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

PERVASIVE AI

Commercial & Enterprise

EPYC™, Ryzen™ PRO CPUs

- **Server**
- **PC**
- **Workstation**

Cloud Data Center

EPYC CPUs , AMD Instinct™ GPUs

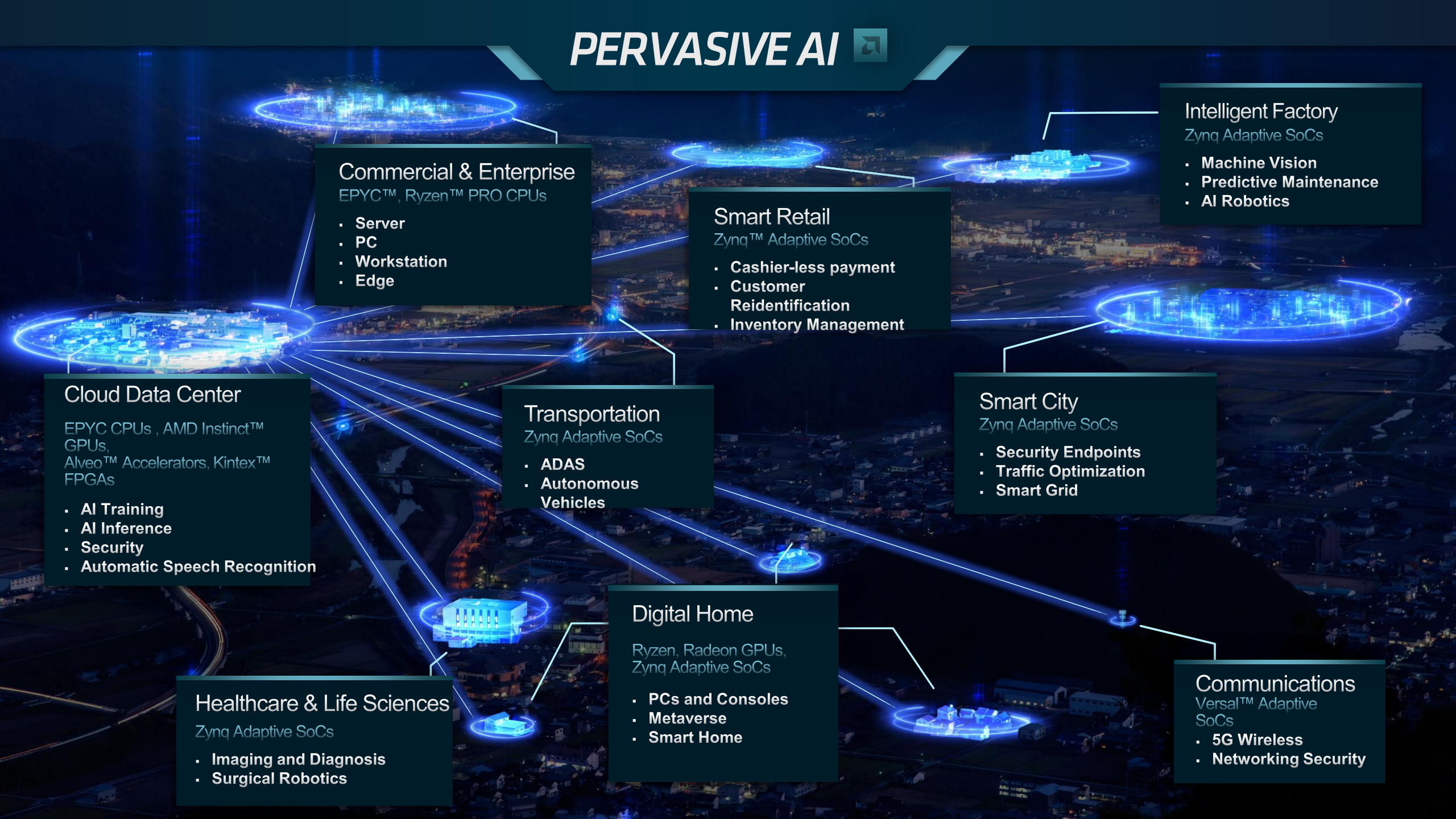
- **AI Training**
- **AI Inference**

Digital Home

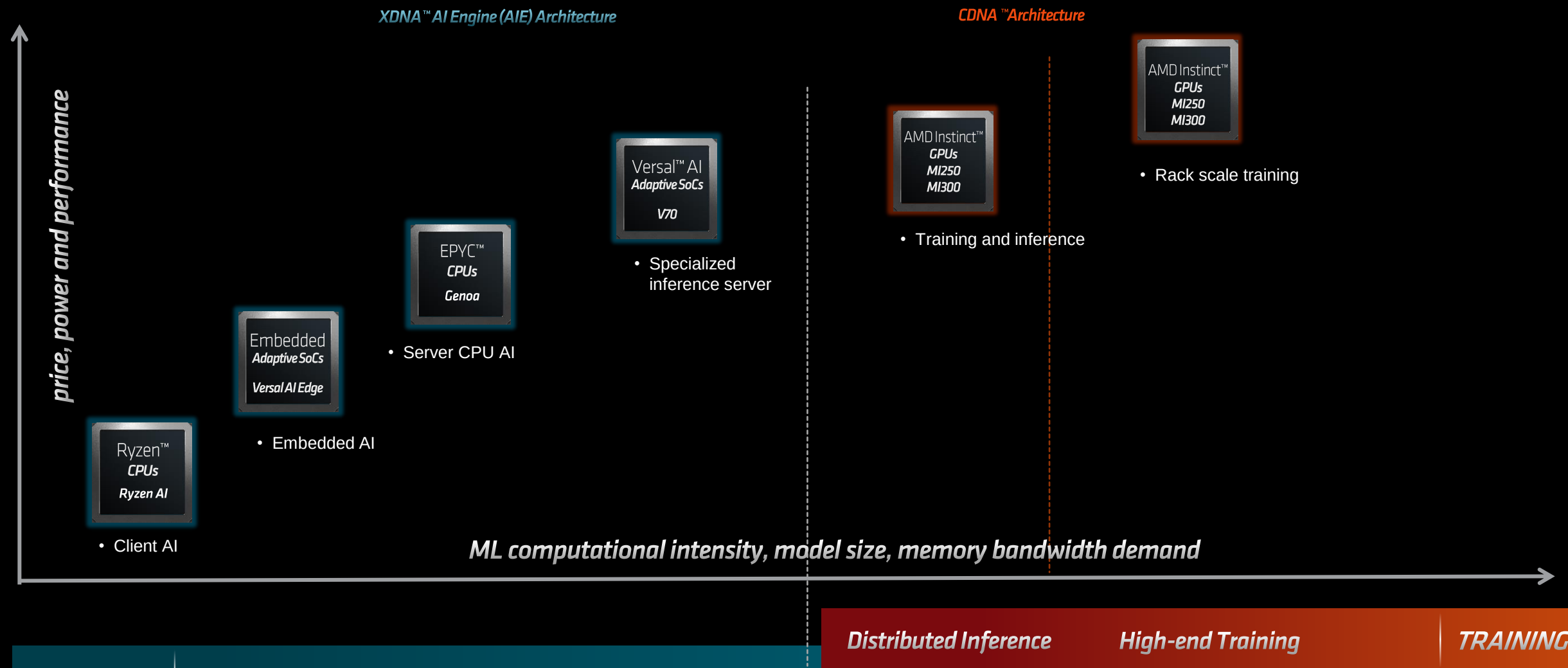
Ryzen CPUs, Radeon™ GPUs

- **PCs and Consoles**
- **Metaverse**


PERVASIVE AI
Commercial & Enterprise
EPYC™, Ryzen™ PRO CPUs
Server
PC
Workstation
Edge
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 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
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

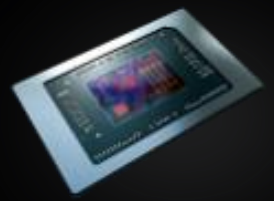
AMD AI PRODUCTS FROM CLIENT TO EMBEDDED TO CLOUD



Roadmap, features, and schedules preliminary and subject to change

RECENT AMD ANNOUNCEMENTS

FROM CLIENT TO CLOUD

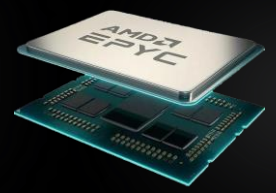


AMD Ryzen™ 7040 Mobile processors with AI accelerator

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

3rd-Gen EPYC

4th-Gen EPYC

Epyc

U50 @ 75W

U30 @ 25W

VCK5000
125 TOPS @ 150W

AIE

V70
400 TOPS* @ 75W

AIE

FPGA + AIE

2020 2021 2022 2023 2024

* 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

TensorFlow
PyTorch

Automatic model conversion

FP32 Trained Model → **Layer Partition** → **Quantize (INT8 / BF16)** → **Compile** → **Inference**

Nvidia GPU

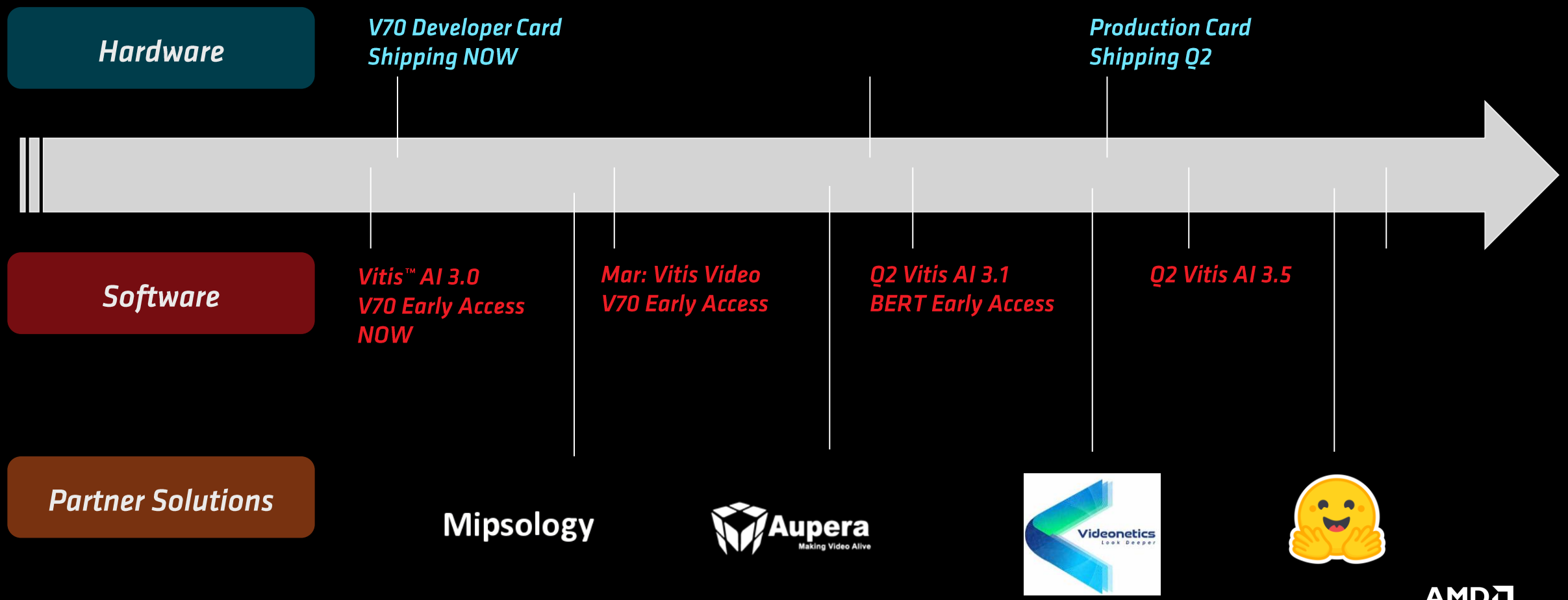
```python
import torch
import torchvision
import numpy as np
import cv2
import sys
from torchvision.models import resnet50, ResNet50_Weights

weights = ResNet50_Weights.DEFAULT

model = torchvision.models.resnet50(weights = weights).eval().cuda()
model.load_state_dict(torch.load('./resnet50.pth'))

img = cv2.imread('./cat.jpg')
img = cv2.resize(img, (224, 224))
mean = np.array([0.406, 0.456, 0.485]).astype(np.float32)
std = np.array([0.225, 0.224, 0.229]).astype(np. float32)
img = (img / 255. - mean) / std
img = np.expand_dims(img, axis=0)
img = np.transpose(img, (0, 3, 1, 2)).astype(np. float32)

inputimage = torch.from_numpy(img).cuda()

out = model(inputimage)
top_k = int(sys.argv[1])
cls_values, cls_indexes = out.topk(top_k, 1, True, True)
for cls_ids in cls_indexes.tolist()[0]:
    category_name = weights.meta["categories"][cls_ids]
    print(category_name)
```

AMD V70

```python
import torch
import torchvision
import numpy as np
import cv2
import sys
from torchvision.models import resnet50, ResNet50_Weights
import wego_torch

weights = ResNet50_Weights.DEFAULT

model = torch.jit.load('./resnet50_wego_compiled.wg')

img = cv2.imread('./cat.jpg')
img = cv2.resize(img, (224, 224))
mean = np.array([0.406, 0.456, 0.485]).astype(np.float32)
std = np.array([0.225, 0.224, 0.229]).astype(np. float32)
img = (img / 255. - mean) / std
img = np.expand_dims(img, axis=0)
img = np.transpose(img, (0, 3, 1, 2)).astype(np. float32)

inputimage = torch.from_numpy(img)

out = model(inputimage)
top_k = int(sys.argv[1])
cls_values, cls_indexes = out.topk(top_k, 1, True, True)
for cls_ids in cls_indexes.tolist()[0]:
    category_name = weights.meta["categories"][cls_ids]
    print(category_name)
```

Drop-in Replacement from Nvidia GPUs

Alveo™ V70 ACCELERATOR CARD SCHEDULE



ALVEO™ V70 CAR CLASSIFICATION DEMO

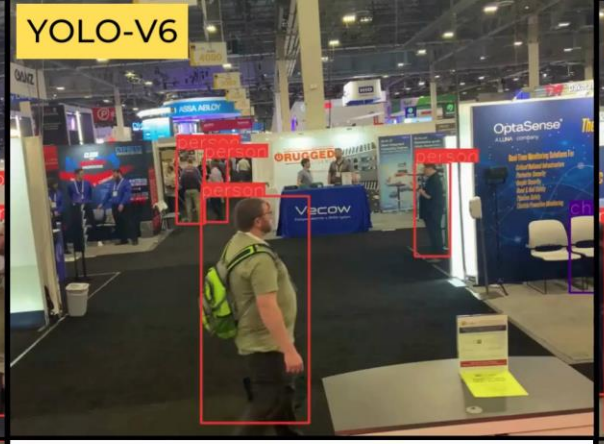
Multi -Stream AI Car Classification with VVAS

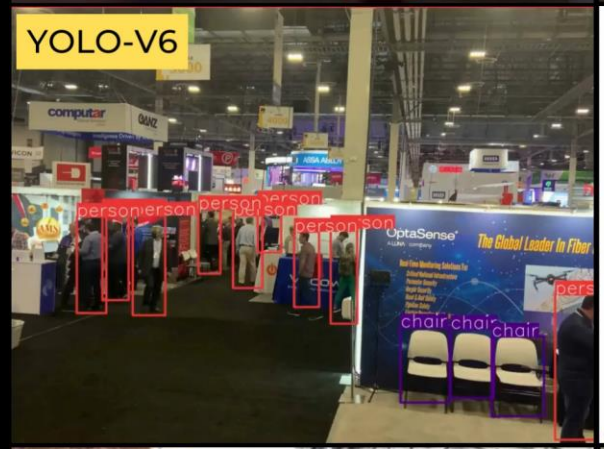
AI models	Yolov3 + ResNet18 x 9 (average)
Other workloads	H.264 FHDdecode + CV (crop, resize, color correction)
Performance	224 FPS (16 video feeds @ 14 FPS)
Availability	Contact: V70_solutions@amd.com

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	Contact: zebra@mipsology.com



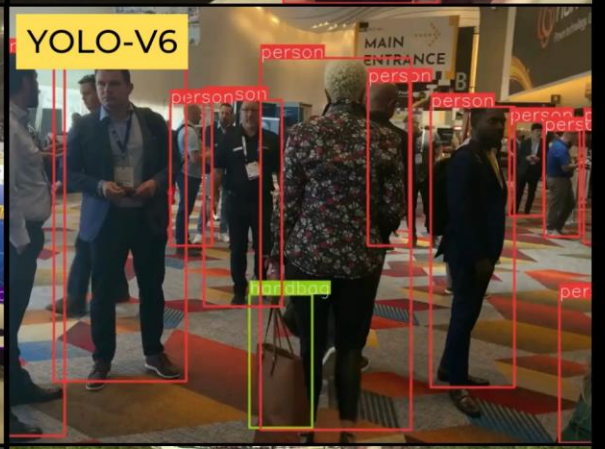


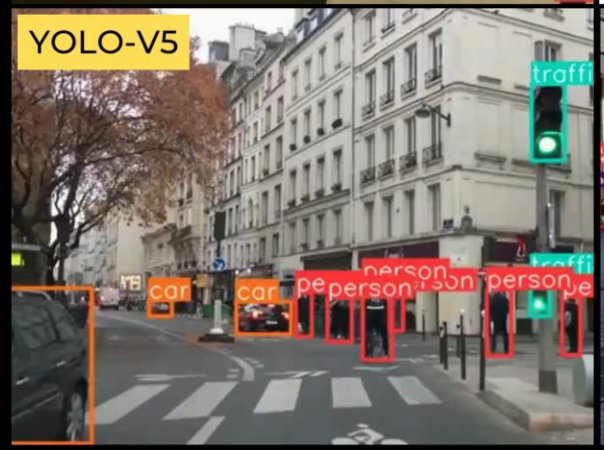
Zebra®

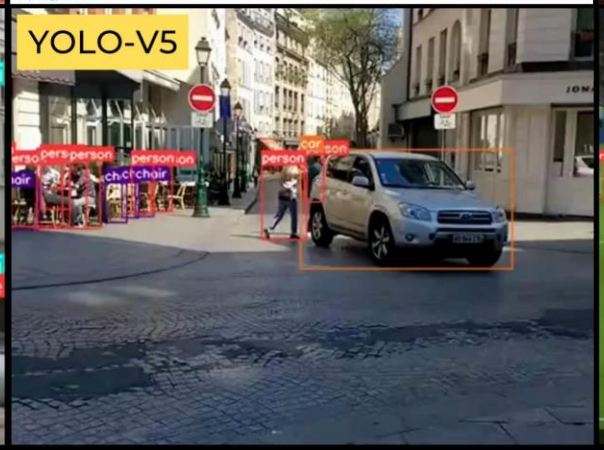
by Mipsology

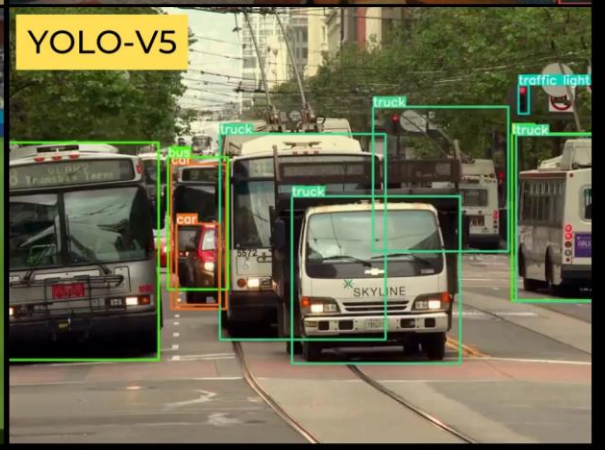
Plug and Play

AI Inference Acceleration









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


AMD