C-DAC

- Centre for Development of Advanced Computing (C-DAC) is the premier R&D organization of the Ministry of Electronics and Information Technology (MeitY) for carrying out R&D in IT, Electronics and associated areas.

Persistent memory based Storage node for HPC domain

Harikrishnan B, C-DAC

Shibu R M, C-DAC

HPC storage in General

- I/O is recognized as a performance bottleneck in HPC domains
- Disparity between computation and I/O capacity on future HPC machines to be increased.
- Need to re-evaluate old architectures
- New storage architectures and technologies to address capacity and speed requirements

Persistent memory and Storage

Persistent memory and Storage

- Storage architectures based on persistent memory
- Improves performance
- Persistent memory moves the storage closer to compute.
- Persistent memory technology allow programs to access data as memory, directly byte-addressable, while the contents are non-volatile, preserved across power cycles.
- Persistent memory has aspects that are like memory, and aspects that are like storage and is used as a third tier, in conjunction with memory and storage.

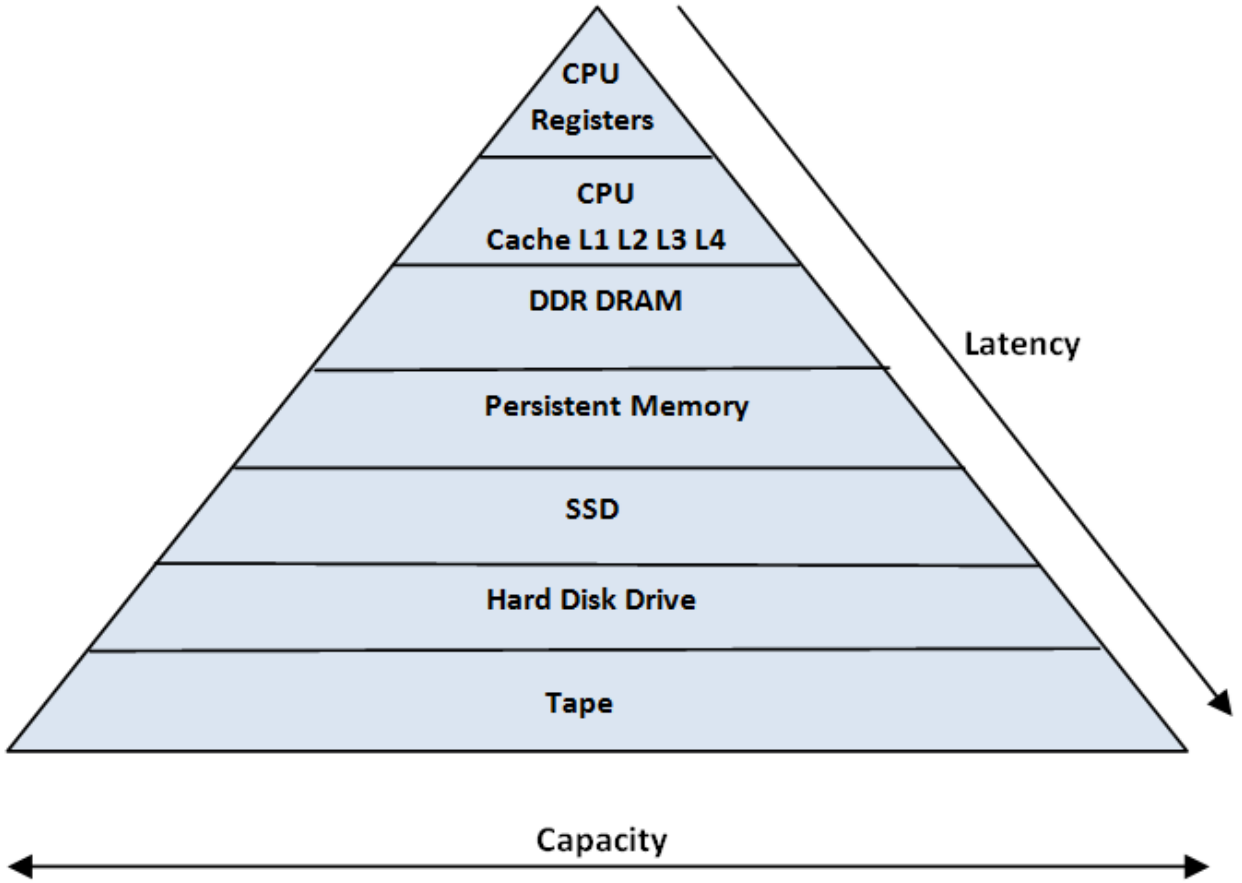
Persistent memory and Storage

- The compute and storage locality can be approached in many ways

- Stays in the host CPU (traditional)
- Compute offloaded to accelerator (FPGA, ASIC's)
- Compute moved to storage (computational storage)
- Storage moved closer to compute (persistent memory)

Memory-storage hierarchy

Memory-storage hierarchy



Memory-storage hierarchy

- Persistent memory provides applications with a new tier for data placement
- Offers greater capacity than DRAM and significantly faster performance than storage
- Applications can access persistent memory like they do with traditional memory
- To the operating system, the persistent memory looks like conventional block storage

Some of the Companies Investing in Persistent Memory

- Avalanche Technology Inc
- Crossbar Inc
- Cypress Semiconductor Corporation
- Everspin Technologies Inc
- Hewlett Packard Enterprise
- Intel Corporation
- Micron Technology Inc
- Samsung Electronics Co. Ltd
- Western Digital Corporation

Optane Persistent Memory

- Presently available persistent memory in the market from Intel
- Intel® Optane™ memory based on 3D XPoint™ technology
- 3D Xpoint - Non-volatile memory (NVM) technology - Developed jointly by Intel and Micron Technology
- Optane persistent memory has two modes of operation - Volatile and Persistent

Storage Node

Storage Node

- A typical application of persistent memory - use in the HPC storage node
- Contribute to the improvement of the storage bottlenecks on future HPC machines
- Storage node based on Intel® Xeon® and Persistent Memory
- Populated with a combination of DRAM and Intel Optane DC persistent memory
- Intel Xeon - 6 DDR channels – Two DIMM slots each
- Each channel having one Intel Optane PMEM should have one DRAM also

Storage Node

- Optane persistent memory - Two modes of operation
- Volatile mode Persistent mode
- Persistence enabled in Persistent mode
- Volatile Mode is used by unmodified applications constrained by main memory size and only require volatile memory
- Persistent mode provides non-volatile access to persistent memory that helps reduce I/O bottlenecks

Storage Node

- Persistent mode - applications and operating systems should be aware of the two types of memory in the platform
- Persistent mode - Requires an operating system or virtualisation environment enabled with a persistent memory-aware file system
- Each OS vendor provides native tools for persistent memory management.

Storage Node

- Distributed Asynchronous Object Storage (DAOS) - Open source software-defined scale-out object store
- DAOS takes advantage of next-generation NVM technology, like Storage Class Memory (Intel Optane) and NVM Express™ (NVMe™)
- DAOS + Intel Optane persistent memory +NVMe SSD's - Optimum performance
- Fast I/O and data persistence of Optane persistent memory - Alleviate bottlenecks and drive storage performance in distributed environments

Storage Node

DAOS sidestep locking contention seen with the parallel filesystems used at HPC sites

Data written in persistent media (SSD's / HDD's) as blocks

When data smaller than size of the block - share a block

Two compute nodes writing to same block – one locked out

Parallel actions get serialized

Persistent Memory - byte addressable

Bottleneck with blocks avoided -can do byte-granular I/O - no longer have different I/Os done serially

Storage Node

Storage nodes in HPC cluster outfitted with Optane persistent memory modules and NVMe SSD's

DAOS - Metadata and small read/writes into the persistent memory - deals with the locking issue

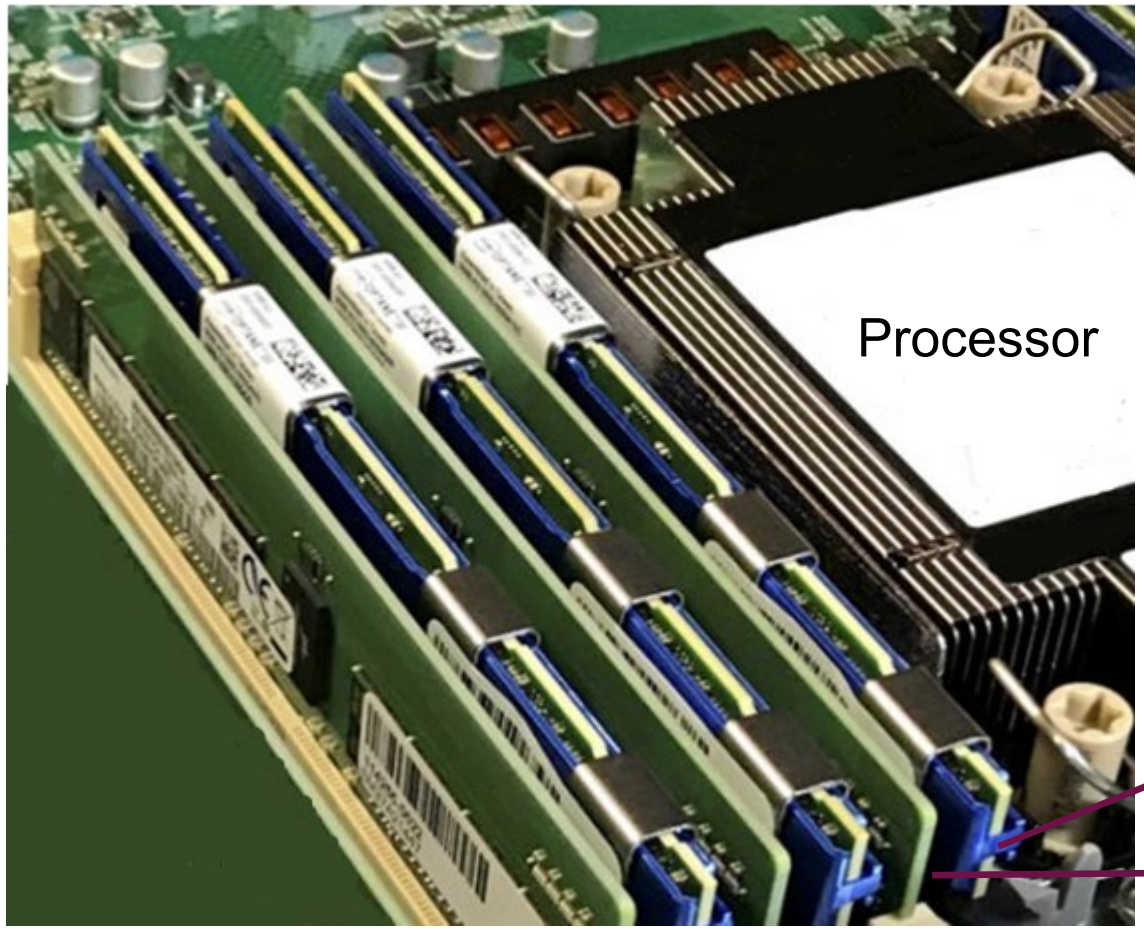
Block-friendly larger I/O operations - To NVM-Express SSD

Data access - orders of magnitude faster than in existing storage systems

Achieve high bandwidth and IOPS

Minimum 6% ratio of SCM (Storage Class Memory – Persistent Memory) to SSD – DAOS to store its internal metadata in SCM

Storage Node



Persistent Memory

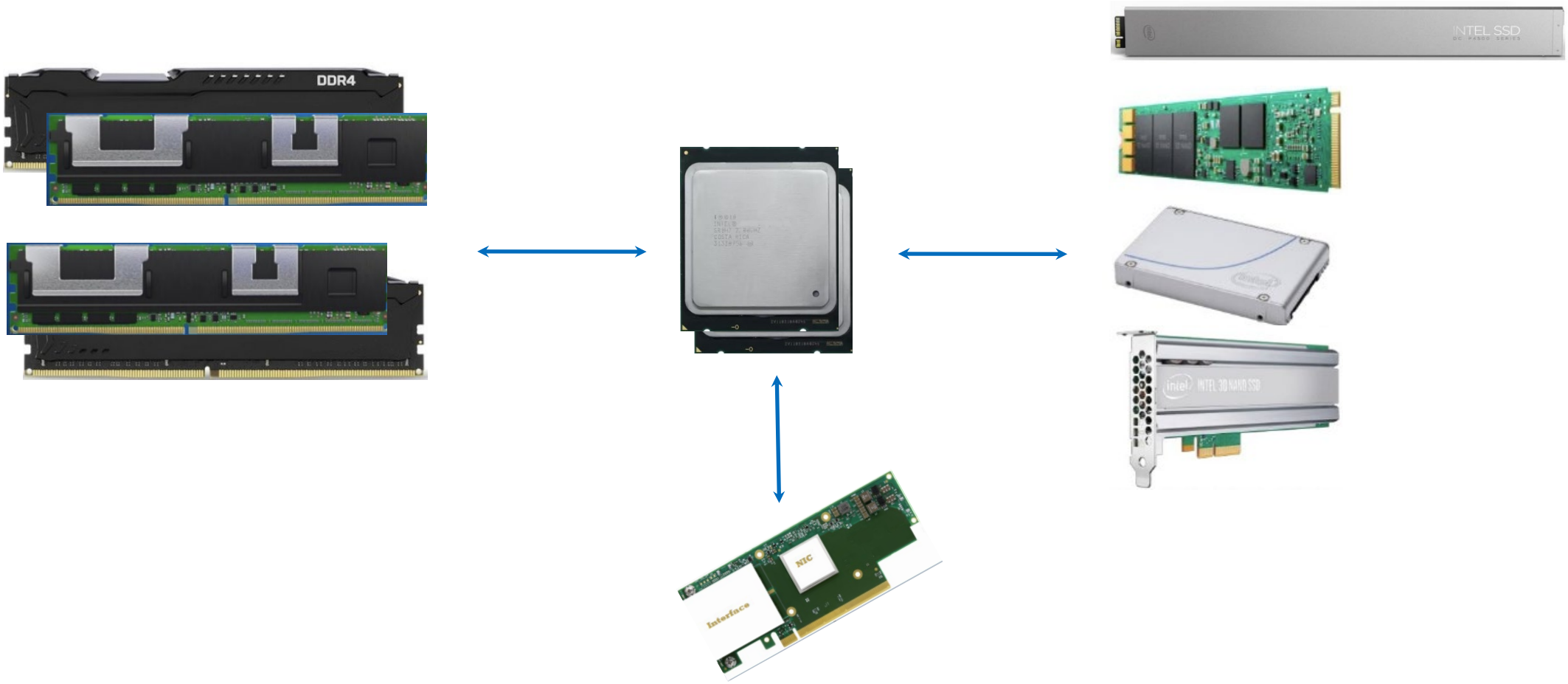
+

DRAM Memory

Storage Node

- NVMe SSDs - Allows storage I/Os to saturate the PCIe bus with a bigger data pipeline.
- NVMe SSDs - E1.S, E1.L, AIC, U.2, M.2 – Form factor
- Persistent memory based storage node can deliver storage I/O that is faster (from milliseconds (ms) to tens of microseconds (µs)) compared to traditional storage.

Storage Node



Storage Node

- Operating System Support for optane persistent memory based storage node
  - CentOS 7.6 or later
  - RHEL 7.6 or later
  - SLES 12 SP4 or later
  - SLES 15 or later
  - Ubuntu 18.04 LTS
  - Ubuntu 18.10 or later
  - VMWare ESXi 6.7 U1 or later
  - Windows 10 Pro for Workstation Version 1809 or later
  - Windows Server 2019 LTSC
  - Windows Server 2019 or later

Please take a moment to rate this session.

Your feedback is important to us.

