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.
Intel® Optane™
DC Persistent Memory Architecture

A Breakthrough with a New Interface Protocol, Memory Controller, Media, and Software Stack
Memory-Storage Gap

- **Memory Sub-System**
  - **DRAM (HOT TIER)**: pico-secs, <100 nanoseconds
  - **SSD (WARM TIER)**: nano-secs, <100 milliseconds
  - **HDD / TAPE (COLD TIER)**: <100 milliseconds

**Access Distribution**
- **Hot data**: more often
- **Cooler data**: less often

- **Data Access Frequency**
- **Intel® 3D Nand SSD**: 10s GB, <100 nanoseconds
- **Network Storage**: 10s TB, <100 microseconds
- **10s TB**: <100 milliseconds
Close Memory – Storage Gap

Optimize performance given cost and power budget

Move Data Closer to Compute
Maintain Persistency

Intel® 3D Nand SSD
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

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

Scalable
Memory layers can be stacked
in a 3D manner

Breakthrough
Material Advances
Compatible switch and memory cell
materials

High Performance
Cell and array architecture
that can switch fast

First Generation Capacities:
128 GB
256 GB
512 GB
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

 DDR4 Slot on Host CPU

 Interface to Host CPU

 DCPMM Memory Interface

 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

 ECC/Scrambler

 Error Handling Logic

 Media Queue

 Optane™ Media Devices

 Caps for Flashes

 Media Channel

 Optane™ Media Devices

 Media Channel
Intel® Optane DC Persistent Memory SW
Enabling Stack

Management UI
Management Library
Application
Standard Raw Device Access
Standard File API
Generic NVDIMM Driver
File System
Pmem-Aware File System
MMU Mappings
Load/Store
"DAX"

Persistent Memory

file
memory
Asynchronous DRAM Refresh (ADR) - Power Fail Protected Feature

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
App Direct Mode – Persistent Memory

- 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
2. PERFORMANCE

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

**1000x lower latency**

Note 4K granularity gives about same performance as 256B

**Read idle latency**

Lower is better

Smaller granularity (vs. 4K)

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)

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
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.
## Memory Mode Performance/Load/Locality

<table>
<thead>
<tr>
<th></th>
<th>2-2-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Platform</td>
<td>Neon city</td>
</tr>
<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>Miss rate (%)</th>
<th>R</th>
<th>W2</th>
</tr>
</thead>
<tbody>
<tr>
<td>~0</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>~10</td>
<td>1.0</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>~25</td>
<td>4.5</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>~40</td>
<td>9.0</td>
<td>4.5</td>
<td></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

<table>
<thead>
<tr>
<th>VM size</th>
<th>DRAM baseline</th>
<th>MM capacity</th>
<th>Throughput vs. DRAM</th>
<th>Summary</th>
<th>VM's</th>
</tr>
</thead>
<tbody>
<tr>
<td>45GB</td>
<td>768GB</td>
<td>1TB</td>
<td>111%, meets SLA</td>
<td>42% more VM's @lower cost</td>
<td>14-&gt;20</td>
</tr>
<tr>
<td>90GB</td>
<td>768GB</td>
<td>1TB</td>
<td>147%, meets SLA</td>
<td>42% more VM's @lower cost</td>
<td>7-&gt;10</td>
</tr>
</tbody>
</table>

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
## Configuration 1 - 1LM

<table>
<thead>
<tr>
<th>Test by</th>
<th>Intel</th>
<th>Intel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test date</td>
<td>02/22/2019</td>
<td>02/22/2019</td>
</tr>
<tr>
<td>Platform</td>
<td>Neoncity</td>
<td>Neoncity</td>
</tr>
<tr>
<td># Nodes</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td># Sockets</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel® Xeon® Platinum 8276, 165W</td>
<td>Intel® Xeon® Platinum 8276, 165W</td>
</tr>
<tr>
<td>Cores/socket, Threads/socket</td>
<td>28/56</td>
<td>28/56</td>
</tr>
<tr>
<td>HT</td>
<td>On</td>
<td>On</td>
</tr>
<tr>
<td>BIOS version</td>
<td>PLYXCRB1.86B.0573.D10.1901300453</td>
<td>PLYXCRB1.86B.0573.D10.1901300453</td>
</tr>
<tr>
<td>AEP FW version – E.g. 5336</td>
<td>5346 (QS AEP)</td>
<td>5346 (QS AEP)</td>
</tr>
<tr>
<td>System DDR Mem Config: slots / cap / run-speed</td>
<td>12 slots / 32GB / 2666</td>
<td>12 slots / 16 GB / 2666</td>
</tr>
<tr>
<td>System DCPMM Config: slots / cap / run-speed</td>
<td>12 slots / 32GB / 2666</td>
<td>12 slots /128GB,256GB,512GB/ 2666</td>
</tr>
<tr>
<td>Total Memory/Node (DDR, DCPMM)</td>
<td>768, 0</td>
<td>192, 1TB,1.5TB,3TB, 6TB</td>
</tr>
<tr>
<td>NICs</td>
<td>2x40GB</td>
<td>2x40GB</td>
</tr>
<tr>
<td>OS</td>
<td>Fedora-27</td>
<td>Fedora-27</td>
</tr>
<tr>
<td>Kernel</td>
<td>4.20.4-200.fc29.x86_64</td>
<td>4.20.4-200.fc29.x86_64</td>
</tr>
<tr>
<td>AEP mode: ex. MM or AD-volatile (replace DDR) or AD-persistent (replace NVME)</td>
<td>1LM</td>
<td>Memory Mode (2LM)</td>
</tr>
<tr>
<td>Workload &amp; version</td>
<td>Redis 4.0.11</td>
<td>Redis 4.0.11</td>
</tr>
<tr>
<td>Other SW (Frameworks, Topologies…)</td>
<td>memtier_benchmark-1.2.12 (80/20 read/write) ; 1K record size</td>
<td>memtier_benchmark-1.2.12 (80/20 read/write) ; 1K record size</td>
</tr>
<tr>
<td>VMs (Type, vcpu/VM, VM OS)</td>
<td>KVM, 1/VM, centos-7.0</td>
<td>KVM, 1/VM, centos-7.0</td>
</tr>
</tbody>
</table>

## Configuration 2 – Memory Mode (2LM)
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
Redis Example (with Software Change)

- 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

![Diagram showing the process of moving data from DRAM to PMEM](Image)

**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

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

- Optimized for performance and orders of magnitude faster than NAND
  - Memory mode for large affordable volatile memory
  - App Direct mode for persistent memory