How Persistent Memory can Benefit Artificial Intelligence and Machine Learning Applications

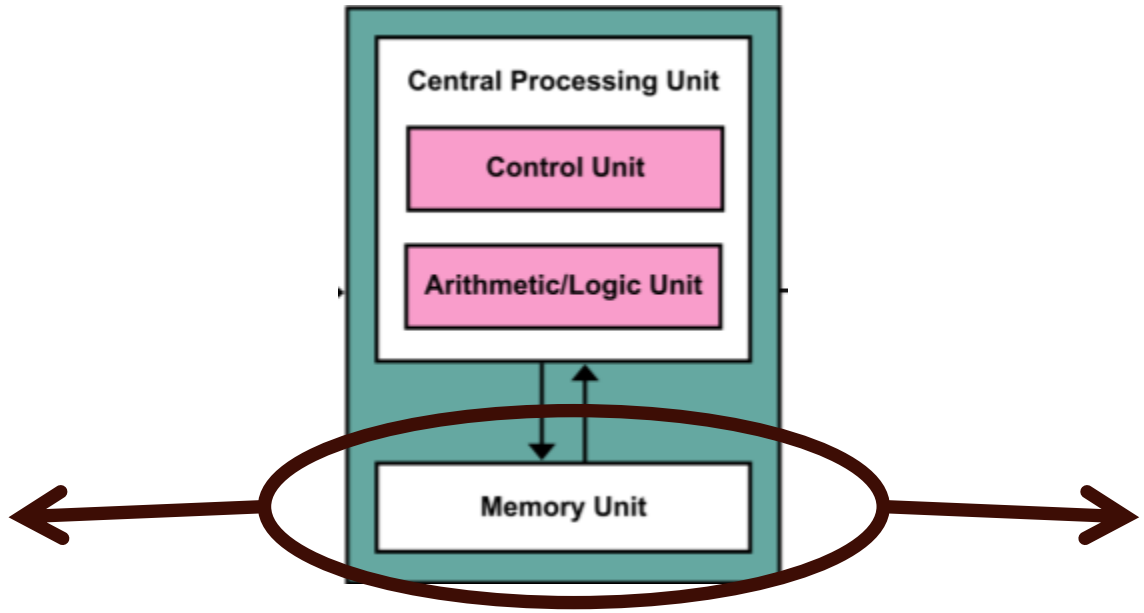
Arthur Sainio, Jim Fister
Storage Networking Industry Association (SNIA)

Introduction

- Explosion of data creation for use by Artificial Intelligence (AI) and Machine Learning (ML) applications
- But traditional systems are not designed to address the challenge of accessing large and small data sets
- The key hurdle is reducing the overall time to discovery and insight based on data intensive ETL (Extract, Transform, Load); and checkpoint workloads
- Artificial intelligence and machine learning applications are starting to take advantage of persistent memory to eliminate bottlenecks and accelerate performance

Dominant Memory Technologies

REASONING



DRAM

Working memory

NAND

Storage

The future of **DRAM** is …… DRAM

The future of **NAND** is …… NAND

There are shortcomings for both

Persistent Memory has emerged to fill the gaps

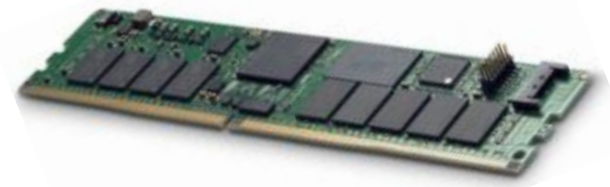
What is Persistent Memory?

Persistent Memory is:

- Byte-addressable and accessed by memory semantics (Load/Store)
- Fast (low-latency, faster than block-accessed media)
- Persistent (non-volatile)

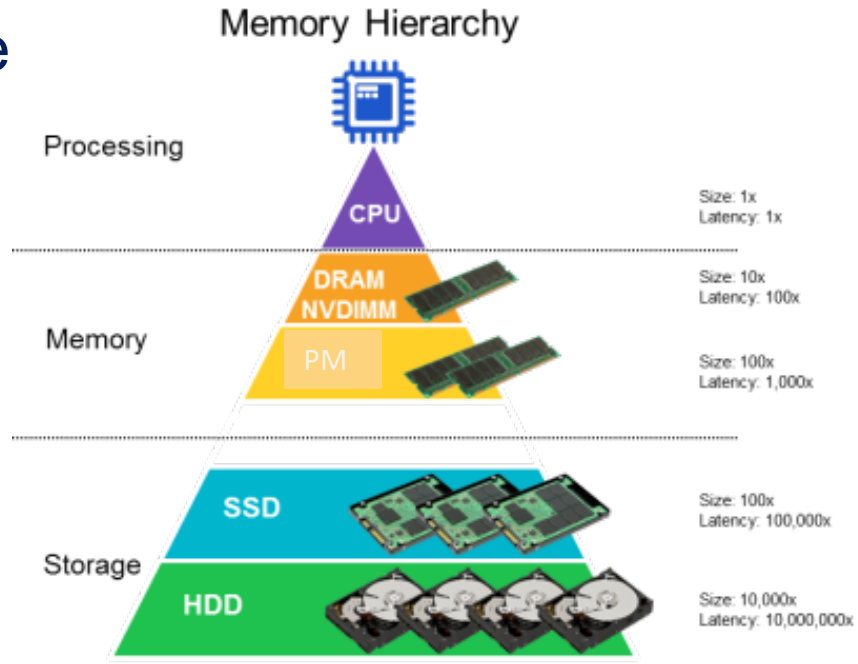
Persistent Memory includes:

- Persistent Memory devices: PM Media or PM Devices (aka Emerging Non-Volatile Memory)
- Persistent Memory modules/cards: NVDIMM-N, NVDIMM-P, byte-addressable memory cards
- Persistent Memory: used like storage in architecture of systems and software, can be main memory



Why Persistent Memory?

- For system acceleration!
- For very low latency tiering, caching, write buffering metadata storage, and in-memory database (i.e.,NVDIMMs)
- Persistent Memory as a fast access tier in your storage application
- High capacity PM makes it possible to run multi-TB databases completely in memory
- Speed of non-volatile memory changes dynamics of storage industry
- Instant, byte-level persistence enables new database algorithms for storing machine learning data sets



Persistent Memory Use Cases

Enterprise & Software Defined Storage

Tiering, caching, write buffering, meta data storage

Traditional & In-Memory Database

Log acceleration

Journaling, recovery time, tables

High-Performance Computing

Check point acceleration and/or elimination

High-Performance Data Analytics

AI / ML Workflows

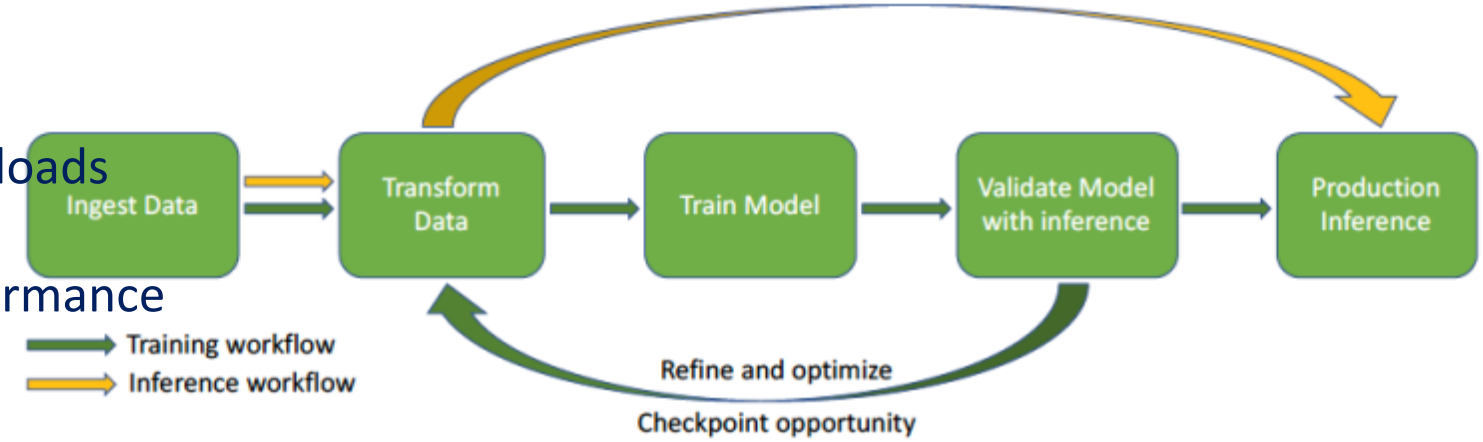
Checkpointing

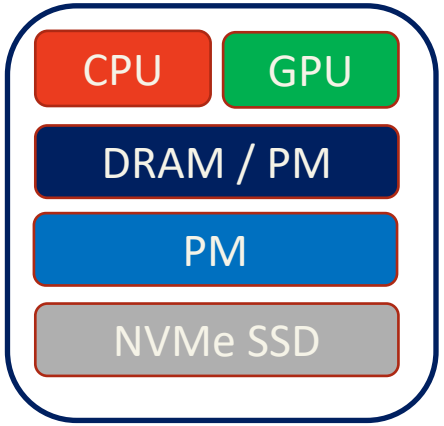
Spark Acceleration

Data Intensive Workflows

Why Persistent Memory in AI / ML?

- Challenge: Reducing overall time to discovery and insight based on Data Intensive ETL and Checkpoint Workloads
- Demanding I/O and computational performance for GPU accelerated ETL
- Varying I/O and computational performance is driven by bandwidth and latency
- Generate metadata databases using emerging Computational Storage PM solutions as an integrated AI inference engine

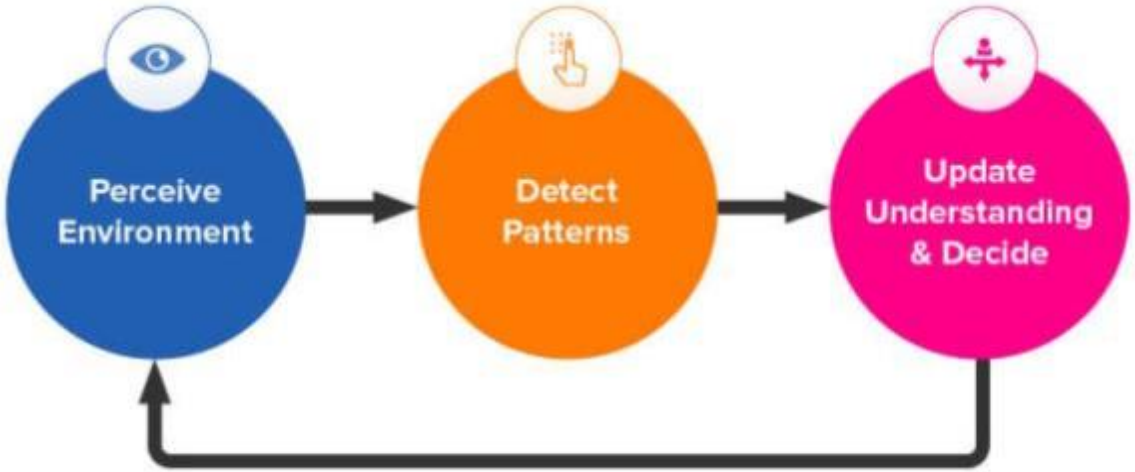




AI Training Challenges

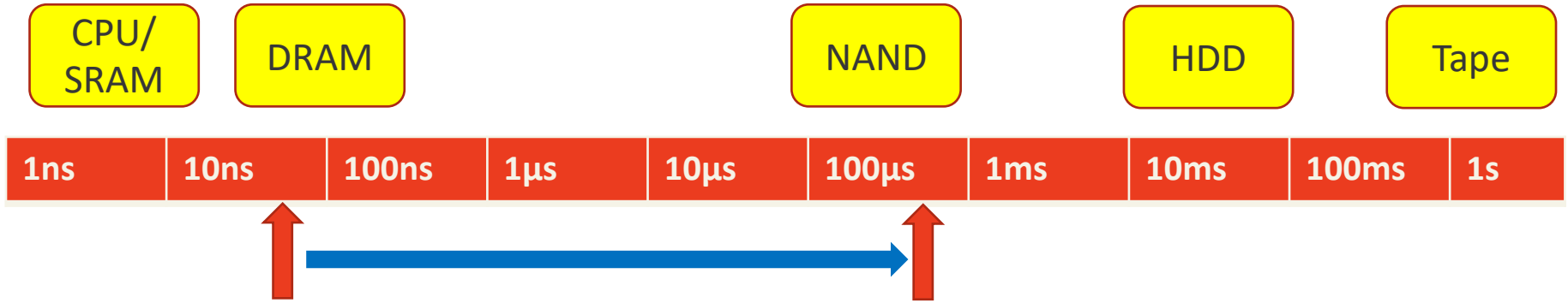
Challenge

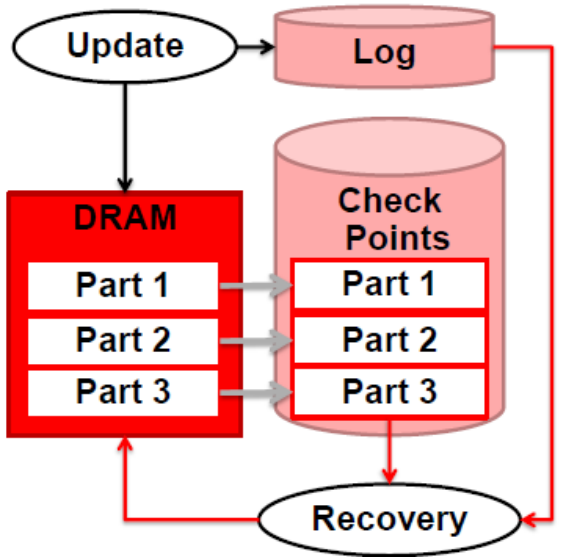
- Model training takes a long time to complete for datasets
- Data preprocessing and importing can take a long time
- Failure recovery is painful without frequent checkpointing
- Delays model deployment



Checkpointing Today

- Checkpointing - Taking a snapshot of the DBMS state
- By taking checkpoints periodically, DBMS can reduce the work to be done during restart in the event of a subsequent crash
- Checkpointing is done in storage (SSD, NAND)

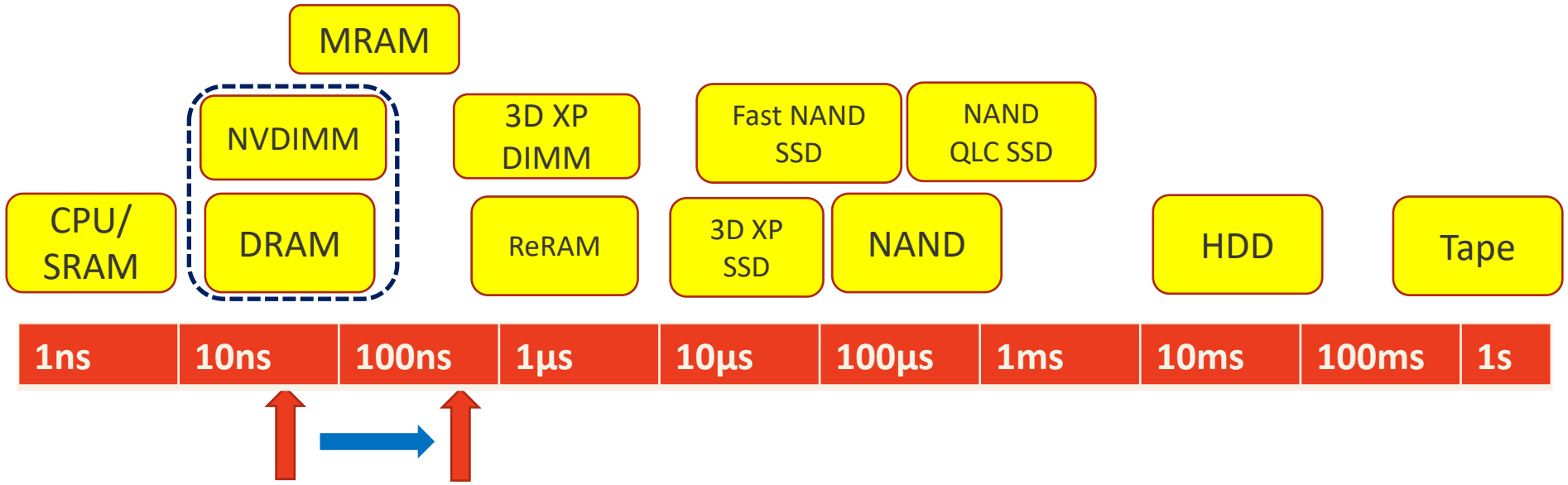


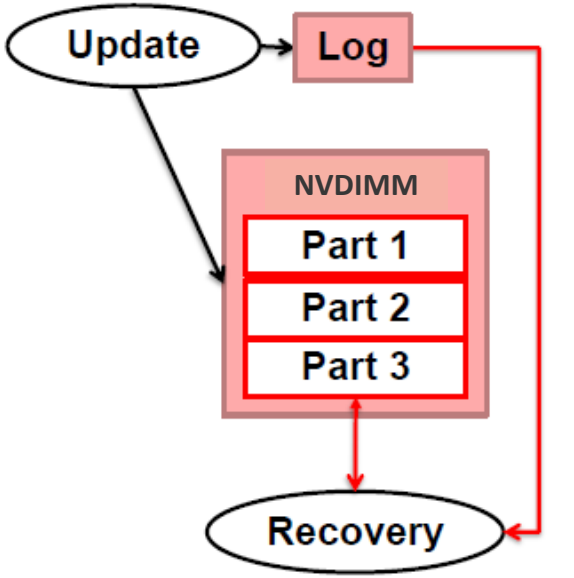


- But checkpointing takes time (I/O + NAND latency + points of failure)

Checkpointing with Persistent Memory

- Persistent Memory options have emerged to reduce and even eliminate checkpointing. Some in use now…..
- Checkpointing is an ideal use-case for NVDIMMs

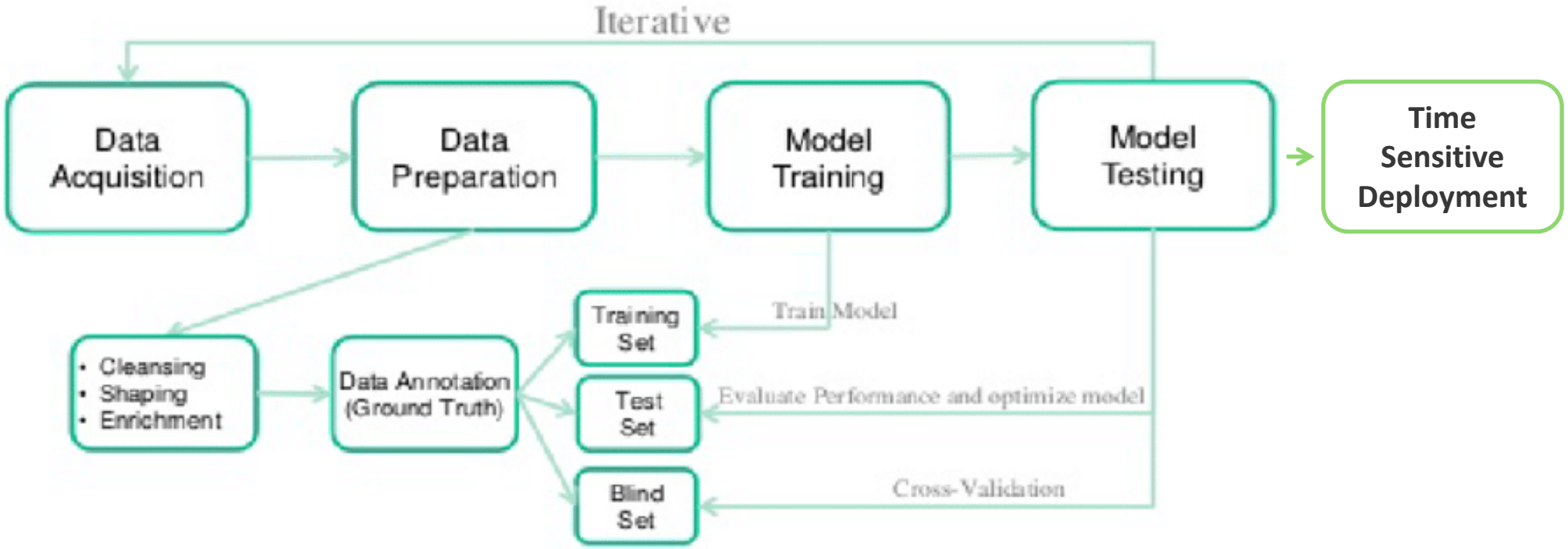




- NVDIMMs allow checkpointing to be done at DRAMs speeds (ns vs. µs)

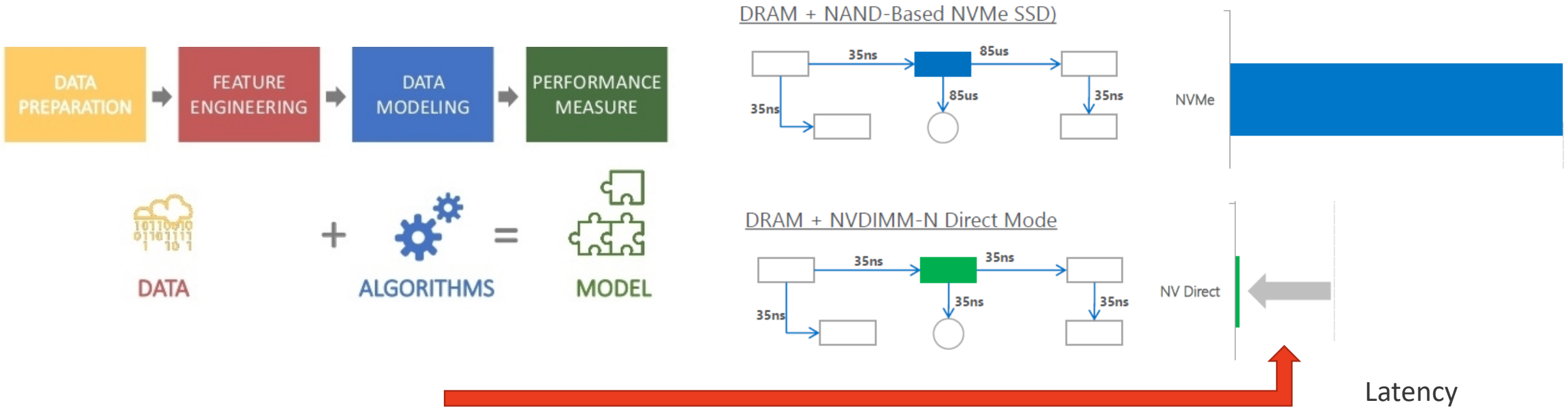
Machine Learning

- Data acquisition, preparation, model training, testing done in storage
- Data sets cannot risk being lost or else the model training and testing process needs to restart



Source; IBM

Machine Learning with Persistent Memory



- Dramatic acceleration of the ML process can be achieved by using fast Persistent Memory vs. writing to storage

Source; Performance chart from Micron

Machine Learning Dataset Size Examples

Range between 850KB to 2TB

Google-Landmarks Dataset
Google
20 hours | 107 MB | 8.2 | 5 Files (CSV)

Pizza Restaurants and the Pizza They Sell
Datafiniti
2 months | 850 KB | 7.6 | 2 Files (CSV)

Transit systems of world
citylines.co
4 months | 3 MB | 7.6 | 7 Files (CSV)

Bitcoin Blockchain
Google BigQuery
5 months | 871 GB | 6.8 | BigQuery

Chicago Taxi Trips
City of Chicago
a year | 32 GB | 7.1 | BigQuery

Google Patents Public Data
Google BigQuery
10 month | 2 TB | 7.1 | BigQuery

Credit Card Fraud Detection
Machine Learning Group - ULB
a year | 66 MB | 8.5 | 1 File (CSV)

Hand Gesture Recognition Database
GTI
a year | 1 GB | 8.8 | 1 File (other)

Shared Cars Locations
DoiT International
6 months | 78 MB | 8.2 | 1 File (CSV)

Intel Image Classification
Puneet Bansal
6 months | 344 MB | 8.8 | 3 Files (other)

Clinical Trials on Cancer
auriml
2 years | 42 MB | 6.5 | 1 File (CSV)

AMEX, NYSE, NASDAQ stock histories
Jiun Yen
3 months | 502 MB | 7.6 | 2 Files (CSV, other)

Machine Learning with Persistent Memory

for smaller data sets

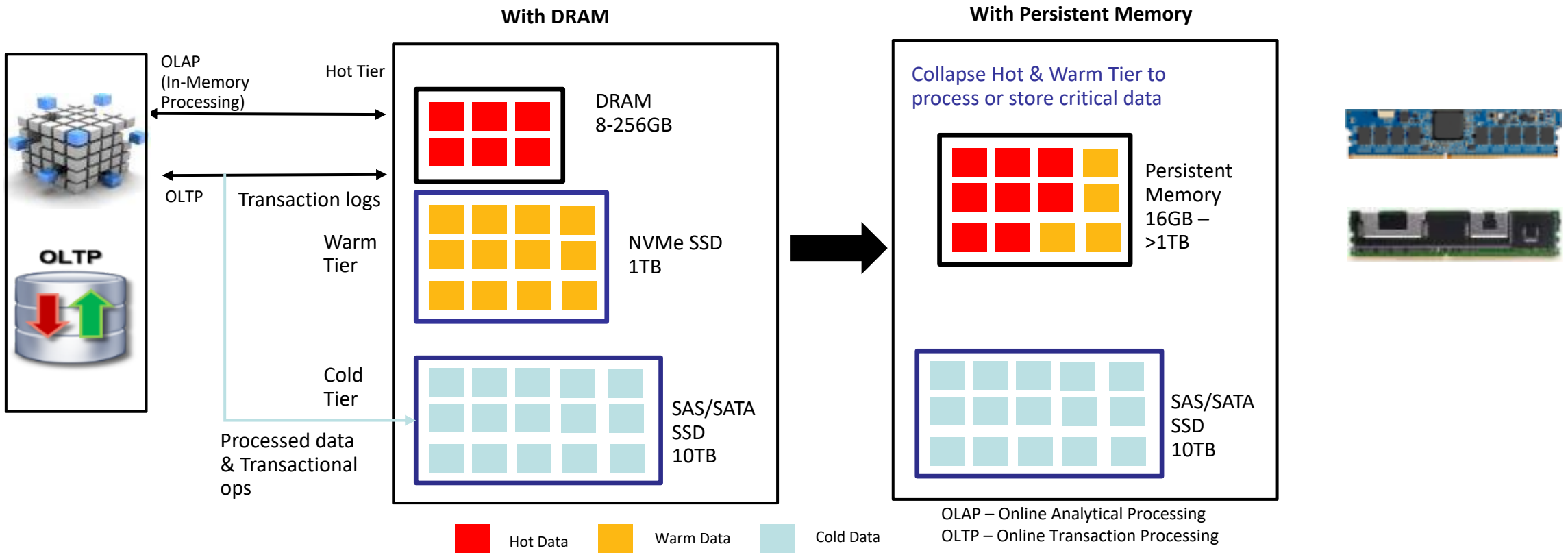
- GPU servers run algorithms which are integral for ML
- Adding NVDIMMs protect GPU servers from loosing ML data. Lost data would cause need to restart work.
- Multiple servers/nodes will be needed
- Industry standard servers can support twelve 16GB NVDIMMs (192GB per server/node)
- NVDIMMs add persistence capability to a rack
- Reduces read latency from 100’s µs to ~300ns

Machine Learning with Persistent Memory

for larger data sets

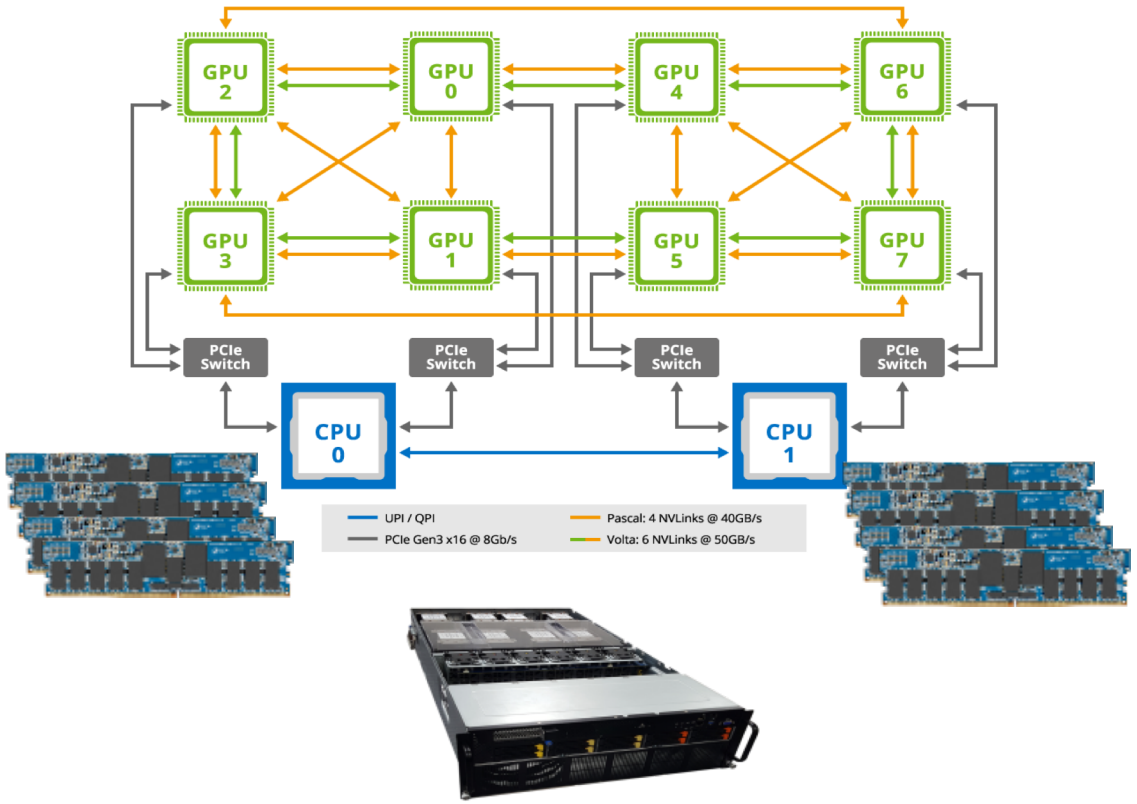
- Using larger byte-addressable Persistent Memory in AFAs
- Optane™ DC PM expands memory on the DDR bus
- Arch, software, hardware total effort
- Reduces read latency from 100’s µs to ~15µs
- DRAM and Optane operate as “near memory”
- Intel proprietary

Evolution of In-Memory Apps with Persistent Memory



Source: Modified from VMWare, FMS 2018

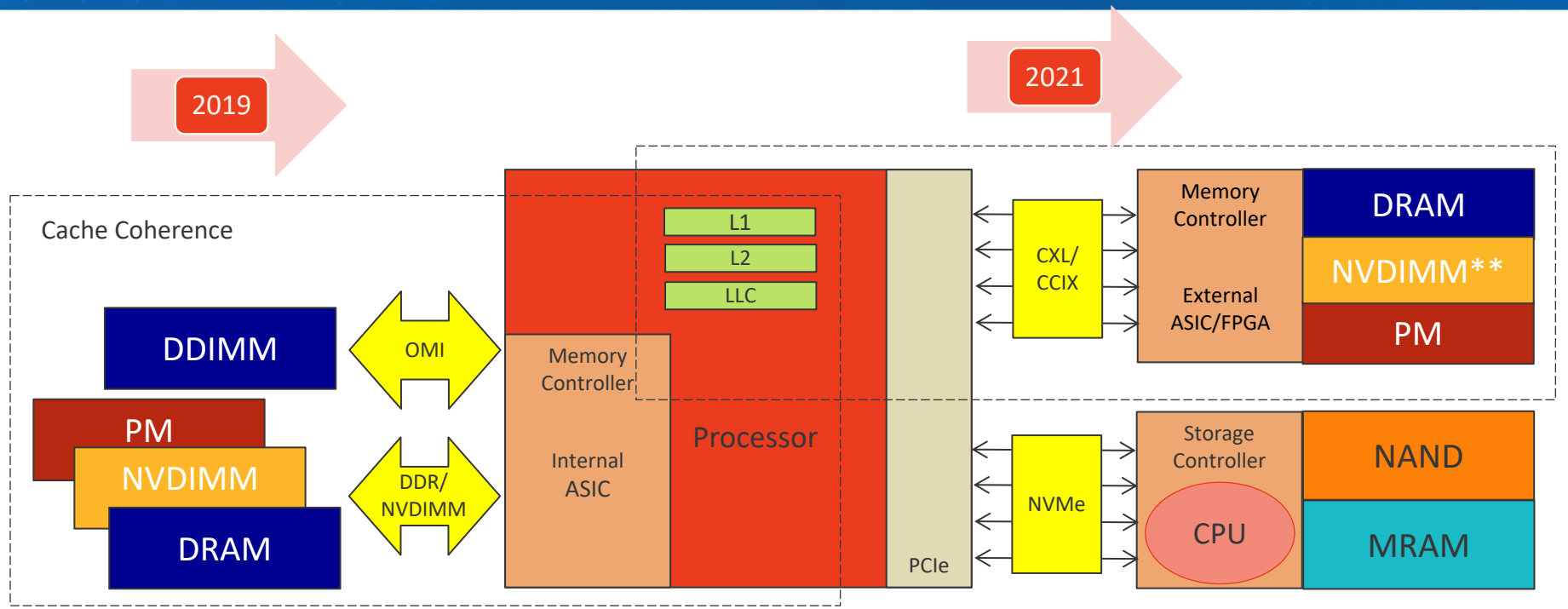
Persistent Memory Optimized Server for AI & ML



Get the data out onto the network as quickly as possible!

Example System – AI / ML Performance at Scale	
Processor/ Chipset	2x Intel® Xeon® Scalable Processor family, TDP up to 205W ***Support for Cascade Lake**
Data and Storage Layer	24x DIMM slots, 6 DPC, DDR4 - **DRAM** - **NVDIMM or** - **Persistent Memory (i.e., Optane)** 10 x 2.5" hot-swappable HDD/SSD bays - **4 x U.2 (Secure) NVMe devices only** - **6 x 2.5" (Secure) SATA/SAS devices**
PCIe Accelerated GPU and Networking	**8X NVIDIA V100 SXM2 w/ NVLINK** **4x 100G Low Latency High Speed Network** 2x 25GbE Ethernet
Workloads and Verticals	RAPIDS Open GPU Data Science TensorFlow + TensorRT DALI nvJPEG APACHE Spark

Memory Expansion Today and Tomorrow
including Persistent Memory

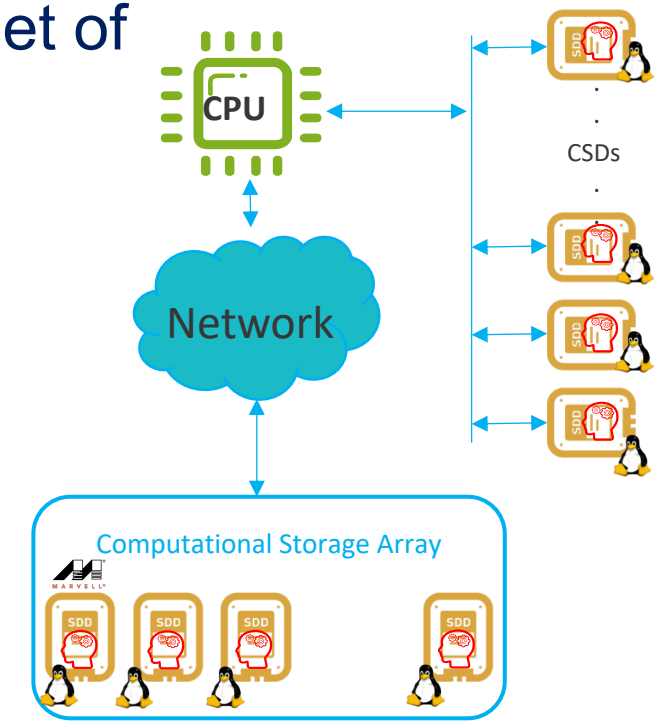


** Functionally equivalent to NVDIMM-N, not in DIMM form factor

- Today memory is direct-attached to the CPU
- New emerging interfaces will add high-speed differential CPU-attach options
- Systems will be aware of what type of memory or storage is available and how it is connected
- ***Lots of new types of memory, persistent memory and storage products are possible!***

Computational Storage and AI Inferencing

- Generating metadata database (e.g tags) over a large set of unstructured data locally with an integrated AI inference engine
- Operation may be:
  - Triggered by a host processor
  - Done offline as a background task (batches)
- Metadata database may be then used by upper layer big data Analytics software for further processing
- Can work both on direct attached storage or on remote over the network storage
- Examples: Video search, Ad insertion, Voice call analysis, Images, Text scan, chatbot, etc



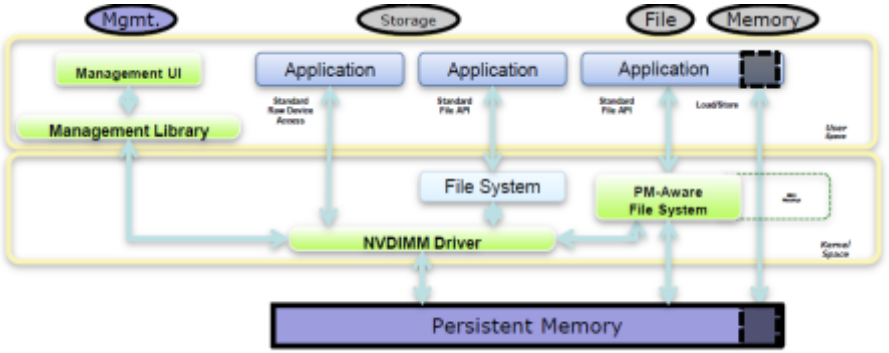
SNIA

Source; SNIA Computational Storage TWG

Persistent Memory Standards and Industry Enablement

- PM SW and Programming:
  - SNIA NVM Programming Model
  - pmem.io and PMDK for libraries and tools for implementing the NVM Programming Model

SNIA



- PM HW:
  - JEDEC Hybrid DIMMs and interface standards

JEDEC

- PM Fan-out interfaces / Fabric:
  - Low latency fabrics supporting PM directly in-system or remote

CXL Compute Express Link GEN Z CCIX SEAMLESS ACCELERATION OpenCAPI™

So NOW You Wanna' Program Persistent Memory...

Variety of open-source tools and libraries

- Persistent Memory Development Kit (PMDK)
- Direct programming models
- Multiple open-source file systems
- Similar Windows/Linux architecture models

Programming or Experience Opportunities

- Persistent Memory Hackathons
- NVDIMM Programming Challenge
- Persistent Memory Summit
- Persistent Programming In Real Life (PIRL)

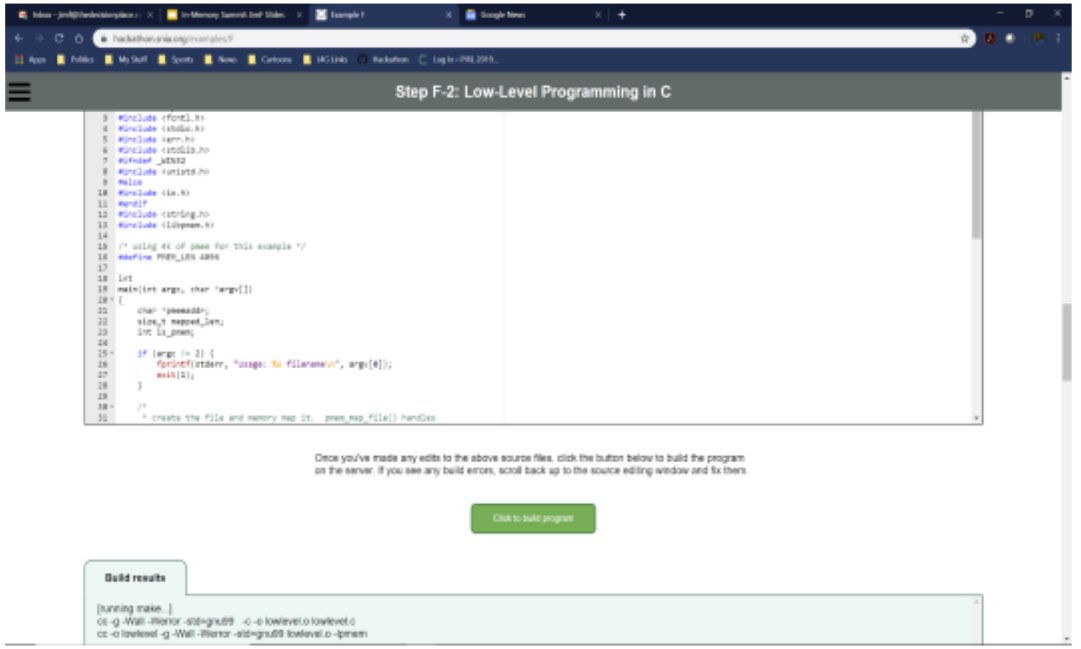
SNIA 2019 PERSISTENT MEMORY HACKATHON

pmhackathon@snia.org

SNIA Hackathon Program

Hackathon/Workshop

- Offering opportunities worldwide in 2020
- PM Summit, SDC Europe, ...
- Host your own hackathon



NVDIMM Programming Challenge

- Ongoing through at least Q1'20
- Online system configured with PM, it's an "online hackathon"
- Video tutorials coming in 2020
- Interesting tools/applications will get online and conference exposure

pmhackathon@snia.org

Additional meet-up and conference options throughout 2020

In-Memory Computing SUMMIT NORTH AMERICA 2019

Thank You!

SNIA®