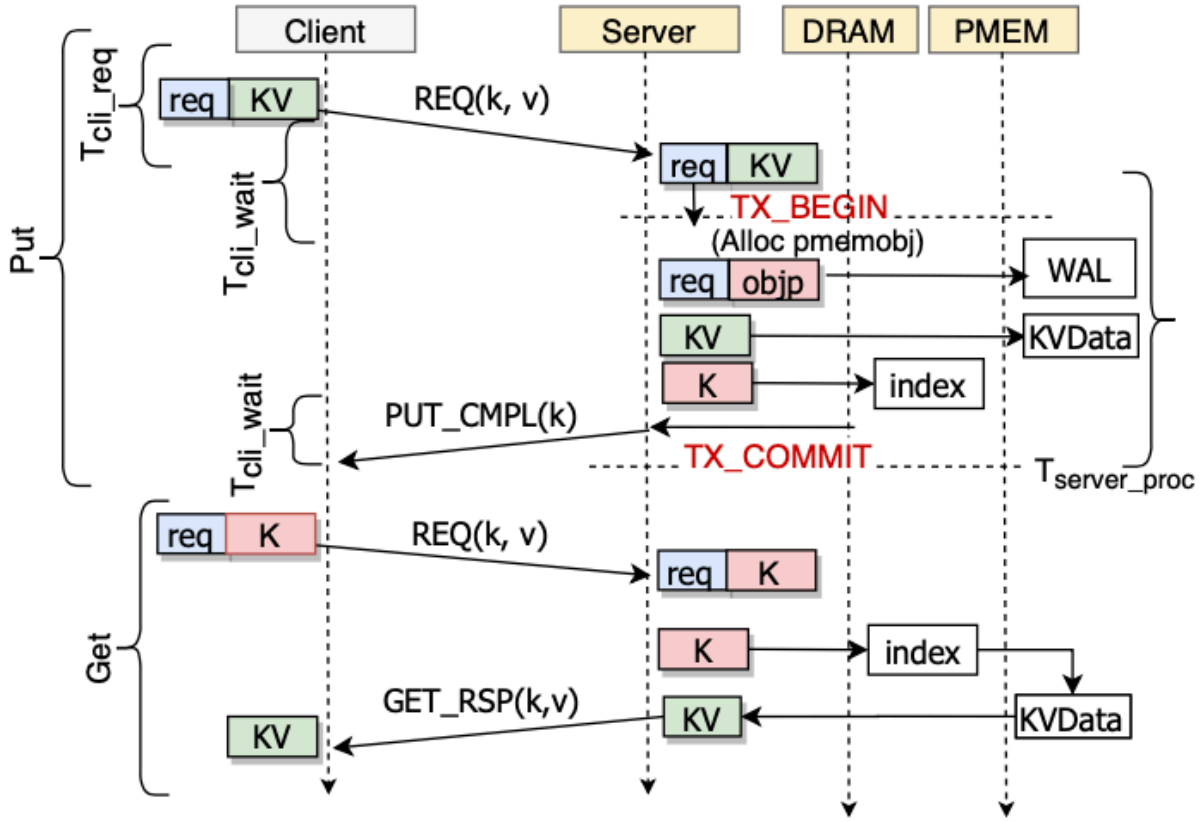


Fig. 1. PUT and GET in Pmem-Redis

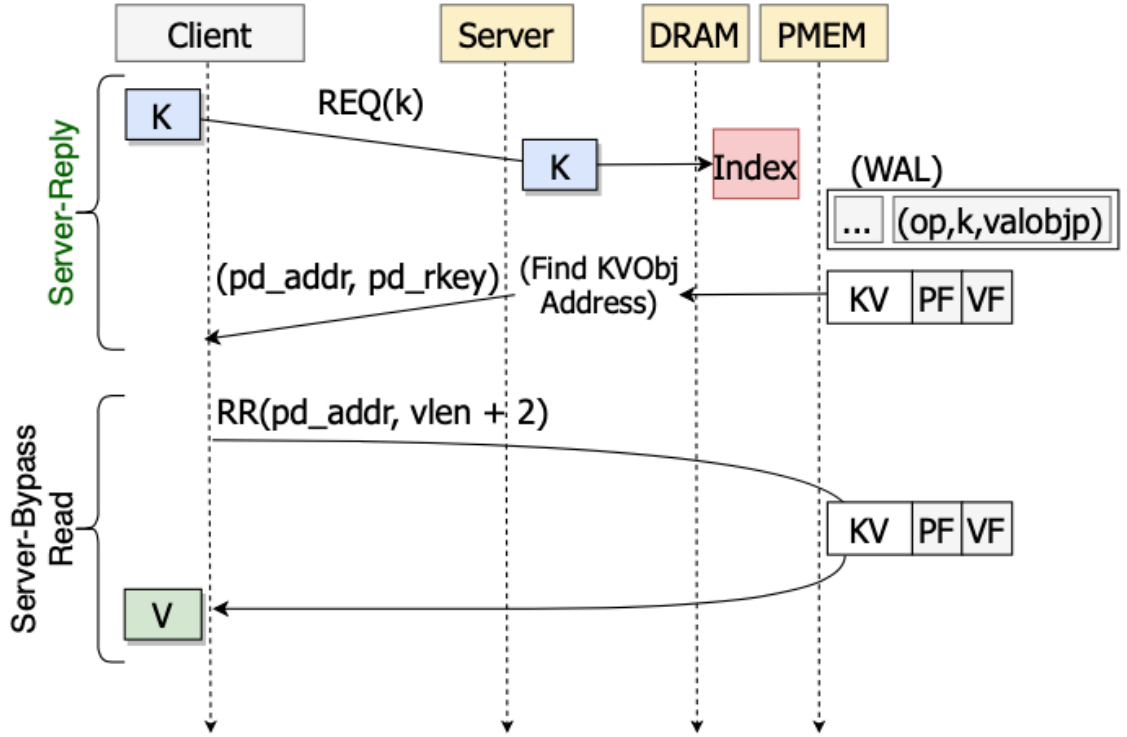


Fig. 4. RDMP-KV Protocol for GET ('RR' means RDMA-Read)

RDMP-KV: Designing Remote Direct Memory Persistence based Key-Value Stores with PMEM

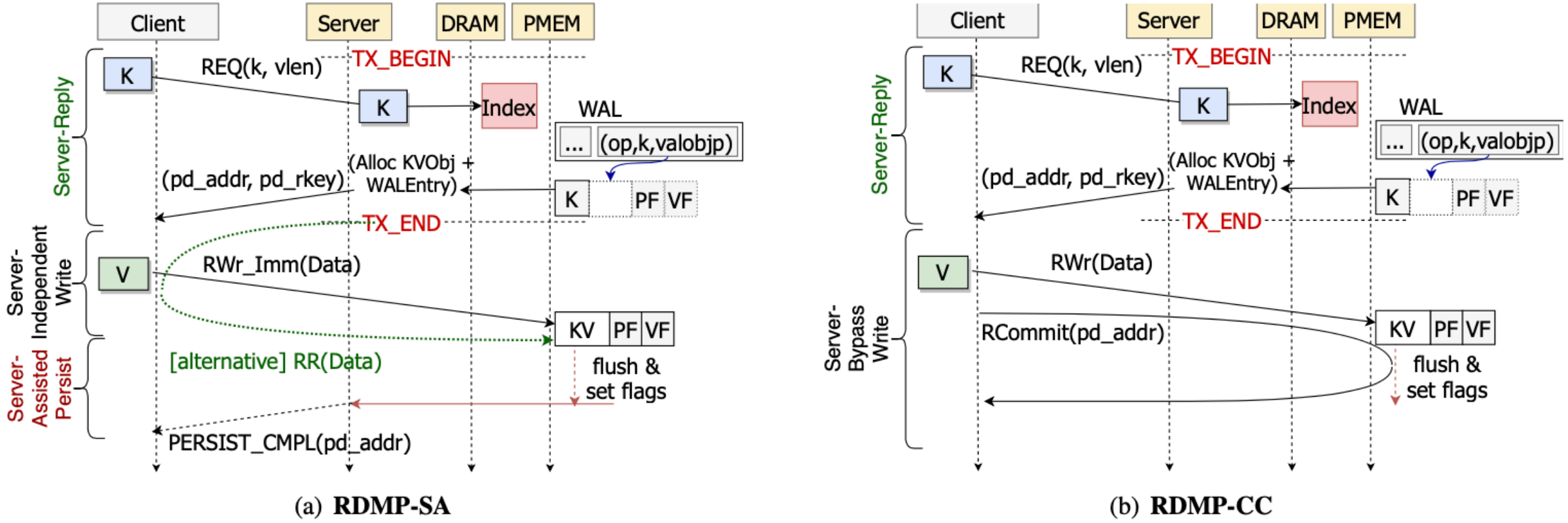


Fig. 5. RDMP-KV Protocols for PUT ('RWr': RDMA-Write, 'RWr_Imm': RDMA-Write-with-Immediate)

RDMP-KV: Designing Remote Direct Memory Persistence based Key-Value Stores with PMEM

测试结果

优于不针对PMEM优化的通信机制如HERD

优于其他RDMA-over-PMEM架构如Octopus和Forca

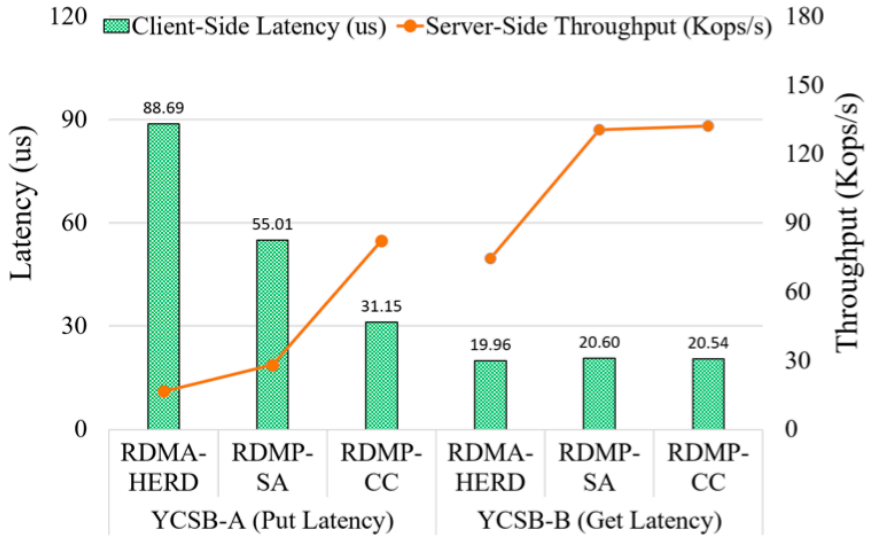


Fig. 9. Performance with YCSB Workloads A (50:50 read:write) and B (95:5 read:write). Contrasting RDMP-KV vs. RDMA-HERD over Pmem-Redis running 96 Clients over 24 server instances

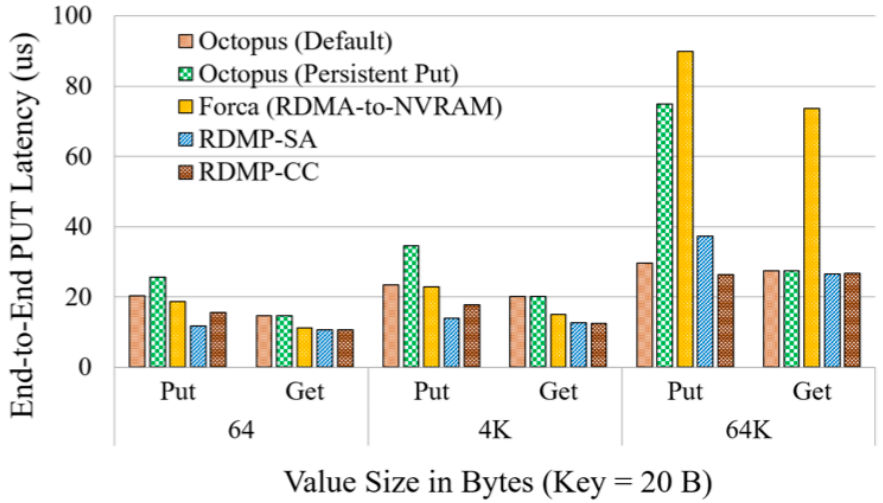


Fig. 10. Comparison with state-of-the-art RDMA-enable persistent memory systems. Contrasting end-to-end PUT/GET latency RDMP-KV with Octopus FS [22] and Forca [14]

Disaggregating Persistent Memory and Controlling Them Remotely: An Exploration of Passive Disaggregated Key-Value Stores

动机

存算分离，有助于提升可扩展性和资源利用率，同时降低成本。

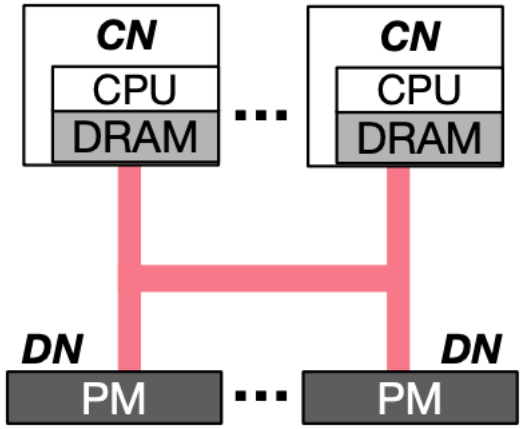
贡献

提出pDPM（passive disaggregated persistent memory）模型：PM资源配置在单独的数据节点上，计算节点通过RDMA远程管理PM。

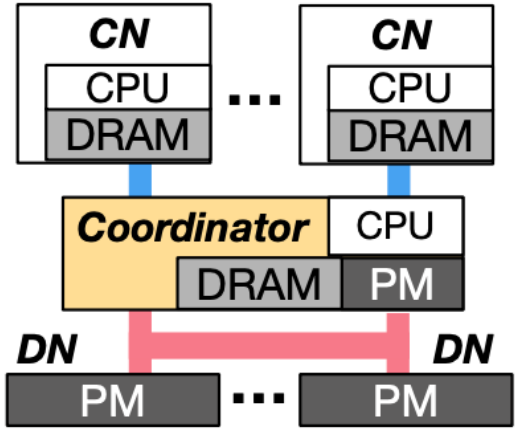
设计实现了三种pDPM架构的KV存储系统。

Disaggregating Persistent Memory and Controlling Them Remotely: An Exploration of Passive Disaggregated Key-Value Stores

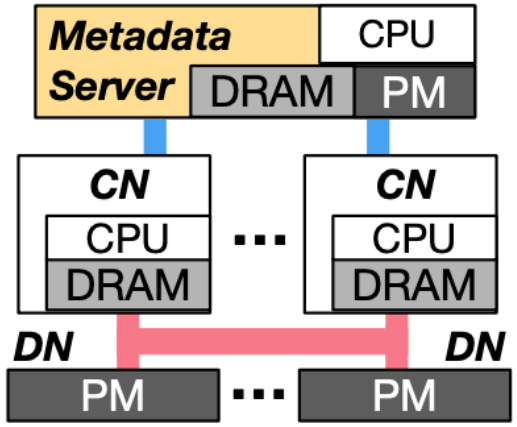
三种pDPM系统：在何处管理数据？



pDPM-Direct
CN管理数据



pDPM-Central
Coordinator管理数据



Clover
混合方法

CN：计算节点；DN：数据节点；红色连线：单边RDMA；蓝色连线：双边RDMA

Disaggregating Persistent Memory and Controlling Them Remotely: An Exploration of Passive Disaggregated Key-Value Stores

测试结果

优于非分离式系统

成本大幅降低的同时性能接近aDPM系统

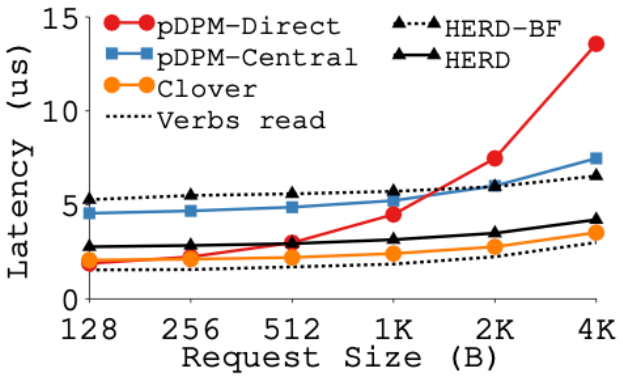


Figure 5: **Read Latency.**

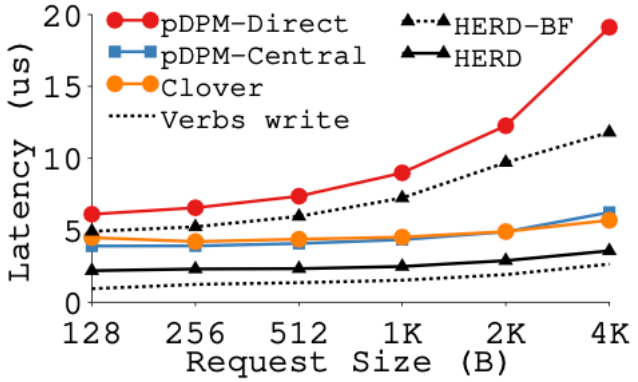


Figure 6: **Write Latency.**

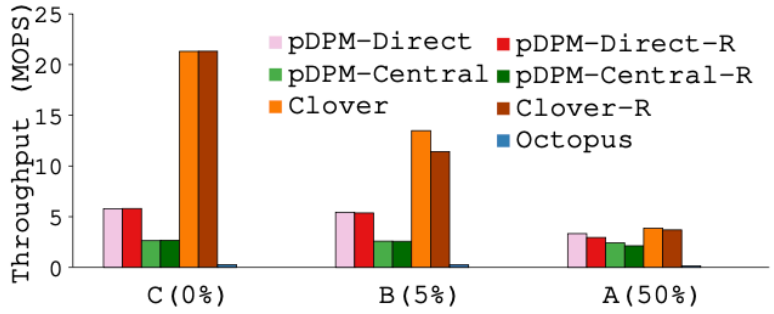


Figure 7: **Throughput Comparison with YCSB.**
Running YCSB on four CNs and four DNs.

TH-DPMS: Design and Implementation of an RDMA-enabled Distributed Persistent Memory Storage System

动机

数据中心中的文件系统和KV存储等不同接口分别实现了相似的功能（空间分配、崩溃一致性等）可以将其统一协调。

利用RDMA和NVM的优秀硬件性能要求更快的软件设计。

贡献

设计了管理分布式PM的抽象层pDSM (persistent distributed shared memory)，并在其基础上实现了文件系统和KV存储的接口，整个系统称为TH-DPMS (TsingHua Distributed Persistent Memory System).

TH-DPMS: Design and Implementation of an RDMA-enabled Distributed Persistent Memory Storage System

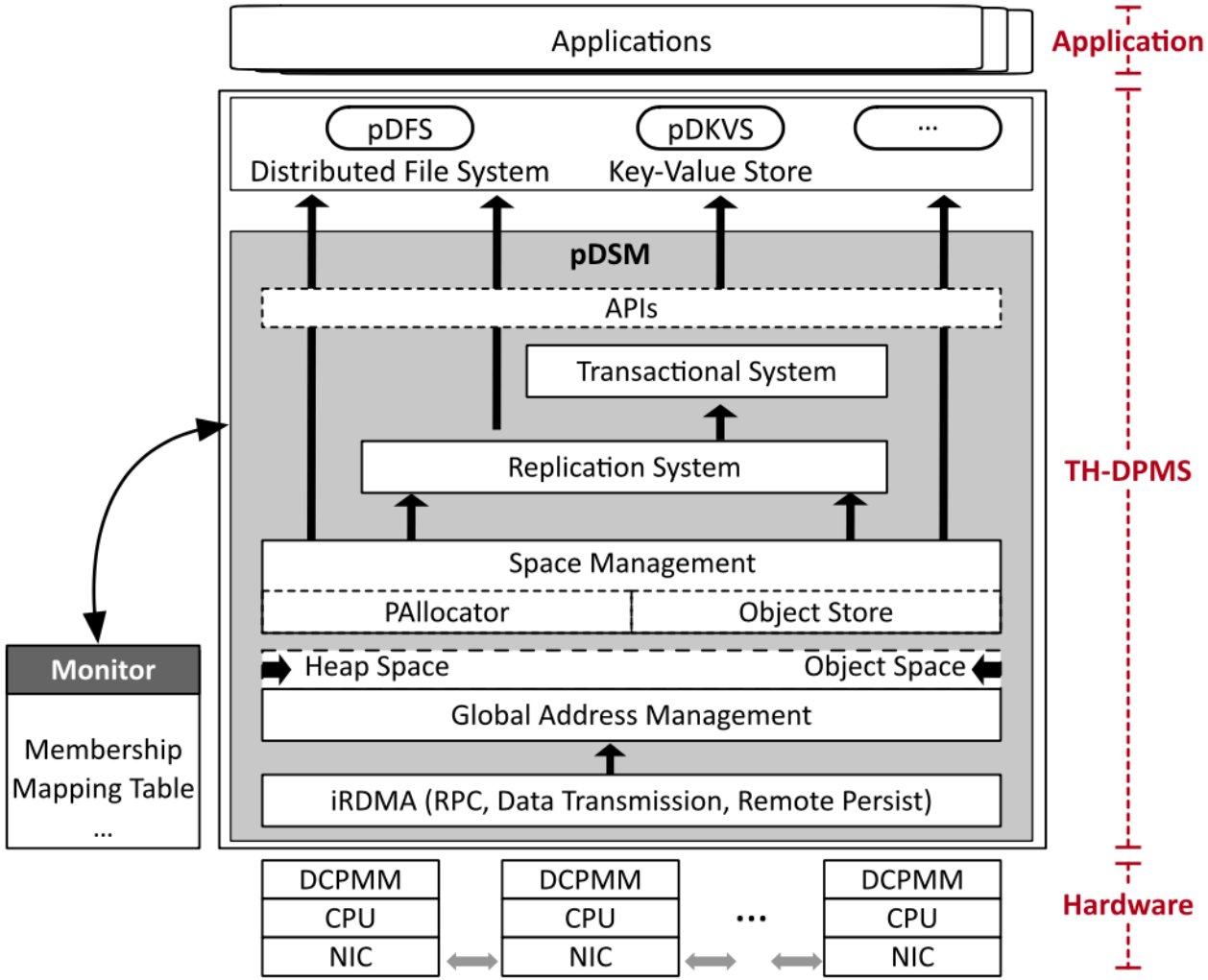


Fig. 2. Software architecture of TH-DPMS.

TH-DPMS: Design and Implementation of an RDMA-enabled Distributed Persistent Memory Storage System

测试结果

接近甚至优于数据全部保存在DRAM中的系统

在较大数据规模下表现优秀

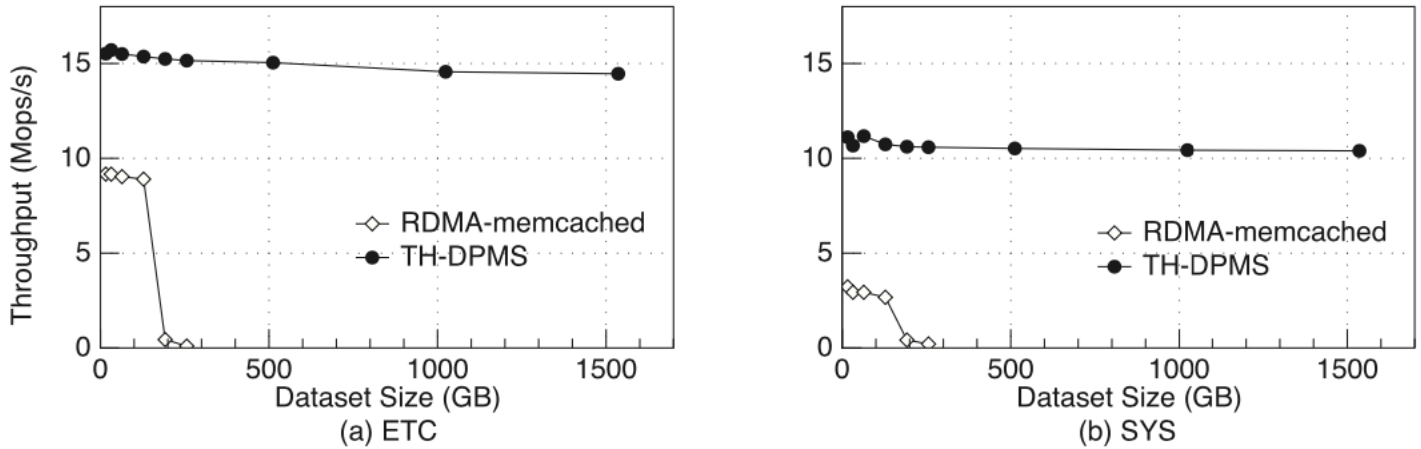


Fig. 9. The performance of pDKVS with varying workload size.

THANKS

汇报人：刘旭冉