Container World

#CONTAINERWORLD

Boosting Performance of Data Intensive Applications via Persistent Memory

Arthur Sainio

Co-Chair SNIA NVDIMM SIG

Container World

#CONTAINERWORLD

Agenda

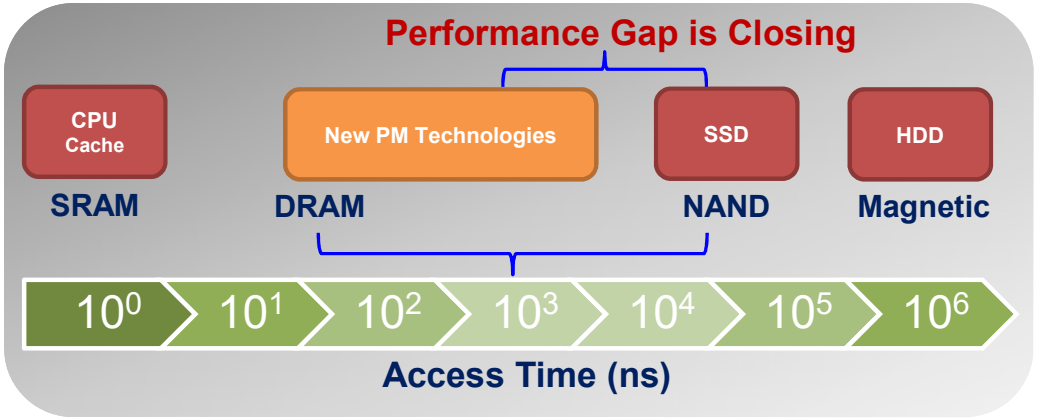
- How are NVDIMMs a revolutionary technology which will boost the performance of next-generation server and storage platforms?
- What are the ecosystem enablement efforts around NVDIMMs that are paving the way for plug-n-play adoption?
- What are the use cases and performance metrics of NVDIMMs?
- What would customers, storage developers, and the industry like to see to fully unlock the potential of NVDIMMs?
- What is the Storage Networking Industry Association (SNIA) doing to advance persistent memory?

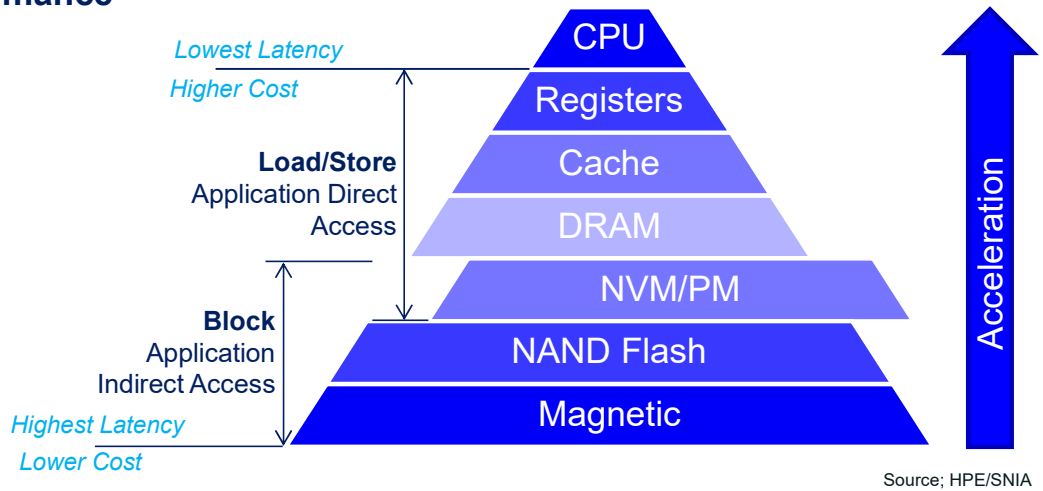
Container World

#CONTAINERWORLD

Memory – Storage Hierarchy

- Data-intensive applications need fast access to storage
- Persistent memory is the ultimate high-performance storage tier
- NVDIMM-N are a practical next-step for boosting performance





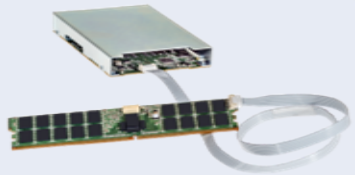
Source; HPE/SNIA

NVDIMM Types

NVDIMM-N

Standardized

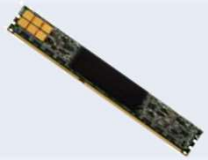
- Host has direct access to DRAM
- Cntlr moves DRAM data to Flash on power fail
- Requires backup power (typically 10's of seconds)
- Cntlr restores DRAM data from Flash on next boot
- Communication through SMBus (JEDEC std)



NVDIMM-F

Vendor Specific

- Host accesses Flash through controller
- Block-access to Flash, similar to an SSD
- Enables NAND capacity in the memory channel (even volatile operation)
- Communication through SMBus (JEDEC std TBD)



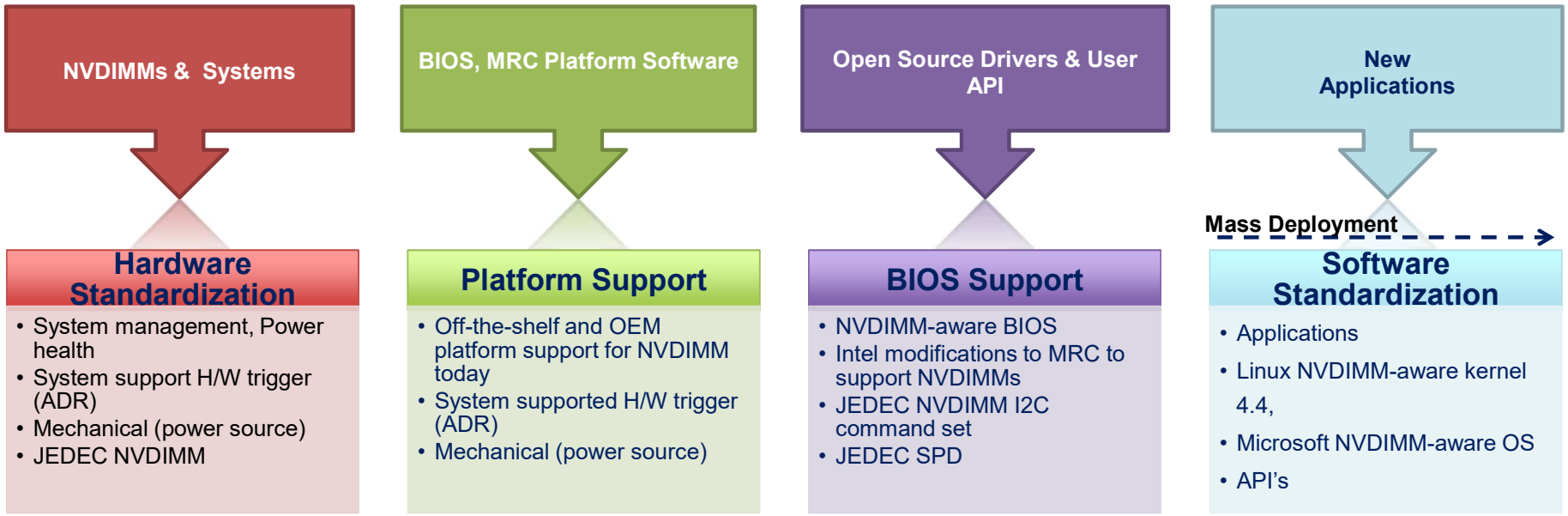
NVDIMM-P

Proposals in Progress

- Combination of -N and -F
- Host accesses memory through controller
- Definition still under discussion
- Sideband signaling for transaction ID bus
- Extended addressing for large linear addresses

DDR5 or
COMING SOON?

NVDIMM-N Ecosystem

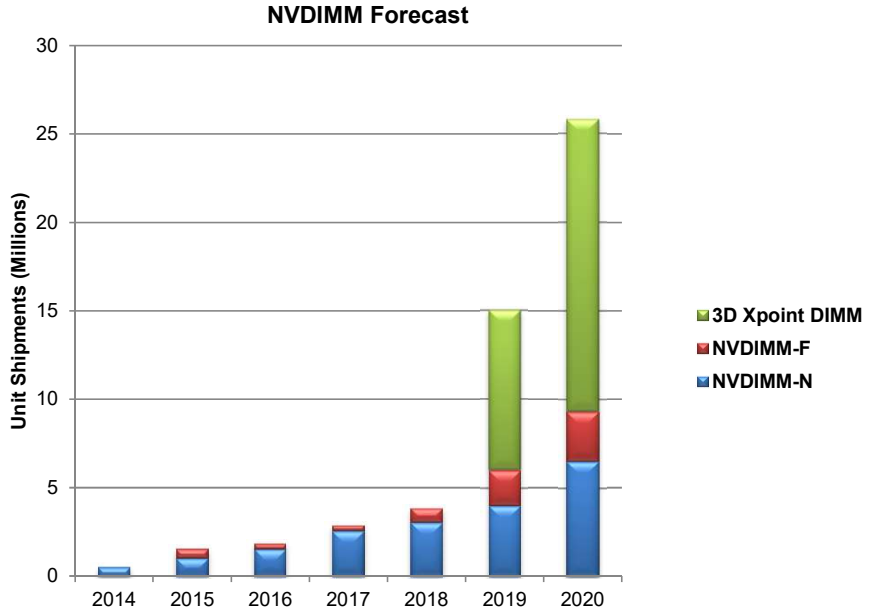


Container World

#CONTAINERWORLD

NVDIMM Outlook

- NVDIMM-N forecast based on trends and the ongoing release of more NVDIMM-N-enabled systems
- 3D Xpoint DIMM forecast may be optimistic. Assumes all 3D Xpoint parts sell in a DIMM form factor and they arrive on time and have no issues
- NVDIMM-P No forecast yet
- NVDIMM-F based on prior forecasts (from Objective Analysis)
- NVDIMM types will co-exist and support different persistent memory requirements



Objective Analysis, Jan 2017

Container World

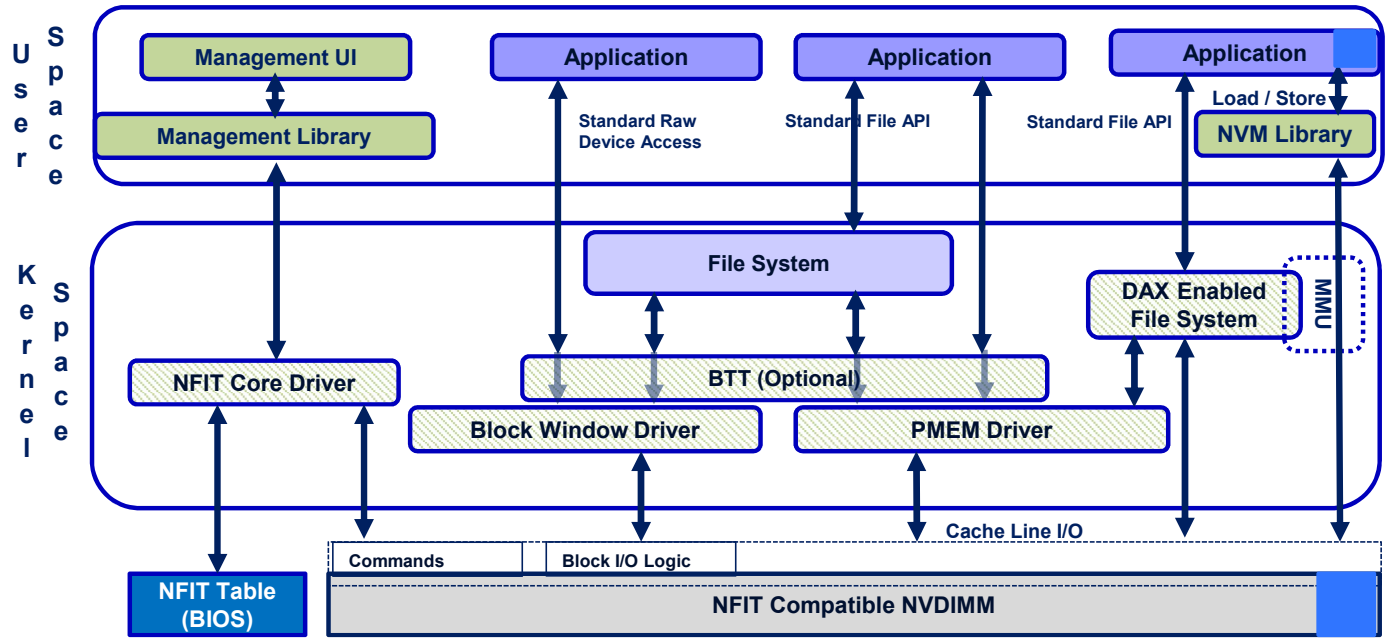
#CONTAINERWORLD

NVDIMM-N Applications

- **In Memory Database: Journaling, reduced recovery time, Ex-large tables**
- **Traditional Database: Log acceleration by write combining and caching**
- **Enterprise Storage: Tiering, caching, write buffering and meta data storage**
- **Virtualization: Higher VM consolidation with greater memory density**
- **High-Performance Computing: Check point acceleration and/or elimination**

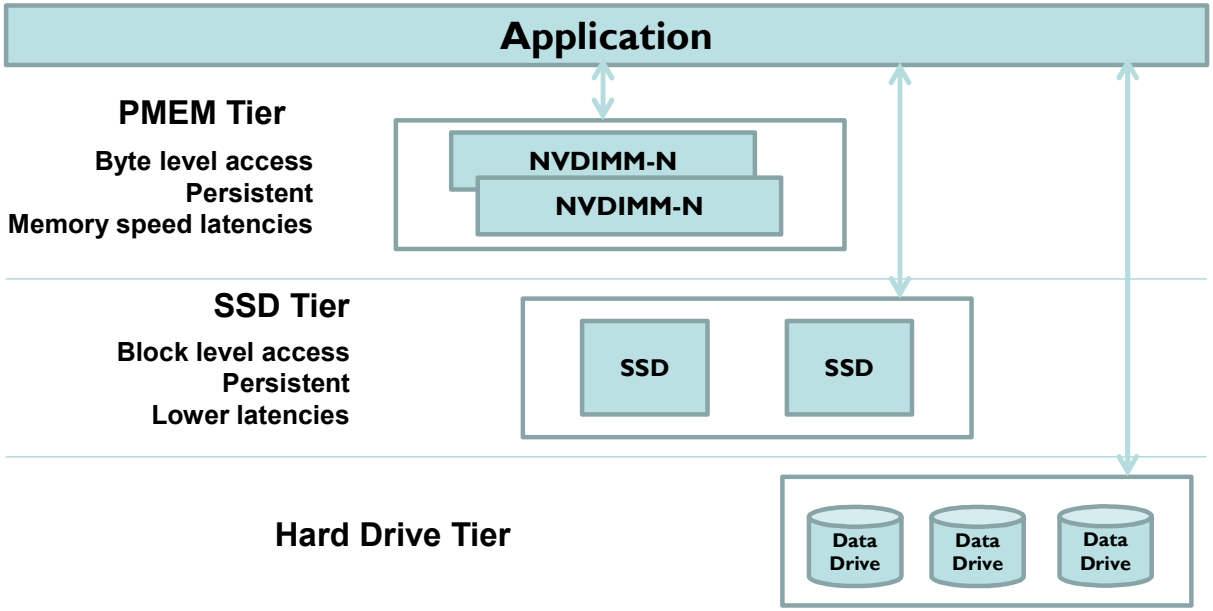
- **NVRAM Replacement: Higher performance enabled by removing the DMA setup/teardown**
- **Other: Object stores, unstructured data, financial & real-time transactions**

NVDIMM Software Architecture



Source; PMEM.IO

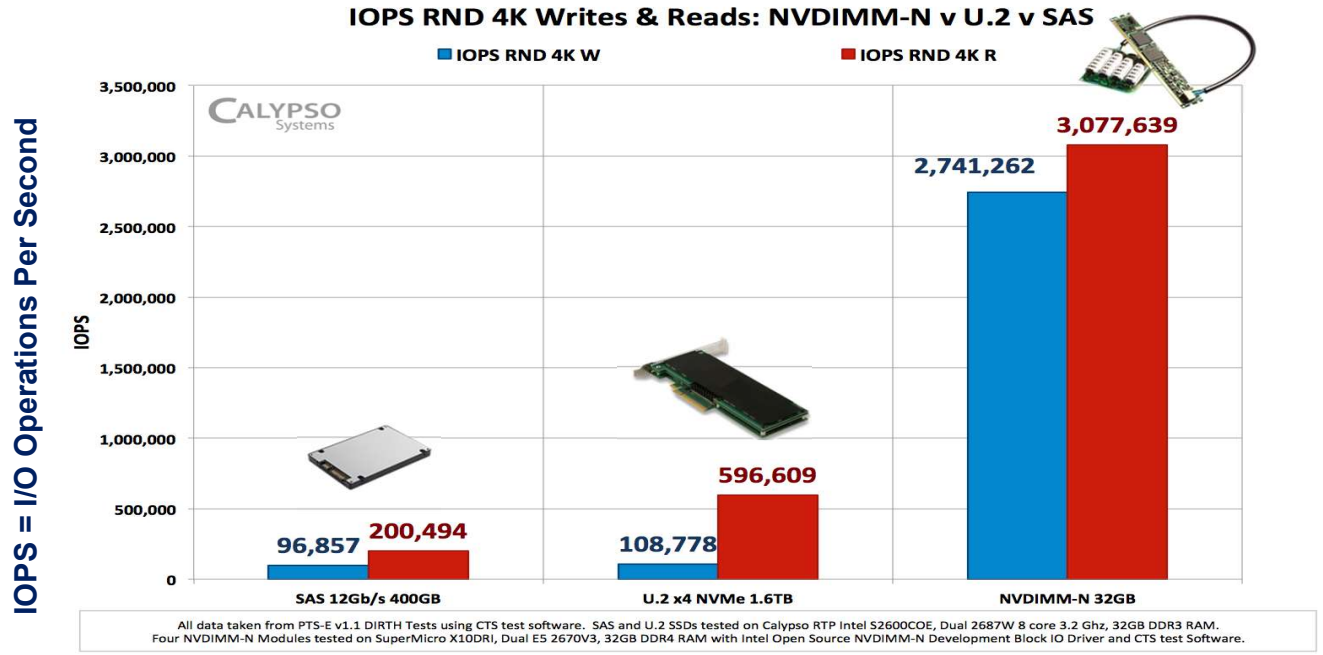
NVDIMM Use Case
Application Persistent Data Tier



Container World

#CONTAINERWORLD

NVDIMM Benchmarks



Source; Calypso

Container World

#CONTAINERWORLD

Linux Kernel 4.4+ - NVDIMM-N OS Support

- Linux 4.2 + subsystems added support of NVDIMMs. Mostly stable from 4.4
- NVDIMM modules presented as device links: /dev/pmem0, /dev/pmem1
- QEMO support (experimental)
- XFS-DAX and EXT4-DAX available

https://www.kernel.org/doc/Documentation/nvdimm/nvdimm.txt
http://pmem.io/documents/NVDIMM_Namespace_Spec.pdf

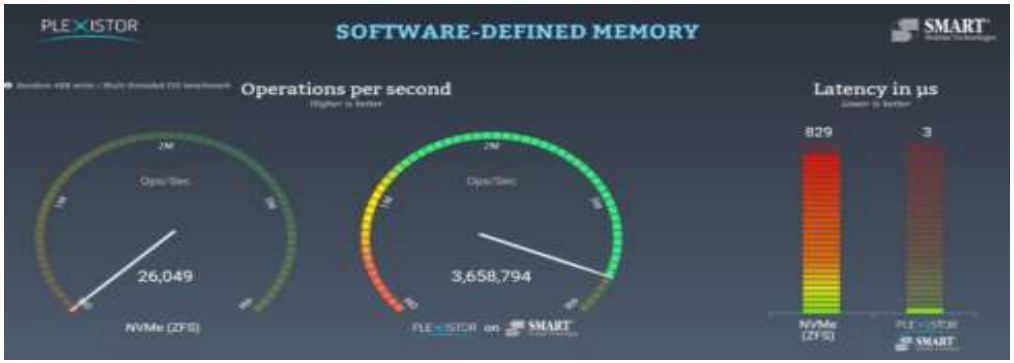
DAX	File system extensions to bypass the page cache and block layer to memory map persistent memory, from a PMEM block device, directly into a process address space.
BTT (Block, Atomic)	Block Translation Table: Persistent memory is byte addressable. Existing software may have an expectation that the power-fail-atomicity of writes is at least one sector, 512 bytes. The BTT is an indirection table with atomic update semantics to front a PMEM/BLK block device driver and present arbitrary atomic sector sizes.
PMEM	A system-physical-address range where writes are persistent. A block device composed of PMEM is capable of DAX. A PMEM address range may span an interleave of several DIMMs.
BLK	A set of one or more programmable memory mapped apertures provided by a DIMM to access its media. This indirection precludes the performance benefit of interleaving, but enables DIMM-bounded failure modes.

Container World

#CONTAINERWORLD

NVDIMM-N Benchmark Demo

- Showing performance benchmark testing using a SDM (Software Defined Memory) file system
- Compares the performance between four 16GB DDR4 NVDIMMs and a 400GB NVMe PCIe SSD
- The NVDIMMs create a byte-addressable section of persistent memory within main memory allowing for high-speed DRAM access to business-critical data
- Demo
  - Motherboard - Supermicro X10DRi
  - Intel E5-2650 V3 processor
  - Four 16GB NVDIMMs and supercap modules
  - Four 16GB RDIMMs
  - One 400GB NVMe PCIe SSD
  - Plexistor SDM file system



Container World

#CONTAINERWORLD

Microsoft WS 2016 - NVDIMM-N OS Support

- Windows Server 2016 supports DDR4 NVDIMM-N
- Block Mode
  - No code change, fast I/O device (4K sectors)
  - Still have software overhead of I/O path
- Direct Access
  - Achieve full performance potential of NVDIMM using memory-mapped files on Direct Access volumes (NTFS-DAX)
  - No I/O, no queueing, no async reads/writes
- More info on Windows NVDIMM-N support:
  - https://channel9.msdn.com/events/build/2016/p466
  - https://channel9.msdn.com/events/build/2016/p470

Microsoft

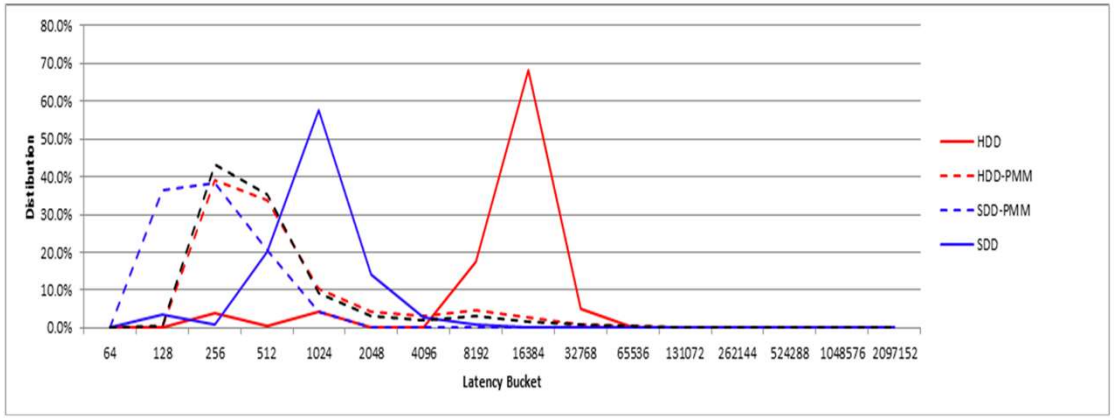
4K Random Write	Thread Count	IOPS	Latency (us)
NVDIMM-N (block)	1	187,302	5.01
NVDIMM-N (DAX)	1	1,667,788	0.52

Source; Microsoft, HPE

Application Benefits – Windows Examples

Tail of Log in SQL 2016

- **Writes updates to SQL log through persistent memory first**
- **Uses memory instructions to issue log updates to persistent memory directly**
- ***Utilizes memory-mapped files on NTFS Direct Access (DAX) volume***



		HDD	HDD-PMM	SDD-PMM	SDD
64	us]	0.0%	0.0%	0.0%	0.0%
128	us]	0.0%	0.1%	**36.3%**	3.5%
256	us]	3.9%	**39.2%**	**38.3%**	0.9%
512	us]	0.4%	**34.0%**	**20.7%**	**20.1%**
1024	us]	4.4%	**10.4%**	4.5%	**57.6%**
2048	us]	0.0%	4.2%	0.1%	**14.2%**
4096	us]	0.1%	3.0%	0.0%	2.6%
8192	us]	**17.6%**	4.7%	0.0%	0.9%
16384	us]	**68.2%**	2.6%	0.0%	0.2%
32768	us]	**5.0%**	1.0%	0.0%	0.0%
65536	us]	0.3%	0.6%	0.0%	0.0%
131072	us]	0.1%	0.1%	0.0%	0.0%
262144	us]	0.0%	0.0%	0.0%	0.0%
524288	us]	0.0%	0.0%	0.0%	0.0%
1048576	us]	0.0%	0.0%	0.0%	0.0%
2097152	us]	0.0%	0.0%	0.0%	0.0%

Source; Microsoft

Container World

#CONTAINERWORLD

NVDIMM Benchmarks

Microsoft

Flash Memory SUMMIT Performance Comparison

4K random writes
1 Thread, single core

	IOPS	Avg Latency (ns)	MB / Sec
NVMe SSD	14,553	66,632	56.85
Block Mode NVDIMM-N	148,567	6,418	580.34
DAX Mode NVDIMM-N	1,112,007	828	4,343.78

~10x

~8x

Santa Clara, CA
August 2016

14

Hewlett Packard Enterprise

Flash Memory SUMMIT Comparing NVDIMM Performance to Flash

Performance Measurement	NVDIMM vs SAS SSD	NVDIMM vs PCIe Workload Accelerator
IOs Per Second (IOPs)	34x more IOPs	24x more IOPs
Bandwidth	16x greater Bandwidth	6x greater Bandwidth
Latency	81x lower Latency	73x lower Latency

HPE NVDIMM technology promises to unlock new levels of HPE ProLiant performance

Hewlett Packard Enterprise

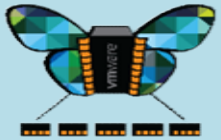
Source; Microsoft, HPE

Container World

#CONTAINERWORLD

VMware NVDIMM Program for ISVs

vSphere-based NVDIMM Emulation Vehicle



- Available Now
- Emulates all of the capabilities of NVDIMMs from different vendors
- Works with off-the-shelf commercial servers

To Get Emulation Vehicle

Join VMware NVDIMM Program

Contact VMware: PMEM@vmware.com

↓

Sign program documents

↓

Get free emulation vehicle; free support from VMware & NVDIMM partner

↓

Reference ISV (e.g. quote, logo, etc.)

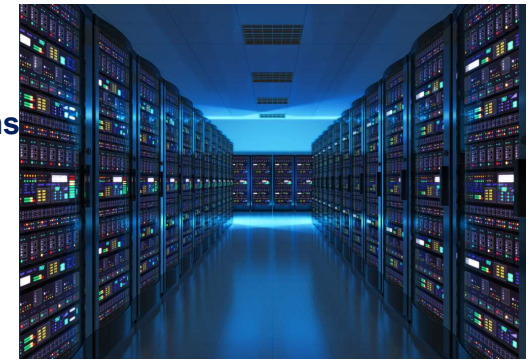
Source; VMWare

Container World

#CONTAINERWORLD

What Customers, Storage Developers, and the Industry Would Like to See to Fully Unlock the Potential of NVDIMMs

- Standardization and Interoperability
  - Standard server and storage motherboards enabled to support all NVDIMM types
  - Standardized BIOS/MRC, driver, and library support
  - Interoperability between MBs and NVDIMMs
  - Standardized memory channel access protocol adopted by Memory Controller implementations
  - O/S recognition of APCI 6.0 (NFIT) to ease end user application development
- Features
  - Data encryption/decryption with password locking JEDEC standard
  - Standardized set of OEM automation diagnostic tools
  - NVDIMM-N Snapshot: JEDEC support of NMI trigger method alternative to ADR trigger
- Performance
  - Standardized benchmarking and results
  - Lower latency I/O access < 5us



Container World

#CONTAINERWORLD

SNIA-At-A-Glance

160
unique member companies

3,500
active contributing members

50,000
IT end users & storage pros worldwide

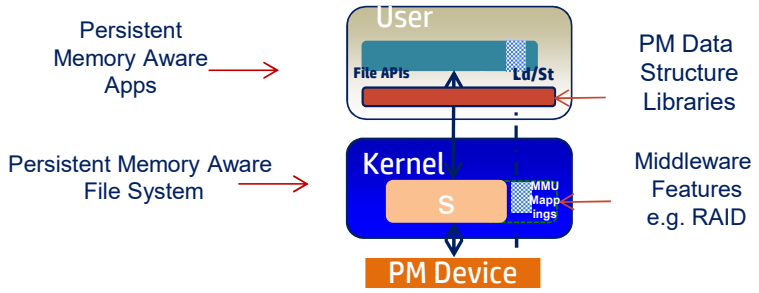
Learn more: **snia.org/technical**

@SNIA

SNIA Activities Advancing Persistent Memory Access and Use

SNIA NVM Programming Model – Enabling Persistent Memory Access

- Describes application visible behaviors
- Allows API's to align with OS's
- Exposes opportunities in networks and processors
- SNIA 2017 work activity
  - V1.2 of Model in progress
  - V1.1 and 1.0 of Model available at snia.org/forums/sssi/nvmp
  - Atomicity and Remote Access WP published
  - Security Threat Model WIP



SNIA NVDIMM Special Interest Group – Powerful Persistent Memory is Here

- SIG contributes to:
  - Common PM Specifications
  - Common PM Messaging
  - Common PM Taxonomy
  - PM Ecosystem Development

NVDIMM-N

- ✓ Memory-mapped DRAM
- ✓ JEDEC-ratified
- ✓ Easily exploited in **Microsoft Windows Server 16** and **Linux** for extremely high performance read/write workloads, such as SQL

Container World

#CONTAINERWORLD

Thank You!

Learn more about Persistent Memory, including NVDIMMs, at

www.snia.org/nvdimm