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

The future of DRAM is ...... DRAM

NAND Storage

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
• Dramatic acceleration of the ML process can be achieved by using fast Persistent Memory vs. writing to storage
<table>
<thead>
<tr>
<th>Dataset Description</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google-Landmarks Dataset</td>
<td>107 MB</td>
</tr>
<tr>
<td>Pizza Restaurants and the Pizza They Sell</td>
<td>850 KB</td>
</tr>
<tr>
<td>Transit systems of world</td>
<td>3 MB</td>
</tr>
<tr>
<td>Bitcoin Blockchain</td>
<td>871 GB</td>
</tr>
<tr>
<td>Chicago Taxi Trips</td>
<td>32 GB</td>
</tr>
<tr>
<td>Google Patents Public Data</td>
<td>2 TB</td>
</tr>
<tr>
<td>Credit Card Fraud Detection</td>
<td>65 MB</td>
</tr>
<tr>
<td>Hand Gesture Recognition Database</td>
<td>1 GB</td>
</tr>
<tr>
<td>Shared Cars Locations</td>
<td>78 MB</td>
</tr>
<tr>
<td>Intel Image Classification</td>
<td>344 MB</td>
</tr>
<tr>
<td>Clinical Trials on Cancer</td>
<td>42 MB</td>
</tr>
<tr>
<td>AMEX, NYSE, NASDAQ stock histories</td>
<td>502 MB</td>
</tr>
</tbody>
</table>
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

With DRAM:
- OLAP (In-Memory Processing)
- OLTP
- Transaction logs
- Warm Tier
- Cold Tier
- DRAM 8-256GB
- NVMe SSD 1TB
- SAS/SATA SSD 10TB
- Hot Data
- Warm Data
- Cold Data

With Persistent Memory:
- Collapse Hot & Warm Tier to process or store critical data
- Persistent Memory 16GB – >1TB
- SAS/SATA SSD 10TB
- OLAP – Online Analytical Processing
- OLTP – Online Transaction Processing

Source: Modified from VMWare, FMS 2018
Persistent Memory Optimized Server for AI & ML

Example System – AI / ML Performance at Scale

<table>
<thead>
<tr>
<th>Processor/Chipset</th>
<th>2x Intel® Xeon® Scalable Processor family, TDP up to 205W *Support for Cascade Lake</th>
</tr>
</thead>
</table>
| 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

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

- **PM HW:**
  - JEDEC Hybrid DIMMs and interface standards

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

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

Additional meet-up and conference options throughout 2020
Thank You!