GAUDI®2

GAUDI®2 AI MEZZANINE CARD

Gaudi®2

Gaudi®2 is Habana's second-generation AI deep learning Training and Inference Processor Mezzanine card.

The processor is built on the high-efficiency architecture of Gaudi®, now in 7nm process technology, to deliver leaps in performance, scalability, and power efficiency.

The heterogeneous compute architecture is built from a powerful centralized GEMM engine and a cluster of twenty-four 4th generation Tensor Processor Cores (TPCs). Gaudi2 supports FP32,TF32,FP16,BF16 and the new FP8 datatype. Gaudi2 also includes an independent media engine capable of decoding and postprocessing compressed media directly.

Gaudi2 integrates 96GB of HBM2e memories on chip, at 2.45GB/sec bandwidth, and integrates 48MB of SRAM.

With its unique, native integration, of 24x100GbE NICs on-chip, Gaudi2 offers the ability to scale using standard Ethernet, eliminating the need for extra components in the system, and completely avoids using proprietary interfaces, for scaling up or scaling out.

The Gaudi2 Mezzanine card part number is HL-225H.

System designers who plan to build an 8-Gaudi server can also purchase an HLBA-225H baseboard.

Processor Technology	Gaudi® HL-2080
Host Interface	PCIe Gen 4.0 x 16
Memory	96GB HBM2E
TDP	600W
Scale-Out Interconnect	ROMA (RoCE v2) 24x 100 Gbps
Form Factor	OCP Accelerator Module V1 .1 Compliant

Technology Innovation

The Gaudi2 processor features a unique combination of technology innovations such 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-Die ROMA (RoCE v2) and interface directly 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 the 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 start migrating their existing models to run on Gaudi2 with minimal code changes. https://developer.habana.ai/ is the hub for Habana developers from where they can find a wealth of information to get started with training on Gaudi AI processors, including tutorials, reference models, how-to guides, documentation and so on. It also hosts a Forum for the Habana developer community.

For more details on Gaudi's performance and scaling, see the Habana Gaudi2 Whitepaper.