

Mipsology Zebra™: Fast Migration of Neural Network Inferencing

INTRODUCTION

FPGAs: Ideally suited for Inference Acceleration

Migration of AI inference workloads from GPU/CPU to FPGAs has been seen as notoriously difficult, time-consuming and cumbersome. Mipsology makes migration to FPGAs quick, simple, and cost-effective, allowing customers to realize the benefits of high throughput and deterministic low latency delivered by Xilinx Alveo™ accelerator cards.

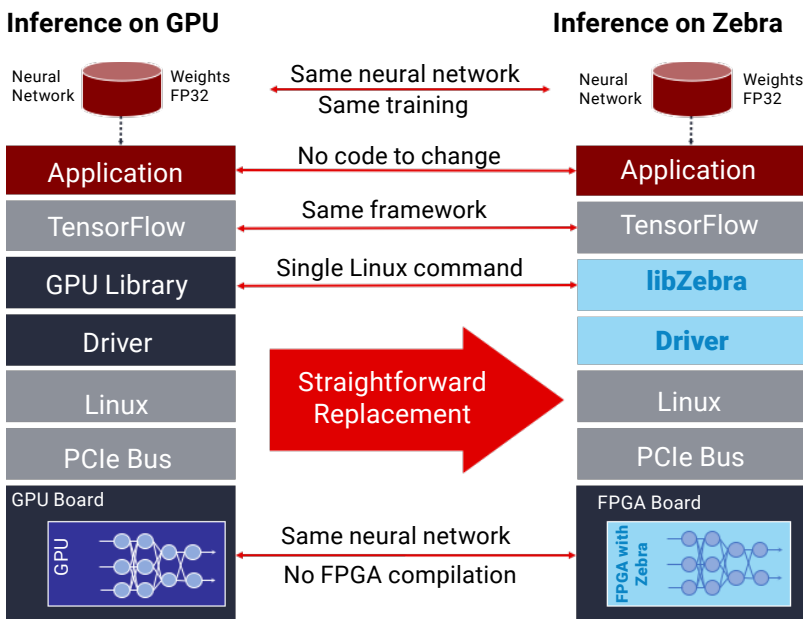
SOLUTION OVERVIEW






Zebra Makes Migration from CPU/GPU Easy

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

Migration is a snap: simply type a single Linux command. No knowledge of FPGA technology, compilation, or any changes to the environment or the application are required.

Zebra frees AI engineers to focus on application development, while enjoying unmatched performance.



-  **The Fastest Inference**
-  **Support All Neural Networks**
-  **Extremely Simple to Use**
-  **No Code Change**
-  **Scalable, Flexible & Adaptable**

Effortless Migration From GPU/CPU

- ▶ No neural network change
- ▶ No new training
- ▶ No lines of code to add
- ▶ No FPGA knowledge needed
- ▶ No FPGA compilation
- ▶ No transition effort

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, sigmoid...
- 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, Caffe, MXNet
- No change to source code required

Precision

- 8-bit
- Automatic proprietary quantization

Migration from GPU/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 and Cooling

- From a few watts in the field to 140W in data centers

RESULTS

Neural Network	Performance On Alveo Accelerator					
	Large Batch			Non-Batch		
	U250	U200	U50_LV	U250	U200	U50_LV
ResNet50 ¹	5078	3195	1755	3285	1940	1157
ResNet152 ¹	1907	1172	642	1528	892	527
InceptionV3 ¹	2452	1463	877	2048	1194	738
InceptionV4 ¹	1286	795	450	1154	698	405
VGG16 ¹	850	581	369	556	342	219
Yolo-V2 ²	588	411	244	585	411	243
Yolo-V3 ²	254	180	110	253	178	110

Notes:

Performance measured as Frames per Second(FPS)

1. ImageNet Dataset

2. Coco Dataset

TAKE THE NEXT STEPS > [Learn more about Alveo accelerator cards.](#)
[Learn more about Mipsology www.mipsology.com](#)
[Contact Mipsology sales](#)

