Converging Memory and Storage

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

- 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](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

© 2019 Storage Networking Industry Association. All Rights Reserved.
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

10s GB
<100ns

10s TB
<100µsecs

10s TB
<10 msecs

MEMORY

DRAM HOT TIER
CAPACITY GAP

STORAGE PERFORMANCE GAP

3D NAND SSD WARM TIER

COST PERFORMANCE GAP

HDD / TAPE COLD TIER

STORAGE

Solution Must Meet:
- Capacity
- System Performance
- System Fit
A CONVERGENT MEMORY

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

PLATFORM LEVEL INNOVATION ENABLES SYSTEM FIT

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

MEDIA MEDIA MEDIA MEDIA MEDIA
Intel® Optane™ SSD

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

Lower is better

Higher is better

Source – Intel-tested: Measured using FIO 3.1. Common Configuration - Intel 2U Server System, OS CentOS 7.5, kernel 4.17.6-1.176.165.64, CPU 2 x Intel® Xeon® Gold 6154 @ 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.

© 2019 Storage Networking Industry Association. All Rights Reserved.
System Level Performance

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

Higher is better
Lower is better

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

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

(70Read/30Write Random 4KB)

60x QOS advantage
3

6x IOPs advantage at QD=1
2

Source – Intel tested: Measured using FIO 3.1. Common Configuration - Intel® 2U Server System, OS CentOS 7.5, kernel 4.17-6.1-64.el7, CPU 2 x Intel Xeon® E5-2650 v3 @ 2.30GHz (10 cores), RAM 256GB DDR4 @ 2666MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P4600 1.6TB. Intel® Microcode: 0x2000043; System BIOS: 0.0.01.0013; ME Firmware: 04.00.04.294; BMC Firmware: 1.51.0700005; RAID Controller: 3.1. The benchmark results may need to be revised as additional testing is conducted. Performance results are based on tests conducted 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.

Source – Intel tested: 4K 70/30 RW Performance at Low Queue Depth. Test and System Configuration: CPU: Xeon® Skylake Gold 6140 FC-LGA14B 2.3GHz 24.75MB 140W 18 cores CD8067303405200 , CPU Sockets: 2, RAM Capacity: 32G, RAM Model: DDR4, RAM Stacking: NA, DIMM Slots Populated: 2 slots, PCIe Attach: CPU (not PCH lane attach), Chipset: Intel C620 chipset BIOS: SE5C620.86B.00.01.0013.030920180427 , Switch/Mux/Network Adapter: Custom, Ethernet 10GBase-T 800mm straight SFF-8611 to right angle SFF-8611 Intel® AXXCBL800CVCR, OS: CentOS 7.5, kernel 4.14.50(LTS), FIO version: 3.5; NVMe Driver: Inbox, C-states: Disabled, Hyper Threading: Disabled, CPU Governor (through OS): Performance Mode. EIST (Speed Step), Intel Turbo Mode=Disabled, and P-states = Enabled. The benchmark results may need to be revised as additional testing is conducted. Performance results are based on tests conducted as of July 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Source – Intel tested: 4K Read latency under 300MB/s VMB Workload. Measured using FIO 2.15. Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Common Configuration - Intel® 2U Server System, OS CentOS 7.5, kernel 4.17-6.1-64.el7, CPU 2 x Intel Xeon® E5-2650 v3 @ 2.30GHz (10 cores), RAM 256GB DDR4 @ 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 FIO 2.15. System BIOS: 0.0.01.0013; ME Firmware: 04.00.04.294; BMC Firmware: 1.51.0700005; RAID Controller: 3.1. The benchmark results may need to be revised as additional testing is conducted. Performance results are based on tests conducted as of July 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

© 2019 Storage Networking Industry Association. 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)

Idle Average

1. Source—Intel tested. Average read latency measured at queue depth 1 during 4K random write workload. Measured using FIO 3.1. Common Configuration – Intel 2U Server System, OS CentOS 7.5, kernel 4.17.6-1.el7, 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 1.6TB.

2. See Notes on previous slide.
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

10s GB <100ns

1 Intel® Optane™ SSD: 1sTB <10μsecs

IMPROVING SSD PERFORMANCE

3D NAND SSD WARM TIER

HDD / TAPE COLD TIER

STORAGE PERFORMANCE GAP

CAPACITY GAP

DRAM HOT TIER

© 2019 Storage Networking Industry Association. 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 */

**Intel® Optane™ SSD**

**Intel® Optane™ DC Persistent Memory**

4-10μs for Linux¹

---


© 2019 Storage Networking Industry Association. All Rights Reserved.
PERSISTENT MEMORY PLATFORM SUPPORT

- BIOS
- OPERATING SYSTEM
- SNIA NVM PROGRAMMING MODEL
- APPLICATION

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

© 2019 Storage Networking Industry Association. All Rights Reserved.
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
- ~100ns to ~350ns for 64B

**Memory Subsystem Idle Avg. is About**

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.

2 App Direct Mode, NeonCity, LBG B1 chipset, CLX BD 28 Core (QDF Q0Y2), Memory Conf 102GB DDR4 (per socket) 2933 MT/s, Optane DCPMM 128GB, BIOS 561.D09, BKC version WW48.5BKC, Linux OS 4.18.8-100.fc27, Spectre/Meltdown Patched (1,2,3, 3a)

© 2019 Storage Networking Industry Association. 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
**APP DIRECT USAGE EXAMPLE**

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

© 2019 Storage Networking Industry Association. 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
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
NAND TECHNOLOGY ADVANCEMENT

2018 3D QLC
1024 Gb/Die

64 LAYERS

256X INCREASE IN AREAL DENSITY


2D SLC 4 Gb
AREAL DENSITY

2D MLC 128 Gb
3D TLC 384 Gb
3D QLC 1024 Gb

E1.L
32TB

INTEL SSD

© 2019 Storage Networking Industry Association. 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
COMPLETE IN PERFORMANCE, CAPACITY, FIT

MEMORY

PERSISTENT MEMORY

STORAGE

GROWING MEMORY CAPACITY

IMPROVING SSD PERFORMANCE

DELIVERING EFFICIENT STORAGE

INTEL® QLC 3D NAND SSD

DRAM HOT TIER

HDD / TAPE COLD TIER

10s GB
<100ns

100s GB
<1usec

1s TB
<10µsecs

10s TB
<100µsecs

10s TB
<10 msecs

© 2019 Storage Networking Industry Association. All Rights Reserved.
Converging Memory and Storage

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