intel ai

3D View of Intel AI

Maciej Hoffmann – NEX & DCAI Sales Account Manager
Łukasz Chrzanowski – Industry Technical Sales Specialist

Legal Notices and Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:

Bringing AI Everywhere

AI is transforming how we work and live everyday

From facial recognition to personalized learning and modern GenAI, Intel is putting AI to work across your enterprise.

Facial Recognition

Personalized Learning

AI Based Rendering

Video Conference

Purchase Recommendation

Code Generation

Robotics Vision

Inventory Management

Recommendation Systems

AI is evolving rapidly

Underlying data technologies....

- Structured Data
- Unstructured Data
- Data Fabrics
- Synthetic Data

58% of CEOs from leading public companies actively investing in AI[2]

70% By 2027, GenAI will be used to identify and replace legacy apps, reducing modernization costs by 70%[3]

$300B worldwide GenAI spending set to exceed $300B by 2026[1]

50% of edge deployments will involve AI by 2026[5]

80% of PCs to be AI PCs by 2028[6]

of enterprises will use Gen AI by 2026[7]

AI as disruptive as the Internet

Generative AI predicted to add up to $4.4T of value to global economy by 2040[4]

AI inferencing driving up compute costs; exceeding the pace of Moore's Law

Growth of **large model** sizes (1T+ parameter models)

Growth of **smaller, nimbler models** (~10B parameters)

1. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights
2. https://chiefexecutive.net/the-rise-of-the-ai-ceo/.
3. Gartner's Top Strategic Predictions for 2024 and Beyond.
4. Worldwide Artificial Intelligence Spending Guide (IDC)
5. Gartner®, Building an Edge Computing Strategy, Thomas Bittman, 12 April 2023. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All right reserved
6. Source: Boston Consulting Group
7. Gartner news release – Oct. 10, 2023.

The rapid growth of AI across the enterprise

Machine Learning	Deep Learning	Generative AI	AI Everywhere
Demand forecast	CV quality detection	Architecture designs	Sustainable AI
Anomaly detection	NLP – chatbots	Code generation	Process automation
Classifications	Medical image diagnosis	Marketing copy or images	Federated learning
Recommendations	...	Digital assistants	Path to AGI?
Fraud detection		...	...
...			

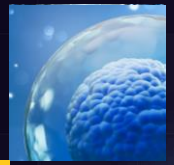
What AI needs

- Data, compute, networking, memory and algorithms
- Speed training and fine tuning on large and nimble compute clusters (from weeks to days to hours)
- Responsibly deploy and inference anywhere on all devices (milliseconds)

Why is AI challenging?

Complexity	Rapidly growing number of methods, capabilities, data types and sizes, and infrastructure requirements to run AI
Costs	Increasing costs due to increased compute demand as AI becomes more widely adopted and consumed
Operationalizing	Many steps and skill sets required to get AI from proof of concepts through to production in a scalable, sustainable process
Data security and privacy	Activating sensitive or regulated data globally while remaining secure and compliant
Human impact	Ensuring AI technology advances responsibly, ethically and equitably with a comprehensive approach that lowers risks, improves lives and optimizes benefits

Bringing AI everywhere



+

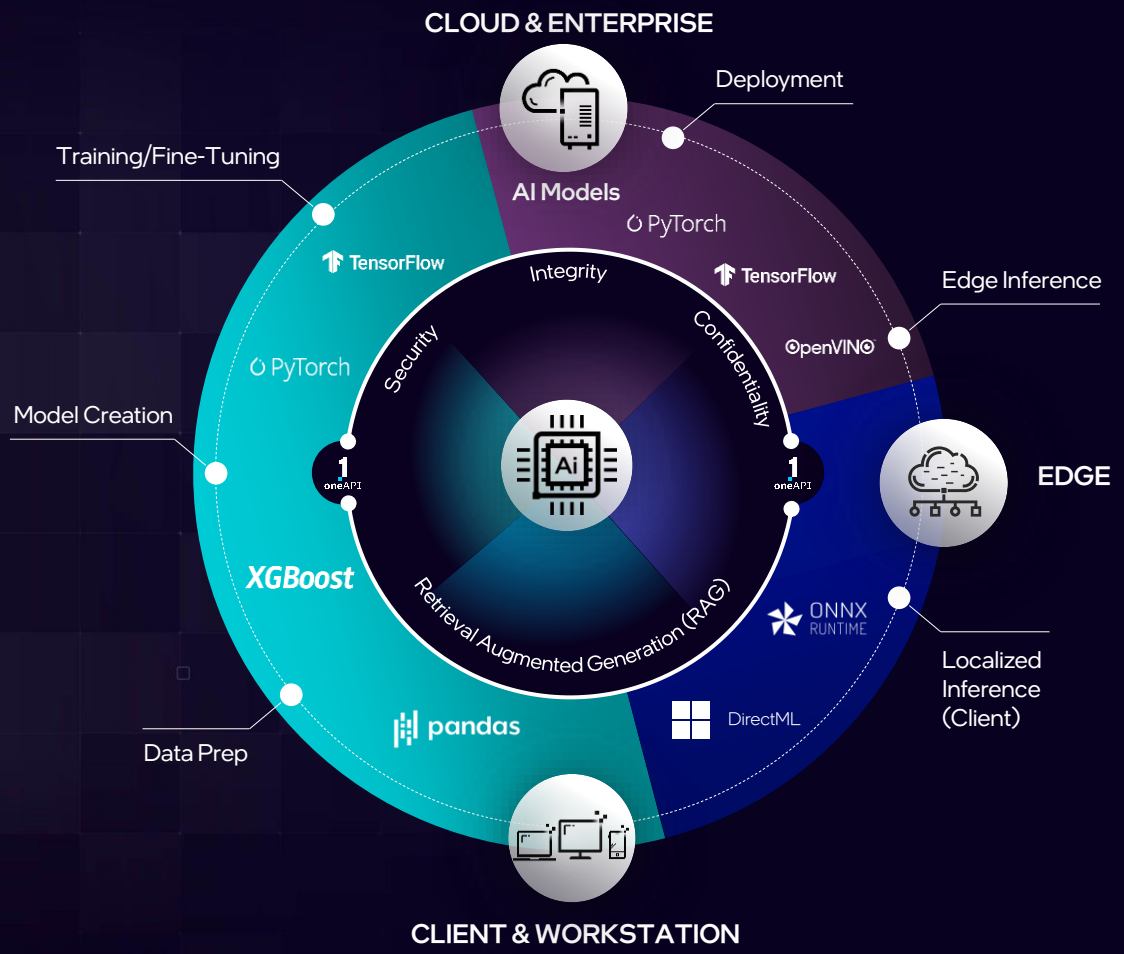
$

Intel® AI product positioning

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

AI Continuum

Bringing AI everywhere



Bringing AI everywhere

Complete AI systems strategy

Open Ecosystem Stack

- Open & Easy
- Secure & Responsible
- Scalable & Reliable
- Accessible

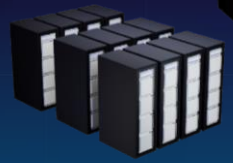
Bringing AI everywhere

Intel AI technology solutions

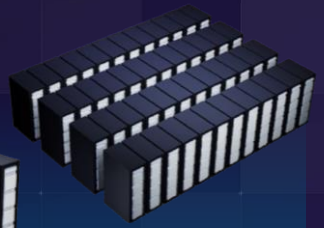
AI PC Node
Light Inference

Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inf.



Super Cluster
Training, Tuning, Peak Inf.



Mega Cluster
Large Scale Training & Inference

AI PC
Broadest AI SW ecosystem

Enterprise & Edge
Open standards, "Ready to Use"

Data Center
AI open, scalable systems & reference arch

AI PC powered by Intel® Core™ Ultra processors

3 Powerful AI Engines

GPU

High Throughput
Ideal for AI-accelerated digital content creation

NPU

Low Power
Ideal for sustained AI workloads and AI offload for battery life

CPU

Fast Response
Ideal for low-latency AI workloads

Unmatched AI Compute

With Intel Core Ultra 200V Series Processors

Up to **120** platform TOPS

GPU	Up to **67** TOPS	XMX & DP4a	Gaming & creator AI
NPU	Up to **48** TOPS	Dense vector & matrix math	AI assistants & creation
CPU	Up to **5** TOPS	VNNI & AVX	Light AI workloads

intel.

See intel.com/performanceindex for details.

Scaling the NPU

NPU 3

Increase number of engines

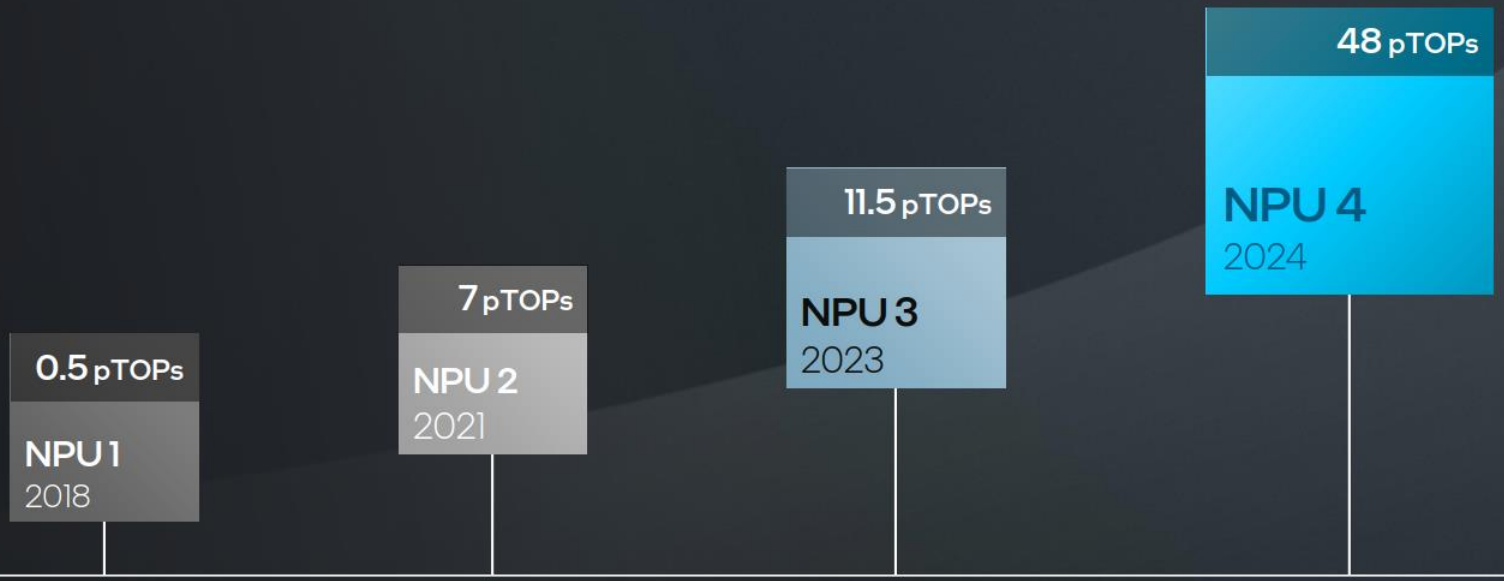
Increase frequency

Improve architecture

NPU 4

Continuous NPU Improvements

Across 4 generations of IP



All tops are Int8 on high end SKU, will vary based on SKU

Operation Types

Overview

	Scalar	Vector	Matrix
Complexity	1	N	N^2
Example functions	Conditional Looping	SoftMax Activation functions	Convolution Matrix multiplication
Occurrence in AI	Low	Very high	Very high

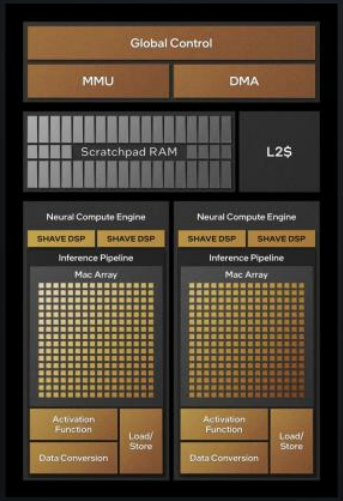
What is a TOP?

Trillions of **Operations** per **Second**

1 Multiply & 1 Accumulate (MAC)

Clock Frequency (Hz)

How Many AI TOPS in Meteor Lake's NPU?



Operations × **Frequency** ÷ **Trillions** = **TOPs**

- Operations: 1 Multiply & 1 Accumulate (MAC)
- Frequency: Clock Frequency (Hz)
- Trillions: 10^{12}

How Many AI TOPS in Meteor Lake's NPU?

Operations

2 ops/MAC *4096 MAC/cycle

×

Frequency

1.4 GHz

÷

Trillions

10^{12}

=

11.5 TOPs

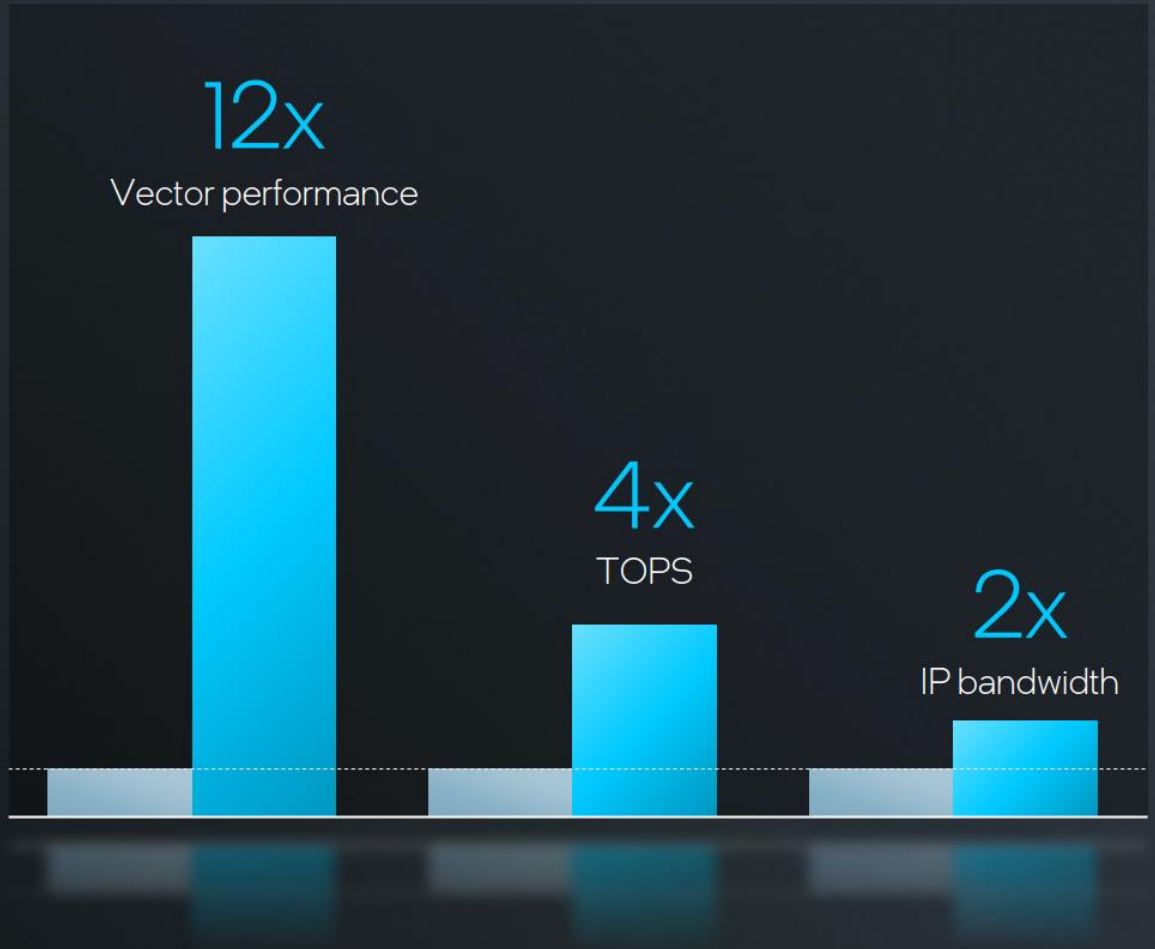
intel.

NPU 4

Performance

Intel - NPU 4

Intel - NPU 3



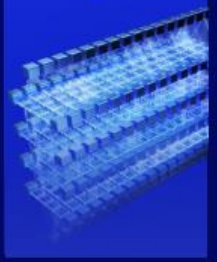
See backup for details.

Next Gen Xe2 GPU

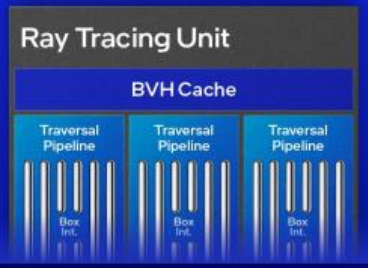
Major leap in graphics performance

up to 67 TOPS

New XMX engines



8 Larger ray tracing units



8 2nd gen Xe cores

Xe2 vector engines

1.5x better vs. Meteor Lake GPU

intel. ARC™ Software stack

eDP 1.5

Enhanced XeSS kernels

8 MB L2 cache

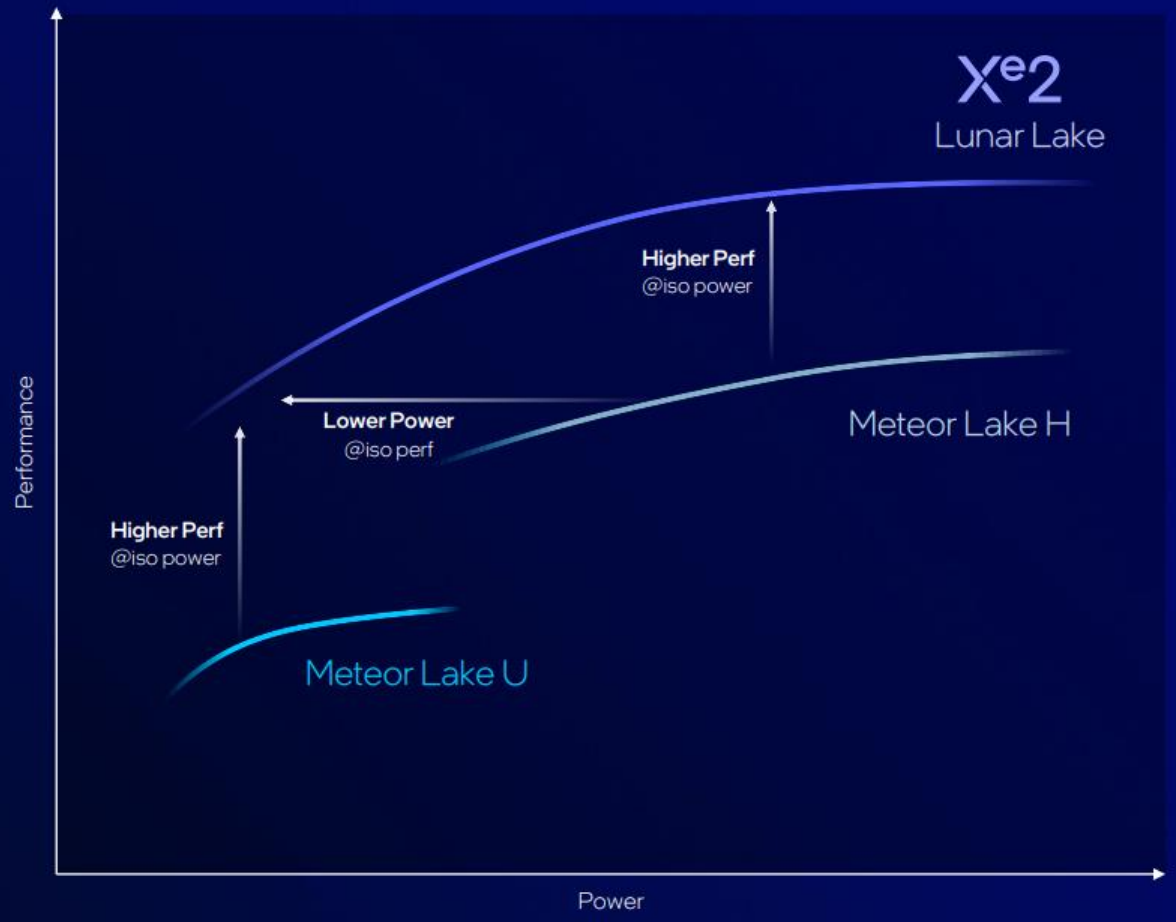
intel.

Next Gen Xe2 GPU

Major leap in graphics performance

~1.5x

vs. previous gen



See backup for workloads and configurations. Results may vary.

intel CORE ULTRA

AI PC Momentum

>8M

AI PCs Shipped to Date

40Mu

Shipped by End of This Year

Intel Core Ultra

200S Series Processors

Sales and
Pre-orders Start

Oct 24

intel.

Under embargo until October 10, 2024, at 8:00 AM Pacific

Intel Core Ultra 200S Series

The complete enthusiast solution

Enthusiast Gaming

Flagship Gaming FPS

Parity vs. Intel® Core™ i9-14900K and AMD Ryzen™ 9 9950X, geomean 31 games

+

Fastest Multithread

up to **+13%**

Faster CPU compute perf

vs. AMD Ryzen™ 9 9950X, average of four nT workloads

Cooler and Quieter

up to **17°C**

Lower CPU package temperatures

vs. Intel® Core™ i9-14900K during active PC gaming

Ultra Efficient Gaming

up to **165W**

Lower system power

vs. Intel® Core™ i9-14900K during active PC gaming

Expanding AI PC to Desktop

36

Platform TOPS

Across the platform with VNNI, DP4a, and NPU acceleration

intel. As of October 2024, among desktop processors targeting ~125W TDP. Results may vary based on use, configurations, and other factors. See intel.com/performanceindex for details.

Under embargo until October 10, 2024, at 8:00 AM Pacific

Bringing AI everywhere

Intel AI for the enterprise & edge

AI PC
Node
Light Inference

AI PC

Broadest AI SW Ecosystem

Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inf.

ENTERPRISE & EDGE

Open Standard, "Ready to Use"

Super Cluster
Training, Tuning, Peak Inf.

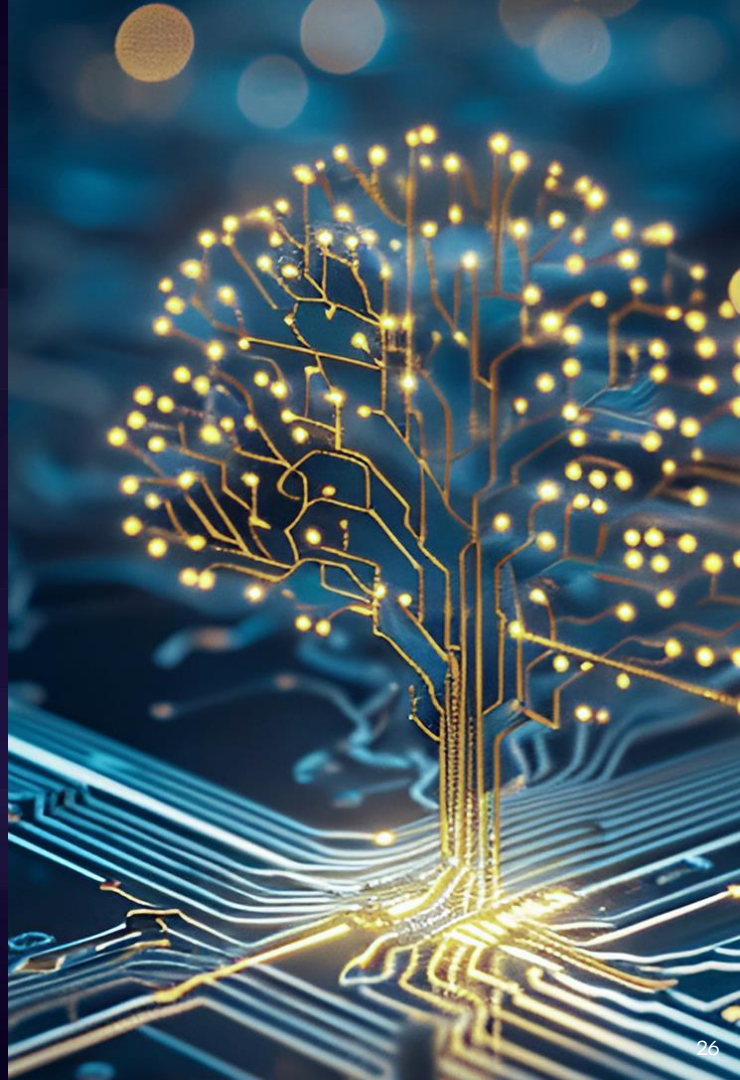
Mega Cluster
Large Scale Training & Inference

DATA CENTER

AI Open, Scalable Systems & Reference Arch

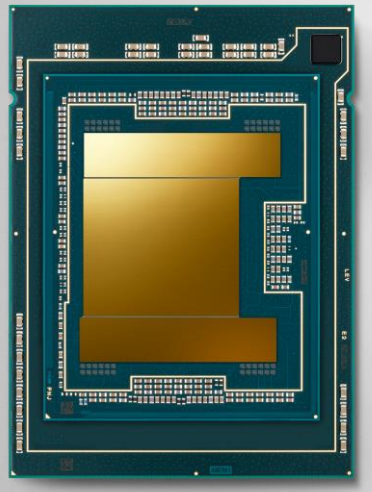
Intel® Xeon® 6 Processor

Performance & Efficiency



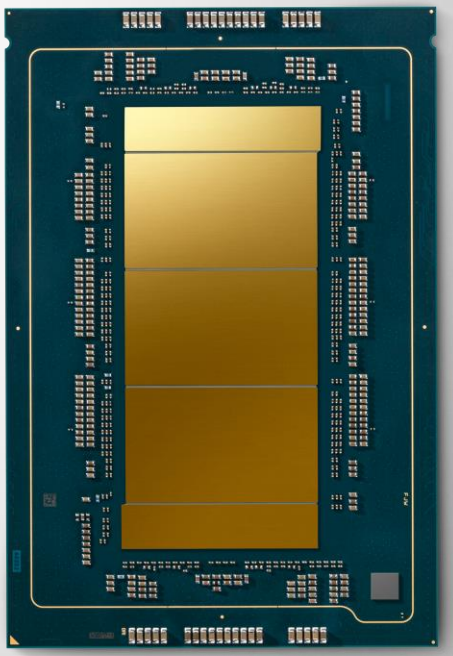
Intel® Xeon® 6 Processors | 6700 and 6900 Platform Series

Shared underlying platform delivering new levels of hardware optimization



Intel® Xeon® 6700-series

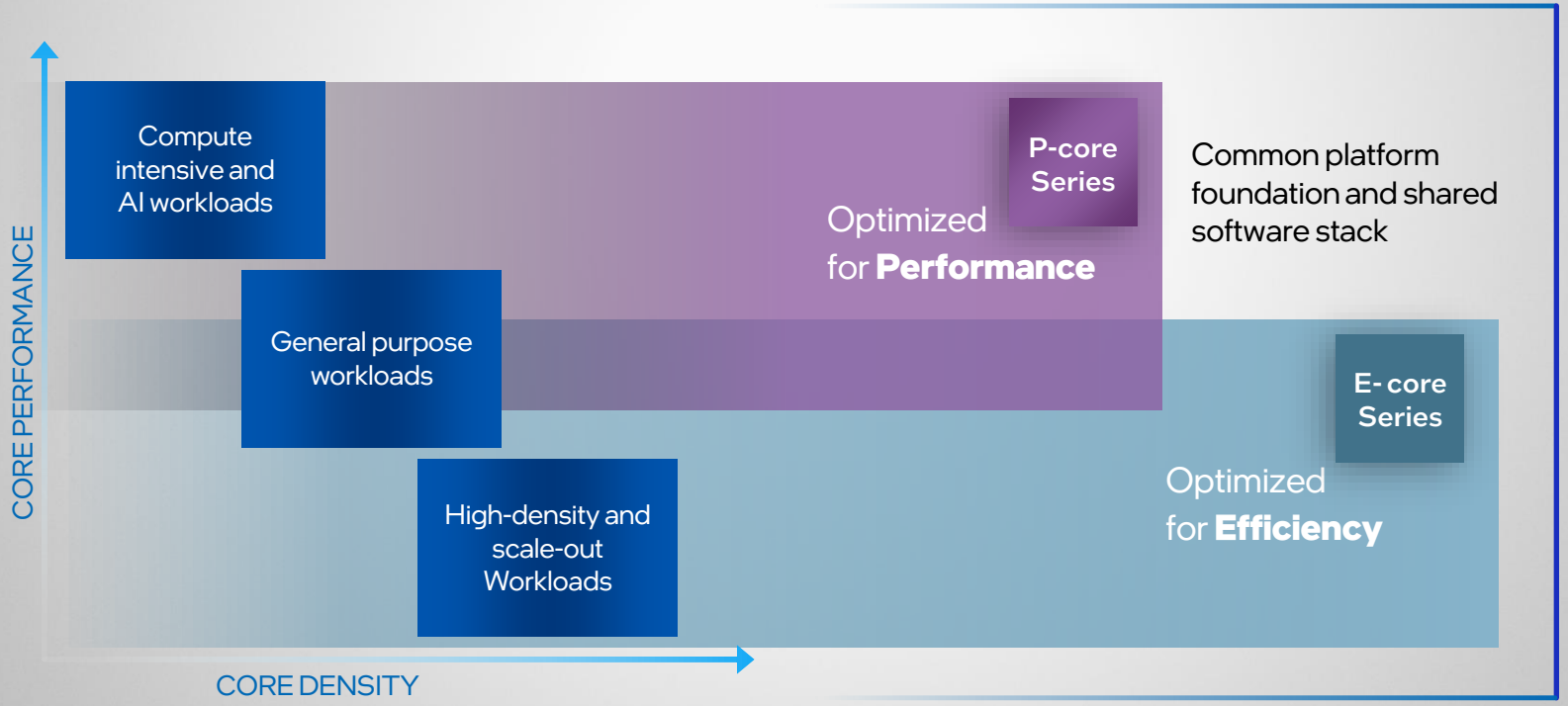
- P-core and E-core SKU selections
- Increased core counts
- Increased memory bandwidth with DDR5
- Multiplexed Rank DIMM (MRDIMM)
- Increased inter-socket bandwidth with UPI 2.0
- Compute Express Link® (CXL®) 2.0
- Increased I/O bandwidth on PCIe 5.0
- Increased shared LLC
- Intel® Accelerator Engines
- Increased Intel® VMD domains
- HW-enhanced security
- Common OS and firmware



Intel® Xeon® 6900-series

Intel® Xeon® 6 Designed to Address Market Needs

The best processors to meet diverse performance and efficiency requirements



Intel® Xeon® 6 Processors | 6700 and 6900 Platform Offerings

Scalability and flexibility across a wide range of optimized products

		Socket Support	Max TDP	Mem Channels	PCIe/ CXL	UPI Links
6700 Series	Up to **144** Efficient-cores Up to **86** Performance-cores	**1S/2S** and **4S/8S** (P-core only) support	Up to **350W** per CPU	**8** channel memory Up to **6400 MT/s** DDR5 memory **8000 MT/s** MCR DIMM memory (P-core)	Up to **88** lanes PCIe 5.0 /CXL® 2.0 **(up to 136 lanes for 1S designs)**	**4** UPI 2.0 links, up to **24 GT/s**
6900 Series	Up to **288** Efficient-cores Up to **128** Performance-cores	**1S/2S** support	Up to **500W** per CPU	**12** channel memory Up to **6400 MT/s** DDR5 memory **8800 MT/s** MCR DIMM memory (P-core)	Up to **96** lanes PCIe 5.0 /CXL® 2.0	**6** UPI 2.0 links, up to **24 GT/s**

Intel® Software Guard Extensions (Intel® SGX)

Intel® Trust Domain Extensions (Intel® TDX)

Intel® Accelerator Engines

Increased shared LLC

Common OS and firmware

Simplify Development and Deployment

Building efficiency and ease of use with a common software stack

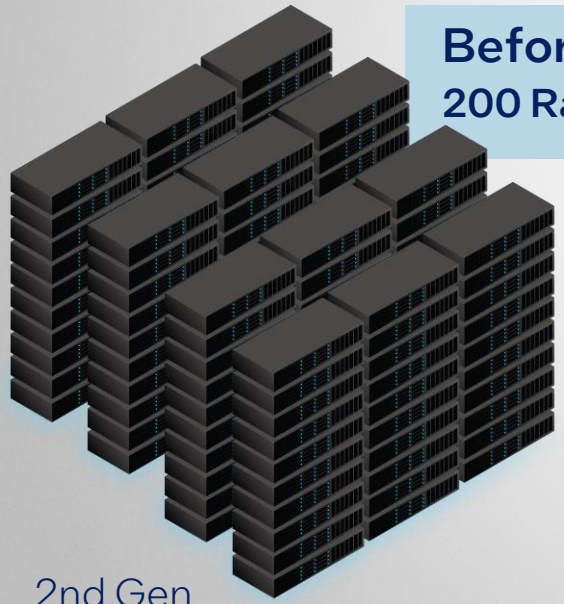
Category	Software Stack Component	Efficient-core (E-core)	Performance-core (P-core)
Instruction Set and Extensions	Base x86 ISA	x	x
	Intel® Advanced Vector Extensions 2 (Intel® AVX2)	x	x
	Intel® Advanced Vector Extensions 512 (Intel® AVX-512)		x
	Intel® Advanced Matrix Extensions (Intel® AMX)		x
OS and Hypervisor	Linux kernel and commercial Linux	x	x
	Windows	x	x
	VMware ESXi	x	x
Applications and Libraries	Database incl. common libraries (ex. ZStd)	x	x
	Network & media incl. common libraries (ex. DPDK)	x	x
	General compute & storage incl. libraries (ex. SPDK)	x	x

Addressing Unique Workload Requirements

P-core			Workloads	E-core		
	Modeling & Simulation	CAE	**HPC**			
			Web & Microservices	Cloud-Native	Consumer Digital Services	Application DevOps
CRM, ERP	Big Data	In-Memory Analytics	**Database & Analytics**	Unstructured Databases	Scale-Out Analytics	
Generative AI	Deep Learning Machine Learning	Inference	**AI**			
HCI	Virtualization	Storage	**Infrastructure & Storage**	Storage		
	CDN	Media & Gaming	**Networking**	Network Microservices	Cloud-Native CDN	5G Core
	Video	Edge Analytics	**Edge**	Virtual Protection Relay		

Today's AI Data Center

Process the same media streams/second in less space and at lower power to enable new AI projects



Before
200 Racks

2nd Gen
Intel® Xeon® Processor

3:1

Rack Consolidation[1]

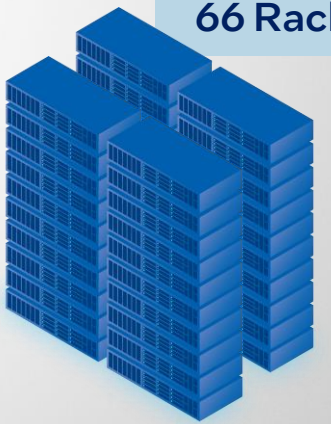
Over 4 years

80k MWh[1]

Fleet energy saved

34k mt[1]

Reduced CO2 Emissions



Now
66 Racks

Intel® Xeon® 6700E

See [7T2] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary

Intel® Xeon® 6 processor compared to 5th Gen Intel® Xeon® Scalable processor

Artificial intelligence	SQL performance	HPC	Microservices	Analytics	Media
Up to **3x** better Llama2 performance with Intel® AMX[4]	Up to **2x** better HammerDB MySQL performance[5]	Up to **2.5x** better performance for HPCG benchmark with MRDIMM[6]	Up to **1.5x** better performance/watt for server-side Java throughput[7]	Up to **1.6x** better performance/watt for MySQL OLTP[8]	Up to **1.5x** higher AVC performance/watt[9]
P-core			E-core		

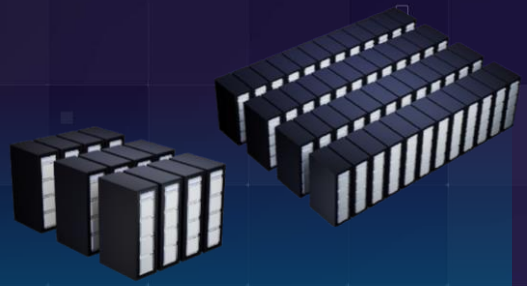
Bringing AI everywhere

Intel AI across the data center

AI PC
Node
Light Inference

Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inf.



Super Cluster
Training, Tuning, Peak Inf.

Mega Cluster
Large Scale Training & Inference

AI PC

Broadest AI SW Ecosystem

ENTERPRISE & EDGE

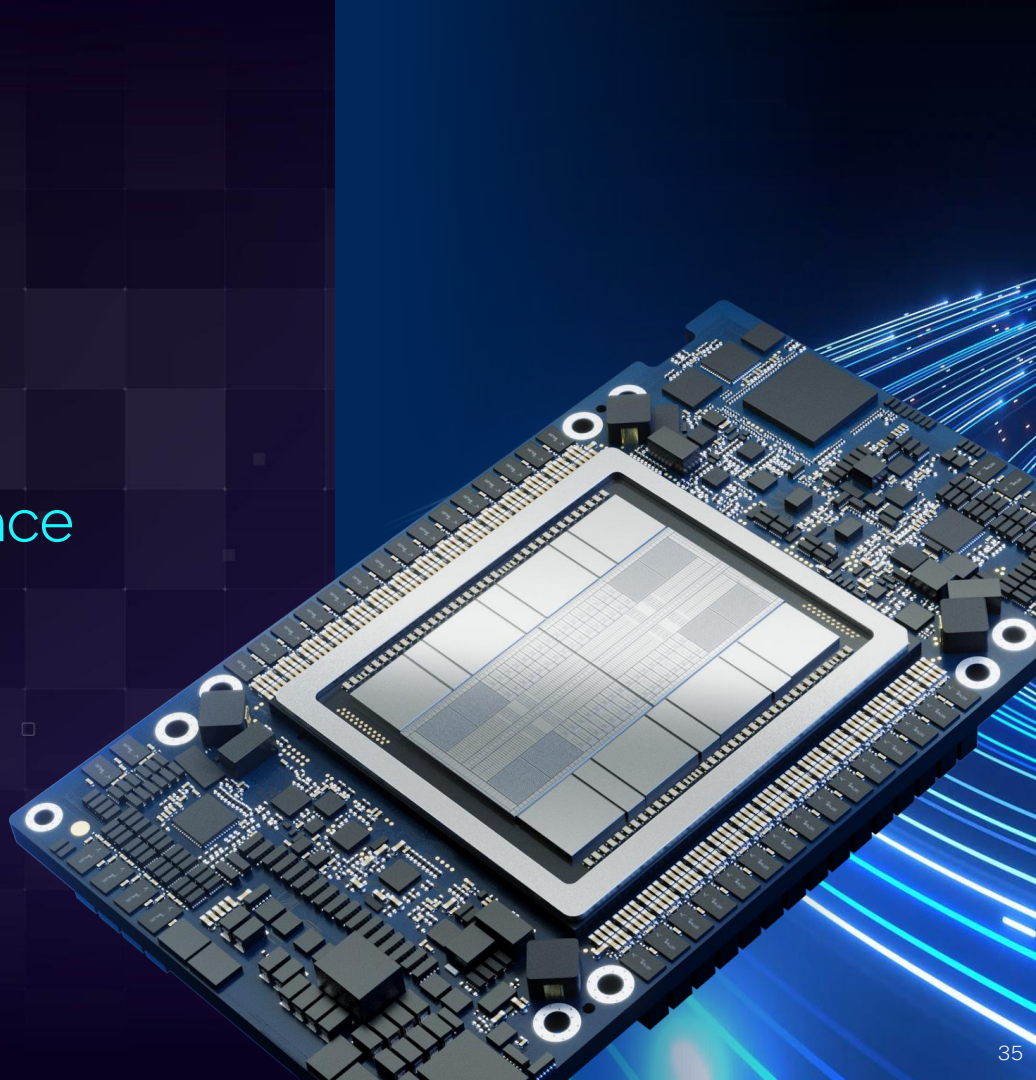
Open Standard, "Ready to Use"

DATA CENTER

AI Open, Scalable Systems & Reference Arch

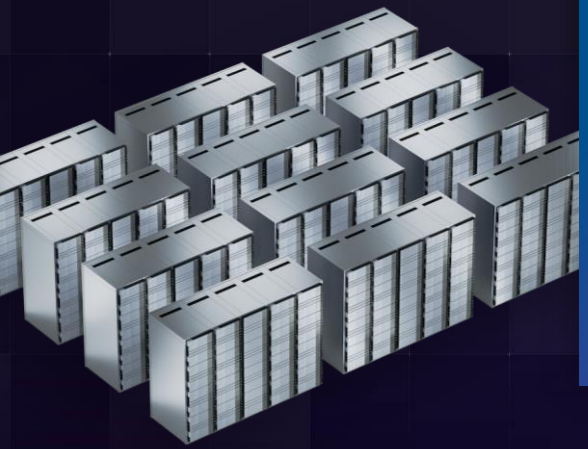
Intel® Gaudi® AI Accelerator

Giant leap in performance and productivity for Generative AI



Training Performance at Scale

Intel® Gaudi® 3 AI accelerator



40% faster[1]

Time-to-train vs. H100

8192 Accelerator Cluster

GPT3-175B

TTT projection

15% faster[2]

Training throughput vs. H100

64 Accelerator Cluster

LLAMA2-70B

TTT projection

[1]Source for Nvidia H100 GPT 3 performance https://mlcommons.org/benchmarks/training/, v3.1, closed division round. Accessed Apr 30th, 2024
Intel Gaudi 3 measurements and projections by Habana Labs, Apr 2024; Results may vary

[2]Sourcefor Nvidia H100 LLAMA2-70B performance https://developer.nvidia.com/deep-learning-performance-training-inference/training , Apr 29th 2024 → "Large Language Model" tab.
Intel Gaudi 3 measurements and projections by Habana Labs, Apr 2024; Results may vary

intel GAUDI

Scalable Systems

Reference Architectures*

1 Node
(x8 Accelerators)

FP8 Compute	14.7 PF
Memory Capacity	1024 GB
Networking B/W	9.6 TB/s

64-node Cluster
(x512 Accelerators)

FP8 Compute	940 PF
Memory Capacity	65.5 TB
Networking B/W	614 TB/s

512-node Cluster
(x4096 Accelerators)

FP8 Compute	7.52 EF
Memory Capacity	525.3 TB
Networking B/W	4.915 PB/s

1024-node Cluster
(x8192 Accelerators)

FP8 Compute	15 EF
Memory Capacity	1 PB
Networking B/W	9.830 PB/s

Performance

Inference - Lower latency , higher throughput
Training - Faster Time to train, larger model sizes

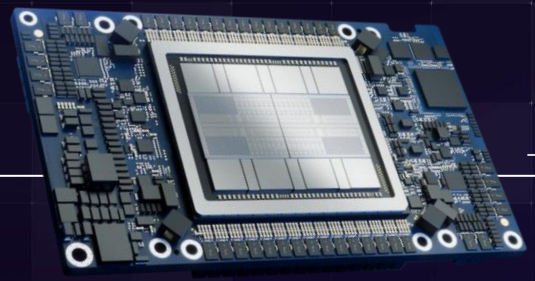
*Visuals for illustrative purposes, not actual systems.

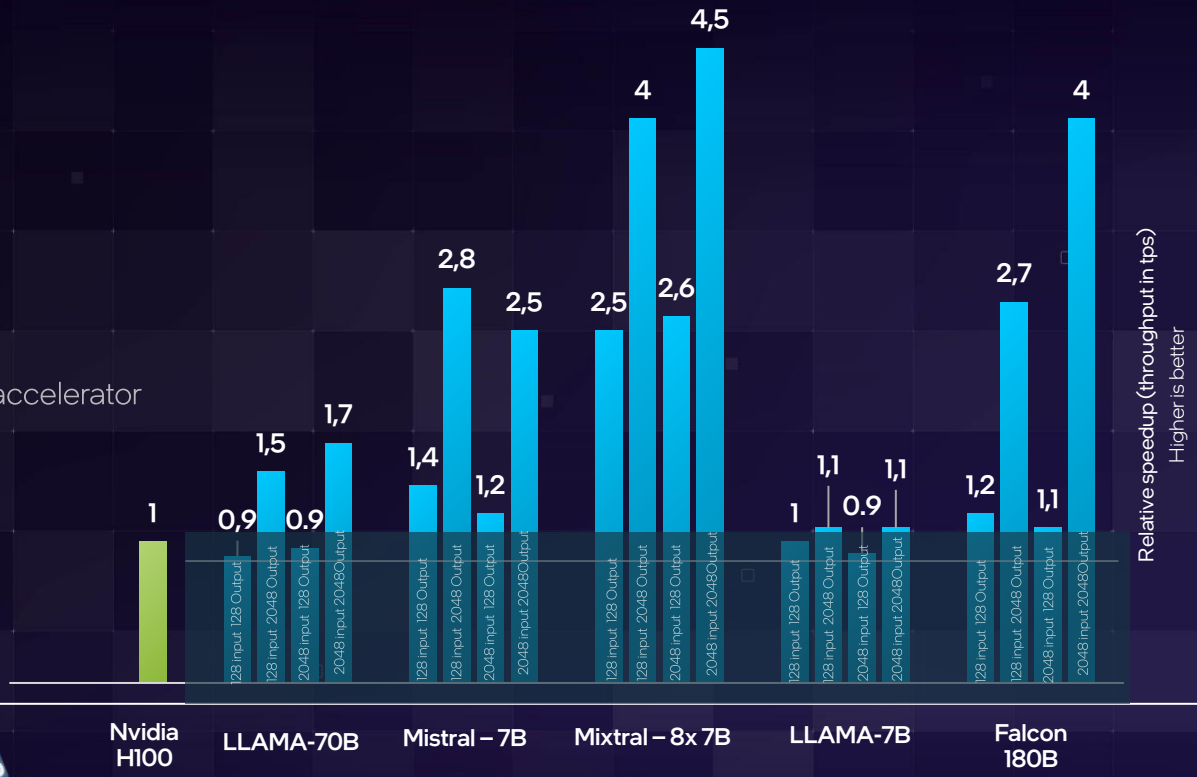
Peak projected performance, memory capacity & B/W, networking scale-up/scale-out B/W Performance varies by use, configuration and other factors. Results may vary

intel GAUDI

2x faster inferencing

Average projection for Intel® Gaudi® 3 accelerator vs. Nvidia H100, running common Large Language Models*

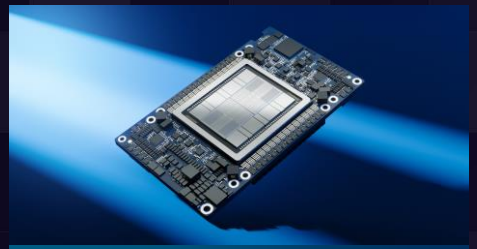




source for Nvidia performance: Overview — tensorrt_llm documentation (nvidia.github.io), May, 2024. Reported numbers are per GPU.
Intel Gaudi 3 projections by Habana Labs, Apr 2024; Results may vary

Bringing AI everywhere

Intel® Gaudi® 3 AI Accelerator



Mezzanine Card

HL-325 OAM-Complaint

Universal Baseboard

HLB-325

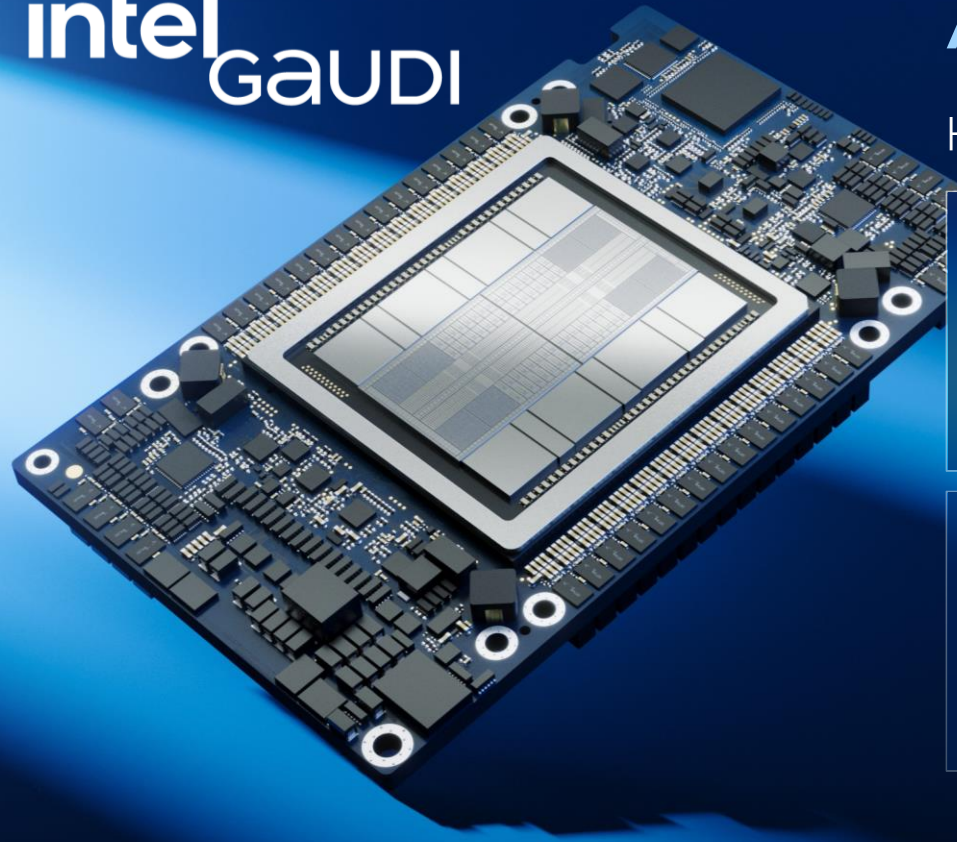
PCIe

HL-338 Add-in Card

intel GAUDI

Accelerator Card

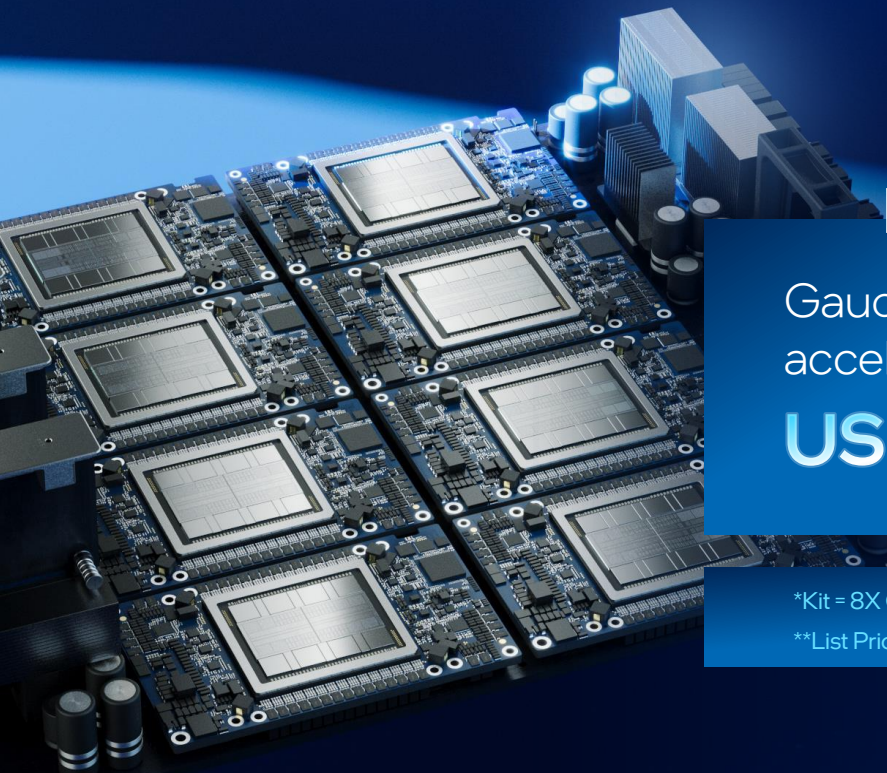
HL-325L (OAM-Compliant)



183 TFLOPS FP8	**128GB** HBM2e	**3.7TB/s** HBM Bandwidth
8 Matrix Multiplication Engines	**24** 200 GbE RDMA NICs	**1.2TB/s** bi-directional networking

intel GAUDI

Making Gen AI more accessible



Addressing Cost Barriers

Gaudi 3 AI accelerator kit*

USD 125K**

Gaudi 2 AI accelerator kit*

USD 65K**

*Kit = 8X Gaudi AI accelerators + Universal Baseboard (UBB)

**List Price (for reference only)

Pricing guidance for cards and systems is for modeling purposes only. Please consult your original equipment manufacturer (OEM) of choice for final pricing.
Results may vary based upon volumes and lead times.

intel® tiber™ Developer Cloud

Accelerate AI development using Intel-optimized software on the latest Intel® Xeon® processors, Intel® Gaudi® accelerators and Intel® Data Center GPUs.

cloud.intel.com

Get started with Intel

Get hands-on experience with the latest Intel® technologies. Empower your AI skills with Intel.

Early technology access

Evaluate pre-release Intel platforms and Intel-optimized software stacks.

Deploy AI at scale

Speed up AI deployments with the latest tools and libraries on Intel® Developer Cloud.

Advancing patient care with AI in Intel® Core™ Ultra processors

CPU-powered ultrasound imaging applications delivers more accessible and cost-effective imaging technology.

Situation

Samsung Medison is a pioneer in healthcare innovation. Their ultrasound imaging applications use AI for the most effective patient care.

Challenge

Previously, their applications were run on previous generation Intel Core processors accelerated by a competitor discrete GPU.

Solution

Samsung tested new Intel Core Ultra processors with built-in GPU engines. They saw significant AI performance improvements when compared to their previous gen CPU + dGPU combo. With Intel Core Ultra, Samsung Medison can offer advanced AI features in their next-gen ultrasound devices based solely on the CPU.

SAMSUNG MEDISON

Get the details:

Learn more

intel CORE ULTRA

Deploying high-performance and cost-efficient AI at scale

The value and performance acceleration that the combination of Intel® Xeon® processors and Intel® Software brings to the entire AI lifecycle

Situation

The Netflix performance engineering team deploys AI to improve subscriber experience, from generating better recommendations to optimizing video delivery.

Challenge

Supporting the wide variety of devices and network conditions requires encoding multiple bitstreams for every title, and every subscriber is presented with a personalized home page and recommendations. These large-scale AI deployments must be performant yet cost-efficient.

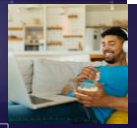
Solution

Netflix has realized large savings in cloud infrastructure costs by using Intel-optimized software, such as the Intel® oneAPI Deep Neural Network Library (oneDNN), XGBoost, and Intel® vTune™ Profiler, to get the most performance out of Intel® Xeon® processors without having to offload to more expensive GPUs.

NETFLIX

Case study

Learn more



intel®

Thank you

intel®