Breakthrough Data-Centric Computing with a New Memory Tier

Alper Ilkbahar  
Vice President & General Manager  
Data Center Group, Intel Corporation
Big and Affordable Memory
High Performance Storage
Direct Load/Store Access
Native Persistence

128, 256, 512GB Modules
DDR4 Pin Compatible
Hardware Encryption
High Reliability

LAUNCHED APRIL 2ND
NOW SHIPPING IN VOLUME
Introducing a New Tier
Introducing a New Tier
Introducing a New Tier

MEMORY TIER

PERSISTENT MEMORY

STORAGE TIER

2ND GEN INTEL® XEON® SCALABLE PROCESSOR

APPLICATION

PM AWARE
FILE SYSTEM

FILE SYSTEM

DRIVER

KERNEL SPACE

USER SPACE

Flash Memory Summit

intel
Persistent Memory as Storage
Persistent Memory: Low Latency

More Bandwidth:
Up to 3.7X read/write bandwidth vs NVMe SSDs, with one module; more with multiple modules

Lower Latency:
Orders of magnitude lower latency than NVMe SSDs
- 1000X lower latency than NAND NVMe SSD at 1GB/s

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 33 for configurations.
MORE users per system 1

MORE update transactions (ops/sec) 1

Vs. comparable server system with DRAM and NAND NVMe drives when using Apache* Cassandra-4.0

1 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.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 35-37 for configurations.
Data Replication with Persistent Memory

Traditional Data Replication

Multiple data hops:
1. Processors move data to remote memory
2. Remote processor moves to SSD

Software overhead for network and storage drivers

Data Replication with Persistent Memory

Single hop:
1. RDMA moves data from local to remote persistent memory with minimal software overhead
Data Replication with Persistent Memory

Average Latency Components
(4KB Sequential Writes)

- Replication to Remote Intel Optane DC Persistent Memory
- DRBD Replication to NVMe SSD

Average latency in microseconds (lower is better)

Network / Fabric
NVM Media
Software

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.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 38 for configurations.
Customer Example: Oracle Exadata

Preview: Exadata – Persistent Memory Accelerator for OLTP

- Exadata Storage Servers will add Persistent Memory Accelerator in front of Flash memory
- RDMA bypasses the software stack, giving 10X faster access latency to remote Persistent Memory
- Persistent Memory mirrored across storage servers for fault-tolerance
- Persistent Memory used as a shared cache effectively increases its capacity 10x vs using it directly as expensive storage
- Log Writes will use RDMA to achieve super fast commits

System Configuration: 2x Intel® Xeon Cascade Lake 24 cores with 768G RAM; 2x Intel® Xeon Cascade Lake 16 cores with 192G RAM + 1.5TB DCPMM. Oracle Linux 7 UEKS U2 4.14.35-1902

10X lower latency

Slide courtesy of Oracle

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
Persistent Memory as Main Memory
Cost vs Performance Framework

- **Storage Costs ($/GB)**
  - Main Memory (DRAM)
  - Secondary Storage (NVMe NAND)

- **Execution Costs ($/ops)**

- **Total Burden**
Cost vs Performance Framework

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.
Cost vs Performance Framework

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.
Cost vs Performance Framework

1 Assumes 10% of data in the DRAM with 90% cache hit rate.

Based on Intel internal testing.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.
Cost vs Performance Framework

Efficient caching architectures try to follow the lowest line along.

In-memory architectures follow the red line.

Efficient caching architectures try to follow the lowest line along.

1 Assumes 10% of data in the DRAM with 90% cache hit rate.

Based on Intel internal testing.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.
Cost vs Performance Framework

- Secondary Storage (NVMe NAND)
- Main Memory (DRAM)
- Persistent Memory (Optane DCPM)

1 Assumes 10% of data in the DRAM with 90% cache hit rate.

Based on Intel internal testing.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.
Cost vs Performance Framework

RELATIVE COST

RELATIVE PERFORMANCE (OPS/SEC)

1 Assumes 10% of data in the DRAM with 90% cache hit rate.

Based on Intel internal testing.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.
SAP HANA and Persistent Memory

- SAP HANA controls what is placed in Persistent Memory and what remains in DRAM.
  - Volatile data structures remain in DRAM.
  - Column Store Main moves to Persistent Memory:
    - More than 95% of data in most HANA systems.
    - Loading of tables into memory at startup becomes obsolete.
    - Lower TCO, larger capacity.
  - No changes to the persistence.

- More than 95% of data in most HANA systems in persistent memory.

- SAP BW for SAP HANA query/hour results@6.5B
  - DRAM: 4,062
  - Persistent Memory: 3,825

- Performance delta vs all-DRAM system: 5.9%

- Faster start times for less downtime:
  - Traditional System: 50 mins
  - Persistent Memory: 4 mins, 12.5x improvement

- Increased memory capacity while reducing TCO:
  - Memory capacity per CPU: >3 TB

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 38 for configurations.
Vibrant Software Ecosystem

Support announced by ISV or open source enabled by Intel
Breakthrough Data-Centric Computing with a New Memory Tier
Intel Data Management Platform (DMP)

**Compute Nodes with up to 6TB of Intel Optane DC PM**
- All random I/O serviced by Intel Optane DC persistent memory
- Minimal DRAM for hot indexes
- Page and block caches turned OFF
- Checkpoints from persistent memory into storage

**Storage Nodes with up to 1PB Intel Ruler QLC SSDs**
- Disaggregated with NVMe-oF and RDMA
- Sequential accesses through periodic checkpoints and snapshot images
- Integrated cloud storage (S3)

**Infrastructure Nodes**
- Interconnected with 100Gb ethernet
- Kubernetes orchestration

For more information: www.usenix.org/conference/vault19/presentation/cohen
Video shown – see separate link
Cost vs Performance Framework for DMP

Relative Cost

Low-Cost Storage Tier

Relative Performance (ops/sec)

Performance Tier: Persistent Memory

Cache Tier

Based on Intel internal testing.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.
Join the Persistent Memory Revolution
Join the Persistent Memory Revolution
Join the Persistent Memory Revolution
Notices & Disclaimers

- Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.
- No product or component can be absolutely secure.
- 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.
- 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
- Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo
- Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microarchitectures. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.
- 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.
- Intel, the Intel logo, Intel Xeon, and Optane DC Persistent Memory are trademarks of Intel Corporation in the U.S. and/or other countries.
- *Other names and brands may be claimed as property of others.
- © 2019 Intel Corporation.
Back-up
Systems & Configurations
Data Management Platform Configurations

Node Information:
- os_release: Fedora 30 Server
- kernel: 5.1.18-300.fc30.x86_64
- cpu_type: Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz
- cpus_total: 36
- dimm_count: 6
- dimm_size: 32GB
- memory: 192GB

D. Lomet: Cost/Performance in Modern Data Stores, How Data Caching Systems Succeed, DaMoN, 2018
Persistent Memory: Low Latency “P4800 and P4610”

SSDs

**1000X Claim:** 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® 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 or component can be absolutely secure.

**3.7X Claim:** Tested by Intel on single DIMM configuration; Test date 02/20/2019. Platform Neon City; Chipset LBG B1; CPU CLX B0 28 Core (8276), 1S; DDR speed 2666 MT/s; Intel Optane DC PMEM 256GB, 18W; Memory configuration 1 channel, 32GB DDR4 (per socket), 128GB Intel Optane DC PMEM (per socket); Intel Optane DC PMEM FW 5336; BIOS 573.D10; BKC version WW08 BKC, Linux OS 4.20.4-200.fc29; Spectre/Meltdown patched (1,2,3,3a); Performance Tuning QoS Disabled IODC=5(AD)
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. 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 information go to www.intel.com/benchmarks. *Three 9s and five 9s availability assumes bi-weekly maintenance restarts.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>NVMe</th>
<th>DCPMM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test by</strong></td>
<td>Intel/Java Performance Team</td>
<td>Intel/Java Performance Team</td>
</tr>
<tr>
<td><strong>Test date</strong></td>
<td>22/02/2019</td>
<td>22/02/2019</td>
</tr>
<tr>
<td><strong>Platform</strong></td>
<td>S2600WFD</td>
<td>S2600WFD</td>
</tr>
<tr>
<td><strong># Nodes</strong></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong># Sockets</strong></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>8280L</td>
<td>8280L</td>
</tr>
<tr>
<td><strong>Cores/socket, Threads/socket</strong></td>
<td>28/56</td>
<td>28/56</td>
</tr>
<tr>
<td><strong>ucode</strong></td>
<td>0x4000013</td>
<td>0x4000013</td>
</tr>
<tr>
<td><strong>HT</strong></td>
<td>On</td>
<td>On</td>
</tr>
<tr>
<td><strong>Turbo</strong></td>
<td>On</td>
<td>On</td>
</tr>
<tr>
<td><strong>BIOS version</strong></td>
<td>SE5C620.86B.0D.01.0286.011120190816</td>
<td>SE5C620.86B.0D.01.0286.011120190816</td>
</tr>
<tr>
<td><strong>DCPMM BKC version</strong></td>
<td>NA</td>
<td>WW52 -2018</td>
</tr>
<tr>
<td><strong>DCPMM FW version</strong></td>
<td>NA</td>
<td>5318</td>
</tr>
<tr>
<td><strong>System DDR Mem Config: slots / cap / run-speed</strong></td>
<td>12 slots / 16GB / 2666</td>
<td>12 slots / 16GB / 2666</td>
</tr>
<tr>
<td><strong>System DCPMM Config: slots / cap / run-speed</strong></td>
<td>12 slots / 512GB</td>
<td>12 slots / 512GB</td>
</tr>
<tr>
<td><strong>Total Memory/Node (DDR, DCPMM)</strong></td>
<td>192GB, 0</td>
<td>192GB, 0</td>
</tr>
<tr>
<td><strong>Storage - boot</strong></td>
<td>1x Intel 800GB SSD OS Drive</td>
<td>1x Intel 800GB SSD OS Drive</td>
</tr>
<tr>
<td><strong>Storage - application drives</strong></td>
<td>4x P4610 1.61TB NVMe</td>
<td>12x512GB DCPMM</td>
</tr>
<tr>
<td><strong>NIC</strong></td>
<td>1x Intel X722</td>
<td>1x Intel X722</td>
</tr>
<tr>
<td><strong>OS</strong></td>
<td>Red Hat Enterprise Linux Server 7.6</td>
<td>Red Hat Enterprise Linux Server 7.6</td>
</tr>
<tr>
<td><strong>Kernel</strong></td>
<td>4.19.0 (64bit)</td>
<td>4.19.0 (64bit)</td>
</tr>
<tr>
<td><strong>Mitigation log attached</strong></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>DCPMM mode</strong></td>
<td>App Direct, Persistent Memory</td>
<td></td>
</tr>
<tr>
<td><strong>Run Method</strong></td>
<td>5 minute warm up post boot, then start performance recording</td>
<td>5 minute warm up post boot, then start performance recording</td>
</tr>
<tr>
<td><strong>Iterations and result choice</strong></td>
<td>3 iterations, median</td>
<td>3 iterations, median</td>
</tr>
<tr>
<td><strong>Dataset size</strong></td>
<td>Two 1.5 Billion Partitions (Insanity schema)</td>
<td>Two 1.5 Billion Partitions (Insanity schema)</td>
</tr>
<tr>
<td><strong>Workload &amp; version</strong></td>
<td>Read Only, Mix 80% Read/20% Updates, Updates only</td>
<td>Read Only, Mix 80% Read/20% Updates, Updates only</td>
</tr>
<tr>
<td><strong>Compiler</strong></td>
<td>ANT 1.9.4 compiler for Cassandra</td>
<td>ANT 1.9.4 compiler for Cassandra</td>
</tr>
<tr>
<td><strong>Libraries</strong></td>
<td>NA</td>
<td>PMDK 1.5, LLPL (latest as of 2/20/1019)</td>
</tr>
<tr>
<td><strong>Other SW (Frameworks, Topologies...)</strong></td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

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. 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 information go to www.intel.com/benchmarks.
## 2/3 Cassandra Configuration Settings

### Software

<table>
<thead>
<tr>
<th>Software</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cassandra</td>
<td>NVME uses 3.11.3 released version, DCPMM uses 4.0 trunk with persistent memory modifications: <a href="https://github.com/shyla226/cassandra/tree/13981_llp_engine">https://github.com/shyla226/cassandra/tree/13981_llp_engine</a></td>
</tr>
<tr>
<td>PMDK</td>
<td>1.5</td>
</tr>
<tr>
<td>LLPL</td>
<td><a href="https://github.com/pmem/llpl/">https://github.com/pmem/llpl/</a> pulled 2/20/19</td>
</tr>
<tr>
<td>Java</td>
<td>Java™ SE Runtime Environment 1.8.0_201 Java HotSpot™ 64-bit Server VM (build 25.201)</td>
</tr>
</tbody>
</table>

### Cassandra Settings

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software Version</td>
<td>Cassandra NVME uses 3.11.3 released version, DCPMM uses 4.0 trunk with persistent memory modifications: <a href="https://github.com/shyla226/cassandra/tree/13981_llp_engine">https://github.com/shyla226/cassandra/tree/13981_llp_engine</a></td>
</tr>
<tr>
<td>PMDK</td>
<td>1.5</td>
</tr>
<tr>
<td>LLPL</td>
<td><a href="https://github.com/pmem/llpl/">https://github.com/pmem/llpl/</a> pulled 2/20/19</td>
</tr>
<tr>
<td>Java</td>
<td>Java™ SE Runtime Environment 1.8.0_201 Java HotSpot™ 64-bit Server VM (build 25.201)</td>
</tr>
</tbody>
</table>

### Cassandra Settings

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yaml modifications</td>
<td>concurrent_read/concurrent_write 168/168 for DCPMM concurrent_read/concurrent_write 56/32 for NVME</td>
</tr>
<tr>
<td>Jvm.options (comment out CMS section in file)</td>
<td>-Xms64G -Xmx64G -Xmn48G for DCPMM, no read cache -Xms32G -Xmx32G -Xmn24G for NVME, more read cache -XX:+UseAdaptiveSizePolicy for both</td>
</tr>
<tr>
<td>Number of Cassandra Processes, DataBases, Clusters</td>
<td>2 independent Cassandra processes each with a database, each process running 1 node cluster configuration</td>
</tr>
<tr>
<td>Cassandra Database per Application</td>
<td>cqlstress-insanity-example.yaml schema, with 1.5 Billion partition per database(3.0 Billion Total)</td>
</tr>
<tr>
<td>Cassandra Application pinned to CPU</td>
<td>numactl -m 0 -C 0-27,56-83 for socket 0 numactl -m 1 -C 28-55,84-111 for socket 1</td>
</tr>
<tr>
<td>Cassandra-Stress Command to Populate Database</td>
<td>cassandra-stress user profile=${CASSANDRA_HOME}/tools/cqlstress-insanity-example.yaml ops{&quot;insert=1} n=1500000000 cl=ONE no-warmup -pop seq=1..1500000000 -mode native cql3 -node &lt;ip_addr&gt; -rate threads=&lt;variable&gt;</td>
</tr>
<tr>
<td>Cassandra-Stress Command to Read Database</td>
<td>cassandra-stress user profile=${CASSANDRA_HOME}/tools/cqlstress-insanity-example.yaml ops{&quot;simple=1} duration=30m cl=ONE -pop dist=UNIFORM(1..1500000000) -mode native cql3 -node &lt;ip_addr&gt; -rate threads=&lt;variable&gt;</td>
</tr>
</tbody>
</table>
### Methodology:
- Adjust the Cassandra-stress load (number of client threads) to get maximum throughput where the 99th latency is less than 20ms.
- This method has been accepted by our partners (Apple, Netflix and others).

### Two different way of classifying the speed up:
- Increased throughput speedup, for example maximum seen for the read workload of 8.13 times more throughput with DCPMM vs NVMe.
- Increased number of supported clients threads, for example maximum seen for the update workload of 9.09 times more client threads supported for similar SLA with DCPMM vs NVMe.

<table>
<thead>
<tr>
<th>Workload</th>
<th>NVMe Throughput (op/sec)</th>
<th>NVMe 99th latency (ms)</th>
<th>NVMe client load (# threads)</th>
<th>DCPMM Throughput (op/sec)</th>
<th>DCPMM 99th latency (ms)</th>
<th>DCPMM client load (# threads)</th>
<th>Throughput Speedup with DCPMM</th>
<th>Client Load Increase with DCPMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read</td>
<td>66,018</td>
<td>17.6</td>
<td>800</td>
<td>537,121</td>
<td>19.0</td>
<td>5600</td>
<td>8.13X</td>
<td>7.00X</td>
</tr>
<tr>
<td>Mix (80/20)</td>
<td>76,747</td>
<td>18.5</td>
<td>800</td>
<td>491,831</td>
<td>16.6</td>
<td>4400</td>
<td>6.40X</td>
<td>5.50X</td>
</tr>
<tr>
<td>Update</td>
<td>54,013</td>
<td>18.4</td>
<td>440</td>
<td>390,935</td>
<td>16.1</td>
<td>4000</td>
<td>7.23X</td>
<td>9.09X</td>
</tr>
</tbody>
</table>
Data Replication with Persistent Memory (slide 10):


SAP HANA (slide 20):

**5.9% SAP HANA** claim based on source: SAP* BW for SAP HANA* @ 6.5B initial records - https://www.sap.com/dmc/exp/2018-benchmark-directory/#/bwh. Baseline: 4s Intel® Xeon® Platinum 8280L with DRAM, Certification #2019022, Benchmark score: Runtime of Data Load/Trans (18821 secs), Query Executions per Hour (4062), Total Runtime of Complex Query (107 secs). New config: 4s Intel® Xeon® Platinum 8280L with Intel® Optane™ DC persistent memory;, Certification #2019020, Benchmark score: Runtime of Data Load/Trans (21533 secs), Query Executions per Hour (3825), Total Runtime of Complex Query (135 secs).

**12.5X and 95% Claim:** https://blogs.sap.com/2018/12/03/sap-hana-persistent-memory/