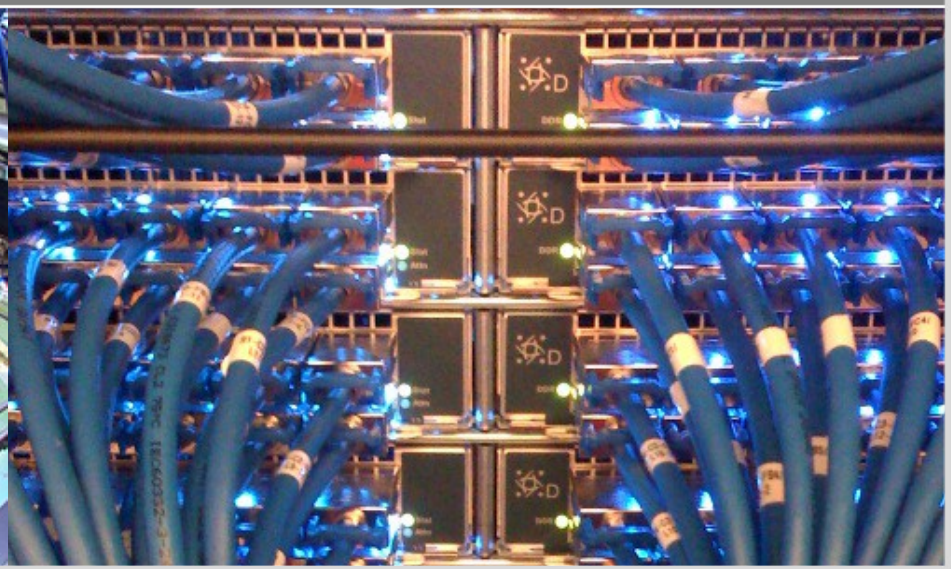
Karlsruher Institut für Technologie

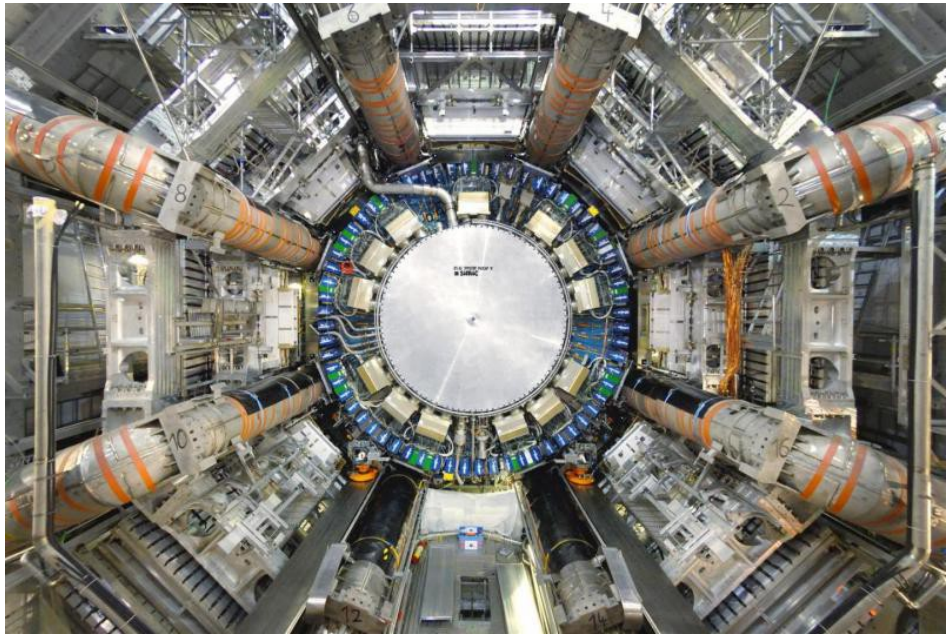
High-speed interconnects for DAQ applications

Timo Dritschler

IPE, Institute for data processing and electronics

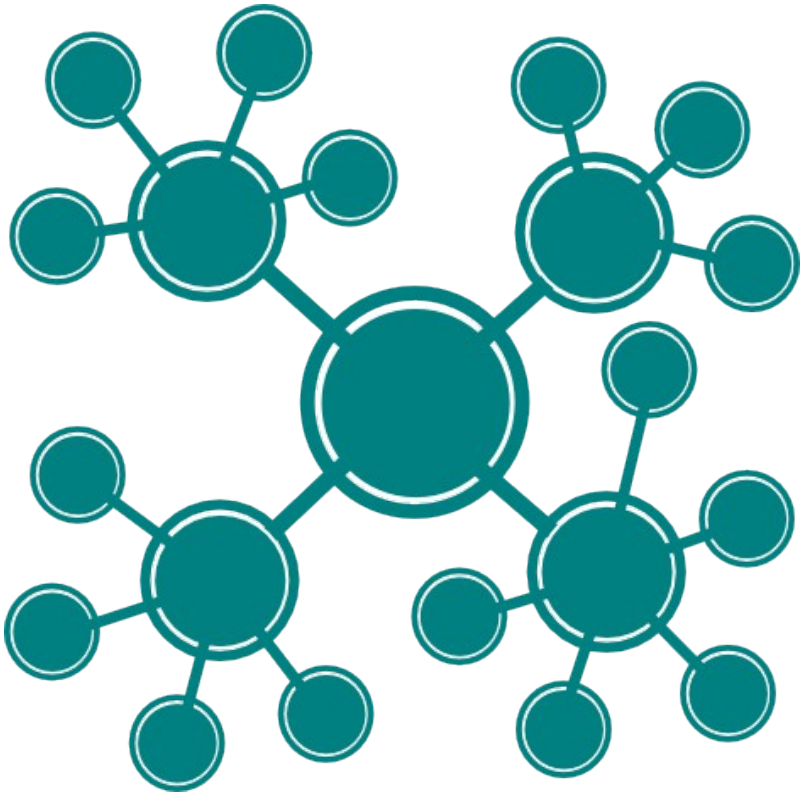


The challenge of modern large experiments



© CERN, the European Organization for Nuclear Research

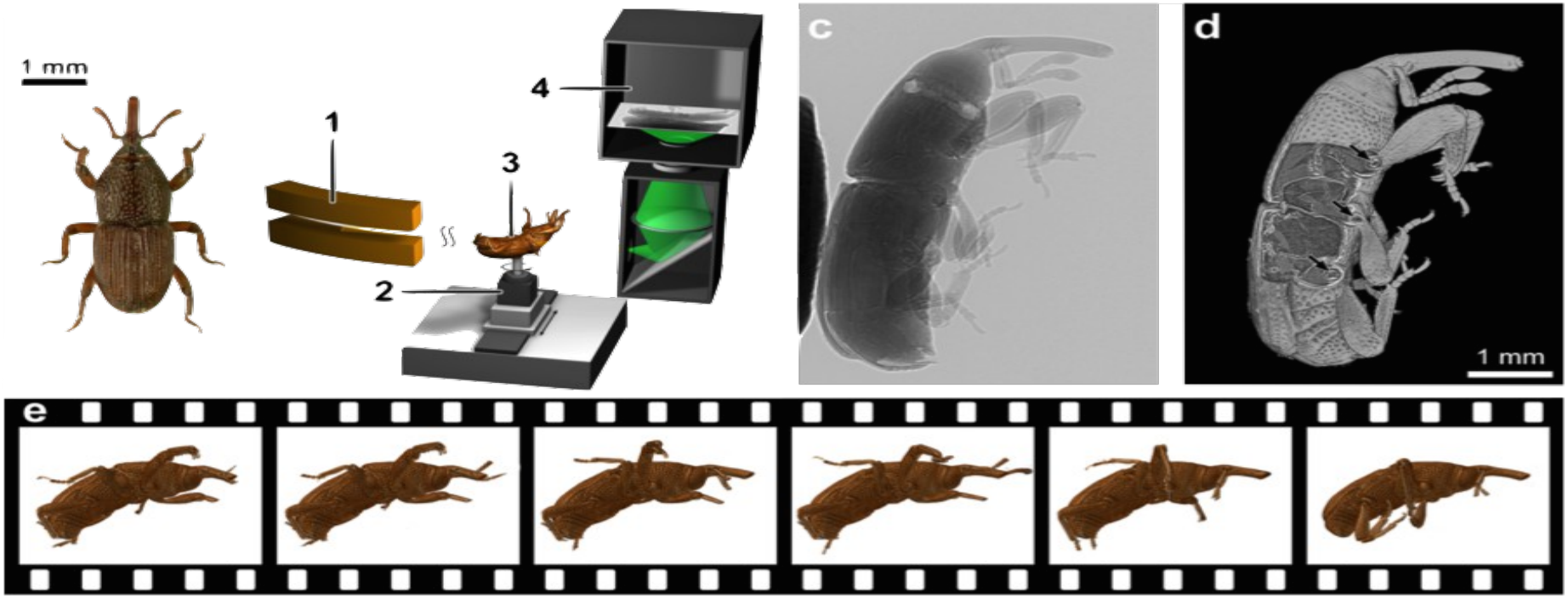
"The detector generates unmanageably large amounts of raw data: about 25 megabytes per event [...], multiplied by 40 million beam crossings per second in the center of the detector. This produces a total of **1 petabyte of raw data per second.**" - "ATLAS Experiment", Wikipedia, 2015



- Data of large experiments can no longer be handled by one single machine
- Cooperation of multiple machines is necessary
- Each machine needs to be provided with data
- Communication overhead grows with number of machines and size of dataset

Networks and interconnects between machines are of huge importance for modern high-performance computing!

KIT UFO Project as an example

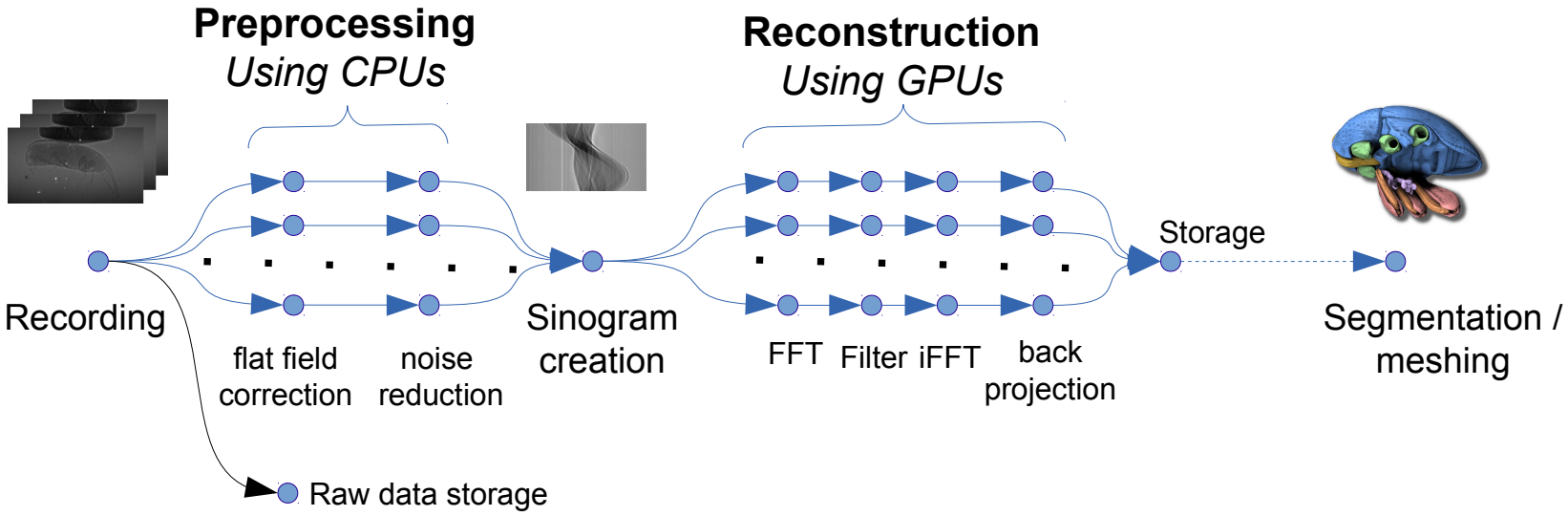


dos Santos Rolo, Tomy, et al. "In vivo X-ray cine-tomography for tracking morphological dynamics." Proceedings of the National Academy of Sciences111.11 (2014): 3921-3926.

Ultra-**F**ast X-ray Imaging of Scientific Processes with **O**n-line Assessment and Data-driven Process Control

- Fast cameras at approx 5000FPS at 1MP
- Streaming interface with **50Gbit/s** bandwidth
- Soft real-time reconstruction and evaluation of recorded data using GPU computing

KIT UFO Framework



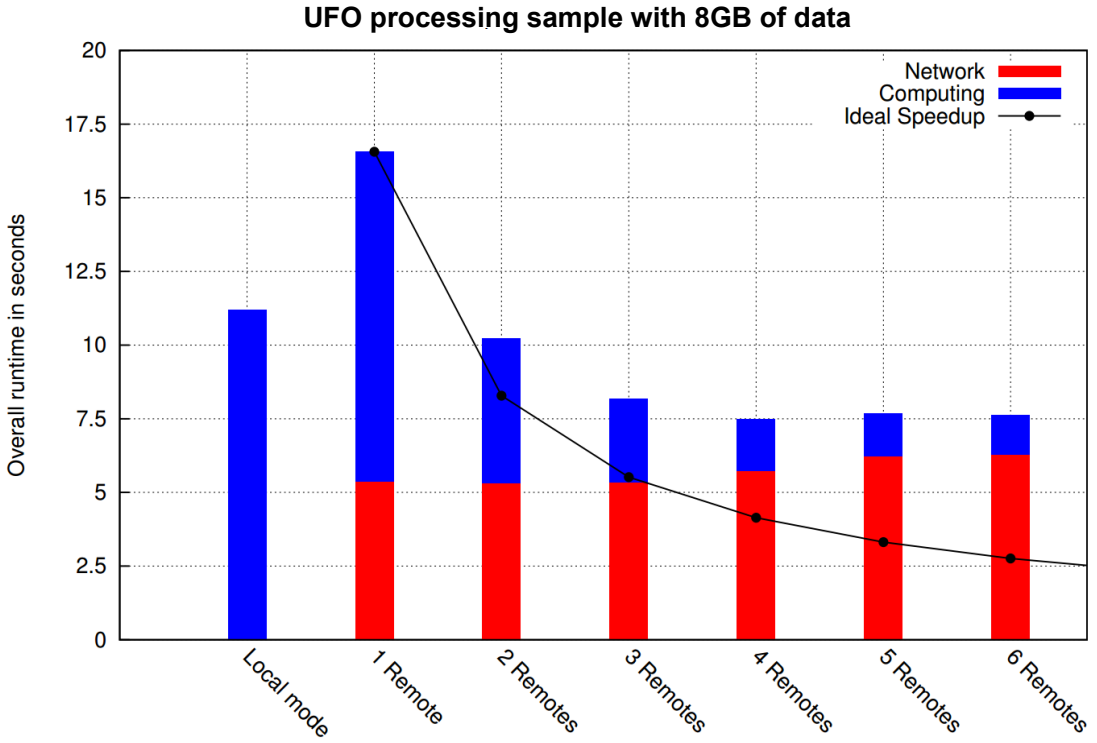
KIT UFO Framework

- Distributed computing framework
- Plug-in based and extensible algorithms
- GPU-enabled
- Automatically distributes work across the network
- Simple markup-file-based configuration

Available on Github:
https://github.com/ufo-kit

KIT UFO framework scalability

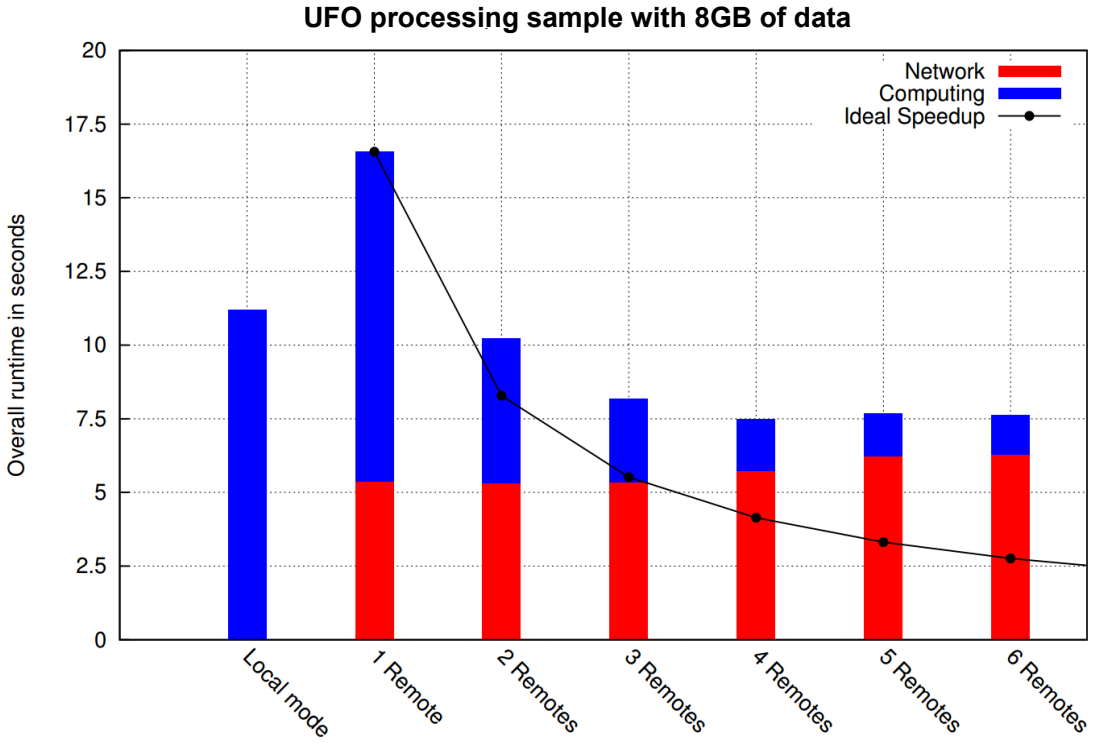
- Early test with MPI based communication
- Computing part scaled nicely
- However, network communication did not scale



Graphic by Timo Dörr, *'Concepts and evaluation of communication patterns for digital image processing in heterogeneous distributed systems'*, 2014

KIT UFO framework scalability

- Early test with MPI based communication
- Computing part scaled nicely
- However, network communication did not scale



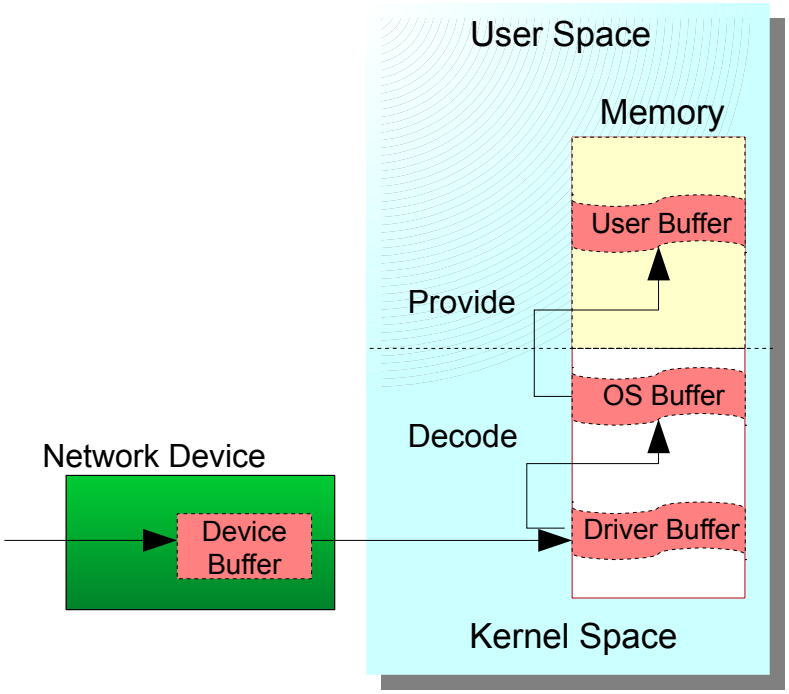
Graphic by Timo Dörr, *'Concepts and evaluation of communication patterns for digital image processing in heterogeneous distributed systems'*, 2014

Throughput and efficiency of the network is crucial!

How do we increase the efficiency of our network?

Key Technology RDMA

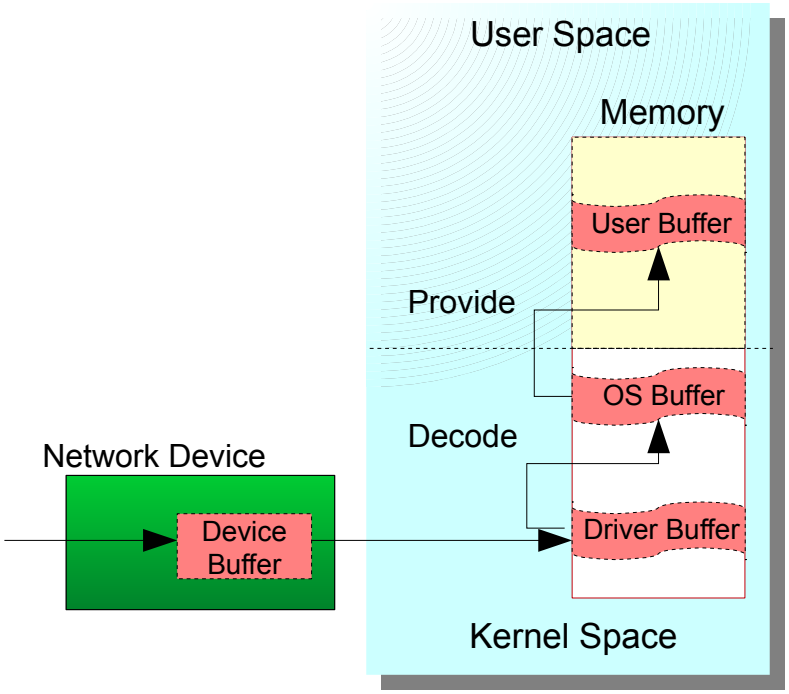
Classical network transfer



- Up to four copies
- High latency

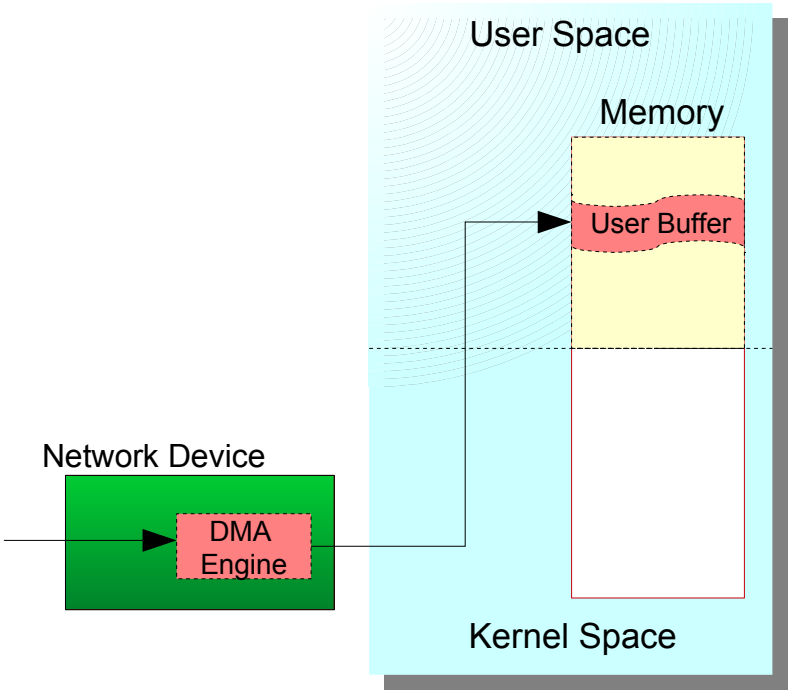
Key Technology RDMA

Classical network transfer



- Up to four copies
- High latency

RDMA (Remote Direct Memory Access)



- Only one implicit 'copy' from DMA engine
- Low latency

RDMA capable interconnects

Feature	InfiniBand (FDR)	Ethernet (ROCE)	PCIe (Gen. 3)
Nominal bandwidth	**14Gbit/s per x1** (Max x12 at 164Gbit/s)	**10/40/100 Gbit/s**	**~0.8Gbit/s per x1** (Max x16 at 126Gbits/s)
Nominal Latency	0.7µs	1.3µs	~ 1µs *
Max cable length	5m (Copper) 300m (Optical)	30m (Copper) 40km (Optical)	7m (Performs best on cables < 1m)
Next Generation	EDR at **25Gbit/s** per x1 (End of this Year?)	**400Gbit** (Expected 2017)	Gen. 4 at **~1.96GBit/s** per x1 (Expected past 2014)

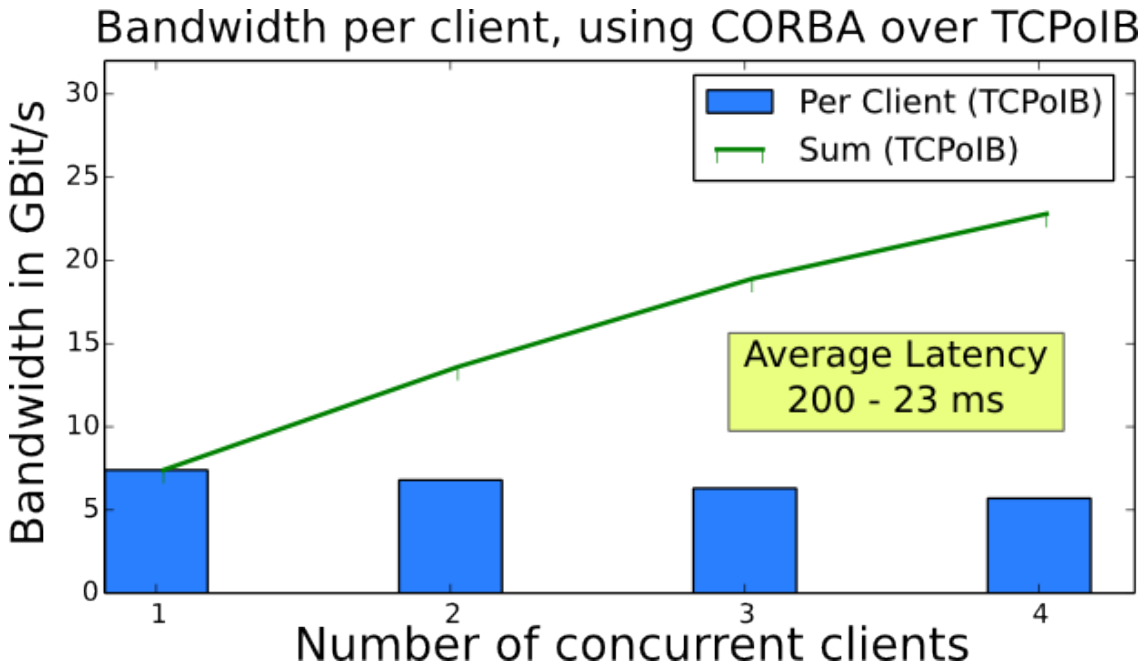
*Technically, PCIe (since it is a 'local' bus system) has not latencies other than connection negotiation

RDMA capable interconnects

Feature	InfiniBand (FDR)	Ethernet (ROCE)	PCIe (Gen. 3)
Nominal bandwidth	**14Gbit/s per x1** (Max x12 at 164Gbit/s)	**10/40/100 Gbit/s**	**~0.8Gbit/s per x1** (Max x16 at 126Gbits/s)
Nominal Latency	0.7µs	1.3µs	~ 1µs *
Max cable length	5m (Copper) 300m (Optical)	30m (Copper) 40km (Optical)	7m (Performs best on cables < 1m)
Next Generation	EDR at **25Gbit/s** per x1 (End of this Year?)	**400Gbit** (Expected 2017)	Gen. 4 at **~1.96GBit/s** per x1 (Expected past 2014)

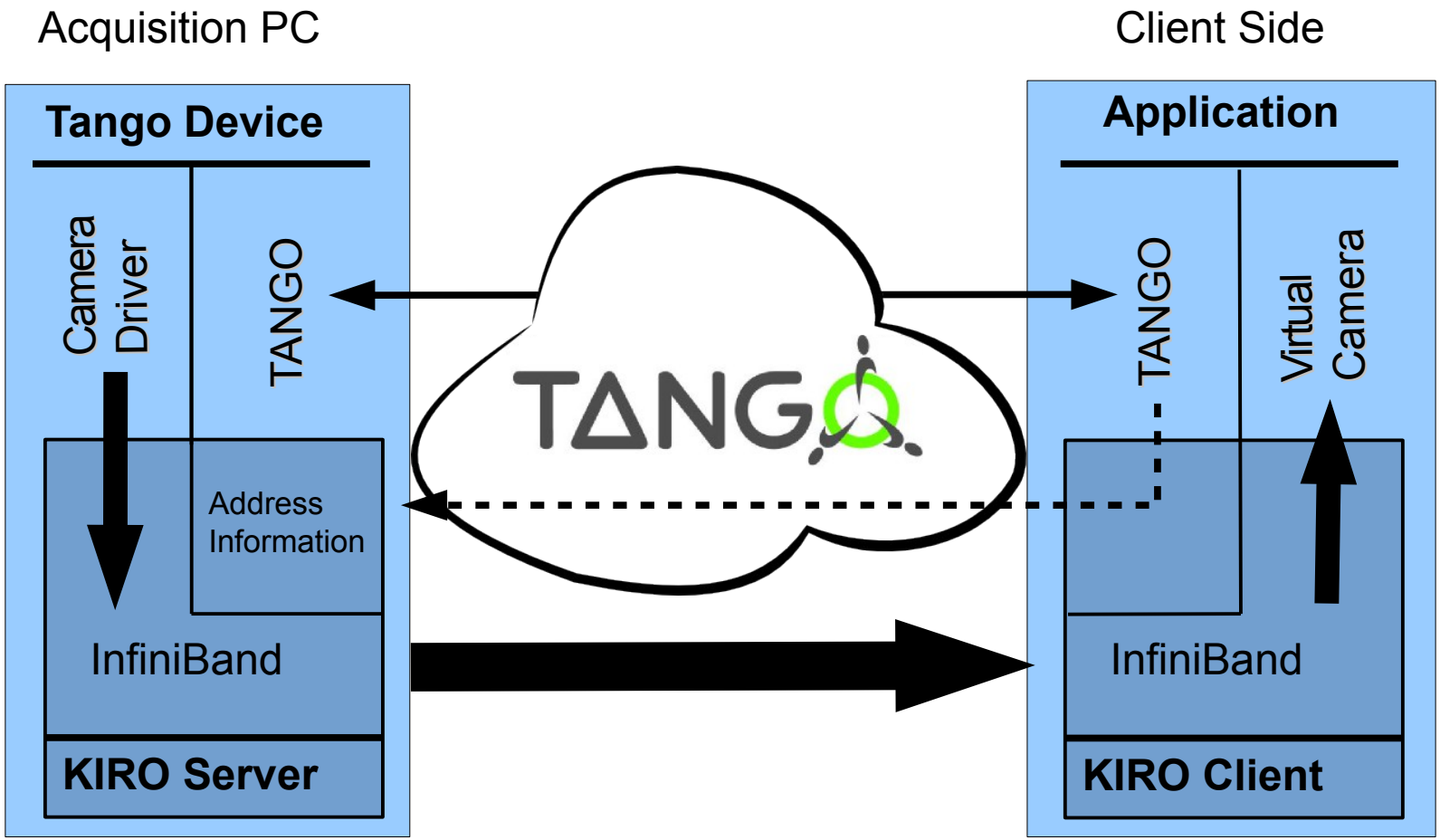
*Technically, PCIe (since it is a 'local' bus system) has not latencies other than connection negotiation

A usecase example for InfiniBand

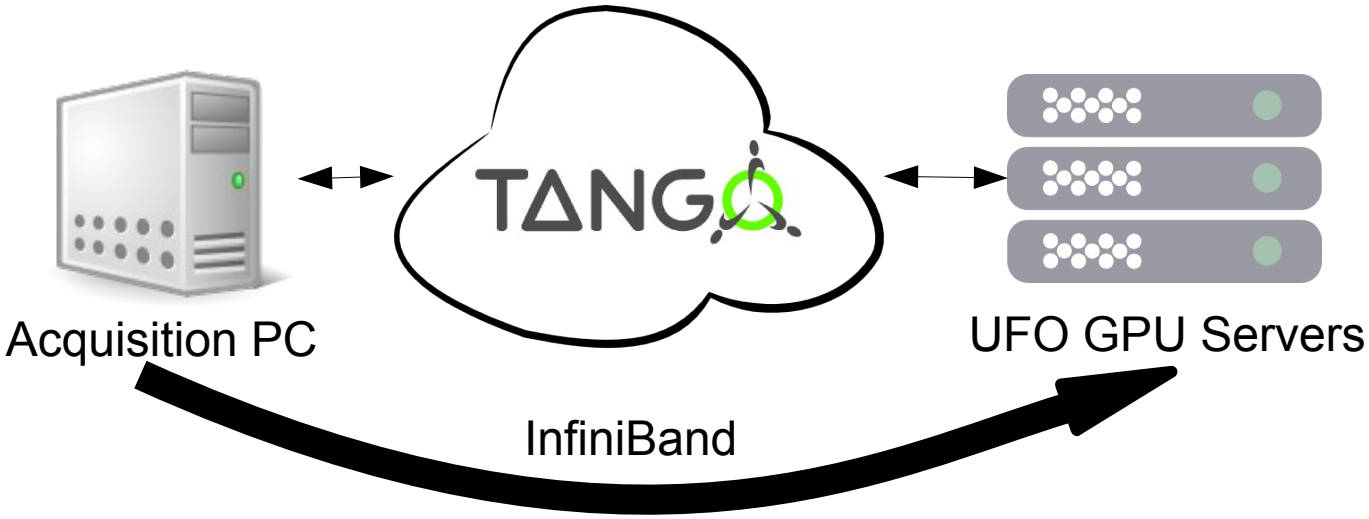


(Max. 32Gb/s)

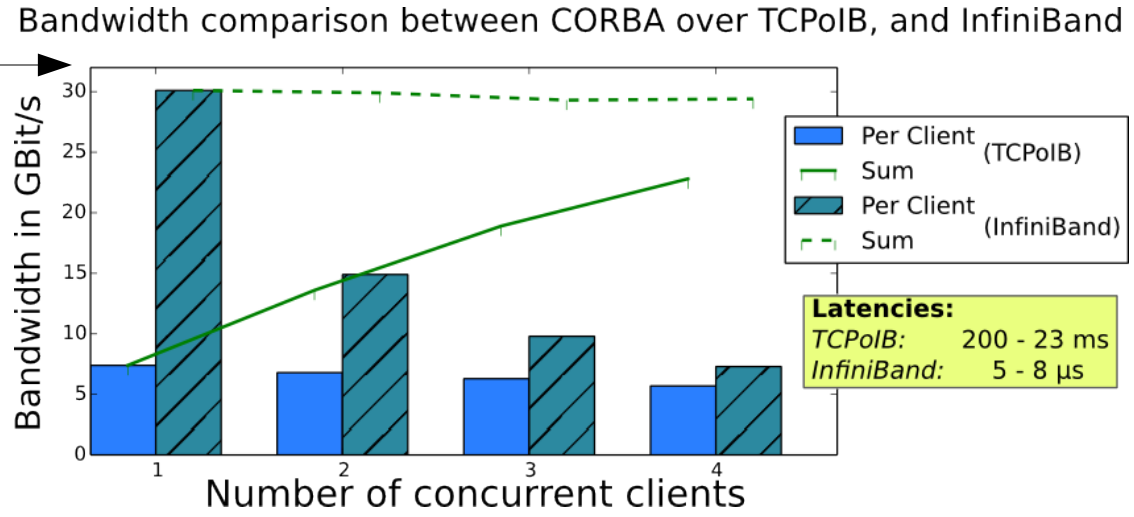
Software Architecture using KIRO



An usecase example for InfiniBand



(Max. 32Gb/s)



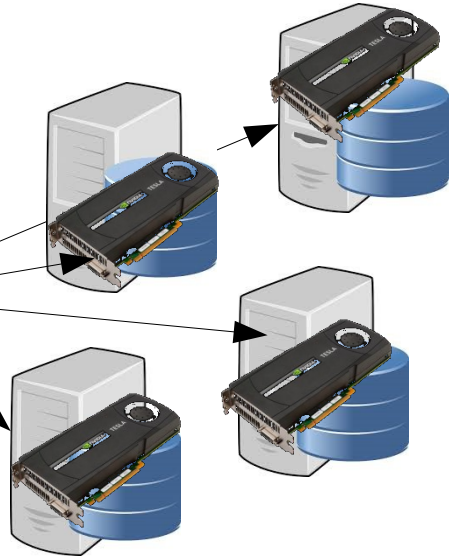
Next Step: UFO & KIRO

KIT UFO Framework

UFO Camera

InfiniBand switch

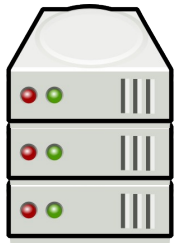
KIRO



GPU-cluster connected by InfiniBand

InfiniBand for fast storage

iSER
(InfiniBand SCSI Extension for RDMA)

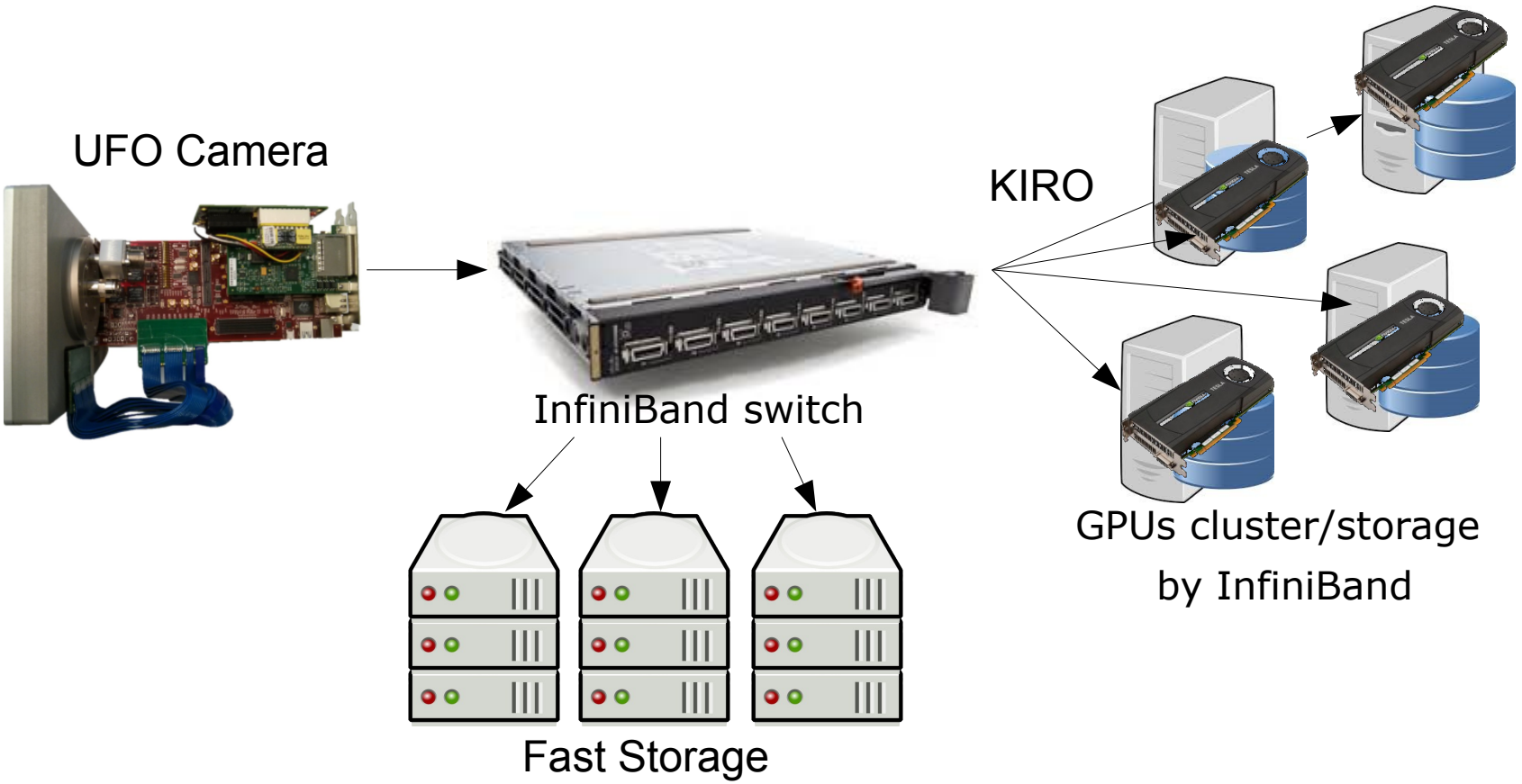


Fast Storage

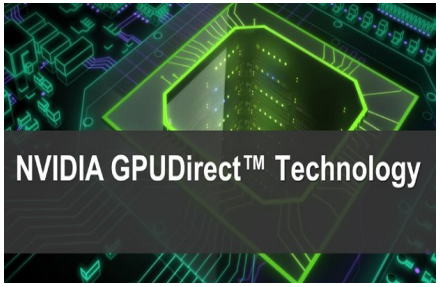
Graphics Courtesy of Suren Chilingaryan

Fast storage via InfiniBand

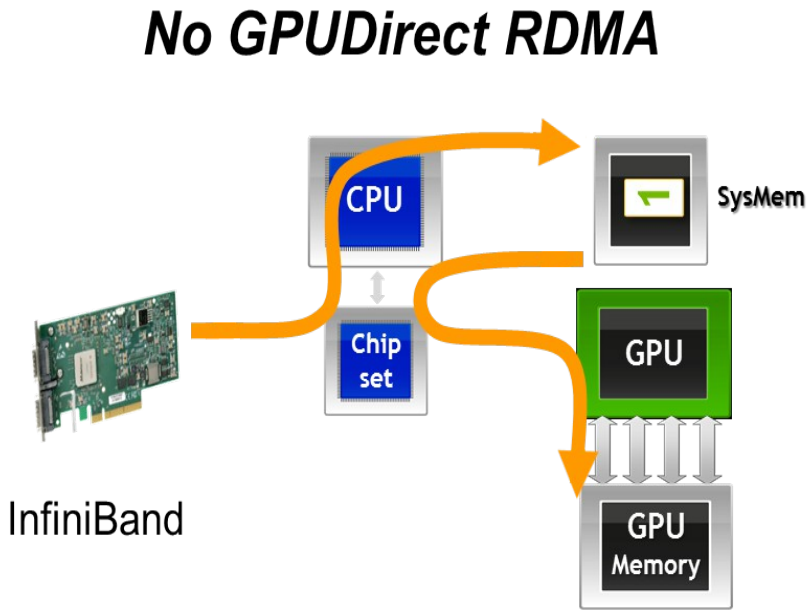
KIT UFO Framework



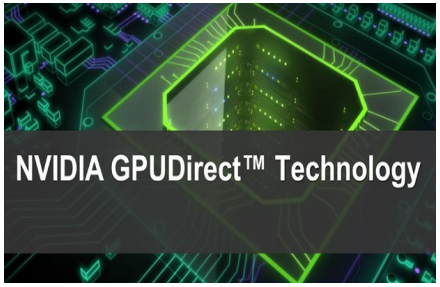
Key technology GPUDirect



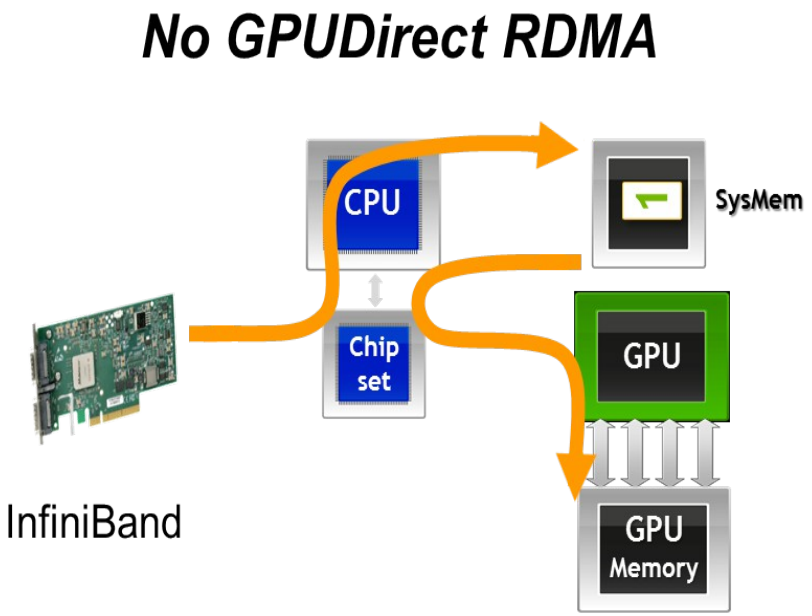
When doing conventional RDMA transfer that is meant for GPU computation, at least two copy operations are required.

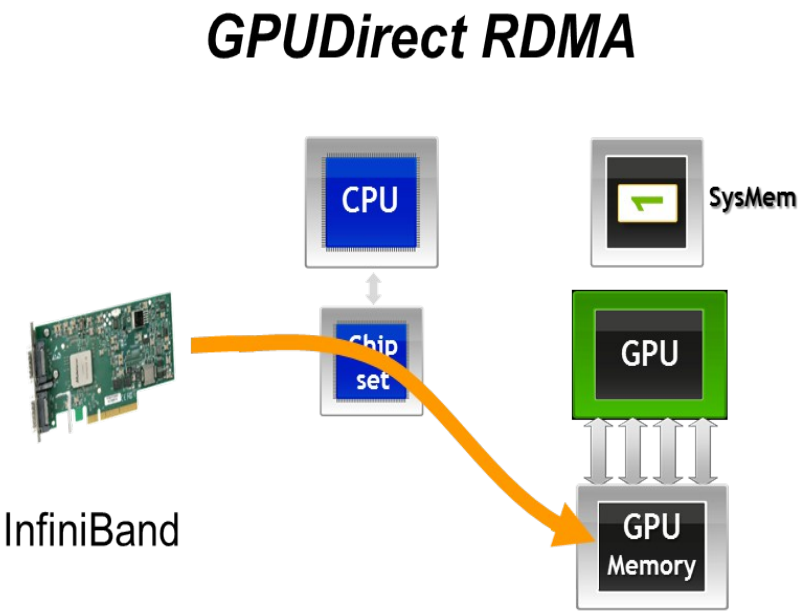


Key technology GPUDirect

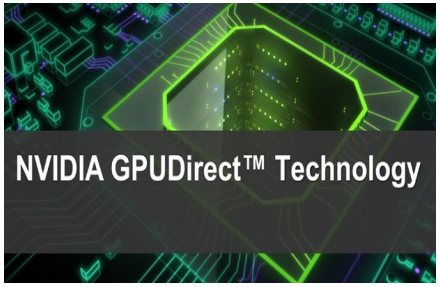


With GPUDirect technology, the GPU can become the target of RDMA operations and save valuable memory bandwidth.

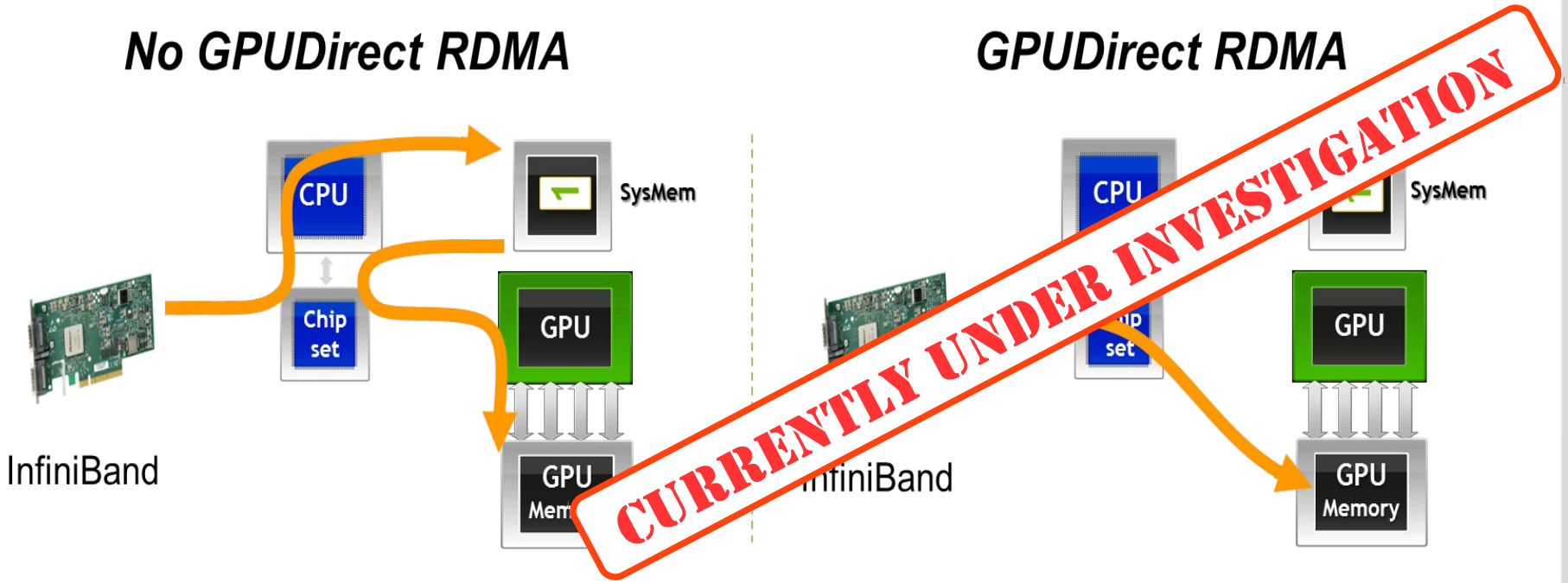




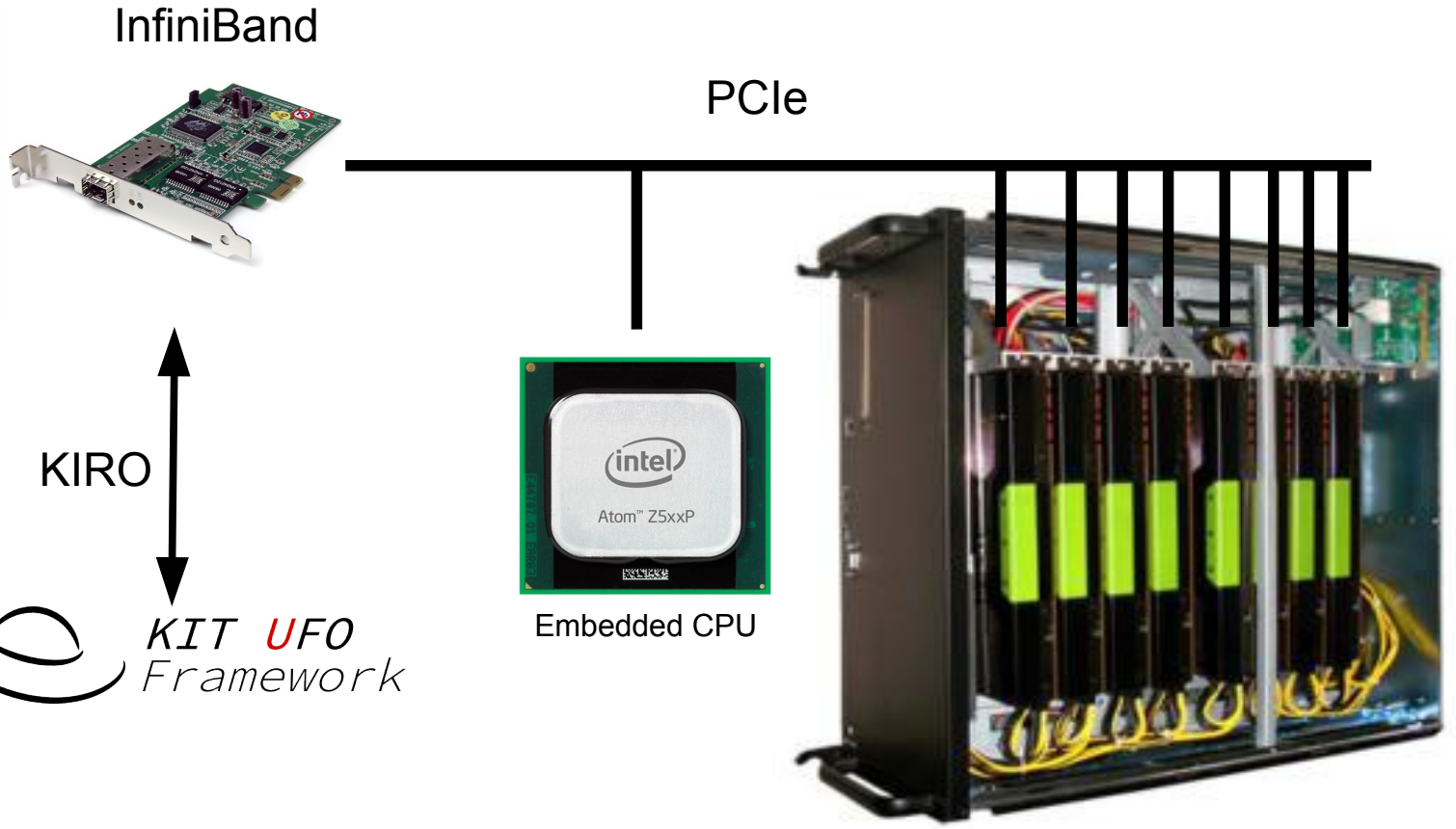
Key technology GPUDirect



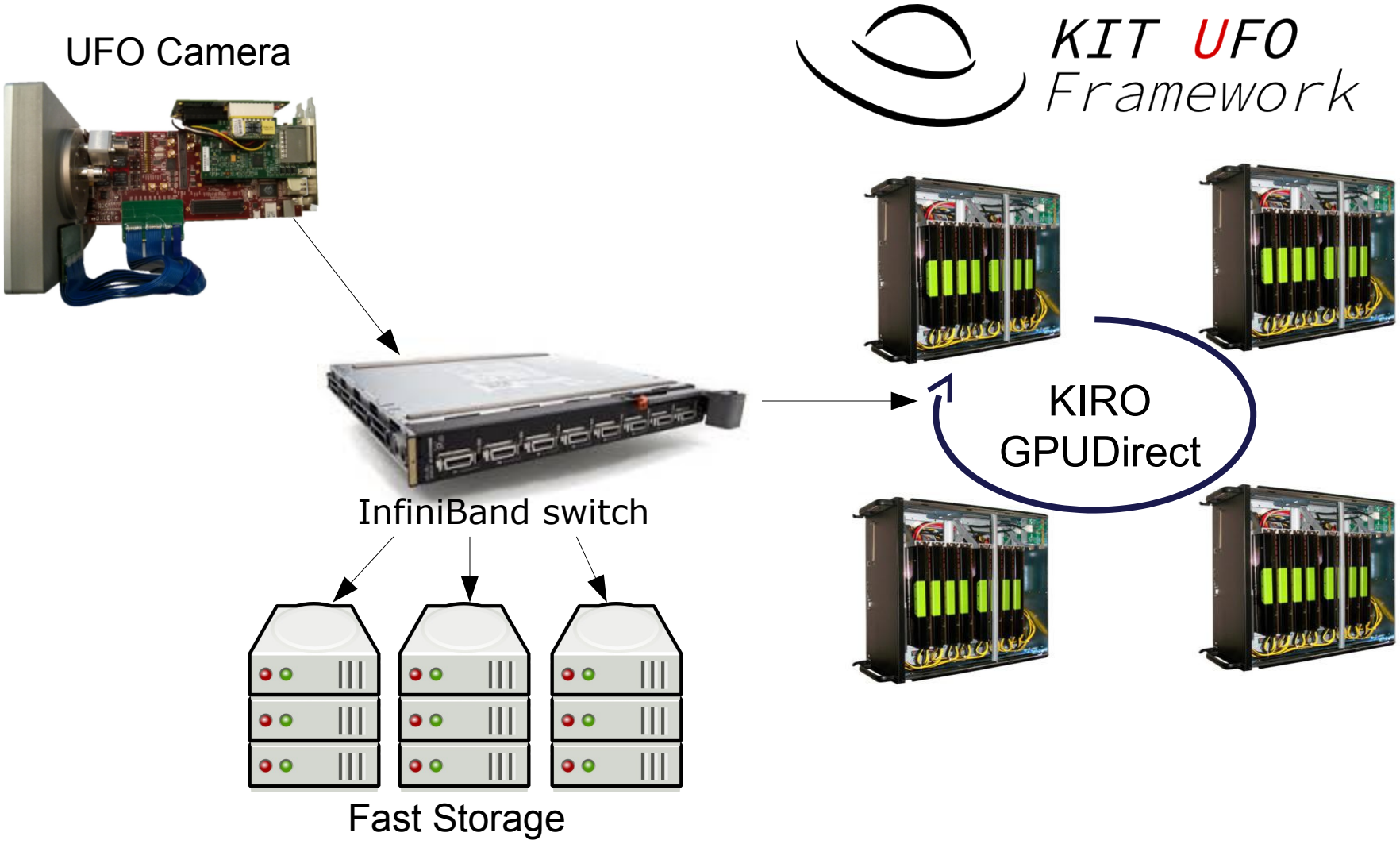
With GPUDirect technology, the GPU can become the target of RDMA operations and save valuable memory bandwidth.



Putting everything together



The future


UFO Camera
KIT UFO Framework
KIRO
GPUDirect
InfiniBand switch
Fast Storage

Karlsruher Institut für Technologie

Thank you for your attention!