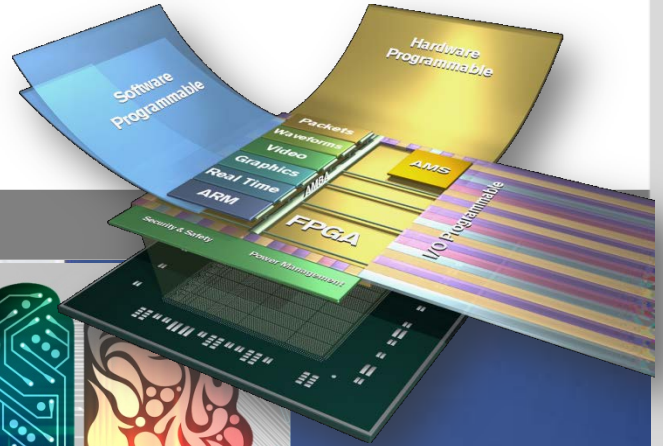
Deep machine learning implementation in FPGA

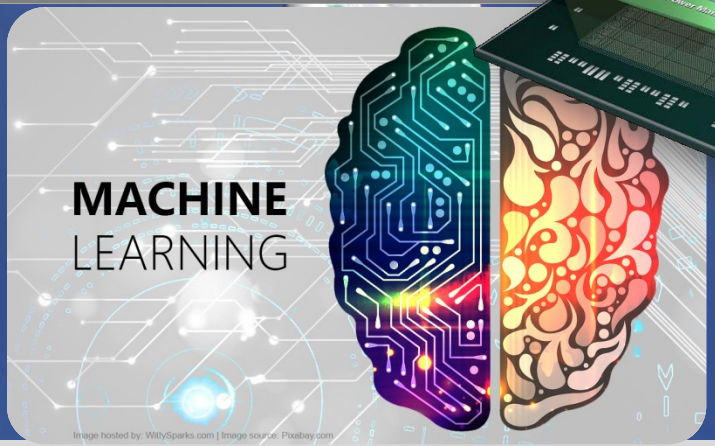
HAP Workshop | Big Data Science in Astroparticle Physics,

RWTH Aachen University, 20th February 2018

M. Caselle, W. Wang and A. Haungs

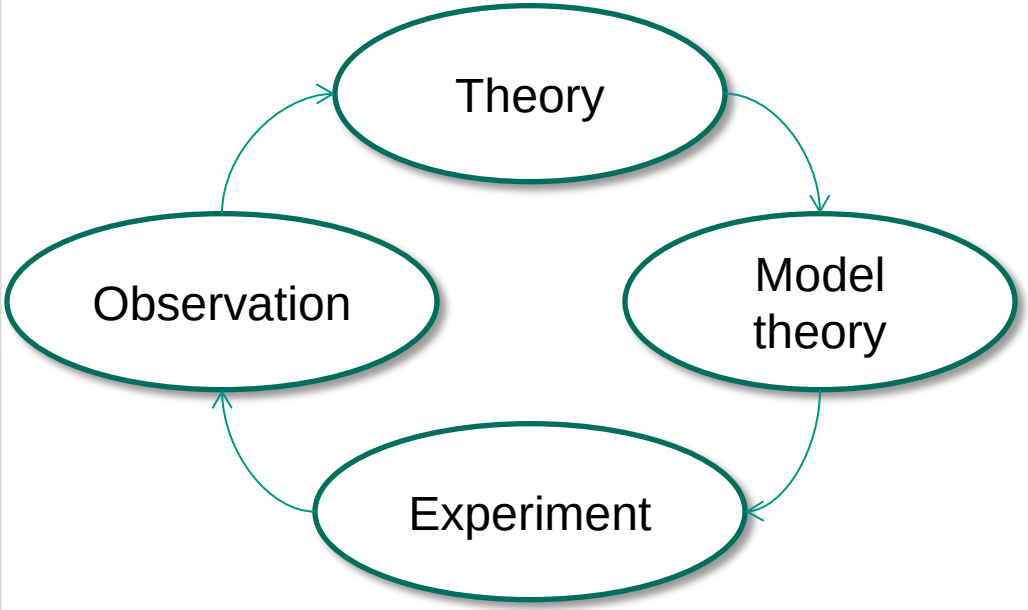
INSTITUTE FOR DATA PROCESSING AND ELECTRONICS (IPE)





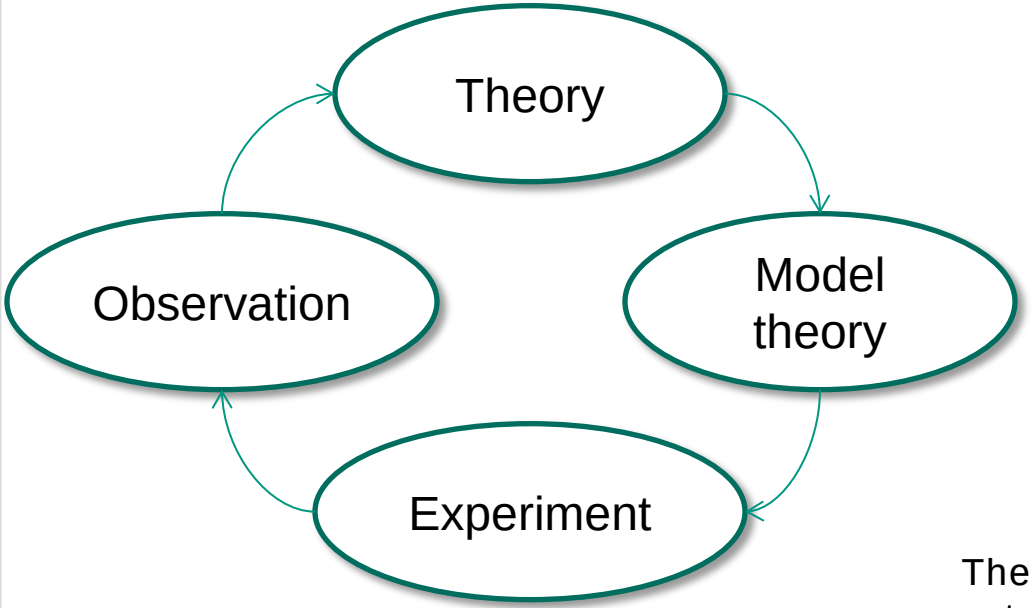
Galileo and Scientific Method

Since Galileo….
Theories allow us to make experimental predictions and vice versa

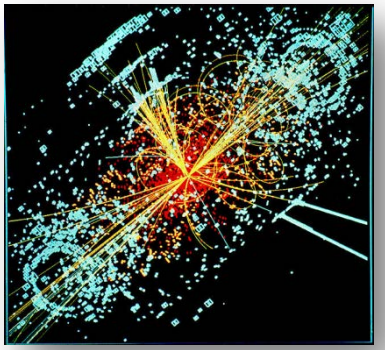


Galileo and Scientific Method

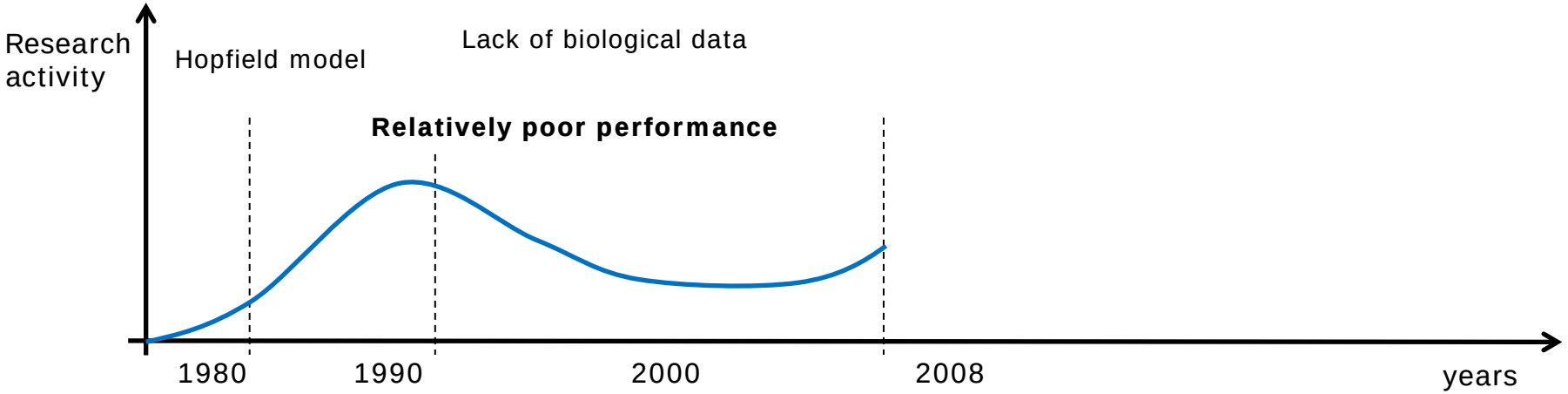
Since Galileo….
Theories allow us to make experimental predictions and vice versa



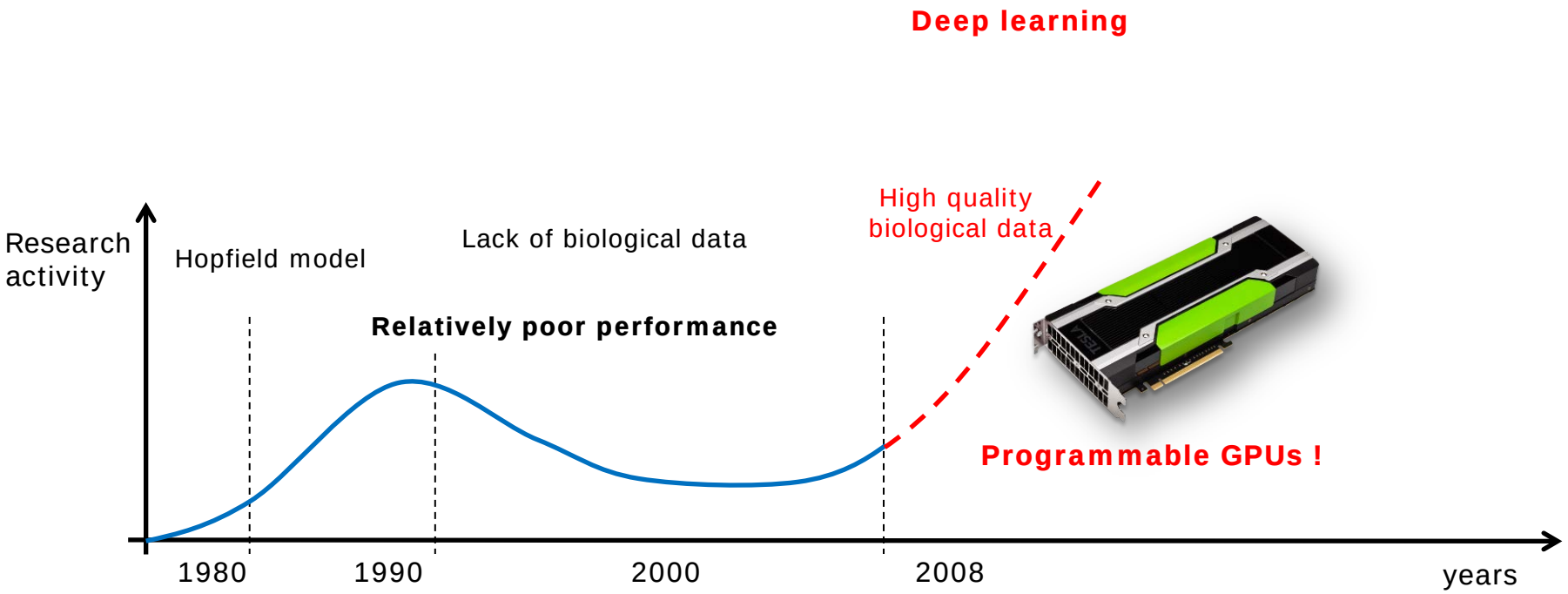
The last big success of the traditional science is the Higgs boson



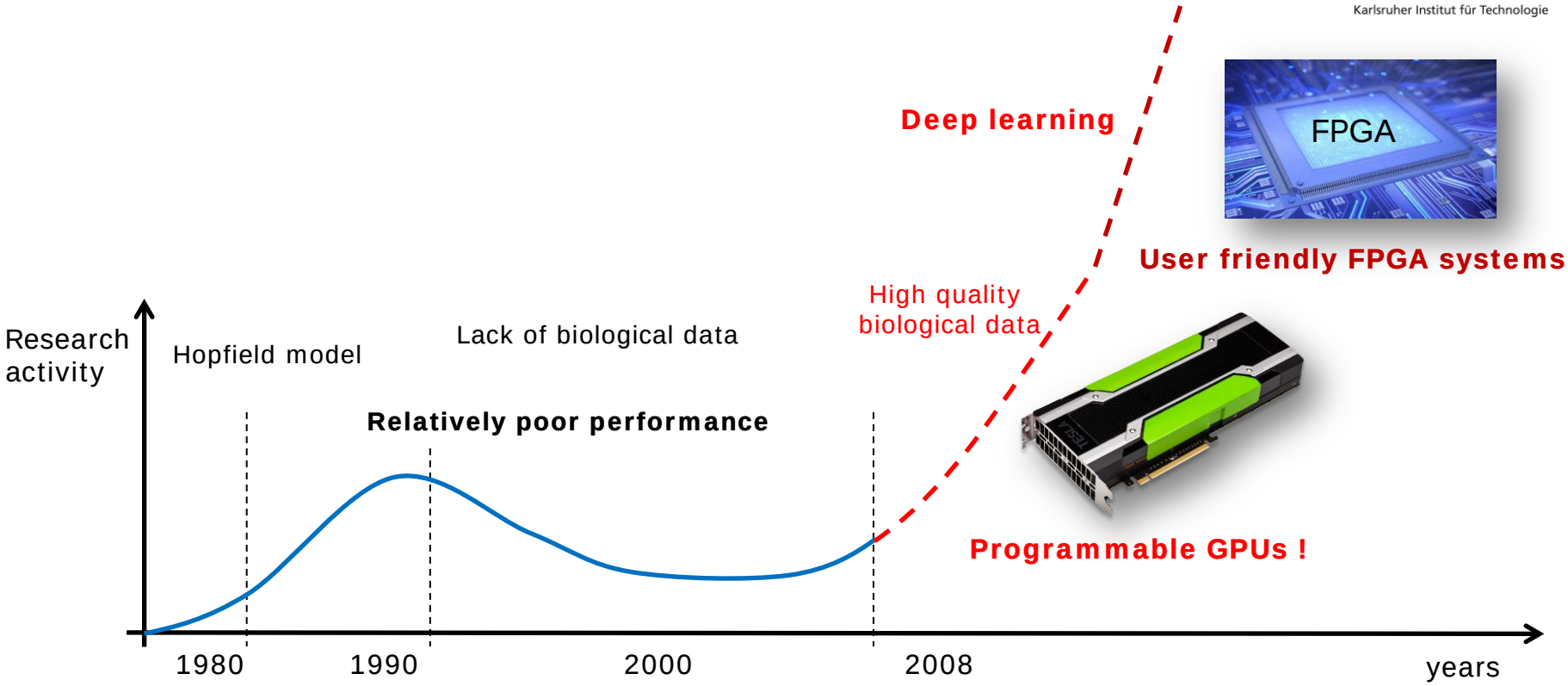
Artificial Intelligence



Artificial Intelligence



Artificial Intelligence



Artificial Intelligence

We are working on heterogeneous FPGA – GPU architectures

Deep learning

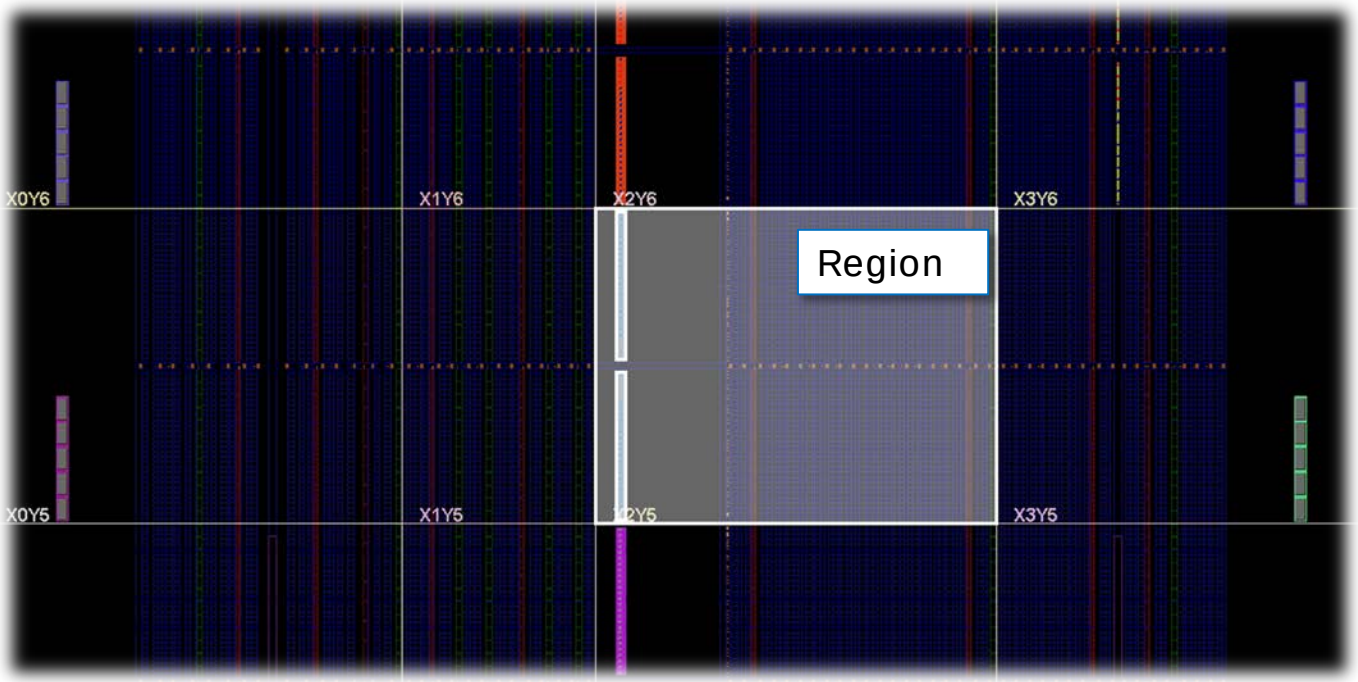
User friendly FPGA systems

High quality biological data

Programmable GPUs !

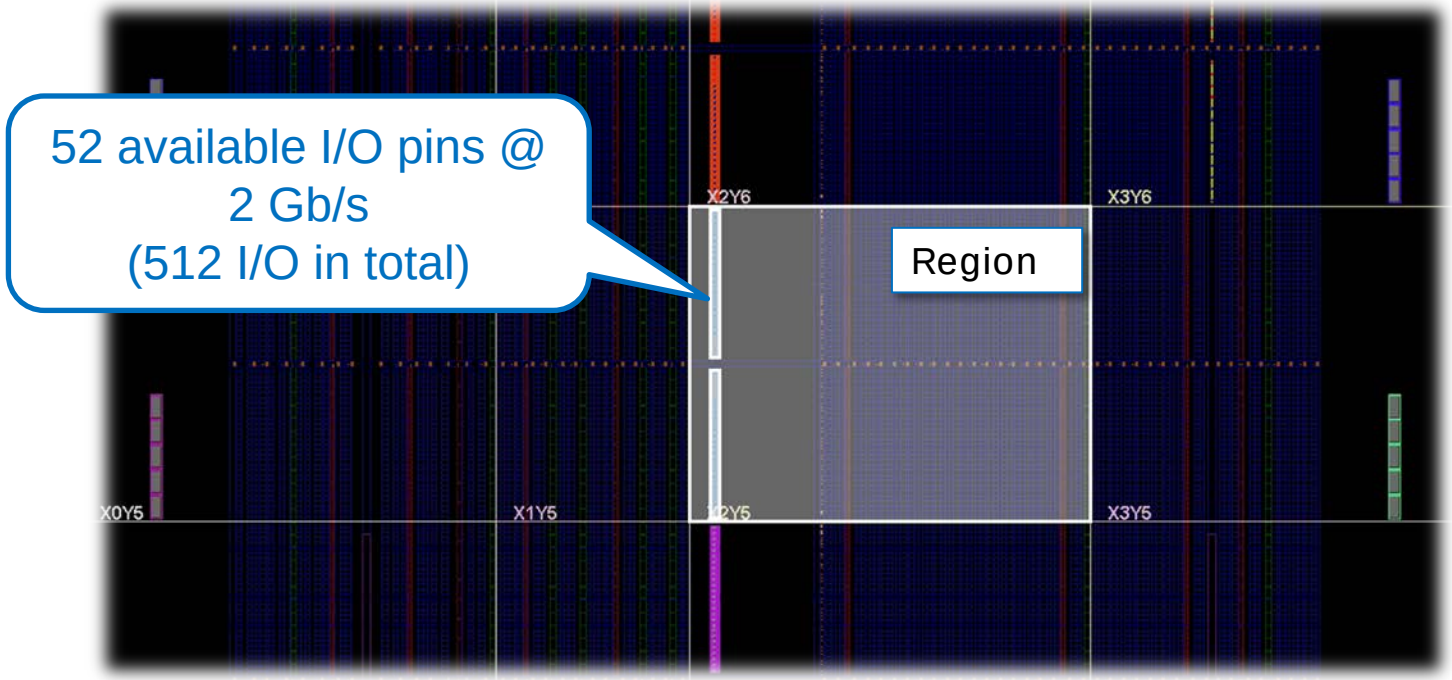
Modern FPGA – ZYNQ Ultrascale+

44 regions in total



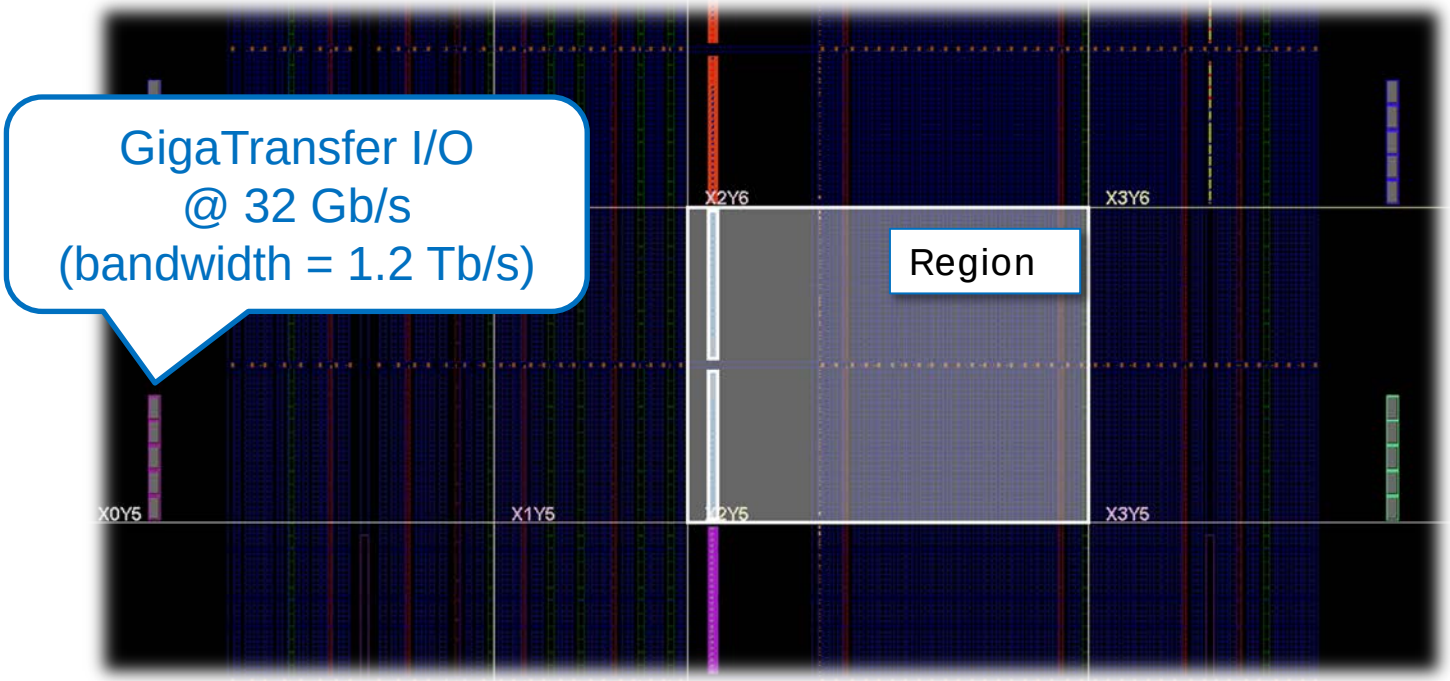
Modern FPGA – ZYNQ Ultrascale+ XCZU11

44 regions in total

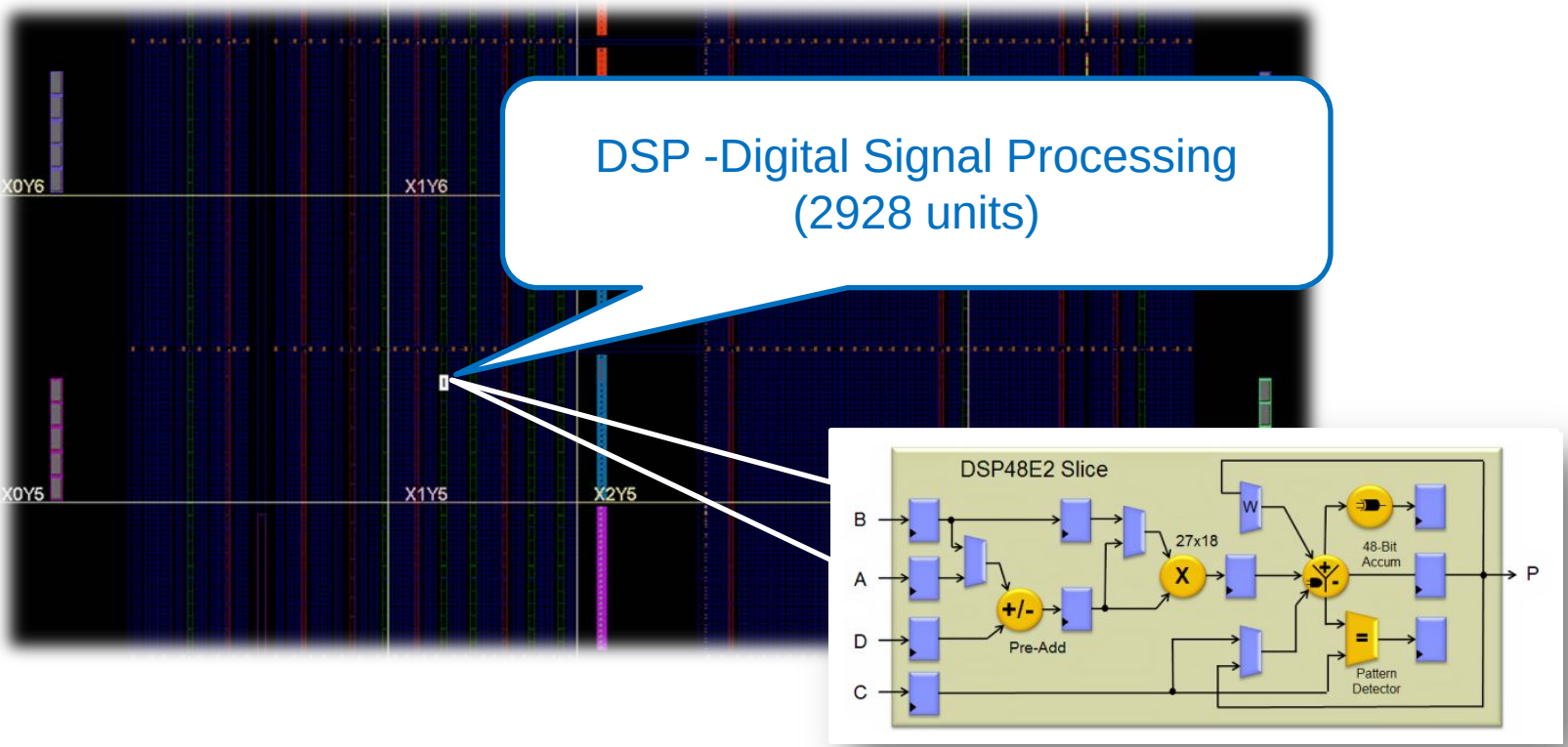


Modern FPGA – ZYNQ Ultrascale+ XCZU11

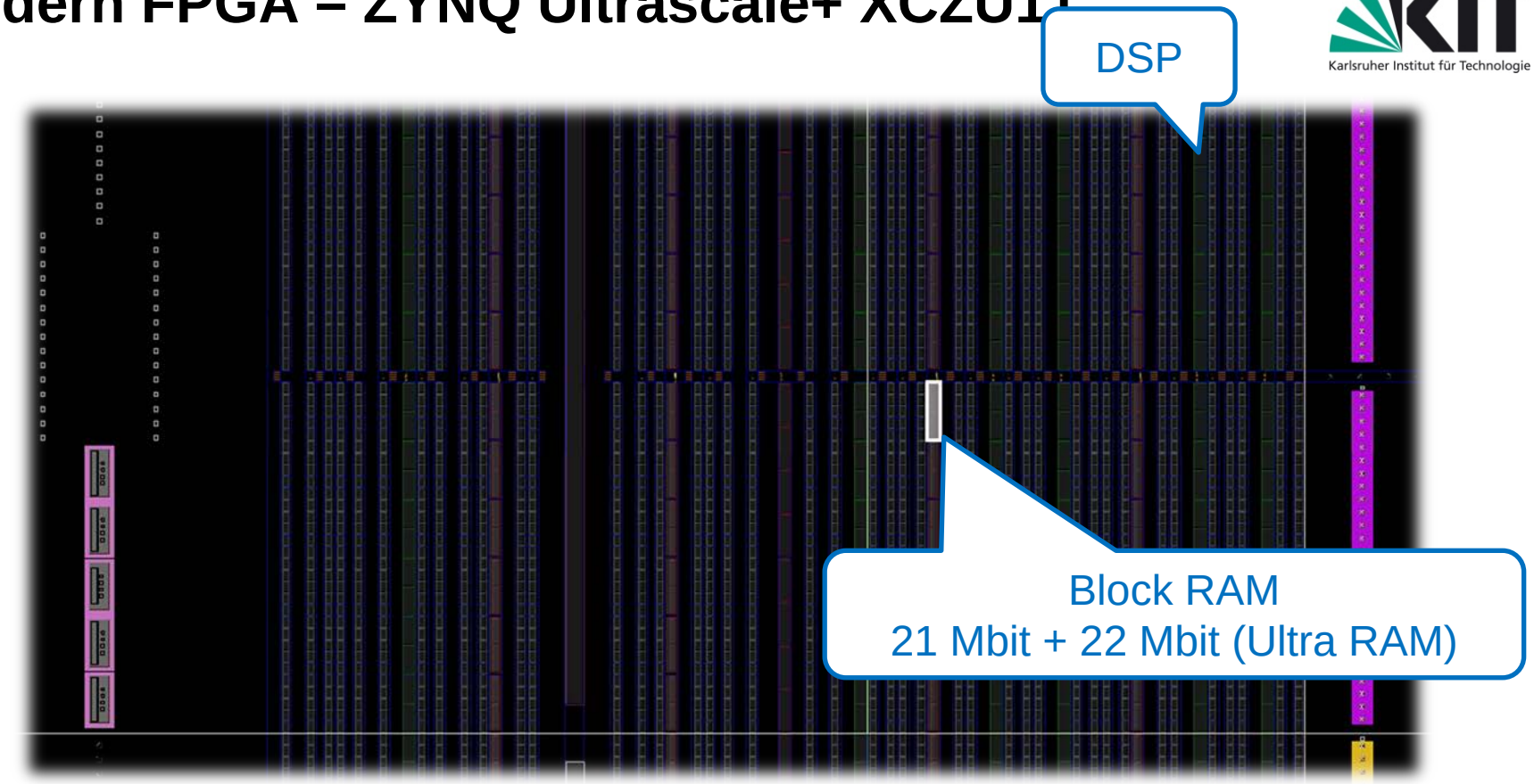
44 regions in total



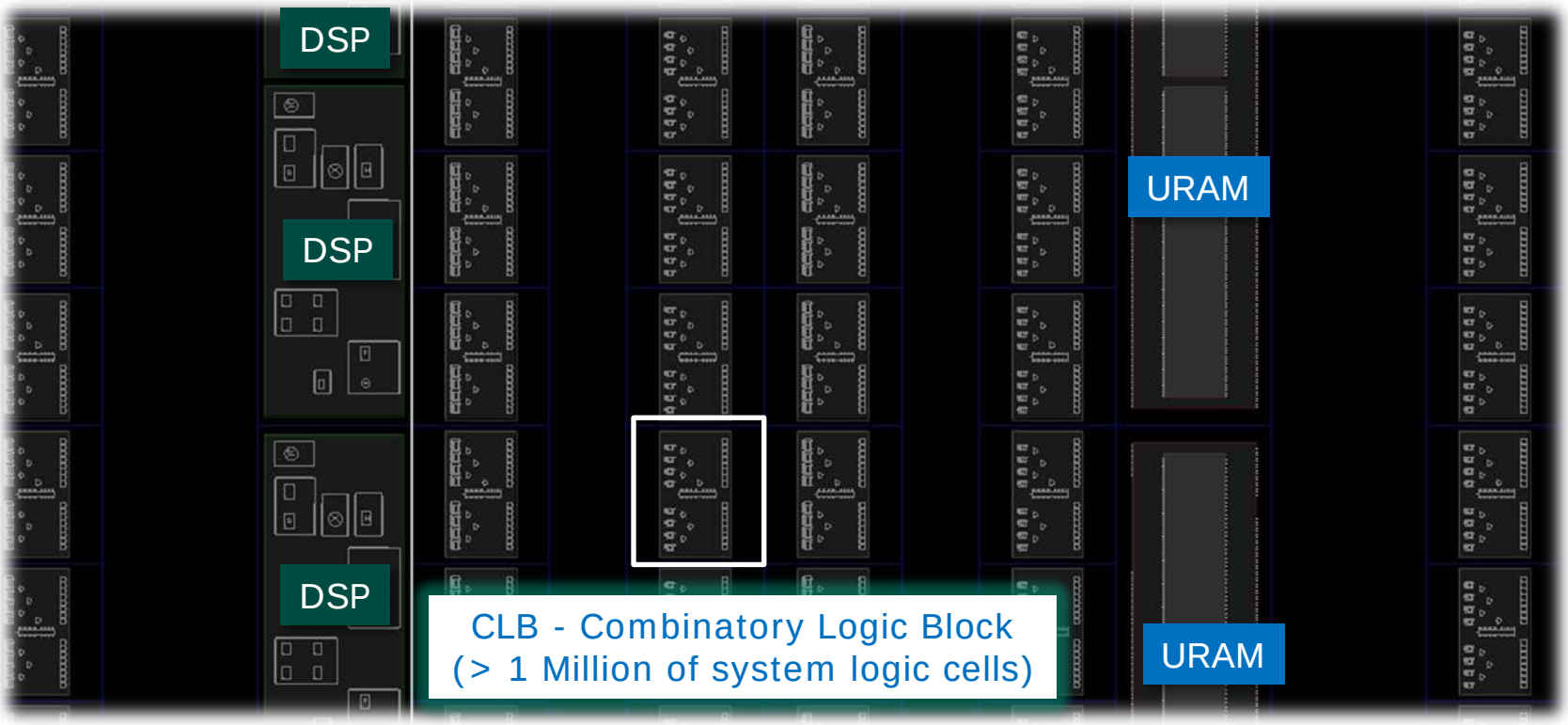
Modern FPGA – ZYNQ Ultrascale+ XCZU11



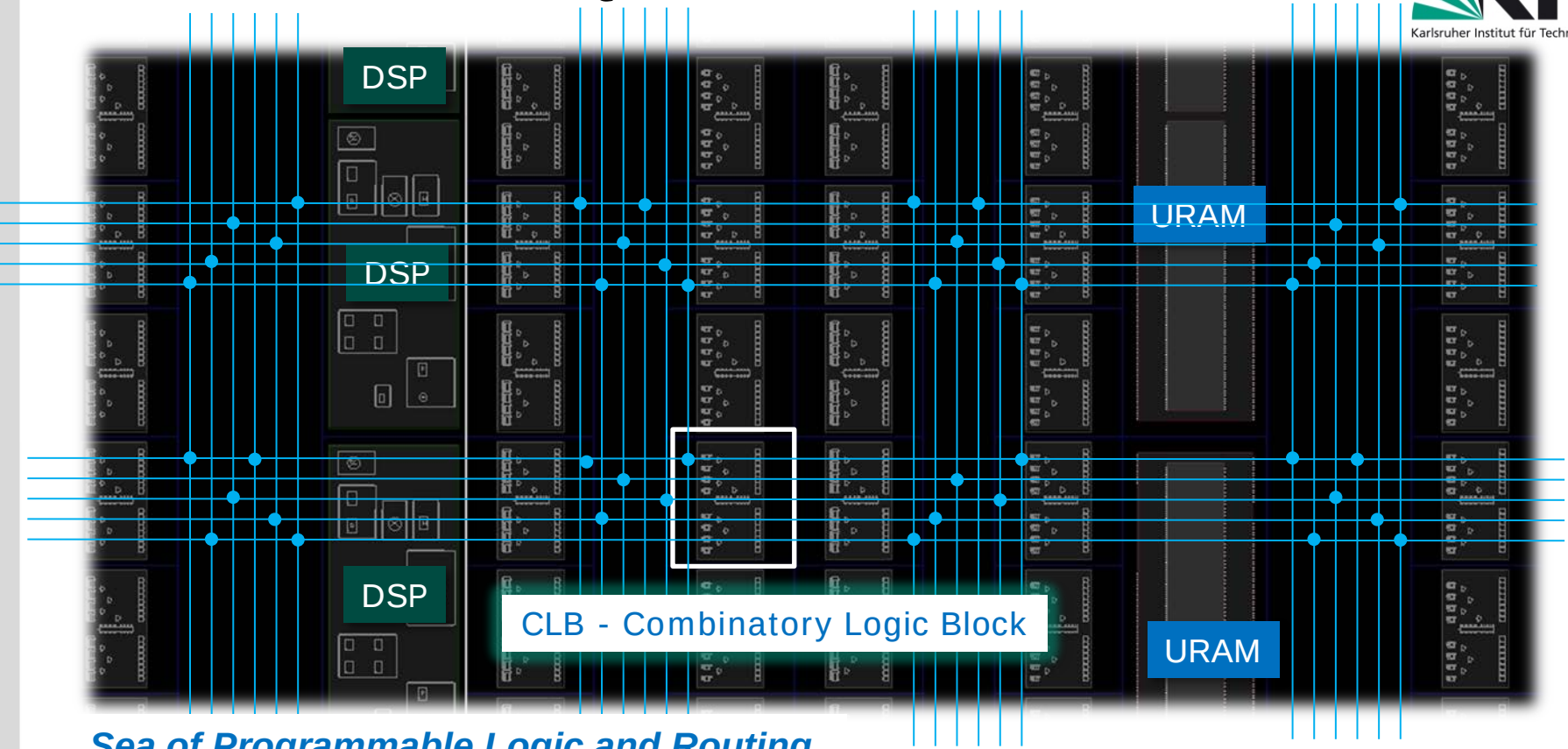
Modern FPGA – ZYNQ Ultrascale+ XCZU11



Modern FPGA – ZYNQ Ultrascale+ XCZU11

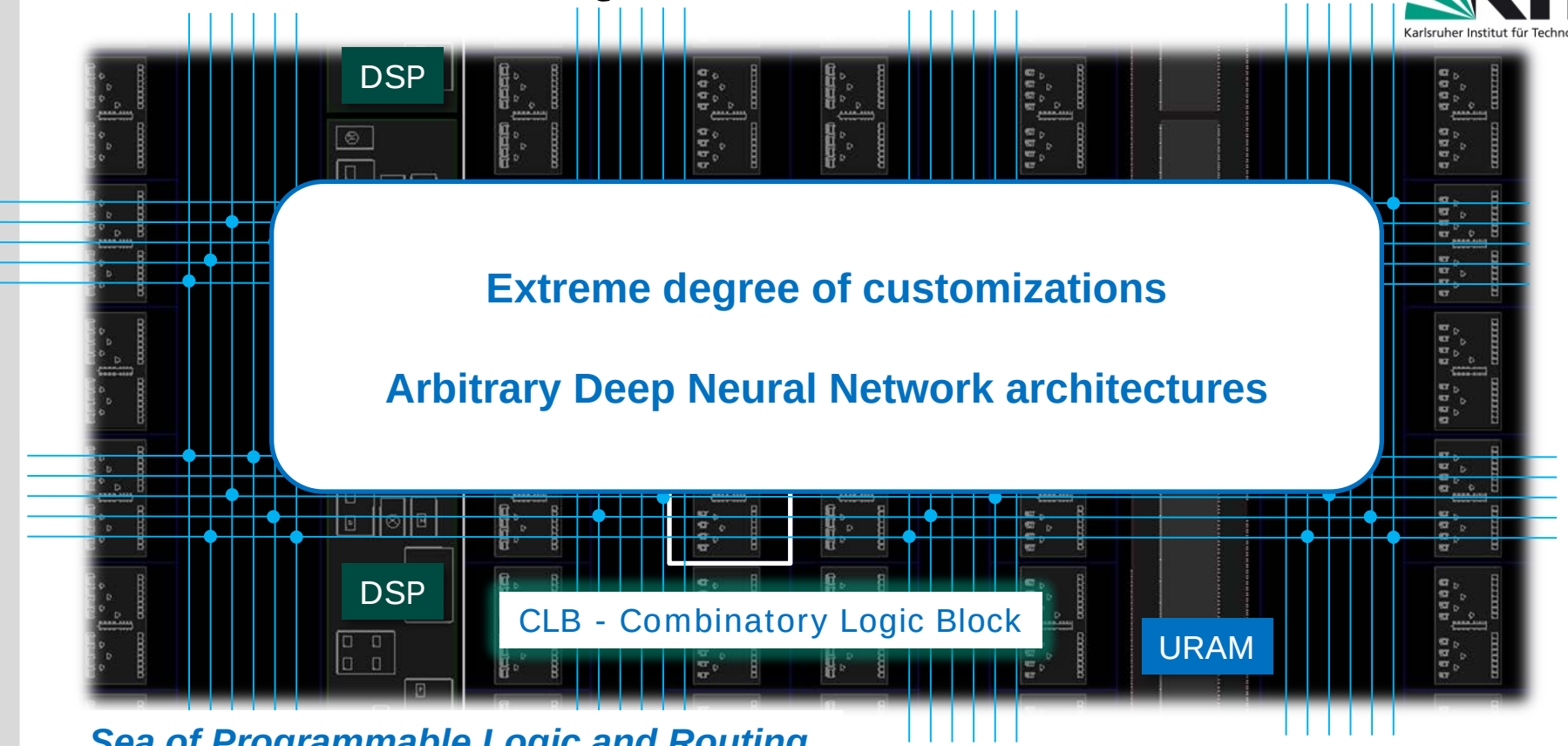


Modern FPGA – ZYNQ Ultrascale+ XCZU11



Sea of Programmable Logic and Routing

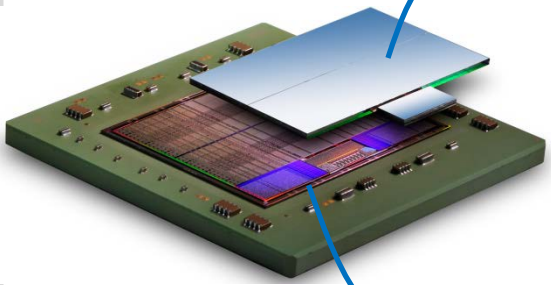
Modern FPGA – ZYNQ Ultrascale+ XCZU11



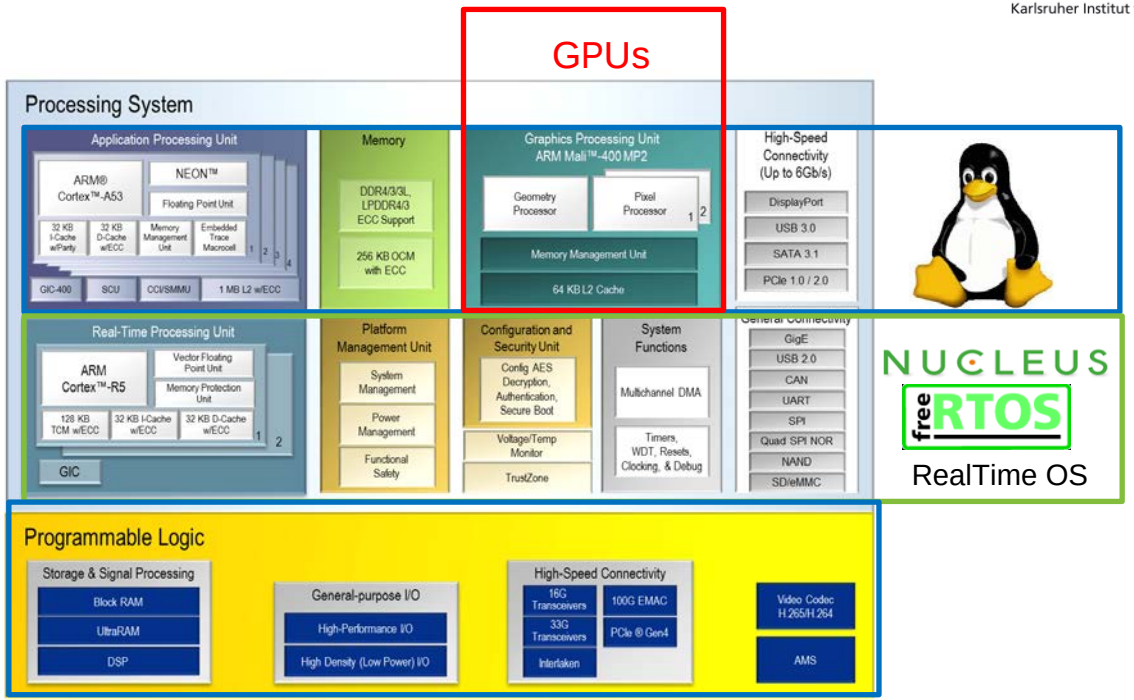
Sea of Programmable Logic and Routing

Not only FPGA…

Multi-Processor System-on-Chip

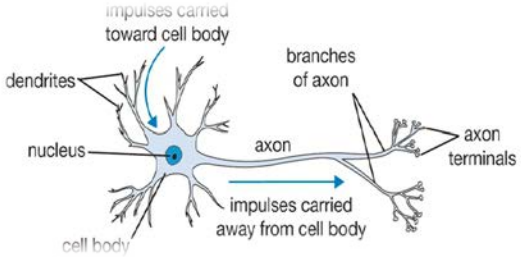


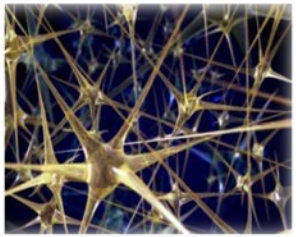
FPGA



Programmable gate logic and Linux O.S. combined on an unique system

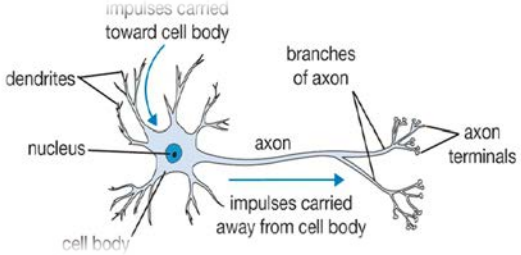
Deep Neural Network on FPGA

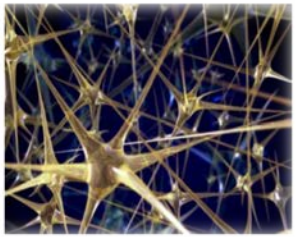




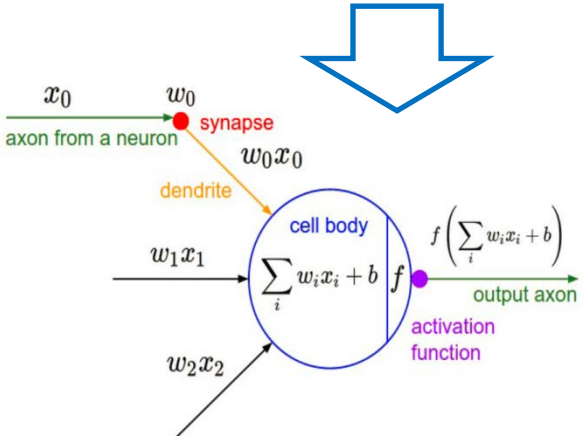
The basic unit is a **neuron,** neurons are connected with **synapses**.
Neurons receive input and produce output along **axon**, which interact with the dendrites of other neurons via **synaptic weights**.
Synaptic weights – learnable

Deep Neural Network on FPGA



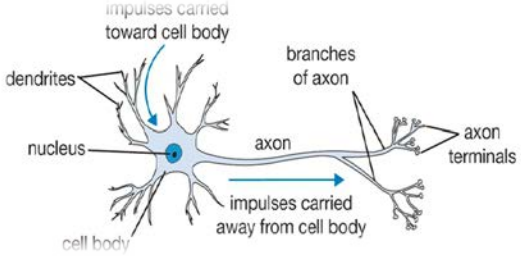


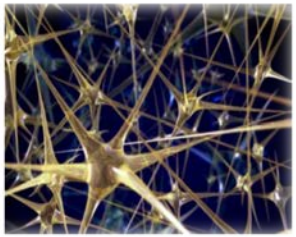
The basic unit is a **neuron,** neurons are connected with **synapses**.
Neurons receive input and produce output along **axon**, which interact with the dendrites of other neurons via **synaptic weights**.
Synaptic weights – learnable



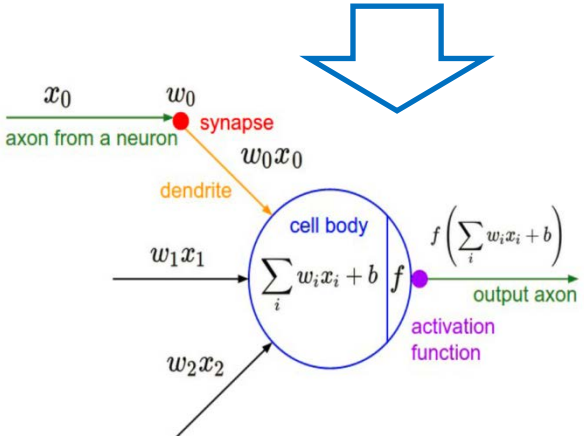
Model of neuron cell

Deep Neural Network on FPGA

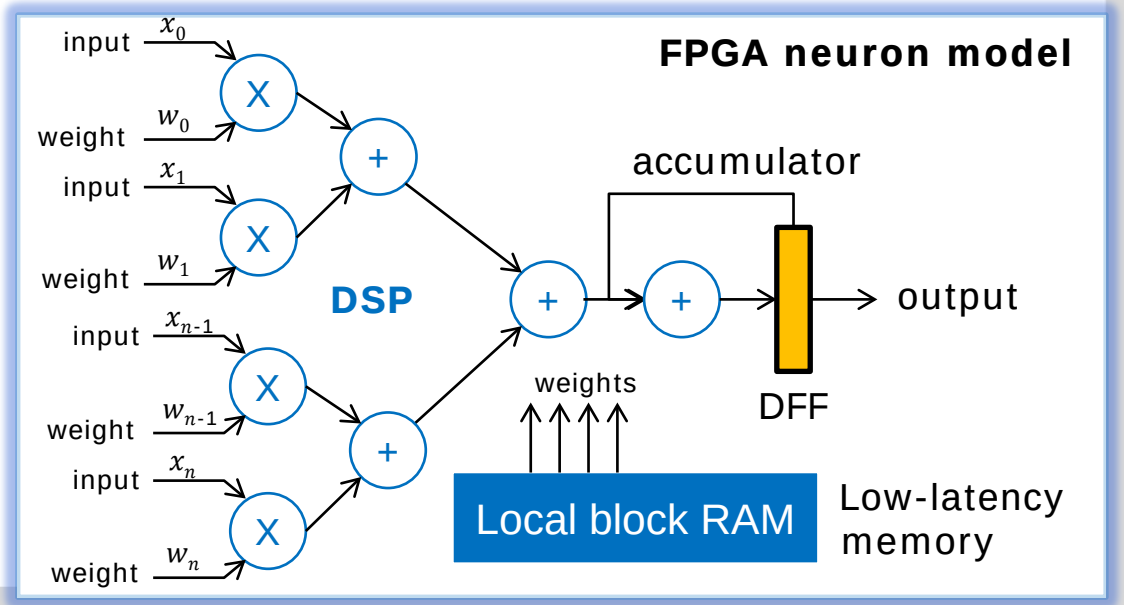




The basic unit is a **neuron,** neurons are connected with **synapses**.
Neurons receive input and produce output along **axon**, which interact with the dendrites of other neurons via **synaptic weights**.
Synaptic weights – learnable



Model of neuron cell



Comparison – CPU, GPU and FPGA

Feature	Analysis	Winner
DNN Training	GPU floating point capabilities are greater	GPU
DNN Inference	FPGA can be customized, and has lower latency	FPGA
Large data analysis	CPUs support largest memory and storage capacities. FPGAs are good for inline processing.	CPU/FPGA
Timing latency	Algorithms implemented on FPGAs provide deterministic timing, can be an order of magnitude faster than GPUs	FPGA
Processing/Watt	Customized designs can be optimal	FPGA
Processing/$$	GPUs win because of large processing capabilities. FPGA configurability enables use in a broader acceleration space.	GPU/FPGA
Interfaces	FPGA can implement many different interfaces	FPGA
Backward compatibility	CPUs have more stable architecture than GPUs. Migrating RTL to new FPGAs requires some work.	CPU
Ease of change	CPUs and GPUs provide an easier path to changes to application functionality.	GPU/CPU
Customization	FPGAs provide broader flexibility	FPGA
Size	CPU and FPGA's lower power consumptions leads to smaller volume solutions	CPU/FPGA
Development	CPUs are easier to program than GPUs, both easier than FPGA	CPU

Reference: *Allen Rush, Ashish Sirasao, Mike Ignatowski. "Unified Deep Learning with CPU, GPU, and FPGA Technologies »*

Comparison – FPGA, GPU and FPGA

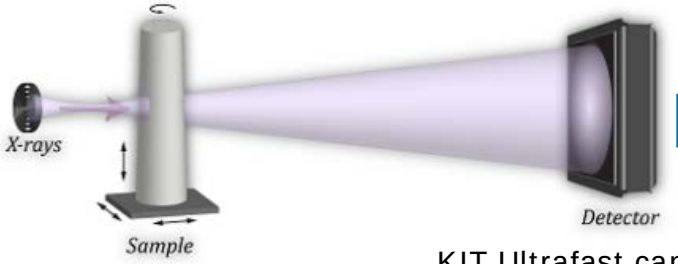
Feature	Analysis	Winner
DNN Training	GPU floating point capabilities are greater	GPU
DNN Inference	FPGA can be customized, and has lower latency	FPGA
Large data analysis	CPUs support largest memory and storage capacities. FPGAs are good for inline processing.	CPU/FPGA
Timing latency	Algorithms implemented on FPGAs provide deterministic timing, can be an order of magnitude faster than GPUs	FPGA
Processing/Watt	Customized designs can be optimal	FPGA
Processing/$$	GPUs win because of large processing capabilities, configurability enables use i	
Interfaces	FPGA can implement many d	
Backward compatibility	CPUs have more stable archi RTL to new FPGAs requires s	
Ease of change	CPUs and GPUs provide an ea application functionality.	GPU/CPU
Customization	FPGAs provide broader flexibility	FPGA
Size	CPU and FPGA's lower power consumptions leads to smaller volume solutions	CPU/FPGA
Development	CPUs are easier to program than GPUs, both easier than FPGA	CPU

There are complementary strengths and weaknesses, there are no obvious "one size fits all" solutions.

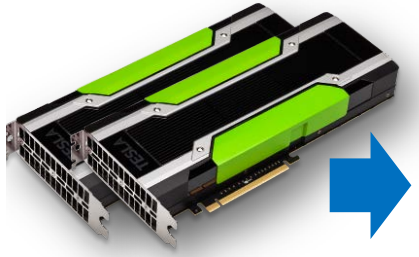
Reference: *Allen Rush, Ashish Sirasao, Mike Ignatowski. "Unified Deep Learning with CPU, GPU, and FPGA Technologies »*

KIT Big Data and Data AcQuisition systems

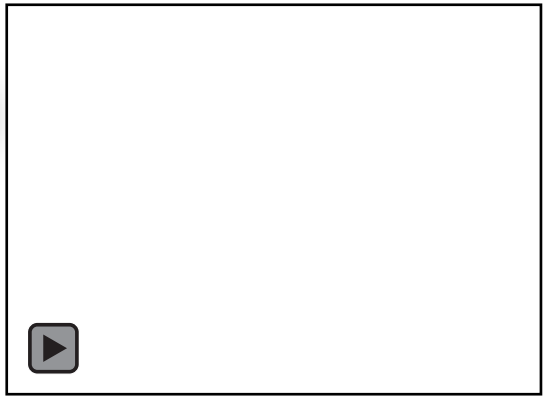
Example: Ultrafast X-ray computer tomography @ KARA (KIT)



KIT Ultrafast camera
Frame rate: 5 – 100 kfps



GPUs cluster



Reconstruction and semi-automatic segmentation by GPUs

Feedback to sample manipulator

UFO experimental station:

high **spatial resolution** → sub-µm in 2D and 3D

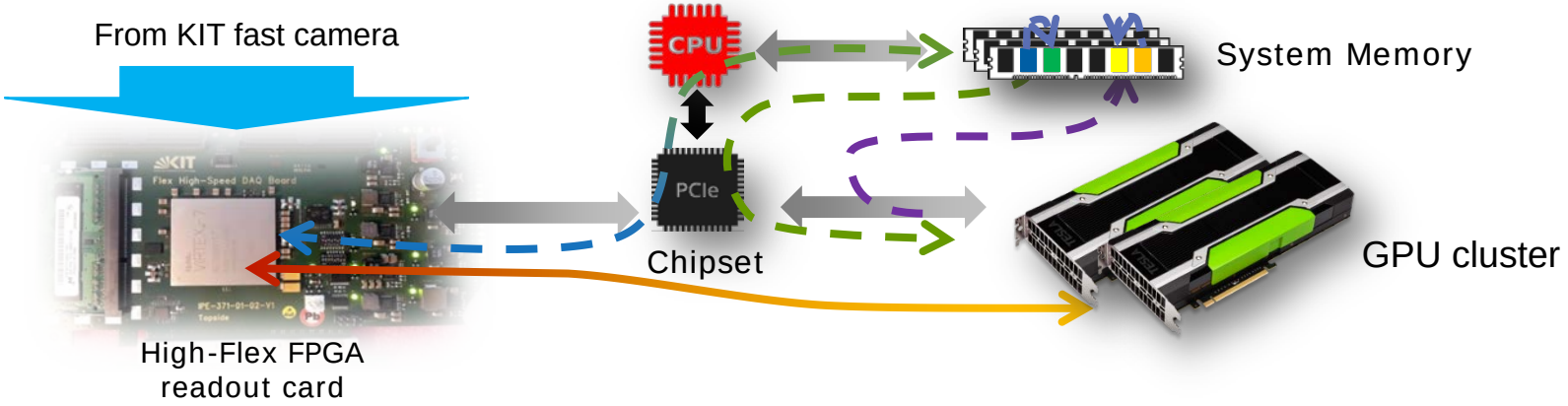
+

high ***time resolution*** → µs / ms to give insight in the temporal structure evolution

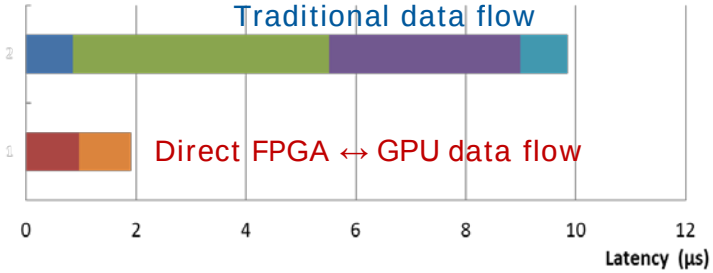
M. Caselle et al., IEEE-RT
DOI:10.1109/TNS.2013.2252528 (2013)

High-performance heterogeneous FPGA-GPU DAQ

Heterogeneous FPGA/GPU-based readout system, the UFO DAQ platform, has been developed



Round-trip time: FPGA -> GPU -> FPGA

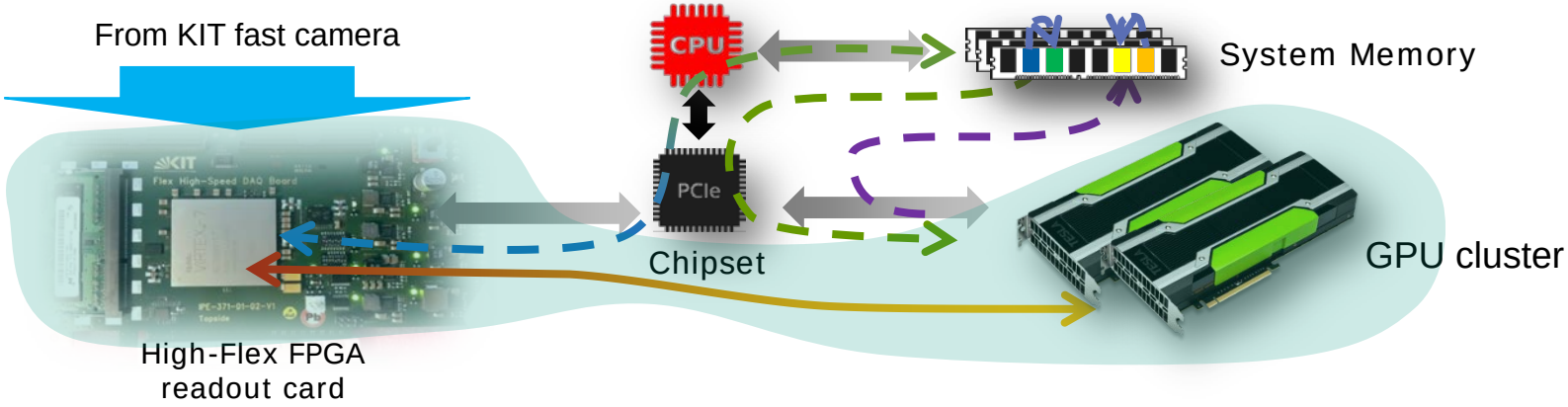


- Direct FPGA ↔ GPU communication enables real-time data processing
- DMA working close to theoretical limit of data link
- Data latency 5 times better than other DMA architectures

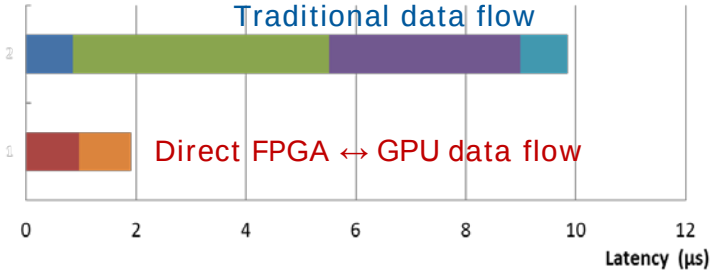
M. Caselle et al., JINST 12 C03015 (2017)

High-performance heterogeneous FPGA-GPU DAQ

Heterogeneous FPGA/GPU-based readout system, the UFO DAQ platform, has been developed



Round-trip time: FPGA -> GPU -> FPGA



- Direct FPGA ↔ GPU communication enables real-time data processing
- DMA working close to theoretical limit of data link
- Data latency 5 times better than other DMA architectures
- Because very low-latency → ***FPGA and GPU operating as an unique system***

M. Caselle et al., JINST 12 C03015 (2017)

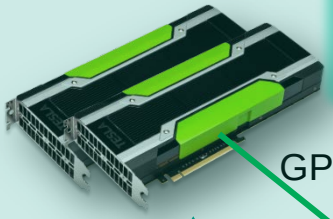
High-performance Deep Machine Learning platform

Karlsruher Institut für Technologie

Using heterogeneous FPGA-GPU to combine the strengths of both FPGA and GPU devices

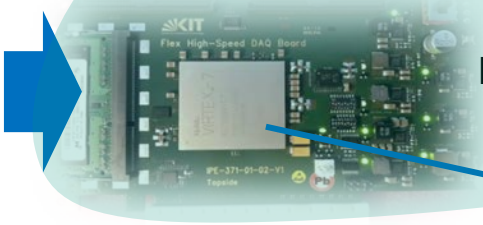
Low-latency, customized designs, many interfaces (input data), broader flexibility, etc.

GPU floating point and large processing capabilities are excellent, easy to program, etc.

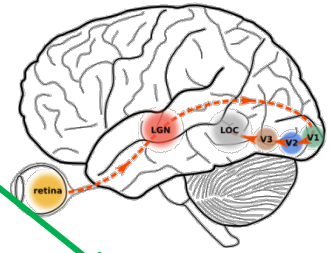


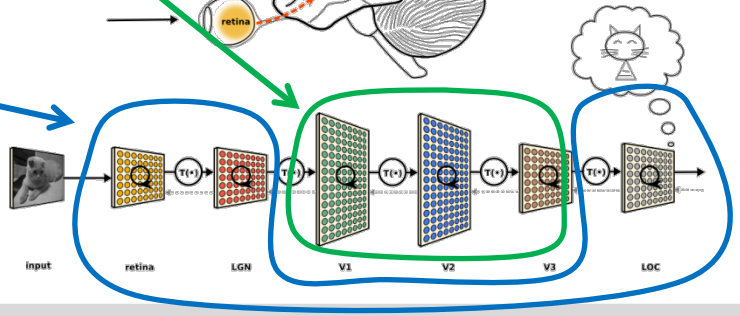
GPU cluster

Data source



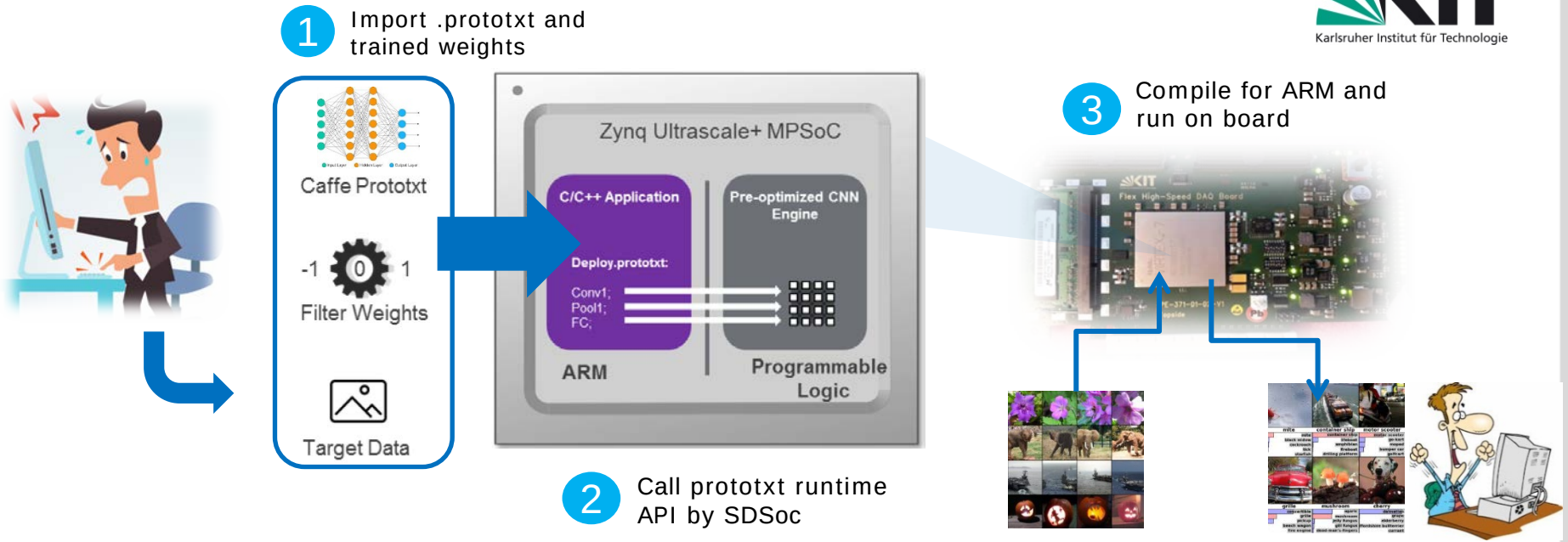
heterogeneous





Fast feedback to experimental station based on machine learning

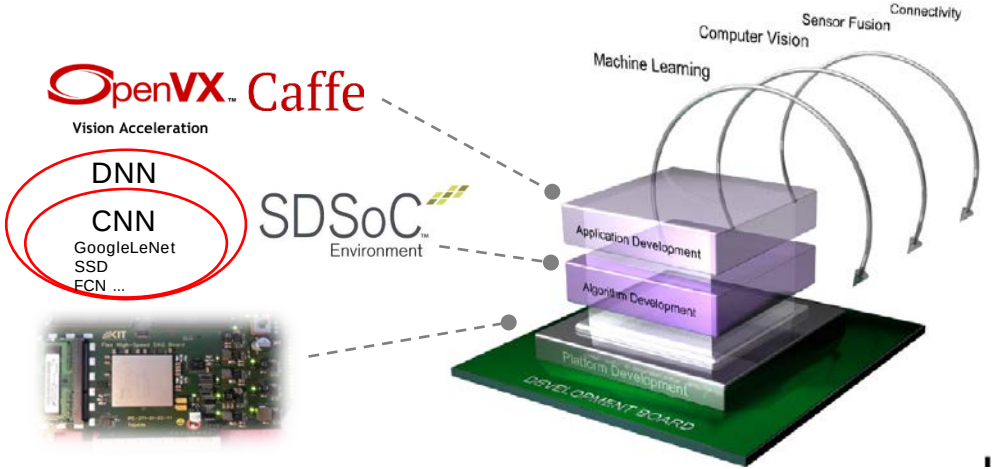
Machine Learning - FPGA design flow



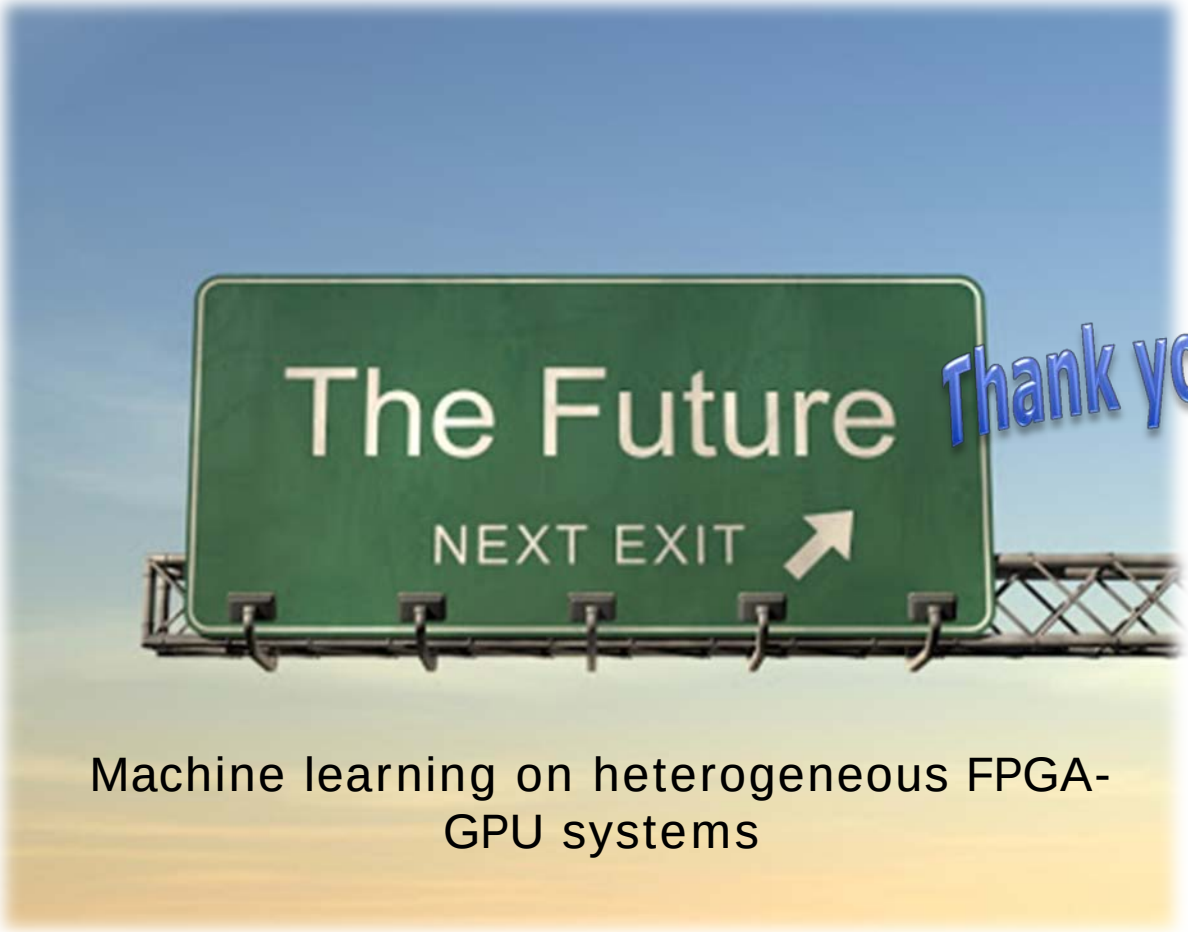
- Full software stack for deploying machine learning
- SDSoC (Software Defined System-on-Chip) generates netlist (RTL) or full FPGA bitstream file with the Convolution Neural Network
- Compiles only ARM software code in minutes. No hardware compilation

Conclusions

- Deep learning machine open new windows into science where FPGA and GPU well complement each other
- High-data throughput and low-latency heterogeneous FPGA-GPU system has been developed which can operate as unique system
- *SDSoC* (Software Defined System-on-Chip) environment removes major barrier on hardware development → allowing developers with little or no FPGA expertise to program FPGA by high level programming languages like C/C++, OpenCL, Python, etc.
- *SDSoC (ReVISION)* includes support for the most popular neural networks → AlexNet, GoogLeNet, SqueezeNet, SSD, FCN, etc.
- FPGA Reconfigurability → allowing to change only the DNN/CNN on FPGA without touching the rest of FPGA logic
- FPGA Responsiveness → high throughput system, deterministic low-latency from data source to inference & control



Reference: https://www.xilinx.com/products/design-tools/embedded-vision-zone.html



Thank you for your attention

Machine learning on heterogeneous FPGA-GPU systems

Michele.caselle@kit.edu