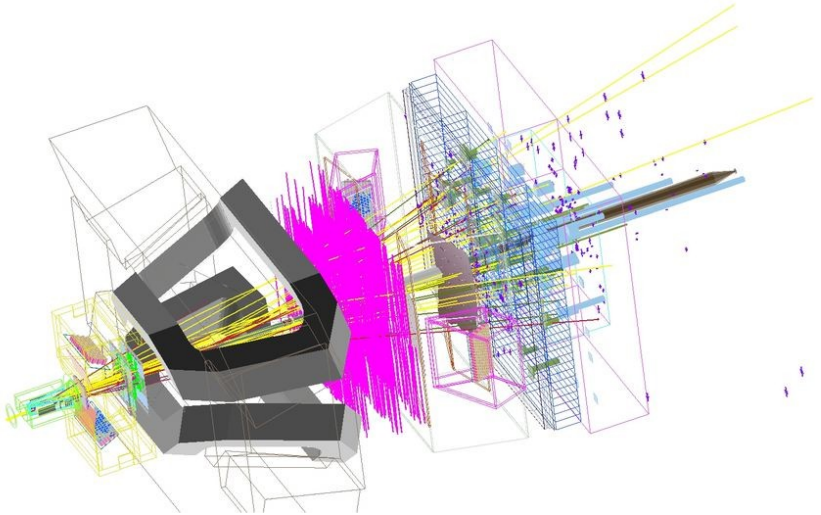
The LHCb PCIe Readout

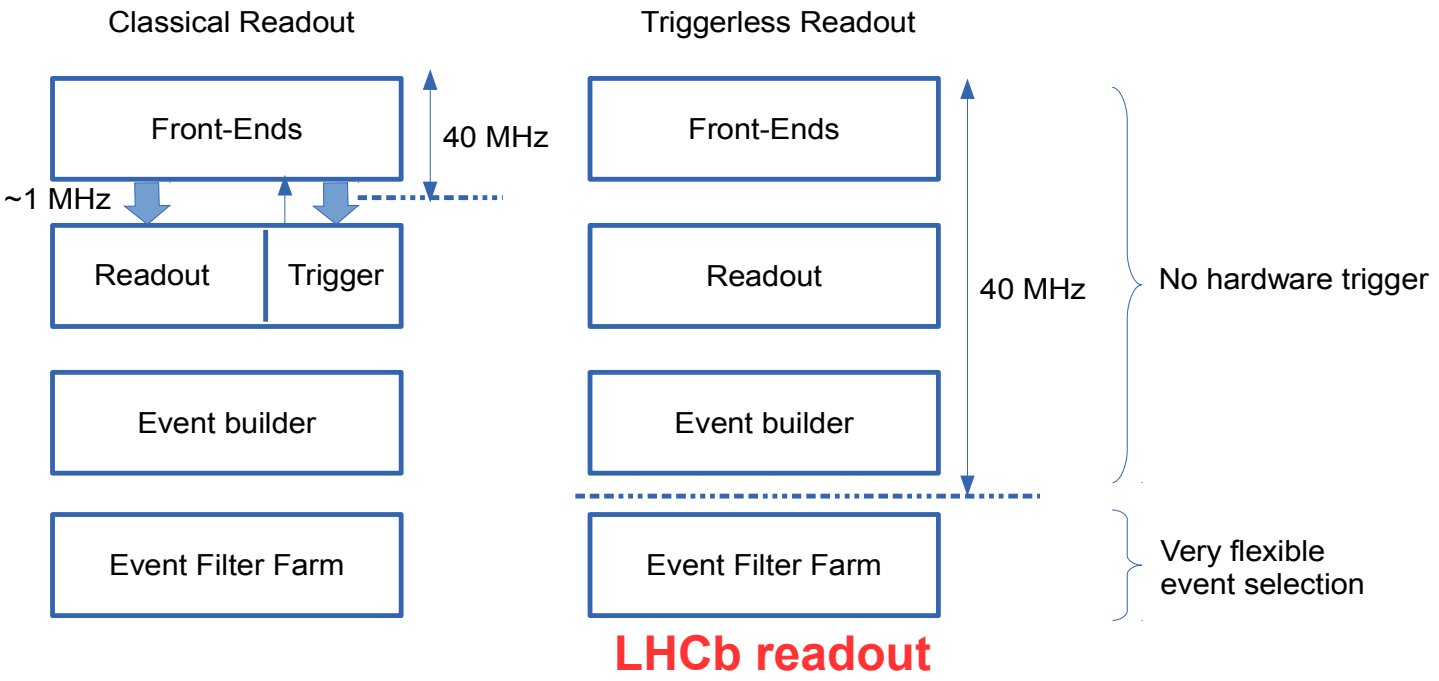


J.P. Cachemiche, on behalf of the LHCb collaboration

Outline

- LHCb PCIe Readout key points
- Hardware design
- Firmware design
- Perspectives

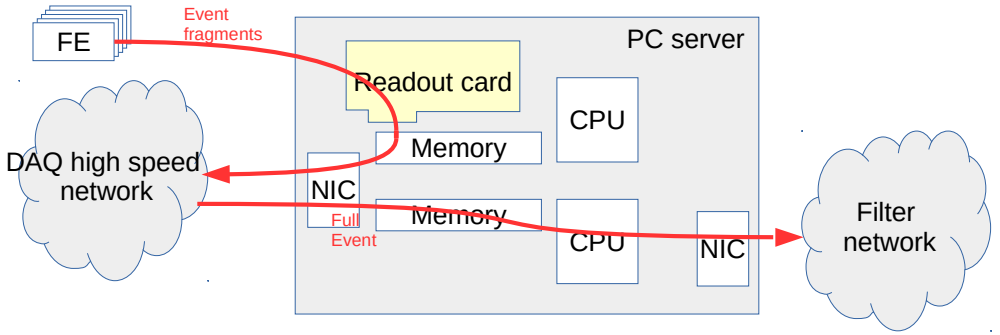
LHCb Upgrade key features



- LHCb uses a triggerless readout
- All event fragments routed at 40 MHz up to the farm

LHCb Upgrade key features

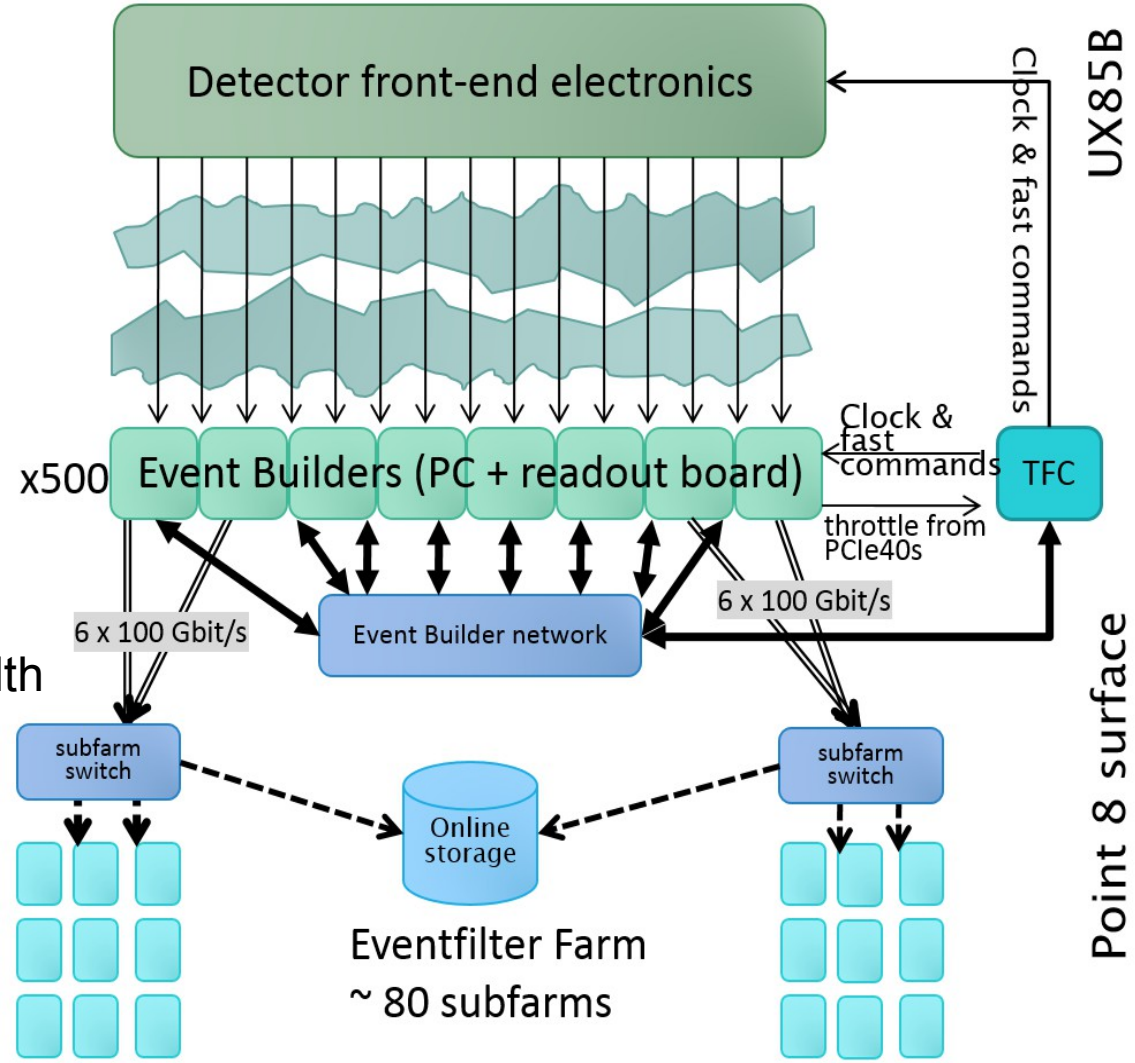
- Event building done by tightly coupled acquisition boards, CPUs and high speed network
- Event building and event reconstruction in real time
- Triggering replaced by filtering of reconstructed events
  - Now possible due to internal CPU architecture evolution and affordable high speed switches



- No intermediate back-end stage
- Readout card implemented as a PCIe card

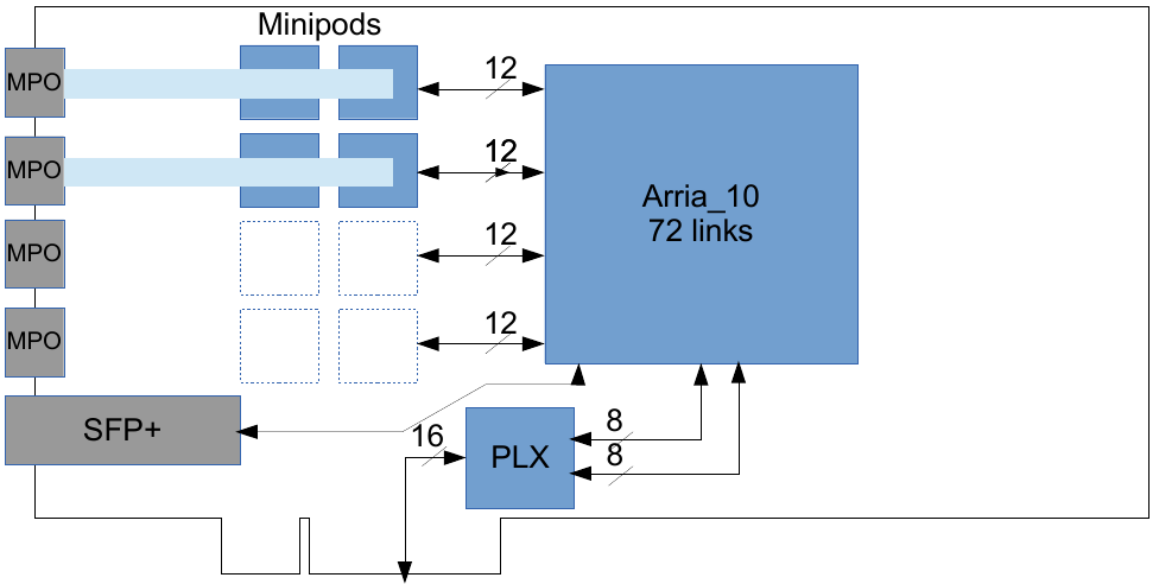
Architecture

- Readout located on surface
  - Distance between FE and RO : ~350m
- ~15000 optical links
- ~ 500 readout boards
- ~24 links in average on each board
- ~100 kbytes per event
- ~32 Tb/s aggregate bandwidth

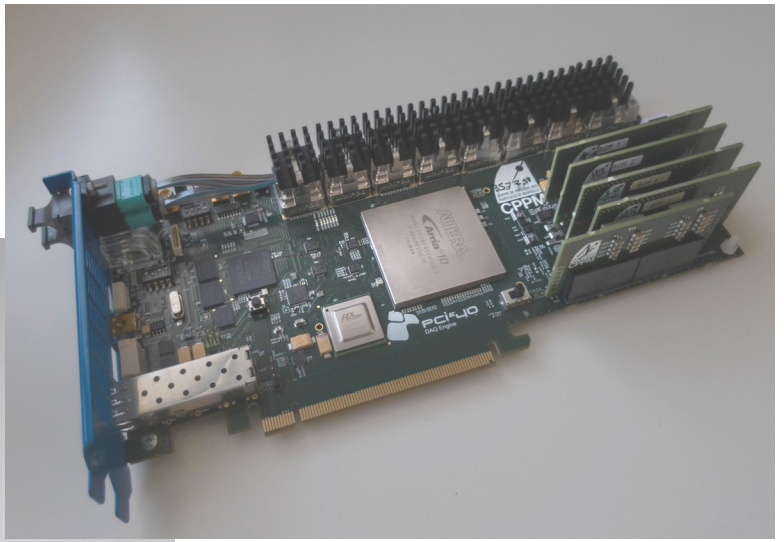


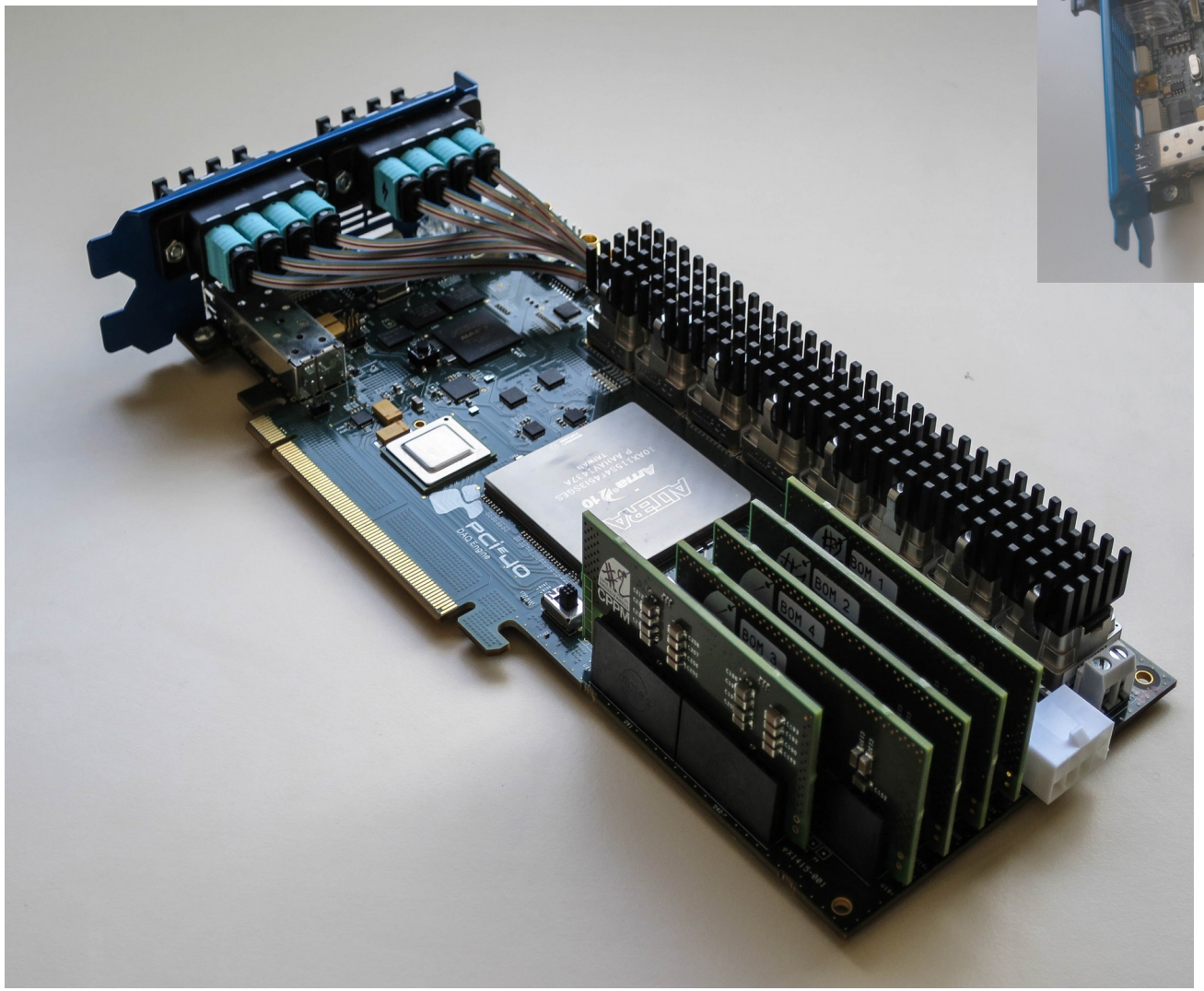
The readout board : PCIe40

- Features :
  - 48 bidirectional links running at up to 10 Gbits/s
  - 1 bidirectional link running at 10 Gbits/s devoted to time distribution (can use GBT or 10G PON protocol)
  - Sustained 100 Gbits/s interface with CPU through PCIe
  - 1 large FPGA 1.15 million cells (Arria10 10AX115S4F45E3SG)
  - Fast reprogramming through PCIe
- Can be mapped over Acquisition, Slow Control or Timing and Fast control by reprogramming the FPGA



PCIe40 prototype



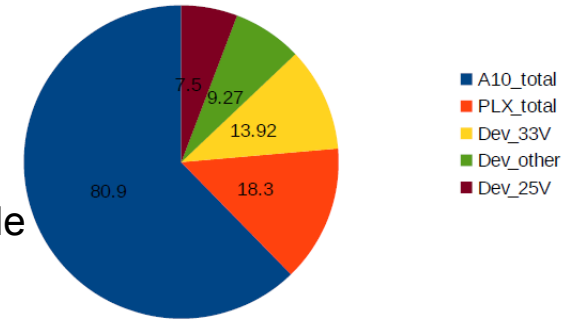


Hardware design

Power supply

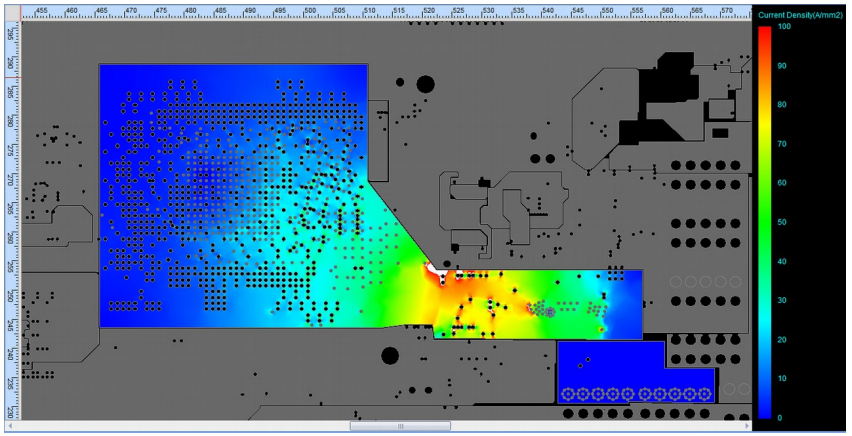
Power consumption of large FPGAs very high

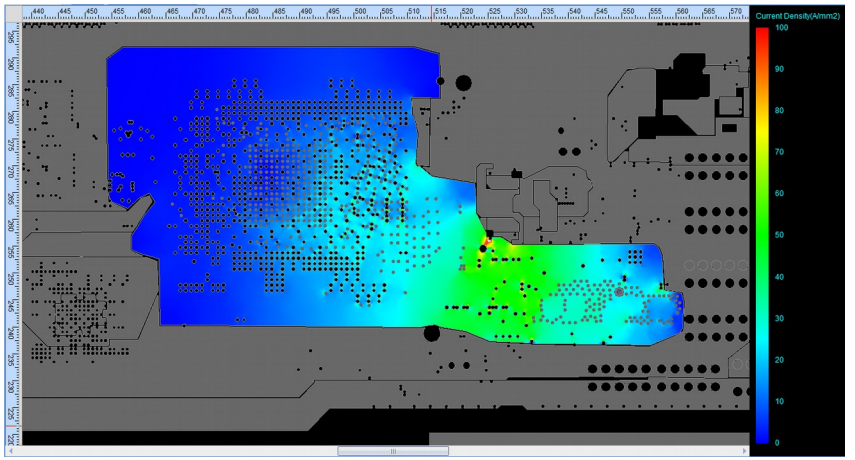
- Up to 60 A on the core !
- Power consumption
  - FPGA estimated at ~ **80 W**
  - Card estimated at ~ **157 W** with Engineering Sample
- Limited thickness for the stackup



Power consumtion estimation
100 % logic cells, 50 % toggle rate, 250 MHz operation ES chip

- Requires careful power supply design
- Simulation of current flow in power planes, vias, thermal drains





After optimization

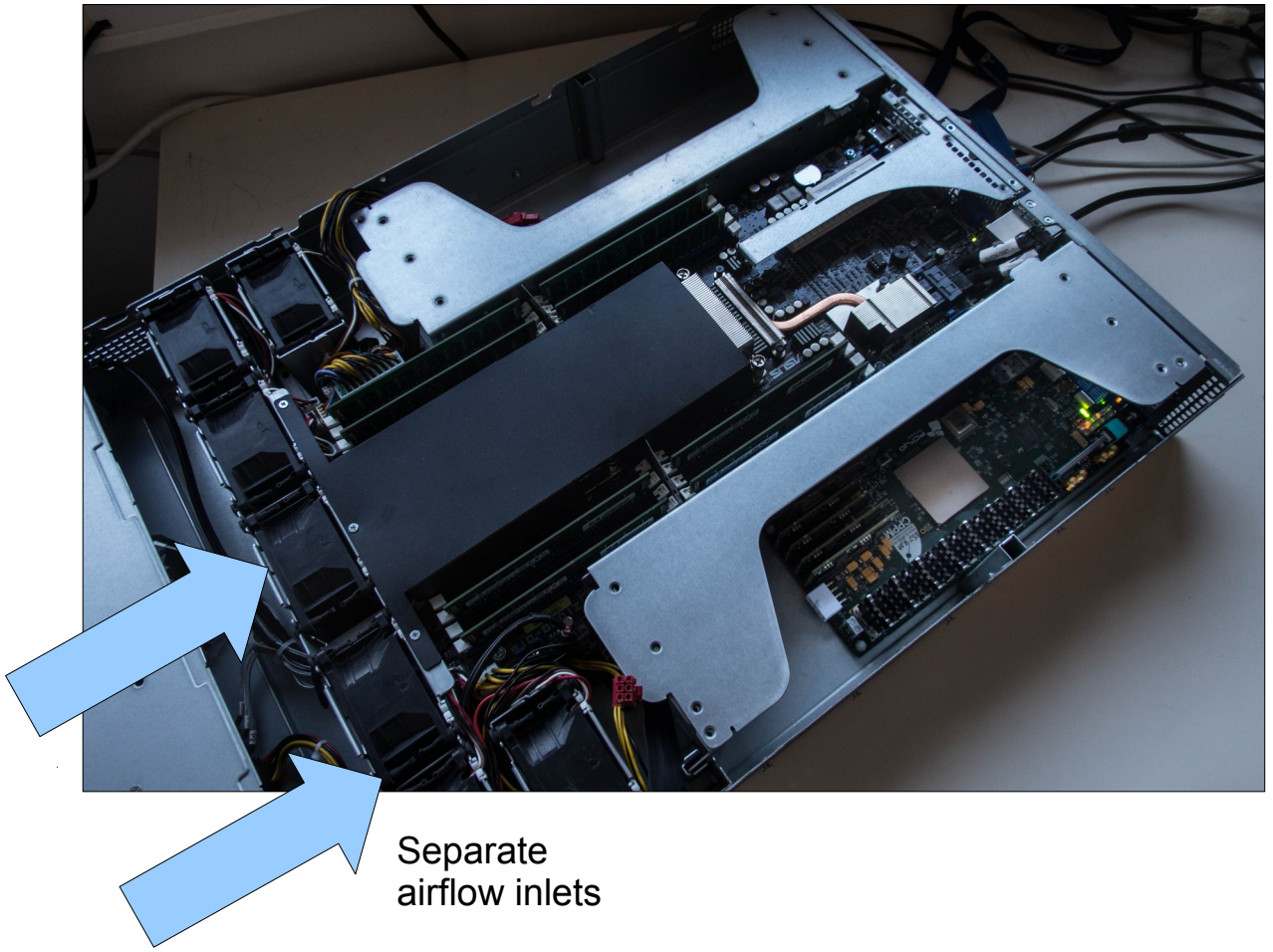
PCIe validation

- No physical loopback possible
- No measurement points on PC side
- Signal path unknown
- Test qualification tools extremely expensive

- Pragmatic approach based on software BER measurements and eye diagram measurement with on-die instrumentation

Cooling

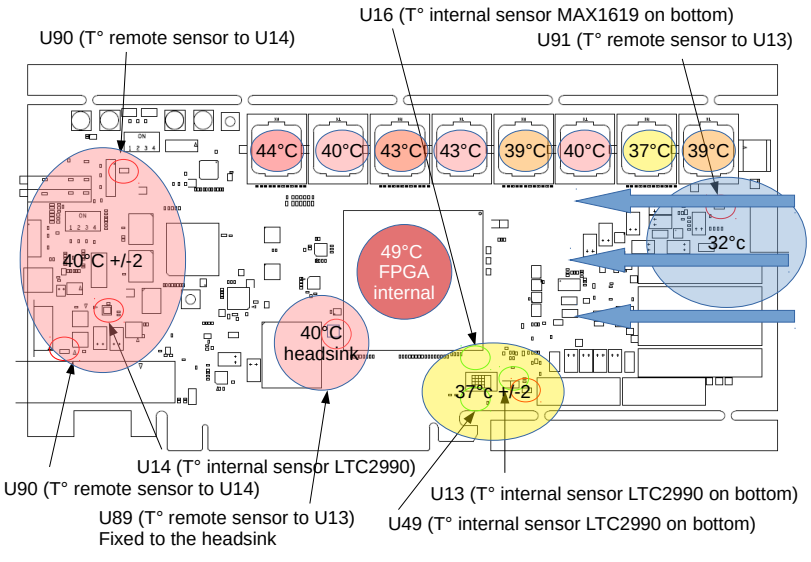
- PC environment not as well defined as xTCA systems
- Very well cooled PC server has been specified

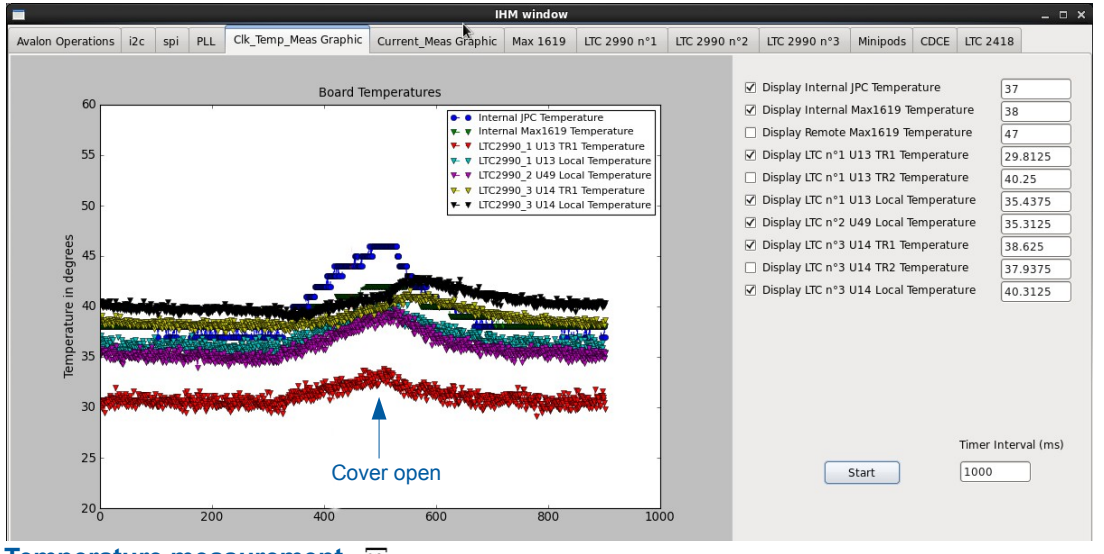


Cooling again ...

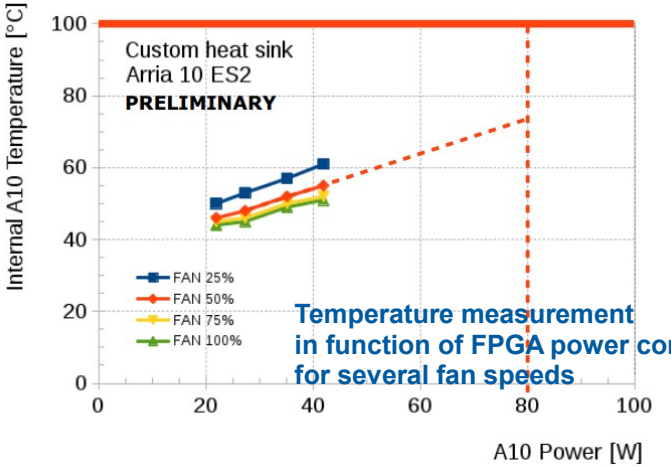
- Voltage, current and temperature sensors spread all over the board
- We can monitor the temperature vs power consumption vs airflow
- Will help to design a custom heat-sink dissipating up to 160 W

PCIe 40 Temperature sensors location





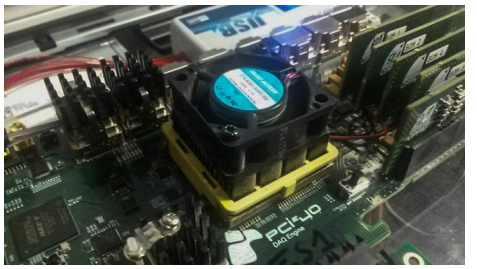
Temperature measurement on several sensors in function of time



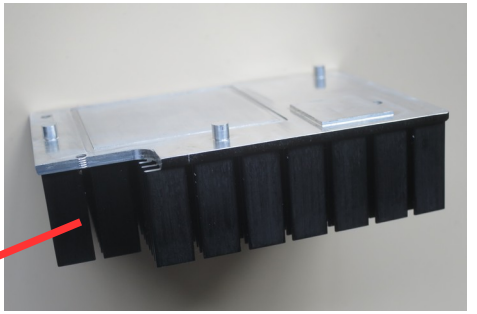
Temperature measurement in function of FPGA power consumption for several fan speeds

Cooling solutions

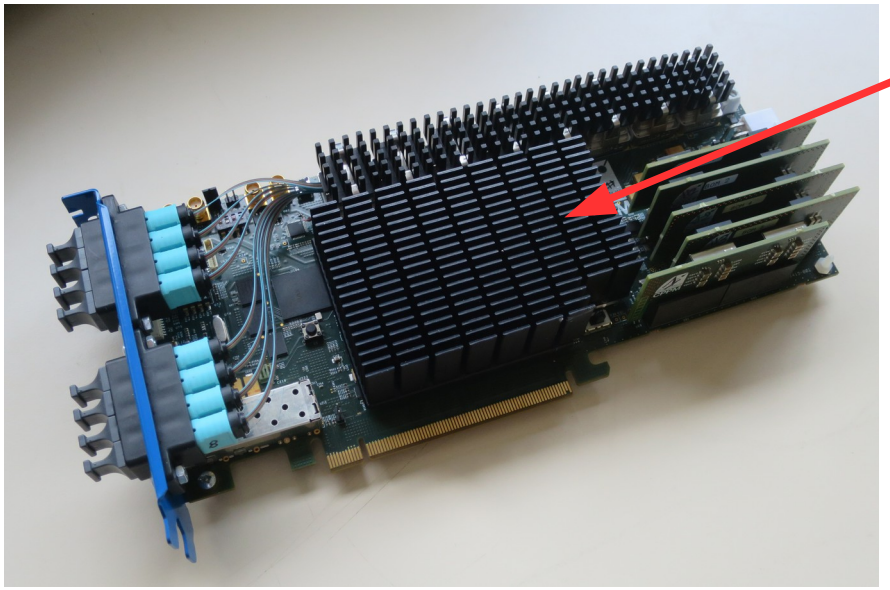
- Tests ongoing with several kinds of heat-sinks
  - Standard heat-sink with fan
  - Custom passive heat-sink
  - Custom liquid cooled heatsink
- Final solution will be decided after trials

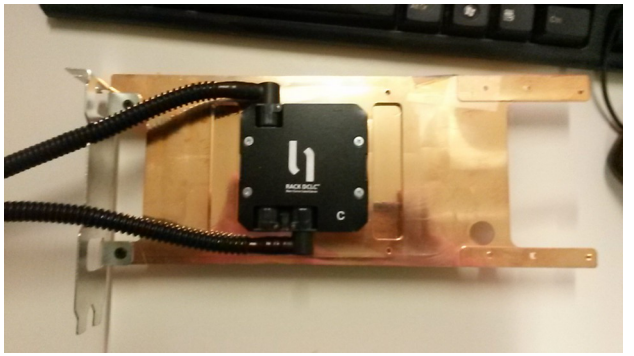


Heatsink with fan



Custom passive heatsink





Custom liquid-cooled heatsink

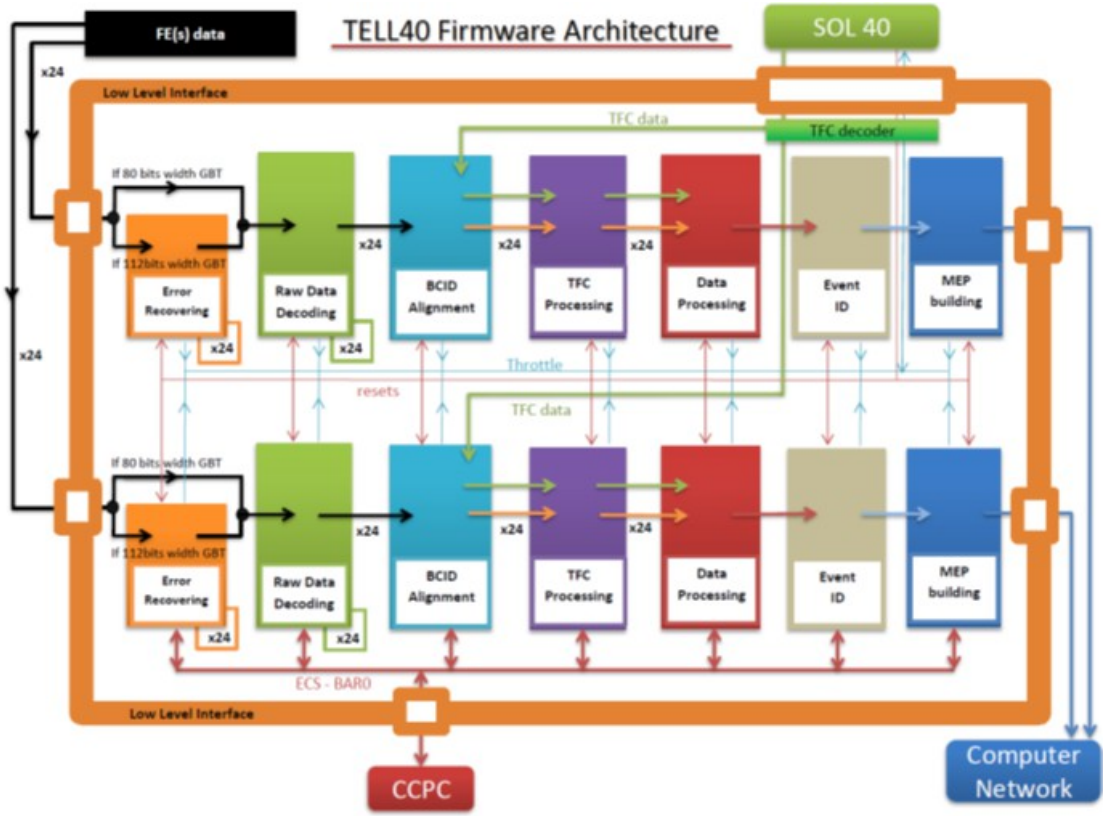
Firmware design

FPGA firmware development

- Core firmware design team spread over CPPM, LAPP, CERN
- Should be extended later on over 7 more groups
  - Requires specific methodologies and organization
    - Simulation environment to check every modification
    - Setup tested with realistic FE data simulation and data injection
    - Quality insurance by regression tests
    - Use of a collaborative code management tool : GitLab
    - Formal decision process before to push
    - Ticketting system for bugs

FPGA firmware layers

- Very large number of control registers (~10000) on the board
- All controls and initializations masked to the user by a hardware abstraction layer called LLI (Low Level Interface)
- User code could be run on any kind of FPGA
- First delivered MiniDAQ based upon Stratix V, Future one based on Arria10 : nearly no change for core code

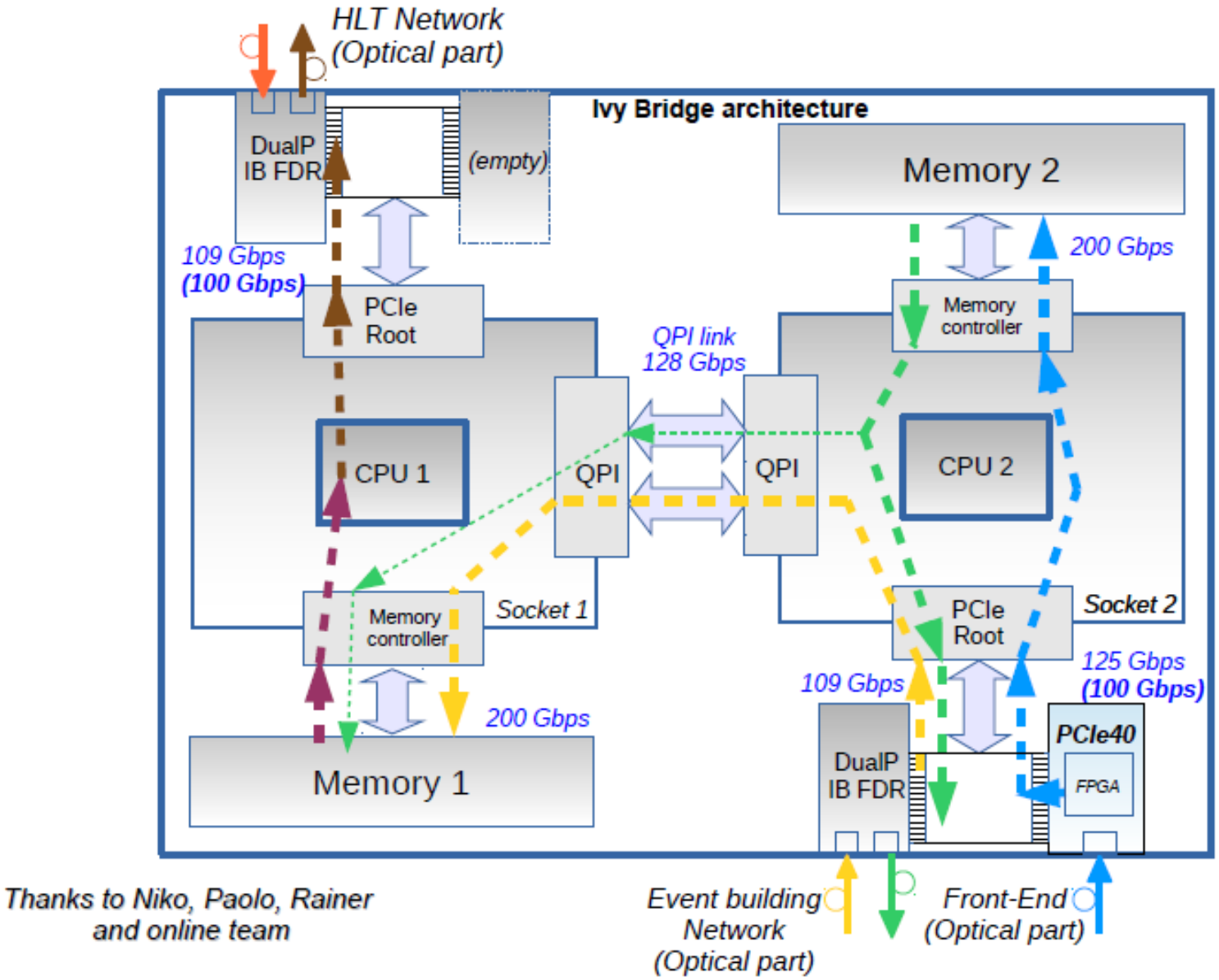


Perspectives and conclusion

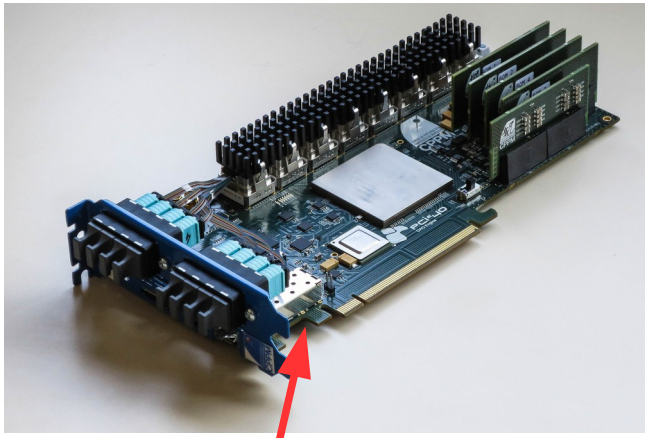
- 2 Engineering-Sample-based prototypes fully operational and fully qualified
- 25 cards currently under production :
  - Available mid-2016
- Methodologies to :
  - Develop and integrate a large number of firmwares
  - Automatically test a large number of cards
- We will go soon in an industrialization phase targetting a full production over 2017-2018
  - LHCb needs ~ 500 PCIe40 cards including spares for its readout
- Card selected by ALICE for its CRU (Common Readout Unit)

More information

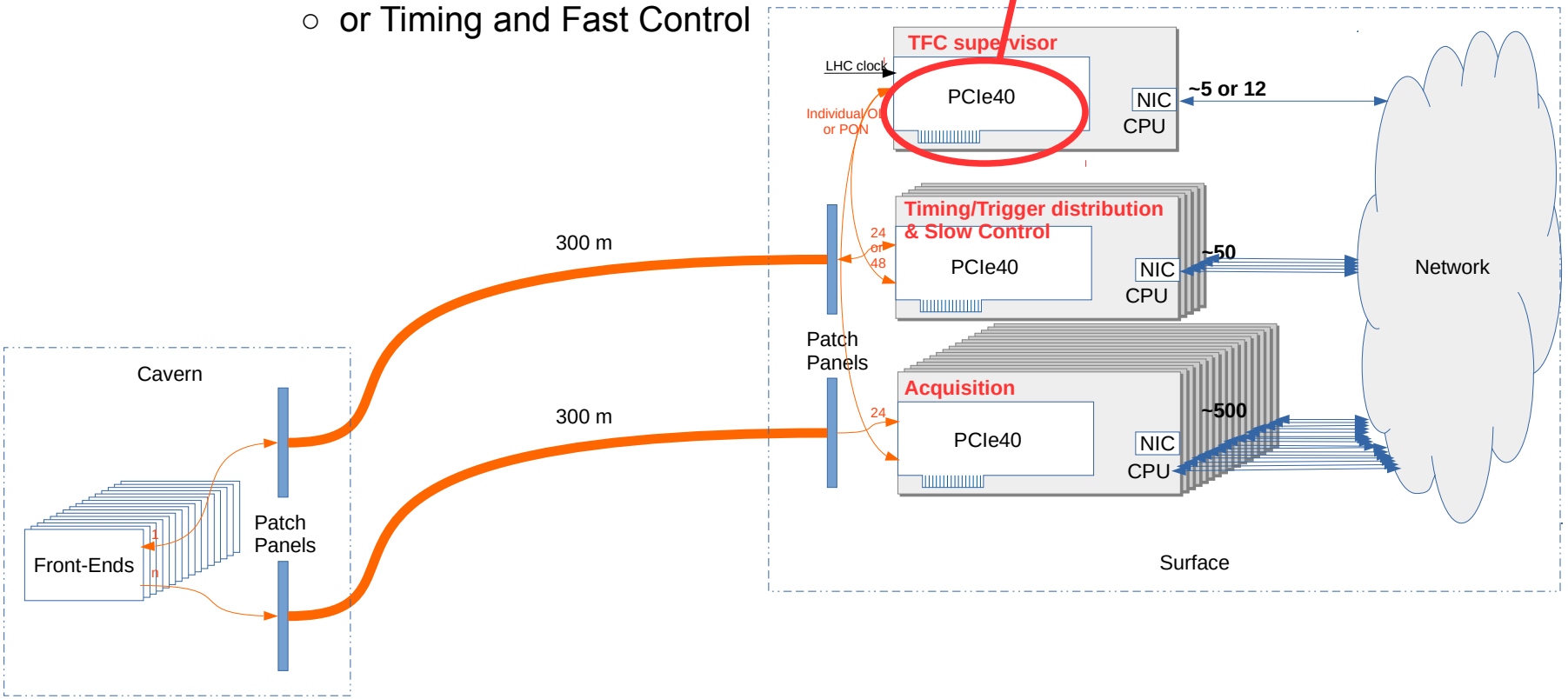
Data path in the computer



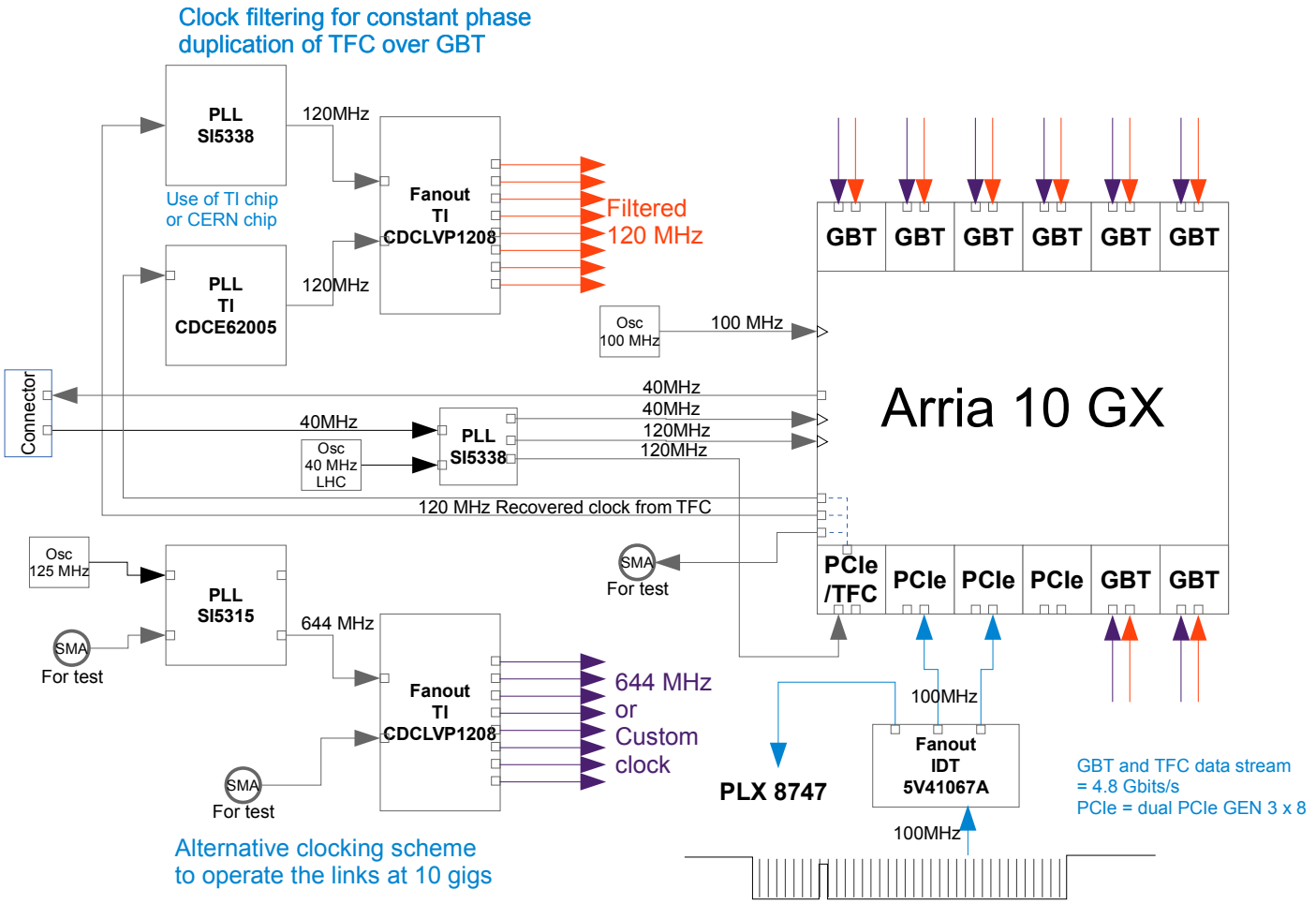
Instances



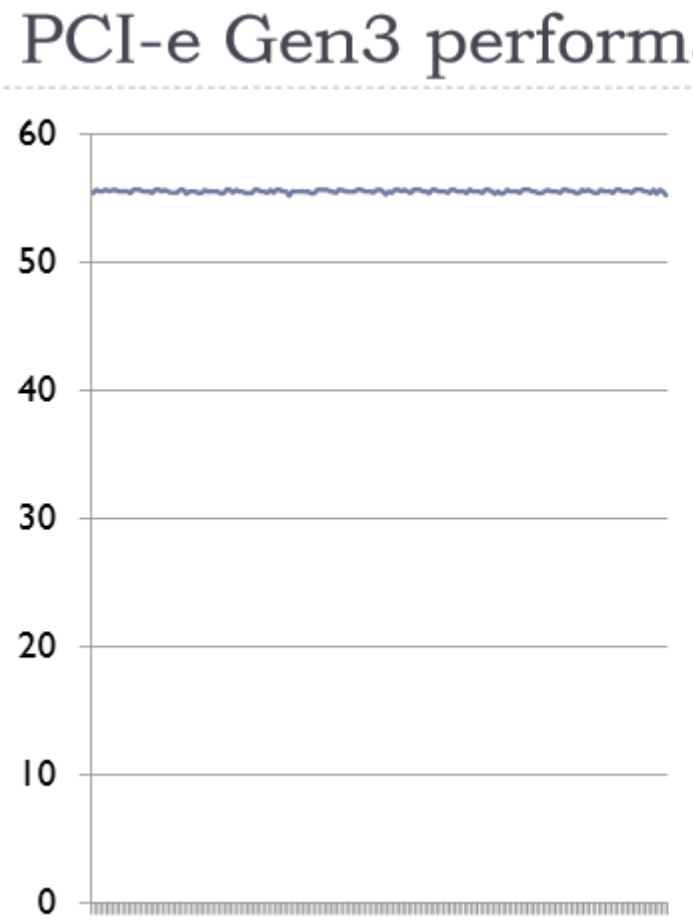
- Generic card for LHCb : PCIe40
- Same card used to implement
  - DAQ,
  - Slow Control
  - or Timing and Fast Control



Clock distribution



PCIe GEN3 performance

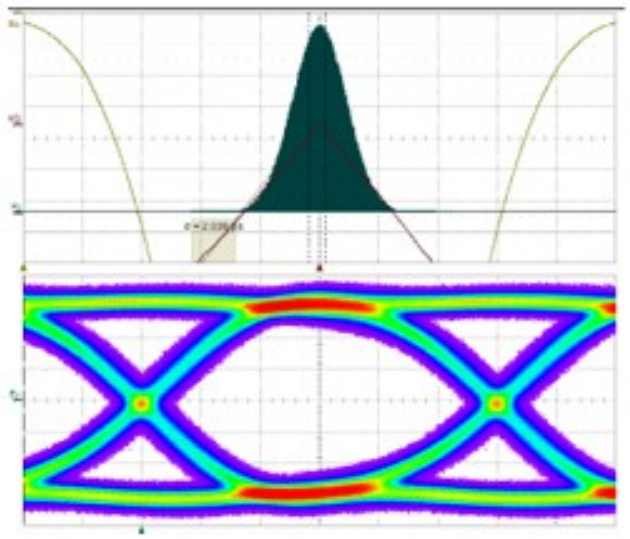


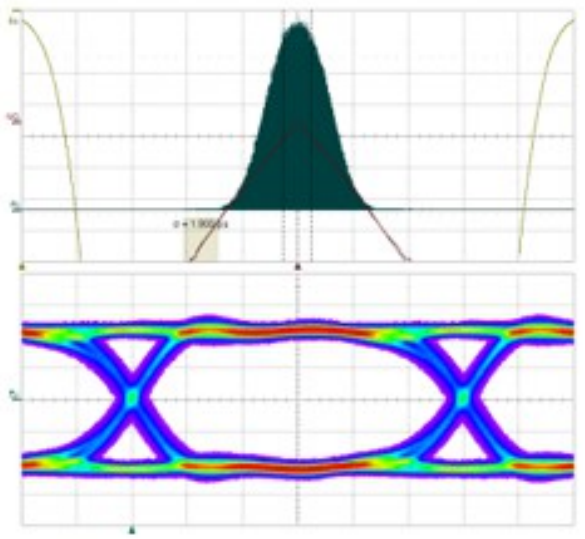
- Avg. BW **55.51 Gbps**
  - MSI overhead unaccounted for in previous calculation
  - Additional Hard-IP dead time unaccounted for in calculation
- Excellent performance stability over time

DMA memory requirements

- Events are committed to buffer in 32 Byte chunks
- Event header parsing + pass-through memory write
- FPGA buffer is divided in DMA sections
  - 4KiB sections → 88% bandwidth
  - ≥ 8KiB sections → 87% bandwidth
- 16 sections → just 64 KiB on the FPGA
  - DMA does not need an external memory interface

Eye diagrams





Measurements at 10.0Gbit/s
Total jitter ~ 36.82ps
Random Jitter ~ 2.22ps
Deterministic Jitter ~ 5.6ps

Measurements at 5.0Gbit/s
Total jitter ~ 37ps
Random Jitter ~ 2.1 ps
Deterministic Jitter ~ 3.44 ps