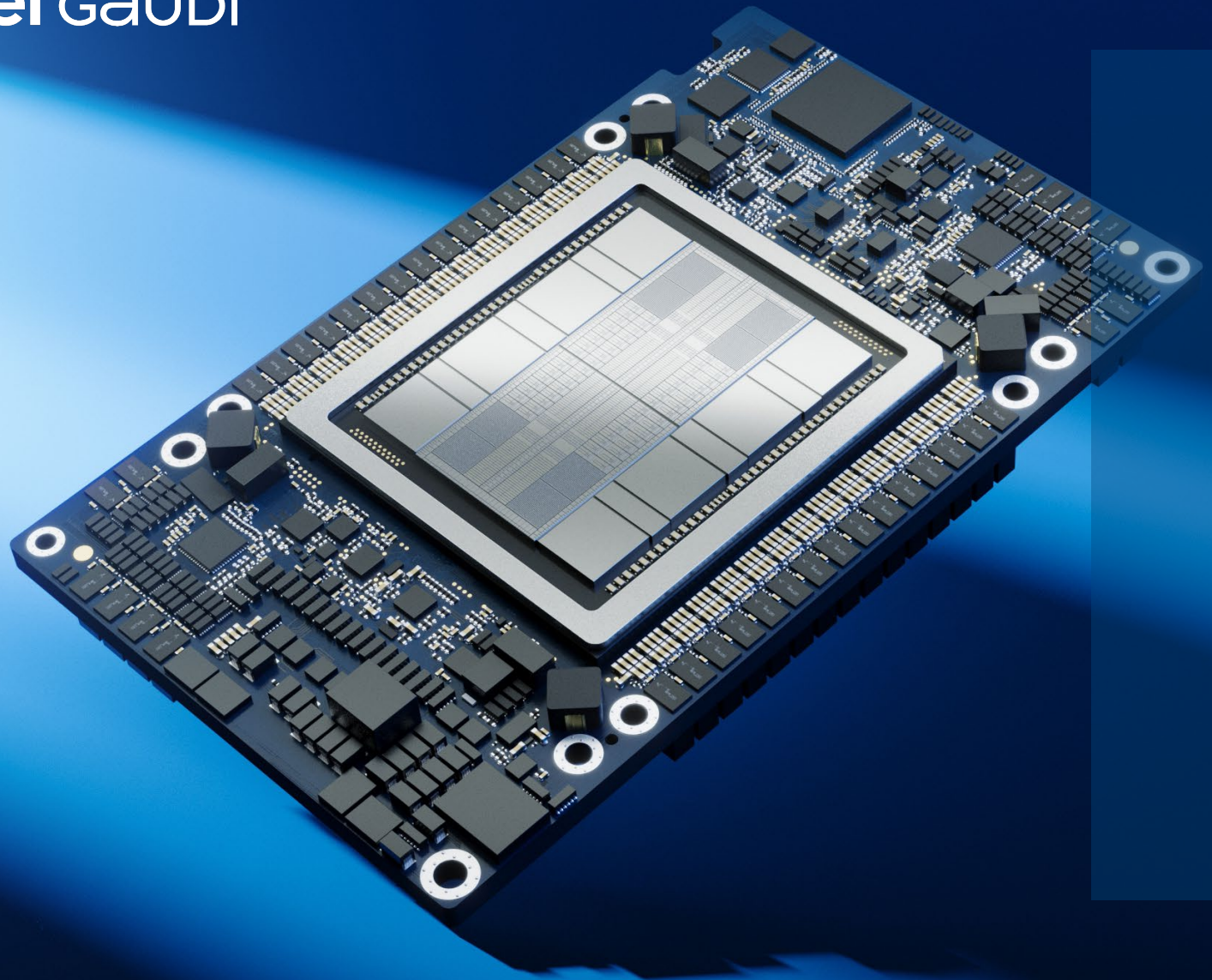
intel GAUDI



Intel® Gaudi® 3 AI Accelerator

Performance and Positioning

Version 1.1

Agenda

- **Performance Testing Overview**
- **Intel Gaudi 3 On-Prem Testing**
  - Llama 3.1 Inference Performance
  - Llama 3.1 Price Performance
- **Intel Gaudi 3 on IBM Cloud**
  - IBM Granite Performance Testing
  - Mixtral Performance Testing
  - Llama Performance Testing
  - Performance per Dollar

Intel Gaudi 3 Performance Testing with Signal65

Signal65 is a third-party performance testing and validation lab

- Provides independent, unbiased performance results

Conducted two studies to evaluate the competitive performance of Intel Gaudi 3 AI Accelerator

- Intel Gaudi 3 On-Prem Testing
- Intel Gaudi 3 on IBM Cloud

Gen AI Inference Parameters

Model Architecture

How data flows, process information, and interconnect

Transformers, MoE, GNN, Stable Diffusion, etc.

Models & Model Size

Pre-trained models based on training data sets – based on parameter sizes

Small to Large models 1B to 1T

Llama 3b/8b/ 70b/405b

Data Types

data types and precisions are used to balance accuracy, speed, and memory efficiency.

Int4, FP8, BF16

Context Length

Maximum number of tokens processed in single request

Input & Output

Sizes: 128 – 2k-16K-128K- more

Serving Frameworks

tool or system that helps run and deliver AI models so they can be used in real-world applications

vLLM, nVidia Triton Inference Serving, TGI, etc

Inference Performance

Throughput

Tokens per second

TFFT

Time to First Token (ms)

TPOT

Time per output token / Intra-token latency

Concurrency

# of parallel queries/users

Delivering Business Results Efficiently

Companies that can provide a service more efficiently than their competitors have an advantage

AI is inherently a scale-out workload – speed of individual AI accelerators can be overcome by using more AI accelerators

Goal is to deliver results quickly and efficiently - Performance and Cost are both important

Intel Gaudi 3 On-Prem Testing

Performance & Price-to-Performance

Signal65 Findings

With the use of AI growing at unprecedented rates, there is increasing demand for more choice by companies for hardware accelerators. On-Prem enables enterprises to securely leverage their private data to build customized solutions using that data.

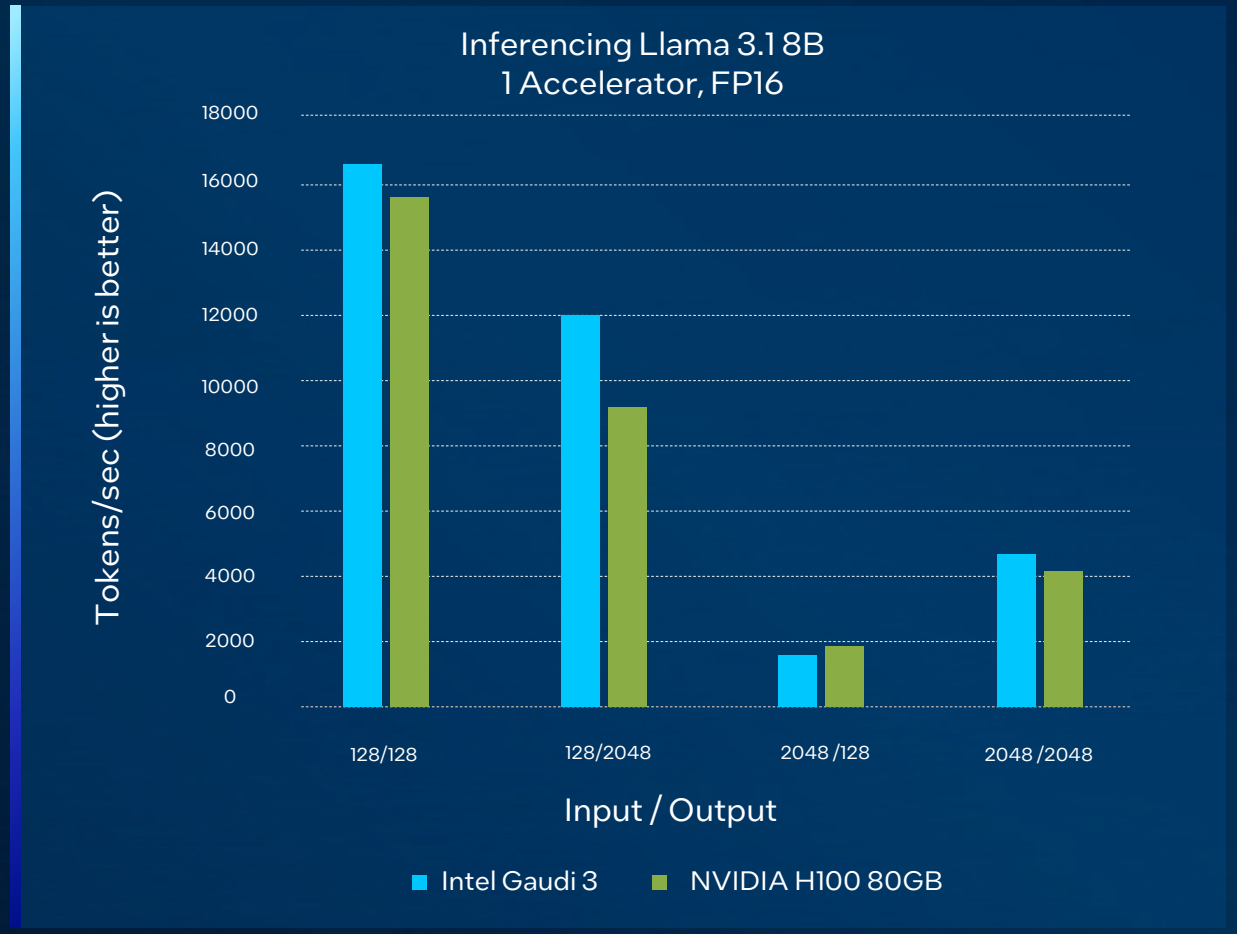
We have seen Intel's continuing focus, and a steady cadence of hardware and software releases focused on this market, and we look forward to seeing further optimizations in future planned releases.

Intel Gaudi 3 AI Accelerator – On-Prem Performance

Llama 3.1 8B Inferencing: 1 Accelerator using FP16

Competitive Inferencing Performance:

up to **30% higher** throughput vs NVIDIA H100



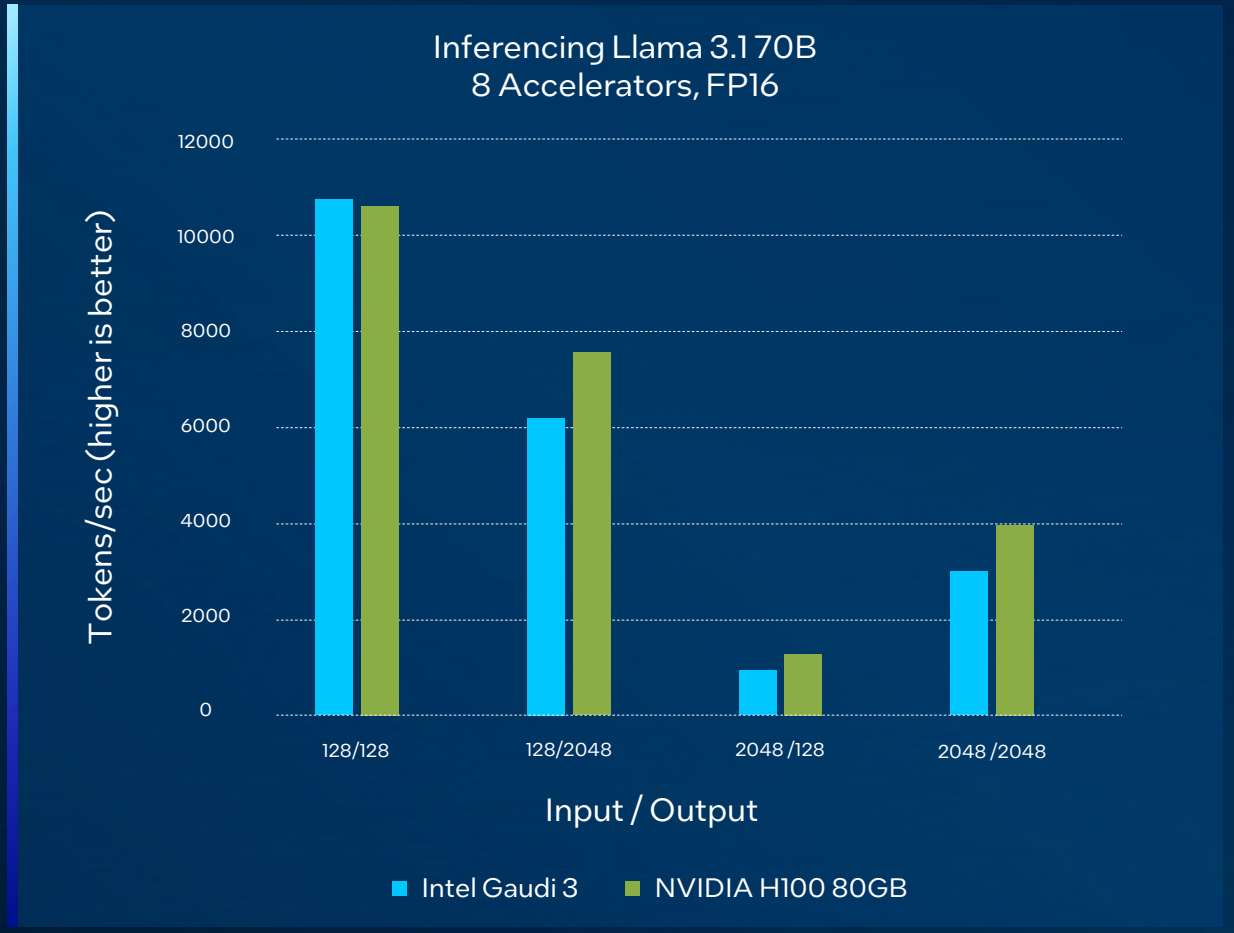
*Source: NV H100 comparison based on Signal65 Report: The New AI Accelerator Economic Lanscape, February 13, 2025.
Reported numbers are inferencing results for Llama 3.1 8B on Intel® Gaudi® 3 vs NVIDI H100 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

Intel Gaudi 3 AI Accelerator – On-Prem Performance

Llama 3.1 8B Inferencing: 1 Accelerator using FP16

- Comparison of Gaudi 3 to H100

  Token processing rate (tokens / s)
- Gaudi 3 had similar performance as the H100



*Source: NV H100 comparison based on Signal65 Report: The New AI Accelerator Economic Lanscape, February 13, 2025.
Reported numbers are inferencing results for Llama 3.1 8B on Intel® Gaudi® 3 vs NVIDI H100 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

Intel Gaudi 3 AI Accelerator

Price-to-performance Analysis

Intel Gaudi 3 AI Accelerator Pricing Details

Pricing Note

- Pricing is based on pricing of XPX-XH20 NVIDIA H100 system from Thinkmate.com, January 2025.
- Gaudi 3 Pricing based upon Intel published pricing to OEMs. January 2025.

Price Calculations	Supermicro GPX XH20	Supermicro Gaudi 3 XH20
GPU	8 x H100	8 x Gaudi 3
Full System	$300,107.00	$157,613.22
8 x GPUs	$267,493.78	$125,000.00
Base System Cost	$32,613.22	$32,613.22
Cost / GPU Only	$33,436.72	$15,625.00
System $ / GPU	$37,513.38	$19,701.65

Intel-commissioned study by Signal65, published February 13, 2025. Costs and Results may vary

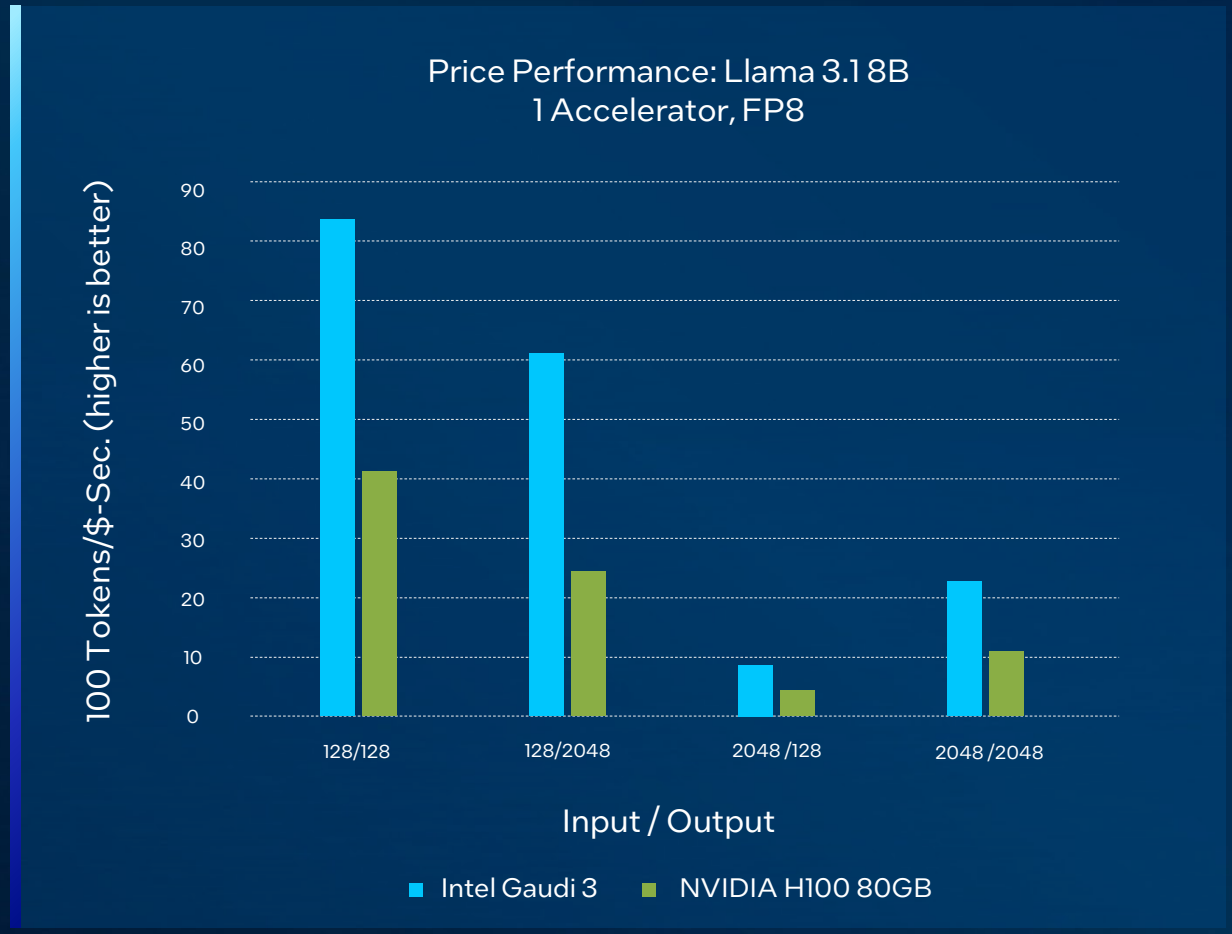
Source: Signal65 Report

Intel Gaudi 3 AI Accelerator – On-Prem Price / Performance

Llama 3.1 8B Inferencing: 1 Accelerator using FP8

Intel Gaudi 3 Outperformed NVIDIA H100 in **Every Test Case**

Price performance advantage ranged from **40%** up to **150%**



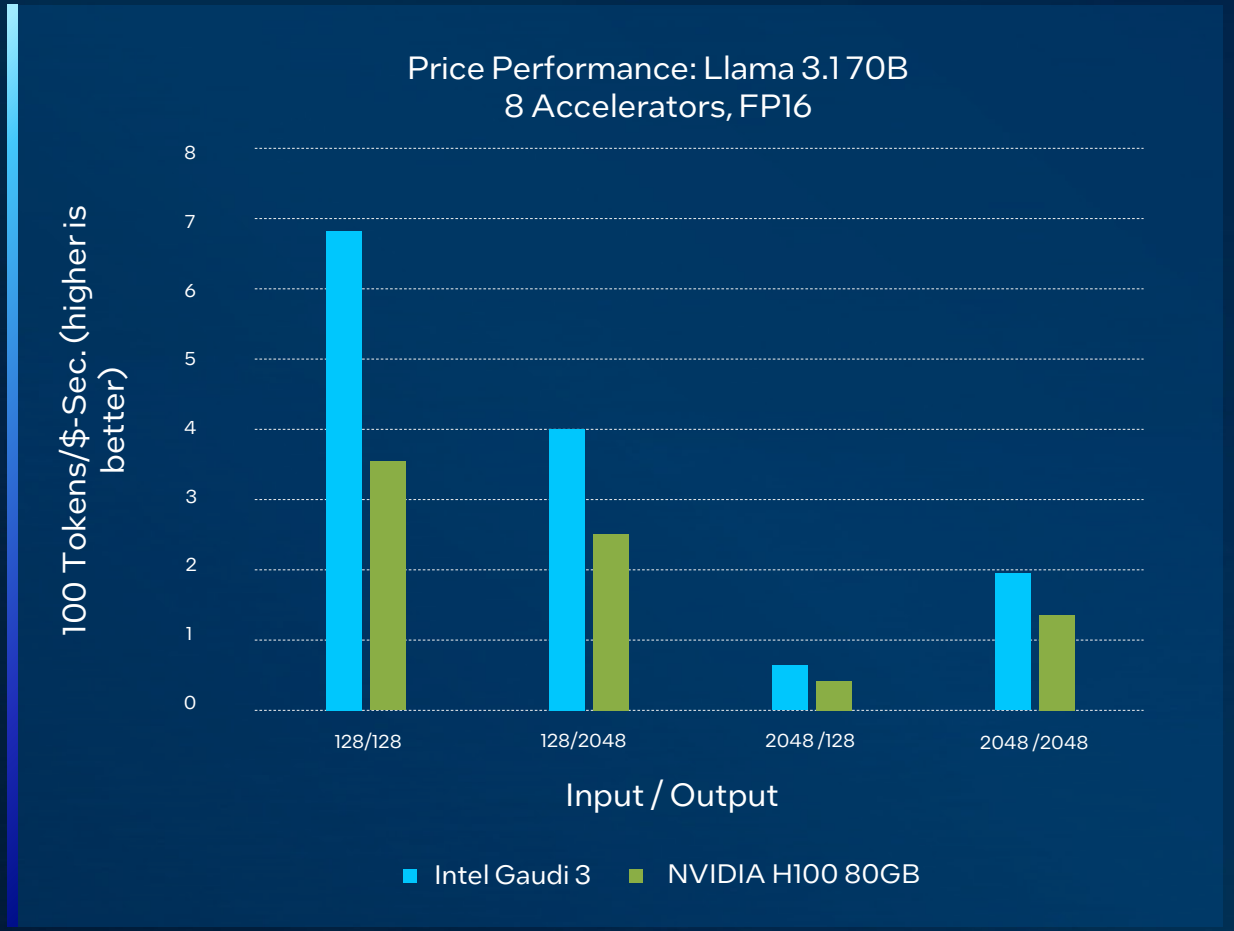
*Source: NV H100 comparison based on Signal65 Report: The New AI Accelerator Economic Lanscape, February 13, 2025.
Reported numbers are inferencing results for Llama 3.1 8B on Intel® Gaudi® 3 vs NVIDI H100 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

Intel Gaudi 3 AI Accelerator – On-Prem Price / Performance

Llama 3.1 70B Inferencing: 8 Accelerator using FP16

Intel Gaudi 3 Outperformed NVIDIA H100 in **Every Test Case**

Price performance advantage ranged from **44%** up to **92%**



*Source: NV H100 comparison based on Signal65 Report: The New AI Accelerator Economic Lanscape, February 13, 2025.
Reported numbers are inferencing results for Llama 3.1 8B on Intel® Gaudi® 3 vs NVIDI H100 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

Intel Gaudi 3 On IBM Cloud

Performance Testing
April 2025

Signal65 Findings

While testing in IBM Cloud, Intel Gaudi 3 instances showed consistently higher performance than NVIDIA H100 instances.

Additionally, Gaudi 3's performance was highly competitive with NVIDIA H200 instances.

Importantly Gaudi 3 instances were 30% less expensive than H100 on IBM Cloud.

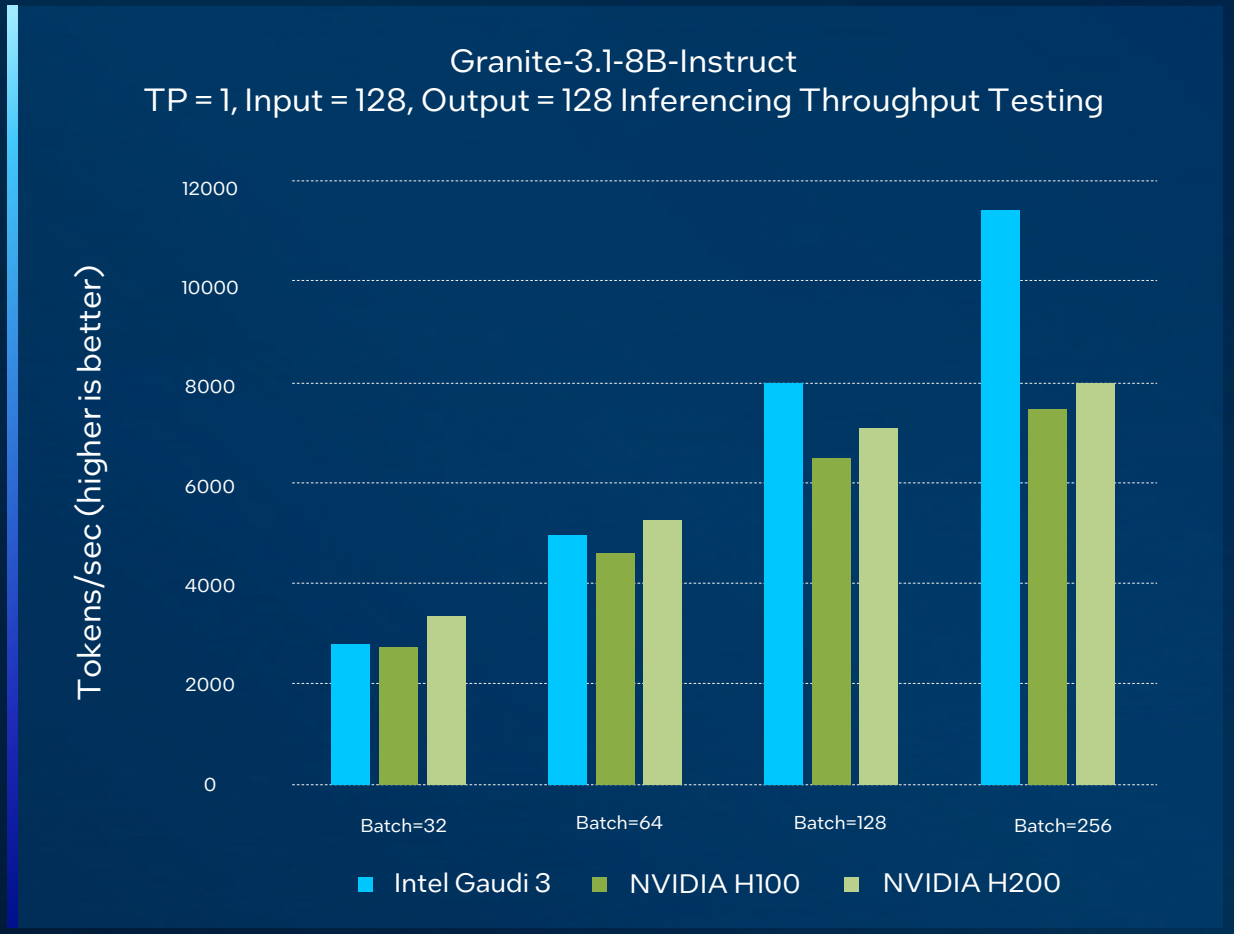
Intel Gaudi 3 on IBM Cloud

IBM Granite Performance Testing

Up to **43% higher** throughput than NVIDIA H200

Up to **52% higher** throughput than NVIDIA H100

For lightweight AI Use Cases



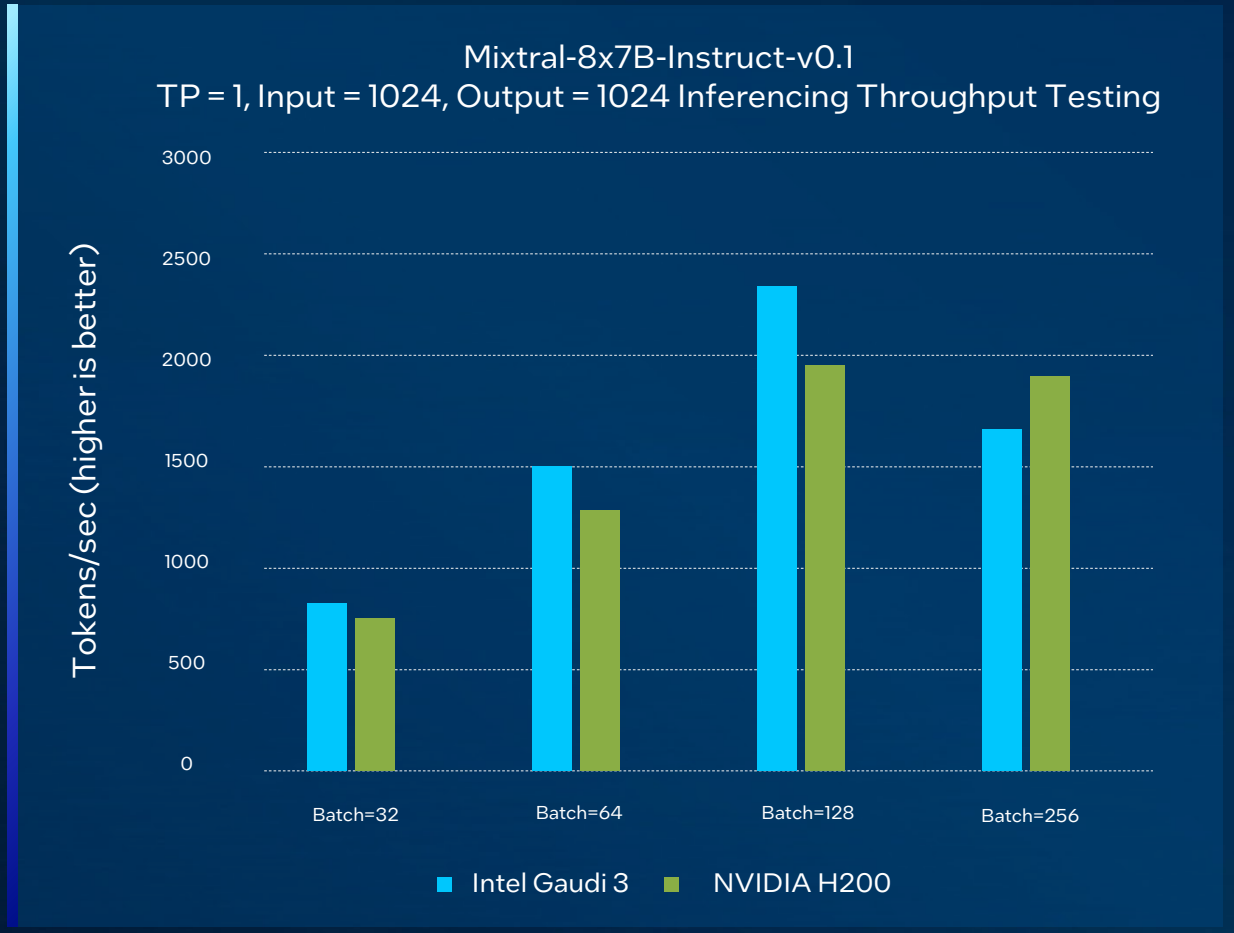
*Source: NV H100 and H200 comparisons based on Signal65 Lab Insight: Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud. April 2025.
Reported numbers are inferencing results for IBM Granite-3.1-8B-Instruct on Intel® Gaudi® 3 vs NVIDI H100 GPU and NVIDIA H200 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis.
Results may vary.

Intel Gaudi 3 on IBM Cloud

IBM Granite Performance Testing

Up to **20% higher** throughput than NVIDIA H200

For Balanced AI Use Cases



*Source: NVH200 comparisons based on Signal65 Lab Insight: Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud.April 2025.
Reported numbers are inferencing results for Mixtral-8x7B-Instruct-v0.1 on Intel® Gaudi® 3 vs NVIDIA H200 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

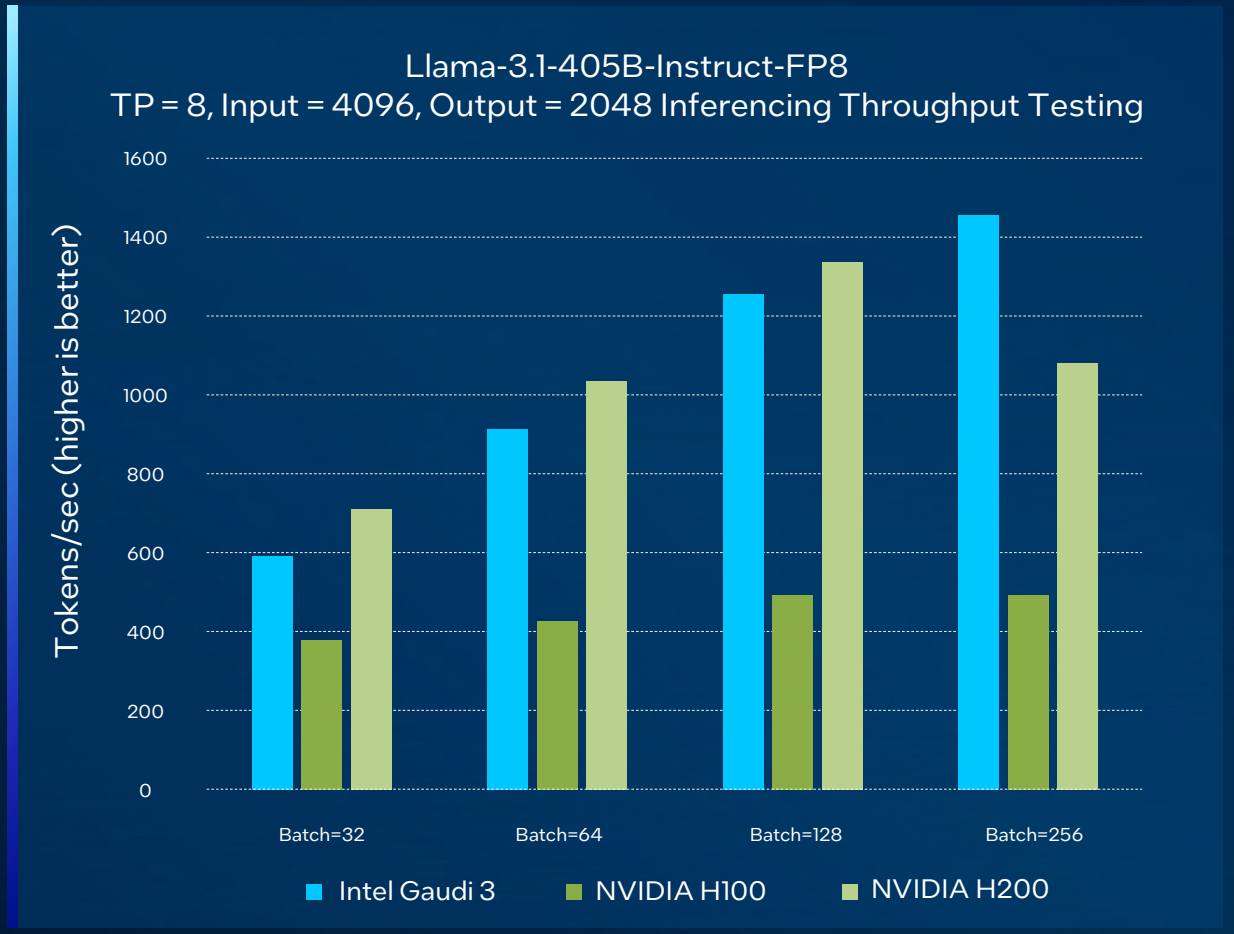
Intel Gaudi 3 on IBM Cloud

Llama Performance Testing

Up to **36% higher** throughput than NVIDIA H200

Up to **200% higher** throughput than NVIDIA H100

For Large AI Workloads



*Source: NVH200 comparisons based on Signal65 Lab Insight: Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud.April 2025.
Reported numbers are inferencing results for Mixtral-8x7B-Instruct-v0.1 on Intel® Gaudi® 3 vs NVIDIA H200 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis. Results may vary.

Intel Gaudi 3 on IBM Cloud

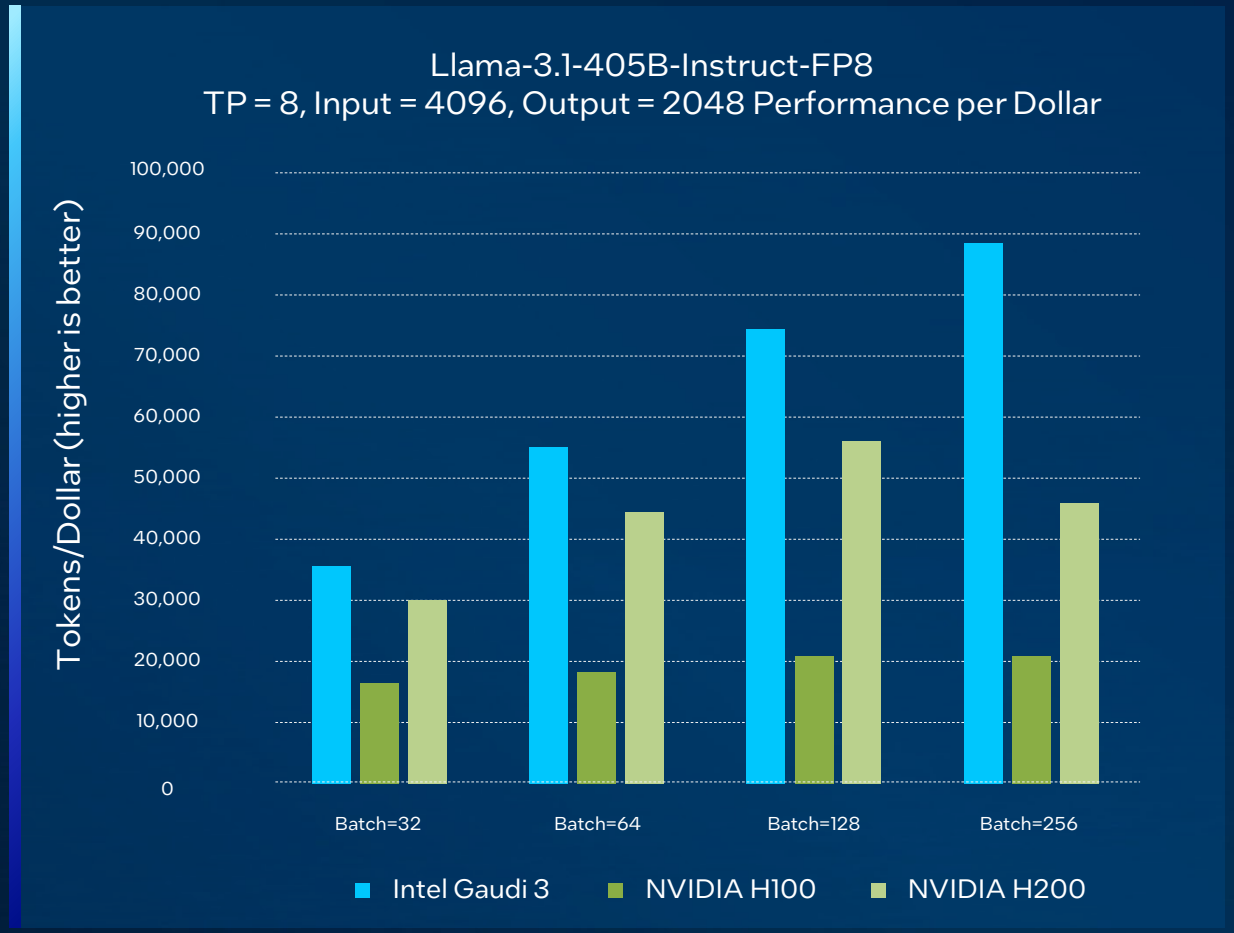
Llama Performance per Dollar

Gaudi 3 achieves more tokens per dollar

Up to **335% more** than NVIDIA H100

Up to **92% more** than NVIDIA H200

For Large AI Workloads



*Source: NV H100 and H200 comparisons based on Signal65 Lab Insight: Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud.April 2025.
Reported numbers are inferencing results for Meta Llama-3.1-405B-Instruct-FP8 on Gaudi® 3 vs NVIDI H100 GPU and NVIDIA H200 GPU. Results may vary.
Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information from IBM Cloud, accessed March 21, 2025. Results may vary.

Appendix

Data sources

Independent third-party testing

Intel Gaudi 3 On-Prem Testing	Intel Gaudi 3 on IBM Cloud Testing
https://signal65.com/research/ai/the-new-ai-accelerator-economic-landscape/	https://signal65.com/research/ai/intel-gaudi-3-accelerates-ai-at-scale-on-ibm-cloud/

Intel Gaudi 3 on IBM Cloud

Test Configurations

	Gaudi 3 on IBM Cloud	NVIDIA H100 on IBM Cloud	NVIDIA H200 on IBM Cloud
Operating Environment			
OS	Ubuntu 22.04	Ubuntu 22.04	Ubuntu 22.04
Accelerator Drivers	Habana 1.20.0-543	Nvidia 570.124.06	Nvidia 570.124.06
Runtime Environment			
Python version	3.10	3.10	3.10
PyTorch	2.6.0+hpu_1.20.0-543.git4952fce	2.5.1+cu124	2.5.1+cu124
Inferencing Server	VLLM v0.6.6.post1	VLLM v0.6.6.post1	VLLM v0.6.6.post1

All testing outlined in this report was completed between March 20th and April 11th, 2025. Pricing information used in this report was sourced from IBM Cloud, accessed on March 21, 2025. Pricing for each IBM Cloud instance discussed in this report can be found here:

NVIDIA H200 | NVIDIA H100 | Intel Gaudi 3

Intel® Gaudi® 3 AI Accelerator: Delivering Price Performance Advantage

Up to
43%

Higher throughput
(tokens per second)

on IBM Granite-3.1-8B-Instruct
vs. NVIDIA **H200**
with small context sizes

Up to
335%

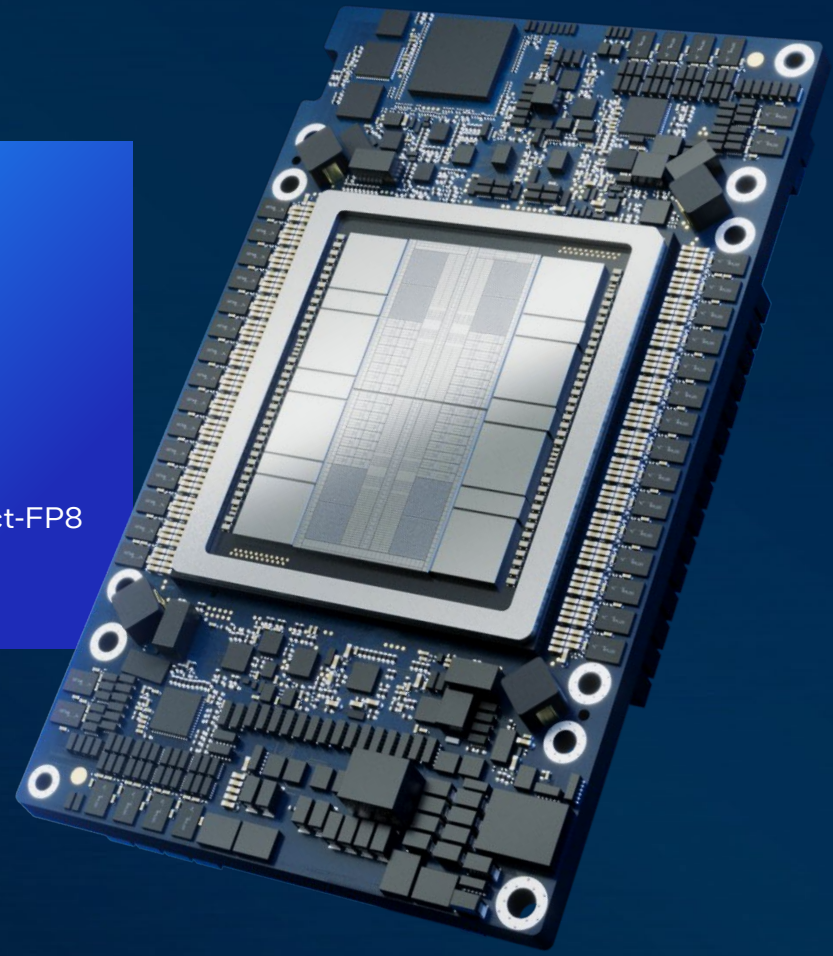
More cost efficient
(tokens per dollar)

on Llama-3.1-405B-Instruct-FP8
vs. NVIDIA **H100**
with large context sizes

Up to
92%

More cost efficient
(tokens per dollar)

on Llama-3.1-405B-Instruct-FP8
vs. NVIDIA **H200**
with large context sizes



Source: Signal65 Lab Insights Report - Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud, Intel-commissioned study by Signal65, published April 2025. See Signal65 report source for workloads and configurations. Results may vary.

Legal Notices and Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:


intel