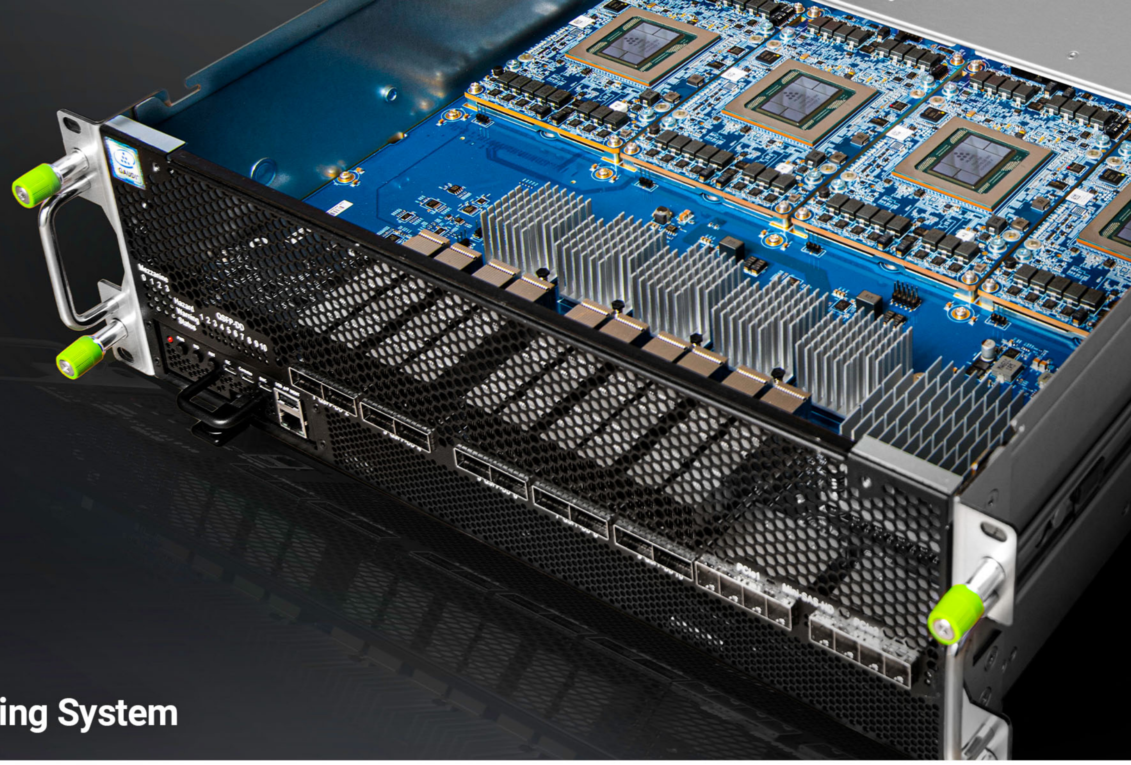


An Intel Company

GAUDI®

GAUDI® HLS-1H AI Training System

AI Performance with Ethernet Scale

Habana Labs System 1H (HLS-1H) brings to data centers a new level of AI compute performance and power efficiency, together with massive scale-out capability.

The HLS-1H incorporates four Gaudi HL-205 mezzanine cards, a Gen 4.0 PCIe switch and is built to be managed by an external host CPUs of your choice.

The HL-205 is OCP-OAM (Open Compute Project Accelerator Module) specification compliant. Each card incorporates the Gaudi HL-2000 processor that integrates 32GB HBM2 memory, and ten ports of 100GbE RoCE v2 RDMA.

The HLS-1H interfaces are 2x16 PCIe Gen4 that can be connected to an external Host server, and up to 40X100Gb Ethernet links (using 10 QSFP-DD connectors). The external Ethernet links can be connected to any switching hierarchy. Such configuration can be optimized to implement extra-large Model Parallelism in large scale and can easily handle Data Parallelism or a combination of Model and Data parallelism.

The GAUDI® processor delivers new levels of throughput and power efficiency on key benchmarks, thanks to innovative programmable architecture that is purpose-built for AI training, and is capable of scaling to a large number of processors while maintaining high throughput.

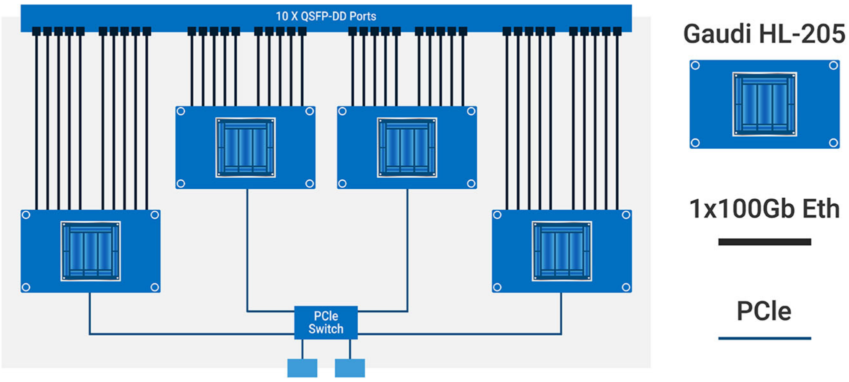
AI PROCESSING	4 x Gaudi HL-205
HOST INTERFACE	2 x PCIe Gen4.0
MEMORY	128GB HBM2
MAX POWER USAGE	2.3 KW
SCALE-OUT INTERFACE	RDMA (RoCE v2) 40X100Gbps 10 X SQFP-DD
SYSTEM DIMENSIONS	3U Height, 19” / 21”
OPERATING TEMP	5C to 35C

habana® An Intel Company

The HLS-1H is a successor server product to the HLS-1 which contains all-to-all 11.2 Terabits/sec of interconnect inside the box and designed for topologies which requiring mainly internal Ethernet communication. For more information regarding the HLS-1 system please refer to the HLS-1 product brief.

HLS-1H Block Diagram

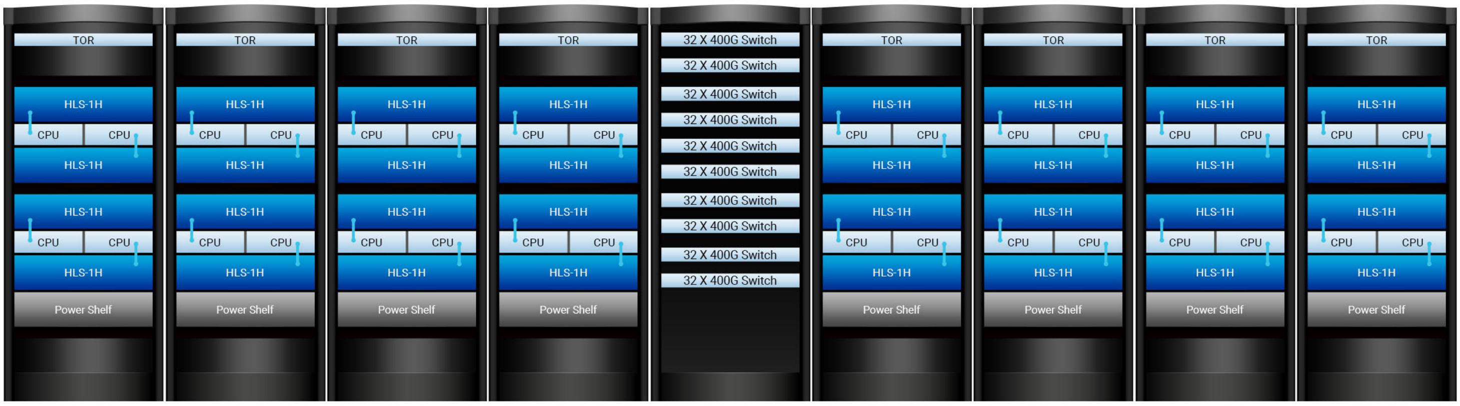
The HLS-1H interfaces are 2x16 PCIe Gen4 cables that can be connected to an external Host server, and up to 40X100Gb Ethernet links (using 10 QSFP-DD connectors). The external Ethernet links can be connected to any standard switching hierarchy. Such configuration can be optimized to implement extra-large Model Parallelism in large scale and can easily handle Data Parallelism or a combination of Model and Data parallelism.



HLS-1H Block Diagram

HLS-1H in a POD

A full POD based on the Gaudi HLS-1Hs standard interfaces provides unparalleled modularity and flexibility to efficiently support the growing demands of AI training compute infrastructure.
The example here shows 16 Gaudi HLS-1H systems (four systems per rack), each HLS-1H server is connected to a host CPU and scale-out its 40x100Gb networking links in one hop using ten standard ethernet switches (32 x 400G) to enable symmetric non-blocking communication of the POD total of 128 Gaudi processors. Larger clusters can be built in the same manner forming a much larger training farm, that can hold hundreds or thousands of Gaudi processors.



HLS-1H based POD

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