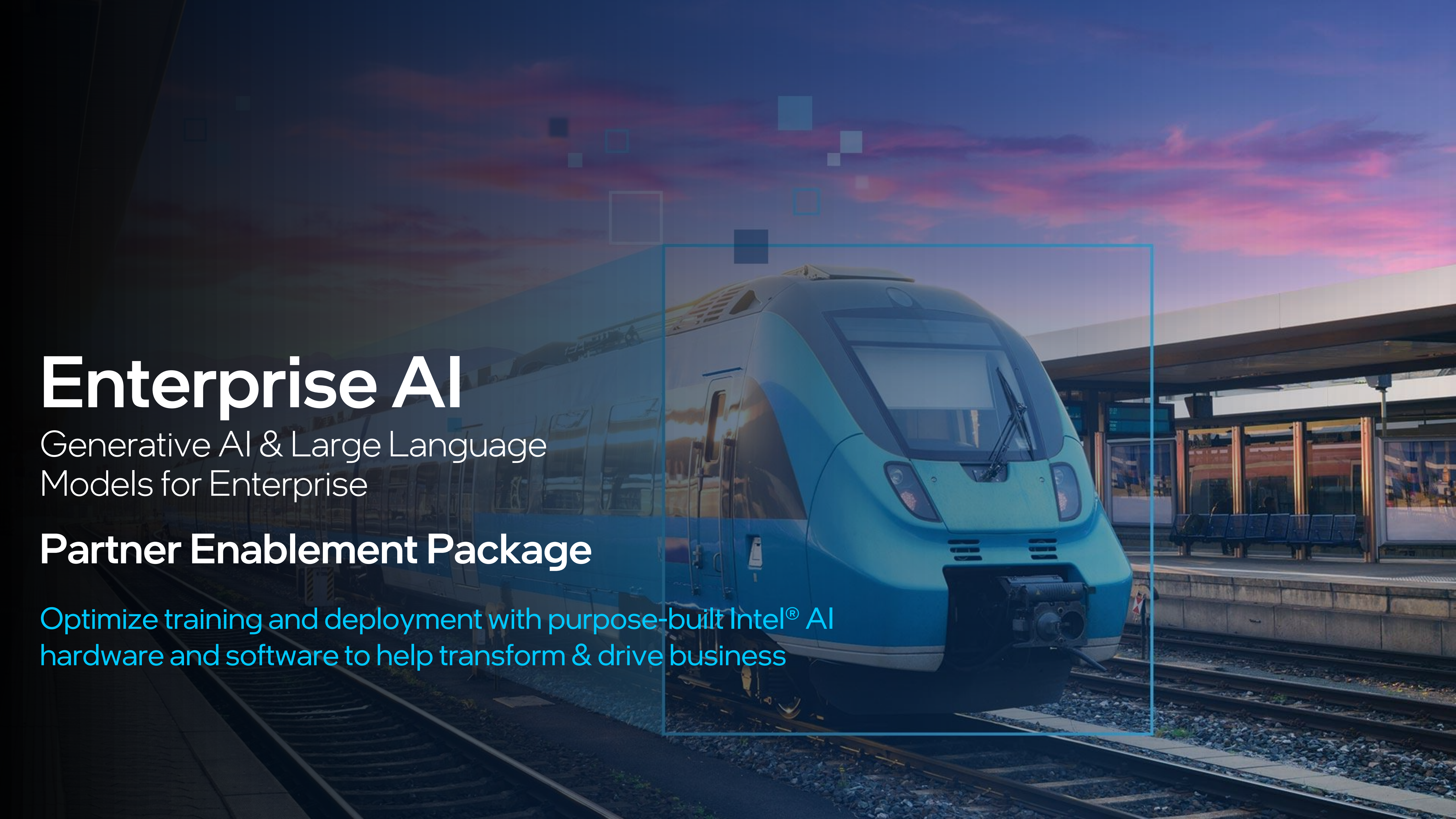


Enterprise AI

Generative AI & Large Language Models for Enterprise

Partner Enablement Package

Optimize training and deployment with purpose-built Intel® AI hardware and software to help transform & drive business

Contents

- Why Partner with Intel on Generative AI
- Generative AI Landscape
  - What is Generative AI
  - What are some of the GenAI challenges today?
- Intel AI Overview
- Intel AI Solutions
  - AI PC
  - Edge AI
  - Data Center and Cloud AI
- Call to Action & Resources

Why Partner With Intel?

Transform 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 enable AI securely and responsibly.

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

[Assess Today's Enterprise AI Opportunity Landscape]

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

Generating Value for Customers with Intel AI Solutions

Intel's approach enables a broad, open ecosystem of AI players to offer solutions that satisfy enterprise-specific GenAI needs

NAVER

To develop a powerful large language model (LLM) for the deployment of advanced AI services globally, from cloud to on-device. NAVER has confirmed Intel Gaudi's foundational capability in executing compute operations for large-scale transformer models with outstanding performance per watt.

BOSCH

To explore further opportunities for smart manufacturing, including foundational models generating synthetic datasets of manufacturing anomalies to provide robust, evenly-distributed training sets (e.g., automated optical inspection).

IBM

Using 5th Gen Intel® Xeon® processors for its watsonx.data™ data store and working closely with Intel to validate the watsonx™ platform for Intel Gaudi accelerators.

OLA

To pre-train and fine-tune its first India foundational model with generative capabilities in 10 languages, producing industry-leading price/performance versus market solutions. Krutrim is now pre-training a larger foundational model on an Intel® Gaudi® 2 cluster.

seekr

Leader in trustworthy AI runs production workloads on Intel Gaudi 2, Intel® Data Center GPU Max Series and Intel® Xeon® processors in the Intel® Tiber™ Developer Cloud for LLM development and production deployment support.

IFF

Global leader in food, beverage, scent and biosciences will leverage GenAI and digital twin technology to establish an integrated digital biology workflow for advanced enzyme design and fermentation process optimization.

airtel

Embracing the power of Intel's cutting-edge technology, Airtel plans to leverage its rich telecom data to enhance its AI capabilities and turbo charge the experiences of its customers. The deployments will be in line with Airtel's commitment to stay at the forefront of technological innovation and help drive new revenue streams in a rapidly evolving digital landscape.

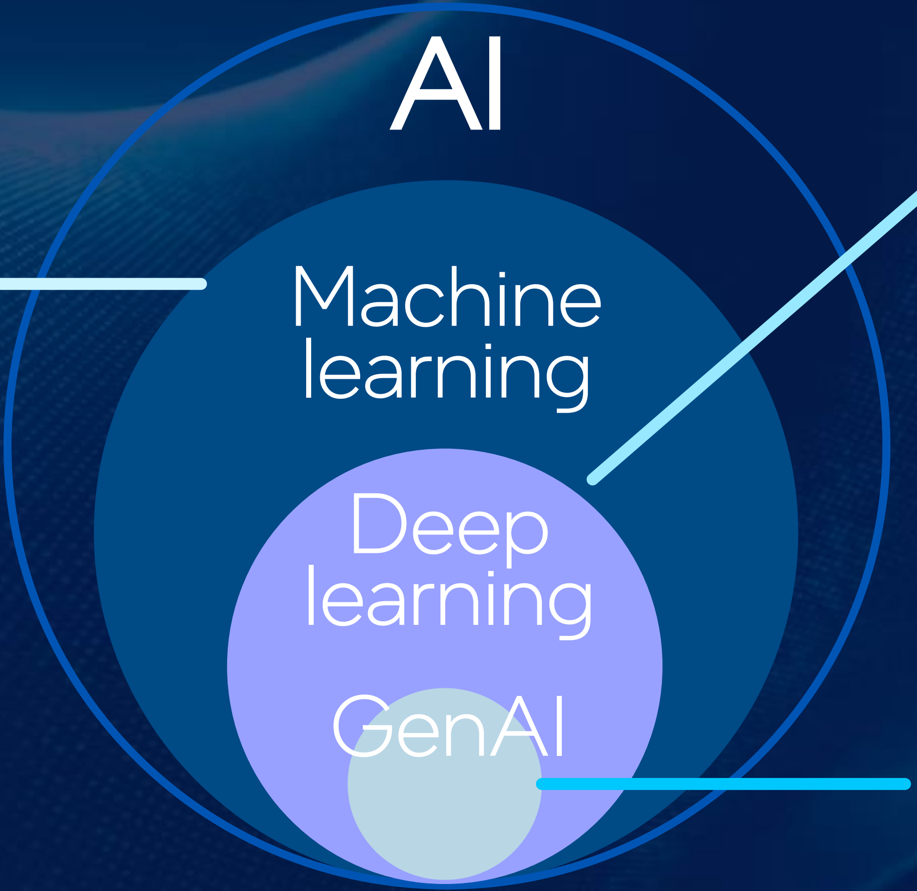
Infosys

Global leader in next-generation digital services and consulting announced a strategic collaboration to bring Intel technologies including 4th and 5th Gen Intel Xeon processors, Intel Gaudi 2 AI accelerators and Intel® Core™ Ultra to Infosys Topaz – an AI-first set of services, solutions and platforms that accelerate business value using generative AI technologies.

Ecosystem Rallies to Develop Open Platform for Enterprise AI

Generative AI Landscape

Understanding AI segmentation



Regression

Classification

Clustering

Decision Trees

Data Generation

✓ Practical to reverse engineer
✓ Tabular/limited dataset
✓ Good enough accuracy
✓ Fully-explainable

Image Processing

Computer Vision

Natural Language Processing

Recommender Systems

✓ Difficult problem to reverse engineer
✓ Large, uniform dataset
✓ Highest accuracy

Subset of AI that focuses on creating new, original content

✓ GenAI algorithms use advanced techniques like deep learning and neural networks to produce realistic and coherent outputs

What is Generative AI?

Generative AI (GenAI) is a subset of AI that focuses on creating new, original content.

It involves the training and deployment of AI models to generate data such as images, text, or audio that closely resemble examples from the training dataset.

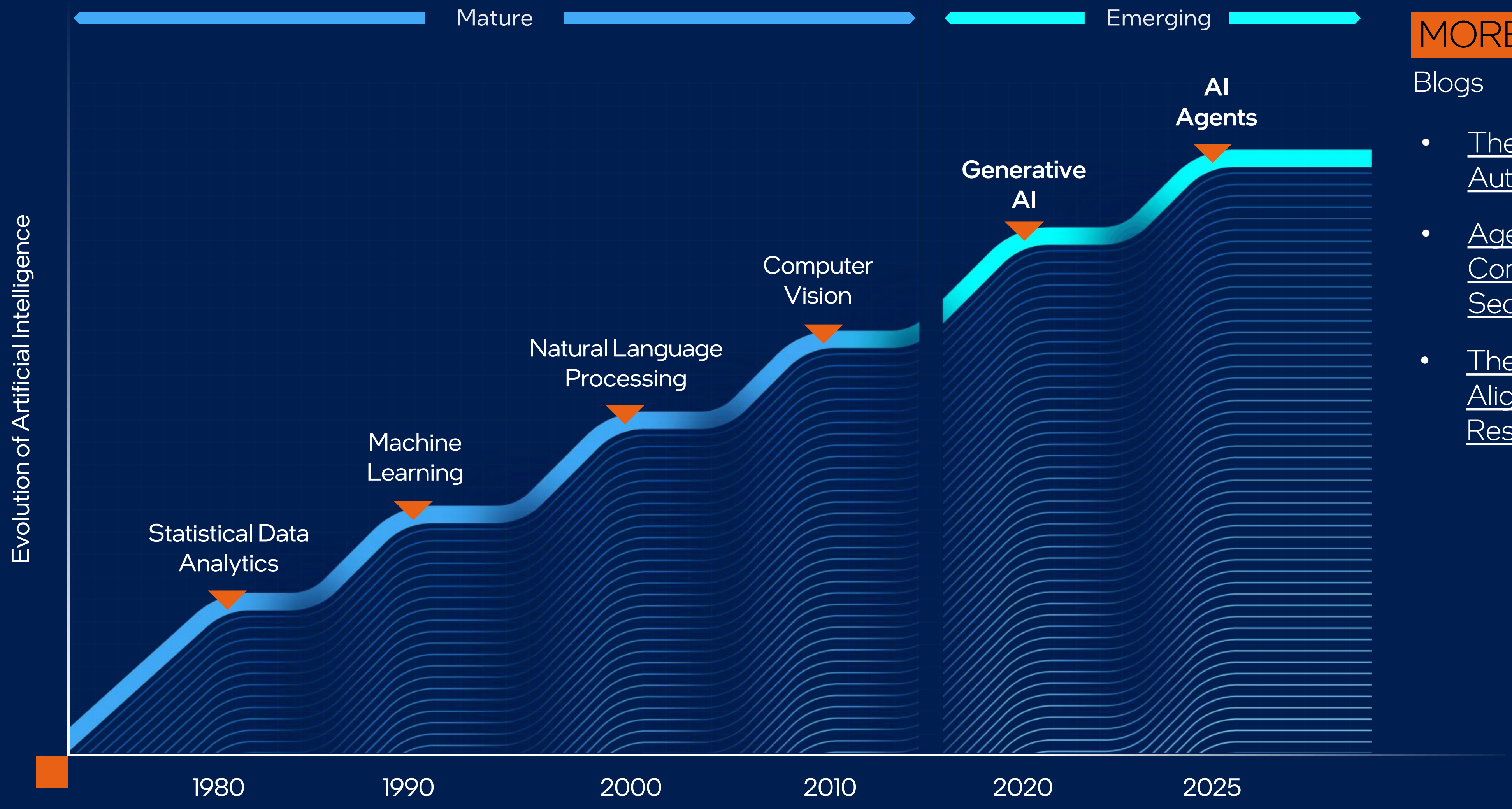
GenAI algorithms use advanced techniques like deep learning and neural networks to produce realistic and coherent outputs that enable applications like image synthesis, text generation, and even creative artwork.

Learn More

READ MORE

Capture the Power of Generative AI

Evolution of AI Applications in Enterprise Use Cases



MORE INFO >

Blogs

- The Rise of Agentic AI: Autonomous Problem Solvers
- Agentic AI and Confidential Computing: A Perfect Synergy for Secure Innovation
- The Urgent Need for Intrinsic Alignment Technologies for Responsible Agentic AI

AI Forecast 2025 to 2027 by Industry Vertical

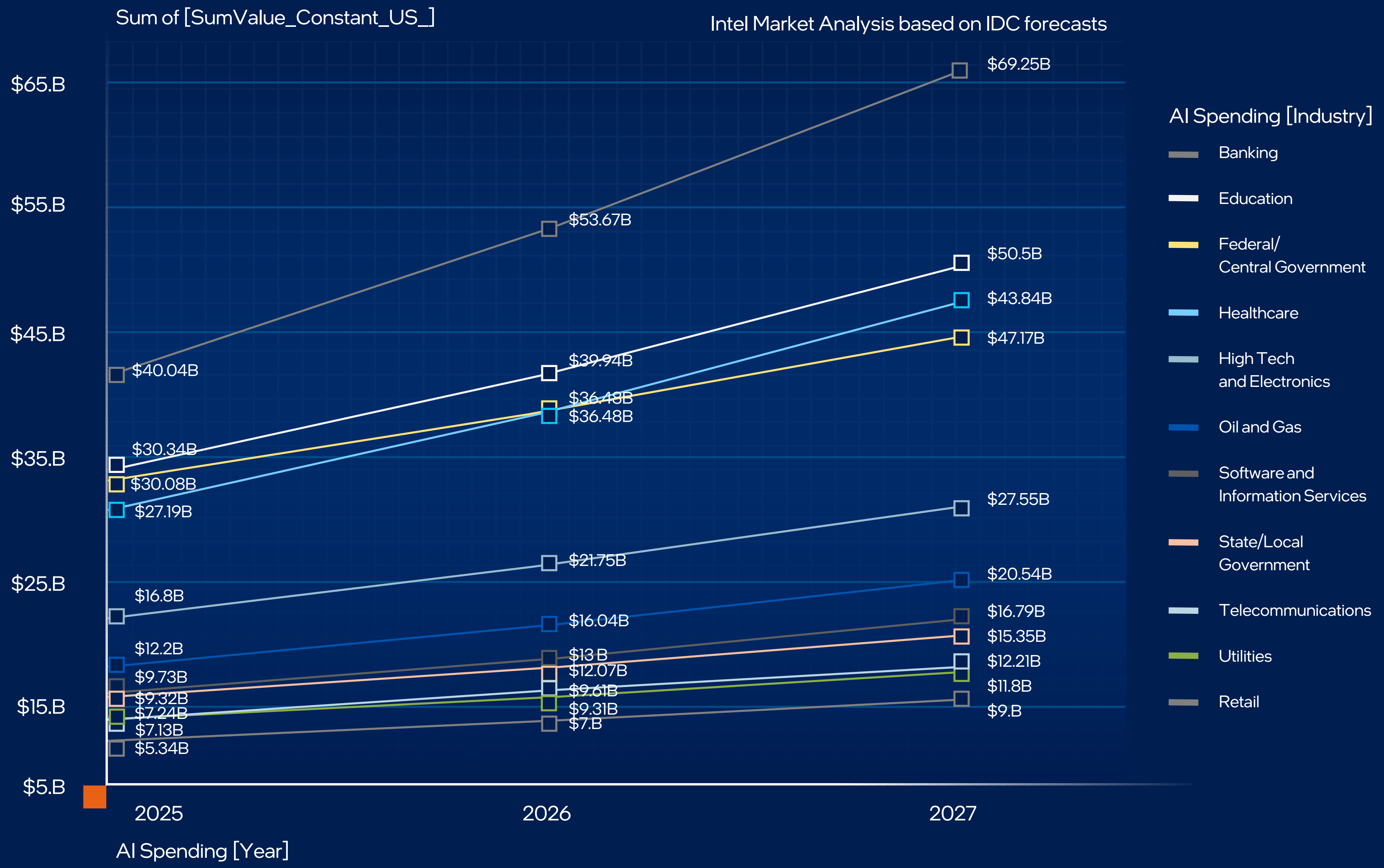
Intel will deliver Enterprise use cases in multiple industries

For Investment

- Banking leads with $69.25B expected in 2027
- While Retail, Healthcare & Software follow
- Telecom leads the rest

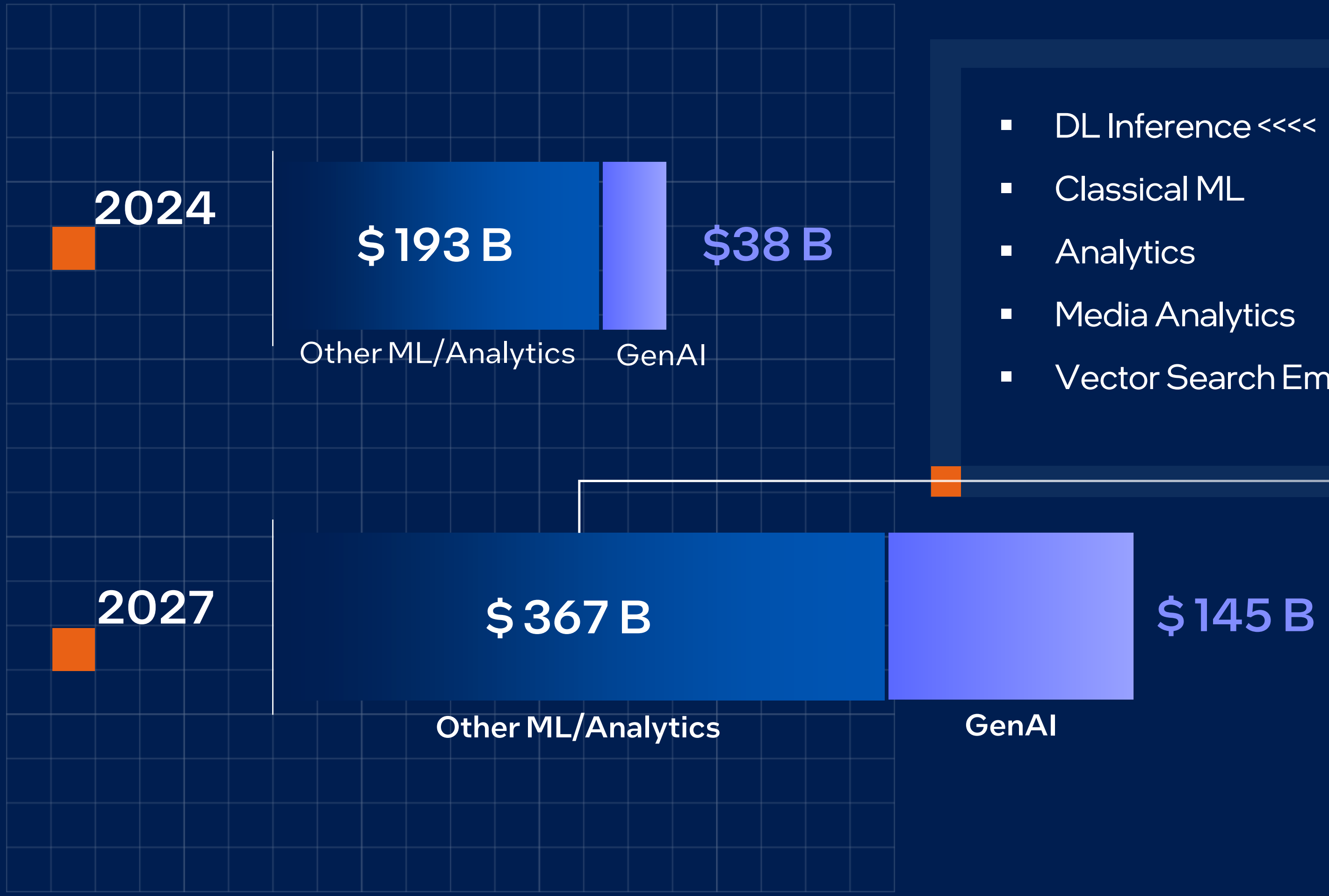
Growth wise

- Healthcare leads with ~73.5% expected in 2027
- Education sees the least investment, but 68% growth expected in 2027!



Source: Intel Market Analysis based on IDC forecasts

Gen AI and ML/Analytics Continues to Grow



- DL Inference <<<<
- Classical ML
- Analytics
- Media Analytics
- Vector Search Embedding for Enterprise RAG

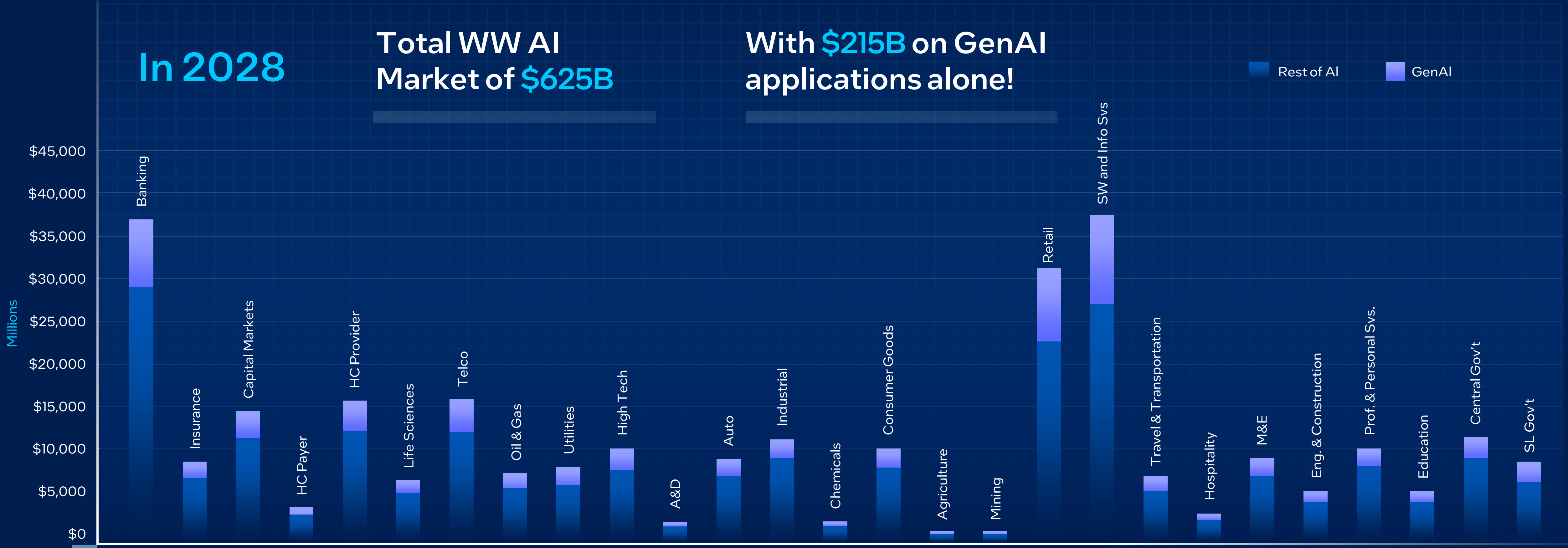
https://www.gartner.com/en/documents/4558899
https://www.idc.com/fdsfile/download/0cb3a164-e692-412b-a7ce-514e5f3bcee3

Expectations vs Reality: GenAI vs AI Overall

In 2028

Total WW AI Market of $625B

With $215B on GenAI applications alone!



Source: IDC's Worldwide AI and Generative AI Spending Guide, August (v2 2024) Preliminary Data Subject to Change

Each Vertical is Deploying AI into Many Use Cases

Example
AI Use Cases
by Industry

Education	**Teacher Assistant**	Student Study Buddy	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 Help Sentiment Tool**	Image Shopping Aid
Government	Gov Services Chatbot	Document Search Summarization	**Live Language Translation**
Energy	**Energy Consumption Forecasting**	Operational Performance	Energy Trading Assistant
Automotive	Autonomous 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

How Will Enterprises Use GenAI?

Consumer Goods & Retail

- Virtual fitting rooms
- Delivery and installation
- In-store product-finding assistance
- Demand prediction and inventory planning
- Novel product designs

Healthcare & Medicine

- Assist busy front-line staff
- Transcribe and summarize medical notes
- Chatbots to answer medical questions
- Predictive analytics to inform diagnosis and treatments

Manufacturing

- Expert copilot for technicians
- Conversational interactions with machines
- Prescriptive and proactive field service
- Natural language troubleshooting
- Warranty status and documentation
- Understanding process bottlenecks, devising recovery strategies

Media & Entertainment

- Intelligent search, tailored content discovery
- Headline and copy development
- Real-time feedback on content quality
- Personalized playlists, news digests, recommendations
- Interactive storytelling via viewer choices
- Targeted offers, subscription plans

Financial Services

- Uncovering trading signals, alerting traders to vulnerable positions
- Accelerating underwriting decisions
- Optimizing and rebuilding legacy system
- Reverse-engineering banking and insurance models
- Monitoring for potential financial crimes and fraud
- Automating data gathering for regulatory compliance
- Extracting insights from corporate disclosures

Source: Compiled by MIT Technology Review Insights, based on data from "Retail in the Age of Generative AI,"9 "The Great Unlock: Large Language Models in Manufacturing,"10 "Generative AI Is Everything Everywhere, All at Once," and "Large Language Models in Media & Entertainment,"12 Databricks, April–June 2023.

Specialized GenAI Models

The answer for the "masses"

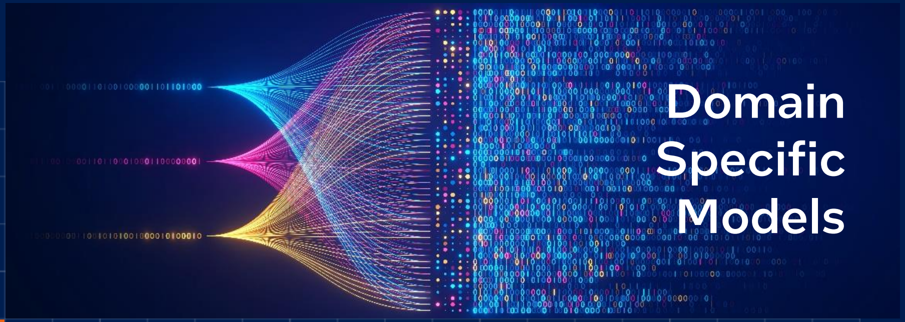
Large Frontier Models

Advantages

- + Incredible all-in-one, out-of-the-box versatility: text, programming, continual natural language conversation and plain summarization
- + Surprisingly, compelling outcomes

Challenges

- - Big (>400B parameters)
- - Expensive to train/inference
- - Hallucinations; lack of explainability, intellectual property issues
- - Frozen in time (sampling)



Domain Specific Models

Advantages

- + 10-100x smaller models while maintaining/improving accuracy, 3B – 80B Parameters
- + Economical on general-purpose compute
- + Correctness; Source attribution; Explainability
- + Utilizing private/enterprise data
- + Continuously updated information
- + Fast RAG accurate deployments

Challenges

- - Reduced range of tasks

Intel AI Overview

Intel's AI Strategy

Open	Less cost, No lock-in.
Innovation	AI PC to Edge to Datacenter & Cloud
Efficient	Performance per $ & per W leadership
Secure	Data as your IP & Models as your IP

PUBLIC

- Power Your AI Transformation with Intel

Building an open & accessible ecosystem

That drives innovation through shared access and exchange of ideas

ACCESS NOW >

Optimize enterprise AI results with Intel

Driving an open AI application ecosystem

By contributing to PyTorch and other leading AI frameworks with optimizations through communities like Hugging Face

PyTorch

Hugging Face

Open AI Platforms & Software

Making AI software simpler with approaches like oneAPI and the Open Platform for Enterprise AI

oneAPI

Open Platform for Enterprise AI

Open Standards & Protocols for scalable AI

From foundational protocols like ethernet & CXL to scalable interfaces with PCIe & UCIe

UCIe Universal Chiplet Interconnect Express

CXL Compute Express Link

Intel Ecosystem Delivers Ready-to-use Enterprise AI Applications from Intel Optimized Priority AI ISVs

	Application AI ISVs
Enterprise	Adobe Iterate.ai SAP ITERNAL ARQIT altra VNClagoon
Pub Sector	ARQIT KAMIWAZA VNClagoon ITERNAL
Retail & Hospitality	EPIC iO centific PreciTaste sodaclick WaitTime
Financial Services	KAMIWAZA ITERNAL
Health & Life Sciences	GE HealthCare EPIC iO ITERNAL
Industrial / Mfg	Eigen Innovations Industrial Machine Vision Without Limits BOSCH VISTRY

AI ISV Partner Catalogue

intel ai

Intel Ecosystem Delivers Ready-to-use Enterprise AI Applications from Intel Optimized Priority AI ISVs

	Application AI ISVs					
Generative AI	Iterate.ai	ITERNAL	KAMIWAZA	llmware	VISTRY	iGenius
Data / Vector DB	DATASTAX	Pinecone	sas	redis		
AI Security	CROWDSTRIKE	FORTINET	BUFFERZONE	Dynamo AI		
Infrastructure	Microsoft	NetApp	PURESTORAGE	Red Hat	vmware by Broadcom	Canonical

AI ISV Partner Catalogue

Software Resources to Simplify Generative AI Training and Deployment

Open-Source Model

- a BigScience initiative BLOOM 176B params 59 languages Open-access — 176B
- BioGPT — Domain 1.5B
- Stable Diffusion — Image
- Llama2 GPT-J MPT Falcon — 7-65B LLM
- Stanford Alpaca — Fine-tuned 7B LLM
- Knowledge base

Open Software

- OpenVINO™
- Intel® Extension for PyTorch (IPEX) — PyTorch
- Intel® Extension for Transformers (ITREX)
- Intel® Extension for DeepSpeed (IDEX) — DeepSpeed
- haystack
- fastRAG

GenAI Platform

- intel. Gaudi
- intel. Xeon
- intel Core Ultra

READ MORE [Unlock Generative AI with Ubiquitous Hardware and Open Software]

Why Intel's Open-Source SW Approach is Suited to Your AI Business Needs

Avoid Vendor Lock-in

open-source standards-based software

MODIN
SciPy
dmlc XGBoost
PyTorch
pandas
NumPy
scikit learn
TensorFlow
APACHE Spark
Numba
SigOpt AutoML
oneAPI

Leverage Intel's Hardware portfolio

optimized for AI use cases

intel. XEON
intel CORE ULTRA
intel. ARC GRAPHICS
intel. DATA CENTER GPU
intel. GAUDI
intel. ETHERNET

For tomorrow's AI, create new opportunities from the client and edge, to the data center and cloud, with **software optimized hardware and open standards**

Simplifying Enterprise Generative AI Adoption and Reducing the Time to Production of Hardened, Trusted Solutions

[Open Platform For Enterprise AI]

Open Platform for Enterprise AI

Ecosystem Participants of OPEA

aiven	AMD	anyscale	ArangoDB	Articul8	BONC 东方国信	ByteDance	CDW	Canonical	China unicom中国联通
clarifai	CLOUDERA	Corsha	Couchbase	DATASTAX	Datastrato	docker	Domino	dstack	expanso
FRONTIERX	Haystack by deepset	Hugging Face	H3C	Infosys Navigate your next	intel	Iterate.ai	JFrog	KX	LlamaIndex
MariaDB Foundation	MINIO	mongoDB	neo4j	OpenEuler	pathway	Plum AI	prediction Guard	Qdrant	Red Hat
Redis	Rivos	SAP	sas	TURINTECH	vectara	vmware by Broadcom	wipro	Yellowbrick	zensar
				zilliz	ZTE				

Hugging Face Partnership for Generative AI

Hugging Face

To facilitate generative AI and language AI training and innovation, Intel teamed up with Hugging Face, a popular platform for sharing AI models and data sets. Most notably, Hugging Face is known for its transformers library built for NLP.

intel. xeon

Intel has worked with Hugging Face to build state-of-the-art hardware and software acceleration to train, fine-tune, and predict with transformer models.

The hardware acceleration is driven by Intel® Xeon® processors, while the software acceleration is enabled by our portfolio of optimized AI software tools, frameworks, and libraries.

intel. GAUDI

Intel® Gaudi® deep learning accelerators are also paired with Hugging Face open-source software through the Optimum Habana Library to enable developer ease of use on thousands of models optimized by the Hugging Face community.

Get started with Intel® Gaudi® using Hugging Face

CASE STUDY > seekr Building Trustworthy LLMs for Evaluating & Generating Content at Scale

Responsible AI for Enterprise

Generative AI models learn from vast amounts of data available on the internet, which can contain biases present in society and may inadvertently apply these biases. LLMs can be manipulated to generate or spread misinformation, phishing emails, or social engineering attacks.

LLMs can often have "hallucinations" and generate inaccurate information, which can be particularly problematic in industries like healthcare, where models can influence diagnostic and therapeutic decisions and potentially harm patients.

Learn More

Minimizing the Risks of Generative AI

SOLUTIONS:

Companies and individuals working on AI technology need to make sure their software is developed and deployed according to ethical AI principles

The open-source Intel® Explainable AI Tools allow users to run post hoc model distillation and visualization to examine the predictive behavior of both TensorFlow* and PyTorch* models

LLMs are typically trained on large public datasets and then fine-tuned on potentially sensitive data (e.g. financial and healthcare)

Technologies like Intel's Open Federated Learning (OpenFL) incorporate confidential computing so that LLMs can be safely fine-tuned on sensitive data, which in turn improves the generalizability of models while reducing hallucinations and bias

Intel AI Solutions

AI PC

Edge AI

Data Center & Cloud AI

Intel Ecosystem Delivers Ready-to-use Enterprise AI Applications

AI PC

300 AI-enabled features from 100+ software vendors for

- Productivity
- Collaboration
- Creativity and content creation
- Security and manageability

https://aiswcatalog.intel.com/explore

Edge AI

Hundreds of AI solutions from Intel's partner ecosystem covering

- Financial services
- Energy and utilities
- Healthcare
- Industrial/manufacturing
- Retail
- Smart city/building
- Transportation

Data center and cloud AI

Coming soon: AI software catalog for common use cases

- Chat Q&A
- Code generation
- Code translation
- FAQ generation
- Content summarization
- Visual Q&A
- Audio Q&A

ACCESS NOW > Partner Guide: Assessing Today's Enterprise AI Opportunity Landscape

Intel Hardware Portfolio

Build, optimize and run AI at any scale

Intel provides for the entire AI workflow from the Data Center, Cloud and Network, to the Client and Edge

ACCESS NOW >

- The AI Guide: Drive Revenue Potential with AI
- Selling Intel® AI Hardware: A Conversation Guide

AI PC

Broadest AI SW ecosystem

AI PC
Light inference

Edge AI

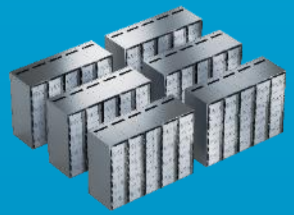
Flexible, edge node reference architectures

Node
Fine-tuning, inference

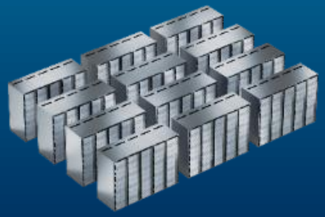
Cluster
Light training, tuning, peak inf.

Data Center & Cloud AI

Open, scalable systems & reference architecture



Super Cluster
Training, tuning, peak inf.



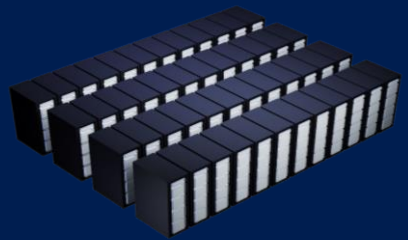
Mega Cluster
Large-scale training & inference

Scalable Systems for AI

Training and Fine-Tuning → Inference and Deployment

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

Cloud Data Center | Edge | Client

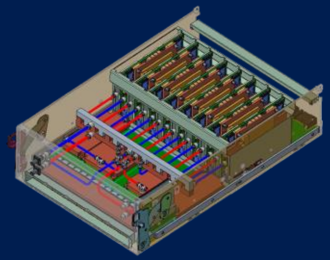


Cluster and Data Center Scale

intel GAUDI

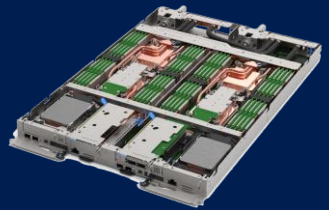
Multi-node Deployment per Rack

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

The AI Hierarchy: Mapping ML, Deep Learning, and GenAI with Intel

Discover how Intel® processors fuel AI workloads across inference, training, and next-gen generative AI applications

intel XEON

30-year history of machine learning on CPUs

- Regression
- Classification
- Clustering
- Decision Trees
- Data Generation

intel XEON

Meet Customer SLAs with general-purpose hardware

- Image Processing
- Computer Vision
- Natural Language Processing
- Recommender Systems

intel XEON

Inference

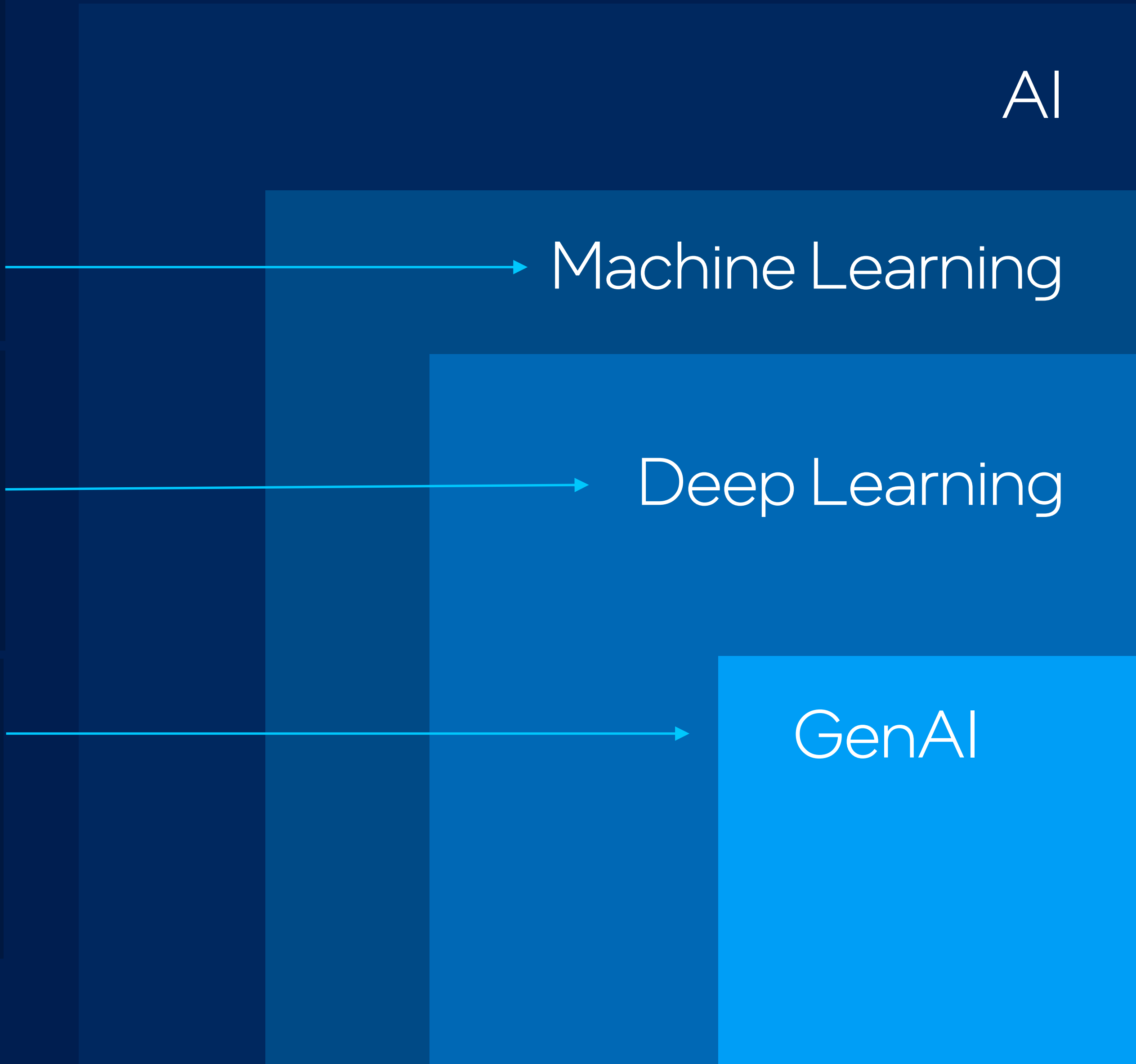
<20B Parameters

intel GAUDI

Training + Inference

Small to Large LLMs

- Subset of AI that focuses on creating new, original content



AI

Machine Learning

Deep Learning

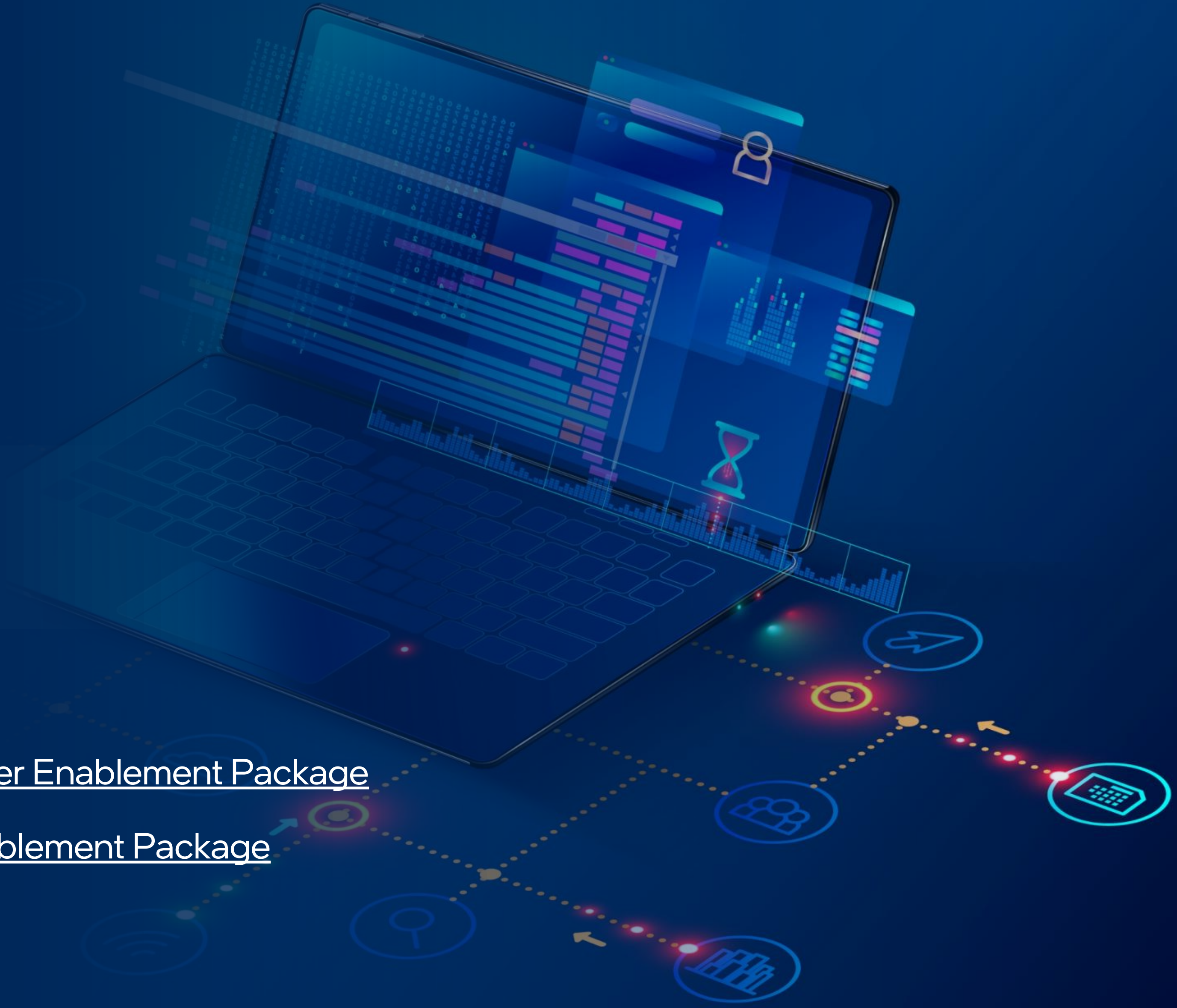
GenAI

Intel AI Solutions

AI PC

intel CORE ULTRA

intel ARC GRAPHICS



AI PC

ENABLEMENT PACKAGES

- Commercial PC Refresh Partner Enablement Package
- Security for AI PC Partner Enablement Package

Age of the AI PC

with Intel® Core™ Ultra Processor

Delivering AI PCs to Market

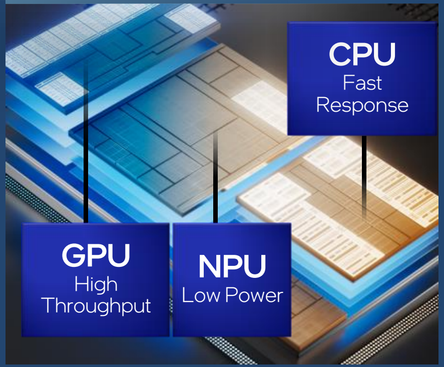
40M+ systems in 2024

100M+ systems by 2025

AI-Powered Capabilities

100+ AI experiences

CPU + GPU + NPU

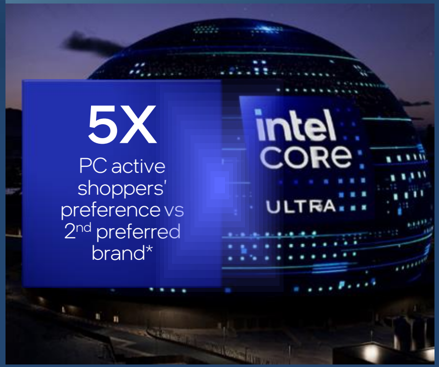


CPU Fast Response

GPU High Throughput

NPU Low Power

Unrivalled Go to Market



5X PC active shoppers' preference vs 2nd preferred brand*

* Paid research, Q1, 2023: Consumer Brand Tracker 2023 conducted in 12 markets. n=4,490
All product plans and roadmaps subject to change without notice.

Commercial AI Use Cases

New and Enhanced AI Experiences on Intel® Core™ Ultra Processors

Productivity & Accessibility

- AI LLM writing assistant
- Live captions/Transcription
- Real-time ASL-to-text translation
- AI-powered gesture/voice control

Copilot OMNIBRIDGE™ Cephable.

Collaboration

- Smart framing
- Background removal
- Eye tracking & noise suppression

zoom Windows Studio Effects (WSE)

Security

- Anti-phishing
- Ransomware & deepfake detection
- Data protection

Microsoft Defender CROWDSTRIKE McAfee proofpoint.

Content Creation

- AI avatar spokesperson
- AI-enhanced presentation

Adobe CANVID CyberLink DEEPBRAIN AI

Verticals

- AI tools for enterprises & students
- AI-driven super-resolution upscale at endpoint for VD
- Data visualizations & business insights

Power BI +ableau citrix. Google LLMWare.ai

Life Cycle Management

- Endpoint anomaly detection
- Remote screen analysis
- Data-driven insights on Digital Employee Experience (DEX)

Lakeside Microsoft Intune GoTo Workforce Solutions ivanti

Growing AI PC momentum

Intel® Core™ Ultra for Generative AI

Intel's most power-efficient client processor ushers in the age of the AI PC

Major Improvements in Efficiency and Performance

AI EFFICIENCY

up to **70%**

faster generative AI performance[2]

POWER SAVINGS

up to **25%**

reduction in power consumption[3]

READ MORE

- Product Brief
- Website

Intel® Core™ Ultra features Intel's first client on-chip AI accelerator — the neural processing unit, or NPU — to enable a new level of power-efficient AI acceleration with **2.5x better power efficiency** than the previous generation[1]

Both the Intel® Core™ Ultra H and U generation of chips include two new Low Power Island (LP-E) cores for low intensity workloads, with two Neural Compute Engines within the Intel AI NPU designed to tackle generative AI inferencing.

Accelerating AI Innovation

Intel is working with leading industry ISVs to optimize your experience with AI.

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.

Learn More

[1] As measured by Perf/Watt on UL Procyon AI benchmark while running an int8 model on Intel® Core ™ Ultra 7 165H NPU vs. Intel® Core ™ i7-1370P GPU.

[1,2,3] See www.intel.com/PerformanceIndex for workloads and configurations. Results may vary.

Decentralizing Generative AI (GenAI) Inference

On-device deployment of lightweight open-source GenAI models, including Large Language Models (LLMs), can improve accessibility and latency

Why it matters

Low Cost

Localized deployments of on-device GenAI applications do not require incremental subscriptions or fees. By optimizing lightweight models for existing underlying hardware configurations (i.e., leveraging sunk costs), organizations can achieve competitive GenAI performance at a fraction of the cost associated with cloud-based instances.

Education:

UNICEF estimates that nearly 1 billion children may face unreliable internet access and a shortage of qualified teachers. Low cost, offline access to GenAI-powered educational tools presents an opportunity to bridge this gap.

Data Privacy

On-device GenAI is complementary to agentic AI workflows, offering a local inference tier that can function offline at low latency and act upon private data enclaves for sovereign data intelligence.

Healthcare:

In crisis response situations or remote locations without reliable internet, healthcare providers could access GenAI tools that support diagnostics, triage or secure access to critical patient records.

Latency

While cloud-based GenAI scales for high-end workloads, it often entails network latency. Lightweight models with low latency and on-device compute offer a hybrid approach; AI workloads can be allocated between the cloud and ondevice depending on user requirements, resource utilization and desired quality or user experience.

Automotive:

Low-latency, on-device GenAI tools could automate certain workflows at the point of decision-making or help streamline repairs at the point of service.

READ MORE >

Decentralizing Generative AI (GenAI) Inference On Device — White Paper

ISV Case Study

VNCLagoon: On-device, AI-powered Chatbot

Situation

VNCLagoon is a comprehensive suite of communication and collaboration tools designed to adapt to the unique needs of any organization. The new VNCLagoon Chatbot - powered by Llama-3-8b-Instruct-64K, Llama 72B and TinyLlama models - acts a smart, responsive AI agent capable of handling complex inquiries with ease within the VNCLagoon Suite.

Challenge

From source analysis to text translation, the Chatbot reduces time spent on menial tasks to improve workplace efficiency. Built with extensive security requirements in mind, VNCLagoon suite and the integrated Chatbot utilize end-to-end data protection hosted locally within the platform to safeguard sensitive information, thereby addressing common privacy & security concerns.

Solution

By leveraging Intel® Core™ Ultra processors and Intel OpenVINO toolkit, this solution provides users with faster processing (inference) speed and quicker response times (by delivering more tokens per second), as well as improved runtime memory usage.

Case Study

intel.

Intel and VNClagoon -
High-speed GenAI processing with security by design

The challenge: Most generative AI models nowadays run in large cloud environments, which makes them vulnerable to malicious attacks. It's time to run them on your local or closed systems, where you can safely integrate your data and get the best performance out of your own algorithms – on a deep-state encryption and hardware level, utilizing your own PCs and your own data centers.

Learn more about this collaboration between Intel and VNClagoon.

GenAI is changing the way the world works
– it makes us even more relying on data as fuel for success in the business world.

Generative AI is promising exponential growth in solving

In a world run on data, tasks that seemed impossible to

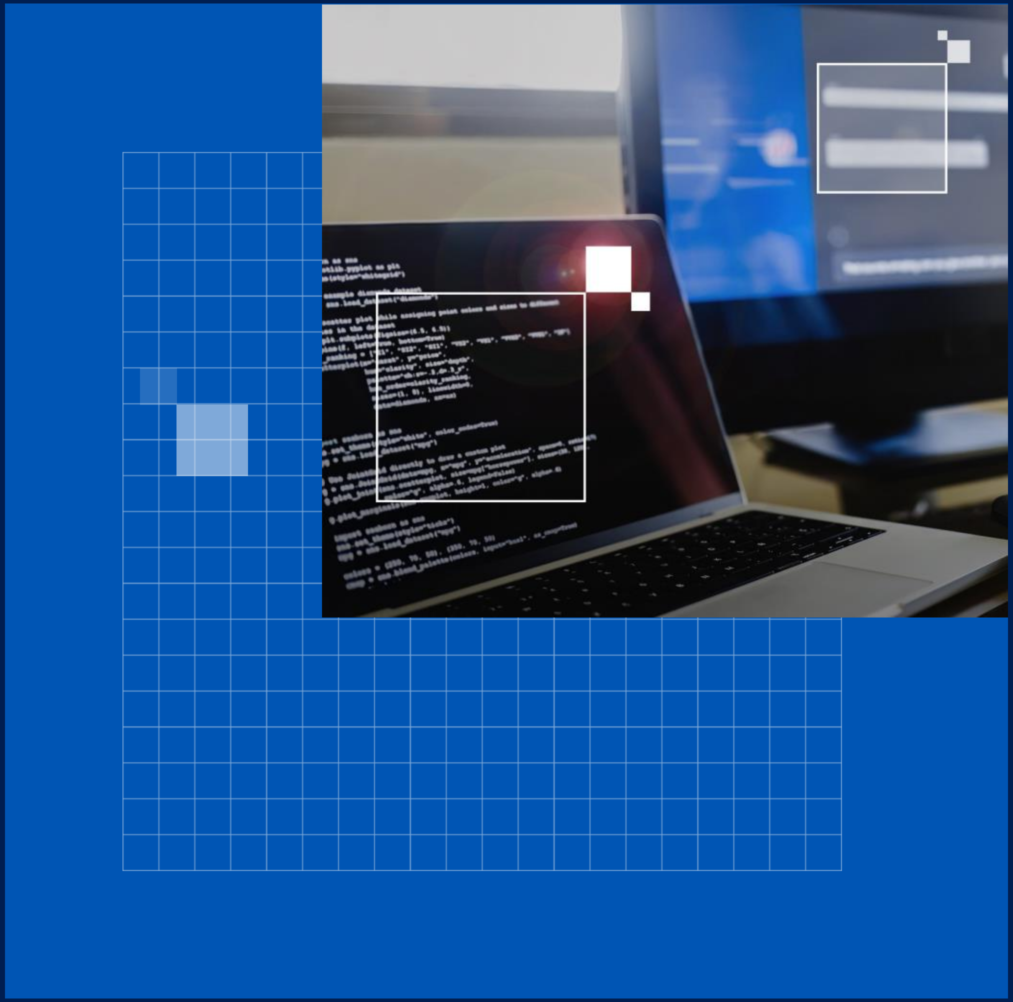
Case study

Download

ISV Case Study

Iterate Generate: Customized LLMs & Gen AI Assistants

Situation	Iterate Generate is an LLM-powered AI Assistant that can optionally be deployed both on-prem and in the cloud, ensuring security and enabling the use of enterprise and external data. Generate boosts productivity with features like Text Augmentation to re-tone, rephrase, or rewrite text, Service Pilot to summarize and suggest follow-up emails, and Document Search to summarize data, analyze trends, and enable LLM queries in an easy chat interface. It has a modular drag-and-drop, low-code AI environment that allows enterprises to develop AI applications quickly and easily.
Challenge	Generate enables new use cases that can be executed directly on-device (laptop), such as text augmentation, chatbots etc. These new usages can significantly increase enterprise user productivity and efficiency, while addressing privacy and security concerns.
Solution	Developed for Intel® Core Ultra processors and optimized with the Intel OpenVINO Inference Engine, Generate provides users with better cost efficiency and faster inference speed while maintaining accuracy, scalability, security, and flexibility. All of this is achieved due to enhanced CPU load performance and better memory efficiency.



Intel AI Solutions

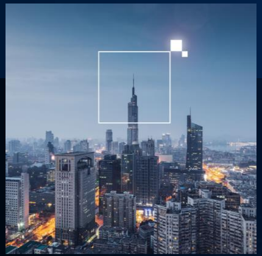
Edge AI

intel CORE ULTRA

intel ARC GRAPHICS

intel. XEON

Edge AI



ENABLEMENT PACKAGES

- Edge AI Partner Enablement Package

Empowering Industry Verticals

Accelerating digital and AI transformation

Health, Education & Consumer Industries

Health IT, medical imaging, digital signage, point-of-sale, interactive flat panel displays, multifunctional printers

Cities & Critical Infrastructure

Smart cities, safety and security, road and rail infrastructure, airports, electric vehicle charging, government edge infrastructure

Federal & Industrial

Manufacturing, robotics, commercial buildings, warehousing and logistics, utilities, integrated energy companies, military, aerospace, government

- Interactive kiosk
- Interactive flat panel display
- Digital signage
- Point of sale
- Mobile POS
- Industrial PC
- Edge server
- Orchestrated compute
- Ai box / nvr
- Security Camera
- Print imaging
- Medical Imaging

Accelerate Edge Workloads with 5th Gen Intel® Xeon® processors

Achieve incredible performance for demanding emerging AI and edge workloads. 5th Gen Intel® Xeon® processors boost AI performance and energy efficiency, improve operational efficiencies, and enable confidential computing in edge deployments.

Up to **1.59x** average performance gain[1]

Up to **1.29x** average performance-per-watt gain[2]

Up to **2.81x** higher real-time inference performance for image classification[3]

Up to **5.28x** higher real-time inference performance for object detection[4]

vs. 3rd Gen Intel® Xeon® Gold 6348 processors

CPU REFRESH

Access the Intel® Xeon® Processor Advisor Suite to calculate the best route to lower TCO and path to ROI

[1, 2, 3, 4] For workloads and configurations, visit intel.com/processorclaims: 5th Gen Intel® Xeon® processors. Results may vary.

NEW: Intel® Xeon® 6 SoC

Trusted Xeon® cores in a dense, integrated System-on-a-Chip (SoC) package designed to address space and power constraints

Acceleration

Media, network, and AI accelerators

Integration

Intel® QuickAssist Technology and Intel® Ethernet in one BGA package

Long life and power optimization

IO die with Intel 4 process for the highest efficiency and density and long-life options to support edge requirements

LEARN MORE

🟧 Product Brief

ISV Case Study

Retail: AI-Powered Automated Ordering

Situation

Sodaclick is an AI-powered voice and digital signage platform that enables quick service restaurants (QSRs) to deploy automated, contactless, and multilingual voice AI for ordering systems, drive-thrus, and self-service kiosks. By leveraging cloud-based AI models and Intel-powered edge computing, Sodaclick ensures real-time, high-accuracy voice interactions that enhance operational efficiency and customer engagement.

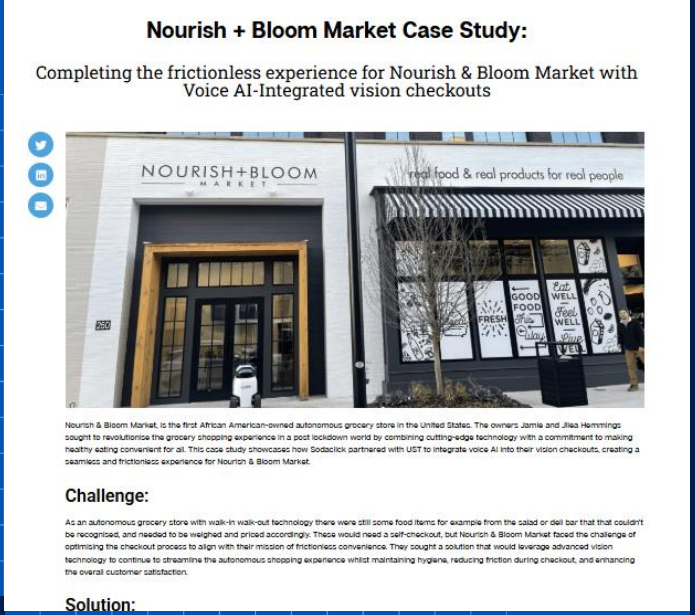
Challenge

QSRs and retail businesses struggle with long wait times, order inaccuracies, and labor shortages, making efficient customer interactions a critical challenge. Traditional touch-based interfaces and manual ordering processes can lead to inefficiencies, bottlenecks, and increased operational costs.

Solution

By integrating Intel's AI acceleration technologies, Sodaclick developed an advanced voice AI solution capable of processing natural language orders with high accuracy. Intel-powered processors and edge AI enable real-time voice processing, reducing latency and ensuring a seamless, human-like interaction. This collaboration resulted in:

- A faster and more efficient ordering process, reducing customer wait times
- Enhanced order accuracy, minimizing errors and improving service quality
- A contactless experience that meets evolving consumer expectations for safety and convenience
- Scalable AI-driven solutions that adapt to different industries beyond QSR, including retail and hospitality



Nourish + Bloom Market Case Study:

Completing the frictionless experience for Nourish & Bloom Market with Voice AI-Integrated vision checkouts

Nourish & Bloom Market, is the first African American-owned autonomous grocery store in the United States. The owners Jamie and Jilea Hemmings sought to revolutionise the grocery shopping experience in a post lockdown world by combining cutting-edge technology with a commitment to making healthy eating convenient for all. This case study showcases how Sodaclick partnered with UST to integrate voice AI into their vision checkouts, creating a seamless and frictionless experience for Nourish & Bloom Market.

Challenge:

As an autonomous grocery store with walk-in walk-out technology there were still some food items for example from the salad or deli bar that that couldn't be recognised, and needed to be weighed and priced accordingly. These would need a self-checkout, but Nourish & Bloom Market faced the challenge of optimising the checkout process to align with their mission of frictionless convenience. They sought a solution that would leverage advanced vision technology to continue to streamline the autonomous shopping experience whilst maintaining hygiene, reducing friction during checkout, and enhancing the overall customer satisfaction.

Solution:

Case study

Download

More Info:
https://sodaclick.com/case-studies/
https://www.insight.tech/retail/qsrs-voice-ai-will-now-take-your-order-with-sodaclick-3

ISV Case Study

Retail: Commercial Kitchen Automation

Situation

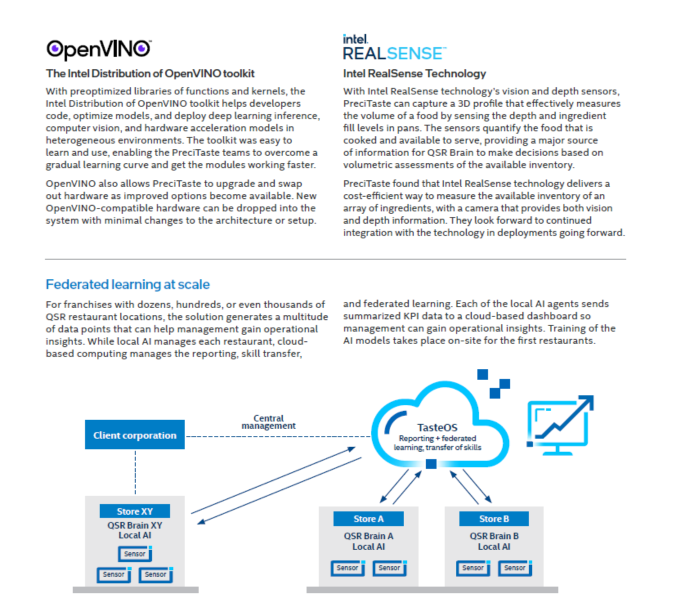
PreciTaste's QSR Brain is an AI machine vision system for digital management in the food service industry. QSR Brain senses and digitizes restaurant operations and uses AI inferencing to guide food production at retail locations. With this combination of AI inferencing technology at the edge and point-of-sale (POS) data augmented by computer vision–detected customer and vehicle sensing, QSR Brain addresses many challenges faced by this industry today.

Challenge

Manually run QSRs face production inefficiencies that impact freshness, service times, and waste. With traditional methods, food production is scheduled beforehand or is reactive once orders start coming in. Crew members must often decide between overproduction, which leads to waste and stale food, and underproduction, which leads to slow service and stockouts. With notoriously high staff turnover and minimal training, crew members need a more-precise way to plan the timing and quantities of food preparation, and managers need processes that are more sustainable.

Solution

Optimized with the Intel® Distribution of OpenVINO™ toolkit, QSR Brain also uses Intel® RealSense™ cameras in the kitchen to capture both depth and visual information. Using this solution is having a marked operational impact at major QSR franchisors by providing a more sustainable solution. The company's customers include 4 of the largest 10 QSR restaurants in the US. Overall, their use of the solution ultimately resulted in faster service, more-efficient management of food production, and labor efficiencies. One restaurant reported doubling its operating profits.



OpenVINO™

The Intel Distribution of OpenVINO toolkit

With preoptimized libraries of functions and kernels, the Intel Distribution of OpenVINO toolkit helps developers code, optimize models, and deploy deep learning inference, computer vision, and hardware acceleration models in heterogeneous environments. The toolkit was easy to learn and use, enabling the PreciTaste teams to overcome a gradual learning curve and get the modules working faster.

OpenVINO also allows PreciTaste to upgrade and swap out hardware as improved options become available. New OpenVINO-compatible hardware can be dropped into the system with minimal changes to the architecture or setup.

intel. REALSENSE™

Intel RealSense Technology

With Intel RealSense technology's vision and depth sensors, PreciTaste can capture a 3D profile that effectively measures the volume of a food by sensing the depth and ingredient fill levels in pans. The sensors quantify the food that is cooked and available to serve, providing a major source of information for QSR Brain to make decisions based on volumetric assessments of the available inventory.

PreciTaste found that Intel RealSense technology delivers a cost-efficient way to measure the available inventory of an array of ingredients, with a camera that provides both vision and depth information. They look forward to continued integration with the technology in deployments going forward.

Federated learning at scale

For franchises with dozens, hundreds, or even thousands of QSR restaurant locations, the solution generates a multitude of data points that can help management gain operational insights. While local AI manages each restaurant, cloud-based computing manages the reporting, skill transfer, and federated learning. Each of the local AI agents sends summarized KPI data to a cloud-based dashboard so management can gain operational insights. Training of the AI models takes place on-site for the first restaurants.

More Info:
https://precitaste.com/

ISV Case Study

Health & Life Sciences: Surgical Intelligence

Situation

Caresyntax offers a Data-Driven Surgical Intelligence Platform that leverages advanced AI capabilities to enhance surgical procedures and operating room (OR) efficiency. It provides real-time decision support, personalized surgical playbooks, and turn-by-turn guidance for OR staff. This collaboration aims to improve patient outcomes, streamline workflows, and address critical staffing shortages in healthcare.

Challenge

The healthcare industry faces significant challenges due to acute staffing shortages, particularly in surgical departments. High turnover rates among nurses and clinical support staff have led to increased workloads, longer hours, and operational inefficiencies. These issues not only strain existing staff but also jeopardize patient safety and the financial viability of healthcare organizations, as operating rooms contribute to over 50% of hospital revenues.

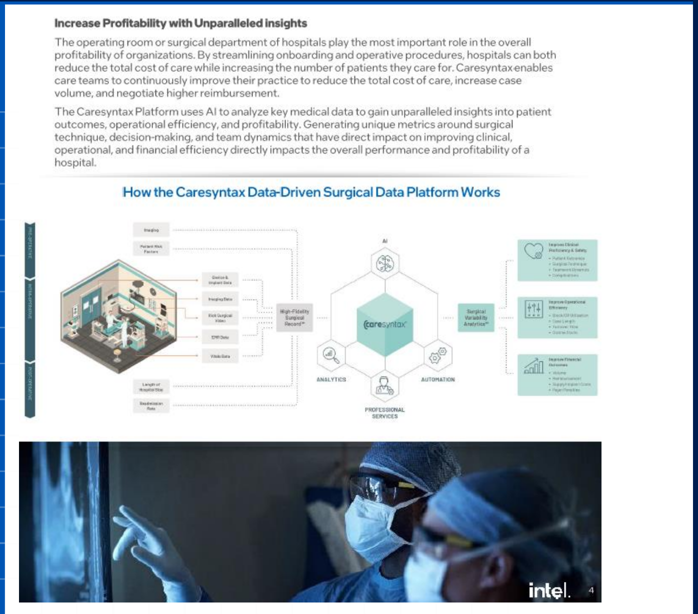
Solution

By integrating the Intel® Distribution of OpenVINO™ Toolkit and using Intel® Core™ Processors, Caresyntax can deliver various benefits to its customers:

Rapid Onboarding & Training of new clinicians and traveling nurses, ensuring they are well-prepared to support surgeons during procedures.

Improved Decision Support via real-time processing of surgical data to provide actionable insights, aiding surgeons in making informed decisions and reducing cognitive load.

Enhanced Operational Efficiency through analyzing large volumes of surgical data and identifying opportunities to streamline workflows, optimize resource utilization, and improve overall OR efficiency.



Increase Profitability with Unparalleled insights

The operating room or surgical department of hospitals play the most important role in the overall profitability of organizations. By streamlining onboarding and operative procedures, hospitals can both reduce the total cost of care while increasing the number of patients they care for. Caresyntax enables care teams to continuously improve their practice to reduce the total cost of care, increase case volume, and negotiate higher reimbursement.

The Caresyntax Platform uses AI to analyze key medical data to gain unparalleled insights into patient outcomes, operational efficiency, and profitability. Generating unique metrics around surgical technique, decision-making, and team dynamics that have direct impact on improving clinical, operational, and financial efficiency directly impacts the overall performance and profitability of a hospital.

How the Caresyntax Data-Driven Surgical Data Platform Works

Case study

Download

Intel AI Solutions

Data Center and Cloud AI

Data Center & Cloud AI

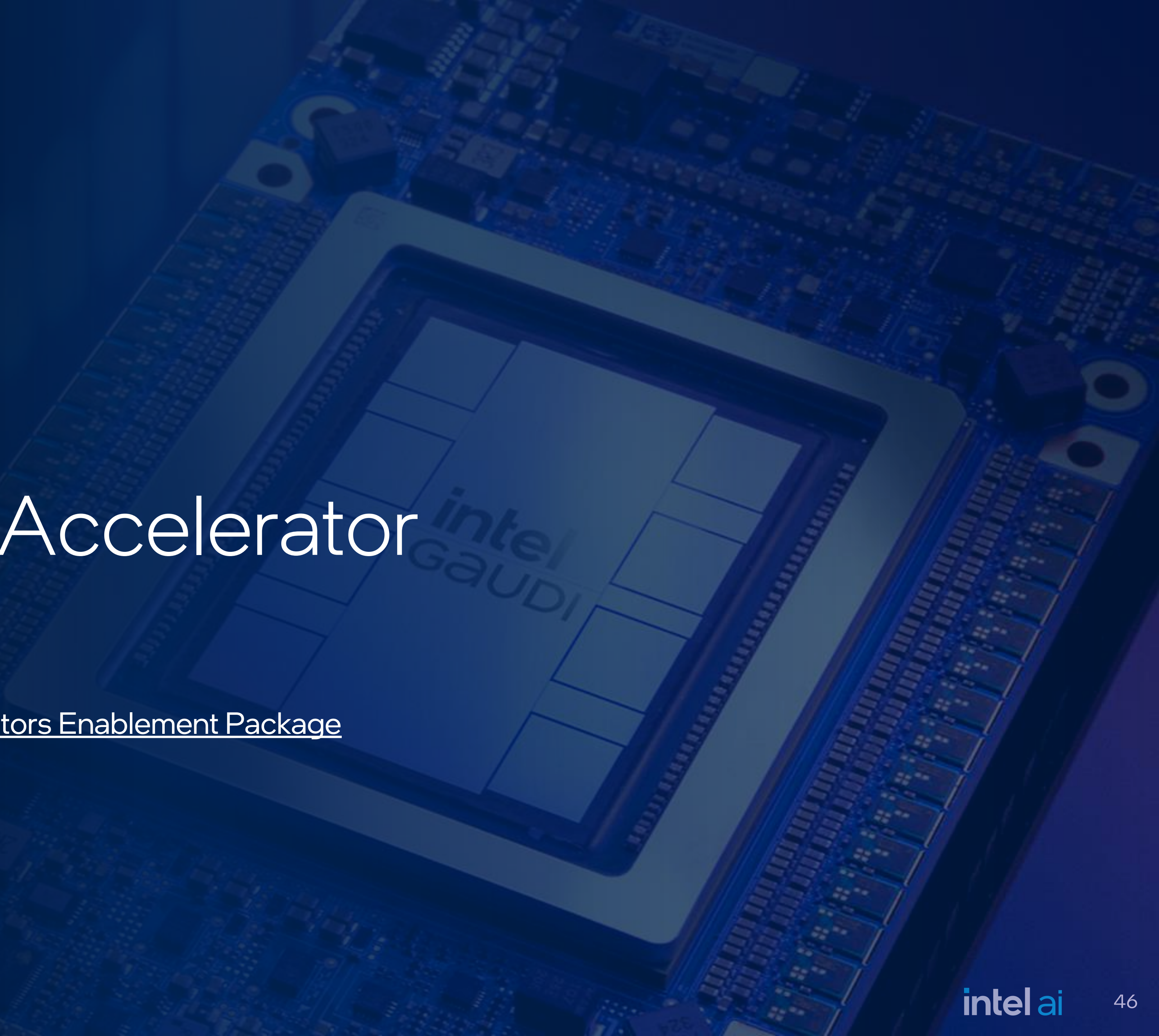
intel GAUDI

intel XEON

ENABLEMENT PACKAGES

- AI Everywhere Partner Enablement Package

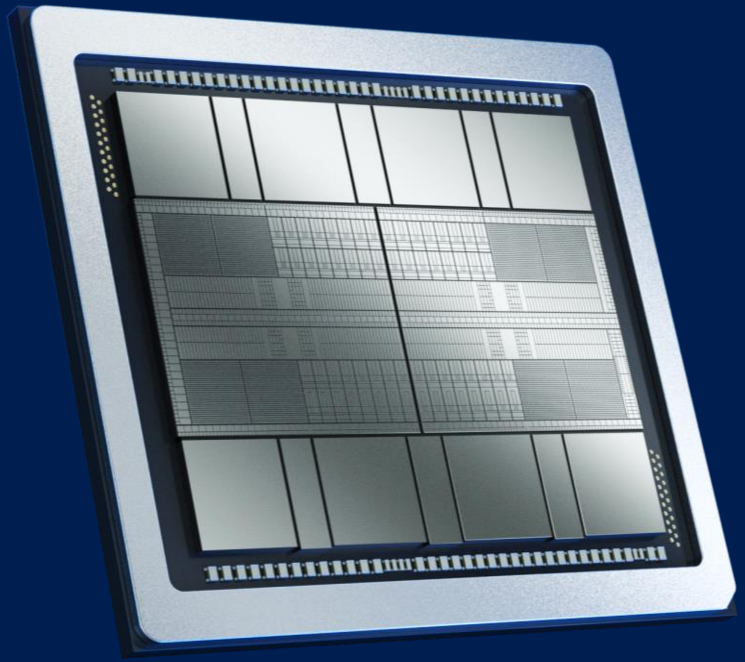
Intel® Gaudi® 3 AI Accelerator



ENABLEMENT PACKAGES

- Enterprise AI with Intel® Gaudi® AI Accelerators Enablement Package

Intel® Gaudi® 3: Architected for Gen AI Performance & Productivity



Increased memory for LLM efficiency and cost effectiveness

128GB
HBM capacity, 3.7 TB/s B/W

96MB
SRAM, 12.8 TB/s SRAM B/W

Massive, flexible, on-chip networking

Open standard vs. proprietary InfiniBand

24 x 200 GbE
Industry-standard RoCE Ethernet ports

PCIe 5
x 16

Designed for AI

Driving greater efficiency & performance

64
Tensor Processor Cores

8
Matrix Math Engines

Introducing Intel® Gaudi® 3 AI Accelerator

The Intel® Gaudi® 3 AI accelerator is designed to provide state-of-the-art data center performance for all large AI workloads, from generative applications such as large language models (LLMs) and diffusion models to multimodal model AI solutions.

High Parallel Processing Power: Intel® Gaudi® 3 is designed to handle massive parallel processing tasks efficiently, making it well-suited for training large neural networks.

Optimized Acceleration: Intel® Gaudi® 3 provides specialized acceleration for AI tasks, ensuring faster training times and more efficient computation.

High Memory Bandwidth: With its high memory bandwidth, Intel® Gaudi® 3 can manage the large datasets and numerous parameters required for Deep Learning.

Energy Efficiency: Intel® Gaudi® 3 is built with energy efficiency in mind, reducing power consumption and lowering operational costs.

AI-Specific Design: Intel® Gaudi® 3 is tailored specifically for AI workloads. This means it cannot be used for tasks like graphics processing or blockchain mining. This specialization ensures superior performance and efficiency for AI applications.

Visit the website: www.intel.com/gaudi3

WATCH NOW > Intel® Gaudi® 3 explained in 60 seconds

Intel® Gaudi® 3 Benefits

More choice

versus single GPU provider

Better price-performance than competitors

Simple adoption

for new or existing models

Migrate your models with as few as 3 - 5 lines of code

Improved efficiency

across business challenges

Integration of open-source frameworks

Massively scalable

while containing costs

Readily scales Gen AI workloads to thousands of nodes

Open model

software and networking

Community-based stack using industry-standard frameworks

Future-ready

to preserve investments

Software-compatible with next-generation Intel GPUs

WATCH NOW > Intel® Gaudi® 3 AI Accelerator Explainer Video

- ✓ On-premise deployment from single systems to large clusters
- ✓ Cloud-on-demand instances from top-tier cloud providers
- ✓ Train and deploy Gen AI models up to 1TB+ parameters
- ✓ Developed partner ecosystem for enhanced supply-chain options

How Intel® Gaudi® 3 Addresses Enterprise Challenges

Need more choice

other than single-source GPUs

- ✓ Intel® Gaudi® 3 outperforms H100 performance of LLMs for inferencing[1]
- ✓ Lower hardware cost and no CUDA licensing costs
- ✓ Industry-standard high speed ethernet

Locked-in

with proprietary software and networking

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

Ability to scale

while containing costs of infrastructure

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

Maximize efficiency

yet still solves business challenges

- ✓ Higher performance per watt than H100[1]
- ✓ Higher price-performance over H100[1]
- ✓ Integration of open software frameworks drives developer productivity

[1] https://www.forbes.com/sites/karlfreund/2024/04/09/intel-launches-gaudi-3-with-impressive-gen-ai-performance/

Intel® Gaudi® 3 AI Accelerators Benchmarks

Outstanding Results vs Nvidia H100[1]

1.8x perf/$[1]

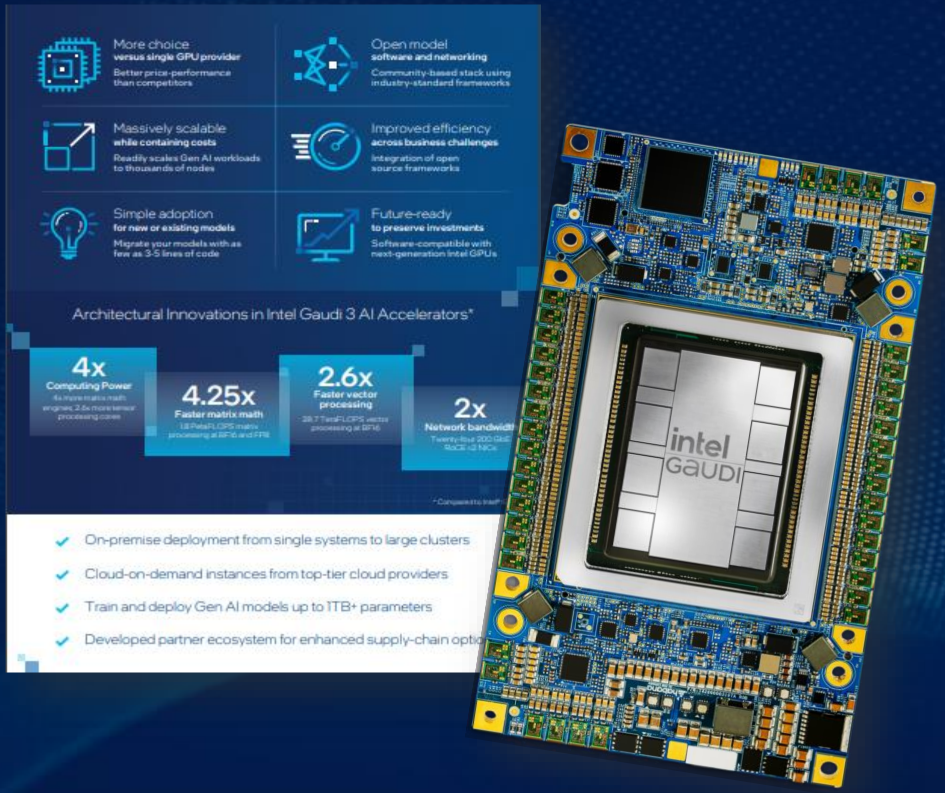
(Inference Throughput, LLaMA 3 8B)

~2x perf/$[2]

(Inference Throughput, LLaMA 2 70B)

LEARN MORE

- Quick Reference Guide
- Enterprise Sales Deck
- White Paper



More choice
versus single GPU provider
Better price-performance than competitors

Open model
software and networking
Community-based stack using industry-standard frameworks

Massively scalable
while containing costs
Readily scales Gen AI workloads to thousands of nodes

Improved efficiency
across business challenges
Integration of open source frameworks

Simple adoption
for new or existing models
Migrate your models with as few as 3-5 lines of code

Future-ready
to preserve investments
Software-compatible with next-generation Intel GPUs

Architectural Innovations in Intel Gaudi 3 AI Accelerators*

4x Computing Power
4x more matrix math engines, 2x more tensor processing cores

4.25x Faster matrix math
1.8 PetaFLOPS matrix processing at BF16 and FP8

2.6x Faster vector processing
28.7 TeraFLOPS vector processing at BF16

2x Network bandwidt
Twenty-four 200 GbE RoCE v2 NICs

*Compared to Intel®

- ✓ On-premise deployment from single systems to large clusters
- ✓ Cloud-on-demand instances from top-tier cloud providers
- ✓ Train and deploy Gen AI models up to 1TB+ parameters
- ✓ Developed partner ecosystem for enhanced supply-chain optio

All public performance benchmarks are here > https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html

[1] Input-output sequences: 128-2048tps on 2 accelerators/GPUs. Hardware: Two Intel Gaudi 3 AI Accelerators (128 GB HBM) vs two Nvidia H100 GPU (80 GB HBM).

[2] Input-output sequences: 128-2048tps on 1 accelerator/GPU. Hardware: One Intel Gaudi 3 AI Accelerators (128 GB HBM) vs one Nvidia H100 GPU (80 GB HBM).

[1,2] Intel results obtained on September 9th 2024. Intel measured results vs H100 data sources: https://github.com/NVIDIA/TensorRT-LLM/blob/main/docs/source/performance/perf-overview.md Software: Intel Gaudi software release 1.18.0. See Nvidia link for H100 software details Results may vary. Pricing estimates based on publicly available information and Intel internal analysis

Handle Demanding Enterprise Generative AI with Intel® Gaudi® Accelerators

Support a wide range of industry AI models and frameworks with cost-effective solutions for GenAI compute and enable scale-out with industry-standard Ethernet networking

Example Intel® Gaudi® use cases include:

- Chatbot assistants
- Code generation
- Content summarization
- Language translation
- Speech recognition
- Enterprise retrieval-augmented systems (RAG)
- Computer vision
- Image, video, and audio generation

Twice the throughput at a comparable price

LLAMA2-70B

~2x

higher inference performance throughput

Intel® Gaudi® 3 AI accelerator vs Nvidia H100[36]

Faster training and improved throughput		Faster model execution and support larger models	
FP8 GEMM FLOPs	BF16 GEMM FLOPs	High-bandwidth Memory (HBM)	HBM Capacity
2x	4x	1.5x	1.33x
FP8 GEMM floating operations per second (FLOPs)	**faster BF16 FLOPs**	**faster HBM bandwidth**	**larger HBM capacity**
Intel Gaudi 3 AI accelerator vs. Intel Gaudi 2[37]	Intel Gaudi 3 AI accelerator vs. Intel Gaudi 2[37]	Intel Gaudi 3 AI accelerator vs. Intel Gaudi 2[37]	Intel Gaudi 3 AI accelerator vs. Intel Gaudi 2[37]

Case Studies

AI Sweden Adopts Intel® Xeon® Processors and Intel® Gaudi® Accelerators for Virtual Assistant

"We need powerful AI infrastructure to run our enormous language models. **Working closely with Intel's team to deploy and optimize the Intel® Gaudi® accelerators made our prototype project possible.** A common digital assistant for the public sector has the potential to benefit employees daily. We hope our work can serve as a template for other countries seeking to tackle similar challenges."

Jonatan Permert, AI Transformation Strategist, AI Sweden

CASE STUDY > AI Sweden Prototypes a Virtual Assistant

AI SWEDEN

Deep Learning Capabilities of the Intel® Gaudi® 2 AI Processor Power Social Counterfactual Breakthrough

"By probing six models using data-intensive methods, the team **mitigated biases by as much as 20%**."

Vasudev Lal Principal Research Scientist of Cognitive AI at Intel Labs

CASE STUDY > Intel Labs Mitigates AI Bias in Foundational Multimodal Models by 20 Percent

intel.

Building Trustworthy LLMs for Evaluating & Generating Content at Scale

"This strategic collaboration with Intel allows Seekr to build foundation models **at the best price and performance** using a super-computer of 1,000s of the latest Intel Gaudi chips…"

CASE STUDY > Seekr, Intel® Gaudi® 2 and Intel® Tiber™ AI Cloud

seekr

Availability

Dell PowerEdge XE9680

Air-cooled
Dell AI Factory

Shipping Q1'25

Supermicro X14

Air-cooled
Equipped with Intel® Xeon® 6 processors

Shipping Q1'25

HPE Proliant Compute XD680

Air-cooled

Shipping Q1'25

Intel® Gaudi® 3 on IBM Cloud

Flexible consumption & user experience

VPC Virtual Servers

Red Hat Enterprise Linux AI servers
or
Accelerated Intel Gaudi 3 virtual servers for non-RHEL AI workloads

ROKS & IKS Clusters

OpenShift AI clusters
or
IKS or OpenShift Clusters with Intel® Gaudi® 3 accelerated workers

Deployable Architectures

Production ready, pre-configured RAG solution

watsonx

As SaaS with no exposure to underlying infra
or
As Software in private datacenter

IBM Cloud Data Center Locations for Intel® Gaudi® 3

Dallas (DAL)
Frankfurt (FRA)
Washington D.C.(WDC)

Select availability in US/EMEA early 2025

Regional expansion plans TBD

MORE INFO >

- Infographic
- Brief
- Video

DENVR
dataworks

DELIVERING CLOUD SERVICES with
SIMPLICITY, FLEXIBILITY, SAVINGS

Denvr Dataworks: brings choice and increased efficiency

with Intel® Gaudi® 2 and Intel® Gaudi® 3

Accelerate time-to-market, increase ROAI

Training
-as-a-Service

Inference
-as-a-Service

RAG
-as-a-Service

Model
-as-a-Service

Intel® Xeon® Processors

ENABLEMENT PACKAGES

- AI on Intel® Xeon® Partner Enablement Package

Report: CPUs are Key to Enterprise AI

Intel® Xeon® - The Processor Designed for AI

Efficiently run AI inference

5th Gen Intel® Xeon® processor

The flexibility of Intel® Xeon® with the built-in DL performance of an AI accelerator

- Up to **29% higher training** and up to **42% higher inference** performance than our previous generation[1]
- Up to **2.69x higher performance than AMD** EPYC 9654 (96C) and 9754 (128C) processors[2]

Build and deploy AI everywhere

Intel AI software suite of optimized open-source frameworks and tools

Enables out of the box AI performance and E2E productivity

- **5x improvement** on GPT-J in 10 weeks through software optimizations alone[3]
- Optimizing larger models up to **70B parameters** to meet customer SLAs
- Optimized 300+ DL models and 50+ ML and Graph Models

Open Ecosystem

Extensive Intel AI products and partnership

Accelerate end customer time to market

READ MORE

Product Brief

intel xeon

5th Gen Intel® Xeon® Processors

With faster memory and a larger last-level cache, 5th Gen Intel Xeon processors deliver up to 42 percent better AI performance than the previous generation and up to 14x better AI performance than 3rd Gen Intel Xeon processors.[1,2]

Trusted performance. Exceptional efficiency.

The digital universe is buzzing with data. From telemetry data collected at the edge to online transactions to cloud data stores, our world is brimming with information. But raw data alone, regardless of its volume, is not what propels us into the future. To truly innovate, we must convert this abundance of data into intelligence.

Intel enables this transformation and helps businesses solve some of the world's most complex problems with its software and hardware, accelerating time to actionable insights. Customers access the performance and power efficiency they need to optimally run diverse and demanding workloads, using one of the most extensive product portfolios on the market. Together with its broad ecosystem of partners, Intel enables businesses to meet every challenge head-on.

Now Intel's customers can harness the power of 5th Gen Intel Xeon processors to turn data into knowledge and drive innovation forward. These processors deliver impressive performance-per-watt gains across all workloads, plus outsized performance and lower total cost of ownership (TCO) for AI, databases, networking, storage, and high-performance computing (HPC).[3] They offer more compute, larger shared last-level cache, and faster memory at the same power envelope as the previous generation.[4] They are also software- and platform-compatible with the previous generation of 4th Gen Intel Xeon processors, so you can minimize testing and validation when deploying new systems for AI and other workloads.

The processor designed for AI

AI is transforming how we work. Businesses save money and time and create new opportunities using state-of-the-art AI models for language, text-to-image, and code generation. With AI acceleration in every core, 5th Gen Intel Xeon processors with built-in AI accelerators like Intel® Advanced Matrix Extensions (Intel® AMX), faster memory, and larger last-level cache are ready to handle demanding AI workloads with improved performance for deep learning (DL) inference and training.

Benefit from up to 14x higher performance on PyTorch for inference and training with 5th Gen Intel Xeon processors, compared to 3rd Gen Intel Xeon processors.[5] Realize up to 2.5x higher performance per watt with 5th Gen Intel Xeon processors with Intel AMX, compared to 4th Generation AMD EPYC processors.[6]

With sizable large-language models (LLMs) of up to 20 billion parameters, 5th Gen Intel Xeon processors meet the service-level agreements (SLAs) for delivering real-time user experiences with less than 100 ms second-token latency on LLMs.[7]

1. Based on performance gains of 1.1x to 1.29x for training (ResNet50v1.5, BERT-Large, SSD-ResNet34, RNN-T, MaskRCNN, and DLRM) and 1.19x to 1.42x for inference (ResNet50v1.5, BERT-Large, SSD-ResNet34, RNN-T (BF16 only), Resnext101 32x16d, MaskRCNN (BF16 only), DistilBERT) compared to 4th Gen Intel® Xeon® processor. See A15-A16 at intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.
2. Based on performance gains of 1.19x to 2.69x with Intel® Advanced Matrix Extensions (Intel® AMX) for inference on GPT-J, LLaMA-2 13B, DLRM, DistilBERT, BERT-Large, and ResNet50v1.5 compared to AMD EYPC 9654 and 9754. See A201, A202, A208-A211 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary..

Intel® AMX accelerates DEEP LEARNING use cases

Intel® Advanced Matrix Extensions (AMX)

BF16, INT8, and FP16 precision

- Recommender Systems
- Natural Language Processing
- Image Recognition Object Detection

Intel® Advanced Vector Extensions (AVX-512)

FP32 and FP64 precision

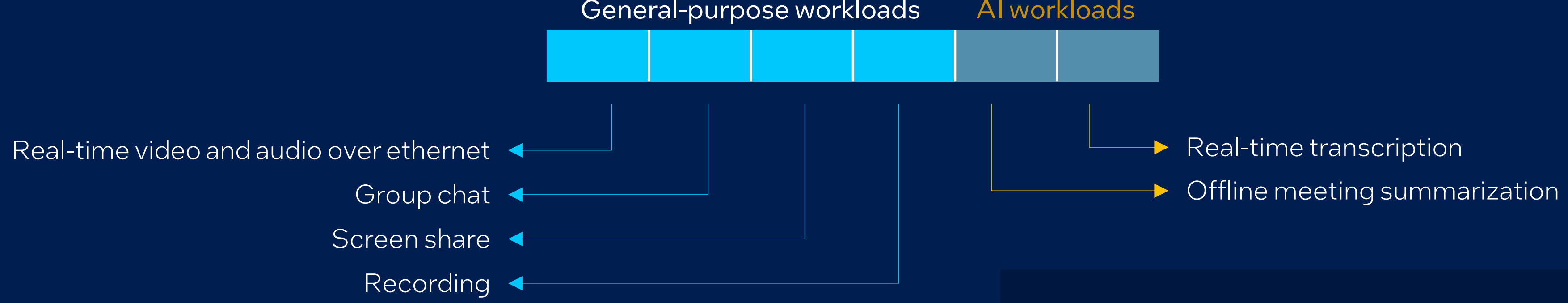
- Data Analytics
- Classical Machine Learning

Many DL workloads are "mixed precision" and
5th Gen Xeon® can seamlessly transition between AMX and AVX-512 as needed

Intel® Xeon® Processor delivers TCO Value for Mixed General-Purpose and AI Workloads

Case Study: Video Conferencing Service

Workloads



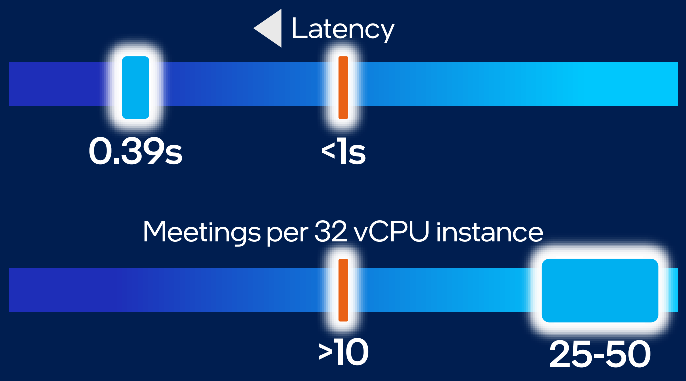
General-purpose workloads

- Real-time video and audio over ethernet
- Group chat
- Screen share
- Recording

AI workloads

- Real-time transcription
- Offline meeting summarization

Intel performance



Customer SLA

<1 sec latency

>10 Meetings per 32 vCPU instance

Intel engineers collaborating with the customer's technical team were able to beat the latency SLA at 0.39 seconds and were able to support anywhere between 25-50 meetings on a single 32 vCPU instance.

See more AI Customer Success Stories on Intel® Xeon®

5th Gen Intel® Xeon® Outperforms Competition Around The Clock

1.70x
on HammerDB MySQL OLTP

2.26x
on offline batched image classification inference

2.34x
on batched recommendation system inference

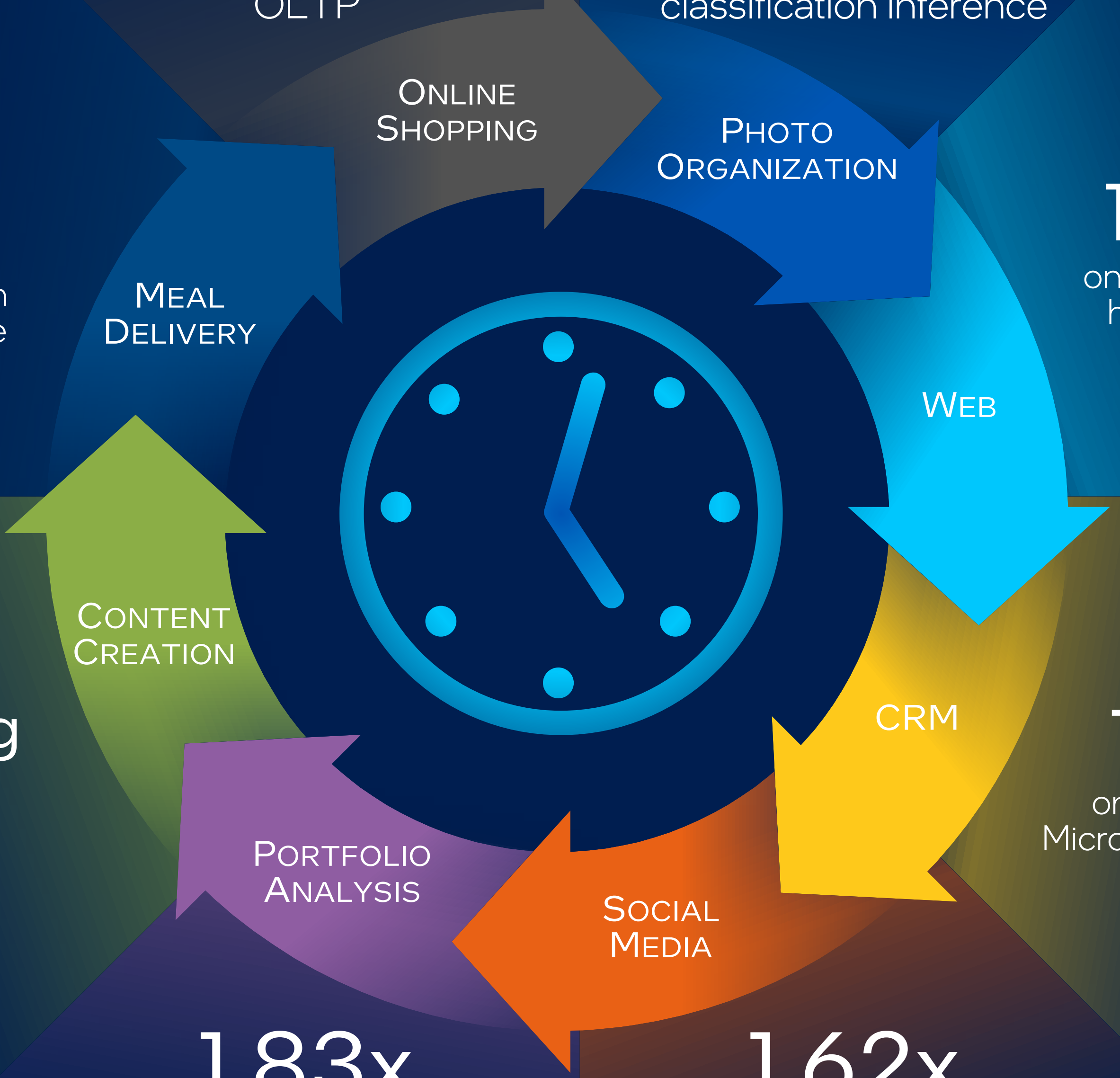
1.66x
on NGINX TLS handshakes

Delivering Gen AI
on Llama2 13B inferencing

1.93x
on HammerDB Microsoft SQL Server + Backup

1.83x
on Monte Carlo simulations

1.62x
on RocksDB



5th Gen Intel® Xeon® Benchmarks

Compares are relative to 4th Gen EPYC 9654 on AI;
relative to 4th Gen EPYC 9554 on all else
See backup for workloads and configurations. Results may vary.

intel ai

Get the Most Efficient Text Processing for Enterprise AI Workloads

Handle increased throughput and achieve faster request service for NLP and RAG with Intel® Xeon® processors

Natural Language Processing (NLP)

Intel® Xeon® 6 vs. AMD EPYC Turin

BERT-large

Up to **1.85x**

higher batch inference performance

with Intel Xeon 6980P compared with AMD EPYC 9755[19]

Intel Xeon 6 vs. AMD EPYC Genoa

BERT-large

Up to 2.69x higher batched NLP inference performance

5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. AMD EPYC 9654[20]

5th Gen Intel Xeon vs. previous gen

BERT-large

Up to 9.9x higher BERT-large real-time inference performance

5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd gen Intel Xeon processor with FP32[21]

Intel Xeon 6 vs. previous gen

BERT-large

Up to 1.88x higher BERT-large inference performance

Intel Xeon 6980P vs. prior generation[22]

Vector Search for Retrieval-Augmented Systems (RAG)

Intel® Xeon® 6 vs. AMD EPYC Turin

Wiki-45M dataset using Intel® Scalable Vector Search (SVS) technology

Up to **2.59x**

higher similarity search performance

Intel® Xeon® 6767P vs. AMD EPYC 9575F[23]

Intel Xeon 6 vs. previous gen

Redis VecSim

Up to 1.45x higher similarity search performance

Intel® Xeon® 6767P vs. Intel Xeon 8592+[24]

intel ai

AI Case Studies on Intel® Xeon® Processors

REAL WORLD RESULTS

Healthcare

Winning Health has introduced the WiNGPT solution based on 5th Gen Intel® Xeon® Scalable processors, through working with Intel, the **inference performance has been increased by over 3X** compared with the platform based on the 3rd Gen Intel® Xeon® Scalable processors

卫宁健康 WINNING HEALTH

READ ARTICLE

Energy

Storm Reply chose the new **Amazon EC2 C7i instances supported by 4th Gen Intel® Xeon® Scalable processors** and Intel libraries for LLM modeling. After a HW evaluation process, they **matched the price-performance ratio of GPU-based options by using CPU-based instances.**

READ ARTICLE

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 ARTICLE

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.

Introducing Intel® Xeon® 6 Processors

P-core

Optimized
for performance
in compute-intensive
and AI workloads

Common
platform foundation and
shared
software stack

E-core

Optimized
for efficiency
in high-density and
scale-out workloads

VISIT NOW > Intel® Xeon® 6 Website

Intel® Xeon® is the Most Deployed Host CPU for AI Accelerated Systems*

The role of a host CPU in maximizing the performance of an AI accelerated system:

- Data preprocessing to prepare data for training models
- Data transmission to GPU for parallel processing
- Manages check-pointing to system memory
- Inherent flexibility to process mixed workloads running on same infrastructure

Host CPU Technical Requirements	Intel® Xeon® 6 Winning Features
Superior I/O performance	20% more lanes with up to 192 PCIe 5.0 lanes resulting in higher i/o bandwidth[1]
High single threaded performance	High per core performance and max turbo frequency
High memory bandwidth and capacity	30% higMem capacity expansion with CXL her bandwidth w/ MRDIMMs[2]
RAS for large scale systems	Advanced RAS support (ex: PCIe enhanced Downstream Port Containment - eDPC)
Flexible form factor support	Both DC-MHS and NVIDIA® MGX™ supported

* Source: IDC Server Tracker report, based on Q1'24 system volume

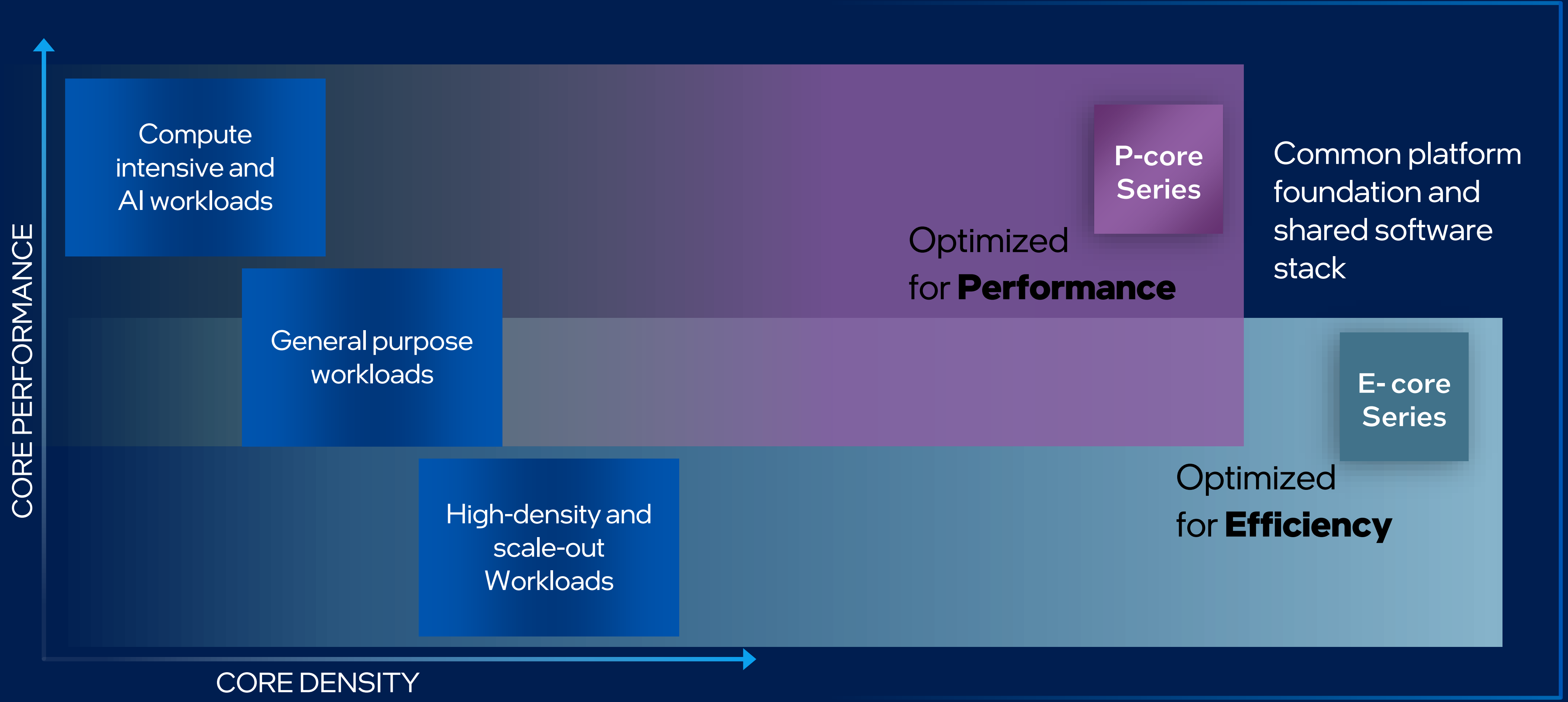
[1] compared to DDR5 6400 RDIMM

[2] 2S system as compared to 5th Gen Intel Xeon processor

* NVIDIA logo, MGX and HGX are trademarks of NVIDIA and/or its subsidiaries

Intel® Xeon® 6 Processors

The best processors to meet diverse performance and efficiency requirements


CORE PERFORMANCE
CORE DENSITY
Compute intensive and AI workloads
General purpose workloads
High-density and scale-out Workloads
P-core Series
Optimized for **Performance**
E- core Series
Optimized for **Efficiency**

Common platform foundation and shared software stack

LEARN MORE >

Product Brief

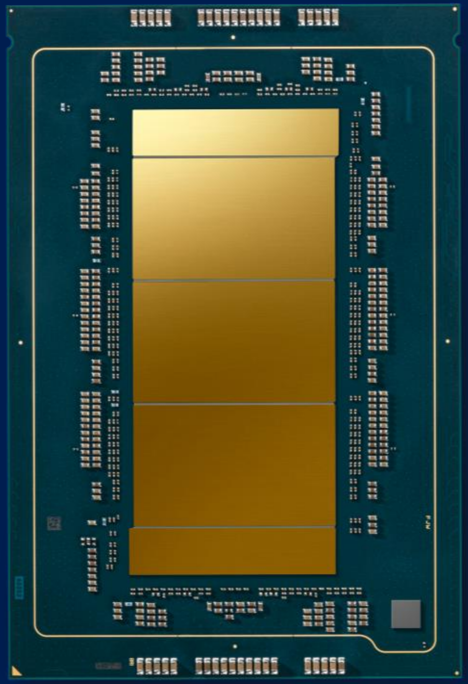
Infographic

Intel® Xeon® 6 processor with P-cores

AI | HPC | IaaS | General Compute

Intel® Xeon® 6 processors with P-cores

- Industry-leading Performance-cores (P-cores) are architected for compute-intensive workloads which benefit from multiple data elements being processed in parallel
- Choose from a range of SKUs with up to 128 cores and 12 memory channels for higher overall performance
- Maximize data throughput with the latest DDR5 and Multiplexed Combined Rank (MCR) DIMMs
- Scale AI everywhere with Intel Advanced Matrix Extensions (Intel AMX) to accelerate inferencing for INT8, BF16, and newly supported FP16 datatypes



2x

higher AI inference performance vs. 5th Gen Intel® Xeon® processors[1]

Up to

2.3x

higher HPC performance vs. 5th Gen Intel® Xeon® processors[1]

2x

higher average performance for general compute vs. 5th Gen Intel® Xeon® processors[1]

SOLUTION BRIEF >

- Intel® Xeon® 6 with P-cores for the Cloud

PRODUCT BRIEF >

- Intel® Xeon® 6 Processors with Performance-Cores (P-Cores) Deep Dive
- Intel® Xeon® 6700 with P-cores

[1] See [9G10, 9H10, 9A10] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

intel ai

Intel® Xeon® 6 with Performance Cores (P-cores)

Server Consolidation

Up to **17:1** server consolidation

Free up rack space and reduce data center footprint

Up to **94%** reduction in server count

Simplify and lower data center operations cost

Up to **77%** reduction in CO_2 emissions and power

Lower power bills and reach sustainability goals

Up to **87%** reduction in TCO

Lower overall cost to reinvest for growth

With more cores, double the memory bandwidth, and AI acceleration in every core, Intel® Xeon® 6 processors with P-cores provide **twice the performance** for the widest range of workloads, including AI and high-performance computing (HPC).[1]

Lower your total cost of ownership (TCO) by migrating from 2nd Gen Intel® Xeon® processors (4208) to Intel® Xeon® 6 processors with P-cores (6952P).[1]

Recover your cost in **4 months**[1]

[1] See Optimize Your Data Center with Intel Xeon for TCO Savings for full configuration details. Results may vary.

Intel® Xeon® 6 with Efficient-cores (E-cores)

Free up space and power in the data center for new AI projects

200 racks

2nd Gen Intel® Xeon® Processor

3:1

Rack consolidation

Over 4 years

80k MWh

Fleet energy saved

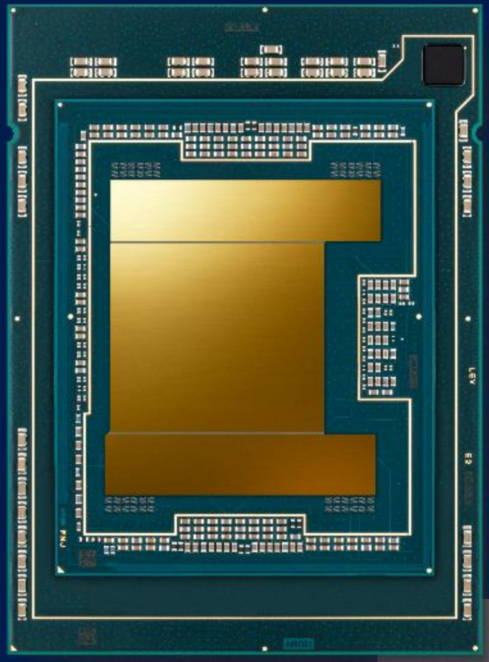
34k mt

Reduced CO2 emissions

66 racks

Intel® Xeon® 6700E

See [7T2] at intel.com/processorclaims: Intel® Xeon® 6.
Your costs and results may vary.

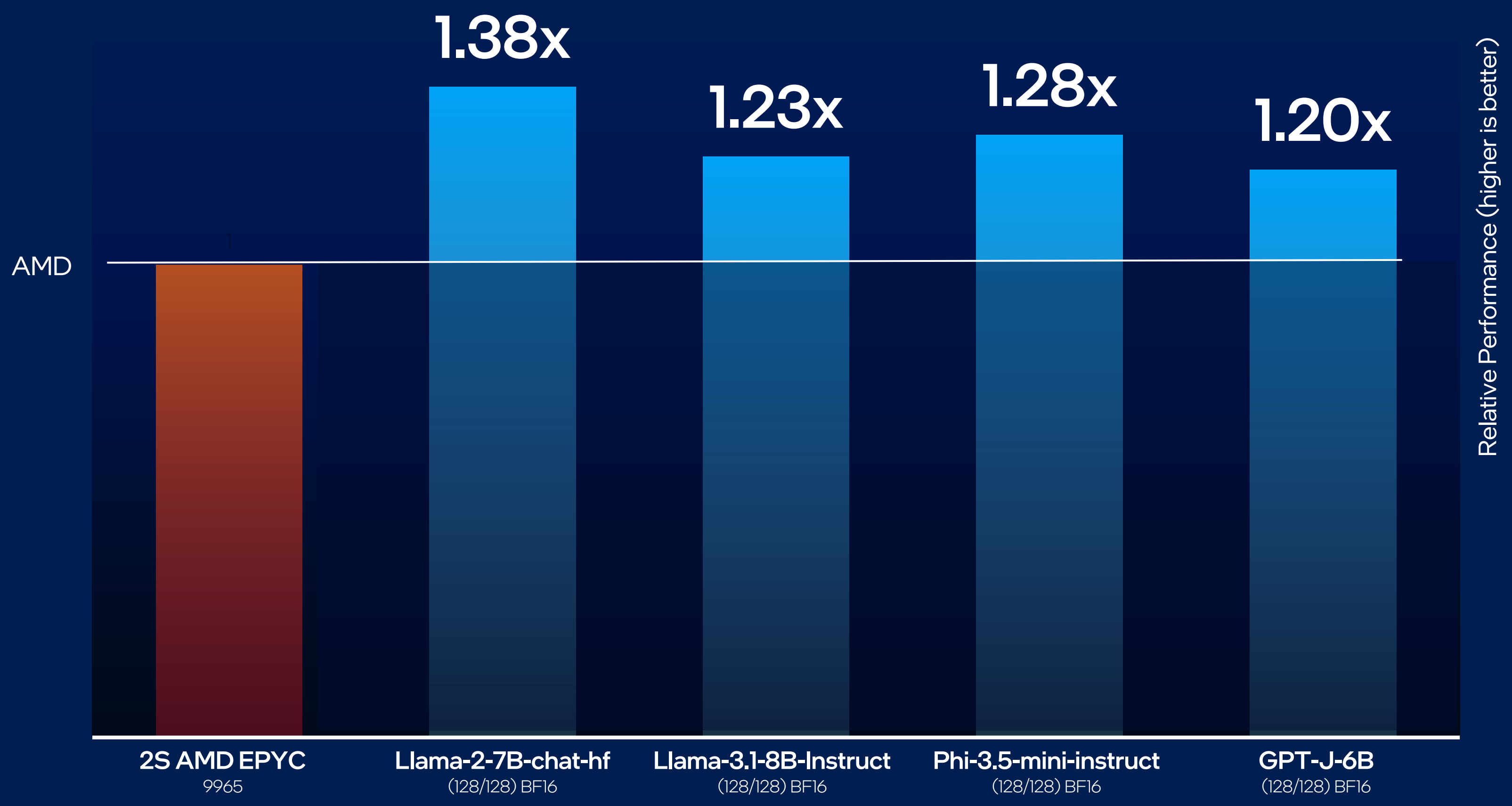


DEEP DIVE >

- Intel® Xeon® 6 Processors with Efficient-Cores (E-Cores)

Leadership in Small Model Performance

Intel® Xeon® 6900P series processors with MRDIMM support outperform the best AMD EPYC 9005



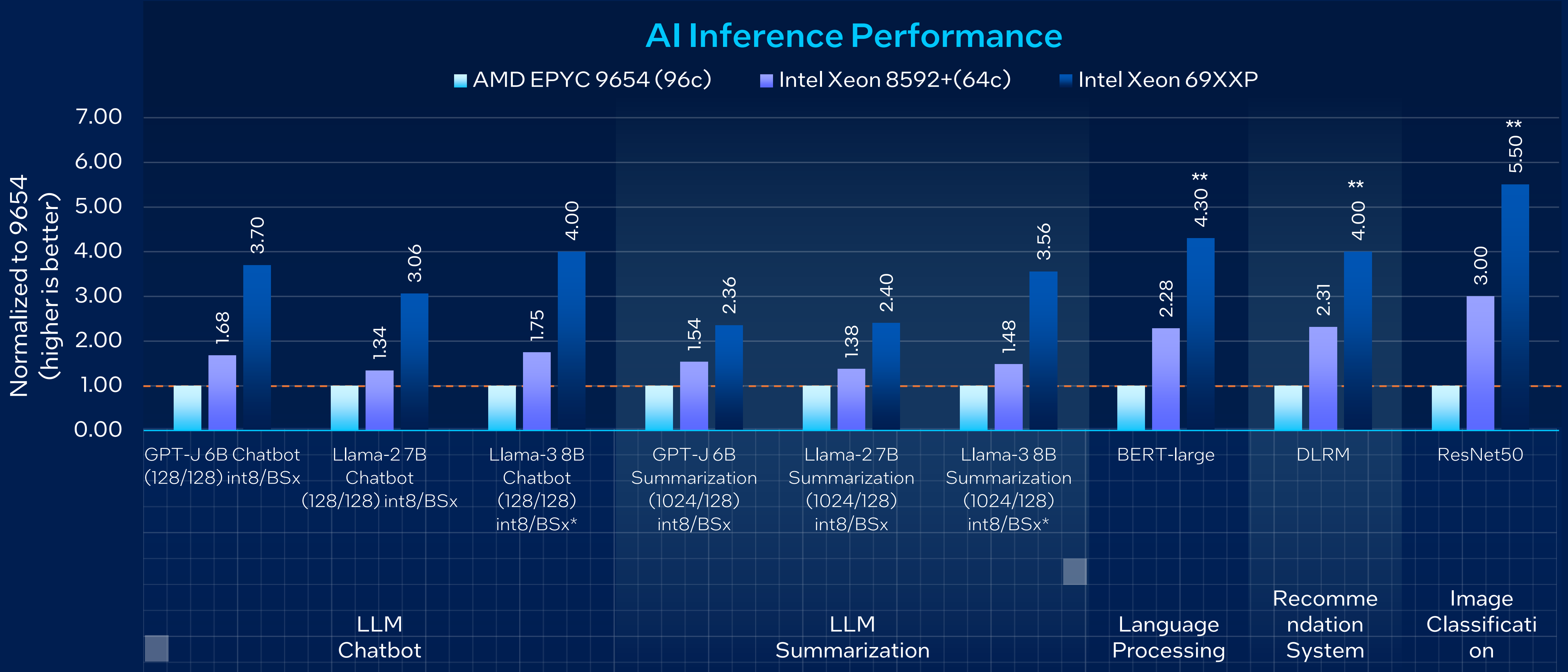
🟧 2S AMD EPYC 9965 (192 cores)

🟦 2S Intel Xeon 6980P (128 cores)

See [9A231] intel.com/processorclaims: Intel Xeon 6. Results may vary

World’s Best CPU for AI Inference

Continued leadership in AI with 5.5x higher performance vs. competition



**Intel Xeon result based on 128C 6980P. All other 69XXP results on 96C 6972P.

See backup for workload and configurations. Results may vary.
* AMD EPYC 9654 with BF16 on Llama3 8B Chatbot; did not meet SLA on INT8

Call to Action & Resources

Accelerate Enterprise AI Development with Intel® Tiber™ AI 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.

Read the Technical Article ›

Get Started ›

Call to Action

EDUCATION

Understand how Intel® technology can be used for Generative AI, and the scope upon which Intel® Xeon® and Intel® Gaudi® product lines can help you win more business

Get Started

ENGAGEMENT

Get started with

Intel® Tiber™ AI Cloud

Accelerate and scale AI with the latest hardware and software innovations in this development environment

CONTACT

Reach out to your **Intel Representative** for more information

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.

AI Enablement Zones

Access a comprehensive resource hub designed to help grow your business and solve your customers' most pressing business challenges. Find exclusive, value-added technical and sales enablement resources to help you build and sell solutions with Intel technology.

AI PC

Technical Enablement

Sales & Marketing Enablement

Edge AI

Technical Enablement

Sales & Marketing Enablement

GenAI

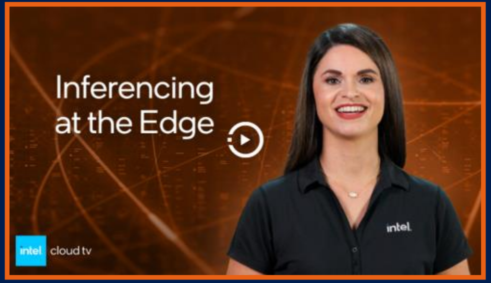
Technical Enablement

Sales & Marketing Enablement

Training Videos

Intel® Cloud TV

Your GenAI Opportunity with Intel® Gaudi® AI Accelerators



Gain Insights Using Data Inferencing at the Edge

Creating Competitive Advantage with AI in the Cloud

AI Inferencing Using Cloud Technologies

Selling the AI Experience

Get on the Fast Path to Scale AI Everywhere

Principles of AI Competencies

intel
partner
solution pro
Principles of
AI Everywhere

Principles of AI Everywhere Competency

In this curriculum, you'll delve into Deep Learning, Machine Learning, and Generative AI, and learn to navigate AI challenges using industry models tailored to data parameters. Learn how to assess customer needs effectively by applying the ADDS Methodology to offer tailored solutions from Intel's diverse portfolio, including CPU, GPUs, accelerators, technologies, software, and toolkits, for ease of AI solution deployments.

Enroll ›

intel
partner
solution pro
Intel® Gaudi®
AI Accelerators

Intel® Gaudi® AI Accelerators Competency

The Intel® Gaudi® AI Accelerator curriculum equips you with practical, in-depth knowledge about AI accelerators, including hardware, building clusters, software tools, cloud access, AI use cases, workloads, and ecosystem positioning. Learn how to boost performance, scale efficiently, and drive innovation with Intel® Gaudi® accelerators, designed to help you unlock powerful insights and deliver greater value to your customers.

Enroll ›

intel
partner
technical pro
Principles of AI
Software & Ecosystem

Principles of AI Software & Ecosystem Competency

From this curriculum, you will learn how to expedite AI development using open standards and harness data to drive business transformation. Explore a wide range of security solutions within the broad Intel AI ecosystem to ensure data integrity and protection. Delve into the breadth of Intel's AI-based products with a deep focus on Intel's AI software stack, toolkits, and Intel Developer Cloud for ease of AI solution deployments.

Enroll ›

Additional Trainings

Technical

Asset Type	Title and Link
Training Course	Improving LLMs with Prompt Economization and In-Context Learning
Training Course	Streamline AI for Data Generation and Large Language Models
Training Course	Applied Deep Learning with TensorFlow*
Training Course	Small and Nimble – the Fast Path to Enterprise GenAI
Training Course	The Next Wave of GenAI - Domain-Specific LLMs
Guide	A Developer's Guide to Getting Started with Generative AI: A Use-Case-Specific Approach
Training Course	Taking AI on Intel® Xeon® Processors Into the Solution Space
Guide	A Developer's Guide to Adapting to Enterprise AI
Training Course	Streamline AI for Data Generation and Large Language Models
Training Course	LLMs and RAG in GenAI
Training Course	Stable Diffusion and Hugging Face in GenAI

Additional Trainings

Non-Technical

Asset Type	Title and Link
Video Series	Embracing Generative AI
Training Course	Small and Nimble – the Fast Path to Enterprise GenAI
Training Course	The Next Wave of GenAI - Domain-Specific LLMs
Training Course	Principles of AI Everywhere Competency
Training Course	Principles of AI Software & Ecosystem Competency
Training Course	Engaging the AI Ecosystem: Win with Software, Scale with SIs and Sell the Solution
Training Course	Generative AI and Large Language Models for the Real World
Training Course	Foundations of GenAI

Additional Resources

Asset Type	Title and Link
Webinar	Generative AI Webinar Series
Webinar	Bringing GenAI Everywhere
Podcast	How Copilot, ChatGPT, Stable Diffusion and Generative AI Will Change How We Develop, Work and Live
Business Brief	Deploy AI Everywhere
Blog Series	Tuning and Inference for Generative AI with 4th Generation Intel Xeon Processors
Solution Brief	Deploy and Scale Generative AI Inference with Lenovo ThinkSystem SR650 V3 / 4th Gen Intel Xeon Processors
Solution Brief	New Intel and VMware Technologies Turbocharge Lenovo ThinkAgile VX V3 Systems
Tech Article	Accelerate Llama 2 with Intel® AI Hardware and Software Optimizations
Research PR	10% of Organizations Surveyed Launched GenAI Solutions to Production in 2023
Fireside Chat Video	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
Supermicro Benchmarks	Habana Claims Validation
Hugging Face Benchmarks	Benchmarks
Training / Webinar	Cloud Solution Architect (CSA) Tech Talk: AI with Habana
White Paper	Enterprise AI is all about the Developer
Infographic	CPUs are Key to Enterprise AI

Additional Resources

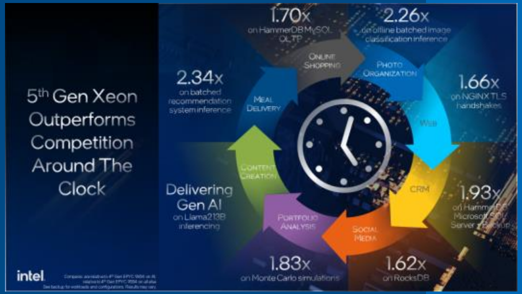
Asset Type	Title and Link
Solution Brief	Streamline AI Adoption and Deployment Using Intel Enterprise AI with Red Hard OpenShift AI
Guide	The AI Guide
Reference Kit	AI Unstructured Text Data Generation
White Paper	Zoho is Optimizing and Accelerating Video AI Workloads
White Paper	Seekr Develops Trustworthy AI Screening System
Solution Brief	Security in Education: AI and Confidential Computing Help Make Secure Remote Exams a Reality
Case Study & Video	Nature Fresh Farms Utilizes AI from Seed to Store
Case Study	QMed Asia Drives Early-Stage Cancer Detection Rate
Case Study & Video	MetaApp Revamps AI-Based Recommendation System
Solution Brief	Optimizing AI Model Training and Refinement for Automated Optical Inspection (AOI)
Blog	Prompt-Driven Efficiencies for LLMs
Solution Brief	Driving Enterprise RAG Innovation with Intel® Xeon® Processors
White Paper	Improving Intel Technical Sellers' Effectiveness and Customer Engagement with Help of a Generative AI Chatbot

Notices and Disclaimers

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

intel ai

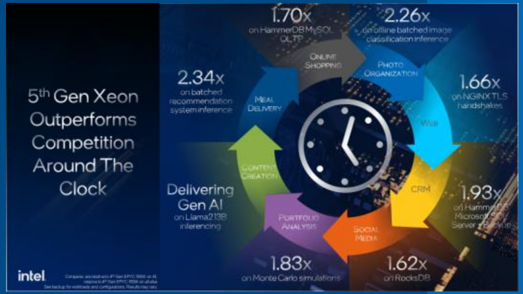
Resources and Configurations


5th Gen Xeon Outperforms Competition Around The Clock
1.70x on HammerDB MySQL OLTP
2.26x on offline batched image classification inference
2.34x on batched recommendation system inference
1.66x on NGINX TLS handshakes
Delivering Gen AI on Llama213B inferencing
1.93x on HammerDB Microsoft SQL Server
1.83x on Monte Carlo simulations
1.62x on RocksDB
Online Shopping
Photo Organization
Meal Delivery
Web
Content Creation
CRM
Portfolio Analysis
Social Media
intel
See backup for workloads and configurations. Results may vary.

- ## 5th Gen Intel® Xeon® Outperforms Competition Around The Clock

- ResNet50v1.5
- Intel® Xeon® 8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. TensorFlow= Intel TF 2.13, OneDNN=3.2, Python 3.8, AI Model=ResNet50v1.5 Large(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BFloat16, INT8-AMX (BS >1),, Test by INTEL as of 10/10/2023.
- AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, ZenDNN 4.1 TensorFlow 2.12.1, Python 3.8, AI Model=ResNet50v1.5 Large(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using (BS >1),Test by INTEL as of 09/11/23.
- NGINX TLS
- Intel® Xeon® 8592+: 1-node, 2x 5th Gen Intel® Xeon® Scalable processor (64 core) with integrated Intel Quick Assist Technology (Intel QAT), QAT device utilized=4 (1 active socket), HT On, Turbo Off, SNC On, with 1024GB DDR5 memory (16x64 GB 5600), microcode 0x21000161, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, IPP Crypto 2021.8, IPsec MB v 1.4, QAT _Engine v 1.4.0, QAT Driver 20.l.1.1..20-00030, TLS 1.3 Webserver: ECDHE-X25519-RSA2K, tested by Intel October 2023.
- AMD EPYC 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost Off, NPS1, Total Memory 1536GB (24x64GB DDR5 4800), microcode 0xa10113e, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, TLS 1.3 Webserver: ECDHE-X25519-RSA2K,tested by Intel October 2023.

Resources and Configurations


5th Gen Xeon Outperforms Competition Around The Clock
1.70x on HammerDB MySQL OLTP
2.26x on offline batched image classification inference
2.34x on batched recommendation system inference
1.66x on NGINX TLS handshakes
Online Shopping
Photo Organization
Meal Delivery
Web
Content Creation
CRM
Portfolio Analysis
Social Media
Delivering Gen AI on Llama213B inferencing
1.93x on HammerDB Microsoft SQL Server + Backup
1.83x on Monte Carlo simulations
1.62x on RocksDB
intel

▪ 5th Gen Intel® Xeon® Outperforms Competition Around The Clock

- HammerDB Microsoft SQL Server + Backup
- Intel® Xeon® 8592+: 1-node, 2x 5th Gen Intel® Xeon® Scalable processor 8592+ (64 cores) with integrated Intel Quick Assist Technology (Intel QAT), Number of IAA device utilized=8(2 sockets active), HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 5600), microcode 0x21000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 7x 3.5T INTEL SSDPE2KE032T807, QATZip 2.0.W.1.9.0-0008, Microsoft Windows Server Datacenter 2022, Microsoft SQL Server 2022, SQL Server Management Studio 19.0.1, HammerDB 4.5, tested by Intel October 2023.
- AMD EPYC 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost On, NPS1, Total Memory 1536GB (24x64GB DDR5 4800), microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 7x 3.5T INTEL SSDPE2KE032T807, Microsoft Windows Server Datacenter 2022, Microsoft SQL Server 2022, SQL Server Management Studio 19.0.1, HammerDB 4.5, tested by Intel October 2023.
- RocksDB
- Intel® Xeon® 8592+: 1-node, 2x 5th Gen Intel® Xeon® Scalable processor 8592+ (64 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 5600), microcode 0x21000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, QPL v1.2.0, accel-config-v4.0, iaa_compressor plugin v0.3.0, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 64 RocksDB instances, tested by Intel October 2023.
- AMD EPYC 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost On, NPS1, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 28 RocksDB instances, tested by Intel October 2023.
- Monte Carlo
- Intel® Xeon® 8592+: 1-node 2x Intel® Xeon® 8592+, HT On, Turbo On, SNC2, 1024 GB DDR5-5600, ucode 0x21000161, Red Hat Enterprise Linux 8.7, 4.18.0-425.10.1.el8_7.x86_64, Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of October 2023.
- AMD EPYC 9554: 1-node, 2x AMD EPYC 9554, SMT On, Turbo On, CTDP=360W, NPS=4, 1536GB DDR5-4800, ucode= 0xa101111, Red Hat Enterprise Linux 8.7, Kernel 4.18,Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of March 2023

Resources and Configurations

5th Gen Xeon Outperforms Competition Around The Clock

5th Gen Intel® Xeon® Outperforms Competition Around The Clock

- DLRM
- Intel® Xeon® 8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. Framework=Pytorch 2.1, IPEX=2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BS>1, Precision=INT8-AMX, Test by INTEL as of 10/10/2023.
- AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.1, IPEX=2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BS>1, Precision=INT8. Test by INTEL as of 09/11/23.
- HammerDB MySQL
- Intel® Xeon® 8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 8 [0], DSA 8 [0], IAX 8 [0], QAT 8 [0], Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.0-84-generic, HammerDB Mv4.4, MySQL 8.0.33. Test by Intel as of 10/04/23.
- AMD EPYC 9554: 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 0 [0], DSA 0 [0], IAX 0 [0], QAT 0 [0], Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, HammerDB v4.4, MySQL 8.0.33. Test by Intel as of 10/05/23.