Low Latency AI Inference Acceleration with Mipsology Zebra and Xilinx Alveo

Andy Luo

AI Product Maketing

XILINX

AI Applications Powered by Xilinx

AI Proliferation

Data Center

5G

Autonomous Driving

Security

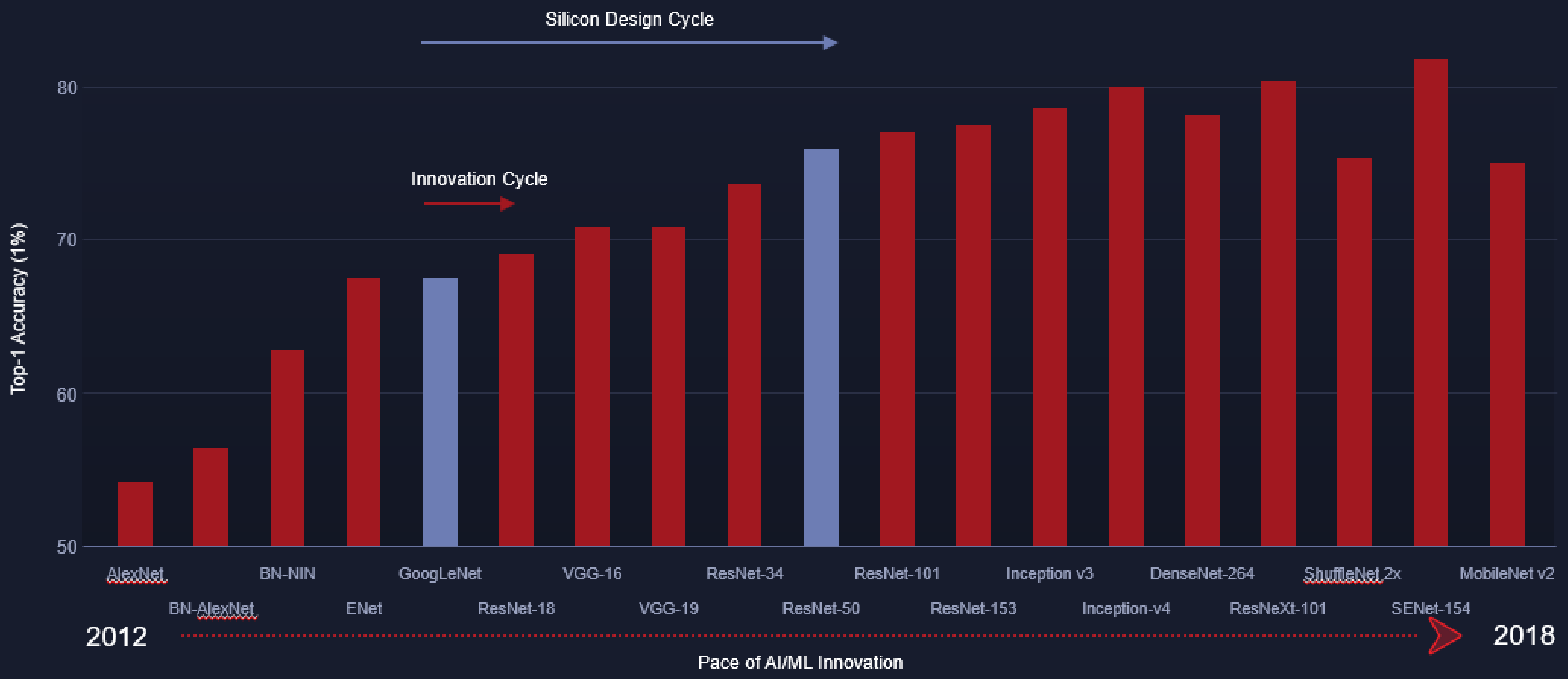
Genomics

Video Analytics

Healthcare

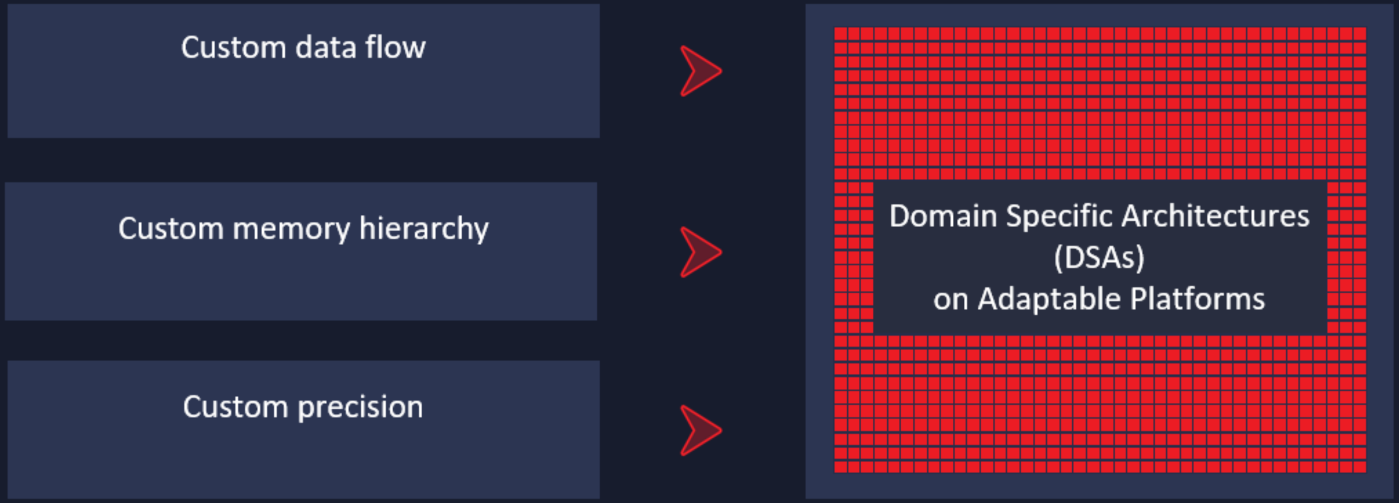
Finance

AI is Evolving Rapidly



XILINX.

Adapt to Your AI Workloads

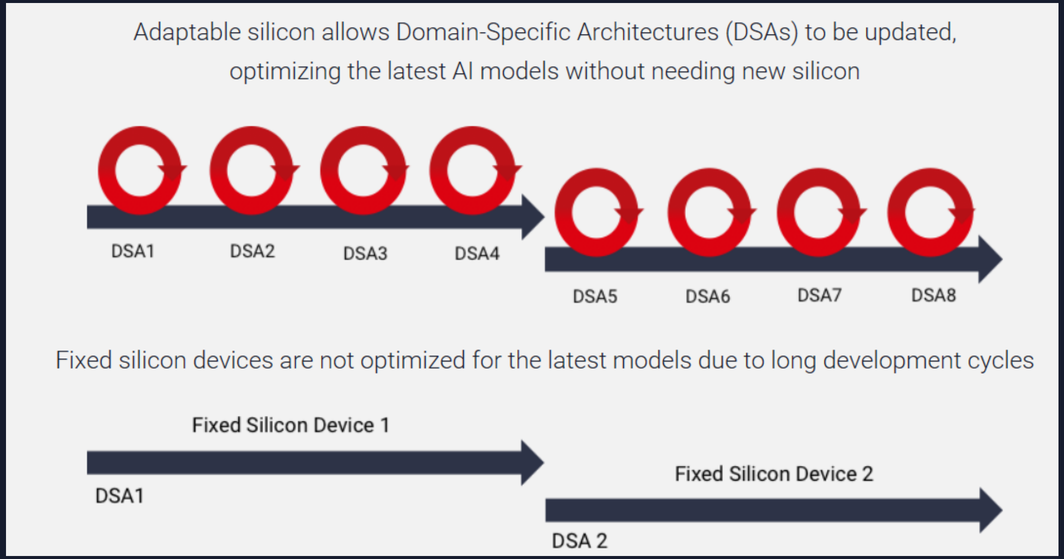


Custom data flow

Custom memory hierarchy

Custom precision

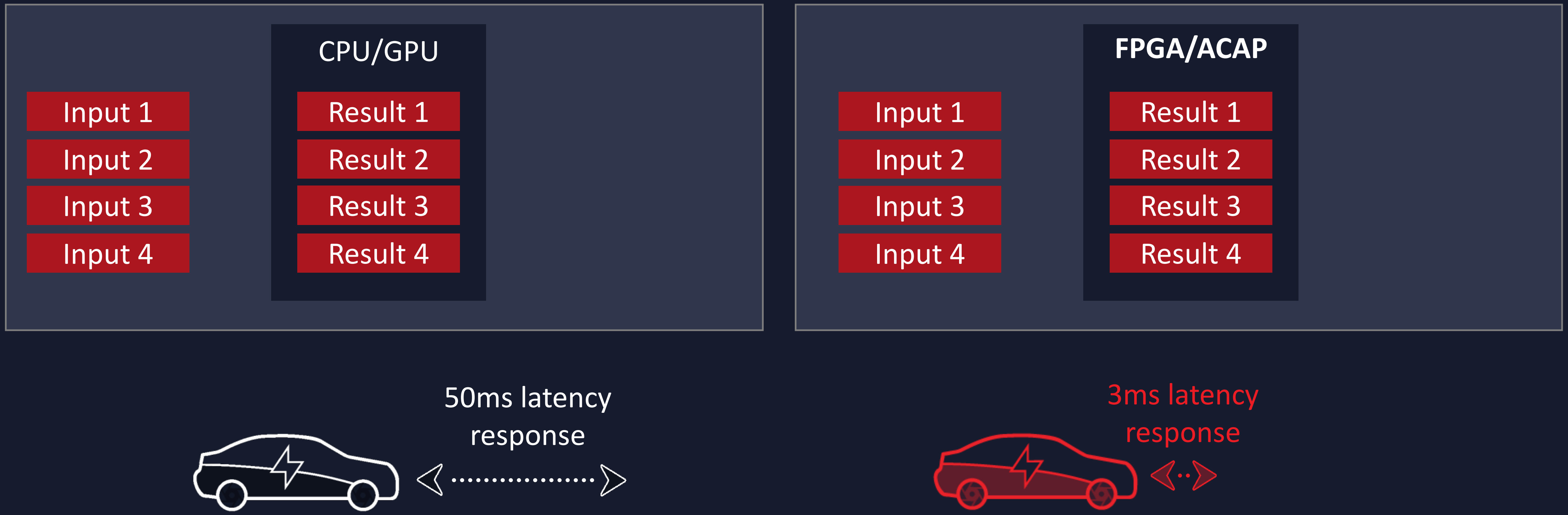
Domain Specific Architectures (DSAs) on Adaptable Platforms



Adaptable silicon allows Domain-Specific Architectures (DSAs) to be updated, optimizing the latest AI models without needing new silicon

Fixed silicon devices are not optimized for the latest models due to long development cycles

Low Latency AI Inference



CPU/GPU

Input 1

Input 2

Input 3

Input 4

Result 1

Result 2

Result 3

Result 4

50ms latency response

FPGA/ACAP

Input 1

Input 2

Input 3

Input 4

Result 1

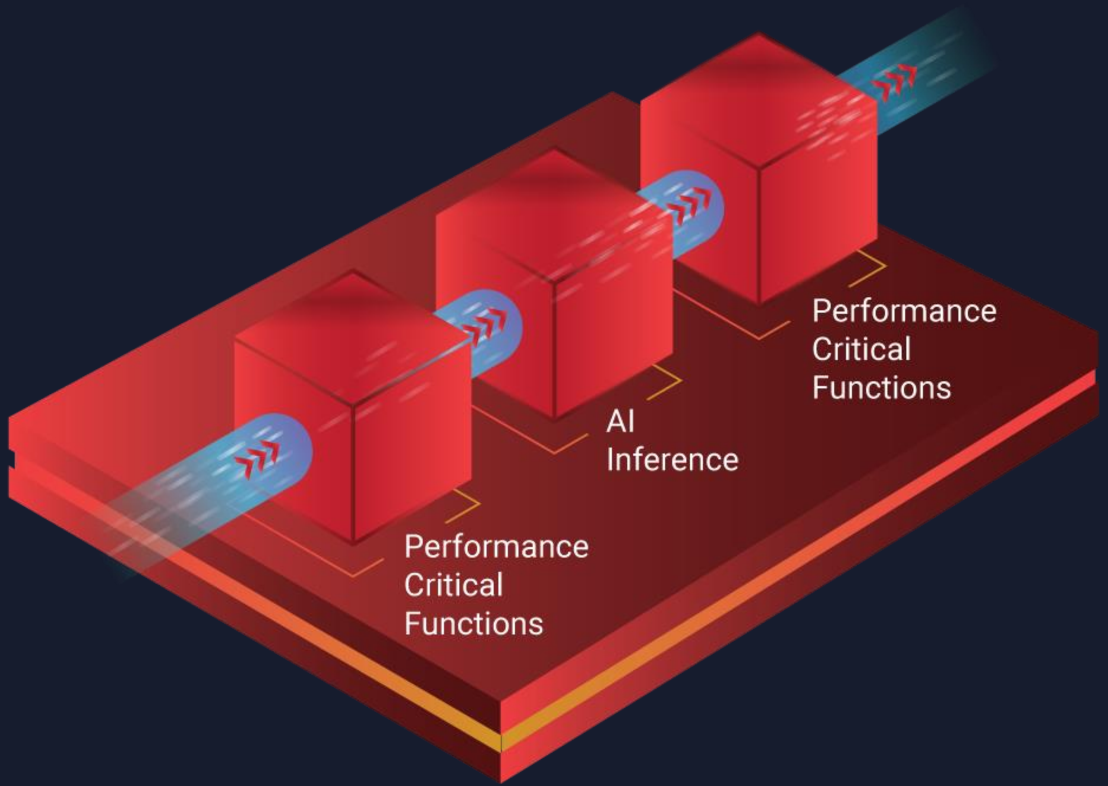
Result 2

Result 3

Result 4

3ms latency response

Adapt to Your AI Application

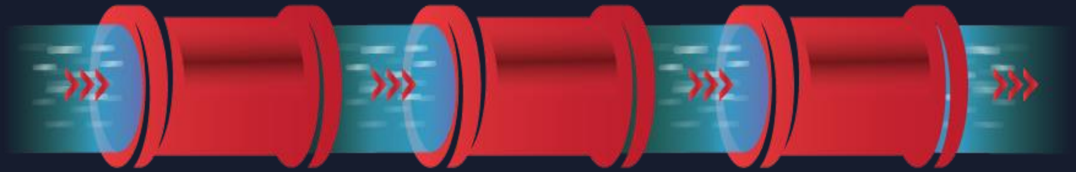


Performance Critical Functions

AI Inference

Performance Critical Functions

Xilinx – Matched Throughput

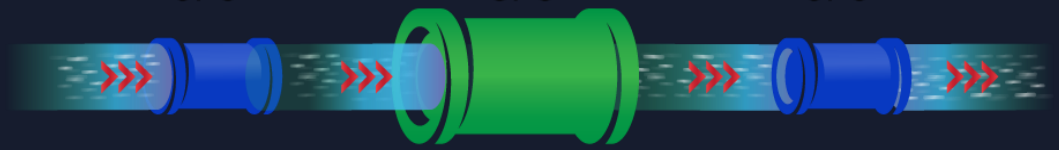


Performance Critical Functions

AI Inference

Performance Critical Functions

GPU & CPU – Mismatched Throughput



Performance Critical Functions

AI Inference

Performance Critical Functions

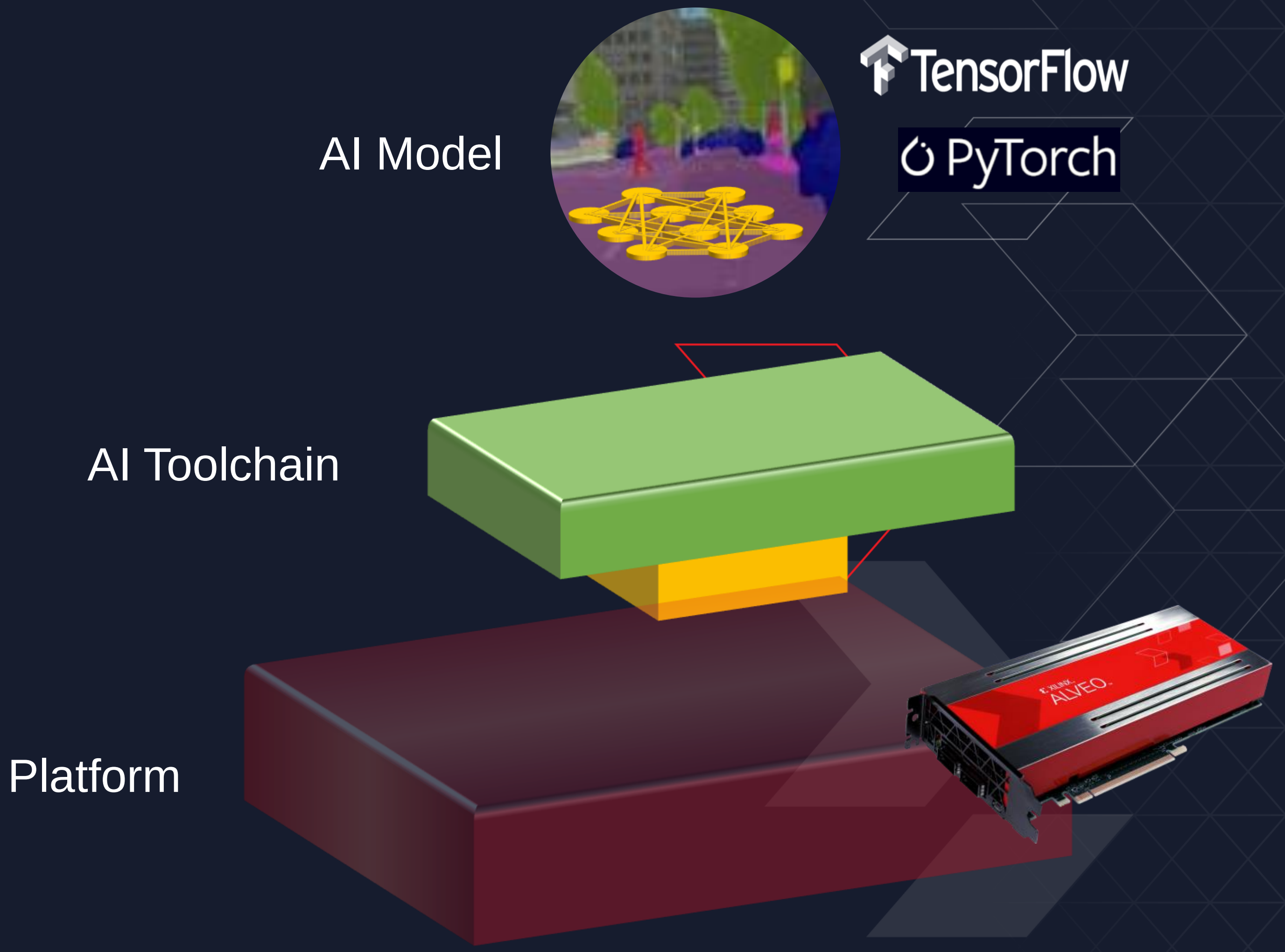
How to Deploy AI Inference to Xilinx Platform

AI DSA

Direct Framework Compilation

Minutes of Compile Times

MIN HRS



AI Model

AI Toolchain

Platform

Adaptable.

Intelligent.

XILINX®

Mipsology

Zebra™ by Mipsology

Accelerating computation for machine learning

Presented by: Ludovic (Ludo) Larzul, Founder and CEO, Mipsology

Date: February 25, 2020

We Focus on Acceleration, You Focus on Your Application!

Welcome

Low Latency AI Inference Acceleration with Zebra by Mipsology and Xilinx Alveo

Ludovic Larzul, Mipsology CEO

Robert Lara, Mipsology Sr. Director

Mario Trentini, Mipsology Sr. Director

For more information email us at zebra@mipsology.com

What is Zebra by Mipsology?

Zebra™ software accelerates inference computation faster and easier for machine-learning AI based systems

Zebra works on most Xilinx Alveo Cards

What Makes Zebra Unique

Ease of Use

With a single command and zero change, any engineer can replace CPU or GPU by Zebra without any knowledge of FPGA.

Large Support of Neural Network

With a large support of neural network, its start from the same training, results with similar accuracy, Zebra does not require scarce AI resources to do any change.

Best Performance

Zebra delivers immediately the best performance at low latency from small to large FPGA, matching AI needs from embedded to data centers.

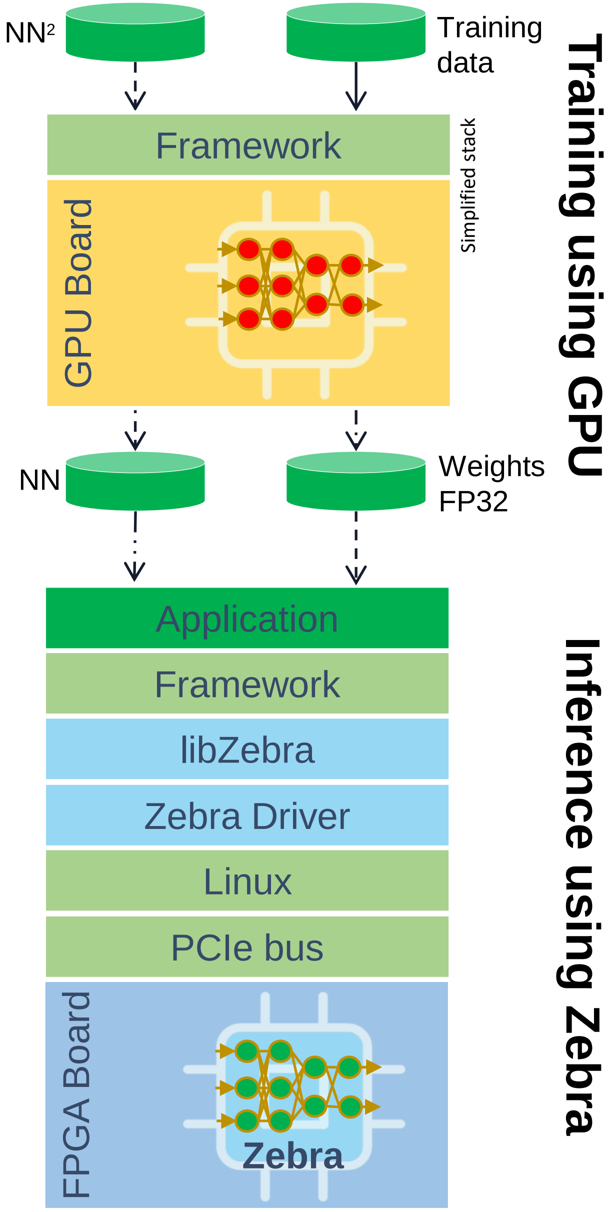
Lower Cost of Ownership

Quick transition for low NRE and large choice of FPGA and cards, with long life span to lower the TCO.

How does Zebra work?

- ✓ Keep your GPU for training
- ✓ Keep your same neural network for GPU/CPU
- ✓ Keep your same application software for GPU/CPU
- ✓ Keep your same framework for GPU/CPU
- ✓ **Type a Single Linux Command**
- ✓ Zebra will do automatic proprietary quantization
- ✓ Experience best performance with low latency

- ✓ Of course, Zebra works on Xilinx Alveo Boards



Training using GPU

NN^2

Training data

Framework

GPU Board

Simplified stack

NN

Weights FP32

Inference using Zebra

Application

Framework

libZebra

Zebra Driver

Linux

PCIe bus

FPGA Board

Zebra

TensorFlow

Caffe

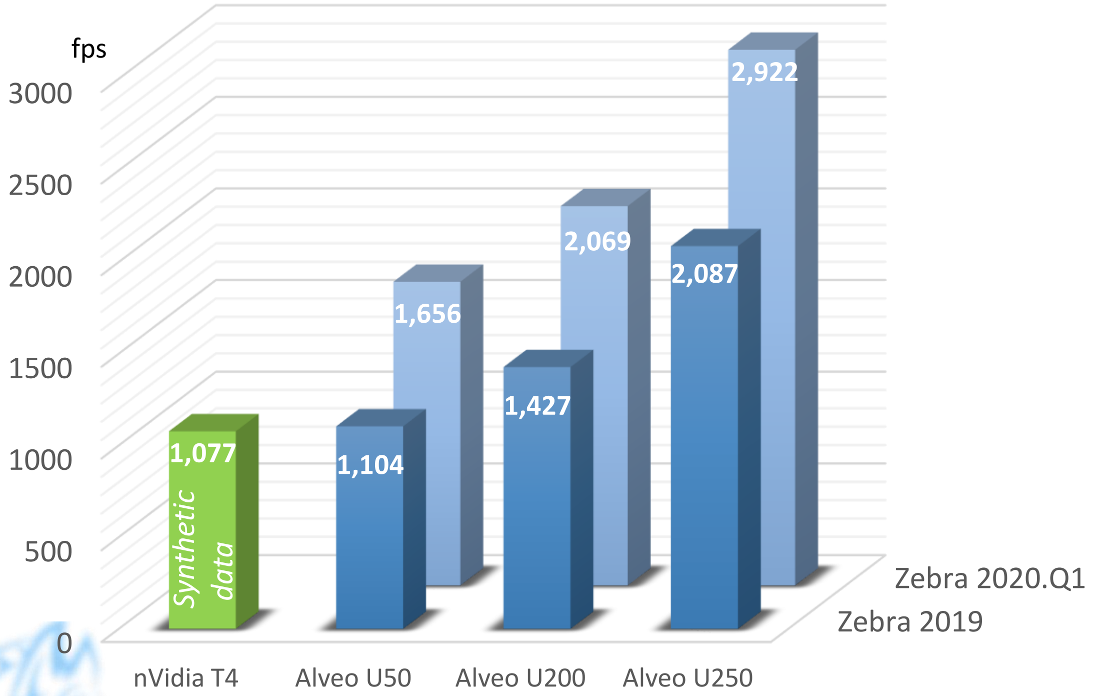
PyTorch

mxnet

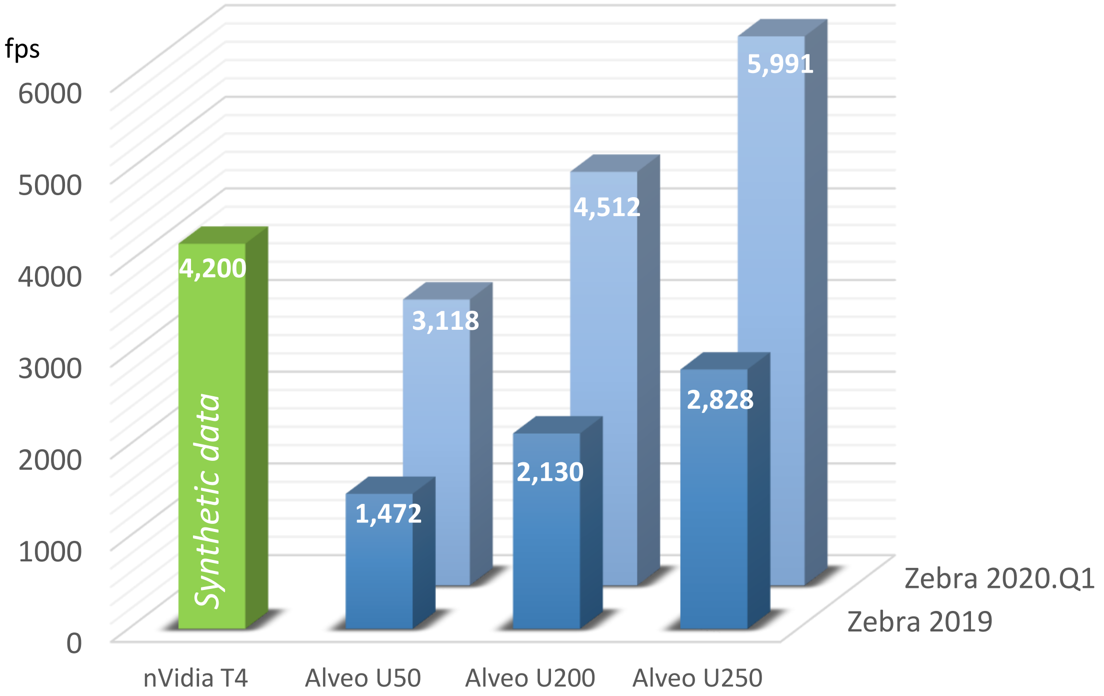
ONNX

Zebra Metrics – Best Performance

ResNet50 Performance on actual data, single-image batch

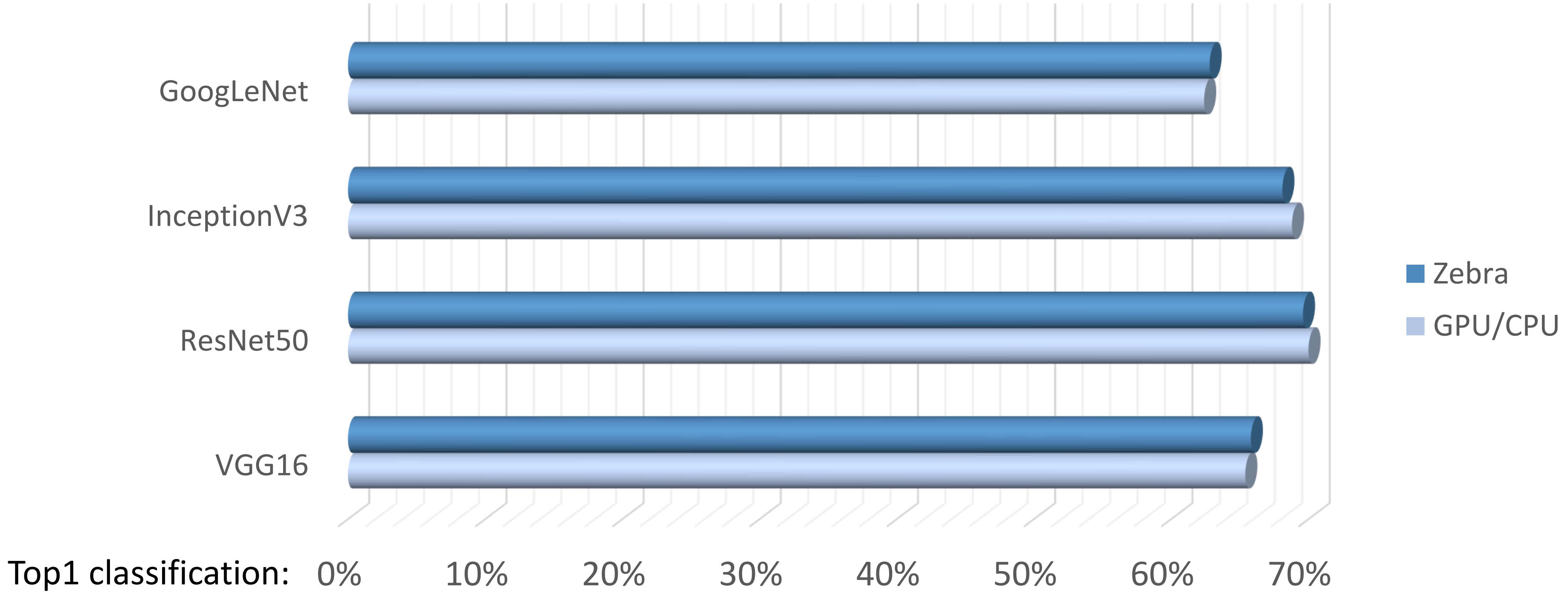


ResNet50 Performance on actual data, multi-image batch



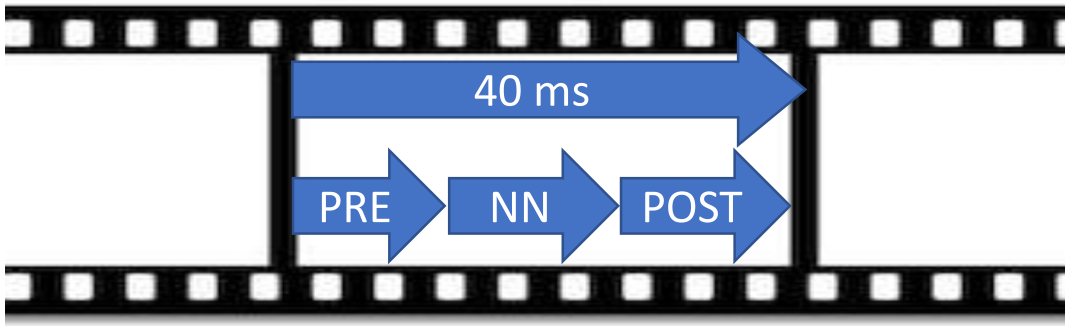
Zebra 2019: batch below 64. Production version. All performance measured.
Zebra 2020.Q1: batch below 64. Alpha version, production end of Q3. Performance measured on Alveo U250, scaled for U200 and U50.
nVidia T4: performance for same latency as Alveo. ResNet50 performance number from nVidia website measured with synthetic data. nVidia does not publish Yolo performance.

Zebra Metrics –
Keeps Accuracy with Proprietary quantization

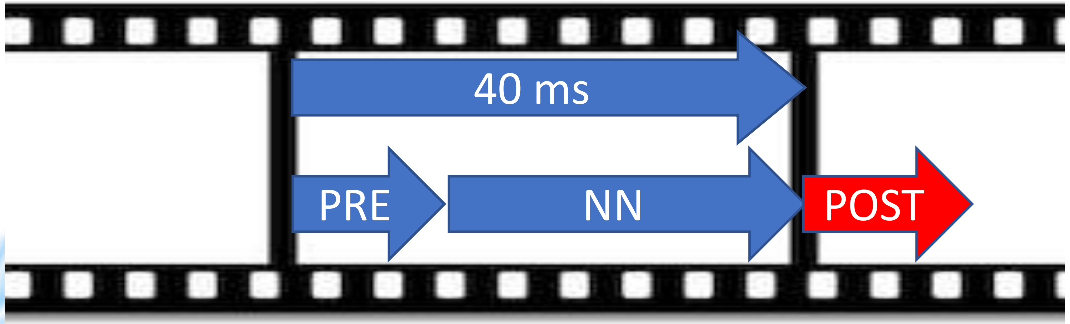


Zebra enables real-time processing of cameras at full speed

At 25 frame per second, a reactive application has 40ms to process an image from a camera:

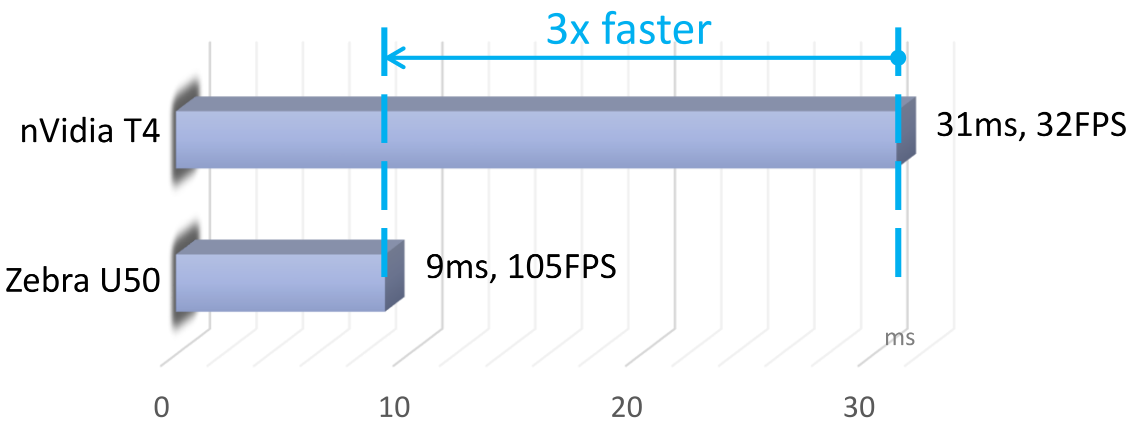


Long NN processing means late actions:



- ➢ A slow robot moving at 12ft/s → 1 inch per 10ms
- ➢ A car moving at 65mph → 8 inches per 10ms

Zebra on Alveo U50 reacts **3x faster than best performing GPU**.



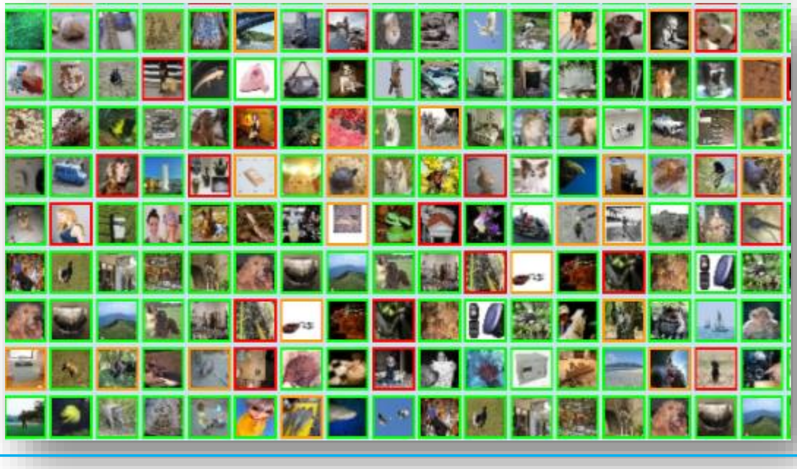
Latency: LOWER IS BETTER

(computing time for an image, based on customer Yolo-class CNN)

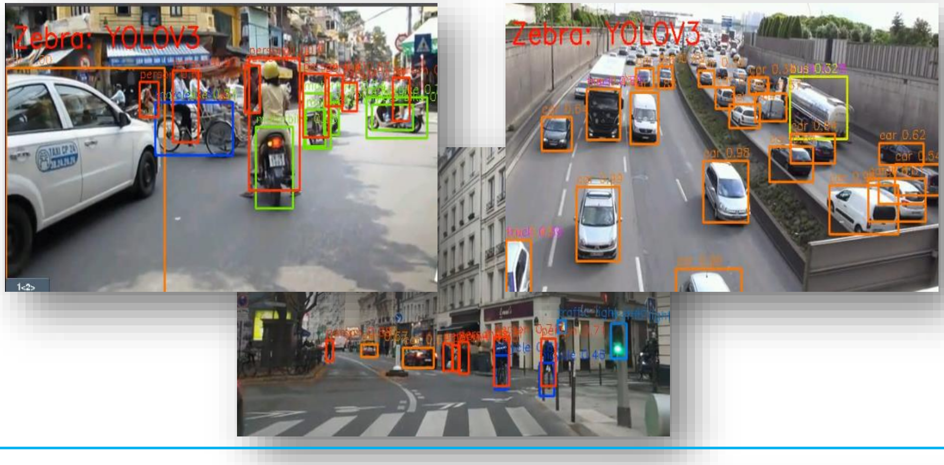
Zebra for All Applications

Tested networks: AlexNet, CaffeNet, GoogLeNet, inceptionV3, inceptionV4, ResNet50, ResNet152, NiN, VGG16, VGG19, YoloV1, YoloV2, YoloV3, VDSR, SR_ResNet, MobileNet, SSD, and many custom ones

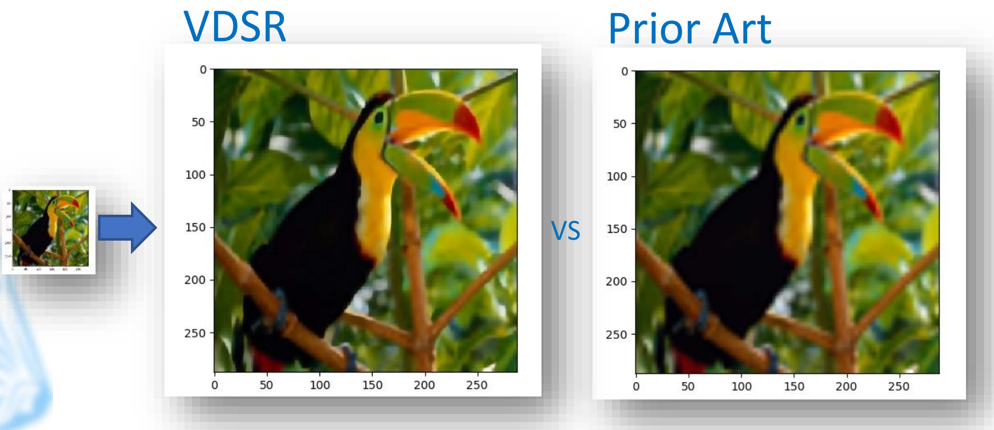
Classification



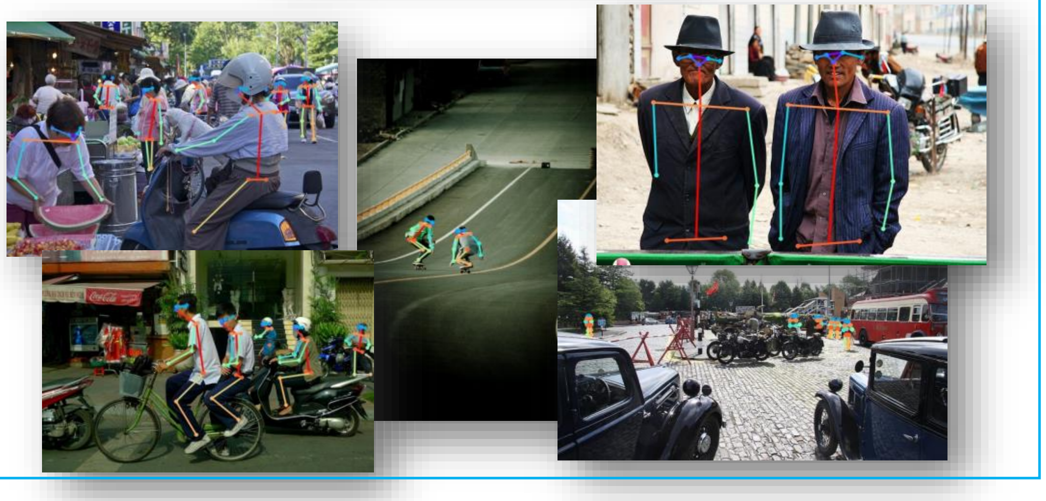
Segmentation



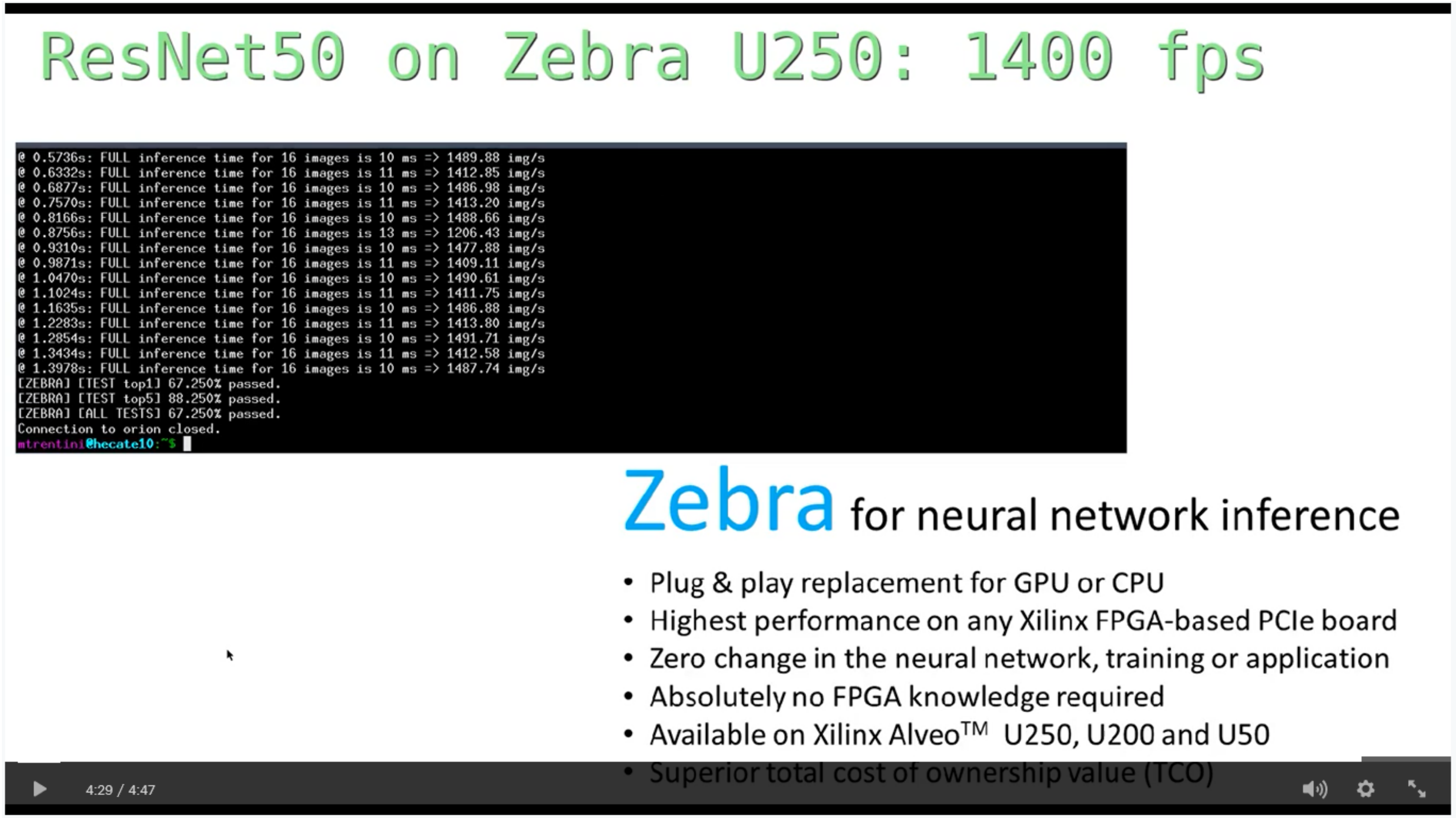
Super Resolution



Body Positioning



Zebra Demo on Alveo U50



ResNet50 on Zebra U250: 1400 fps

```
@ 0.5736s: FULL inference time for 16 images is 10 ms => 1489.88 img/s
@ 0.6332s: FULL inference time for 16 images is 11 ms => 1412.85 img/s
@ 0.6877s: FULL inference time for 16 images is 10 ms => 1486.98 img/s
@ 0.7570s: FULL inference time for 16 images is 11 ms => 1413.20 img/s
@ 0.8166s: FULL inference time for 16 images is 10 ms => 1488.66 img/s
@ 0.8756s: FULL inference time for 16 images is 13 ms => 1206.43 img/s
@ 0.9310s: FULL inference time for 16 images is 10 ms => 1477.88 img/s
@ 0.9871s: FULL inference time for 16 images is 11 ms => 1409.11 img/s
@ 1.0470s: FULL inference time for 16 images is 10 ms => 1490.61 img/s
@ 1.1024s: FULL inference time for 16 images is 11 ms => 1411.75 img/s
@ 1.1635s: FULL inference time for 16 images is 10 ms => 1486.88 img/s
@ 1.2283s: FULL inference time for 16 images is 11 ms => 1413.80 img/s
@ 1.2854s: FULL inference time for 16 images is 10 ms => 1491.71 img/s
@ 1.3434s: FULL inference time for 16 images is 11 ms => 1412.58 img/s
@ 1.3978s: FULL inference time for 16 images is 10 ms => 1487.74 img/s
[ZEBRA] [TEST top1] 67.250% passed.
[ZEBRA] [TEST top5] 88.250% passed.
[ZEBRA] [ALL TESTS] 67.250% passed.
Connection to orion closed.
mtrentini@hecate10:~$ 
```

Zebra for neural network inference

- Plug & play replacement for GPU or CPU
- Highest performance on any Xilinx FPGA-based PCIe board
- Zero change in the neural network, training or application
- Absolutely no FPGA knowledge required
- Available on Xilinx Alveo^{TM} U250, U200 and U50
- Superior total cost of ownership value (TCO)

Q&A

Mipsology

Request our Zebra Solution Overview

Zebra Solution Overview
https://mipsology.com/product/#download

Visit us at www.mipsology.com   Or email us at zebra@mipsology.com