SNIA PERSISTENT MEMORY PM SUMMIT

JANUARY 24, 2018 | SAN JOSE, CA

Windows Support for PM

Tom Talpey, Microsoft

Agenda

- Windows and Windows Server…
  - PM Industry Standards Support
  - PMDK Support
  - Hyper-V PM Support
  - SQL Server PM Support
  - Storage Spaces Direct PM Support
  - SMB3 and RDMA PM Support

Windows PM Industry Standards Support

- JEDEC JESD 245, 245A: Byte Addressable Energy Backed Interface
  - Defines the host to device interface and features supported for a NVDIMM-N
- UEFI 2.5 – 2.7
  - Label format
  - Block Translation Table (BTT): sector atomicity support
- ACPI 6.0 – 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

Windows PMDK Support

- PMDK open source library available on Windows
  - Formerly Intel “NVML” (ref: Andy Rudoff’s talk earlier today)
  - http://pmem.io
  - Source and **prebuilt binaries** available via GitHub

    https://github.com/pmem/pmdk/
- Application API’s for efficient use of PM hardware
  - Most PMDK libraries (libpmem, etc) feature-complete on Windows
  - Underlying implementation uses memory mapped files
    - Access via native Windows DAX
  - Libraries work in both PM and non-PM hardware environments
  - Use case: simplified cross-platform application development

Windows Hyper-V PM Support

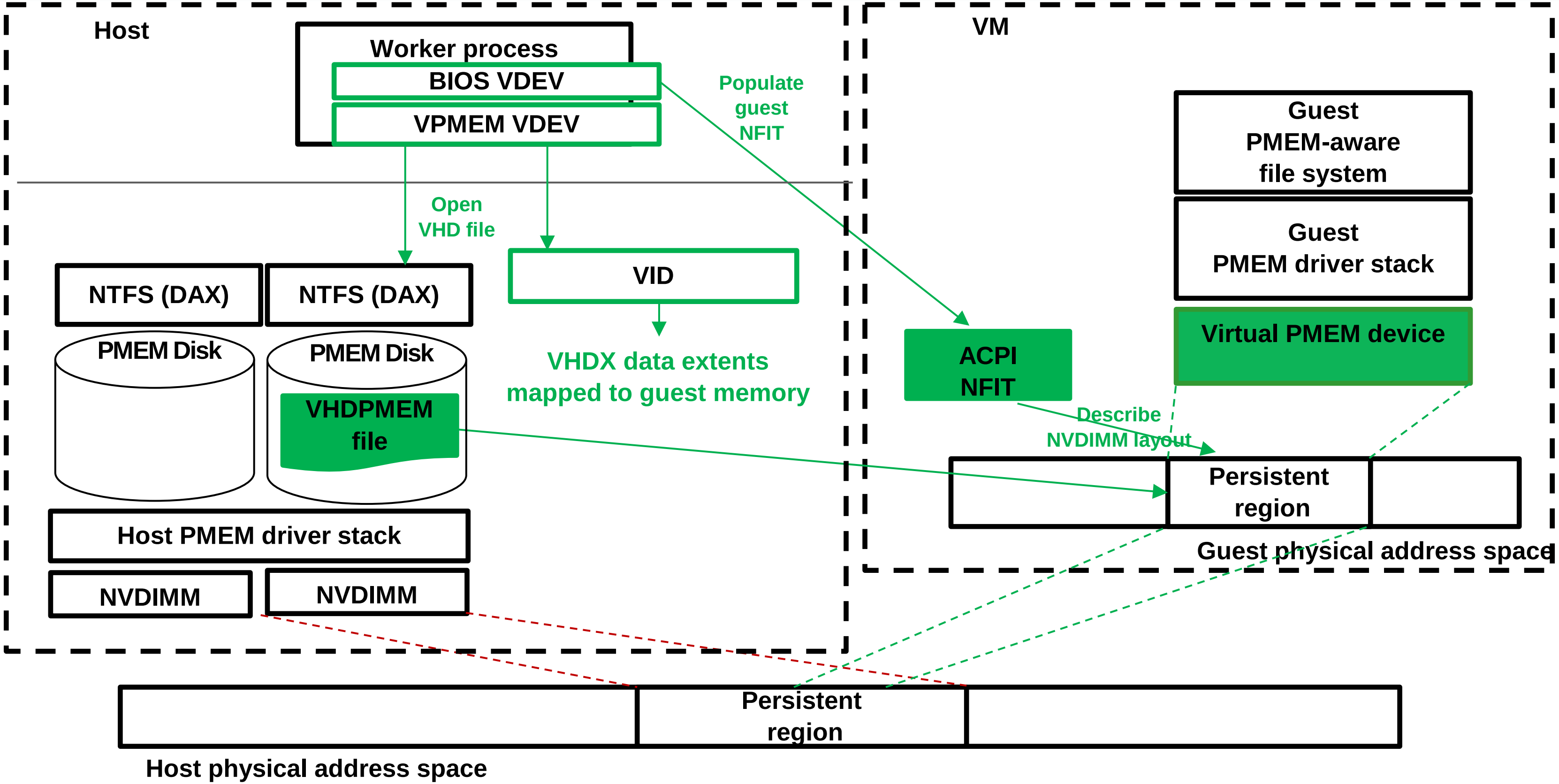
- Supported in Windows Fall 2017 release
- Windows & Linux guests in generation 2 VMs see virtual PMEM (vPMEM) devices
- New VHD file type: **.VHDPMEM**
  - Can convert between VHDX and VHDPMEM formats
    - Using **convert-vhd** PowerShell cmdlet
  - Admin decides at create time if the VHDPMEM file has a BTT
  - VHDPMEM files can also be mounted as a SCSI device on the host for read/write access
- Each vPMEM device is backed by one .VHDPMEM file

Windows Virtual PM (“vPMEM”) Details

- Windows Fall 2017 release basic enabling of persistent memory directly into a Hyper-V VM
- Windows Server Insider Preview 17074 “RS4” further updates
  - https://blogs.windows.com/windowsexperience/2018/01/16/announcing-windows-server-insider-preview-build-17074/

- DAX and BTT programming models, including Win32 APIs and PMDK, supported from guest
- Uses large pages automatically, when available
- Feature capacity limits
  - Maximum vPMEM device size is 256GB (1TB in preview)
  - Maximum VM vPMEM allocation is 256 GB (1TB in preview)
  - Minimum vPMEM device size is 16MB
  - Maximum number of vPMEM devices allocated to a VM is 128
- New features in Preview
  - **Migration of VMs with vPMEM allocations**
- All management implemented through PowerShell
  - Note, some VM meta-operations not yet supported:
    - Checkpoints, Backup, Save/Restore

vPMEM Component Level View



Windows vPMEM Configuration Example

- Use the **New-VHD** cmdlet to create a persistent memory device for a VM.

```
PS> New-VHD d:\VMPMEMDevice1.vhdpmem -Fixed -SizeBytes 4GB
```

- Use the **New-VM** cmdlet to create a Generation 2 VM with specified memory size and path to a VHDX image.

```
PS> New-VM -Name “ProductionVM1” -MemoryStartupBytes 1GB -VHDPath c:\vhd\BaseImage.vhdx
```

- Use **Add-VMPmemController** cmdlet to add a persistent memory controller to the VM.

```
PS> Add-VMPmemController ProductionVM1
```

- Use **Add-VMHardDiskDrive** to attach persistent memory device to the VM’s controller.

```
PS> Add-VMHardDiskDrive ProductionVM1 PMEM -ControllerLocation 1 -Path D:\VPMEMDevice1.vhdpmem
```

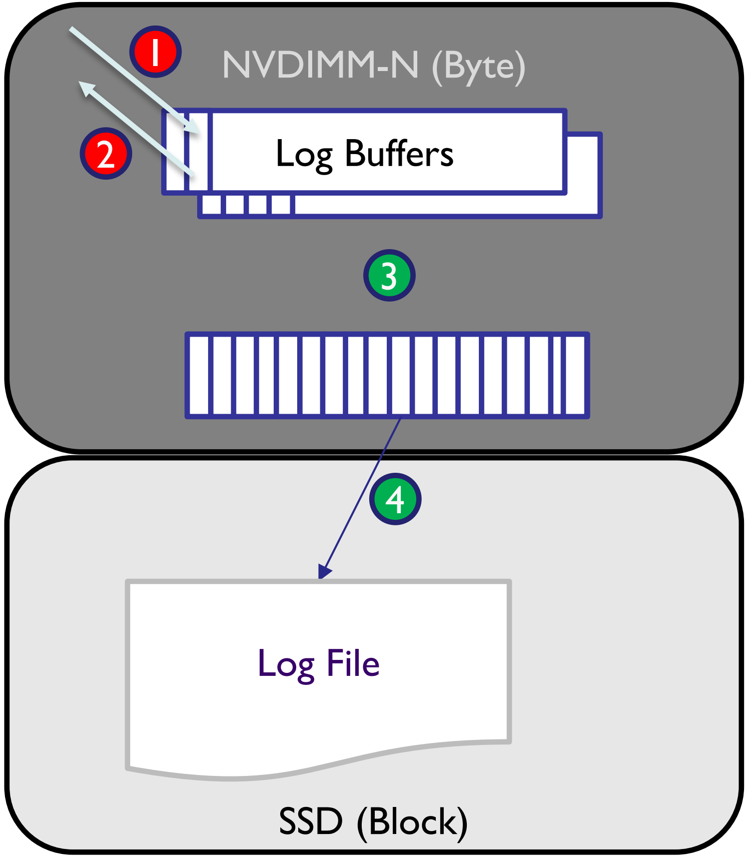
SQL Server 2016 PM Support: Tail-of-Log

- **Problem**
  - DB Transactions gated by log write speed
  - The faster the log, the more DB updates possible
- **Opportunity**
  - Accelerate Log Commits
  - Accelerate DB
- **Approach**
  - **Log on PM**

With "Tail of Log":

1. Copy log records into buffer, building up block **persistently in PM**
2. Complete transaction when commit arrives
3. Close log block when full
4. Schedule I/O to re-persist full block to SSD

Red indicates the critical path for a transaction, accelerated by PM



Accelerating SQL Server 2016 with PM

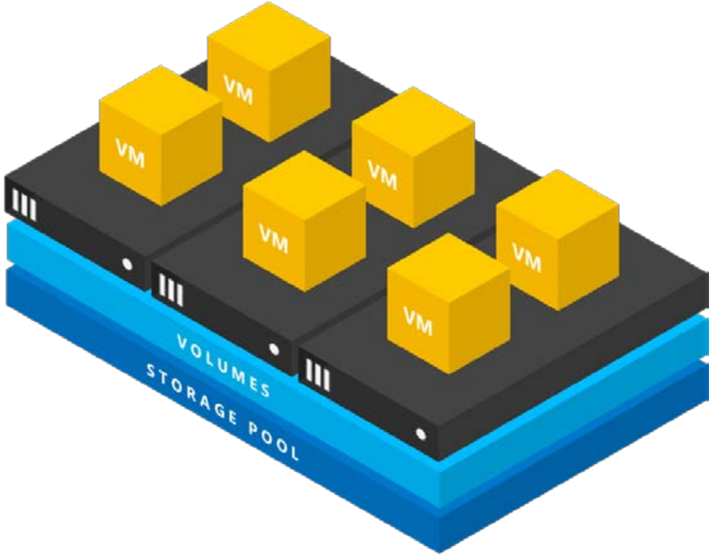
- SQL Server 2016 can use a byte-addressable log
  - Commit @ DRAM speed!
- Enabled through DAX volumes on PM in Windows
- Accelerate In-Memory DB updates by up to 2x

Configuration	HK on NVMe (block)	HK on NVDIMM-N (DAX)
Row Updates / Second	63,246	124,917
Avg. Time / Txn (ms)	0.379	0.192

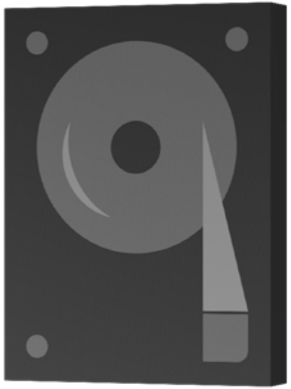
Configuration: Row Size: 32B, Table Size: 5GB, Threads:24, Batch Size: 1

Windows Storage Spaces Direct PM Support

- Set of servers acting together to expose unified internal services
  - Hosting multiple configurations and multiple scenarios, e.g.
    - Scaleout File Server
    - SQL Server
    - Hyperconverged compute/storage (shown)
- Highly available and scalable
- **PM** and NVMe devices for better performance and efficiency
- SATA devices for lower cost
- Ethernet/RDMA storage fabric
- Available in WS2016 Datacenter
  - Update previewing in “Insider 17074”

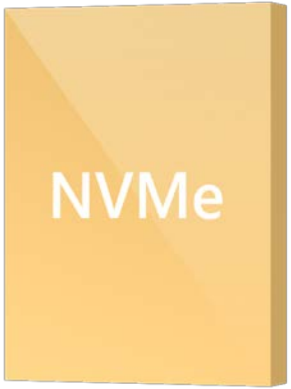


Storage Spaces Direct Device Support



HDD: Any Hard Disk Drive connected via SATA or SAS

SSD: Any other Solid-State Drive connected via SATA or SAS

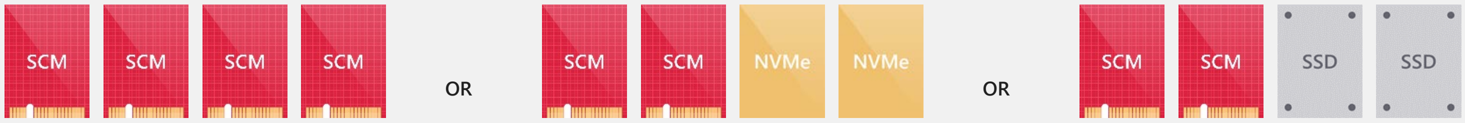


NVMe: Non-Volatile Memory Express, connected via PCIe bus (AIC, M.2 and U.2)

PM: (aka “**SCM”** Storage-Class Memory), Non-volatile storage connected via CPU memory bus

Storage Device Configurations

PM pooled as capacity devices

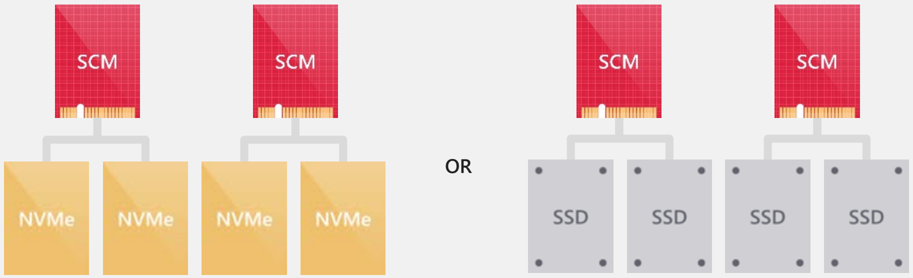


PM *for* Capacity

PM + NVMe *for* Capacity

PM + SSD *for* Capacity

PM tiered as caching devices



PM *for* Cache **NVMe** *for* Capacity

PM *for* Cache **SSD** *for* Capacity

Windows Storage Spaces Direct PM Support

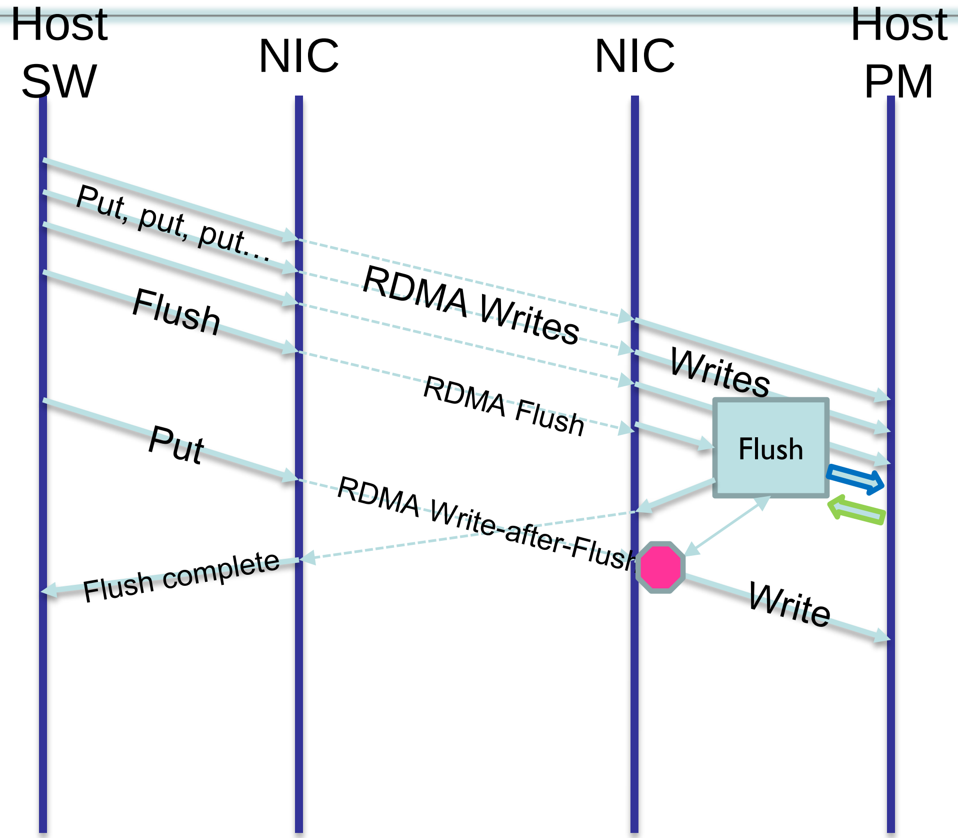
- Initially block emulation
- PM as cache or capacity

- Future – further PM usage envisioned

Windows RDMA-to-PM Support

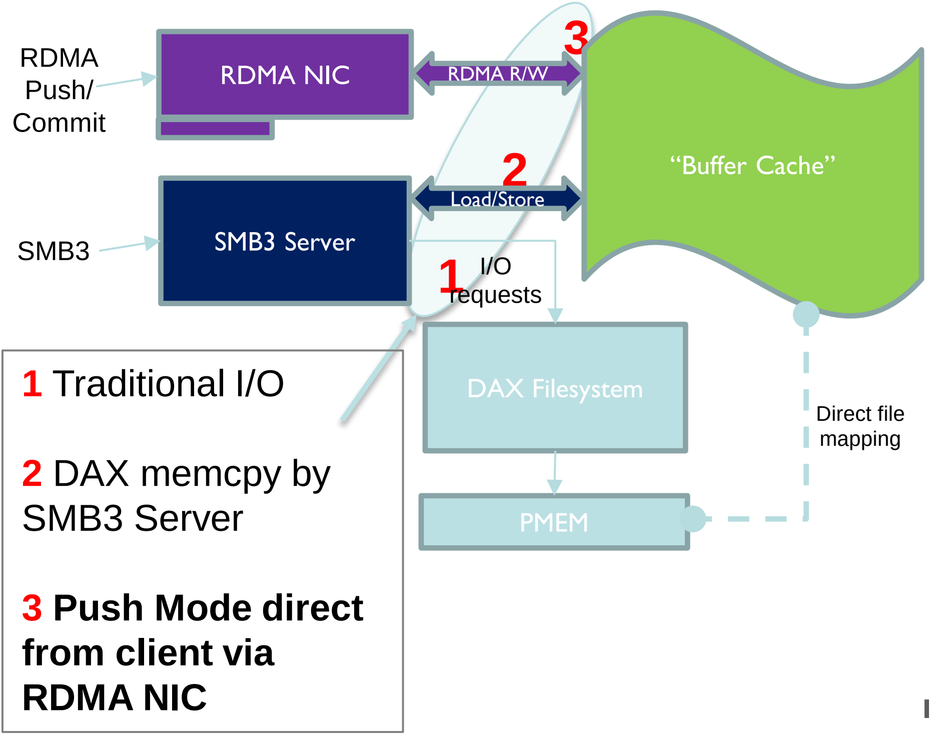
- Industry continues to converge on “RDMA Flush”
- Additional optimization semantics also discussing
  - Write-after-flush
  - Integrity, security
  - See Storage Developer Conference 2017, and prior
- Windows and Windows SMB3 presentations
  - Storage Developer Conference, etc

RDMA Flush and Write-after-Flush Extensions



SMB3 PM Support (including “Push Mode”)

- Enables zero-copy remote read/write to DAX file
  - Ultra-low latency and overhead
  - Optimal for replication
- Implementable in phases
- SDC 2017 “What’s new in SMB3” presentation
  - Modes 1 and 2
- Future
  - Mode 3!



1 Traditional I/O

2 DAX memcpy by SMB3 Server

3 Push Mode direct from client via RDMA NIC

Summary

- PM well-supported by Windows
- PM also well-supported by Windows Services
- Watch for further adoption and integration