Karlsruhe Institute of Technology

Institute for Data Processing and Electronics Karlsruhe Institute of Technology

Cloud platform for high data rate detector instrumentation

Research Center

Karlsruhe University

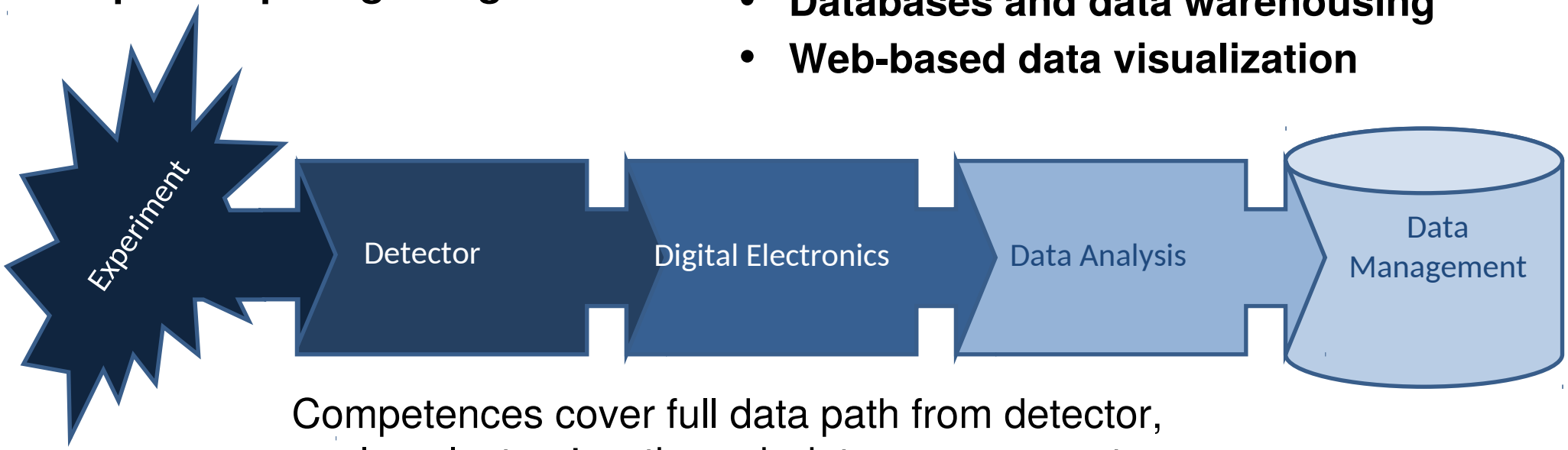
IPE Competences

Experiments

- Astroparticle & High Energy Physics
- Atmosphere and Climate
- Nuclear Fusion
- Electrical Storage Systems
- Photon Science
- Ultrasound Tomography
- Nano- and Microsystems
- Supercomputing & Big Data

Tools and Technologies

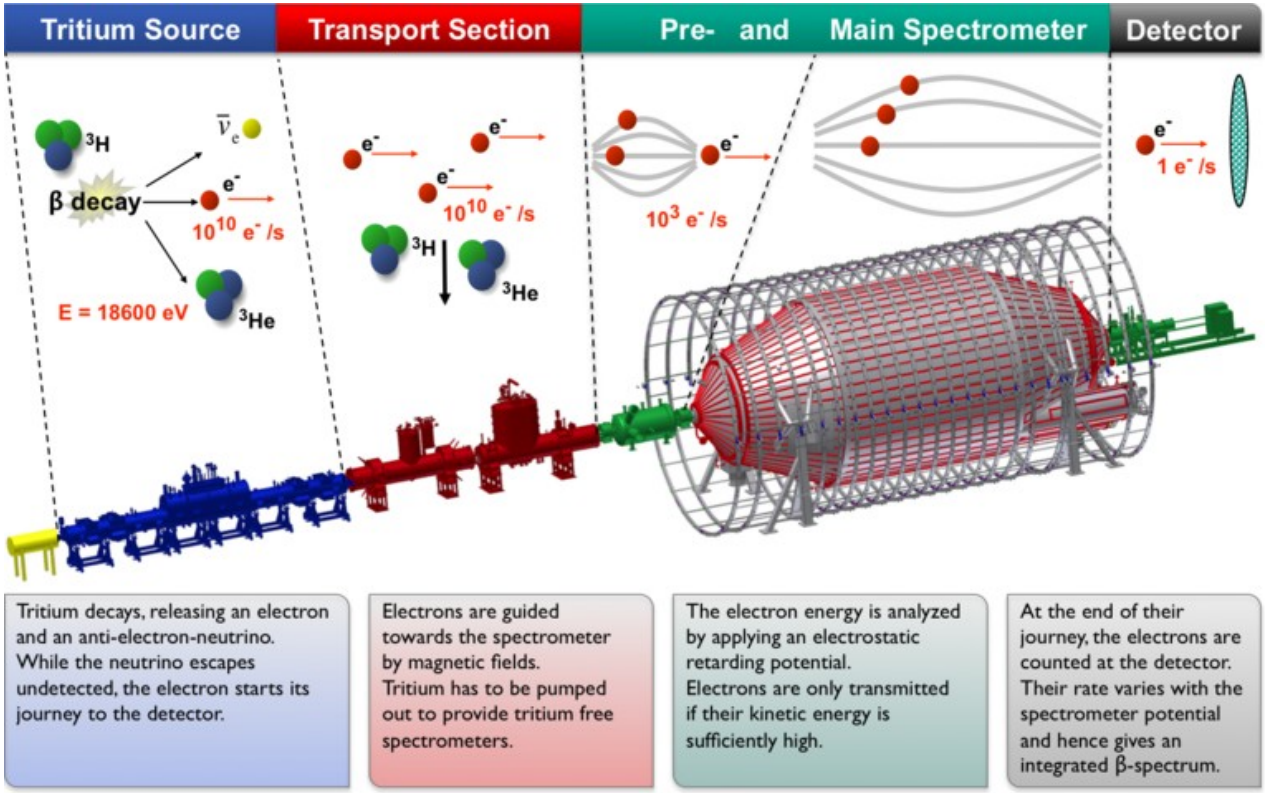
- High-speed DAQ Electronics
- High-performance and GPU computing
- Software optimization
- Databases and data warehousing
- Web-based data visualization

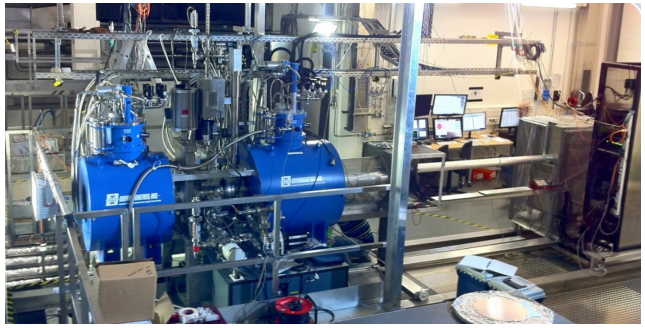


Competences cover full data path from detector, analog electronics, through data management

Prof. Dr. Marc Weber
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

KATRIN



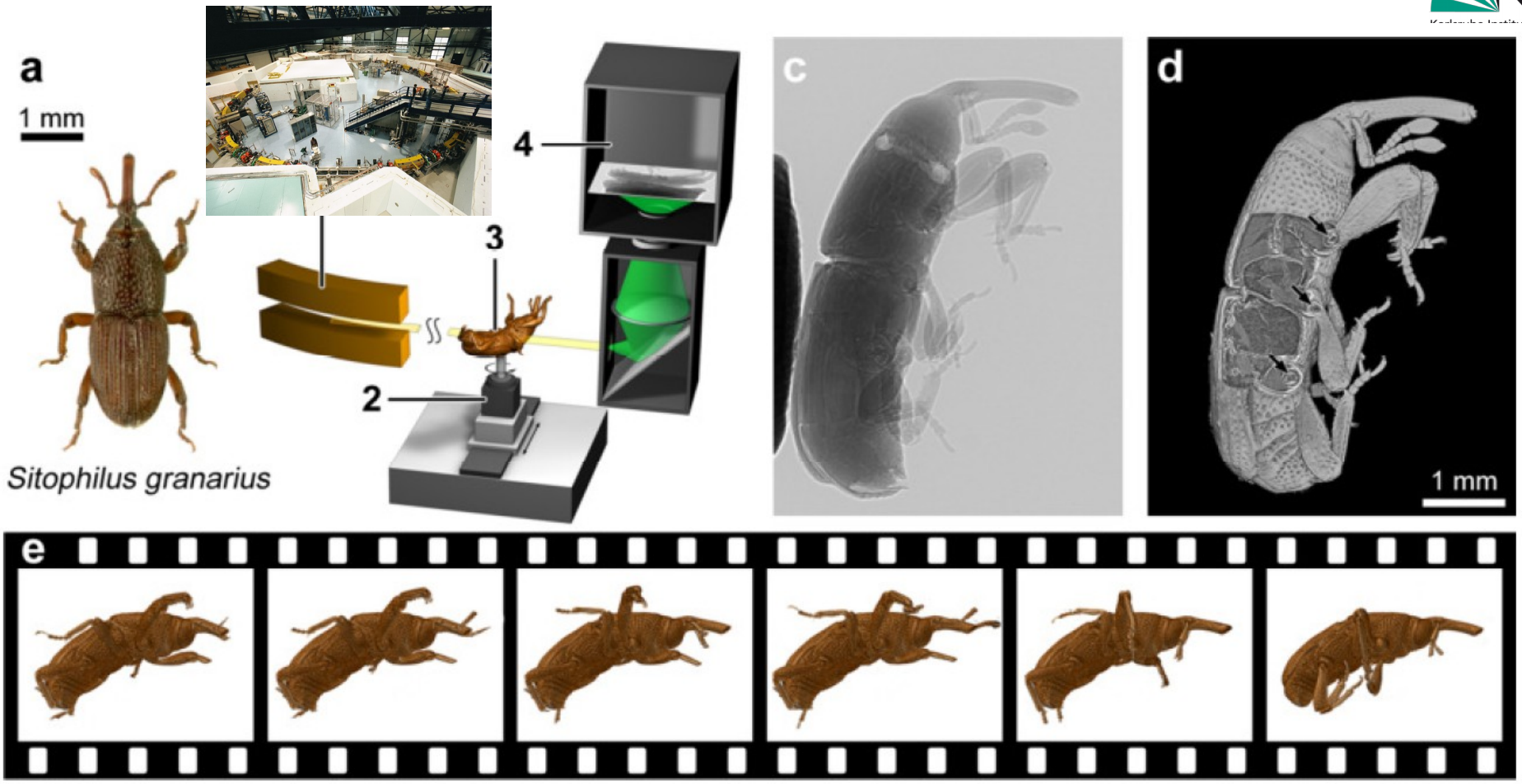


KATRIN Detector

KATRIN experiment is aimed to measure the mass of electron neutrino with sub-eV precision

KATRIN Main Spectrometer

Synchrotron 4D cine-tomography



In vivo X-ray 4D cine-tomography experiment. (a) Photograph of Sitophilus granarius, dorsal view. (b) Experimental set-up for ultra-fast X-ray microtomography showing bending magnet (1), rotation stage (2), fixed specimen (3) and detector system (4). (c) Radiographic projection. (d) 3D rendering of the reconstructed volume with thorax cut open and revealing hip joints (arrows). (e) In vivo cine-tomographic sequence of moving weevil.

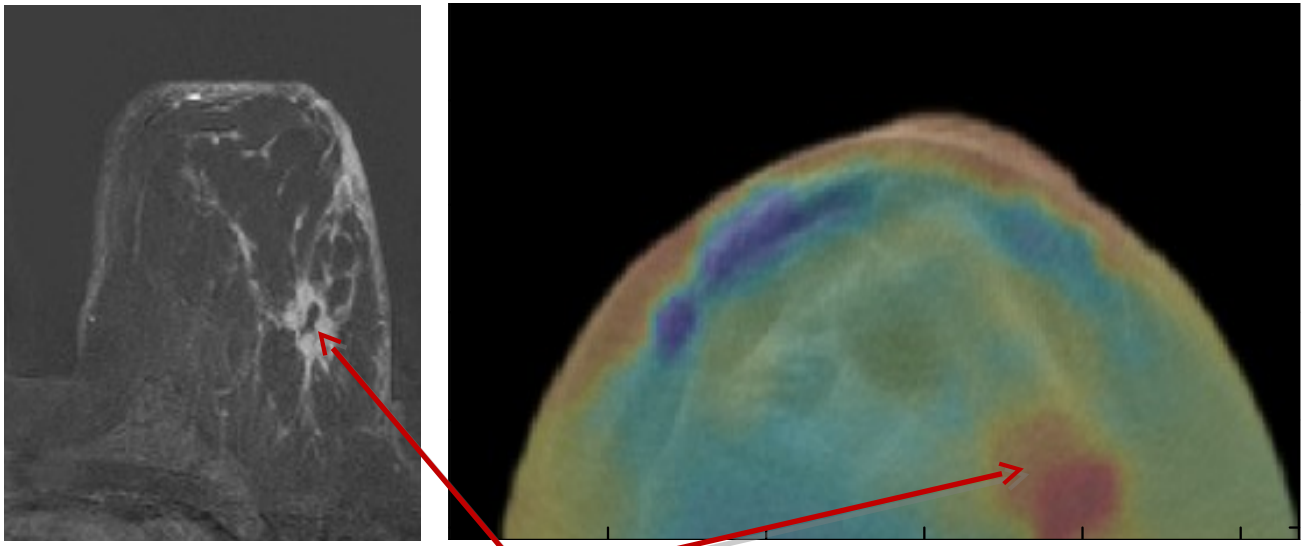
S. Chilingaryan et. all

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

USCT: 3D Ultrasound Computer Tomography

- Goal: Novel and unique imaging device for early breast cancer detection with resolution of MRI and cost of X-ray

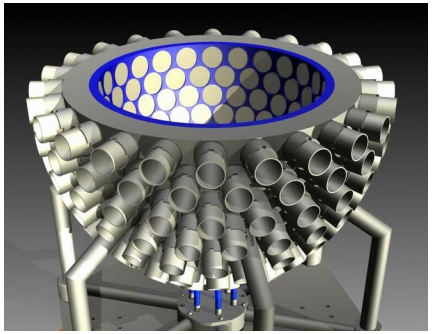
- Basic idea: Surround object with ultrasound transducers in a fixed setup



Ground Truth: MRI

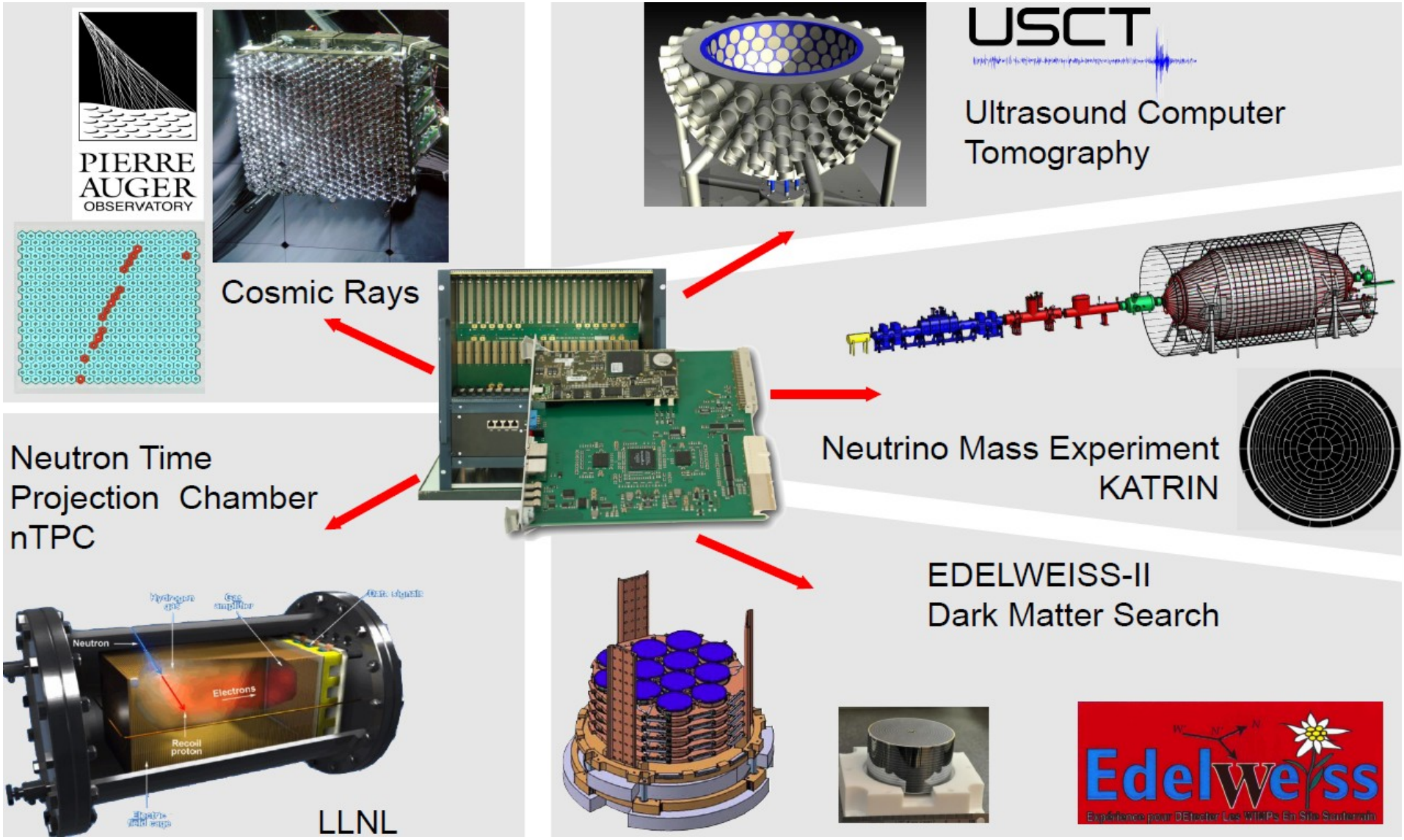
Tumor

USCT (Reflection + Speed of sound imaging)

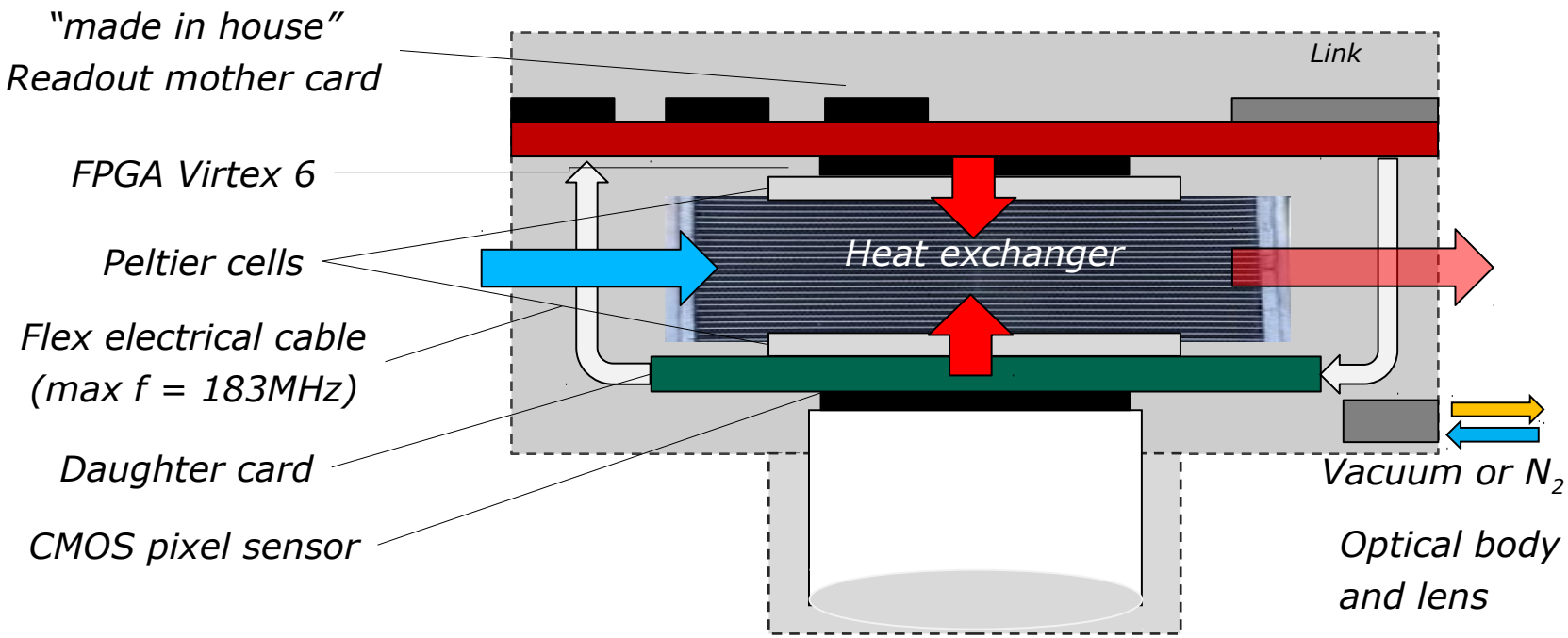


Sensor Basin

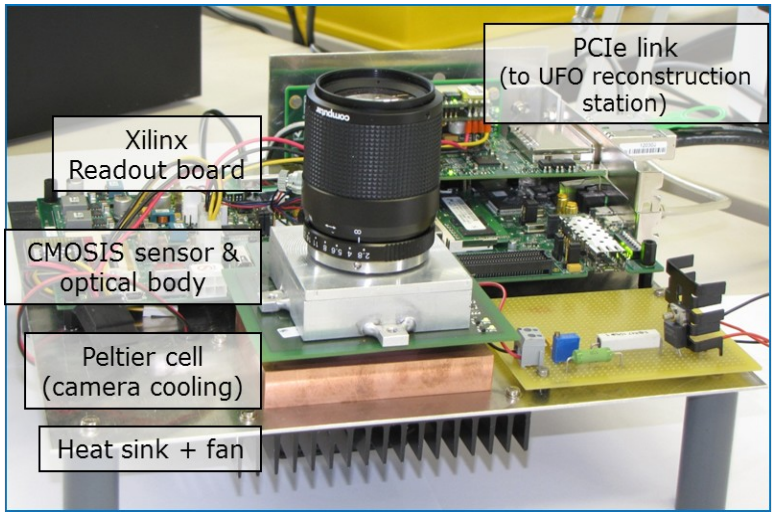
DAQ Electronics



DMA: Programmable Streaming Camera



- ***Exchangeable CMOS sensor***
- ***1Mpix, 5000 fps, 10 bits***
- ***Self-trigger & Data compression***
- ***On-line elaborations and control***
- ***Full Programmability***
- ***Infiniband connection***

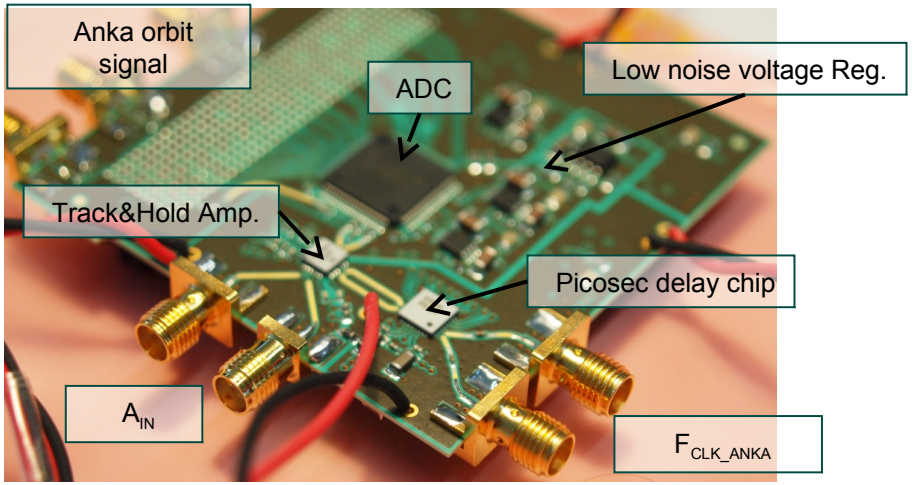


First Prototype

S. Chilingaryan et. all

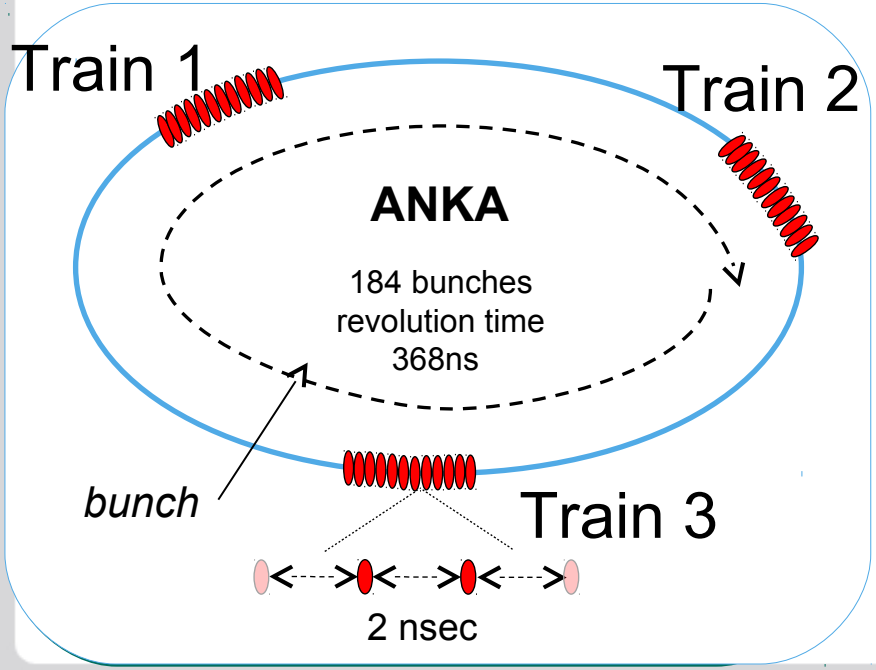
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Hot-electron bolometer

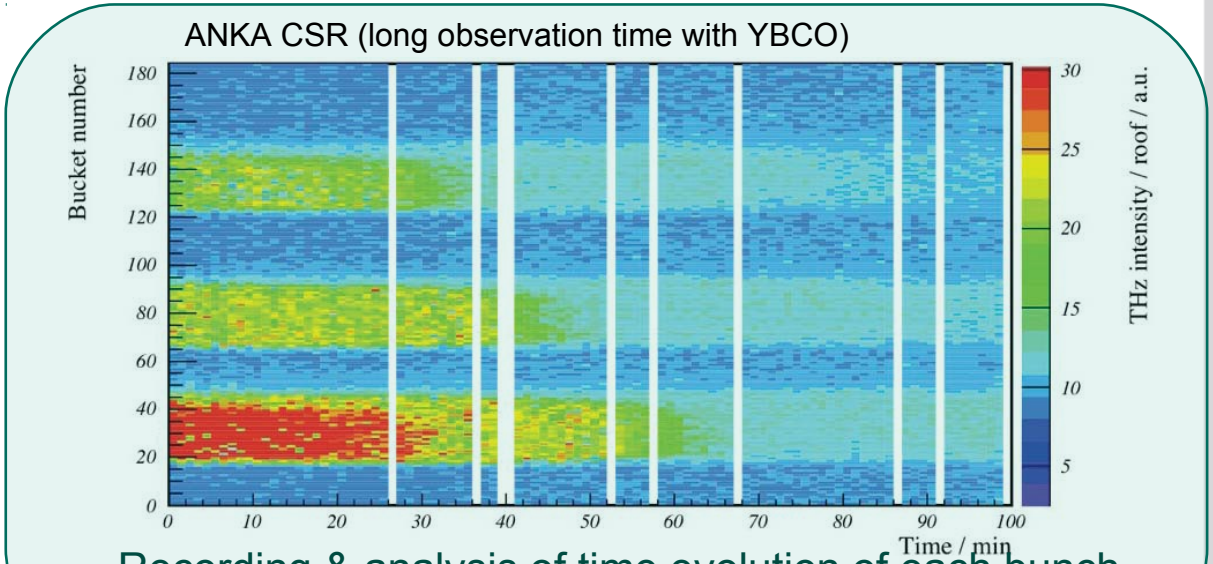


Goal: Measure properties of electron bunches in synchrotron storage rings

- Measure peak amplitude with resolution down to meV
- Measure pulse width of each bunch with resolution down to picoseconds



ANKA test beam



Recording & analysis of time evolution of each bunch in a multi-bunches accelerator filling-scheme

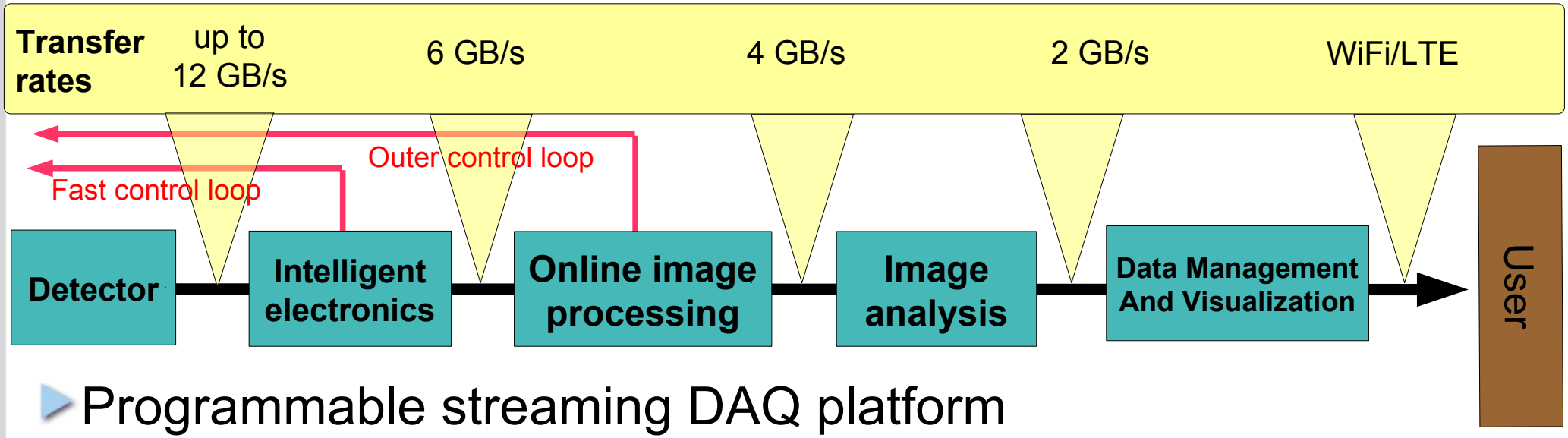
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

New challenges for DAQ Software

- **New detectors: Extreme data rates**
  - Can’t store all the data: online data reduction is needed
  - Moving between sites is slow: remote analysis services are needed
- **Increased automation: High throughput of samples/runs**
  - Detect the problems already during acquisition
  - Automate curation of the stored data
- **Uneven resource utilization: High investments and power balance**
  - Multiple experiment phases: Acquisition, analysis, curration, etc.
  - Huge load spikes before meetings and conferences
- **More complex data processing chains**

DAQ platform for high-speed detectors

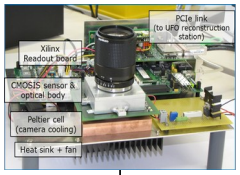
A scalable platform for rapid deployment of heterogeneous, high-bandwidth DAQ systems with online monitoring and control



- Programmable streaming DAQ platform
- Multiple levels of scalability to adapt for fast sensors
- Build of readily available cheap of-the-shelf components
- Modular image-processing framework supporting diversity of the hardware platforms and based on open standards
- An extensible library of efficient algorithms
- Remote data analysis services

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Detector Connection



	PCI express (external)	Infiniband	Ethernet
Status:	Used	Supported	In development
Bandwidth:	12 → 24 GB/s	Up to 24 GB/s	Up to 24 GB/s
Latency:	1 – 3 us	2 – 4 us	Protocol dep.
Length:	Up to 10m	Up to 200m	Internet
Hot-Plug:	Yes, difficult	Yes	Yes, easy

Easier integration,
Complex development

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Alps: Advanced Linux PCI Services

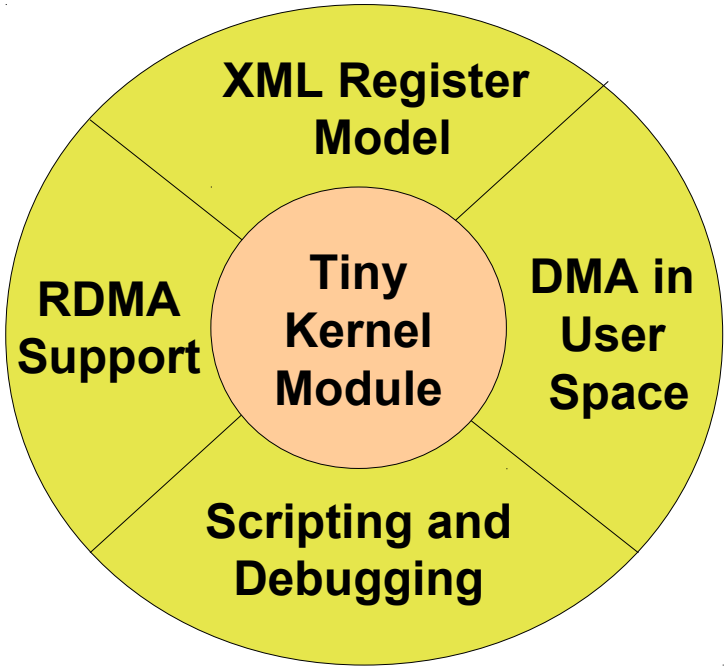
Karlsruhe Institute of Technology

A Linux driver platform for custom PCI electronics

Motivation

- Synchronization Software and Hardware development
- Easy hardware debugging
- Keeping drivers up to date with latest Linux kernels
- Multiple common components

Only a simple XML model is required to start with the new hardware



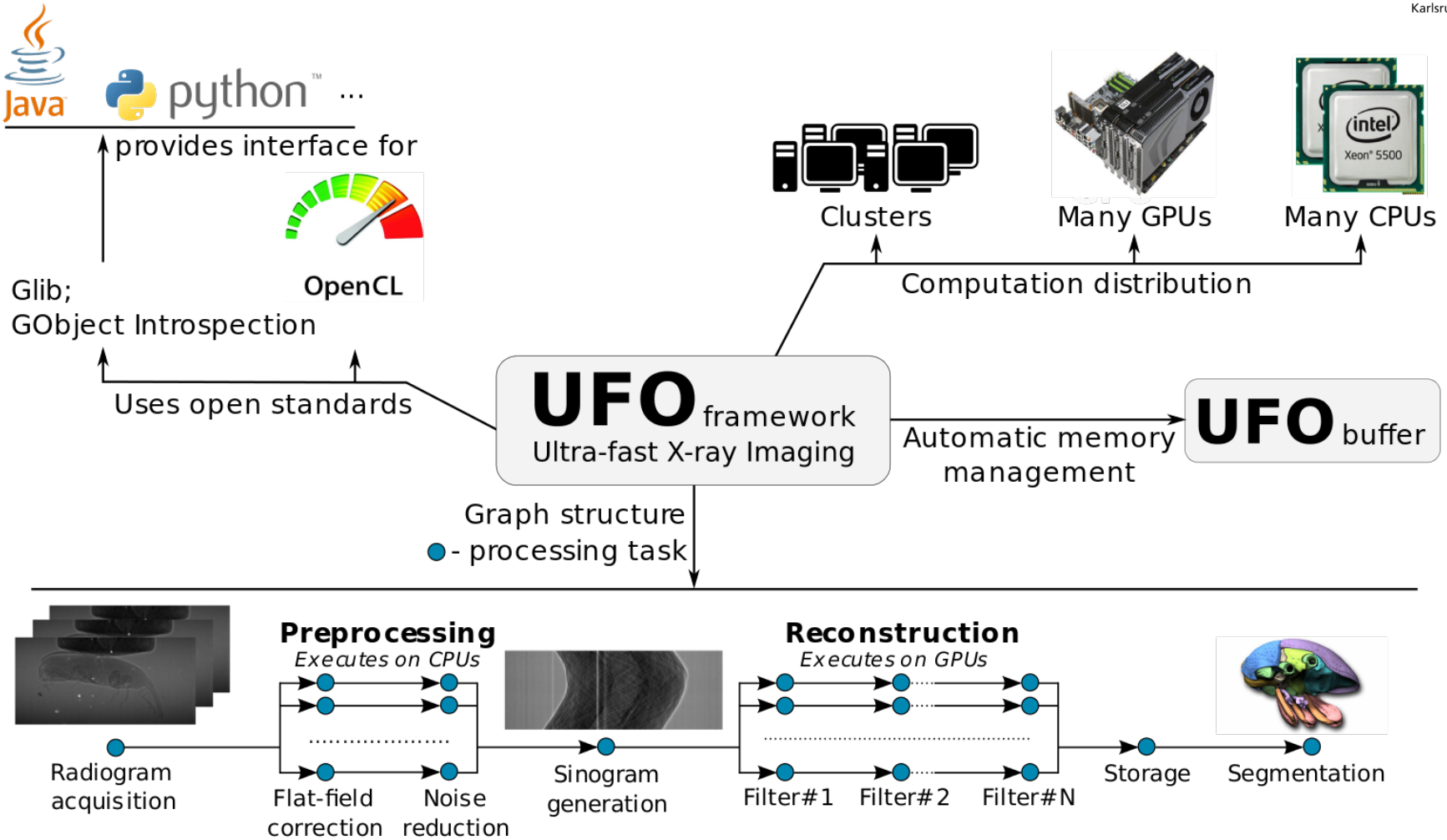
Components

- PCI driver
- User-space library
- Register Protocols
- DMA Engines
- RDMA Support
- Command line utility
- Scripting
- Web API

S. Chilingaryan, A. Kopmann

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

UFO Image Processing Framework



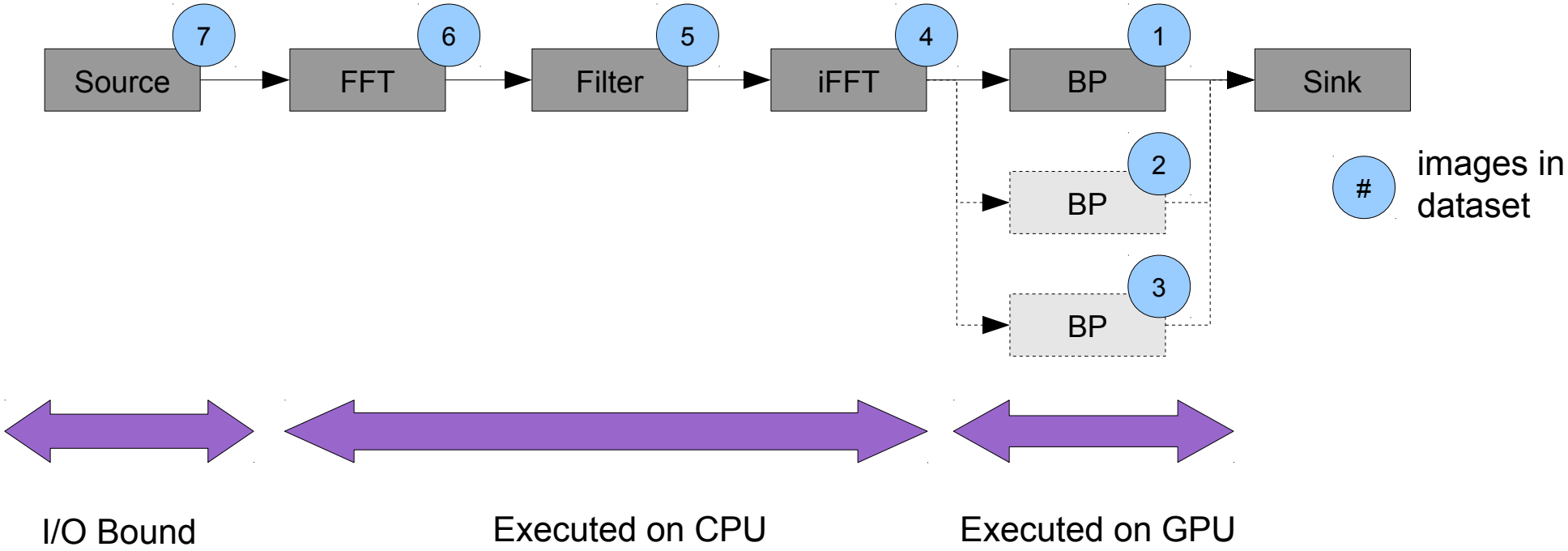
Fully pipelined architecture supporting diversity of the hardware platforms and based on open standards for easy algorithms exchange. Easy prototyping with Python and other scripting languages.

S. Chilingaryan et. all

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Pipelines for Heterogeneous Architectures

- Connect filters to achieve the desired data flow and results
- Each filter may process data on a GPU, FPGA, or CPU with OpenCL
- Run-time schedules distribution of data and execution of tasks
- Execution is pipelined for efficient use of available resources
- Duplicates sub path for multi-GPU execution



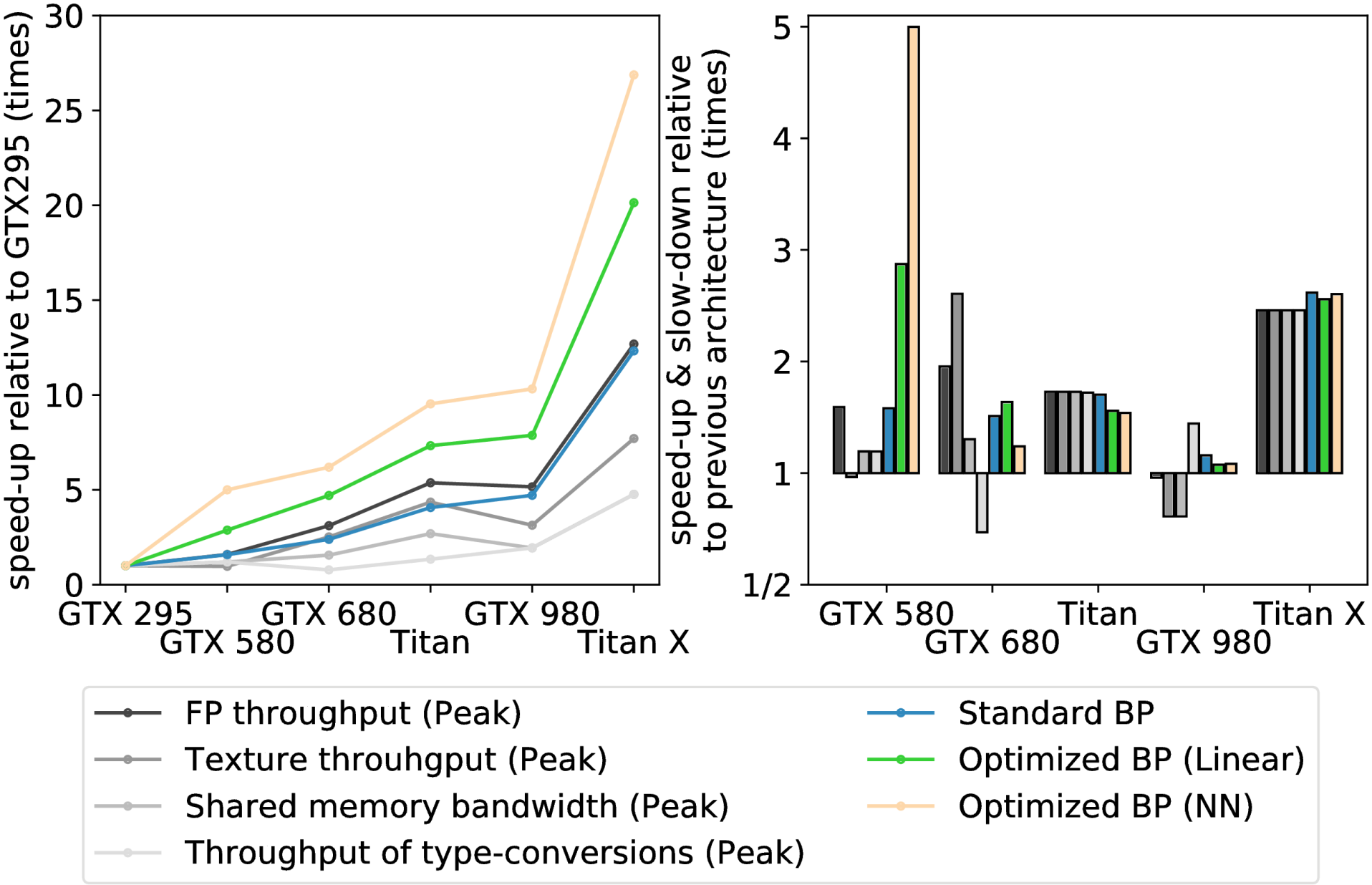
Tunning the code for GPU architectures

GPU architecture is complex (and rapidly evolving) and includes different type of computational units and caches

- **Functional Units**
  - Core Units (Floating-point and some integer operations)
  - Double and half-precision op.
  - Special Function Units (SFU)
  - Load-Store Units
  - Texture Units
- **On-chip Memories**
  - Shared Memory
  - Local Memory Cache
  - Constant Memory Cache
  - Texture Cache

Multiprocessor on NVIDIA Pascal GPU

Use-case: Tomographic Reconstruction

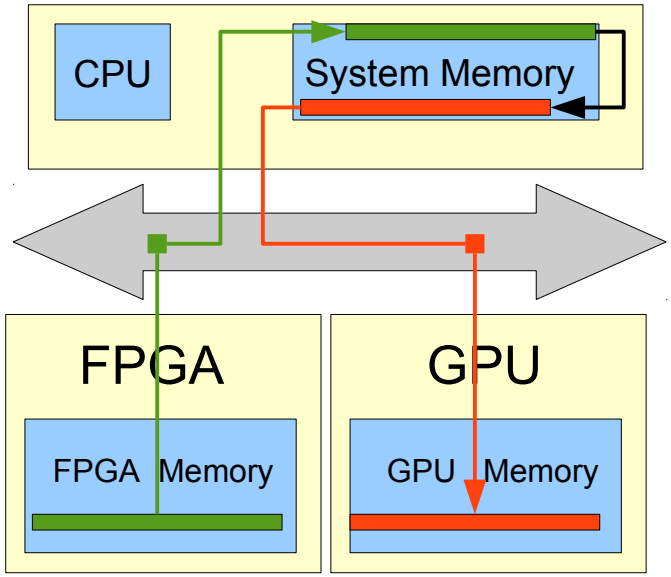


S. Chilingaryan et. all

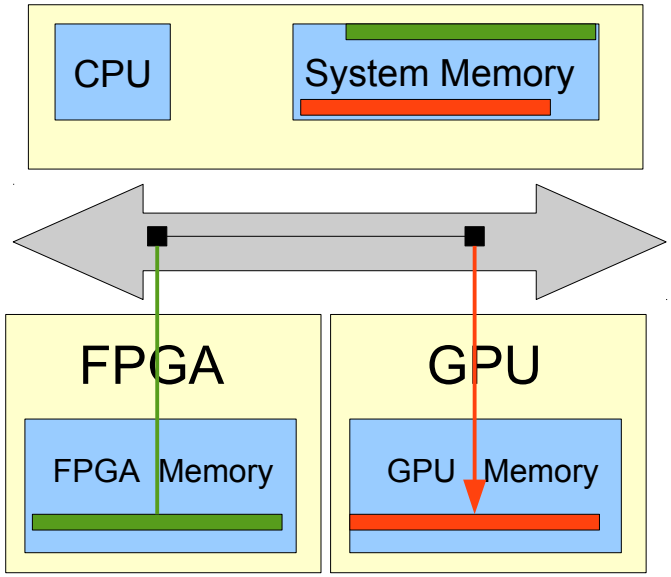
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Improving device communication

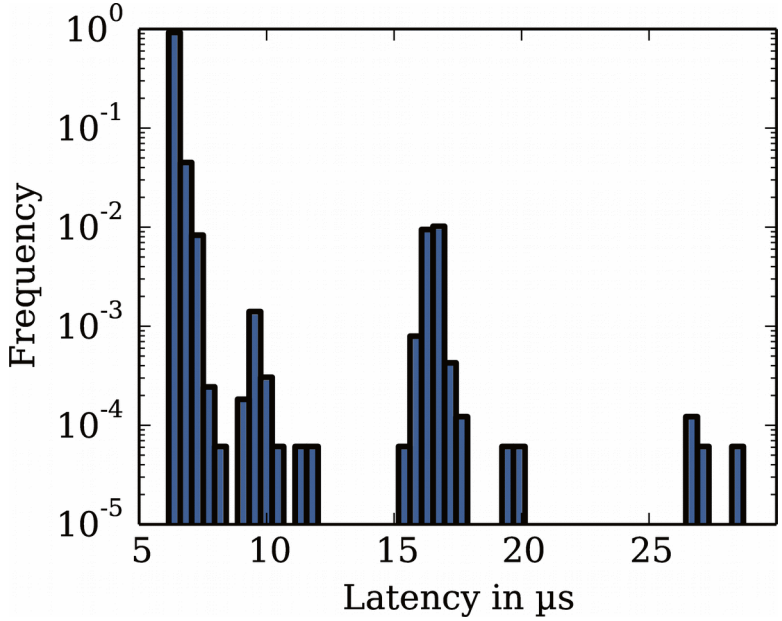
+ 4 memory accesses



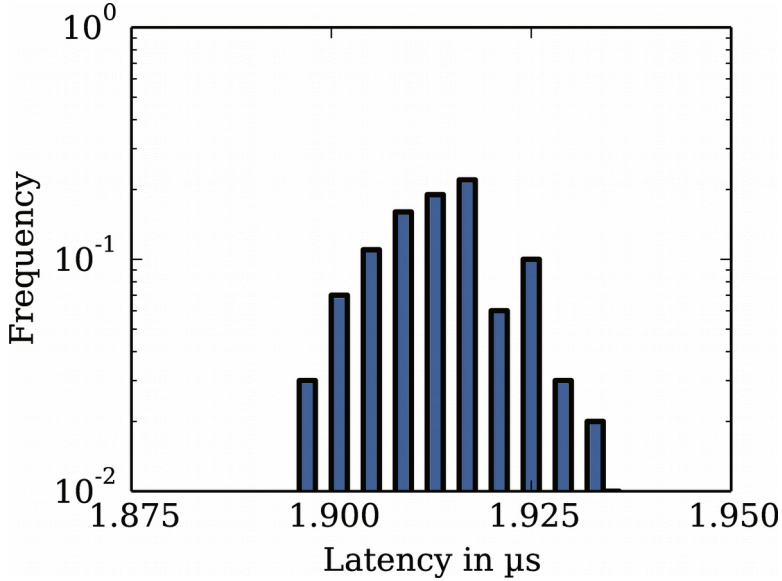
PCIe bus



Standard



GPUDirect

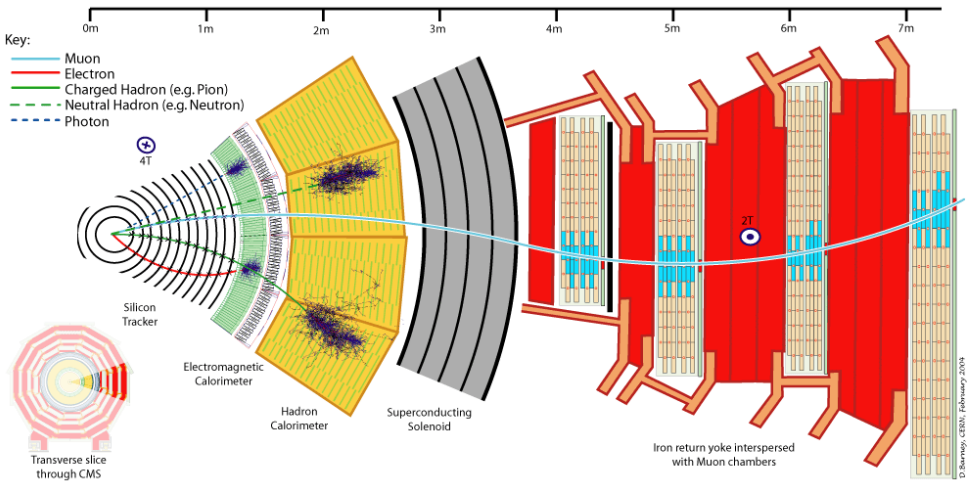


S. Chilingaryan et. all

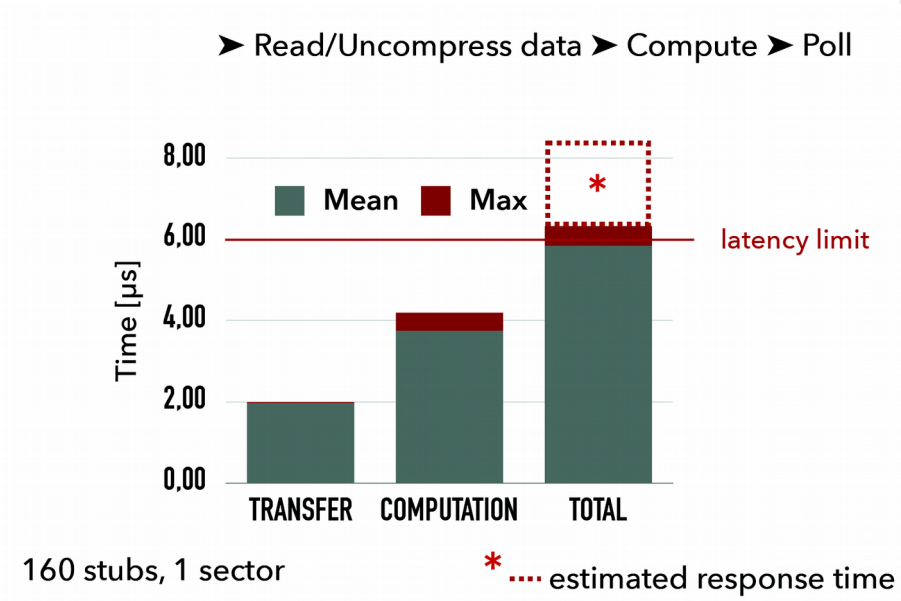
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Use case: GPU Trigger for CMS

Adopt FPGA algorithm for GPU with Hough transform to identify track candidates within 6 us and with high throughput



Achieved Performance of the prototype



- Collisions every 25ns (approx. **100TBit/s**)
- Detector is split into sections at approx. **150Gbit/s** data rate
- Pipeline gives us **6µs** for Track-Finding

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Low-latency Ethernet Communication

- ROCe encapsulates Infiniband headers in the payload of UDP packet which can traverse standard Ethernet infrastructure.
- 4791 port in UDP header indicates ROCe packet and the UDP payload, then, includes additionally an Infiniband header and checksum
- Infiniband header contains ID of remote queue and a virtual address to read/write the data from/to.

UDP Packet (1500 – 9000 bytes)

Eth Header	IP header	UDP header	Payload	Eth CRC

ROCe Packet

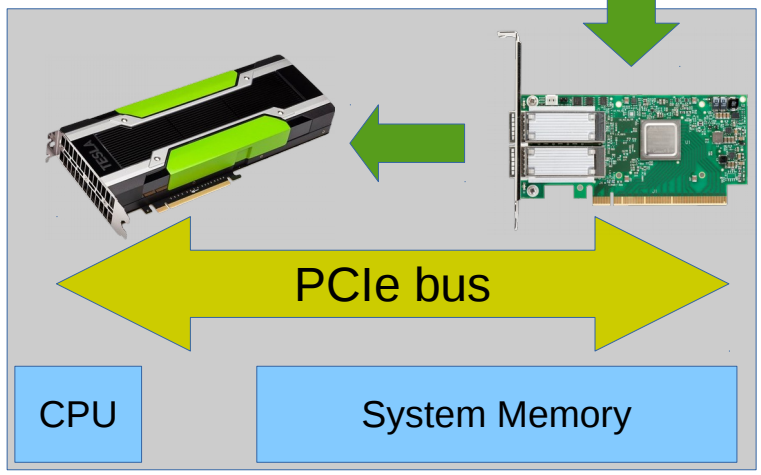
Eth Header	IP header	UDP header	IB header	Payload	IB CRC	Eth CRC

Contains remote queue and virtual address

Port 4791 indicates ROCe packet

100m / 10km

Traverses Ethernet switches and routers



<u>Possible collaboration:</u>
Xilinx implementation available at ESRF

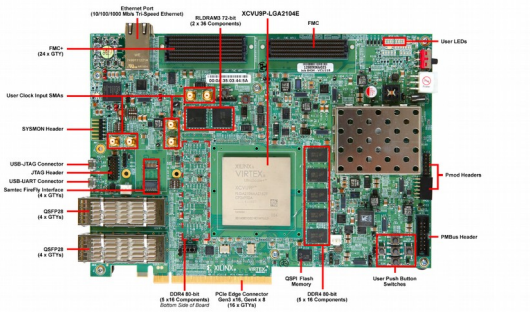
Cluster-based Processing

- Master server only configures the hardware and the data processing-nodes, but doesn't receive any data
- Processing nodes send UDP packet with buffer addresses to the FPGA which responds with the data in round-robin fashion.

Master Server

UDP: Control-connection
Set and read hardware registers, no data streaming

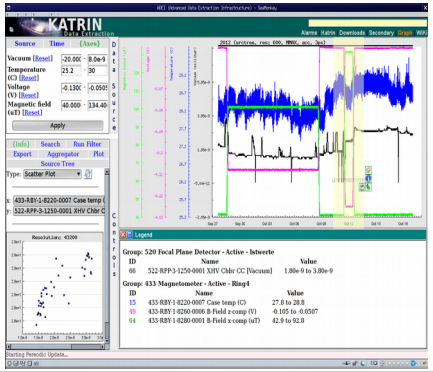
Processing Servers



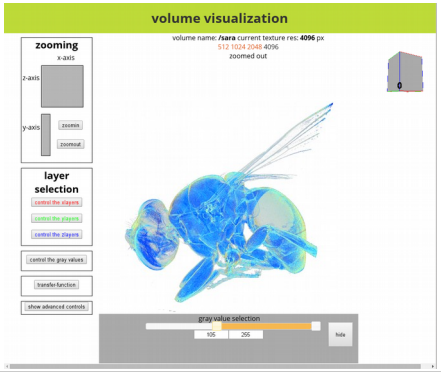
ROCe: Stream data to multiple servers
in round-robin fashion

UDP: report address of free buffers to FPGA

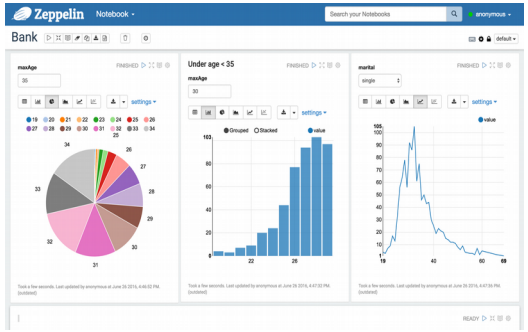
ADEI Data Management Platform



Time Series Inspection



Volume Visualization



Data Mining and Analysis

Single Point of Access

Visualization

Data Analysis

Data Mining

Remote Access

ADEI

Data Inspection

High Availability

Data Safety

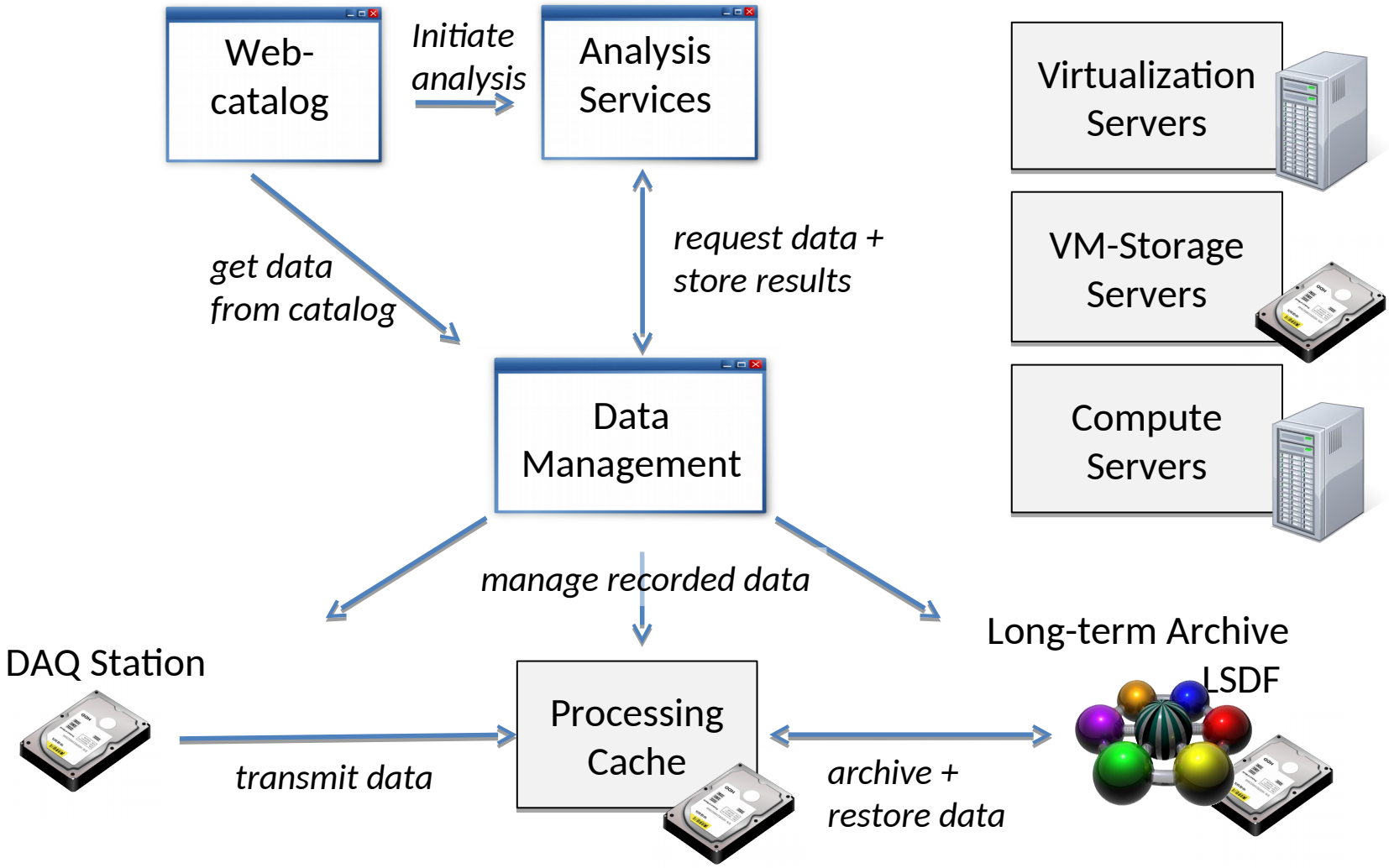
Data Abstraction

Multiple Subsystems Generating Data

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Remote Data Analysis

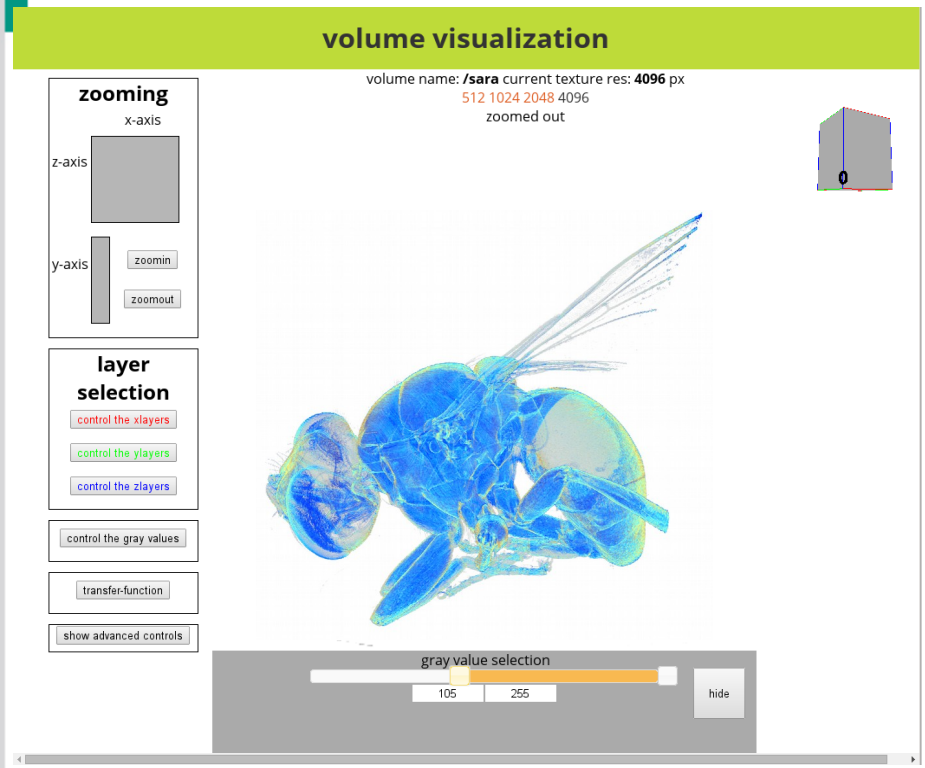
Rapid growth of data volumes requires on-site data processing and reduction services.



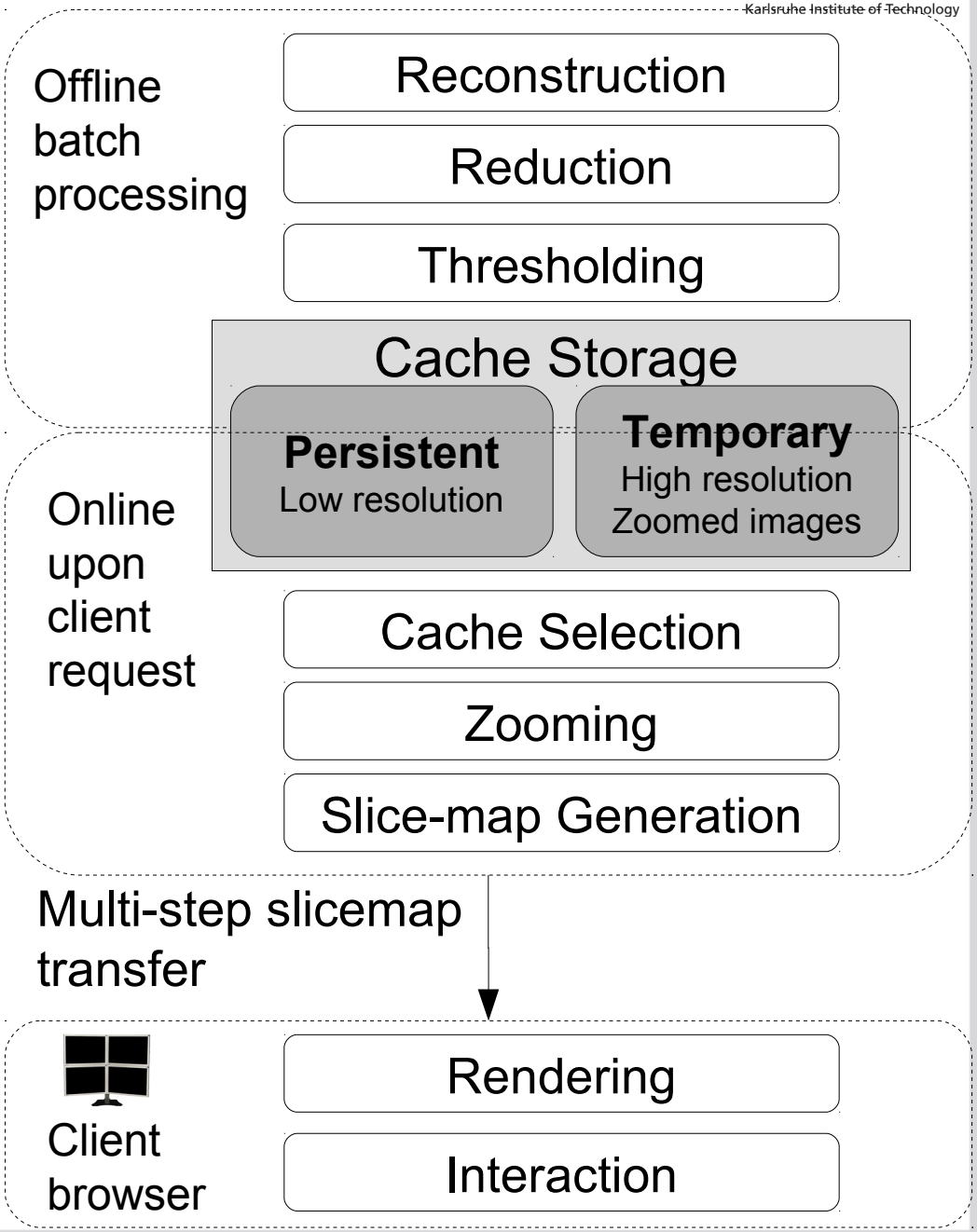
S. Chilingaryan et. all

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

WAVe: Web-based volume visualization



- Balance between offline jobs, online jobs, and client-side rendering
- Multi-step transfer to allow quick preview and improve resolution later
- Multiple zooming levels for inspecting fine details
- High-quality cuts
- Multi-modality rendering support



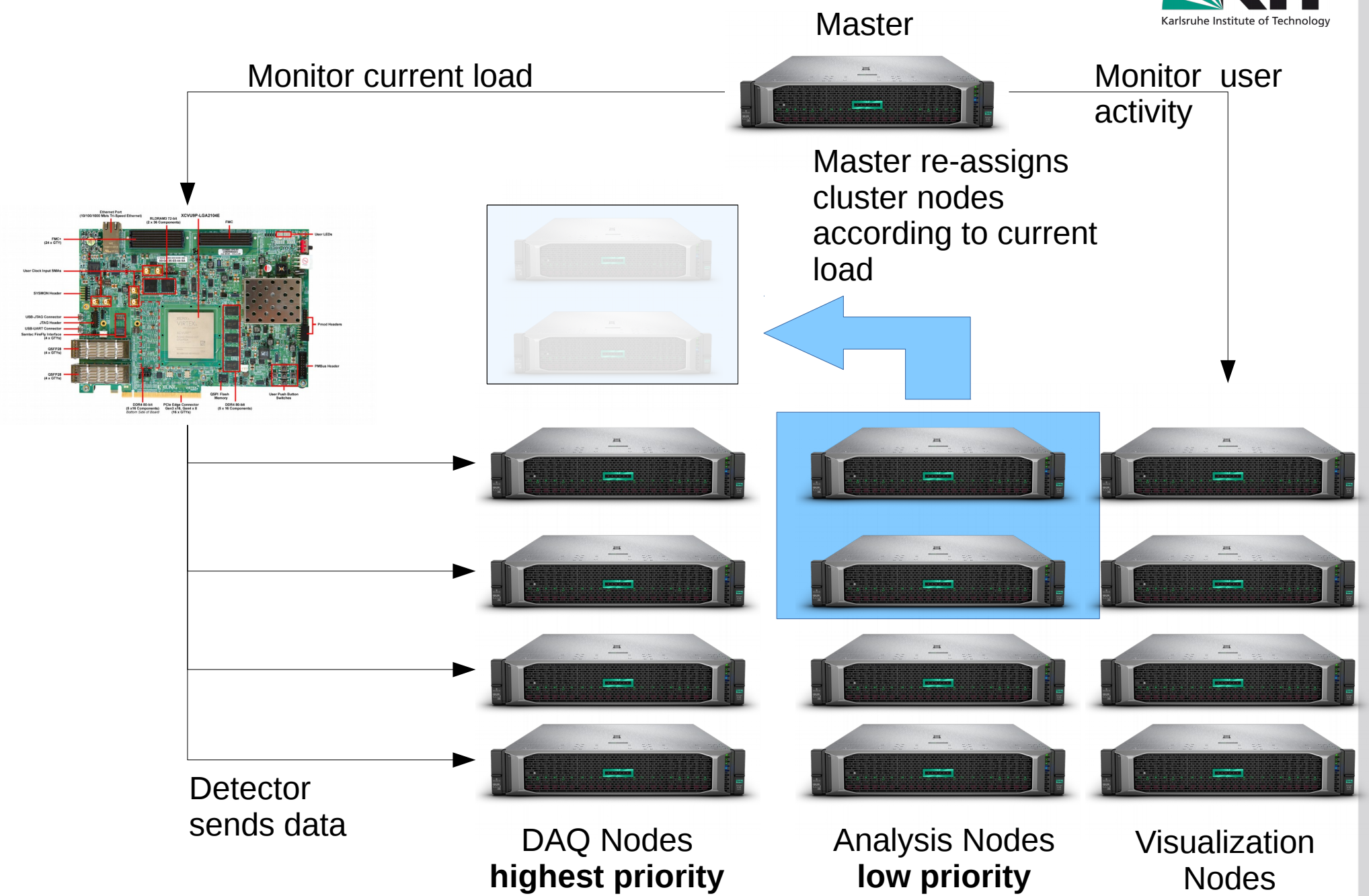
S. Chilingaryan et. all

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Experiment Life-cycle

- **Multiple-phases in experiments**
  - Data Acquisition Phase
    - Data reduction; Real-time reconstruction, monitoring, and slow control
  - Offline Data Analysis Phase
    - Quality control and automated data preparation (i.e. Registration, fully automated segmentation, generation of previews, etc.)
  - Interactive Remote Analysis Phase
    - Data Visualization , user-assisted analysis
- **Improving utilization of IT infrastructure**
  - Similar resources required during all phases: GPUs, Storage, ….
  - Readout nodes can be used for offline analysis when detector is not streaming data
  - Not critical if offline analysis is executed few hours or days later
- **Priorities**
  - **Highest**: Readout
  - **Normal**: Monitoring and serving interactive user requests
  - **Idle**: Offline analysis and data pre-processing.

Re-balancing load

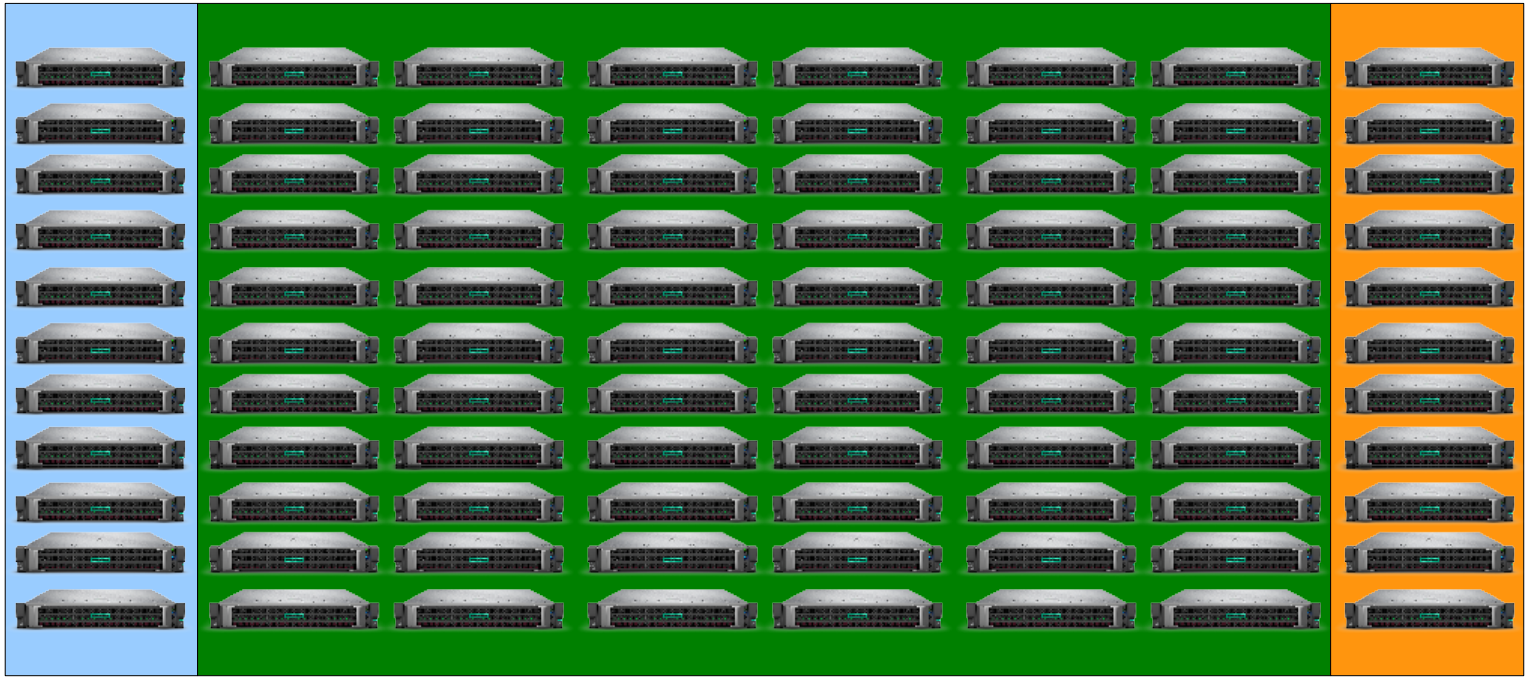


Data Acquisition Phase

DAQ Nodes

Interactive Nodes

Night



DAQ Nodes

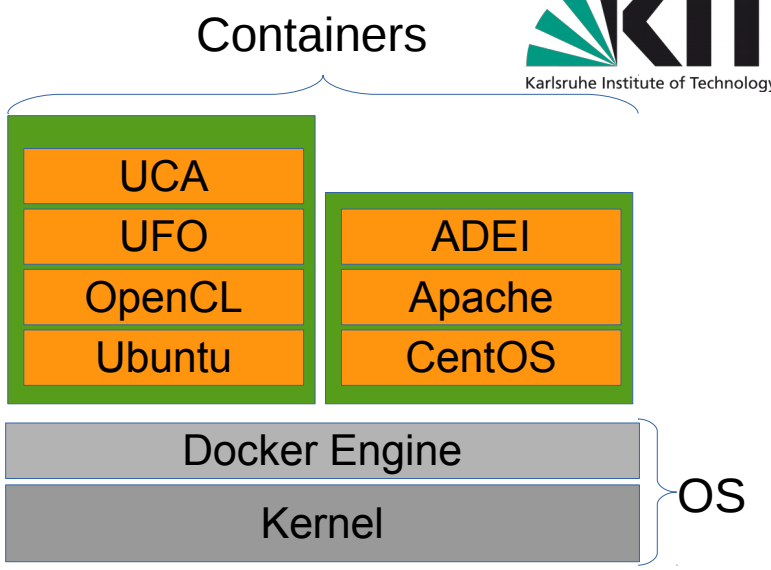
Analysis Nodes

Interactive Nodes

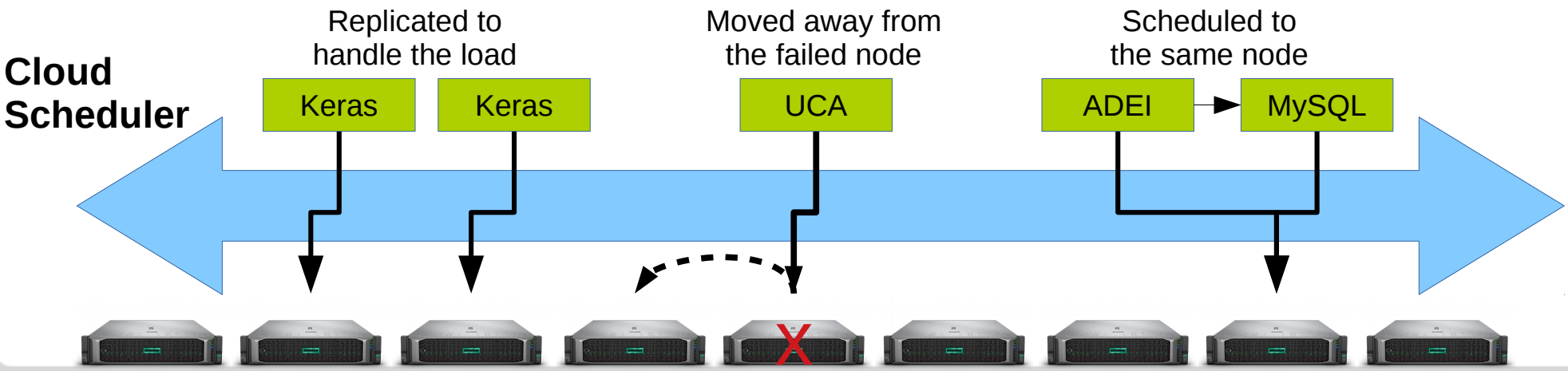
Remote users connect

DAQ Nodes Analysis Nodes Interactive Nodes

Software Platform

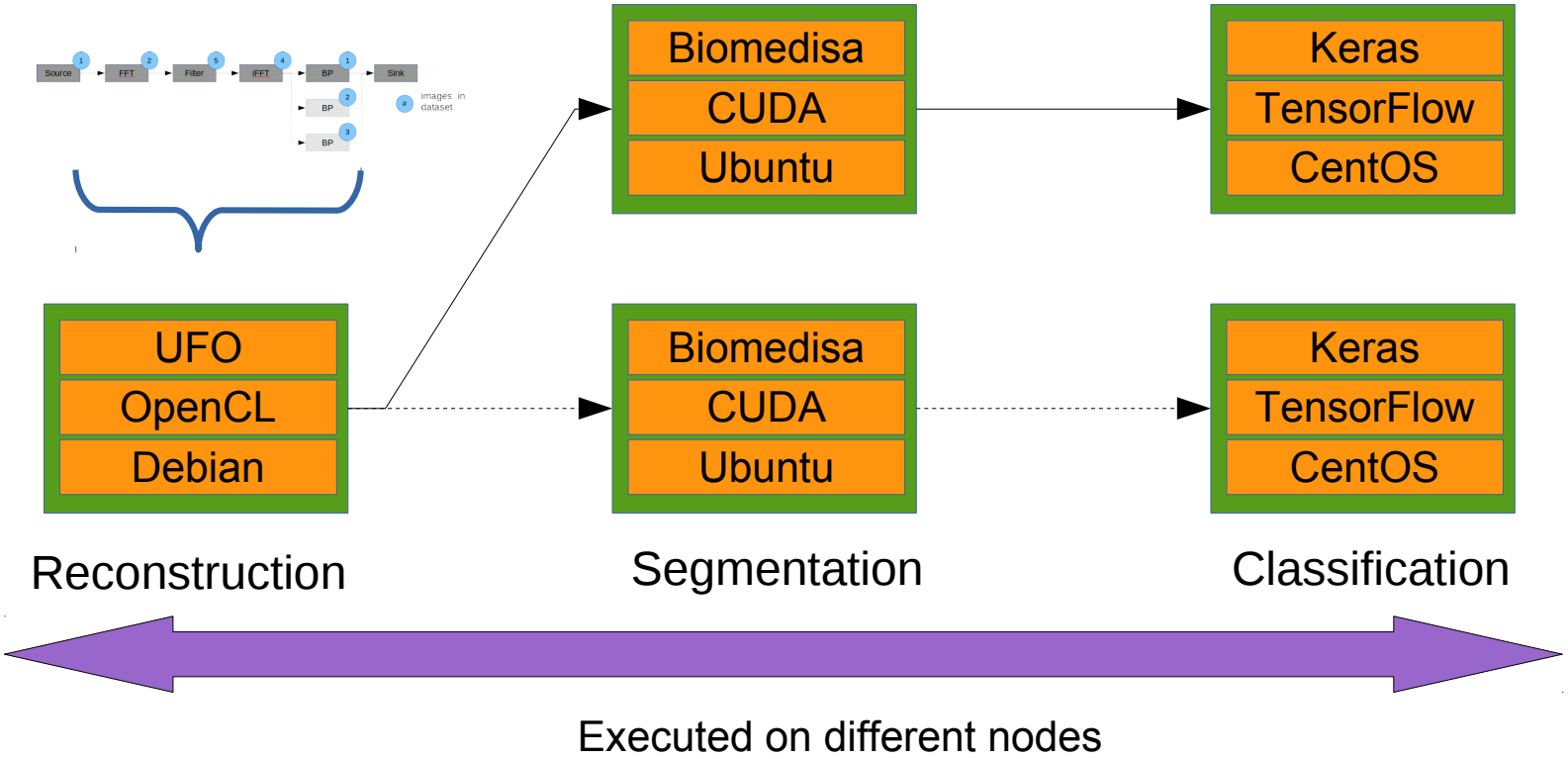


- **Containers**
  - Pack application with all dependencies
  - Isolation (problems & resources)
  - Low overhead
- **Private Cloud Infrastructure**
  - Load-balancing: Stops / starts additional replicas according to the load
  - High-availability: Restart failed services, migrate from failed nodes
  - Resource management: Allocate nodes to apps, set memory/cpu limits
  - Security: Allows to share hardware without sharing the data

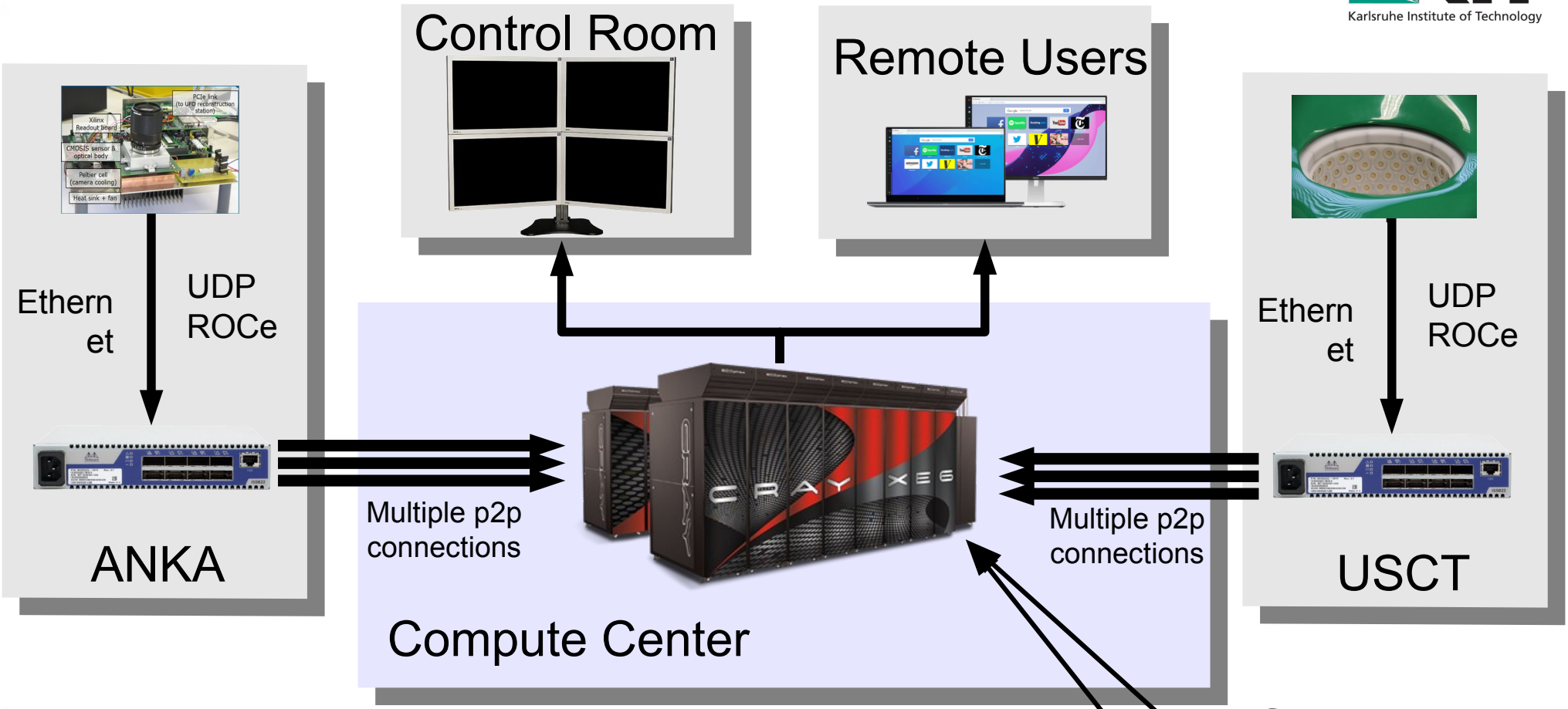


Container-Native Workflow

- Connect containers to achieve the desired data flow and results
- Each filter may process data on a GPU or CPU with CUDA/OpenCL/OpenMP/...
- Scientific Workflow Engine schedules execution of task
- Execution is distributed for efficient use of available nodes and network bandwidth
- Duplicates sub path for multi-node execution
- Base on the existing Scientific Workflow Engine, like Project Argo or Pegasus



DAQ Cloud



- Cloud-based control system based on standard Ethernet
- GPU computing and cheap of-the-shelf components
- Both online and offline processing is performed in the Private Cloud to simplify maintenance and improve resource utilization
- Multiple experiments can be served by the same cloud to further improve resource utilization
- Distributed data processing framework based on the Containers and RDMA

Slow Devices

Katrin cFP

Tristan cFP

Collaboration with CRD

- Acquisition, organization, and visualization of the data from ASEC and SEVAN networks
- High-speed imaging of lightning strikes on Nor-Amberd and Aragats stations
- Drones to monitor electric field in the thunder clouds (?)
- Co-supervision of Master and PhD students (?)