Karlsruhe Institute of Technology

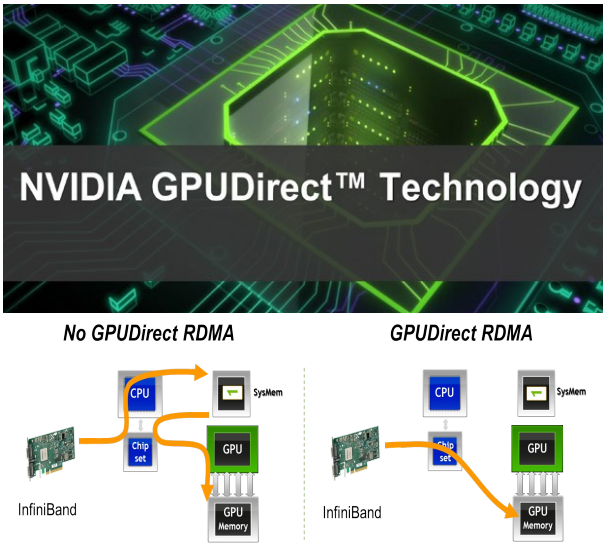
Institute for Data Processing and Electronics
Prof. Dr. Marc Weber

Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen, Germany
http://www.ipe.kit.edu

Optimizing high speed data transfer and processing of DAQ systems with NVIDIAs GPUDirect

Diploma thesis for Max Richelmann

Recent data acquisition systems are characterized by increasing data rates and the need for efficient online analysis and monitoring. Conventional CPUs are no longer able to handle the increased computational demands of scientific processes. In the field of high performance computing, GPUs with their modern and simple methods to utilize parallel processing make for an easily accessible alternative to classical CPU computing. Unfortunately, the gap between computational capabilities of GPU systems and throughput of system memory has grown tremendously and becomes the main factor limiting performance. This is especially harmful for PCI-express (PCIe) based data acquisition systems using multiple GPU cards for data processing. Using standard approaches to handle PCIe devices, the data will be copied into the system memory, sometimes multiple times, at each stage of data processing pipeline. For instance, the standard pipeline consisting of 3 stages (data readout from the frame-grabber card, preprocessing on GPU, and dispatch to the remote server over network or Infiniband interface) will include 4 copies in system memory at least and usually more depending on the hardware and software configuration.



NVIDIAs GPUDirect technology realizes RDMA to the GPU memory and reduces data transport overhead.

Recently NVIDIA revealed the GPUDirect for RDMA technology to relieve the load on the system memory. The GPUDirect/RDMA technology enables point to point transfers between PCIe devices and NVIDIA GPU on the same bus bypassing the system memory entirely. The alternative technology is called GPUDirect for Video and developed by NVIDIA specially for high-speed frame grabbers. For his Diploma work, the student is supposed to:

- Compare GPUDirect/RDMA and GPUDirect/Video technologies ,
- Provide GPUDirect-enabled drivers for our FPGA-based data acquisition platform (FDAP),
- Investigate if similar approach could be used to transfer data between FDAP and Infiniband adapters directly ,
- Evaluate the technology in terms of latency and throughput compared to the existing drivers,
- Check if and how the GPUDirect technology can be used with UFO parallel processing framework to split load across the nodes in GPU cluster.

The performance benefit of technology should be demonstrated for realistic scenarios like ultra high-speed X-ray tomography.

Required Skills: Good knowledge of C/C++ programming language as well es Linux kernel and driver development. Knowledge of parallel programming models is a plus.

Experience Gained: Parallel programming, GPU programming, RDMA data transfer mechanisms.

Contacts:

Suren Chilingaryan <suren.chilingaryan@kit.edu>, phone: +49 721 / 608 26579

Timo Dritschler <timo.dritschler@kit.edu>, phone: +49 721 / 608 25693