Product Brief

Intel® Gaudi® 3 AI Accelerator
HL-338 PCIe Card

Intel Gaudi 3 AI Accelerators now available as a PCIe Card

The Intel® Gaudi® 3 AI Accelerator PCIe Add-In Card (HL-338) delivers scalable, cost-efficient AI performance in a standard form factor.

Introducing the Intel® Gaudi® 3 AI Accelerator PCIe Card (HL-338), engineered to deliver inference and fine tuning for the most popular AI models and usages in a standard PCIe Gen 5.0 compliant CEM Spec revision 5.1. Enjoy the cost-effective performance and easy deployment of AI inference for more uses to innovate faster.

Designed from the ground up for Inference

The inference accelerator is built on the 5nm process technology generation high-efficiency heterogeneous architecture. The HL-338 is a full-height, dual-slot PCIe card with a length of 10.5" and a card level TDP of 600 watts, providing customers with the flexibility to integrate into new or existing AI server designs. The Intel® Gaudi® 3 AI accelerator features 8 MME engines and 64 fully programmable Tensor Processor Cores (TPCs). The TPCs are natively designed to accelerate and be optimized for a wide array of deep learning workloads. The accelerator card is equipped with 128GB of HBM2e memory and 96MB of on-die SRAM which can handle small to large LLM models ideal for enterprises looking for low latency.

Scalability

The Intel® Gaudi® 3 PCIe card offers an array of configurations that enables scalability with a PCIe 5.0 x16 slot offering up to 128 GB/s of bandwidth, or through a top bridge setup that aggregates 4 cards to achieve 900 GB/s bandwidth. The Intel® Gaudi® 3 accelerator integrates a dedicated media processor for image and video decoding and pre-processing. The RoCE v2 RDMA ports on the HL-338 are exposed through a gold-finger connector, which can utilize the top bridges (HLTB-304A/HLTB-304B) to connect the 4 HL-338 cards.

Compute Technology: Multiple data types, TPCs and MMEs

Based on the proven architecture of first-gen Intel® Gaudi® and Intel® Gaudi® 2, Intel® Gaudi® 3 accelerators support the most advanced data types for AI, including FP8, BF16, FP16, TF32 and FP32. Leverage Intel's fully programmable TPC and GEMM Engine to run your complex matrix multiplications. The TPC core was designed to support inference workloads. It is a VLIW SIMD vector processor with an instruction set and hardware tailored to serve these workloads efficiently.

Memory improving LLM performance

Memory bandwidth and capacity are as important as compute capability as it enables to transfer and manage large workloads. The Intel® Gaudi® 3 accelerator incorporates HBM2e memory technology, supporting extremely high memory capacity of 128GB and total throughput of 3.7 TB/s. The HBM controller is optimized for both random access and linear access, providing strong throughput in all access patterns.

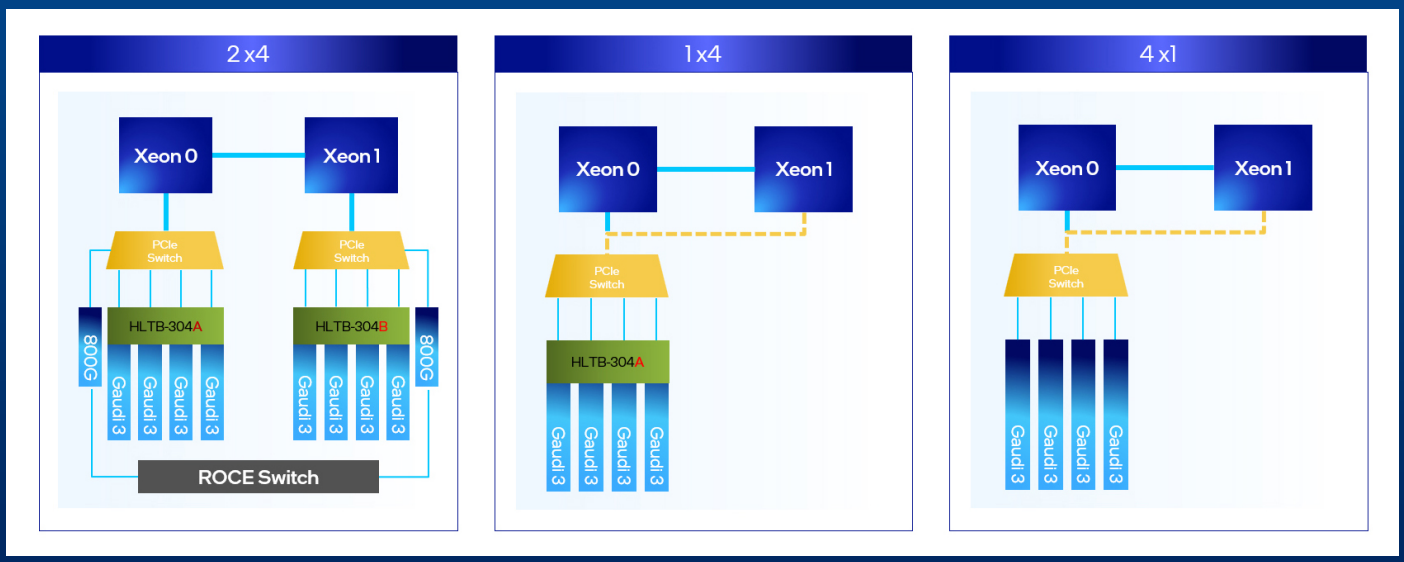
Intel Gaudi 3 PCIe Card topologies

There are three topologies that are available with HL-338 PCIe cards. PCIe card offers a modular set of configurations, reducing the risk of overprovisioning, making it a great choice for early pilots or phased deployments.

The 2x4 topology allows scale-up and utilizes a configuration of two sets of four HL-338 PCIe cards (quad) within a single node. Each quad is scaled up using 2 top boards, (HLTB-304A and HLTB-304B) that manages data between the 4 cards. The HLTB-304A board connects the first quad, while the HLTB-304B, in conjunction with HLTB-304A, connects the second quad, totaling eight cards (2 x4 configuration). Each top board scale up bandwidth provides 900 GB/s per card. These quads can interface with a PCIe switch in a dual-socket server. Quad to quad scale up within a single node can be implemented via host-NIC and ROCE switch connectivity. To ensure rapid data transfer, reduced latency, and seamless scalability, Intel recommends providing at least 800Gbps of network connectivity per bank of 4 cards for scale-up connections between the quads. NIC must support LibFabric and GDR based on DMA-buff protocol. The two sets of the four PCIe cards provide 2* 512GB HBM2e pooled memory, ideal for storing and processing large AI datasets.

Utilize the 1x4 topology to harness AI capabilities with 512GB of pooled HBM2e memory (128GB per card), ideal for medium to large-sized models. This setup includes a set of 4 HL-338 PCIe cards, scaled up through the HLTB-304A top board, connecting to a single CPU. For 1x4 topology, the NIC and the PCIe switch is arbitrary.

The 4x1 configuration is perfect for running different AI models concurrently on each card. In this setup, four HL-338 PCIe cards are connected to a single CPU without a top bridge, functioning as four independent PCIe card.



Open software ecosystem: Familiar frameworks with no lock-in

The Intel Gaudi 3 software stack plays a critical role in unlocking the full return-on-investment potential of the hardware. It helps enable efficient deployment and scale of AI models and applications. Flexibility, open integration and seamless deployment streamline delivery of real-world impact.

- **Seamless Integration**
  Intel Gaudi 3 AI accelerators natively integrate with PyTorch, vLLM and Hugging Face, ensuring effortless support for popular inferencing models and building AI services.
- **Scalable, Open Ecosystem**
  Intel Gaudi 3 AI accelerators support an open ecosystem, allowing AI developers to innovate freely without vendor lock-in, ensuring flexibility for future advancements.
- **FP8 Quantization**
  Intel Gaudi 3 AI accelerators offer automated FP8 quantization, significantly improving throughput while maintaining model accuracy — ideal for large language and multimodal models.

Go Deeper: Software

Intel® Gaudi® software provides model references, libraries, containers and tools to streamline every stage of developing, training and deploying AI solutions.

Learn more: https://www.intel.com/content/www/us/en/developer/platform/gaudi/overview.html#gs.mesgac

Datasheet Technical Specifications

Architecture	5th Generation Tensor Processor Core
TDP	600W (air cooling)
PCIe	FH & 10.5" in length, Double Width (x16 PCIe Gen 5.0)
PCIe Peak BW	128 GB/s bidirectional
Data Types	FP32, BF16, FP16 & FP8 (both E4M3 and E5M2)
HBM	8 x HBM2E
HBM Capacity	128 GB
HBM Peak BW	3.7 TB/s
On-die-SRAM	96 MB
On-die-SRAM BW	19.2 TB/s
System config (1x4)	1 group of x4 via Top Board (HLTB-304A)
System config (2x4)	2 groups of x4 via Top Boards (HLTB-304A & HLTB-304B)
Scale-out support	Via Host-NIC

Addressing the needs of today's AI Data Center

The Gaudi 3 family delivers new degrees of performance and efficiency that can help businesses meet their growing AI demands. Access the Intel® Gaudi® 3 AI Accelerators Webpage to explore the different Intel Gaudi 3 options, including the cluster reference design guide, for product and solutions recommendations.

intel GAUDI

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for additional details.

No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Printed in USA 0225/DR/PRW/PDF Please Recycle 360231-002US