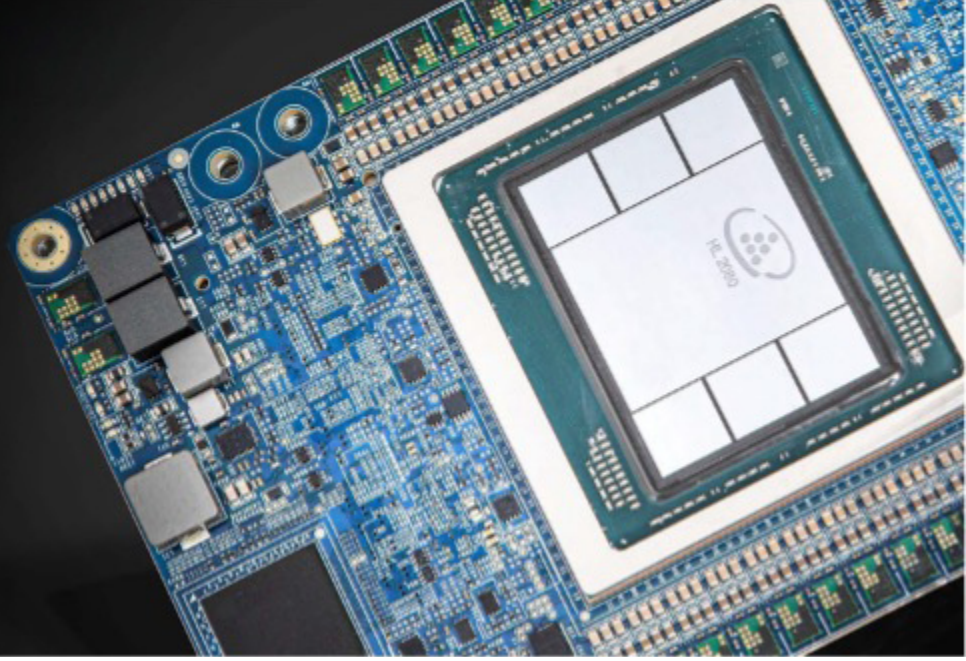
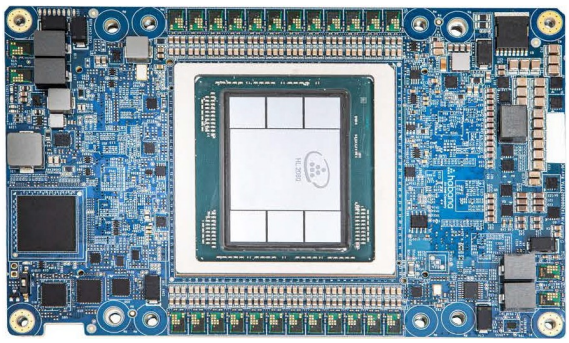
habana
An Intel Company

intel
GAUDI



Intel® Gaudi®2 AI Accelerator
HL-225H Mezzanine Card



The Intel® Gaudi®2 AI mezzanine card, the HL-225H, is designed for massive scale out in data centers. The training and inference accelerator is built on the high-efficiency architecture of first-generation Intel® Gaudi®, now in 7nm process technology, to deliver leaps in performance, scalability and power efficiency. The HL-225H complies with the OCP OAM v1.1 (Open Compute Platform-Open Accelerator Module) specification, giving customers system design flexibility with choice among products conforming to the spec. The HL-2080 processor features 24 fully programmable Tensor Processor Cores (TPCs) natively designed to accelerate a wide array of deep learning workloads, while giving the user the flexibility to optimize and innovate to address to their requirements. It also integrates 96 GB of HBM2E memory and 48 MB SRAM and supports card level TDP of 600 watts.

The Intel Gaudi2 AI accelerator offers unmatched scalability of 2.4 Terabits networking capacity with native integration of 24x100 GB RoCE v2 RDMA ports to enable inter-Gaudi 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 they require. The Intel Gaudi2 accelerator integrates dedicated media processor for image and video decoding and pre-processing.

	Intel® Gaudi®2 AI Accelerator HL-225H Mezzanine Card
PROCESSOR INTERFACE	Intel® Gaudi® HL-2080
HOST INTERFACE	PCIe Gen 4.0 x 16
MEMORY	96GBHBM2E
TDP	600W
SCALE-OUT INTERCONNECT	ROMA (RoCE v2) 24x100 Gbps
FORM FACTOR	OCP Accelerator Module V1.1 Compliant

Technology Innovation

The Intel® Gaudi®2 AI accelerator features a unique combination of technology innovations, as a high-performance and fully programmable AI processor with high memory bandwidth/capacity and scale-out based on standard Ethernet technology. With its wide array of connectivity options, the Intel Gaudi2 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, shipping first-gen Intel Gaudi architecture, Intel Gaudi2 accelerators leverage Intel's fully programmable TPC and GEMM Engine, supporting the most advanced data types for AI: 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 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 Gaudi2 accelerator incorporates the most advanced HBM memory technology, supporting extremely high memory capacity of 96GB and total throughput of 2.4TB/s. Intel Gaudi's 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 Gaudi2 accelerator is "the only AI deep learning processor to integrate on-chip RDMA over converged Ethernet (RoCEv2) to interface with mature and widely used Ethernet networking. The HL-2080 chip interconnect technology is based on 48 pairs of 56Gbps Tx/Rx PAM4 SerDes configured as 24 ports of 100Gb Ethernet.

SynapseAI® Software Suite

Designed to facilitate ease of use and high-performance training on Intel® Gaudi® AI accelerators, the SynapseAI® Software Suite enables efficient mapping of neural network topologies onto Intel Gaudi family of hardware. The software suite includes Intel Gaudi's 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 SynapseAI Profiler. SynapseAI is integrated with popular frameworks, TensorFlow and 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 Gaudi2 accelerators with minimal code changes. Intel's Habana Developer Site 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 Habana developer community.

For more details on Intel Gaudi performance and scaling, see the Intel Gaudi2 AI Accelerator Whitepaper.

habana
An Intel Company