[Public]

AMD
XILINX

Adaptive Computing, a Disruptive Technology - AMD Unified AI Stack

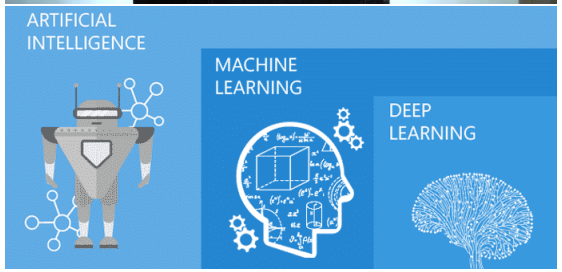
Jens Stapelfeldt – Sr. DBM
Data Center AI & Compute Markets

Jens.Stapelfeldt@amd.com
LinkedIn: www.linkedin.com/in/JensStapelfeldt

[Restricted - AMD Internal Distribution Only]

Jens Stapelfeldt TSL – EMEA – AMD / Xilinx

Jens.Stapelfeldt@amd.com
LinkedIn: www.linkedin.com/in/JensStapelfeldt



Jens technical background:

- 1996 /1998 - 3 Y ASIC designer
- 1998 - 12 Y Business Manager CE and Trainer for Doulos in CE and ARM ATC WW
  - Design methodology (Verification, SystemC,..)
  - (V)HDL, FPGA, ARM Architecture (ARM7 - Cortex-M to Cortex-A)
- 2010 - 5 Y Sr. Embedded FAE at TI
  - Sitara, DaVinci, OMAP, Keystone II
  - Industrial App. (IoT, Industrial Ethernet (EtherCAT, ProfiNet,..), Camera Vison ..)
- **Since Oct. 2016** Tech Sales Lead for **AMD/Xilinx in EMEA**
  - Supporting DACH, EE, Russia, Israel, India
  - 2019 BDM Data Center
  - TSL DACH, Nordic UK & Ireland
  - 2022 Sr. BDM Data Center AI & Compute Markets
- **2016-2018 Part time MBA** in Bremen with weeks in Hong Kong, Cambridge, Dublin Business school!
- **Since Oct. 2018 guest lecturer** for “Digital Signal Processing (VHDL / FPGA design”).

MBA: In May 2018 I finished my MBA with Master Thesis in International Marketing looking into about 600 AI/ML Start-ups in EMEA!
LinkedIn Article: https://www.linkedin.com/pulse/analysing-europes-aiml-start-up-landscape-using-jens-stapelfeldt/

© Copyright 2022 AMD

[Public]

High Performance and Adaptive Computing

Cloud, Network, Hyperscale & Supercomputing

5G & Comms Infrastructure

AI & Analytics Everywhere

Adaptive Intelligent Systems

Gaming, Simulation and Visualization

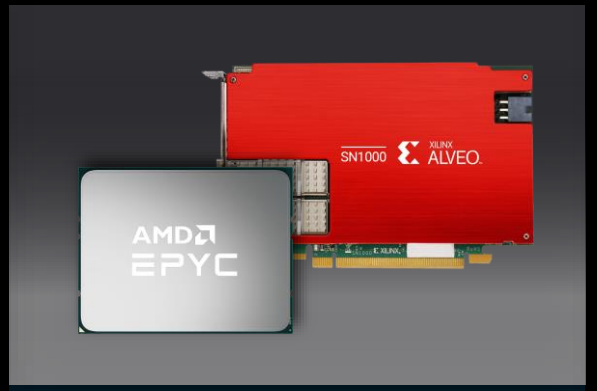
Smarter Client Devices & Edge

At The Center of Today's Intelligent World

AMD

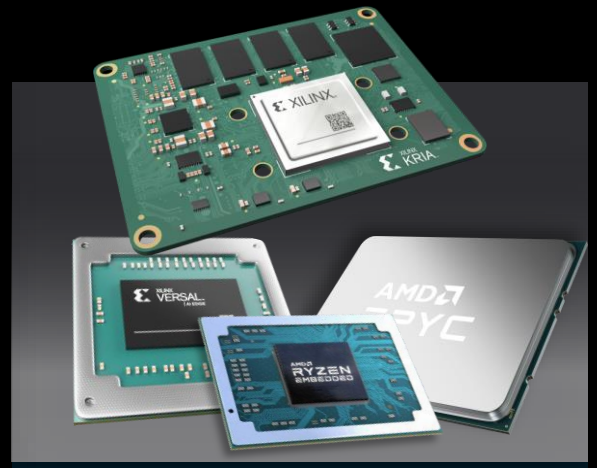
AMD + XILINX AI OPPORTUNITIES

AI Inference and Training



Data Center and Communications

Automotive



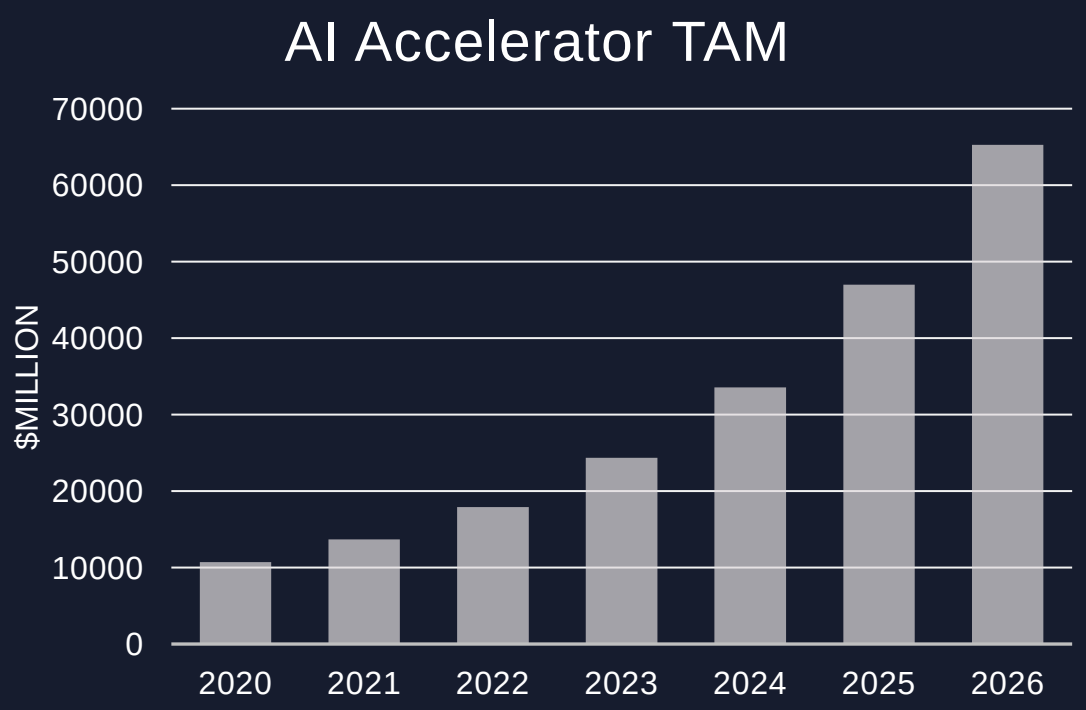
Embedded

- AI OPPORTUNITIES in all areas

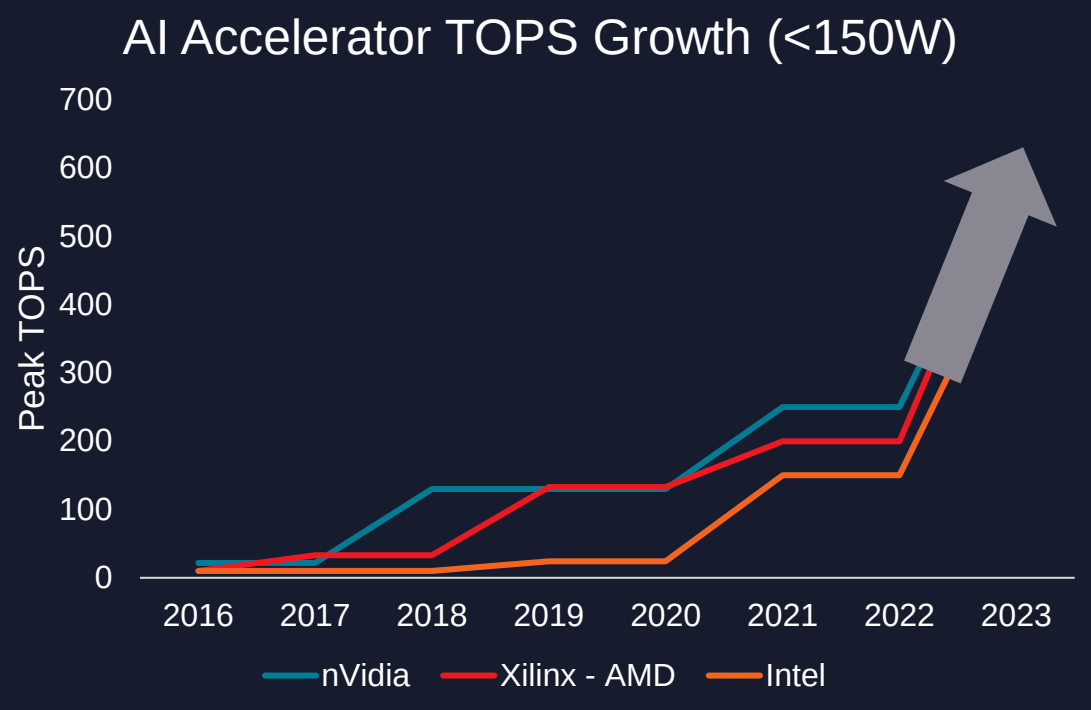
AMD

[AMD Official Use Only - General]

AI Accelerator Demands Major Innovation



DC AI Capex Growing at 36.7% CAGR, projected to $65B TAM by 2026



AI Accelerator Peak TOPS Growing Exponentially to Keep Up with the Model Innovation

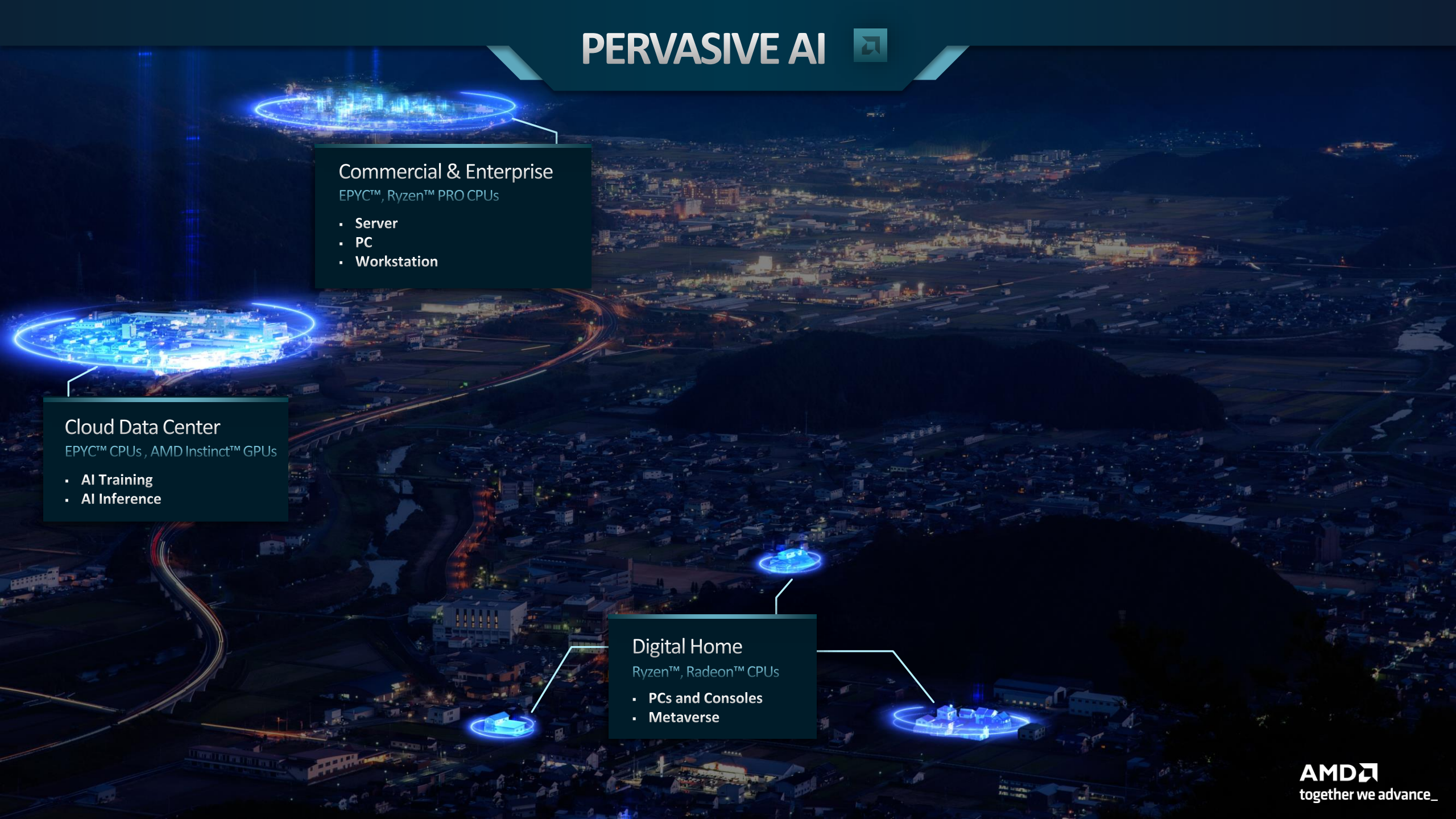
Source: Markets and Markets (Data Center Accelerator Market Global Forecast to 2026)

**2022-2023 is projected based on history*
*** Only including AI cards with 150W or less power consumption*

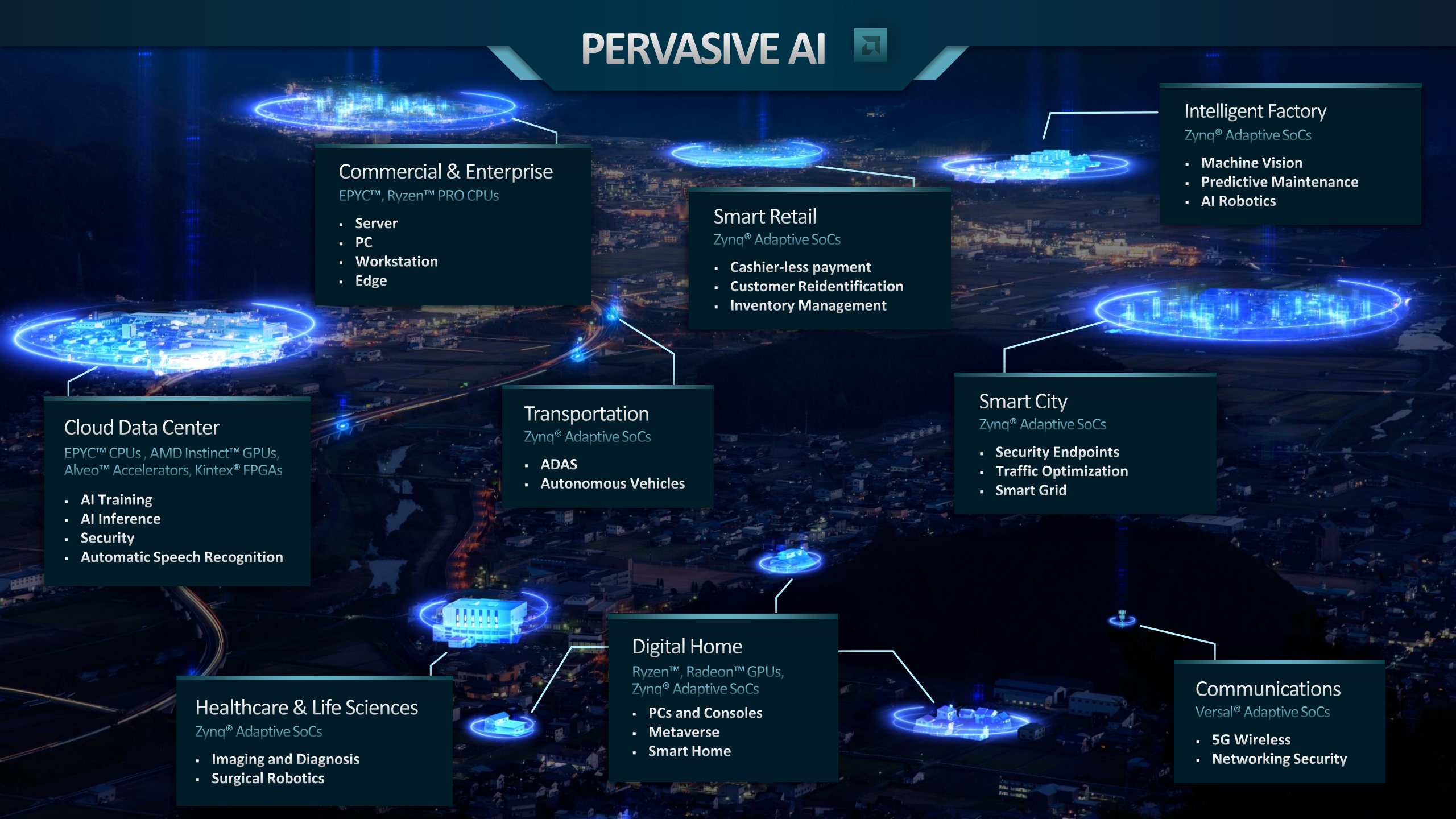
AMD
XILINX

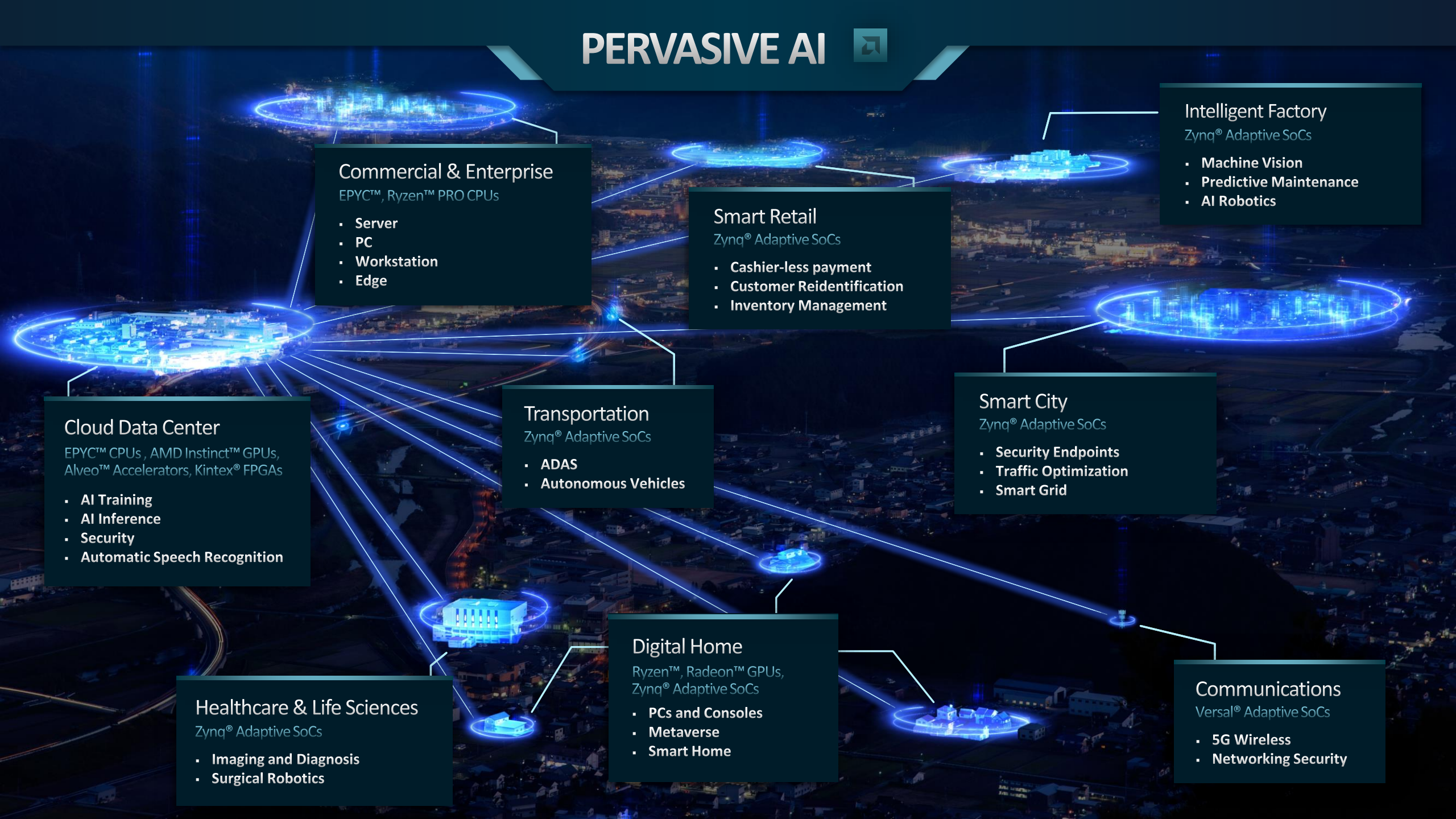
AMD
together we advance_

PERVASIVE AI



AMD
together we advance_


PERVASIVE AI
Commercial & Enterprise
EPYC™, Ryzen™ PRO CPUs
Server
PC
Workstation
Edge
Intelligent Factory
Zynq® Adaptive SoCs
Machine Vision
Predictive Maintenance
AI Robotics
Smart Retail
Zynq® Adaptive SoCs
Cashier-less payment
Customer Reidentification
Inventory Management
Cloud Data Center
EPYC™ CPUs , AMD Instinct™ GPUs, Alveo™ Accelerators, Kintex® FPGAs
AI Training
AI Inference
Security
Automatic Speech Recognition
Transportation
Zynq® Adaptive SoCs
ADAS
Autonomous Vehicles
Smart City
Zynq® Adaptive SoCs
Security Endpoints
Traffic Optimization
Smart Grid
Digital Home
Ryzen™, Radeon™ GPUs, Zynq® Adaptive SoCs
PCs and Consoles
Metaverse
Smart Home
Healthcare & Life Sciences
Zynq® Adaptive SoCs
Imaging and Diagnosis
Surgical Robotics
Communications
Versal® Adaptive SoCs
5G Wireless
Networking Security


PERVASIVE AI
Commercial & Enterprise
EPYC™, Ryzen™ PRO CPUs
Server
PC
Workstation
Edge
Intelligent Factory
Zynq® Adaptive SoCs
Machine Vision
Predictive Maintenance
AI Robotics
Smart Retail
Zynq® Adaptive SoCs
Cashier-less payment
Customer Reidentification
Inventory Management
Cloud Data Center
EPYC™ CPUs , AMD Instinct™ GPUs, Alveo™ Accelerators, Kintex® FPGAs
AI Training
AI Inference
Security
Automatic Speech Recognition
Transportation
Zynq® Adaptive SoCs
ADAS
Autonomous Vehicles
Smart City
Zynq® Adaptive SoCs
Security Endpoints
Traffic Optimization
Smart Grid
Digital Home
Ryzen™, Radeon™ GPUs, Zynq® Adaptive SoCs
PCs and Consoles
Metaverse
Smart Home
Healthcare & Life Sciences
Zynq® Adaptive SoCs
Imaging and Diagnosis
Surgical Robotics
Communications
Versal® Adaptive SoCs
5G Wireless
Networking Security

PERVASIVE AI STRATEGY

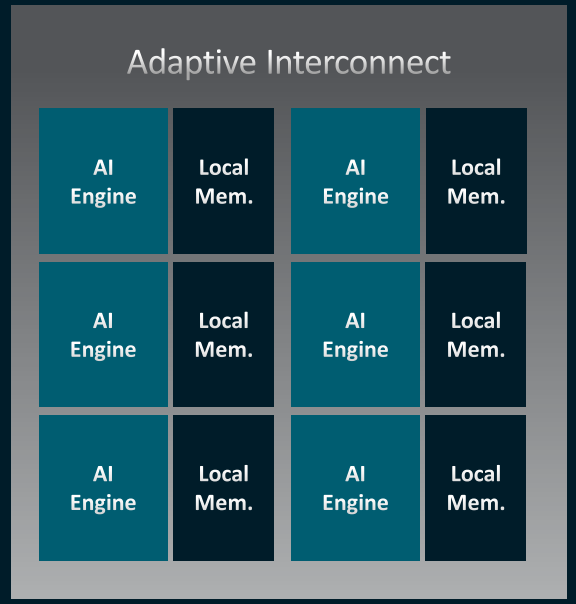
- Leadership AI IP with AMD CDNA™ and scalable AI Engine architecture
- Broaden AMD AI product portfolio across cloud, edge, and endpoint applications
- Unified AI stack to empower developers across the AMD portfolio

AMD
together we advance_

AMD XDNA: ADAPTIVE ARCHITECTURE IP

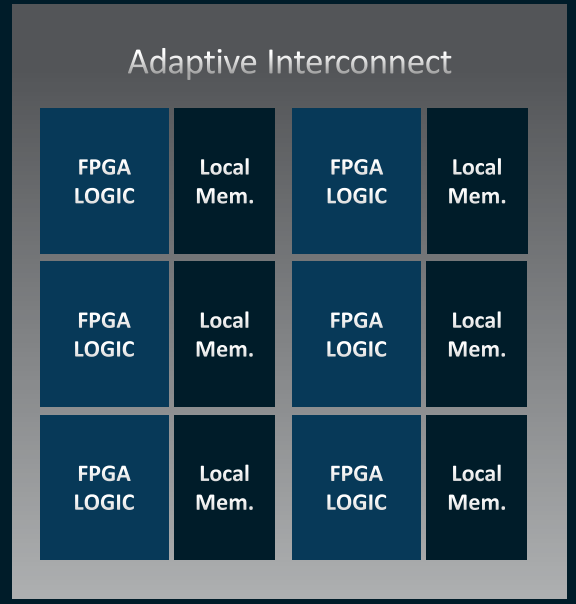
- Dataflow architecture optimal for AI and signal processing applications
- Highly-scalable array of engines with local memory and data movers
- Leverages deep expertise of compiling algorithms to FPGAs and adaptive SoCs

AI Engine
(AIE)



High-Performance and Energy Efficiency for AI and Signal Processing

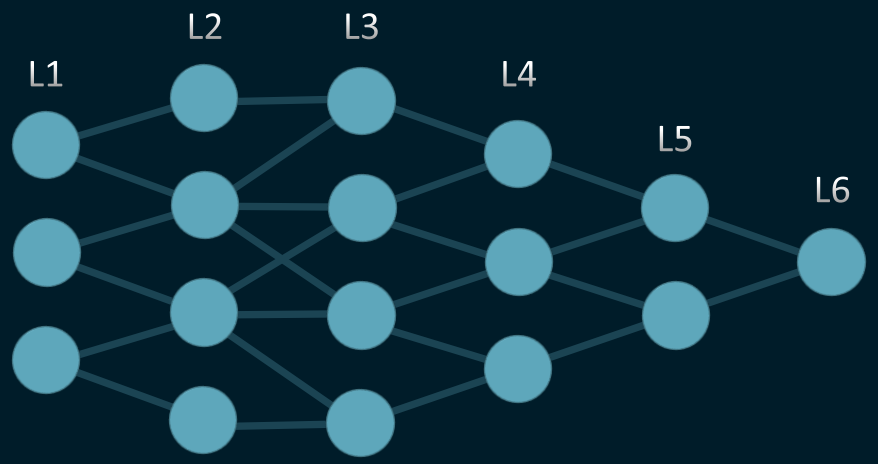
FPGA Fabric
(FPGA)



Leading FPGA for broad set of applications and AI

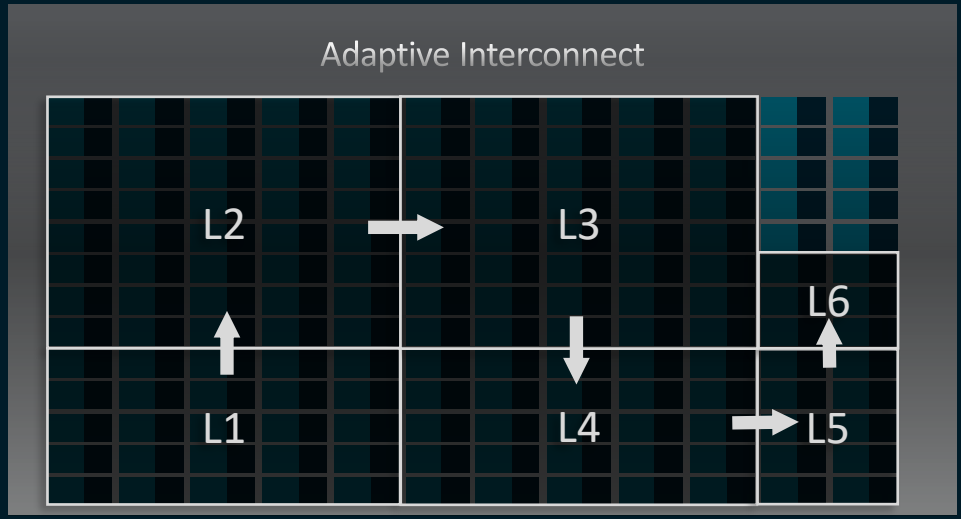
AMD AI ENGINE: ADAPTIVE DATAFLOW PROCESSOR

Deep Neural Network (DNN)



Data “flows” from layer to layer, connections between layers are often “sparse”

DNN Runs Optimally on AIE

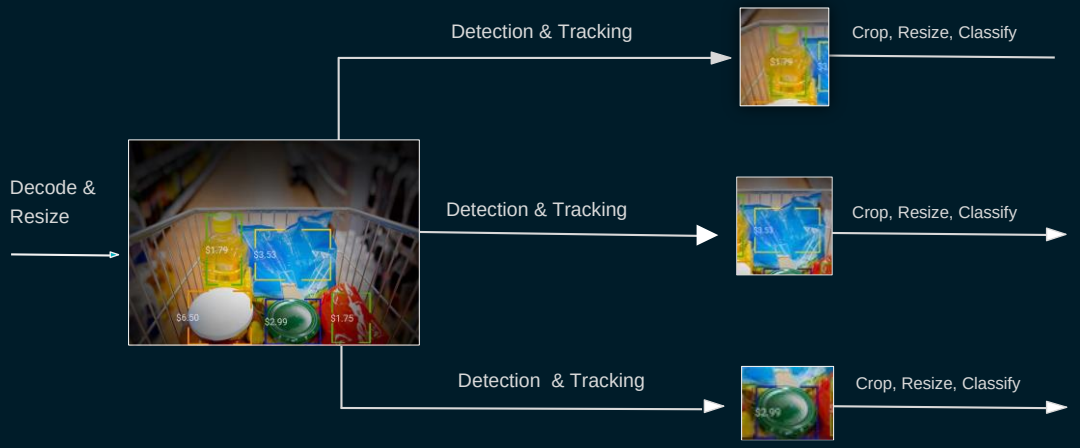


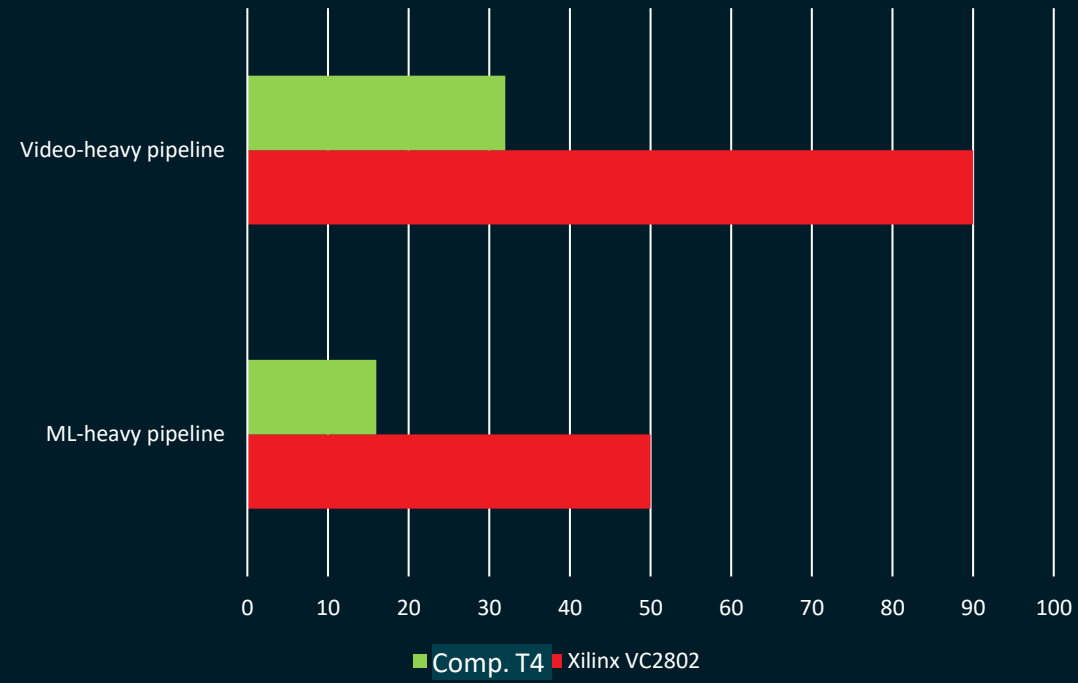
Dataflow architecture, sparsity, efficient datatypes deliver high performance and low power

High-Performance, Energy Efficient, and Customizable for AI Workloads

Example 1: Video Analytics

ML-Heavy: H.264 Decode + CV + Yolov3 + 9 x RN18
Video-Heavy: H.264 Decode + CV + tinyYolov3 + 9 x RN50



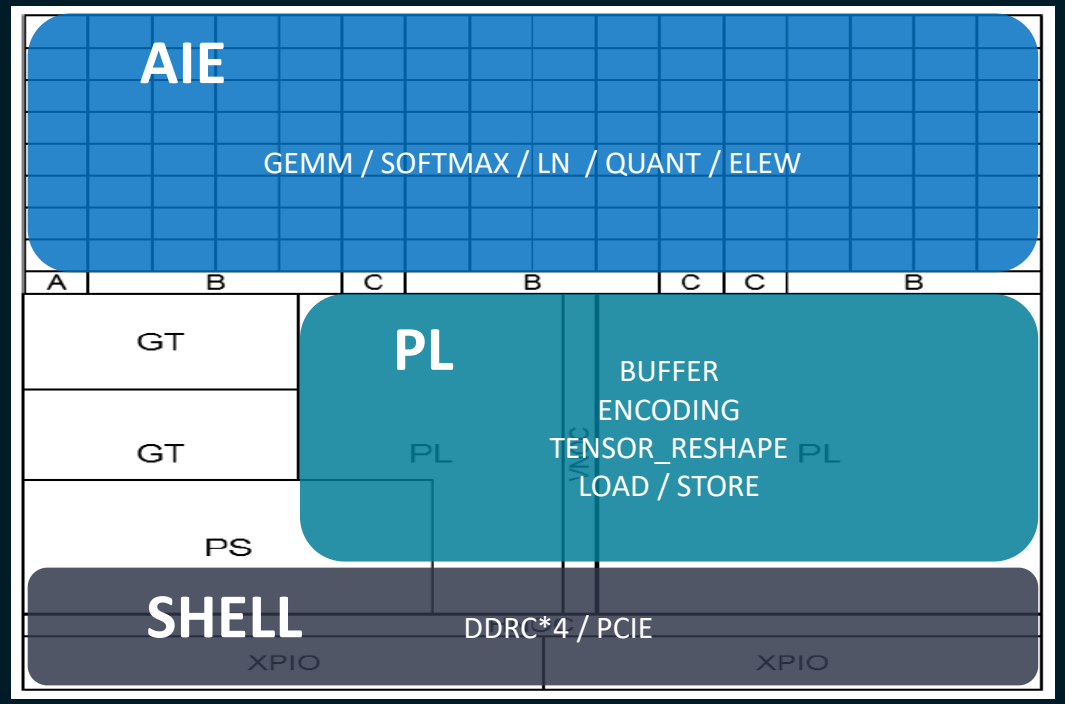


~ 3x competitor in End-to-End Throughput

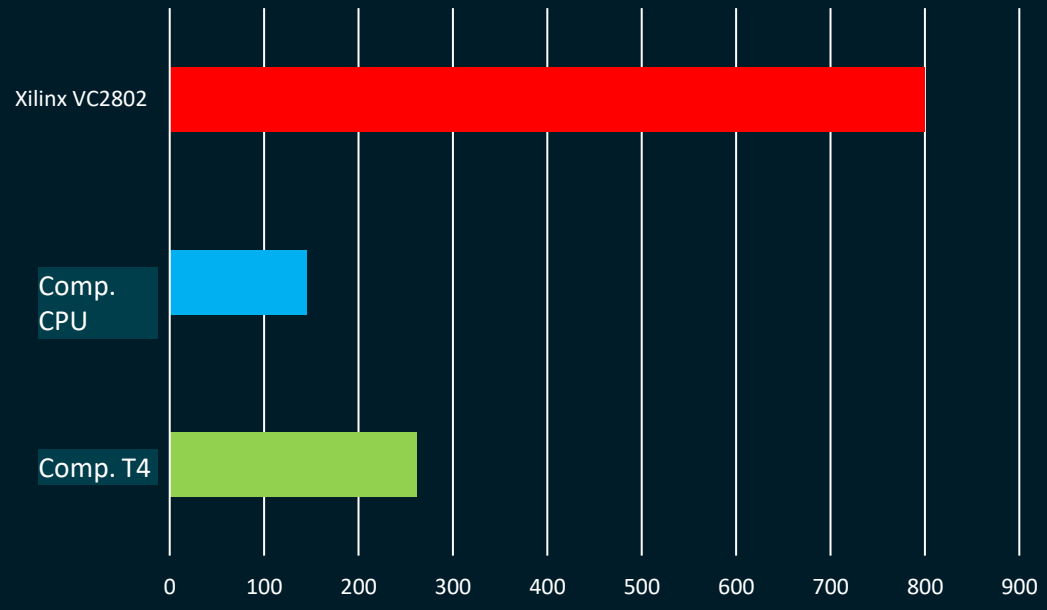
AMD
together we advance_

Example 2: Natural Language

Transformer-based XDNA Design



BERT-large (Sentences / sec)

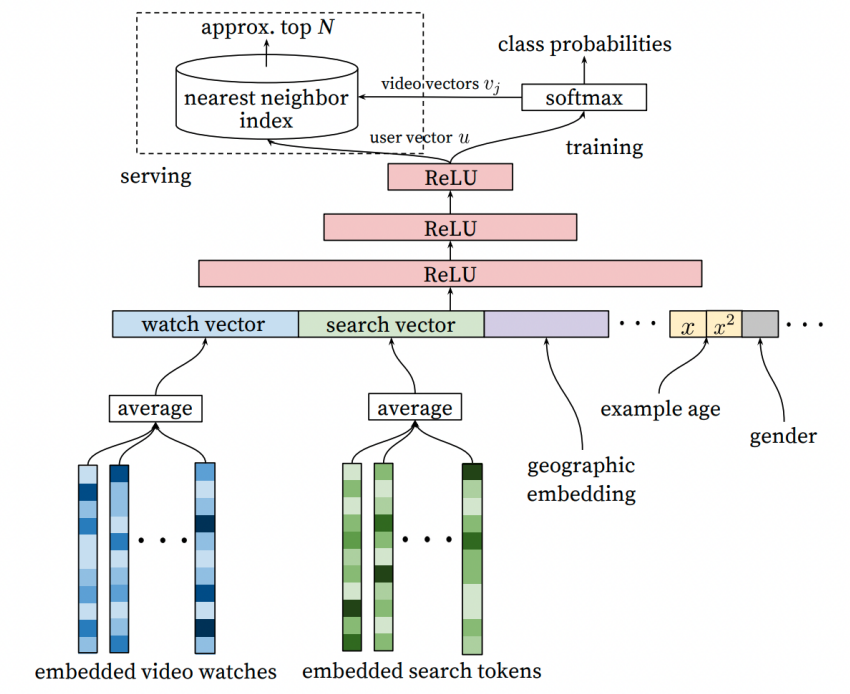


Flexible Architecture for GEMM and Non-Linear Operations

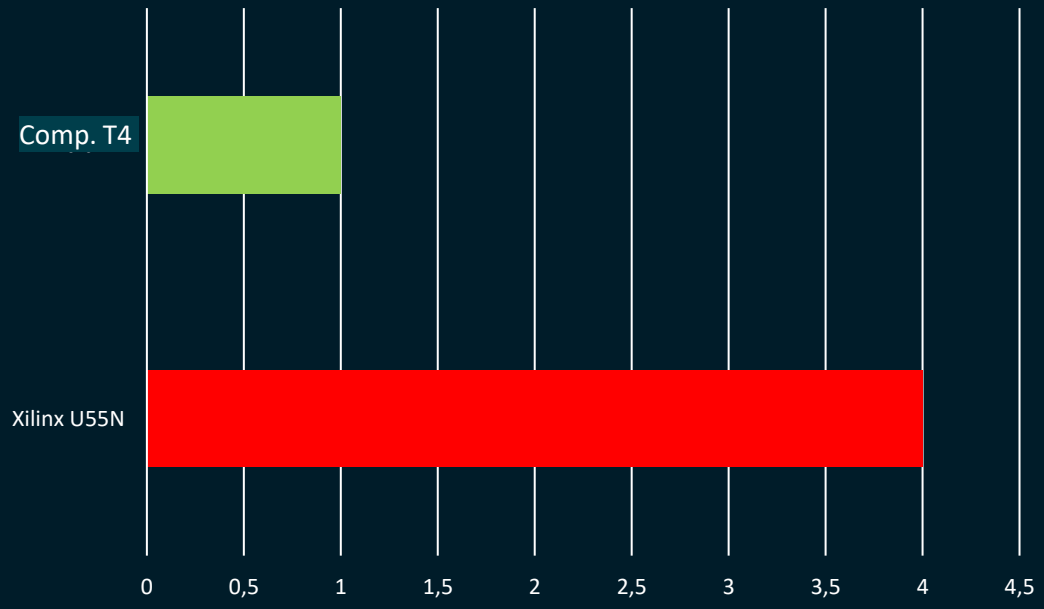
AMD
together we advance_

Example 3: Recommendation AI

HBM-based Pipeline to Hide Compute Latency



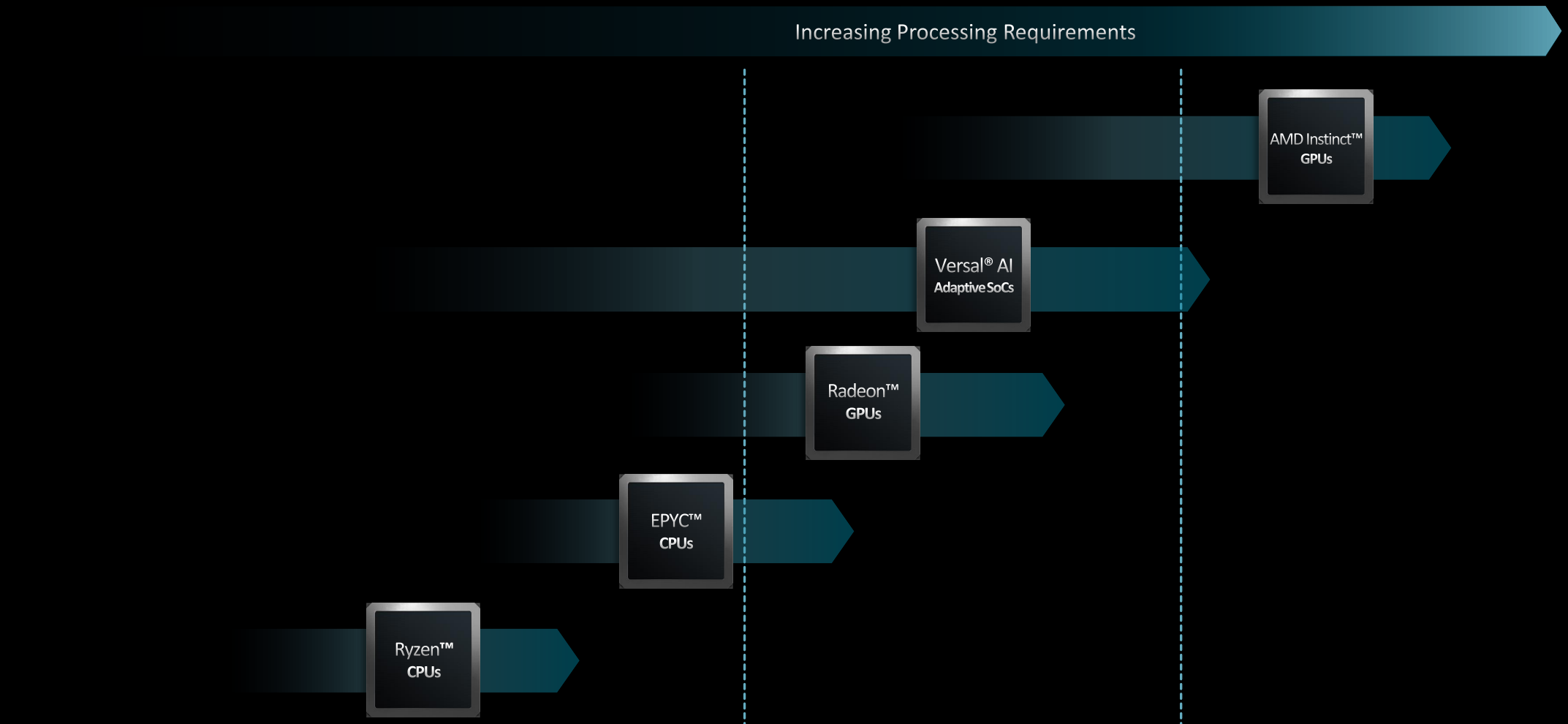
Vector Retrieval (Speed-up factor)



4x competitor in End-to-End Throughput at 10 ms Latency

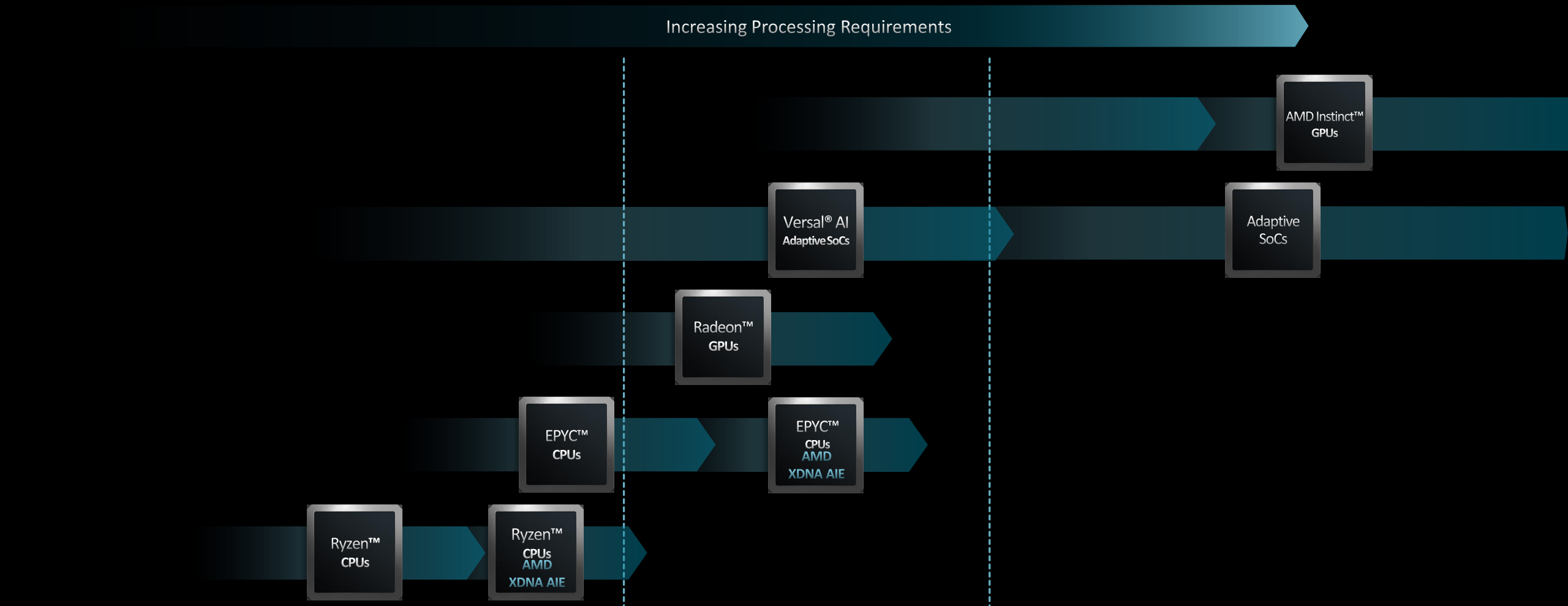
AMD
together we advance_

AI APPLICATION COVERAGE



Training	Small-medium models (EPYC™ CPUs only)	Medium-large models	Large-very large models
Inference	Mid-size models, mid-range performance and memory BW, high memory capacity **(e.g., face recognition, recommendation systems)**	Larger models, high performance and more memory BW **(e.g., video analytics)**	Very large models, efficient data types, and high memory BW **(e.g., NLP, large recommendation systems)**

AI APPLICATION COVERAGE



Training	Small-medium models (EPYC™ CPUs only)	Medium-large models	Large-very large models
Inference	Mid-size models, mid-range performance and memory BW, high memory capacity **(e.g., face recognition, recommendation systems)**	Larger models, high performance and more memory BW **(e.g., video analytics)**	Very large models, efficient data types, and high memory BW **(e.g., NLP, large recommendation systems)**

AMD AI SOFTWARE TODAY

CPU Stack

- Optimized Inference Models
- WinML ONNX Runtime | Pytorch | TensorFlow
- ML Graph Compiler
- ZenDNN, AOCL Optimized Library
- ZEN Studio (AOCC)
- Windows Runtime | Linux Runtime

Ryzen™ CPUs | EPYC™ CPUs

ROCm™ Platform

- Optimized Inference Models
- ONNX Runtime | Pytorch | TensorFlow
- MIGraphX: AI Development Tools
- MIGraphX: ML Graph Compiler
- MIOpen, ROCBLAS
- ROCm™ HIP Compiler and tools
- Runtime

AMD Instinct™ GPUs | Radeon™ GPUs

Vitis™ AI Platform

- Optimized Inference Models
- ONNX Runtime | Pytorch | TensorFlow
- Vitis AI Development & Deployment Tools
- Vitis AI ML Graph Compiler
- Vitis ML Libraries
- Vitis™ SW Platform (AIE Compiler, HLS and tools)
- Runtime
- CNN Overlay | Transformer Overlay | ...

Versal® Adaptive SoCs | Zynq® Adaptive SoCs

AMD UNIFIED AI STACK 1.0

Unified Inference Frontend (UIF) for AI Developers

Optimized Inference Models

WinML | ONNX Runtime | Pytorch | TensorFlow

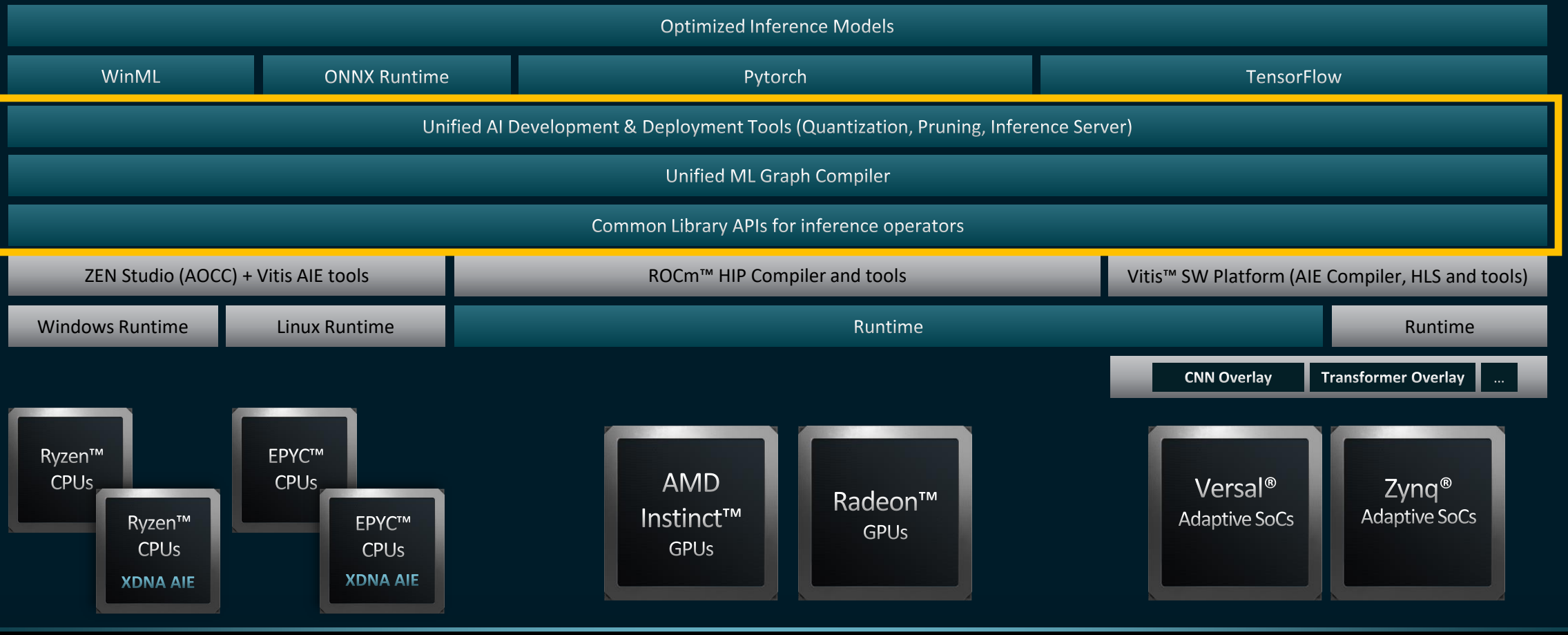
Unified AI Development & Deployment Tools (Quantization, Pruning, Inference Server)

ML Graph Compiler	MIGraphX: ML Graph Compiler	Vitis AI ML Graph Compiler
ZenDNN, AOCL Optimized Library	MIOpen, ROCBLAS	Vitis ML Libraries
ZEN Studio (AOCC)	ROCm™ HIP Compiler and tools	Vitis™ SW Platform (AIE Compiler, HLS and tools)
Windows Runtime / Linux Runtime	Runtime	Runtime
		CNN Overlay / Transformer Overlay / ...

Ryzen™ CPUs	EPYC™ CPUs	AMD Instinct™ GPUs	Radeon™ GPUs	Ryzen™ CPUs XDNA AIE	Versal® Adaptive SoCs	Zynq® Adaptive SoCs

AMD UNIFIED AI STACK 2.0

Seamless Workload Partitioning with UIF, Graph Compiler and Library APIs

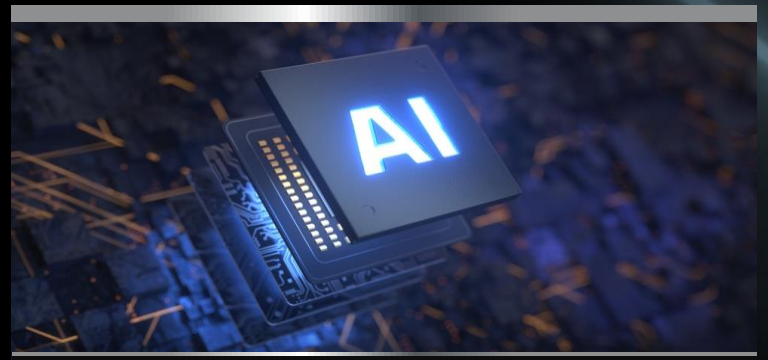


AMD
together we advance

AI INNOVATION

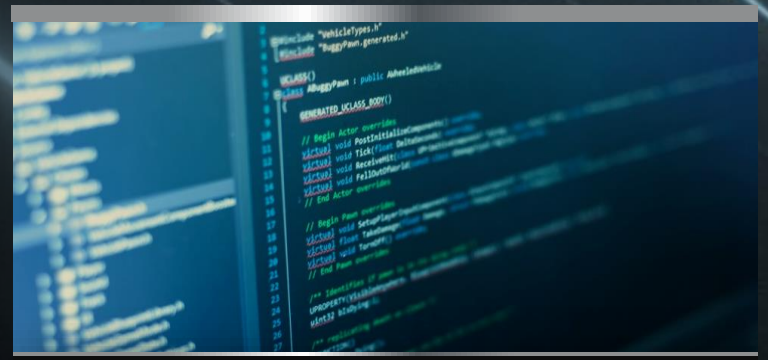
Large Model Training

Exponential growth in
generative model size



Sustainable Inference

OPEX and CAPEX
optimized deployment



Data Science First

Common environment for
CPU, GPU and FPGA

SUMMARY

TAM Expansion

AI, Data Center, Automotive and Embedded

Adaptive Computing

XDNA Architecture: Best of ASICs and FPGAs

Pervasive AI

Unified AMD Hardware and Software Roadmap