Flash Memory Summit 2019

Persistent Memory and NVDIMMs

Presented by Arthur Sainio
SNIA PM&NVDIMM SIG Co-Chair
Director, SMART Modular Technologies

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.

© 2019 Storage Networking Industry Association. All Rights Reserved.

Contents

- Persistent Memory
- NVDIMM
- Applications
- Standards Work
- More on Persistent Memory at FMS

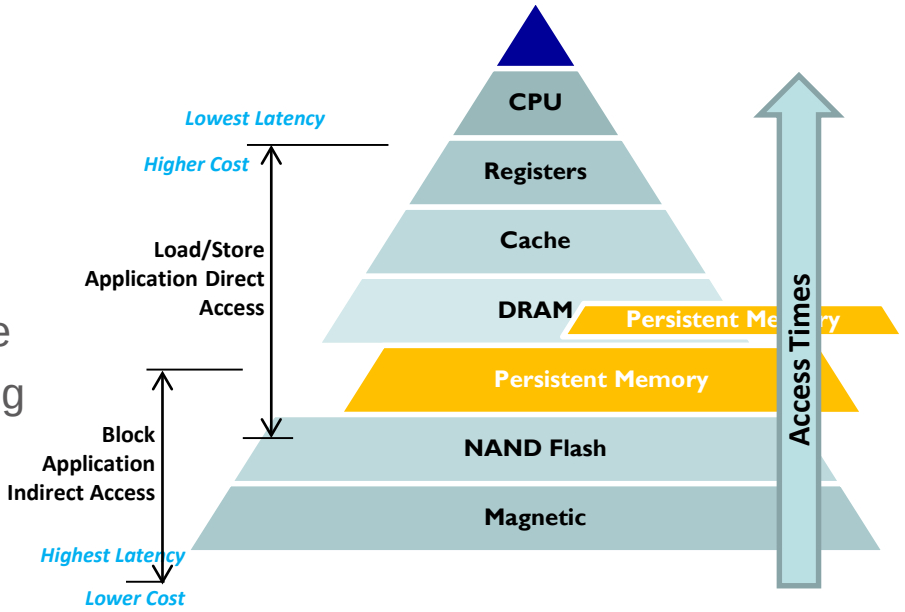
© 2019 Storage Networking Industry Association. All Rights Reserved.

Persistent Memory

What Is Persistent Memory? Why Is It Important?

- Persistent Memory is
  - Non-Volatile
  - Byte Addressable
  - Low Latency (<1µs)
  - Densities greater than or equal to DRAM (for wide-scale adoption)

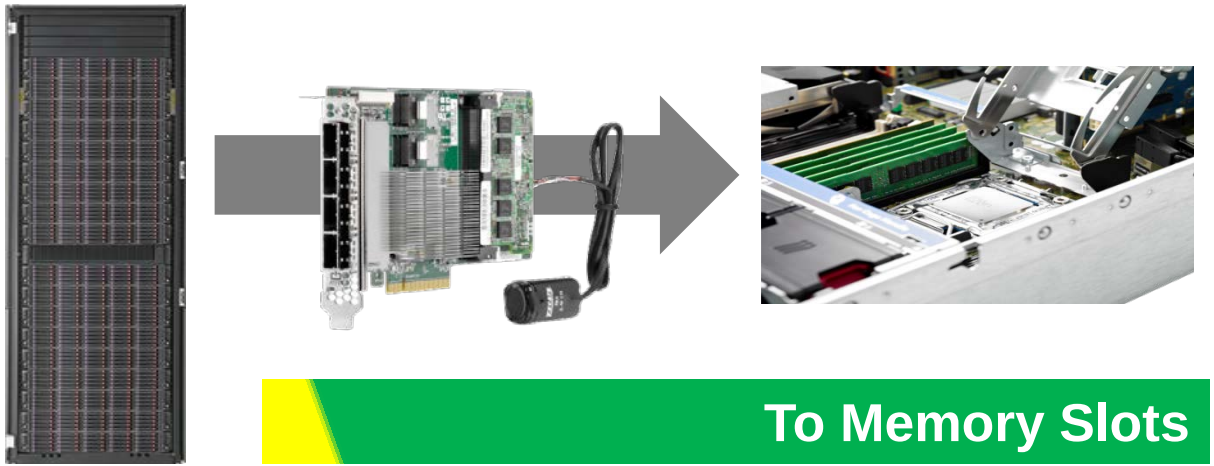
- Why is it important?
  - Dramatically increases system performance
  - Enables a fundamental change in computing architecture
  - Apps, middleware and OSs are no longer bound by file system overhead in order to run persistent transactions



© 2019 Storage Networking Industry Association. All Rights Reserved.

Persistent Memory Vision

Persistent Memory Brings Storage



To Memory Slots

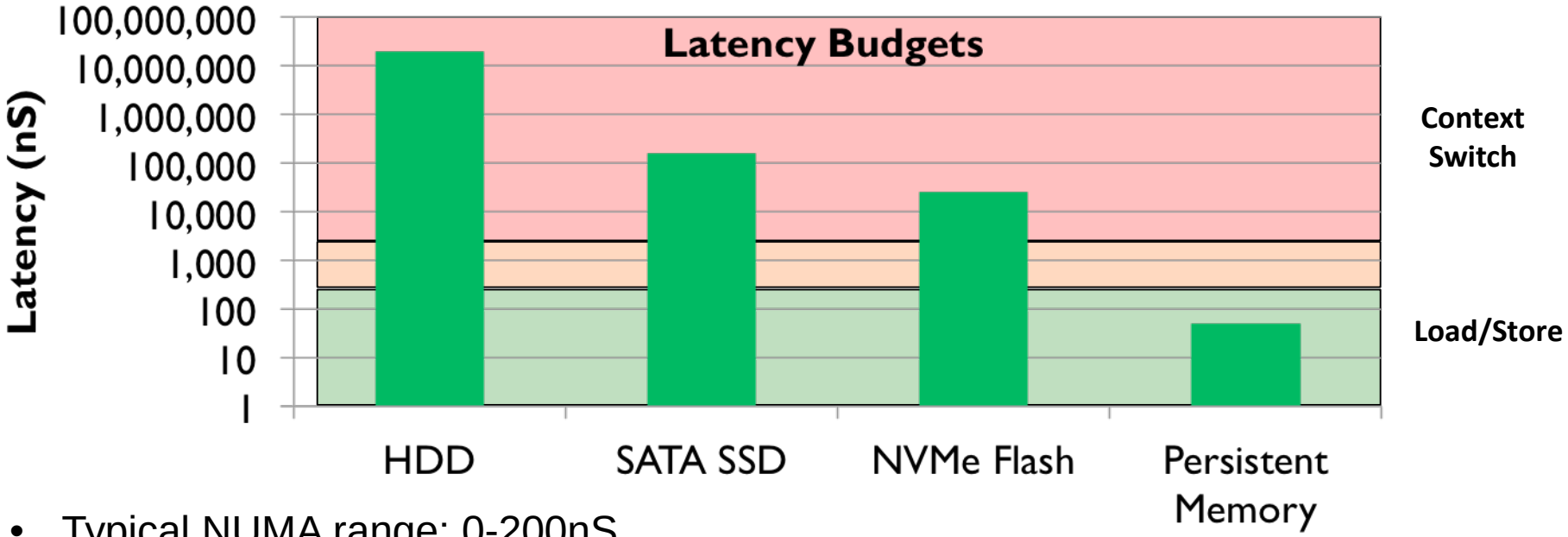
Fast
Like Memory

Persistent
Like Storage

- For system acceleration
- For real-time data capture, analysis and intelligent response

© 2019 Storage Networking Industry Association. All Rights Reserved.

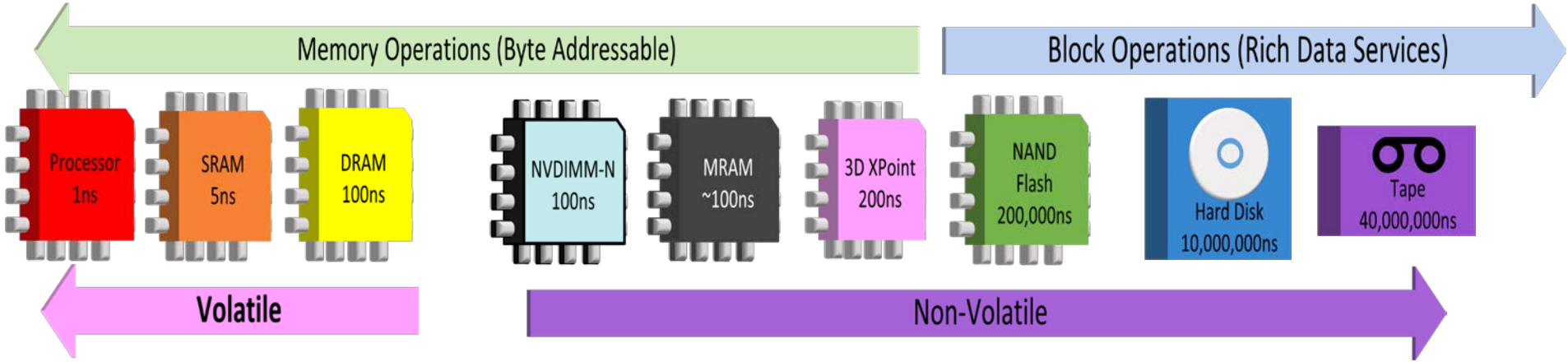
Storage vs. Memory



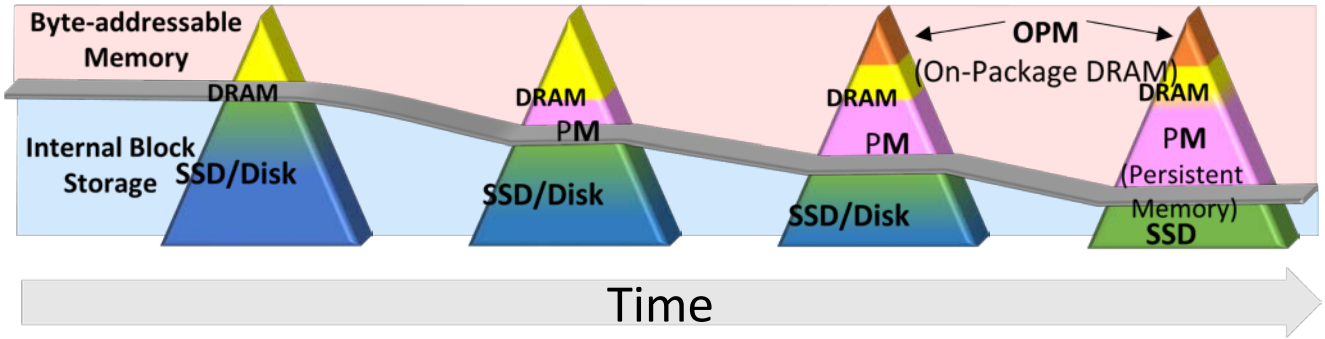
- Typical NUMA range: 0-200nS
- Typical context switch range: above 2-3µS

© 2019 Storage Networking Industry Association. All Rights Reserved.

Memory and Storage are Converging



- Memory semantic operations become predominant (Volatile & Non-Volatile)



Source: Modified from Gen-Z Consortium 2017

© 2019 Storage Networking Industry Association. All Rights Reserved.

Emerging Technologies

- Multiple Persistent Memory technologies nearing commercialization
- Phase Change (PCM): a middle ground between DRAM and Flash
- MRAM: DRAM replacement? density past 8Gb, lower idle power
- ReRAM: Flash replacement? High density, better endurance
- CNTRAM: Carbon Nanotube based memory – another DRAM replacement?

© 2019 Storage Networking Industry Association. All Rights Reserved.

Source: FMS, Jim Harrell, Aug'18

Technology Comparison

Technology	FeRAM	MRAM	ReRAM	PCM	3D Xpoint	NAND Flash	DRAM NVDIMM
Endurance	10^{12}	10^{12}	10^{6}	10^{8}	10^{6} - 10^{7}	10^{3}	10^{15}
Byte Addressable	yes	yes	yes	yes	yes	no	yes
Latency R/W	70ns-100ns	70ns/70ns	100ns/100µs	20ns/65ns	100ns/500ns	10µs/10µS	40-140ns
Power Consumption	Low	Medium/ Low	Low	Medium	Medium	Low	Medium
Interface	DRAM	DDR3 DDR4	Flash-like	Proprietary	Proprietary	Toggle ONPHI	DDR3 DDR4
Density Path	Low	Gigabit+	Terabit	64Gb+	64Gb+	Gigabit+	Gigabit+

© 2019 Storage Networking Industry Association. All Rights Reserved.

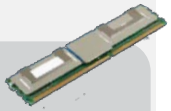
Existing and Emerging Variations of Persistent Memory Products

NVDIMM-N
(DRAM/NAND)

Performance Optimized PM

RDIMM/LRDIMM
(DRAM)

NVDIMM-P
(DRAM/NVM)



DIMMs
(NVM)

Cost Optimized PM

Cost

NVME SSD
(NVM)

NVME SSD
(NAND)

SATA/SAS SSD
(NAND)

Hard Disk Drive
(Spinning Media)

Nanoseconds **Microseconds** **Milliseconds**

Access Time

© 2019 Storage Networking Industry Association. All Rights Reserved.

Source: Modified from Micron, FMS 2018

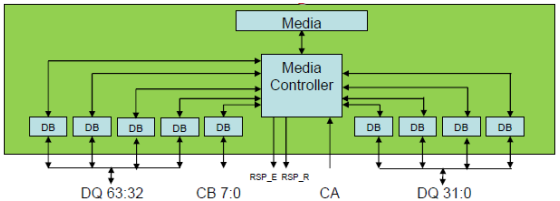
NVDIMM

Persistent Memory - NVDIMMs

NVDIMM-N

- Host has direct access to DRAM
- Controller moves DRAM data to Flash on power fail
- Requires backup power
- CNTLR restores DRAM data from Flash on next boot
- Communication through SMBus
- Byte-addressable DRAM for lowest latency with NAND for persistence backup

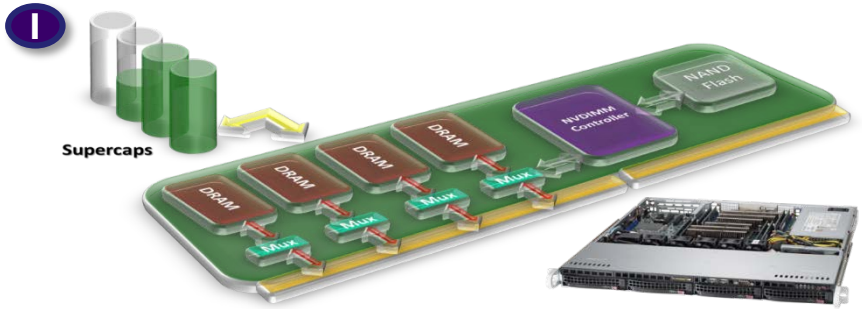
NVDIMM-P



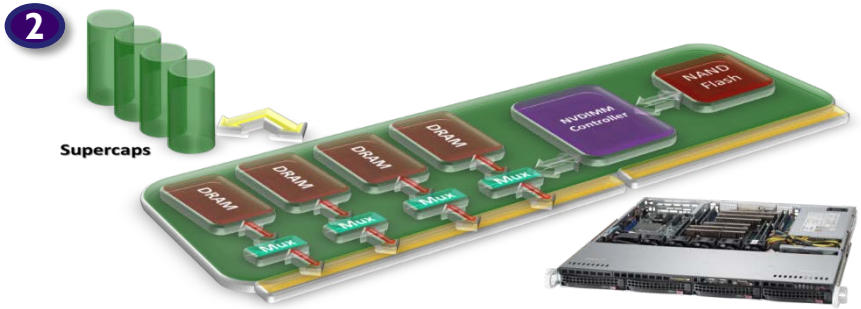
- NVDIMM-P interface specification targeting persistent memories and high capacity DRAM memory on DDR4 and DDR5 channels
- Extends the DDR protocol to enable transactional access
- Host is decoupled from the media
- Multiple media types supported
- Supports any latency (ns ~ us)
- JEDEC specification publication in 2019

© 2019 Storage Networking Industry Association. All Rights Reserved.

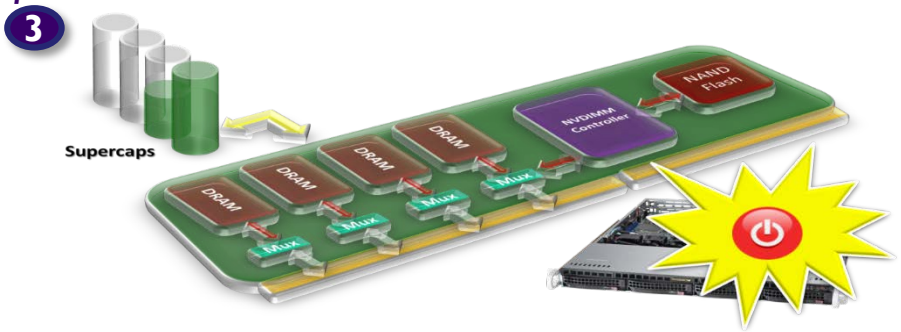
NVDIMM-N How It Works



- *Plugs into JEDEC Standard DIMM Socket*
- *Appears as standard RDIMM to host during normal operation*
- *Supercaps charge on power up*



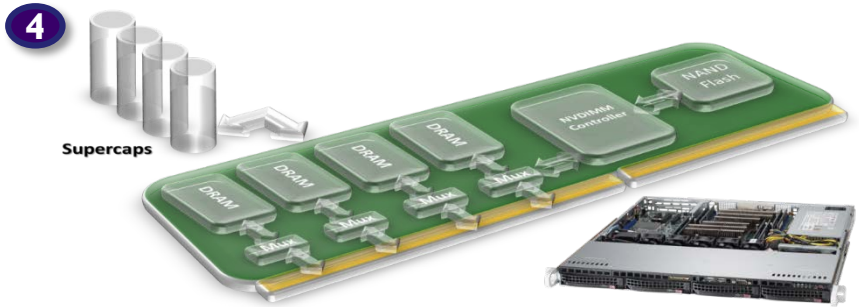
- *When health checks clear, NVDIMM can be armed for backup*
- *NVDIMM can be used as persistent memory space by the host*



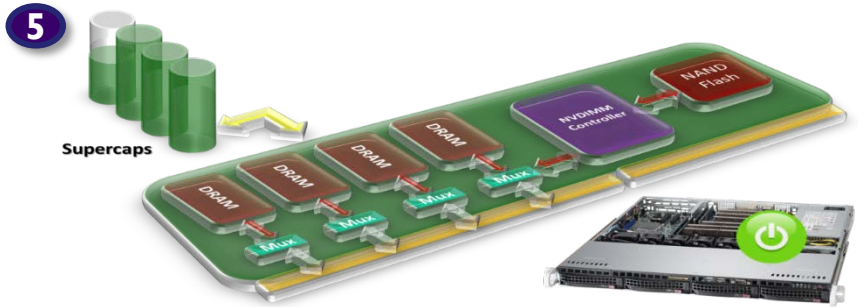
- *During unexpected power loss event, DRAM contents are moved to NAND Flash using Supercaps for backup power*

© 2018 Storage Networking Industry Association. All Rights Reserved.

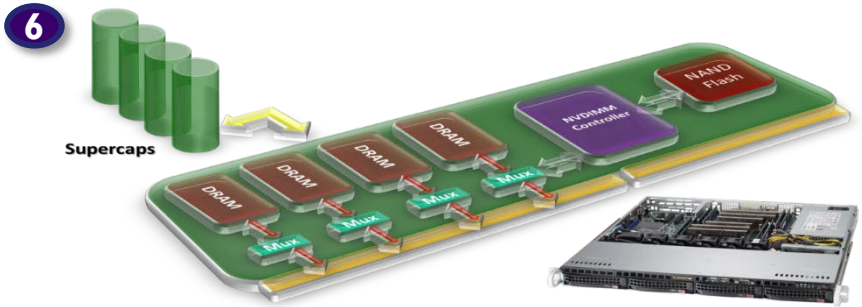
NVDIMM-N How It Works



- *When backup is complete, NVDIMM goes to zero power state*
- *Data retention = NAND Flash spec (typically years)*



- *When power is returned, DRAM contents are restored from NAND Flash*
- *Supercaps re-charge in minutes*

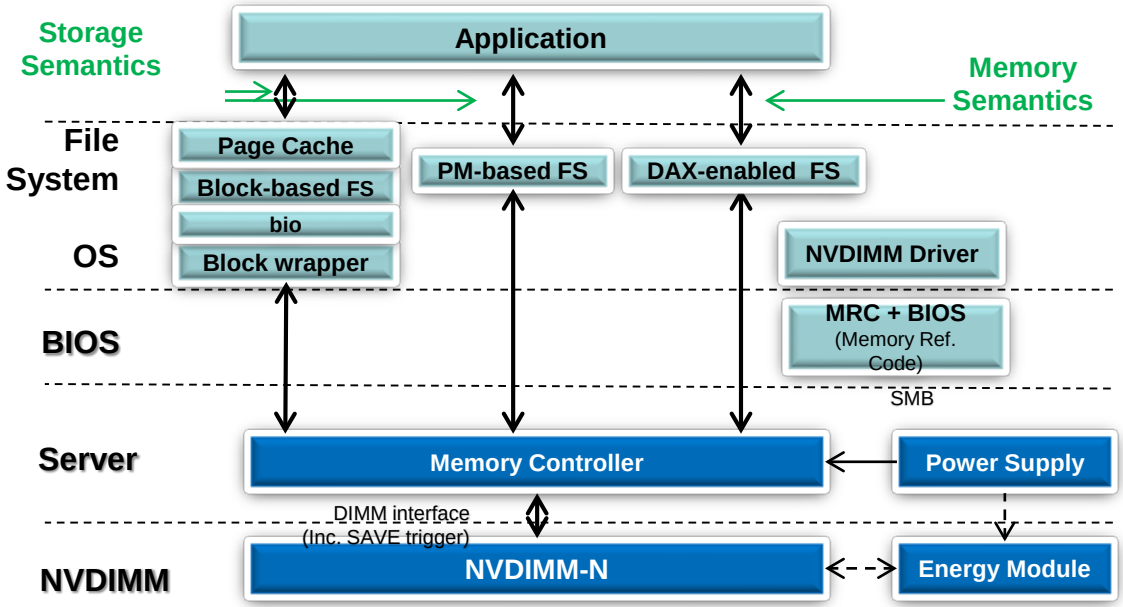


- *DRAM handed back to host in restored state prior to power loss*

© 2018 Storage Networking Industry Association. All Rights Reserved.

NVDIMM Ecosystem

- Standardized through NFIT and JEDEC
- Linux 4.4+ kernels have the software stack
- Open source library is available for applications



© 2019 Storage Networking Industry Association. All Rights Reserved.

NVDIMM-N BIOS/MRC Support Functions

NVDIMMs BIOS/MRC (Memory Reference Code)

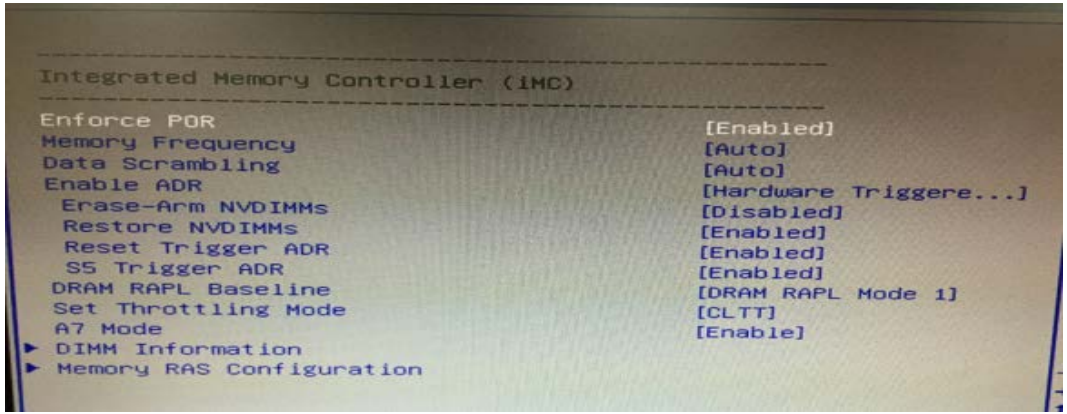
1. Detect NVDIMMs
2. Setup Memory Map
3. ARM for Backup
4. Detect AC Power Loss or BMC/CPLD Triggered ADR
5. Flush Write Buffers
6. RESTORE Data On Boot
7. Enable I2C R/W Access

© 2018 Storage Networking Industry Association. All Rights Reserved.

Source: SNIA Persistent Memory and NVDIMM SIG

Additional BIOS Settings

- BIOS also presents various menu options to setup NVDIMM operation
- Configuration:
  - Erase-Arm NVDIMM
  - Restore NVDIMM
  - Reset Trigger ADR
  - S5 Trigger ADR



© 2018 Storage Networking Industry Association. All Rights Reserved.

Linux Kernel 4.4+ NVDIMM-N OS Support

SNIA®

- Linux 4.2 + subsystems added support of NVDIMMs. Mostly stable from 4.4 **(now at 5.0)**
- NVDIMM modules presented as device links: /dev/pmem0, /dev/pmem1
- QEMO support (experimental)
- XFS-DAX and EXT4-DAX available

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.

© 2019 Storage Networking Industry Association. All Rights Reserved.

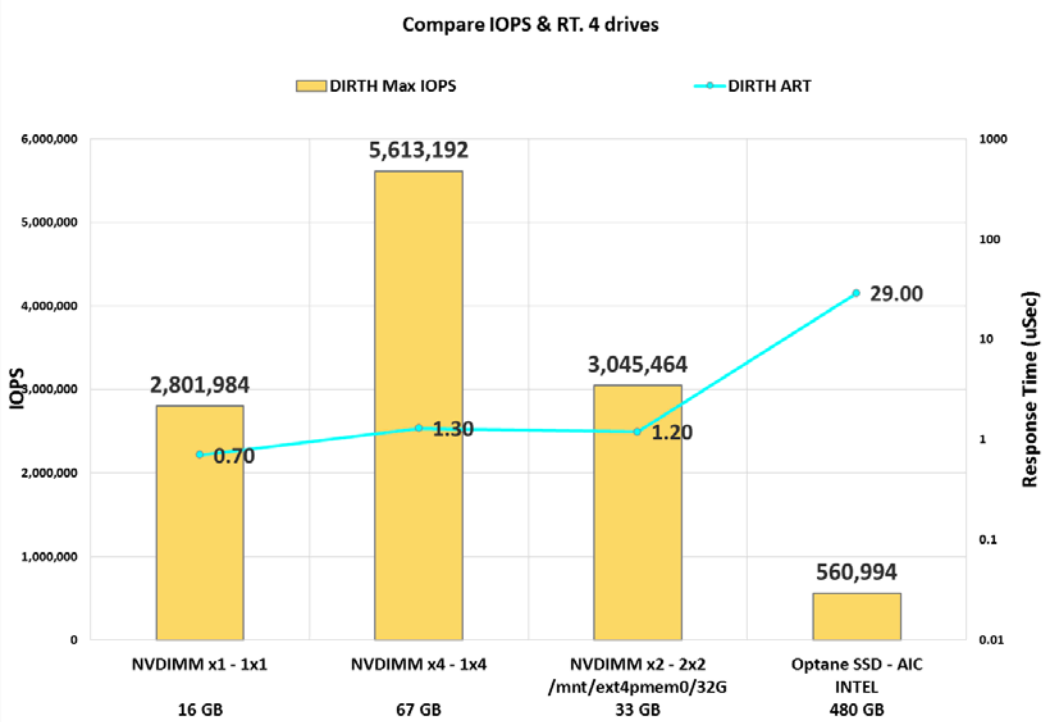
Windows NVDIMM-N OS Support

- Windows Server since 2016 supports DDR4 NVDIMM-N (now 2019)
- 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

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

© 2019 Storage Networking Industry Association. All Rights Reserved.

NVDIMM Performance Comparison



- **Test Platform: Supermicro X11DRI 16GB DDR4 2400 Mhz RDIMM RAM, Intel XEON 8160 2.1 Ghz 24 core, 16 GB DDR4 JEDEC NVDIMM-N. 480GB Optane SSD**
- **Software: Ubuntu 16.04.3 LTS Linux 4.10.0-28; DAX File System**
- **Test Software: Calypso CTS 7.0 fe 1.26.25 be 1.9.317**

© 2019 Storage Networking Industry Association. All Rights Reserved.

How NVDIMM-N's Improve Performance

- NVDIMM-Ns are byte addressable. This allows databases to be built in memory
- With direct access to records this removes disk IO and all the overhead that involves
- A memcached structure is much faster than the best solid-state solution with updates just requiring a register-to-memory computer instruction instead of the file stack and interface overhead
- Since this looks like DRAM to the system, using RDMA to create redundancy and cluster sharing is a given, with existing designs working just fine

© 2019 Storage Networking Industry Association. All Rights Reserved.

NVDIMM-N Performance

- NVDIMMs provide 34 times the number of IOPS compared with standard SSDs, with 16 times the bandwidth and 81 times lower latency
- Streaming data applications can be architected to greatly benefit from this marriage of memory and storage

Source: Microsoft

© 2019 Storage Networking Industry Association. All Rights Reserved.

Persistent Memory Applications

What Applications use PM?

- Applications that have a large working set of data with a need for persistence
  - Using NVMe or standard SSDs add latency
  - Decreasing the latency to avoid disk access
- In Memory Databases
  - Application driven data locality
  - Newer DB adaptations beginning to use PM
- Productivity Improvements
  - Software infrastructure is enabled
  - Standard libraries are available

© 2019 Storage Networking Industry Association. All Rights Reserved.

NVDIMM Use Cases

Enterprise Storage
Tiering, caching,
write buffering,
meta data storage

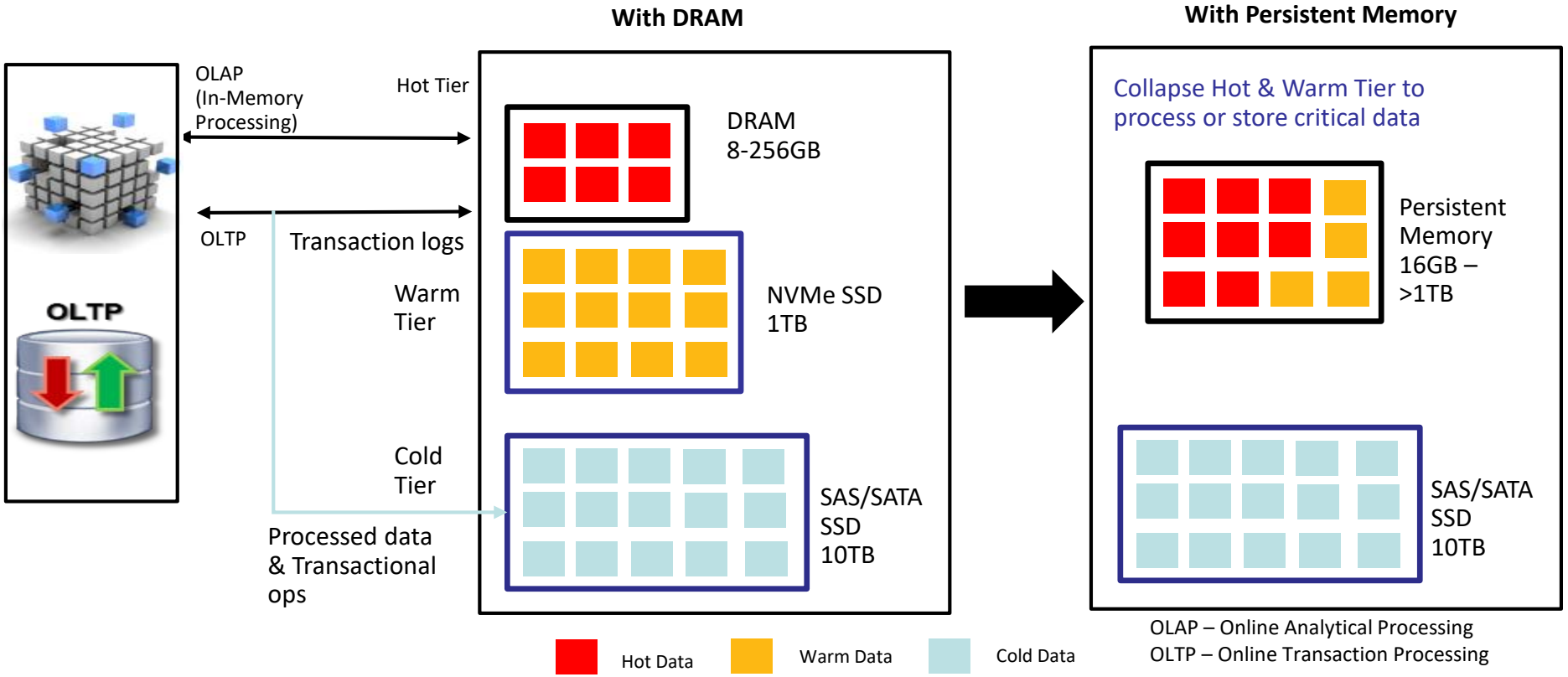
Traditional Database
Log acceleration
by write combining
and caching

In-Memory Database
Journaling,
recovery time, tables

High-Performance Computing
Check point
acceleration
and/or elimination

© 2019 Storage Networking Industry Association. All Rights Reserved.

Evolution of In-Memory Apps with Persistent Memory



© 2019 Storage Networking Industry Association. All Rights Reserved.

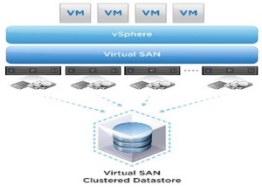
Source: Modified from VMWare, FMS 2018

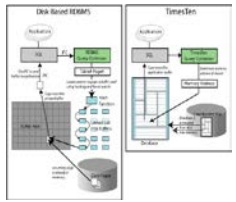
PM System Support and Applications

A sample of companies showing PM support

© 2019 Storage Networking Industry Association. All Rights Reserved.

Persistent Memory Adds Value Across Diverse Applications





Relational Database

- MSFT SQL
- MySQL
- Maria DB
- Oracle

Log acceleration: write combining and caching

Scale-out Storage

- Vmware VSAN
- MSFT Azure
- Store Virtual

Tiering, caching, write buffering, meta-data storage

Virtual Desktop Infrastructure

- Vmware VDI
- Citrix HDI

Higher VM consolidation

Big Data

- Mongo DB
- Cloudera
- HortonWorks
- Hadoop
- Cassandra
- MSFT SQL Hadoop

Higher performance

In Memory Database

- SAP HANA
- MSFT SQL Hekaton
- XAP Gigaspace

Journaling, Transaction logs

Middleware

- Java
- .NET

Optimized abstraction

HPC

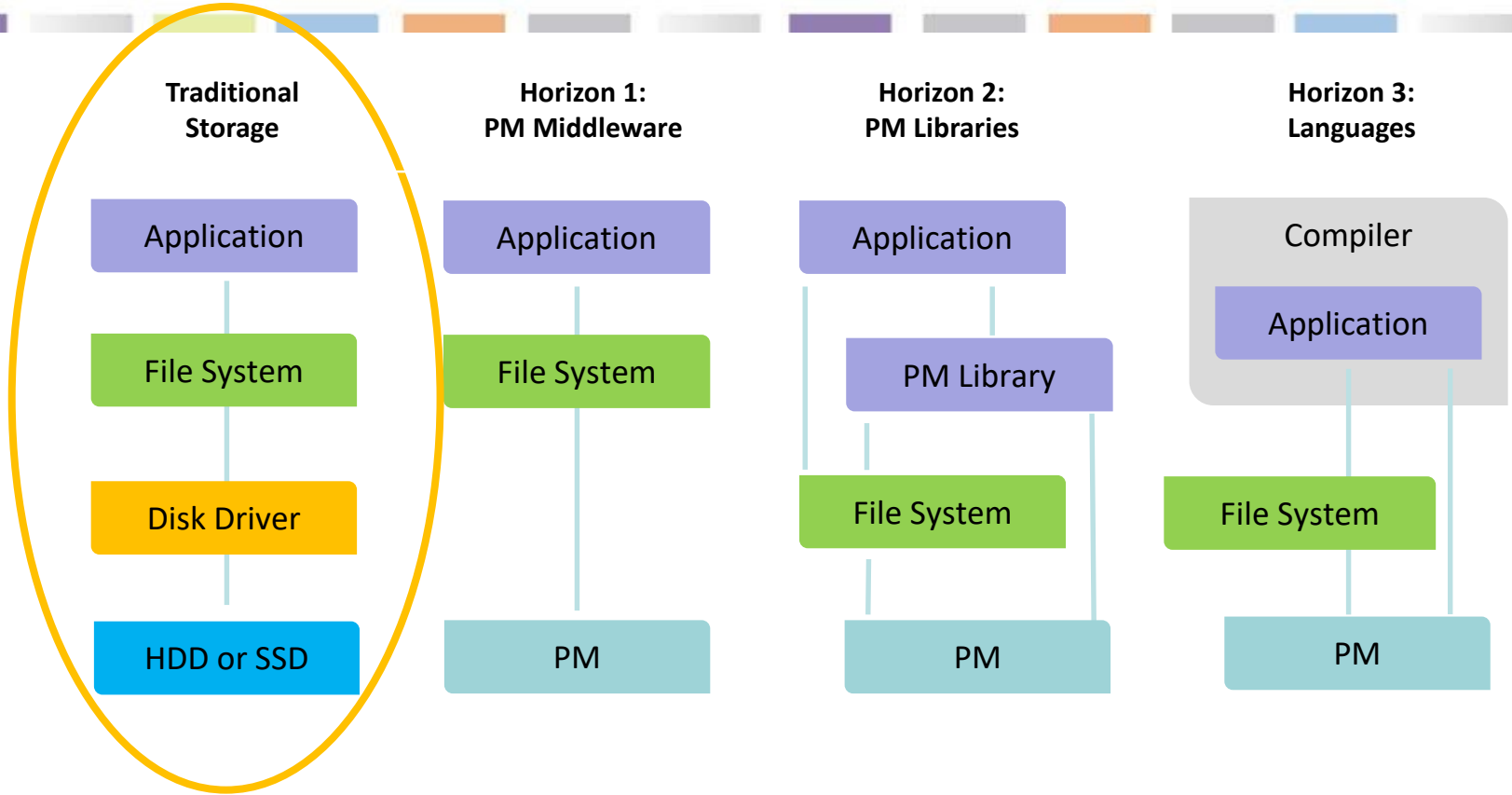
- HPC

Check point acceleration

© 2019 Storage Networking Industry Association. All Rights Reserved.

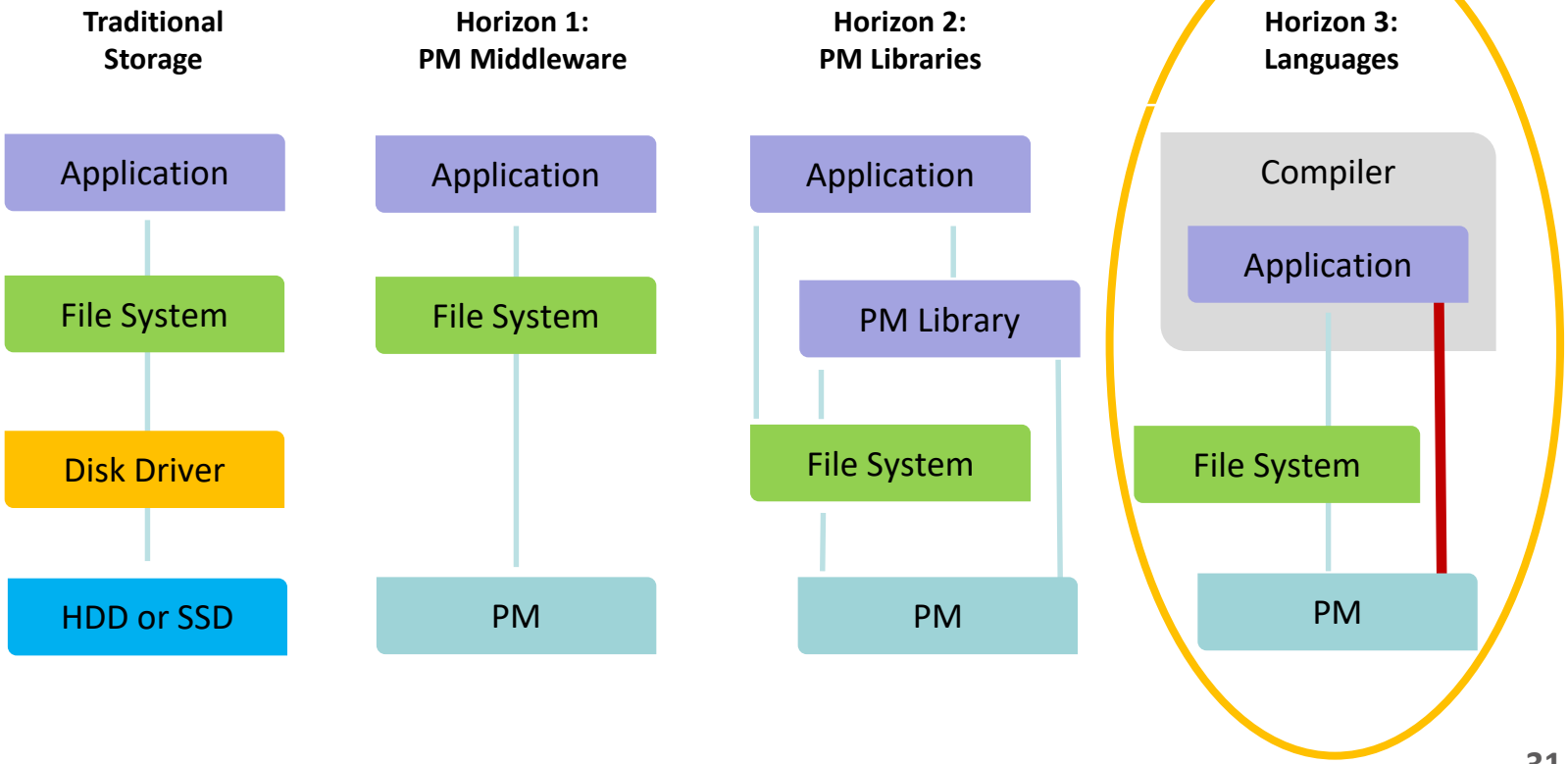
Source: Micron HPE Discover 2018

Application Horizons - Today



© 2019 Storage Networking Industry Association. All Rights Reserved.

Application Horizons – Ultimate Goal



© 2019 Storage Networking Industry Association. All Rights Reserved.

NVDIMM-N and 3DXPoint Applications

- Many NAND flash storage array vendors are using NVDIMM-N modules for write acceleration and commit logging
  - These applications do not require a density higher than multiple GBs so they are well-suited for NVDIMM-N
- 3DXPoint is well-suited for PM applications like In-Memory databases that need 100’s of GBs to TBs of persistent memory that is used in combination with DRAM

© 2019 Storage Networking Industry Association. All Rights Reserved.

Example: Need for In Memory Persistent Database

DreamWorks

- 600TB's of data in one film
- Many small items in a large working set
- Substantial re-use and repeat file I/O
- Expensive to compute and convert
- Distributed clients doing similar things
- Writes are immutable; lockless updates

Goal with PM

- NVDIMMs in each workstation and server
- Accelerate local workflows
- Cluster of Persistent Memory servers
- Software stack that provides RPM-as-a-Service
- A way for apps to persist things and reduce trips through the storage stack
- A way for apps to find and get things
- That behaves like named shared memory

© 2019 Storage Networking Industry Association. All Rights Reserved.

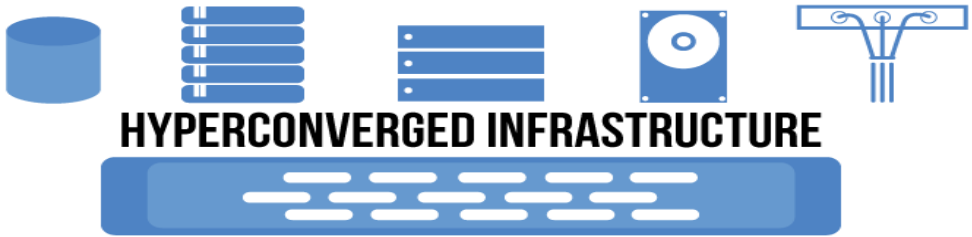
Source: Modified from DreamWorks, FMS, Aug'18

Example: Using Persistent Memory to Accelerate HCI Storage Performance

Differentiated value with Persistent Memory in HCI storage tier

Create a new persistent memory tier for metadata (benefits ALL apps)

1. Read-modify-write with persistent memory as byte addressable is 100X faster than block storage
2. Faster metadata access for dedup, checksum etc results in reduced CPU utilization and higher IOPS for all apps
3. Faster reboots due to persistence of metadata in persistent memory (save time for not having to rebuild metadata from logs)



© 2019 Storage Networking Industry Association. All Rights Reserved.

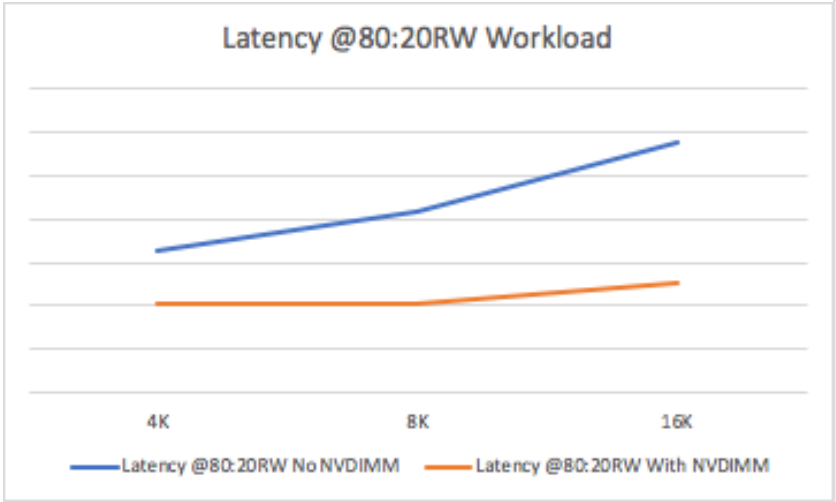
Source; Modified from VMWare, FMS 2018

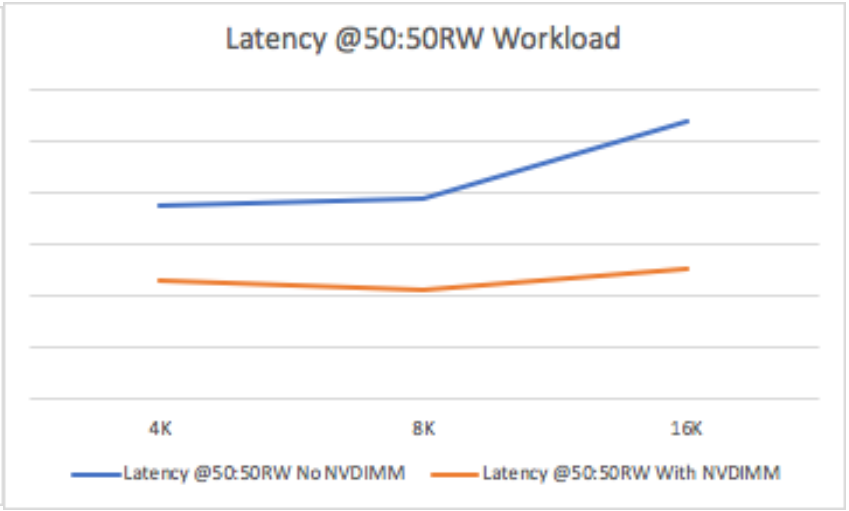
Example: WDC IntelliFlash Write Cache

- Separate logging for incoming writes
- Write is acknowledged after persisting to the write cache
- Coalesced I/O is flushed to drives after dedupe and compression
- Uses high performance media as the latency is crucial for many applications like DBT and OLTP
- Best fit for NVDIMM

© 2019 Storage Networking Industry Association. All Rights Reserved.

Example Results – Latency Comparison





- All flash array with 24TB capacity
- iSCSI protocol
- fio with 4 clients and 8 LUNs

Source: WDC, FMS 2018

© 2019 Storage Networking Industry Association. All Rights Reserved.

Infrastructure Changes

- Operating System
- File system changes for memory mapped files
- Memory Management software
- Hypervisors
- Allocation of Persistent Memory to Guests
- Coordinating with Guest's use of Persistent Memory
- User space libraries supporting Persistent Memory
- Support for legacy interfaces with Persistent Memory-aware implementations
- Securing application data in a multi-tenant environment

© 2019 Storage Networking Industry Association. All Rights Reserved.

Persistent Memory Standards

- **JEDEC JESD 245, 245B: Byte Addressable Energy Backed Interface**
  - Defines the host to device interface and features supported for a NVDIMM-N

- **ACPI 6.2**
  - NVDIMM Firmware Interface Table (NFIT)
  - NVM Root and NVDIMM objects in ACPI namespace
  - Address Range Scrub (ARS)
  - Uncorrectable memory error handling
  - Notification mechanism for NVDIMM health events and runtime detected uncorrectable memory error

Source: SNIA Persistent Memory Summit January 2018

© 2019 Storage Networking Industry Association. All Rights Reserved.

NVDIMMs with Encryption

- Application Examples
  - Financial – high-speed trading, OLTP
  - Public – DoD
  - Health – medical records
  - Private - corporate IT departments

- With block access NVDIMMs the controller chip can manage encryption in the same way as SSDs
- With byte access NVDIMMs, the host memory controller needs to provide encryption support
  - A key is supplied by the host to support backup with encryption (which could impact performance)
- During a system power loss, in-flight data written from the DRAM to the Flash will be encrypted
- NVDIMM encryption standardization is in process with JEDEC

© 2019 Storage Networking Industry Association. All Rights Reserved.

Tools and Utilities for Managing NVDIMMs

- Even though NVDIMMs are JEDEC-standard there are no open source utilities
  - Write/read data patterns
  - Backup/restore automated testing
  - Firmware updates (hdparm for SSDs)
  - Status updates (i.e., smartmontools for SSDs)

- Backup power is not standardized
- How can we come up with an open source management utility for NVDIMMs?

© 2019 Storage Networking Industry Association. All Rights Reserved.

Key Takeaways

- Any latency-sensitive data that is continuously changing can benefit with Persistent Memory
- Workloads are being re-architected to use large amounts of data placed in local memory
- Data reload times are significant, driving a need to retain data through a power failure
- Applications help drive demand for Persistent Memory
- Standardization enables wide adoption of Persistent Memory-aware applications
- SNIA Persistent Memory and NVDIMM SIG is driving education and adoption

© 2019 Storage Networking Industry Association. All Rights Reserved.

SNIA Persistent Memory & NVDIMM SIG

- ✓ Educates on the types, benefits, value, and integration of Persistent Memories
- ✓ Communicates usage of the NVM Programming Model developed to simplify system integration of current and future PM technologies
- ✓ *Influences and collaborates with middleware and application vendors to support Persistent Memories*
- ✓ Develops user perspective case studies, best practices, and vertical industry requirements
- ✓ Coordinates with industry standards groups and promote industry standards related to PM and NVDIMM
- ✓ Synchronizes and communicates a common Persistent Memory taxonomy

© 2019 Storage Networking Industry Association. All Rights Reserved.

Full agenda details and complimentary registration available at *snia.org/pm-summit*

© 2019 Storage Networking Industry Association. All Rights Reserved.

Persistent Memory Events at FMS

- Monday August 5
  - Persistent Memory Programming Tutorial and Introduction to the Hackathon – (1:00 – 5:00 pm, room 209/210)
- Tuesday, August 6
  - FMS Persistent Memory Hackathon hosted by SNIA (drop in any time 8:30 am – 7:00 pm with your laptop – Great America Ballroom Foyer)
  - PMEM-101-1 – Persistent Memory Part 1 – Advances in Persistent Memory (8:30 – 10:50 am)
    - Talks on PM standards, PM and Compute Express Link (CXL) and a state of the Union, followed by a look at new and emerging media
  - PMEM-102-1 – Persistent Memory Part 2 - Software and Applications (3:40 -6:00 pm)
    - Talks on where applications are going, how to program, providing native support in software, & new applications like neuromorphic computing
- Wednesday, August 7
  - FMS Persistent Memory Hackathon hosted by SNIA (drop in any time 8:30 am – 7:00 pm with your laptop – Great America Ballroom Foyer)
  - PMEM-201-1 – Persistent Memory Part 3 – Remote Persistent Memory (8:30 – 10:50 am)
    - Talks on the case for use cases, and Gen-Z and CXL, followed by a panel on hows and whys on remote PM
  - PMEM-202-1 – Persistent Memory Part 4 – Current Research in Persistent Memory
    - The latest from Ghent University, University of Virginia, Los Alamos National Labs, and Webfeet Research on what you will see in the future

Visit SNIA in booth 820 where they will have a persistent memory demo and information on computational storage and storage management

© 2019 Storage Networking Industry Association. All Rights Reserved.

Thanks for Attending

Visit www.snia.org/pm

for Persistent Memory videos, webcasts, and presentations