Choose the Best Accelerated Technology

AI on Intel Architecture

Akash Dhamasia – AI Software Solutions Engineer
akash.dhamasia@intel.com
March 7th 2024

intel.

Agenda

- Introduction to Intel® AI
  - Intel® Hardware for AI
  - Some Usecases
  - Demo
  - Intel AI® Software Stack
- Conclusion

Notices & Disclaimers

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

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

Your costs and results may vary.

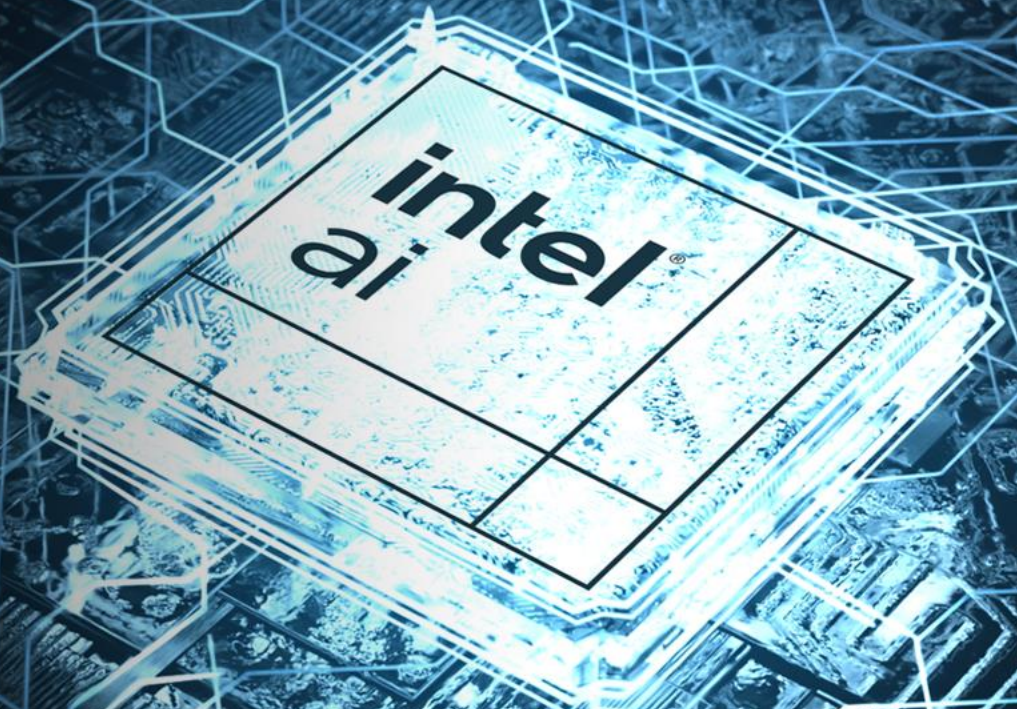
Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Introduction to Intel® AI

The AI Revolution Is Happening.

And you're already ready.



MACHINE LEARNING , DEEP LEARNING & BEYOND

CLASSICAL MACHINE LEARNING

How do you engineer the best features?

$N \times N$

$(f_1, f_2, \ldots, f_K)$

Roundness of face
Distance between eyes
Nose width
Eye socket depth
Cheek bone structure
Jaw line length
Etc.

CLASSIFIER ALGORITHM

SVM
Random Forest
Naïve Bayes
Decision Trees
Logistic Regression
Ensemble methods

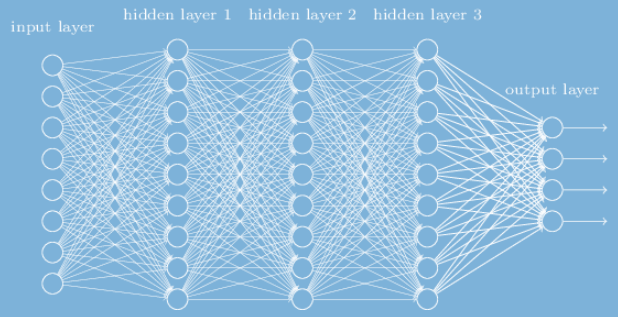
Akash

DEEP LEARNING & BEYOND

How do you guide the model to find the best features?

$N \times N$

NEURAL NETWORK



Akash

AI
Model Support

Computer Vision Image Classification	ResNet-50 v15	ResNeXt-101	ResNet-101	EfficientNet-B7	SE-ResNeXt50	TSM	
	Adorym	CosmoFlow	RegNetY-32Y	ResNeXt-101	Candle Uno	Swin Transformer	
Image Segmentation	Cosmic Tagger	Mask R-CNN	DenseNet169	FFN	3D-Unet		
	PointNet	DeepCAM	DRN-D-54	ResNeXt3D-101			
Object Detection	SSD-ResNet34	SSD-ResNet50	EfficientDet	ShuffleNet	YOLO-v3	YOLO-v4	RetinaNet-ResNet50
	Deep Fusion	CascadeRCNN-	MobileNet v3	SSD-MobileNet	MMA	ResNet101-FPN	
NLP Language Modeling	BERT-Large	Stable Diffusion	ALBERT	FastFormers	Transformer-LT	Big Bird	Faster Transformer
	BERT-base	GP-J	BLOOM	DistilBERT	RoBERTa	XLNet	
Speech Recognition	RNN-T	LAS - Listen Attend & Smell	Wave2Vec	QuartzNet			
Speech Synthesis	FastSpeech2	Tacotron-2 with LPCNet					
Recommendation	DLRM	DSSM	ESSM	Wide & Deep	DeepFM		
	DIN	AttRec	DIEN	MMOE			

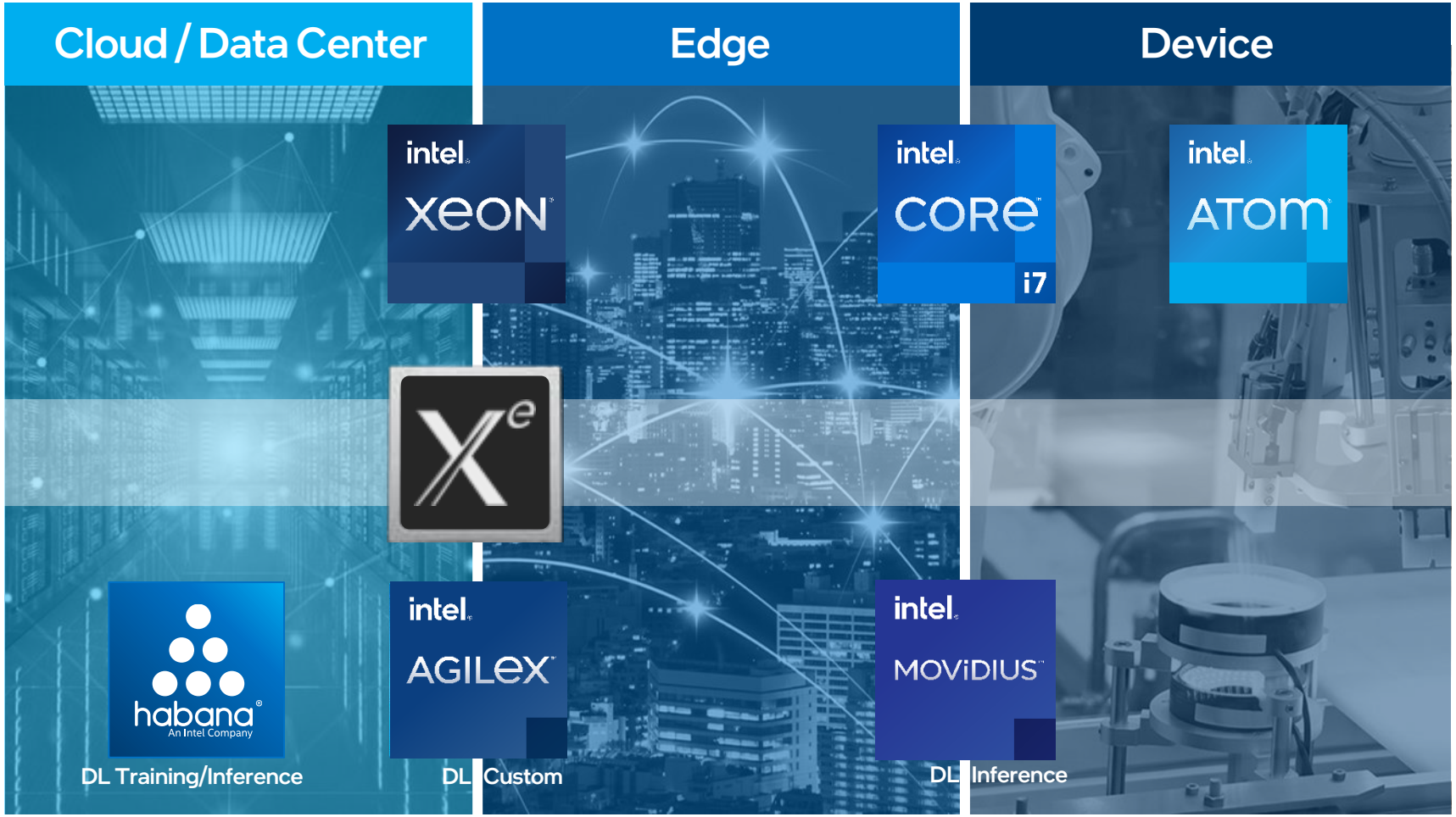
intel.

Flexible AI Acceleration

CPU *only*
Built-in AI acceleration for mainstream AI use cases

CPU + GPU
When compute is dominated by AI, HPC, graphics, and/or real-time media

CPU + custom
When compute is dominated by deep learning (DL)



Intel® AI Portfolio

Open Software Environment	oneAPI, OpenVINO	PyTorch, TensorFlow, ONNX RUNTIME, deepspeed, XGBoost, pandas
Deep Learning Acceleration	intel habana Gaudi	Gaudi: Dedicated Deep Learning Training and Inference
General Acceleration	intel DATA CENTER GPU FLEX SERIES	Cloud Gaming, VDI, Media Analytics, Real-Time Dense Video
	intel DATA CENTER GPU MAX SERIES	Parallel Compute, HPC, AI for HPC
General Purpose	intel xeon, intel xeon MAX SERIES, intel ethernet	Real-Time, Medium Throughput, Low Latency, and Sparse Inference
	intel xeon, intel xeon MAX SERIES, intel xeon W, intel ethernet	Medium to Small Scale Training and Fine Tuning
	intel core ULTRA, intel xeon D, intel DATA CENTER GPU FLEX SERIES, intel ARC GRAPHICS, intel core, intel ethernet	Edge and Network AI Inference
	intel core, intel core ULTRA w/VPU, intel ARC GRAPHICS	Client AI Usages

Intel® Hardware for AI

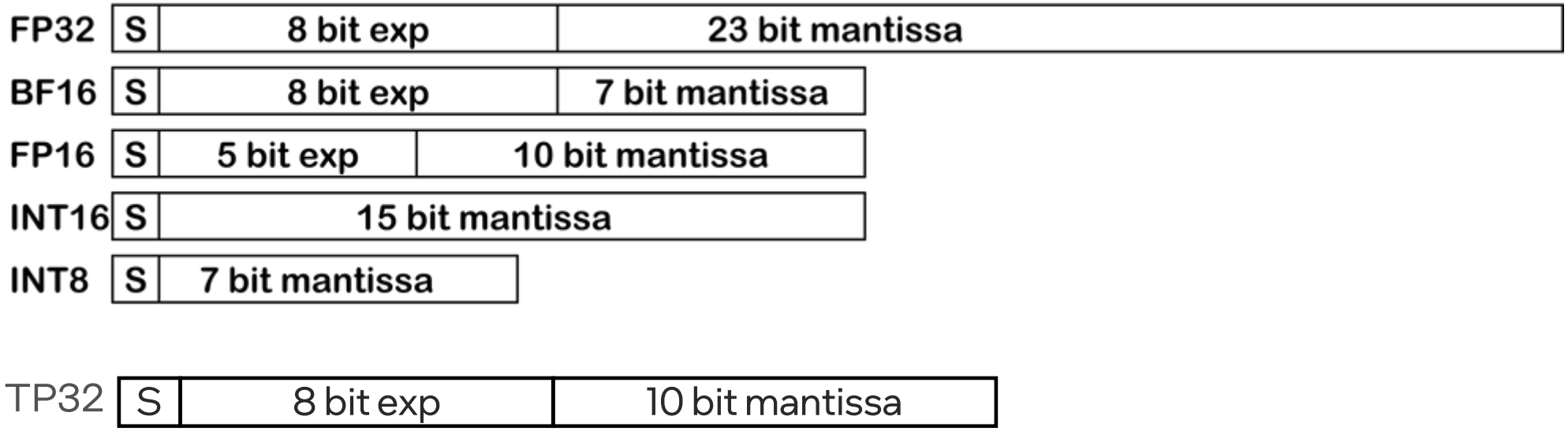
Data Precision

Data Precision

- Data precision:

  Number of bits used to store numerical values in memory

- Commonly found types of precision in Deep Learning:



Ref. https://www.intel.com/content/www/us/en/developer/videos/how-to-use-new-built-in-acceleration-engines.html#gs.wige3o

INT8/BF16 on Artificial intelligence/Machine Learning

- F32 is the default datatypes used in AI/ML for inference, which has a high memory footprint and higher latency.

- Low-precision models are faster in computation. To optimize and support these:
  - HW needs special features/instructions
  - Intel provide those in the form of Intel AMX/Intel XMX.

- SYCL Joint Matrix is the coding abstraction to invoke Intel AMX/Intel XMX, which ensures portability and performance of the code

Ref. https://www.intel.com/content/www/us/en/developer/videos/how-to-use-new-built-in-acceleration-engines.html#gs.wige3o

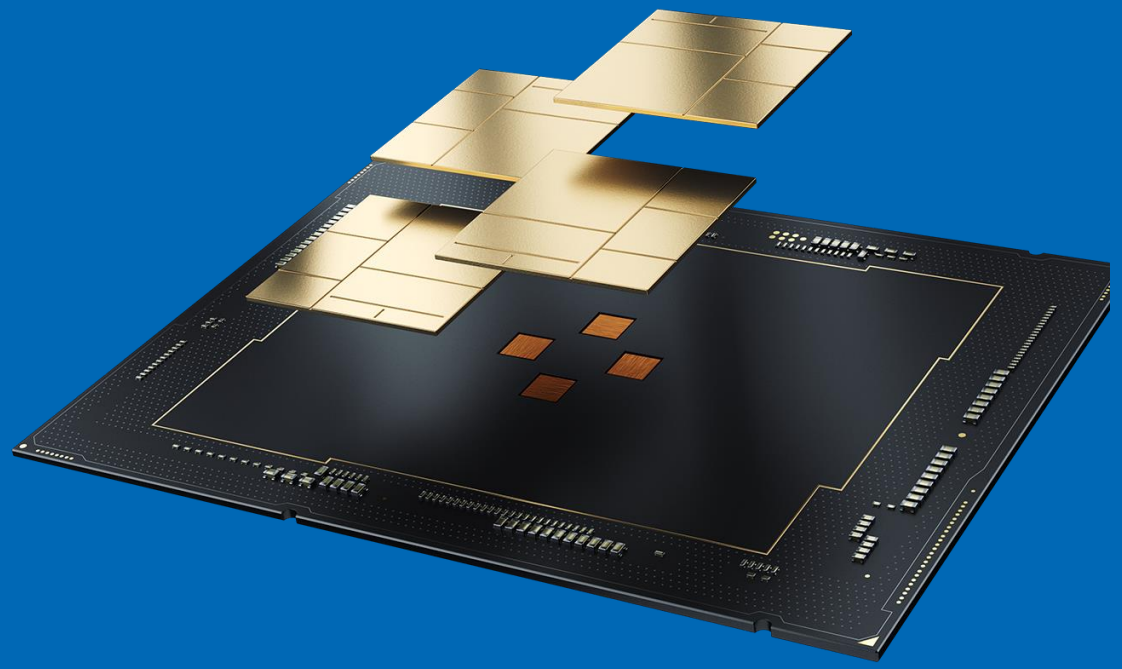
Introduction to Intel® Advanced Matrix Extension and Intel® Xe Matrix Extensions

Instruction Set	**Hardware support**	Description
Intel® AMX	Intel® Xeon 4th Generation Scalable CPUs (Formerly code-named Sapphire Rapids)	Intel® Advanced Matrix Extension are extensions to the x86 instruction set architecture (ISA) for microprocessors using 2-dimensional registers called tiles upon which accelerators can perform operations. Supports INT8/BF16
Intel® XMX	Intel® Data Center GPU Max (Formerly code-named Ponte Vecchio) or Intel® Data Center GPU Flex Series	Intel® Xe Matrix Extensions also known as DPAS specializes in executing dot product and accumulate instructions on 2D systolic arrays Supports U8,S8,U4,S4,U2,S2, INT8 FP16, BF16, TF32

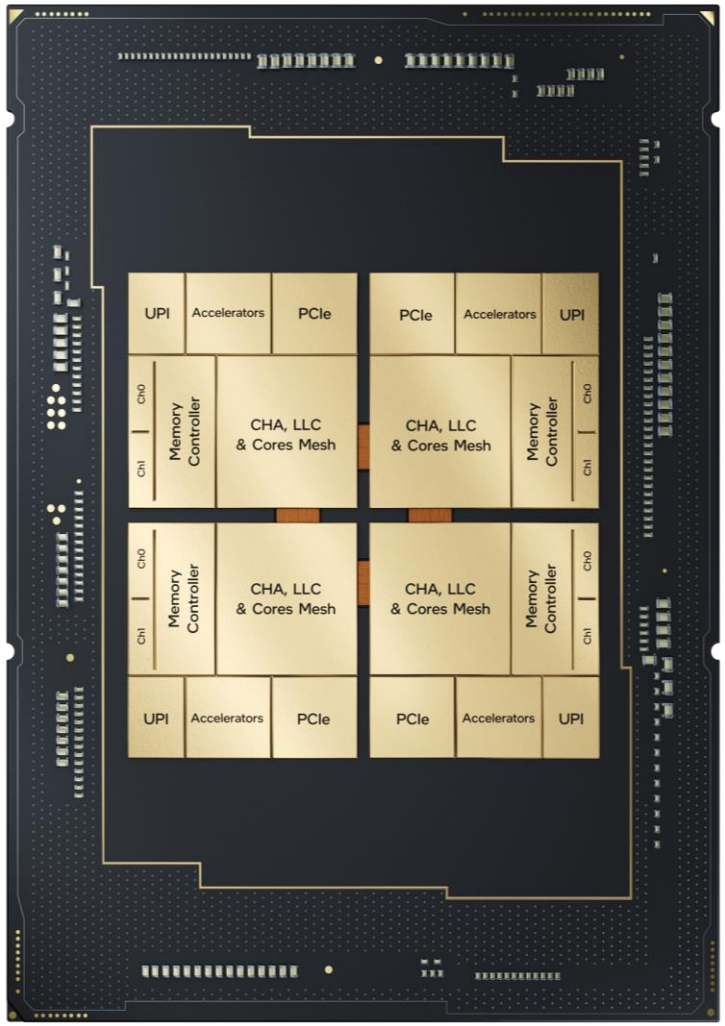
Both these Instruction Sets require Intel® oneAPI Base Toolkit 2023.0.0 and above for compilation

Ref. https://www.intel.com/content/www/us/en/developer/videos/how-to-use-new-built-in-acceleration-engines.html#gs.wige3o

4th Gen Intel® Xeon® Scalable Processor (Sapphire Rapids)



4th Generation Intel® Xeon® Scalable Processor
codenamed **Sapphire Rapids**



3-10x speedup and 7.7x performance/watt[1]
for INT8/ BF16 models
with Built-In AI Acceleration

4th Gen Intel® Xeon® Scalable Processor with Intel® Advanced Matrix Extensions acceleration vs. 3rd Gen Intel® Xeon® Scalable Processors

Intel® AI software
300+ DL Models
50+ optimize ML and Graph Models
Optimizations up-streamed
Intel® AI Developer Tools

2x PCI Express 5.0 Bandwidth
Compared to 3rd Gen Intel® Xeon® Scalable Processors

OneAPI AI Ecosystem
Use any popular DL, ML, and Data processing library and framework, operating system, and virtual machine manager

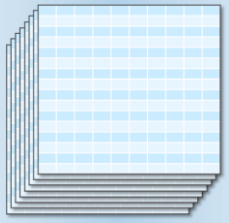
1.5x DDR5 Memory Bandwidth and Capacity
Compared to 3rd Gen Intel® Xeon® Scalable Processors

Up to 512 GB/Socket Protected Memory Enclave—Intel® Security Guard Extensions
Confidential AI supported in BigDL and OpenVINO™ toolkit

1 See [A16, A17, A33] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Intel® Advanced Matrix Extensions (Intel® AMX)

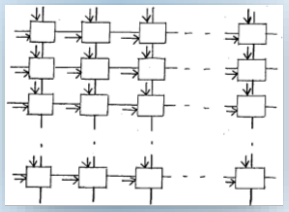
"Tiles"



2D Register Files

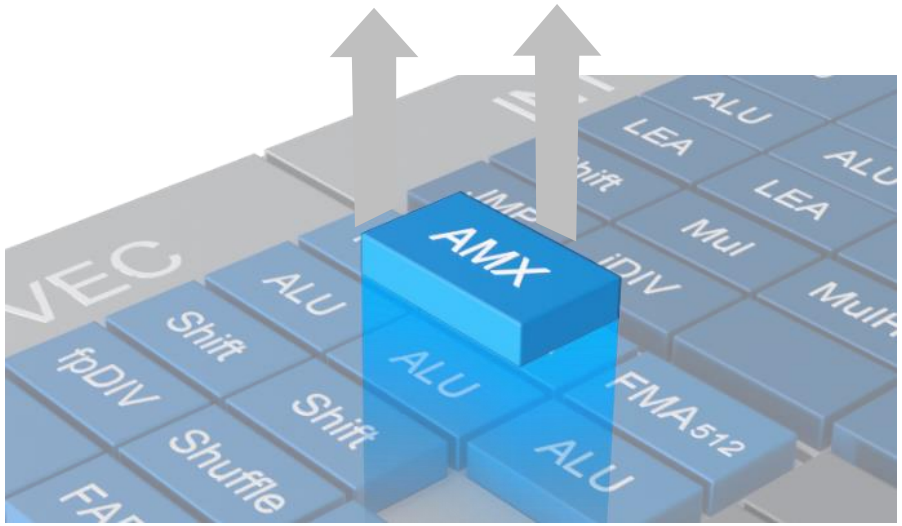
+

"TMUL"



Tile Matrix Multiply

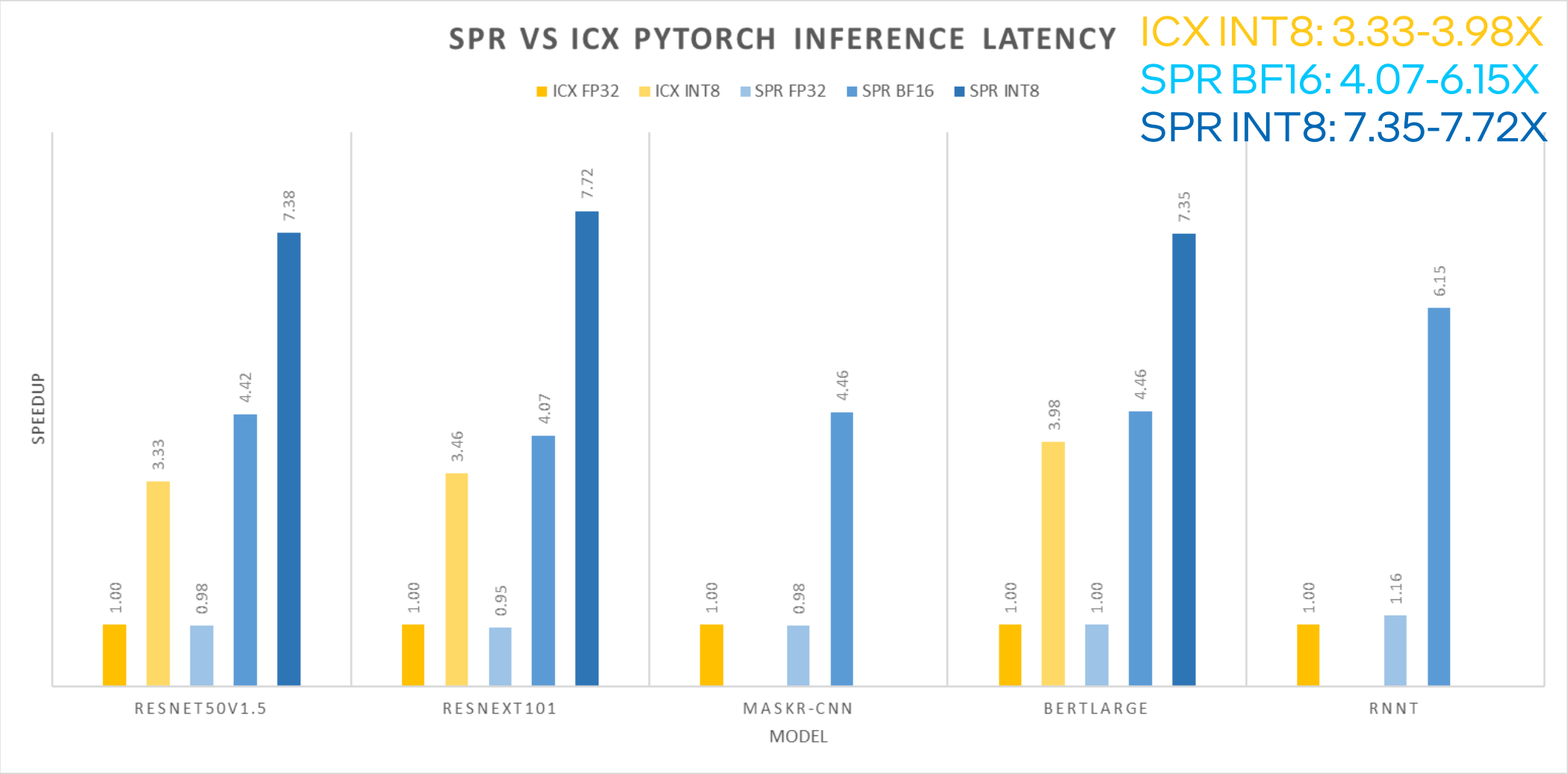
Store bigger chunks of **data**



Instructions that compute larger matrices in a single operation

PyTorch Benchmark: SPR vs ICX Inference (Batch Size = 1)

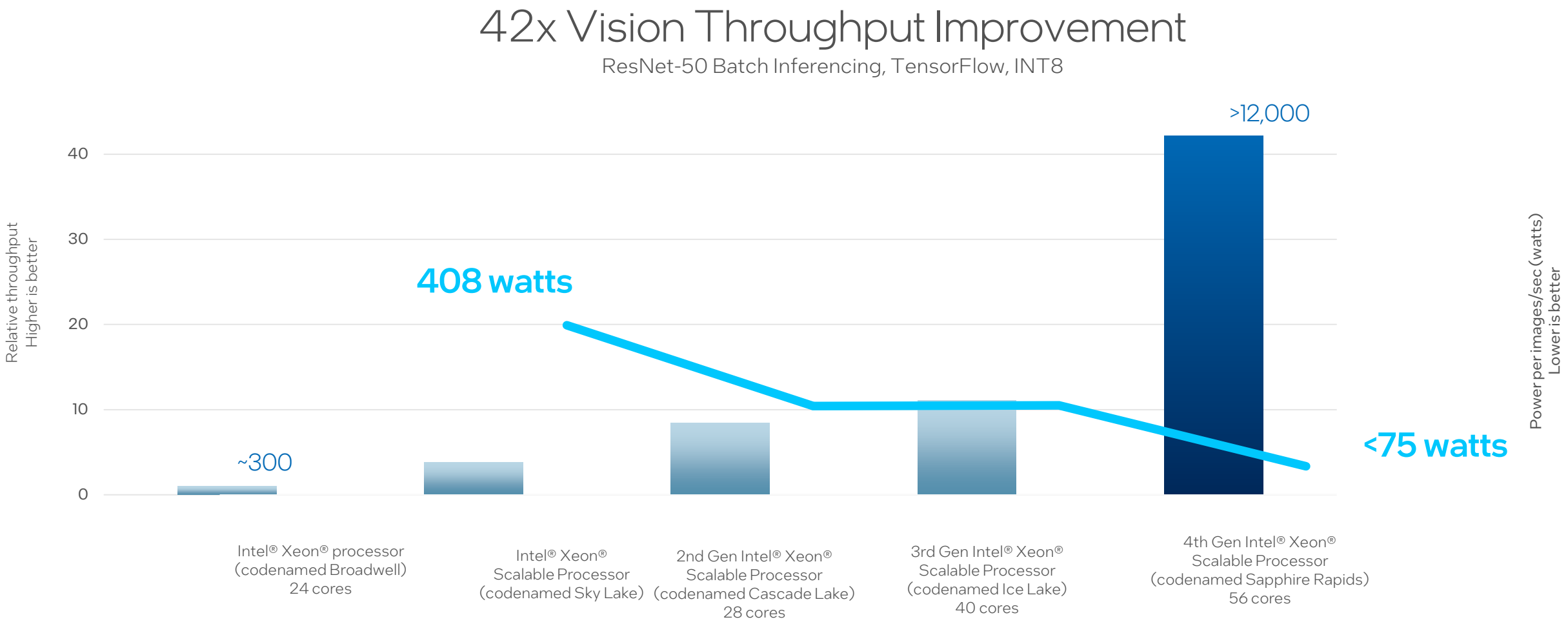
Inference latency speedup: the higher the better



Benchmark data for the Intel® 4th Gen Xeon Scalable Processors can be found here.

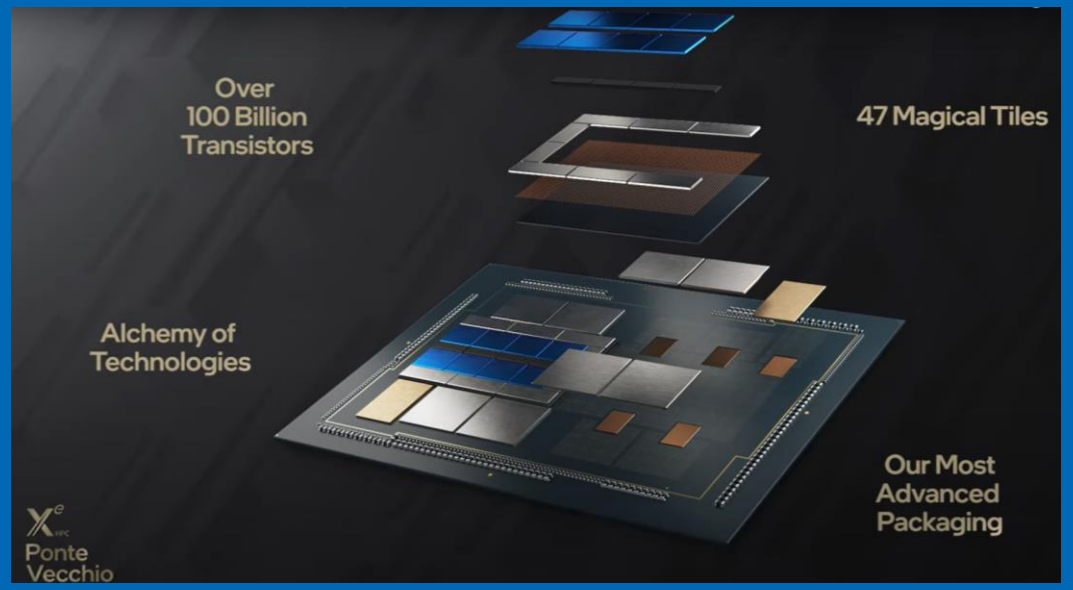
Configurations slide at the end(slide 106).

Real Workloads: SLA Compliance with Outstanding Performance Per Watt



Performance varies by use, configuration and other factors. See backup for configurations. Results may vary.

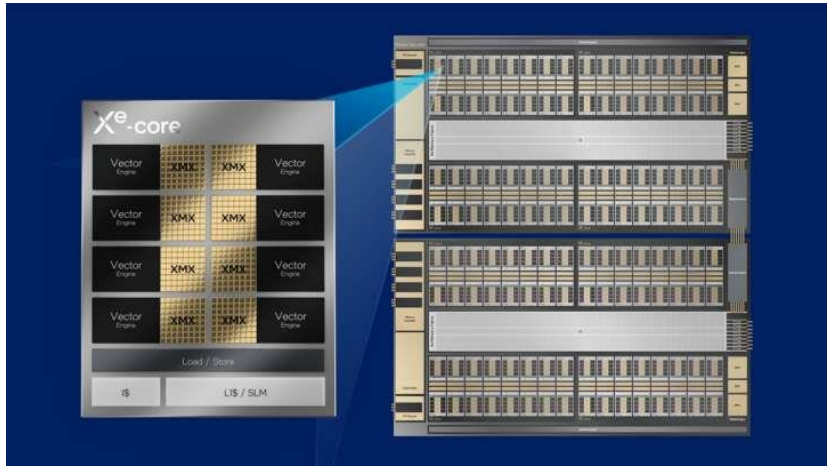
Intel® Data Center GPU Max Series



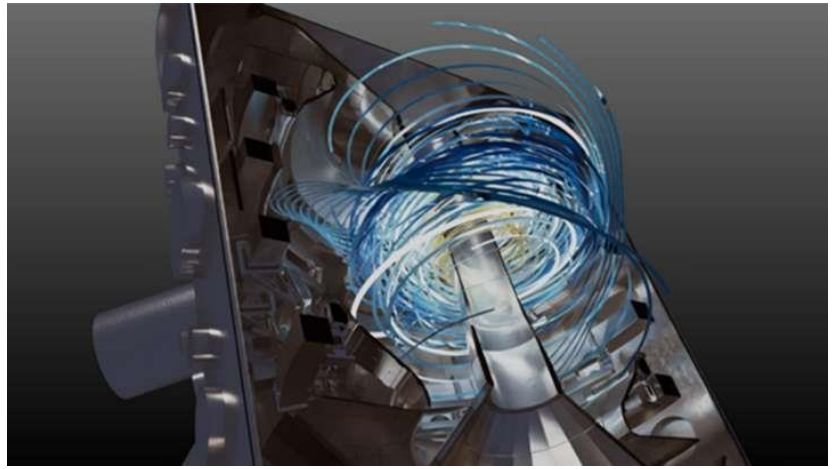
The Intel® Data Center GPU Max Series is designed to take on the most challenging high-performance computing (HPC) and AI workloads. The Intel® Xe Link high-speed, coherent, unified fabric offers flexibility to run any form factor to enable scale up and scale out.

Intel® Data Center GPU Max Series

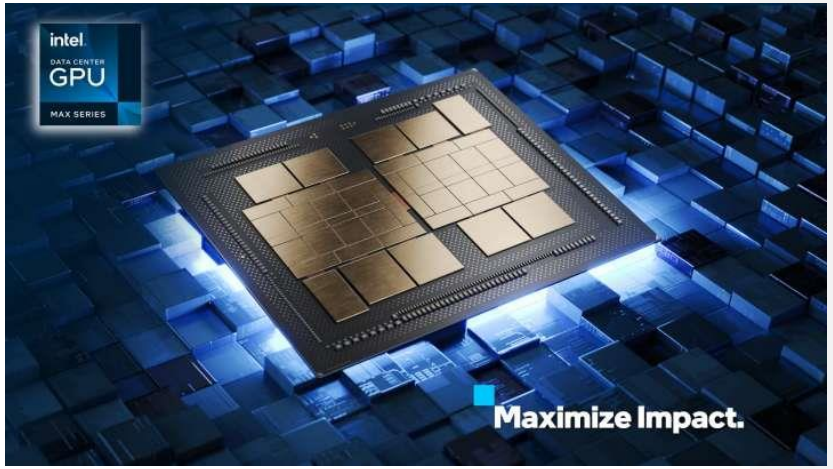
Leadership Performance for Data-level Parallel AI Workloads



Intel® Xe Matrix Extensions (XMX)



Built-in Ray Tracing Acceleration



Up to 408MB of L2 Cache

Maximize Impact

The Intel® Data Center GPU Max Series accelerates science and discovery with breakthrough performance.

Up to **2x** performance gain on HPC and AI workloads over competition due to the Intel® Max Series GPU large L2 cache.[1]

Up to **12.8x** performance gain over 3rd Gen Intel® Xeon® processors on LAMMPS workloads running on Intel® Max Series CPUs with kernels offloaded to six Intel® Max Series GPUs, optimized by Intel® oneAPI tools.[2]

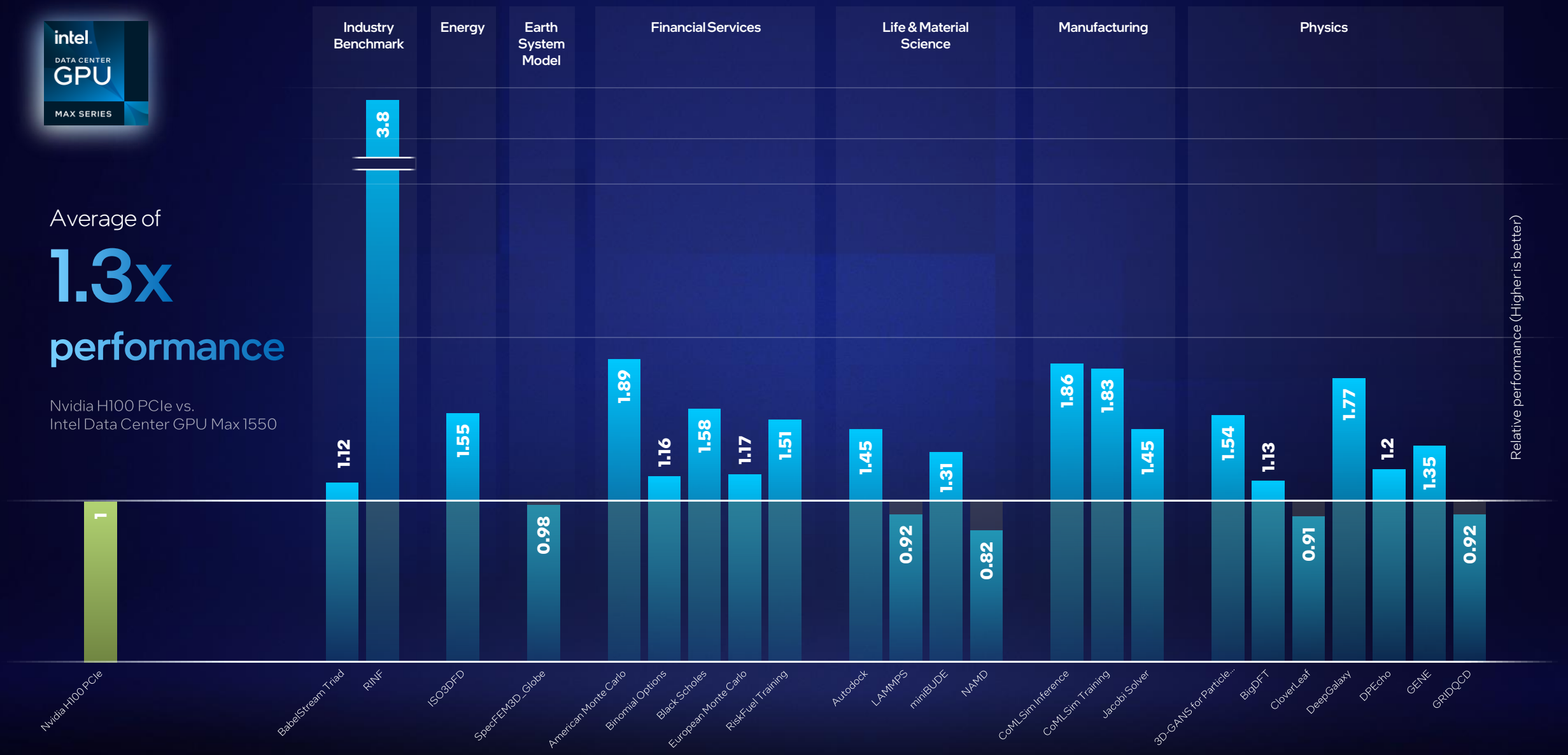
Up to **256x** Int8 operations per clock. Speed AI training and inference with up to 256 Int8 operations per clock with the built-in Intel® XMX.

intel. DATA CENTER GPU MAX SERIES

Average of

1.3x performance

Nvidia H100 PCIe vs.
Intel Data Center GPU Max 1550



Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

intel.

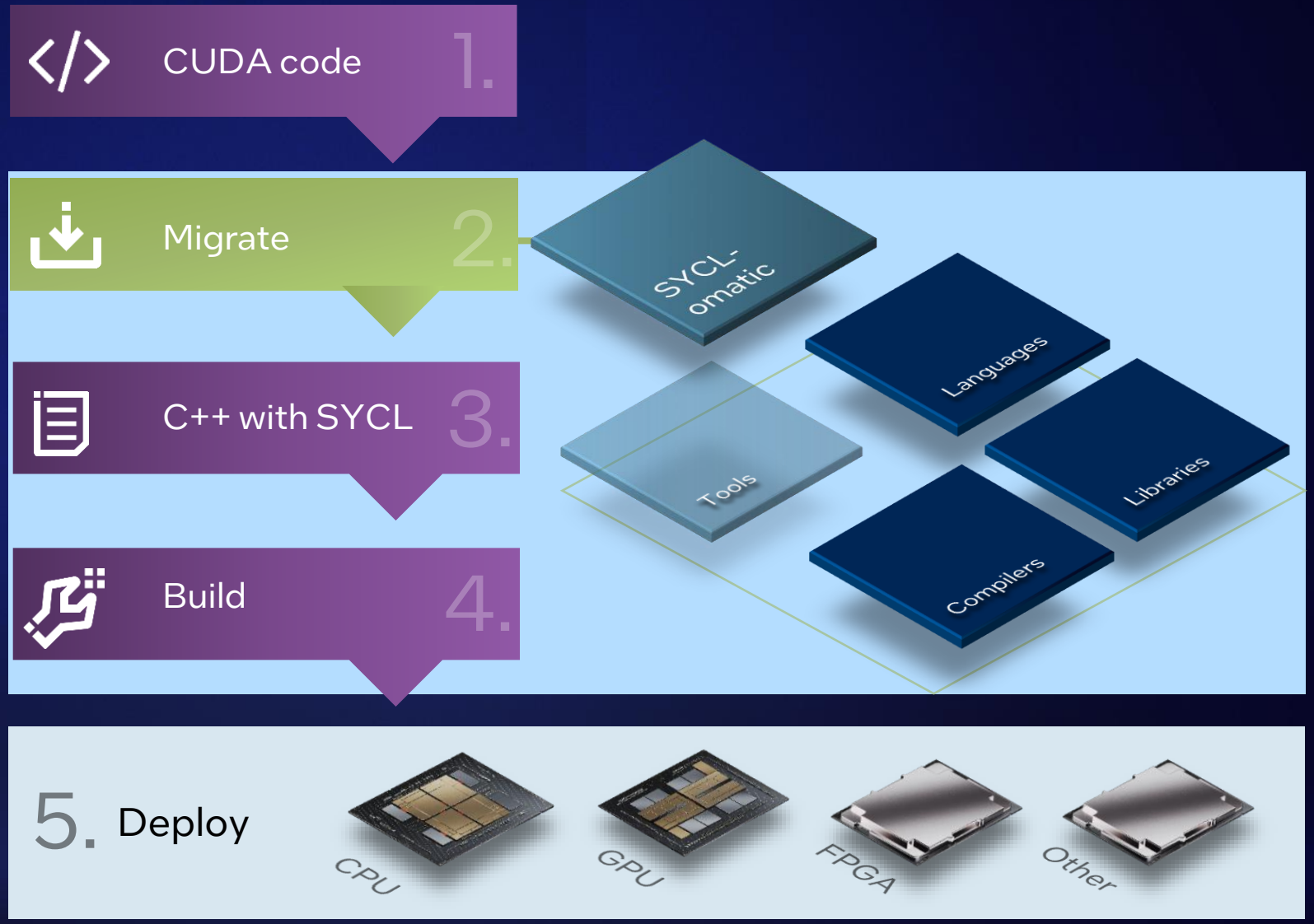
SYCLomatic

easily port CUDA* code to SYCL* and C++ to accelerate cross-architecture programming

Open Source

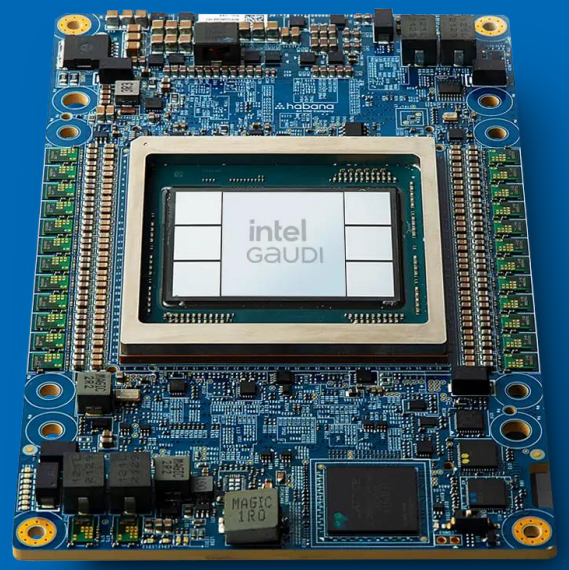
For more info visit:

https://github.com/oneapi-src/SYCLomatic



intel.

Intel®Gaudi® AI accelerator



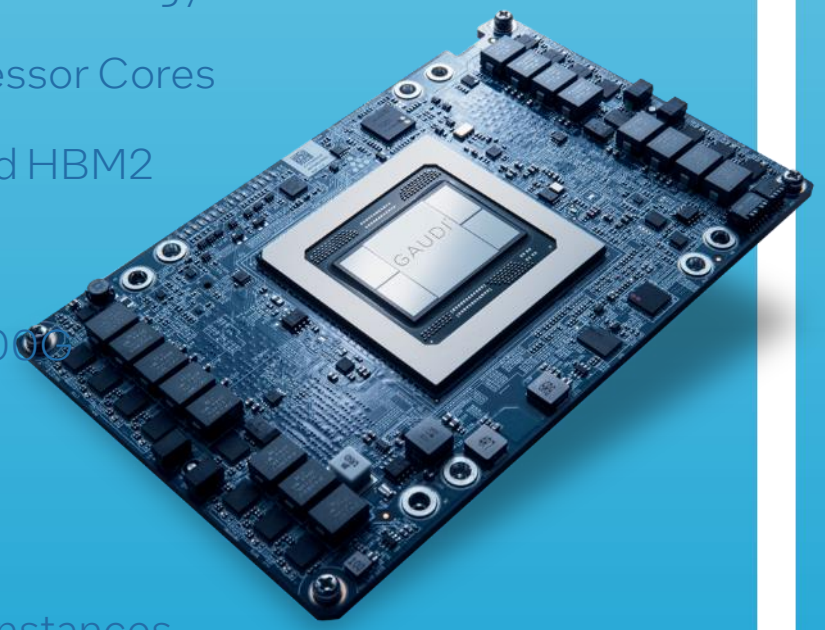
High Performance Acceleration for GenAI and LLMs

Habana Deep Learning Solutions

GAUDI®

High-performance, high-efficiency (price/performance)

- 16nm process technology
- 8 Tensor Processor Cores
- 32 GB on-board HBM2
- 24 SRAM
- 10 integrated 100G Ethernet ports



In the cloud:

- Amazon EC2 DL1 Instances

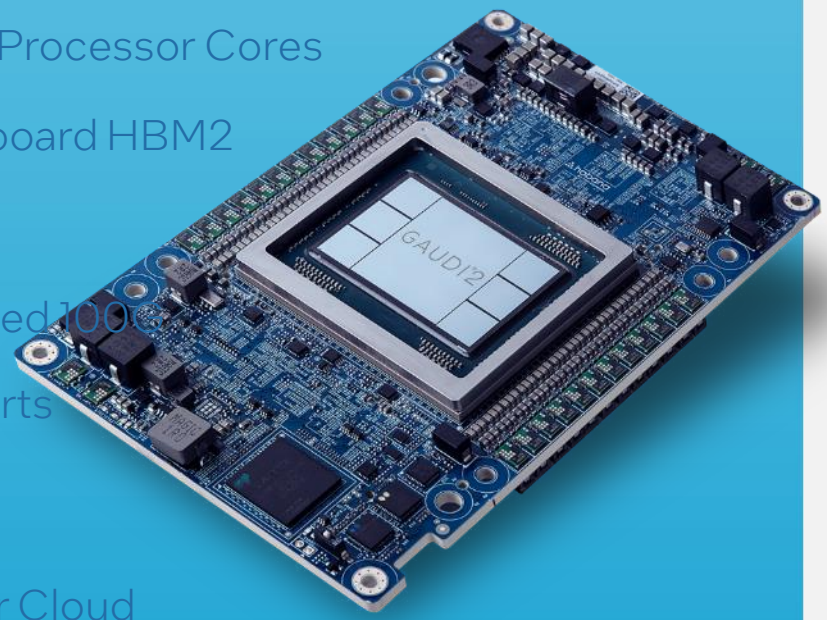
On-premises:

- Supermicro X12 Gaudi Server with 3rd Gen Xeon CPU

GAUDI®2

Higher performance, high-efficiency; optimized speed, memory, scalability for large scale models

- 7nm process technology
- 24 Tensor Processor Cores
- 96 GB on-board HBM2
- 48 SRAM
- 24 integrated 100G Ethernet ports



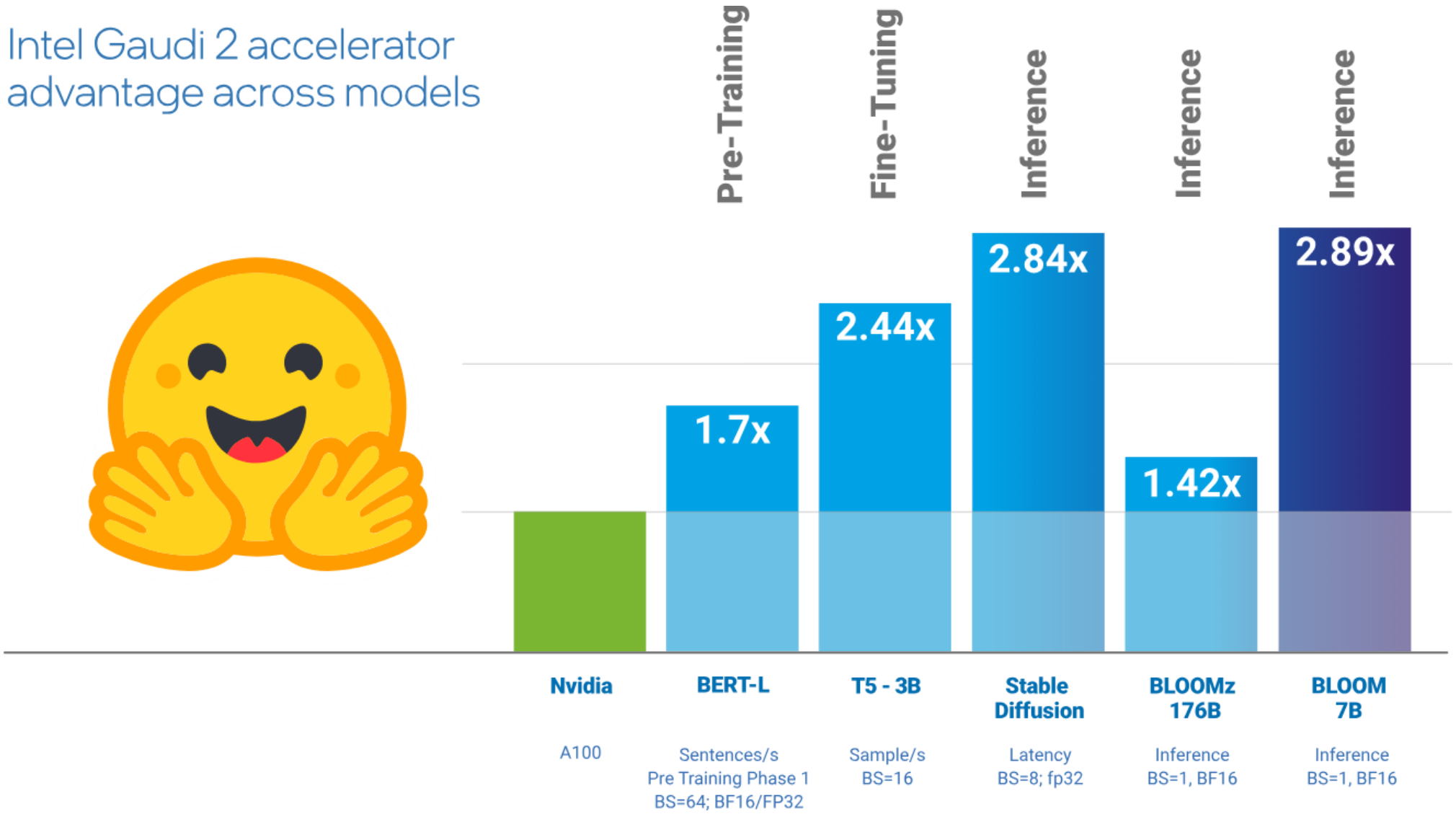
In the cloud:

- Intel Developer Cloud

On-premises:

- Supermicro Gaudi2 Server with 3rd Gen Xeon CPU

Intel® Gaudi2 accelerator Performance



Visit https://habana.ai/habana-claims-validation for workloads and configurations. Results may vary https://huggingface.com/blog/habana-gaudi-2-benchmark https://huggingface.co/blog/habana-gaudi-2-bloom

Enabling the Software Ecosystem through Robust Partnerships

intel. GAUDI

+

Hugging Face

PyTorch Lightning

cnvrg.io

DeepSpeed

Red Hat

Intel® Gaudi Software integrated with Industry leading ecosystem partners for Generative AI & Deep Learning

Docker container images
Intel® Gaudi Developer Site
Intel® Gaudi GitHub
Intel® Gaudi Developer Forum
Intel® Gaudi Software Documentation

Use cases

oneAPI Powered AI Reference Kit

Focusing on tackling deployment challenges with most popular AI use cases

SCAN ME

Finance & Insurance

- Claim Document Automation
- Fraud detection in credit card transactions
- Default Risk Prediction
- Disaster appraisal process

Health & Life Sciences

- Medical Imaging Diagnosis
- Disease Prediction
- AI Transcribe for Therapists

Process Automation

- Visual Process Discovery
- Intelligent Document Indexing
- Invoice-to-Cash Automation
- Historical Assets Document Processing

Customer Care

- Purchase Prediction
- Customer Segmentation
- Customer Churn Prediction
- Customer Care Chatbot
- Demand Forecasting
- Product Recommendations
- Order to delivery Forecasting

Synthetic Data

- AI Synthetic Data (Structured)
- AI Synthetic Data (Unstructured – Text)
- AI Synthetic Data (Unstructured – Image)
- AI Synthetic Data (Unstructured – Voice)

Manufacturing & Utilities

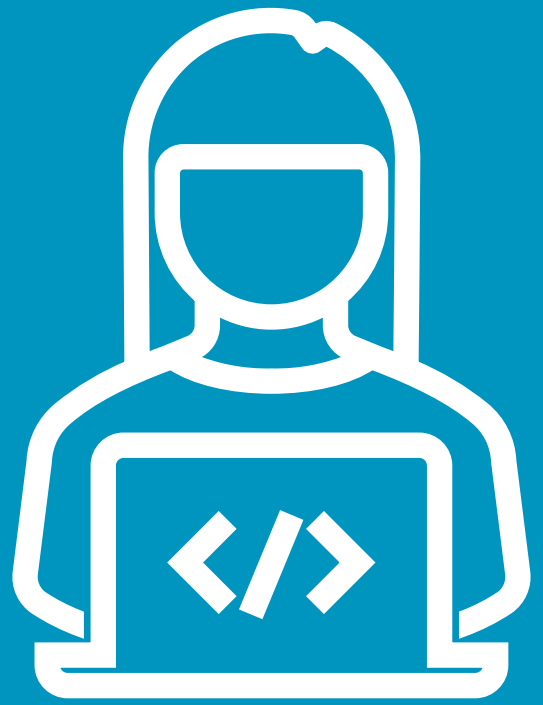
- Drone Navigation Segmentation
- Power Line Fault Detection
- Predictive Asset Analytics
- Visual Quality Inspection
- Digital Twin
- Engineering Design
- Traffic Camera object detection
- Aerodynamics / Fluid Flow Profiling

Tech & Security

- Vertical Search Engine
- Network Intrusion Detection
- Data Protection
- IoT (Data Streaming Anomality Detection)

- More info at https://www.intel.com/aireferencekit
- Downloads available from GitHub at https://github.com/oneapi-src

Demo: Stable Diffusion on Intel® Max Series GPU



Stable Diffusion(SD) Use case

Create Your Own Stable Diffusion

Few-shot Fine-tuning

5 minutes

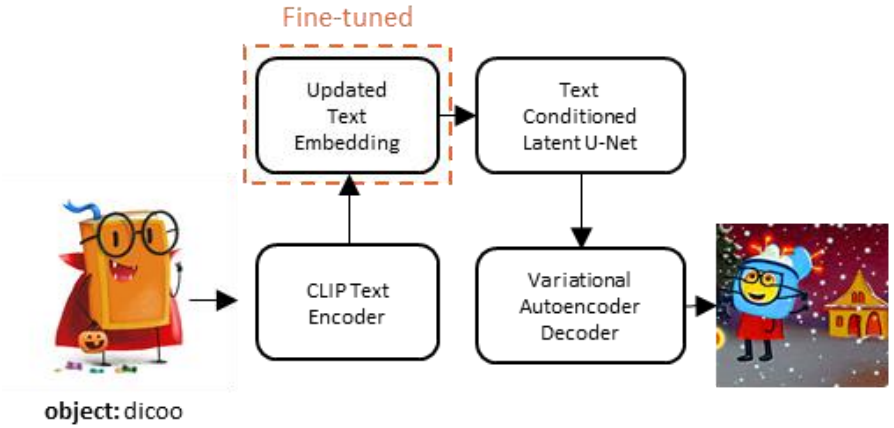
4 SPR nodes

Accelerated Stable Diffusion Inference

Low-precision Inference

5 seconds

1 SPR chip



Optimizations upstreamed to Hugging Face Diffusers and Optimum-Intel

[Try SD demo here: https://huggingface.co/spaces/Intel/Stable-Diffusion-Side-by-Side](https://huggingface.co/spaces/Intel/Stable-Diffusion-Side-by-Side)

Ref: https://venturebeat.com/ai/unlocking-generative-ai-with-ubiquitous-hardware-and-open-software/

Model Zoo for Intel® Architecture

Available on GitHub

Runs out-of-the-box

PyTorch use cases

- Image Recognition, Image Segmentation, Language Modeling/Translation, Object Detection, Recommendation, Text-to-Speech, Shot Boundary Detection, AI Drug Design
- Supported on dGPU: INT8 inference on ResNet50v1.5, SSD-MobileNet, Yolo V4

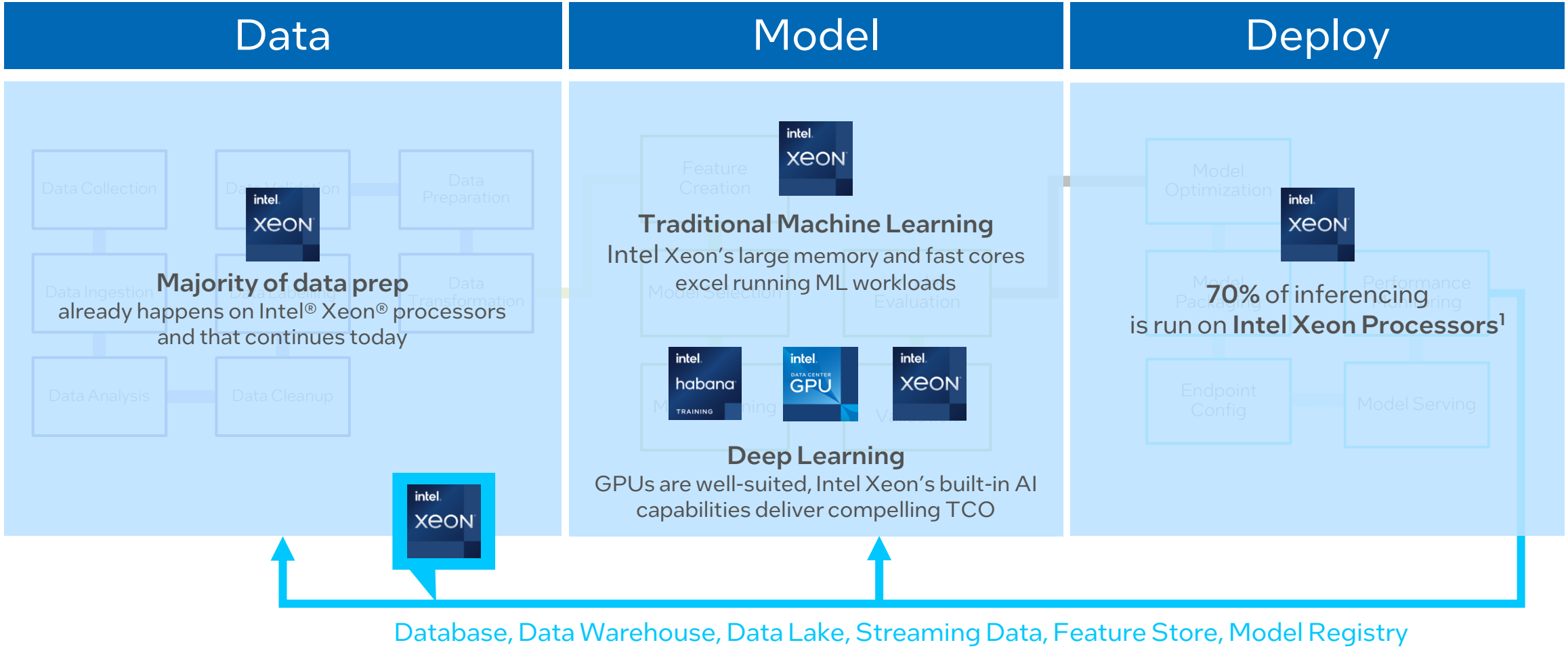
Model Zoo: https://github.com/intelAI/models/tree/master

Image Recognition

Model	Framework	Mode	Model Documentation	Benchmark/Test Dataset
DenseNet169	TensorFlow	Inference	FP32	ImageNet 2012
Inception V3	TensorFlow	Inference	Int8 FP32	ImageNet 2012
Inception V4	TensorFlow	Inference	Int8 FP32	ImageNet 2012
MobileNet V1*	TensorFlow	Inference	Int8 FP32 BFloat16	ImageNet 2012
ResNet 101	TensorFlow	Inference	Int8 FP32	ImageNet 2012
ResNet 50	TensorFlow	Inference	Int8 FP32	ImageNet 2012
ResNet 50v1.5	TensorFlow	Inference	Int8 FP32 BFloat16 dGPU Int8	ImageNet 2012
ResNet 50v1.5 Sapphire Rapids	TensorFlow	Inference	Int8 FP32 BFloat16	ImageNet 2012
ResNet 50v1.5	TensorFlow	Training	FP32 BFloat16	ImageNet 2012
Inception V3	TensorFlow Serving	Inference	FP32	Synthetic Data
ResNet 50v1.5	TensorFlow Serving	Inference	FP32	Synthetic Data
GoogLeNet	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
Inception v3	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
MNASNet 0.5	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
MNASNet 1.0	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
ResNet 50	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
ResNet 50	PyTorch	Training	FP32 BFloat16	ImageNet 2012
ResNet 101	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
ResNet 152	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
ResNext 32x4d	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
ResNext 32x16d	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
VGG-11	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
VGG-11 with batch normalization	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
Wide ResNet-50-2	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
Wide ResNet-101-2	PyTorch	Inference	FP32 BFloat16	ImageNet 2012
ResNet 50 v1.5	PyTorch	Inference	dGPU Int8	ImageNet 2012

Intel AI® Software Stack

The AI Pipeline Runs on Intel



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

Intel AI Software

Data → Model → Deploy

AI Platforms, MLOPs
(Productivity)

AI Libraries, Tools, Frameworks
(Most AI developers operate here)

Low-level Performance Libraries, Compilers, Kernel
(Most hardware enabling and performance tuning happens here)

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

AI Platforms & Kits

Most Popular Tools and Frameworks

Performance Libraries

Note: not all components are necessarily compatible with all other components in other layers

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

AI Platforms & Kits

Most Popular Tools and Frameworks

SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL | SynapseAI™

Note: not all components are necessarily compatible with all other components in other layers

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions, XMX – Xe Matrix Extensions

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

AI Platforms & Kits

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, OpenFL, scikit learn, ONNX, OpenXLA, Triton, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML

Automate Low-Precision Optimization

SigOpt

Intel Neural Compressor

w/ Intel Optimizations

SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL | SynapseAI™

intel. ATOM | intel. CORE | intel. XEON | intel. ARC GRAPHICS | intel. DATA CENTER GPU | intel. habana

Note: not all components are necessarily compatible with all other components in other layers

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions, XMX – Xe Matrix Extensions

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

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, OpenFL, scikit learn, ONNX, OpenXLA, Triton, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML

Automate Low-Precision Optimization

SigOpt

Intel Neural Compressor

w/ Intel Optimizations

SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL | SynapseAI™

intel. ATOM | intel. CORE | intel. XEON | intel. ARC GRAPHICS | intel. DATA CENTER GPU | intel. habana

Note: not all components are necessarily compatible with all other components in other layers

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions, XMX – Xe Matrix Extensions

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

Connect AI to Big Data — Spark — BigDL (previously "Analytics Zoo")

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, OpenFL, scikit learn, ONNX, OpenXLA, Triton, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML — SigOpt

Automate Low-Precision Optimization — Intel Neural Compressor

w/ Intel Optimizations

SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL | SynapseAI™

intel. ATOM | intel. CORE | intel. XEON | intel. ARC GRAPHICS | intel. DATA CENTER GPU | intel. habana

Note: not all components are necessarily compatible with all other components in other layers

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions, XMX – Xe Matrix Extensions

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

Connect AI to Big Data — Spark — BigDL (previously "Analytics Zoo")

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, OpenFL, scikit learn, ONNX, OpenXLA, Triton, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML

Automate Low-Precision Optimization

SigOpt

Intel Neural Compressor

OpenVINO™ Toolkit

Write Once

Deploy Auto-Optimized

Anywhere

w/ Intel Optimizations

SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL | SynapseAI™

intel. ATOM | intel. CORE | intel. XEON | intel. ARC GRAPHICS | intel. DATA CENTER GPU | intel. habana

Note: not all components are necessarily compatible with all other components in other layers

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions, XMX – Xe Matrix Extensions

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

Container Repository	oneAPI powered	MLOps	Developer Sandbox	Annotation/Training/Optimization
oneContainer	**AI Reference Kits**	**Cnvrg.io**	**Intel® Developer Cloud**	**Intel® GETi**

Connect AI to Big Data — Spark — BigDL (previously "Analytics Zoo")

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, OpenFL, scikit learn, ONNX, OpenXLA, Triton, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML — **SigOpt**

Automate Low-Precision Optimization — **Intel Neural Compressor**

OpenVINO™ Toolkit

Write Once

Deploy Auto-Optimized

Anywhere

w/ Intel Optimizations

SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL | SynapseAI™

intel. ATOM | intel. CORE | intel. XEON | intel. ARC GRAPHICS | intel. DATA CENTER GPU | intel. habana

Note: not all components are necessarily compatible with all other components in other layers

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions, XMX – Xe Matrix Extensions

Intel® oneAPI Toolkits

oneAPI

Intel® oneAPI Base Toolkit

intel. 1 oneAPI BASE TOOLKIT

A core set of high-performance libraries and tools for building C++, SYCL, C/OpenMP, and Python applications

Add-on Domain-specific Toolkits

intel. 1 oneAPI HPC TOOLKIT

For HPC developers

Intel® oneAPI Tools for HPC
Deliver fast Fortran, OpenMP & MPI applications that scale

intel. 1 oneAPI RENDERING TOOLKIT

For visual creators, scientists & engineers

Intel® oneAPI Rendering Toolkit
Accelerate visual compute, deliver high-performance, high-fidelity visualization applications.

intel. 1 oneAPI IoT TOOLKIT

For edge & IoT developers

Intel® oneAPI Tools for IoT
Build efficient, reliable solutions that run at network's edge

Toolkits powered by oneAPI

intel. AI ANALYTICS TOOLKIT

For AI developers & data scientists

Intel® AI Analytics Toolkit
Accelerate machine learning & data science pipelines end-to-end with optimized DL & ML frameworks & high-performing Python libraries

OpenVINO™

For deep learning inference developers

Intel® OpenVINO™ toolkit
Deploy high performance inference & applications from edge to cloud

Download at **intel.com/oneAPI**
Or visit Intel® DevCloud for oneAPI

Intel Has the Developer Tools Companies Use to Scale AI Everywhere

Upstream

Integrated acceleration to popular open source software

Modin, XGBoost, TF, PT, PDPD, MxNet, more …

Intel Extension

Easily pluggable extensions to open source software

Scikit-Learn Extension, Optimized Analytics Package, IPEX, more …

Intel Distro

Intel Optimized Distributions of open source software

Modin, Intel TF, IDP

Intel Tools

Tools / Kits which improve productivity and perf on Intel HW

AIKIT, OpenVINO™, BigDL, oneContainer Portal, Cnvrg.io, Intel Neural Compressor, SigOpt, SynapseAI®

Across major software channels (PyPI, Anaconda, Intel, Apt, Yum, Docker)

Intel AI Software by Platform

Intel® Xeon® Scalable Processor
Intel® Data Center GPU
Intel® Gaudi® Processors for DL

Category	Software	Open Source	Optimizations Upstreamed*	Intel Extension**	Intel Distribution	Intel Tool / Kit
Orchestration	Cnvrg.io	No				Xeon, DGPU, Gaudi
Toolkits	BigDL	Yes				Xeon
	OpenVINO	Yes			Xeon, DGPU	Xeon, DGPU
Optimization	Neural Compressor	Yes				Xeon, DGPU
	SigOpt	Yes				Xeon, DGPU, Gaudi
DL Frameworks	TensorFlow	Yes	Xeon	Xeon, DGPU	Gaudi	
	PyTorch	Yes	Xeon	Xeon, DGPU	Gaudi	
	ONNX	Yes	Xeon			
	PDPD	Yes	Xeon			
	DeepSpeed	Yes		DGPU		
	OpenFL	Yes	Xeon			
ML Frameworks	XGBoost	Yes	Xeon			
	Scikit-Learn	Yes		Xeon, DGPU		
	CatBoost	Yes	Xeon			
	LightGBM	Yes	Xeon			
Data Preprocessing	Modin (for Pandas)	Yes	Xeon		Xeon	
	Intel® Distribution for Python	Yes			Xeon, DGPU	
	Spark	Yes	Xeon	Xeon		
AI Compilers	Triton	Yes		DGPU		
	OpenXLA	Yes		DGPU		

* Intel strives to upstream as many optimizations for as many hardware targets as soon as possible
** Access more Intel optimizations and target hardware support through API-compliant extensions

Intel® AI Analytics Toolkit

Powered by oneAPI

Accelerate end-to-end AI and data analytics pipelines with libraries optimized for Intel® architectures

Who Uses It?

Data scientists, AI researchers, ML and DL developers, AI application developers

Top Features/Benefits

- Deep learning performance for training and inference with Intel optimized DL frameworks and tools
- Drop-in acceleration for data analytics and machine learning workflows with compute-intensive Python packages

Learn More: software.intel.com/oneapi/ai-kit

Deep Learning

- Intel® Optimization for TensorFlow
- Intel® Optimization for PyTorch
- Intel® Neural Compressor
- Model Zoo for Intel® Architecture

Data Analytics & Machine Learning

Accelerated Data Frames

- Intel® Distribution of Modin
- OmniSci Backend

Intel® Distribution for Python

- XGBoost
- Scikit-learn
- Daal-4Py
- NumPy
- SciPy
- Pandas

Samples and End2End Workloads

CPU GPU

Supported Hardware Architechures[1]

Hardware support varies by individual tool. Architecture support will be expanded over time.
Other names and brands may be claimed as the property of others.

Get the Toolkit HERE or via these locations

- Intel Installer
- Docker
- Apt, Yum
- Conda
- Intel® DevCloud

High-Performance Deep Learning Using Intel® Distribution of OpenVINO™ toolkit - **Powered by oneAPI**

A toolkit for fast, more accurate real-world results using high-performance AI and computer vision inference deployed into production on Intel XPU architectures (CPU, GPU, FPGA, VPU) from edge to cloud

Who needs this product?

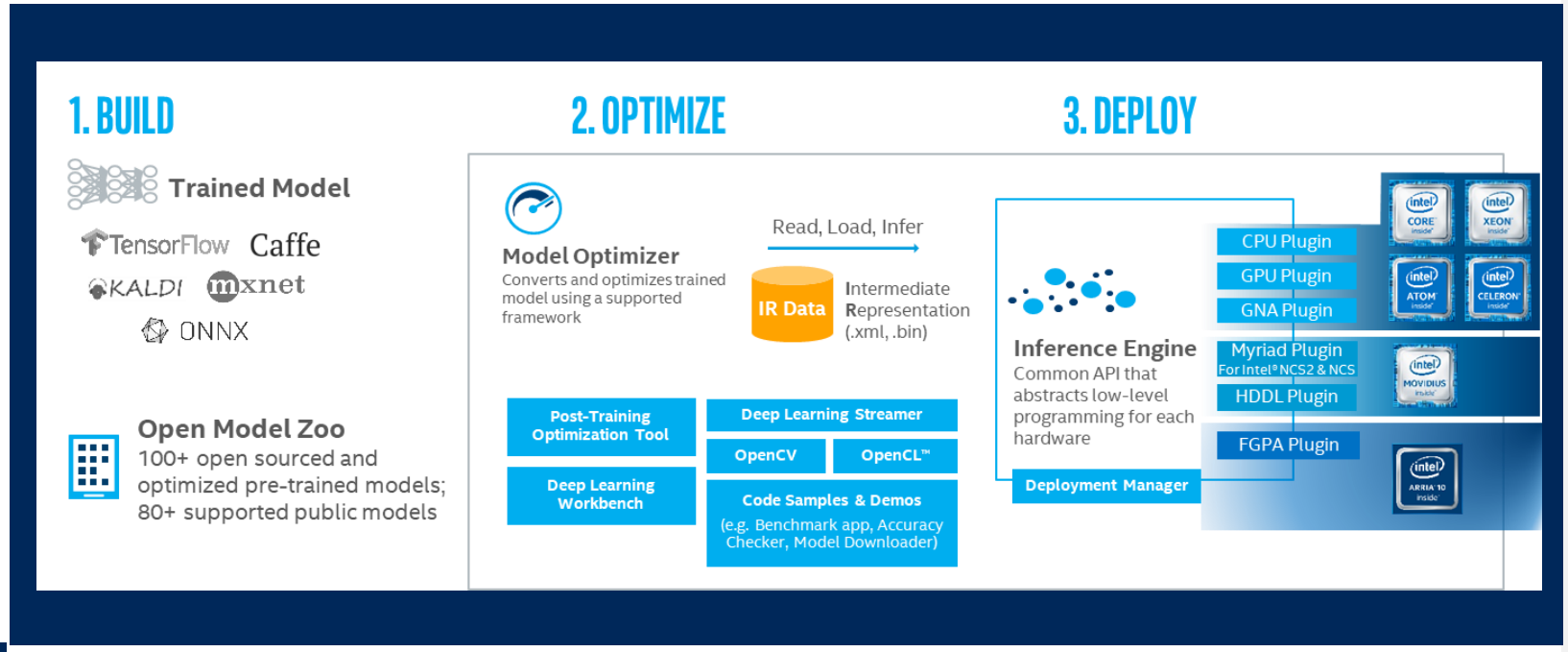
AI application developers, OEMs, ISVs, System Integrators, Vision and Media developers

Top Features/Benefits

High-performance, deep learning inference deployment

Streamlined development; ease of use

Write once, deploy anywhere



Proven, industry-leading accelerated technology

software.intel.com/openvino-toolkit

Simplified Download Experience for Intel AI Software

AI Tools Selector (Preview)

Achieve End-to-End Performance for AI Workloads, Powered by oneAPI

Overview | Download | Documentation

Presets

Data Analytics | Classical Machine Learning | Deep Learning | Inference Optimization

Python* Versions

Python* 3.9 | Python 3.10

Package Type

conda* | pip | Docker*

Deep Learning Framework Optimizations

- [x] Intel® Extension for TensorFlow*
- [x] Intel® Extension for PyTorch*

Intel®-Optimized Tools & Libraries

- [ ] Intel® Optimization for XGBoost*
- [ ] Intel® Extension for Scikit-learn*
- [ ] Intel® Distribution of Modin*
- [ ] Intel® Neural Compressor

SDKs & CLIs

- [ ] cnvrg.io™ SDK V2 in Python*

AI Tools: Deep Learning

All packages are for Linux* only.

Install a Docker* Container

```
docker pull intel/deep-learning:2023.2-py3.9
```

Installation Instructions for Docker*

You must install Docker to run the containers. For complete instructions, visit the Docker website.

Installation Instructions

Working with Preset Containers

Offline Installer

All AI tools are available for offline installation using a stand-alone installer. Choose this option if your target installation environments are behind a firewall, you need to manage versions, or for other purposes.

Download

In This Package

Intel® Distribution for Python* is a cluster of packages, including the Python Interpreter and compilers, that are optimized via Intel® oneAPI Math Kernel Library (oneMKL) and Intel® oneAPI Data Analytics Library (oneDAL) to make Python applications more efficient.

Intel® Extension for TensorFlow* is a heterogeneous, high-performance, deep learning extension plug-in based on a TensorFlow* PluggableDevice interface that enables access to Intel CPU and GPU devices with TensorFlow for AI workload acceleration.

Intel® Extension for PyTorch* extends PyTorch with up-to-date feature optimizations for an extra performance boost on

OpenVINO

OpenVINO™ toolkit: An open source toolkit that makes it easier to write once, deploy anywhere.

Overview | What's New | Get Started | Industry | Download

Choose a Preferred Package

You can customize the selections to fit your needs.

Sign up for the latest product releases, news, and tips.

Version

2023.1.0 (Recommended) | 2022.3.1 (Latest LTS release, includes NCS2/HDDL support) | 2021.4.2 (Previous LTS release)

Operating System

Windows | macOS | Linux

Previous Releases

Distribution

OpenVINO Archives | PIP (Python API only) | GitHub (Source)

Gitee (Source) | Docker | Conda

vcpkg (Source)

Try in the Intel® Developer Cloud

We have simplified the install options (example: consolidation of Runtime and Development Tools): Learn more

Download

Installation Instructions

Get Started Guide

Notebooks

Troubleshooting Guide

Download Archives

Advanced Optimization tool available separately: Learn about NNCF

System Requirements

For a complete list of supported hardware, see the system requirements.

Supported Operating System	Python Version (64 bit)
Windows® 10 (64 bit)	3.7, 3.8, 3.9, 3.10, 3.11
Windows® 11 (recommended for 12th Generation Intel® Core™ processors)	3.7, 3.8, 3.9, 3.10, 3.11

https://www.intel.com/content/www/us/en/developer/tools/oneapi/ai-tools-selector.html
https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/download.html?VERSION=v_2023_1_0&OP_SYSTEM=WINDOWS&DISTRIBUTION=ARCHIVE

oneAPI Available on

Intel® DevCloud for oneAPI

A development sandbox to develop, test and run workloads across a range of Intel CPUs, GPUs, and FPGAs using Intel's oneAPI software.

Get Up & Running In Seconds!

Sign up at:
software.intel.com/devcloud/oneapi

intel DevCloud

1 Minute to Code

No Hardware Acquisition

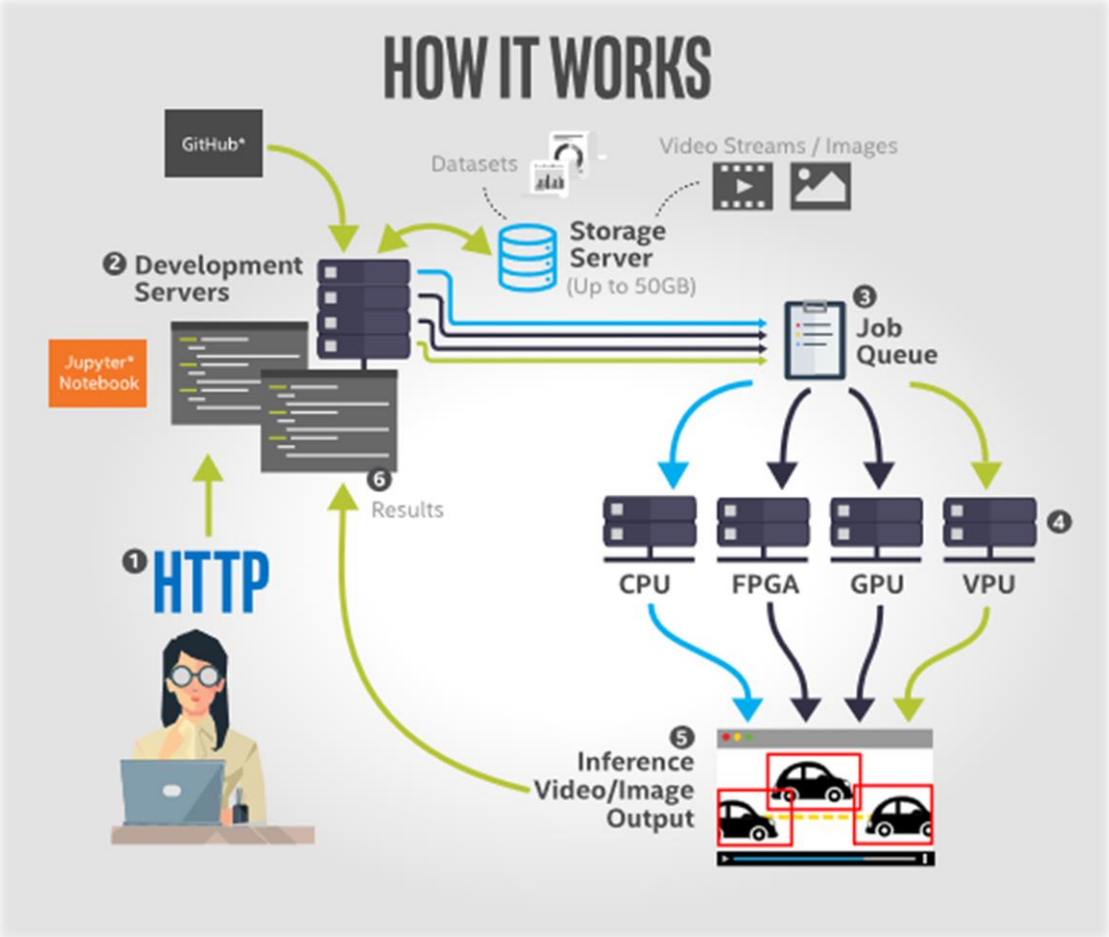
No Download, Install or Configuration

Easy Access to Samples & Tutorials

Support for Jupyter Notebooks, Visual Studio Code

Accelerate Time to Production with Intel® DevCloud for the Edge

See immediate AI Model performance across Intel's vast array of Edge Solutions



- **Instant, Global Access**
  Run AI applications from anywhere in the world
- **Prototype on the Latest Hardware and Software**
  Develop knowing you're using the latest Intel technology
- **Benchmark your Customized AI Application**
  Immediate feedback - frames per second, performance
- **Reduce Development Time and Cost**
  Quickly find the right compute for your edge solution

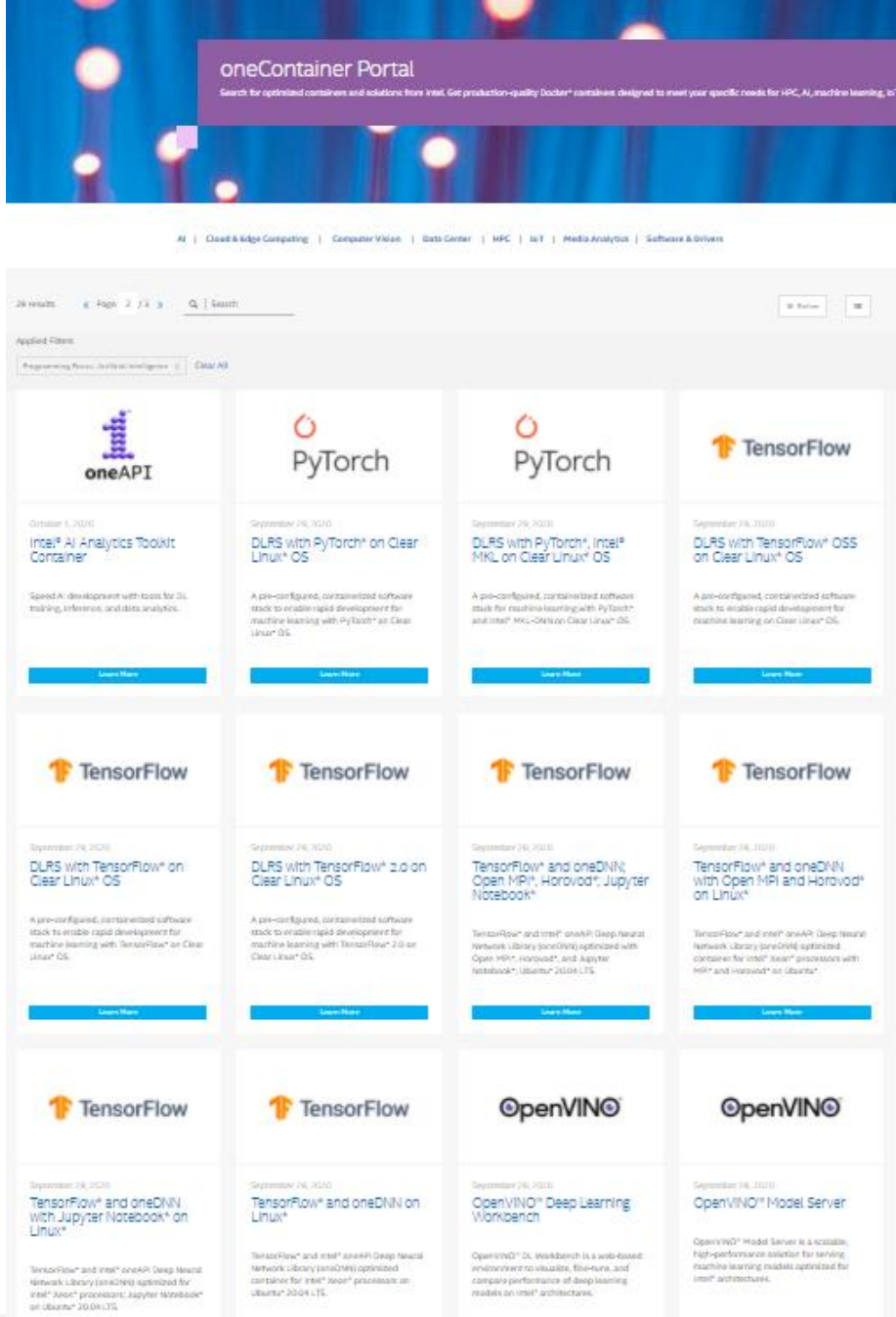
Sign up now for access

AI Containers for Flexibility

- Optimized, validated, deployable AI containers
- Available via Docker containers. Will expand to include Kubernetes orchestrations, Helm charts
- Access from oneContainer Portal
  - Include containers with ready-to-use AI software stacks
  - And containers with full AI workloads (including models)

Topology	Frameworks
DLRM	PYT
ResNet50	PYT, TF, OV
BERT-large	PYT, TF, OV
Transformer-LT	PYT, TF
MobileNet-v1	PYT, TF, OV
SSD-Mobilenet-v1	PYT, TF, OV
SSD-Resnet34	PYT, TF, OV
WaveNet*	TF

Topology	Framework
Mask R-CNN	PYT, TF, OV
RNN-T	PYT, TF, OV
3D-UNet	TF, OV
DIEN	TF
Wide & Deep	PYT, TF
RNX101	
Yolo-V3	PYT, TF, OV
NCF*	TF



Which Toolkit Should I Use

Use Both!

Intel® oneAPI Analytics Toolkit & Intel® Distribution of OpenVINO™ toolkit

Toolkits are complementary to each other and recommendation is to use them both based on your current phase of AI Journey

- *I am **exploring and analyzing data;** I am **developing models***
- *I want **performance and compatibility** with frameworks and libraries I use*
- *I would like to have **drop-in acceleration** with little to no additional code changes*
- *I prefer **not to learn any new tools** or languages*

Data Scientist/ML Developer
Intel® oneAPI AI Analytics Toolkit

- *I am **deploying models***
- *I want **leading performance and efficiency** across multiple target HW*
- *I'm concerned about **having lower memory footprint**, which is critical for deployment*
- *I am **comfortable with learning and adopting a new tool or API** to do so*

App Developer
Intel® Distribution of OpenVINO™ toolkit

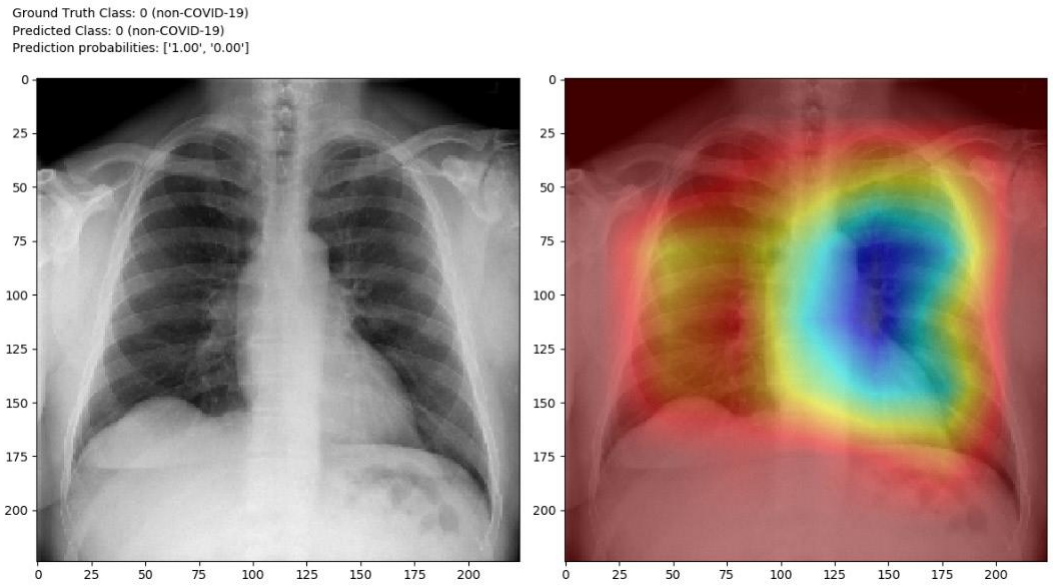
If you prefer working on primitives and to optimize kernels and algorithms directly using oneAPI libraries (oneDNN, oneCCL & oneDAL), then use Intel® oneAPI Base Toolkit

Accrad AI-based Solution Helps Accelerate COVID-19 Diagnosis

Optimized by Intel® oneAPI Analytics Toolkit & Intel® Distribution of OpenVINO™ toolkit

CheXRad helps radiologists and physicians identify COVID-19, viral pneumonia and other diseases on chest X-ray images, and predict the need for ventilators.

- *CheXRad* comes pre-configured with a COVID-19 and viral pneumonia classification neural network.
- To architect, train and validate the neural network, Accrad used **Intel Tensorflow from AI Analytics Toolkit** and the **Intel oneAPI DevCloud** to develop the model.
- To optimize its model for deployment, Accrad used **OpenVINO™ toolkit** and **Intel® DevCloud for Edge**.
- *CheXRad* could classify pathologies in 140 chest x-rays in just **90 seconds** —up to **160x faster** than radiologists, at comparable levels of accuracy, sensitivity and specificity.



Learn more in this solution brief

LLMs in Enterprise with Intel

Select a Pre-trained model

- Stable LM
- Vicuna
- Falcon
- GPT-J
- Llama

Hugging Face
TensorFlow Hub
PYTORCH HUB

Bring & process your proprietary data

MODIN

AutoML & Hyperparameter Tuning

- Learning Rate
- Epochs
- Other
- Batch Size
- Accuracy Target
- Other

SIGOPT

Fine Tune

- Domain Adaptation
- PEFT
- QLORA
- Feature Extraction
- LORA

SIGOPT
Transformers
intel Transformer Extension

Model Compression

- Quantization
- Distillation
- Pruning

Optimum
intel Neural Compressor

Deploy on XPU

Deploy on Intel® Xeon Scalable processor with Built-In AI Accelerator

Optimized Models & Spaces

Dolly	LLAMA2	MPT	LDM3D	Whisper	*Hundreds of thousands more...*

Intel Optimized Hugging Face Libraries & Tools

Transformers	Diffusers	Accelerate	PEFT	Optimum
Fine Tuning for NLP,CV	Generative Use Cases	Fine Tuning at Scale	Efficient Fine Tuning	Performance Optimization

Foundational Stack

TensorFlow PyTorch ONNX RUNTIME OpenVINO

Fine Tuning workflows on Hugging Face Platform optimized OOB for Intel products

https://huggingface.co/Intel

Save Time with One-line Code Changes

More model experimentation for higher accuracy

Engineer Data → Create Machine Learning & Deep Learning Models

~90x

pandas MODIN

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

~38x

scikit learn

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

More Acceleration

Quick Start Guide

See link below for workloads and configurations. Results may vary
https://www.intel.com/content/www/us/en/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html

Conclusion

Key Takeaways & Call to Action

- Intel toolkits are **FREE**, complementary & work seamlessly together
- They help achieve performance & efficiency across different stages of AI Journey
- Recommend the toolkits based on current phase of customer pipeline

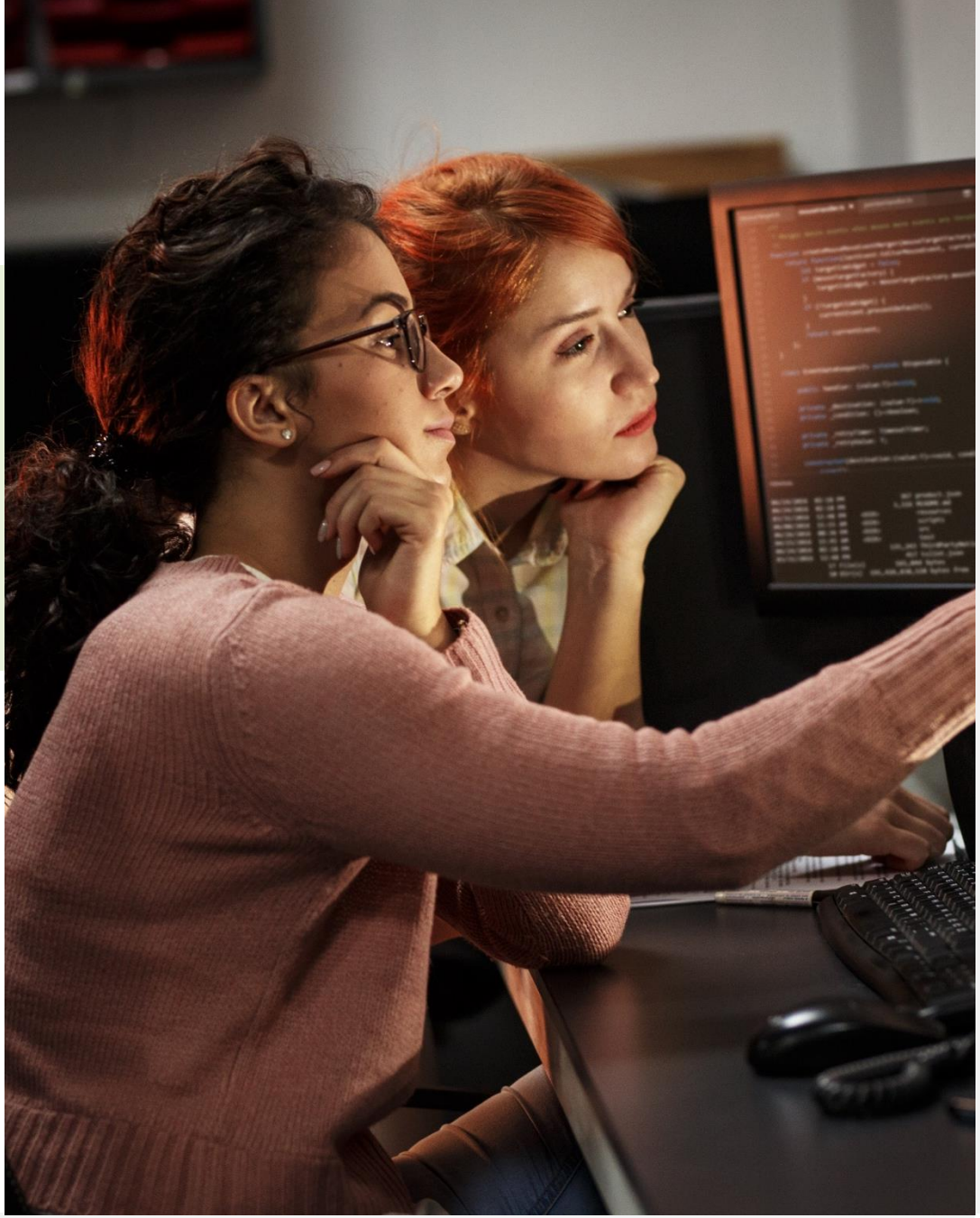
Download the toolkits

Intel® oneAPI AI Analytics Toolkit

Intel® Distribution of OpenVINO ™ toolkit

Intel® oneAPI Base Toolkit

Learn more about Intel® oneAPI Toolkits
intel.com/oneAPI-AllToolkits



Thank you for your attention!