HELMHOLTZ KIT JÜLICH Forschungszentrum

RESEARCH FOR GRAND CHALLENGES

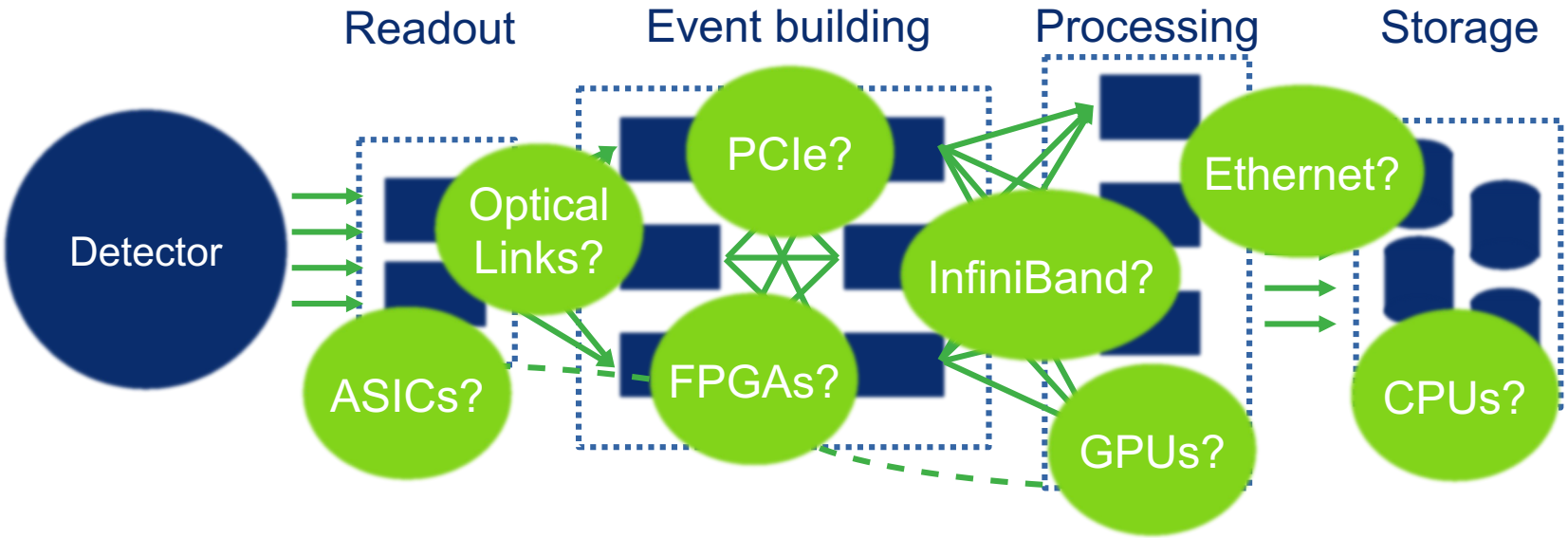
T. Dritschler, M. Caselle, L. Ardila, S. Chilingaryan,
A. Ebersoldt, J. Hurst, N. Karcher, A. Kopmann, O. Sander, T. Stockmann(*)

(Commercial) DMA technologies to realize a flexible DAQ system for Tera-Scale experiments

Karlsruhe Institute of Technology (KIT)
(*)Forschungszentrum Jülich

Challenges for modern DAQ Systems

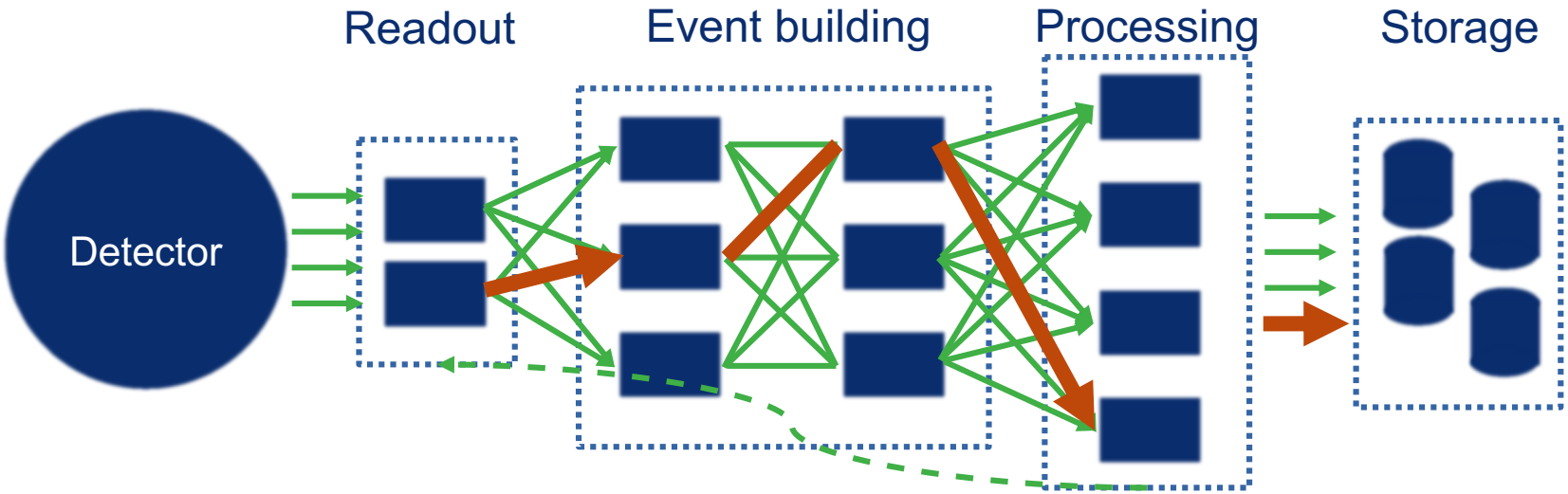
Complex data paths and heterogeneous architectures



- Advancements in Detector Technologies lead to highly complex DAQ systems
- This complexity results in heterogeneous systems with many different components

Challenges for modern DAQ Systems

Complex data paths and heterogeneous architectures

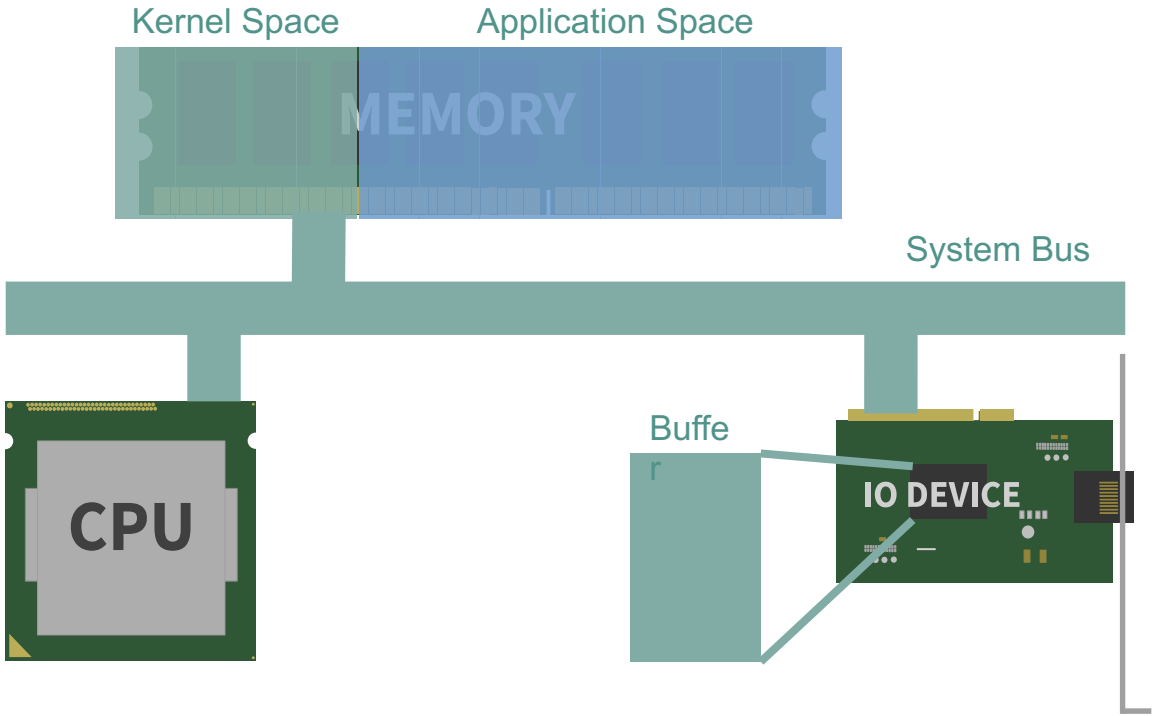


- Development and maintenance of complex DAQ systems is expensive and time consuming
- Using commercially available standards and „off-the-shelves“ components might mitigate these efforts

- Can we design scalable and maintainable DAQ using mostly commercial components?

Data transfer inside of conventional computing systems

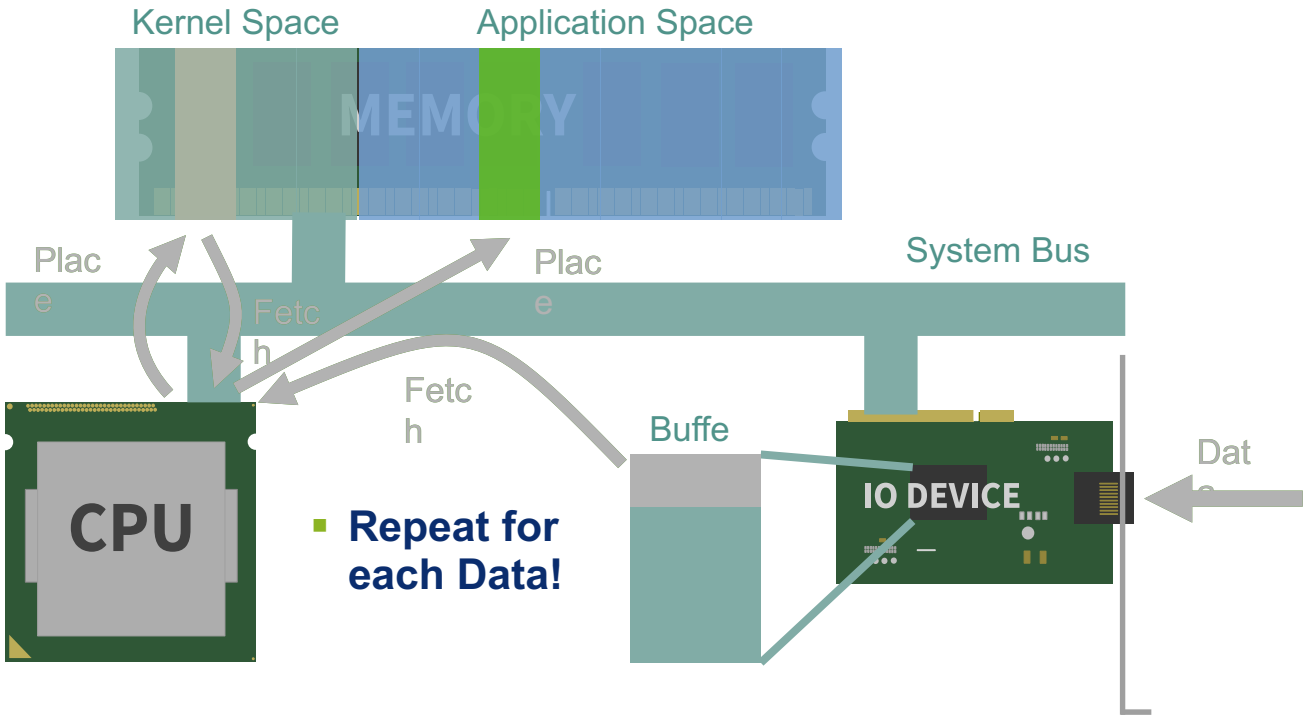
CPU involvement in data movement



Data transfer inside of conventional computing systems

CPU involvement in data movement

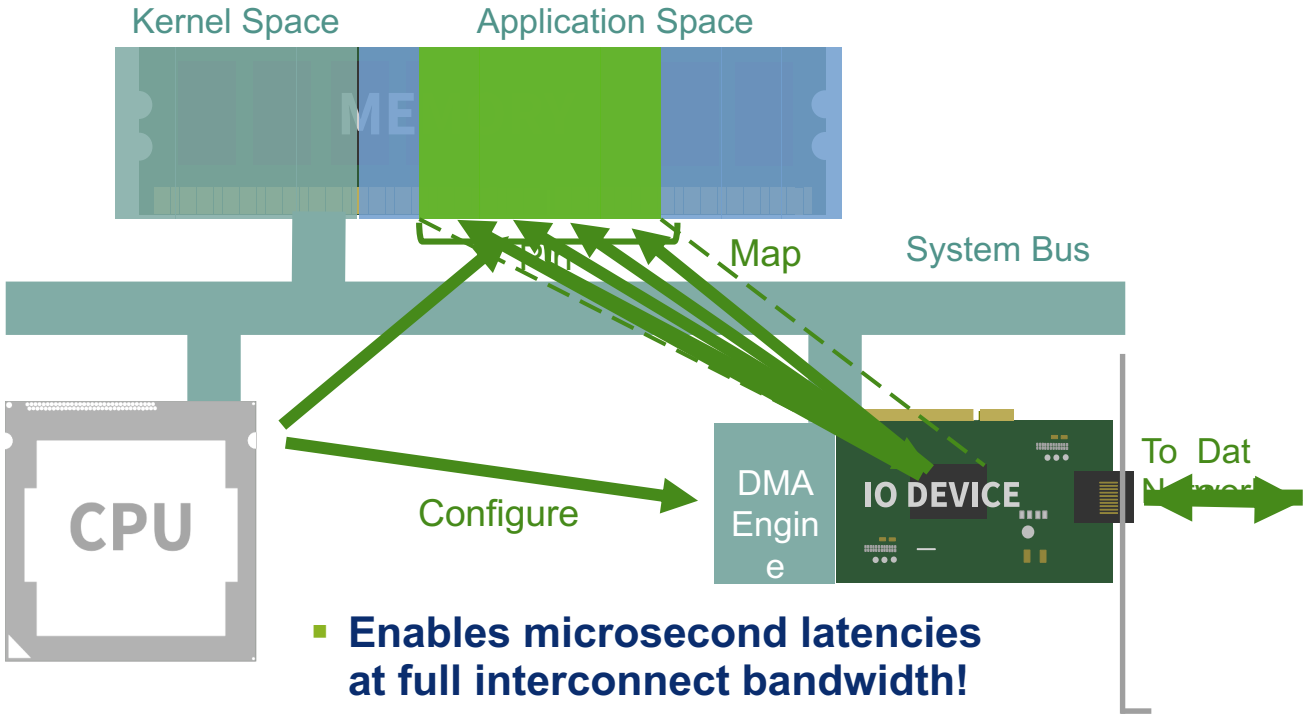
- Data arrives and is placed in device buffer
- CPU Fetches data from buffer
- CPU Places data into Driver memory
- CPU Fetches data from Driver
- CPU Places data into application space



Direct Memory Access (DMA) enabled data transfer

Freeing up CPU and reducing copy efforts

- CPU „reserves“ memory in application space (Pinning)
- CPU configures the device’s DMA engine, effectively „Mapping“ the pinned memory to the device
- Data arrives and is placed right into destination memory

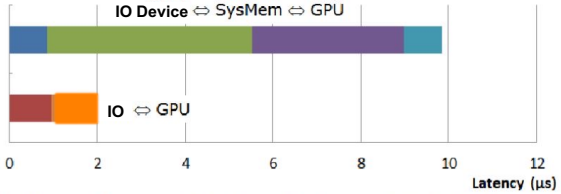


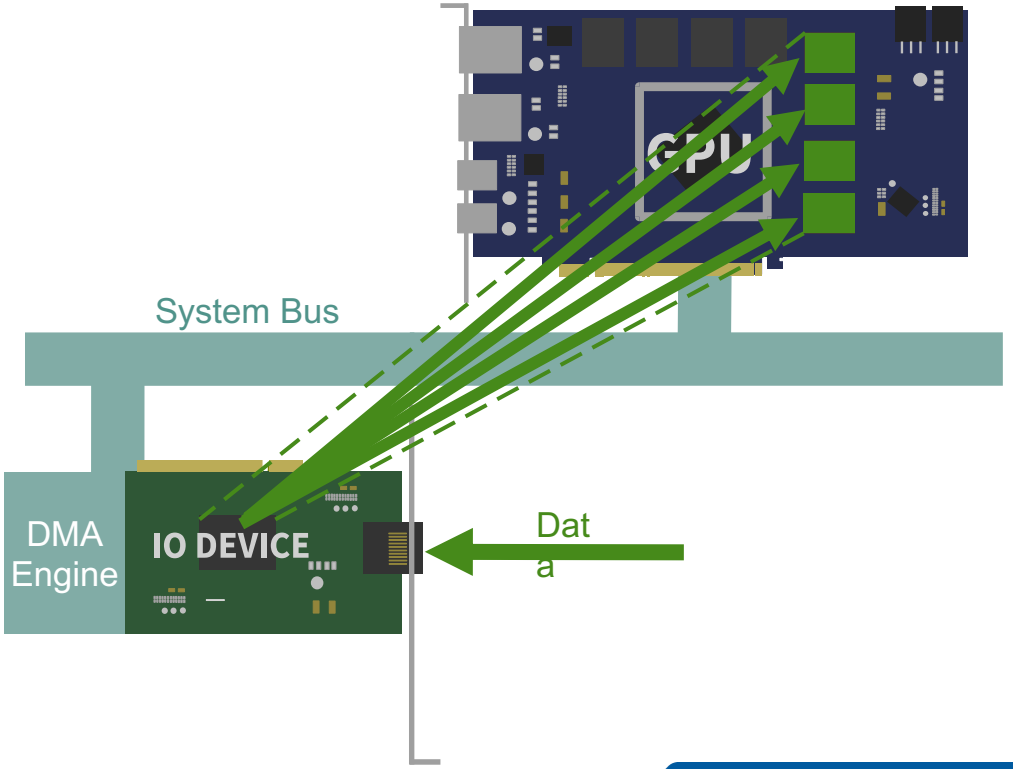
- **Enables microsecond latencies at full interconnect bandwidth!**

RDMA enabled GPGPU computing

Efficient use of GPUs for On-Line and In-Line data processing

- DMA is possible for GPU memory as well
- Analogous to main memory DMA
- Pin → Map → Place
- Significantly reduces latency!





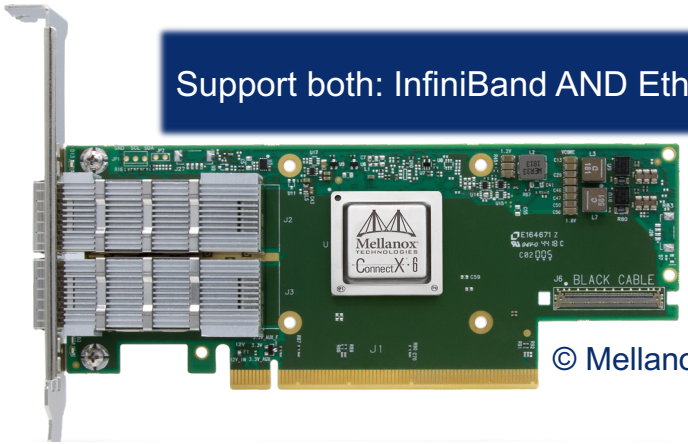
InfiniBand, the ‚de facto' standard for RDMA

Commercial High-Speed, Low-Latency data transfer

ConnectX-6

General Specs	
Ports	Single, Dual
Port Speed (Gb/s)	10, 25, 40, 50, 100, 200
PCIe	2x Gen3 x16; Gen4 x16
Connectors	QSFP56; SFP-DD
Message Rate (DPDK) (million msgs/sec)	215
RoCE Latency at Max Speed	0.78

Support both: InfiniBand AND Ethernet!



© Mellanox

InfiniBand

From Wikipedia, the free encyclopedia

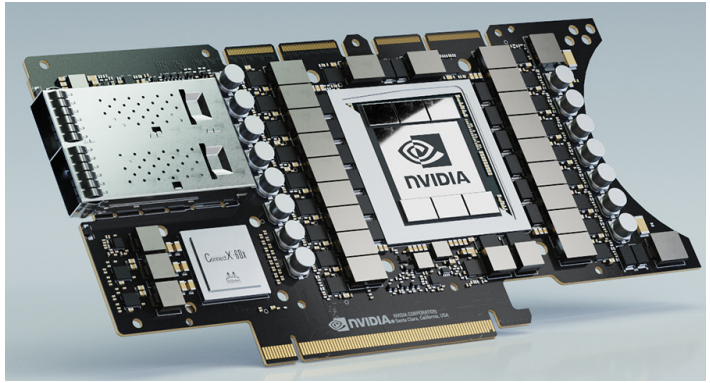
InfiniBand (**IB**) is a computer networking communications standard used in high-performance computing that features very high throughput and very low latency. It is used for data interconnect both among and within computers. InfiniBand is also used as either a direct or switched interconnect between servers and storage systems, as well as an interconnect between storage systems. It is designed to be scalable and uses a switched fabric network topology.

As of 2014, it was the most commonly used interconnect in supercomputers. Mellanox manufactures InfiniBand

Next generation RDMA capable computing accelerators

Closing the gap between networking and computing

- Commercially available components
- Software Programmable
- Low development effort
- Easy to upgrade to newer generations, once they become available
- (R)DMA Capable
- Optimized for Machine Learning and AI applications!



© Nvidia

Nvidia EGX A100:
Ampere GPGPU,
200Gbps Network,
Ethernet + InfiniBand

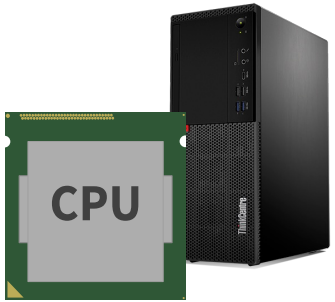
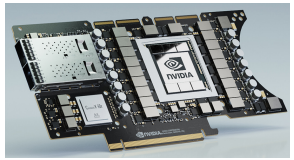
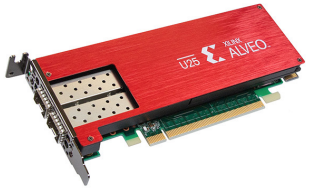
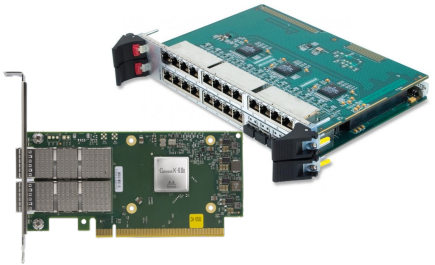
© Xilinx

Xilinx ALVEO SmartNIC:
Up to 1.1M LUTs
Up to 100Gbps Ethernet

Communication between heterogeneous components

One protocol for all the most common components?

- Ethernet is one of the most common commercially available interconnects

PC / CPU	GPU	FPGA	Network Infrastructure
			
Ethernet ✓	Ethernet ✓	Ethernet ✓	Ethernet ✓

- Is there a way to benefit from RDMA using Ethernet?

RoCE (RDMA over Converged Ethernet)

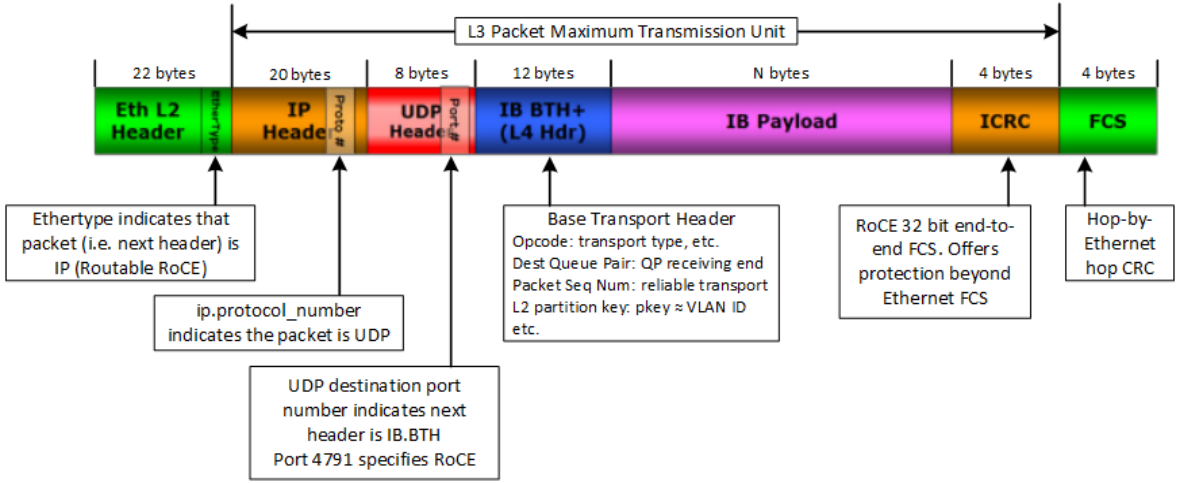
Enabling RDMA benefits for conventional Ethernet networks

RDMA over Converged Ethernet

From Wikipedia, the free encyclopedia

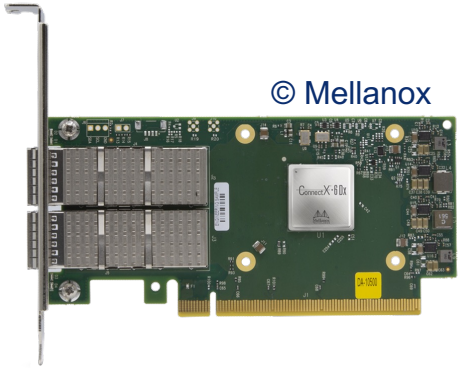
RDMA over Converged Ethernet (**RoCE**) is a network protocol that allows remote direct memory access (RDMA) over an Ethernet network. It does this by encapsulating an IB transport packet over Ethernet.

- RoCE uses UDP packets and encapsulates InfiniBand “instructions” into the UDP payload
- Soft- and Hardware implementations available!



KIRO: KIT RDMA Programming Library

Integrating RMDA capabilities into software



GitHub

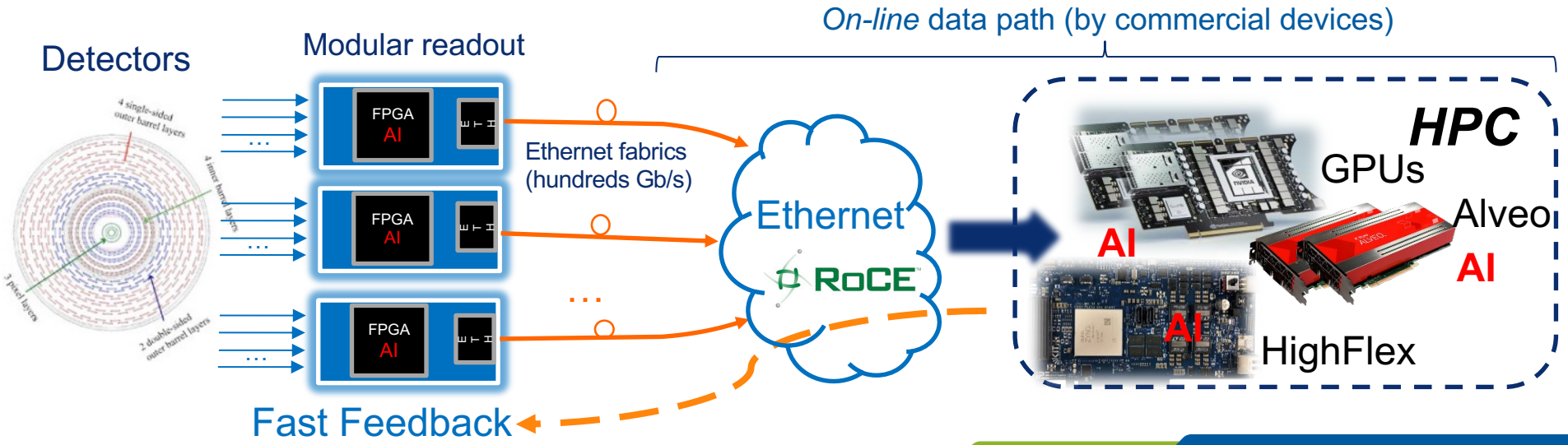
Clone it on Github!
https://github.com/ufo-kit/kiro

- **Works for both: InfiniBand and RoCE!**
  - **KIRO Server/Client:**
  - Unidirectional data transfer from Server to Client (Clients “Pull” from Server)
  - Fixed-Size memory segment
  - Supports multiple connected clients per server
  - **KIRO Messenger (Beta):**
  - Layers bi-directional point-to-point messaging on top of KIRO
  - Messengers can connect to multiple peers
  - Fully RDMA enabled arbitrary sized memory exchange (“Push” and “Pull”)

Novel DAQ architecture using commercial HPC Components

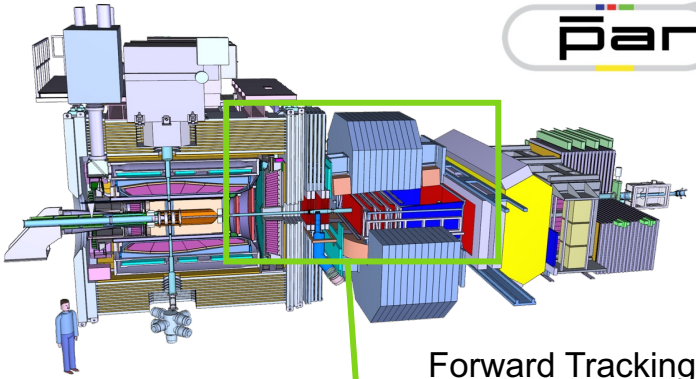
High-performance distributed ML for physics experiments

- Versatile DAQ optimized for detector and AI applications
- High-performance ML inference on modern programmable hardware platforms
- Novel heterogeneous FPGA/GPU architecture based on emerging Ethernet protocols



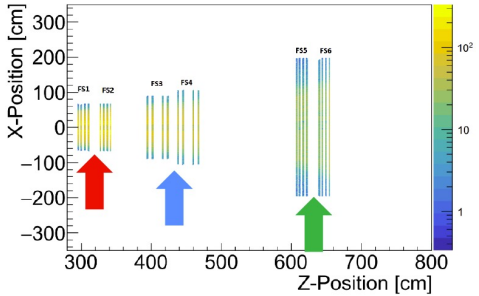
Machine Learning Application Example

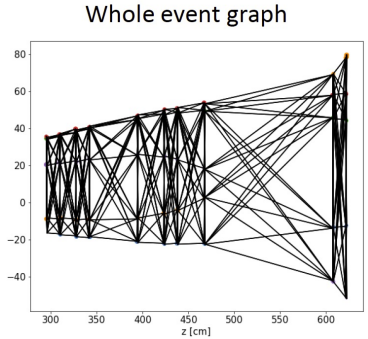
PANDA Tracking and Event Selector

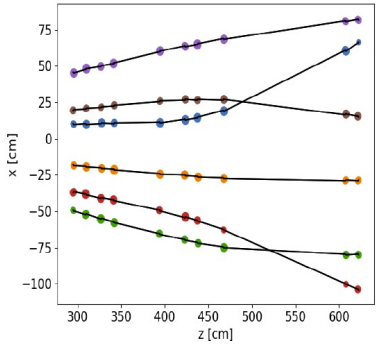


Forward Tracking Stations

- 1 & 2, 3 & 4, 5
- 3&4 inside 2 Tm dipole field







Graph Neural Network

- Connection of all hits from one layer with all hits in next layer

Most probable connections survive

- 6 charged tracks

In Total:

- Track finding efficiency 96%

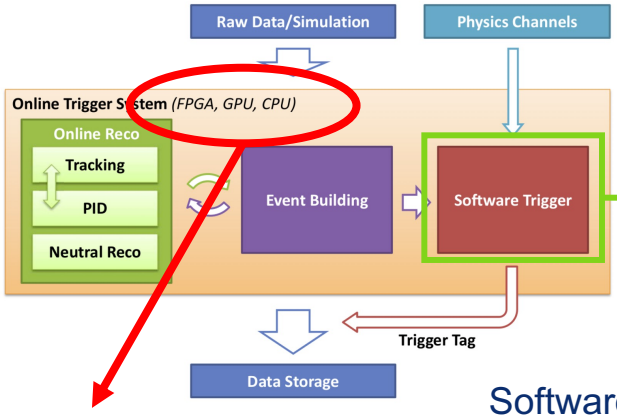
Waleed Esmail

Machine Learning Application Example

PANDA Tracking and Event Selector

JÜLICH Forschungszentrum

PANDA online event selector



- Heterogeneous system
- Unifiable communication

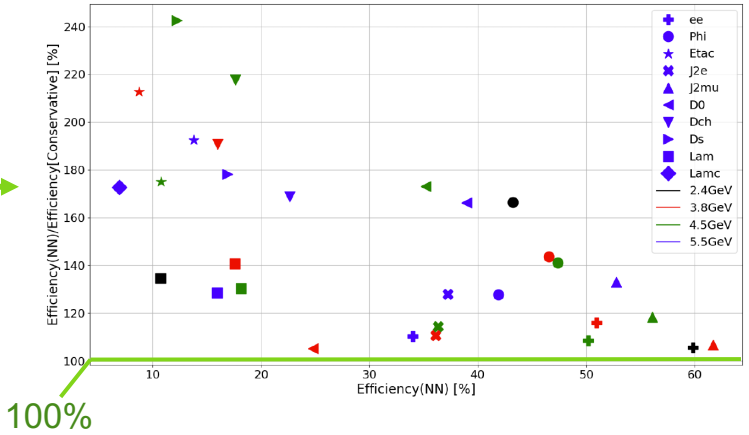
RoCE

- Proposed for PANDA DAQ

Best network: CNN

Software trigger selects physical interesting channels from background for data storage

Efficiency improvement for different physics channels
CNN vs. Classical method



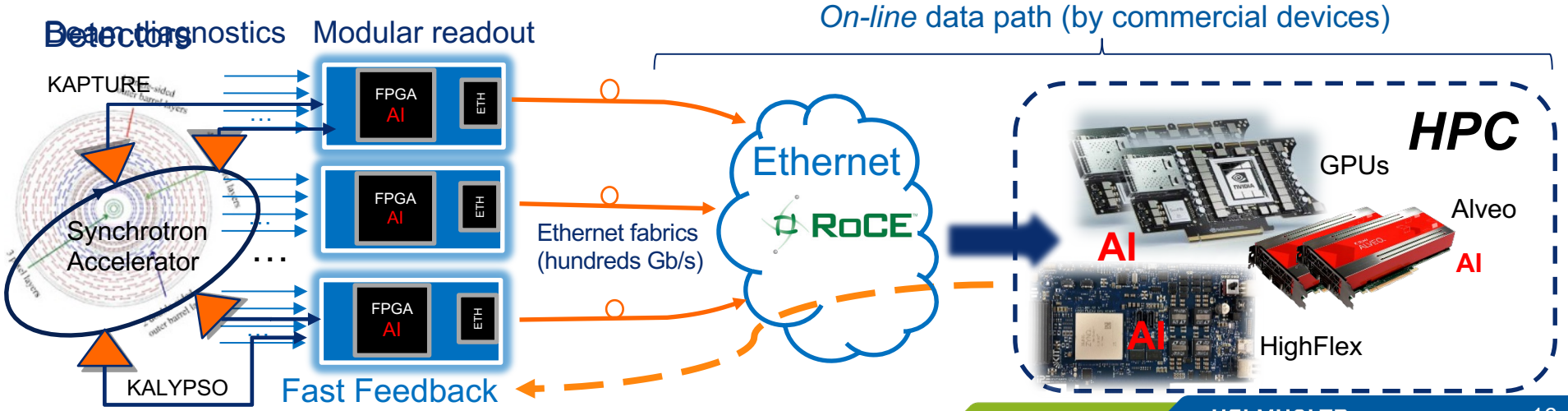
All data points above 100% efficiency vs. conventional approach → Indicates Neural Network performs better accross all metrics

Peiyong Jiang

Heterogeneous FPGA – GPU architecture for beam physics

High-performance distributed ML for physics experiments

- Versatile DAQ optimized for detector and AI applications
- High-performance ML inference on modern programmable hardware platforms
- Novel heterogeneous FPGA/GPU architecture based on emerging Ethernet protocols

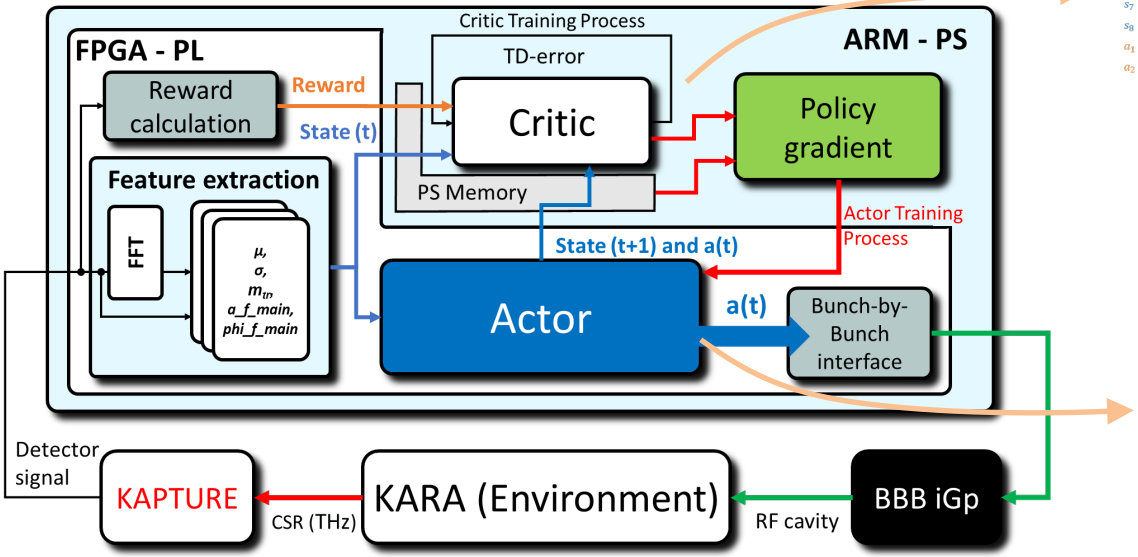


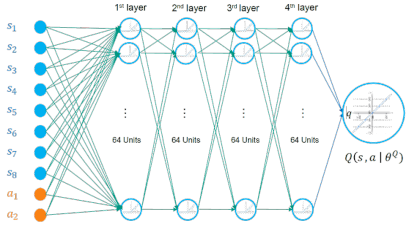
Hardware implementation

As seen in: **„Accelerated Deep Reinforcement Learning for Fast Feedback of Beam Dynamics at KARA“** **A. Ebersoldt**

Reinforcement Learning on modern programable device

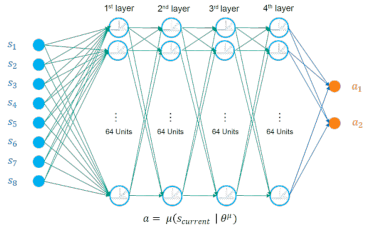
Deep Deterministic Policy Gradient (DDPG)





Critic

Critic neural network and the policy gradient are implemented as bare-metal application on ARM processor



Actor

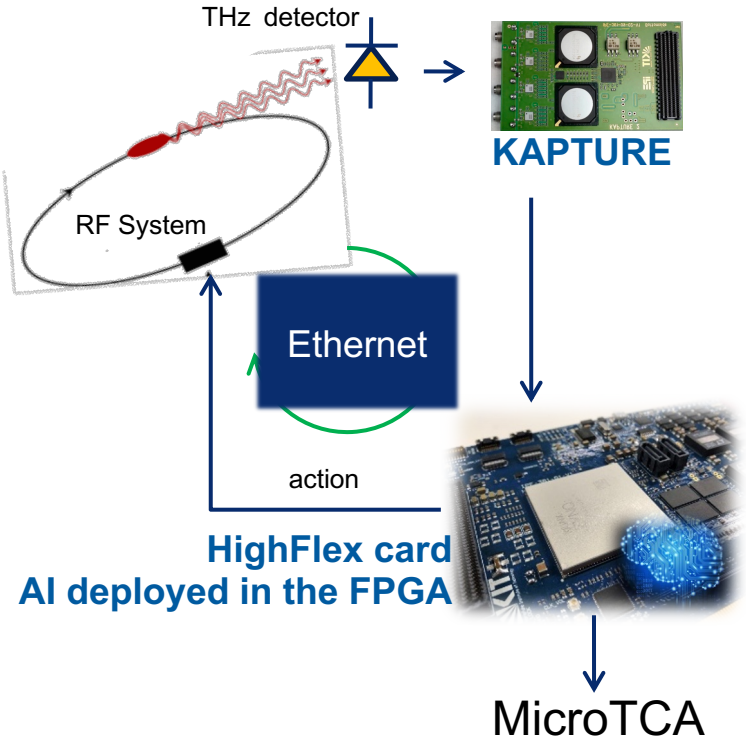
Actor consists of four fully-connected dense layers, each layer consists of 64 neurons using a rectified linear unit (ReLU) as activation function in **FPGA**

Control of the complex beam with ML

Machine Learning toward Autonomous Accelerators

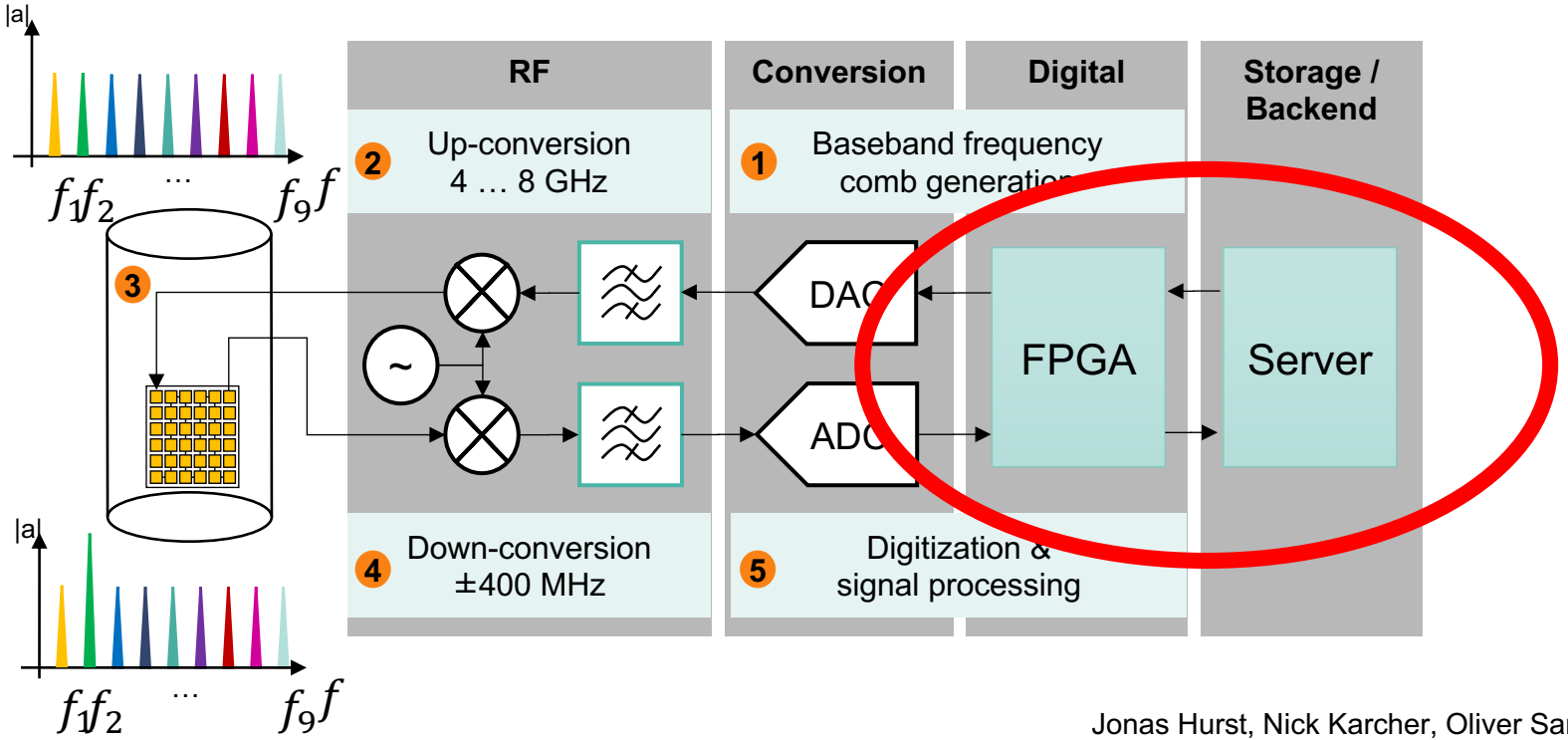
A. Ebersoldt

- Closed feedback loop at KARA:
  - **Detection** of signals with THz detectors and KAPTURE @ 500 MPulse/s
  - **Data processed** by Reinforcement Learning on FPGA
  - **FPGA action** as special RF signal modulation is sent to the kicker cavity
- **Goal**: total latency of control feedback loop ***<< 1 ms***
- **Target applications**: KARA, FLUTE, ARES and more
- **Status**: first beam control on FPGA developed within *AMALEA* → will continue in ACCLAIM (Helmholtz Innovation fund)



Application: RoCE in FPGA for the ECHo experiment

ECHo Experiment to measure the Neutrino mass

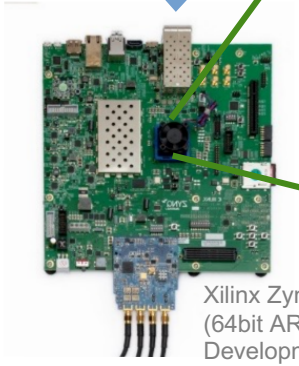


Jonas Hurst, Nick Karcher, Oliver Sander

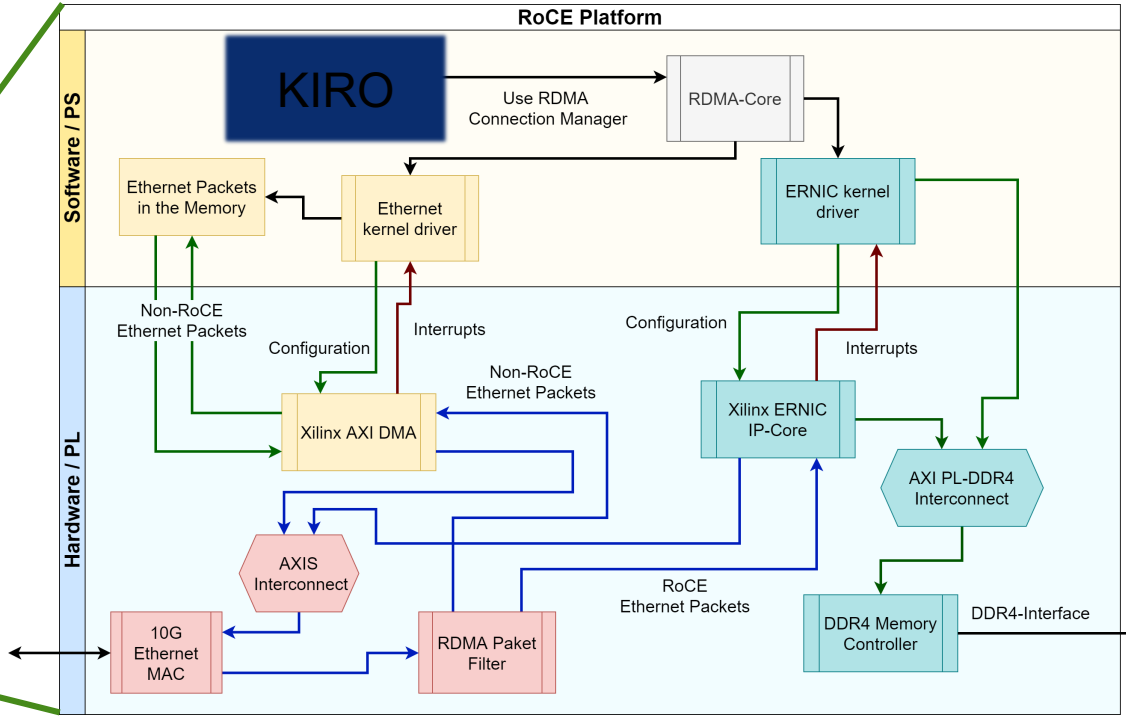
Application: RoCE in FPGA for the ECHo experiment

Xilinx ERNIC IP Core implementation and integration

RoCE



Xilinx Zynq US+
(64bit ARM + FPGA)
Development Board



Master Thesis: „**Entwicklung eines breitbandigen Datentransfers (DMA) für die Messdatenakquise**“ (Jonas Hurst, Nick Karcher, Oliver Sander)

Conclusion

Application of industrial RDMA standards for modern DAQ

Drawbacks:

- Integration into existing infrastructure requires (complex) retooling
- Difficult to deploy and operate from within virtualized software environments
- Full performance can only be reached when using dedicated networking hardware

Benefits:

- Computing Accelerator integration → Full support for HPC and on-line processing
- Readily available Hard- and Software → Drastically reduces development efforts
- Single, commercial components → Easy maintenance and easy upgrades to newer versions
- Highly versatile → Unify communication between all components: CPU, GPU, FPGA