RWTH AACHEN UNIVERSITY

INTERNSHIP REPORT

Industry internship at Enclustra FPGA Solutions, Räffelstrasse 28, CH-8045 Zürich

presented by
Arthur Ruder
Matr.-Nr. 310697

Supervised by Jelena Dragas

Aachen, 06.09.2019

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
APU	Application Processing Unit
ASIC	Application Specific Integrated Circuit
AXI	Advanced eXtensible Interface
CPU	Central Processing Unit
DNNDK	Deep Neural Network Development Kit
DSP	Digital Signal Processor
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
IP	Intellectual Property
MPSoC	Multiple Processor System on Chip
OS	Operating System
PCIe	Peripheral Component Interconnect Express
PL	Programmable Logic
PS	Processing System
RPU	Real-time Processing Unit
SDR	Software Defined Radio
SoC	System on Chip
USB	Universal Serial Bus

1. Week

The first two days of the week were spent getting to know all colleagues and familiarize myself with internal processes and guidelines. Zurich is the headquarters of Enclustra GmbH and therefore the majority of hardware and software design is being done here. Around forty people, most of which are hardware and software engineers, work in the Zurich office. The company itself is divided into two areas, *Field Programmable Gate Array (FPGA)* Design Center and FPGA Solution Center. The former is offering customer-specific design services implementing applications on FPGAs and providing support and custom *Intellectual Property (IP)* components. Areas of expertise include wired networks and switching, wireless communications (*Software Defined Radio (SDR)*), smart cameras, embedded interfaces (*Peripheral Component Interconnect Express (PCIe)*, *Universal Serial Bus (USB)*, *Advanced eXtensible Interface (AXI)*, ethernet, etc.), test and measurement (sensors, data acquisition, *Digital Signal Processor (DSP)*) and drive/motion control. The latter designs custom FPGA/*System on Chip (SoC)* modules and IP solutions. Several base board families and FPGA module families are developed and supported which can be adapted to the needs of the application by offering different performance key points. Reference designs for each combination of base board and module are provided as a starting point for customers. Furthermore, my task was to do market research on artificial intelligence and artificial intelligence on FPGAs especially. The four key platforms for artificial neural network applications are shown in 1.1. A qualitative design trade-off is shown on the x- and y-axis in terms of power efficiency and performance versus flexibility and ease-of-use. As Enclustras focus is on the embedded market, the market survey has been mainly on *Graphics Processing Units (GPUs)*, FPGAs and *Application Specific Integrated Circuits (ASICs)* as full blown *Central Processing Units (CPUs)* are too inefficient for embedded applications. Possible competitors as well as toolkits provided by FPGA manufacturers such as Intel, Xilinx and Lattice have been evaluated. The results have been presented in a meeting in which a discussion has been held, where Enclustras products and services can fit. One of

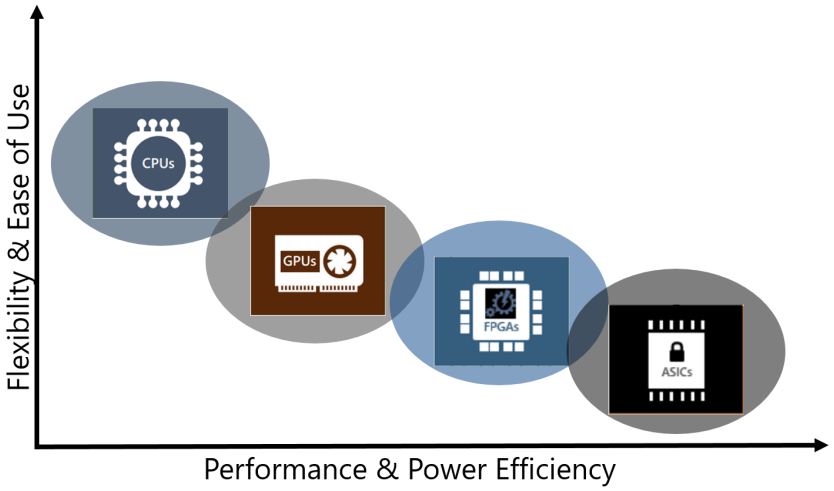

Figure 1.1.: Hardware platform overview

the results was to start planning an *Artificial Intelligence (AI)* demonstrator using Enclustra hardware. The purpose of this demonstrator was to showcase machine learning applications running on Enclustra hardware. As a preliminary step it was decided to check the Xilinx *Deep Neural Network Development Kit (DNNDK)* samples on an evaluation board, the ZCU 104.

2. Week

At the beginning of the week a task unrelated to AI was given to check upon internal documentation and customer support. Together with another recently hired employee, a day of out-of-box testing was scheduled. An Enclustra base board (Mercury+ PE1-400) together with a fitting FPGA module (Mercury+ XU1) featuring a Xilinx *Multiple Processor System on Chip (MPSoC)*. Some details will be given for this specific FPGA family as the Zynq-7000 SoC and the Zynq-MPSoC Xilinx product family are unique in the way dedicated ARM processors are combined with traditional FPGAs. A high level overview is shown in figure 2.1. It shows the difference between

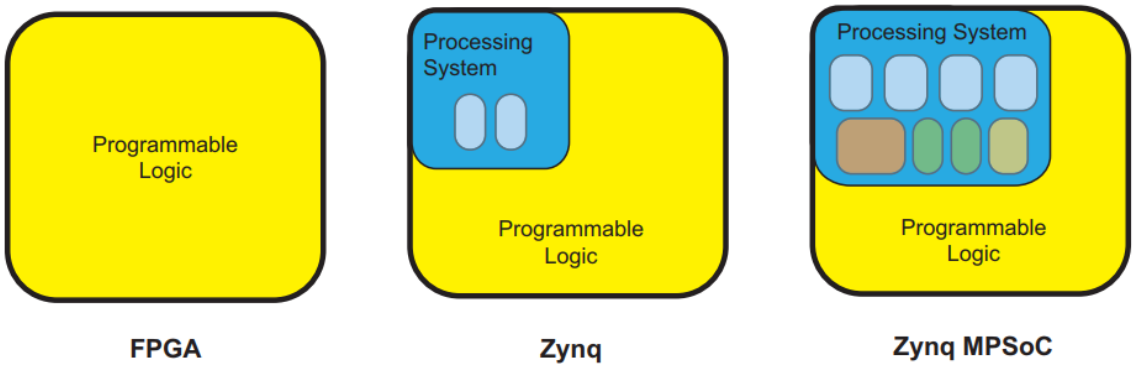

Figure 2.1.: High level ZYNQ family overview [1]

traditional FPGAs and the Zynq product family. The main benefit here is the division between *Processing System (PS)* and *Programmable Logic (PL)*. This allows to combine the benefits of traditional processing units with the flexibility of FPGAs. Custom logic and all peripheral devices can be implemented in the PL part, while system control and even complete *Operating Systems (OSs)* (such as embedded Linux) can be done in the PS. As this system is completely integrated in one package, the communication between the two fabrics is extremely fast and can be done using AXI interfaces. The MPSoC family even integrates several different types of processing units, *Application Processing Units (APUs)*, GPUs, ***VPU!**s (**VPU!**s)* and *Real-*

time Processing Units (RPUs). The Enclustra module uses an MPSoC Xilinx FPGA and our task was to go through the whole process of bringing up the base board together with a corresponding module to test customer experience using the provided documentation, user manual and reference design. The goal was to find unclear instructions in the documentation and provide feedback as to the overall experience bringing up the hardware out-of-the-box. First, the hardware reference design was loaded in the Vivado Design Suite and the bit stream generated. After, the hardware description file was exported so it can be used using the Xilinx ***SDK!*** *(**SDK!**).* This allows to create applications in C/C++ against the custom hardware design. All of the provided sample applications have been tested and verified. Some unclear instructions were identified and discussed with the employee in charge to improve customer experience. The rest of the week was spend updating the internal Wiki page for AI. Furthermore, the DNNDK sample applications were tested on the ZCU 104 evaluation board. The provided examples include several state-of-the-art neural networks demonstrating key applications for neural network inference, such as image classification, face detection, object detection and pose detection. As only the image classification example worked directly for this particular evaluation a fix needed to be found. Another task was to introduce the topic of AI to the whole company as *Artificial Neural Networks (ANNs)* was a completely new design field for a majority of the technical staff. Two PowerPoint presentations should be prepared, namely 'Introduction to AI' and 'Introduction to AI on FPGA'. I started with the preparation of the first one in parallel with finding a bug fix for the other DNNDK sample applications, as these should be part of the second presentation.

3. Week

The main focus of this week was research and starting to layout the first presentation. It was assumed that the audience is technology savvy but has no particular background in AI. Thus, the presentation had to introduce the whole field and key concepts that enabled the rise of AI applications in recent years. The first draft of the presentation was discussed in a meeting and some changes were made to the overall structure, the content and the degree of complexity.The rough structure of the presentation is as follows:

- **Motivation**: To get the viewers interest it was shown that AI applications are already part of daily life for everyone. This was achieved by showing that all of the major companies such as Google, Apple, Facebook, Microsoft as well as Tesla, Netflix and Amazon use AI in their datacenters and products and allocate huge resources to AI research. The importance of AI was further enhanced by showing the rapid growth of annually published AI papers and startups developing AI systems. The trend from 1995 to 2015 resembles almost exponential growth in AI research and products.

- **Definition**: As AI has become such a buzz word in media a definition of the term was needed and what part of AI is actually used in all of the common applications. ANN that perform typical computer vision and language processing tasks are all part of ***ML! (ML!)***, which is a subset of AI. **ML!** itself can then be divided into further subsets using roughly three learning methods, supervised learning, unsupervised learning and reinforcement learning. As supervised learning is the most commonly used method, the presentation focused on this method used to train ANN. Furthermore, the different parts that comprise a ANN are introduced, namely the neuron and how neurons are formed into layers. These layers are then stacked together to form an ANN.

- **Key concepts**: An explanation of supervised learning was given with two distinct examples, linear regression and deep learning neural networks to

illustrate the idea behind supervised learning: predict a value y given an input x by deploying a function $f(x)$. This function $f(x)$ is acquired by deploying a learning algorithm and usage of a so called training set, consisting of input pairs x and y. The concept of inference and training were also explained with an emphasis on inference. Once a network is trained, only inference needs to be run, so this is the crucial part application wise.

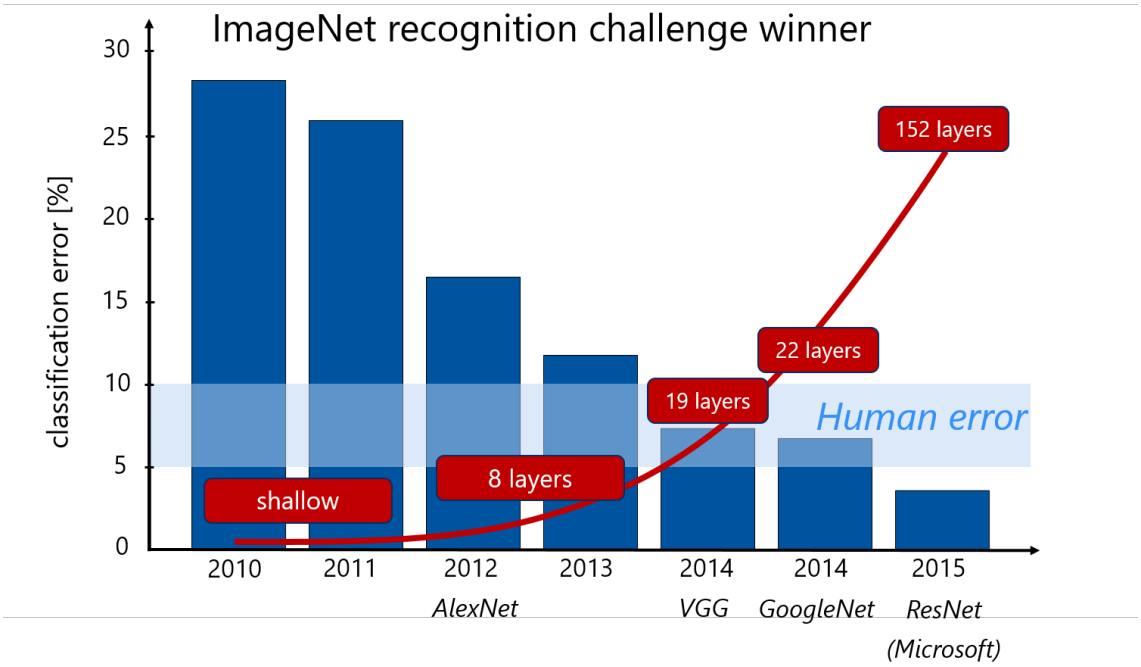

Figure 3.1.: ILSVRC winners

- **Deep Learning**: Figure 3.1 shows the winners of the state-of-the-art ImageNet Large Scale Visual Recognition Challenge. It illustrates that the breakthrough in performance came not only from more sophisticated networks, but mainly from stacking different kinds of layers deeply, hence the name. The presentation was ended with the question, what hardware is best suited for ANN applications.

Alongside preparing the presentation the error preventing the more sophisticated examples was identified as the board crashed while performing tasks related to video analysis (face detection, pose detection, etc.). At first, it was suspected there were some problems with heat management and using the system monitor the temperature of the FPGA was investigated during operation. As this seemed to be well within allowed borders specified by the Xilinx data sheet, other causes had to be found.

4. Week

At the beginning of the week the first presentation 'Introduction to AI' was held before the technical stuff of the company. The general background and principles of **ML!** have been introduced and an outlook given to the second presentation, which would go more into detail about the actual hardware realization. The rest of the week was spent going through various tutorials provided by Xilinx to familiarize myself with the workflow and the DNNDK toolkit. As the state of tools used for AI applications on FPGA is still in flux, several approaches needed to be evaluated:

- **DNNDK workflow**: Version 2.08 of the toolkit supported only the Caffe neural network training framework and needs a network description file and the trained weights as input. The key component here is the ***DPU!*** *(**DPU!**)* IP core provided by the DNNDK toolkit. This core is integrated via Vivado into the block design of the hardware and can be configured and adjusted for several performance and power profiles.
- **DNNDK *SDSoC! (SDSoC!)***: Another option is to abstract away the whole Vivado block design process and use Xilinx **SDSoC!** to implement the whole system in a higher programming language, C++. Supported functions can then be flagged as being executed in the PL part of the system. This approach makes using a traditional ***HDL!*** *(**HDL!**)* obsolete and is deemed more accessible. This approach uses the established Xilinx reVision stack for development providing high level ***API!****s (**API!**s)* for computer vision.

Furthermore, the IP core provided by the DNNDK toolkit was studied in more detail. A new base board was in the production state phase and the idea was to have a **ML!** design ready to showcase the capabilities of the new board. Figure 4.1 shows an example hardware design with integrated **DPU!** module. In this example a camera is connected via the ***MIPI!*** *(**MIPI!**)* ***CSI2!*** *(**CSI2!**)* interface to the PS. ***DMA!*** *(**DMA!**)* is usually used in conjunction with an AXI interconnect to communicate with the PS. The captured images are used as the input to the neural

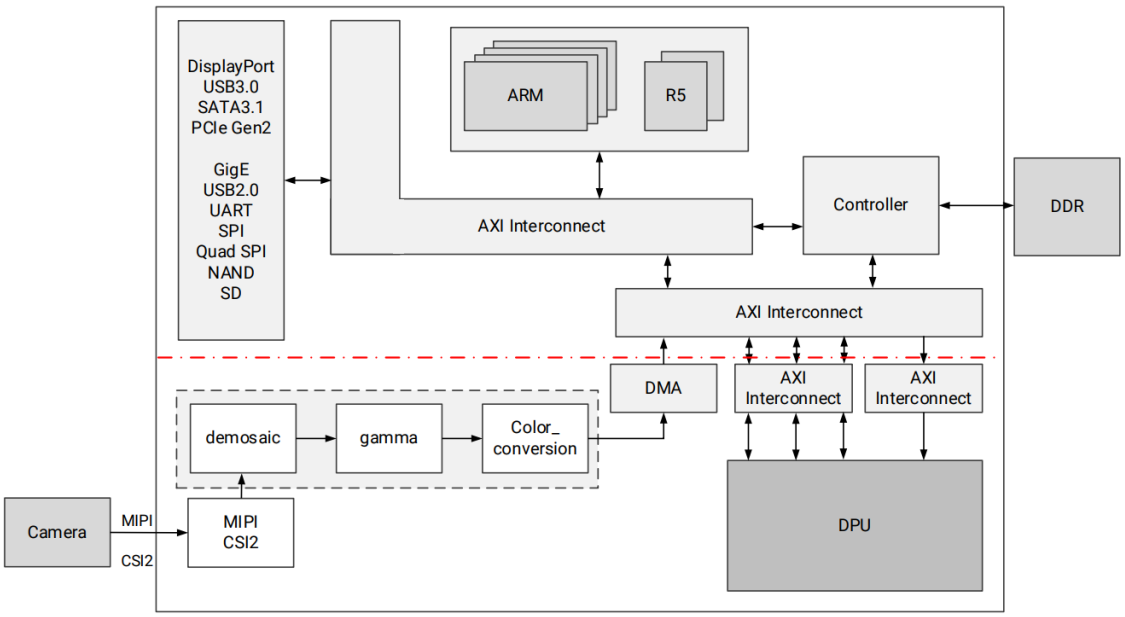

Figure 4.1.: Example system with integrated **DPU!**

network and the **DPU!** itself can be viewed as a co-processor to the PS implemented in the PL fabric.

Bibliography

[1] Loise Crocket et al. *Exploring Zynq MPSoC: With PYNQ and Machine Learning Applications*. Strathclyde Academic Media, 2019.

A. Workday reports