Hardware and Software AI Acceleration Powered by oneAPI

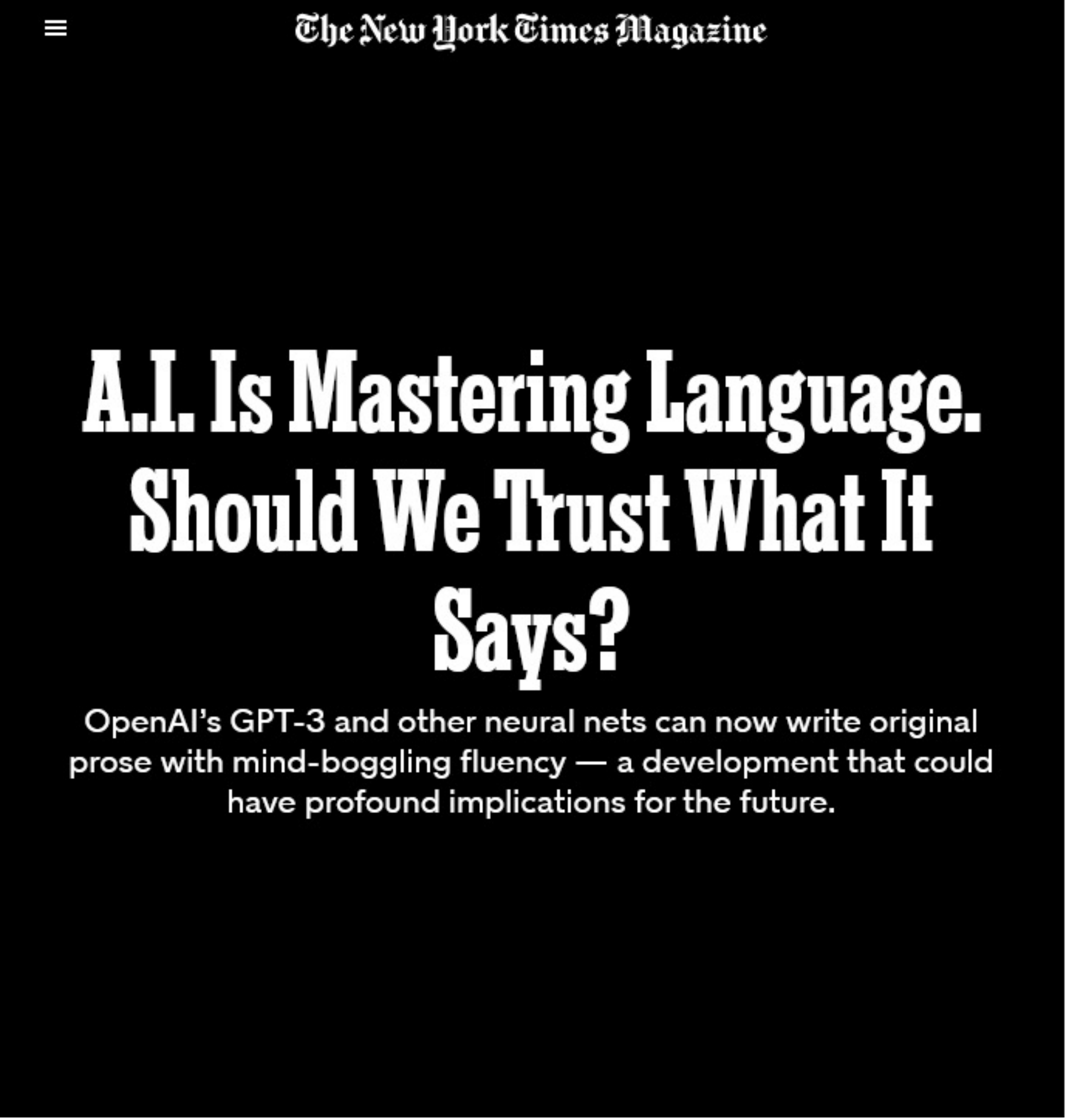
Andres Rodriguez, PhD

Intel Fellow, Chief AI Architect

Contents

- Why invest in AI?
- Intel's AI software and hardware offerings
- Intel Xeon Scalable Processors
- Intel GPU Max Series
- Intel Habana Gaudi
- Ecosystem Programs

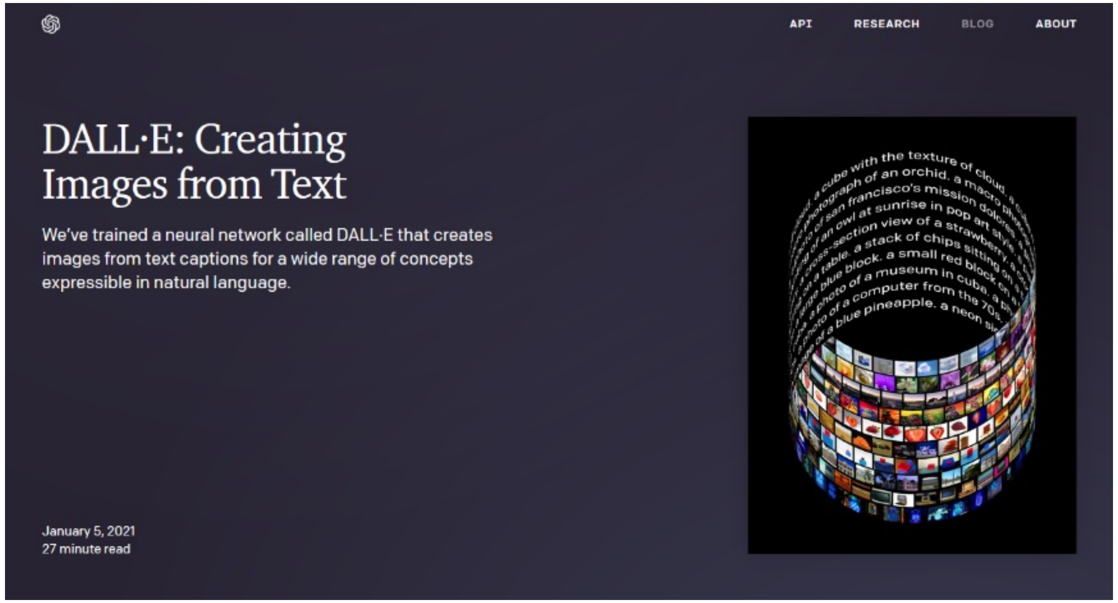
Why Invest in AI?



The New York Times Magazine

A.I. Is Mastering Language. Should We Trust What It Says?

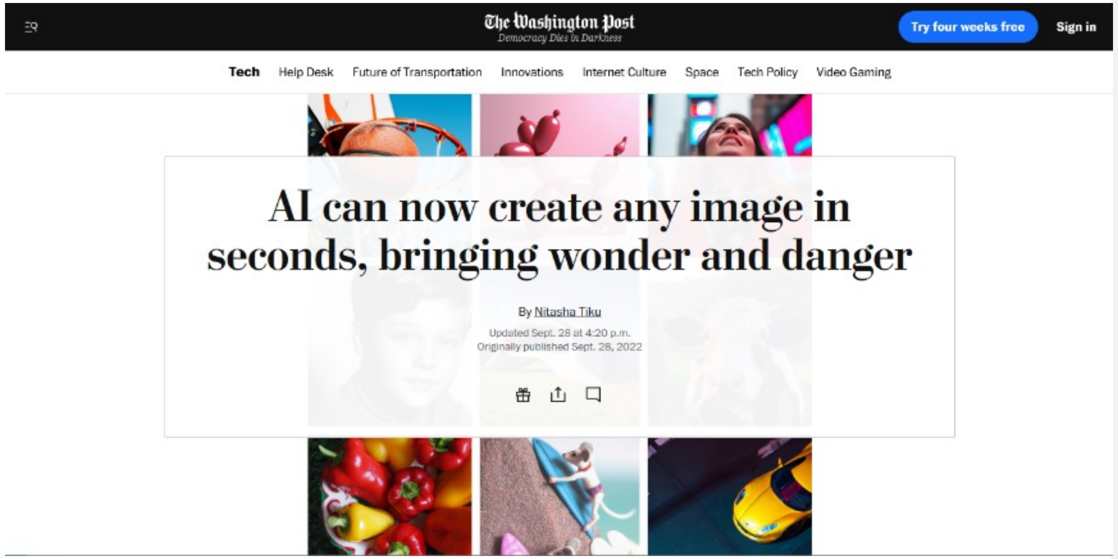
OpenAI's GPT-3 and other neural nets can now write original prose with mind-boggling fluency — a development that could have profound implications for the future.



DALL·E: Creating Images from Text

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.

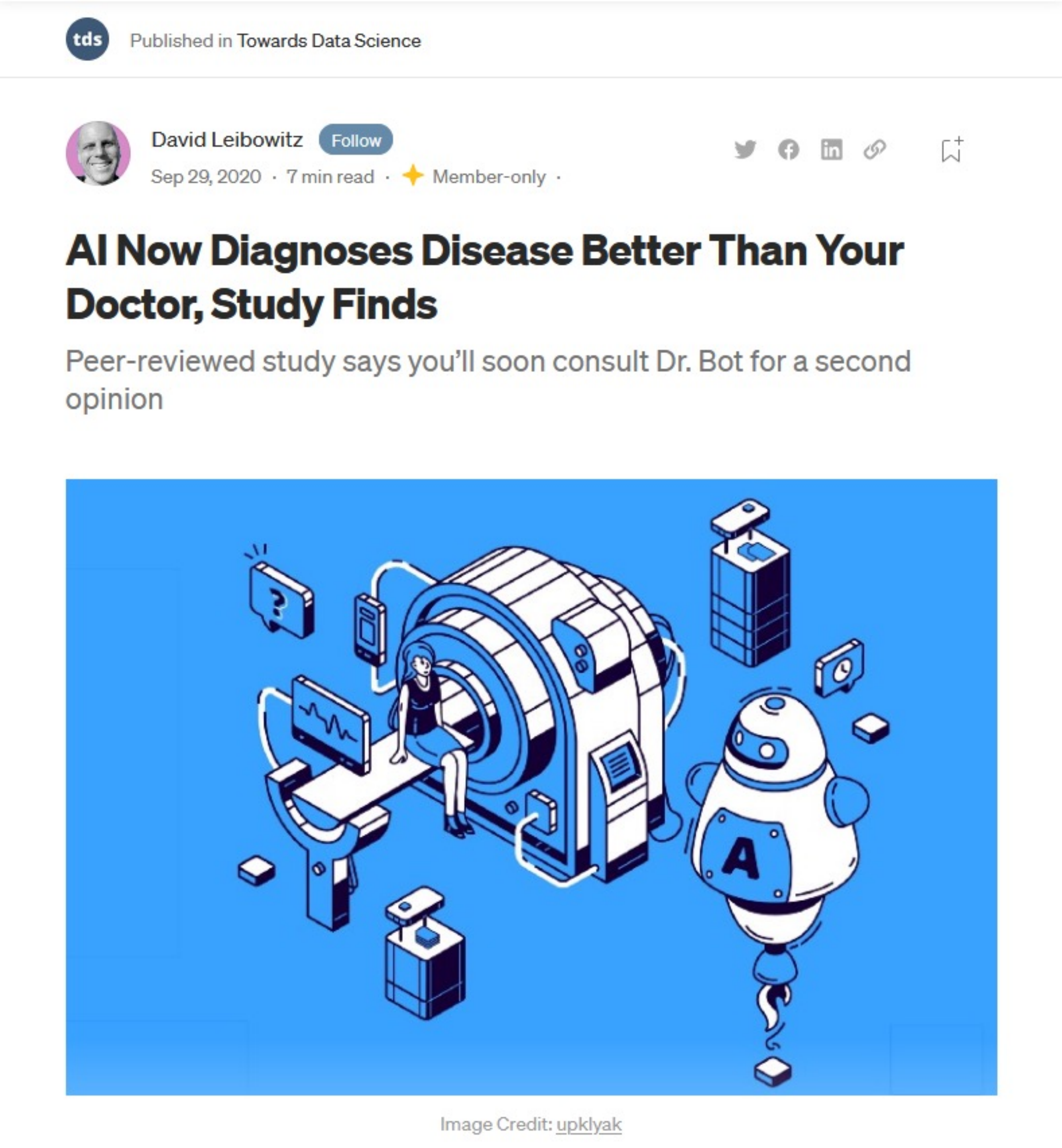
January 5, 2021
27 minute read



AI can now create any image in seconds, bringing wonder and danger

By Nitasha Tiku

Updated Sept. 28 at 4:20 p.m.
Originally published Sept. 28, 2022



Published in Towards Data Science

David Leibowitz Follow

Sep 29, 2020 · 7 min read · ✦ Member-only ·

AI Now Diagnoses Disease Better Than Your Doctor, Study Finds

Peer-reviewed study says you'll soon consult Dr. Bot for a second opinion

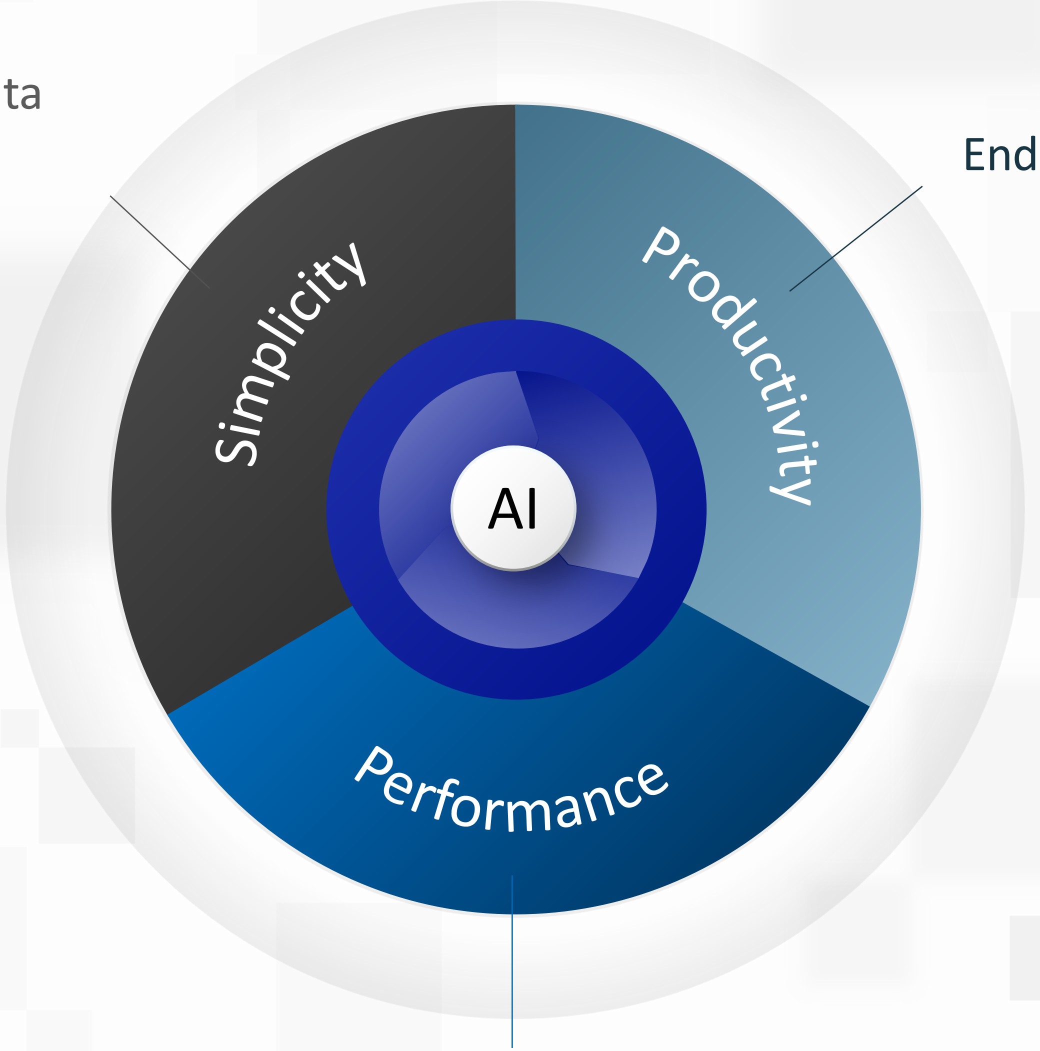
Image Credit: upklyak

Intel’s AI Software and Hardware Offerings

Intel AI Software Strategy

to deliver

Simplicity, Productivity, and Performance



To go from Data to Solutions

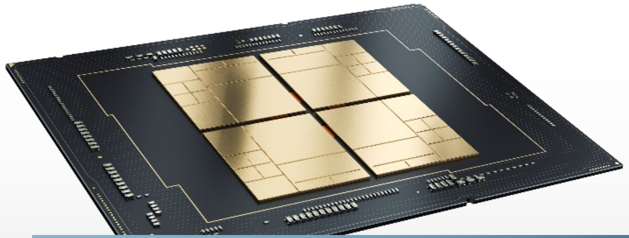
By Optimizing End-to-End Workflows

For Every AI Workload

Intel AI and HPC Hardware Portfolio

4th Gen Intel Xeon Scalable Processor

Architected for AI



New Advanced Matrix Extensions

Intel® AMX

Up to

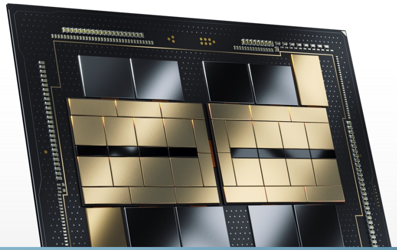
8x gen-on-gen

compute increase

Intel Datacenter GPU

Previously codenamed "Ponte Vecchio" (PVC)

Super-Charged GPU for HPC & AI



Xe Matrix Extensions

Intel® XMX

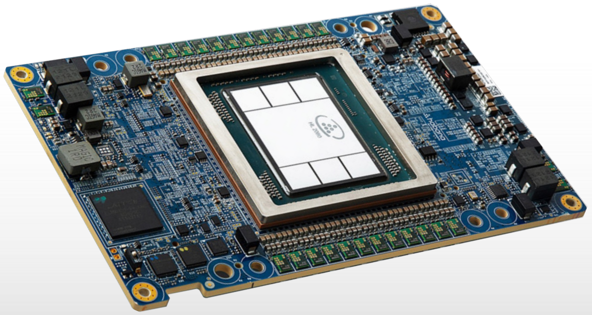
Outperforms Nvidia A100

Training & Inference

PVC B step @ 1.4GHz vs. A100 (80G) Resnet 50

GAUDI®2

Dedicated Deep Learning



~2X Training vs. Nvidia A100

BERT, ResNet-50 Throughput

Performance varies by use, configuration and other factors. PVC performance shared at ISC 2022
Learn more at www.Intel.com/PerformanceIndex.

oneAPI

An Open Project & Intel’s Product

Open Specification for Accelerated Computing

Standards-Based Data Parallel Language

Standard Interfaces for Common Accelerator Libraries

Open-source implementations on diverse non-Intel CPU, GPU, FPGA, and AI solutions

Intel’s Implementation of the oneAPI Specification

First Customer Shipment – Dec 2020

Productive, Performant, Cross-Platform

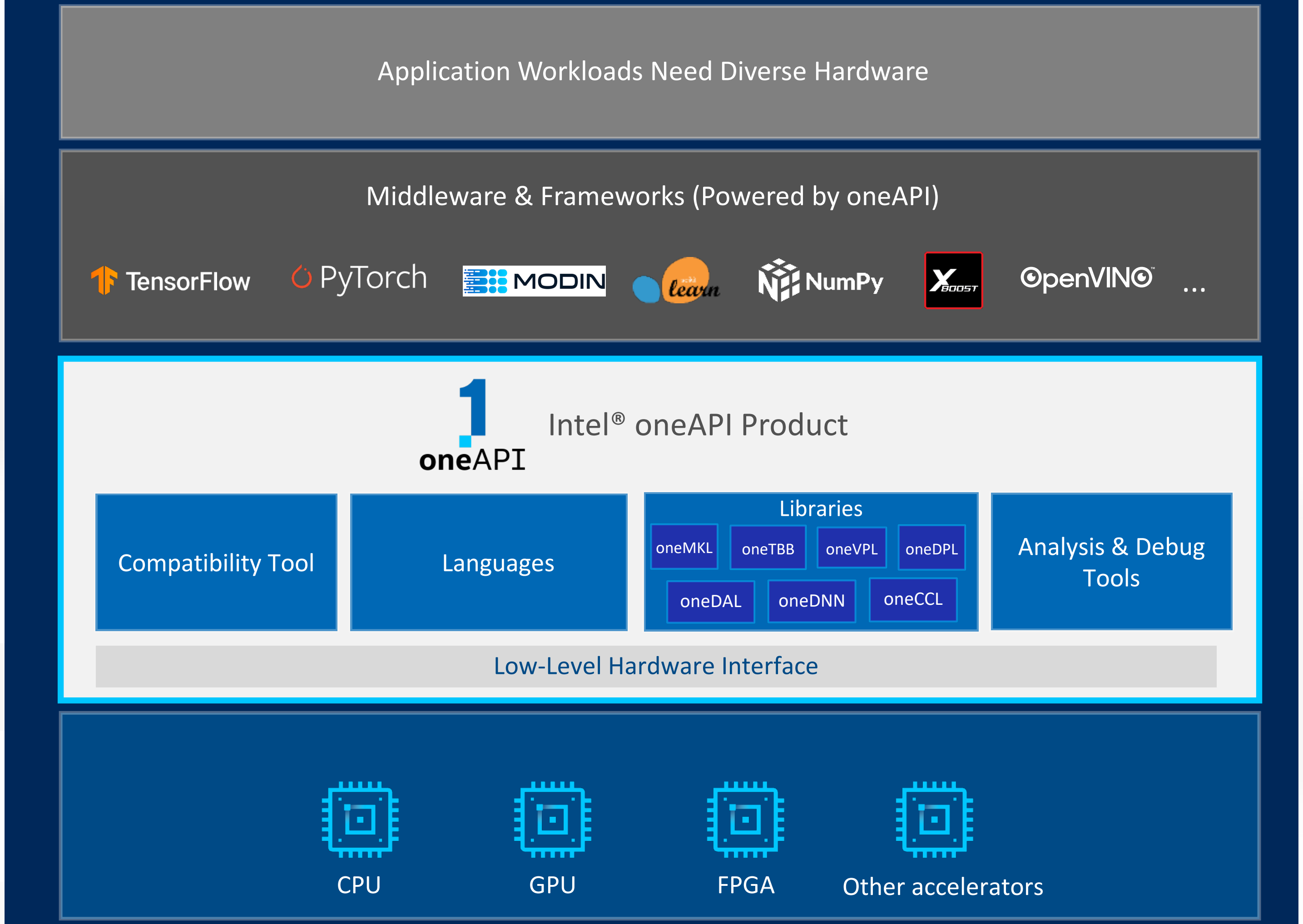
Supports Intel CPU, GPU (integrated & discrete), and FPGA today

Realizing the vision of productive programming for accelerators, free from proprietary lock-in

Intel® oneAPI Software Tools for AI & Analytics

Popular AI frameworks and middleware are extended and optimized using one or more of the oneAPI industry specification elements

Can target CPUs, GPUs, and other accelerators



Application Workloads Need Diverse Hardware

Middleware & Frameworks (Powered by oneAPI)

TensorFlow PyTorch MODIN learn NumPy OpenVINO …

oneAPI Intel® oneAPI Product

Compatibility Tool

Languages

Libraries

oneMKL oneTBB oneVPL oneDPL

oneDAL oneDNN oneCCL

Analysis & Debug Tools

Low-Level Hardware Interface

CPU

GPU

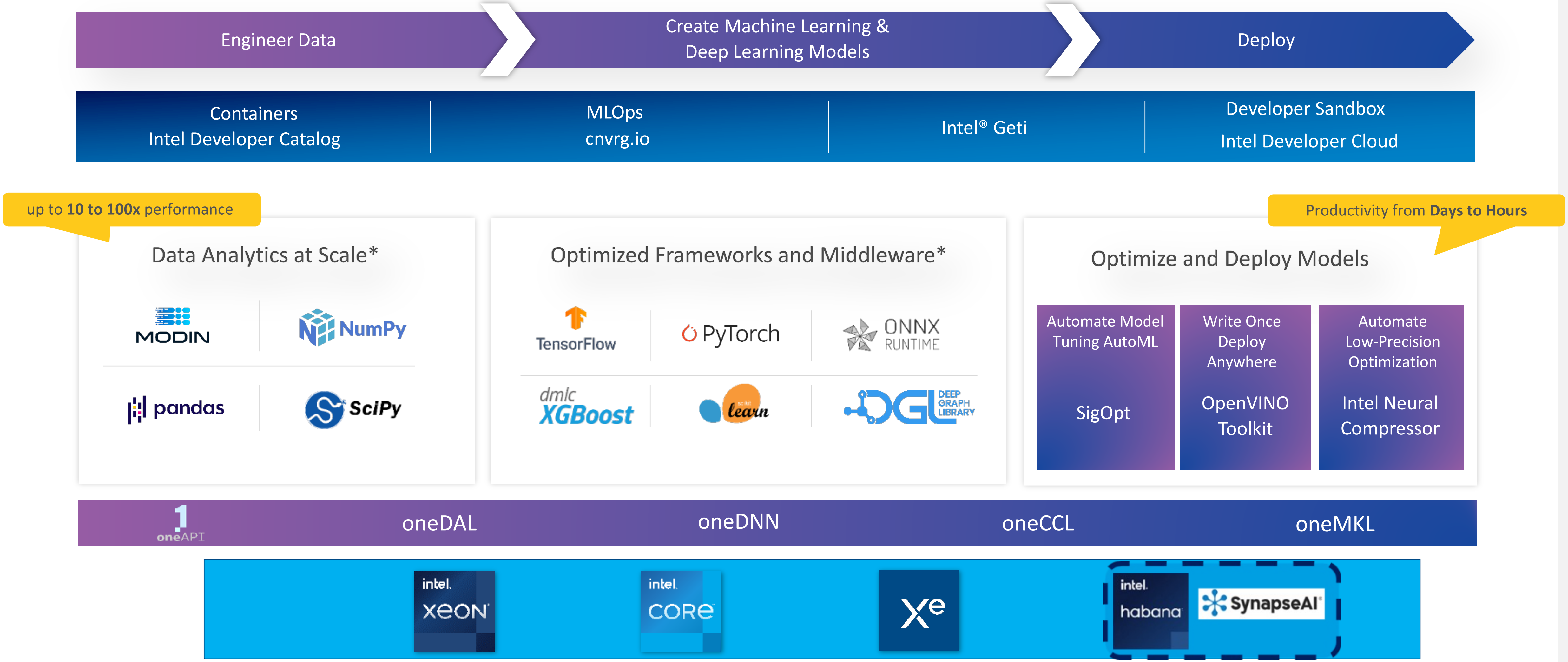
FPGA

Other accelerators

Visit software.intel.com/oneapi for more details

Some capabilities may differ per architecture and custom-tuning will still be required. Other accelerators to be supported in the future.

Enabling End-to-End Software and Solutions Powered by oneAPI



Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

Containers
Intel Developer Catalog

MLOps
cnvrg.io

Intel® Geti

Developer Sandbox
Intel Developer Cloud

up to **10 to 100x** performance

Productivity from **Days to Hours**

Data Analytics at Scale*

MODIN
NumPy
pandas
SciPy

Optimized Frameworks and Middleware*

TensorFlow
PyTorch
ONNX RUNTIME
dmlc XGBoost
scikit learn
DGL DEEP GRAPH LIBRARY

Optimize and Deploy Models

Automate Model Tuning AutoML
SigOpt

Write Once Deploy Anywhere
OpenVINO Toolkit

Automate Low-Precision Optimization
Intel Neural Compressor

oneAPI
oneDAL
oneDNN
oneCCL
oneMKL

intel. xeon
intel. core
Xe
intel. habana
SynapseAI

* Other names and brands may be claimed as the property of others

AI Models Growing in Complexity and Diversity

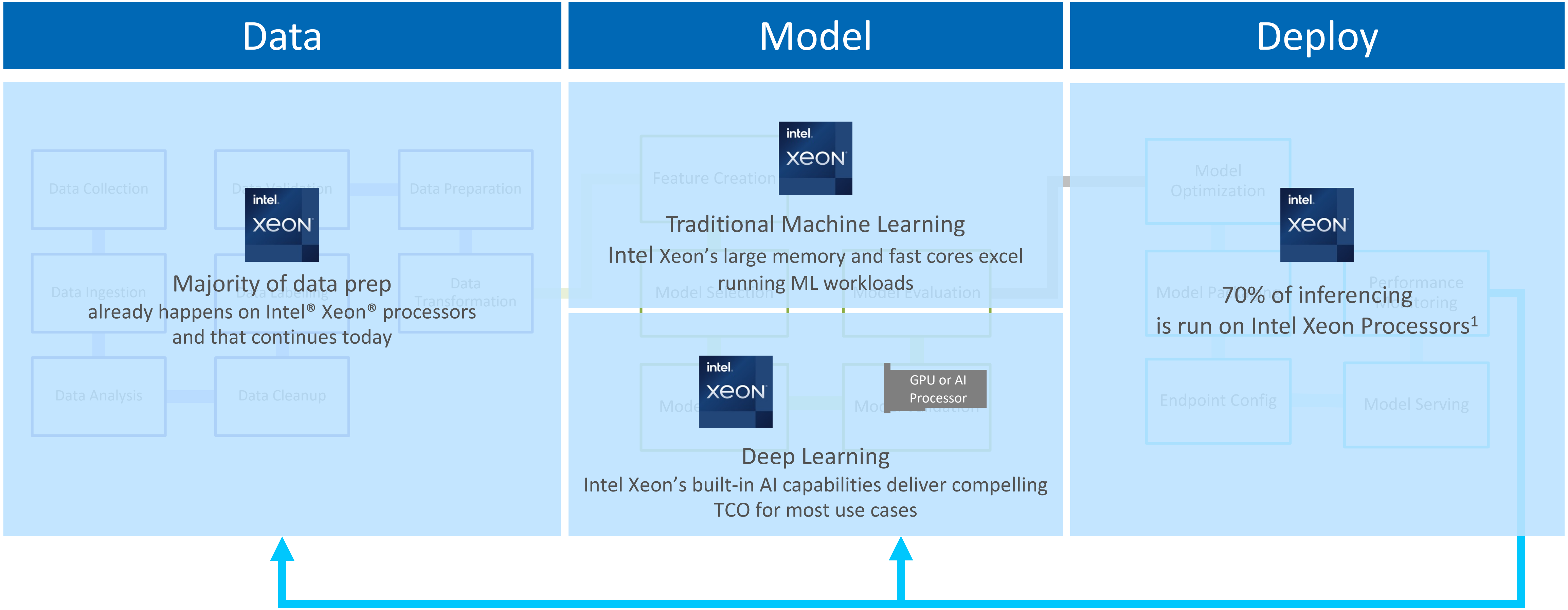
Solution…

Start with Intel Xeon Scalable Processors with built-in HW and SW acceleration

- Availability
- Ease-of-use and use-of-programmability
- Fast cores
- Large memory capacity
- Matured & robust SW stack
- Data pre-processing, AI compute, and post-processing in the same HW
- HW acceleration: AVX512, Intel DL Boost (VNNI), Intel AMX
- SW acceleration: TensorFlow, PyTorch, ONNX Runtime, XGBoost and more …

Use Intel’s discrete accelerators to train large models in less time

The AI Pipeline Runs on Intel Xeon



Data

Majority of data prep
already happens on Intel® Xeon® processors
and that continues today

Model

Traditional Machine Learning
Intel Xeon's large memory and fast cores excel running ML workloads

Deep Learning
Intel Xeon's built-in AI capabilities deliver compelling TCO for most use cases

Deploy

70% of inferencing
is run on Intel Xeon Processors[1]

Database, Data Warehouse, Data Lake, Streaming Data, Feature Store, Model Registry

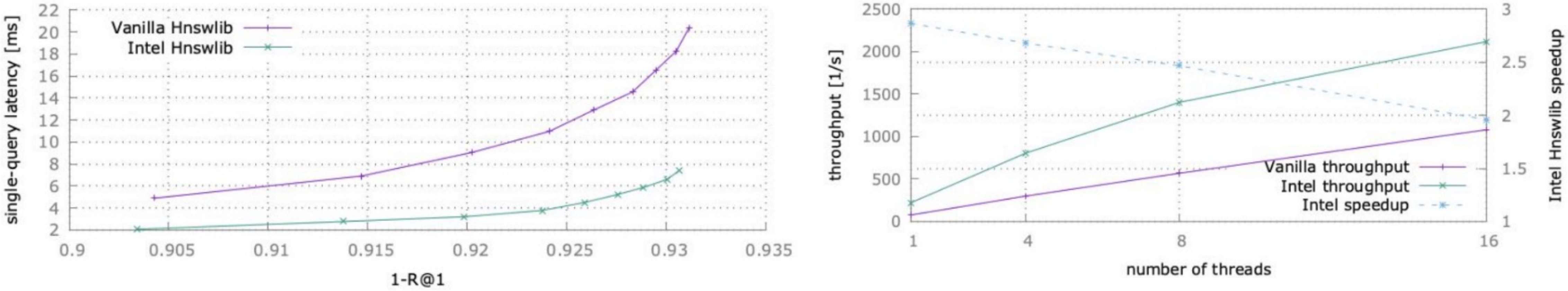
[1] Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2021.

Improved Search

Improved Ranking and Similarity → More relevant search results

- Leveraged DL Boost on Xeon & SW acceleration

2.5x Speedup of Search Latency and Throughput

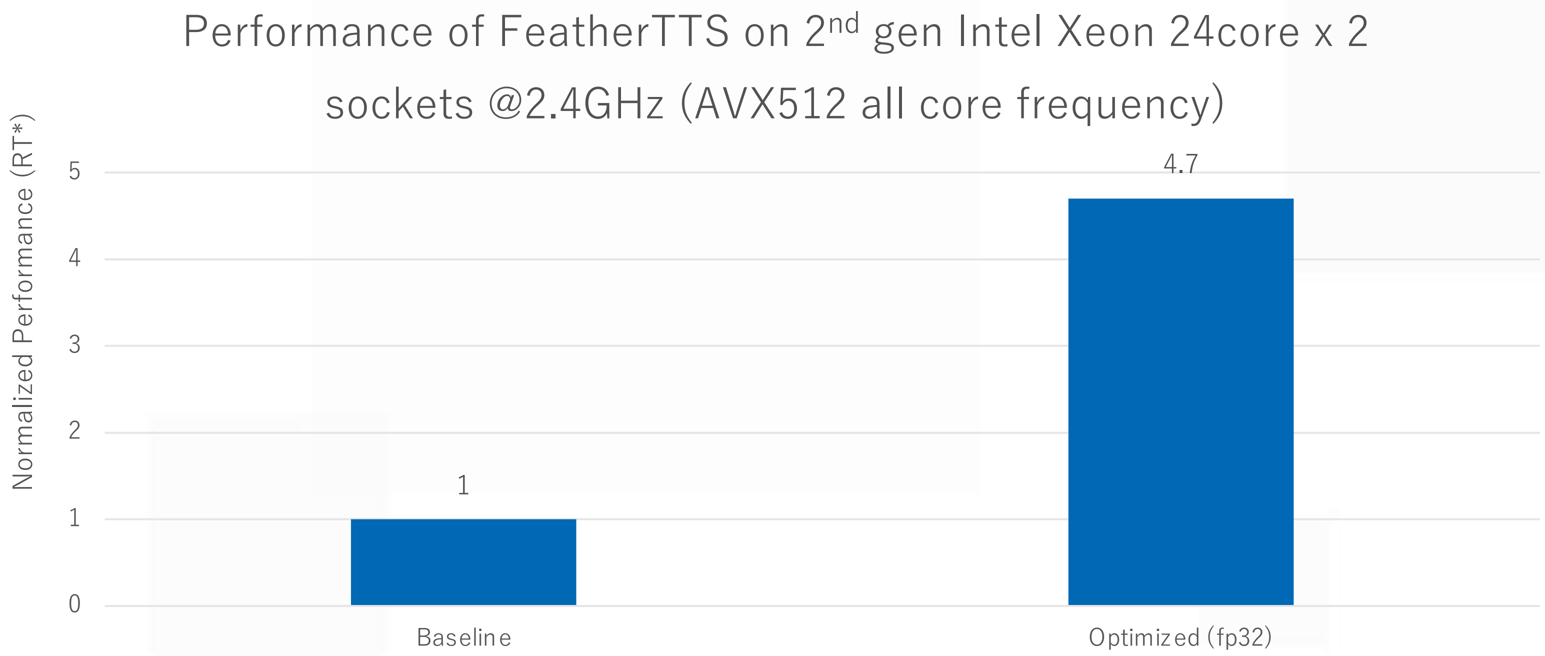


Source (eBay talk): https://player.vimeo.com/video/602002518

Improved Text-to-Speech

Vocoder acceleration → Higher-quality speech synthesis

Tencent Cloud



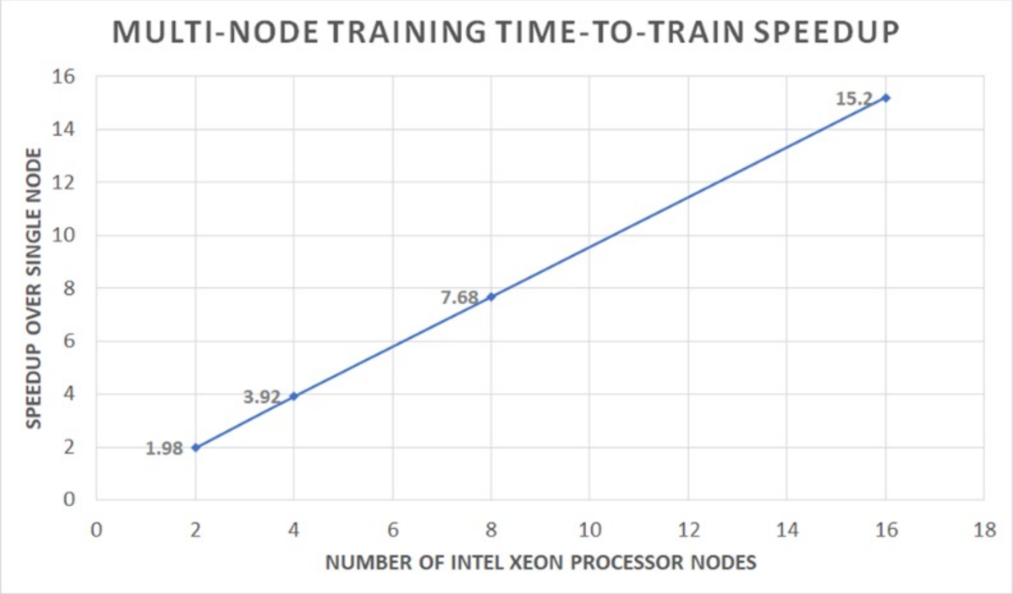
Co-authored paper: https://www.isca-speech.org/archive/pdfs/ssw_2021/tian21_ssw.pdf

Reinforcement Learning Distributed Training

Tencent AI Lab

Efficient RL training on widely available CPUs → Lower operating costs

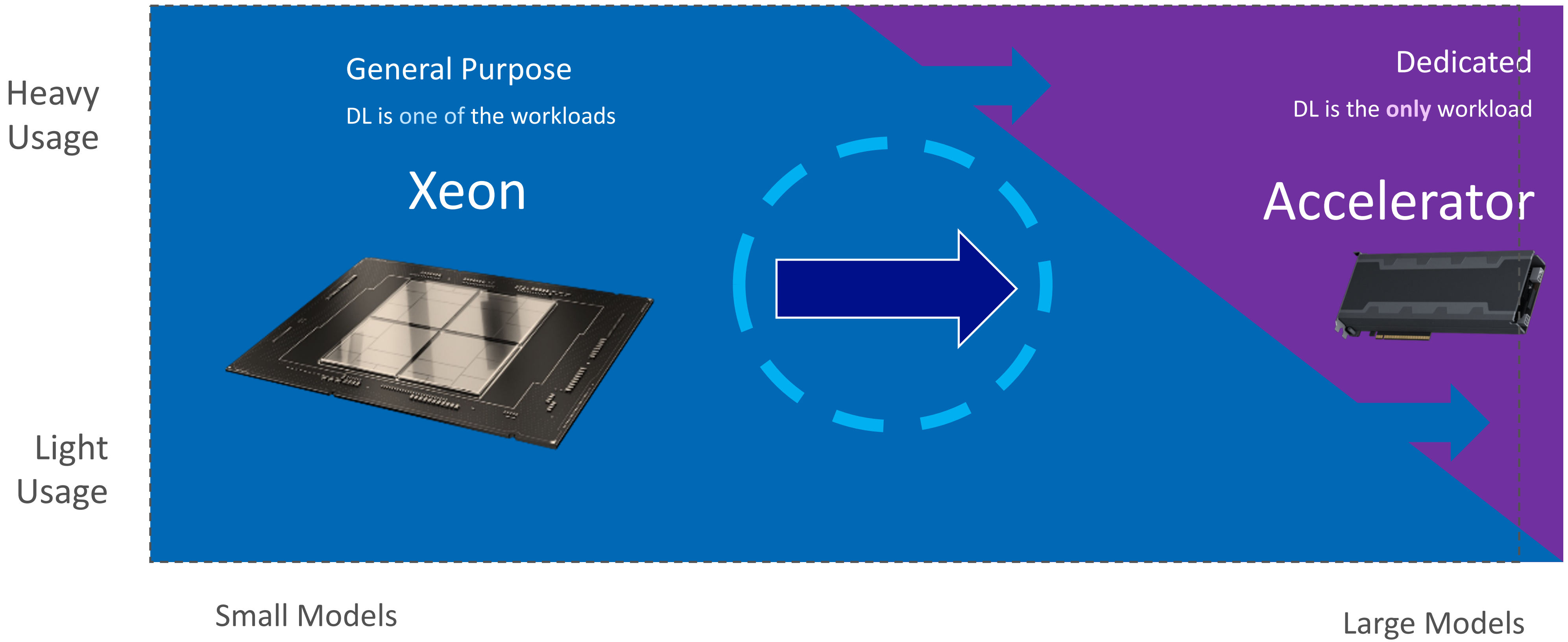
- Tencent's Honor of Kings is the most popular MOBA game in the world
- AI player is trained on 16-node CPU cluster to scale to multiple RL learners
- 15.2x speedup over single node



Joint blog: https://medium.com/intel-analytics-software/distributed-training-on-intel-xeon-scalable-processors-1b335ccf911b

4th Gen Intel® Xeon® Scalable Processors

HW AI Accelerators Built-in Expands the Deep Learning Reach

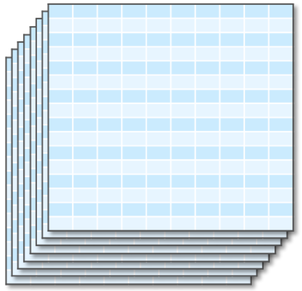


Start with the Xeons you know!

4th Gen Intel® Xeon® AMX Components

“Tiles”

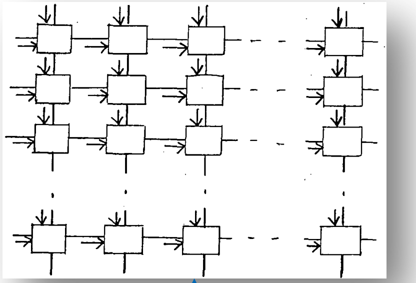
2D Register Files



+

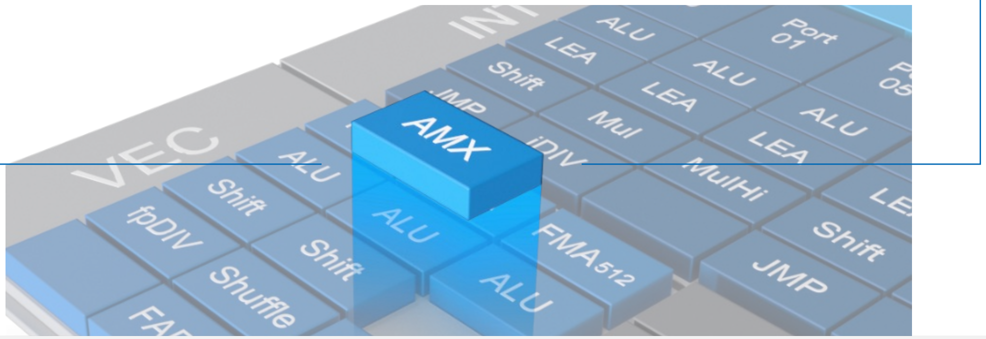
“TMUL”

Tile Matrix Multiply



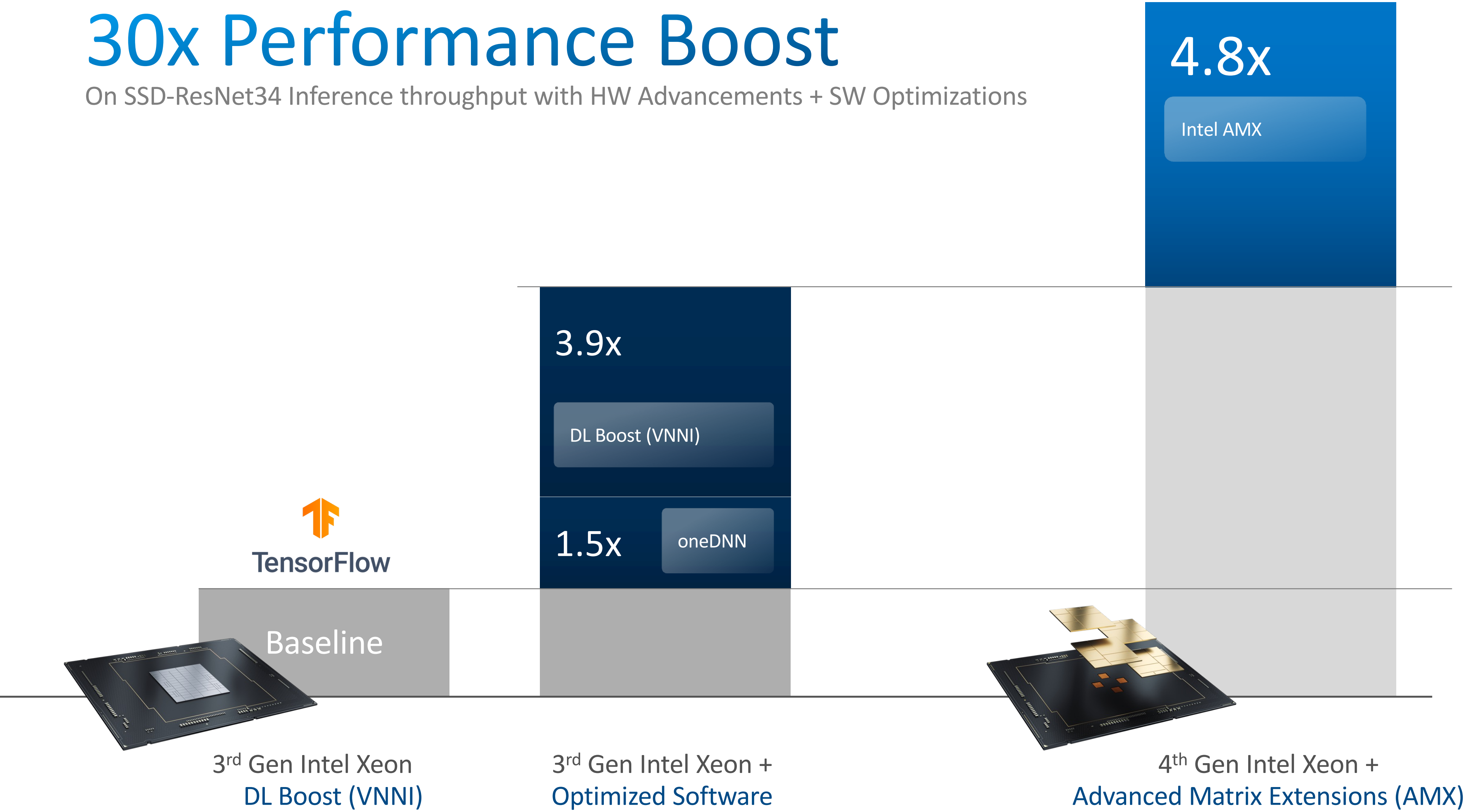
Store bigger chunks of **DATA** in each core

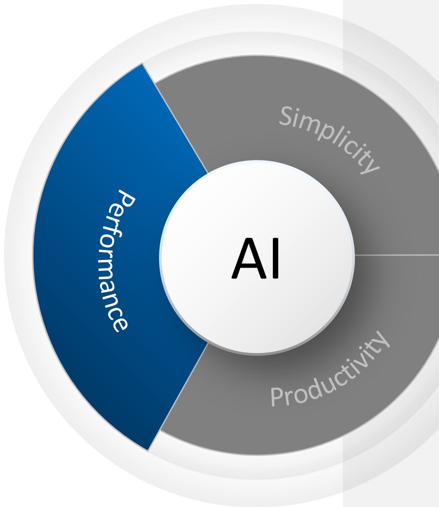
INSTRUCTIONS that compute larger matrices in a single operation



30x Performance Boost

On SSD-ResNet34 Inference throughput with HW Advancements + SW Optimizations

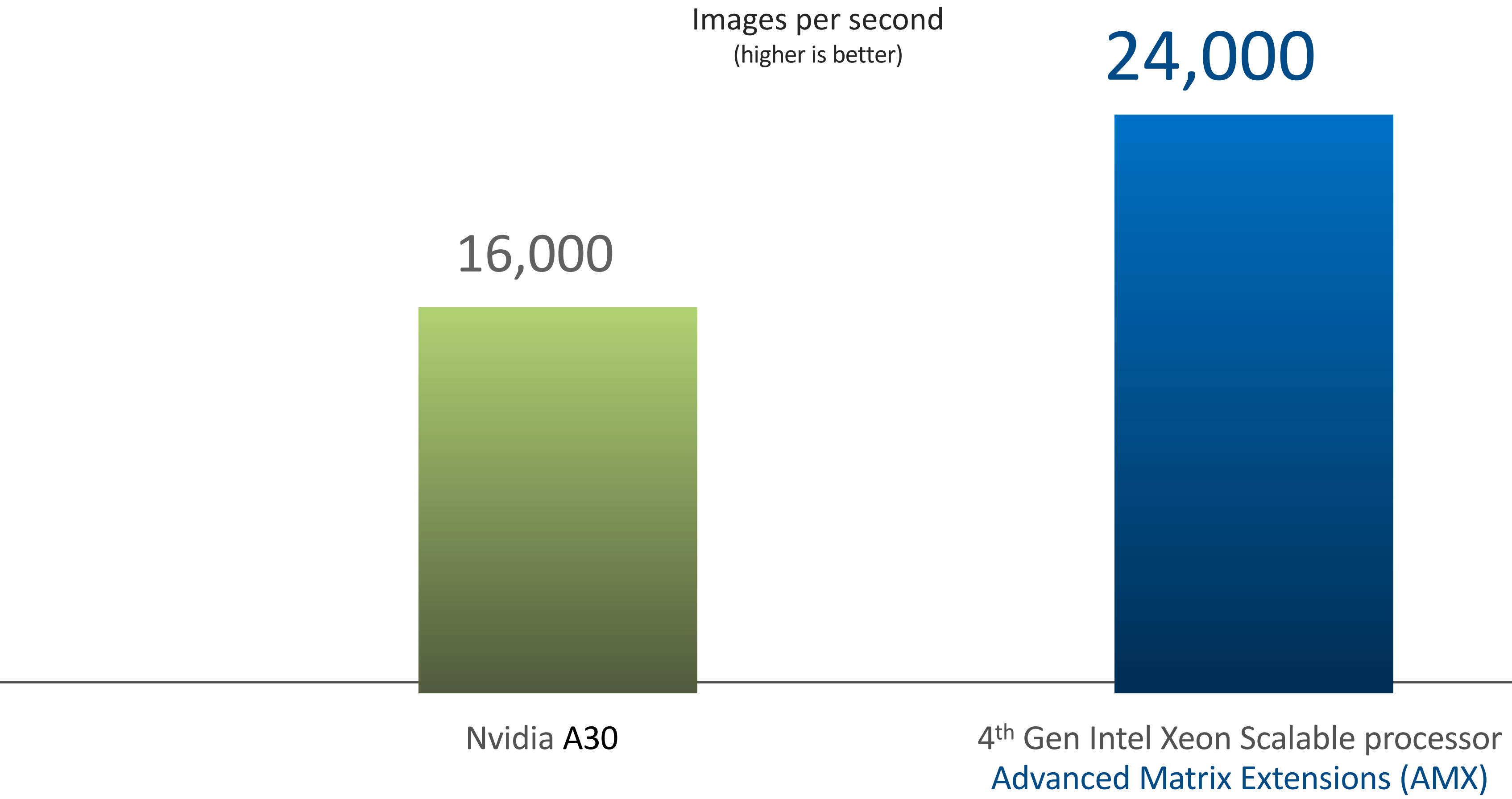


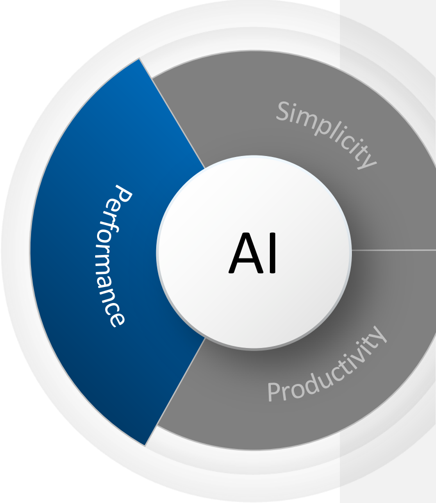


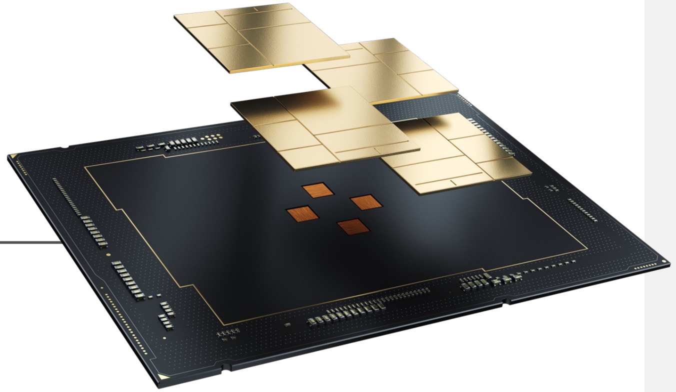
SSD-ResNet-34 Inference Throughput (Batch Size =1) For workloads and configurations visit www.intel.com/InnovationEventClaims. Results may vary.

Intel Xeon with AMX 1.5x Faster

vs. Nvidia A30 on Resnet50







Based on pre-production measurements. See backup for workloads and configurations or visit www.intel.com/innovationeventclaims. Results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

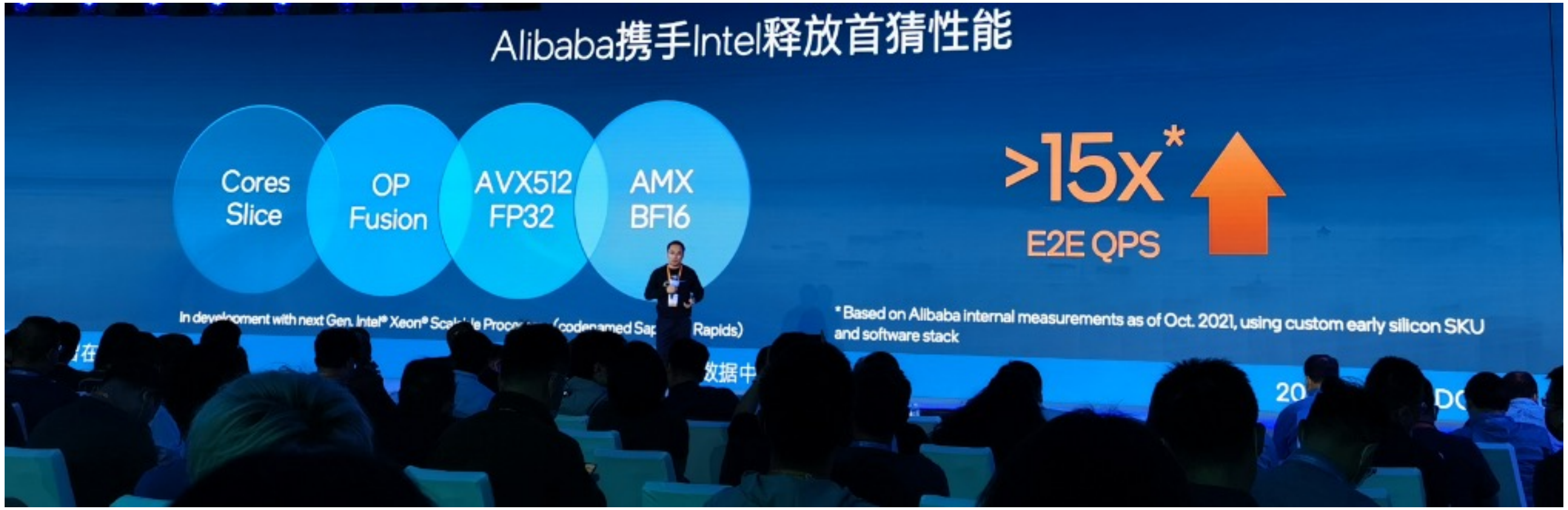
Intel Xeon Processors Targeted AI Use Cases

- Deep learning (DL) inference for all models
- DL training for small and medium models
- DL fine-tuning / transfer learning models
- All traditional ML inference and training workloads
- Infrequent DL large-model training

Alibaba

Early 4th gen Intel Xeon Scalable processor validation in production environment

- Alibaba's custom SW stack
  - oneDNN AMX-BF16, Eigen, graph fusions, and parallelism op optimizations
- 15.9x gains over 3rd gen Intel Xeon Scalable processors using early samples



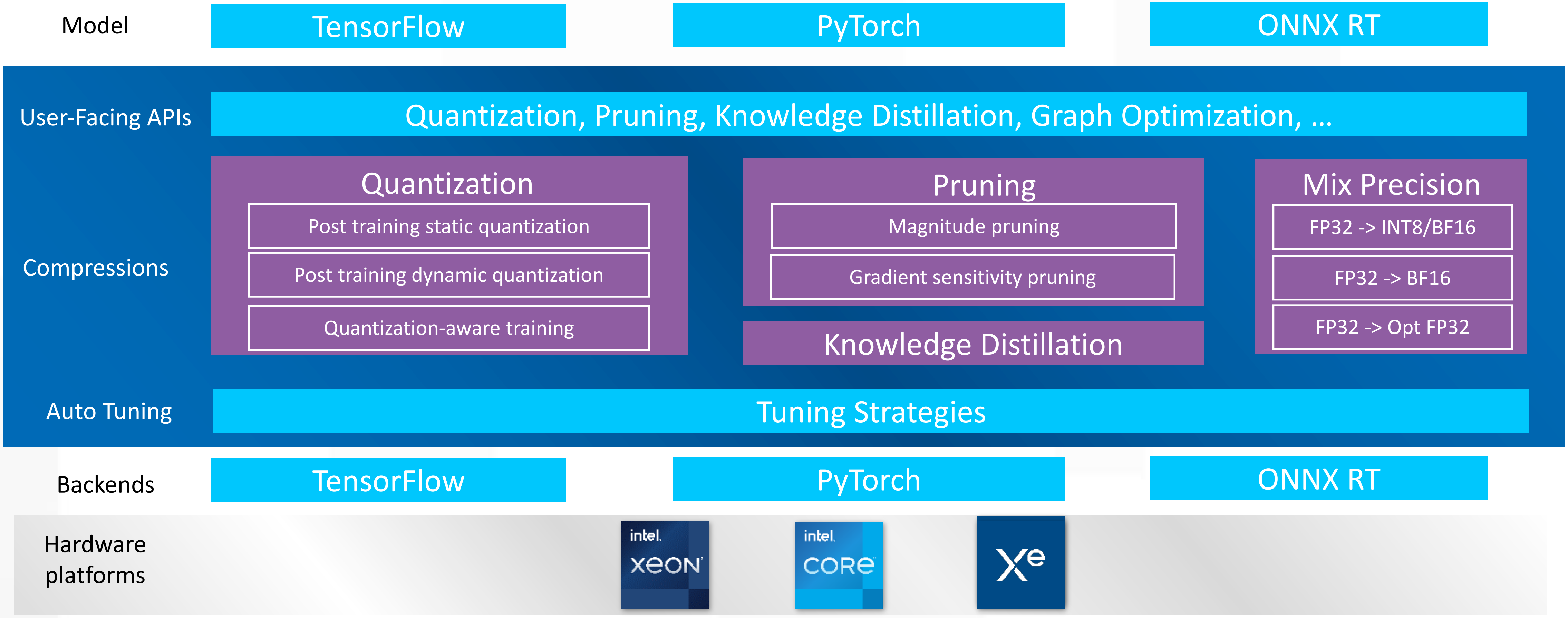
In development with next Gen. Intel® Xeon® Scal le Proc (codenamed Sap Rapids)

*Based on Alibaba internal measurements as of Oct. 2021, using custom early silicon SKU and software stack

Fangzhi, Alibaba Director

Intel® Neural Compressor Infrastructure

Quickly deploy low-precision inference solutions on popular deep learning frameworks such as TensorFlow and PyTorch



Model: TensorFlow | PyTorch | ONNX RT

User-Facing APIs: Quantization, Pruning, Knowledge Distillation, Graph Optimization, …

Compressions:

- Quantization
  - Post training static quantization
  - Post training dynamic quantization
  - Quantization-aware training
- Pruning
  - Magnitude pruning
  - Gradient sensitivity pruning
- Knowledge Distillation
- Mix Precision
  - FP32 -> INT8/BF16
  - FP32 -> BF16
  - FP32 -> Opt FP32

Auto Tuning: Tuning Strategies

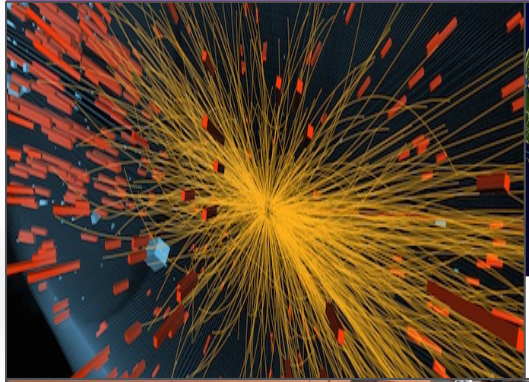
Backends: TensorFlow | PyTorch | ONNX RT

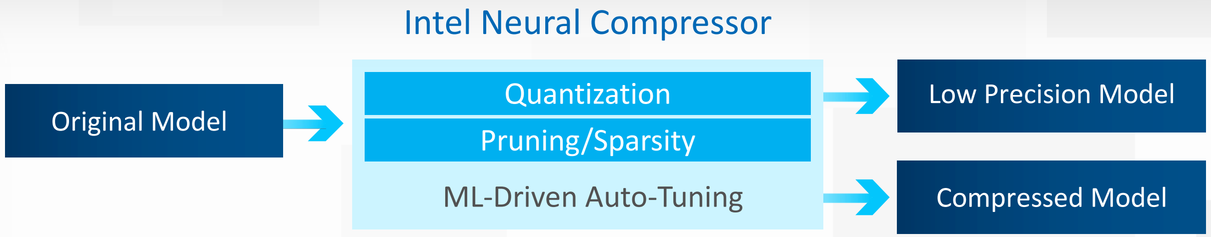
Hardware platforms: intel. XEON | intel. CORE | Xe

AI & HPC: CERN Large Hadron Collider (LHC)

with Intel Neural Compressor for 10X Productivity

High Performance AI Inferencing made Easy





Intel Neural Compressor

Original Model

Quantization

Pruning/Sparsity

ML-Driven Auto-Tuning

Low Precision Model

Compressed Model

- Simulations are essential to all high energy physics experiments
- Complex physics and geometry modeling requires >50% power of worldwide LHC Computing Grid (WLCG)

- Deep Generative Adversarial Networks (GAN) models can replace Monte Carlo simulation to significantly save computation needs and ensure computing requirements remain manageable

- Faster inference via Intel Neural Compressor allows GAN models to generate data on the fly delivering more timely simulations

Sources: Conference paper: 10th International Conference on Pattern Recognition Applications and Methods
Blog: https://www.nextplatform.com/2021/02/01/cern-uses-dlboost-oneapi-to-juice-inference-without-accuracy-loss

oneAPI Deep Neural Network Library (oneDNN)

Integrated into PyTorch and TensorFlow

oneDNN

- Open-sourced supporting Intel and non-Intel hardware products
- Implements rich operators, including convolution, matrix multiplication, pooling, batch normalization, activation functions, recurrent neural network (RNN) cells, and long short-term memory (LSTM) cells
- Supports key data type formats, including 16- and 32-bit floating points, bfloat16, and 8-bit integers
- Accelerates inference performance with automatic detection of Intel® Deep Learning Boost technology

TensorFlow

oneDNN included in TF >= 2.5

Turn on: export TF_ENABLE_ONEDNN_OPTS=1

oneDNN default in TF >= 2.9

PyTorch

oneDNN default in PyTorch >= 1.0

Intel Extension for PyTorch for additional optimization and INT8 quantization

Unlocking TensorFlow for All

TensorFlow + intel ai

Increased Performance by default on CPU

oneAPI & TensorFlow 2.9

No code change

3x perf

Extending Architecture Support

oneAPI

CPU

intel. CORE

intel. XEON

GPU

Xe

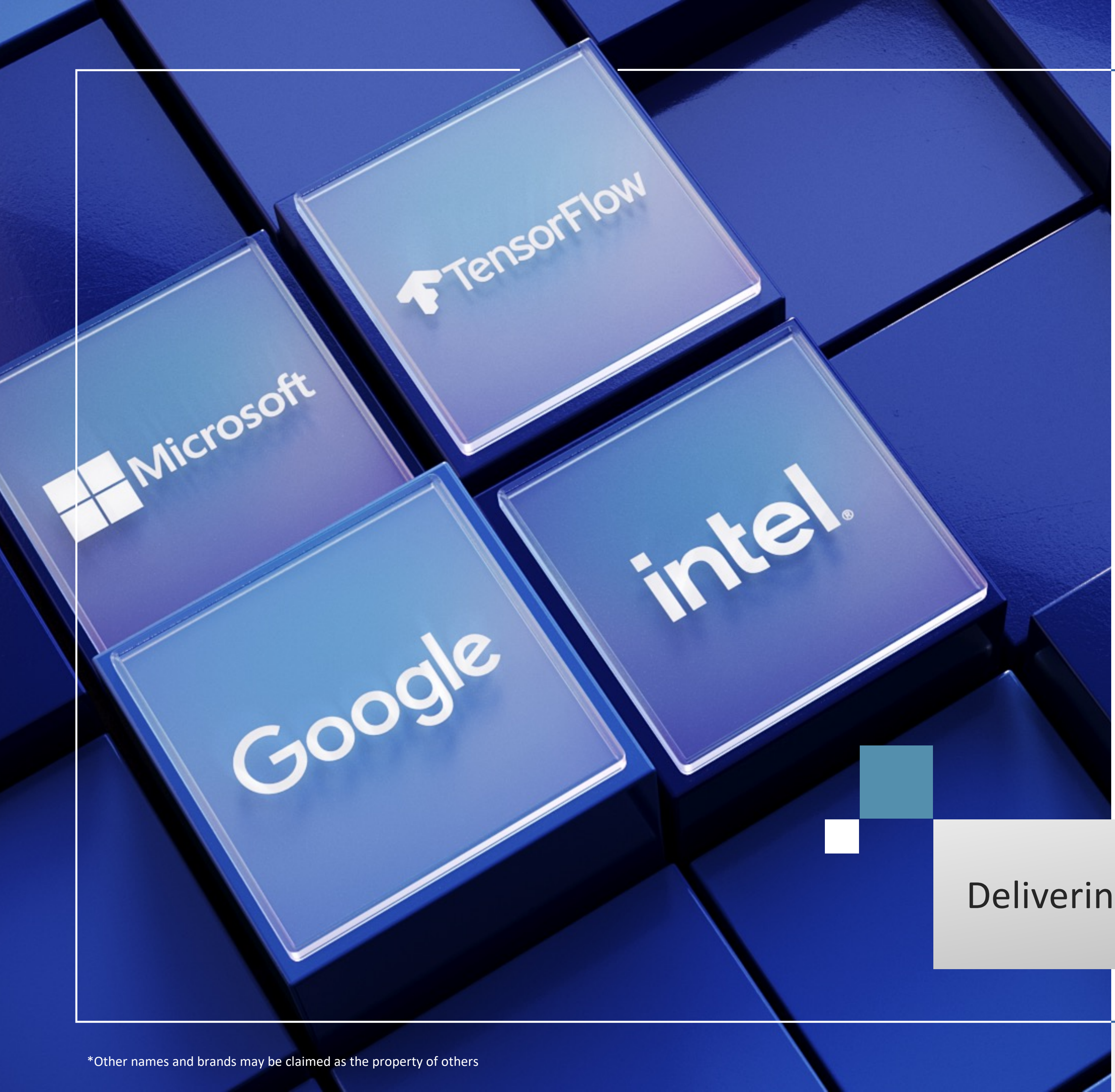
Accelerator

intel. habana

3rd Party *

arm

Source: https://www.intel.com/content/www/us/en/newsroom/news/intel-onednn-speeds-ai-optimizations-in-tensorflow.html
*Other names and brands may be claimed as the property of others



Intel takes ownership

of all future Windows builds of TensorFlow

Delivering more AI performance to more devices

*Other names and brands may be claimed as the property of others

One Line of Code

Unlocks End-to-End Performance Gains

```
import modin.pandas as pd
```

pandas MODIN

Engineer Data

up to 90x

performance

```
from sklearnex import patch_sklearn
patch_sklearn()
```

Intel Extension for scikit-Learn

scikit learn

Create ML and DL Models

up to 38x

performance

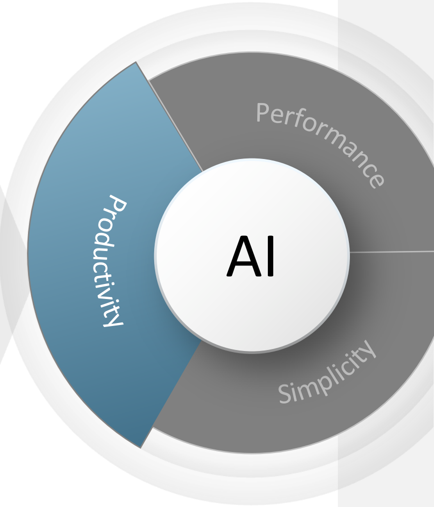
No Change Needed

TensorFlow

Deploy

up to 3x

throughput



Source: https://techdecoded.intel.io/resources/one-line-code-changes-to-boost-pandas-scikit-learn-and-tensorflow-performance/#gs.bzkn2n
* Other names and brands may be claimed as the property of others

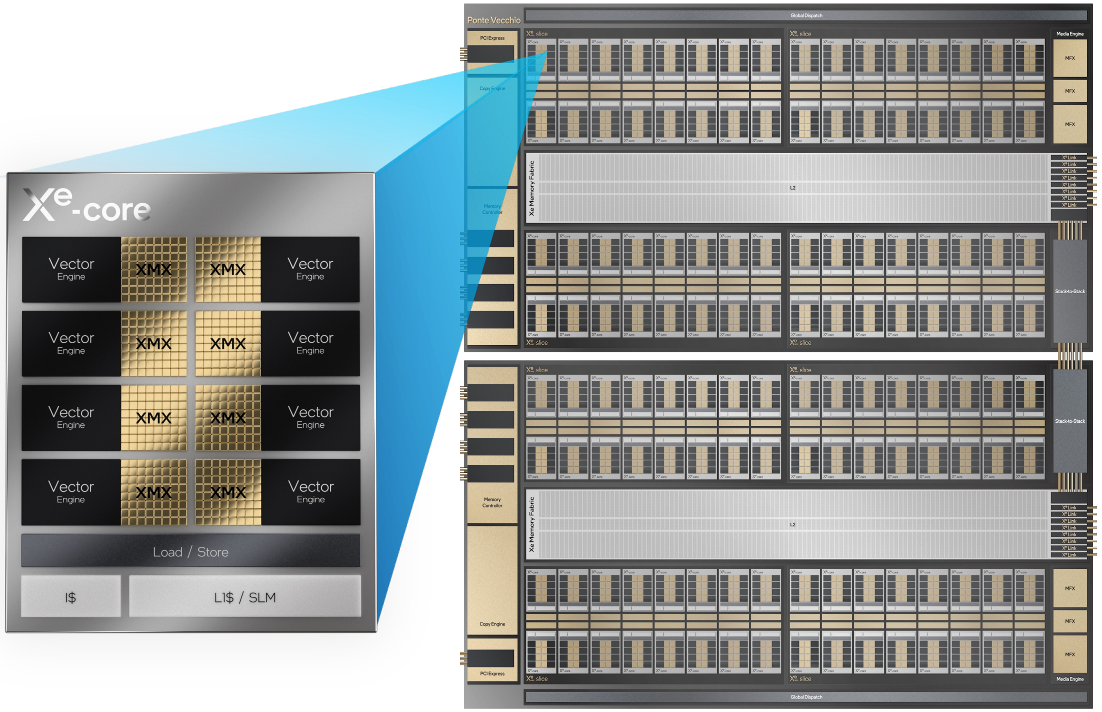
Intel GPU Max Series

Intel Datacenter GPU Max Series

General Compute Accelerator

X^e HPC **Architecture**

2 Stacks	128 X^e - cores 8 Hardware Contexts
8	HBM2e controllers
16	X^e Links



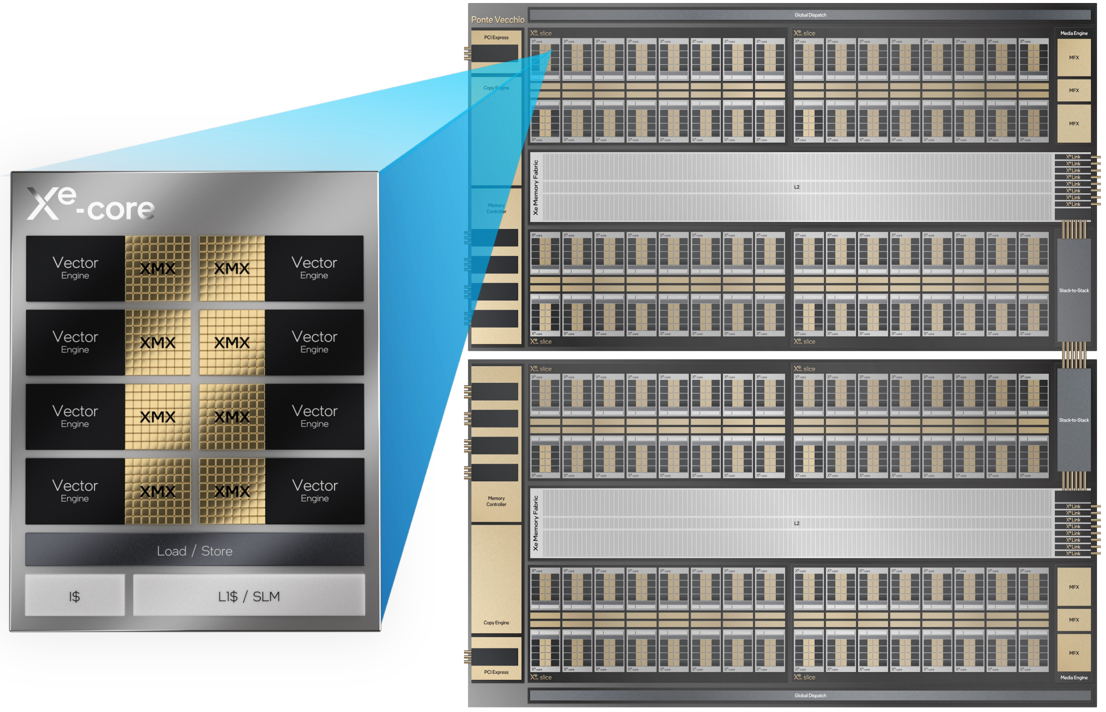
X^e core

Intel Datacenter GPU Max Series

Intel Datacenter GPU Max Series - Throughput

Peak Throughput	2-Stack GPU
FP64	52 TFLOPS
FP32	52 TFLOPS
XMX Float 32 (TF32)	419 TFLOPS
XMX BF16	839 TFLOPS
XMX FP16	839 TFLOPS
XMX INT8	1678 TOPS

XMX: X^e Matrix Extensions



X^e core

Intel Datacenter GPU Max Series

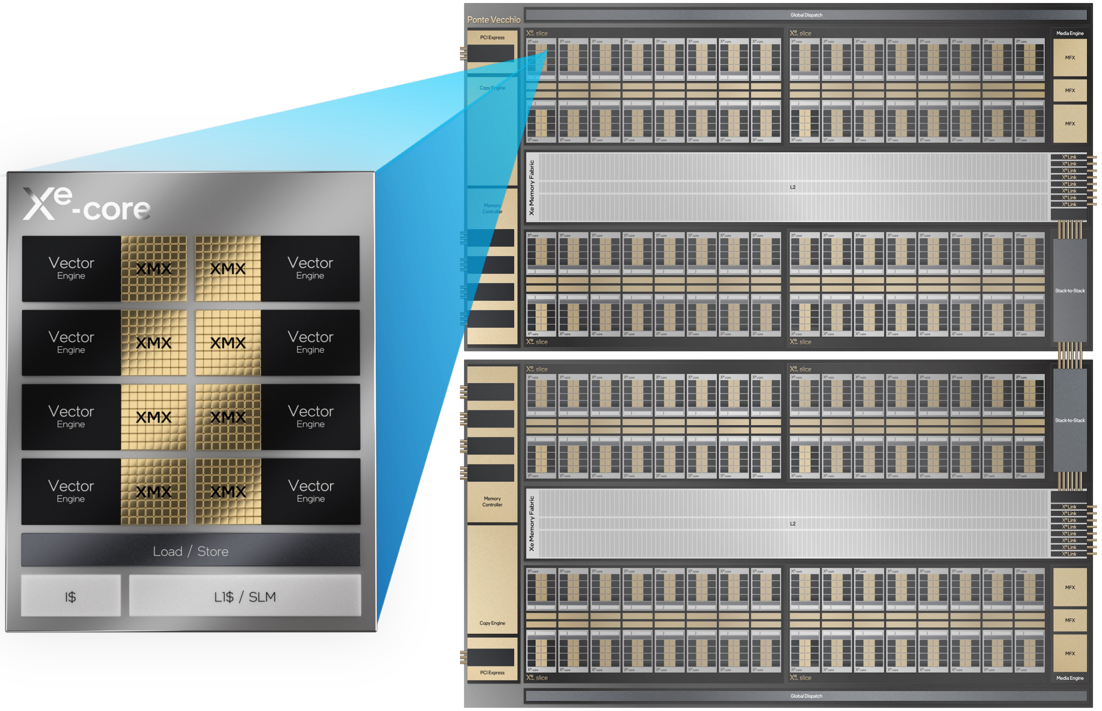
Intel Datacenter GPU Max Series - Memory Hierarchy

Large bandwidth and cache bring data close to compute

2-Stack GPU	Register File	L1 Cache	L2 Cache	HBM
Maximum Size	64 MB	64 MB	408 MB	128 GB
Peak Read Bandwidth	419 TB/s	105 TB/s	13 TB/s	3.2 TB/s

Maximum Size: 1:1 1:~6

Peak Read Bandwidth: 4:1 8:1 4:1



X^e core

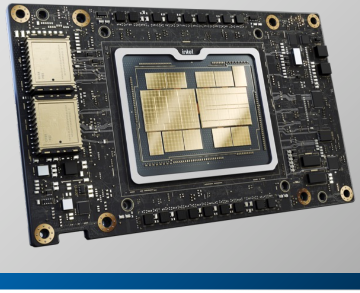
Intel Datacenter GPU Max Series

Accelerated Compute Systems

- x4 subsystem supports all-to-all connection across X^e Links
- OAMs support all-to-all topologies for both 4 GPU and 8 GPU platforms

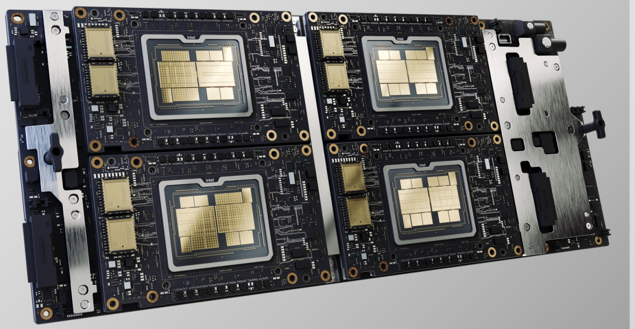
Intel Datacenter GPU Max Series

OAM



Intel Datacenter GPU Max Series

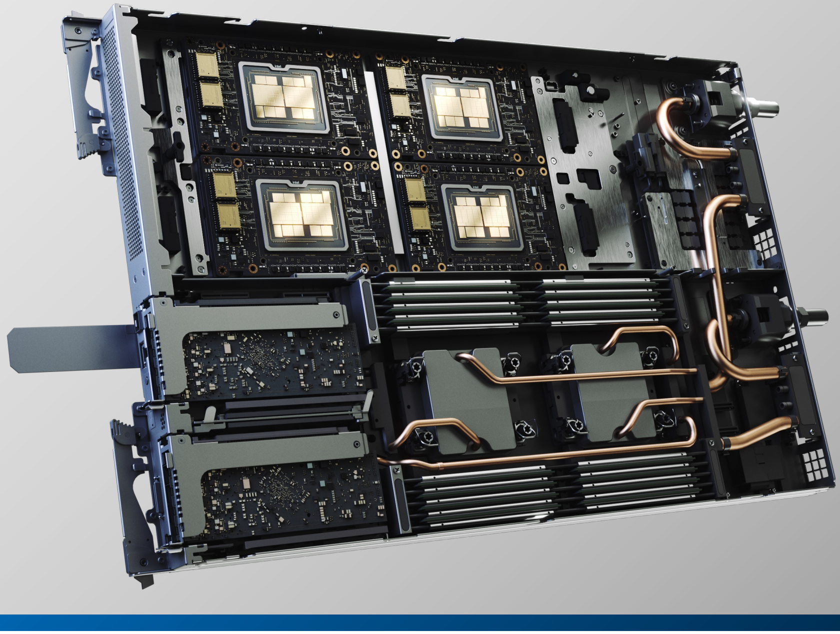
x4 Subsystem with X^e Links



Intel Datacenter GPU Max Series

x4 Subsystem with X^e Links

+ 2S 4th gen Intel Xeon Scalable processor



15+

Intel® Data Center GPU Max Series System Designs

From industry-leading solution providers

Hewlett Packard Enterprise

DELL Technologies

Lenovo

Atos

inspur

SUPERMICRO

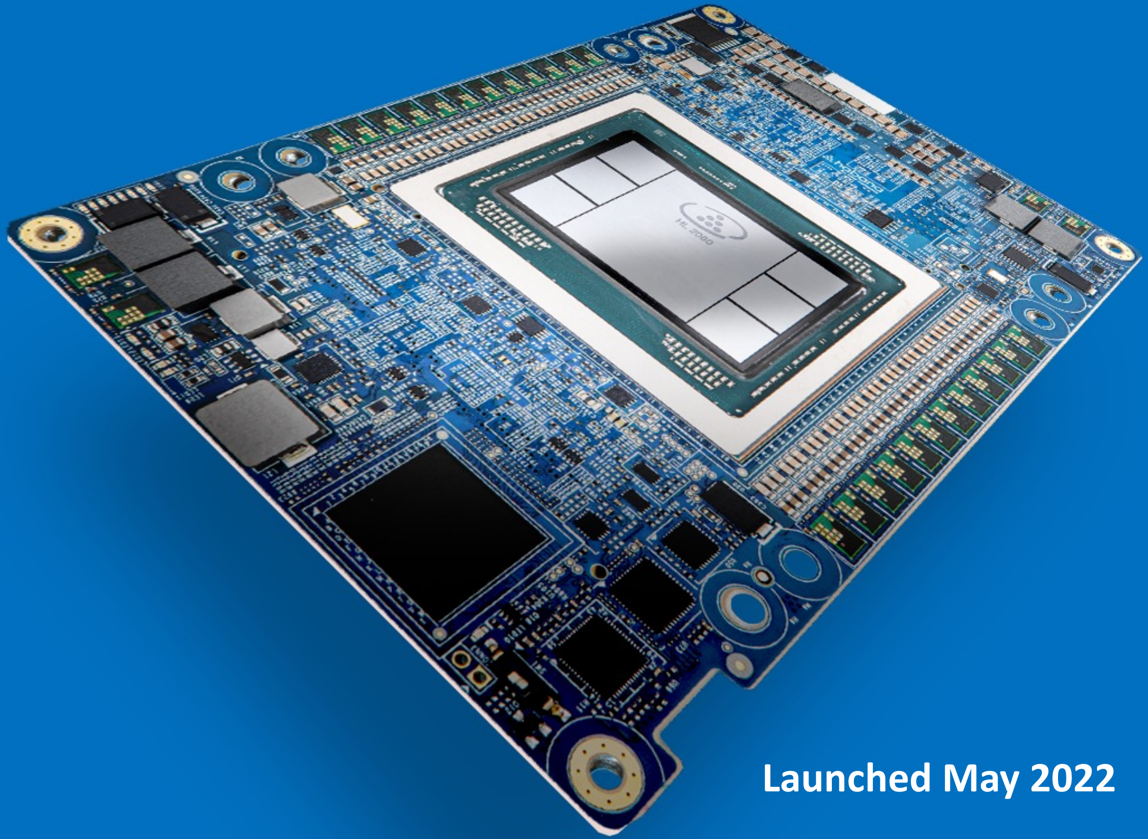
QCT

Intel Habana Gaudi

GAUDI®2

Purpose-built for Deep Learning

Leadership Performance
& Cost Performance Advantage



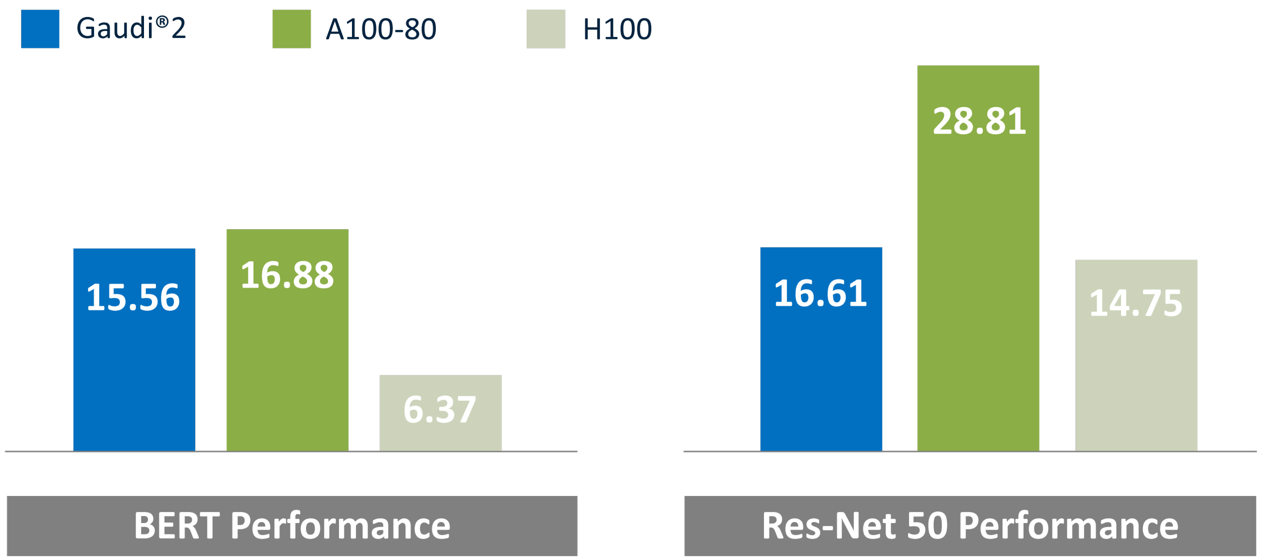
Launched May 2022

MLPerf 2.1 Gaudi2 vs Competition

Gaudi®2 Comparative Performance

Performance based on 8 AI processors

Time-to-train (minutes): lower is better



- Gaudi2 outperformed A100-80 GB for BERT and ResNet-50
- Habana results using standard BF16 datatype
  - H100 BERT result is generated with FP8
- Habana results in available category
  - H100 results in preview category
- Habana's MLPerf optimizations included in SynapseAI software releases*
  - Users can get out-of-box performance

Source: mlperf.org. Click here for results

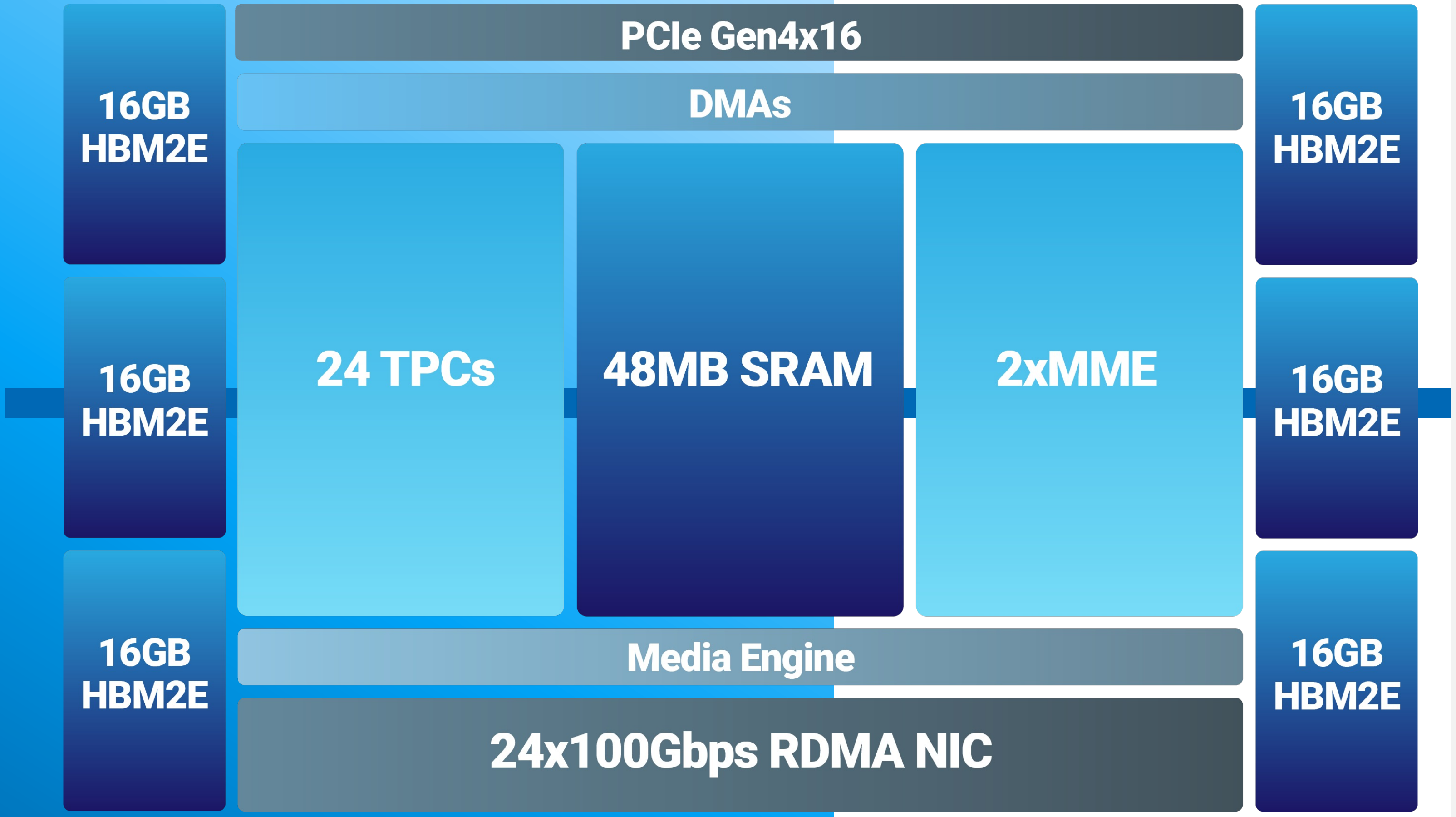
* MLPerf 2.1 related optimizations will be available in upcoming SynapseAI release (version 1.8.0)

An Array of Architectural Advances

GAUDI®2

Purpose-built to accelerate deep learning workloads

- Heterogeneous compute architecture enables high-efficiency on large DL workloads
- Software-managed memory architecture (HBM + SRAM + local memory)
- Integrates multiple 100Gb Ethernet RoCE ports on-chip for higher scaling efficiency
- Industry standard interfaces and no vendor lock-in



7nm process technology

Supermicro Gaudi2 On-premises

Supermicro Gaudi®2 AI Training Server

- Featuring 8 Gaudi2 processors
- Dual 3rd Gen Xeon Scalable processors
- 24 x 100 GbE integrated onto Gaudi2
- Available this quarter

https://www.supermicro.com/en/accelerators/intel#habana-gaudi-intro

Inspur x Gaudi2 On-premises

Gaudi®2
...second-
...designed for high-
...ent of large-scale workloads.
...compliant Mezzanine Cards, TDP 600W
Xeon® Sapphire Rapids Host CPUs, up to 350W
configuration, support 8 400G CX7, OCP and Smart NICs
AI Mezzanine
rator Module V1 .1 compliant
M2e memory at 2.45GB/sec bandwidth
AM
bit Ethernet RDMA over Converged Ethernet (RoCE) ports integrated on-chip
edia Processing
sec on ResNet-50 of Training throughput on single Gaudi2
rver
SONY
habana

inspur

- Inspur OAM Server with Gaudi2
- Dual 4th gen Intel® Xeon® processors (Sapphire Rapids)

https://www.inspursystems.com/blog/deepening-ai-training-inference-inspur-habana-labs-partnership/

HLS-Gaudi2

- Developed and deployed in Habana's R&D clusters
- Intended also for customer evals
- System also available from ODM Wiwynn

https://habana.ai/wp-content/uploads/2022/09/HLS-Gaudi2-Datasheet-Aug-2022.pdf

https://www.wiwynn.com/hubfs/Whitepapers/Future-Ready_Cooling_Solutions_Whitepaper_221013.pdf

Easily Get Started with TensorFlow Models

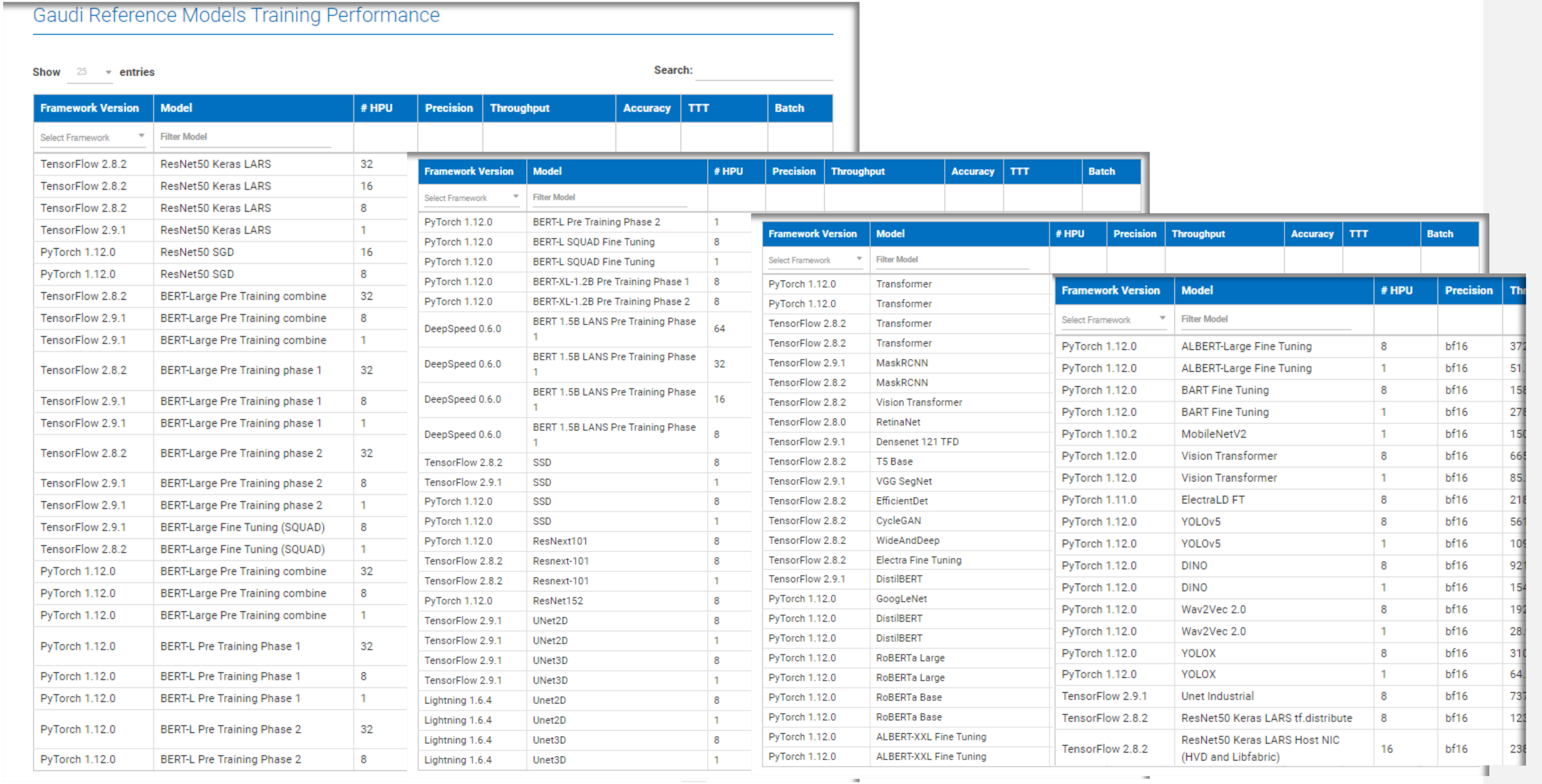
```
import tensorflow as tf

from TensorFlow.common.library_loader import load_habana_module
load_habana_module()

(x_train, y_train), (x_test, y_test) = tf.keras.datasets.mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
         tf.keras.layers.Flatten(input_shape=(28, 28)),
         tf.keras.layers.Dense(10),
])
loss = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)
optimizer = tf.keras.optimizers.SGD(learning_rate=0.01)

model.compile(optimizer=optimizer, loss=loss, metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5, batch_size=128)
model.evaluate(x_test, y_test)
```

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 use Gaudi

Gaudi Reference Models Training Performance



Show 25 entries

Search:

Framework Version	Model	# HPU	Precision	Throughput	Accuracy	TTT	Batch
Select Framework	Filter Model						
TensorFlow 2.8.2	ResNet50 Keras LARS	32					
TensorFlow 2.8.2	ResNet50 Keras LARS	16					
TensorFlow 2.8.2	ResNet50 Keras LARS	8					
TensorFlow 2.9.1	ResNet50 Keras LARS	1					
PyTorch 1.12.0	ResNet50 SGD	16					
PyTorch 1.12.0	ResNet50 SGD	8					
TensorFlow 2.8.2	BERT-Large Pre Training combine	32					
TensorFlow 2.9.1	BERT-Large Pre Training combine	8					
TensorFlow 2.9.1	BERT-Large Pre Training combine	1					
TensorFlow 2.8.2	BERT-Large Pre Training phase 1	32					
TensorFlow 2.9.1	BERT-Large Pre Training phase 1	8					
TensorFlow 2.9.1	BERT-Large Pre Training phase 1	1					
TensorFlow 2.8.2	BERT-Large Pre Training phase 2	32					
TensorFlow 2.9.1	BERT-Large Pre Training phase 2	8					
TensorFlow 2.9.1	BERT-Large Pre Training phase 2	1					
TensorFlow 2.9.1	BERT-Large Fine Tuning (SQUAD)	8					
TensorFlow 2.8.2	BERT-Large Fine Tuning (SQUAD)	1					
PyTorch 1.12.0	BERT-Large Pre Training combine	32					
PyTorch 1.12.0	BERT-Large Pre Training combine	8					
PyTorch 1.12.0	BERT-Large Pre Training combine	1					
PyTorch 1.12.0	BERT-L Pre Training Phase 1	32					
PyTorch 1.12.0	BERT-L Pre Training Phase 1	8					
PyTorch 1.12.0	BERT-L Pre Training Phase 1	1					
PyTorch 1.12.0	BERT-L Pre Training Phase 2	32					
PyTorch 1.12.0	BERT-L Pre Training Phase 2	8					

Framework Version	Model	# HPU	Precision	Throughput	Accuracy	TTT	Batch
Select Framework	Filter Model						
PyTorch 1.12.0	BERT-L Pre Training Phase 2	1					
PyTorch 1.12.0	BERT-L SQUAD Fine Tuning	8					
PyTorch 1.12.0	BERT-L SQUAD Fine Tuning	1					
PyTorch 1.12.0	BERT-XL-1.2B Pre Training Phase 1	8					
PyTorch 1.12.0	BERT-XL-1.2B Pre Training Phase 2	8					
DeepSpeed 0.6.0	BERT 1.5B LANS Pre Training Phase 1	64					
DeepSpeed 0.6.0	BERT 1.5B LANS Pre Training Phase 1	32					
DeepSpeed 0.6.0	BERT 1.5B LANS Pre Training Phase 1	16					
DeepSpeed 0.6.0	BERT 1.5B LANS Pre Training Phase 1	8					
TensorFlow 2.8.2	SSD	8					
TensorFlow 2.9.1	SSD	1					
PyTorch 1.12.0	SSD	8					
PyTorch 1.12.0	SSD	1					
PyTorch 1.12.0	ResNext101	8					
TensorFlow 2.8.2	Resnext-101	8					
TensorFlow 2.8.2	Resnext-101	1					
PyTorch 1.12.0	ResNet152	8					
TensorFlow 2.9.1	UNet2D	8					
TensorFlow 2.9.1	UNet2D	1					
TensorFlow 2.9.1	UNet3D	8					
TensorFlow 2.9.1	UNet3D	1					
Lightning 1.6.4	Unet2D	8					
Lightning 1.6.4	Unet2D	1					
Lightning 1.6.4	Unet3D	8					
Lightning 1.6.4	Unet3D	1					

Framework Version	Model	# HPU	Precision	Throughput	Accuracy	TTT	Batch
Select Framework	Filter Model						
PyTorch 1.12.0	Transformer						
PyTorch 1.12.0	Transformer						
TensorFlow 2.8.2	Transformer						
TensorFlow 2.8.2	Transformer						
TensorFlow 2.9.1	MaskRCNN						
TensorFlow 2.8.2	MaskRCNN						
TensorFlow 2.8.2	Vision Transformer						
TensorFlow 2.8.0	RetinaNet						
TensorFlow 2.9.1	Densenet 121 TFD						
TensorFlow 2.8.2	T5 Base						
TensorFlow 2.9.1	VGG SegNet						
TensorFlow 2.8.2	EfficientDet						
TensorFlow 2.8.2	CycleGAN						
TensorFlow 2.8.2	WideAndDeep						
TensorFlow 2.8.2	Electra Fine Tuning						
TensorFlow 2.9.1	DistilBERT						
PyTorch 1.12.0	GoogLeNet						
PyTorch 1.12.0	DistilBERT						
PyTorch 1.12.0	DistilBERT						
PyTorch 1.12.0	RoBERTa Large						
PyTorch 1.12.0	RoBERTa Large						
PyTorch 1.12.0	RoBERTa Base						
PyTorch 1.12.0	RoBERTa Base						
PyTorch 1.12.0	ALBERT-XXL Fine Tuning						
PyTorch 1.12.0	ALBERT-XXL Fine Tuning						

Framework Version	Model	# HPU	Precision	Th
Select Framework	Filter Model			
PyTorch 1.12.0	ALBERT-Large Fine Tuning	8	bf16	372
PyTorch 1.12.0	ALBERT-Large Fine Tuning	1	bf16	51.
PyTorch 1.12.0	BART Fine Tuning	8	bf16	158
PyTorch 1.12.0	BART Fine Tuning	1	bf16	278
PyTorch 1.10.2	MobileNetV2	1	bf16	150
PyTorch 1.12.0	Vision Transformer	8	bf16	665
PyTorch 1.12.0	Vision Transformer	1	bf16	85.
PyTorch 1.11.0	ElectraLD FT	8	bf16	218
PyTorch 1.12.0	YOLOv5	8	bf16	561
PyTorch 1.12.0	YOLOv5	1	bf16	109
PyTorch 1.12.0	DINO	8	bf16	921
PyTorch 1.12.0	DINO	1	bf16	154
PyTorch 1.12.0	Wav2Vec 2.0	8	bf16	192
PyTorch 1.12.0	Wav2Vec 2.0	1	bf16	28.
PyTorch 1.12.0	YOLOX	8	bf16	310
PyTorch 1.12.0	YOLOX	1	bf16	64.
TensorFlow 2.9.1	Unet Industrial	8	bf16	737
TensorFlow 2.8.2	ResNet50 Keras LARS tf.distribute	8	bf16	123
TensorFlow 2.8.2	ResNet50 Keras LARS Host NIC (HVD and Libfabric)	16	bf16	238

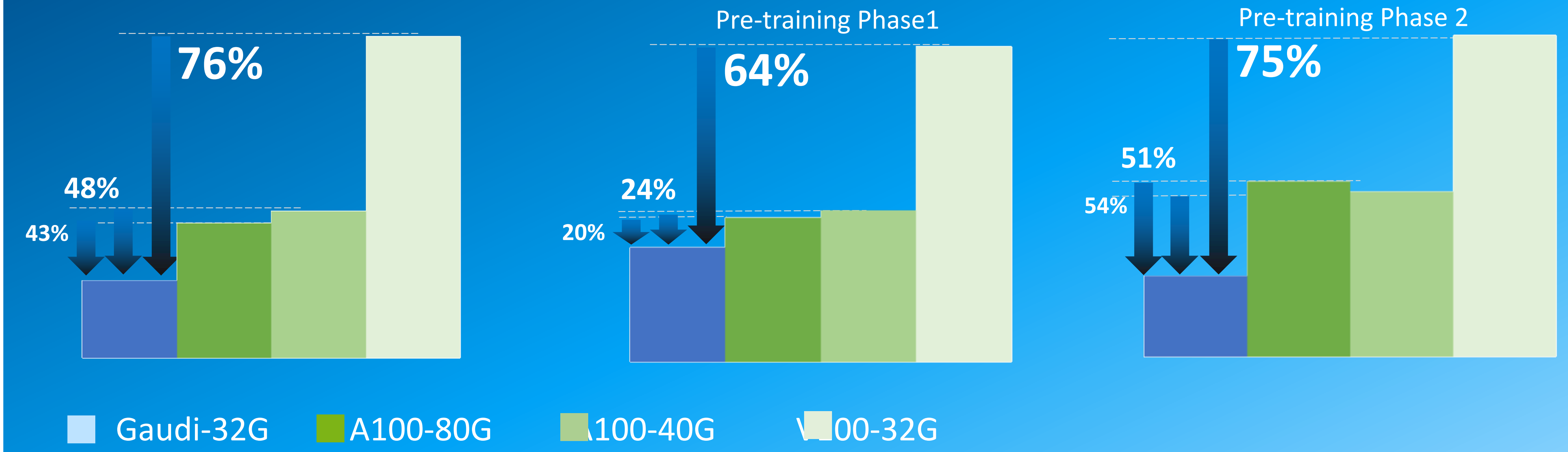
Customer Cost Savings on Amazon EC2 DL1 Instances

ResNet50 $/image

(lower is better)

BERT-Large $/seq

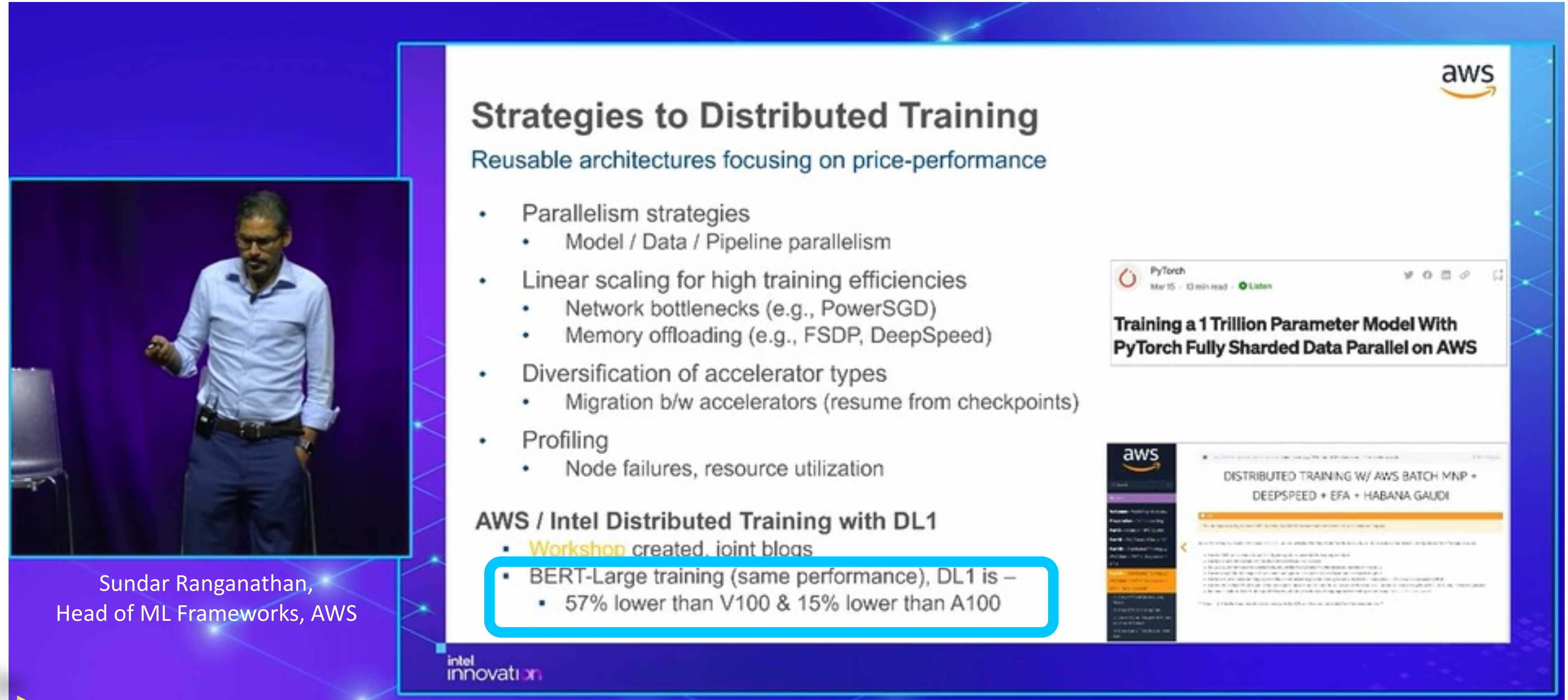
(lower is better)



Cost savings based on Amazon EC2 On-Demand pricing for P3dn, P4d, P4de and DL1 instances respectively. Performance data collected and measured using the following resources:

Habana BERT-Large Model: https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/nlp/bert
Habana ResNet50 Model: https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/computer_vision/Resnets/resnet_keras
Habana SynapseAI Container: https://vault.habana.ai/ui/repos/tree/General/gaudi-docker/1.7.0/ubuntu20.04/habanalabs/tensorflow-installer-tf-cpu-2.8.3
Habana Gaudi Performance: https://developer.habana.ai/resources/habana-training-models/
A100 / V100 Performance: https://ngc.nvidia.com/catalog/resources/nvidia:bert_for_tensorflow/performance, https://ngc.nvidia.com/catalog/resources/nvidia:resnet_50_v1_5_for_tensorflow/performance,
results published for DGX A100-40G and DGX V100-32G

AWS Distributed Training with DL1



Strategies to Distributed Training

Reusable architectures focusing on price-performance

- Parallelism strategies
  - Model / Data / Pipeline parallelism
- Linear scaling for high training efficiencies
  - Network bottlenecks (e.g., PowerSGD)
  - Memory offloading (e.g., FSDP, DeepSpeed)
- Diversification of accelerator types
  - Migration b/w accelerators (resume from checkpoints)
- Profiling
  - Node failures, resource utilization

AWS / Intel Distributed Training with DL1

- Workshop created, joint blogs
- BERT-Large training (same performance), DL1 is –
  - 57% lower than V100 & 15% lower than A100

PyTorch
Mar 15 · 13 min read · Listen

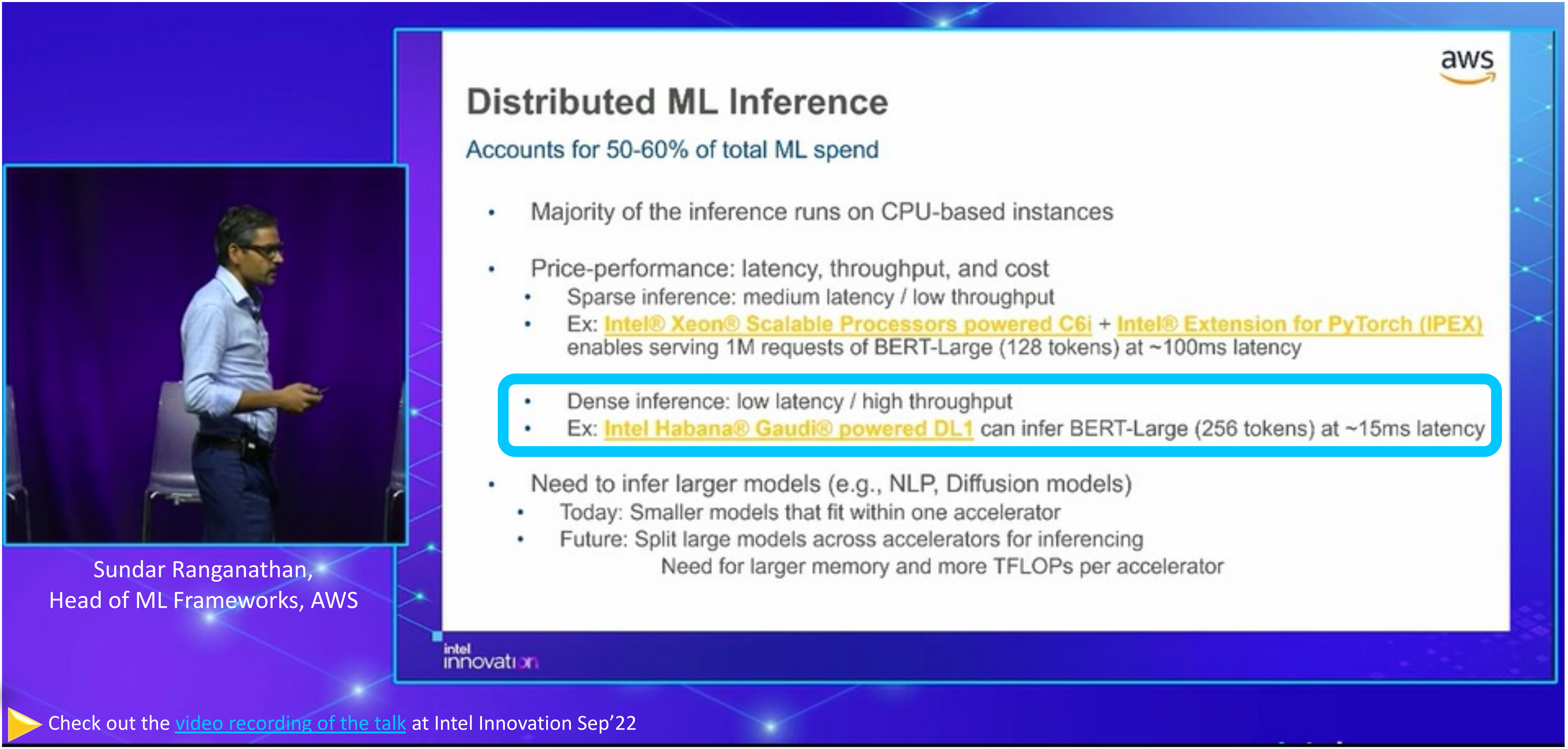
Training a 1 Trillion Parameter Model With PyTorch Fully Sharded Data Parallel on AWS

DISTRIBUTED TRAINING W/ AWS BATCH MNP + DEEPSPEED + EFA + HABANA GAUDI

Sundar Ranganathan,
Head of ML Frameworks, AWS

Check out the video recording of the talk at Intel Innovation Sep’22

AWS Distributed Inference with DL1



aws

Distributed ML Inference

Accounts for 50-60% of total ML spend

- Majority of the inference runs on CPU-based instances
- Price-performance: latency, throughput, and cost
  - Sparse inference: medium latency / low throughput
  - Ex: Intel® Xeon® Scalable Processors powered C6i + Intel® Extension for PyTorch (IPEX) enables serving 1M requests of BERT-Large (128 tokens) at ~100ms latency
  - Dense inference: low latency / high throughput
  - Ex: Intel Habana® Gaudi® powered DL1 can infer BERT-Large (256 tokens) at ~15ms latency
- Need to infer larger models (e.g., NLP, Diffusion models)
  - Today: Smaller models that fit within one accelerator
  - Future: Split large models across accelerators for inferencing

    Need for larger memory and more TFLOPs per accelerator

intel innovation

Sundar Ranganathan,
Head of ML Frameworks, AWS

Check out the video recording of the talk at Intel Innovation Sep'22

intel.

Detecting COVID19 in Frontal Chest X-ray Images

> 60% cost savings with DL1 vs. p3dn.24xlarge

leidos

"The rapid-pace R&D required to tame COVID demonstrates an urgent need our medical and health sciences customers have for fast, efficient deep learning training of medical imaging data sets--when hours and even minutes count—to unlock disease causes and cures. We expect Gaudi2, building on the speed and cost-efficiency of Gaudi1, to provide customers with dramatically accelerated model training, while preserving the DL efficiency we experienced with first-gen Gaudi."

Chetan Paul, CTO Health and Human Services at Leidos

Mobileye

Custom object detection

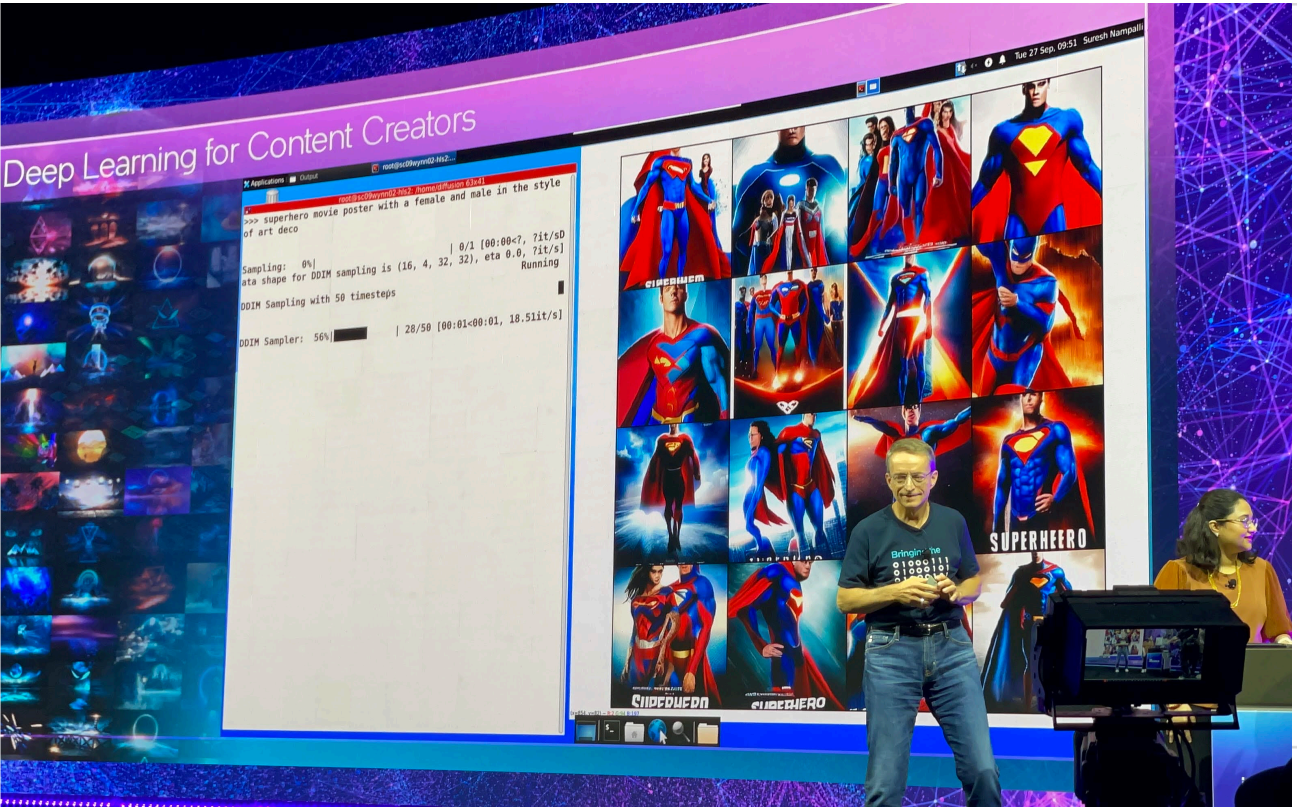
(2D and 3D) models trained on Gaudi

mobileye
An Intel Company

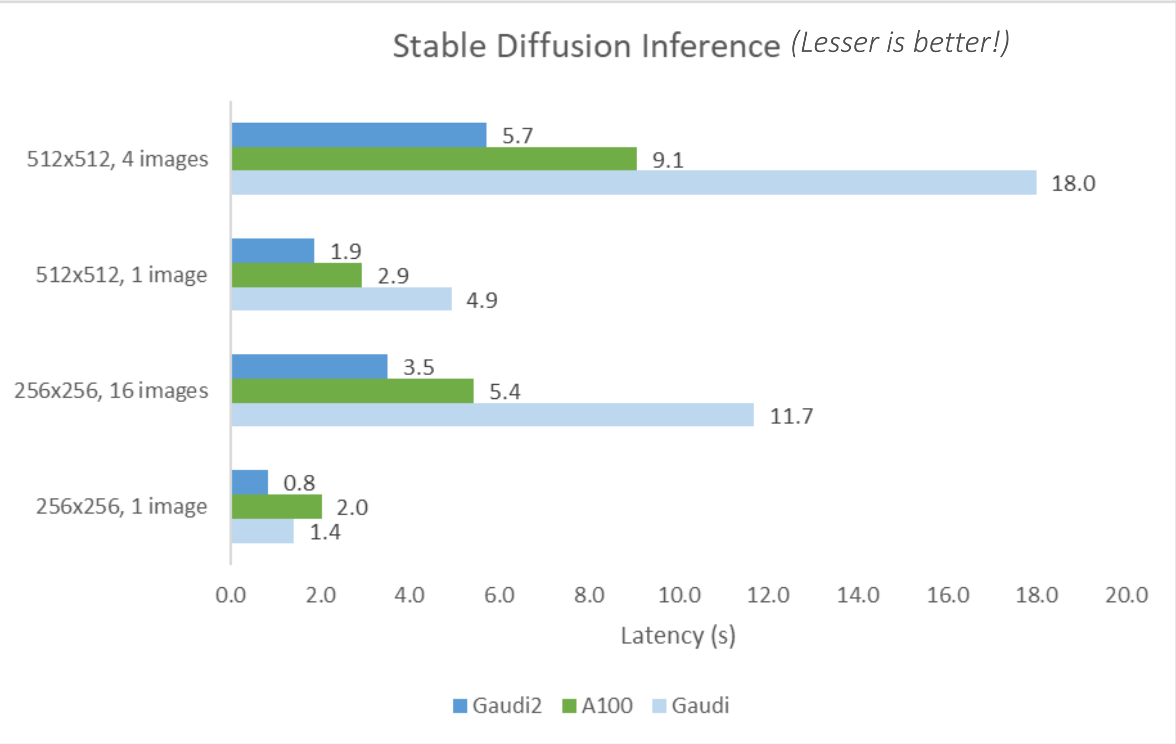
"On our own models the increase in price performance met and even exceeded the published 40% mark."

Chaim Rand, Mobileye

Diffusion Model Inference on Gaudi


Deep Learning for Content Creators
Applications
Output
root@sc09wynn02-hls2:...
root@sc09wynn02-hls2: /home/diffusion 63x41
Tue 27 Sep, 09:51 Suresh Nampalli
SUPERHEERO

```
>>> superhero movie poster with a female and male in the style
of art deco
                                        | 0/1 [00:00<?, ?it/sD
Sampling:   0%|
ata shape for DDIM sampling is (16, 4, 32, 32), eta 0.0, ?it/s]
                                                       Running
DDIM Sampling with 50 timesteps

DDIM Sampler:  56%|█████      | 28/50 [00:01<00:01, 18.51it/s]
```



Stable Diffusion Model based on https://github.com/pesser/stable-diffusion

Check out Pat Gelsinger's keynote featuring Gaudi2 stable diffusion demo at Intel Innovation in Sep'22

Multi-modal Deep Learning on Gaudi

Large scale models no longer limited to language
Foundation models now handle multiple input modalities (vision + language)
SynapseAI supports training and inference

- Multi-modal Understanding with Transformer-based models
  - Bridge-Tower model (MSFT Research & Intel Labs) trained on **512x Gaudi**
  - Video Retrieval Using Multilingual Knowledge Transfer (Intel Labs & UNC Chapel Hill)

- Multi-modal Generation with Diffusion-based models
  - V-diffusion
  - K-diffusion
  - Stable diffusion

new

Hugging Face + intel

Democratizing Accelerated Transformers

on Intel Platforms

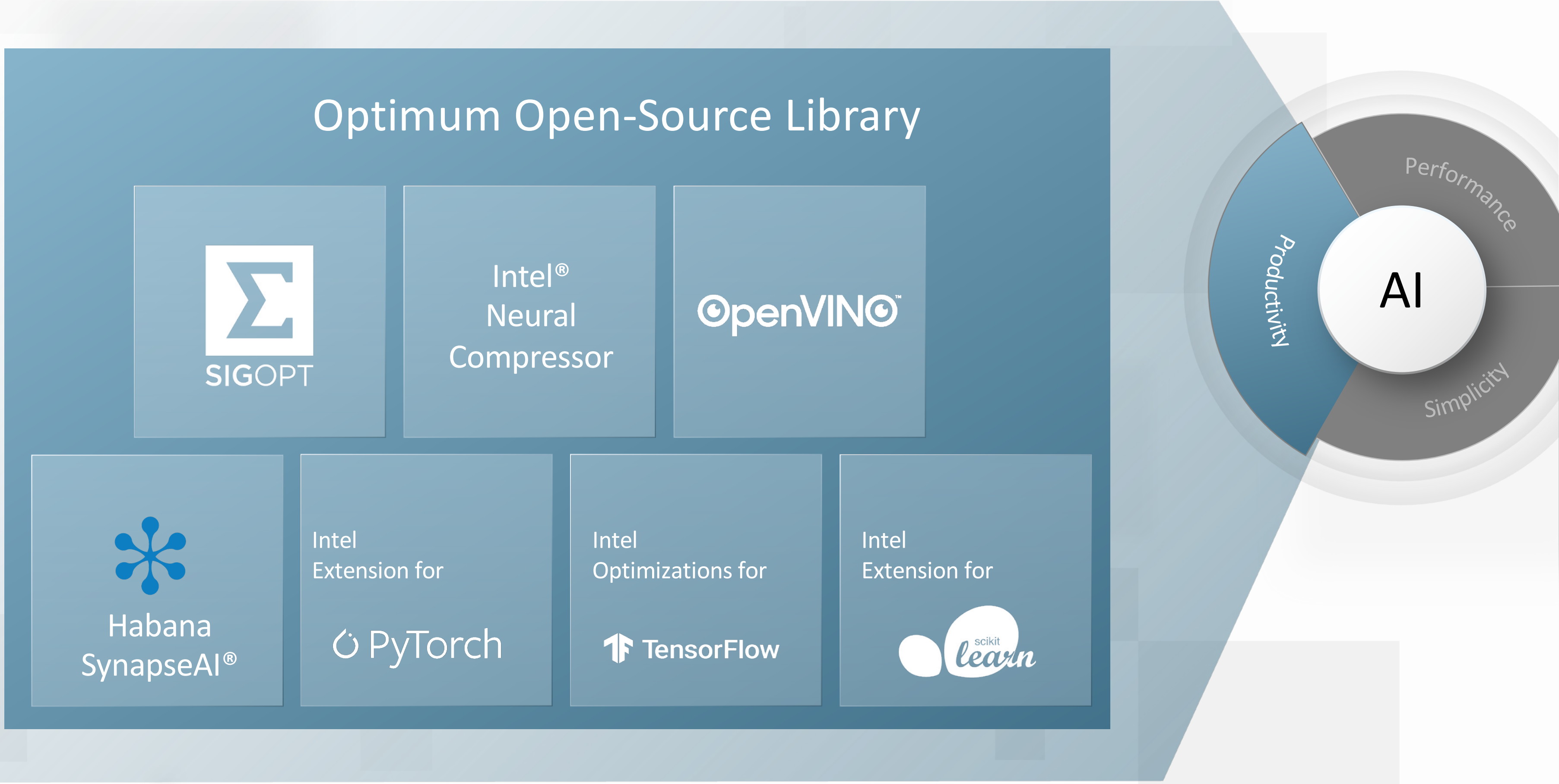
Inference Optimization Process from Days to Hours with Up to 4x Performance Speedup

intel. xeon

Distributed Training through efficient compute scaling

GAUDI

habana An Intel Company



Optimum Open-Source Library

SIGOPT

Intel® Neural Compressor

OpenVINO™

Habana SynapseAI®

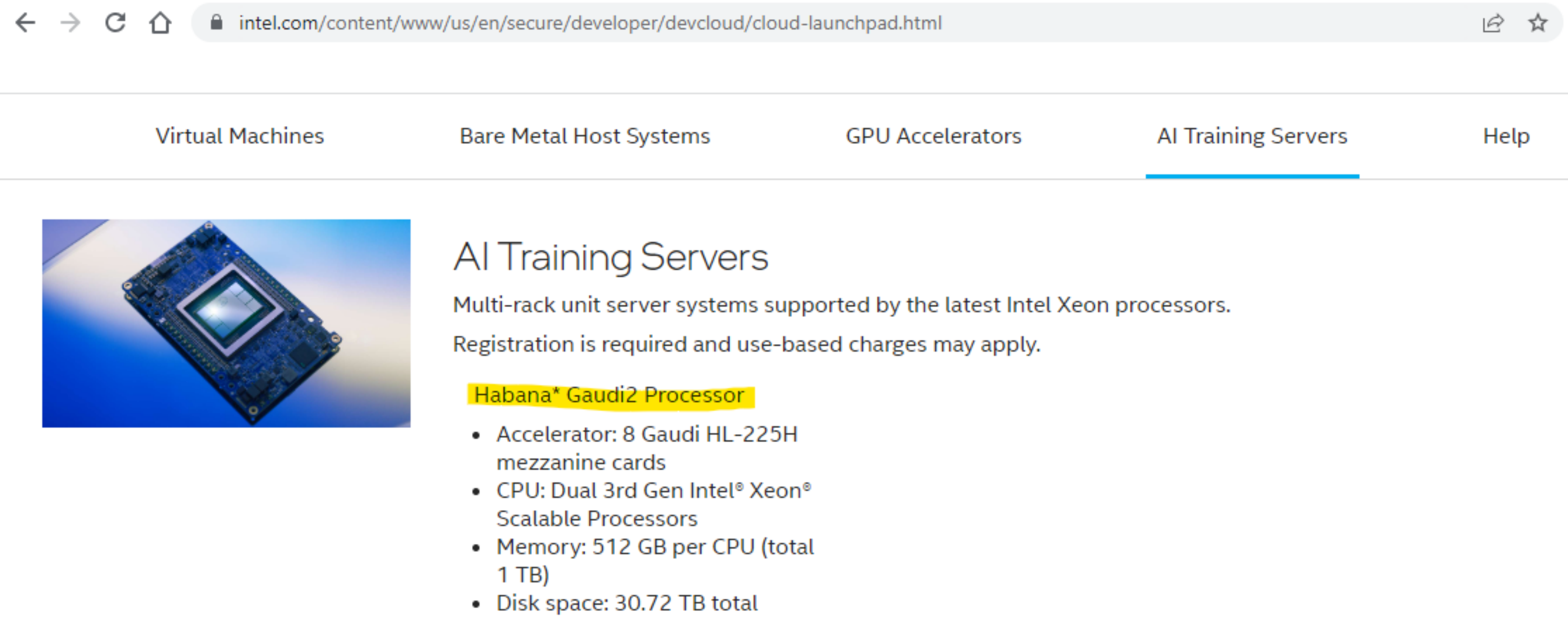
Intel Extension for PyTorch

Intel Optimizations for TensorFlow

Intel Extension for scikit learn

* Other names and brands may be claimed as the property of others

Gaudi 2 Processors Now Available on Intel DevCloud



AI Training Servers

Multi-rack unit server systems supported by the latest Intel Xeon processors.

Registration is required and use-based charges may apply.

Habana* Gaudi2 Processor

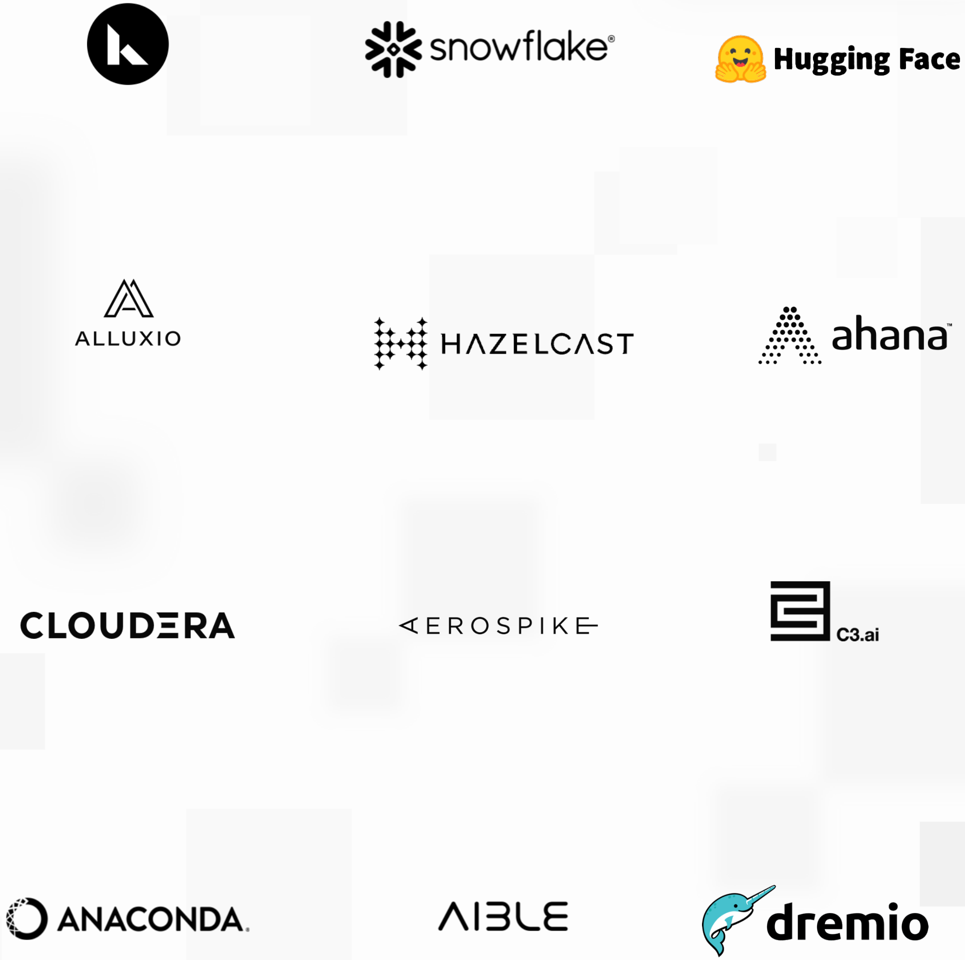
- Accelerator: 8 Gaudi HL-225H mezzanine cards
- CPU: Dual 3rd Gen Intel® Xeon® Scalable Processors
- Memory: 512 GB per CPU (total 1 TB)
- Disk space: 30.72 TB total

Ecosystem Programs

Intel® Disruptor Initiative

The Intel Disruptor Initiative participants are companies that are pushing the limits of innovation. Intel supports its members by driving growth through technical enablement and multi-channel go to market activities.

+ Many Additional Participants



https://www.intel.com/content/www/us/en/partner-alliance/membership/select-benefit/disruptors/overview.html

intel
AI Builders

AI Solutions Accelerated by Intel

Vertical Partners

Retail
- AiHello
- PROPORTUNITY
- nine TECH
- W locaweb
- ReadSense
- Build
- SKY REC
- MALONG
- GIGASPACES
- bigdatacorp
- ARDIC

Healthcare
- Laura
- SUBTLE MEDICAL
- lu
- DATACOGN
- SIG TUPLE
- John Snow LABS
- JP健信
- OrboGraph
- Carestream
- SARADA
- HEARTVISTA
- ubotica

BFSI
- nuveo
- IoT
- NEXT labs
- Alpaca
- Allgoo
- Phrase
- Fluid AI
- WorkFusion

Transportation
- WI-TRONIX
- GLOBAL EDGE
- Playment
- ERA

Media, Travel & Entertainment
- WINIMY.AI
- ZIVA
- pimloc
- Gramener
- keemotion
- Taboola
- ShotTracker
- sensifai

Sec & Govt
- DEEPGLINT
- POLSEC
- simbiose social

SW Tools & Services
- IFDAQ
- deepsense.ai
- UiPath
- ittiam
- ANACONDA

Prof. Services
- seedlink
- finch soluções
- kenoby
- readr
- LEGAL LABS
- voiceitt

Agri
- LeBov
- tbit
- Asquared IoT
- Aerobotics

Telecom
- TATA CONSULTANCY SUBEX
- Aricent
- CertusNet

Horizontal Partners

- COLFAX
- inspur
- EXXACT
- EQUUS COMPUTE SOLUTIONS
- Lenovo
- NCR
- Lambda
- L&T Technology Services
- accenture
- Mobiliya
- HCL
- NTT DATA
- wipro
- Tech Mahindra
- ALTOROS
- valtech_
- DELL EMC
- hp

BI & Analytics
- PROPHETSTOR
- SIGOPT
- DataRobot
- bluedata
- skymind
- MACNICA DHW

Vision
- altia SYSTEMS
- Matroid
- imagga
- restb.ai
- CrowdAI
- ENTROPIK TECH
- VISIONGENE
- gestoos

Conv. Bots
- avaamo
- SET SAIL
- [24]7.ai
- nama

NLU/NLP
- JQUANT
- mihup
- verbio
- 云知声 Unisound
- AIS
- FORMCEPT
- gnani.ai
- Aimirim
- gamalon

AI Tools & Consulting
- Mighty AI
- ALEGION
- TIVIT
- QuikFynd
- LEAPMIND
- iMerit
- minds.ai
- Julia computing
- DARWIN AI
- nologin
- KUNUMI
- axondata
- cloudera

AI PaaS
- KuberLab
- H2O.ai
- Paperspace
- LA DRONES
- NIMBIX
- DOMINO
- RocketML
- JETWARE
- clusterone

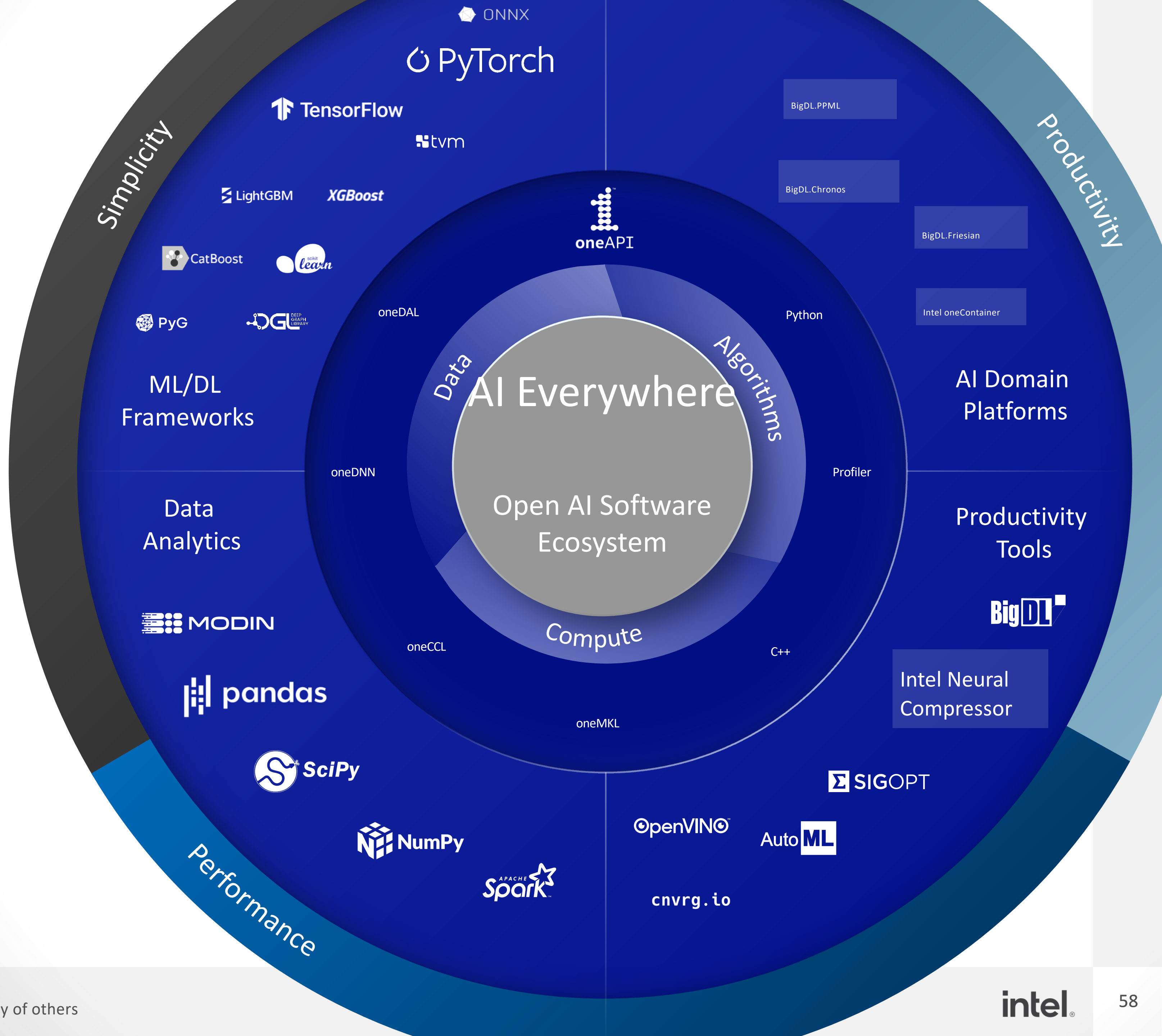
Big Data
- moBagel
- labs CLIMATEMPO

* Other names and brands may be claimed as the property of others

intel

Let’s work together to bring AI Everywhere

Visit **developer.intel.com/ai** for more info



Simplicity

Productivity

Performance

ONNX

PyTorch

TensorFlow

tvm

LightGBM

XGBoost

CatBoost

learn

PyG

DGL

ML/DL Frameworks

Data Analytics

MODIN

pandas

SciPy

NumPy

Spark

oneAPI

oneDAL

oneDNN

oneCCL

oneMKL

Python

Profiler

C++

Data

Algorithms

Compute

AI Everywhere

Open AI Software Ecosystem

BigDL.PPML

BigDL.Chronos

BigDL.Friesian

Intel oneContainer

AI Domain Platforms

Productivity Tools

BigDL

Intel Neural Compressor

SIGOPT

OpenVINO

AutoML

cnvrg.io

* Other names and brands may be claimed as the property of others

intel®