CONTINUITY — EBV Elektronik I An Avnet Company I — INNOVATION

ACCELERATING NEURAL NETWORK DRIVEN IMAGE CLASSIFICATION USING AN FPGA WITH A BINARY NEURAL NETWORK

ACCELERATING NEURAL NETWORK DRIVEN IMAGE CLASSIFICATION USING AN FPGA WITH A BINARY NEURAL NETWORK

Image Classification using a GPU and a Convolutional neural network delivers great performance but also creates some challenges if you want to use this type of machine learning in an edge application like a smart camera. Size, power consumption and long-term availability are a few we will briefly discuss.

We will explain how Xilinx Research created a framework to speed-up and shrink a convolutional neural network so it can fit a small FPGA using a Binary neutral network implementation.

We will explain the implementation in the Zynq UltraScale+ MPSoC and give some details on the used Ultra96 board, which is build according to the consumer 96boards specification.

THE NAME OF THE GAME IS „ACCELERATION“

Breaking the MegaFLOP boundary in 1992

My first accelerator was a Cyrix FPU for my Intel 286 that bolded a 387 in a 287 socket using a special socket that increased the clock going in to the FPU.

Today the FPU that comes for free in the Quad Core ARM Cortex A53 that is used in the delivers above 1 GFLOP

Neural Network Based Image classifcation required performance in TFLOPs though.

MOVING FROM INT16 TO AVX512

PC performance based on Intel CPU‘s and Nvidia GPU‘s

Moore‘s Law is well known and often miss used.

Adding more cores / threads and introducing SIMD options covered the clockspeed reaching it‘s limits.

Let‘s simply take the ALU, FPU and the SIMD accelerators as a reference.

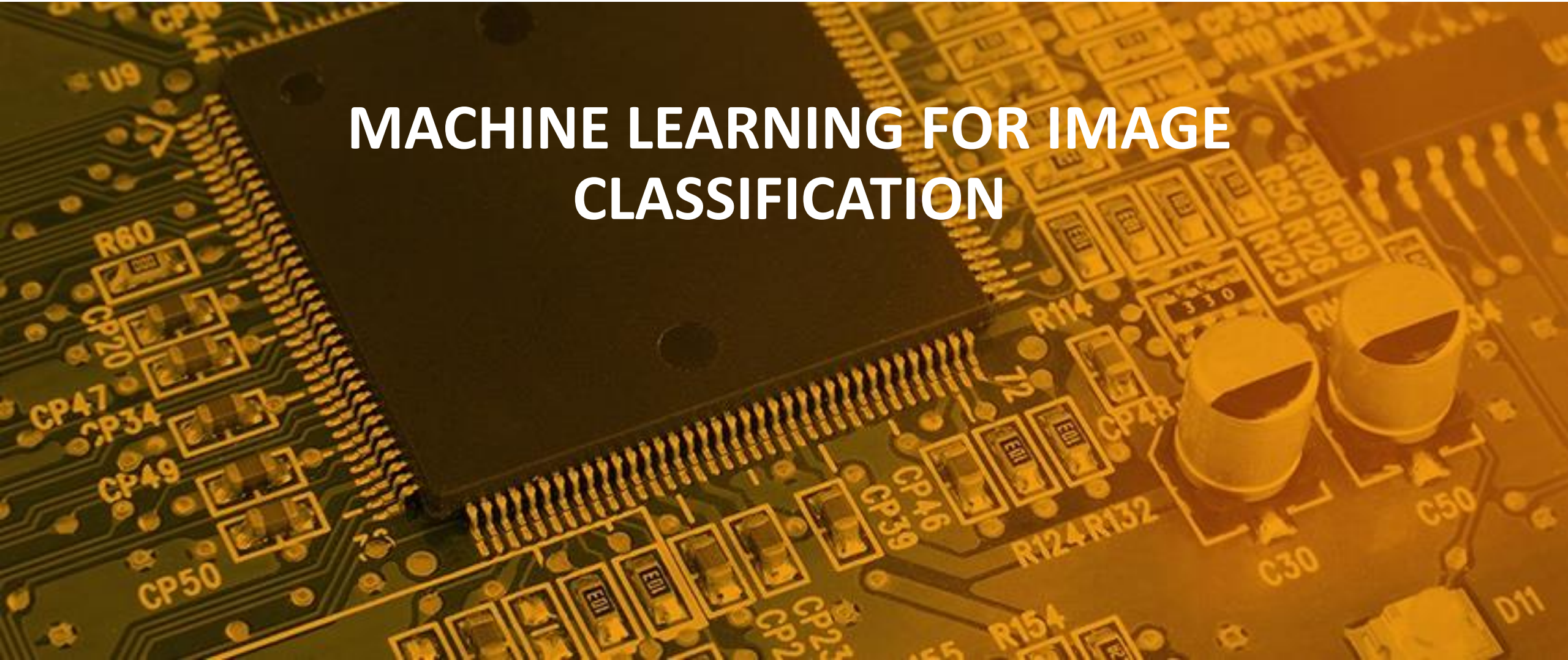
AVX512 is the latest addition and only available in Intel CPU‘s used in the datacenter.
Simular to the High End GPU of Nvidia. Only the datacenter card get the highest end floating point performance.

16 MHz with a single 16-bit ALU to 3 GHz with 512-bit SIMD gives a speed up of 64000 times.
The required acceleration is roughly a milions times though…

EBV Elektronik
An Avnet Company

CONTINUITY — INNOVATION

MACHINE LEARNING FOR IMAGE CLASSIFICATION

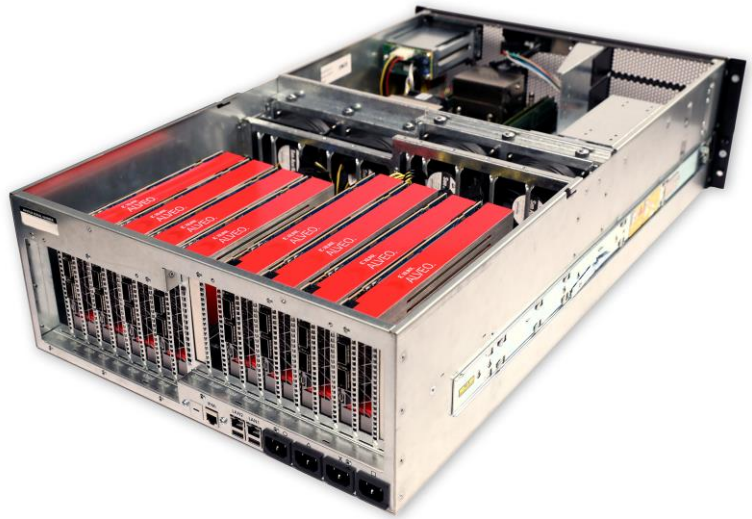


MACHINE LEARNING FOR IMAGE CLASSIFICATION

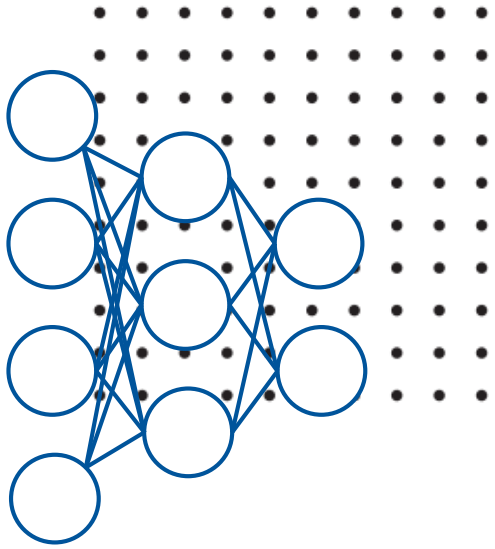
Training is done mostly in the cloud using GPU‘s

Inference now typically requires the edge device to upload the image to the cloud and get metadata returned from the cloud.

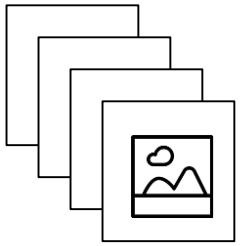
Here FPGA‘s can accelerate by using the Xilinx ML Suite that targets FaaS providers and the Alveo Accelerator cards.



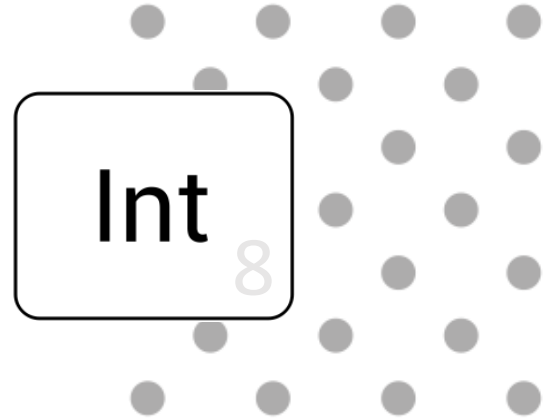
QUANTIZER: GETTING THE NETWORK READY FOR FPGA



1) Provide FP32 network and model



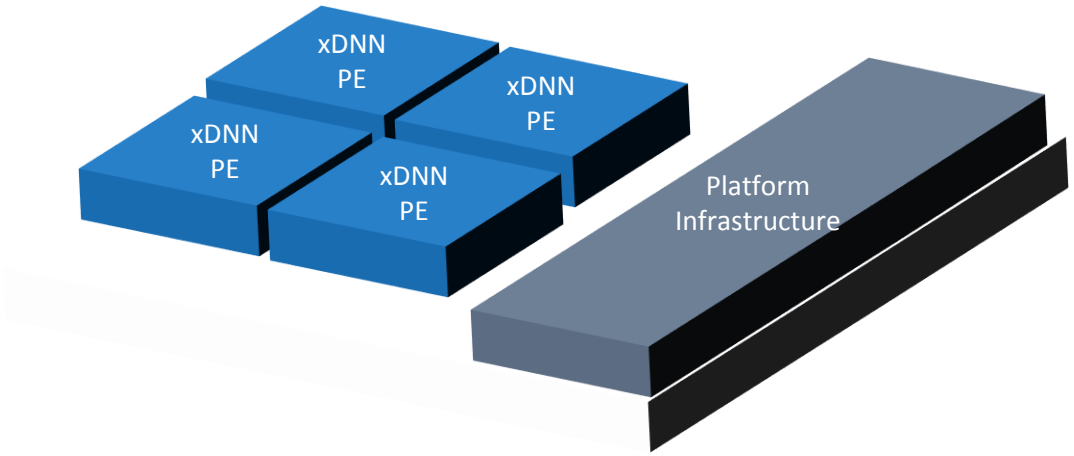
2) Provide a small sample set, no labels required
   - 16 to 512 images



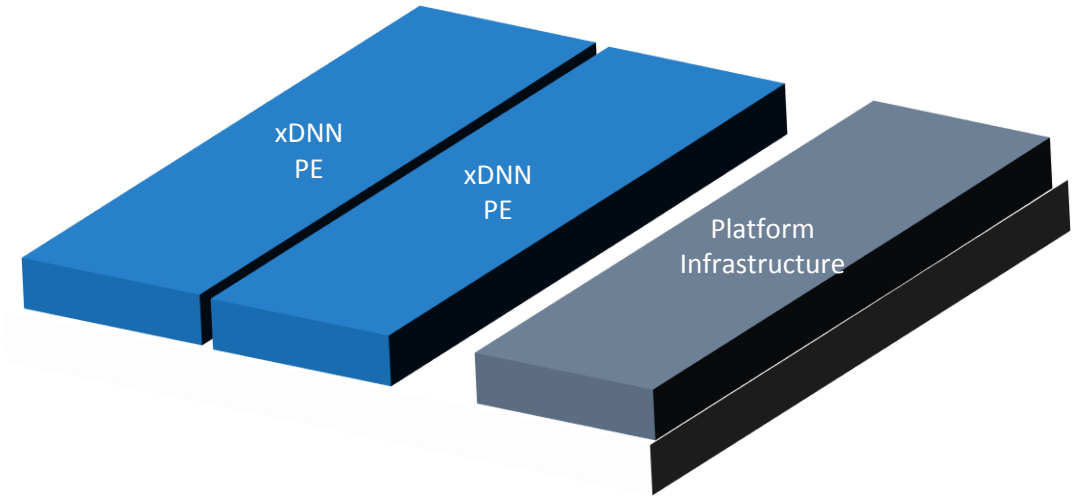
3) Specify desired precision
   - Quantizes to <8 bits to match FPGA's DSP

FPGA PROVIDES ADAPTABLE IMPLEMENTATION OPTIONS

Throughput, Multi-Network Optimized

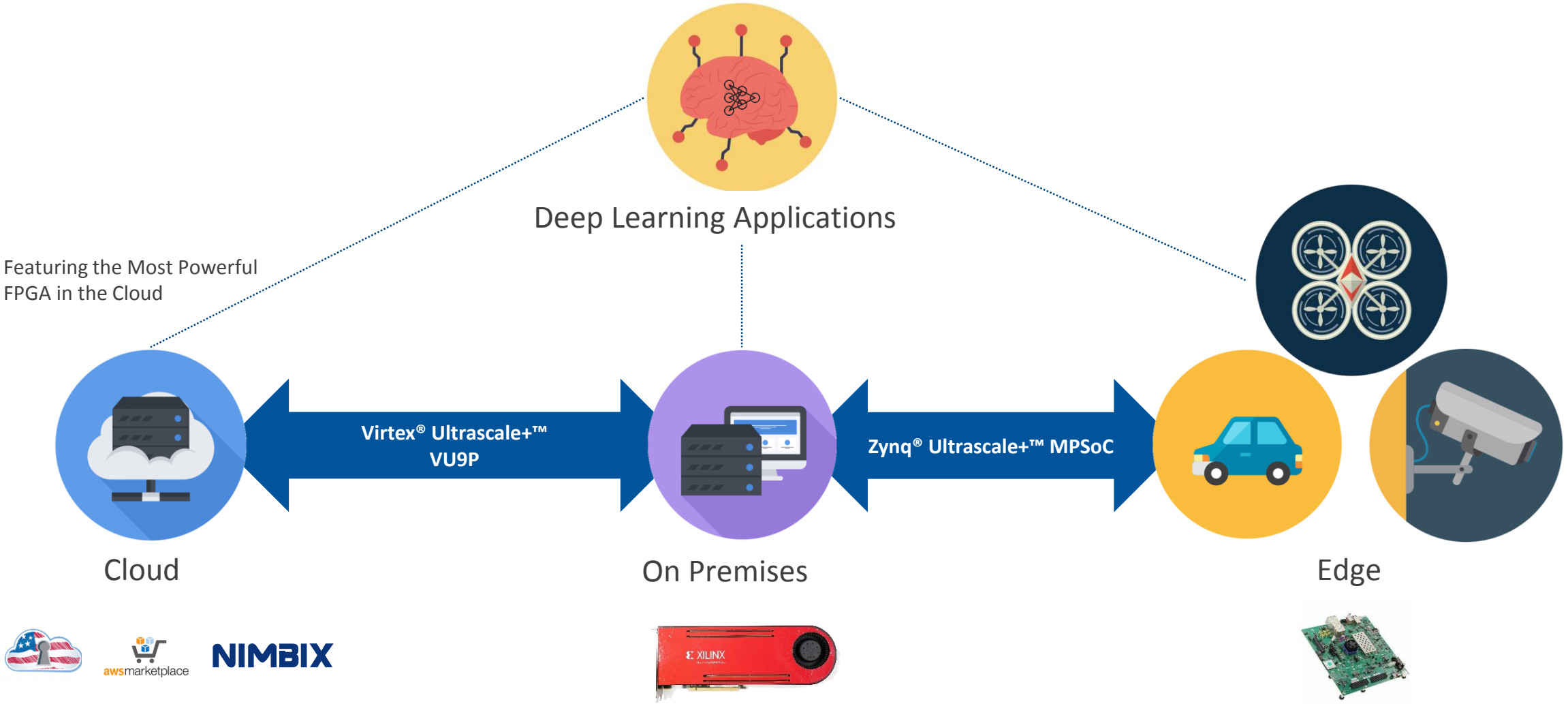


Latency, High Res Optimized



Overlay Name	DSP Array	#PEs	Cache	Precision	GOP/s	Optimized For	Examples Networks
Overlay_0	28x32	4	4 MB	Int16	896	Multi-Network, Maximum Throughput	ResNet50 (224x224)
Overlay_1	28x32	4	4 MB	Int8	1,792	Multi-Network, Maximum Throughput	ResNet50 (224x224)
Overlay_2	56x32	1	5 MB	Int16	1,702	Lowest Latency	Yolov2 (224x224)
Overlay_3	56x32	1	5 MB	Int8	3,405	Lowest Latency	Yolov2 (224x224)

ACCELERATING AI INFERENCE INTO YOUR CLOUD APPLICATIONS



WHY EDGE COMPUTE FOR ML ?

Power

Latency

The number of devices

Volume of the data

Privacy

Security

Connection issues

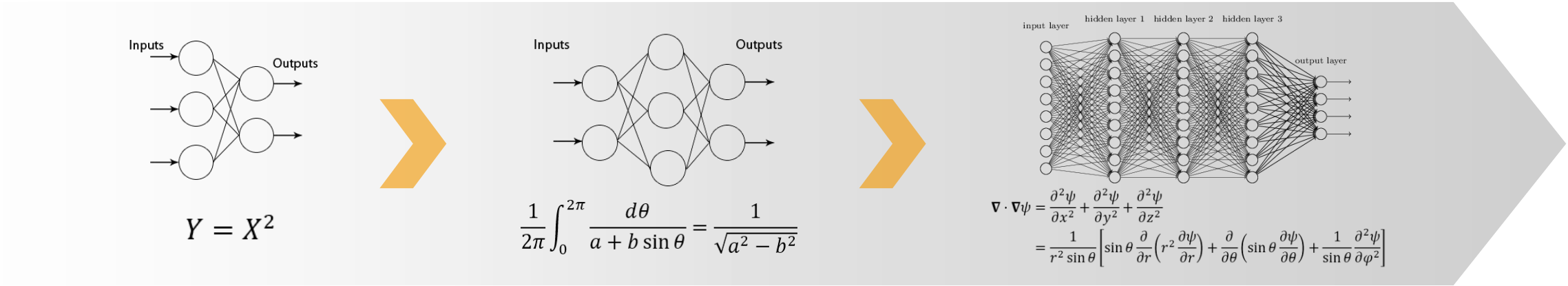
Moving DNN compute to the edge devices requires a lot of optimization for the power budgets and the the performance levels available at the edge.

The communication from the edge to the cloud has many issues as well that come in to play when making devices smart as well.

HOW FPGA'S CAN DELIVER ML TO THE EDGE

Research has shown that through quantization and pruning the neural network the computational intensisty can be lowered drasticly.

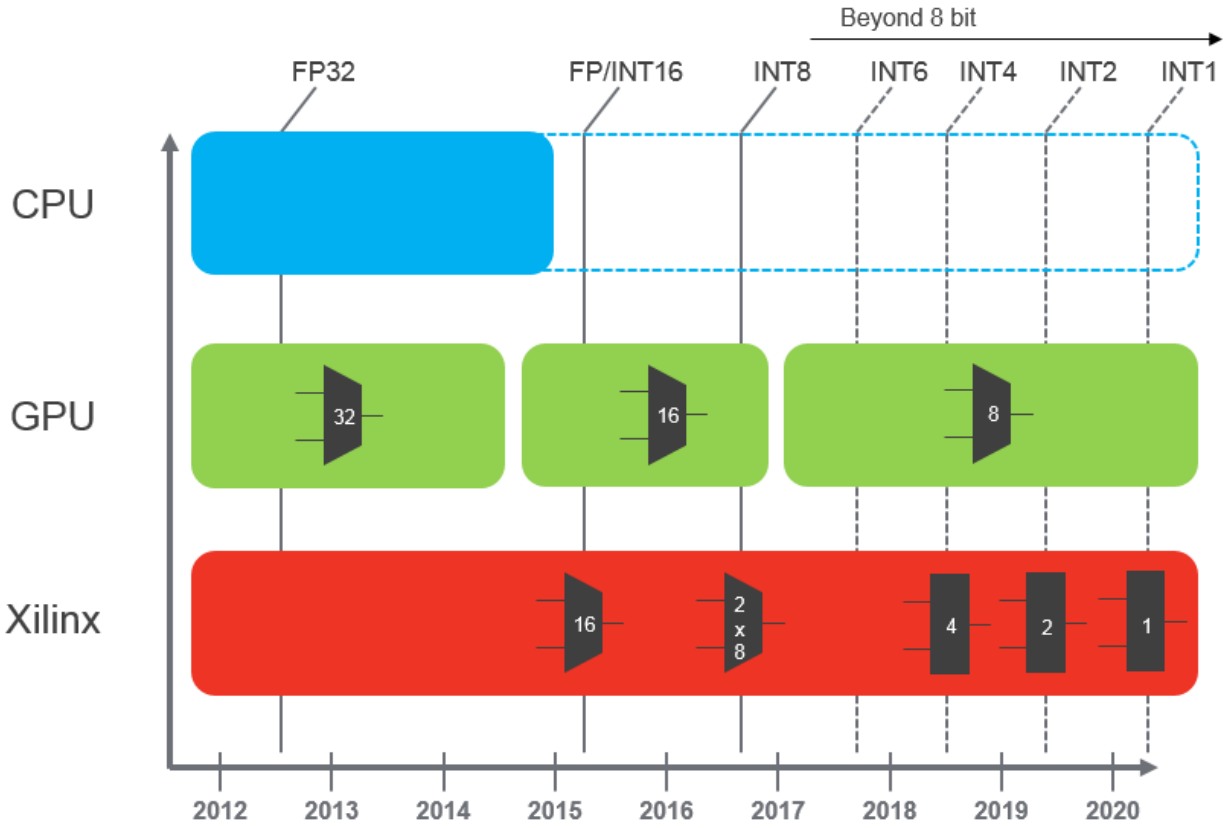
The ML started of with 32-bit floats but is now quicky adapting INT16 and INT8 in order to further reduce the performance requirements

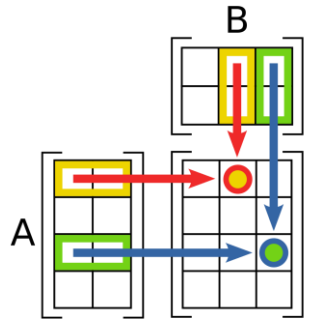


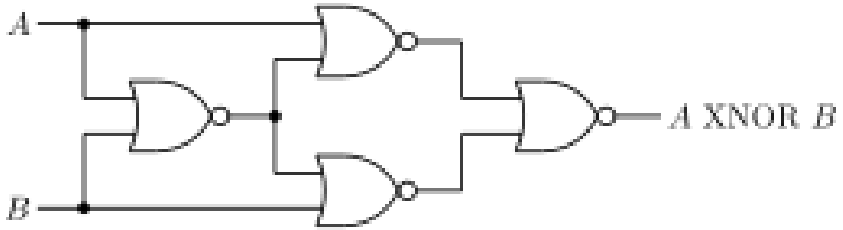
FPGA fabric has the ability to easily adapt from 512-bits math down to 1-bit logic with ease.

Further research in ML has shown that even binary neural networks can keep their accuracy because neural networks can cope with noise quite OK.

UNDOING MOORE'S LAW AND GOING BACK TO BINARY





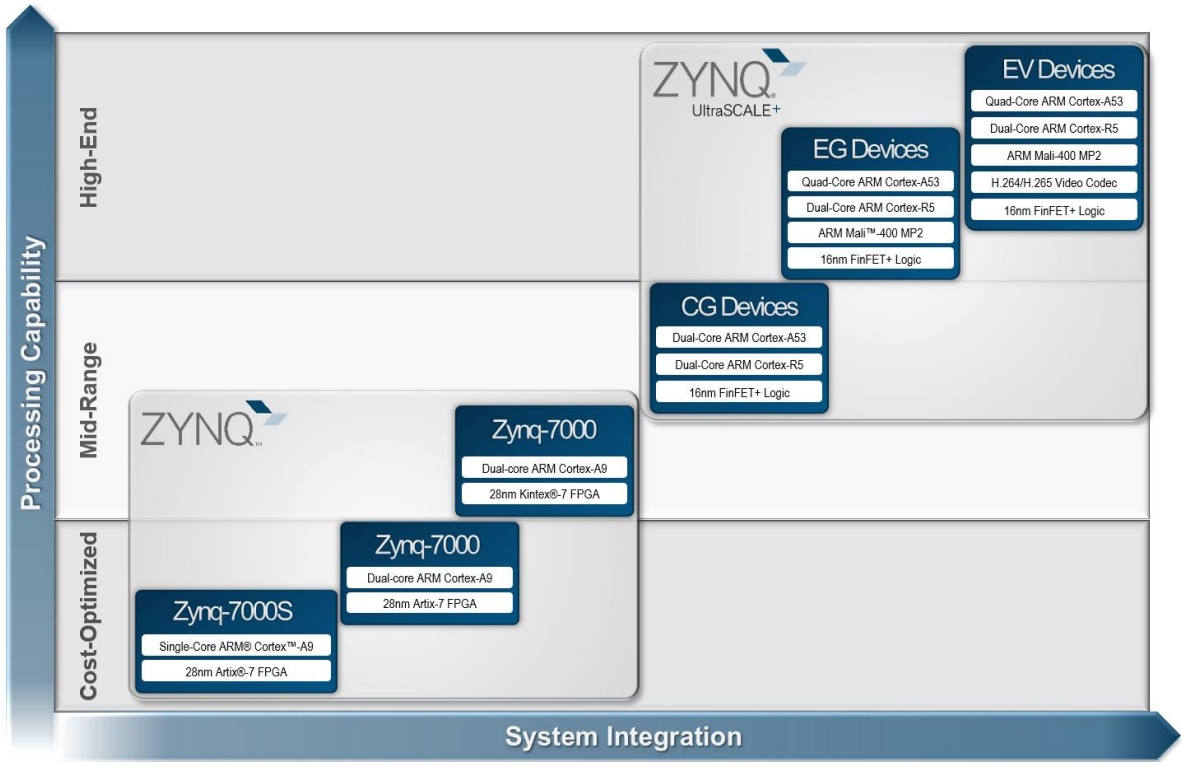


FPGA‘S MAKE ML SCALEABLE

Current solutions use SoC with powerfull GPU or Neural Network processors. They are like a one-size fit all solution. That makes them more expensive, bulky and consume too much power.

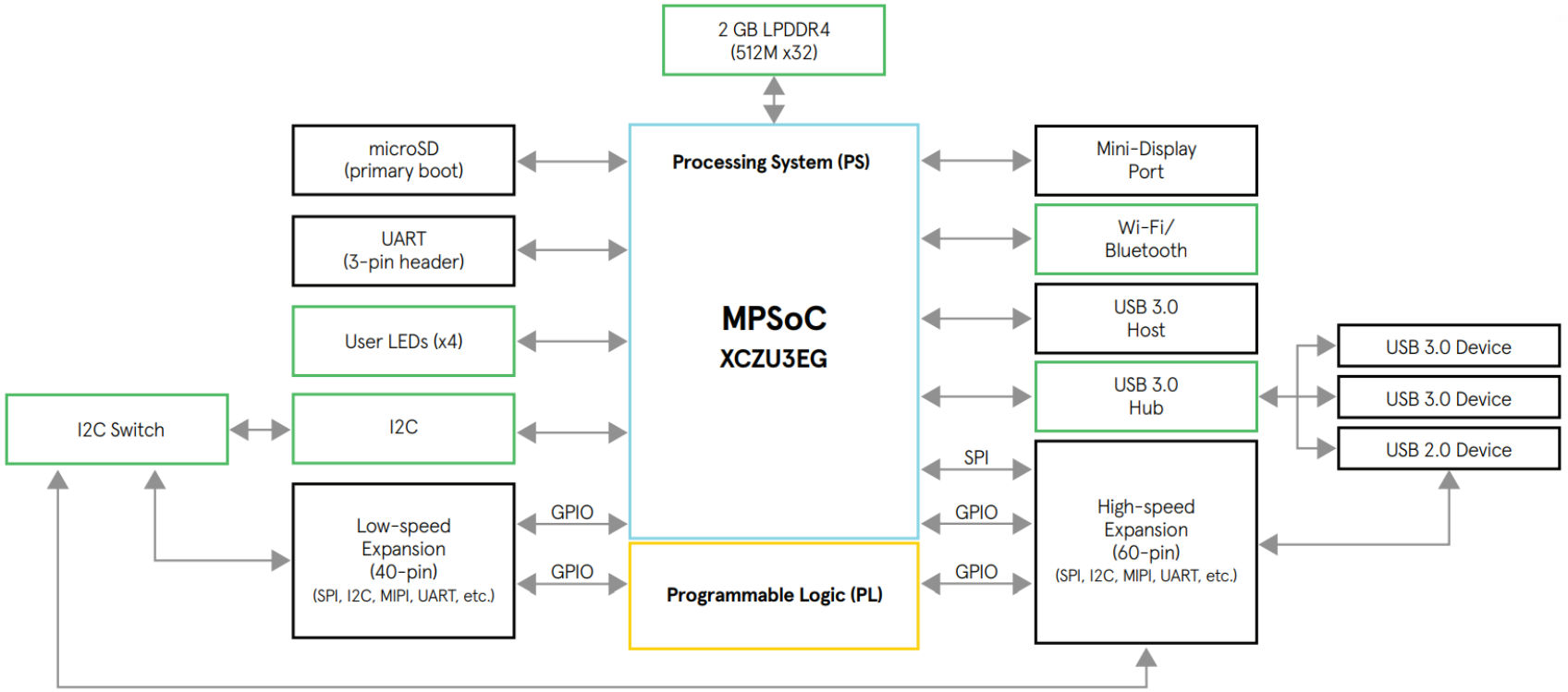
With Xilinx Zynq devices you get embedded processing tailered for your needs with sizeable logic, DSP, internal & external memory sizes and I/O.

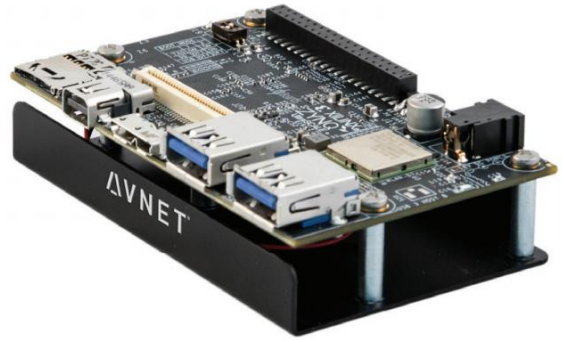
These I/O‘s that fit all kinds of sensor, camera‘s, communication standards and display options.



BRING ON THE ULTRA96

BLOCK DIAGRAM



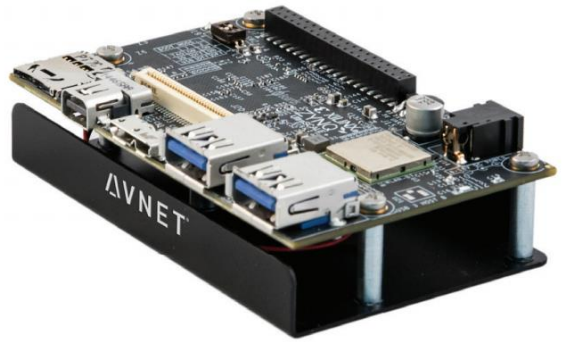


The Ultra96 runs the Accelerated Image Classifier that was build based on BNN research.

The BNN runs in the PL portion.

The images are from file or USB camera.

GERMAN TRAFFIC SIGNS DEMO

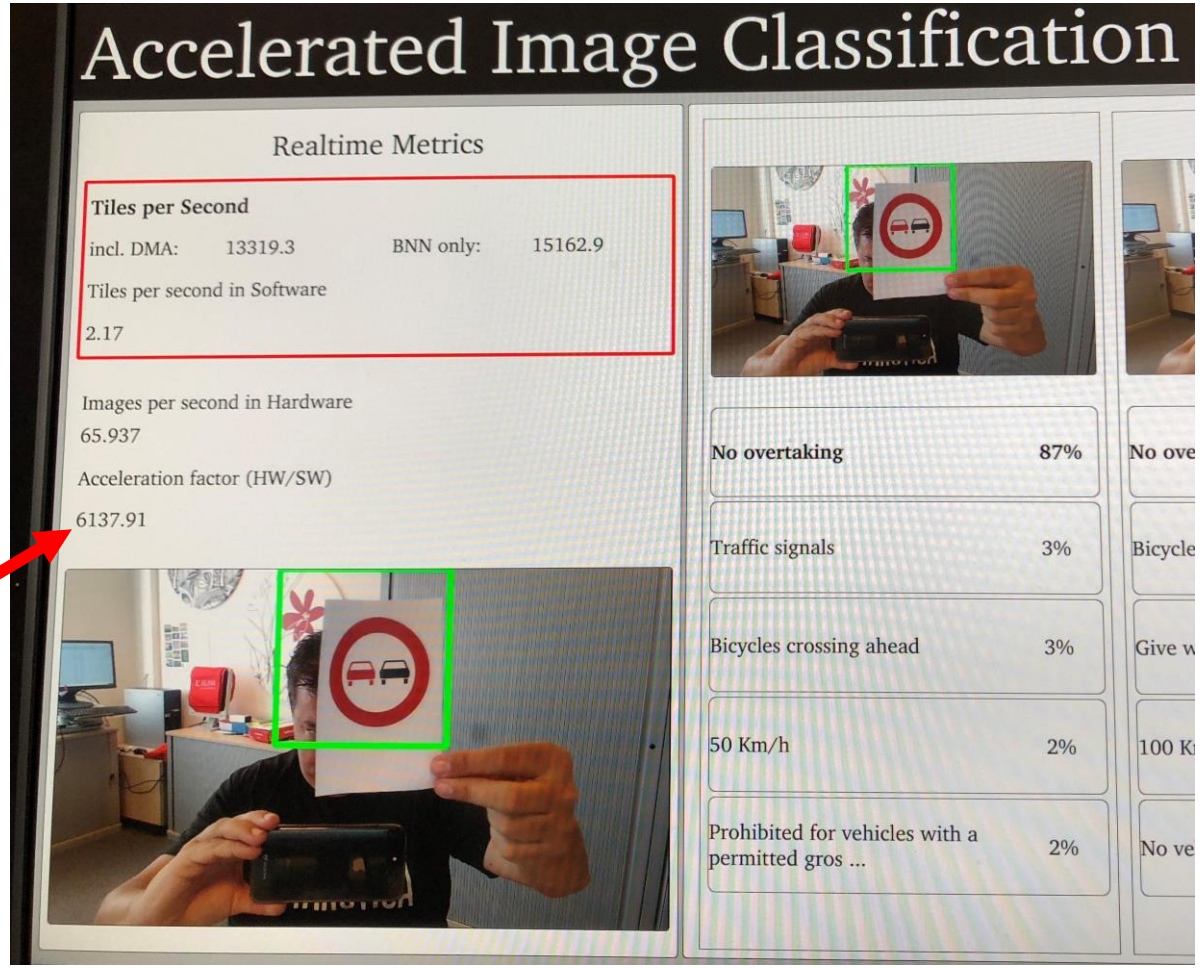


An extreme example of the power of an FPGA. The Quad Core ARM Cortex A53 can process 2.17 tiles per second.

The BNN in the FPGA handles 13319.3 tiles per second including the DMA transfers.

That more than 6000 times faster.

This is no push button example. Many research and training has been done, but it works !



FINN – PYNQ

The demo is part of the FINN framework. Xilinx published all there work based on that and opensourced it on Github

It is now also available for their Python initative where you can use Jupyter Notebook to create an interactive environment for Zynq with Python

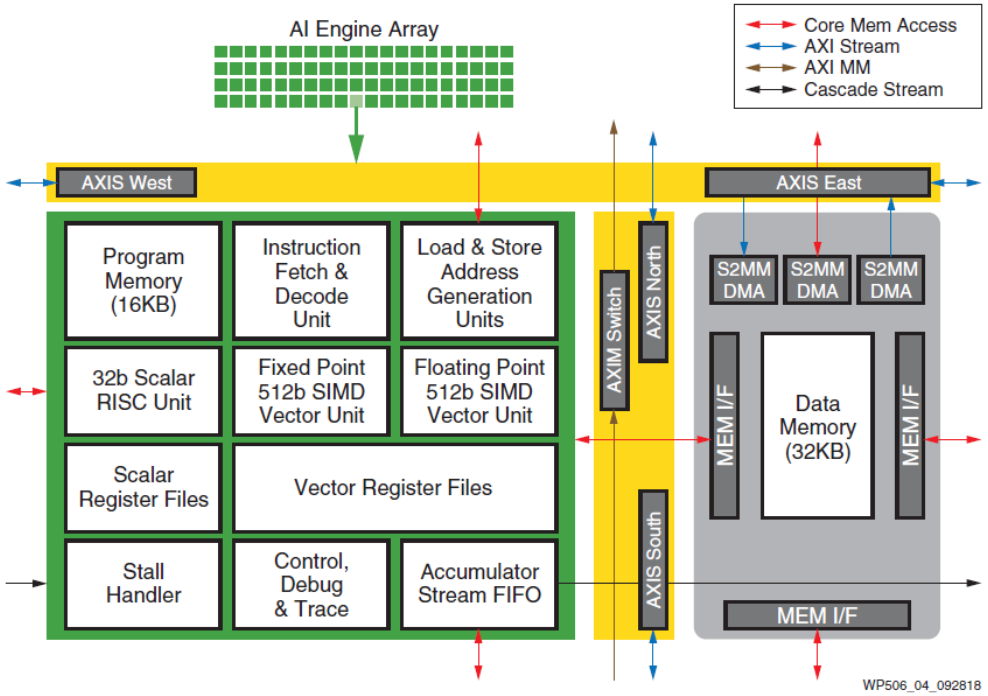
FINN

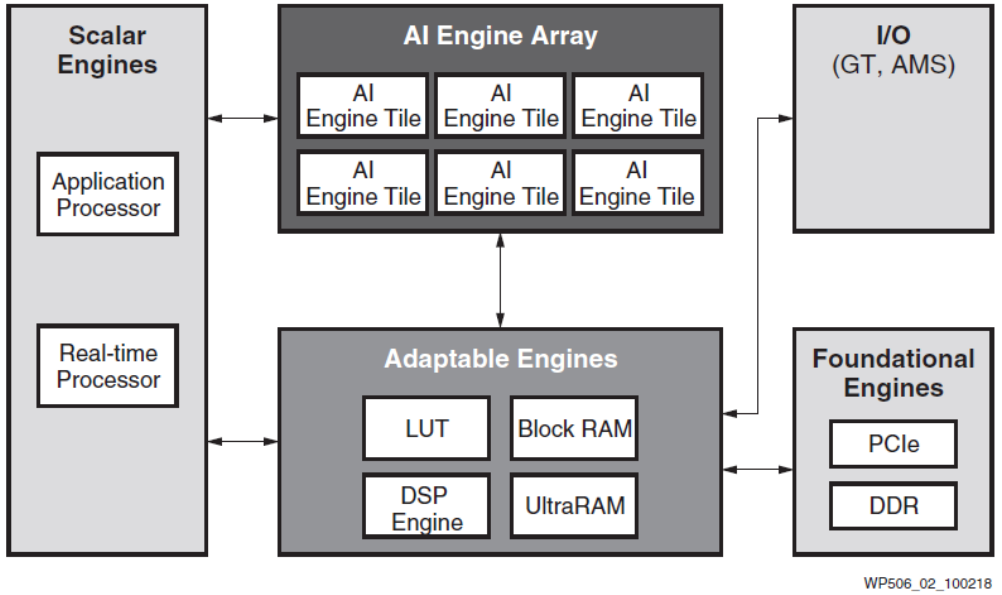
https://github.com/Xilinx/FINN/

PYNQ

http://www.pynq.io/

THE FUTURE – XILINX VERSAL ACAP





The new 7-nm based Xilinx Versal ACAP (Adaptive Compute Acceleration Platform) devices will have large numbers of AI Engines that will be able to perform 128 8-bit MACs per clock cycle that will run above 1 GHz.
Check out the White Paper on this topic here:
https://www.xilinx.com/support/documentation/white_papers/wp506-ai-engine.pdf

EBV FRANCHISE PARTNERS

ams Sensing is life.
BROADCOM
EBV Chips
FIT Foxconn Interconnect Technology
Infineon
IDT
ISSI
maxim integrated
MICROCHIP
Micron
nexperia
NXP
ON Semiconductor
OSRAM
SAMSUNG
SIMCom a SUNSEA AIOT company
life.augmented
TOSHIBA
VISHAY
XILINX

Status: September 2018

EBV Elektronik
An Avnet Company

EBV – THE TECHNICAL SPECIALIST

Inspire Innovation, Expand Capabilities

Applications and Technologies Expertise

International Networking and Ecosystem Partners

Live Demos, Workshops, New Product Updates

Deep Design Support, Technical Advisors

D&E event 2018

Questions ? Ask them now or try the following media later !

- For further questions you can contact me on the following media:
- Email - karl.deboois@ebv.com
- Visit - Planetenbaan 116 3606AK Maarssen-Broek
- Telefone (landline) +31 346 583 031
- Telephone (mobile) +31 6 2242 6907
- Fax - +31 346 583025
- ebv.avnet.com

- Or try Twitter – DutchXilinxGuy or LinkedIn - https://www.linkedin.com/in/karldeboois/


CONTINUITY
EBVElektronik
I An Avnet Company I
INNOVATION
DISTRIBUTION IS TODAY.
TOMORROW IS EBV.