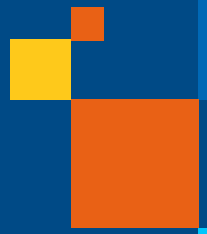
Artificial Intelligence (AI)

Partner Enablement Package

Addressing customers' business challenges with Intel based solutions



intel®

Contents

Bringing AI Everywhere

Why Partner With **Intel**?

At Intel, our goal is to improve lives and outcomes for everyone and every enterprise on this planet

But we aren't doing this alone!

Together with our partners, we are creating real value for our customers by **bringing AI everywhere** and minimizing the risks in AI solution deployment

When you partner with Intel, you partner with a complete AI ecosystem

Our broad portfolio of AI-enabling technologies and collaboration with hardware, software, and solution ecosystem partners delivers real world solutions and differentiated business outcomes for industries, companies, and communities.

Helping you to grow your business.

Join Us On the Journey to Bring AI Everywhere

Enterprise AI Value Proposition

Transforming your business with Enterprise AI

In today's hypercompetitive environment, **enterprises that embrace AI are pulling ahead.**

Businesses across industries are reimaging every aspect of operations to understand how AI can augment or even automate workflows.

At Intel, embedding AI into the fabric of the enterprise is our unique expertise.

From AI PCs that transform productivity, to years of expertise in understanding which use cases return the most value, Intel is your trusted partner to bring AI everywhere, securely and responsibly.

It's time to think differently about your Enterprise AI.

This Enablement Package will help you understand how businesses across markets can gain significant value from partnering with Intel in AI for long-term success

Intel AI Industry Impact

CLUB 386

"Understanding the once-in-a-lifetime business opportunity that runs into a total addressable market (TAM) counted in the tens of billions of dollars, Intel has been busy building the infrastructures required for pervasive AI, across all industries and business segments"

Windows Central

"Buckle up, if the industry is to be believed, 2024 is the year of the AI PC, and it all starts with Intel."

EXP review Extreme Performance

"We are beginning to sense that Intel has progressively created its own advantages in AI after it concurrently released performance enhanced AI PC and new data center CPU."

tom's HARDWARE

"Critically, Intel's new chips have also arrived on schedule, a much-needed confirmation that the company's turnaround remains on track."

omicrono

"Intel knows that AI will be everywhere by 2024 and wants its processors to be the basis for all the software technology that will flood the Internet and computer operating systems such as Windows. With this, you will be able to re-edit your favorite songs with just a few clicks or model the photos of a trip easily and quickly from your computer. Intel Core Ultra will turn every person into an individual artist, writer and musician."

siliconANGLE

"Besides throwing out impressive numbers and claims, Intel offered some concrete, real-world examples of the kinds of AI workloads its new silicon will enable. For example, restaurants will be able to guide diners' menu choices, based on their individual budget and dietary needs, while manufacturers will be able to build new systems that catch quality and safety issues on the factory floor. Advanced AI powered by Intel's silicon will also lead to the creation of more effective ultrasound systems that can catch problems that a human doctor might miss."

Industry Momentum

11-15-23 | 12:21 PM | AI DECODED

How Intel is quietly gearing up to become a player in the AI arms race

The 55-year-old company that powered the personal computing revolution has developed a compelling alternative to Nvidia's scarce and expensive GPUs.

Lenovo WaitTime

Lenovo and Intel AI solution for queue and crowd analytics

WaitTime's patented AI powers an intuitive guest interface and operator's platform to observe, measure, and maximize delivery for customer spaces using both real-time data and historical analysis.

Taboola

"[For inferencing] you need a processor that is agile, that is flexible and that can do multiple things. So it's not the core count, it's what the cores can do that counts."

Ariel Pisetzky, VP of IT & Cyber, Taboola

ThirdAI

"Our solution aims to bring an entire ecosystem of AI products which are completely built on CPU. We specifically use a lot of the high-performance compute CPU servers like the latest generation Xeon servers. And we have trained a lot of our models on those. Specifically, we've trained large language models similar to what you hear from like the GPT models of the world."

Vinod Iyengar, VP of product and go to market at ThirdAI

aws

Amazon's Elastic Compute Cloud (EC2) Custom Instances Powered by 4th Gen Intel® Xeon Processors

The new M7i-flex, M7i, C7i, R7iz, and R7i instances powered by 4th Gen Intel Xeon processors and Intel® Accelerator Engines, help accelerate AI with a pervasive platform, choice, and open access, from ML/DL to Gen AI, to drive better predictive business decisions.

DELL Technologies

"Combining the reliability of Dell PowerEdge servers – the industry's top selling portfolio – with Intel technologies for general purpose and accelerated compute provides powerful system for optimized AI. PowerEdge systems with Xeon and Gaudi will support AI workloads ranging from large scale training to base level inferencing."

Jeff Clark, Vice Chairman and COO
Dell Technologies

Hewlett Packard Enterprise

HPE expands partnership with VMware and Intel to accelerate AI for all organizations

activeloop

"Implementing your solution very effectively on CPUs not only makes the solution much more efficient, cheaper, cost effective, etc., but also it makes it very easy for customers to deploy."

Davit Buniatyan, CEO Activeloop

Google Cloud

Generally Available: C3 VMs with 4th Gen Intel Xeon and industry-leading price-performance

By adopting C3 VMs with Intel Sapphire Rapids and the new AMX instruction set for AI, we are seeing 2x performance for some of our inline models, compared to the previous generation N2…"

Suiqang Deng, Distinguished Engineer
Palo Alto Networks

Microsoft Azure

Microsoft Azure Adds Confidential VMs to Expand Options for Confidential Computing

Powered by 4th Gen Intel® Xeon Processors featuring Intel® Trust Domain Extensions, the new confidential VMs enable organizations to bring confidential workloads to the cloud, ensuring that data and applications stay private and encrypted even while in use.

Mark Russinovich, Chief Technology Officer, Microsoft Azure

DataRobot

"Our video analytics runs on Intel processors, so we are heavily using the CPU for inference. The most exciting part is how can we make AI inference more accessible to the customer and easier to use for the customer."

Raj Ramanujam, SVP of Global Alliances and Channels, Datarobot

Sensormatic

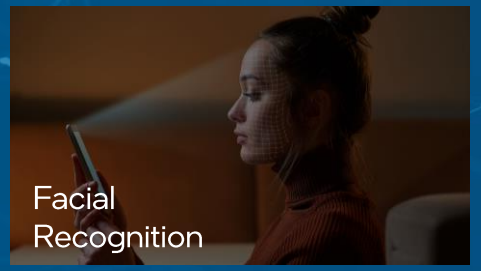
""Intel allows us that low latency to run at the edge with high performance."

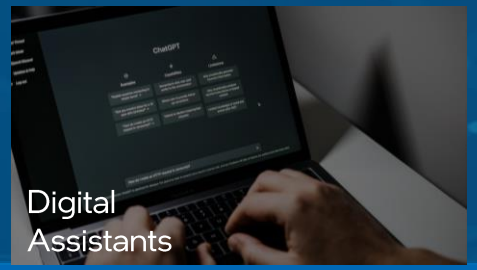
Dustin Ares, GM for video analytics, A.I. and Incubation ,Sensormatic

AI is Transforming Business Worldwide

How can businesses benefit?

Your business can leverage AI to increase profits and improve efficiency





Gartner AI Hype Cycle

The 2023 Gartner Hype Cycle™ for Artificial Intelligence (AI) identifies innovations and techniques that offer significant, and even transformational benefits while also addressing the limitations and risks of fallible systems.

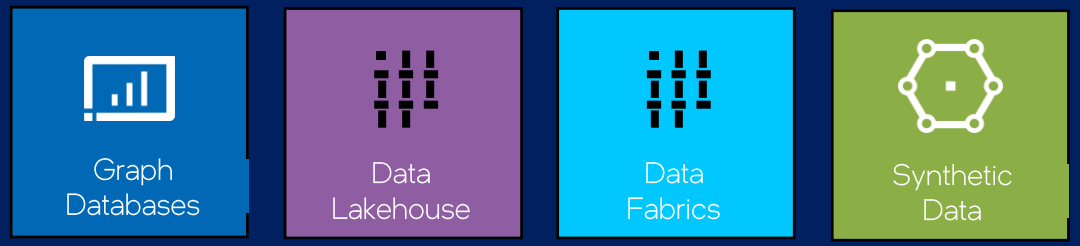
AI strategies should consider which offer the most credible cases for investment.

"Early adoption of these innovations will lead to significant competitive advantage and ease the problems associated with utilizing AI models within business processes." *Gartner Director Analyst, Afraz Jaffri*

What's New in Artificial Intelligence from the 2023 Gartner Hype Cycle

AI is Evolving Rapidly

Underlying data technologies:



$300B Worldwide GenAI spending set to exceed $300B by 2026

AI everywhere
By 2026

More than **50%** of enterprise-managed data will be created & processed outside the data center or cloud

58% of CEOs from leading public companies are actively investing in AI

50% of edge deployments will involve AI

AI as disruptive as the Internet

Generative AI predicted to add up to $4.4T of value to global economy by 2040[2]

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)

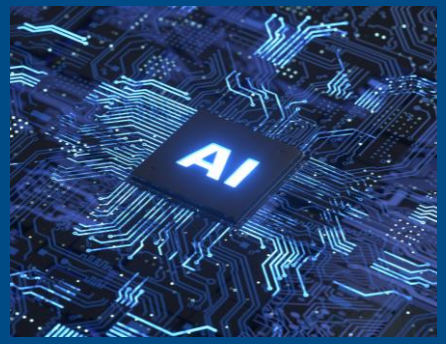
https://chiefexecutive.net/the-rise-of-the-ai-ceo/
https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/

By 2026, at least 50% of edge computing deployments will involve machine learning (ML), compared to 5% in 2022 (Building an Edge Computing Strategy, April 2023)

[1] https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights
[2] Worldwide Artificial Intelligence Spending Guide (IDC)

What Are Some of the AI Challenges Today?

Why Partner with Intel



GPU Availability

Intel CPU alternative to global GPU shortage

Naver's AI server switch comes as global information technology firms are increasingly disgruntled with Nvidia's GPU price hikes and a global shortage of its GPUs

Vendor Lock-in

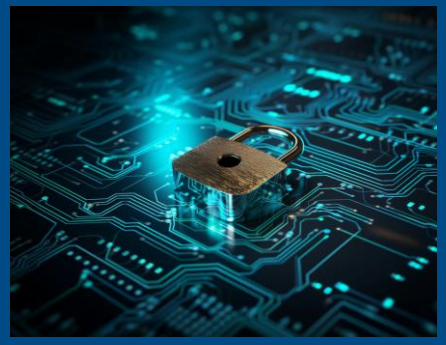
Avoid Vendor Lock-in with open-source standards-based software

Intel works with all the industry standard open frameworks and libraries to optimize for highest performance and ensure a quality out-of-the-box experience on Intel technologies

Cost

Intel is delivering better price and performance on 4th Gen Intel® Xeon®

In real work applications, Intel is disrupting the industry and democratizing AI by delivering better performance, lower price and a more balanced platform for AI inference



Secure AI

Intel Offers the Most Comprehensive Security Portfolio

Intel security capabilities let you set the trust boundary appropriate to your workloads, helping protect sensitive data, content, and software IP from advanced attacks, tampering, and theft

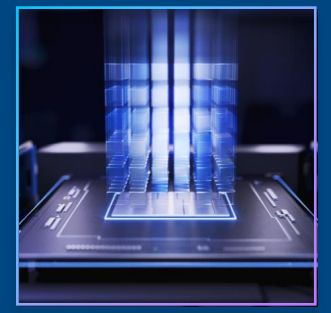
How Intel is Bringing AI Everywhere

World-changing technology that improves the life of every person on the planet

Intel's Unique Value

- Open approach
- Expertise across hardware & software
- Ecosystem
- Execution

Intel's broad portfolio of AI-enabling technologies, unique vision of future AI-enhancing innovations, and unrivaled support for an open ecosystem are helping **bring AI everywhere** that benefits everyone



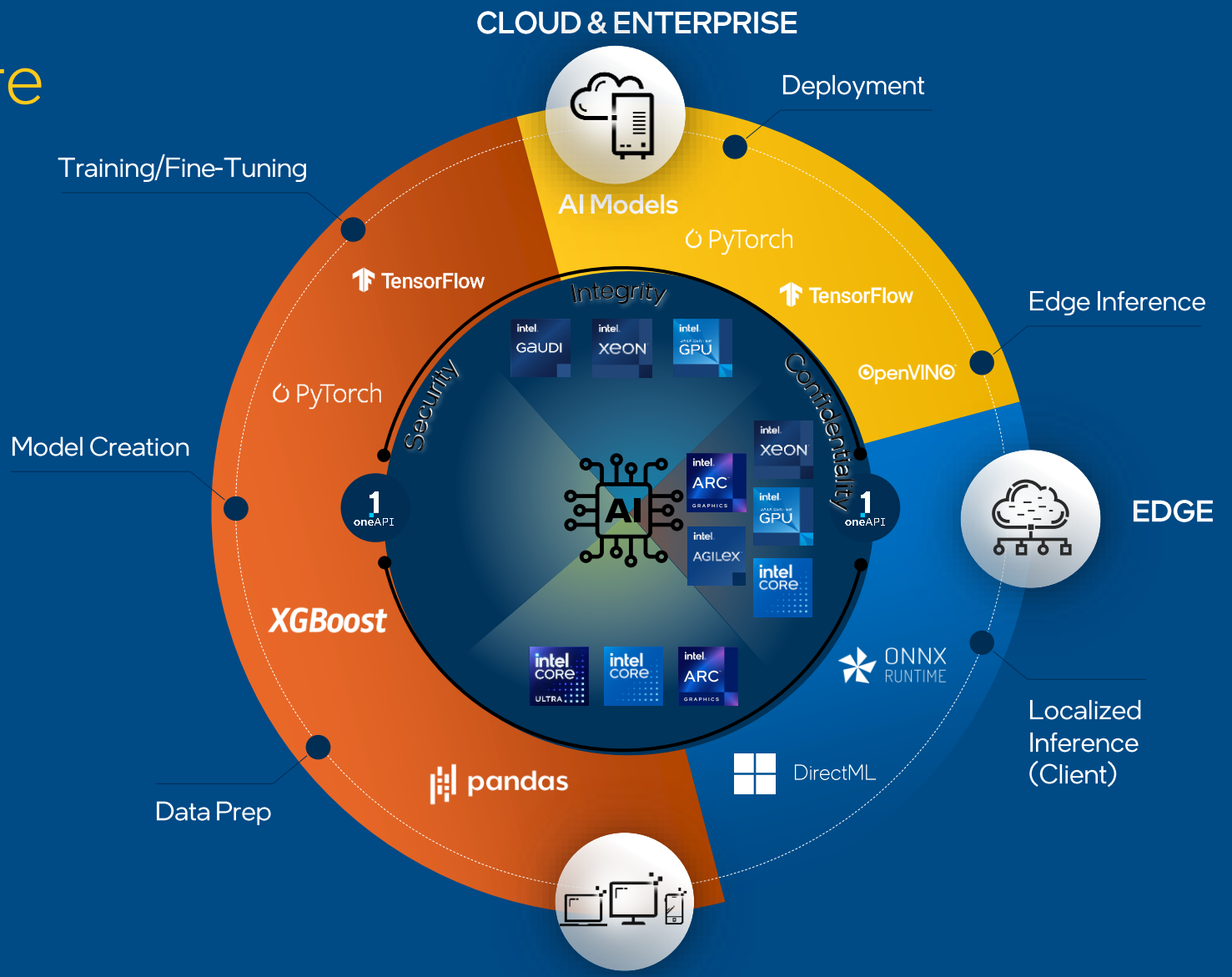
- Scaling AI across the full spectrum of workloads, making it accessible to individuals and organizations
- Heterogenous architectures, open standards, and solutions that let customers confidently secure diverse AI workloads across the data center, cloud, on PCs and at the edge

AI Continuum

Enables AI Everywhere

Intel is your trusted partner to **bring AI everywhere** and support your business through every step of the AI journey

From the data center, cloud and network, to the client and edge



Note: Intel® Core Ultra integrates NPU low power inference engine from Meteor Lake onwards.

Responsible AI with Intel

Enhancing Accessibility

For many individuals with disabilities, independence and autonomy can be a challenge. AI is helping to change that by creating products that offer alternative solutions to everyday barriers

Creating Environmental Solutions

Using AI technology, researchers can better understand how our environment works and develop solutions to build a better future

Expanding Access to Education

Intel is dedicated to responding to the global AI skill gap with programs like AI for Youth and AI for Future Workforce, preparing students for the digital revolution

Advancing Healthcare

AI is now commonly used in healthcare and life sciences, from improving patient care to developing preventive disease research

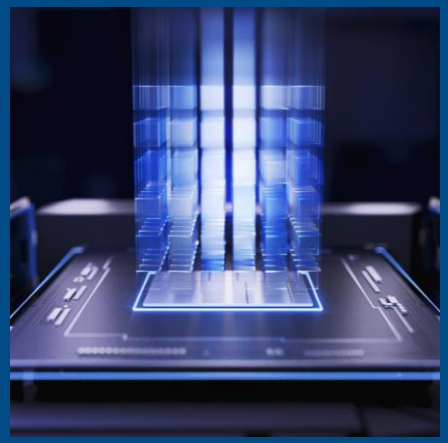
Improving Safety

From enabling automated vehicles to drive successfully to reducing child exploitation, AI technology is helping make society safer

How Intel is Powering AI at Every Stage

Maximize Value

Do Enterprise AI right: Realize your vision at a lower cost



Accelerate Innovation

Get to market fast by avoiding the "AI failure" trap

From AI PCs that transform productivity, to years of expertise in understanding which use cases return the most value, Intel is your trusted partner to bring AI everywhere, securely and responsibly.

It's time to think differently about your Enterprise AI

Maximize Value

Realize your vision without over-investing

REAL WORLD RESULTS

Healthcare

BeeKeeperAI is helping **improve disease prediction, diagnosis, and treatment** by securely collaborating on private data, validating algorithms in **half the time, at half the cost**, using Intel confidential computing[1]

BeeKeeperAI®

READ THE WHITE PAPER

Samsung Medison achieved **more accessible and cost-effective AI medical imaging** by speeding up throughput for real-time ultrasound imaging by **22%** and **25%** with Intel® Core™ Ultra processors, compared to previous generation processors with a discrete GPU[2]

SAMSUNG MEDISON

Media & Entertainment

Gunpowder accelerated rendering times for **stunning visual effects** while lowering costs with as much as **52% better performance per dollar** compared to previous-gen instances with Intel® Xeon® processors[3]

GUNPOWDER®

READ THE CASE STUDY

Netflix delivered **fast and seamless streaming experiences** with **2x better AI-enabled video encoding** and significant cloud savings by upgrading AWS EC2 instances. Netflix achieved a **3.5x performance improvement per CPU** with Intel® Xeon® CPUs and software optimizations, at a lower cost than with GPUs[4]

NETFLIX

READ THE BLOG

Professional Services

Ropers Majeski increased worker **productivity** by **18.5%**, saving an average of **75 minutes** per user per day by automating email processing, document filing, and report generation with built-in AI acceleration from Intel® Xeon® CPUs[5]

ROPERS MAJESKI

READ THE CASE STUDY

Retail

Meituan uses vision AI services to **improve a wide range of customer experiences**, and achieved **70% cost savings** by migrating from GPUs to Intel® Xeon® CPUs and software for AI inference[6]

美团 Meituan

READ THE CASE STUDY

[1, 3, 4, 5, 6] See respective papers and blogs (linked above) for configuration details. Results may vary.

[2]Based on internal test results conducted by Samsung Medison. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Results may vary. For details, see: intel.com/content/www/us/en/products/details/embedded-processors/core-ultra.html

Accelerate Innovation

Get to market fast—avoid the "AI failure" trap

REAL WORLD RESULTS

Healthcare

Siemens Healthineers is helping practitioners provide **faster, more efficient treatments to patients**, accelerating medical imaging inference time by **35x** while meeting sustainability objectives using Intel® Xeon® CPUs and OpenVINO™ software[1]

SIEMENS Healthineers

READ THE CASE STUDY

Hospitality

Hellometer's AI timer, optimized with OpenVINO™ software, enables restaurant operators to **improve drive-thru service speed and increase sales** by **47 seconds** on average, translating into about **$130k** in added revenue per location[3]

Hellometer

READ MORE

Manufacturing

The **Siemens SIMATIC Industrial PC** is empowering **better product quality and yield in the automated factory**, delivering up to **4.25x** higher inference performance for object detection and up to **3x** higher inference performance for image classification compared to the previous generation using Intel® Xeon® processors for IoT and OpenVINO™ software[2]

SIEMENS

Consulting

BCG and Intel's GenAI solution helps employees **quickly find information relevant to their work and generate business insights**. The solution, powered by Intel® Xeon® CPUs, Intel® Gaudi® accelerators, and hybrid cloud-scale software, delivered a **25%** growth in result relevancy and a **39%** increase in improved work completion rates[4]

BCG

READ MORE

[1, 3, 4] See respective papers and blogs (linked above) for configuration details. Results may vary.

[2]Based on internal test results conducted by Siemens. See N21 and N22 at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Why Intel's Open AI Approach is suited to your AI business needs

Avoid Vendor Lock-in

open-source standards-based software

Leverage Intel's Hardware portfolio

optimized for AI use cases

Create new opportunities from the client and edge to the data center and cloud with **hardware optimized by software and open standards** for tomorrow's AI

Intel's AI Strategy

What Intel brings to accelerate AI innovation

AI Applications & Software

Open | Productive | Accessible | New Algos

Driving performance at scale

Intel® Developer Cloud | **Hybrid AI** | OpenVINO™

with open standards and software

Data Center	Networking	Client & Edge
Scalable Systems	Open Standards	AI PC
Accelerators, Xeon	Network Infrastructure	NPU, GPU, CPU

built into and accelerating every platform

Advanced High-Performance Technologies

Open AI Systems Foundry

with advanced, responsible processes

Ethical Leadership Foundation

that keep AI data trusted and secure.

Deploy AI Everywhere: Whitepaper Video

Intel AI Portfolio

Take advantage of hardware and software optimized for all your AI compute needs

Open Software Environment

oneAPI · OpenVINO · PyTorch · XGBoost · MODIN · TensorFlow · ONNX RUNTIME · deepspeed

Deep Learning Acceleration

- Intel Gaudi — Dedicated Deep Learning Training and Inference

General Acceleration

- Intel Data Center GPU Flex Series — Cloud Gaming, VDI, Media Analytics, Real-Time Dense Video
- Intel Data Center GPU Max Series — Parallel Compute, HPC, AI for HPC

General Purpose

- Intel Xeon, Intel Xeon Max Series — Real-Time, Medium Throughput, Low Latency, and Sparse Inference
- Intel Xeon, Intel Xeon Max Series, Intel Xeon W — Medium to Small Scale Training and Fine Tuning
- Intel Core i7, Intel Xeon D, Intel Data Center GPU Flex Series — Edge and Network AI Inference
- Intel Core i7, Intel Arc Graphics — Inference on Client

Intel® AI Software Portfolio

Engineer Data → Create Models → Optimize & Deploy

MODIN · SciPy · pandas · NumPy · Spark · Numba

dmlc XGBoost · PyTorch · ONNX RUNTIME · OpenVINO · scikit learn · TensorFlow · DirectML · Write Once Deploy Anywhere · SigOpt AutoML · intel Neural Compressor

Data Analytics at Scale†

Machine & Deep Learning Frameworks, Optimization and Deployment Tools†

oneAPI

- Intel® oneAPI Deep Neural Network Library
- Intel® oneAPI Collective Communications Library
- Intel® oneAPI Math Kernel Library
- Intel® oneAPI Data Analytics Library

Open, cross-architecture programming model for CPUs, GPUs, and other accelerators

CLOUD & DATA CENTER · CLIENT & WORKSTATION · EDGE

intel AI ANALYTICS TOOLKIT

Accelerate end-to-end data science and AI

Intel® Tiber™ Developer Cloud (formerly; Intel® Developer Cloud) and Intel® Developer Catalog

Try the latest Intel tools and hardware, and access optimized AI Models

Intel Tiber AI Studio (formerly; Cnvrg.io)

Full stack ML operating system

Intel® Geti

Annotation/training/optimization platform

Hugging Face

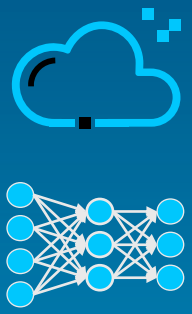
Intel optimizations and fine-tuning recipes, optimized inference models, and model serving

Note: components at each layer of the stack are optimized for targeted components at other layers based on expected AI usage models, and not every component is utilized by the solutions in the rightmost column

† This list includes popular open-source frameworks that are optimized for Intel hardware

Hybrid AI: Seamless Edge-to-Cloud Coordination

Cloud



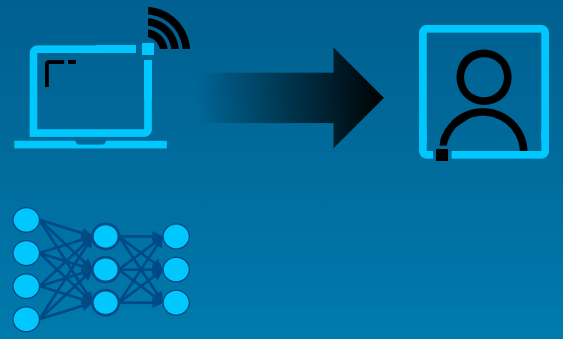
LLM

Powerful LLM in public cloud for inferencing non-confidential data

Cloud-Edge handshake with non-confidential data

Near-real-time inferencing on client and edge combines with deeper cloud-based insight, using common tools and silicon without manual coding

Client and Edge



LLM

Local LLM inferencing protected data

Personalized outcome from both LLMs

HEALTHCARE

Use Generative AI to automate creation of personalized emails to patients while protecting privacy

RETAIL

Inference video data at the Edge and gain insights from Generative AI without backhauling costs

ENTERPRISE

Use Generative AI for productivity gains without exposing confidential information to public cloud

Accelerate AI Development with Reference Kits

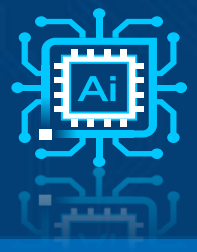
Optimized AI reference kits help developers and data scientists innovate faster

Built on the oneAPI open, standards-based, heterogeneous programming model and components of Intel's end-to-end AI software portfolio, such as Intel® AI Analytics Toolkit and the Intel® Distribution of OpenVINO™ toolkit, the reference kits enable AI developers to streamline the process of introducing AI into their applications, enhancing existing intelligent solutions and accelerating deployment.

The result is proven performance improvements with a shorter, more productive workflow versus a traditional model development workflow

Using the **AI reference kit** designed to set up interactions with an enterprise conversational AI chatbot, users can experience inferencing in batch mode **up to 45% faster** with oneAPI optimizations

The **AI reference kit** designed to automate visual quality control inspections for life sciences demonstrated training **up to 20% faster** and inferencing **55% faster** for visual defect detection with oneAPI optimizations



To enable developers to predict utility asset health and deliver higher service reliability, there is an **AI reference kit** that provides **up to a 25% increase** in prediction accuracy

Press Release

AI Reference Kits eBook

Stay Secure

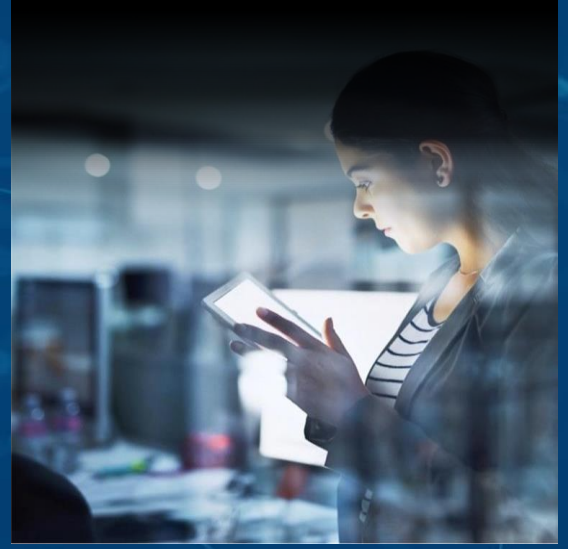
Protect your AI initiative and comply with regulations with built-in security features

Security

Protect sensitive data & models

Compliance

Comply with security and privacy regulations



Confidentiality

Engage multi-party AI without exposing private data

Intel Offers the Most Comprehensive Security Portfolio

Intel® Software Guard Extensions (Intel® SGX)

Application isolation

Intel® Trust Domain Extensions (Intel® TDX)

Virtual machine isolation

Intel® Tiber™ Trust Services

formerly; Intel® Trust Authority

Independent trust verification services for multi-cloud & hybrid cloud

Software Solutions, Cloud, OEM and System Integrator Ecosystem

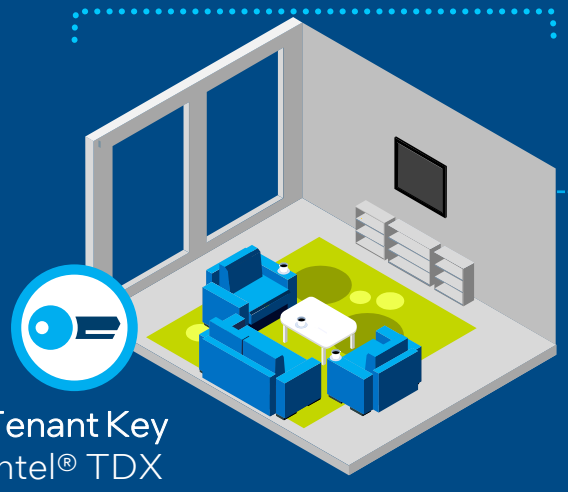
Intel Security-First Development & Lifecycle Support

*Intel® TDX available through select cloud providers

Protecting Data in Memory

Today: If you can snoop the memory, you can see everything passing through, including private keys used to decrypt data

Analogy: VM

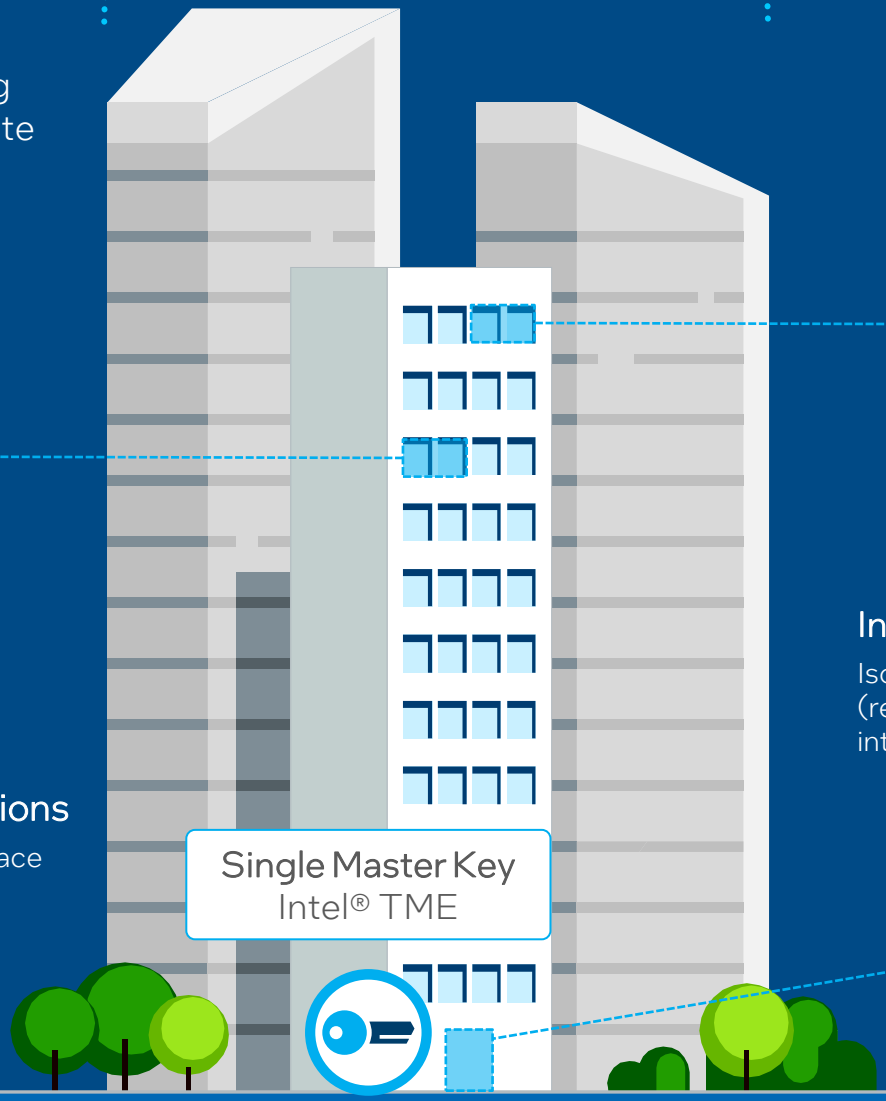


Tenant Key
Intel® TDX

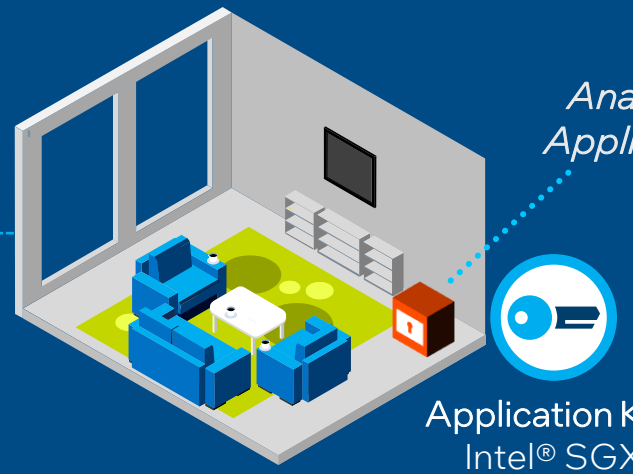
Intel® TDX = Intel® Trust Domain Extensions

Separate keys for separately encrypting each VM space (requires only OS/VMM to be feature-aware)

Analogy: System Memory



Analogy: Application



Application Key
Intel® SGX

Intel® SGX = Intel® Software Guard Extensions

Isolation for individual application data spaces (requires application code modifications or abstraction interface)

Intel® TME = Intel® Total Memory Encryption

A single key to encrypt full system memory (no OS/App mods required)

Accelerate Innovation and Enhance Data Protection with Intel® Security Engines

Confidential Computing with the Intel® Xeon® Scalable platform

Put data into action while helping to keep it private

Maintain performance while helping preserve data confidentiality and code integrity with Intel® Security Engines on Intel® Xeon® CPUs

READ THE PRODUCT BRIEF

With Intel you can get better insights for critical business outcomes

READ THE BUSINESS BRIEF

Embrace Confidential Computing with Intel® SGX and Intel® TDX

Enabling the AI Ecosystem

Drive new opportunities and key business outcomes with optimized performance using the modern software tools preferred by AI developers

Open
Programmability

PyTorch
TensorFlow
python
Hugging Face
UXL
SYCL

Choice
Compatibility

oneAPI
OpenVINO
SYCLomatic
OpenXLA
OpenAI Triton
MLIR

Trust
AI inference

Secure AI

Secure workloads

Secure models

Secure data in use

Scaled
Develop and test

Intel® Tiber™ Developer Cloud
(formerly; Intel® Developer Cloud)

cloud.intel.com

Small, medium and large models

Full system, full cluster

Latest Intel CPUs, accelerators and software

Open, accelerated, connected computing for AI

Multivendor

Multiarchitecture

Hardware / Architecture

intel Xeon
intel Gaudi
intel Data Center GPU Flex Series
intel Data Center GPU Max Series
intel Core Ultra
intel Arc Graphics
intel Ethernet

[AI Everywhere Customer Stories]

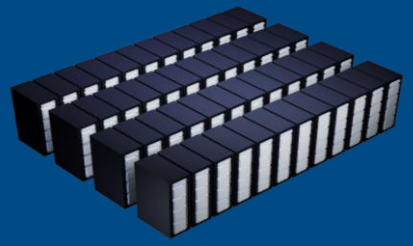
Intel Provides for the Entire AI Workflow from the Data Center, Cloud and Network, to the Client and Edge

Training and Fine-Tuning

Training | Peak Inference | Mainstream Inference/ Fine-Tuning | Baseline Inference | Endpoint Inference

Inference and Deployment

Cloud Data Center | Edge | Client

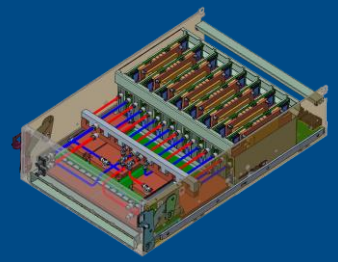


Cluster and Data Center Scale

intel. GAUDI

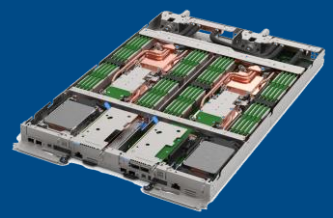
Multi-node Deployment per Rack

intel. GAUDI | intel. XEON | intel. XEON MAX SERIES



Multi-GPU or Multi-socket CPU

intel. XEON | intel. XEON MAX SERIES



Single-Socket CPU or Single GPU

intel. XEON | intel. CORE ULTRA

Client CPU

intel. CORE ULTRA | intel. ARC GRAPHICS

intel. ETHERNET

CLOUD & DATA CENTER

Bringing AI everywhere

Running AI: Cloud & Data Center

Intel® Gaudi® 2: Ideal for Efficient Training & Inference of Foundation Models

Gaudi2 is architected for deep learning performance, efficiency and scalability to meet the demands of large-scale foundation models like LLMs (GPT) and GAIs (Stable Diffusion)

Requirements	Intel® Gaudi® 2
Speed	**1.5-2x faster** than A100 for both training and inference
Memory	Each Gaudi® 2 device features **96 GB on-chip high bandwidth memory** making it easier fit large foundation models in memory, and train and deploy them at scale
Scalability	Scaling efficiency with **24x 100 GbE ports integrated on-chip**, direct all-to-all connectivity between 8 cards in a server, and open ROCEv2 based communication within and across servers.
Ease of Use	Migrate or build models **with minimal code changes** with SynapseAI, PyTorch and DeepSpeed
Power Efficiency	**~1.8x higher throughput/Watt vs A100**
Cost-Efficiency	Based on purpose-built 1st-gen Gaudi® architecture that yields up to **40% better price performance** than A100 on Amazon cloud

Source: https://habana.ai/habana-claims-validation/

Accelerating Generative AI and Large Language Models with Intel® Gaudi®2

[Intel® Gaudi® 2 Remains Only Benchmarked Alternative to NV H100 for GenAI Performance]

Intel® Gaudi® 2 delivers leading performance and optimal cost savings for AI training

[Press Release]

The Gaudi® 2 deep learning accelerator performs competitively on deep learning training and inference, with up to **2.4x faster performance than Nvidia A100**[1]

[Newsroom]

[Tech Article]

Gaudi® 2 delivers compelling performance vs. Nvidia's H100[2,3] **for GPT-3 and GPT-J**

[Newsroom]

[MLCommons Announcement]

[WATCH NOW ›]

Intel webinar recording discussing the cutting-edge capabilities of the Intel® Gaudi® 2 AI processor in capturing the potential of Generative AI and Large Language Models (LLMs)

[CASE STUDY ›]

AWS instances featuring Intel® AI acceleration technologies, with Optimum Intel and Optimum Habana libraries, give companies powerful tools for generative AI implementation

[1] Performance varies by use, configuration, and other factors; workloads and configuration details available at: intel.com/performanceindex Results may vary.

[2] Performance varies by use, configuration, and other factors; workloads and configuration details available at: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Results may vary.

Deep Learning Innovation: Intel, Habana Labs and Hugging Face

The focus of Intel's ongoing work with Hugging Face is to scale adoption of training and inference solutions optimized on latest Intel® Xeon® Scalable and Intel® Gaudi® and Intel® Gaudi® 2 processors

habana

Hugging Face

The collaboration brings the most advanced deep learning innovation from the Intel® AI Toolkit to the Hugging Face open source ecosystem and informs innovation drivers in future Intel® architecture

Intel, Habana Labs and Hugging Face Advance Deep Learning Software

Getting Started with Transformers

Faster Training and Inference: Intel® Gaudi® 2 vs Nvidia A100 80GB Benchmarks

Democratized AI: Intel, Habana Labs and Hugging Face

habana

Hugging Face

20% faster Intel® Gaudi® 2 running inference on a 176 billion parameter model than Nvidia's A100[1]

1.8x advantage in throughput-per-watt over a comparable A100 server when running a popular computer vision workload on an Intel® Gaudi® 2 server[1]

LEARN MORE

Taking on the Compute and Sustainability Challenges of Generative AI

Hugging Face and Intel - Driving Towards Practical, Faster, Democratized and Ethical AI solutions

How Democratized Large Language Models Boost AI Development

[1] Performance varies by use, configuration, and other factors; workloads and configuration details available at: Supermicro L12 Validation Report of Gaudi2 HL-225H SYS-820GH-THR2, Oct. 20, 2022

5th Gen Intel® Xeon®: The Processor Designed for AI

With AI acceleration in every core, 5th Gen Xeon processors address demanding end-to-end AI workloads before customers need to add discrete accelerators

Higher Performance on AI Inference

up to **42%**

vs. prior generation[1]

General Compute Performance Gains

average **21%**

vs. prior generation[1]

Faster Natural Language Processing

up to **23%**

vs. prior generation[1]

Leadership performance with the world's best CPU for AI

5th Generation Intel® Xeon® Scalable Processors with Intel® AMX Outperform AMD EPYC[2]

5th Gen Intel® Xeon® Scalable Processors with Intel Accelerator Engines Outperform AMD EPYC[3]

COMPETITOR BRIEFS

Website

Product Brief

LEARN MORE

[1,2,3] For workloads and configurations, visit intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

Intel® Xeon®: CPU Performance Leadership in Real World AI Applications

In real work applications, Intel is disrupting the industry and democratizing AI by delivering a better performance, lower price and more balanced platform for AI inference with:

- Larger cache that helps with data locality and large memory capacity that allows to solve larger problems
- Higher core frequency, multiple scalar ports and out-of-order execution that helps accelerate compute that is single threaded or multi-threaded but scalar
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) that helps with non-DL vector compute
- Intel® Advanced Matrix Extensions (Intel® AMX) that is built-in hardware support for AI acceleration

READ MORE

Infographic

Full Tech Article

Debunking the GPU Myth: How CPUs with Built-In Accelerators Revolutionize AI

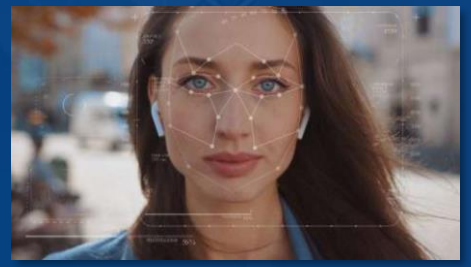
4th Gen Intel® Xeon® Scalable Processors with Accelerators for AI Inference

Accelerators like Intel® AVX-512 and Intel® AMX are designed to improve performance, reduce latency and increase memory bandwidth, making them well suited for running demanding Inference AI workloads

Built-in Accelerators and Why You Should Use Them

Intel® Advanced Matrix Extensions (Intel® AMX)

significantly accelerates deep learning training and inference, ideal for workloads like natural language processing, recommendation systems and image recognition



Website | Solution Brief
Video | User Guide

Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

can accelerate classical machine learning and other workloads in the end-to-end AI workflow, such as data prep

Website Solution Brief
Video User Guide and Downloads

READ THE ARTICLE

Taboola Improves Content Recommendation Engines

4th Generation Intel® Xeon® Scalable processors with Intel® AMX outperforms AMD EPYC

Drive Revenue Growth and Improve Customer Experience with Faster, More Personalized AI

- Better inform business decisions to drive revenue growth
- Improve customer retention and acquisition
- Increase engagement and improve conversion rates
- Reduce repetitive tasks, costs, and time for your business

READ THE INFOGRAPHIC

Discover how 4th Gen Intel® Xeon® Scalable processors with Intel® Advanced Matrix Extensions (Intel® AMX) outperform AMD EPYC

AI Workload: VMware on Intel® Xeon® Scalable Processors

vmware®

Intel® Advanced Matrix Extensions (Intel® AMX)

"You can run your entire end to end AI pipeline — data prep, training, optimization, inference – using CPUs with built-in AI acceleration."

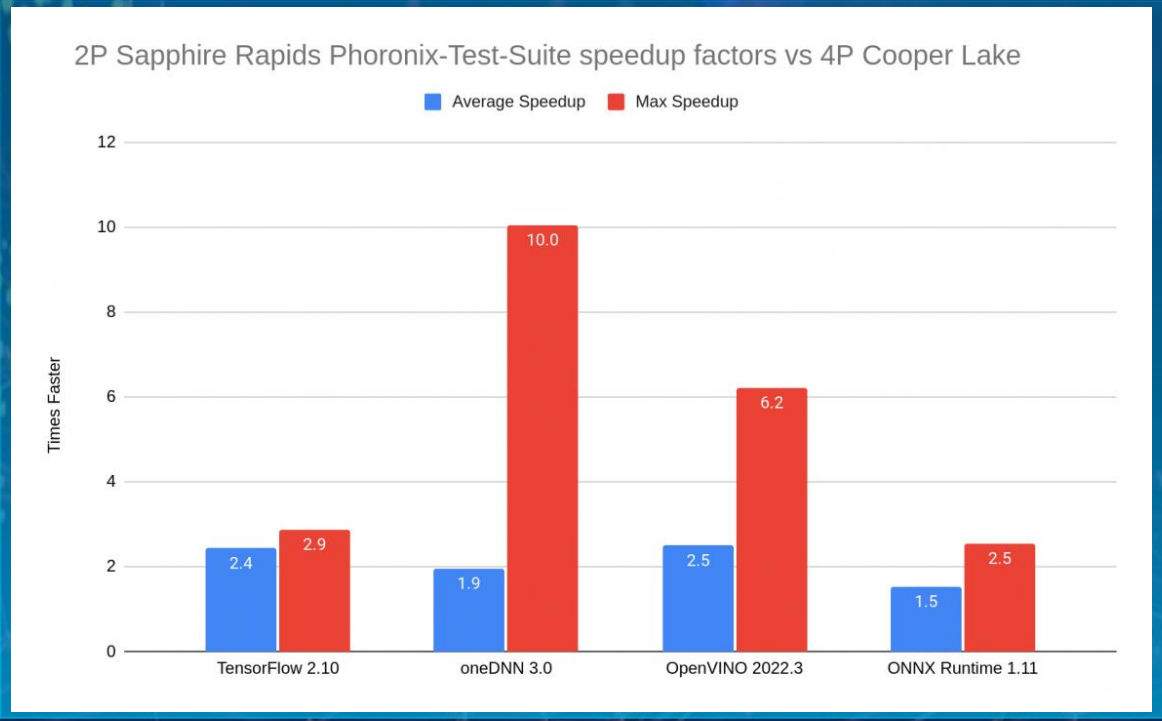
"One thing you can do to increase the performance of your AI/ML workloads is to let the CPU's AMX instructions do some of that AI/ML work, lessening the need for expensive and hard-to-procure GPUs."

Full Article from Earl Ruby, Staff Engineer at VMware

AI Workload: Red Hat on Intel® Xeon® Scalable Processors

Red Hat Enterprise Linux achieves significant performance gains with 4th Generation Intel® Xeon® Scalable Processors

Intel® AMX



Our results show 4th gen speedup factors ranging from an average of **1.5x** **up to 10x faster**[1]

[1] https://www.redhat.com/en/blog/red-hat-enterprise-linux-achieves-significant-performance-gains-intels-4th-generation-xeon-scalable-processors

Case Studies

	Challenge	Solution / Results	Intel Products	More info
Tencent Cloud *Search engine for cloud compute service*	How to handle large-scale queries and respond promptly with the search results	Tencent can use the optimized BERT model to deliver better service experiences and to help reduce TCO	4th Gen Xeon® + Intel® AMX	Case Study
美团 Meituan *Leading retail technology company*	Cost effective vision AI services	Meituan increased the overall efficiency of its online resources by over 3x and saved 70% on service costs	4th Gen Xeon® + Intel® AMX + Intel® IPP + Intel® Extension for PyTorch (Intel® IPEX)	Case Study
SIEMENS *Medical Image Processing*	Improving efficiency of radiation therapy professionals	Supporting radiation therapy professionals with AI-based auto contouring technology increases workload efficiency, improve consistency, and help free up staff to focus on value adding work	4th Gen Xeon® + Intel® AMX + OpenVINO™	Case Study Video
Alibaba Cloud *Leading Cloud Computing Provider*	Improve performance of address-purification services	Faster end-to-end performance translates to better business results for Alibaba's customers in logistics, e-commerce, energy, retail, and finance. Using a built-in accelerator helps Alibaba control TCO	4th Gen Xeon® + Intel® AMX + Intel® oneDNN	Case Study

See case study links for workloads and configurations. Results may vary.

Testimonials on Intel's AI Technology

FOX SPORTS

"We've shaved weeks off of setup time"

"For us, Intel® Xeon® processors are a cornerstone of how we deploy technology. We run only on Intel® Xeon® CPUs, and that gives us the ability to run everywhere: in VMs, in dedicated on-premises bare metal, in the cloud."

Case Study

SIEMENS

35x speedup in AI inference time for auto contouring algorithms compared to previous gen[1]

20% reduction in energy consumption compared to previous gen[2]

Case Study Video

[12]See case study links for workloads and configurations. Results may vary.

Intel® Data Center GPU Max Series and Flex Series: Breakthrough Performance

Faster Results on AI and HPC Workloads

1.7x
better average performance
vs. NVIDIA A100 across 23 workloads in 7 verticals[1]

1.3x
better average performance
vs. NVIDIA A100 across 23 workloads in 7 verticals[1]

Achieve Leadership Performance and Lower TCO

up to 50%
lower TCO
for VDI vs/ NVIDIA A16 with no contracts or licensing[2]

up to 30%
greater visual inference
performance vs. NVIDIA A10[3]

More Competitive Benchmarks

[1,2,3]Performance varies by use, configuration, and other factors. Workload and configuration details available at: Infographic

Intel® Data Center GPU Max Series: Breakthrough Performance

Intel's highest performing, highest density discrete GPU

Intel's foundational GPU compute building block features:

- Up to 408 MB of L2 cache based on discrete SRAM technology, 64 MB of L1 cache and up to 128 GB of high-bandwidth memory
- Up to 128 ray tracing units built into each Max Series GPU for accelerating scientific visualization and animation
- AI-boosting Intel® Xe Matrix Extensions (XMX) with deep systolic arrays enabling vector and matrix capabilities in a single device
- oneAPI standards-based, multiarchitecture programming and tools, which boost performance and productivity and overcome proprietary programming model lock-in
- Strong performance highlighted by
  - up to 12.8x performance gain over 3rd Gen Intel® Xeon® processors on LAMMPS (large-scale atomic/ molecular massively parallel simulator) workloads running on Xeon® Max CPU with kernels offloaded to six Max Series GPUs and optimized by Intel oneAPI tools[1]

Intel® Data Center GPU Max Series is designed for **breakthrough performance** in data intensive computing models used in **AI and HPC.** Intel® Max Series GPUs enable **greater flexibility and modularity** in the construction of the SOC.

Product Brief Website Tech Article

The entire Intel® Max Series product family is unified by oneAPI for a common, open, standards-based programming model to unleash productivity and performance.

Using oneAPI optimized deep learning frameworks and machine learning libraries, developers can realize drop-in acceleration for data analytics and machine learning workflows.

1 Performance varies by use, configuration, and other factors. Workload and configuration details available at Product Brief

Case Study: Aurora Supercomputer on Intel® Data Center GPU Max Series

Solving the world's most challenging problems…faster

The **U.S. Department of Energy's Aurora Supercomputer** at Argonne National Laboratory (ANL) is expected to be one of the industry's first supercomputers to feature over 1 exaflop of sustained double-precision performance and over 2 exaflops of peak double-precision performance. **Aurora will also be the first to showcase the power of pairing Max Series GPUs and CPUs in a single system**, with more than 10,000 blades, each containing six Max Series GPUs and two Xeon® Max CPUs

Aurora Blade for Machine Learning Demo

EDGE

Running AI: Edge

Bringing AI everywhere

Create real value for your business by bringing smart processes to the intelligent edge

Leveraging the combined power of Edge Computing and AI creates better outcomes for businesses and improved experiences for customers

Processing Data at the Edge

By 2025, 75% of enterprise-generated data will be created and processed at the edge[1]

Lower latency

Edge computing does not require a round trip to the cloud before it is stored and processed, leading to reduced time to insights and greater efficiency

Reduced risk

Data at the edge is stored and processed on the IoT device itself, enabling rapid responses to events in real-time to better mitigate business risks and enhance security

Decreased costs

Edge computing keeps data at the edge for more cost-efficient storage and processing while enabling faster insights that streamline business processes

Computing and processing at the edge creates the opportunity to leverage data where it is created

[1]Source: https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders

Edge AI Enables Transformational Use Cases Across Industries

Case Studies

- VSBLTY
- Convergint
- Taco Bell

Education

- Enhance learning environments with key insights into classroom environments
- Create more engaging lesson plans with behavioral insights
- Improve school security with AI-based video monitoring

Energy

- Reduce environmental impact with AI based equipment monitoring
- Reduce energy costs with automated monitoring

Government

- Optimize personnel management
- Enhance building safety and security with AI-based video analytics
- Reduce energy waste

Health & Life Sciences

- Reduce time to insights for diagnoses and medical tests
- Improve accuracy of diagnoses
- Provide better patient care

Manufacturing

- Improve quality control processes
- Ensure employee health and safety
- Reduce maintenance costs
- Enable predictive maintenance

Retail

- Increase store traffic with engaging and personalized advertising
- Increase sales with intelligent promotions

Transportation

- Increase logistics accuracy and efficiency
- Reduce shipping costs and returns with intelligent package handling

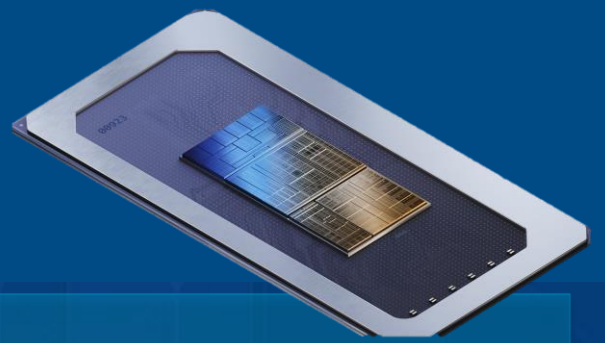
Hospitality

- Personalize and improve customer experiences at restaurants
- Streamline QSR ordering and customer queuing
- Optimize food preparation and avoid wastage

Intel is using Edge AI to improve experiences across diverse environments

Supercharge AI and Graphics at the Edge

Deploy edge solutions with advanced AI and graphics performance in power efficient BGA packaging



Enable Edge Innovation

up to **1.5x**
the AI performance
vs. previous generation[1]

up to **2.56x**
the AI performance / watt
vs. previous generation[1]

up to **1.81x**
the graphics performance
vs. previous generation[1]

Power the most-demanding edge workloads across industries

MORE INFORMATION

Infographic
Website

- Cities & Critical Infrastructure
- Healthcare
- Retail & Entertainment
- Manufacturing

[1]Performance compared to 13th Gen Intel® Core™ processors. For workloads and configurations, visit intel.com/PerformanceIndex. Results may vary

CLIENT &
WORKSTATION

Bringing AI everywhere

Running AI: Client PC

Use Cases: AI on the PC

Creator: Photo & video search & editing

Faster, more natural filters, higher quality previews & faster export times with automated, quicker searches.

Collaboration/streaming

New AI capabilities for next-gen video conferencing, streaming and collaboration, preserving battery life.

AI on the PC

Mainstream gaming

New AI features for in-game, 3D animation for added realism, transcription & chat translation.

Productivity

AI assistants for writing, creating, coding and offline features, like text & grammar prediction.

Creator: Text to image

New AI effects & features for creating images with just a few descriptive words – marketing, advertising, design.

Accessibility

AI-assisted audio-visual capabilities for diverse user needs, making it easier to create and be productive on the PC.

"Unlocking the mundane"

*Intel Dec. 14, 20323, AI Everywhere event news release

Intel Core Ultra Ushers in the Age of the AI PC

The heart of the AI PC, the new processor takes advantage of AI capabilities across operating systems and applications

First processor built on Intel 4 process technology

Largest architectural shift in 40 years

Built in Intel® Arc™ GPU3 that features up to eight Xe-cores

up to **2x** graphics performance over the previous generation[1]

Intel's newest NPU, Intel® AI Boost, built for longer-running AI workloads at low power

up to **2.5x** better power efficiency than the previous generation[1]

Michelle Johnston Holthaus, Intel Executive Vice President and General Manager of Client Computing Group

"The launch of Intel Core Ultra represents the unmatched scale and speed at which Intel is enabling AI on the PC. By 2028, AI PCs will comprise 80% of the PC market[2] and together with our vast ecosystem of hardware and software partners, Intel is best positioned to deliver this next generation of computing."

Website

Product Brief

LEARN MORE

[1]For workloads and configurations, visit intel.com/processorclaims: Intel Core Ultra 7 165H performance. Results may vary.

[2] Source: Boston Consulting Group

Case Study: 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

AI PC Acceleration Program

The AI PC Acceleration Program aims to connect independent hardware vendors (IHVs) and independent software vendors (ISVs) with Intel resources including artificial intelligence (AI) toolchains, training, co-engineering, software optimization, hardware, design resources, technical expertise, co-marketing, and sales opportunities.

Intel leads the way in AI

Engaging with

100+

AI ISV partners for AI PC Optimization

More than

300

AI-accelerated ISV features throughout 2024

More than

100m

Processors with AI accelerators through 2025

Reach out today to learn more!

ai.pc.acceleration.program@intel.com

Call to Action

Accelerate AI Development with Intel® Tiber™ Developer Cloud (formerly; Intel® Developer Cloud)

Learn, prototype, test, and run applications and workloads on a cluster of the latest Intel® hardware and software

Accelerate and **scale AI** with the latest hardware and software innovations in this development environment.
Gain more compute power and choices to **fine-tune your software** and **generative AI.**

Get Started with Intel

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

Early Technology Access

Evaluate prerelease Intel platforms and associated Intel-optimized software stacks.

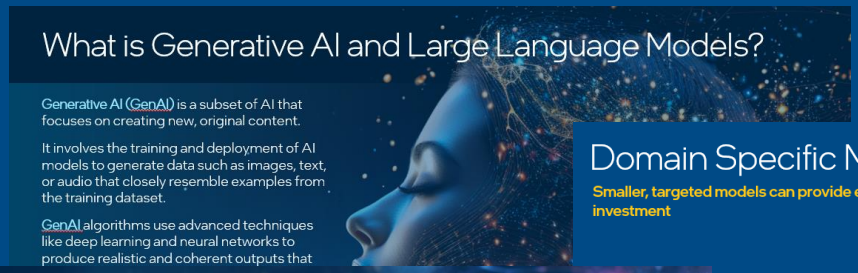
Deploy AI at Scale

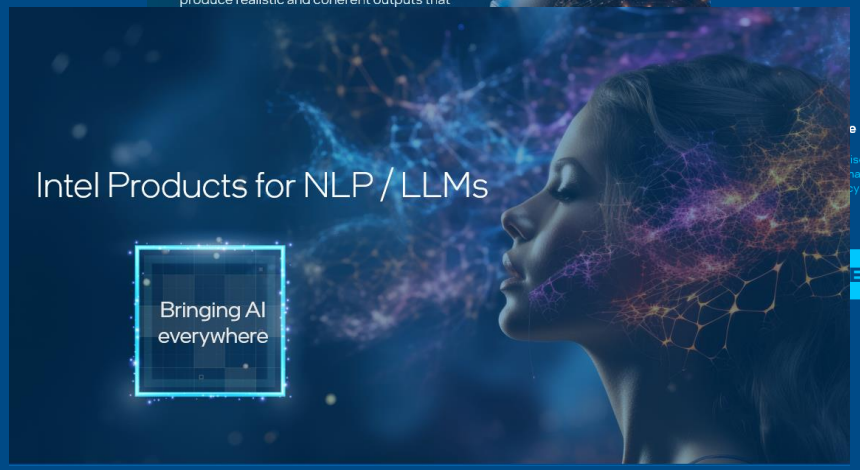
Speed up AI deployments with the latest machine learning toolkits from Intel and libraries hosted on Intel® Developer Cloud.

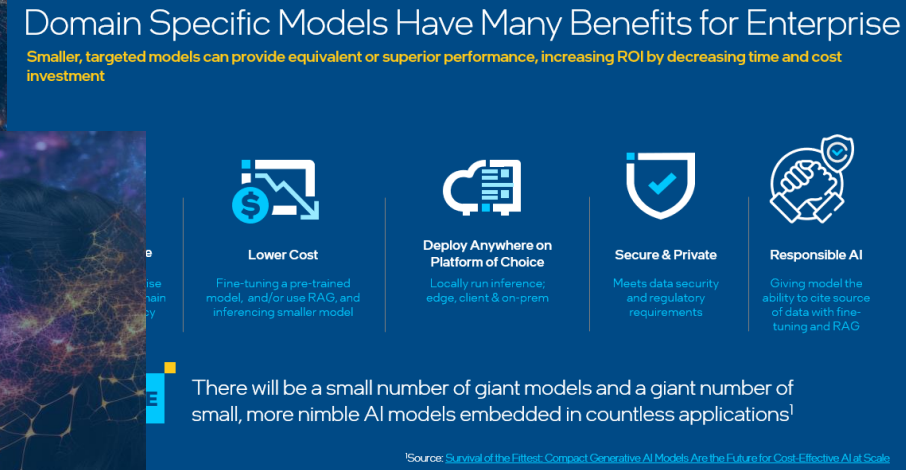
Get Started

Enterprise AI Partner Enablement Package

Check out the new Enterprise AI Partner Enablement Package to learn more about how businesses across markets can gain significant value from Generative AI, in particular domain-specific models, for long-term success







Access Now

Call to Action

Education

Understand how Intel technology can be used for your AI needs and is best suited to help you win more business

Learn more with

AI Training Assets

Engagement

Get started with a Technical Domain Meeting

To schedule a Technical Disclosure, send email to:
cloud.insider.program@intel.com

How Intel® Partner Alliance can help

Get Started with Intel® Partner Alliance

Intel Partner Alliance membership gives you exclusive business-building opportunities, like entry to our global marketplace, advanced training, and promotional support – all tailored to your needs

Training and Competencies

Admission to Intel® Partner University provides you with specialized training on advanced technologies, competency programs and rewards for learning

Marketing Resources

Entry to the Intel® Solutions Marketplace and the Intel® Marketing Studio helps you create more demand for your products and services

Valuable Rewards

Earn points for your qualifying activities, advance your membership status and get access to additional resources to build your business

If you're not already a Member

Join Now

Benefits of a Membership

Earn Points

One of the most popular and differentiated benefits within Intel® Partner Alliance are points we award partners to recognize their business results with Intel and their engagement in high priority activities. There are over 1,000 ways to earn points within Intel Partner Alliance, and 100's of redemption opportunities.

Cloud Insider Community

Intel® Cloud Insider Community offers continuously refreshed, world-class cloud content and tools. Members have the opportunity to connect with peers and the ecosystem to take innovative, joint cloud solutions to market

Learn More

Industry Insights

Gold and Titanium members can access specifically curated quarterly industry insights to help fuel their growth

Learn More

Financial Incentives

Membership unlocks powerful marketing development funds and incentive programs to accelerate your product marketing success

Speak to your Intel Representative to learn about Intel® Partner Alliance Accelerator Initiatives and more Financial Incentives

Resources

How to Access Intel® Partner Alliance Customer Support

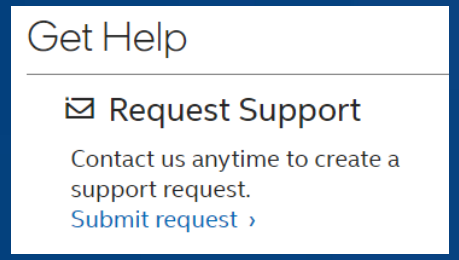
Intel Virtual Assistant

This Chat Bot, located in the bottom-right corner of each Partner Alliance webpage, provides self-help to most questions or a quick link to a live support agent.

Get Help "Blade"

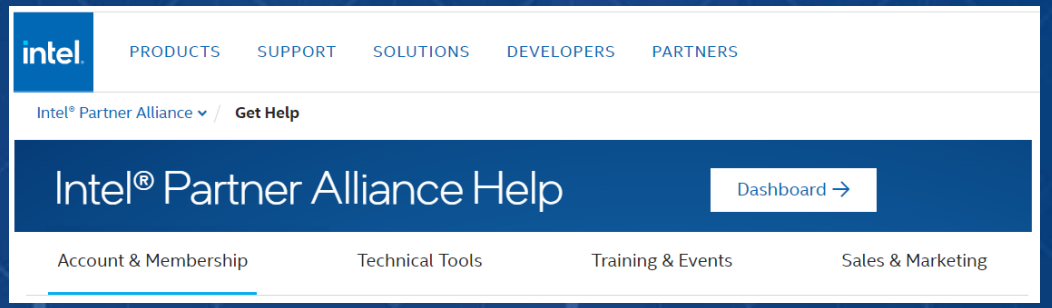
Submit an online support request.

This link is found on the footer of most pages within the Partner Alliance website.



Partner Alliance "Get Help" page

The Get Help page provides detailed self-help guides on most of the tools and benefits available to Partner Alliance members.



Cloud TV

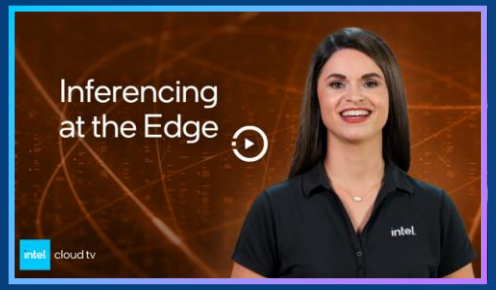
Intel® Cloud TV explores cloud computing news, trends, and strategies to drive your success

AI in the Cloud

Your GenAI Opportunity with Intel® Gaudi® AI Accelerators

AI Inferencing Using Cloud Technologies

Creating Competitive Advantage with AI in the Cloud



Gain Insights Using Data Inferencing at the Edge

Get on the Fast Path to Scale AI Everywhere

AI Activation Zones

Digital-first AI workspaces that curate critical resources, tools and benefits - activating partners to build, market, and sell solutions based on Intel technology

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

Additional Resources

Product Briefs

4th Gen Intel® Xeon® Scalable Processors

Intel AI Engines for Intel® Xeon® CPUs boost performance of the entire AI pipeline

Tech Papers

Accelerated AI Inference with Confidential Computing

Scalable End-to-End Enterprise AI on 4th Gen Intel® Xeon®

Simplify Your AI Initiatives with Technology Innovators and Intel® Technologies

Infographics

Deploy High-Performance AI Rapidly and Cost Effectively

Faster ROI from AI

Case Studies

Fujitsu | Siemens | BCM | ai.io

Use Case Solution Briefs

JelloX- Enabling a New Era of Digital Pathology with the MetaLite Digital Pathology Edge

Aurify Systems - Empowering Businesses with Accessible AI for Data-Driven Decision Making

Optalio - Bergi-Plast Optimizes Their Manufacturing Operations and Improves Quality Control with Optalio's AI and Advanced Vision Analytics Platform

Videos

Intel AI Pipeline Video

Intel® AMX: The Next Big Step in AI

Intel AI Accelerators Video

4th Gen Xeon Cloud AI Video

Additional Resources

Performance Index

4th Generation Intel® Xeon® Scalable Processors

Catalogue

AI Inference Software & Solutions Catalogue

Business Reports

Hype Cycle for Artificial Intelligence, 2022

Unlock Digital Transformation in a Digital-First Economy: Become an Artificial Intelligence Disruptor

4th Gen Intel Xeon Scalable Processors Primed to Accelerate Data Center Performance and Capabilities

Additional Training

In-deck links to Online Trainings

Training Assets

Additional AI Training Assets

Artificial Intelligence - General

- Artificial Intelligence: Workload Acceleration with 4th Gen Intel® Xeon® Processor
- Deep Dive into Securing On-Demand AI Workloads with Fortanix Confidential AI
- Why Intel AI in the Cloud?
- AI Cloud Deployment Options
- CSP AI Portfolios
- Achieve AI Performance from Data Center to Edge
- Introduction to 4th Gen Intel® Xeon® Platform

Enterprise / GenAI

- AI in the Cloud Competency
- Optimizing AI for Intel® Hardware with Hugging Face
- How to Set Up Cloud-based Distributed Training to Fine-tune an LLM
- Improving LLMs with Prompt Economization and In-Context Learning
- Streamline AI for Data Generation and Large Language Models
- Natural Language Processing
- Applied Deep Learning with TensorFlow*
- Small and Nimble – the Fast Path to Enterprise GenAI
- The Next Wave of GenAI - Domain-Specific LLMs
- Creating Competitive Advantage with AI in the Cloud
- Embracing Generative AI

Additional AI Training Assets

Edge AI

- Data Inferencing at the Edge
- Intel® Edge AI Certification
- AI on the Edge with Computer Vision
- AI from the Data Center to the Edge: An Optimized Path Using Intel® Architecture
- Choosing the Right Path to Edge Computing
- Intel® AI Products & Roadmap: Winning the AI Ecosystem to Drive Intel® Based Outcomes
- Top 3 Reasons to Elevate Edge AI & Graphics with Intel® Core™ Ultra Processors

AI PC

- AI on the PC Course
- Intel® Core™ Ultra Processor - Architecture Overview
- Intel® Core™ Ultra Processor - Power Efficiency
- Intel® Core™ Ultra Processor - Graphics and Media Deep Dive
- Intel® Core™ Ultra Processor - Meteor Lake - Artificial Intelligence (AI) Deep Dive
- What's New: Selling Commercial AI PCs
- Top 3 Reasons to Elevate Edge AI & Graphics with Intel® Core™ Ultra Processors

Legal Notices and Disclaimers

Notices and Disclaimers.


intel