Intel's Exciting Persistent Memory Technologies and VMware vSphere

Andy Rudoff, Intel
Richard Brunner, VMware, Inc.
Disclaimer

This presentation may contain product features or functionality that are currently under development.

This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.

Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.

Technical feasibility and market demand will affect final delivery.

Pricing and packaging for any new features/functionality/technology discussed or presented, have not been determined.
Business Agility 101: The Speed of Analysis

Today:
Find the needle in the haystack

Tomorrow:
Find it faster in bigger haystacks at lower cost
Problem: Local Storage Latency

• Speed of Analysis is dependent on Storage and Database Access Latency.
• A key component is local storage latency.
  – Current solutions cannot compete with the latency of DRAM.
• What if you could move storage closer to where the analysis is being done?
  – So close that the data can be accessed by a processor as if it were DRAM-like.
  – With reduced latency and byte-granular access.
Problem: Local Storage Latency

- What if you could move storage closer to where the analysis is being done?
  - So close that the data can be accessed by a processor as if it were DRAM-like.
  - With reduced latency and byte-granular access.

- You can with Byte-Addressable Persistent Memory (PMEM).

- PMEM is a fundamental change in Storage & Database architecture.

- This year PMEM solutions will offer high-capacity and performance.

- VMware vSphere 6.7 brings the agility benefits of PMEM to the data center.
What is PMEM (Persistent Memory)?

Non-Volatile Memory

NVM

Block-Mode Access

Random Access

(byte/word addressable)

NVM Express: NVM block-mode attached via PCIe.

NVM block-mode implemented in DIMM form factor.

NVMe

NVDIMM-F

NVRAM

NVM with Random-Access Memory.

PMEM

Persisted Memory: NVRAM that is direct-mapped & addressable by CPU Load/Store Instructions from non-privileged Apps.

DIMM-compatible

PMEM

NVDIMM-N

3D-XPoint

NVDIMM-P

Byte-Addressable Persistent Memory (PMEM) is Storage with these characteristics:

- DRAM latency and bandwidth → A few hundred nanoseconds or less on average
- DRAM granularity and access → byte-level access in the normal system memory map
- DRAM model for App → regular, non-privileged, load/store CPU instructions
- DRAM model for OS → paged/mapped by OS just like Memory Mapping DRAM

VMware vSphere 6.7 enables PMEM by supporting a virtualized NVDIMM device.
How Does PMEM Change & Benefit the Data Access Model?

*Read-Modify-Write (RMW) Byte Example (Greatly Simplified*)

**External Device Block Storage Model**
- 64-Byte Cacheline
- 4-KByte block
- DRAM
- SSD Storage Controller & Media

(RMW duration can be a few hundred microseconds)

**Byte-Addressable PMEM Model**
- Processor Package
- PMEM

(RMW duration can be a few hundred nanoseconds)
Updating storage at finer-granularity can reduce storage overhead.

**PMEM benefits persistent workloads that need reduced latency & more flexible data access:**
- Traditional Database: Log Acceleration by caching and combining data writes
- In-Memory Database: Journaling, Logging, Reduced Recovery time
- Enterprise Storage: Fast-Caching Layer
- High-Performance Computing: In-memory check-pointing

* = Storage optimization can remove some of the penalty, but the basic flow is still the same. Note CPU Cache skipped due to complexity of example.
What is Persistent Memory?
(and Why do you care?)
The Memory Timeline

- 1947: Ram
- 1956: PROM
- 1961: SRAM
- 1966: DRAM
- 1971: EPROM
- 1988: NOR Flash Memory
- 1989: NAND Flash Memory
- 2016: 3D XPoint™
The Memory Timeline

1955-1975
- Magnetic Core Memory

1947
- Ram

1956
- PROM

1961
- SRAM

1966
- DRAM

1971
- EPROM

1988
- NOR Flash Memory

1989
- NAND Flash Memory

2016
- 3D XPoint™

©2018 VMware, Inc.
Big and Affordable Memory
High Performance Storage
Direct Load/Store Access
Native Persistence

128, 256, 512GB
DDR4 Pin Compatible
Hardware Encryption
High Reliability

Now shipping samples
broad developer engagement
Optimized System Interconnect

Reach full potential of 3D XPoint™ Technology by connecting it as Memory

Sources: “Storage as Fast as the rest of the system” 2016 IEEE 8th International Memory Workshop and measurement; Intel® Optane™ SSD measurements and Intel P3700 measurements, and technology projections
Intel® Optane™ DC Persistent Memory – Database Restart

Database Restart From Minutes to Seconds!
The Value of Persistent Memory

Data sets addressable with no DRAM footprint
• At least, up to application if data copied to DRAM

Typically DMA (and RDMA) to pmem works as expected
• RDMA directly to persistence – no buffer copy required!

The “Warm Cache” effect
• No time spend loading up memory

Byte addressable

Direct user-mode access
• No kernel code in data path
How does software use persistent memory?
The Storage Stack (50,000ft view...)

- Management UI
- Application
  - Standard Raw Device Access
  - Standard File API
- Application
- Management Library
- File System
- Driver
- Storage

User Space

Kernel Space
Making Persistent Memory Visible

Application

Middleware
Libraries

File Systems

In-kernel Usages
Block Storage

Driver

File System

Increasing barrier to adoption

Intel OPTANE DC Persistent Memory

VMware, Inc.
Making Persistent Memory Visible

- Application
- Middleware
- Libraries
- File Systems
- In-kernel Usages
- Block Storage

Increasing value to application

Intel OPTANE DC Persistent Memory

USER SPACE

KERNEL SPACE
Memory-Mapped Files

What are memory-mapped files really?
• Direct access to the page cache
• Storage only supports block access (paging)

With load/store access, when does I/O happen?
• Read faults/Write faults
• Flush to persistence

Not that commonly used or understood
• Quite powerful
• Sometimes used without realizing it
OS Paging

![Diagram of OS Paging]

- **User Space**
  - Application
  - load/store access
- **Kernel Space**
  - Application
  - page fault access
  - DRAM
  - NVDIMM

VMworld 2018 Content: Not for publication or distribution
The SNIA NVM Programming model
The SNIA NVM Programming model
The SNIA NVM Programming model
The SNIA NVM Programming model

USER SPACE

file

memory

KERNEL SPACE

Management Library

Management UI

Application

Standard Raw Device Access

Standard File API

File System

Generic NVDIMM Driver

Application

Load/Store

“DAX”

Application

Standard File API

pmmem-Aware File System

MMU Mappings
Can we make persistent memory easier to use?
Division of Roles & Responsibilities in the Server Technology Stack

- **Order 10⁶ developers**
  - Minimal to no HW knowledge
  - Goals are stability and SW features
  - Continue to trade-off performance for developer efficiency

- **Order 10³ developers**
  - Minimal HW knowledge
  - Goal to provide stable SW platform for ISVs

- **Order 10² relevant developers**
  - Good HW knowledge
  - Goal to isolate layers above from HW changes
  - Goal to innovate and expose innovation to customers
  - Innovation often requires changing interfaces

Persistent memory breaks this industry model limiting developer adoption and restricting our ability to innovate in the HW platform
The persistent memory development kit – PMDK
http://pmem.io

PMDK is a collection of libraries

• Developers pull only what they need
  – Low level programming support
  – Transaction APIs
• Fully validated
• Performance tuned

Open source & product neutral
Using Cassandra as an example

App

Unmodified App, uses Cassandra APIs

Cassandra

Use Java containers to create pmem-aware Cassandra

PCJ

Caller just sees the same APIs, uses them as before
Provide Java containers

libpmemobj

Provide transactions, persistent memory allocator

libpmem

Abstract away hardware details

pmem-aware File System

Expose Persistent Memory as memory-mapped files (DAX)
Using Cassandra as an example

**SNIA Programming Model**

- **App**: Unmodified App, uses Cassandra APIs
- **Cassandra**: Use Java containers to create pmem-aware Cassandra
- **PCJ**: Caller just sees the same APIs, uses them as before
  - Provide Java containers
- **libpmemobj**: Provide transactions, persistent memory allocator
- **libpmem**: Abstract away hardware details
- **pmem-aware File System**: Expose Persistent Memory as memory-mapped files (DAX)
Using Cassandra as an example

- **App**: Unmodified App, uses Cassandra APIs
- **Cassandra**: Use Java containers to create pmem-aware Cassandra
- **PCJ**: Caller just sees the same APIs, uses them as before; Provide Java containers
- **libpmemobj**: Provide transactions, persistent memory allocator
- **libpmem**: Abstract away hardware details
- **SNIA Programming Model**
  - **pmem-aware File System**: Expose Persistent Memory as memory-mapped files (DAX)
- **PMDK**: Intel Optane DC Persistent Memory

Using Cassandra as an example.
VMware vSphere PMEM Support
How VMware vSphere 6.7 Works with PMEM Solutions

Legacy OS & Application Usage

- Native:
  - Can use PMEM as block storage device with special driver.

- Virtualized:
  - Can use PMEM as block storage device with special driver in VM.
    - With vSphere 6.7, no special driver is required in the VM.
    - Guest Storage can be mapped to PMEM outside of VM.
How VMware vSphere 6.7 Works with PMEM Solutions

New OS & Application Usage
- Native & Virtualized
  - Can use a direct load/store model with little OS overhead
- All the benefits of VMware vSphere Virtualization can be available:
  - Multiple workloads using PMEM
  - Live VM Migration across servers
  - Check-pointing
  - Boost for Legacy VMs/Workloads
  - And More …
Key Characteristics of PMEM Solutions (1)

1. The persistence technology, latency, and reliability.
   • Note: in some cases access is to the DRAM image of the data, thus latency is that of DRAM.
   • DRAM caches can also be used with certain technologies.
   • Note: in some cases writes only happen to the “media” on shutdown, thus significantly reducing P/E cycles.

2. The capacity of the solution.
   • Per DIMM can span from 8 GiB to 100s of GiB.

3. The software access model.
   • Can it support both legacy block-mode and new load/store access?

4. The right metric for assessing Total Cost of Ownership for your workloads.
   • A traditional Cost / GB metric may not take into account the significantly reduced latency and bandwidth of PMEM.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Program/Erase Cycles (P/E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDD</td>
<td>$10^{16}$</td>
</tr>
<tr>
<td>NAND Flash SLC</td>
<td>$10^{5}$</td>
</tr>
<tr>
<td>NAND Flash MLC</td>
<td>$10^{3}$</td>
</tr>
<tr>
<td>3D-XPoint[1]</td>
<td>$10^{8}$</td>
</tr>
<tr>
<td>PCRAM[2]</td>
<td>$10^{8}$</td>
</tr>
<tr>
<td>ReRAM[2]</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>HDD[2]</td>
<td>$10^{16}$</td>
</tr>
<tr>
<td>DRAM</td>
<td>$10^{18}$</td>
</tr>
</tbody>
</table>

1. VMware’s estimate on 3D XPoint endurance. Estimate is based on Intel’s claim that 3D XPoint is capable of up to 1000 times greater endurance than NAND.
VMware Supported (or Intended) PMEM Technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Capacity</th>
<th>Latency</th>
<th>Endurance</th>
<th>Access</th>
<th>Cost per GiB</th>
<th>Energy Source</th>
<th>Availability</th>
<th>VMware Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NVDIMM-N</strong></td>
<td>8-32 GiB per DIMM</td>
<td>DRAM (10's of ns)</td>
<td>NAND (10^4 to 10^5 PE cyc)</td>
<td>Load/Store</td>
<td>~ $45/GiB[1]</td>
<td>Battery or equivalent</td>
<td>Now</td>
<td>Supports Dell NVDIMM-N &amp; HPE NVDIMM-N</td>
</tr>
<tr>
<td><strong>Intel® Optane™ DC Persistent Memory (3D-XPoint [1])</strong></td>
<td>128 GiB – 512 GiB[2]</td>
<td>100's of ns[3]</td>
<td>NAND &lt; Endurance &lt; DRAM</td>
<td>Load/Store (Also Block)</td>
<td>&lt; DRAM (See later this presentation)</td>
<td>Light-Weight Energy Source</td>
<td>Intends to support Intel® Optane™ DC Persistent Memory at Launch</td>
<td></td>
</tr>
</tbody>
</table>

1. Based on 16-GiB NVDIMM-N for Dell PowerEdgeR740xd at www.dell.com
2. VMware's estimate on 3D XPoint capacity. Estimate is based on Intel's claim that 3D XPoint is ~10x capacity of DRAM
3. VMware's estimate on 3D XPoint performance. Estimate is based on Intel's claim that 3D XPoint is ~10x latency of DRAM
4. VMware's estimate on 3D XPoint endurance. Estimate is based on Intel's claim that 3D XPoint is capable of up to 1000x endurance than NAND.
5. How persistence is ensured during a power fail.
   • All PMEM solutions need a way of ensuring that the last set of updates have made it.
   • Persistence Domain size depends on SW requirements, platform power budget, and PMEM energy source.
   • Some PMEM solutions can ensure from the CPU cache outward; others from just DRAM.

6. The Server HW &FW support of PMEM.
   – Correct and tested UEFI/ACPI interfaces are critical to Operating Systems & Hypervisors.
   – Well-tested error and power-fail handling are critical.
   – From VMware’s POV, this is so important that we will only certify a PMEM solution in combination with the platform that supports it.
VMware vSphere Support for PMEM...

Goal: Your applications whether modified to use NVDIMMs or legacy can take advantage of PMem on VMware vSphere!
Enter maintenance mode (vacate powered off VMs also)
Summary

- VMware vSphere supports PMEM today.
  - Intends to support Intel® Optane™ DC Persistent Memory at Intel launch
- All the benefits of VMware vSphere Virtualization will be available:
  - Multiple workloads using PMEM
  - Live VM Migration across servers even with different PMEM types
  - Boost for Legacy VMs/Workloads
- PMEM will experience a capacity explosion when Intel® Optane™ DC Persistent Memory launches
- The software eco-system is ready now, key ISVs have already converted
- Now there are 3 storage tiers to consider when spec’ing out your new Intel 2019 servers
  - Making the right choices can relay benefit TCO
    - Stop at the Intel booth (#1212) for demonstrations of VMware® ESX and Intel® Optane™ DC Persistent Memory
Learn more about PMEM @VMworld 2018 US

- Extreme Performance Series: vSphere PMEM = Storage at Memory Speed [VIN2183BU]
  - Thursday, Aug 30, 10:30 a.m. - 11:30 a.m.

- Intel's Exciting Persistent Memory Technologies and VMware vSphere [VIN2963BU]
  - Tuesday, Aug 28, 3:30 p.m. - 4:30 p.m.

- Persistent Memory Deep Dive: HW and SW Architecture for VMware Technologies [CTO2860BU]
  - Wednesday, Aug 29, 11:30 a.m. - 12:30 p.m.

- Enhance your vSphere and vSAN implementations w/ Persistent Memory and SSDs [HCI3697BUS]
  - Tuesday, Aug 28, 2:00 p.m. - 3:00 p.m.

Stop at the Intel booth (#1212) for demonstrations of VMware® ESX and Intel® Optane™ DC Persistent Memory
POSSIBLE BEGINS WITH YOU