

INTEL GAUDI 3 ACCELERATOR

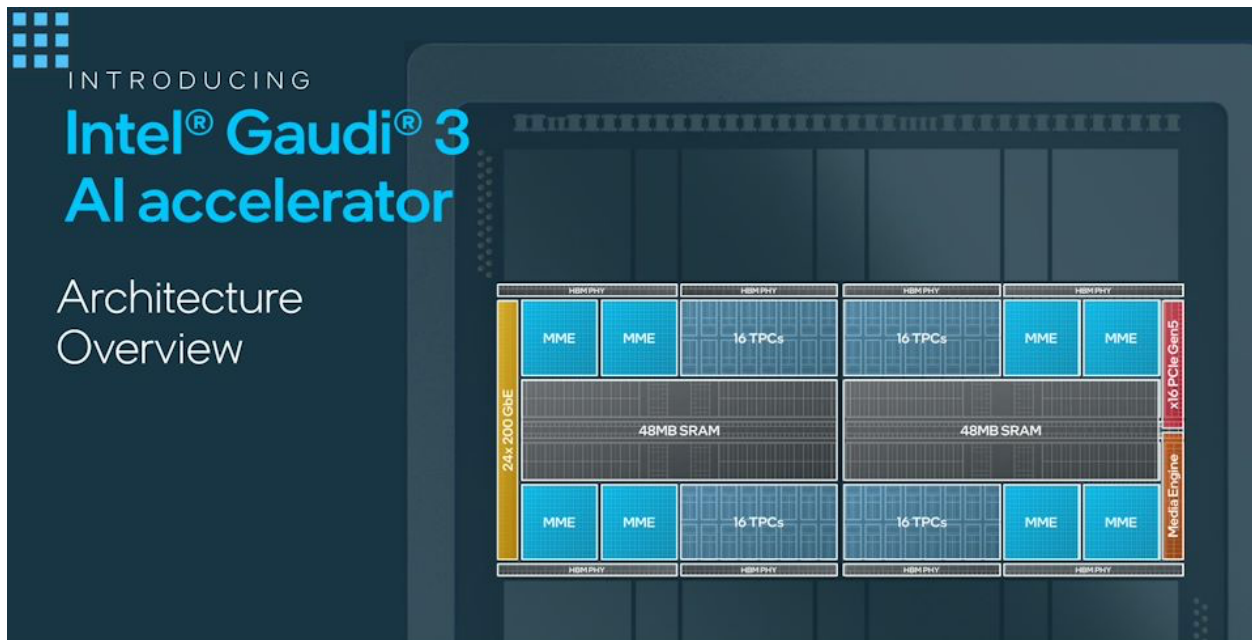
STEVE MCDOWELL, CHIEF ANALYST
ARIL 9, 2024

CONTEXT

Intel [announced](#) its long-anticipated new Intel Gaudi 3 AI accelerator at its Intel Vision event. The new accelerator offers significant improvements over the previous generation Gaudi 3 processor and promises to challenge Nvidia's current generation accelerators in training and inference for LLMs and multimodal models.

INSIDE GAUDI 3

Gaudi 3 dramatically increases AI compute capabilities, delivering substantial improvements over Gaudi 2 and competitors, particularly in processing BF16 data types, which are crucial for AI workloads.



ARCHITECTURAL ENHANCEMENTS

Manufactured using a 5nm process technology, Gaudi 3 incorporates significant architectural advancements, including more Tensor Processor Cores (TPCs) and Matrix Multiplication Engines (MMEs). This provides the computing power necessary for the parallel processing of AI operations, significantly reducing training and inference times for complex AI models.

The accelerator features a heterogeneous compute architecture, combining 64 custom AI TPCs and eight powerful MMEs. Each MME can perform a large number of parallel operations, enhancing the accelerator's capability to manage complex matrix operations fundamental to deep learning algorithms.

Gaudi 3 expands its hardware capabilities with more Matrix Math Engines and Tensor Cores than its predecessor, Gaudi 2. Specifically, it increases from 2 to 4 Matrix Math Engines (MMEs) and 24 to 32 Tensor Cores, bolstering its processing power for AI workloads.

Gaudi 3 boasts an FP8 precision throughput of 1835 TFLOPS, doubling the performance of Gaudi 2. It also significantly enhances BF16 performance, although specific throughput figures for this improvement were not disclosed.

Gaudi 3 has 128GB of HBM_e2 memory, offering 3.7TB/s of memory bandwidth and 96MB of onboard static RAM (SRAM). This massive memory capacity and bandwidth supports processing large datasets efficiently, which is crucial for training and running large AI models.

It features twenty-four 200Gb Ethernet ports, significantly enhancing its networking capabilities. This ensures scalable and flexible system connectivity, allowing for the efficient scaling of AI compute clusters without being locked into proprietary networking technologies.

NETWORKING CAPABILITIES

High-speed, low-latency networking is critical when building clusters of accelerators to solve large training tasks. While Nvidia is building its accelerators using proprietary interconnects like its NVLink, Intel is all-in on standard ethernet-based networking.

The advanced networking features in Gaudi 3 are built around the accelerator's Network Interface Controllers (NICs). These are pivotal for enabling efficient scale-out of deep neural network (DNN) training across multiple devices.

Here are the critical aspects of Gaudi 3's networking:

- **Ethernet Connectivity:** Gaudi 3 utilizes Ethernet connectivity for its NIC ports, offering an aggregated bandwidth of 4.8Tb/s in each direction. This choice

allows seamless integration into common cloud infrastructures and supports multiple port configurations.

- **RDMA and RoCE v2:** The NICs provide the compute engine with Remote Direct Memory Access (RDMA), featuring high bandwidth and low latency over a reliable connection without requiring software intervention. Gaudi 3 implements the RoCE v2 specification, which combines the low latency RDMA capabilities of the InfiniBand protocol with the widespread adoption and infrastructure of Ethernet, explicitly tailored for DNN applications and large-scale deployments.
- **MPI Collective Operations Mapping:** One of the challenges in DNN training over distributed systems is the efficient execution of MPI collective operations, which do not naturally map to RDMA's read-and-write operations. Gaudi 3 addresses this by implementing a hardware-based solution that manages the rendezvous flow on the sender side, ensuring data is sent to the receiver efficiently, minimizing latency, and bypassing memory copies.
- **Offloading and Congestion Control:** Intel Gaudi 3 offloads collective operations to hardware, allowing the NICs to manage these operations directly. This offloading significantly reduces CPU overhead and ensures full utilization of the network's bandwidth. For congestion control, Gaudi 3 goes beyond traditional methods by implementing time-based schemes like SWIFT, which use delay (RTT calculation) as a congestion indication signal, providing finer control over congestion than ECN alone.
- **Multi-Path Load Balancing:** To optimize the network bandwidth and mitigate congestion in large cluster configurations, Gaudi 3 introduces a sophisticated load balancing system. This system considers the load on each path and adapts to maintain high bandwidth utilization and low latency, which are essential for sustaining performance in distributed AI training tasks.

The networking innovations in Gaudi 3, from its enhanced RDMA over Ethernet capabilities to its advanced congestion control and load balancing mechanisms, are designed to ensure scalability and efficiency. These features enable linear scalability over thousands of Gaudi accelerators, making it an ideal platform for building extensive DNN training clusters capable of tackling some of the most challenging AI workloads.

PERFORMANCE

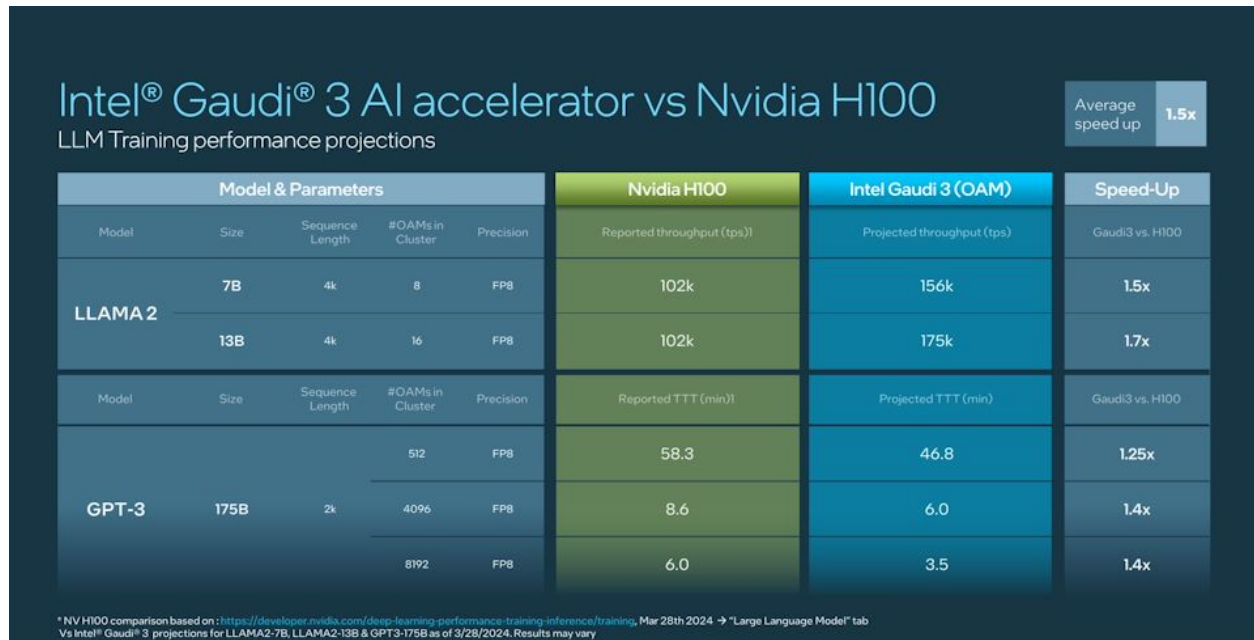
Intel's Gaudi 3 AI accelerator shows robust performance improvements across several key areas relevant to AI training and inference tasks, particularly for LLMs and multimodal models.

Here's an overview of its performance characteristics:

- **Quadruple AI Compute for BF16:** Gaudi 3 shows a 4x increase in AI compute capabilities for BF16 precision over Gaudi 2, significantly boosting the efficiency and speed of training and inference tasks. .

- 1.5x Increase in Memory Bandwidth:** The accelerator features a substantial 1.5 times increase in memory bandwidth, enabling faster data transfer rates and improved handling of large datasets.
- 2x Networking Bandwidth:** With double the networking bandwidth, Gaudi 3 facilitates enhanced system scalability and interconnectivity, allowing for massive system scale-out capabilities and the efficient linking of multiple AI accelerators for larger, more complex AI computations.

Intel projects that Gaudi 3 will significantly outperform competing products like Nvidia's H100 and H200 in training speed, inference throughput, and power efficiency for various parameterized models.



Intel® Gaudi® 3 AI accelerator vs Nvidia H100

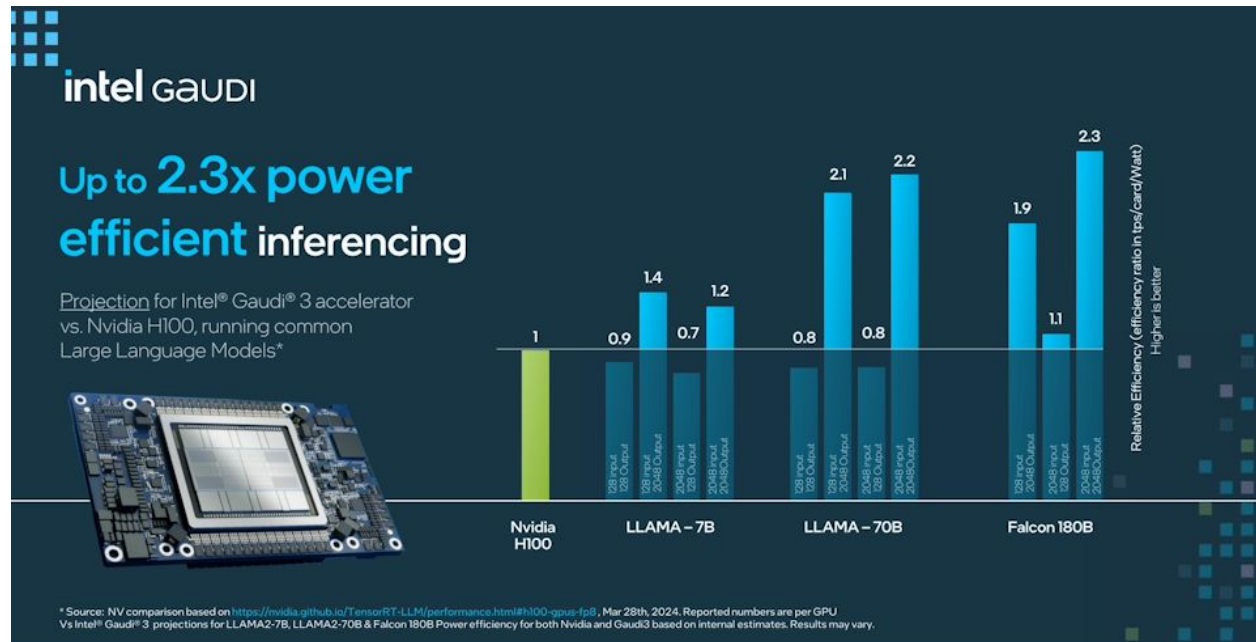
LLM Inference performance projections

Average speed up **1.5x**

Model & Parameters				Nvidia H100		Intel Gaudi 3 (OAM)		Speed-Up
Model	# OAM	Input Length	Output Length	Batch Size ¹	Reported TPT ¹ (tps)	Batch Size	Projected TPT (tps)	Gaudi3 vs. H100
LLAMA-7B	1	128	128	896	20,241	1536	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	1024	6,538	4096	5,794	0.9x
	4	128	2048	512	10,872	1024	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	4096	5,111	1.2x
	8	128	2048	1024	6688	4096	17,798	2.7x
	4	2048	128	64	456	512	507	1.1x
	4	2048	2048	64	1000	512	4,047	4.0x

¹Source: 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.

Intel also predicts Gaudi 3 will deliver an average 50% faster training time and superior inference throughput and power efficiency against leading competitors for several parameterized models. This includes a greater inference performance advantage on longer input and output sequences.



PACKAGING

Intel delivers Gaudi 3 in multiple packaging solutions designed to cater to a broad range of system designs and applications:

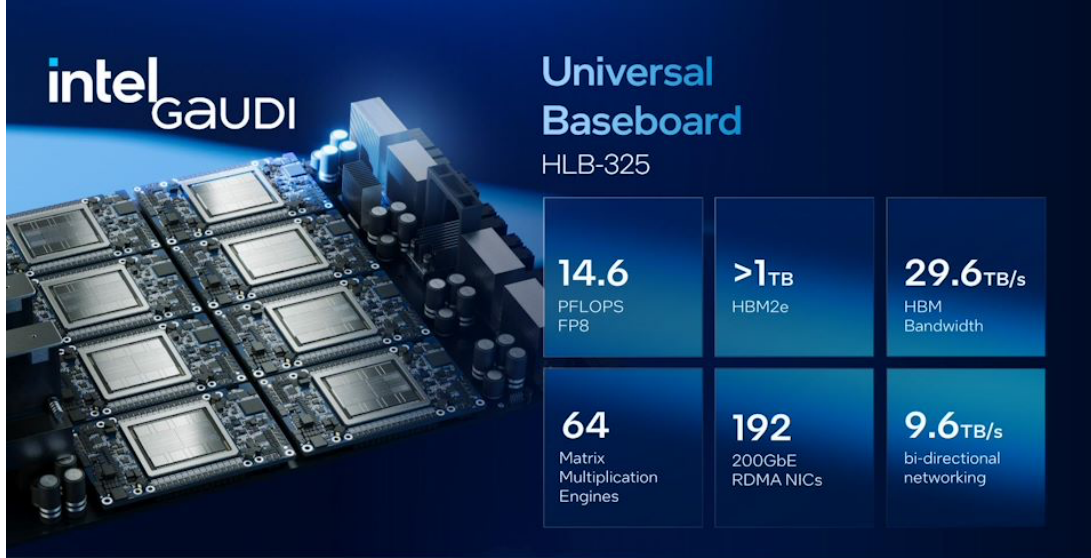
- **Open Accelerator Module (OAM):** Gaudi 3 is available in an Open Accelerator Module (OAM) form factor, ensuring compatibility and interoperability across different vendors' AI accelerators and servers. This form factor enables efficient thermal management and power delivery, catering to high-density AI compute environments.



intel.vision

Intel Confidential – Embargoed until 4/9/2024 at 8:35am PT

- **Universal Baseboard (UBB):** Intel also offers the Gaudi 3 accelerator in a UBB configuration. This setup integrates eight Gaudi 3 accelerators onto a single motherboard. The UBB design facilitates tight interconnection among the accelerators and provides six high-speed OSFP (Octal Small Form-factor Pluggable) ports for external connectivity at 800 Gb/sec speeds. This configuration is analogous to Nvidia's HGX system boards, supporting dense, high-performance AI compute clusters.



intel GAUDI

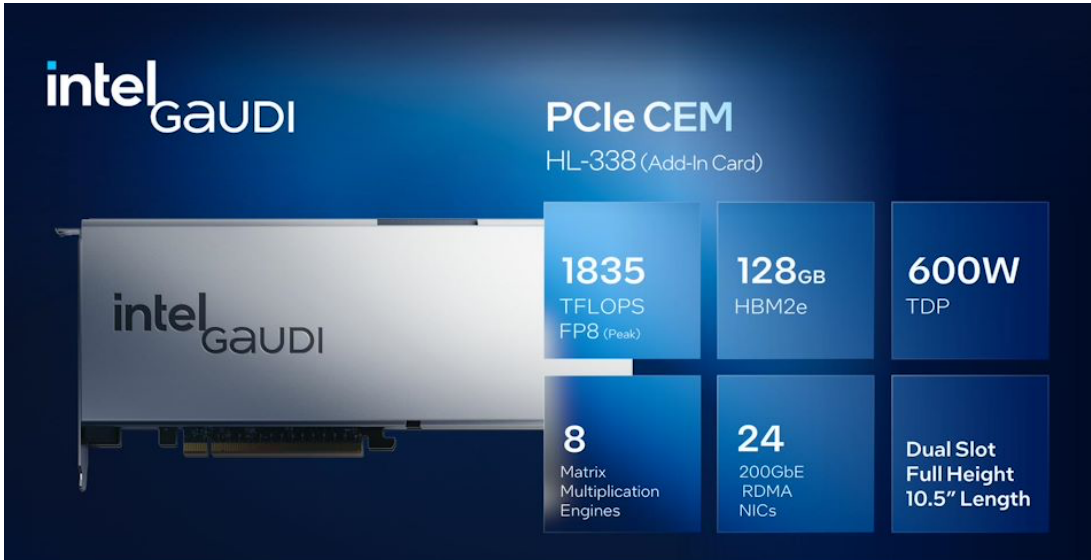
Universal Baseboard
HLB-325

14.6 PFLOPS FP8	>1TB HBM2e	29.6TB/s HBM Bandwidth
64 Matrix Multiplication Engines	192 200GbE RDMA NICs	9.6TB/s bi-directional networking

intel.VISION

Intel Confidential – Embargoed until 4/9/2024 at 8:35am PT

- PCI-Express 5.0 x16:** For broader compatibility and ease of deployment in existing server infrastructures, Intel is releasing a PCIe gen 5.0 x16 variant of the Gaudi 3 accelerator. This is designed for passive cooling and dual-width form factor slots, making it suitable for direct insertion into servers that support PCIe 5.0 slots.



intel GAUDI

PCIe CEM
HL-338 (Add-In Card)

1835 TFLOPS FP8 (Peak)	128GB HBM2e	600W TDP
8 Matrix Multiplication Engines	24 200GbE RDMA NICs	Dual Slot Full Height 10.5" Length

intel.VISION

Intel Confidential – Embargoed until 4/9/2024 at 8:35am PT

Intel developed air-cooled and liquid-cooled versions of the Gaudi 3 accelerators to meet varying thermal management needs. The air-cooled versions utilize large, skyscraper-like heatsinks for effective heat dissipation in environments where liquid cooling may not be feasible.

Meanwhile, the liquid-cooled variants are equipped with mounted cold plates, offering enhanced cooling efficiency for high-density and high-performance computing tasks.

ANALYSIS

Intel's Gaudi 3 AI accelerator is a strategic move by Intel to gain a greater position in the supply-hungry AI accelerator market, directly challenging Nvidia to address the burgeoning demand for advanced AI compute solutions.

Intel built a compelling solution, bringing substantial performance improvements over Gaudi 2 and delivering a solution that will challenge the market. The 4x AI compute for BF16, 1.5x increase in memory bandwidth, and 2x networking bandwidth improvements position the Gaudi 3 as a powerful solution for the needs of next-generation AI applications.

Intel's emphasis on open community-based software and industry-standard Ethernet networking addresses critical market needs for flexibility and scalability without vendor lock-in. This approach differentiates Intel from Nvidia and aligns with the broader industry trend toward open standards and interoperability.

Intel's partnerships with Dell Technologies, HPE, Lenovo, and Supermicro for the Gaudi 3 rollout set Intel up for success. If Intel can deliver the accelerators to the market on schedule and the promised performance claims hold, then Intel is poised to realize significant growth in the accelerator market. The same is also true for AMD and its MI300x accelerator.

Gaudi 3 isn't just about the current generation of AI accelerators but also sets the stage for Intel's next-generation GPU, Falcon Shores. By integrating the Intel Gaudi and Intel Xe IP with a single GPU programming interface, Falcon Shores is expected to further Intel's capabilities in AI and HPC.

The launch of the Gaudi 3 AI accelerator is a significant milestone for Intel, highlighting its technological advancements, strategic market positioning, and commitment to addressing the evolving needs of the AI industry. By offering substantial performance improvements, embracing open standards, and establishing strategic OEM partnerships, Intel is challenging the status quo in the AI accelerator market and positioning itself as a leader in the next wave of AI infrastructure.



© Copyright 2024 NAND Research.

NAND Research is a registered trademark of NAND Research LLC, All Rights Reserved.

This document may not be reproduced, distributed, or modified, in physical or electronic form, without the express written consent of NAND Research. Questions about licensing or use of this document should be directed to info@nand-research.com.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.