signal65

The New AI Accelerator Economic Landscape

AUTHOR

Russ Fellows
VP, Labs | Signal65

IN PARTNERSHIP WITH

intel

FEBRUARY 2025

Overview

Without question, one of the leading topics of conversation within the world of IT is that of artificial intelligence, with almost no subject having gained traction faster in the past 30 years. To further support this claim, The Futurum Group's Intelligence division surveyed over 100 CIO's from Fortune 500 and Global 2000 organizations.

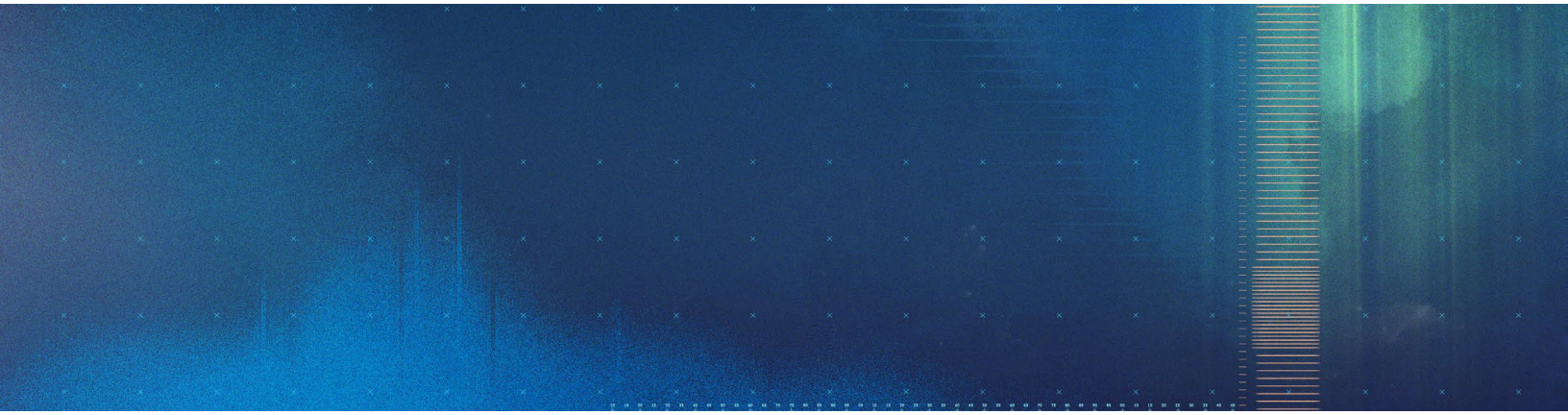
According to these results, nearly 80% of companies are in the process of conducting AI pilot projects. Futurum's research asked what challenges these IT leaders were facing, and over 50% listed adopting emerging technologies such as AI as their biggest challenge. Additionally, when examining the key issues driving their IT purchasing, issues including modernization, innovation and AI adoption were all in the top five most cited criteria.

This paper seeks to build upon our earlier analysis of the Intel® Gaudi® 2 AI Accelerator and its potential impact on enterprises. In that study we analyzed reported usage of Gaudi 2, including MLCommons results for standard AI training workloads. In this new research, we detail the results of our hands-on testing of Intel® Gaudi® 3 AI Accelerator vs. a leading competitor while running inferencing workloads on two different Llama 3.1 Large Language Models (LLMs).

To provide meaningful, real-world data points for IT and business executives, Signal65 developed an AI testing platform to run and measure AI workload performance. The Signal65 AI test suite was developed in collaboration with Kamiwaza – a commercial AI inferencing platform. The Signal65 AI test-suite leverages the Kamiwaza stack to accurately measuring AI LLM inferencing performance on various hardware and software platforms.

Key takeaways that are detailed in this report:

- Intel Gaudi 3 performance was similar to Nvidia H100 across a set of LLM inferencing tests
    - Intel Gaudi 3 ranged from 15% lower to 30% higher performance than the H100
    - Intel Gaudi 3 outperforms H100 for small inputs and large output inferencing sessions, while Nvidia outperforms Intel Gaudi 3 with large inputs and small outputs
- When considering pricing, Intel Gaudi 3 produced more work per dollar than Nvidia H100, across the same set of inferencing workloads
    - Intel Gaudi 3's advantage ranged from 10% up to 2.5x
    - Intel Gaudi 3 excelled at combinations of small inputs and large outputs, while having less advantage when processing large inputs and small outputs



The Enterprise AI Landscape

Although AI has been top of mind for companies over the past year, the vast majority of firms are still early in their journey to adoption of AI. As companies begin pilot projects, they are focused on how to leverage their corporate data and other knowledge sources to enhance existing foundational LLMs for production use.

Data privacy, governance and other items are significant concerns, and a reason why many companies are investigating the use of on-premises AI tools in addition to cloud-based solutions. Maintaining control of training and run-time inferencing data sets, along with the ability to establish guard-rails and ethical AI practices are all seen as issues that demand greater control over the data, tool chains and infrastructure utilized.

While inferencing may be accomplished with minimal hardware for single interactive sessions, deployments at scale typically require the use of hardware accelerators, particularly when using features such as Retrieval Augmented Generation (RAG) and similar techniques. As a result, companies should closely evaluate the price – performance of AI accelerators for inferencing workloads, which may dictate the overall ROI of any AI application as it moves into production.

Signal65 Comment: ***Privacy and security are top concerns for executives from companies, regardless of their size. Moreover, the ability to run AI workloads on private clouds, and do so efficiently will become a point of differentiation for companies.***

LLM Inferencing

Utilizing a LLM to produce useful results from a trained model is known as inferencing. The process of LLM inferencing typically occurs in two stages: Prefill and Decode. These stages work together to generate responses to input prompts, with both components working together.

First, Prefill converts text into an AI representation, known as a token. This tokenization process occurs on the CPU with the token sent to an AI accelerator, where an output is generated and then decoding occurs. The model continues this process iteratively, with each new token influencing the generation of the next. Finally, after this process concludes, the generated sequence is converted from tokens back into readable text.

The primary tools frameworks utilized for this process are specialized software stacks that are optimized for the inferencing process. Several examples include the open-source project vLLM, TGI from Hugging Face, along with specialized versions for specific AI accelerators. Nvidia has an optimized inferencing stack known as TensorRT-LLM. Intel also has an optimized software stack, known as Optimum Habana.

Mapping Test Cases to Enterprise Use

Signal65's testing focused on four different combinations, or workload patterns, expressed as input and output token sizes. In general, these are designed to be similar to different real-world use cases that enterprises may experience during production deployment. What is most likely is that usage would not fit any one combination, as input and output token sizes would occur across a broad range of values. However, these four combinations are designed to show likely scenarios.

In general, small token input scenarios would correspond to short input commands, without a great deal of context, such as an interactive chat. The use of Retrieval Augmented Generation (RAG) adds significant context and tokens to the input, and thus using RAG for chat sessions results in long input tokens with small output token count. Using RAG for content creation, or iterative refinement of a document or programming code would result in long input and output token workloads.

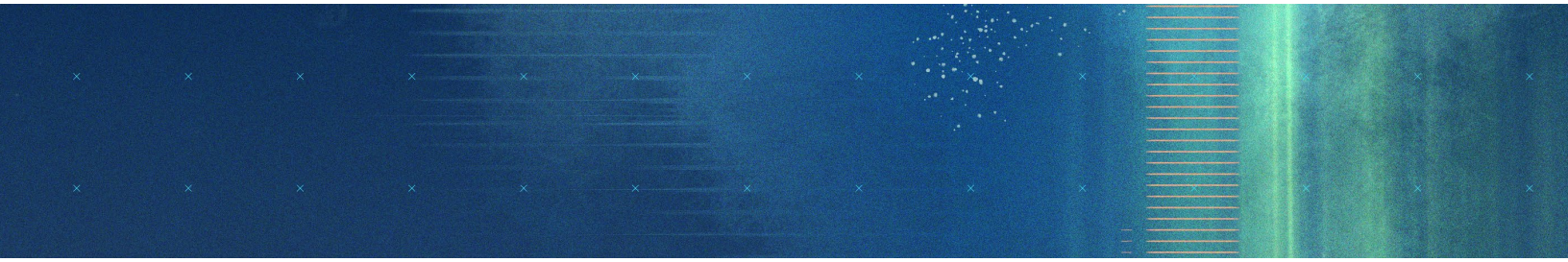
Signal65's analysis of common scenarios shows that longer context input and output combinations are the most likely scenario, and chat without RAG being the least likely. The other two scenarios together would make up the other use cases. The expected percentages are based upon discussions with clients, along with our own use of LLMs.

Input / Output Token Size	Expected % of Workload	Workload Scenario Example
128 / 128	10%	Interactive Chat Session, without RAG.
128 / 2048	25%	Content generation, e.g. document or code generation
2048 / 128	25%	Content review, with brief analysis only
2048 / 2048	40%	Interactive content creation w/ modification and review

Table 1: *Inferencing Workload Types and % of Totals (Source Signal65)*

As seen above in Table 1, the two scenarios with longer outputs would together make up 65% of the total usage, while the two scenarios with shorter outputs would comprise the remaining 35%. This is important, due to the fact that overall Intel Gaudi 3 outperforms the Nvidia H100 with larger output token workloads.

Moreover, for the most common workloads expected in enterprises, the Gaudi 3 accelerator shows performance advantages compared to the Nvidia H100. Detailed results for each of these workloads, along with the corresponding price / performance results will be presented.



AI Inferencing Test Review

In order to process input and submit data to AI accelerators efficiently, the inferencing software creates tokens from input data and then sends those tokens in large groups (called batching) in order to help increase overall token processing rates.

As described previously, several LLM inferencing stacks were available to choose from. The inferencing frameworks we investigated included the following:

- TGI – for both H100 and Gaudi 3
- vLLM – for both H100 and Gaudi 3
- Nvidia H100: Nvidia's TensorRT-LLM inferencing stack
- Intel Gaudi 3: Optimum Habana inferencing stack

Note: *The optimal solution for each accelerator was chosen. We used TensorRT-LLM for Nvidia H100 testing, and Optimum Habana for Intel Gaudi 3 testing.*

As seen in Figure 1, the Signal65 / Kamiwaza AI test suite enables the testing of inferencing performance of various LLM models across multiple GPU's and optionally multiple nodes. The hardware utilized for inferencing is agnostic at the level of submitting requests.

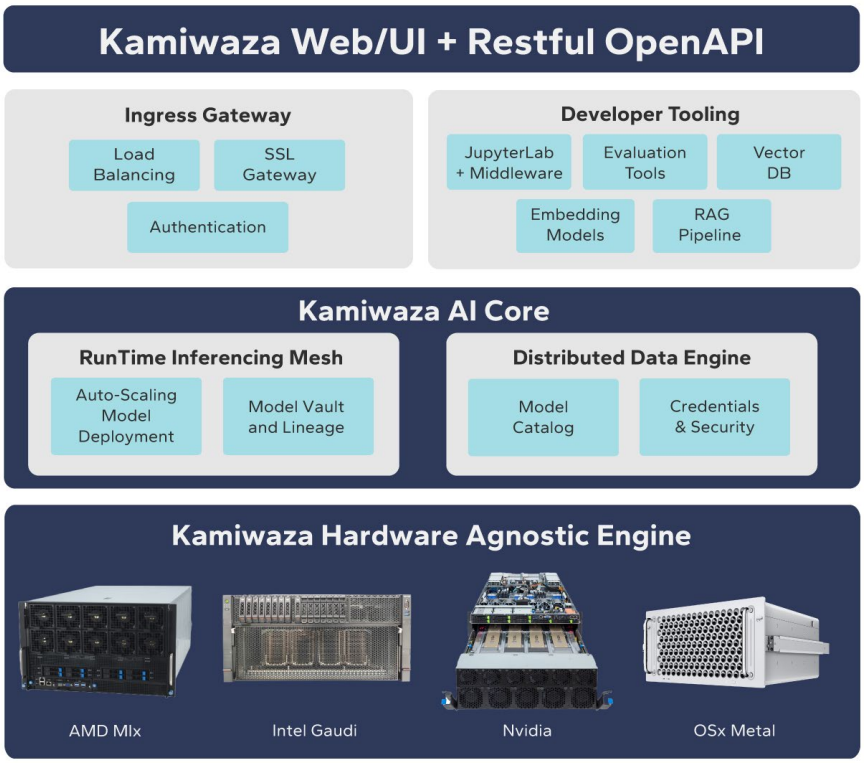


Figure 1: *Signal65 / Kamiwaza Inferencing Platform*

These frameworks are basic utilities only. The Signal65 / Kamiwaza Bench provided automations and tools to provide a benchmarking experience (from batch experiment configuration to exec automation to logging to scoring and visualization).

The testing methodology utilized was to compare the two hardware AI accelerators using two different open-source large language models. For testing a single AI accelerator, we chose to utilize Llama 3.1 8B model, which can fit fully within a single accelerator with 48 GB or more memory capacity.

To properly exercise an eight-accelerator system, we used the Llama 3.1 70B model, which was then loaded across 8 accelerators during inference testing. All inferencing was done in batch mode to maximize accelerator throughput.

The majority of our testing was performed at so called "full weight" or FP16 data size, without the use of quantization. Our focus was on recreating common scenarios, which led us to primarily test full-weight models as these are shown to produce significantly better, i.e. higher accuracy results than those using quantized data sizes.

For both the 8B and 70B models, we ran a range of input and output token size tests. For brevity we are showing only 4 combinations. In all cases, the input vs. output size is shown as two values in the format (input / output).

Cost Analysis

To provide a price vs. performance analysis, Signal65 gathered pricing data for the two competing solutions. First, we obtained a configuration quote from a publicly accessible reseller, Thinkmate.com, which provided detailed pricing data for a GPU server with 8 Nvidia H100 GPU's. The details are shown below in Table 2. Additionally, we utilized published Intel pricing for the Gaudi 3 accelerators at the time of their release, which was quoted as having a "List price of $125,000" according to several sources, which are provided in the Appendix.

Moreover, we constructed a system price for a Gaudi 3 - XH20 system by using the base system price of $32,613.22 shown below and then adding in the reported cost of the 8 Intel Gaudi 3 accelerators which was $125,000 to arrive at a total system price of $157,613.22, compared to a full system cost of $300,107.00 for the identical system with 8x Nvidia H100 GPU system.

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

Table 2: *Pricing Details for H100 vs. Gaudi 3 AI Server as of 1/10/25 (Source: Thinkmate.com)*

Performance Comparison

One important aspect of testing is the use of the word "performance", as that term can apply to two completely different methods of measuring AI accelerators. One measure of performance is to measure the accuracy of the results provided. This is an important factor and is sometimes called "model performance." However, *accuracy was not the focus* of this lab validation. Instead, we describe performance by measuring the rate at which tokens were processed, reported in tokens per second, resulting in a token rate of the solution.

Additionally, to ensure that model accuracy was not being sacrificed while trying to achieve higher token processing rates, we measured model accuracy of both accelerators utilizing several well-known tests. The results showed that both the Intel Gaudi 3 and Nvidia H100 produced accuracy results with no statistical differences. That is, although there were slight differences in the reported accuracy, these differences were within the margin of error for our measurements. The accuracy results are provided in the Appendix.

Quantized Model Comparison

We start our comparison by looking at a use case that may not be commonly deployed; however, these results are often published due to their higher throughput rates compared to inferencing models at "full weight" or FP16 data type. The results below utilize a smaller, so called "quantized" data size of FP8, which enables faster inferencing performance, at the expense of model and result quality. These results are relevant for some users and are shown for completeness.

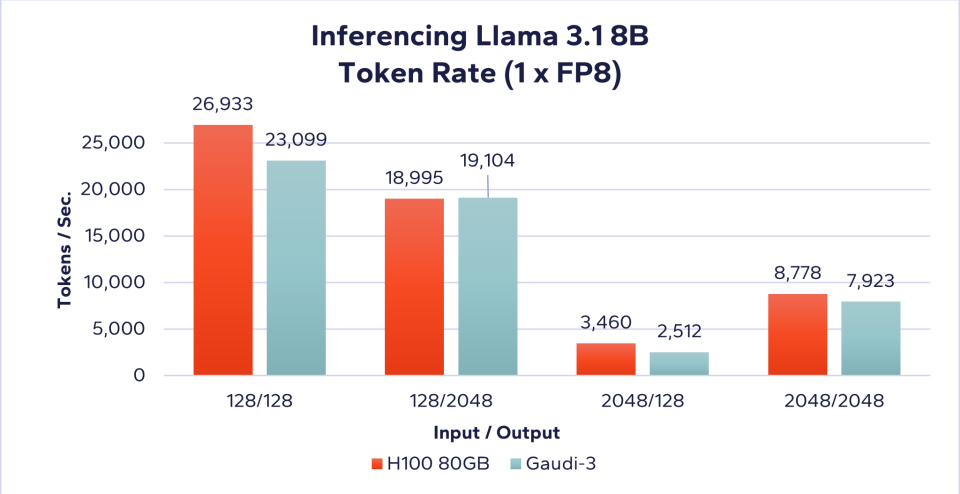


Figure 2: *Inferencing Performance Comparison at 8bits – FP8 (Source Signal65)*

In the chart above, the denotation of "1 x FP8" indicates a single accelerator card was used, and inferencing was performed using the FP8 data type. These results highlight how Nvidia H100's support for quantized, FP8 data type can provide inferencing speed benefits vs. Intel Gaudi 3 accelerators. However, it is evident that Gaudi 3 results are similar to those of the H100 accelerator even though the H100 was optimized to run FP8 data types.

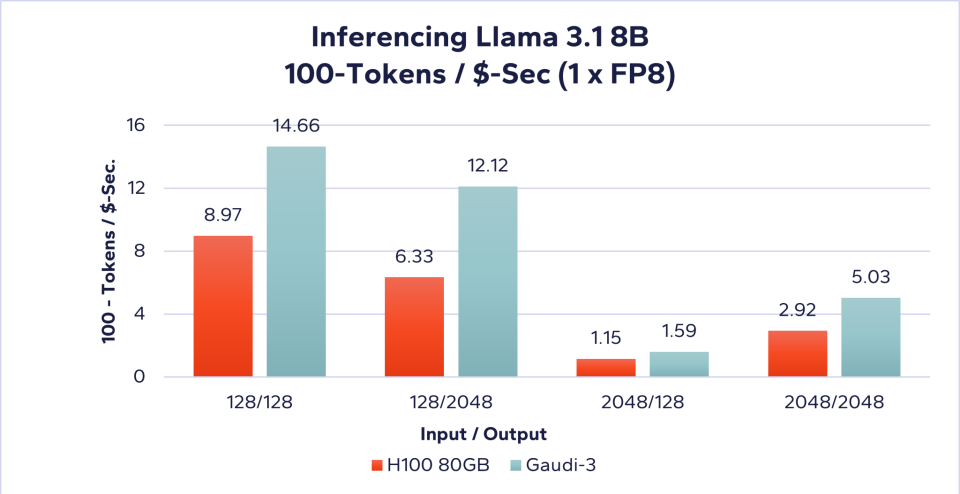


Figure 3: *Token rate per unit cost Comparison at 8bits – FP8 (Source Signal65)*

As seen above in Figure 3, when evaluating the number of tokens processed per unit cost, where higher (more tokens) is better, we see that across all 4 workload combinations, Intel's Gaudi 3 provides better results.

As an example, using the data for 128 input tokens and 128 output tokens, the left most set of bars in Figure 2, and the cost from Table 1, we have the following calculations:

- Nvidia H100: 128/128 Perf = (26,933 tokens / sec) / $300,107.00 = 0.089744 * 100 = 8.97
- Gaudi 3: 128/128 Perf = (23,099 tokens / sec) / $157,613.22 = 0.1466 * 100 = 14.66

Full Weight Llama Performance

In Figure 4, we show a comparison of the Nvidia H100 80 GB accelerator and an Intel Gaudi 3 accelerator running the Llama 3.1 8B LLM using a single accelerator, using a 16-bit data type. Importantly, Nvidia uses "FP16" while Intel uses "BF16" which are equivalent, but slightly different representations using 16 bits of precision. As shown, Gaudi 3 was roughly equivalent to the H100 across the four workloads, with Gaudi 3 performing better for workloads with smaller input vs output ratios, and the H100 performing better for workloads with larger input vs. output ratios.

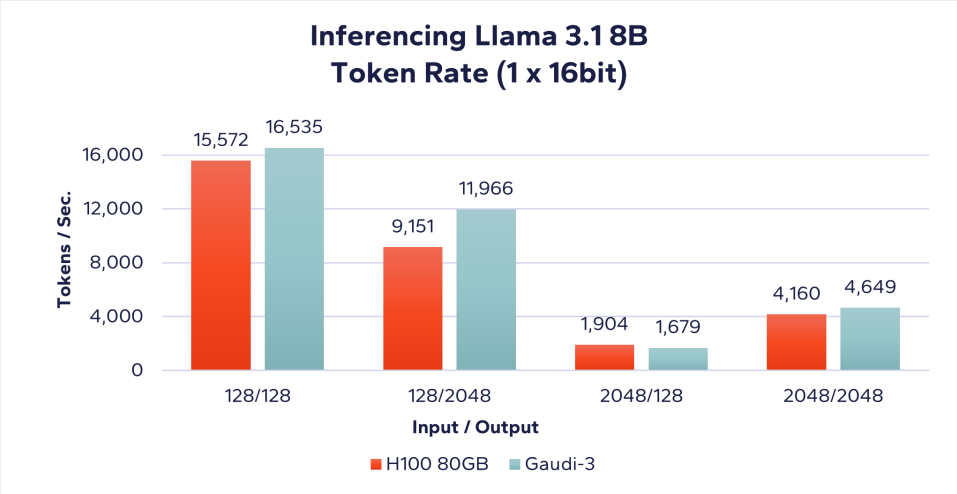


Figure 4: *Llama 8B – 1 Accelerator Performance Comparison at 16bits (Source Signal65)*

Next, we evaluated the AI accelerator performance across the same 4 workload scenarios, using the larger Llama 3.1 70B model, which requires the use of multiple accelerators to operate, due to the memory requirements. In Figure 5, we show the performance for 8 accelerators, again comparing Nvidia H100 to Intel Gaudi 3. The label of (8 x 16bit) indicates 8 accelerators were used with either FP16 or BF16 data type.

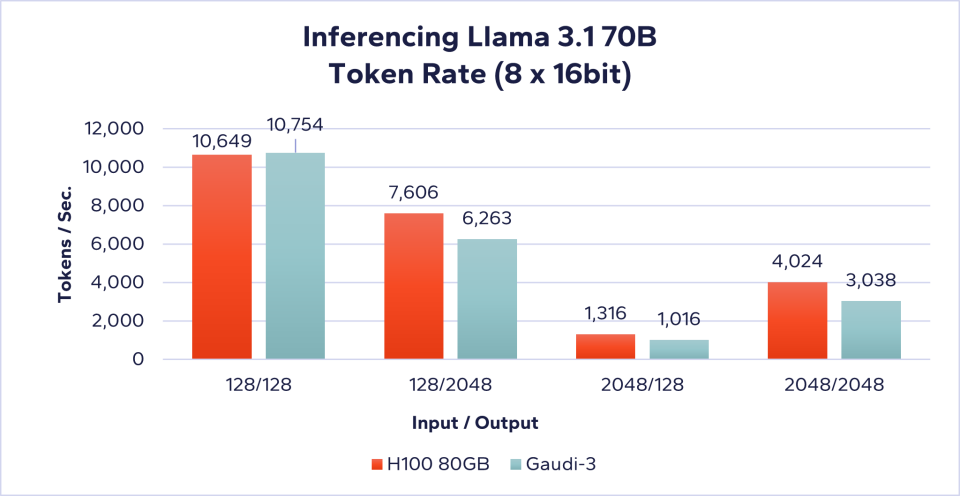


Figure 5: *Llama 70B – 8 Accelerator Performance Comparison at 16bits (Source Signal65)*

Again, the results are similar, with Nvidia performing somewhat better for those workloads with a higher ratio of input size to output sizes.

Performance vs. Cost

As described previously, one of the most important considerations for many companies is the rate of token processing compared to the cost. In this paper, performance per cost is expressed as the number of tokens per second processed per unit cost.

First, in Figure 6, we analyze the results of a single accelerator running the Llama 3.1 8B model, this time considering cost. The results are presented in terms of 100's of tokens processed per second, per dollar. Thus, higher is better, representing more tokens processed per unit of cost.

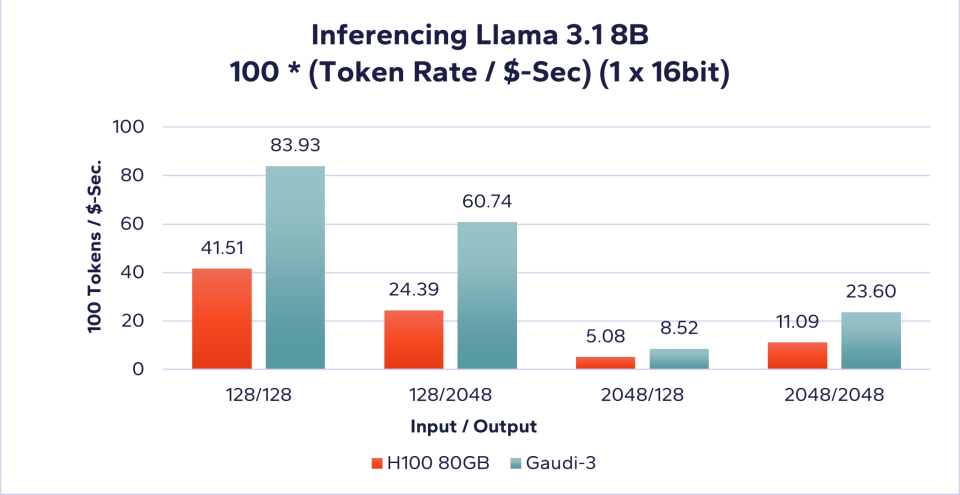


Figure 6: *Llama 8B – 1 Accelerator Token Rate per $ Comparison at 16bits (Source Signal65)*

Next, Figure 7 presents the performance per dollar for the larger, Llama 3.1 70B model. Again, this workload was run in full 16bit precision across eight AI accelerators.

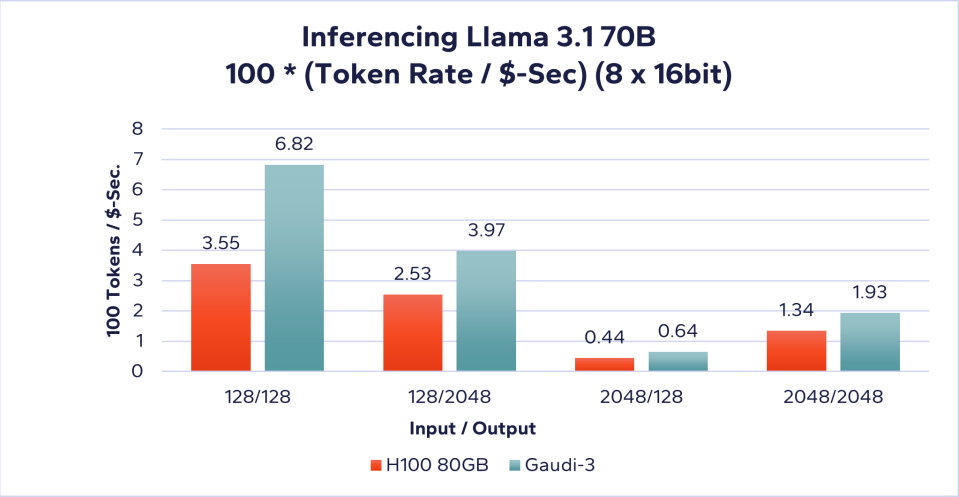


Figure 7: *Llama 70B – 8 Accelerator Token Rate per $ Comparison at 16bits (Source Signal65)*

Performance Summary

As evidenced by the numerous data points, when looking only at performance, the Nvidia H100 and Intel Gaudi 3 provide similar inferencing speed across the tested set of Llama 3.1 workloads. In some cases, Nvidia has a slight performance advantage and in other scenarios, Intel Gaudi 3 performed better.

Based upon our pricing data, we found that Intel's Gaudi 3 delivered from 10% better, up to 2.5X better performance per dollar than the Nvidia H100.

Final Thoughts

Enterprises are rapidly developing applications that leverage AI to improve their productivity. However, as AI enhanced applications become more prevalent, competitive pressures will move from simply having operational AI applications, to one of competitive differentiation based upon quality and the cost effectiveness of delivering these applications.

To date, much of the AI hype and reporting on the AI landscape has focused on the hyper-scale deployments, and their thousands of AI accelerators used to develop and train the latest AI models. While hyper scalers have the resources to do this work, it is not feasible or cost effective for enterprises to spend time developing and training foundational transformer or diffusion models.

***Signal65 Comment:* With the use of AI growing at unprecedented rates, there is increasing demand for more choice by companies for hardware accelerators. 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.**

Moreover, the predominant use case for enterprises will be production deployments, running inferencing workloads. Our focus on these workloads using the Signal65 benchmark suite is designed to provide meaningful insights into both performance and price / performance metrics, that will enable corporate executives to make informed purchasing decisions when choosing their AI inferencing platforms.

While the Nvidia H100 provides some modest performance advantages compared to Intel® Gaudi® 3 AI Accelerator, when evaluating these two options with their cost differences included, Intel's Gaudi 3 shows clear price / performance advantages across the range of inferencing workloads we have shown.

Appendix

Testing Details

One important aspect of testing throughput for inferencing, is the batch size used for testing. It is well understood that increasing the batch size up to some values approaching 128 or larger can significantly improve the total throughput, as measured by the token rate, or tokens per second. Throughout testing, the batch sizes were varied in order to maximize throughput. Moreover, batch sizes used for different input / output token size combinations were different for both the Nvidia GPUs and the Intel Gaudi 3 GPUs.

Accuracy Test Results

SIGNAL65-KAMIWAZA BENCH: Accuracy Measurement			
Llama 3.1 8B Instruct		**Nvidia GPU**	**GAUDI HPU**
Test	**Description**		
Basic Coherence	Simple questions, simple answers	81.95	82.6
SimpleBench pass@1	Very hard human reasoning test	10	0
SimpleBench pass@5		12	10
SimpleBench pass@10		9	10
HumanEval pass@1	Coding test	55.12	54.76
HumanEval pass@10		60.37	61.59
MMLU Pro: Bio	Multi-domain knowledge: Biology	68.06	67.64

Gaudi 3 Pricing References

CRN Announcement
https://www.crn.com/news/components-peripherals/2024/intel-reveals-8-chip-gaudi-3-platform-price-upending-industry-norm-of-secrecy

Tech Radar Announcement
https://www.techradar.com/pro/intel-discloses-list-prices-of-its-gaudi-3-and-gaudi-2-ai-accelerators-and-were-in-for-a-shock-rivals-to-iconic-nvidias-h100-gpu-have-a-much-better-performance-per-dollar-ratio-but-will-it-matter

Next Platform Analysis
https://www.nextplatform.com/2024/06/13/stacking-up-intel-gaudi-against-nvidia-gpus-for-ai/

Important Information About this Report

CONTRIBUTORS

Russ Fellows
VP, Labs | 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