Technical Paper

Generative AI Compute
AI for Enterprise

Intel® Gaudi® 3 AI Accelerator

This technical paper introduces the next-generation AI accelerator from Intel®: the Intel® Gaudi® 3 AI Accelerator.

Introduction

Deep Learning and Artificial Intelligence workloads continue to demand higher performance and lower power consumption. This technical paper introduces the next generation AI accelerator from Intel and Habana Labs, an Intel company: the Intel® Gaudi® 3 AI accelerator. The new accelerator features the 5th generation of heterogenous AI acceleration architecture. The Intel® Gaudi® 3 accelerator was designed to provide state-of-the-art datacenter performance for all AI workloads, from generative applications such as large language models (LLMs) and diffusion models (image generation such as Stable Diffusion) to standard object recognition, classification, and voice dubbing.

The Intel® Gaudi® 2 AI accelerator, introduced in 2022, is supported by the Intel® Gaudi® software suite, which integrates the PyTorch framework. With the Intel® Gaudi® 3 AI accelerator we provide the next level of AI performance and power efficiency. Advancing from the Intel® Gaudi® 2 accelerator 7nm process, the Intel® Gaudi® 3 accelerator is manufactured in TSMC 5nm process, which provides improved area density and power efficiency.

Intel® Gaudi® 3 AI accelerator continues to push the boundaries of what is possible in performance and power efficiency. Built on the Intel® Gaudi® 2 accelerator architecture, Intel® Gaudi® 3 accelerator provides significant boosts in compute, memory bandwidth, and architectural efficiency.

The Intel® Gaudi® 3 AI accelerator features two compute dies, which together contain 8 MME engines, 64 TPC engines and 24x 200 Gbps RDMA NIC ports. In addition, the total of 8 HBM2e chips comprise a 128 GB unified High Bandwidth Memory (HBM).

The Intel® Gaudi® 3 AI accelerator excels at training and inference with 1.8 PFlops of FP8 and BF16 compute, 128 GB of HBM2e memory capacity, and 3.7 TB/s of HBM bandwidth.

Table of Contents

Overview

AI applications increasingly demand faster and more energy-efficient hardware solutions and the Intel® Gaudi® 3 accelerator was designed to answer the demand. With more than 2x FP8 GEMM FLOPs and more than 4x BF16 GEMM FLOPs compared to the Intel® Gaudi® 2 accelerator, Intel® Gaudi® 3 accelerator continues to provide state-of-the-art AI training performance. With 1.5x faster HBM bandwidth and 1.33x larger HBM capacity, the Intel® Gaudi® 3 accelerator provides an order-of-magnitude improvement in large language model inference performance compared to the Intel® Gaudi® 2 accelerator.

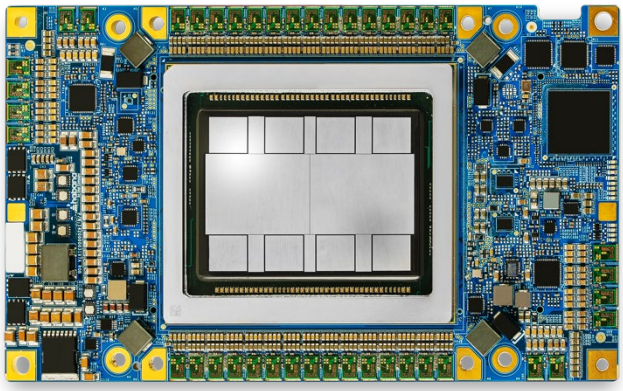

Figure 1. Intel® Gaudi® 3 OAM Module

The Intel® Gaudi® 3 accelerator (Figure 1) features two identical compute dies, connected through a high-bandwidth, low-latency interconnect over an interposer bridge. The die-to-die connection is transparent to the software, providing performance and behavior equivalent to that of a large unified single die.

The Intel® Gaudi® 3 accelerator compute architecture is heterogeneous and includes two main compute engines – a Matrix Multiplication Engine (MME) and a fully programmable Tensor Processor Core (TPC) cluster. The MME is responsible for doing all operations that can be lowered to Matrix Multiplication, like fully connected layers, convolutions and batched-GEMMs. The TPC, a Very Long Instruction Word (VLIW) Single-Instruction Multiple-Data (SIMD) processor tailor-made for deep learning applications, is used to accelerate all non-GEMM operations.

Intel® Gaudi® Accelerator Product Line

Intel® Gaudi® 2 to Intel® Gaudi® 3 Accelerator Feature Comparison

Feature / Product	Intel® Gaudi® 2 Accelerator	Intel® Gaudi® 3 Accelerator
BF16 MME TFLOPs	432	1835
FP8 MME TFLOPs	865	1835
BF16 Vector TFLOPs	11	28.7
MME Units	2	8
TPC Units	24	64
HBM Capacity	96 GB	128 GB
HBM Bandwidth	2.46 TB/s	3.7 TB/s
On-die SRAM Capacity	48 MB	96 MB
On-die SRAM Bandwidth	6.4 TB/s	12.8 TB/s
Networking	600 GB/s bidirectional	1200 GB/s bidirectional
Host Interface	PCIe Gen4 X16	PCIe Gen5 X16
Host Interface Peak BW	64 GB/s (32 GB/s per direction)	128 GB/s (64 GB/s per direction)
Media	8 Decoders	14 Decoders

Table 1. Intel® Gaudi® 2 and Intel® Gaudi® 3 Accelerators

Figure 2. Intel® Gaudi® Accelerator Product Line

Intel® Gaudi® 3 Performance Improvements

Intel® Gaudi® 3 accelerator is the third generation of the Intel Gaudi AI accelerator family. The large HBM capacity and bandwidth allow the Intel® Gaudi® 3 accelerator to achieve state-of-the-art performance GenAI training and inference. The charts below compare Intel® Gaudi® 3 accelerator to the second-generation accelerator and show how advancements introduced in Intel® Gaudi® 3 accelerator result in significant speedup in key AI workloads: training, inference throughput and inference latency.

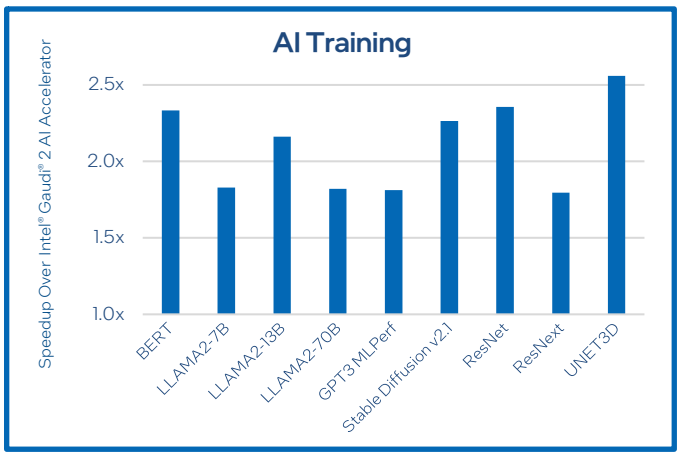

Figure 3. Performance Speedup vs. Intel® Gaudi® 2
Projected theoretical peak throughput of Intel Gaudi 3 accelerator vs.2nd generation.

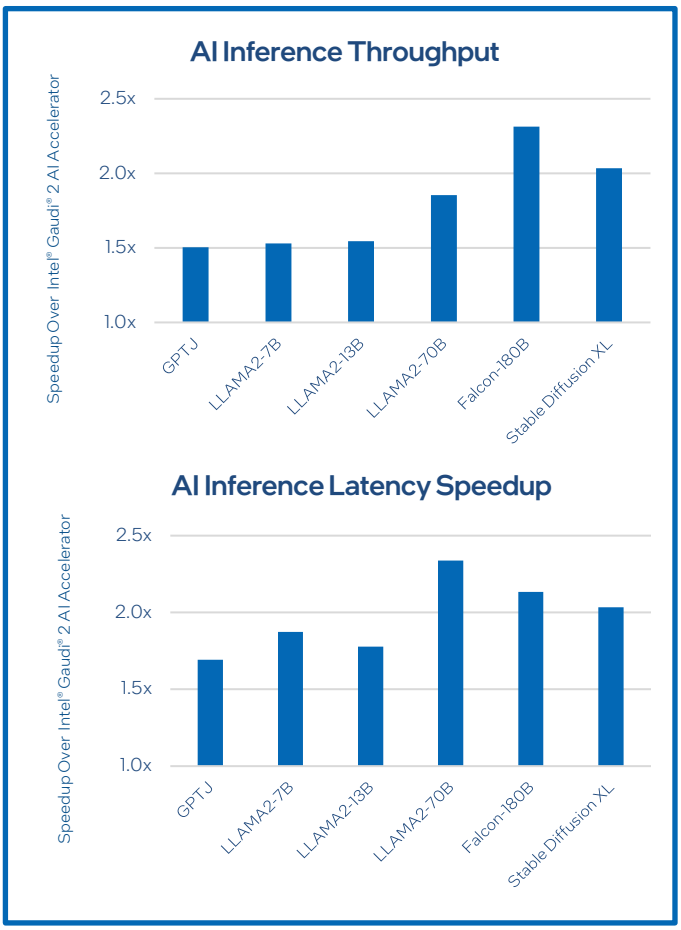

Figure 4. Intel® Gaudi® 3 AI Inference Throughput and Latency Speedups vs. Intel® Gaudi® 2
Projected theoretical peak throughput of Intel Gaudi 3 accelerator vs.2nd generation.

In training scenarios, virtually all of the advanced capabilities of Intel® Gaudi® 3 accelerator over the previous generation come into play. Since training scenarios are compute-intensive, the increased compute ratio provides immediate gain. The increased HBM bandwidth allows larger compute to manifest the increased compute power. In addition, the larger HBM capacity also contributes to improved performance. Larger HBM capacity allows increased batch size, enabling higher compute utilization and it allows avoiding re-computation of certain parts of the workload or avoiding model-parallel splits, which add networking operations during runtime.

In general, LLM inference throughput is determined by the available HBM bandwidth, which is used for reading the model parameters and context window. When comparing Intel® Gaudi® 3 accelerator to Intel® Gaudi® 2 accelerator, we see that for small LLMs (13B-sized model or smaller), speedup is similar to the ratio of HBM bandwidths between the two generations of accelerators, roughly 1.5x. However, when comparing larger LLM models, like LLama-70B and Falcon-180B, we see improvements that are greater than the HBM bandwidth ratios and surpass a 2x ratio. The larger improvement is due to the larger memory capacity that is available for Intel® Gaudi® 3 accelerator This larger capacity allows use of increased batch size and therefore more samples processed per given amount of time.

Intel® Gaudi® 3 accelerator LLM Training and Inference Performance Compared to NVIDIA H100

Intel® Gaudi® 3 accelerator was built to provide competitive performance on key AI workloads, as well as compelling price-performance relative to the NVIDIA H100.

Tables 2 and 3 compare projected Intel® Gaudi® 3 performance on LLM inference workloads against NVIDIA H100 and H200. The results of H100 and H200 were published by NVIDIA with the source links available below. We see that on average, Gaudi® 3 is expected to provide 50% higher throughput than H100 and 30% higher throughput on the inference scenarios below. Gaudi® 3 achieves the large gains compared to H100 not only due to its large HBM bandwidth and more efficient architecture, but also due to the larger HBM capacity, which allows use of larger batch sizes, thereby achieving larger HBM bandwidth utilization and overall inference throughput.

Table 4 shows 40% average inference power efficiency advantage of Intel® Gaudi® 3 accelerator vs. NVH100.

Table 5 compares Intel® Gaudi® 3 accelerator against H100 for pre-training throughput of Llama 7B and 17B workloads for 8 and 16 card scenarios. On average, Intel® Gaudi® 3 accelerator is expected to achieve 60% higher performance than H100. The main reason for the large gaps may be attributed to avoiding re-computation of the intermediate results that are generated during the back-propagation process. This is possible due to the 128 GB of HBM capacity of Intel® Gaudi® 3.

Table 6 shows pre-training time-to-train of GPT3 175B on MLPerf benchmark, on average, Intel® Gaudi® 3 is expected to achieve 25%- 40% higher faster time-to-train on large-scale pre-training. In this case, one reason for the larger gain is avoiding the usage of pipeline-parallelism by Intel® Gaudi® 3, while H100 uses pipeline-parallelism due its limited HBM capacity. With pipeline parallelism, the larger the number of devices, the higher the overhead.

Model & Execution Parameters				H100		Intel® Gaudi® 3 OAM		
Model	OAM (# Devices)	Input Length	Output Length	Batch Size[1]	Reported Throughput[1] (tps)	Batch Size	Projected Throughput (tps)	Gaudi® 3 H100 Speedup (x)
LLAMA-7B	1	128	128	896	20,241	1,536	21,201	1.0x
	1	128	2048	120	6,922	220	7,934	1.1x
	1	2048	128	64	2,170	120	2,002	0.92x
	1	2048	2048	56	2,816	120	3,168	1.1x
LLAMA-70B	2	128	128	1,024	6,538	4,096	5,794	0.9x
	4	128	2048	512	10,872	1,024	16,128	1.5x
	2	2048	128	96	694	220	655	0.9x
	2	2048	2048	64	2040	256	3,382	1.7x
Falcon-180B	4	128	128	512	4192	4,096	5,111	1.2x
	8	128	2048	1,024	6688	4,096	17,798	2.7x
	4	2048	128	64	456	512	507	1.1x
	4	2048	2048	64	1000	512	4,047	4.0x
							Average Speedup	1.5x

Table 2. LLM Inference Performance of Intel® Gaudi® 3 vs. NVIDIA H100*

*Source: NV H100 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , Mar 28th, 2024. Reported numbers are per GPU.Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B projections. Results may vary.

Model & Execution Parameters				H200		Intel® Gaudi® 3 OAM		
Model	OAM (# Devices)	Input Length	Output Length	Batch Size[1]	Reported Throughput[1] (tps)	Batch Size	Projected Throughput (tps)	Gaudi® 3 H200 Speedup (x)
LLAMA-7B	1	128	128	896	20,548	1,536	21,202	1.0x
	1	128	2048	120	8,343	220	7,934	1.0x
	1	2048	128	84	2,429	120	2,000	0.8x
	1	2048	2048	56	3,530	120	3,167	0.9x
LLAMA-70B	1	128	128	512	3,844	1,024	3,374	0.9x
	2	128	2048	512	8,016	512	6,947	0.9x
	1	2048	128	64	421	128	352	0.8x
	1	2048	2048	64	1,461	96	1,662	1.1x
Falcon-180B	4	128	128	1024	4,464	4,096	5,111	1.1x
	4	128	2048	1024	3,960	1,024	8,923	2.3x
	4	2048	128	64	472	512	507	1.1x
	4	2048	2048	64	1,076	512	4,048	3.8x
							Average Speedup	1.3x

Table 3. LLM Inference Performance of Intel® Gaudi® 3 vs. NVIDIA H200*

Source: NV H200 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8, Mar 28th, 2024. Reported numbers are per GPU.vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B projections. Results may vary.