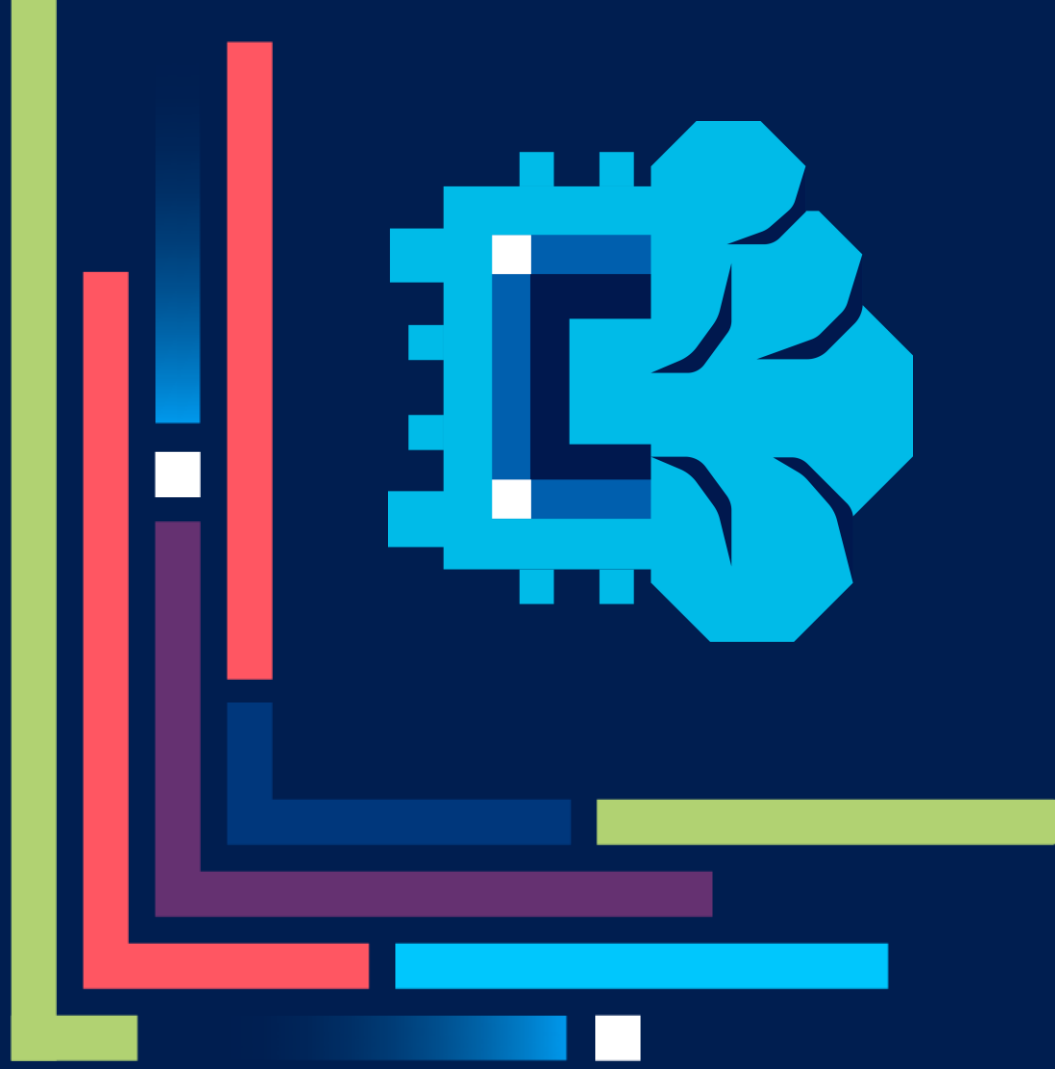
intel ai summit
英特爾AI科技論壇

Bringing AI Everywhere



Accelerating GenAI & LLM with Intel Gaudi Solutions

Leo Zhao

AI Software Evangelist

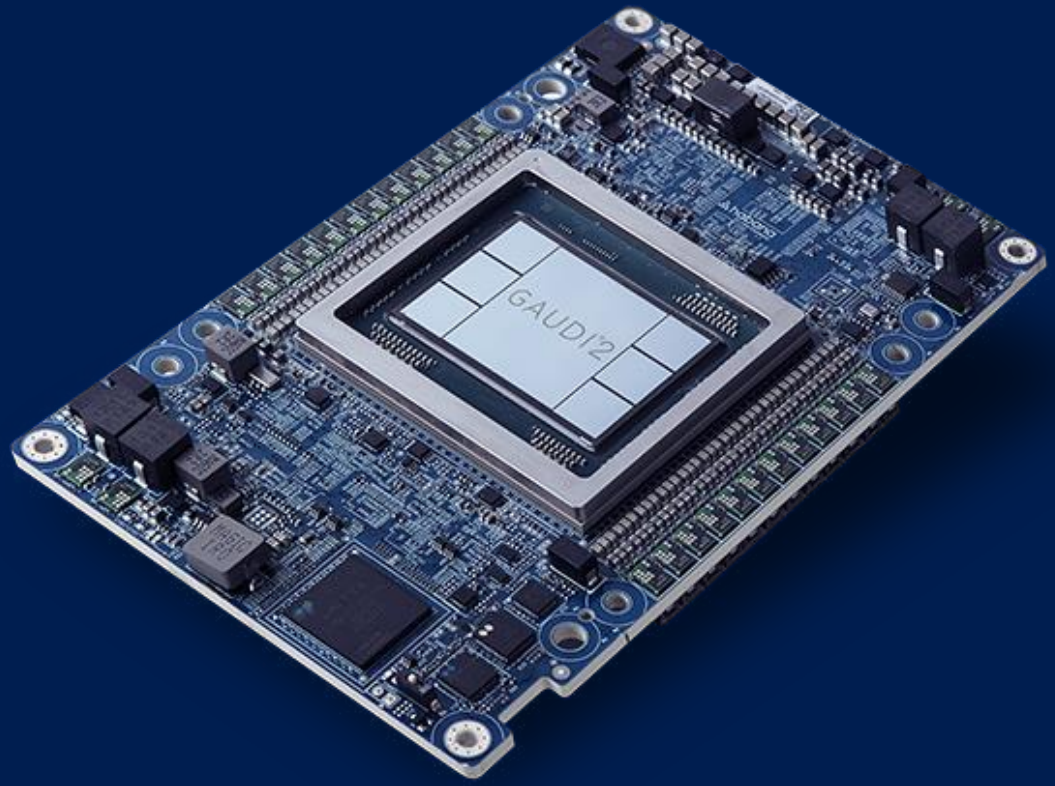
March 27th, 2024

Brief of Intel® Gaudi® Solution

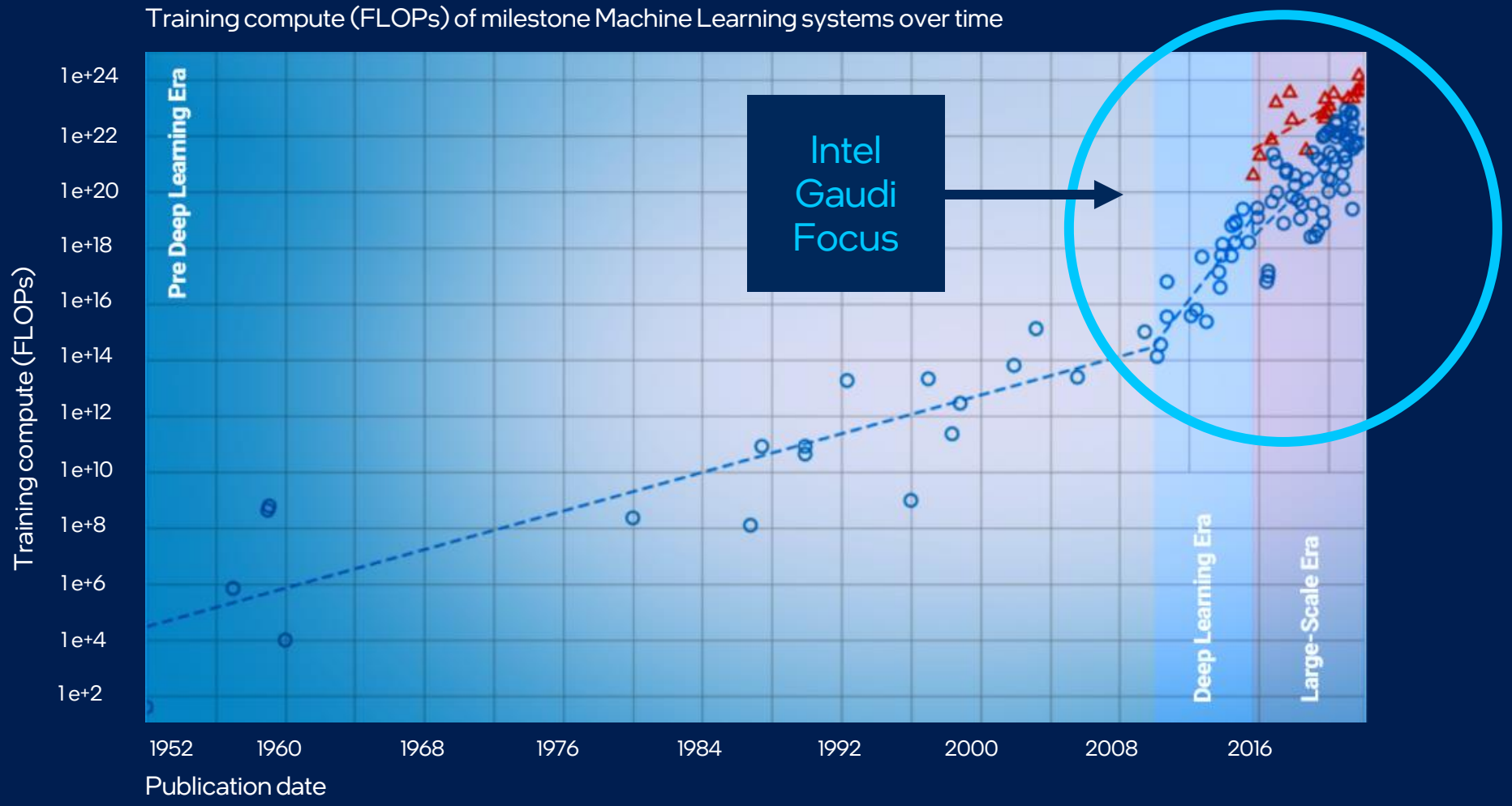
intel Ai summit

GenAI and LLMs require…

- Deep learning performance
- Flexible, efficient scalability
- Developer ease of use



Intel® Gaudi® AI Accelerator: Built for the Deep Learning and Large-Scale Eras



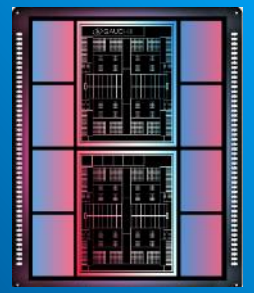
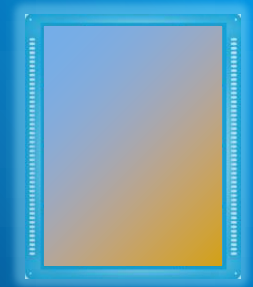
intel AI summit

Intel® AI Portfolio

Software	oneAPI, OpenVINO, PyTorch, deepspeed, TensorFlow, XGBoost, pandas, SynapseAI	
Deep Learning Acceleration	Intel Gaudi	Dedicated Deep Learning Training and Inference
General Acceleration	Intel Data Center GPU Flex Series	Cloud Gaming, VDI, Media Analytics, Real-Time Dense Video
	Intel Data Center GPU Max Series	Parallel Compute, HPC, AI for HPC
General Purpose	Intel Xeon, Intel Xeon Max Series, Intel Ethernet	Real-Time, Medium Throughput, Low Latency, and Sparse Inference
	Intel Xeon, Intel Xeon Max Series, Intel Xeon W, Intel Ethernet	Medium- to Small-Scale Training and Fine Tuning
	Intel Core Ultra, Intel Xeon D, Intel Data Center GPU Flex Series, Intel Arc Graphics, Intel Core, Intel Ethernet	Edge and Network AI Inference
	Intel Core, Intel Core Ultra, Intel Arc Graphics	Client AI Usages

intel AI summit

Intel® Gaudi® AI Accelerator Roadmap

Intel Gaudi Accelerator	Intel Gaudi 2 Accelerator	Intel Gaudi 3 Accelerator	Intel next-gen accelerator
			
(16nm)	(7nm)	(5nm)	
Available Now	Available Now	2024	2025

intel Ai summit

Intel® Gaudi® 2 Accelerator: Architected for GenAI and Large Language Models

Heterogenous AI-custom compute engine

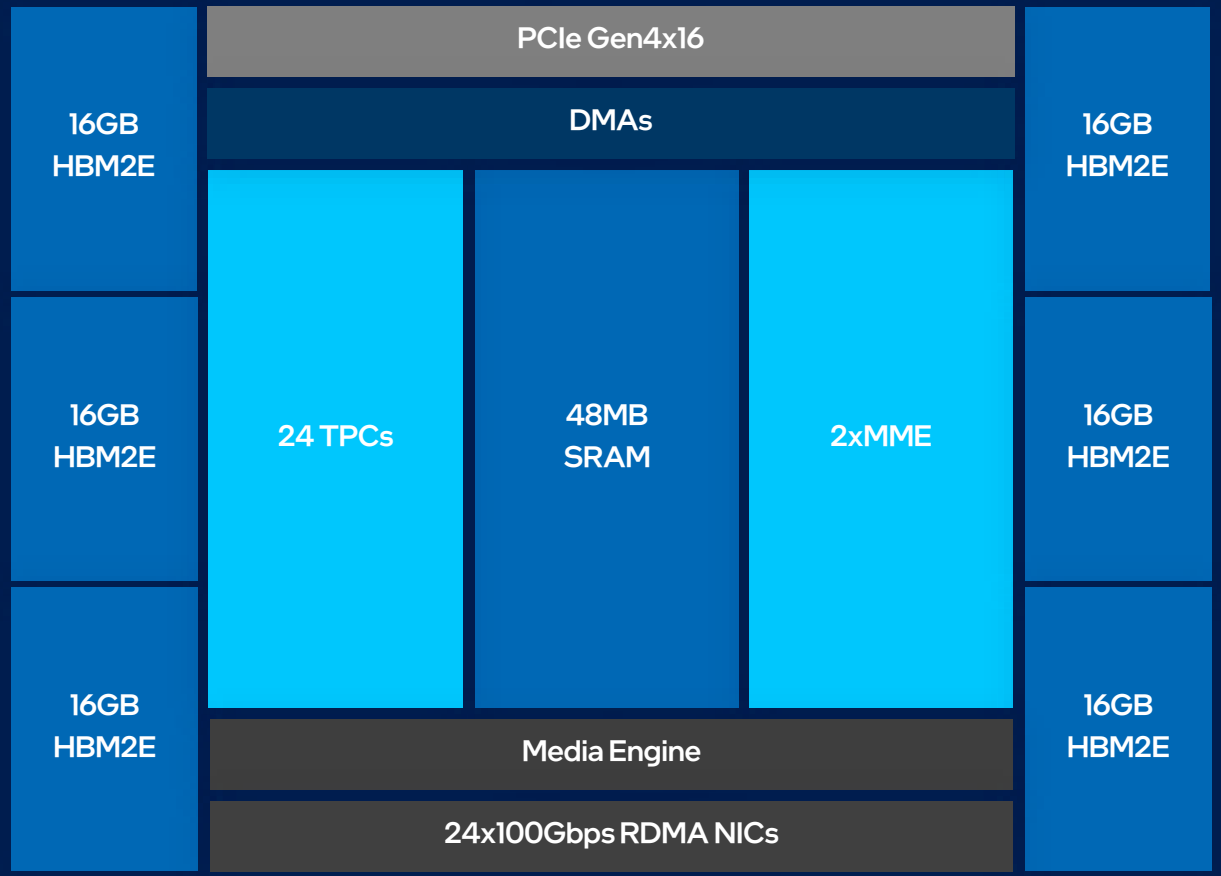
- 24 Tensor processor cores
- Dual matrix multiplication engines

Leading on-chip memory capacity

- 96 GB HBM2E
- 48 MB SRAM

Massive, flexible on-chip networking

- 24x 100 GbE RoCE ports on every Intel Gaudi2 chip

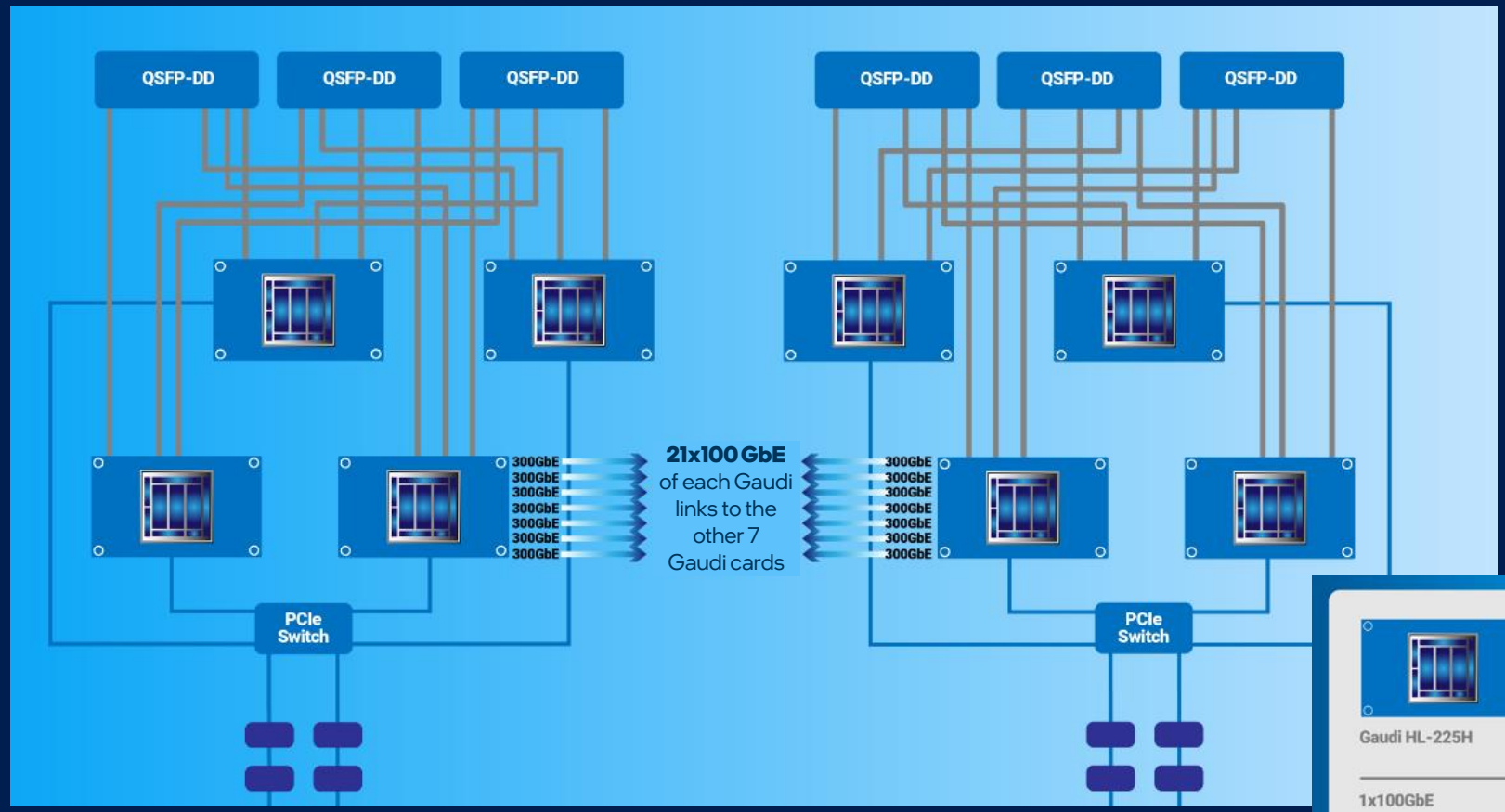


intel Ai summit

Intel® Gaudi® 2 Server: Designed for Flexible, Efficient Scalability

HLS-Gaudi 2 Reference Server featuring...

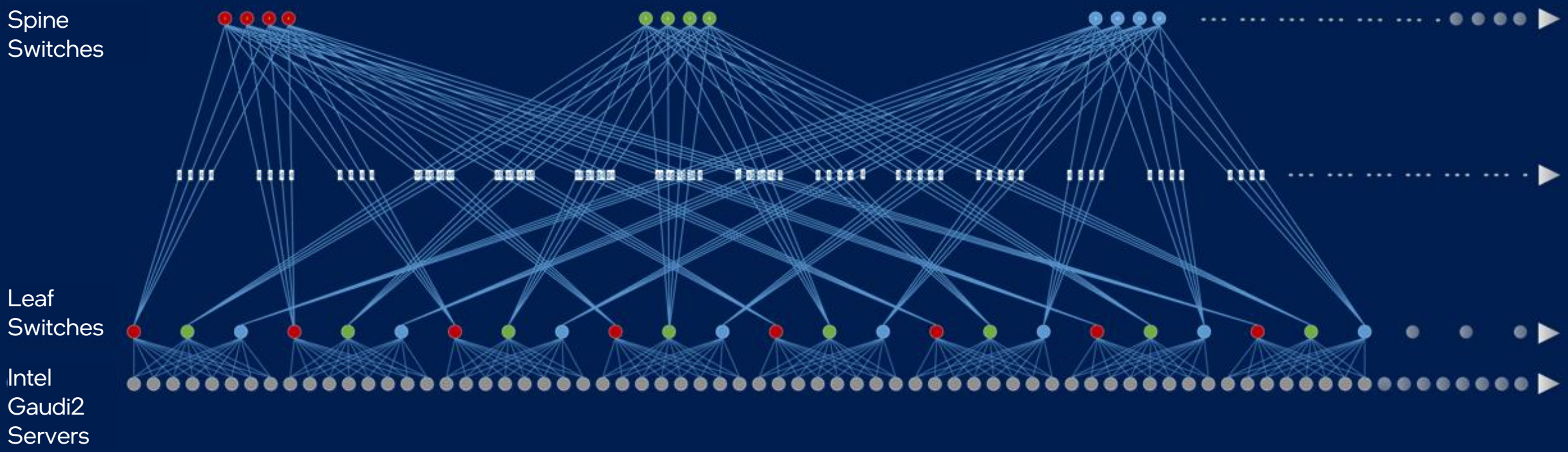
- 8 Intel Gaudi 2 mezzanine cards
- 24x 100 GbE ports per card
  - 21 for all-to-all connectivity to other 7 Intel Gaudi processors within the server
  - Three to scale out
  - Through six QSFP-DD ports
- Dual-socket Host CPU: Intel® Xeon® Scalable processor



Dual-socket Intel Xeon Scalable Processor Host CPU

intel Ai summit

Near-Infinite System Scale-Out Possibilities to Meet Demands for the Largest GenAI Models



Intel® Gaudi® Performance in MLPerf

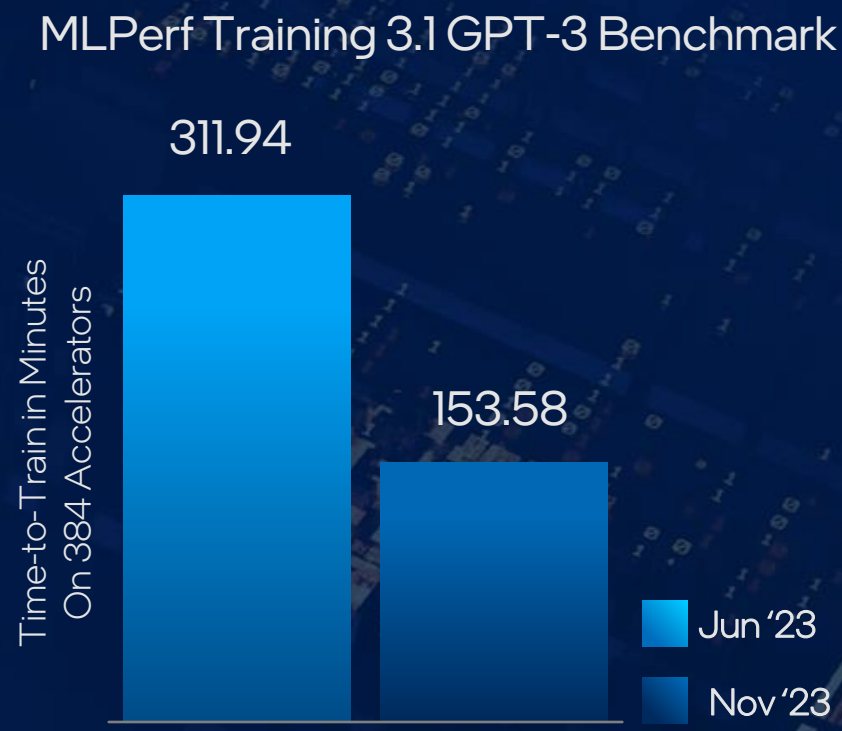
intel Ai summit

Intel® Gaudi® 2 Accelerator Performance Doubled with FP8

We projected for customers +90% performance gain with FP8

Delivered more than 2x

Committed to credible, reliable performance projections and delivering on them



intel Ai summit

Performance source: MLPerf Training 3.1 results https://mlcommons.org/benchmarks/training/

40-50% Better Price-Performance on GPT-3 vs. H100

Based on the most recent MLPerf Training benchmark and estimated range of server pricing, Intel Gaudi2 AI accelerator delivers between 40 and 50% better price performance than Nvidia H100.

- 1 of only 2 merchant silicon submissions for GPT-3
- H100/FP8 outperformed Gaudi2/BF16 on BERT
- Intel® Gaudi® 2 ResNet result near H100 submission

intel Ai summit

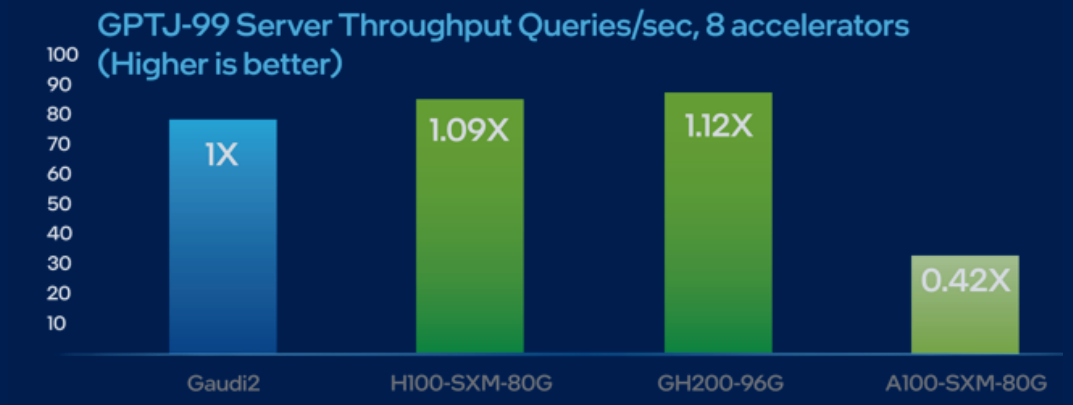
Pricing information for range of servers based on Semianalysis GPU Utils for cloud (with pricing) and for servers as of December 2023.
Performance source: MLPerf Training 3.1 results https://mlcommons.org/benchmarks/training/

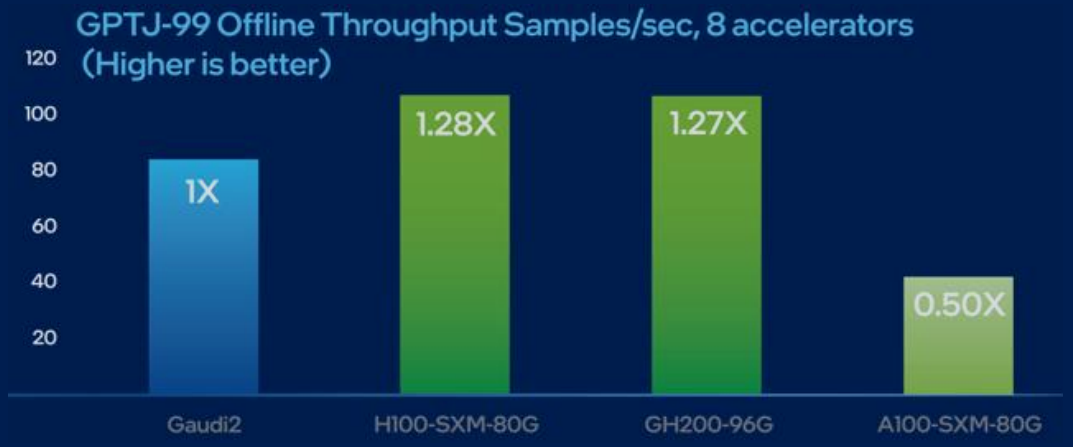
Near-Parity GPT-J Inference Performance vs. H100

Intel® Gaudi® 2 with FP8 achieved accuracy of

99.9%

- Intel Gaudi 2 throughput:
  -9% (server) and
  -28% (offline) vs H100
- Vs. A100: 2.4x (Server) and 2x (Offline)





intel Ai summit

Performance source: MLPerf Inference 3.1 results https://mlcommons.org/benchmarks/inference-datacenter/

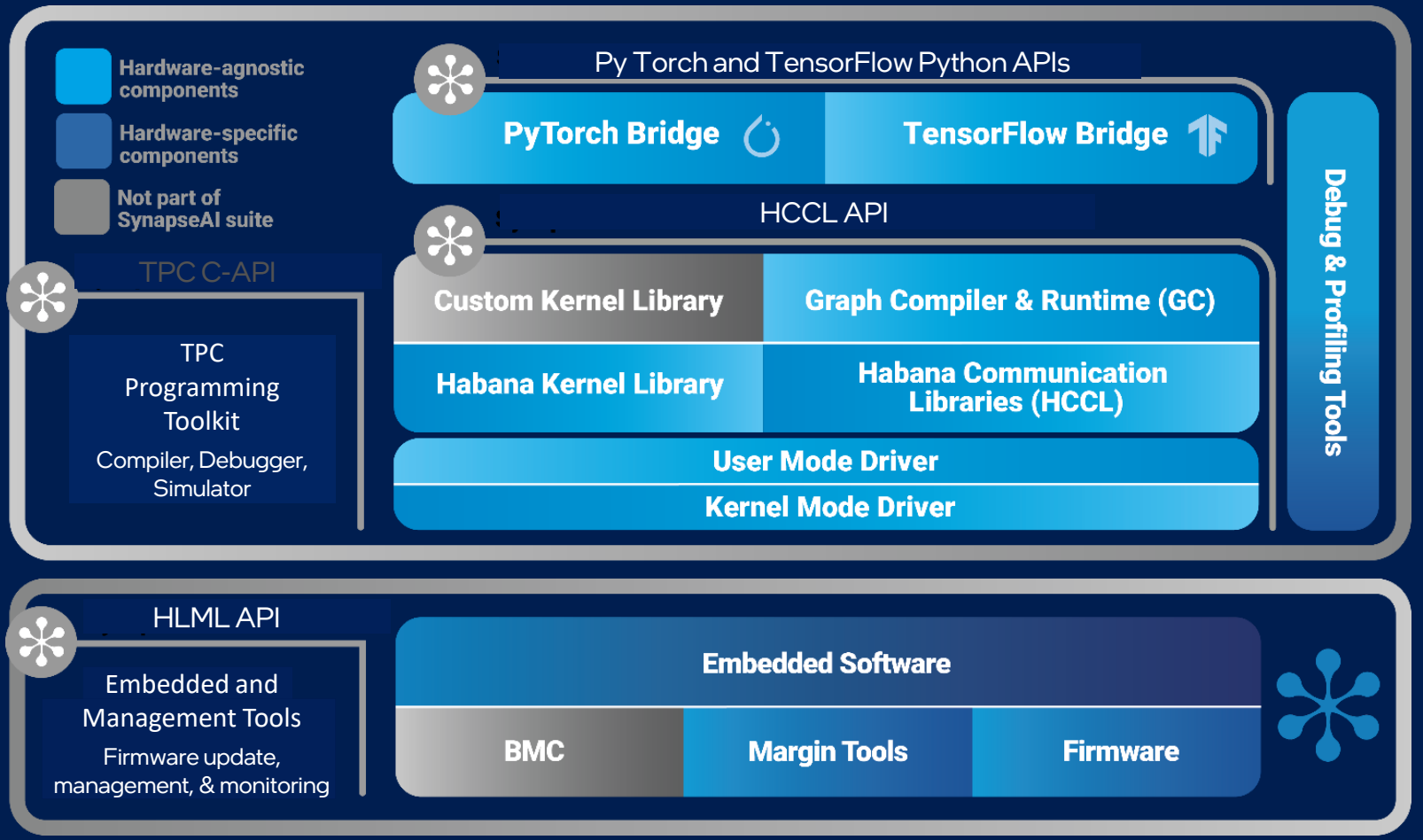
Intel® Gaudi® Software Ecosystem

intel Ai summit

Software Optimized for Intel® Gaudi® AI Accelerator Performance and Usability

SynapseAI®

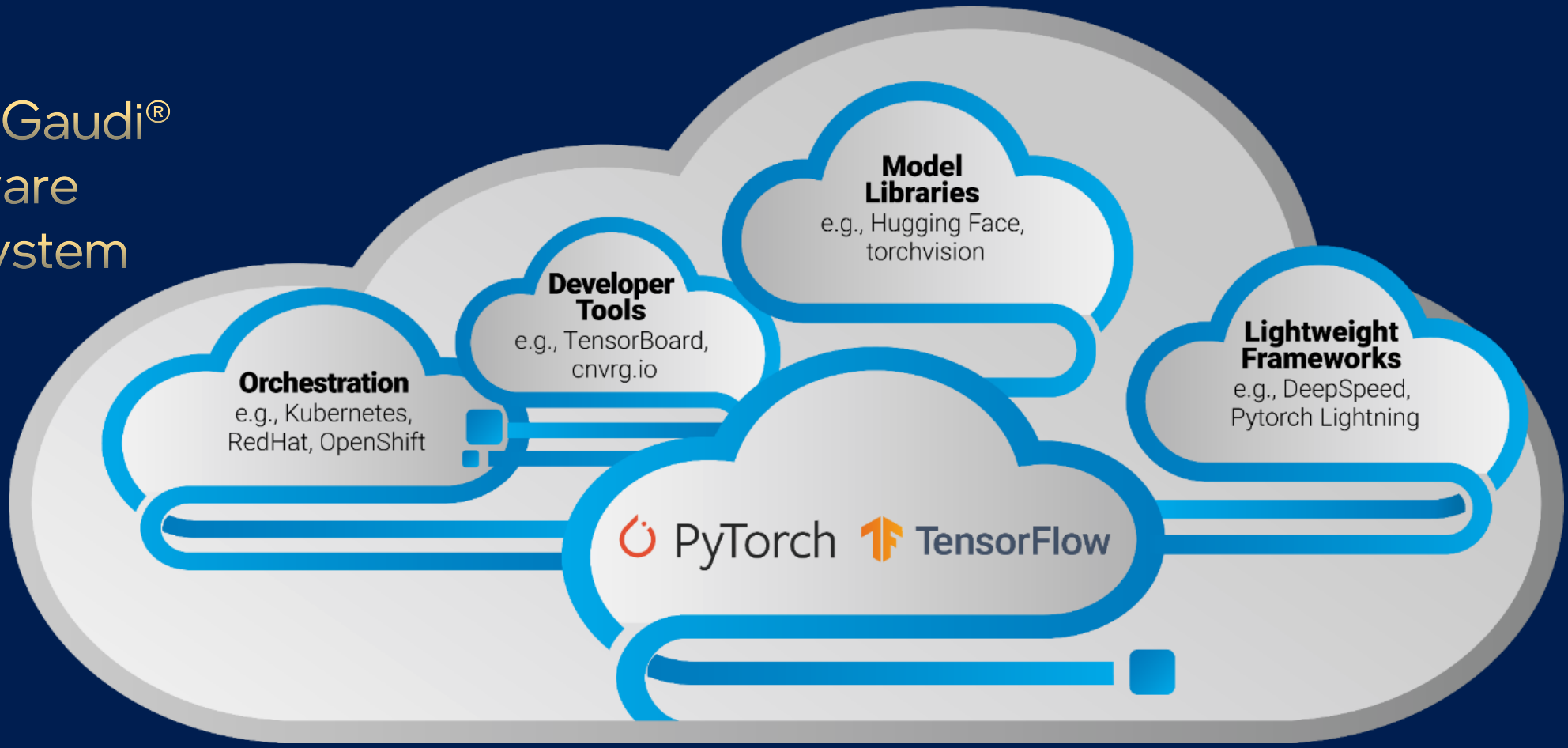
Will evolve to merge with OneAPI, delivering forward compatibility for Gaudi software to Falcon Shores



intel AI summit

Access to Thousands of AI Models to Ease Development Process

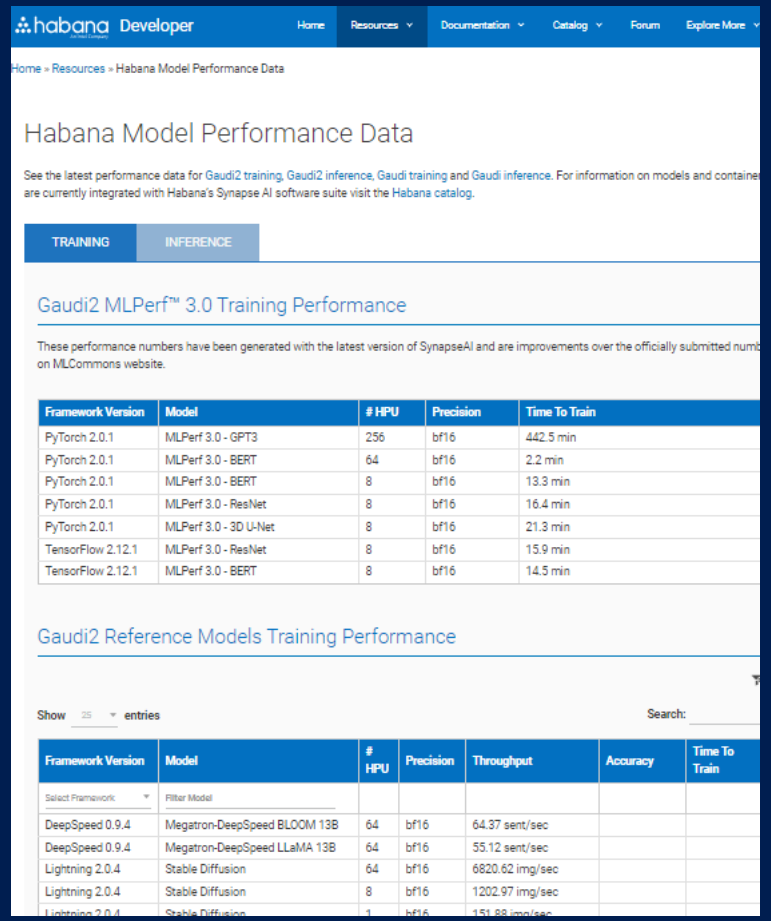
Intel® Gaudi® Software Ecosystem



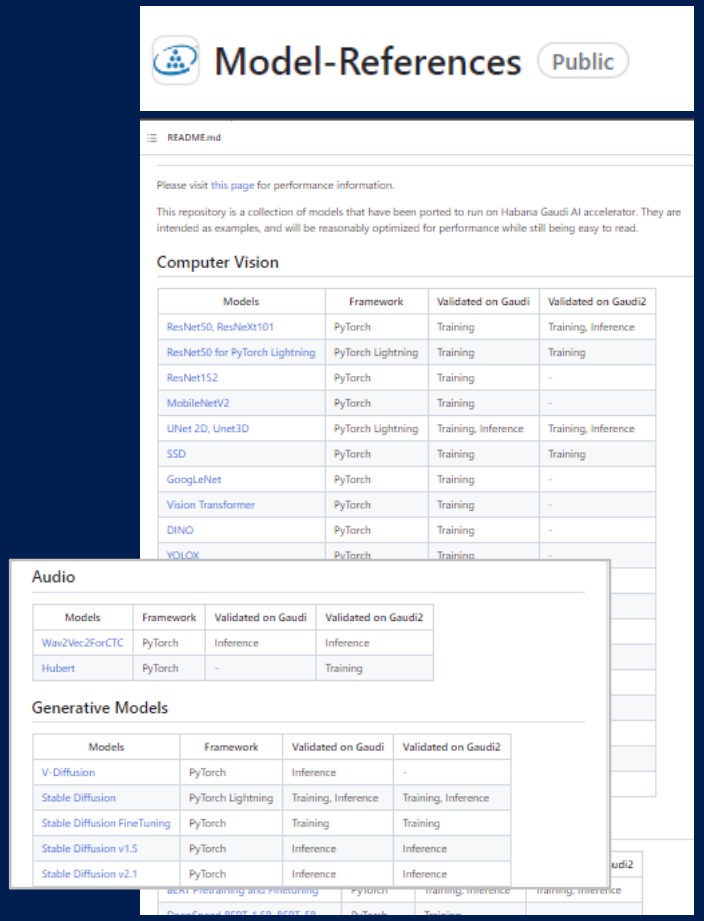
intel Ai summit

Developer Resources

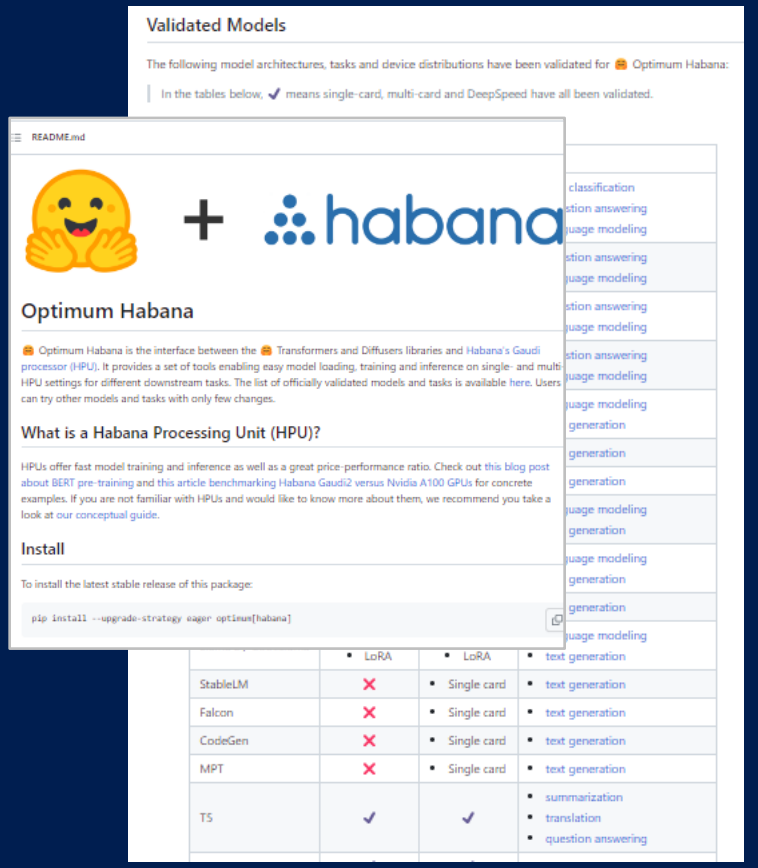
Intel® Gaudi® Developer Site



Intel Gaudi GitHub



Intel Gaudi Optimum Library on Hugging Face Hub



intel AI summit

Train on Gaudi2

Native PyTorch

based model migration

Tools:

- GPU migration tools
- HPU graph
- FP8

References:

- Habana Model References

Megatron & DeepSpeed

based model migration

Tools:

- 3D Parallelism
- DP Zero
- LLM pretrain
- FP8

References:

- Habana Megatron-DeepSpeed
- Habana DeepSpeed

Huggingface

based model migration

Tools:

- HF API compatibility
- One step train
- DP Zero
- FP8

References:

- Optimum Habana

Inference on Gaudi2

Model / Framework based	Tools	References
Native PyTorch	GPU migration tools DeepSpeed Tensor parallelism FP8 quantization tool HPU graph	Habana Model References
Huggingface	HF API compatibility One step inference DeepSpeed Tensor parallelism FP8 quantization tool HPU graph	Optimum Habana
vLLM / TGI	To be released soon	Habana vLLM / TGI-gaudi
Ray.io	To be released soon	

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 Gaudi

intel Ai summit

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

Getting Started With Hugging Face on Gaudi

```python
from optimum.habana import GaudiConfig, GaudiTrainer, GaudiTrainingArguments
from transformers import BertTokenizer, BertModel
…
tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
model = BertModel.from_pretrained("bert-base-uncased")
gaudi_config = GaudiConfig.from_pretrained("Habana/bert-base-uncased")
args = GaudiTrainingArguments(
    output_dir="/tmp/output_dir",
    use_habana=True,
    use_lazy_mode=True,
)
trainer = GaudiTrainer(
    model=model,
    gaudi_config=gaudi_config,
    args=args,
    tokenizer=tokenizer,
)
trainer.train()
```

Supported Model Architectures include:

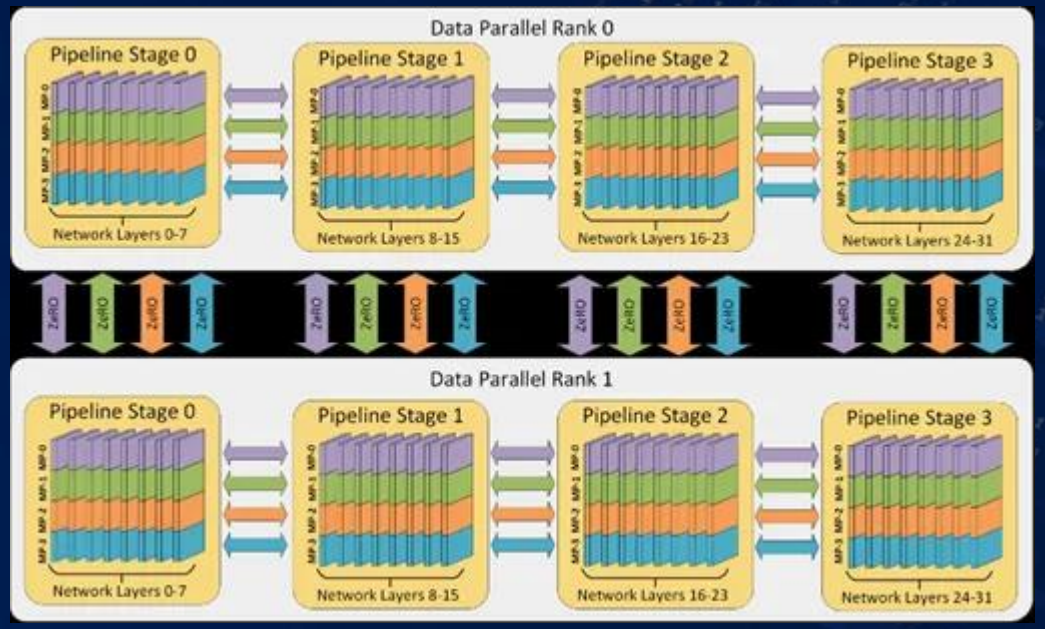
- NLP: BERT, AlBERT, DistilBERT, RoBERTa, T5, GPT2, BLOOM, LLaMA, MPT, Mixtral,…
- Vision: VIT, SWIN, Stable-diffusion-diffusers
- Audio: Wav2vec2

Optimum-Habana documentation: https://huggingface.co/docs/optimum/habana_index

Train models with 4D parallelism on Gaudi2

By using Megatron-DeepSpeed, train LLaMA on large scale Gaudi2 systems.

- Data Parallelism
- Tensor Parallelism
- Pipeline Parallelism
- Sequence Parallelism



Example: HL_HOSTSFILE=scripts/hostsfile HL_SEQ_PARALLEL=1 HL_MICRO_BATCH=1 HL_NUM_NODES=32 HL_PP=8 HL_TP=8 HL_DP=4 scripts/run_llamav2.sh

https://github.com/HabanaAI/Model-References/tree/master/PyTorch/nlp/DeepSpeedExamples/Megatron-DeepSpeed

FP8 Training with Intel Gaudi Transformer Engine

1. Import TE and use TE modules in your model, e.g. a te.Linear

```
import torch
import habana_frameworks.torch.hpex.experimental.transformer_engine as te

# Set dimensions.
in_features = 768
out_features = 3072
hidden_size = 2048

# Initialize model and inputs.
model = te.Linear(in_features, out_features, bias=True)
inp = torch.randn(hidden_size, in_features, device="hpu")
```

2. Wrap the forward pass of the training with fp8_autocast

```
from habana_frameworks.torch.hpex.experimental.transformer_engine import recipe

# Create an FP8 recipe. Note: All input args are optional.
fp8_recipe = recipe.DelayedScaling(margin=0, interval=1)

# Enable autocasting for the forward pass
with te.fp8_autocast(enabled=True, fp8_recipe=fp8_recipe):
    out = model(inp)

loss = out.sum()
loss.backward()
```

https://docs.habana.ai/en/latest/PyTorch/PyTorch_FP8_Training/index.html

Customer examples in GenAI & LLM

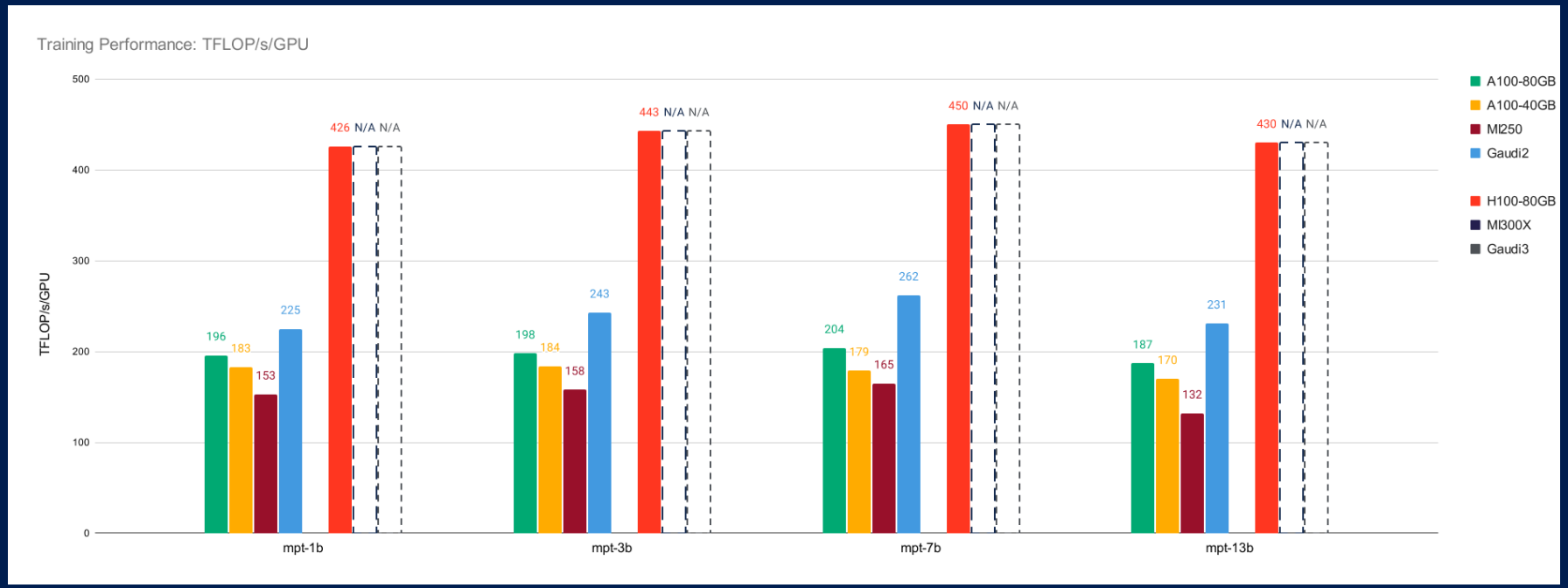
intel Ai summit

Intel Gaudi 2 Training New MPT Model from MosaicML: Outperforms H100 for price-performance

3rd party evaluation by Databricks' MosaicML:

https://www.databricks.com/blog/llm-training-and-inference-intel-gaudi2-ai-accelerators

- Delivers 260 TFLOPs-per-Gaudi2: ~55% training performance relative to H100
  - With its cost advantage over H100, Gaudi 2 outperforms H100 for price-performance training MPT



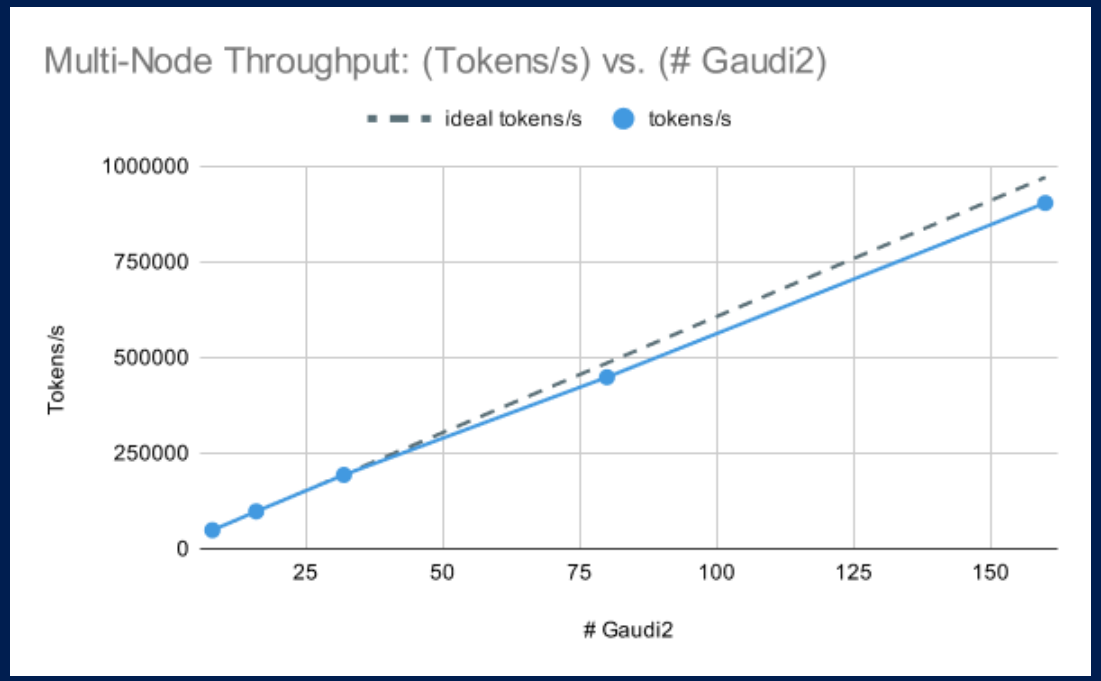
"The MI300x and Gaudi 3 are not yet profiled but are expected to be competitive with H100. "

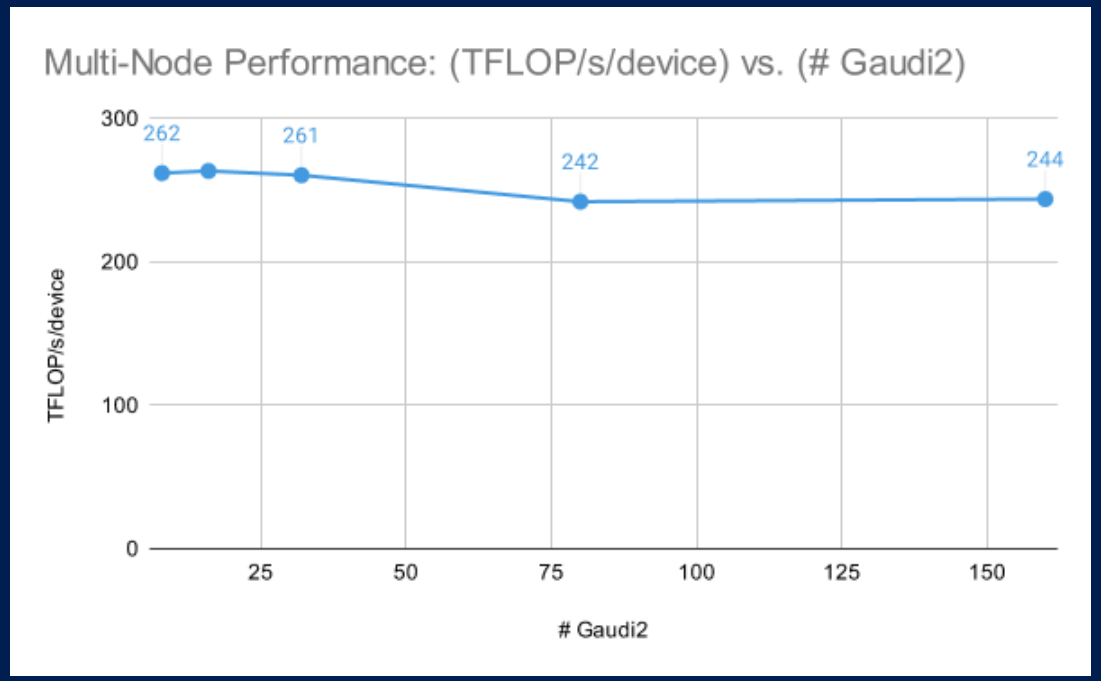
The MPT-7B model is the first entry in MosaicML's Foundation Series. MPT-7B is a transformer trained from scratch on 1T tokens of text and code. It is open source, available for commercial use, and MosaicML claims it matches the quality of LLaMA-7B. MPT-7B was trained on the MosaicML platform in 9.5 days with no human intervention. See configuration info in backup.

See back-up for test configurations. Results may vary.

MosaicML Evaluation: Gaudi2 near-linear scaling training MPT

"MPT-7B multi-node training performance: As we scale from 8xGaudi2 to 160xGaudi2, we see a near-linear increase in throughput and nearly constant TFLOP/s/device. Note that the global train batch size is held constant at 1920 samples, so these plots demonstrate strong scaling." https://www.databricks.com/blog/llm-training-and-inference-intel-gaudi2-ai-accelerators



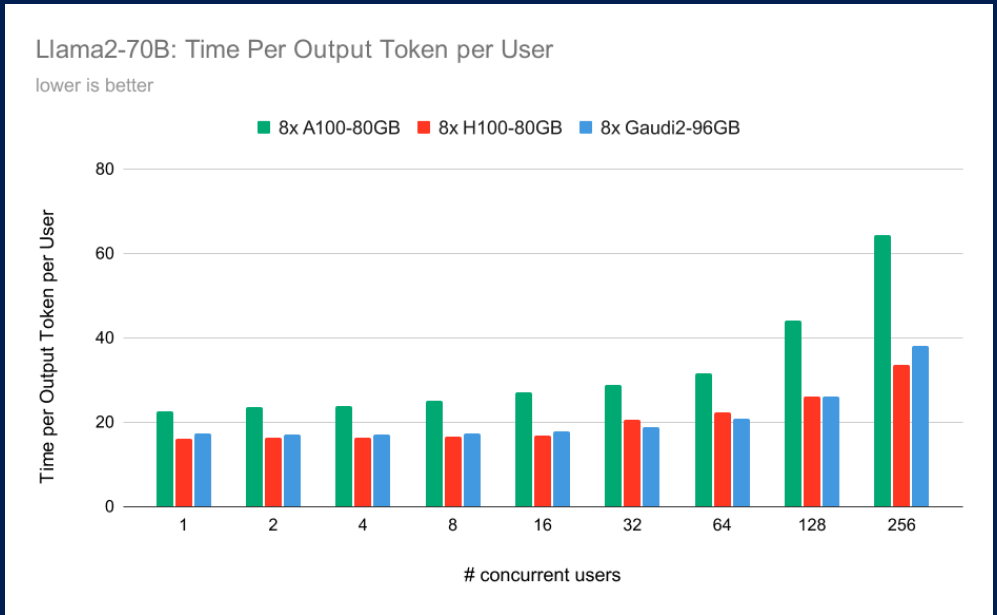


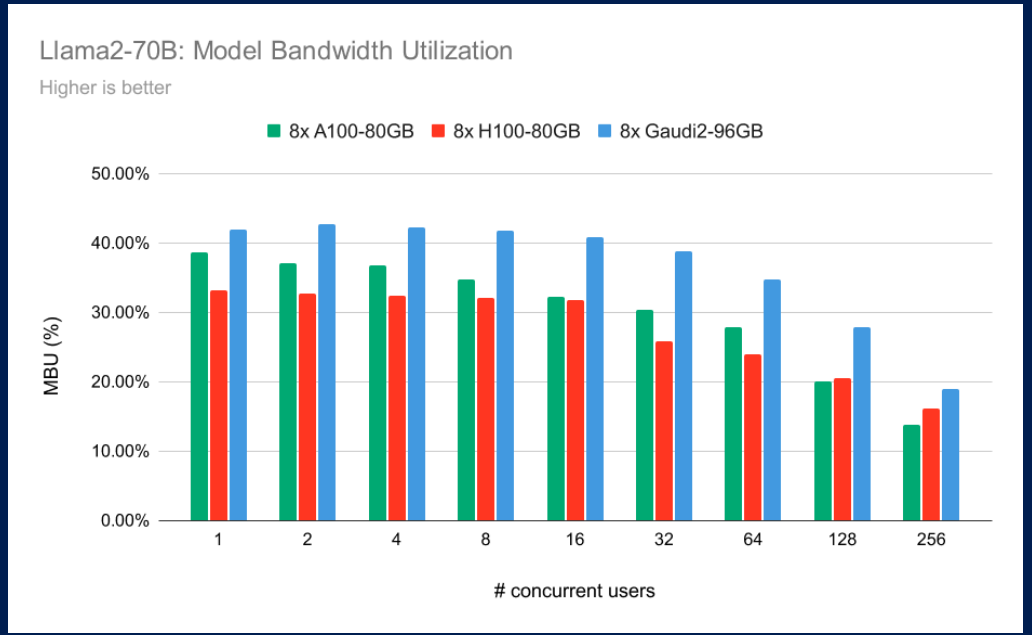
See back-up for test configurations. Results may vary.

intel Ai summit

MosaicML Evaluation: Gaudi2 Inference of LLama2-70B

https://www.databricks.com/blog/llm-training-and-inference-intel-gaudi2-ai-accelerators





"The TPOT for Gaudi2 is better than H100 under most of the user loads."

For decoding latency, the most expensive phase of LLM inference:

With 2450 GB/s of HBM2 memory bandwidth, Gaudi 2 matches H100 decoding latency with 3350 GB/s of HBM3 bandwidth

See back-up for test configurations. Results may vary.

Stability.ai Cluster Built on Intel® Gaudi® 2 AI Accelerator

"Today, we're excited to announce that we just secured a design win that's a pretty big deal. A large AI supercomputer will be built entirely on Intel's Xeon processors and 4,000 Intel Gaudi2 AI Hardware accelerators. Stability AI is now our anchor customer. Given the magnitude of the Stability AI build out, it will be a top 15 AI supercomputer in the world."

– Pat Gelsinger

intel GAUDI

intel AI summit

Stability.ai: Benchmarking Compute Solutions

- Training throughput of 2B MMDiT on Stable Diffusion 3

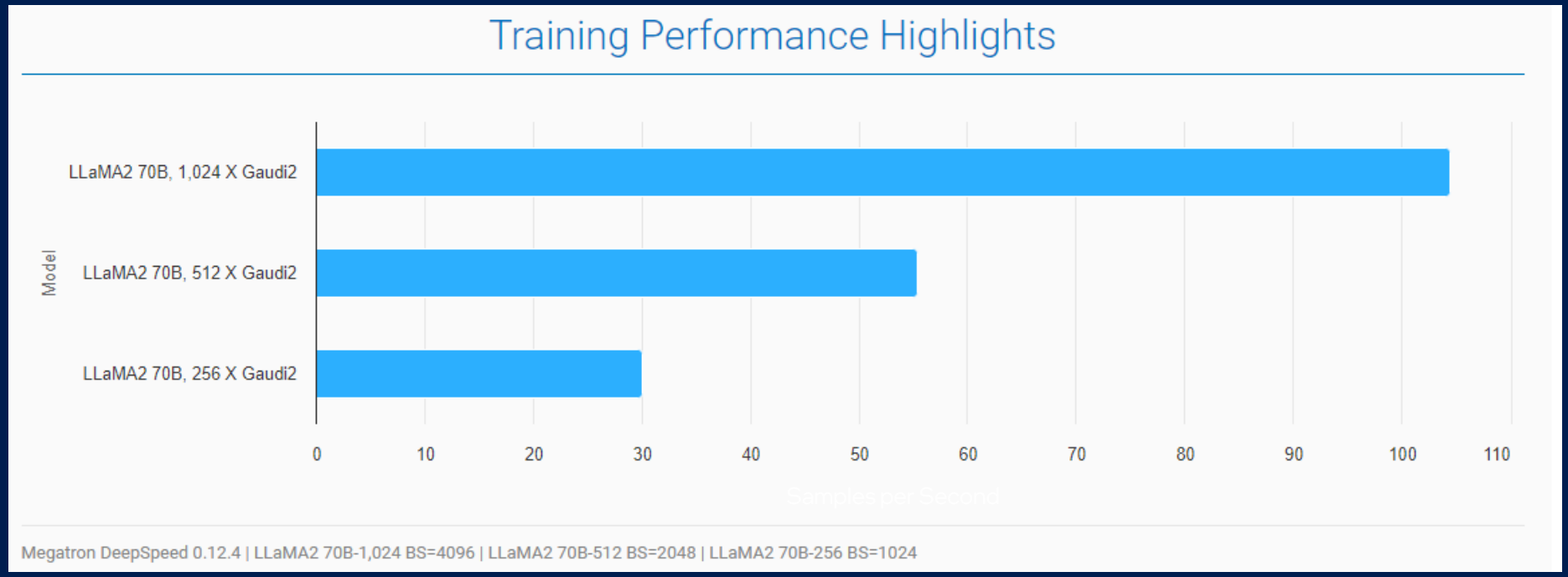
Device	Attention	# Nodes	# Accelerators (total)	Batch Size per Accelerator	Total Batch Size	Images / sec (100-MA)
Gaudi2	FusedSDPA	2	16	32	512	1,254
Gaudi2	FusedSDPA	2	16	16	256	927
H100-80GB	xFormers	2	16	16	256	595
A100-80GB	xFormers	2	16	16	256	381

Device	# Nodes	# Accelerators (total)	Batch Size per Accelerator	Total Batch Size	Images / sec	Images / sec / device
Gaudi2	32	256	16	4096	12,654	49.4
A100-80GB	32	256	16	4096	3,992	15.6

https://stability.ai/news/putting-the-ai-supercomputer-to-work

LLaMA2-70B Training on Gaudi2

- Train LLaMA2-70B on 256x, 512x, 1024x Gaudi2



intel Ai summit

Notices and Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:

intel Ai summit

Thank You!