intel ai summit
英特爾 AI 科技論壇

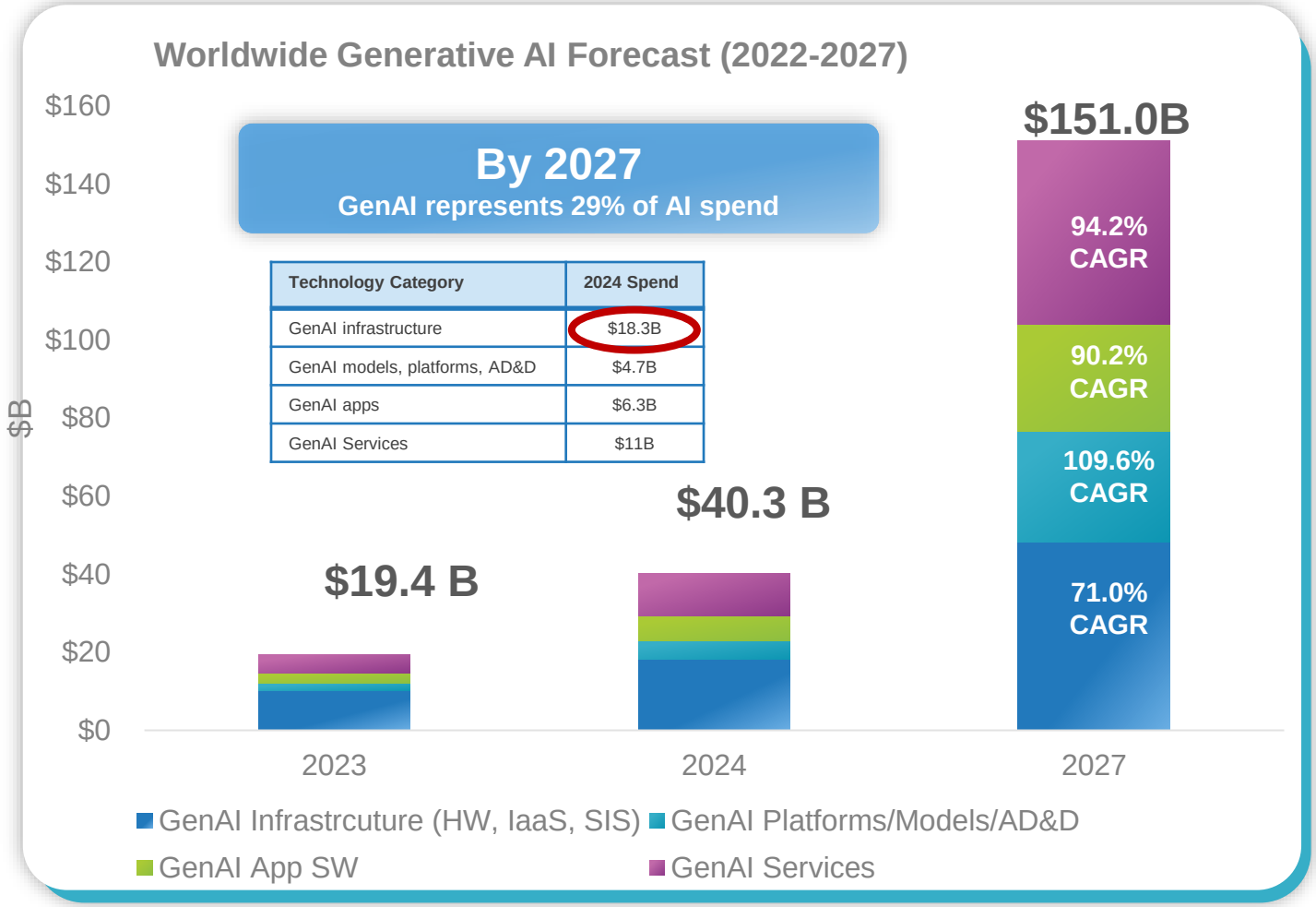
Bringing AI Everywhere

Enabling the AI continuum in every platform...from client and edge to data center and cloud.

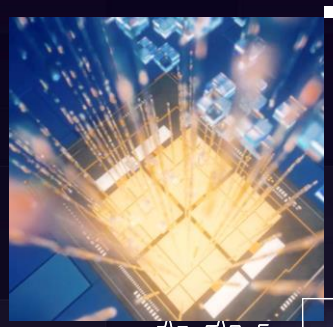
Alex Cheng 鄭智成
英特爾 業務暨行銷事業群商用業務總監

March 27th, 2024

The Generative AI Market Opportunity



Source: Source: IDC, GenAI Forecast Update December 2023



Bringing AI everywhere

+

$

Intel® AI product positioning

Enabling AI in every platform...from client and edge to data center and cloud.

AI use cases

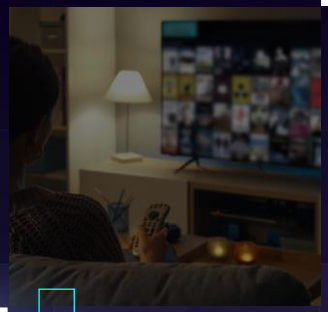
AI on the PC

Creator: Photo & video search & editing

Faster, more natural filters, higher quality previews & faster export times with automated, quicker searches.

Collaboration/streaming

New AI capabilities for next-gen video conferencing, streaming and collaboration, preserving battery life.



Mainstream gaming

New AI features for in-game, 3D animation for added realism, transcription & chat translation.

Productivity

AI assistants for writing, creating, coding and offline features, like text & grammar prediction.

Creator: Text to image

New AI effects & features for creating images with just a few descriptive words – marketing, advertising, design.

Accessibility

AI-assisted audio-visual capabilities for diverse user needs, making it easier to create and be productive on the PC.

"Unlocking the mundane"

With over 100 ISV's and 300 AI features planned for 2024, the Intel® Core™ Ultra offers a comprehensive AI PC experience with improved power & performance vs. prior generation*

*Intel Dec. 14, 20323, AI Everywhere event news release

AI PC powered by Intel® Core™ Ultra processors

3 Powerful AI Engines

GPU

High Throughput
Ideal for AI-accelerated digital content creation

NPU

Low Power
Ideal for sustained AI workloads and AI offload for battery life

CPU

Fast Response
Ideal for low-latency AI workloads

AI at the Edge

Manufacturing

Using edge analytics and machine learning, Audi's Neckarsulm factory implemented a solution to inspect 100 percent of the five million welds they make each day while inferring the results of each weld within 18 msecs. Audi was able to reduce labor costs by 30 to 50[1] percent and free employees for more valuable work at the company.

Retail

Nearly 80 percent of shoppers now prefer nontraditional checkout.[2] Nourish + Bloom is capitalizing on this trend with their autonomous grocery store. Just download the app, scan into the store, load groceries, walk out, and get a credit card charge. Customer journeys are faster and store personnel spend more time with customers and stocking shelves, according to data collected on each customer visit.

Cities & transportation

Ferrovial began the AIVIA Smart Roads initiative—aspiring to build roadways where conventional and autonomous vehicles can coexist. Along with Intel, Microsoft Azure, Capgemini, and others, Ferrovial is designing a ruggedized 5G- and AI-enabled roadside unit to relay near-real-time information to vehicles, warning of potential collisions, wrong-way driving, traffic conditions, and hazards, including pedestrians.

Healthcare

Primary ciliary dyskinesia (PCD) is a rare, inherited condition that causes defects in human cilia. It can cause respiratory distress early in life, leading to a lifetime of congestion, coughing, and chronic infections. Detection traditionally requires manual acquisition of cilia cross-sectional images for analysis using electron microscopy. To make it easy for medical professionals to apply their knowledge to an AI-based solution, the Royal Brompton team uses the Intel® Geti™ computer vision platform.

With over 84,000 edge AI deployments and counting, Intel offers a wide range of multipurpose compute and accelerators to deploy edge AI virtually anywhere.

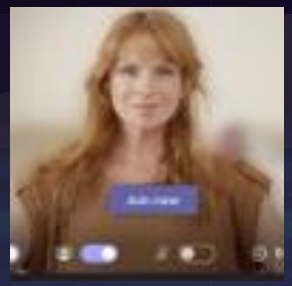
1. "Intel® Helps Audi Achieve Precision Manufacturing & Industrial Automation," Intel. Results may vary
2. "New Study Finds Self-Service Checkout Options Gaining Favor Across Demographic Groups." PYMNTS, Sept. 23, 2021, https://www.pymnts.com/news/retail/2021/new-study-finds-self-service-checkout-options-gaining-favor/

AI on the PC

Enhanced audio effects

Elevated video collaboration & streaming

Creator and gaming effects



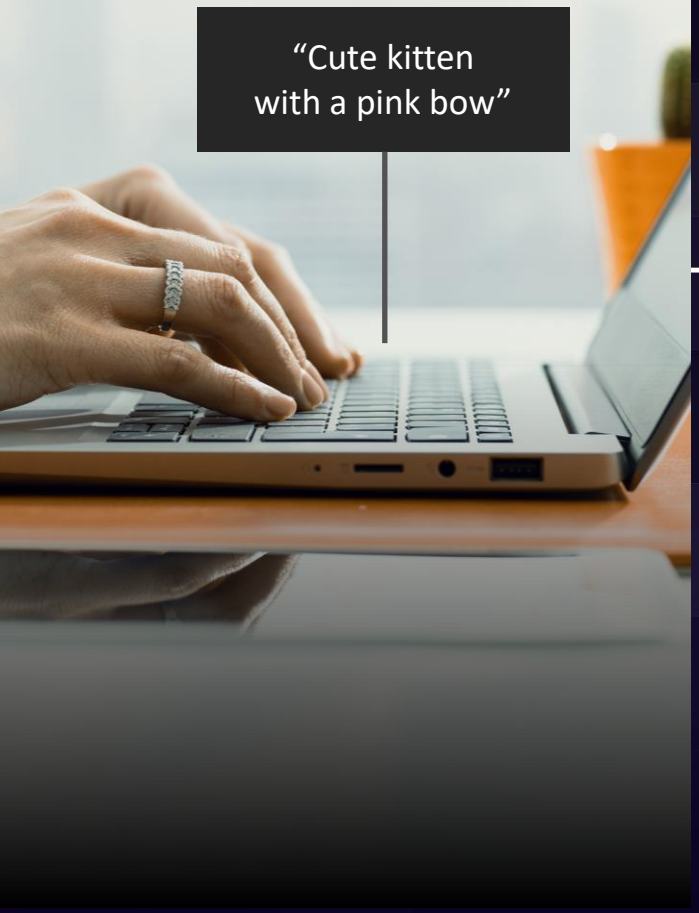
AI Assistants know your daily context

More creative, productive, & collaborative

Across everything you do

Today – Enhancements → **Tomorrow –** Everything

Stable Diffusion



"Cute kitten with a pink bow"

Text Prompt Understanding

Text Encoder

Image Creation "Diffusion"

Unet+ Unet-

Image Decoder

VAE

Output: Image

Four Models
Multiple Iterations of Unet in Diffusion Stage

Text Prompt Understanding

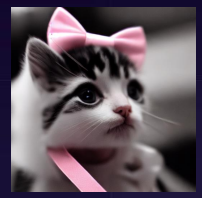
Text Encoder

Image Creation "Diffusion"

Unet+ Unet-

Image Decoder

VAE



UNet Step 1

UNet Step 2

...

UNet Step 4

UNet Step 5

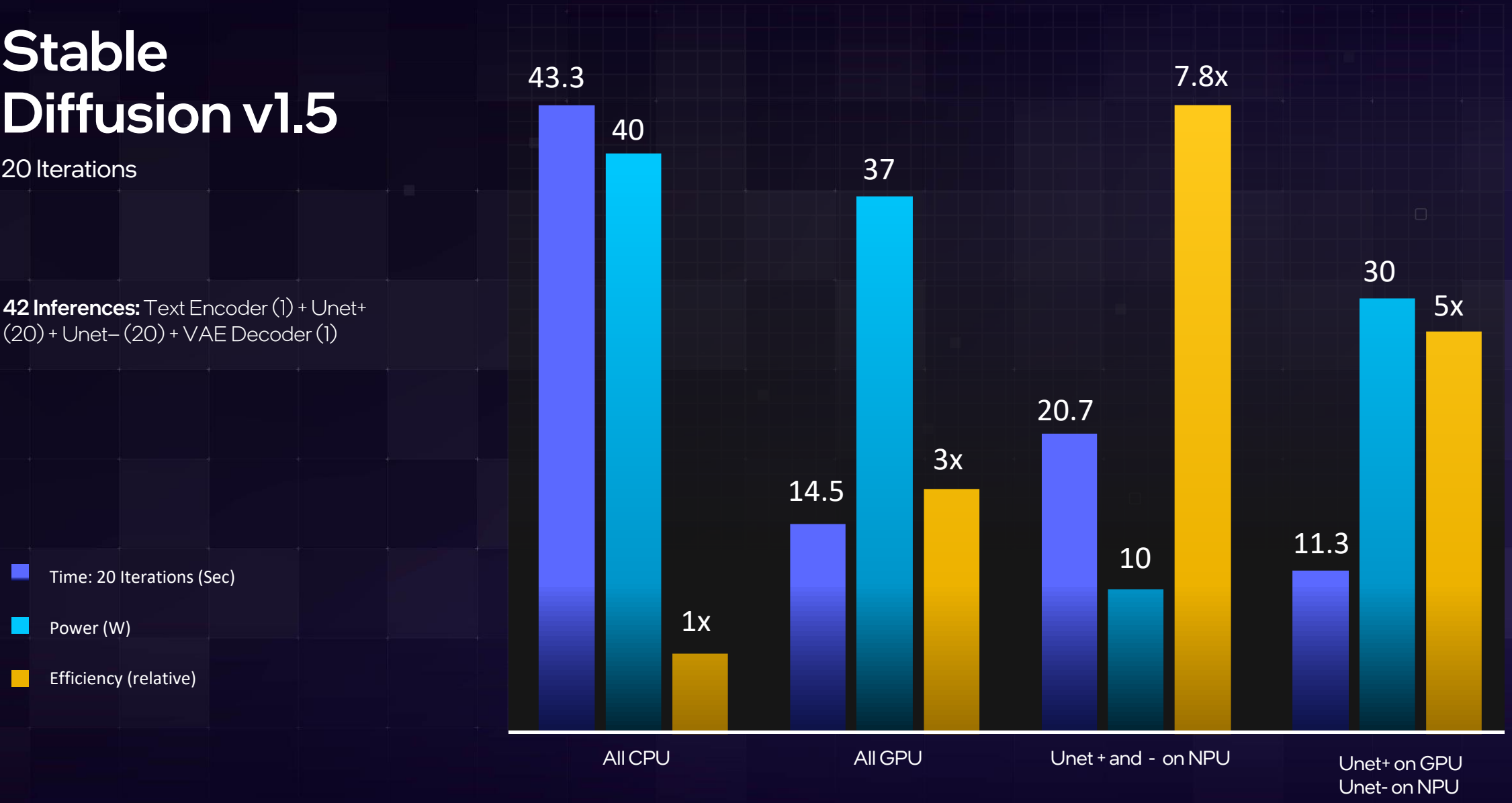
...

UNet Step 20

Stable Diffusion v1.5

20 Iterations

42 Inferences: Text Encoder (1) + Unet+ (20) + Unet– (20) + VAE Decoder (1)



*Based on internal estimates. Learn more at www.intel.com/PerformanceIndex. Results may vary.

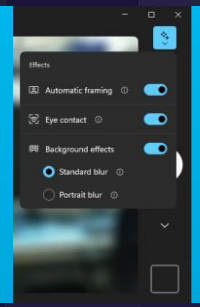
Embracing and enabling an open ecosystem **for innovation and scale**

Intel + Microsoft

Applications

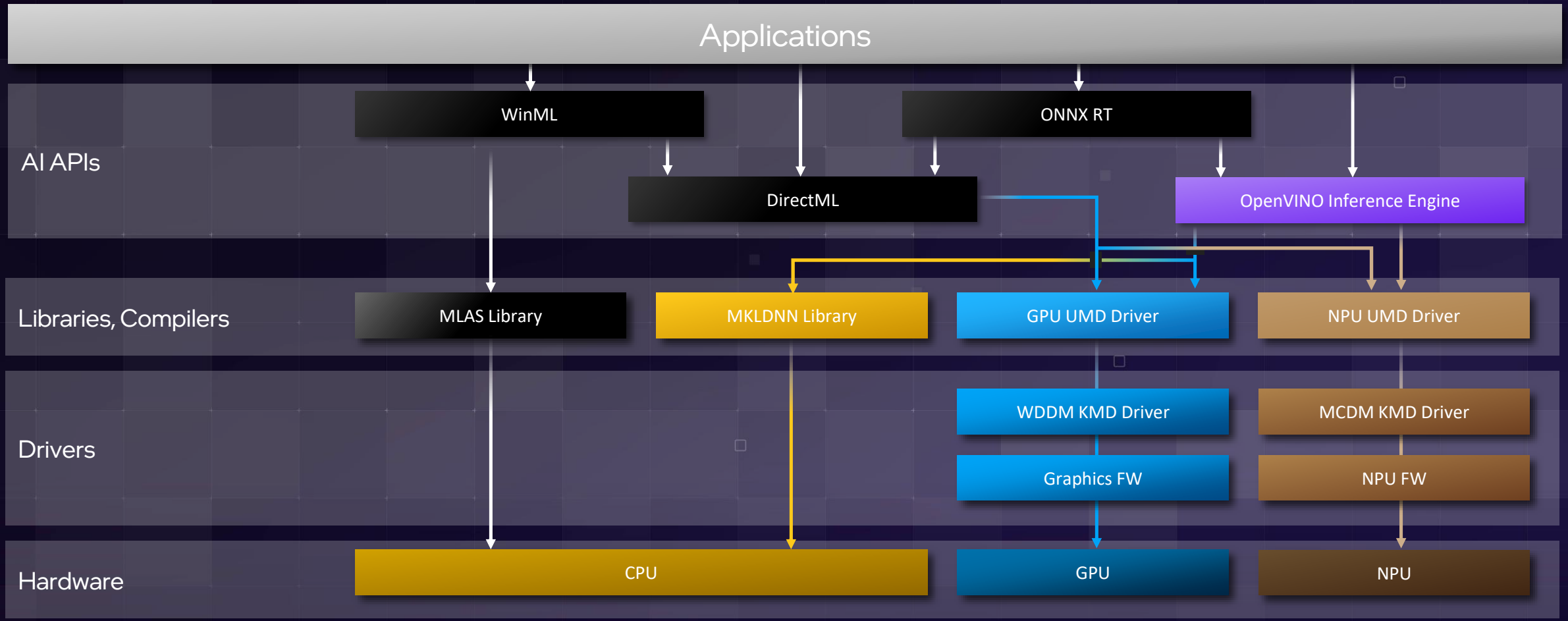
Office

Windows Studio Effects



DirectML

AI Software Stack



AI Software Stack

Microsoft Teams

Windows Studio Effects

Audio (APO)

Vision

AI APIs

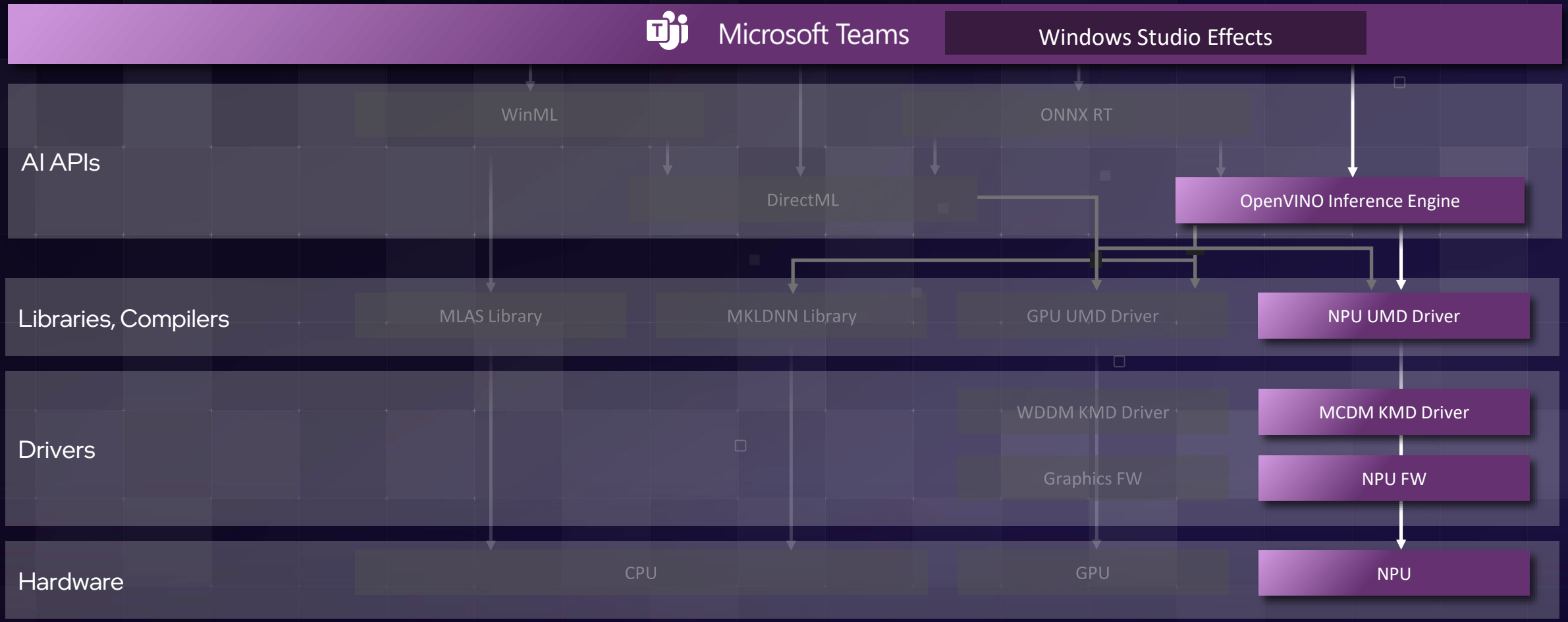
WinML

ONNX RT

DirectML

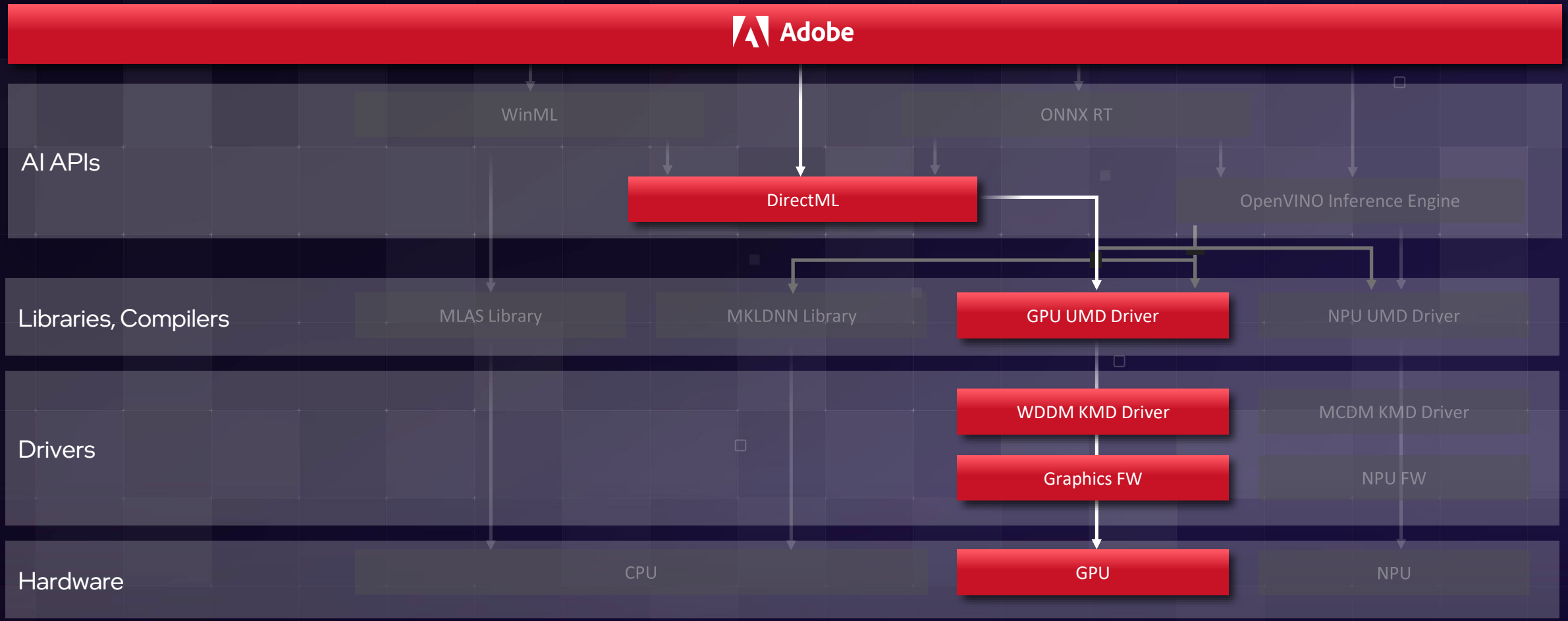
OpenVINO Inference Engine

AI Software Stack



AI Software Stack

AI Software Stack



AI Software Stack

Video Analytics

Decode → Process → Detect → Track → Classify

AI APIs

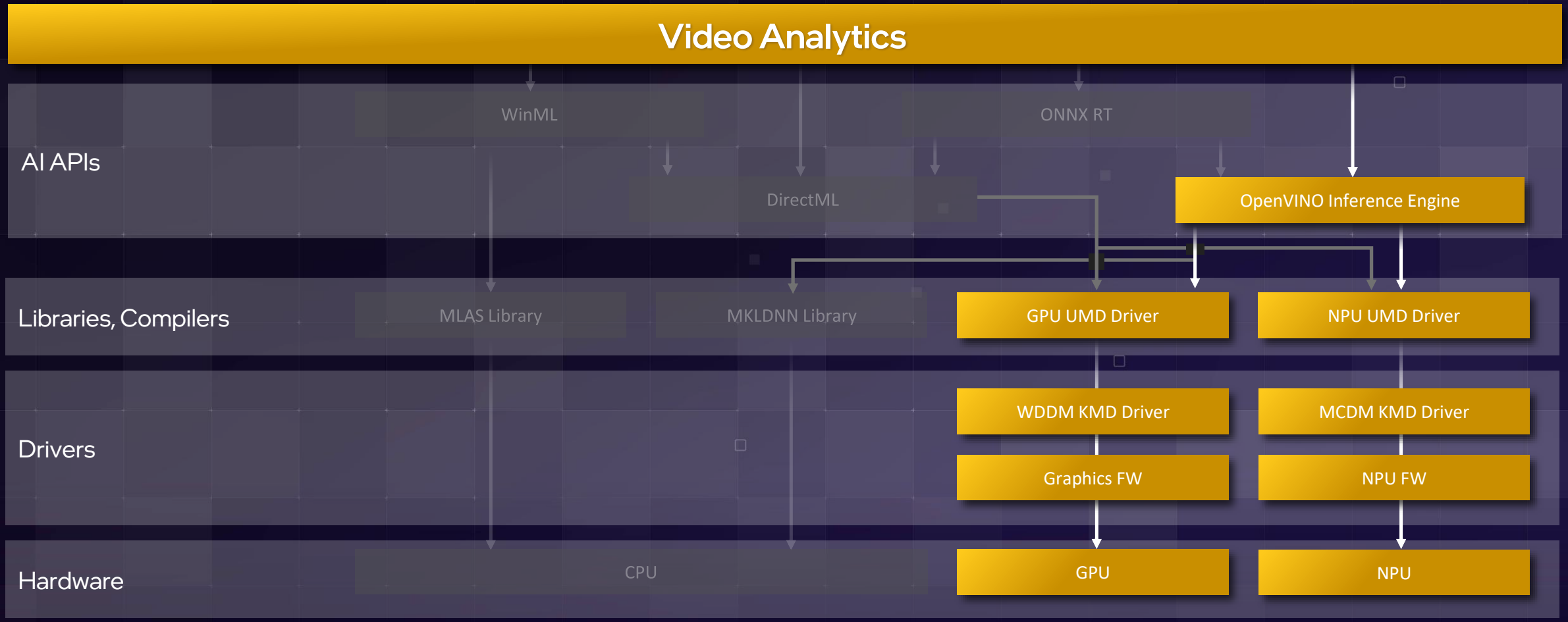
WinML

DirectML

ONNX RT

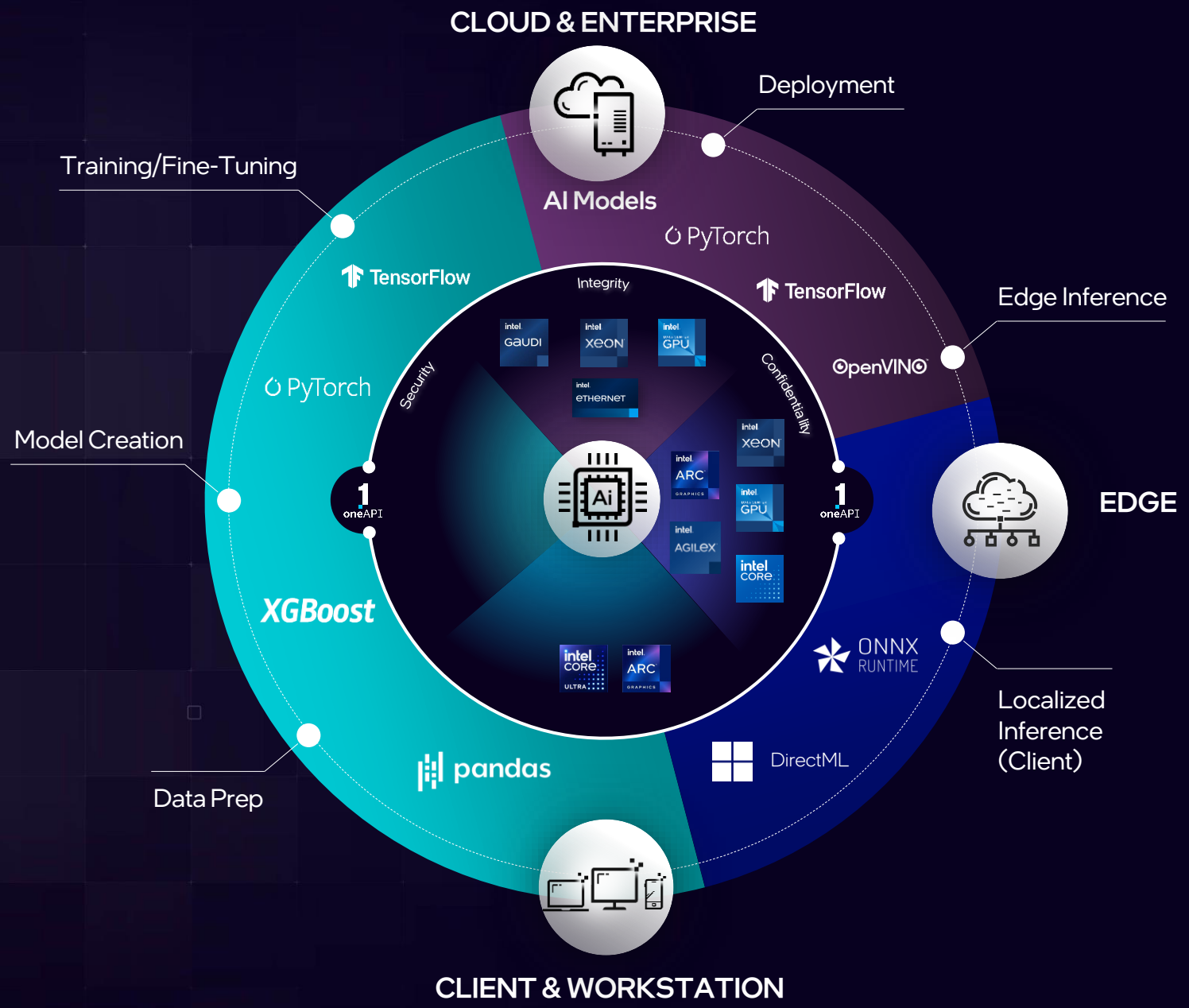
OpenVINO Inference Engine

AI Software Stack



AI Continuum

Bringing AI everywhere



Note: Intel® Core™ Ultra processors integrate NPU low power inference engine from 15th Gen processors onwards.

Bringing AI everywhere

Intel AI portfolio

Open Software Environment

oneAPI · OpenVINO · TensorFlow · PyTorch · ONNX Runtime · XGBoost · MODIN · DeepSpeed

Deep Learning Acceleration

Intel Gaudi — Dedicated Deep Learning Training and Inference

General Acceleration

Intel Data Center GPU Flex Series — AI Visual Inference, VDI, Media Analytics

Intel Data Center GPU Max Series — Parallel Compute, HPC, AI for HPC, Data Center

General Purpose

Intel Xeon · Intel Xeon Max Series — 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

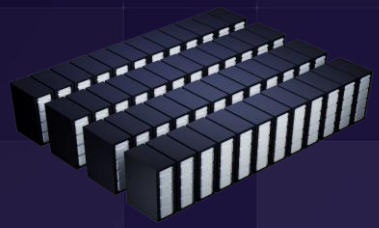
Intel Core Ultra · Intel Xeon D · Intel Data Center GPU Flex Series · Intel Arc Graphics · Intel Core · Intel Agilex · Intel Ethernet — Edge and Network AI Inference

Intel Core Ultra · Intel evo Edition · Intel Arc Graphics — Inference on Client

Building scalable AI systems







Enterprise Training & Deployment of LLMs

	Intel® Core™ Processors Intel® Core™ Ultra	Xeon Workstations Servers and racks 4th/ 5th Gen Intel® Xeon® SP
Training Parameters	N/A	~ 1B + E2E data Pipeline
Fine-tuning Parameters	N/A	<~ 10B 1 - 8+ nodes
Inference	~20 TOPS	< 20B *

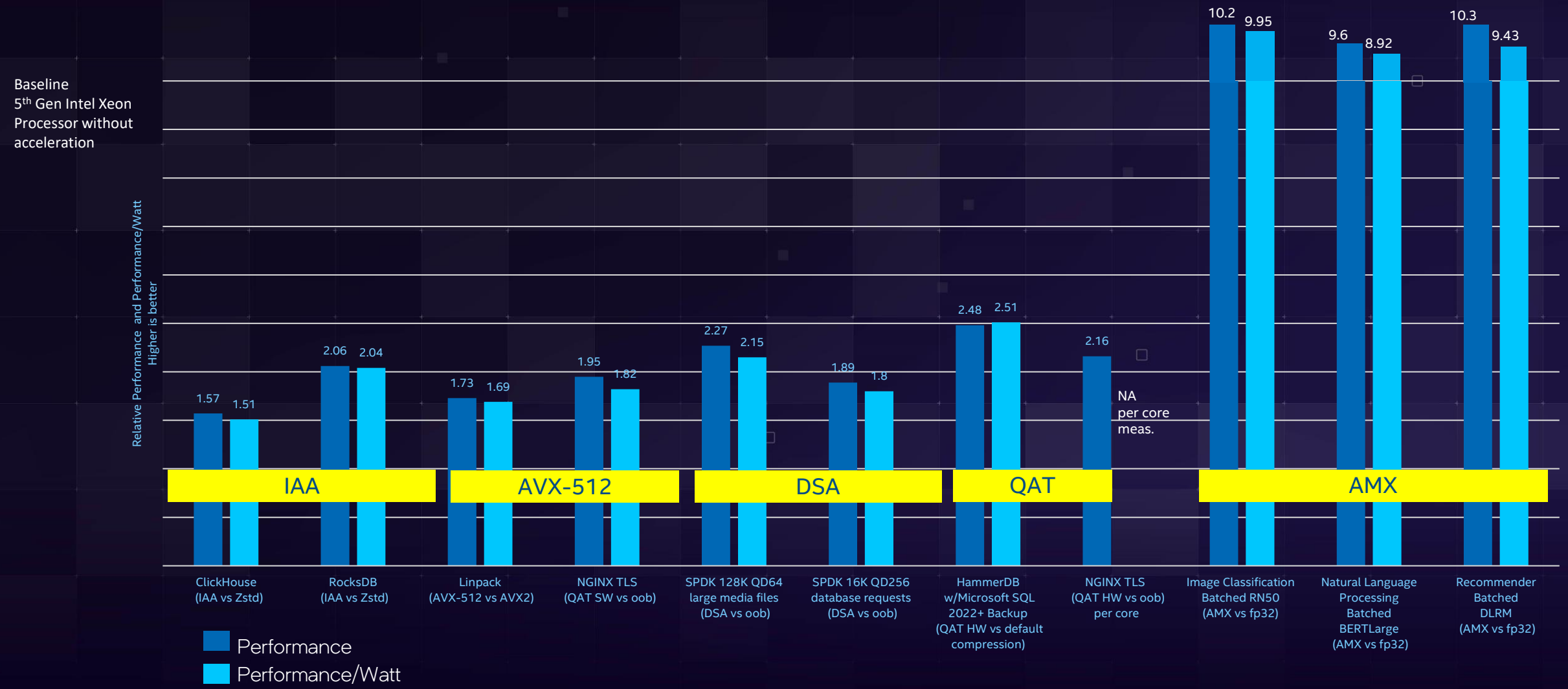
Dedicated DL training and inference of LLMs

	Gaudi®2 Server Dual-socket Xeon SP with 8 Gaudi® 2 devices	1 MegaPOD 8 Gaudi®2 Servers + 3 400G leaf switches	MegaPOD Cluster Sized as needed
Training Parameters	~ 20B	10Bs Shared across models of many sizes	Large Scale Distributed Training 10Bs - 1T+
Fine-tuning Parameters	~ 70B		Largest foundational models
Inference	~ 350B	10B-100Bs Shared across models of many sizes	Distributed Inference Shared across models of many sizes

* Gaudi2 numbers are estimated based on 16-bit precision. 8-bit precision allows for larger models.

A More Energy-Efficient Server Architecture

Intel® Accelerator Engines Raise Performance Per Watt Ceilings – 5th Gen Intel® Xeon® Scalable Processors



[1]Source: Intel Claims D1-D2 (IAA) , N16 (DSA), D5 (QAT), A17,19-20 (AMX). Results may vary

Source: Intel testing. See https://edc.intel.com/content/www/us/en/products/performance/benchmarks/5th-generation-intel-xeon-scalable-processors/ Results may vary

Hybrid AI: Seamless edge-to-cloud coordination

Cloud

LLM

Powerful LLM in public cloud for inferencing non-confidential data

Cloud-Edge handshake with non-confidential data

Near-real-time inferencing on client and edge combines with deeper cloud-based insight, using common tools and silicon without manual coding

Client and Edge

LLM

Local LLM inferencing protected data

Personalized outcome from both LLMs

HEALTHCARE

Use Generative AI to automate creation of personalized emails to patients while protecting privacy

RETAIL

Inference video data at the Edge and gain insights from Generative AI without backhauling costs

ENTERPRISE

Use Generative AI for productivity gains without exposing confidential information to public cloud

Intel AI software portfolio

Engineer Data

MODIN · SciPy · pandas · NumPy · Spark · Numba

Data Analytics at Scale†

Create Models

dmlc XGBoost · PyTorch · ONNX RUNTIME · OpenVINO · scikit learn · TensorFlow · DirectML · SigOpt AutoML · intel Neural Compressor · W3C WebAssembly WebGPU WebNN

Machine & Deep Learning Frameworks, Optimization and Deployment Tools†

Optimize & Deploy

oneAPI

- Intel® oneAPI Deep Neural Network Library
- Intel® oneAPI Collective Communications Library
- Intel® oneAPI Math Kernel Library
- Intel® oneAPI Data Analytics Library

Open, multiarchitecture programming model for CPUs, GPUs, and other accelerators

CLOUD & ENTERPRISE

CLIENT & WORKSTATION

EDGE

- **intel AI TOOLS** — 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

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

intel® Developer Cloud

Accelerate AI development using Intel-optimized software on the latest Intel® Xeon® processors, Intel® Data Center GPUs, and Intel® Gaudi® 2 accelerators.

cloud.intel.com

Get started with Intel

Get hands-on experience with the latest Intel® technologies. Empower your AI skills with Intel.

Early technology access

Evaluate pre-release Intel platforms and Intel-optimized software stacks.

Deploy AI at scale

Speed up AI deployments with the latest tools and libraries on Intel® Developer Cloud.

AI outcomes

+

AI innovation across the Data Center

Education	Teacher Assistant	Student Study Buddy	Parent Chat Portal
Health	Drug Discovery	Doctor Co-pilot	Patient Family Chatbot
Finance	Algorithmic Trading	Customer Portfolio Assistant	Risk / Credit Assessment
Retail	Product Promotion	Customer Interface and Sentiment Tool	Image Shopping Aid
Government	Gov Services Chatbot	Document Search Summarization	Live Language Translation
Energy	Energy Consumption Forecasting	Operational Performance	Energy Trading Assistant
Automotive	Autonomous Car Development	Multi-language in car aid	Supply Chain Optimization
Manufacturing	Factory Automation	Predictive Maintenance	Precision Agriculture
Telco	Personalized Customer Services	Network Automation	Operational Performance

AI outcomes

Advancing patient care with AI in Intel® Core™ Ultra processors

CPU-powered ultrasound imaging applications delivers more accessible and cost-effective imaging technology.

Situation

Samsung Medison is a pioneer in healthcare innovation. Their ultrasound imaging applications use AI for the most effective patient care.

Challenge

Previously, their applications were run on previous generation Intel Core processors accelerated by a competitor discrete GPU.

Solution

Samsung tested new Intel Core Ultra processors with built-in GPU engines. They saw significant AI performance improvements when compared to their previous gen CPU + dGPU combo. With Intel Core Ultra, Samsung Medison can offer advanced AI features in their next-gen ultrasound devices based solely on the CPU.

SAMSUNG MEDISON

Get the details:

Learn more

intel CORE ULTRA

Better customer experience with computer vision-based automation

In a drive-thru, time is of the essence. If a line is too long, guests will find something else. That's why over 20 of the world's best service brands are utilizing Hellometer's computer vision-based restaurant automation solution. Based on Intel® Core mobile processors with built-in AI acceleration and OpenVINO software, Hellometer is the world's first AI timer for quick service restaurants, using cameras to monitor and report on each guest's experience. The Hellometer enables restaurant operators to improve service speed by 47 seconds on average, or about $130k in added revenue per location.

Hellometer

Case study:

Learn more

AI outcomes

Meituan aligns compute to **business needs**

High growth business in food delivery and eCommerce with applications in merchant registration, QR code bike lock, package scanning, identity verification and more.

Situation

High growth business in food delivery and eCommerce reallocates GPU workloads to 4th Gen Xeon to lower AI inference costs.

Challenge

Fast growing business enabled by computer vision with increasing compute costs.

Solution

Meituan moved over 400 models from GPU to 4th Gen Intel® Xeon® with Intel® AMX, Intel Integrated Performance Primitives (Intel IPP), and Intel Extension for PyTorch (Intel IPEX)

美团 Meituan

70%
cost savings[1]

Case study

Learn more

Up to
4.13x
performance improvement[2]

AI outcomes

Preserving data privacy while accelerating healthcare innovation

Confidential computing platforms (CCPs), with memory encryption and privacy preserving analytics, can support healthcare organizations by helping protect data at rest and in use.

Situation

Novartis Biome develops diagnostic models and therapies for rare diseases. Rare disease information is sparse and dispersed across multiple hospitals and research institutions.

Challenge

Patient information is private and highly regulated. Hospitals do not want to move date off-prem or disclose private records to BeeKeeperAI or Novartis.

Solution

An Intel® SGX-enabled BeeKeeperAI node installed on-prem at each hospital analyzes private data and updates master model weights in the cloud. Neither Novartis nor BeeKeeperAI personnel ever see, or store, regulated health records.

NOVARTIS

BeeKeeperAI™

+ Regulated data

Case study

Learn more

AI transformation with the CPU

South Korea's web portal provider, Naver Corp., has replaced the main chip supplier of its artificial intelligence server for its map service, Naver Place, with Intel AI technology.

Situation

Naver is advancing its AI-powered location information provision service that run on graphic processing unit (GPU)-based servers.

Challenge

Naver is one of many global information technology firms increasingly disgruntled with Nvidia's GPU price hikes and a global shortage of its GPUs.

Solution

Naver has replaced the main chip supplier of its artificial intelligence server for its map service with Intel's Intel's central processing unit (CPU)-based server, 4th Gen Intel® Xeon® processors, after a month of test runs.

NAVER

Case study

Learn more

AI outcomes

Keeping AI data secure across the enterprise

Pioneering solution powered by Intel's AI supercomputer unlocks business value with custom datasets while maintaining high levels of security and data privacy.

Situation

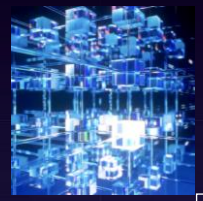
Intel and Boston Consulting Group were looking to advance generative AI with custom solution for enterprise clients that keeps private data in their trusted environments.

Challenge

Generative AI, supported traditionally by proprietary hardware and models, requires truly open access that enables more secure and scalable choice.

Solution

BCG collaborated with Intel to deliver an GenAI solution based on Intel's supercomputer powered by Intel® Xeon® Scalable processors and Intel® Gaudi® accelerators, as well as hybrid cloud-scale software. Users have reported step improvements, including a 25% growth in result relevancy and a 39% increase in improved work completion rates.



BCG + intel®

Case study

Learn more

AI outcomes

Deploying high-performance and cost-efficient AI at scale

The value and performance acceleration that the combination of Intel® Xeon® processors and Intel® Software brings to the entire AI lifecycle

Situation

The Netflix performance engineering team deploys AI to improve subscriber experience, from generating better recommendations to optimizing video delivery.

Challenge

Supporting the wide variety of devices and network conditions requires encoding multiple bitstreams for every title, and every subscriber is presented with a personalized home page and recommendations. These large-scale AI deployments must be performant yet cost-efficient.

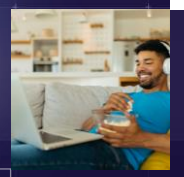
Solution

Netflix has realized large savings in cloud infrastructure costs by using Intel-optimized software, such as the Intel® oneAPI Deep Neural Network Library (oneDNN), XGBoost, and Intel® vTune™ Profiler, to get the most performance out of Intel® Xeon® processors without having to offload to more expensive GPUs.

NETFLIX

Case study

Learn more



AI outcomes

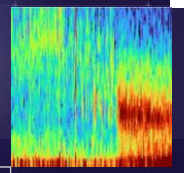
Optimizing diagnostic delivery & performance with AI software

Improving efficiency and build times of deep-learning models to broaden the system's applications to a wider range of psychiatric conditions and diseases.

宏智生醫科技股份有限公司
HippoScreen Neurotech Corp

Case study

Learn more



Situation

HippoScreen developed the Stress EEG Assessment (SEA) System, which helps doctors more accurately diagnose mental health conditions based on AI analysis of 90-second brainwave signals, providing a probability that an individual is suffering from depression.

Challenge

Developing the AI model to make it applicable in real clinical conditions and able to accommodate variation in the data requires finding the right combination of parameters and feature sets, which could take days to iterate on.

Solution

HippoScreen utilized PyTorch* Optimizations from Intel, Intel® Extension for Scikit-learn*, and Intel® vTune™ Profiler to speed training time by 2.4x on Intel® Xeon® processors.

Business outcome

Maximize value

Choose the hardware and software optimized for all your AI compute needs and available today.

Unlock new and enhanced experiences with

the AI PC: 300+ AI-accelerated ISV features throughout 2024

Accelerate AI with the broadest hardware portfolio that

matches compute and connectivity with your complete AI needs

intel XEON
intel GAUDI
intel DATA CENTER GPU
intel CORE ULTRA
intel ARC GRAPHICS
intel ETHERNET

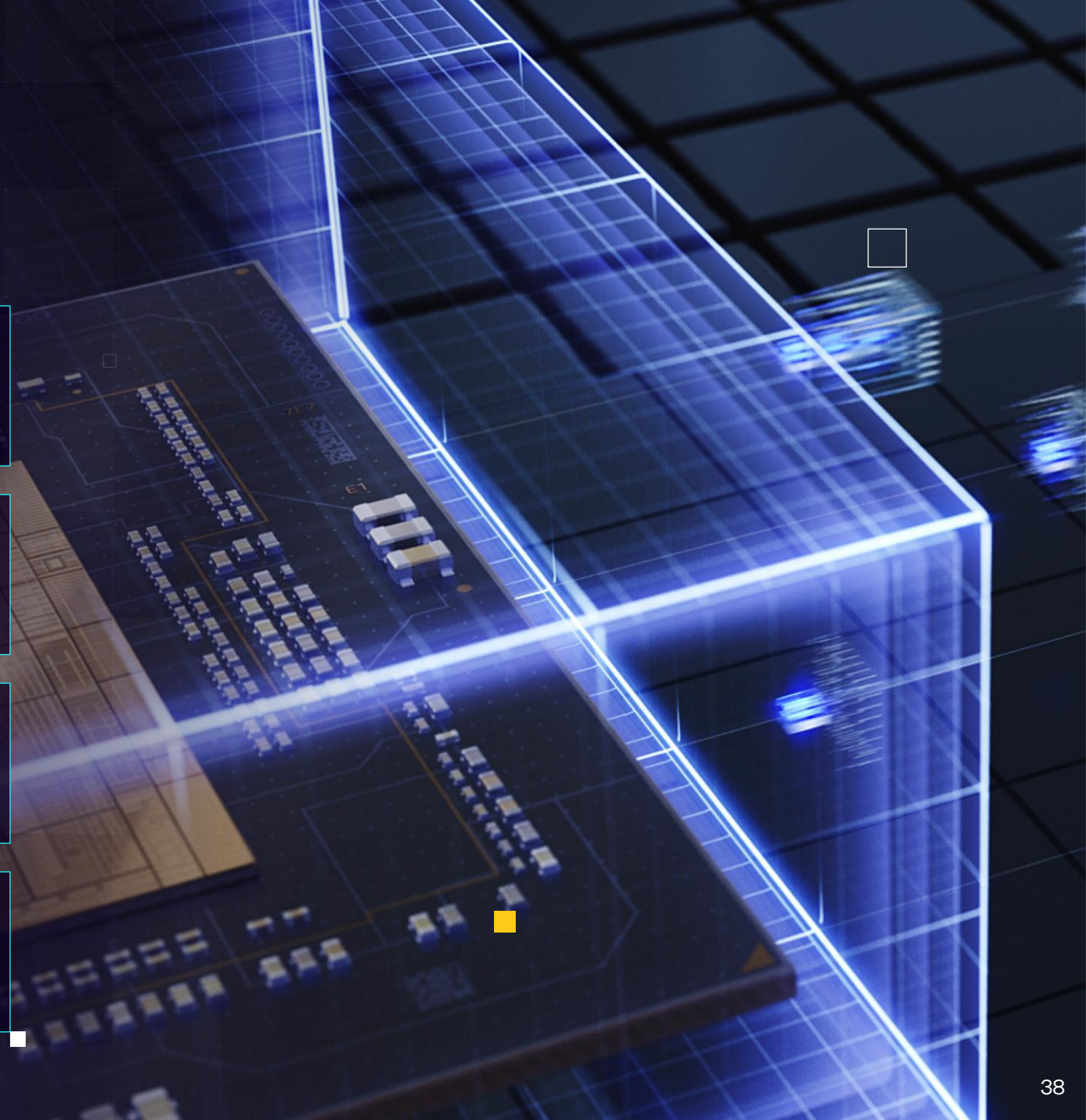
Create new opportunities from the client and edge to the data center & cloud

with hardware optimized by software **and open standards** for tomorrow's AI

Bringing AI everywhere

What's next?

1. Learn more at Intel.com/AI
2. Accelerate with Intel's AI Software Tools
3. Test the latest Intel AI hardware and software on the Intel Developer Cloud
4. Stay current with Intel's latest AI News



intel

Thank you

Bringing AI everywhere

Notices and Disclaimers

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

intel Ai summit

Thank You!