INTRODUCTION TO PERSISTENT MEMORY

Presenter: Wojciech Golab
wgolab@uwaterloo.ca

PODC 2019

Toronto

August 2nd, 2019

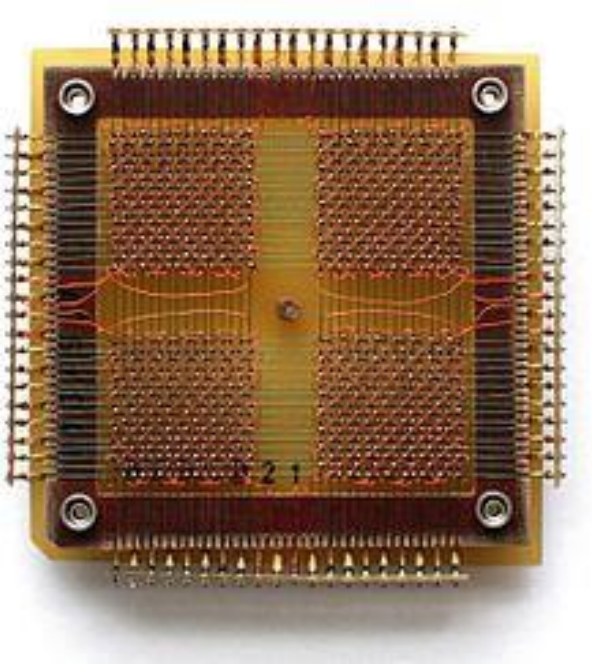
UNIVERSITY OF WATERLOO

OUTLINE

- **History**
- **Memory Hierarchy**
- **Cost and Performance**
- **Hardware and Software Support**
- **Research Directions**

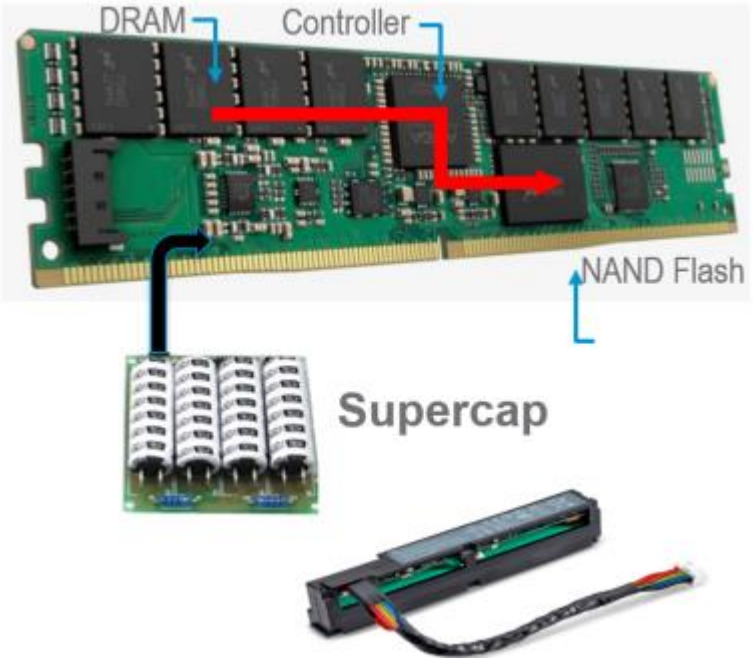
HISTORY

MAGNETIC CORE MEMORY



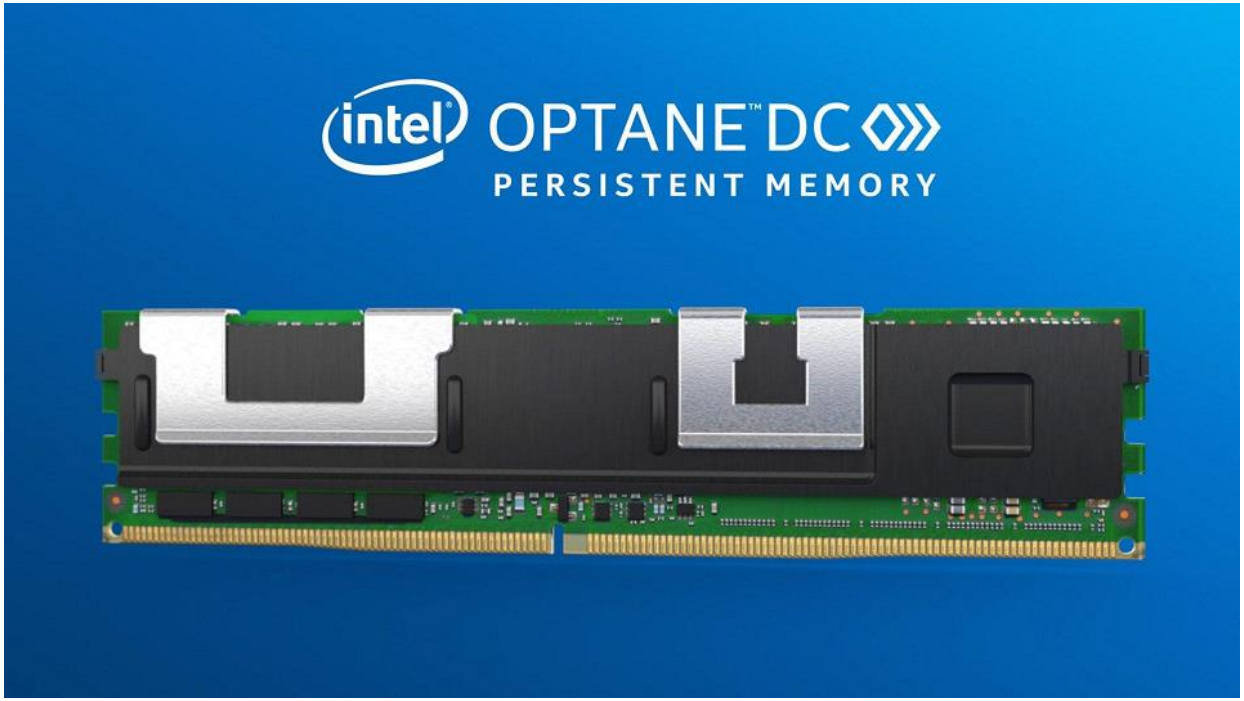
Source: https://en.wikipedia.org/wiki/Magnetic-core_memory

DRAM + FLASH



Source: https://storagehub.vmware.com/t/vsphere-storage/vsphere-6-7-core-storage-1/pmem-persistant-memory-nvdimm-support-in-vsphere/

3D XPOINT



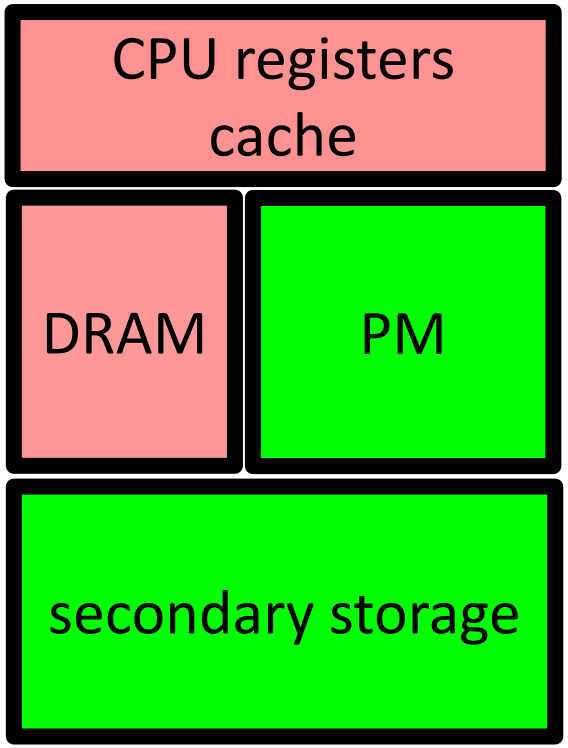
Source: https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html

MEMORY HIERARCHY

HYBRID MEMORY HIERARCHY

Volatile:

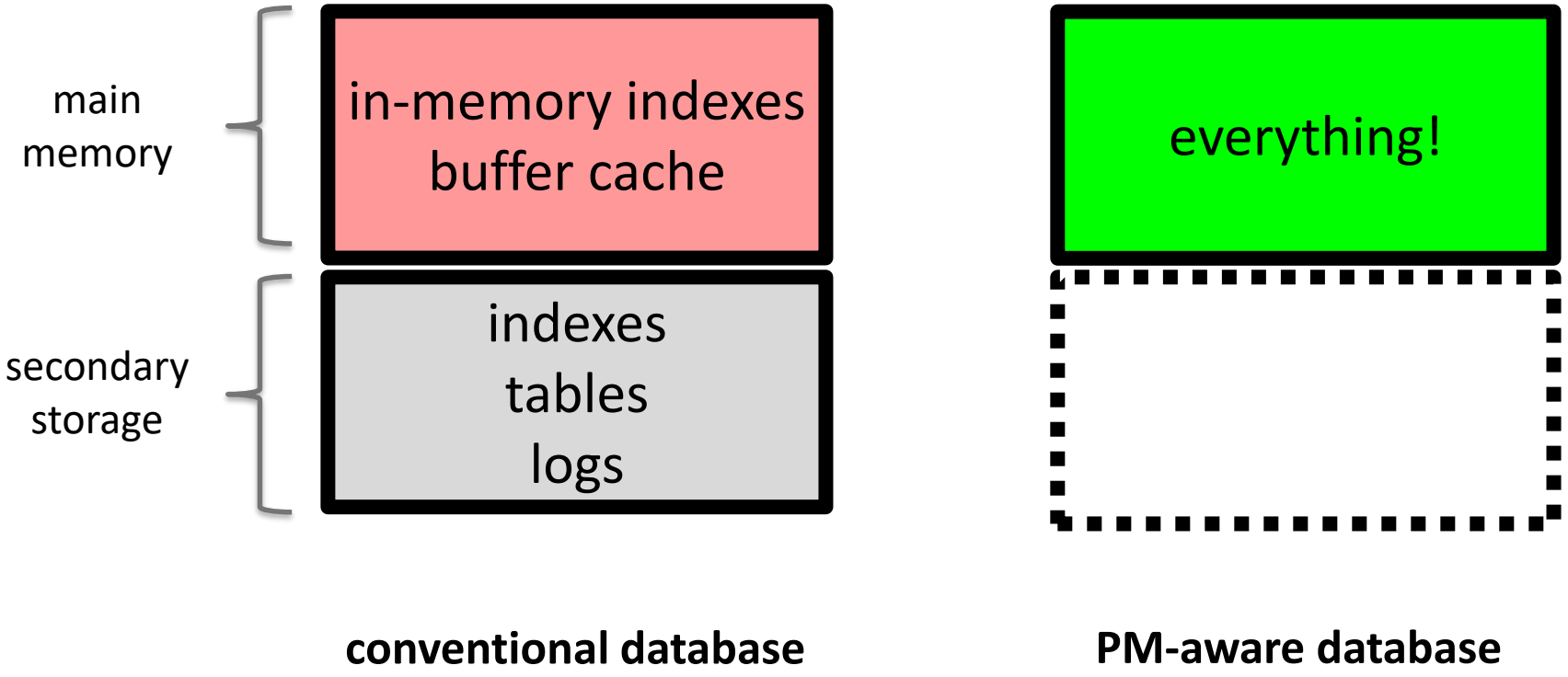
Non-volatile:



Optane DC
Persistent Memory

Optane Memory

WHAT CHANGES?



PARADIGM SHIFT

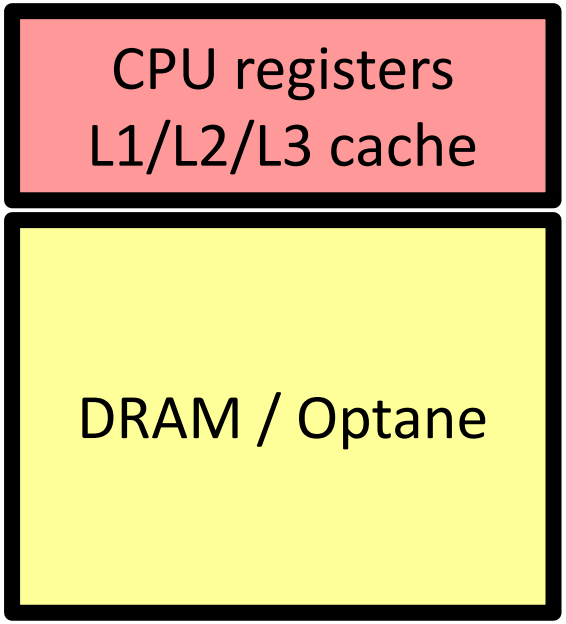
conventional in-memory data structure

recoverable in-memory data structure

OPTANE DCPM ACCESS MODES

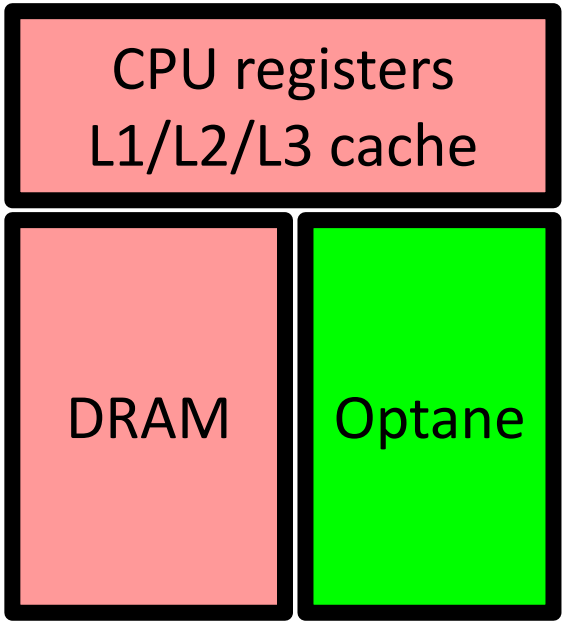
Mode	OS support required	Persistent?
Memory	NO	NO
App Direct	YES	YES

OPTANE DCPM ACCESS MODES



Memory Mode

(DRAM replacement)



App Direct Mode

(SSD replacement)

COST AND PERFORMANCE

COST

Product	128GB	256GB
Optane DCPM	$7/GB	$10/GB
DRAM (single-stick)	$34/GB	$95/GB
DRAM (multi-stick)	$5/GB	$7/GB
SSD	<$1/GB	<$1/GB

PERFORMANCE (LATENCY)

Product	Random Read	Random Write
Optane DCPM	~ 300ns	~ 100ns
DRAM	~ 100ns	~ 100ns
SSD	10-100µs	10-100µs

PERFORMANCE (BANDWIDTH)

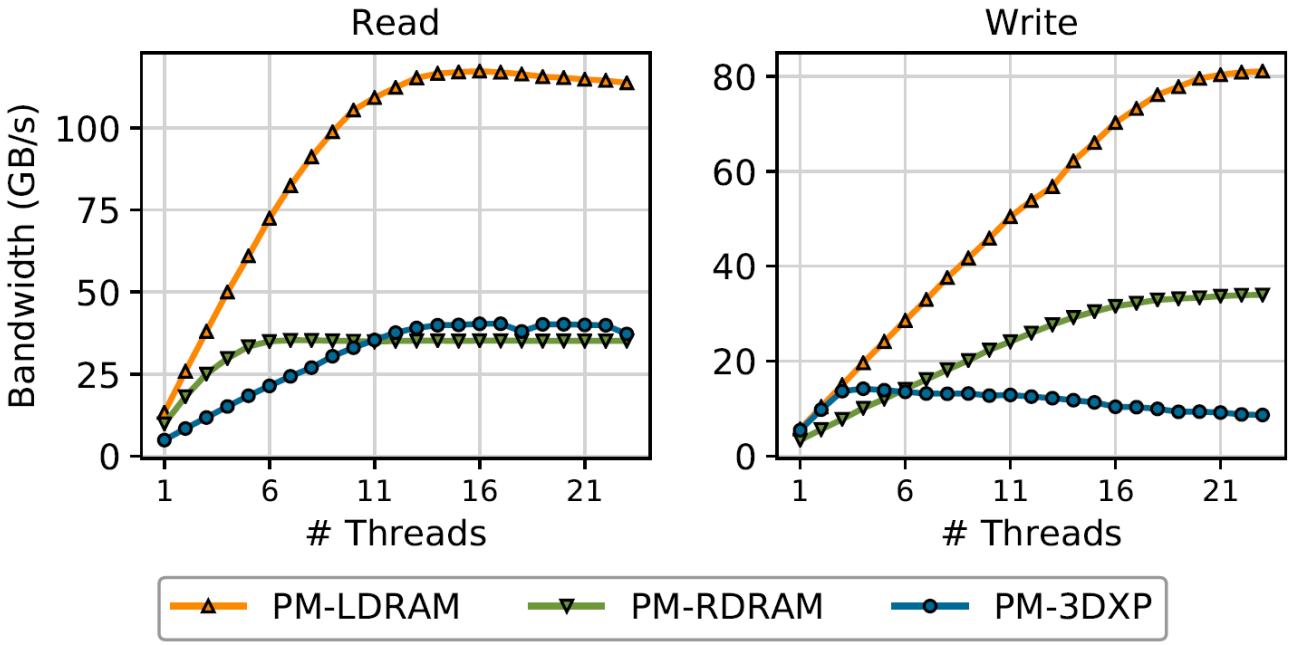


Figure 1: **Optane DC Sequential Bandwidth** The data show read (left) and write (right) bandwidth for an array of six Optane DC PMMs compared to a similar array of six DRAM DIMMs. Optane DC bandwidth is lower and, for writes, reaches saturation with fewer threads.

Izraelevitz et al. Basic Performance Measurements of the Intel Optane DC Persistent Memory Module. CoRRabs/1903.05714 (2019).

OPTANE VERSUS FLASH

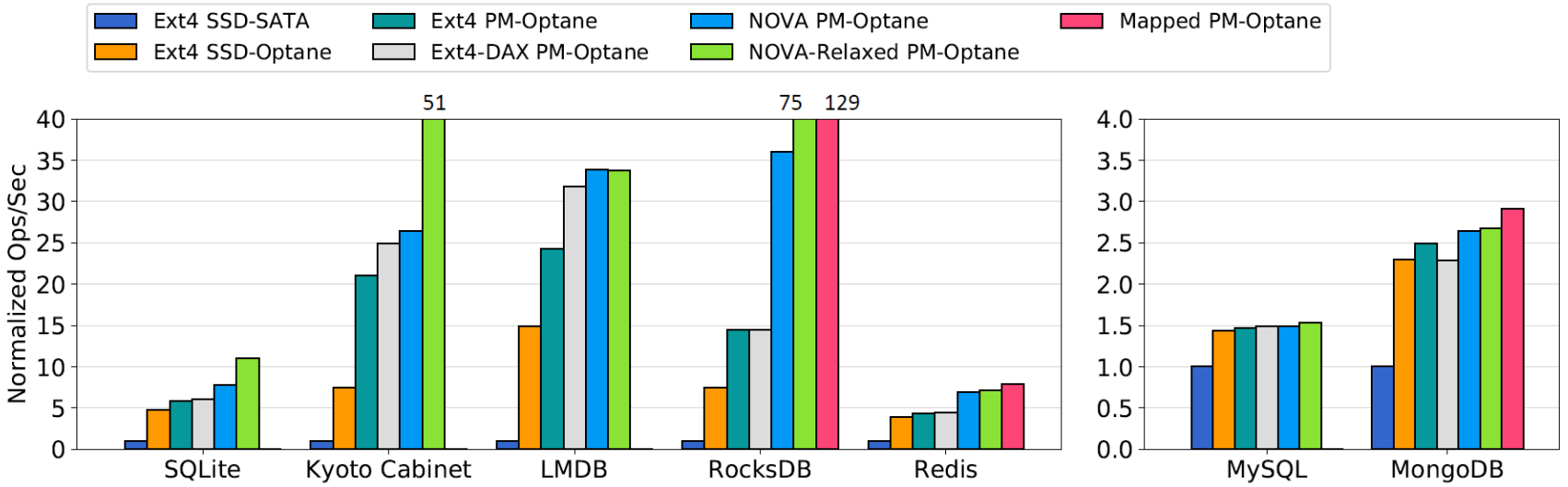


Figure 6: **Application Performance on Optane DC and SSDs** These data show the impact of more aggressively integrating Optane DC into the storage system. Replacing flash memory with Optane DC in the SSD gives a significant boost, but for most applications deeper integration with hardware (i.e., putting the Optane DC on a DIMM rather than an SSD) and software (i.e., using an PMEM-optimized file system or rewriting the application to use memory-mapped Optane DC) yields the highest performance.

Izraelevitz et al. Basic Performance Measurements of the Intel Optane DC Persistent Memory Module. CoRRabs/1903.05714 (2019).

OPTANE DCPM VERSUS DRAM

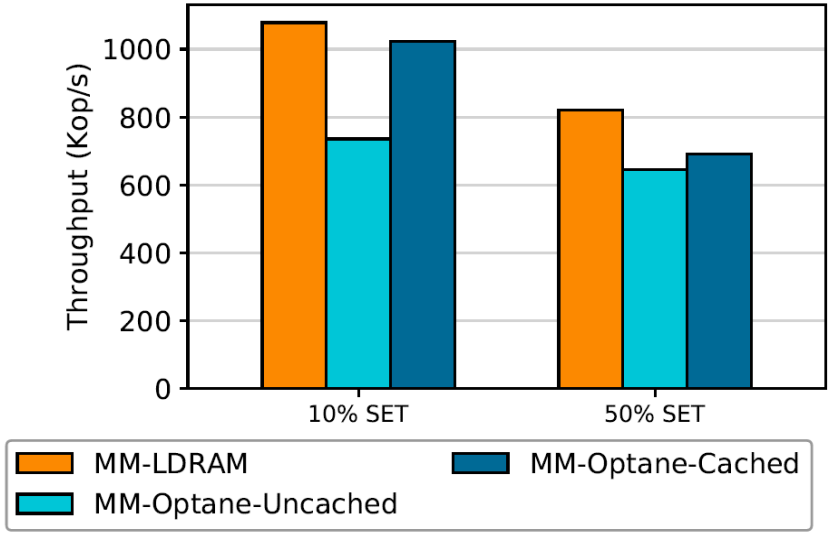


Figure 19: **Memcached on read/write workloads** This graph shows memcached throughput for different mixes of operations. Note that the DRAM cache is effective in hiding read latency, but has more trouble hiding write latency (see data in csvroot/memory/memcached.csv).

Izraelevitz et al. Basic Performance Measurements of the Intel Optane DC Persistent Memory Module. CoRRabs/1903.05714 (2019).

WHO BENEFITS?

- Category 1: disk-intensive systems (gain performance)
  - » databases
  - » storage systems
- Category 2: memory-intensive systems (gain fault tolerance, sacrifice performance)
  - » memcached

PURCHASING CONSIDERATIONS

- Optane DCPM presently requires Intel's 2nd generation Xeon Scalable processors (a.k.a. Cascade Lake).
- Only certain models support Optane DCPM (mainly Gold and Platinum SKUs).
- Vendor-recommended configurations include Optane + DRAM in various proportions.
- Cost: about $20-50k for a dual-socket server.

SYSTEM SUPPORT

PERSISTENT STORE OPERATIONS

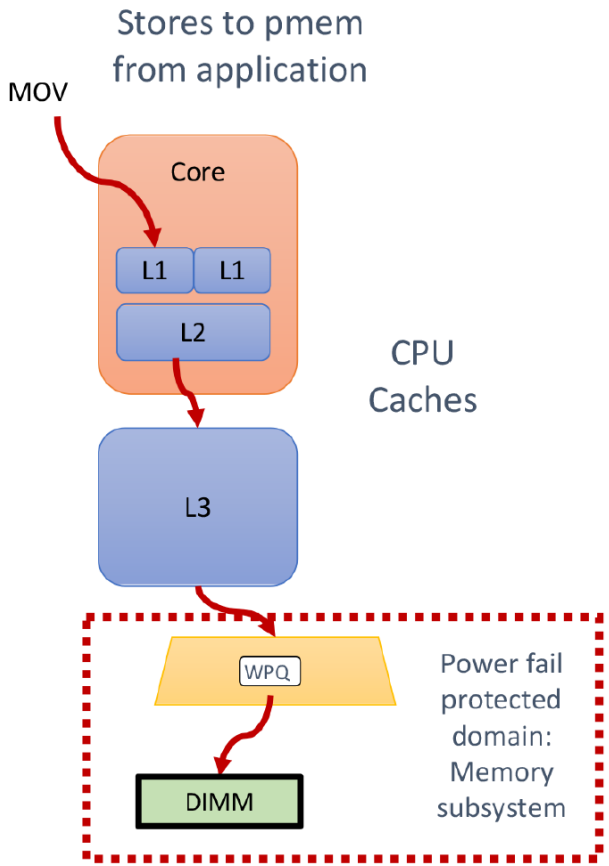


Figure 2: The path taken by a store, and the persistence domain (dashed box)

Rudoff. Persistent Memory Programming. ;login: 42(2) (2017).

SUPPORT FROM HARDWARE

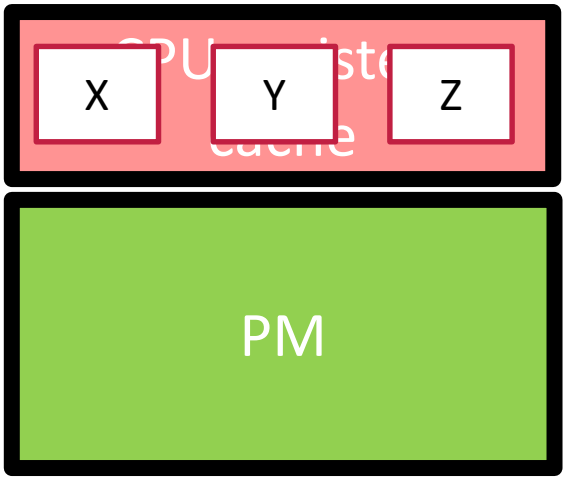
- Detect different memory types and configure the access mode.
- Flush Write Pending Queue (WPQ) to the PM DIMM in the event of power loss.
- Order writes and flush individual cache lines:
  - » SFENCE (store fence)
  - » CLFLUSH, CLFLUSHOPT (cache line flush)
  - » CLWB (cache line write back)

SUPPORT FROM OS

- Expose PM regions to applications as files.
- Memory-map files into app's address space.
  - » mmap, munmap
  - » msync ← replace with CLWB if possible!

SUPPORT FROM LIBRARY

- Wrapper around CLFLUSH and CLWB.
- Persistent heaps, objects, and pointers.
- Atomic multi-word operations or transactions.



1. write X
2. write Y
3. write Z
4. crash

RESEARCH DIRECTIONS

THEORY

- Specifying, implementing, and verifying PM-aware synchronization constructs.
  - » X-linearizable objects
  - » mutual exclusion
  - » consensus
- Understanding how the possibility of recovering from crash failures affects the complexity and computability of shared memory synchronization problems.

PRACTICE

- Designing PM-aware concurrent data structures, synchronization primitives, and transactional memories.
- Testing open-source implementations for bugs.