

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

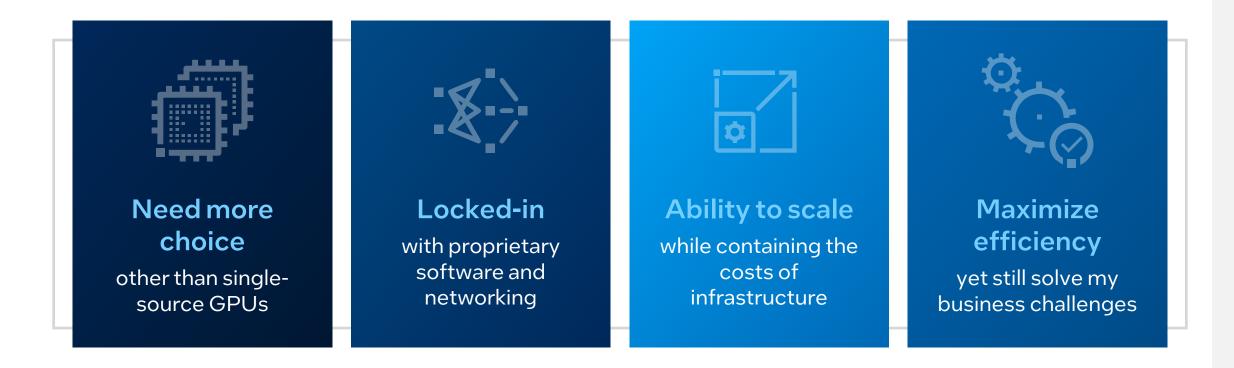
Intel[®] Gaudi[®] 3 Al accelerator Version 2.6 Table of Contents

Intel® Gaudi® 3 AI Accelerators

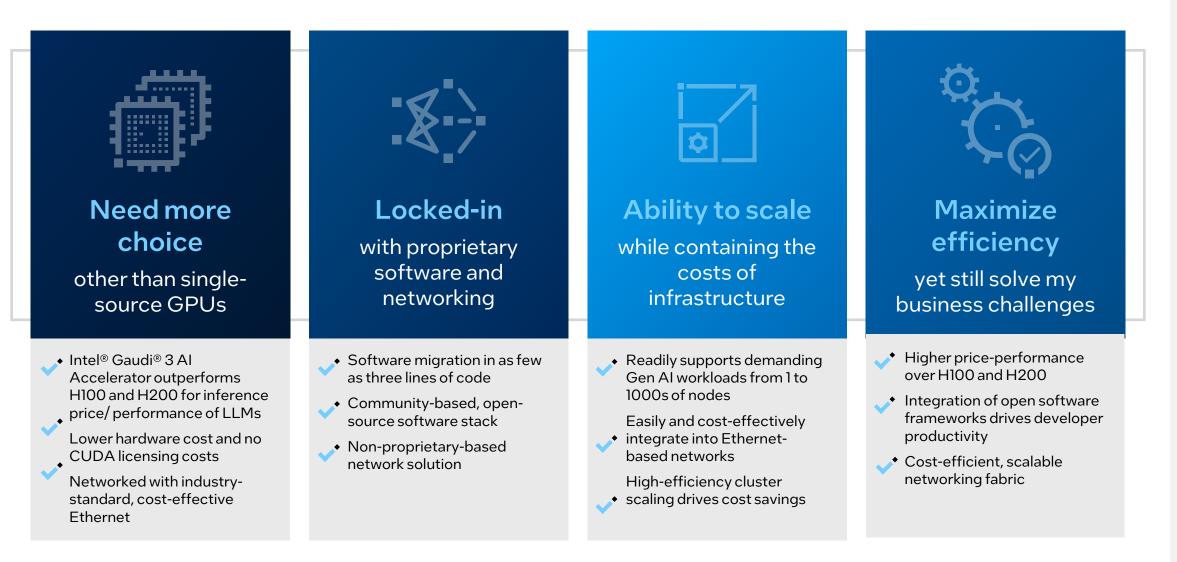
- Customer Challenges
- Introducing Intel[®] Gaudi[®] 3 AI Accelerator
- Product Line Details
- Performance
- Software Support
- Availability and Customer Momentum
- Open and Efficient Scalability

Customer Challenges

Customer Challenges with AI Compute Solutions



How Intel® Gaudi® 3 AI Accelerator Addresses Enterprise Challenges



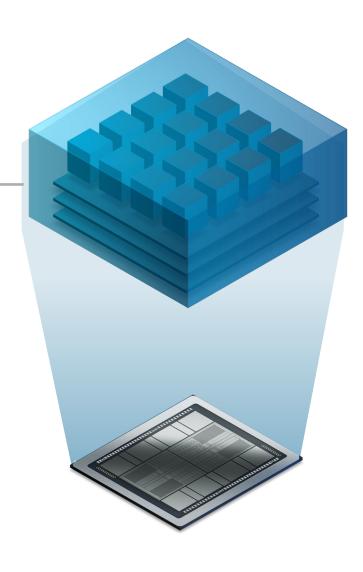
5

intel Gaudi

Broad Al Application Support

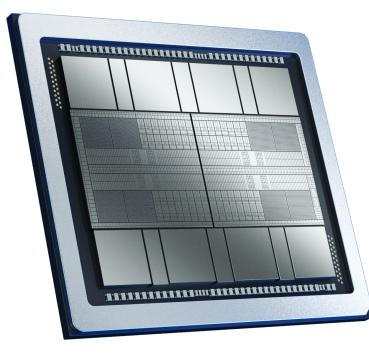
With focus on multi-modal, LLM, and RAG

AI	Applicatio	ns
	AI Functions	
3D Generation	Text Generation	Classification
Video Generation	Sentiment	Translation
Image Generation	Summarization	Q&A
C	Core Capabilitie	S
Mu	ulti-modal Models	
	LLM	
	RAG	



Introducing Intel® Gaudi® 3 Al Accelerator

Architected for Gen Al Performance & Productivity



Increased memory for LLM efficiency and cost effectiveness

128gb

HBM capacity, 3.7 TB/s B/W SRAM B/W

96MB SRAM, 12.8 TB/s

Massive, flexible, on-chip networking

Open standard vs. proprietary InfiniBand

 24×200 GbE

Ethernet ports

PCle 5

Industrystandard RoCE x 16

Designed for AI

Driving greater efficiency & performance

64 Tensor Processor Cores

8 Matrix Math Engines

8

Intel[®] Gaudi[®] 3 Al Accelerator Product Specs Advances over Intel Gaudi 2 Al accelerator

	Feature	Intel Gaudi 2 Al accelerator	Intel Gaudi 3 Al accelerator
Availability		Now	Now
	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
Architectural	PCle	xl6 Gen 4	x16 Gen 5
Features	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 Al Accelerator Gen Al. Your way.



Competitive Gen Al Price-Performance

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



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 server³

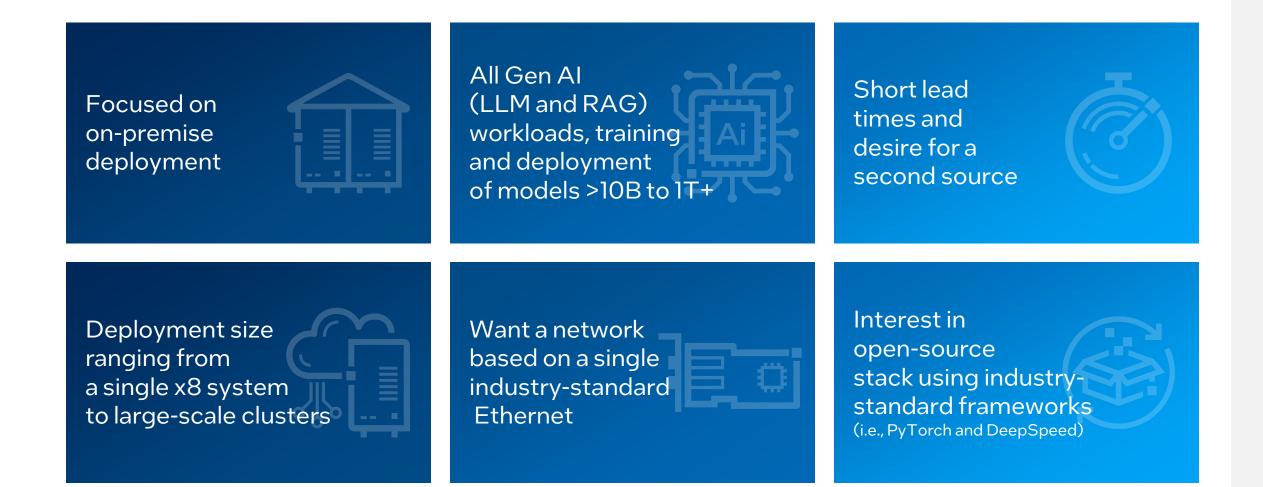


Open Development on Gen AI platforms

- 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 Al Accelerator



Intel[®] Gaudi[®] 3 AI Accelerator Spec and Block Diagram

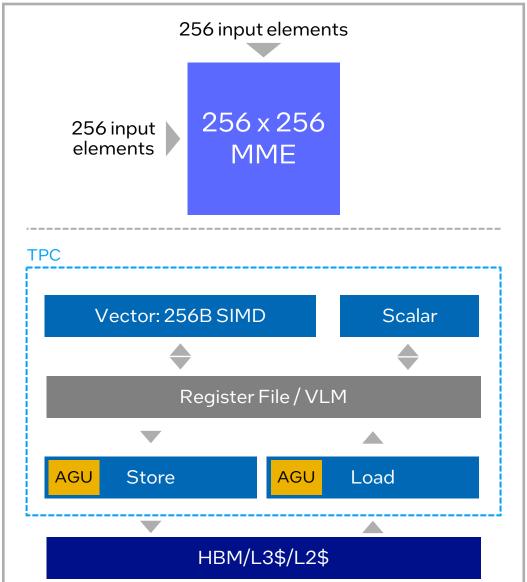
Feature/Product	Intel [®] Gaudi [®] 3 Accelerator
BF16 Matrix TFLOPs*	1678
FP8 Matrix TFLOPs*	1678
BF16 Vector TFLOPs	28.7
MME Units	8
TPC Units	64
HBM Capacity	128 GB
HBM Bandwidth	3.67 TB/s
On-die SRAM Capacity	96 MB
On-die SRAM Bandwidth RD+WR (L2 Cache)	19.2 TB/s
Networking	1200 GB/s bidirectional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

_		HBM PI	ΗY	HBM PHY	НВМРНУ	ΗB	MPHY		
x8 PCle Gen5		MME	MME	16 TPCs	16 TPCs	MME	MME		x8PCle Gen5
	12x 200 GbE		48MB	SRAM	48MB	SRAM		12x 200 GbE	
Media Engine		MME	MME	16 TPCs	16 TPCs	MME	MME		Media Engine
		HBM PI	HY	HBM PHY	HBM PHY	HB	MPHY		

Matrix Multiplication and Vector	or 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 movem increasing efficiency	ient,
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-accele	erated address

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

generation



Memory Sub-System

Unified Memory Space of L2 / L3 / HBM

Near Memory Compute:

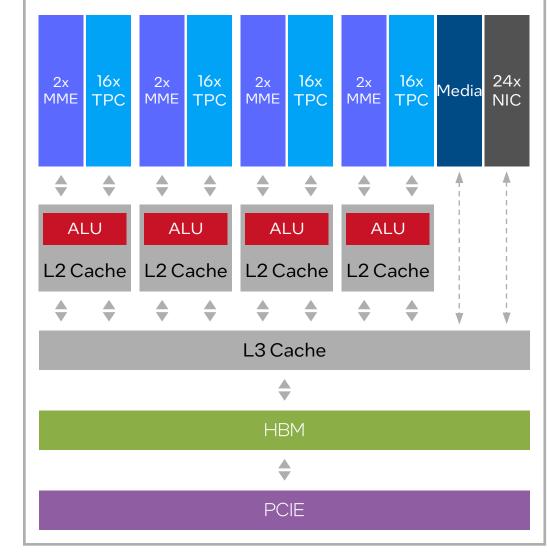
- Add/Sub
- Max / Min

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

Memory Sub-System Logical View



MME-TPC Parallelism via Pipelining

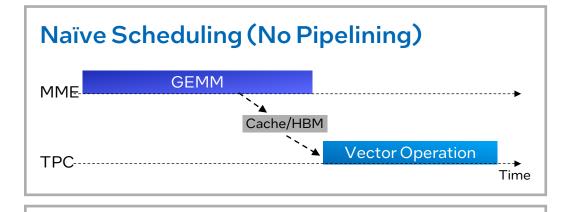
Graph Compiler orchestrates MME & TPC parallelism

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

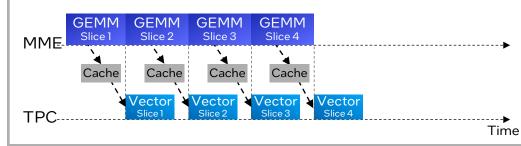
Slice size is determined to balance between the following:

- 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

Execution Sub-Graph Model Parameters Input Activations GEMM GEMM Output Vector Op Vector Output



Efficient Scheduling (With Pipelining)



Product Line Details

intel Gaudi Product Line





Accelerator Card

HL-325L OAM-Compliant

Universal Baseboard

HLB-325



Enabling customer infrastructure choice

intel Gaudi

.....

Accelerator Card HL-325L (OAM-Compliant)

28 _{GB}	3.
BM2e	HBN
	Ran

3.7TB/s HBM Bandwidth

8 Matrix Multiplication

Engines

24 200 GbE RDMA NICs **1.2**TB/s Bi-directional Networking

intel Gaudi

00

ö

Universal Baseboard HLB-325

64 Matrix Multiplication Engines

192 200 GbE RDMA NICs 9.6TB/s Bi-directional

Bi-directiona Networking

Networking Intel[®] Gaudi[®] 3 Accelerator reference based-Server



Collective operations execute with low control overhead

Intel[®] Gaudi[®] 3 Servers feature:

2x Intel® Xeon® Host CPUs

8x Intel Gaudi 3 OAM Cards

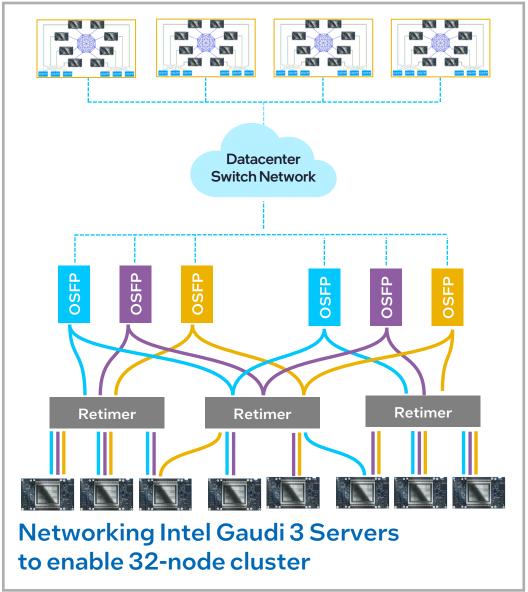
Peer-to-peer (P2P) connection between each pair of Intel Gaudi 3 cards

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 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)

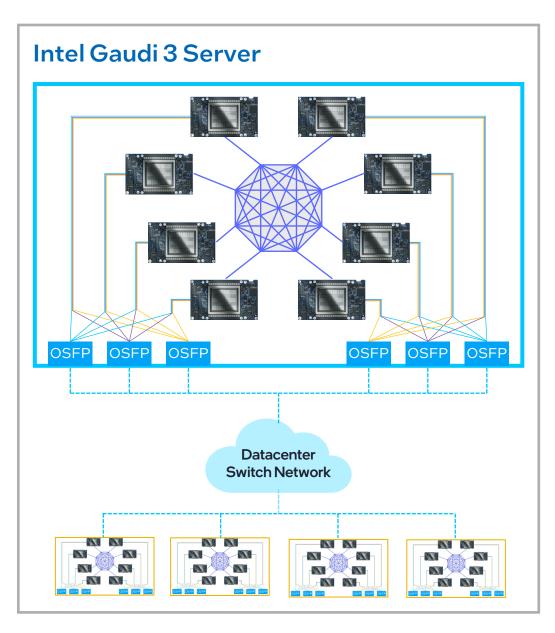
Peer-to-peer (P2P) connection between each pair of Intel Gaudi 3 cards

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)



Performance

All Public Performance benchmarks are here.....

Developers 🗸 / Hardware Platforms 🗸 / Intel® Gaudi® Al Accelerators 🗸 / Model Performance Data

Model Performance Data for Intel[®] Gaudi[®] 3 AI Accelerators

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 \rightarrow

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)
27b 1 fp8 128 128 1,536 19,810	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

intel gaudi

Software Support

Intel® Gaudi® Software Suite

Integrates the main Gen AI frameworks used today

Supports FP16/BF16 \rightarrow 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)

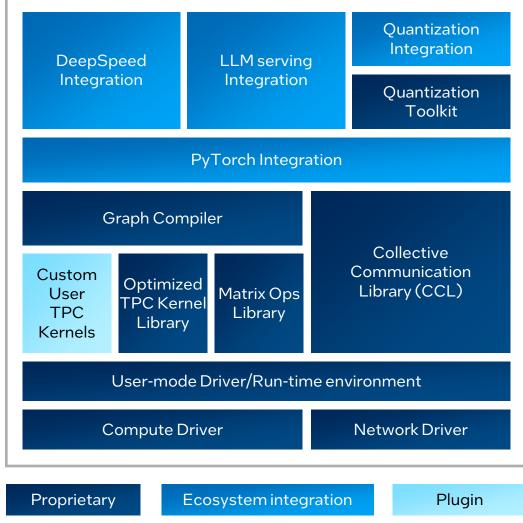
Several sources for TPC Kernels

Intel® Gaudi® software optimized TPC kernel library

Custom user kernels

MLIR-based fused kernels: generated during graph compilation

Layered View of Intel® Gaudi® Software Suite



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 Al

Composable Microservices with OPEA Enterprise RAG use case solutions

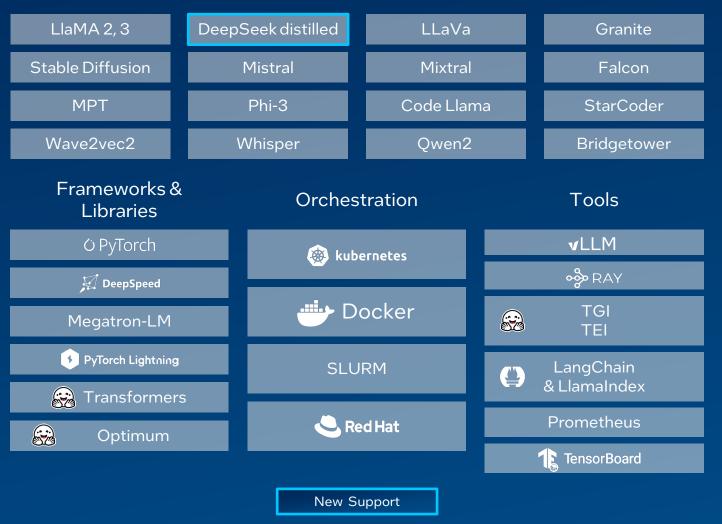
Customers & Developers can experience the Intel Gaudi platform on Intel® Tiber™ Developer Cloud

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



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
		MLPerfv3.0	MLPerfV3.1			MLPerfv4.0				
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/70E Mixtral 8x7B	
Less than 6 mo param LLM	nths to inference	on 100B+								
Less than 12 mo	onths to large-sca	ale training on 100	B+ 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*

29

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

abana Mo	odel Performance	Data	a								
als arra rra											
	ce data for Gaudi2 training, Gaudi2 infer vith Habana's Synapse AI software suite				nce. For info	rmation on mo	dels and cor				
contentity integrated in		that the fit	iouria cutare	9.							
TRAINING	INFERENCE										
Gaudi2 MLPe	erf™ 3.0 Training Perfor	mance	÷								
hese performance nu n MLCommons webs	imbers have been generated with the lat	est version	of Synapse	Al and are imp	rovements	over the official	ly submitted				
TIMECOTTINICITS WEDG	we.										
Framework Version	Model	# HPU	Precis	ion T	ime To Train						
PyTorch 2.0.1	MLPerf 3.0 - GPT3	256	bf16	4	42.5 min						
PyTorch 2.0.1	MLPerf 3.0 - BERT	64	bf16	2	2 min						
PyTorch 2.0.1	MLPerf 3.0 - BERT	8	bf16	1	3.3 min						
PyTorch 2.0.1	MLPerf 3.0 - ResNet	8	bf16	1	6.4 min	1					
PyTorch 2.0.1	MLPerf 3.0 - 3D U-Net	8	bf16	21.3 min							
Fytototi 2.0.1	MLPerf 3.0 - ResNet	8	bf16	1	5.9 min						
TensorFlow 2.12.1	Willreit 3.0 - Nebivel										
	MLPerf 3.0 - BERT	8	bf16	1	4.5 min						
TensorFlow 2.12.1		8	bf16	1	4.5 min						
TensorFlow 2.12.1 TensorFlow 2.12.1	MLPerf 3.0 - BERT			1	4.5 min						
TensorFlow 2.12.1 TensorFlow 2.12.1				1	4.5 min						
TensorFlow 2.12.1 TensorFlow 2.12.1	MLPerf 3.0 - BERT			1	4.5 min						
TensorFlow 2.12.1 TensorFlow 2.12.1	MLPerf 3.0 - BERT			1.	4.5 min						
TensorFlow 2.12.1 TensorFlow 2.12.1 Gaudi2 Refer	MLPerf 30-BERT			1.	4.5 min	Sear	rch:				
TensorFlow 2.12.1 TensorFlow 2.12.1 Gaudi2 Refer	MLPerf 3.0 - BERT	Perform	nance		4.5 min						
TensorFlow 2.12.1 TensorFlow 2.12.1 Gaudi2 Refer	MLPerf 30-BERT	Perform		Throughput	4.5 min	Sear	rch: Time To Train				
TensorFlow 2.12.1 TensorFlow 2.12.1 Gaudi2 Refer how entri Framework Version	MLPerf 3.0 - BERT	Perform	nance		4.5 min		Time Te				
TensorFlow 2.12.1 TensorFlow 2.12.1 Caudi2 Refer how <u>15 *</u> entrin Framework Version Salect Parmework *	MLPerf 30 - BERT rence Models Training F es Model	Perforn # HPU	nance				Time Te				
TensorFlow 2.12.1 TensorFlow 2.12.1 Gaudi2 Refer how 25 * entri Framework Version Salect Paramotic * DeepSpeed 0.9.4	MLPer 3.0 - BERT ence Models Training F es Model Filter Model	Perform	nance Precision	Throughput	iec		Time Te				
TensorFlow 2.12.1 TensorFlow 2.12.1 Saudi2 Refer how <u> </u>	MLPerf 3.0 - BERT ence Models Training F es Model Filter Model Megatron-DeepSpeed BLOOM 138	Perform	Precision bf16	Throughput 64.37 sent/s	sec sec		Time Te				

Intel Gaudi Al Accelerator GitHub

https://github.com/HabanaAl

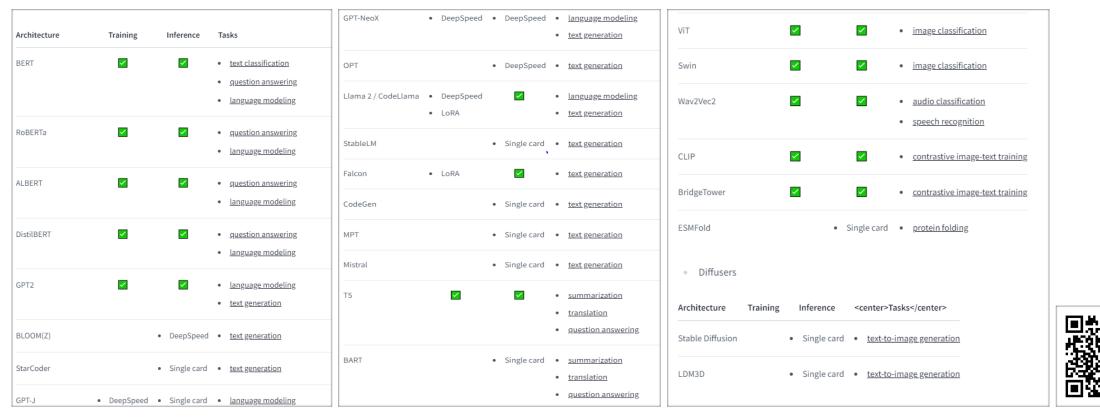
README								
Please visit <u>this page</u> I								
This repository is a co are intended as examp								
Computer Visio								
•								
Models			Framework		alidated on Gaud		lidated on Gaudi2	
ResNet50, ResNeXt		PyTorch		raining		iining, Inference		
ResNet50 for PyTore	h Lightning		yTorch Lightning		raining		iining	
ResNet152			PyTorch		raining	-		
MobileNetV2			PyTorch		raining	-		
UNet 2D, Unet3D			PyTorch Lightning		raining, Inference		iining, Inference	
SSD			PyTorch		raining	Training		
GoogLeNet			PyTorch		raining	-		
Vision Transformer			PyTorch		raining	-		
DINO			PyTorch		raining			
YOLOX			PyTorch		raining			
ResNet50 Keras			TensorFlow	Training			iining	
ResNeXt101			[ensorFlow				iining	
DenseNet			TensorFlow		raining	-		
Vision Transformer			[ensorFlow	1	raining			
Natural Langua	ige Proc	es	sing					
Mode	ls		Framework	Val	idated on Gaudi	Valic	lated on Gaudi2	
BERT Pretraining an	d Finetuning		PyTorch	Tra	ining, Inference	Train	ing, Inference	
DeepSpeed BERT-1.		3	PyTorch		ining			
BART			PyTorch		ining			
Audio								
Models	Framewor	k	Validated on G	audi	Validated on Ga	udi2		
Wav2Vec2ForCTC	PyTorch		Inference		Inference			
Generative Mo	dolo							
Generative Mo	uels							
Models		Fr	amework	Valida	ited on Gaudi	/alidat	ed on Gaudi2	
						Training, Inference		
Stable Diffusion	P	yTor	ch Lightning	Trainir	ng, Inference	Training	, Inference	

Intel Gaudi Al Accelerator Hugging Face

ted Models			
ving model architecture	s, tasks and device	distributions have	been validated for 🦲 Optimum Habana:
ables below, 🗸 means	single-card, multi	card and DeepSpe	eed have all been validated.
formers:			
Architecture	Training	Inference	Tasks
BERT	~	~	text classification question answering language modeling
RoBERTa	1	J	question asswering language modeling
ALBERT	~	~	question answering language modeling
DistilBERT	~	~	question answering language modeling
GPT2	~	~	Ianguage modeling text generation
BLOOM(Z)	×	DeepSpeed	text generation
StarCoder	×	Single card	text generation
GPT-J	DeepSpeed	Single card DeepSpeed	language modeling text generation
GPT-NecX	DeepSpeed	 DeepSpeed 	Ianguage modeling text generation
OPT	×	DeepSpeed	text generation
Llama 2 / CodeLlama	DeepSpeed LoRA	DeepSpeed LoRA	Language modeling text generation
StableLM	×	Single card	text generation
Falcon	×	Single card	E README.md
CodeGen	×	Single card	
MPT	×	Single card	
TS	1	~	😔 + 🎎 habanc
			Optimum Habana
			Optimum Habana is the interface between the Transformers and Diffuses libraries and Habana's Gaudi processor (HPU). It provides a set of tools enabling easy model loading, training and inference on single- and mult HPU entitings for different downstrema trainist. The far of officially validated models and tasks is available here. User can try other models and tasks with only few changes.
			What is a Habana Processing Unit (HPU)?
			HUs offer fast model training and inference as well as a great price-performance ratio. Check out this blog post about BRT pre-training and this article benchmarking Habana Gaudi2 versus hirdia A100 GPUs for concrete examples. If you are not familiar with HPUs and would like to know more about them, we recommend you take a look at our conceptual guide.
			Install
			To install the latest stable release of this package:
			pip installupgrade-strategy eager optimum[habana]

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:



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

Performance Transparency on Every Model

	™ 3.1 Train	ning Performanc	;e				Gaudi2 Large	Language	s Models Infe	erence Perform	iance			Gaudi2 Reference M	odels Tr	aining Performanc	e			Gaudi Reference	e Models Ir	nference F	Performance			
		nerated with the latest version	on of SynapseAI and are	improvements	over the officially su	bmitted numbers				Max				Show 50 v entries				Search		Model	± UD	U Precisio	m Throughput		aloncy ¹¹¹ Rateb	Size Framework Version
posted on MLCommons we	vebsite.						Model	# HPU Precis	sion Input Output	Token Throughput	t Latency**	** Batch Fr	ramework Version							Bloom-1768-BeamSearch-8	16	bf16	10.51 token/		5.1 ms 1	DeepSpeed 0.12.4
							(Length				Model	# HPU Prec	sion Throughput	Accuracy Ti	ne To Train Size	Framework Version	Bloom-1768-Greedy	16	bf16	11.92 token/		3.38 ms 1	DeepSpeed 0.12.4
Model	E F	HPU Precision	Time To Train	1	Fran	nework Version	Falcon-78	1 bf16	100 8k	8k 110.7 toker	n/sec 9.03 ms		optimum Habana 1.9.0						Megatron DeepSpeed	Bloom-1768-Sampling	16	bf16	7.98 token/s		25.26 ms 1	DeepSpeed 0.12.4
MI Perf 3 1 - GPT3	38-		153.58 min*	_			Bloom-78-Greedy	1 bf16		2k 721.56 toks				DeepSpeed Chat LLaMA 7B Step1	8 bf16	870 sec/iter	ppt 1.61	8	0.12.4	Bloom-7B (512 token)	1	bf16	42.84 token/		3.34 ms 1	
MLPerf 3 1 - GPT3	25		223.75 min**				Bloom-7B-Greedy	1 fp8			en/sec 5.15 ms		eepSpeed 0.12.4,	DeepSpeed Chat LLaMA 78 Step2	8 bf16	770 sec/iter	acc: 81	4	Megatron DeepSpeed 0.12.4	Stable Diffusion v2.1 (512x51	2)	bf16	0.36 img/set	277	777.77	Lightning 2.1.2
MLPerf 3.1 - Stable Diffusion			19.4 min**		DvTr	orch Lightning 2.1.2	GPT-J	8 bf16	6 100	100 562.23 toks	en/sec 7.11 ms	4 0	ptimum Habana 1.9.0						0.12.4 Megatron DeepSpeed					rc ms	15	
MLPerf 3.1 - ResNet	8		16.22 min		170	Ch Dyning 2.1.2	LLaMA 2-7B	1 fp8	1K 3k	4k 1101.7 toks	en/sec 10.89 ms	12 Or	optimum Habana 1.9.0	DeepSpeed Chat LLaMA 78 Step3	8 bf16	7.8 sec/iter	ema: 2.7	4	0.12.4	Stable Diffusion v2.1 (768X76	i8) 1	bf16	0.13 img/set		692.3 ms 1	Lightning 2.1.2
MLPerf 3.1 - BERT	8		14.25 min				LLaMA 2-7B	1 fp8	2k 6k	8k 551.84 toks			optimum Habana 1.9.0	Stable Diffusion	64 bf16	10658.26 img/sec		32	Lightning 2.1.2	Bert	1	bf16	147.17 toker		63.12 ms 24	
MEPELI 3.1 * DENI	•	UTU	14.201111				LLaMA 2-7B	1 fp8	4k 12k	16k 273.32 toks			optimum Habana 1.9.0	Stable Diffusion Fine Tuning	1 bf16	70 img/sec		7	Lightning 2.1.2	Unet2D	1	bf16	1364.2 img/s		6.9 ms 64	Lightning 2.1.2
The 0770	with on the sector	vas taken using a pre-launch	and the first state of the first				LLaMA 2-78 Falcon-408	1 bf16 8 bf16	1k 3k 100 8k	4k 361.14 toks 8k 61.85 toker			optimum Habana 1.9.0 Optimum Habana 1.9.0	Stable Diffusion Fine Tuning Textual Inversion	1 bf16	20.58 img/sec		7	Lightning 2.1.2	Unet3D	1	bf16	52.68 img/sr	sc 37.	7.96 ms 2	Lightning 2.1.2
		nd Stable Diffusion were tak						0 0110	100 01				eepSpeed 0.12.4.	ResNet50 LARS	32 bf16	185594.14 img/sec	76.33 6.	i2 min 256								
The OP I measurement i	with 250 cards an	The Stable Diffusion were tak	an doing the synapsemin	10.0 Software	COLDUK		LLeMA 2-70B	8 fp8	24 24	4k 4910.4 toks	en/sec 56.41 ms	a 2// Op	optimum Habana 1.9.0	ResNet50 LARS	8 bf16	47104.44 img/sec		.53 min 256					10 A 4			
							LLaMA 2-70B	8 fp8	2k 6k	8k 2859.5 toks	en/sec 26.92 ms	. 77 De	eepSpeed 0.12.4, ootimum Habana 1.9.0	ResNet50 LARS	1 bf16	6072.03 img/sec		256		Hugging Face C	iptimum Ha	abana Ga	iudi Inference	Performar	nce	
Gaudi2 Large L		Models Training	Performance				(a	16k 1470.6 toks			ReepSpeed 0.12.4,	BERT Pre Training Phase 1	32 bf16	32249.28 sent/sec	1.4975064 27	3.25 min 64							_	
Jadulz Larye Li	anguage n	woucis maining	renormance				LL8MA 2-70B	8 fp8	2k 14k	16k 1470.6 toks	en/sec 25.83 ms	00 Op	optimum Habana 1.9.0	BERT Pre Training Phase 1	8 bf16	1151.95 sent/sec	77	64		See the Examples page for	information on how	v to run each of	the Tasks, including m	odel naming and b	hyperparameter v	usage.
							LLaMA 2-70B	8 fp8	2k 30k	32k 775.4 toker	n/sec 24.5 ms	19 De	eepSpeed 0.12.4, optimum Habana 1.9.0	BERT Pre Training Phase 1	1 bf16	9126.16 sent/sec		64								
			_	_		and the second s	(Do	DeepSpeed 0.12.4.	BERT Pre Training Phase 2	32 bf16	10812.96 sent/sec	1.3324460 91	21 min 16		Show 25 + entries						
Model	# HPU Precision	1 Throughput	Sequence Length TP,	PP, DP	Batch Size	Framework Version	LL8MA 2-70B	8 bf16	2k 2k	4k 3225.6 toks	en/sec 66.96 ms	s 216 Op	Optimum Habana 1.9.0				98			Madel	# HPU Precision	Throughput	Latency	Batch Task		Framework Version
11.111.100		(0.10	2.048 2.2				LLeMA 2-70B	8 bf16	2k 6k	8k 1229.1 toks	in/sec 24.4 ms		eepSpeed 0.12.4,	BERT Pre Training Phase 2	8 bf16	348.71 sent/sec		16		Model	Precision	rnrougnput	Latency	Batch Tas		ramework version
	64 bf16	68.12 samples/sec 30 samples/sec	2,048 2, 2 4,096 8.8		256	DeepSpeed 0.12.4 DeepSpeed 0.12.4							optimum Habana 1.9.0 NeepSpeed 0.12.4,	BERT Pre Training Phase 2 BERT SQUAD Fine Tuning	1 bf16 8 bf16	2779.63 sent/sec 2054.6 sent/sec	90.83 5.	16 12 min 24		BERT	1 bf16	39.55 token/se	ec 101.12 ms	4 lang	odeling (Optimum Habana 1.9.0
							LL8MA 2-70B	8 bf16	2k 14k	16k 566.9 toker		s 10 Op	optimum Habana 1.9.0	BERT SQUAD Fine Tuning	1 bf16	2034.6 sent/sec 281.49 sent/sec	90.63 5.	24							estion-	
	512 bf16	55.4 samples/sec	4,096 8, 8		2,048	DeepSpeed 0.12.4	LLeMA 2-13B	1 bf16	2K 2K	4k 125.66 toks	en/sec 15.91 ms		optimum Habana 1.9.0	ResNext101	8 bf16	22146.5 img/sec	78.03 10	2 min 256		BERT	1 bf16	126.85 token/1		e ans	iswering	Optimum Habana 1.9.0
	1,024 bf16	104.4 samples/sec			4,096	DeepSpeed 0.12.4	Bloomz-1768	8 bf16	6 100	100 36.36 toker	n/sec 27.5 ms		leepSpeed 0.12.4, optimum Habana 1.9.0	ResNext101	1 bf16	2841.31 img/sec		256		BERT	1 bf16	107.76 token/s		8 text		Optimum Habana 1.9.0
Bloom-13B 6	64 bf16	72.5 samples/sec	2,048 2, 2	2, 16	1,024	DeepSpeed 0.12.4	Bloom-1768-Greedy	8 fnR		4K 199.39 toks	en/sec 40.12 ms		eepSpeed 0.12.4	SSD	8 bf16	16555.1 img/sec	22.95 9.	i8 min 128		BART-Greedy	1 bf16	2.96 token/sec				Optimum Habana 1.9.0
							Bloom-1768-Greedy	8 bf16		4K 394.47 toks			eepSpeed 0.12.4	SSD	1 bf16	2098.21 img/sec		128		ESMFold	1 bf16	14.17 token/se	ec 70.54 ms	1 prof	otein-folding (Optimum Habana 1.9.0
P, PP, DP = These are the	Tensor Parallel, Pi	Pipeline Parallel and Data Pa	rallel parameters for the	Megatron Deep	pSpeed training		Bloom-1768-Greedy	8 bf16		8K 196.21 toks	en/sec 50.96 ms		eepSpeed 0.12.4	Transformer	8 bf16	1074849 token/sec	27.9 24	2.05 min 8192		Stable Diffusion v2.1	1 bf16	0.35 token/sec	c 11173.18 ms		xt to image	Optimum Habana 1.9.0
							Bloom-1768-Greedy	8 bf16		16K 80.79 toker	n/sec 49.51 ms		leepSpeed 0.12.4	Transformer	1 bf16	133257.66 token/sec	70.44	8192	1	image size 512x512 T5-Small Translation					eneration	
Jugging Ease C	Ontinum I	Habana Gaudi2	Informon Der	formore	~		Bloom-1768-Greedy	8 bf16		32K 25.81 toker			eepSpeed 0.12.4	Unet2D Unet2D	8 bf16 1 bf16	16999.22 img/sec 2697.3 img/sec	72.61 13	.92 min 64	Lightning 2.1.2 Lightning 2.1.2	Greedy	1 bf16	15.48 token/se	ec 258.36 ms	4 tran	anslation C	Optimum Habana 1.9.0
rugging race u	phannan F		anerence Peri	ormance	e		Bloom-1768-Sampling Bloom-1768-	8 bf16		1k 19.59 toker			eepSpeed 0.12.4	Unet2D Unet3D	1 bf16 8 bf16	2697.3 img/sec 256.76 img/sec	74.26 19	77 min 2	Lightning 2.1.2 Lightning 2.1.2		1 bf16	529.7 token/se	ec 7.68 ms	4 spe	eech-	Optimum Habana 1.9.0
							Bloom-1768- BeamSearch-8	8 bf16		512 30.57 toker	n/sec 32.7 ms	1 DF	eepSpeed 0.12.4	Unet3D	1 bf16	32.69 img/sec	13	2	Lightning 2.1.2		0016	529.7 token/se	ec v.odims	4 reor	cognition	Optimum Habaria 1.9.0
e the Examples page for	information on he	how to run each of the Task:	a, including model namin	ig and hyperpa	arameter usage.									L				-		Wav2Vec 2.0 Speech Classification	1 bf16	9.39 token/set	c 425.62 ms	4 spei	cognition	Optimum Habana 1.9.0
how 25 v entries							Hugging Face	9 Optimum	i Habana Ga	udi2 Training I	erforman	се		Gaudi Reference Mo	dels Trair	ning Performance								reco	ognidul	
Model #	HPU Precision	Max Token Throu	ughput Latency	Batch	task	Framework Version	See the Examples page	for information o	in how to run each of 1	the Tasks, including mod	el naming and hy	perparameter usage	e.													
							Show 25 - entries	s				Search:		Show 25 - entries				Search:								
StableDiffusion v2.1 ((512x512) 1	bf16	1.24	es/sec 3223.2 ms	4	stable-diffusion	PyTorch Lightning 2.1.2												Time To Batch								
						DeepSpeed	Model	A HELL Brook	sion Throughout	Accuracy Time To	Bato h Tas		Framework Version	Modul	# HPU	Precision Throughput	Accuracy	Time To Batch Train Size	Framework Version							
OPT 1	bf16	980.9 token		1	text-generation	0.12.4, Optimum		سر اندر ا		Train	Size			ResNet50 Keras LARS	32	bf16 48176.47 img/sec	75.86	19.66 min 256								
		token	/360			Habana 1.9.0	Liama2.708 Eine Tuning	8 bf16	2.43 sentences/s	Gec 2.12 49.21	min 10 lang	quage-modeling		ResNet50 Keras LARS	8	bf16 12305.57 img/sec	76.16	69.86 min 256								
						Habana 1.9.0	a monactive rine funning		2.45 senterices/s		to tang	yougo modeling														
StarCoder 1	bf16					DeepSpeed	Llama2-708 Fine Tuning (LoRA)						Optimum Habana 1.9.0	ResNet50 Keras LARS	1	bf16 1624.15 img/sec		256								
		65.51	token/sec 15.26 ms	1	text-generation	DeepSpeed 0.12.4, Optimum	Liama2-708 Fine Tuning (LoRA) Liama1-78 Fine Tuning (LoRA)		143.02 sentences/sec	0.93 5.5 min	n 64 lang	iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0		1	bf16 1624.15 img/sec bf16 4805.33 sent/sec		256 1803.26 64								
		105.2	20	1		DeepSpeed 0.12.4, Optimum Habana 1.9.0	Liama1-7B Fine Tuning (LoRA)	8 bf16	sentences/sec	0.93 5.5 min			Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine	1 32	bf16 4805.33 sent/sec		256 1803.26 min 64								
MPT-7B 1		105.3	38 0.40	1	text-generation text-generation	DeepSpeed 0.12.4, Optimum	(LORA) Liama1-78 Fine Tuning (LoRA) Falcon-1808 Fine Tuning (LoRA)	8 bf16 8 bf16	sentences/sec 1.53 sentences/s	0.93 5.5 min		iguage-modeling		ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine	1 32 8	bf16 4805.33 sent/sec bf16 1221.73 sent/sec										
Bert (Text	bf16	1932 105.3 token 186.2	38 9.48 ms n/sec 40.04 ms		text-generation	DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana	Liama1-7B Fine Tuning (LoRA)	8 bf16 8 bf16	sentences/sec 1.53 sentences/s	0.93 5.5 min	min 1 lang		Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training combine	8 1	bf16 4805.33 sent/sec bf16 1221.73 sent/sec bf16 153.34 sent/sec	Loss									
Bert (Text 1	bf16	1932 105.3 token	38 9.48 ms n/sec 40.04 ms	1		DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	(LORA) Liama1-7B Fine Tuning (LORA) Falcon-180B Fine Tuning (LORA) Falcon-40B Fine Tuning (LORA)	8 bf16 8 bf16 8 bf16 8 bf16	sentences/sec 1.53 sentences/s	0.93 5.5 min sec 1.05 270.13 1.19 16.8 mi	min 1 lang	iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training combine BERT Pre Training Phase 1	1 32 8 1 32	bf16 4805.33 sent/sec bf16 1221.73 sent/sec bf16 153.34 sent/sec bf16 5757.69 sent/sec	Loss: 1.49	min 64 64 64								
tert (Text 1 Classification)	bf16 bf16	1932 105.3 token 186.2 token	38 9.48 ms n/sec 42.94 ms		text-generation text-classification language-	DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana	(LORA) Liama1-78 Fine Tuning (LoRA) Falcon-1808 Fine Tuning (LoRA)	8 bf16 8 bf16	sentences/sec 1.53 sentences/s 28.83 sentences/sec 16.64 sentences/sec	0.93 5.5 min sec 1.05 270.13	min 1 lang	iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training combine BERT Pre Training Phase 1 BERT Pre Training Phase 1	8 1	bf16 4805.33 sent/sec bf16 1221.73 sent/sec bf16 153.34 sent/sec bf16 5757.69 sent/sec bf16 1467.43 sent/sec	Loss: 1.49	min 64 64 1348.41 64								
Bert (Text Classification) 1 Bert (Language Modeling) 1	bf16 bf16	1932 105.3 token 186.2 token 80.72	38 9.48 ms 7/sec 42.94 ms 2 token/sec 49.55 ms		text-generation text-classification language- modeling	DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	(LORA) Liama1-7B Fine Tuning (LORA) Falcon-180B Fine Tuning (LORA) Falcon-40B Fine Tuning (LORA)	8 bf16 8 bf16 8 bf16 8 bf16	28.83 sentences/sec 28.83 sentences/sec 16.64	0.93 5.5 min sec 1.05 270.13 1.19 16.8 mi 0.53 14.13 n	imin 1 lang in 1 lang min 4 lang	iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training combine BERT Pre Training Phase 1	8 1	bf16 4805.33 sent/sec bf16 1221.73 sent/sec bf16 153.34 sent/sec bf16 5757.69 sent/sec		min 64 64 1348.41 min 64 64 64								
Bert (Text 1 Classification) 1 Bert (Language 1 Modeling) 1 Bert (Question 1	bf16 bf16	1932 105.3 1932 105.8 188.2 105.8 100.8 100.8 100.8 100.8 100.8 100.8 100.8 100.8 100.8 100.8 100.8 10	38 9.48 ms 26 42.94 ms ztoken/sec 49.55 ms 54 13.24 ms		text-generation text-classification language- modeling question-	DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana	Liama1-78 Fine Tuning (LoRA) Falcon-1808 Fine Tuning (LoRA) Falcon-408 Fine Tuning (LoRA) GPTJ-CLM GPTJ-CLM	 8 b116 	sentences/sec 1.53 sentences/s 28.83 sentences/sec 16.64 sentences/sec 290.8	0.93 5.5 min sec 1.05 270.13 1.19 16.8 mi 0.53 14.13 n 0.17 28.75 n	min 1 lang nin 1 lang nin 4 lang nin 2 lang	iguage-modeling iguage-modeling iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training combine BERT Pre Training Phase 1 BERT Pre Training Phase 1	8 1	bf16 4805.33 sent/sec bf16 1221.73 sent/sec bf16 153.34 sent/sec bf16 5757.69 sent/sec bf16 1467.43 sent/sec	Loss:	min 64 64 1348.41 64 64 64 454.85								
Bert (Text Classification) 1 Bert (Language 1 Modeling) 1 Bert (Question 1 Answering) 1	bf16 bf16 bf16 bf16 bf16	1932 105.3 token 186.2 token 80.72 599.5 token	38 9.48 ms 26 42.94 ms 21 oken/sec 49.55 ms 54 13.34 ms	8	text-generation text-classification language- modeling question- answering	DeepSpeed 0.12.4.0ptimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	(LORA) Liama 1-78 Fine Tuning (LORA) Falcon-1808 Fine Tuning (LORA) Falcon-408 Fine Tuning (LORA) GPTJ-CLM	8 bf16 8 bf16 8 bf16 8 bf16 8 bf16	sentences/sec 1.53 sentences/sec 28.83 sentences/sec 16.64 sentences/sec 290.8 sentences/sec 473.05 sentences/sec	0.93 5.5 min sec 1.05 270.13 1.19 16.8 mi 0.53 14.13 n	min 1 lang nin 1 lang nin 4 lang nin 2 lang	iguage-modeling iguage-modeling iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNetSO Karas LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training Combine BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 2	8 1 32 8 1	bf16 4805.33 sent/sec bf16 1221.73 sent/sec bf16 153.34 sent/sec bf16 5757.69 sent/sec bf16 1467.43 sent/sec bf16 1467.43 sent/sec bf16 1941.84 sent/sec bf16 1941.84 sent/sec		min 64 64 1348.41 min 64 64 64								
Bert (Text Classification) 1 Bert (Language Modeling) 1 Bert (Question Answering) 1	bf16 bf16 bf16 bf16 bf16	1932 105.3 token 186.2 token 80.72 599.5 token	38 9.48 ms 26 42.94 ms ztoken/sec 49.55 ms 54 13.94 ms		text-generation text-classification language- modeling question-	DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana	Liama1-78 Fine Tuning (LoRA) Falcon-1808 Fine Tuning (LoRA) Falcon-408 Fine Tuning (LoRA) GPTJ-CLM GPTJ-CLM	 8 b116 	sentences/sec 1.53 sentences/s 28.83 sentences/sec 16.64 sentences/sec 290.8 sentences/sec	0.93 5.5 min sec 1.05 270.13 1.19 16.8 mi 0.53 14.13 n 0.17 28.75 n	min 1 lang nin 1 lang min 4 lang nin 2 lang n 40 con	iguage-modeling iguage-modeling iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training combine BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 1	8 1 32 8 1	bf16 4805.33 sent/sec bf16 1221.73 sent/sec bf16 1231.73 sent/sec bf16 153.34 sent/sec bf16 5757.69 sent/sec bf16 1467.43 sent/sec bf16 1467.43 sent/sec bf16 1467.43 sent/sec	Loss:	min 64 64 1348.41 64 64 64 454.85								
Bert (Text Jassification) 1 Bert (Language Vodeling) 3ert (Question 1 Howsening) 1	bf16 bf16 bf16 bf16 bf16 bf16	1932 105.3 token 186.2 token 80.72 599.5 token	38 9.48 ms 26 92.94 ms 26 42.94 ms 2 token/sec 49.55 ms 54 13.34 ms coken/sec 614.91 ms	8 4 8 4	text-generation text-classification language- modeling question- answering language- modeling constrastive-	DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana	LORA) Liama 17 B Fine Tuning (LORA) Falcon-160B Fine Tuning (LORA) GPTJ-CLM GPTJ-CLM BridgeTower GPT2	8 616 8 616 8 616 8 616 8 616 8 616 8 616 8 616 8 616 8 616 9 616 9 616 9 616	sentences/sec 1.53 sentences/sec 28.83 sentences/sec 16.64 sentences/sec 290.8 sentences/sec 473.05 sentences/sec 575.41 sentences/sec	0.93 5.5 min sec 1.05 270.13 1.19 16.8 mi 0.53 14.13 n 0.17 28.75 n	min 1 lang in 1 lang min 4 lang nin 2 lang n 40 con 4 lang	iguage-modeling iguage-modeling iguage-modeling iguage-modeling intrastive-image-text iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Karas LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 2 BERT Pre Training Phase 2	8 1 32 8 1	bf16 4905.33 sent/sec bf16 1221.73 sent/sec bf16 153.34 sent/sec bf16 5757.69 sent/sec bf16 1467.43 sent/sec bf16 1467.43 sent/sec bf16 1467.43 sent/sec bf16 1467.43 sent/sec bf16 1918.48 sent/sec bf16 1911.84 sent/sec bf16 1911.84 sent/sec	Loss: 1.33	min 64 64 1348.41 64 64 64 454.85								
Bert (Text Classification) 1 Bert (Language Modeling) 1 Bert (Question Answering) 1 Bart 1	bf16 bf16 bf16 bf16 bf16 bf16	1932 105.3 104.4 186.2 1	38 9.48 ms 7/sec 9.48 ms 25 42.94 ms 2 token/sec 49.55 ms 54 13.34 ms ocken/sec 614.91 ms 65 46.52 mc	8	text-generation text-classification language- modeling question- answering language- modeling	DeepSpeed 0.12.4, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	Liama 1-78 Fine Tuning (LoRA) Falcon-1808 Fine Tuning (LoRA) Falcon-488 Fine Tuning (LoRA) GPTJ-CLM OPTNEOX-208-CLM BridgeTower	8 b116	sentences/sec 1.53 sentences/s 28.83 sentences/sec 10.64 sentences/sec 290.8 sentences/sec 473.05 sentences/sec 575.41 sentences/sec 86.59 sentences/sec	0.93 5.5 min sec 1.05 270.13 1.19 16.8 mi 0.53 14.13 n 0.17 28.75 n	min 1 lang in 1 lang min 4 lang nin 2 lang n 40 con 4 lang	iguage-modeling iguage-modeling iguage-modeling iguage-modeling intrastive-image-text	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Keras LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training combine BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 2 BERT Pre Training Phase 2 BERT Pre Training Phase 2	8 1 32 8 1	b116 4805.33 set//sec b116 1221.73 set//sec b116 1221.73 set//sec b116 153.34 set//sec b116 5757.69 set//sec b116 1407.43 set//sec b116 148.41 set//sec b116 1911.84 set//sec b116 1911.84 set//sec b116 50.55 set//sec	Loss: 1.33	min 64 64 1348.41 64 64 64 64 454.85 8 8 8 8								
And the set of the set	br16 br16 br16 br16 br16 br16	1932 105 5 1932 106 m 1862 80 72 80 72 599 5 106 m 6.5 fc 50 m 6.5 fc 50 m	38 9.48 ms 7/sec 9.48 ms 25 42.94 ms 2 token/sec 49.55 ms 54 13.34 ms ocken/sec 614.91 ms 65 46.52 mc	8 4 8 4	text-generation text-classification language- modeling question- answering language- modeling constrastive-	DeepSpreed 0.124, Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana	LORA) Liama 17 B Fine Tuning (LORA) Falcon-160B Fine Tuning (LORA) GPTJ-CLM GPTJ-CLM BridgeTower GPT2	8 616 8 616 8 616 8 616 8 616 8 616 8 616 8 616 8 616 8 616 9 616 9 616 9 616	sentences/sec 1.53 sentences/sec 28.83 sentences/sec 10.64 sentences/sec 290.8 sentences/sec 473.05 sentences/sec 575.41 sentences/sec	0.93 S.Smin BEC 1.05 270.13 1.19 16.8 mil 0.53 0.53 14.13 m 0.17 0.17 28.75 m 7.4 min 0.1 - - 0.1 - -	min 1 lang in 1 lang min 4 lang nin 2 lang n 40 con 4 lang	iguage-modeling iguage-modeling iguage-modeling iguage-modeling intrastive-image-text iguage-modeling iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Karas LARS BERT Pre Taining combine BERT Pre Taining combine BERT Pre Taining combine BERT Pre Taining Phase 1 BERT Pre Taining Phase 1 BERT Pre Taining Phase 2 BERT Pre Taining Phase 2 BERT Pre Taining Phase 2 BERT Pre Taining Phase 2	8 1 32 8 1	b116 4805.33 sert/sec b116 1221.73 sert/sec b116 1221.73 sert/sec b116 5757.69 sert/sec b116 1467.43 sert/sec b116 1467.43 sert/sec b116 1487.81 sert/sec b116 1481.81 sert/sec b116 50.55 sert/sec b116 60.56 sert/sec b116 442.51 sert/sec b116 60.46 sert/sec b116 60.46 sert/sec	Loss: 1.33	min 64 64 1348.41 64 64 64 64 454.85 8 8 8 8								
art (Text Zlassification) 1 Jert (Language Modeling) 1 Jert (Question 1 Inswering) 1 Sart 1 BridgeTower 1 ESMFold 1	bf16 bf16 bf16 bf16 bf16 bf16 bf16 bf16	1932 105 5 1932 106 8 186 2 80 72 90 90 7 10 4 10 4 10 4 10 4 10 4 10 4 10 4 10 4	38 9.48 ms 1/sec 9.48 ms 26 42.94 ms 21 token/sec 49.55 ms 54 13.34 ms oken/sec 614.91 ms token/sec 48.53 ms token/sec 272.47 ms	8 4 8 4 4 16	text-generation text-classification language- modeling question- answering language- modeling constrastive- image-text protein-folding	DeepSpreed 0.124.0ptimum Habana 15.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	LORA) Liama 17 B Fine Tuning LorAA) Priscos 1086 Fine Tuning LorAA GPTJ-CLM OPTNEOX-208-CLM BridgeTower OPT2 GPT2-3L ALBERTLange	8 bf16	sentences/sec 1.33 sentences/sec 28.83 sentences/sec 16.64 sentences/sec 290.8 sentences/sec 473.05 sentences/sec 575.41 sentences/sec 86.59 sentences/sec 280.45	0.93 S.Smin tec 1.05 27013 1.19 3.68 min 0.53 1.413 m 0.17 28.75 m 1.19 1.413 m 1.19 3.67 min 1.19 1.19 1.19 3.74 min 1.19 1.19 1.19 1.19 1.19 1.19	min 1 lang in 1 lang min 4 lang min 2 lang n 40 con 4 lang 4 lang in 32 que	guage-modeling iguage-modeling iguage-modeling iguage-modeling intrastive-image-text iguage-modeling iguage-modeling estilon-answering	Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Karas LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training Combine BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 2 BERT Pre Training Phase 3 BERT Pre Training Phase 4 BERT Phase	8 1 32 8 1	b116 4005.33 sent/sec b116 1221.73 sent/sec b116 153.34 sent/sec b116 153.34 sent/sec b116 159.76.69 sent/sec b116 167.43 sent/sec b116 194.74 sent/sec b116 194.84 sent/sec b116 191.84 sent/sec b116 50.56 sent/sec b116 60.56 sent/sec b116 60.56 sent/sec b116 60.56 sent/sec b116 60.56 sent/sec	Loss: 1.33	min 64 64 1348.41 64 64 64 64 454.85 8 8 8 8								
Bert (Text Classification) 1 Bert (Language Modeling) 1 Bert (Uestion Answering) 1 BridgeTower 1 BirdgeTower 1 ESMFold 1	bf16 bf16 bf16 bf16 bf16 bf16 bf16 bf16	1932 105 3 1932 105 3 1962 105 4 1864 0 00.72 990 9 105 4 105 4 105 4 105 4 105 4 105 4 105 4 105 5 105 5 1	28 9.48 ms 1/sec 9.48 ms 26 42.94 ms 21 token/sec 49.55 ms 54 13.34 ms token/sec 614.91 ms 65 48.53 ms token/sec 272.47 ms 34 4.3 ms	8 4 8 4 4 16	text-generation text-classification language- modeling question- answering language- modeling constrastive- image-text	DeepGpeed 0.124. Aptimum Habana 1.30 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana	LIAMA THE Flucture OFTL CMA OFTL CMA OFTL CMA OFTL CMA ALBETT-LIAP ALBETT-SOL	8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 9 8 8	sentences/sec 1.53 sentences/sec 28.83 sentences/sec 16.64 sentences/sec 290.8 sentences/sec 290.8 sentences/sec 473.05 575.41 sentences/sec 575.41 sentences/sec 280.45 sentences/sec	0.93 S.Smin BEC 1.05 270.13 1.19 16.8 mil 0.53 0.53 14.13 m 0.17 0.17 28.75 m 7.4 min 0.1 - - 0.1 - -	min 1 lang in 1 lang min 4 lang min 2 lang n 40 con 4 lang 4 lang in 32 que	iguage-modeling iguage-modeling iguage-modeling iguage-modeling intrastive-image-text iguage-modeling iguage-modeling	Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Karas LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training Combine BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 2 BERT Pre Training Phase 3 BERT Pre Training Phase 4 BERT Phase	8 1 32 8 1	b116 4005.33 sert/sec b116 1221.73 sert/sec b116 123.74 sert/sec b116 575.69 sert/sec b116 575.69 sert/sec b116 134.4 sert/sec b116 134.743 sert/sec b116 134.18 sert/sec b116 131.144 sert/sec b116 0.50 sert/sec b116 0.51 sert/sec b116 50.51 sert/sec b116 50.51 sert/sec b116 50.51 sert/sec	Loss: 1.33	mn 64 64 1348.41 64 64 64 64 64 64 64 8 8 13.08 mn 24 24 2280.8 64								
art (Text Jassification) 1 14604eling) 1 14604eling) 1 1 14604eling) 1 1 14604eling) 1 1 14604eling) 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	br16 br16 br16 br16 br16 br16 br16 br16	1932 105.3 1932 105.4 184.6 184.6 184.6 184.6 184.7 18	38 9.48 ms 1/13ec 9.48 ms 26 42.94 ms 210ken/sec 42.95 ms 210ken/sec 49.55 ms 54 13.34 ms 60ken/sec 614.91 ms 65 48.53 ms 10ken/sec 272.47 ms 10ken/sec 4.3 ms	8 4 8 4 4 16	text-generation text-classification language- modeling question- answering language- modeling construstive- image-text protein-folding text-generation	DeepSpeed 0.124.0ptimum Habana 1.50 Optimum Habana 1.90 Optimum Habana 1.90 Optimum Habana 1.90 Optimum Habana 1.90 Optimum Habana 1.90 Optimum Habana 1.90 Optimum Habana 1.90	LORA) Liama 17 B Fine Tuning LorAA) Priscos 1086 Fine Tuning LorAA GPTJ-CLM OPTNEOX-208-CLM BridgeTower OPT2 GPT2-3L ALBERTLange	8 bf16	sentences/sec 1.53 sentences/se 28.83 sentences/sec 16.64 sentences/sec 200.8 sentences/sec 473.05 sentences/sec 575.41 sentences/sec 28.59 sentences/sec 3103 sentences/sec sentences/sec sentences/sec	0.93 S.Smin tec 1.05 27013 1.19 3.68 min 0.53 1.413 m 0.17 28.75 m 1.19 1.413 m 1.19 3.67 min 1.19 1.19 1.19 3.74 min 1.19 1.19 1.19 1.19 1.19 1.19	min 1 lang in 1 lang min 4 lang min 2 lang n 40 con 4 lang 4 lang in 32 que in 12 que	guage-modeling iguage-modeling iguage-modeling iguage-modeling intrastive-image-text iguage-modeling iguage-modeling estilon-answering	Optimum Habana 1.9.0 Optimum Habana 1.9.0	ResNet50 Karas LARS BERT Pre Training combine BERT Pre Training combine BERT Pre Training Combine BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 1 BERT Pre Training Phase 2 BERT Pre Training Phase 3 BERT Pre Training Phase 4 BERT Phase	8 1 32 8 1	b116 4005.33 sert/sec b116 1221.73 sert/sec b116 123.74 sert/sec b116 575.69 sert/sec b116 575.69 sert/sec b116 134.4 sert/sec b116 134.743 sert/sec b116 134.18 sert/sec b116 131.144 sert/sec b116 0.50 sert/sec b116 0.51 sert/sec b116 50.51 sert/sec b116 50.51 sert/sec b116 50.51 sert/sec	Loss: 1.33	mn 64 64 1348.41 64 1348.41 64 64 64 64 64 8 8 13.08 min 24 13.08 min 24 24 32								
art (Text Jassification) 1 14604eling) 1 14604eling) 1 1 14604eling) 1 1 14604eling) 1 1 14604eling) 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 14604eling 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	bf16 bf16 bf16 bf16 bf16 bf16 bf16 bf16	1932 105.3 1932 105.4 166.2 164.4 80.72 599.9 164.4 6.5 to 25.9 25.4 25.4 25.4 25.4 25.4 25.4 25.4 25.4	B 9.48 ms Nracc 9.48 ms 20 42.94 ms 21 42.94 ms 2 token/sec 49.55 ms 54 3.34 ms ochen/sec 61.91 ms Nrisec 61.91 ms token/sec 272.47 ms 34 43.ms	8 4 8 4 4 16	text-generation text-classification language- modeling question- answering language- modeling constrastive- image-text protein-folding	DeepGpeed 0.124. Aptimum Habana 1.30 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana 1.90 Opfimum Habana	LIAMA THE Flucture OFTL CMA OFTL CMA OFTL CMA OFTL CMA ALBETT-LIAP ALBETT-SOL	8 bf16	sentences/sec 1.53 sentences/se 28.83 sentences/sec 290.8 sentences/sec 273.41 sentences/sec 280.45 sentences/sec 280.45 sentences/sec	0.93 S.Smin 1.05 27013 1.19 16.8mi 0.53 14.13 n 0.17 20.5n 0.17 20.5n 0.17 20.5n 0.17 20.5n 0.17 20.5n 0.17 20.5n 0.18 1.19 0.19 1.19	Imm Imm Imm 1 Im	guisge-modeling guisge-modeling guisge-modeling guisge-modeling ntrastive-image-text iguisge-modeling guisge-modeling estion-answering estion-answering	Optimum Habana 1.0.	Restance Draws LAMS BERT Pro Training combine BERT Pro Taining combine BERT Pro Taining combine BERT Pro Taining Press BERT Pro Taining P	8 1 32 8 1	th16 4005.33 ent/bits th16 122.73 ent/bits th16 152.73 ent/bits th16 152.34 ent/bits th16 157.55 ent/bits th16 157.55 ent/bits th16 147.43 ent/bits th16 147.43 ent/bits th16 147.43 ent/bits th16 147.43 ent/bits th16 42.53 sent/bits th16 42.53 sent/bits th16 53.53 ent/bits th16 13.43 ent/bits th16 53.53 ent/bits th16 13.53 ent/bits th16 13.53 ent/bits th16 13.53 ent/bits th17 53.58 ent/bits th16 13.53 ent/bits	Loss: 1.33 90.68 77 71.21	mn 64 64 1346.41 64 454.85 8 8 8 13.08 mi 24 24 2200.8 min 64 531.06 64								
and ref (reat Datasification) 1 bert (Language 1 bert (anguage 1 bert (anguage 1 ange rowering) 1 ange rower 1 ange rower 1 ange rower 1 ange rower 1 analytic (Anguage) 1	bf16 bf16 bf16 bf16 bf16 bf16 bf16 bf16	1922 105.5 1922 105.5 1960 106.0 1960 106.0 1999 5 1999 5 999 5 999 5 999 5 999 5 999 5 999 5 1999 106.0 1999 106.0 1000000000000000	38 0 11/126 42.94 ms] 21/126 42.94 ms] 21/126 49.55 ms] 21/126 13.34 ms] 04/04/126 61.94 ms] 15/126 48.53 ms] 15/126 272.47 ms] 15/126 4.3 ms] 15/126 6.12 ms]	8 4 8 4 16 1 1 1 1 1	text-generation text-classification language- modeling question- answering language- modeling constrastive- image-text protein-folding text-generation text-generation	DeepGpeed 0.124,0ptmum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0 Optimum Habana	LLANA, 178 Fine Tuning LLANA, 178 Fine Tuning LLANA, 178 Fine Tuning LLANA, 108 Fine Tuning OFT2 XLLANA, 108 Fine Tuning OFT2 XLLANA, 108 Fine Tuning ALBERT Sale	8 bf16	sentences/sec 1.53 sentences/se 28.83 sentences/sec 16.64 sentences/sec 200.8 sentences/sec 473.05 sentences/sec 575.41 sentences/sec 28.59 sentences/sec 3103 sentences/sec sentences/sec sentences/sec	0.93 S.Smin eet 1.55 270.13 1.19 16.8 min 5.8 min 0.53 1.4 13 min 5.8 min 0.17 28.75 min 28.75 min 1 0.17 20.11 min 1 0.17 20.11 min 1 0.14 0.01 min 1 0.14 0.01 min 1 0.14 0.01 min 1 0.14 0.01 min 1 0.14 0.11 min	Imm 1 Iang ini 1 Iang min 4 Iang min 2 Iang min 2 Iang min 2 Iang min 2 Iang min 4 Iang ini 32 Que ini 32 Que ini 24 Que ini 24 Que	guage-modeling inguage-modeling inguage-modeling inguage-modeling intrastive-image-text inguage-modeling iguage-modeling iguage-modeling estion-answering estion-answering	Optimum Habana 1.9.	Restartion Journey LANS BERT Pre Training combine BERT Pre Training combine BERT Pre Training Combine BERT Pre Training Press 1 BERT Pre Training Press 2 BERT Pre Training Pr	8 1 32 8 1	b16 40533 sent/sec b16 121.73 anticles b17 121.73 anticles b18 527.95 sent/sec b16 133.44 sent/sec b16 143.41 sent/sec b16 144.73 sent/sec b16 144.14 sent/sec b16 144.14 sent/sec b16 44.64 sent/sec b16 43.64 sent/sec b16 43.44 sent/sec b16 43.44 sent/sec b16 43.44 sent/sec b16 43.46 sent/sec b16 53.84 reline b16 54.44 sent/sec b17 54.54 sent/sec b18 54.54 sent/sec b	L088: 1.33 90.68 77	mn 64 64 1348.41 64 454.85 8 64 64 64 8 8 8 13.08 min 24 22 2200.8 64 531.06 256 435.41 28								
Interf (Text Classification) 1 Interf (Text Modelino) 1 Interf (Linguage Interf (Destribution) 1 Interf (Destribution) 1	br16 br16 br16 br16 br16 br16 br16 br16	1922 105.5 1922 105.5 1960 106.0 1960 106.0 1999 5 1999 5 999 5 999 5 999 5 999 5 999 5 999 5 1999 106.0 1999 106.0 1000000000000000	38 9.48 ms h/tac 9.48 ms star 42.94 ms 2 token/sec 49.55 ms 54 13.34 ms coken/sec 61.491 ms 55 48.53 ms token/sec 27.47 ms 34 43.38 ms 14/ysec 43.98 ms	8 4 8 4 16 1 1 1 1 1	text-generation text-classification language- modeling question- answering language- modeling construstive- image-text protein-folding text-generation	Desployed D12.4, Optimum Habana 19.0 Optimum Habana 1.90	LLARA, 12 Fine Tuning LLARA, 12 Fine Tuning (LLARA), 12 Fine Tuning (LLARA), 100 Fine Tuning (LLARA), 100 Fine Tuning (LLARA), 100 Fine Tuning (LLARA), 100 Fine Tuning OFT2 CLAM OFT2 CLAM OFT2 CLAM OFT2 CLAM OFT2 CLAM ALBERT Large Fine Tuning CluphoBERTs	Present Present 2 8 Present 8 Present Present 9 Present Present	entences/sec 1.53 sentences/s 28.83 sentences/sec 16.64 solo sentences/sec 577.41 sentences/sec 577.41 sentences/sec 200.43 sentences/sec 30.03 sentences/sec 31.	0.93 5.5 min 105 2.013 110 5.6 min 0.33 1.4 min 0.7 2.05 min 110 5.7 min 110 1.7 min 110 1.7 min 110 1.8 min 111 1.9 min	mmin 1 Iang min 1 Iang min 4 Iang min 2 Iang min 2 Iang data 400 con 4 Iang in 32 que in 32 que in 24 que in 24 que in 24 que	guage-modeling guage-modeling guage-modeling guage-modeling guage-modeling mtrastive-image-text guage-modeling estion-answering estion-answering estion-answering mtrastive-image-text	Optimum Habina 1.0. Optimum Habina 1.0.	Restance Draws LAMS BERT Pro Training combine BERT Pro Taining combine BERT Pro Taining combine BERT Pro Taining Press BERT Pro Taining P	8 1 32 8 1	th16 4005.33 ent/bits th16 122.73 ent/bits th16 152.73 ent/bits th16 152.34 ent/bits th16 157.55 ent/bits th16 157.55 ent/bits th16 147.43 ent/bits th16 147.43 ent/bits th16 147.43 ent/bits th16 147.43 ent/bits th16 42.53 sent/bits th16 42.53 sent/bits th16 53.53 ent/bits th16 13.43 ent/bits th16 53.53 ent/bits th16 13.53 ent/bits th16 13.53 ent/bits th16 13.53 ent/bits th17 53.58 ent/bits th16 13.53 ent/bits	Loss: 1.33 90.68 77 71.21	mn 64 64 1348.41 64 454.85 8 8 13.08 mi 24 24 2280.8 64 531.06 45 531.06 128								
Image: Provide and the set of th	bf16 bf16 bf16 bf16 bf16 bf16 bf16 bf16	1992 1055 1068 1068 1068 107 107 107 107 107 107 107 107 107 107	38 0 11/126 42.94 ms] 21/126 42.94 ms] 21/126 49.55 ms] 21/126 13.34 ms] 04/04/126 61.91 ms 51/126 48.53 ms] 101/126 272.47 ms 31/146 4.3 ms] 101/126 61.22 ms] 201/126 8.12 ms]	8 4 8 4 16 1 1 1 1 1	text-generation text-classification language- modeling question- answering language- modeling constrastive- image-text protein-folding text-generation text-generation	Desployed 012.4. Optimum Habana 19.0 Optimum Habana 19.0 Optimum Habana 19.0	LLang, 12 First Turing LLang, 12 First Turing (LLAN, 12 First Turing 14 Fir	8 bT16	sentences/sec 1.53 sentences/s 28.83 sentences/sec 16.64 sentences/sec 20.8 sentences/sec 20.8 473.05 80.59 sentences/sec 20.9 80.59 sentences/sec 3100 sentences/sec 3100 sentences/sec	0.93 S.Smin eet 1.55 270.13 1.19 16.8 min 5.8 min 0.53 1.4 13 min 5.8 min 0.17 28.75 min 28.75 min 1 0.17 20.11 min 1 0.17 20.11 min 1 0.14 0.01 min 1 0.14 0.11 min	mmin 1 Iang min 1 Iang min 4 Iang min 2 Iang min 2 Iang data 400 con 4 Iang in 32 que in 32 que in 24 que in 24 que in 24 que	guisge-modeling guisge-modeling guisge-modeling guisge-modeling guisge-modeling guisge-modeling guisge-modeling estion-answering estion-answering estion-answering estion-answering	Optimum Habina 1.0 Optimum Habina 1.9.0 Optimum Hab	Restance Draws LAMS BERT Pro Training combine BERT Pro Taining combine BERT Pro Taining combine BERT Pro Taining Press BERT Pro Taining P	8 1 32 8 1	b16 40533 sent/sec b16 121.73 anticles b17 121.73 anticles b18 527.95 sent/sec b16 133.44 sent/sec b16 143.41 sent/sec b16 144.73 sent/sec b16 144.14 sent/sec b16 144.14 sent/sec b16 44.64 sent/sec b16 43.64 sent/sec b16 43.44 sent/sec b16 43.44 sent/sec b16 43.44 sent/sec b16 43.46 sent/sec b16 53.84 reline b16 54.44 sent/sec b17 54.54 sent/sec b18 54.54 sent/sec b	90.68 77 71.21 78.56	mn 64 64 1348.41 64 454.85 8 64 64 64 8 8 8 13.08 min 24 22 2200.8 64 531.06 256 435.41 28								
interfeat interfeat Classification) 1 Modeling) 1 Modeling) 1 Bard 1 BindgeTower 1 StableLM-38 1 StableLM-38 1 StableLM-38 1 StableLM-38 1 T5-38 Summarization 1 T5-39 Green 1 T5-39 Green 1 T5-39 Green 1	br16 br16 br16 br16 br16 br16 br16 br16	1992 1992 1993 1992 1993 1993 1993 1993 1993 1999 1993 1999 1993 1999 1993 1999 1993 1999 1993 1999 1993 1994 1993 2048 1202 2048 1202 1994 1202 1	B 44 ma 94 ma 42 94 ma 2 token/se 42 94 ma 2 token/se 40 55 ma 1334 ma 64 94 ma 55 64 53 ma 55 43 53 ma 54 43 53 ma 54 43 53 ma 54 43 53 ma 52 43 53 ma 54 43 53 ma 52 43 53 ma 54 43 53 ma 52 81 25 ma 54 103 51 37 ma 55 420 87 ma	8 8 4 8 4 16 1 1 1 1 1 1	text-generation text-classification language- modeling question- answering language- modeling constructions protein-folding text-generation text-generation	Desplored 0.12.4.00tmm Habana 19.0 Optimum Habana 1.9.0 Optimum Habana 1.9.0	LLARA, 12 Fine Tuning LLARA, 12 Fine Tuning (LLARA), 12 Fine Tuning (LLARA), 100 Fine Tuning (LLARA), 100 Fine Tuning (LLARA), 100 Fine Tuning (LLARA), 100 Fine Tuning OFT2 CLAM OFT2 CLAM OFT2 CLAM OFT2 CLAM OFT2 CLAM ALBERT Large Fine Tuning CluphoBERTs	Present Present 2 8 Present 8 Present Present 9 Present Present	entinces/sec 1.53 sentences/se 28.83 sentences/sec 10.64 solutions sentences/sec 577.41 sentences/sec 28.0.45 sentences/sec 28.0.45 sentences/sec sentences/se	0.93 S.Smin 105 S.2013 119 S.2013 0.32 S.13 0.33 S.13 0.41 S.20 0.7 S.25 0.7 S.25 0.7 S.25 0.7 S.25 0.8 S.20 1.9 S.25 0.20 S.26 0.25 S.26	mm 1 Imp in 1 Imp in 1 Imp in 4 Imp ini 2 Imp ini 4 Imp ini 4 Imp ini 32 Que ini 24 Que	guage-modeling guage-modeling guage-modeling guage-modeling guage-modeling mtrastive-image-text guage-modeling estion-answering estion-answering estion-answering mtrastive-image-text	Optimum Habina 1.0. Optimum Habina 1.0.	Restarts Duras LARS BERT Pri Taning combine BERT Pri Taning combine BERT Pri Taning combine BERT Pri Taning Prise BERT Pri Taning Prise 2 BERT Pri Taning Pri School Pri Scho	8 1 32 8 1	b116 4005 33 sent/sec 1016 122.73 sent/sec 1016 133 sent/sec 1016 533 sent/sec 1016 533 sent/sec 1016 533 sent/sec 1016 533 sent/sec 1016 513 sent/sec 1016 1014 sent/sec 1016 10114 sent/sec 1016 405 sent/sec 1016 102 sent/sec 1016 405 sent/sec	90.68 77 71.21 78.56 ec 28.2	mm 64 64 64 mm 64 mm 64 100 64 1100 64 22 22 233 64 534 64 1200 64 534 64 534 64 534 64 534 64 534 64 534 64 534 64 535 64 535 64 535 64 534 64 535 64 535 64 535 64 535 64 535 64 535 64 535 64 536 64 537 64 536 64 537 64 536 64 537 64 <tr< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr<>								
BridgeTower 1 BridgeTower 1 BridgeTower 1 BridgeTower 1 BrabelLM-78 1 StabiLLM-78 1 StabiLLM-78 1 Disdeminization 1 Disdeminization 1 HT-55 Barmarization 1	br16 br16 br16 br16 br16 br16 br16 br16	1992 1992 1993 1993 1993 1993 1993 1993	38 0.48 ms 1/126 42.94 ms 1/126 42.94 ms 1/126 42.94 ms 54 0.45 ms 55 0.49 ms 65/126 0.43 ms 1024/126 4.53 ms 1024/126 2.72 ms 34 - 22 0.12 ms 1025 ms 1.03 ms 1026 ms 1.03 ms 1026 ms 40.87 ms 1026 ms 40.87 ms 1026 ms 40.87 ms	8 8 4 8 4 16 1 1 1 1 1 1	text-generation text-classification language- modeling question- answering language- modeling constructions protein-folding text-generation text-generation	Desp5peed 0.12.4.09tmm Habana 19.0 0ptimum Optimum Habana 1.9.0	LDAM Then Turing LDAM Then Turing LDAM Then Turing LDAM The Turing CHA Tu	and and and and and and and and and and and and and and and and and and and and and	sentences/sec 1.53 sentences/s 28.83 sentences/sec 20.03 sentences/sec sentences/sec sentences/sec 20.04 sentences/sec 20.03 sentences/sec 20.040 sentences/sec 20.0	0.93 5.5 min 1.05 2.013 1.19 1.64 mi 0.53 1.13 mi 0.17 2.05 mi	1 leng in 1 leng in 1 leng in 2 leng in 4 lang in 4 lang in 4 lang in 4 lang in 12 que in 12 que in 64 que in 8 que	gauge-modeling gauge-	Optimum Hadawa 19.0 Optimum Hadawa 19.0	Restart Draws LMS ERF Ps Training combine ERF Ps Training combine ERF Ps Training combine ERF Ps Training Press ERF Ps Training Pres	8 1 32 8 1	015 40553 sentitive 016 123.75 antitive 127.75 antitive 34 antitive 116 123.84 antitive 116 347.95 antitive 116 347.95 antitive 116 347.94 antitive 116 347.94 antitive 116 347.94 antitive 116 425.55 sentitive 116 425.55 sentitive 116 426.55 sentitive 116 426.55 sentitive 116 53.54 antitive 116 53.94 antitive 117 7.41 antitive 118 126.25 antitive 119 126.25 antitive 116 126.25 antitive 117 126.25 antitive 116 126.25 antitive 116 126.23 antitive 116 512.27 antitive	400	mn 64 1464 64 1464 64 64 64 64 64 64 64 64 64 64	Uphing 112							
Interfact I Classification) I Ber (Lancales Electron) I Modeling) I BridgeTower I BridgeTower I StableLM-78 I StableLM-78 I T5-38 Summarization I T5-38 Generalization I T5-38 Generalization I HT-75 Smither I	bris bris bris bris bris bris bris bris	1992 105.5 1992 106.5 18.2 18.2 99.9 99.9 10.4 10.4 10.4 20.4 20.4 20.4 20.4 10.4 10.4 10.4 10.4 10.4 10.4 10.4 1	38 44 ma 1/1/26C 42.94 ms 36/1/1/26C 42.94 ms 34 40.55 ms 34 13.34 ms 45 61.91 ms 55 42.91 ms 51/16C 42.91 ms 34/16C 12.47 ms 34/16C 43.91 ms 52 4.3 ms 52 4.12 ms 1024:res 4.32 ms 1024:res 4.20.87 ms 10540:res 3.87 ms	8 8 4 8 4 16 1 1 1 1 1 1	ted generation ted classification indefinition modeling selector- sessering modeling	Developed 013.0 (offmum Habana 10.0 (offmum Habana	LUAN) 10 Protections LUAN) 10 Protections LUAN) Factor LUB Free Turing (LUAN) GPT3 CLM GPT3 CLM GPT3 CLM GPT2 CLM GPT2 GPT2 GPT2 GPT2 GPT2 GPT2 LUAN EETT Sale EETT Sale EETT False EETT False EETT False	end end 0 end end	entennees/see 1.53 sentennees/s 2.58 sentennees/s 2.60 all 2.60 all 2.6	0.93 S.Smin 105 S.2013 119 S.2013 0.32 S.13 0.33 S.13 0.41 S.20 0.7 S.25 0.7 S.25 0.7 S.25 0.7 S.25 0.8 S.20 1.9 S.25 0.20 S.26 0.25 S.26	1 leng in 1 leng in 1 leng in 2 leng in 4 lang in 4 lang in 4 lang in 4 lang in 12 que in 12 que in 64 que in 8 que	guage-modeling guage-modeling guage-modeling guage-modeling guage-modeling mtrastive-image-text guage-modeling estion-answering estion-answering estion-answering mtrastive-image-text	Optimum Habina 1.0.	Restantion Durans LAMS BERT Pro Training combine BERT Pro Training combine BERT Pro Training combine BERT Pro Training Press 1 BERT Pro Training Press 2 BERT Pre	8 1 32 8 1	b15 4055 33 sent/sec b16 122.73 sent/sec b17 122.73 sent/sec b18 527.95 sent/sec b19 527.95 sent/sec b16 133.45 sent/sec b17 154.18 sent/sec b18 191.14 sent/sec b19 52.35 sent/sec b16 62.55 sent/sec b16 52.55 sent/sec b17 52.55 sent/sec b16 52.55 sent/sec b17 52.55 sent/sec b16 52.55 sent/sec b16 52.55 sent/sec b16 53.55 sent/sec b17 53.55 sent/sec b16 53.55 sent/sec b17 53.55 sent/sec b16 53.55 sent/sec b17 53.55 sent/sec b17 53.55 sent/sec b17 53.55 sent/sec <t< td=""><td> Loss: 1.33 90.68 77 71.21 78.56 ec 28.2 72.6 74.21 </td><td>mm 64 44 54 54 54 54 54 54 54 54 5</td><td>Uphing 212 Uphing 212</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></t<>	 Loss: 1.33 90.68 77 71.21 78.56 ec 28.2 72.6 74.21 	mm 64 44 54 54 54 54 54 54 54 54 5	Uphing 212 Uphing 212							
BridgeTower 1 BridgeTower 1 BridgeTower 1 BridgeTower 1 BrabelLM-78 1 StabiLLM-78 1 StabiLLM-78 1 Disdeminization 1 Disdeminization 1 HT-55 Barmarization 1	bris bris bris bris bris bris bris bris	1992 105.5 1992 106.5 18.2 18.2 99.9 99.9 10.4 10.4 10.4 20.4 20.4 20.4 20.4 10.4 10.4 10.4 10.4 10.4 10.4 10.4 1	38 0.48 ms 1/126 42.94 ms 1/126 42.94 ms 1/126 42.94 ms 54 0.45 ms 55 0.49 ms 65/126 0.43 ms 1024/126 4.53 ms 1024/126 2.72 ms 34 - 22 0.12 ms 1025 ms 1.03 ms 1026 ms 1.03 ms 1026 ms 40.87 ms 1026 ms 40.87 ms 1026 ms 40.87 ms	8 8 4 8 4 16 1 1 1 1 1 1	text-generation text-classification language goestion- answering language modeling constrative- image-text protein-folding text-generation text-generation summarization summarization	Desp5peed 0.12.4.09tmm Habana 19.0 0ptimum Optimum Habana 1.9.0	LDAM Then Turing LDAM Then Turing LDAM Then Turing LDAM Then Turing LDAM The Turing LDAM Turing <	and and and and and and and and and and and and and and and and and and and and and	entennes/see 1.53 sentennes/se 28.83 sentennes/sec 200.8 200.8 200.8 200.8 200.8 200.8 200.8 200.8 200.45 200.45 200.45 200.45 200.45 200.45 200.45 200.8 200.	0.91 0.5 min 1.05 2.013 1.10 3.6 min 0.20 3.13 0.21 3.0 min 0.22 3.0 min 0.23 3.0 min 0.24 3.0 min 0.25 3.0 min 0.25 3.0 min 0.25 3.0 min 0.25 3.0 min	mm 1 leng in 1 leng mm 2 leng mm 2 leng mm 4 lang mm 4 lang mm 4 lang mm 2 que mm 2 que mm 12 que mm 64 que mm 64 que mm 64 que mm 2 que mm 2 que mm 2 que	gauge-modeling gauge-	Optimum Hadawa 19.0 Optimum Hadawa 19.0	Restart Draws LMS ERF Ps Training combine ERF Ps Training combine ERF Ps Training combine ERF Ps Training Press ERF Ps Training Pres	8 1 32 8 1	015 40553 sentitive 016 123.75 antitive 127.75 antitive 34 antitive 116 123.84 antitive 116 347.95 antitive 116 347.95 antitive 116 347.94 antitive 116 347.94 antitive 116 347.94 antitive 116 425.55 sentitive 116 425.55 sentitive 116 426.55 sentitive 116 426.55 sentitive 116 53.54 antitive 116 53.94 antitive 117 7.41 antitive 118 126.25 antitive 119 126.25 antitive 116 126.25 antitive 117 126.25 antitive 116 126.25 antitive 116 126.23 antitive 116 512.27 antitive	400	mn 64 1464 64 1464 64 64 64 64 64 64 64 64 64 64								

Availability and Customer Momentum

Announcing general availability

DCLTechnologies



Dell PowerEdge XE9680

Air-cooled Dell Al Factory

Shipping Q225





intel Gaudi

Available today on the Intel® Tiber™ Al Cloud

Speed up your development for Gen Al 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

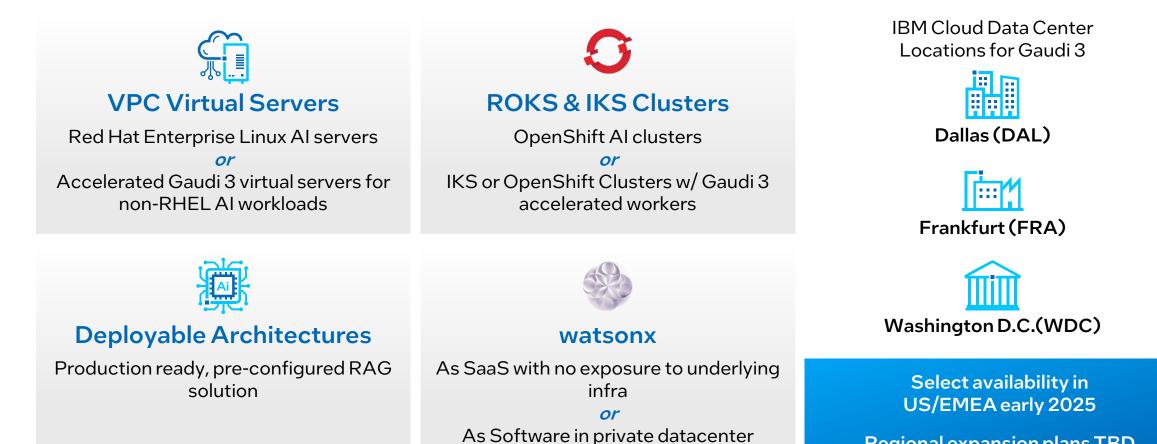
Select availability of Intel Gaudi 3 accelerators on Intel® Tiber™ AI Cloud

Visit Intel Tiber AI Cloud

intel Gaudi Growing Customer Momentum



Intel[®] Gaudi[®] 3 on IBM Cloud Flexible consumption & user experience



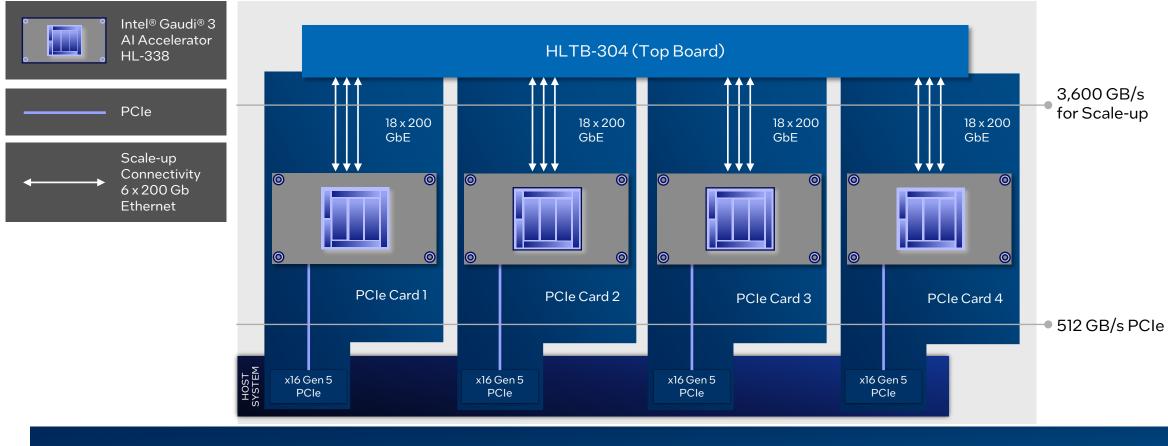
Intel[®] Gaudi[®] Al accelerator - Overview

Regional expansion plans TBD

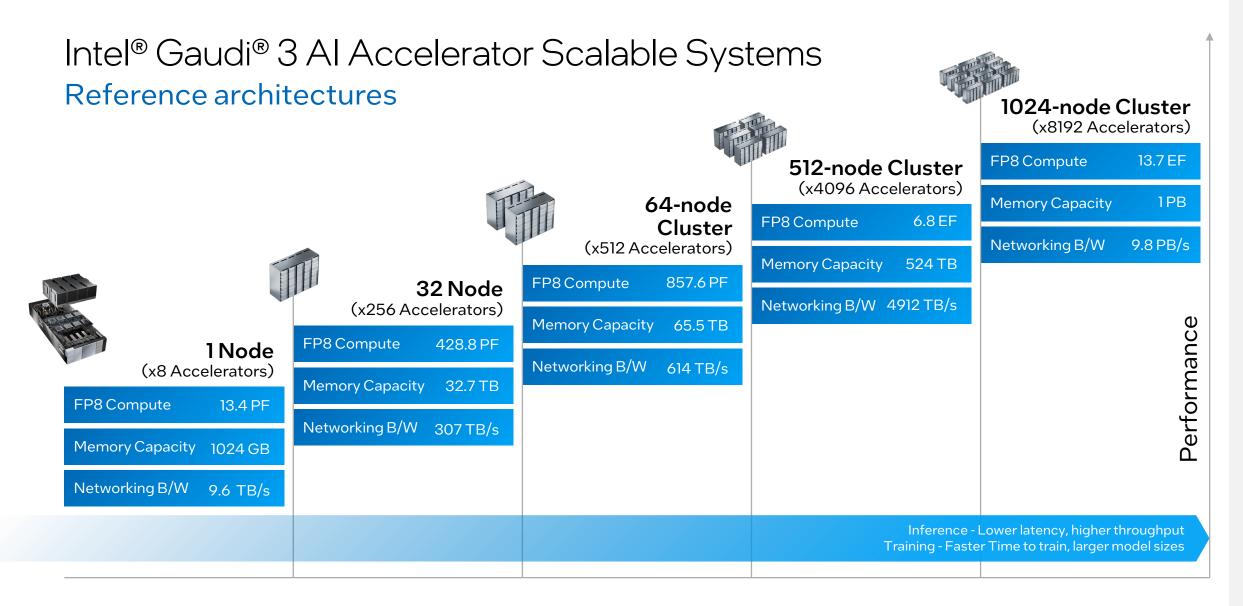
Open and Efficient Scalability

intel Gaud PCIe (HL-338) Block Diagram

4xPCIe cards per system



Ideal for inferencing, fine-tuning & small model training



*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 Al 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

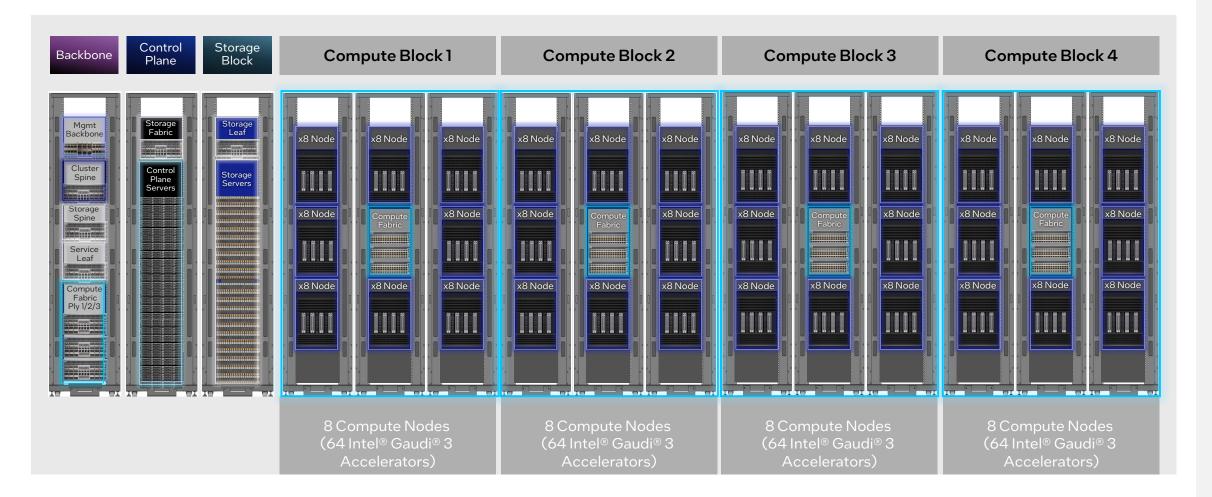
1. Introduction

Contents

Introduction. 1 Reference Design Overview	Al'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 Al infrastructure and concerns about being locked into vendor-specific solutions can slow Al adoption. Fortunately, the market now offers more open industry solutions such as those based on the Intel® Gaudi® Al accelerator product line. Intel Gaudi accelerators are architected for deep learning (DL) and Generative Al, excelling at large language model (LLM) and multi-modal model training and inferencing. Intel Gaudi Al 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 Al compute capability, pricing, energy efficiency, and market availability. ¹
Storage Considerations6 Control Plane Considerations7	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
. Software7 . Summary8	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 <u>www.Intel.com/PerformanceIndex</u>.

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.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

