STRATEGIES ON STORAGE

Author: Frank Hady, PhD, Fellow, Chief Systems Architect, NVM Solutions Group, Intel Corporation

Presenter: Piotr Wysocki, Software Architect, NVM Solutions Group, Intel Corporation

April 2019
PERFORMANCE RESULTS ARE BASED ON TESTING AS OF SPECIFIED DATES AND MAY NOT REFLECT ALL PUBLICLY AVAILABLE SECURITY UPDATES. SEE CONFIGURATION DISCLOSURE FOR DETAILS. NO PRODUCT CAN BE ABSOLUTELY SECURE

THIS PRESENTATION INCLUDES FORWARD-LOOKING STATEMENTS RELATING TO INTEL. ALL STATEMENTS THAT ARE NOT HISTORICAL FACTS ARE SUBJECT TO A NUMBER OF RISKS AND UNCERTAINTIES, AND ACTUAL RESULTS MAY DIFFER MATERIALLY. PLEASE REFER TO INTEL'S MOST RECENT EARNINGS RELEASE, 10-Q AND 10-K FILINGS FOR THE RISK FACTORS THAT COULD CAUSE ACTUAL RESULTS TO DIFFER.
NOTICES AND DISCLOSURES

[Performance results are based on testing as of [INSERT DATE] and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

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 about performance and benchmark results, visit http://www.intel.com/benchmarks.

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

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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel, the Intel logo, Intel Core, Intel Optane, Intel 3D XPoint, Intel Performance Maximizer, and Thunderbolt are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

© Intel Corporation 2019.

*Other names and brands may be claimed as the property of others.
“Ideally one would desire an indefinitely large memory capacity such that any particular ... word would be immediately available. ... It does not seem possible physically to achieve such a capacity. We are therefore forced to recognize the possibility of constructing a hierarchy of memories, each of which has greater capacity than the preceding but which is less quickly accessible.”

Preliminary Discussion of the Logical Design of an Electronic Computing Instrument
Arthur Burks, Herman Goldstine and John von Neumann, 1946
MEMORY AND STORAGE HIERARCHY

- **1 DIMM:**
  - 10s GB
  - <100ns

- **1 SSD:**
  - 10s TB
  - <100µsecs

- **1 HDD:**
  - 10s TB
  - <10 msecs

1000X

- **DRAM HOT TIER**

100X

- **3D NAND SSD WARM TIER**

10X

- **HDD / TAPE COLD TIER**

90/10 LOCALITY “RULE”
10X CAPACITY
1/10TH PERFORMANCE

- 10x performance
DRAM SCALING SLOWED, NAND SCALING KEPT PACE

Source: Intel: Presented at IEEE S3S (SOI-3D-SUBTHRESHOLD) in 2017
PERFORMANCE: TECHNOLOGY SCALING

CONCLUSIONS: EVOLUTIONARY IMPROVEMENTS DELIVER IMPROVED BANDWIDTH
ONLY NEW TECHNOLOGIES CAN DELIVER IMPROVED LATENCY


NAND SSD data points added by Intel Based on product brief specifications for Intel NAND SSDs available at www.intel.com
MEMORY AND STORAGE HIERARCHY GAPS

SOLUTION MUST MEET:
• CAPACITY
• SYSTEM PERFORMANCE
• SYSTEM FIT

MEMORY

DRAM HOT TIER
CAPACITY GAP
10s GB
<100ns

STORAGE

STORAGE PERFORMANCE GAP
3D NAND SSD WARM TIER
10s TB
<100µsecs

COST PERFORMANCE GAP
HDD / TAPE COLD TIER
10s TB
<10 msecs

© ISOTDAQ 2019. All Rights Reserved.
Desirable Attributes: Non-volatile, Low Cost, High Performance

- Memory in atomistic state, not electrostatic  
  → Non-Volatile and Scalable
- Simple scalable structure + 3D technology  
  → Large Memory Capacity
- Fast switching materials + local low resistance metal interconnect  
  → Immediately Available
- Individual Cell Access  
  → Word Access
INTEL® OPTANE™ TECHNOLOGY: BUILDING BLOCKS

INTEL MEMORY AND STORAGE CONTROLLERS

INTEL INTERCONNECT IP

INTEL® SOFTWARE

CPU

LIBRARIES & DRIVERS

OPTIMIZED INTERFACE

PARALLEL ACCESS HARDWARE ONLY RD/WR
NEW MEDIA MANAGEMENT

LOW LATENCY WRITE-IN-PLACE

PLATFORM LEVEL INNOVATION ENABLES SYSTEM FIT

© ISOTDAQ 2019. All Rights Reserved.
Higher is better  Lower is better

Latency vs. Load: NAND SSD vs. Intel® Optane™ DC SSD

(Intel® DC P4610 3.2TB vs. Intel® Optane™ SSD DC P4800X 375GB)

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 www.intel.com/benchmarks.

Source – 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 P4600 1.6TB. 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.

© ISOTDAQ 2019. All Rights Reserved.
Latency Improvement Requires System Innovation

Latency vs. Load: NAND SSD vs. Intel® Optane™ SSD

(Intel® DC P4610 3.2TB vs. Intel® Optane™ SSD DC P4800x 375GB)

<table>
<thead>
<tr>
<th>Storage</th>
<th>Idle Avg. is About</th>
<th>Optane SSD latency balanced between SSD and System</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAND SSD latency</td>
<td>80µs FOR 4KB</td>
<td>Hardware Latency</td>
</tr>
<tr>
<td>Intel® Optane™ SSD latency</td>
<td>10µs FOR 4KB</td>
<td>Software Latency</td>
</tr>
</tbody>
</table>

1Source: Inte­l­test­ed. Aver­age read la­te­ncy mea­sured at quies­cent de­pth 1 dur­ing 4K ran­dom write work­load. Mea­sured us­ing RD 3.1. Com­mon Con­fig­ura­tion — Intel 2U Ser­ver Sys­tem, OS CentOS 7.5, kernel: 4.17-1.62.4-64, CPU 2 x Intel® Xeon® 6154 Gold @ 3.0GHz (18 cores), RAM 256GB, DOM @ 2666MHz. Con­fig­ur­a­tion — Intel® Optane™ SSD DC P4800X 375GB and Intel SSD DC P4600 1.6TB. Latency — Aver­age read la­te­ncy mea­sured at QD1 dur­ing 4K Random Write opera­tions using RD 3.1. Intel Microcode: L0; System BIOS: 00.01.0013; ME Firmware: 0400042594; BMC Firmware: 143.16179659; RSCDR: 1.43. SSD tested were commercially available at time of test. The benchmark results may need to be re­vised as additional testing is conducted. Performance results are based on testing as of July 24, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. This product can be evaluated as a storage solution. Software and workloads used in performance tests may have been configured for performance only. Software may be modified for performance on Intel platforms. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operating systems, and functions. Any change to any of these 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 www.intel.com/benchmarks.

© ISOTDAQ 2019. All Rights Reserved.
INTEL® OPTANE™ SSDS PUT STORAGE BACK IN THROUGHPUT/LATENCY BALANCE


NAND and Optane SSD data points added by Intel Based on product brief specifications for Intel NAND and Optane SSDs available at www.intel.com
MEMORY AND STORAGE HIERARCHY

MEMORY

STORAGE

IMPROVING SSD PERFORMANCE

10s GB <100ns

1 Intel® Optane™ SSD: 1sTB <10µsecs

3D NAND SSD WARM TIER

HDD / TAPE COLD TIER

STORAGE PERFORMANCE GAP

DRAM HOT TIER

CAPACITY GAP

© ISOTDAQ 2019. All Rights Reserved.
**LOW LATENCY SOFTWARE PATH**

**STORAGE**
- Read(fileptr,offset) /* OS call */
- Write(fileptr,offset) /* OS call */

**PERSISTENT MEMORY**
- ld(address) /* CPU opcode */
- st(address) /* CPU opcode */

4-10µs for Linux

Intel® Optane™ SSD

Intel® Optane™ DC Persistent Memory

---


© ISOTDAQ 2019. All Rights Reserved.
PERSISTENT MEMORY PLATFORM SUPPORT

Direct Load/Store Access
Native Persistence
128, 256, 512GB
DDR4 Pin Compatible

- BIOS
- OPERATING SYSTEM
- SNIA PM PROGRAMMING MODEL
- APPLICATION
LOW LATENCY SYSTEM ACCESS TO PERSISTENT MEMORY

IDLE AVERAGE RANDOM READ LATENCY

- Storage with NAND SSD
- Storage with Intel® Optane™ SSD
- Memory subsystem with Intel® Optane™ DC Persistent Memory

Storage Idle AVG. is about 10µs for 4KB
Memory subsystem Idle AVG. is about ~100ns to ~350ns for 64B

Source: Intel-tested: Average read latency measured at queue depth 1 during 4k random write workload. Measured using FIO 3.1. comparing Intel Reference platform with Optane™ SSD DC P4800X 375GB and Intel® SSD DC P4600 1.6TB compared to SSDs commercially available as of July 1, 2018. Performance results are based on testing as of July 24, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

App Direct Mode, NeonCity, LBG B1 chipset, CLX B0 38 Core (QDF QQY2), Memory Conf 192GB DDR4 (per socket) 2666 MT/s, Optane DCPMM 128GB, BIOS 561.090, BKC version WW48.5 BKC, Linux OS 4.18.8-100.fc27, Spectre/Meltdown Patched 1,2,3, 3a}

© ISOTDAQ 2019. All Rights Reserved.
App Direct Mode provides the persistent memory programming model
- Reported to OS by ACPI
- Linux and Windows expose via "DAX" file systems

Several use cases supported by OS & PMDK APIs
- Persistent memory, non-paged (no DRAM footprint when accessed)
- Volatile App Direct, an explicit pool of volatile memory
- Storage over App Direct, a very fast SSD built on persistent memory
Developer placed data structures

“SAP HANA knows which data structures benefit most from persistent memory. SAP HANA automatically detects persistent memory hardware and adjusts itself by automatically placing these data structures on persistent memory, while all others remain in DRAM”

- Column Store Main in Persistent Memory
  - 90% of the data footprint
  - Nonvolatile – no initial load time
- High perf, volatile in DRAM
- SSDs still used for row store, column delta, replication, backups...

Source: “SAP HANA & Persistent Memory”
- Andreas Schuster

© ISOTDAQ 2019. All Rights Reserved.
Memory Mode provides familiar volatile memory programming model
- Additional layer of caching: DRAM as WB cache
- Hardware managed, software sees very high capacity memory (6 TB)

Range of use cases supported
- No software change – big memory
- Applications/Algorithms changes for new hierarchy/capacity
MEMORY MODE USAGE EXAMPLE

VMware vSphere* using memory mode:

“When used in memory mode, the new Intel memory technology can greatly increase the memory capacity available to software in a platform when compared with the capacity of DRAM. This increase in capacity requires no changes to your existing software, operating systems, or virtual machines.”

- Developer allocates VM memory images in “memory”
- Platform memory controller caches active VM data in DRAM for use

Source: “Extending Memory Capacity with VMware vSphere and Upcoming Intel Optane Memory Technology” – Rich Brunner

*Performance results have been estimated based on SAS internal tests as of 11/05/2018 using future version of VMware vSphere, SAS Viya® 400GB Gradient Boosting Models running Linux with Intel® Optane™ DC persistent memory vs. DRAM-based server and may not reflect all publicly available security updates. As measured by VMware on system listed as 2-CPU socket server, Intel® Cascade lake, future version of VMware vSphere, 6TB Intel® Optane™ DC Persistent Memory in Memory Mode, versus 2-CPU socket server, Intel® Cascade lake, future version of VMware vSphere, 1.5TB DDR4 DRAM 3x 3.6 TB SSD. Performance results are based on testing as of [INSERT DATE] and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

© ISOTDAQ 2019. All Rights Reserved.
COMPLETE IN PERFORMANCE, CAPACITY, FIT

MEMORY

PERSISTENT MEMORY

STORAGE

DRAM HOT TIER

10s GB <100ns

100s GB <1usec

3D NAND SSD WARM TIER

STORAGE PERFORMANCE GAP

HDD / TAPE COLD TIER

Growing Memory Capacity

© ISOTDAQ 2019. All Rights Reserved.
NAND TECHNOLOGY ADVANCEMENT

© ISOTDAQ 2019. All Rights Reserved.
THE FUTURE OF DATA CENTER STORAGE & MEMORY

DATA/METADATA CACHE IN PM

w/INTEL® OPTANE™ DC PERSISTENT MEMORY

DATA STORAGE IN 1PB IN 1U

w/INTEL® 3D NAND SSDs

© ISOTDAQ 2019. All Rights Reserved.
COMPLETE IN PERFORMANCE, CAPACITY, FIT

MEMORY

PERSISTENT MEMORY

GROWING MEMORY CAPACITY

STORAGE

IMPROVING SSD PERFORMANCE

DELIVERING EFFICIENT STORAGE

INTEL® QLC 3D NAND SSD

HDD / TAPE COLD TIER

DRAM HOT TIER

10s GB <100ns

100s GB <1µsec

1s TB <10µsecs

10s TB <100µsecs

10s TB <10 msecs