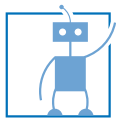




Proposal for a Bachelor/Master Thesis at the
Institute of Robotic and Embedded system

Direct GPU/FPGA Communication

Advisors: Biao Hu
Dr. Kai Huang

Professor: Prof. Dr. Alios Knoll

Project Description

In the context of TU9 project, research at the Institute of Robotics and Embedded Systems is dealing with the challenge of providing high-performance ECUs as an enabling technology applicable in the automotive field, which will be a heterogenous system with multi-core CPU, FPGA, and GPU.

A key requirement of such heterogeneous computing system is the ability to transfer data between components at high bandwidth and low latency. Several GPGPU abstractions support explicit transfers between the CPU and GPU[1, 2, 3], and it has recently been shown that this is also possible between CPU and FPGA [4]. However, the GPU-FPGA communication still relies on cumbersome approach. Existing facilities may be used to implement GPU to FPGA communication by transferring data through CPU memory as illustrated by the red line in Fig. 1. Data must traverse through the PCI Express switch twice and suffer the latency penalties of both the operating system and the CPU memory hardware using the red indirect path. We refer to this as a GPU-CPU-FPG transfer. This additional indirection adds communication latency and operating system overhead to the computation, as well as consuming bandwidth that can otherwise be used by other cluster elements sharing the same communication network.

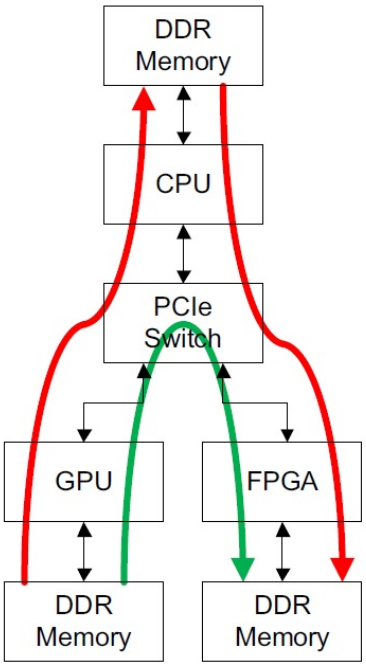
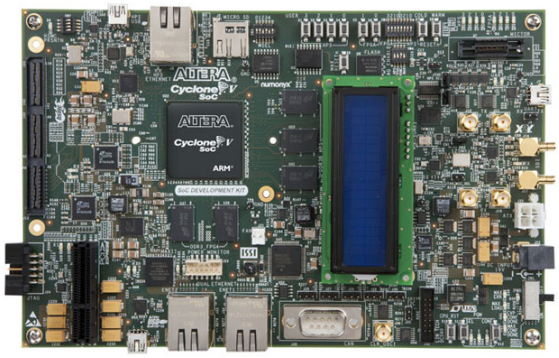


Figure 1: Conceptual Models of GPU to FPGA Transfers

Therefore, our aim is to describe and evaluate a mechanism for implementing the green line in Fig. 1 for direct, bidirectional GPU-FPGA communication over the PCIe bus [5]. As illustrated, data moves through the PCIe switch once and is never copied into system memory, thus enabling more efficient communication between these disparate computing elements. We refer to this as GPU-FPGA transfer.

The project will include the following phases:

- Investigate the current technique of GPU-FPGA communication.
- Design a new GPU-FPGA communication driver for Cyclone V SoC board (Fig. 2a) and Nvidia Geforce GTX 680 (Fig. 2b).
- Evaluate the communication driver with some test cases.



(a) Cyclonev SoC dev board

(b) Geforce Gtx 680

Figure 2: Hardware pictures

This work will be carried out in TU9 project to which the Institute of Robotics and Embedded Systems is contributing in terms of hardware and software design and performance analysis.

Kind of Work

- 10% theory
- 40% implementation
- 30% evaluation
- 20% documentation

Requirements

- good knowledge on communication theory and computer architecture.
- good knowledge of FPGA design flow with VHDL or Verilog HDL (or probably CUDA or OpenCL).
- ability to acquaint yourself with an unfamiliar software design environment.

Contact

Dr. Kai Huang
MI 03.07.042
huangk@in.tum.de
Phone: +49.89.289.18111

Biao Hu
MI 03.07.059
hub@in.tum.de
Phone: +49.89.289.18128

References

[1] K. Group, "OpenCL: The open standard for parallel programming of heterogeneous systems," http://www.khronos.org/opencl/.

[2] M. Corporation, "DirectCompute," http://blogs.msdn.com/b/chuckw/archive/2010/07/14//directxompute.aspx.

[3] nVidia Corporation, "nVidia CUDA API Reference Manual, Version4.1," http://ww.nvidia.com/CUDA.

[4] R. Bittner, "Speedy bus mastering pci express," in *Field Programmable Logic and Applications (FPL), 2012 22nd International Conference on*, Aug 2012, pp. 523–526.

[5] A. Goldhammer and J. Ayer Jr, "Understanding performance of pci express systems," *Xilinx WP350, Sept*, vol. 4, 2008.