SNIA PERSISTENT MEMORY + SUMMIT 2021 COMPUTATIONAL STORAGE

FROM DATACENTER TO EDGE : VIRTUAL EVENT
APRIL 21-22, 2021

The Persistent Memory Connection

How to Attach PMEM in Computing Systems?

Jonathan Hinkle

Executive Director and Distinguished Researcher – System Architecture, Lenovo

The Persistent Memory Connection

Agenda:

- Background basics and the lurking problem
- Persistent Memory is part of the solution, but… how to attach it?
- Step one – Start with the existing interfaces and slots for memory today
- Step two – How can we optimize PM in computing systems and solve next generation system problems?
- Preliminary Results

Computing System Basic Architecture

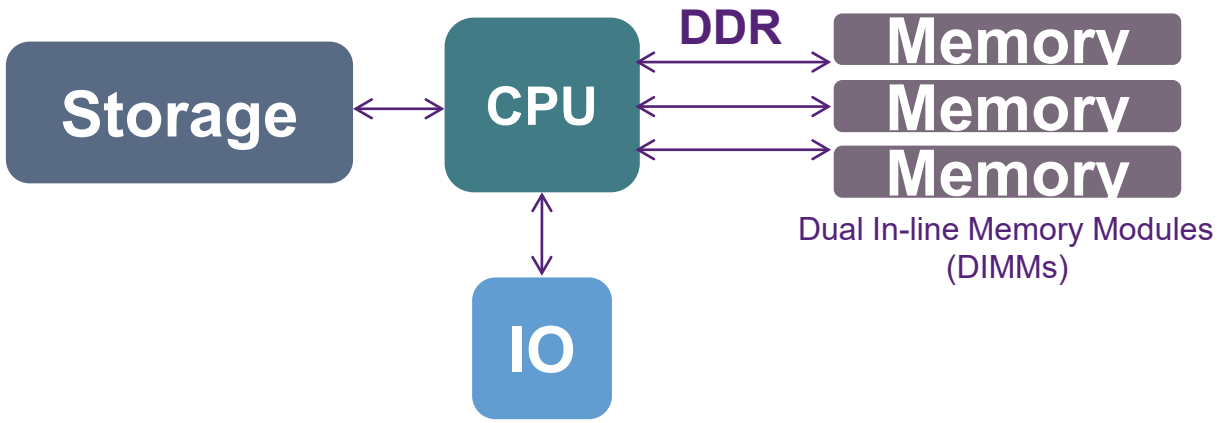
Data Center systems

DIMM

Key memory needs:

High capacity

High performance

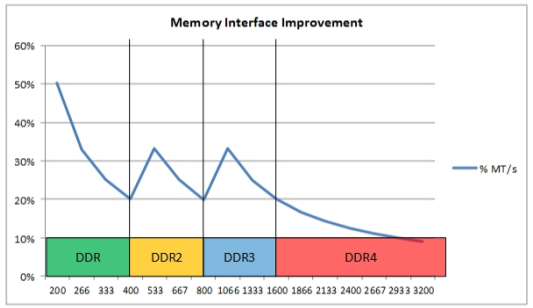


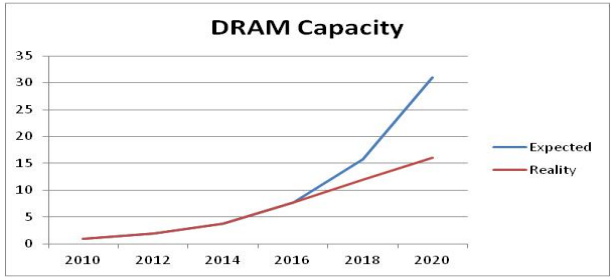
Dual In-line Memory Modules (DIMMs)

- Both client and data center systems typically have similar architectures, though different in scale
- Systems include CPU, memory, storage and IO as main functional sub-systems
- For decades, memory in datacenter systems has been added on the DDR electrical interface as DRAM modules called **DIMMs**

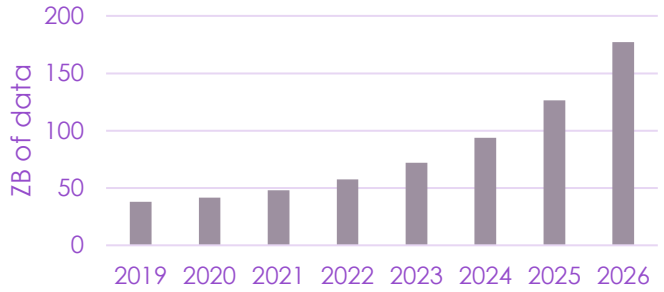
Datacenter System Memory

DRAM scaling is significantly slowing along with Moore's law.





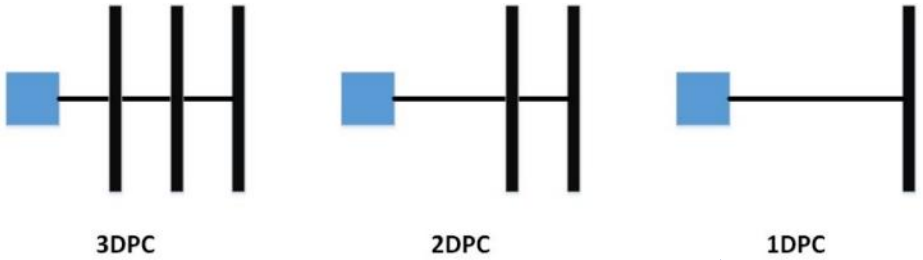
Annual Data Growth



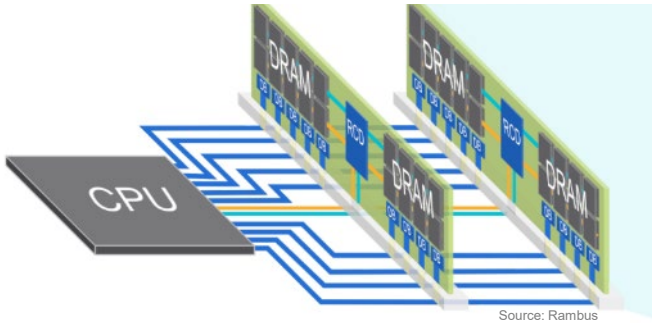
is driving

increasing demand for memory capacity

Data Center Systems must now often sacrifice performance to add memory capacity.



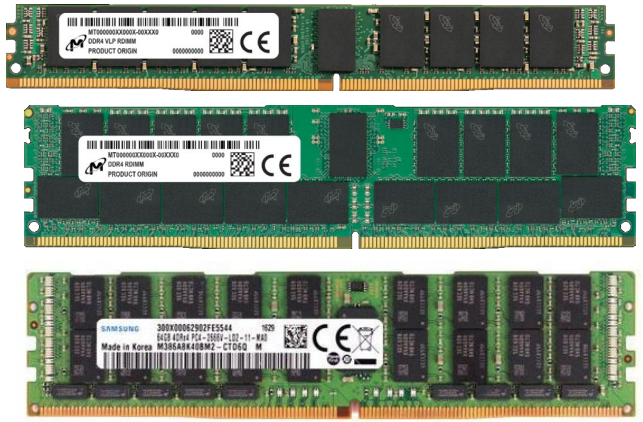
DIMMs Per Channel with frequency

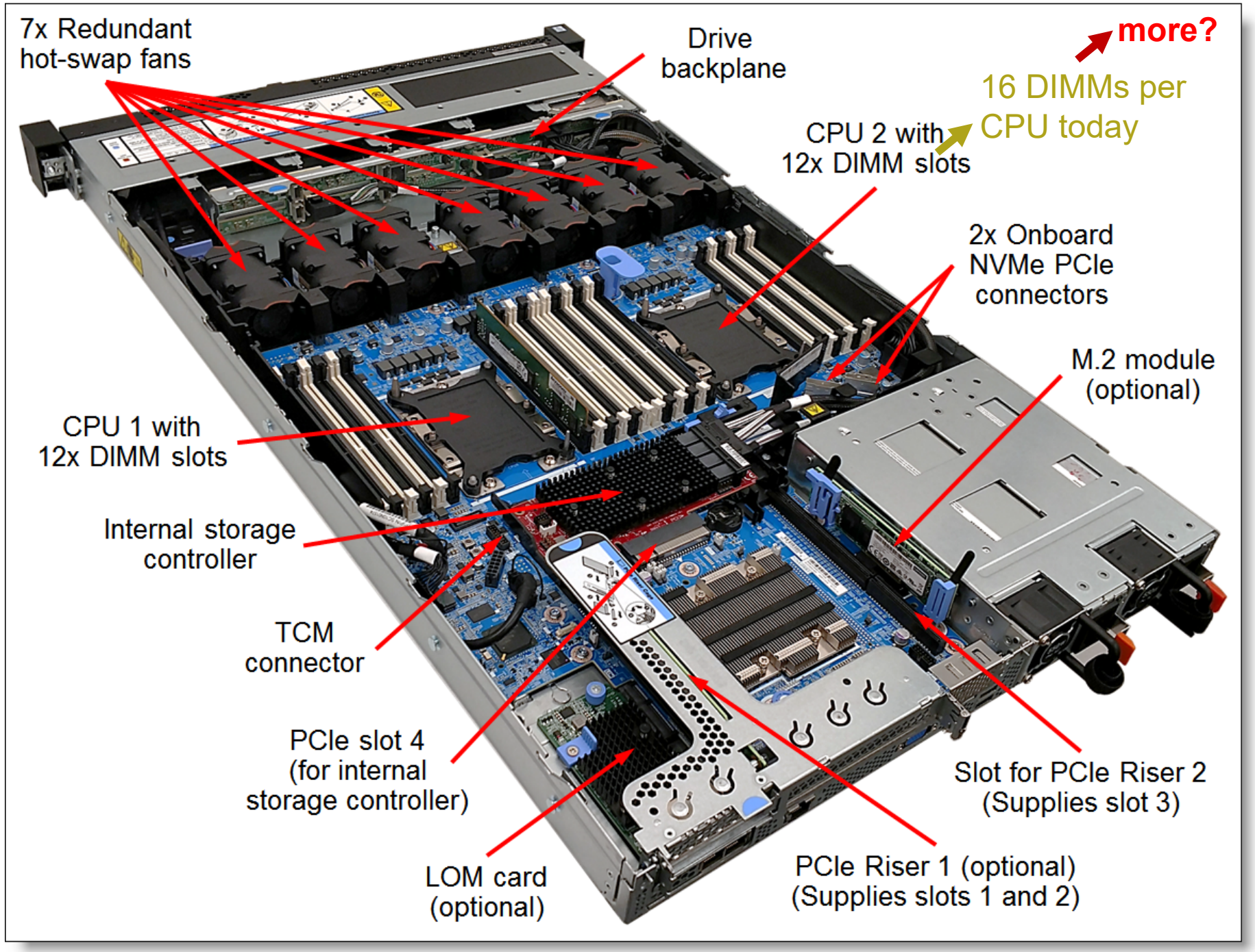


Source: Rambus

Next generation memory like DDR5 DIMMs is evolutionary and does not fix this issue. We still must address this fundamental problem of memory speed vs. capacity.

Systems are also hitting significant challenges in physical space and cooling





Implications for Datacenter

- Without a strong response, systems won't keep the current pace of improvement and significant value that can be extracted from the data will be lost.

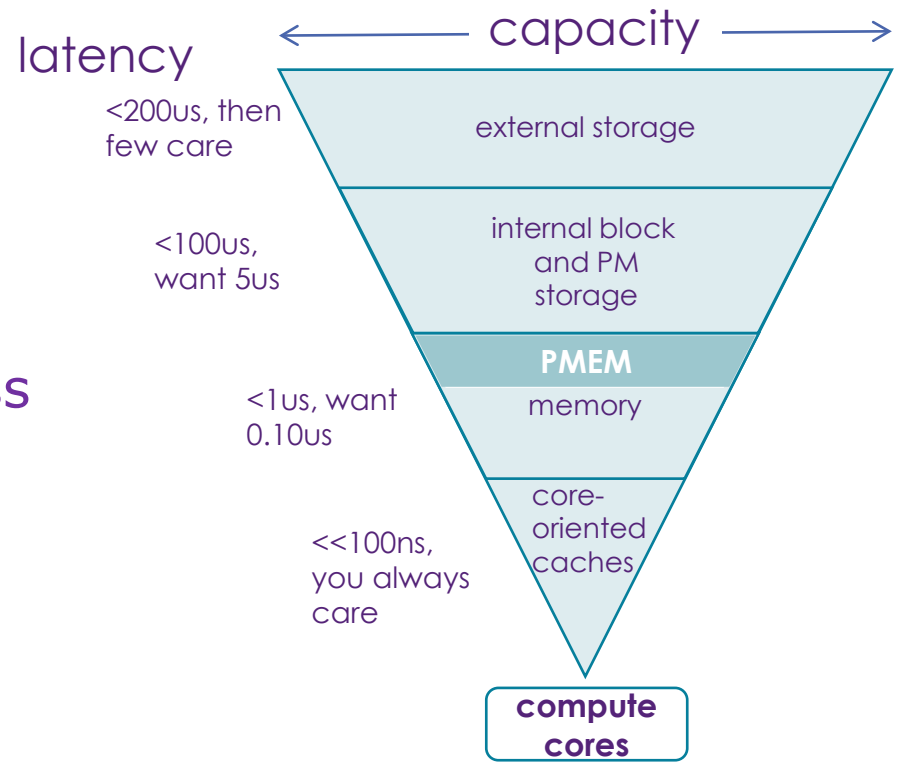
- Considering very strong data growth (Big Data) and further rising numbers of users, cloud and enterprise could significantly miss meeting global IT needs.

Must improve:

Memory { Performance, Capacity, Power/cooling/space }

Persistent Memory – part of the solution

- Persistent memory promises to help fill the gaps in future progress and drive further system improvement.
- Often Persistent Memories offer new advantages for main memory versus only DRAM, such as **higher capacity** and **lower cost per capacity** to help address upcoming system challenges.
- Beyond those advantages, the **non-volatility** itself provides a **new disruptive benefit** and enables significant speed-ups in application performance along with new system capabilities.
- Still, new Persistent Memory technologies often don't follow the same rules as DRAM for endurance, error rate and management required, power, size, and timings along with different performance and capacity.



basic data store hierarchy

How shall we plug them into our systems?

Step one

Start with existing interfaces and slots in systems

Attach as a NVMe drive

- PCIe-attached SSDs are easiest method to plug in Persistent Memory
  - Just use as a very fast storage device
  - Can leverage in existing systems with traditional slots or scale-out with next gen form factors
  - Existing PCIe interface to system and known NVMe driver / software

- As an SSD, PM can scale-out well like flash drives, but for long time (5+ years) will be too expensive/capacity for main storage, like all-flash vs HDD many years ago.
- The block IO interface and software stack significantly constrain the potential performance of byte-addressable Persistent Memory.

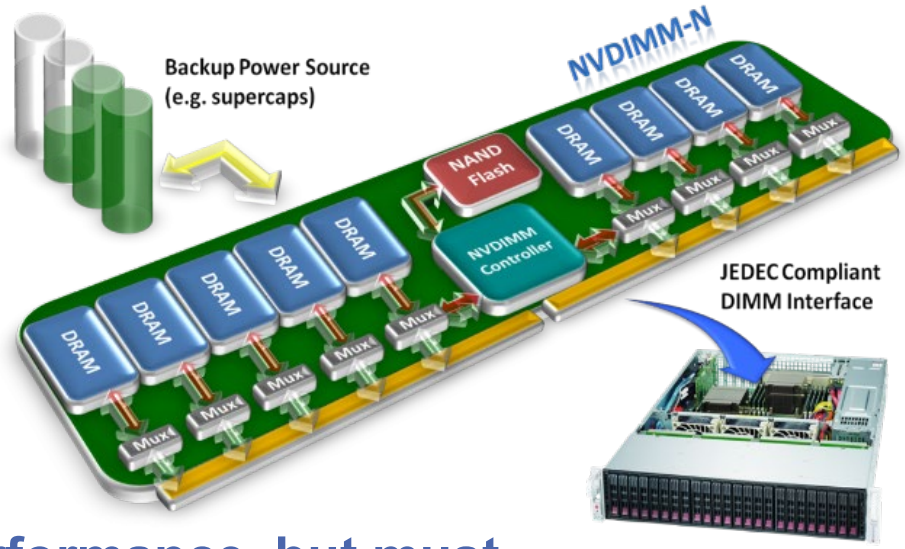
Would like this ease of use and scaling, but with full performance of Persistent Memory

Attach as a DIMM: NVDIMM-N

NVDIMM-N: 1st Persistent Memory Module with DRAM and Flash, providing persistent memory with the capacity and performance of DRAM.

- It doesn't slow down the memory sub-system
- Plug into same memory controller – transparent and play by DDR rules
- Very useful for storage metadata, database journaling and recovery logs

- Unfortunately, the module cost = DRAM + NAND Flash + NVDIMM logic + backup energy source
- NVDIMM-N capacity is low compared to TBs of storage, so not as useful for caching.
- Can't access NVM (Flash) during run-time – paying for it, but can't use it most of the time.



Source: SNIA NVDIMM Cookbook

Would love to have this performance, but must be able to scale capacity and at lower cost

Use Existing System Interfaces and Slots

Recap of Existing system interfaces and device slots:

- PCIe/NVMe drive slots (Flash memory/storage): great scalability, simplicity and flexibility, but high added latency and no byte-access for memory (limits performance)
- DDR4 DIMM slots (DRAM/main memory): lowest latency interface and massive bandwidth, but limited capacity and scalability (limits capacity)

Plug into existing system slots: ✓completed

…but leaves a lot of the PM value on the floor

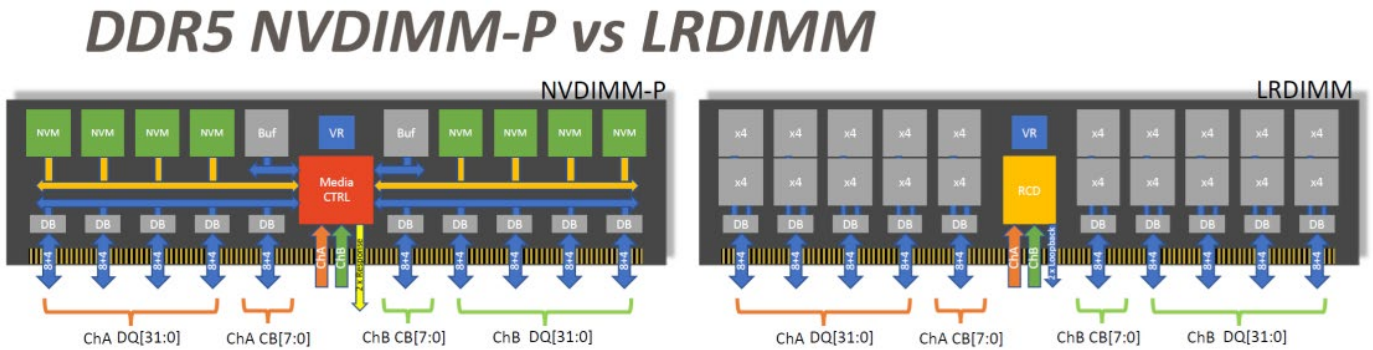
Step two

How can we optimize PM in Computing Systems?

NVDIMM-P *(and Intel Optane DIMMs)*

NVDIMM-P: Persistent Memory module and system interface optimized for attaching emerging Persistent Memory in DIMM slots and to run alongside DDR DRAM modules.

- Fully byte-addressable at lowest latency and massive bandwidth for highest performance
- Hybrid NVDIMM-P modules can leverage PM and DRAM
- 2-10X+ capacity per slot over DRAM, often at lower cost



Source: JEDEC

New interface designed to optimally attach PM as a DDR DIMM

DDR4 NVDIMM-P JEDEC standard

NVDIMM-P standard enables:

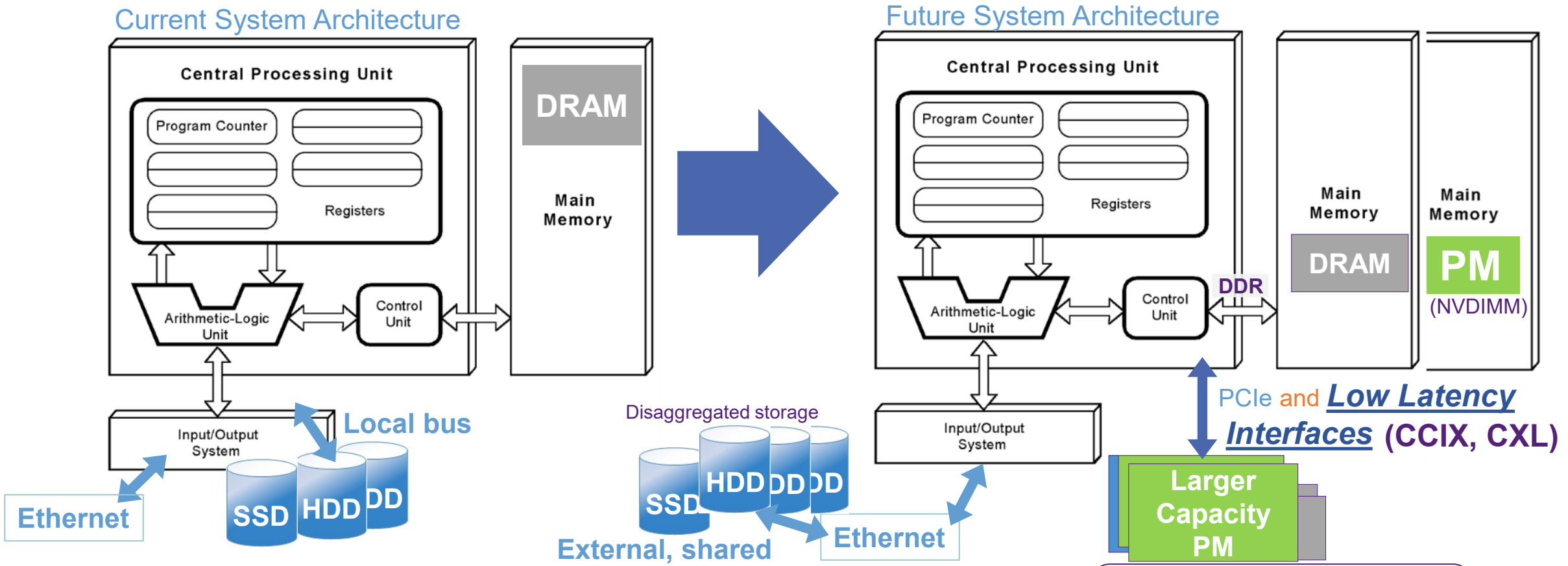
- **Persistence**: lowest latency and high bandwidth access to Persistent Memory modules in a system
- **Abstraction of Memory Media**: enabling most any memory media on the DDR channel
- **Higher Memory Capacity**: supports expanded memory addressing
- **Plug and Play Interoperability**: physically plugs into standard dual in-line memory module (DIMM) sockets and run-time interoperable with DDR DRAM DIMMs on the same bus

Key features include:

- Fully compatible with existing DDR channels (Physicals, Electricals, Protocol, Clocking)
- Minimal to no pin adder in CPU socket
- Protocol support of non-deterministic latency for data reads
- Transactional operations for ensuring data is preserved in Persistent Memory
- Latency support from NAND Flash down to DRAM latency (at module level)
- Self-contained reliability + link error protection

New NVDIMMs provide needed capacity at highest performance, but systems are still space constrained.. how can we scale and get even more memory?

A New Systems Architecture for DRAM and Persistent Memory Expansion



Persistent Memory (PM): Promises *lower* latency than NAND Flash and/or higher capacity than DRAM.

NVMe SSDs + ***scalable, expanded memory***

New low-latency fabric technologies enable a new method of expanding system memory resources.

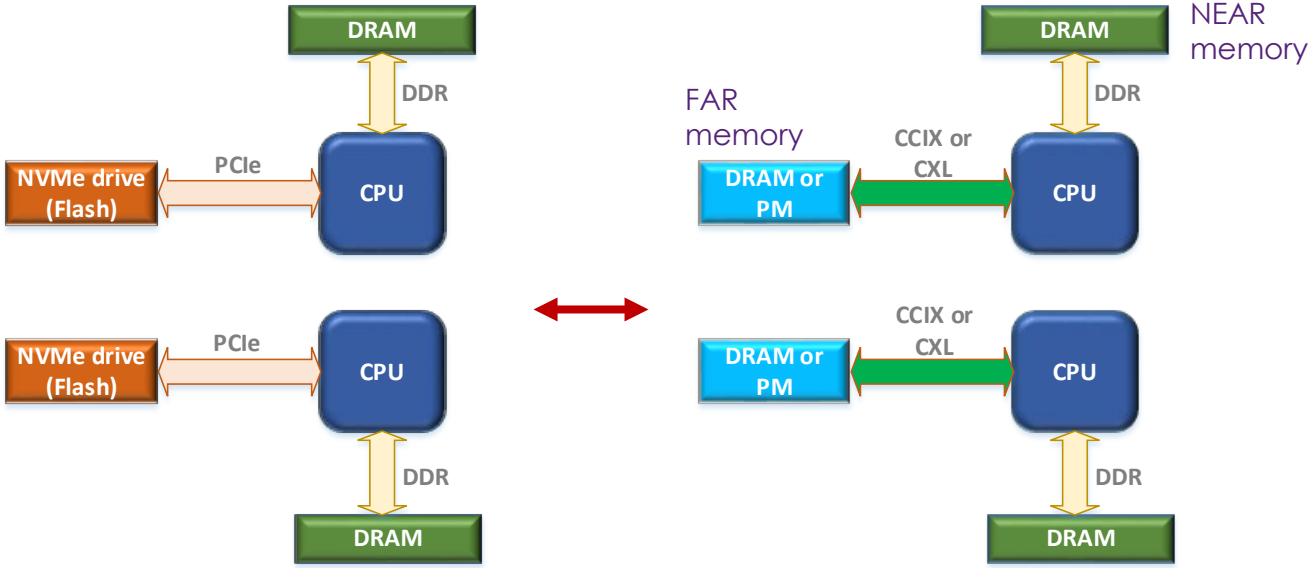
Low-latency fabric interfaces

- Emerging computer interfaces such as CCIX and CXL can provide time to access latency that is low enough for data in memory.
- There are new benefits that can be realized because these interfaces can leverage the same wires and interconnect as PCIe, enabling interchangeable slots for devices.
- Form factors for NVMe drives (attached by PCIe) such as EDSFF* are perfect fit to enable scaling of memory devices in a system.

CCIX is available today for prototyping, so our current lab work is based on it. Memory expansion with this system architecture works well in the same method with either CXL or CCIX.

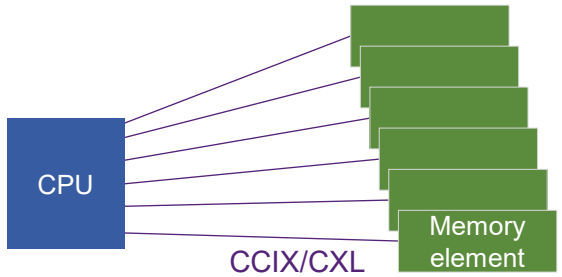
© 2021 SNIA Persistent Memory + Computational Storage Summit. All Rights Reserved.

Lenovo Research System Architecture for Memory Expansion



EDSFF E1.S drive slots for PCIe-attached NVMe Flash

Same slots leveraged for CCIX or CXL attached DRAM & PM



Link interface enables further distance connections to more memory elements
- additional memory capacity

Interface latency of nanoseconds enables appropriate performance
- at memory latency

As devices added to scale capacity, access bandwidth grows as well.
- scale-out bandwidth

Memory accesses can still be made in 100's of ns instead of microseconds, while enabling expansion for up to TBs of data without scaling up cost.

This translates into higher system performance and lower cost for customers running memory and storage intensive applications.

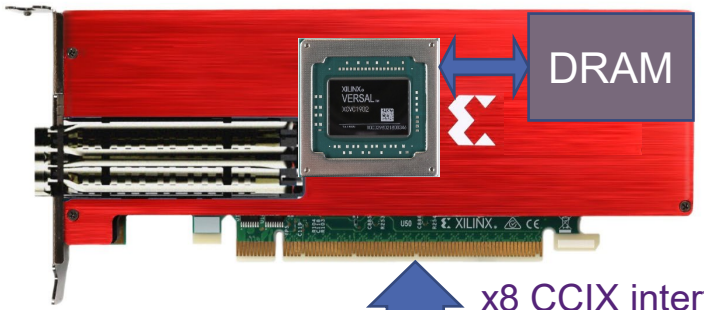
Results!

First look at some measured results for expanded memory

Lenovo Research 1st Prototype System

XILINX®

Versal ACAP
CCIX to DDR4
controller



x8 CCIX interface @16Gbps
(BW ~ x4 @ 32Gbps)

AMD server

Investigation of prominent data center workloads that could benefit this most:

- Relational Database
- In-memory computing
- Big Data
- Software Defined Storage

MySQL redis OpenFOAM GIGASPACES

Initial MySQL results: 2X workload performance

Lenovo Research investigation - results

Formulus Black's FORSA software is a production, optimized ramdrive.

It provides a POSIX-compliant block interface for workloads to use the FAR memory like a normal storage device.

FIO: 4K random read, QD=1 direct=1

Ave Lat us	NVME QEMU RAW img		Lenovo prototype	
Num jobs	**1 VM**	**2 VM**	**1 VM**	**2 VM**
1	136.9	161.2	19.1	19.2
2	140.7	171.3	20.7	26.3
4	143.2	219.6	25.5	44.6
8	182.5	290.3	40.6	83.6
16	224.6	334.1	72.7	160.1

BW MB/s	NVME QEMU RAW img		Lenovo prototype	
Num jobs	**1 VM**	**2 VM**	**1 VM**	**2 VM**
1	29.8	50.6	196	409
2	57.8	95	366	601
4	114	156	585	716
8	178	237.6	752	774
16	290	402	835	812

Block device performance

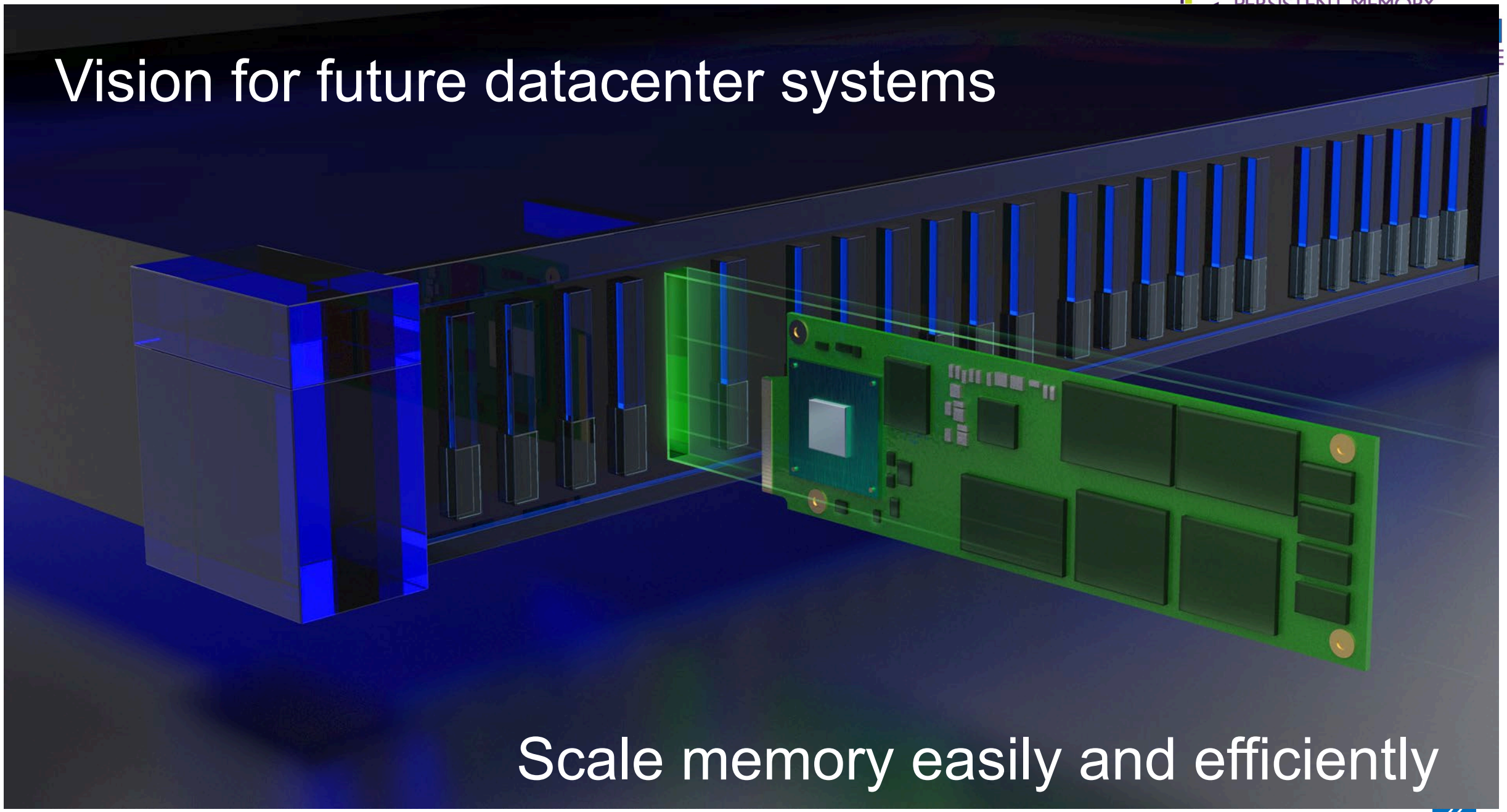
Summary

- We've attached Persistent Memory in the existing slots in systems and the 1st wave of adoption is complete.

- The industry has now completed definition of standard interfaces that are optimal for attaching Persistent Memory: NVDIMM-P and CXL.

- The 2nd wave of Persistent Memory adoption is now ready to begin, helping meet the needs for next generation memory performance and capacity.

Vision for future datacenter systems



Scale memory easily and efficiently

Thank you

Please visit www.snia.org/pm-summit for presentations