Mipsology

Zebra for Deep Learning Inference

Fast execution, easy deployment, endless adaptability

- Seamless and quick replacement for CPU or GPU
- Supports any Convolutional Neural Network and several Frameworks
- Runs on all Xilinx FPGAs, from small to large

INTRODUCTION

FPGAs for AI

Today, FPGAs are more popular than ever. Bolstered by the unfolding of Artificial Intelligence, a field of virtually infinite possibilities, they are becoming a viable and efficient alternative to CPUs and GPUs.

FPGAs are attractive because of their raw processing power and adaptability. Their massive parallelism supports tens of thousand operations, hundreds of million times every second. Their fast re-programmability allows them to be reconfigured at any point up to and after the end-product has been deployed. Both characteristics are essential for accelerating Deep Learning Inference in data centers.

The Other Side of the Coin

But the benefits don't come for free. Programming an FPGA requires uncommon knowledge and expertise possessed only by few highly specialized hardware designers.

ZEBRA OVERVIEW

Zebra by Mipsology

Mipsology is a startup developing state-of-the-art FPGA-based accelerators for Deep Learning Inference. It was founded in 2015 by a team of engineers and scientists who created a family of world-leading FPGA-based supercomputers over the past 20 years.

The team devised the technology that eliminates the challenging task of programming an FPGA for DL Inference and made it invisible to AI engineers. Called Zebra, the accelerator has been conceived as a replacement for CPUs/GPUs. It provides very fast speed of execution, is easy to deploy, and accommodates a range of Neural Networks and Frameworks.

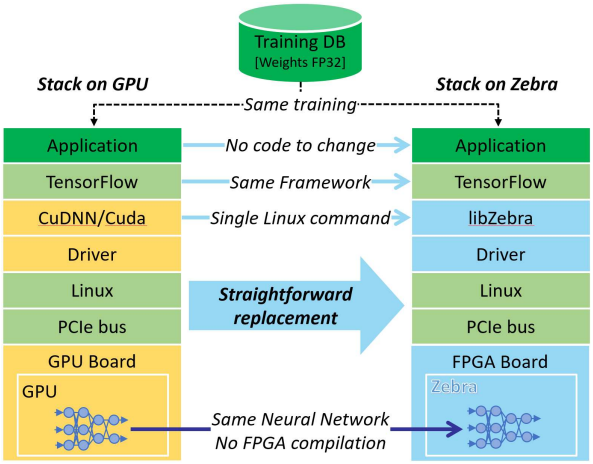
How to Replace GPUs with Zebra

The deployment of Zebra is simple and straightforward. It does not require knowledge of FPGA technology, compilation or changes to the environment nor to the application.

Adopting Zebra is swift, seamless and painless. Zebra lets the AI engineers continue to work as before just at significant higher performance.

Zebra can replace GPUs or complement CPUs in data centers, and thanks to its lower power needs Zebra works also at the edge or just under a desk.

The Zebra application stack corresponds one-for-one to a GPU stack.



ZEBRA SPECIFICATIONS

Supported Neural Networks

- Delivered with many pretrained networks: AlexNet, CaffeNet, GoogLeNet V1, Inception V3, Inception V4, VGG16, VGG19, ResNet50, ResNet152, Nin…
- Supports custom CNN without modification
- Supported layers: convolutions, fully connected, max/average pooling, concat, LRN, relu, softmax, batch norm, scale, add eltwise, etc
- Up to 3 billion weights in a single network
- Up to 1 million layers
- Tens of thousands of filters per convolution
- Input images up to 1360x1360 with 3 color channels

Supported Frameworks

- Caffe, Caffe2, MXNet or TensorFlow
- No code change required, no extra compiler

Precision

- 8-bit or 16-bit integers with automatic quantization

Hardware

- Runs on FPGAs of any size: KU40, KU115, KU15P, VU9P, VU13P, ZU7EV, and more
- Alveo™ U200, Alveo™ U250, VCU1525, KCU1500

Power & Cooling

- No specific power/HVAC requirements
- Typically 40W on KU115, 75W on VU9P

MIGRATION FROM CPU or GPU

- Uses trained parameters from GPU training
- No proprietary training or re-training needed
- Offers same accuracy as GPU or CPU
- Fully off-load CPU from all neural network computing

CONCLUSION

Mipsology's Zebra uses FPGAs to accelerate inference of convolutional neural networks (CNN), faster than GPUs and CPUs, without requiring any change. Replacing CPUs and GPUs with Zebra is a plug&play process, effortless, painless and quick.

Zebra has been optimized to deliver the highest performance on any FPGA and still kept simple for AI engineers to deploy it on their applications within minutes.

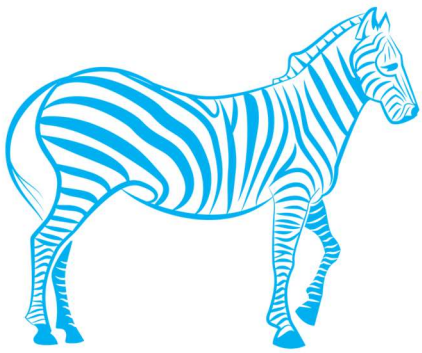
We focus on acceleration,
You focus on your application!

TAKE THE NEXT STEP

Use Zebra in the Cloud
Zebra runs on AWS EC2 F1 instances:
https://aws.amazon.com/marketplace/search/results?searchTerms=zebra

Use Zebra on Xilinx® Alveo™ Data Center Accelerator Cards – www.xilinx.com/alveo
Contact Mipsology to get a license for your U200/U250.

Use Zebra on your own FPGA board
Zebra can also run on your board, contact Mipsology!



Mipsology

24, rue Emile Baudot
91120 Palaiseau
France

www.mipsology.com
zebra@mipsology.com