SOLUTION BRIEF

DeepAI: Integrated AI Training and Inference for the Edge

INTRODUCTION

AI training demands massive compute resources that typically utilize expensive and power-hungry GPUs. Consequently, deep learning is customarily performed in the cloud or in large on-prem data centers. Training new models takes days and weeks to complete, and inference queries suffer from long latencies of the round-trip delays to and from the cloud.

DeepAI's solution addresses these challenges by providing an integrated ML training and inference solution on Xilinx Alveo U50 accelerator cards. This solution is deployed at the edge to deliver high throughput performance and low latency.

SOLUTION OVERVIEW

The data for updating training models and for inference queries is generated mostly at the edge – in stores, factories, terminals, office buildings, hospitals, city facilities, 5G cell sites, vehicles, farms, homes and hand-held mobile devices. Transporting the rapidly growing data to and from the cloud or data center leads to unsustainable network bandwidth, high cost and slow responsiveness, as well as compromises data privacy and security and reduces device autonomy and application reliability.

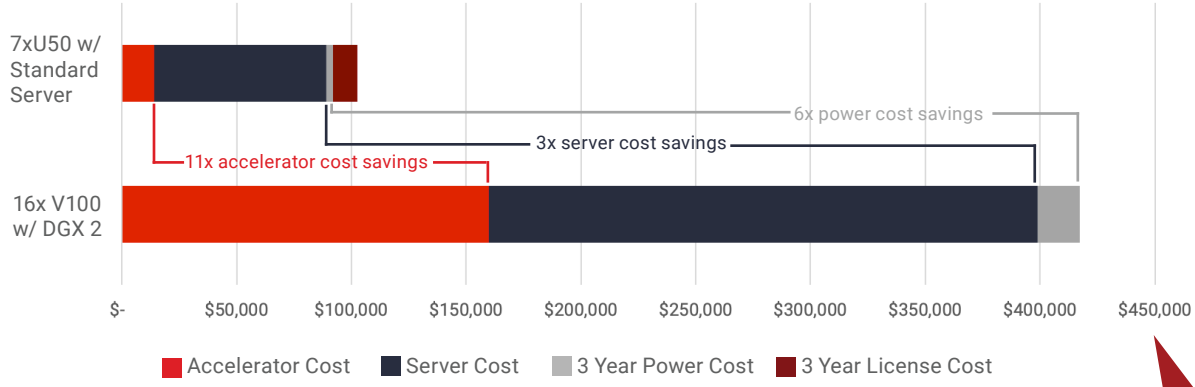
Deep-AI unique and efficient deep learning solution for the edge, allows for integrated ML training & inference workloads to be deployed on Alveo U50, removing the need to go to the cloud or datacenter for retraining on high-end GPUs.

Deep-AI's training solution delivers 2x the performance of NVIDIA V100 GPU at a ⅓ of the power and ¼ of the cost. With training done at the Edge, customers use one system to train in parallel to online inference.

Features and Benefits

- Quantized training at 8-bit fixed point
- Training at high sparsity ratios
- Same hardware accelerator for training and inference
- 2x faster training vs Nvidia v100
- 4x lower TCO vs Nvidia v100
- Training output is inference ready
- Seamless transition between training and inference
- Scalable, secure, and reliable

DeepAI vs Nvidia V100 TCO Analysis



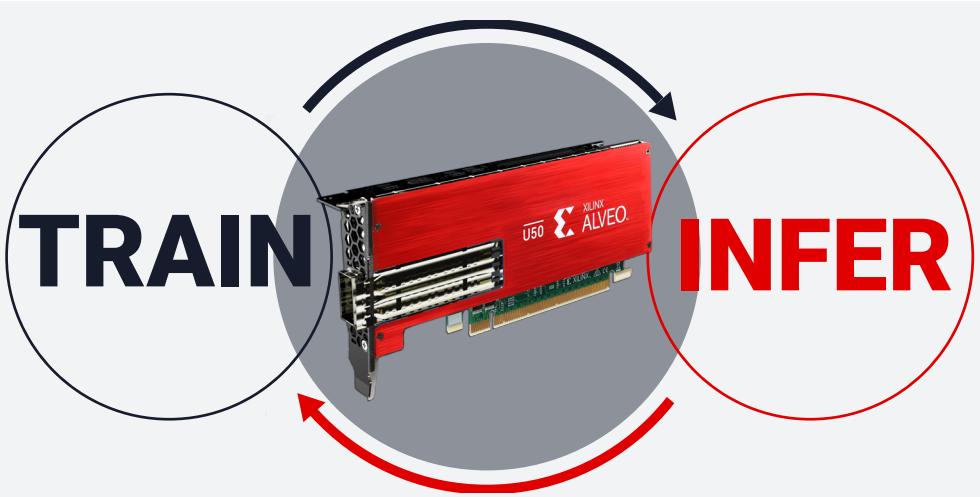
XILINX

Adaptable. Intelligent.

SOLUTION DETAILS

Deep-AI's solution runs on the Xilinx Alveo U50 data center accelerator card. The same accelerator is used for inference and retraining of the deep learning model, allowing an on-going iterative process that keeps the model updated to the new data that is continuously generated.

Deep-AI's software solution ensures the underlying Alveo U50 accelerator is completely under-the-hood and transparent to data scientists and developers designing their AI applications.



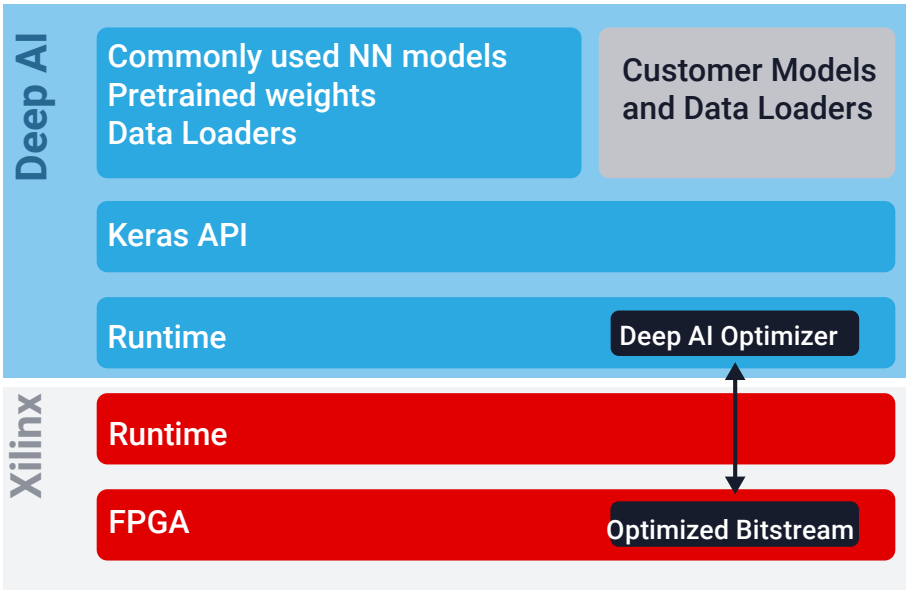
- **Training output feeds directly to inference**
- **No manual post-training processing or calibration**
- **No loss of accuracy from training to inference**

Supported Frameworks

- Tensorflow, PyTorch, Keras

Supported neural networks

- CNNs for Imaging applications popular neural layers, convolutions, max/average pooling, residual shortcut, batch norm
- Resnet and Mobilenet for Classification
- Yolo, TinyYolo and SSD for Object Detection
- Multilayer Perception(MLP)



Notes:
Performance based on training ResNet50 with ImageNet dataset
V100 training performance = 360 images/sec at FP32, benchmarked on AWS
U50 training performance = 800 images/sec
Assume standard server can host up to 8 U50 cards
Assume fair market listing price for all the accelerators
Standard servers using matching HW(CPU, Memory, Storage, etc.) configurations as DGX2

TAKE THE NEXT STEPS > For more information or to request a demo visit: www.deep-aitech.com
Or contact DeepAI Sales

XILINX.

Adaptable. Intelligent.