Karlsruhe Institute of Technology

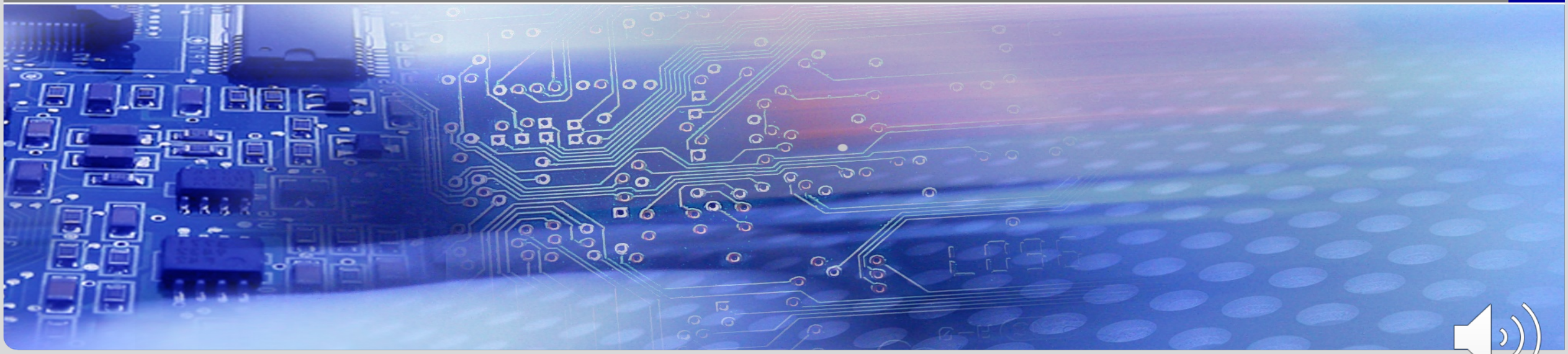
Neuromorphic FPGA Integration –

HPC, Reliability and NN as Key Enablers

Jürgen Becker

Karlsruhe Institute of Technology – KIT

Institute for Information Processing Technologies (ITIV)



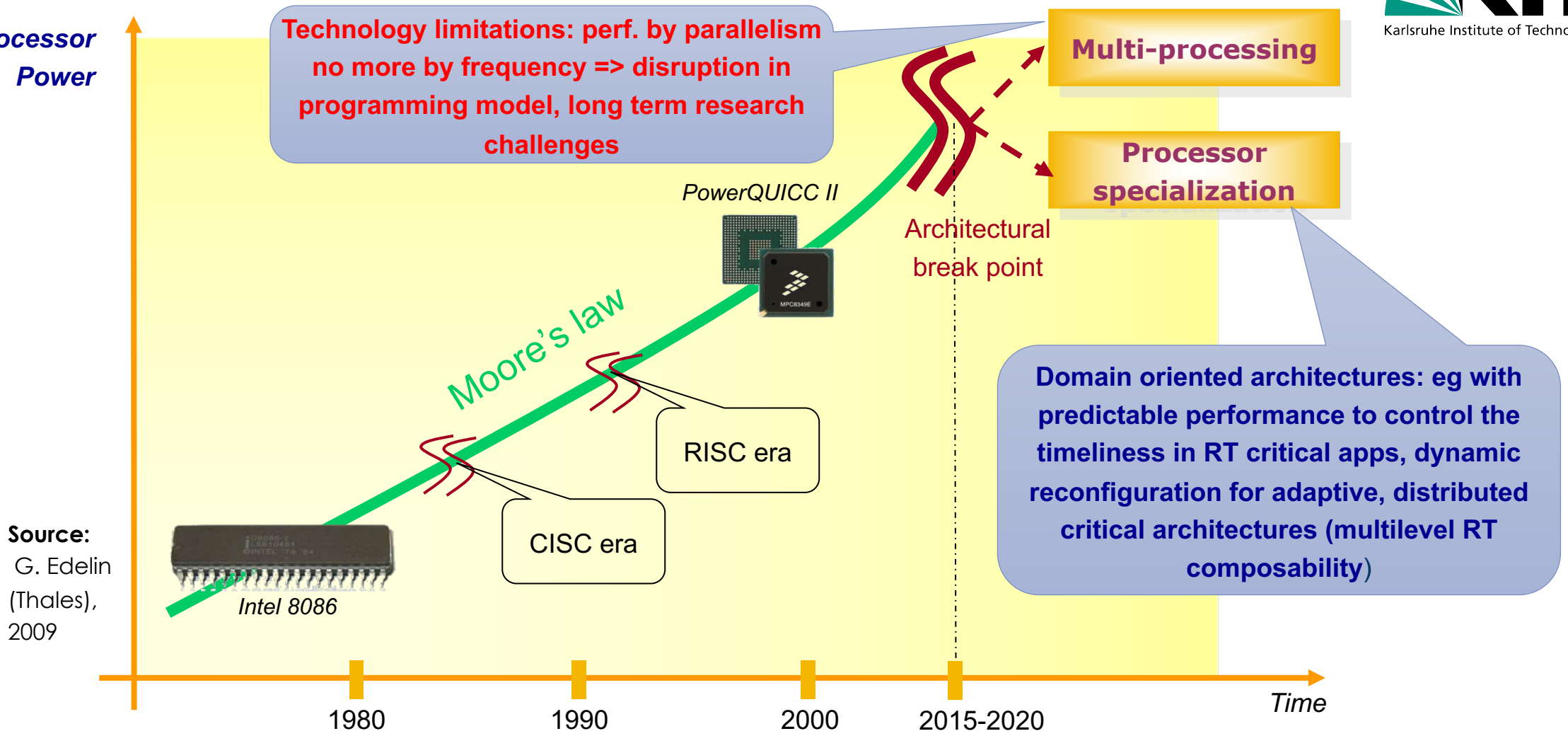
Context and Motivation

- **Current embedded systems are subject to challenging requirements:**
  - Increased **performance** is necessary to facilitate the execution of computationally intensive algorithms (machine learning, big data, …)
  - **Power and energy consumption** must be minimized to facilitate the constraints of mobile and wireless devices (internet of things, …)
  - A sufficient degree of **dependability** is necessary to employ digital systems in safety-critical environments (autonomous driving, …)

- Existing technology must evolve in order to meet these requirements
  ⇒ ***Open Hardware*, *HPC*, *AI* and *Reliability* can play key roles in this research process**

Current embedded systems are subject to challenging requirements:

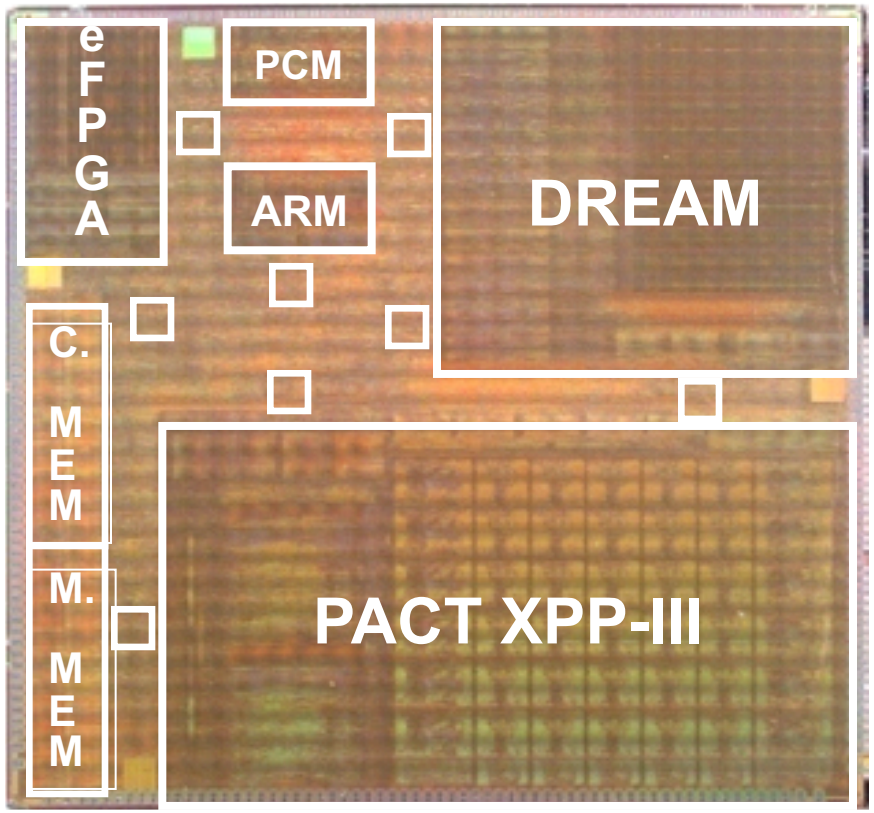
Motivation: Processor Developments ....



Major architecture disruption: multiprocessing & specialization have strong impact on software

Former EU Project: *MORPHEUS* – Early Heterogenous Architectural SoC Solution …

Chip Integration Results



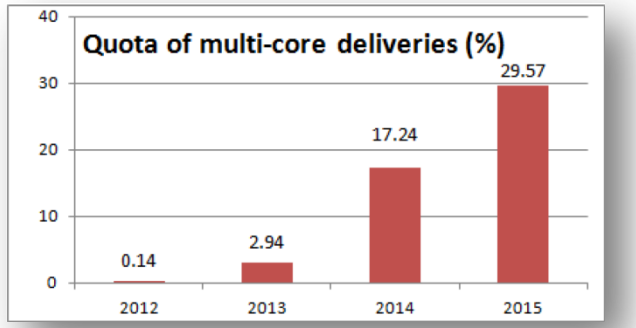
Technology:	**STM-CMOS090GP**
Supply voltage:	**[.9V : 1.2V], ref 1V**
Transistor count:	**97 M**
Chip area:	**110 mm2**
Static power:	**235 mW**
Dynamic power:	
Standby:	**700 mW**
peak:	**3100 mW**
Pinout:	**256, 163 I/O**

Main (System) DOMAIN:
Frequency@1V: 250MHz
Dynamic power: 2.8 mW/MHz

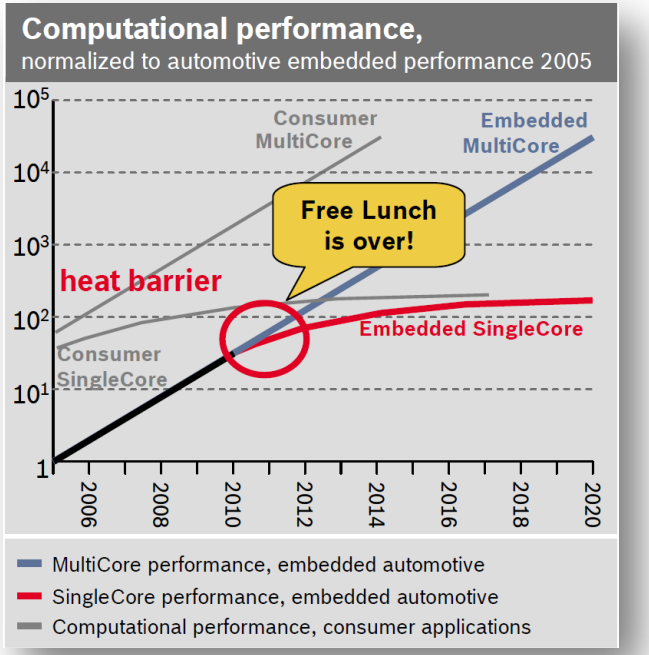
Source: G. Edelin, Thales, 2009

XPP subsystem:	**DREAM subsystem:**	**eFPGA subsystem:**
Macro area: 8.5x5 mm2	**Macro area: 5x4.5 mm2**	**Macro area: 1.8x2.8 mm2**
Max Freq@1V: 150 MHz	**Max Freq@1V: 200 MHz**	**Max Freq@1V: 100 MHz**
Dynamic power: 13 mW/MHz	**Dynamic power: 1.7 mW/MHz**	**Dynamic power: 1.3 mW/MHz**

Embedded Computing Performance Needs



[Source: Shared SW development in multi-core automotive context, L. Michel, et. al, 2016]



[**Source:** The Challenge of Mastering Parallelism in Real-Time Systems, J. Haerdtlein, 2014]

Quota of deliveries based on multi-core CPU at VW/AUDI (not yet in safety critical applications):

[Source: Audi AG]

[Source: Airbus Group]

Specified Requirements about …

- … **functional Safety** (*Safety*) …
- … **Data Security** (*Security*) …
- … „**Space, Weight & Power**“ (SWaP) …

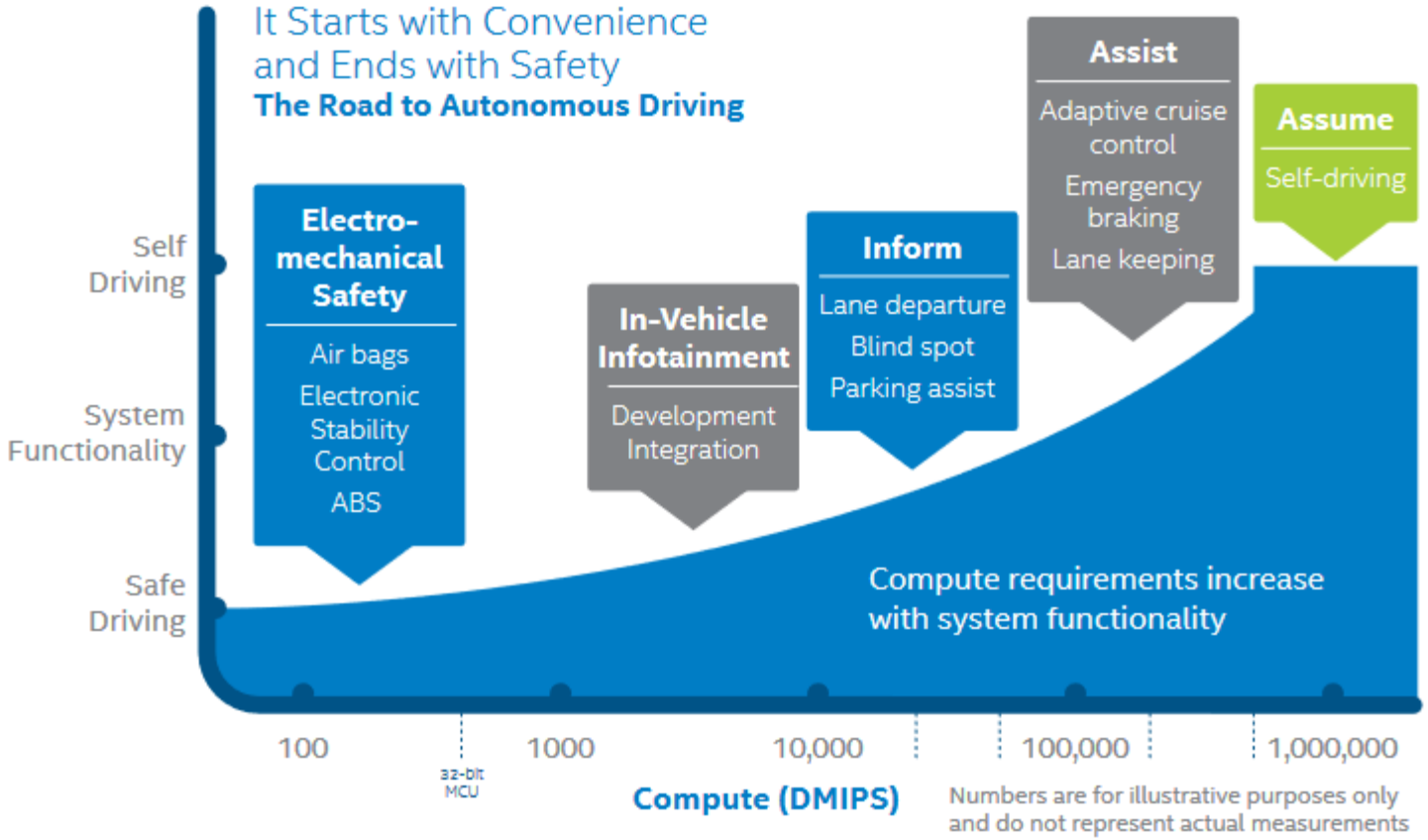
… dedicated and new Solutions needed in the Context of Multi-Core & MP Systems

Safety/Security Codesign is one new Challenge
-> additional Requirements for Platform Development!

Automotive Technology Developments

-> **<u>Autonomous Cars</u>: Embedded HPC Demand**

- Trend towards autonomous driving is driven by performance requirements in the TOPs range



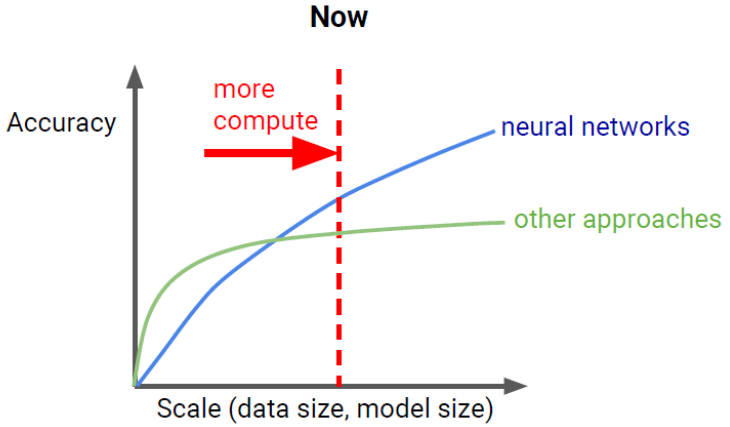
(1) https://www.intel.com/content/www/us/en/automotive/driving-safety-advanced-driver-assistance-systems-self-driving-technology-paper.html

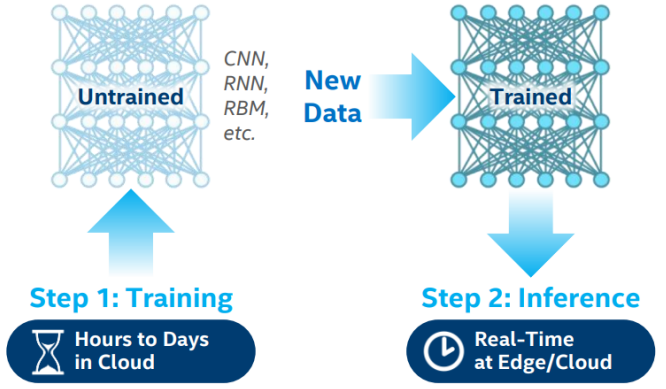
Neuromorphic Options – Machine Learning (ML)

- **Machine Learning for large scale Data Processing**
  - Cost-efficient
  - Reasonably precise
  - Enabled by availability of processing resources

- **Push towards closer Integration Sensors and Embedded Devices**
  - Dedicated ***Neural Processing***
    - Apple A11 Bionic
    - Google PixelVision

- **Traditional methods do not scale ...**
  - **ML as option -> system approach!**

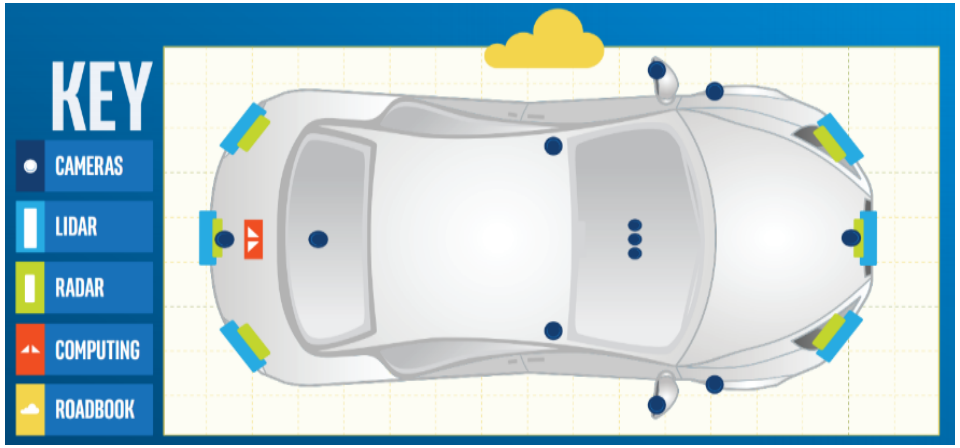




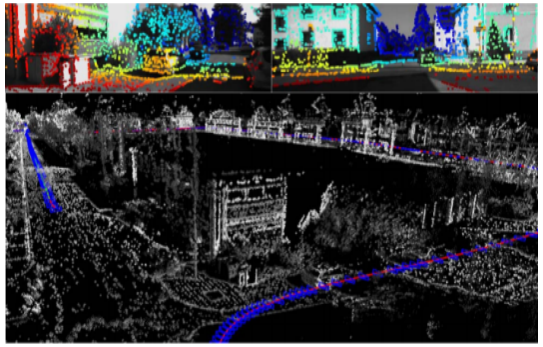
Motivation – eHPC in Automotive
Innovation Driver -> Autonomous Driving

Performance Challenges in different Domains

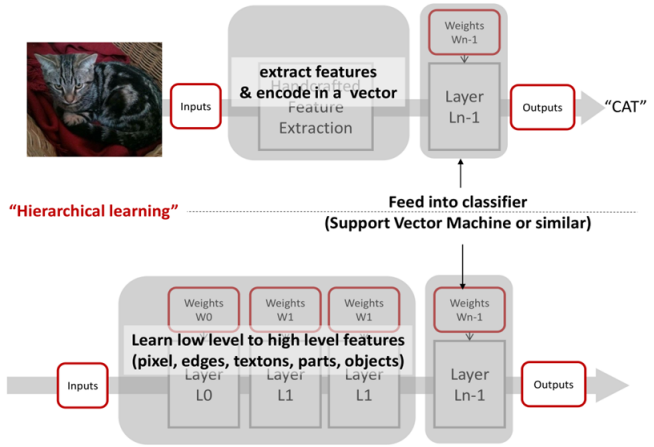
- **Computer Vision**
  - Recognition (e.g. Intel i7)
  - Image Classification (Xilinx Everest)
  - Semantic Segmentation (NVIDIA PX2)
- **Data fusion**
  - Cameras
  - Lidar, Radar
- **Connectivity / 5G (Intel Go)**



Taken from (1).



Taken from (2).



Taken from (3).

Taken from (4).

(1) https://newsroom.intel.de/news/sensors-the-eyes-and-ears-of-autonomous-vehicles/
(2) https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8237683
(3) Xilinx, Inc. **Architectures for Accelerating Deep Neural Networks.**
(4) https://blogs.nvidia.com/blog/2016/01/05/eyes-on-the-road-how-autonomous-cars-understand-what-theyre-seeing/

Motivation

- The trend towards autonomous driving is accompanied by **performance requirements in the TOPs range**

- ***Machine Learning*** can help to automate this process
- However, they require billions of operations
  - E.g. 1.2 billion MACs in MobileNets

→ ***Regular CPUs*** and ***GPUs*** can not fulfil the growing requirements anymore (memory wall, power wall, etc…)

→ A clear trend towards **dedicated hardware accelerators** is observed

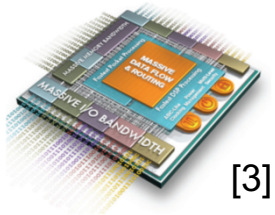
Deep Learning Hardware Evolution

Where Deep Learning has been traditionally happening …

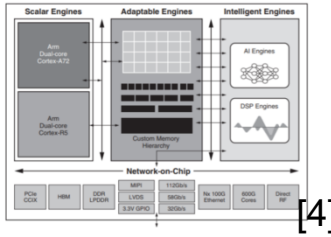
Where state-of-the-art Development is moving towards …

[1]

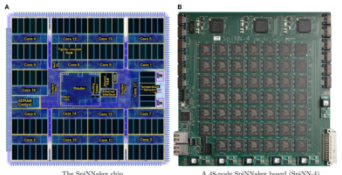
[2]



[3]



[4]


[5]

CPU | GPU | Soft IP | Hard IP | ASIC

Flexibility ⟵ ⟶ Efficiency

FPGA Systems
Microsoft Brainwave
DeePhi/Xilinx
Baidu SDA
Intel/Altera
IBM PowerAI
Amazon AWS F1

Coprocessor/ Heterogeneous
Google TPU
Xilinx ACAP
Intel Nervana
Wave Comp.

Dedicated Circuit
ARM SpiNNaker
Memristor
PCM
RRAM
New Technologies

[1] Paul's Hardware
[2] Nvidia Inc. "GPU-BASED DEEP LEARNING INFERENCE"
[3] Xilinx Inc. "Virtex UltraScale+ FPGAs Product Brief"
[4] Xilinx Inc. "Versal: The First Adaptive Compute Acceleration Platform (ACAP) "
[5] Human Brain Project ; https://www.humanbrainproject.eu

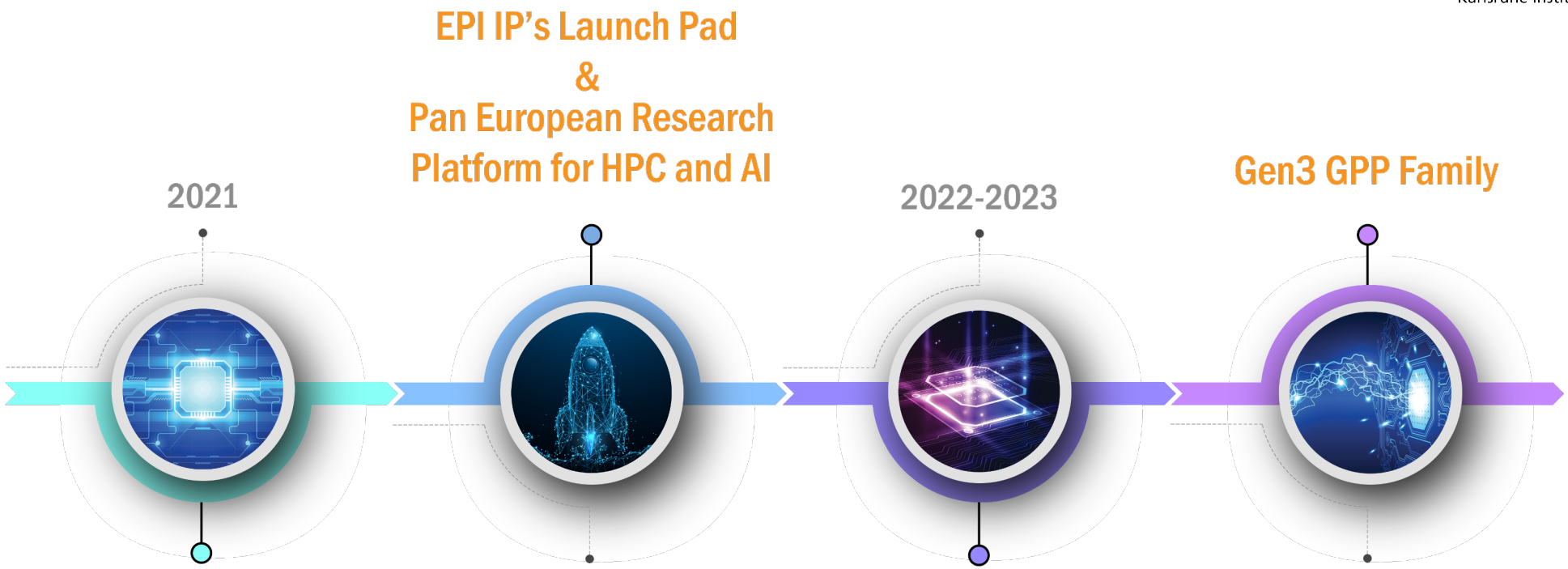
EPI – European Processor Initiative

- Project timespan: Dec. 2018 to End of 2021
- **Target markets:**
  - HPC and Automotive
- **Proposal drivers:**
  - Create a competitive European HPC and Automotive platform

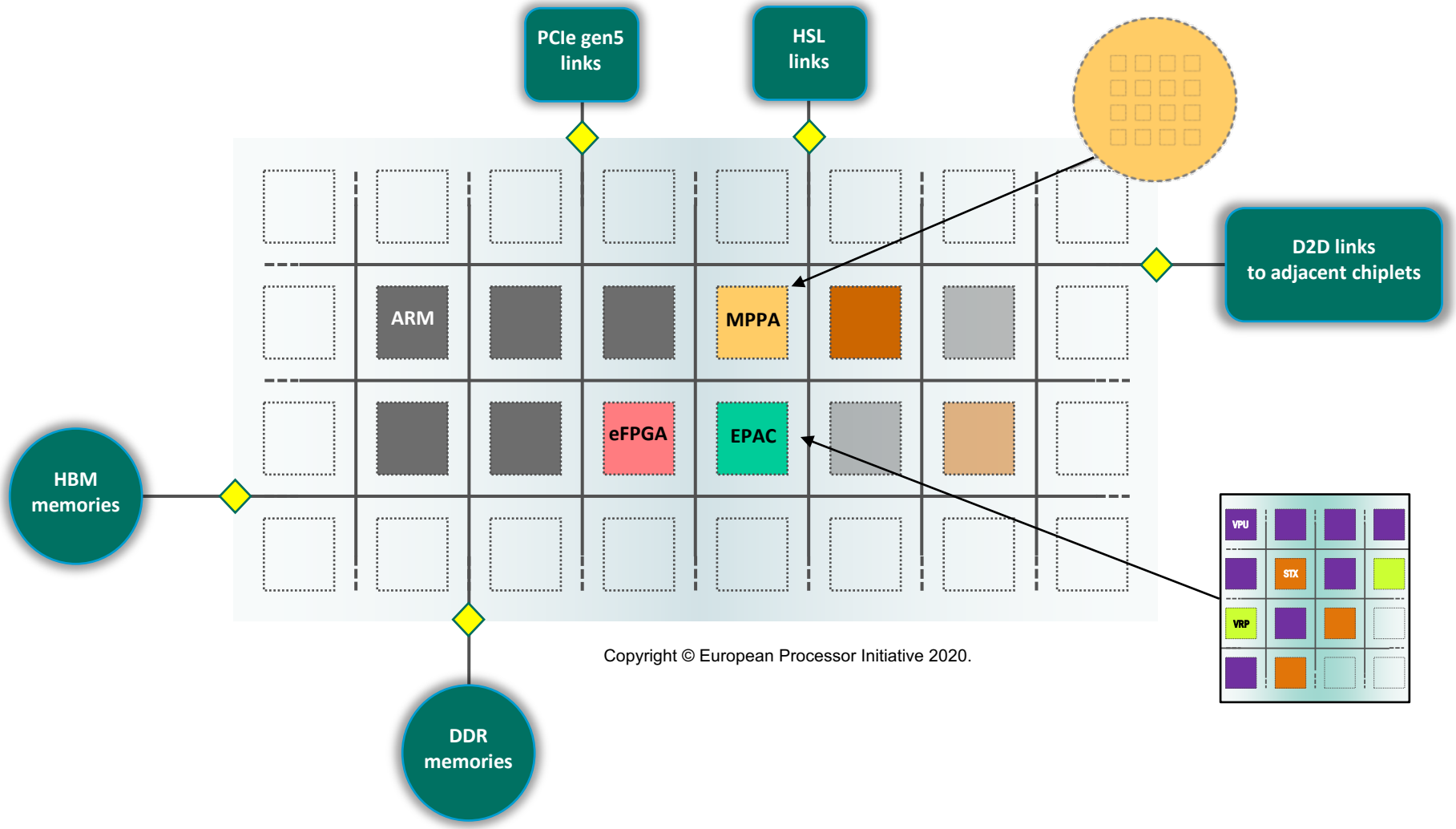
→ **Mission:** independable EU Exascale machine by 2023

More Info: https://european-processor-initiative.eu

EPI Roadmap



EPI – Common Architecture



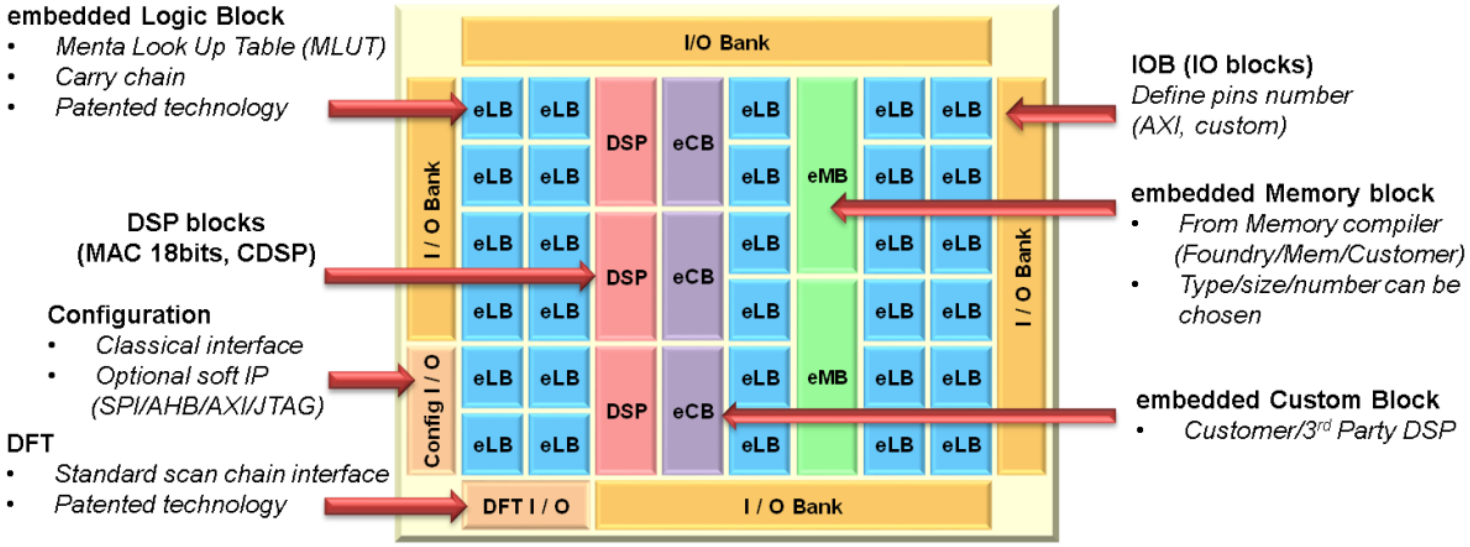
- **MPPA:** Multi-Purpose Processing Array
- **eFPGA:** embedded FPGA
- **EPAC:** EPI Accelerator

EPI – automotive

- EPI for future autonomous driving systems
- Connected mobility (Edge, Cloud)
- **EPI**: A powerful data fusion platform –the automotive embedded HPC platform
- **EPI** heterogeneous multicore architecture can provide enough performance and low power consumption in parallel

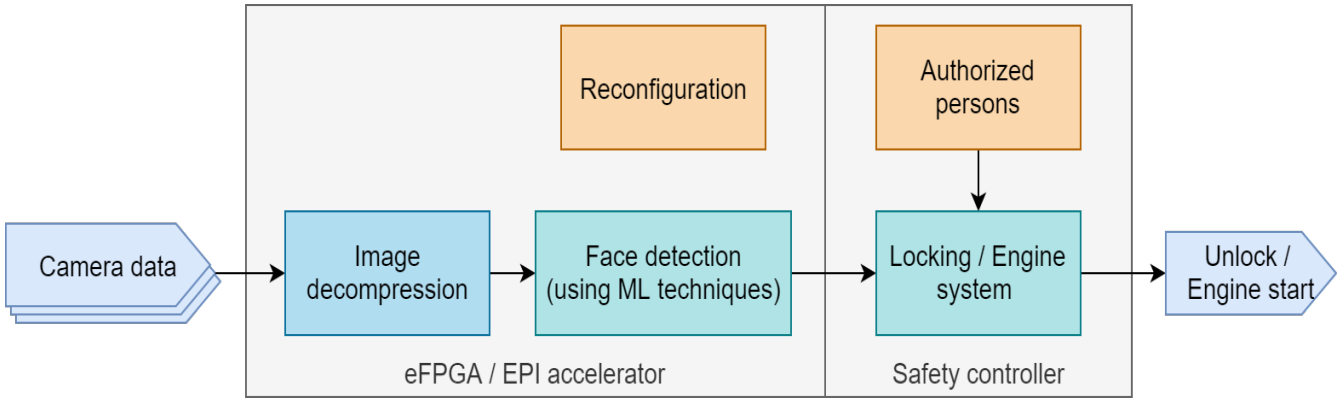
EPI – eFPGA

- Provided by Menta S.A.S.
- Optimized for general purpose and automotive applications
- **Application scenario in EPI: Run-time reconfigurable crypto and general purpose accelerators in a low power domain**

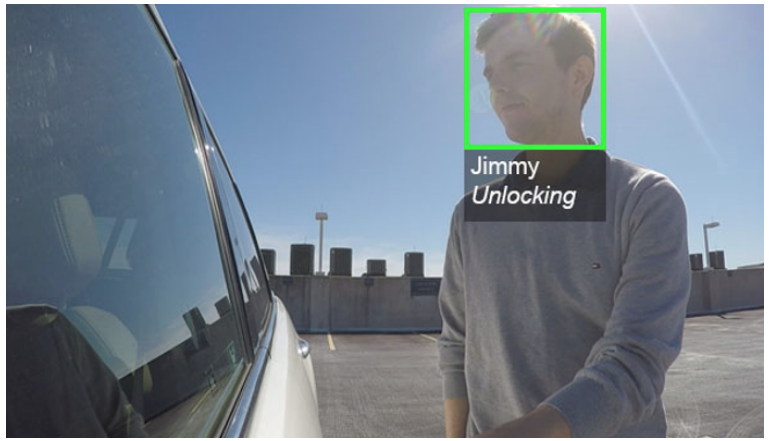


ML Application 1: *EPI* – Face Detection

- ***Face Detection*** for access control = unlock the car
- Use of state of the art machine learning algorithms
- Implementation on the ***eFPGA*** and an *AURIX MCU*

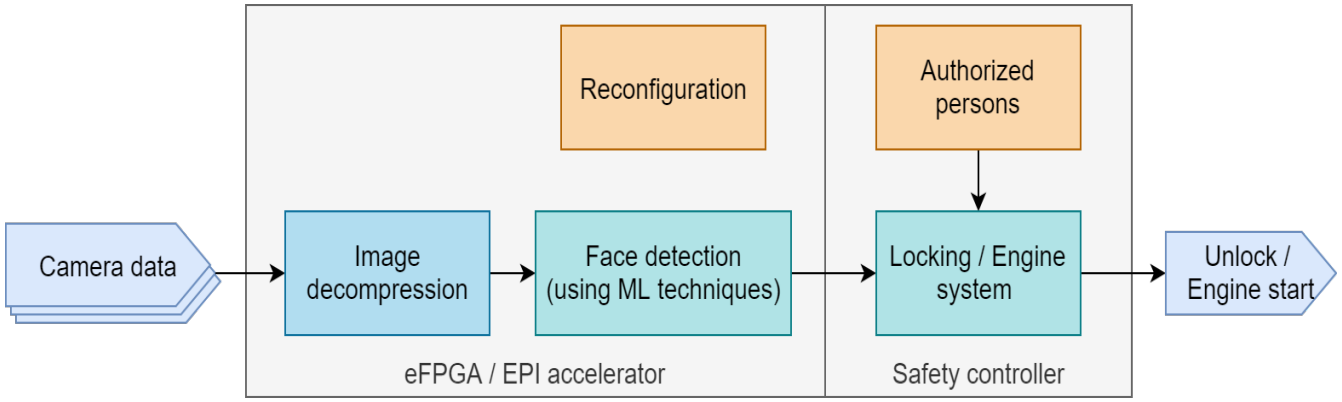


System Overview

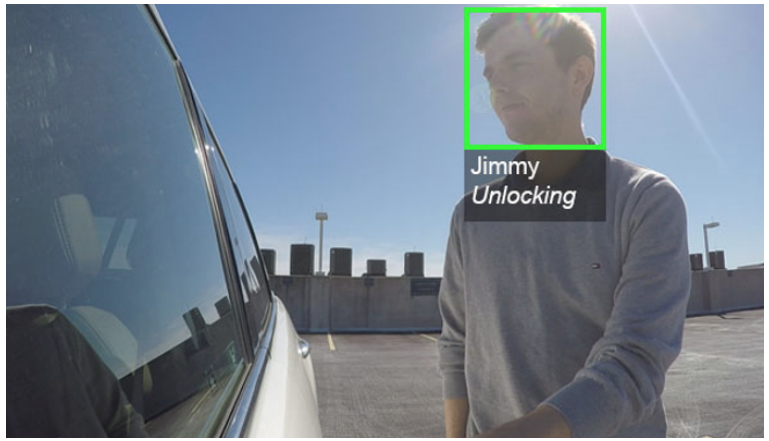


ML Application 1: *EPI* – Face Detection

- **Face Detection on the *eFPGA***
  - Use case demonstrates ***Neural Networks (NN)*** running on the EPI platform
  - First presented on european hardware
  - Low power domain of the *eFPGA* can be leveraged when car is standing
  - **More complex implementation: *Security* feature -> hard to reverse-engineer**

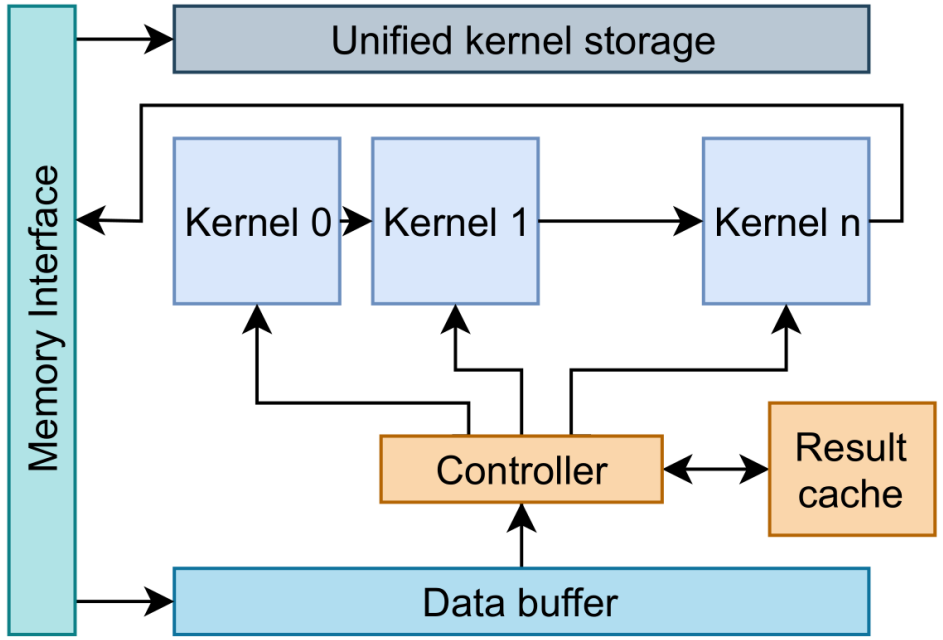


System Overview



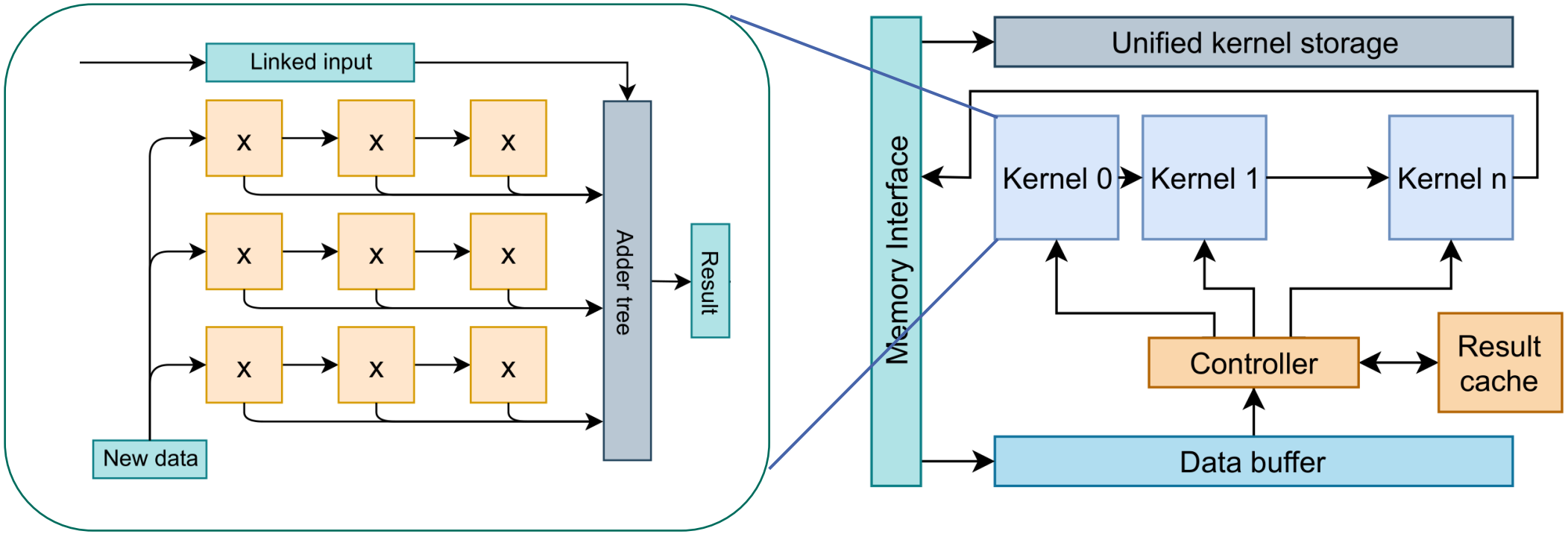
EPI – Image Processing Architecture

- Algorithm is based on ***Convolutional Neural Networks*** (***CNNs***)
- Design of a low *latency hardware accelerator*



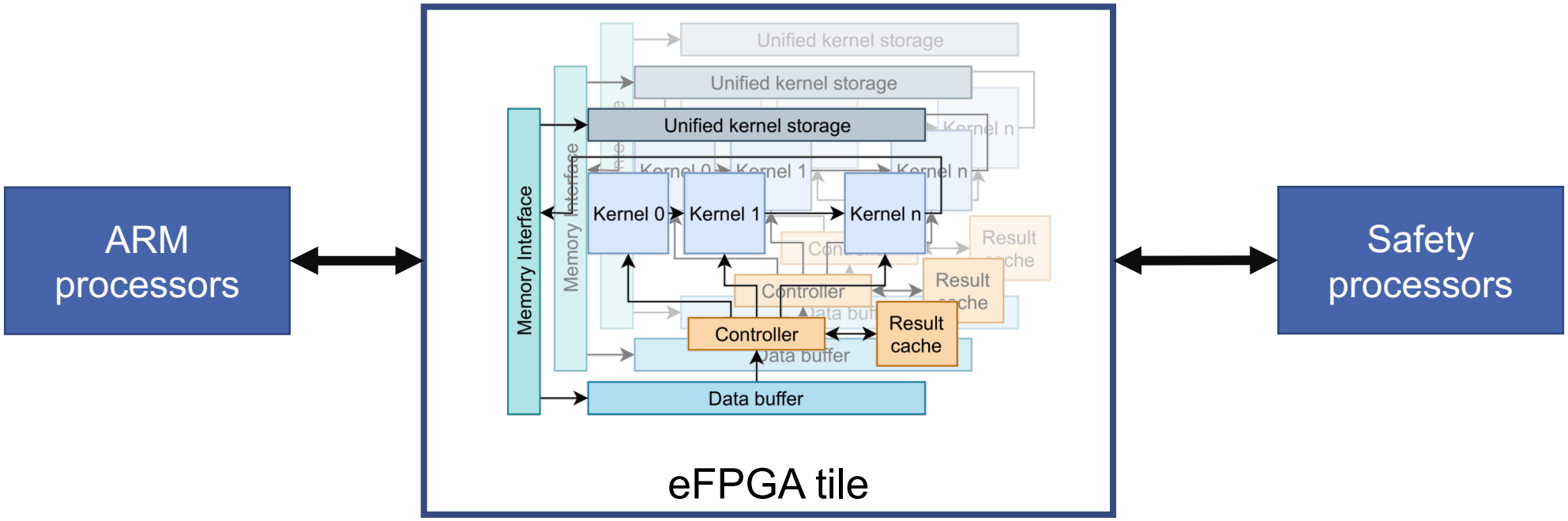
EPI – Image Processing Architecture

- The algorithm is based on ***Convolutional Neural Networks*** (***CNNs***)
- We designed a *low latency hardware accelerator*



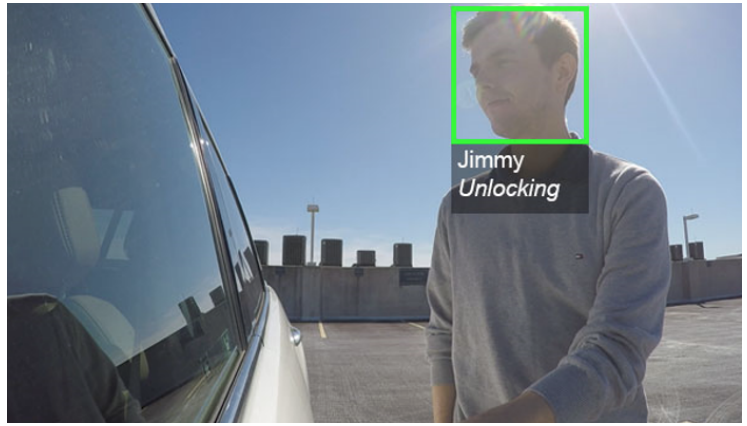
EPI – Image Processing Architecture

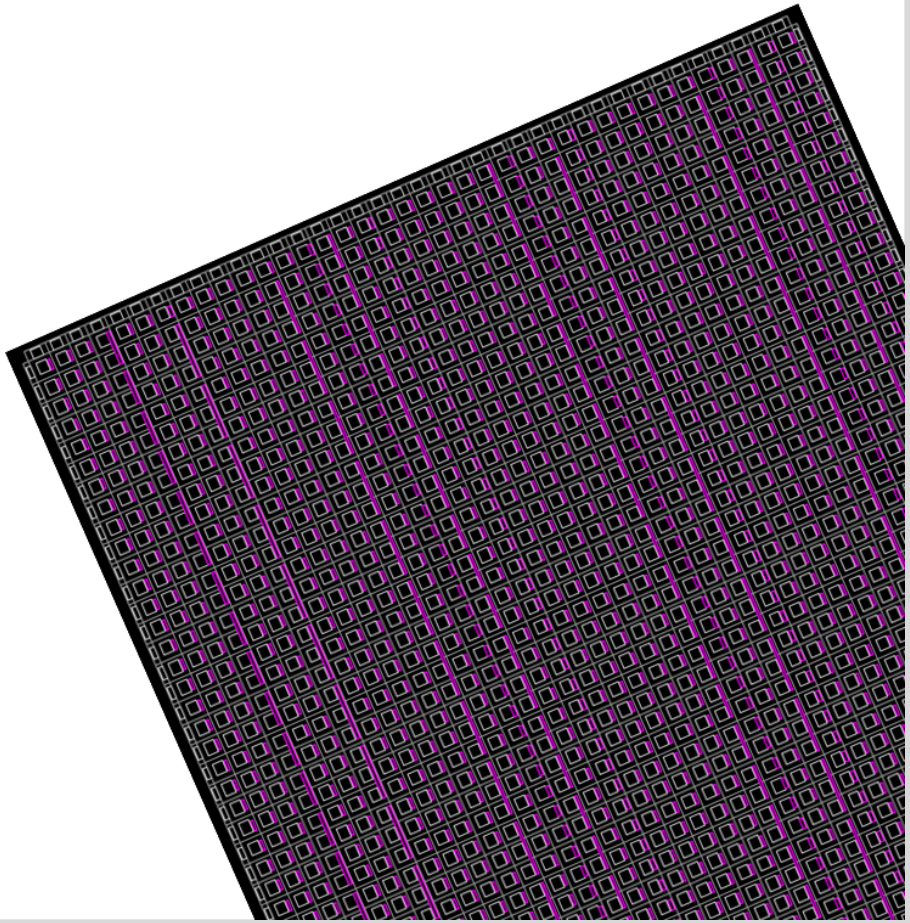
- The architecture is scalable for different ***eFPGA*** sizes
  - With this we explored the design space of the *eFPGA*



EPI – Image Processing – Results

- **We used our architecture to guide the eFPGA architecture**
  - On 3mm² we can fit 52 kernels
  - With about 70% utilization
  - The eFPGA is still completely reconfigurable for any architecture





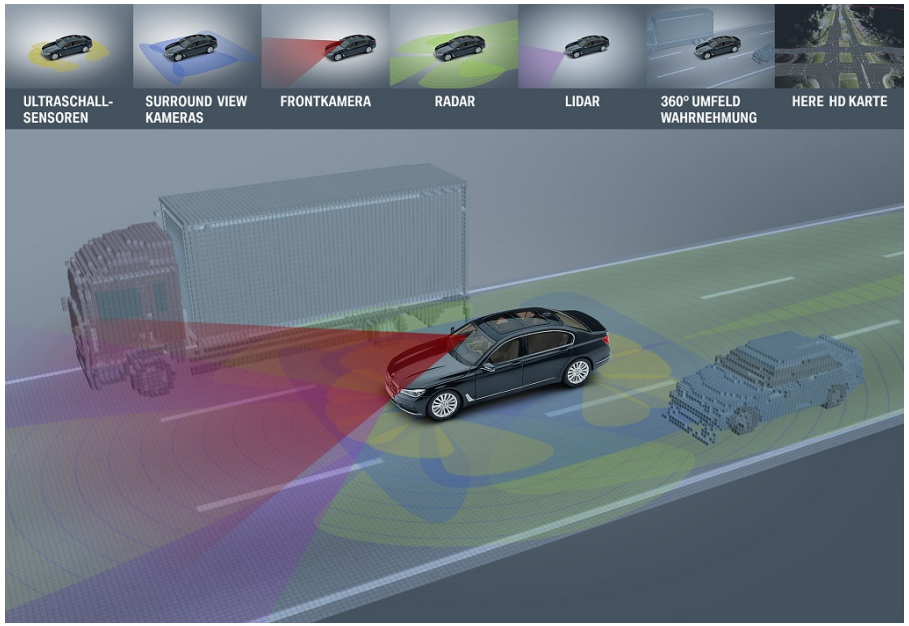
ZuSE-KI-mobil ... National Project

- **Efficient *CNN-based accelerator SoC*** for performance demanding applications such as ***embedded AI*** and **sensor fusion tasks** in autonomous cars

- New ***AI-Processor*** **types** need new approaches to guarantee **functional safety**

- Idea of a **scalable platform** to target various fields of embedded applications

ZuSE



PARTNERS

TECHNISCHE UNIVERSITÄT DRESDEN

Karlsruher Institut für Technologie

Dream CHIP

T3-technologies

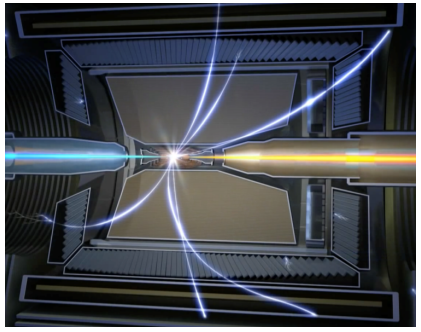
Leibniz Universität Hannover

[1] https://www.autonomes-fahren.de/bmws-bericht-zum-automatisierten-fahren/

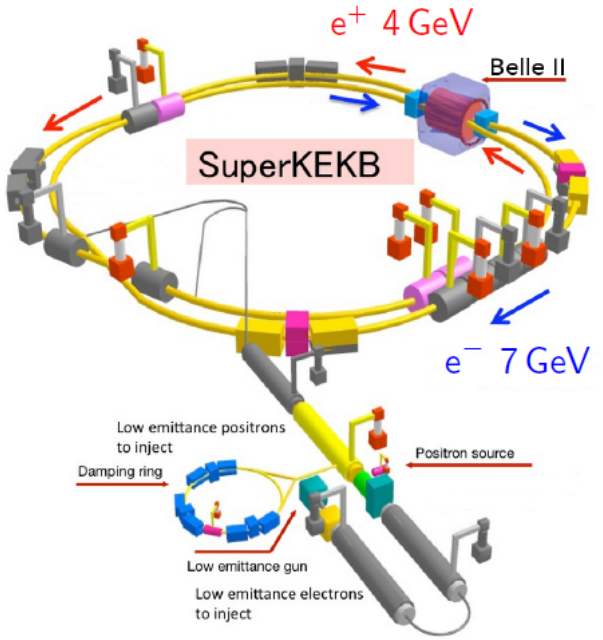
ML Application 2: *High Energy Physics*

- ***Particle Accelerator*** **experiments for discovery of new physics**
- **Modern experiments generating massive data rates:**
  - **~130 Gbyte/s** at Belle II at Tsukuba, Japan (newest experiment)
  - **~60 Tbyte/s** at High Luminosity-CMS at CERN, Switzerland

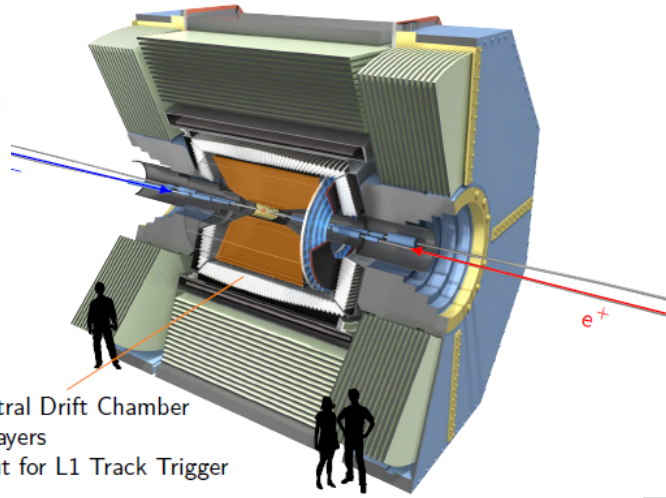
Complete *Readout* of detector is very expensive



Collision example at Belle II



Belle II Experiment



- Asymmetric accelerator at Tsukuba, Japan (60 km north of Tokyo)
- **World Record *Luminosity* (June 2020)**

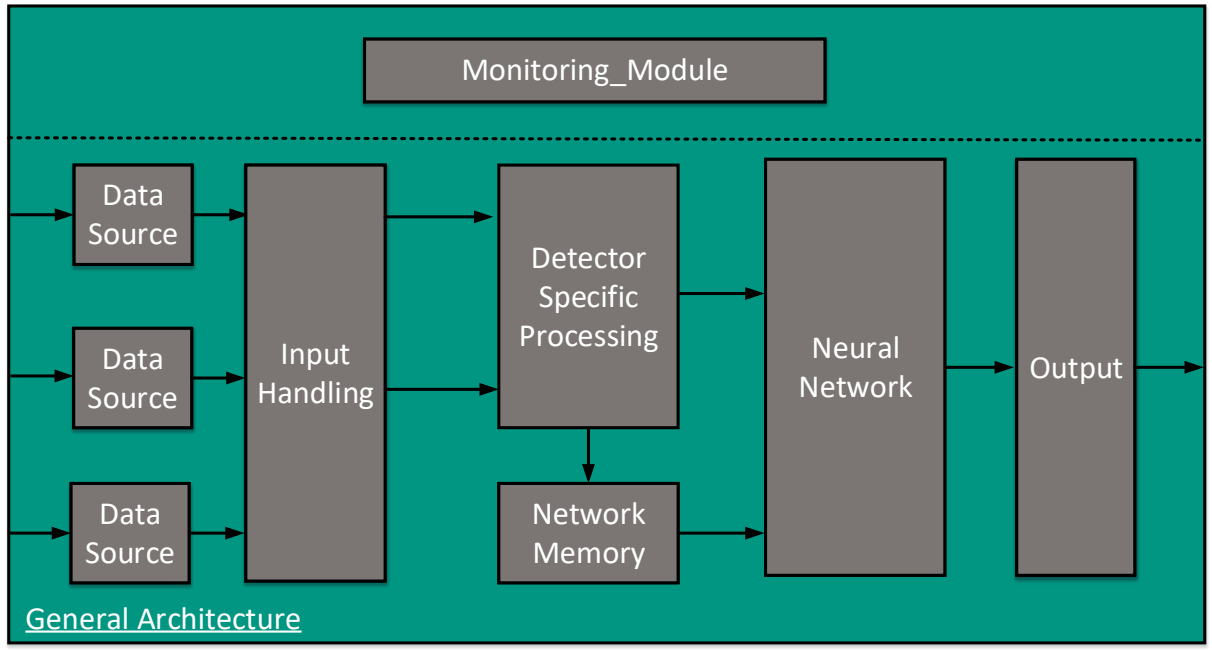
ML and FPGAs for *High Energy Physics*

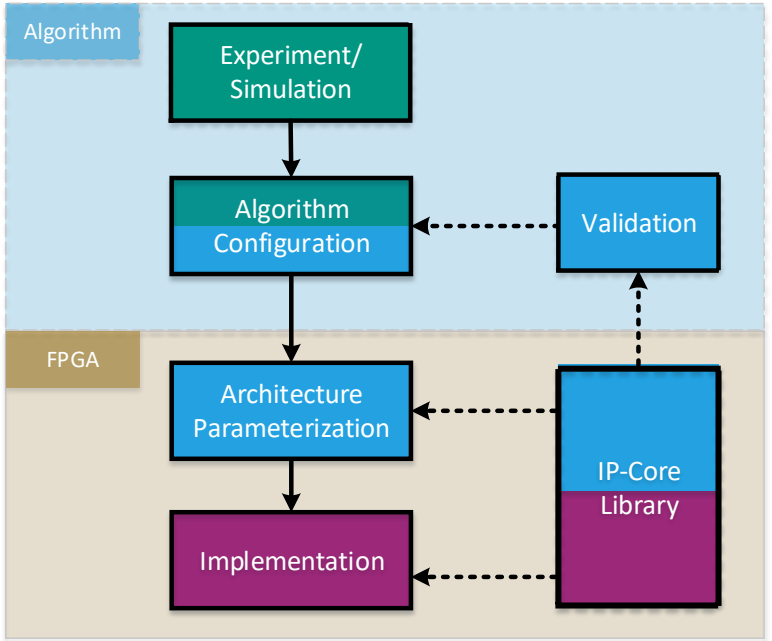
Karlsruhe Institute of Technology

How to make use of ML and FPGAs to reduce data rates ?

Use flexibility of FPGA‘s architecture and ML‘s algorithmic structure

Architecture and framework to adapt to current needs of the experiment in a semi-automated design flow

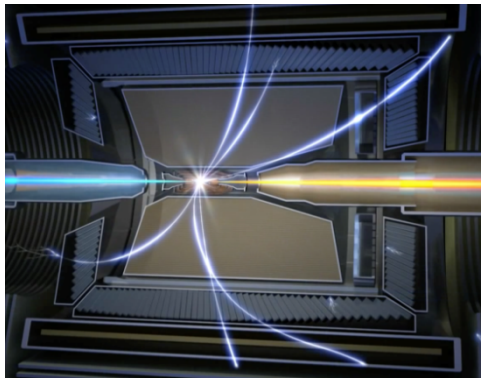


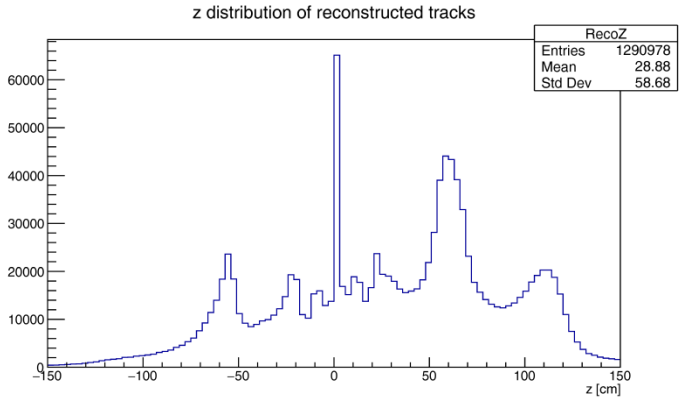


ML Application - Trigger Systems

Data rate Belle II = Readout_Rate * Readout_Size

Readout_Rate ~ 130 kHz



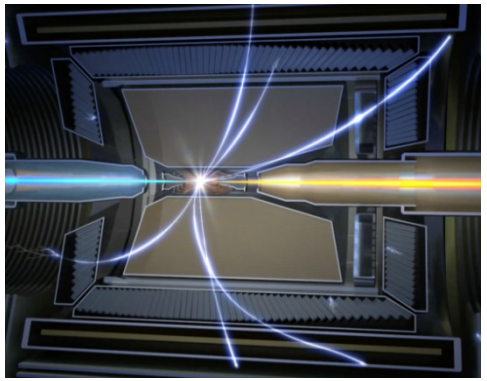


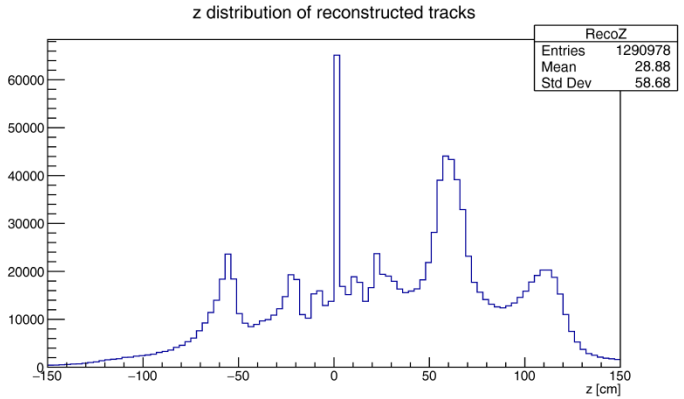
Target = 30 kHz * ~0.1 MByte/Event

ML Application - Trigger Systems

Data rate Belle II = Readout_Rate * Readout_Size

Readout_Rate ~ 130 kHz

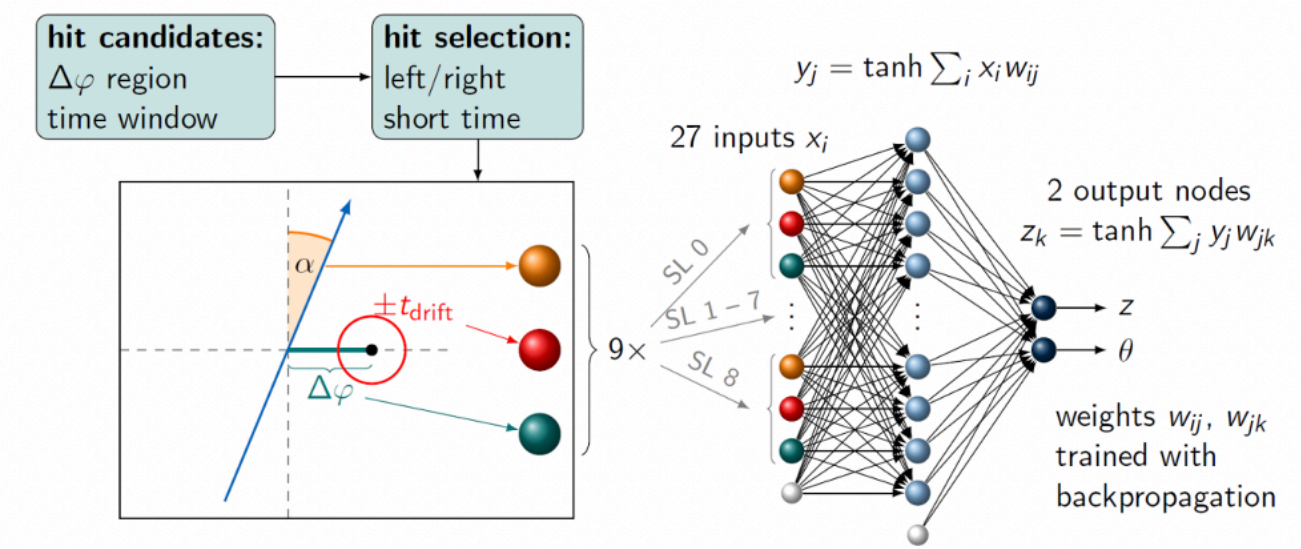




Neural z-Vertex Trigger

300 ns Latency for an estimation

32 MHz Input frequency

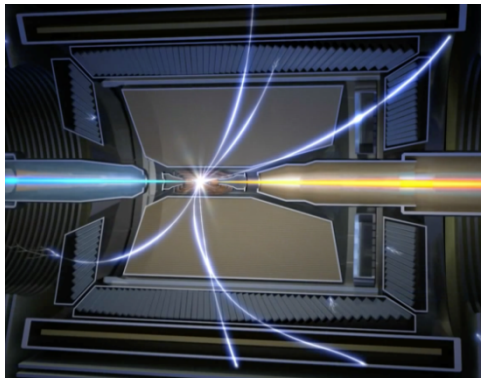


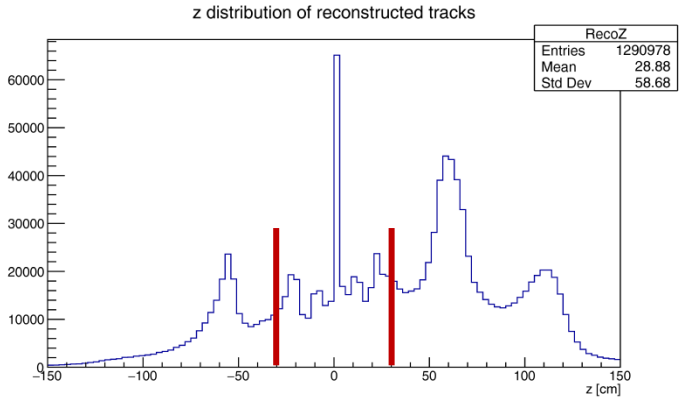
Target = 30 kHz * ~0.1 MByte/Event

ML Application - Trigger Systems

Data rate Belle II = Readout_Rate * Readout_Size

Readout_Rate ~ 130 kHz

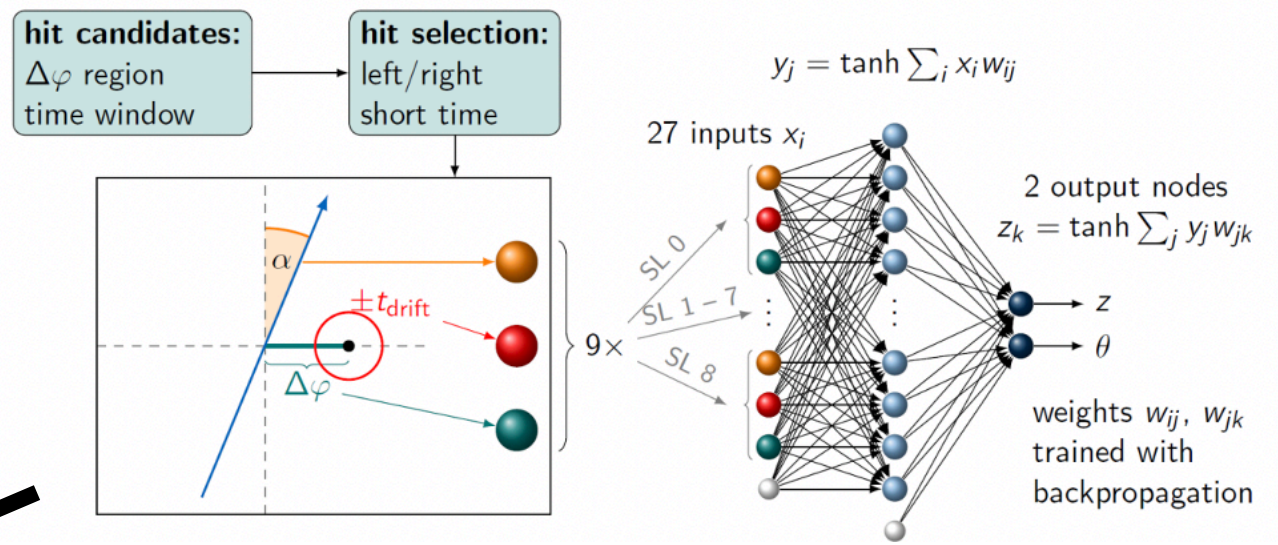




Neural *z-Vertex Trigger*

300 ns Latency for an estimation

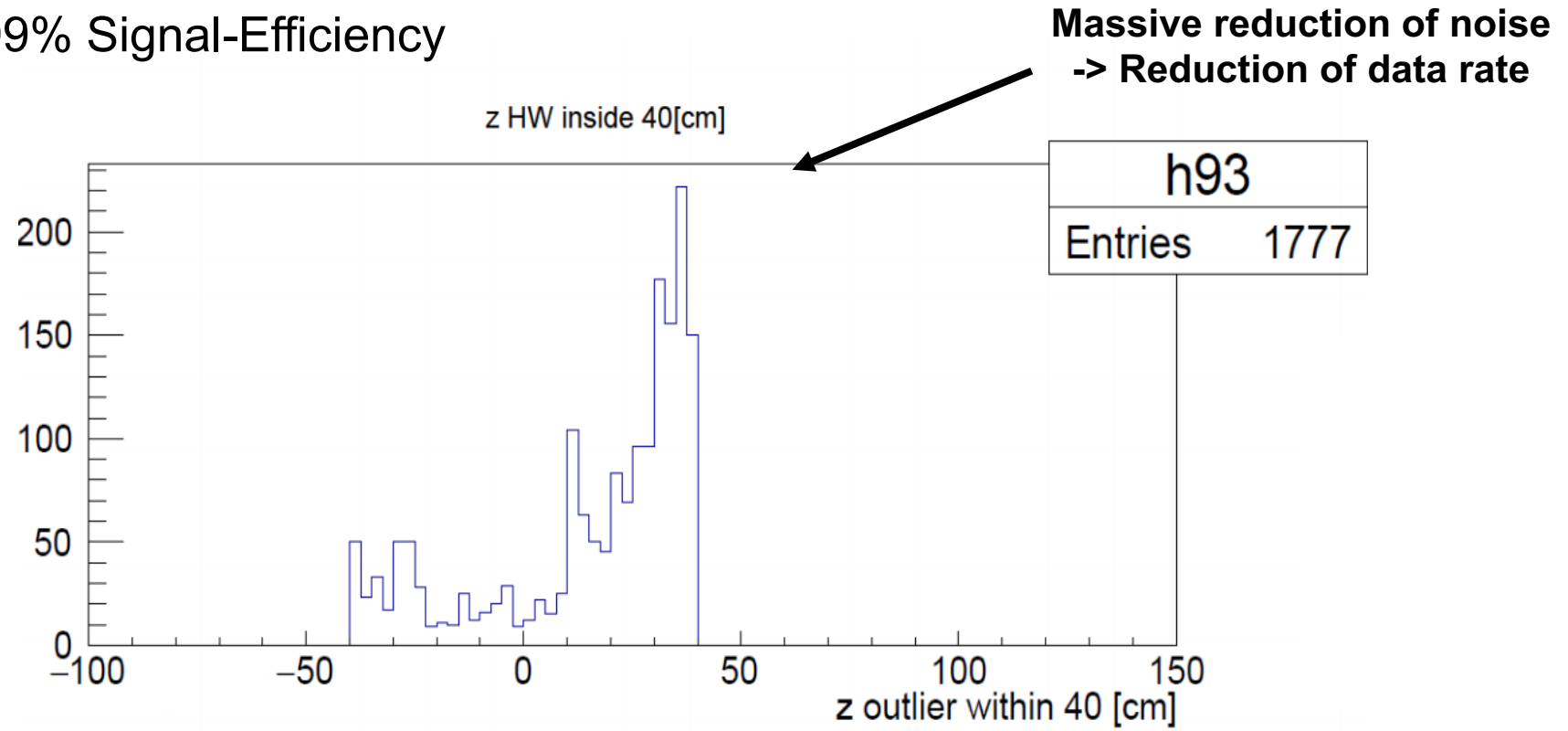
32 MHz Input frequency



Target = 30 kHz * ~0.1 MByte/Event

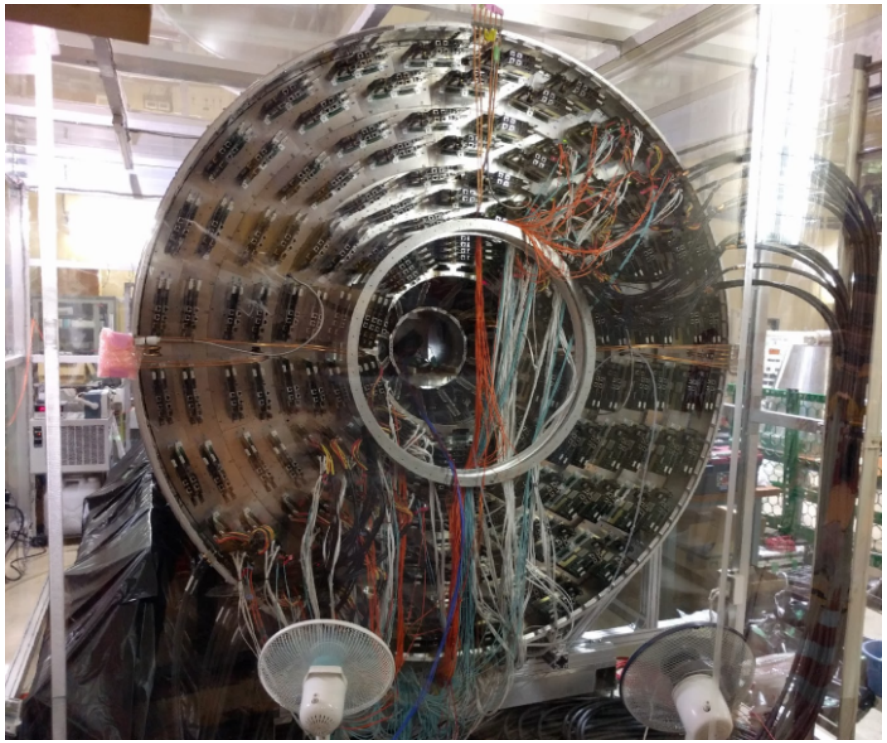
ML Performance in *z-Vertex Trigger System*

- **Operational at the Belle II experiment**
- **Latest results from June operation**
  - Average ~37% Reduction of Noise -> Readout_Rate
  - Average ~99% Signal-Efficiency



Flexibility Example – *Early Testing*

- **Trade-Off between Preprocessing und *Neural Network***
  - Less neurons/resources for neural network -> Use free resources to extend scalable preprocessing

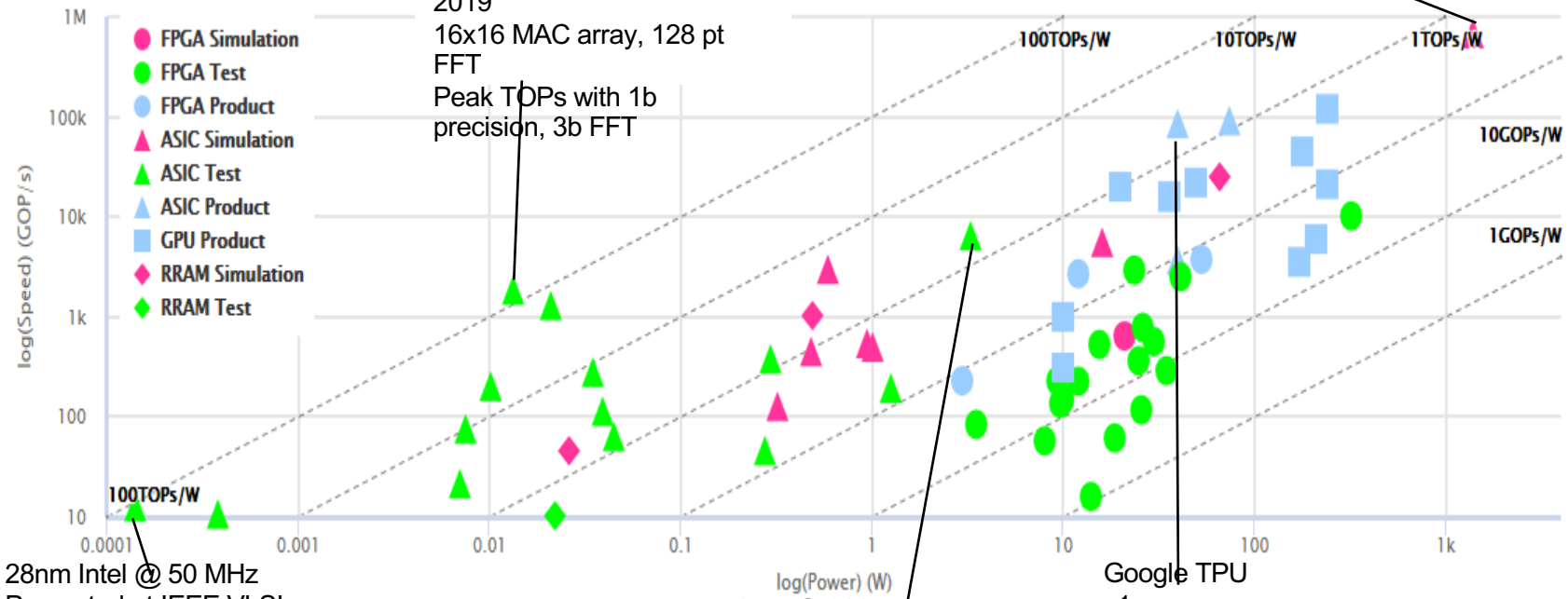


Detector outside for the accelerator for early testing

Testing with incomplete *Data Readout*

Conclusions & Outlook

State of the art Research



Neural Network (NN) Accelerator Comparison

- Power in Watt vs. „Speed“ in GOP/s

Source: https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

State of the art Chips …

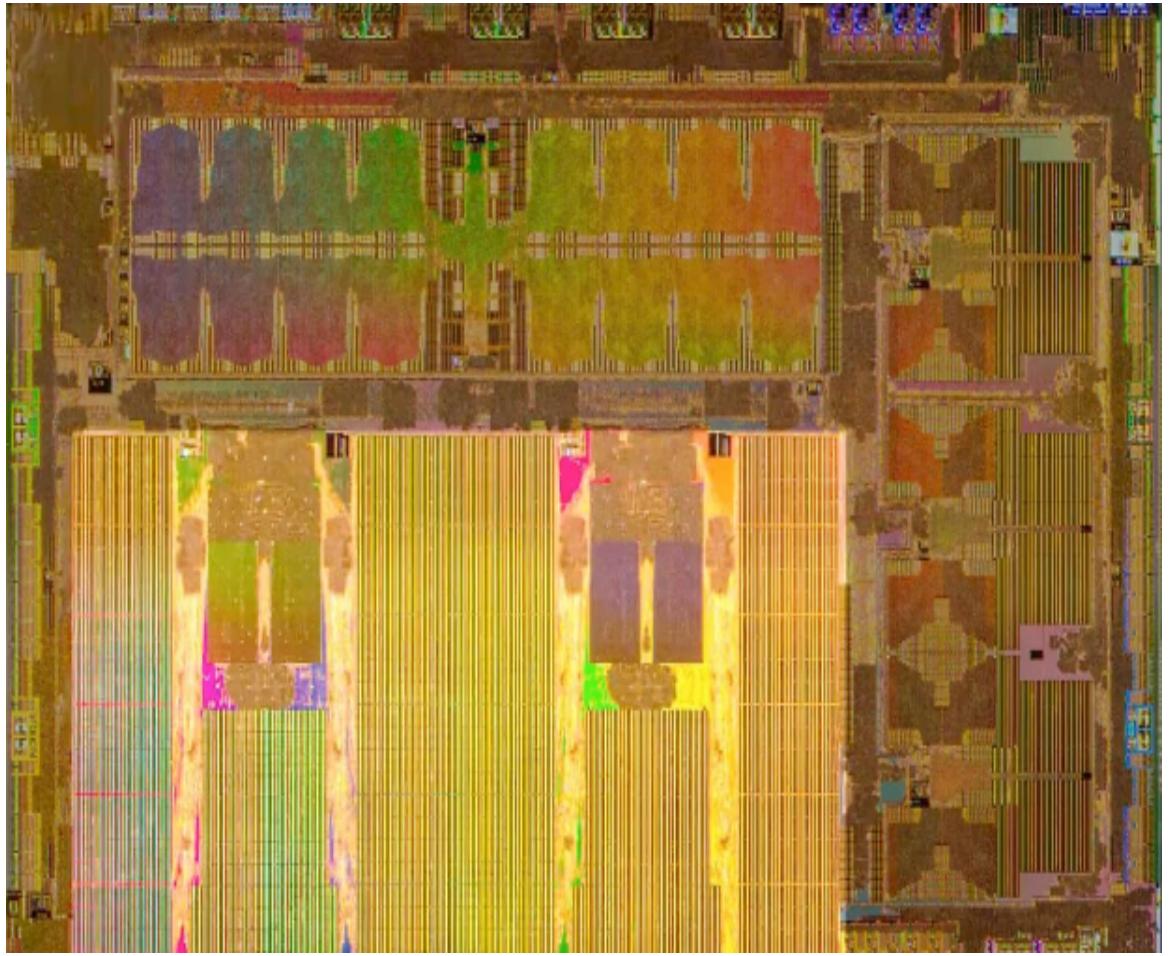
Tesla – Full Self Drive Chip

- Six ARM CPUs
- Dedicated GPU
- **Two NPUs**

  ***Neural Processing Unit* (NPU)**

  - ***Systolic Array***

    with Memory
- 36 TOPs @ 32 W

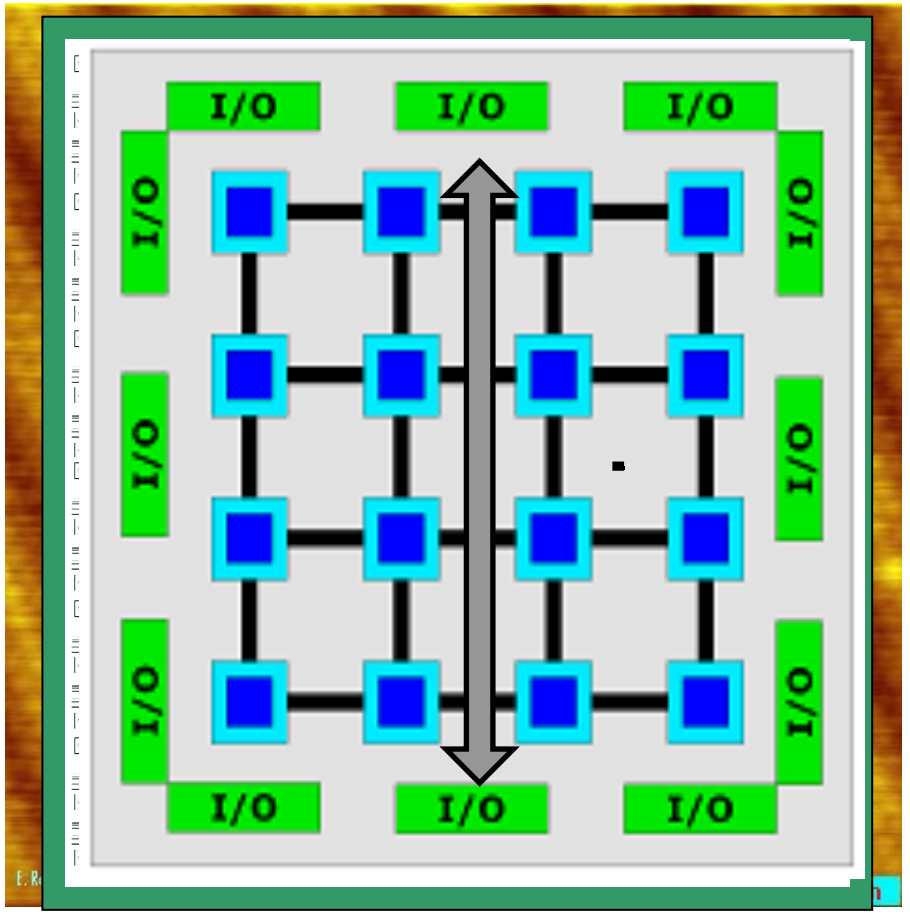


Enabler: Embedded Reliable *Neuromorphic* Performance

Karlsruhe Institute of Technology

Technologies & Functional Intelligence:

- Technologies:
  - Silicon, Optics, Nano, … Quantum ...
- Parallel Programming & Abstraction Layers
- Dynamic Networks & Offline/Online Trade-offs
- Correctness & Reliability
- Verifiable *ML-Models* & Metrics
- *Safety* & *Security* Codesign
- Cross-layer System Integration

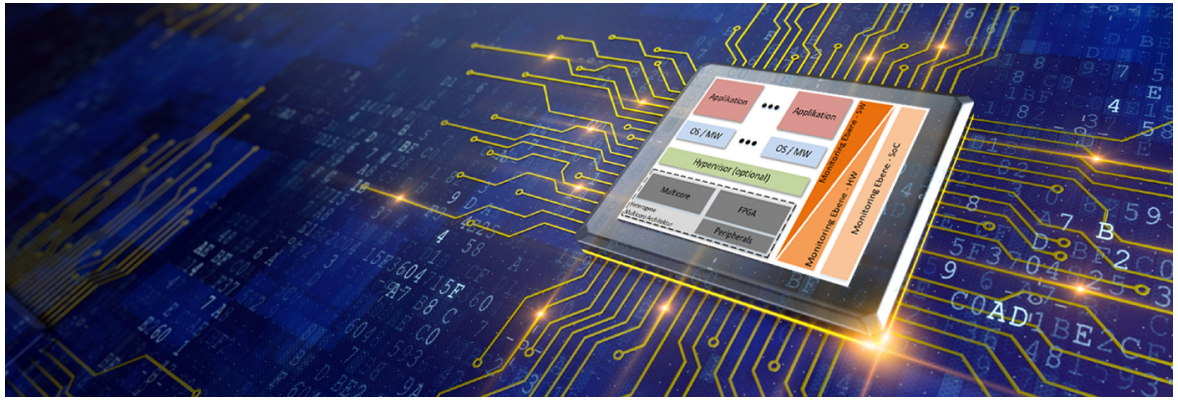


Research, Education & Innovation

- Hardware/Software/Architecture/Algorithms Codesign & *ML*
- Tools & Methods for Programming & Resource Management
- Challenges: Embedded Integration of *ML*, *HPC* & *Reliability* -> *Open Hardware* as possible Enabler …

Thanks for your attention!

… Silicon will live for ever!



Contact :

Prof. Dr.-Ing. Dr.h.c. Jürgen Becker

Karlsruhe Institute of Technology (KIT)
Institute for Information Processing Technologies (ITIV)
Engesserstr. 5
76131 Karlsruhe, Germany
www.itiv.kit.edu
Email: becker@kit.edu