Karlsruhe Institute of Technology

S. Chilingaryan, A. Kopmann, M. Vogelgesang

–

Cluster Architecture
for Development in IPE

--

GPU Server

Interconnects

Realization

Offline Reconstruction Station

SuperMicro 7046GT-TRF (Dual Intel 5520 Chipset)
CPU: 2 x Xeon E5540 (total 8 cores at 2.53 Ghz)
GPUs: 2 x GTX 580 + 2 x GTX295 External
Memory: 96 GB / 12 DDR3 slots (192GB max)

PCIe 2 x16 (8 GB/s)

PCIe 2 x16 (8 GB/s)

2 x GTX 580
(Fermi Architecture)

PCIe 2 x16 (8 GB/s):

PCIe 2 x16 (8 GB/s):

2 x GTX295
(Dual GT100)

PCIe 2 x4 (2 GB/s):

PCIe 2 x4 (2 GB/s):

PCIe 1 x4 (1 GB/s):

PyHST

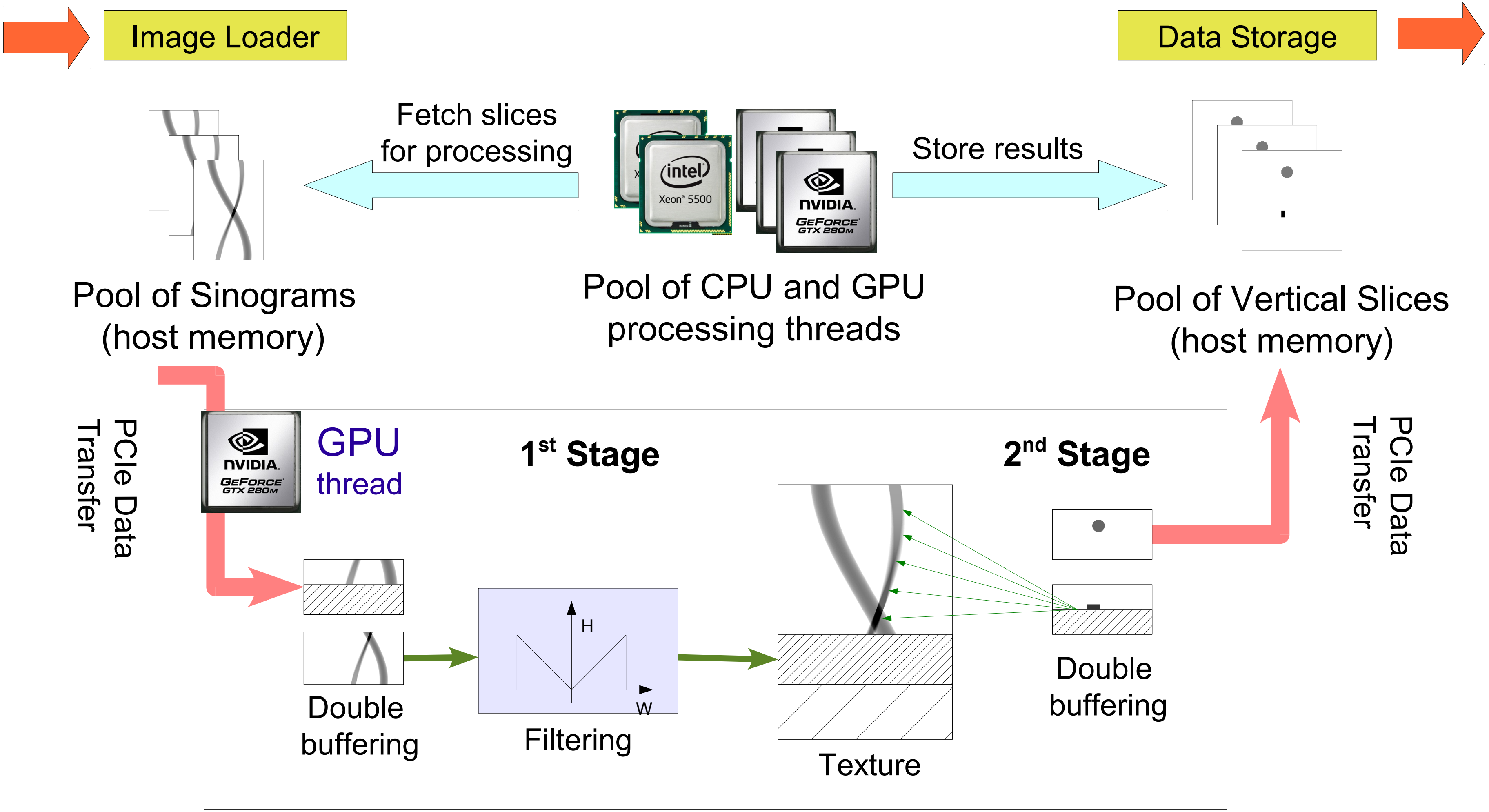


Image Loader

Data Storage

Fetch slices for processing

Store results

Pool of Sinograms (host memory)

Pool of CPU and GPU processing threads

Pool of Vertical Slices (host memory)

PCIe Data Transfer

GPU thread

1st Stage

2nd Stage

Double buffering

Filtering

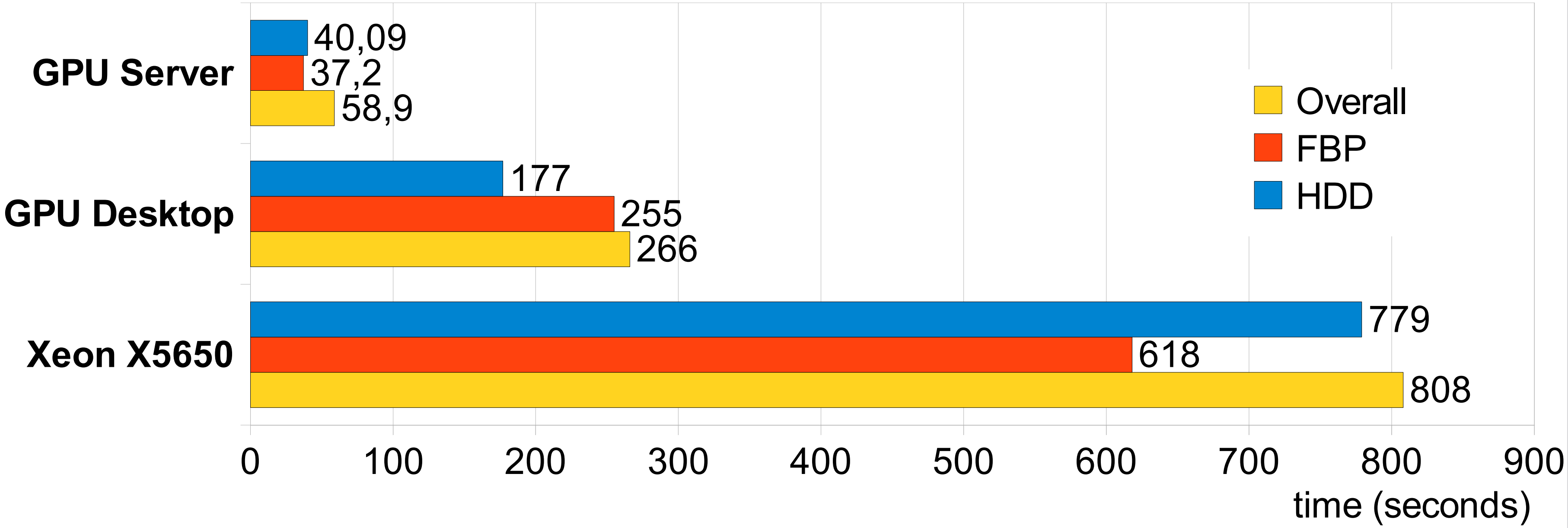
Texture

Double buffering

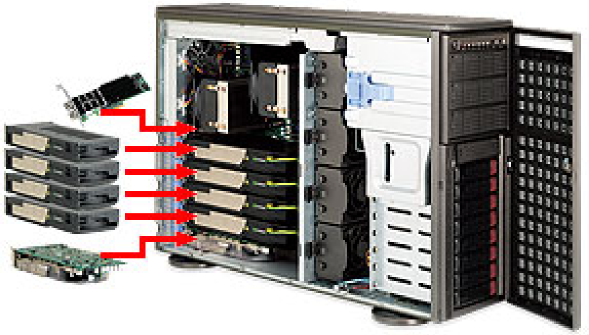
PCIe Data Transfer

Performance: GPU vs. CPU

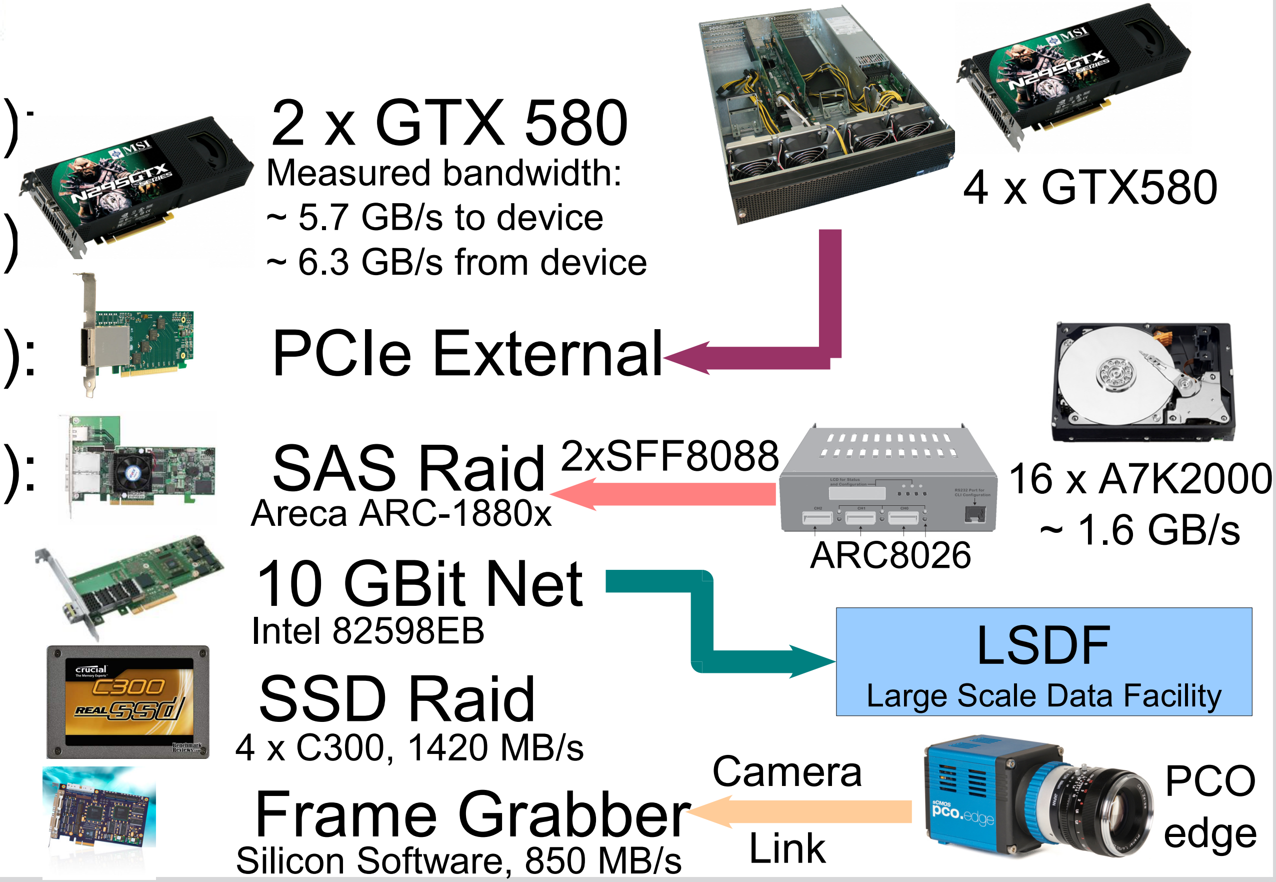
	Xeon Server	GPU Desktop	GPU Server
Type of Computation	CPU / Xeon X5650 12 cores, 2.66 GHz	**GeForce GTX 280** **1 core**	**2 x GTX295 + 2 x GTX580** **6 cores**
CPU	2 x Xeon E5650	Core2 E6300	2 x Xeon E5540
Memory	16GB DDR3	4GB DDR2	96GB DDR3
HDD/SSD	Hitachi A7K2000	**2 x Intel X25-E**	**4 x Crucial RealSSD C300**
Price	5500$ (2000$ CPUs)	1500$ (400$ GPU)	9000$ (2000$ GPU, 1200$ SSD)
Software	SuSe 11.3, CUDA 3.2, MKL 10.2.1, gcc4.5 -O3 -march=nocona -mfpmath=sse		



Scalable Real-Time Station



SuperMicro 7046GT-TRF (Dual Intel 5520 Chipset)
CPU: 2 x Xeon X5650 (total 12 cores at 2.66 Ghz)
GPUs: 2 x GTX 580 + 4 x GTX580 External
Memory: 96 GB / 12 DDR3 slots (192GB max)



PCIe 2 x16 (8 GB/s)
PCIe 2 x16 (8 GB/s)
2 x GTX 580
Measured bandwidth:
~ 5.7 GB/s to device
~ 6.3 GB/s from device

4 x GTX580

PCIe 2 x16 (8 GB/s): PCIe External

PCIe 2 x16 (8 GB/s): SAS Raid
Areca ARC-1880x

2xSFF8088

ARC8026

16 x A7K2000
~ 1.6 GB/s

PCIe 2 x4 (2 GB/s): 10 GBit Net
Intel 82598EB

LSDF
Large Scale Data Facility

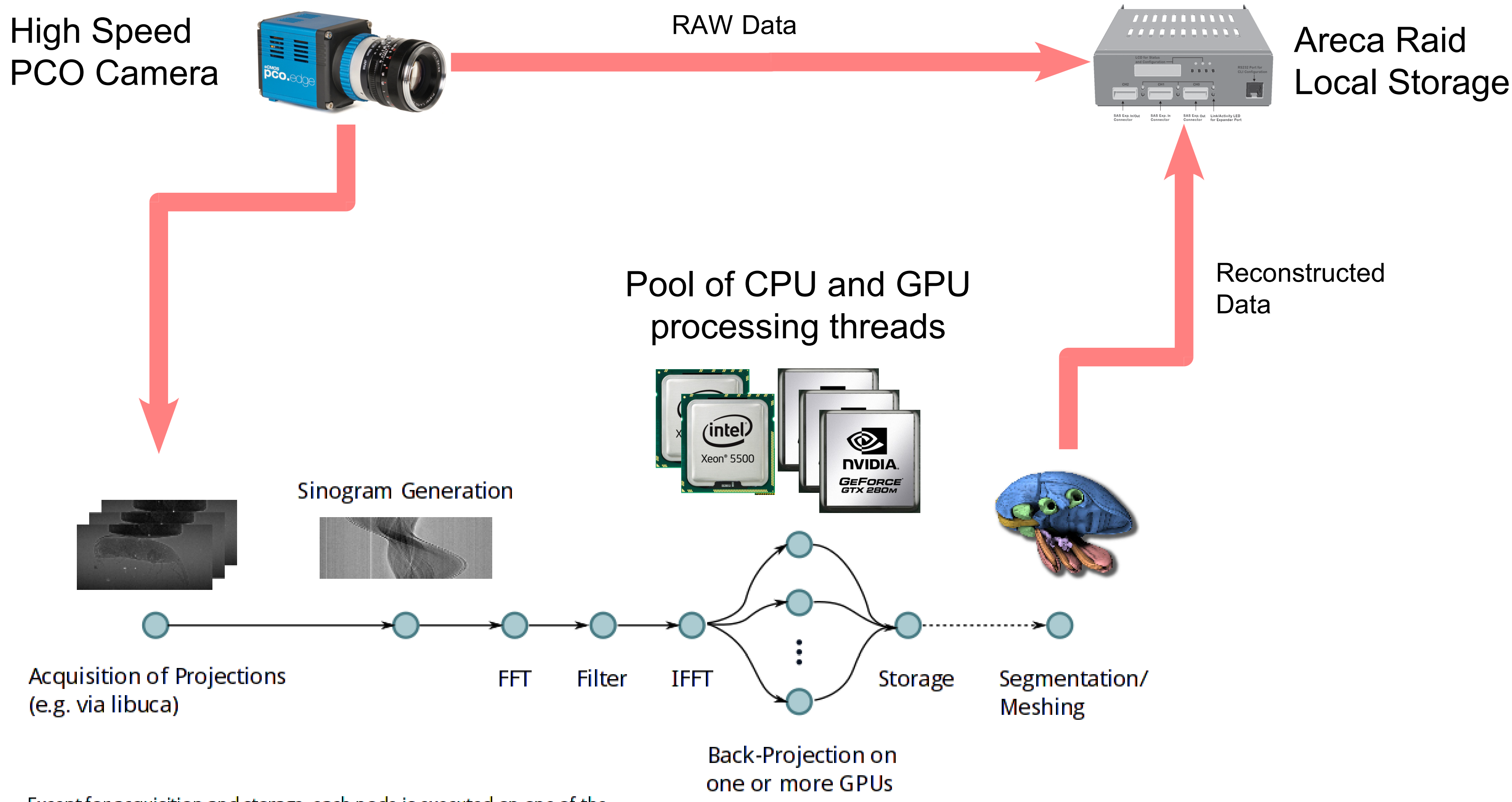
PCIe 2 x4 (2 GB/s): SSD Raid
4 x C300, 1420 MB/s

PCIe 1 x4 (1 GB/s): Frame Grabber
Silicon Software, 850 MB/s

Camera Link

PCO edge

UFO Framework



High Speed PCO Camera

RAW Data

Areca Raid Local Storage

Pool of CPU and GPU processing threads

Reconstructed Data

Sinogram Generation

Acquisition of Projections (e.g. via libuca)

FFT

Filter

IFFT

Storage

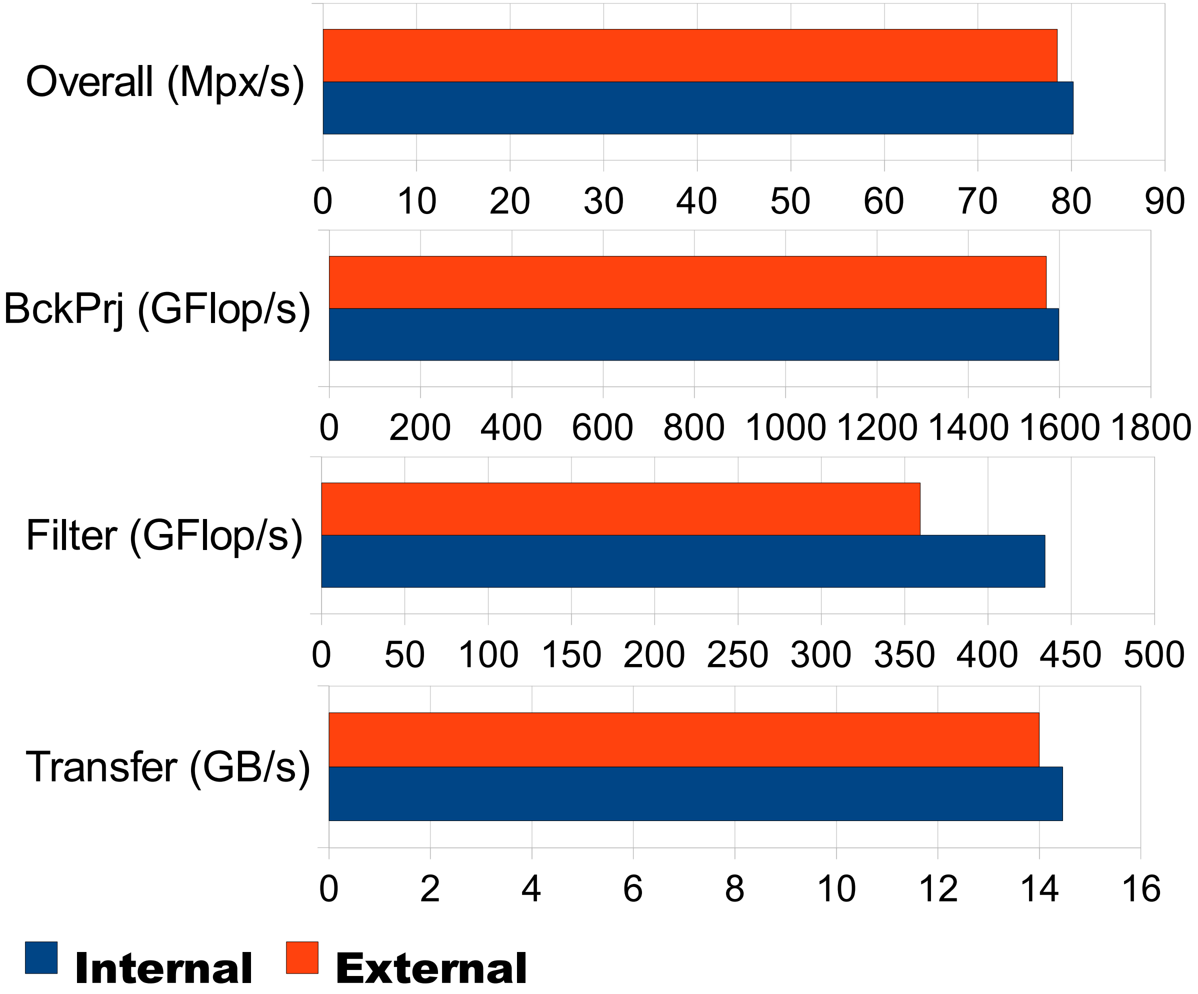
Segmentation/ Meshing

Back-Projection on one or more GPUs

Except for acquisition and storage, each node is executed on one of the available GPUs according to a heuristic.

External GPU Box

Can we breach 12 GPU barrier?



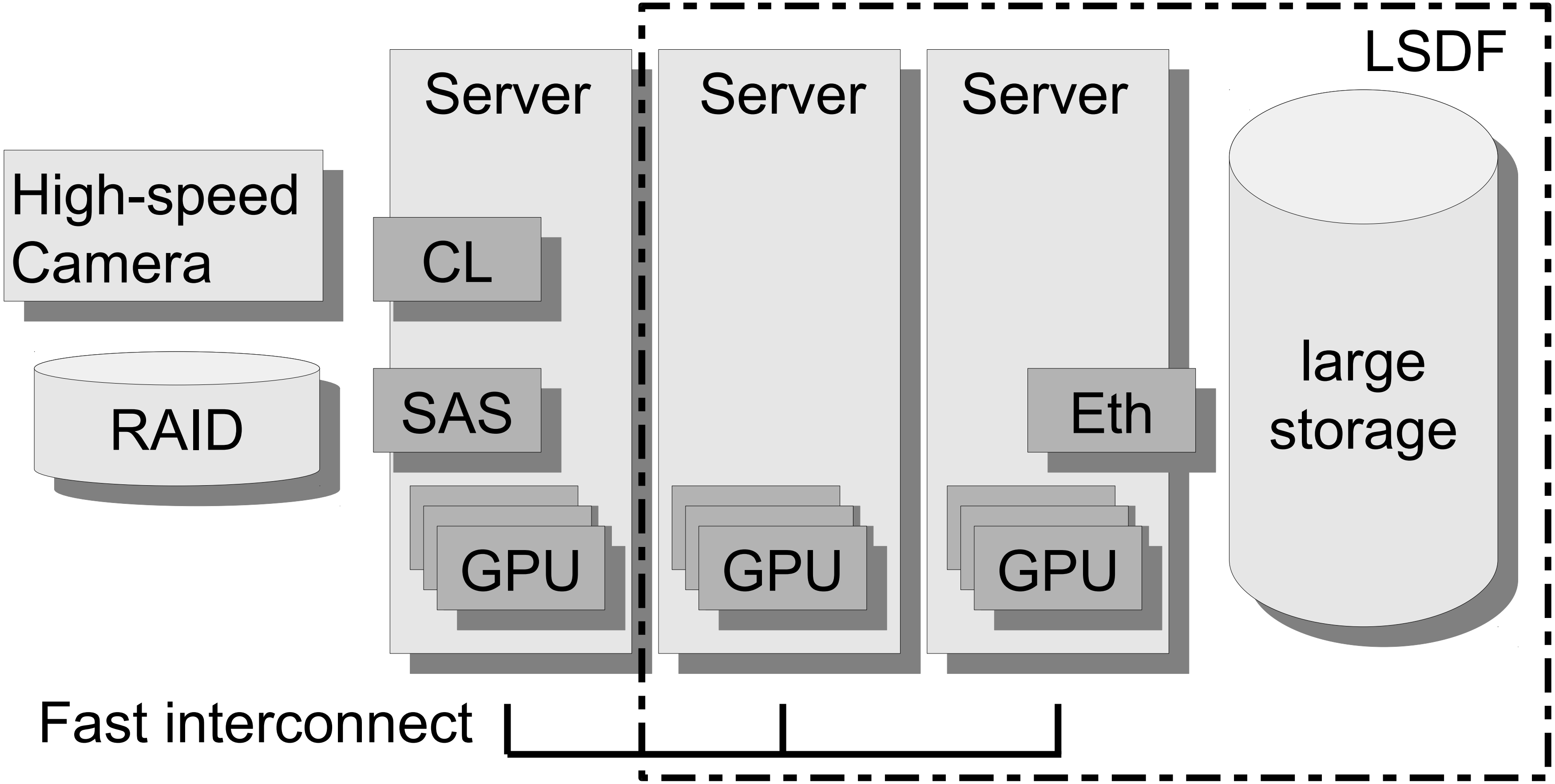
External GPU Box

PCIe Interface Card (16x)

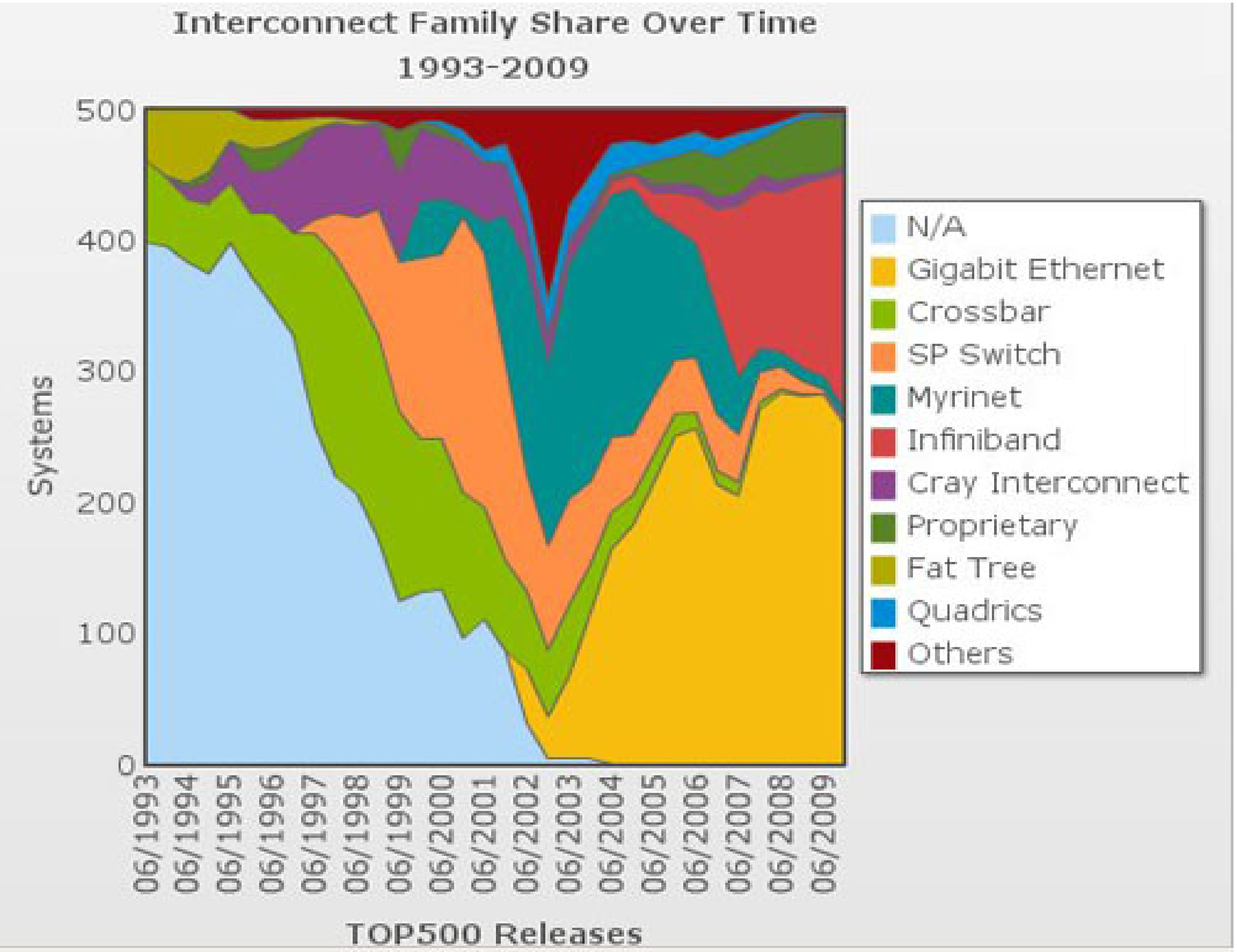
4 External GPUs

4600 EUR

GPU-Cluster


LSDF
Server
Server
Server
High-speed Camera
CL
SAS
Eth
large storage
RAID
GPU
GPU
GPU
Fast interconnect

Fast Interconnect



http://www.hpccommunity.org/content/ethernet-cluster-170/

Latency and Bandwidth

	Latency	Bandwidth
DDR3 Memory (PC1600)	~ 10ns	100 Gb/s chan. (i.e. 400Gb/s)
PCIe 2.1 x16	100-400ns ~7us CUDA	64Gb/s
QDR Infiniband (x4)	100ns ~2us MPI	32Gb/s
10GBit Ethernet	~500ns ~10us MPI	10Gb/s

GPUDirect

Without GPUDirect

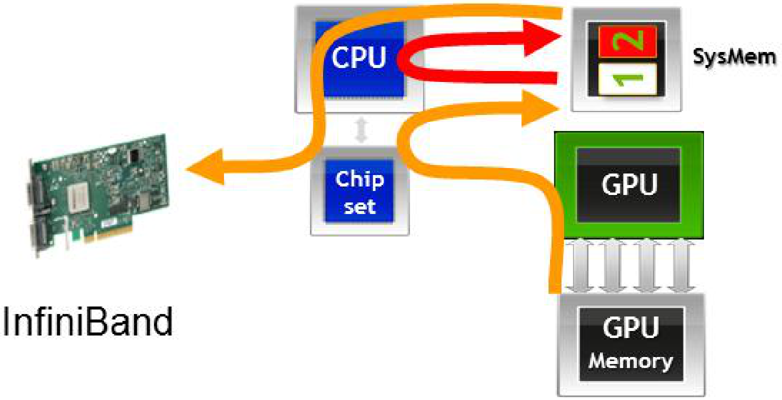
Same data copied three times:

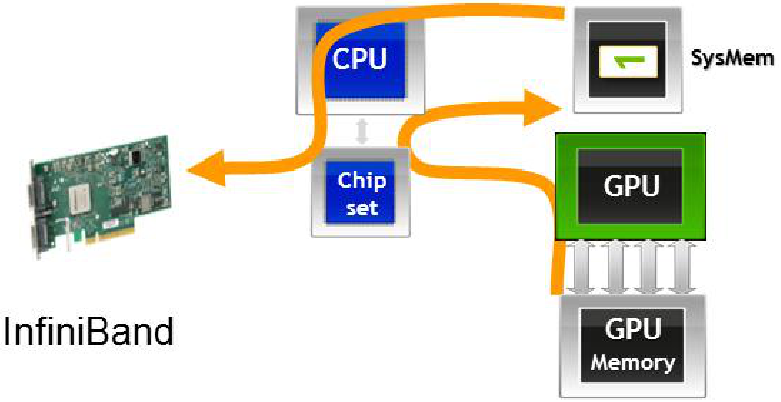
1. GPU writes to pinned sysmem1
2. CPU copies from sysmem1 to sysmem2
3. InfiniBand driver copies from sysmem2

With GPUDirect

Data only copied twice

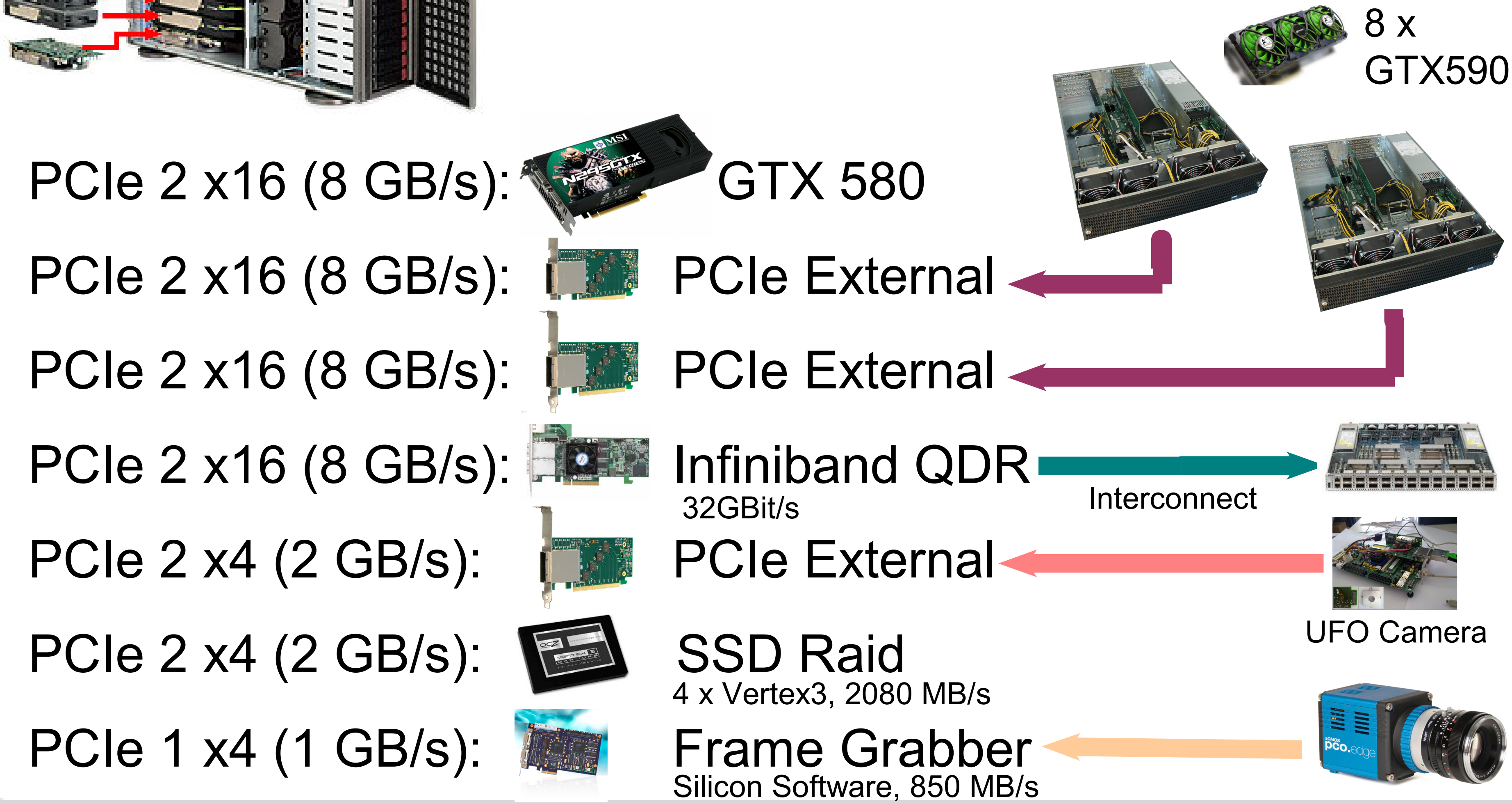
Sharing pinned system memory makes sysmem-to-sysmem copy unnecessary





Server Upgrade

SuperMicro 7046GT-TRF (Dual Intel 5520 Chipset)
CPU: 2 x Xeon X5650 (total 12 cores at 2.66 Ghz)
GPUs: GTX 580 + 8 x GTX590 External (17 cores)
Memory: 96 GB / 12 DDR3 slots (192GB max)



8 x GTX590

PCIe 2 x16 (8 GB/s): GTX 580

PCIe 2 x16 (8 GB/s): PCIe External

PCIe 2 x16 (8 GB/s): PCIe External

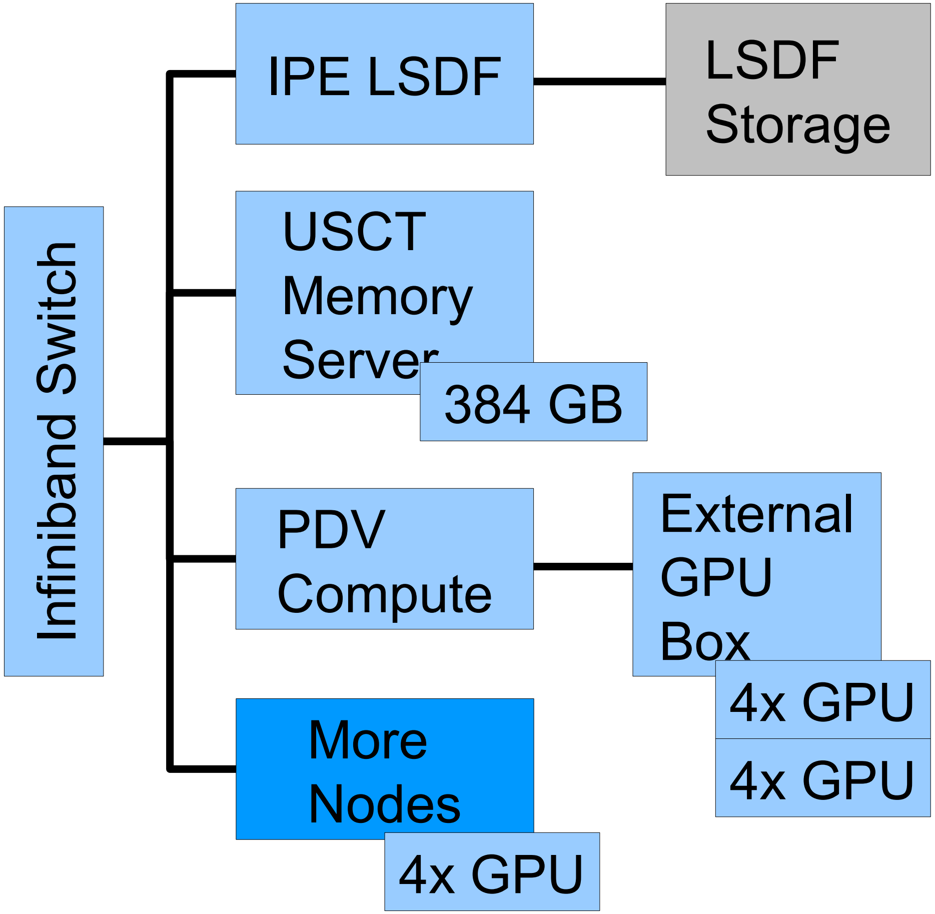
PCIe 2 x16 (8 GB/s): Infiniband QDR
32GBit/s
Interconnect

PCIe 2 x4 (2 GB/s): PCIe External
UFO Camera

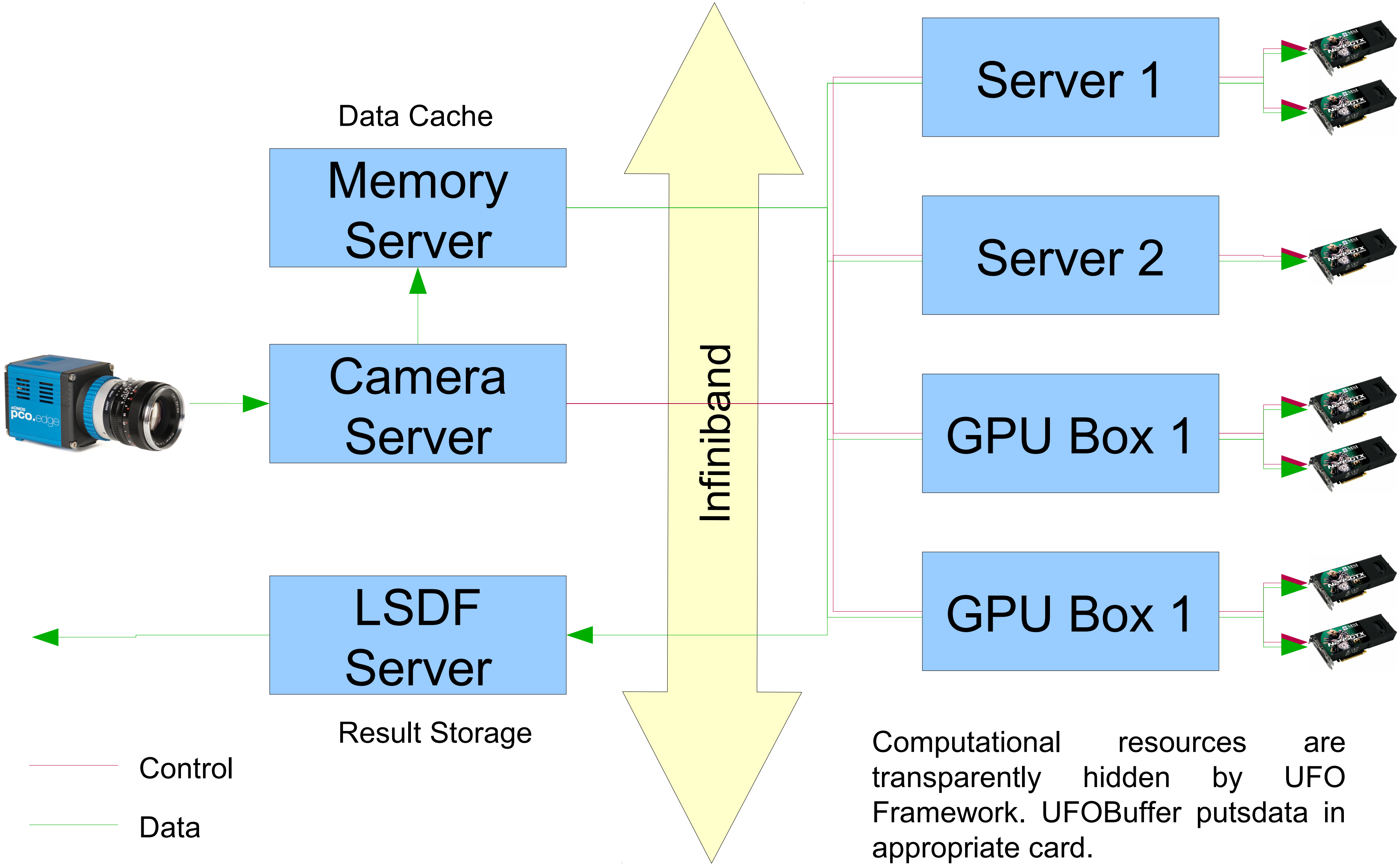
PCIe 2 x4 (2 GB/s): SSD Raid
4 x Vertex3, 2080 MB/s

PCIe 1 x4 (1 GB/s): Frame Grabber
Silicon Software, 850 MB/s

Servers


Infiniband Switch
IPE LSDF
LSDF Storage
USCT Memory Server
384 GB
PDV Compute
External GPU Box
4x GPU
4x GPU
More Nodes
4x GPU

Resource Handling



Computational resources are transparently hidden by UFO Framework. UFOBuffer putsdata in appropriate card.

What to do with the setup?

- GPU Performance in a box? Can a limit of 12th GPU cores be breached? Scalability?
- Comparison of Infiniband communication models.
- Remote GPU abstraction in the UFO Framework.
- Reconstruction performance of Local GPUs vs. Remote GPUs
- Scalability of cluster setup?
- How we can use GPUDirect to accelerate reconstruction? Is integration with UFO camera possible?
- NUMA architecture for filter scheduling: there are different distances between data and GPUs (Direct PCIe transfer, Shared PCIe transfer, Infiniband + Direct transfer, Infiniband + Shared transfer)