Planning For Persistent Memory In The Data Center

Sarah Jelinek/Intel Corporation

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Abstract

Preparing for persistent memory

- Persistent memory is a disruptive technology that will require a lot of thought for integration into the data center.
  - Persistent memory(Non-volatile memory) is a new class of storage. It provides unique attributes that should be considered when deploying this storage class in the data center. We will examine and explore potential ways to deploy persistent memory: Public cloud, private cloud, hybrid cloud and enterprise data centers.

Learning objectives

- What is persistent memory?
- How could it fit into the cloud and enterprise data center?
- What are the advantages of integrating persistent memory into the data center?
- Impact on existing software?

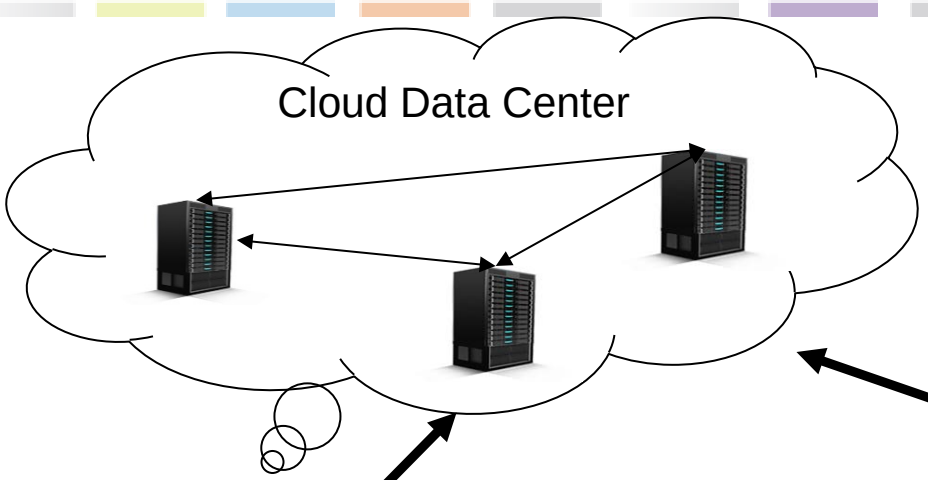
What is Persistent Memory?

- Storage Class Memory
- Byte addressable memory
  - Persistent
    - Unlike DRAM, data on persistent memory can be accessed using memory instructions or memory APIs even after the end of the process that created or last modified them
    - Do not lose state on power loss
- Must be reasonable to stall a CPU and waiting for a load to finish
  - Not NAND based non-volatile storage
- Similarities to DRAM:
  - Same density
  - Near DRAM speed
- Differences to DRAM:
  - Lower cost, large capacity
- Can DMA to it

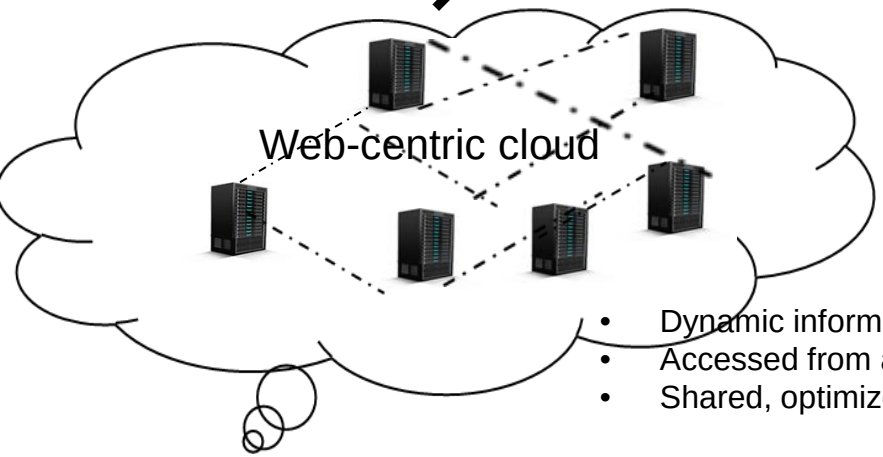
Things To Consider

- Why would you want to put persistent memory in the data center?
  - In-memory computing is one of today's hottest data center trends
  - DRAM is expensive and consumes more power.
  - Persistent memory will come in very large capacities.

Cloud Service Provider



- Request driven
- Virtualized infrastructure for multiple workloads- massively scalable
- Optimized for security, transactions, data integrity



- Dynamic information delivery
- Accessed from anywhere
- Shared, optimized for traffic, scalable

Data Center

- Mission critical apps
- Controlled access
- Optimized for security and integrity

Breaking It Down

- Pubic and private cloud
  - Looking at cloud data center, not from HW perspective but from the viewpoint of applications we are seeing a much different landscape than in the past.
  - The objective is not to improve the speed of today's disk centric architectures. The *GOAL* is to keep entire data sets memory-resident.
    - Memcached, Redis, Spark
  - How is this achieved today?
    - Many data centers are putting large amounts of flash behind DDR DRAM DIMMs.
    - On the server a thin layer of DRAM fronts flash
    - Deeper in the data center, high-capacity, high-reliability SSDs of genuinely massive capacity—tens or hundreds of terabytes (TBs)—support the server DIMMs through RDMA transactions that bypass the operating system and hypervisor altogether to minimize latency. Some architects call this disaggregation—pulling the mass storage out of deep centralized pools and spreading it across the data center, as near as possible to the servers.

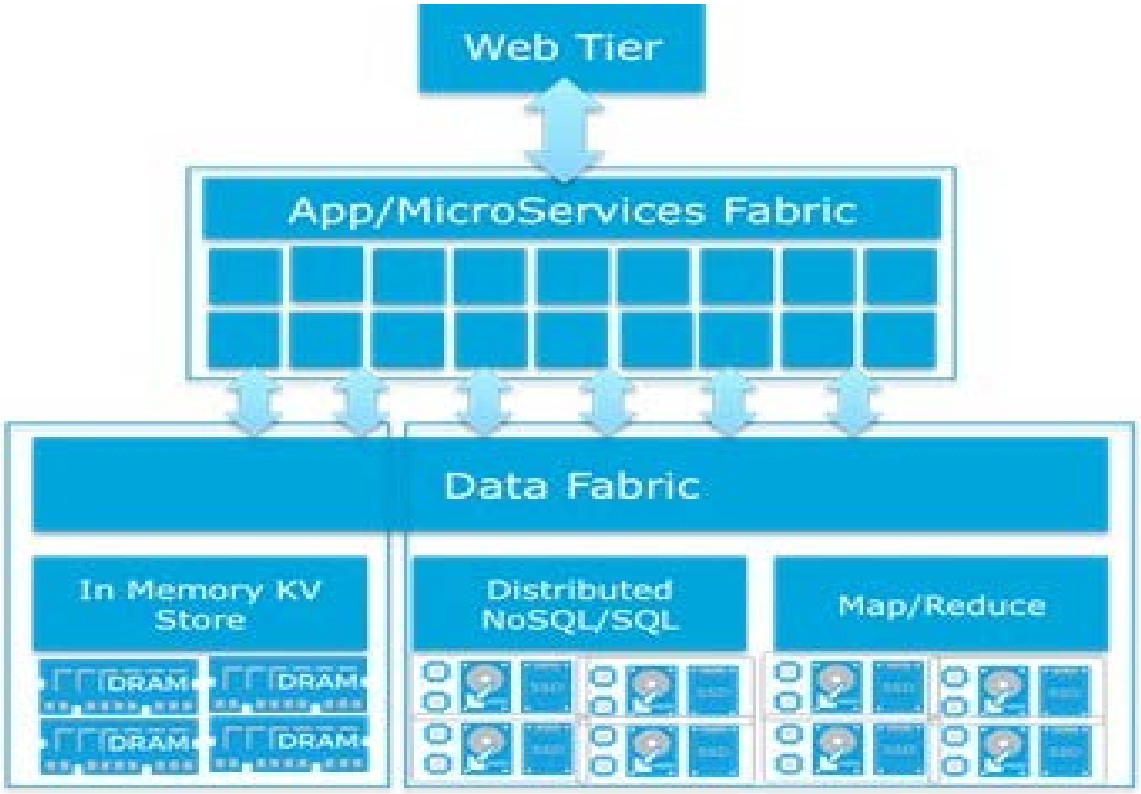
Breaking It Down

- Pubic and private cloud
  - However, what if persistent memory were available?
    - New memory DIMMs that support persistent memory together with a controller chip that will implement extensions to the DDR4 memory will form exactly the layer of dense persistent memory software developers have been designing for.
    - The transition to software like Spark that keep entire data sets memory-resident create a natural opportunity to do some operations in the DIMMs themselves, rather than in CPUs.
    - Much of the computational aspects of the applications will be done inside a persistent cache.
  - Bringing in persistent memory to the cloud
    - Provides lower cost, guaranteed low latency, large capacity and the potential to create large masses of memory for applications such as Spark, Redis and memcached .
      - Remember the goal of these new class of applications is to operate on the full data-sets in memory!
    - Does persistent memory support RDMA? Yes.
    - Can it be used just like DRAM? Yes.

Breaking It Down

- Public and Private Cloud
  - Bringing in persistent memory
    - Will the existing applications need to be modified to use persistent memory?
      - No, but if an application wants to make use of the persistent feature of this new memory tier then *maybe* it would have to be modified to understand persistent memory.
        - Persistence == Correct and consistent

Enterprise Data Applications

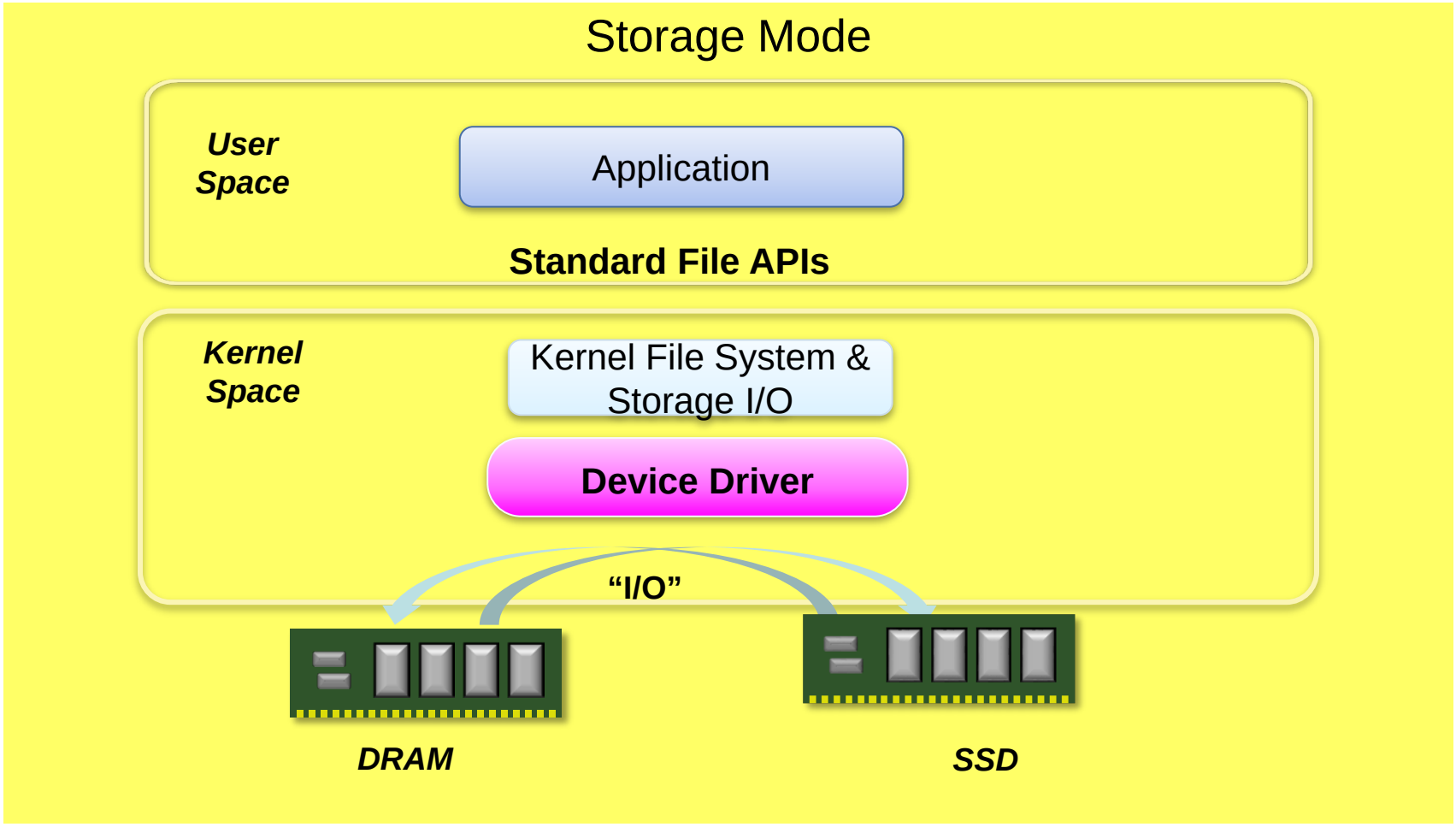


Breaking It Down

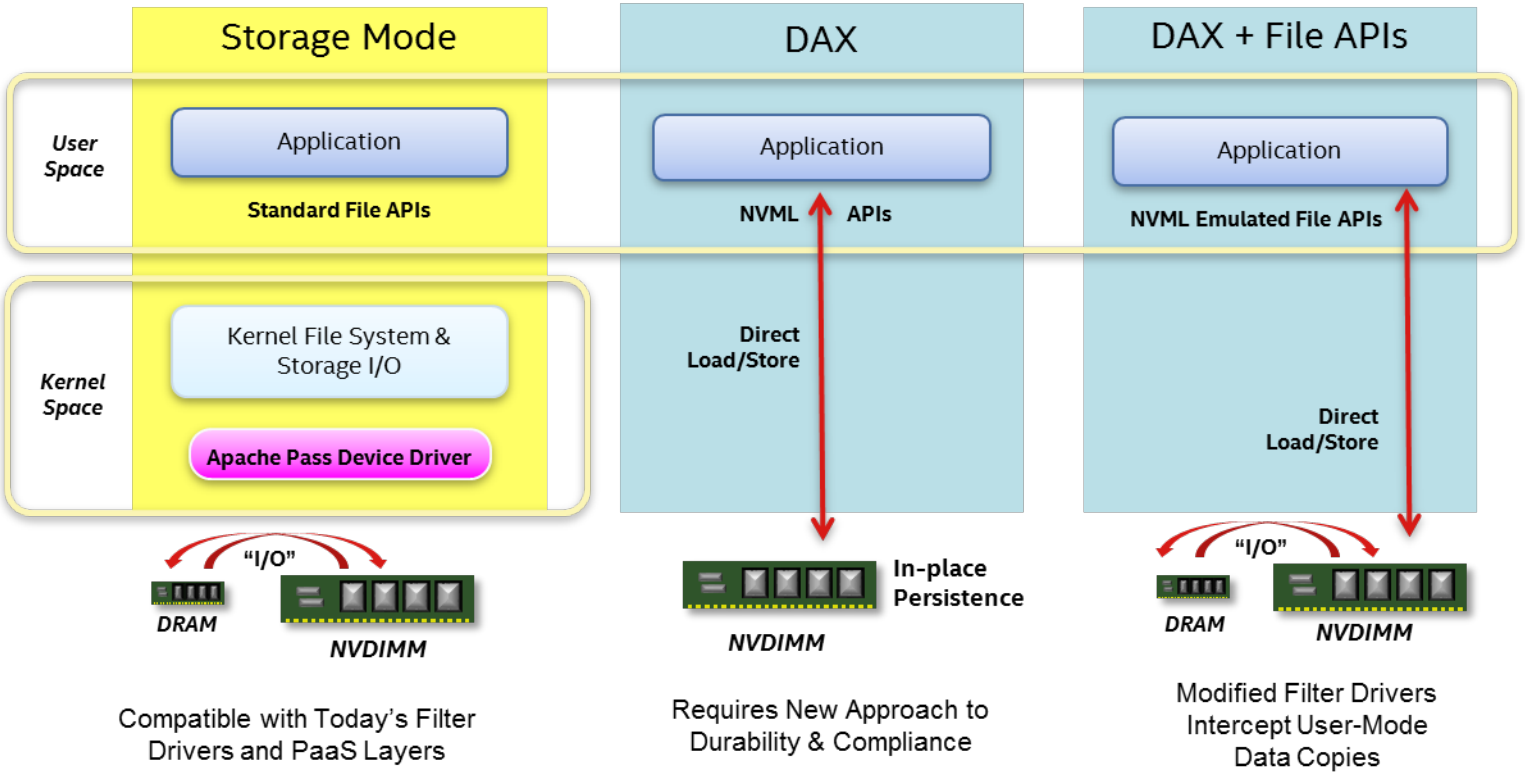
Enterprise Data Centers

- Organizations are seeking both faster and better alternatives to processing an ever-increasing amount of data at high speed.
- Software-Defined Memory provides growing data centers the ability to run both-in memory and traditional enterprise applications on a single infrastructure.
  - SDM is the convergence of Storage and Memory, which for decades were two separate computing domains.
  - For persistent memory it would be PMEM + SSD
  - With the emergence of PMEM SDM becomes increasingly a prominent part of the data center.
- As data loads grow, and applications become increasingly latency-sensitive.
- Modern data center applications can benefit from more memory.
- Current trend is hyper-convergence of servers and storage into a single block.

Enterprise storage I/O today



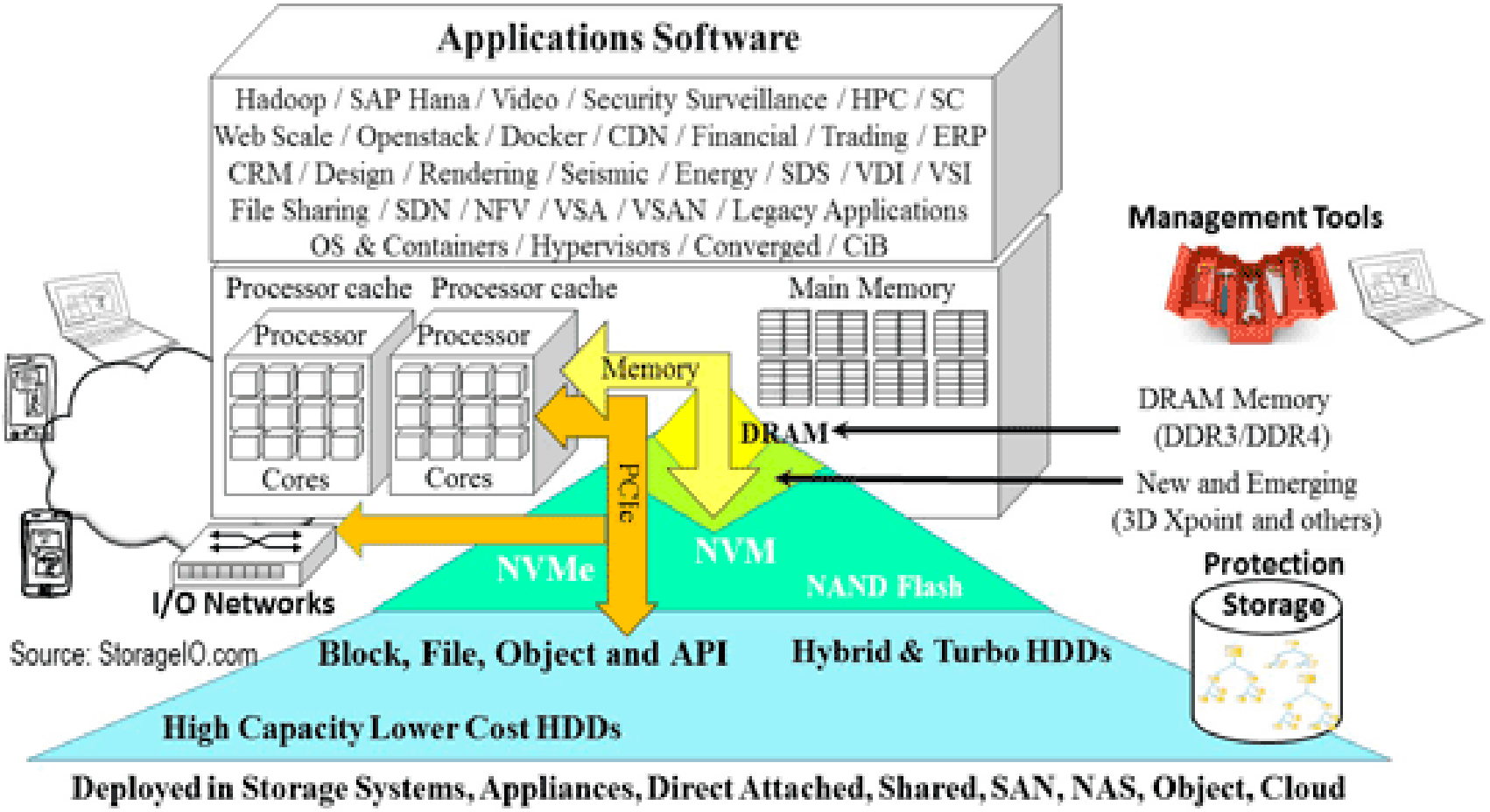
What if we integrate persistent memory?



Breaking It Down

- Persistent Memory could be integrated today in standard storage mode.
  - No application changes required
  - However, it's an inefficient use of persistent memory
- Persistent Memory could be integrated using the new Linux Driver + the NVML open source libraries
  - Application changes required. Best performance using properties of Persistent Memory and accessing it via direct load and store.
- Or… Persistent Memory could be integrated using the new Linux Driver + NVML open source libraries + a File API layer(In development)
  - No application changes required(for those that use standard file system interfaces).
  - Utilizes NVML, which provides direct load store access.

Server Storage I/O memory HW and SW hierarchy



Attribution & Feedback

The SNIA Education Committee thanks the following Individuals for their contributions to this Tutorial.

Authorship History

Name/Date of Original Author here:
Sarah Jelinek 2015

Additional Contributors

Andy Rudoff
Paul Von Behren
Chet Douglas

Please send any questions or comments regarding this SNIA Tutorial to ***tracktutorials@snia.org***