STORAGE INDUSTRY SUMMIT

Convergence of Storage and Memory Developing the Needed Ecosystem

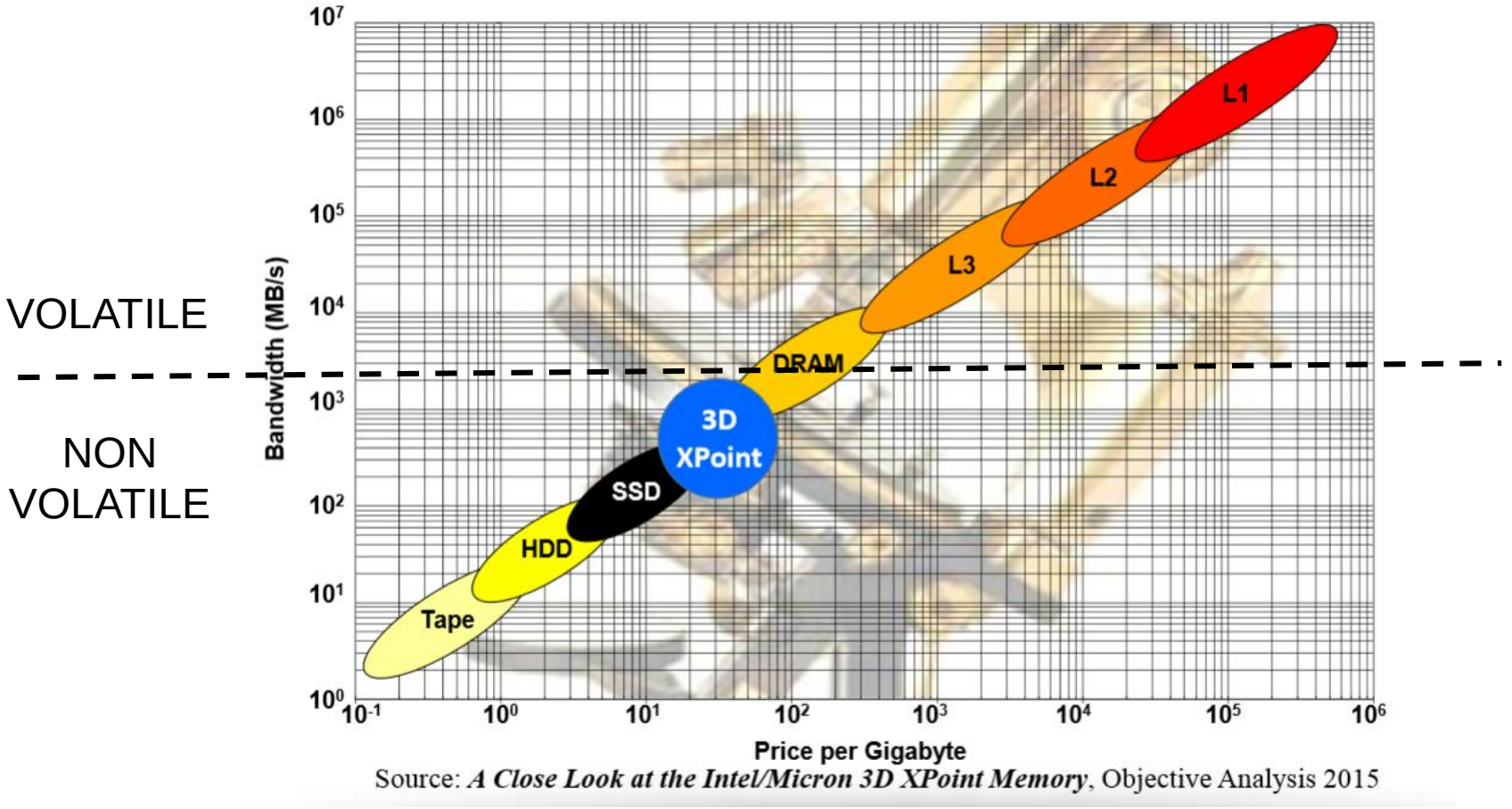
JANUARY 20, 2016, SAN JOSE, CA

Kevin Deierling
Mellanox Technologies
VP, Marketing

Persistent Memory over Fabric

SNIA SSSI | SOLID STATE STORAGE

Memory Hierarchy



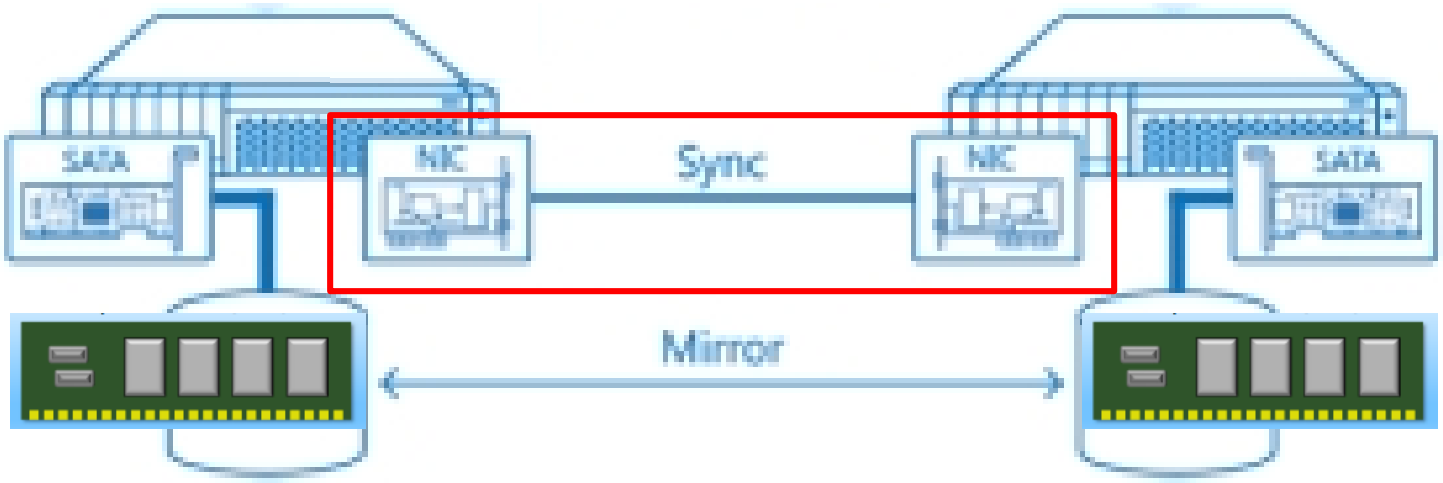
- Ideal memory is non-volatile, < 1ns access time, & is free … but we live in the real world

There is Always More Data than Memory

- Nature Abhors a Vacuum (Aristotle)
- Corollary: Computer Scientists Abhor Unused Memory (Dilbert?)

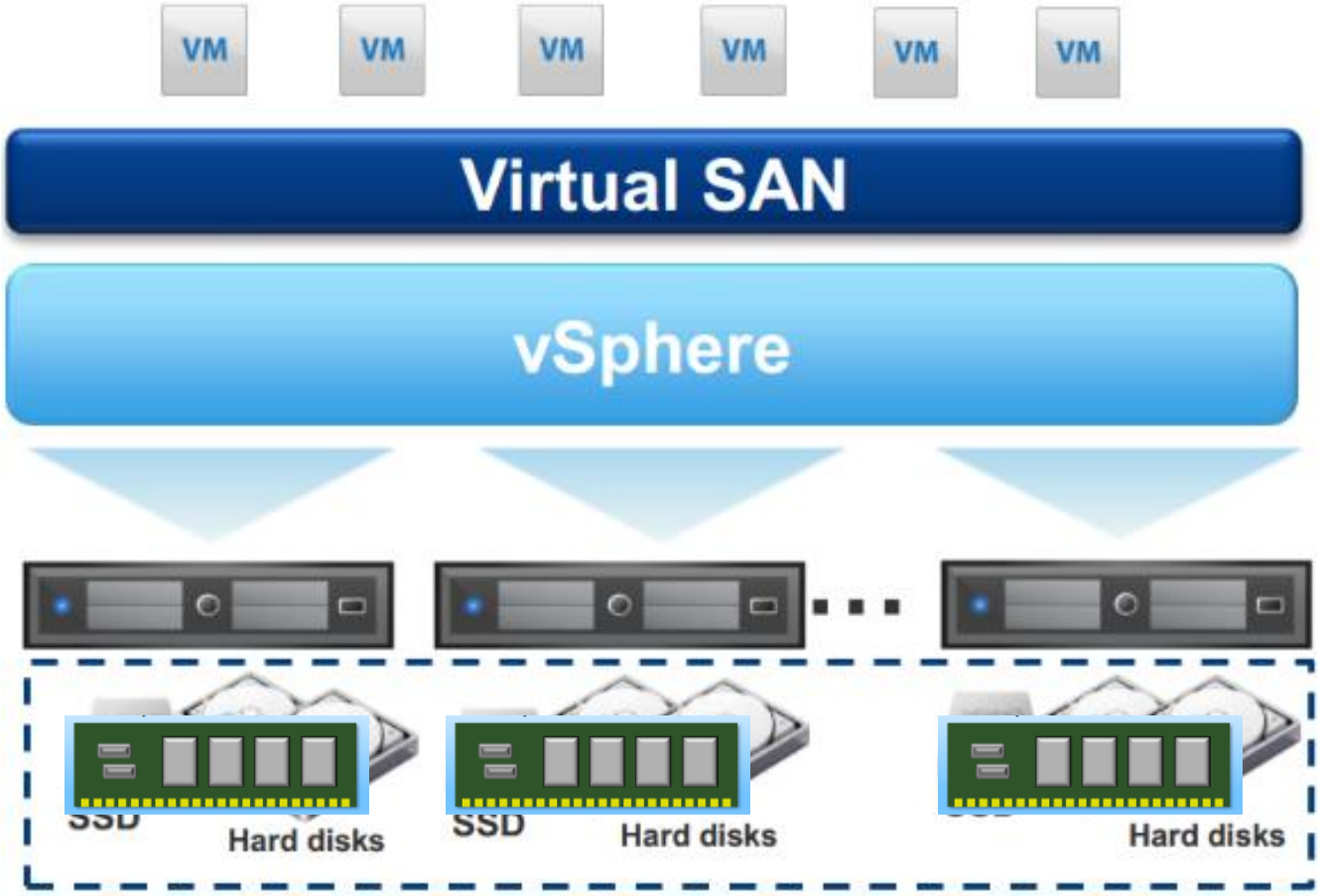
Challenges to Network PM

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



- PM is really fast so needs low-latency networking

Hyper Converged Infrastructure

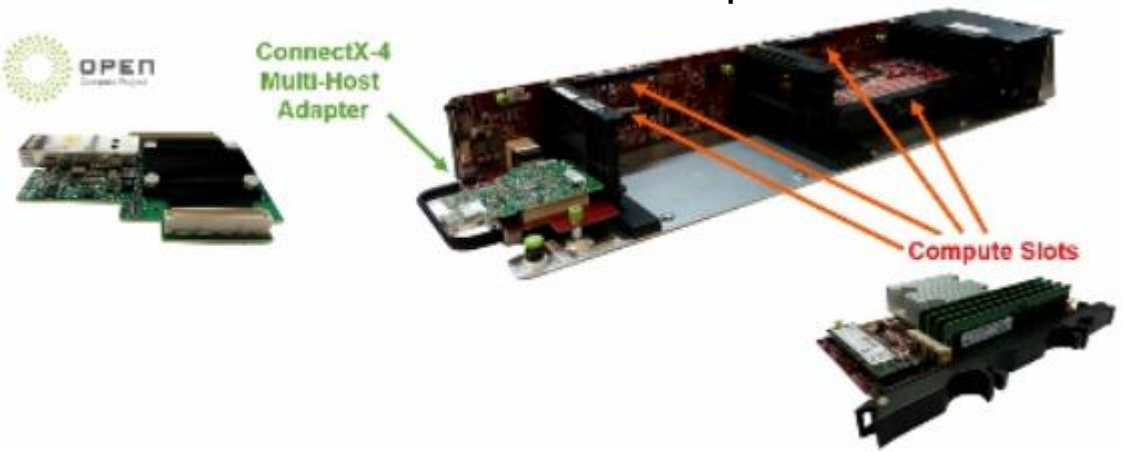


Server Disaggregation

Network

Compute

Persistent Memory

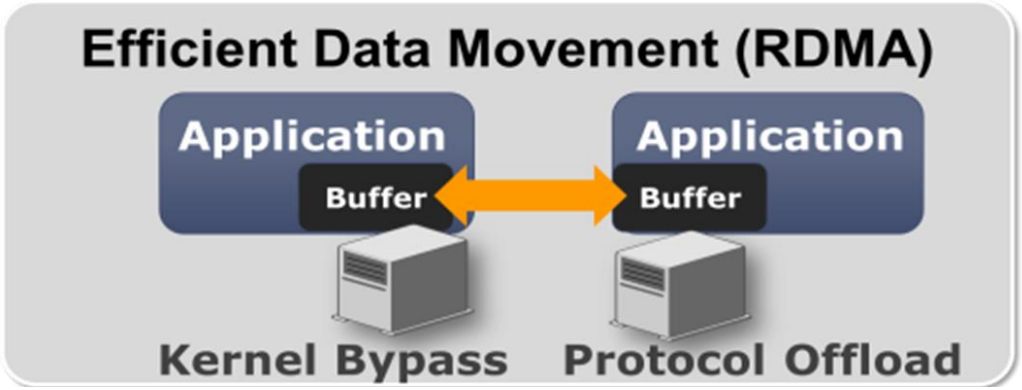


- Efficient resource utilization
- Evolve components to use case
- Smarter technology refreshes
- Speed of innovation

Dissagregation Requires High Performance Network

Networked PM Needed

- Networking PM Required
  - For efficiency, scalability, and reliability
- Name:
  - Phase Change Memory over Fabrics (PMoF?) (PMf?)
- Protocol:
  - RDMA: Remote Direct Memory Access



NVMf 1.0 Performance

SNIA™ Solid State Storage Initiative

Canonical Latency (4K)
17us

	IOPs	Latency
iodepth = 8	2.2 M	55 us
iodepth = 16	2.3 M	108 us
iodepth = 32	2.4 M	208 us

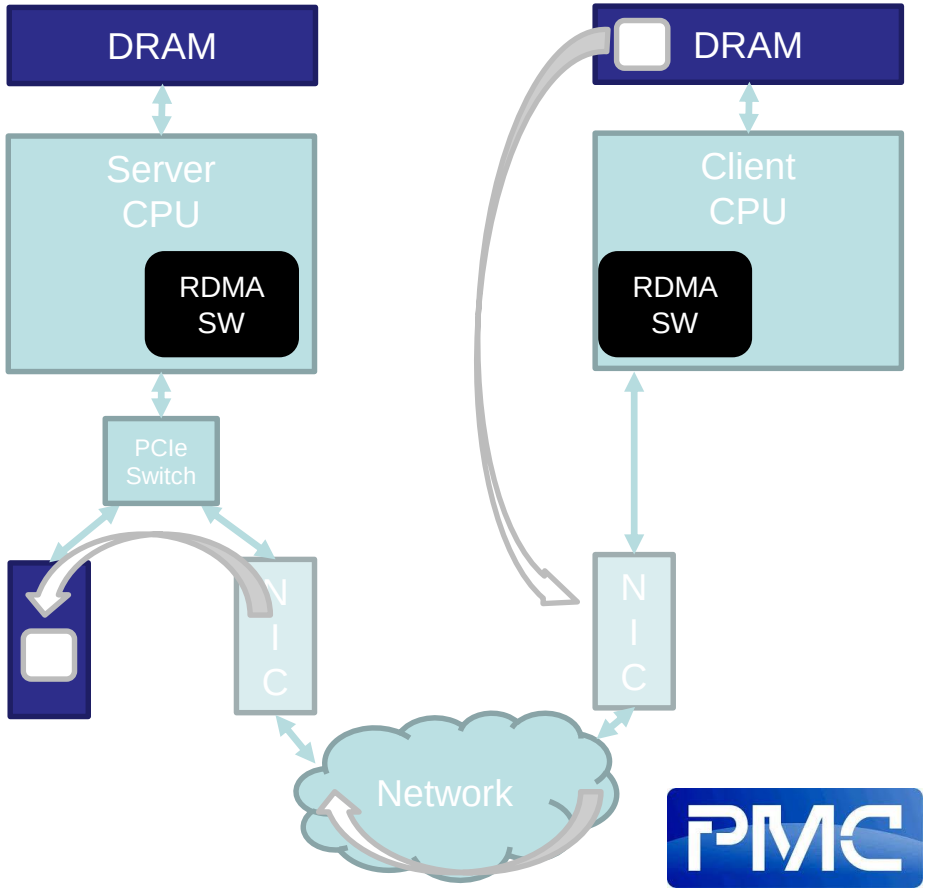
	Bandwidth
BS = 4K	40.6 Gb/s
BS = 8K	43.8 Gb/s
BS = 16K	49.5 Gb/s

Setup
Xeon CPU E5-2670 0 @ 2.60GHz
2 CPUs, 8 cores each, no HT
ConnectX-3 56Gbps
1 port back-to-back
1 LUN
16 jobs (threads), 1 per core

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

Peer-Direct NVRAM over RDMA Fabrics

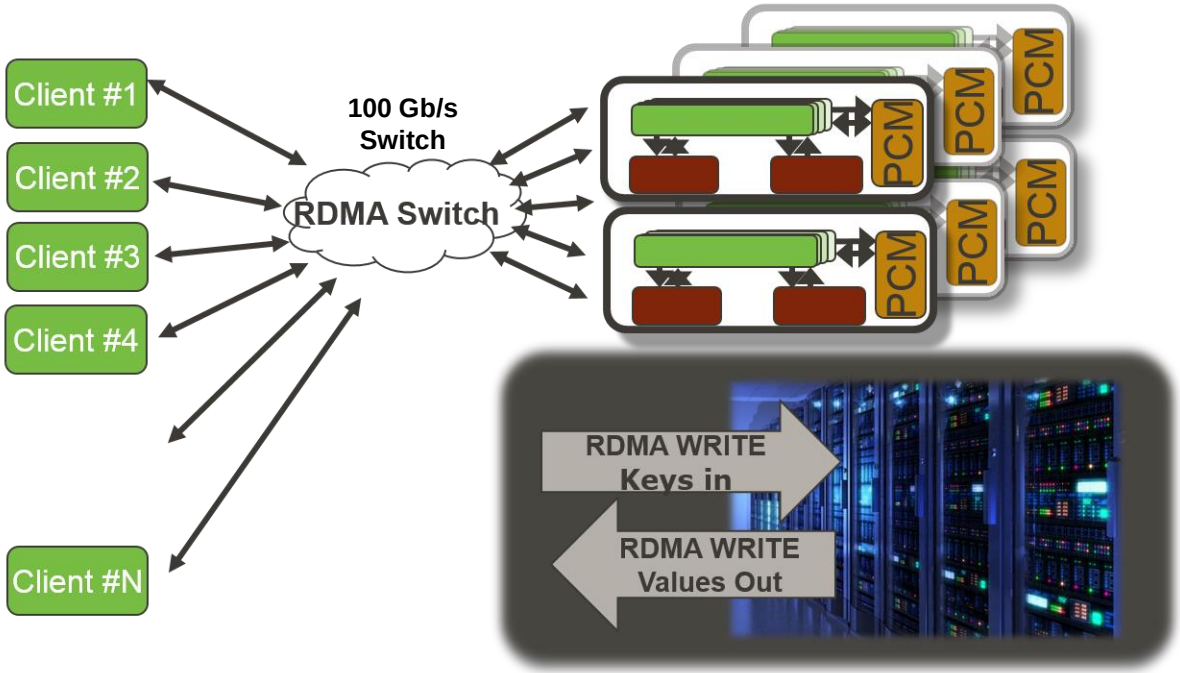
- Dev Platform
  - Mellanox RoCE
  - PMCS NVRAM Card
- Peer-Direct
  - I/O ops bypass host CPU
  - Reduced server load & DRAM bandwidth
- 7us latency 4KB IO
  - Client-Server to Mem
- Network latency critical

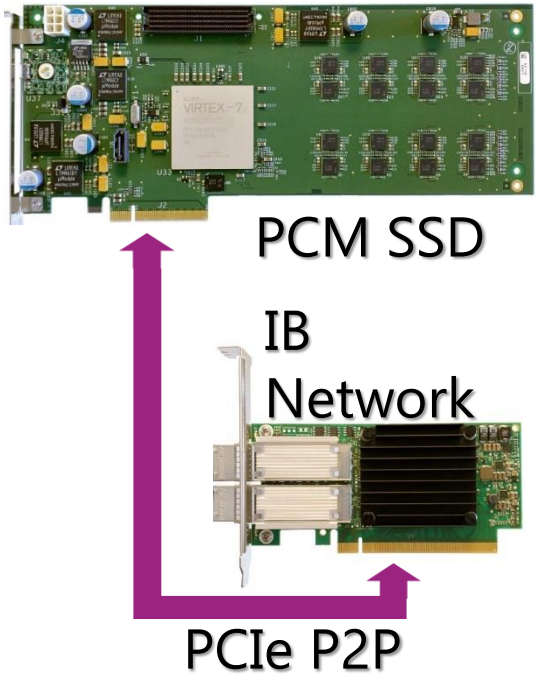


Live Proof of Concept Demo at Flash Memory Summit 2015

HGST PCM Remote Access Demo

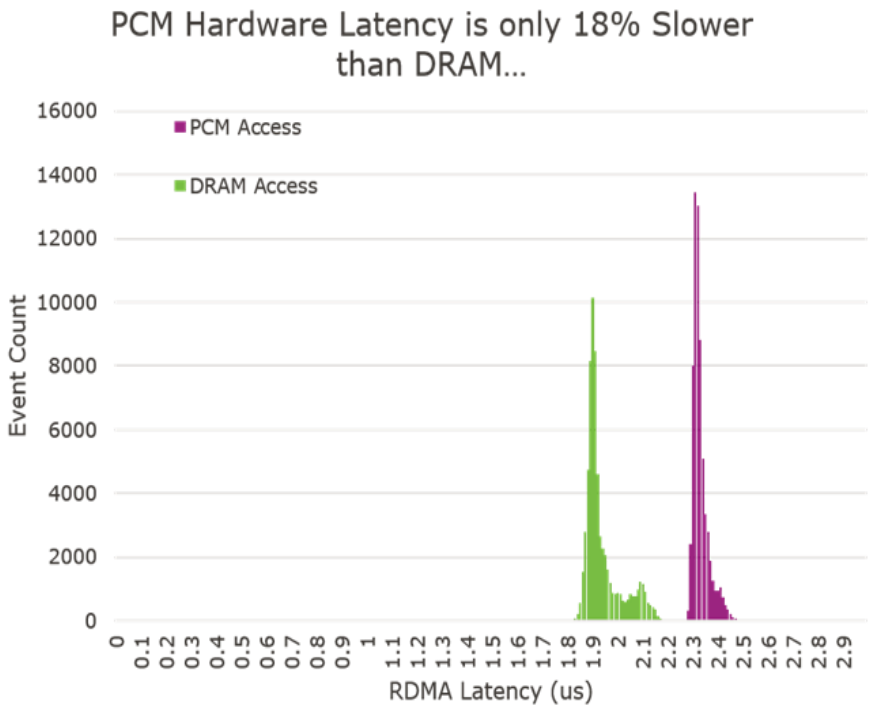
Key-Value Store fetch from Non-Volatile Storage in ~ 5 µs, comparable to cutting-edge DRAM systems

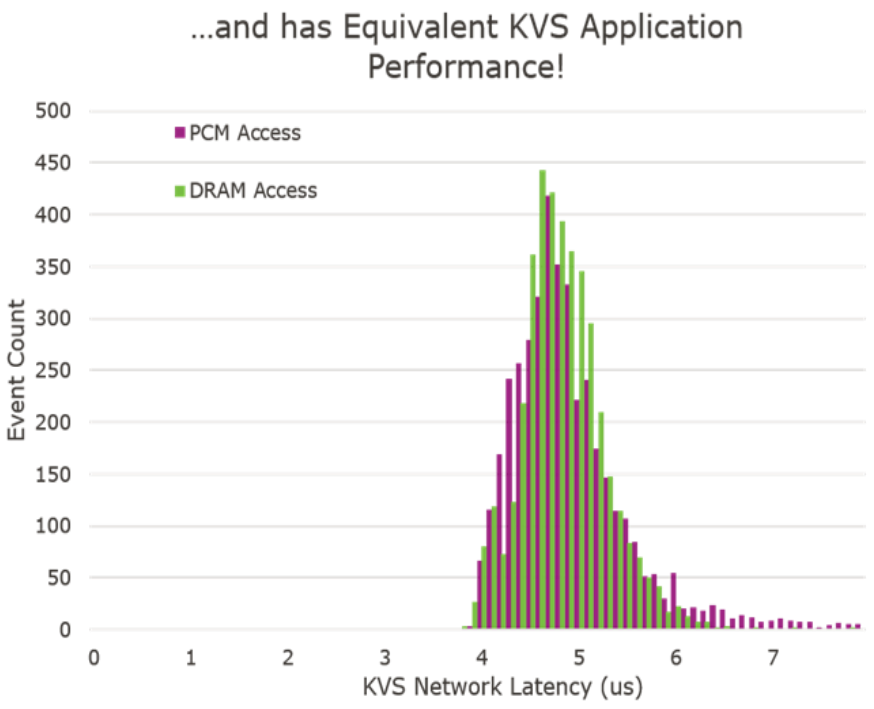




- HGST live demo at Flash Memory Summit
- PCM based Key-Value store over 100Gb/s RDMA

App Level Performance



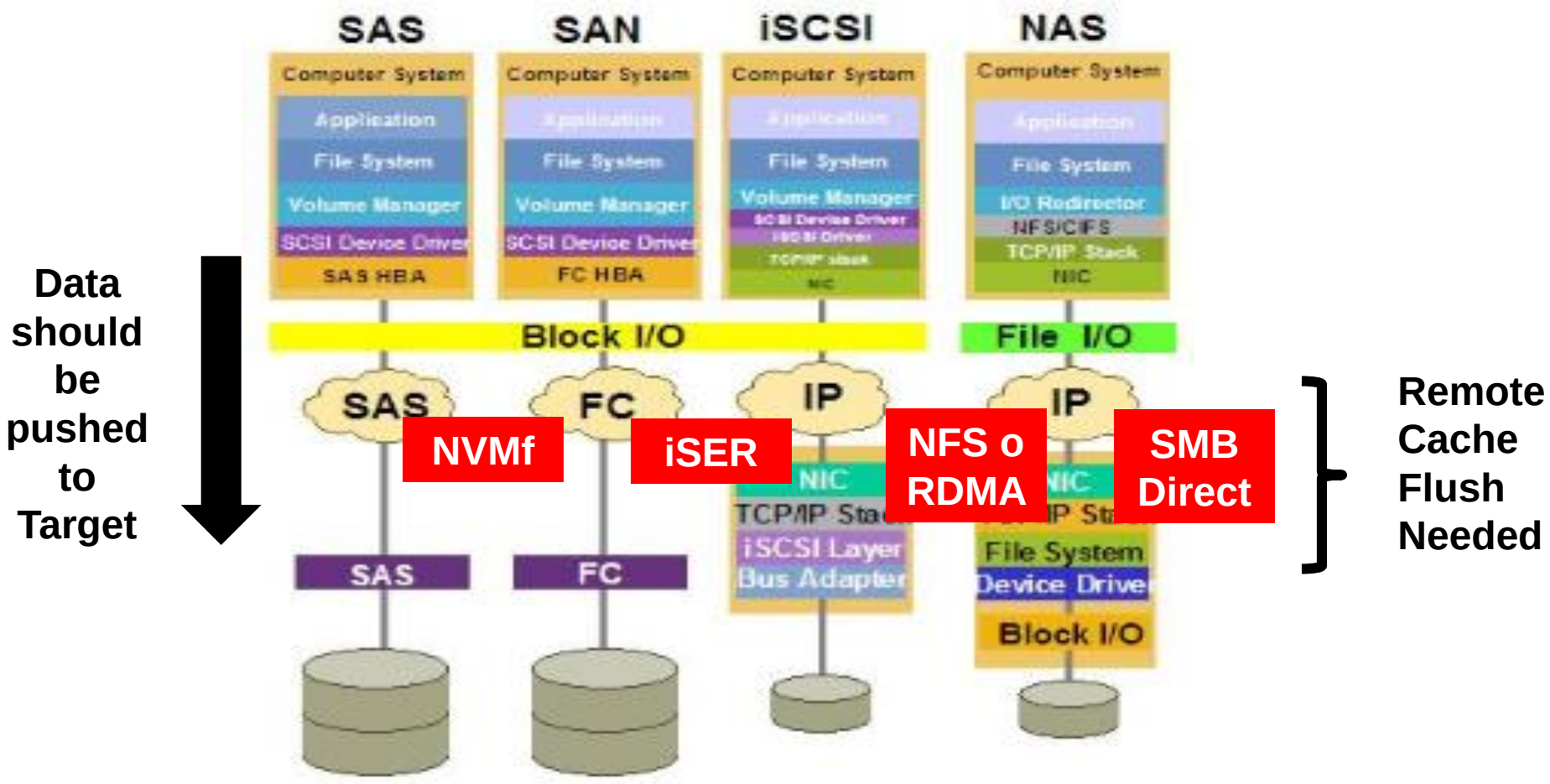


- PCM is slightly slower than DRAM but …
- Equivalent application level perf

Standards Work in Progress

- SNIA Document: NVM PM Remote Access for High Availability
  - Looking at NVM.PM.FILE mode of the SNIA NVM Programming Model
  - Focus on RDMA as the transport
    - RDMA agnostic
  - Remote flush
  - Assurance of remote durability
  - Error handling
  - Multi-tenant security

Remote PM Extensions Needed!



Summary

- PM is great technology but limited if trapped in server
- Needs to be networkable to achieve its full potential
- Many use cases driving PM over Fabric
  - Hyper Converge Infrastructure
  - Server disaggregation
- RDMA is the obvious protocol choice
- PM access times will make synchronous programming models pervasive
- Remote cache flush and the ability to push data to the targets is needed