VMware vSphere Virtualization of PMEM (PM)

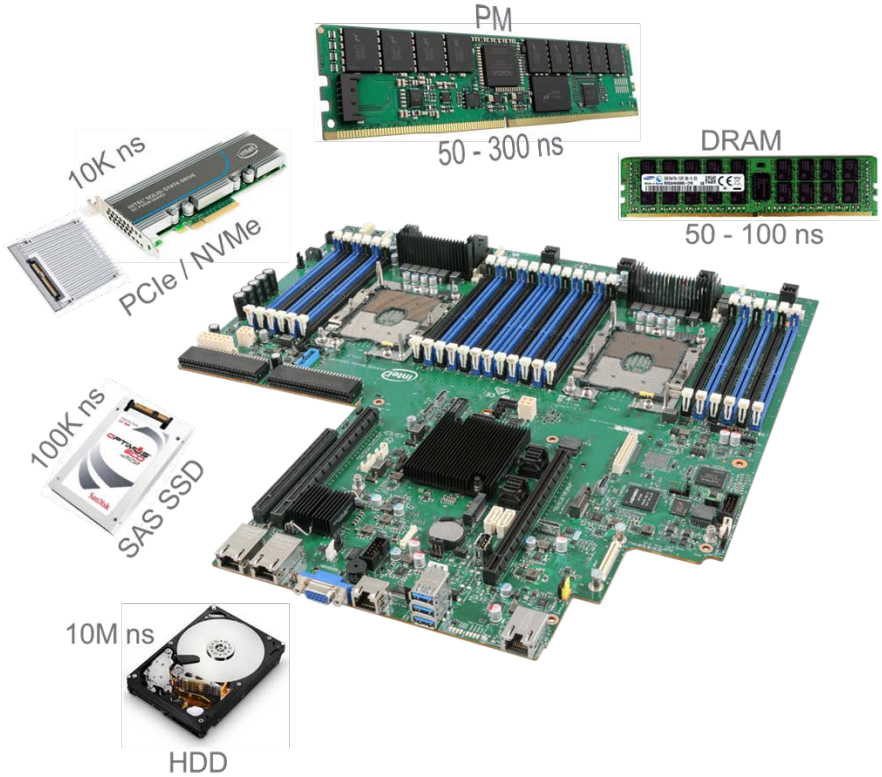
Richard A. Brunner, VMware

Disclaimer

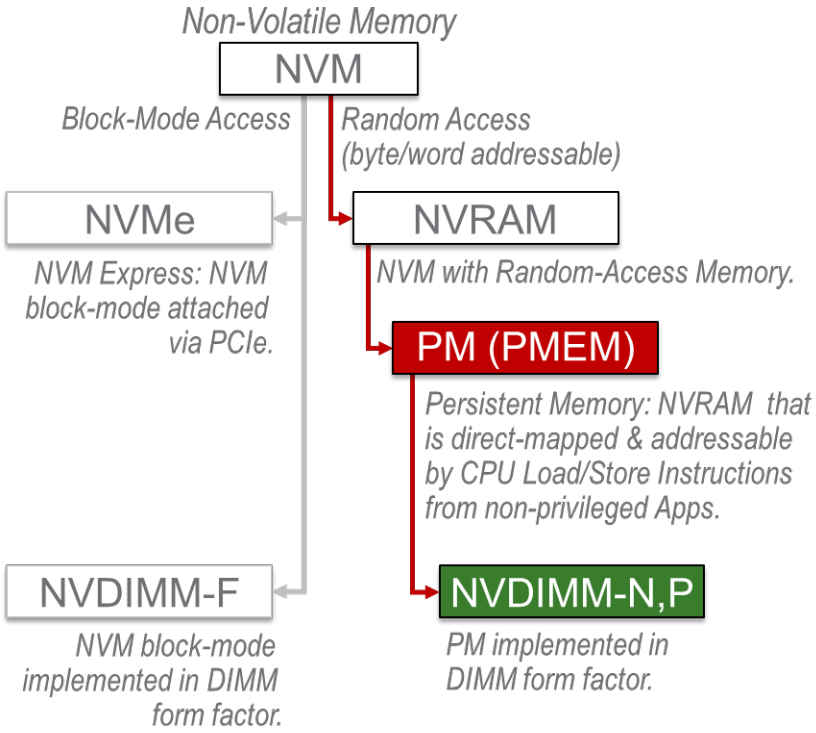
- This presentation may contain product features that are currently under development.
- This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.
- Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.
- Technical feasibility and market demand will affect final delivery.
- Pricing and packaging for any new technologies or features discussed or presented have not been determined.

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 (PM).
- PM is a fundamental change in Storage & Database architecture.
- This year PM solutions will offer high-capacity and performance.
- Future VMware vSphere will bring the agility benefits of PM to the data center.



What is PM (Persistent Memory)?



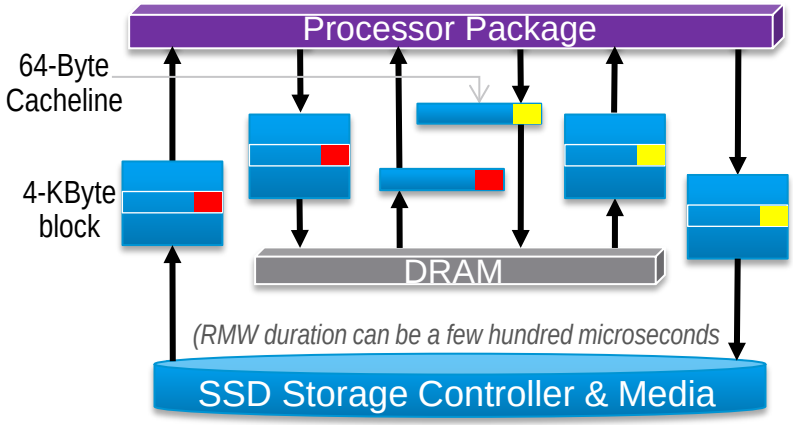
Byte-Addressable Persistent Memory (PM) is Storage with these characteristics:		
DRAM latency and bandwidth	→	A few hundred nanoseconds or less on average
DRAM granularity and access	→	byte-level access in the normal system memory map
DRAM model for App software	→	regular, non-privileged, load/store CPU instructions
DRAM model for OS Memory Mapping	→	paged/mapped by OS just like DRAM

A future version of VMware vSphere intends to enable **PM** by supporting a virtualized ***NVDIMM*** device.

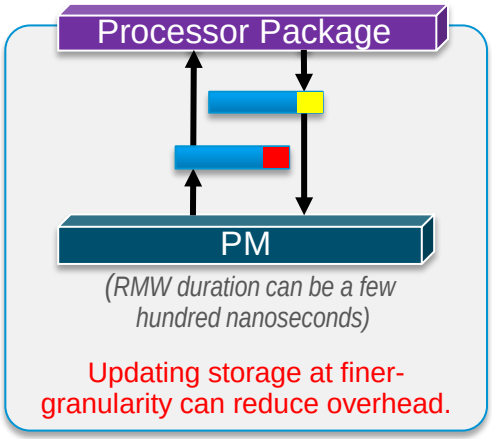
How Does PM Change the Data Access Model?

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

External Device Block Storage Model



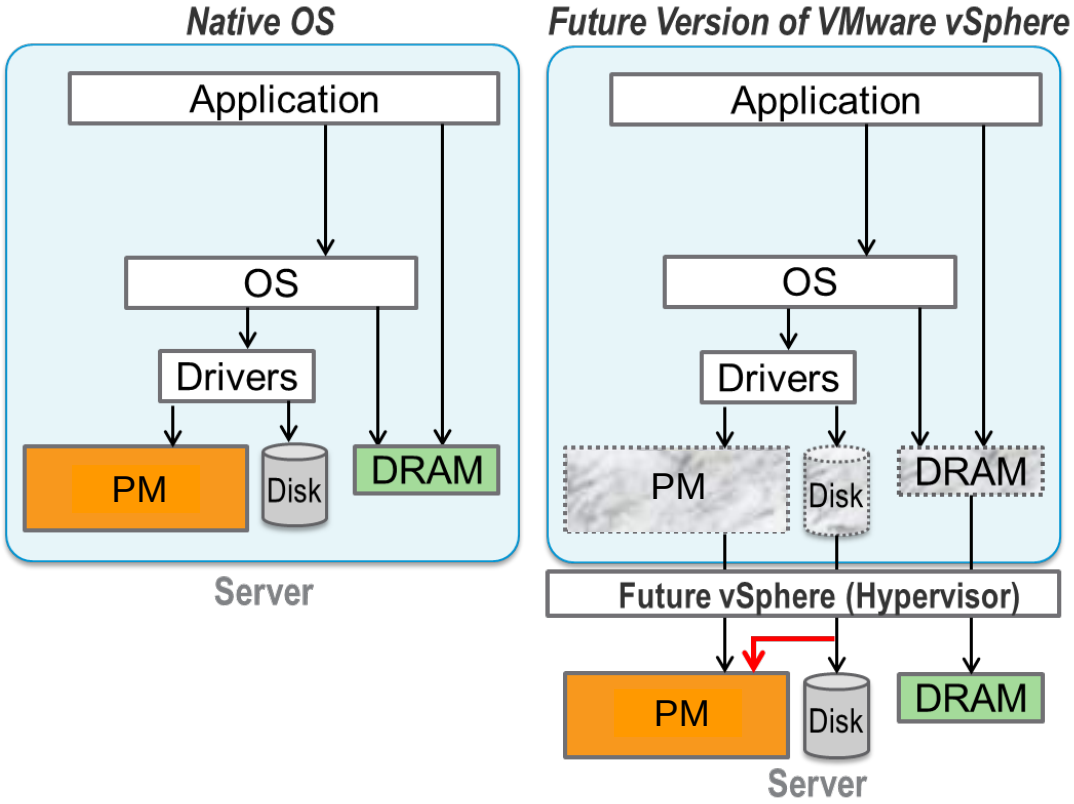
Byte-Addressable PM Model



PM benefits persistent workloads with 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.

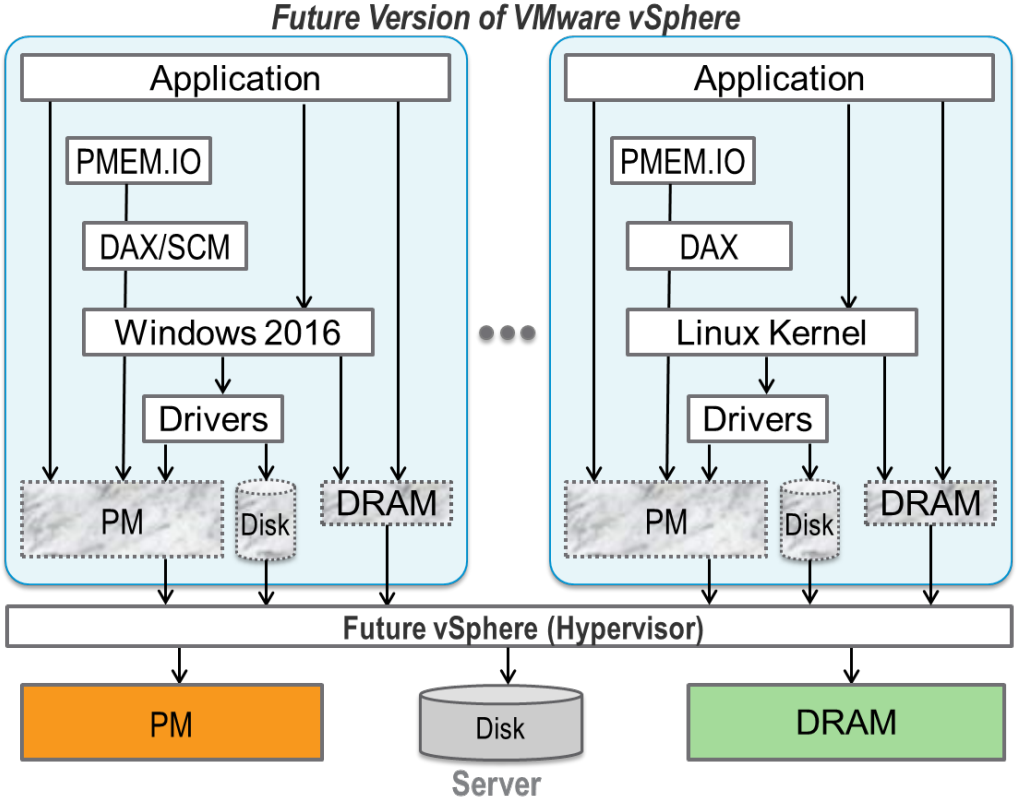
How VMware vSphere Can Work with PM Solutions



Legacy OS & Application Usage

- Native:
  - Can use PM as block storage device with special driver.
- Virtualized:
  - Can use PM as block storage device with special driver in VM.
  - *With a future version of vSphere, no special driver is required in the VM.*
  - *Guest Storage can be mapped to PM outside of VM.*
  - *No change to the guest OS or Application stack*

How VMware vSphere Can Work with PM Solutions

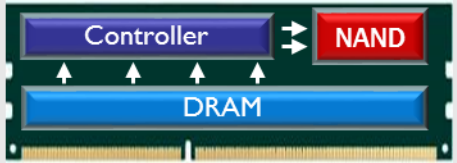


New OS & Application Usage

- 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 PM
  - Live VM Migration across servers
  - Check-pointing
  - Boost for Legacy VMs/Workloads
  - And More …

PM Technology

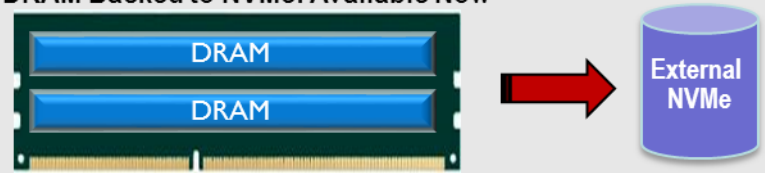
NVDIMM-N: Available Now



Memory-Mapped DDR4 DRAM, backup to NAND Flash on Shutdown & Power Fail

Capacity:	8-32 GiB per DIMM
Latency:	DRAM (10's of ns)
Endurance:	NAND (10^3 to 10^5 PE cyc)
Cost per GiB:	~ $80/GiB[5]
Energy Source:	Battery or equivalent
VMware:	Intend to support Dell & HPE NVDIMM-N

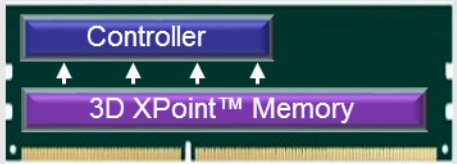
DRAM Backed to NVMe: Available Now



Memory-Mapped DDR4 DRAM, backup to NVMe devices on Shutdown & Power Fail

Capacity:	RDIMM (128 GiB per RDIMM)
Latency:	DRAM (10's of ns)
Endurance:	NVMe Media
Cost per GiB:	TBD Less than $80/GIB ??
Energy Source:	Battery or equivalent
VMware:	Intend to support HPE Scalable PMEM

Intel DIMMs Based on 3D XPoint™ Technology

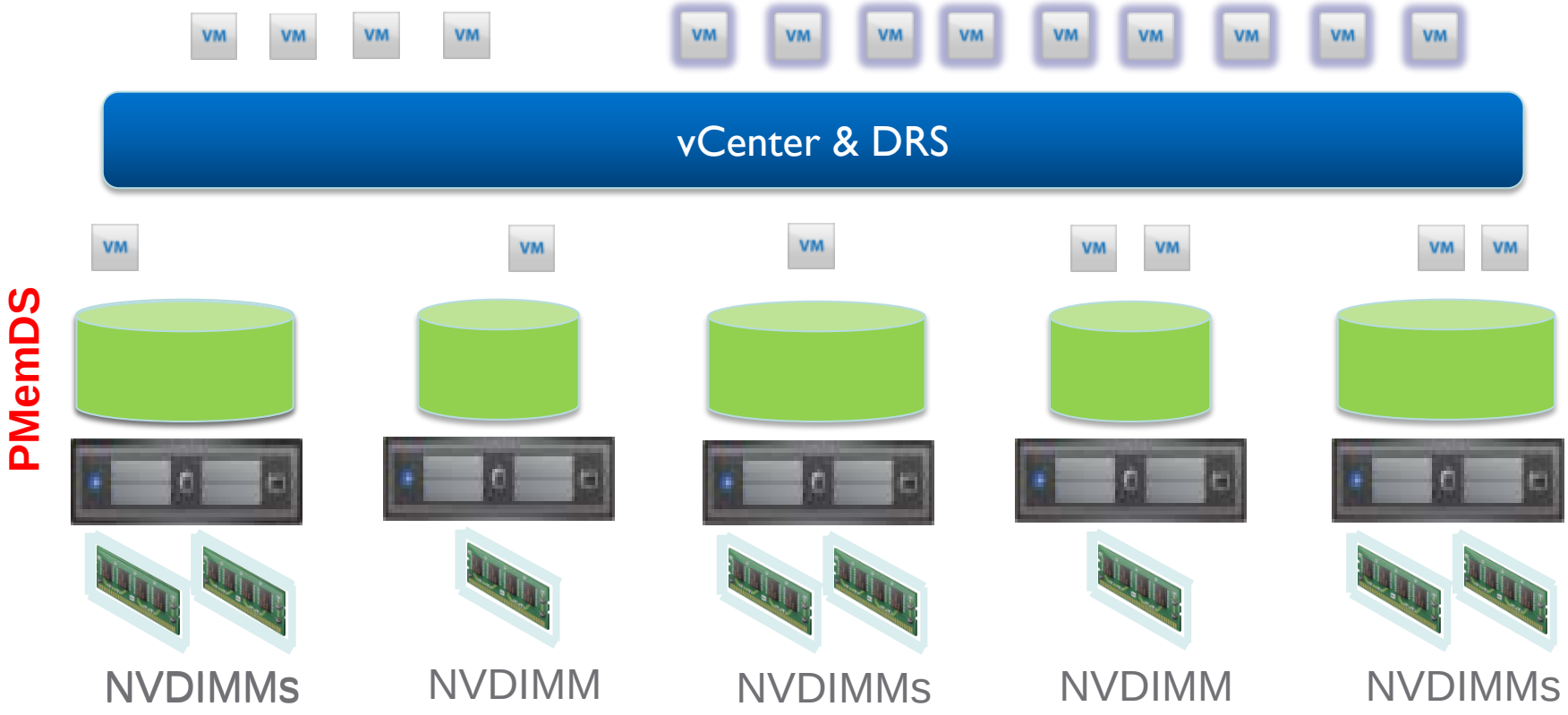


Memory-Mapped DDR4-extended Storage, No backup needed

Est. Capacity:	256 GiB – 512 GiB[1]
Est. Latency:	100's of ns[2]
Est. Endurance:	1000 x NAND[4]
Est. Access:	Load/Store (Also Block)
Est. Cost per GiB:	< DRAM[3] (Today 128 GiB RDIMM is ~$13/GiB)
Est. Energy Source:	Light-Weight Energy Source
VMware:	Intend to support Intel 3D Xpoint™

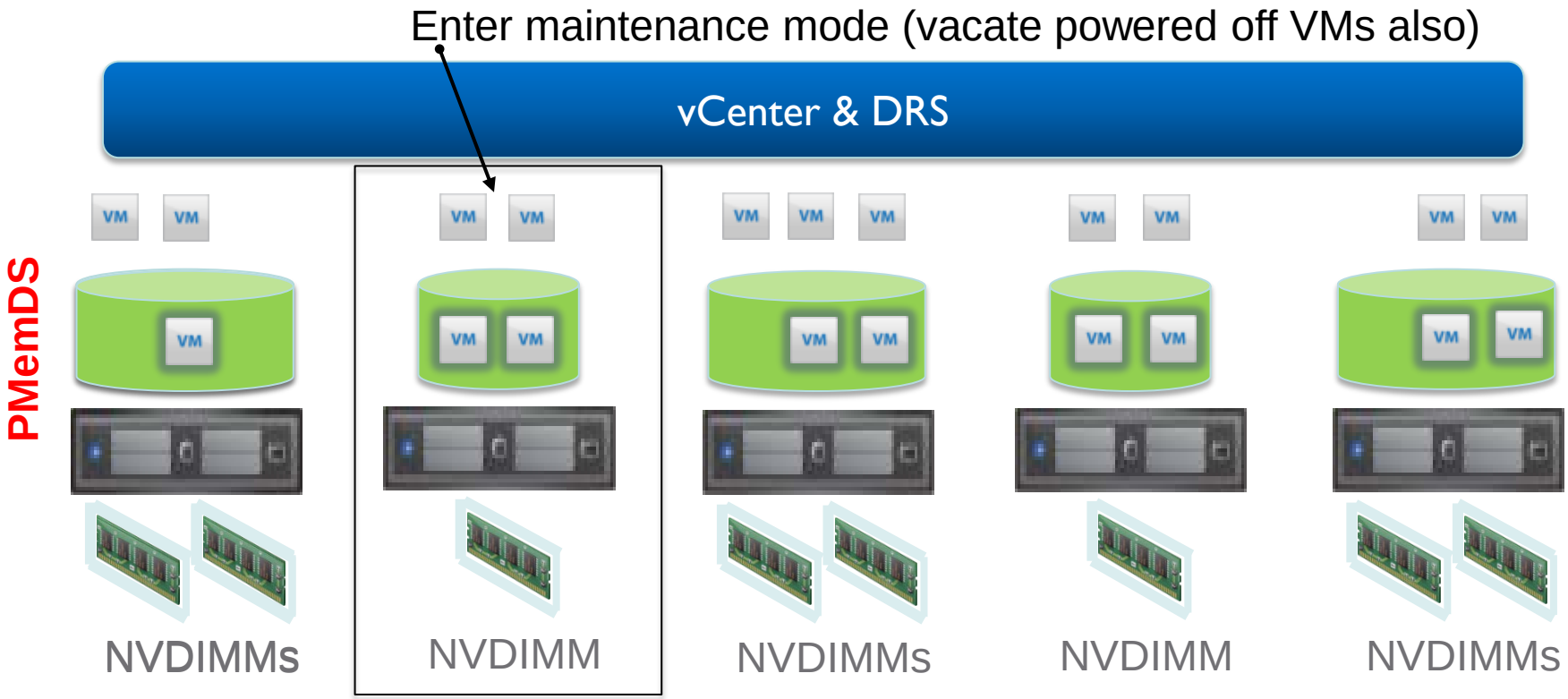
Please See Footnotes on Last Page

vSphere Support For Persistent Memory



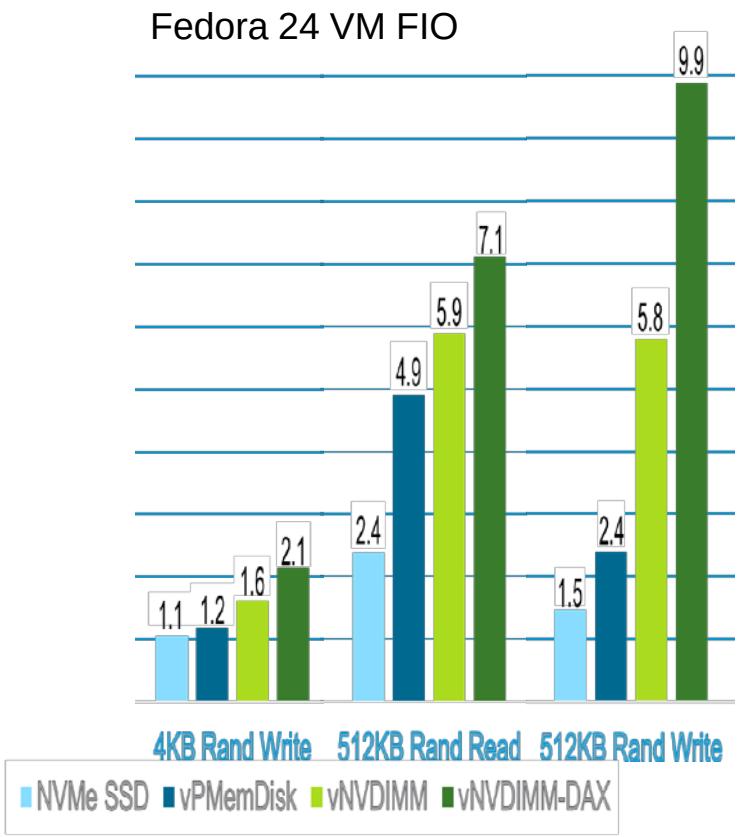
Slide from Rajesh Venkatasubramanian, VMware

vSphere Support For Persistent Memory



Slide from Rajesh Venkatasubramanian, VMware

Performance - Single Thread Using Various Virtual Devices



NVMe SSD: vSCSI Emulation to NVMe SSD; Legacy & New OS

- Block based file read/write via page cache (RAM)
- mmap backed by page cache (RAM)
- File system writes page cache (RAM) to disk on fsync
- Disk write executes guest OS SCSI stack

vPMemDisk: vSCSI Emulation to PM; Legacy & New OS

- Block based file read/write via page cache (RAM)
- mmap backed by page cache (RAM)
- File system writes page cache (RAM) to disk on fsync
- Disk write executes guest OS SCSI stack

vNVDIMM (block access): PM mapped into New OS

- Block based file read/write via page cache (RAM)
- mmap backed by page cache (RAM)
- File system writes page cache (RAM) to PM on fsync

vNVDIMM-DAX (Direct Access): PM mapped into New OS

- File read/write directly to PM pages
- mmap directly maps PM pages to application
- No need for fsync

Slide from Rajesh Venkatasubramanian, VMware

Conclusion

- Byte-Addressable PM is a fundamental change in storage & database architecture.
- Multiple PM technologies are on the market in 2018, more coming.
- Legacy OS & Apps can get some uplift.
- New OS & Apps re-coded for SNIA programming model will see much better uplift.
- Future VMware vSphere will unlock PM data chained to a server and bring the agility benefits of PM to the data center.

Footnotes

1. VMware's estimate on Intel 3D XPoint capacity. Estimate is based on Intel's claim that 3D XPoint is ~10x capacity of DRAM
2. VMware's estimate on Intel 3D XPoint performance. Estimate is based on Intel's claim that 3D XPoint is ~10x latency of DRAM
3. Intel Developer Forum 2015 San Francisco, Diane Bryant Keynote
4. VMware's estimate on Intel 3D XPoint endurance. Estimate is based on Intel's claim that 3D XPoint is capable of up to 1000 times greater endurance than NAND.
5. Based on 8-GiB RDIMM – https://www.cdw.com/shop/products/HPE-DDR4-8-GB-DIMM-288-pin/4097720.aspx and 8-GiB NVDIMM-N – https://www.cdw.com/shop/products/HPE-DDR4-8-GB-NVDIMM-N-288-pin/4077823.aspx