



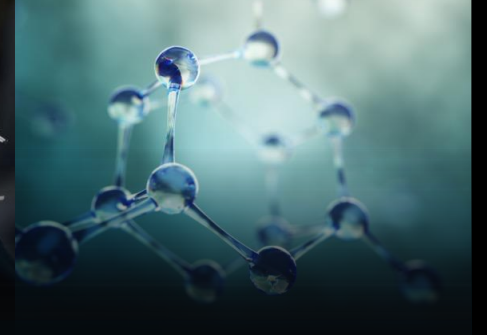
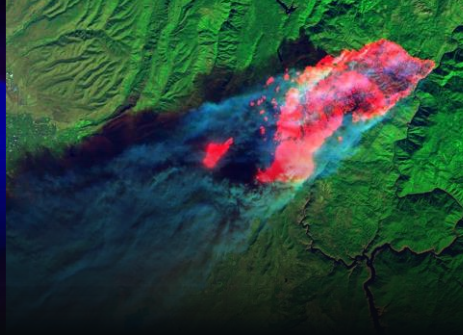
AMD ALVEO V70 AI 推理加速卡, 促進 AI 的未來

***ADVANCING THE FUTURE OF AI AT AMD WITH
THE ALVEO V70 AI INFERENCE ACCELERATOR
CARD***

Jason Chen
Technical Sales Manager

AMD 
together we advance_

AI IS THE DEFINING MEGATREND IN TECHNOLOGY



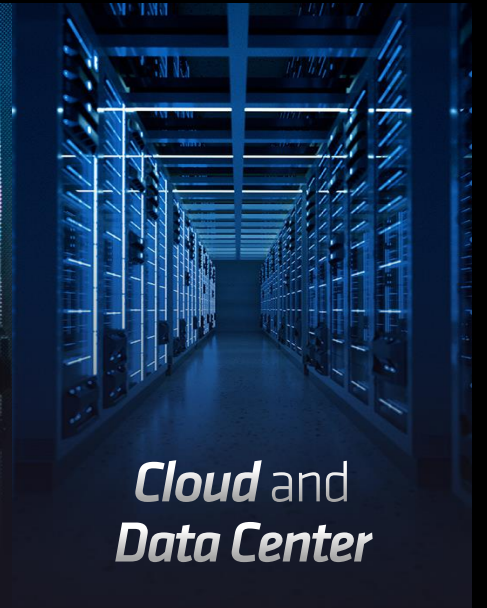
Unlocking the potential of AI
requires AI engines everywhere



Endpoint



Edge



*Cloud and
Data Center*

PERVASIVE AI

Commercial & Enterprise

EPYC™, Ryzen™ PRO CPUs

- Server
- PC
- Workstation

Cloud Data Center

EPYC CPUs, AMD Instinct™ GPUs

- AI Training
- AI Inference

Digital Home

Ryzen CPUs, Radeon™ GPUs

- PCs and Consoles
- Metaverse

PERVASIVE AI

Commercial & Enterprise

EPYC™, Ryzen™ PRO CPUs

- Server
- PC
- Workstation
- Edge

Smart Retail

Zynq™ Adaptive SoCs

- Cashier-less payment
- Customer Reidentification
- Inventory Management

Intelligent Factory

Zynq Adaptive SoCs

- Machine Vision
- Predictive Maintenance
- AI Robotics

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 CPUs, Radeon™ GPUs,
Zynq Adaptive SoCs

- PCs and Consoles
- Metaverse
- Smart Home

Healthcare & Life Sciences

Zynq Adaptive SoCs

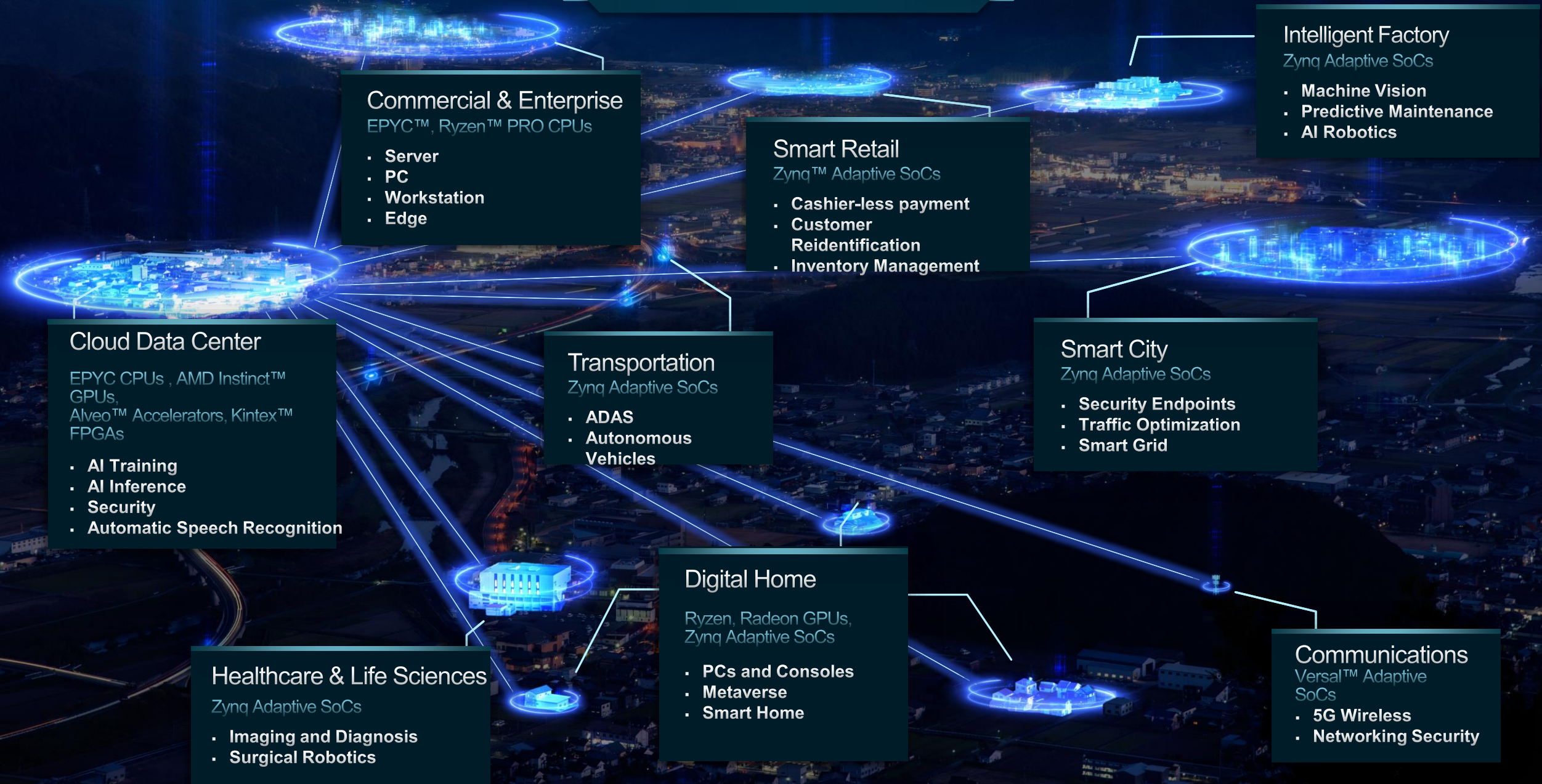
- Imaging and Diagnosis
- Surgical Robotics

Communications

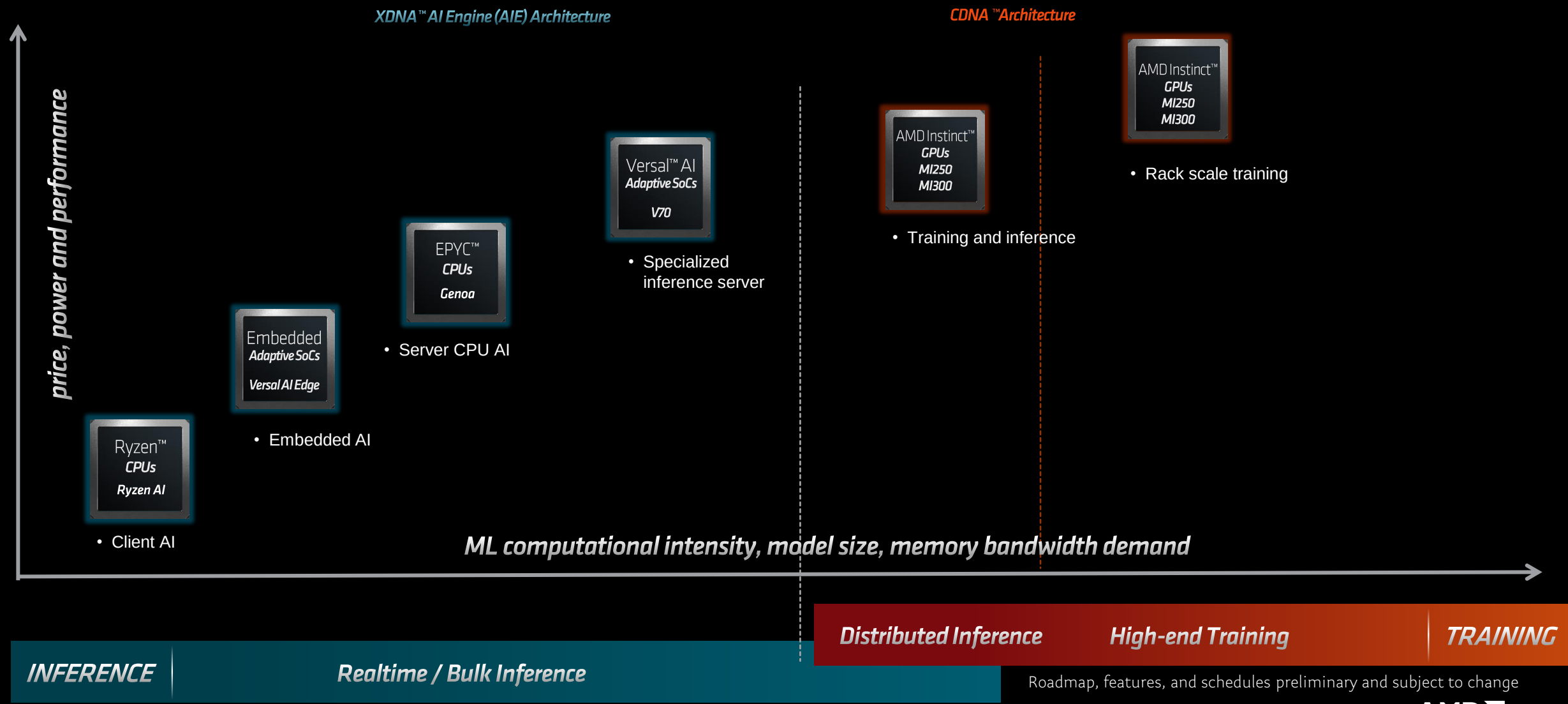
Versal™ Adaptive SoCs

- 5G Wireless
- Networking Security

PERVASIVE AI



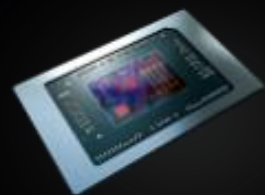
AMD AI PRODUCTS FROM CLIENT TO EMBEDDED TO CLOUD



Roadmap, features, and schedules preliminary and subject to change

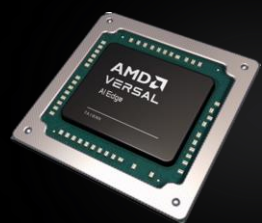
RECENT AMD ANNOUNCEMENTS

FROM CLIENT TO CLOUD



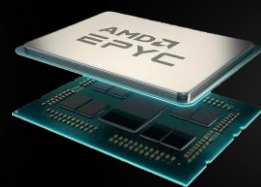
AMD Ryzen™ 7040
Mobile processors
with AI accelerator

First x86 with
integrated AI
accelerator



AMD Embedded
Versal™ AI Edge
Zynq™ MPSoC

AI + Sensor
Fusion for
Embedded



AMD
EPYC™
Processors

Leadership
Server
Solution



AMD
Alveo™
Accelerators

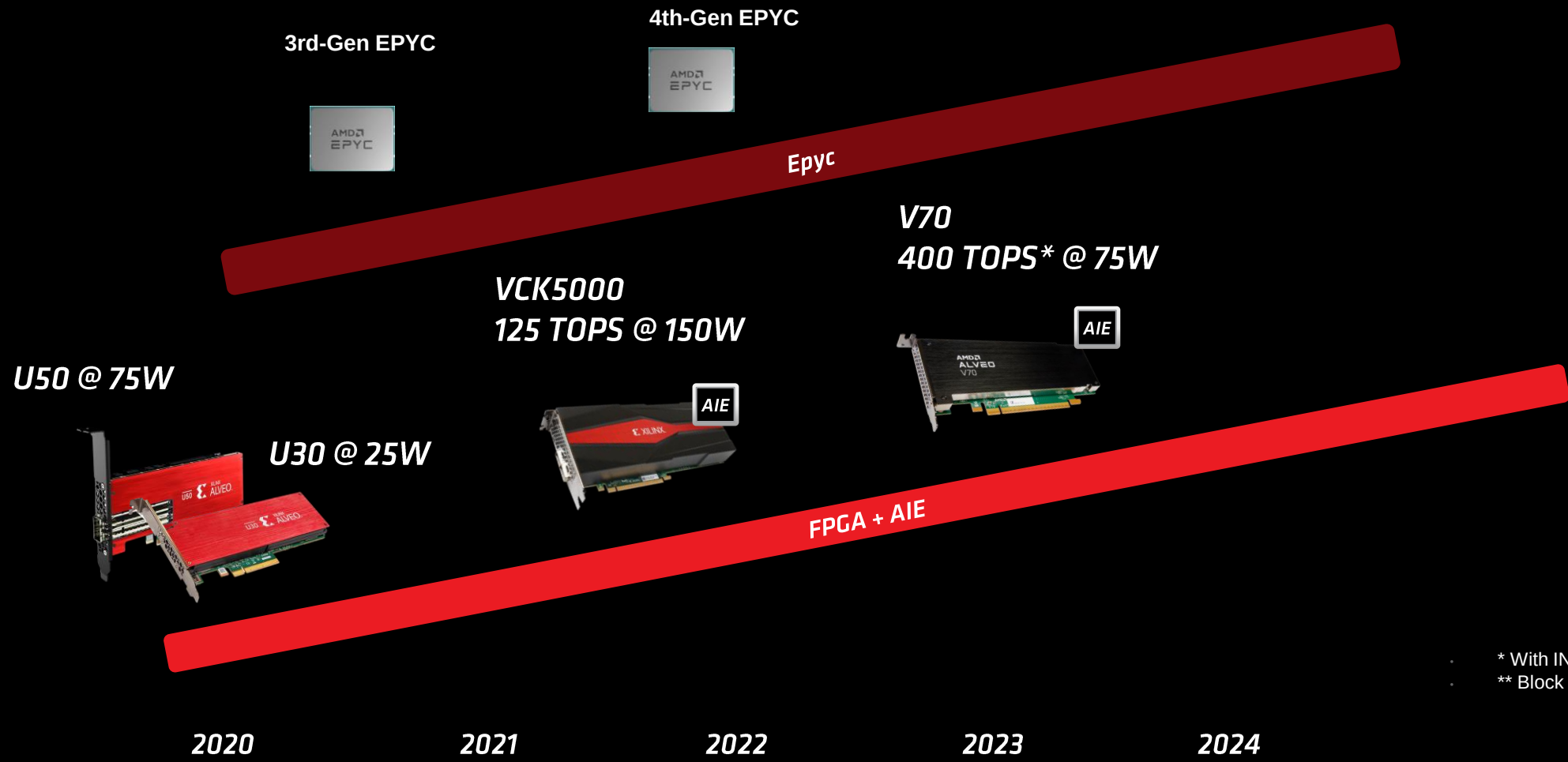
AI Inference
Optimized
Solutions



AMD Instinct™
Accelerators

Data Center
HPC and AI
Solutions

DATA CENTER XDNA™ AI INFERENCE ROADMAP



* With INT8 + sparsity
** Block FP13 datatype



AMD ALVEO™ V70

Designed for Efficient AI inference acceleration

AMD
XDNA
AI Engine

400 TOPS
of AI compute

PCIe® 5.0

75W TDP

FEATURE	DETAILS
Part #	A-V70-P16G-ES3-G
Form Factor	HHHL
Memory Bandwidth	47600 GB/s for internal memory 76.8 GB/s for external DDR
High Density Video Decoder	96 channels of 1920x1080p H.264/H.265

* Using 50% weight sparsity

** 1920x1080p10 H.264/H.265



**Gigabyte G293-Z43
AI Inference Server**

**16 Units
AMD ALVEO™ V70
Accelerator Cards**

High Density of AI Inference Server

FOCUS APPLICATIONS

VIDEO ANALYTICS

- ❑ Face Recognition
- ❑ Object Detection
- ❑ Traffic Monitoring
- ❑ License plate Recognition

VIDEO CONTENT ANALYSIS

- ❑ Short Video Uploads
- ❑ Video Censoring

NATURAL LANGUAGE PROCESSING

- ❑ Chatbots
- ❑ Text Processing
- ❑ Search Engine
- ❑ Voice UI
- ❑ Realtime Automatic Speech Recognition (ASR)

VIDEO ANALYTICS

Lower your cost of operations with V70

❑ *Very competitive cost per channel*

- ❑ *Minimum 50% reduction**
- ❑ *Lower cost servers*
- ❑ *CAPEX & OPEX reduction*

❑ *High channel density per card*

- ❑ *Up to 96x1080p10***

❑ *High performance at low power*

- ❑ *Deterministic performance*
- ❑ *Very low latency jitter*
- ❑ *Max 75W power*

❑ *Frictionless Migration*

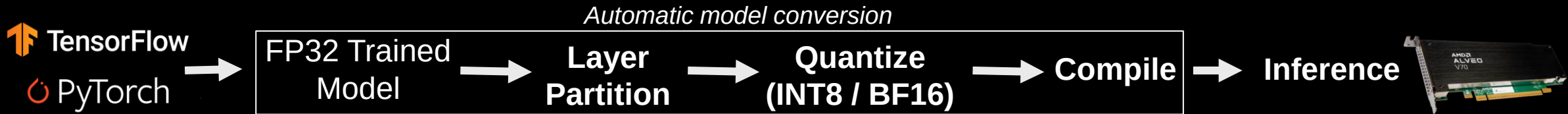
- ❑ *Easy 3 step migration*



* Over previous generation devices

** 1920x1080p10 H.264/H.265

MIGRATING AI MODELS FROM NVIDIA GPUS



```
1 import torch
2 import torchvision
3 import numpy as np
4 import cv2
5 import sys
6 from torchvision.models import resnet50, ResNet50_Weights
7
8 weights = ResNet50_Weights.DEFAULT
9
10 model = torchvision.models.resnet50(weights = weights).eval().cuda()
11 model.load_state_dict(torch.load('./resnet50.pth'))
12
13 img = cv2.imread('./cat.jpg')
14 img = cv2.resize(img, (224, 224))
15 mean = np.array([0.406, 0.456, 0.485]).astype(np.float32)
16 std = np.array([0.225, 0.224, 0.229]).astype(np.float32)
17 img = (img / 255. - mean) / std
18 img = np.expand_dims(img, axis=0)
19 img = np.transpose(img, (0, 3, 1, 2)).astype(np.float32)
20
21 inputimage = torch.from_numpy(img).cuda()
22
23 out = model(inputimage)
24 top_k = int(sys.argv[1])
25 cls_values, cls_indexes = out.topk(top_k, 1, True, True)
26 for cls_ids in cls_indexes.tolist()[0]:
27     category_name = weights.meta["categories"][cls_ids]
28     print(category_name)
```

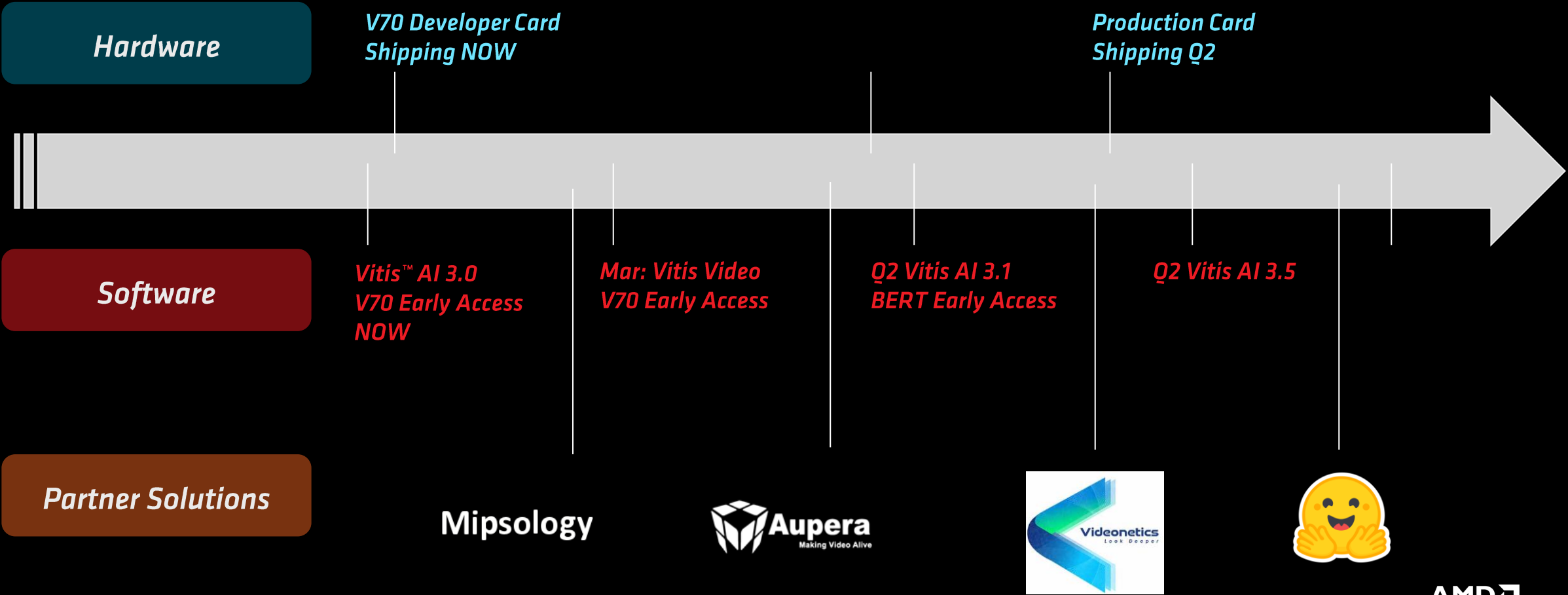
Nvidia GPU

```
1 import torch
2 import torchvision
3 import numpy as np
4 import cv2
5 import sys
6 from torchvision.models import resnet50, ResNet50_Weights
7+ import wego_torch
8
9 weights = ResNet50_Weights.DEFAULT
10
11+ model = torch.jit.load('./resnet50_wego_compiled.wg')
12
13 img = cv2.imread('./cat.jpg')
14 img = cv2.resize(img, (224, 224))
15 mean = np.array([0.406, 0.456, 0.485]).astype(np.float32)
16 std = np.array([0.225, 0.224, 0.229]).astype(np.float32)
17 img = (img / 255. - mean) / std
18 img = np.expand_dims(img, axis=0)
19 img = np.transpose(img, (0, 3, 1, 2)).astype(np.float32)
20
21+ inputimage = torch.from_numpy(img)
22
23 out = model(inputimage)
24 top_k = int(sys.argv[1])
25 cls_values, cls_indexes = out.topk(top_k, 1, True, True)
26 for cls_ids in cls_indexes.tolist()[0]:
27     category_name = weights.meta["categories"][cls_ids]
28     print(category_name)
```

AMD V70

Drop-in Replacement from Nvidia GPUs

Alveo™ V70 ACCELERATOR CARD SCHEDULE



ALVEO™ V70 CAR CLASSIFICATION DEMO

Multi -Stream AI Car Classification with VVAS

AI models	Yolov3 + ResNet18 x 9 (average)
Other workloads	H.264 FHDdecode + CV (crop, resize, color correction)
Performance	224 FPS (16 video feeds @ 14 FPS)
Availability	Contact: V70_solutions@amd.com



ALVEO™ V70 OBJECT DETECTION DEMO

High-Density Real-Time AI Object and People detection with Partner Mipsology

AI models	YoloV5, V6 (directly from internet)
Other workloads	CV (crop, resize, color correction)
Performance	12x 1080p at 30FPS
Availability	<u>Contact: zebra@mipsology.com</u>

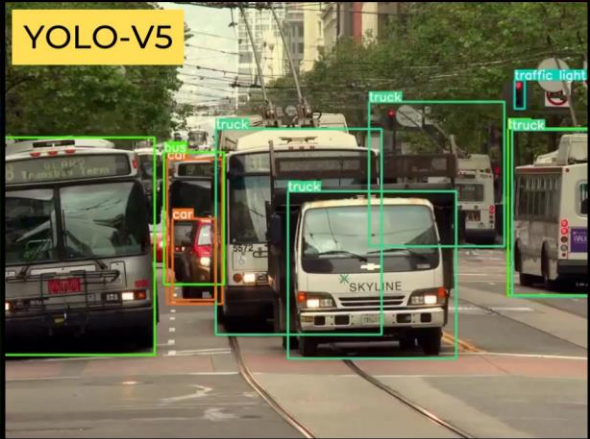
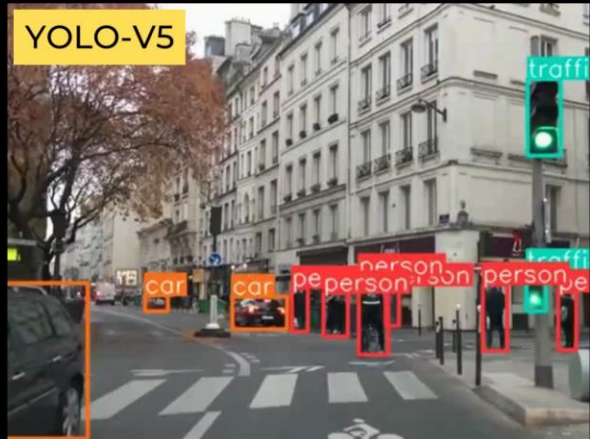
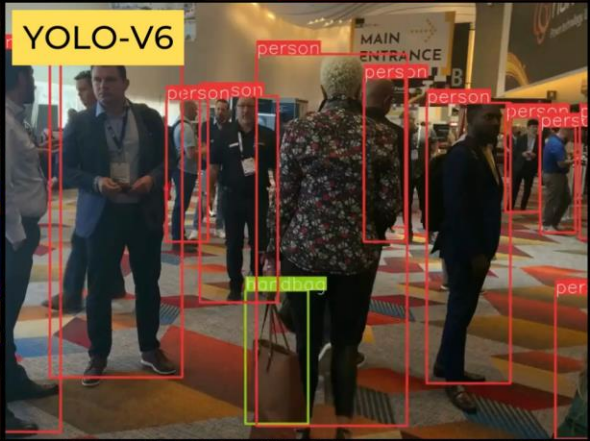




Zebra®
by Mipsology

**Plug and Play
AI Inference Acceleration**

Mipsology © 2022



Collateral

Available Demos

[Product Page](#)

[EA Site](#)

[Contact V70 Team](#)

Multi -Stream AI Car Classification with VVAS	
AI models	Yolov3 + ResNet18 x 9 (average)
Other workloads	H.264 FHDdecode + CV (crop, resize, color correction)
Performance	224 FPS (16 video feeds @ 14 FPS)
Availability	Contact: V70_solutions@amd.com

High-Density Real-Time AI People/Object Detection with Partner Mipsology	
AI models	YoloV5, V6 (directly from internet)
Other workloads	CV (crop, resize, color correction)
Performance	12x 1080p at 30FPS

Product Brief

PRODUCT BRIEF

ALVEO™ V70 AI INFERENCE ACCELERATOR CARD

OVERVIEW

With intelligent connected devices in our homes, cars, offices, factories, cities, and in the cloud, the cost for this proliferation of AI enabled applications is an exponential increase in the data-processing and energy efficiency requirements placed on the chips powering these devices. The challenge is not just how to deploy the AI model, but how to deploy the AI application most efficiently. The best implementation of an AI application doesn't need to be the fastest, it needs to be the most efficient, yet remain flexible. AMD XDNA architecture built on adaptive computing platforms provide the best of both worlds for AI inference workloads - from cloud, edge or endpoint.

The Alveo™ V70 accelerator card is the first AMD Alveo production card leveraging AMD XDNA™ architecture with AI Engines providing a tightly integrated heterogeneous compute platform for CNN, RNN, and NLP acceleration targeting cloud and edge applications. V70 is designed to be the most energy efficient AI inference card in the AMD portfolio tuned for video analytics and natural language processing workloads and offers industry standard framework support, directly compiling models trained in TensorFlow and PyTorch. The card is a PCIe™ based half height, half length, single slot card that supports passive cooling for closed-loop thermal control in the server PCIe expansion slot. The card is equipped with a 7nm Versal™ ACAP device which has an integrated AI Engine core to complement adaptable and scalar engines and 16 GB of DDR4 memory. Providing low power and a low-profile form factor, the V70 helps reduce cost per AI channel and provides high channel density for video applications.



TARGET APPLICATIONS

KEY VALUE PROPS

- Helps reduced cost per AI channel
- Entire pipeline execution on accelerator
- High channel density for video analytics applications

TARGET USERS

- Data Scientist
- AI Developers
- Software Developers
- Data Center Infrastructure Engineers

USE CASES

- Video Content Analysis
- Natural Language Processing
- Smart City/ Smart Retail
- Recommendation Ranking
- Medical Imaging

AMD

READY TO CONNECT? VISIT www.xilinx.com/products/specialty-aip/v70/guest.html

1

