An FPGA platform for ultra-fast data acquisition

M. Caselle, M. Balzer, S. Chilingaryan, A. Kopmann, U. Stevanovic, M. Vogelgesang

FPGAs in Research - Applications, Technologies and Tools, Forschungszentrum Jülich, 3-4 December 2012
Ultra Fast X-ray Imaging (ANKA/UFO experimental station)

UFO → Ultra-Fast X-ray Imaging of Scientific Processes with On-line Assessment and Data-driven Process Control

**High spatial resolution** (<1 µm) included 2D and 3D visualizations

**Time resolution** (2D: ≈10kHz, 3D: ≈10Hz) to give insight in the temporal structure evolution and thus access to dynamics of processes

**Main application fields:** medical diagnostics, biology, non-destructive testing, materials research and etc.

**Requirements:**

- **High granularity** and **low noise** monolithic silicon pixel detector, few µm pixel pitch, several MPixel matrix operating at several kframes/sec

  *High readout bandwidth up to 50Gb/s with GPU* (3D-tomography reconstruction)
KIT-IPE – Readout concept of high data throughput for scientific applications

**Concept:**

Data source

- X-ray detector
- CMOS image sensor
- Fast ADC

DAQ Boards

- Connection
- FPGA
- Memory

Real time data elaboration
Data reduction
High-throughput data flow

Driver

- GPU/CPU algorithms

Fast Data storage

Feedback loops

Up to 10 GB/s

Up to 4 GB/s

Up to 0.25 GB/s

GPUs/CPUs infrastructure

Small PCIe backplane

Under developing by Data processing group in KIT-IPE

Data storage LSDF
**KIT-IPE – Readout concept of high data throughput for scientific applications**

**Concept:**
- Data source
  - X-ray detector
  - CMOS image sensor
  - Fast ADC
  - ...  
- FPGA & Readout Board
  - Small backplane
  - Mother readout board
  - Daughter sensor board

**Implementation:**
- UFO Camera
  - FPGA Virtex 6
  - PCIe link to DAQ

**System Overview:**
- Real time data elaboration
- Data reduction
- High-throughput data flow
- GPUs/CPU infrastructure
- Data storage
  - LSDF
  - Up to 10GB/s
  - Up to 4GB/s
  - Up to 0.25GB/s

**Feedback Loops:**
- Driver
  - GPU/CPU algorithms
  - Fast Data storage
- Under developing by Data processing group in KIT-IPE

**This talk is focus on FPGA & Readout Board**
Flexible high-throughput FPGA platform

PC DAQ

FPGA internal architecture

CPU

Chipset root port

memory

Optical/Electrical
X4 lanes @ 5Gb/s

FSM Master control

User bank register

FIFO

SerDes input stage
(KIT_ipcore)

Data Source (Detector)

Remote Detector Control

Remote Detector Control

X lanes @ 500Mbit/s

...
Flexible high-throughput FPGA platform

PC DAQ

CPU

Chipset root port

memory

FPGA internal architecture

FSM Master control

On-line parallel data processing

DDR interface (KIT_ipcore)

User bank register

DDR3 memory (800MHz @ 64bit)

FIFO

SerDes input stage (KIT_ipcore)

Data Source (Detector)

Remote Detector Control

Optical/Electrical
X4 lanes @ 5Gb/s

Remote Detector Control

User bank register

DDR interface (KIT_ipcore)

FSM Master control

On-line parallel data processing

X lanes @ 500Mbit/s

...
Three logic cores have been developed for a flexible high-throughput platform:

- PCI-e Bus Master DMA readout architecture
- Multi-port high speed DDR3 interface
- Configurable 2..16 bits “SerDes” (Serializers/Deserializers) architecture

- PCI Express/DMA Linux 32-64 bits driver with ring buffer data management
- Integration in the parallel GPU/CPU computing framework
PCIe-Bus Master DMA readout architecture

- Bus Master DMA operating with 4 lanes PCIe @ Gen2 (250MHz)
- Two individual engines for write/read from FPGA (User logic) to PC centre memory
- IN and OUT FIFO-like interface (for User logic)
- FIFO used to decouple the time domain between DMA and User custom logic

DMA performance with PCIe X4 @ GEN2

- TX (PC -> FPGA)
- RX (FPGA -> PC)
Preliminary, PCIe-Bus Master DMA new architecture

Disadvantage of IP-cores from external vendors, are:

1) expensive (35k€ for North-West DMA and 10-60k€ for EZDMA/QuickPCIe-IP by PLDA)
2) for unique FPGA family (Virtex 6, speed grade -2)
Preliminary, PCIe-Bus Master DMA new architecture

Disadvantage of IP-cores from external vendors, are:

1)   expensive (35k€ for North-West DMA and 10-60k€ for EZDMA/QuickPCIe-IP by PLDA)
2)   for unique FPGA family (Virtex 6, speed grade -2)

New, KIT-IPE Bus Master DMA engines operating with x8 lanes PCI Express @ GEN 2

IN/OUT data at 128 bit @ 250MHz → internal bandwidth of 32 Gb/s in Read/Write

Comparison (NW-DMA vs. KIT-DMA)

Data valid for X58 PCIe chipset

FPGA ring buffer management → on-going

Software 64bit@linux driver → under optimization (≈ 32Gb/s)
Two-ports DDR3 memory interface architecture

Why a two-ports DDR3 memory controller .. ?

The Xilinx Multi-port Memory Controller (IP-Core) is limited in the maximum data throughput (less than 2GB/s for each port) & complex user interface.

Ref. LogiCORE IP Multi-Port Memory Controller (MPMC) (v6.03.a), DS643 March 1, 2011
Two-ports DDR3 memory interface architecture

Why a two-ports DDR3 memory controller .. ?

The Xilinx Multi-port Memory Controller (IP-Core) is limited in the maximum data throughput (less than 2GB/s for each port) & complex user interface.

Ref. LogiCORE IP Multi-Port Memory Controller (MPMC) (v6.03.a), DS643 March 1, 2011

Bandwidth 51Gb/s, limited by FPGA speed grade (Virtex 6, speed grade -1)

Two operations are possible in same/different segmentation/s (each operation ~ 25Gb/s)

Data interface FIFO-like, minimum control signals are required

FIFO used to decouple the time domain between Memory Controller and custom User logic

Configurable user define data width N and M → 32/64/128/512 bits
A configurable “SerDes” input stage architecture

Why not a Xilinx ISERDERSE stage .. ?

- Limited parallel data width (output) not more than 10bits (for two ISEDERSE in cascade configuration) and not dynamically configurable. The FSM Alignment is not included in the Xilinx tools.

Ref. Virtex-6 FPGA Select IO resources user guide. ug361 (v1.3) august 16, 2010.
A configurable “SerDes” input stage architecture

Why not a Xilinx ISERDERSE stage .. ?

- Limited parallel data width (output) not more than 10bits (for two ISEDERSE in cascade configuration) and not dynamically configurable. The FSM Alignment in not included in the Xilinx tools.

Ref. Virtex-6 FPGA Select IO resources user guide. ug361 (v1.3) august 16, 2010.

Configurable 2 to 16bit parallel data output “SerDes “ logic with MSB Alignment State Machine

- Individual clock-to-data time tuning by IODELAY (time step of 75psec)
- I/O clock buffer located in the centre of the FPGA bank
- Regional buffer synchronous to parallel data out
- “SerDes” input stage fully configurable by User
Future developments for high speed readout systems

Requirements:

- Real-time FPGA + GPU data elaboration → high data throughput (range of 64Gb/s)
- Data source and FPGA readout board located far from DAQ system
- Using commercial/well-known protocol for ease interface with commercial devices/boards

Two different approaches are possible:

- **Peer – to – peer (P2P) streaming data transfer**
  (based on new generation of PCI express protocol)

- **Point – to – node (net) for distributed GPU/CPU High Performance Computing (HPC) clusters**
IPE - PCI Express Readout card - Overview

✓ PCIe GEN3 optical/electrical data transmission (8 lanes x 8GT/s)

**Readout Board - Concept**

- **DDR3 Memory**
- **User logic**
- **EndPoint PCIe Integrated block x8 lanes GEN2**
- **Multi-port PCIe switching X8 lanes GEN3**
- **No DMA is needed**

**Data Source**

**FPGA - Virtex 6**

- Electrical cable (up to 5m)
- Optical cable (up to 30m)

**MiniPOD X12 lanes optics cable for PCIe GEN3 (8 GT/s per lane)**

- 64 Gb/s (W) + 64Gb/s (R) → full-duplex mode
- FPGA Real Time process → close to data source
IPE - PCI Express Host card - Overview

- PCIe host board with high speed data recording

PCIe – host card

- Fully configurable data flow

FPGA

DMA integrated

NAND flash SSD
Up to 2TB @ 64 Gbit/s

To PC memory system for GPU data elaborations

X16 lanes PCIe slot
High bandwidth readout system based by InfiniBand

Data Source → Input stage → FPGA → Memory → QSFP+

Optical or electrical data link up to 100m

- 40Gb/s → InfiniBand, in house
- 120Gb/s → InfiniBand available soon
- 384Gb/s → in the next two years

8-port switch capable of up to 640Gb/s

InfiniBand GPU cluster under developing in KIT-IPE by Data processing group IPE-KIT

Heterogeneous FPGA + CPU + GPU

InfiniBand Router

QDR 40Gbps InfiniBand protocol

Ultra-low latency for high cluster performance

Optionally

μ/ATCA

InfiniBand DAQ cluster

FPGAs in Research - Applications, Technologies and Tools, Forschungszentrum Jülich, 3-4 December 2012. M. Caselle
InfiniBand readout Board - Overview

High Speed connectors
From data Source
(HPC Samtec or similar)

- IP based application layer → possible (i.e. TCP, UDP, SSH, FTP ..)
- The InfiniHost provide the PHY, Link and Transaction layers for InfiniBand
- Remote DMA for fast data transfer → intranet communication
Conclusion and What’s next

Logic cores for high data throughput platform → employed in several scientific applications:


- **A X-ray camera for phase contrast tomography** (M. Caselle, A. Kopmann, Felix Beckmann (HZG), Joerg Burmester (HZG) KIT and HZG

- **A X-ray camera for high spatial resolution tomography** (M. Caselle, M. Balzer, A. Kopmann, V. E. Asadchikova) Shubnikov Institute of Crystallography, Russian Academy of Sciences, Moscow, Russia

- **A readout electronics for Ultrafast electron beam X-ray tomography system "ROFEX" in HZDR** (proposal under discussion)

- **New KIT-DMA (32Gb/s) engines → developed and tested**

- **Driver 64bit@Linux → under optimization**

What’s next

- **Design & production of readout board based by:**
  - PCIe GEN3 optical communication
  - InfiniBand protocol
  - Integration in the GPU/CPU compute infrastructure
Frame rate from 500 to 2Kfps

Bandwidth: 8 Gb/s. Future upgrade: 50Gb/s

Recording & analysis of time evolution of each bunch in a multi-bunches accelerator filling-scheme

Bandwidth: 6Gb/s: Future upgrade to 24Gb/s

Thank you for your attention
Backup slides ..
Credit-based link-level flow control

- Link Flow control assures NO packet loss within fabric even in the presence of congestion
- Link Receivers grant packet receive buffer space credits per Virtual Lane
- Flow control credits are issued in 64 byte units
InfiniBand: application layers and latency

Ref: Introduction to InfiniBand™ for End Users,
InfiniBand Trade Association
3855 SW 153rd Drive Beaverton, OR 97006

Designing with InfiniBand

UDP or TCP, FTP, ssh ...

Schlumberger ECLIPSE
(FOURMILL)
UFO architecture - overview

- Smart high-speed camera
- FPGA
- On-line Processing
- Fast Data Link
- Memory
- Detector
- Fast-reject Trigger
- SW control loops
- 2D and 3D image-based control loop
- GPU server
- Online monitoring and evaluation
- Raw Data Processing
- Data Evaluation
- Post Processing
- Visualization
- Optical link
- Large scale data facility (LSDF)

- High speed & bandwidth, full programmable camera (continuous data acquisition at full speed)
- Optimized image processing algorithm using GPU computing
- Fast HW loop: On-line image-based self-event trigger architecture (Fast reject)
- SW control loops: based on 2D and 3D data evaluation:
  - 2D data → camera calibration, autofocus, self-alignment & etc.
  - 3D data reconstructed → like optical flow, etc.
The main features already implemented and tested, include:

- **Fully configurable camera** → adjustable image exposure time and dynamic range, analog and digital pixel features as pixel threshold, mask, analog gain, etc.
- **Continuous data acquisition at full speed**
- **On-line image-based self-event trigger architecture (Fast reject)**
- **Region-of-interest readout strategy using self-event trigger information**
- **Easily extendable to any available CMOS image sensor**
Readout electronics for Coherent Synchrotron Radiation

- Measure of the peak amplitude of each bunch (resolution few mV)
- Measure of the pulse width of each bunch (resolution few psec)
- Measure of the relative time jitter between electron bunches (res. few psec)

**Strategy:**
Digitalize each pulse with 4 samples + pulse reconstruction & Constant Fraction Discriminator (CFD) for precise pulse timestamp.

![Analog signal (single bunch) (output of amplifier)]

**ANKA CSR (long observation time with YBCO):**

- Recording & analysis of time evolution of each bunch in a multi-bunches accelerator filling-scheme