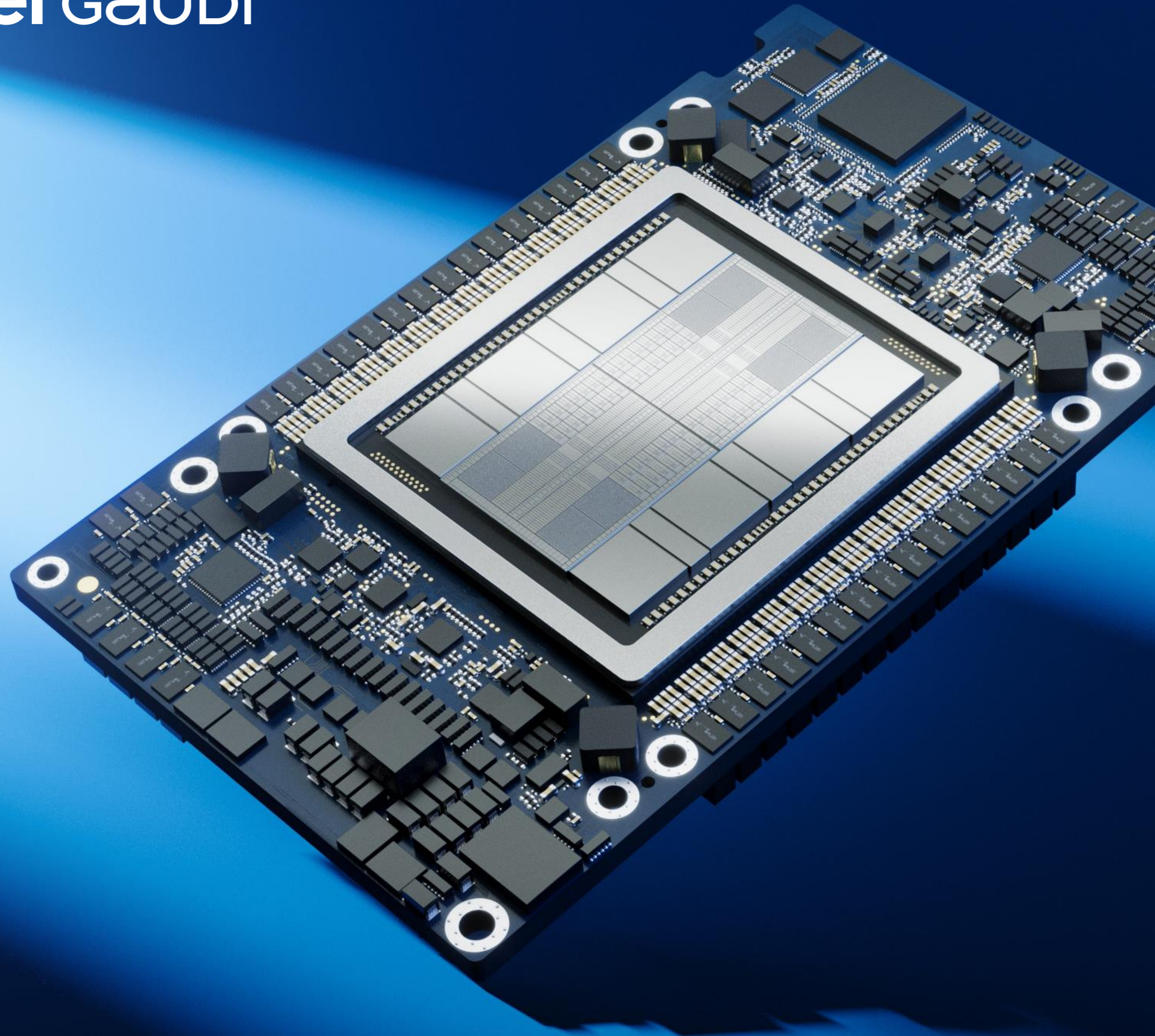


intel gaudi



Programming Novel AI Accelerators for Scientific Computing Intel®

Intel® Gaudi® 3 AI accelerator

Milind S. Pandit (milind.s.pandit@intel.com)

INTEL® GAUDI® 3 AI ACCELERATOR

Accelerator Card

OAM-Compliant (HL-325L)



INTEL® GAUDI® 3 AI ACCELERATOR

PCIe CEM

Add-In Card (HL-338)

| | | |
|---|----------------|--|
| 5th Generation Tensor Processor Core | Architecture | 5th Generation Tensor Processor Core |
| 900W | TDP | 600W |
| OAM-Compliant (HL-325L) | Form Factor | FH 10.5" in length, Double Width (x16 PCIe Gen 5.0) |
| 128 GB/s bidirectional | PCIE Peak BW | 128 GB/s bidirectional |
| FP32, BF16, FP16 & FP8 (both E4M3 and E5M2) | Data Types | FP32, BF16, FP16 & FP8 (both E4M3 and E5M2) |
| 8 x HBM2E | HBM | 8 x HBM2E |
| 128 GB | HBM Capacity | 128 GB |
| 3.7 TB/s | HBM Peak BW | 3.7 TB/s |
| 96 MB | On-die-SRAM | 96 MB |
| 19.2 TB/s | On-die-SRAM BW | 19.2 TB/s |
| x8 UBB | Card Config(s) | 1x4 (Bridged), 2x4 (Bridged), 1 or 4 cards (Unbridged) |
| 1200 GB/s bidirectional (24x200 GbE) | Networking | 900 GB/s bidirectional (18x200 GbE) |

Gaudi 3 Performance

intel

PRODUCTS

SUPPORT

SOLUTIONS

DEVELOPERS

PARTNERS

FOUNDRY

ENGLISH

Search Intel.com

Developers / Hardware Platforms / Intel® Gaudi® Software / Models / Overview / Model Performance Data for Inference on Intel Gaudi 3 Accelerator

intel GAUDI

Models

Overview

Catalog

Performance Data

Hugging Face*

Model Optimization and Debugging

Model Reference

Model Performance Data for Intel® Gaudi® 3 AI Accelerators

These performance numbers are measured using the latest Intel® Gaudi® software release version 1.20, unless otherwise noted.
Note All models for both training and inference are using the PyTorch* 2.6.0 framework. Other applicable frameworks used for training or inference are noted for each model.

[Explore Intel® Gaudi® 2 Accelerator Performance Data](#)

INFERENCE

Large Language Models (LLM) for Throughput with Intel® Gaudi® 3 Accelerators

Search Table

☒ Model

☒ Precision

☒ Input Length

☒ Output Length

☒ HPU

☒ Batch Size

☒ Throughput

| Model | Precision | Input Length | Output Length | #HPU | Batch Size | Throughput (tokens/sec) |
|-------------|-----------|--------------|---------------|------|------------|-------------------------|
| LLaMA 2 70b | fp8 | 128 | 128 | 2 | 1750 | 4853 |
| LLaMA 2 70b | fp8 | 128 | 2048 | 2 | 512 | 6835 |
| LLaMA 2 70b | fp8 | 2048 | 128 | 2 | 242 | 506 |
| LLaMA 2 70b | fp8 | 2048 | 2048 | 2 | 241 | 2859 |

Feedback

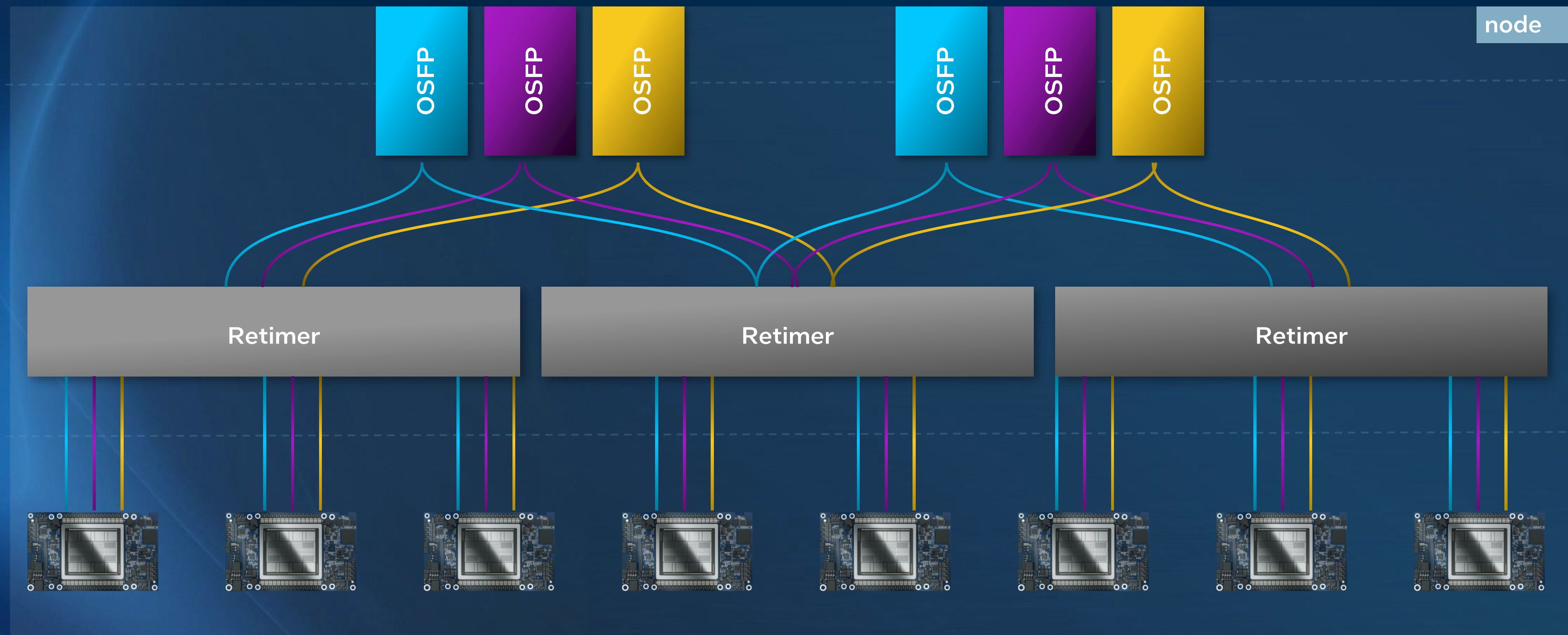
Gaudi 3 Performance

- 20%-43% more tokens per second than H200
- 1.2X tokens-per-dollar compared to H200
- 3.35X tokens-per-dollar compared to H100

Source: Intel® Gaudi® 3 AI Accelerator at Scale on IBM Cloud,
<https://signal65.com/research/ai/signal65-lab-insight-intel-gaudi-3-accelerates-ai-at-scale-on-ibm-cloud/>

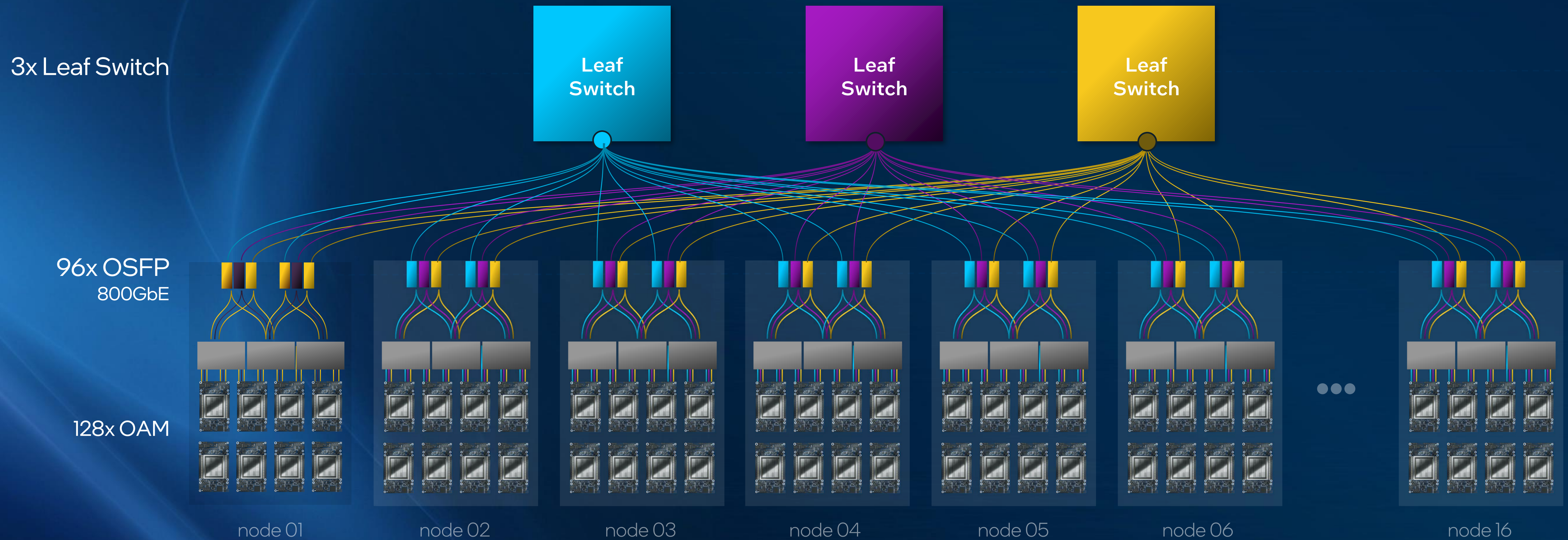
Intel Gaudi 3 Accelerator Scale-out for GenAI Requirements

Node Level Architecture



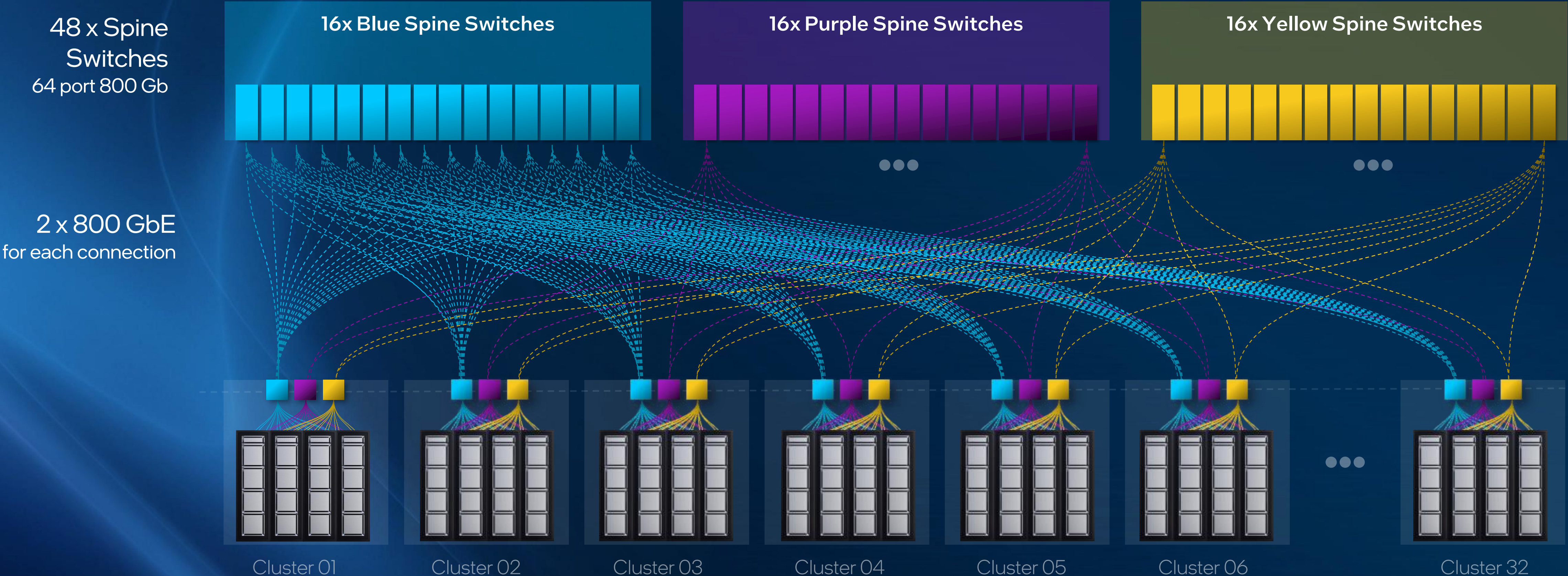
Intel Gaudi 3 Accelerator Scale-out

Sub-Cluster Level Architecture (16 Nodes)



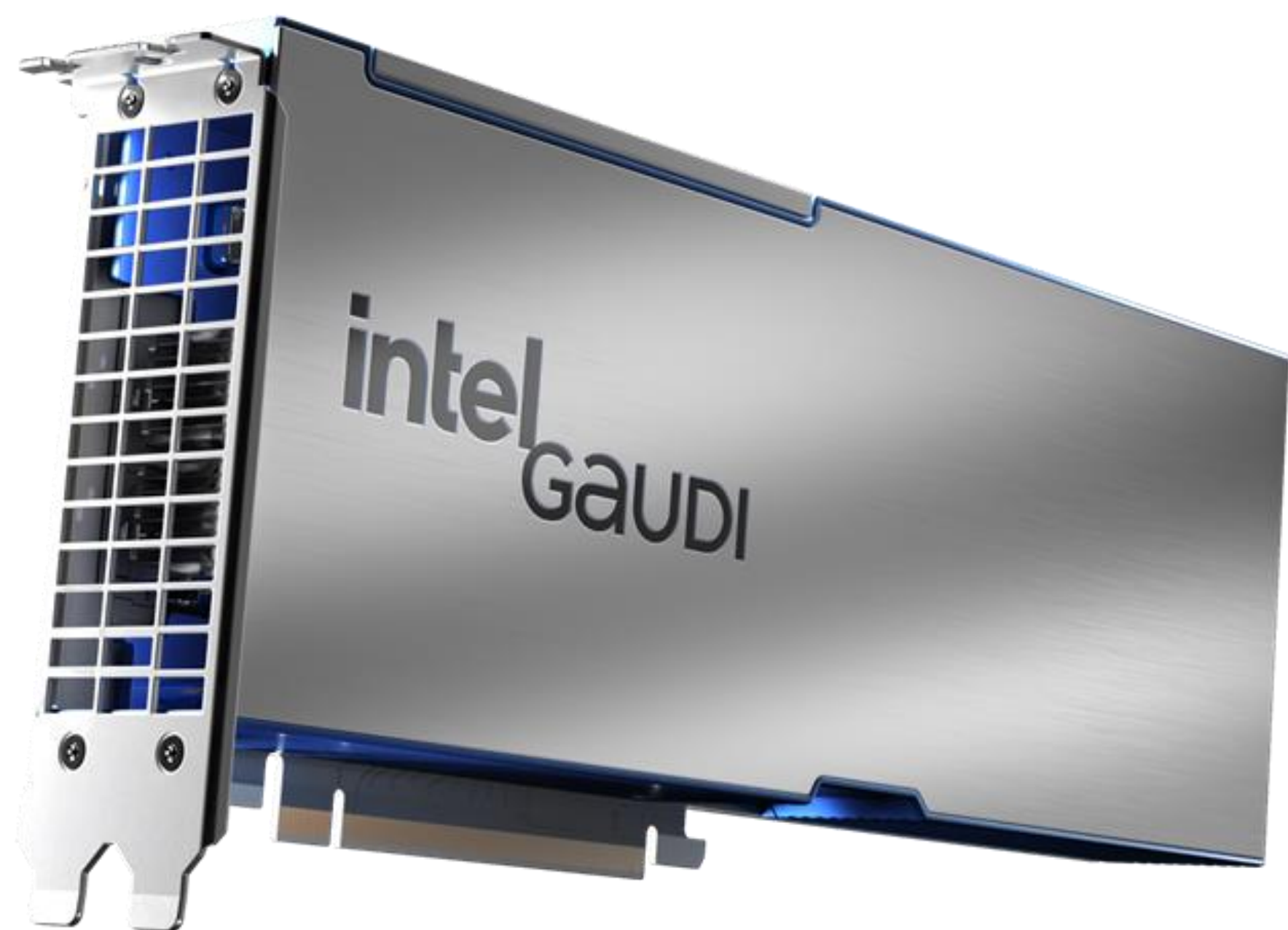
Intel Gaudi 3 accelerator Scale-out meets GenAI Requirements

Cluster Level Architecture (512 Nodes)



Delivering Price Performance Advantage

intel GAUDI



up to **1.7x** tokens/sec

Inference Throughput
Gaudi 3 PCIe Card
Vs H100 NVL

up to **3.0x** perf/\$

Inference Throughput
Gaudi 3 PCIe Card
Vs H100 NVL

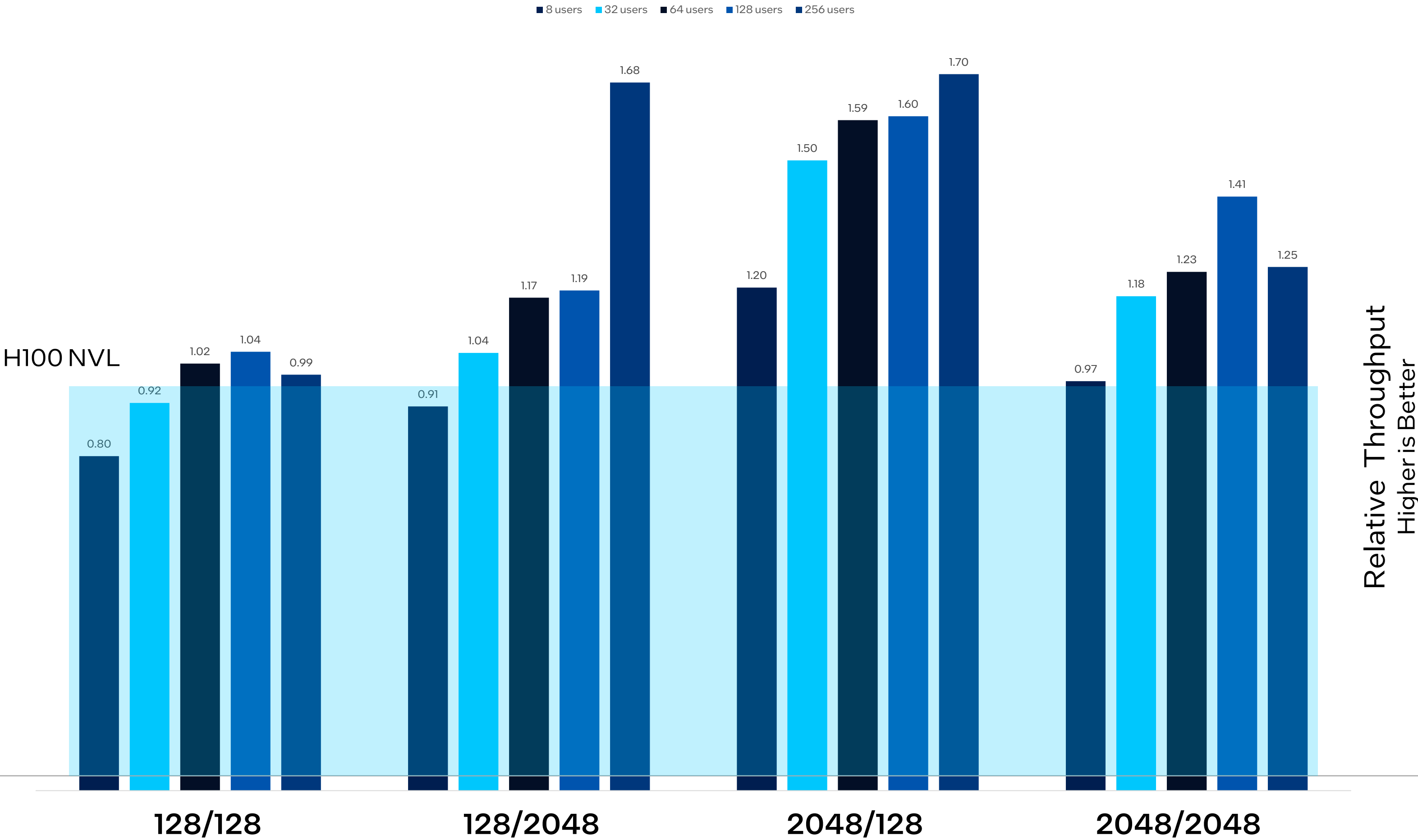
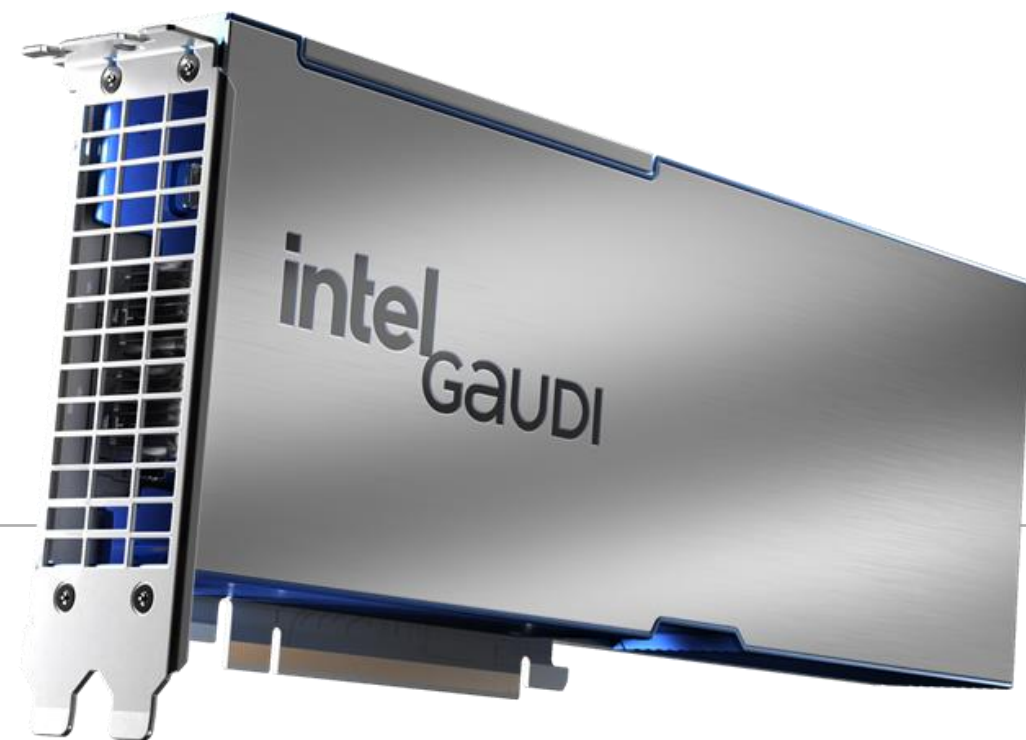
Online inference performance measured as output token throughput at FP8 precision using LLAMA 3.3 70B 2048/128 with vLLM shows 1.7x better throughput on Gaudi 3 PCIe vs H100 with two cards.
Pricing estimates based on publicly available information and Intel internal analysis as of 8/27/2025

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

1.2x tokens/sec

Geometric mean of throughputs

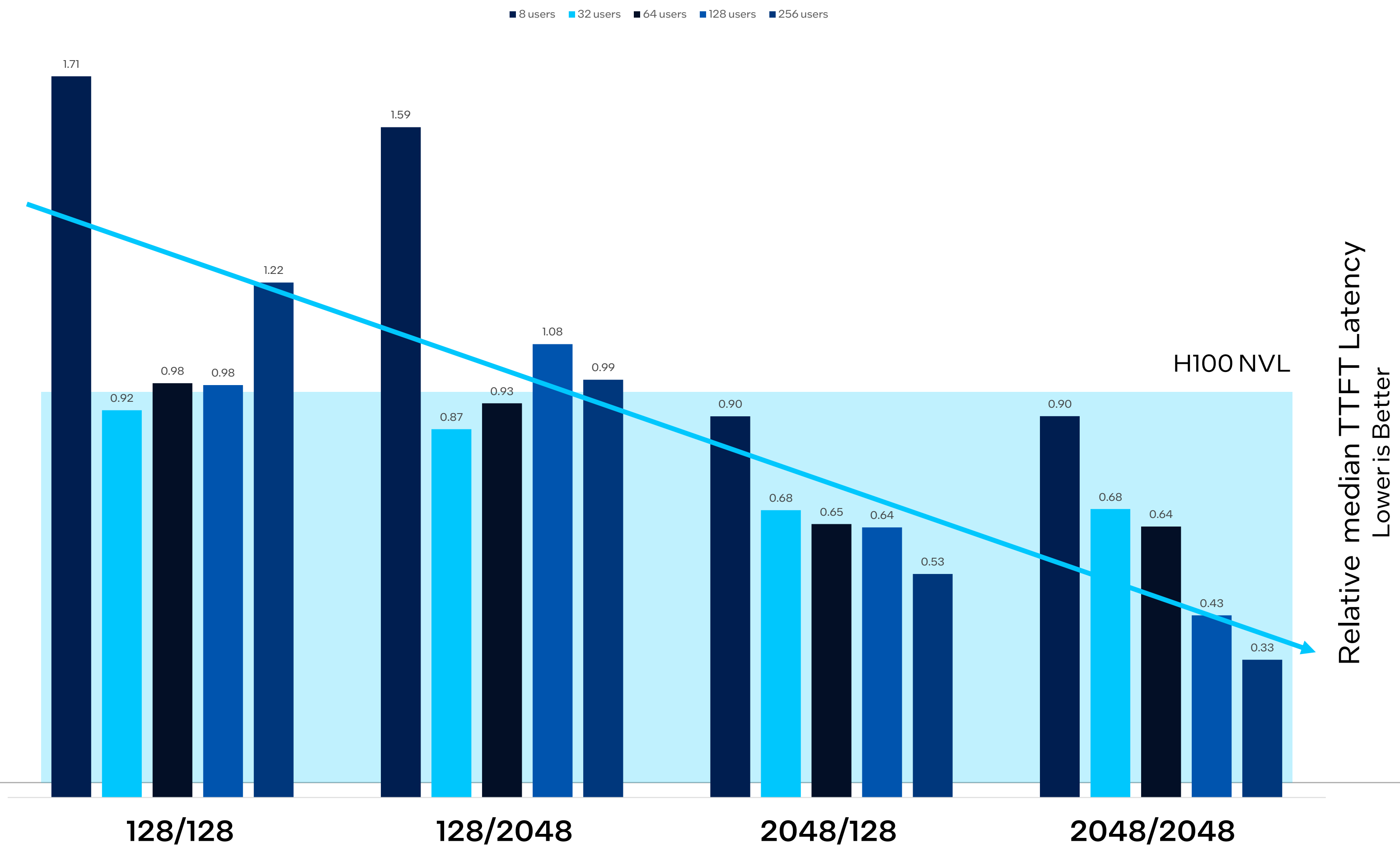
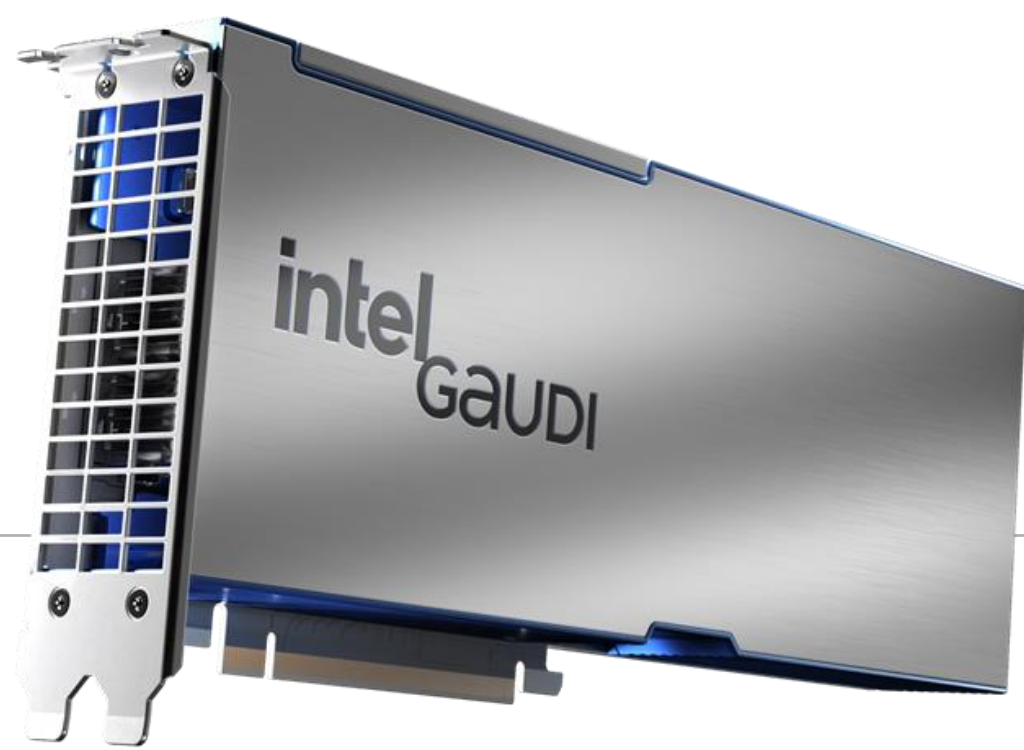
Gaudi 3 PCIe Card
Vs H100 NVL



Claim: Throughput measured as Geometric mean of throughputs across scenarios 2048/128, 2048/2048 representing various real world use cases and Concurrency spanning 8,32,64,128,256 users

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

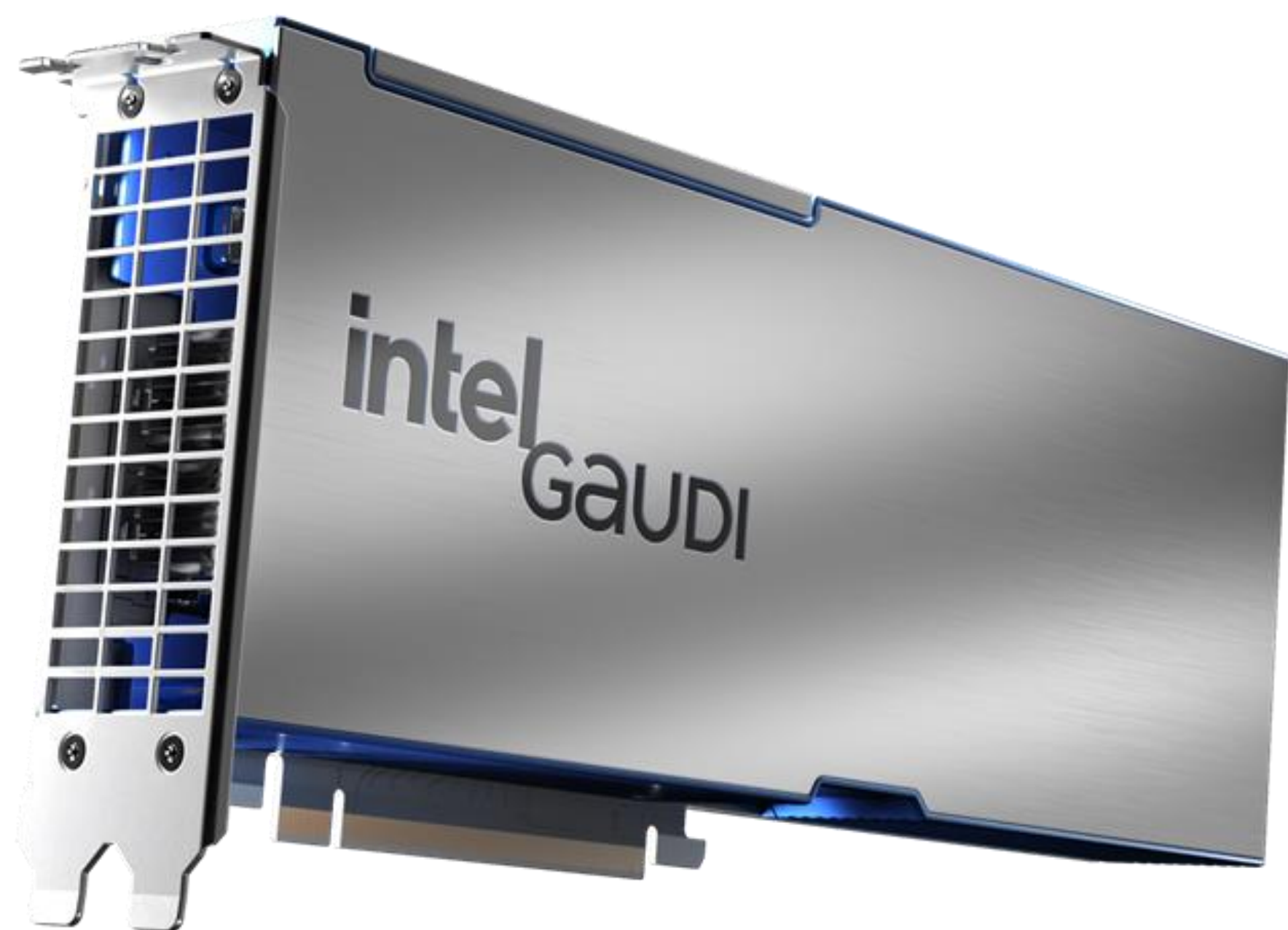
Delivers **lower latency**
at **higher context sizes**
vs **H100 NVL**



Claim: Online inference performance measured as time to first token at FP8 precision using LLAMA 3.3 70B with vLLM shows to be in the range of 1.71x to 0.33x on Gaudi 3 PCIe vs H100 with two cards

Delivering Price Performance Advantage vs H200 NVL

intel GAUDI



up to **2.15x** tokens/sec

Inference Throughput
Gaudi 3 PCIe Card
Vs H200 NVL

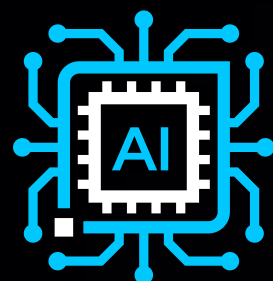
up to **6x** perf/\$

Inference Throughput
Gaudi 3 PCIe Card
Vs H200 NVL

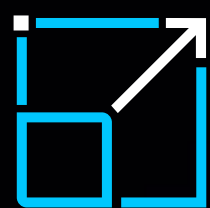
1. See backup for workloads and configurations. Your costs and results may vary.
Online inference performance measured as output token throughput at FP8 precision using LLAMA 3.3 70B 2048/128 with 256 user using vLLM shows 2.15x better throughput on Gaudi 3 PCIe vs H200NVL with four cards.
Pricing estimates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

Can AI inference scalability be simpler? **It can!**

Scale your AI to meet your needs with flexible, performant Intel® Gaudi® 3 PCIe cards.



Intel® Gaudi® 3 PCIe cards provide a modular AI growth path without overcommitment, making them a great choice for early pilots or phased AI deployments.

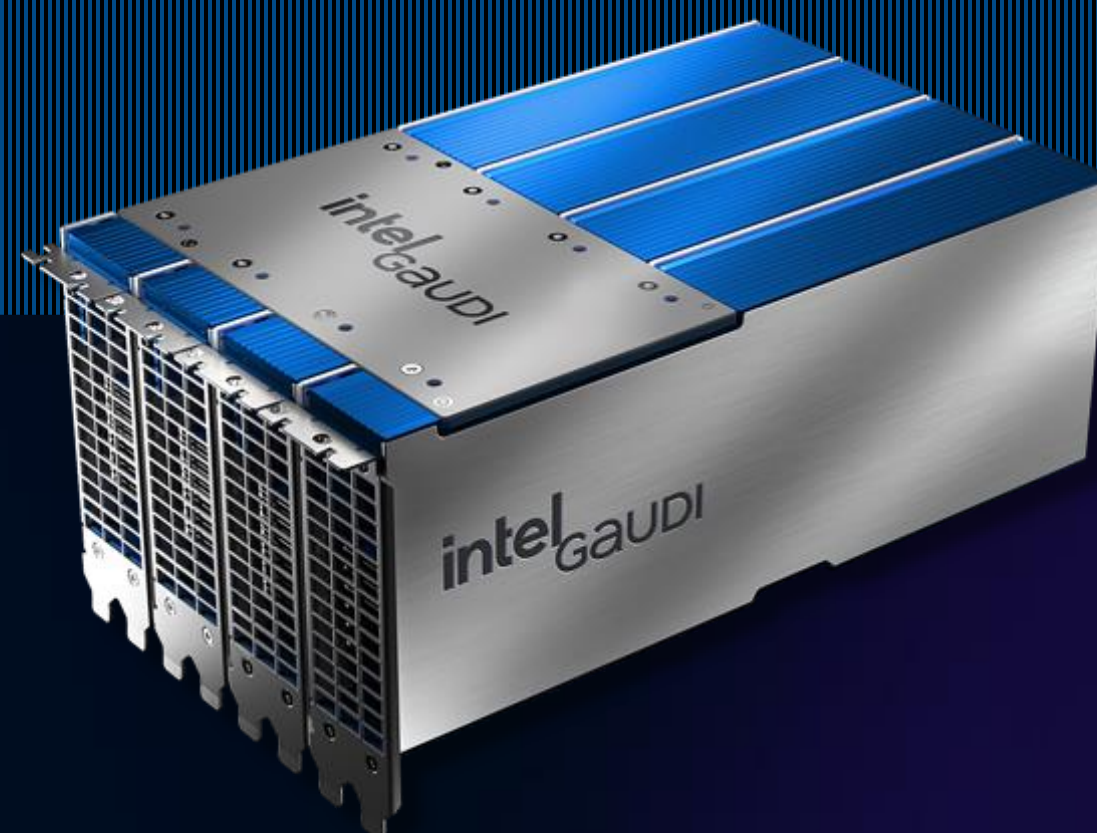


Scale up your AI infrastructure from one to up to eight Intel® Gaudi® 3 PCIe cards per system to meet your specific AI needs.

Inference - Lower latency, increased users, higher throughput



1 or 4 cards
(unbridged)



1x4 (1 top bridge)



2x4 (2 top bridges)

Intel® Gaudi® 3 AI accelerator Software Suite

Integrates the main Gen AI frameworks used today

Supports BF16 and FP8 quantization

vLLM Integration

- Upstream support in vLLM mainline
- Continuous batching, paged attention
- Multi-step scheduling
- Speculative decoding

Gaudi PyTorch Bridge

- Open sourced
- Work with the latest upstream PyTorch
- TPC fuser

Layered View of Intel® Gaudi® Software Suite

LLM Serving:



Quantization
Integration

Quantization Toolkit
(INC)

PyTorch Integration | PyTorch Bridge ([Github Repository](#))

Graph Compiler

Custom
user
TPC kernels

Optimized
TPC kernel
library

Matrix ops
library

Collective
Communication
Library (HCCL)

User-mode driver/run-time environment

Compute Driver

Network Driver

Ways to Run Models on Gaudi 3



Intel [Model References](#) GitHub

A collection of deep neural network (CV, NLP, Diffusion, LLM) models that have been migrated to run on Intel Gaudi AI accelerators, including examples for training, fine-tuning, and inference



Hugging Face

Start with examples of training, fine-tuning, and inference or use the [Optimum Habana](#) library with any transformer (NLP, code generation, translation, Q&A) model



vLLM Fork

[Instructions](#) for serving models in vLLM on Gaudi

Other PyTorch Models

[Migrate models](#) built for CPUs or GPUs.

intel®