intel GAUDI

Enterprise AI with Intel® Gaudi® AI Accelerators

Partner Enablement Package

Bringing Choice to Gen AI with Performance, Scalability and Efficiency

intel.

Contents

Why Partner with Intel?

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

But we aren't doing this alone!

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

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

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

Helping you to grow your business.

Join Us On the Journey to Bring AI Everywhere

GenAI Market Opportunity

Generative AI is poised to be a **$1.3 trillion** market by 2032 and could expand to **10-12%** of total IT expenditure[1]

WATCH NOW >

Your GenAI Opportunity with Intel® Gaudi® AI Accelerators

Rising demand for generative AI products could add about **$280 billion** of new software revenue[1]

[1]Bloomberg: Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds

Challenges with AI Compute Solutions

Need more Choice

other than single-source GPUs

Locked-in

with proprietary software and networking

Ability to Scale

while containing costs of infrastructure

Maximize efficiency

yet still solves business challenges

The Need for a Better Approach

Unlocking the power of GenAI with LLMs, RAG, and Multimodal models

As models grow in size and complexity, the need for hardware designed specifically for AI workloads has never been more critical.

It is crucial that organizations avoid vendor lock-in, maintaining the flexibility to adapt to changing needs and innovations without being tied to a single proprietary solution.

READ THE REPORT

AI and ROI – Systems that offer:

- cost-effective scaling
- quick model convergence
- minimized energy consumption
- rapid availability

...can unlock AI's full potential for enterprises, driving tangible business outcomes while keeping both CAPEX and OPEX in check

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

AI PC

Broadest AI SW ecosystem

AI PC
Light inference

Edge AI

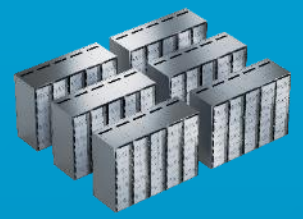
Flexible, edge node reference architectures

Node
Fine-tuning, inference

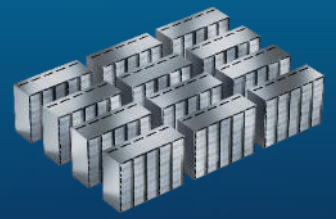
Cluster
Light training, tuning, peak inf.

Data Center & Cloud AI

Open, scalable systems & reference architecture



Super Cluster
Training, tuning, peak inf.



Mega Cluster
Large-scale training & inference

ACCESS NOW >

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

Scalable Systems for AI

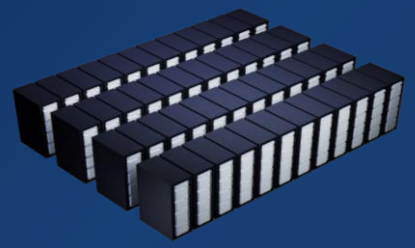
For dedicated AI deployments, Intel® Xeon® processors paired with Intel® Gaudi® accelerators will deliver optimal TCO

Training and Fine-Tuning

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

Inference and Deployment

Cloud Data Center | Edge | Client

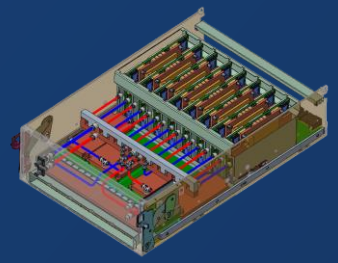


Cluster and Data Center Scale

intel. GAUDI

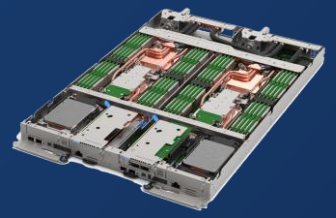
Multi-node Deployment per Rack

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



Multi-GPU or Multi-socket CPU

intel. XEON | intel. XEON MAX SERIES | intel. DATA CENTER GPU FLEX SERIES



Single-Socket CPU or Single GPU

intel. XEON | intel. DATA CENTER GPU FLEX SERIES | intel CORE ULTRA

Client CPU

intel CORE ULTRA | intel. ARC GRAPHICS

intel. ETHERNET

ACCESS NOW > AI Workloads on Intel® Xeon® Processors Partner Enablement Package

Intel® AI Software is Enterprise Ready

Data Acquisition, Preprocessing → Feature Engineering → Model Training, Inference

Data Analytics at Scale

SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, OpenVINO, scikit learn, HOROVOD, LightGBM, XGBoost, deepspeed

with Intel Optimizations (ITEX and IPEX)

oneAPI: oneDNN, oneMKL, oneDAL, Intel OpenMP, oneCCL, Intel® MPI

Model Development

- Kubeflow Pipelines
- Training Operator (MPI)
- Jupyter Notebooks

Model Serving

- OpenVINO Model Server
- TensorFlow Serving

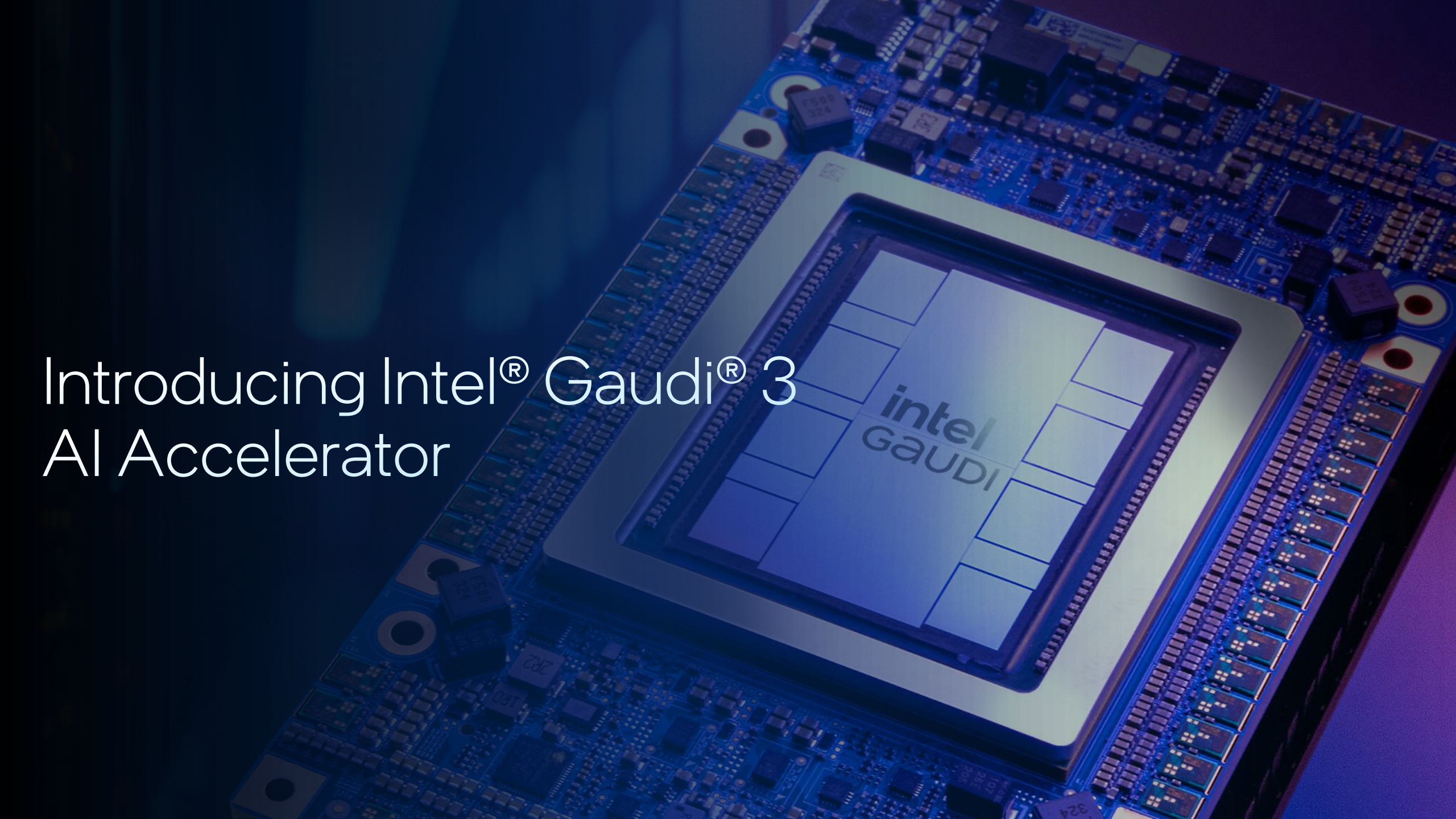
SW Infrastructure

- Kubernetes, Red Hat OpenShift
- Container plug-ins – runtime, storage, network
- VMWare vSphere (vSphere Networking + AVI)

HW Infrastructure

- Intel CPUs with built-in AI accelerators
- Intel GPUs
- Intel® Ethernet

Introducing Intel® Gaudi® 3 AI Accelerator



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

intel.
GAUDI

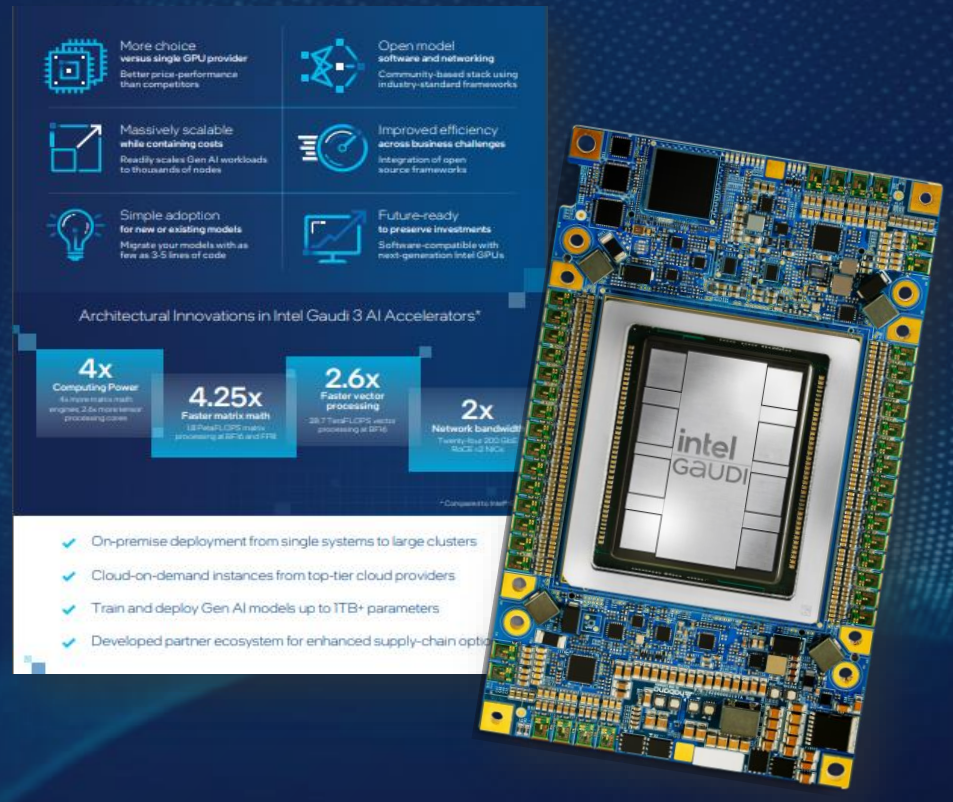
Outstanding Results vs Nvidia H100[1]

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

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

Empowering AI with Intel® Gaudi® 3: The Software Edge

Seamless Integration

Optimized Graph Compiler

Cutting-Edge Kernel Fusion

Scalable, Open Ecosystem

FP8 Quantization

Energy-Efficient Performance

Intel® Gaudi® Software Stack

Enable DeepSpeed on Intel Gaudi Processors

- DeepSpeed Integration
- LLM serving Integration
- Quantization Integration
- PyTorch Integration
- Graph Compiler
  - Customer custom TPC kernels
  - Optimized TPC kernels library
  - Matrix ops library
- Collective Communication Library
- User-mode driver / run-time environment
- Compute Driver
- Network Driver

Intel® Gaudi® 3 AI Accelerator

Legend

- Proprietary
- Ecosystem Integration
- Plugin

OPEA: Open Ecosystems Reduce Barriers to Enterprise AI Production Software

Open Platform for Enterprise AI

Simplify development, production, & adoption of Enterprise GenAI apps

Framework of composable building blocks for Generative AI solutions

For example: LLMs, data stores, and prompt engines

Blueprints of end-to-end workflows

For example: RAG applications, Agentic systems, AI avatar

CONSTRUCTION

Assessment for **grading Gen AI systems**

For example: performance, features, trustworthiness and enterprise-grade readiness

EVALUATION

WATCH NOW > Elevate Your AI Expertise with Intel® Gaudi® Accelerators

OPEA: Partners

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
prediction Guard
qdrant
Redis
Red Hat
SAP
Rivos
sas
TURINTECH
vectara
vmware by Broadcom
wipro
Yellowbrick
zensar
zilliz
ZTE

Partners as of January 2025

OPEA by the numbers

695	332	35K	600+	51
Github Followers	Stars on GenAI Examples Repo	Users to OPEA.dev	Registered for OPEA virtual events	Partners and growing...

See backup for the latest OPEA news

Ecosystem Momentum

Thad Omura,
Chief Business Officer

"Our strategic collaboration with Intel is essential as we deliver high-performance and reliable PCIe and CXL solutions for our hyperscale and OEM customers driving large AI infrastructure deployments. The combination of our Aries PCIe/CXL Smart DSP Retimers and Leo CXL Smart Memory Controllers with Intel® Gaudi® 3 AI accelerators and Intel® Xeon® 6 processors provides a robust solution for scaling modern AI and cloud workloads."

Arun Narayanan,
SVP, Servers and Networking

"As organizations look to power their advanced compute needs, Dell Technologies and Intel are delivering the technology to perform complex tasks and innovate using AI with the Dell PowerEdge XE9680 server with Intel® Gaudi® 3 AI accelerators."

Trish Damkroger,
Senior Vice President & General Manager, HPC & AI Infrastructure Solutions

"For over 30 years, Hewlett Packard Enterprise and Intel have worked closely to codevelop and deliver breakthrough innovation, from the edge to exascale. We look forward to continuing our partnership by supporting Intel® Gaudi® 3 on future HPE systems for significant performance to accelerate scientific discovery and innovation."

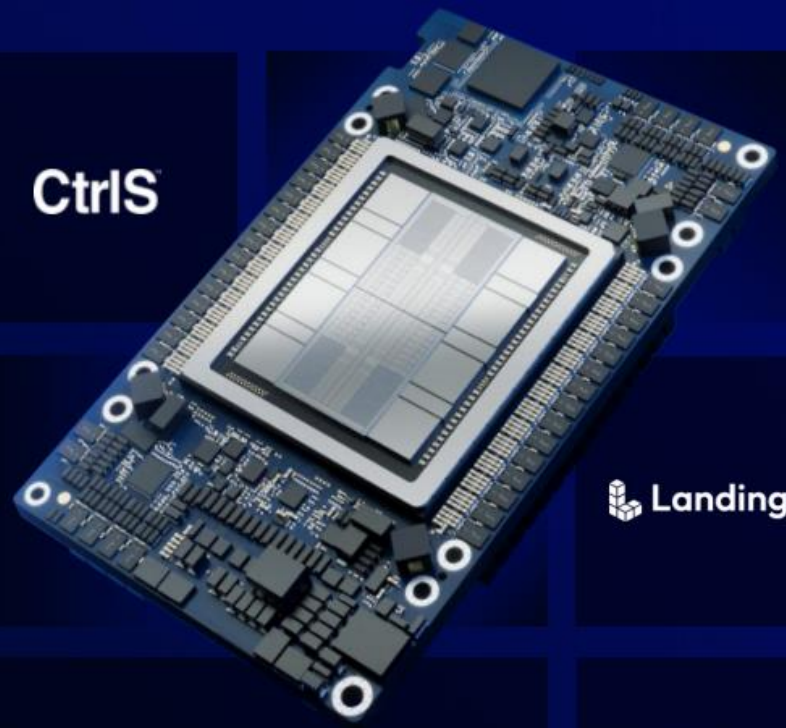
Steven Huels, Vice President and General Manager, AI Engineering

"We are pleased to collaborate with Intel to deliver end-to-end AI solutions, based on Intel® Gaudi® 3 AI accelerators and backed by Red Hat OpenShift AI and Red Hat Enterprise Linux AI, to help organizations accelerate their AI roadmaps."

Ray Pang, Supermicro SVP, Technology and Business Enablement

"Building on the successful deployment of the world's largest Intel® Gaudi® 1 and Gaudi® 2 clusters, Supermicro is now pleased to offer the industry's first and only Intel® Xeon® 6 based Gaudi 3 system powered by the Xeon 6900 series with P-cores."

Growing Customer Momentum



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

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™

OEM General Availability

DELL Technologies

Dell PowerEdge XE9680

Air-cooled
Dell AI Factory

Shipping Q1'25

SUPERMICRO

Supermicro X14

Air-cooled
Equipped with Intel® Xeon® 6 processors

Shipping Q1'25

Hewlett Packard Enterprise

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

Get Started with Intel® Tiber™ AI Cloud

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

intel® tiber™ AI Cloud

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.

intel. GAUDI

AVAILABLE NOW

Select availability for Intel® Gaudi® 3 AI Accelerators

Sign up today

Summary

- Understand your partner / customers' usage model and performance needs FIRST
- When AI is just another workload in a mixed general purpose and AI environment, lead with the Intel® Xeon® processors that are already running your customers' business
- For dedicated AI deployments, Intel® Xeon® processors paired with Intel® Gaudi® accelerators will deliver the optimal TCO
- When deploying Intel® Gaudi®, refresh older Intel® Xeon® processors to free up power and space then add Intel® Gaudi® AI Accelerators for deep learning AI training & inference

Call to Action

1 Get Started with Intel® Tiber™ AI Cloud

intel. tiber™ AI Cloud

2 Deploy Intel® Gaudi® 3 AI Accelerators via OEM Designs

Intel is working with OEM partners to bring Intel® Gaudi® 3 AI accelerators to on-prem deployments. For more information on purchasing, please reach out to your OEM partner or Intel representative.

DELL Technologies · Hewlett Packard Enterprise · SUPERMICRO · ASUS · GIGABYTE · Inventec · QCT · wistron · wiwynn

3 Experience Intel® Gaudi® 3 AI Accelerators in the Cloud

IBM Cloud
Coming early 2025

DENVR dataworks
Learn more

Developer Resources

Create, Migrate, and Optimize Your AI Models with Intel® Gaudi® AI Accelerators

Discover the resources, guidance, tools, and support needed to more easily and flexibly build new AI models, migrate existing ones, and optimize model performance to meet your requirements. Access the latest Intel® Gaudi® software to build or update your infrastructure.

Get Access

Connect to Intel® Tiber™ AI Cloud for Intel® Gaudi® 2 AI accelerators or Amazon EC2* DL1 for first-gen Intel Gaudi accelerators.

Get Started

Find detailed instructions and videos to get started with GPU migration, working with Hugging Face models, and new customer onboarding.

Tutorials

Step-by-step tutorials that walk you through creating and training your models.

Model Optimization & Debugging

Optimize, fine-tune, debug, and profile your model to meet your performance targets.

Performance Data

Review training and inference model performance data on the Intel Gaudi AI accelerator.

Documentation

Access the most recent documentation or repositories on GitHub*.

Additional Resources
Intel® Gaudi® 3 AI Accelerator 32-Node Cluster Reference Design White Paper
Intel® Gaudi® 3 AI Accelerator 325-L OAM Mezzanine Card Product Brief
Intel® Gaudi® 3 AI Accelerator HL-338 PCIe Add-In Card Product Brief
Intel® Gaudi® 3 AI Accelerator HLB-325 Baseboard Product Brief

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.

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

Sign up to Intel® Partner Alliance for full access or select one of the Enablement Zones if you are already a member

Training – Intel® Partner University

intel. partner solution pro
Intel® Gaudi® AI Accelerators

Intel® Gaudi® AI Accelerators Competency

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.

intel. partner solution pro
Principles of AI Everywhere

Principles of AI Everywhere Competency

Delve into Deep Learning, Machine Learning, and Generative AI, and learn to navigate AI challenges using industry models tailored to data parameters.

intel. partner technical pro
Principles of AI Software & Ecosystem

Principles of AI Software & Ecosystem Competency

Learn how to expedite AI development using open standards and harness data to drive business transformation.

Additional Training

Stable Diffusion and Hugging Face in GenAI	https://partneruniversity.intel.com/learn/courses/17689/url

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®

OPEA: Today

- LlamaIndex and LangChain integration that enables OPEA as a backend
- The OPEA project has reached over 50 partners!
- OPEA is now available on the **AWS marketplace**, part of our goal to reach developers where they are
- **Amazon** has contributed Opensearch with **Bedrock** (managed LLM service) integration due with the

OPEA 1.3 release (managed LLM service)

- **Infosys** has been a key contributor to OPEA 1.2 with two key contributions including;
  - Azure automated deployment for OPEA applications
  - Elasticsearch vector database integration
- **OPEA awareness and adoption is growing** end users are looking to o replace Azure OpenAI service citing

TCO and data confidentiality as the primary reasons

- **AMD** has continued their strong collaboration with the project with several contributions **validating more**

GenAI examples on ROCm hardware

- Dell and H3C have plans in place to create appliances that are '**Powered by OPEA**'