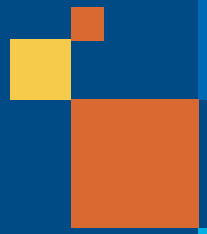
AI on Intel Architecture

Dr. Séverine Habert
Severine.habert@intel.com



Let’s play a little game: which image is real?

ALL images fully generated by Stable Diffusion SDXL

AI has made incredible progress in the last years

Typical Domains of AI (for the last 10 years)

COMPUTER VISION

- Ability to understand the visual world

Classification

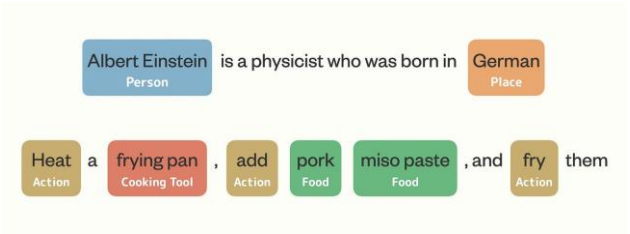
Object Detection

Instance or Semantic segmentation

NATURAL LANGUAGE PROCESSING (NLP)

- Ability to understand the written world

Translation



Entity Name Recognition

Generative AI

- End of 2022, ChatGPT was released, and the generative AI (genAI) craze started!
- Generative AI is the ability for the AI model to create contents (text, image , music, code …)

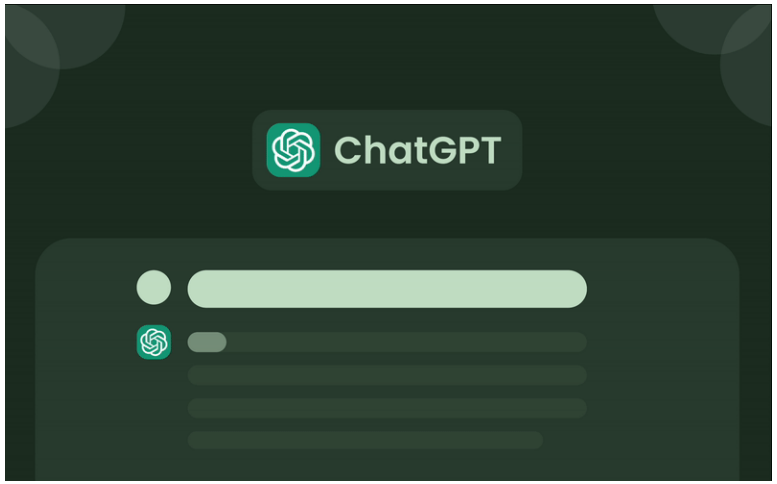
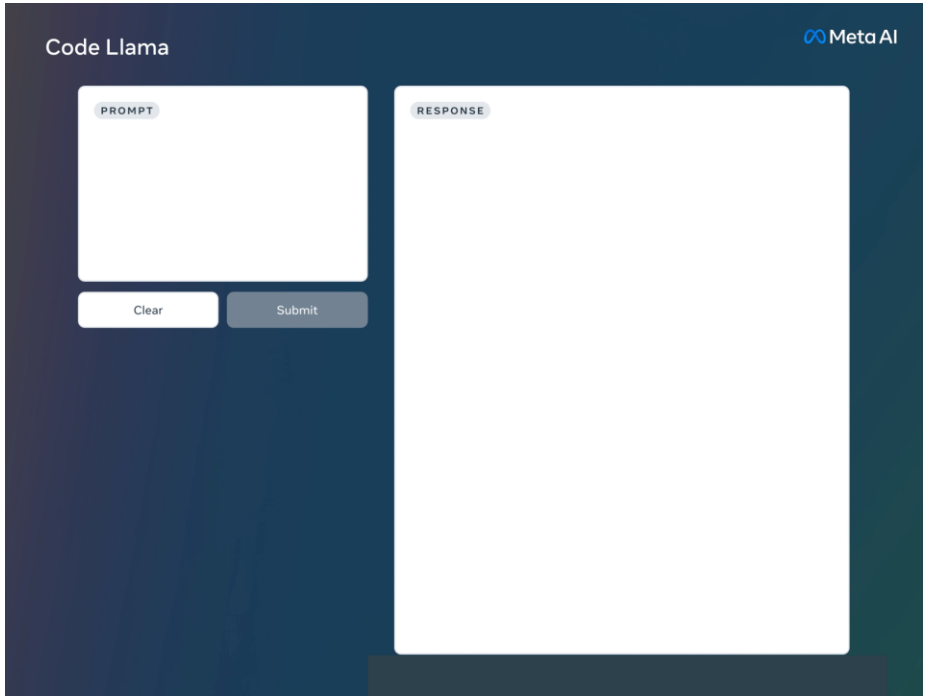


Image fully generated by Stable Diffusion SDXL, a text-to-image AI

The new trend – generative AI

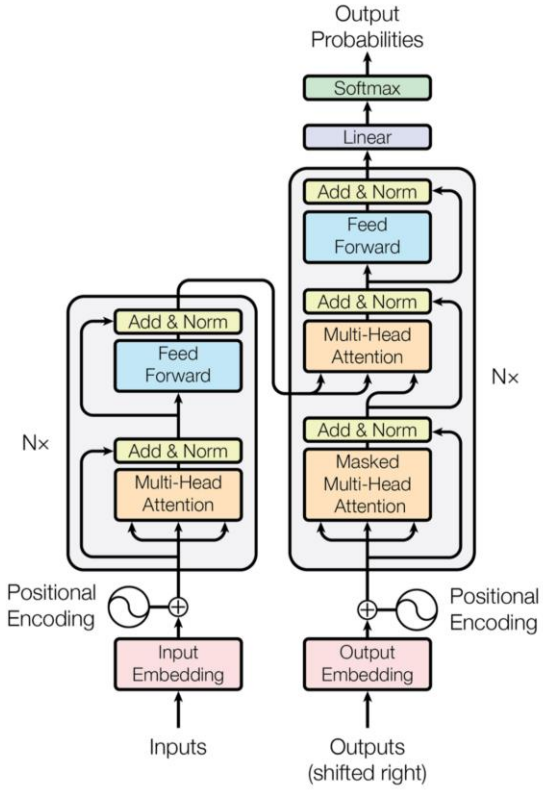
Art generation

Code generation



Transformers

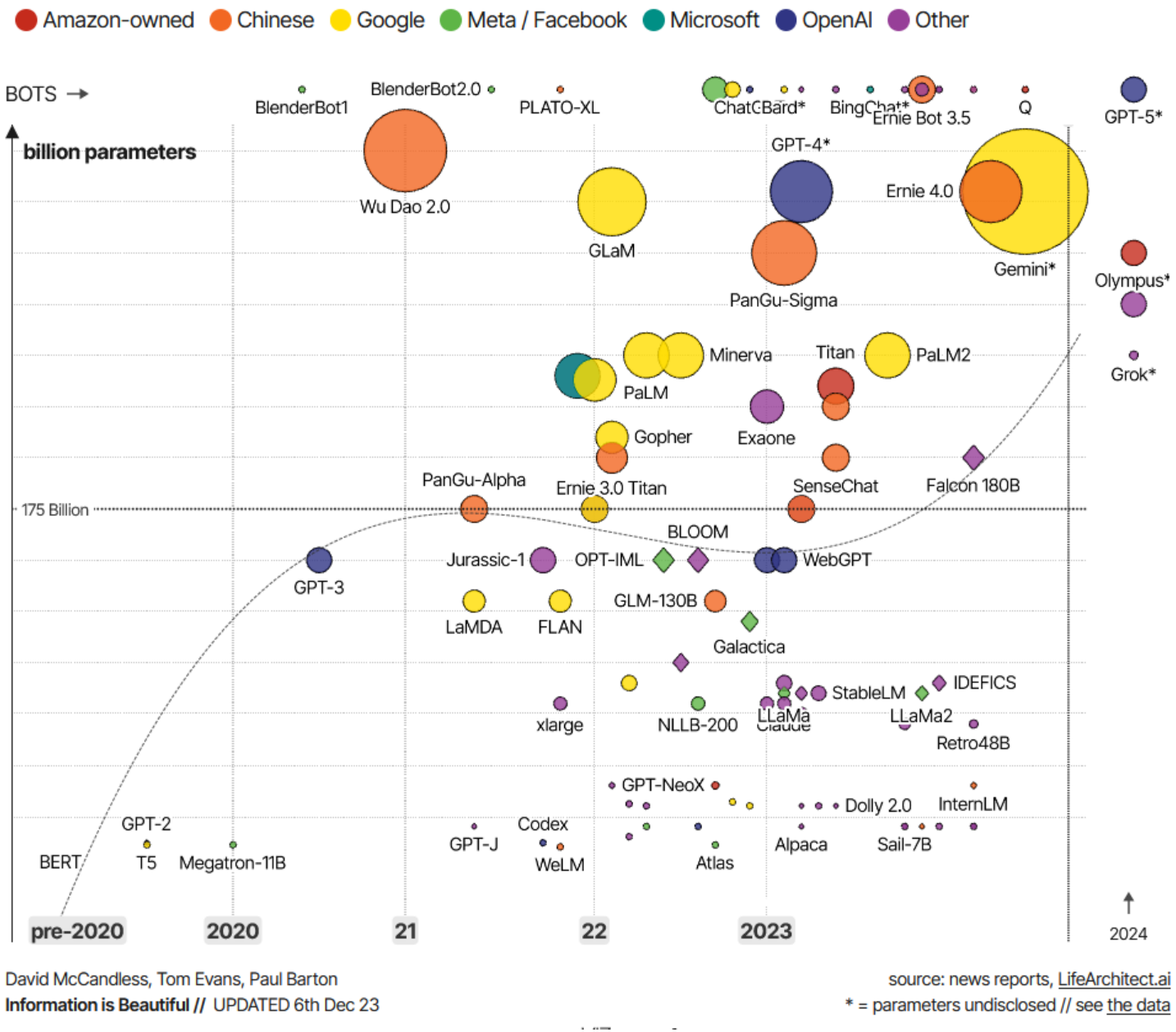
- Transformer architecture is the base for NLP and genAI (e.g., BERT, LLM, …)
  - Composed of 2 building blocks: encoder and decoder



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

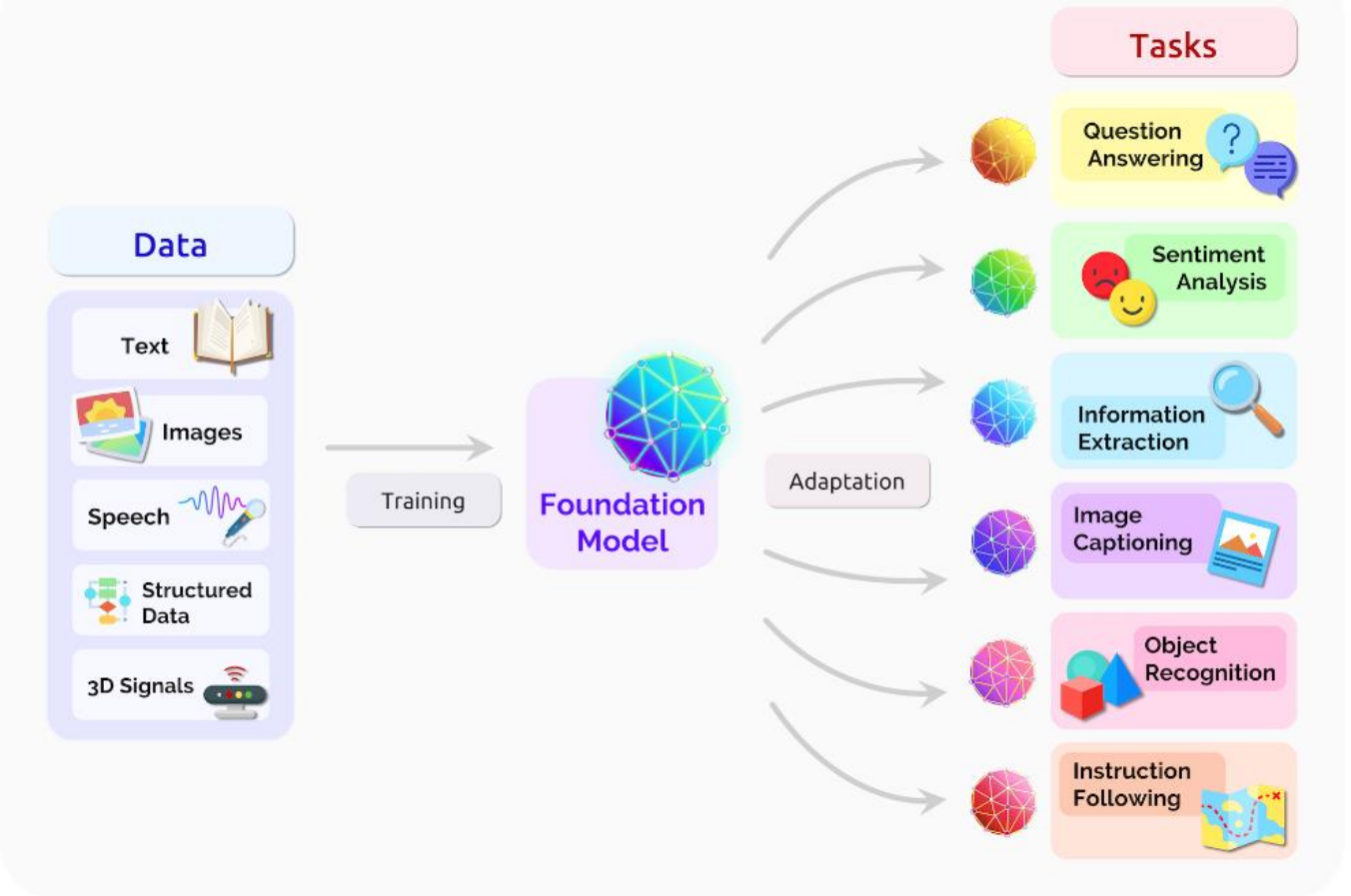
Model Size, way up

- Model keep growing in size
- Trained in self-supervised manner on web-scale quantities of unlabeled data
- Starting with Bert was 0.3B in 2018
- This is logarithmic scale!



Foundational Model

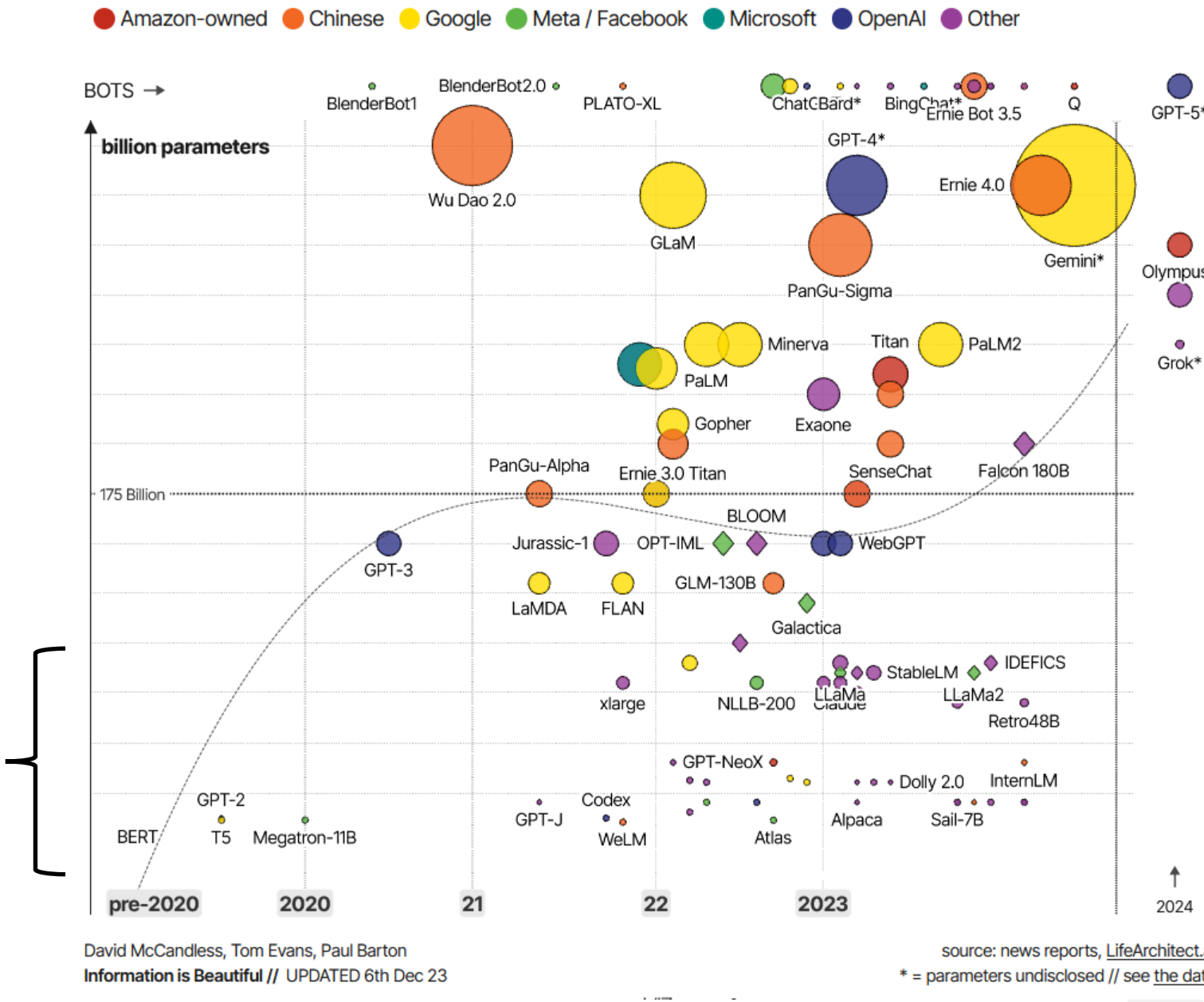
- From the first iterations of Large Language Model such as GPT, model started to be large enough to abstract concepts and language
- They are coined as Foundational Models
- This ONE model can then be adapted to a wide range of downstream tasks



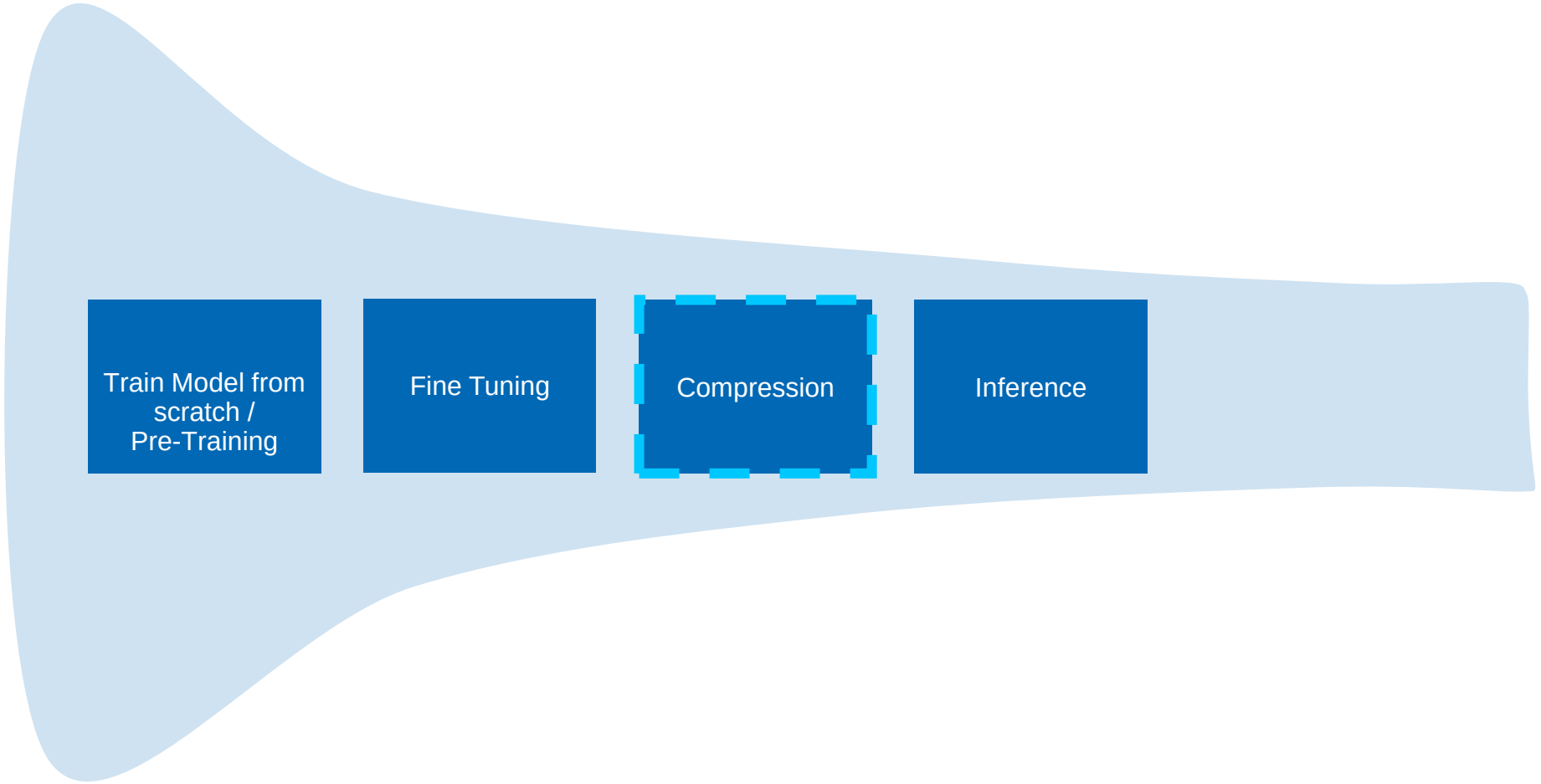
Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Recent trend is to scale down

- Push from the open-source community
- Models are trained on better quality (and smaller) dataset
- Trained on “smaller” infrastructures
- Mixture of Experts are also the trend



Deep Learning Funnel Pipeline



HW for AI

Intel Generative AI Products

LFM pre-training and batch inferencing

<20b parameters (training)

Fine tune on Gaudi 2

Fine tune on GPU Max

<10b parameters (fine tuning in hours)

Fine tune on single or dual socket workstation (4-8TB) **and Servers** (8TB per node, multi-node)

DL inference on Gaudi 2

Fine tune on GPU Max

Inference on Xeon, Core, ARC

Intel® Xeon® Scalable Processors

The Only Data Center CPU with Built-in AI Acceleration

Intel Advanced Vector Extensions 512

Intel Deep Learning Boost (Intel DL Boost)

Intel Advanced Matrix Extensions

Shipping

Cascade Lake

New Intel DL Boost (VNNI)
New memory storage hierarchy

Ice Lake

January 2023

Sapphire Rapids

Intel Advanced Matrix Extensions (AMX) extends built-in AI acceleration capabilities on Xeon Scalable

December 2023

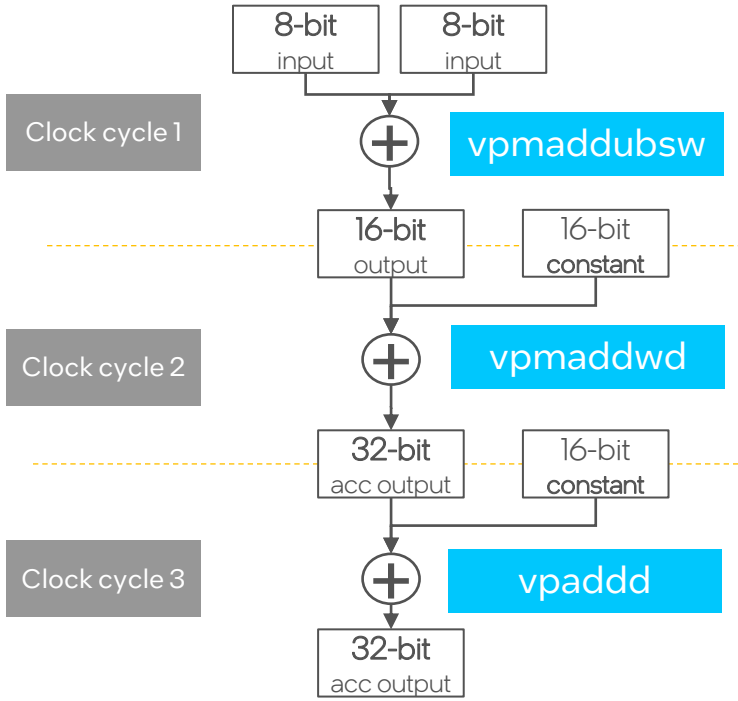
Emerald Rapids

Leadership performance

One Processor for Scalar, Vector, and Matrix

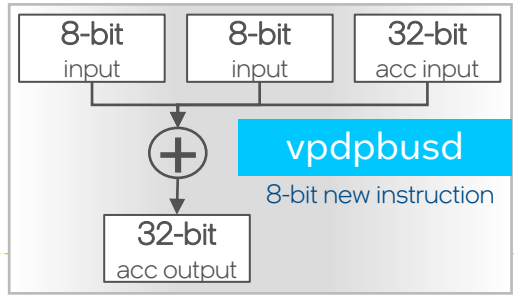
Intel® AVX-512

85 int8 ops/cycle/core
with 2 FMA



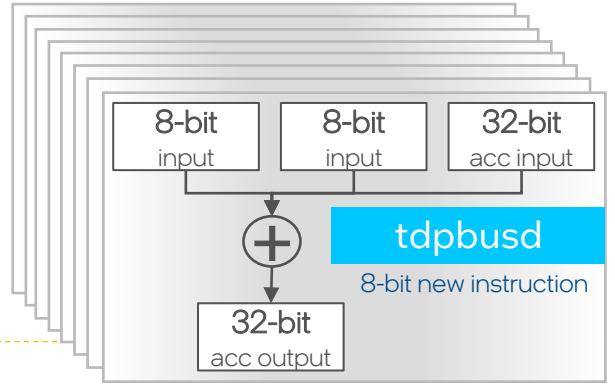
Intel® AVX-512 (VNNI)

256 int8 ops/cycle/core
with 2 FMAs



Intel® AMX

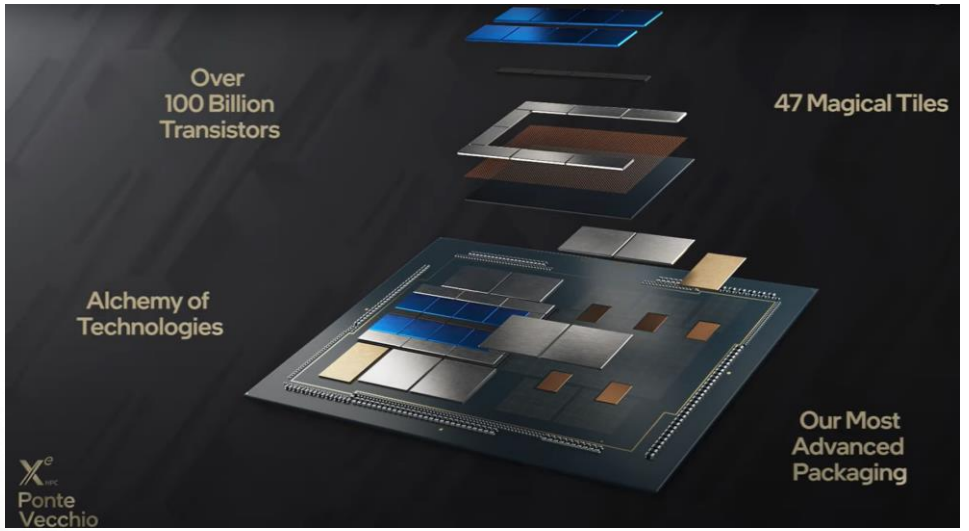
2048 INT8 ops/cycle/core
Multi-fold MACs in one instruction



Xe HPC (Ponte Vecchio)

Leadership Performance for Data-level Parallel AI Workloads

>40 active tiles, over 100 billion transistors integrated into a single package



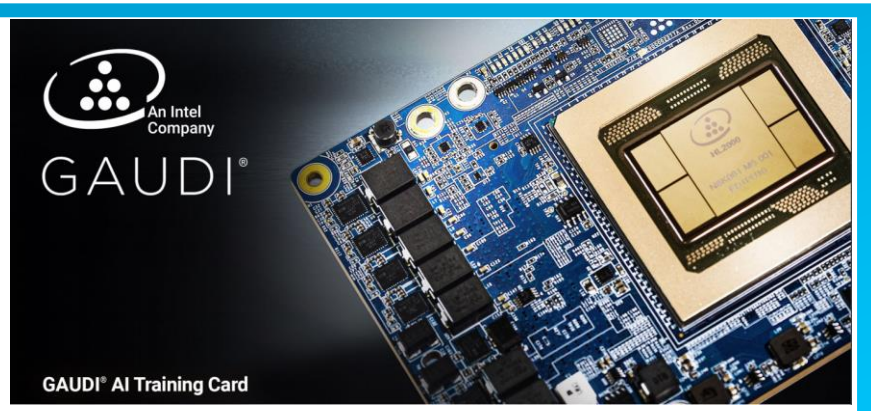
Powering New Phase of SuperMUC-NG at Leibniz Supercomputing Centre (LRZ)

https://www.youtube.com/watch?v=JzbN1IOAcwY

Habana – an Intel Company

intel habana

Deep Learning ASIC for Training and Inference



All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

Intel HW @ LRZ

SuperMUC-NG Phase 2, LRZ, Germany

- Among the fastest HPC systems in Germany as of today
- Broad AI/HPC userbase to drive life & environmental sciences
- Software stack enabled by oneAPI for easy portability
- Lenovo ThinkSystem SD650-I V3 Neptune DWC servers
- Hot water cooling to increase efficiency and lower the TCO
- Assembled in Europe to reduce carbon footprint and shipping timelines

Compute			Fabric	
480 CPUs	960 GPUs	240 Nodes	12 TB/s Peak Injection Bandwidth	6 TB/s Peak Bisection Bandwidth

Memory				Storage		
123 TB DDR Capacity	147 TB/s Peak DDR BW	123 TB HBM Capacity	3.1 PB/s Peak HBM BW	1 PB DAOS Capacity	750 GB/s DAOS Bandwidth	42 DAOS Nodes

Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

intel Lenovo

Intel® Tiber™ Developer Cloud

Boutique AI Cloud Services

Build & Deploy AI at Scale

Develop models, applications & solutions, deploy production workloads at scale. Deliver **10-100X** more performance using common tools.

Maximize AI Compute Resources

Choose the best accelerator for every use case for optimal price-performance.

Customers gained **up to 400%** cost savings for select AI workloads vs. on-prem or another CSP.[2]

Open Software, Open Platform Advantage

Provides choice in hardware. Supports a wide range of optimized models, frameworks & tools. Compatible with open source LLMs & GenAI products.

Managed, Cost-Efficient Infrastructure

Cloud interface(s)	**Intel Cloud Portal & APIs**
Cloud Services	**AI Infrastructure & Platform Services** (BMaaS, VMs, Intel® Kubernetes Service, storage…)
Cloud Infrastructure	**Core Services**

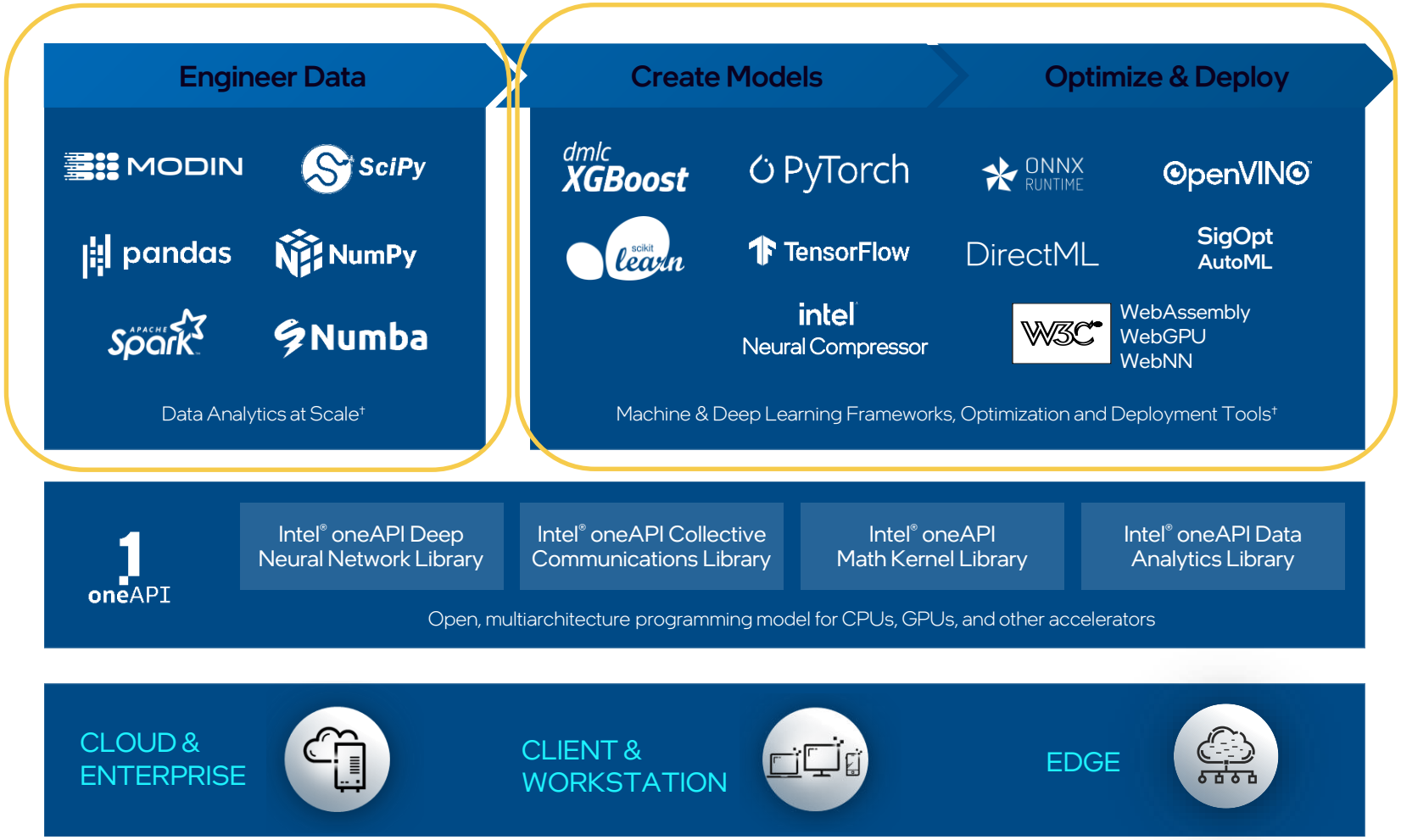
Used by developers, companies & partners

Intel® Tiber™ Portfolio of Business Solutions

1. Performance varies by use, configuration, & other factors. Performance results are based on testing as of dates shown in configurations & may not reflect all publicly available updates. Learn more at www.Intel.com/PerformanceIndex & intel.com/content/www/us/en/developer/articles/technical/software-ai-accelerators-ai-performance-boost-for-free.html.
2. Prediction Guard case study, CIO.com Seekr article. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

SW for AI

Intel AI software portfolio



Accelerate end-to-end data science and AI

Intel® Developer Cloud and Intel® Developer Catalog

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

cnvrg.io

Full stack ML operating system

Intel® Geti™

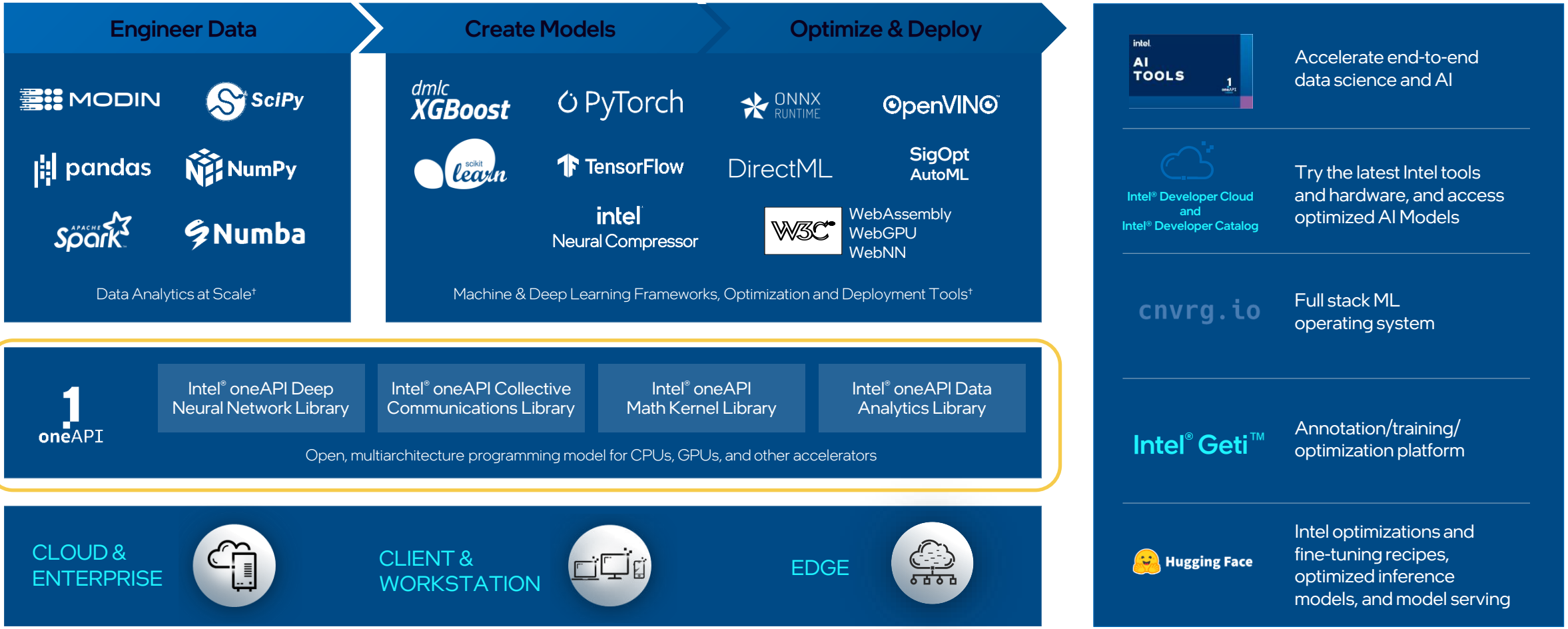
Annotation/training/ optimization platform

Hugging Face

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

This list includes pod for open-source frameworks that are optimized for Intel hardware.

Intel AI software portfolio



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

oneAPI

One Programming Model for Multiple Architectures & Vendors

Freedom to Make Your Best Choice

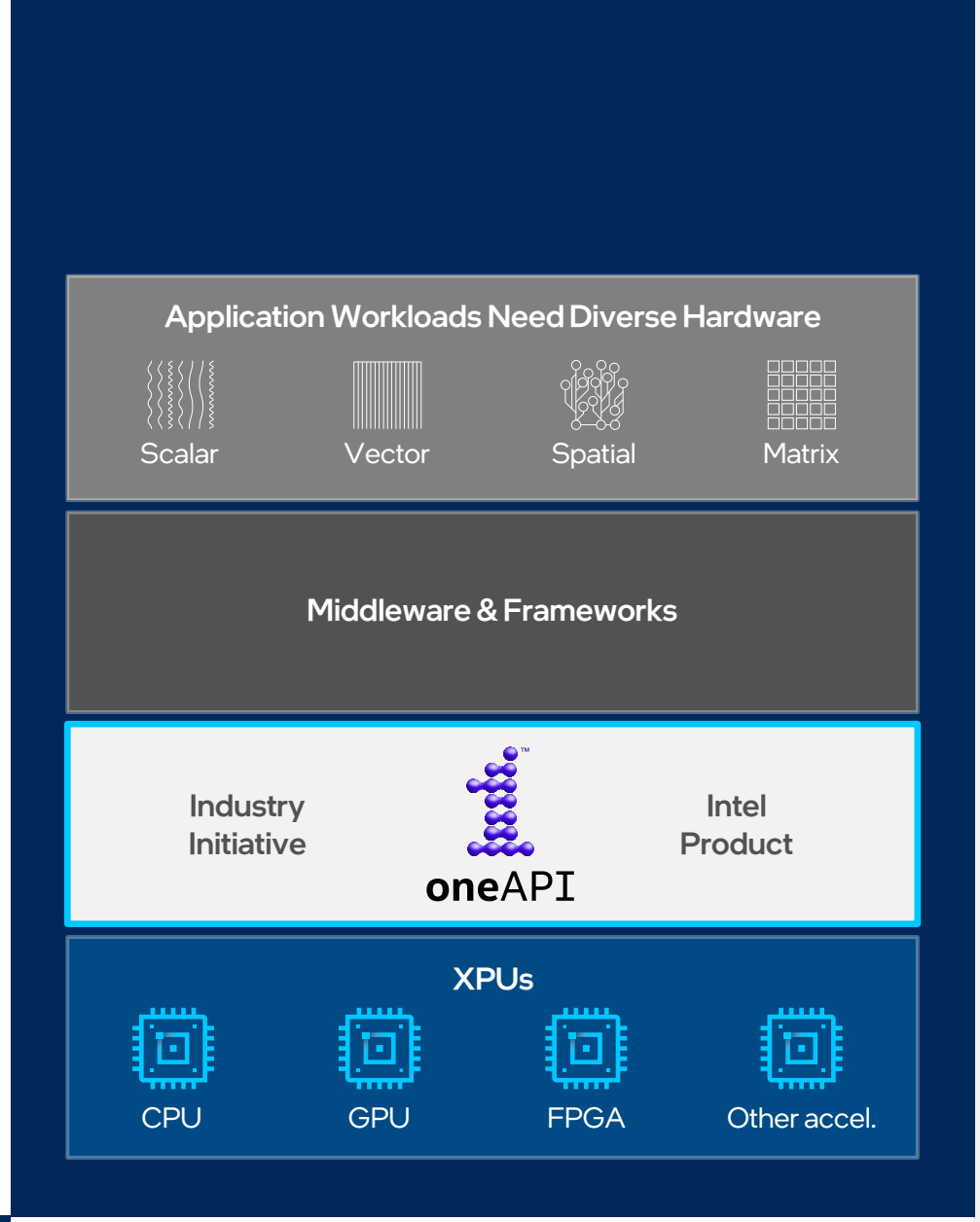
- Choose the best accelerated technology the software doesn't decide for you

Realize all the Hardware Value

- Performance across CPU, GPUs, FPGAs, and other accelerators

Develop & Deploy Software with Peace of Mind

- Open industry standards provide a safe, clear path to the future
- Compatible with existing languages and programming models including C++, Python, SYCL, OpenMP, Fortran, and MPI



Intel's oneAPI Ecosystem

Built on Intel's Rich Heritage of CPU Tools Expanded to XPUs

oneAPI

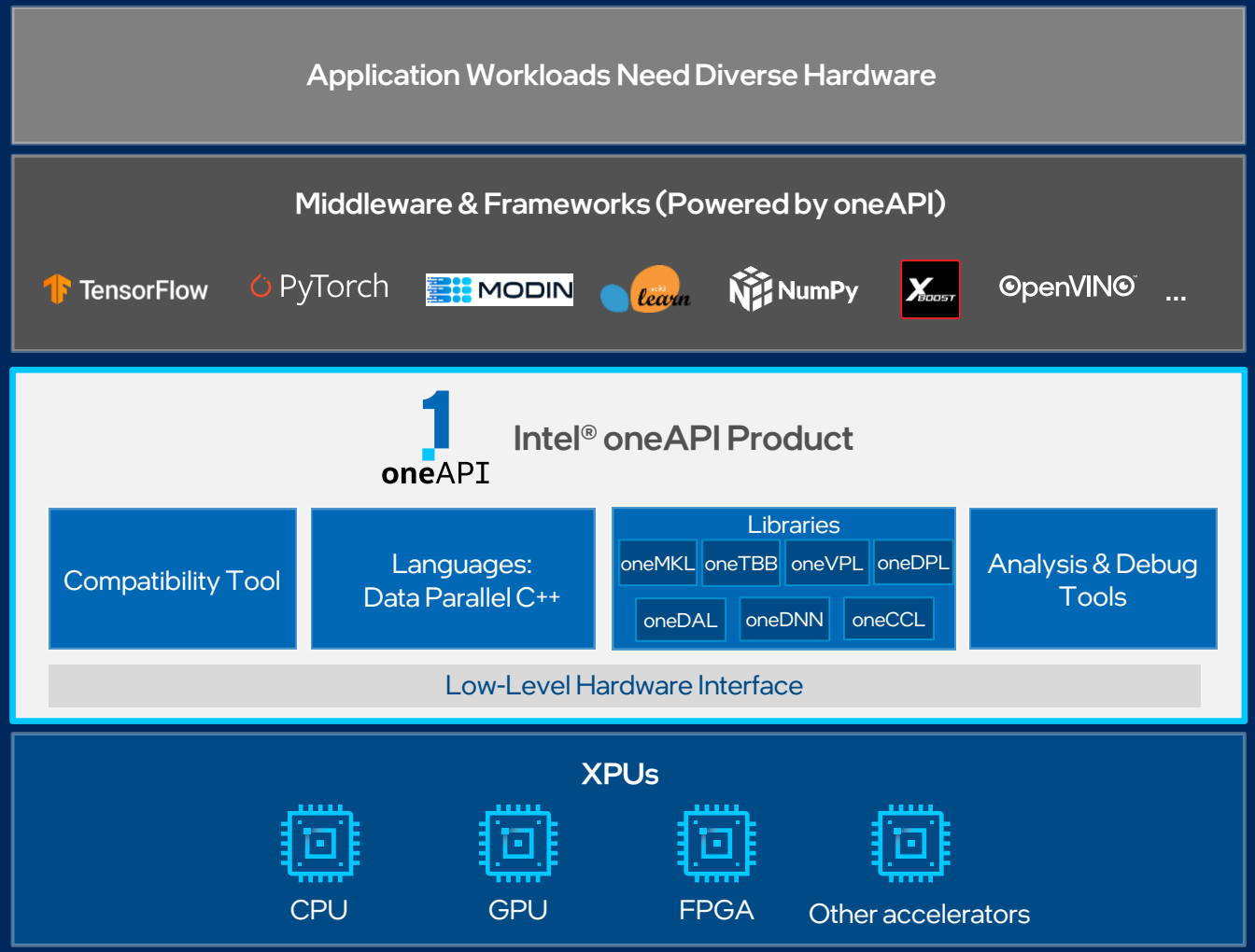
A cross-architecture language based on C++ and SYCL standards

Powerful libraries designed for acceleration of domain-specific functions

A complete set of advanced compilers, libraries, and porting, analysis and debugger tools

Powered by oneAPI

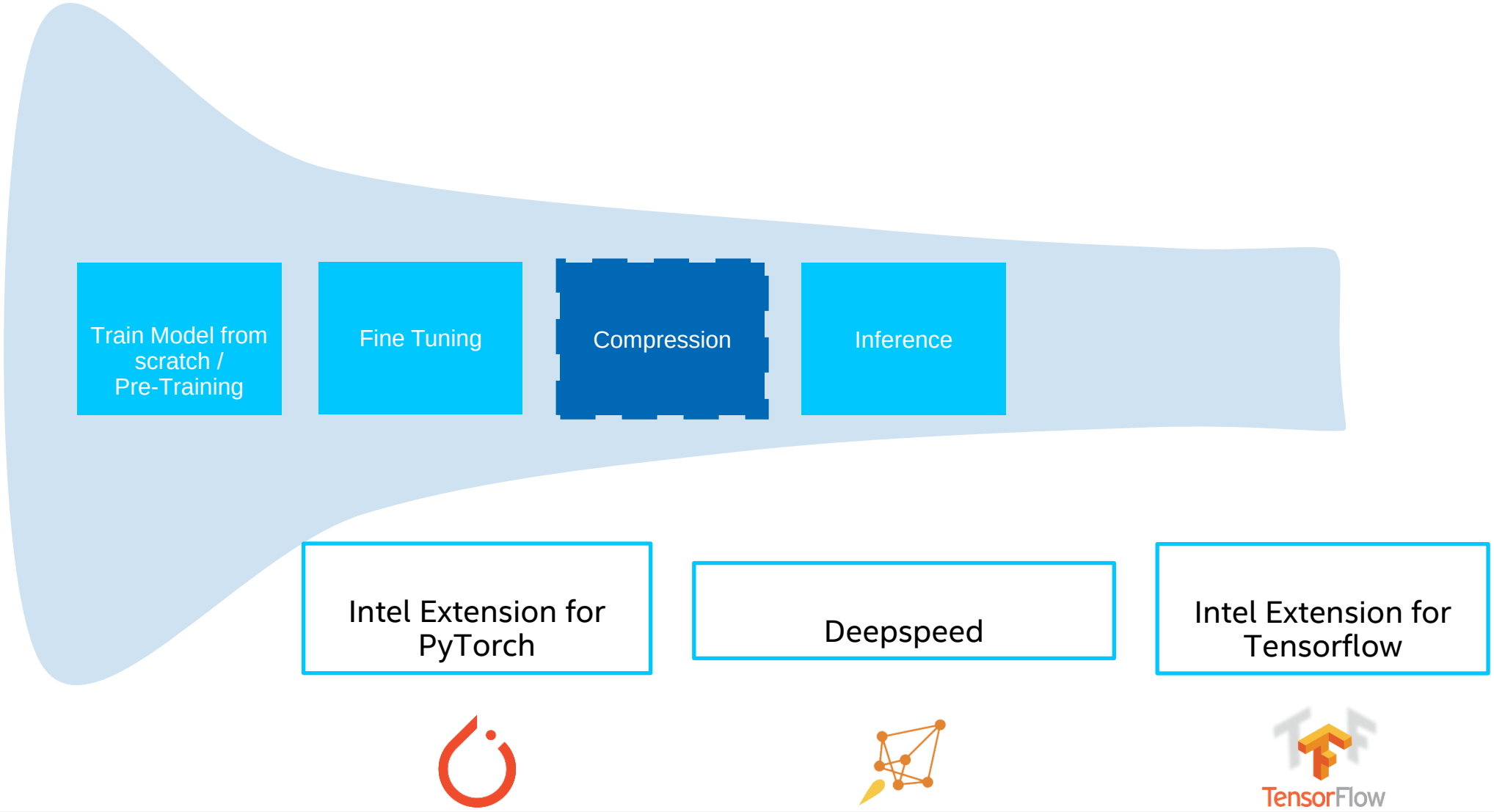
Frameworks and middleware that are built using one or more of the oneAPI industry specification elements, the DPC++ language, and libraries listed on oneapi.com.



Available Now

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.

Deep Learning Funnel Pipeline




intel

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.

Configurations

Deep Learning Training and Inference Performance using Intel® Optimization for PyTorch with 3rd Gen Intel® Xeon® Scalable Processors
ResNet50/ResNext101 (FP32/BF16): batch size 128/instance, 4 instances.
ResNet50/ResNext101 dataset (FP32/BF16): ImageNet Dataset
DLRM batch size (FP32/BF16): 2K/instance, 1 instance
DLRM dataset (FP32/BF16): Criteo Terabyte Dataset
DLRM batch size (INT8): 16/instance, 28 instances, dummy data.
Tested by Intel as of 6/2/2020.
Intel® Xeon® Platinum 8380H Processor, 4 socket, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0015.P96.2005070242 (ucode: 0x700001b), Ubuntu 20.04 LTS, kernel 5.4.0-29-generic
PyTorch: https://github.com/pytorch/pytorch.git
Intel Extension for PyTorch: https://github.com/intel/intel-extension-for-pytorch.git
gcc: 8.4.0,
Intel® oneAPI Deep Neural Network Library (oneDNN) version: v1.4
ResNet50: https://github.com/intel/optimized-models/tree/master/pytorch/ResNet50
ResNext101 32x4d: https://github.com/intel/optimized-models/tree/master/pytorch/ResNext101_32x4d
DLRM: https://github.com/intel/optimized-models/tree/master/pytorch/dlrm

Inference Throughput FP32 vs Int8 optimized by Intel® Optimization for Tensorflow and Intel® Low Precision Optimization Tool (part of the Intel® oneAPI AI Analytics Toolkit)
Tested by Intel as of : 10/26/2020: TensorFlow v2.2 (https://github.com/Intel-tensorflow/tensorflow/tree/v2.2.0); Compiler: GCC 7.2.1; DNNL(https://github.com/oneapi-src/oneDNN) v1.2.0 75d0b1a7f3586c212e37acebbb8acd221cee7216; Dataset: ImageNet/Coco/Dummy, refer to each model README; Precision: FP32 and Int8
Platform: Intel® Xeon® Platinum 8280 CPU; #Nodes: 1; #Sockets: 2; Cores/socket: 28; Threads/socket: 56; HT: On; Turbo: On; BIOS version: SE5C620.86B.02.01.0010.010620200716; System DDR Mem Config: 12 slots / 16GB / 2933; OS: CentOS Linux 7.8; Kernel: 4.4.240-1.el7.elrepo.x86_64

Stock scikit-learn vs Intel-optimized scikit-learn
Testing by Intel as of 10/23/2020. Intel® oneAPI Data Analytics Library 2021.1 (oneDAL), scikit-learn 0.23.1, Intel® Distribution for Python 3.8; Intel® Xeon® Platinum 8280LCPU @ 2.70GHz, 2Sockets, 28 cores per socket, 10M samples, 10 features, 100 clusters, 100 iterations, float32

XGBoost CPU vs GPU
Test configs: Tested by Intel as of 10/13/2020;
CPU: c5.18xlarge AWS Instance (2 x Intel® Xeon Platinum 8124M @ 18 cores, OS: Ubuntu 20.04.2 LTS, 193 GB RAM. GPU: p3.2xlarge AWS Instance (GPU: NVIDIA Tesla V100 16GB, 8 vCPUs), OS: Ubuntu 18.04.2 LTS, 61 GB RAM. SW: XGBoost 1.1:build from sources. compiler – G++ 7.4, nvcc 9.1. Intel® Data Analytics Acceleration Library (Intel® DAAL): 2019.4 version; Python env: Python 3.6, Numpy 1.16.4, Pandas 0.25, Scikit-lean 0.21.2.

XGBoost fit CPU acceleration
Test configs: Tested by Intel as of 10/13/2020; c5.24xlarge AWS Instance, CLX 8275 @ 3.0GHz, 2 sockets, 24 cores per socket, HT:on, DRAM (12 slots / 32GB / 2933 MHz); SW: XGBoost 0.81, 0.9, 1.0 and 1.1:build from sources. compiler – G++ 7.4, nvcc 9.1. Intel® DAAL: 2019.4 version; Python env: Python 3.6, Numpy 1.16.4, Pandas 0.25, Scikit-lean 0.21.2.

End-to-End Census Workload Performance
Tested by Intel as of 10/15/2020. 2x Intel® Xeon® Platinum 8280 @ 28cores, OS: Ubuntu 19.10.5.3.0-64-generic Mitigated, 384GB RAM. SW: Modin 0.8.1, scikit-learn 0.22.2, Pandas 1.0.1, Python 3.8.5, Daal4Py 2020.2 Census Data, (21721922, 45). Dataset is from IPUMS USA, University of Minnesota, www.ipums.org . Version 10.0.

Tiger Lake + Intel® Distribution of OpenVINO™ toolkit vs Coffee Lake CPU

System Board	Intel prototype, TGL U DDR4 SODIMM RVP	ASUSTeK COMPUTER INC. / PRIME Z370-A
CPU	11th Gen Intel® Core™ -5-1145G7E @ 2.6 GHz.	8th Gen Intel® Core™ i5-8500T @ 3.0 GHz.
Sockets / Physical cores	1 / 4	1 / 6
HyperThreading / Turbo Setting	Enabled / On	Na / On
Memory	2 x 8198 MB 3200 MT/s DDR4	2 x 16384 MB 2667 MT/s DDR4
OS	Ubuntu* 18.04 LTS	Ubuntu* 18.04 LTS
Kernel	5.8.0-050800-generic	5.3.0-24-generic
Software	Intel® Distribution of OpenVINO™ toolkit 2021.1.075	Intel® Distribution of OpenVINO™ toolkit 2021.1.075
BIOS	Intel TGLIFUI1.R00.3243.A04.2006302148	AMI, version 2401
BIOS release date	Release Date: 06/30/2020	7/12/2019
BIOS Setting	Load default settings	Load default settings, set XMP to 2667
Test Date	9/9/2020	9/9/2020
Precision and Batch Size	CPU: INT8, GPU: FP16-INT8, batch size: 1	CPU: INT8, GPU: FP16-INT8, batch size: 1
Number of Inference Requests	4	6
Number of Execution Streams	4	6
Power (TDP Link)	28 W	35W