RASHPA
a RDMA-based data transfer platform for high-performance DAQ applicatons

Future detectors for the European XFEL workshop

18-19 Sep 2023

Pablo Fajardo
Detector & Electronics group
***Instrumentation Services and Development Division*, ESRF**

ESRF | The European Synchrotron

RASHPA development:
Fabien LeMentec
Nicolas Janvier
Wassim Mansour
Aurèlien Bideuad
Raphaël Ponsard

DISTRIBUTED DATA ACQUISITION FOR ON-LINE PROCESSING

Motivation

- A single data receiver approach is no longer sustainable with increasing data throughputs.
- Need of tools to implement an efficient scheme for multi-receiver data distribution and processing.

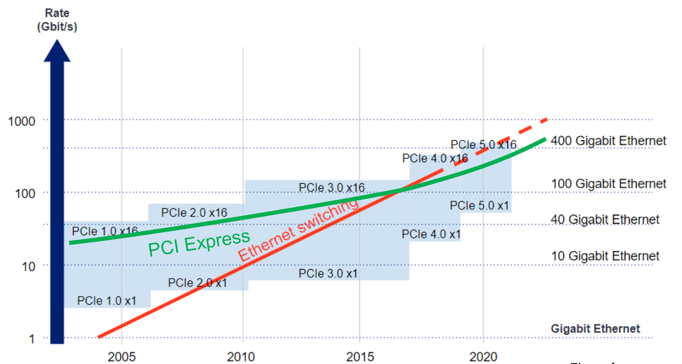
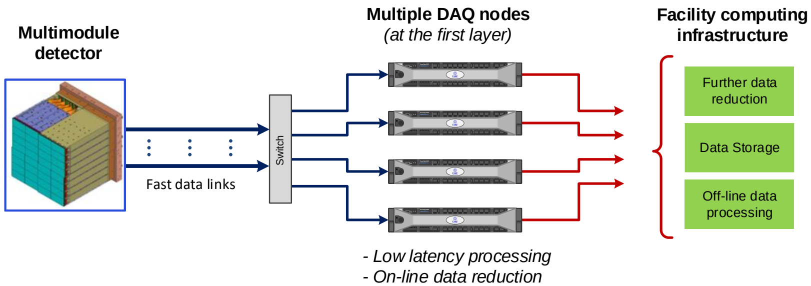


Figure from: wwwplda.com



The ESRF approach to build fully distributed DAQ systems relies on two components:

- **RASHPA** : Transfer and redistribution from the detector data into the processing nodes
- **Lima2** : Software framework for building detector servers with parallel processing (MPI based)

ABOUT RASHPA

What is RASHPA?

- A **data dispatching** layer that transfers data from the detector to the **first level of computing** nodes
- A component to be integrated in a DAQ system for **high data throughput** detectors
- Uses remote direct memory access (**RDMA**)
- Designed/optimized for **modular/segmented** detectors and **2D detector data** (images)

Some features:

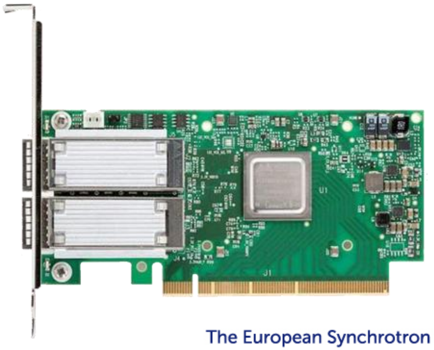
- **Zero-copy**: data is pushed by the detector modules into destination buffers, no software intervention
- Uses a **switchable network** supporting RDMA transfers, the number of processing nodes is arbitrary
- Data is dispatched and redistributed in a **configurable** way, to match the application needs
- **Fully COTS backend** components: no dependence with specific hardware, the achievable performance can follow the evolution of the market

DATA COMMUNICATION

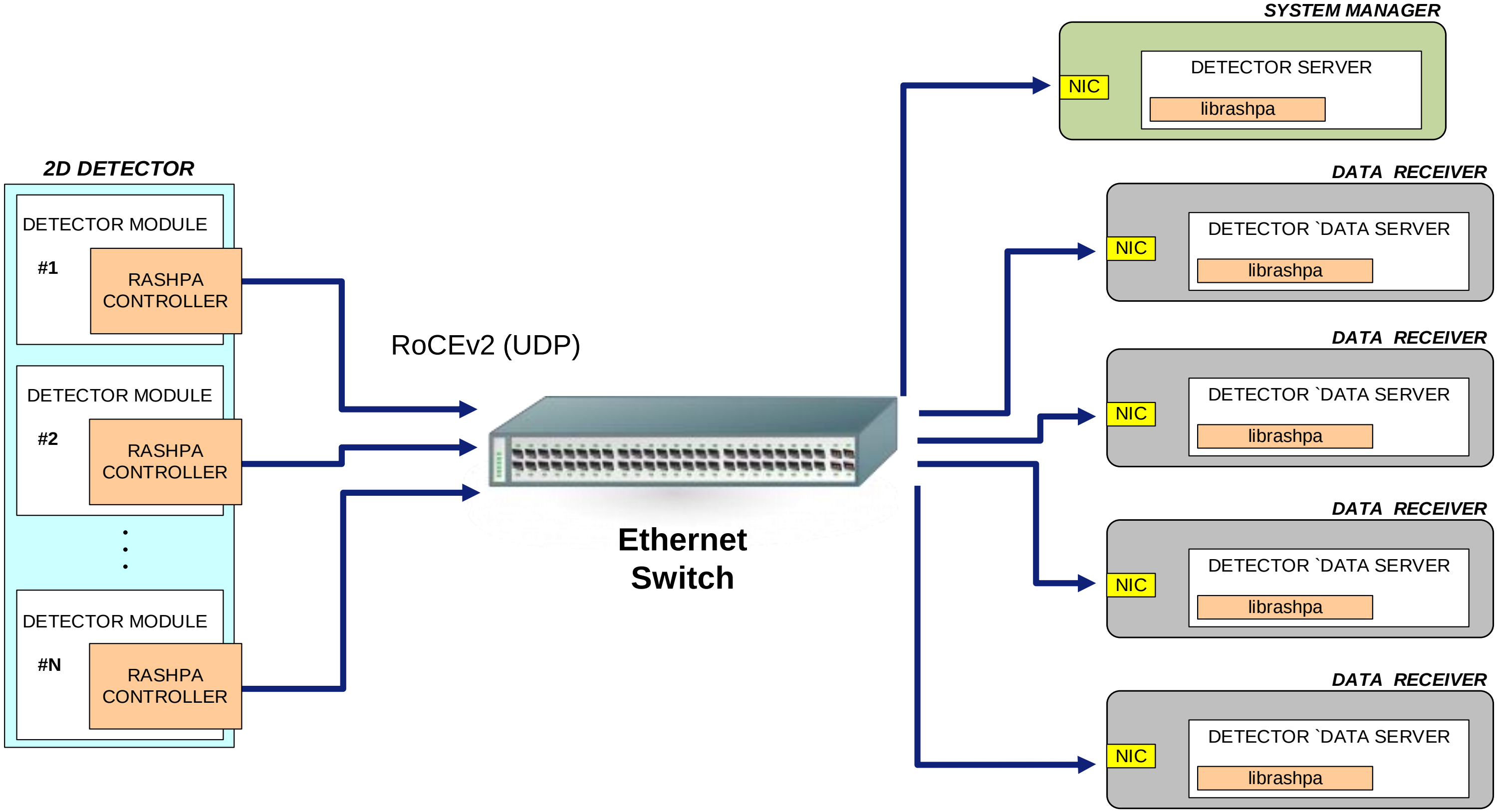
- Data transmission protocol (RASHPA functional requirements)
  - Support of RDMA routable/switchable data transfers
  - Event dispatch mechanism
  - High data bandwidth

- Our implementation is based on **RoCEv2**
  - Other candidates: InfiniBand, iWARP (and even PCIe over cable)

- About RoCEv2:
  - RoCE = ***RDMA over Converged Ethernet***
  - Protocol **developed by Mellanox** (NVIDIA) for low-latency high-bandwidth communication in **data centers**
  - Supported in **Linux** (and all major OS)
  - Encapsulates RDMA packets in **UDP** datagrams (IP routable protocol)
  - Major manufacturers of network interfaces (NIC) **support RoCEv2** :
    - Intel, Broadcom, Mellanox, Marvell, …



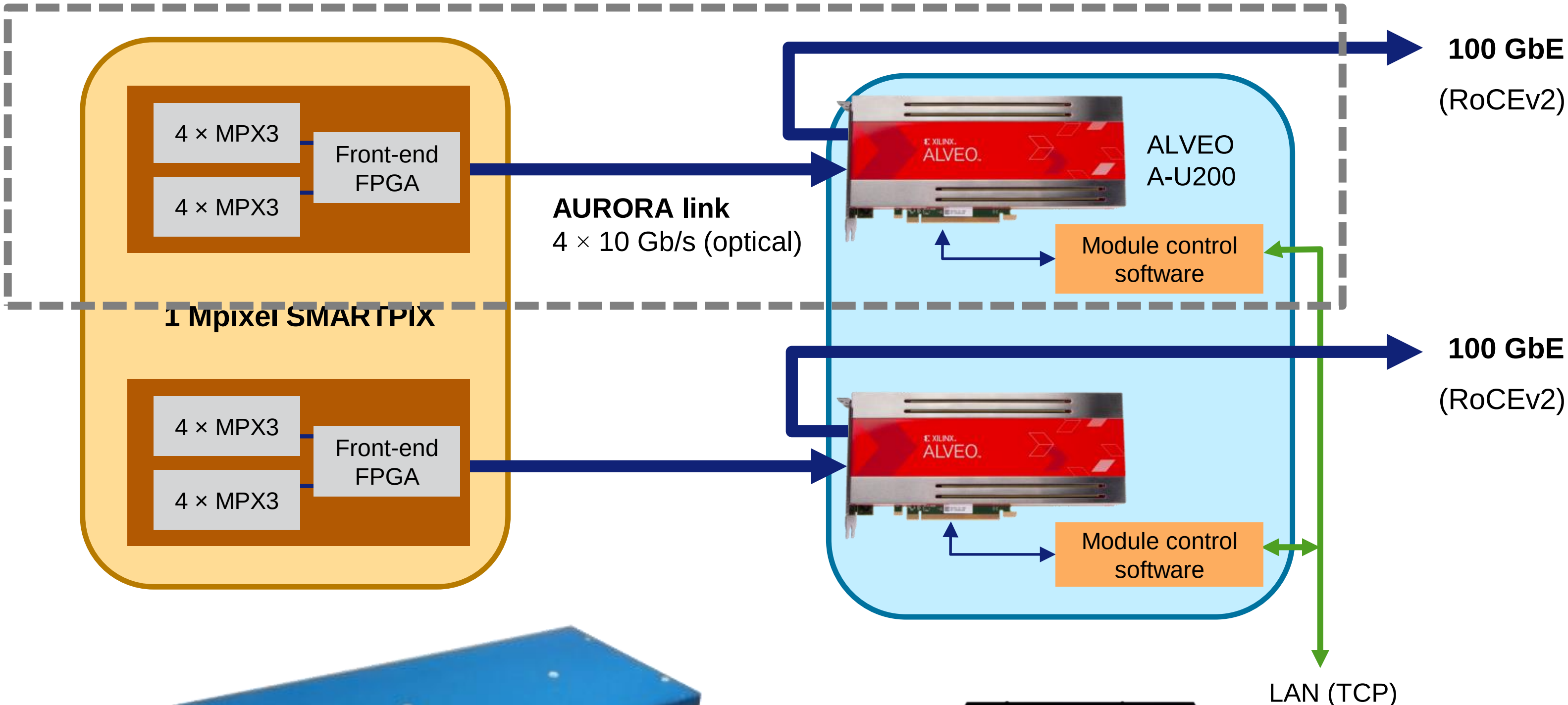
A RASHPA SYSTEM



THE DATA TRANSFER PROCESSES (DTP)

- Image frames (data blocks) are transferred from the detector modules to the destination buffers
  - The destination buffers can be **any (virtual) memory area** in the data receiver nodes… (RAM, GPU, …)
- A DTP must be configured by defining a set of rules:
  - **What to transfer**: which data blocks (images) , which fraction of the data block (ROI)
  - **Where**: which destination buffers in the data receivers and how to dispatch the data between them
- Several data transfer processes can run in parallel on the same data
  - **Multiple DTPs** allow to distribute image regions into groups data receivers, to separate processing
  - The same data can be retransmitted in more than one DTP
- The geometric reconstruction of image data from multiple modules is implicit in the DTP operation
- In an active DTP, new data is transferred as soon as they are available
  - Asynchronous **events** are issued to inform the data receivers and system manager of **progress and errors**
  - A 'data credit' mechanism is envisaged **to slow down** the data transfer (not implemented so far)

A RASHPA-BASED DETECTOR (EXAMPLE)

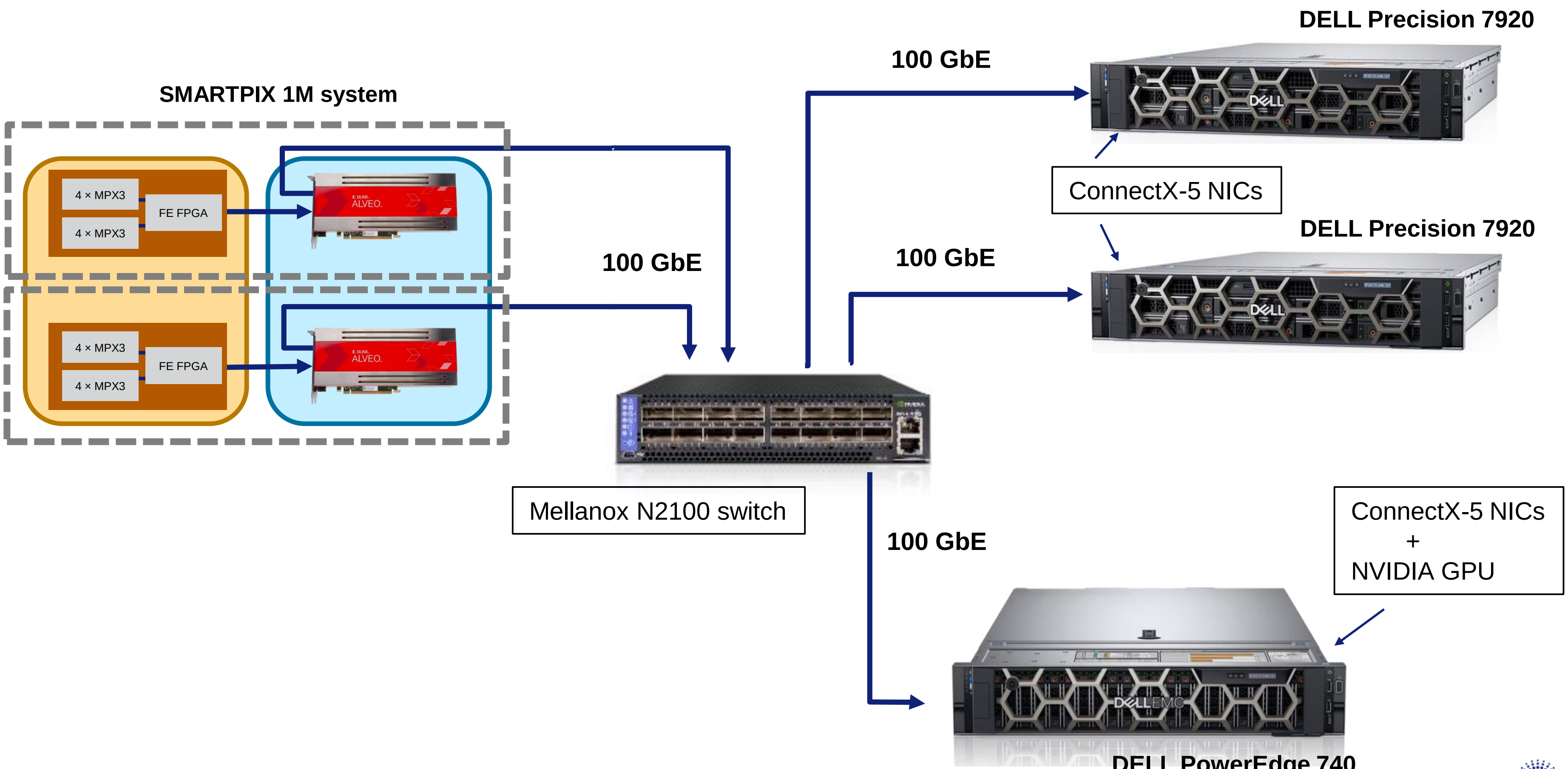


SMARTPIX 1M

3000 fps (6 GB/s)

NETSTOR enclosure

A RASHPA-BASED SYSTEM (EXAMPLE)



OTHER CONSIDERATIONS

Among the potentially useful aspects of RASHPA

- **Frees CPU resources** for other purposes (on-line data processing, compression, etc.)

- **Generic** design for standardization and reusability
  - Is **scalable** by construction
  - The current implementation (RASHPA controller and librashpa) is **not tied to a specific detector**
  - “RASHPA” **learns** about the detector and the computing node topology **at configuration time**

- **Low-latency** data transfer (sub-millisecond) opens the door to fast **experimental feedback**
  - For instance one could use data information to **change dynamically** the experiment conditions:
    - To correct the sample/beam position based on the image patterns
    - To reduce sample radiation damage (adapt the beam intensity and the scanning sequence)
  - A proof-of-concept at an ESRF beamline (microfocus diffraction) is under study
  - A simple demonstration of fast feedback capabilities was performed with an X-ray lab source by closing the position control loop of a mechanical device by means of a SMARTPIX/RASHPA GPU-based metrology chain.

THANK YOU FOR YOUR ATTENTION