Data transfer and DAQ system in ELI Beamlines

Pavel Bastl, Petr Pivoňka ELI Beamlines/Institute of Physics, Prague, Czech Republic

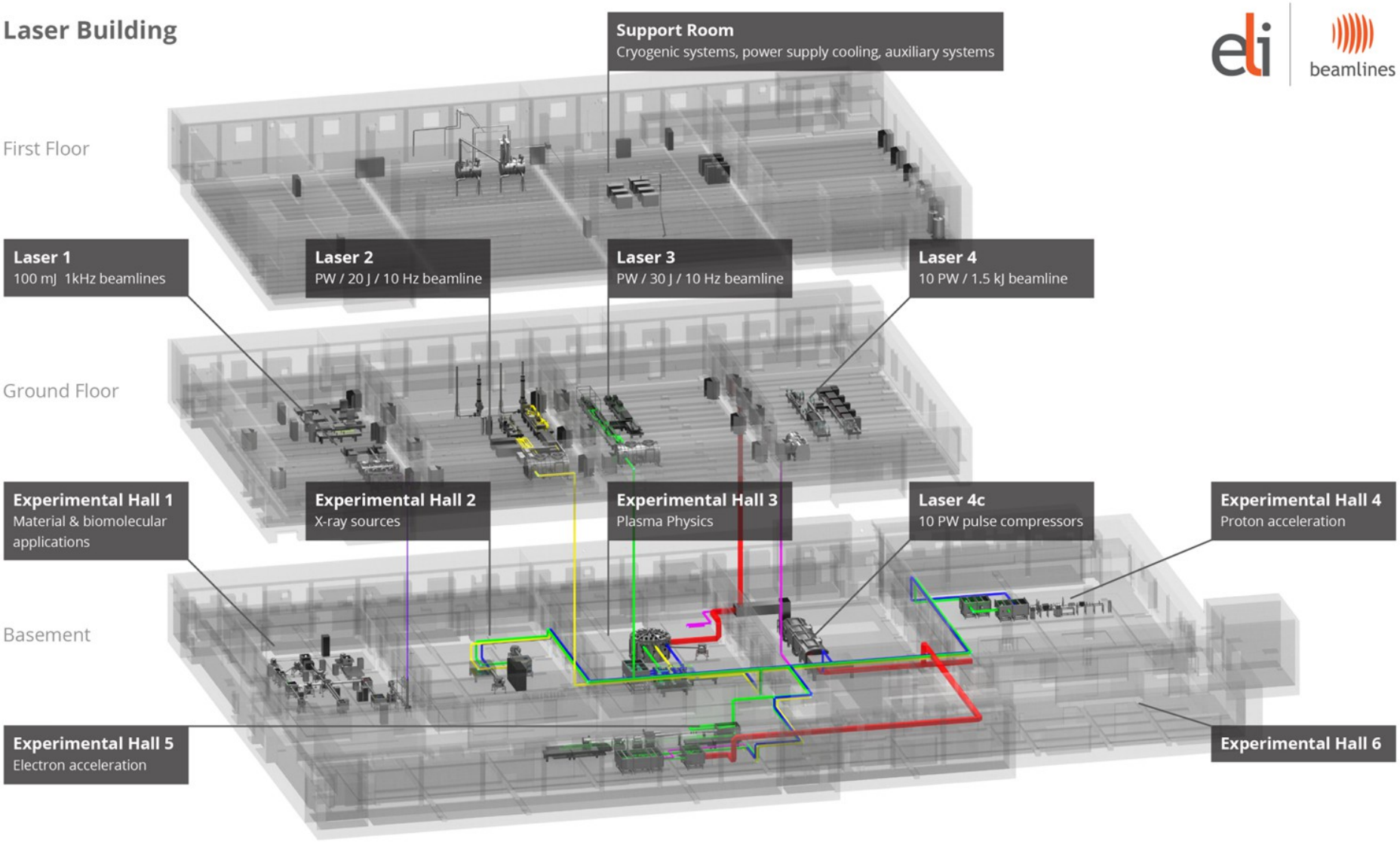
FZÚ

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

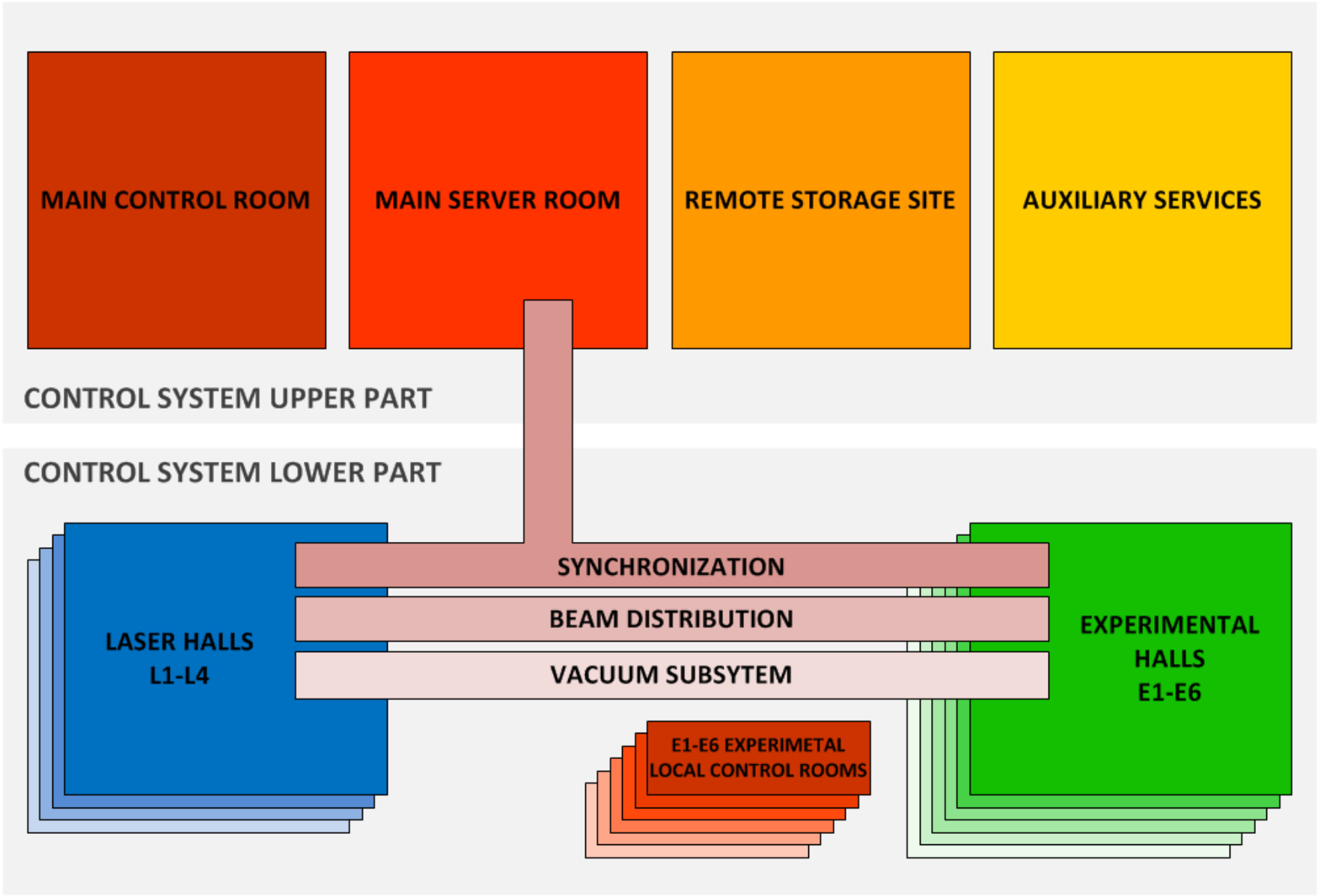
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

ELI Bemlines

ELI Bemlines



CS Structure



EUROPEAN UNION
European Structural and Investing Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION
YOUTH AND SPORTS

ELI DAQ Requirements

Data Transfer and DAQ – Operational environment

- High data rep rate
  - Single shot → 10Hz → 10kHz
- Two types of detectors
  - 1D detector → digitizers, high speed rep rates 10Gs/s multi channel, 2 bytes/per sample
  - 2D detectors → cameras etc, medium range of data throughput
- Distributed DAQ system
- EMP

ELI DAQ Requirements

Data transfer and DAQ – Basic requirements

- High speed data storage → First bunch
- Normal data storage → SSD/HDD
- Long term data storage → Tapes
- On-line data processing → GPU/FPGA
- Off-line data processing → HPC/GPU
- Synchronization with timing → Time stamps
- Data availability
  - For the users → Access to offline data (Data storage – SSD/HDD/Tapes), Access to online data (Screens)
  - For control system processes → here the online data processing and propper timing plays the key role

ELI DAQ Requirements

Data transfer and DAQ – Use case 1: High speed digitizer

- 2 channel 5Gs/s repetition rate → 10Gs/s
- 12 bits resolution → 2 bytes per sample
- 20GB/s raw throughput

Data transfer and DAQ – Use case 2: CCD/CMOS camera / 2D detector

- Up to 5MPixel maximal expected repetition rate 1kHz 5GPixels / s
- 12 bits resolution → 2 bytes per pixel
- 20GB/s raw throughput

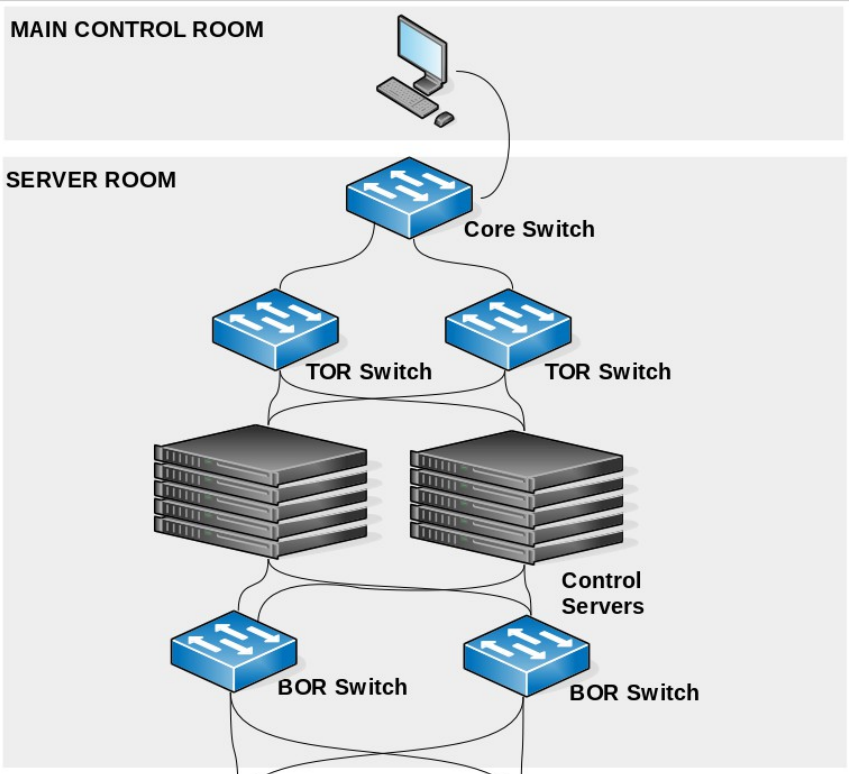
Devices can spread over ELI Beamlines facility based on particular application and experiment → We are not precluding concrete setup, the DAQ system shall be universal enough

Two main approaches to solve these requirements

- High speed low latency network with high data storage
- Online data processing

CS Overview

TOP LEVEL



Main Control room

CORE Switch – Cisco Nexus 7700, 10/40Gb/s, 100Gb/s ready

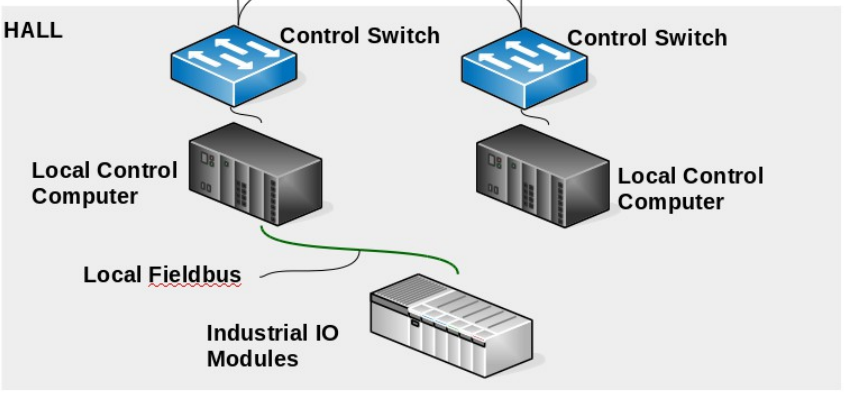
TOR Switch – Cisco Nexus 5672, 10/40Gb/s

Top level control – Server Room

- **Control servers** – Butch of 10 servers
- **Lenovo Systems x3650m5** 2U servers
  - 24 cores
  - 256GB RAM
- **Virtualization** – Private Cloud

BOR Switch – Cisco Nexus 56128, 10/40Gb/s

LOCAL LEVEL



CONTROL Switch – Cisco Catalyst 2960X, 1/10Gb/s

Local level control – Halls&Plant rooms

- **Industrial control**, undemanding applications
- **Advanced control**, challenging applications, with high demands on data rates

CS Local Control

µTCA Basic Components

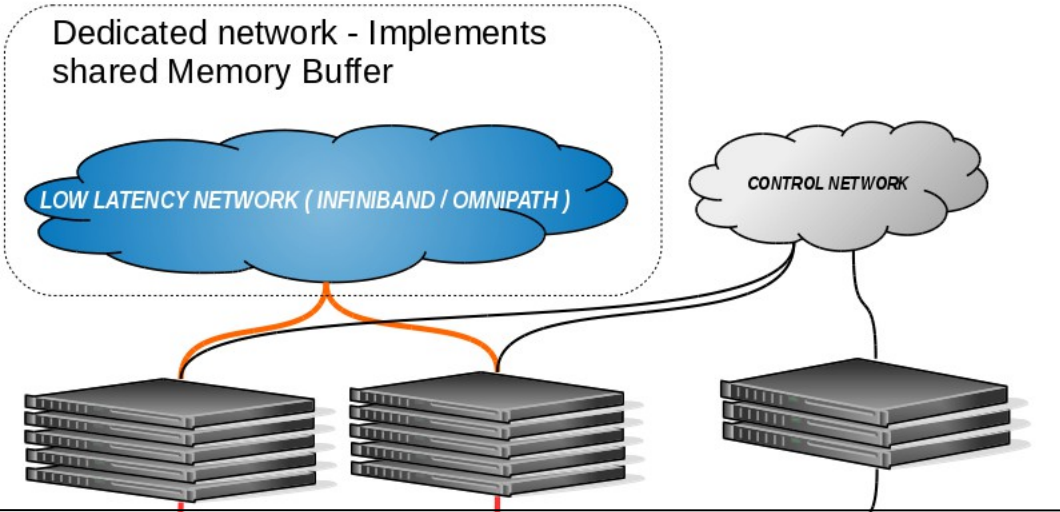
- **Chassis –** part which defines the size of the system
- **Backplane –** flexible architecture, 1Gb ethernet, 8 bidirectional fat-pipes, timing support, AMC-to-AMC and more
- **µTCA (MicroTCA Controller Hub) –** 'defines' the system, including fat-pipe implementation
- **Cooling Unit –** up to two fully controlled redundant units
- **Power Supply –** up to two redundant fully controlled power modules, up to 1kW
- **AMC/RTM Cards –** application cards, more than one computer system can be installed in chassis

Three Levels of µTCA in ELI Beamlines

- **Class A (µTCA.4)** – NATIVE R10-WR – High Demand Control, 12AMC, 12RTM, Redundant MCH, Redundant Power Supply
- **Class B (µTCA.4)** – NATIVE R2-WR – Advanced Control, 6AMC, 4RTM
- **Class C** – NATIVE A1 – 'Standard' local Controller, 6AMC, Mid-Size, Single-Width

DAQ Overview

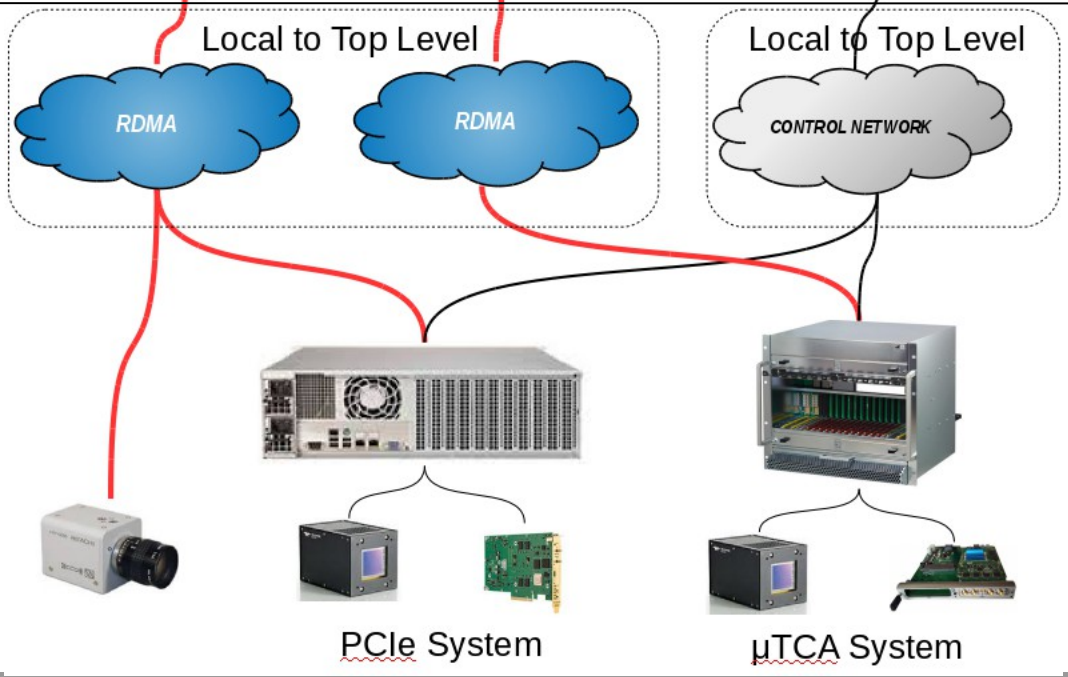
TOP LEVEL



Top Level of DAQ System

- 2x Blade Server
  - 10Gb/s ethernet (mezzanine) each
  - 56Gb/s Infiniband (mezzanine) each
  - 24 cores each Blade
  - 768GB RAM each Blade
  - 4x PCIe slot (2 slots x8, 2 slots x16)

LOCAL LEVEL

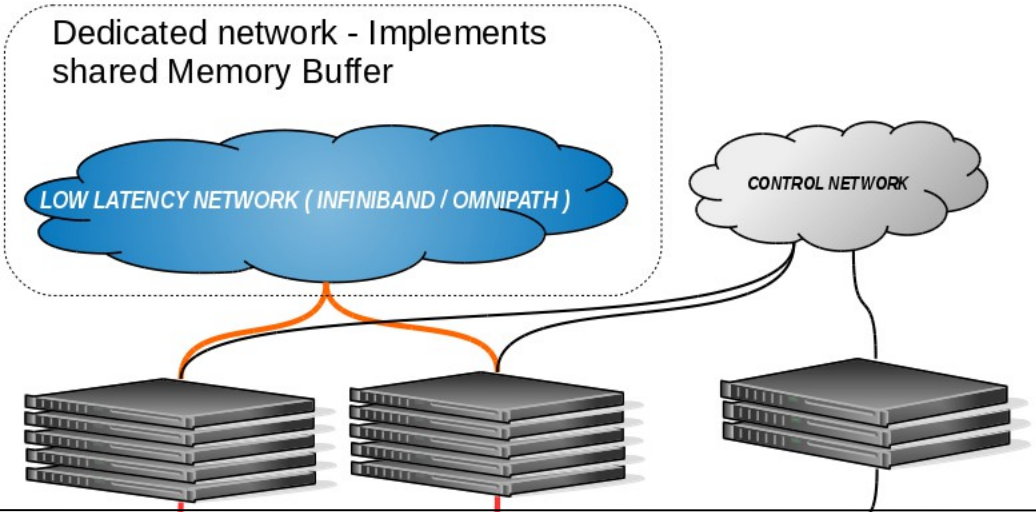


Local Level of DAQ System

- Local PCIe Server
  - 4x 10Gb/s ethernet with RDMA
  - 24 cores
  - 128GB RAM each
  - 10x PCIe slots x8
- µTCA.4
  - 12x AMC, 12x RTM
  - 4x 10Gb ethernet to RDMA
  - MCH-PHYS-80 PCIe on Fat-Pipes, PCIe connection with PCIe local server through optical fibers

DAQ Overview

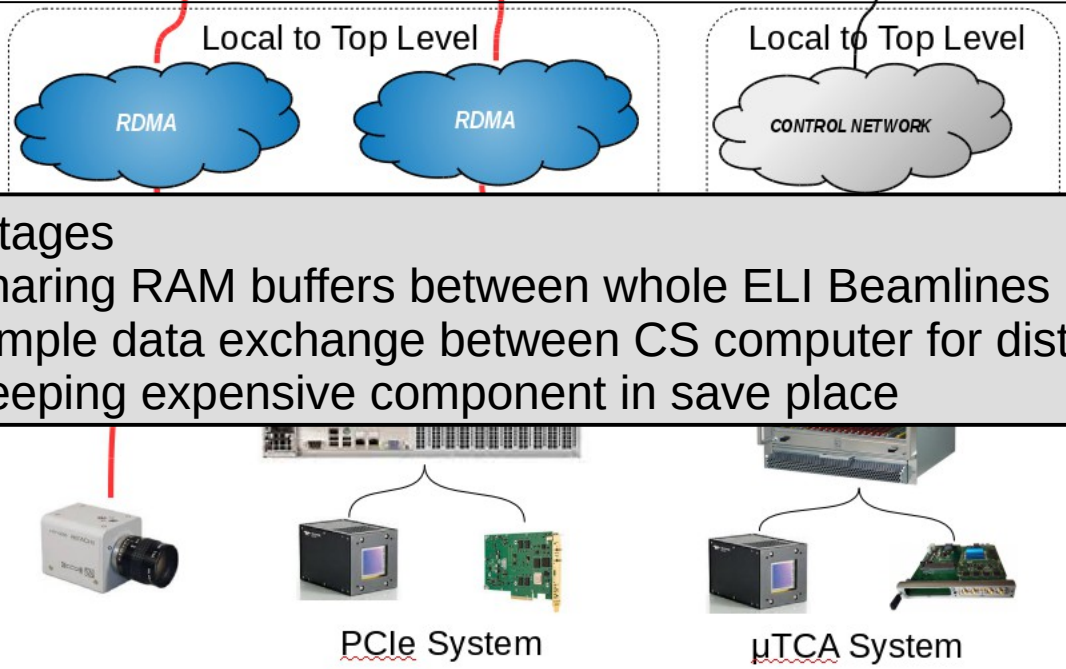
TOP LEVEL



Top Level of DAQ System

- 2x Blade Server
  - 10Gb/s ethernet (mezzanine) each
  - 56Gb/s Infiniband (mezzanine) each
  - 24 cores each Blade
  - 768GB RAM each Blade
  - 4x PCIe slot (2 slots x8, 2 slots x16)

LOCAL LEVEL



Local Level of DAQ System

- Local PCIe Server
  - 4x 10Gb/s ethernet with RDMA
  - 24 cores
  - 4x 10Gb ethernet to RDMA
  - MCH-PHYS-80 PCIe on Fat-Pipes, PCIe connection with PCIe local server through optical fibers

Advantages

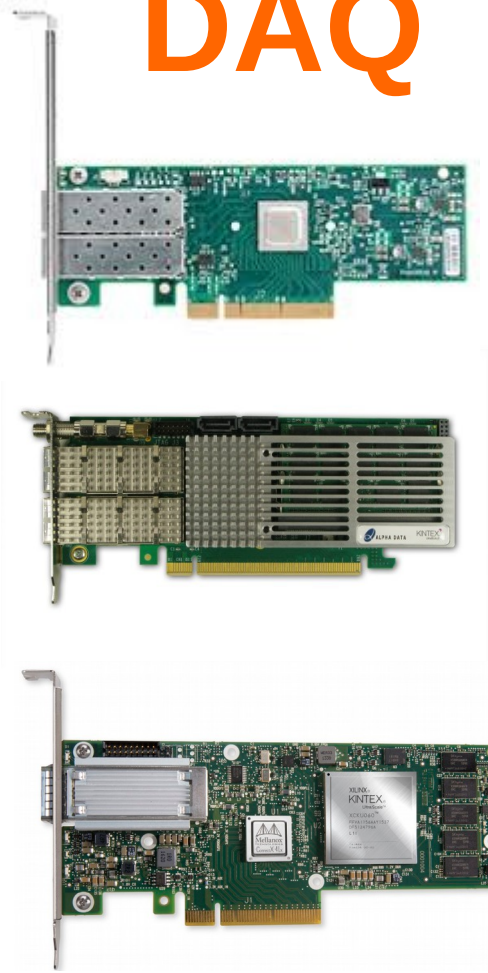
- Sharing RAM buffers between whole ELI Beamlines building
- Simple data exchange between CS computer for distributed control
- Keeping expensive component in save place

DAQ Top Level

Blade Server With PCIe Slots

Infiniband and Ethernet Switches are integrated inside the chassis.

+



NIC – Our 'standard'
- 2x 10GBASE-X
- RDMA Support

FPGA XCKU060
- 2x QSFP
- SDAccell support
- CAPI support
- 8GB DDR RAM

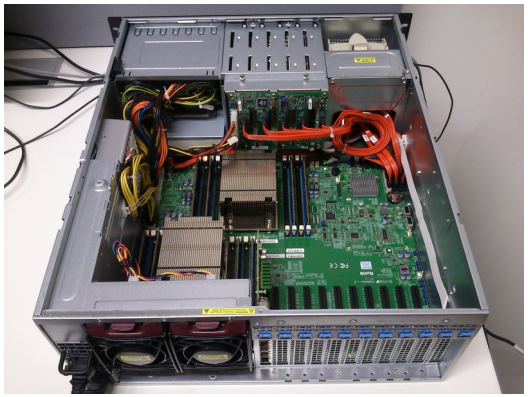
FPGA+NIC
- 1X QSFP (4xSFP+)
- Connect-X NIC Offload chip
- RDMA Support
- XCKU060
- 2GB DDR RAM

DAQ Local Level

PCIe System

PCIe Local Server

- 24 core
- 128GB RAM
- 10x PCIe x8



NIC AOC-STG-b4S

- 4x 10GBASE-X
- RDMA support

FMC Carrier

Kintex UltraScale XCKU085

- Acquisition and local processing
- VITA 57.4 compatible with 57.1
- 16GB DDR RAM
- 1x SFP+

µTCA.4 System

µTCA.4 Chassis

- 12 AMC
- 12 RTM
- WR Support
- MCH-PHYS-80
- PCIe interconnection

FMC Carrier

Artix/Kintex XC7A200, XC7K325

- Acquisition and local processing

Compatible RTM

Module

- 8x SFP+

NIC Vadatech AMC211

- 2x 10GBASE-X

DAQ Local Level

µTCA.4 Digitizer ADQ7

- 2 channels 5Gs/s or 1 channel 10Gs/s
- 14 bit resolution
- Includes SDK & Necessary IP cores
- XILINX FPGA with user IP core support
- 4Gbytes of internal buffer

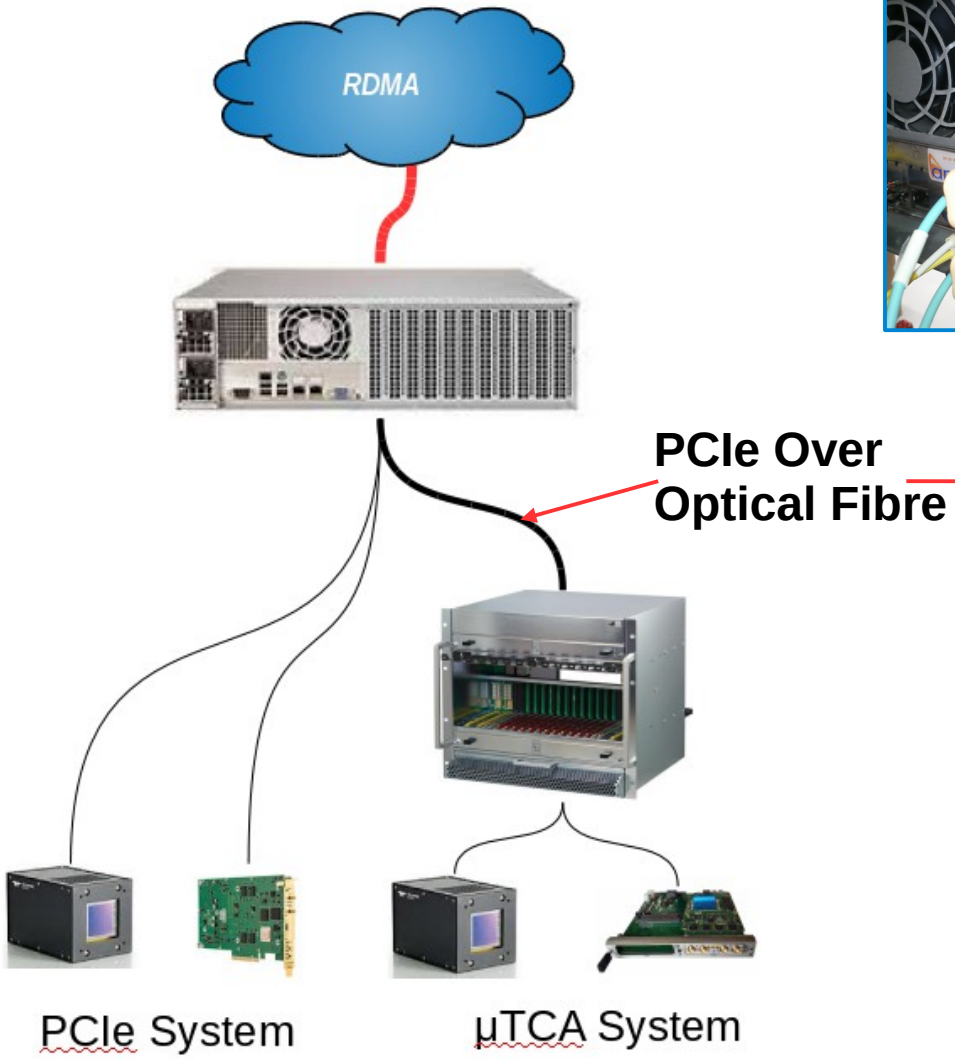
10Gb interface

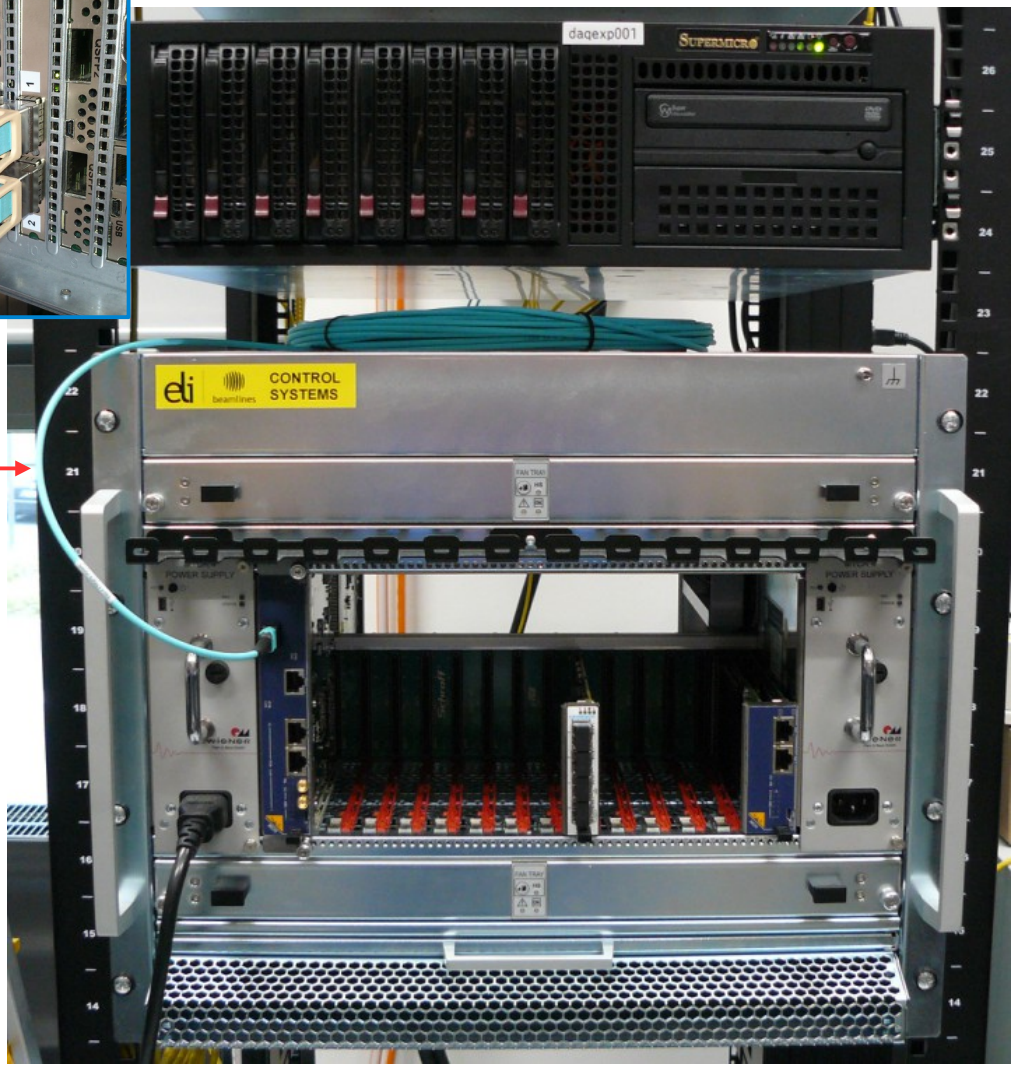
- Direct integration to ELI DAQ system

1Gb interface

- Timing system support

DAQ Local Setup



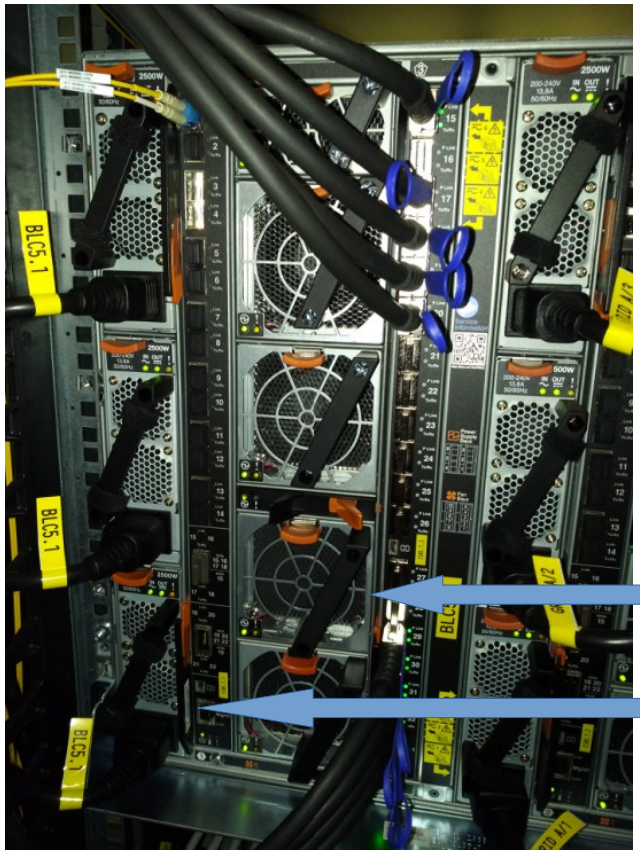


CS & DAQ Installations

CS & DAQ Installations

Core Switch 40/10 Gb

TOR Switch 40/10 Gb



40/10 Gb Ethernet switch

56 Gb FDR Infiniband switch

CS & DAQ Installations

EUROPEAN UNION
European Structural and Investing Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION
YOUTH AND SPORTS

Further development

Coherent Interconnection / Interfaces

There are three different technologies

- CCIX – Cache Coherent Interconnect for Accelerators (https://www.ccixconsortium.com)
- GenZ - (https://genzconsortium.org)
- OpenCAPI – Open Coherent Accelerator Interface (https://opencapi.org)

Above mentioned technologies provide tightly coupled interfaces between processor, accelerators and memory → **extension of currently used approach**

Comparison:

CCIX, GEN-Z, OpenCAPI: OVERVIEW & COMPARISON, by Bred Benton (AMD), OpenFabrics 13th Annual Workshop 2017

Superscalar Processor

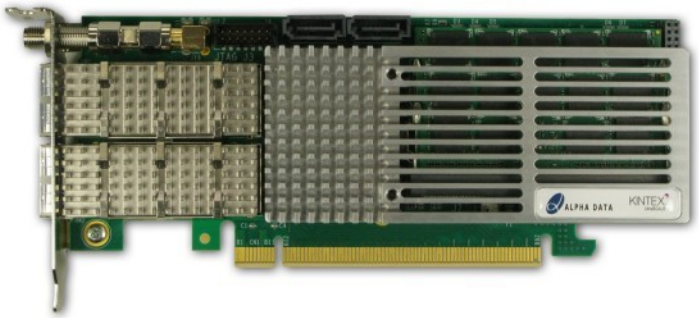
- Actually Power 8
- Power 9 processors are coming to the market

Further development

Actual status of technologies in ELI Beamlines

CAPI Based Hardware Installed

- CAPI Enabled card from Alpha Data (Including IP Cores and SDK)



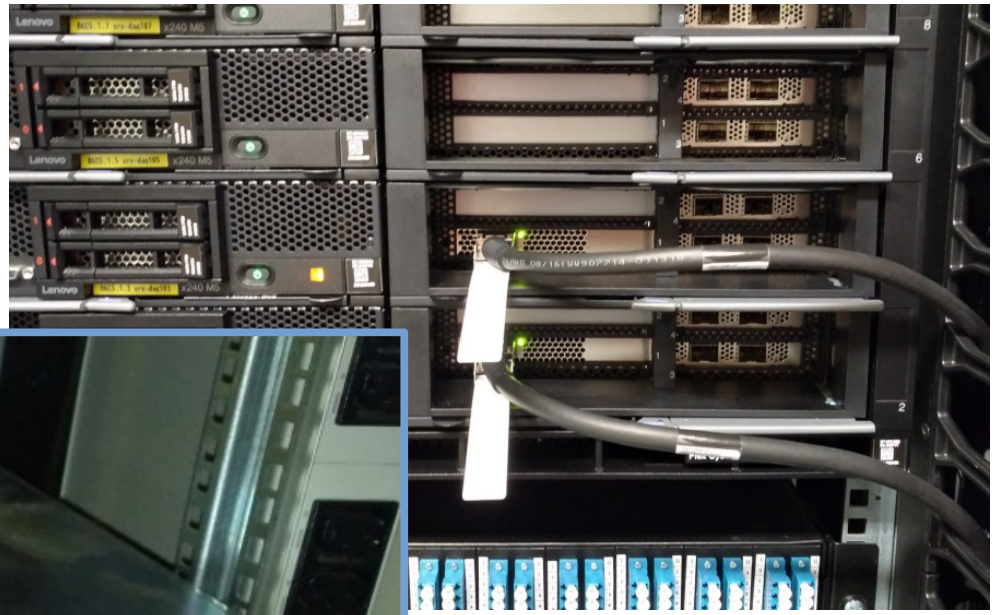
- CAPI Enabled Power server IBM (Power S812L)

Further development

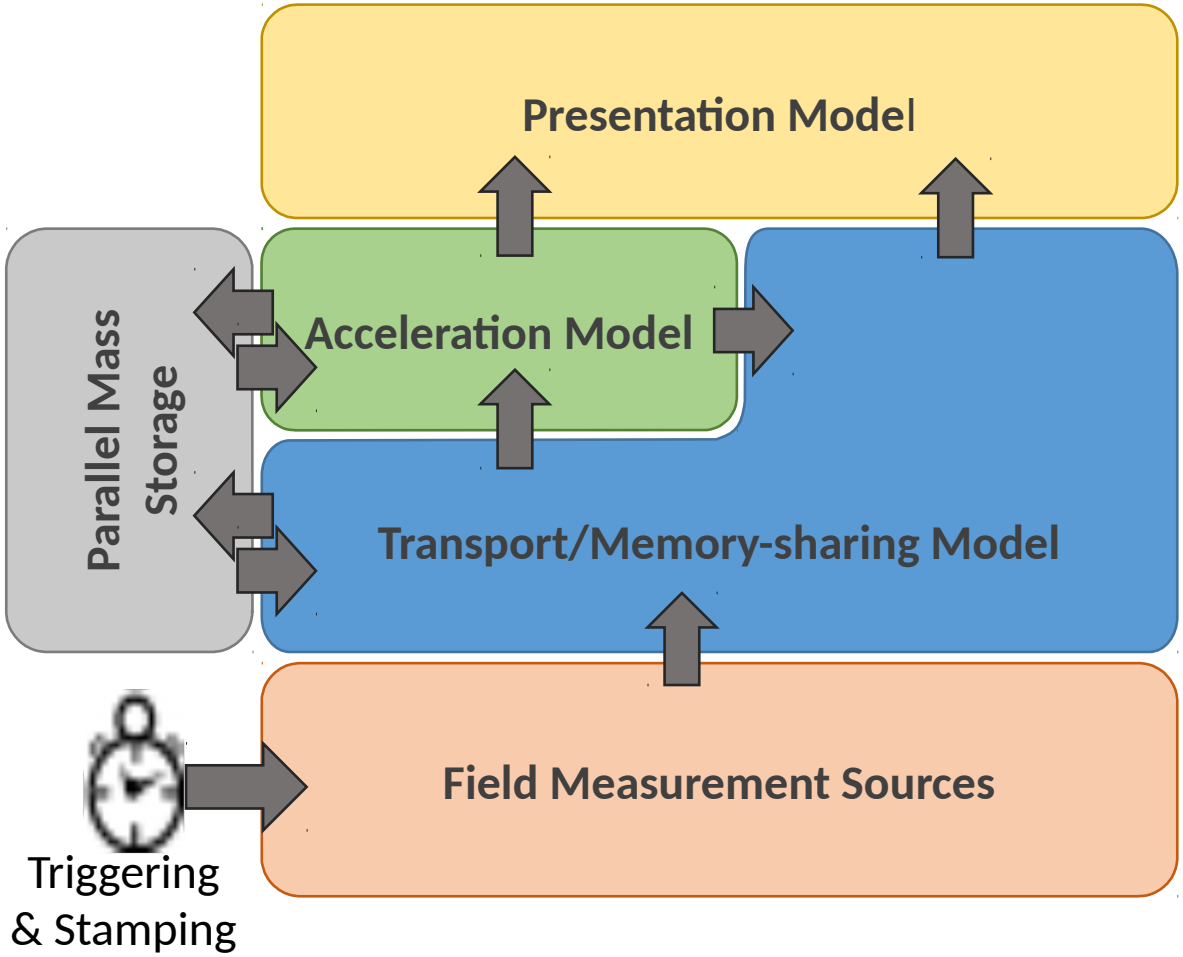
Actual status of technologies in ELI Beamlines

Mellanox Infiniband & Intel Omnipath

- 48ports / 100Gb Intel Omnipath switch for testing purposes



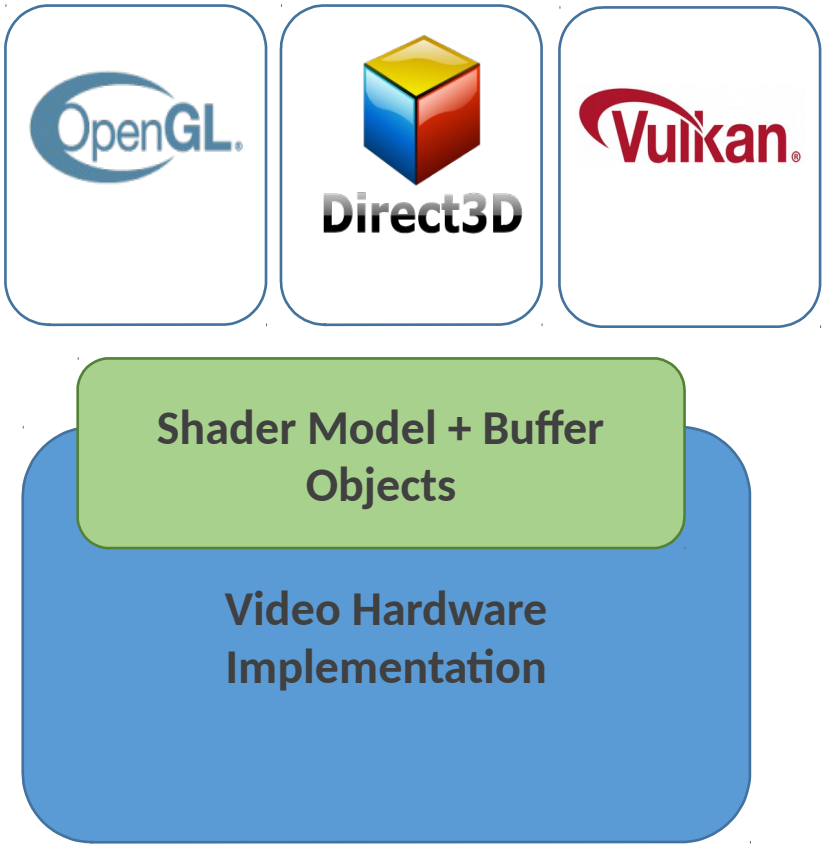
DAQ / Software Architecture



Key Features:

- Zero Copy
- Minimum Data Refactoring
- Lightweight Design
- Maximum Hardware Offload
- Low Latency Media

DAQ / Presentation Model



- For now we are going with OpenGL
- Future switch to Vulkan depends on its success (mixed reactions from the community so far ...)

Robust data adoption needs strict rules:

- do not use interleaved data
- use primitives with adjacency
- offload as much as possible (geometry shaders)
- OpenGL/OpenCL interops not always benefit

Discussion whether go with COTS or in-house dev:

- ParaView, VTK, ITK ... extensive frameworks with loads of dependencies left often unused
- few hundred lines of custom code may do things faster and easier to deploy

Thank you for attention!