Product Brief

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

The Intel® Gaudi® 3 AI accelerator PCIe card (HL-338) is designed to deliver performance and efficiency in a standard PCIe card form factor. The training and inference accelerator is built on the Intel® Gaudi® 5th generation high-efficiency heterogeneous architecture, now in 5nm process technology with state-of-the-art performance, scalability and power efficiency. The HL-338 is a full-height, dual-slot PCIe card with a length of 10.5" and a card level TDP of up to 600 watts, providing customers with the flexibility to integrate into new or existing AI server designs. The Intel® Gaudi® 3 AI processor features 8 MME engines and 64 fully programmable Tensor Processor Cores (TPCs). The TPCs are natively designed to accelerate a wide array of deep learning workloads while also providing the flexibility to optimize and innovate. The accelerator card is equipped with 128 GB of HBM2E memory and 96 MB of on-die SRAM.

The Intel® Gaudi® 3 AI accelerator offers unmatched scalability with 9.6 Terabits per second bi-directional networking capacity with native integration of 24x200 GbE RoCE v2 RDMA ports, enabling all-to-all communication via direct routing or via standard Ethernet switching. This on-chip networking integration gives customers capacity and flexibility to build systems of any scale. 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 HLTB-304 to connect 4 HL-338 cards, as well as 2 QSFP-112 connectors placed directly on the card.

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

Host Interface	Memory	TDP	Scale-Out Interconnect	Form Factor
PCIe Gen 5.0 x 16	128GB HBM2E	600W	RDMA (RoCE v2) 24x200 Gbps via 6 OSFP connectors	Full-height, Double-wide, 10.5" length PCIe Card

Technology Innovation

The Intel® Gaudi® 3 AI accelerator features a unique combination of technology innovations. As a high-performance and fully programmable AI processor, the accelerator is equipped with high memory bandwidth and capacity designed for efficient scale-out based on standard Ethernet technology. With its wide array of connectivity options, the Intel® Gaudi® 3 accelerator enables system integrators to build training systems of any scale, from a single server to complete racks using a variety of Ethernet switches and scale-out topologies, all while using the same standards-based, scale-out technology.

Compute Technology

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

Memory

Memory bandwidth and capacity are as important as compute capability. The Intel® Gaudi® 3 accelerator incorporates the most advanced HBM memory technology, supporting extremely high memory capacity of 128GB and total throughput of 3.7 TB/s. The cutting-edge HBM controller is optimized for both random access and linear access, providing record-breaking throughput in all access patterns.

Scale Out with Integrated RDMA

The Intel® Gaudi® 3 accelerator is the only AI deep learning processor to integrate on-chip RDMA over converged Ethernet (RoCE v2) to interface with industry standard Ethernet networking. The Intel® Gaudi® 3 AI accelerator chip interconnect technology is based on 48 pairs of 112Gbps Tx/Rx PAM4 SerDes configured as 24 ports of 200 Gb Ethernet.

Intel® Gaudi® Software Suite

Designed to facilitate ease of use and high-performance training on Intel® Gaudi® AI accelerators, the Intel® Gaudi® software suite enables efficient mapping of neural network topologies onto the Intel® Gaudi® family of hardware. The software suite includes the Intel® Gaudi® graph compiler and runtime, performance optimized TPC kernel library, firmware and drivers, and developer tools such as the TPC programming tool kit for custom kernel development and Intel® Gaudi® Profiler. Intel® Gaudi® software is integrated with popular frameworks, such as PyTorch, and optimized for training on the Intel® Gaudi® family of AI accelerators. Data scientists and developers can migrate their existing models to run on Intel® Gaudi® 3 accelerators with minimal code changes.

Intel® Gaudi® AI accelerator developer website is the hub where developers can find a wealth of information to get started with training on Intel® Gaudi® AI processors, including tutorials, reference models, how-to guides, documentation and more. It also hosts a Forum for the Intel® Gaudi® developer community.