signal65

SIGNAL65 LAB INSIGHT

Intel® Gaudi® 3 AI Accelerator at Scale on IBM Cloud

AUTHOR

Mitch Lewis
Performance Analyst | Signal65

IN PARTNERSHIP WITH

intel.

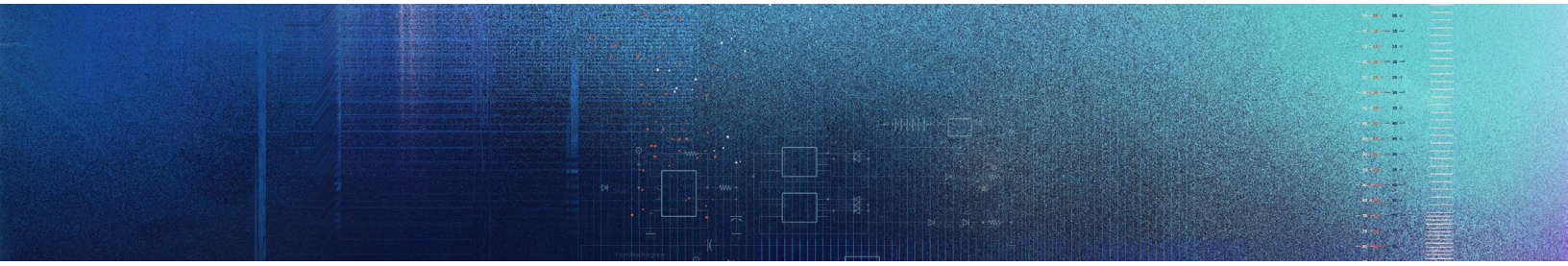
APRIL 2025

Executive Summary

Over the last few years, generative AI has demonstrated its immense potential as a revolutionary technology. AI-powered applications have demonstrated the ability to enhance automation, streamline workflows, and rapidly increase innovation. Further, the technology has proven to be broadly applicable, with opportunity for the creation of new, intelligent applications across virtually every industry. While the value of generative AI is apparent, the powerful hardware required to run such applications is often a barrier. As AI is increasingly moving from an experimental trend to the backbone of real world applications, IT organizations are challenged with balancing the necessary performance with economic considerations of AI hardware, and doing so at scale.

This paper outlines how Intel Gaudi 3 AI accelerators hosted on IBM Cloud can assist organizations in overcoming these challenges, and further, evaluates the performance and economics of Gaudi 3 compared to other leading solutions available on IBM Cloud. To evaluate performance, Signal65 conducted comprehensive AI inference testing utilizing multiple Large Language Models (LLMs) running on Intel Gaudi 3, NVIDIA H100, and NVIDIA H200 IBM Cloud instances. Key findings of this analysis include:

- Up to 43% more tokens per second than NVIDIA H200 when running IBM Granite-3.1-8B-Instruct for small AI workloads.
- Up to 20% more tokens per second than NVIDIA H200 when running Mixtral-8x7B-Instruct-v0.1 for balanced AI workloads.
- Up to 36% more tokens per second than NVIDIA H200 when running Llama-3.1-405B-Instruct-FP8 with large context sizes.
- Up to a 120% increase in tokens per dollar than NVIDIA H200 when running Mixtral-8x7B-Instruct-v0.1 and up to 92% more tokens per dollar than NVIDIA H200 when running Llama-3.1-405B-Instruct-FP8.
- Up to a 335% increase in tokens per dollar compared to NVIDIA H100 when running Llama-3.1-405B-Instruct-FP8.



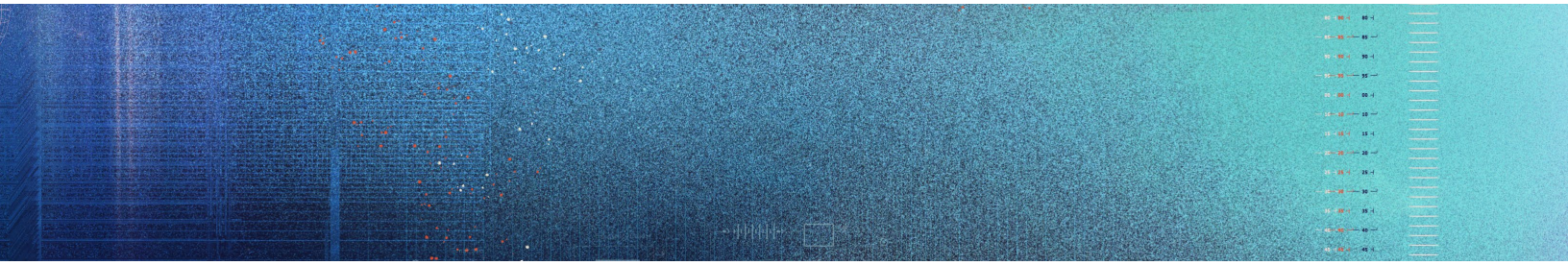
Challenges of Deploying AI at Scale

The field of AI has undergone rapid advancement in recent years with the emergence of generative AI and LLMs. The broad applicability of the technology has made AI a strategic priority for organizations across industries. While early adoption of AI has been largely experimental, consisting of small scale deployments and proof-of-concepts, organizations are increasingly seeking to deploy large scale, production ready AI applications.

Deploying AI at scale, however, creates a significant challenge for IT organizations. Meeting the performance requirements of AI inferencing workloads typically requires GPUs or other specialized AI accelerators. Due to the high demand for AI driven solutions, such hardware is often costly and difficult to obtain, impacting both the timelines and budgets of IT organizations.

A previous Signal65 report exploring the performance and economics of AI hardware found Intel Gaudi 3 AI accelerators to achieve competitive performance to NVIDIA GPUs while offering notable economic benefits. For enterprise organizations deploying AI inferencing workloads, Gaudi 3 offers a compelling solution to balance both performance and cost considerations.

Meanwhile, public cloud environments offer enterprises a solution to quickly access and easily scale AI infrastructure, however the high cost of cloud hosted GPU instances continues to be challenge. A recent collaboration between Intel and IBM has made IBM Cloud the first cloud provider to support Intel Gaudi 3 accelerators, changing the landscape of cloud hosted AI infrastructure and presenting enterprises with the flexibility to leverage the economic benefits of Gaudi 3 in the cloud.



Testing Overview

To evaluate the competitive performance and economic positioning of AI accelerators available on IBM Cloud, Signal65 conducted a series of AI inferencing tests on Intel Gaudi 3, NVIDIA H100, and NVIDIA H200 instances on IBM Cloud. To represent a broad range of enterprise AI use cases, testing was conducted across three distinct models and various input size, output size, and batch size configurations. All testing was completed using vLLM as an inferencing server, due to its high performance and broad model and hardware compatibility. In addition, Signal65 has noted increasing adoption of vLLM by AI developers, making it well suited for this performance evaluation.

Models tested included:

- IBM Granite-3.1-8B-Instruct
- MistralAI Mixtral-8x7B-Instruct-v0.1
- Meta Llama-3.1-405B-Instruct-FP8

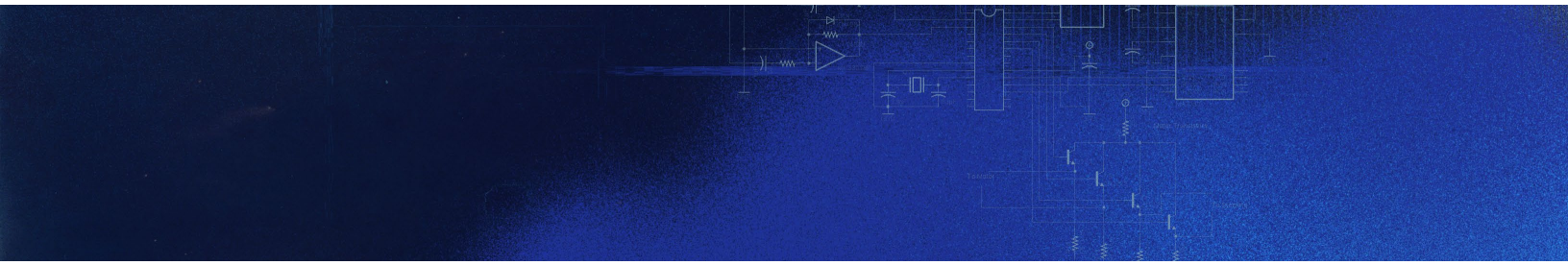
The chosen models represent a selection of three popular open source models, covering a wide range of sizes, architectures, and potential use cases.

For each model, an array of input and output token context size combinations were tested, to evaluate performance of different AI workload scenarios. An overview of input and output token sizes tested can be seen below:

	Input Size	Output Size	Workload Examples
Short Input / Short Output	128	128	Text classification, short question and answer chat applications
Medium Input / Medium Output	1024	1024	Standard chat applications, code generation
Long Input / Short Output	2048	128	Summarization or classification of articles, emails, or full length documents
	4096	128	
Long Input / Long Output	2048	2048	Multi-turn chat applications with long context windows, RAG, complex code generation, full length document translation
	4096	2048	

Figure 1: *Input / Output Sizes*

Testing for each model and context size combination was completed at various batch sizes spanning from 32 up to 256, using a metric of tokens per second to measure the throughput achieved by each solution. In general, tokens per second typically increase alongside batch sizes, however, increased latency can also become a limiting factor for enterprise applications. While larger batch sizes are commonly used in model training, batch sizes beyond 256 are typically less relevant for real-time or latency sensitive AI inferencing applications. By scaling up to a batch size of 256, this testing aimed to provide realistic performance metrics across a range of possible enterprise use cases, while still stressing system throughput capabilities.



Performance Testing Results

Granite

IBM Granite-3.1-8B-Instruct is a relatively small model created by IBM with enterprise use cases in mind. As an 8 billion parameter model, Granite-3.1-8B-Instruct fits comfortably within the memory limits of modern AI accelerators, enabling enterprises to deploy single-card inferencing. To evaluate a realistic enterprise scenario, testing was completed on each system with a single card (TP = 1).

As a relatively small model, Granite-3.1-8B-Instruct is easily deployable and fine-tuned for enterprise specific applications. IBM Granite models have also been designed with a strong focus on data security and compliance, making them well suited for enterprise workloads leveraging internal or private data, as found in a previous evaluation conducted by Signal65.

Potential enterprise use cases for Granite align closely with several of the input and output sizes tested. The short input / short output size tested aligns well with lightweight, enterprise focused workloads, such as short-form question and answer applications, or prompt completion.

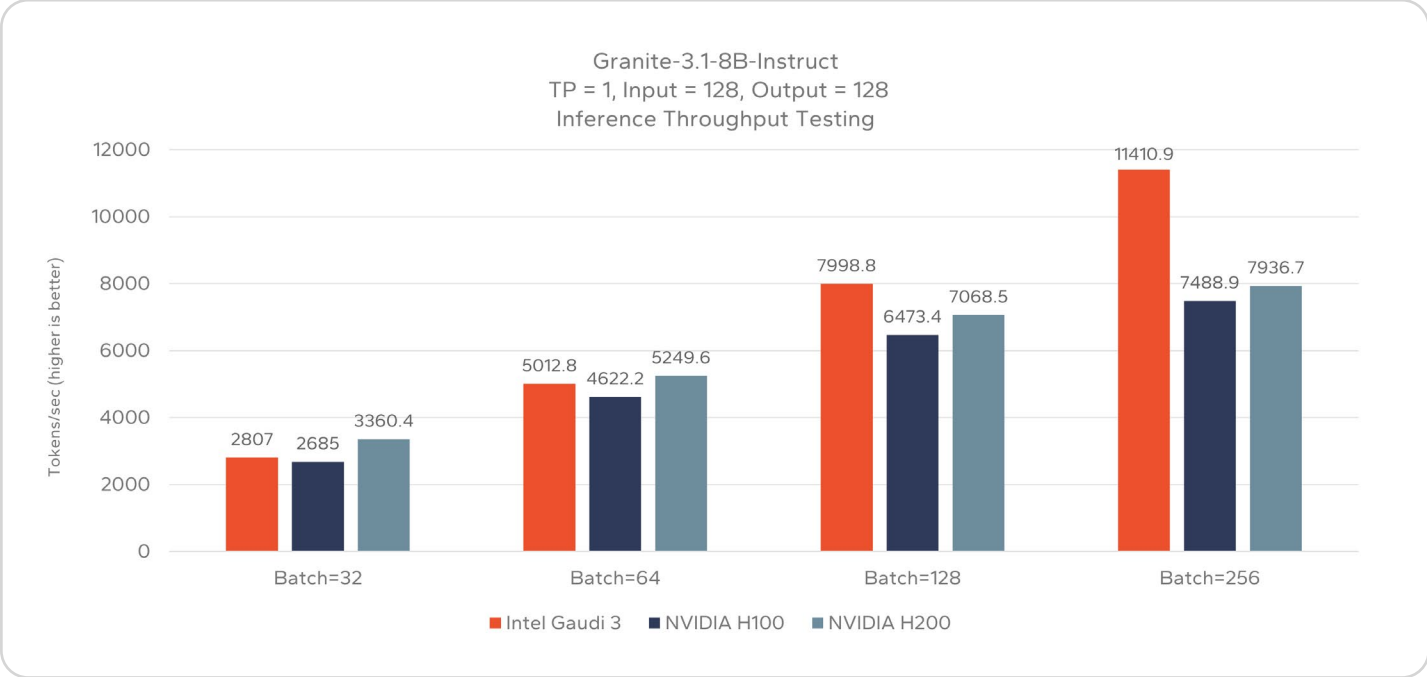


Figure 2: *IBM Granite Tokens per second vs Batch Size (short input / short output)*

When evaluating the performance results for short input / short output workloads, Intel Gaudi 3 achieved a higher tokens per second rate than NVIDIA H100 at all batch sizes tested. When compared to NVIDIA H200, Gaudi 3 achieved competitive performance at smaller batch sizes, and outperformed NVIDIA H200 at batch sizes of 128 and longer. For the largest batch size tested, Gaudi 3 achieved 43% more tokens per second than NVIDIA H200 and 52% more tokens per second than NVIDIA H100.

The medium input / medium output context length tested may represent enterprise AI applications handling mid-length enterprise documents, such as emails, reports, or log files. Testing of a medium input / medium output context length again showed Intel Gaudi 3 to have a notable performance advantage over NVIDIA H100 at all batch sizes. As with the short input / short output testing, testing of a medium context length found Gaudi 3 to outperform NVIDIA H200 at larger batch sizes.

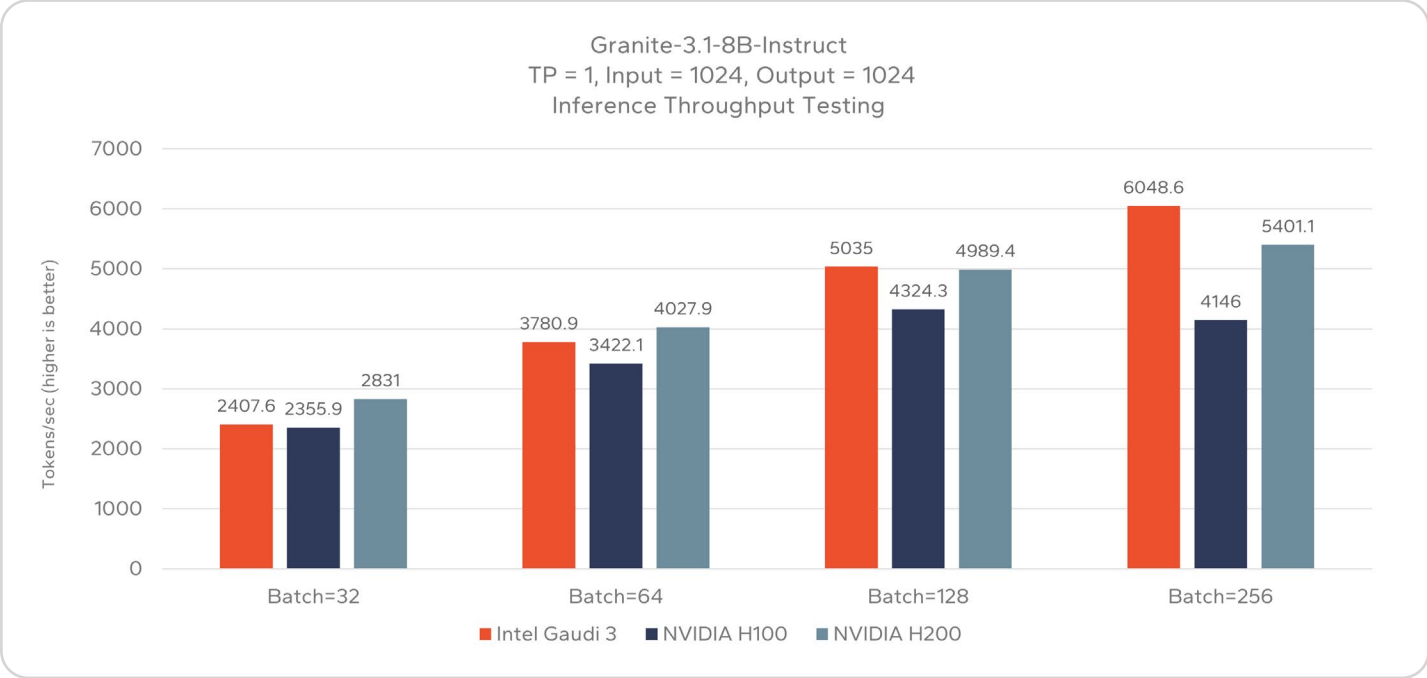


Figure 3: *IBM Granite Tokens per second vs Batch Size (medium input / medium output)*

For fields with complex documents, such as legal contracts, an enterprise focused model such as IBM Granite may be well suited for classification or summarization purposes, which can be represented by long input / short output tests.

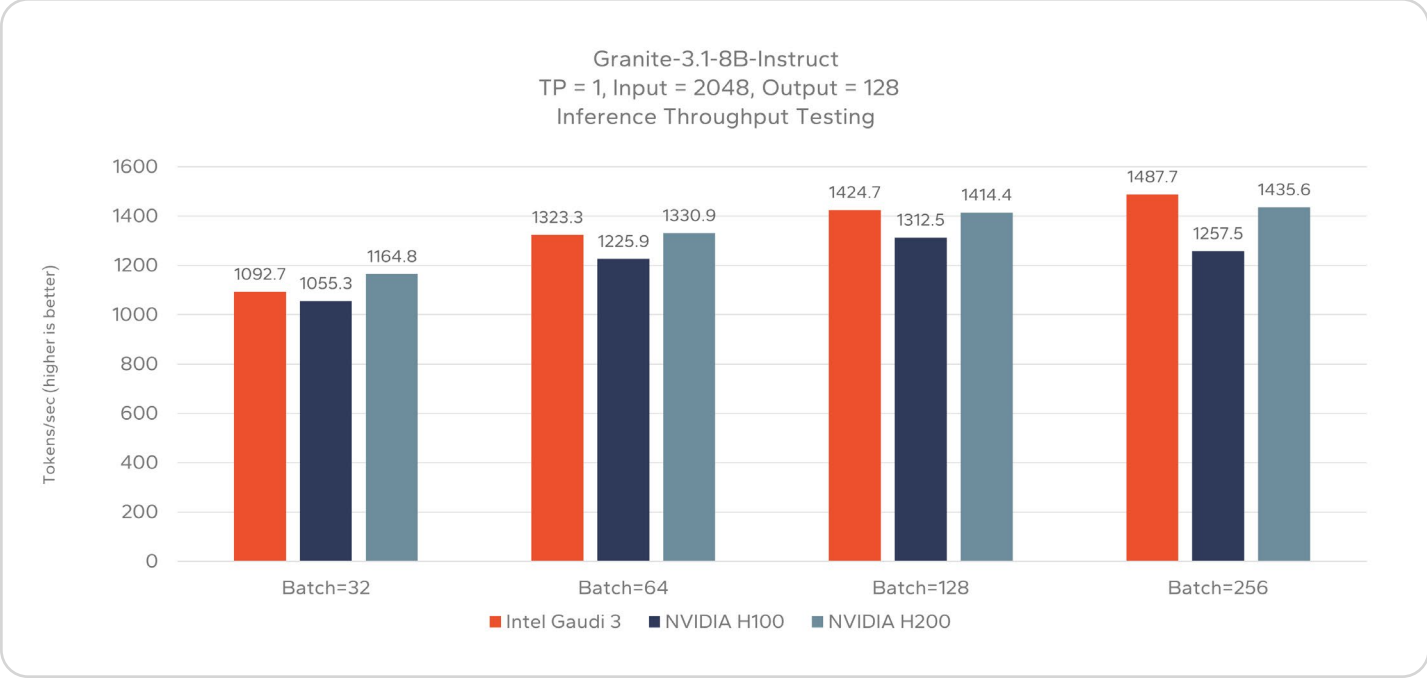


Figure 4: *IBM Granite Tokens per second vs Batch Size (long input / short output)*

When tested with a long input / short output, Gaudi 3 achieved up to 18% more tokens per second than NVIDIA H100. Compared to NVIDIA H200, performance was competitive at all batch sizes tested, ranging from within 6% at a batch size of 32 to 3.6% higher at a batch size of 256.

Mixtral

Mixtral-8x7B-Instruct-v0.1 uses a mixture of experts (MOE) architecture to balance performance and resource efficiency. The MOE architecture provides a mid-sized model capable of achieving the performance and accuracy of larger models by activating a subset of available experts. In the case of Mixtral-8x7B-Instruct-v0.1, two out of 8 experts are activated, each with 7B parameters. This makes it a compelling model for achieving specific tasks where real time inference performance is crucial, such as customer facing chat bots or translation tools.

As with Granite, testing of Mixtral was performed on a single card (TP = 1). It should be noted, however, that results for NVIDIA H100 have been omitted, as it was unable to support the model on a single card due to insufficient memory capacity.

The flexibility Mixtral-8x7B-Instruct-v0.1 makes it suitable for a wide range of tasks with medium to long context sizes. The 1024 / 1024 and 2048 / 2048 context lengths represent applications requiring a balance of both input and output, at various lengths. These context lengths can represent applications spanning a broad range of standard AI use cases including chat interactions, document editing, and code generation.

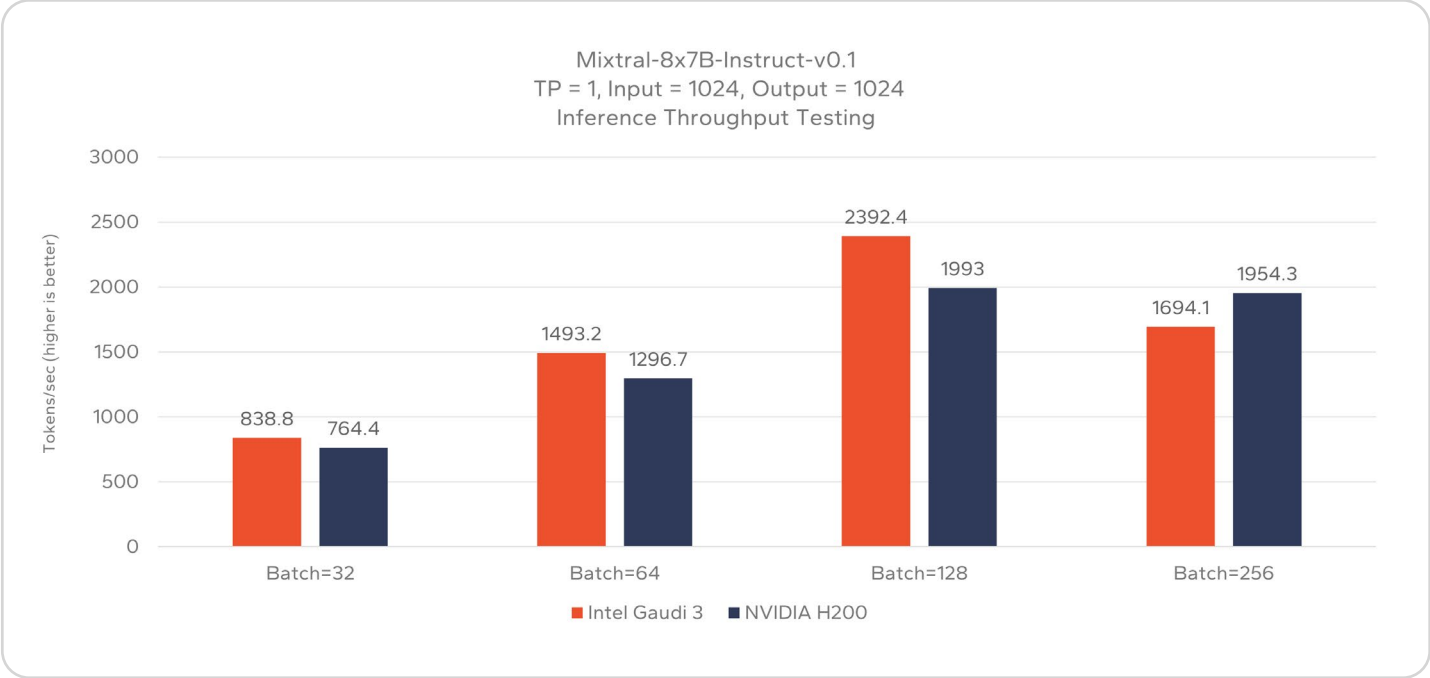


Figure 5: *Mixtral Tokens per second vs Batch Size (medium input / medium output)*

In contrast to the trend found during testing Granite, the 1024 / 1024 and 2048 / 2048 context length Mixtral inferencing tests showed Gaudi 3 to achieve more tokens per second than NVIDIA H200 at smaller batch sizes, with NVIDIA H200 achieving more tokens per second at the larger batch sizes. At the medium sized 1024 / 1024 context length, Gaudi 3 achieved 15% more tokens per second at a batch size of 64 and 20% more tokens per second at a batch size of 128.

At the larger 2048 / 2048 context length, Gaudi 3 achieved more tokens per second at batch sizes of 32 and 64, while NVIDIA H200 gained an advantage at the larger batch sizes of 128 and 256.

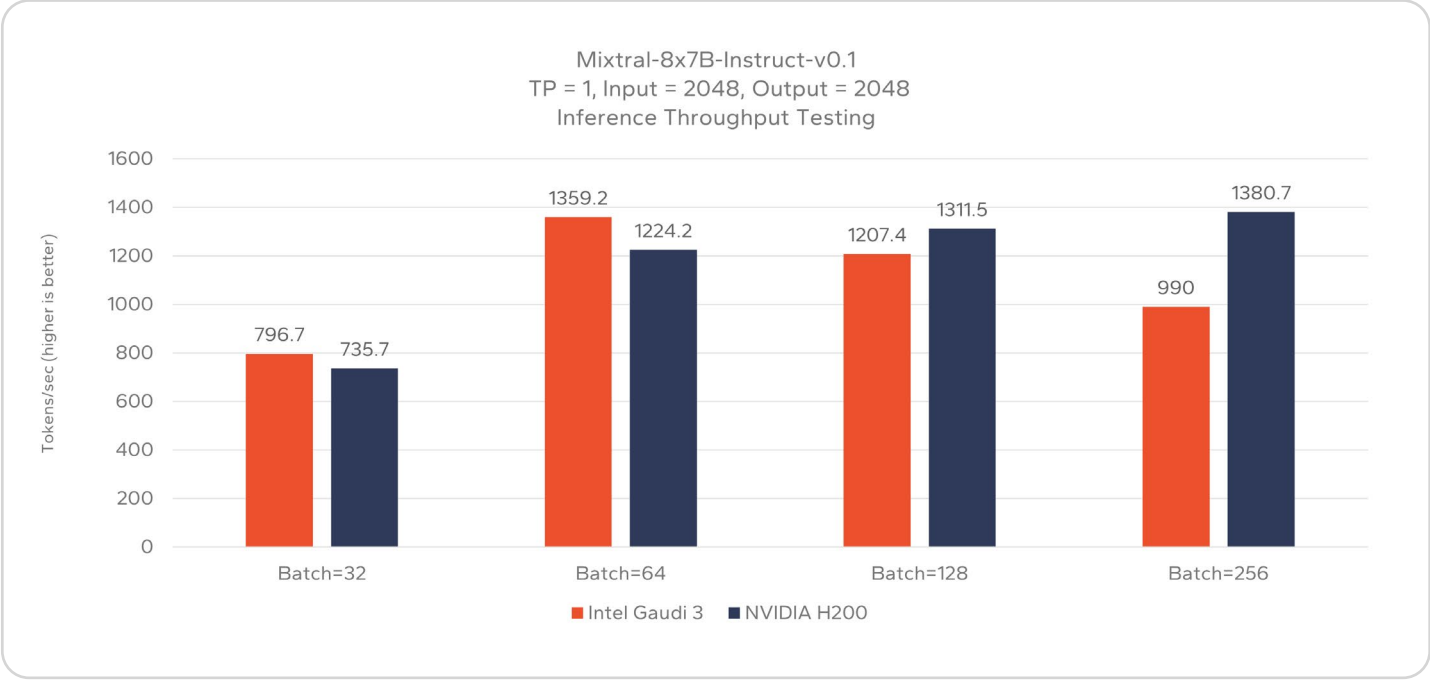


Figure 6: *Mixtral Tokens per second vs Batch Size (long input / long output)*

Mixtral's MOE approach additionally offers a model capable of handling more complex tasks, with larger context windows that may otherwise be reserved for larger, more resource intensive models. With the rapid development of generative AI, there is a growing requirement for AI workloads to handle increasingly long inputs for complex tasks, long chat histories, or inclusion of additional context from methods such as Retrieval Augmented Generation (RAG). This requirement can be evaluated with tests utilizing larger input sizes, such as 4096.

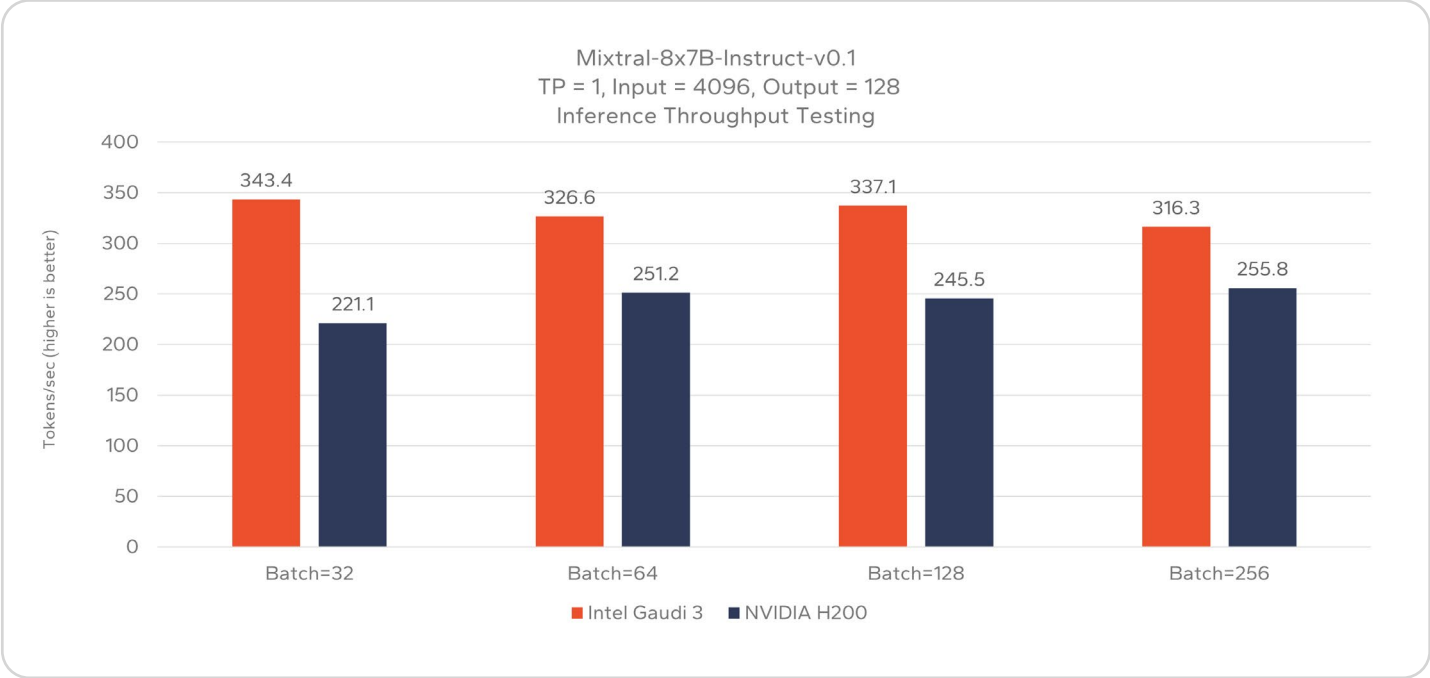


Figure 7: *Mixtral Tokens per second vs Batch Size (long input / short output)*

Inference testing with a 4096/128 context length replicates use cases in which the model is provided a long context, but only requires a relatively short result, such as a summary or a label. For Mixtral, this test showed a clear advantage for Gaudi 3, outperforming NVIDIA H200 at all batch sizes tested from 32 to 256. Gaudi 3 had its largest advantage at a batch size of 32, with a 55% increase in tokens per second, and its smallest advantage at a batch size of 256 with a 23% increase in tokens per second.

Alternatively, extended context windows with long output sizes may be required for applications providing multi-turn chat, writing assistants, or complex code generation. Performance of such applications can be evaluated through inference testing of long context lengths, such as 4096 / 2048.

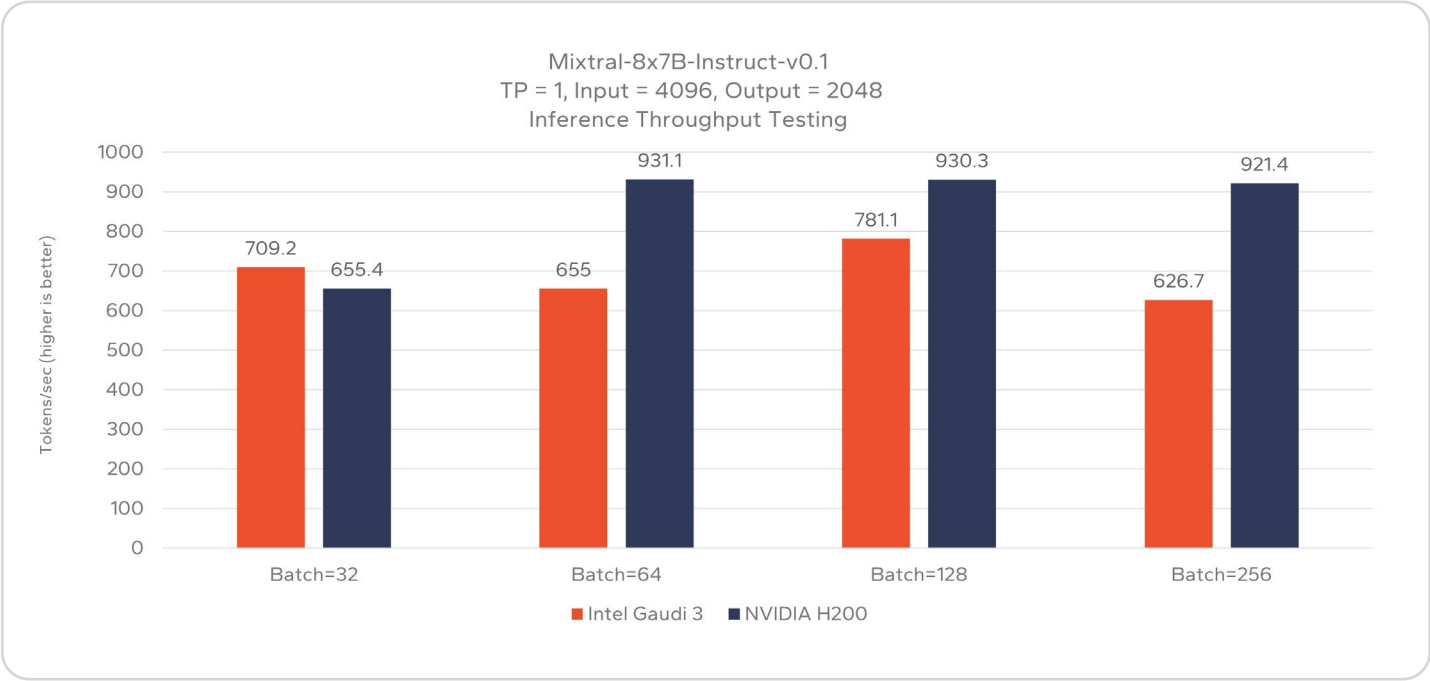


Figure 8: *Mixtral Tokens per second vs Batch Size (extended input / long output)*

Testing of Mixtral at a 4096 / 2048 context length showed Gaudi 3 to achieve a slightly more tokens per second than NVIDIA H200 at a batch size of 32, however, NVIDIA H200 demonstrated an advantage at higher batch sizes.

Llama

Llama-3.1-405B-Instruct-FP8 represents one of the largest and most accurate models currently available. While the large size requires greater resource requirements, Llama-3.1-405B-Instruct-FP8 provides state of the art performance for complex AI workloads which may include significant tool calling, reasoning, or multi-modal understanding.

Testing utilized an 8-bit floating point (FP8) quantized version of the full Llama-3.1-405B-Instruct model to optimize for performance and memory considerations. Testing in each environment was conducted across all 8 available cards (TP = 8).

While Llama-3.1-405B-Instruct-FP8 is well equipped to support large context windows, evaluation of the medium sized 1024 / 1024 context length provides a baseline of performance for applications with balanced input and output lengths.

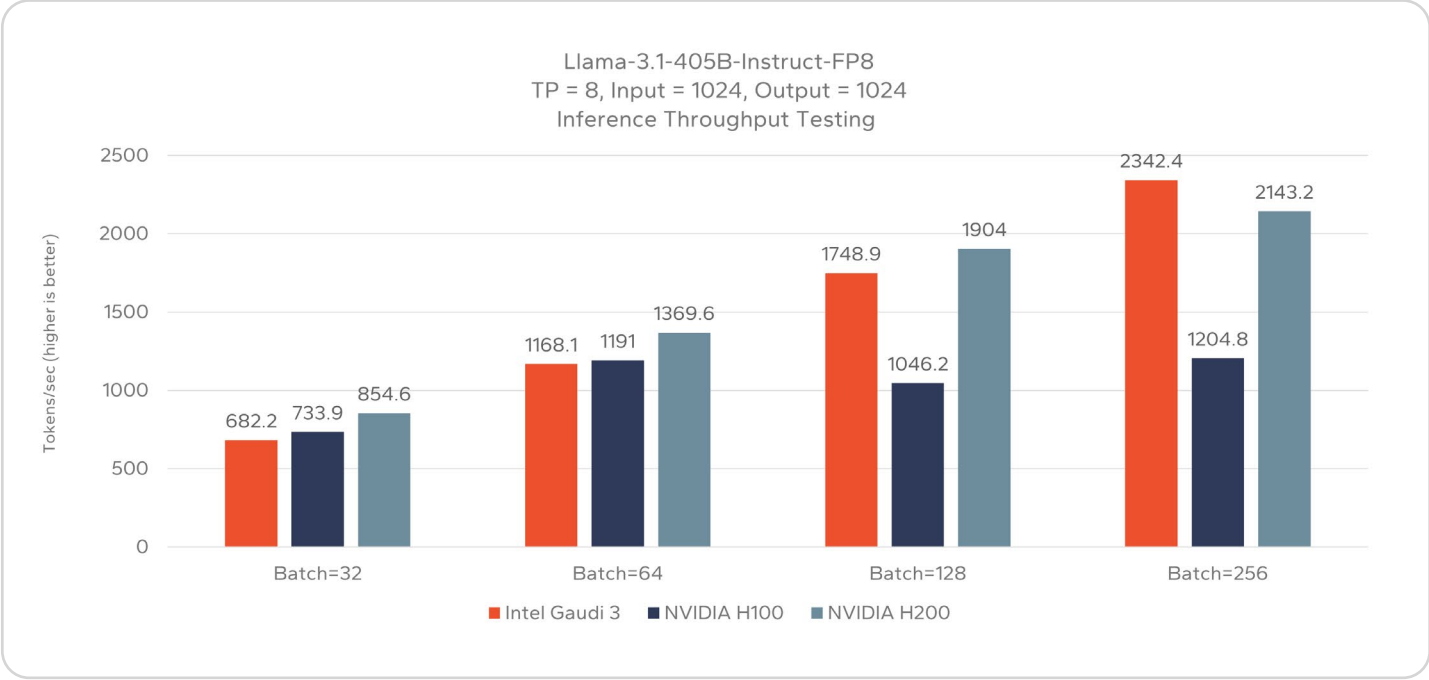


Figure 9: *Llama Tokens per second vs Batch Size (medium input / medium output)*

As can be seen in the medium input / medium output tests conducted, all three environments remained competitive at batch sizes of 32 and 64, with both NVIDIA GPUs achieving slightly more tokens per second. As the batch sizes increased, however, Gaudi 3 gained a significant advantage over NVIDIA H100, which became constrained by key-value cache memory limits. This bottleneck on NVIDIA H100 resulted in more frequent re-computation of key-value attention states, ultimately limiting the systems performance. Gaudi 3 achieved 67% more tokens per second at a batch size of 128 and 97% more tokens per second at a batch size of 256. When tested at a batch size of 256, Gaudi 3 additionally achieved an advantage compared to NVIDIA H200.

While the medium sized 1024 / 1024 context length tests provide a representation of a wide range of common AI applications, a large model such as Llama-3.1-405B-Instruct-FP8 is often selected to handle more complex use cases with very long context windows, such as RAG.

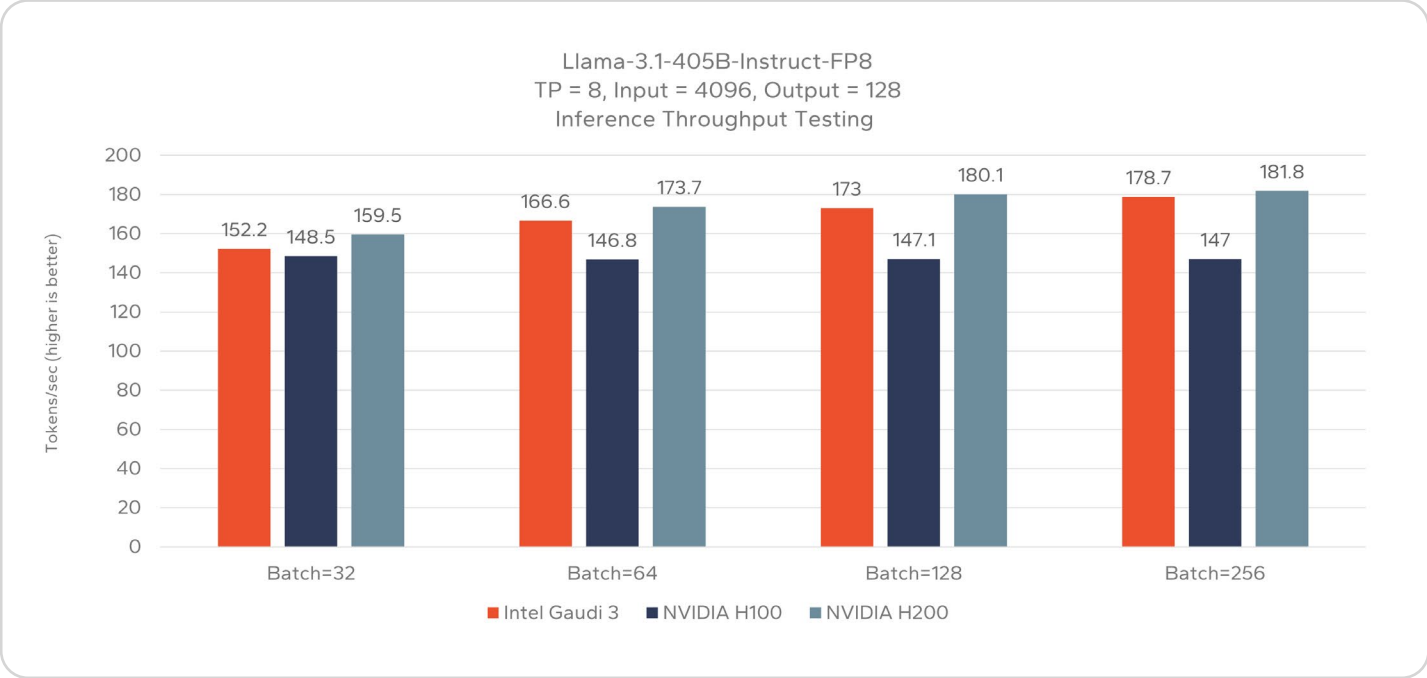


Figure 10: *Llama Tokens per second vs Batch Size (extended input / short output)*

When evaluating tests configured with a 4096 / 128 context length, representing an extended input size and a short output, Gaudi 3 can be seen to be highly competitive across all batch sizes tested. Gaudi 3 achieved more tokens per second than NVIDIA H100 at all batch sizes, while performing within 5% and 2% of NVIDIA H200.

Testing the same extended input size with large outputs showcases the performance for more complex applications, where a model like Llama-3.1-405B-Instruct-FP8 is most well suited. Testing configured with a 4096 / 2048 context length shows Gaudi 3 to achieve a significant tokens per second advantage compared to NVIDIA H100, ranging from an increase of 55% at a batch size of 32, to an increase of over 200% at a batch size of 256. Compared to NVIDIA H200, Gaudi 3 maintained competitive performance at batch sizes of 32, 64, and 128, while achieving a 36% advantage at a batch size of 256.

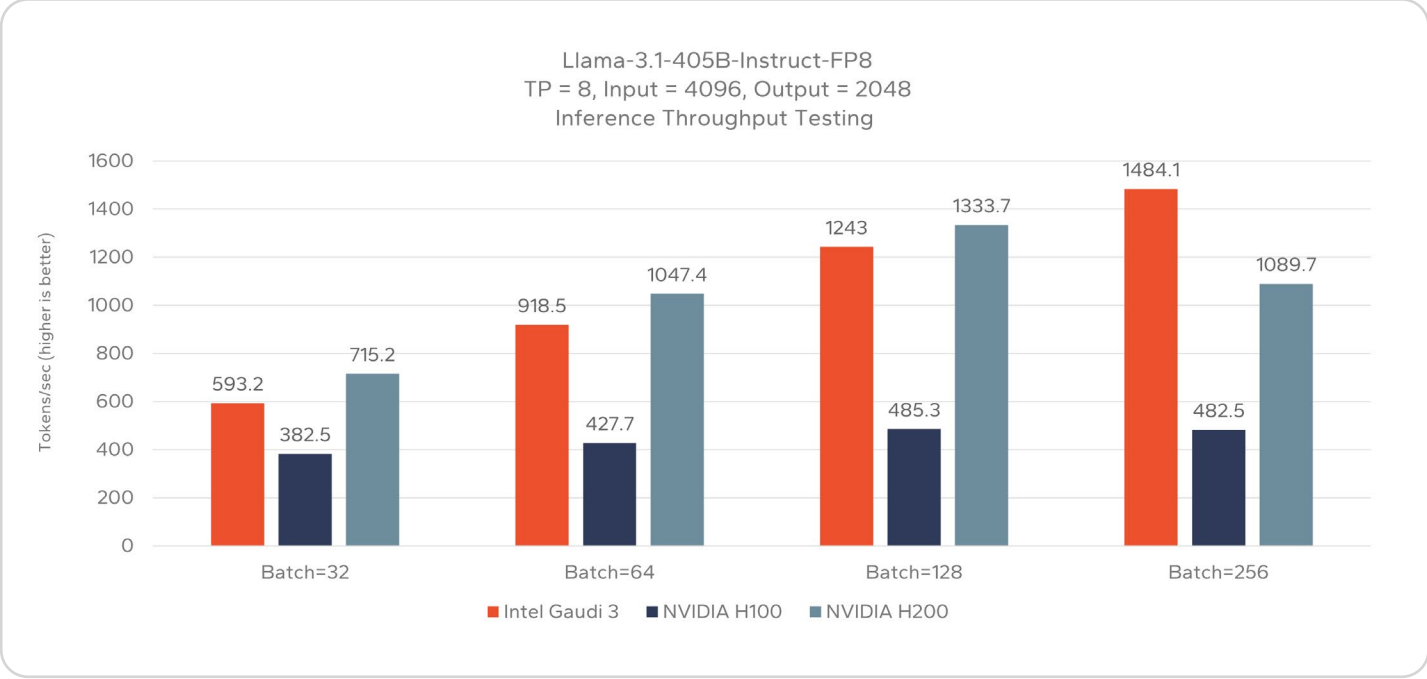


Figure 11: *Llama Tokens per second vs Batch Size (extended input / long output)*

Performance per Dollar Evaluation

Raw performance testing alone demonstrated Gaudi 3 is capable of achieving highly competitive performance compared to both NVIDIA H100 and H200 for a wide range of models and application scenarios. In general, Gaudi 3 outperformed NVIDIA H100, while comparisons to NVIDIA H200 were largely dependent on input / output size and batch size configurations. To fully evaluate the competitive value of utilizing each accelerator on IBM Cloud, however, organizations should consider the economics of each solution alongside its performance.

The pricing for each IBM Cloud instance tested, as of the time of testing, is shown below:

Device	Input Size
Intel Gaudi 3	$60
NVIDIA H100	$85
NVIDIA H200	$85

Figure 12: *IBM Cloud Pricing (IBM Cloud Pricing Accessed March 21, 2025)*

Notably, Gaudi 3 instances are available for a lower hourly price, approximately 30% less than either NVIDIA GPU instance. To further understand the economic value that each solution is capable of providing enterprise organizations, however, both price and performance should be considered together. This can be achieved using a tokens-per-dollar metric derived from the tokens per second measured combined with the hourly pricing of each IBM Cloud instance.

When evaluating the performance per dollar of Granite at the medium-sized 1024 / 1024 context length, Gaudi 3 outperformed both NVIDIA H100 and NVIDIA H200 at each batch size tested. Notably, performance testing shows Gaudi 3 achieved fewer tokens per second than NVIDIA H200 at batch sizes of 32 and 64, however, when considering the cost of each solution, Gaudi 3 achieved between 20% and 32% more tokens per dollar at these batch sizes. At a batch size of 256, where Gaudi 3 was found to have the largest performance advantage, it achieved a 58% increase in tokens per dollar compared to NVIDIA H200, and over 2x more tokens per dollar compared to NVIDIA H100.

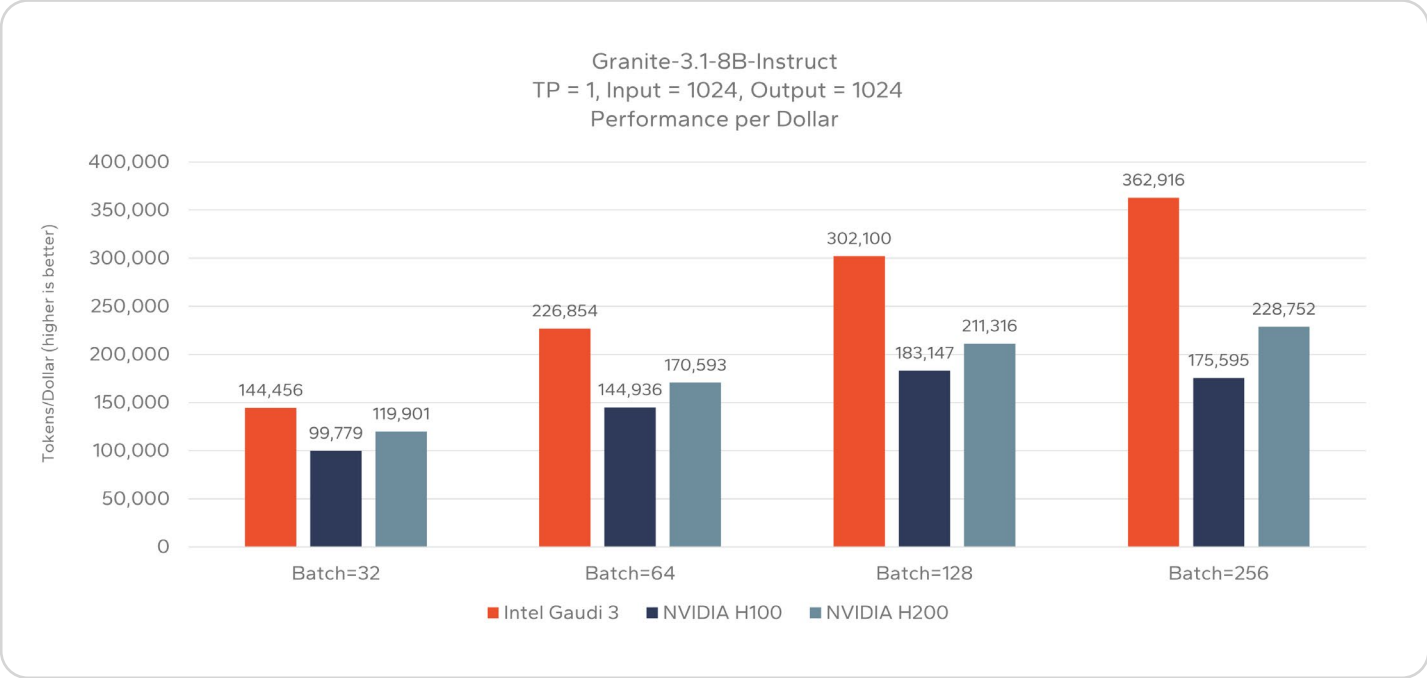


Figure 13: *Granite Performance per Dollar vs Batch Size (medium input / medium output) - IBM Cloud Pricing Accessed March 21, 2025*

When testing Mixtral for long input sizes, and short output sizes, Gaudi 3 achieved more tokens per second than NVIDIA H200 at all batch sizes. The impact of this becomes increasingly apparent when evaluating the performance per dollar metrics of this same testing, where Gaudi 3 achieved between 75% and 120% more tokens per dollar.

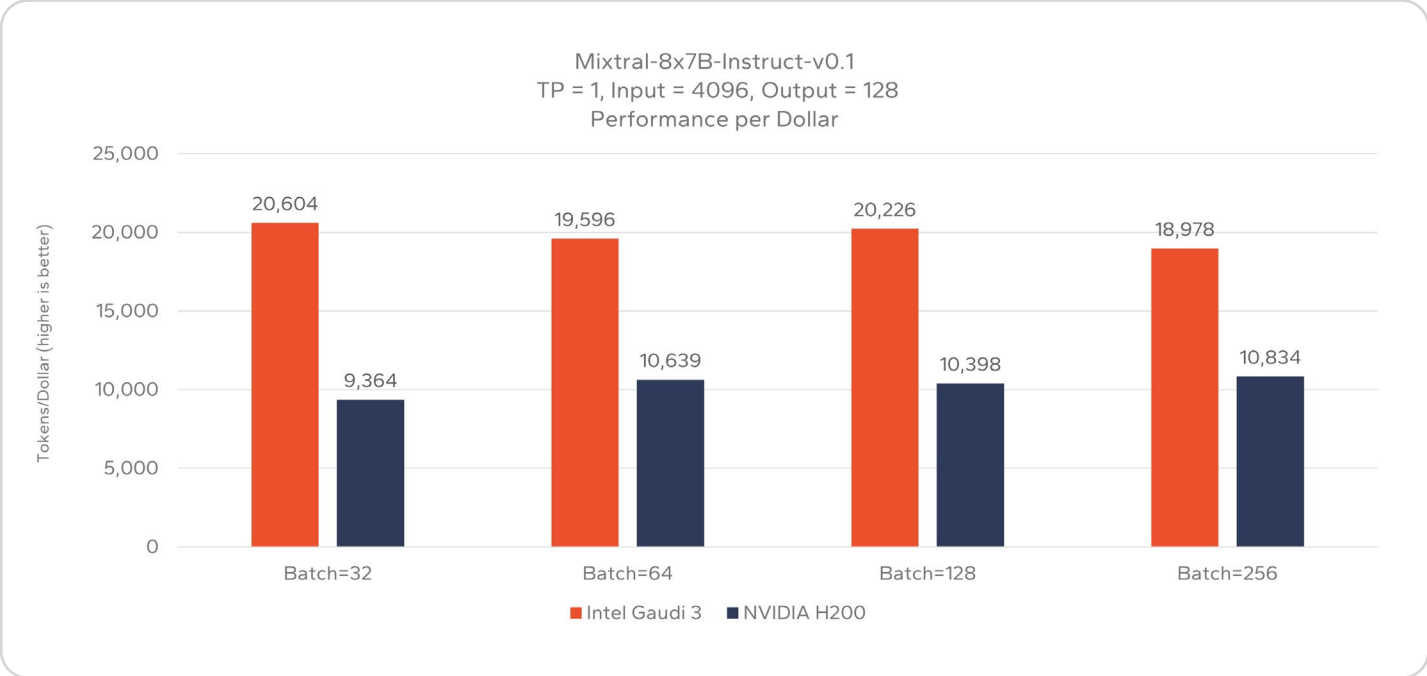


Figure 14: *Mixtral Performance per Dollar vs Batch Size (extended input / short output) - IBM Cloud Pricing Accessed March 21, 2025*

Since Gaudi 3 instances on IBM Cloud are priced at a lower hourly rate than NVIDIA GPUs, they will achieve a clear performance per dollar advantage in all scenarios where Gaudi 3 achieved the highest overall performance. More notable, however, may be evaluating the comparative performance per dollar metrics in situations in which Gaudi 3 did not achieve the most overall tokens per second. Performance testing of Mixtral with a context length of 4096 / 2048 showed Gaudi 3 to have an advantage at a Batch size of 32, while recording fewer tokens per second than NVIDIA H200 at batch sizes of 64, 128, and 256. When evaluating the comparative price performance for these same tests, however, the results are much more competitive.

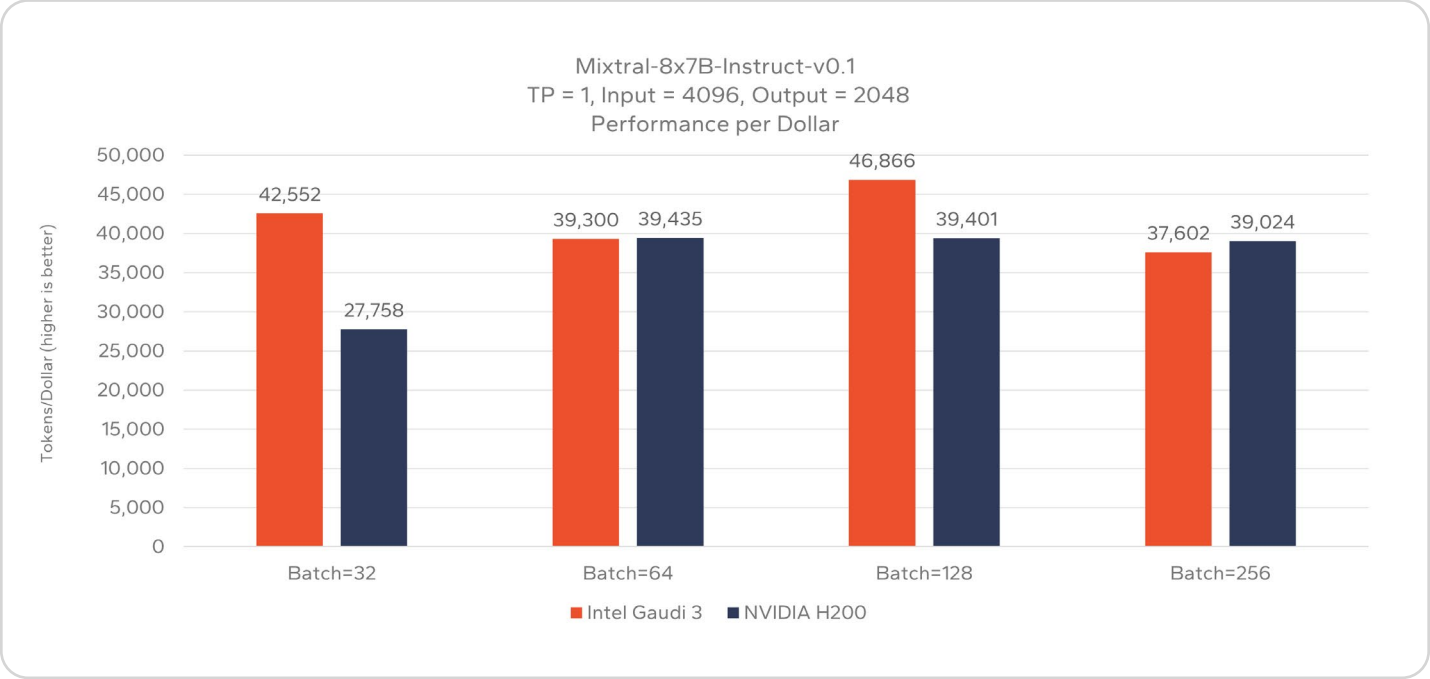


Figure 15: *Mixtral Performance per Dollar vs Batch Size (extended input / long output) - IBM Cloud Pricing Accessed March 21, 2025*

At batch sizes of 64 and 256, NVIDIA H200 was found to achieve a slight advantage, ranging from less than 1% to approximately 4% more tokens per dollar. For a batch size of 128, however, Gaudi 3 achieved nearly 19% more tokens per dollar than NVIDIA H200, highlighting how Gaudi 3 on IBM Cloud can create a significant economic advantage, even in scenarios where it was not found to have the highest overall performance.

Performance per dollar calculations additionally show a strong economic advantage for Gaudi 3 when applied to inference testing of Llama. Performance testing of Llama at a 4096 / 128 context length showed competitive results across all three solutions, with NVIDIA H200 leading. When evaluating the tokens per dollar achieved for the same test, however, Gaudi 3 significantly outperformed both NVIDIA systems, achieving between 45% and 72% more tokens per dollar than NVIDIA H100 and between 35% and 39% more tokens per dollar than NVIDIA H200.

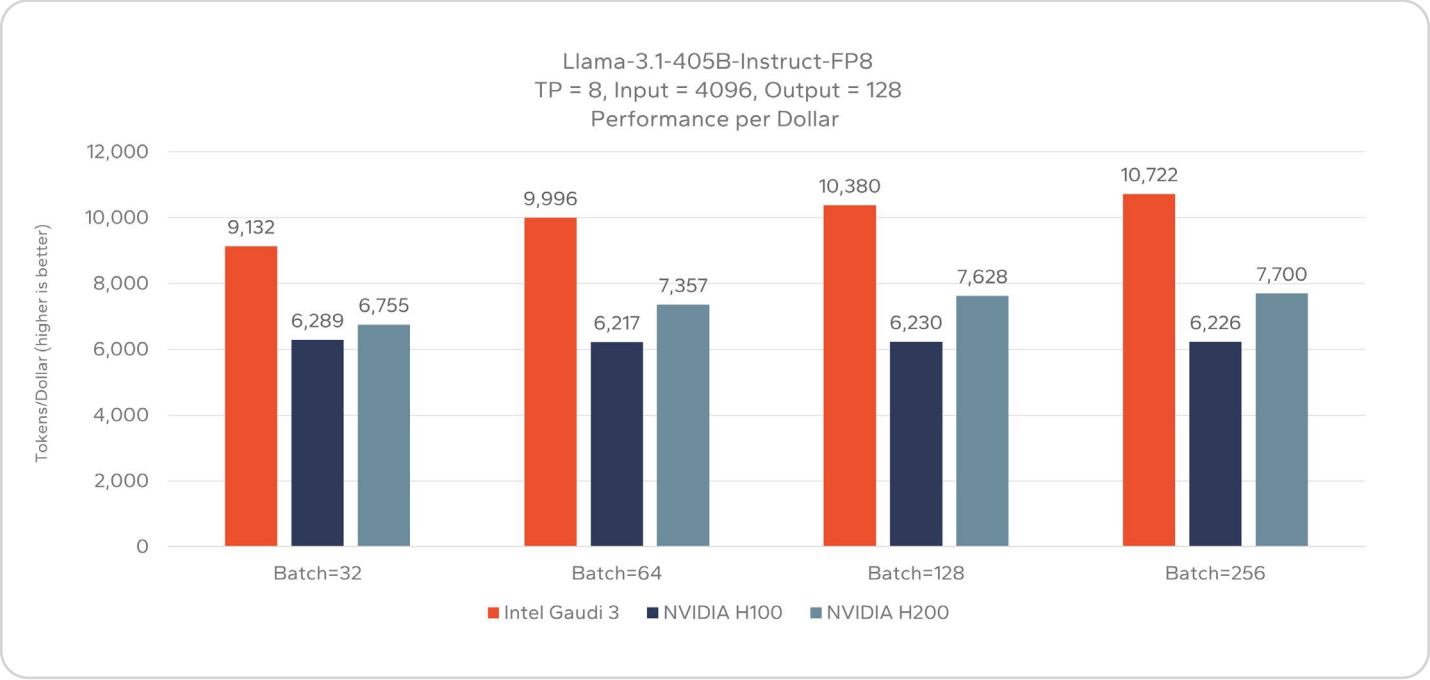


Figure 16: *Llama Performance per Dollar vs Batch Size (extended input / short output) - IBM Cloud Pricing Accessed March 21, 2025*

The long input / long output testing performed on Llama, represents some of the most computationally demanding, cutting edge AI workloads currently being deployed by enterprises. For these tests, the performance per dollar calculations clearly showcase how Gaudi 3 can enable organizations to run such workloads cost effectively.

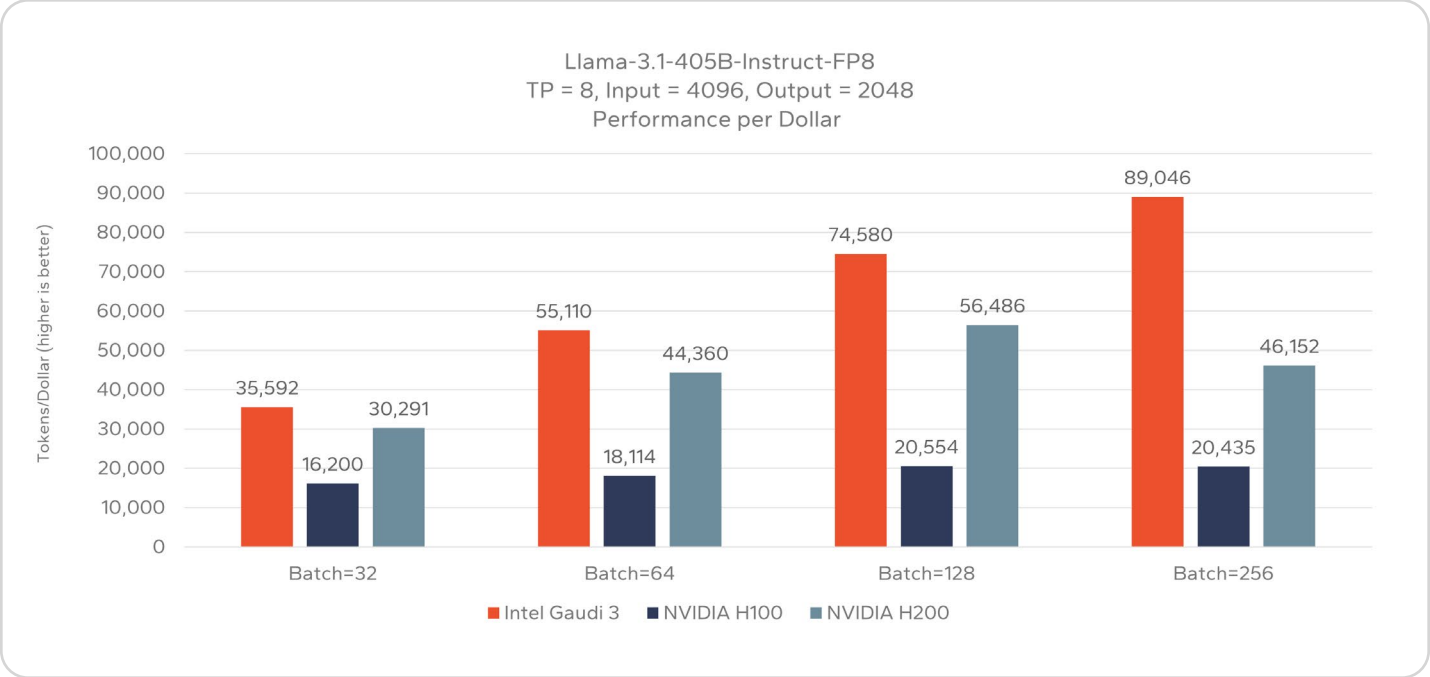


Figure 17: *Llama Performance per Dollar vs Batch Size (extended input / long output) - IBM Cloud Pricing Accessed March 21, 2025*

When evaluating tokens per second, Gaudi 3 has a large performance advantage over NVIDIA H100 for these test configurations. When evaluating the performance per dollar results, Gaudi 3 gains an even greater advantage, ranging from 119% more tokens per dollar at a batch size of 32 up to 335% more tokens per dollar at a batch size of 256. Gaudi 3 also achieved a performance per dollar advantage at each batch size over NVIDIA H200, where the raw performance results were much more competitive. Compared to NVIDIA H200, Gaudi 3 achieved between 17% and 92% more tokens per dollar.

Signal65 Evaluation

AI inference testing, and subsequent economic analysis, found Intel Gaudi 3 to be a highly competitive solution for running AI workloads on IBM Cloud. This testing highlights the broad flexibility of Gaudi 3 to support a wide range of model sizes and AI workload scenarios.

When compared to NVIDIA H100, Gaudi 3 consistently achieves a performance advantage for both Granite and Llama models across a wide range of test configurations. For Mixtral, it should be noted that Gaudi 3 provides the flexibility to run the model on a single card, which could not be achieved by NVIDIA H100.

Performance results for Gaudi 3 and NVIDIA H200 are highly competitive, with performance advantages varying between the two solutions depending on specific test configurations. While raw performance is competitive between the two solutions, Gaudi 3 was found to achieve significant economic advantages over NVIDIA H200 when considering the performance results alongside the cost of each solution on IBM Cloud.

Upon completion of this analysis, Signal65 views Intel Gaudi 3 on IBM Cloud to be an attractive offering for IT organizations who are challenged to balance the performance and cost requirements of their AI applications. Intel Gaudi 3 on IBM Cloud provides AI accelerators capable of delivering high performance inferencing across several popular LLMs, outperforming NVIDIA H100, and rivaling the performance of NVIDIA H200. When considering the pricing of each accelerator on IBM Cloud, Intel Gaudi 3 was found to offer a significant price-performance advantage over both competitive solutions when measuring tokens per dollar, enabling IT organizations to achieve more with their AI applications while spending less.

The availability of Intel Gaudi 3 on IBM Cloud enables organizations to quickly access powerful AI accelerators and scale their AI applications as needed. As enterprises increasingly transition their AI solutions from proof-of-concepts to real world applications, Intel Gaudi 3 on IBM cloud provides a high performance, cost effective platform to deploy AI at scale.

Testing Configuration

	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

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Performance Analyst | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

intel

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION
Signal65 I signal65.com