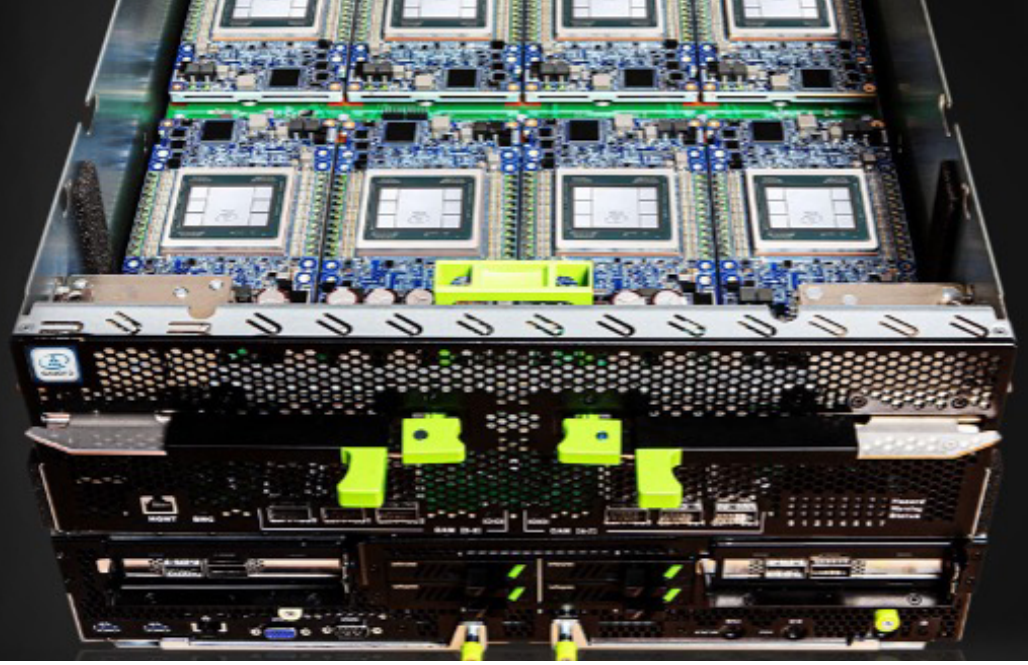
habana
An Intel Company

intel
GAUDI



Intel® HLS-Gaudi®2 AI Accelerator Server

AI Performance with Ethernet Scale

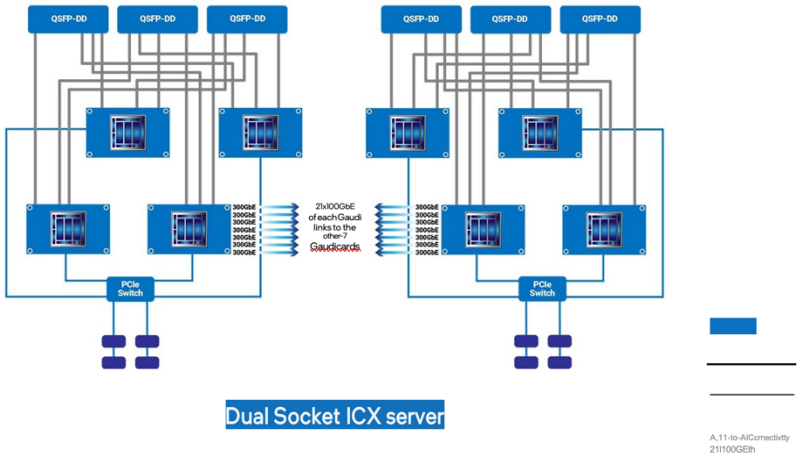
The Intel® HLS-Gaudi®2 system brings to data centers a new level of deep learning performance and scalability. The system incorporates eight Intel® Gaudi®2 mezzanine cards, two Gen 4.0 PCIe switches and a standard dual-socket Intel® Xeon® 3rd Generation Scalable processor as CPU host subsystem with NVME storage and host connectivity. The Intel Gaudi2 server is OCP-OAM (Open Compute Project Accelerator Module) 1.1 specification compliant. Each card incorporates the Intel® Gaudi® HL-2080 processor that integrates 96GB HBM2E memory, and 24 NICs of 100GbE RoCE v2 RDMA.

Intel Gaudi2 AI accelerators deliver 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 Intel Gaudi2 cards are interconnected internally with non-blocking all-to-all connectivity dedicating 21x100GbE RoCE ports from every Intel Gaudi2 to the other 7 processors. In addition, the system offers 24x100GbE RoCE RDMA for further scaling-out, to racks and clusters of Intel 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 thousands of Intel Gaudi2s.

	Intel® HLS-Gaudi®2 AI Accelerator Server
AI PROCESSING	8 x Intel Gaudi2 AI accelerators
HOST	Integrated Dual-Socket XEON Icelake subsystem
MEMORY	768GB HBM2E
MAX POWER USAGE	7.5KW
SCALE-OUT INTERCONNECT	RDMA (RoCE v2) 24x100 GbE 6xQSFP-DD
SYSTEM DIMENSIONS	19”
SOFTWARE	SynpaseAI®, TensorFlow & PyTorch

Intel® HLS-Gaudi®2 Block Diagram

The system contains eight Intel® Gaudi®2 OAM Mezzanine cards and an integrated dual socket ICX host server subsystem. The all-to-all connectivity allows training across all eight Intel Gaudi2 processors without requiring an external Ethernet switch. PCIe can be dedicated to Host communication. The eight Mezzanine cards, each with x16 PCIe 4.0 are connected via PCIe switches to the CPU headnote (Dual socket Intel® 3rd Generation Xeon® 8380 CPUs with 1TB memory, 4 x 7.68TB NVME SSDs and two 2x100GbE PCIe Host NICs).



HLS-Gaudi2 Block Diagram

Intel® HLS-Gaudi®2 in a Rack

A full rack based on the Intel HLS-Gaudi2 standard interfaces provides unparalleled modularity and flexibility to efficiently support the growing demands of AI compute infrastructure. The example here shows how four Intel Gaudi2 systems (32 Intel Gaudi2 accelerators in total) can be connected to a single Ethernet data switch. That switch can be further connected to other racks to form a much larger training farm, that can hold hundreds or thousands of Intel Gaudi2 accelerators.

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 Intel Gaudi2 accelerator. 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 Intel Gaudi performance and scaling, see the Intel Gaudi2 AI Accelerator Whitepaper.

habana An Intel Company