AI Update from AI PC to Accelerators

Gretchen Stewart, PE – AI Solution Architect

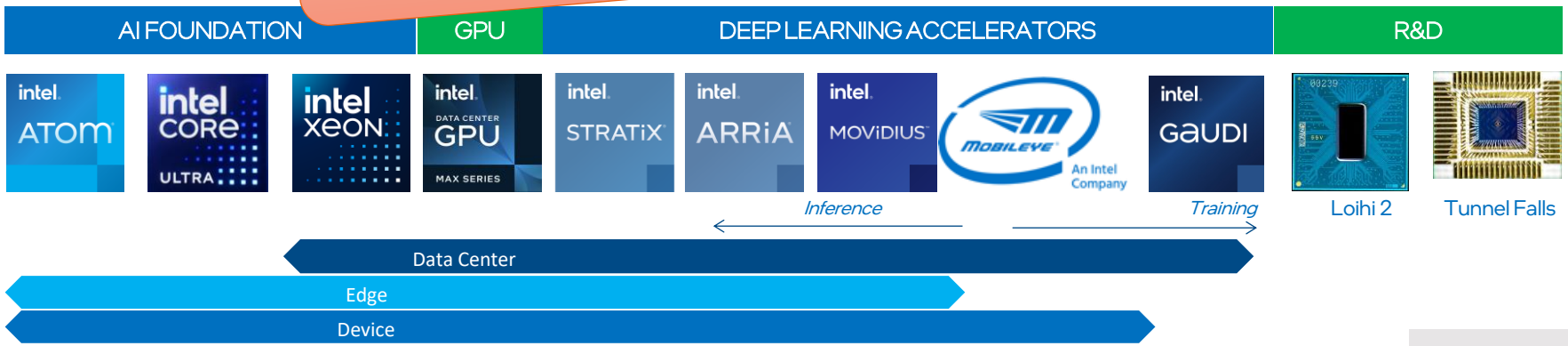
gretchen.stewart@intel.com

Present – 1-28-25

intel.

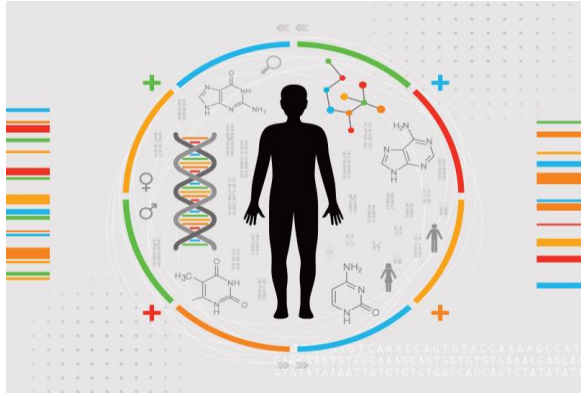
Lead with Software coupled with Fit for Purpose AI Compute

Critical to bring algorithms and technology to the data



Code reuse to ensure best fit architecture

- Software Optimizations – PyTorch, TensorFlow, Python, & more built on **OneAPI**
- Open Programming Model – Open Source



Embracing Open Standards

Open Ecosystem

x86 Ecosystem Advisory Group
intel. AMD

Ultra Ethernet Consortium

Ultra Accelerator Link
(UALink)

Building for Scale and Innovation

Open Software

PyTorch

Open Platform for Enterprise AI

oneAPI

UXL
Unified Acceleration Foundation

Simplify development, production, & adoption across heterogenous compute

Open Platforms

OCP DC-MHS
Data Center Modular Hardware System

OCP Open Rack

Liquid and Immersion Cooling techniques

Common Modular Building Blocks that enable efficiency and sustainability

Intel continues to champion Industry Open Choice

Intel® AI Software Portfolio

Engineer Data → Create Models → Optimize & Deploy

MODIN | SciPy | pandas | NumPy | Spark | Numba

Data Analytics at Scale†

dmlc XGBoost | PyTorch | ONNX RUNTIME | OpenVINO (Write Once Deploy Anywhere) | scikit learn | TensorFlow | DirectML | SigOpt AutoML | intel Neural Compressor

Machine & Deep Learning Frameworks, Optimization and Deployment Tools†

oneAPI

Intel® oneAPI Deep Neural Networks Library | Intel® oneAPI Collective Communications Library | Intel® oneAPI Math Kernel Library | Intel® oneAPI Data Analytics Library

Open, cross-architecture programming model for CPUs, GPUs, and other accelerators

CLOUD & ENTERPRISE | CLIENT & WORKSTATION | EDGE

Intel Developer Cloud and Intel Developer Catalog

Try the latest Intel tools and hardware, and access optimized AI Models

intel oneAPI AI ANALYTICS TOOLKIT

Accelerate End-to-End Data Science and AI

Hugging Face

Intel optimizations and fine-tuning recipes, optimized inference models, and model serving

intel® Geti

Annotation/Training/ Optimization Platform

Note: components at each layer of the stack are optimized for targeted components at other layers based on expected AI usage models, and not every component is utilized by the solutions in the rightmost column

† This list includes popular open source frameworks that are optimized for Intel hardware

oneAPI

Specification and Open Source

Freedom to Make Your Best Choice

- An open alternative to single-vendor/proprietary lock-in enables easy architecture retargeting
- Open, standards-based programming (C++ with SYCL) so software investments continue to add value in future hardware generations

Performance – Realize All the Hardware Value

- Expose and exploit all the cutting-edge features and maximize performance across CPUs, GPUs, FPGAs, and other accelerators.
- Powerful libraries for acceleration of domain-specific functions

Productivity – Develop Performant Code Quickly

- One programming model for all – easy integration with existing code including migration of CUDA code to SYCL
- Based on familiar C++ – no need to learn a new language
- Interoperable with existing HPC standards including Fortran, C/C++, OpenMP, and MPI, as well as Python with a rich set of optimized Python libraries

oneAPI

Open industry initiative driving a vendor-neutral software ecosystem for multiarchitecture accelerated computing.

Now governed by the Linux Foundation.

UXL
Unified Acceleration Foundation

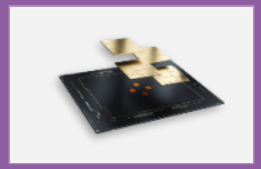
Middleware and Frameworks

PyTorch | TensorFlow | scikit learn | OpenVINO | NumPy | XGBOOST

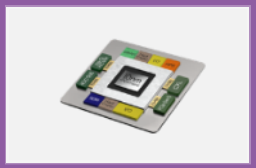
oneAPI Industry Specification

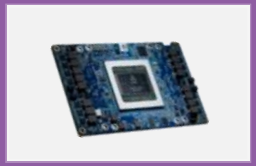
Direct Programming	API-Based Programming		
SYCL (C++)	Math oneMKL	Threading oneTBB	Parallel STL oneDPL
	Analytics/ ML oneDAL	DNN oneDNN	ML Comm oneCCL

Low-Level Hardware Interface (oneAPI Level Zero)


CPU


GPU


FPGA

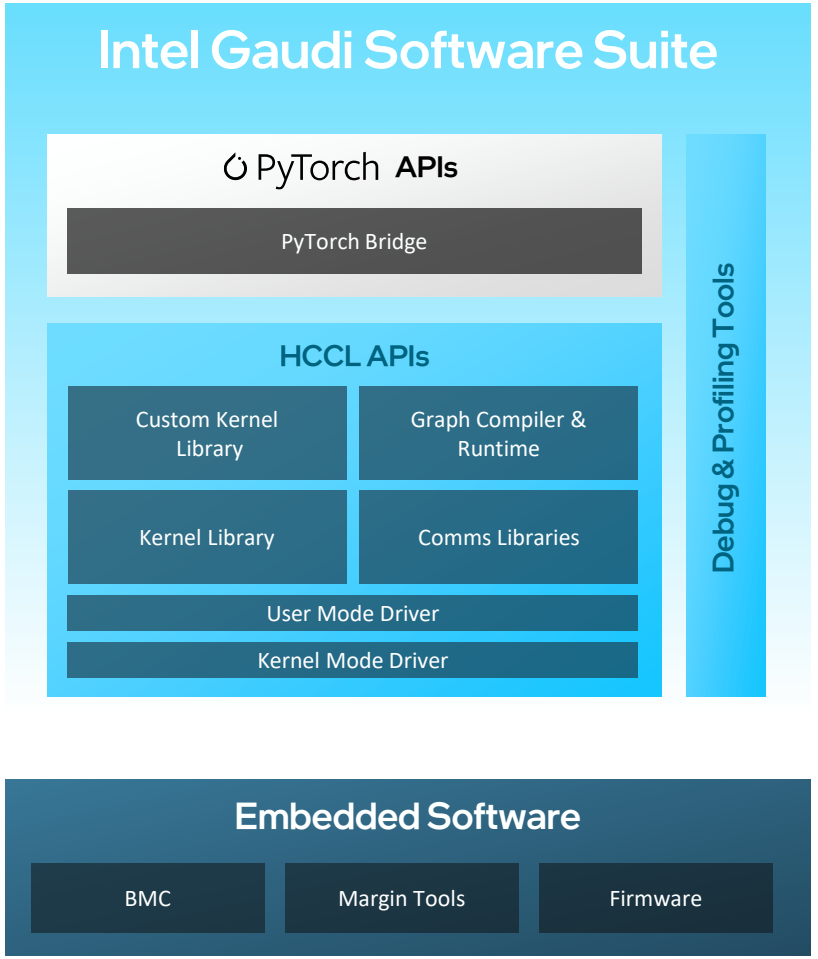

Other Accelerators

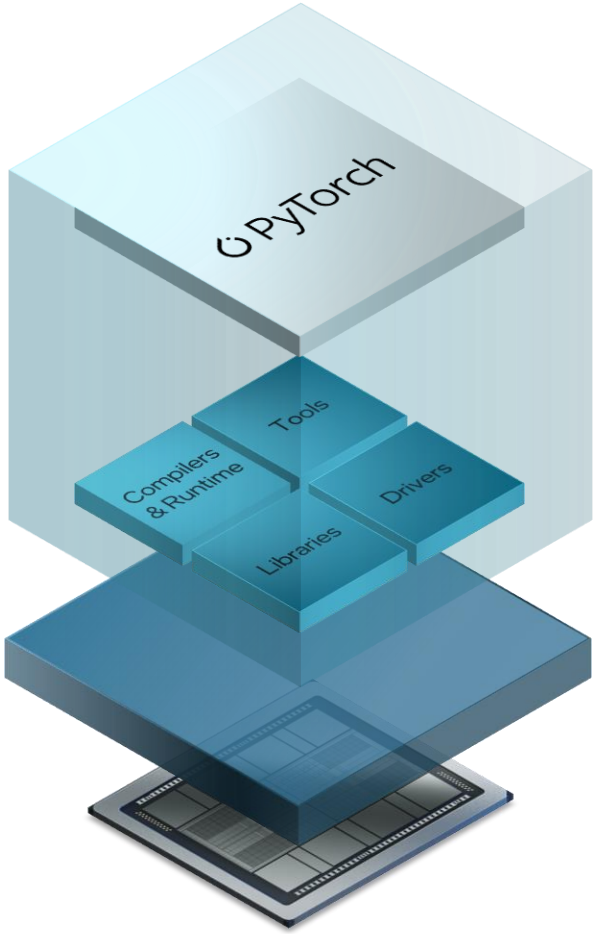
Visit oneapi.io or https://uxlfoundation.org/ for more details

intel GAUDI

Intel Gaudi Software Suite

Ecosystem access to hundreds of 1000s of Gen AI models that run on Gaudi to ease development





Available for download or in the cloud

Run the tools locally

visit intel.com/oneAPI

- Downloads
- Repositories
- Containers

Code Samples, Quick-start Guides, Webinars, Training

Run the tools in Intel® Developer Cloud*

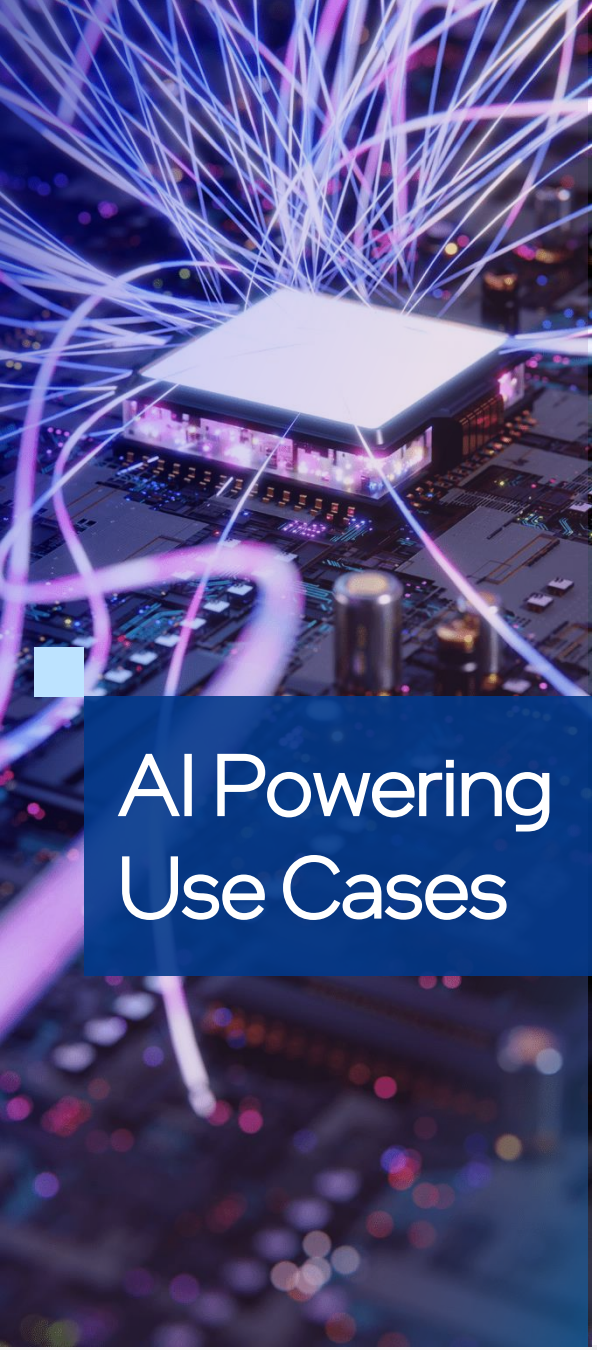
visit cloud.intel.com

- No hardware acquisition
- No download, install or configuration
- Sample code & documentation
- Ready-to-use deployment & development environments
- Access to cutting edge learning resources.

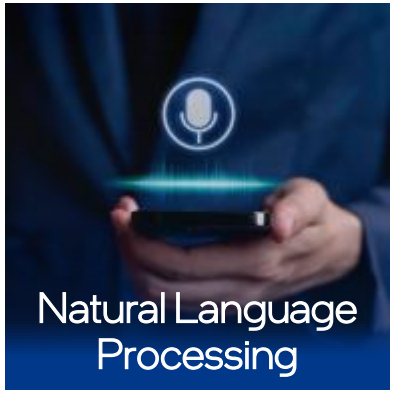
Professional and Community Support Available

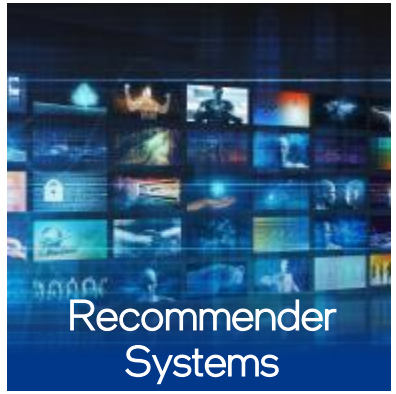
- Download or run tools in the cloud for free
- Every paid version of Intel® oneAPI Base, HPC, and Rendering Toolkit products includes Priority Support
- Intel® Developer Cloud offers Free, Premium (individual), and Enterprise (team) service tiers

*2024.0 tools coming soon to Intel Developer Cloud.

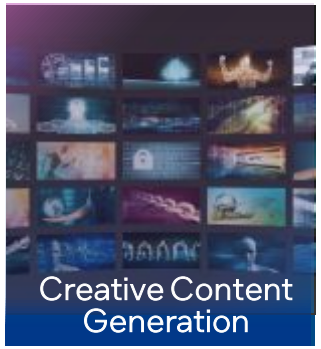


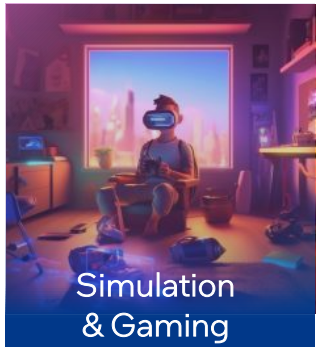
AI Powering Use Cases

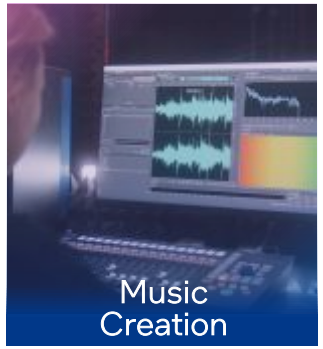


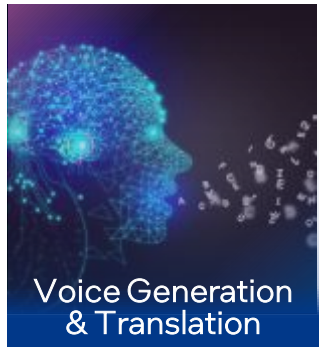


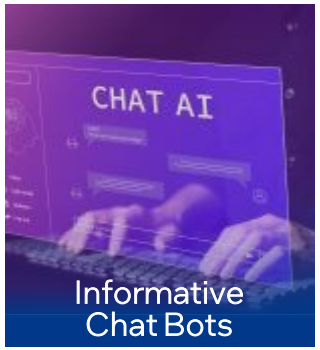
Generative AI and Large Language Models

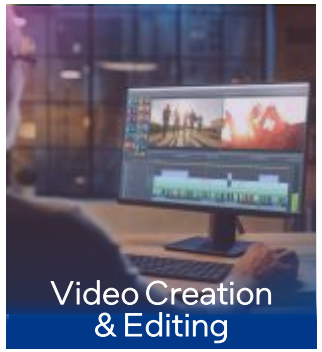


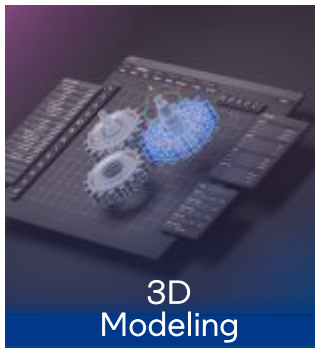


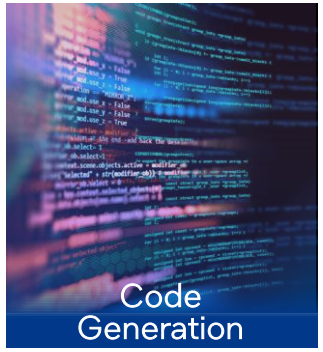


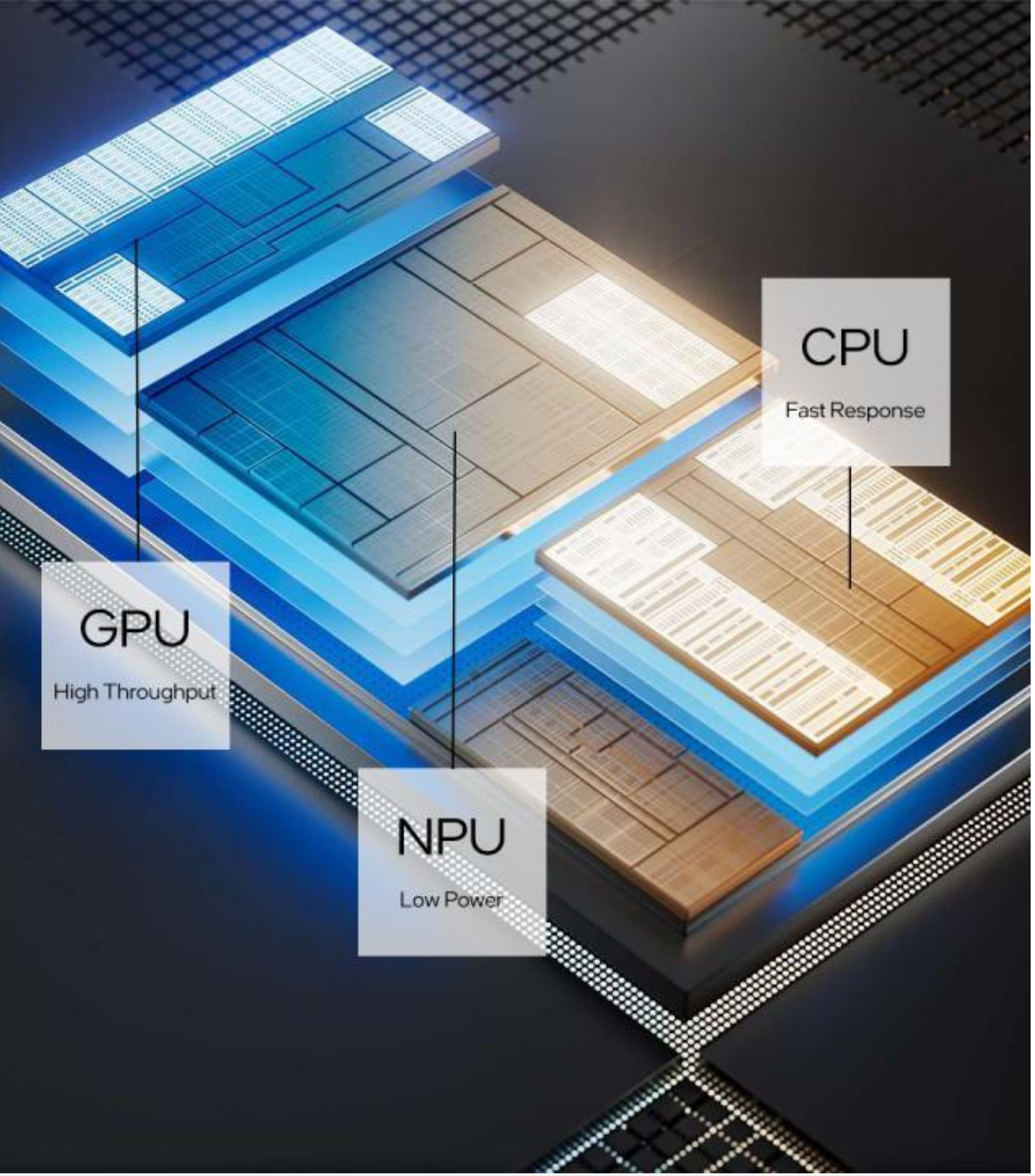












Flexible Applications' Development

Performance Optimization

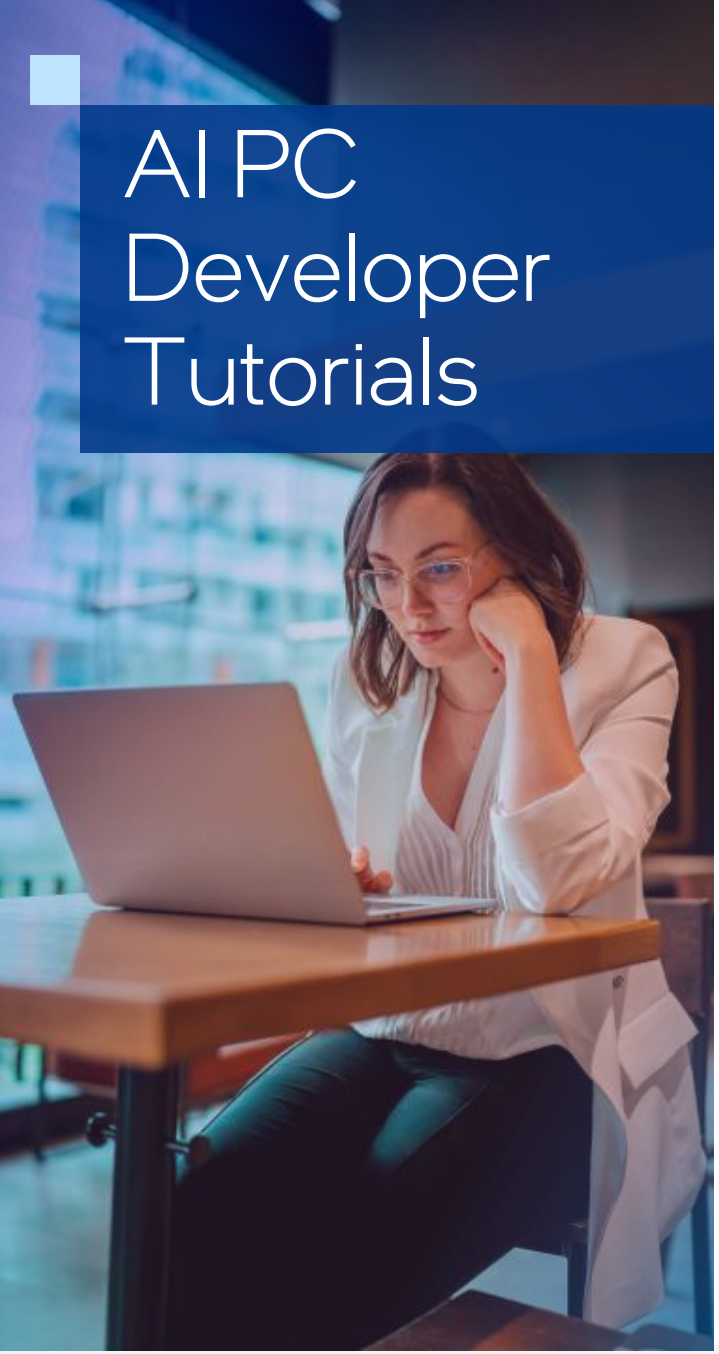
Let your software automatically select the most suitable hardware accelerators, or define a list, allowing applications to achieve better performance

Load Balance Inference

Ensure that available hardware resources are utilized effecting, preventing bottlenecks and maximizing throughput

Flexible Software with Auto Plugin

With hardware technology advancing, auto can adapt to these technologies without rewriting application software



AI PC Developer Tutorials

Kickstart your AI PC applications with GenAI and LLM Jupyter Notebooks

Large Language Model (LLM) Chatbot	▪ Craft chatbots powered by an LLM using the OpenVINO toolkit.
LLM Instruction Following	▪ Run an instruction-following text-generation pipeline.
Latent Consistency Models (LCM)	▪ Learn about image generation using the LCM and the OpenVINO toolkit.
Distil-Whisper Model	▪ Experience automatic speech recognition with this model and the OpenVINO toolkit.
Stable Diffusion* v2	▪ Venture into text-to-image generation and infinite zoom capabilities with Stable Diffusion* v2 and the OpenVINO toolkit.
Bootstrapping Language-Image Pretraining (BLIP)	▪ Use BLIP for visual language processing tasks like visual question answering and image captioning.
MusicGen	▪ Discover a single-stage, autoregressive transformer model that produces high-quality music samples based on text descriptions or audio prompts.
YOLOv8* Optimization	▪ Learn how to convert and optimize YOLOv8* models.

Gen AI Use Cases and Models Supported

- Video Generation
- 3D Modeling
- Image Generator
- Image Segmentation

Visual

- CLIP
- BLIP
- FILM
- Pix2Pix
- Riffusion
- ControlNet
- Zero Scope
- QR code monster
- Segment Anything Model (SAM)
- Latent Consistency Models (LCM)
- Würstchen
- DeepFloyd IF
- DeciDiffusion
- Stable Diffusion

- Chat Bot
- Code Generation
- Search
- Text Classification
- Content Creation
- Instruction Following

Language

- GPT J
- Notus
- LLaVa
- Llama 2 & 3
- BLOOM
- chatGLM
- chatGLM3
- Baichuan 2
- Neural Chat
- StableLM-Epoch-3B
- StableLM-tuned-alpha-3b
- MPT
- Dolly
- Youri
- Qwen
- Mistral
- Zephyr
- RedPajama
- LLM chatbot

- Music Generation
- Text to Audio
- Audio to Text
- Single Voice Conversion

Audio

- BARK
- VITS
- SoftVC
- Whisper
- MusicGen
- AudioLDM
- Distil-Whisper

Example model support includes, but not limited to:

Introducing the OpenVINO Generative AI Github* Repository

Intel vPro® Platform with Intel® Core™ Ultra

Delivers productivity, security, manageability and stability

intel vPRO

Up to 36% lower processor power gen-over-gen

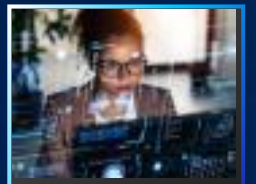
Up to 12x Workstation performance gen-over-gen with Intel® Arc™ Pro drivers

Unmatched ISV ecosystem partnership – **100+ ISVs delivering new experiences, Windows 11 Pro & Copilot**

Single use of Intel vPro® to support a PC remotely can **save carbon emissions equal to 2 years of use** of that PC

Intel® Core™ Ultra ushers in the AI PC era for commercial customers, **enabling IT to transition with confidence**

Intel vPro® can provide **213% ROI over a 3-year period**



Up to 70% of attack surface reduction

Up to 47% better productivity vs 3-year-old PC

Up to 2.2x AI performance gen-over-gen

See www.intel.com/PerformanceIndex for workloads and configurations. Results may vary.

Bringing AI Everywhere

5th Gen Intel® Xeon® Scalable Processors

Intel Acceleration products

intel.

Intel® Xeon® - The Processor Designed for AI

Run any AI code, every workload

5th Gen Intel® Xeon® Scalable Processor

The flexibility of Xeon with the built-in DL performance of an AI accelerator

- Up to **29% higher training** and up to **42% higher inference** performance than our previous generation[1]
- Up to **2.69x higher performance than AMD** EPYC 9654 (96C) and 9754 (128C) processors[2]

Build and deploy AI everywhere

Intel AI software suite of optimized open-source frameworks and tools

Enables out of the box AI performance and E2E productivity

- **5x improvement** on GPT-J in 10 weeks through software optimizations alone[3]
- Optimizing larger models up to **70B parameters** to meet customer SLAs
- Optimized 300+ DL models and 50+ ML and Graph Models

Implement pre-built solutions

Extensive Intel AI products and partnership

Accelerate end customer time to market

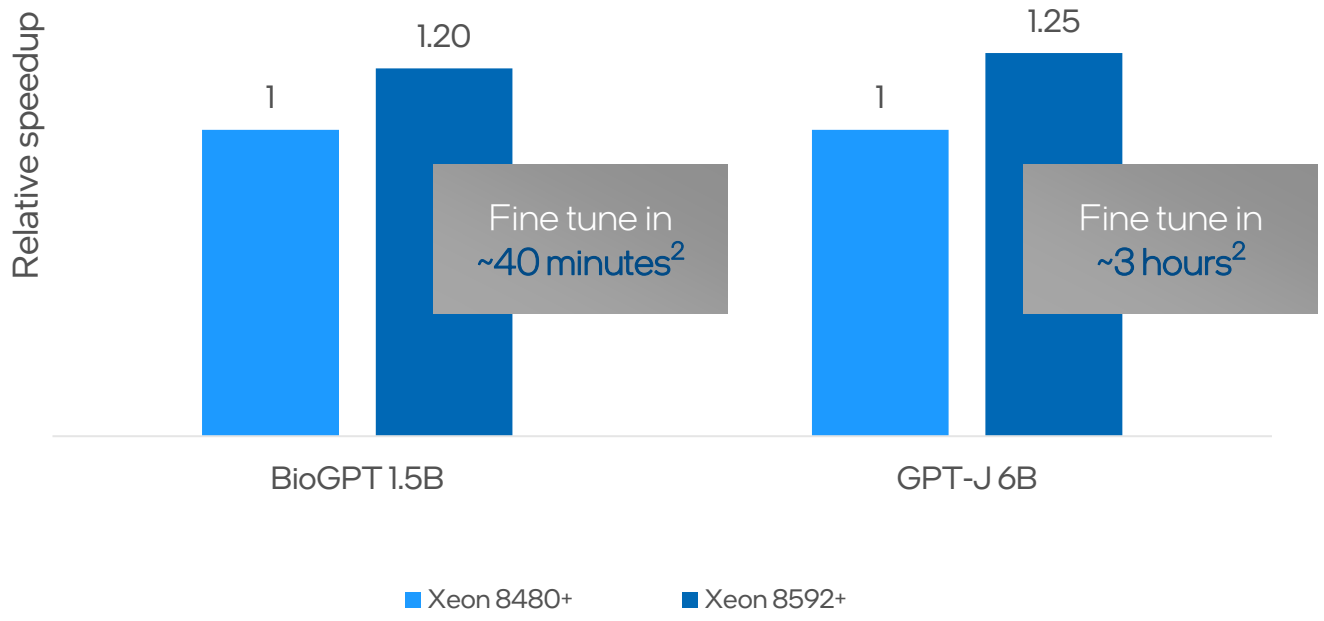
Numenta with Gallium Gaming:
5x Faster than GPUs with Numenta Platform for Intelligent Computing (NuPIC) on Xeon[3]

1. Based on performance gains of 1.1x to 1.29x for training (ResNet50v1.5, BERT-Large, SSD-ResNet34, RNN-T, MaskRCNN, and DLRM) and 1.19x to 1.42x for inference (ResNet50v1.5, BERT-Large, SSD-ResNet34, RNN-T (BF16 only), Resnext101 32x16d, MaskRCNN (BF16 only), DistilBERT) compared to 4th Gen Intel® Xeon® processor. See A15-A16 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
2. Based on performance gains of 1.19x to 2.69x with Intel® Advanced Matrix Extensions (Intel® AMX) for inference on GPT-J, LLaMA-2 13B, DLRM, DistilBERT, BERT-Large, and ResNet50v1.5 compared to AMD EYPC 9654 and 9754. See A201, A202, A208-A211 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
3. See backup for workloads and configurations. Results may vary.

Fine Tune in minutes to hours on 5th Gen Intel® Xeon® Scalable processors

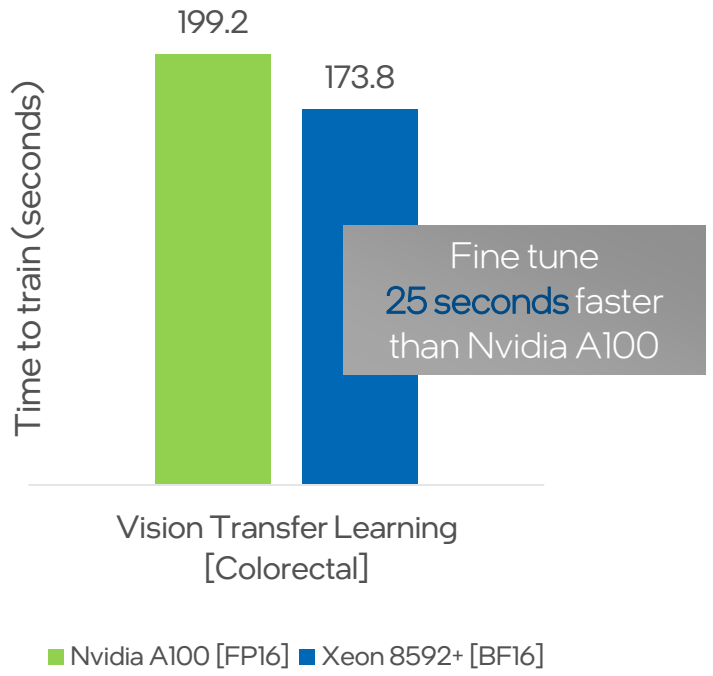
Fine tuning time-to-train speedup
Intel® Xeon® Platinum 8592+ processor
vs. prior generation (BF16, IPEX)[1]

Higher is better



5th Gen Xeon can outperform
Nvidia A100 for vision transfer learning
(Colorectal dataset)[2]

Lower is better



1 See [A9, A10] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
2 See backup for workloads and configurations. Results may vary.

Intel® Advanced Matrix Extensions (Intel® AMX) Acceleration Engine

What is Intel® AMX?

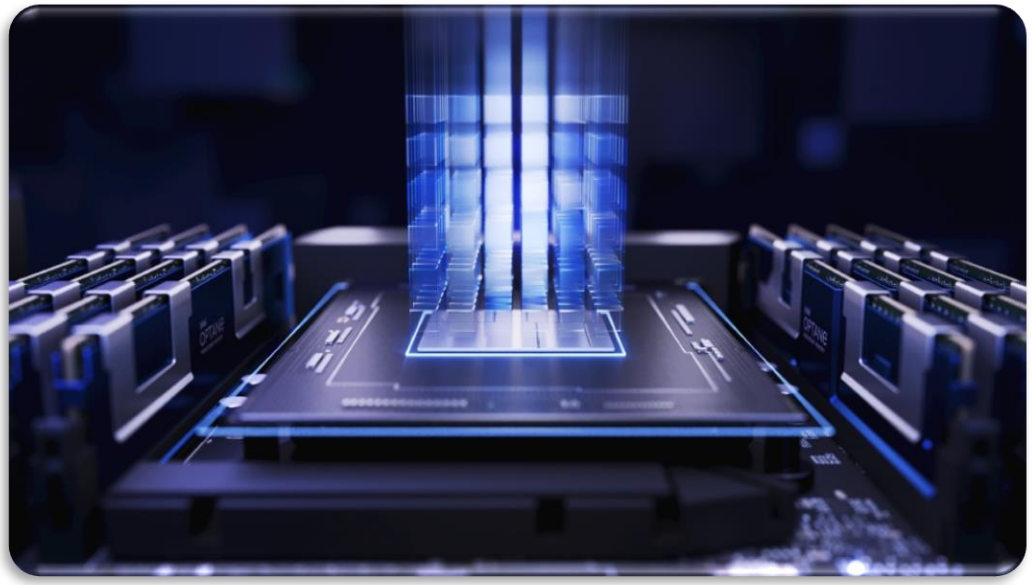
- Intel® AMX is a built-in accelerator that improves the performance of **deep learning** training and inference on 4th Gen Intel® Xeon® processors
- Advanced matrix multipliers are integrated into **EVERY** core

Business Value

- Help to **lower customers' TCO** as it raises the bar for where they can meet AI SLAs without the need for a discrete accelerator

Software Support

- Works **out-of-box** on industry-standard frameworks, toolkits and libraries such as PyTorch, TensorFlow, and OpenVINO
- **vSphere 8 supports Intel AMX**



PyTorch Training and Inference

10x

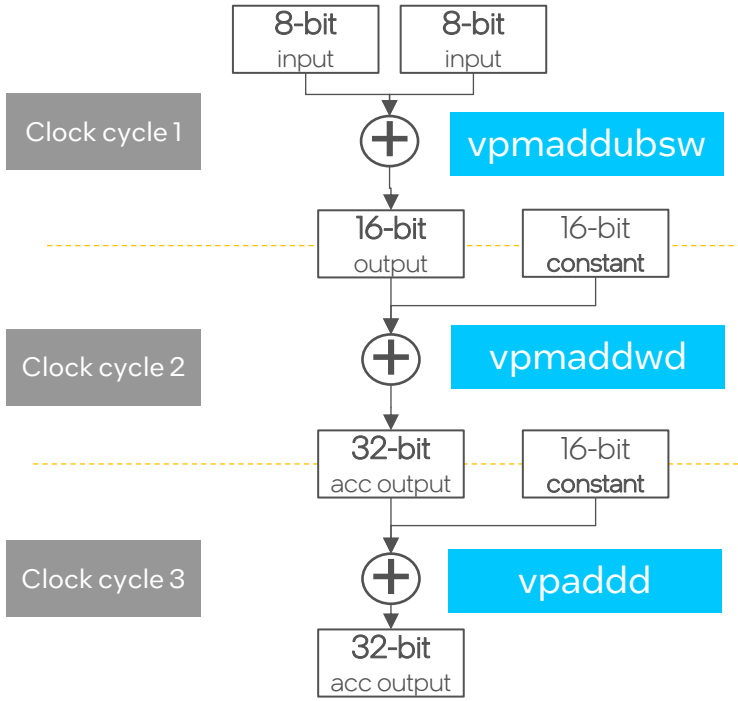
PyTorch for both real-time inference and training performance with built-in Intel AMX (BF16) on 4th Gen Intel® Xeon® Scalable processors vs. the prior generation (FP32)

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

One Processor for Scalar, Vector, and Matrix

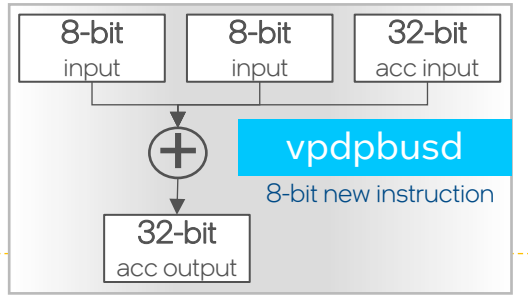
Intel® AVX-512

85 int8 ops/cycle/core
with 2 FMA



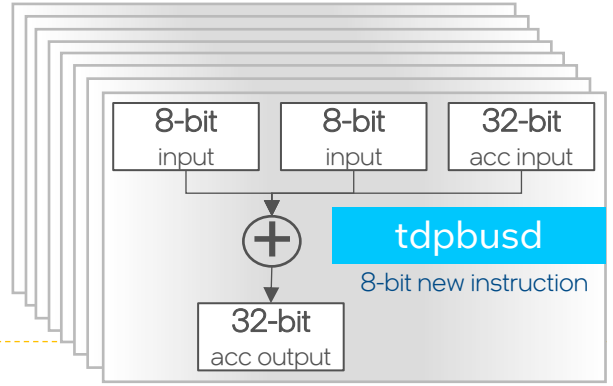
Intel® AVX-512 (VNNI)

256 int8 ops/cycle/core
with 2 FMAs

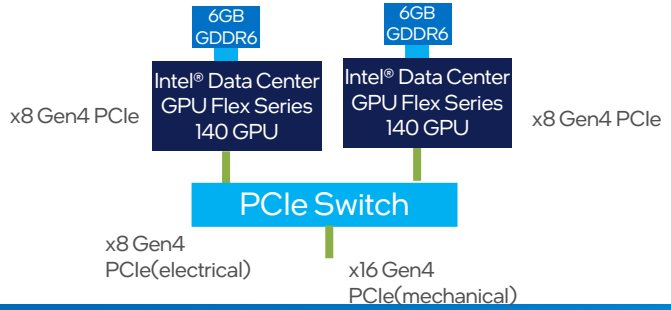


Intel® AMX

2048 INT8 ops/cycle/core
Multi-fold MACs in one instruction



Intel® Data Center GPU Flex Series 140

Card Design	Intel® Data Center GPU Flex Series 140 
Card TDP	**Board Power: 75W**
Card Specifications GPU	Half height, half length, single-wide, Passive cooling Intel® Data Center GPU Flex Series 140
GPU's Per Card	2
Memory w/ECC	Capacity: 12GB (6GB/GPU) Mem xfer Rate: 1750GT/s Mem Bus Width: 96 bits/GPU
Fixed Function Media Units (Per Card)	4 (2 per GPU): 28 transcode streams H.265 1080p60 1:1
Supported Usecases	Media transcode, Visual Inference/Media Analytics, VDI
GPU Throughput (Peak)	•FP32: 8.0 TFLOPs •FP16: 52 TFLOPs •INT8: 105 TOPs •INT4: 210 TOPs
Product Availability	3 years *
Operating System	Linux: Ubuntu, RHEL Windows: WinServer 2019 & 2022, WinClient
Host CPU Support	Whitley- Ice Lake (ICX) & Eagle Stream- Sapphire Rapids (SPR)
Branding/Channel Partners	Intel Branded Card

Designed for high-density use cases

Smaller card, low profile, and lower power

- **VDI** - uses SRIOV with no SW licensing fee leading to lower TCO
- **Media Delivery** - Deliver highest **Media** density/performance and TCO leadership within the Flex Series GPUs
- **Visual Inference/Media Analytics -** good solutions where primarily work is media processing/transcode and light/occasional inference

* From launch date Sep'22

Intel® Data Center GPU Flex Series 170

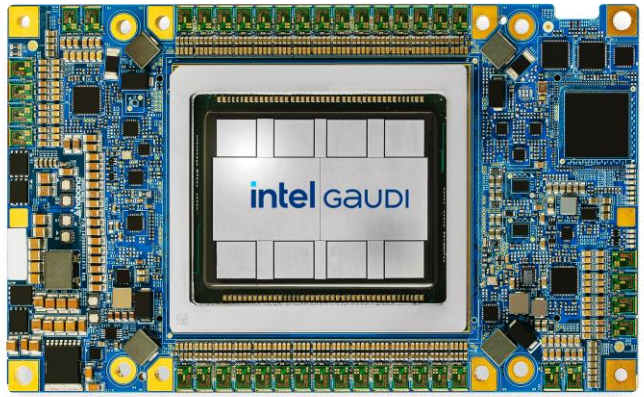
Card Design	Intel® Data Center GPU Flex Series 170 16GB GDDR6 Intel® Data Center GPU Flex Series 170 GPU x16 Gen4 PCIe
Card TDP	**Board Power: 150W**
Card Specifications	Full Height, ¾ length, single-wide, Passive cooling
GPU	Intel® Data Center GPU Flex Series 170 GPU
GPU's Per Card	1
Memory w/ECC	Capacity: 16GB Mem xfer Rate: 2250GT/s Mem Bus Width: 256 bits
Fixed Function Media Units (Per Card)	2 (2 per GPU): 14 transcode streams H.265 1080p60 1:1
Supported Usecases	Media transcode, Visual Inference/Media Analytics, VDI
GPU Throughput (Peak)	• FP32: 16.8 TFLOPs • FP16: 128 TFLOPs • INT8: 256 TOPs • INT4: 512 TOPs
Product Availability	3 years *
Operating System	Linux: Ubuntu RHEL Windows: WinServer 2019 & 2022, WinClient
Host CPU Support	Whitley- Ice Lake (ICX) & Eagle Stream- Sapphire Rapids (SPR)
Branding/Channel Partners	Intel Branded Card

For high performance use cases
Delivers highest workload performance within the Flex Series GPUs

- **VDI** - uses SRIOV with no SW licensing fee leading to lower TCO
- **Visual Inference/Media Analytics** - and balances dedicated media and matrix compute assets for good visual inference performance
- **Inference** run real time and batch plus acceleration for running medium sized LLMs (upto ~16B like Llama 2)

* From launch date Sep'22

Architected for Gen AI Performance & Productivity – Gaudi3



Designed for AI

diving greater efficiency & performance

64
Tensor Processor Cores (5th gen)

8
Matrix Math Engines

Increased memory for LLM efficiency and cost effectiveness

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

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

Massive, flexible on-chip networking

Open standard vs. proprietary InfiniBand

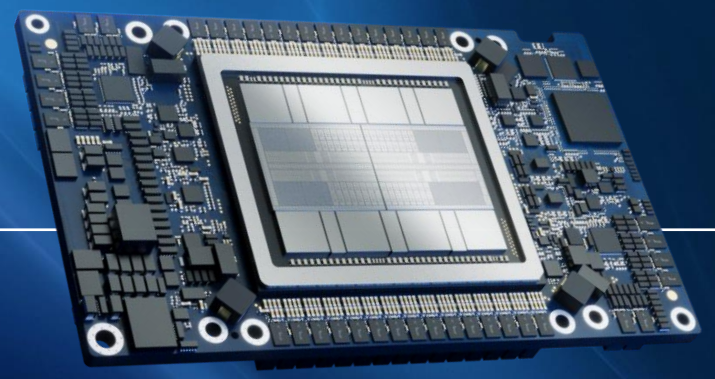
24x 200 GbE
industry-standard RoCE Ethernet ports

PCIe 5
x 16

intel GAUDI

1.5x faster time-to-train

Average projection for Intel Gaudi 3 accelerator vs. Nvidia H100, running common Large Language Models*



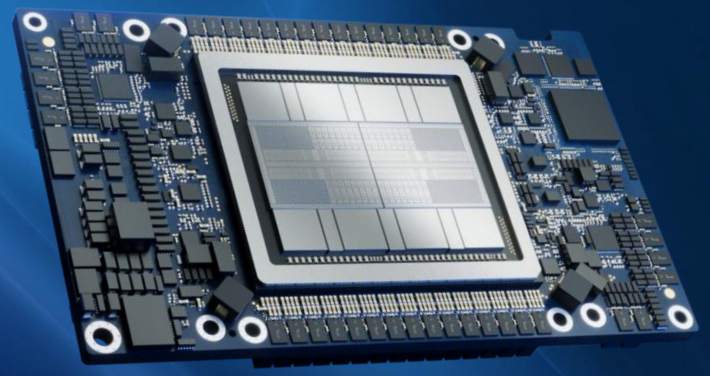
* NV H100 comparison based on : https://nvidia.github.io/TensorRT-LLM/performance/perf-overview.html, May 28th 2024 → "Large Language Model" tab
Vs Intel Gaudi 3 projections for LLAMA2-7B, LLAMA2-13B & GPT3-175B as of 3/28/2024. Gaudi 3 performance was projected using static batch assumptions. Results may vary.

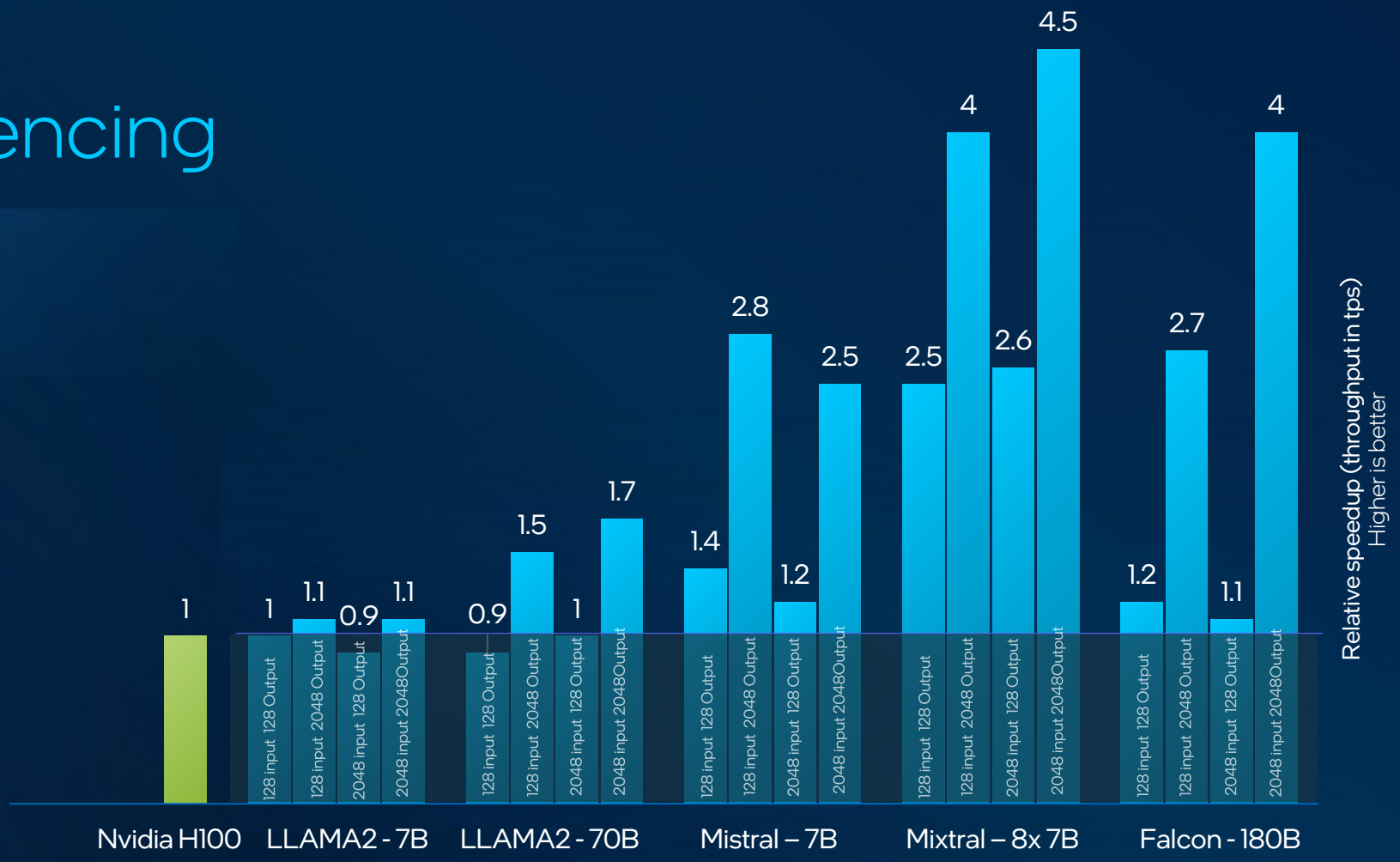
Link to data

intel GAUDI

2x faster inferencing

Average projection for Intel Gaudi 3 accelerator vs. Nvidia H100, running common Large Language Models*





Source for Nvidia performance: Overview — tensorrt_llm documentation (nvidia.github.io), June 2024. Reported numbers are per GPU.
Intel Gaudi 3 projections by Habana Labs, using static batch assumptions, April 2024. Results may vary.

Link to data

Enterprise Customer Proof Point: Boston Consulting Group

BCG deployed **semantic knowledge discovery solution** using our GenAI platform, with data privacy at the core

Delivered a new search paradigm with meaningful impact

Chat-based, semantic querying

Page-level results w/ insight summary

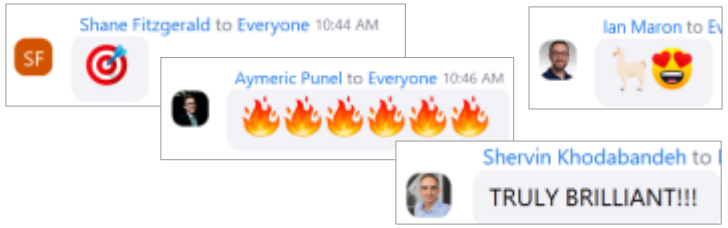
41% in user satisfaction

25% in search result relevance

39% in work completion rate

Viral users' interest and engagement – scaling solution is a non-issue

Vibrant discussion and excitement during live demo



"BOOM!!!! I found one of the hardest things to find on KD, I never found it despite searching multiple times [with the old tool]"

- Managing Director & Senior Partner

"This is life changing. Search will never be the same again. My clients are going to be blown away!"

- Managing Director & Senior Partner

Ability to deliver speed, scale, and privacy

0 Number of times Intel saw customer's raw data due to encrypted tokenization

6 Weeks to train model and prepare for enterprise employee use

8 TB of text and image confidential data to train LLM

12 Weeks in total to provide solution to enterprise employees

Learn more about our collaboration with BCG here- https://www.prnewswire.com/news-releases/intel-and-bcg-announce-collaboration-to-deliver-enterprise-grade-secure-generative-ai-301821547.html

Delivering Customer Value

NETFLIX

2x Improved performance[1]

Using Xeon and its built-in AI accelerator engines, Netflix deployed a convolutional neural network for intelligent content downscaling.

"Performance improvements mean huge savings in cloud infrastructure cost. "

Amer Ather
Cloud and Studio
Performance Engineer

Numenta

Gallium STUDIOS

5x Faster than GPUs with Numenta Platform for Intelligent Computing (NuPIC) on Xeon[2]

NuPIC with Intel Xeon enables Gallium Studios to run AI models with incredible performance on CPUs in their new breakthrough game, Proxi.

"Proxi uses LLMs to deliver AI simulated agents generated from players' uploaded memories. Since partnering with Numenta, we have been able to increase speed and accuracy and lower costs"

Lauren Elliott
CEO, Gallium Studios

1 See https://community.intel.com/t5/Blogs/Tech-Innovation/Tools/Deploying-AI-Everywhere-at-Netflix/post/1528334
2 See backup for workload and configurations. Results may vary.

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 AGILEX, intel ethernet — Edge and Network AI Inference
	intel core, intel core ULTRA w/VPU, intel ARC GRAPHICS — Client AI Usages

intel®