DELL EMC

Whitepaper

Deep Learning Inferencing with Mipsology using Xilinx ALVEO™ on Dell EMC Infrastructure

Revision: **1.1**
Issue Date: **11/5/2019**

Abstract

This blog evaluates throughput, efficiency and ease-of-use of Deep Learning inference performed by Mipsologys' Zebra software stack running on FPGA-based Xilinx ALVEO™ U200 installed in a Dell EMC PowerEdge R740/R740xd server. The objective is to show how the Zebra stack from Mipsology can deliver high inferencing throughput without requiring any effort.

Revisions

Date	Description
24 October 2019	Initial release

Acknowledgements

This paper was produced by the following people:

Name	Role
Bhavesh Patel	Server Advanced Engineering, Dell EMC
Ludovic Larzul	CEO, Mipsology

Overview of Deep Learning

The deployment of a Deep Learning (DL) algorithm proceeds in two stages: training and inference. As illustrated in Figure 1, training configures the parameters of a neural network model implementing an algorithm via a learning process based on a large dataset over several training iterations and loss function [1]; the larger the dataset, the higher the accuracy of the model. The output of this stage, the learned model, is then used in the inference stage to speculate on new data.

There are two major differences between training and inference: training employs *forward propagation* and *backward propagation* (two classes of the deep learning process), whereas inference mostly consists of forward propagation [2].

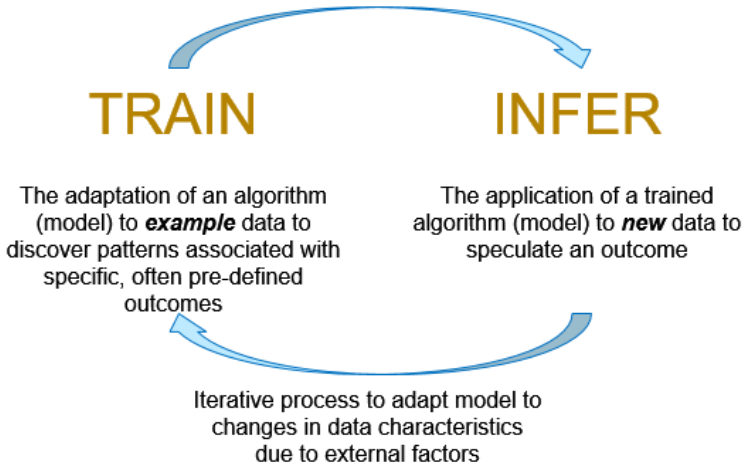


Figure 1. Deep Learning phases.

Deep Learning Inferencing

Upon completing training, a model can be *deployed* on a variety of hardware accelerators such as CPUs, GPUs, FPGAs or special purpose devices to perform a specific business-logic function or task such as identification, classification, recognition and segmentation [Figure 2].

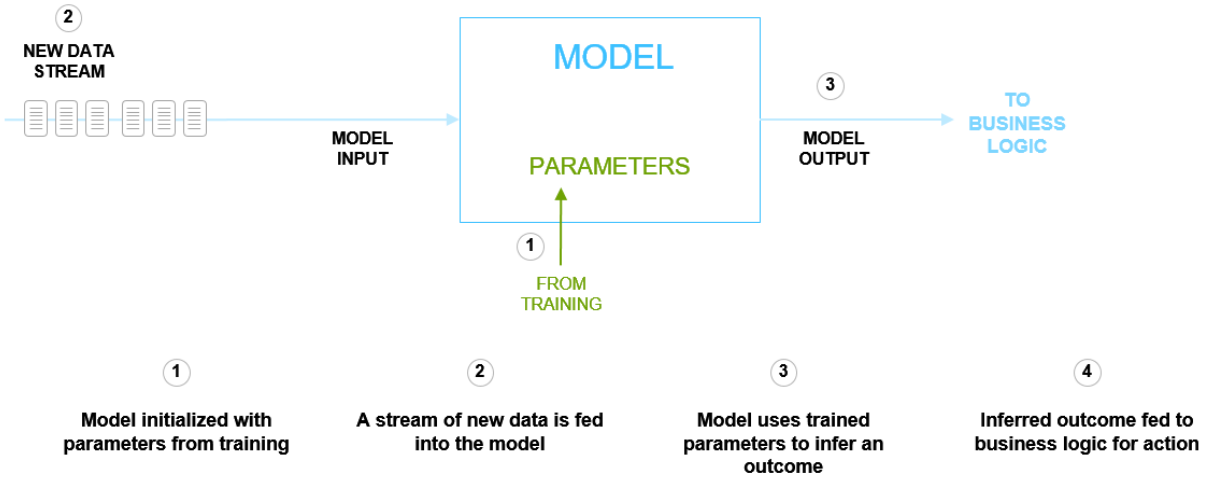


Figure 2. Inference Flow.

The hardware accelerator must meet a set of basic requirements:

- Deliver high throughput of computations with low latency.
- Support a neural network as defined by AI scientists without changes to avoid a time-consuming re-design or never-ending training.
- Adaptable to accommodate different loads without long delay to execute different models.
- Be a proven hardware solution that can run 24/7 without interruptions.
- Be flexible to support the constantly evolving neural network technology.

FPGA devices are perfect choices for the tasks. They deliver high processing power, especially in fixed point computation, provide high adaptability, and consume low power. DL is evolving rapidly, making the FPGA an excellent choice to accommodate new requirements avoiding silicon re-spins, thus drastically reducing the cost of ownership. FPGAs come in various sizes making them ideal for most of the places where inference happens, from IoT devices, to embedded applications, to field, data centers and Clouds.

However, for all their noteworthy characteristics, FPGA's *have traditionally been* complex to program, requiring uncommon skills and knowledge, *but new solutions available in the marketplace are making implementation much easier".*

The focus of this blog is on FPGA-accelerated inference of (Convolutional Neural Networks) CNNs, specifically on how Zebra from Mipsology enables ALVEO™ boards to perform the task in Dell PowerEdge Servers.

Why FPGA?

FPGAs achieve high computation throughput via a robust set of resources that comprises substantial reprogrammable lookup tables (LUT) to implement millions of equivalent Boolean-logic functions, a large assembly of multipliers/adders (MAC), numerous embedded memories to accommodate a broad variety of logic circuitry. They can also support a high number of off-chip memories if necessary. A series of auxiliary logic, such I/O interfaces, etc., complete the device. The overall fabric is ideal for parallel processing as required by (neural networks) NN. See figure 3.

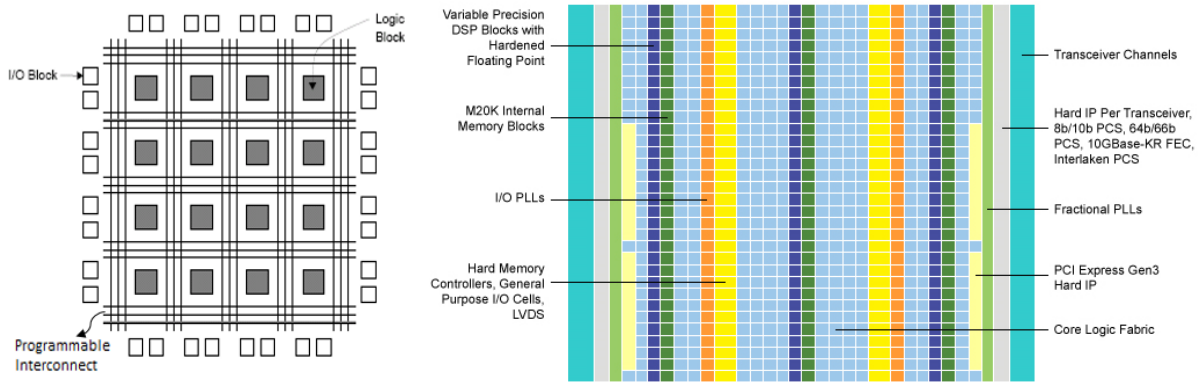


Figure 3. (Left) Arrayed building blocks are connected via interconnect wires; (Right) Fully featured FPGAs include a variety of advanced building blocks.

While CPUs/GPUs operate at the byte level, FPGAs function at the bit level, giving the user the ability to design logic to perfectly fit a task. Users can handle irregular parallelism and fine-grained computations much better with FPGAs than with CPUs/GPUS. FPGAs are ideal for processing both sparse data and compact data types.

The re-programmability of the FPGA permits an unusual degree of customization after the hardware was manufactured. The ability to tune the underlying hardware architecture and select any computation quantization desired allows for FPGA-based platforms to support state-of-the-art deep learning innovations as they emerge.

As mentioned before, the cornucopia of resources and the benefits come at a price. Not only do FPGAs require unique skills and expertise to program and deploy them, but their compilation process is very slow, and they need a laborious tuning process to achieve high frequency computation.

Xilinx ALVEO™ U200

The Xilinx® ALVEO™ U200 and U250 data center accelerator cards are peripheral component interconnect express (PCIe®) Gen3 x16 compliant cards featuring the Xilinx Virtex® UltraScale+™ technology. These cards accelerate compute-intensive applications such as machine learning, data analytics and video processing. The ALVEO™ U200 and U250 data center accelerator cards are available in passive and active cooling configurations. See figure 4 and 5.

Figure 4. Xilinx ALVEO™ U200 data center accelerator card

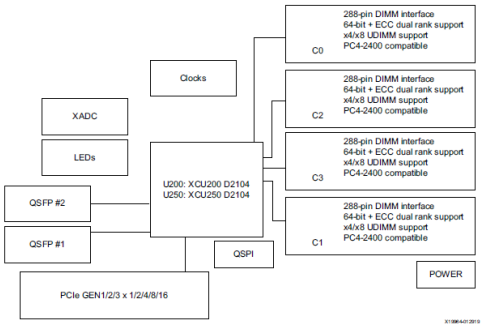


Figure 5. U200 Block Diagram.

Zebra Acceleration Stack from Mipsology

Mipsology, an innovative AI/DL startup, conceived a software stack called Zebra for ultra-fast inference acceleration of CNN. Based on FPGAs, Zebra sits on top of the FPGAs and conceals them to the user.

The Mipsology Zebra stack outshines its competitors with several advantages: exceptional throughput, very low latency, very high efficiency, and remarkable ease-of-use.

The following lists the main Zebra's capabilities:

Zebra Delivers High Performance, Low Latency and High Efficiency

Taking advantage of the rich set of FPGA resources and executing them at high frequency (650MHz), Mipsology implemented Zebra to deliver high throughput on all CNN. Several computational optimizations within Zebra push its efficiency to a level not achievable with conventional hardware accelerators. The typical pruning required by other FPGA solutions is not useful with Zebra to enhance performance, avoiding re-training of the NN.

The FPGA architecture also allows Zebra to deliver low latency for application requiring real time response.

Further, Mipsology releases periodic software updates to improve performance. Ultimately, the user can preserve the hardware investment and deliver more throughput by a simple software upgrade.

Zebra Accelerates Inference Calculation of Any Neural Network

Zebra accelerates a large set of the layers commonly used in CNNs. No change is required to the neural network or its training to run on Zebra. This gives Zebra the ability to supports a wide range of CNNs, from the most popular neural networks found on the internet to custom designed CNNs for commercial use. All the layers are accelerated within the FPGA to maximize

the performance and free the CPU. To accommodate the progress of ML technology, Mipsology R&D expands the acceleration to new layers on a regular basis.

When in the course of a project a CNN evolves, it can be processed on Zebra on-the-spot once trained, drastically simplifying the deployment of new CNN versions in data centers, at the edge, on a desktop, or in embedded applications.

Once a CNN has been trained on GPUs or CPUs, it can be processed on Zebra as is; eliminating the need for re-training and for new tools to migrate the neural network.

Zebra Works with Most Popular Neural Frameworks

Zebra is integrated in Caffe, Caffe2, MXNet, TensorFlow, PyTorch and ONNX to accelerate any application without changing the sources code, without adding a proprietary API, and most likely without even recompiling the application.

Zebra Accelerates CNN in Datacenters, on The Edge, in the Desktop, in Embedded Applications

Whether installed on PCIe-boards designed for Data Centers or for desktops, or encapsulated in devices for edge computing, the combination of Zebra with FPGAs accelerates any CNN application.

The board configured with Zebra can compute a CNN in a data center when processing massive amounts of data, or in the field for local processing. One-size-fits-all so there is no need to duplicate the efforts designing ML for data center and for the edge.

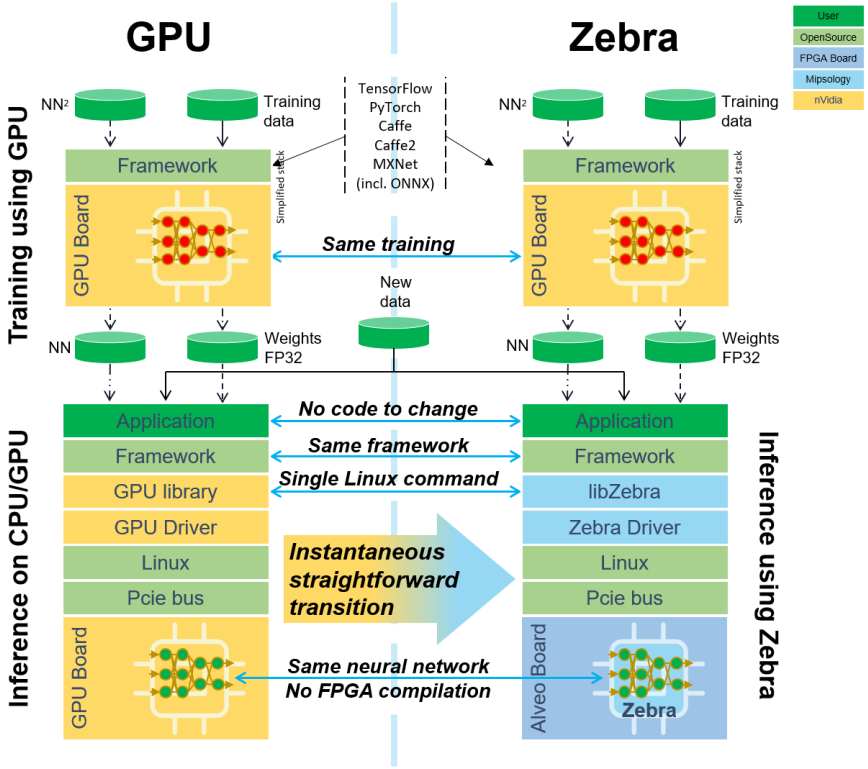


Figure 6. Mipsology' s Zebra stack

Zebra Is Easy to Use

Deploying Zebra is a "plug play" process. Plug an ALVEO™ Board into a PC running Linux, issue one single Linux command to configure it, and you are ready to go. See figure 6.

There is no R&D cost in using Zebra. No extra work is required to make Zebra compute a neural network nor is any proprietary tool needed to understand how to migrate the neural network.

Zebra Uses 8-bit or 16-bit Fixed-point Math and Optimizes Calculation Precision

Zebra executes inference calculations using 8-bit or 16-bit fixed-point integers. Mipsology has devised a quantization algorithm that converts floating points training parameters into fixed points parameters in minutes maintaining similar accuracy achieved in training. The quantization in Zebra does not require users specific effort nor does it require to retrain the network for lower precision.

Zebra Scales in Size and Power

Zebra runs one or many FPGA boards of any size, scaling from supporting massive computing in datacenters to limited computing at the edge or in embedded applications.

Zebra Supports Multiple Users and Multiple Neural Networks on the Same FPGA Board

The unique architecture of Zebra supports multiple users and multiple networks, simultaneously, on the same ALVEO™ board. The resource sharing is selected when using the board, maximizing the investment in hardware to avoid under-utilized computing resources. Each processing session owns a "full stack", making call through its framework, and executing its own neural network with its own weights.

Zebra Does Not Require Knowledge of FPGA Technology

Zebra removes the burden to learn the FPGA technology. It does not require knowledge of any hardware design language, new design tools, or to understand hardware-level details. Delivered with pre-compiled FPGA binary files, it removes the need to learn FPGA coding and compilation.

Zebra Boasts the Lowest Cost-of-Ownership

The collective characteristics of Zebra make it the accelerator with the lowest cost-of-ownership (COO).

The ease-of-use, scalability, inherent lower power consumption and the ability to support large number of neural networks make the Zebra stack running on Xilinx FPGA a pretty good proposition.

There is no additional R&D effort to run the same trained neural network on various sizes of Zebra accelerators or in different locations, from data center to embedded. If the size of a neural network grows during the life-time of the product, a simple upgrade of the hardware, keeping the same Zebra stack, will accommodate the increased computational power of a larger neural network. The ongoing enhancement of Zebra will reduce the computing needs, letting the user do more computing on the same hardware without incurring in additional cost. As Zebra runs the same neural networks as a CPU or a GPU, without migration and changes, it is also possible to mix the resources and maximize the investment in hardware to accommodate usage peaks.

Dell EMC PowerEdge Servers

The evaluation was based on Dell EMC PowerEdge R740/R740xd servers to host the Xilinx ALVEO™ boards. The PowerEdge R740/R740xd is a general-purpose platform with highly expandable memory (up to 3TB) and impressive I/O capability to match both read-intensive and write-intensive operations. The R740 is capable of handling demanding workloads and applications such as data warehouses, E-commerce, databases, high-performance computing (HPC), and Deep Learning workloads.

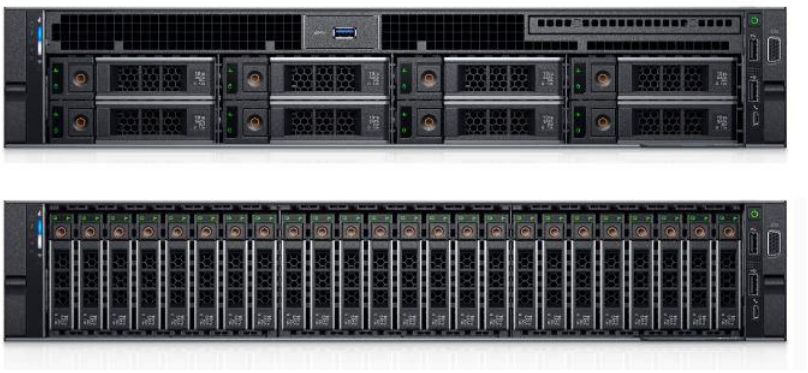


Figure 7. Dell PowerEdge R740/R740xd.

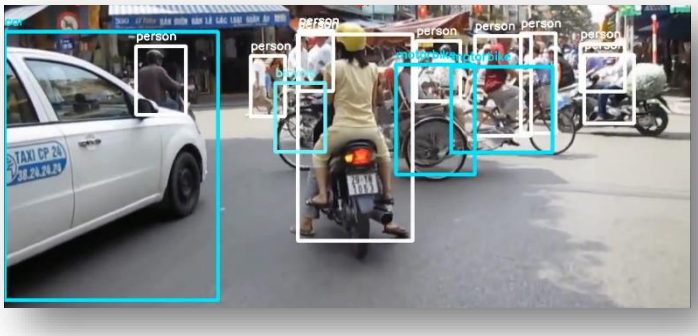
The PowerEdge R740/R740xd is Dell EMC's latest two-socket, 2U rack server designed to run complex workloads using highly scalable memory, I/O capacity, and network options. The R740/R740xd features the Intel Xeon processor scalable family, up to 24 DIMMs, PCI Express (PCIe) 3.0 enabled expansion slots, and a choice of network interface technologies to cover NIC (Network interface card) and rNDC (rack network daughter card). In addition to the R740's capabilities, the R740xd adds unparalleled storage capacity options, making it well-suited for data intensive applications that require greater storage, while not sacrificing I/O performance. See figure 7.

Zebra Applications

Zebra is a perfect fit for CNN acceleration for still and video image processing in a wide range of applications. Zebra can ideally replace CPUs and GPUs without added engineering costs.

Here are few examples of applications:

Video surveillance



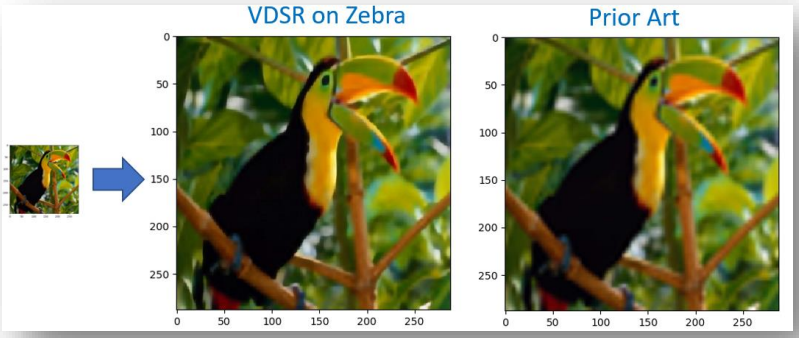
Video surveillance can be implemented via multiple cameras with a single ALVEO/Zebra combo installed in a Dell PowerEdge R740 server. All cameras would run concurrently in real time, not sequentially as on GPU boards. In the case of GPUs, the above scenario would need multiple GPU boards or significant R&D effort, dramatically increasing the cost and complicating the deployment. With multiple neural networks running in parallel on a single board, scaling up/down is considerably simplified. Not to mention the increased portability for edge computing vis-à-vis multiple GPU boards that would further call for additional power and need for cooling.

Smart City Surveillance

A subclass of video surveillance, monitoring a city to assure its safety can be carried out by an ALVEO/Zebra combo. For example, in a surveillance system installed in a city with hundreds of main street intersections, cameras scrutinize the traffic crossing at each of the intersection. The ALVEO/Zebra combo installed in a set of Dell PowerEdge R740 servers in a surveillance center can receive the feeds via a 4G/5G connections, process them to identify any event that needs attention. Local processing can also be performed using the same Zebra technology in a smaller Dell computer near the intersection.

Image resolution enhancement

An ALVEO/Zebra combo installed in a Dell EMC PowerEdge R740 server can be used for generating high-quality high-resolution images from low res images by mapping a very deep super-resolution (VDSR) algorithm onto Zebra. Super resolution algorithms are particularly demanding in processing power as they must generate high resolution images. On a Dell PowerEdge R740, Zebra running on ALVEO™ can deliver real time video VDSR movie. The same processing can be applied in video surveillance to enhance the images that humans have to analyze to identify people.



Applications over internet using image segmentation and classification

Many applications over internet rely on identifying objects, people, text, scenes, activities, body positioning, or globally the content of images or videos. Relying on all sorts of convolutional neural networks, their processing requires high bandwidth and flexibility to adapt to the demand. Dell PowerEdge R740 installed in a data center or a Cloud, equipped with ALVEO running Zebra can easily process the load. Zebra's ability to run multiple networks in parallel simplifies the deployment of applications. The low latency of FPGAs helps reducing the response time of reactive applications. Transitioning from a farm running inference on CPU/GPU to ALVEO configured with Zebra is straightforward since Zebra supports the same neural networks without modifications.

Automation of Quality Control in Manufacturing

An ALVEO/Zebra combo installed in a Dell EMC PowerEdge R740 server can automate quality control in manufacturing, dramatically increasing the accuracy of the monitoring. By installing one of them next to the production line, or few of them in a more central location, the quality of the manufacturing process can increase considerably, and the overall cost noticeably reduced.

Evaluation Methodology

The evaluation was performed via a setup made up by a combination of hardware and software processing an image-classification application. The hardware accelerator consisted of a Xilinx ALVEO™ U200 FPGA board hosted by PowerEdge R740/R740xd servers, running Linux.

The software included Mipsology' s Zebra that computed inference of eight trained deep learning models or convolutional neural networks running TensorFlow framework.

The image-classification application comprised a complete set of ImageNet images.

Evaluation Results

Ease of use

Switching from CPU/GPU to Zebra running on an ALVEO™ board deployed for inference was surprisingly simple, carried out via a single Linux command. No FPGA tools or knowledge was necessary.

During the evaluation, eight neural networks were executed without any changes, proving that the Zebra/ALVEO U200 is the most versatile FPGA solution for neural networks.

The application was based on TensorFlow workflow. No change was required to the application, nor to the neural network, nor to the training, making the transition effortless and free from engineering resources.

The quantization used did not require any user intervention, which emphasizes the ease-of-use of Zebra.

Performance

Eight popular neural networks were processed during the evaluation. The performance of each of them when run on the Zebra mapped on an ALVEO U200 installed in a Dell PowerEdge R740/R740xd servers is charted in figure 8.

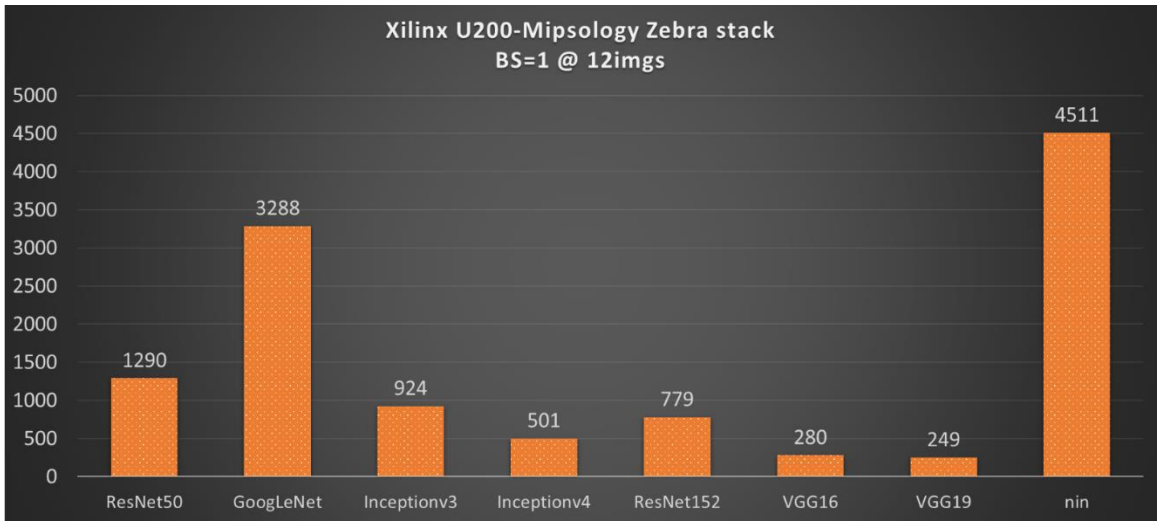


Figure 8. Comparing deep learning models using Mipsology Zebra stack.

Accuracy:

Running the complete set of images from ImageNet, the evaluation showed a difference of accuracy of less than 1%. For some of the NN the results were better than using FP32. However, it should be noted that the purpose of the evaluation was not to reproduce the best accuracy results for each network, rather to compare Zebra against a CPU/GPU solution running FP32 with no changes to the networks.

Table I summarizes the accuracy achieved by Zebra based on int8 computations (obtained from FP32 training and using Zebra quantization) compared to the accuracy obtained by a GPU/CPU platform based on FP32 computations. The results were obtained with the same networks and the same application, without modifications.

Neural Network	Zebra (int8)		GPU/CPU (FP32)		Difference	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
CaffeNet	53.3%	75.4%	53.4%	77.0%	-0.1%	-1.6%
GoogLeNet	62.8%	84.0%	62.4%	84.8%	+0.4%	-0.7%
InceptionV3	68.1%	88.5%	68.8%	89.4%	-0.7%	-0.9%
InceptionV4	72.2%	90.8%	72.9%	91.5%	-0.7%	-0.7%
ResNet50	69.6%	88.6%	70.0%	89.4%	-0.4%	-0.8%
VGG16	65.8%	85.9%	65.4%	86.3%	+0.4%	-0.4%
VGG19	66.2%	86.2%	65.8%	86.7%	+0.4%	-0.5%

Table I. Summary of accuracy of results obtained in the evaluation

Conclusions

The purpose of the evaluation of PowerEdge R740/R740xd Dell servers hosting a Xilinx ALVEO™ U200 boards configured with Mipsology Zebra was two-fold:

- Estimate the throughput of the setup in accelerating a set of images from ImageNet without impact on the accuracy of the computation.
- Establish the ease-of-use of Zebra deployment.

Both objectives have been met successfully. On the one hand, the setup produces an acceleration platform for neural network inference. On the other hand, Zebra does not require changes to the neural network, does not need re-training of the NN, and it operates within the popular frameworks without any extra work. Zebra conceals the FPGAs, paving the path for data scientists and engineers to use their knowledge, talent, and expertise without spending efforts necessary with alternative products.

References

- Mipsology - https://mipsology.com/
- Dell PowerEdge R740/R740xd - https://www.dell.com/en-us/work/shop/povw/poweredge-r740
- Dell EMC Accelerator site: https://www.dellemc.com/en-us/servers/server-accelerators.htm