INTEL® OPTANE™ DATA CENTER
PERSISTENT MEMORY

Architecture (Jane) and Performance (Lily)

Presenters: Lily Looi, Jianping Jane Xu

Co-Authors: Asher Altman, Mohamed Arafa, Kaushik Balasubramanian, Kai Cheng, Prashant Damle, Sham Datta, Chet Douglas, Kenneth Gibson, Benjamin Graniello, John Grooms, Naga Gurumoorthy, Ivan Cuevas Escareno, Tiffany Kasanicky, Kunal Khochare, Zhiming Li, Sreenivas Mandava, Rick Mangold, Sai Muralidhara, Shamima Najnin, Bill Nale, Jay Pickett, Shekoufeh Qawami, Tuan Quach, Bruce Querbach, Camille Raad, Andy Rudoff, Ryan Saffores, Ian Steiner, Muthukumar Swaminathan, Shachi Thakkar, Vish Viswanathan, Dennis Wu, Cheng Xu

08/19/2019
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks.

Configurations on slides 18 and 20.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer.

Performance results are based on testing as of Feb. 22, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

*Other names and brands may be claimed as property of others.

Intel, the Intel logo, Xeon, the Xeon logo, Optane, and the Optane logo are trademarks of Intel Corporation in the United States and other countries.

© Intel Corporation.
1. Intel® Optane™ DC Persistent Memory Architecture

A Breakthrough with a New Interface Protocol, Memory Controller, Media, and Software Stack
MEMORY- STORAGE GAP

Access Distribution

Hot data

Cooler data

Data Access Frequency

Hot data → less often → Cooler data

Access Distribution

Memory-Storage Gap

DRAM HOT TIER

SSD WARM TIER

INTEL® 3D NAND SSD

HDD / TAPE COLD TIER

Memory Sub-System

SSD

Network Storage

10s GB < 100 nanoseconds

10s TB < 100 micro seconds

10s TB < 100 milliseconds

Intel® 3D Nand SSD

CPU LLC Core L1

CPU Core L2

CPU Core L2

CPU Core L2

CPU Core L2

CPU Core L2

CPU Core L2

CPU Core L2

CPU Core L2

CPU Core L2

Intel" Logo

Hot data

Cooler data
CLOSE MEMORY – STORAGE GAP

Optimize performance given cost and power budget

Move Data Closer to Compute
Maintain Persistency

Data Access Frequency

Hot data
Cooler data

SSD
WARM TIER

INTEL® 3D NAND SSD

HDD / TAPE
COLD TIER

DRAM
HOT TIER
**INTEL® OPTANE™ MEDIA TECHNOLOGY**

**High Resistivity – ‘0’**
**Low Resistivity – ‘1’**

**Attributes**
- Non-volatile
- Potentially fast write
- High density
- Non-destructive fast read
- Low voltage
- Integrate-able w/ logic
- Bit alterable

**First Generation Capacities:**
- 128 GB
- 256 GB
- 512 GB

**Cross-Point Structure**
Selectors allow dense packing
And individual access to bits

**Breakthrough Material Advances**
Compatible switch and memory cell materials

**Scalable**
Memory layers can be stacked in a 3D manner

**High Performance**
Cell and array architecture that can switch fast
1. DQ buffers presents a single load to the host
2. Host SMBus: SPD visible to the CPU, Optane Controller plays thermal sensing (TSOD) functionality
3. Address Indirection Table
4. Integrated PMIC controlled Optane Controller
5. On DIMM Firmware storage
6. On-DIMM Power Fail Safe with auto-detection
INTEL® OPTANE™ DC PERSISTENT MEMORY CONTROLLER ARCHITECTURE

DDM4 SLOT ON HOST CPU

Interface to Host CPU

DCPMM Memory Interface

Addr Mapping Cache

Address Mapping Logic

Encrypt/Decrypt

Uctrl

Media Management

Power & Thermal Mgmt

DRNG

Key Mgmt

Scheduler

Read Queue

Write Queue

ECC/Scrambler

Error Handling Logic

Refresh Engine

Optane™ Media Channel

Optane™ Media Devices

Caps for Flushes
INTEL® OPTANE DC PERSISTENT MEMORY SW ENABLING STACK

- MANAGEMENT UI
- MANAGEMENT LIBRARY
- APPLICATION
  - Standard Raw Device Access
- APPLICATION
  - Standard File API
- APPLICATION
  - Standard File API
  - Load/Store
- PMEM-AWARE FILE SYSTEM
- MMU Mappings
- GENERIC NVDIMM DRIVER
- FILE SYSTEM
- PERSISTENT MEMORY

“DAX”
1. AC power loss to de-assert the PWROK
2. Platform logic then asserts the ADR_Trigger
3. PCH starts the ADR programmable timer
4. PCH assertion to SYNC message
5. PCU in processor detects SYNC message bit and sends AsyncSR to MC
6. MC flushes Write pending queue (WPQ)
7. After ADR timer expires, PCH asserts ADR_COMPLETE pin
MEMORY MODE

- Large Memory Capacity
- No software/application changes required
- To mimic traditional memory, data is “volatile”
  - Volatile mode key cleared and regenerated every power cycle
- DRAM is ‘near memory’
  - Used as a write-back cache
  - Managed by host memory controller
  - Within the same host memory controller, not across
  - Ratio of far/near memory (PMEM/DRAM) can vary
- Overall latency
  - Same as DRAM for cache hit
  - DC persistent memory + DRAM for cache miss
PMEM-aware software/application required
- Adds a new tier between DRAM and block storage (SSD/HDD)
- Industry open standard programming model and Intel PMDK

In-place persistence
- No paging, context switching, interrupts, nor kernel code executes

Byte addressable like memory
- Load/store access, no page caching
- Cache Coherent
- Ability to do DMA & RDMA

CPU CACHES
Minimum required power fail protected domain:
Memory subsystem

App Direct Mode - Persistent Memory

SW makes sure that data is flushed to durability domain using CLFLUSHOPT or CLWB
2. PERFORMANCE

Intel® Optane™ DC Persistent Memory for larger data, better performance/$, and new paradigms
INTEL® OPTANE™ DC Persistent Memory Latency

Latency vs. Load - P4800X vs. P4610 vs. Intel Optane DC Persistent Memory

(70%Read/30%Write Random, 4K) for SSD's, 256B for Intel Optane DC PMM

1000x lower latency

Intel DC P4610 NVMe SSD

Intel Optane DC SSD
P4800X

Read idle latency

Lower is better

Smaller granularity (vs. 4K)

Note 4K granularity gives about the same performance as 256B

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
Performance can vary based on:

- 64B random vs. 256B granularity
- Read/write mix
- Power level (programmable 12-18W, graph is 18W)

Intel® Optane™ DC Persistent Memory Latency
Ranges from 180ns to 340ns (vs. DRAM ~70ns)

Read idle latency
Ranges from 180ns to 340ns (vs. DRAM ~70ns)

Latency vs. Load - P4800X vs. P4610 vs. Intel Optane DC Persistent Memory
(70% Read/30% Write Random, 4kB) for SSD's, 64B and 256B for Intel Optane DC PMM

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
MEMORY MODE TRANSACTION FLOW

- 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
  - Other factors:
    - #reads > #writes
    - Config vs. Workload size

MEMORY MODE PERFORMANCE VS. LOCALITY & LOAD

- Synthetic traffic generator represents different types of workloads
- Vary size of buffers to emulate more or less locality
  - Very large data size (much larger than DRAM cache) causes higher miss rate

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
<table>
<thead>
<tr>
<th>System Platform</th>
<th>Neon city</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>CLX-B0</td>
</tr>
<tr>
<td>CPU per Node</td>
<td>28core/socket, 1 socket, 2 threads per core</td>
</tr>
<tr>
<td>Memory</td>
<td>6x 16GB DDR + 6x 128GB AEP QS</td>
</tr>
<tr>
<td>SUT OS</td>
<td>Fedora 4.20.6-200.fc29.x86_64</td>
</tr>
<tr>
<td>BKC</td>
<td>WW08</td>
</tr>
<tr>
<td>BIOS</td>
<td>PLYXCRB1.86B.0576.D20.1902150028 (mbf50656_0400001c)</td>
</tr>
<tr>
<td>FW</td>
<td>01.00.00.5355</td>
</tr>
<tr>
<td>Security</td>
<td>Variants 1,2, &amp; 3 Patched</td>
</tr>
<tr>
<td>Test Date</td>
<td>4/5/2019</td>
</tr>
</tbody>
</table>

MLC parameters: --loaded_latency --d<varies> -t200

<table>
<thead>
<tr>
<th>Buffer size (GB) per thread</th>
<th>2-2-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miss rate (%)</td>
<td></td>
</tr>
<tr>
<td>~0</td>
<td>0.1</td>
</tr>
<tr>
<td>~10</td>
<td>1.0</td>
</tr>
<tr>
<td>~25</td>
<td>4.5</td>
</tr>
<tr>
<td>~40</td>
<td>9.0</td>
</tr>
</tbody>
</table>
Enable More Redis VM Instances with Sub-MS SLA

1. One Redis Memtier instance per VM
2. Max throughput scenario, will scale better at lower operating point

**Redis VM's Meeting SLA**

- 2 VMs per core
- 1 VM per core

**VM size**

<table>
<thead>
<tr>
<th>Capacity (GB)</th>
<th>Summary</th>
<th>VM’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>45GB</td>
<td>111%, meets SLA</td>
<td>14-&gt;20</td>
</tr>
<tr>
<td>90GB</td>
<td>147%, meets SLA</td>
<td>7-&gt;10</td>
</tr>
</tbody>
</table>

**Throughput vs. DRAM**

Throughput: Higher is better, Latency: lower is better (must be 1ms or less)

For more complete information about performance and benchmark results, visit [www.intel.com/benchmarks](http://www.intel.com/benchmarks).
## REDIS CONFIGURATION

<table>
<thead>
<tr>
<th>Configuration 1 - 1LM</th>
<th>Configuration 2 – Memory Mode (2LM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test by</strong></td>
<td>Intel</td>
</tr>
<tr>
<td><strong>Test date</strong></td>
<td>02/22/2019</td>
</tr>
<tr>
<td><strong>Platform</strong></td>
<td>Neoncity</td>
</tr>
<tr>
<td><strong># Nodes</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong># Sockets</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>Intel® Xeon® Platinum 8276, 165W</td>
</tr>
<tr>
<td><strong>Cores/socket, Threads/socket</strong></td>
<td>28/56</td>
</tr>
<tr>
<td><strong>HT</strong></td>
<td>On</td>
</tr>
<tr>
<td><strong>BIOS version</strong></td>
<td>PLYXCRB1.86B.0573.D10.1901300453</td>
</tr>
<tr>
<td><strong>BKC version – E.g. ww47</strong></td>
<td>WW06</td>
</tr>
<tr>
<td><strong>AEP FW version – E.g. 5336</strong></td>
<td>5346 (QS AEP)</td>
</tr>
<tr>
<td><strong>System DDR Mem Config: slots / cap / run-speed</strong></td>
<td>12 slots / 32GB / 2666</td>
</tr>
<tr>
<td><strong>System DCPMM Config: slots / cap / run-speed</strong></td>
<td>12 slots / 32GB / 2666</td>
</tr>
<tr>
<td><strong>Total Memory/Node (DDR, DCPMM)</strong></td>
<td>768, 0</td>
</tr>
<tr>
<td><strong>NICs</strong></td>
<td>2x40GB</td>
</tr>
<tr>
<td><strong>OS</strong></td>
<td>Fedora-27</td>
</tr>
<tr>
<td><strong>Kernel</strong></td>
<td>4.20.4-200.fc29.x86_64</td>
</tr>
<tr>
<td><strong>AEP mode: ex. MM or AD-volatile (replace DDR) or AD-persistent (replace NVME)</strong></td>
<td>1LM</td>
</tr>
<tr>
<td><strong>Workload &amp; version</strong></td>
<td>Redis 4.0.11</td>
</tr>
<tr>
<td><strong>Other SW (Frameworks, Topologies...)</strong></td>
<td>memtier_benchmark-1.2.12 (80/20 read/write); 1K record size</td>
</tr>
<tr>
<td><strong>VMs (Type, vcpu/VM, VM OS)</strong></td>
<td>KVM, 1/VM, centos-7.0</td>
</tr>
</tbody>
</table>
APP DIRECT MODE TRANSACTION FLOW

- Traditional read to page fault (disk):
  1. Software
  2. 4K transfer from disk
  3. Request returned

- App Direct access memory directly
  - Avoids software and 4K transfer overhead
  - Cores can still access DRAM normally, even on same channel
• Reduce TCO by moving large portion of data from DRAM to Intel® Optane™ DC persistent memory
• Optimize performance by using the values stored in persistent memory instead of creating a separate copy of the log in SSD (only pointer written to log)
  • Direct access vs. disk protocol

Moving Value to App Direct reduces DRAM and optimizes logging by 2.27x
(Open Source Redis Set, AOF=always update, 1K datasize, 28 instances)

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
**SPARK SQL OAP CACHE**

- Intel® Optane™ DC persistent memory as cache
- More affordable than similar capacity DRAM
- Significantly lower overhead for I/O intensive workloads

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Query time</th>
</tr>
</thead>
<tbody>
<tr>
<td>768GB DRAM</td>
<td>1417s</td>
</tr>
<tr>
<td>192GB DRAM 1TB App Direct</td>
<td>171s</td>
</tr>
</tbody>
</table>

8X improvement in Apache Spark* sql IO intensive queries for Analytics

3TB scale factor

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
SUMMARY

- Intel® Optane™ DC Persistent Memory closes the DDR memory and storage gap
- Architected for persistence
- Provided large capacity scales workloads to new heights
- Offered a new way to manage data flows with unprecedented integration into system and platform
- Optimized for performance and orders of magnitude faster than NAND
  - Memory mode for large affordable volatile memory
  - App Direct mode for persistent memory