Persistent Memory in ML

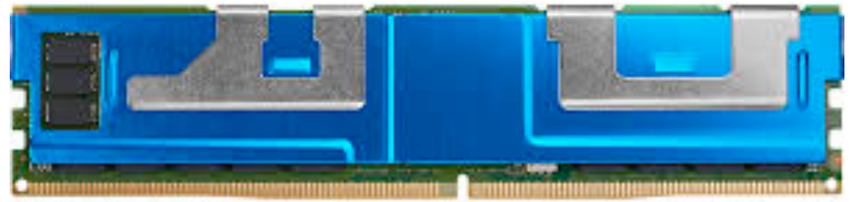
Mijin An
meeeeejin@gmail.com

Introduction

- Explosion of data creation for use by AI and ML applications
- But traditional systems are not designed to address the challenge of accessing large and small data sets
- AI and ML applications are starting to take advantage of **persistent memory** to eliminate bottlenecks and accelerate performance

What is Persistent Memory?

- **Byte-addressable** and accessed by memory semantics (Load/Store)
- **Low latency** (faster than block-accessed media)
- **Persistent** (non-volatile)
- e.g., NVDIMM, Intel Optane, …



Memory-Storage Hierarchy

Volatile Memory
Load/Store Instructions
Cache Line Granularity

$56 ~ $80 (~N)
40~140ns

NVDIMM

intel OPTANE DC PERSISTENT MEMORY

Non-Volatile Storage
Load/Store Instructions
Cache Line Granularity

$6.58~$10.42
100~340ns

Non-Volatile Storage
I/O Commands
Block Granularity

Cost ($/GB)

CPU Registers — ~ 0.1ns

CPU Caches (L1, L2, L3, L4) — 1 ~ 10ns

DDR DRAM — 80 ~ 100ns

Persistent Memory — < 1us

Solid State Drive (SSD) — 10 ~ 100us

Hard Disk Drive (HDD) — ~ 10ms

Tape — ~ 100ms

Latency

Capacity

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

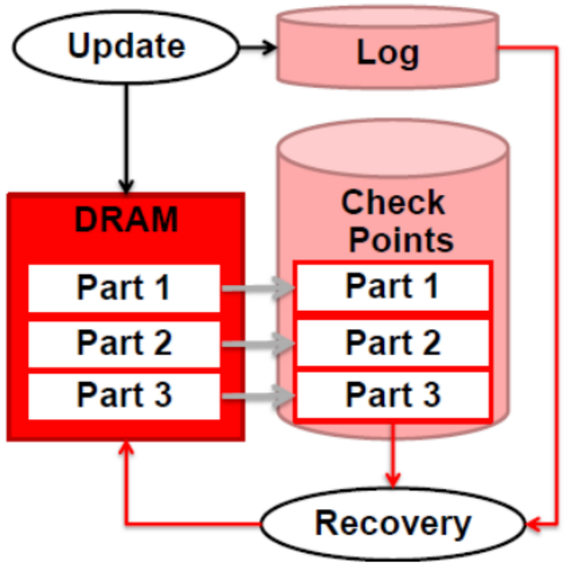
Data Intensive Workflows

Why Persistent Memory in 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

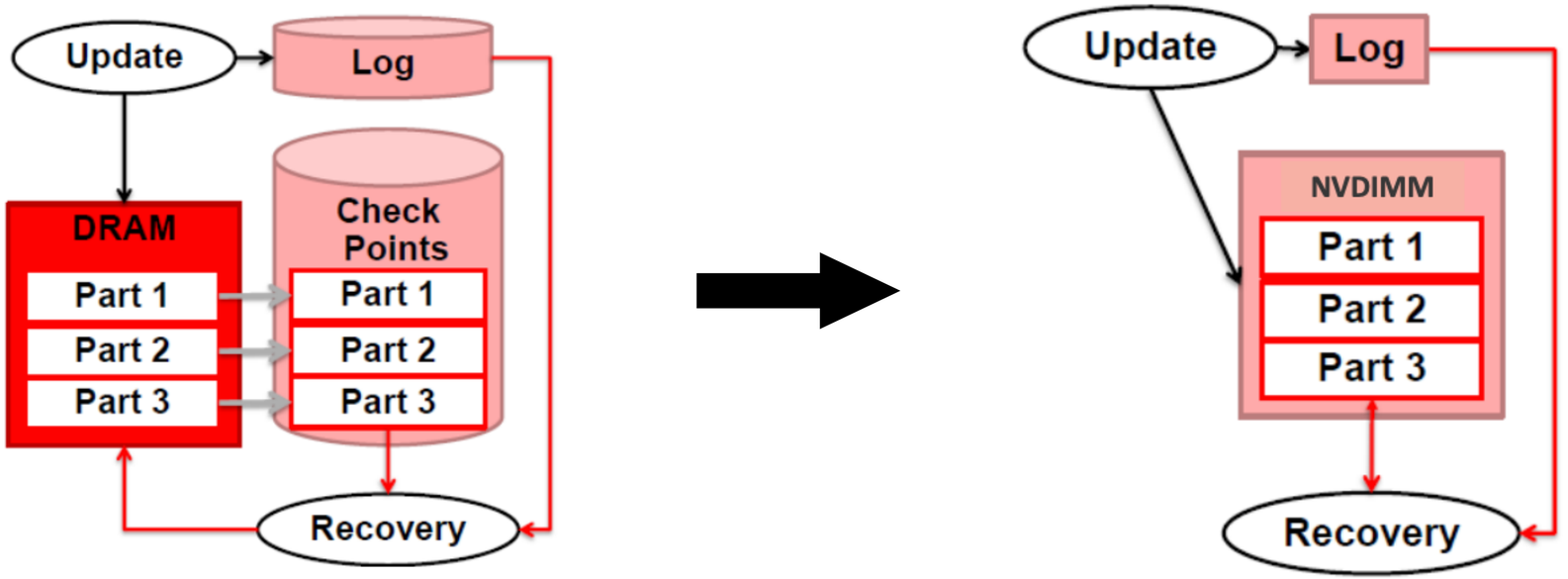
Checkpointing Today

- 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 + fsync + NAND latency)



Checkpointing with Persistent Memory

- Checkpointing is an ideal use-case for NVDIMMs
- NVDIMMs allow checkpointing to be done at DRAM's speeds (**ns** vs. µs)



ML with Persistent Memory

- It is essential to make fast ETL processes, where move vast amount of information from data lakes to the faster storage and then into the GPU complex

- Dramatic acceleration of the ML process can be achieved by using fast persistent memory (vs. writing to NAND storage)

Reference

[1] Pete Warden and Daniel Situnayake, “TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers”, O'Reilly Media, 1st Edition (December 31, 2019)

[2] “PERSISTENT MEMORY”, Flash Memory Summit 2020, https://www.flashmemorysummit.com/opt_persistent_memory.html

[3] “Intel® Optane™ Technology: Memory or Storage?”, Intel, https://www.intel.com/content/dam/www/public/us/en/documents/technology-briefs/what-is-optane-technology-brief.pdf

[4] Arthur Sainio, “NVDIMM – CHANGES ARE HERE SO WHAT'S NEXT?”, In-Memory Computing Summits 2016, https://www.snia.org/sites/default/files/SSSI/NVDIMM%20-%20Changes%20are%20Here%20So%20What's%20Next%20-%20final.pdf

[5] “Faster Access to More Data”, Intel, https://www.intel.com/content/www/us/en/architecture-and-technology/optane-technology/faster-access-to-more-data-article-brief.html

[6] Raghu Kulkarni, “Persistent Memory and NVDIMMs”, Flash Memory Summit 2018, https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2018/20180806_PreConC_Kulkarni.pdf

[7] Arthur Sainio, Jim Fister, “How Persistent Memory can Benefit Artificial Intelligence and Machine Learning Applications”, Storage Networking Industry Association (SNIA), In-Memory Computing Summit 2019, https://www.snia.org/sites/default/files/SSSI/Benefits%20of%20Persistent%20Memory%20for%20AI%20and%20ML%20A.%20Sainio%20-%20J.%20Fister%20%20final.pdf

[8] Gary Hilson, “Micron Puts SSD into AI Mix”, EE Times, https://www.eetimes.com/micron-puts-ssd-into-ai-mix