Intel for HPC & AI

Igor Freitas

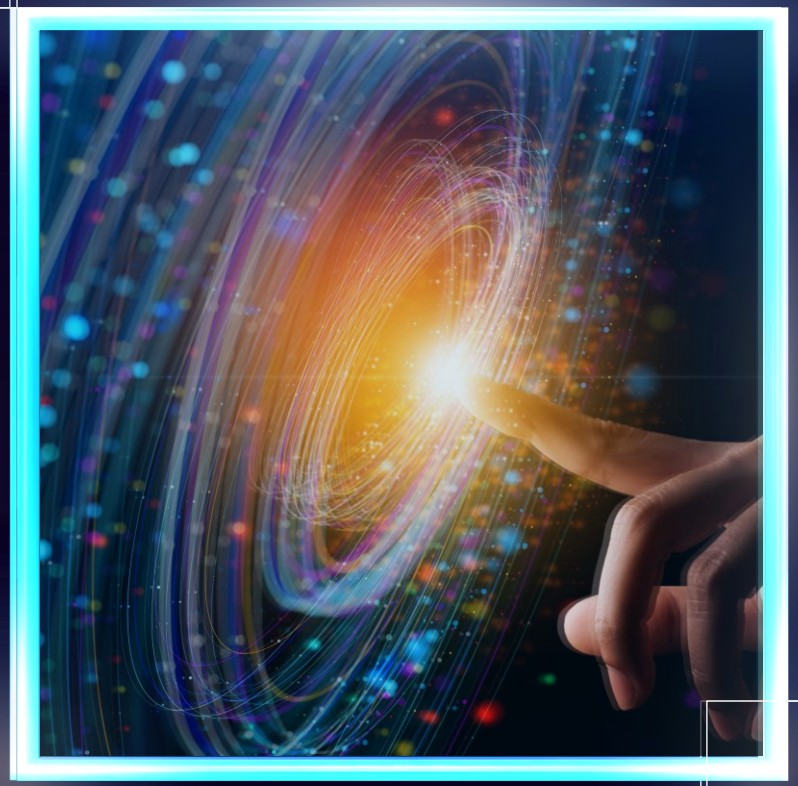
intel.

Lenovo

intel

Bringing AI everywhere.

Enabling the AI continuum in every platform...
from client and edge to data center and cloud.



Bringing AI Everywhere

with an open systems-based approach

Vertical Solutions

Open Application Ecosystem
Software Vendors & System Integrators

Open Software Ecosystem
Data management, AI models & frameworks, infrastructure foundational software

Open Infrastructure Ecosystem
OEMs, ODMs, CSPs, OSVs

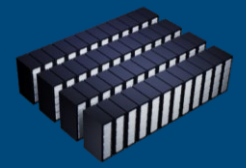
AI PC Node

Node / Server

Rack

Cluster

Super Cluster



Mega Cluster

intel core ULTRA

intel vPRO i7 intel vPRO i9 intel vPRO i5

An Evo Design

AI PC

intel core ULTRA

intel ARC GRAPHICS

intel xeon

intel ethernet

Enterprise & Edge

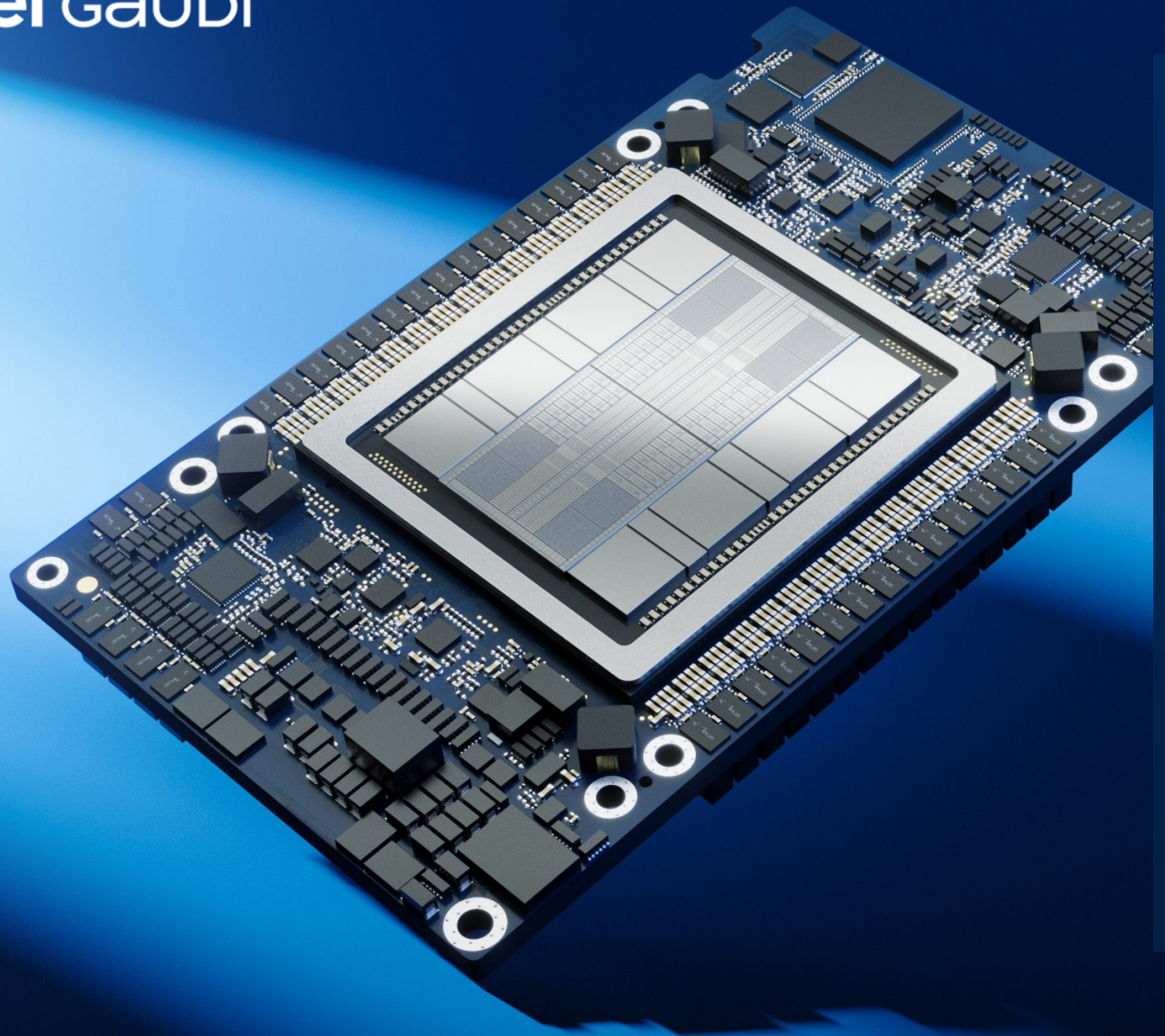
intel IPU

intel xeon

intel gaudi

Data Center

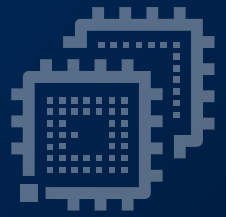
intel GAUDI



Bringing Choice to Gen AI with Performance, Scalability and Efficiency

Intel® Gaudi® 3 AI accelerator

How Intel Gaudi addresses Enterprise challenges



Need more Choice

other than single-source GPUs

- Gaudi 3 outperforms H100 performance of LLMs for inferencing
- Lower hardware cost and no CUDA licensing costs
- Industry-standard high speed ethernet

Locked-in

with proprietary software and networking

- Software migration in as few as three lines of code
- Community-based open-source software stack
- Non-proprietary based network solution

Ability to Scale

while containing costs of infrastructure

- Readily supports demanding Gen AI workloads from 1 to 1000s of nodes
- Easily and cost-effectively integrate into Ethernet-based networks
- High-efficiency cluster scaling drives cost savings

Maximize efficiency

yet still solve my business challenges

- Higher performance per watt than H100
- Higher price-performance over H100
- Integration of open software frameworks drives developer productivity

General availability starting Q4 ’24 from our Partners

DELL Technologies

Hewlett Packard Enterprise

SUPERMICRO

Dell PowerEdge XE9680

Air-cooled
Dell AI Factory

SHIPPING OCTOBER 2024

HPE ‘XX’ (to be announced)

Air-cooled & Liquid-cooled

SHIPPING DECEMBER 2024

Supermicro X14

Air-cooled & Liquid-cooled
Equipped with Xeon 6

SHIPPING OCTOBER 2024

ODM Partners | Shipments starting Q1 2025

ASUS GIGABYTE TECHNOLOGY ingrasys Inventec QCT Quanta CLOUD TECHNOLOGY wistron wiwynn

intel | IBM Cloud

Intel® Gaudi® 3 AI accelerator on IBM Cloud to Drive Enterprise AI Efficiency

And support Gaudi 3 with IBM's watsonx AI & data platform

ENABLING ENTERPRISE AI CUSTOMERS TO

Cost-effectively
scale Gen AI, lowering TCO

Securely scale
AI across hybrid cloud environments

Ease development
on open community-based platforms

DENVR dataworks

DELIVERING CLOUD SERVICES with
SIMPLICITY, FLEXIBILITY, SAVINGS

Denvr Dataworks: brings choice and increased efficiency

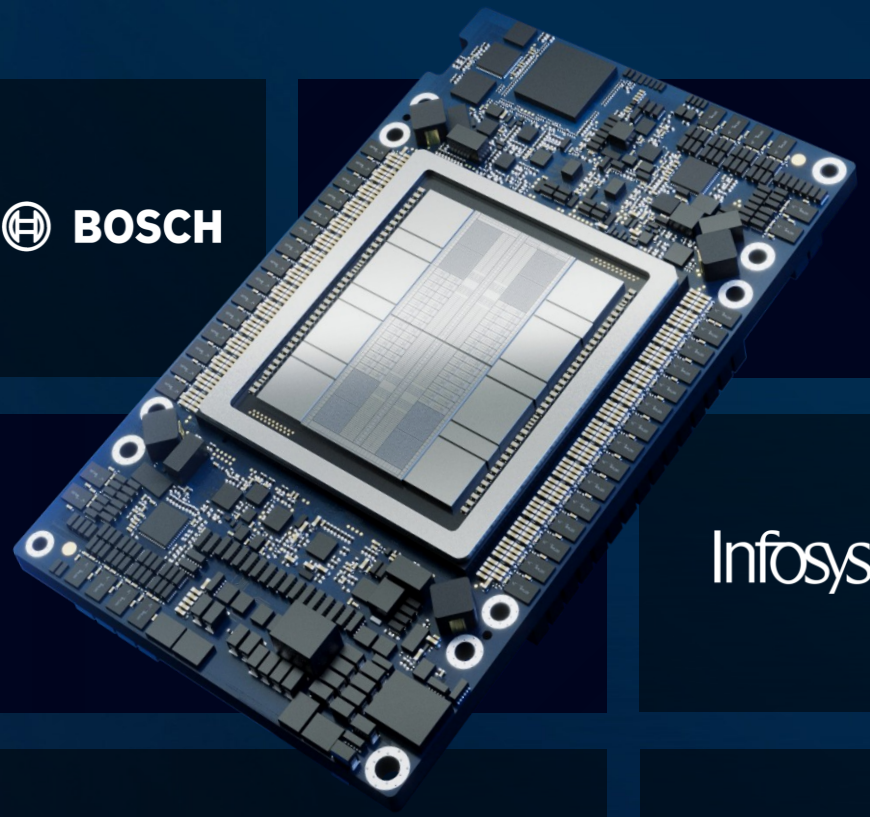
with Gaudi 2 today, Gaudi 3 coming soon

Accelerate time-to-market, increase ROAI

Training -as-a-Service	**Inference** -as-a-Service
RAG -as-a-Service	**Model** -as-a-Service

intel GAUDI

Growing Customer Momentum



AI SWEDEN
Articul8
BOSCH
CtrlS
DENVR dataworks
IBM
iff
Infosys
LandingAI
LUMEN
NAVER
NIQ
predictionGuard
40 Advent International
salesforce
seekr

Delivering Price Performance Advantage

1.19x

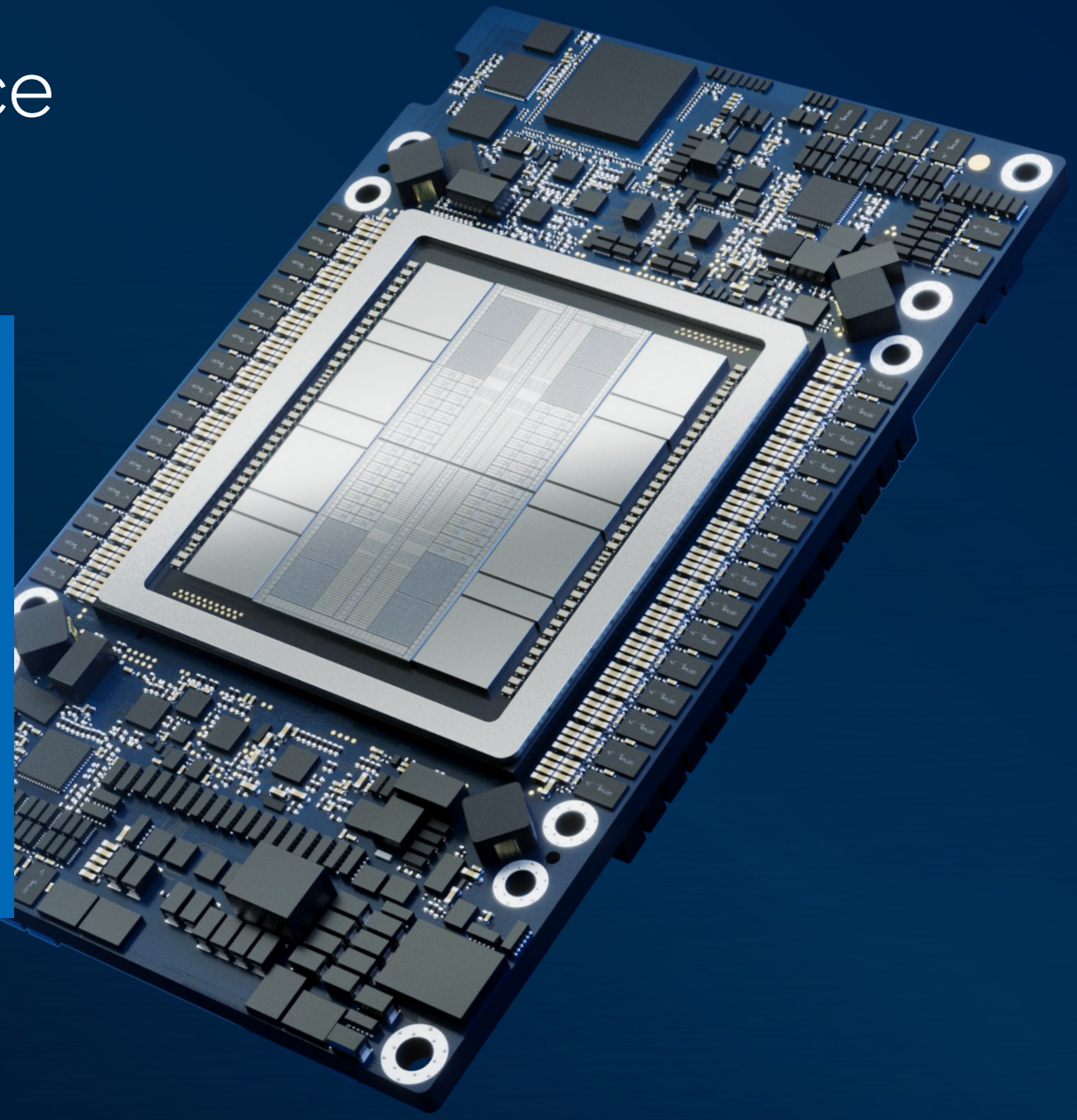
Inference Throughput
LLaMA 2 70B

Intel® Gaudi® 3 AI accelerator
Vs H100

~2x perf/$

Inference Throughput
LLaMA 2 70B

Intel® Gaudi® 3 AI accelerator
Vs H100



Source Intel measured results vs H100 data sources: https://github.com/NVIDIA/TensorRT-LLM/blob/main/docs/source/performance/perf-overview.md input-output sequences: 128-2048tps on 2 accelerators/GPUs. Intel results obtained on September 9th 2024. Hardware: Two Intel Gaudi 3 AI Accelerators (128 GB HBM) vs two Nvidia H100 GPU (80 GB HBM). Software: Intel Gaudi software release 1.18.0. See Nvidia link for H100 software details. Results may vary. Pricing estimates based on publicly available information and Intel internal analysis

Intel® Gaudi® 3 AI Accelerator Competitive Analysis

Inference Workload Comparison – Intel Performance Projections[1]

Model & Execution Parameters				H100 SXM (April 30th 24)		Gaudi3 OAM					Gaudi3 PCIe				
Model	TP (# Devices)	Input Length	Output Length	Batch Size	Reported Throughput[1] (tps)	Batch Size	Projected Throughput (tps)	Projected Power per Card (W)	Power Efficiency: tps/W	Gaudi3 OAM vs. H100 SXM Speedup (x)	Batch Size	Projected Throughput (tps)	Projected Power per Card (W)	Power Efficiency: tps/W	Gaudi3 PCIe vs. H100 SXM Speedup (x)
LLAMA-7B	1	128	128	896	19,854	1,536	21,201	782	27.1	1.1x	1,536	15,891	600	26.5	0.8x
	1	128	2048	120	6,944	220	7,934	519	15.3	1.1x	220	8,029	517	15.5	1.2x
	1	2048	128	84	2,163	120	2,002	725	2.8	0.9x	120	1,557	570	2.7	0.7x
	1	2048	2048	56	2,826	120	3,168	533	5.9	1.1x	120	3,148	519	6.1	1.1x
Mistral-7B	1	128	128	896	20,404	896	28,462	900	31.6	1.4x	896	18,928	600	31.5	0.9x
	1	128	2048	120	8,623	120	16,920	616	27.5	2.0x	120	20,823	600	34.7	2.4x
	1	2048	128	84	2,405	84	2,477	829	3.0	1.0x	84	1,943	600	3.2	0.8x
	1	2048	2048	56	3,731	56	6,603	603	11.0	1.8x	56	8,785	579	15.2	2.4x
Mixtral-7x8B	2	128	128	512	10,510	512	18,883	681	13.9	1.8x	512	16,836	590	14.3	1.6x
	2	128	2048	128	6,816	128	8,908	486	9.2	1.3x	128	8,498	457	9.3	1.2x
	2	2048	128	64	1,176	64	1,981	690	1.4	1.7x	64	1,604	538	1.5	1.4x
	2	2048	2048	32	2,634	32	2,404	453	2.7	0.9x	32	2,253	420	2.7	0.9x
LLAMA-70B	2	128	128	1024	6,428	4,096	5,794	816	3.6	0.9x	4,096	4,182	600	3.5	0.7x
	4	128	2048	512	10,900	1,024	16,128	702	5.7	1.5x	1,024	12,798	600	5.3	1.2x
	2	2048	128	96	692	220	655	861	0.4	0.9x	220	447	600	0.4	0.6x
	2	2048	2048	64	2022	256	3,382	687	2.5	1.7x	256	2,922	600	2.4	1.4x
Falcon-180B	4	128	128	512	4400	4,096	5,111	900	1.4	1.2x	4,096	3,309	600	1.4	0.8x
	8	128	2048	1024	6696	4,096	17,798	900	2.5	2.7x	4,096	11,463	600	2.4	1.7x
	4	2048	128	64	448	512	507	898	0.1	1.1x	512	335	600	0.1	0.7x
	4	2048	2048	64	984	512	4,047	871	1.2	4.1x	512	2,707	600	1.1	2.8x
									Average Speedup:	1.5x					1.3x

1 - Source for Intel Gaudi 3 compared to Nvidia H100 performance: (Nvidia) https://nvidia.github.io/TensorRT-LLM/performance/perf-overview.html, May 2024. Reported numbers are per GPU. Intel Gaudi 3 projections by Intel, April 2024. Also see *Providing High Performance GenAI at Significantly Lower Total Cost with Intel Gaudi AI Accelerators*, https://www.intel.com/content/www/us/en/newsroom/news/computex-2024-ai-everywhere-power-performance-affordability.html. Results that are based on systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications or configurations..

Intel® Gaudi® 3 AI Accelerator Competitive Analysis

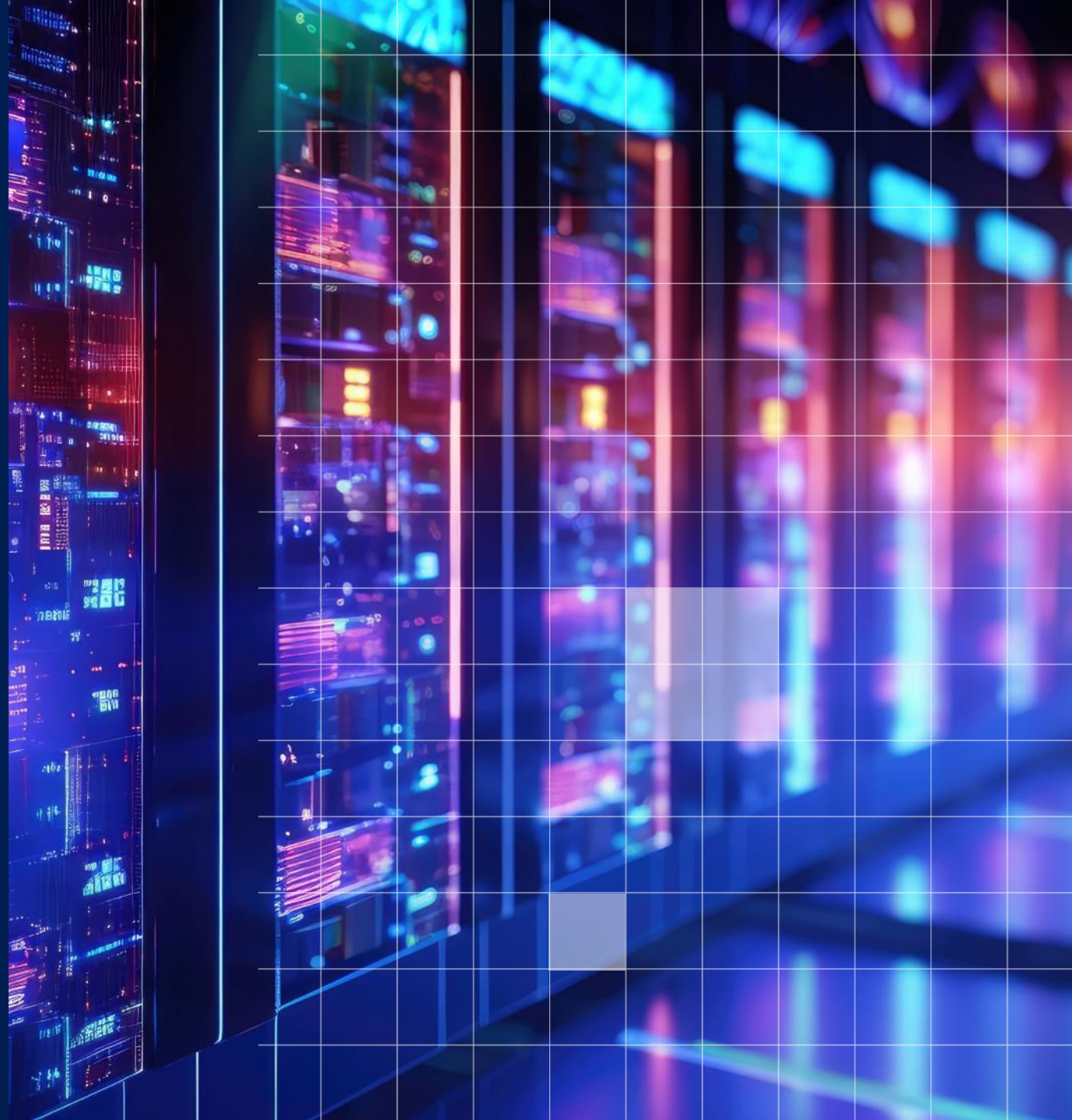
Training Comparison – Intel Performance Projections[1]

Gaudi3 OAM & PCIe: vs H100												
Model & Execution Parameters				H100	Gaudi3 OAM				Gaudi3 PCIe			
Model	Sequence Length	# Cards in Cluster	Precision	Reported Throughput (tps)[1]	Projected Throughput (tps)	Projected Power per Card (W)	Power Efficiency: ktps/W	Gaudi3 vs. H100 Speedup (x)	Projected Throughput (tps)	Projected Power per Card (W)	Power Efficiency: tps/W	Gaudi3 vs. H100 Speedup (x)
LLAMA2-7B	4k	8	FP8	130k	156k	900	21.7	**1.2x**	95k	600	19.7	**0.7x**
LLAMA2-13B	4k	16	FP8	133k	175k	900	12.2	**1.3x**	111k	600	11.6	**0.8x**
GPT3-175B	4k	64	FP8	110k	127k	900	2.2	**1.2x**	76k	600	2.0	**0.7x**
							Average Gain vs. H100	1.2x				0.8x

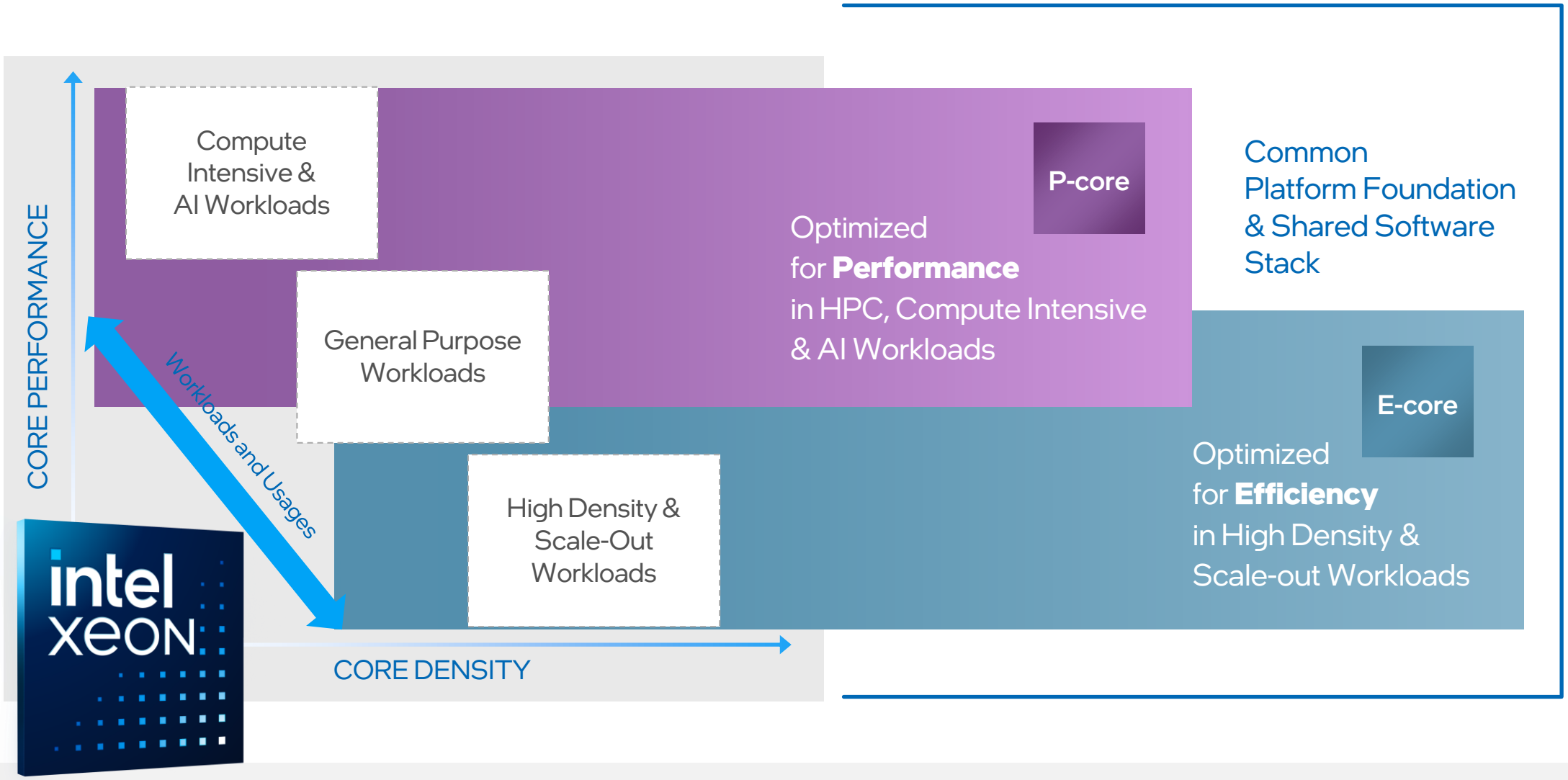
1 - Source for Intel Gaudi 3 compared to Nvidia H100 performance: (Nvidia) https://nvidia.github.io/TensorRT-LLM/performance/perf-overview.html, May 2024. Reported numbers are per GPU. Intel Gaudi 3 projections by Intel, April 2024. Also see *Providing High Performance GenAI at Significantly Lower Total Cost with Intel Gaudi AI Accelerators*, https://www.intel.com/content/www/us/en/newsroom/news/computex-2024-ai-everywhere-power-performance-affordability.html. Results that are based on systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications or configurations.

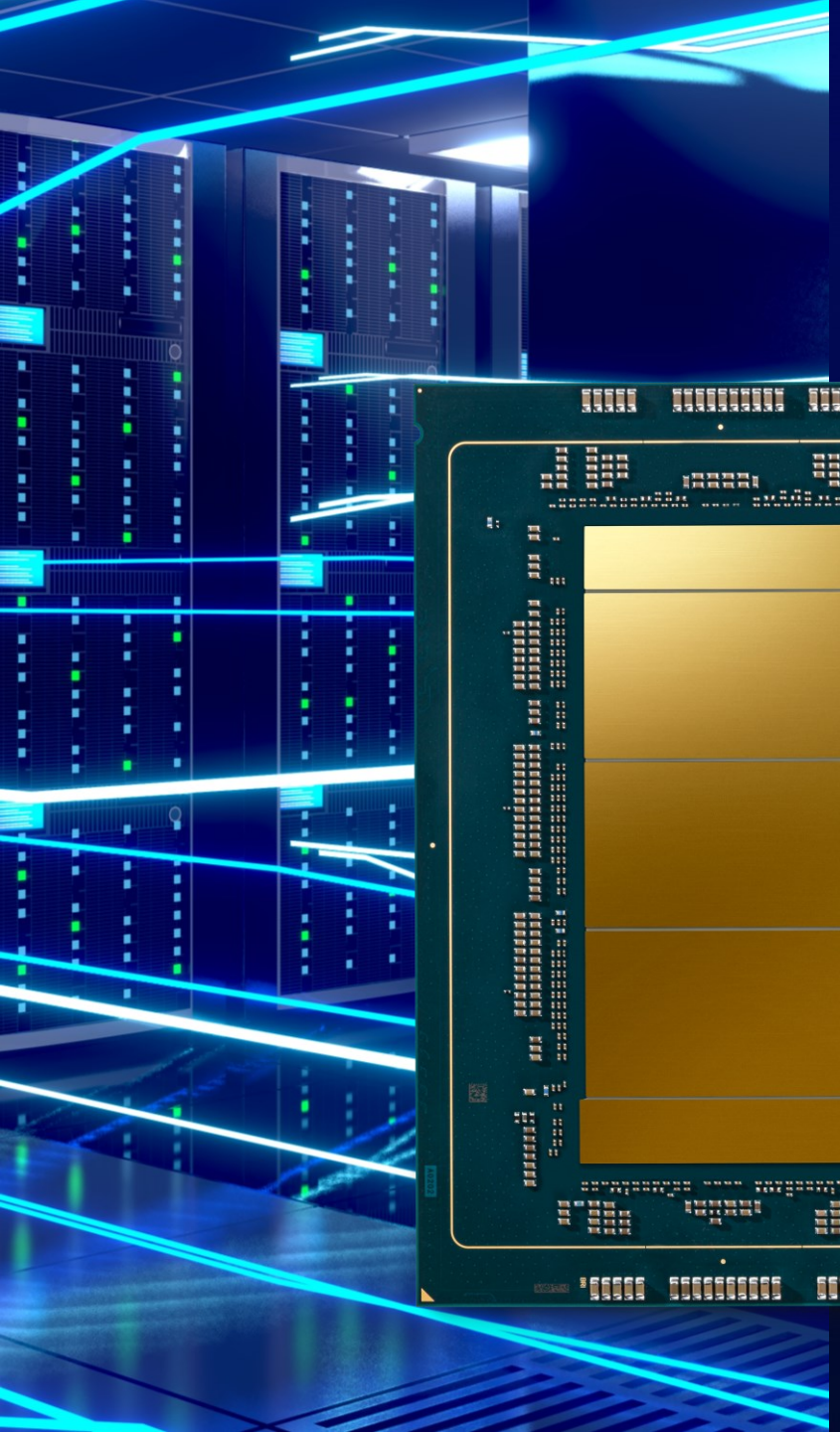
Intel® Xeon® 6

intel ai



Why P and E Cores.....





Intel Xeon 6

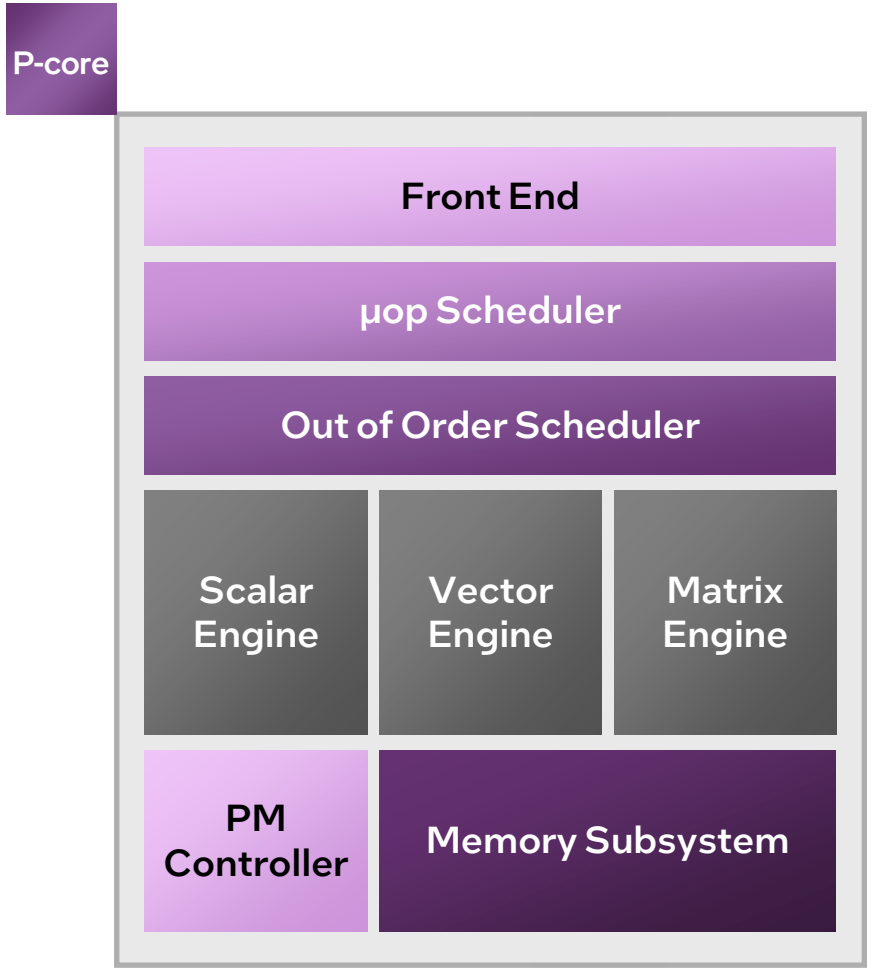
with Performance Cores (P-cores)

6900P Enhancements

- Up to **6400 MT/s** DDR5
- **8800 MT/s** MRDIMM memory
- Up to **128** performance cores
- **6 UPI 2.0** links, up to **24 GT/s**
- Up to **96 lanes** PCIe 5.0/CXL® 2.0
- L3 cache as large as **504 MB**
- Intel Advanced Matrix Extensions (Intel AMX) with **FP16 support**

Intel® Xeon® 6 processors with P-cores

Performance Optimized Core



Proven Intel® Xeon® Processor Architecture

- Optimized for high performance per core
- Built on the latest Intel 3 process technology
- Improved power efficiency

New Software Capabilities

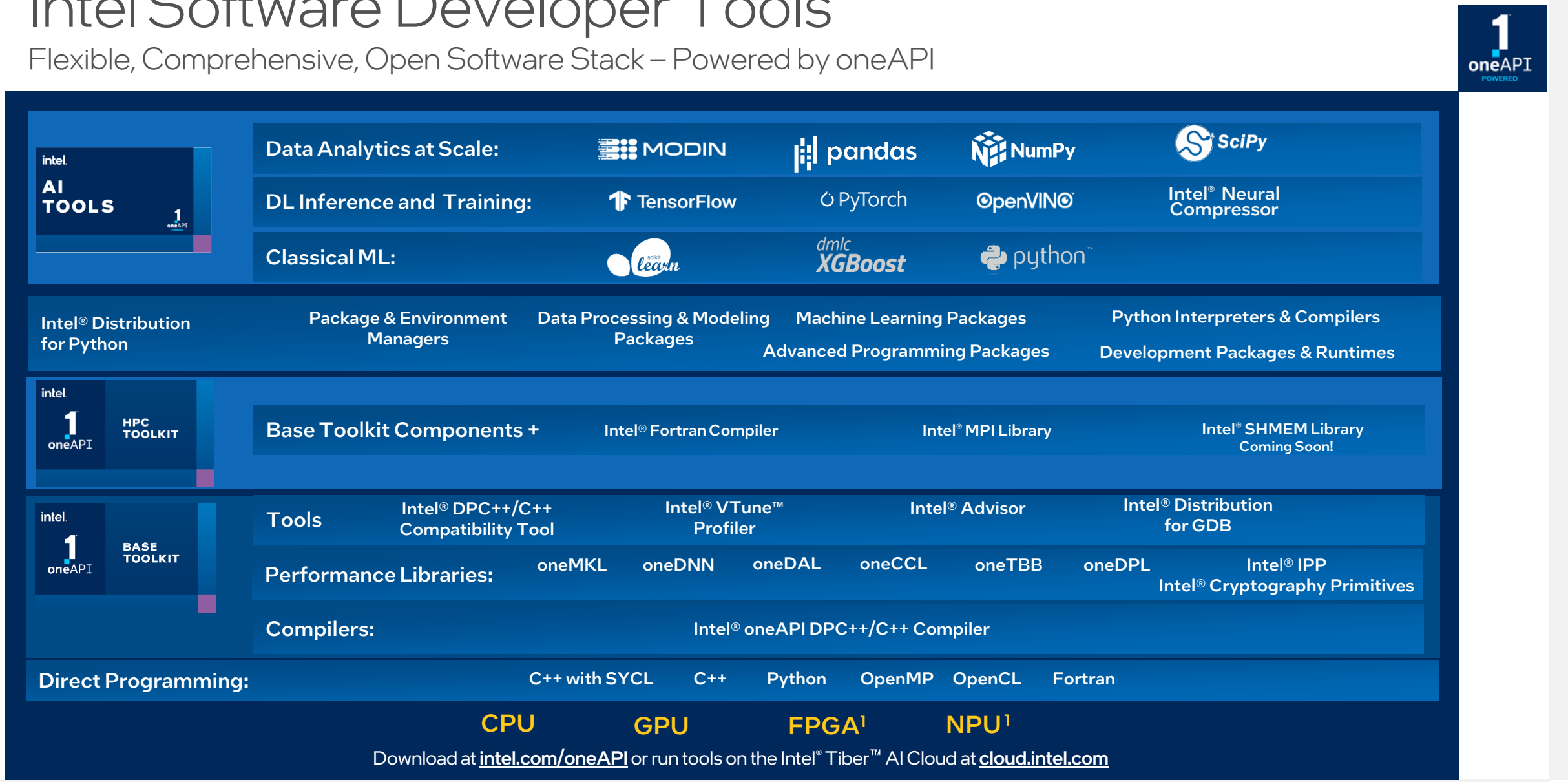
- Matrix engine supports Intel Advanced Matrix Extensions (Intel AMX) and Intel Advanced Vector Extensions 512 (Intel AVX-512) for AI
- More memory encryption keys with 256b strength
- Code SW prefetch and taken branch hints
- Per-thread memory bandwidth allocation
- L2 cache allocation and code/data prioritization

Enhanced µArch

- 64KB, 16-way I-cache
- Improved branch predictor and miss-recovery
- 3-cycle FP multiplication
- More outstanding memory requests and prefetch capabilities

Intel Software Developer Tools

Flexible, Comprehensive, Open Software Stack – Powered by oneAPI



*Other names and brands may be claimed as the property of others. SYCL is a trademark of the Khronos Group Inc.
1. Limited software support available

Intel® AVX -512

Built-in acceleration and outstanding performance

Intel® Advanced Vector Extensions 512 (Intel® AVX-512) is an instruction set that speeds data-heavy, double-precision codes frequently used in scientific computing by widening the register to 512 bits, Intel AVX-512 lets you operate on more data at once, speeding results. This wider vectorization speeds computation processes per clock, increasing frequency over the prior generation.

Intel AVX-512 takes advantage of the processor's increased memory bandwidth, new core architecture, and improved frequency management. Additionally, 2 *FMA throughput is now available across the Platinum, Gold, and Silver SKUs[1].

Four Turbo Frequency Levels

- Enable the usage of high-power instructions such as AVX512 Heavy and AMX while delivering high frequencies for low-power instructions such as SSE, AVX2, etc.

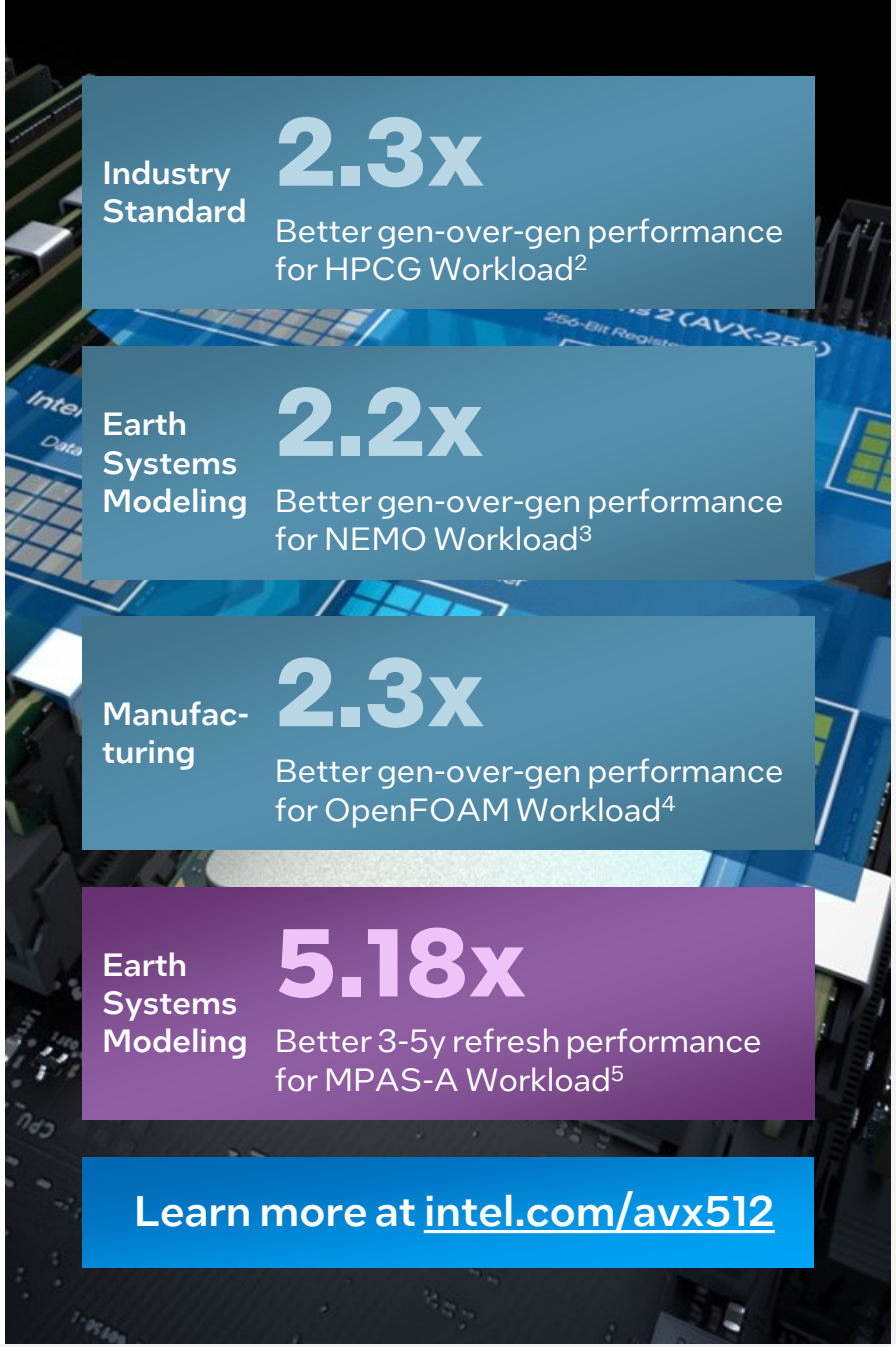
Very Fast License Transition Latencies

- Intel Xeon 6 processors can follow phases for mixed license workloads much closer than any other Intel Xeon product.

Turbo Frequency Selection Based on Instruction Density

- Prevents frequency jitters on mixed license workloads and provides much higher frequency if usage of heavy instructions is sparse.

[1] Check full sku configurations for feature availability.
[2, 3, 4] See "Summary HPC Configurations" for workloads and configurations. Results may vary. Testing done on early pre-production platform and engineering samples. Production 128c SKU will have frequencies up to 4 bins higher than the product tested. Final performance can vary.
[5] See "Summary HPC Configurations" for workloads and configurations. Results may vary. Intel® Xeon® 6900P Series Processor data was collected on pre-production platforms and may not include all final optimizations. The production 128c SKU will have AVX frequencies up to 4 bins higher than the product tested. Final performance can vary. For 4th Gen Intel® Xeon® CPU, MPAS-A, and NEMO were measured on Intel® Xeon® 8480+. The remaining workloads were measured on Intel® Xeon® 8490H.
This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark.



Intel® Advanced Matrix Extensions (Intel® AMX) Tiled Matrix Multiplication Accelerator

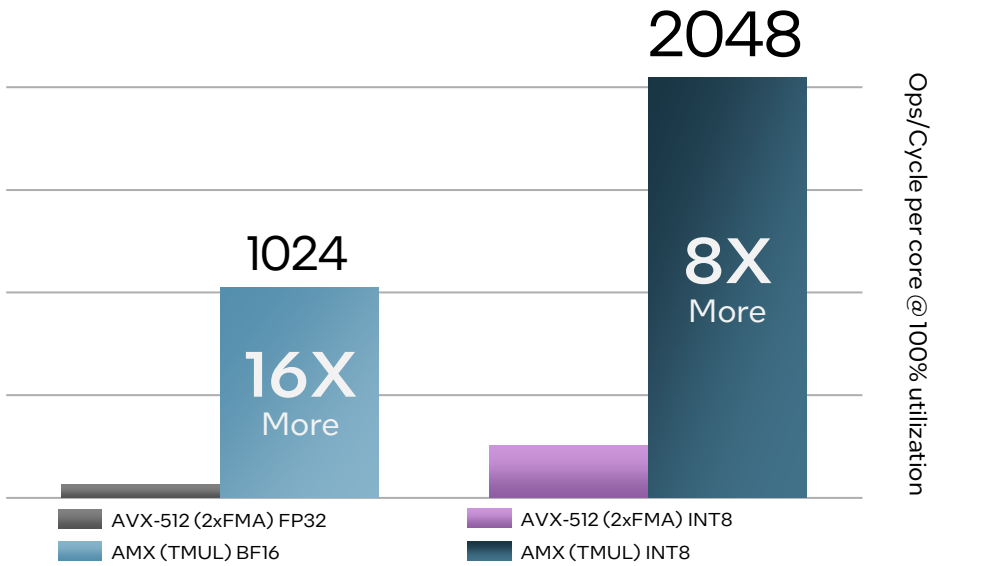
Intel® AMX architecture has two components:

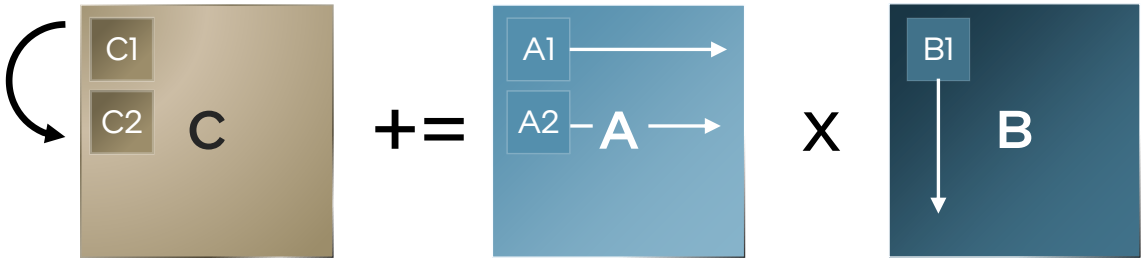
Tiles

- A new expandable 2D register file – 8 new registers, 1Kb each: T0-T7
- Register file supports basic data operators – load/store, clear, set to constant, etc.

TMUL

- Set of matrix multiplication instructions, the first operators on TILEs
- A MAC computation grid calculates "tiles" of data
- TMUL – performs Matrix ADD-Multiplication (C=+A*C) using three Tile registers (T2=+T1*T0)
- TMUL requires TILE to be present
- TILES declares the state and is OS-managed by XSAVE architecture





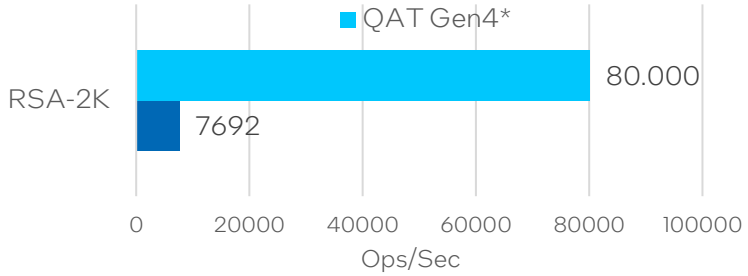
Express more work per instruction and per µop – save power for fetch/decode/OOO

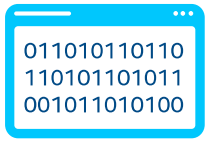
Intel® QuickAssist Technology (Intel® QAT)

QuickAssist Technology Gen 4 (QAT 4)

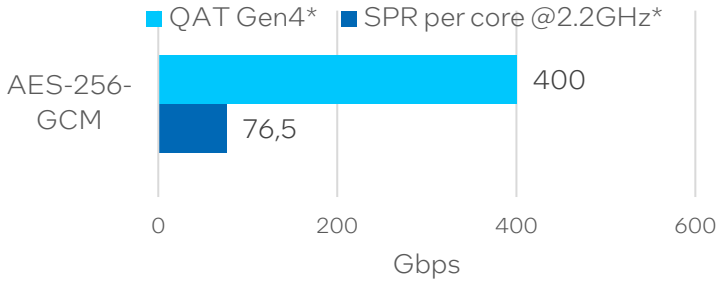
By offloading bulk cryptography, public key cryptography & compression to the Quick Assist Technology (QAT) accelerator, businesses can improve the performance of their infrastructure without investing in specialized hardware. QAT enables significant gains in CPU efficiency, data footprint reduction, power utilization and application throughput.

Public Key Cryptography

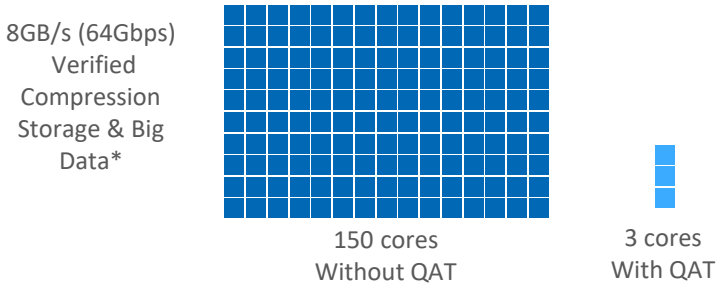




Cryptographic Ciphers & Hash/Auth



Compression/ Decompression



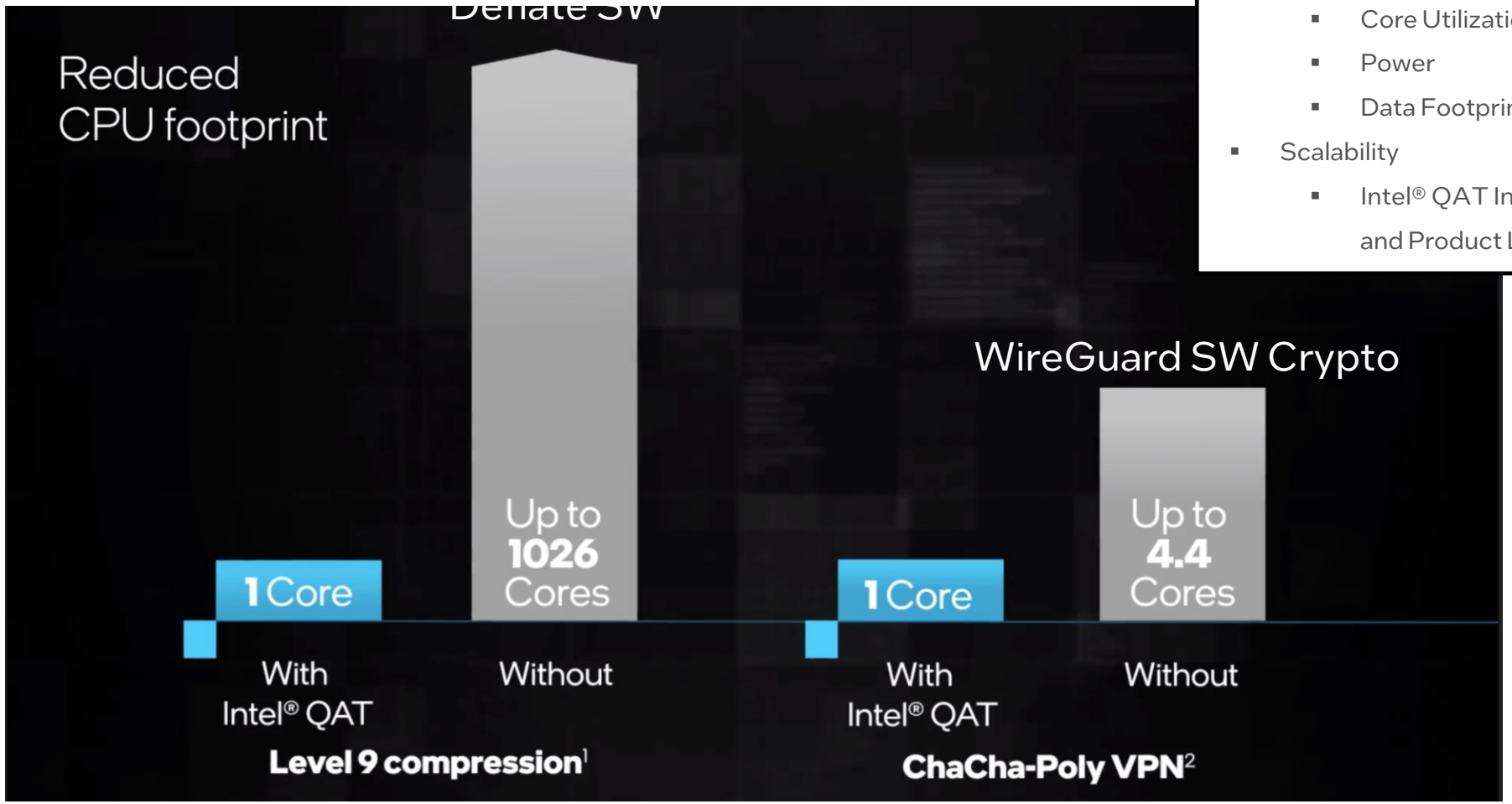
QAT OVERVIEW:

- Intel® QAT is an encryption and compression offload engine that delivers 200Gbs Crypto, 160Gbs verified compression, 100kops PFS ECDHE & RSA 2K Decrypt
- Application developers can access QuickAssist features through the Intel QuickAssist API. The API enables easy interfacing between the customer application and the QuickAssist acceleration driver.
- Target use cases include distributed storage systems (Ceph), file systems (BTRFS, ZFS), MSFT Azure Cosmos DB, Rocks DB, data lakes, Apache Spark, Hadoop, RDBMS and http compression.
- Up to 4 QAT instances per socket
- Scalable I/O Virtualization (SIOV) will not be OS enabled for Intel® QAT. Single Root I/O Virtualization (SR-IOV) will continue to be supported.

*Results shown are Sapphire Rapids (SPR)-based (not Xeon 6 CPUs) and were estimated using internal Intel analysis. They are provided for informational purposes only. Differences in system hardware or software design or configuration may affect actual performance.

Efficiencies of Intel QAT

- Networking & Compression Application Benefits
- Performance
  - Throughput
- Efficiency
  - Core Utilization
  - Power
  - Data Footprint
- Scalability
  - Intel® QAT Integration throughout the SKU Stack and Product Lines



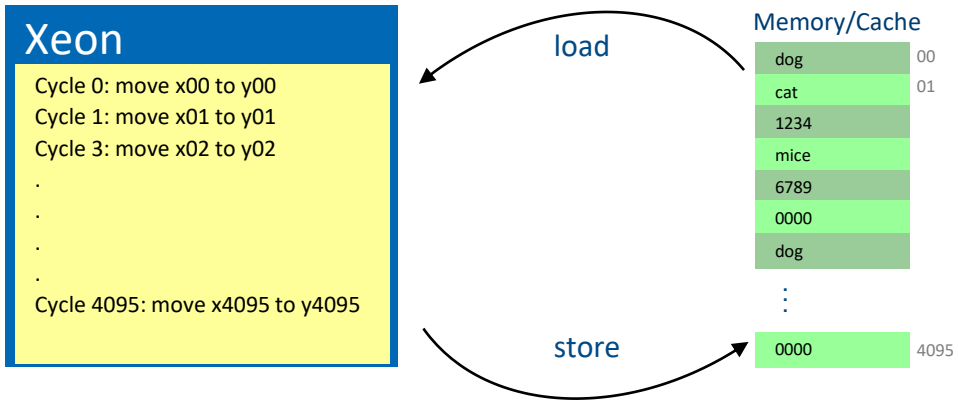
Intel® Data Streaming Accelerator (DSA)

Data Streaming Accelerator 2.0 (DSA 2.0)

Frequent data movement and transform operations in server workloads introduce overheads by consuming CPU cycles otherwise used for application processing. DSA is a high-performance integrated data mover accelerator on Xeon SoCs, capable of low-overhead offload of data movement, freeing CPU cores for higher workload performance and efficiency.

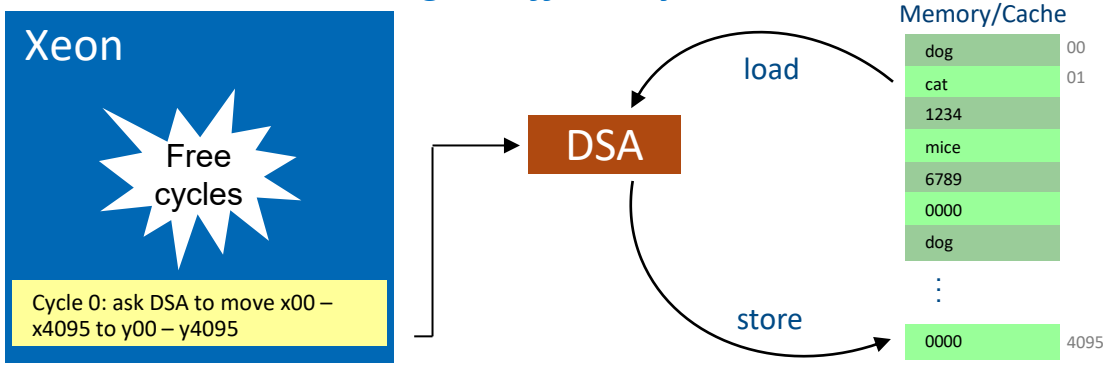
Memory Copy Without DSA

Core moves data. Core cycles consumed.



Memory Copy With DSA

DSA HW moves data with higher efficiency



DSA Overview:

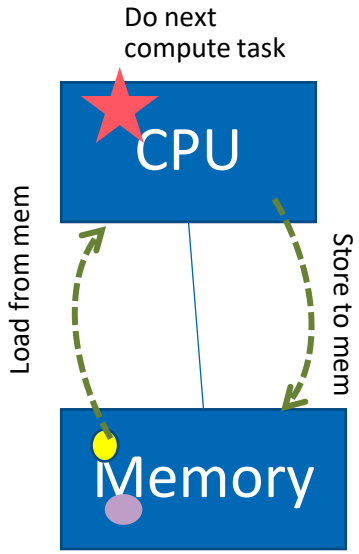
- Offloads data movement and transform operations (move, fill, compare, CRC, DIF, delta, flush) from CPU cores to accelerator HW
- Broad usages across network acceleration, storage acceleration, data center tax reduction, memory-tiering, and workload acceleration
- Up to 4 DSA instances per socket
- Supported by Intel® DML, Intel® MPI, Libfabric, SPDK and DPDK libraries
- Supported on bare metal only (no virtualization support)

New Xeon 6 Enhancements:

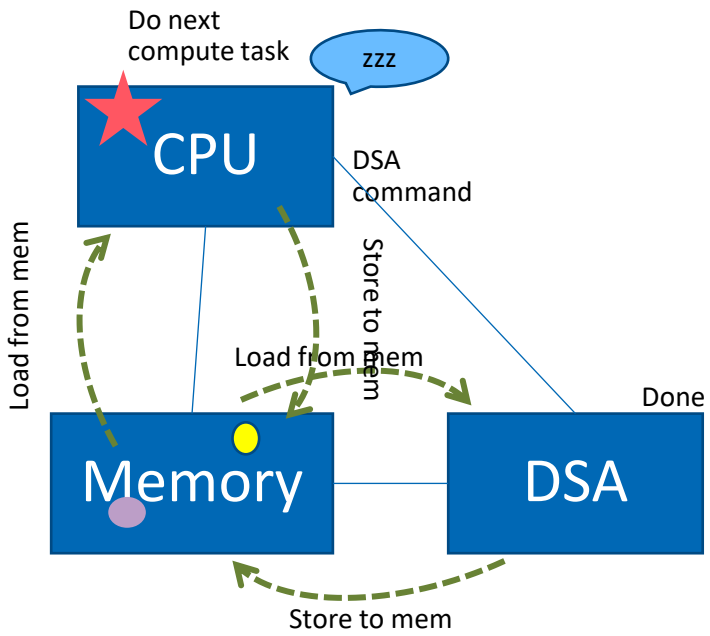
- Bandwidth doubles to ~60 GB/s per direction per instance for high-speed data ingestion, staging and replication
- Inter-address-space data movement extensions
- 64-bit CRC and DIX operations

Public spec
Open source blog

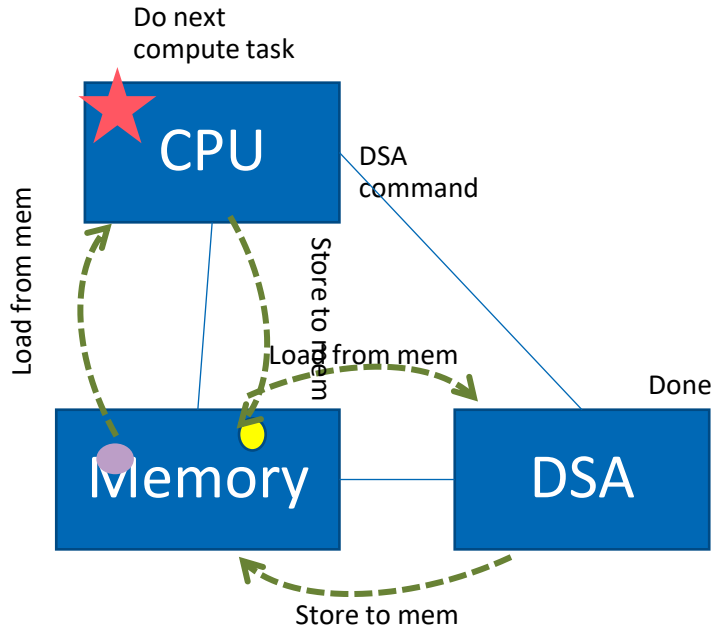
CPU → DSA Offload Storyline



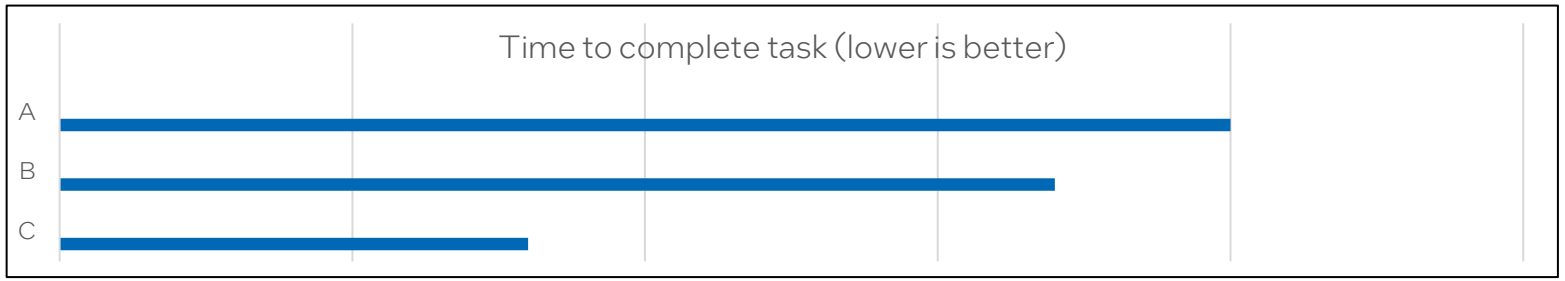
Ⓐ Data movement and compute operations execute serially on CPU core(s)



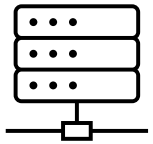
Ⓑ Data movement offload to DSA followed serially by compute on CPU core



Ⓒ Concurrent execution of data movement using DSA and compute on CPU core



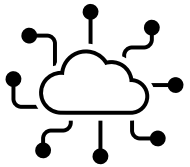
Intel® DSA – Usages & SW abstractions



Network Acceleration

Storage Acceleration

Datacenter Tax Reduction



Workload Acceleration

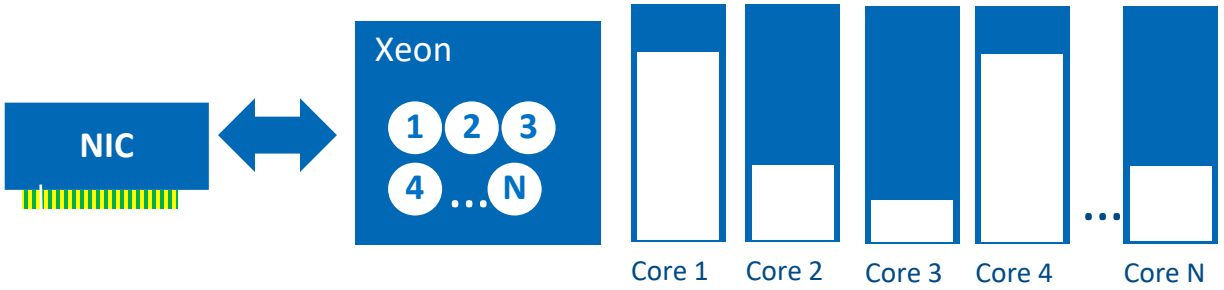
Example Usages			
• vSwitch (virtual switch) acceleration for VM network • Container networking acceleration • User space network stack acceleration for NGINX	• Data integrity check & CRC acceleration • Data replication over NTB acceleration • I/O acceleration with io_uring	• VM boot acceleration with memory zeroing offload • Tiered memory acceleration with CXL attached memory • Live migration acceleration for virtual machines	• Media transport acceleration for visual cloud applications • oneCCL acceleration for AI training applications • MPI acceleration for HPC applications
Higher-Level Software Libraries/Stacks			
Open VSwitch Calico	Enterprise Storage Stacks Distributed Storage (DAOS)	OS Memory Manager	Media Transport Library Intel oneCCL
Low-Level Libraries			
DPDK VPP	SPDK Linux io_uring	OS Kernel DMA APIs	OFI libfabric library Intel MPI Intel Data Mover library Intel DSA Proxy library

Intel® Dynamic Load Balancer

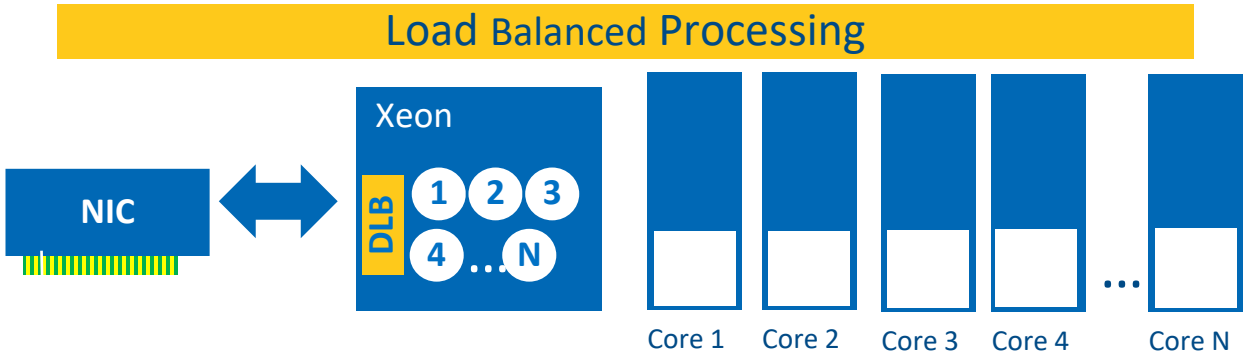
Dynamic Load Balancer 2.5 (DLB 2.5)

Processing network traffic, particularly extremely large continuous streams, aka Elephant flows, can occupy a disproportionate amount of a core(s) resources which causes uneven packet handling performance. DLB dynamically distributes and re-orders traffic to maintain reliable and consistent results.

Without DLB: CPU Utilization Per Core



With DLB: CPU Utilization Per Core



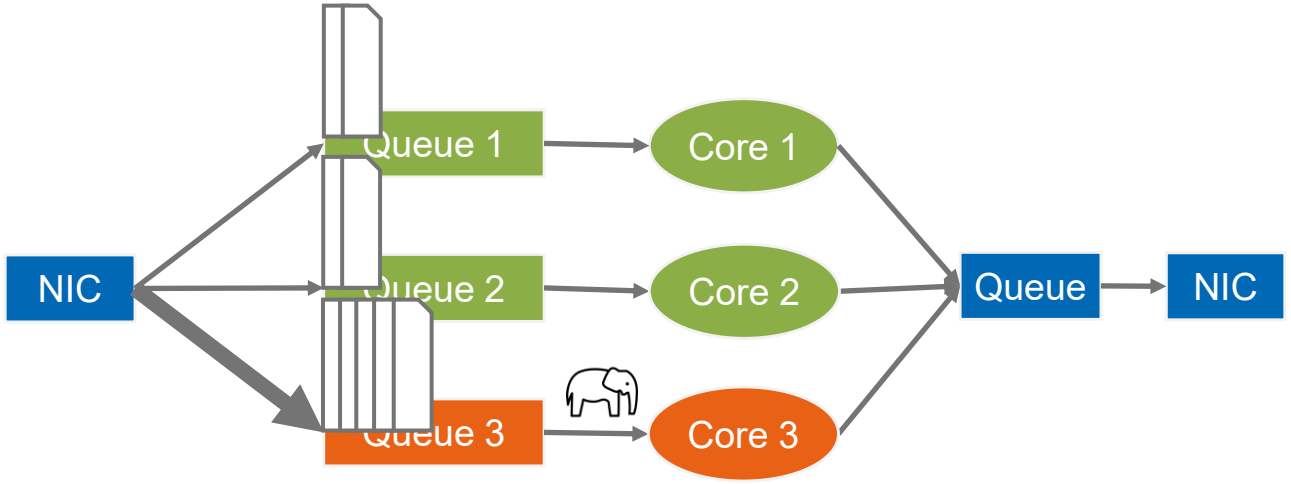
DLB OVERVIEW:

- Dynamic Load Balancing: Adjusted distribution in real-time as core loads vary.
- Dynamic Network Processing Reordering: Restores the order of networking data packets after processing.
- Target use cases include IPSec security gateway, VPP router, UPFc, vSwitch and Elephant flow handling.
- Up to 4 DLB instances per socket.

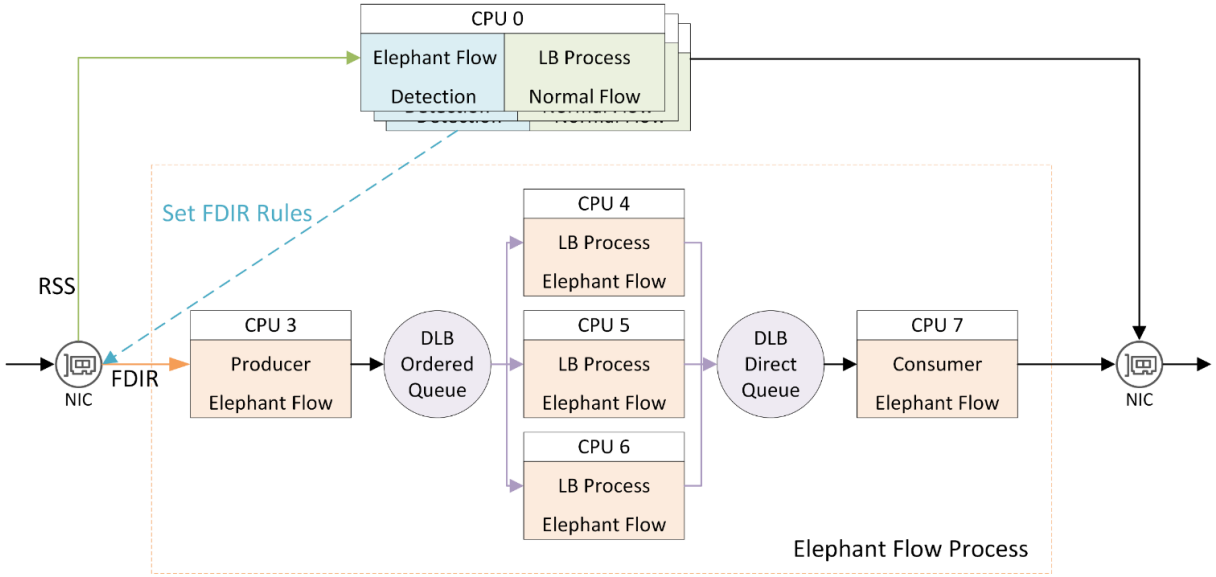
NEW Xeon 6 ENHANCEMENTS:

- Adds a merged credit scheme which simplifies software and reduces offload (API) cost.
- QoE weighting which allows for better quality load balancing (helps reduce latency spikes).

Elephant Flow Management in Security Load Balancer

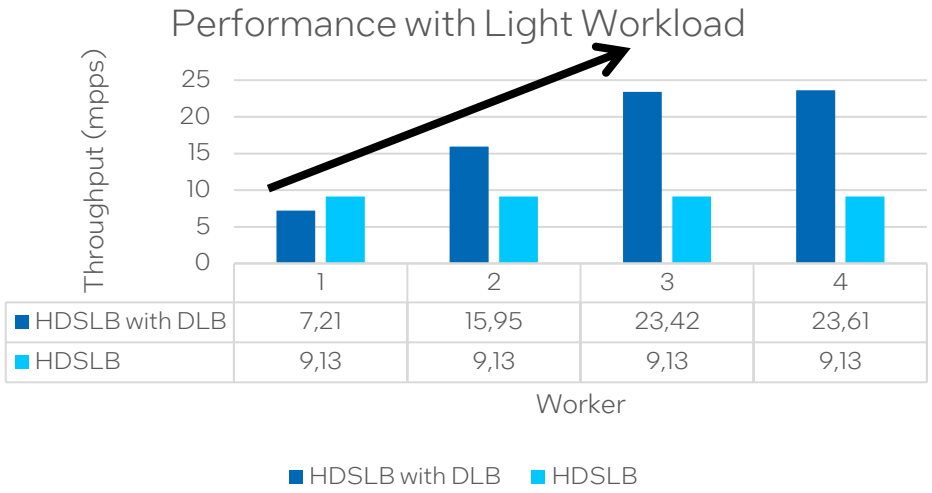


With DLB:



DLB PoC with leading cloud vendors

Shows greater throughput with the addition of DLB distributing packets across cores.



Why it matters to customers:

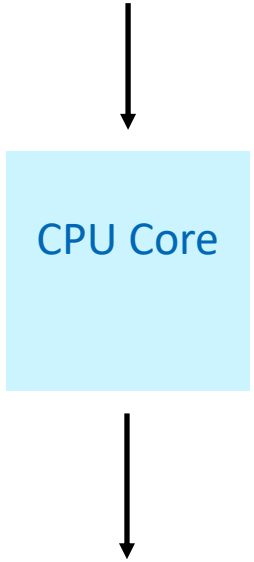
Higher Performance – More Efficient distribution of flows to workers results in higher performance vs SW solutions.

In-Memory Analytics Accelerator (IAA)

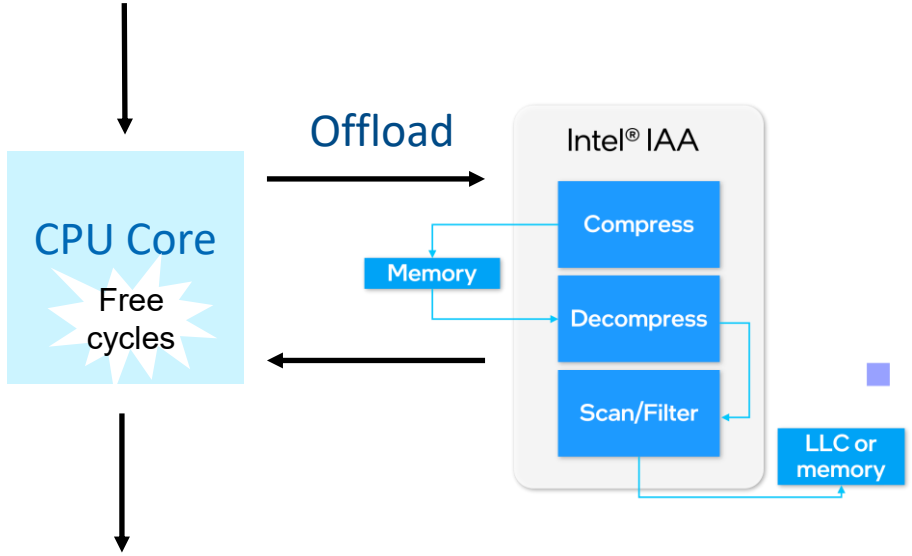
In-Memory Analytics Accelerator (IAA)

IAA increases analytic queries per second while decreasing memory footprint, resulting in faster power efficient analyses and better CPU utilization.

Without IAA:
Run compression / analytic algorithms on core without acceleration



With IAA:
Run algorithms on IAA



IAA Overview

- The Intel In-Memory Analytics Accelerator (Intel IAA) is a hardware accelerator that provides high throughput compression and decompression, analytic primitive functions (Scan/Filter) and data integrity Cyclic Redundancy Check (CRC)
- IAA targets big data applications and in-memory analytic databases, as well as application-transparent usages such as memory page compression
- Up to 4 IAA instances per socket
- Supported by Intel® Query Processing Library (Intel® QPL)
- Supported on bare metal only (no virtualization support)

New Xeon 6 Enhancements

- Device bandwidth increases 2X
- Improved Deflate algorithm decompress

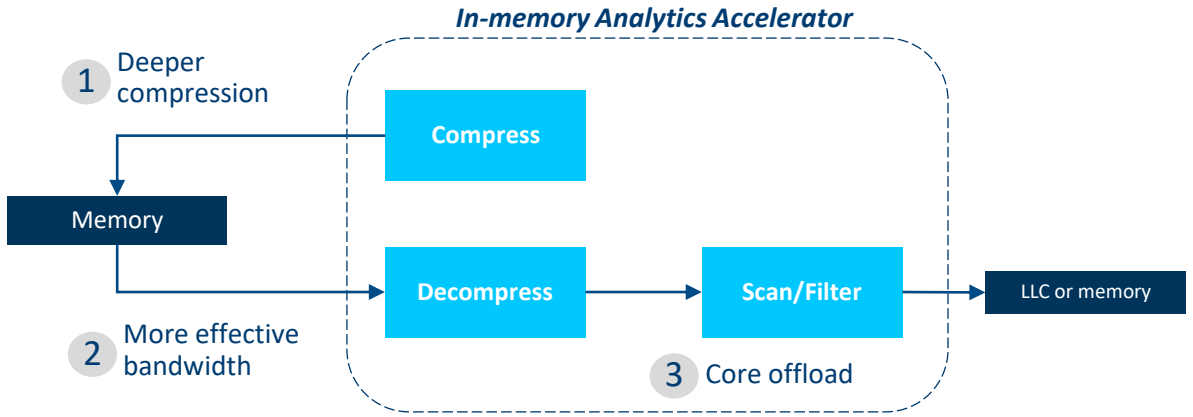
Intel® In-Memory Analytics Accelerator

Increasing query throughput and decreasing memory footprint for analytics

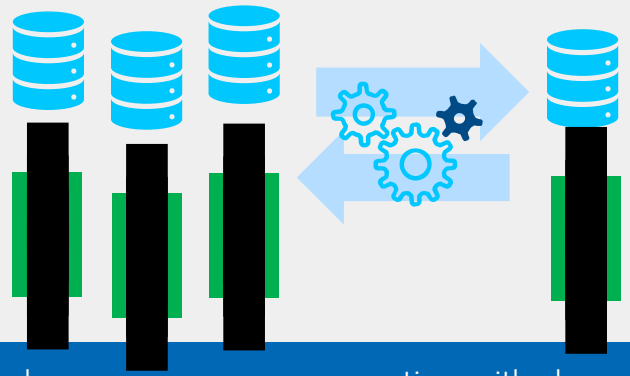
Intel In-Memory Analytics Accelerator (Intel® IAA)* integrates hardware acceleration of compute intensive workloads

Accelerates operations foundational to data analytics workloads by offloading to Intel IAA hardware device(s)

Enables significant gains in CPU efficiency, reduction of memory footprints, and in-memory database performance



Compression/ Decompression



Reduces memory consumption with deeper compression vs. software-only methods

Analytics Primitives
Scan, filter, etc.

Row ID	Fruit	State	Price $
1	Orange	FL	0.85
2	Apple	NC	0.45
4	Peach	SC	0.60
5	Grape	CA	1.25
18	Lemon	FL	0.25
19	Strawberry	CA	2.45
23	Blueberry	ME	1.5

Identifies relevant data in large data sets, accelerating database query throughput

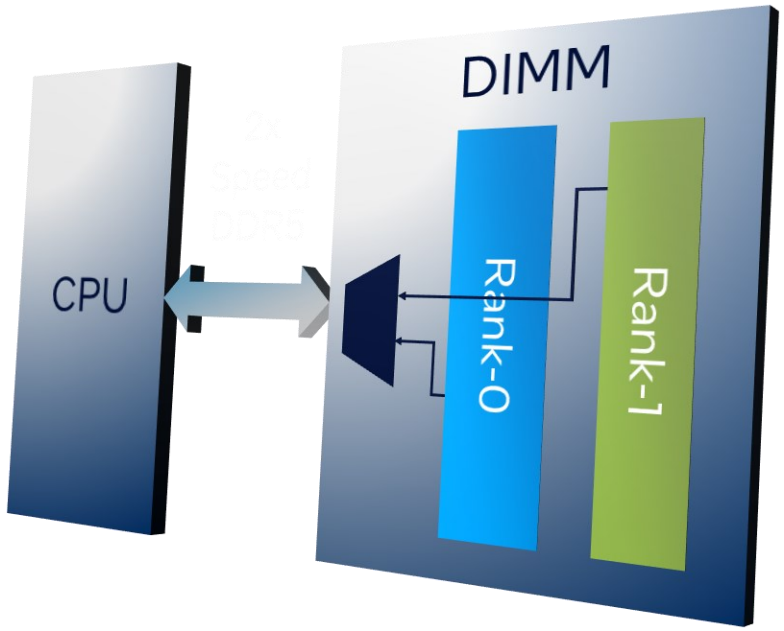
CRC Calculations

Detects accidental changes to raw data to help prevent data corruption

*Naming Note: Intel IAA previously referenced in earlier customer communications as IAX

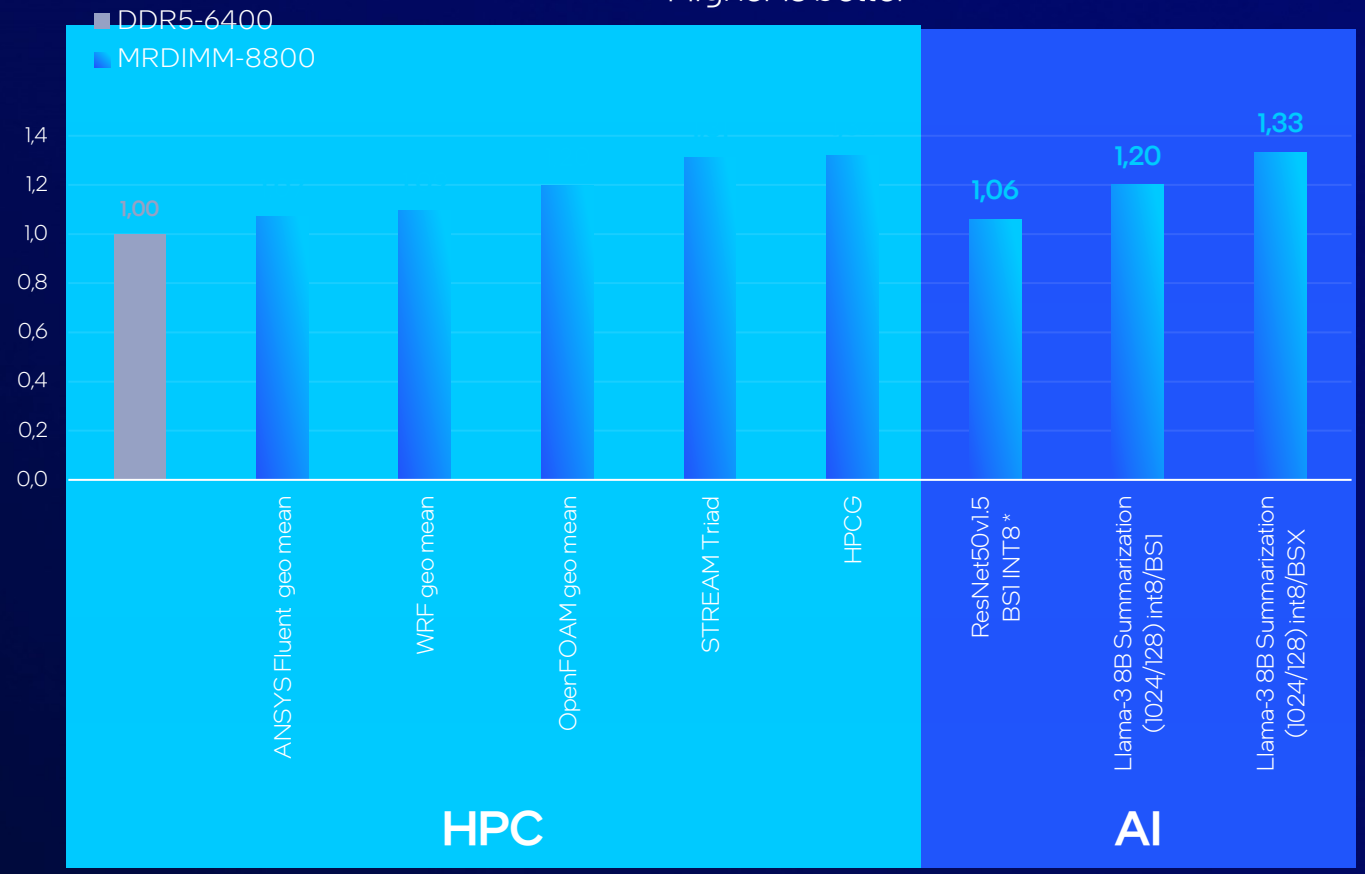
Multiplexed Rank DIMMs

First to market on Intel Xeon 6 processors with P-core



MRDIMMs deliver up to **8800 MT/s** data rate on Intel Xeon 6 with P-cores

Intel® Xeon® 6 with P-cores (128c)
MRDIMM-8800 Performance Gains Over DDR5-6400
Higher is better



See intel.com/processorclaims: Intel Xeon 6. Results may vary. * 6972P (96c) used.
This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark

NEMO Success with Intel® Xeon® 6 Processor

Key uses for NEMO are simulating:

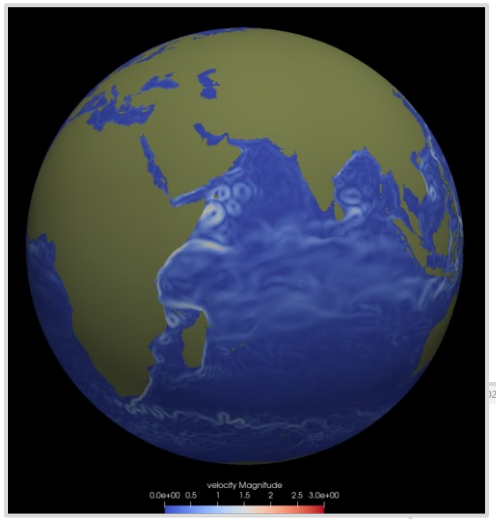
- Volume weighted ocean temperature
- Global average sea level change
- Sea surface temperature, comparing to well-known datasets from observational sources such as HadISST
- Sea water salinity and sea ice time evolution
- Other thermodynamics and biogeochemistry metrics of ocean and sea-ice

Intel ran NEMO global ocean circulation model on Intel® Xeon® 6 processors with P-cores and 5th Gen Intel Xeon CPUs

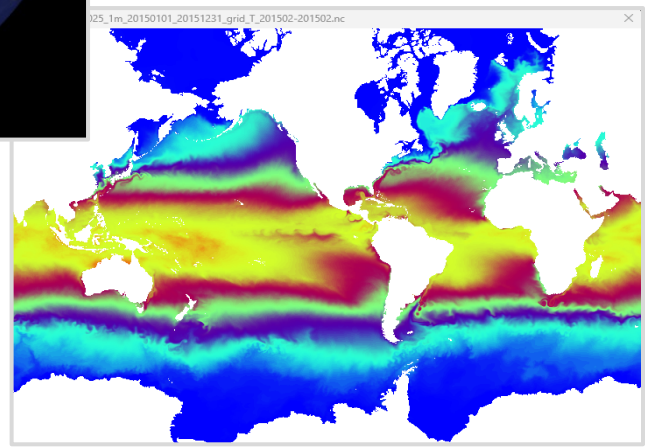
Intel® Xeon® 6 processor with P-cores CPU combined with MRDIMM:

2.3x faster[1]

vs. 5th Gen Intel® Xeon® CPU with traditional DDR memory



Visualization of NEMO output result files is done with open-source software (NCView/ParaView). Below is an example of visualization screenshot, which shows the global sea surface temperature distribution - Intel Xeon 6900 processor with P-cores takes 2.3x less wall time[1]



Georgia Tech

[1] See "Memory Performance" in backup for workloads and configurations. Results may vary.

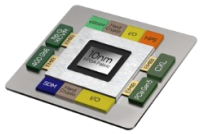
The Challenge

Today, researchers and developers wanting to utilize the power of the highest-performance processors are forced to compromise on openness.

CPU

GPU

FPGA



Other Accelerators

48% of developers target heterogeneous systems that use more than one kind of processor or core[1]

Developer Challenges: Multiple Architectures, Vendors, and Programming Models

This means:

- You lose the freedom to integrate different technologies together.
- You lose the freedom to choose the best processor for your workload.
- It's much harder to bring new innovations in performance to market.

Source 1: Evans Data Global Development Survey Report 23.2 2023

Vision to Reality

Unify the heterogeneous compute ecosystem around open standards

2014 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024

SYCL™

Language for open portability

oneAPI™

Build out an open ecosystem at scale

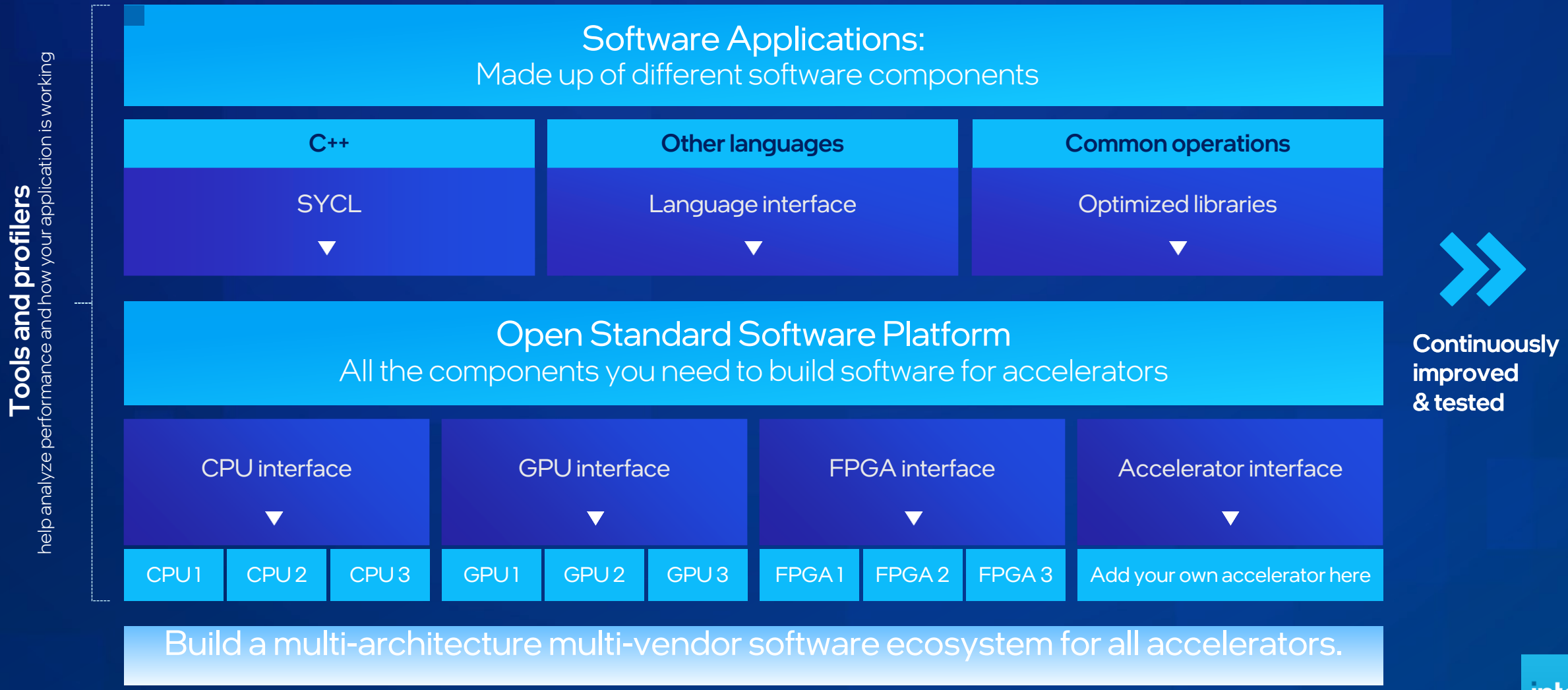
- oneDPL — Data Parallel C++ Library
- oneDNN — Deep Neural Network Library
- oneCCL — Collective Communications Library
- oneDAL — Data Analytics Library
- oneTBB — Threading Building Blocks
- oneMKL — Math Kernel Library
- Level Zero

UXL — Unified Acceleration Foundation

Open industry governance

intel.

The Solution: An Open Software Platform



Progress: Ecosystem Momentum

60+

CUDA+ to SYCL+ Catalog of Ready-to-Use Applications.[1]

- Artificial Intelligence
- Image and Video Processing
- Healthcare
- High-Energy Physics
- Computing
- Biology
- Geology and Geography
- Finance
- Aerodynamics and Fluid Dynamics
- Math

Source: CUDA* to SYCL* Catalog of Ready-to-Use Applications.
https://www.intel.com/content/www/us/en/developer/tools/oneapi/training/migrate-cuda-to-sycl-library.html#gs.8nyhtc

Software Porting Momentum

HPC at Argonne National Laboratory

it starts with

Source: Image source: https://www.anl.gov/article/us-department-of-energys-incite-program-seeks-proposals-for-2024-to-advance-science-and-engineering

Aurora Programming Models

Reliance on Open Standards

- ALL applications on Aurora use Open Standard & portable programming models allowing researchers flexibility to do what they need.
- Argonne actively participates in driving open standard programming models & languages: SYCL, OpenCL, C++, OpenMP, MPI.
  - Approx. ½ leverage SYCL either directly or through a higher-level programming model: Kokkos, OCCA, RAJA
  - 100% of applications that use the GPU do so through SPIR-V
  - openCL used directly or as an alternative backend for SYCL, OpenMP
  - LLVM ecosystem is the basis of Aurora software tool chain – performant compilers and runtimes
  - PyTorch and TensorFlow models and ecosystem supported and accelerated with oneAPI
  - Scalable AI using DDP, DeepSpeed, Horovod accelerated with oneCCL
  - Python ecosystem, including NUMBA, supported and accelerated using oneAPI

Probing the Universe with Machine Learning-Enhanced, Extreme-Scale Cosmological Simulations

Application code: CRK-HACC

- CRK-HACC simulates the formation of large-scale structures in the Universe over cosmological time.
- Employs n-body methods for gravity and a novel formulation of Smoothed Particle Hydrodynamics (CRK-SPH) for baryons.
- A mixed-precision C++ code, with FLOPS-intense sections implemented using architecture-specific programming models in FP32.

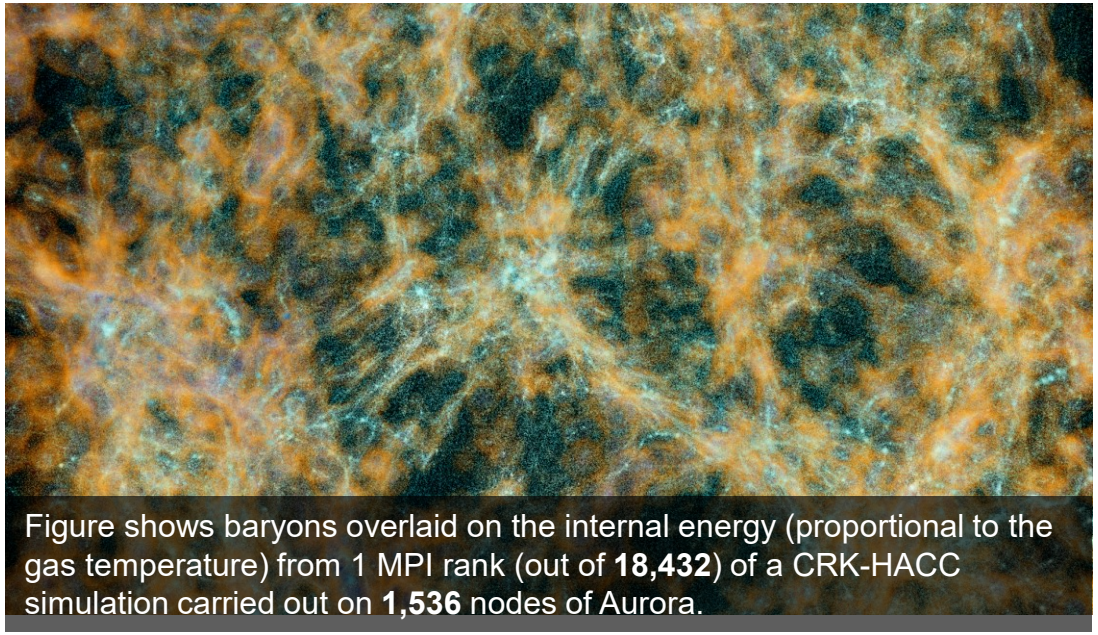
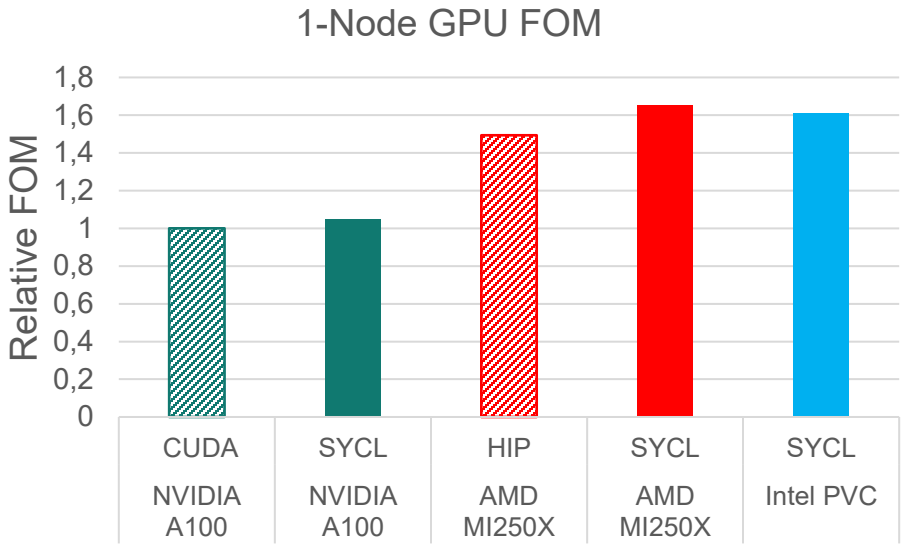


Figure shows baryons overlaid on the internal energy (proportional to the gas temperature) from 1 MPI rank (out of **18,432**) of a CRK-HACC simulation carried out on **1,536** nodes of Aurora.



- CUDA and HIP implementations are maintained as a single source with macros, SYCL kernels were translated from CUDA using *SYCLomatic*. The SYCL sub-group "shuffle" operation was optimized[1] for the Intel PVC GPU.
- Relative Figure-of-Merit (FOM) is the number of particle-steps per second, with the A100 CUDA implementation run on Polaris (A100) as baseline.

E. Rangel, S. J. Pennycook, A. Pope, N. Frontiere, Z. Ma, and V. Madananth. A Performance-Portable SYCL Implementation of CRK-HACC for Exascale. https://doi.org/10.1145/3624062.3624187

We have a dream!

"one code, many architectures"

Learnings using SYCL on Aurora

THE SETUP for One Node

"Many vendors, one code, happy end"

1. Polaris
   - 1-socket AMD EPYC 7543P, 32 cores + 4 NVIDIA A100-SXM4-40GB
2. Frontier
   - 1-socket AMD EPYC 7A53 64 cores + 4 AMD Instinct MI250X
3. Aurora
   - Dual-socket Intel Xeon CPU Max 9470C, 52 cores + 6 Intel GPU Max 1550

MAJOR RESULTS & FINDINGS

SYCLomatic: CUDA migration to SYCL

1. Diagnostics alerts of non compatible code between CUDA vs SYCL

2. CUDA intrinsics could be safely removed

3. Math funcionts with different precision

4. Adjustments in some wapers for HIP/CUDA

5. Kernels as a Function Object to allow C++ features: templates, class, inheritance

MAJOR RESULTS & FINDINGS

First run, first learning!

1. **First version shows SYCL outperforming CUDA and HIP by far, why?**
   - oneAPI DPC++ compiler set “fast math” by defaul, but nvcc and hipcc do not

2. **But some questios remains:**
   - Even with “fast math” for all compilers, SYCL still outperforms, why?
   - Based on theoretical peak performance, Intel and AMD GPUs were supposed to deliver more FLOPS!

MAJOR RESULTS & FINDINGS

WARPS and SYCL sub-groups

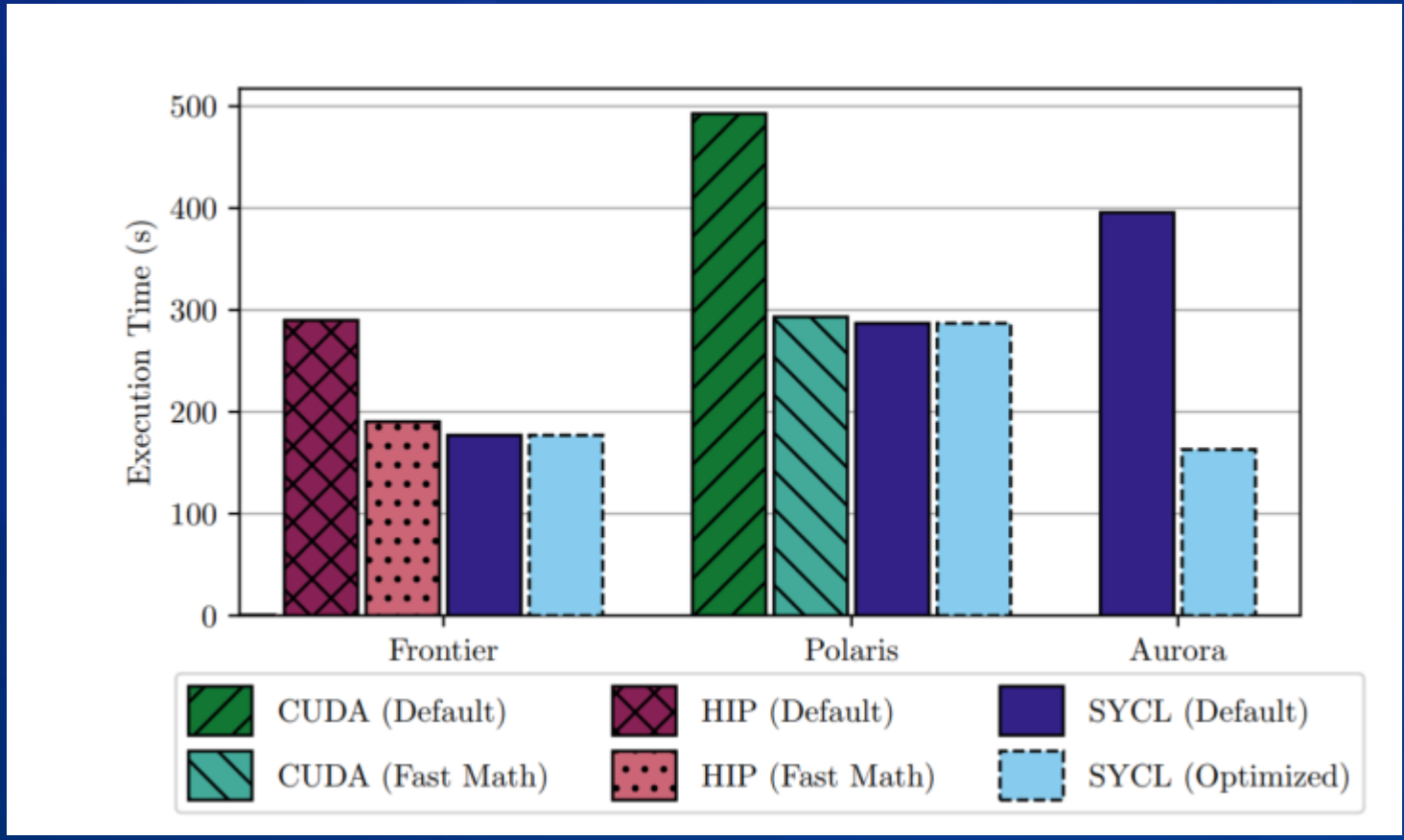
- NVIDIA has a warp size of 32 threads
- SYCL has work-groups divided in sub-groups ranging from 16, 32 to 64 threads
- AMD GPUS supports 32 or 64
- Intel GPUs supports 16 and 32

Setting a sub-group of 16 on Intel GPUs resulted in a 2.4x speedup

Setting sub-group size via macros for each hardware is important!

MAJOR RESULTS & FINDINGS

The initial performance of the migrated SYCL code compared to CUDA, HIP, and our optimized SYCL code *



*** Source: E. Rangel, S. J. Pennycook, A. Pope, N. Frontiere, Z. Ma, and V. Madananth. A Performance-Portable SYCL Implementation of CRK-HACC for Exascale. https://doi.org/10.1145/3624062.3624187**

Courtesy of Argonne National Laboratory

intel

MAJOR RESULTS & FINDINGS

SYCL version of CRK-HAAC resulted:

1. Performance portability of 96%
2. Code Divergence near Zero
3. Viable programming model for performance-portable applications
4. Developers have choice about Portability vs Performance

Intel Software Developer Tools Use Cases

Multiarchitecture Performance & Productivity Value for Customers

HPC

Argonne rolled out Aurora performance using Intel® Data Center GPU Max Series

TACC's Frontera Supercomputer uses oneAPI to accelerate exascale scientific computing

Univ.of Cambridge strives for zettascale using oneAPI

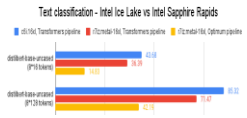
Accelerating Google Cloud for HPC Video | Podcast

Alibaba Cloud DP Technology 深势科技

Using Alibaba's E-HPC Cloud Service and Intel® hardware and software, DP Technology achieves 45.2% performance improvement.

AI/ML/DL

Red Hat optimizing Data Science Workflows



Scaling HuggingFace Transformer and Optimum Performance with Intel AI

Optimize performance of IBM Watson with NLP & NLU

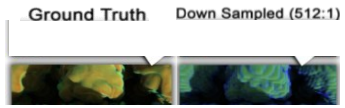
PyTorch Foundation

Advance PyTorch through Intel Optimizations

Rendering

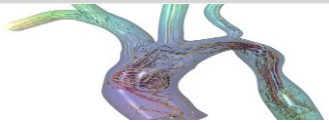
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Univ. of Tennessee used oneAPI to enable a cloud-based Rendering-as-a-Service (RaaS) environment



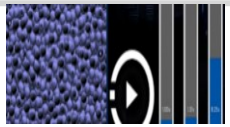
Univ. of Calif. at Davis increased performance by 3x & delivered 100x data compression for scientific rendering[1]

Stephen Hawking Centre for Cosmology Visualizes Cosmos Physics

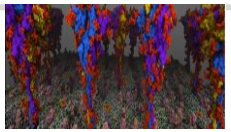


Duke University, Oak Ridge and Argonne National Lab 10x Performance gain using Intel GPUs

CUDA* Code Migration to SYCL*



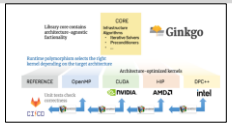
Preparing NAMD molecular dynamics for Aurora Supercomputer



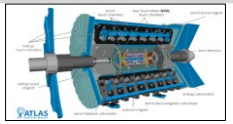
LAMMPS: Speeding Exascale Material Discovery

University of Stockholm GROMACS 2022

University of Utah Massive Image Dataset Binding Using SYCL



Ginkgo and oneAPI accelerate numerical simulation using Intel GPUs



The ATLAS Experiment Implements Heterogeneous Particle Reconstruction with Intel® Tools

1. See Notices & Disclaimers for configuration details. Refer to software.intel.com/articles/optimization-notice for more information regarding performance & optimization choices in Intel software products. For workloads and configurations visit www.Intel.com/PerformanceIndex. Results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. *Other names and brands may be claimed as the property of others. SYCL is a trademark of the Khronos Group Inc.

Available for download or in the cloud

Run the tools locally

visit intel.com/oneAPI

Downloads

Repositories

Containers

Code Samples, Quick-start Guides, Webinars, Training

Run tools in Intel® Tiber™ AI Cloud

cloud.intel.com

- No hardware acquisition
- No download, install or configuration
- Sample code & documentation
- Ready-to-use deployment (AI & compute) & development environments
- Access to cutting-edge learning resources

Professional and Community Support Available

- Download or run tools in the cloud for free
- Every paid version of Intel® oneAPI Base, HPC, and Rendering Toolkit products includes Priority Support
- Intel Tiber AI Cloud offers standard, Premium and Enterprise service tiers

Intel® Tiber™ AI Cloud

A platform where you can develop and deploy AI models, applications and services at scale with best price-performance.

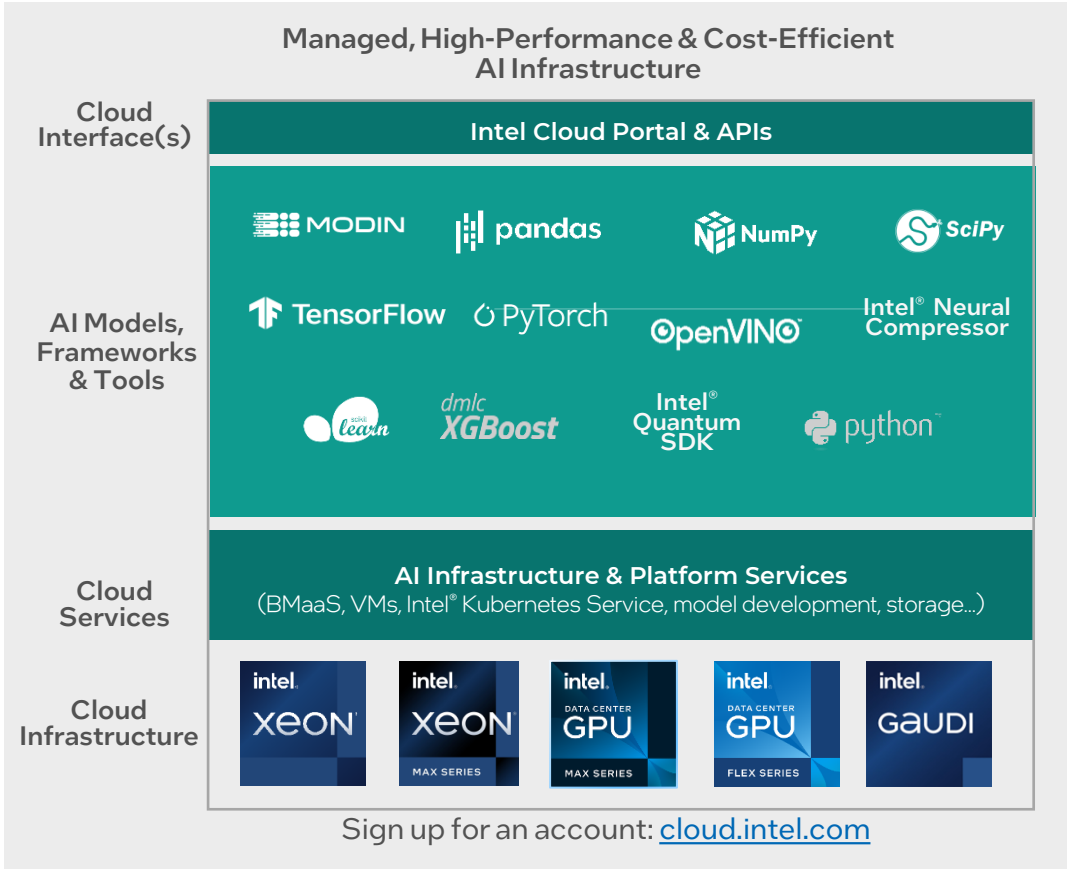
- Build and deploy AI at scale
- Maximize AI compute resources
- Evaluate hardware
- Open Software, Open Ecosystem Advantage

For Developers Provides an **easy path for developers** to access and use powerful Intel® architecture (CPUs, GPUs, Intel® Gaudi® AI accelerators) and Intel-optimized AI software.

For Enterprises Create **new software, use for compute, deploy AI services. Accelerate adoption and deployment** of Intel products.

For Partners Provides performance **and cost-optimized Intel AI compute services** to their customers.

Free cloud credits may be available
Contact your Intel representative



Sign up for an account: cloud.intel.com

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

Intel® Tiber™ AI Cloud

Build & deploy AI models, applications and services at scale with best price-performance

For Developers

Get started with Intel and Intel-powered AI

Evaluation

Education

Development

Research

For Companies

Deploy AI at scale

Certification and benchmarking

AI training and inference production workloads

For Partners

AI Infrastructure services for SaaS providers

Intel compute services for third-party AI SaaS

Cloud interface(s)

Intel Cloud Portal and APIs

Cloud Services

AI Infrastructure and Platform Services
(BMaaS, VMs, Intel® Kubernetes Service, model development, storage…)

Cloud Infrastructure

Intel® Tiber™ AI Cloud Service Tiers

Standard

Free

Explore Intel AI products

Included with standard:

- ✓ Evaluate latest Intel products
- ✓ Develop AI skills
- ✓ Access to cutting-edge learning resources
- ✓ Intel Community Support

Premium

Pay as you go compute

Single user access to the latest Intel products

Included with standard:

- ✓ Early pre-release hardware access
- ✓ AI / ML software toolkits
- ✓ Intel Premium Support

Enterprise

Committed use with discounts

Team access to latest Intel products

Included with standard:

- ✓ Billed subscription for teams
- ✓ Deploy inference on your own infrastructure.
- ✓ Use CPU, GPU, AI accelerators, ***
- ✓ 24 x 7 Intel premium plus support.

Professional and Community Support Available

Priority Support for Intel Toolkits

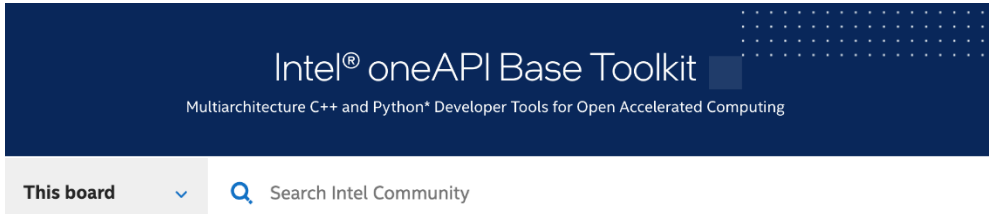
Every paid version of Intel® oneAPI Base, HPC, and Rendering Toolkit products includes Priority Support

- **Direct and private interaction** with Intel's support engineers, including the ability to submit confidential support requests
- **Accelerated response time** for toolkit-related technical questions and other product needs
- **Free download access** to all new product updates and continued access to older versions of the product
- **Ability to influence** product features and quality
- **Priority Support** for escalated defects
- **Access to a vast library** of self-help documentation that builds off decades of experience in creating high-performance code
- **Additional services at reduced cost**, including on-site or online training and consultation by Intel technical consulting engineers

Free Community Support

Connect with the Intel Community in public **Developer Software Forums**

- Supported by community technical experts and monitored by Intel Engineers
- Answers to commonly asked questions
- Access to online tutorials and self-help forums
- Troubleshooting guidance from fellow developers



Notices and Disclaimers

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein. The products described 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.

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

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 http://www.intel.com/benchmarks .

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure. Results have been estimated or simulated. Performance varies by use, configuration and other factors. Intel technologies may require enabled hardware, software or service activation. Your costs and results may vary. Learn more at www.Intel.com/PerformanceIndex

Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com.

All information provided is subject to change at any time, without notice. Intel may make changes to manufacturing life cycle, specifications, and product descriptions at any time, without notice. The information herein is provided "as-is" and Intel does not make any representations or warranties whatsoever regarding accuracy of the information, nor on the product features, availability, functionality, or compatibility of the products listed. Please contact system vendor for more information on specific products or systems.

All product plans and roadmaps are subject to change without notice. Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or visiting www.intel.com/design/literature.htm .


intel