habana GAUDI™

Gaudi HLS-1 AI Training System

AI Performance with Ethernet Scale

Habana Labs System 1 (HLS-1) brings to data centers a new level of AI compute performance and power efficiency, together with unprecedented scalability.

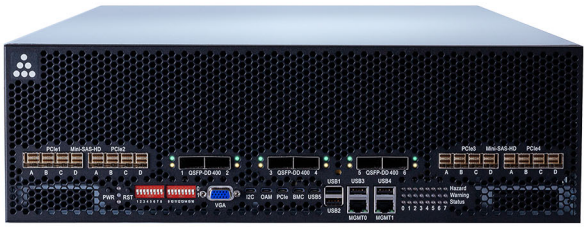
The HLS-1 incorporates eight Gaudi HL-205 mezzanine cards, two Gen 4.0 PCIe switches and is built to be managed by an external host CPU's of your choice.

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

The HLS-1 contains all-to-all 11.2 Terabits/second of internal interconnect, thus not requiring use of an external Ethernet switch.

The HLS-1 offers 24 x 100GbE RoCE RDMA for further scaling beyond eight Gaudis, to racks and clusters of Gaudis by utilizing external off-the-shelf Ethernet switching. Various system architectures can be built using HLS-1 (or similar Gaudi-based servers) to reach thousands of processors.



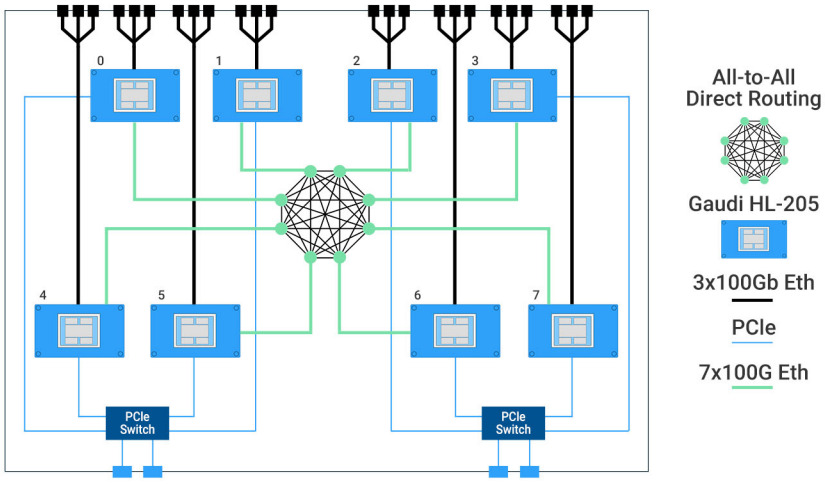
AI PROCESSING	8 X Gaudi HL-205
HOST INTERFACE	4 X PCIe Gen 4.0
MEMORY	256GB HBM2
MEMORY BANDWIDTH	8TB/s
ECC PROTECTED	Yes
MAX POWER USAGE	3 kW
SCALE-OUT INTERFACE	RDMA (RoCE v2) 24X100Gbps 6 X QSFP-DD
SYSTEM DIMENSIONS	3U Height, 19"
OPERATING TEMP	5C to 35C

HLS-1 Block Diagram

The HLS-1 system contains eight HL-205 OAM Mezzanine cards and dual PCIe switches. The all-to-all connectivity allows training across all eight Gaudi processors without requiring an external Ethernet switch.

PCIe can be dedicated to Host communication.

Customers can choose any desired ratio of CPU to AI acceleration, depending on their application, decoupling the CPU server from the acceleration system, and upgrading them independently over time.



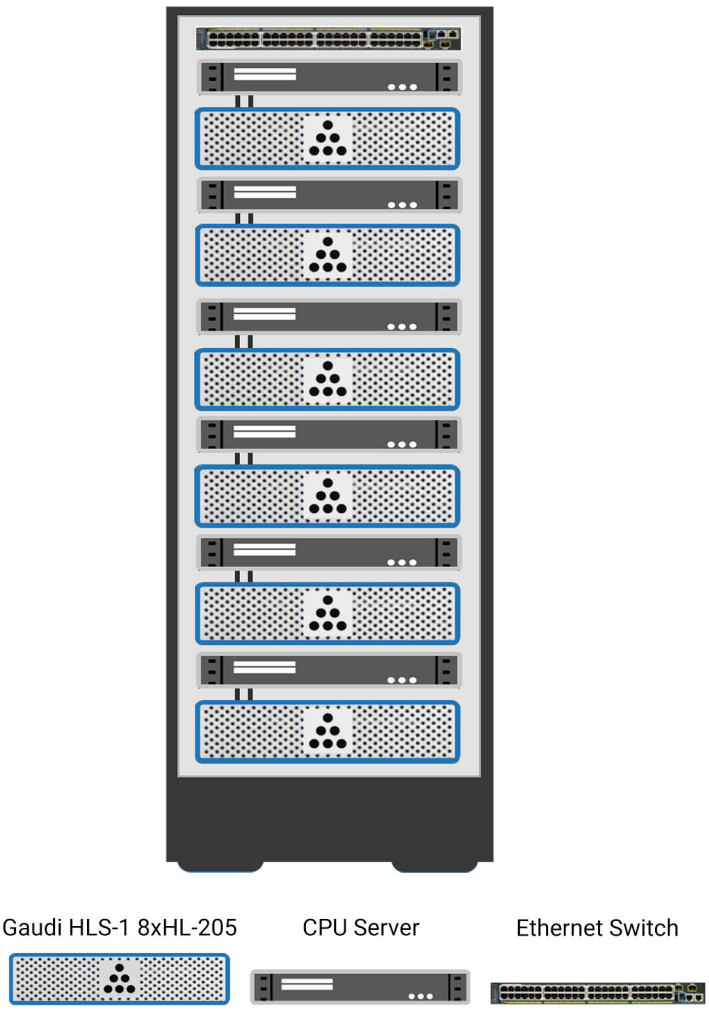
HLS-1 in a Rack

A full rack based on the Gaudi HLS-1's standard interfaces provides unparalleled modularity and flexibility to efficiently support the growing demands of AI compute infrastructure.

The example here shows how six Gaudi systems (48 Gaudi processors in total) can be connected to a single Ethernet switch. That switch can be further connected to other racks, in order to form a much larger training farm, that can hold hundreds or thousands of Gaudi processors.

The HLS-1 is based on the modular HL-205 OCP-OAM compliant mezzanine cards. Customers may choose to design their own system, using a different balance between internal connectivity and external scale-out connectivity, to fit their workloads.

In this example, each HLS-1 is connected to the switch with four cables, carrying a total of 16x100GbE, two ports per Gaudi processor. Customers can easily adapt the connectivity capacity they use (number of cables and switch capacity) to their workloads and other system constraints.



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

© 2019 Habana Labs Ltd. All rights reserved. Habana Labs, Habana, the Habana Labs logo, Goya, Gaudi, Pure AI, TPC and SynapseAI are trademarks or registered trademarks of Habana Labs Ltd. All other trademarks or registered trademarks and copyrights are the property of their respective owners.