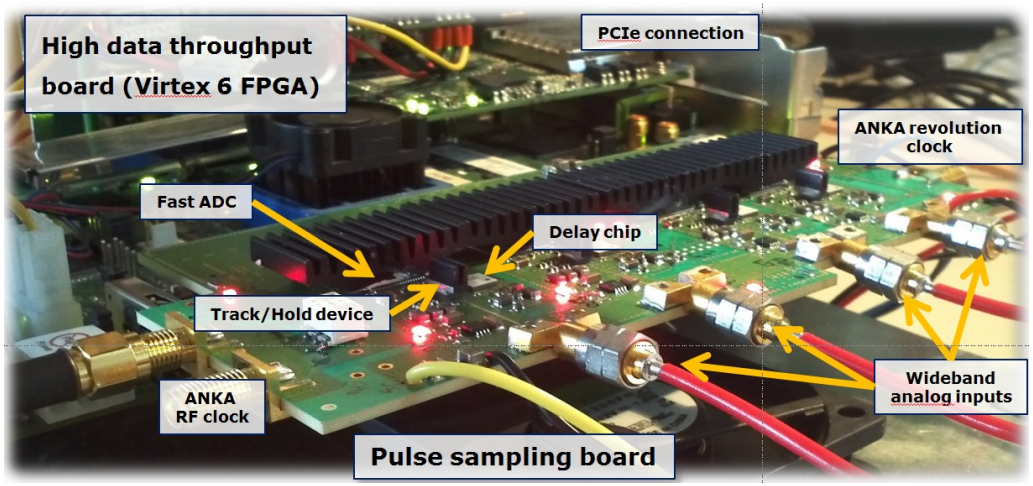
Karlsruhe Institute of Technology

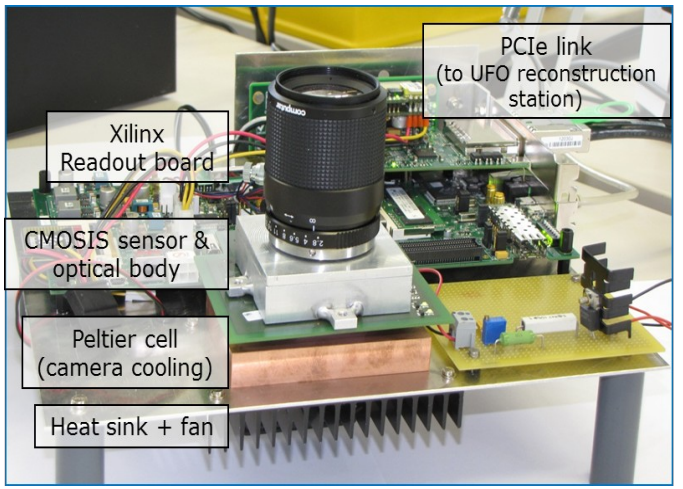
Efficient GPU-enabled computing infrastructure for rapid prototyping high-speed scientific detectors

S. Chilingaryan, M. Caselle, T. Dritschler, T. Farago,
A. Kopmann, U. Stevanovic, M. Vogelgesang

Hardware, Software, and Network Organization



Picosecond Sampling Electronics for Terahertz Synchrotron Radiation

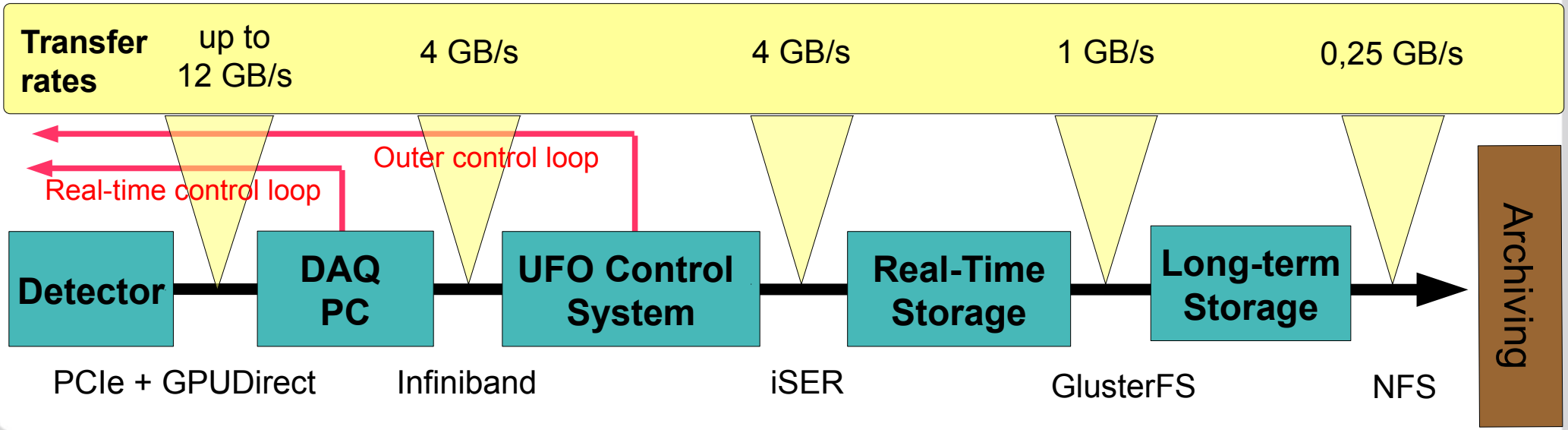


Prototype of Streaming PCIe Camera for scientific applications

Requirements

- **Handling of Sensors with data rates up to ~ 12 GB/s (8-12 bit)**
- **Real-time control loop based on 2D Images + Online compression**
  - In-flow 12 GB/s, unpacked up to 16 GB/s (16 bit)
- **Slow control loop based on 3D Tomographic Images**
  - In-flow 4 GB/s, unpacked 32 GB/s (single-precision floating-point)
- **Raw data storage at full speed, i.e. 4 GB/s**
- **Long-term storage at 1 GB/s**
- **Integration with Tango Control System**
- **Low administrative effort**

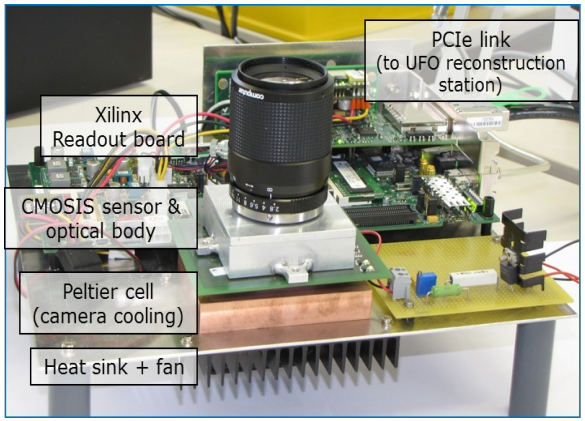
Only a few dozens GB/s max, we are not aiming to XFEL size systems



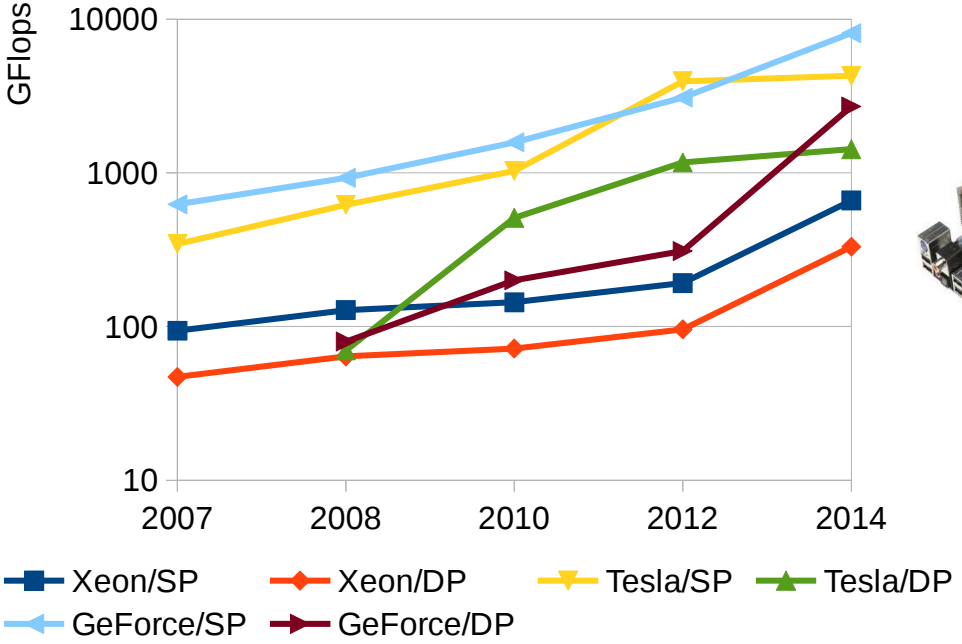
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Concepts

- **Programmable DAQ electronics with PCI-express interface**
- **Distributed control system based on Infiniband interconnects**
- **GPU-based computing**
- **Multiple levels of scalability**
- **Cheap off-the-shelf components**



Prototype of Streaming PCIe Camera for scientific applications

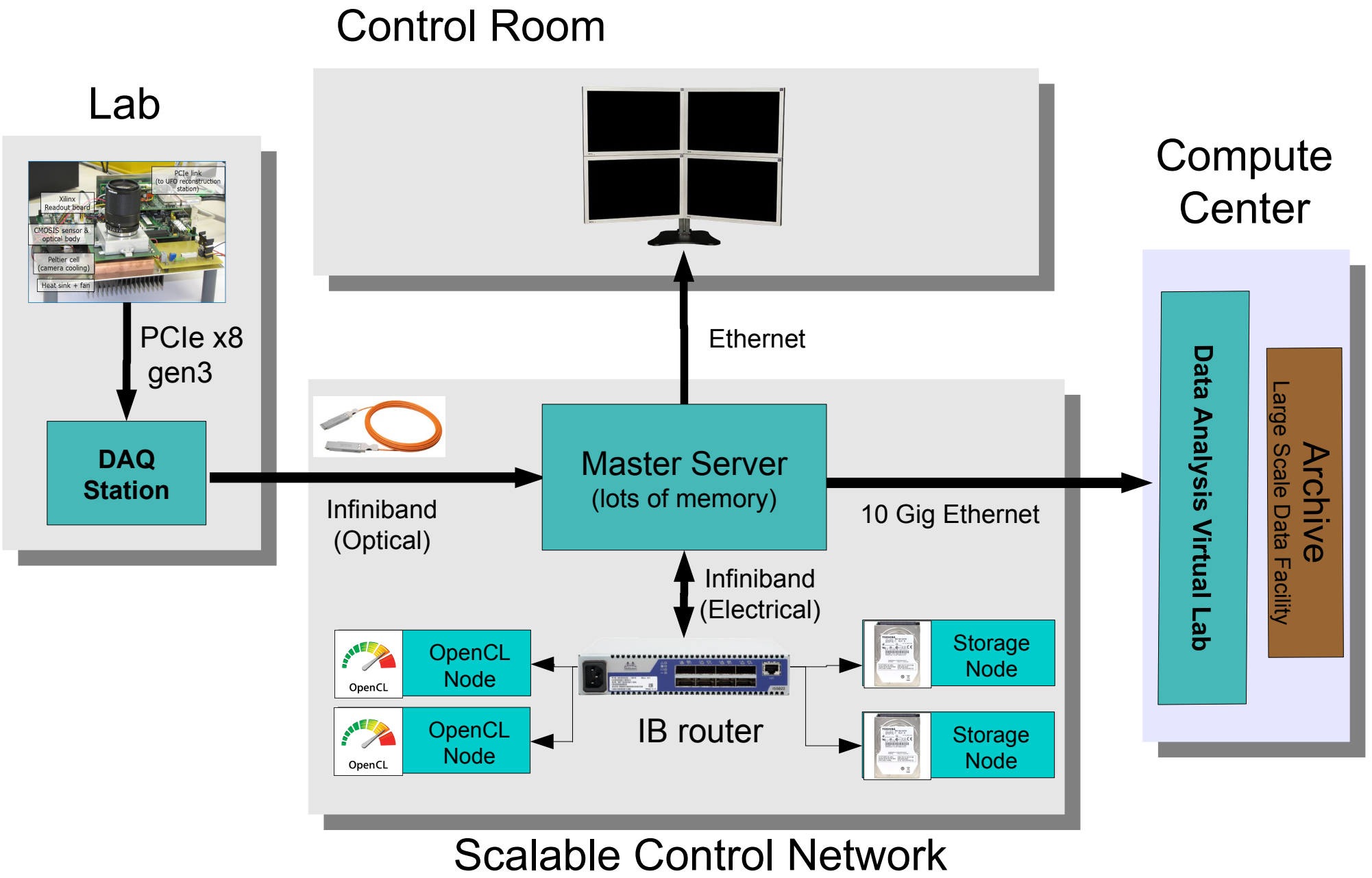


Historical trends of CPU and GPU performance

Easily scalable
Quad-SLI
35 Tflops
for ~ 5000 EUR

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Scalable Control Network



Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

DAQ Station

Activities

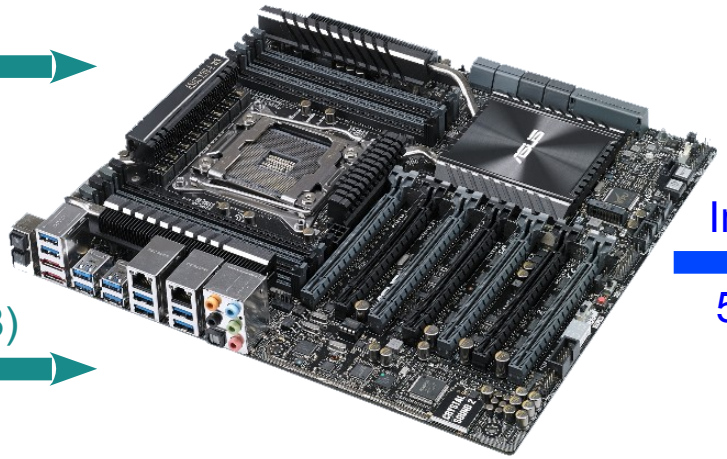
- Data decoding and reduction
- Fast control loop (2 – 5 us)
- Data streaming

Requirements

- High-speed 4-channel memory
- IPMI-based remote control
- Optional fast SSD-based storage
- 3x high speed PCI express slots
- Integrated PLX switch

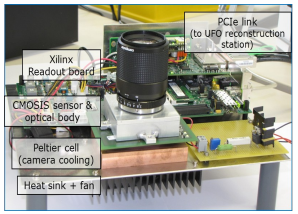
Camera Link

850 MB/s



Infiniband

56 Gbit/s



PCI express (x8)

Up to 8 GB/s

Asus X99-E WS (Intel X99 Chipset)
NO IPMI Remote Control

Asus X99 WS/IPMI (Intel X99 Chipset)
NO PLX Chip

CPU: Xeon E5-1630v3 (total 4 cores at 3.7 Ghz)
GPUs: NVIDIA Tesla K40
Memory: 32 GB (128GB max)
Infiniband: Mellanox ConnectX-3 VPI (FDR)

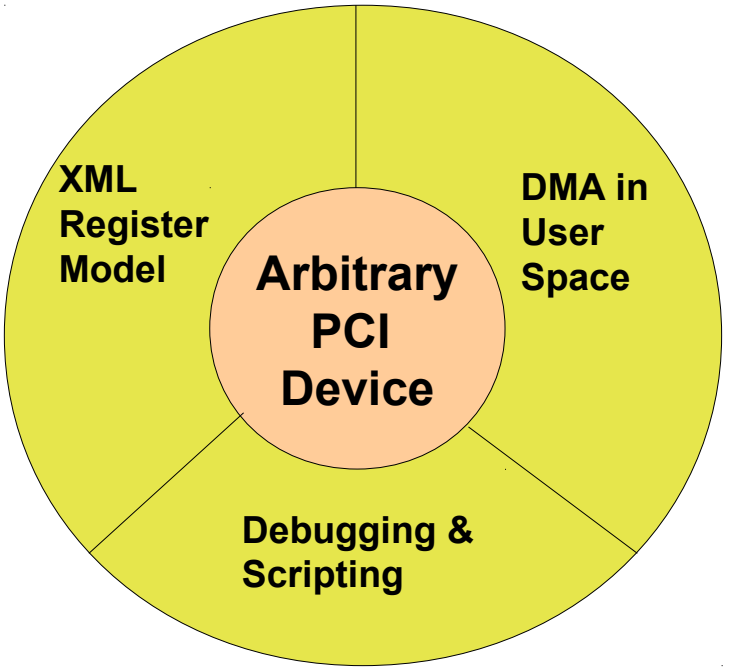
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

ALPS – Advanced Linux PCI Services

A reusable components for custom PCI electronics

Requirements

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



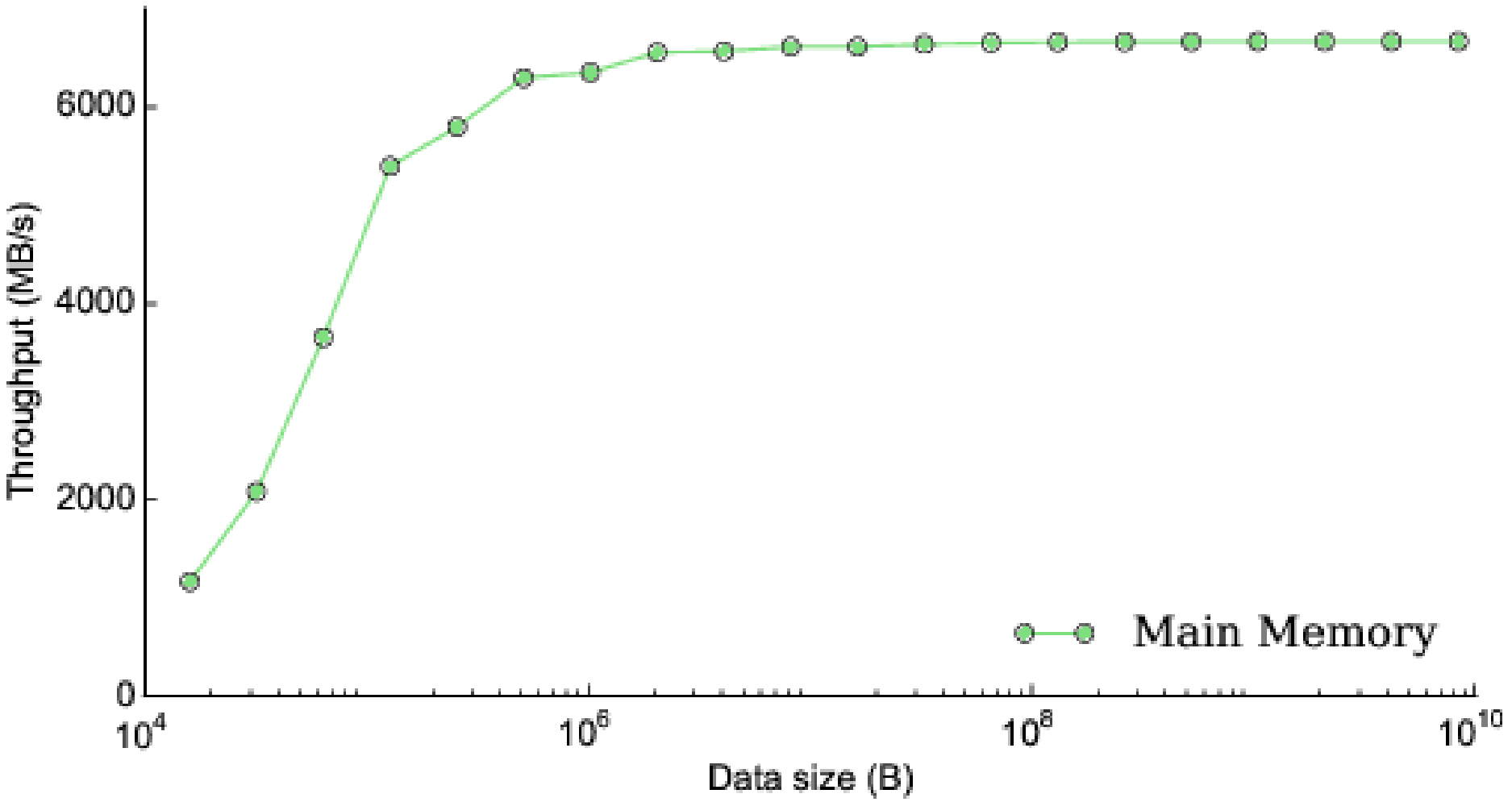
Components

- PCI driver
- Register Model
- DMA Engine
- Custom Event Plugins
- Web API
- Scripting
- RDMA

S. Chilingaryan, A. Kopmann

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

DMA Performance



PCIe x8 gen3

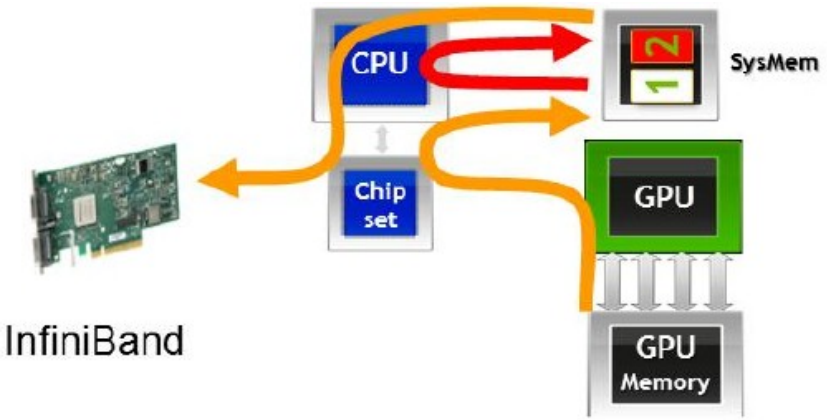
S. Chilingaryan, A. Kopmann

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

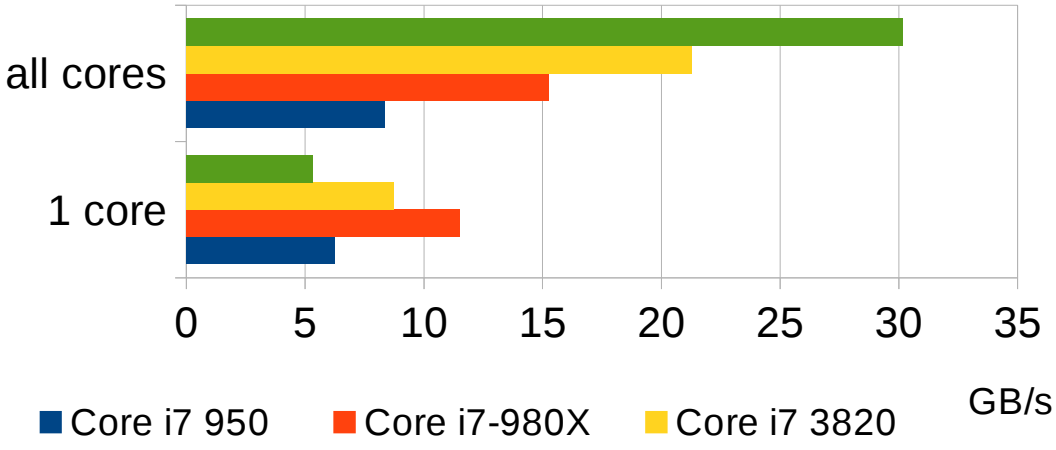
Remote DMA Technologies

We expect up to 12 GB/s from FPGA.

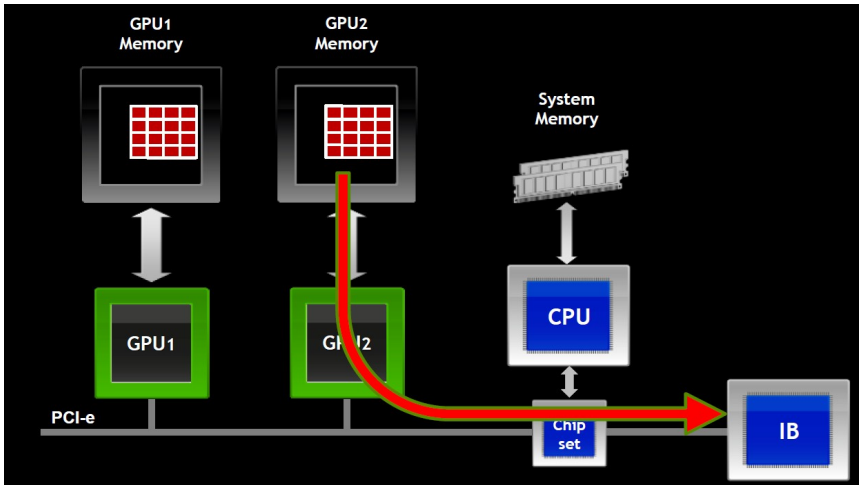
Standard

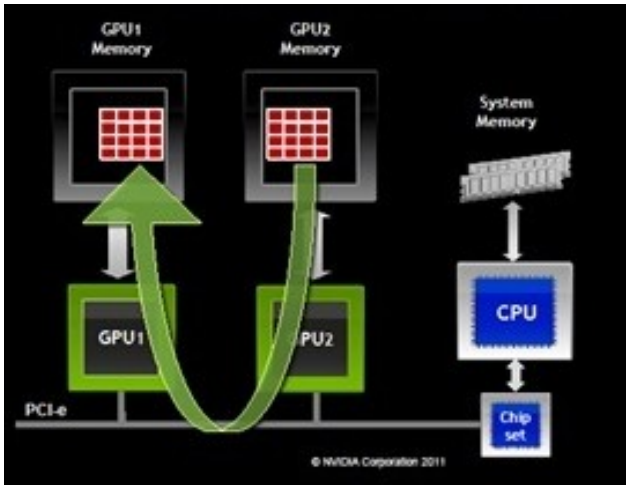


Memcopy performance



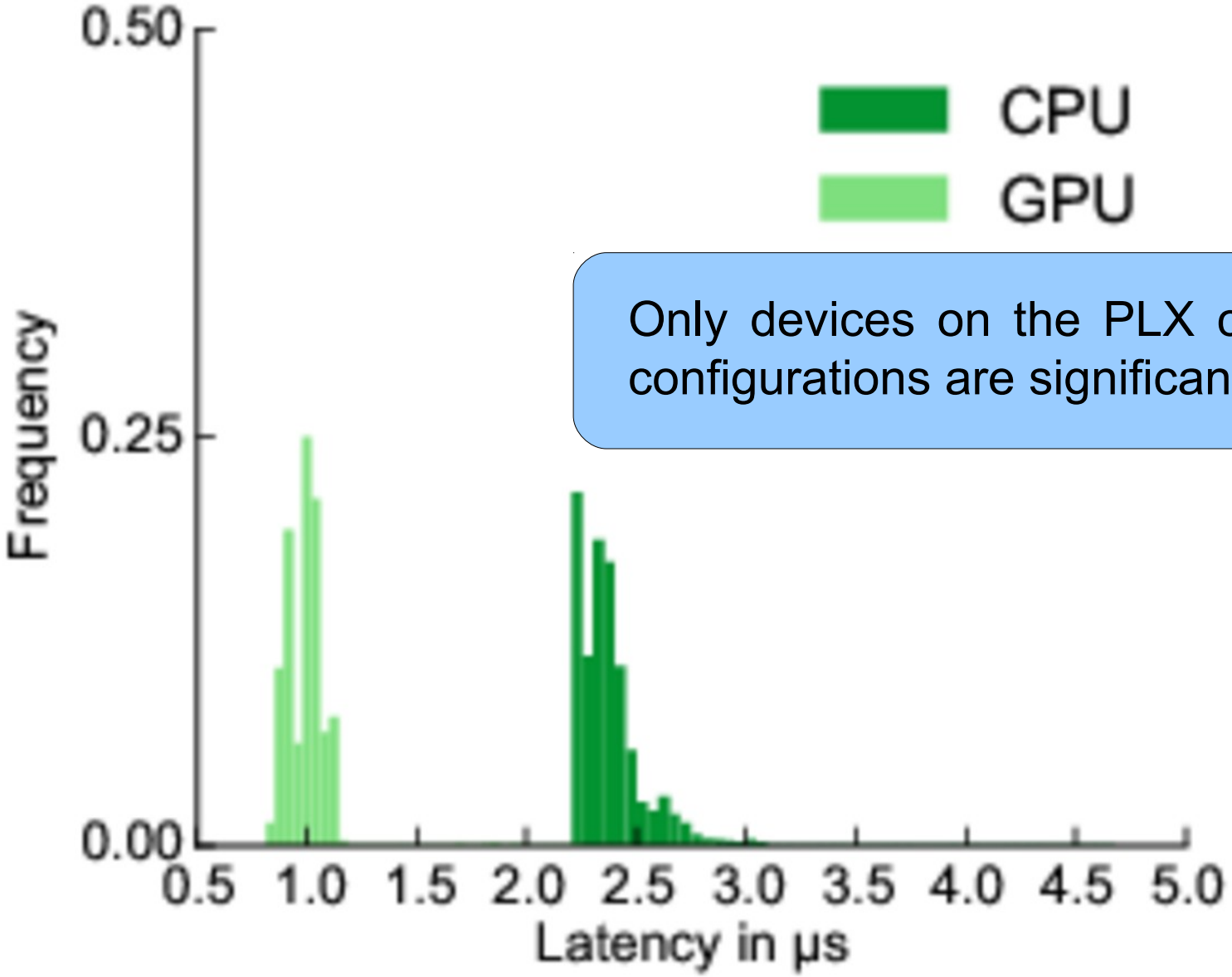
NVIDIA GPUDirect / AMD Direct GMA





Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

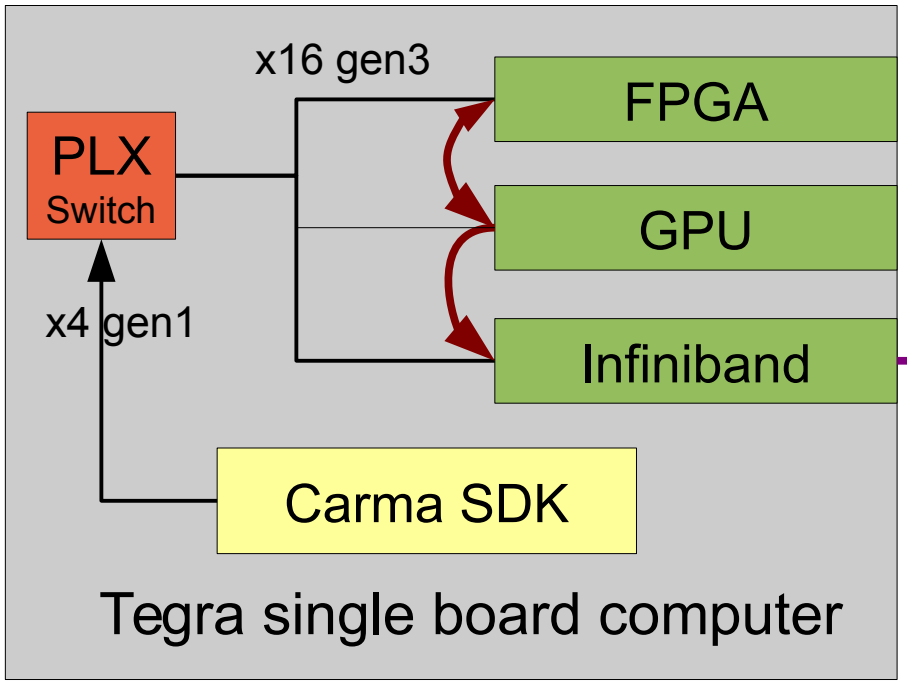
RDMA Latencv



Only devices on the PLX chip, other configurations are significantly slower

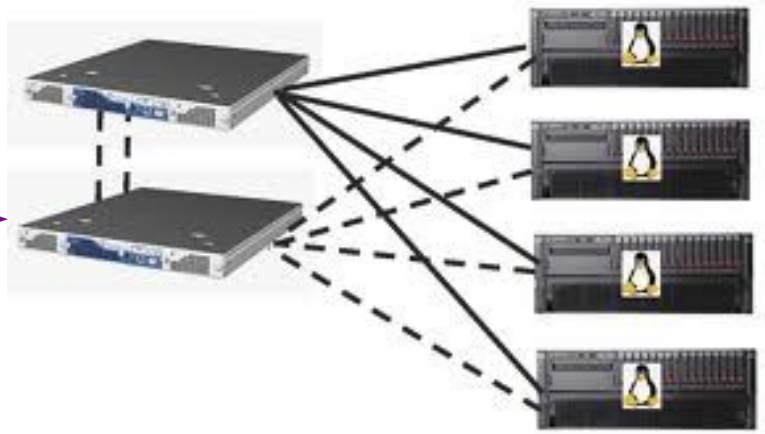
FPGA (PCIe x8 gen3) and AMD FirePro W9100

Integrating DAQ station and Electronics



DAQ board

Optical Infiniband
QDR 40 Gbps
up to 100m



InfiniBand DAQ cluster

Future: directly send the data to the GPU node processing this data.

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Master Server

DAQ Station

FDR Infiniband

56 Gbit/s

(or in small setup electronics directly)

Ethernet

10 Gb/s

Storage

LSDF
Large Scale Data Facility

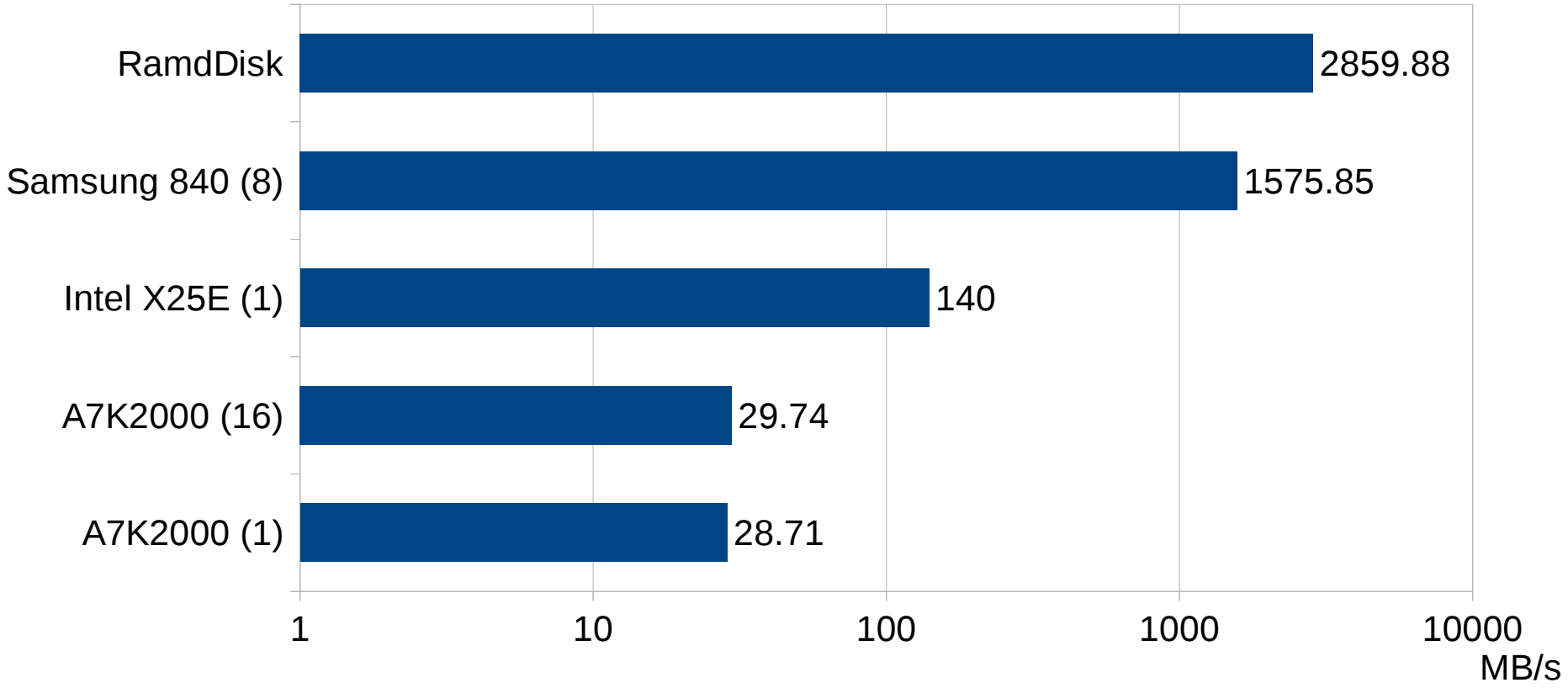
External PCIe x16 (16 GB/s)

SFF8088 (2.4 GB/s)

SuperMicro 7048GR-TRF (Intel C612 Chipset)
CPU: 2 x Xeon E5-2680v3 (total 24 cores at 2.5 Ghz)
GPUs: 7 x NVIDIA GTX Titan
Memory: 512GB DDR4
Network: Intel 82598EB (10 Gb/s)
Infiniband: 2 x Mellanox ConnectX-3 VPI
Storage: Areca ARC-1880-ix-12 SAS Raid
8 x Samsung 850 Pro 1TB (Raid0)
16 x Hitachi A7K200 (Raid6)

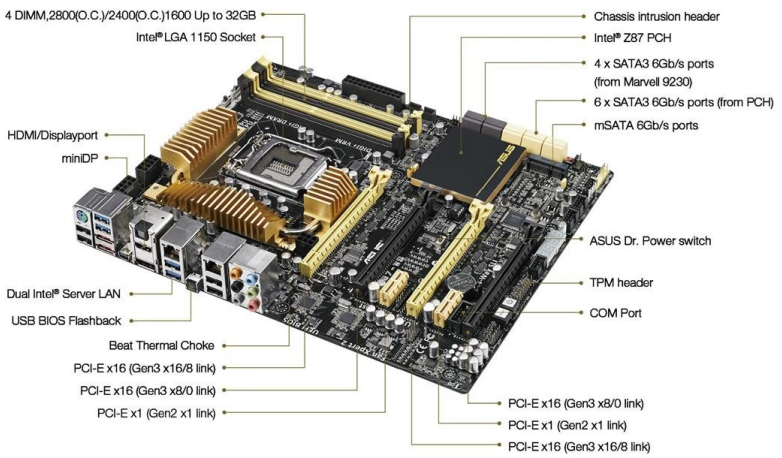
- High amount of memory
- Fast SSD-based Raid for overflow data
- Easy scalability with external PCI express and SAS

Caching large data sets



Using SSD drives may significantly increase random access performance to the data sets which are not fitting in memory completely. The big arrays of magnetic hard drives will not help unless multiple readers involved.

Cluster Node

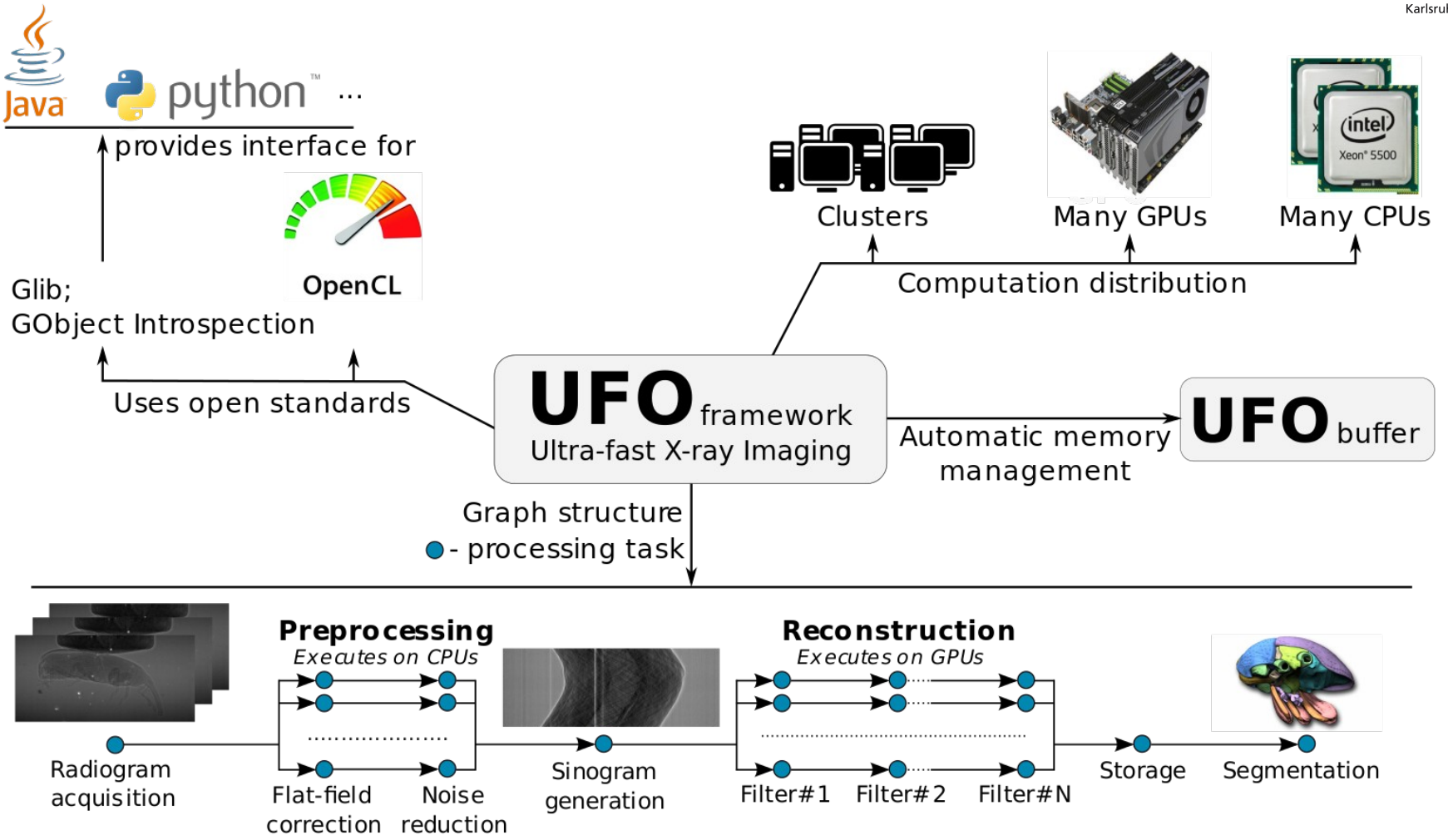


Asus Z87-WS (Intel Z87 PCH Chipset)
CPU: Core i5-4670 (total 4 cores at 3.4 Ghz)
GPUs: 3 x NVIDIA GTX Titan
Infiniband: Mellanox ConnectX-3 VPI
Memory: 16 GB (32GB max)

NVIDIA GTX Titan
Memory: 6 GB at 288 GB/s
Single-precision Gflops: 4500
Double-precision Gflops: 1500

- 4-Way SLI
- Low Price

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.

Storage Protocols

Network FS
NFS
Samba
SSHFS

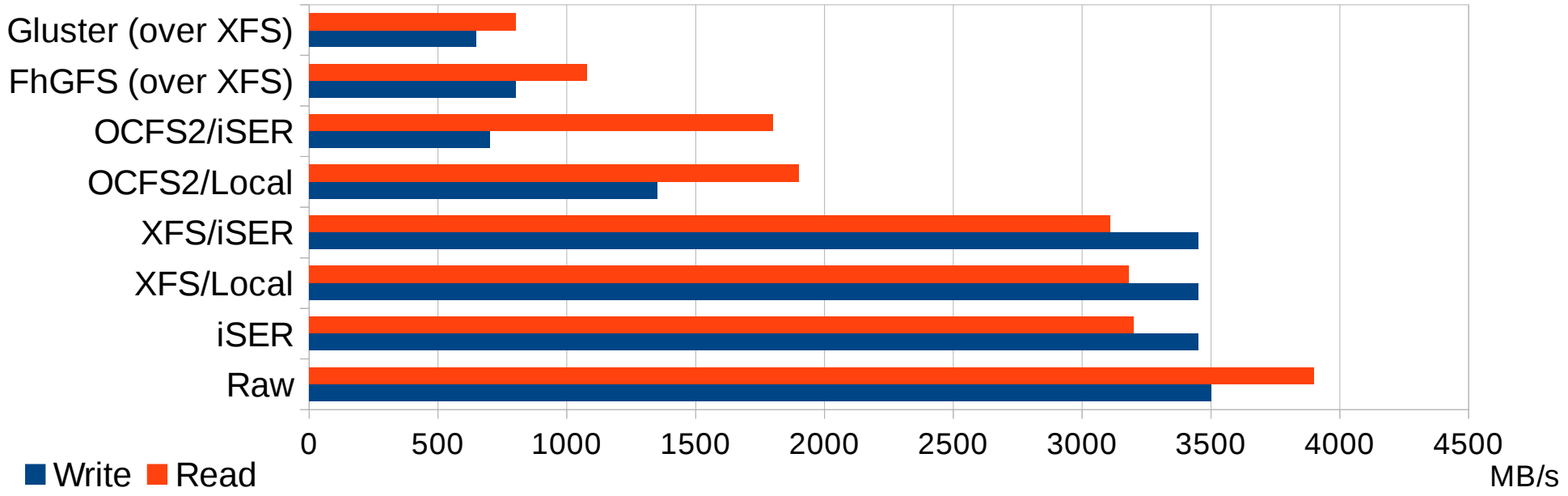
Slow

Cluster FS
Lustre (patched kernel)
Gluster
BeeGFS (close-sourced)

Slow if few nodes

Network Devices
iSCSI (slow)
iSER

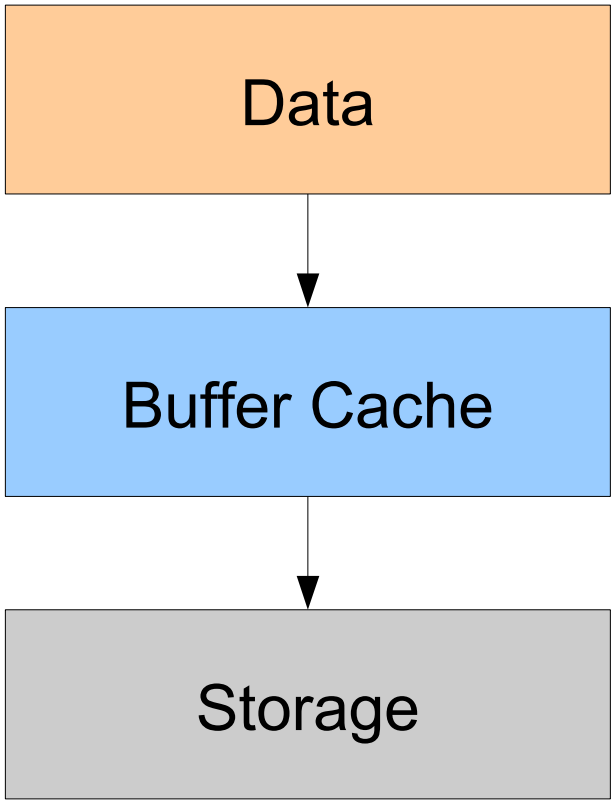
OCFS2



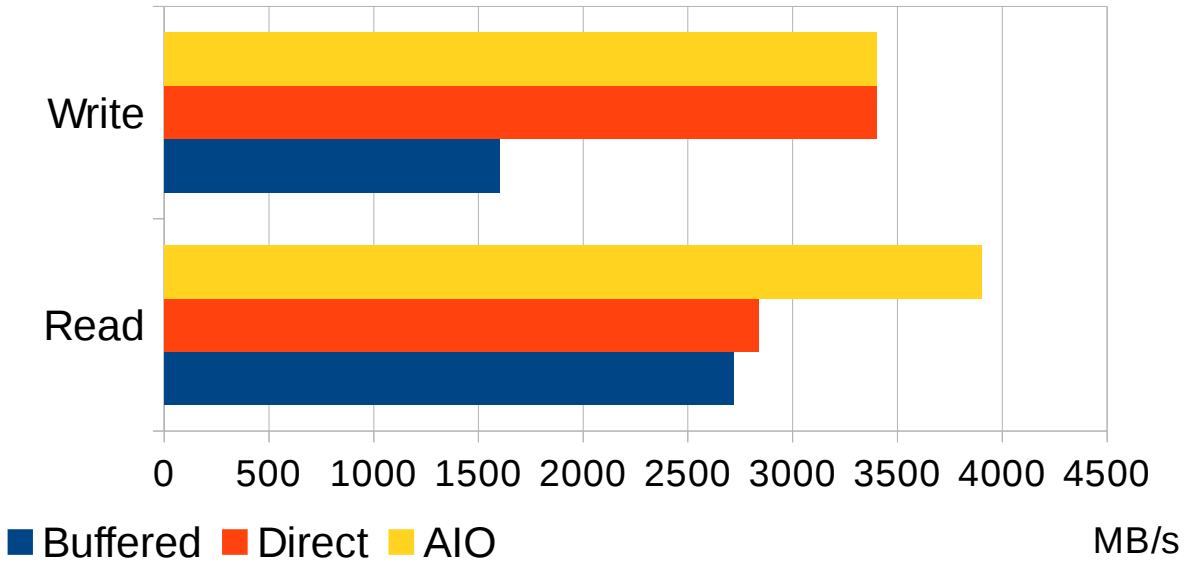
S. Chilingaryan et. all

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

High-speed Data Streaming



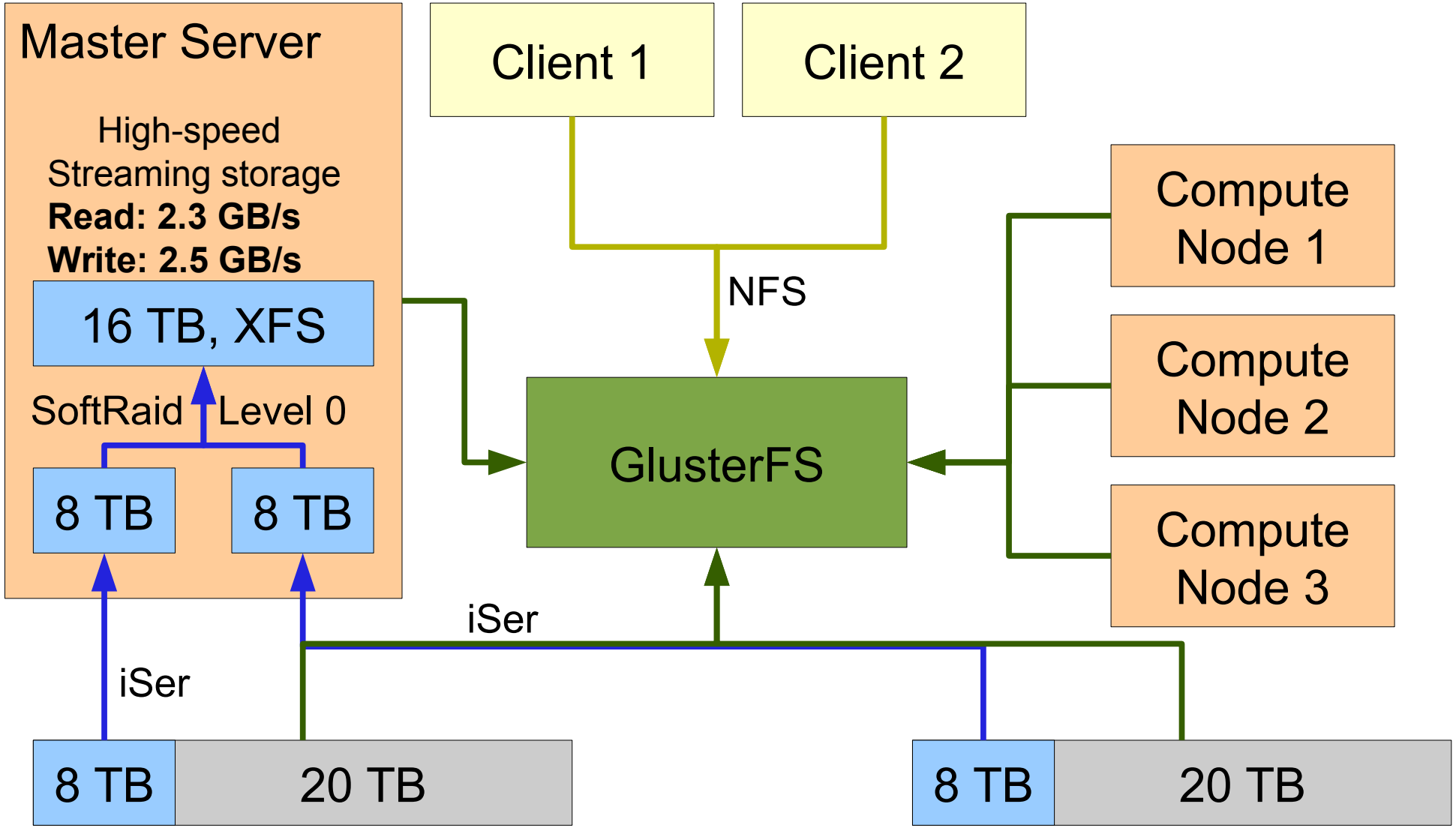
Default data flow in Linux



- Buffer cache significantly limits maximal write performance
- Kernel AIO may be used to program IO scheduler to issue read requests without delays

Optimizing I/O for maximum streaming performance using a single data source/receiver

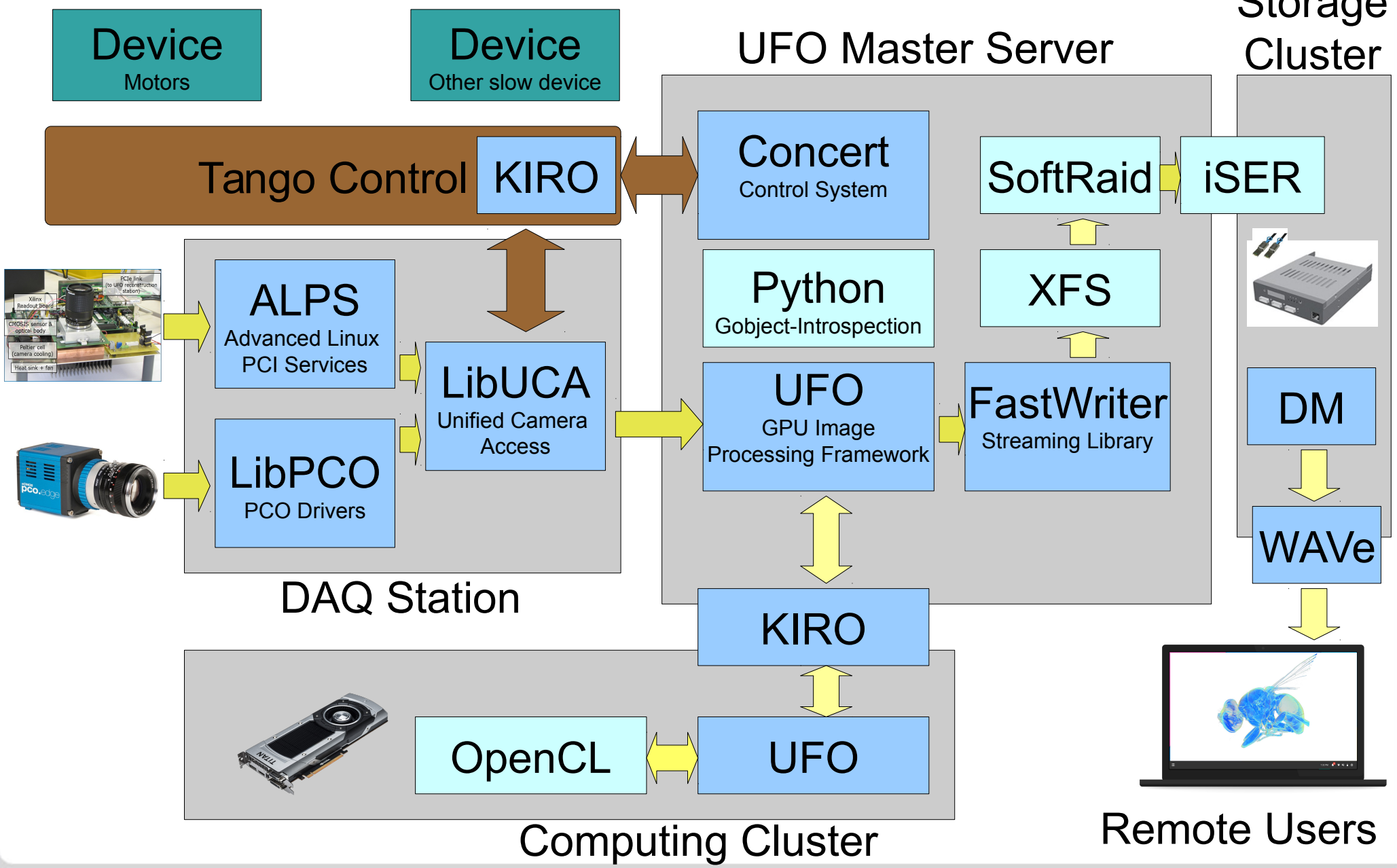
Storage Subsystem



Storage Node 1
Raid6: 16 Hitachi 7K300, 28TB

Storage Node 2
Raid6: 16 Hitachi 7K300, 28TB

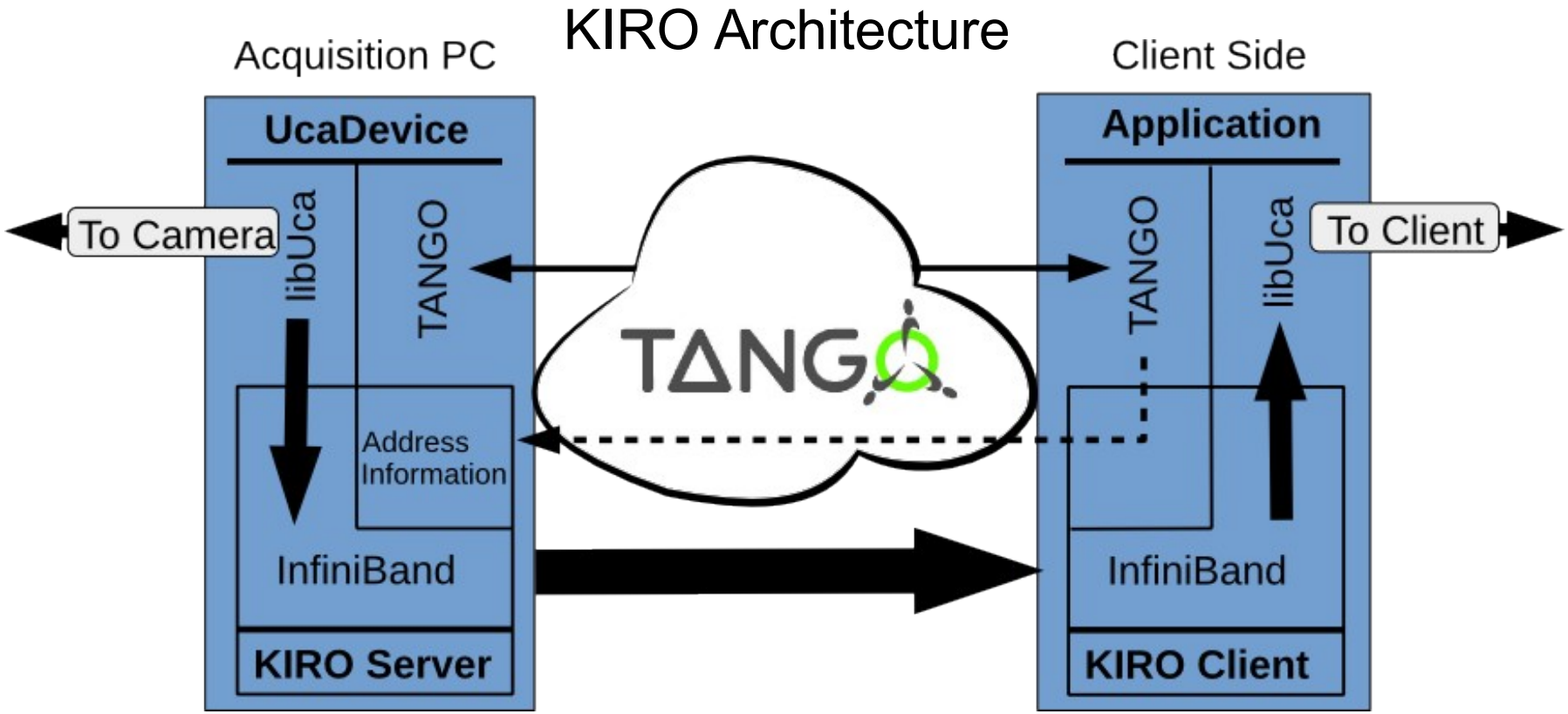
Software Stack

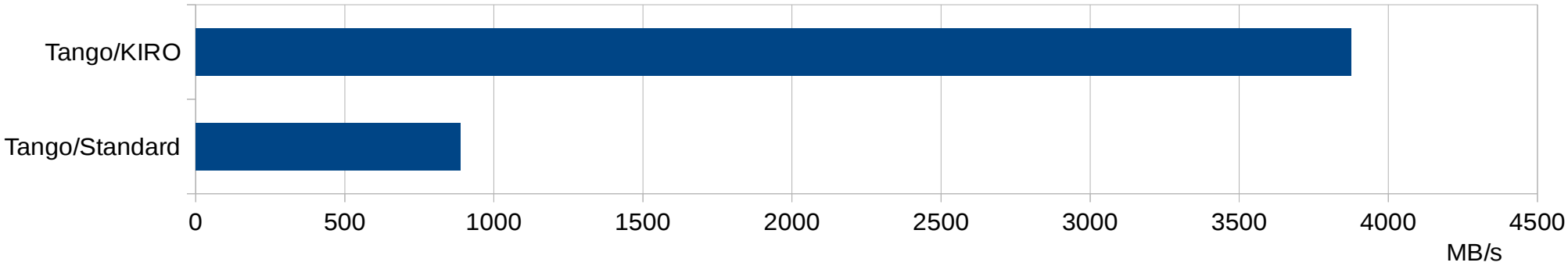


Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

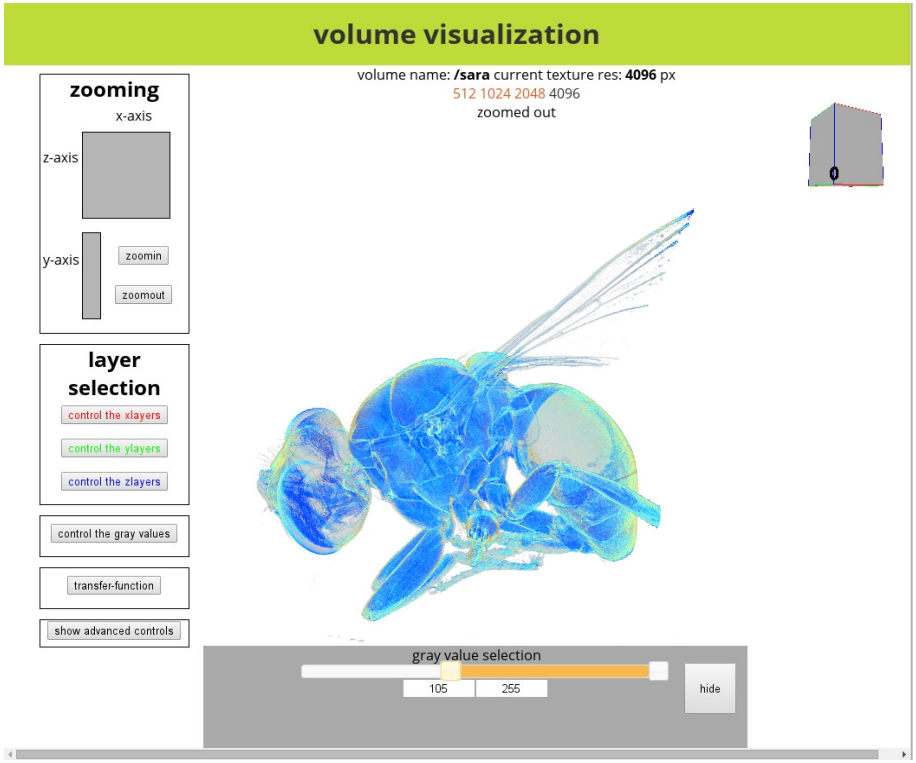
KIRO: High-speed RDMA library

Tango over Corba over TCP over Infiniband is slow



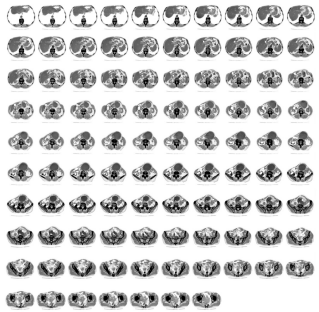


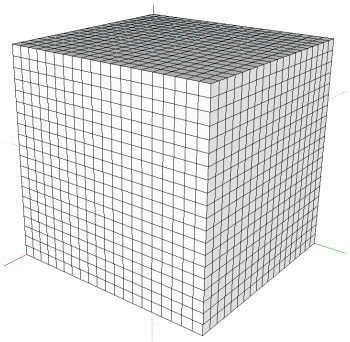
WAVe: Web-based volume visualization



Ray-casting approach

A preview slice-maps pre-generated using UFO framework





Optimized storage layout for fast zooming

- Working on majority mobile platforms with descent GPUs
- Multiple zooming levels for inspecting fine details
- High-quality cuts
- Automatic thresholding-based segmentation
- Multi-modality rendering support

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Summary

- **Alps (Advanced Linux PCI Services)**
  - Easy integration of new PCIe electronics
  - High-speed DMA engine with direct PCIe communication
  - Advanced scripting and debugging support
- **Scalable hardware platform for image-based control**
  - Only off-the-shelf components are used
  - Easily scalable from single PC to the GPU cluster
  - Distributed over large area using optical Infiniband Links
- **Fully-pipelined parallel image-processing framework**
  - Easily extensible library of algorithms
  - Tuning for various parallel architectures
  - Good scalability using native Infiniband transport
- **Storage, visualization, and remote data analysis**
  - Reliable storage for data streaming at rates up to 4 GB/s
  - High quality web-based visualization of large volumes
  - Virtualization environment for remote image segmentation