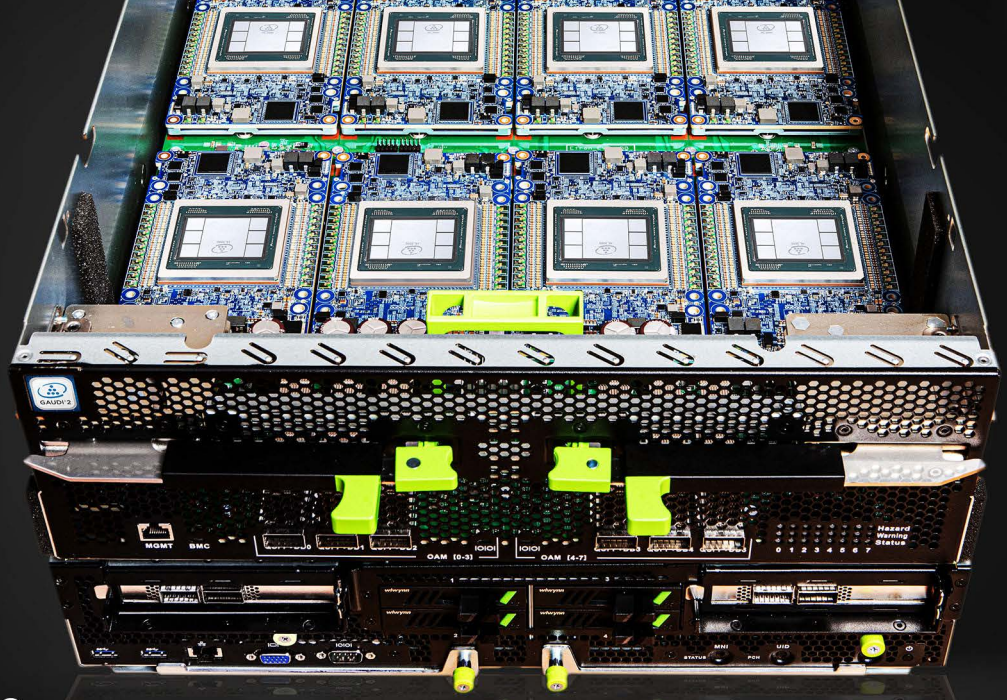
An Intel Company

GAUDI®2



HLS-GAUDI®2 Deep Learning Server

AI Performance with Ethernet Scale

Habana Labs HLS-Gaudi2 system brings to data centers a new level of deep learning performance and scalability.

The system incorporates eight Gaudi2 mezzanine cards, two Gen 4.0 PCIe switches and a standard dual-socket Xeon Icelake CPU host subsystem with NVME storage and host connectivity.

The Gaudi2 is OCP-OAM (Open Compute Project Accelerator Module) 1.1 specification compliant. Each card incorporates the Gaudi HL-2080 processor that integrates 96GB HBM2E memory, and twenty four NICs of 100GbE RoCE v2 RDMA.

The Gaudi2 processor delivers (as demonstrated in its May 2022 announcement) leadership Training performance for key vision and language workloads thanks to innovative, programmable architecture that is purpose-built for AI training and inference, and is capable of scaling to a large number of processors using standard interfaces and with full software stack, reference models and how-to guides.

The 8 Gaudi2 cards are interconnected internally with a non-blocking all-to-all connectivity using 3x100GbE RoCE ports from every Gaudi2 to each of the other 7 processors. In addition, the system offers 24x100GbE RoCE RDMA for further scaling-out, to racks and clusters of Gaudi2-based nodes by utilizing external off-the-shelf Ethernet switching. Various cluster architectures can be built using similar servers, to scale the AI training and inference cluster utilizing thouhsands of Gaudi2s.

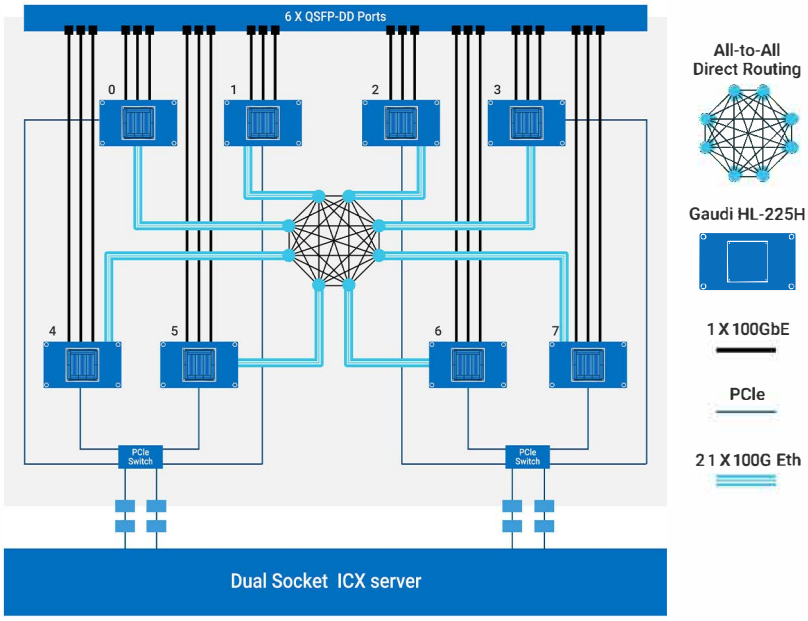
AI PROCESSING	8 x Gaudi2
HOST	Integrated Dual-Socket XEON Icelake subsystem
MEMORY	768GB HBM2E
MAX POWER	9.8 KW
SCALE-OUT INTERFACE	RDMA (RoCE v2) 24 X 100GbE 6 X QSFP-DD
SYSTEM DIMENSIONS	19”
SOFTWARE	SynapseAI TensorFlow & PyTorch

HLS-Gaudi2 Block Diagram

The system contains eight Gaudi2 OAM Mezzanine cards and an integrated dual socket ICX host server subsystem. 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.

The eight Mezzanine cards, each with 1x6 PCIe 4.0 are connected via PCIe switches to the CPU headnote (Dual socket Intel Xeon IceLake 8380 CPUs with 1TB memory, 4 x 7.68TB NVME SSDs and two 2x100GbE PCIe Host NICs)



HLS-Gaudi2 Block Diagram

HLS-Gaudi2 in a Rack

A full rack based onthe HLS-Gaudi2's standard interfaces provides unparalleled modularity and flexibility to efficiently support the growing demands of AI compute infrastructure.

The example here shows how four Gaudi systems (32 Gaudi processors in total) can be connected to a single Ethernet data 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.

For Maximum scale-out bandwidth, each server is connected to a standard Ethernet switch with six cables, carrying a total of 24x100GbE, three ports per Gaudi2 processor. Customers can easily adapt the connectivity they use (using two, four or six cables and switch capacity) to their workloads and other system constraints.

For more details on Gaudi2's performance and scaling, see our habana.ai site.