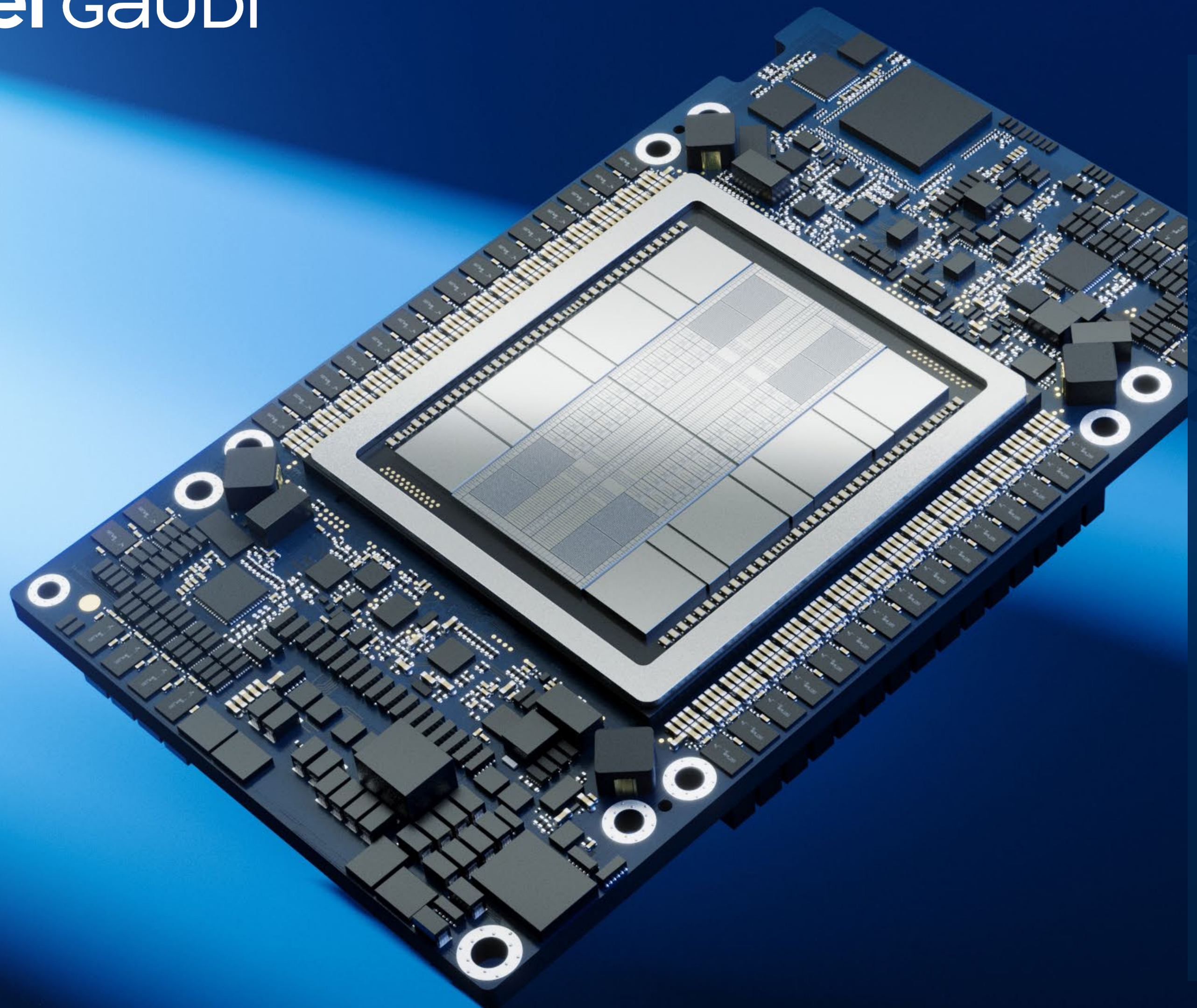
intel GAUDI



Bringing Choice to Gen AI with Performance, Scalability, and Efficiency

Intel® Gaudi® 3 AI accelerator

Version 2.6

Table of Contents

Intel® Gaudi® 3 AI Accelerators

Customer Challenges

Customer Challenges with AI Compute Solutions

Need more choice
other than single-source GPUs

Locked-in
with proprietary software and networking

Ability to scale
while containing the costs of infrastructure

Maximize efficiency
yet still solve my business challenges

How Intel® Gaudi® 3 AI Accelerator Addresses Enterprise Challenges

Need more choice
other than single-source GPUs

- Intel® Gaudi® 3 AI Accelerator outperforms H100 and H200 for inference price/ performance of LLMs
- Lower hardware cost and no CUDA licensing costs
- Networked with industry-standard, cost-effective Ethernet

Locked-in
with proprietary software and networking

- Software migration in as few as three lines of code
- Community-based, open-source software stack
- Non-proprietary-based network solution

Ability to scale
while containing the costs of infrastructure

- Readily supports demanding Gen AI workloads from 1 to 1000s of nodes
- Easily and cost-effectively integrate into Ethernet-based networks
- High-efficiency cluster scaling drives cost savings

Maximize efficiency
yet still solve my business challenges

- Higher price-performance over H100 and H200
- Integration of open software frameworks drives developer productivity
- Cost-efficient, scalable networking fabric

intel GAUDI

Broad AI Application Support

With focus on multi-modal, LLM, and RAG

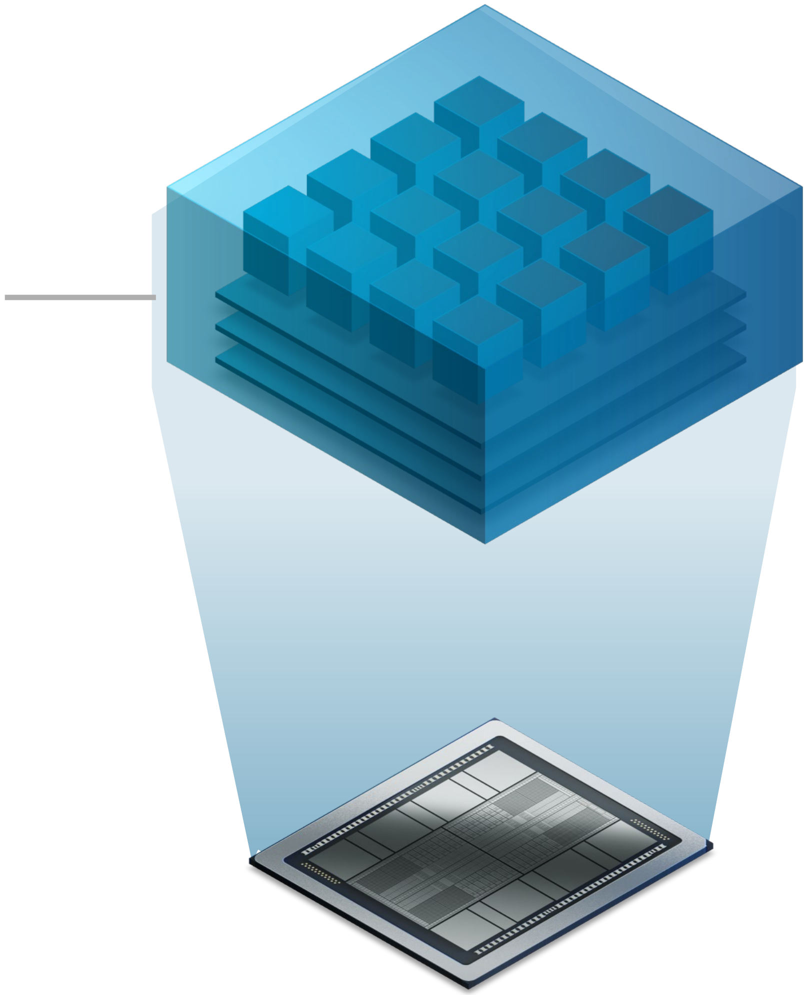
AI Applications

AI Functions

- 3D Generation
- Text Generation
- Classification
- Video Generation
- Sentiment
- Translation
- Image Generation
- Summarization
- Q&A

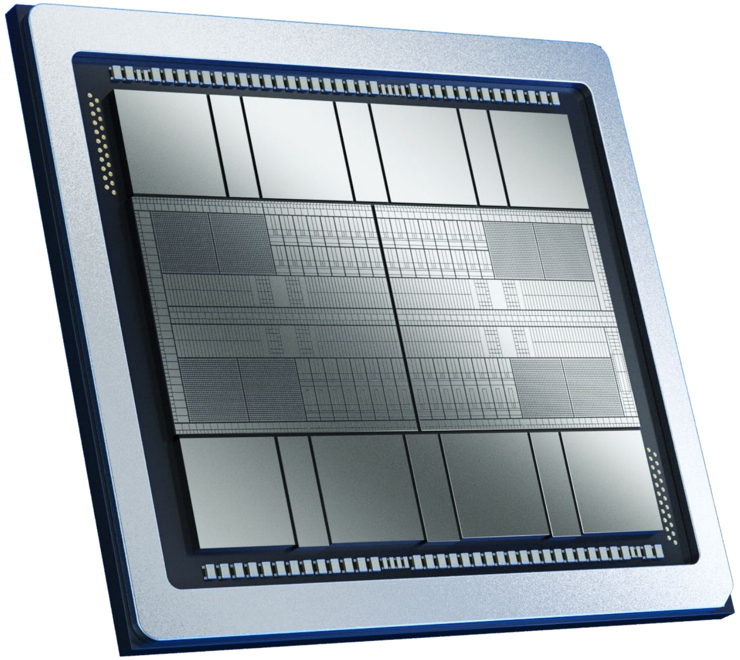
Core Capabilities

- Multi-modal Models
- LLM
- RAG



Introducing Intel® Gaudi® 3 AI Accelerator

Architected for Gen AI Performance & Productivity



Increased memory for LLM efficiency and cost effectiveness

128GB
HBM capacity, 3.7 TB/s B/W

96MB
SRAM, 12.8 TB/s SRAM B/W

Massive, flexible, on-chip networking

Open standard vs. proprietary InfiniBand

24 x 200 GbE
Industry-standard RoCE Ethernet ports

PCIe 5
x 16

Designed for AI

Driving greater efficiency & performance

64
Tensor Processor Cores

8
Matrix Math Engines

Intel® Gaudi® 3 AI Accelerator Product Specs

Advances over Intel Gaudi 2 AI accelerator

	Feature	Intel Gaudi 2 AI accelerator	Intel Gaudi 3 AI accelerator
Availability		Now	Now
Architectural Features	HBM Capacity	96 GB	128 GB
	HBM Bandwidth	2.45 TB/sec	3.7 TB/sec
	Industry Standard Ethernet	24 x 100GbE Ports	24 x 200GbE Ports
	Thermal Design Power	600W	900W
	PCIe	x16 Gen 4	x16 Gen 5
	OCP OAM Version Baseboard	OAM 1.1 x8 Baseboard (P/N HLBA-225)	OAM 2.0 x8 Baseboard (P/N HLB-325)
	Process Technology	7nm	5nm
	Data-types	FP32, TF32, Bfloat16, FP16, FP8, INT32, INT16, INT8	With Much Higher TFLOPs (4x 16bit, 2x FP8)
	Cooling	Air	Air: HL-325L

Intel® Gaudi® 3 AI Accelerator

Gen AI. Your way.

Competitive Gen AI Price-Performance

- For the latest public performance data....
- Visit: https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html

Freedom to Scale without Lock-in

- Open standard ethernet networking vs. proprietary InfiniBand
- 24x200 GbE ports of industry-standard RoCE on every Intel® Gaudi® 3 AI Accelerator
- 33% more I/O peak throughput vs H100 for massive scale-up within the server[3]

Open Development on Gen AI platforms

- Integrated open-source PyTorch framework with optimized model library on Hugging Face
- Migrate models on open software from H100 with as few as 3 lines of code

1-2 Source: Intel measured results vs H100 data sources: https://github.com/NVIDIA/TensorRT-LLM/blob/main/docs/source/performance/perf-overview.md for 128-2048 input-output sequences
Intel results obtained in September 9th 2024. Results may vary. Pricing estimates based on publicly available information and Intel internal analysis
3 900 GB/s NVLink connectivity on H100 vs. 1200 GB/s on Intel® Gaudi® 3 AI Accelerator

Ideal Customer Fits for Intel® Gaudi® 3 AI Accelerator

Focused on on-premise deployment

All Gen AI (LLM and RAG) workloads, training and deployment of models >10B to 1T+

Short lead times and desire for a second source

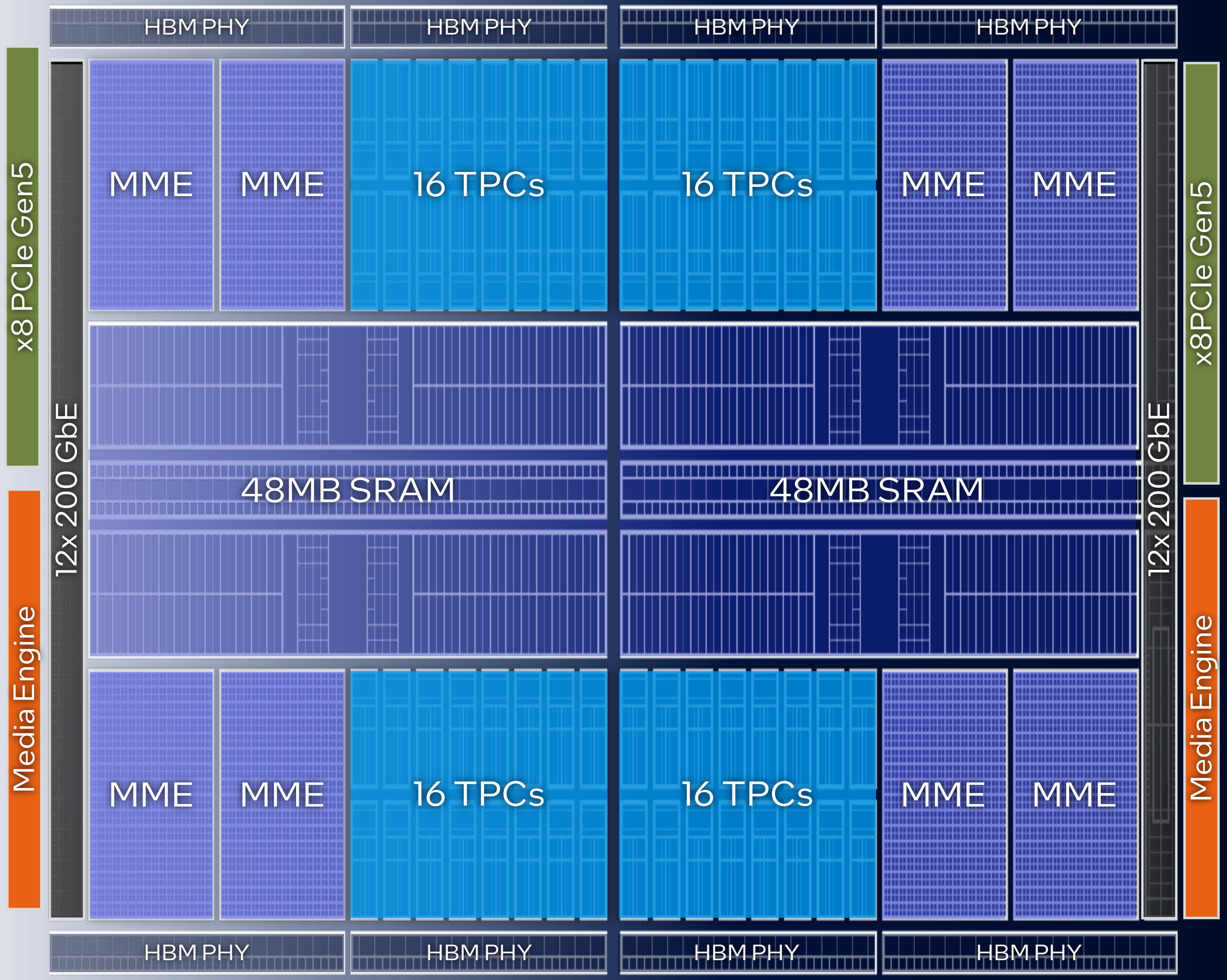
Deployment size ranging from a single x8 system to large-scale clusters

Want a network based on a single industry-standard Ethernet

Interest in open-source stack using industry-standard frameworks (i.e., PyTorch and DeepSpeed)

Intel® Gaudi® 3 AI Accelerator Spec and Block Diagram

Feature/Product	Intel® Gaudi® 3 Accelerator
BF16 Matrix TFLOPs*	1678
FP8 Matrix TFLOPs*	1678
BF16 Vector TFLOPs	28.7
MME Units	8
TPC Units	64
HBM Capacity	128 GB
HBM Bandwidth	3.67 TB/s
On-die SRAM Capacity	96 MB
On-die SRAM Bandwidth RD+WR (L2 Cache)	19.2 TB/s
Networking	1200 GB/s bidirectional
Host Interface	PCIe Gen5 x16
Host Interface Peak BW	128 GB/s bidirectional
Media Engine	Rotator + 14 Decoders (HEVC, H.264, JPEG, VP9)



*TFLOPs projected maximum

Matrix Multiplication and Vector Engines

Matrix Multiplication Engine (MME): designed for AI efficiency

Configurable, not programmable

Each MME is a large-output stationary systolic array
- 256x256 MAC structure w/ FP32 accumulators
- 64k MACs/cycle for BF16 and FP8

Large systolic array reduces intra-chip data movement, increasing efficiency

Internal pipeline to maximize compute throughput

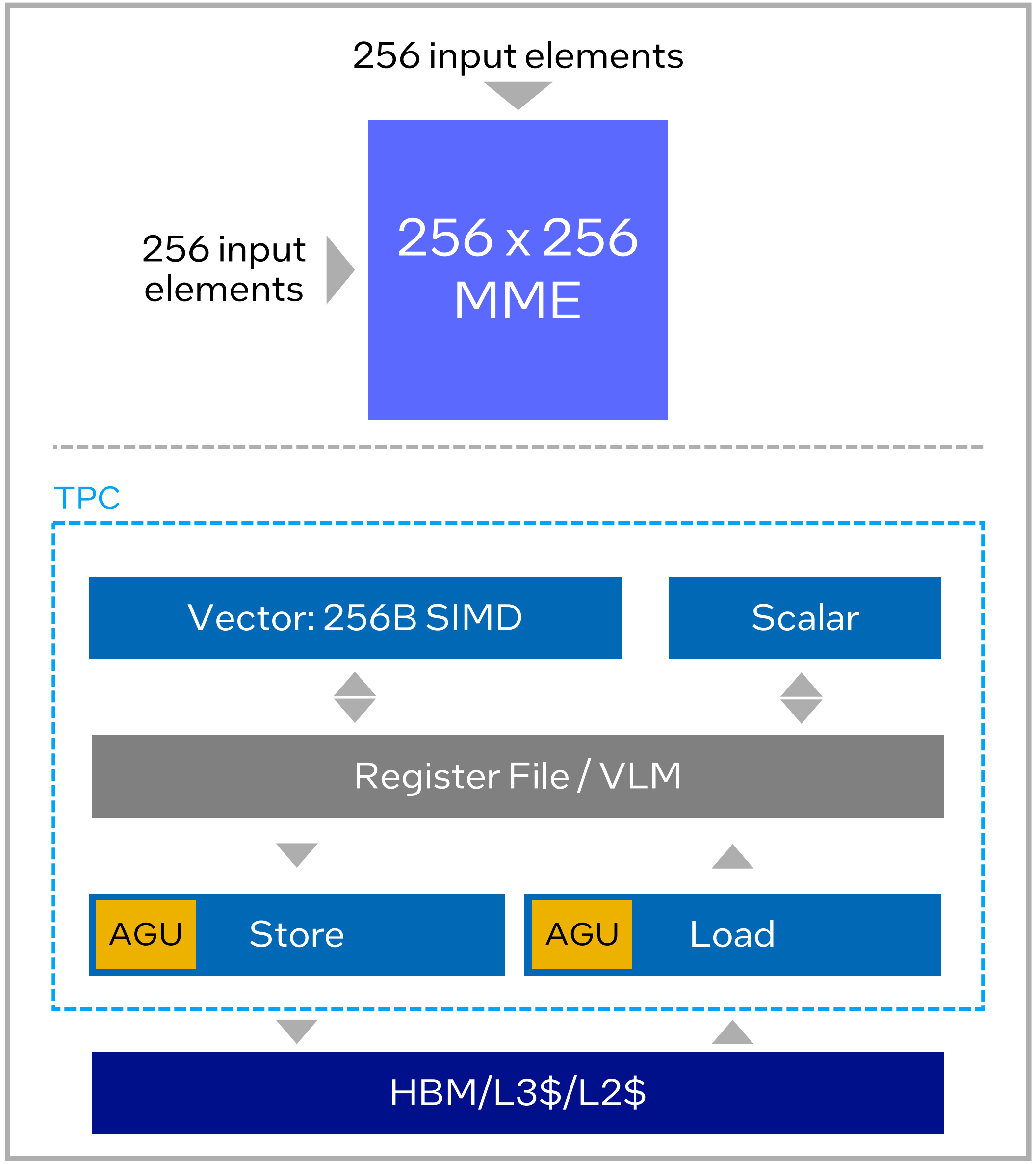
Tensor Processing Core (TPC): 256B-wide SIMD Vector Processor

Programmable: C enhanced with TPC intrinsics

VLIW with 4 separate pipeline slots: Vector, Scalar, Load & Store

Integrated Address Generation Unit for HW-accelerated address generation

Supports main 1/2/4-Byte datatypes: Floating Point and Integer



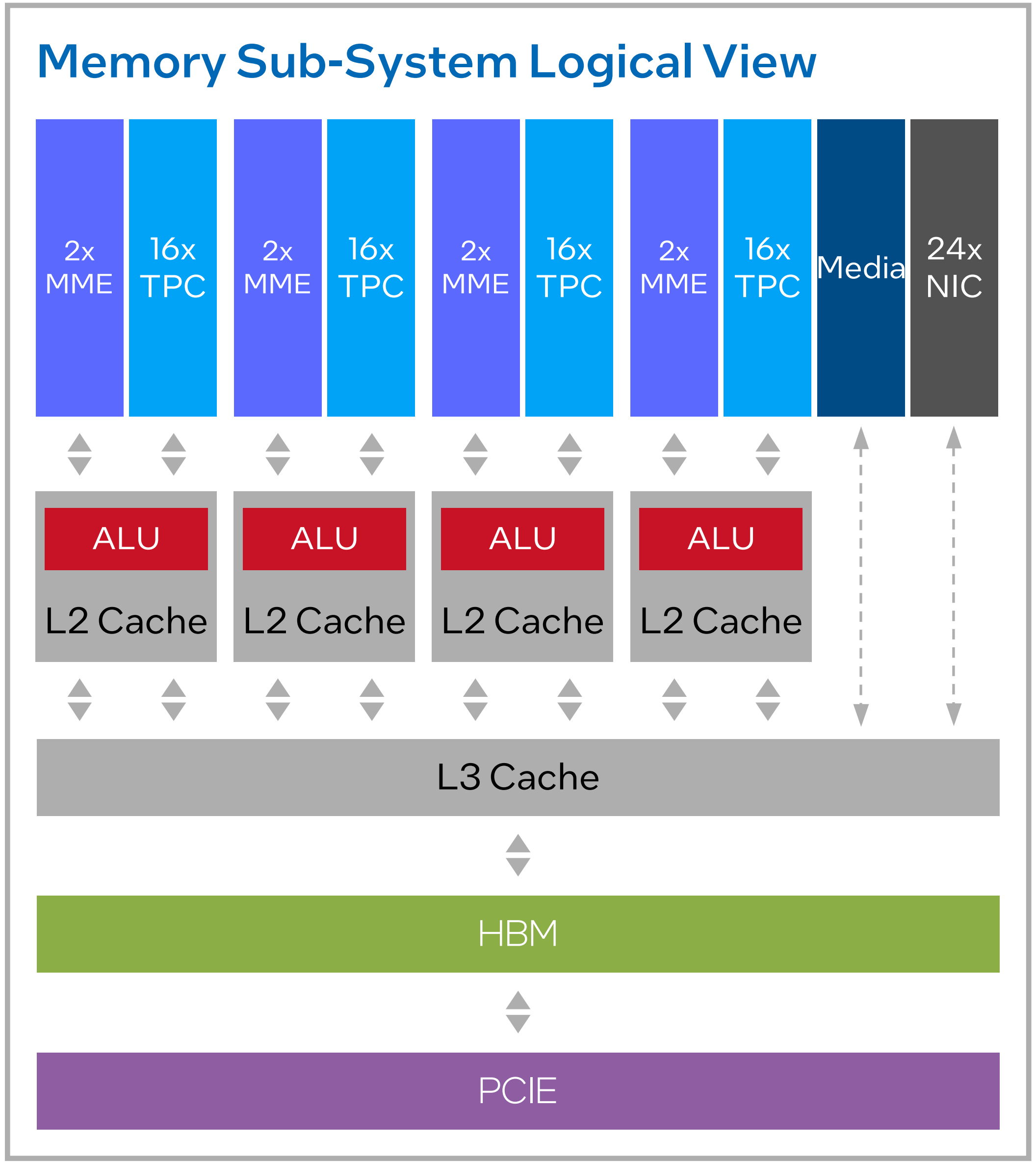
Memory Sub-System

Unified Memory Space of L2 / L3/ HBM

Near Memory Compute:
- Add / Sub
- Max / Min

Usage of Memory Context ID (MCID) to tag cache lines with shared algorithmic usage

Cache Directives:
- No-$, L2$, L3$, L2$+L3$
- Discard: Invalidate all same-MCID CLs
- Degrade: Reset same-MCID CLs hit count



MME-TPC Parallelism via Pipelining

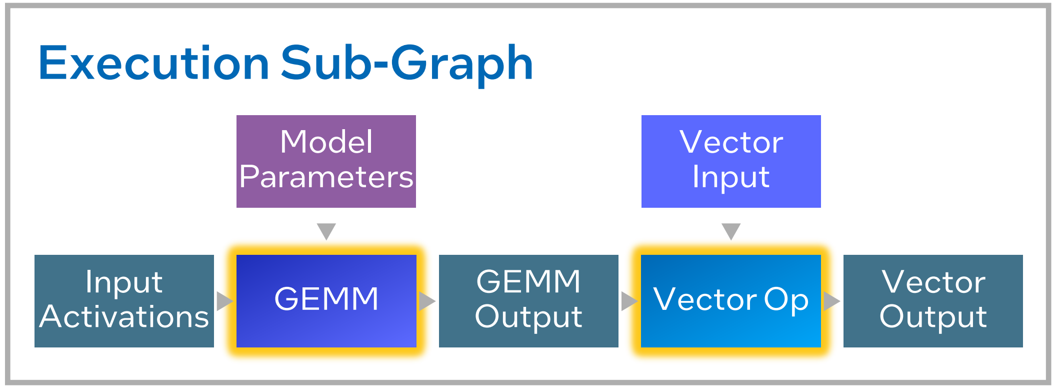
Graph Compiler orchestrates MME & TPC parallelism

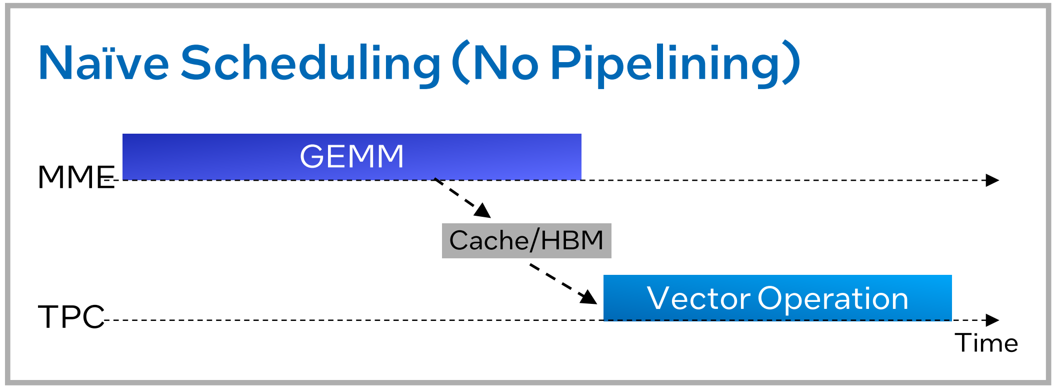
- Long chunks of work are split to smaller independent slices
- Pipeline through cache with producer → consumer relation

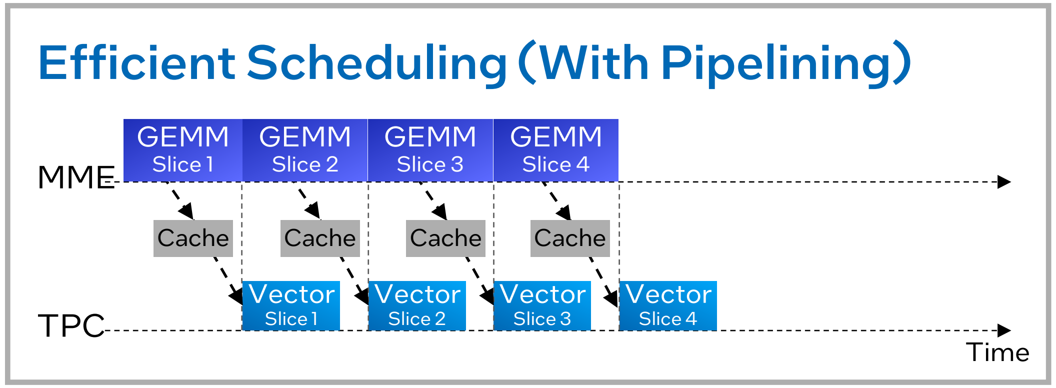
Slice size is determined to balance between the following:

- High compute utilization
- Maximize engine parallelism
- Fit within the cache capacity

- NOC fabric was designed to support the parallel work of MME and TPC
- Pipelining is the main enabler for reaching high compute utilization

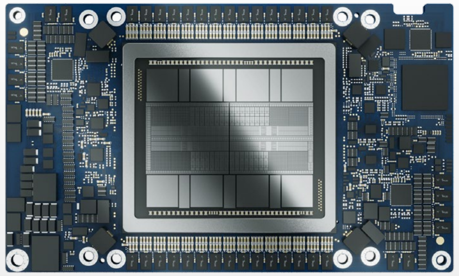






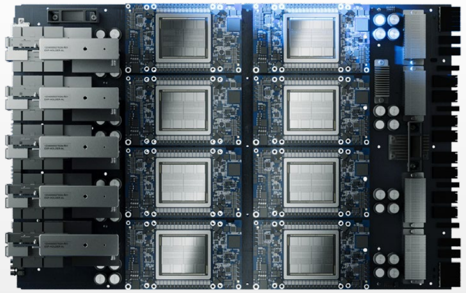
Product Line Details

intel GAUDI Product Line



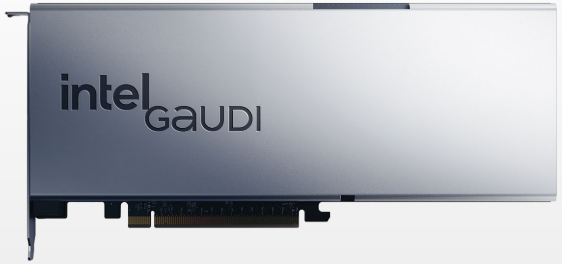
Accelerator Card

HL-325L OAM-Compliant



Universal Baseboard

HLB-325

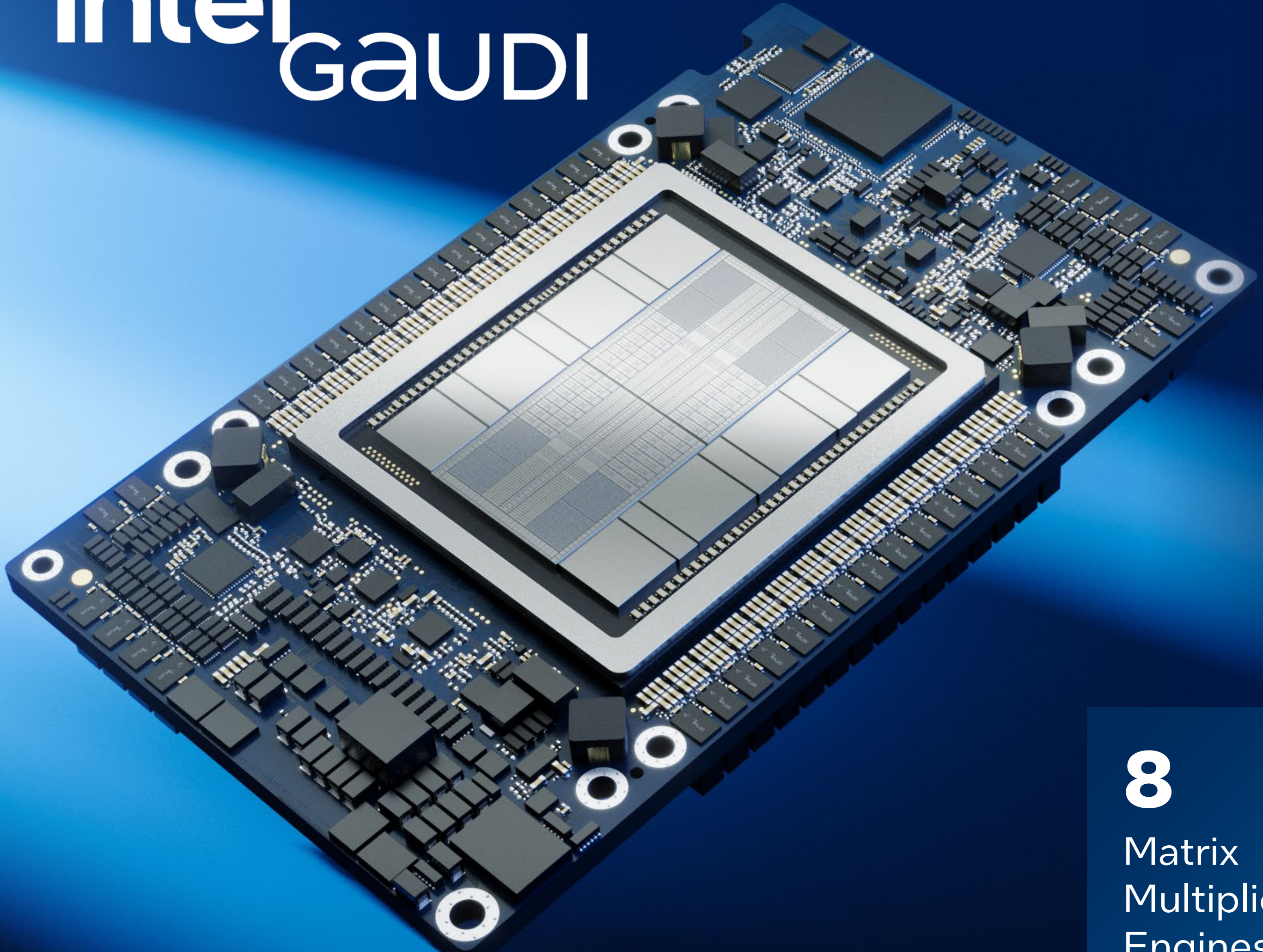


PCIe

HL-338 Add-In Card

Enabling customer infrastructure choice

intel GAUDI



Accelerator Card

HL-325L (OAM-Compliant)

128GB
HBM2e

3.7TB/s
HBM Bandwidth

8
Matrix Multiplication Engines

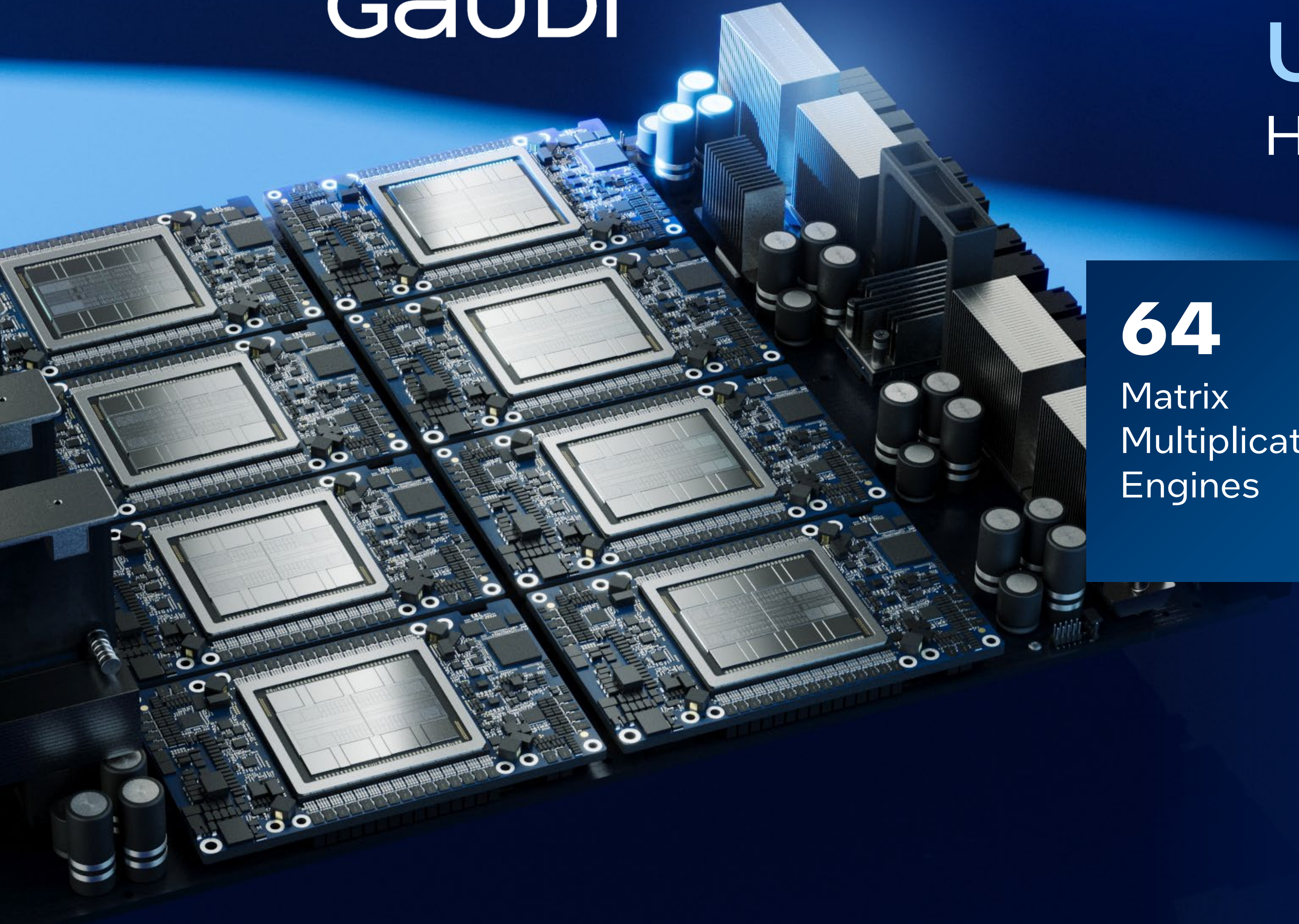
24
200 GbE RDMA NICs

1.2TB/s
Bi-directional Networking

intel GAUDI

Universal Baseboard

HLB-325



64
Matrix Multiplication Engines

192
200 GbE RDMA NICs

9.6TB/s
Bi-directional Networking

Networking Intel® Gaudi® 3 Accelerator reference based-Server

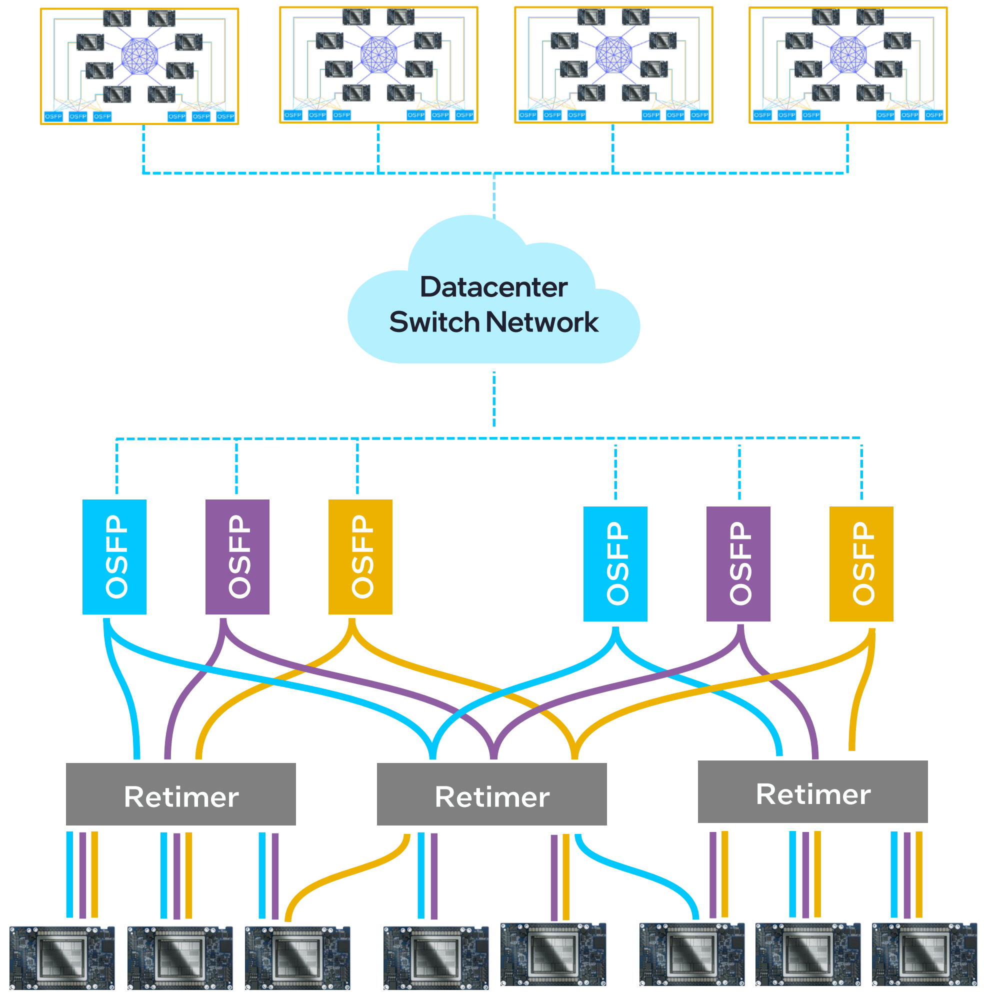
Network ports exposed as NICs to driver

NICs are activated via RDMA verbs over Device Virtual Space

Collective operations execute with low control overhead

Intel® Gaudi® 3 Servers feature:

- 2x Intel® Xeon® Host CPUs
- 8x Intel Gaudi 3 OAM Cards
- Peer-to-peer (P2P) connection between each pair of Intel Gaudi 3 cards
- Intel Gaudi 3 NICs are used both for scale-up and for scale-out
- No need for a network switch inside the node to support scale-up
- Scale-up BW: Total of 8.4TB/s bi-directional
- Scale-out BW: Total of 1.2TB/s bi-directional



Networking Intel Gaudi 3 Servers to enable 32-node cluster

Networking with Intel® Gaudi® 3 AI Accelerator

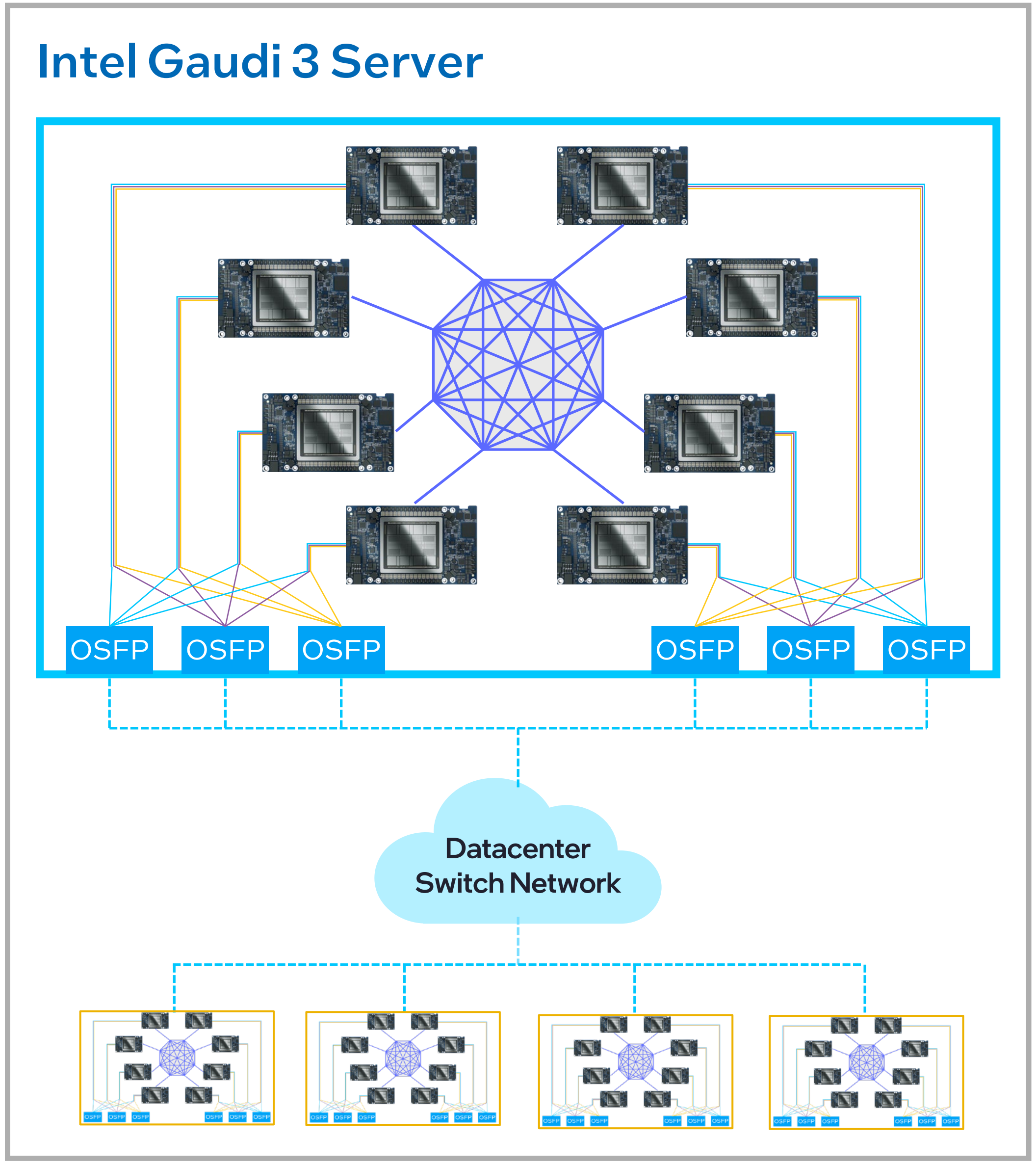
Network ports exposed as NICs to driver

NICs are activated via RDMA verbs over Device Virtual Space

Collective operations execute with low control overhead

Intel® Gaudi® 3 Reference Server

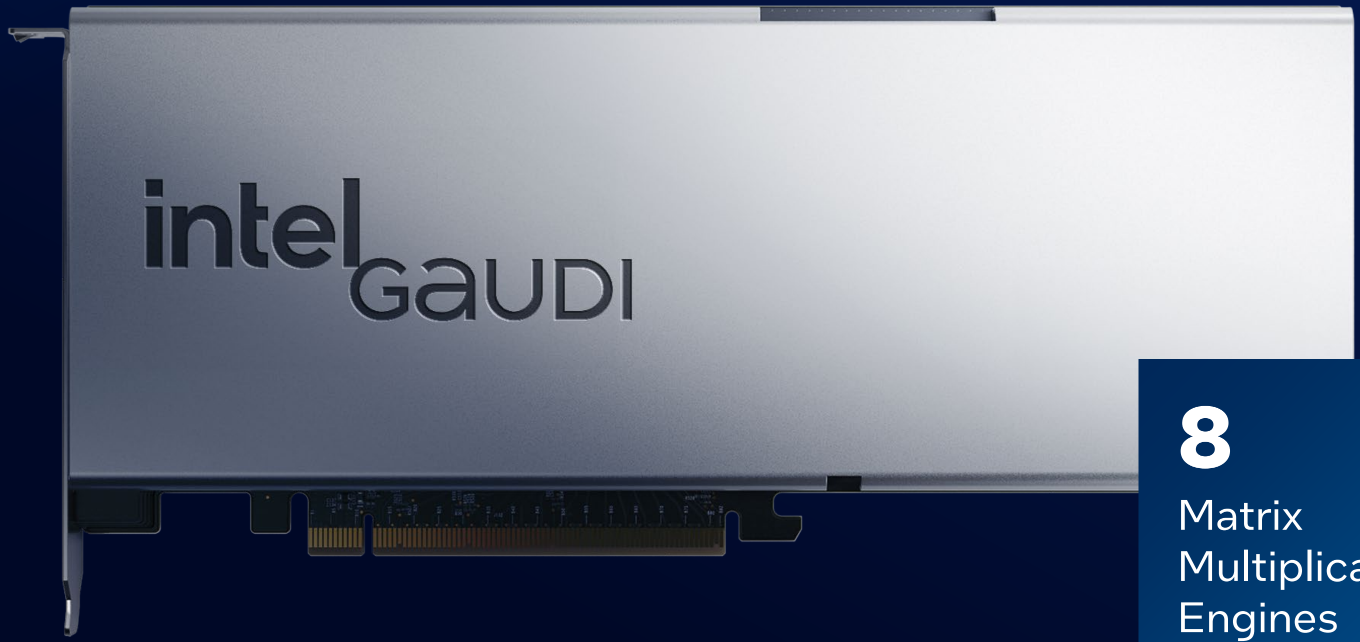
- 2x Intel® Xeon® Host CPUs
- 8x Intel Gaudi 3 Accelerators (OAM Cards)
- Peer-to-peer (P2P) connection between each pair of Intel Gaudi 3 cards
- Intel Gaudi NICs are used both for scale-up and for scale-out
- No need for a network switch inside the node to support scale-up
- Scale-up BW: Total of 8.4Tb/s bi-directional
- Scale-out BW: Total of 9.6Tb/s bi-directional



intel GAUDI

PCIe CEM

HL-338 (Add-In Card)



128GB
HBM2e

600W
TDP

8
Matrix Multiplication Engines

18
200 GbE RDMA NICs

Dual Slot Full Height 10.5" Length

Performance

All Public Performance benchmarks are here……

Developers / Hardware Platforms / Intel® Gaudi® AI Accelerators / **Model Performance Data**

Model Performance Data for Intel® Gaudi® 3 AI Accelerators

intel GAUDI

These performance numbers are measured using the latest SynapseAI* software release version 1.18.0, unless otherwise noted.

Note All models for both training and inference are using the PyTorch* 2.4.0 framework. Other applicable frameworks used for training or inference are noted for each model.

View Intel Gaudi 2 Performance Data →

Feedback

Inference

Large Language Models (LLM) for Throughput with Intel Gaudi 3 Accelerator

Search Table

☑ Model ☑ #HPU ☑ Precision ☑ Input Length ☑ Output Length ☑ Batch Size ☑ Throughput

Model	# HPU	Precision	Input Length	Output Length	Batch Size	Throughput (tokens/sec)
LLaMA 2 7b	1	fp8	128	128	1,536	19,810

https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html

Software Support

Intel® Gaudi® Software Suite

Integrates the main Gen AI frameworks used today

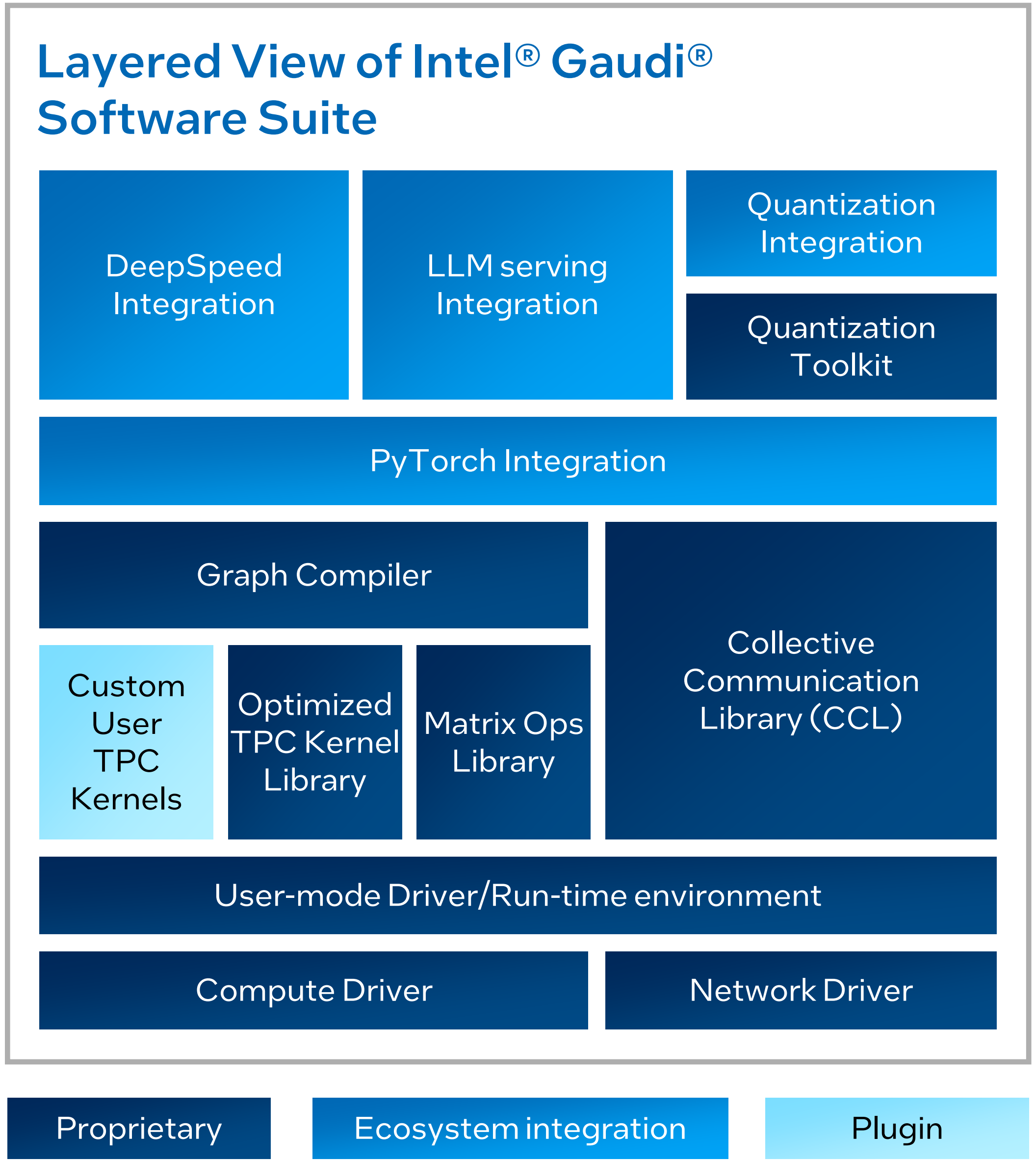
Supports FP16/BF16 → FP8 quantization

Main proprietary SW layers

- Graph Compiler: Handles all engine dependency and scheduling logic
- Matrix operations: Configuring the MME
- TPC kernels: All non-Matrix operations
- Collective Communication Library (CCL)

Several sources for TPC Kernels

- Intel® Gaudi® software optimized TPC kernel library
- Custom user kernels
- MLIR-based fused kernels: generated during graph compilation



intel GAUDI

Extensive Model & Framework Support

500K+ Transformer and Diffusion models on Hugging Face are easily enabled on the Intel® Gaudi® platform

Open Platform for Enterprise AI

Composable Microservices with OPEA Enterprise RAG use case solutions

Customers & Developers can experience the Intel Gaudi platform on

Intel® Tiber™ Developer Cloud

intel GAUDI Software Suite

Enterprise RAG Use Cases
Chat QnA, Code Gen, Code Trans, Doc Summarization, Visual QnA, Audio QnA, FAQ Gen

Representative Models on Intel® Gaudi® Platform

LlaMA 2, 3	DeepSeek distilled	LLaVa	Granite
Stable Diffusion	Mistral	Mixtral	Falcon
MPT	Phi-3	Code Llama	StarCoder
Wave2vec2	Whisper	Qwen2	Bridgetower

Frameworks & Libraries	Orchestration	Tools
PyTorch	kubernetes	vLLM
DeepSpeed	Docker	RAY
Megatron-LM	SLURM	TGI TEI
PyTorch Lightning	Red Hat	LangChain & LlamaIndex
Transformers		Prometheus
Optimum		TensorBoard

New Support

GPU ecosystem has been working on these capabilities for 3+ years. We are quickly catching up and will keep pace with the community going forward.

Intel® Gaudi® Software Stack:

Keeping pace with state-of-the-art model optimization

	Jul 23	Sep 23	Nov 23	Feb 24	Mar 24	Jun 24	Aug 24	Oct 24	Dec 24	Feb 25
Release **Frameworks**	**1.11:** DeepSpeed-Chat support & PEFT LORA	**1.12:** Inference FP8 support	**1.13:** Training FP8 support	**1.14:** Quantization Toolkit DeepSpeed ZeRO ++	**1.15:** FSDP Support TGI-Gaudi support RHEL 9.2	**1.16:** vLLM support Ray.io support Slurm workload manager support	**1.17:** UINT4 support OpenShift support Intel Neural Compressor	**1.18:** FP8 vLLM LoRA vLLM Video Media	**1.19:** Stock PyTorch Support Fused MoE UNIT4 vLLM	**1.20:** vLLM Pipeline Parallelism Dynamic FP8 Quantization
		MLPerf v3.0	**MLPerf** V3.1			**MLPerf** v4.0				
Models		GPT-J 6B inference	GPT3-175B FP8 On 384x Gaudi2 Stable Diffusion On 64x Gaudi2 Plus: LLAMA2 70B training on 256x Gaudi2, BLOOM 176B FP8 Inference On 8x Gaudi2	Training Llama v2 70B on 1024x Gaudi 2s Inference Llama v2 70B FP8 with Flash Attention	Mixtral	LLAMA2 70B Inference on Gaudi2 Stable Diffusion XL Inference on Gaudi2	Inference LLaMA3.1 8B / 70B Gaudi 3	Training LLaMA 3.1 8B/70B on Gaudi 2	Improved performance inc. LLaMA 3.18B/70B Megatron-LM pretraining of LLaMA 3.1 B/70B Mixtral 8x7B	

Less than 6 months to inference on 100B+ param LLM

Less than 12 months to large-scale training on 100B+ param LLM

Easily *Get Started* with PyTorch Models

```
import torch
import torch.nn as nn
import torch.optim as optim
import torch.nn.functional as F
import torchvision
import torchvision.transforms as transforms
import os

# Import Habana Torch Library
import habana_frameworks.torch.core as htcore


# neural network model
class SimpleModel(nn.Module):
...

# training loop
def train(net,criterion,optimizer,trainloader,device):
...
        loss.backward()

        # API call to trigger execution
        htcore.mark_step()

        optimizer.step()

        # API call to trigger execution
        htcore.mark_step()


def main():
...

    # Target the Gaudi HPU device
    device = torch.device("hpu")
```

*Minimal code to start using Intel Gaudi AI Accelerators**

Migrating Python APIs with GPU dependencies

```
import torch
import torch.nn as nn
import torch.optim as optim
import torch.nn.functional as F
import torchvision
import torchvision.transforms as transforms
import os

# Import GPU Migration Package:
import habana_frameworks.torch.gpu_migration

# Import Habana Torch Library
import habana_frameworks.torch.core as htcore


# neural network model
class SimpleModel(nn.Module):
...

# training loop
def train(net,criterion,optimizer,trainloader,device):
...
        loss.backward()

        # API call to trigger execution
        htcore.mark_step()
        optimizer.step()

        # API call to trigger execution
        htcore.mark_step()


def main():
...

    # Target the Gaudi HPU device
    device = torch.device("hpu")
```

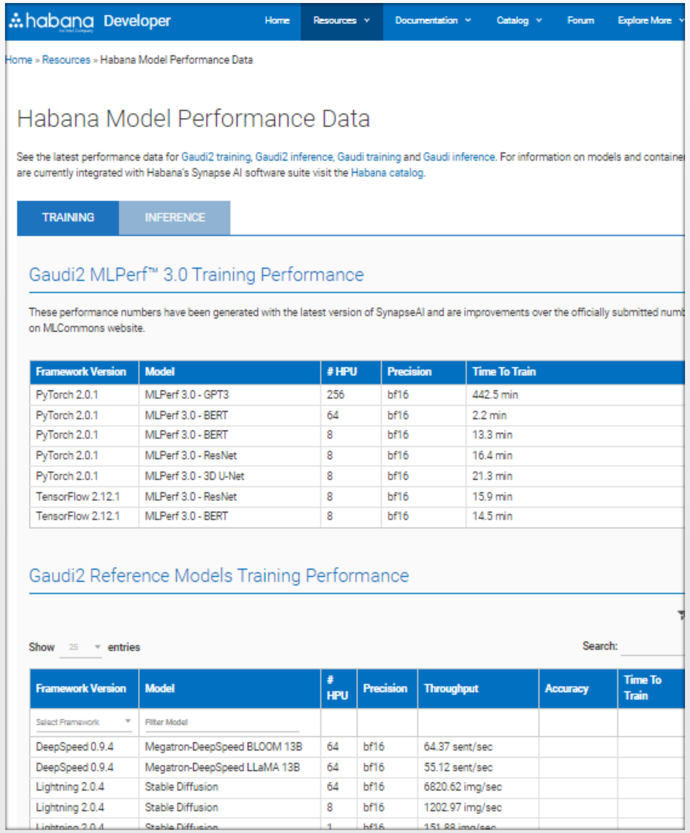
Simplifies replacing Python API calls that have dependencies on GPU libraries with HPU-specific API calls

Specific API calls from following Python libraries are mapped to equivalents in SynapseAI:

- torch.cuda
- torch APIs with GPU related parameters. For example, torch.randn(device="cuda").
- pytorch_lightning
- apex
- pynvml

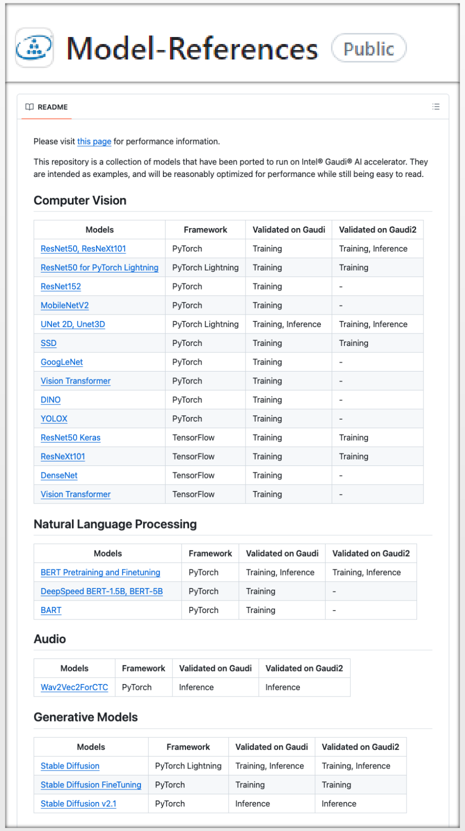
Developer Resources

Intel® Gaudi® AI Accelerator Developer Site

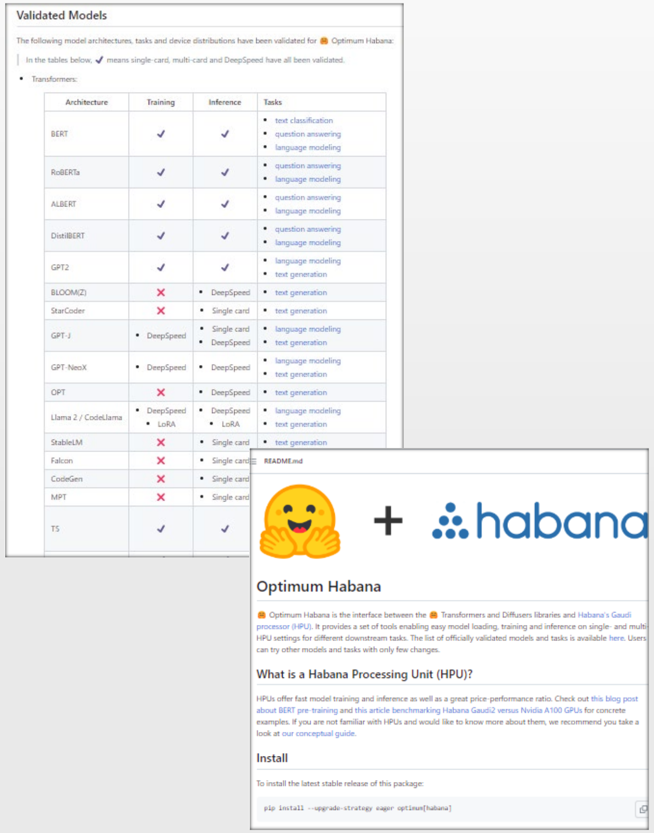


Intel Gaudi AI Accelerator GitHub

https://github.com/HabanaAI



Intel Gaudi AI Accelerator Hugging Face



Intel® Gaudi® AI Accelerator Hugging Face

Transformers & diffusers models

All published Intel Gaudi model architectures, tasks and device distributions have been validated for 🤗 Optimum Habana:

Architecture	Training	Inference	Tasks
BERT	✅	✅	• text classification • question answering • language modeling
RoBERTa	✅	✅	• question answering • language modeling
ALBERT	✅	✅	• question answering • language modeling
DistilBERT	✅	✅	• question answering • language modeling
GPT2	✅	✅	• language modeling • text generation
BLOOM(Z)		• DeepSpeed	• text generation
StarCoder		• Single card	• text generation
GPT-J	• DeepSpeed	• Single card	• language modeling
GPT-NeoX	• DeepSpeed	• DeepSpeed	• language modeling • text generation
OPT		• DeepSpeed	• text generation
Llama 2 / CodeLlama	• DeepSpeed • LoRA	✅	• language modeling • text generation
StableLM		• Single card	• text generation
Falcon	• LoRA	✅	• text generation
CodeGen		• Single card	• text generation
MPT		• Single card	• text generation
Mistral		• Single card	• text generation
T5	✅	✅	• summarization • translation • question answering
BART		• Single card	• summarization • translation • question answering
ViT	✅	✅	• image classification
Swin	✅	✅	• image classification
Wav2Vec2	✅	✅	• audio classification • speech recognition
CLIP	✅	✅	• contrastive image-text training
BridgeTower	✅	✅	• contrastive image-text training
ESMFold		• Single card	• protein folding

- Diffusers

Architecture	Training	Inference	<center>Tasks</center>
Stable Diffusion		• Single card	• text-to-image generation
LDM3D		• Single card	• text-to-image generation

In the tables above, ✅ means single-card, multi-card, and DeepSpeed have all been validated

https://huggingface.co/docs/optimum/habana/index

Performance Transparency on Every Model

Gaudi2 MLPerf™ 3.1 Training Performance

These performance numbers have been generated with the latest version of SynapseAI and are improvements over the officially submitted numbers posted on MLCommons website.

Model	# HPU	Precision	Time To Train	Framework Version
MLPerf 3.1 - GPT3	384	fp8	153.58 min*	
MLPerf 3.1 - GPT3	256	fp8	223.75 min**	
MLPerf 3.1 - Stable Diffusion v2	64	bf16	19.4 min**	PyTorch Lightning 2.1.2
MLPerf 3.1 - ResNet	8	bf16	16.22 min	
MLPerf 3.1 - BERT	8	bf16	14.25 min	

* The GPT3 measurement with 384 cards was taken using a pre-launch version of the SynapseAI 1.13.0 Software stack
** The GPT measurement with 256 cards and Stable Diffusion were taken using the SynapseAI 1.13.0 Software stack

Gaudi2 Large Language Models Training Performance

Model	# HPU	Precision	Throughput	Sequence Length	TP, PP, DP	Batch Size	Framework Version
LLaMA 13B	64	bf16	68.12 samples/sec	2,048	2, 2, 16	256	DeepSpeed 0.12.4
LLaMA 2 70B	256	bf16	30 samples/sec	4,096	8, 8, 4	1,024	DeepSpeed 0.12.4
LLaMA 2 70B	512	bf16	55.4 samples/sec	4,096	8, 8, 8	2,048	DeepSpeed 0.12.4
LLaMA 2 70B	1,024	bf16	104.4 samples/sec	4,096	8, 8, 16	4,096	DeepSpeed 0.12.4
Bloom-13B	64	bf16	72.5 samples/sec	2,048	2, 2, 16	1,024	DeepSpeed 0.12.4

TP, PP, DP = These are the Tensor Parallel, Pipeline Parallel and Data Parallel parameters for the Megatron DeepSpeed training

Hugging Face Optimum Habana Gaudi2 Inference Performance

See the Examples page for information on how to run each of the Tasks, including model naming and hyperparameter usage.

Show 25 entries

Model	# HPU	Precision	Max Token Sequence Length	Throughput	Latency	Batch	task	Framework Version
StableDiffusion v2.1 (512x512)	1	bf16		1.24 images/sec	3223.2 ms	4	stable-diffusion	PyTorch Lightning 2.1.2
OPT	1	bf16		980.99 token/sec	1.01 ms	1	text-generation	DeepSpeed 0.12.4, Optimum Habana 1.9.0
StarCoder	1	bf16		65.5 token/sec	15.26 ms	1	text-generation	DeepSpeed 0.12.4, Optimum Habana 1.9.0
MPT-7B	1	bf16	1932	105.38 token/sec	9.48 ms	1	text-generation	Optimum Habana 1.9.0
Bert (Text Classification)	1	bf16		186.26 token/sec	42.94 ms	8	text-classification	Optimum Habana 1.9.0
Bert (Language Modeling)	1	bf16		80.72 token/sec	49.55 ms	4	language-modeling	Optimum Habana 1.9.0
Bert (Question Answering)	1	bf16		599.54 token/sec	13.34 ms	8	question-answering	Optimum Habana 1.9.0
Bart	1	bf16		6.5 token/sec	614.91 ms	4	language-modeling	Optimum Habana 1.9.0
BridgeTower	1	bf16		329.65 token/sec	48.53 ms	16	constrastive-image-text	Optimum Habana 1.9.0
ESMFold	1	bf16		3.67 token/sec	272.47 ms	1	protein-folding	Optimum Habana 1.9.0
StableLM-3B	1	bf16	2048	232.34 token/sec	4.3 ms	1	text-generation	Optimum Habana 1.9.0
StableLM-7B	1	bf16	2048	123.02 token/sec	8.12 ms	1	text-generation	Optimum Habana 1.9.0
T5-3B Summarization 1024-128 Beam4	1	bf16		0.96 token/sec	1035.19 ms	1	summarization	Optimum Habana 1.9.0
T5-3B Summarization Greedy	1	bf16		2.37 token/sec	420.87 ms	1	summarization	Optimum Habana 1.9.0
HF-T5-Small-Translation-Greedy	1	bf16		1031.6 token/sec	3.87 ms	4	translation	Optimum Habana 1.9.0
Wav2vec(Audio Classification)	1	bf16		15.26 token/sec	262.12 ms	4	audio-classification	Optimum Habana 1.9.0
Wav2vec(Speech	1	bf16		23.14 token/sec	172.83 ms	4	speech-	Optimum Habana

Gaudi2 Large Languages Models Inference Performance

Model	# HPU	Precision	Input length	Output Length	Max Token Sequence Length	Throughput	Latency***	Batch	Framework Version
Falcon-7B	1	bf16	100	8k	8k	110.7 token/sec	9.03 ms	1	Optimum Habana 1.9.0
Bloom-7B-Greedy	1	bf16			2k	721.56 token/sec	11.08 ms	8	
Bloom-7B-Greedy	1	fp8			2K	194.12 token/sec	5.15 ms	1	
GPT-J	8	bf16	6	100	100	562.23 token/sec	7.11 ms	4	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-7B	1	fp8	1K	3k	4k	1101.7 token/sec	10.89 ms	12	Optimum Habana 1.9.0
LLaMA 2-7B	1	fp8	2k	6k	8k	551.84 token/sec	10.87 ms	6	Optimum Habana 1.9.0
LLaMA 2-7B	1	fp8	4k	12k	16k	273.32 token/sec	10.97 ms	3	Optimum Habana 1.9.0
LLaMA 2-7B	1	bf16	1k	3k	4k	361.14 token/sec	11.07 ms	4	Optimum Habana 1.9.0
Falcon-40B	8	bf16	100	8k	8k	61.85 token/sec	16.16 ms	1	Optimum Habana 1.9.0
LLaMA 2-70B	8	fp8	2k	2k	4k	4910.4 token/sec	56.41 ms	277	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-70B	8	fp8	2k	6k	8k	2859.5 token/sec	26.92 ms	77	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-70B	8	fp8	2k	14k	16k	1470.6 token/sec	25.83 ms	38	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-70B	8	fp8	2k	30k	32k	775.4 token/sec	24.5 ms	19	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-70B	8	bf16	2k	2k	4k	3225.6 token/sec	66.96 ms	216	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-70B	8	bf16	2k	6k	8k	1229.1 token/sec	24.4 ms	30	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-70B	8	bf16	2k	14k	16k	566.9 token/sec	26.45 ms	15	DeepSpeed 0.12.4, Optimum Habana 1.9.0
LLaMA 2-13B	1	bf16	2k	2k	4k	125.66 token/sec	15.91 ms	2	Optimum Habana 1.9.0
Bloomz-176B	8	bf16	6	100	100	36.36 token/sec	27.5 ms	1	DeepSpeed 0.12.4, Optimum Habana 1.9.0
Bloom-176B-Greedy	8	fp8			4K	199.39 token/sec	40.12 ms	8	DeepSpeed 0.12.4
Bloom-176B-Greedy	8	bf16			4K	394.47 token/sec	53.23 ms	21	DeepSpeed 0.12.4
Bloom-176B-Greedy	8	bf16			8K	196.21 token/sec	50.96 ms	10	DeepSpeed 0.12.4
Bloom-176B-Greedy	8	bf16			16K	80.79 token/sec	49.51 ms	4	DeepSpeed 0.12.4
Bloom-176B-Greedy	8	bf16			32K	25.81 token/sec	38.74 ms	1	DeepSpeed 0.12.4
Bloom-176B-Sampling	8	bf16			1k	19.59 token/sec	51.04 ms	1	DeepSpeed 0.12.4
Bloom-176B-BeamSearch-8	8	bf16			512	30.57 token/sec	32.7 ms	1	DeepSpeed 0.12.4

Hugging Face Optimum Habana Gaudi2 Training Performance

See the Examples page for information on how to run each of the Tasks, including model naming and hyperparameter usage.

Show 25 entries Search:

Model	# HPU	Precision	Throughput	Accuracy	Time To Train	Batch Size	Task	Framework Version
Llama2-70B Fine Tuning (LoRA)	8	bf16	2.43 sentences/sec	2.12	43.21 min	10	language-modeling	Optimum Habana 1.9.0
Llama1-7B Fine Tuning (LoRA)	8	bf16	143.02 sentences/sec	0.93	5.5 min	64	language-modeling	Optimum Habana 1.9.0
Falcon-180B Fine Tuning (LoRA)	8	bf16	1.53 sentences/sec	1.05	270.13 min	1	language-modeling	Optimum Habana 1.9.0
Falcon-40B Fine Tuning (LoRA)	8	bf16	28.83 sentences/sec	1.19	16.8 min	1	language-modeling	Optimum Habana 1.9.0
GPTJ-CLM	8	bf16	16.64 sentences/sec	0.53	14.13 min	4	language-modeling	Optimum Habana 1.9.0
GPTNEOX-20B-CLM	8	bf16	290.8 sentences/sec	0.17	28.75 min	2	language-modeling	Optimum Habana 1.9.0
BridgeTower	8	bf16	473.05 sentences/sec		7.4 min	40	contrastive-image-text	Optimum Habana 1.9.0
GPT2	8	bf16	575.41 sentences/sec			4	language-modeling	Optimum Habana 1.9.0
GPT2-XL	8	bf16	86.59 sentences/sec			4	language-modeling	Optimum Habana 1.9.0
ALBERT-Large	8	bf16	2280.45 sentences/sec	91.84	2.08 min	32	question-answering	Optimum Habana 1.9.0
ALBERT-XXL	8	bf16	446.96 sentences/sec	94.89	7.01 min	12	question-answering	Optimum Habana 1.9.0
BERT Base	8	bf16	3103 sentences/sec	85.42	1.2 min	24	question-answering	Optimum Habana 1.9.0
BERT-Large Fine Tuning	8	bf16	2256.81 sentences/sec	93.29	1.91 min	24	question-answering	Optimum Habana 1.9.0
ClipRoBERTa	8	bf16	1160.11 images/sec		27.86 min	64	contrastive-image-text	Optimum Habana 1.9.0
DistilBERT	8	bf16	9985.41 sentences/sec	82.5	0.56 min	8	question-answering	Optimum Habana 1.9.0
Flan-T5 XXL	8	bf16	28.59 sentences/sec	36.64	387.47 min	22	question-answering	Optimum Habana 1.9.0
RoBERTa Base	8	bf16	6563.31 sentences/sec	92.15	0.75 min	12	question-answering	Optimum Habana 1.9.0
RoBERTa Large	8	bf16	2234.79 sentences/sec	94.52	1.93 min	12	question-answering	Optimum Habana 1.9.0
Swin Transformer	8	bf16	5753.05 images/sec			64	image-classification	Optimum Habana 1.9.0

Gaudi2 Reference Models Training Performance

Show 50 entries Search:

Model	# HPU	Precision	Throughput	Accuracy	Time To Train	Batch Size	Framework Version
DeepSpeed Chat LLaMA 7B Step1	8	bf16	870 sec/iter	ppl: 1.61		8	Megatron DeepSpeed 0.12.4
DeepSpeed Chat LLaMA 7B Step2	8	bf16	770 sec/iter	acc: 81		4	Megatron DeepSpeed 0.12.4
DeepSpeed Chat LLaMA 7B Step3	8	bf16	7.8 sec/iter	ema: 2.7		4	Megatron DeepSpeed 0.12.4
Stable Diffusion	64	bf16	10658.26 img/sec			32	Lightning 2.1.2
Stable Diffusion Fine Tuning	1	bf16	70 img/sec			7	Lightning 2.1.2
Stable Diffusion Fine Tuning Textual Inversion	1	bf16	20.58 img/sec			7	Lightning 2.1.2
ResNet50 LARS	32	bf16	185594.14 img/sec	76.33	6.62 min	256	
ResNet50 LARS	8	bf16	47104.44 img/sec	76.24	17.53 min	256	
ResNet50 LARS	1	bf16	6072.03 img/sec			256	
BERT Pre Training Phase 1	32	bf16	32249.28 sent/sec	1.497506499	273.25 min	64	
BERT Pre Training Phase 1	8	bf16	1151.95 sent/sec			64	
BERT Pre Training Phase 1	1	bf16	9126.16 sent/sec			64	
BERT Pre Training Phase 2	32	bf16	10812.96 sent/sec	1.332446098	91.21 min	16	
BERT Pre Training Phase 2	8	bf16	348.71 sent/sec			16	
BERT Pre Training Phase 2	1	bf16	2779.63 sent/sec			16	
BERT SQUAD Fine Tuning	8	bf16	2054.6 sent/sec	90.83	5.12 min	24	
BERT SQUAD Fine Tuning	1	bf16	281.49 sent/sec			24	
ResNext101	8	bf16	22146.5 img/sec	78.03	102 min	256	
ResNext101	1	bf16	2841.31 img/sec			256	
SSD	8	bf16	16555.1 img/sec	22.95	9.58 min	128	
SSD	1	bf16	2098.21 img/sec			128	
Transformer	8	bf16	1074849 token/sec	27.9	242.05 min	8192	
Transformer	1	bf16	133257.66 token/sec			8192	
Unet2D	8	bf16	16999.22 img/sec	72.61	13.92 min	64	Lightning 2.1.2
Unet2D	1	bf16	2697.3 img/sec			64	Lightning 2.1.2
Unet3D	8	bf16	256.76 img/sec	74.26	19.77 min	2	Lightning 2.1.2
Unet3D	1	bf16	32.69 img/sec			2	Lightning 2.1.2

Gaudi Reference Models Training Performance

Show 25 entries Search:

Model	# HPU	Precision	Throughput	Accuracy	Time To Train	Batch Size	Framework Version
ResNet50 Keras LARS	32	bf16	48176.47 img/sec	75.86	19.66 min	256	
ResNet50 Keras LARS	8	bf16	12305.57 img/sec	76.16	69.86 min	256	
ResNet50 Keras LARS	1	bf16	1624.15 img/sec			256	
BERT Pre Training combine	32	bf16	4805.33 sent/sec		1803.26 min	64	
BERT Pre Training combine	8	bf16	1221.73 sent/sec			64	
BERT Pre Training combine	1	bf16	153.34 sent/sec			64	
BERT Pre Training Phase 1	32	bf16	5757.69 sent/sec	Loss: 1.49	1348.41 min	64	
BERT Pre Training Phase 1	8	bf16	1467.43 sent/sec			64	
BERT Pre Training Phase 1	1	bf16	184.18 sent/sec			64	
BERT Pre Training Phase 2	32	bf16	1911.84 sent/sec	Loss: 1.33	454.85 min	8	
BERT Pre Training Phase 2	8	bf16	482.51 sent/sec			8	
BERT Pre Training Phase 2	1	bf16	60.56 sent/sec			8	
BERT SQUAD Fine Tuning	8	bf16	404.66 sent/sec	90.68	13.08 min	24	
BERT SQUAD Fine Tuning	1	bf16	53 sent/sec			24	
BART Fine Tuning	8	bf16	1763.9 sent/sec			32	
DINO	8	bf16	937.41 exmpl/sec	77	2280.8 min	64	
MobileNetV2	8	bf16	12049 img/sec	71.21	531.06 min	256	
ResNet152	8	bf16	4985.28 img/sec	78.56	435.41 min	128	
SSD**	8	bf16	3557.6 images/sec			128	
Transformer	8	bf16	186126.33 tokens/sec	28.2	1035.21 min	4096	
Unet2D	8	bf16	5133.72 img/sec	72.6	58.9 min	64	Lightning 2.1.2
Unet3D	8	bf16	61.53 img/sec	74.21	76.5 min	2	Lightning 2.1.2
YOLOX	8	bf16	380.08 img/sec	39.65	2104.73 min	16	
ResNet50 Host NIC (libfabric)	16	bf16	22542.08 img/sec			256	

Gaudi Reference Models Inference Performance

Model	# HPU	Precision	Throughput	Latency***	Batch Size	Framework Version
Bloom-176B-BeamSearch-8	16	bf16	10.51 token/sec	95.1 ms	1	DeepSpeed 0.12.4
Bloom-176B-Greedy	16	bf16	11.92 token/sec	83.38 ms	1	DeepSpeed 0.12.4
Bloom-176B-Sampling	16	bf16	7.98 token/sec	125.26 ms	1	DeepSpeed 0.12.4
Bloom-7B (512 token)	1	bf16	42.84 token/sec	23.34 ms	1	
Stable Diffusion v2.1 (512x512)	1	bf16	0.36 img/sec	2777.77 ms	1	Lightning 2.1.2
Stable Diffusion v2.1 (768X768)	1	bf16	0.13 img/sec	7692.3 ms	1	Lightning 2.1.2
Bert	1	bf16	147.17 token/sec	163.12 ms	24	
Unet2D	1	bf16	1364.2 img/sec	46.9 ms	64	Lightning 2.1.2
Unet3D	1	bf16	52.68 img/sec	37.96 ms	2	Lightning 2.1.2

Hugging Face Optimum Habana Gaudi Inference Performance

See the Examples page for information on how to run each of the Tasks, including model naming and hyperparameter usage.

Show 25 entries

Model	# HPU	Precision	Throughput	Latency	Batch	Task	Framework Version
BERT	1	bf16	39.55 token/sec	101.12 ms	4	language-modeling	Optimum Habana 1.9.0
BERT	1	bf16	126.85 token/sec	63.06 ms	8	question-answering	Optimum Habana 1.9.0
BERT	1	bf16	107.76 token/sec	74.23 ms	8	text-classification	Optimum Habana 1.9.0
BART-Greedy	1	bf16	2.96 token/sec	675.67 ms	2	summarization	Optimum Habana 1.9.0
ESMFold	1	bf16	14.17 token/sec	70.54 ms	1	protein-folding	Optimum Habana 1.9.0
Stable Diffusion v2.1 image size 512x512	1	bf16	0.35 token/sec	11173.18 ms	4	text to image generation	Optimum Habana 1.9.0
T5-Small Translation Greedy	1	bf16	15.48 token/sec	258.36 ms	4	translation	Optimum Habana 1.9.0
Wav2Vec 2.0 ASR	1	bf16	529.7 token/sec	7.68 ms	4	speech-recognition	Optimum Habana 1.9.0
Wav2Vec 2.0 Speech Classification	1	bf16	9.39 token/sec	425.62 ms	4	speech-recognition	Optimum Habana 1.9.0

https://developer.habana.ai/resources/habana-models-performance/

Availability and Customer Momentum

Announcing general availability

DELL Technologies

Dell PowerEdge XE9680

Air-cooled

Dell AI Factory

Shipping Q225

SUPERMICRO

Supermicro X14

Air-cooled

Equipped with Intel® Xeon® 6 processors

Shipping Q125

intel GAUDI

Available today on the Intel® Tiber™ AI Cloud

Speed up your development for Gen AI
with Intel® Gaudi® 2 accelerators, already available in the cloud

Build your software on Intel Gaudi 2 accelerators and migrate your code seamlessly to Intel Gaudi 3 accelerator

Select availability of Intel Gaudi 3 accelerators
on Intel® Tiber™ AI Cloud

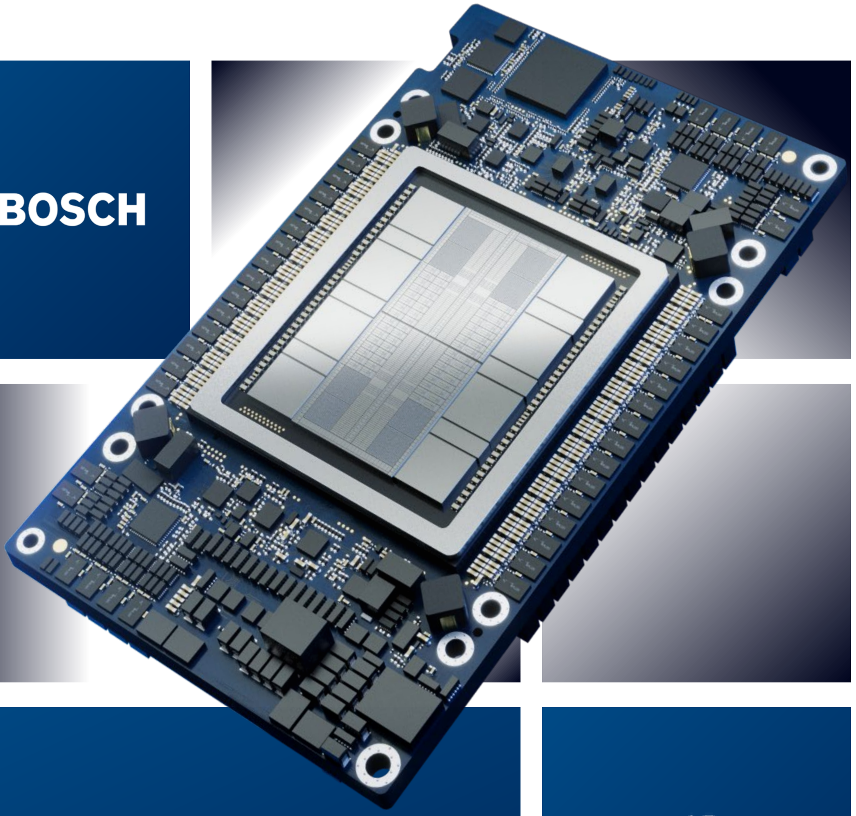
Visit Intel Tiber AI Cloud

intel GAUDI Growing Customer Momentum

AI SWEDEN

Articul8

BOSCH



CtrlS

DENVR dataworks

IBM

iff

LandingAI

LUMEN

NAVER

NIQ

predictionGuard

40 Advent International EST. 1984

salesforce

seekr

Intel® Gaudi® 3 on IBM Cloud

Flexible consumption & user experience

VPC Virtual Servers

Red Hat Enterprise Linux AI servers
or
Accelerated Gaudi 3 virtual servers for non-RHEL AI workloads

ROKS & IKS Clusters

OpenShift AI clusters
or
IKS or OpenShift Clusters w/ Gaudi 3 accelerated workers

Deployable Architectures

Production ready, pre-configured RAG solution

watsonx

As SaaS with no exposure to underlying infra
or
As Software in private datacenter

IBM Cloud Data Center Locations for Gaudi 3

Dallas (DAL)

Frankfurt (FRA)

Washington D.C.(WDC)

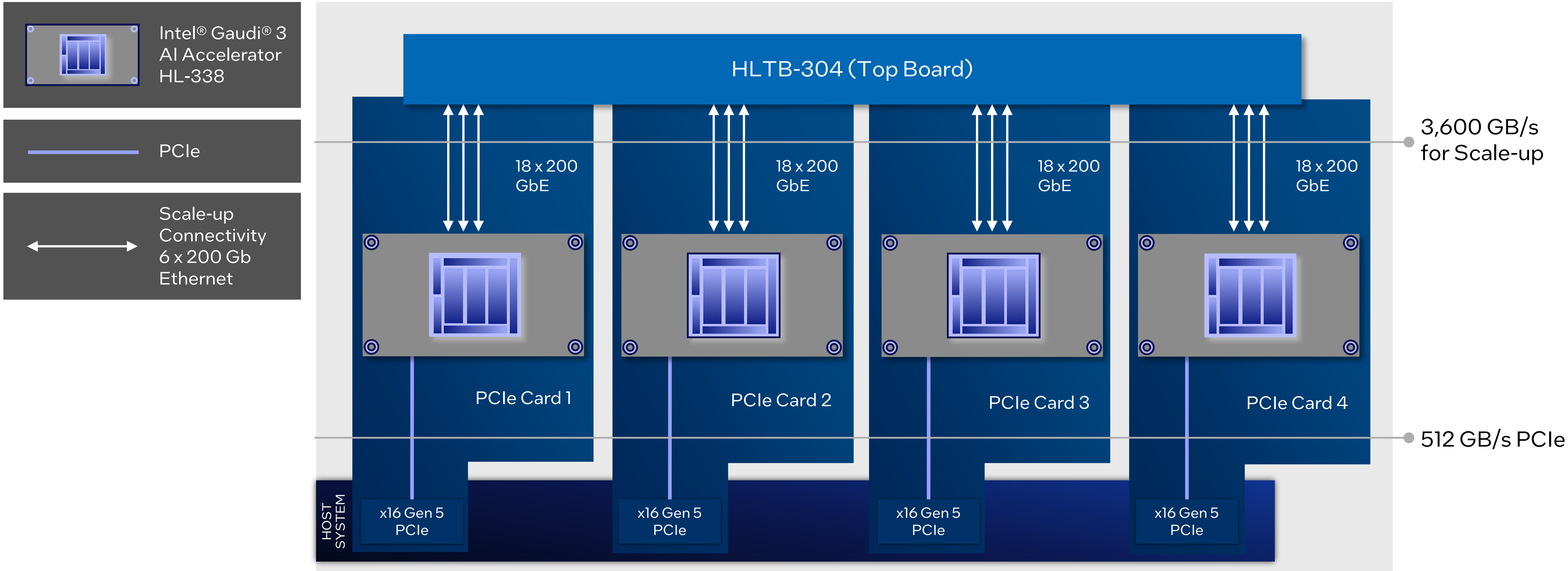
Select availability in US/EMEA early 2025

Regional expansion plans TBD

Open and Efficient Scalability

intel GAUDI PCIe (HL-338) Block Diagram

4xPCIe cards per system



Ideal for inferencing, fine-tuning & small model training

Intel® Gaudi® 3 AI Accelerator Scalable Systems

Reference architectures

1 Node
(x8 Accelerators)

FP8 Compute	13.4 PF
Memory Capacity	1024 GB
Networking B/W	9.6 TB/s

32 Node
(x256 Accelerators)

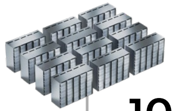
FP8 Compute	428.8 PF
Memory Capacity	32.7 TB
Networking B/W	307 TB/s

64-node Cluster
(x512 Accelerators)

FP8 Compute	857.6 PF
Memory Capacity	65.5 TB
Networking B/W	614 TB/s

512-node Cluster
(x4096 Accelerators)

FP8 Compute	6.8 EF
Memory Capacity	524 TB
Networking B/W	4912 TB/s



1024-node Cluster
(x8192 Accelerators)

FP8 Compute	13.7 EF
Memory Capacity	1 PB
Networking B/W	9.8 PB/s

Performance

Inference - Lower latency, higher throughput
Training - Faster Time to train, larger model sizes

*Visuals for illustrative purposes, not actual systems
Peak projected performance, memory capacity & B/W, networking scale-up/scale-out B/W Performance varies by use, configuration and other factors. Results may vary

Intel® Gaudi® 3 Accelerator based- System Scaling Example

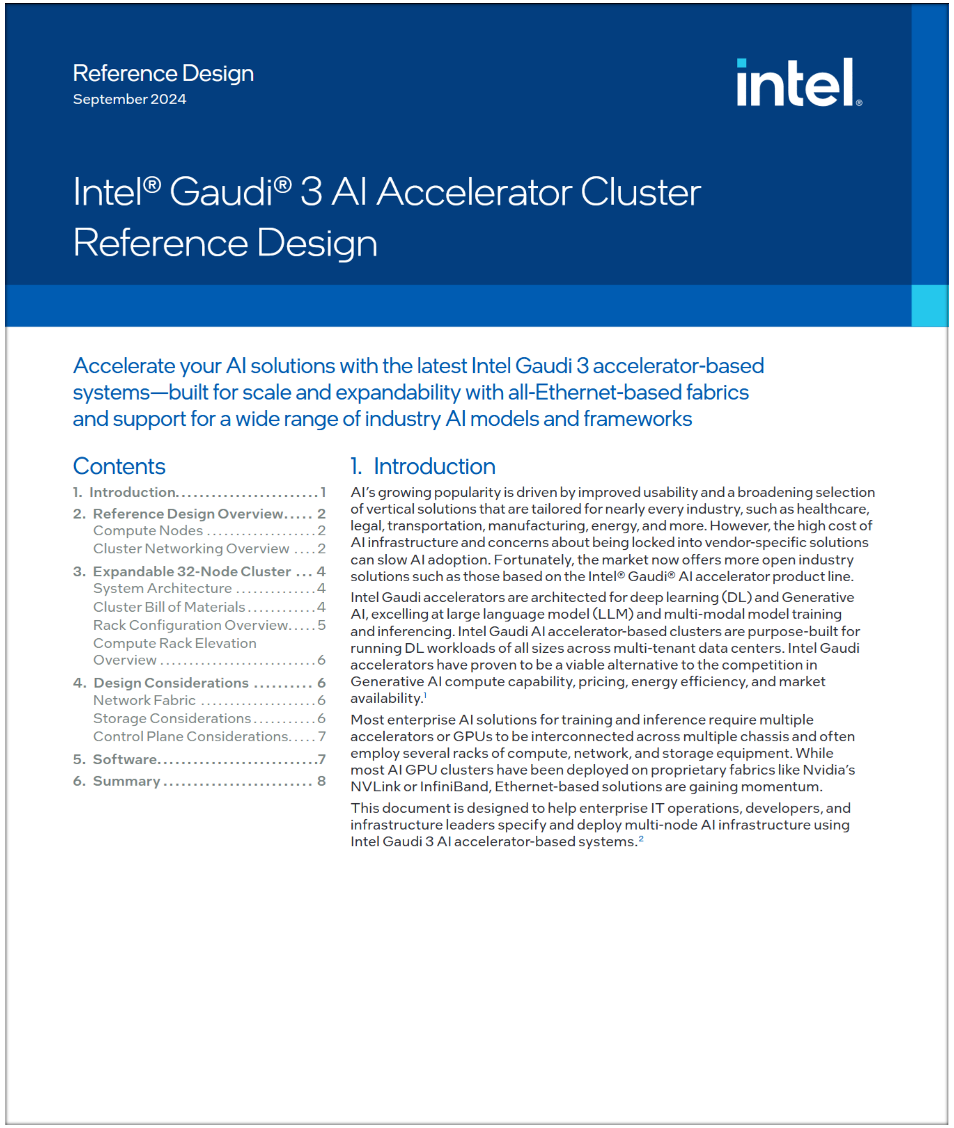
Reference design for 32 Node / 256 Intel® Gaudi® 3 AI accelerator-based cluster

Scalable architecture supports 1024 Nodes & beyond

Compute block: 8 x 8 Intel Gaudi 3 Accelerator based-Nodes on 3-ply Ethernet Fabric

Ethernet switching including Arista

Storage system including Weka



Reference Design

September 2024

intel.

Intel® Gaudi® 3 AI Accelerator Cluster Reference Design

Accelerate your AI solutions with the latest Intel Gaudi 3 accelerator-based systems—built for scale and expandability with all-Ethernet-based fabrics and support for a wide range of industry AI models and frameworks

Contents

1. Introduction

AI's growing popularity is driven by improved usability and a broadening selection of vertical solutions that are tailored for nearly every industry, such as healthcare, legal, transportation, manufacturing, energy, and more. However, the high cost of AI infrastructure and concerns about being locked into vendor-specific solutions can slow AI adoption. Fortunately, the market now offers more open industry solutions such as those based on the Intel® Gaudi® AI accelerator product line.

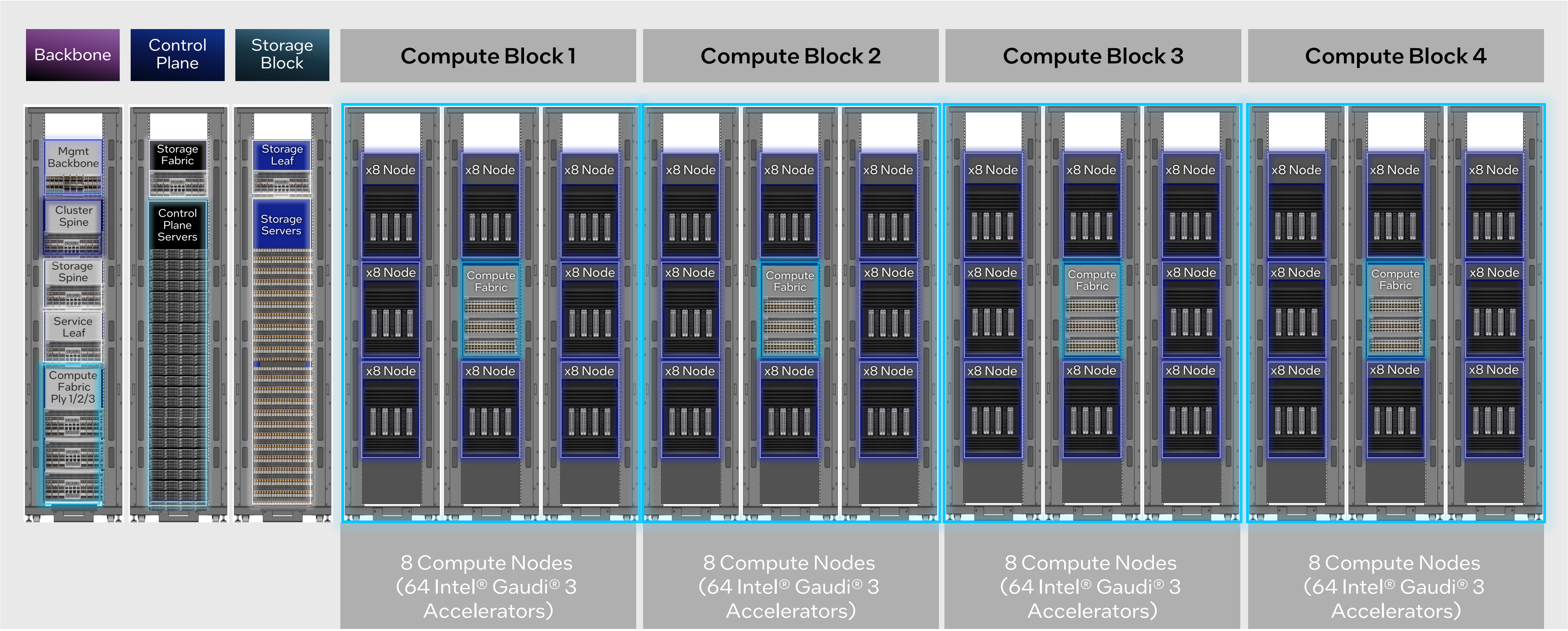
Intel Gaudi accelerators are architected for deep learning (DL) and Generative AI, excelling at large language model (LLM) and multi-modal model training and inferencing. Intel Gaudi AI accelerator-based clusters are purpose-built for running DL workloads of all sizes across multi-tenant data centers. Intel Gaudi accelerators have proven to be a viable alternative to the competition in Generative AI compute capability, pricing, energy efficiency, and market availability.[1]

Most enterprise AI solutions for training and inference require multiple accelerators or GPUs to be interconnected across multiple chassis and often employ several racks of compute, network, and storage equipment. While most AI GPU clusters have been deployed on proprietary fabrics like Nvidia's NVLink or InfiniBand, Ethernet-based solutions are gaining momentum.

This document is designed to help enterprise IT operations, developers, and infrastructure leaders specify and deploy multi-node AI infrastructure using Intel Gaudi 3 AI accelerator-based systems.[2]

Intel® Gaudi® 3 Accelerator Cluster:

32 Node scalable configuration



Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Intel technologies may require enabled hardware, software or service activation.

Availability of accelerators varies depending on SKU. Please contact your Intel sales representative for more information.

Your costs and results may vary.