SNIA NSF | NETWORKING STORAGE

# Extending RDMA for Persistent Memory over Fabrics

**Live Webcast**

**October 25, 2018**

# Today's Presenters

**John Kim**
**SNIA NSF Chair**
**Mellanox**

**Tony Hurson**
**Intel**

**Rob Davis**
**Mellanox**

# SNIA-At-A-Glance

SNIA NSF | NETWORKING STORAGE

## SNIA-At-A-Glance

**170**
industry leading organizations

**3,500**
active contributing members

**50,000**
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**

# SNIA Legal Notice

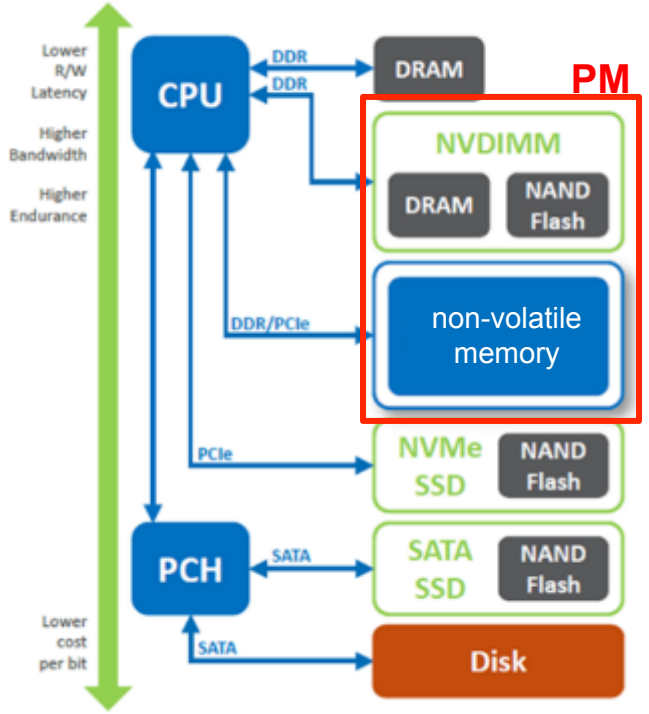
- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

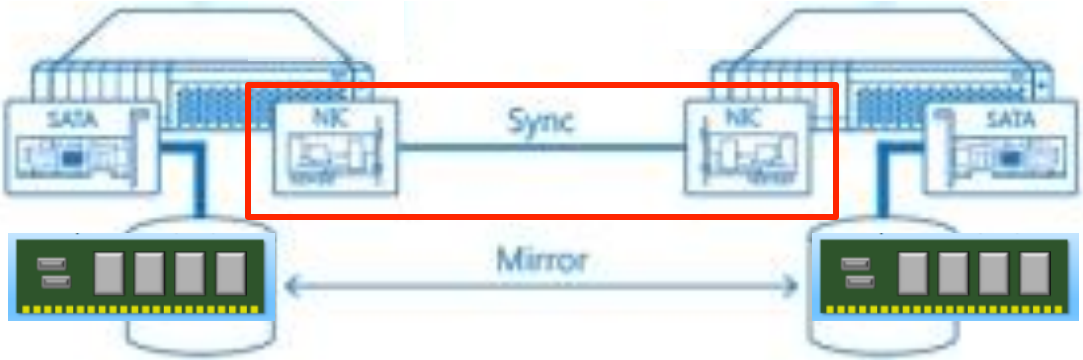
# What is Persistent Memory (PM)?

SNIA NSF | NETWORKING STORAGE

- Persistent Memory (PM) benefits
  - Considerably faster than NAND Flash
  - Performance accessible via PCIe or DDR interfaces
  - Lower cost/bit than DRAM
  - Significantly denser than DRAM



# For Many Storage Applications PM Needs Network Connectivity

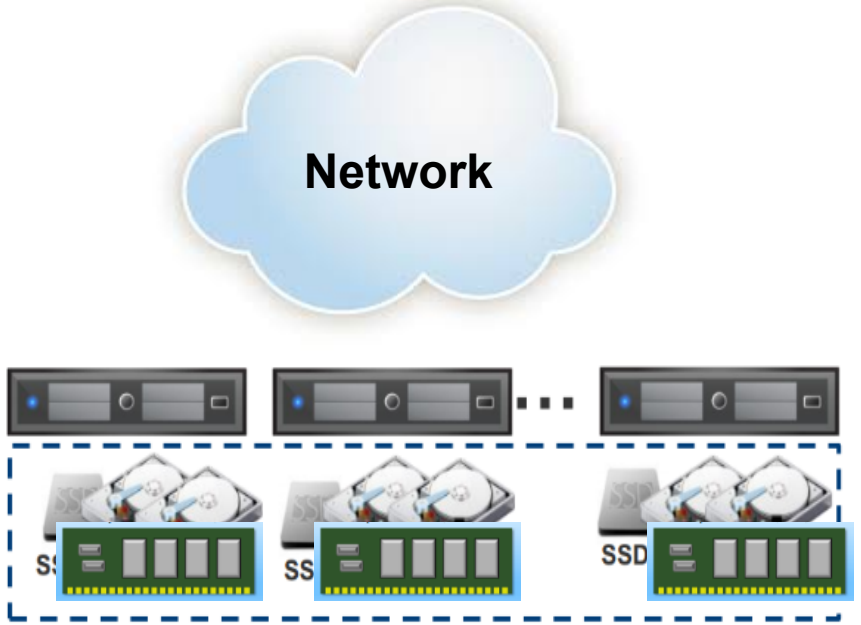



| | PM | NAND |
|---|---|---|
| Read Latency | ~100ns | ~20us |
| Write Latency | ~500ns | ~50us |

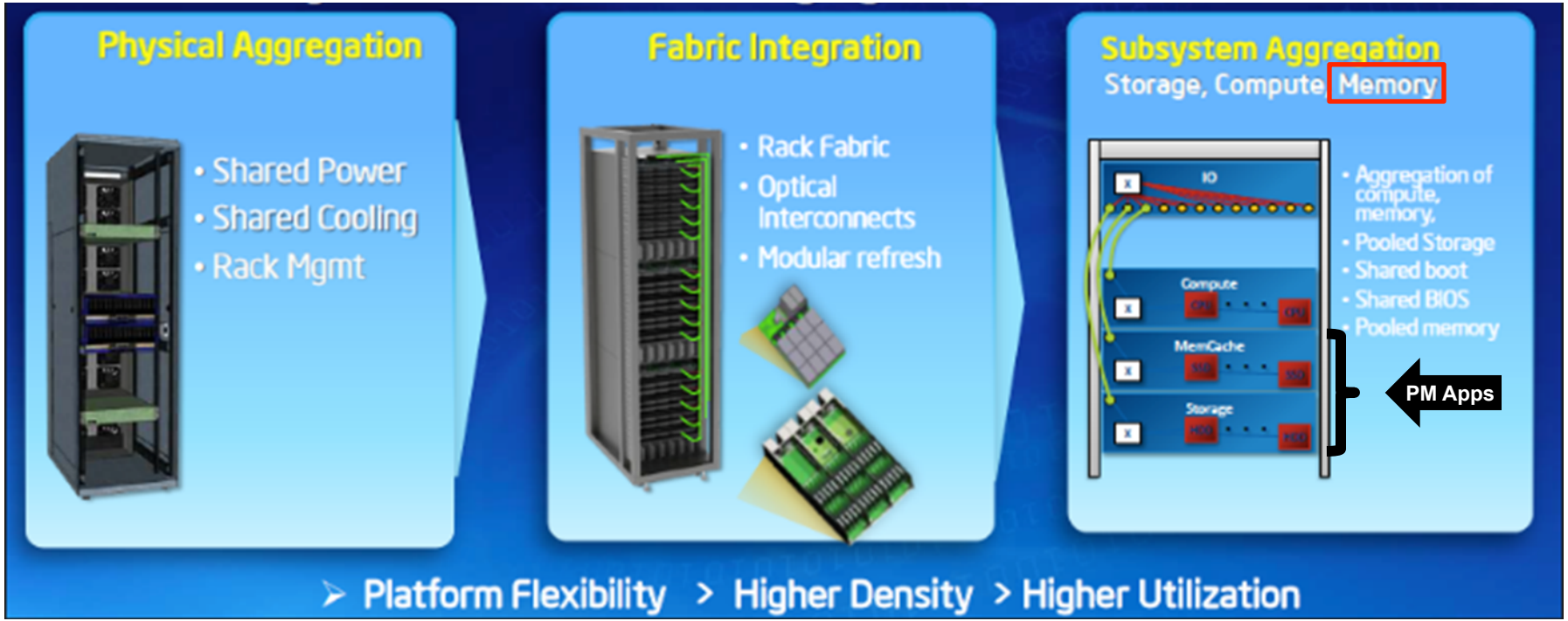
# Applications Benefitting RPM



- In-Memory application scale-out
- Software Defined Storage
- Big Data & NoSQL workloads
- Machine learning
- Hyperconverged Infrastructure
- Complete compute disaggregation



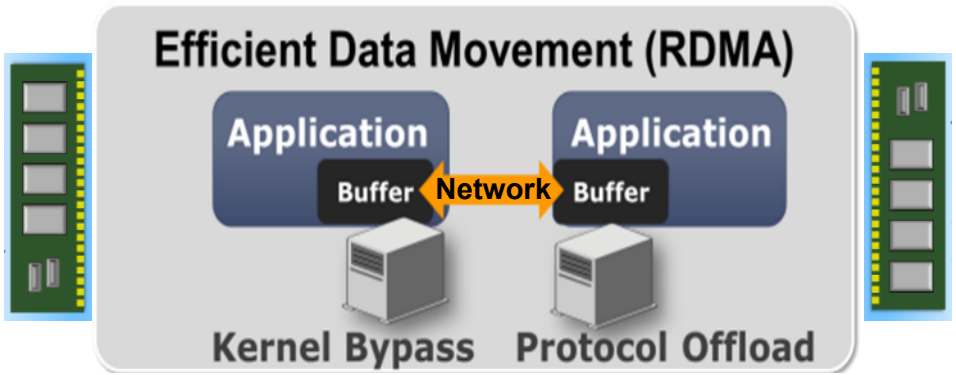
# Compute Disaggregation



# Core Requirements for Networked PM



| | PM | NAND |
|---|---|---|
| Read Latency | ~100ns | ~20us |
| Write Latency | ~500ns | ~50us |




- PM is really fast
  - Needs ultra low-latency networking
- PM has very high bandwidth
  - Needs ultra efficient protocol, transport offload, high bandwidth networks
- Remote accesses must not add significant latency
  - PM networking requires:
    - Predictability, Fairness, Zero Packet Loss
  - Network switches and adapters must deliver all of these

# What is RDMA?



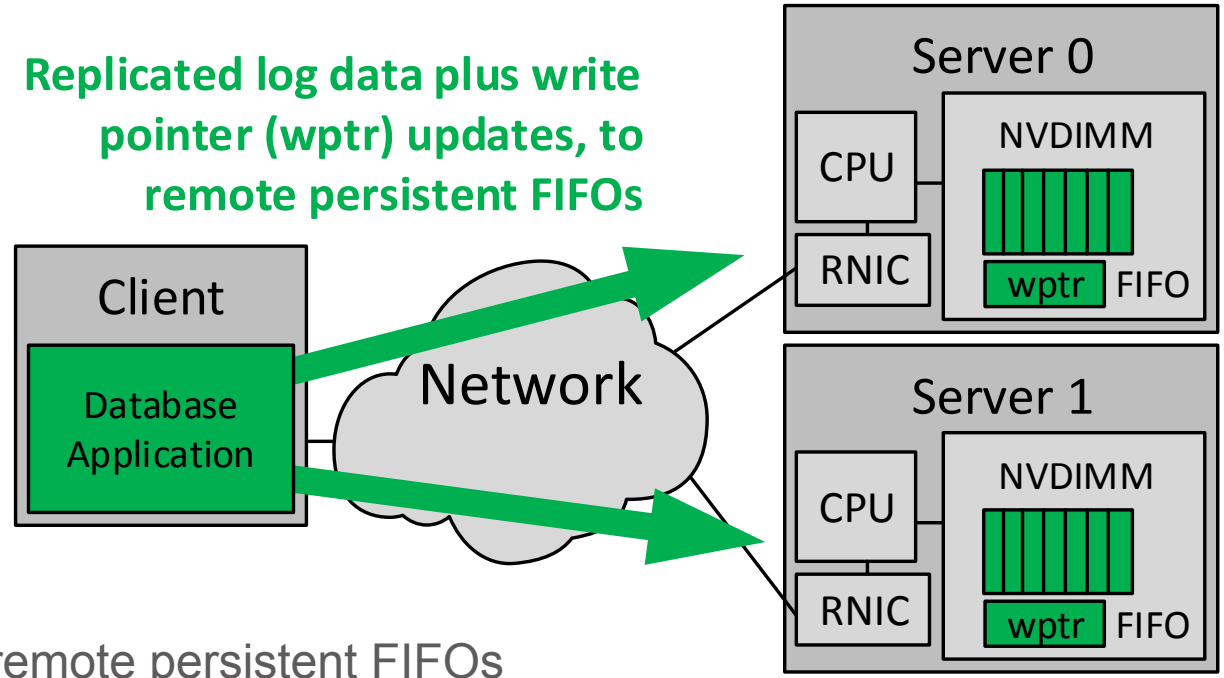
- Remote
  - Data transfers between nodes in a network
- Direct
  - No Operating System Kernel involvement in transfers
  - Everything about a transfer offloaded onto Interface Card
- Memory
  - Transfers between user space application virtual memory
  - No extra copying or buffering
- Access
  - Send, receive, read, write, atomic operations
  - Byte or Block Access

# RDMA Standards Development

- RDMA extensions for RPM being standardized in:
  - InfiniBand Trade Association (IBTA): native IB, RoCE
  - Internet Engineering Task Force (IETF): iWARP
- IBTA/IETF names/terminology may differ:
  - Commit : Flush
  - Atomic Write : Non-Posted Write
- However, extension semantics are similar, such that -
  - IETF will use same Verbs extensions as IBTA

# Application 1: Database Persistent Log

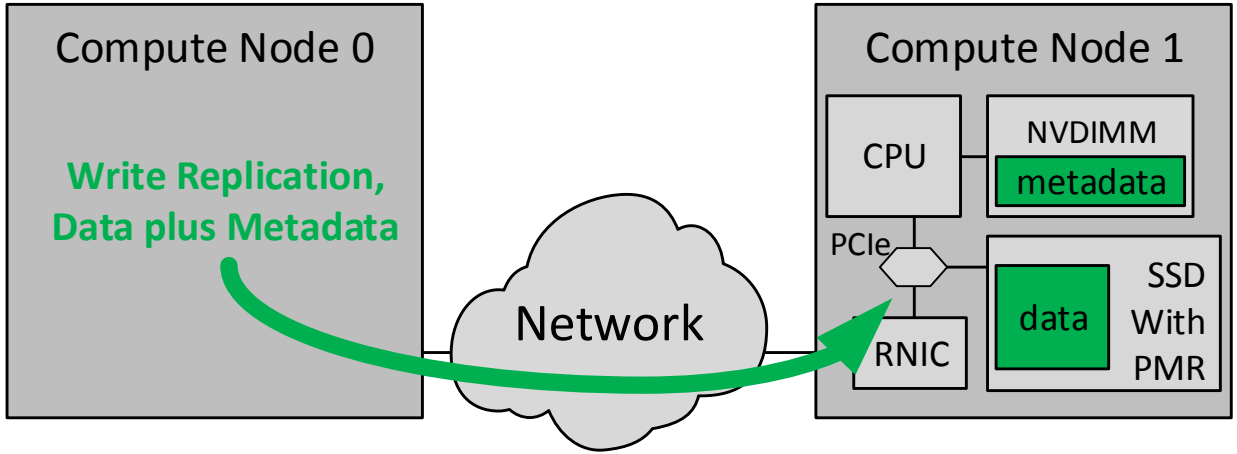





- Logs are remote persistent FIFOs
- Each log entry write: data (~4 KB), followed by write pointer (wptr) update

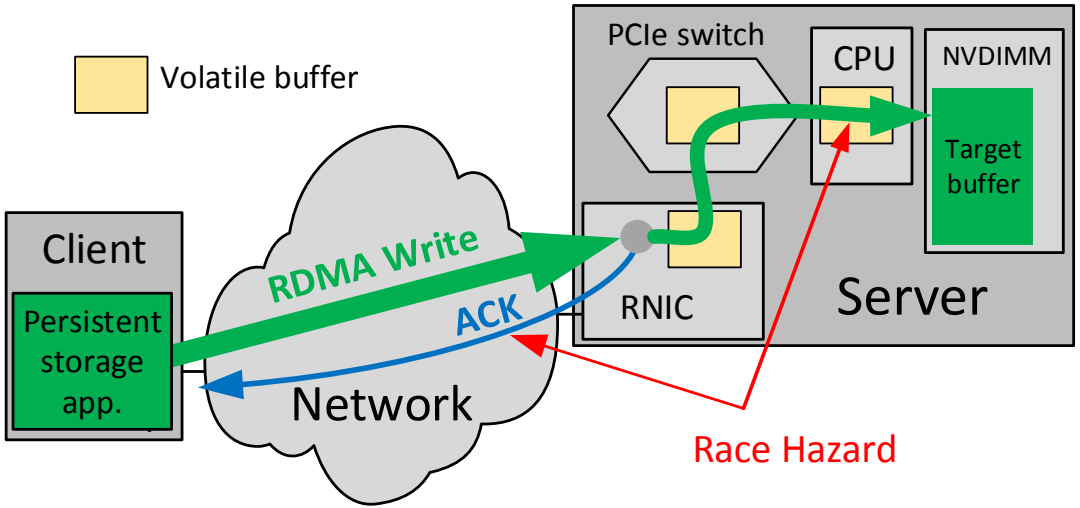
# Application 2: Hot Tier Hyperscale or Hyperconverged Storage





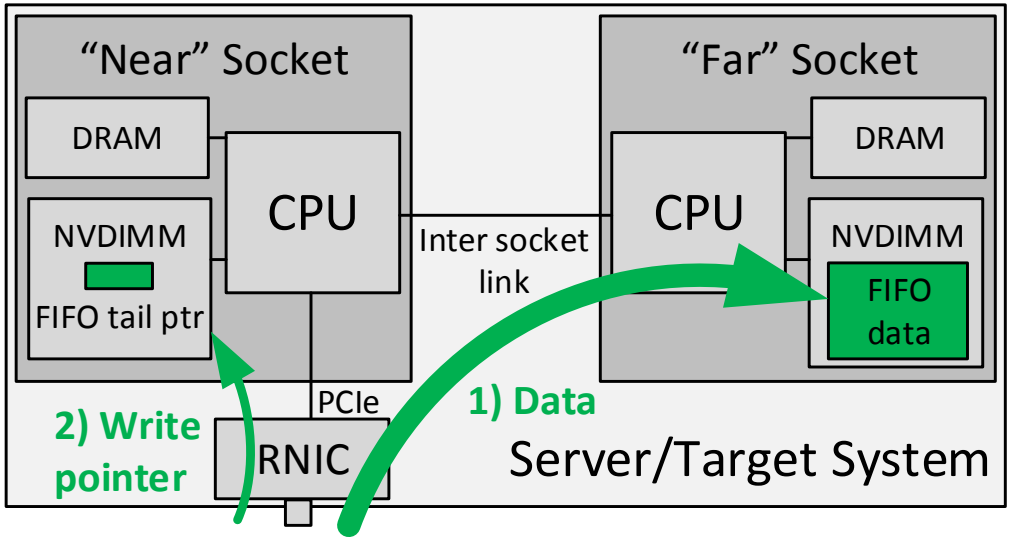

- Storage data writes, replicated for availability: latency-critical, but replicas seldom used
- Great application for emerging NVMe Persistent Memory Regions (PMRs)
  - Target RNIC writes data directly to SSD PMR, bypassing CPU/memory subsystem
- Accompanying metadata may be processed by target CPU

# Why Simple RDMA Writes are Not Enough




- Server RNIC acknowledges (ACK) Write as soon as validates it
- At server, there may be multiple volatile intermediate buffers between RNIC and persistent memory
- ACK might race back to complete Write at client, before actual data makes it to persistence
  - Bad news, if Write subsequently fails within server
- Client requirement: explicit confirmation of persistence, to follow Write

# The Need for RPM Write Ordering



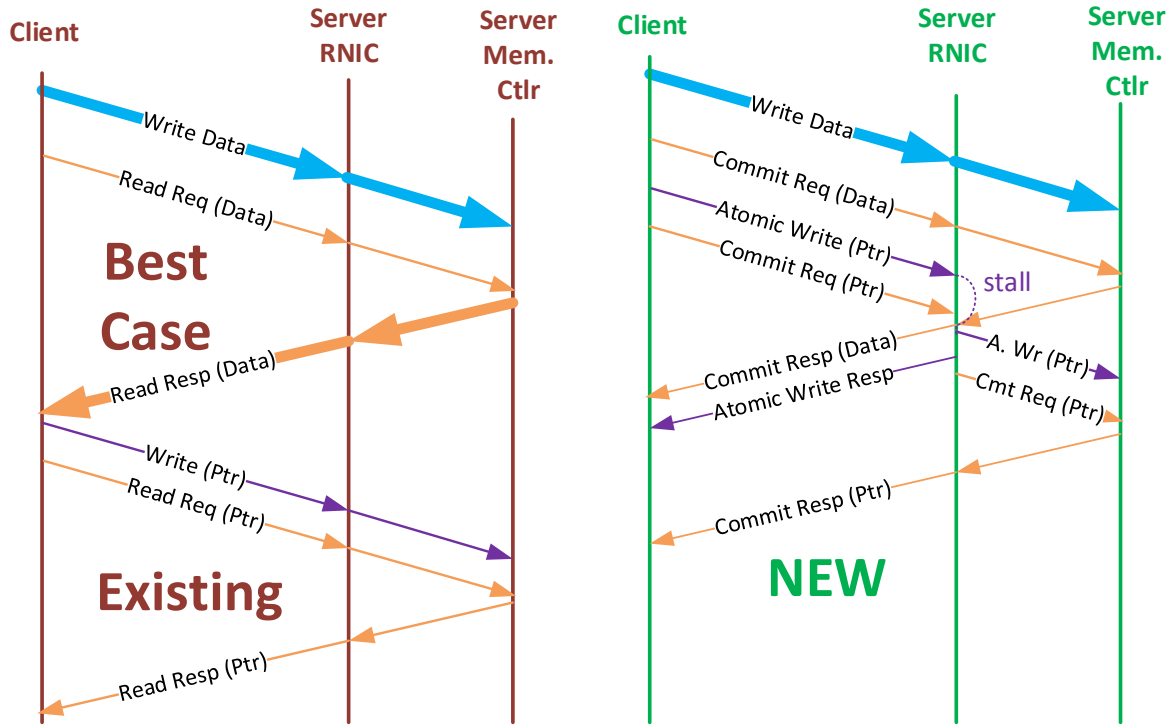

- Remote persistent FIFO, with data NVDIMM further from RNIC than write pointer NVDIMM
- Without explicit write ordering at RNIC, pointer update may race ahead of data update
- Result upon server failure: corrupt FIFO

# RDMA Extensions Summary

SNIA NSF | NETWORKING STORAGE

- Three new RDMA Messages:
  - Commit Request/Response: confirms prior Writes have been flushed and committed to persistence
  - Atomic Write Request/Response: small (8B) all-or-nothing Write, with explicit response
  - Verify Request/Response (IETF only): hash-based integrity check of persistent data
- New Persistence property for data Memory Regions (MRs)
- New API (Verbs) extensions for Messages and MRs

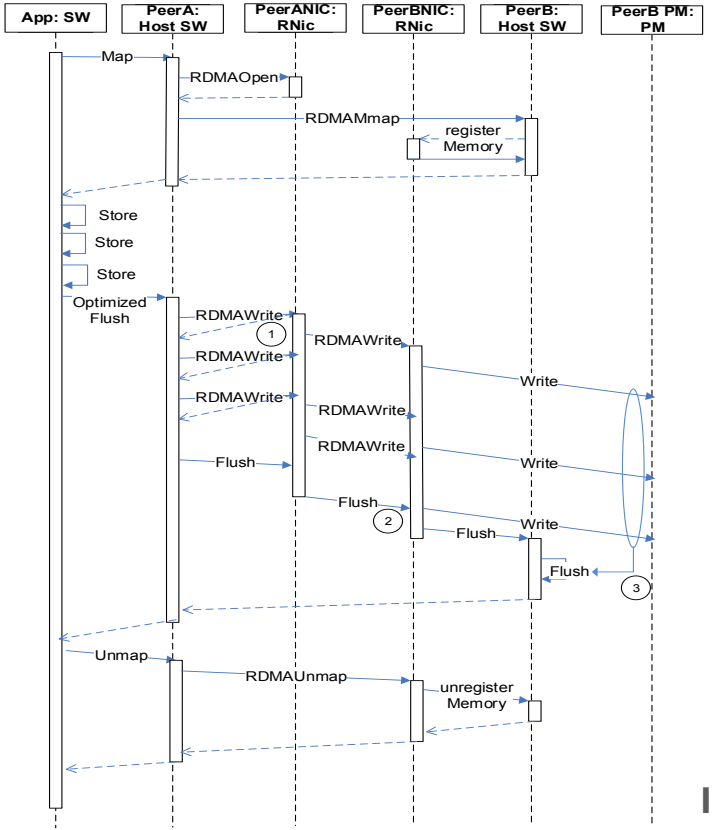
# Accelerating RPM Writes




- New RDMA Messages eliminate one serialized round-trip-time in network
- Commit guarantees persistence (unlike plain Reads on some platforms)
- New Atomic Write guarantees ordering of successive RPM writes
- (Diagram uses IETF terminology)

# Matches SNIA NVM PM Model RDMA to PMEM for High Availability

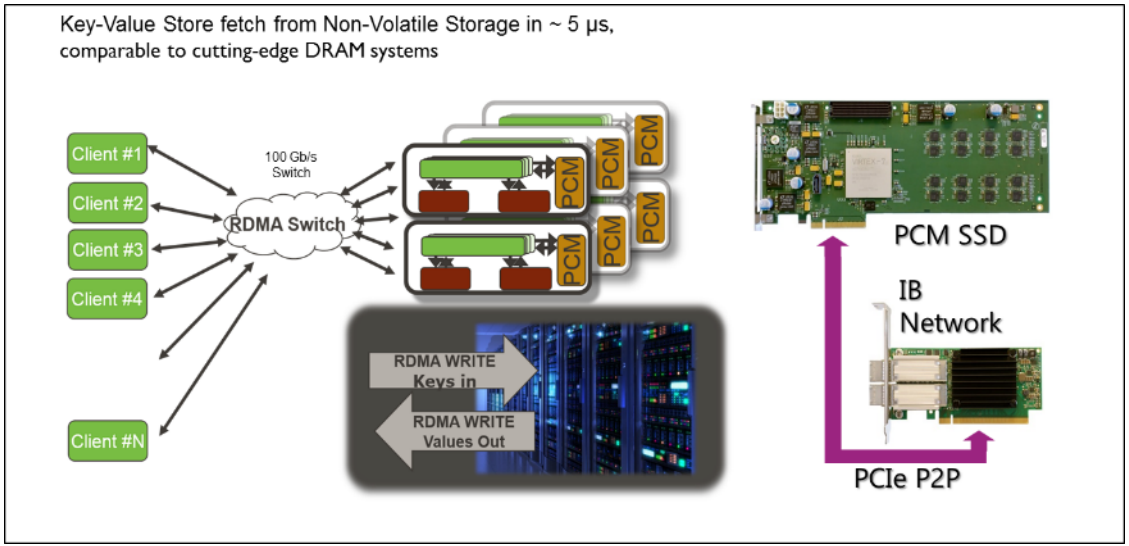
- MAP
  - Memory address for the file
  - Memory address + Registration of the replication
- SYNC
  - Write all the “dirty” pages to remote replication
  - FLUSH (IBTA or Commit, IETF) the writes to persistency
- UNMAP
  - Invalidate the registered pages for replication



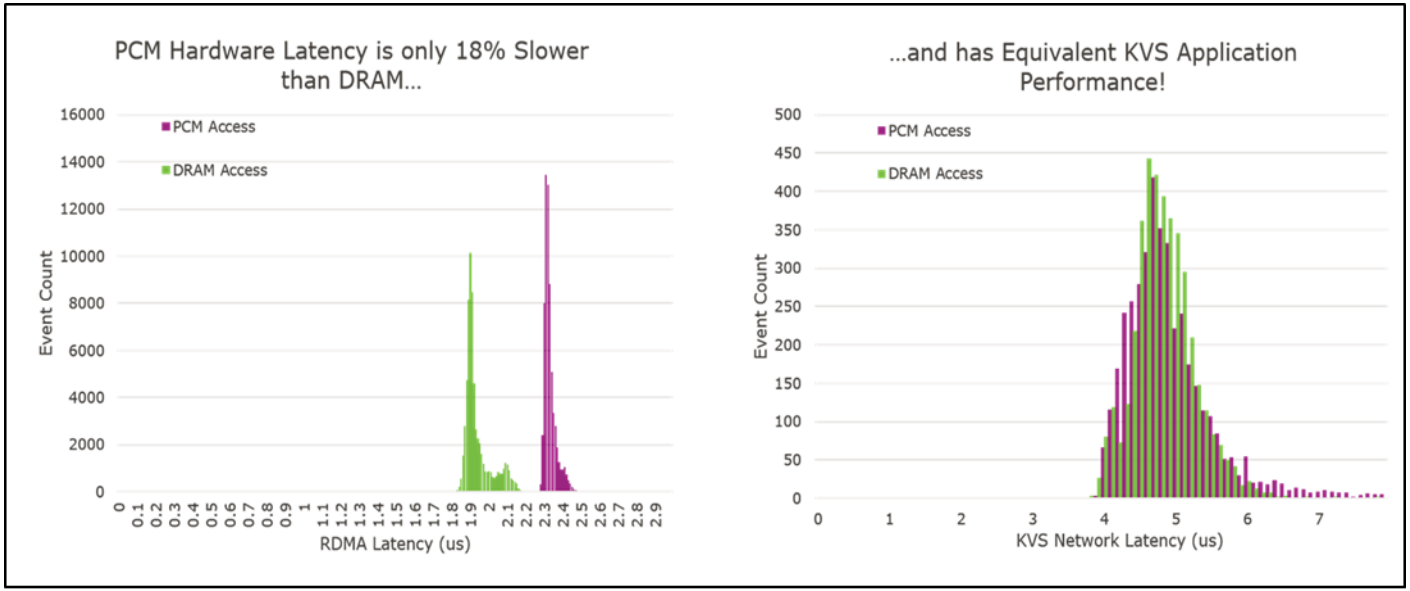
# HGST PM Remote Access Demo



- HGST live demo at Flash Memory Summit
- RPM(=PCM) based Key-Value store over 100Gb/s RDMA



# Application Level Performance




- PCM is slightly slower than DRAM but … equivalent application perf
- https://www.hgst.com/company/media-room/press-releases/HGST-to-Demo-InMemory-Flash-Fabric-and-Lead-Discussions

# After This Webcast

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Networking Storage Forum (NSF) website and available on-demand at www.snia.org/library
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-NSF blog: sniansfblog.org
- Follow us on Twitter @SNIANSF

SNIA NSF | NETWORKING STORAGE

# Thanks!