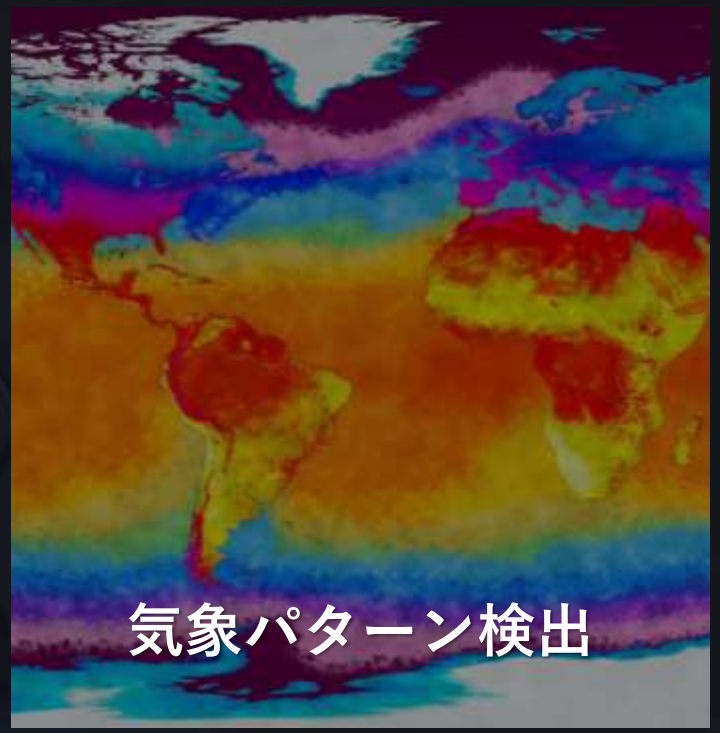
インテルHPC最新技術アップデート

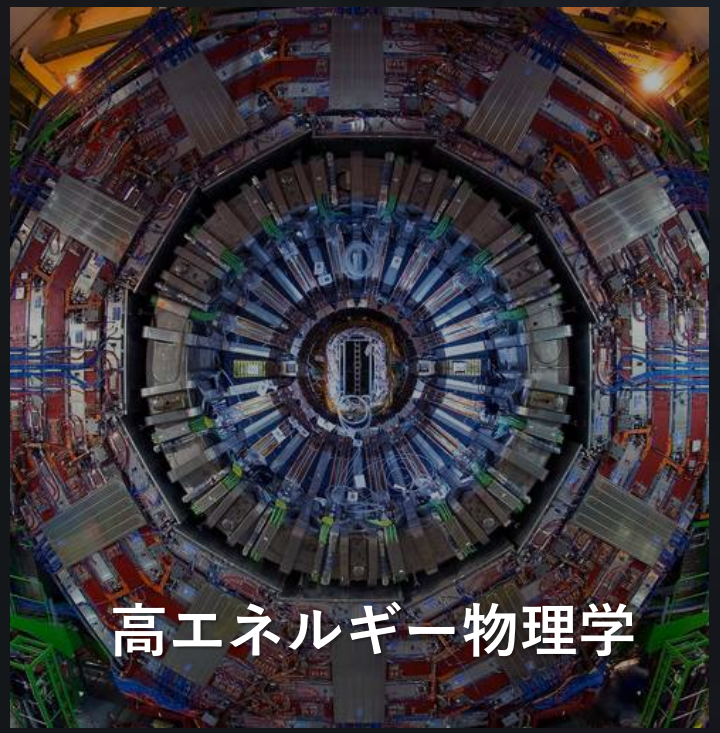
Technical Solution Specialist
Fumiyasu Ishibashi

パラダイムチェンジを牽引 - HPC & AI融合



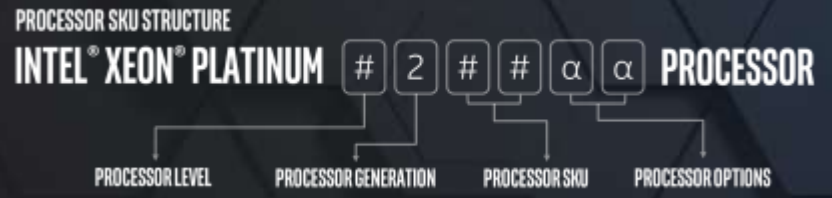
HPC のワークフローに
組み込まれる AI

HPC のシミュレーション
を高速化する AI



HPC のシミュレーション
にとって代わる AI

第２世代
インテル® Xeon®
スケーラブル・プロセッサー



インテル® Xeon® Platinum 9200 プロセッサー	インテル® Xeon® Platinum 8200 プロセッサー	インテル® Xeon® Gold 6200 プロセッサー	インテル® Xeon® Gold 5200 プロセッサー	インテル® Xeon® Silver 4200 プロセッサー	インテル® Xeon® Bronze 3200 プロセッサー
新たなクラスの高度なパフォーマンス					

価値 | トップクラスのワークロード・パフォーマンス | 画期的なメモリー・イノベーション | 人工知能アクセラレーション | ハードウェア支援型セキュリティー | 俊敏性と使用率の向上

http://www.intel.co.jp/xeonscalable/

インテル® Xeon® プロセッサーさらなる発展

唯一、コンバージェンス向けに最適化されたデータセンター CPU

インテル® アドバンスト・ベクトル・エクステンション 512
インテル® ディープラーニング・ブースト (インテル® DL ブースト)
インテル® Optane™ DC パーシステント・メモリー

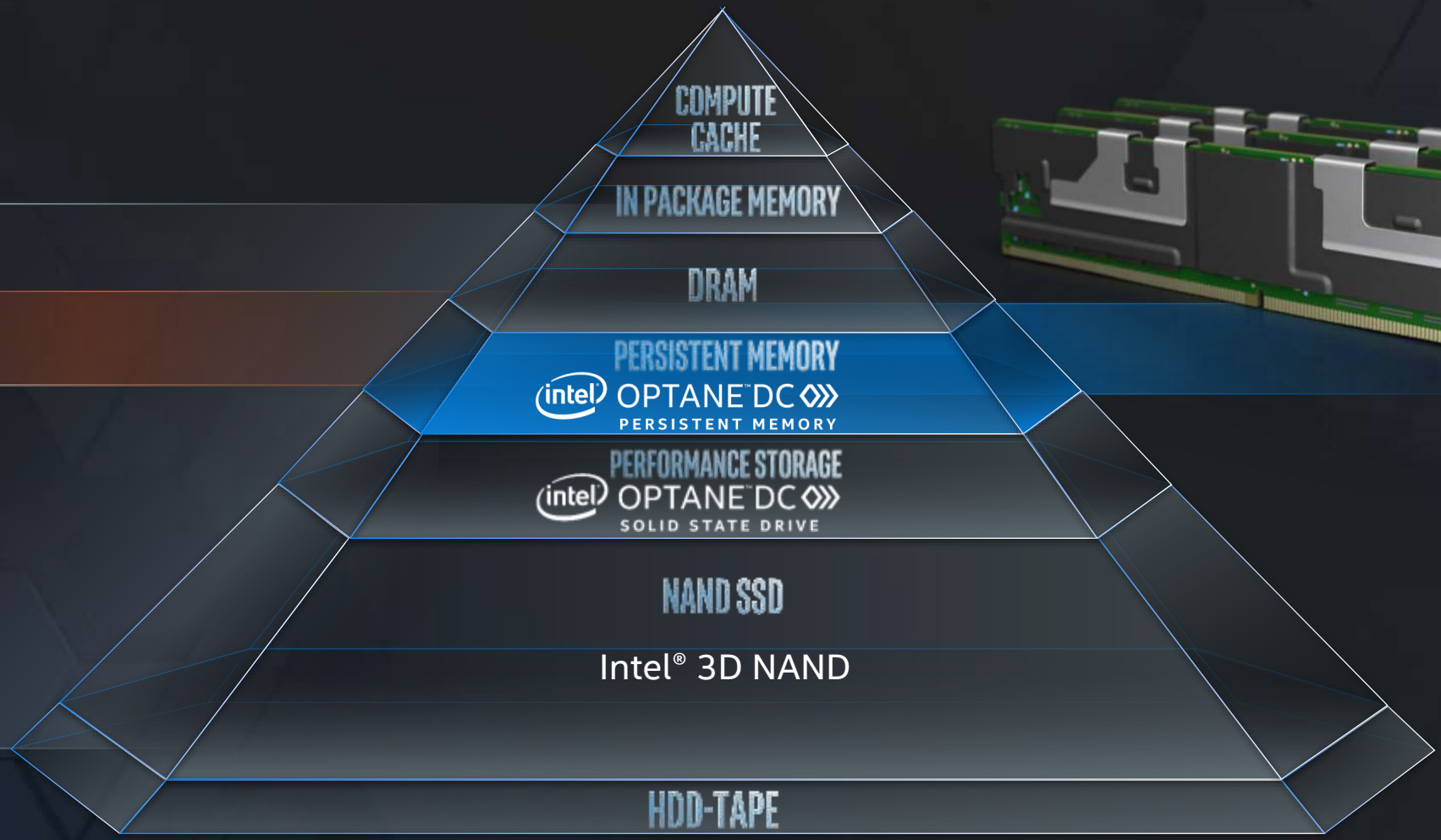
2019年	2020年	2021年
CASCADE LAKE 14NM AI の新たな加速化 (VNNI) メモリーストレージの新しい階層	**COOPER LAKE** 14NM 次世代インテル® DL ブースト (BFLOAT16) **ICE LAKE** 10NM 現在サンプルを出荷開始	**SAPPHIRE RAPIDS** 次世代テクノロジー

業界最先端のパフォーマンス

intel

MEMORY AND STORAGE HIERARCHY GAPS

THE CAPACITY GAP



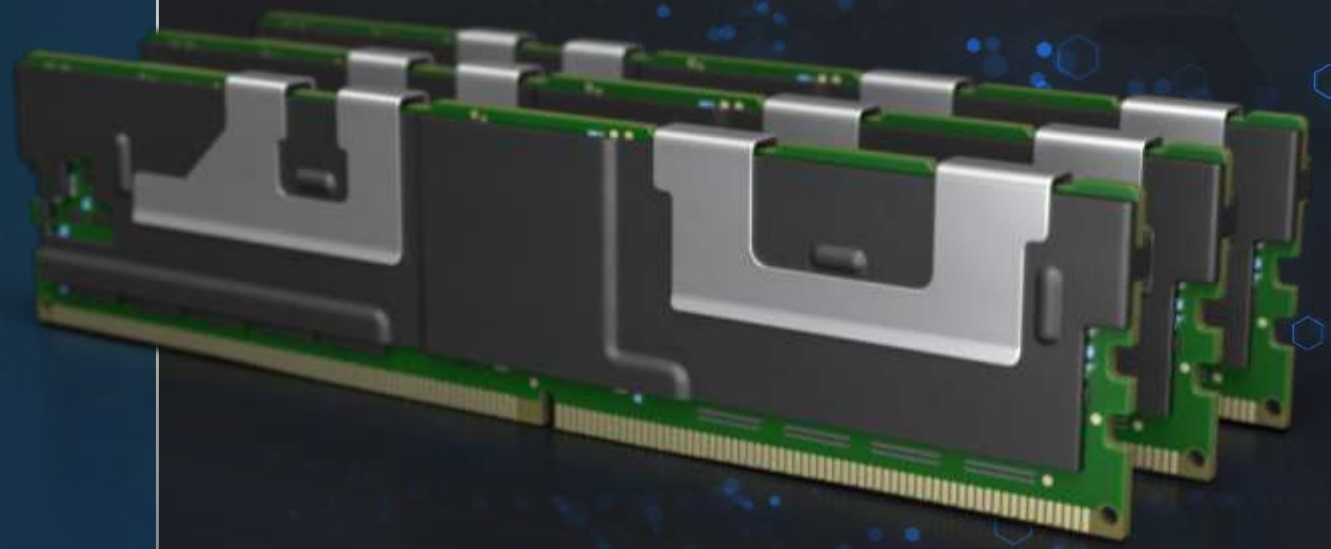
MEMORY

CAPACITY GAP

STORAGE

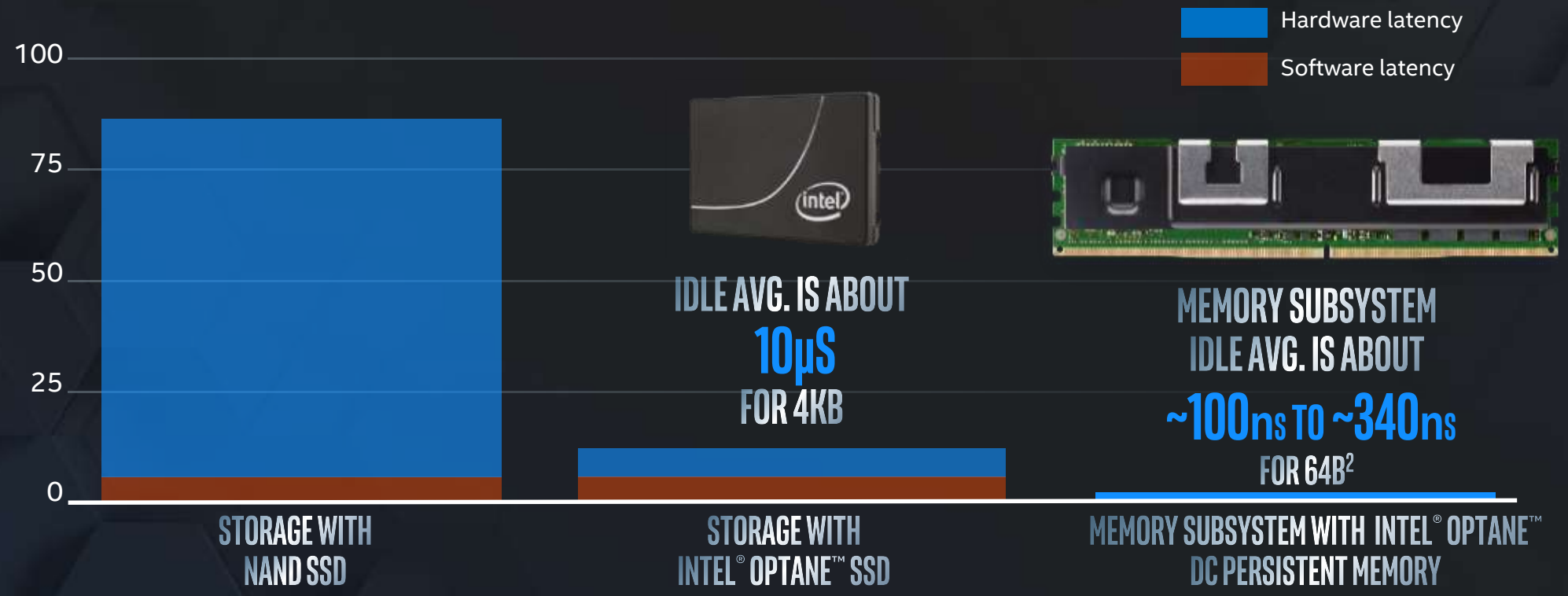
intel OPTANE™ DC PERSISTENT MEMORY

- **BIG AND AFFORDABLE MEMORY**
  128, 256, 512GB MODULES
  DDR4 PIN COMPATIBLE
- **BYTE ADDRESSABLE**
  DIRECT LOAD/STORE ACCESS
- **HIGH PERFORMANCE STORAGE**
  NATIVE PERSISTENCE
- **HIGH RELIABILITY AND SECURITY**
- **TWO OPERATIONAL MODES**



MORE TO BE GAINED BY BEING ON MEMORY BUS

IDLE AVERAGE RANDOM READ LATENCY[1]

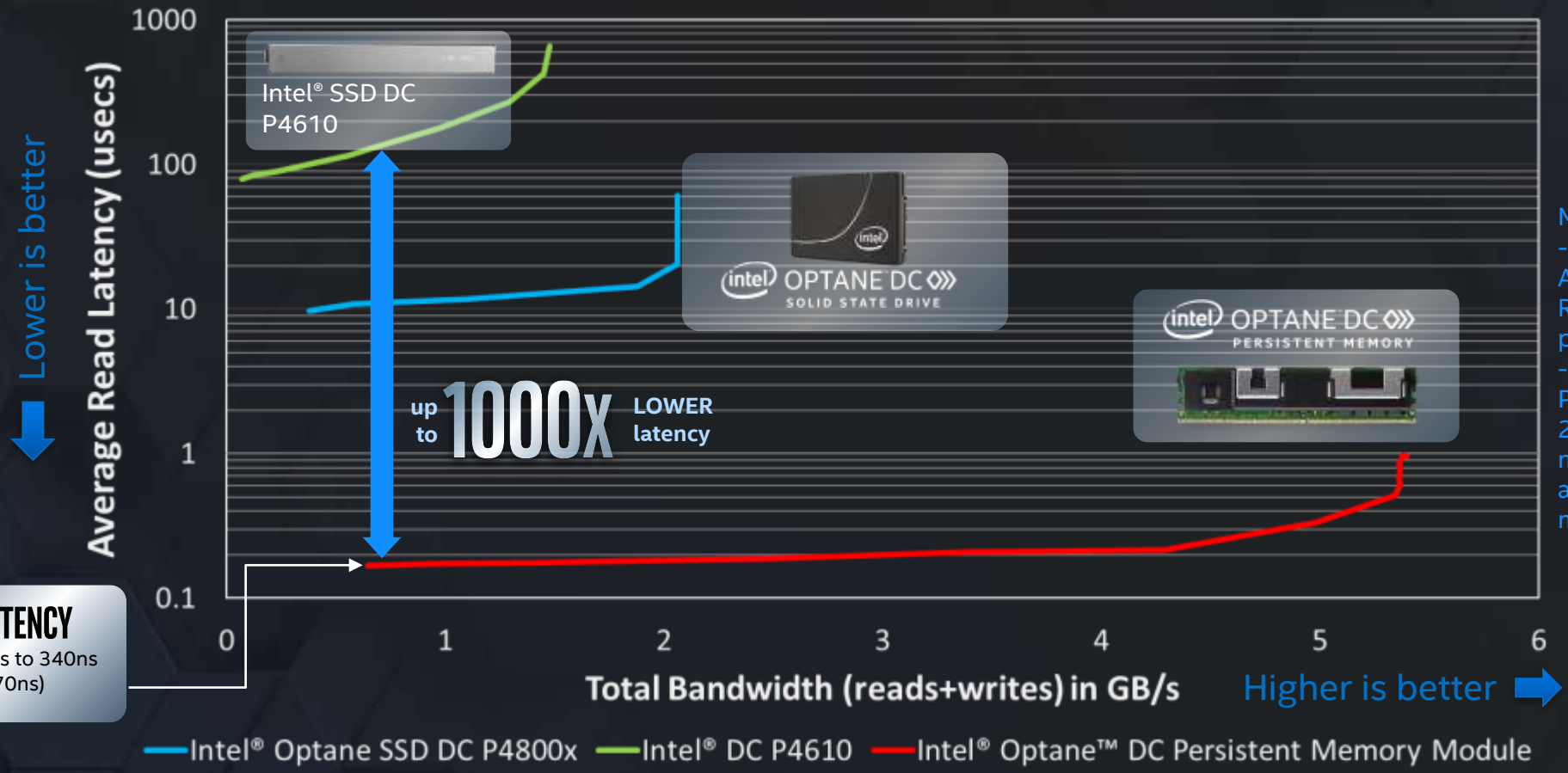


See Aopendix J
Performance results are based on testing as of July 24, 2018 set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

PERFORMANCE

Latency vs. Load

(70Read/30Write Random, 4kB for SSD and 256B for Memory)



READ IDLE LATENCY
Ranges from 180ns to 340ns
(vs. DRAM ~70ns)

Measurement notes:
- For SSDs Random 4KB Accesses over entire SSD, Read latency measured per 4kB access
- For Intel® Optane™ DC Persistent Memory Random 256B accesses Over entire module 256B random accesses w/ read latency measured per 64B access

See Appendix K
Performance results are based on testing as of February 22, 2019 set forth in the Configurations 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.

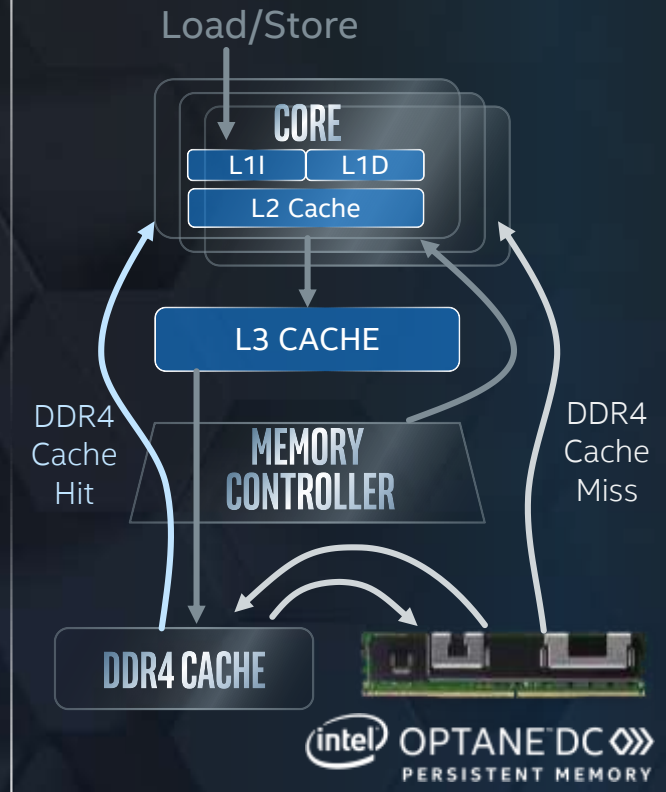
PERFORMANCE DETAILS

- Intel® Optane™ DC persistent memory is programmable for different power limits for power/performance optimization
  - 12W – 18W, in 0.25 watt granularity - for example: 12.25W, 14.75W, 18W
  - Higher power settings give best performance
- Performance varies based on traffic pattern
  - Contiguous 4 cacheline (256B) granularity vs. single random cacheline (64B) granularity
  - Read vs. writes

Granularity	Traffic	Module	Bandwidth
256B (4x64B)	Read	256GB, 18W	8.3 GB/s
256B (4x64B)	Write		3.0 GB/s
256B (4x64B)	2 Read/1 Write		5.4 GB/s
64B	Read		2.13 GB/s
64B	Write		0.73 GB/s
64B	2 Read/1 Write		1.35 GB/s

TWO OPERATIONAL MODES

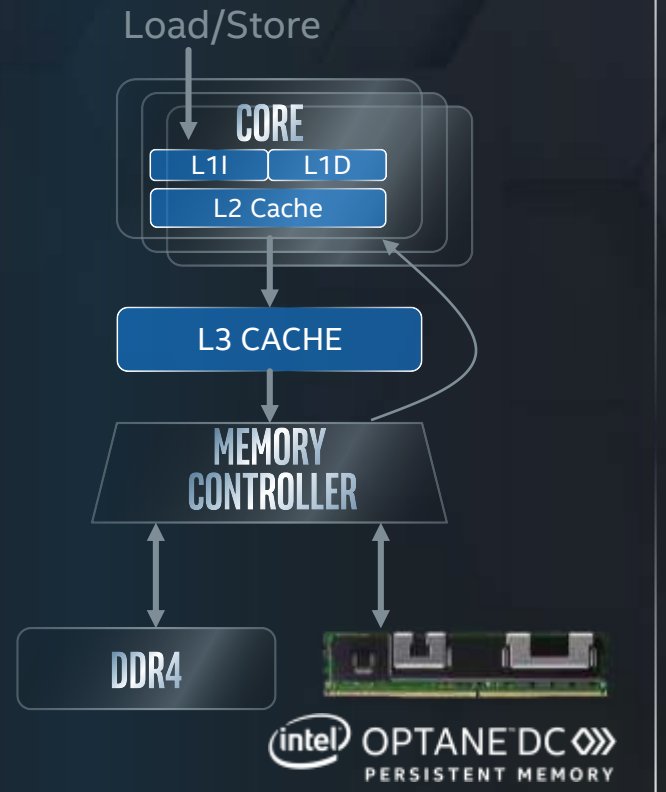
MEMORY MODE



HW MANAGED ⟷ DATA PLACEMENT ⟷ SW MANAGED

VOLATILE ⟷ PERSISTENCY ⟷ PERSISTENT

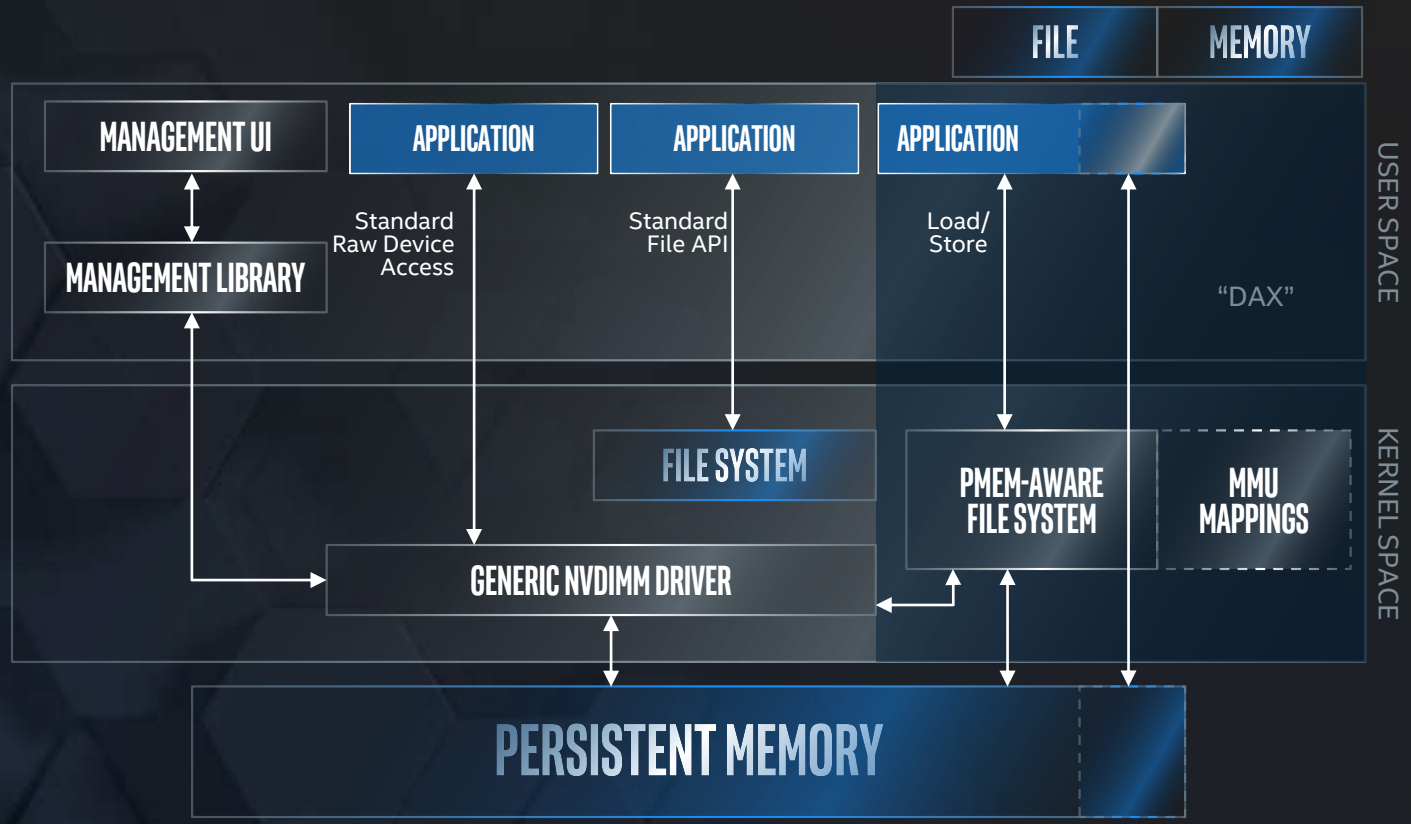
APPLICATION DIRECT MODE



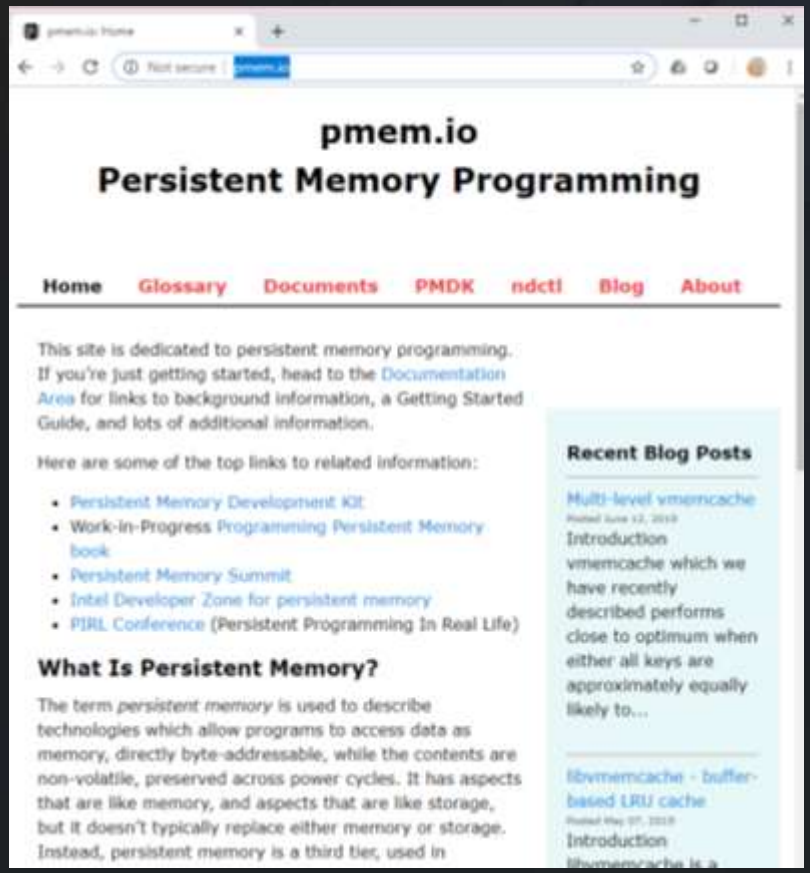
intel

PERSISTENCY: EARLY ENGAGEMENT WITH INDUSTRY

PERSISTENT MEMORY PROGRAMING MODEL
developed through SNIA



PERSISTENT MEMORY DESIGN KIT (PMDK)
available on http://pmem.io



A STRONG MEMORY & STORAGE FUTURE

	TODAY	FUTURE		
	2ND GEN INTEL XEON SCALABLE (CASCADE LAKE)	COOPER LAKE / ICE LAKE	SAPPHIRE RAPIDS	FUTURE INTEL XEON PROCESSOR
intel OPTANE DC PERSISTENT MEMORY	APACHE PASS	BARLOW PASS	3RD GEN DC PERSISTENT MEMORY	4TH GEN DC PERSISTENT MEMORY
intel OPTANE DC SOLID STATE DRIVE	Intel® SSD DC P4800X (COLDSTREAM)	ALDER STREAM	NEXT GENERATION	NEXT GENERATION
INTEL® 3D NAND SSD	INTEL® SSD DC P46XX/P45XX	CLIFFDALE-R/ARBORDALE + (96-L, 144-L)	NEXT GENERATION	NEXT GENERATION

For each processor above, Intel® Optane™ DC Persistent Memory will be supported on select SKUs

BARLOW PASS OVERVIEW – 2ND GENERATION

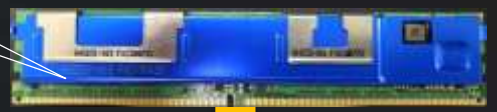
Barlow Pass will be productized with a **Blue colored** Full DIMM Heat spreader for better Identification in the data center.

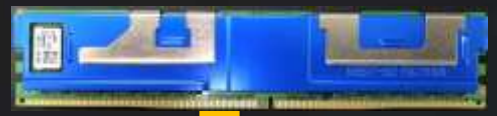
BARLOW PASS
Intel® Optane™ DC persistent memory

15W TDP

3200 MT/S DDR4

UP TO 15% BW Improvement vs AEP





WHITLEY
2 SOCKET PLATFORM

8CH Memory

3200 MT/S DDR4

4TB BPS Per Socket

CEDAR ISLAND
4 SOCKET PLATFORM

6CH Memory

2933 MT/S[1] DDR4

3TB BPS Per Socket

Legend

BPS | DDR4

1. For Cedar Island, 2933 2DPC & 3200 1DPC is POR. Stretch target is 3200 2DPC

INTEL OPTANE DC DESIGN IN ADVOCACY

STORAGE

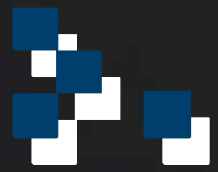
- Write Buffers
- Vaulting
- PMoF
- Caching Layers
- Meta-Data
- Hi-Perf Direct Attached Storage

CLOUD & VMs

- VM and Container Density
- Extended VM Memory Capacity
- Application Density
- Large Memory Java
- Compute-side Storage and Cache

DATABASE

- DB Memory Buffer Pool
- In-Memory DB
- Mixed Logging and memory
- User-Density
- Fast Restart
- Direct-attached DB Storage
- Dedicated Logging



AI/ANALYTICS

- Real Time Analytics
- Machine Learning Analytics
- In-Memory Analytics
- Off-Heap Memory

HPC

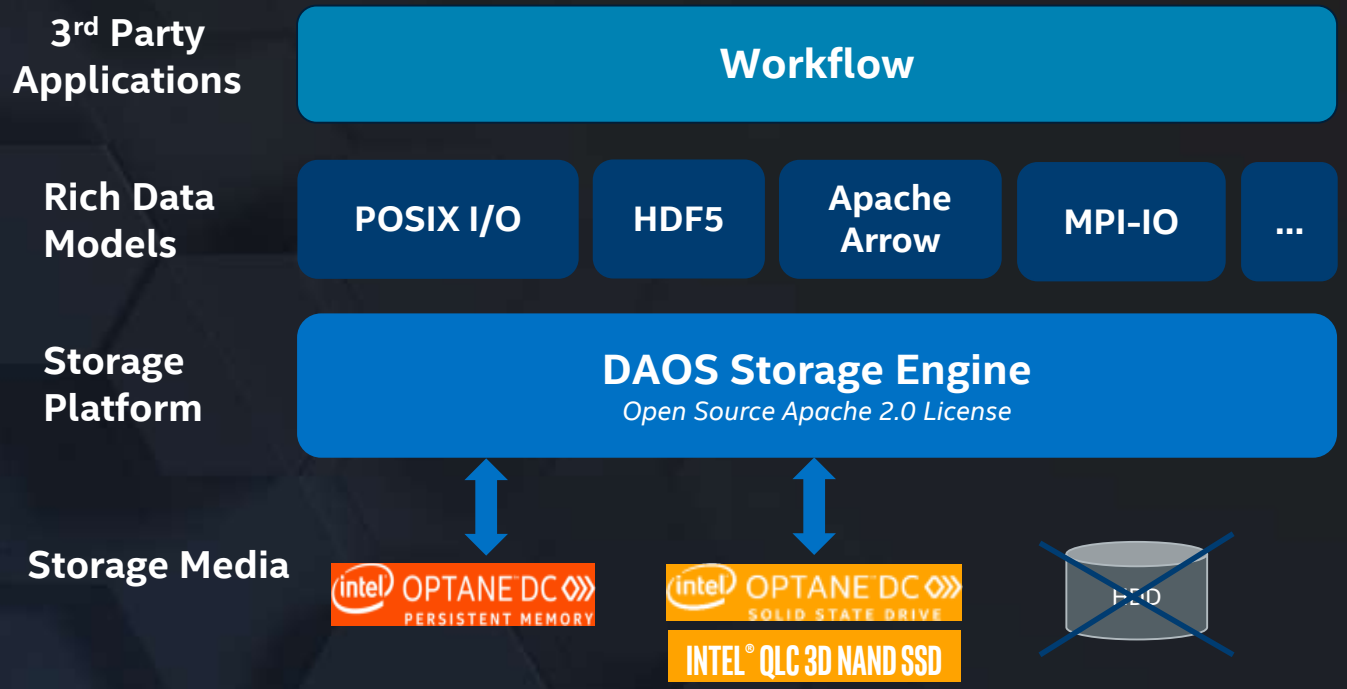
- Larger Memory Nodes
- Fast Checkpoints
- Memory Converged Infrastructure
- Fast Restart
- Scratch & IO Nodes

COMMS

- Content Delivery Network (CDN)
- In-Memory DB: Analytics/AI, Mobility Mgmt
- Resilient VNFs
- Distributed Memory: Cloud Native

VNF – Virtual Network Function
PMoF - Persistent Memory over Fabric

DISTRIBUTED ASYNCHRONOUS OBJECT STORAGE



Benefits

- Built natively over **new userspace** PMEM/NVMe software stack
- **Rich** storage semantics
- High **throughput/IOPS @arbitrary** alignment/size
- **Fine-grained**, **low-latency** & True **zero-copy** I/Os
- **Scalable** communications
- **Software**-managed **redundancy**
- Rely on **COTS** hardware

intel

DAOS ARCHITECTURE

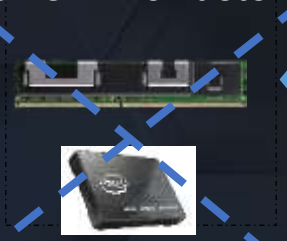
- High-latency communications
- P2P operations
- No HW acceleration

Conventional Storage Systems

Data & Metadata

Block Interface — Linux Kernel I/O

Intel® *3D-XPoint Storage*



Intel® *3D-NAND Storage*

HDD

- Low-latency, high-message-rate communications
- Collective operations & in-storage computing

DAOS Storage Engine

Metadata, low-latency I/Os & indexing/query

Memory Interface — PMDK

Intel® *3D-XPoint Storage*

Bulk data

NVMe Interface — SPDK

3D-NAND/XPoint Storage

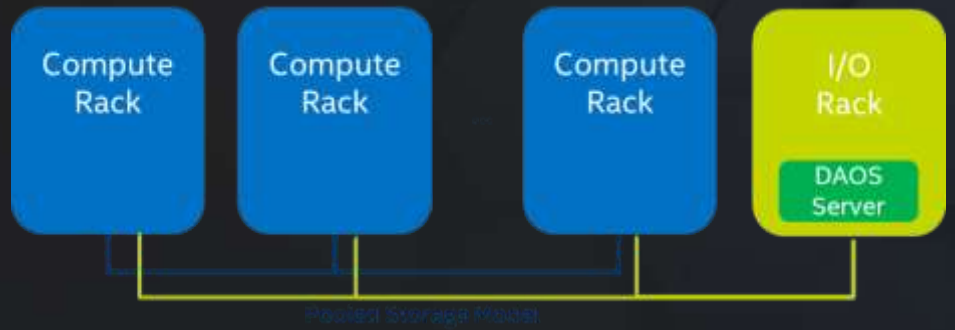
HDD

DEPLOYMENT OPTIONS

DAOS system can be deployed in two different ways:

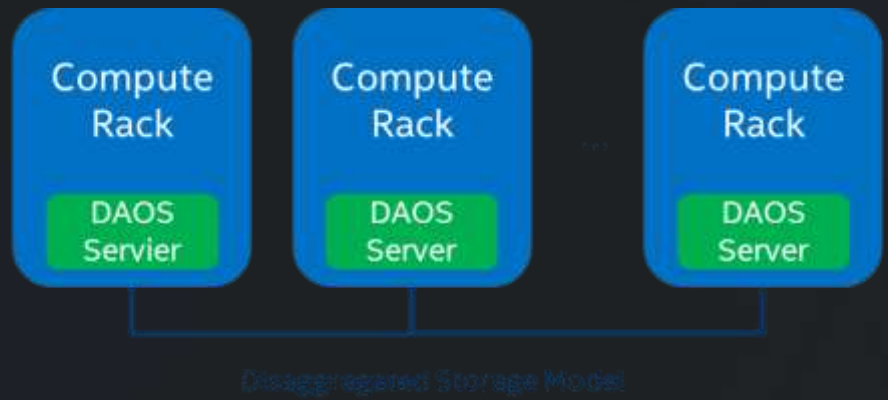
Pooled Storage Model :

The DAOS servers can run on dedicated storage nodes in separate racks. This is a traditional pool model where storage is uniformly accessed by all compute nodes. In order to minimize the number of I/O racks and to optimize floor space, this approach usually requires high density storage servers.



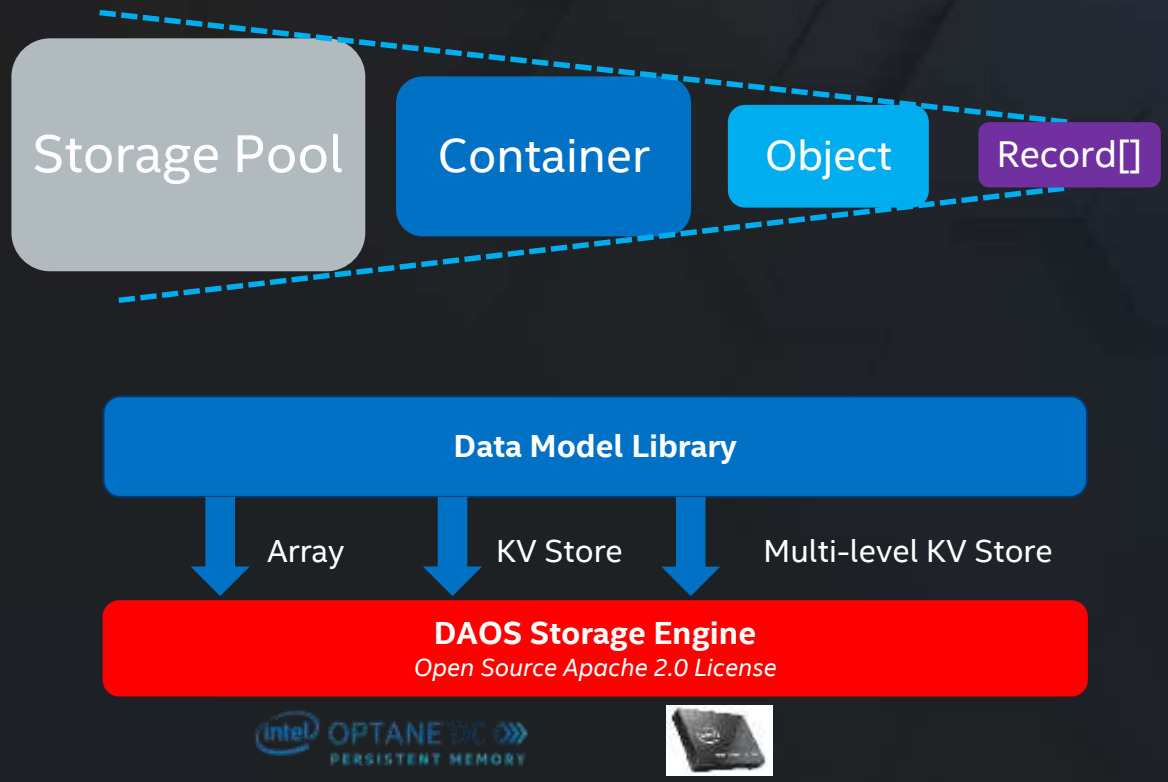
Disaggregated Storage Model :

In the disaggregated model, the storage nodes are integrated into compute racks and can be either dedicated or shared (e.g. in a hyper-converged infrastructure) nodes. The DAOS servers are thus massively distributed and storage access is non-uniform and must take locality into account.



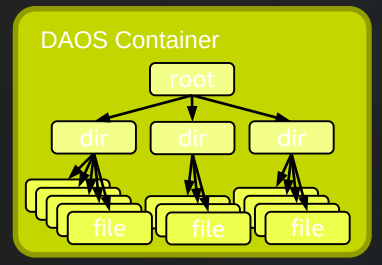
DAOS DATA MODEL

- Non-POSIX rich storage API as the new foundation
  - Scalable storage model suitable for both **structured & unstructured** data
    - key-value stores, multi-dimensional arrays, columnar databases, …
    - Accelerate data analytic/AI frameworks
  - **Non-blocking** data & metadata operations
  - **Extendable** through microservice architecture

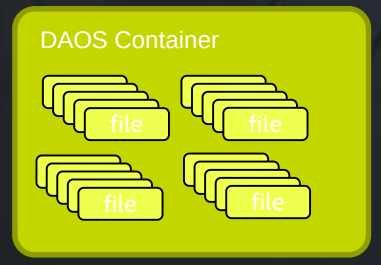


DATASET MANAGEMENT

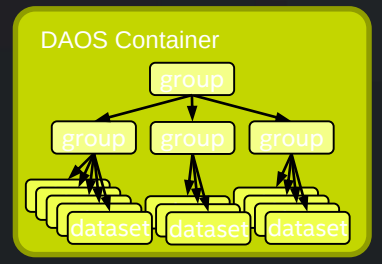
- Aggregate related datasets into manageable and coherent entities
  - Distributed consistency & automated recovery
  - Full Versioning
  - Simplified data management
    - Snapshot
    - Cross-tier Migration
    - Indexing



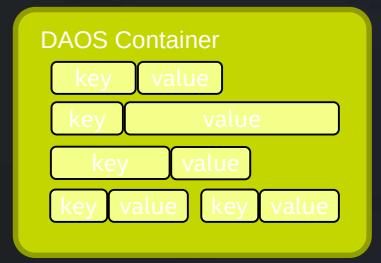
Encapsulated POSIX Namespace



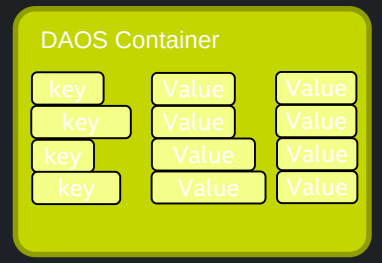
File-per-process



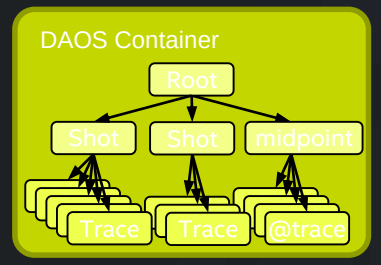
HDF5 « File »



Key-value store

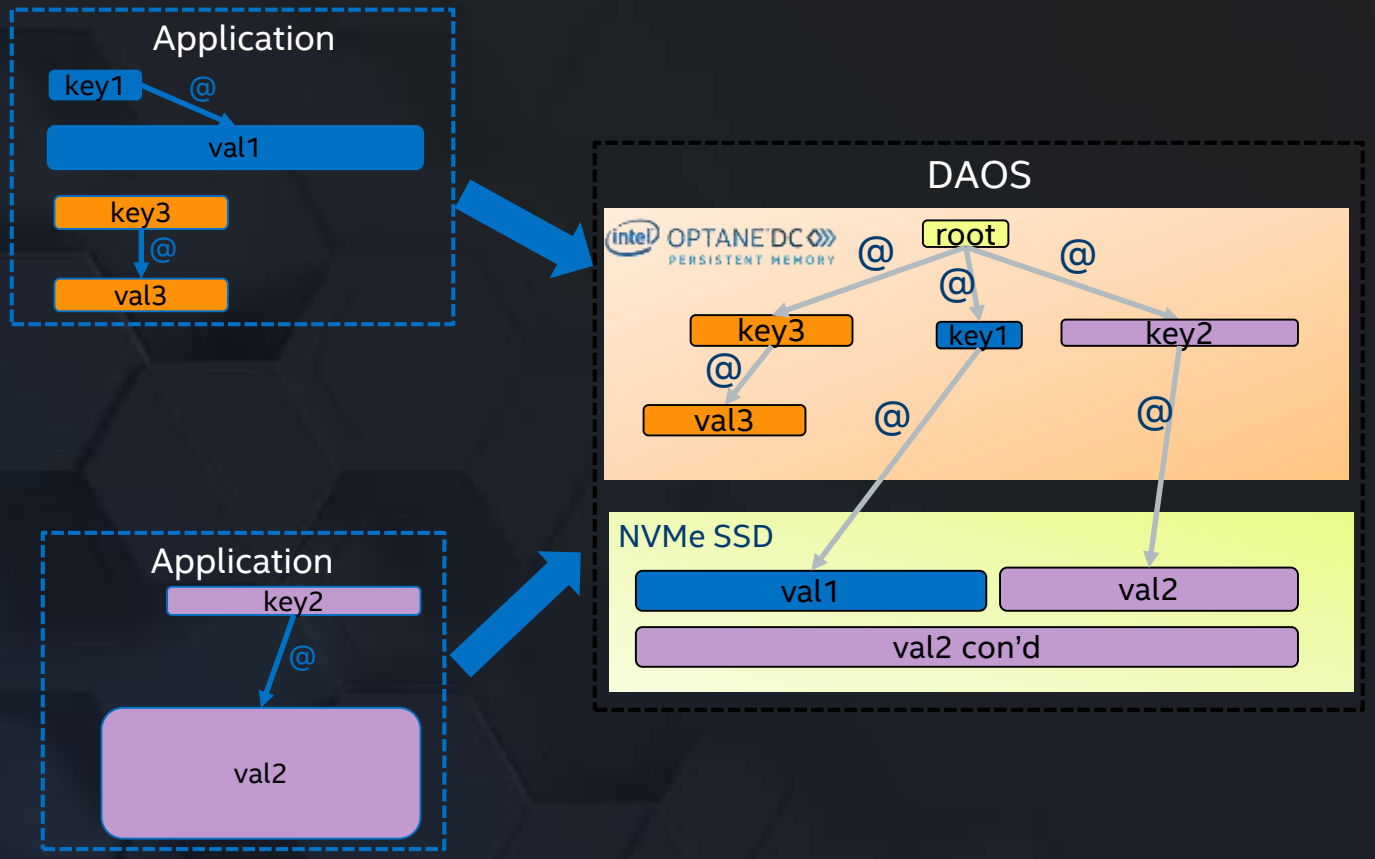


Columnar Database



SEGY

ADVANCED STORAGE API



Fast data retrieval

- Avoid file serialization and offset management
- Keys can be of any size/type
- Keys can be ordered with range query support

Scalable Insert & Fetch

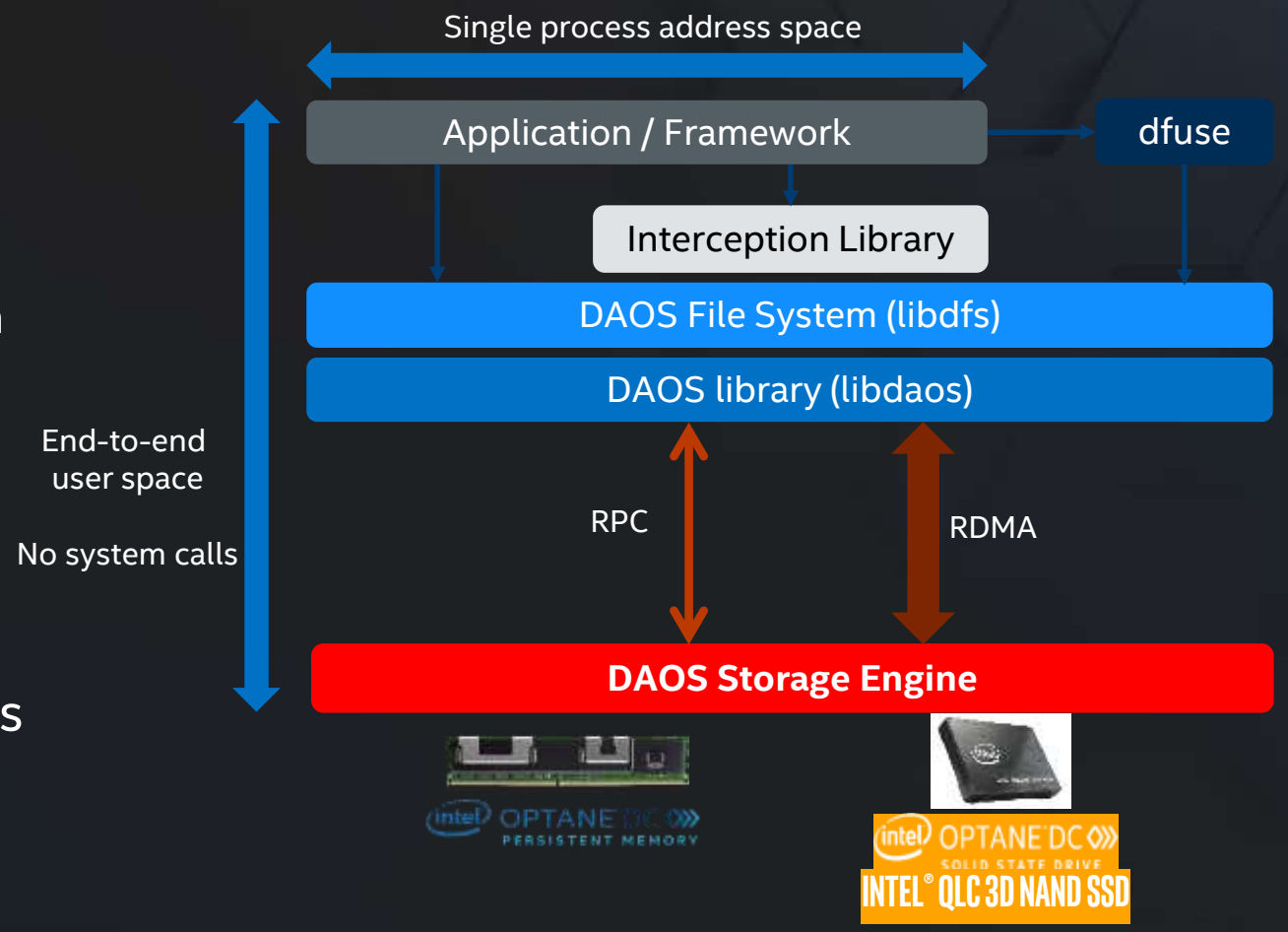
- Allow concurrent access/update
- Unconstrained by POSIX serialization
- Non-blocking
- Distributed transactions keep KV store always consistent

Data indexing

- Enable in-storage computing
- Query & custom index
- Data provenance

POSIX I/O SUPPORT

- DAOS File System (libdfs)
  - Encapsulated POSIX namespace
  - Application/framework can link directly with libdfs
    - ior/mdtest backend provided
    - MPI-IO driver leveraging collective open
    - TensorFlow, …
- FUSE Daemon (dfuse)
  - Transparent access to DAOS
  - Involves system calls
- I/O interception library
  - OS bypass for read/write operations

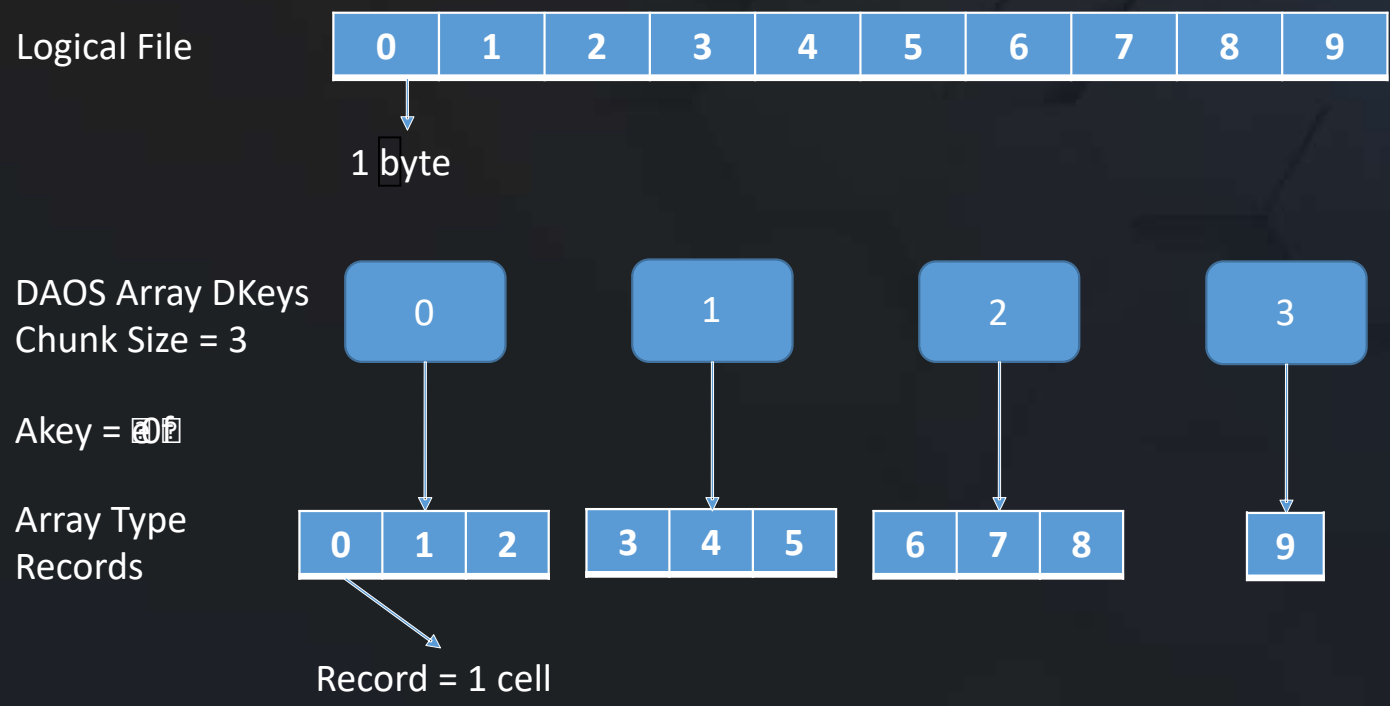


MPI-IO DRIVER FOR DAOS

The DAOS MPI-IO driver is implemented within the I/O library in MPICH (ROMIO).

- Added as an ADIO driver
- Portable to Open-MPI, Intel MPI, etc.
- https://github.com/daos-stack/mpich
- daos_adio branch
- PR to mpich master in review

- 1 MPI File = 1 DAOS Array Object



Application works seamlessly by just specifying the use of the driver by appending “daos:” to the path.

DAOS IN IO-500 2019-11

#1 in 10 Node Challenge


Register Log In

Virtual Institute for I/O

Search

Recent Changes Sitemap

You are here: Virtual Institute for I/O » IO500 » Lists » 2019-11 » **10 Node Challenge**

Ranked List | **10 Node Challenge** | Full List | Historic data | Student Cluster Co...

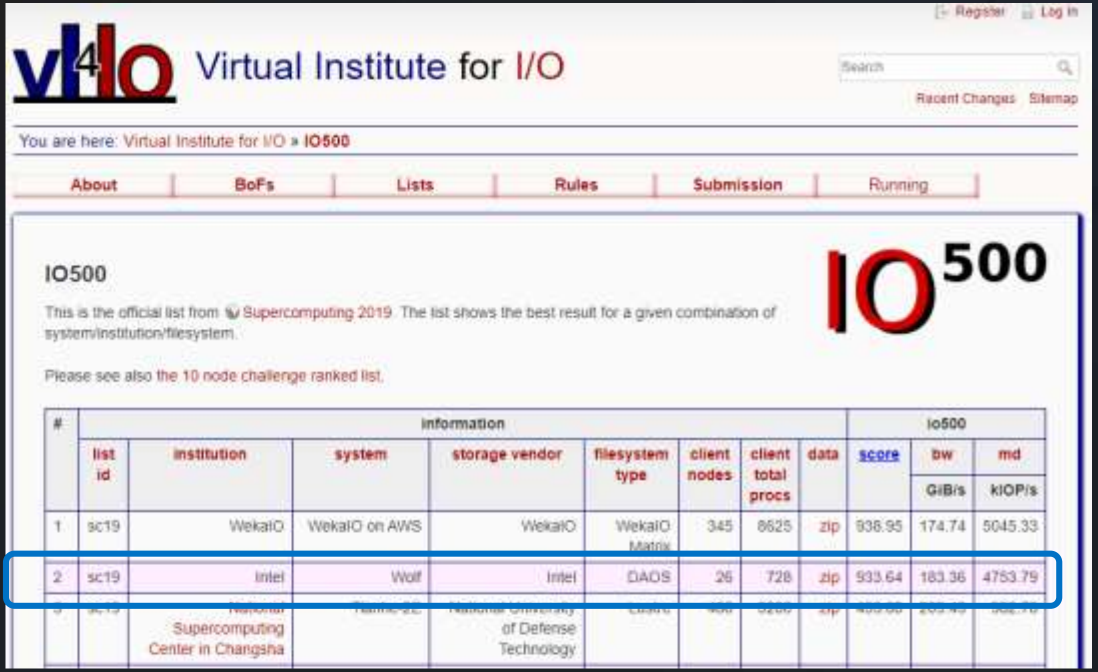
IO500

10 Node Challenge

This is the official list from Supercomputing 2019 for the 10 Node Challenge. The list shows the best result for a given combination of system/institution/filesystem qualifying for the 10 Node Challenge.

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw GiB/s	md kIOP/s
1	sc19	Intel	Wolf	Intel	DAOS	10	310	zip	516.41	123.89	2152.46
2	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
3	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip	156.51	56.22	435.76

#2 in Ranked List


Register Log In

Virtual Institute for I/O

Search

Recent Changes Sitemap

You are here: Virtual Institute for I/O » **IO500**

About | BoFs | Lists | Rules | Submission | Running

IO500

IO500

This is the official list from Supercomputing 2019. The list shows the best result for a given combination of system/institution/filesystem.

Please see also the 10 node challenge ranked list.

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw GiB/s	md kIOP/s
1	sc19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
2	sc19	Intel	Wolf	Intel	DAOS	26	728	zip	933.64	183.36	4753.79
[illegible]	[illegible]	National Supercomputing Center in Changsha	[illegible]	National University of Defense Technology	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

intel

DAOS: PRIMARY STORAGE ON AURORA

Aurora DAOS configuration

- Capacity: 230PB
- Bandwidth: >25TB/s

"Combined in Aurora, the Intel compute system, Cray Slingshot network, and the Intel DAOS storage open new possibilities for accelerating the scientific research needed to solve critical human challenges such as cancer and disease. DAOS enables the creation of new storage data models tailored specifically to applications like the Cancer Distributed Learning Environment (CANDLE) which provide a powerful platform to advance a wide array of scientific challenges using deep learning."

– Rick Stevens, Associate Laboratory Director for Computing, Environment and Life Sciences

"The Argonne Leadership Computing Facility is excited to be the first major production deployment of the DAOS storage system as part of Aurora, an US exascale system coming in 2021. As designed, it will provide us unprecedented levels of metadata operation rates and extremely high bandwidth for I/O intensive workloads."

– Susan Coghlan, ALCF-X Project Director/Exascale Computing Systems Deputy Director

intel

DAOS COMMUNITY ROADMAP

DAOS:
- NVMe & DCPMM support
- Per-pool ACL
- UNS in DAOS via dfuse
- Replication & self-healing (preview)

Application Interface:
- MPI-IO Driver
- HDF5 DAOS Connector
- Basic POSIX I/O support
- Spark

DAOS:
- End-to-end data integrity
- Per-container ACL
- Improved control plane
- Lustre/UNS integration
- Replication & self-healing
- Erasure code (preview)
- Conditional updates

Application Interface:
- POSIX I/O with conditional update support

DAOS:
- Online server addition
- Advanced control plane
- Multi OFI provider support

Application Interface:
- POSIX data mover
- Async HDF5 operations over DAOS

DAOS:
- Erasure code
- Telemetry & per-job statistics
- Distributed transactions

Application Interface:
- POSIX I/O with distributed transaction support
- HDF5 data mover
- Container parking/serialization

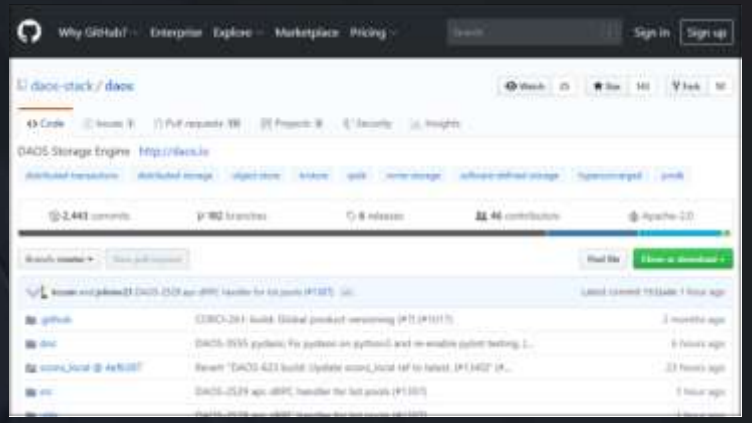
DAOS:
- Progressive layout
- Placement optimizations
- Checksum scrubbing

DAOS:
- Catastrophic recovery tools

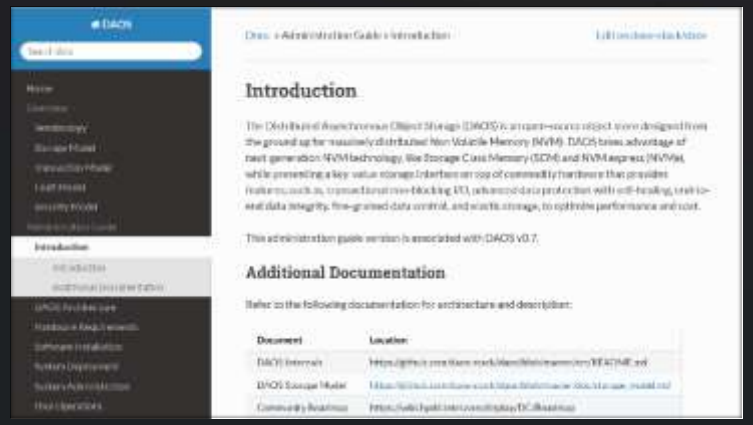
NOTE: All information provided in this roadmap is subject to change without notice.

intel

RESOURCES



[Source Code on GitHub]



[Admin Guide]

[DAOS Solution Brief]

Community mailing list on Groups.io

daos@daos.groups.io

Support

https://jira.hpdd.intel.com

LEGAL DISCLAIMER

All information provided here is subject to change without notice.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

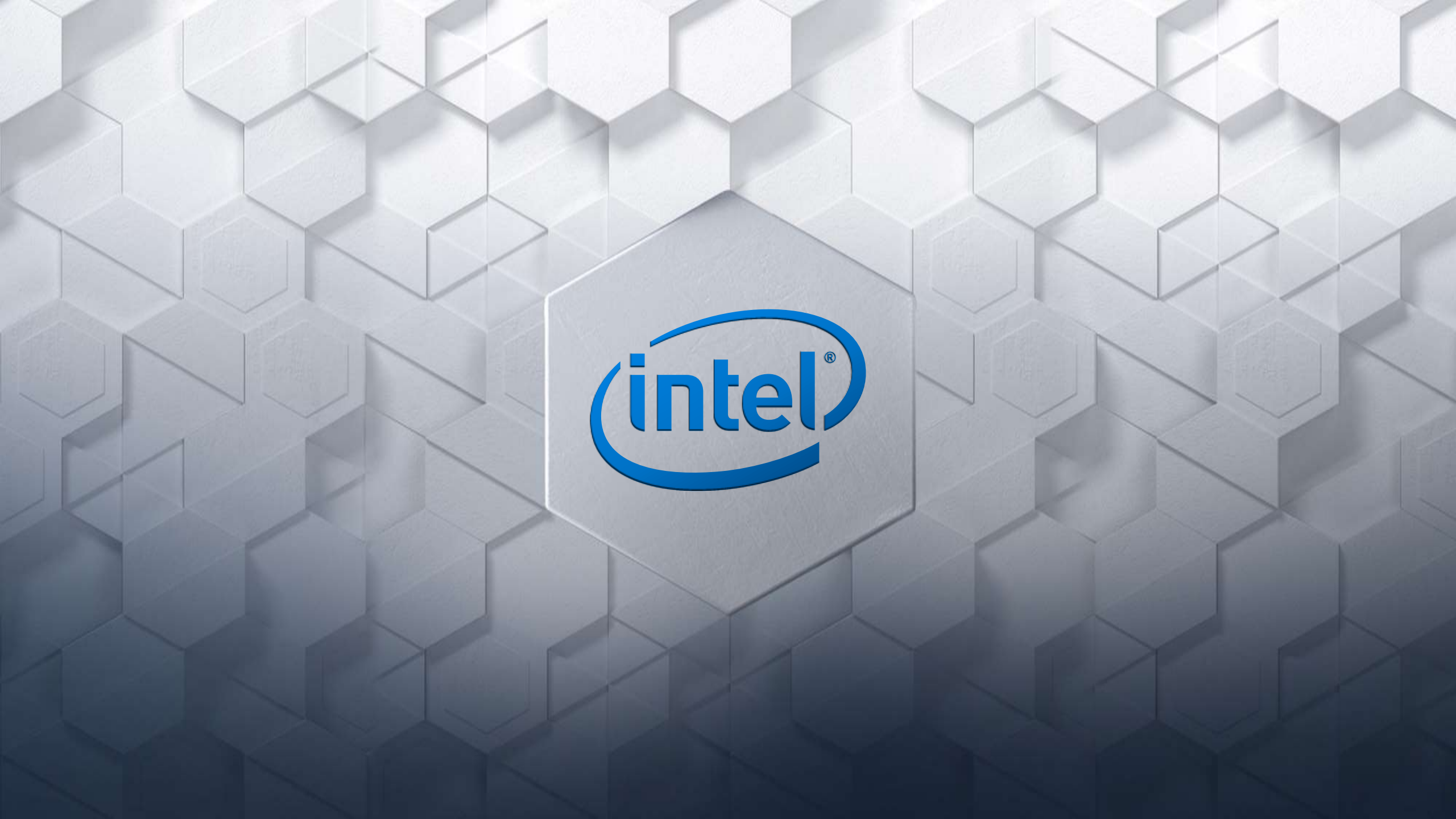
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

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.

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 data. You should review this content, consult other sources, and confirm whether referenced data are accurate.

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.


intel