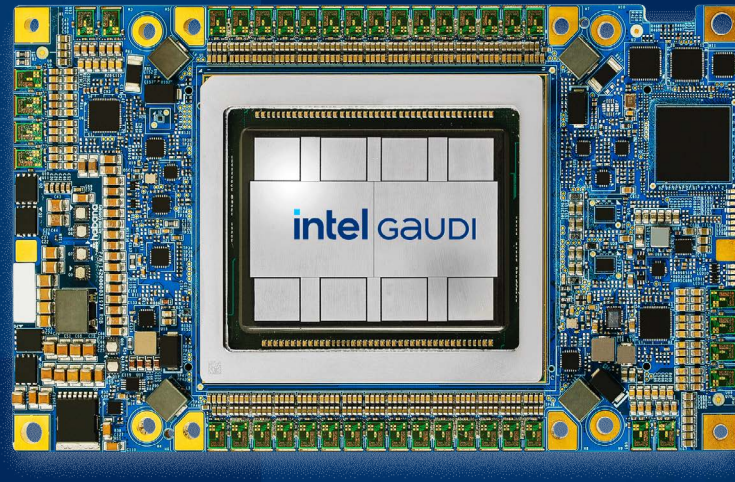




Intel® Gaudi® 3 AI Accelerators

With outstanding price-performance and throughput, plus the ability to scale to thousands of nodes over standard Ethernet, Intel Gaudi 3 AI accelerators bring serious competition to large-scale training and inference.

Avoid the lock-in of proprietary software and networking with an open source software stack. Streamline migration with the tools and libraries needed for your models.



Superior Price-Performance and Throughput Outstanding Results versus Nvidia

Near-parity Inference Throughput with

1.7x perf/\$*

vs. H100

Better performance on larger output sequences

Llama 3 - 8B (1 accelerator)

Llama 2 - 70B (2 accelerators)

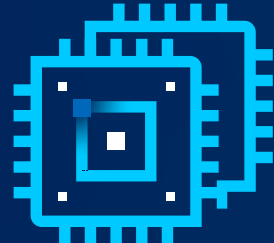
Inference Throughput with

1.6x perf/\$**

vs. H200

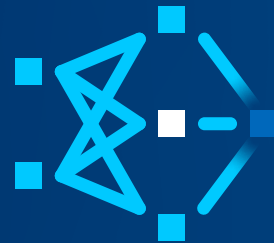
Llama 3 - 8B (1 accelerator)

Llama 2 - 70B (2 accelerators)



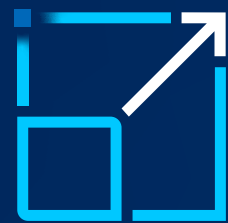
**More choice
versus single GPU provider**

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, 2.6x 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 bandwidth

Twenty-four 200 GbE
RoCE v2 NICs

* Compared to Intel® Gaudi® 2 AI accelerator

- ✓ 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

"As organizations strive to meet their advanced computing demands, Dell Technologies and Intel are delivering the ability to scale along with technology designed to maximize performance per dollar with the industry-leading Dell PowerEdge XE9680 server equipped with Intel® Gaudi® 3 AI accelerators."

– Arun Narayanan, SVP, Servers and Networking, Dell Technologies

"For over 30 years, Hewlett Packard Enterprise and Intel have worked closely to co-develop 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. HPE servers will continue to support the Intel Xeon 6 processor family for various enterprise needs."

– Trish Damkroger, Senior Vice President & General Manager, HPC & AI Infrastructure Solutions at Hewlett Packard Enterprise

"AI innovation is not slowing down, which means organizations need to speed up. Red Hat and Intel remain committed to helping businesses realize the value of AI and generative AI faster and more efficiently through enterprise-ready, AI-enabled hardware and software solutions. 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."

– Steven Huels, Vice President and General Manager, AI Engineering, Red Hat

"This 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® chip, all backed by high bandwidth interconnectivity. Seekr's trustworthy AI products combined with the 'AI-first' Intel Developer Cloud reduces errors and bias, so organizations of all sizes can access reliable LLMs and foundation models to unlock productivity and fuel innovation, running on trusted hardware."

– Rob Clark, President & CTO, Seekr

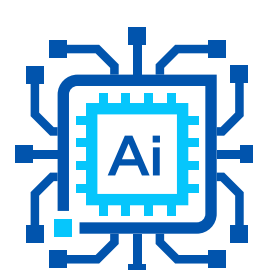
"Building on the successful deployment of the world's largest Intel® Gaudi® 1 and Gaudi® 2 clusters, Supermicro is now pleased to offer the industry's first and only Intel® Xeon® 6 based Gaudi 3 system powered by the Xeon 6900 series with P-cores. Supermicro's Gaudi3 AI server is available in both air-cooled and liquid-cooled total solutions and comes packed with superior hardware performance, an enhanced software stack and tailored vertical solutions to accelerate AI adoption in key enterprise markets such as healthcare, financial services and manufacturing."

– Ray Pang, Supermicro SVP, Technology and Business Enablement

Explore the Intel Gaudi 3 AI Accelerator on Intel® Tiber™ AI Cloud

Accelerate your business, by testing, building
and deploying AI at scale on a managed, high
performance, cost-efficient infrastructure.

cloud.intel.com



Intel Gaudi 3 AI Accelerator Ecosystem Partners



Get started at www.intel.com/gaudi3

*Source: 1.7x perf/\$ NV H100 comparison based on TensorRT-LLM/docs/source/performance/perf-overview.md at main · NVIDIA/TensorRT-LLM · GitHub, October 18, 2024. Reported numbers are per GPU vs Intel® Gaudi® 3 measurements for LLAMA3-8B, LLAMA2-70B. Results may vary. Based on Intel Gaudi software release: 1.18. Refer to this link for the latest published Gaudi3 performance <https://www.intel.com/content/www/us/en/developer/platform/accelerator/ai/ai-model-performance.html>. Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

**Source: NV H200 comparison based on TensorRT-LLM/docs/source/performance/perf-overview.md at main · NVIDIA/TensorRT-LLM · GitHub, October 18, 2024. Reported numbers are per GPU vs Intel® Gaudi® 3 projections for LLAMA3-8B and LLAMA2-70B. Based on Intel Gaudi software release: 1.18. Refer to this link for the latest published Gaudi3 performance <https://www.intel.com/content/www/us/en/developer/platform/accelerator/ai/ai-model-performance.html>. Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

Performance varies by use, configuration and other factors. Learn more at www.intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. 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 Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

0225/JW/RB/US