Challenges and Solutions for Fast Remote Persistent Memory Access

Anuj Kalia (Microsoft Research)
Michael Kaminsky (BrdgAI, CMU)
David G. Andersen (BrdgAI, CMU)

We finally have fast durable storage

	Datacenter network	Solid state drivers	Persistent memory (NVMM)
Latency (μs)	2 μs	10 μs	100 ns
Bandwidth (Gbps)	100 Gbps	20 Gbps	100+ Gbps

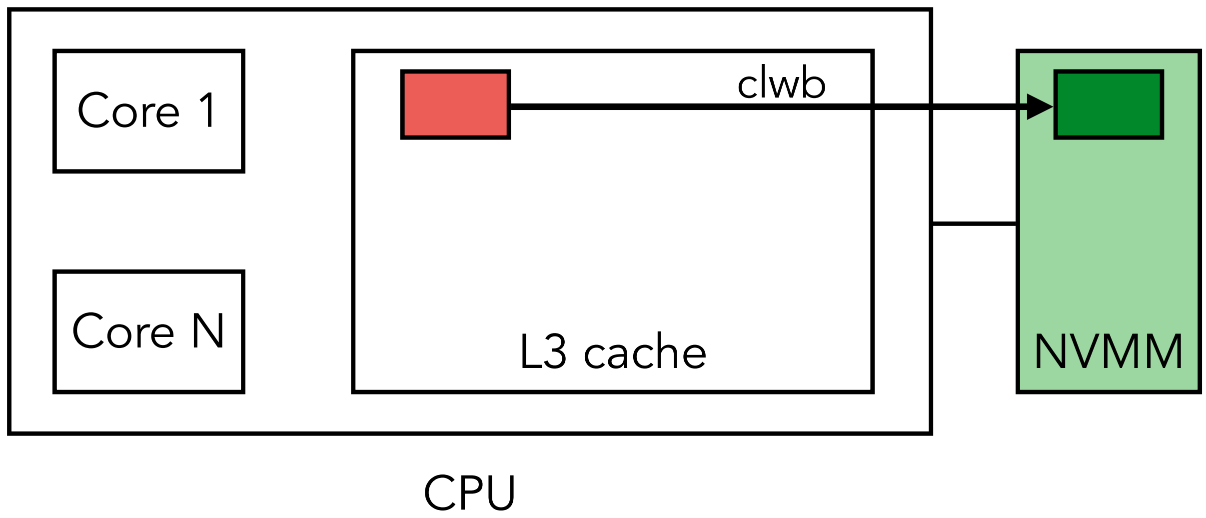
How should we build distributed systems for NVMM?

Recent DRAM-based systems for fast networks provide a blueprint

- Key-value stores: Pilaf [ATC 13], MICA [NSDI 2014], HERD [SIGCOMM 2014], ...
- Transaction processing systems: FaRM [NSDI 14, SOSP 15], DrTM [SOSP 15, OSDI 18], FaSST [OSDI 16], NAM-DB [VLDB 17], ...
- State machine replication: DARE [HPDC 2015], Zookeeper-in-a-box [NSDI 2016], ...

What design decisions need to change if we use NVMM instead of DRAM?

The power-safe domain in NVMM systems



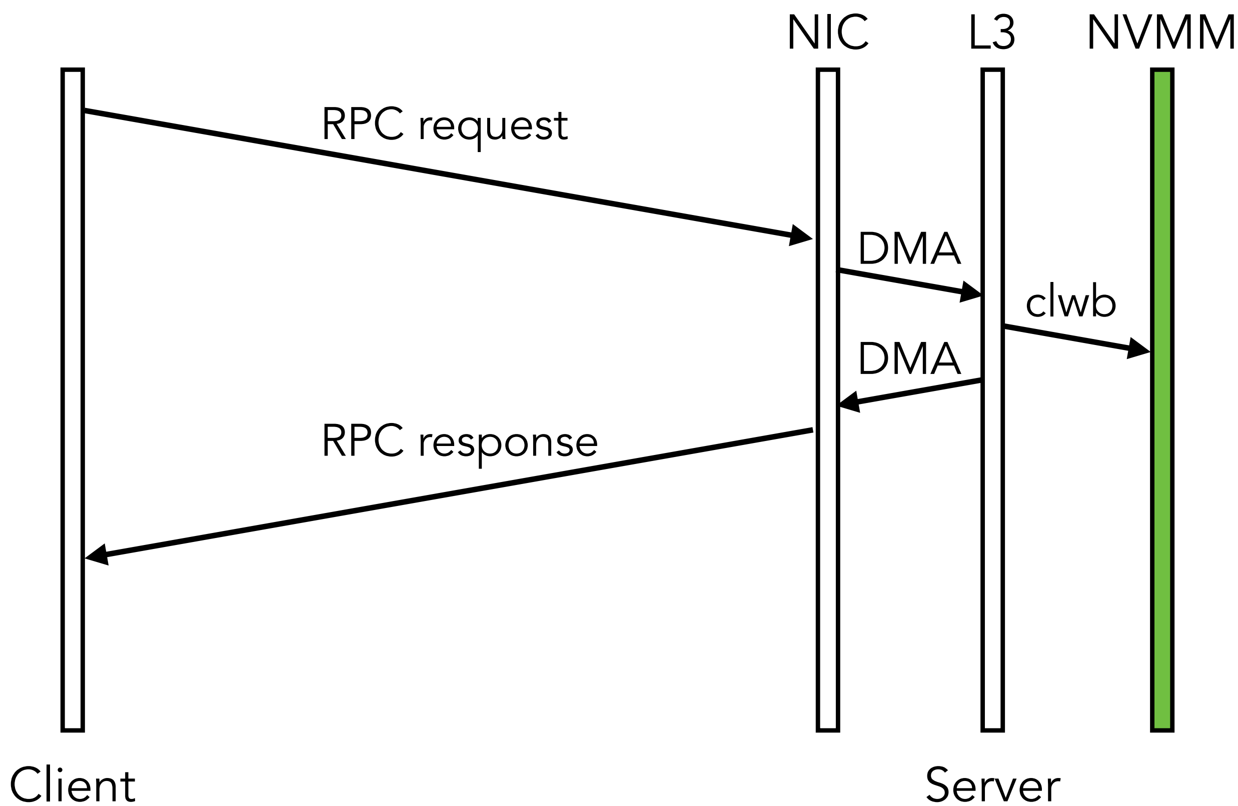
Latency of Remote Persistent Memory Access

Two approaches:

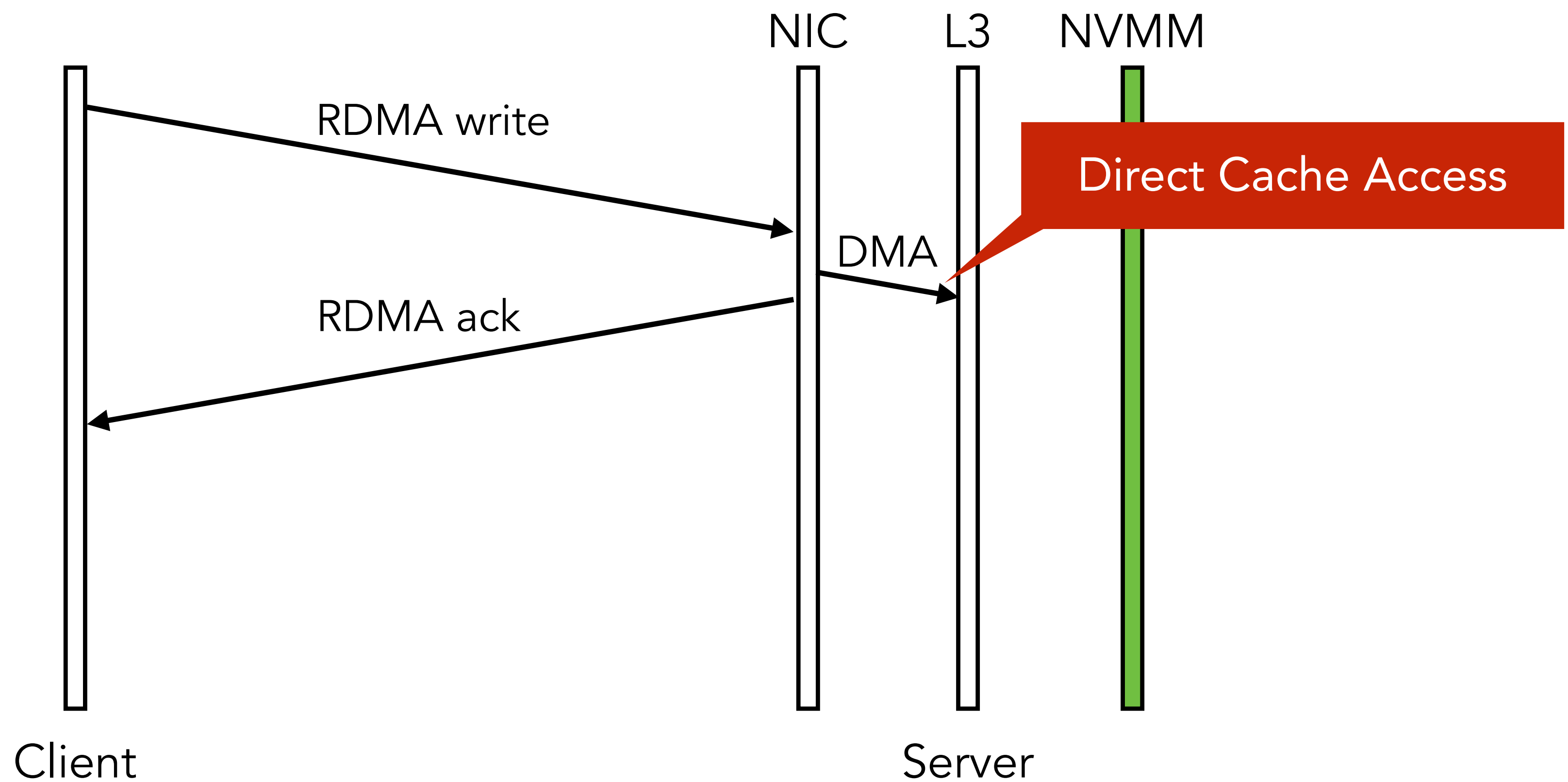
- Remote Procedure Calls (RPCs)
- Remote Direct Memory Access (RDMA)

Finding: RDMA has similar as RPCs for durable writes to remote NVMM

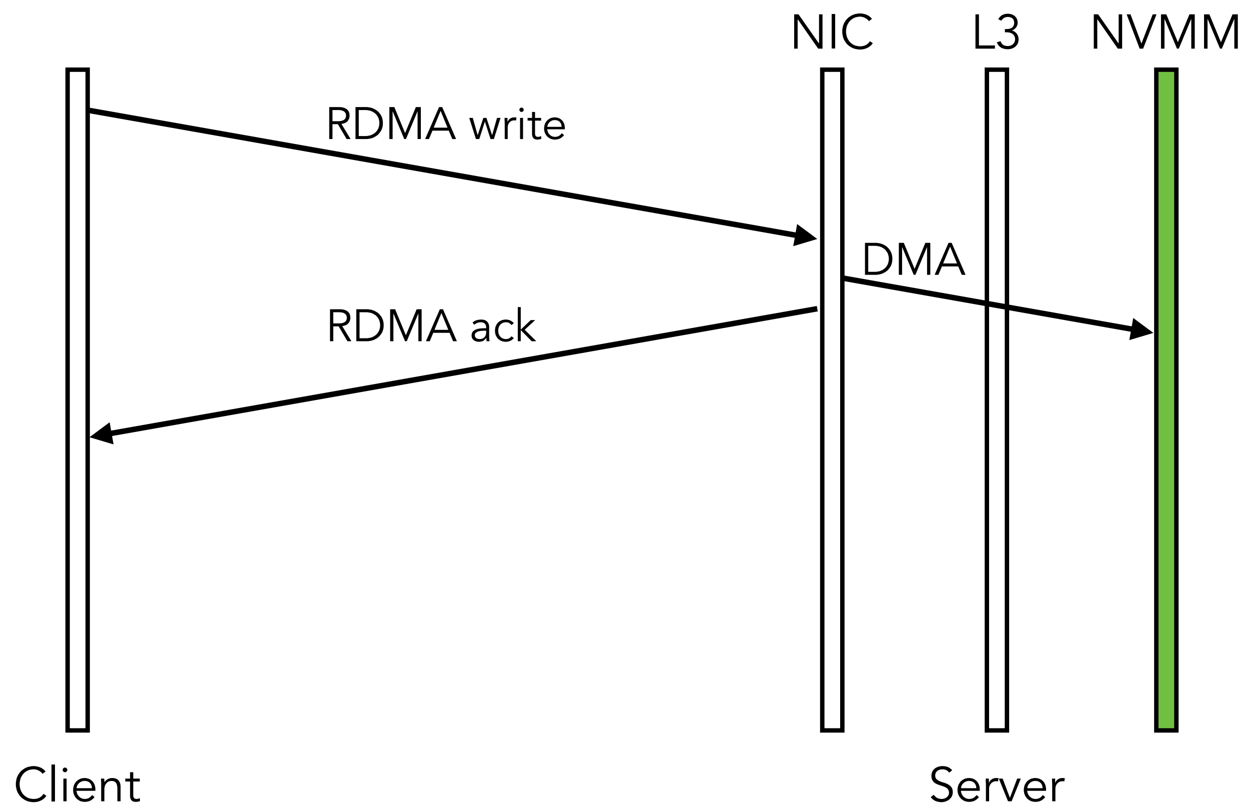
NVMM writes with Remote Procedure Calls (RPCs)



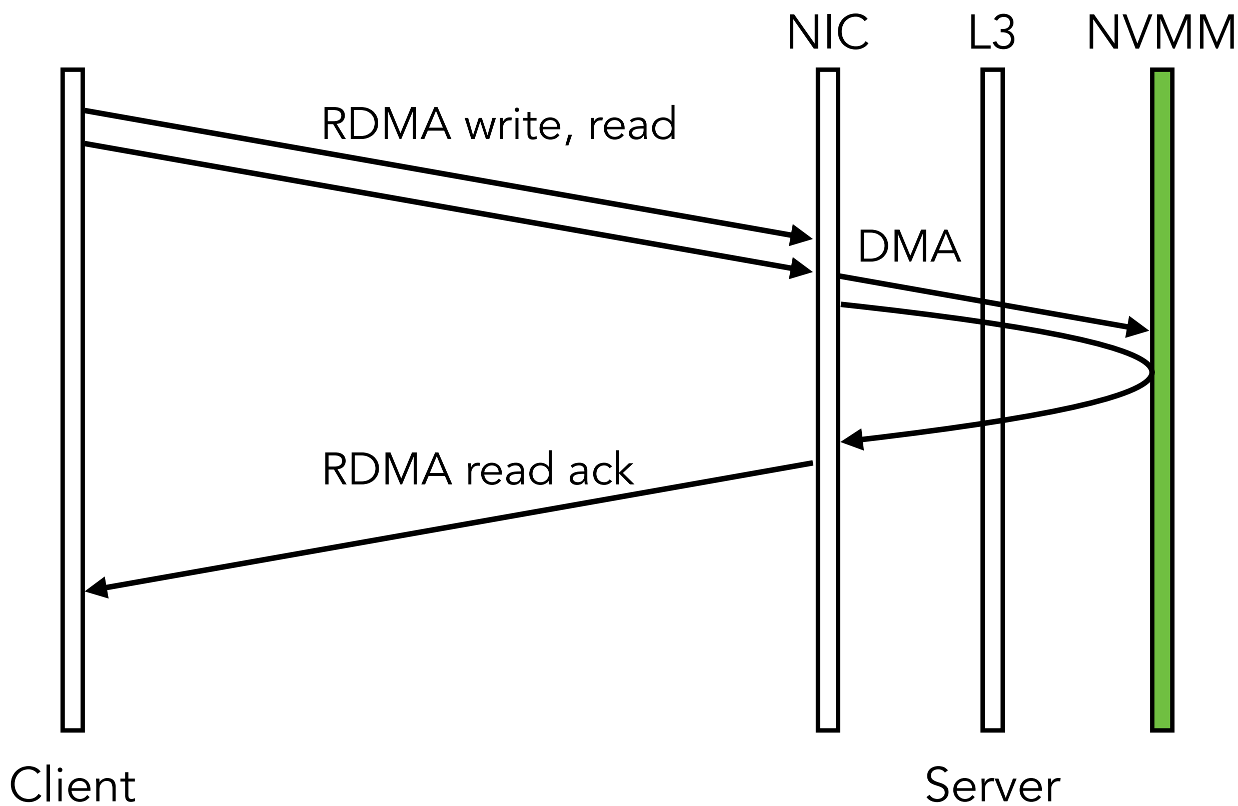
NVMM writes with Remote Direct Memory Access (RDMA)



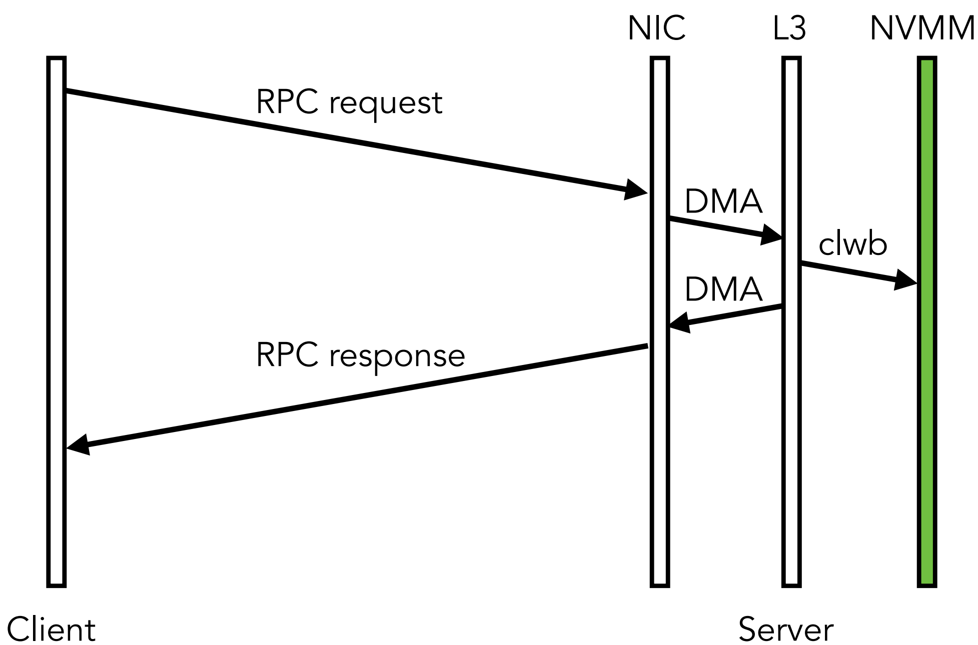
NVMM with RDMA, Direct Cache Access (DCA) disabled

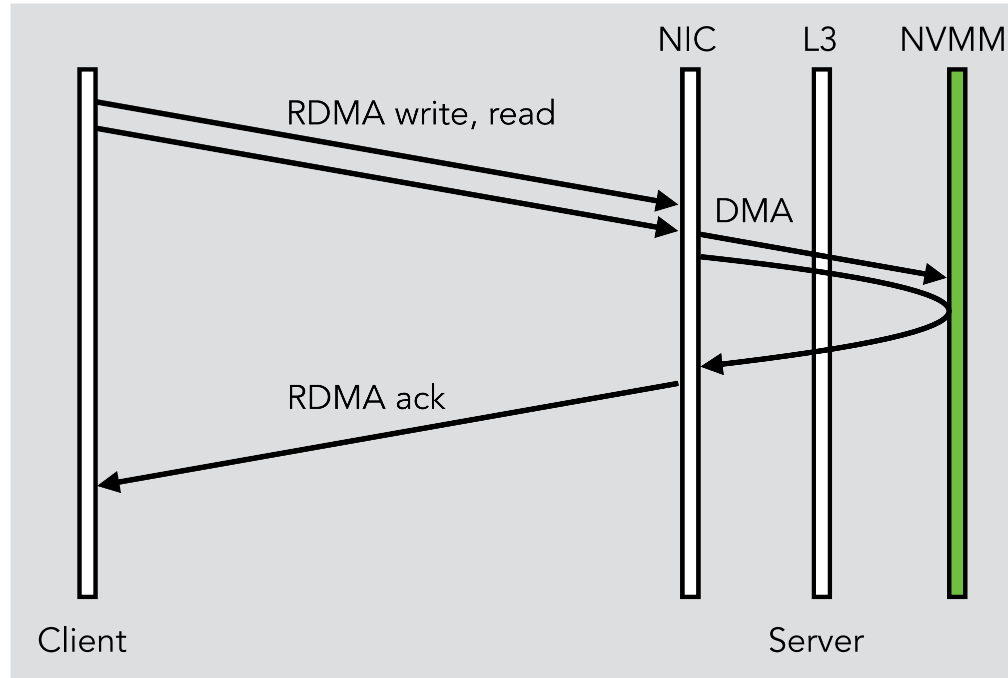


NVMM with Remote Direct Memory Access (RDMA)



NVMM removes latency advantage of RDMA over RPCs





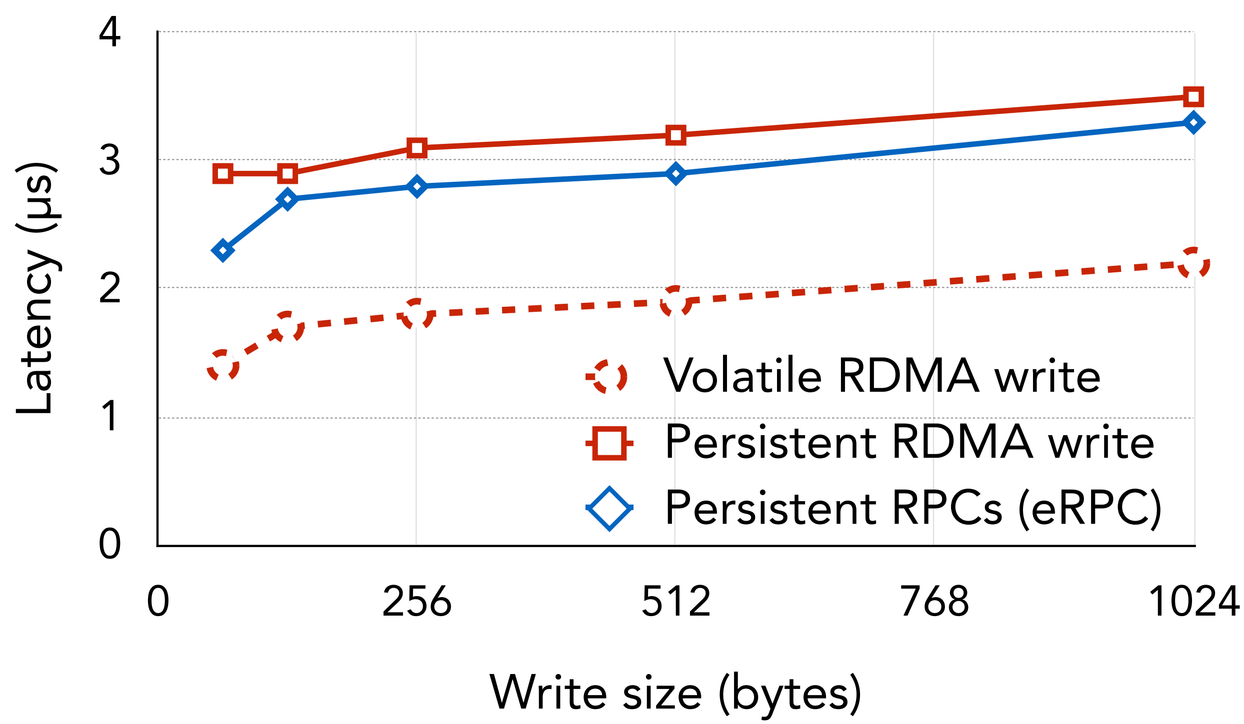
Critical path of persistent RPC:
One network RTT + one PCIe RTT

Critical path of persistent RDMA:
One network RTT + one PCIe RTT

NVMM removes latency advantage of RDMA over RPCs

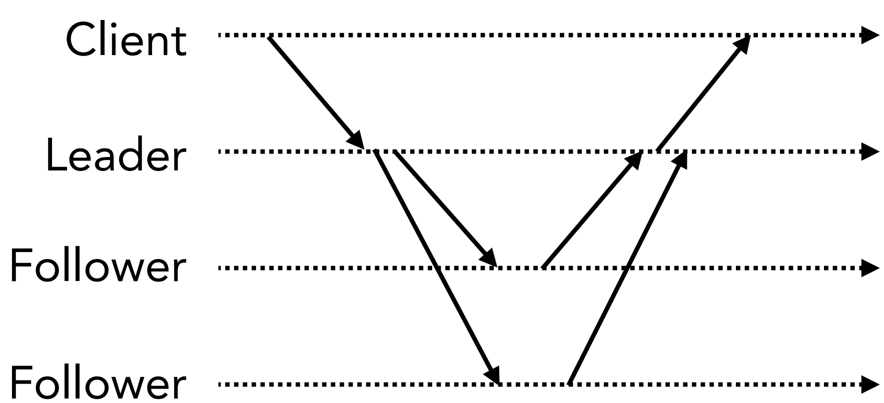
Experiment setup:

- Cascade Lake Xeon CPUs
- 6x 256 GB Optane DIMMs
- 56 Gbps ConnectX-3 InfiniBand

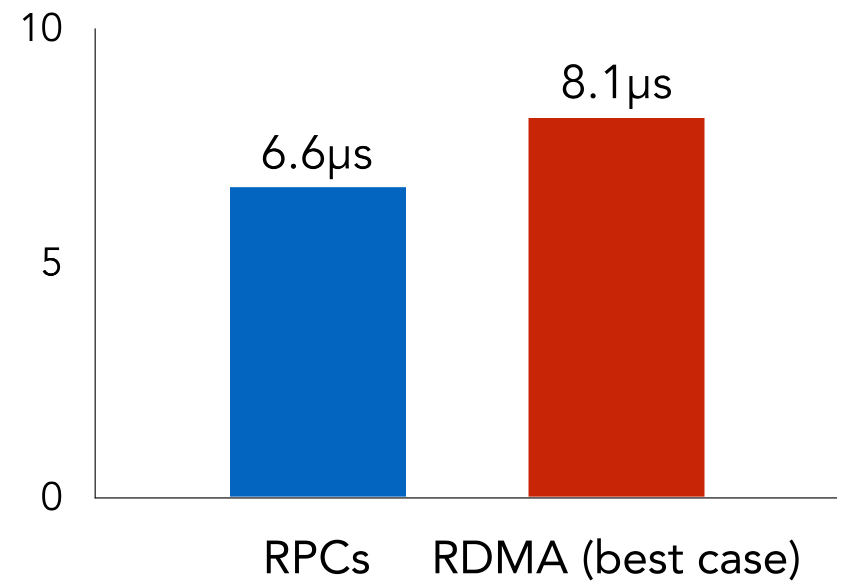


Application: State Machine Replication

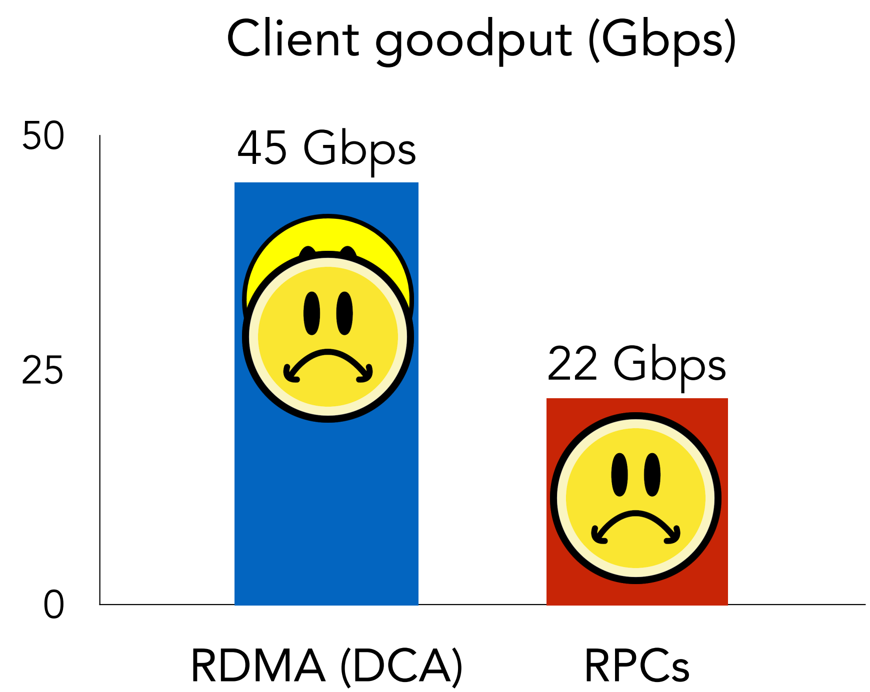
Network messages in Raft SMR



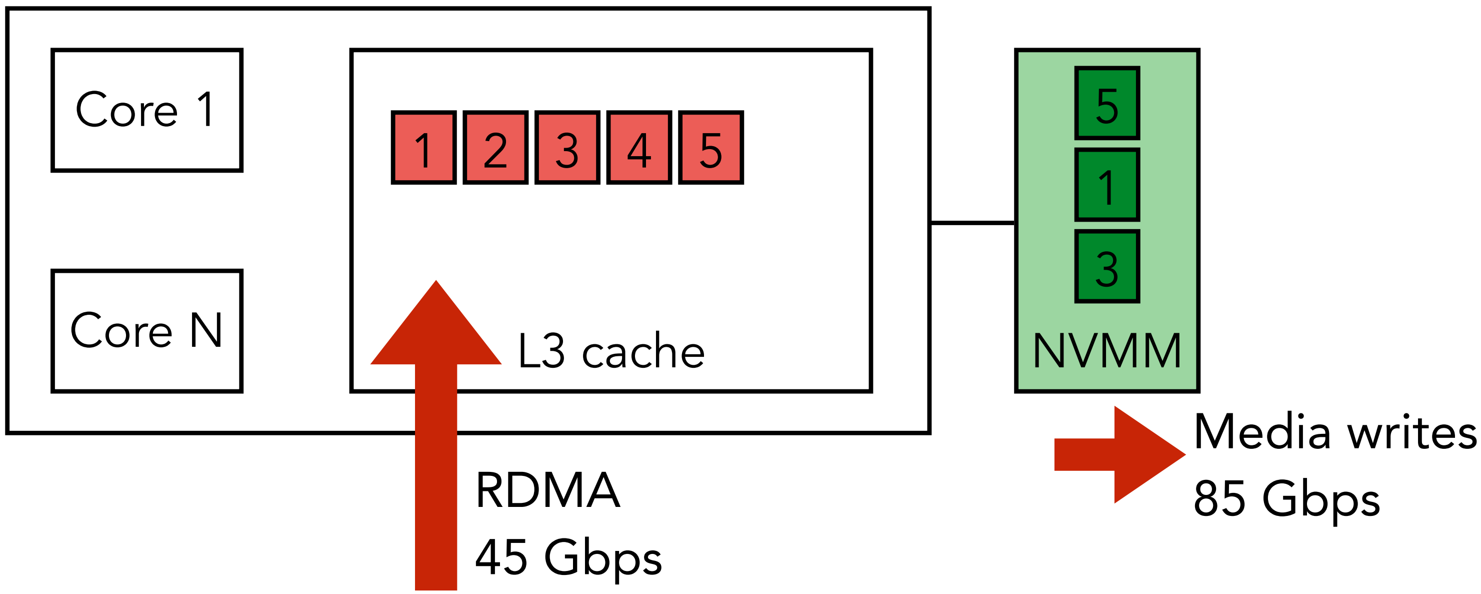
Client latency (median, μs)



Bandwidth of large writes to remote NVMM

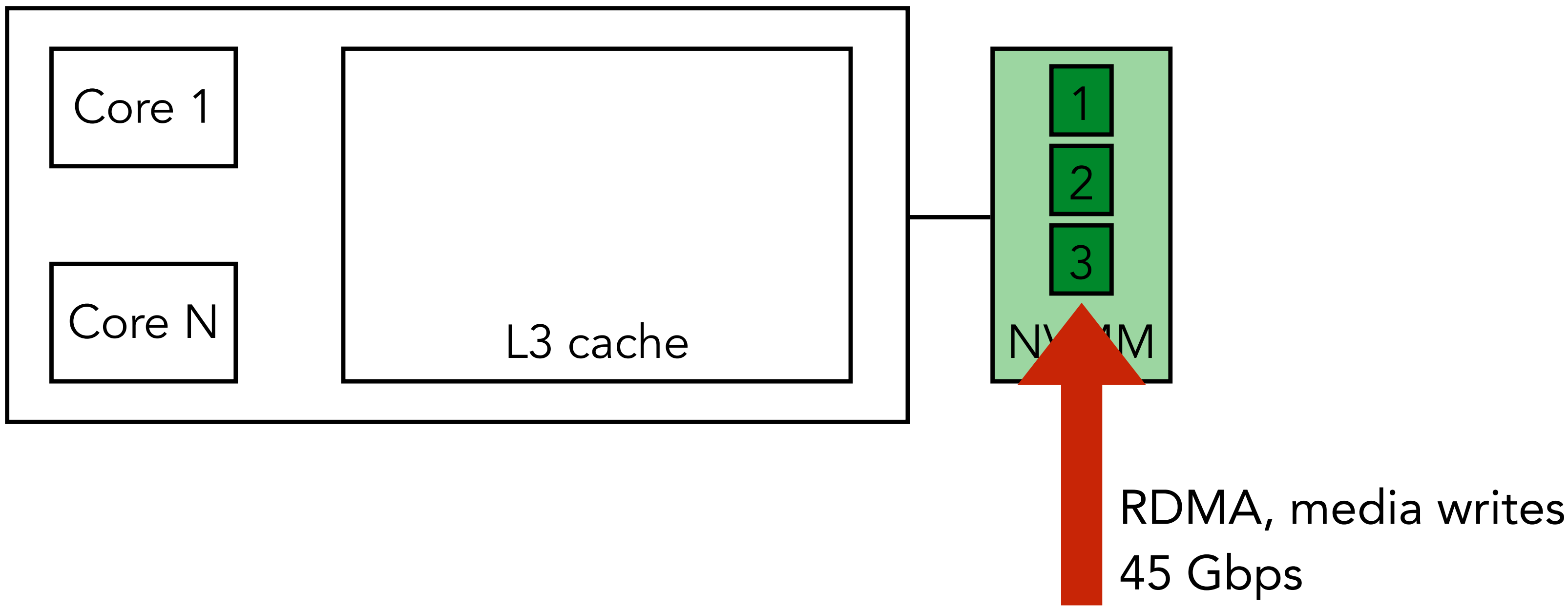


RDMA: Direct Cache Access randomizes NVMM write order



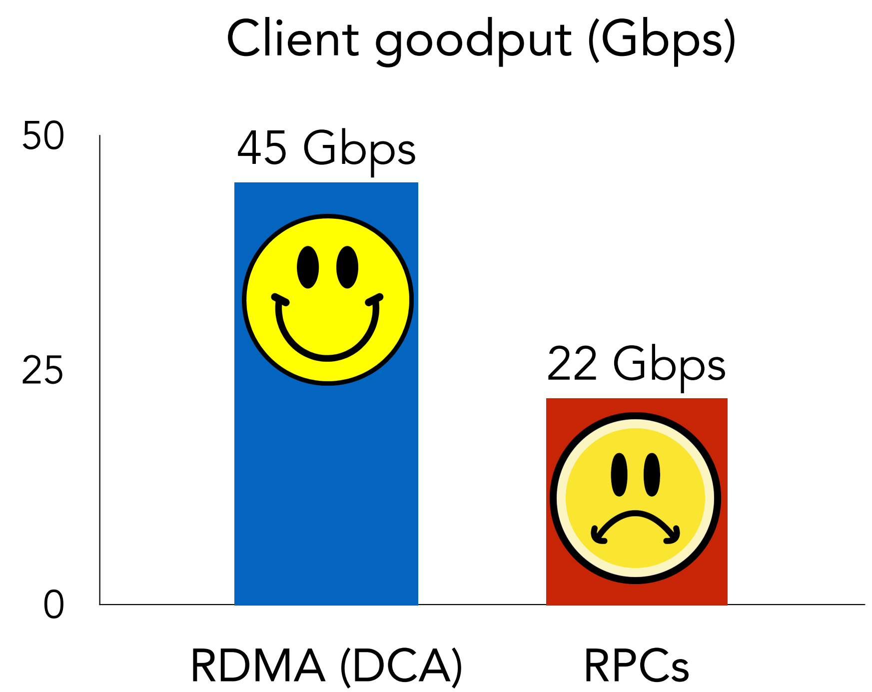
Root: CPU cache line size (64 bytes) < Optane memory erase block size (256 bytes)

Solution: Disable Direct Cache Access

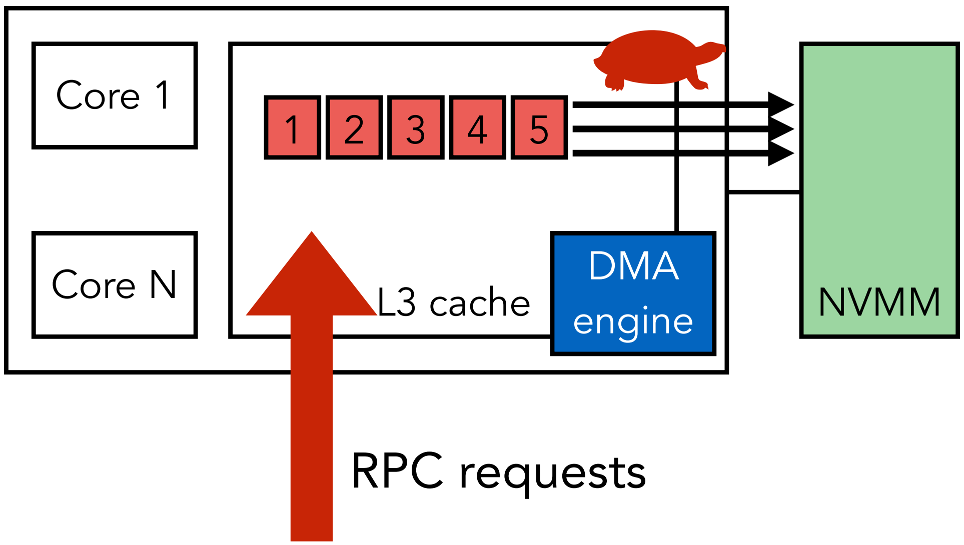


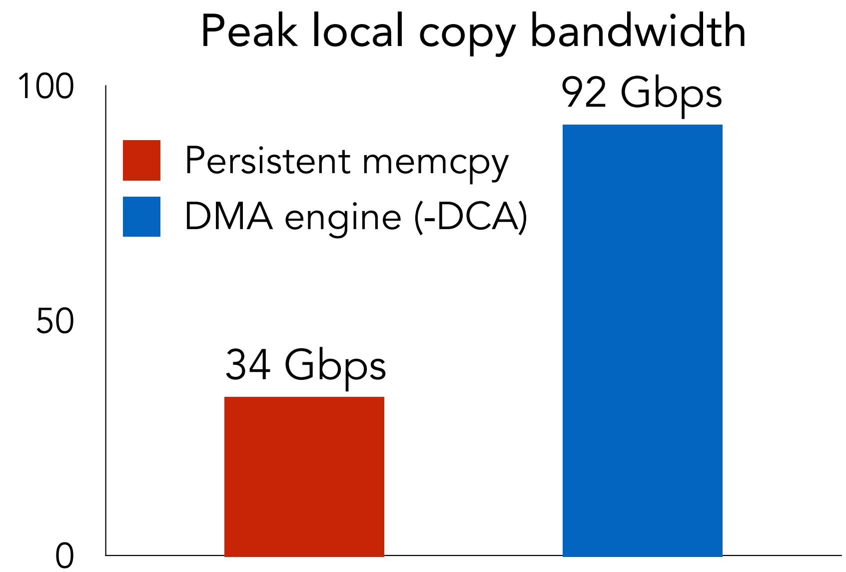
- **Several systems enable Direct Cache Access when using NVMM with RDMA**
- **Disabling a hardware optimization (Direct Cache Access) improves efficiency!**

Bandwidth of large writes to remote NVMM



Problem: CPU cores are slow at writing to NVMM





DMA engine improves bulk write bandwidth with RPCs from 22 Gbps to 48 Gbps

Conclusions

- Designing fast networked systems for NVMM requires attention to new low-level factors

- Our techniques can help while the hardware improves:
  - Better mechanisms for durable RDMA
  - Direct Cache Access without NVMM access order randomization
  - Faster persistent copying for CPU cores
  - See paper for more!

Thank you!