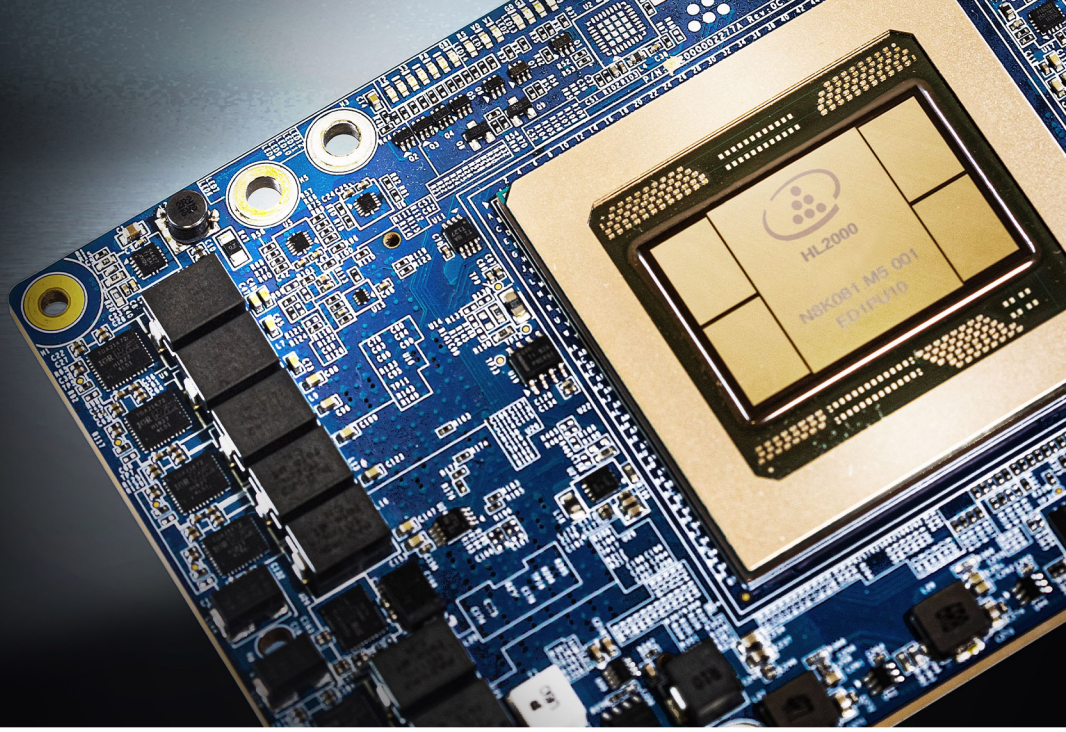
An Intel Company

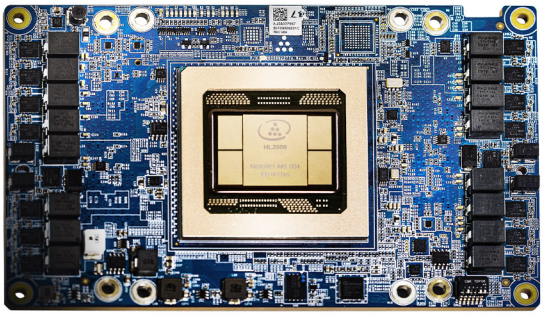
GAUDI®



GAUDI® AI Training Card

The GAUDI® HL-2000 is an advanced AI and Deep Learning Training processor, leveraging purpose-built architecture and delivering superior performance, scalability, power efficiency and cost savings. The Gaudi HL-205 mezzanine card, and the HL-200 PCIe card both incorporate a single GAUDI® HL-2000 Processor that contains a cluster of eight fully programmable Tensor Processing Cores (TPC 2.0). The TPC core is C-programmable, providing the user with maximum flexibility to innovate. The HL-205 is compliant with the OCP-OAM (Open Compute Accelerator mezzanine) specification.

The GAUDI® is designed to accelerate various AI Training workloads such as Image classification, object detection, natural language processing, text to speech, sentiment analysis, recommender systems and many others. The Gaudi HL-2000 includes native on-die integration of RoCE v2 RDMA functionality delivering two Terabits-per-second over 10 x 100GbE or 20 x 50GbE for inter-Gaudi communication by direct routing or via standard Ethernet switching. These engines play a critical role in the inter-processor communication needed during the training process. By integrating this functionality and supporting bi-directional throughput of up to 2 Tb/sec, customers can build systems of any size, and adapt them to their requirements.



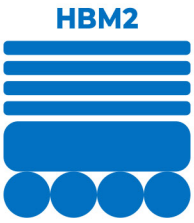
	HL-205 Mezzanine Card	HL-200 PCIe Card
PROCESSOR TECHNOLOGY	Gaudi HL-2000	
HOST INTERFACE	PCIe Gen 4.0 X 16	
MEMORY	32GB HBM2	
TDP	350W	200W
SCALE-OUT INTERCONNECT	RDMA (RoCE v2) 10x100Gbps or 20x50Gbps	8x100Gbps or 16x50Gbps Dual QSFP-DD ports
FORM FACTOR AND SKUS	OCP Accelerator Module Compliant	Full Height/Length HL-200 Passive Cooling

Technology Innovation

GAUDI® introduces 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, GAUDI® enables system integrators to build training systems of any scale, from low-cost servers 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 inference processor architecture, GAUDI® leverages Habana's fully programmable TPC and GEMM Engine, supporting the most advanced data types for AI: BF16 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 HBM2

Memory bandwidth and capacity are as important as compute capability. GAUDI® incorporates the most advance HBM memory technology, supporting extremely high memory capacity of 32GB and total throughput of 1TB/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-Die RDMA (RoCE v2) and interface directly with mature and widely used Ethernet networking. The HL-2000 chip interconnect technology is based on 20 pairs of 56Gbps Tx/Rx PAM4 SerDes that can be configured as 10 ports of 100Gb Ethernet or 20 ports of 50GbE/25GbE, or any combination of both.

SynapseAI® Software Stack and Development Tools

The GAUDI® processor is supported by the SynapseAI® software stack and advanced development tools. SynapseAI is Habana's complete software stack custom designed to support Habana's Gaudi implementation. It is designed for flexibility and ease of development with its C-programmable Tensor Processor Core, custom-developed compiler and runtime and extensive, customizable kernel libraries. In its training incarnation, it is built for seamless integration with existing frameworks that both define a Neural Network for execution and manage the execution Runtime. SynapseAI® can be interfaced directly using either C or Python API. By supporting main frameworks, SynapseAI® enables users to unleash the power of Deep Learning by executing the algorithms efficiently using its high-level software abstraction. As part of its SW package, Habana provides an extensive set of TPC kernel libraries and opens its Tensor-Processing Core (TPC) for user programming, providing a complete TPC tool suite (debugger, simulator, compiler), which enables the user with flexibility to innovate and optimize to any unique end user requirements.

For more details on Gaudi's performance and scaling, see our Whitepaper.