Flash Memory Summit

Persistent Memory Applications

Persistent Memory Track

Panel Chairperson: Steffen Hellmold

VP Technology Strategy, Western Digital

Persistent Memory Application Panel

August 9th 9:45am – 10:50am

Panel Presenters:

- Adrian Jackson, Research Architect, EPCC
- Tom Talpey, Architect, Microsoft
- Dan Williams, Software Engineer, Intel
- Richard Bruner, Principal Engineer, VMware
- Renu Raman, VP Cloud Arch. & Eng., SAP

Chickens and Eggs

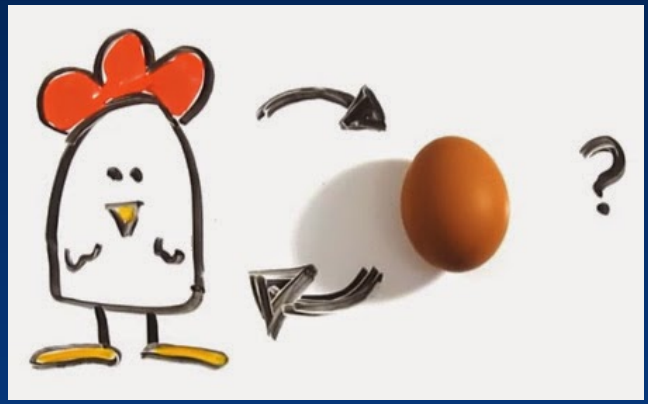
- Adrian Jackson, Research Architect, EPCC

- **Speaker Bio:** Adrian Jackson helps computational scientists use large scale parallel computers as efficiently as possible. Often this involves parallel programming (MPI, OpenMP, etc...), mainly on Fortran codes. He is currently helping design software to exploit NVRAM for HPC systems and developing a prototype HPC system with NVRAM to evaluate the hardware for large scale computational simulation and data analytics.

- **Presentation summary:** In this talk Adrian will discuss the challenges of I/O for large scale parallel computers, the potential benefits that NVRAM can bring, and the challenges that in-node storage presents both to applications and to system ware/system managers.

NEXTGenIO summary

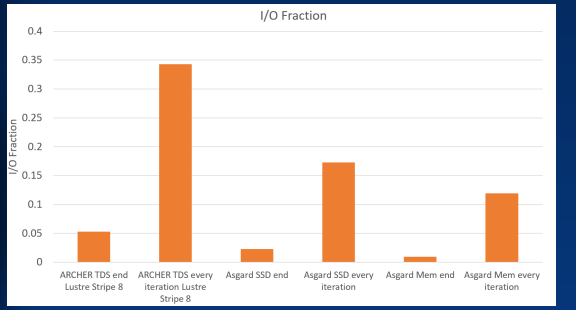
- Project
  - Design, develop, and exploit HPC and HPDA system with NVRAM in compute nodes
  - 36 month duration
  - €8.1 million
  - Approx. 50% committed to hardware development
  - http://www.nextgenio.eu/
  - This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671951.
- Partners

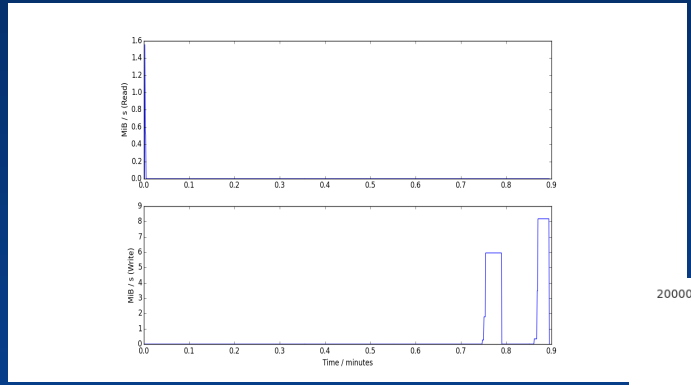


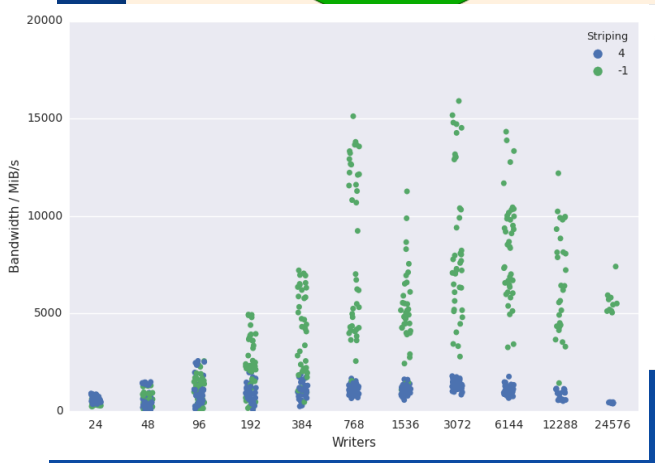
Adrian Jackson
@adrianjhpc
a.jackson@epcc.ed.ac.uk
EPCC, The University of Edinburgh

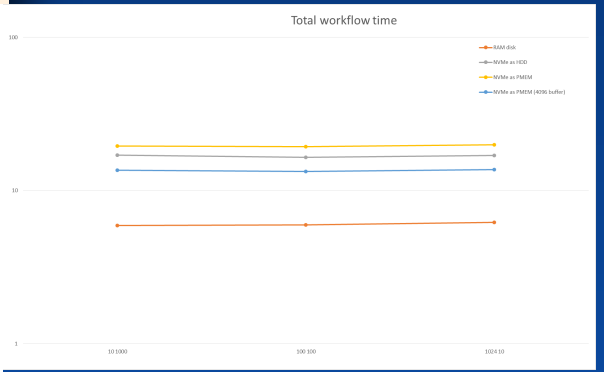
Flash Memory Summit

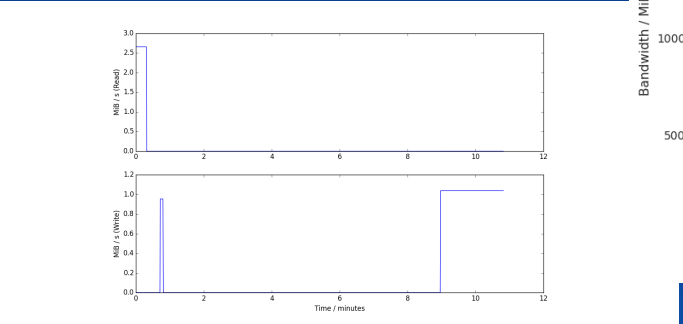
HPC I/O



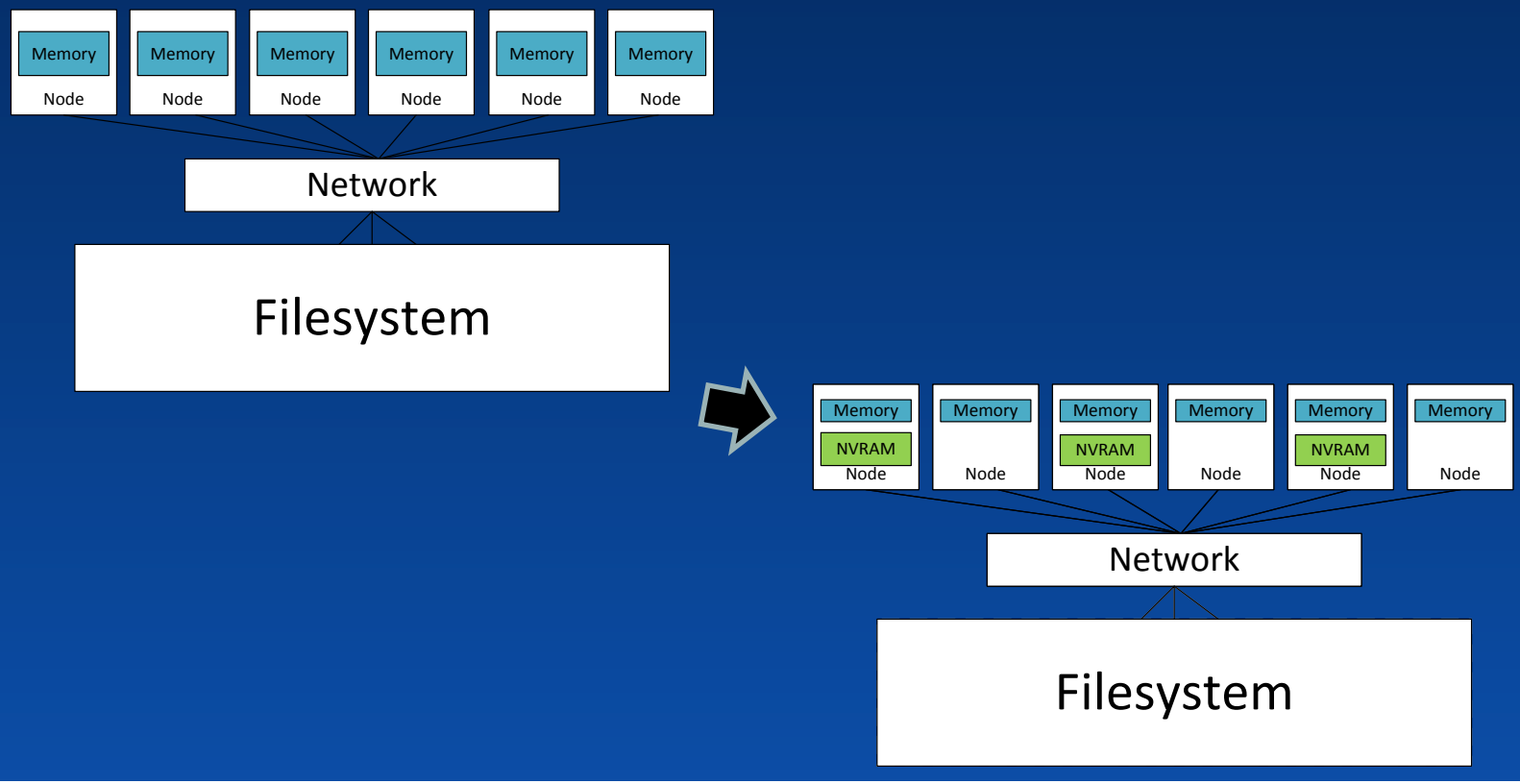






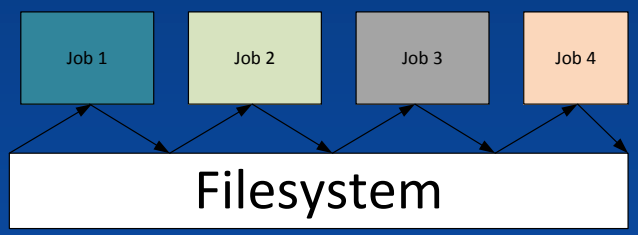


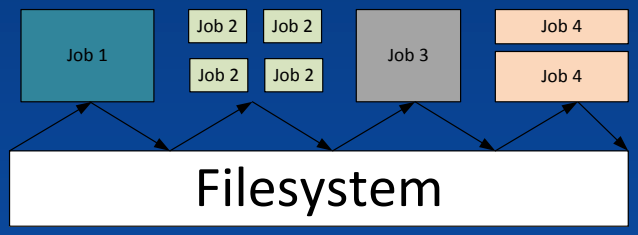
Exploiting distributed storage



Using distributed storage

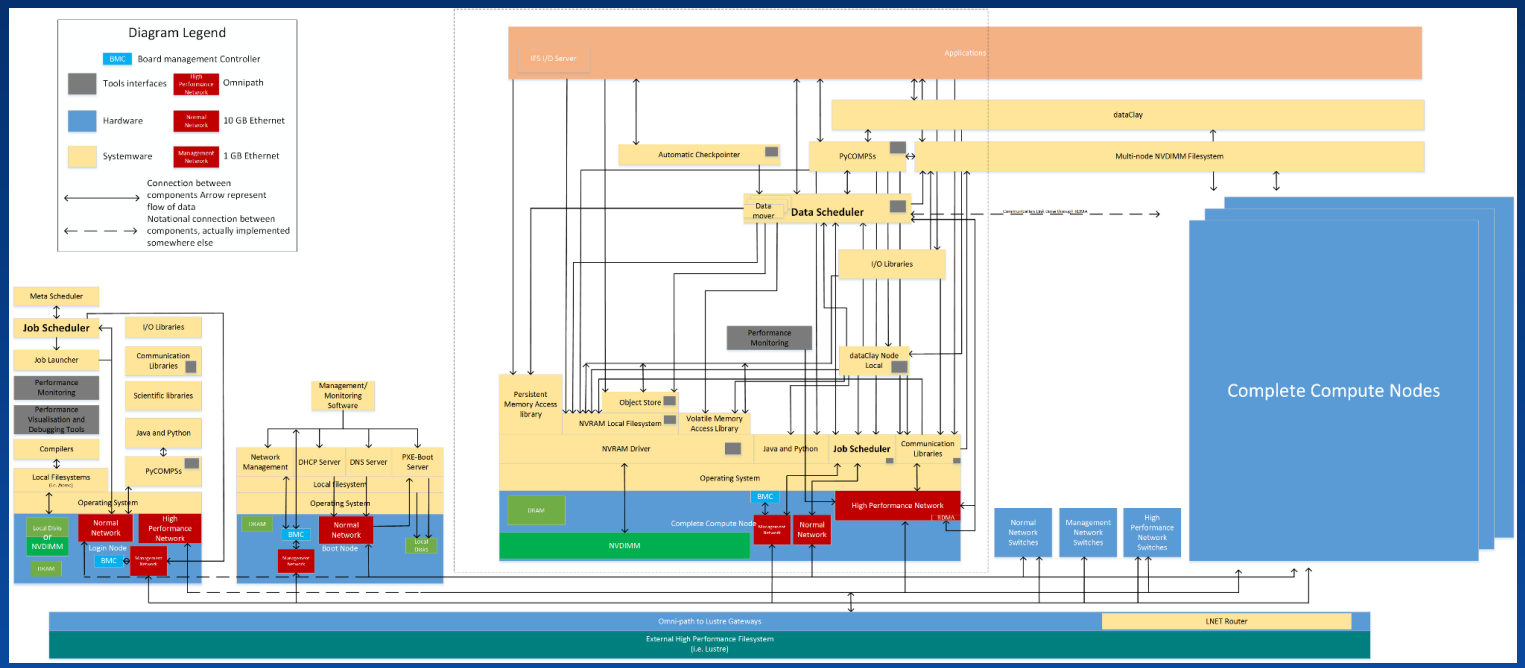
- New usage models
  - Resident data sets
    - Sharing preloaded data across a range of jobs
    - Data analytic workflows
    - How to control access/authorisation/security/etc….?
  - Workflows
    - Producer-consumer model





    - Remove filesystem from intermediate stages

NEXTGenIO Systemware



Microsoft product support for Persistent Memory technology

- Tom Talpey, Architect, Microsoft

- **Speaker Bio:** Tom Talpey is an Architect in the File Server Team at Microsoft. His responsibilities include SMB 3, SMB Direct (SMB over RDMA), and all the protocols and technologies that support the SMB ecosystem. Tom has worked for many years in the areas of network filesystems, network transports and RDMA, and recently has been working on applying these to remote access of Persistent Memory, working within the SNIA Nonvolatile Memory Programming TWG, and others. He is a frequent presenter at SNIA and storage networking events.

- **Presentation summary:** In this brief introduction, Microsoft product support for Persistent Memory technology will be reviewed. This support first appears in Windows 10 Anniversary Update and Windows Server 2016, released last Fall. Microsoft Windows-based application support for persistent memory will also be presented.

Windows Persistent Memory Support

- Persistent Memory is supported in Windows 10 and Windows Server 2016
  - Foundational in **all Windows SKUs**
- Support for JEDEC-defined NVDIMM-N devices in:
  - Windows Server 2016
  - Windows 10 Anniversary Update (Fall 2016)
  - … and all future editions
- Architecture in place to support future PMEM technologies
  - Standards-based driver framework
  - “Storage Class Memory” (SCM) Device, Bus and Disk drivers

Windows PMEM Access Methods

- Direct Access (DAX) Filesystem
  - Mapped files with load/store/flush paradigm
  - Cached and noncached with read/write paradigm
  - Supported via NTFS with “DAX” on-disk flag
    - → `Format n: /dax /q`
    - → `Format-Volume –DriveLetter n –IsDAX $true`
- Block-mode (“persistent ramdisk”)
  - Raw disk paradigm
  - Supported via SCMDisk.sys driver
  - Optional Block Translation Table (“BTT”)
- Application interfaces
  - Mapped and traditional file Win32
  - NVM Programming Library (NVML)
  - “PMEM-aware” open coded

Direct Access Architecture

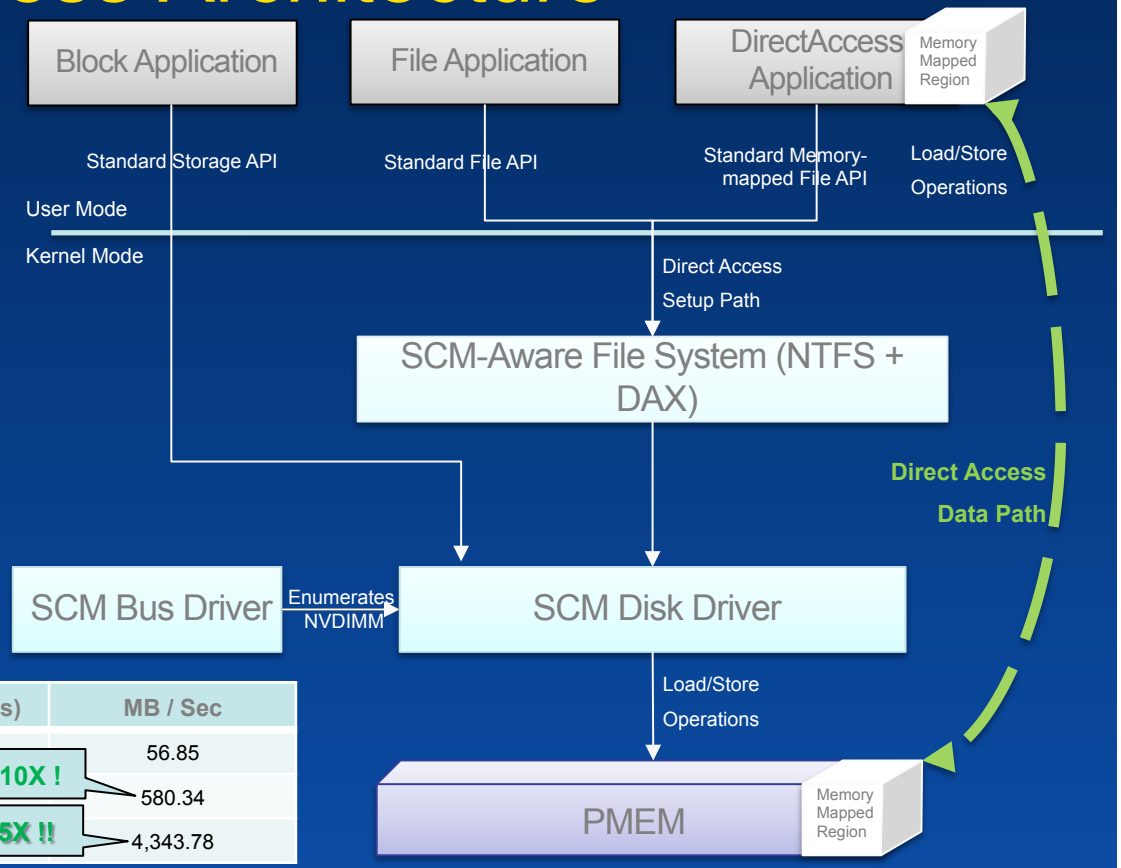
Flash Memory Summit

Overview

- Application has direct access to PMEM via existing memory-mapping semantics
- Updates **directly modify** memory, Storage Stack is not involved
- DAX volumes identified through new flag
- Traditional Block, File Apps can continue to access transparently (but without true direct access)

DAX Characteristics

- True device performance (no software overhead)
- Byte-Addressable
- Filter Drivers may require modification (no I/O requests issued, new volume flag)
- AntiVirus filters can still operate (Windows Defender already updated)



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

10X !

75X !!

Windows PMEM Consumers

- Any Win32 app: Block, File
  - ReadFile, WriteFile, MapViewOfFile / FlushViewOfFile
- Applications coded to NVML Open Source NVM Programming Library v1.3
  - http://pmem.io >> https://github.com/pmem/nvml/releases
- SMB3 / SMB Direct (RDMA)
  - Exports DAX file shares (performance optimizations ongoing)
  - Storage Spaces Direct
    - Block mode PMEM supported in Fall Windows Server update
    - **1 TiB/sec!! >>** https://blogs.technet.microsoft.com/filecab/2016/10/17/storage-spaces-direct-with-persistent-memory/
- Hyper-V
  - Virtual PMEM support in Fall Windows Server update
    - https://blogs.windows.com/windowsexperience/2017/08/08/announcing-windows-server-insider-preview-build-16257/
    - vPMEM in DAX mode (preferred) or optionally BTT mode
- SQL Server
  - "Tail-of-log" (or more) supported on DAX for **2X+ transaction** throughput
    - https://blogs.msdn.microsoft.com/sqlserverstorageengine/2016/12/02/transaction-commit-latency-acceleration-using-storage-class-memory-in-windows-server-2016sql-server-2016-sp1/
- Azure
  - Persistent memory for write commit log

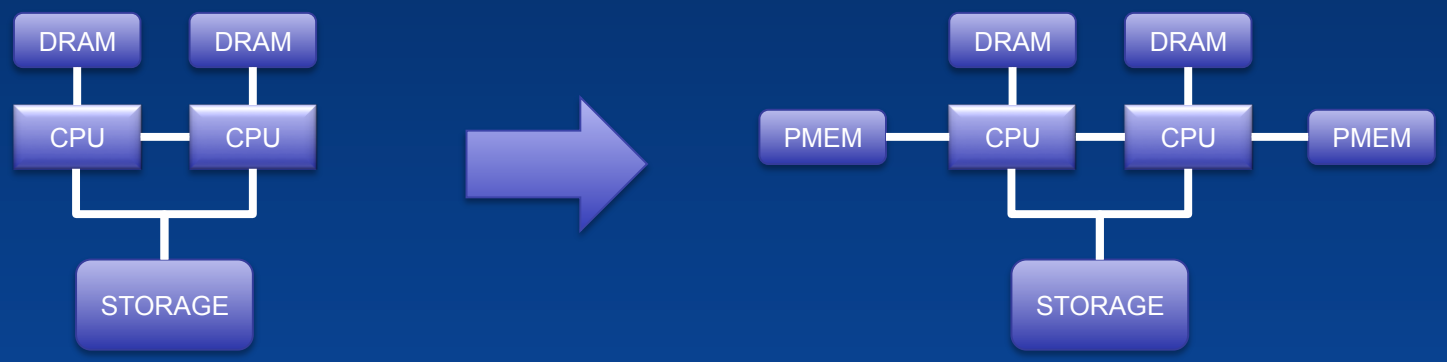
The future is DAX

- Dan Williams, Software Engineer, Open Source Technology Center, Intel

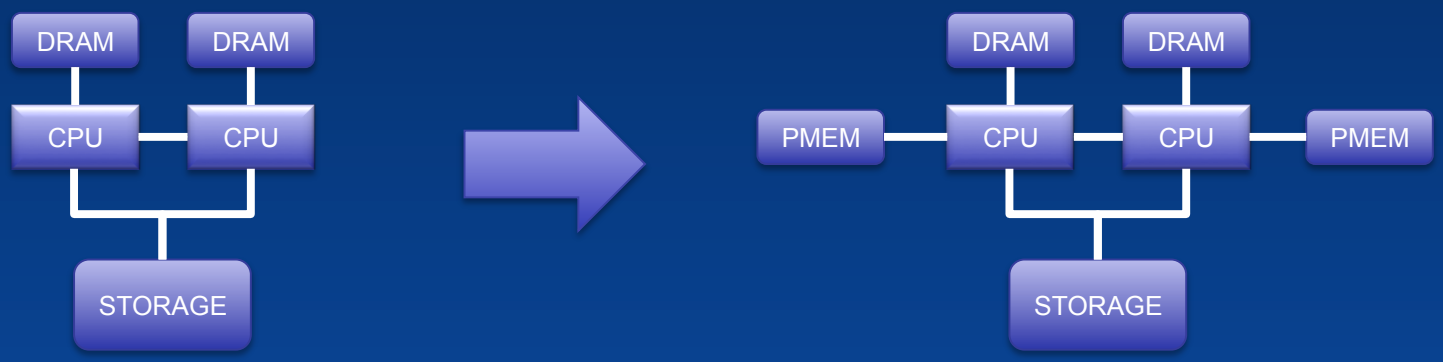
- **Speaker Bio:** Dan Williams is a core Linux kernel developer and a member of the Linux Foundation Technical Advisory Board. He is a lead engineer / maintainer of the Linux “NVDIMM” kernel sub-system and accompanying utilities supporting persistent memory devices. His career enabling storage technologies spans 15 years with time at both Intel and Facebook.

- **Presentation summary:** One need not look much further than the recent UEFI.org specification releases to understand that operating systems will need to contend with a range of performance and feature-differentiated memory types in the near future. DAX is growing set of Linux kernel mechanisms to provide unfettered access to these memories. This talk will introduce some of the challenges and opportunities DAX affords next generation applications.

Memory in an ACPI 6.2 World



Memory in an ACPI 6.2 World



DAX allows an application to find and directly access the memory it needs

DAX Opportunities

- Software defined memory
- Update in place algorithms
- Operating system latency bypass

DAX Challenges

- Software defined memory
- Update in place algorithms
- Operating system latency bypass

VMware's virtualization direction for upcoming PM technology

- Richard Brunner, Principal Engineer, Chief Platform Architect, Server Platform Technologies, VMware

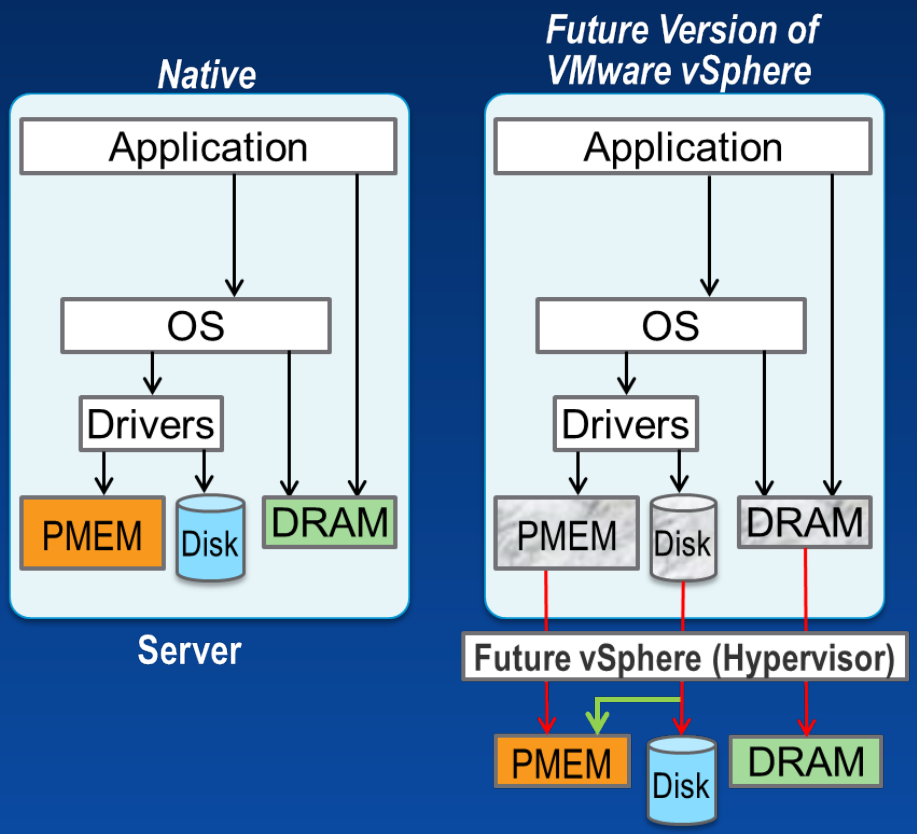
- **Speaker Bio:** Richard A. Brunner is a Principal Engineer and the Chief Platform Architect at VMware. He is responsible for defining and driving future platform features that can benefit virtualization. Richard has over 33 years of experience in the computer industry working as an architect for IBM, DEC, Intel, and AMD. He has been focusing on Persistent Memory as of late.

- **Presentation summary:** Byte-addressable non-volatile memory (PM) products finally meet real-world applications, hypervisors, and soon evaluation by VMware's customers. Virtualization is fundamental to workload agility and consolidation for these new applications. We will briefly review VMware's virtualization direction for upcoming PM technology.

Virtualization direction for upcoming PM technology

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.
  - In a future version of vSphere, the VM doesn't need a special driver.
  - Guest Storage can be mapped to PMEM outside of VM.



Virtualization direction for upcoming PM technology

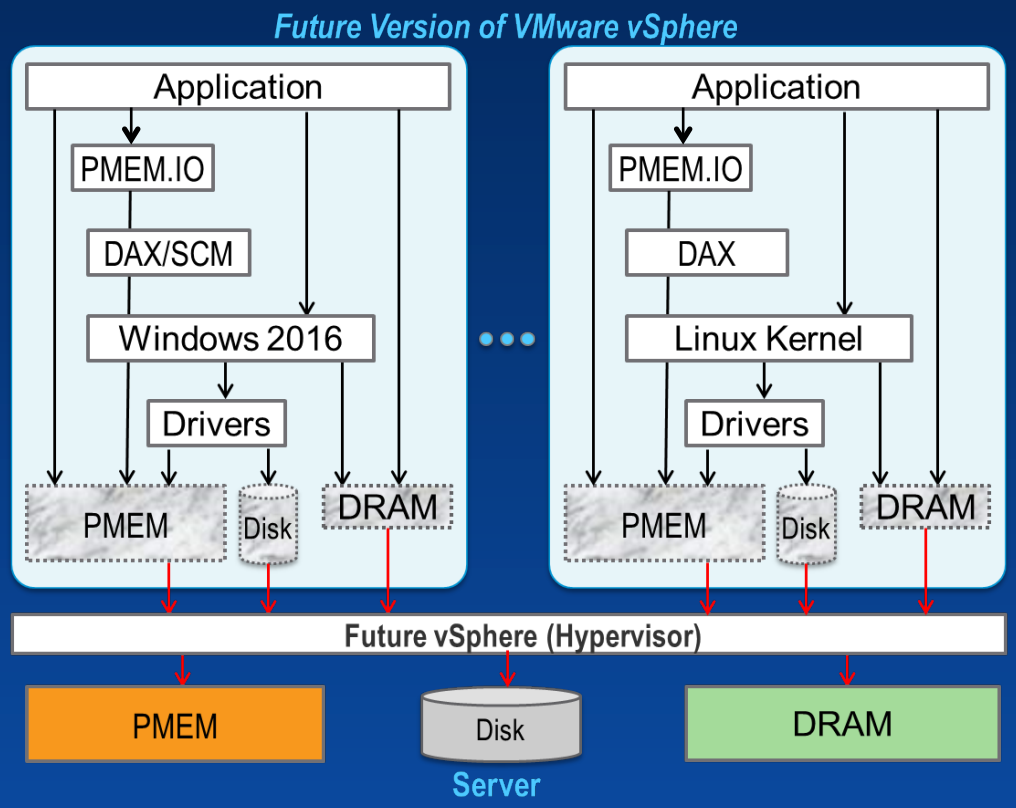
New OS & Application Usage

Native & Virtualized

- Can use a direct load/store model with little OS overhead

All the benefits of VMware vSphere can be available:

- Multiple workloads using PMEM
- Live VM Migration across servers
- Check-pointing
- Boost for LegacyVMs/Workloads
- And More …



HPE Persistent Memory and VMware

Up to 3x
Higher Performance with 1/3 the latency!

Updates/sec – **3x performance gain**

Storage Type	Total Row Updates	Updates/sec	Update Latency (µs)
Log on SSD	725729	36257	28
Log on DAX SCM Volume (8GB)	2106945	105338	9

Future Version of VMware vSphere with HPE ProLiant servers with HPE Persistent Memory **SQL Server 2016 Tail-of-the-Log**

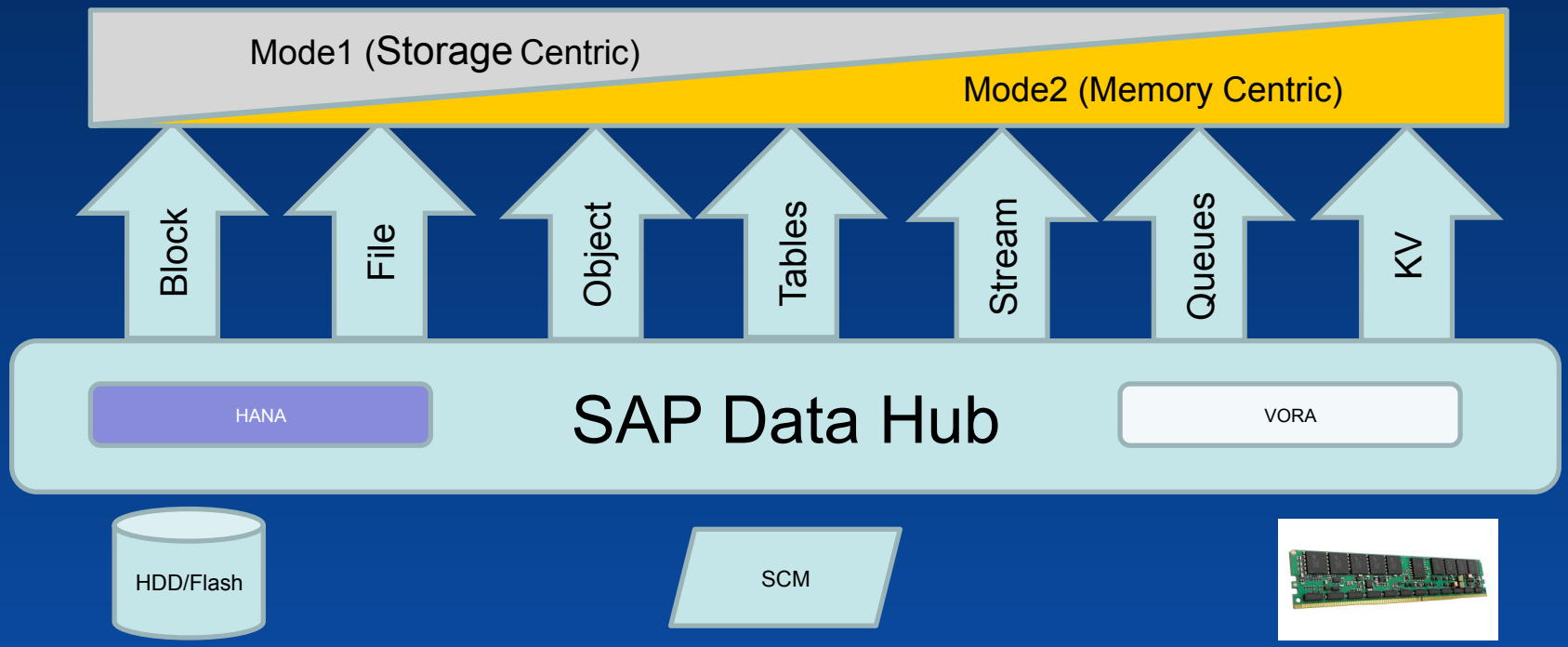
Persistent memory in SAP's data management platform

- Renu Raman, VP Cloud Arch. & Eng., SAP

- **Speaker Bio:** Renu Raman is Vice President in SAP's cloud architecture and engineering group responsible for computing and storage platforms. He has a dual role of both architecting and managing the development and delivery of SAP's compute and storage infrastructure. Prior to SAP, he founded Unity Microsystems and was an executive in residence at Tallwood Venture Capital. Renu's professional career started at Sun Microsystems spanning more than a decade. Renu brings a diverse background ranging from computer system architecture, semiconductors (processors) , Storage, networking and Computer aided design (CAD).

- **Presentation summary:** SAP is the in-memory cloud leader with HANA being the leading in-memory data management platform. This talk with share the architecture of SAP's data management platform and the key technologies and use cases that drive both the design and adoption of the data management platform. With the emergence of Flash and now 3D Xpoint, this talk will talk about current uses of DRAM and Flash and with 3D Xpoint the initial use case and share the architectural directions of SAP's cloud platform with the availability the new memory tier.

SAP Data Strategy



Highly Distributed Data Store enabling continuous availability

Flash Memory Summit

Bringing Persistent Memory to in-memory computing

Opportunities

- Increased scalability
  - Larger memory modules means more memory available per server
- Significant cost savings
  - PM is cheaper than DRAM
- Improved recovery

Challenges

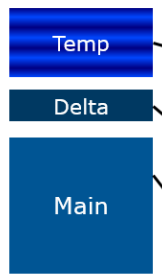
- Higher (than DRAM) latency impacting performance
- New technology, standards still evolving…
  - Means slow, phased implementation with increased complexity and uncertain timelines

Evolving the HANA SW Architecture to incorporate 3D XPoint™ PM technology

Flash Memory Summit

Main Store is the perfect fit for 3DXPoint™ technology!

SAP HANA: Memory Architecture

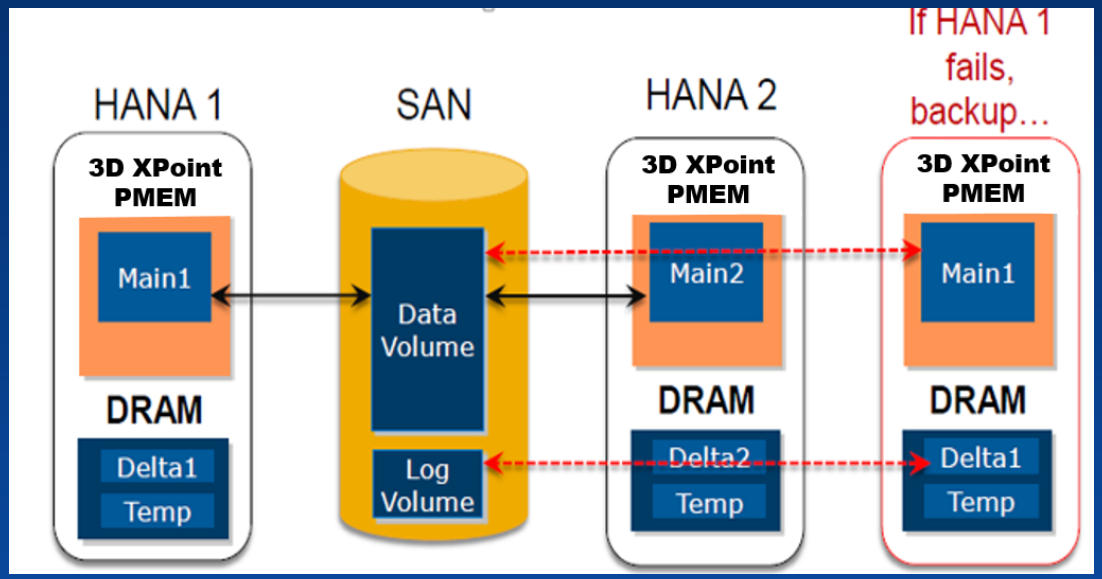


Function	Organization	Access type
Scratch pad for intermediate results	Scratch pad space	Reads and writes
Portions of data that has been modified	~5% of data, light compression	Reads and writes
Data in memory format	~95% of data, 10-20X compressed	Reads (scans)

Main store contains ~95% of the data in highly compressed format.

	Technology differentiators	Why Main is well suited
+	Large capacity	Since 95% of data contained in main, HANA can scale up to larger datasets due to increased memory capacities
+	Persistence	Avoid loading data from storage and reduce downtime
–	Higher latencies	References are in form of scans. Hardware and software prefetchers can hide latencies for such reference patterns

Leveraging 3D XPoint™ PM technology for SAP HANA Main Store: Solution Overview



- Primary data store is data volume (in SAN or local storage)
- Main is in 3D XPoint™ PM instead of DRAM and is now persistent
- On Restart: Main already in 3D XPoint™ PM, no need to load data from SAN
- On HW failure: Backup server loads data from data volume

Leveraging 3D XPoint™ PM technology for SAP HANA Main Store: Pros and Cons

Flash Memory Summit

PROS +

- Significant improvements in database restart time (>5x)
  - No need anymore to load data, Main Store is now persistent as it is in 3D XPoint™ PM instead of DRAM
- Increased memory capacity at a lower cost
  - Potentially significant cost benefits in several areas

CONS −

- New approach has two redundant persistent copies
  - Dual-writes to 3D XPoint™ PM and "SAN"/Disks
    - Necessary for Backup and Recovery using traditional persistence; also to safeguard against corruptions in persistent memory DIMMs
- 3D XPoint™ technology is still evolving, making s/w design tricky
  - When to operate directly on "PM" vs "DRAM" ? (Data storage vs intermediate data handling
    - Both storages co-exist in DDR4 form factor, but the capacity ratio is unknown