Karlsruhe Institute of Technology

Real-Time Trigger and Online Data Reduction based on Machine Learning Algorithms on FPGAs

Steffen Bähr

Institute for Information Processing Technologies (ITIV)

Focus of this talk

- Specialized hardware for neural networks
  - Focus on FPGAs
  - Close to sensor integration
- Choices of possible FPGAs
- Frameworks for architecture configuration
- Trigger application case from the Belle II experiment

The ITIV of KIT

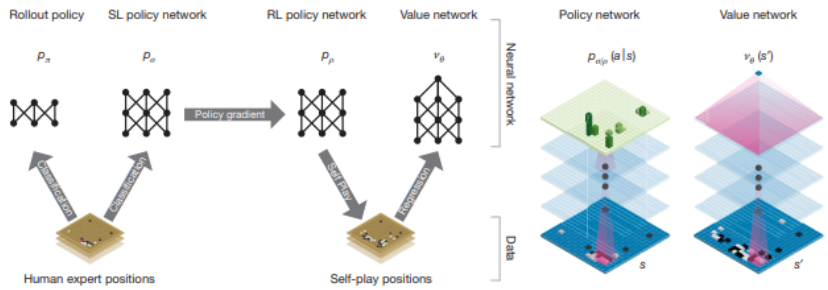
- Electrical Engineering Institute
- Focused on Processor Architectures/ Embedded Computing
- Optical Engineering
- Testing for Autonomous Driving
- …

© ZKM | Zentrum für Kunst und Medien, Foto: Jonas Zilius

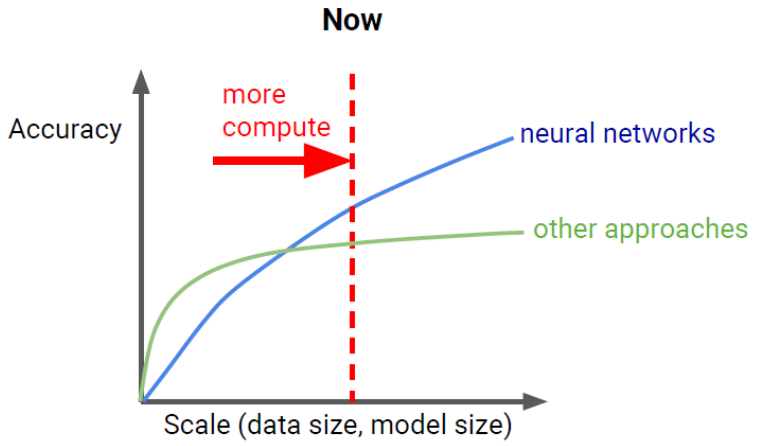
- Founded by Karl Steinbuch
- Established the term Informatik
- Invented the Learning Matrix
- Early Precursor of the Neural Networks

- Lernmatrix discussion paper
- https://www.facebook.com/KIT.Karlsruhe.Official/videos/2165530873702056/

Machine Learning on the Rise

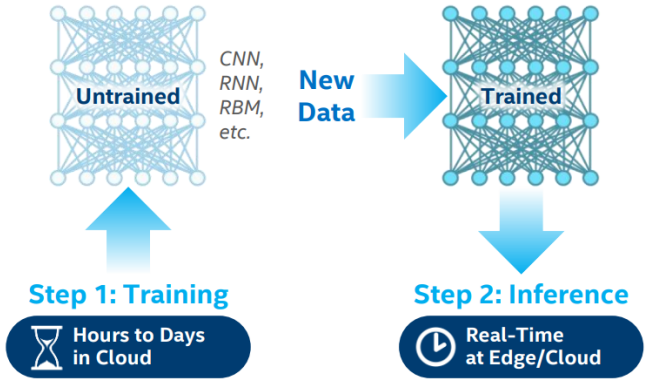


- Google DeepMind AlphaGo [1]
  - First time a human lost to AI playing GO
- Google AI Assistant [2]
  - Automated restaurant reservation
- Machine Learning for large scale data processing [3]
  - Cost-efficient
  - Reasonably precise
  - Enabled by availability of processing resources

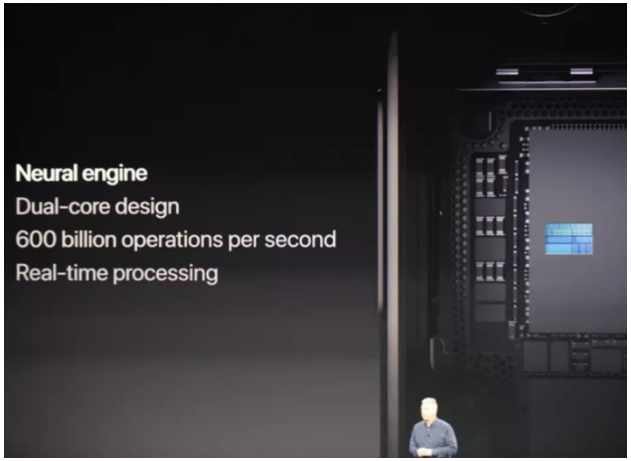


Hardware for Machine Learning

- ML typically associated with large data centers
- Development of dedicated hardware solutions for ML tasks
  - Edge computing
  - Efficient local computing
- Push towards closer integration sensors and embedded devices
  - Dedicated neural processing
    - Image processing to add blur effect
    - Fast unlock mechanisms
    - Apple A11 Bionic [5]
    - Kirin 970 [6]



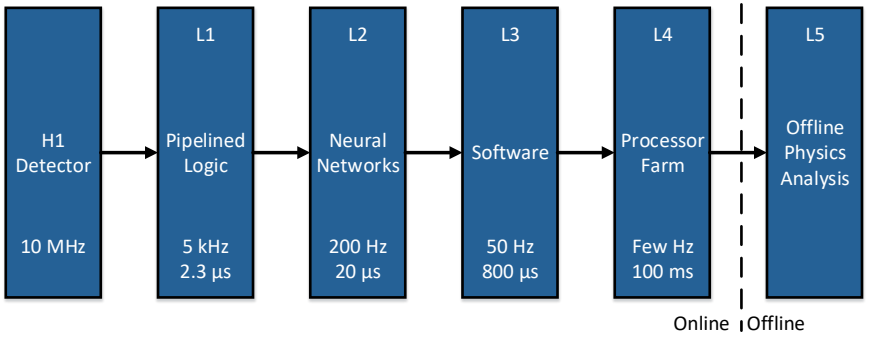
[4]



Data in High Energy Particle Physics

- Generation of massive data rates at modern experiments
  - **480 Tbit/s** at HL-CMS (CERN) [7]
  - **324 Million** channels at 1 GHz at Gigatracker [8]

- Trigger and data reduction systems
  - Trigger decides when to readout
  - Data reduction what to readout
  - Implemented close to the sensors
  - Multi-staged often FPGA-based

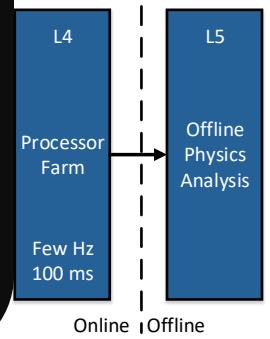


- Steadily increasing complexity
  - Tighter requirements from µs to ns latency
  - Higher amount of data channels to support
  - More complex behavior of the experiment

Data in High Energy Particle Physics

- Generation of massive data rates at modern experiments
  - **480 Tbit/s** at HL-CMS (CERN) [7]
  - **324 Million** channels at 1 GHz at Gigatracker [6]
- Trigger
  - Trig
  - Data
  - Impl
  - Mult
- Steadily increasing complexity
  - Tighter requirements from µs to ns latency
  - Higher amount of data channels to support
  - More complex behavior of the experiment

How to use ML's potential for trigger and data reduction?

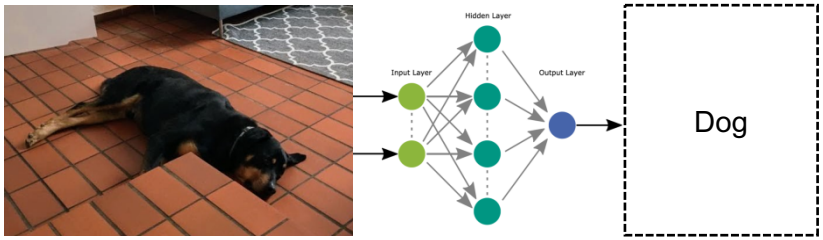


Challenges

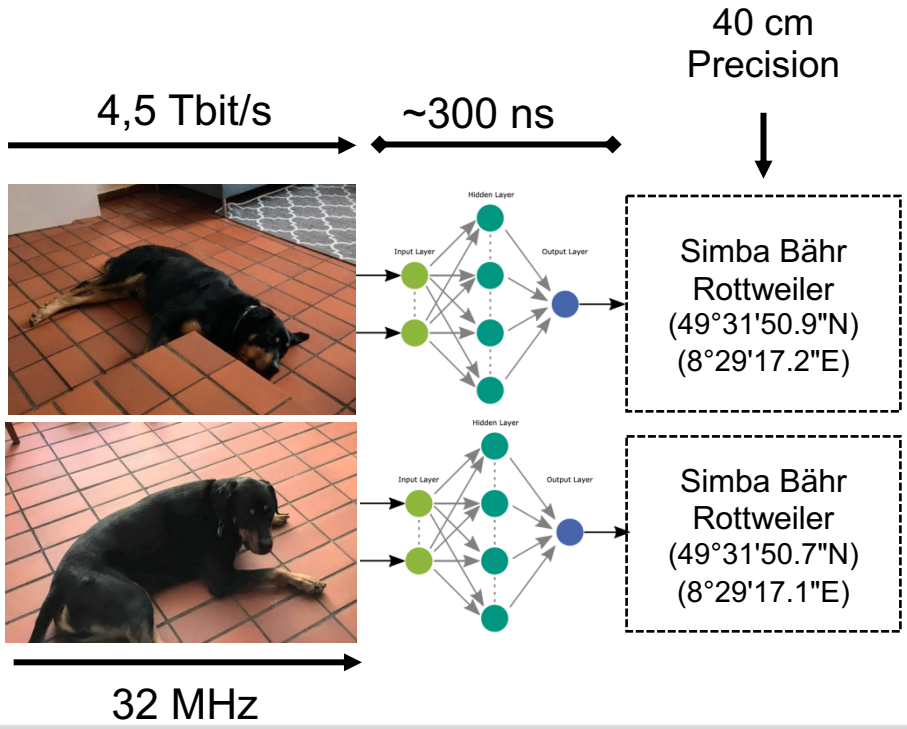
- **Integration**
  - Support of High Data Rates
  - High demand for high-speed IO
  - Different parallel asynchronous sources
- **Functional**
  - High precision classification
  - Online Monitoring for Validation
- **Hardware Architecture**
  - Low-Latency
  - Pipelined and free of dead–time
  - Flexible for changing conditions
  - Real-Time Processing

Challenges

- **Integration**
  - Support of High Data Rates
  - High demand for high-speed IO
  - Different parallel asynchronous sources
- **Functional**
  - High precision classification
  - Online Monitoring for Validation



- **Hardware Architecture**
  - Low-Latency
  - Pipelined and free of dead–time
  - Flexible for changing conditions
  - Real-Time Processing



Presented Application Cases

- **Neural network based trigger system at first level**
  - Architecture fulfilling all requirements
  - Design Flow
  - Integration
  - Experimental Results
  - Upgrade Prototype

- **Online Cluster Analysis for particle identification**
  - Architecture fulfilling all requirements
  - First NeuroBayes implementation
  - Design Flow
  - Integration

MACHINE LEARNING ON SPECIALIZED HARDWARE

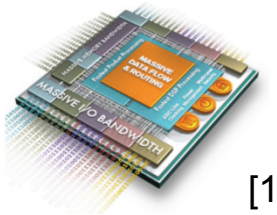
Hardware for Machine Learning

Where Deep Learning has been traditionally happening …

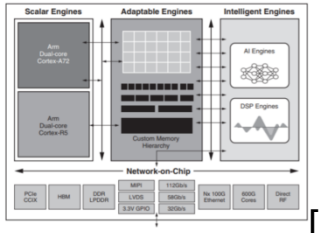
Where state-of-the-art Development is moving towards …

[9]

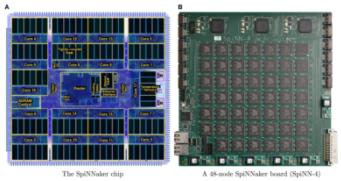
[10]



[11]



[12]



[13]

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

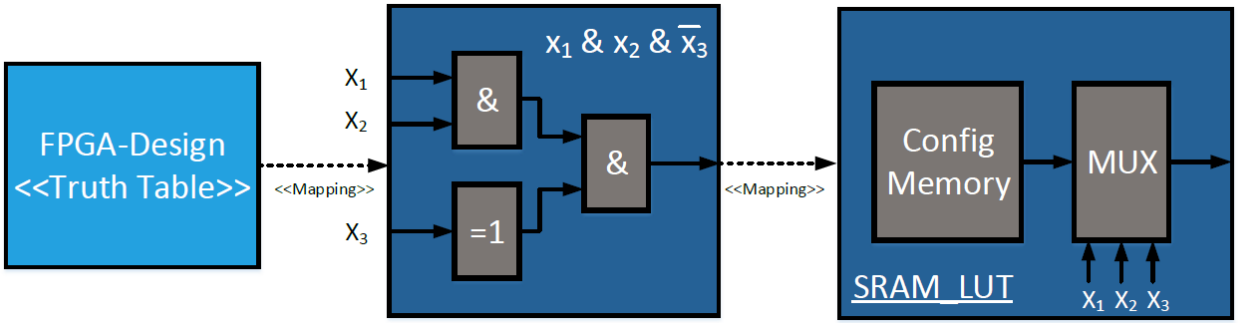
GPP/GPUs for ML

- GPP in ML
  - Highest programmability
  - Limited Parallelism
  - Traditional memory structures

- GPUs in ML
  - Probably highest productivity
    - Programming Tools + Performance
  - Not deterministic latency (internal scheduling)
  - Interfacing often dedicated i.e. PCIe
  - No additional flexibility compared to FPGAs
  - Power Efficiency

FPGAs

- Highly flexible processing resources
  - Can be configured to recreate any logical function
  - Deterministic behaviour
  - Power efficient with sparse operations

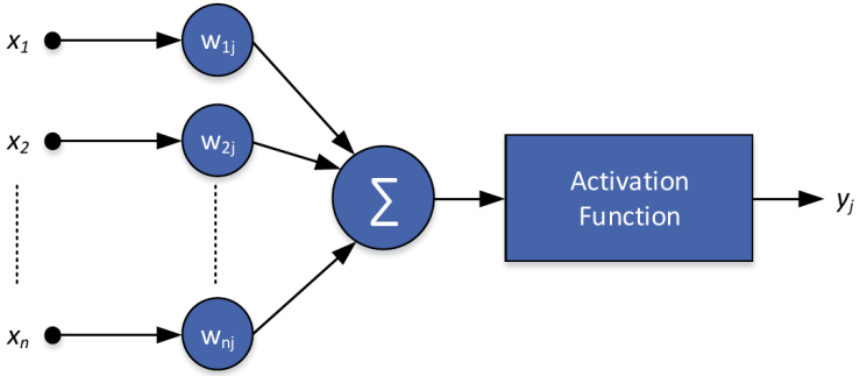


- Special processing units
  - Gigabit Transceivers of for high data rates
  - ADCs and DACs for high frequencies (RF-SoC)
  - Digital Signal Processors (DSP) for multiply and accumulate
  - Local Block RAM Memory (BRAM) for local access

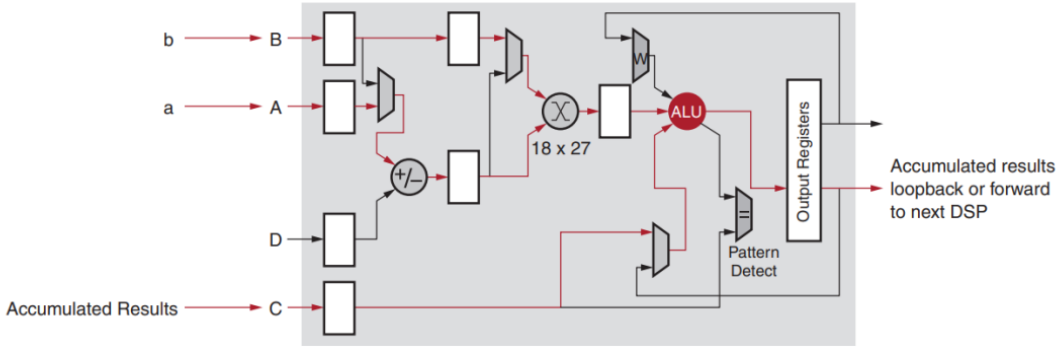
- Best Power-Performance-Flexibility characteristics

Neural Networks on FPGAs

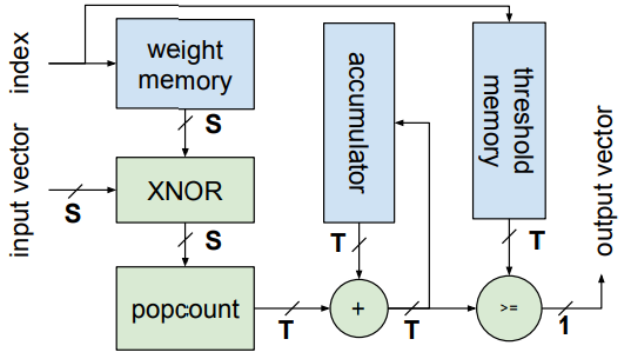
- Artificial Neuron Processing
  - Heavy DSP usage
  - On-chip memory for weights
  - Fixed point operation
    - Only operations with int8
  - Binary networks with XNOR
    - Less Computational Effort
    - Less Accuracy (~3% loss on AlexNet)



[15]

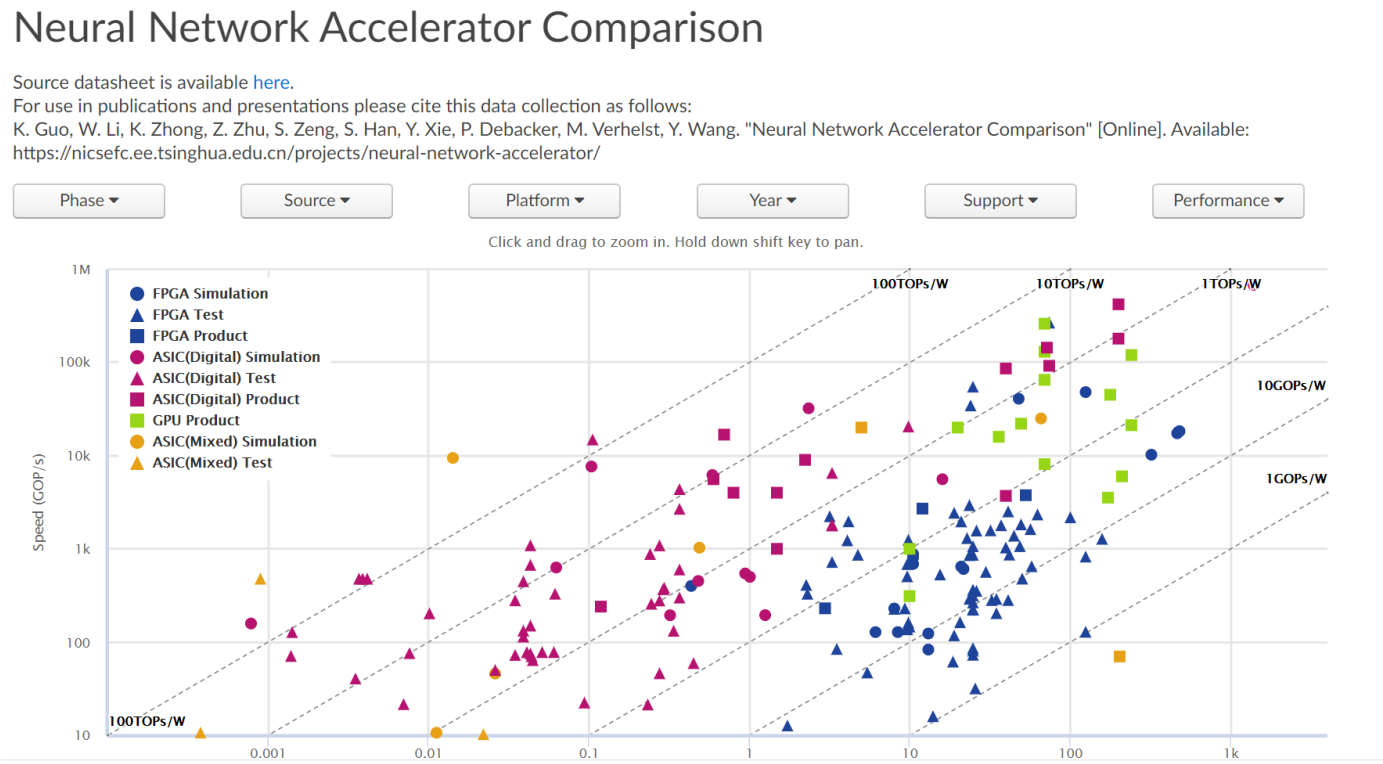


[14]



Neural Network Accelerators

- Study about neural network accelerators by tsinghua university :
  - https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/



FPGA Selection

- Viability of FPGA platform often determined by
  - Number of DSP units and on-chip memory

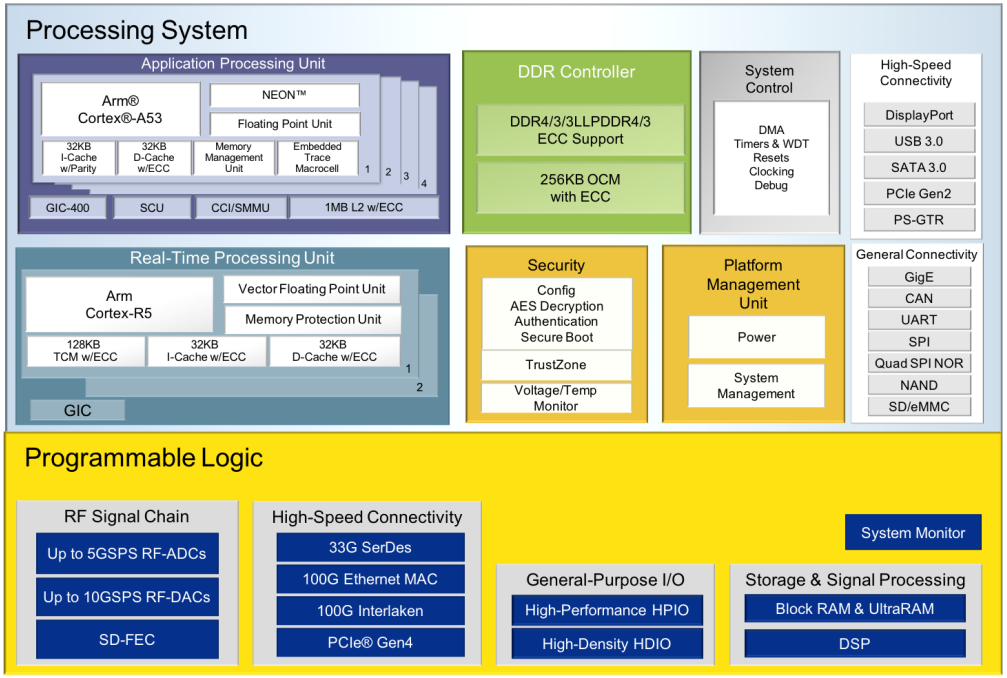
Resource	XC6VHX565T	XC7VX1140T	XCVU190	VU13P
Slices	566,784	1,139,200	1,074,240	1,728,000
DSPs	864	3,360	1,800	12,288
BRAM	32.832 mb	67.680 mb	132.9 mb	454.5 mb
GTX	48	-	-	-
GTH	24	96	60	-
GTY	-	-	60	128

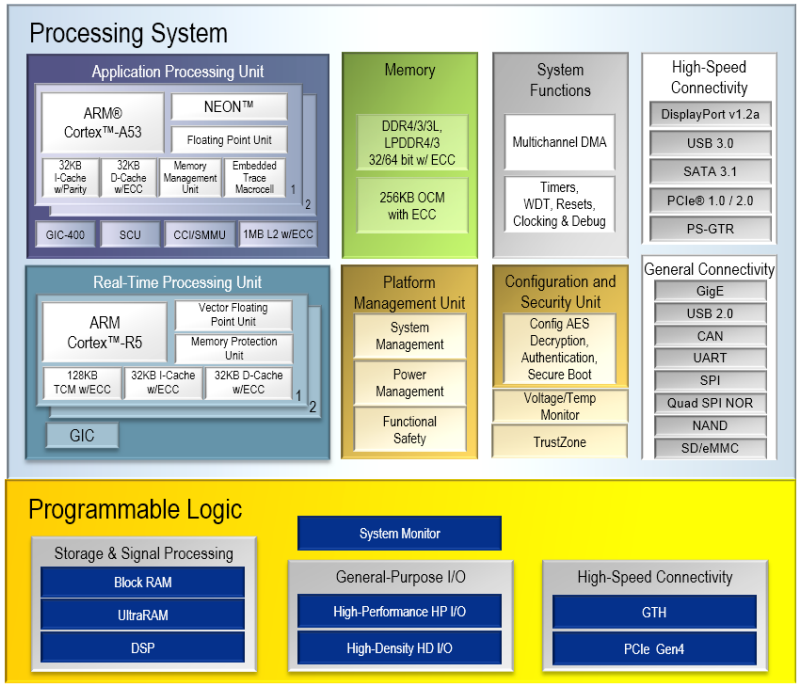
  - On-chip BRAM for lowest latency weight loading
    - Neural Trigger Example :
    - 5 Networks in parallel held in memory
    - 5*18 Bit*2349 = 211 kBit
    - 10% Total memory
  - Parallel MAC operations = #Number of DSPs
    - Neural Trigger Example
    - 2349 total MAC operations
    - 470 DSPs used
    - Latency:
      2349/470 = 5 Clock Cycles
- Secondary features :
  - General Purpose Processing
  - Interfacing

FPGA Selection (II)

- Specialised FPGAs
- ZYNQ Ultrascale+
- RF-SoC

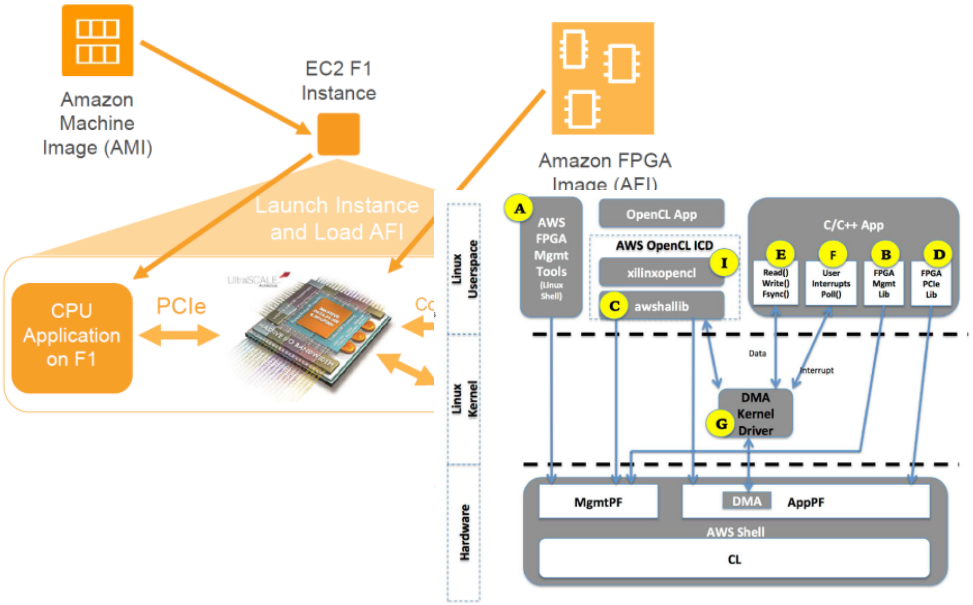
[16]



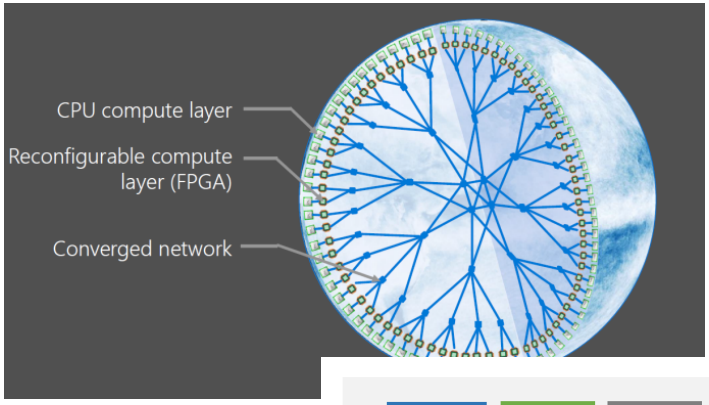


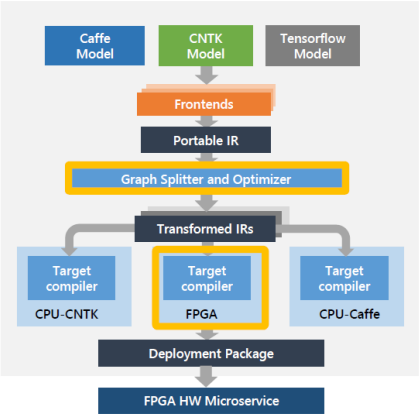
Cloud-based Application

- **Amazon AWS F1** [24]
  - FPGA Based Acceleration in the Cloud
  - SDK for Usability
  - FPGA Deployment scheme for integration into infrastructure



- **Microsoft Brainwave** [25]
  - Large Scale FPGA Network
  - Framework FPGA Mapping
  - ML Tool Support

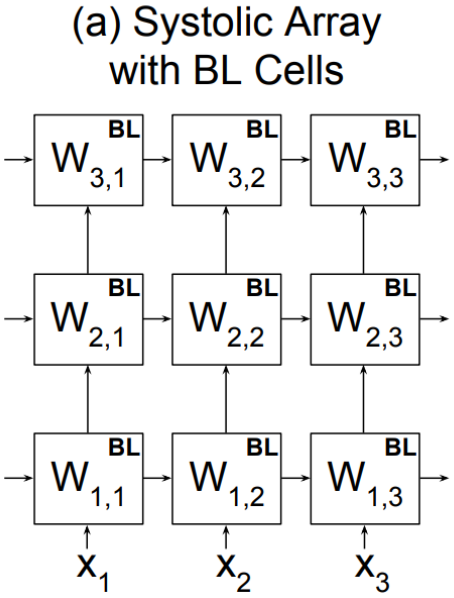




ASICs for ML

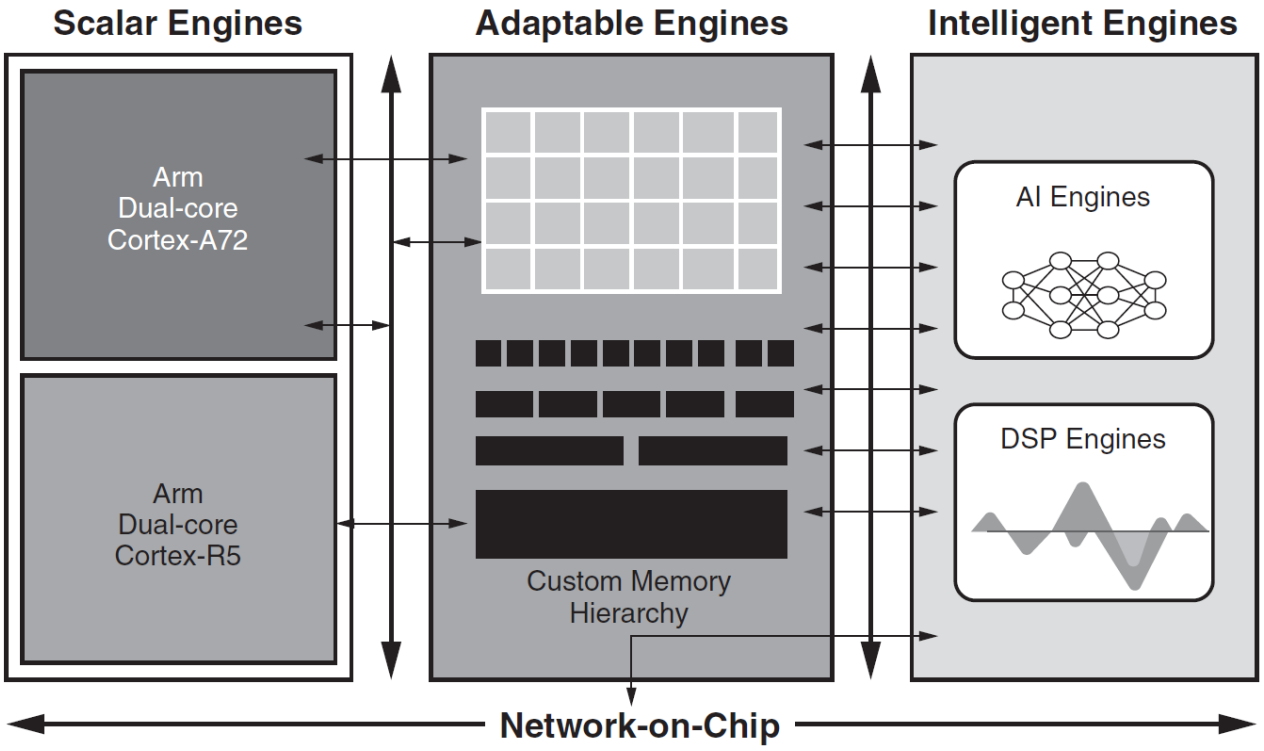
- **NVDLA**
  - AI open source processing Cores
  - Developed by NVIDIA
  - Chosen by ARM for dedicated coprocessing
  - Programming Toolchain
- **Intel Nervana**
- **Google Edge TPU for IoT**
- **…**

- **Tesla ASIC**
  - Dedicated CNN Accelerator
  - Used for a Self-Driving SoC
  - Achieves 36 TOPS



Future platforms – Xilinx ACAPs

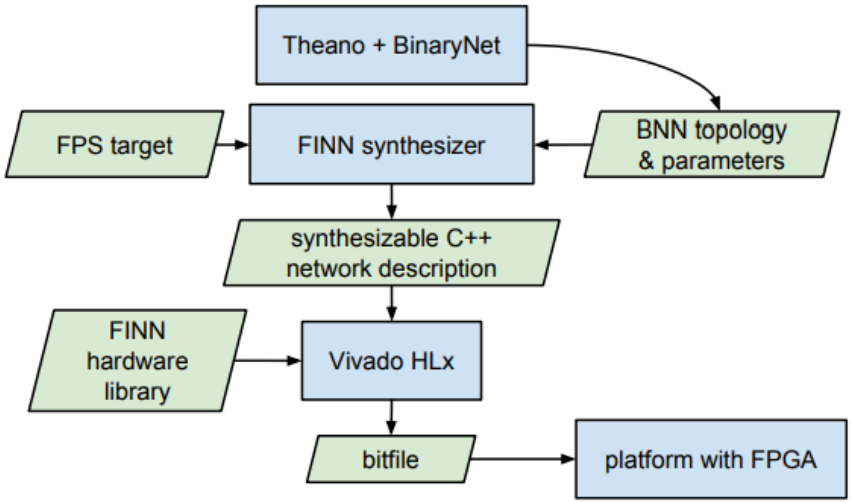
- Combination of different processing structures
  - ARM Processor for general purpose (Slow Control)
  - Flexible FPGA-like array (Preprocessing)
  - Dedicated AI-Acceleration (Network)



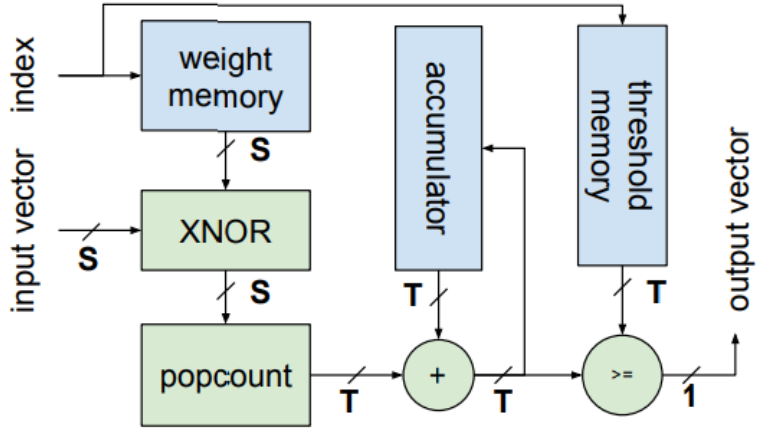
[12]

FRAMEWORKS FOR ARCHITECTURE INFERENCE

FINN : Framework for Fast, Scalable Binarized Neural Network Inference



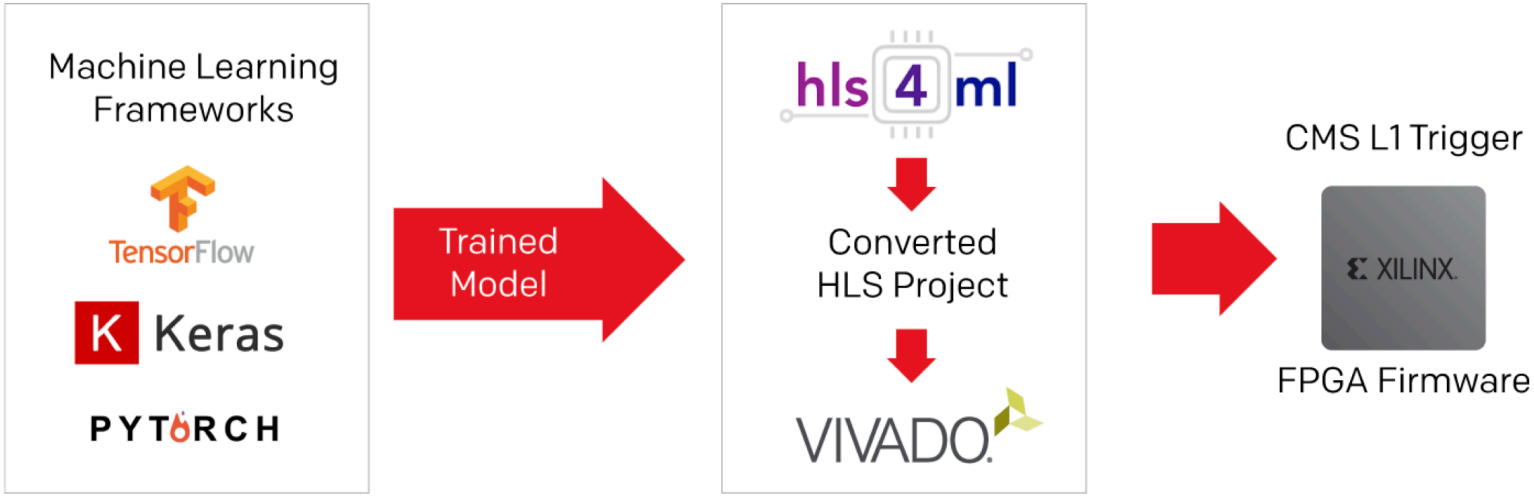
[14]



- Framework co-developed by XILINX, used as a reference
- Integration of popular DNN Tooling (Theano)
- Early performance and efficiency estimation using Roofline model
- Vivado HLS supported network description
- Optimized Binarized Neural Networks Library of IP Cores

HLS4ML

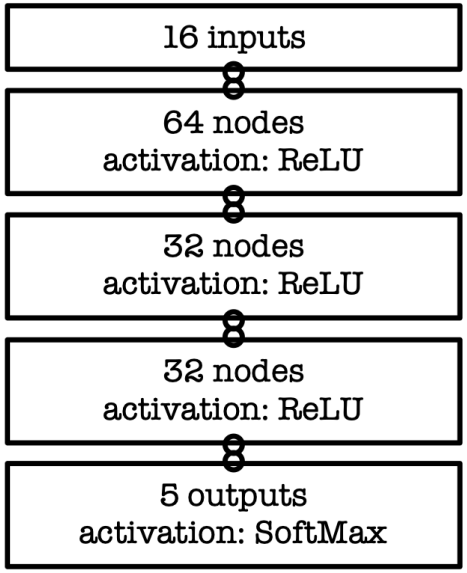
- Vivado HLS Centric approach to implement neural networks [10]
- Support of modern ML frameworks
- Library of algorithms such as CNN, RNN ... Supported
- Efficient compression of the network
- Architecture instantiation
- https://fastmachinelearning.org/hls4ml/



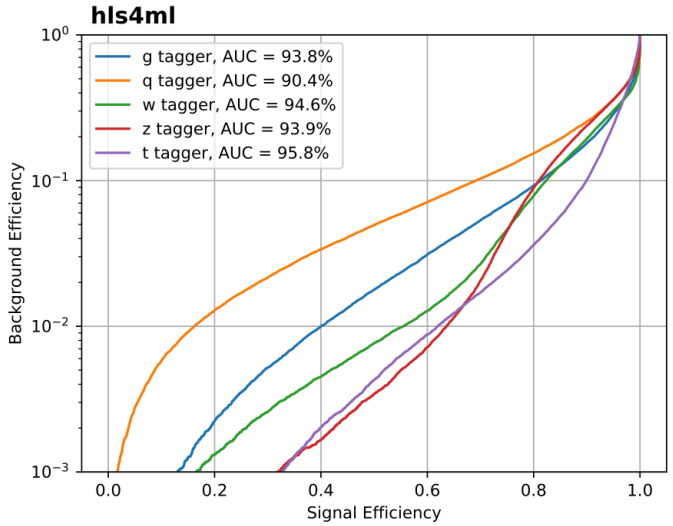
[17]

HLS4ML – Case Study

- Jet substructure classification for LHC



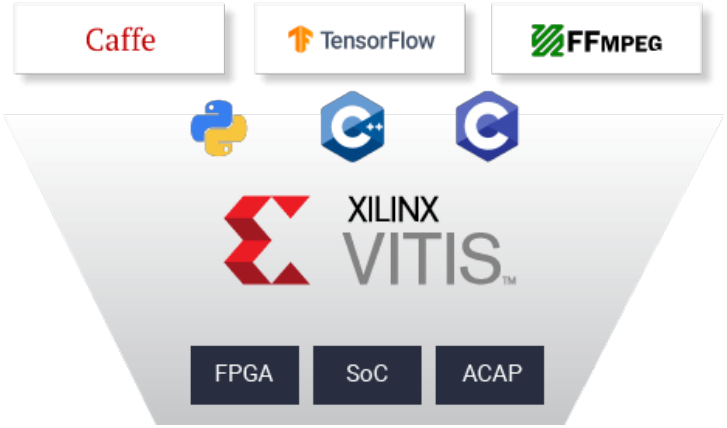
[18]



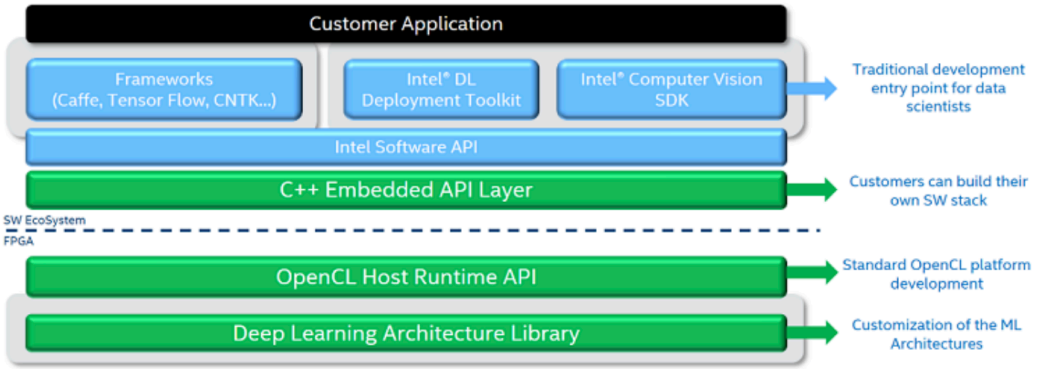
Network	Uncompressed network	Compressed network
AUC / Expected AUC	99.68%	99.55%
Parameters	4389	1338
Compression factor	-	3.3×
DSP48E	3329	954
Logic (LUT + FF)	263,234	88,797
Latency	75 ns	75 ns

Vendor Tools

- Xilinx Vitis Approach
  - Unified Tool Environment
  - Interfacing with ML Tools
  - Support of High-Level Languages
  - Configuration of underlying Hardware
- Intel FPGA DLA



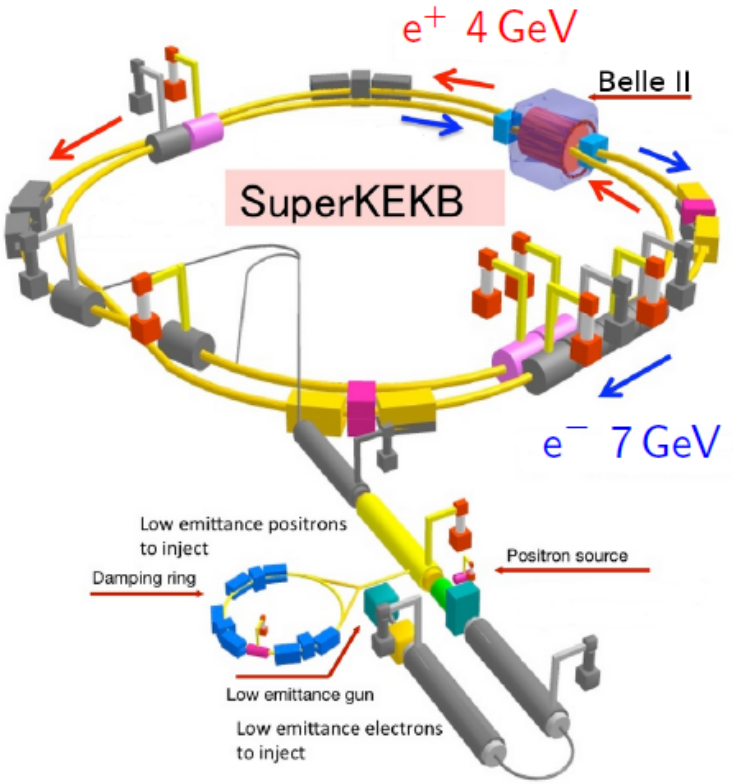
[16]

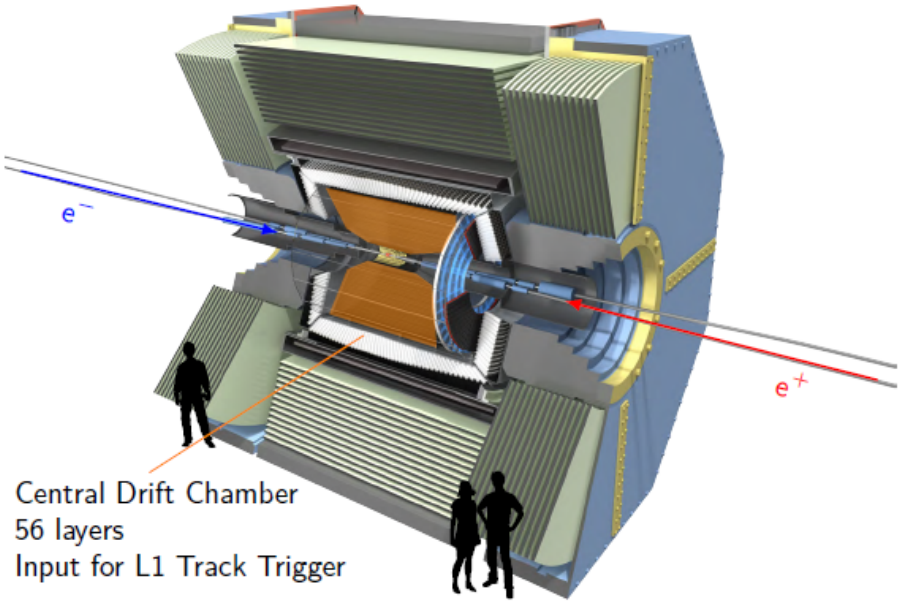


[19]

APPLICATION CONTEXT

Belle II



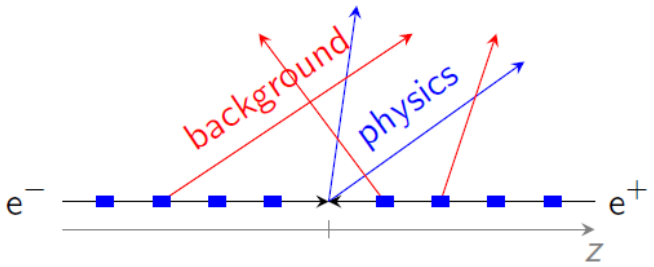


[20]

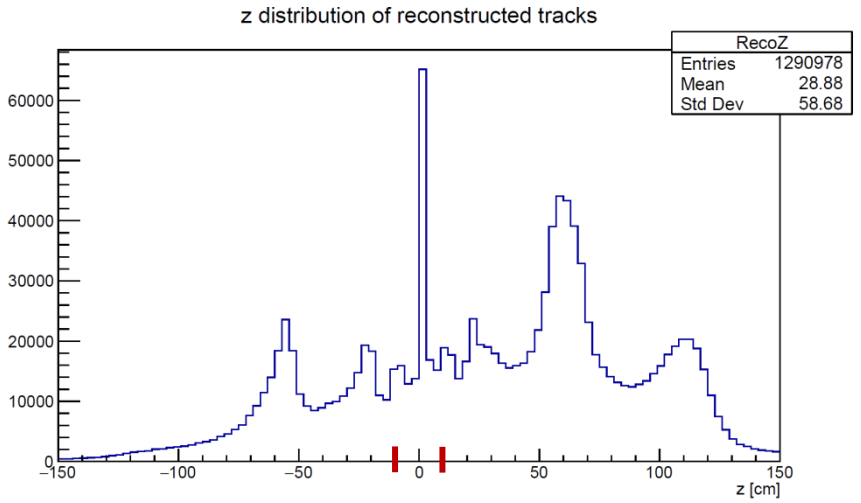
- Located at Tsukuba, Japan
- Asymmetric e^+e^- collider
- First Collisions since April 2018
- Targets world record luminosity

L1 z-Trigger for Belle II

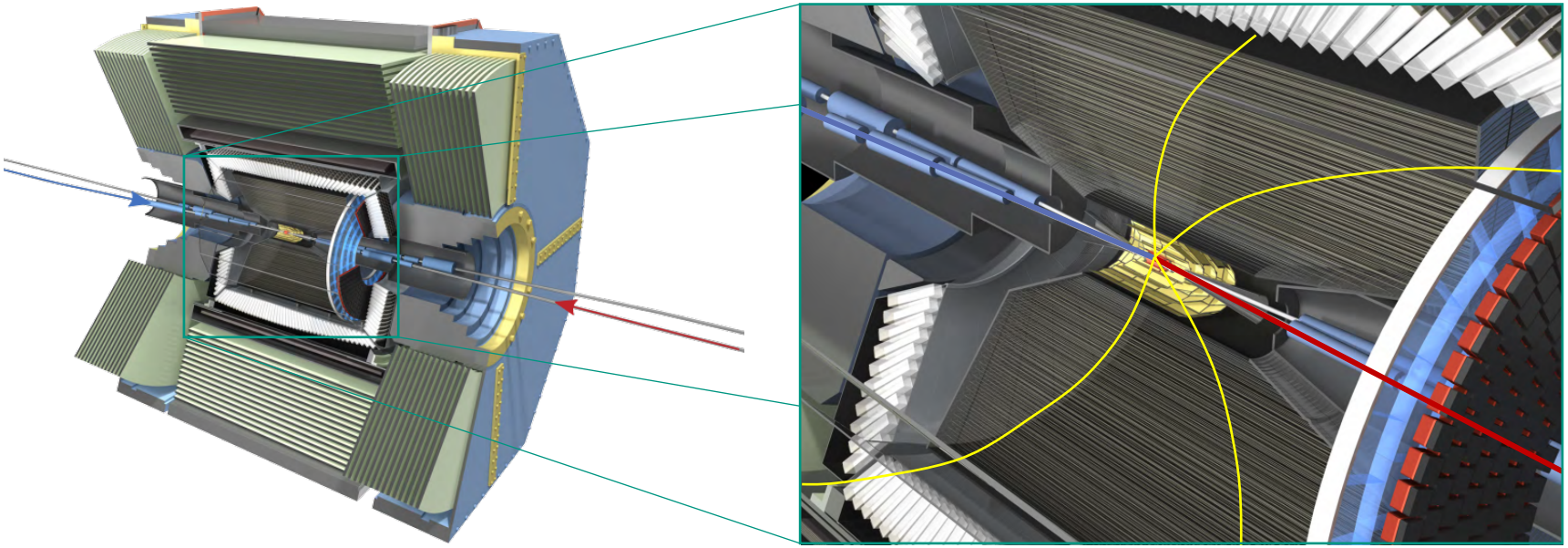
- L1 track suppression
  - Many particles are not related to collisions
  - Tracks outside the point of collision have to be suppressed from 130 kHz to 30 kHz trigger rate
- Neural z-Vertex Trigger
  - Estimation of 3D-Track
  - 300 ns processing latency
  - Dead-time free
  - Multi layer perceptron



[21]

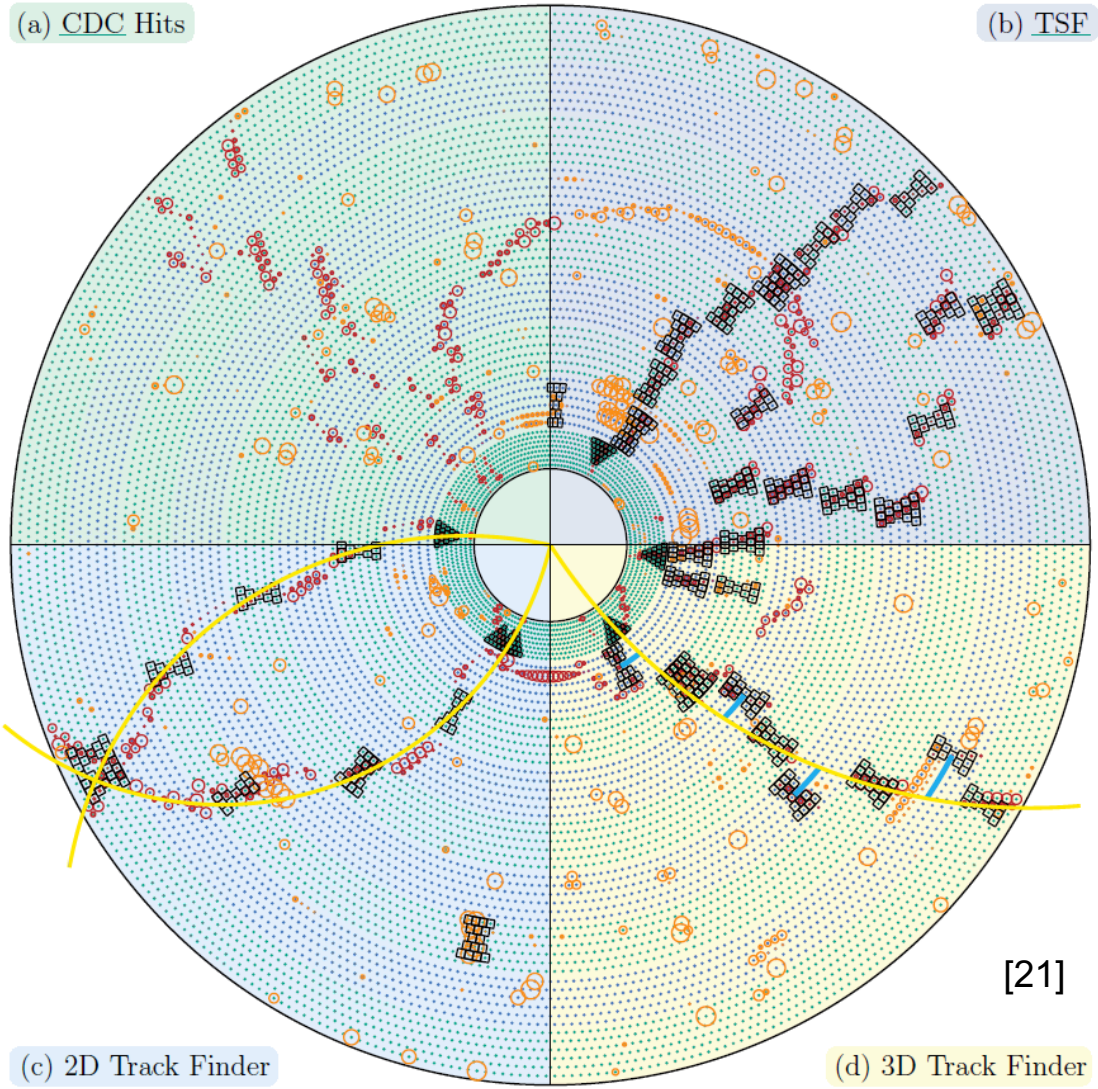


The neural z-Vertex Trigger



CDC Trigger Sub-System

- Drift Chamber of Belle II used for tracking
  - 14336 Wires
  - 56 Layers of Wires
  - 9 SuperLayers
- Alternating Orientation
  - Axial wires parallel to beamline
  - Stereo wires aligned with an angle to the z-Axis
  - Stereos essential for 3D tracking

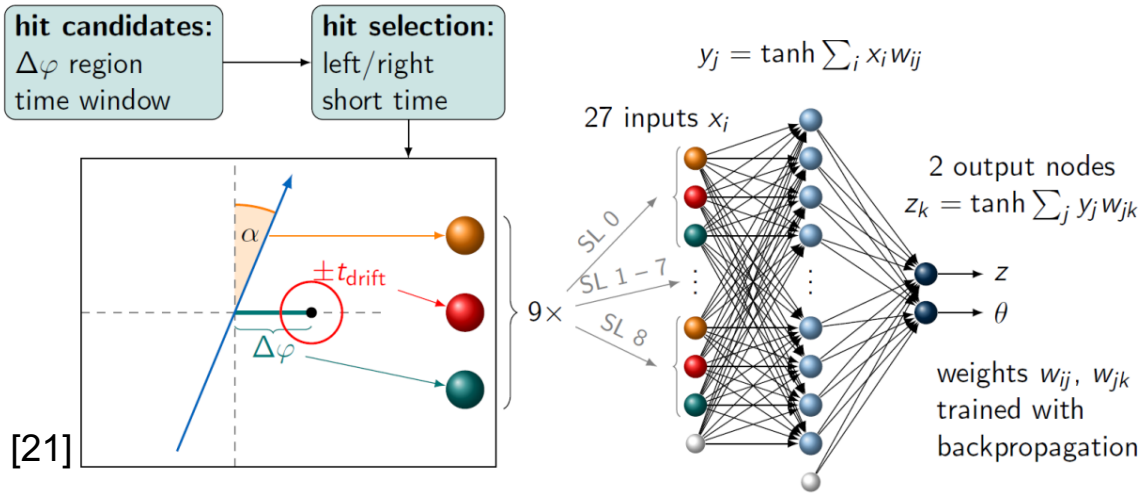


[21]

Neural Trigger Algorithm

- Combination of detector-specific preprocessing and selected neural network
- Each SuperLayer of the CDC (~5 layers of wires) three inputs are generated

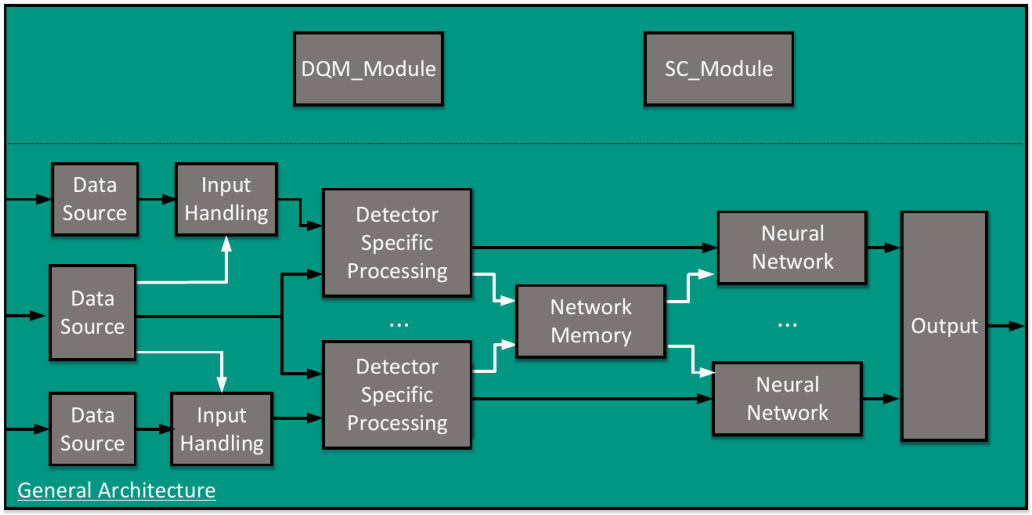
Name	Description
α	Crossing angle of the track relative to the normal of the crossing point of the track with the circular path of the layer.
$\pm t_{drift}$	Drift time of the TS. The sign indicates the direction of the passing particle either left or right. It is not used in case of an unknown direction as defined by the TSF.
φ_{rel}	Azimuth angle of the particle relative to the angle of the sense-wire.



[21]

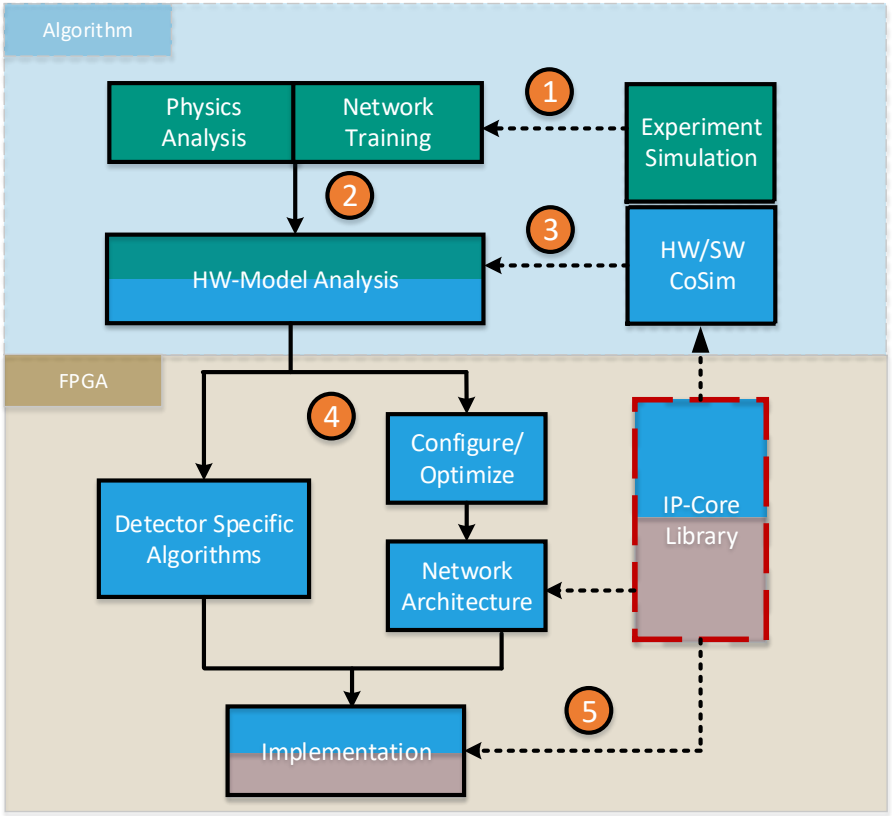
REALISATION

General Architecture Template



- Integration
  - Multiple parallel input sources
  - Distinct detector handling
  - Control flow for activation
- Processing
  - Detector Specific ; Preprocessing for Neural Network
  - Multiple Neural Networks ; Special case networks
  - Fixed point arithmetic
  - Minimum bit widths
  - Retiming enabling
- Monitoring
  - Data Quality Management
  - Slow Control

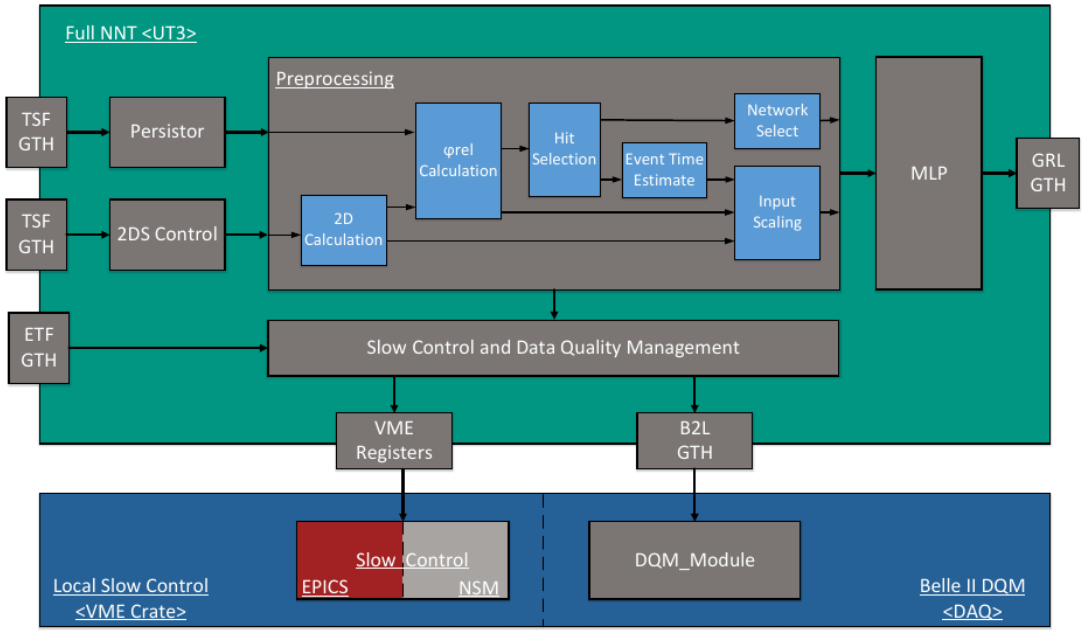
Design Flow for Architecture Inference



- Architecture Configuration Framework
  - Physics domain description of network
  - HW/SW Co-Simulation
  - Generation of all VHDL-Files
  - Semi-Automated configuration of architectural parameters
  - Belle II software framework for validation
  - IP-Core library for neural network

Architecture – Current NNT

- Processing with low latency MACs [Bä,17,18]
  - Multi Layer Perceptron
  - 2 Layers, 81 Neurons, 18 bit weights, 13 bit inputs, 5 weight sets
- Preprocessing
  - Network reduction from O(10^6) to 5
  - Parallel synchronized data paths

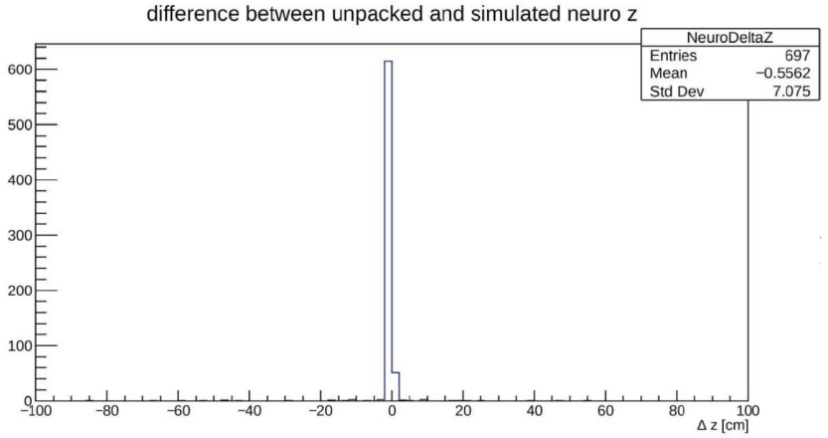


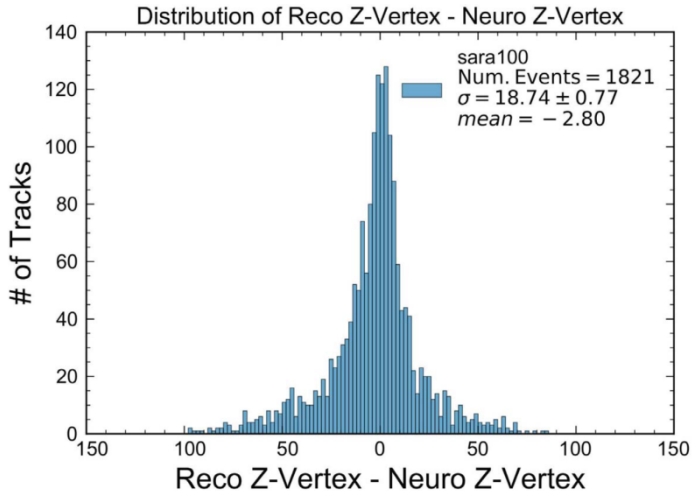
SLICE	Register	DSP	BRAM
46%	14%	53%	49%

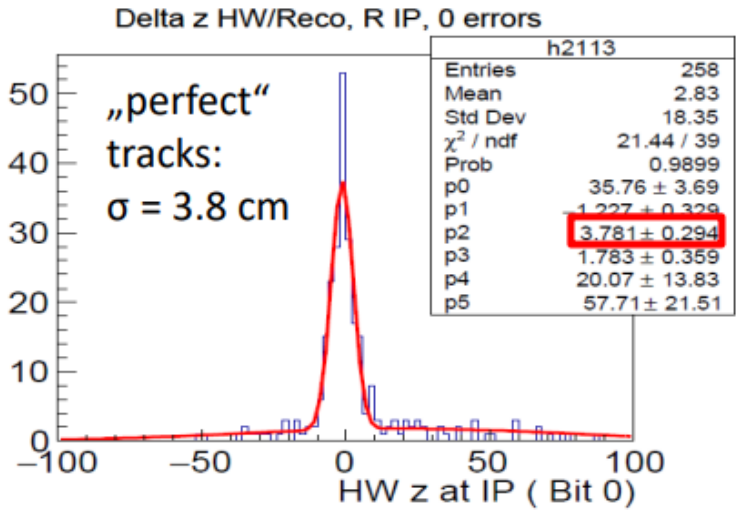
Latency	Latency	Frequency
288ns	32 clock cycles	127 MHz

Evaluation NNT – Recent Runs

- Collision results from december 2019
- Currently active with a z-Cut on 40 cm
- Some tracks with problematic drift times





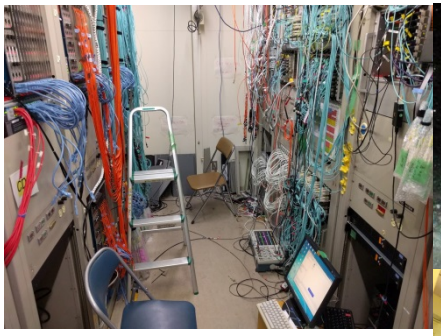


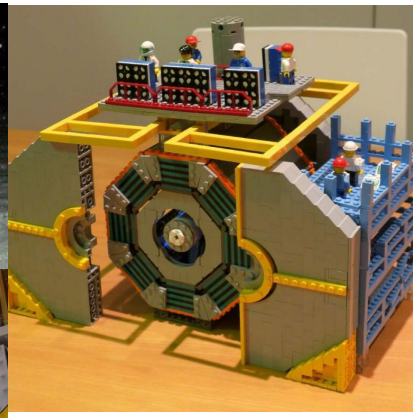
CONCLUSION

Conclusion

- Machine Learning on FPGAs
  - Highly flexible, deterministic, many IOs
  - Supported by tool vendors, open source efforts
  - Frameworks for ML Tools and C++ Integration
  - Multiple future solutions in development

- Neural z-Vertex Trigger
  - Suppression of tracks outside of the z-Vertex
  - FPGA-based implementation
  - Integrated into Belle II Experiment operational within all requirements
  - 40 cm resolution of 3D-Track Parameters





END SLIDE

References (I)

- [1] D. Silver „Mastering the game of Go with deep neural networks and tree search“ Nature 529, 484-489 (2016)
- [2] Google „Google AI“ https://ai.google/
- [3] Dean, J. „Recent Advances in Artificial Intelligence and the Implications for Computer System Design” Hot Chips 2017
- [4] Bradford, D. et al. „Knights Mill: New Inter Processor for Machine Learning“ Hot Chips 2018
- [5] Apple „Apple Special Event 2017“ , https://www.apple.com/de/apple-events/september-2017/
- [6] HiSilicion „Kirin 970 Processor“ http://www.hisilicon.com/en/Media-Center/News/Key-Information-About-the-Huawei-Kirin970
- [7] Amstutz, C. „Evaluation of an Associative Memory and FPGA-based System for the Track Trigger of the CMS-Detector” ; KIT Disseration 2016
- [8] Fiorini, M. et al. “The NA62 gigatracker: Detector properties and pixel readout architectures," Nucl. Instrum. Meth. A 624 (2010) 314.
- [9] Paul’s Hardware
- [10] Nvidia Inc. “GPU-BASED DEEP LEARNING INFERENCE”
- [11] Xilinx Inc. “Virtex UltraScale+ FPGAs Product Brief”

References (II)

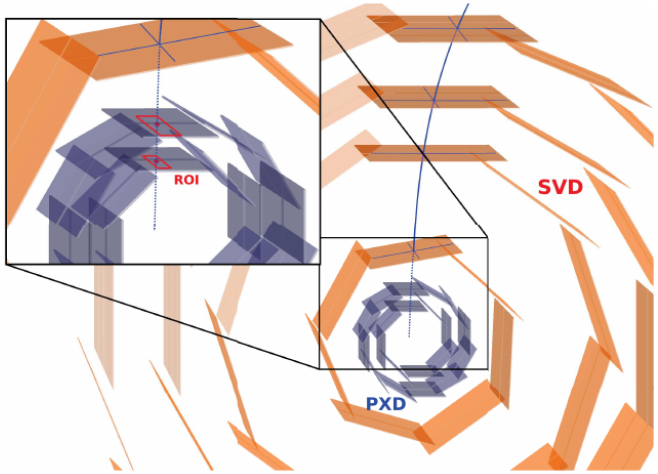
- [12] Xilinx Inc. “Versal: The First Adaptive Compute Acceleration Platform (ACAP) “
- [13] Human Brain Project ; https://www.humanbrainproject.eu
- [14] Y. Umuroglu „FINN: A Framework for fast scalable binarized Neural Network inference“**arXiv:1612.07119**
- [15] Xilinx Inc. “Deep Learning with INT8 Optimization on Xilinx Devices“ WP486
- [16] Xilinx Inc. Xilinx.com Product site
- [17] Xilinx Inc. “Artificial Intelligence Accelerates Dark Matter Search“
- [18] DUARTE, J. et al.: Fast inference of deep neural networks in FPGAs for particle physics. JINST, 13(07):P07027, 2018.
- [19] Intel “Machine Learning on Intel® FPGAs“ White Paper
- [20] ABE, T .: Belle II Technical Design Report 2010
- [21] Neuhaus S. “Track Reconstruction at the First Level Trigger of the Belle II Experiment” 2018
- [22] Schnell M. „Development of an FPGA-based Data Reduction System for the Belle II Depfet Pixel Detector“ – Dissertation Uni Bonn
- [23] Pulvermacher C. „dE/dx particle identification and pixel detector data reduction for the Belle II experiment“ – Master Thesis KIT

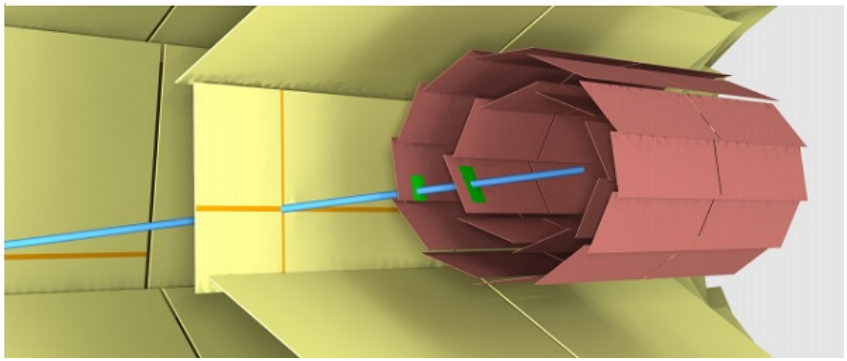
References (III)

- [24] Atomic Rules “WP: EXPERIENCE WITH THE AMAZON F1: A WHITE PAPER”
- [25] J. Fowler et al. “A Configurable Cloud-Scale DNN Processor for Real-Time AI”

Trigger and Data Reduction for Belle II

- Particle identification
  - Current data reduction in Belle II is suppressing interesting particles
  - Parallel approach to rescue data from interesting particles necessary
- Online Cluster Analysis
  - NeuroBayes Algorithm for particle identification
  - 200 million analyses per second
  - 90% noise reduction
  - Maximum signal efficiency

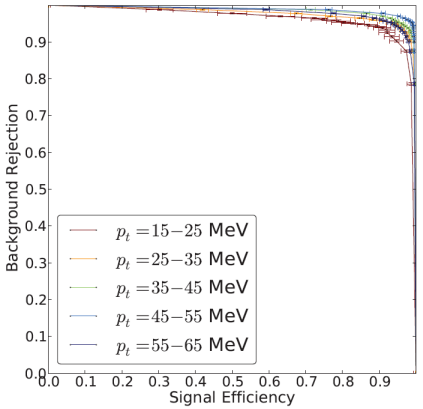


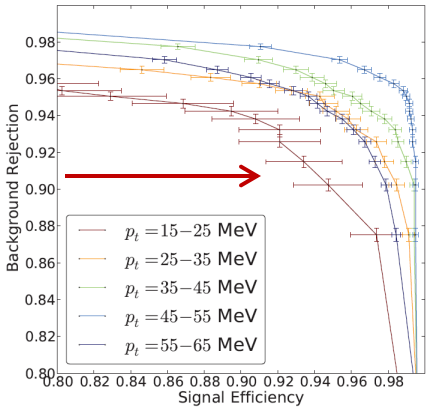


[22]

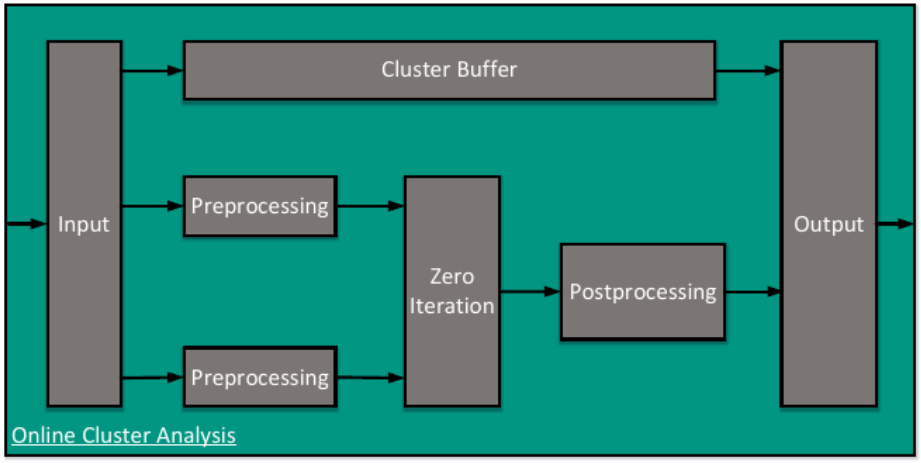
Online Cluster Analysis

- Processing with High-Throughput MAC [Bä,15]
  - NeuroBayes Algorithm
  - Dedicated Preprocessing (CDF)
  - Fix Point operation 0.00001 accurate
- All architectural requirements fulfilled
- Signal Efficiency : 95 %
- Noise Reduction : 90 %





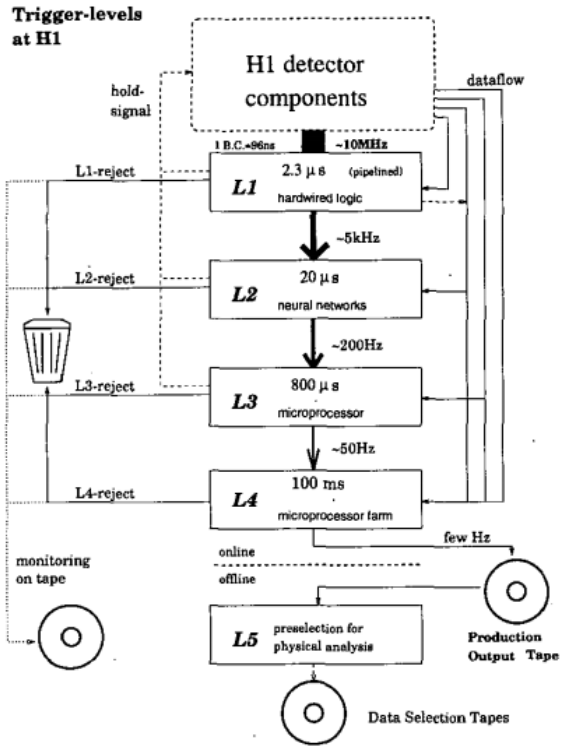
[23]

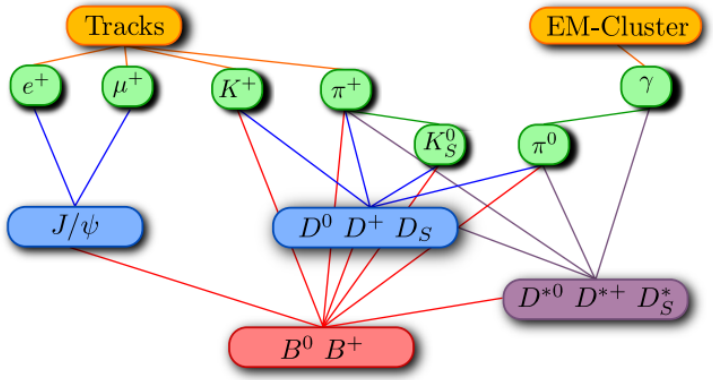


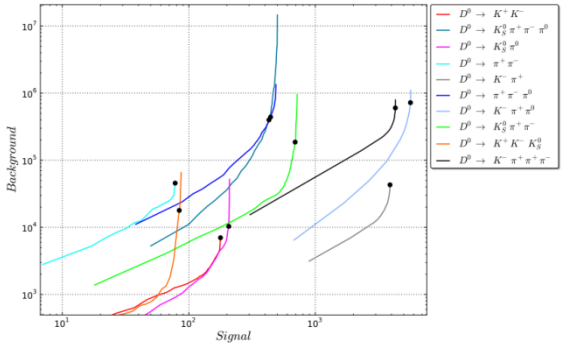
SLICE	Register	DSP
2 % / 52 %	3% / 57 %	3 %

Latency	Latency	Frequency
39ns	13 clock cycles	350 MHz

ML Application in high energy physics



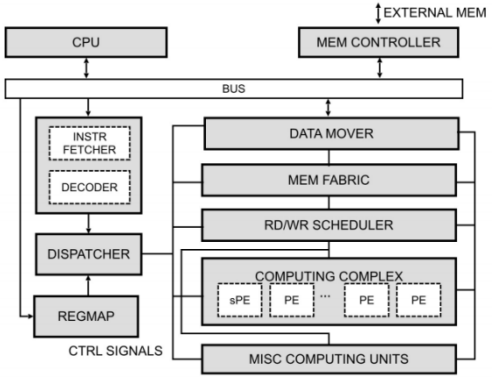




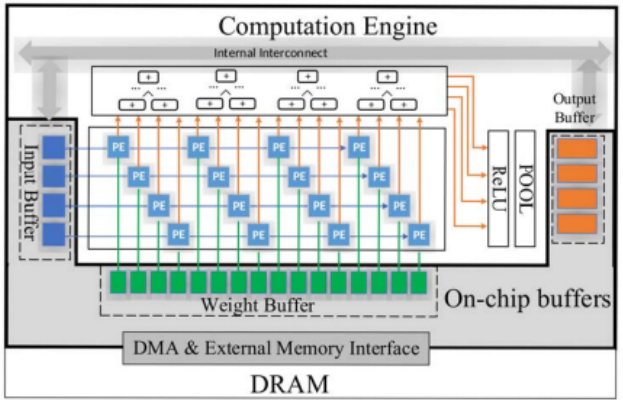
- L2 Neural Network Trigger for the Hera Experiment [5,2003]
  - Custom Processor CNAPS
  - 20 μs Latency budget
- Neural Network based Full Reconstruction for B Mesons [6,2010]
  - Usage of the specialized NeuroBayes Algorithm
  - Offline Analysis

State-of-the-Art – Neural Processing on FPGA

- DeePhi DPU [15]



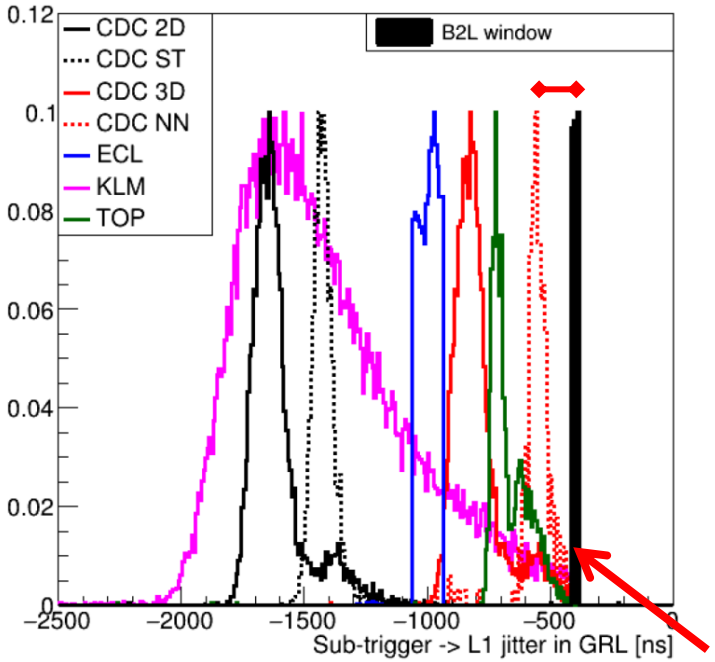
- Caffeine [16]

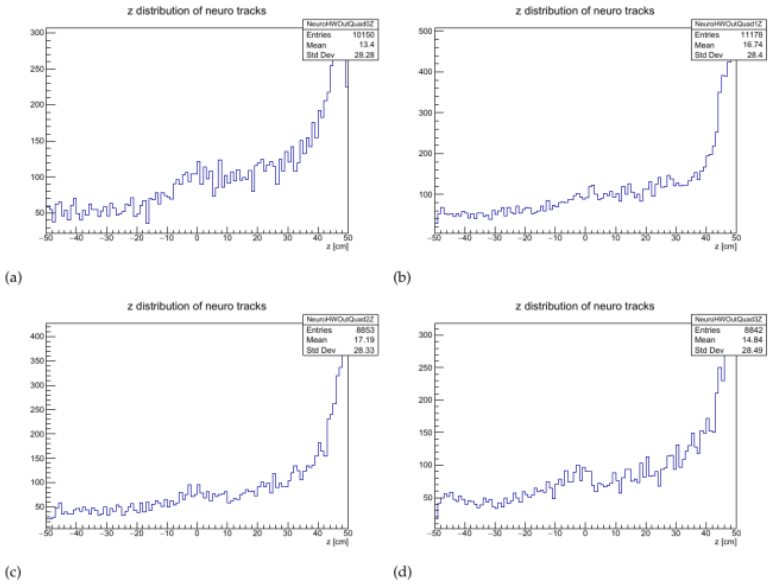


- General solutions with programmable processing cores
  - Off-Chip Memory usage
  - Instruction sets
  - High-Performance compression and optimization
- Dedicated architecture required for physics application
  - Only on-Chip memory due to latency
  - High precision processing required
  - Interchangeable networks without external memory access

Evaluation NNT – Basic Functionality

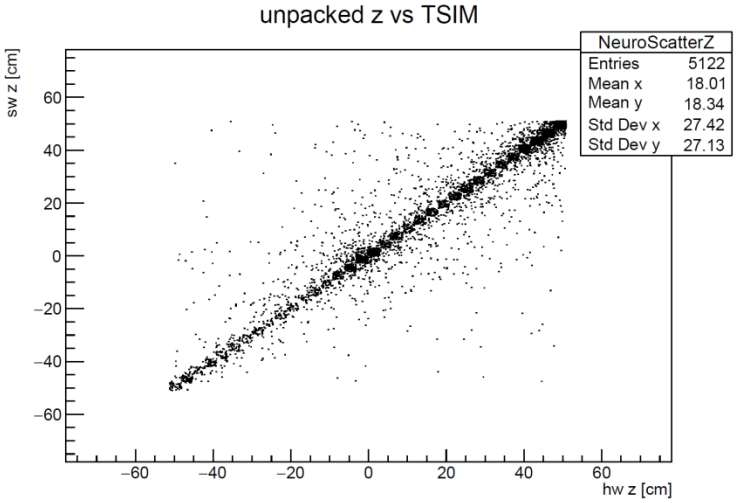
- Latency
  - Close to deadline
  - Still within limits
- Full Coverage
  - Hardware estimations for all quadrants

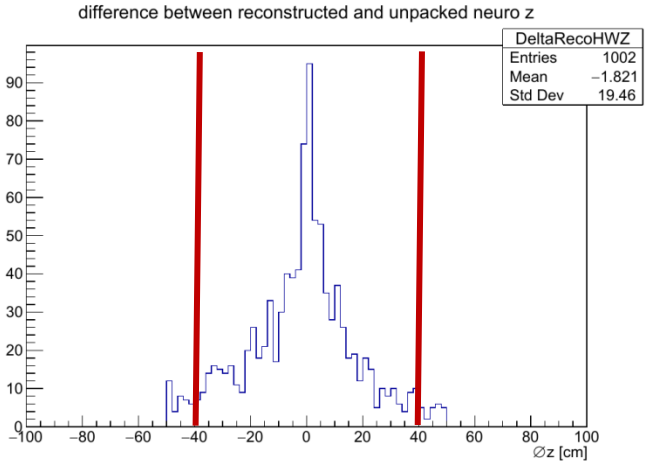


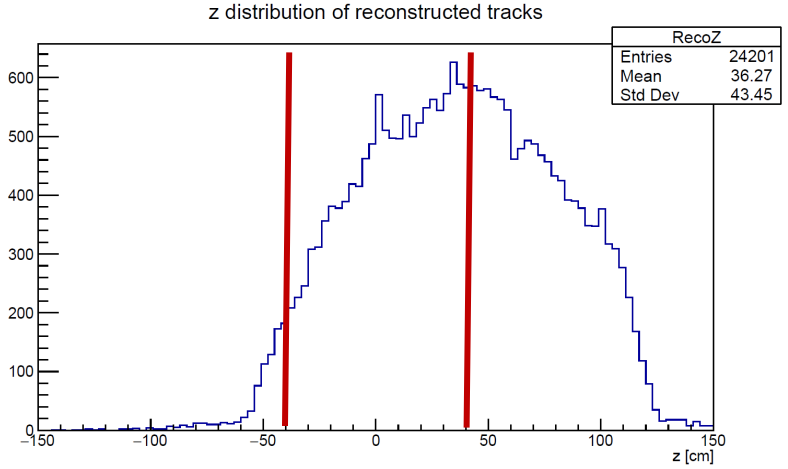


Evaluation NNT – Estimation Functional

- Correlation plot show good match between hardware and software
- Suppression can be applied with 40cm cut, currently sufficient
- Network not trained on real data
  - Worse performance was expected

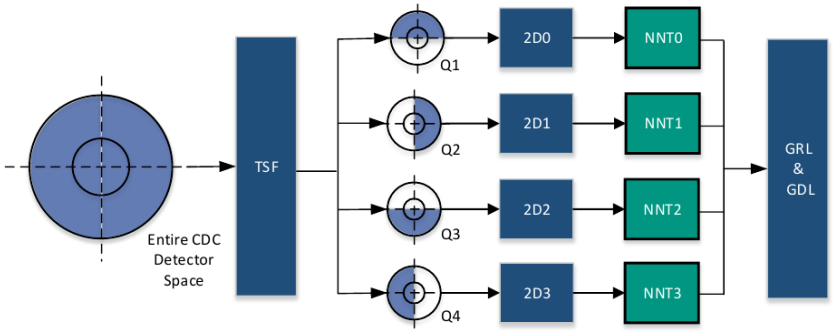




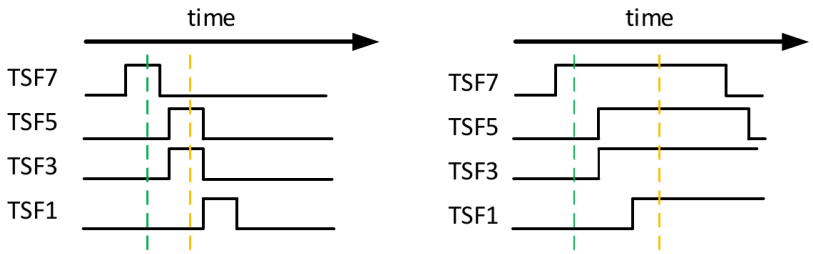


Integration – Application Case NNT

- Interfacing
  - 60 Gigabit Transceivers Lanes required
  - Distributed across 4 boards



- Detector Handling
  - Drift chamber with specific behavior
  - Data arrives over time
  - Generation of Pool of Sensor data within time frame



- Platform Universal Trigger Board 3
  - Virtex-6 380HXT
  - Maximum amount of Gigabit Transceivers