GAUDI®2
HL-225H

An Intel Company

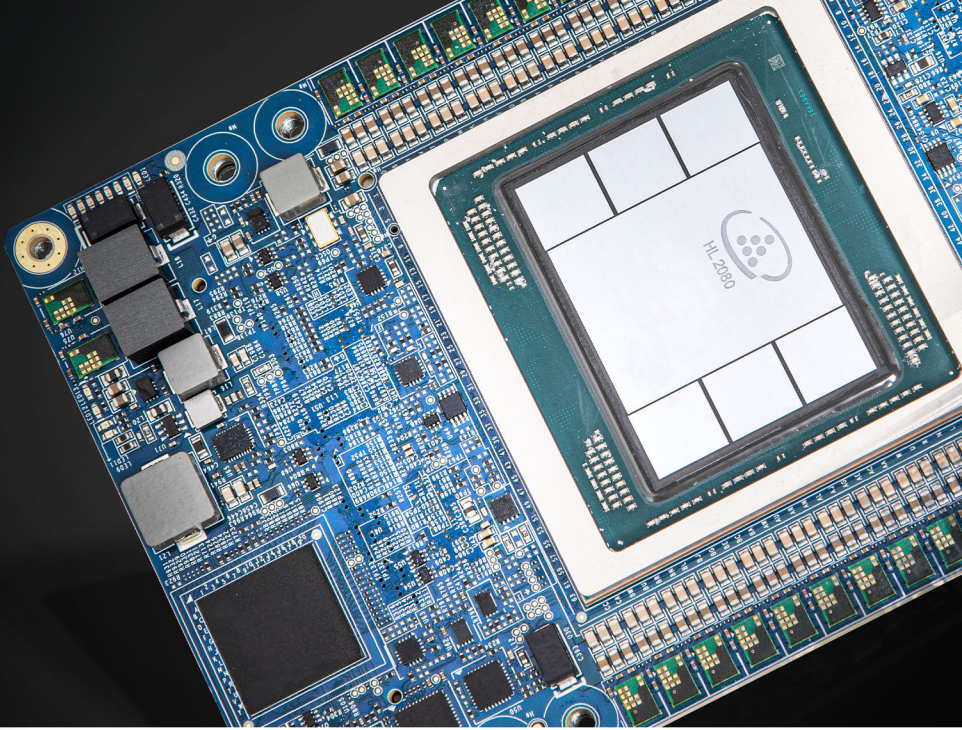**Habana® Gaudi®2 Mezzanine Card**

The Gaudi second-generation AI deep learning mezzanine card, the HL-225H, is designed for massive scale out in data centers. The training processor is built on the high-efficiency architecture of first-generation 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 4th generation 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 Gaudi2 processor offers unmatched scalability of 2.4 Terabits networking capacity with native integration of 24 x 100 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 Gaudi2 processor integrates dedicated media processor for image and video decoding and pre-processing.
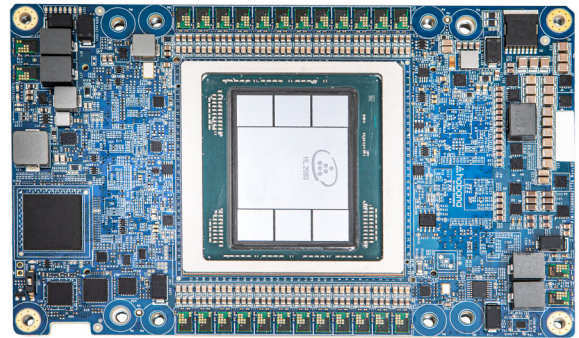
HL-225H Mezzanine Card	
PROCESSOR TECHNOLOGY	Gaudi HL-2080
HOST INTERFACE	PCIe Gen 4.0 x 16
MEMORY	96GB HBM2E
TDP	600W
SCALE-OUT INTERCONNECT	RDMA (RoCE v2) 24 x 100 Gbps
FORM FACTOR	OCP Accelerator Module V1.1 Compliant

habana®
An Intel Company

Technology Innovation

GAUDI®2 processor 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, 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 Architecture

Based on the proven, shipping training processor architecture, GAUDI2 leverages Habana'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. GAUDI2 incorporates the most advanced HBM memory technology, supporting extremely high memory capacity of 96GB and total throughput of 2.4TB/s. 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

GAUDI® is "the only AI training processor to integrate on-chip RDMA (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 Habana's AI processors, SynapseAI® Software Suite enables efficient mapping of neural network topologies onto Gaudi family of hardware. The software suite includes Habana'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 Gaudi family of AI processors. Data scientists and developers can migrate their existing models to run on Gaudi2 with minimal code changes. **The Habana Developer Site** is the hub where developers can find a wealth of information to get started with training on 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 Gaudi's performance and scaling, see the Habana Gaudi2 Whitepaper.

habana
An Intel Company