intel ai summit
英特爾AI科技論壇

Bringing AI Everywhere

運用新一代Intel® Xeon® CPU 技術於雲端世界強化你的AI算力

Luke Tang | 技術專案經理 (Intel)

27th, Mar 2024

Notices and Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:

Intel® Advanced Matrix Extensions (Intel® AMX)

DL Accelerator Performance Built Into Every Core

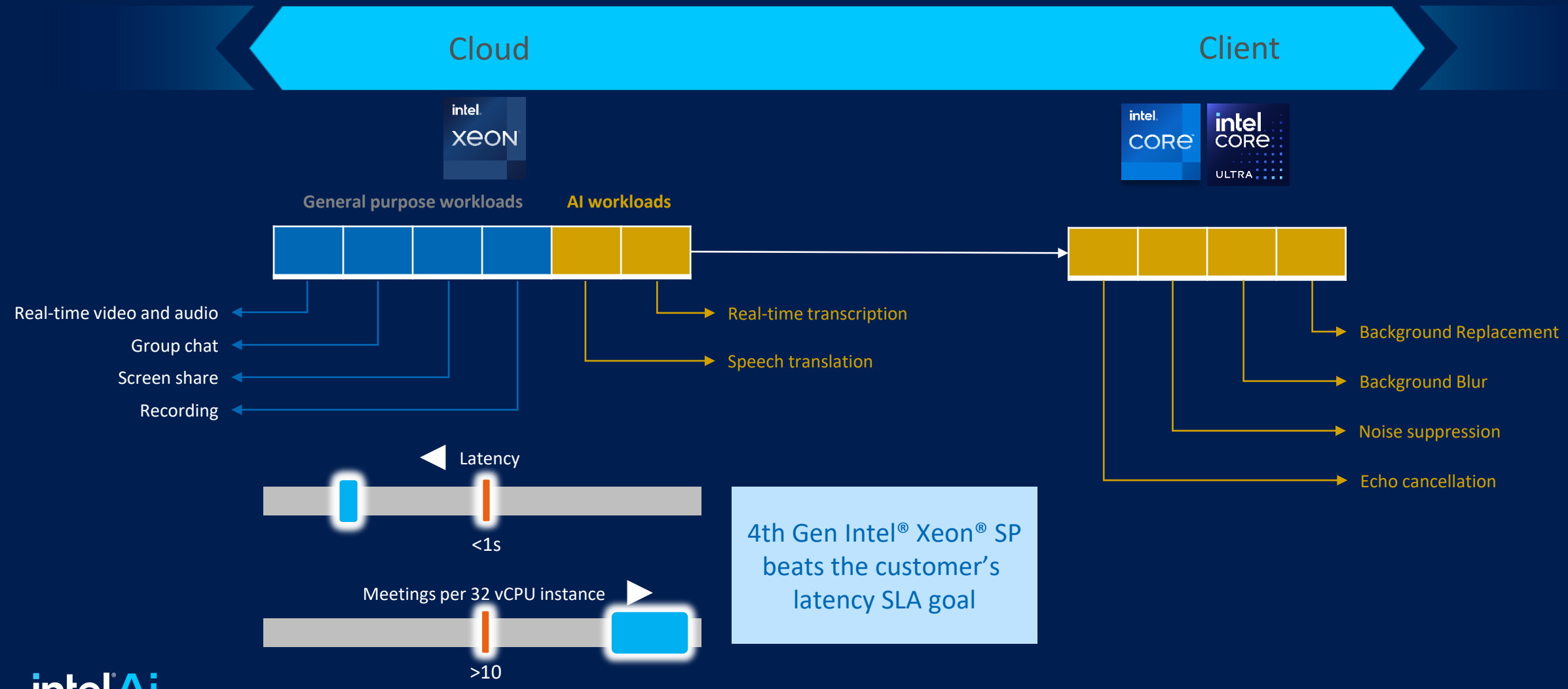
```
ubuntu@ip-10-0-10-191:~$ lscpu
Architecture:            x86_64
  CPU op-mode(s):        32-bit, 64-bit
  Address sizes:         46 bits physical, 48 bits virtual
  Byte Order:            Little Endian
CPU(s):                  2
  On-line CPU(s) list:   0,1
Vendor ID:               GenuineIntel
  Model name:            Intel(R) Xeon(R) Platinum 8488C
    CPU family:          6
    Model:               143
    Thread(s) per core:  2
    Core(s) per socket:  1
    Socket(s):           1
    Stepping:            8
    BogoMIPS:            4800.00
    Flags:               fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr sse sse2 ss ht syscall nx pdpe1gb rdtscp lm constant_tsc arch_perfmon rep_
                          nonstop_tsc cpuid aperfmperf tsc_known_freq pni pclmulqdq monitor ssse3 fma cx16 pdcm pcid sse4_1 sse4_2 x2apic movbe popcnt tsc_deadline_timer aes xsave avx f16c
                         lahf_lm abm 3dnowprefetch invpcid_single ssbd ibrs ibpb stibp ibrs_enhanced fsgsbase tsc_adjust bmi1 avx2 smep bmi2 erms invpcid avx512f avx512dq rdseed adx smap a
                         t clwb avx512cd sha_ni avx512bw avx512vl xsaveopt xsavec xgetbv1 xsaves avx_vnni avx512_bf16 wbnoinvd ida arat avx512vbmi umip pku ospke waitpkg avx512_vbmi2 gfni
                         512_vnni avx512_bitalg tme avx512_vpopcntdq rdpid cldemote movdiri movdir64b md_clear serialize amx_bf16 avx512_fp16 amx_tile amx_int8 flush_l1d arch_capabilities
Virtualization features:
  Hypervisor vendor:     KVM
  Virtualization type:   full
Caches (sum of all):
  L1d:                   48 KiB (1 instance)
  L1i:                   32 KiB (1 instance)
  L2:                    2 MiB (1 instance)
  L3:                    105 MiB (1 instance)
NUMA:
  NUMA node(s):          1
  NUMA node0 CPU(s):     0,1
```

Store bigger chunks of **data**

larger matrices in a single operation

intel Ai summit

Position: AI Value Prop on Intel Xeon CPU



intel AI summit

AI on Xeon with Hyperscalers

Joint Collaboration

intel Ai summit

AWS

GCR Compute GTM

Miley, Shih

intel Ai summit

Demystifying generative AI on AWS

Miley Shih
Compute Go-To-Market Specialist
AWS

Trends in AI/ML innovation

Growth in LLMs

Rapid growth of large language models (LLM) based on transformer architectures

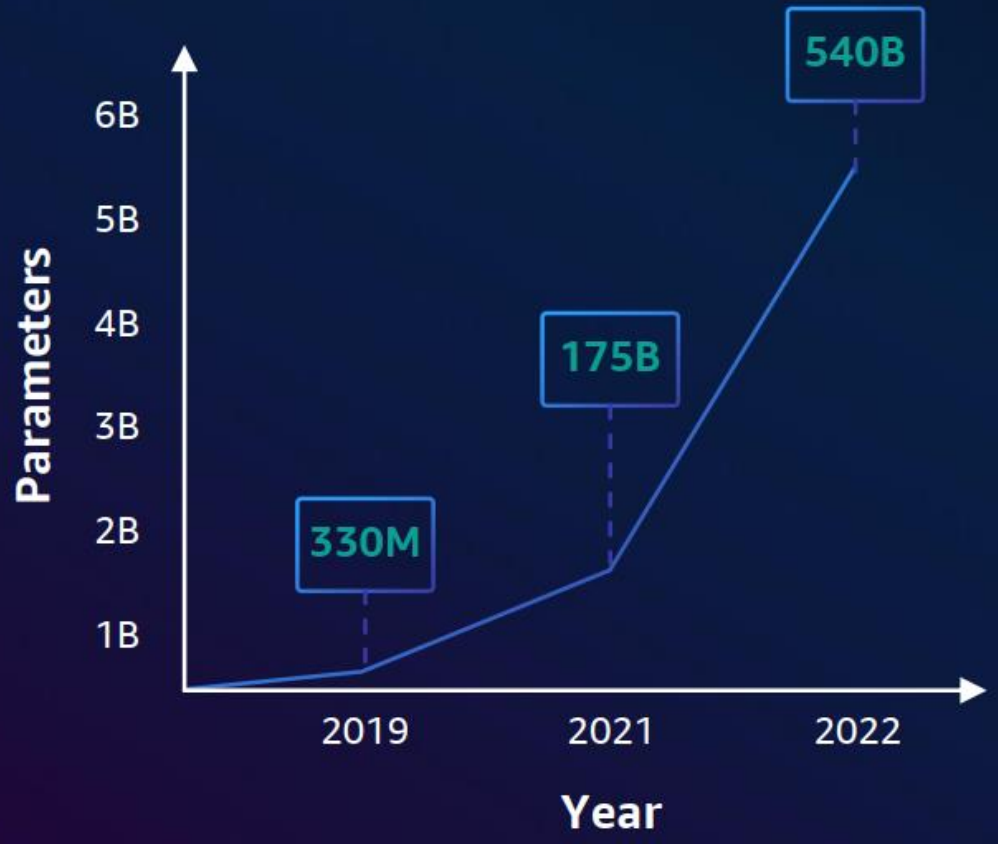
Faster time to solution with pretrained FMs

Data scientists no longer need to train models from the ground up

Open-source momentum

Increased momentum from industry offering open-source, pretrained models

Rise of foundation models – What has changed?



2019–2022

1,600x

increase in size of model as measured by number of parameters

Broadest and deepest compute platform choice

CATEGORIES

General purpose

Burstable

Compute intensive

Memory intensive

Storage (High I/O)

Dense storage

GPU compute

Graphics intensive

+

CAPABILITIES

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High memory footprint
(up to 12 TiB)

Instance storage
(HDD, SSD, NVMe)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 100 Gbps)

Bare Metal

Size
(Nano to 32xlarge)

+

OPTIONS

Amazon EBS

Amazon Elastic Inference

=

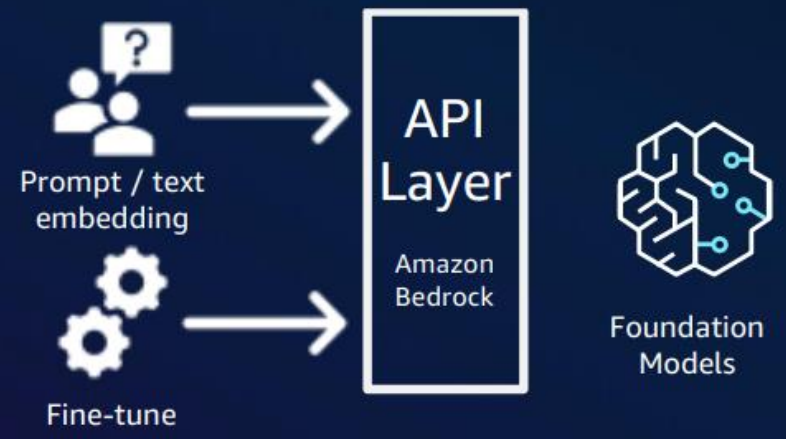
MORE THAN

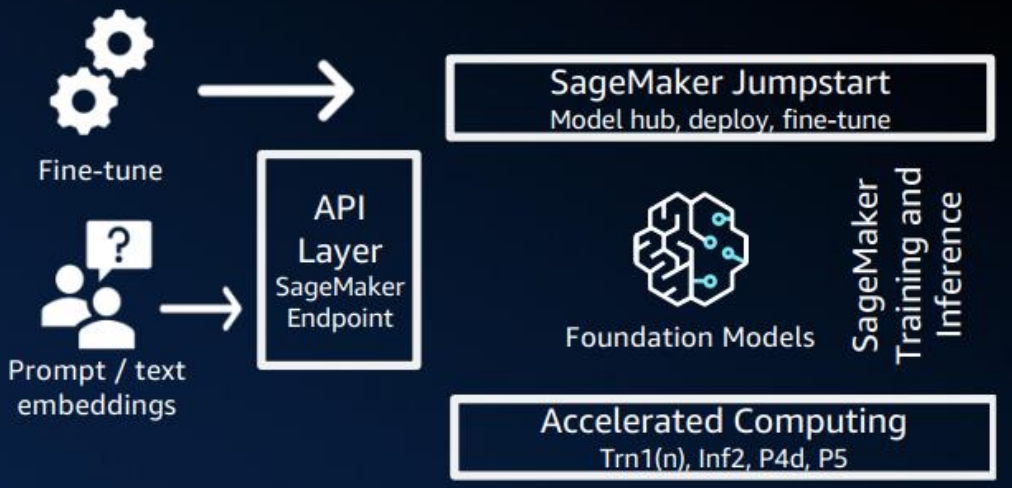
750+

INSTANCE TYPES

for virtually every workload and business need

How do I access foundation models?





Amazon Bedrock

- The easiest way to build and scale generative AI applications with FMs
- Access directly or fine-tune foundation model using API
- Serverless

Amazon SageMaker JumpStart

- ML hub with FMs, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks
- Deploy FM as SageMaker endpoint (hosting)
- Fine-tuning leverages SageMaker training jobs
- Choose SageMaker managed accelerated computing instance

AWS AI/ML stack

AI SERVICES

SPECIALIZED

BUSINESS PROCESSES
- Amazon Personalize
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lookout for Metrics

SEARCH
- Amazon Kendra

CONVERSATION
- Amazon Lex
- Amazon Transcribe Call Analytics
- Contact Lens
- Voice ID

CODE + DEVOPS
- Amazon CodeGuru
- Amazon CodeWhisperer
- Amazon DevOps Guru

INDUSTRIAL
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

HEALTH
- Amazon HealthLake
- Amazon Comprehend Medical
- Amazon Transcribe Medical
- Amazon HealthOmics

CORE

TEXT
- Amazon Translate
- Amazon Comprehend

SPEECH
- Amazon Polly
- Amazon Transcribe

VISION
- Amazon Textract
- Amazon Rekognition
- AWS Panorama

AMAZON SAGEMAKER

CANVAS
No-code ML for business analysts

STUDIO LAB
Learn ML

GROUND TRUTH
Label data

STUDIO IDE

Prepare data | Geospatial ML | Build with notebooks | Train models | Deploy in production

Store features | Tune parameters | Manage and monitor

CI/CD | GOVERNANCE | RESPONSIBLE ML

EDGE MANAGER
Manage edge devices

ML FRAMEWORKS & INFRASTRUCTURE

PyTorch, Apache MXNet, TensorFlow | Amazon EC2 | CPUs | GPUs | AWS Inferentia | AWS Trainium | Habana Gaudi | FPGA

Why customers choose to run ML workloads on Kubernetes

Highly scalable

Improved resource utilization

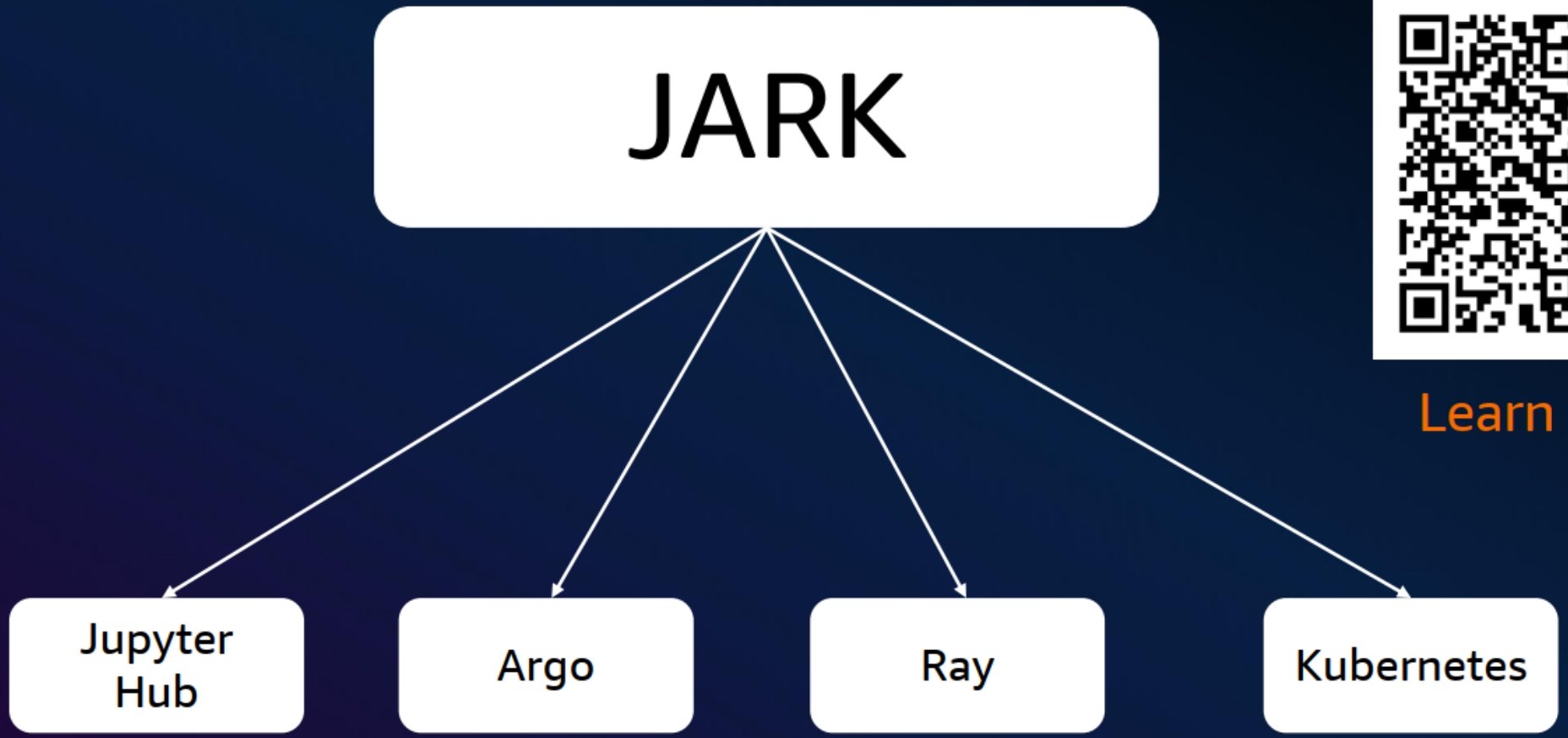
Org standardization

Open source community

Application level challenges for ML workloads

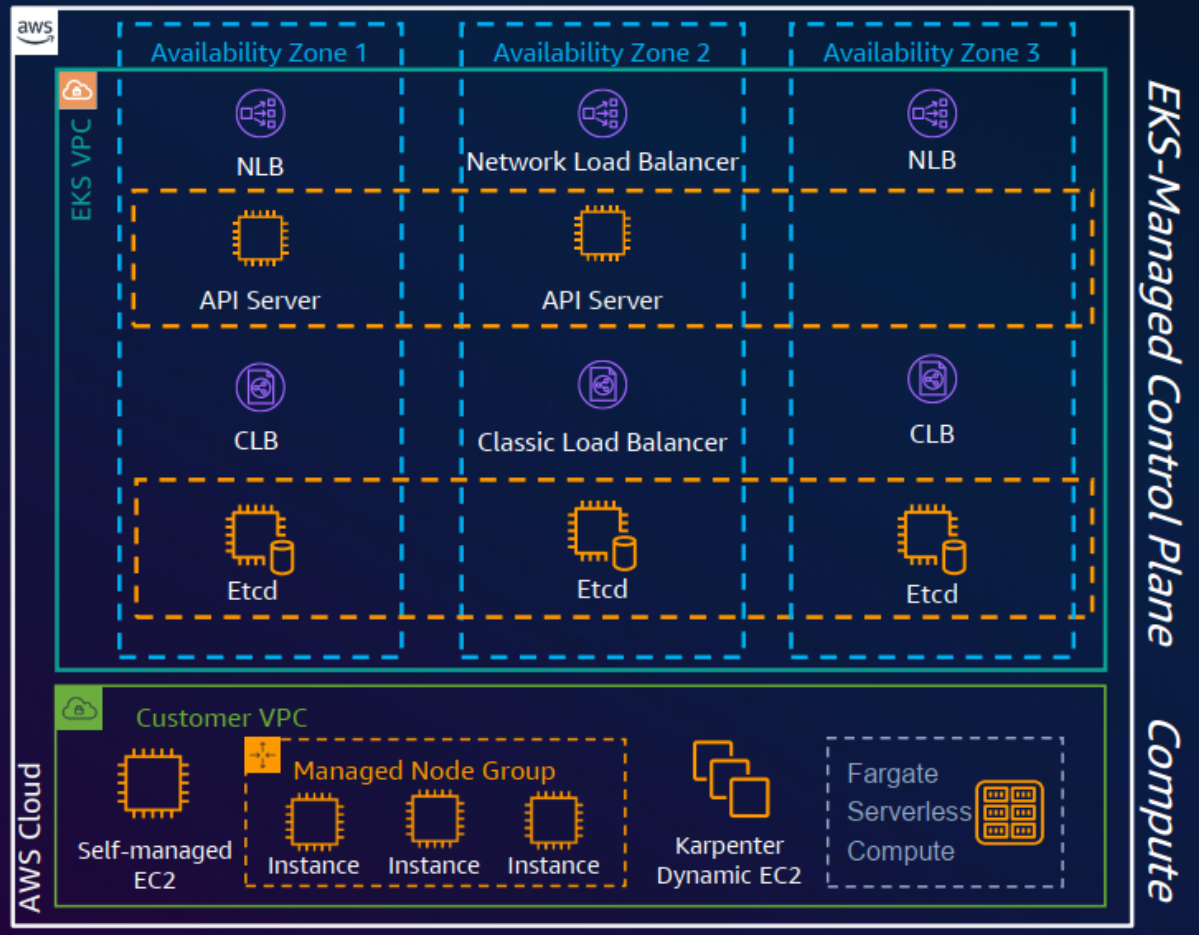
- No K8s built-in ML APIs
- Data scientists are not K8s experts
- Following MLOps best practices

JARK stack on Amazon EKS – Opinionated stack for end-end ML orchestration



Learn more

What does an Amazon EKS cluster look like?



AWS managed control plane

- **Highly available, single-tenant Kubernetes API server and etcd database**

Cluster compute

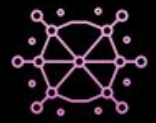
- **Self-managed EC2 instances**
  Run in your account, customer managed, maximum flexibility/configurability
- **EKS managed nodes**
  Run in your account, AWS-managed provisioning and instance lifecycle
- **Karpenter**
  Customer-managed, cutting edge, open-source node provisioning and cluster autoscaling
- **AWS Fargate**
  Serverless, right-sized compute; AWS-managed OS, container runtime; storage/monitoring plugins; granular, pod-based billing

Innovating with Intel

17 years of collaboration with AWS

COLLABORATION

Deep engineering collaboration across the AWS portfolio



INTEGRATION

Over 400 Amazon EC2 instances are powered by Intel processors

FASTEST

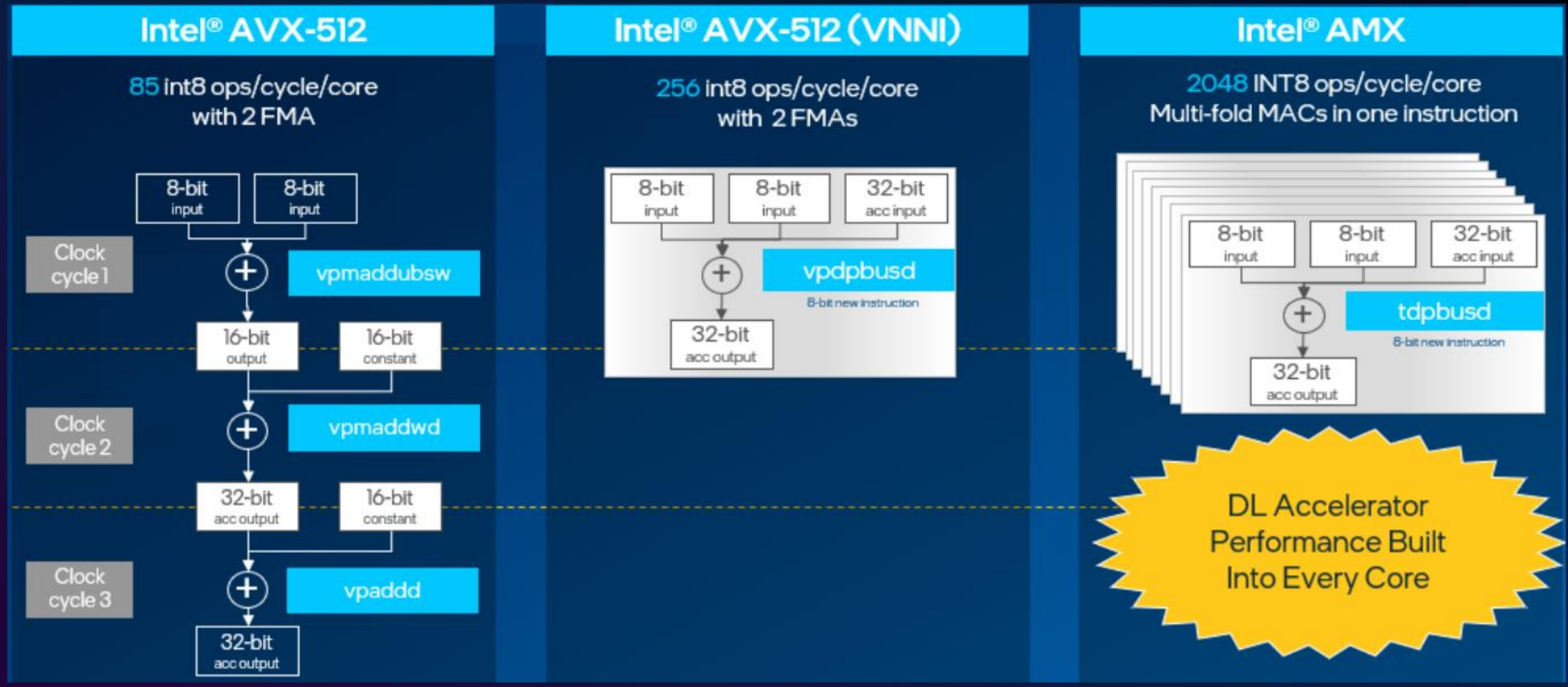
Fastest processor in the cloud and widest selection of Sapphire Rapids instances

RECENT INTEL-BASED INSTANCES

I4i	HPC6id	M7i	M7i-Flex	C7i	R7i	R7iz
STORAGE-OPTIMIZED	HPC-OPTIMIZED	GENERAL PURPOSE	COST-OPTIMIZED GENERAL PURPOSE	COMPUTE-OPTIMIZED	MEMORY-OPTIMIZED	HIGH-PERFORMANCE MEMORY OPTIMIZED

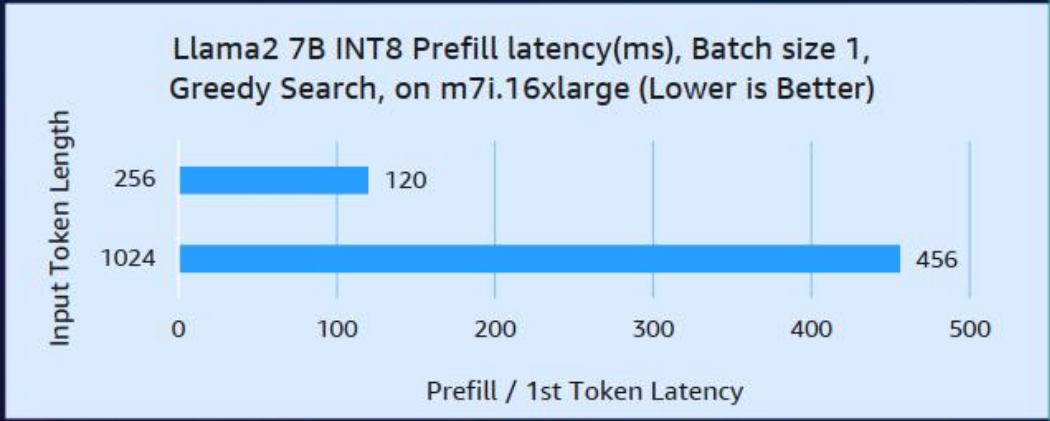
Vectorization on m7i / m7i-flex Instances

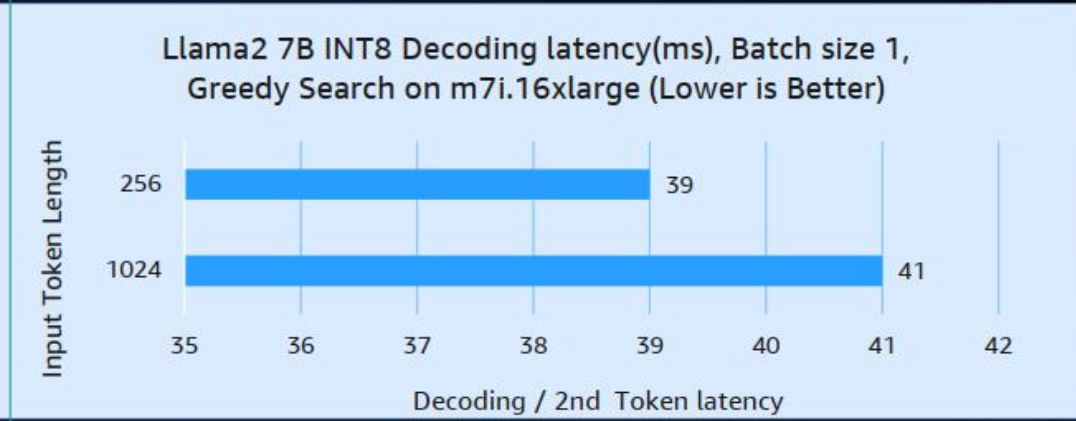
Intel® Advanced Matrix Extensions (Intel® AMX)

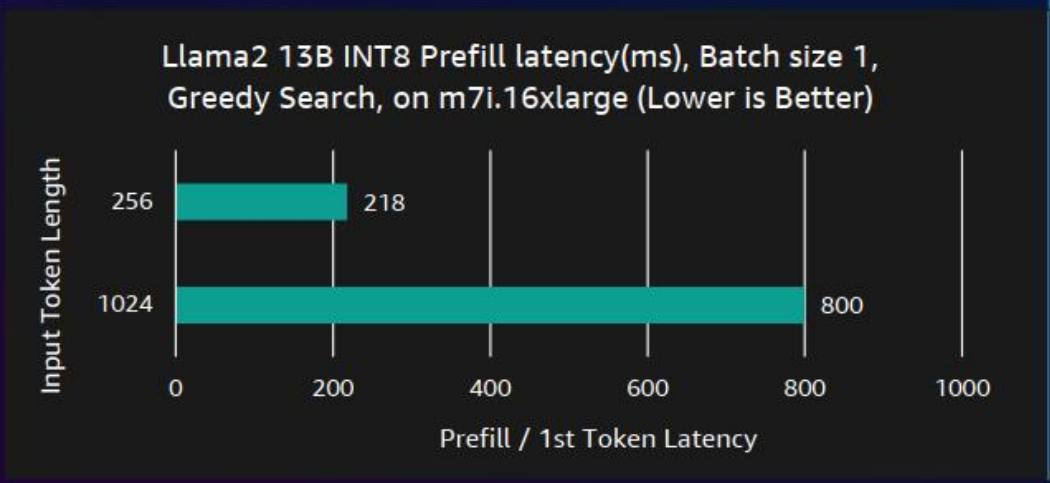


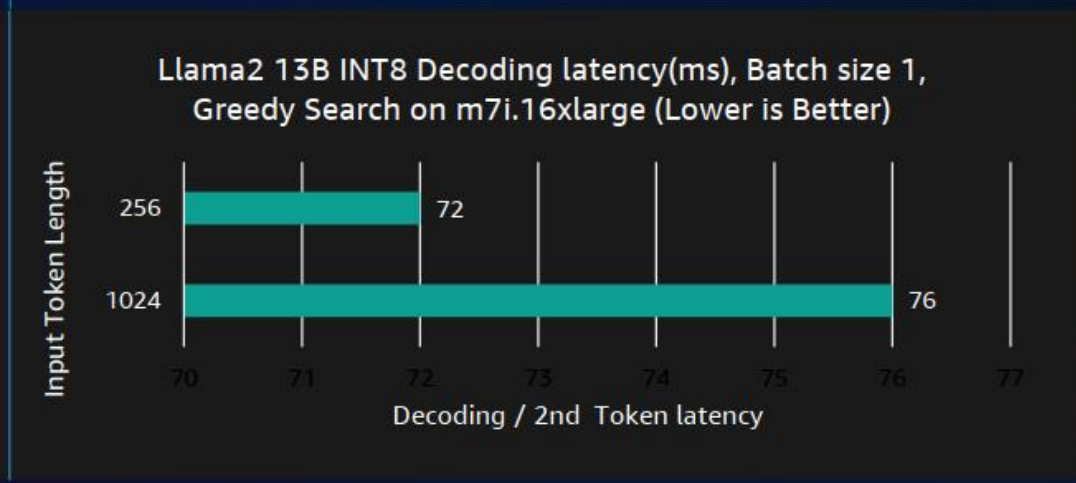
Generative AI and LLM's on M7i with Intel® AMX (Advanced Matrix Extensions)

- Large Language models (LLMs) are trained on >1T tokens (words/sub-words) with billions to a few trillions of parameters.
- Recently, medium-sized models (<13B) showed that they can match the largest models in terms of accuracy in specific use cases
- This data* shows that M7i/C7i/R7i can deliver < 50ms latency for sub-10B parameter models and <100ms latency for sub-20B parameter models









Intel® Neural Chat #1 in 7B Open LLM Leaderboard!

PyTorch + IPEX

Intel® Neural Compressor

Intel® Transformers Extension

*NOTE: GenAI and LLM's are a fast-evolving domain, software optimizations to enable Intel® AMX to handle even larger models are ongoing

Thank You

Google Cloud

Solution Architect

Kimi Lo

intel Ai summit

Google Cloud

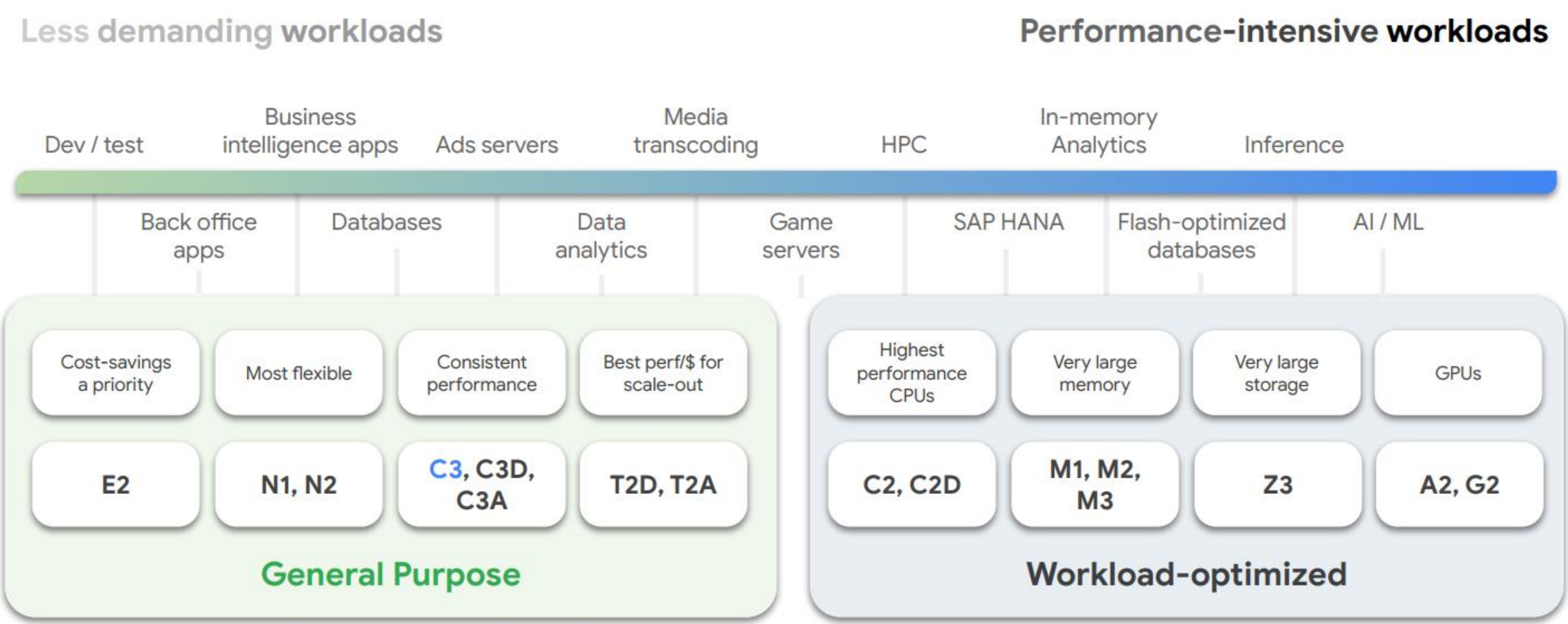
厚實你的雲端 AI 算力 Intel + GCP

Intel AI Summit

GCP Kimi Lo

InfraMod Solution Architect

Our C3 machine series for general-purpose workloads



Target workloads

C3 delivers **highly consistent** and **balanced performance**. This makes it the default choice for general-purpose workloads that cannot tolerate performance variability.

Web & app servers

Databases & caches

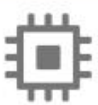
Game servers

Media streaming

Ad servers

Data analytics

Network appliances



ML inference

C3 Advanced Maintenance Experience

Maintenance controls for customer's most sensitive workloads

Initial Uptime

- **Deploy** and get started on your workloads **with** the **confidence**
- We **guarantee** that your workload will run **uninterrupted** from **planned maintenance** monthly

Frequency

- **Consistency** is critical, and providing **reliable maintenance schedules** is key
- We **guarantee** that your workload will **remain uninterrupted** from **planned maintenance** monthly

Notifications

- Stay **informed** of **upcoming** & **ongoing maintenance** of your workloads
- Receive **maintenance notifications** on your workloads up to 7 days in advance

Control

- **Perform** maintenance when it **best fits** your company's **schedule**
- **Simulate** maintenance to **understand** & **prepare** your workloads
- **Integrate** gcloud APIs with your preferred **automation** for **scalable management**

Leading-edge performance

Compute & Memory

- **First cloud** with Intel Sapphire Rapids (up to 176 vCPU)
- DDR5 memory **50% faster** than DDR4 (up to 1.5 TB)
- **Three** memory configs (2, 4, 8GB/vCPU)
- New **AMX accelerator** for up to 12x perf vs. AVX-512

Networking

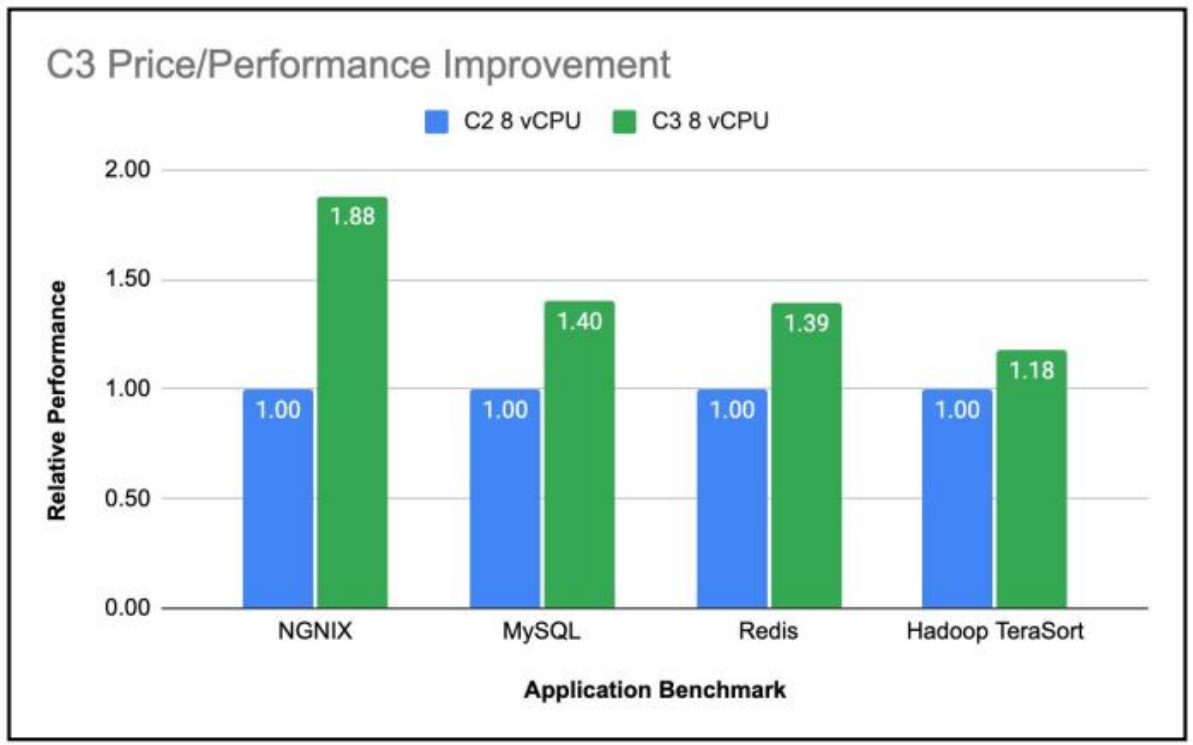
- Google **IPU with network offloads**
- **Dedicated network processing** improves VM consistency & minimizes jitter
- Enables **up to 200 Gbps** (2x C2, N2)
- **3x higher PPS vs. Gen 2** with lower **VM-VM latency** and

Storage

- **Hyperdisk Extreme** with up to 350k IOPS (10x vs. C2)
- **Hyperdisk Balanced** and **Throughput** coming Q3'23
- **Local SSD** on c3-standard with **slice-of-hardware** 3/6/12TB shapes

Performance proof points

C3 vs. C2



C3 price/perf on avg. up to 28% better than AWS C6i

- Intel Sapphire Rapids vs. Ice Lake (2021)
- Web serving and Redis shine

C3 price/perf on avg. up to 35% better than C2 & N2-CLX

- Intel Sapphire Rapids vs. Cascade Lake (2019)
- Web serving, redis, databases, hadoop, game servers

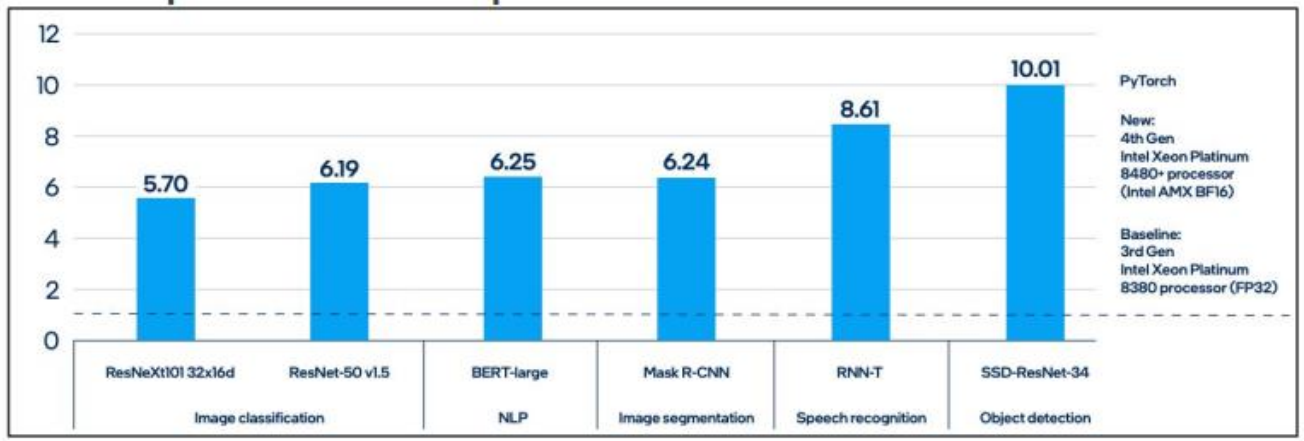
C3 price/perf up to 10% better than N2-ICX

- Intel Sapphire Rapids vs. Ice Lake (2021)
- Web serving, databases shine

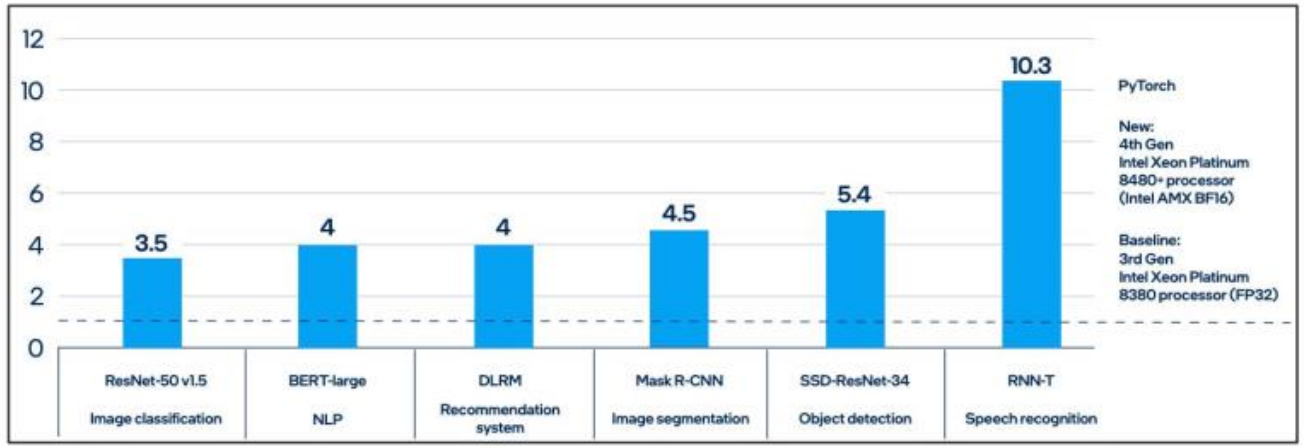
AMX Performance Proof Points

Intel Bare Metal - Source

Improves inference performance 5.7-10x vs. Intel Ice Lake



Improves training performance 3.5-10x vs. Intel Ice Lake

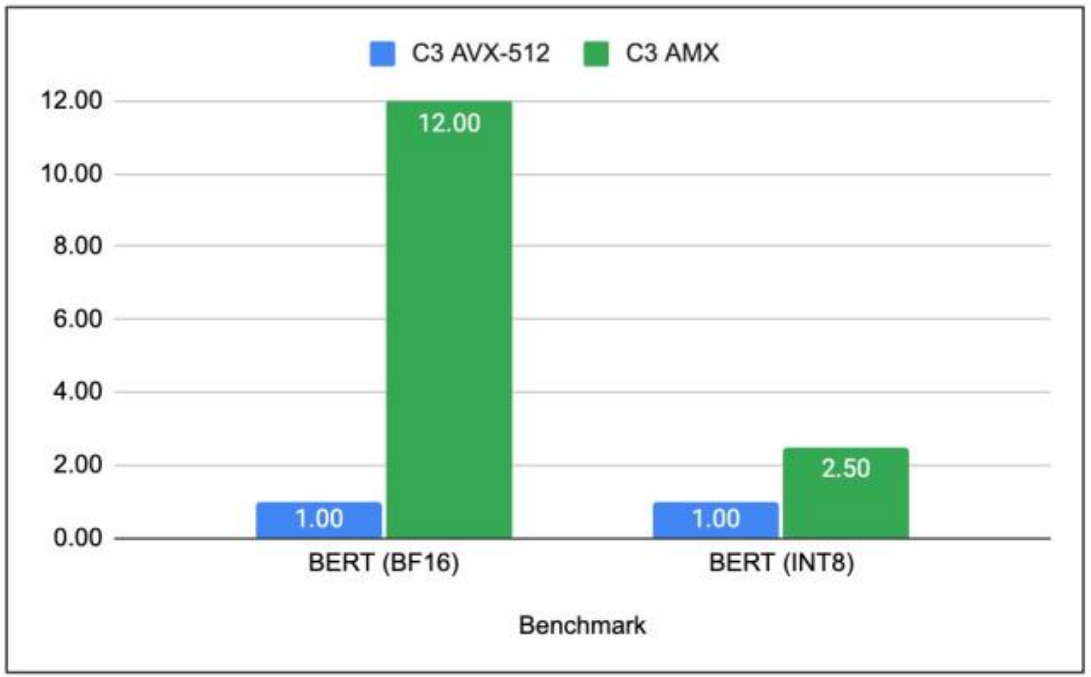


*Benchmarks based on Intel bare metal; GCE VM performance may vary. Source

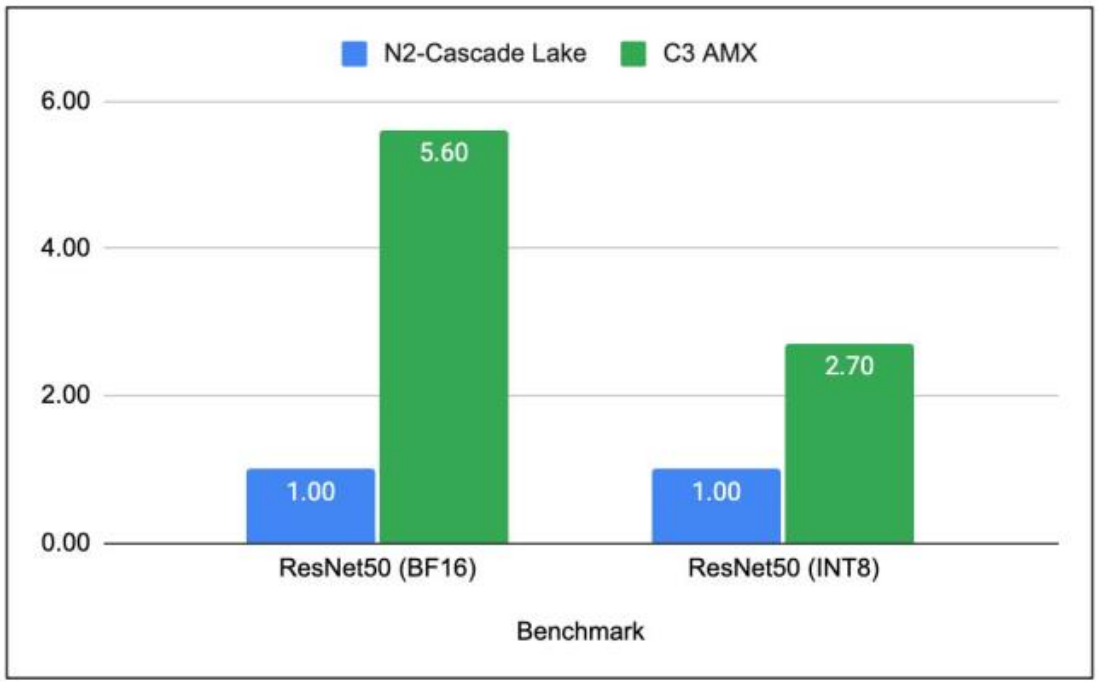
AMX Performance Proof Points

GCE VMs - Source

Improves BERT (NLP) performance up to 12x vs. AVX-512



Improves ResNet (Vision) performance up to 5.6x vs. N2



Intel Advanced Matrix Extensions (AMX)

Available on C3

- **Built-in accelerator for ML training and inference**
- **New to Intel Sapphire Rapids**
- **Target applications**
  - Natural Language Processing
  - Recommendation Systems
  - Image Recognition
  - Object Detection
  - Media/Video Analytics

paloalto NETWORKS

"At Palo Alto Networks, we develop and deploy deep learning models for inline threat detection in our customers' network traffic. Inference latency is critical for our AI workloads. By adopting C3 VMs with Intel Sapphire Rapids and the new AMX instruction set for AI, we are seeing 2x performance for some of our inline models, compared to the previous generation N2 Ice Lake VMs."

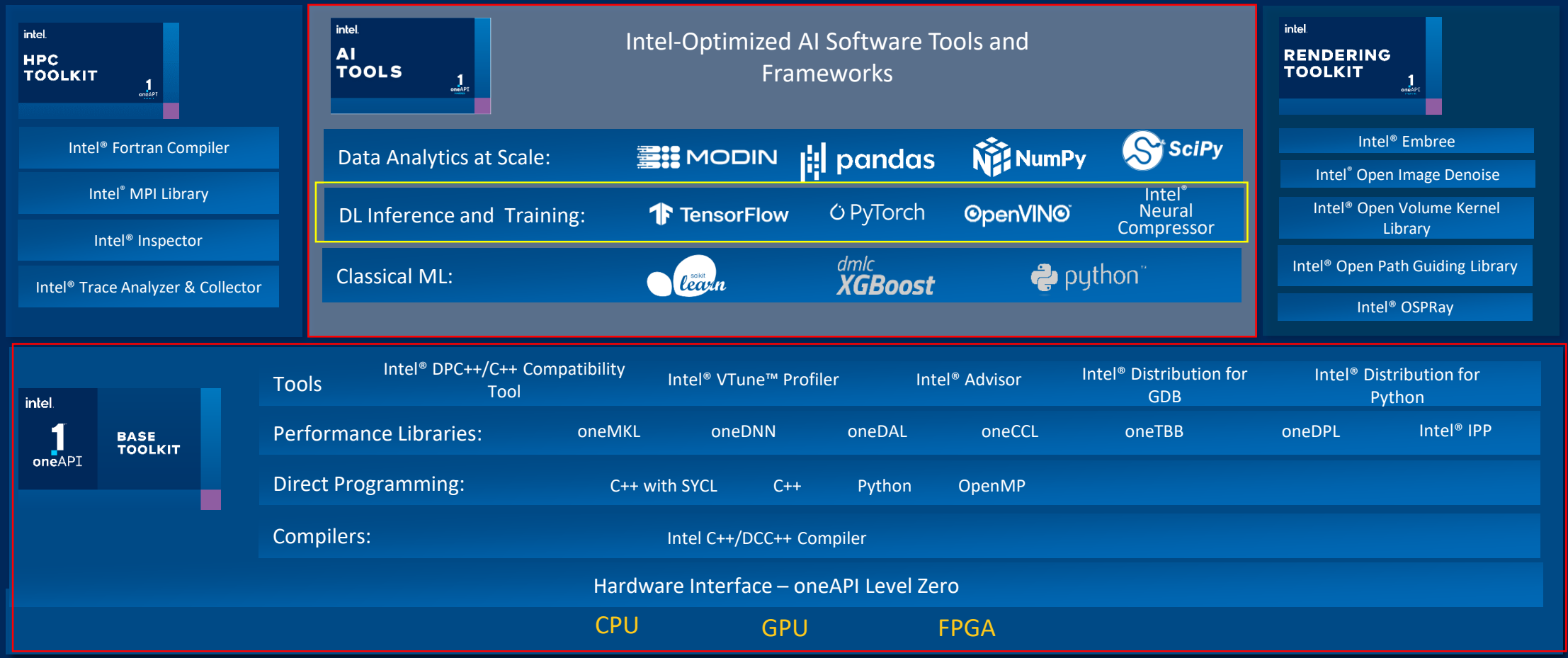
To be continued !!

Thank you.

Google Cloud

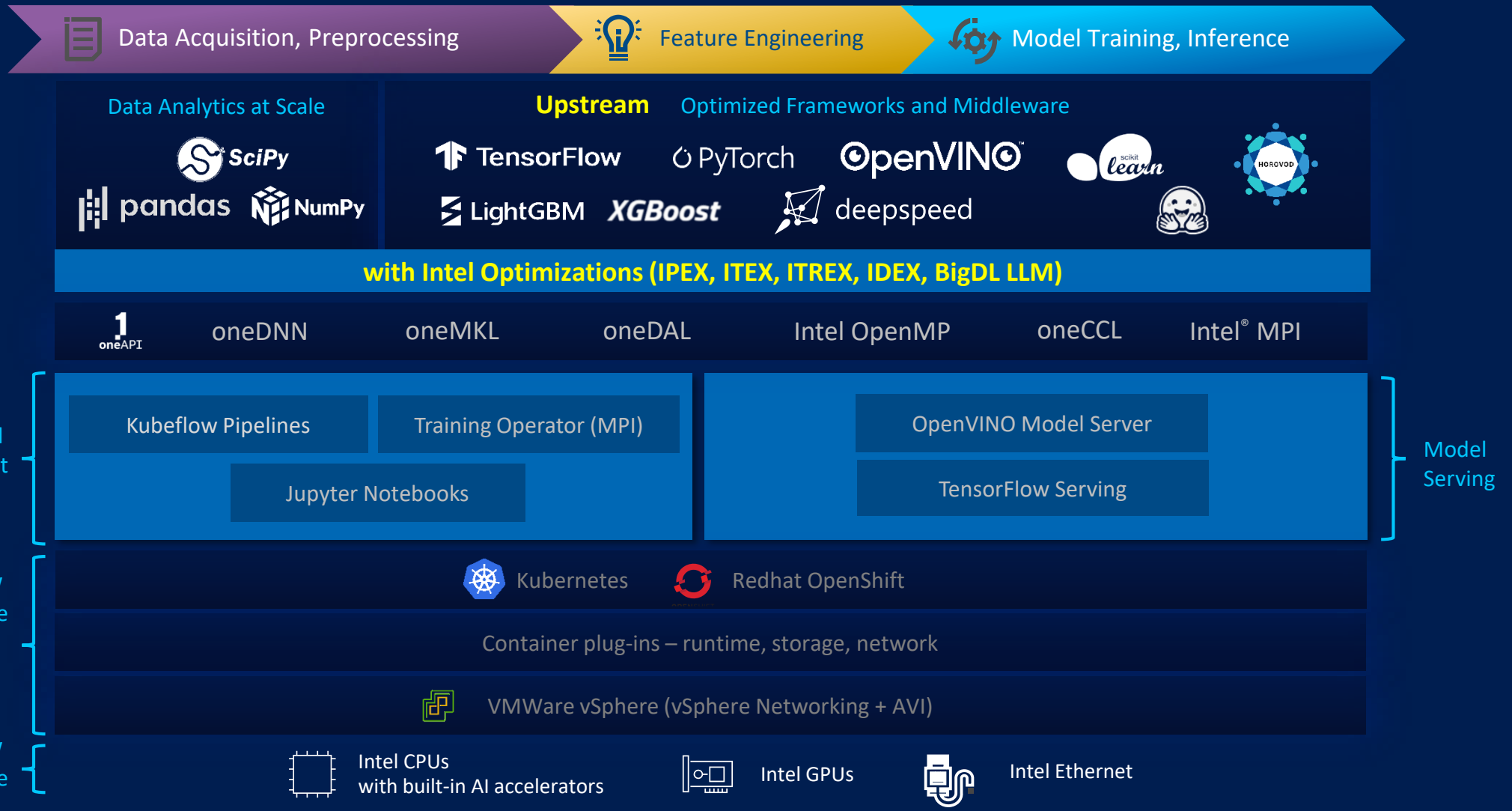
Intel Software Developer Tools

Flexible, Comprehensive, Open Software Stack – Powered by oneAPI



Download at intel.com/oneAPI or run tools on the Intel® Developer Cloud at cloud.intel.com

Intel® AI Software is Enterprise Ready



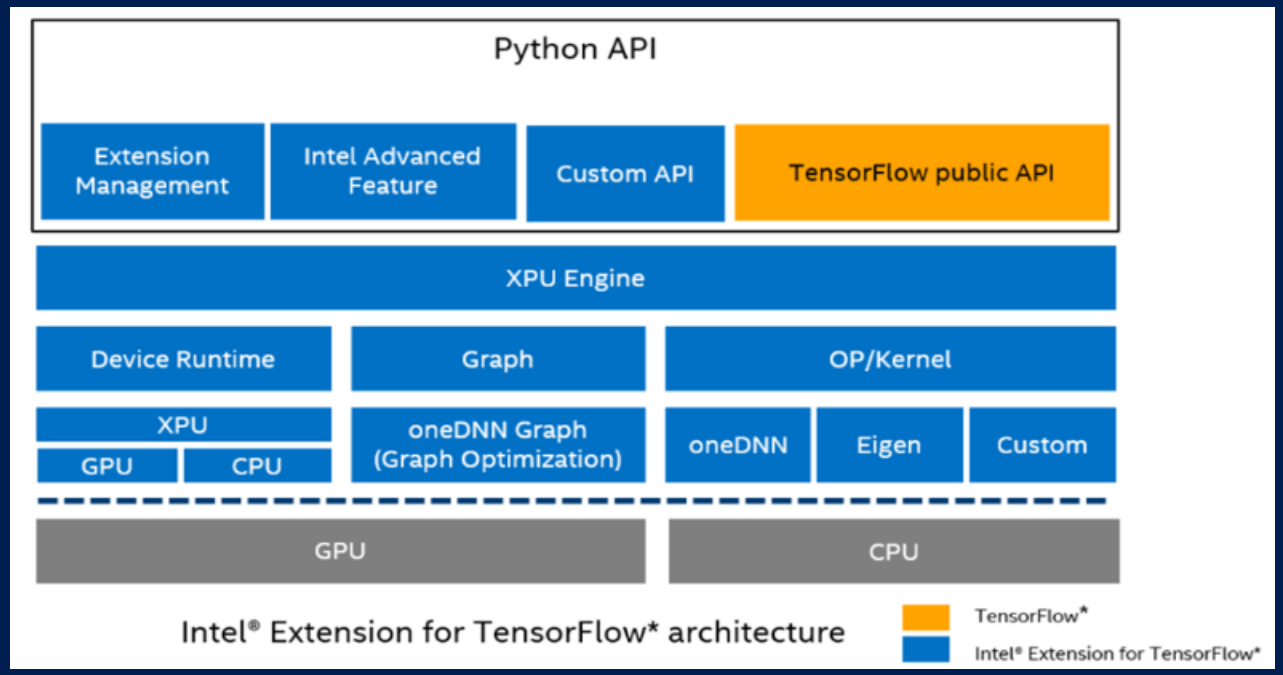
intel AI summit

Framework Level Optimization For AI/LLM Performance on Intel Hardware - IPEX

- Intel® Extension for PyTorch* (IPEX) : Extends PyTorch optimizations for an extra performance boost on Intel hardware
- Optimizations take advantage of Intel® AVX-512, VNNI and Intel® AMX on Intel CPUs as well as Intel Xe Matrix Extensions (XMX) AI engines on Intel discrete GPUs
- Installation: Easy for installing → pip install/docker image deployment
- Utilization: Adoption easily with only few code change needed

Source: https://intel.github.io/intel-extension-for-pytorch/#introduction

Framework Level Optimization For AI/LLM Performance on Intel Hardware – ITEX



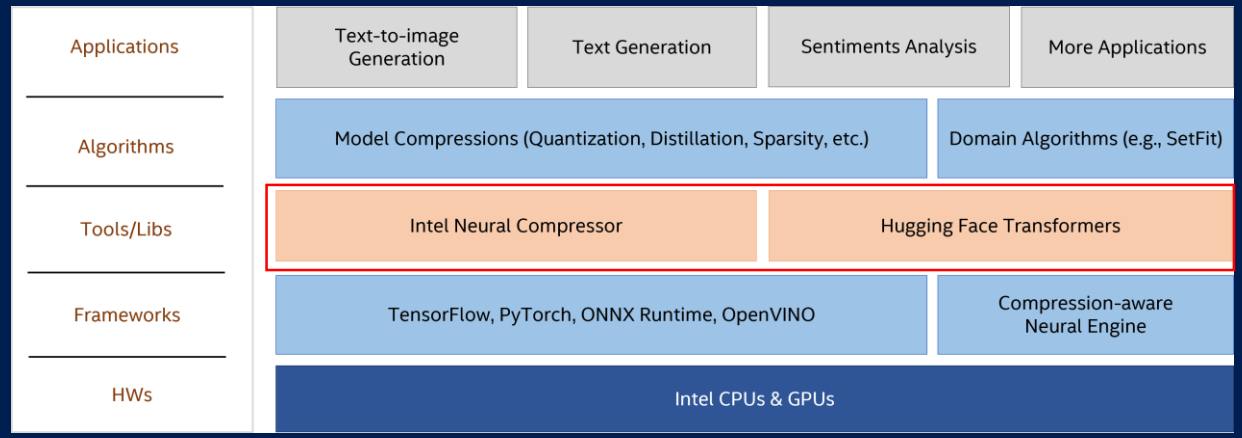
- Intel Extension for TensorFlow (ITEX): Provides users to flexibly plug an XPU into TensorFlow showing the computing power inside Intel's hardware
- Up-streams several optimizations into open source TensorFlow
- With NO CODE change when deploying ITEX to boost AI WL performance on Intel CPU & GPU

Source: https://www.intel.com/content/www/us/en/developer/articles/technical/introduction-to-intel-extension-for-tensorflow.html

Framework Level Optimization For AI/LLM Performance on Intel Hardware - ITREX

- Intel Extension for Transformer (ITREX): Designed to accelerate GenAI/LLM everywhere with the optimal performance of Transformer-based models on various Intel platforms, including

  - Intel CPU
  - Intel GPU
  - Intel Gaudi2



Validated Hardware

Hardware	Fine-Tuning		Inference	
	Full	PEFT	8-bit	4-bit
Intel Gaudi2	✓	✓	WIP (FP8)	-
Intel Xeon Scalable Processors	✓	✓	✓ (INT8, FP8)	✓ (INT4, FP4, NF4)
Intel Xeon CPU Max Series	✓	✓	✓ (INT8, FP8)	✓ (INT4, FP4, NF4)
Intel Data Center GPU Max Series	WIP	WIP	WIP (INT8)	✓ (INT4)
Intel Arc A-Series	-	-	WIP (INT8)	✓ (INT4)
Intel Core Processors	-	✓	✓ (INT8, FP8)	✓ (INT4, FP4, NF4)

Validated Software

Software	Fine-Tuning		Inference	
	Full	PEFT	8-bit	4-bit
PyTorch	2.0.1+cpu, 2.0.1a0 (gpu)	2.0.1+cpu, 2.0.1a0 (gpu)	2.1.0+cpu, 2.0.1a0 (gpu)	2.1.0+cpu, 2.0.1a0 (gpu)
Intel® Extension for PyTorch	2.1.0+cpu, 2.0.110+xpu	2.1.0+cpu, 2.0.110+xpu	2.1.0+cpu, 2.0.110+xpu	2.1.0+cpu, 2.0.110+xpu
Transformers	4.35.2(CPU), 4.31.0 (Intel GPU)	4.35.2(CPU), 4.31.0 (Intel GPU)	4.35.2(CPU), 4.31.0 (Intel GPU)	4.35.2(CPU), 4.31.0 (Intel GPU)
Synapse AI	1.13.0	1.13.0	1.13.0	1.13.0
Gaudi2 driver	1.13.0-ee32e42	1.13.0-ee32e42	1.13.0-ee32e42	1.13.0-ee32e42
intel-level-zero-gpu	1.3.26918.50-736~22.04	1.3.26918.50-736~22.04	1.3.26918.50-736~22.04	1.3.26918.50-736~22.04

Source:
https://github.com/intel/intel-extension-for-transformers/blob/main
https://github.com/intel/intel-extension-for-transformers?tab=readme-ov-file

Framework Level Optimization For AI/LLM Performance on Intel Hardware - IDEX

- Intel® Extension for DeepSpeed (IDEX): Extension that brings Intel GPU (XPU) support to DeepSpeed.
- DeepSpeed would automatically use IDEX when it is installed as a python package.
- After installation, models ported for DeepSpeed Accelerator Interface that run on DeepSpeed could run on Intel GPU device

Contributed HW support

- DeepSpeed now support various HW accelerators.

Contributor	Hardware	Accelerator Name	Contributor validated	Upstream validated
Intel	Intel(R) Xeon(R) Processors	cpu	Yes	Yes
Intel	Intel(R) Data Center GPU Max series	xpu	Yes	No

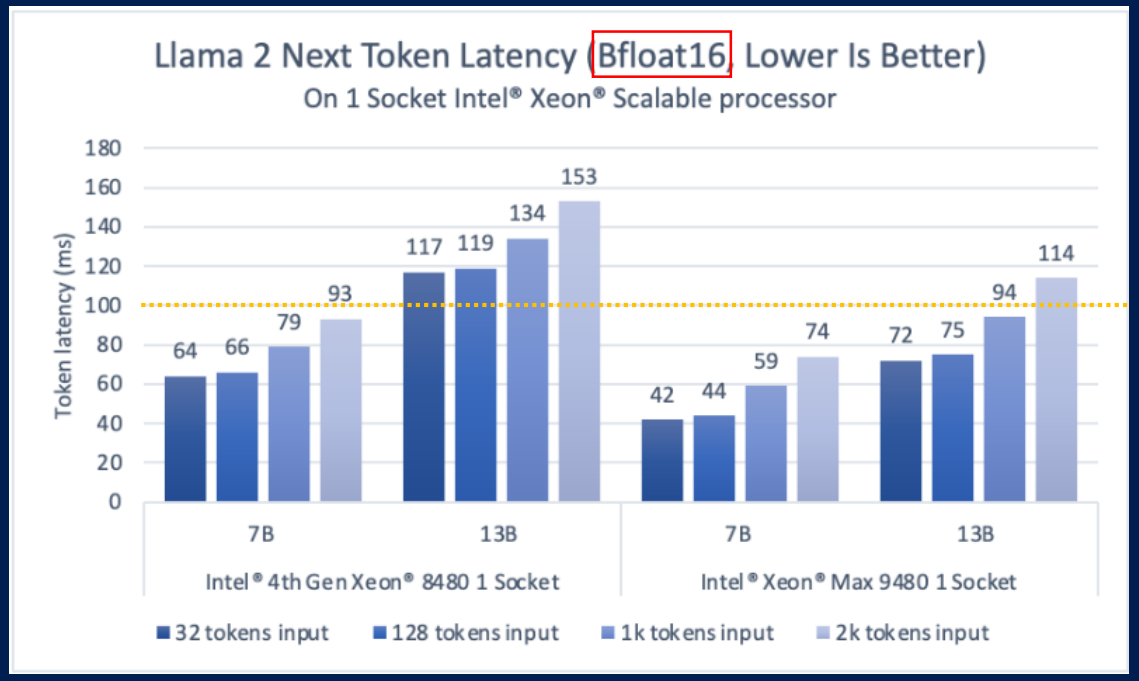
Source:
https://github.com/Microsoft/DeepSpeed
https://github.com/intel/intel-extension-for-deepspeed?tab=readme-ov-file

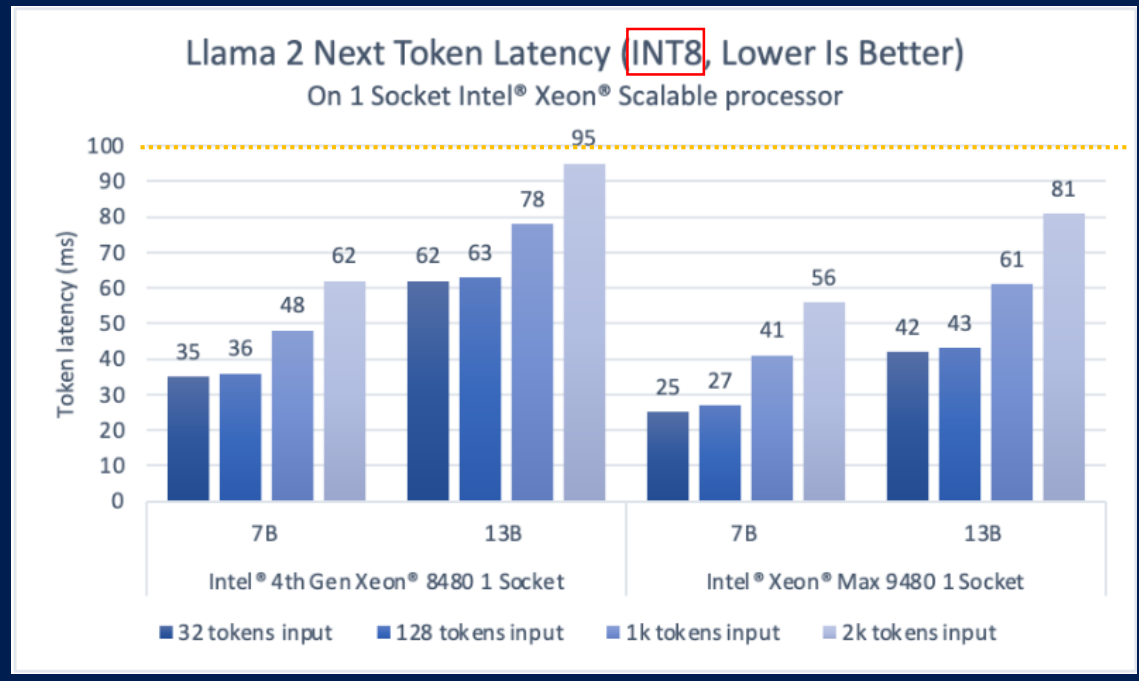
Framework Level Optimization For AI/LLM Performance on Intel Hardware – BigDL LLM

- BigDL LLM: Library for running LLM on Intel XPU (from Laptop to GPU to Cloud) using INT4/FP4/INT8/FP8 with very low latency (for any PyTorch model).

- 40+ model have been optimized/verified on bigdl-llm including LLaMA/LLaMA2, ChatGLM/ChatGLM2, Mistral, Falcon, MPT, Baichuan/Baichuan2, InternLM, QWen

AI on Intel Xeon: Accelerate **Llama 2** with Intel® AI Hardware and Software Optimizations





- Criteria: Next (2nd) Token Latency < 100ms
- Optimized with Intel Extensions for PyTorch (IPEX)

Source: https://www.intel.com/content/www/us/en/developer/articles/technical/accelerate-llama2-ai-hardware-sw-optimizations.html

intel Ai summit

Thank You!


intel