intel GAUDI

Enterprise AI with Intel® Gaudi® AI Accelerators

Partner Enablement Package

Bringing Choice to Gen AI with Performance, Scalability and Efficiency

intel.

Contents

> Generative AI Opportunity

- Gen AI Market
- Challenges with AI Compute Solutions
- The Need for a Better Approach

> Intel® Gaudi® 3 Product Line

- Intel® Gaudi® 3 AI Accelerator & Benefits
- How Intel® Gaudi® 3 AI Accelerator Addresses Enterprise AI Challenges
- The Software Edge
- Open Platform for Enterprise AI (OPEA)
- Intel® Gaudi® 3 PCIe Card
- Intel® Gaudi® 3 PCIe Card Solving Enterprise AI Deployment Challenges

> Ecosystem Adoption

- Ecosystem Momentum
- Case Studies

> Availability

- OEMs
- IBM
- Denvr Dataworks

> Call to Action & Resources

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

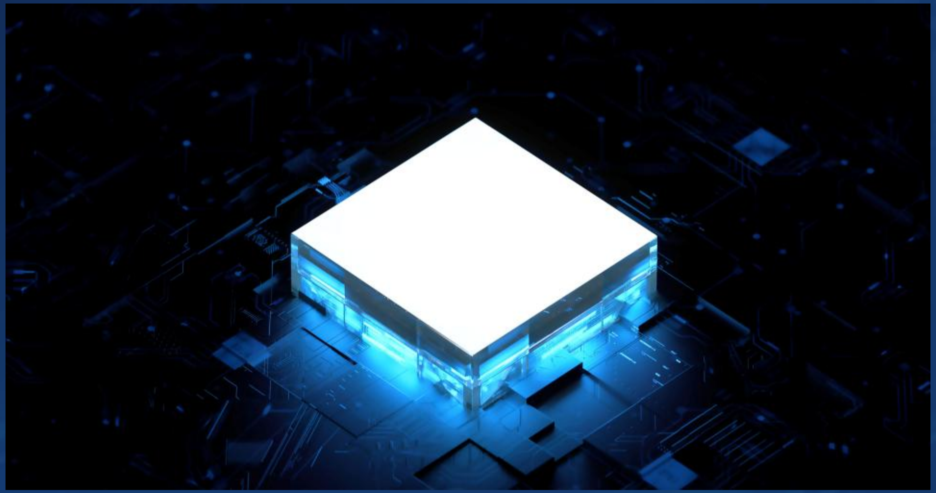
Our broad portfolio of AI-enabling technologies and collaboration with hardware, software, and solution ecosystem partners delivers real world solutions and differentiated business outcomes for industries, companies, and communities.

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]



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.

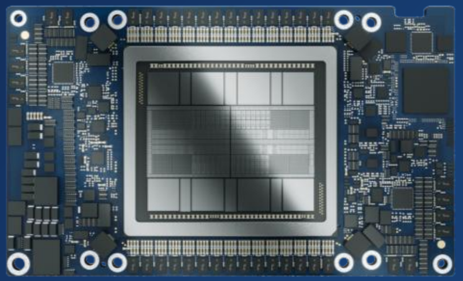
It is crucial that organizations avoid vendor lock-in, maintaining the flexibility to adapt to changing needs and innovations without being tied to a single proprietary solution.

AI and ROI – Systems that offer:

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

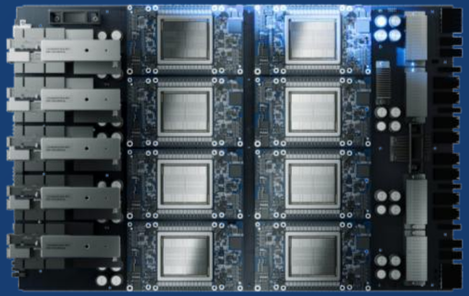
...can unlock AI's full potential for enterprises, driving tangible business outcomes while keeping both CAPEX and OPEX in check

intel GAUDI Product Line



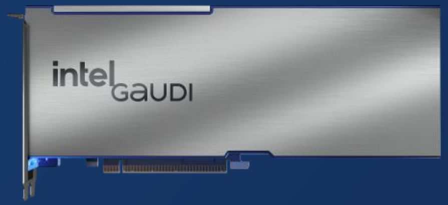
Accelerator Card

HL-325L OAM-Compliant



Universal Baseboard

HLB-325



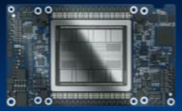
PCIe CEM

HL-338 Add-In Card

Enabling customer infrastructure choice

Scalable AI, Tailored for Your Infrastructure

One architecture. Two deployment paths with open software ecosystem. Optimized to meet your performance, power, and budget needs-from across industries.

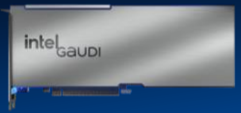


Intel® Gaudi® 3 OAM

Engineered for large scale AI models for hyperscalers and enterprise clusters

- Ideal for running large LLM inferencing in parallel and at scale.
- With strong compute power and networking infrastructure it is ideal for AI research labs, hyperscalers, and cloud buildouts
- Designed to handle high-performance, high-bandwidth demands, 6U–8U chassis provide the infrastructure required for full-speed AI acceleration.

✓ Designed for large models requiring maximum throughput, networking and concurrent users



Intel® Gaudi® 3 PCIe

Right-sized for enterprise AI applications and appliances

- Perfect for GenAI inference, secure on-prem workloads, and cost-sensitive AI deployments.
- With its low power consumption and PCIe form factor, Intel Gaudi 3 PCIe is an ideal fit for enterprises, academic institutions, government agencies, and retail branches looking to scale AI without scaling costs.
- Fits seamlessly into 2U–4U servers, enabling more efficient use of rack space and reducing operational overhead.

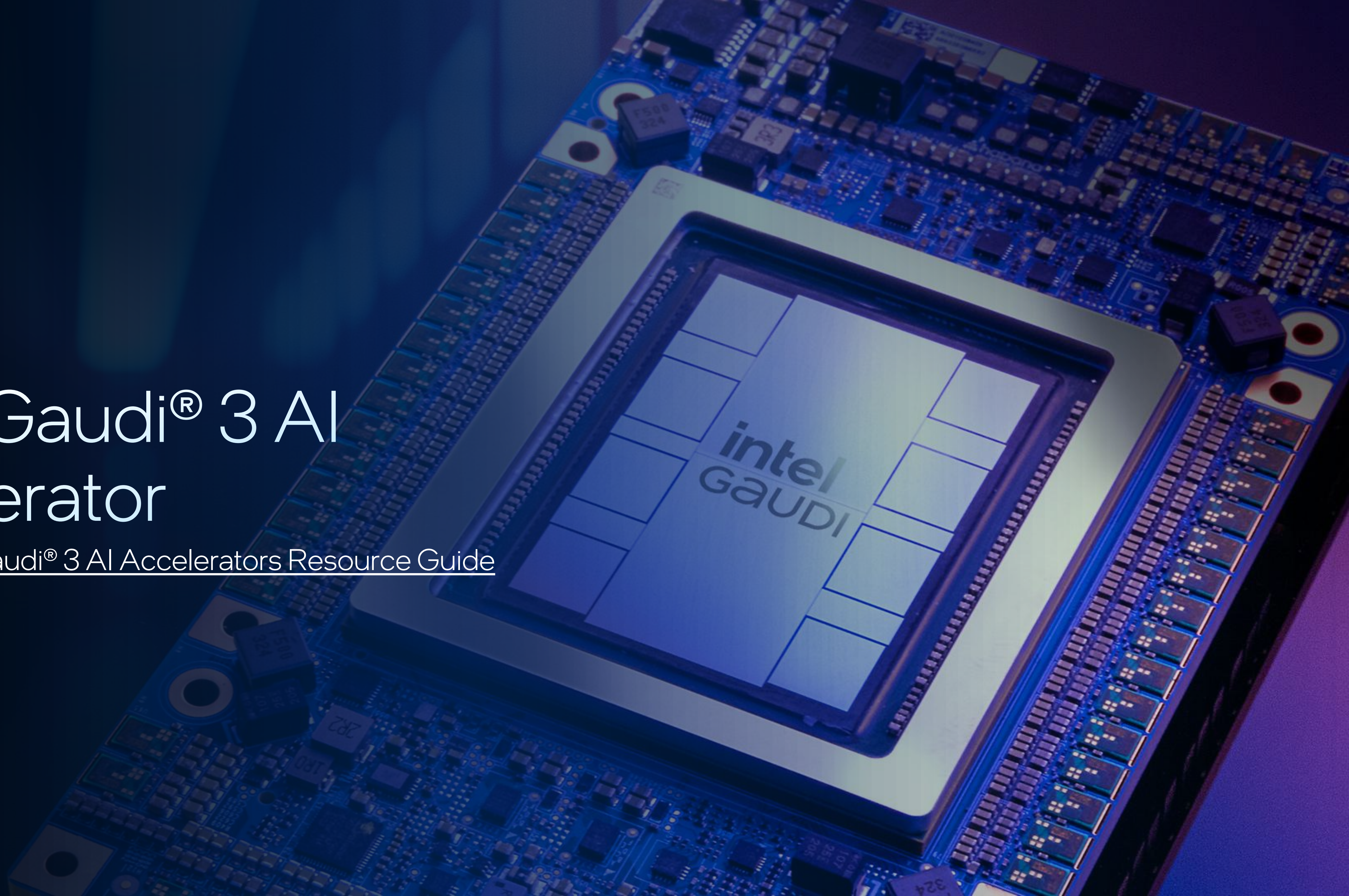
✓ Built for real-world constraints, power, space.

✓ Quick to deploy, priced for small inference models

✓ Flexibility of AI inferencing models with the different configurations

Intel® Gaudi® 3 AI Accelerator

Selling Intel® Gaudi® 3 AI Accelerators Resource Guide



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 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

- ✓ 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

Ebook: Accelerate AI at Scale with Intel® Gaudi® 3 AI Accelerator

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

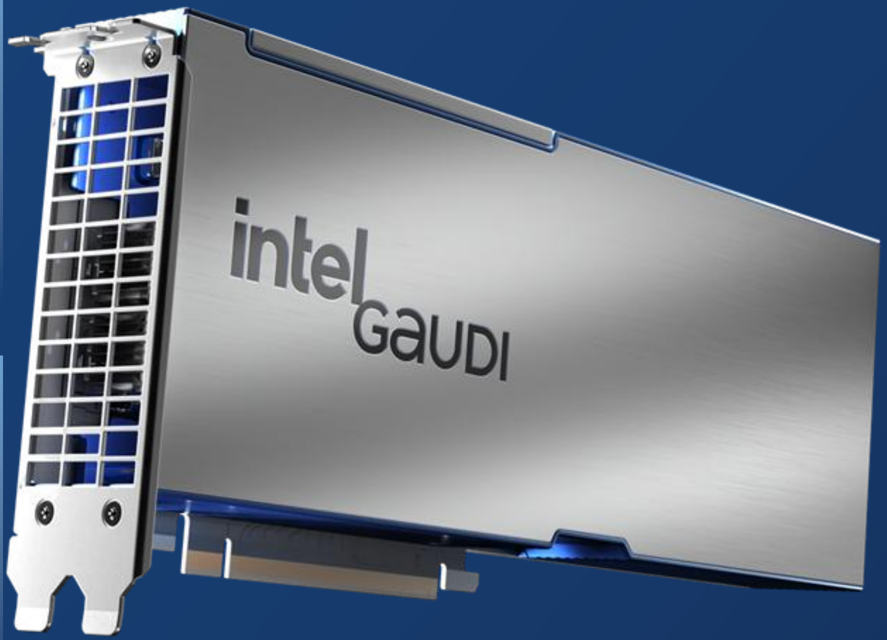
Delivering Price Performance Advantage vs H200 NVL

up to **2.15x** tokens/sec

Inference Throughput
Gaudi 3 PCIe Card
Vs H200 NVL[1]

up to **6x** perf/$

Inference Throughput
Gaudi 3 PCIe Card
Vs H200 NVL[1]



LEARN MORE

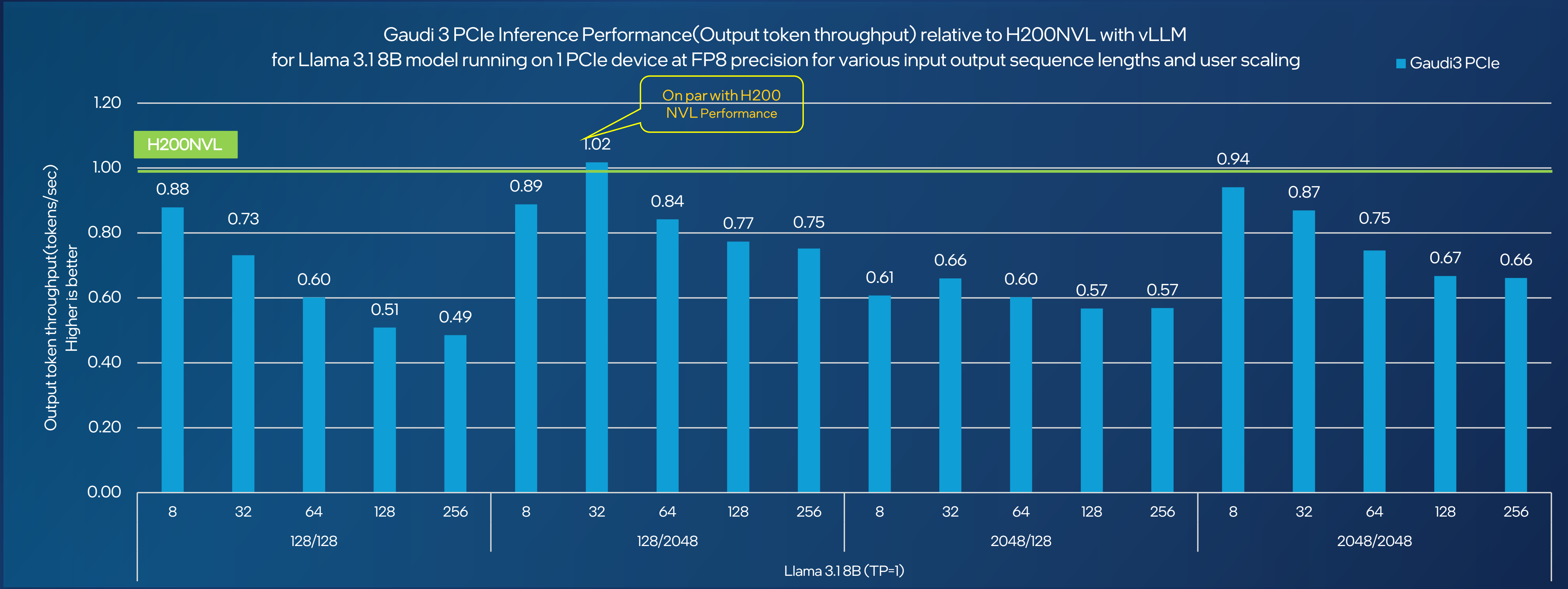
- Infographic
- Enterprise Sales Deck
- White Paper
- Performance and Positioning

Online inference performance measured as output token throughput at FP8 precision using LLAMA 3.3 70B 2048/128 with 256 user using vLLM shows 2.15x better throughput on Gaudi 3 PCIe vs H200NVL with four cards.
Pricing estimates based on publicly available information and Intel internal analysis as of 10/21/2025

[1] See backup for workloads and configurations. Your costs and results may vary.

Llama 3.1 8B Inference at FP8 precision (vLLM)

Gaudi 3 PCIe vs. H200 NVL (1 Card)

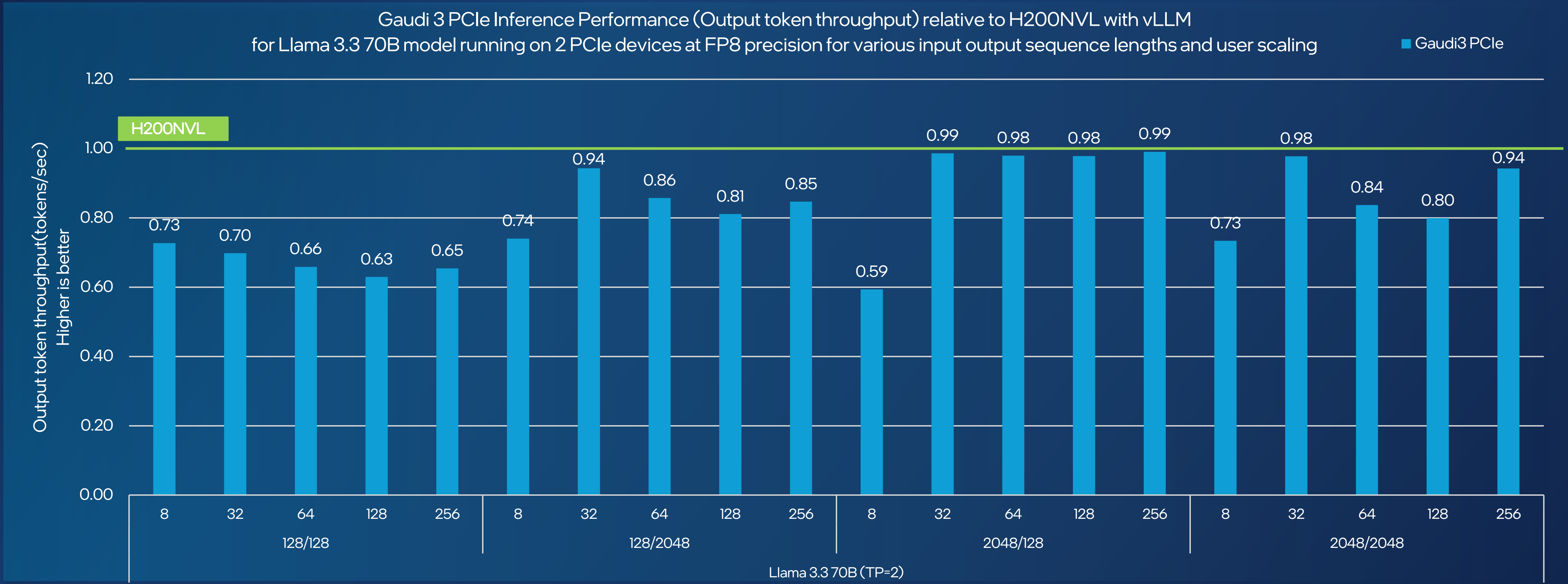


Intel® Gaudi® 3 PCIe output token throughput in above shown scenarios is in the range of **0.49x to 1.02x** and a geomean of **0.70x relative to H200 NVL.**

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.
Results may vary

Llama 3.3 70B Inference at FP8 precision (vLLM)

Gaudi 3 PCIe vs. H200 NVL (2 Cards)



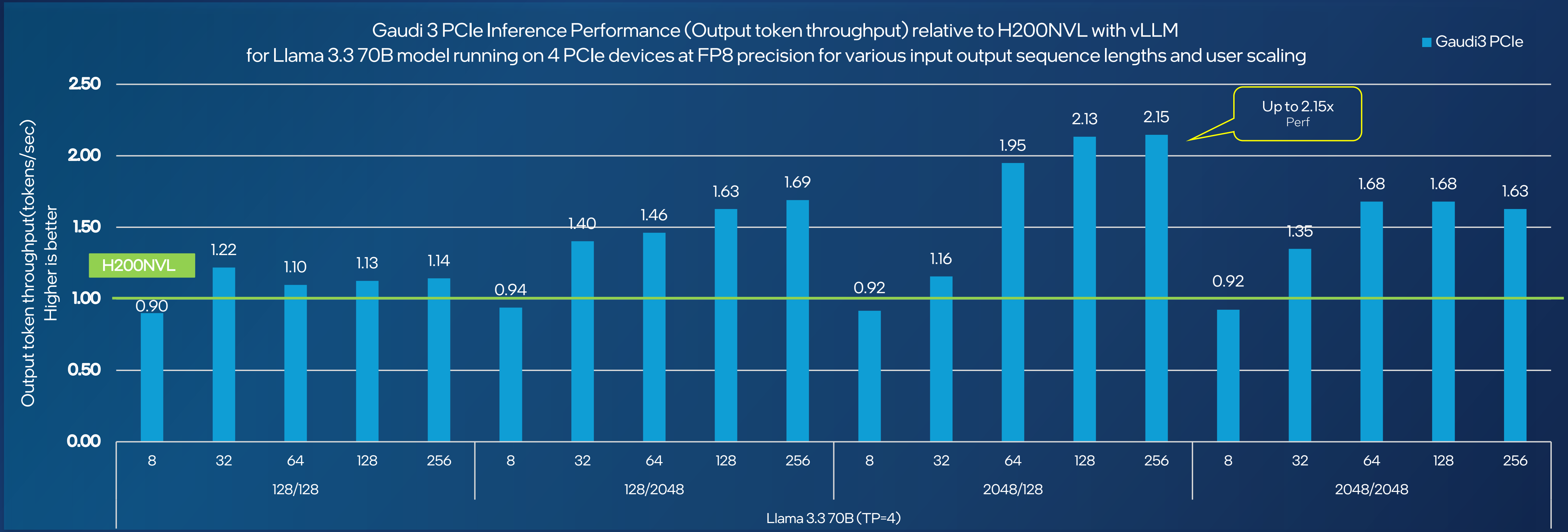
Intel® Gaudi® 3 PCIe output token throughput in above shown scenarios is in the range of **0.59x** to **0.99x** and a geomean of **0.81x** **relative to H200 NVL.**

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.
Results may vary

Llama 3.3 70B Inference at FP8 precision (vLLM)

Gaudi 3 PCIe vs. H200 NVL (4 Cards)

Intel® Gaudi® 3 PCIe offers **up to ~6.0x Perf/$ over H200 NVL *.**



Intel® Gaudi® 3 PCIe output token throughput in above shown scenarios is in the range of **0.90x** to **2.15x** and a geomean of **1.36x** **relative to H200 NVL.**

*Derived from datapoint at 2048/128 context with 256 users for H200 NVL.

Results may vary. Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

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

Seamless Integration

Optimized Graph Compiler

Cutting-Edge Kernel Fusion

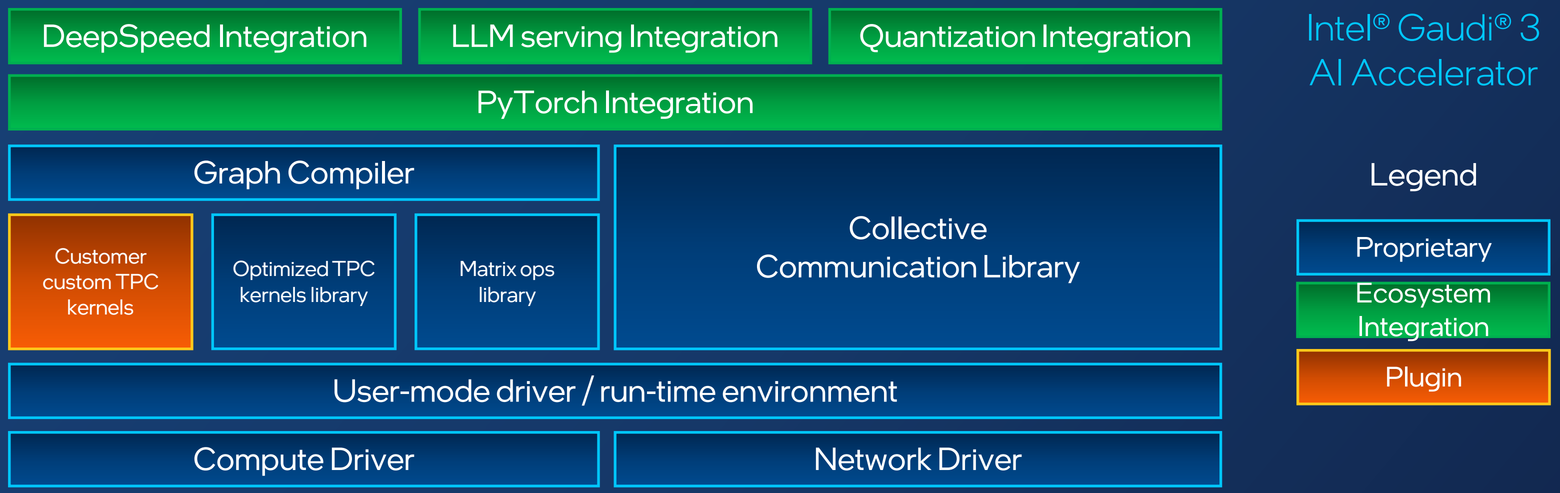
Scalable, Open Ecosystem

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

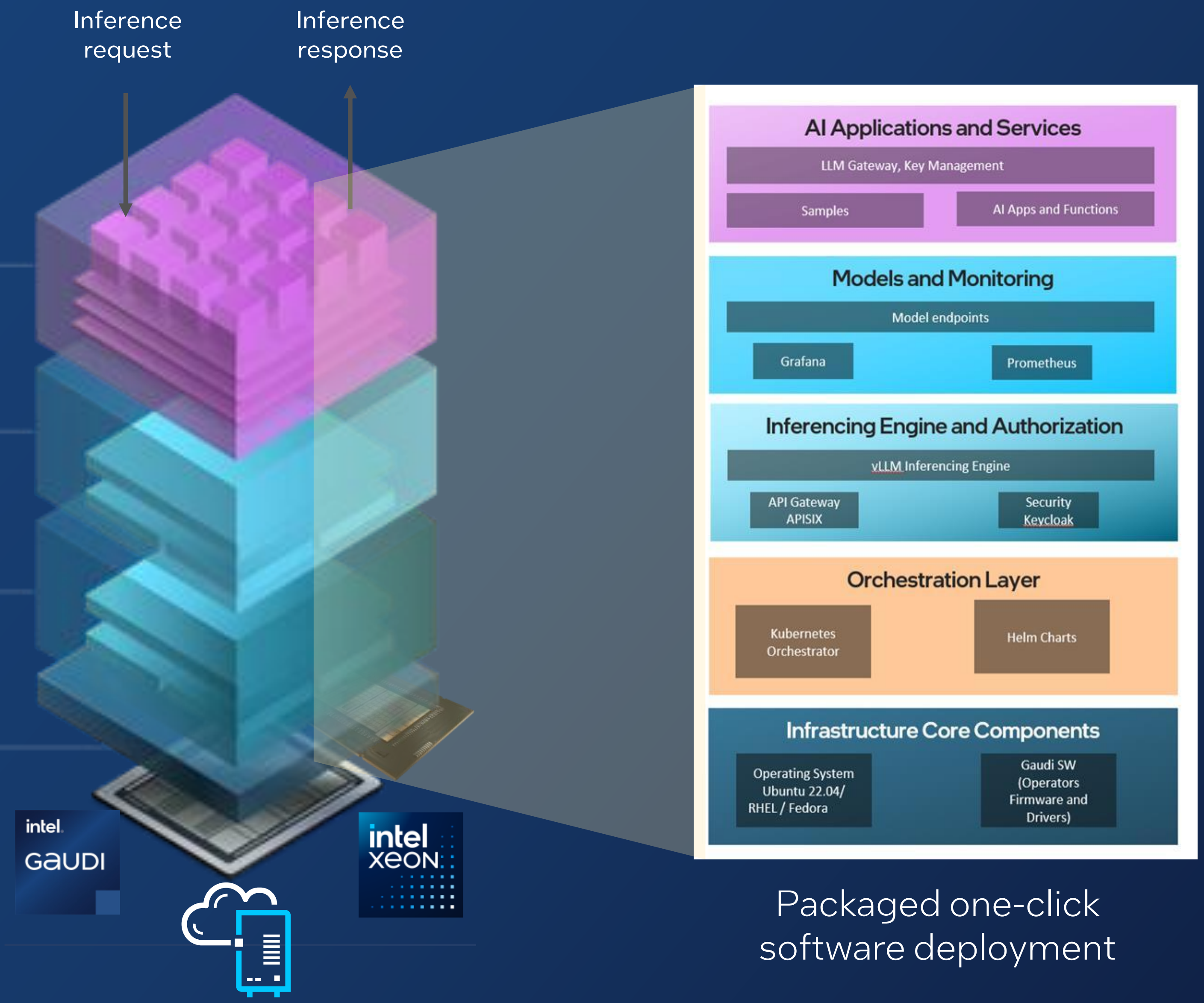
Ecosystem Integration

Plugin

Intel® AI for Enterprise Inference

Provides packages to self-host Intel-accelerated inferencing microservices for GenAI models

- Automated one-click software deployment for scalability and resilience – removes inference management overhead
- Ready-to-use, scalable, and secure industry standard inference APIs
- Free for service providers (CSPs, ISVs, OEMs) and end customers to host and adopt



AI Applications and Services
- LLM Gateway, Key Management
- Samples
- AI Apps and Functions

Models and Monitoring
- Model endpoints
- Grafana
- Prometheus

Inferencing Engine and Authorization
- vLLM Inferencing Engine
- API Gateway APISIX
- Security Keycloak

Orchestration Layer
- Kubernetes Orchestrator
- Helm Charts

Infrastructure Core Components
- Operating System Ubuntu 22.04/ RHEL / Fedora
- Gaudi SW (Operators Firmware and Drivers)

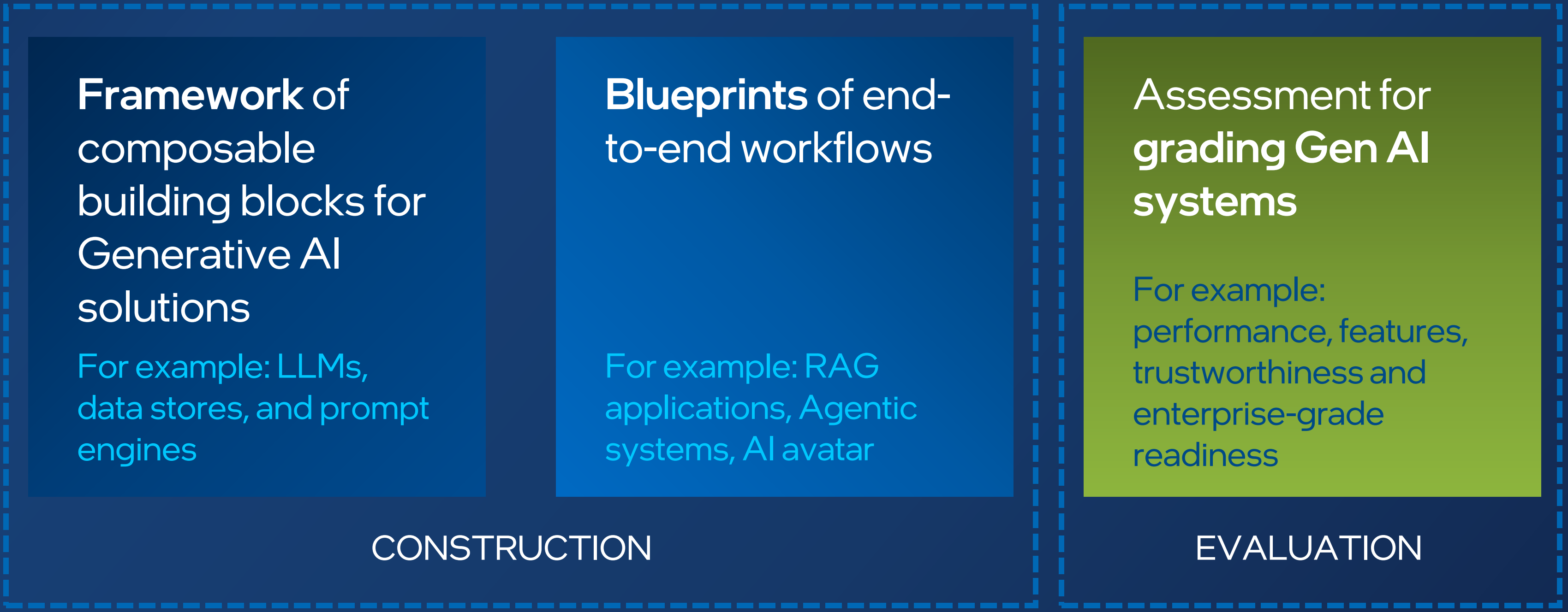
Packaged one-click software deployment

MORE INFO > Accelerate GenAI Deployment with Intel® AI for Enterprise Inference

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

Assessment for **grading Gen AI systems**

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

CONSTRUCTION

EVALUATION

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

OPEA: Partners

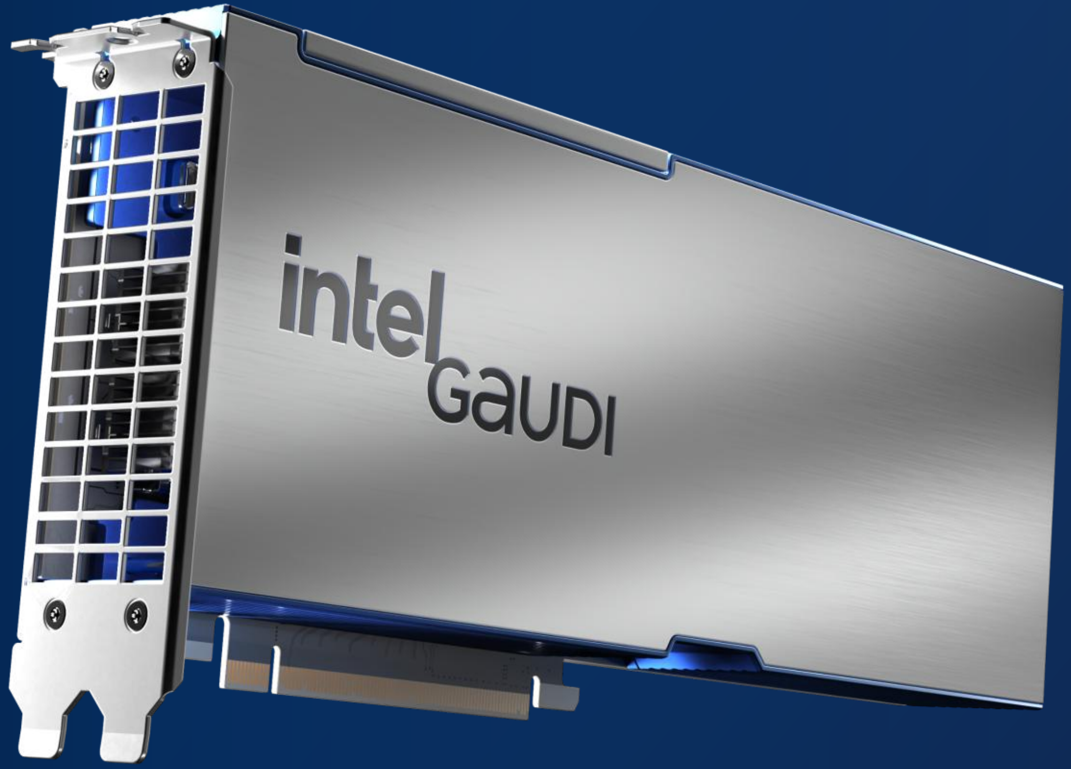
aiven
AMD
anyscale
ArangoDB
Articul8
BONC 东方国信
ByteDance
bud
CDW
Canonical
China unicom 中国联通
clarifai
CLOUDERA
Corsha
Couchbase
DATASTAX
Datastrato
docker
Domino
dstack
expanso
FRONTIERX
Haystack by deepset
Hugging Face
H3C
Infosys Navigate your next
intel
Iterate.ai
JFrog
KX
LlamaIndex
MariaDB Foundation
MINIO
MongoDB.
NetApp
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 August 2025

OPEA by the numbers

941	Github Followers
667	Stars on GenAI Examples Repo
35K	Users to OPEA.dev
600+	Registered for OPEA virtual events
55+	Partners and growing...

Intel® Gaudi® 3 PCIe Card



Reasons to Choose Intel® Gaudi® 3 PCIe

Built for AI Inference

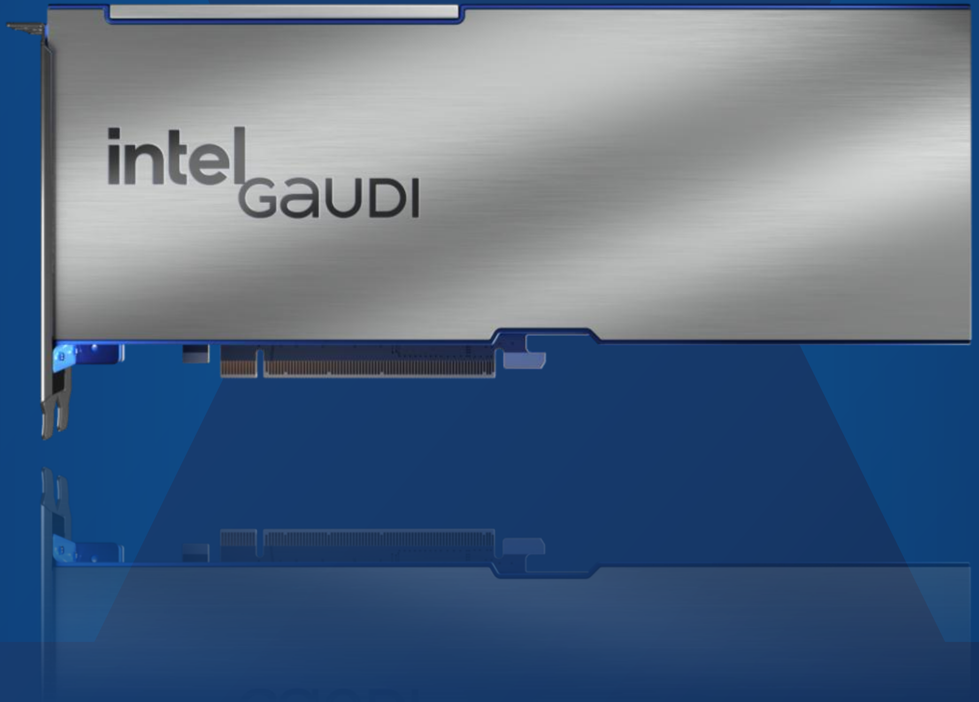
With 128GB HBM, proven on range of LLMs, Multi-Modal Models and Enterprise RAG

Performance/dollar

G3 PCIe delivers up to 3X[1] better performance per dollar than H100 NVL

Flexibility

Scale your PCIe card configurations* as your requirements grow



Scale up Network Bandwidth Advantage

900 GB/s card to card connection

On prem Enterprise AI

PCIe card-based systems designed for on premise deployment to better manage your data and cost.

LLM Ready Software

Gaudi3 provides day-zero access to top models like Llama4, Phi4, Qwen3, Falcon3, plus a diverse catalog of existing models

[1] See backup for workloads and configurations. Your costs and results may vary.

Intel® Gaudi® 3 PCIe: Solving Enterprise AI Deployment Challenges

Intel® Gaudi® 3 PCIe removes traditional AI deployment barriers making enterprise AI flexible, and efficient.

Available with leading OEM partners

Intel® Gaudi® 3 PCIe will be offered preinstalled by OEM partners, ensuring seamless on-premise deployment and support.

Lower Power Footprint

Efficiently support GenAI workloads without exceeding typical rack power budgets — ideal for constrained datacenter environments.

Compact Deployment

Smaller chassis support of 2U- 4U servers per rack, maximizing utilization in space-limited environments.

Hybrid Deployments

Enhance inference efficiency with hybrid deployments that improve latency consistency, optimizes resource allocation, and allows dynamic scalability.

Strategic Partnerships for Scalable Deployment

Intel® Gaudi® 3 AI accelerators are available through a growing network of ecosystem partners, ensuring a flexible, cost-effective path to AI infrastructure — whether deployed on-premises or in the cloud.

Through deep co-engineering efforts with leading OEMs and cloud service providers, Intel delivers optimized, validated stacks that accelerate adoption and performance for enterprise-grade GenAI workloads

On-premises solutions

A growing number of global server OEMs incorporate Intel® Gaudi® 3 AI accelerators into their system offerings, enhancing hardware vendor choice in the open AI ecosystem.

DELL Technologies

SUPERMICRO

QCT™

ASUS

Inventec

ingrasys®

GIGABYTE™

wistron

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

ACCESS NOW >

Customer-Ready Deck

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 Demo Watson Demo

SOLUTION BRIEFS >

- Improving Patient Care and Advancing Scientific Discovery
- Creating a New Vision for AI Solutions in Financial Services

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 inference

Call to Action

1 Get Started with Intel® Partner Alliance

Exclusive access to expert-led trainings, co-marketing, enablement resources and more.
Start earning points today to unlock rewards.

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 SUPERMICRO ASUS GIGABYTE™ Inventec QCT wistron wiwynn

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

IBM Cloud

Learn more

See testing results

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 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 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.

AI PC Enablement Zone

Edge AI Enablement Zone

Generative AI Enablement Zone

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® 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.

Principles of AI Transformation Competency

Enroll ›

Digital Trust for All Competency

Enroll ›

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 confidentialit**y 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**'

Configurations

# Gaudi 3 PCIe: 1-node, 8x(4x used) Gaudi3 600W PCIe cards in 2x4 config, 2x Intel(R) Xeon(R) 6787P cpu, Total Memory 2.0 TB (32x64GB DDR5-6400), BIOS 1.2.2, Microcode 0x010003c2, Network: 8x Mellanox Cx-7, 2x Intel X710 10Gb, Storage: Dell BOSS N1 NVME, Ubuntu 24.04.03 LTS, Kernel 6.8.0-51-generic, Synapse 1.22.0-740. 1x Network Switch Arista DCS-7060DX4-32 QSFP, 8x Network Adapter Mellanox CX-7 MCX75310AAS-NEAT OSFP, 8x OSFP to QSP cables 2.5m Credo CAC425321M1B-C0-HW. Test by Intel as of Nov 01, 2025

# H200NVL System: 1-node, 8x(4x used) H200 [NVIDIA H200 NVL 141GB], 2x Intel(R) Xeon(R) 6747P cpu, Total Memory 3072GB (32x96GB DDR5 6400 MT/s), BIOS 1.2.2, microcode 0x10003c2, 1x Dell Ent NVMe CM7 E3.S RI 3.84TB, Ubuntu 24.04.2 LTS (Kernel: 6.8.0-52-generic), NVIDIA-SMI 580.95.05, Driver Version: 580.95.05, CUDA13.0, VLLM-v0.10.2,nvcr.io/nvidia/tritonserver:25.09-trtllm-python-py3. Test by Intel as of Nov 01, 2025

Configurations

# G3 PCIE: 2x Intel(R) Xeon(R) 6787P CPU @ 2.0GHz , 2.0 Tb RAM, 32x64GB Samsung M321R8GA0PB1-CCPPC, BIOS 1.2.2, Microcode 0x010003c2, 8x Habana labs 600W PCIe cards in 2x4 config, Network: 8x Mellanox Cx-7, 2x Intel X710 10Gb, Storage: Dell BOSS N1 NVME, Ubuntu 24.04.03 LTS, Kernel 6.8.0-51-generic, Synapse 1.21.4. Test by Intel as of Aug 10, 2025

# H100 System: 1-node, 1x Intel(R) Xeon(R) 6761P, 64 cores, ? TDP, HT On, Turbo On, Total Memory 512GB (16x32GB DDR5 6400 MT/s [5200 MT/s]), BIOS 1.0, microcode 0xa0000d1, 2x H100 [NVIDIA H100 NVL 94GB], 2x KINGSTON SNV3S1000G 1TB, Ubuntu 24.04.2 LTS (Kernel: 5.15.0-141-generic), NVIDIA-SMI 570.133.20, Driver Version: 570.133.20, CUDA Version: 12.9. Test by Intel as of Wed July 23, 2025