

Intel Gaudi 3 AI Accelerator

Also known as: Gaudi 3 · HL-338 · Gaudi 3 PCIe card · HL-325L · Gaudi 3 OAM · HLB-325 · Gaudi 3 UBB

Gaudi 3 is Intel's dedicated AI accelerator: two 5nm compute dies with 8 matrix engines (MME), 64 programmable tensor processor cores (TPC), 128 GB of HBM2e at 3.7 TB/s, and 24 x 200 GbE RoCE v2 ports on the chip itself, so scale-up and scale-out both run over standard Ethernet instead of a proprietary fabric. It ships in two form factors: the HL-338, a 600 W, dual-slot, passively cooled PCIe Gen5 card for standard servers, and the HL-325L, a 900 W OCP OAM 2.0 module that goes eight to a node on the HLB-325 universal baseboard. GPU Smith treats the HL-338 as the part to evaluate for single-node inference in existing server fleets, and the OAM/HLB-325 path as the Ethernet-fabric alternative to an 8-GPU HGX node. The catch is software: it runs PyTorch, vLLM and Hugging Face through Intel's Gaudi stack, not CUDA.

128 GB HBM2e Memory · 8 stacks, 3.7 TB/s	1,678 TFLOPS FP8 / BF16 matrix · vendor peak, March 2025 paper; 2024 revision said 1,835	24x 200 GbE RoCE v2 On-die networking · 1,200 GB/s bidirectional, Ethernet only
600 W, dual-slot FHFL PCIe card (HL-338) · passive, PCIe 5.0 x16, 10.5 in	900 W OAM module (HL-325L) · OCP OAM 2.0, eight per HLB-325 baseboard	96 MB On-die SRAM · 19.2 TB/s

COMPUTE01/05	
Process	TSMC 5nm · two compute dies, software sees one device
Matrix engines (MME)	8
Tensor processor cores (TPC)	64 · VLIW SIMD vector cores, programmable
FP8 / BF16 matrix	1,678 TFLOPS · peak, same figure for OAM and PCIe in Intel's table
FP16 / TF32 matrix	459 TFLOPS · peak
FP32 matrix	229 TFLOPS · peak
BF16 / FP16 vector (TPC)	28.7 TFLOPS · FP8 vector 57.3, FP32 vector 14.3
Data types	FP32, TF32, BF16, FP16, FP8 · FP8 in E4M3 and E5M2; no FP4
Media decoders	14 · on-die image and video decode

MEMORY		02/05
Capacity	128 GB HBM2e · 8 x 16 GB stacks	
Bandwidth	3.7 TB/s · peak	
On-die SRAM	96 MB · usable as L2 or last-level cache	
SRAM bandwidth	19.2 TB/s · 12.8 TB/s read / 6.4 TB/s write	
Pooled memory, 4-card quad	512 GB · HL-338 x4 on HLTB-304 top board	
FORM FACTORS		03/05
HL-338 PCIe card	Full-height, dual-slot, 10.5 in · PCIe CEM 5.1, PCIe 5.0 x16, 128 GB/s bidirectional	
HL-338 power	600 W TDP · passive, air-cooled; chassis airflow required	
HL-325L OAM module	OCP OAM 2.0 mezzanine · PCIe 5.0 x16 host link, 48 x 112G PAM4 SerDes	
HL-325L power	900 W TDP · passive; 2024 paper revision listed up to 1.2 kW liquid-cooled, March 2025 revision says 900 W	
HLB-325 baseboard	8x HL-325L · 585 x 417 mm PCB, 54 V and 12 V standby input	
HLTB-304 top board	Bridges 4x HL-338 · HLTB-304A plus HLTB-304B links two quads (8 cards)	
INTERCONNECT (ETHERNET)		04/05
Per-chip NIC	24x 200 GbE RoCE v2 · integrated RDMA engine, no NVLink-style proprietary link	
HLB-325 scale-up	21x 200 GbE per OAM · 3 links to each of the other 7 OAMs, all-to-all on the PCB	
HLB-325 scale-out	3x 200 GbE per OAM · 4.8 Tb/s per baseboard on 6x OSFP 800 GbE	
HL-338 scale-up	18x 200 GbE via top board · 6 links to each of 3 peers, 900 GB/s per card	
HL-338 scale-out	Host NIC only · Intel recommends at least 800 Gb/s per quad; NIC must support LibFabric and GDR	
Reference cluster	512 nodes, 4,096 accelerators · 3-ply leaf/spine on 64-port 800 GbE switches, Intel design	

PLATFORM & SOFTWARE		05/05
Software	Intel Gaudi software stack · PyTorch, vLLM, Hugging Face; automated FP8 quantization	
OEM systems	Dell, HPE, Supermicro · per Intel product page	
Cloud	IBM Cloud, Denvr Dataworks · per Intel product page	
Rack-scale reference design	Up to 64 accelerators per rack · 8.2 TB HBM, liquid-cooled, announced Computex 2025	
Availability	OAM GA October 2024; PCIe announced for H2 2025 · reported (GA) and Intel release (PCIe)	
PROCUREMENT		
Typical lead time		
Not published. OAM systems ship through server OEMs (Intel lists Dell, HPE and Supermicro); Intel said the HL-338 PCIe card would be available in the second half of 2025. A reseller listing for the HL-338 showed out of stock, quote-only, when checked on 2026-10-14.		
Pricing		
HL-338 PCIe card: not published; quote-only. OAM: Intel's June 2024 list price for a kit of eight Gaudi 3 accelerators on a universal baseboard was \$125,000, about \$15,600 per accelerator (derived), reported. That is a 2024 list figure, not a current street price.		
■ The HL-338 is sold as a bare card through distribution, but the reseller listing we checked is non-cancelable and non-returnable; confirm chassis qualification before ordering. ■ A 4-card or 8-card HL-338 build also needs the HLTB-304A (and HLTB-304B for two quads) top boards and a host NIC for scale-out; price those into the quote, they are not optional for pooled-memory serving. ■ The \$125,000 kit price covers accelerators and baseboard only, not the host CPU board, chassis, NICs or switches. ■ Intel's AI roadmap has moved past Gaudi (see engineering notes); ask the vendor for a written support and software-update horizon before a multi-year commitment.		
FIELD NOTES		
■ The headline FP8/BF16 number changed between revisions of Intel's own technical paper: 1,835 TFLOPS in the 2024 copy, 1,678 TFLOPS in the March 2025 revision, for both OAM and PCIe. Use 1,678 and treat both as peak, not sustained. Intel's table gives the 600 W PCIe card the same peak as the 900 W OAM; sustained clocks at 600 W are not published, so benchmark before assuming parity. ■ Ethernet everywhere is the real architectural difference. The 24 on-die 200 GbE ports carry both scale-up and scale-out, so an 8-OAM node plugs into standard 800 GbE switches with no InfiniBand or NVLink switch. On the PCIe card that only applies within a quad via the top board; beyond four cards, traffic goes through a host NIC. ■ An early 2024 HL-338 preview reported external QSFP-DD ports on the card; Intel's current product brief and technical paper list scale-out through the host NIC only. Spec the current part from the brief, not from launch photos. ■ There is no FP4 support, and 128 GB of HBM2e at 3.7 TB/s is a memory generation behind the HBM3 and HBM3e on H100 SXM and later NVIDIA parts. The case for Gaudi 3 is price per accelerator and an open Ethernet fabric, not peak throughput against Blackwell-class parts. ■ Roadmap risk is real. Intel still sells Gaudi 3, but Falcon Shores became an internal development vehicle, Crescent Island (inference GPU, 160 GB LPDDR5X) targets customer sampling in H2 2026, and Jaguar Shores is the longer-term GPU architecture. Intel's named next parts are GPUs, not Gaudi, so do not expect code built on Gaudi-specific kernels to carry forward.		

QUESTIONS WE GET ON THIS PART
How much does Intel Gaudi 3 cost? Intel listed a kit of eight Gaudi 3 OAM accelerators on a universal baseboard at \$125,000 in June 2024, about \$15,600 per accelerator. Pricing for the HL-338 PCIe card is not published; resellers quote it on request.
What is the difference between the HL-338 and the HL-325L? Both use the same Gaudi 3 chip with 128 GB HBM2e. The HL-338 is a 600 W, dual-slot, passively cooled PCIe 5.0 card for standard servers, scaling to four cards over a top board and beyond that through a host NIC. The HL-325L is a 900 W OCP OAM module mounted eight to an HLB-325 baseboard, with all 24 on-die 200 GbE ports wired for all-to-all scale-up and scale-out.
Does Gaudi 3 use NVLink or InfiniBand? Neither. Each chip has 24 x 200 GbE RoCE v2 ports built in, and Intel's cluster reference design uses standard 800 GbE leaf and spine switches. On the eight-OAM baseboard, 21 ports per chip go to the other seven chips and 3 go to six OSFP 800 GbE scale-out connectors.
Does Gaudi 3 run CUDA? No. It runs PyTorch, vLLM and Hugging Face models through Intel's Gaudi software stack. Standard model code usually ports, but custom CUDA kernels must be rewritten, and Intel's next accelerators (Crescent Island, Jaguar Shores) are a different architecture.

REFERENCES

[1] https://cdrdv2-public.intel.com/817488/Gaudi%203%20PCIe%20Product%20Brief_RB_1_V6.pdf

[2] <https://www.presidio.com/wp-content/uploads/2025/06/gaudi-3-ai-accelerator-white-paper.pdf>

[3] <https://www.wwt.com/api-new/attachments/670ed4bbd05857f06688b657/file>

[4] <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi.html>

[5] <https://www.intc.com/news-events/press-releases/detail/1741/computex-2025-intel-unveils-new-gpus-for-ai-and>

[6] <https://www.storagereview.com/news/intel-targets-ai-inference-at-ocp-2025-with-crescent-island-gpu-and-gaudi-3-racks>

[7] <https://www.jonpeddie.com/news/intel-gpu-esque-gaudi-3-to-compete-with-h100-h200-gpus-on-price/>

[8] <https://www.jonpeddie.com/techwatch/intel-resets-its-ai-accelerator-roadmap/>

[9] <https://www.servethehome.com/intel-gaudi-3-going-ga-for-scale-out-ai-acceleration/>

[10] <https://www.sabrepc.com/HL-338-Intel-S117208843>