STORAGE AND MEMORY HIERARCHY IN HPC: NEW PARADIGM AND NEW SOLUTIONS WITH INTEL

DR. JEAN-LAURENT PHILIPPE
Senior EMEA HPC Technical Specialist
Intel Data Center Group

LEGAL DISCLAIMER

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

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.

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.

Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system.

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.

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 the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

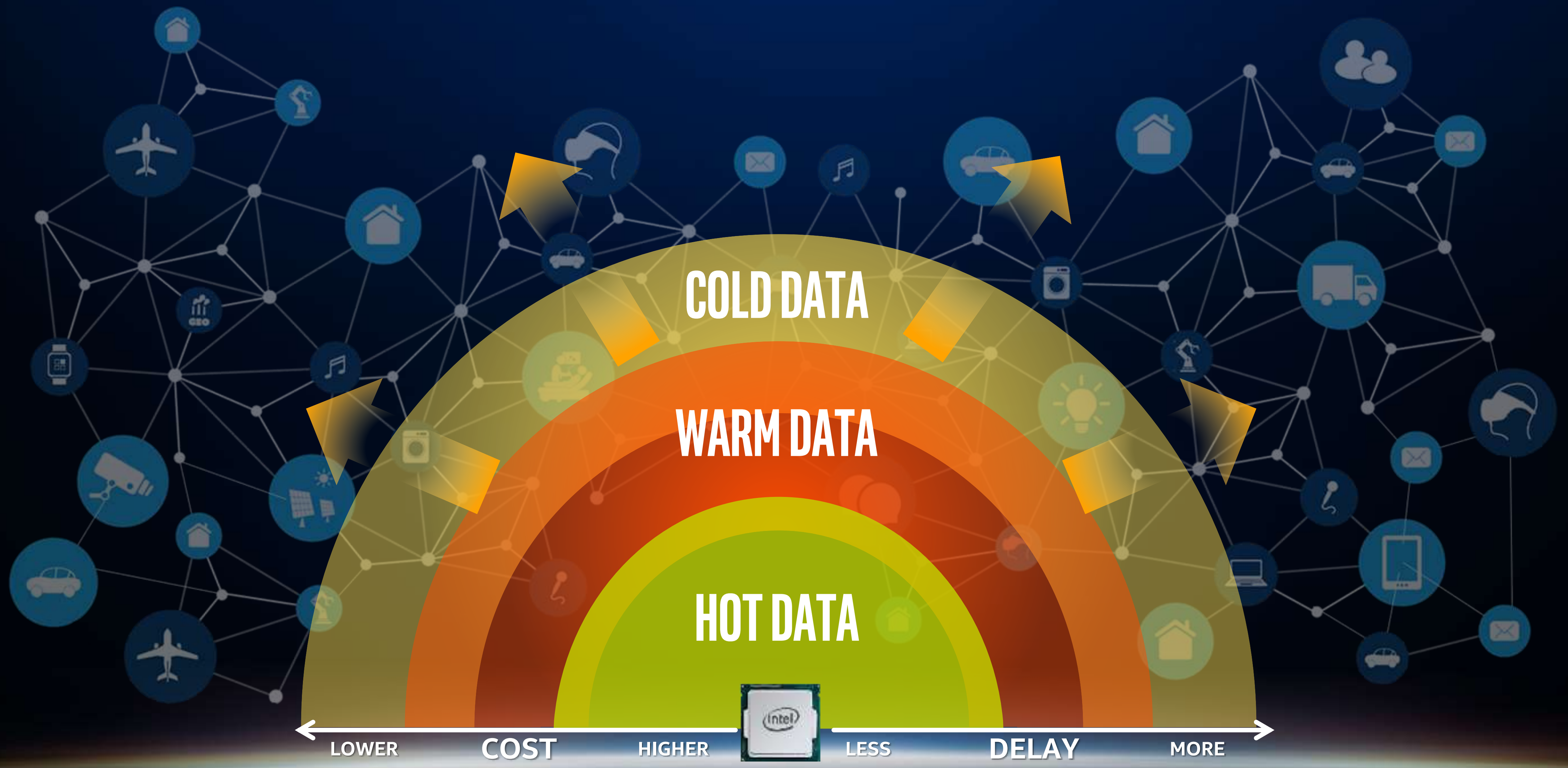
WE ARE IN A DATA-CENTRIC WORLD

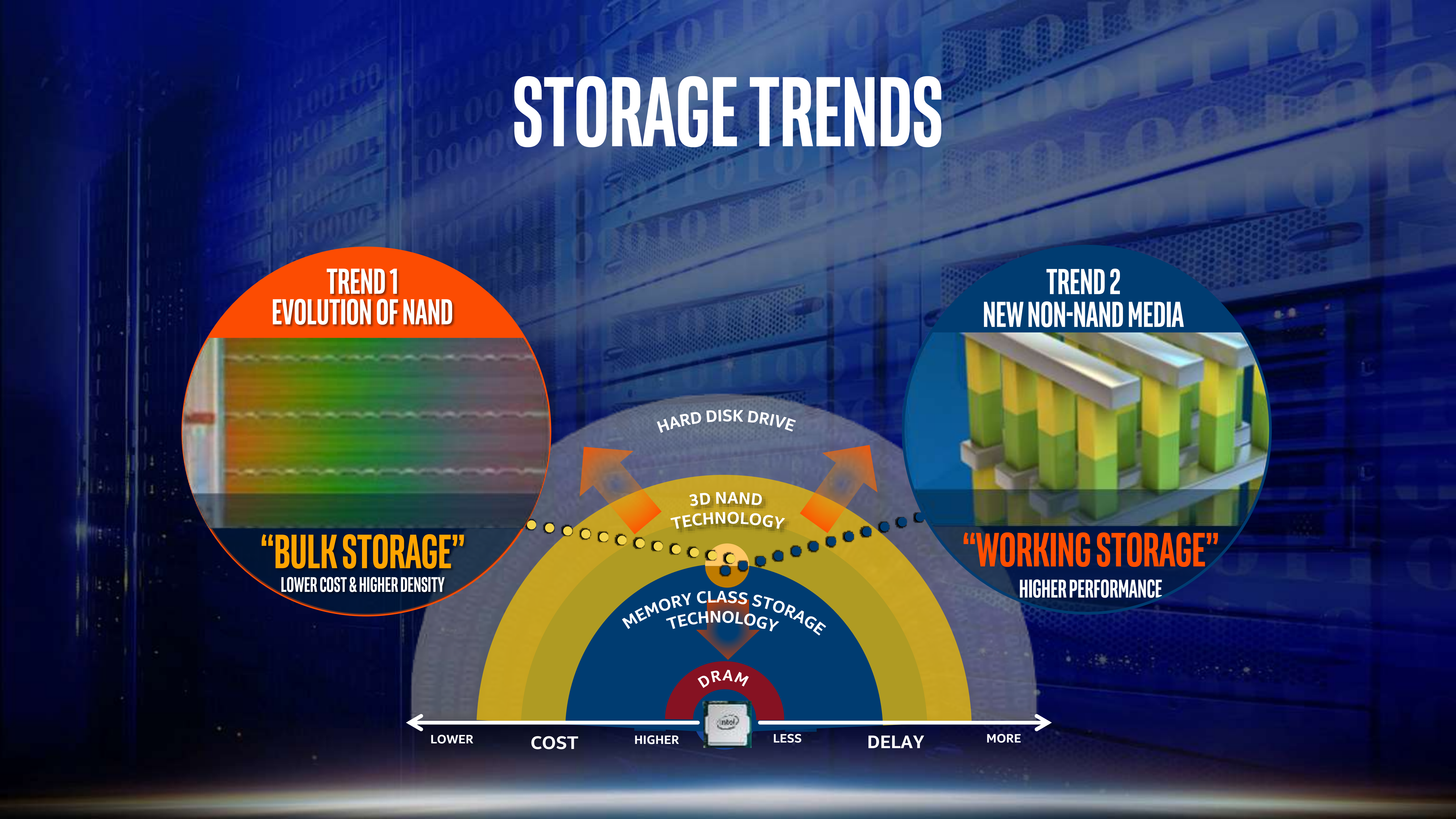
TODAY ALL DATA MUST BE STORED, PROCESSED, ANALYZED & MONETIZED

10x GROWTH

Source: https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

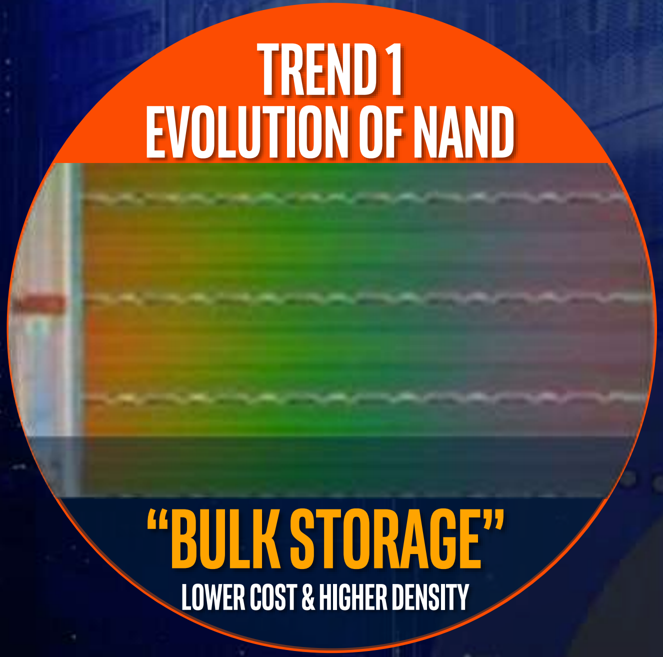
DATA IS STORED BY DIFFERENT TIERS

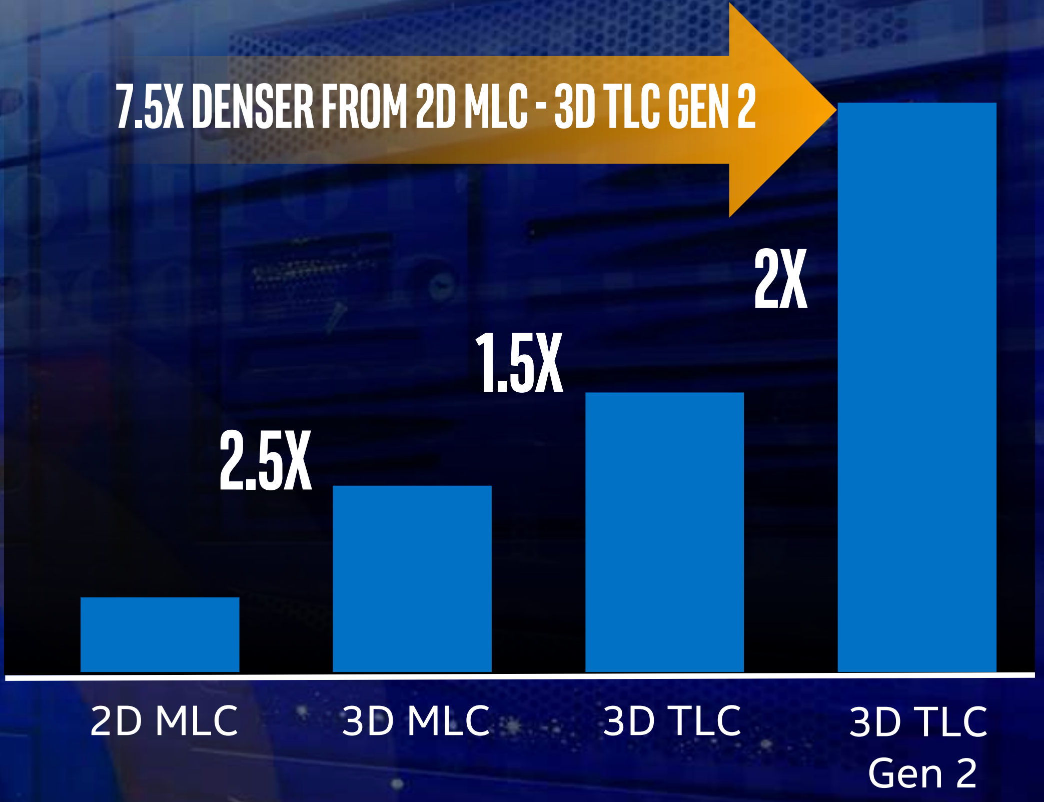



STORAGE TRENDS
TREND 1
EVOLUTION OF NAND
"BULK STORAGE"
LOWER COST & HIGHER DENSITY
TREND 2
NEW NON-NAND MEDIA
"WORKING STORAGE"
HIGHER PERFORMANCE
HARD DISK DRIVE
3D NAND TECHNOLOGY
MEMORY CLASS STORAGE TECHNOLOGY
DRAM
intel
LOWER
COST
HIGHER
LESS
DELAY
MORE

HIGH CAPACITY TRADEOFFS

COST AND DENSITY FOR ENDURANCE AND PERFORMANCE

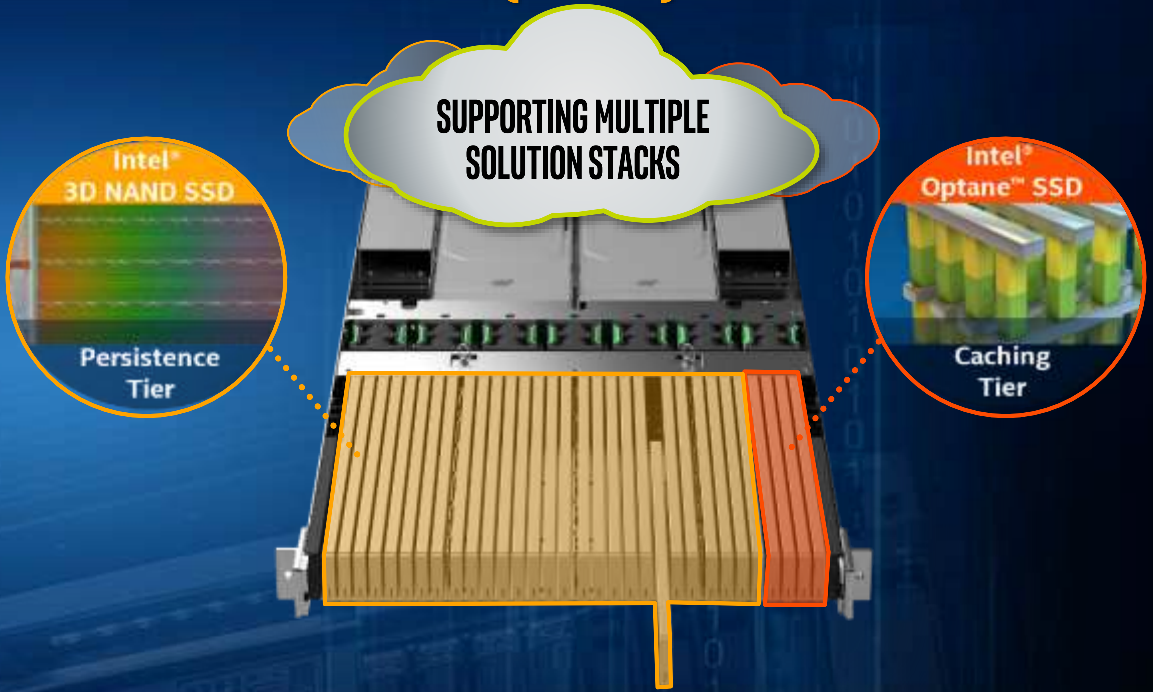




Source: Intel internal analysis of areal density of die sizes measured at Gb/mm2

OEM PLATFORM INNOVATION

ENTERPRISE DATACENTER SSD FORM FACTORS (EDSFF)



1PB IN 42U

WITH 2 TB HDDs

1PB IN 1U

WITH INTEL® 3D NAND SSDs

INTEL® OPTANE™ TECHNOLOGY

PERFORMANCE AND ENDURANCE FOR DENSITY AND COST/GB

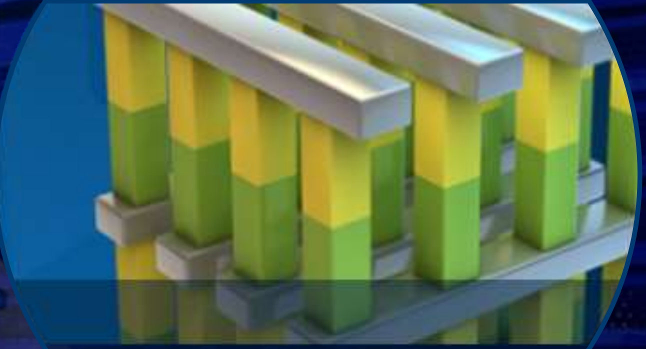
ENDURANCE

QoS

THROUGHPUT (IOPS)

LATENCY

TREND 2
NEW NON-NAND MEDIA



“WORKING STORAGE”

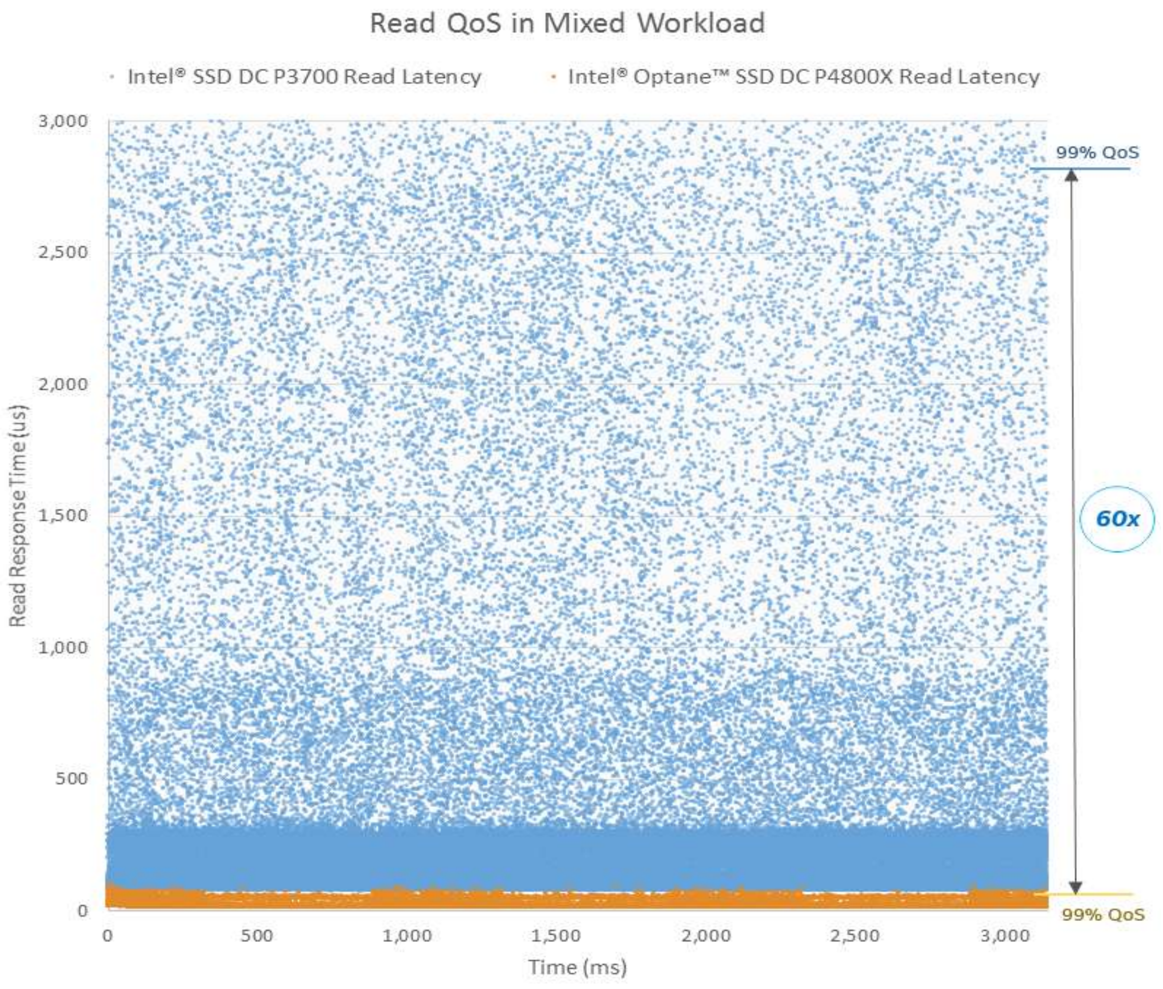
HIGHER PERFORMANCE & ENDURANCE

World's Most Responsive Data Center SSD[1]

Delivering an **industry leading combination of low latency, high endurance, QoS and high throughput**, the Intel® Optane™ SSD is the first solution to **combine the attributes of memory and storage**. This innovative solution is optimized to **break through storage bottlenecks** by providing a new data tier. It accelerates applications for **fast caching and storage, increasing scale per server** and reducing transaction cost. Data centers based on the latest Intel® Xeon® processors can now also **deploy bigger and more affordable datasets** to gain new insights from larger memory pools.

1. Responsiveness defined as average read latency measured at queue depth 1 during 4k random write workload. Measured using FIO 2.15. Common configuration - Intel 2U PCSD Server ("Wildcat Pass"), OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Intel drives evaluated - Intel® Optane™ SSD DC P4800X 375GB, Intel® SSD DC P3700 1600GB, Intel® SSD DC P4600 1600GB. Samsung drives evaluated – Samsung® SSD PM1725a, Samsung® SSD PM1725, Samsung® PM963, Samsung® PM953. Micron drive evaluated – Micron® 9100 PCIe® NVMe™ SSD. Toshiba drives evaluated – Toshiba® ZD6300. Test – QD1 Random Read 4K latency, QD1 Random RW 4K 70% Read latency, QD1 Random Write 4K latency using fio-2.15.

Predictably Fast Service

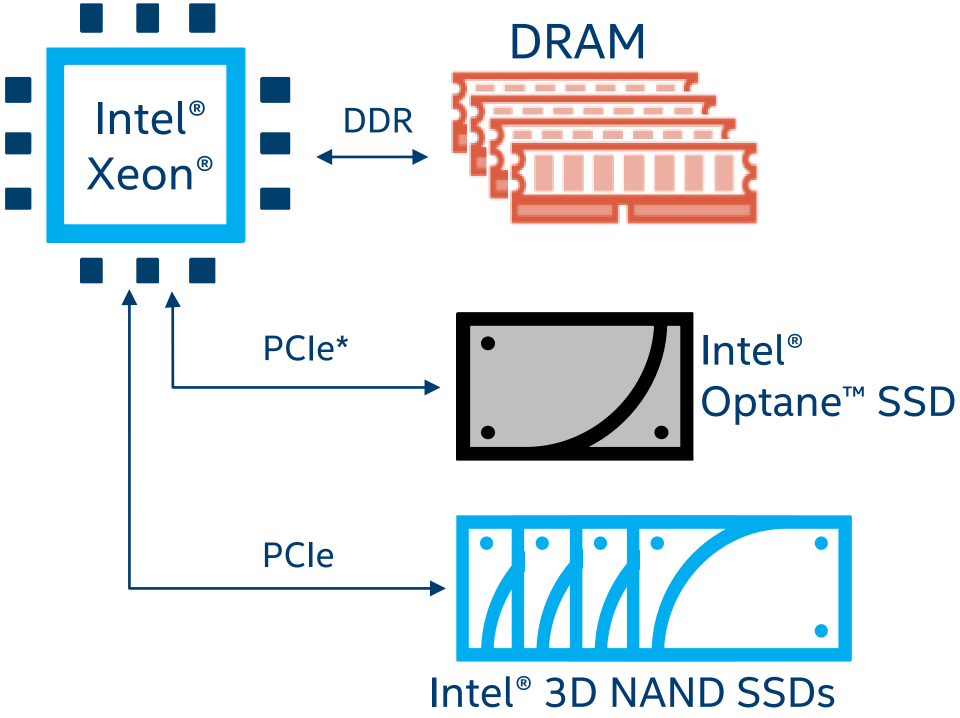


- up to **60X** better at 99% QoS[1]
- Ideal for critical applications with aggressive latency requirements

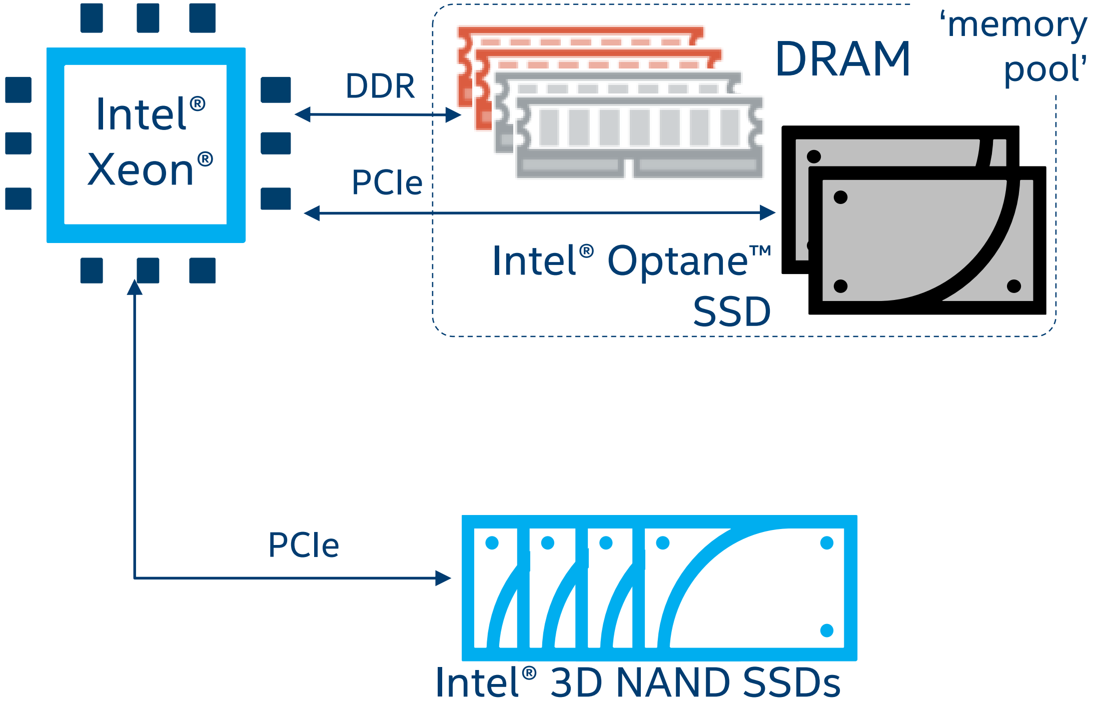
1. Common Configuration - Intel 2U PCSD Server ("Wildcat Pass"), OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Optane Configuration – Intel® Optane™ SSD DC P4800X 375GB. NAND Configuration – Intel® SSD DC P3700 1600GB. QoS – measures 99% QoS under 4K 70-30 workload at QD1 using fio-2.15.

Intel® Optane™ SSD Use Cases

Fast Storage and Cache



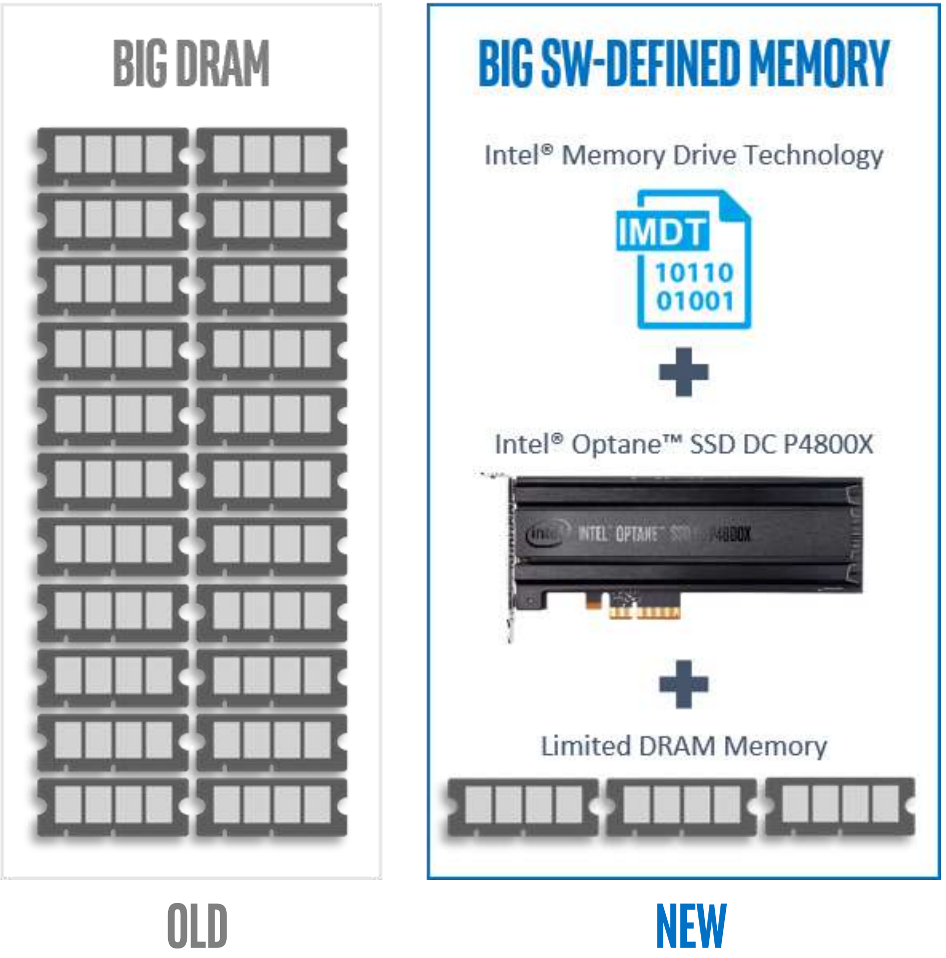
Extend Memory



*Other names and brands names may be claimed as the property of others

INTRODUCING INTEL® MEMORY DRIVE TECHNOLOGY

- Use Intel® Optane™ SSD DC P4800X **transparently as memory**
- Grow **beyond system DRAM capacity**, or **replace high-capacity DIMMs** for lower-cost alternative, with **similar performance**
- **Leverage storage**-class memory today!
  - **No change to software** stack: unmodified Linux* OS, applications, and programming
  - **No change to hardware**: runs bare-metal, loaded before OS from BIOS or UEFI
- **Aggregated single volatile memory pool**

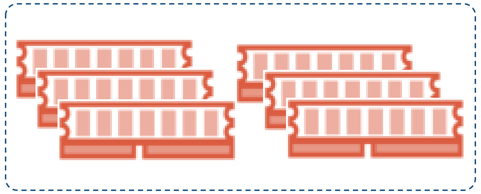


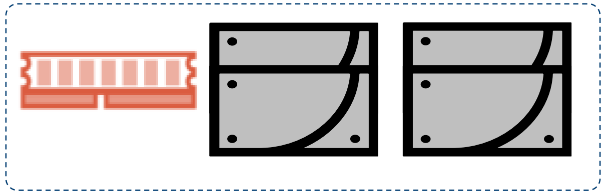
*Other names and brands may be claimed as the property of others

Segmented GEMM benchmark

All DRAM

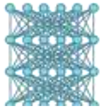
DRAM + Intel® Optane™ SSD + Intel® Memory Drive Technology





Original DGEMM application

2,322 GFLOPS



Optimized application

2,786 GFLOPS

2,605 GFLOPS

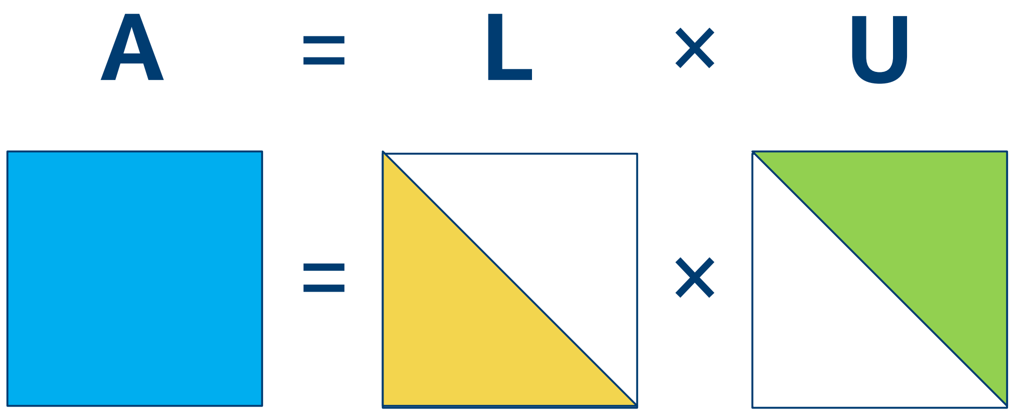
up to **1.1x** **faster matrix multiplication with optimized data locality with no changes[1] in the application**

up to **0.9x** **near DRAM performance of the optimized application**

1. Optane + IMDT configuration – 2 x Intel® Xeon® CPU E5-2699 v4 @ 2.20Ghz, Intel® Server Board S2600WT, 128GB DDR4 + 4* Intel® SSD Optane® (SSDPED1K375GA), CentOS 7.3.1611. All DRAM configuration – 2 x Intel® Xeon® CPU E5-2699 v4 @ 2.20Ghz, Intel® Server Board S2600WT, 768GB DDR4 CentOS 7.3.1611. Test – GEMM(MKL), segment size 18689, factor 22, threads 42, dataset consumed ~650GB.

LU decomposition

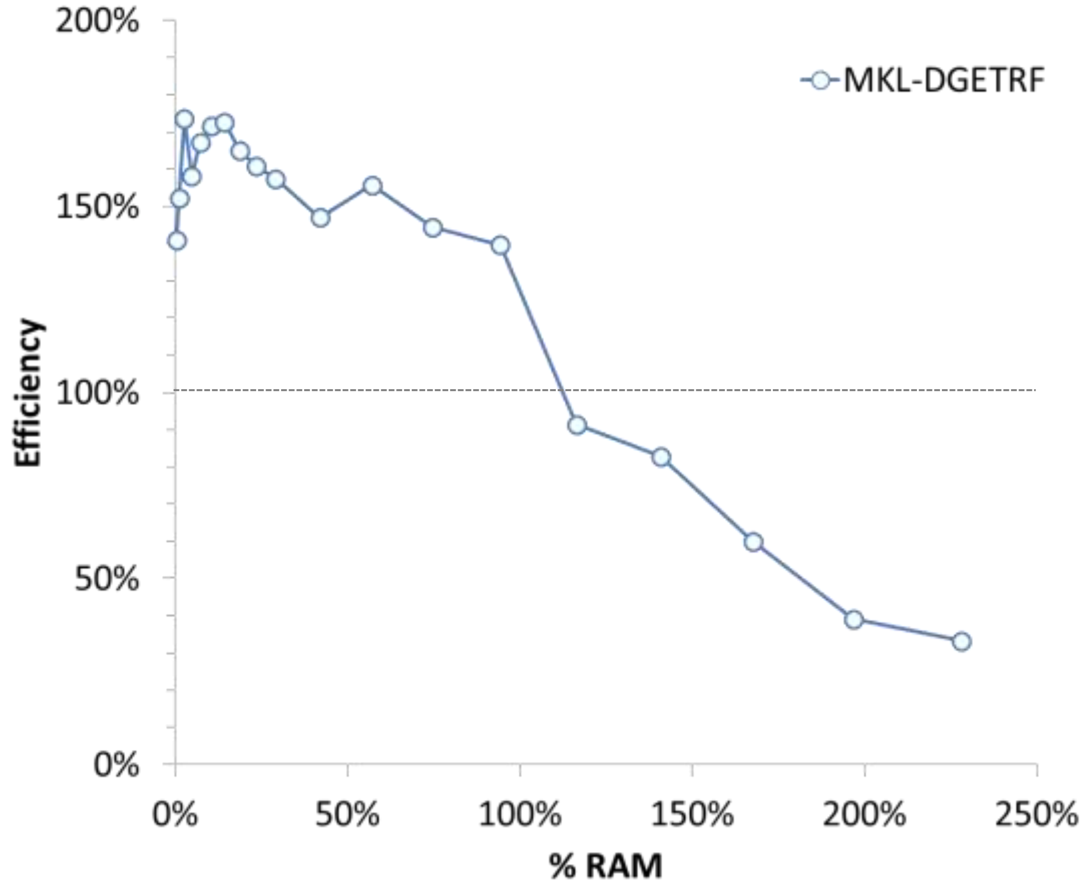
- Factorization of matrix A into product of lower triangular (L) and upper triangular (U) matrices
- A commonly used kernel in many scientific codes:
  - Solving systems of linear equations
  - Matrix inversion
  - Computing determinants
- A kernel in LINPACK benchmark



LU decomposition

Performance results

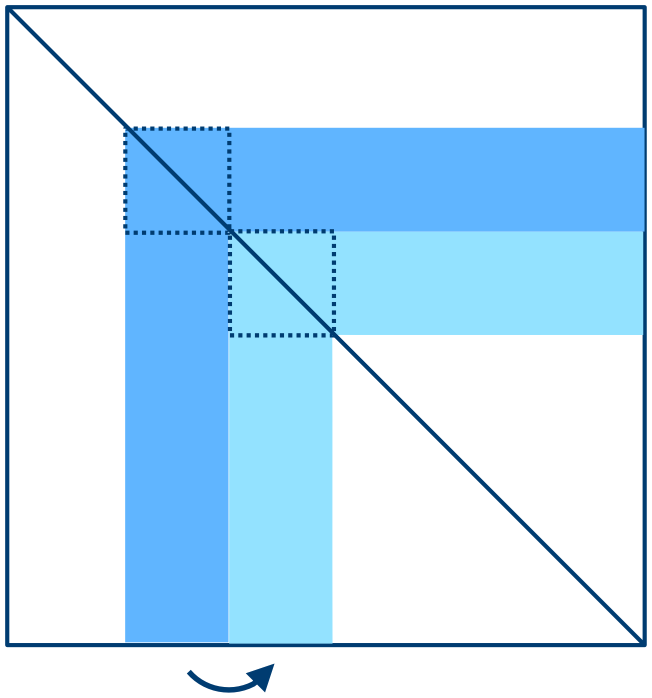
- DDR maximum performance: 850 GFLOPs/s
- Intel® memory drive technology max performance: 1,250 GFLOPs/s
- A huge performance degradation beyond 150% RAM utilization



Can we improve these results?

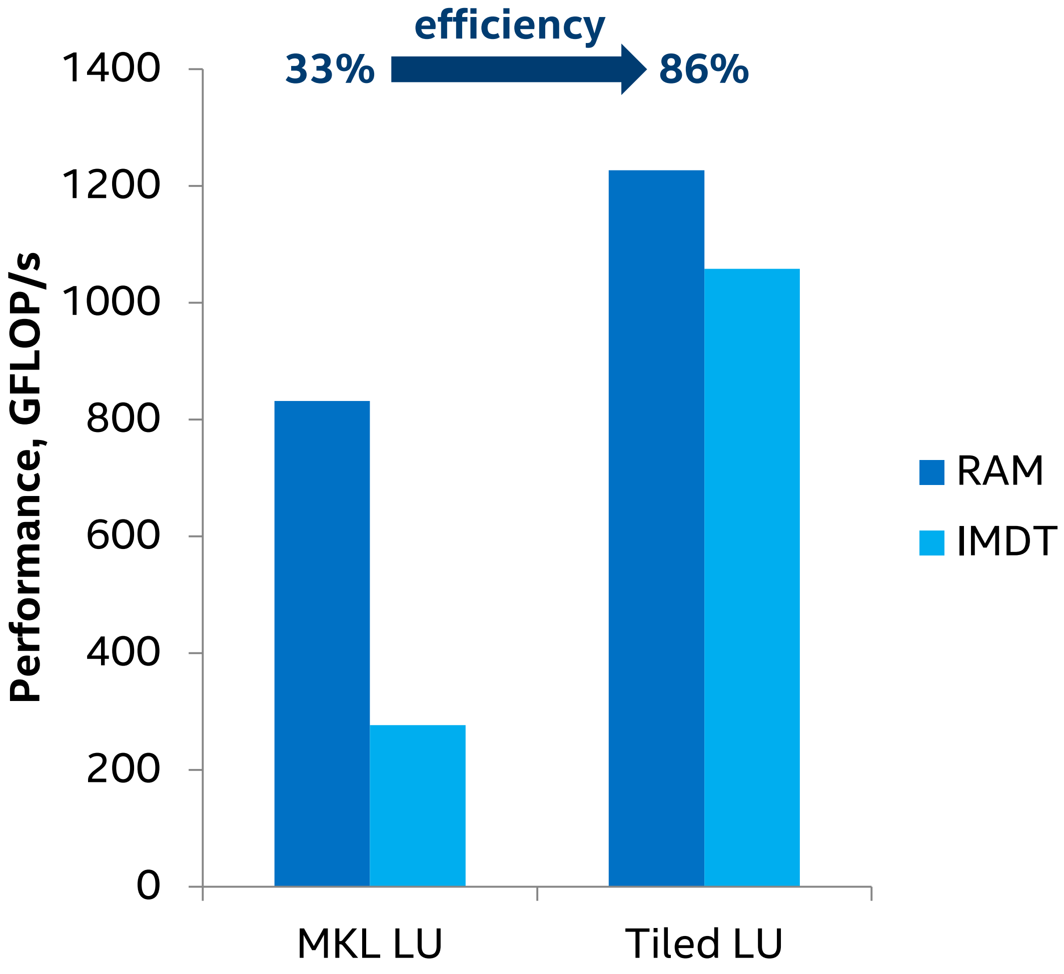
LU decomposition

- Memory access pattern is by column blocks
- Nearby elements are scattered throughout different memory pages
  - 4KB page = 512 double precision numbers
  - A huge data traffic for large matrices ($2 \cdot 10^5$and above)
- There are tiled LU algorithms (e.q. PLASMA)



LU decomposition

- Memory access pattern is by column blocks
- Nearby elements are scattered throughout different memory pages
  - 4KB page = 512 double precision numbers
  - A huge data traffic for large matrices ($2 \cdot 10^5$and above)
- There are tiled LU algorithms (e.q. PLASMA)
- We used a simple implementation from *hetero-streams* code base
- Little performance degradation beyond 100% RAM usage



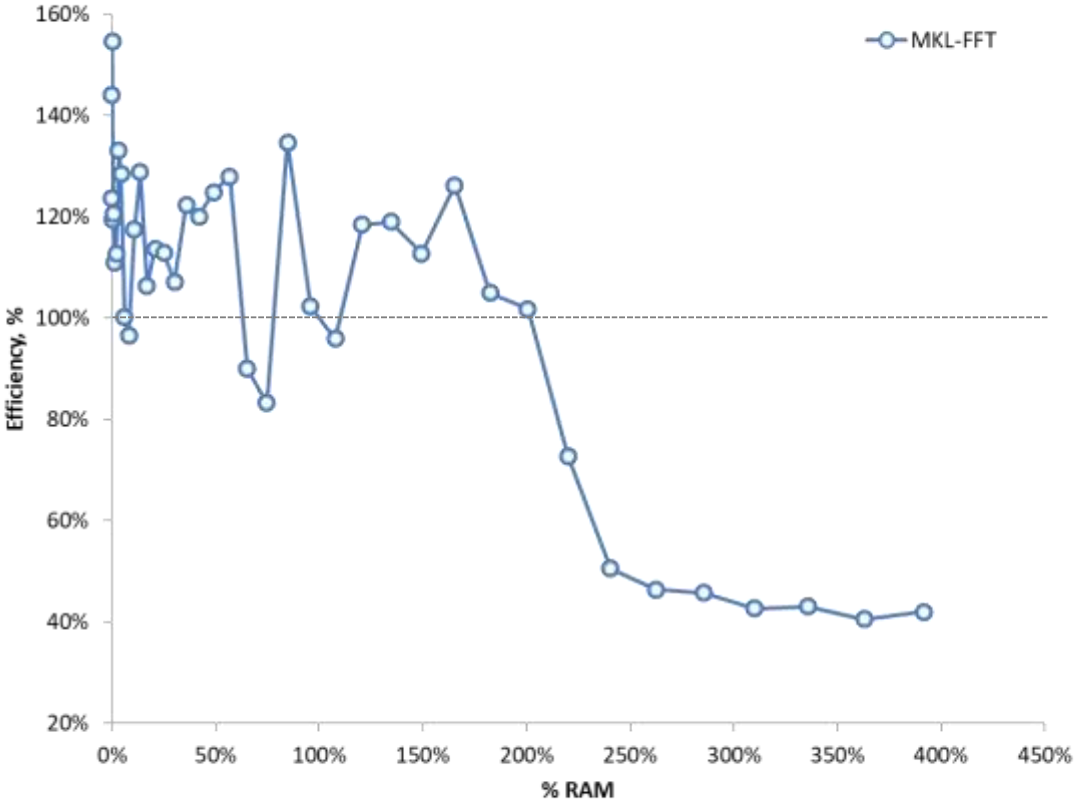
$N = 280000$, 230% RAM

Fast Fourier transformation

- A common used kernel in physics and material science
- Compute bound, but AI grows very slow with problem size
- $O(N \log N)$ time complexity
- “Butterfly” memory access pattern – complex but predictable

Fast Fourier transformation

- Intel® Math Kernel Library DFT kernel
- 3D FFT benchmark, $N \times N \times N$ grid
- Results:
  - 80-130% of DDR performance up to 200% of DDR utilization
  - 40% efficiency over 250% DDR utilization
- 3D FFT can be optimized for NUMA and MDT in a similar way to the LU decomposition
  - by dividing the total memory worked on by all threads at a given time



Lessons learned from benchmarks with Intel® memory drive technology

- Data moving between Intel ® Optane™ SSDs and RAM is very expensive (10 GB/s max):
  - Reuse data as much as possible
    - Arithmetic intensity on DRAM↔MDT level should be ≥500 FLOPs/byte
  - Redesign data structures in you program for locality
  - Work with large data chunks
  - Think about DRAM as a large L4 cache for MDT
- Same optimization principles as on NUMA architectures
- Data-oriented programming is a must
  - It also favors modern hardware architectures

Scientific applications

Computational chemistry:

- LAMMPS* (molecular dynamics)
- GAMESS (two-electron integral kernel)

Astrophysics:

- AstroPhi* (hyperbolic partial differential equation solver)

Sparse linear algebra problems:

- Intel® Math Kernel Library PARDISO

Quantum computing simulator:

- Intel-QS, formerly known as qHipster

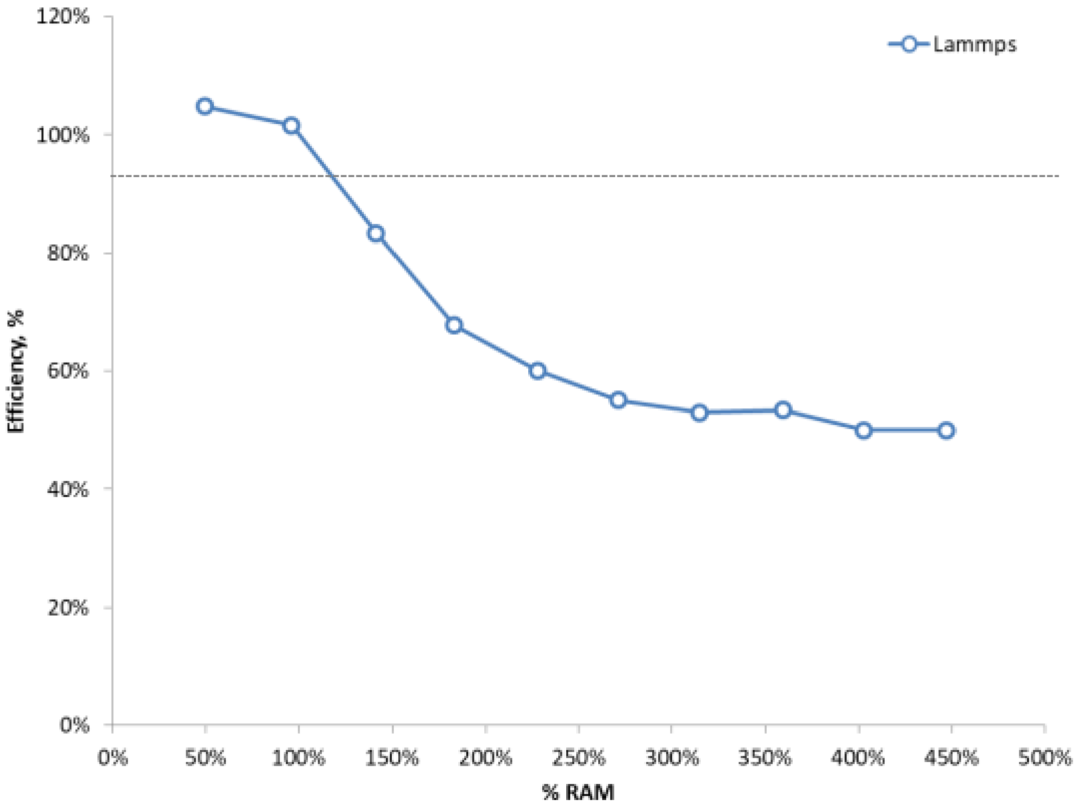
LAMMPS*

Popular molecular dynamics package

- Mostly used in material science
- Force-field based molecular dynamics
  - Partitions the simulation domain (spatial-decomposition) into small 3d sub-domains, each assigned to a CPU
  - Processors communicate and store *ghost* atom information for atoms that border their subdomain

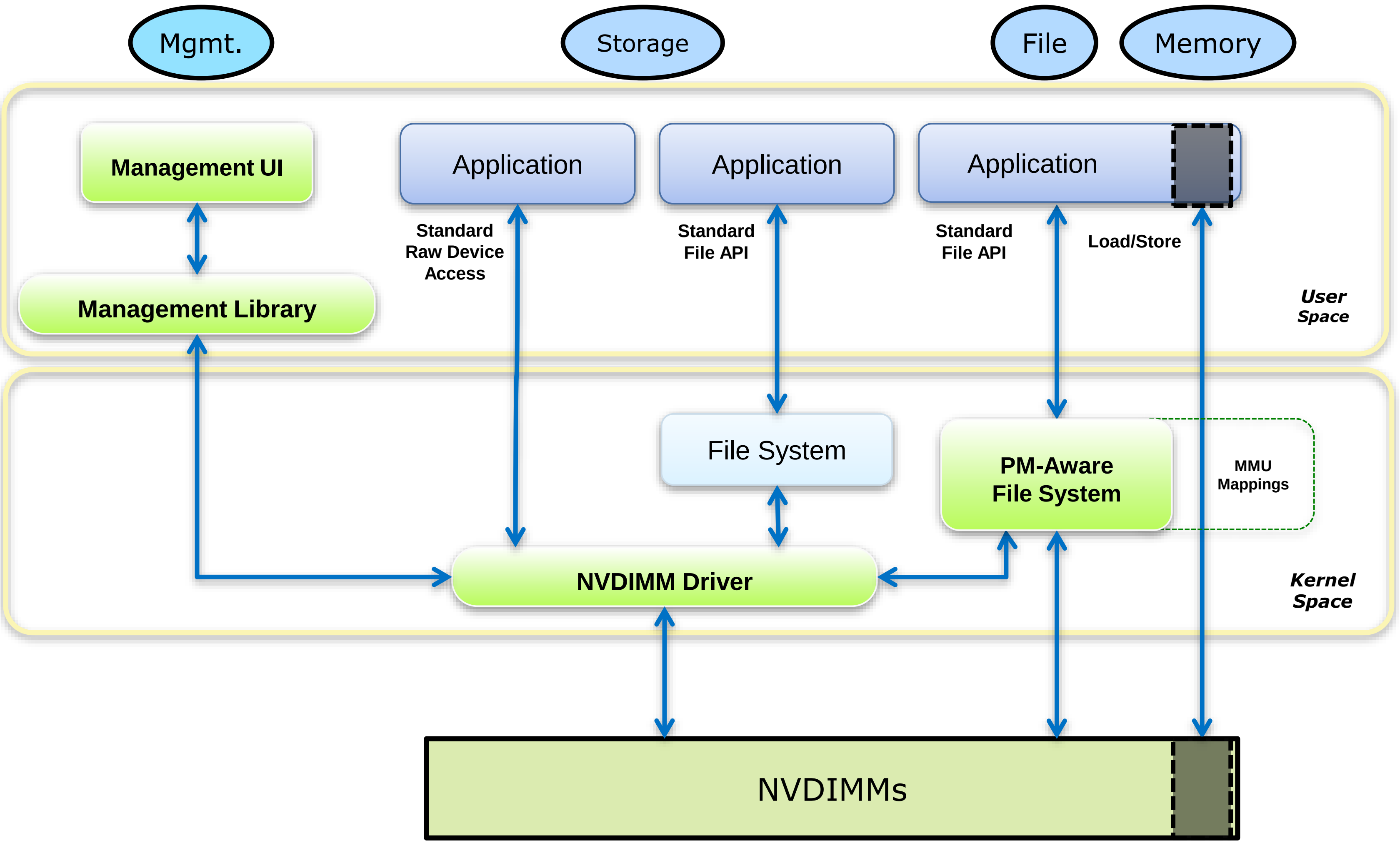
Scaled Rhodopsin benchmark:

- Hundreds of millions atoms
- Major bottleneck is calculation of electrostatic interaction between atoms
- Reasonable efficiency up to 150% RAM
  - 50% efficiency for high memory consumption



FUTURE POSSIBILITIES

The SNIA NVM Programming Model



The Persistent Memory Development Kit

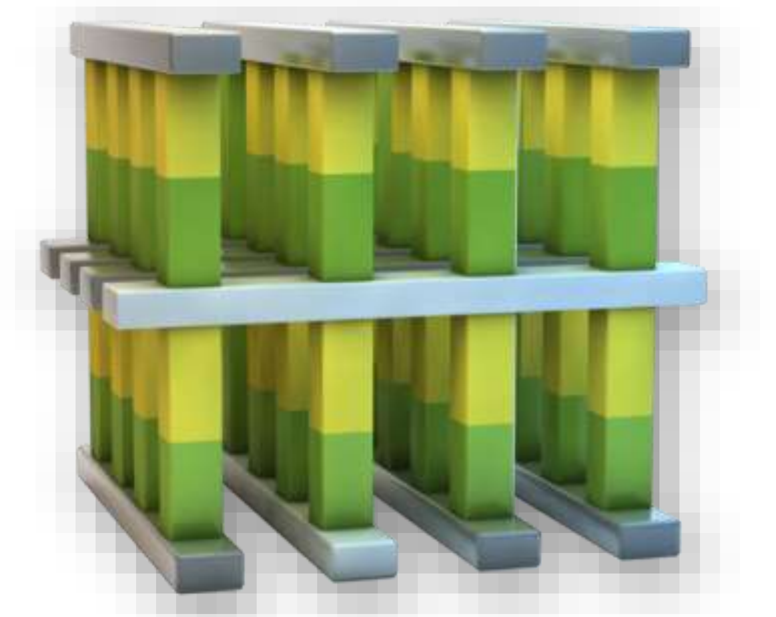
PMDK http://pmem.io

- PMDK is a collection of libraries
  - Developers pull only what they need
    - Low level programming support
    - Transaction APIs
  - Fully validated
  - Performance tuned.
- Open Source & Product neutral

Intel Persistent Memory

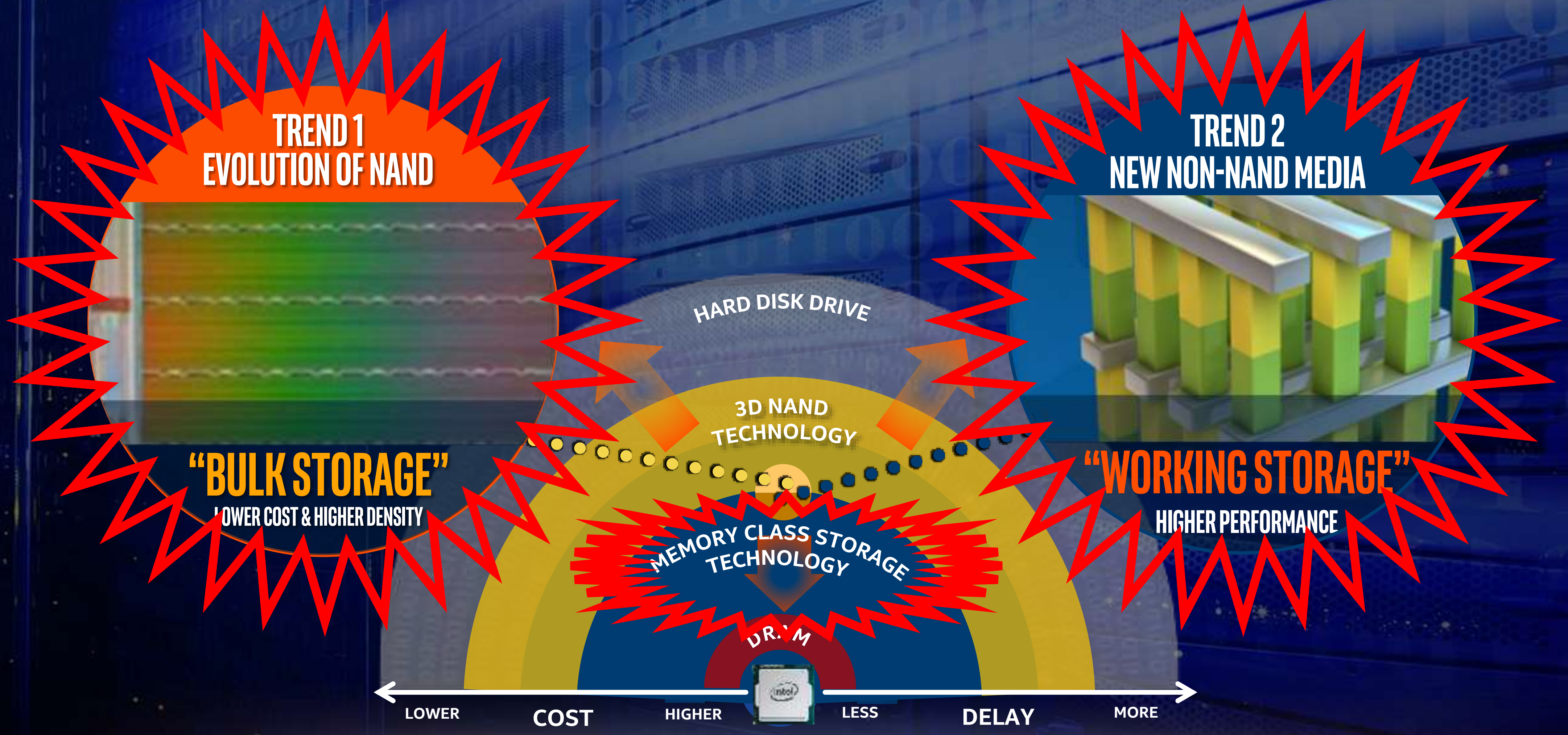
New Type of Memory

- Persistent, Large Capacity & Byte Addressable
  - 6 TB per two-socket system
- DDR4 Socket Compatible
  - Can Co-exist with Conventional DDR4 DRAM DIMMs
- Cheaper than DRAM
- Availability
  - Next Xeon Scalable Platform

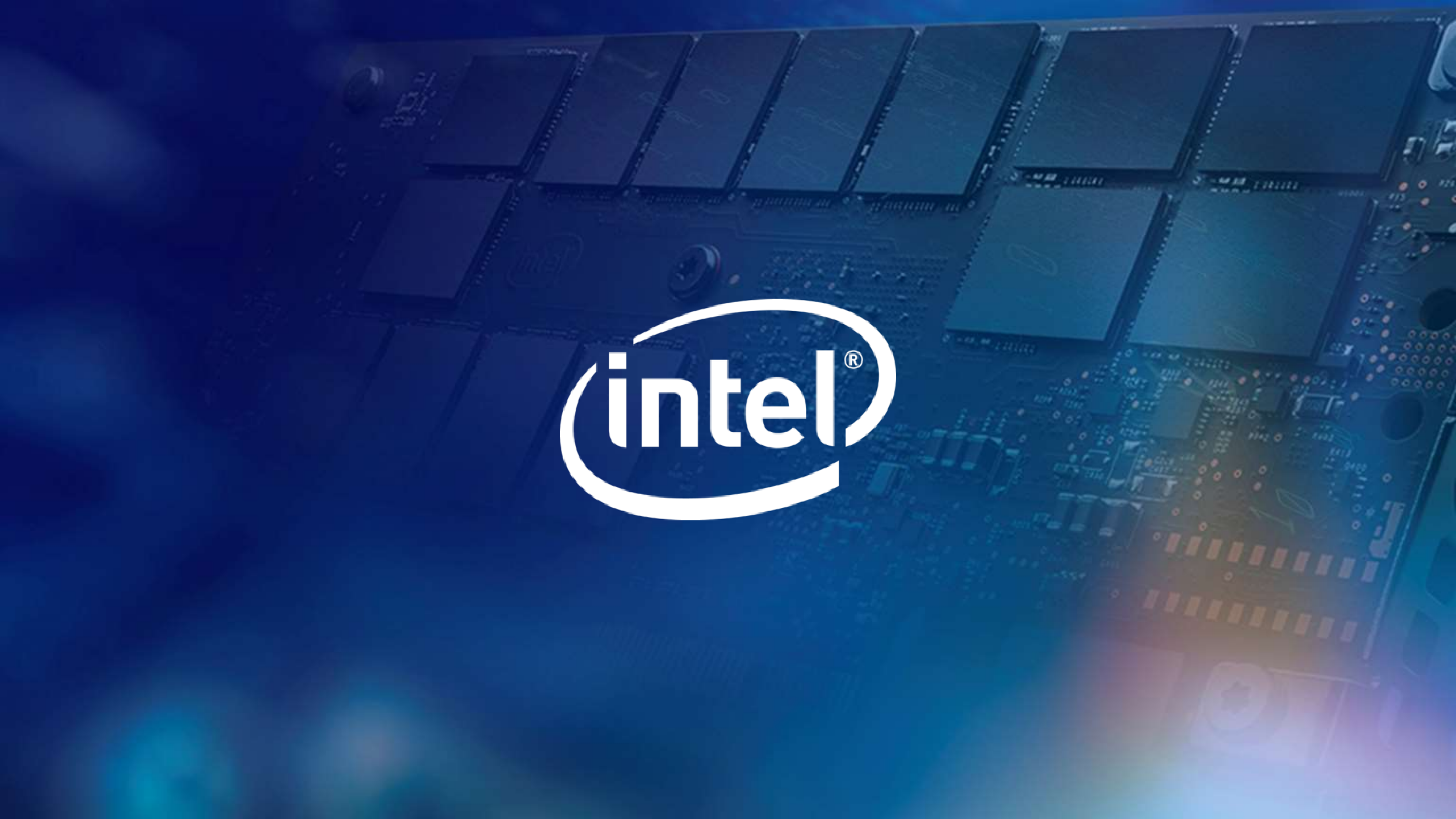


3D XPoint™ technology

NEW STORAGE / MEMORY SOLUTIONS FROM INTEL



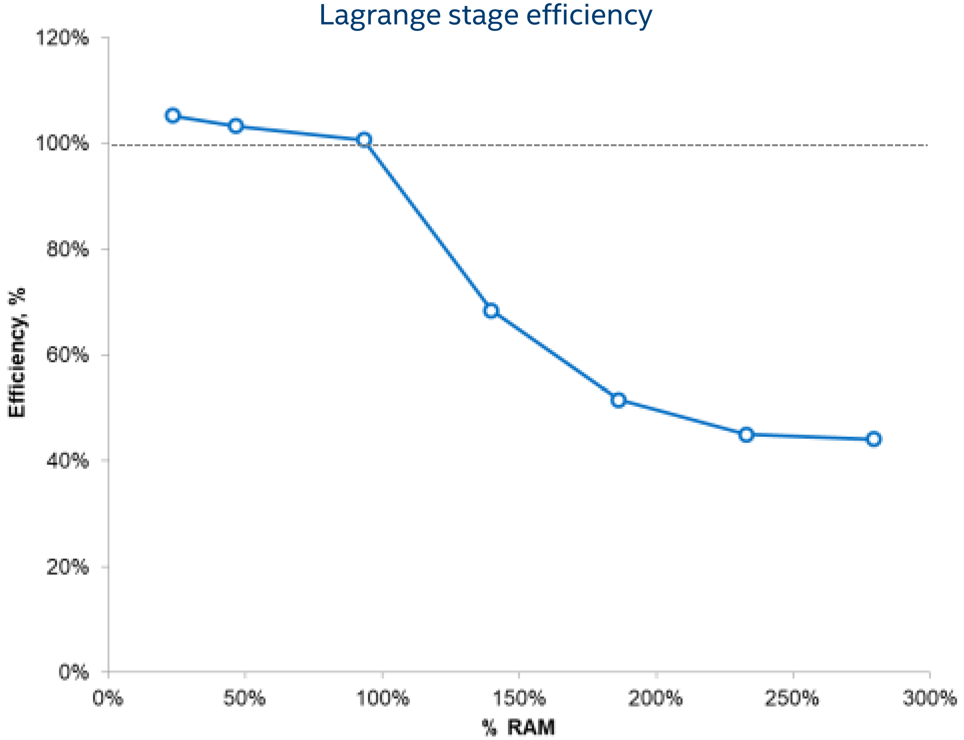
THANK YOU

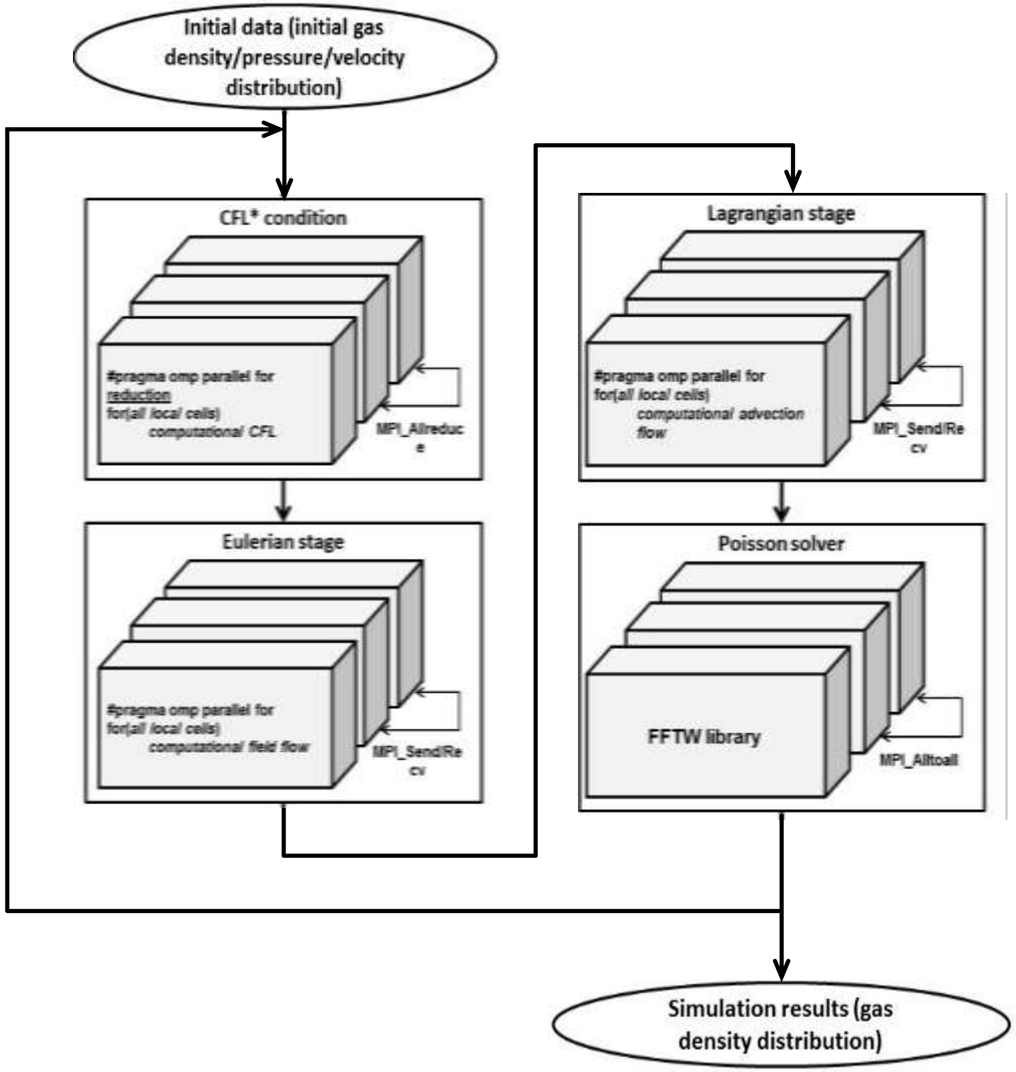

intel®

BACKUP

AstroPhi

- The hyperbolic PDE engine
- Numerical 3D finite difference kernel
- Code is not currently optimized, opportunities for MDT optimization have been identified



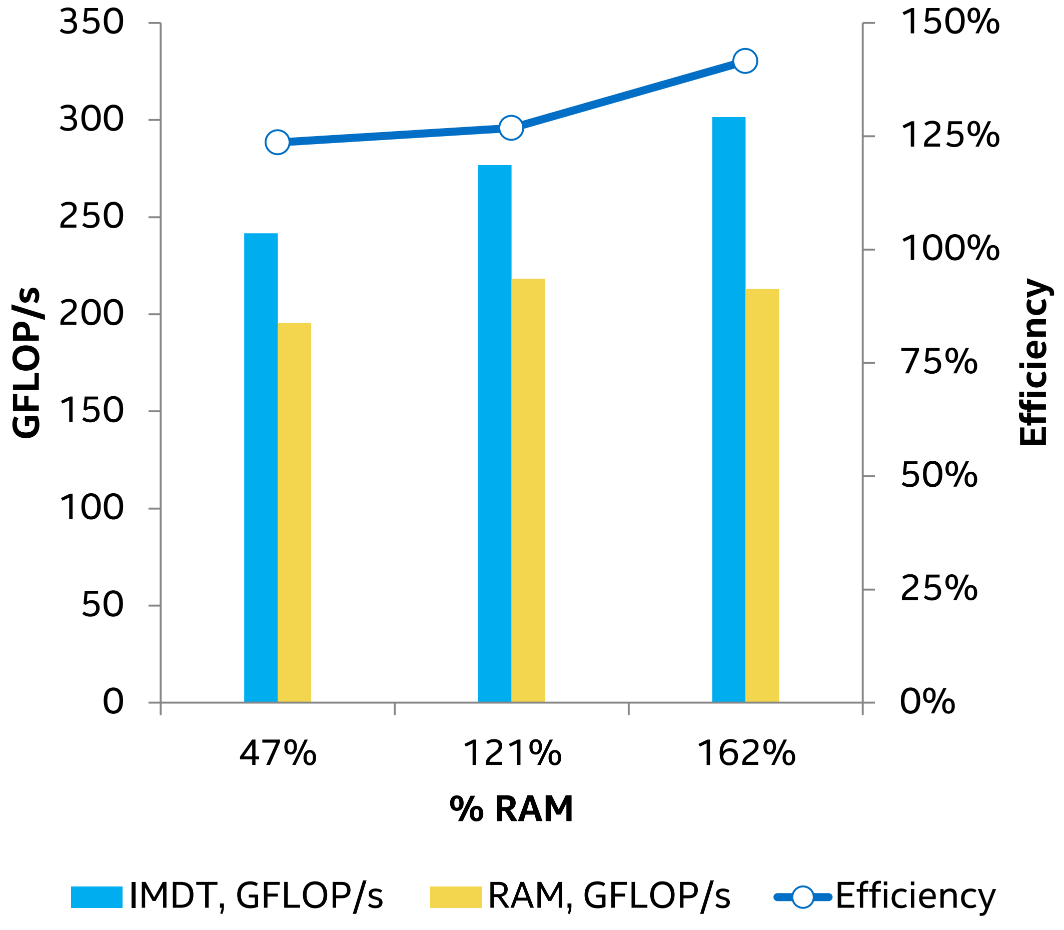


I.M.Kulikov, I.G.Chernykh, A.V.Snytnikov, B.M.Glinskiy, A.V.Tutukov, Comp. Phys. Comm., vol. 186, pp. 71-80, 2015.

PARDISO

- Intel® Math Kernel Library PARDISO – Parallel Direct Sparse Solver
- Solves huge linear algebra problems: hundreds of millions variables
- Test cases:
  - Cholesky factorization of square $N \times N$ matrices
  - Matrix dimensions:
    $N = 10 \cdot 10^7, \quad 20 \cdot 10^7, \quad 25 \cdot 10^7$
  - Number of nonzero elements: $\mathrm{O}(N)$

Intel® memory drive technology provides better than DRAM performance

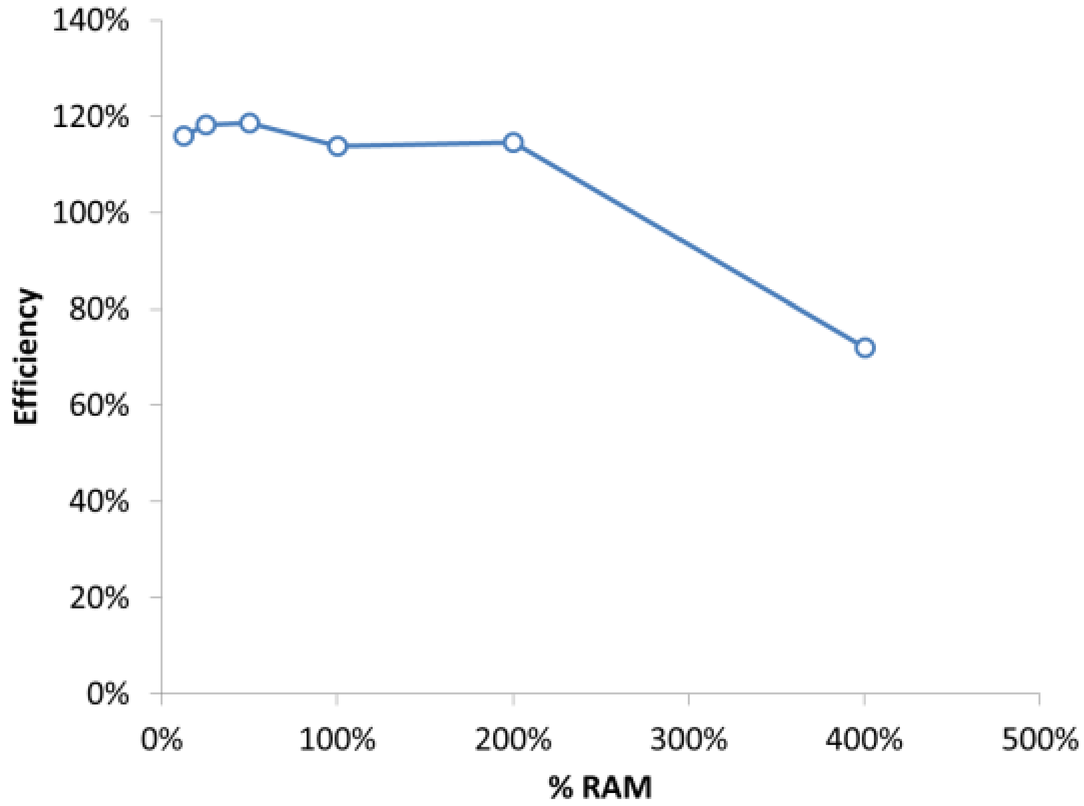


Intel QS

- Quantum computing simulation
  - Application requires more memory as more qubits are simulated
  - Without MDT, scaling beyond a node's capability requires MPI on a cluster

- Test cases:
  - Quantum Fourier transform
  - $N_{qubits} = 30 - 35$

- Good performance up to 4×RAM utilization
  - 35 qubits would require > 1.5TB
  - with MDT a single node can runs 35 qubits – enabling the move to HPC/HTC Cloud, instead of HPC cluster with MPI



GAMESS

Two-electron integrals:

- An important kernel in quantum chemistry
- Used in many quantum chemistry methods
- Different types of two-electron integrals have different efficiency on MDT
- Benchmark details:
  - Rys quadrature ERI kernel from GAMESS
  - Compute and store ERIs to memory

Int. type \ % RAM	27%	82%	109%	164%	273%
(SS|SS)	97%	111%	104%	100%	99%
(PS|SS)	100%	95%	98%	93%	93%
(PS|PS)	95%	94%	97%	87%	71%
(PP|SS)	104%	97%	96%	86%	75%
(PP|PS)	124%	100%	117%	76%	58%
(PP|PP)	146%	106%	106%	79%	50%
(DS|SS)	98%	96%	94%	92%	92%
(DS|PS)	114%	105%	103%	77%	60%
(DS|PP)	140%	87%	98%	91%	53%
(DS|DS)	112%	111%	133%	75%	53%
(DP|SS)	105%	99%	93%	99%	63%
(DP|PS)	130%	107%	101%	68%	45%
(DP|PP)	109%	106%	88%	87%	58%
(DP|DS)	88%	92%	109%	83%	65%
(DP|DP)	124%	121%	115%	88%	42%
(DD|SS)	119%	100%	100%	80%	56%
(DD|PS)	130%	117%	116%	72%	53%
(DD|PP)	88%	135%	109%	89%	54%
(DD|DS)	128%	103%	126%	96%	56%
(DD|DP)	118%	114%	104%	94%	51%
(DD|DD)	111%	115%	113%	83%	53%