Product Brief

Intel® Gaudi® 3 AI Accelerator HL-325L OAM Mezzanine Card

The Intel® Gaudi® 3 AI accelerator mezzanine card (HL-325L) is designed for massive scale out in data centers. 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-325L complies with the OCP OAM v2.0 (Open Compute Platform-Open Accelerator Module) specifications, giving customers system design flexibility with choice among products conforming to the spec. The Intel® Gaudi® 3 processor features 8 MME engines and 64 fully programmable Tensor Processor Cores (TPCs) 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 SRAM and supports card level TDP of up to 900 watts.

The Intel® Gaudi® 3 AI accelerator offers unmatched scalability of 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 dedicated media processor for image and video decoding and pre-processing.

Intel® Gaudi® 3 AI Accelerator HL-325L Mezzanine Card

Host Interface	Memory	TDP	Scale-Out Interconnect	Form Factor
PCIe Gen 5.0 x 16	128GB HBM2E	900W	RDMA (RoCE v2) 24x200 Gbps	OCP Accelerator Module V2.0 Compliant

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 and 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 (RoCEv2) 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.