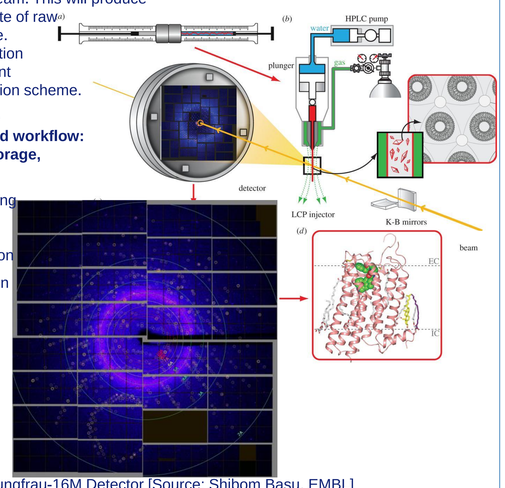
# High-throughput Low-latency Online Image Processing by FPGA/GPU Coprocessors using RDMA Techniques

Raphael Ponsard [a,b*], Jerome Kieffer [b], Nicolas Janvier [b], Vincent Fristot [a] and Dominique Houzet [a]
[a] Grenoble Alpes University, Gipsa-Lab, 11 Rue des Mathématiques, 38400 Saint-Martin-d'Hères, France
[b] ESRF, The European Synchrotron, 71 Avenue des Martyrs, 38000 Grenoble, France. http://www.esrf.eu +33 (0)4 76 88 40 72
Correspondence e-mail: raphael.ponsard@esrf.fr

The continual evolution of photon sources and high-performance X ray **2D image detectors** drives cutting-edge experiments that can produce **very high throughput data streams**. Large data volumes are **challenging to manage and store**. This thesis investigates **direct data placement** (zero-copy) from the detector head to the processing computing units, **bypassing CPU and network stack** (RDMA). An evaluation of Remote Direct Memory Access over Converged Ethernet (**RoCEv2**) and **PCI-e long distance** is presented. As a key contribution, we extend RASHPA with **low latency data processing** using **massively parallel coprocessors**. Scalability and versatility of the proposed system is exemplified with detector simulators and data processing units implementing **Multi-core CPU**, **GPU or FPGA accelerators** as well. **Online data analysis pipeline** is exhibited featuring raw-data preprocessing, rejection and CSR compression suitable for demanding SSX experiments.

## Background of Photon Sciences at Big Data Era

- **Serial Crystallography is among the most demanding case of photon science in terms of data throughput.**
  1. In a typical SSX experiment, a liquid crystalline polymer (LCP) jet propels micro-crystal samples in a pulsed X-ray beam. A rotating chopper produces X-ray pulses synchronous to the data acquisition system.
  2. This enables the collection of alternate dark and signaling images at a up to 2000 images per seconds.
  3. When acquiring 4 M pixel 16-bits images, such a high repetition rate will result in a 128 Gb/s data stream. This will produce nearly one Terabyte of raw data in one minute. Continuous operation requires an efficient online data reduction scheme.




- **New requirements are challenging the standard workflow: acquisition, transfer, storage, batch processing**
  1. Raw data processing
  2. Fast-feedback
  3. Online data rejection
  4. Online compression
  5. Disaggregated storage

PSI Jungfrau-16M Detector [Source: Shibom Basu, EMBL]
(32 modules x 512 x 1024 pixels x 2 kHz x 16 bits and adaptive gain = **62 GB/s**)

## Problem Statement: Where are there Bottlenecks in Data Path ?

- **In the Linux operating system**

Memory isolation between processes and kernel has many advantages for robustness and security but also a cost in term of CPU processing time :

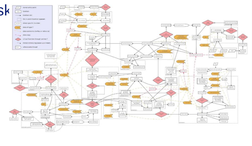
1. Memory copy between kernel space memory and user space memory
2. Context Switches occurring at each system call (time saving/restoring processor state and registers before/after executing kernel code)
3. Virtual to physical memory address translation and multiple look-up table access
4. Devices drivers Interruptions handling at high rate becomes CPU-time consuming

- How mitigates these overheads with user space application :
  1. Using Zero-copy techniques so that data could be written directly to user-space destination buffer
  2. Using Direct Memory Access Engine (DMA) to free processor cores during data transfer. Application shall only provision DMA with Descriptors list (address, length)

- **In standard network protocol stack and NIC hardware**

Handling packet loss and out of order delivery is challenging task retransfer mechanism:

1. TCP/IP is an highly sophisticated software stack putting a lot of pressure on the system.
2. Above 10Gb/s bandwidth, a single CPU core at 2GHz can not cope with data stream.

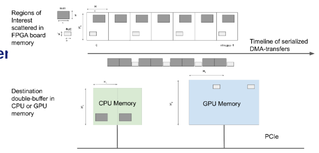


TCP State Machine [Source: Linuxfoundation]

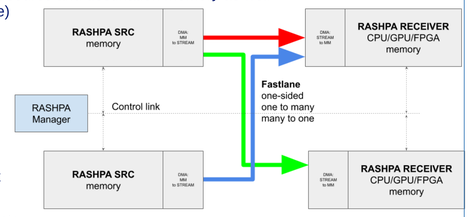
Distributing inbound network traffic across several receive queues (RSS) processed by multiple cores. This solution is at the root of DPDK project, but not adequate in our use case.

## Extending the RASHPA Framework to FPGA/GPU Coprocessors

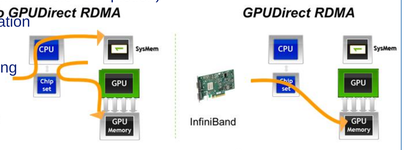
This work is part of a wider RASHPA project [1], a data acquisition framework optimized for 2D X-ray detectors that is sufficiently generic and scalable to be used in a wide diversity of new high performance detector developments.




- **RASHPA distinctive features for direct data placer**
  1. Leveraging RDMA/zero copy techniques
  2. Data movement fully controlled from detector head
  3. Single-sided transfer: no ack, no retransmission (as per detector electronics requirements)
  4. Selectable Dataset of images at source and destination memory as well (and Region of Interest in Image)
  5. Transfer multiple data flow to multiple destinations simultaneously




- **The thesis focus on:**
  1. RoCEv2 implementation study:
     - Feasibility study [3], performance assessment
     - Develop a code base compliant with requirements of Embedded FPGA electronics
     - Propose an event mechanism (RDMA transfer completion)
  2. Online data processing and GPUs integration
     - Proposes a synchronization mechanism to trigger data processing at the end of DMA transfer
     - Reduce kernel launch time, evaluate spinning kernel
     - Direct transfer into GPU memory (GPUDirect technology)
  3. Extension of the RASHPA framework to FPGA coprocessors
     - Design supporting DMA and PCI-e peer device transfer
     - Benefits from High Level Synthesis (Xilinx Vitis_hls workflow) and latency constrained design adapted the to processing of custom data format

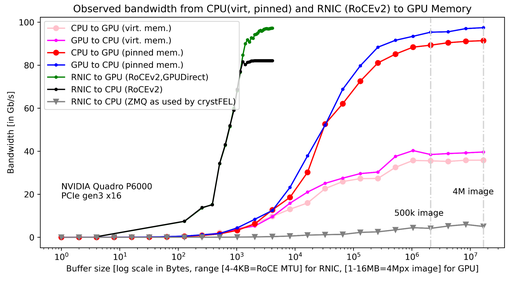




## REMU-RoCEv2: The first RASHPA Detector Emulator and GPU Processing Pipeline

Using dedicated Network Interface Card (NIC) such as Mellanox Connectx-5, it is possible to process RDMA transfer bypassing Linux kernel and even to write straight into GPU accelerator memory (GPUDirect) using PCIe peer device transfer.

The framework benefits from libibverbs RDMA WRITE on UC Queue pair and WRITE_WITH_IMM to implements events.

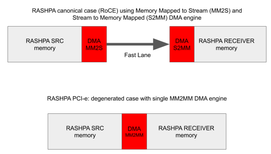


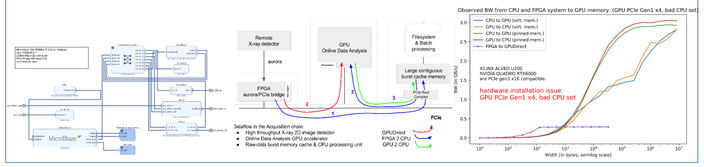

Research work presented in Article [2]

## REMU-PCIe: A RASHPA Source based on Softcore-Controlled DMA

- FPGA based PCI-e detector emulator
- Softcore (Microblaze) controlled DMA
- Aurora – PCI-e bridge
- Large contiguous memory DMA-able buffer allocator (libcdma) in Linux receiver system and GPU device memory (libgdma)

Short presentation in conference [4] and [5]




## Serial Crystallography Online Data Analysis Challenges

- **Efficient computing pipeline**

Stack of Images transfer overlap computations

Synchronization with RDMA transfer completion

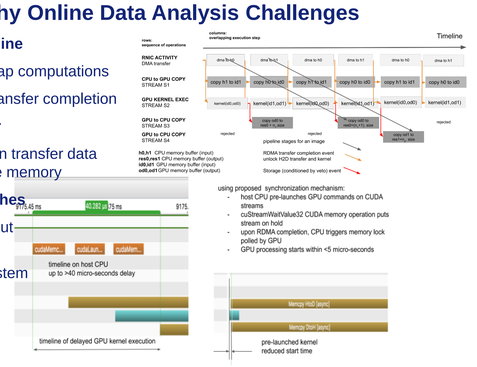
- **PCI-e peer direct transfer**

On compliant device, RNIC can transfer data directly into GPU/FPGA device memory

- **Low latency kernel launches**

Kernel are pre-launched and put on hold by user application

started by RASHPA events system using CUDA special API cuStreamWriteValue32 API




- **Data de-packing and image reconstruction**

Jungfrau detector under development by Paul Scherrer Institute (PSI) is devised with XFEL experiments in mind featuring ultrafast acquisition time, automatic gain switching, …

1. JF has a modular design and up to 32 modules. Each independent module is featuring 1024 x 512 pixels, with one or two 10 Gb/s Ethernet links.
2. Photon intensity computed from raw data pixel by pixel using up to 6 pixel specific pedestal and gain correction factors. Pedestal value, that must be subtracted to raw data, changes after each image in order to mitigate noise drift of the sensor. This pedestal is a measurement of a dark image while the X Ray beam is eliminated by the mean of a rotating chopper between two sampled pictures.

$$Pixel_{i,j}\,[keV] = \frac{\left(Raw_{i,j}\,[ADU] - Pede_{i,j}[ADU]\right)}{Gain_{k,i,j}\left[\frac{ADU}{keV}\right]}$$

- **Image rejection**
  1. If no observable diffraction pattern are found on image (peak counter) then it should be reject by the system. Extremely high rejection rates are expected, up to 90 or 99 %.
  2. The rejection algorithm is a standard deviation computation on each image, at full size or at quadrant level. If there are not found enough observable Bragg peak, i.e. photon intensity over a specific threshold, the image is reputed of poor quality and rejected.
- **Image sparsification and compression**
  1. Compressed Sparse Row (CSR) matrix format using cum-scan generated by PyCuda metacompiler
  2. Azimuthal Integration using PyFAI algorithm
- **Low-latency control loop of X ray experiment**
  1. Find the center of rotation of an image (Auto-correlator)

## Ongoing Works

- **IBM AC922 as RASHPA receiver**

Power9 processor and Tesla GPU are connected by NVLINK2. Address Translation System (ATS) system handles managed memory data buffer (same pointer in CPU or GPU memory) and ensures migration/coherency automatically.

But ATS relies on "On demand Paging" memory registration which requires RC queue pair (with ACK, credits) incompatible with detector electronics as is.

- **XILINX ERNIC IP as RASHPA initiator**

Xilinx IP featuring RoCEv2 available since June 2020

Performance evaluation in the frame of RASHPA Framework

Comparison with home-made IP

Design a RASHPA-compliant test bench

- **FPGA Assisted control loop**

Demonstrate the suitability of FPGA processing using XILINIX Vitis workflow

Demonstrate feasibility of low-latency control in the frame of 2D X ray ultra-fast data acquisition, performing beam alignment with sample using real-time auto-correlator

## Outlooks

Data Plane Development Kit integration as low cost alternative

Evaluation of Fast Storage Solutions using NVMEoF and SSD

CUDA task Graph and continuously spinning evaluation

## References

[1 ] Le Mentec, F., Fajardo, P., Le Caër, T., Hervé, C., & Homs, A. (2013, October). RASHPA: A data acquisition framework for 2D X-ray detectors. In ICALEPCS.
**[2] R. Ponsard**, N. Janvier, J. Kieffer, D. Houzet, and V. Fristot, "RDMA data transfer and GPU acceleration methods for high-throughput online processing of serial crystallography images," *J Synchrotron Rad*, vol. 27, no. 5, Art. no. 5, Sep. 2020, doi: 10.1107/S1600577520008140.
[3] W. Mansour, N. Janvier, and P. Fajardo, "FPGA Implementation of RDMA-Based Data Acquisition System Over 100 GbE," *arXiv:1806.08939 [physics]*, Jun. 2018, Accessed: Feb. 25, 2019. [Online]. Available: http://arxiv.org/abs/1806.08939.
**[4] R. Ponsard**, W. Mansour, N. Janvier, D. Houzet, and V. Fristot, "Online GPU Data Analysis using Softcore-Controlled DMA for 2D X-ray Detector", EuroMicro-DSD2020 Conference
**[5]** W. Mansour, R. Biv, **R. Ponsard**, C. Ponchut, N. Janvier, P. Fajardo, ESRF, FPGA-based real-time image manipulation and advanced data acquisition for 2D X ray detectors RTC2020 (accepted)