

# Artificial Intelligence (AI)

Partner Enablement Package

Addressing customers' business challenges with Intel based solutions

The Intel logo is located in the top right corner of the slide. It consists of the word "intel." in a lowercase, sans-serif font, with a period at the end. The logo is white and is set against a dark blue rectangular background.

# Bringing AI Everywhere

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

# Intel AI Industry Impact



"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"



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



"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."



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



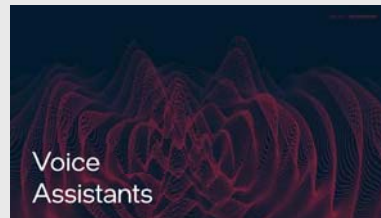
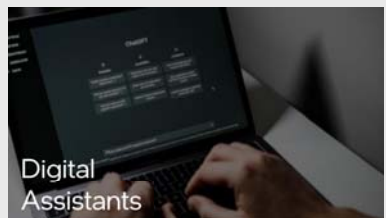
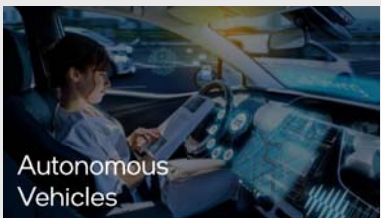
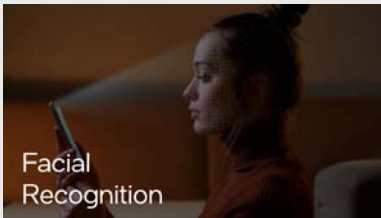
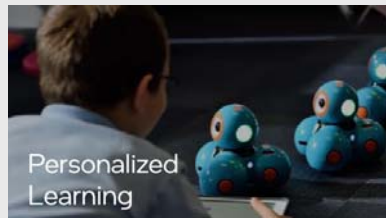
"Intel knows that AI will be everywhere by 2024 and wants its processors to be the basis for all the software technology that will flood the Internet and computer operating systems such as Windows. With this, you will be able to re-edit your favorite songs with just a few clicks or model the photos of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



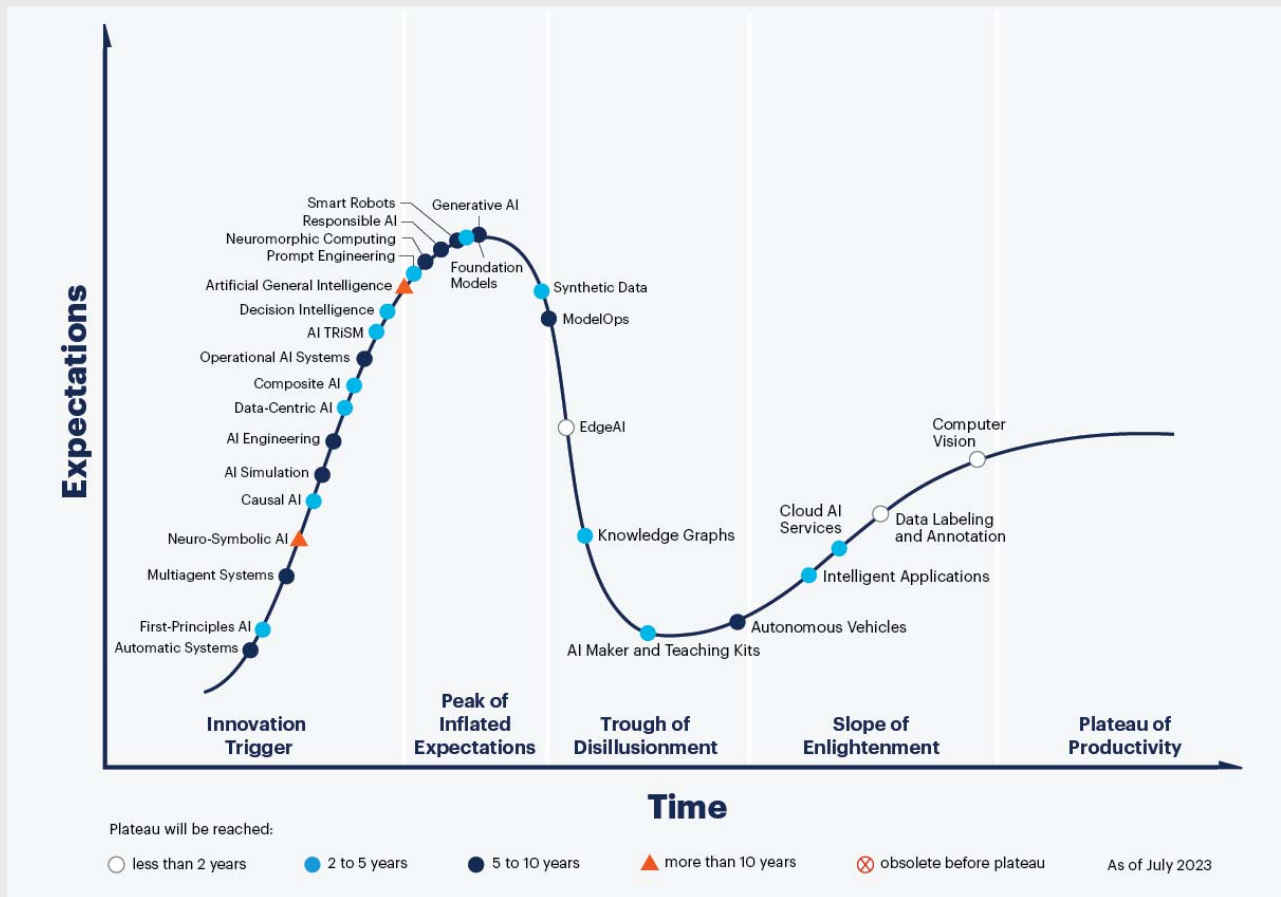
"Besides throwing out impressive numbers and claims, Intel offered some concrete, real-world examples of the kinds of AI workloads its new silicon will enable. For example, restaurants will be able to guide diners' menu choices, based on their individual budget and dietary needs, while manufacturers 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."

# 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.

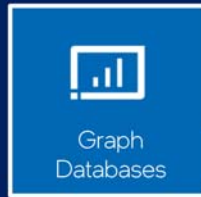
AI strategies should consider which offer the most credible cases for investment.



“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*

# 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<sup>2</sup>

**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)

<sup>1</sup><https://www.mckinsey.com/capabilities/mckinsey>

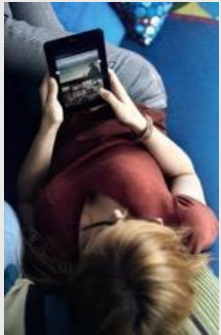
<sup>2</sup>Worldwide Artificial Intelligence Spending Guide (IDC)

<https://chiefexecutive.net/the-rise-of-the-ai-ceo/>  
[https://blogs.gartner.com/andrew\\_white/2021/07/24/bv](https://blogs.gartner.com/andrew_white/2021/07/24/bv)

Stretching to the Digital Edge, July 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.

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)

# Business Opportunity Use Case Examples



## Education

Teacher Assistant

Student Study Support

Parent Chat Portal

## Health

Drug Discovery

Doctor Co-pilot

Patient Family Chatbot

## Finance

Algorithmic Trading

Customer Portfolio Assistant

Risk / Credit Assessment

## Retail

Product Promotion

Customer Interface and Sentiment Tool

Image Shopping Aid

## Government

Gov Services Chatbot

Document Search Summarization

Live Language Translation

## Energy

Consumption Forecasting

Operational Performance

Energy Trading Assistant

## Automotive

Car Development

Multi-language in car aid

Supply Chain Optimization

## Manufacturing

Factory Automation

Predictive Maintenance

Precision Agriculture

## Telco

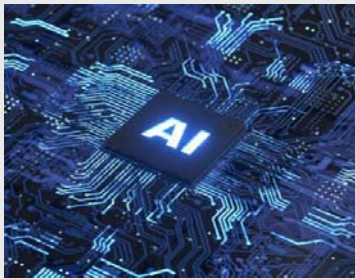
Personalized Customer Services

Network Automation

Operational Performance

# 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

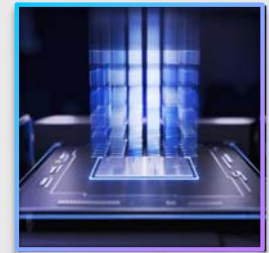
World-changing technology that improves the life of every person on the planet

## 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

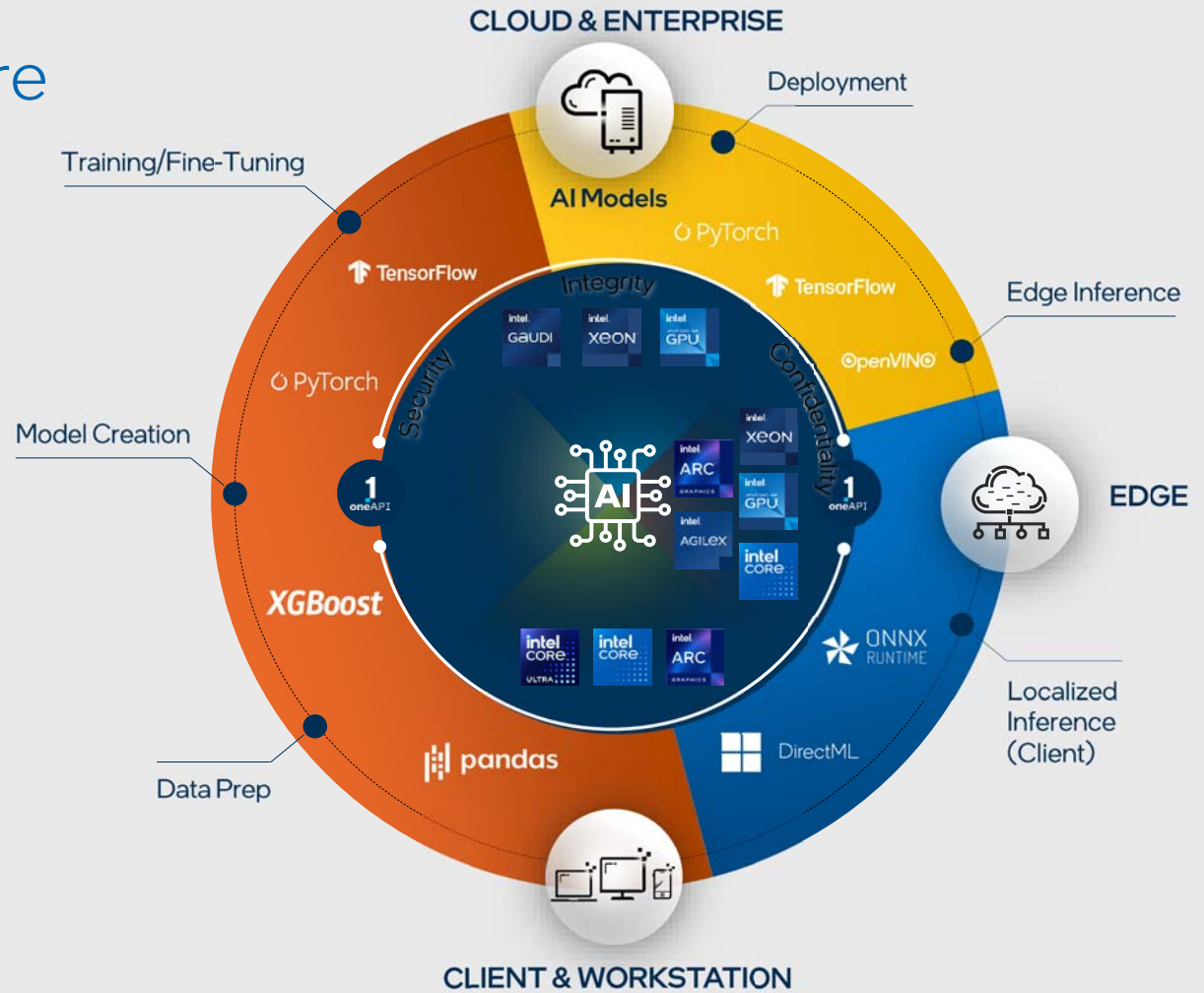


# AI Continuum

## 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



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



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



**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



**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



**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

Intel is your trusted partner to **bring AI everywhere** as we support you through every step of your AI journey



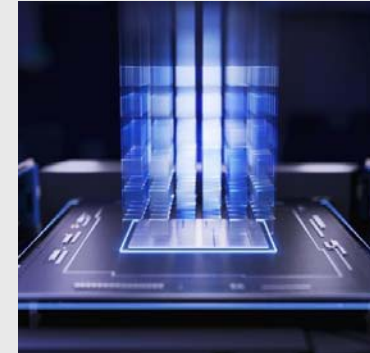
Maximize  
Value



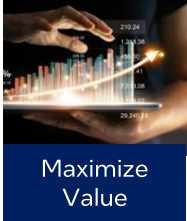
Deploy  
Anywhere



Stay  
Secure



Ecosystem  
Investment

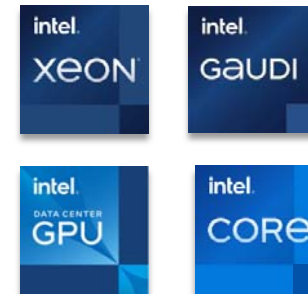


# Maximize Value

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



Maximize Value

# 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

Scalable Systems

Accelerators, Xeon

### Networking

Open Standards

Network Infrastructure

### Client & Edge

AI PC

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.



# Intel AI Portfolio

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

Open Software Environment

1 oneAPI

OpenVINO™

PyTorch

XGBoost

MODIN

TensorFlow

ONNX RUNTIME

deepspeed

Deep Learning Acceleration



Dedicated Deep Learning Training and Inference

General Acceleration



Cloud Gaming, VDI, Media Analytics, Real-Time Dense Video



Parallel Compute, HPC, AI for HPC

General Purpose



Real-Time, Medium Throughput, Low Latency, and Sparse Inference



Medium to Small Scale Training and Fine Tuning



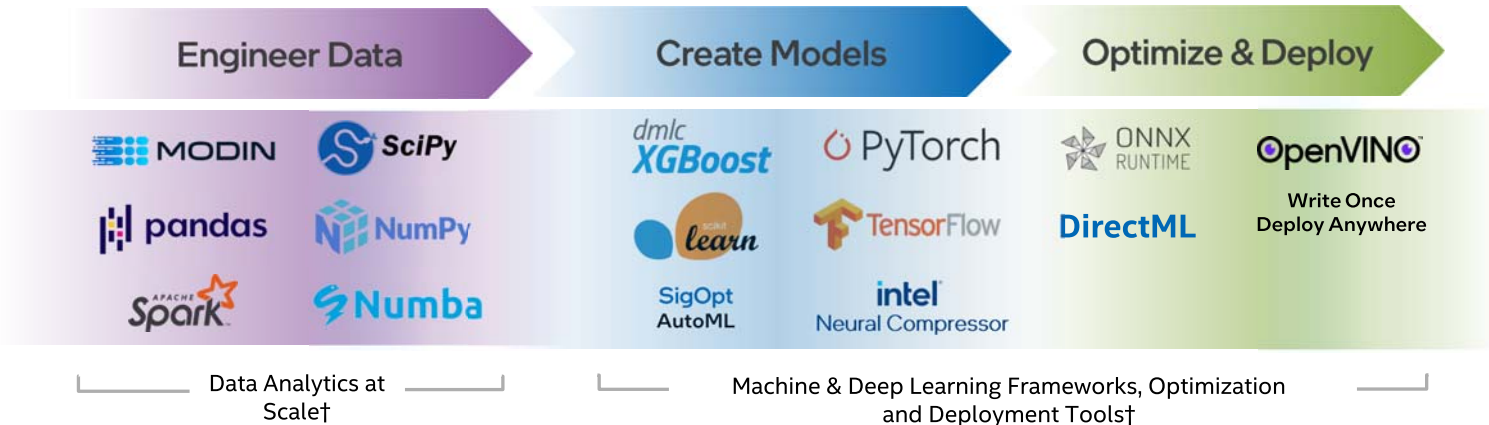
Edge and Network AI Inference



Inference on Client



# Intel® AI Software Portfolio



1  
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 & ENTERPRISE



CLIENT & WORKSTATION



EDGE

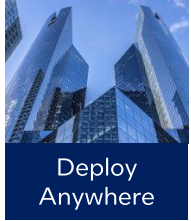


- intel AI ANALYTICS TOOLKIT  
Accelerate end-to-end data science and AI
- Intel® Developer Cloud and Intel® Developer Catalog  
Try the latest Intel tools and hardware, and access optimized AI Models
- cnvrg.io  
Full stack ML operating system
- Intel® Geti  
Annotation/training/optimization platform
- Hugging Face  
Intel optimizations and fine-tuning recipes, optimized inference models, and model serving

† every component is utilized by the solutions in the rig







# Accelerate AI Development with Reference Kits

Optimized AI reference kits help developers and data scientists innovate faster

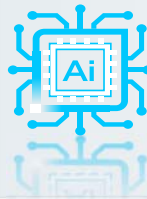
Built on the [oneAPI](#) open, standards-based, heterogeneous programming model and components of Intel's end-to-end AI software portfolio, such as [Intel® AI Analytics Toolkit](#) and the [Intel® Distribution of OpenVINO™ toolkit](#), the reference kits enable AI developers to streamline the process 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





# 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® 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

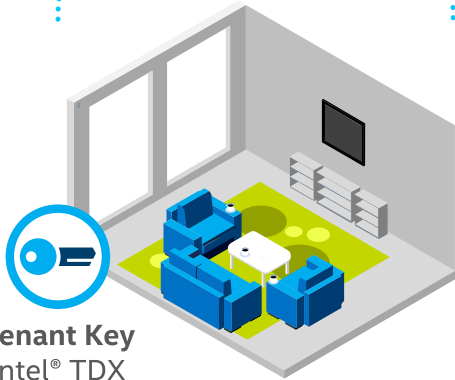


# Protecting Data in Memory

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

*Analogy: System Memory*

*Analogy: VM*

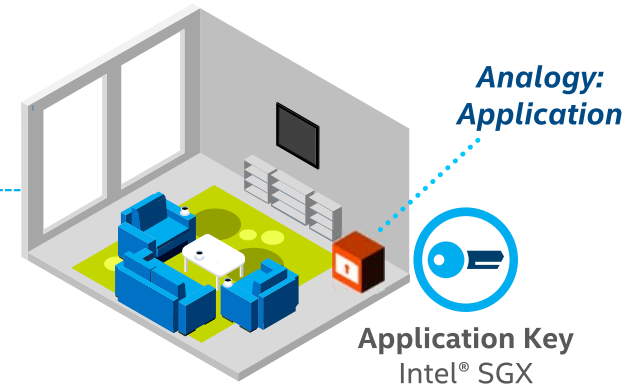


**Tenant Key**  
Intel® TDX

**Intel® TDX = Intel® Trust Domain Extensions**

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

**Single Master Key**  
Intel® TME



*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:



[\*\*Product Brief\*\*](#)

With Intel you can get better insights for critical business outcomes:



[\*\*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

**Choice**  
Compatibility

**Trust**  
AI inference

**Secure AI**

- Secure workloads
- Secure models
- Secure data in use

**Scaled**  
Develop and test

Intel® Developer Cloud

[cloud.intel.com](https://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

Bringing AI  
everywhere

# Running AI on Intel<sup>®</sup> Xeon<sup>®</sup>

CLOUD &  
ENTERPRISE



EDGE





# 5<sup>th</sup> Gen Intel<sup>®</sup> Xeon<sup>®</sup>: 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<sup>1</sup>

General Compute  
Performance Gains

average **21%**  
vs. prior generation<sup>1</sup>

Faster Natural  
Language Processing

up to **23%**  
vs. prior generation<sup>1</sup>

Sandra Rivera, Intel executive vice president and general manager of Data Center and AI Group

*“Designed for AI, our 5th Gen Intel Xeon processors provide greater performance to customers deploying AI capabilities across cloud, network and edge use cases. As a result of our long-standing work with customers, partners and the developer ecosystem, we’re launching 5th Gen Intel Xeon on a proven foundation that will enable rapid adoption and scale at lower TCO.”*

More Information

[Website](#)

[Product Brief](#)





# 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

[Full Tech Article](#)

[Infographic](#)



[Debunking the GPU Myth: How CPUs with Built-In Accelerators Revolutionize AI](#)

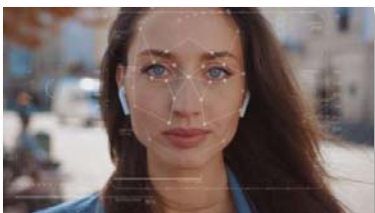


# 4th Gen Intel® Xeon® Scalable Processors with Accelerators for AI Inference

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

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

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 and Downloads](#)

## Intel® Advanced Matrix Extensions (Intel® AMX)

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](#)



# 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



VS



[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

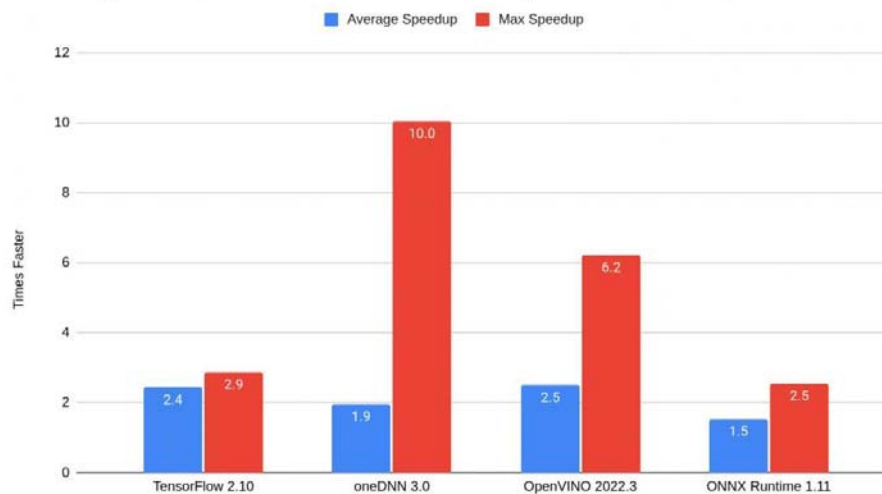


# AI Workload: Red Hat on Intel® Xeon® Scalable Processors

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

**AMX**

2P Sapphire Rapids Phoronix-Test-Suite speedup factors vs 4P Cooper Lake



Our results show 4th gen speedup factors ranging from an average of 1.5x up to 10x faster<sup>1</sup>



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



# Case Studies

 **Tencent Cloud**  
Search engine for cloud compute service

## Challenge

How to handle large-scale queries and respond promptly with the search results

## Solution / Results

Tencent can use the optimized BERT model to **deliver better service experiences** and to help **reduce TCO**

## Intel Products

4<sup>th</sup> Gen Xeon® + Intel® AMX

## More info

[Case Study](#)

 **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**

4<sup>th</sup> 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**

4<sup>th</sup> 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

4<sup>th</sup> Gen Xeon® + Intel® AMX + Intel® oneDNN

[Case Study](#)



# Testimonials on Intel's AI Technology



"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<sup>1</sup>

**20%** reduction in energy consumption compared to previous gen<sup>2</sup>



[Case Study Video](#)

<sup>1,2</sup>See case study links for workloads and configurations. Results may vary.



# Accelerate Critical Edge Workloads with Built-in AI and Security Capabilities

4<sup>th</sup> Gen Intel® Xeon® Scalable processors  
performance compared to 3<sup>rd</sup> Gen

**1.33x**

Higher Performance

**3.01x**

Higher AI inference  
performance with Intel® AMX  
for image classification

**4.25x**

Higher AI inference  
performance with Intel® AMX  
for object detection

4<sup>th</sup> Gen Intel® Xeon® Scalable processors for IoT edge deliver incredible performance, memory, I/O, resource manageability features to support workload consolidation, and **new AI instructions for deep learning training and inference** at the edge

[Learn More](#)

[AI in Production Success Stories](#)



# AI on the PC

CLIENT &  
WORKSTATION



Bringing AI  
everywhere



# 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.



## Mainstream gaming

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



## Creator: Text to image

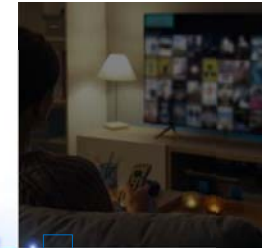
New AI effects & features for creating images with just a few descriptive words – marketing, advertising, design.

# AI on the PC

“Unlocking the mundane”

## Collaboration/streaming

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

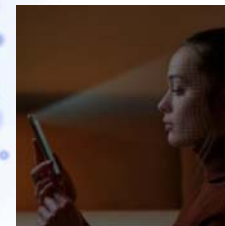


## Productivity

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

## Accessibility

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





# 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<sup>1</sup>

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<sup>1</sup>

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<sup>2</sup> and together with our vast ecosystem of hardware and software partners, Intel is best positioned to deliver this next generation of computing."*

More Information

[Website](#)

[Product Brief](#)

<sup>1</sup>For workloads and configurations, visit [intel.com/processorclaims](https://www.intel.com/processorclaims): Intel Core Ultra 7 165H performance. Results may vary.

<sup>2</sup> 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.

Get the  
details:  
[Learn more](#)



# Running AI on Gaudi2



Bringing AI  
everywhere



# Gaudi2: 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 GAs (Stable Diffusion)

Requirements	Gaudi2
Speed	1.5-2x faster than A100 for both training and inference
Memory	Each Gaudi2 device features <b>96 GB on-chip high bandwidth memory</b> making it easier fit large foundation models in memory, and train and deploy them at scale
Scalability	Scaling efficiency with <b>24x 100 GbE ports integrated on-chip</b> , 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 <b>minimal code changes</b> 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 <b>40% better price performance</b> than A100 on Amazon cloud



# Gaudi2: Accelerating Generative AI and Large Language Models

The **Gaudi2** deep learning accelerator performs competitively on deep learning training and inference, with **up to 2.4x faster performance than Nvidia A100**<sup>1</sup>

GAUDI<sup>2</sup>  
VS  
 NVIDIA.

[1 Press Release](#)

**Habana Gaudi2 and 4th Gen Intel<sup>®</sup> Xeon<sup>®</sup> Scalable processors** deliver leading performance and **optimal cost savings for AI training**<sup>2</sup>



[2 Newsroom](#)

[Tech Article](#)

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



[Webinar](#)



# 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 [Habana Gaudi®](#) and [Gaudi®2](#) processors



**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 on Habana Gaudi](#)

Faster Training and Inference: Habana Gaudi®-2 vs Nvidia A100 80GB  
[Benchmarks](#)





# Democratized AI: Intel, Habana Labs and Hugging Face



**Hugging Face**

**20%** faster Habana® Gaudi®2 running inference on a 176 billion parameter model than Nvidia's A100<sup>1</sup>

**1.8x** advantage in throughput-per-watt over a comparable A100 server when running a popular computer vision workload on a Gaudi2 server<sup>1</sup>

Read the announcement and watch the Fireside Chat here:  
[Taking on the Compute and Sustainability Challenges of Generative AI](#)



**Podcast**

[Hugging Face and Intel - Driving Towards Practical, Faster, Democratized and Ethical AI solutions](#)



**Twitter/X  
Conversation**

[How Democratized Large Language Models Boost AI Development](#)

<sup>1</sup> 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



Bringing AI  
everywhere

# Running AI on Intel® Data Center GPU Max Series



# 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<sup>1</sup>

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](#)

**1**  
oneAPI

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.

<sup>1</sup> 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](#)

# Call to Action

# Call to Action

## Education



Understand how Intel technology can be used for your AI needs and the scope upon which Intel® Xeon® product lines can 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](mailto: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.

### Get Help

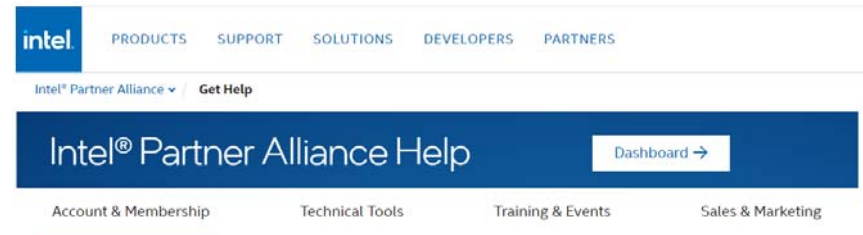
#### ✉ Request Support

Contact us anytime to create a support request.

[Submit request >](#)

## 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



[Sapphire Rapids in the Cloud](#)



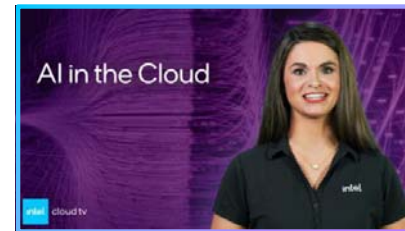
[Supercharging AI with the Cloud](#)



[Get on the Fast Path to Scale AI Everywhere](#)



[AI Inferencing Using Cloud Technologies](#)



[AI in the Cloud](#)

# AI with 4th Gen Intel® Xeon® Scalable processors

## Information and 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](#)



### Infographic

[Deploy High-Performance AI Rapidly and Cost Effectively](#)

[Faster ROI from AI](#)



### Case Studies

[Fujitsu](#) | [Siemens](#) | [BCM](#) | [ai.io](#)



### Videos

[Intel AI Pipeline Video](#)

[Intel® AMX: The Next Big Step in AI](#)

[Intel AI Accelerators Video](#)

[4th Gen Xeon Cloud AI Video](#)



### Briefcase

[Simplify Your AI Initiatives with Technology Innovators and Intel® Technologies](#)

# Additional Resources



## Performance Index

[4th Generation Intel® Xeon® Scalable Processors](#)



## Catalogue

[AI Inference Software & Solutions Catalogue](#)



## Additional Training

[In-deck links to Online Trainings](#)



## 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](#)

# Training Assets

# AI Training Assets

## Artificial Intelligence

[Artificial Intelligence: Workload Acceleration with 4th Gen Intel® Xeon® Processor](#)

ALL

[Deep Dive into Securing On-Demand AI Workloads with Fortanix Confidential AI](#)

DevOps, Cloud Architects

[Why Intel AI in the Cloud?](#)

DevOps, Cloud Architects

[AI Cloud Deployment Options](#)

Cloud Architects, C-Suite

[CSP AI Portfolios](#)

Cloud Architects, C-Suite

[Achieve AI Performance from Data Center to Edge](#)

DevOps, Cloud Architects

[Introduction to 4th Gen Intel® Xeon® Platform](#)

ALL



# Legal Notices and Disclaimers

[Notices and Disclaimers.](#)

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

The Intel logo, consisting of the word "intel" in a lowercase, sans-serif font, is positioned in the bottom right corner of the slide. The logo is white and is set against a dark blue rectangular background.

The image features the Intel logo centered on a dark blue background. The logo consists of the word "intel" in a white, lowercase, sans-serif font, followed by a registered trademark symbol (®). A small, bright blue square is positioned above the letter 'i'. The background is decorated with several semi-transparent, overlapping squares of varying shades of blue, creating a layered, geometric effect.

intel®