Flash Memory Summit

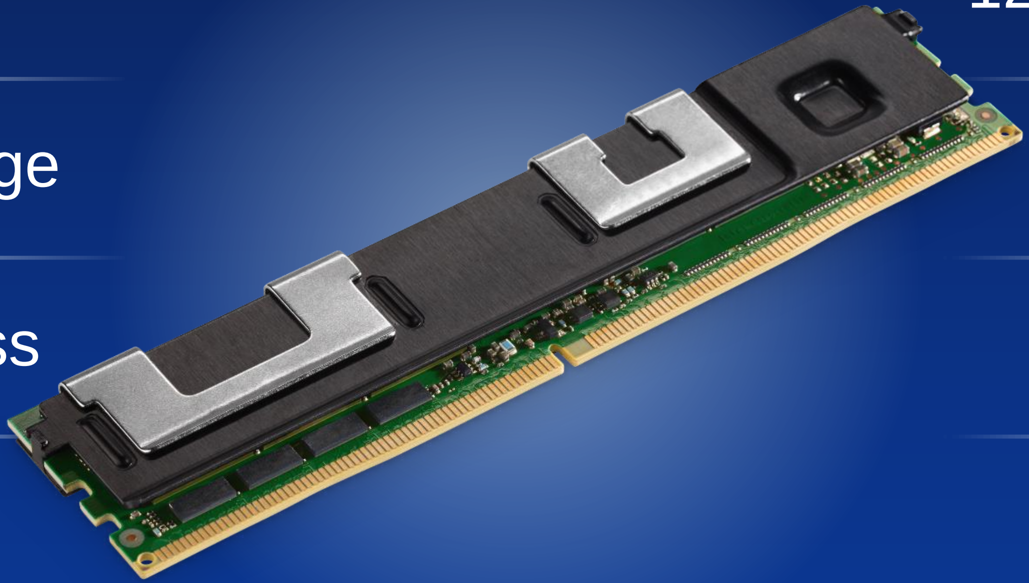
Breakthrough Data-Centric Computing with a New Memory Tier

Alper Ilkbahar
Vice President & General Manager
Data Center Group, Intel Corporation

Intel Optane™ DC Persistent Memory

Intel Xeon Platinum inside

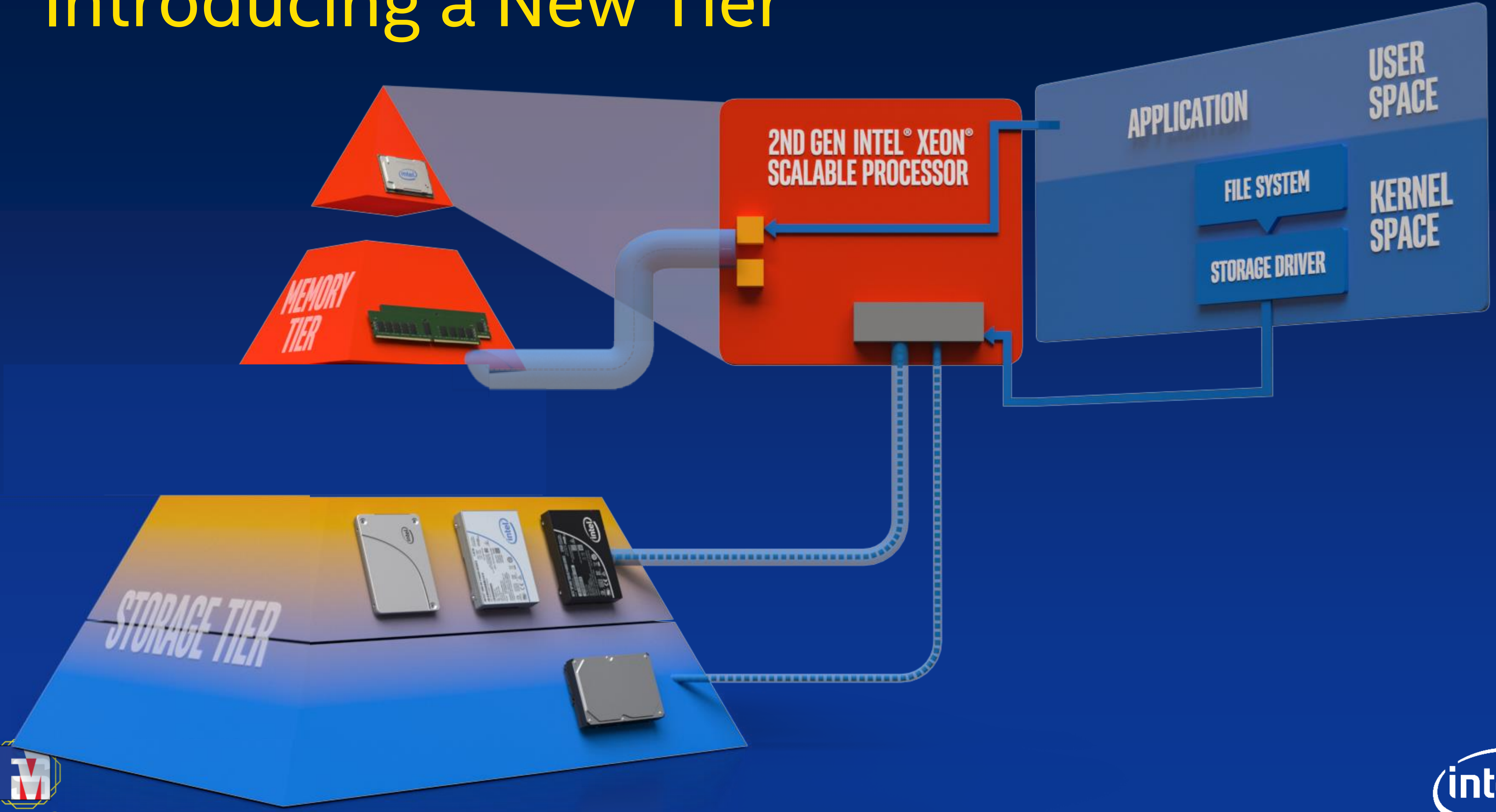
Big and Affordable Memory	128, 256, 512GB Modules
High Performance Storage	DDR4 Pin Compatible
Direct Load/Store Access	Hardware Encryption
Native Persistence	High Reliability



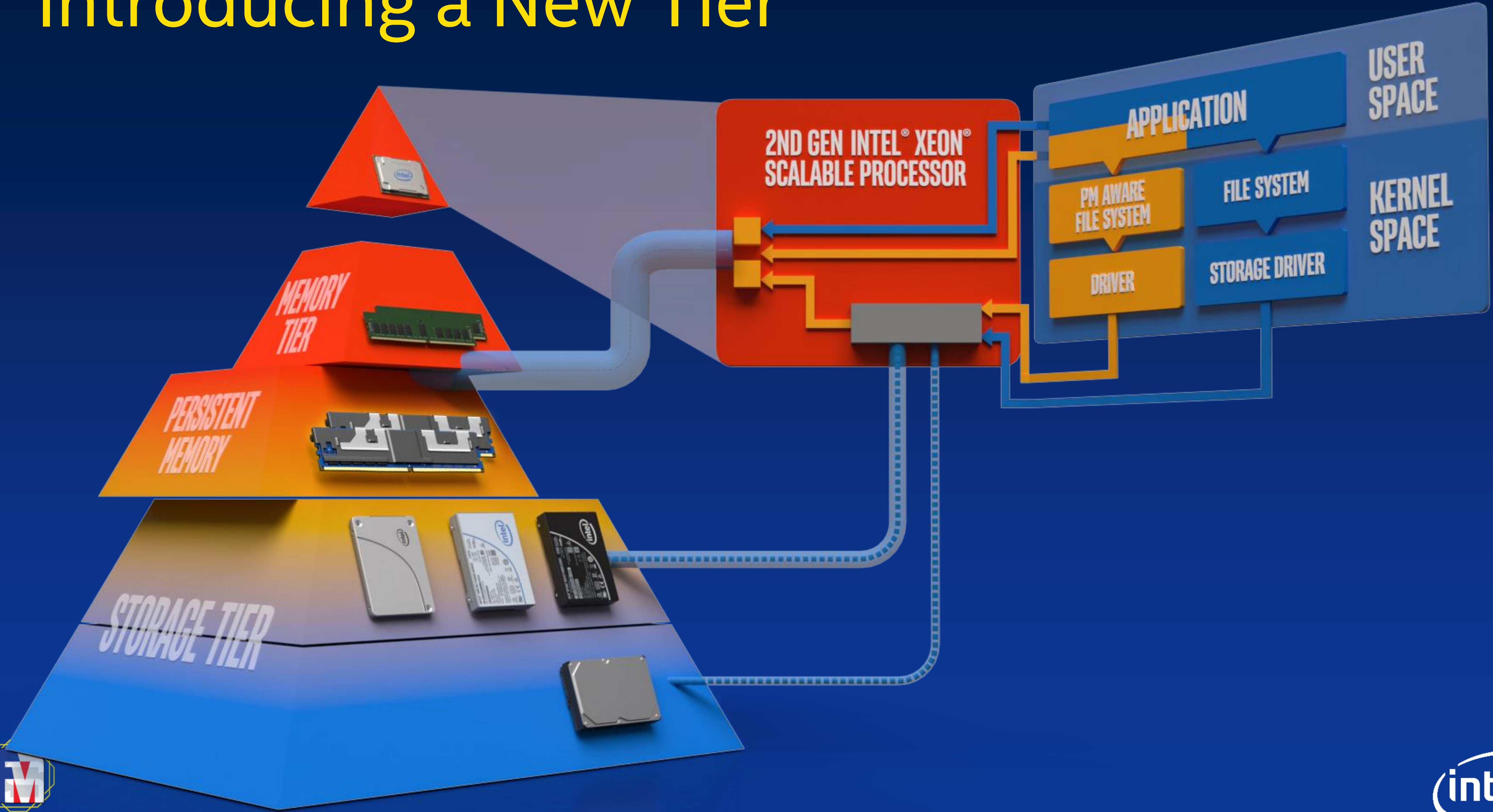
LAUNCHED APRIL 2ND

NOW SHIPPING IN VOLUME

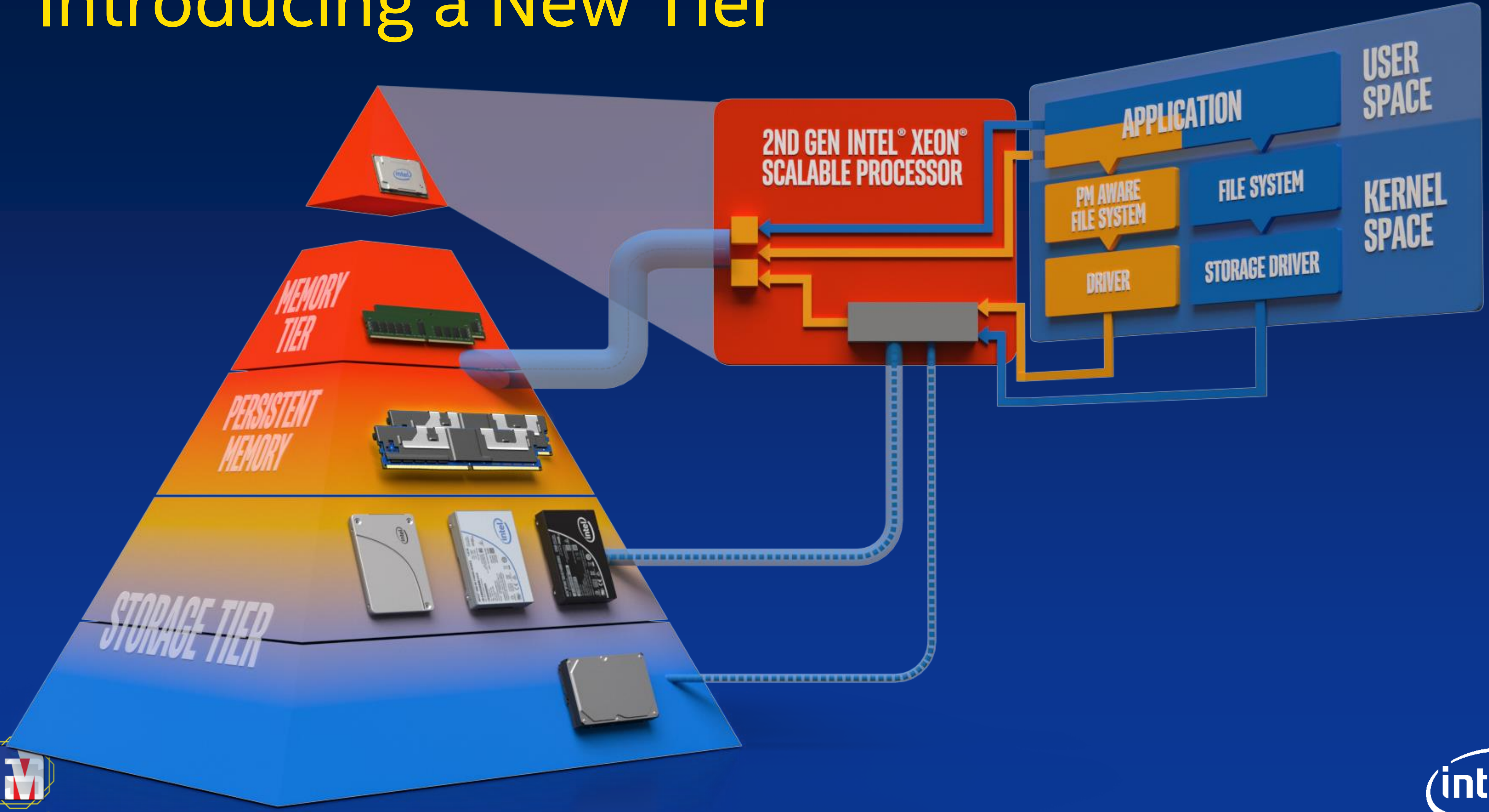
Introducing a New Tier



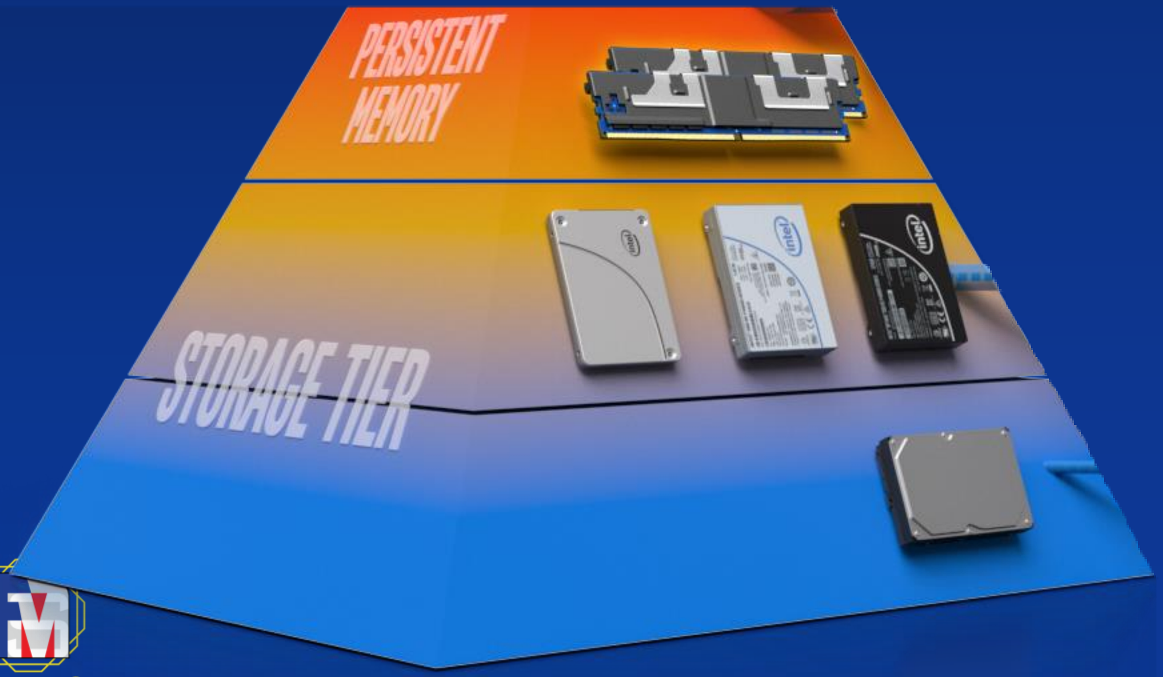
Introducing a New Tier



Introducing a New Tier



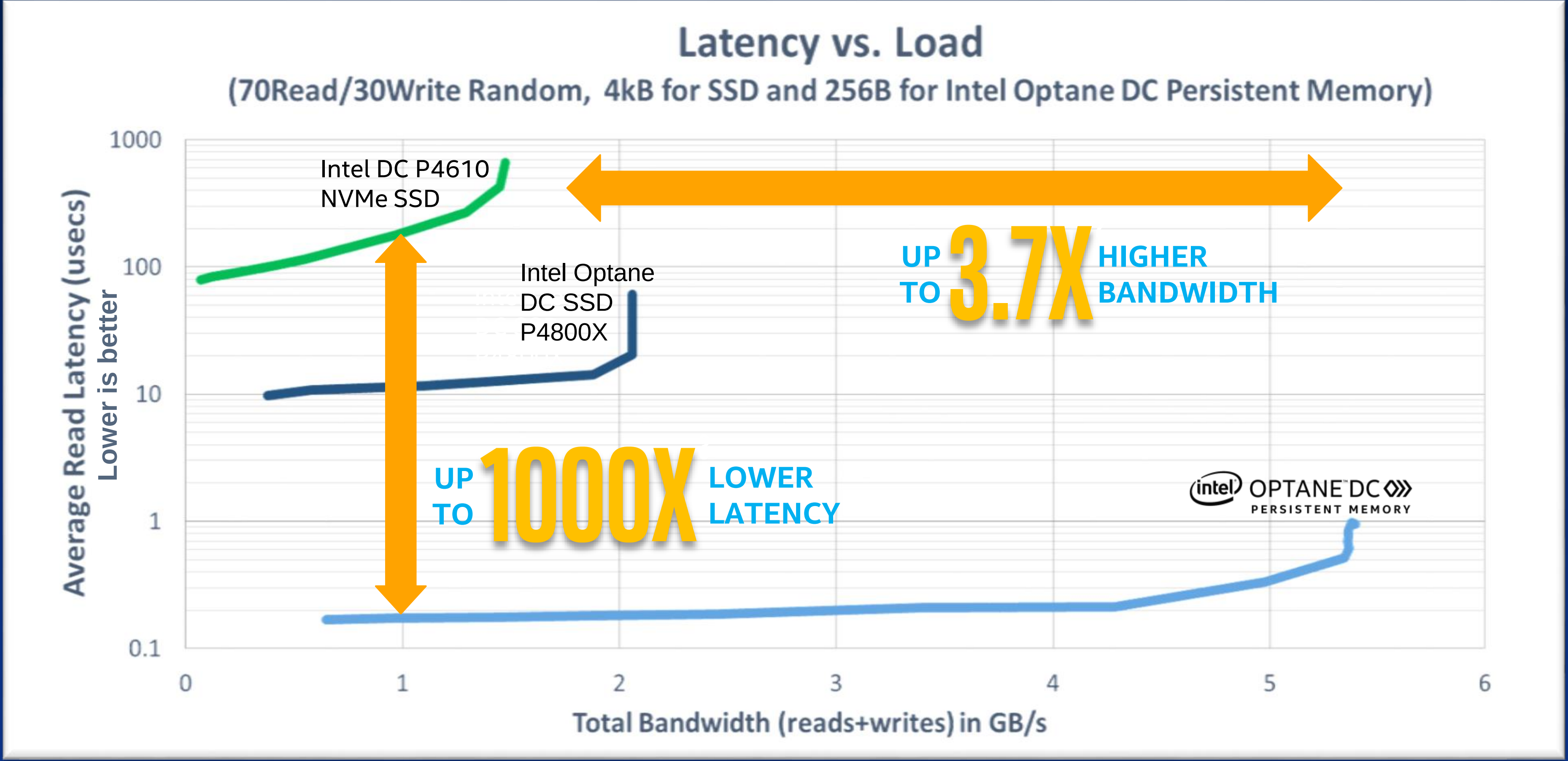
Persistent Memory as Storage



Persistent Memory: Low Latency

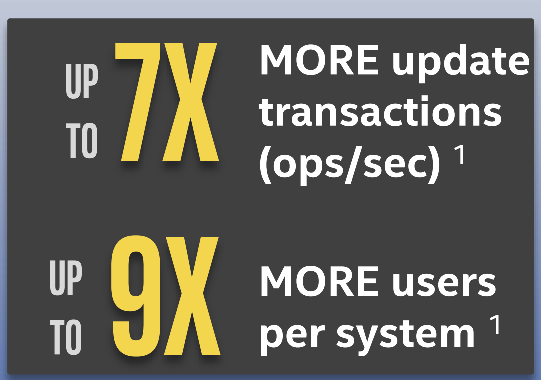
More Bandwidth:
Up to 3.7X read/write bandwidth vs NVMe SSDs, with one module; more with multiple modules

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.

Breaking IO Bottlenecks



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.

Flash Memory Summit

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

intel

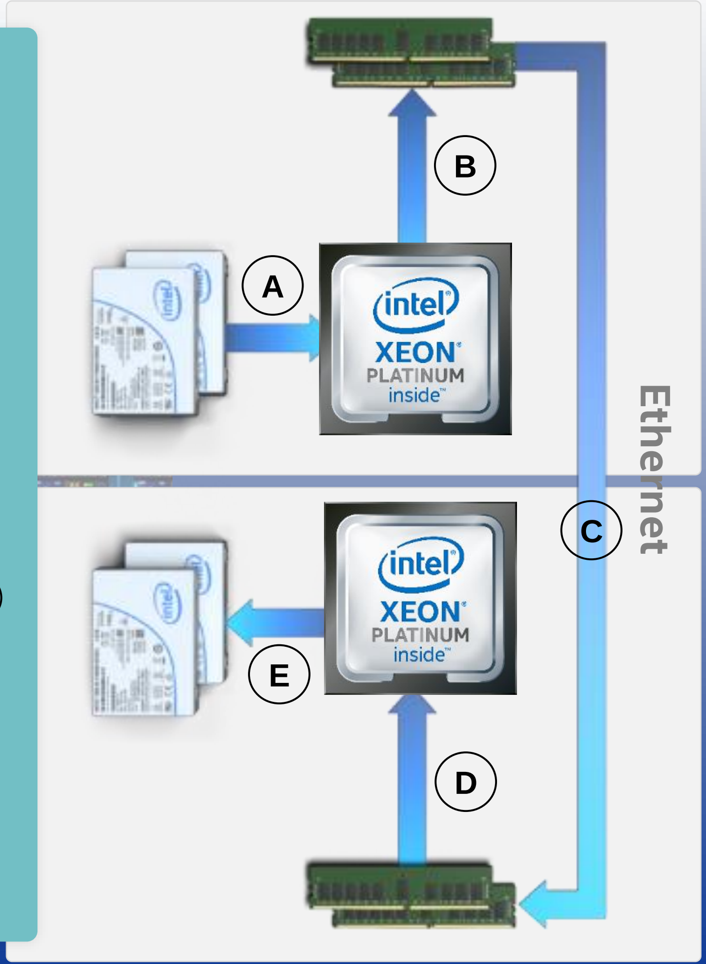
Data Replication with Persistent Memory

Traditional Data Replication

Multiple data hops:

1. Processors move data to remote memory A B C
2. Remote processor moves to SSD D E

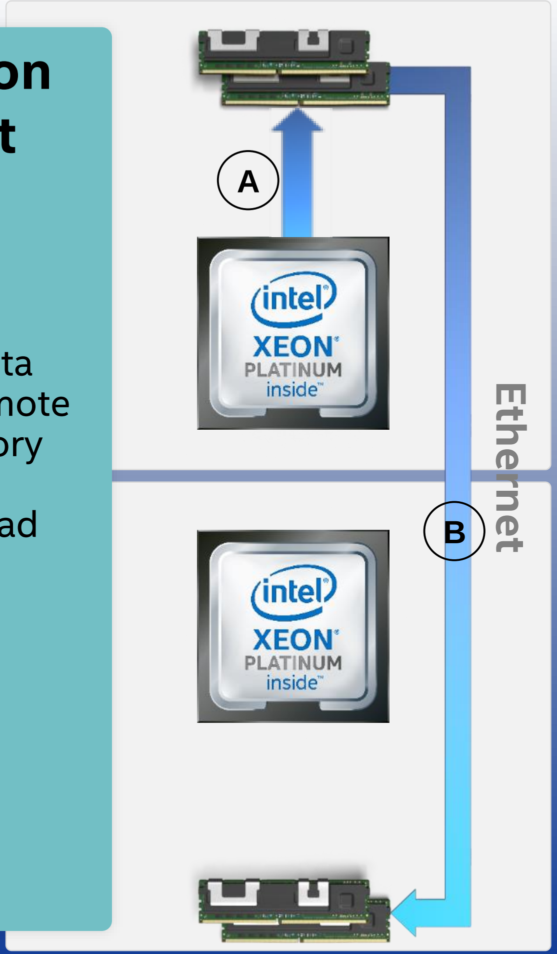
Software overhead for network and storage drivers



Data Replication with Persistent Memory

Single hop:

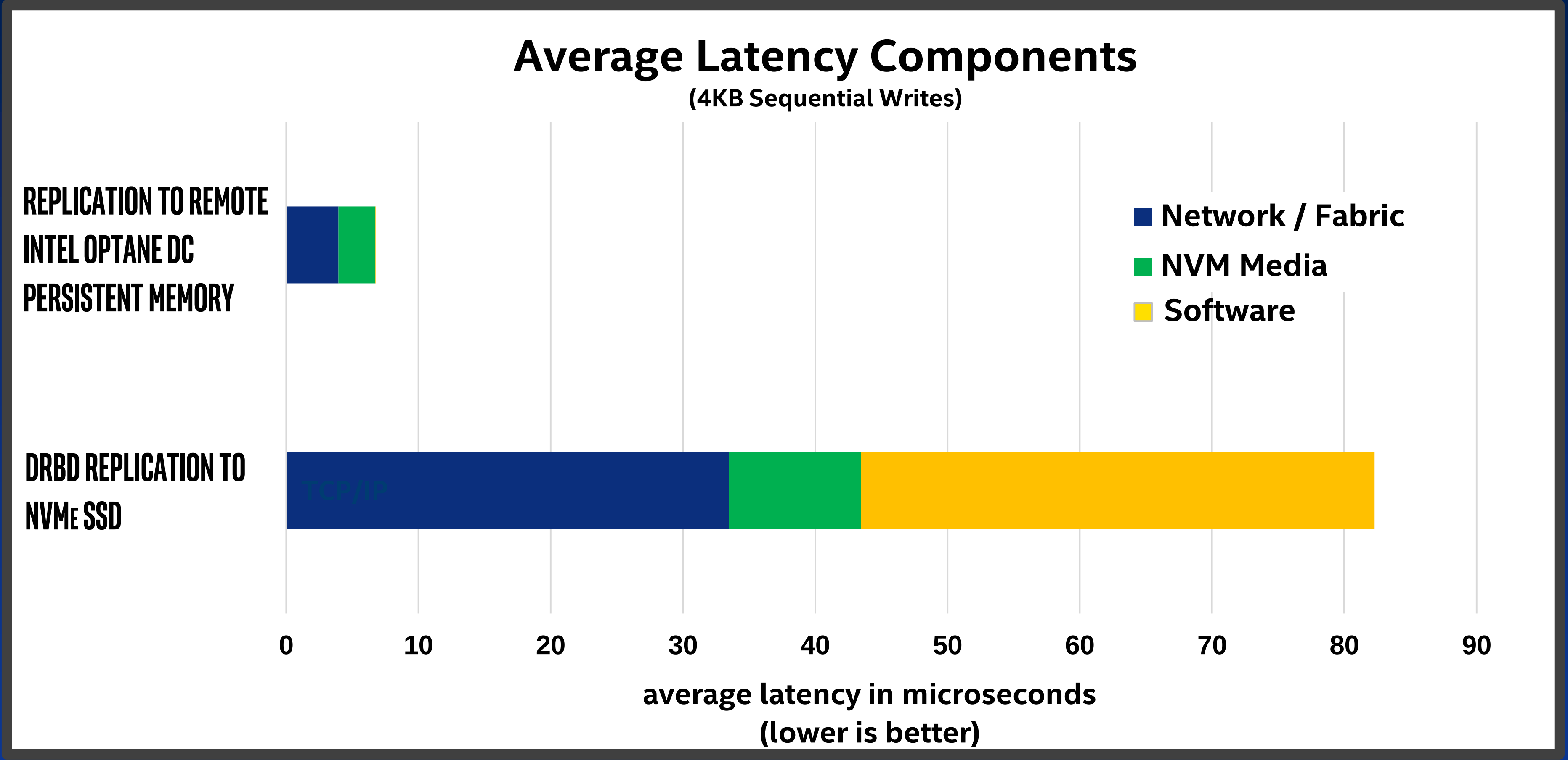
1. RDMA moves data from local to remote persistent memory with minimal software overhead A B



Data Replication with Persistent Memory

UP TO **14X**

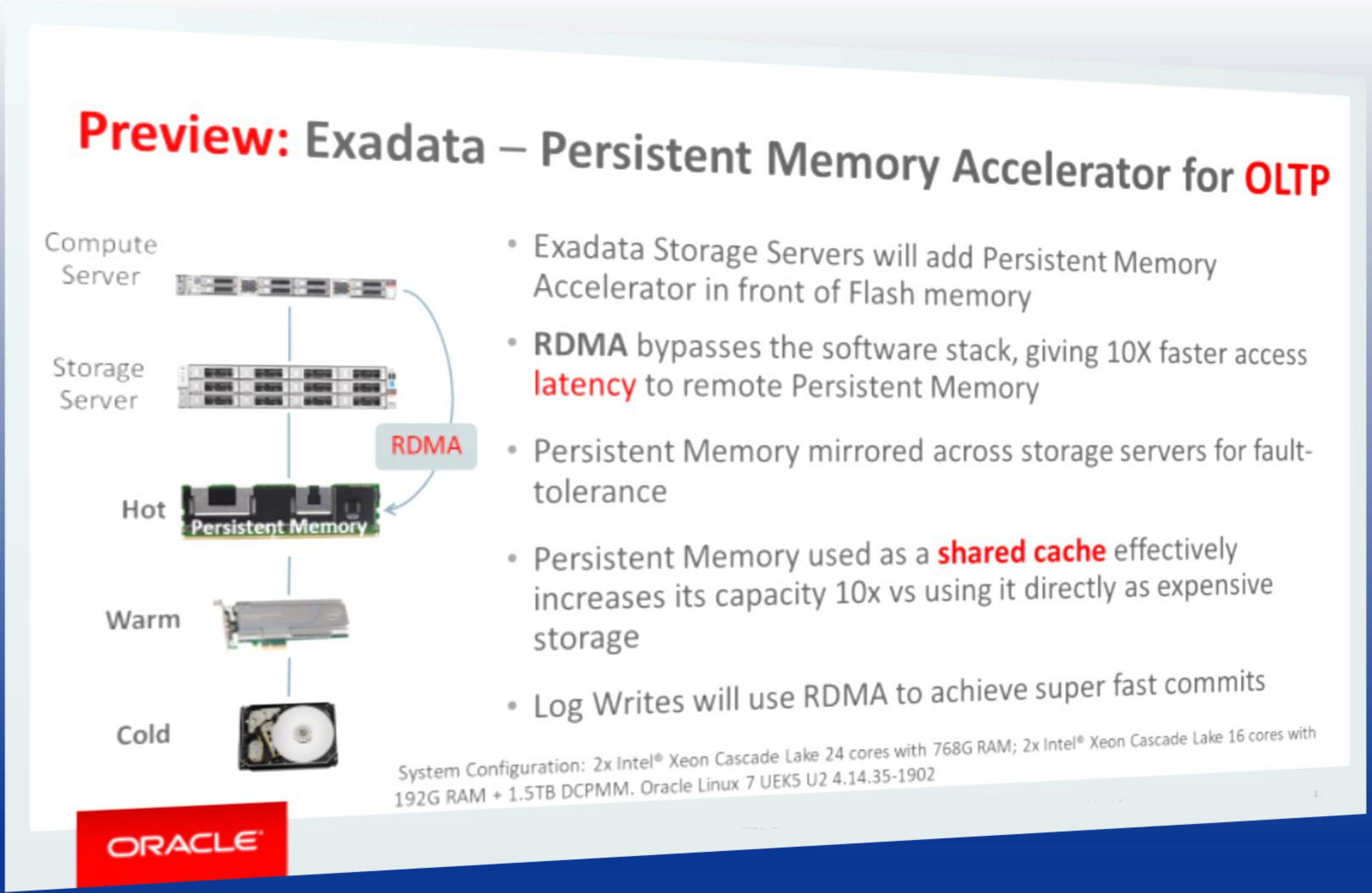
FASTER replication latency [1]



[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 38 for configurations.

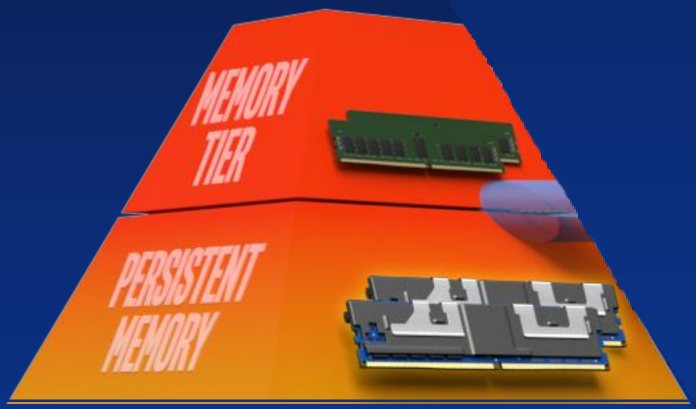
Customer Example: Oracle Exadata



Slide courtesy of Oracle

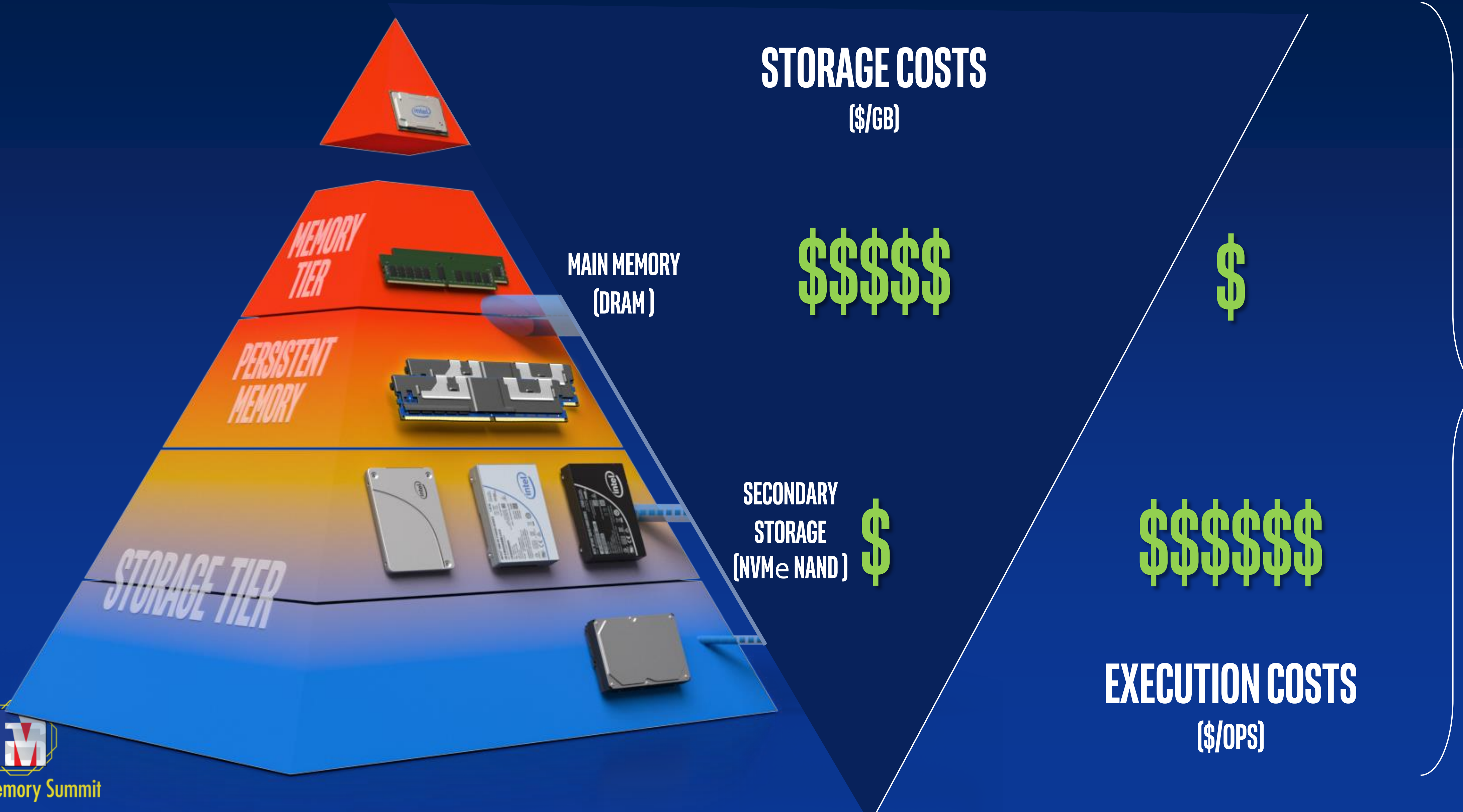
10X
lower latency

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



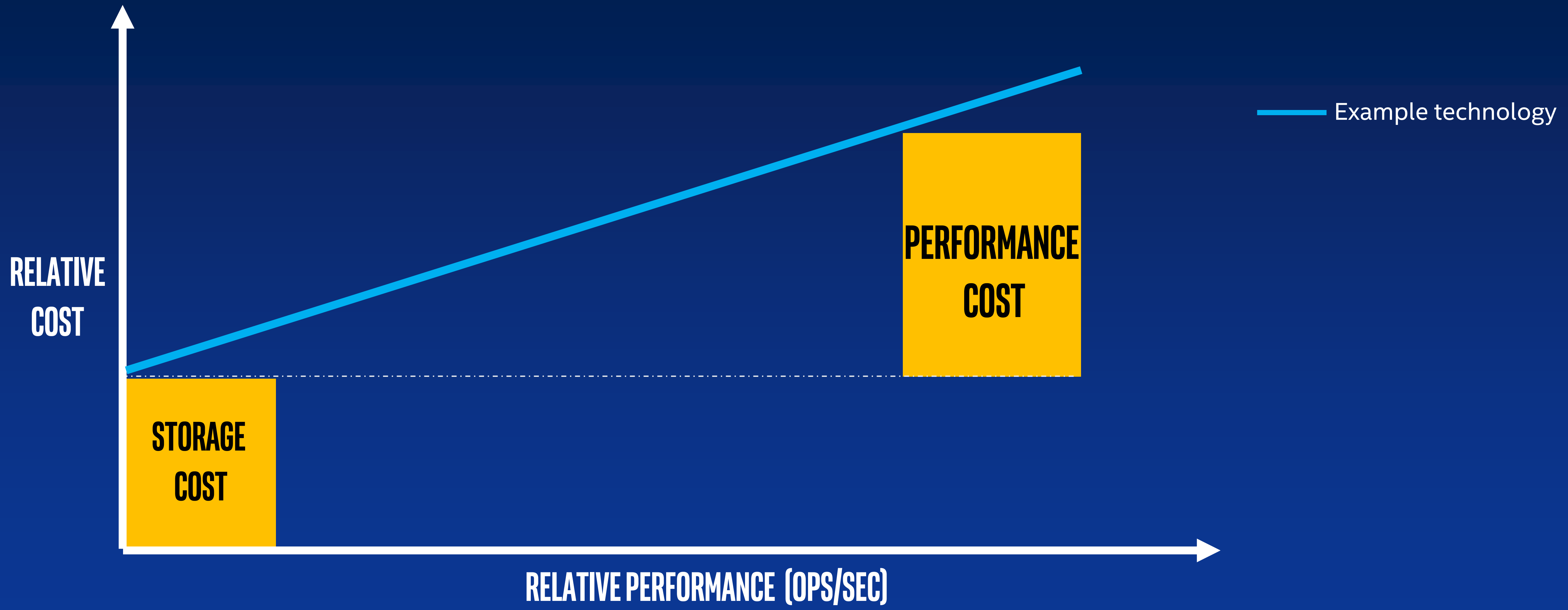
Persistent Memory as Main Memory

Cost vs Performance Framework



Flash Memory Summit

Cost vs Performance Framework

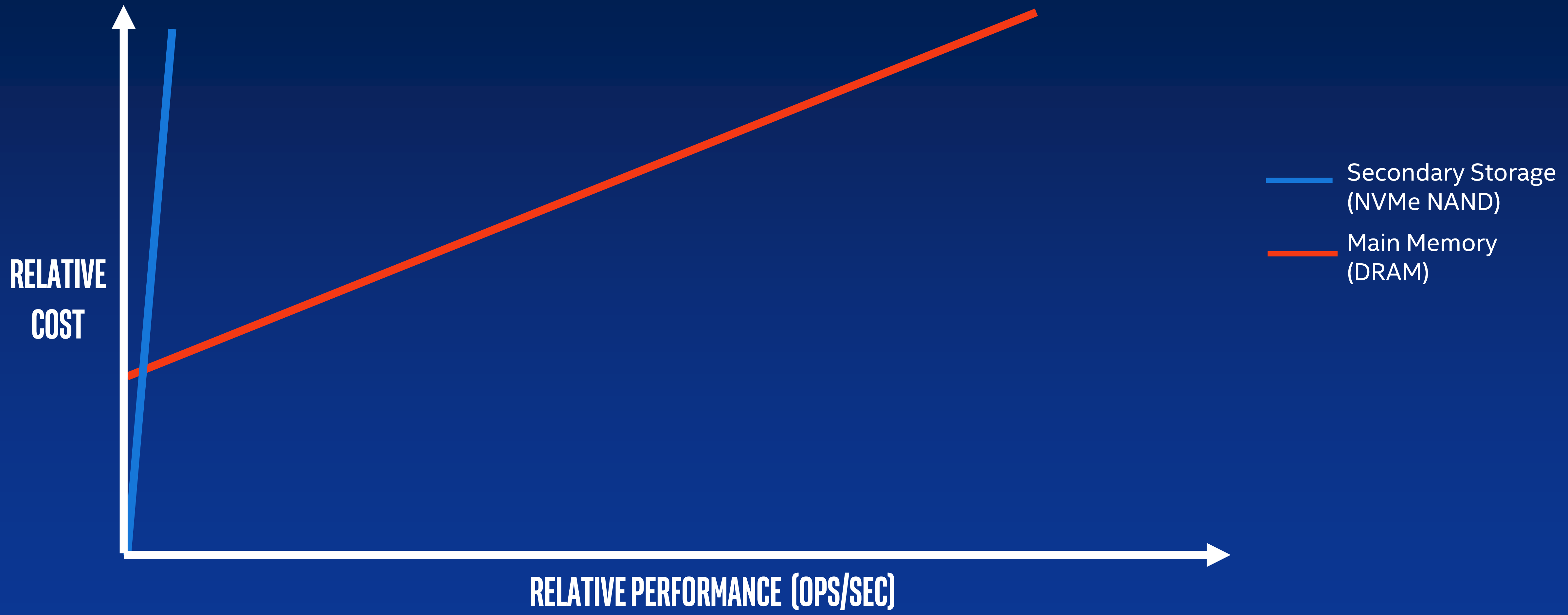


D. Lomet: Cost/Performance in Modern Data Stores, How Data Caching Systems Succeed, DaMoN, 2018; www.hpts.ws/papers/2017/gongshow/06-lomet.pdf

Based on Intel internal testing.

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

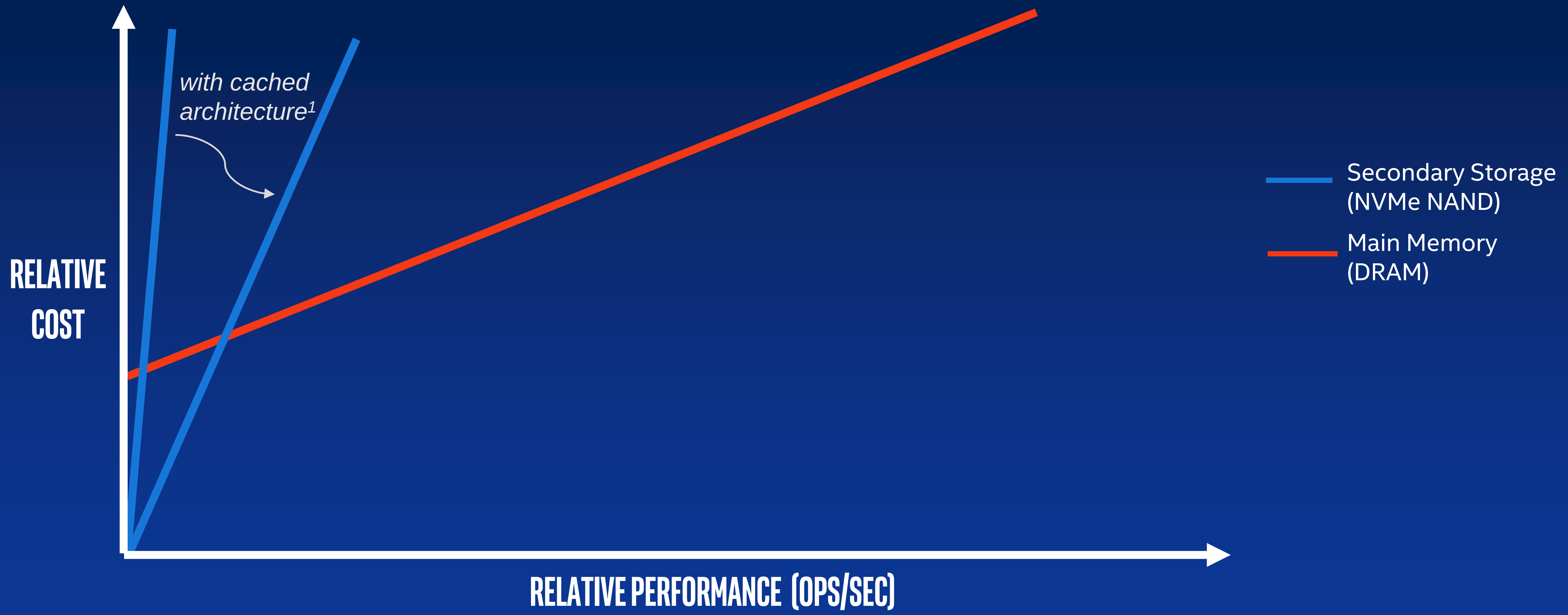
Cost vs Performance Framework



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

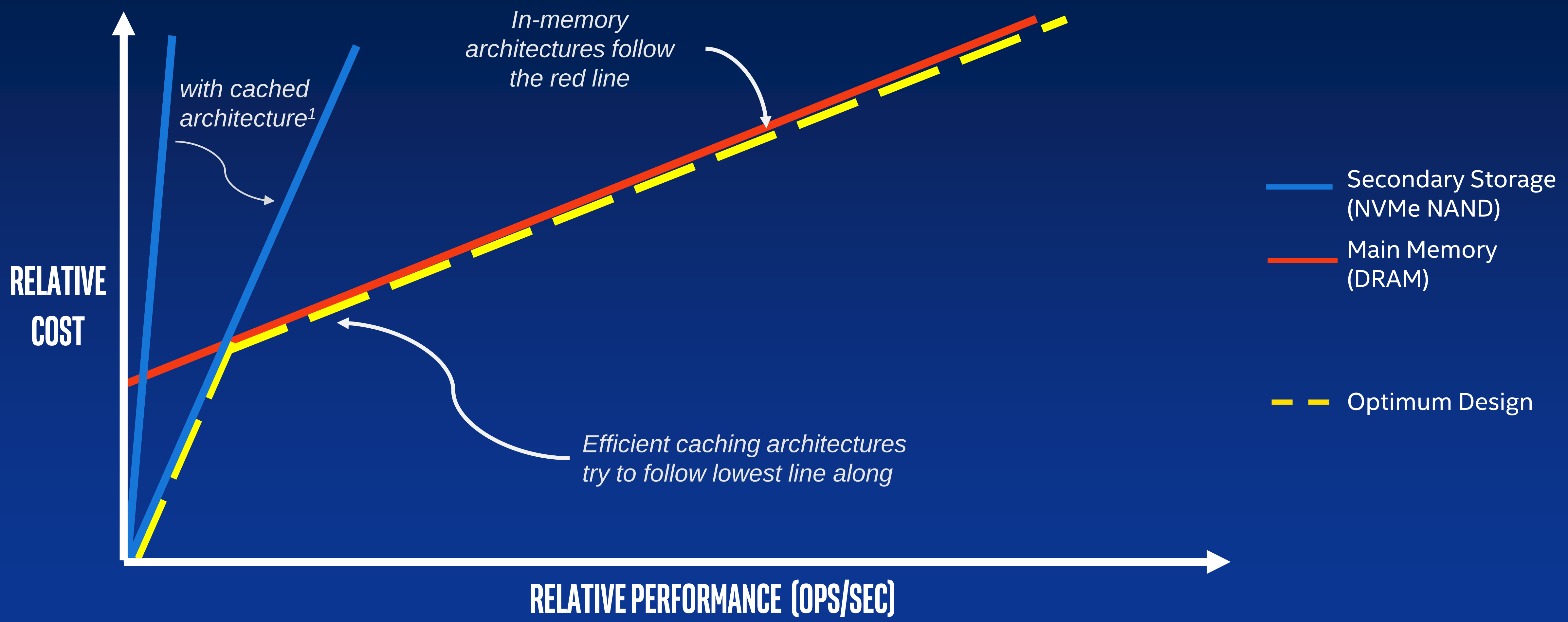


[1] Assumes 10% of data in the DRAM with 90% cache hit rate.

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

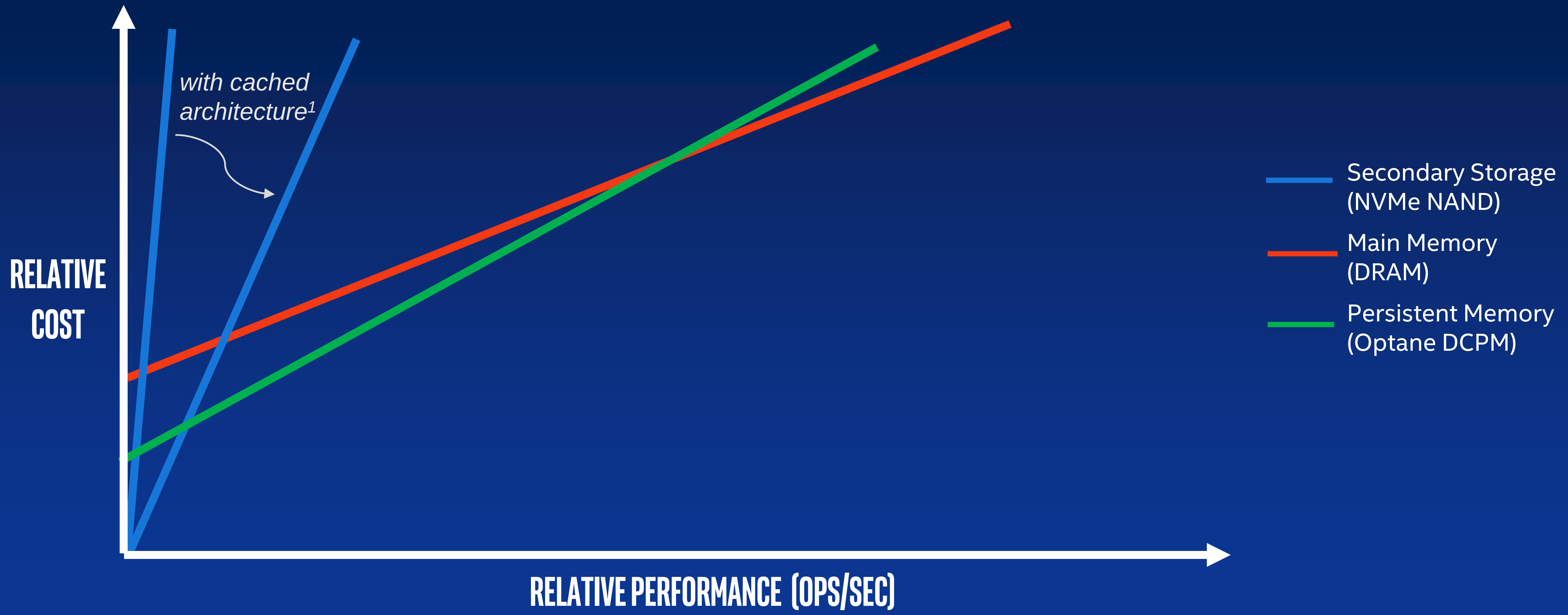


[1] Assumes 10% of data in the DRAM with 90% cache hit rate.

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

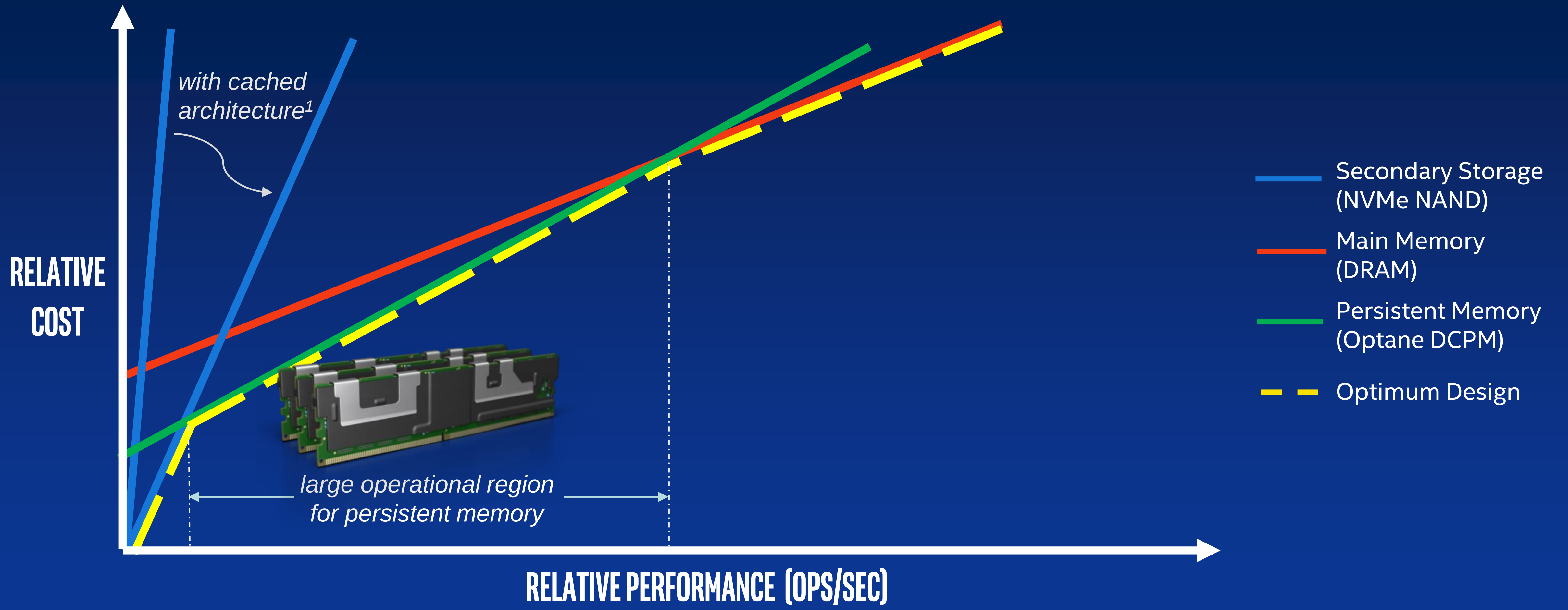


[1] Assumes 10% of data in the DRAM with 90% cache hit rate.

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

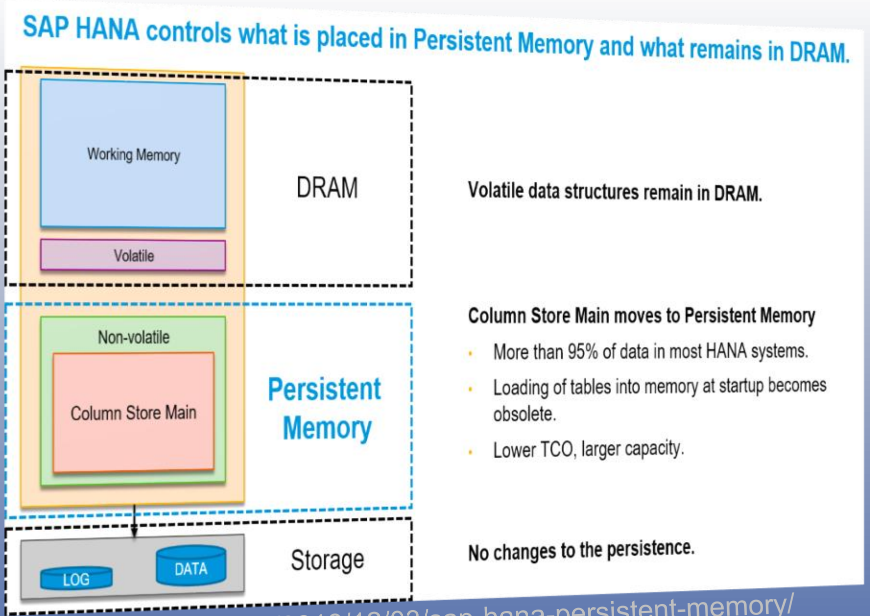


[1] Assumes 10% of data in the DRAM with 90% cache hit rate.

Based on Intel internal testing.

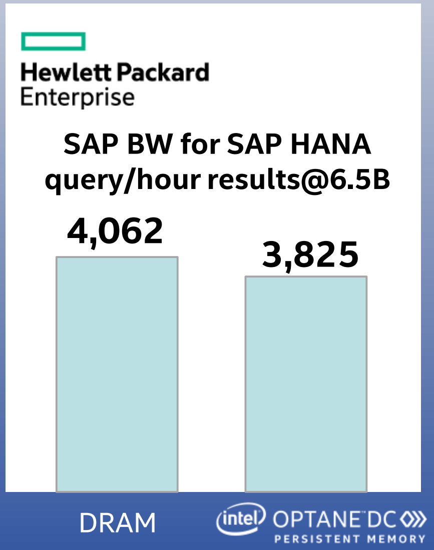
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See slide 32 for configurations.

SAP HANA and Persistent Memory

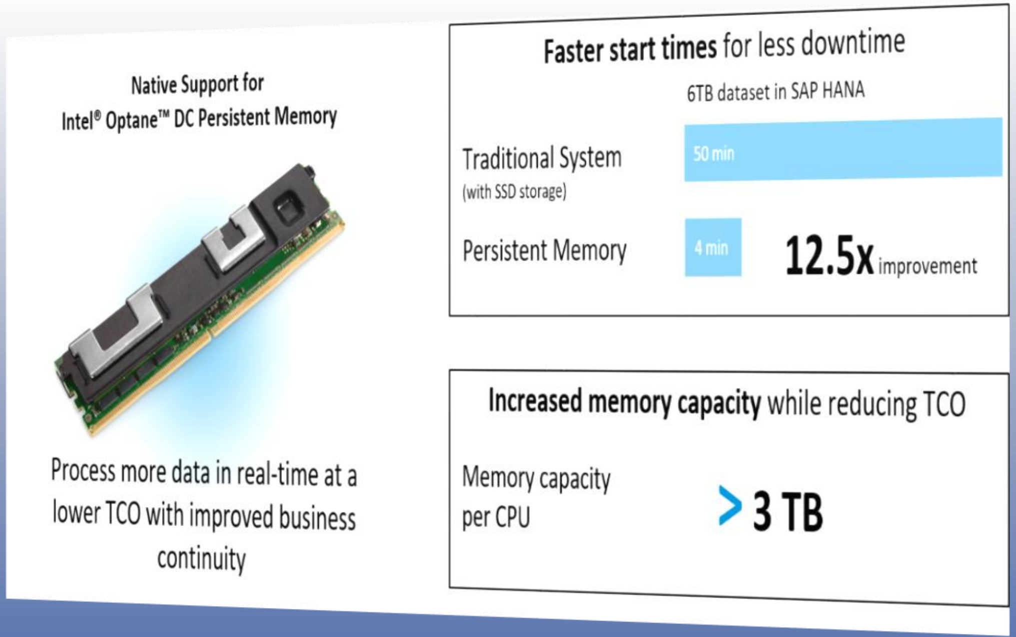


https://blogs.sap.com/2018/12/03/sap-hana-persistent-memory/

MORE THAN **95%** of data in most HANA systems in persistent memory



5.9% Performance delta vs all-DRAM system



12.5X FASTER restart times

Flash Memory Summit

Vibrant Software Ecosystem

AI / ANALYTICS

- GIGASPACES innovate with confidence
- Apache Spark
- sas
- 海鑫科金 Hisign Technology
- aws
- cloudera
- 宝信软件 BAOSIGHT
- AsiaInfo 亚信

DATABASES

- Microsoft
- redis
- MEMCACHED
- ORACLE
- SAP
- ORACLE
- 人大金仓 Kingbase
- SUNJESOFT SUNJESOFT Inc.
- MySQL
- AEROSPIKE
- Cassandra
- ALTIBASE
- kx
- NUODB
- redislabs HOME OF REDIS

INFRASTRUCTURE & STORAGE

- Microsoft
- openstack
- NetApp
- hazelcast
- Virtuozzo
- MemVerge

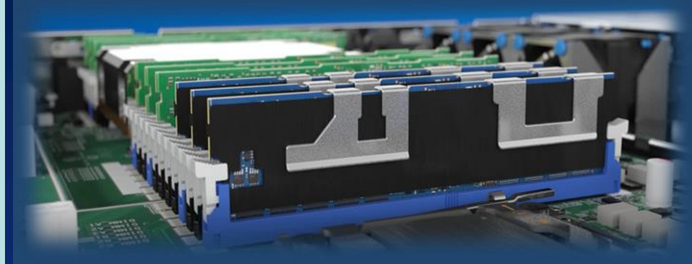
OPERATING SYSTEMS

- Microsoft
- redhat
- SUSE We adapt. You succeed.
- vmware
- ubuntu Delivered by Canonical

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

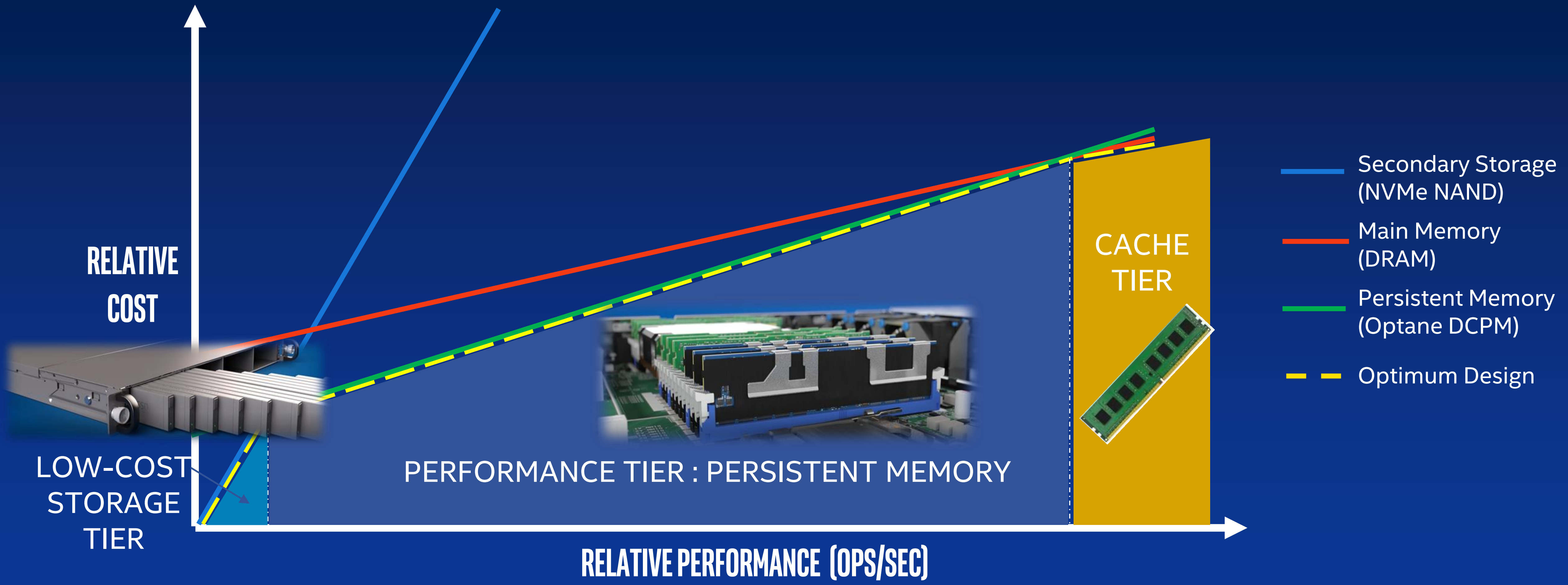
Disaggregated with NVMe oF and RDMA
Sequential accesses through periodic checkpoints and snapshot images
Integrated cloud storage (S3)

INFRASTRUCTURE NODES

Interconnected with 100Gb ethernet
Kubernetes orchestration

Video shown – see separate link

Cost vs Performance Framework for DMP

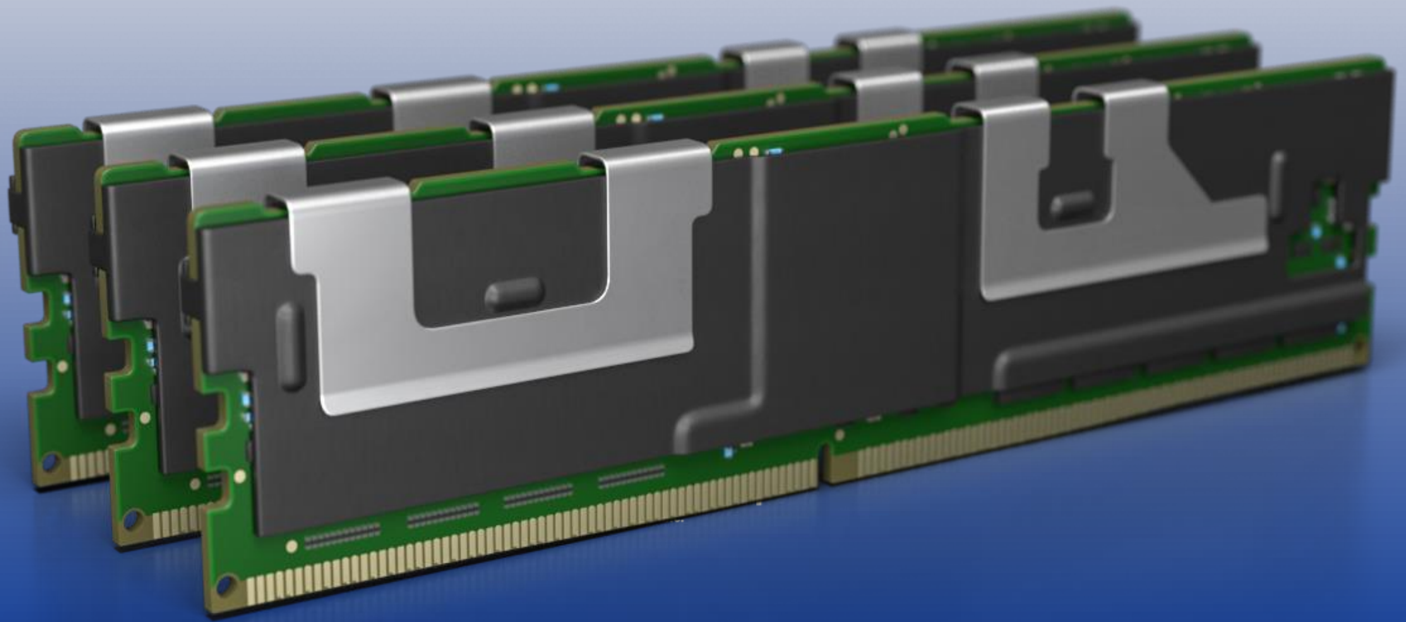


Based on Intel internal testing.

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

OPTANE™ DC

PERSISTENT MEMORY



Join the Persistent Memory Revolution

intel OPTANE™ DC
PERSISTENT MEMORY

accenture
AEROSPIKE
AIC
Alibaba Group 阿里巴巴集团
ALTIBASE
Apache CASSANDRA
Apache Spark
AsiaInfo 亚信科技
Atos
aws

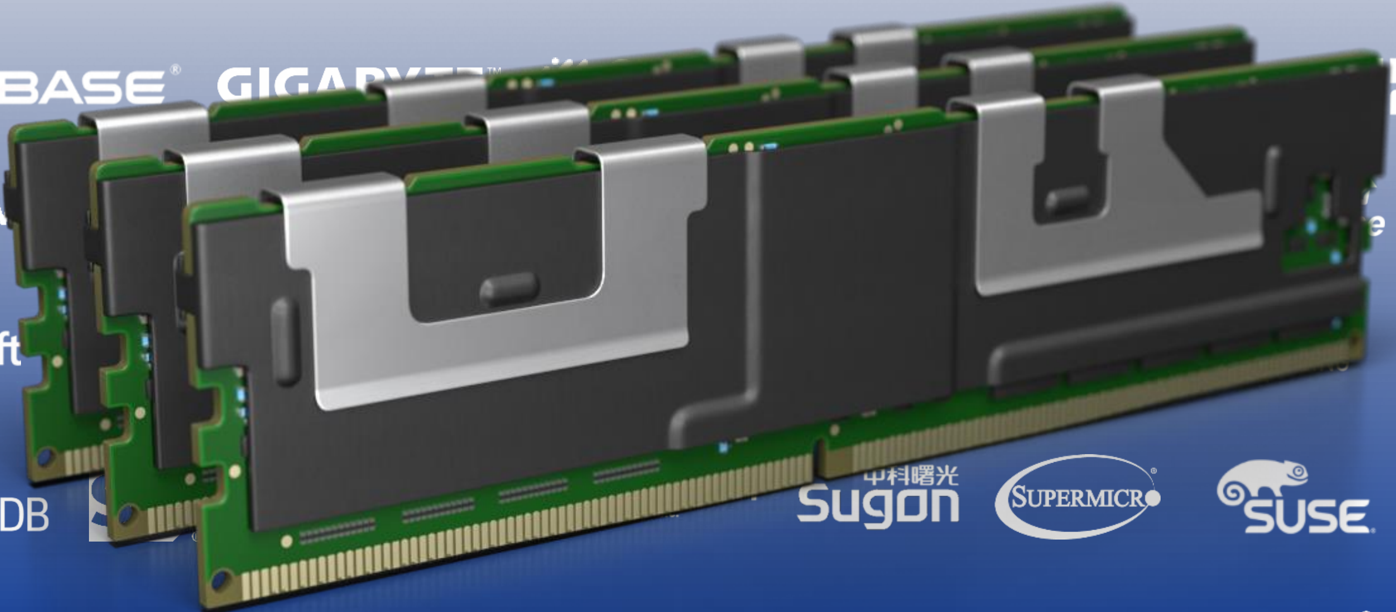
Baidu Cloud
宝信软件 BAOSIGHT
BONC 东方国信
cisco
cloudera
COLFAX Customized Solutions
CRAY
databricks
DATASTAX

DELL EMC
FUJITSU
GBASE
hazelcast
Hewlett Packard Enterprise

海鑫科金 HISIGN TECHNOLOGY
Kingsoft Cloud KSCLOUD
kx

Lenovo
Microsoft
QCT
redhat

redislabs
RocksDB
中科曙光 Sugon
SUPERMICRO
SUSE
Tencent Cloud
THINKMATE HIGH PERFORMANCE COMPUTING



TYAN
ubuntu
UNICOM Engineering, Inc. A Division of UNICOM Global
Virtuozzo
vmware
wiwynn
World Wide Technology

Join the Persistent Memory Revolution

intel® OPTANE™ DC PERSISTENT MEMORY

accenture · AEROSPIKE · AIC · Alibaba Group 阿里巴巴集团 · ALTIBASE · Apache CASSANDRA · APACHE Spark · AsiaInfo 亚信科技 · Atos · aws

Baidu Cloud · 宝信软件 BAOSIGHT · BONC 东方国信 · cisco · cloudera · COLFAX Customized Solutions · CRAY · databricks · DATASTAX

DELL EMC · FUJITSU · GBASE · GIGABYTE · GIGASPACES innovate with confidence · Google Cloud · H3C · hazelcast · Hewlett Packard Enterprise

海鑫科金 HISIGN TECHNOLOGY · HUAWEI · IBM · inspur · Inventec · JABIL · 人大金仓 Kingbase · Kingsoft Cloud KSCLOUD · kx

Lenovo · Microsoft · NARI 国电南瑞 NARI-TECH · NetApp · Neusoft · ORACLE · PENGUIN COMPUTING · QCT · redhat

redislabs · RocksDB · SAP · sas · SUNJESOFT SUNJESOFT Inc. · 中科曙光 Sugon · SUPERMICRO · SUSE · Tencent Cloud · THINKMATE HIGH PERFORMANCE COMPUTING

TYAN · ubuntu · UNICOM Engineering, Inc. A Division of UNICOM Global · Virtuozzo · vmware · wiwynn · World Wide Technology

Join the Persistent Memory Revolution

Flash Memory Summit · intel

intel®

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 microprocessors. 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.
-

Back-up

Systems & Configurations

Data Management Platform Configurations

Node Information:

- os_release: Fedora 30 Server
- kenrel: 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

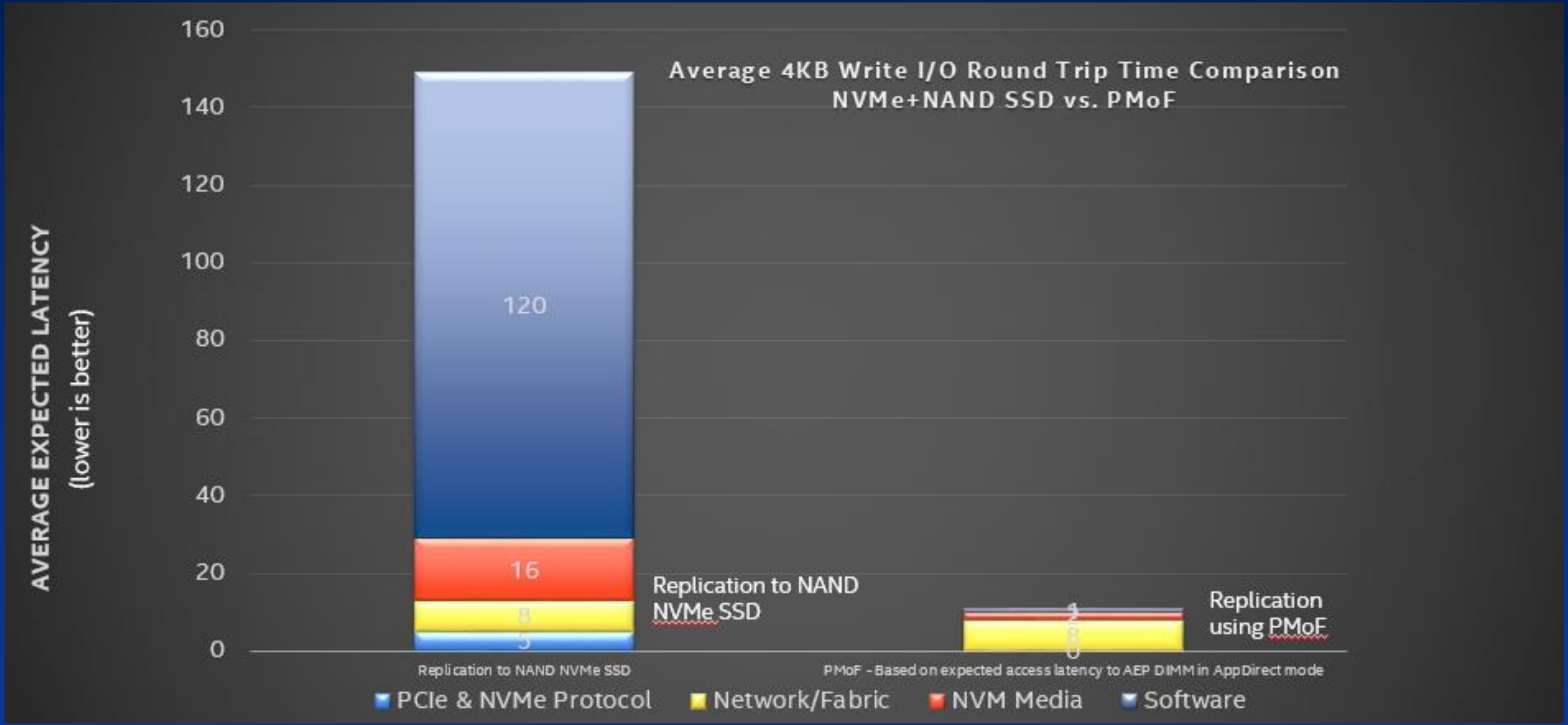
Persistent Memory: Low Latency “P4800 and P4610

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)

Replication Latency with PMoF

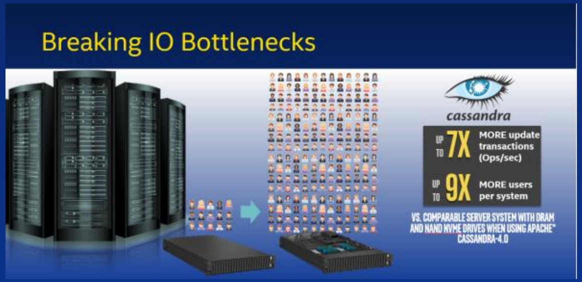


Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. Software and workloads used in performance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

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 sy

Flash Memory Summit

1/3 Cassandra Configuration Summary



Parameter	NVMe	DCPMM
Test by	Intel/Java Performance Team	Intel/Java Performance Team
Test date	22/02/2019	22/02/2019
Platform	S2600WFD	S2600WFD
# Nodes	1	1
# Sockets	2	2
CPU	8280L	8280L
Cores/socket, Threads/socket	28/56	28/56
ucode	0x4000013	0x4000013
HT	On	On
Turbo	On	On
BIOS version	SE5C620.86B.0D.01.0286.011120190816	SE5C620.86B.0D.01.0286.011120190816
DCPMM BKC version	NA	WW52 -2018
DCPMM FW version	NA	5318
System DDR Mem Config: slots / cap / run-speed	12 slots / 16GB / 2666	12 slots / 16GB / 2666
System DCPMM Config: slots / cap / run-speed	-	12 slots / 512GB
Total Memory/Node (DDR, DCPMM)	192GB, 0	192GB, 6TB
Storage - boot	1x Intel 800GB SSD OS Drive	1x Intel 800GB SSD OS Drive
Storage - application drives	4x P4610 1.6TB NVMe	12x512GB DCPMM
NIC	1x Intel X722	1x Intel X722
Software		
OS	Red Hat Enterprise Linux Server 7.6	Red Hat Enterprise Linux Server 7.6
Kernel	4.19.0 (64bit)	4.19.0 (64bit)
Mitigation log attached	Yes	Yes
DCPMM mode	NA	App Direct, Persistent Memory
Run Method	5 minute warm up post boot, then start performance recording	5 minute warm up post boot, then start performance recording
Iterations and result choice	3 iterations, median	3 iterations, median
Dataset size	Two 1.5 Billion Partitions (Insanity schema)	Two 1.5 Billion Partitions (Insanity schema)
Workload & version	Read Only, Mix 80% Read/20% Updates, Updates only	Read Only, Mix 80% Read/20% Updates, Updates only
Compiler	ANT 1.9.4 compiler for Cassandra	ANT 1.9.4 compiler for Cassandra
Libraries	NA	PMDK 1.5, LLPL (latest as of 2/20/1019)
Other SW (Frameworks, Topologies…)	NA	NA

Flash Memory Summit

2/3 Cassandra Configuration Settings

Software	Version
Cassandra	NVME uses 3.11.3 released version, DCPMM uses 4.0 trunk with persistent memory modifications: https://github.com/shyla226/cassandra/tree/13981_llpl_engine
PMDK	1.5
LLPL	https://github.com/pmem/llpl/ pulled 2/20/19
Java	Java™ SE Runtime Environment 1.8.0_201 Java HotSpot™ 64-bit Server VM (build 25.201)

Parameter	Value
Recommended Cassandra Production settings:	https://docs.datastax.com/en/dse/5.1/dse-dev/datastax_enterprise/config/configRecommendedSettings.html

Cassandra Settings	Value
Yaml modifications	concurrent_read/concurrent_write 168/168 for DCPMM concurrent_read/concurrent_write 56/32 for NVME
Jvm.options (comment out CMS section in file)	-Xms64G –Xmx64G –Xmn48G for DCPMM, no read cache -Xms32G –Xmx32G –Xmn24G for NVME, more read cache -XX:+UseAdaptiveSizePolicy for both
Number of Cassandra Processes, DataBases, Clusters	2 independent Cassandra processes each with a database, each process running 1 node cluster configuration
Cassandra Database per Application	cqlstress-insanity-example.yaml schema, with 1.5 Billion partition per database(3.0 Billion Total)
Cassandra Application pinned to CPU	numactl -m 0 -C 0-27,56-83 for socket 0 numactl -m 1 -C 28-55,84-111 for socket 1
Cassandra-Stress Command to Populate Database	cassandra-stress user profile=$CASSANDRA_HOME/tools/cqlstress-insanity-example.yaml ops\(insert=1\) n=1500000000 cl=ONE no-warmup -pop seq=1..1500000000 -mode native cql3 -node <ip_addr> -rate threads=<variable>
Cassandra-Stress Command to Read Database	cassandra-stress user profile=$CASSANDRA_HOME/tools/cqlstress-insanity-example.yaml ops\(simple1=1\) duration=30m cl=ONE -pop dist=UNIFORM\(1..1500000000\) -mode native cql3 –node <ip_addr> -rate threads=<variable>

3/3 Cassandra: Result Summary

- **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

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

Flash Memory Summit

Data Replication with Persistent Memory (slide 10):

14X Claim: System Configuration: 2x Intel® Xeon Cascade (HT on, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled), 384GB DDR4 2933 MT/s, Fedora 29, Linux kernel 4.20.13-200.fc29, Intel DC P3700 Series 400GB SSD, 100GbE Mellanox CX-5, CX-5 FW=16.23.1020, FIO version 3.14, libfabric v1.6.1, OFED drivers 4.6-1.0.1, 4KB sequential write I/O, DRBD version 9.0.19. Production released BKC, https://github.com/speed47/spectre-meltdown-checker

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/

intel