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.

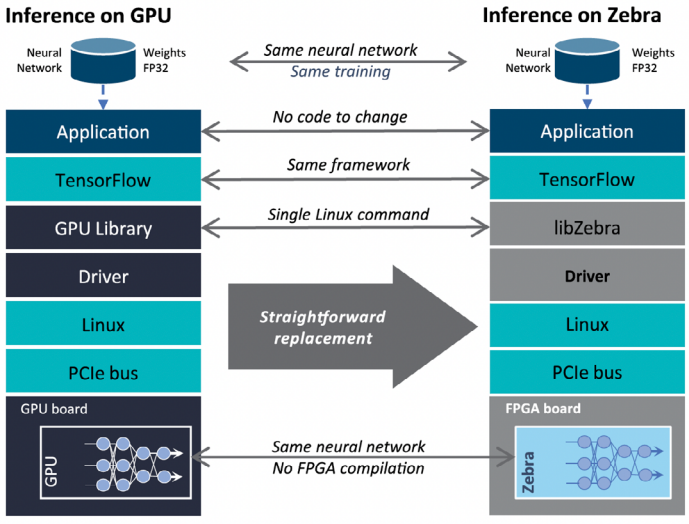
SOLUTION OVERVIEW

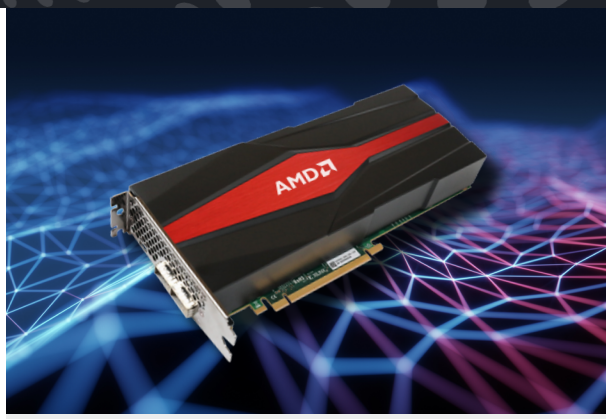
Zebra Software Accelerates Inference

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 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**

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

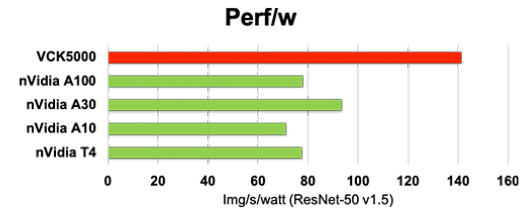
Precision
- 8 bit
- Automatic proprietary quantization

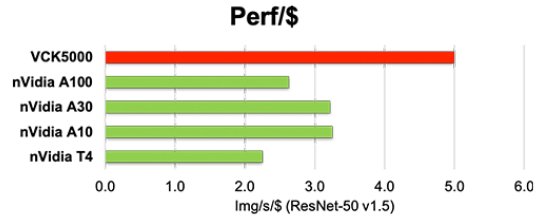
Migration from GPU or CPU
- Trained parameters from GPU/CPU training without changes
- No proprietary training or re training needed, and no pruning required
- Usable immediately
- Similar accuracy as FP32

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.





NEXT STEPS

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

DISCLAIMERS

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

COPYRIGHT NOTICE

© 2023 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Versal, Vitis, Vivado, and other designated brands included herein are trademarks of Advanced Micro Devices, Inc. PCIe is a trademark of PCI-SIG and used under license. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.