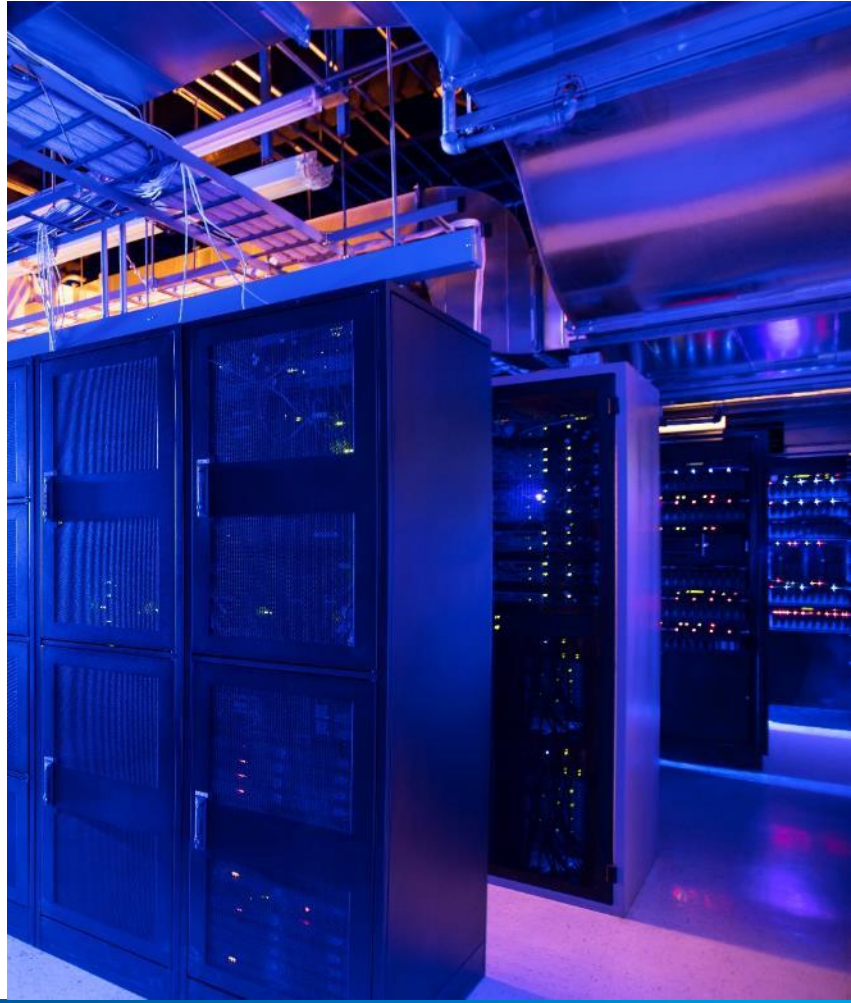
PERSISTENT MEMORY – WHICH MODE DO I WANT? WHERE ARE THE "GOTCHAS" HIDDEN?

Sudha Udanapalli Thiagarajan, Senior Performance Monitoring Engineer

sudha.udanapalli.thiagarajan@intel.com

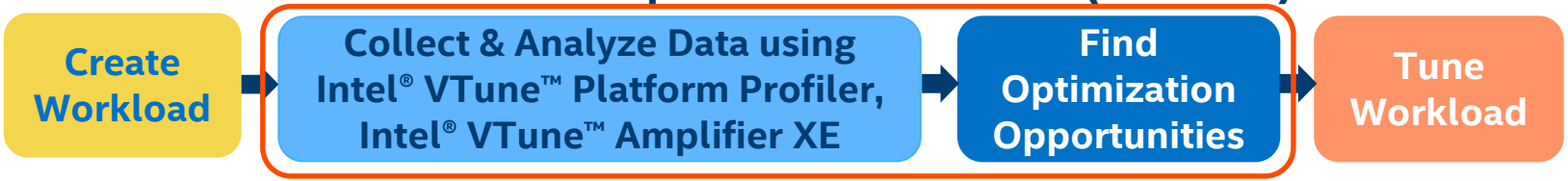
AGENDA

- Intro & Context
- Part 1: Architectural Background
- Part 2: Performance Characteristics
- Part 3: Monitoring & Tuning
- Summary & Resources



FOCUS OF THIS PRESENTATION

General Workload Optimization Process (Iterative):



We **Will** Address

What architectural conditions should I be monitoring, and why?

How is my workload utilizing the hardware?

We **Won't** Address

How do I use PMDK/SPDK?

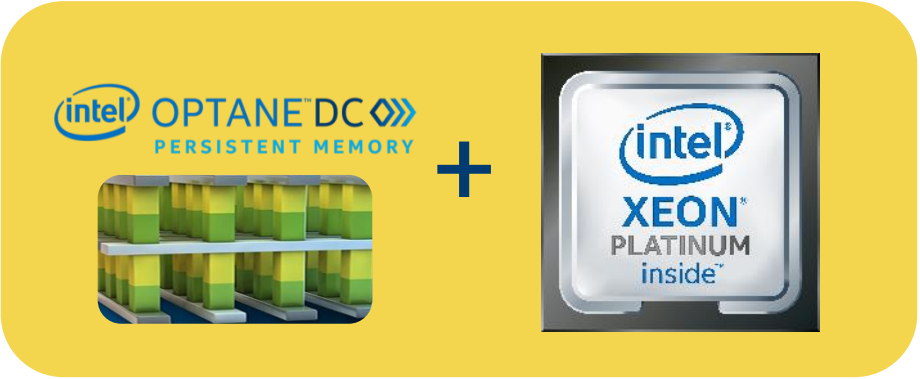
How do I use the Intel analyzer tools to collect data?

Are there better design patterns I could use For more performance?

Could I configure my O/S and environment for better performance?

How do I change my code?

PART 1:
KEY ARCHITECTURAL CONCEPTS:
2ND GENERATION INTEL® XEON® SCALABLE PROCESSORS,
OPTANE™ DC PERSISTENT MEMORY



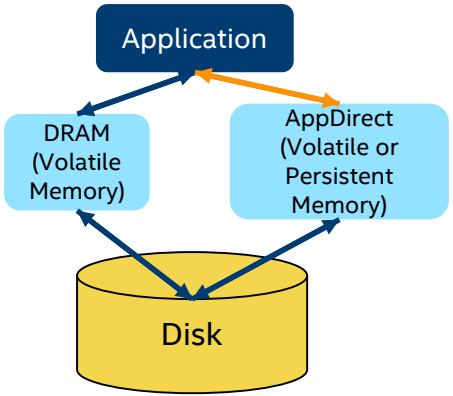
GET READY FOR TERMS, FLOWS, AND DIAGRAMS!

Why?

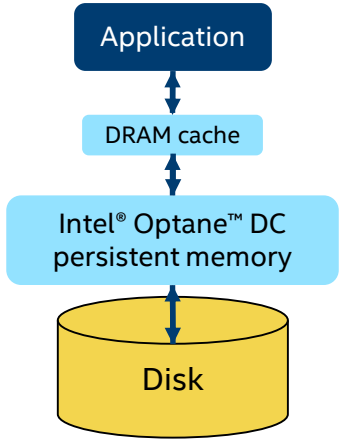
To establish background in the architecture, which will be needed to understand the monitoring and tuning strategies to optimize.

INTEL® OPTANE™ DC PERSISTENT MEMORY MODES

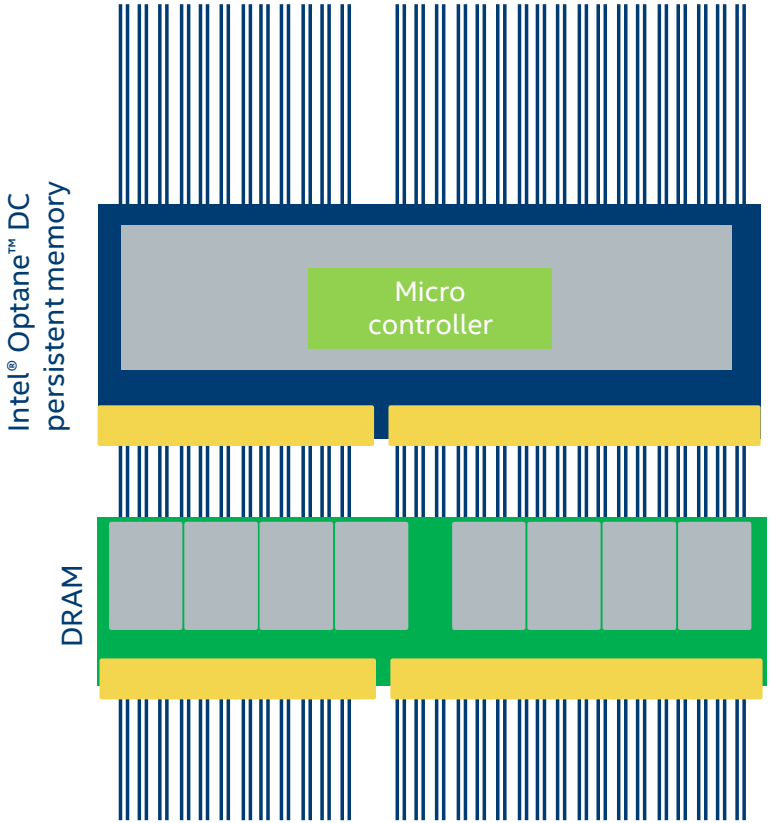
- Application Direct (AppDirect) mode uses Intel® Optane™ DC persistent memory to provide volatile or non-volatile (persistent) memory to applications
  - Intel® Optane™ DC persistent memory modules are accessed by loads and stores or using persistent memory libraries
  - Intel® Optane™ DC persistent memory is treated as a separate region of memory from DRAM



- Memory mode uses Intel® Optane™ DC persistent memory as an additional tier of memory between DRAM and disk
  - DRAM is a **direct-mapped** cache for Intel® Optane™ DC persistent memory, which is transparent to the application developers
  - Intel® Optane™ DC persistent memory tier acts as temporary storage for disk data (not persistent)



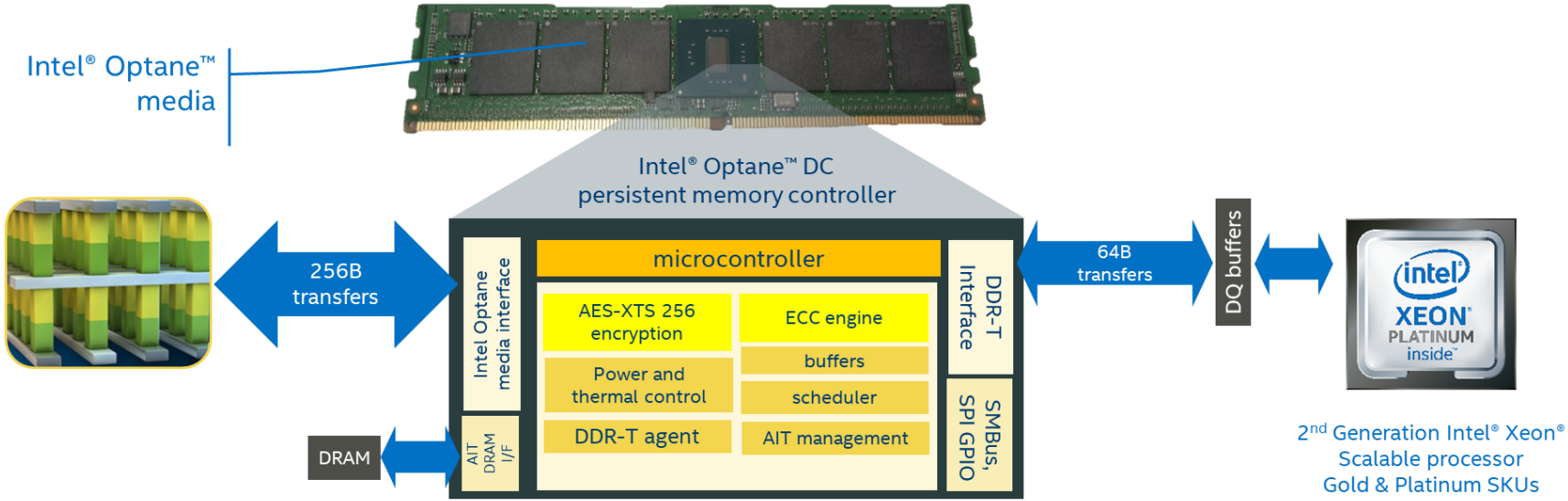
INTEL® OPTANE™ DC PERSISTENT MEMORY MODULE CONCEPTS



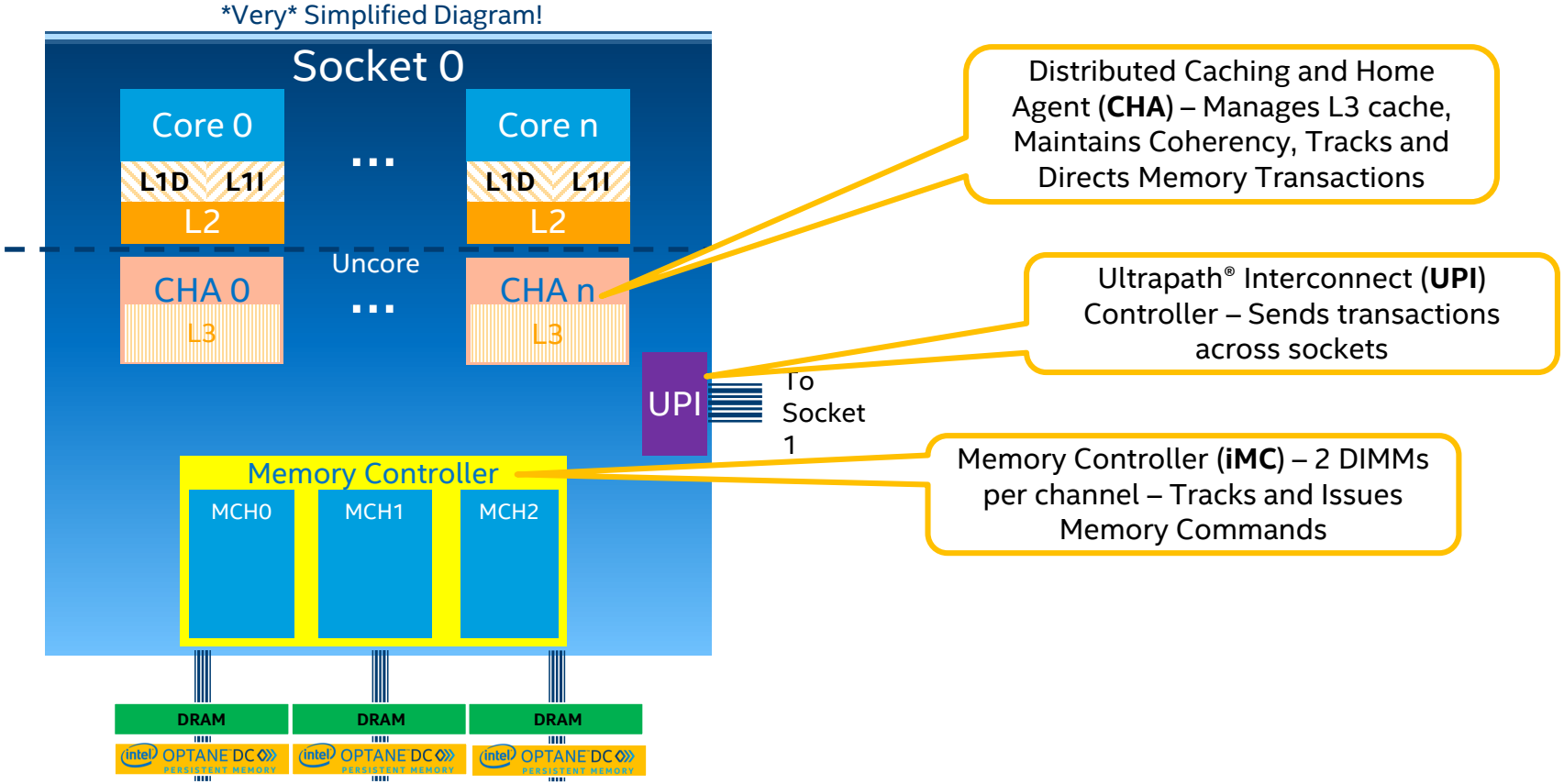
Up to 6 Intel® Optane™ DC persistent memory modules per socket

- 3 modules per memory controller, 2 memory controllers per socket.
- DRAM DIMMS share **the same bus** with Intel® Optane™ DC persistent memory modules.

AN IMPORTANT MICROCONTROLLER CONCEPT

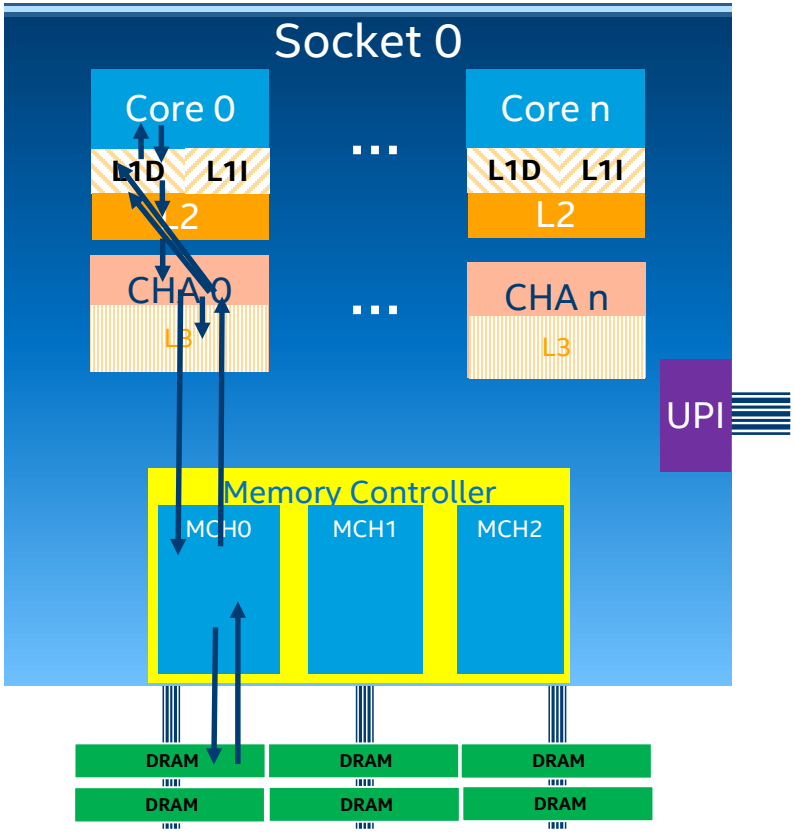


FEW KEY COMPONENTS IN THE SYSTEM



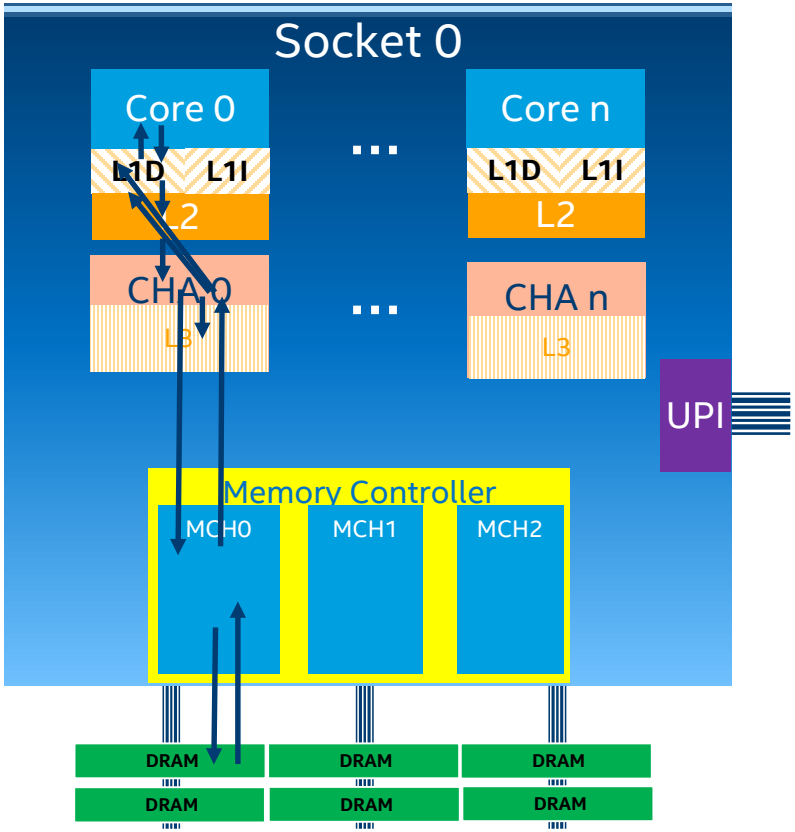
A BASIC MEMORY TRANSACTION TO DRAM

*Simplified! Some steps not shown,
Some components not depicted



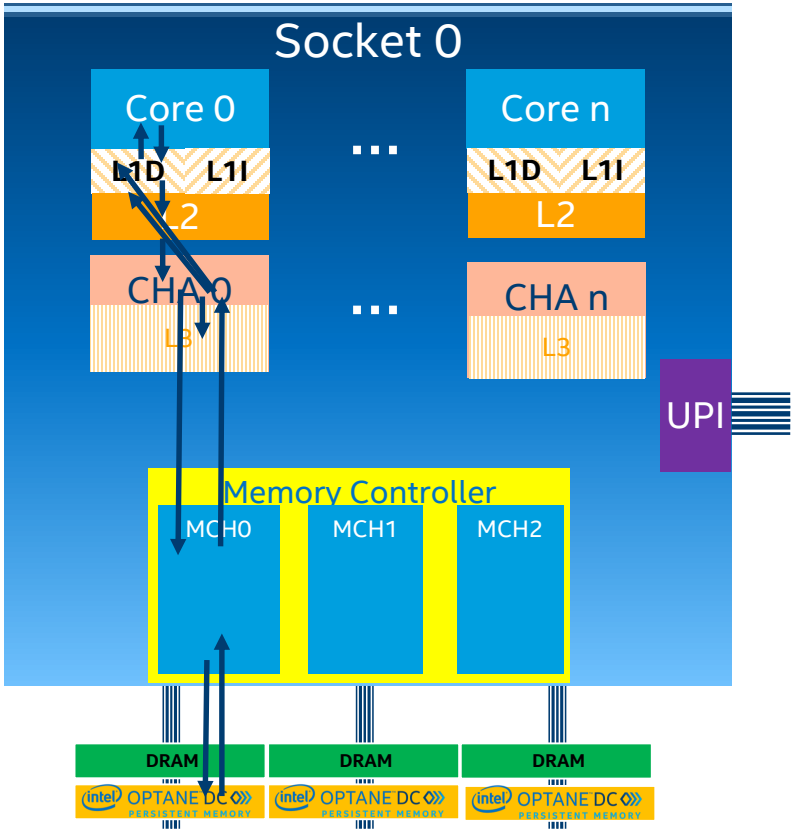
A BASIC MEMORY TRANSACTION TO DRAM

*Simplified! Some steps not shown,
Some components not depicted



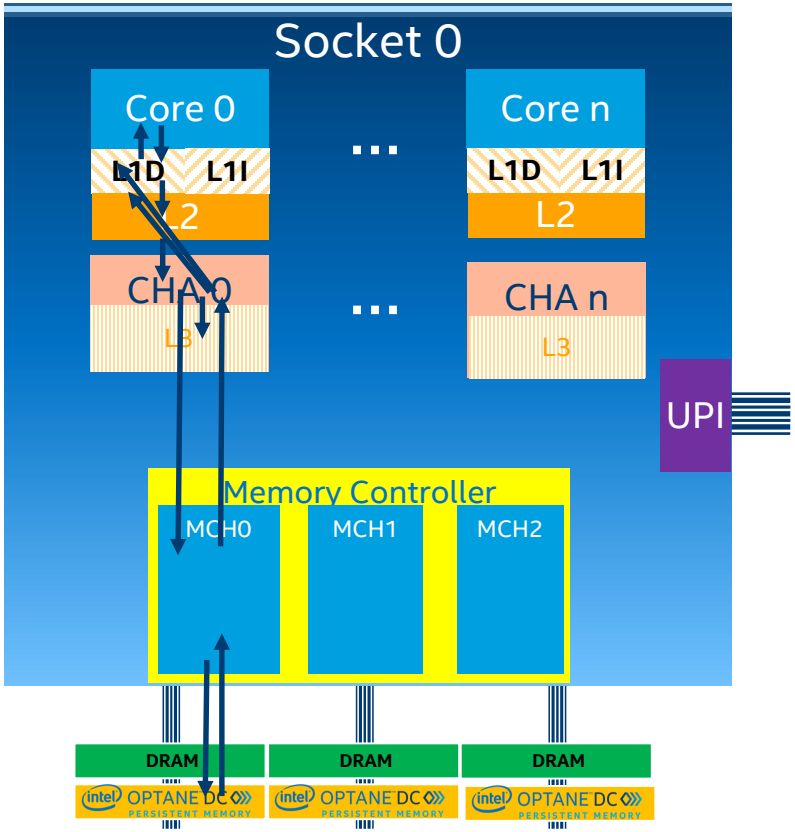
READ TRANSACTION TO APPDIRECT MEMORY

*Simplified! Some steps not shown,
Some components not depicted



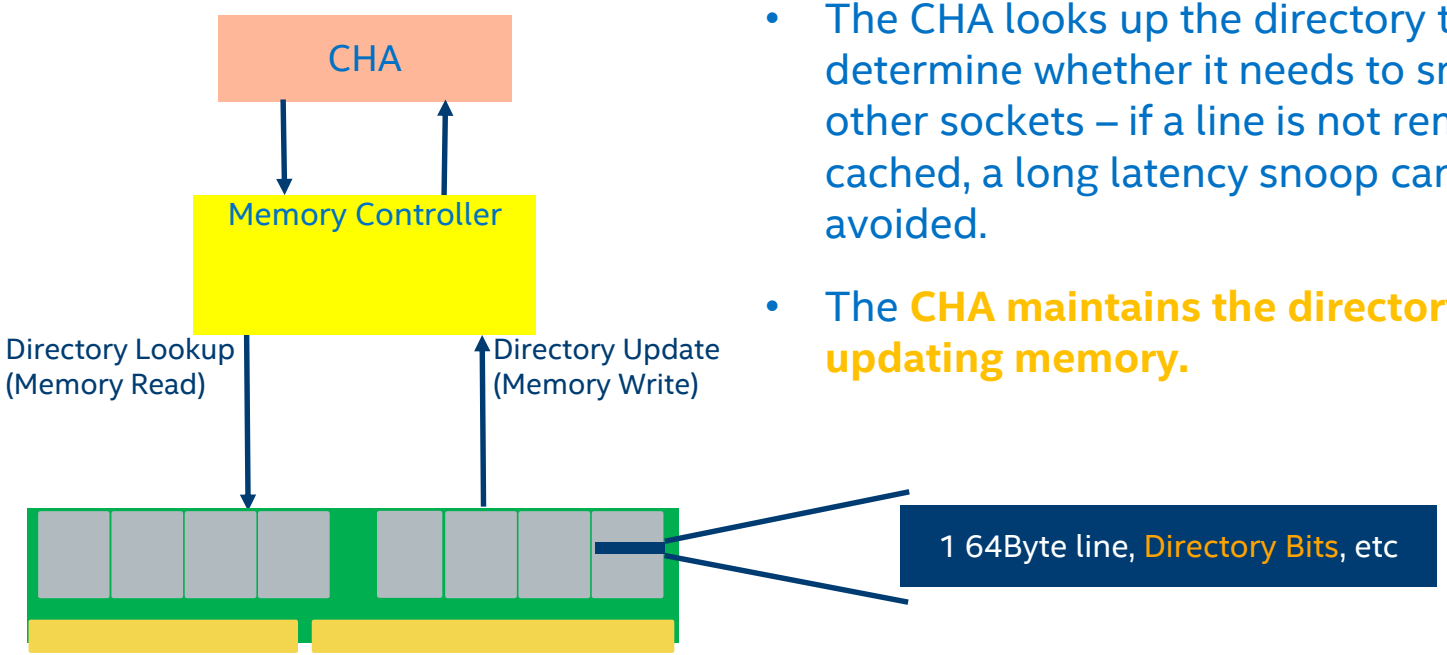
READ TRANSACTION TO APPDIRECT MEMORY

*Simplified! Some steps not shown,
Some components not depicted



ONE IMPORTANT CONCEPT: THE DIRECTORY

- Directory is structure stored **in memory** that tracks whether a line is cached in a remote socket



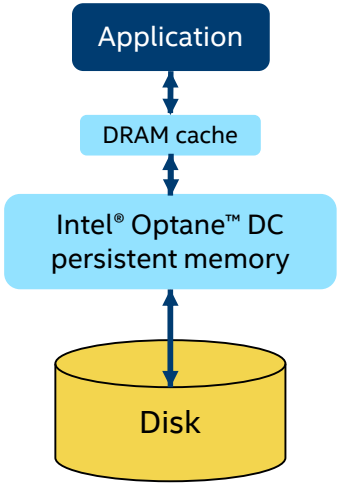
- The CHA looks up the directory to determine whether it needs to snoop other sockets – if a line is not remotely cached, a long latency snoop can be avoided.
- The **CHA maintains the directory by updating memory.**

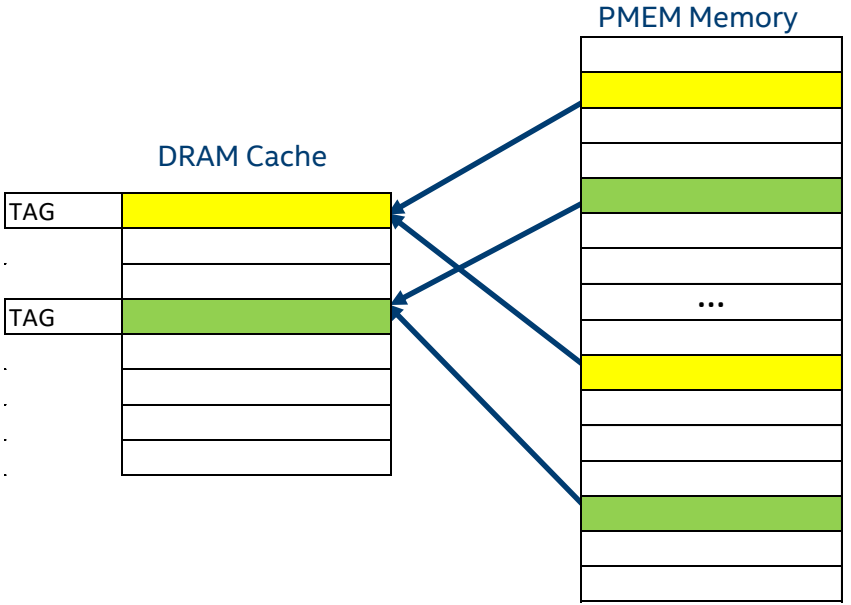
KEY TAKEWAYS OF APPDIRECT MODE

Performance varies by workload

- AppDirect memory can be used for persistence
- Lower latency than disk accesses
- Enable larger memory data structures
  - Move in-memory data structures from disk to AD memory
    - Move hot data structures to DRAM, warm data structures to Intel® Optane DC persistent memory
- Optimize for NUMA awareness

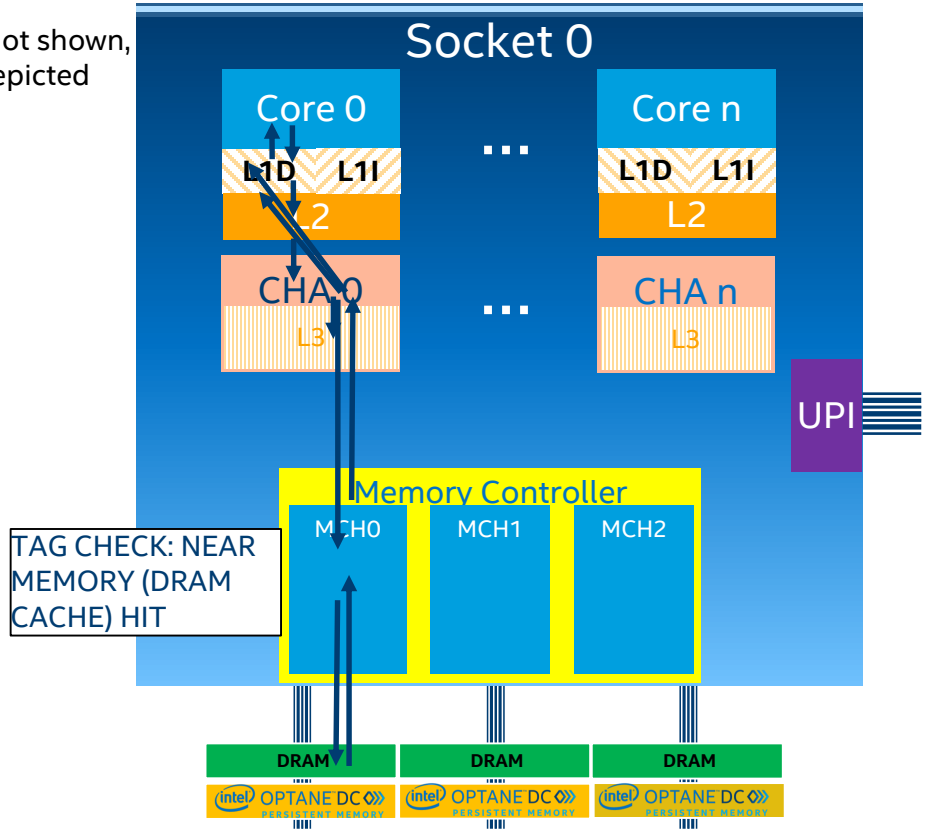
HOW IS INTEL® OPTANE™ DC PERSISTENT MEMORY IS CACHED IN DRAM?





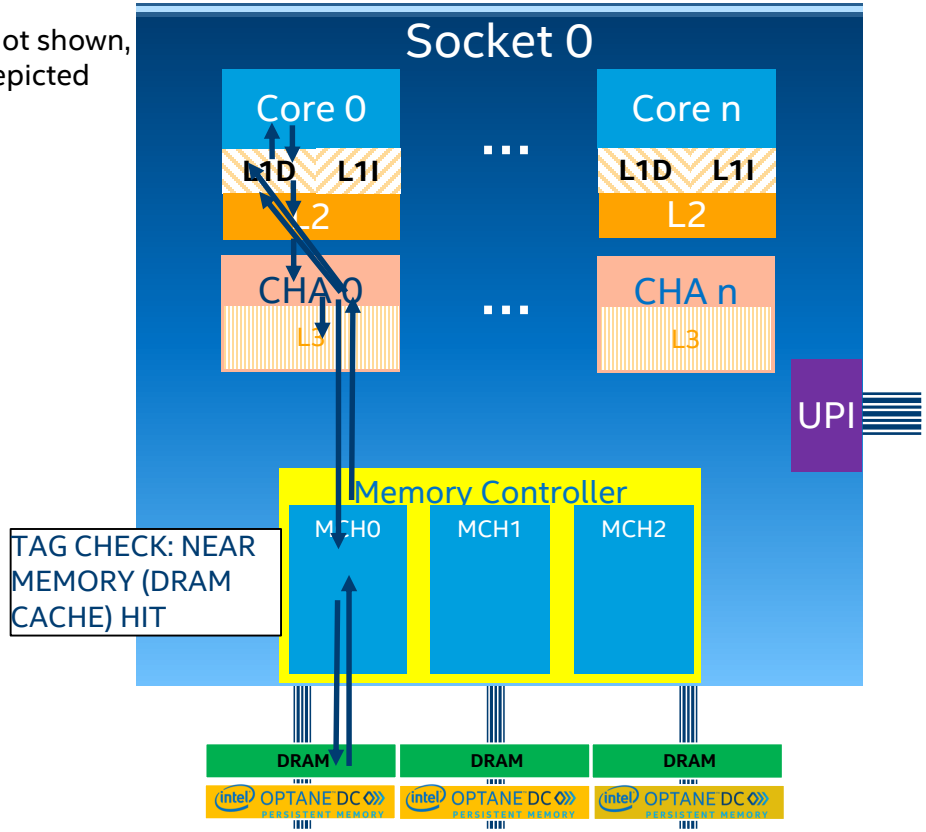
READ TRANSACTION TO MEMORY MODE - DRAM CACHE HIT

*Simplified! Some steps not shown,
Some components not depicted



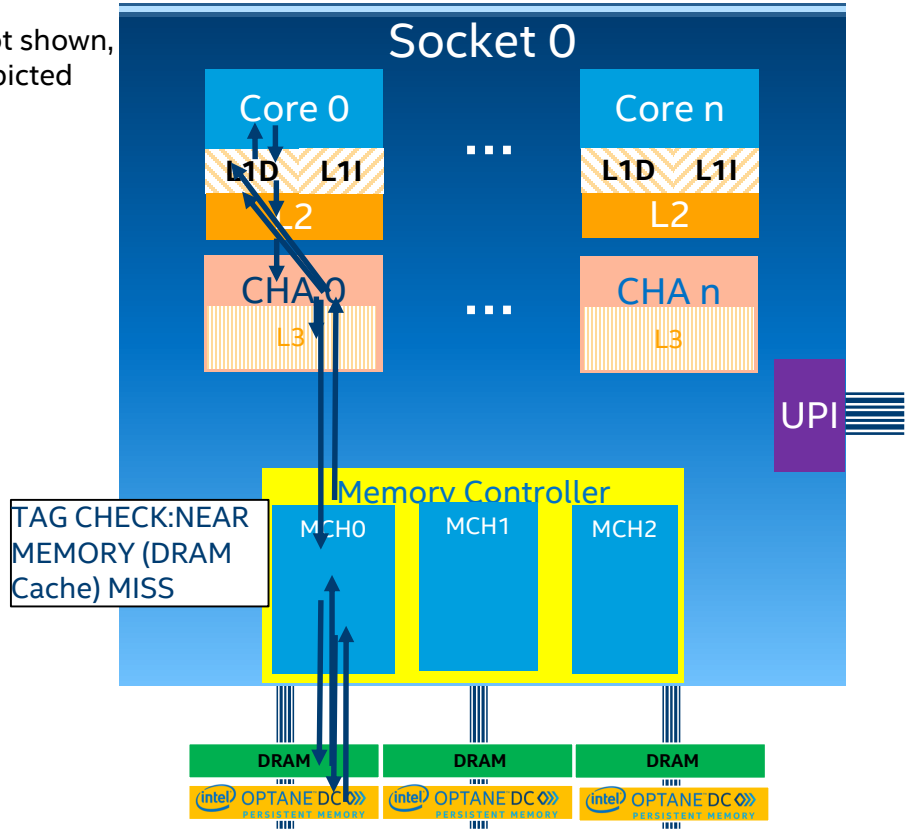
READ TRANSACTION TO MEMORY MODE - DRAM CACHE HIT

*Simplified! Some steps not shown,
Some components not depicted



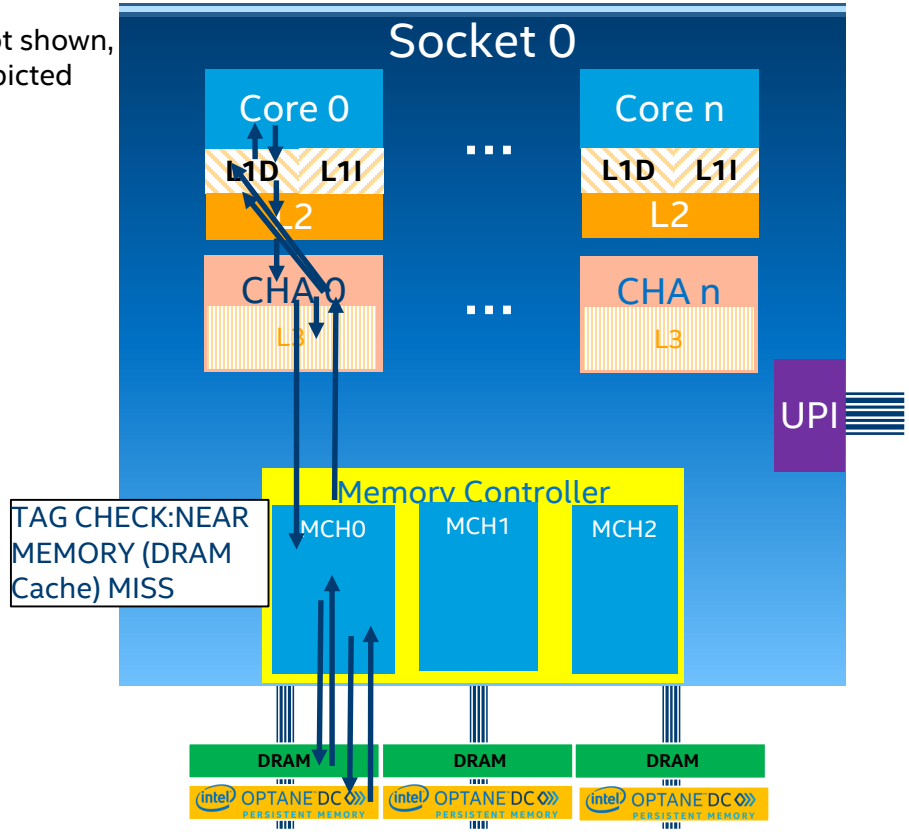
READ TRANSACTION TO MEMORY MODE - DRAM CACHE MISS

*Simplified! Some steps not shown,
Some components not depicted



READ TRANSACTION TO MEMORY MODE - DRAM CACHE MISS

*Simplified! Some steps not shown,
Some components not depicted



KEY TAKEAWAYS OF MEMORY MODE

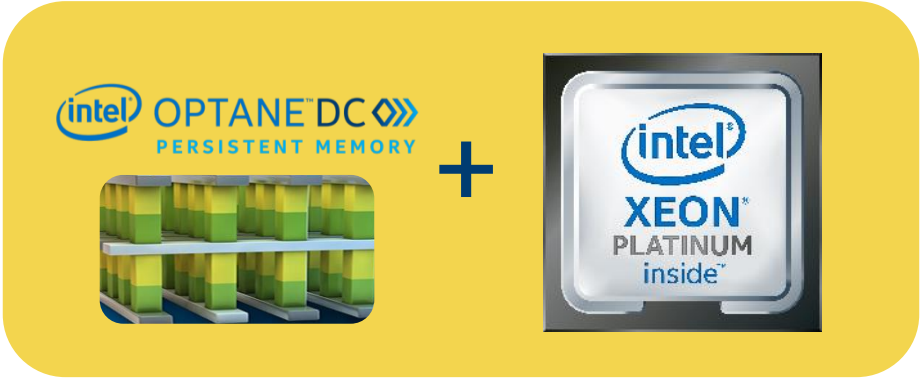
Good locality means near-DRAM performance

- Cache hit: latency same as DRAM
- Cache miss: latency DRAM + Intel® Optane™ DC persistent memory

Performance varies by workload

- Best workloads have the following traits:
  - Good locality for high DRAM cache hit rate
  - Low memory bandwidth demand
  - #reads > #writes
- Also take into consideration
  - Config vs. Workload size
- Optimize for NUMA awareness

PART 2:
PERFORMANCE CHARACTERISTICS OF INTEL® OPTANE™ DC PERSISTENT MEMORY



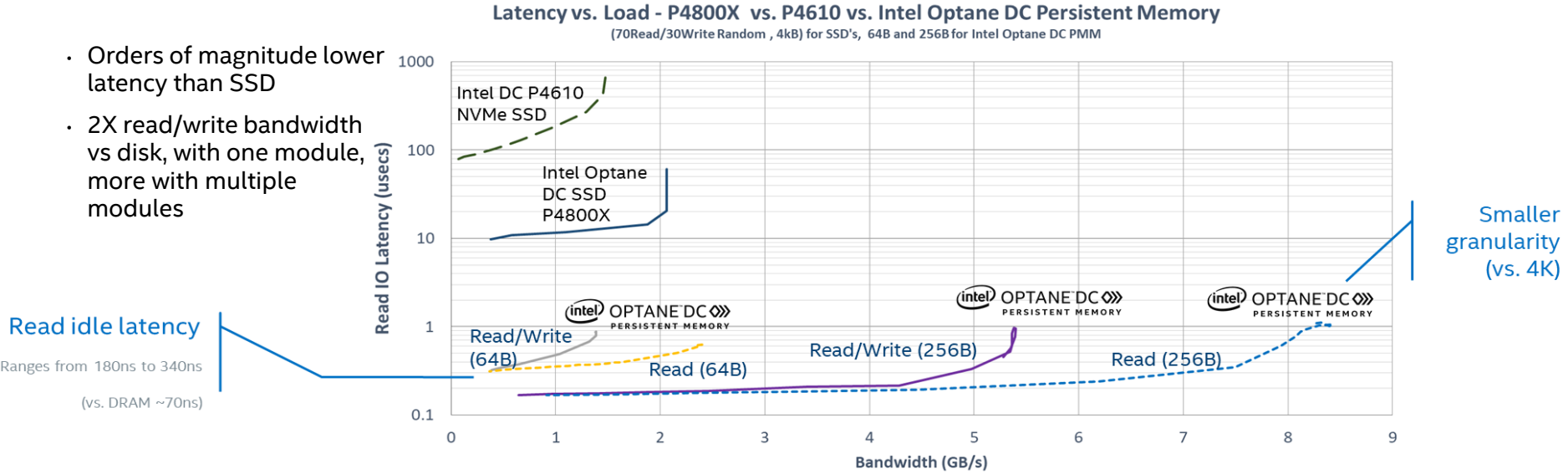
INTEL® OPTANE™ DC PERSISTENT MEMORY PERFORMANCE DETAILS

- Intel® Optane™ DC persistent memory is programmable for different power limits for power/performance optimization
  - 12W – 18W, in 0.25 watt granularity - for example: 12.25W, 14.75W, 18W
  - Higher power settings give best performance
- Performance varies based on traffic pattern
  - Contiguous 4 cacheline (256B) granularity vs. single random cacheline (64B) granularity
  - Read vs. writes

Granularity	Traffic	Module	Bandwidth
256B (4x64B)	Read	256GB, 18W	8.3 GB/s
256B (4x64B)	Write		3.0 GB/s
256B (4x64B)	2 Read/1 Write		5.4 GB/s
64B	Read		2.13 GB/s
64B	Write		0.73 GB/s
64B	2 Read/1 Write		1.35 GB/s

INTEL® OPTANE™ DC PERSISTENT MEMORY LATENCY

- Orders of magnitude lower latency than SSD
- 2X read/write bandwidth vs disk, with one module, more with multiple modules



1. 256B granularity (64B accesses). Note 4K granularity gives about same performance as 256B

PART 3:
MONITORING AND TUNING STRATEGIES

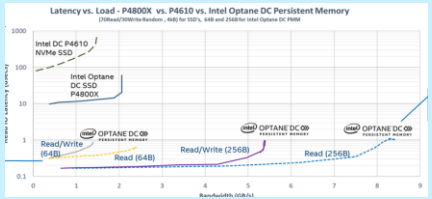
APPDIRECT MODE

OPTIMIZATION OPPORTUNITY: READ/WRITE RATIO

intel OPTANE DC PERSISTENT MEMORY

Workload Behavior:
Write-intensive application

Why?
Reads to Intel® Optane Persistent Memory are faster than writes



How to detect this opportunity?

Use Intel® VTune™ Platform Profiler

- Persistent memory traffic: Read/Write throughput

Use Intel® VTune™ Amplifier

- Persistent memory bound metric under memory access analysis.

Potential Optimizations:

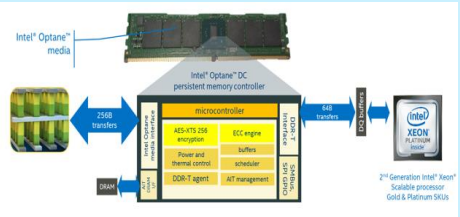
- Move write-sensitive data to DRAM
- Increase Intel® Optane™ DC persistent memory modules. 6 modules per socket is ideal. This will give you more bandwidth.

APPDIRECT MODE
OPTIMIZATION OPPORTUNITY: ACCESS PATTERN

intel OPTANE™ DC PERSISTENT MEMORY

Workload Behavior:
Random media access pattern

Why?
Will not take advantage of 256B transfers



How to detect this opportunity?

Use Intel® VTune™ Platform Profiler

Potential Optimizations:

- Stream data in aligned 256B chunks
- Don't stream from multiple media locations in parallel

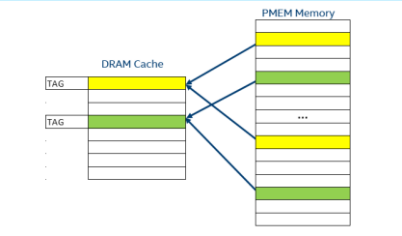
MEMORY MODE

OPTIMIZATION OPPORTUNITY: CACHE LOCALITY

intel OPTANE DC PERSISTENT MEMORY

Workload Behavior:
Sensitivity to DRAM cache misses

Why?
Working set size > DRAM capacity results in conflict misses



How to detect this opportunity?

Use Intel® VTune™ Platform Profiler
Memory Mode analysis
- Near memory cache read miss rate % metric

Use Intel® VTune™ Amplifier XE: Memory Access
- DRAM cache hit ratio under memory-access analysis.

Potential Optimizations:

- Improve DRAM Cache locality
- Consider using AppDirect volatile memory
- Ideal DRAM memory capacity : Intel® Optane DC persistent memory ratio = 1:4 or 1:8

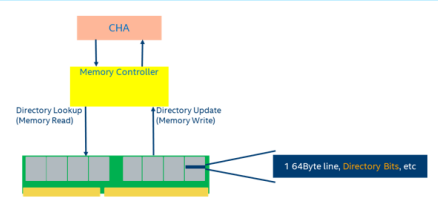
APPDIRECT MODE & MEMORY MODE

OPTIMIZATION OPPORTUNITY: DIRECTORY OVERHEAD

intel OPTANE DC PERSISTENT MEMORY

Workload Behavior:
Unnecessary Cross-socket accesses

Why?
Directory maintenance uses memory bandwidth



How to detect this opportunity?

Use Intel® VTune™ Platform Profiler

- Extra write bandwidth due to Directory Updates metric

Potential Optimizations:

- Optimize for NUMA
- Eliminate false sharing across sockets
- Investigate BIOS settings

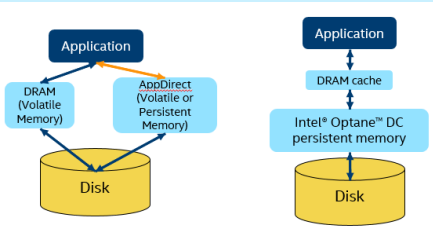
APPDIRECT MODE & MEMORY MODE

OPTIMIZATION OPPORTUNITY: DISK ACCESSES

intel OPTANE DC PERSISTENT MEMORY

Workload Behavior:
Disk I/O

Why?
Accessing data in Intel® Optane DC persistent memory is preferred



How to detect this opportunity?

Use Intel® VTune™ Platform Profiler : I/O view

- IOPS metric

Use Intel® VTune™ Amplifier XE : Input and Output analysis

Potential Optimizations:

- Keep working dataset in Intel® Optane DC persistent memory

SUMMARY

Architected for Persistence, Optimized for Performance

- Offers Persistent capabilities
- Near DRAM latency and faster than SSDs

Flexibility In Operation

- Choose between two modes
  - Memory mode for large affordable volatile memory
  - App Direct mode for persistent memory

Monitor and Tune!

- Take advantage of the modes effectively
- Various software tools available to monitor

RESOURCES: DOWNLOAD THE TOOLS, GET THE TRAINING

Intel® VTune™ Amplifier – Performance Profiler

- software.intel.com/vtune

Intel® Inspector – Persistence and Thread Debugger

- software.intel.com/inspector

Intel® Optane™ DC Persistent Memory Technical Articles:

- software.intel.com/persistent-memory/tools
- software.intel.com/articles/prepare-for-the-next-generation-of-memory

SPDK / DPDK Technical Articles:

- User guide: SPDK IO Data View
  software.intel.com/vtune-amplifier-help-spdk-code-analysis
- IO Issues: Remote socket access
  software.intel.com/vtune-amplifier-cookbook-i/o-issues-remote-socket-accesses

BACKUP

CONFIGURATION FOR “P4800 AND P4610

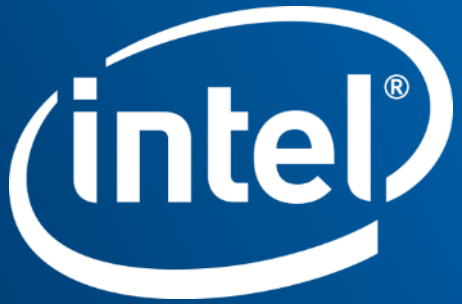
SSDs

Intel-tested: Measured using FIO 3.1. Common Configuration - Intel 2U Server System, OS CentOS 7.5, kernel 4.17.6-1.el7.x86_64, CPU 2 x Intel® Xeon® 6154 Gold @ 3.0GHz (18 cores), RAM 256GB DDR4 @ 2666MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P4610 3.2TB. Intel Microcode: 0x2000043; System BIOS: 00.01.0013; ME Firmware: 04.00.04.294; BMC Firmware: 1.43.91f76955; FRUSDR: 1.43. The benchmark results may need to be revised as additional testing is conducted. Performance results are based on testing as of November 15, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

INTEL® OPTANE™ DC PERSISTENT MEMORY LOADED LATENCY CURVE CONFIG

Component	Single DIMM Config
Test by	Intel
Test date	02/20/2019
Platform	NeonCity
Chipset	LBG B1
CPU	CLX B0 28 Core (8276), 1S
DDR Speed	2666 MT/s
AEP	QS Tranche3, 256GB, 18W
Memory Config	1 channel 32GB DDR4 (per socket) 256GB AEP (per socket)
AEP FW	5336
BIOS	573.D10
BKC version	WW08 BKC
Linux OS	4.20.4-200.fc29
Spectre/ Meltdown	Patched (1,2,3, 3a)
Performance Turning	QoS Disabled, IODC=5(AD)

All SKUs, frequencies, and performance estimates are PRELIMINARY and can change without notice.


intel®