Mipsology

**AMD XILINX**

# Zebra™ by Mipsology: Accelerating Neural Network Inference

## INTRODUCTION

### FPGAs Ideally suited for Inference Acceleration

FPGAs are full of basic computing elements and filled with memories, ideally suitable for high performance and low latency CNN inference computing. They are reprogrammable at the hardware level allowing for continual adaptation. But FPGA programming can be challenging.

## SOLUTION OVERVIEW

### Zebra Software Accelerates Inference

Zebra is the ideal compute engine to accelerate CNN inference. Zebra seamlessly replaces CPUs/GPUs to compute any neural network on an FPGA faster, with lower power consumption, and at lower cost.

Simply type a single Linux command. No knowledge of FPGA technology, compilation, or any changes to the environment or the application are required.

Zebra lets AI engineers focus on application development, while enjoying unmatched performance.
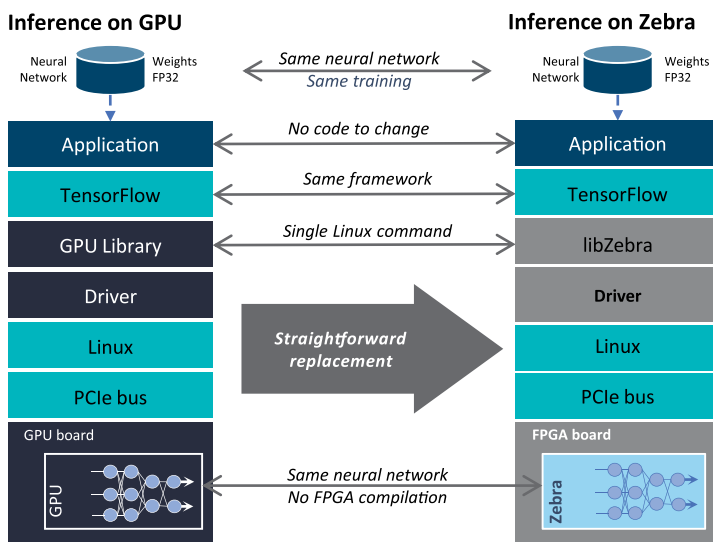
### AMD - XILINX ADVANTAGE

> Delivers high performance from edge to cloud

> Supports most popular deep learning frameworks and broad range of CNN inference

> Easy to use a single Linux command replaces costly CPU or GPU inference nodes

---

> Zero Neural Networks Change

> Zero New Training

> Zero Line of Code to Add

> Zero FPGA Knowledge

> Zero FPGA Compilation

> Zero Transition Effort

... In Short, ZERO Effort



Inference on GPU / Inference on Zebra comparison diagram:

**Inference on GPU** — Neural Network, Weights FP32 → Application → TensorFlow → GPU Library → Driver → Linux → PCIe bus → GPU board (GPU)

**Inference on Zebra** — Neural Network, Weights FP32 → Application → TensorFlow → libZebra → Driver → Linux → PCIe bus → FPGA board (Zebra)

Same neural network / Same training
No code to change
Same framework
Single Linux command
*Straightforward replacement*
Same neural network / No FPGA compilation

**AMD XILINX**

## KEY BENEFITS

### The Fastest Inference

> Computes neural networks highest speed with lowest latency

### Supports All Neural Networks

> Accelerates any CNN, including user defined neural networks

### Extremely Simple to Use

> Deploying Zebra is a "Plug & Play" process

### No Changes to the Software Environment

> Not a single line of code must be changed in the application

### Scalable, Flexible and Adaptable

> Easy replacement of GPUs or complement CPUs

## SOLUTION DETAILS

### Neural Networks

> Supports CNN without modification

> Delivered with tested networks: GoogLeNet V1, Inception V3, Inception V4, VGG16, VGG19, ResNet50, ResNet152, YoloV1, YoloV2, YoloV3, Tiny YoloV2, Tiny YoloV3, VDSR, SSD, MobileNet

> Accelerated layers: convolution, fully connected, max/average pooling, concat, batch norm, scale, add eltwise, reorg, up sampling, depth to space, reduce mean, dilated convolution, squeeze, separable depth wise, clip to value, relu, leaky relu, relu6

> Automatic split of graph

> Up to 3.2 billion weights

> Up to 1 million layers

> Unbounded number of convolutions

> Single or multiple outputs

> Up to 1360x1360x3 input images

> Up to 24 simultaneous independent users

### Supported Frameworks

> TensorFlow, PyTorch , ONNX, and many more

> No change to source code required

### Precision

> 8 bit

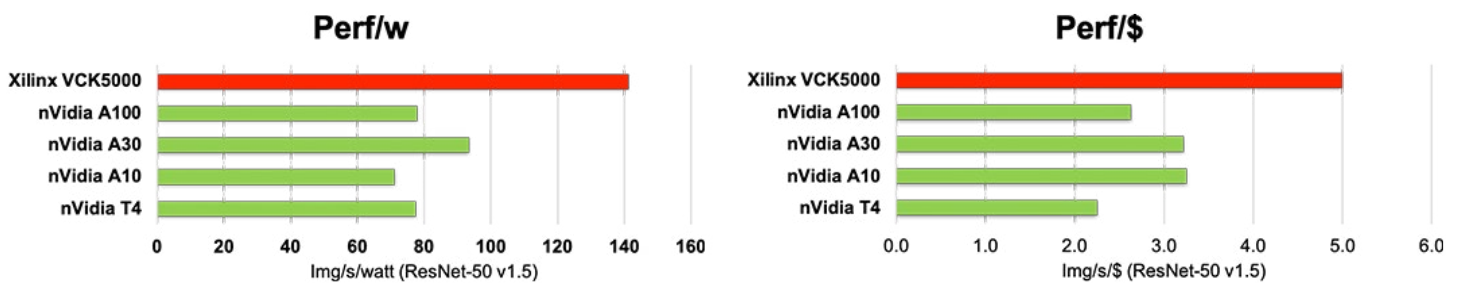> Automatic proprietary quantization

### Migration from GPU or CPU

> Trained parameters from GPU/CPU training without changes

> No proprietary training or re training needed, and no pruning required

> Usable immediately

> Similar accuracy as FP32

### Power & Cooling

> From few watts in the field to 140W in data centers

## RESULTS

Zebra's quantization ensures results stay accurate and does not require retraining the NN.



Perf/w — Img/s/watt (ResNet-50 v1.5): Xilinx VCK5000, nVidia A100, nVidia A30, nVidia A10, nVidia T4

Perf/$ — Img/s/$ (ResNet-50 v1.5): Xilinx VCK5000, nVidia A100, nVidia A30, nVidia A10, nVidia T4

## TAKE THE NEXT STEP

Learn more about Xilinx VCK5000 Development Card  xilinx.com/products/boards-and-kits/vck5000