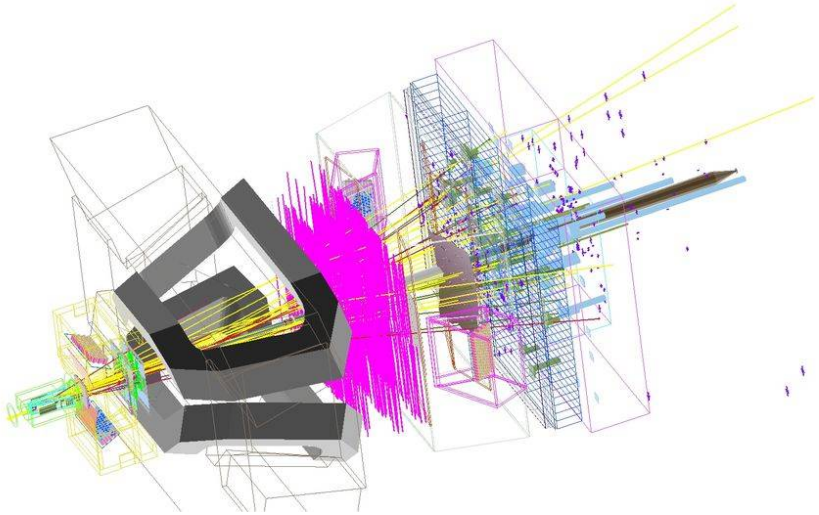
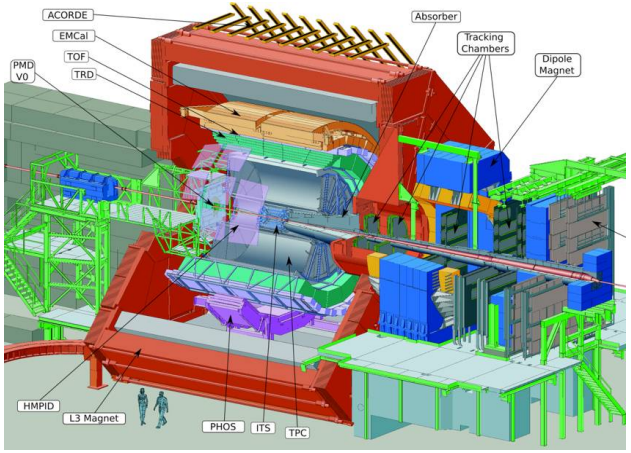
PCIe40: A Common Readout Board for LHCb and ALICE

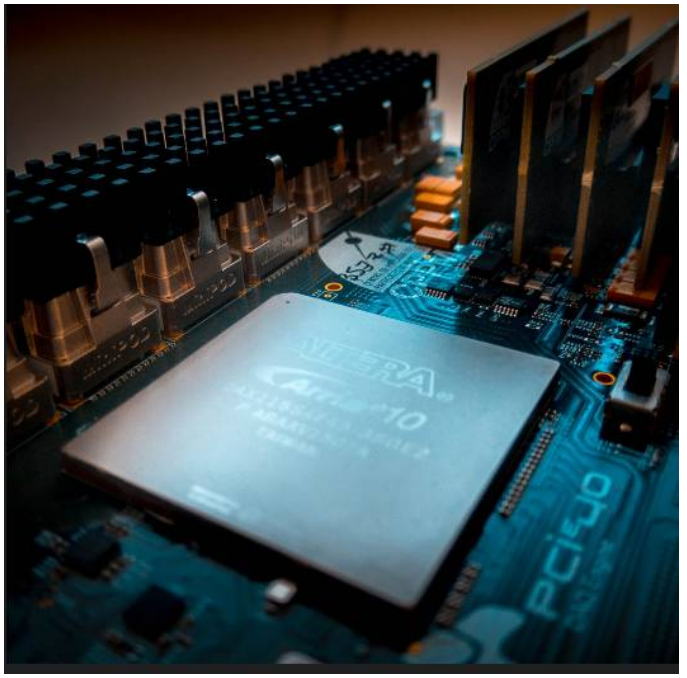




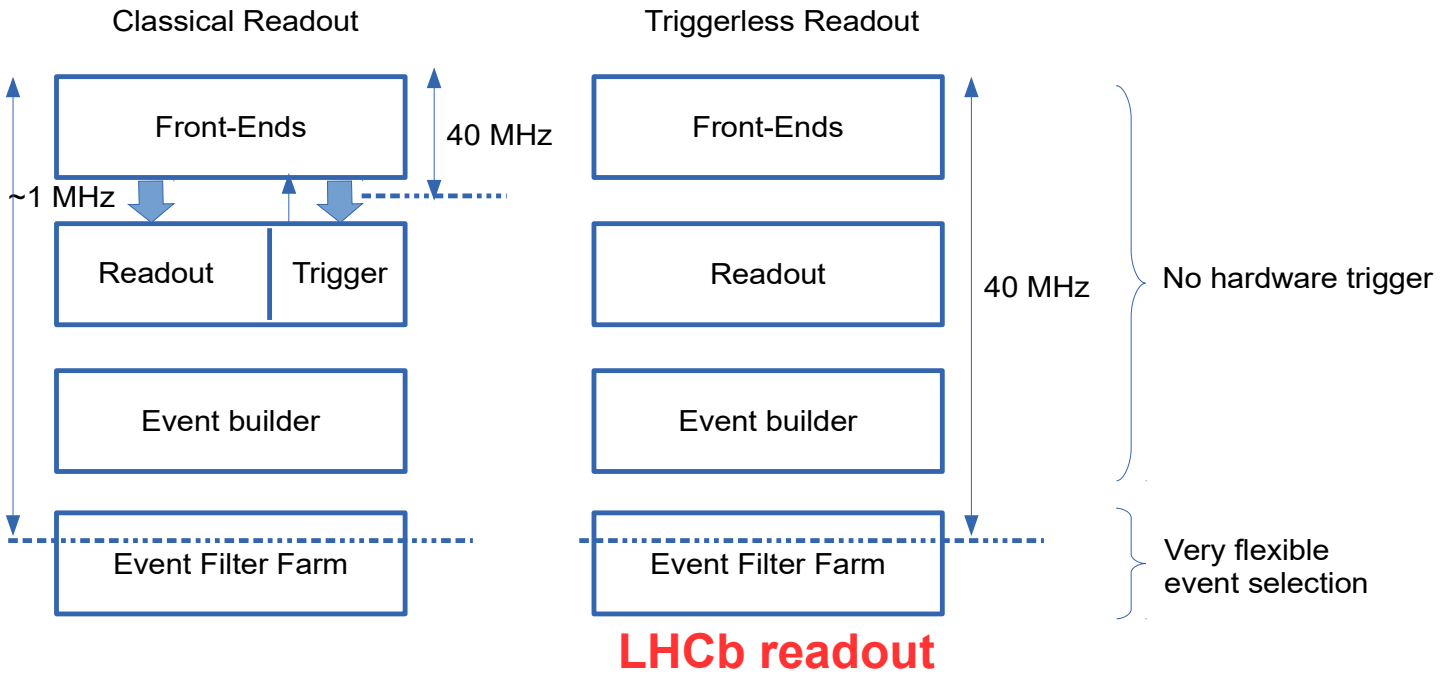
J.P. Cachemiche, on behalf of the LHCb collaboration

Outline

- LHCb and ALICE Readout
- Hardware design
  - Prototype
  - Final card
  - Measurements
- Production
- Firmware design



LHCb Upgrade key features

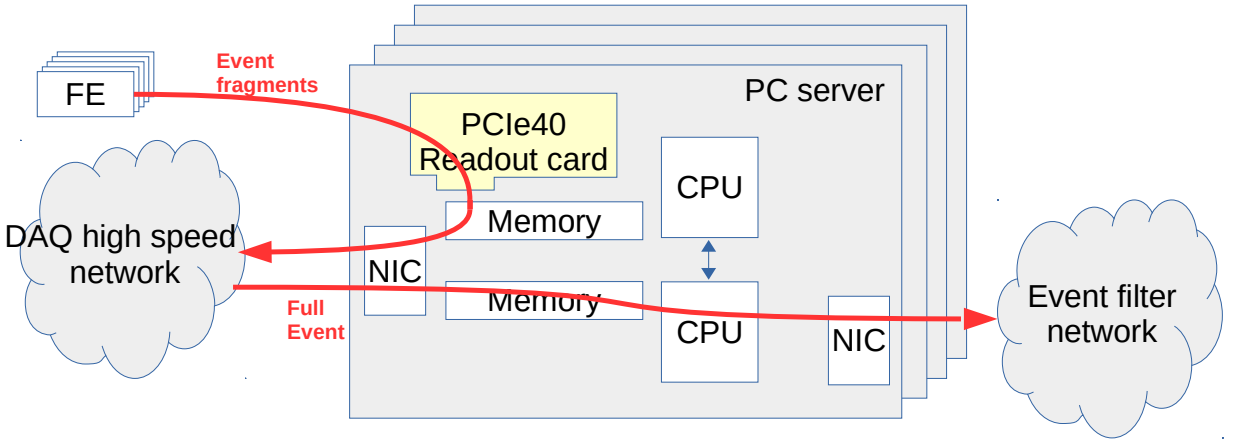


- LHCb uses a triggerless readout
- All event fragments routed at 40 MHz up to the farm

LHCb Upgrade key features

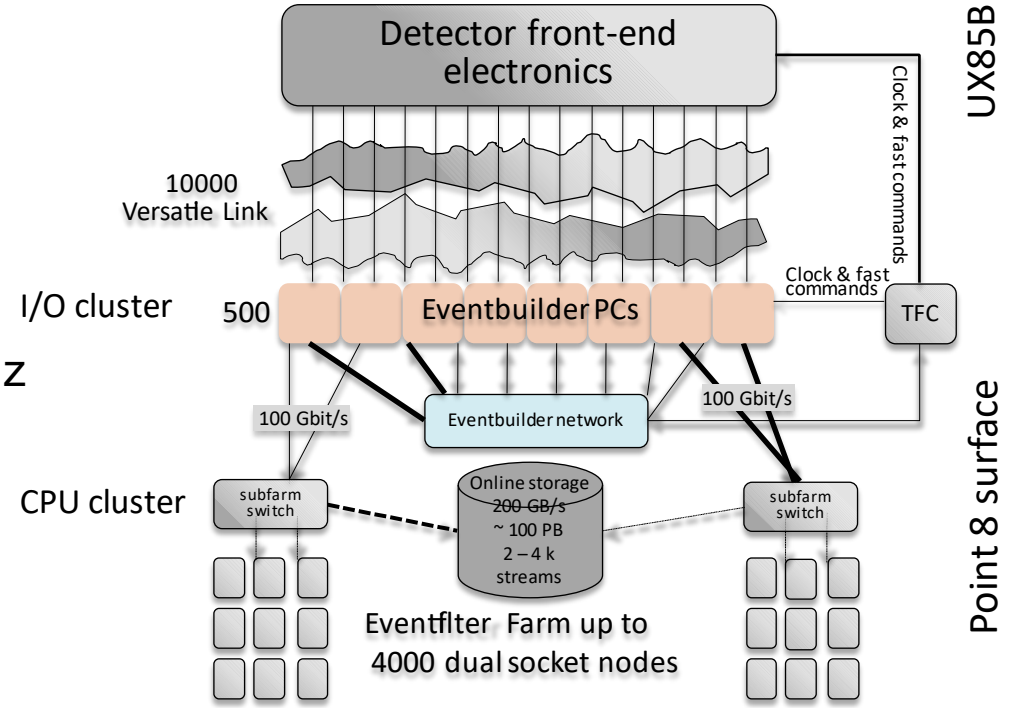
Principle

- Event building done by tightly coupled acquisition boards, CPUs and high speed network
- No intermediate back-end stage
  - Readout card implemented as a PCIe module
- Event building through servers in real time
  - Now possible due to internal CPU architecture evolution
- Event reconstruction with offline quality in real time
- Triggering replaced by filtering of reconstructed events



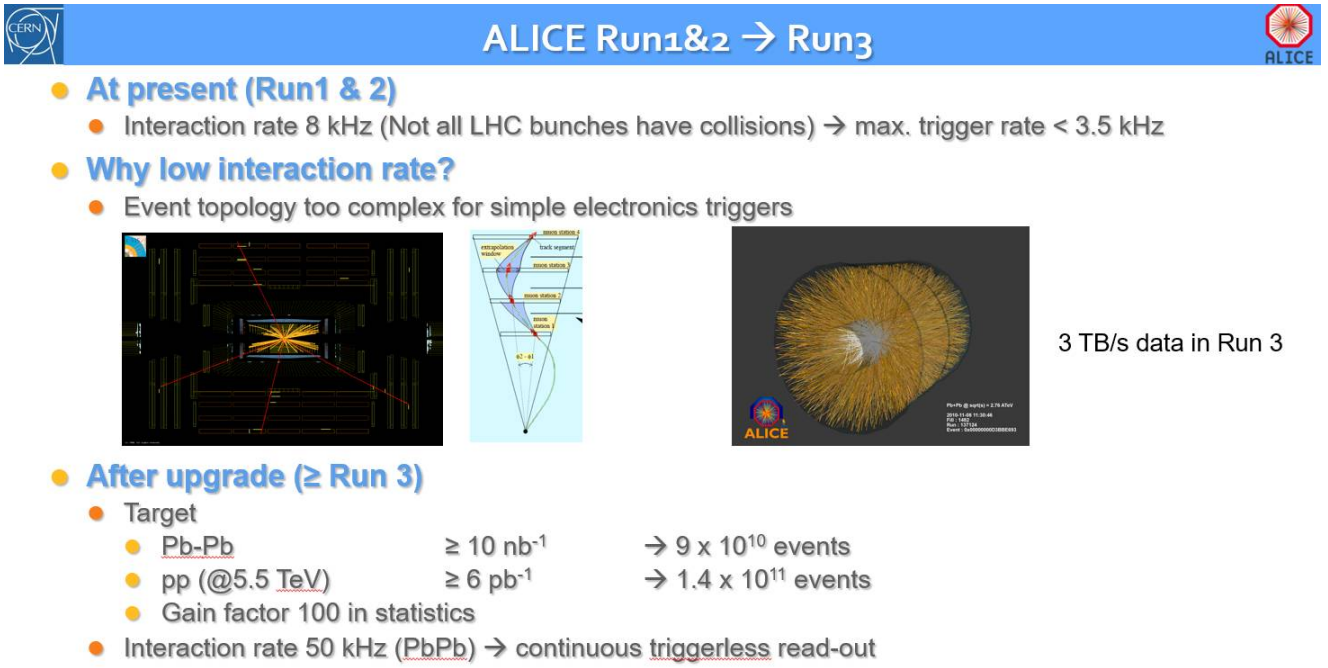
LHCb architecture

- Readout located on surface
  - Distance between FE and RO : ~350m
- ~ 10000 optical links
- ~ 500 readout boards
- ~ 100 TFC/ECS cards
- ~ 100 kBytes per event at 40 MHz
- ~ 32 Tb/s aggregate bandwidth
- ~ 4000 dual CPU nodes



Alice upgrade key features

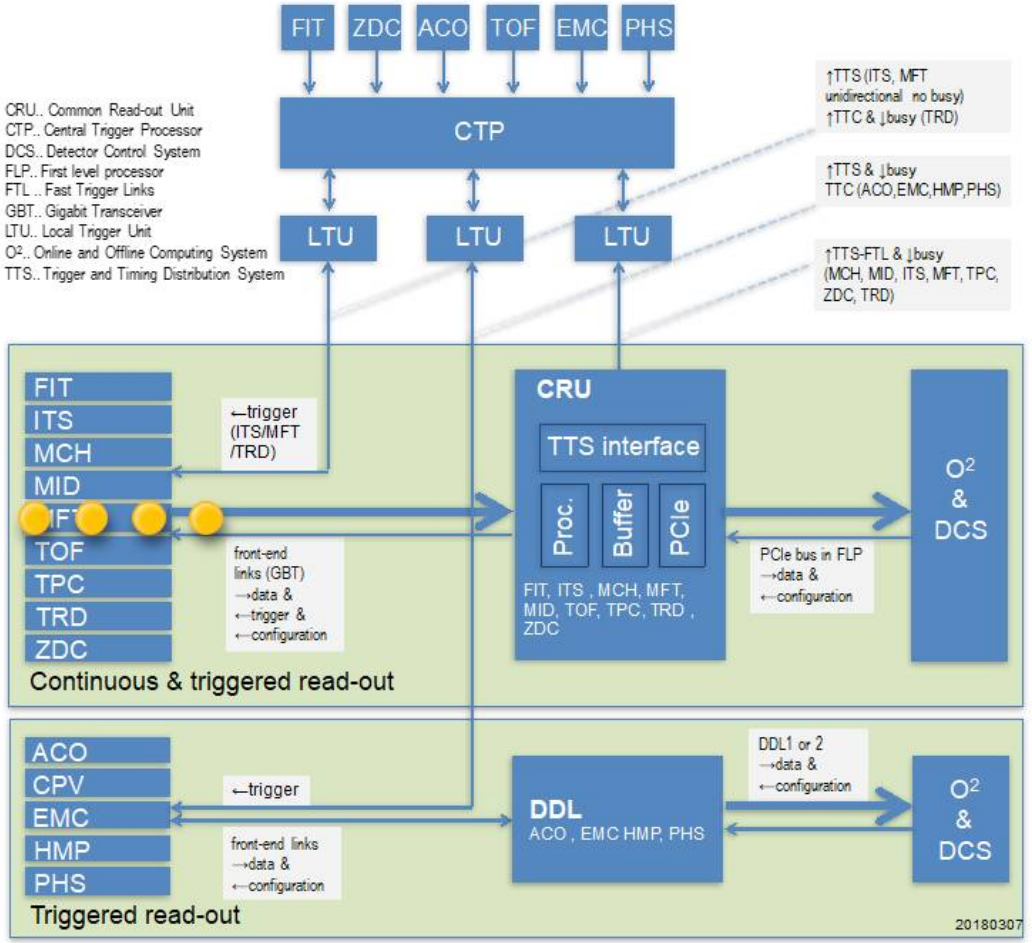
- Event topology too complex for electronics trigger
- 60% of events are kept
  - Low interaction rate + Continuous triggerless readout
- CRU (Common Readout Unit) based on the PCIe40 card
- Acquires and compresses data on the fly



Courtesy Alex Kluge

ALICE architecture

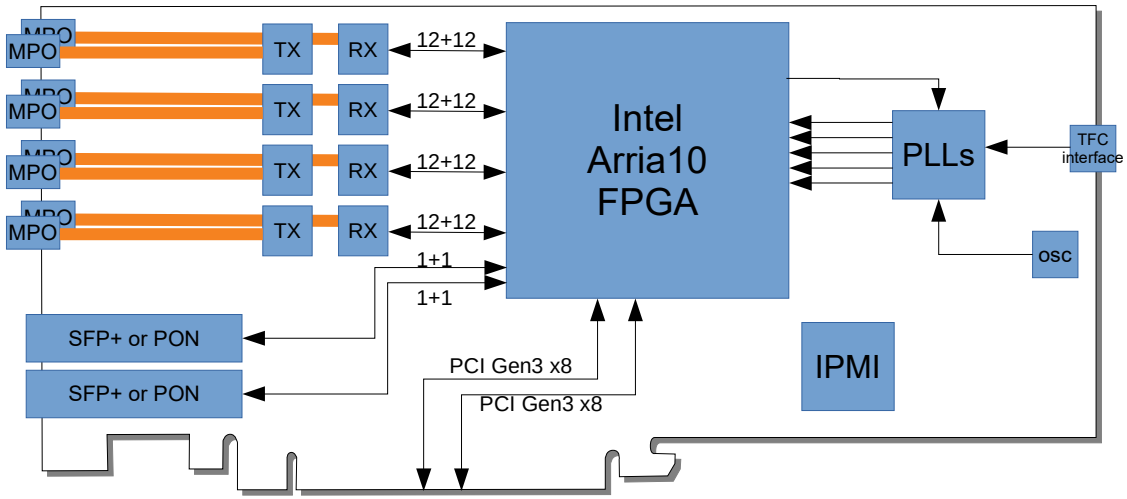
- Readout located on surface
  - Distance between FE and RO : ~120m
- ~ 9000 optical links
- ~ 540 readout boards
- ~ 68 MBytes per event at 50 KHz
- ~ 27 Tb/s aggregate bandwidth
- ~ 1500 GPU based event processing nodes



Courtesy Alex Kluge

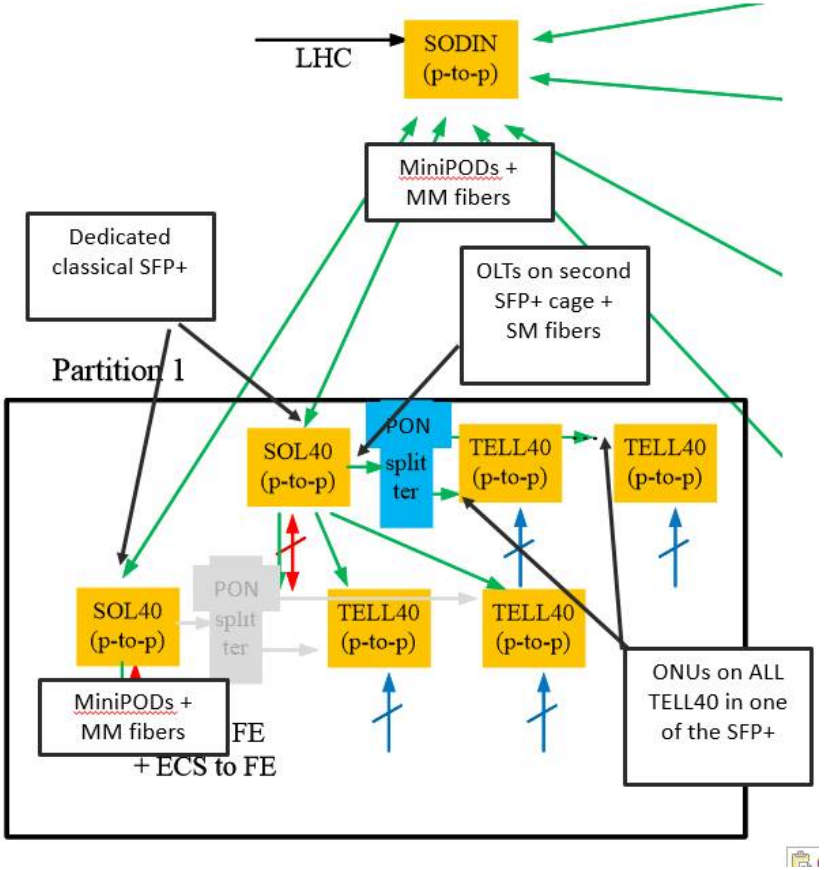
The readout board : PCIe40

- Features :
  - 1 large FPGA 1.15 million cells (Arria10 10AX115S3F45E2SG)
  - 48 bidirectional links running at up to 10 Gbits/s each (minipods)
  - 2 bidirectional links running at up to 10 Gbits/s devoted to time distribution (can use SFP+ or 10G PON devices)
  - Sustained 112 Gbits/s interface with CPU through PCIe
  - No buffer memory : we use the PC memory instead
  - Remote reconfiguration of all the programmable devices
  - Fully instrumented: all voltages, currents and temperatures measured



Versatility

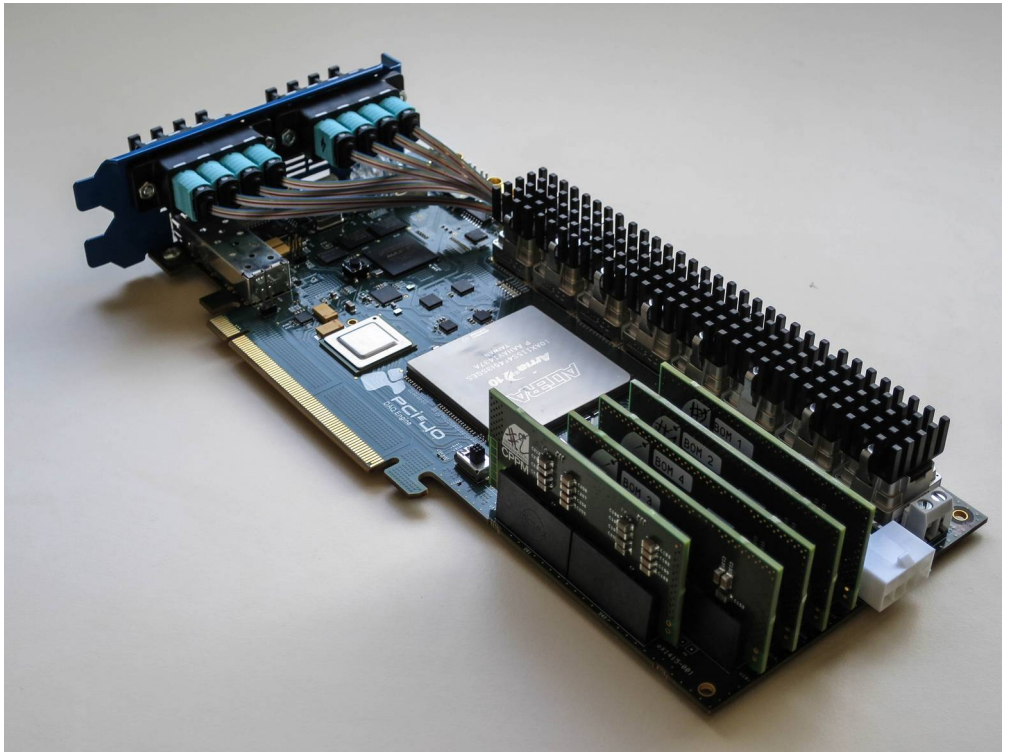
- Can be mapped over several functions by reprogramming the FPGA
- Different names for the same card in LHCb according to its programmation :
  - SODIN : Timing distibution and Fast Control
  - SOL40 : Slow control
  - TELL40 : Acquisition
- Minipods for interfaces with Front Ends
  - GBT protocol at 4.8 Gbits/s
- PON devices for TFC
  - 8B10B protocol at 3.2 Gbits/s



Hardware design

PCIe40 prototype

- First prototype developed in 2016
- 24 copies manufactured for both the LHCb and Alice collaboration
  - Used as « mini DAQ » for debugging front-end cards
  - Programmed to provide acquisition, ECS and TFC in a single firmware



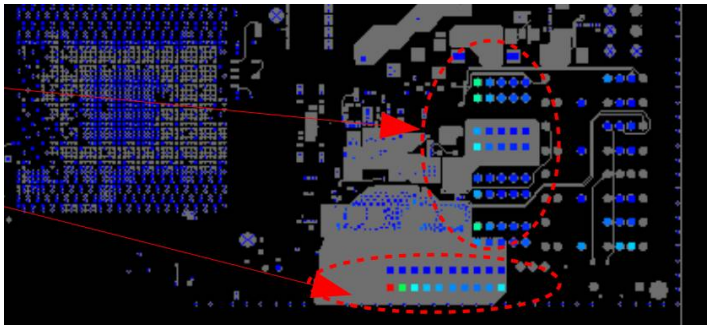
Preparing the final module

Power consumption of large FPGAs very high

- Up to 52 A on the core !
- Power consumption
  - FPGA estimated at ~ 80 W
  - Card estimated at ~ 150 W with Engineering Sample
  - Limited thickness for the stackup

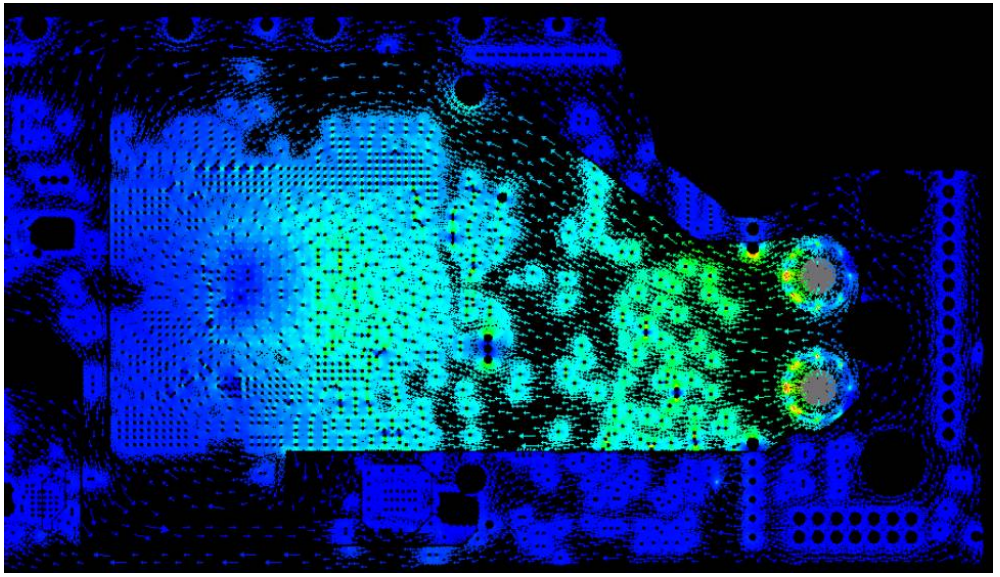
Refining of current flow simulations

- Simulations of current flow showed dangerous hot spots at full load
  - Power planes have been redesigned and vias placement has been optimized
- Current flow through power mezzanine connections not symetric

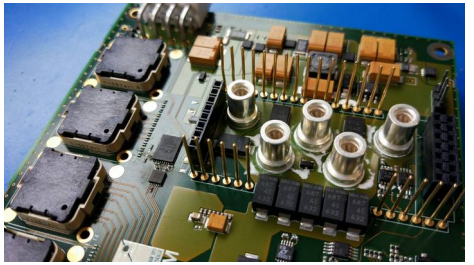


Preparing the final module

Replacement of the 5 vertical mezzanines by a single flat one



Current flow between mezzanine and FPGA with new design



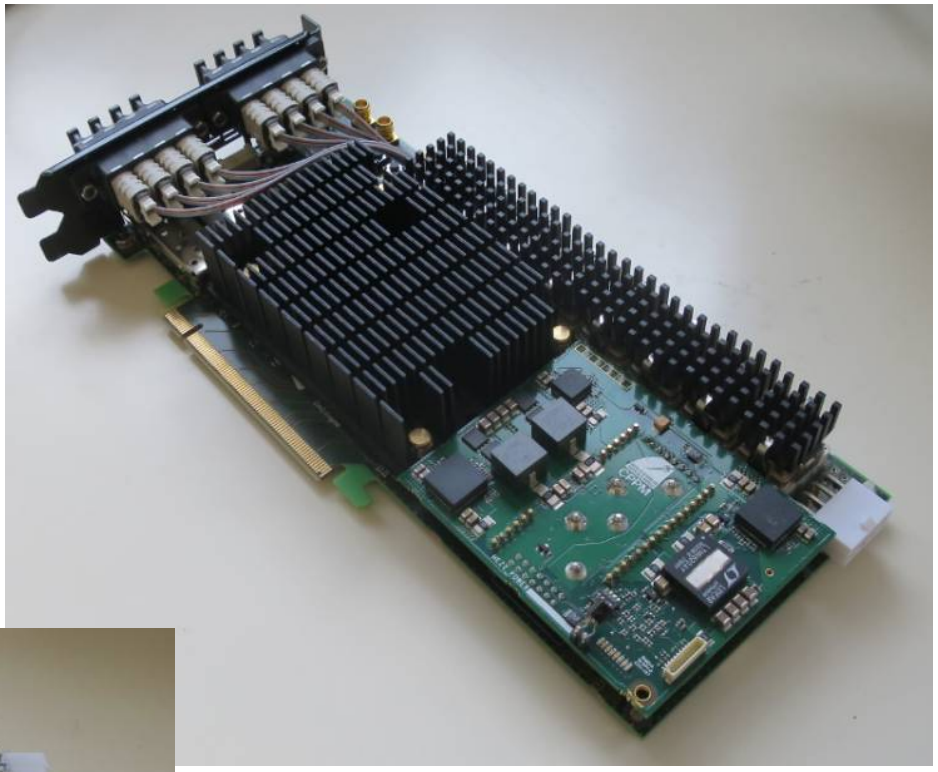
Optimizations

Many improvements

- Cost savings
  - Removal of expensive components (PCIe bridge, Serial Flash and corresponding power supply)
  - One additional SFP+ or PON cage added → less TFC/ECS modules
- Performance improvement
  - Use of new PLLs with a very low jitter compared to previous ones
- Reliability
  - Complete redesign of the power supply due to buggy DCDC converters
  - Optimisation of current flows → avoids local over heatings in the PCB
    → Single power mezzanine now horizontal for symetrical current flow
  - Improvement of power sequencing to ease maintenance and guaranty a longevity of the module → manages now power down
  - Optimization of decoupling → less noise
  - Heat sink redesign for better cooling
- New functionalities
  - Programming speed multiplied by factor 4 with a new embedded USB Blaster II
  - IPMI management : allows the system to adjust the fan speed in function of the temperature or automatically cut the power suply if temperature is too high
  - Serial flash for identificating modules during production

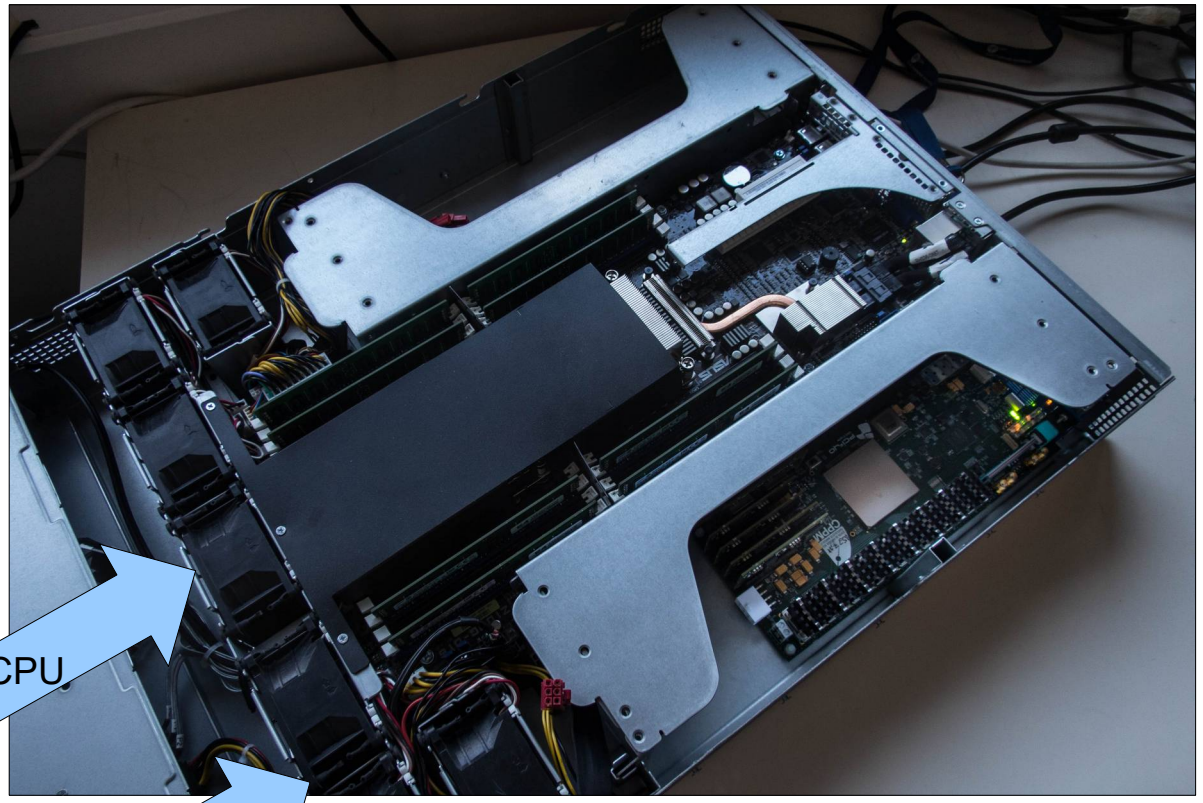
Final module

- Two first modules validated end 2017
- Early duplication by Alice of 28 modules to speed up first production



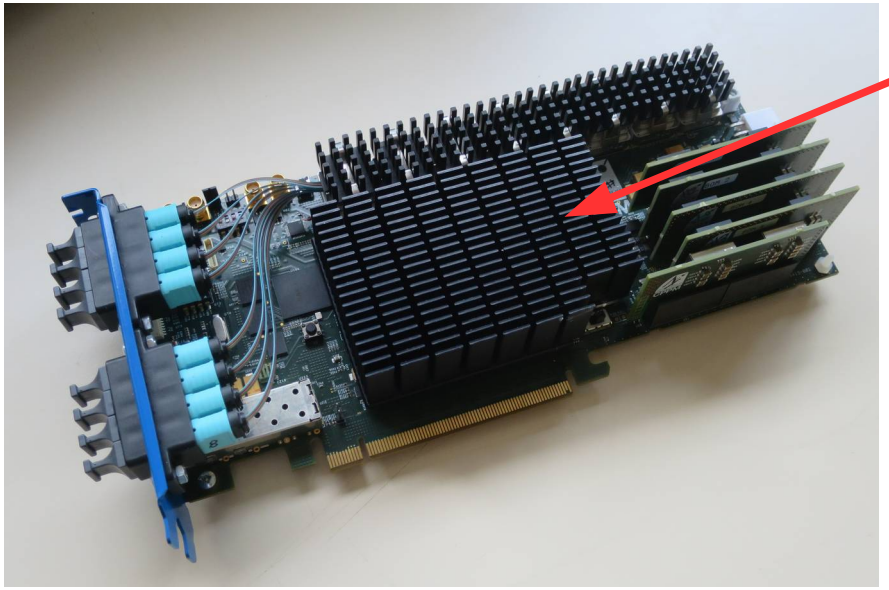
Cooling

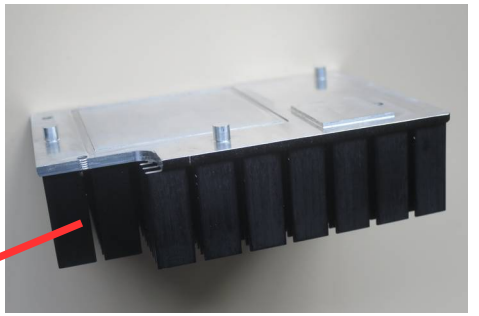
- PC environment not as well defined as xTCA systems
- Very well cooled PC server has been selected



Cooling solution

Use of a custom passive cooling





Custom passive heatsink

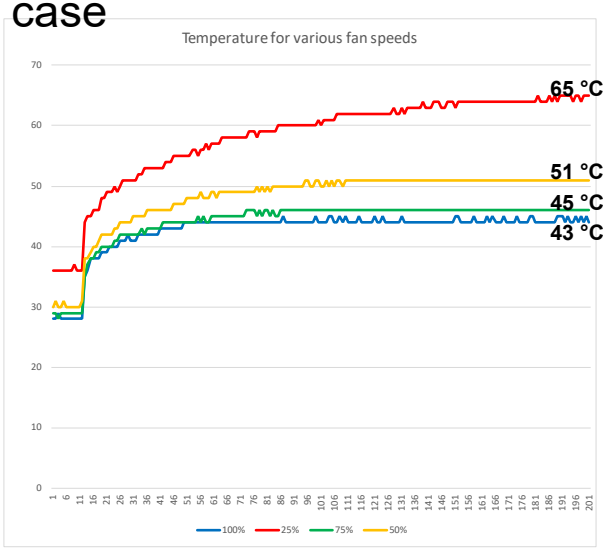
Power consumption and cooling

Power consumption and cooling

- Push the module at the limit of power dissipation
- Principle:
  - Use a « heating function» replicated thousands of times to get an FPGA occupancy of 86%
  - Inject a clock with programmable frequency between 10 MHz and 600 MHz
- Automatic power off if the FPGA temperature overpasses 82°C
- Vary the speed of server fans (25%, 50%, 75%, 100%)
- Measure voltages, currents and temperature in each case

Results obtained with ASUS server

- 2 cards on same side
- Provided that this firmware is representative passive cooling seems sufficient



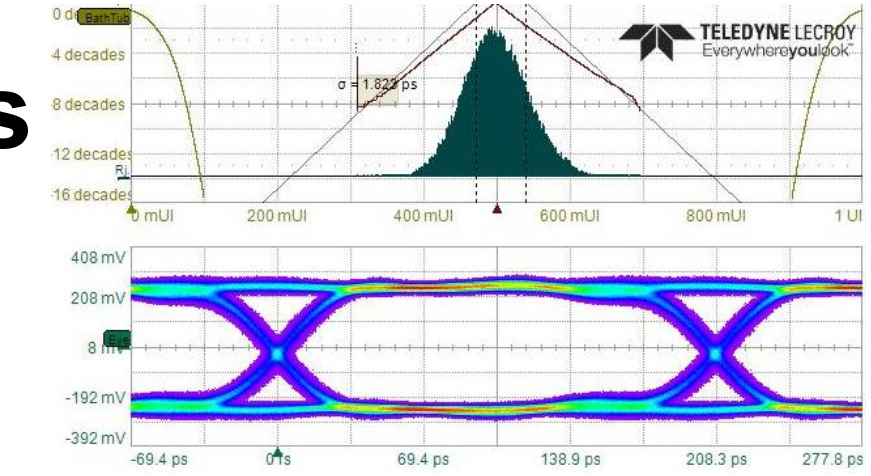
FPGA temperature for several fan speeds

Links measurements

BER << 10^{-16}

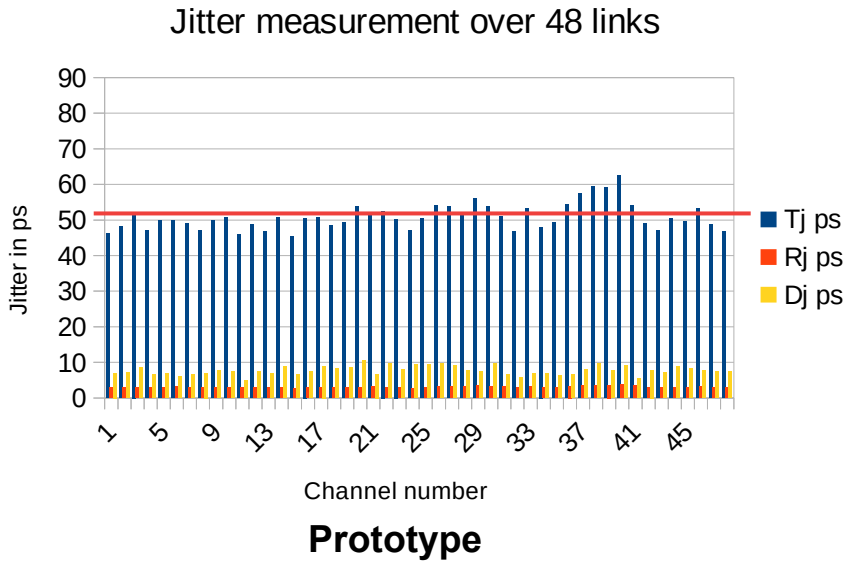
Jitter

- Final card jitter improved vs prototype
  Total jitter goes from 51 ps → 38 ps

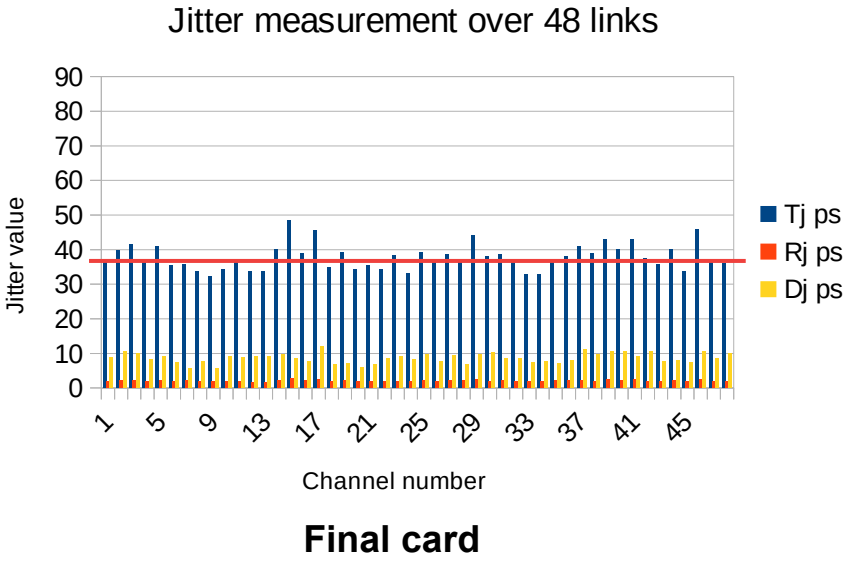


SDA Jitter	Tj(1e-12.0)	Rj(sp)	Dj(sp)	Pj	ISI	DCD	DDj	BitRate
Lane1	36.902 ps	1.995 ps	8.847 ps	6.89 ps	6.902 ps	4.795 ps	11.418 ps	4.80942 Gbit/s
SDA Eye	EyeHeight	EyeOne	EyeZero	EyeAmpl	EyeWidth	EyeCross	EyeAvgPwr	
Lane1	424.5 mV	248.9 mV	-247.7 mV	496.5 mV	192.1 ps	45.52 %	73 µV	

Measurements at reception stage for a PRBS31 pattern running at 4.8 Gbits/s



Prototype



Final card

Production

Production

LHCb production started

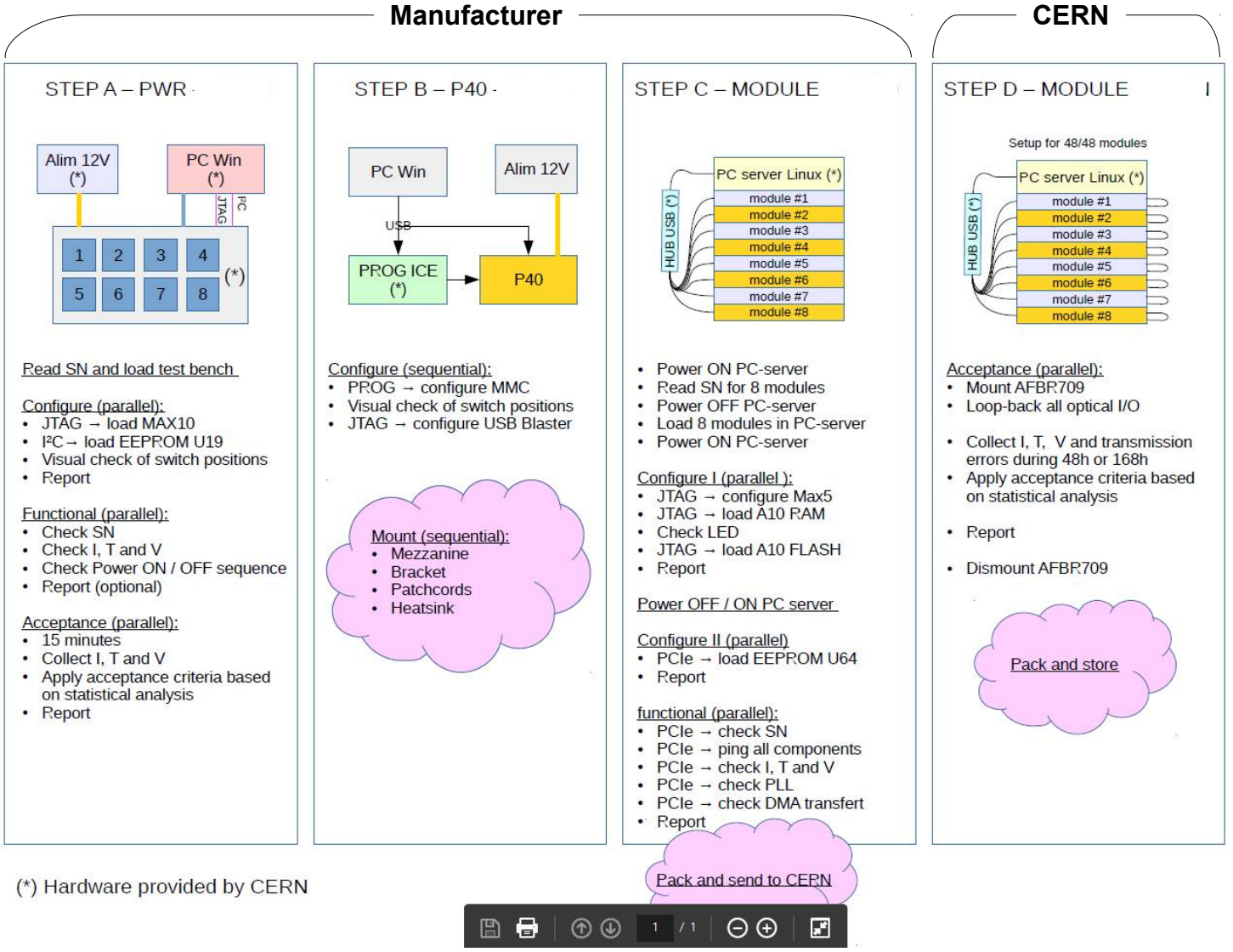
- ~700 modules in 3 batches :
  - Preseries of 24 cards
  - First batch of 330 cards
  - Second batch of 345 cards

- Schedule
  - Preseries July 2018
  - First batch November 2018
  - Third batch April 2019

Alice should follow a similar route

Testing methodology

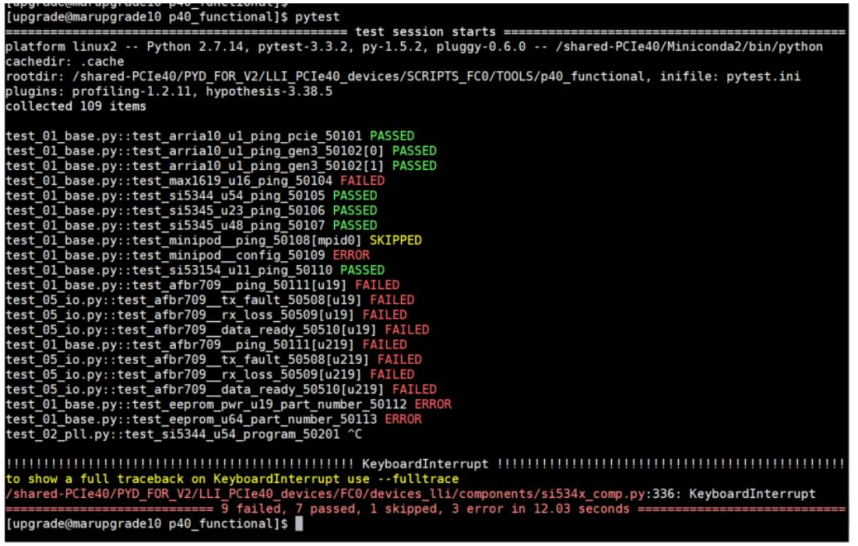
4 steps



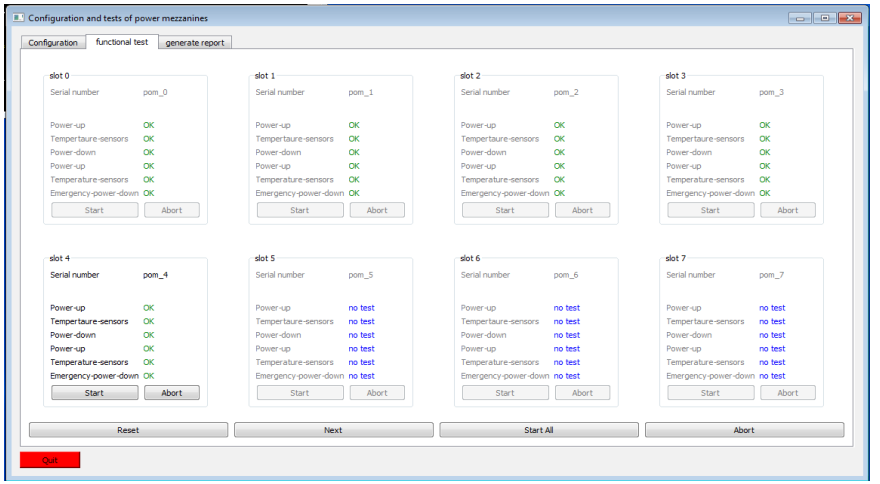
Production tests

Run in assembly company

- Based on Pytest
  - Very flexible command line testing tool
  - Able to test target sub-set of components
  - Object oriented design
  - Can be driven by a GUI
- Fully tests the board
  - 150 unitary tests ran in a few minutes
  - Check the operation of all the devices on the modules
  - Measure voltages, currents, temperatures, frequencies, etc.
  - Produces test reports for each module
- Overall management of reports
  - Reports directly sent to CERN data base



Expert interface

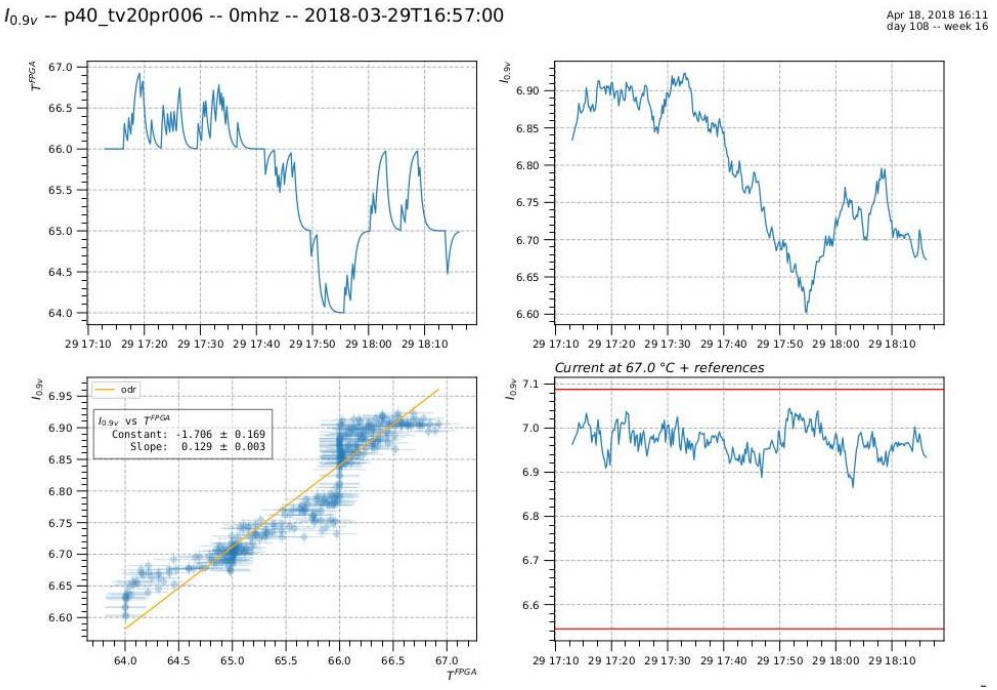


Operator interface

Acceptance tests

Run at CERN

- Duration 24 or 168 hours
  Allow to eliminate early failures
- Rely on Pytest
- Possible post processing of results
  - ~ 20 parameters currently used
  - ~ 60 parameters completely logged

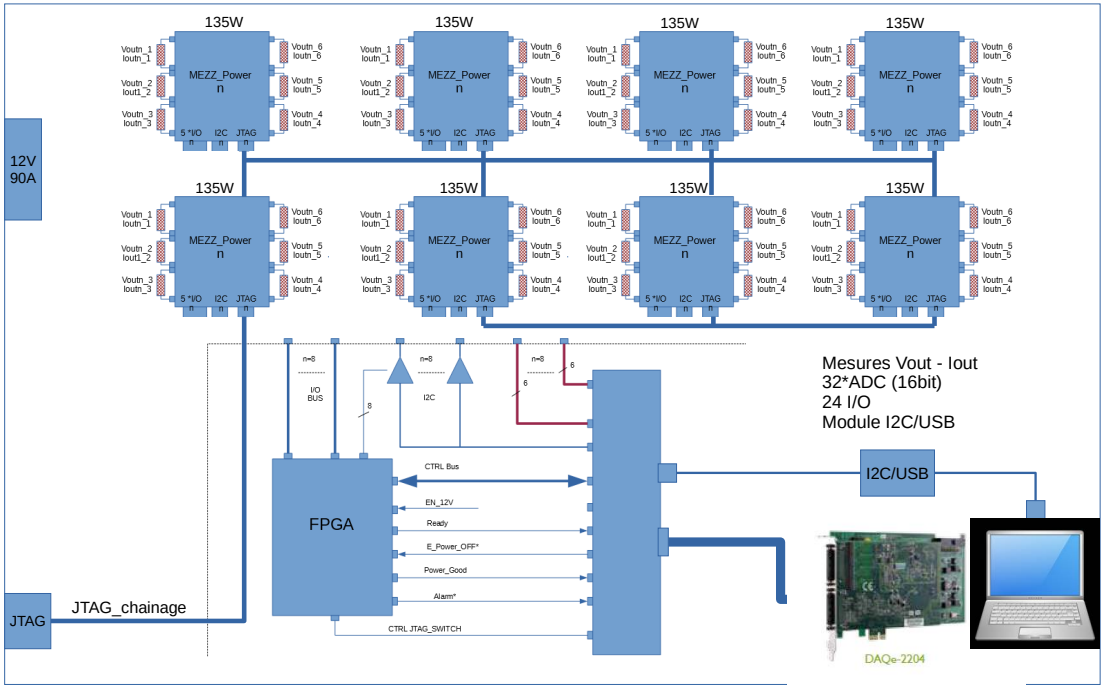


	obs	rsquared	alpha	cl-alpha	current	cl-current	p-value	sigma	cl-sigma
0	0.9V	0.004	-0.104	[-0.577, 0.369]	8.193	[8.17, 8.217]	0.0	0.115	[0.094, 0.137]
1	1.02Vccr	0.073	0.026	[0.001, 0.051]	6.33	[6.326, 6.334]	0.475	0.015	[0.012, 0.018]
2	1.02Vcct	0.041	0.018	[-0.006, 0.041]	1.977	[1.975, 1.979]	0.502	0.006	[0.005, 0.007]
3	1.8V	0.019	0.021	[-0.021, 0.063]	7.011	[7.007, 7.014]	0.809	0.01	[0.008, 0.012]
4	1.8Va10	0.021	0.02	[-0.017, 0.057]	3.627	[3.625, 3.63]	0.729	0.009	[0.008, 0.011]
5	1.8Vccpt	0.004	0.01	[-0.031, 0.05]	1.458	[1.454, 1.462]	0.845	0.01	[0.008, 0.012]
6	2.5V	0.035	0.074	[-0.032, 0.179]	2.805	[2.8, 2.809]	0.003	0.025	[0.02, 0.029]
7	3.3V	0.001	-0.008	[-0.091, 0.074]	1.929	[1.923, 1.936]	0.881	0.02	[0.017, 0.024]
8	12V	0.018	0.069	[-0.069, 0.207]	2.883	[2.874, 2.892]	0.523	0.035	[0.029, 0.042]
9	12Vatx	0.017	-0.003	[-0.01, 0.004]	-0.021	[-0.021, -0.02]	0.032	0.002	[0.001, 0.002]

Production setup for testing mezzanines

Need to speed up the tests

- Goal is to test 8 cards at once
- Specific test bench designed at CPPM
  - Connected to commercial ADC card driven by a Windows PC
  - Allows to test the cards at full load

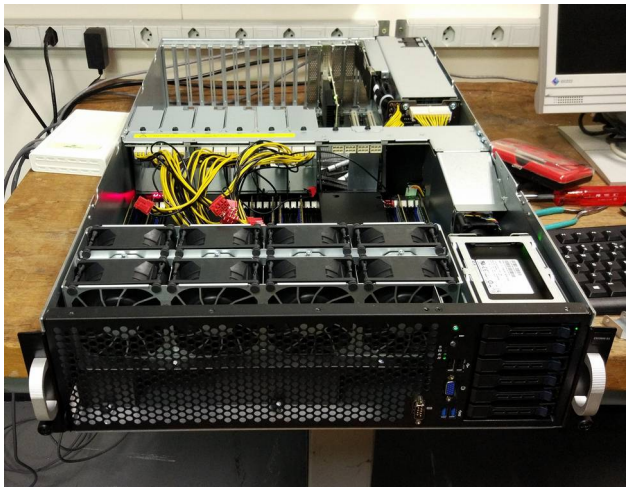


Production setup for testing modules

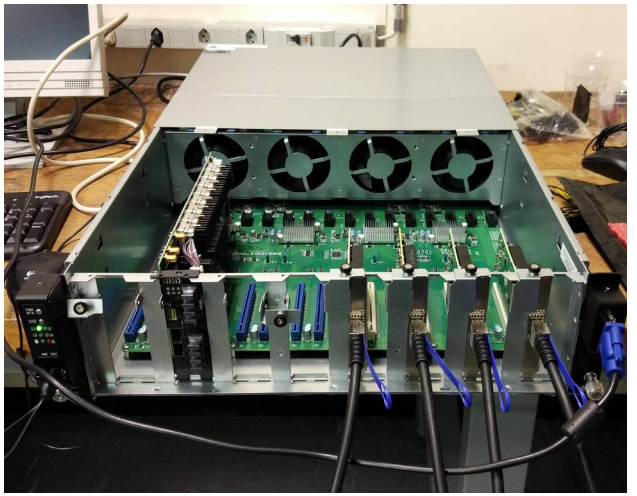
Same approach for the full module

- PCIe crate expander or servers
- On going evaluation

Cubix crate expander



ASUS server

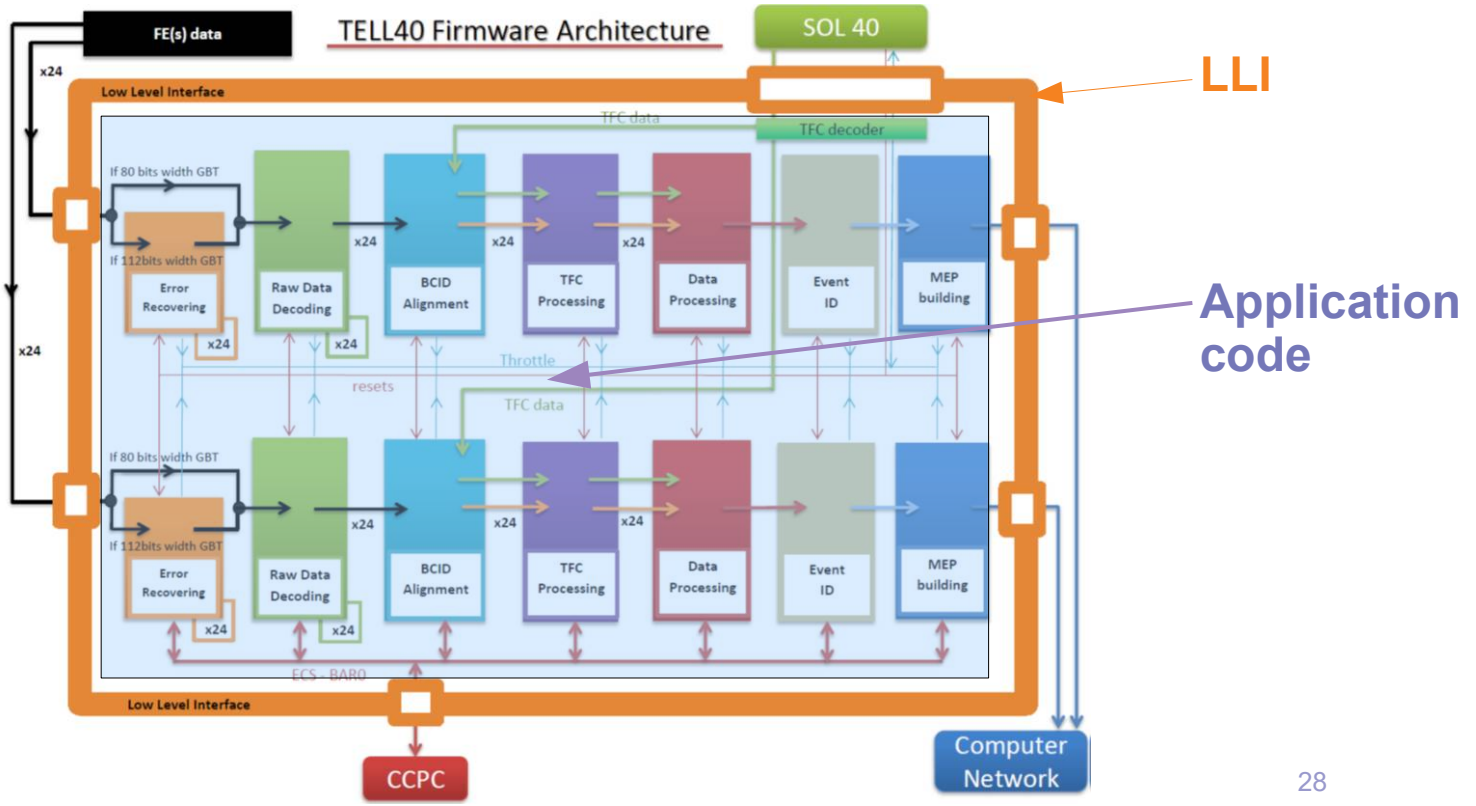


ASRock server

Firmware

LHCb 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)
- Very simple interface for **Application code** mostly drawing from and pushing data to FIFO-like interfaces
- Similar approach by Alice but they wrote their own code

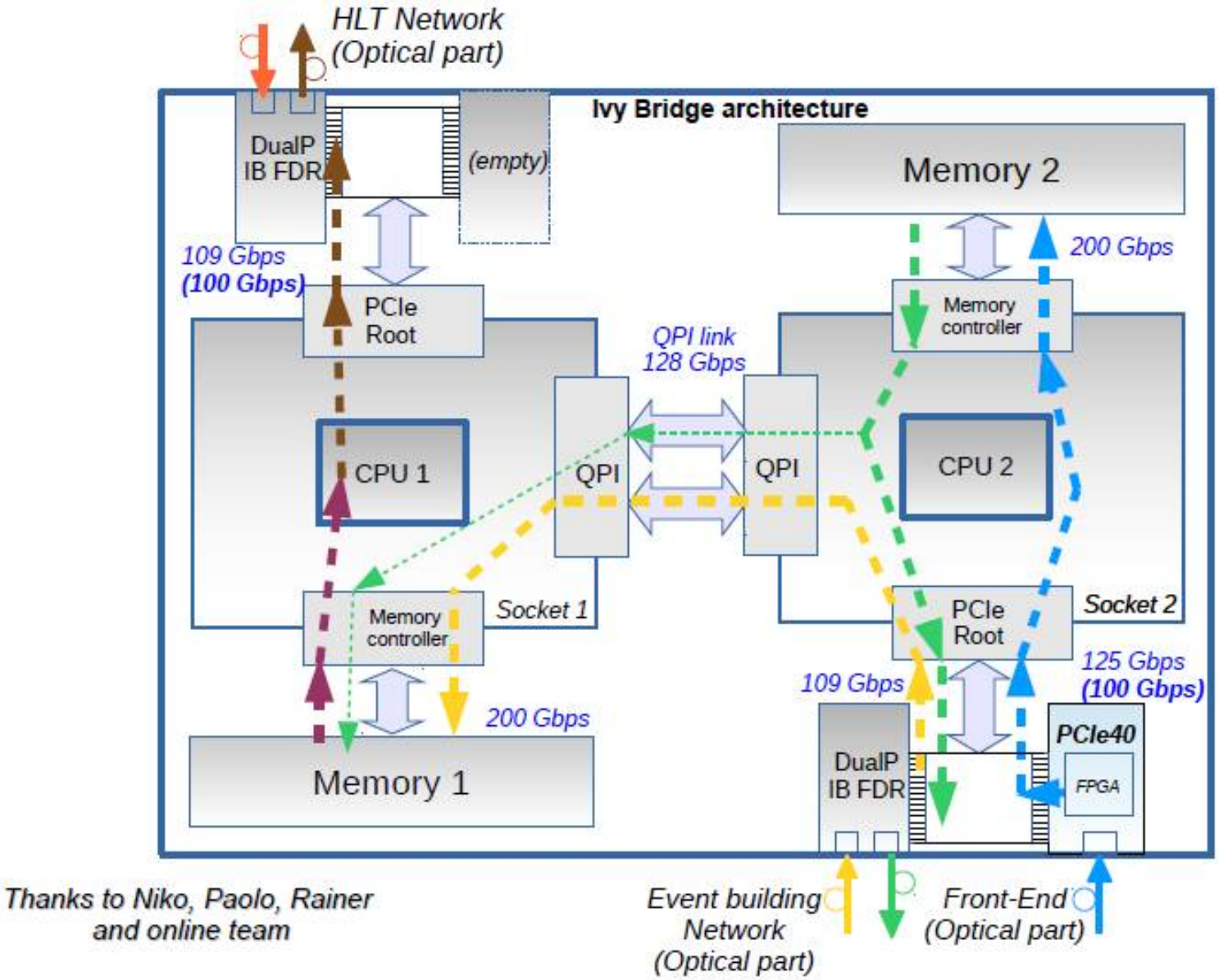


Conclusion

- Cards adressing many needs in our community
  - Large acquisition capability
  - Manages timing distribution
  - High processing power
  - Powerful interface between dedicated Front-Ends and commercial computer CPUs
- Flexible enough to used in many ways
  - 3 functions in LHCb (DAQ, ECS, TFC)
  - Can fit ALICE needs as well
  - Also selected for the readout of the µ3E experiment
- Lots of effort spent for optimizing the card for production
  - Automatic testing
  - Parallel testing
  - Long time acceptance testing
  - Automatic recording

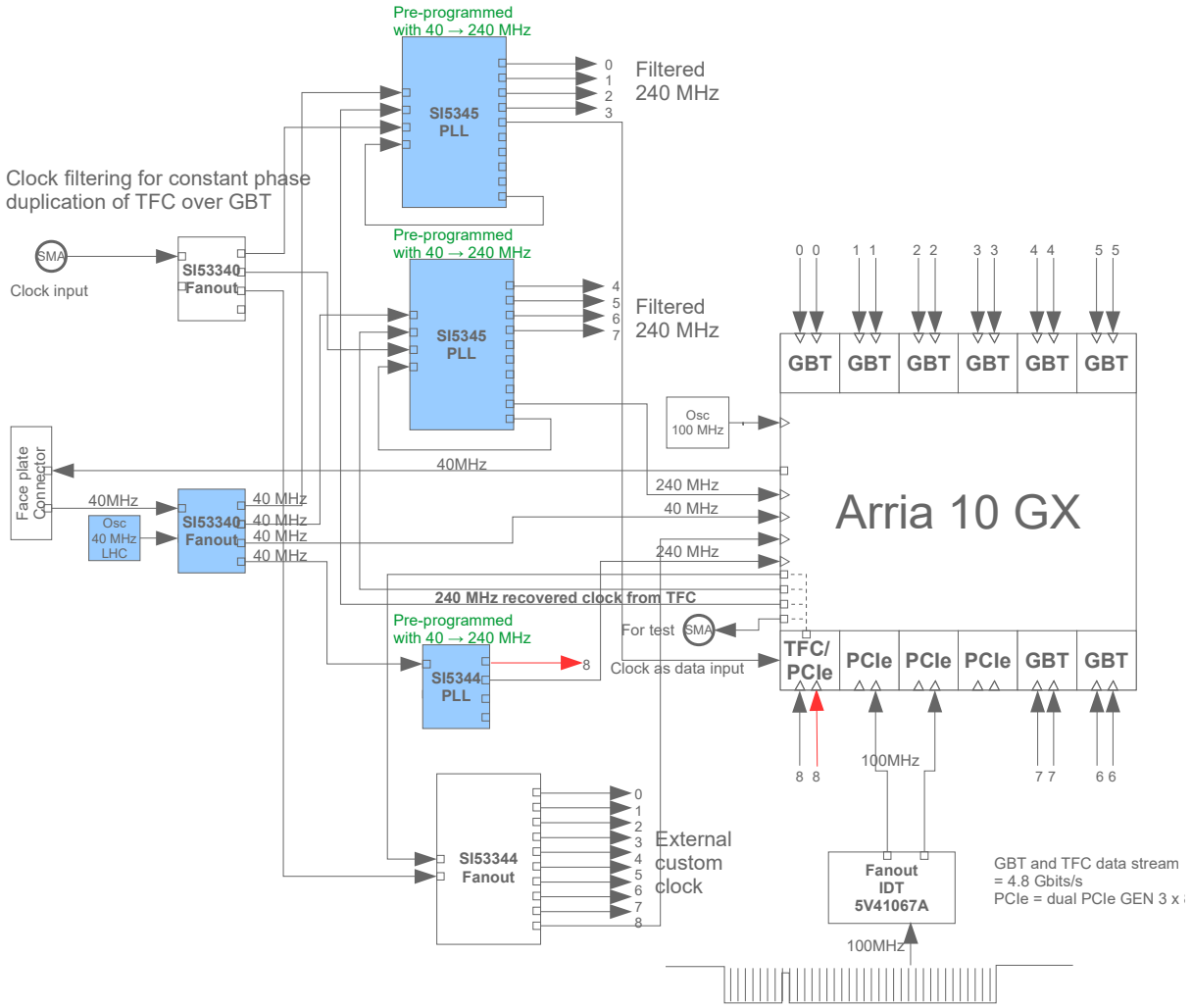
More information

Data path in the computer

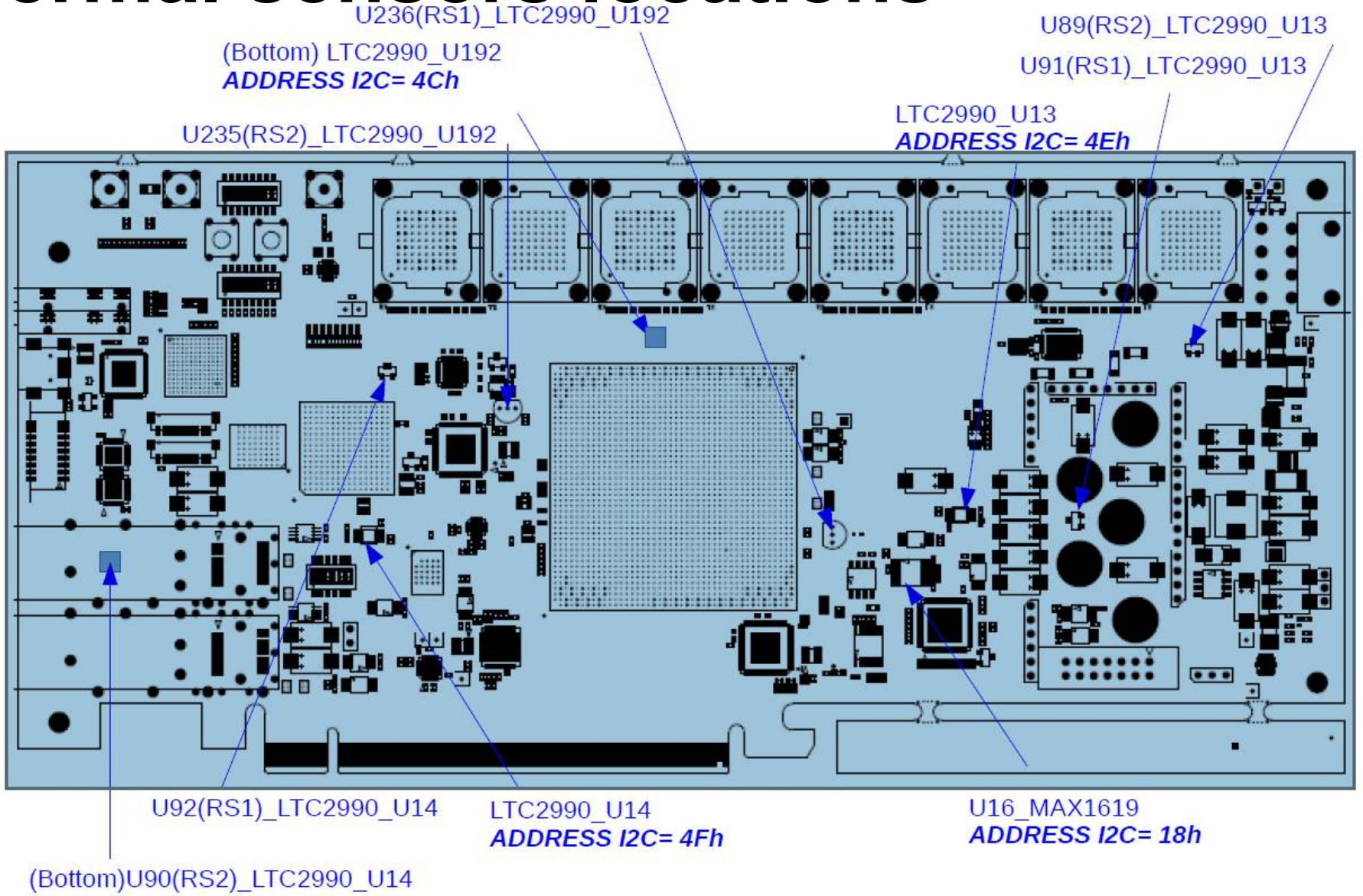


Clock distribution

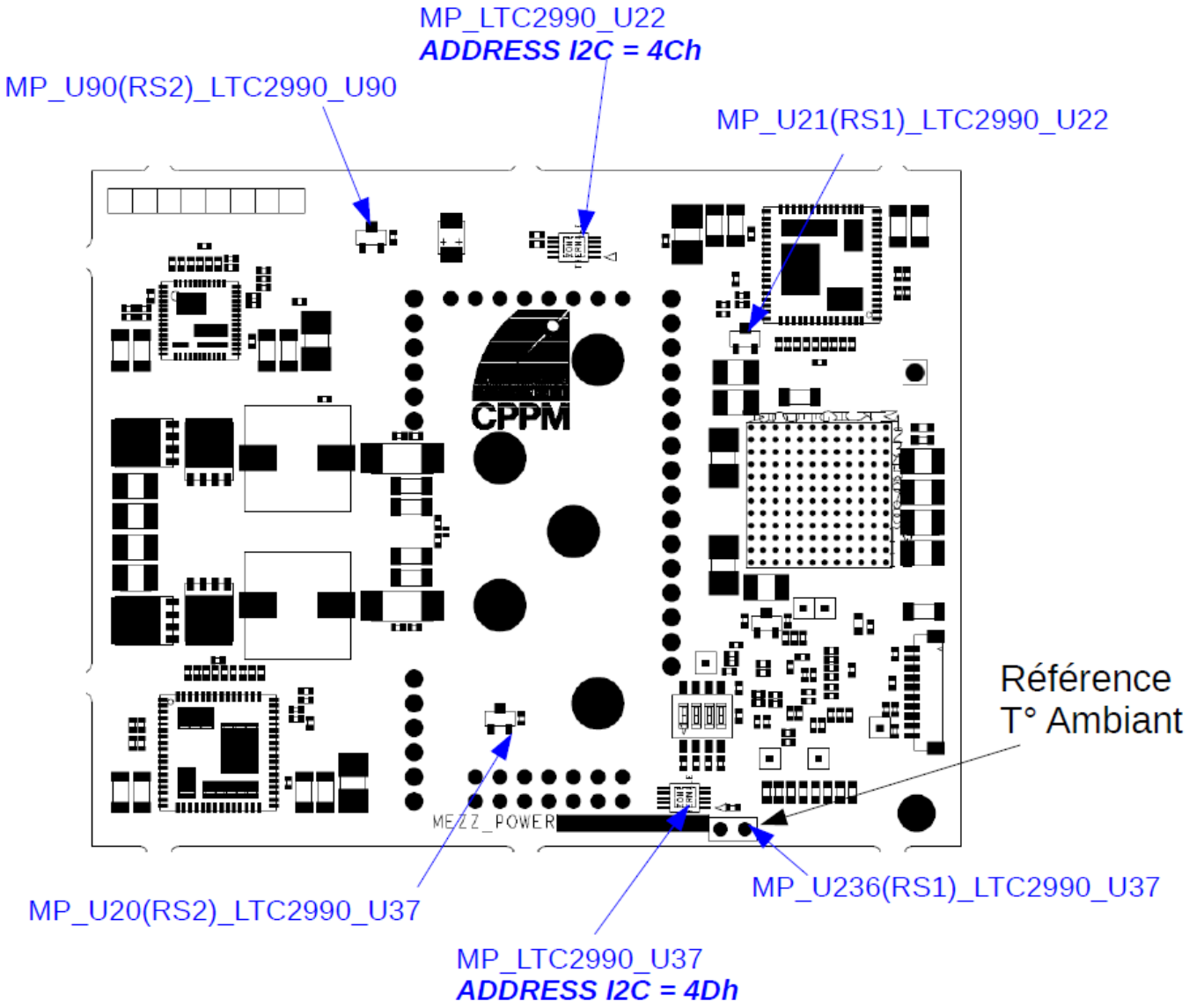
Clock Tree PCIe40V2



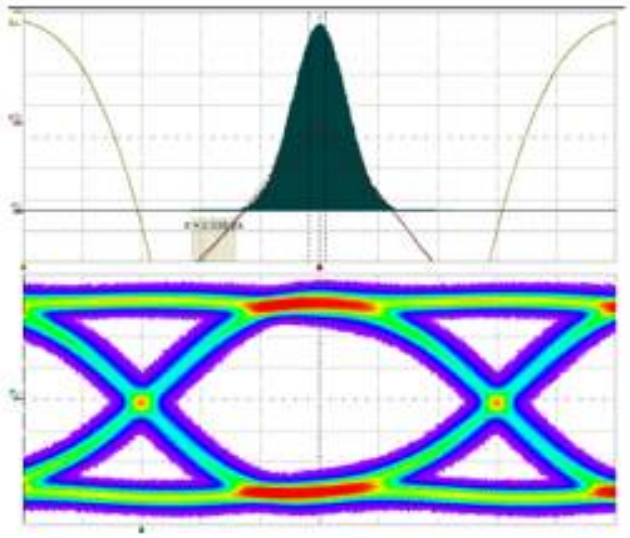
Thermal sensors locations

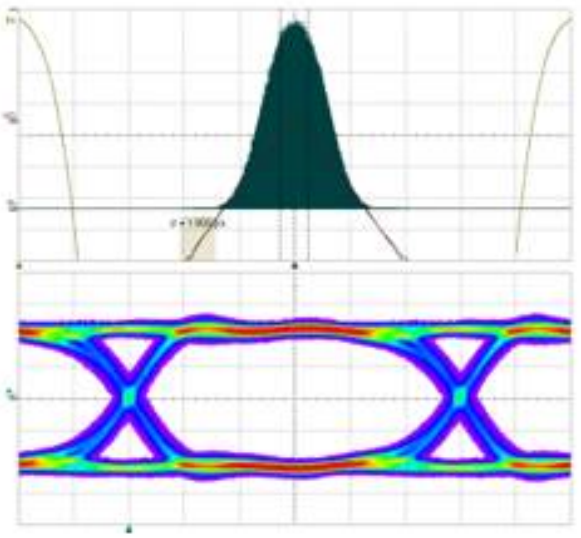


Thermal sensors locations



Eye diagrams

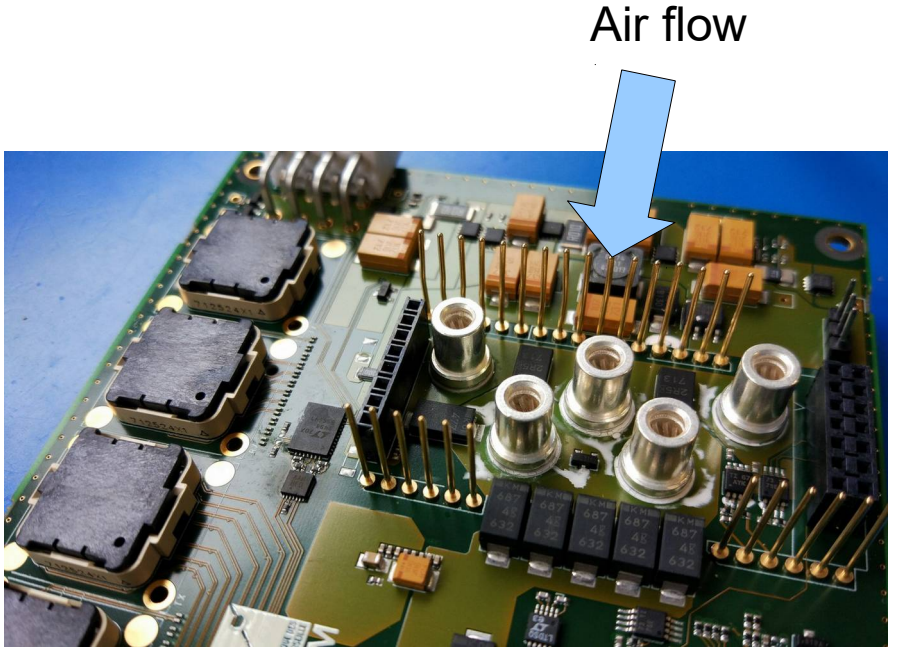




Measurements at 10.0Gbit/s
Total jitter ~ 36.82ps
Random Jitter ~ 2.22ps
Deterministic Jitter ~ 5.6ps

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

Mezzanine connector

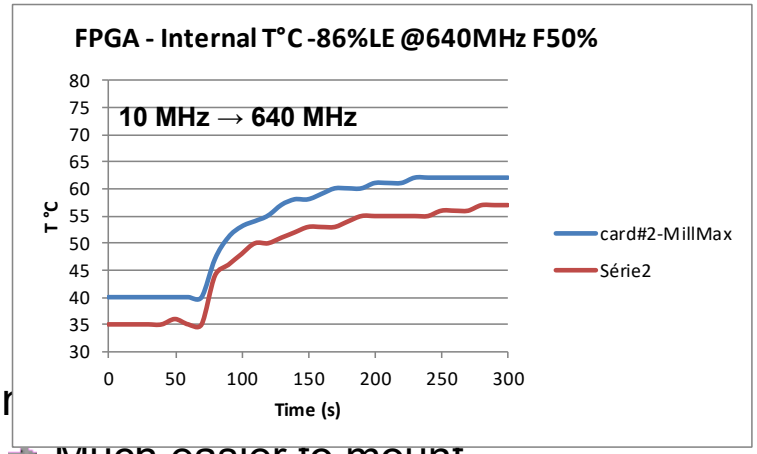


Two choices : Samtec or Millmax

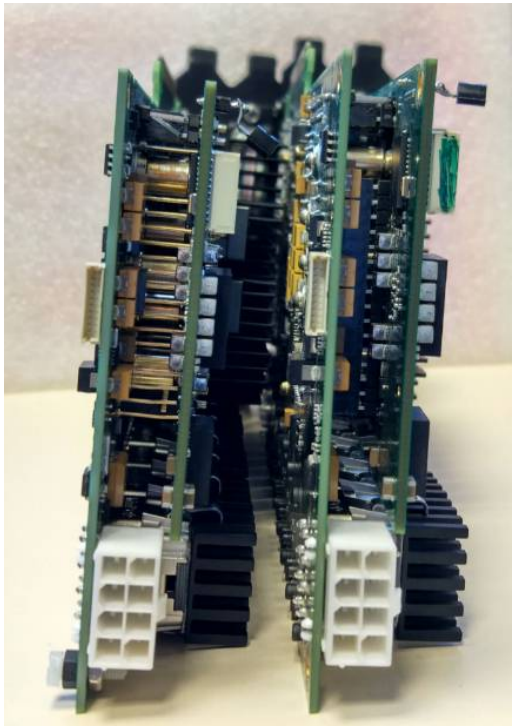
- Samtec : classical « full » connectors
- Millmax « transparent » connectors to let the air flow under the mezzanine

Cooling tests made with both solutions

- Counter intuitive results : Millmax card hotter than Samtec one (~5 to 6°C)
  - Venturi effect ?

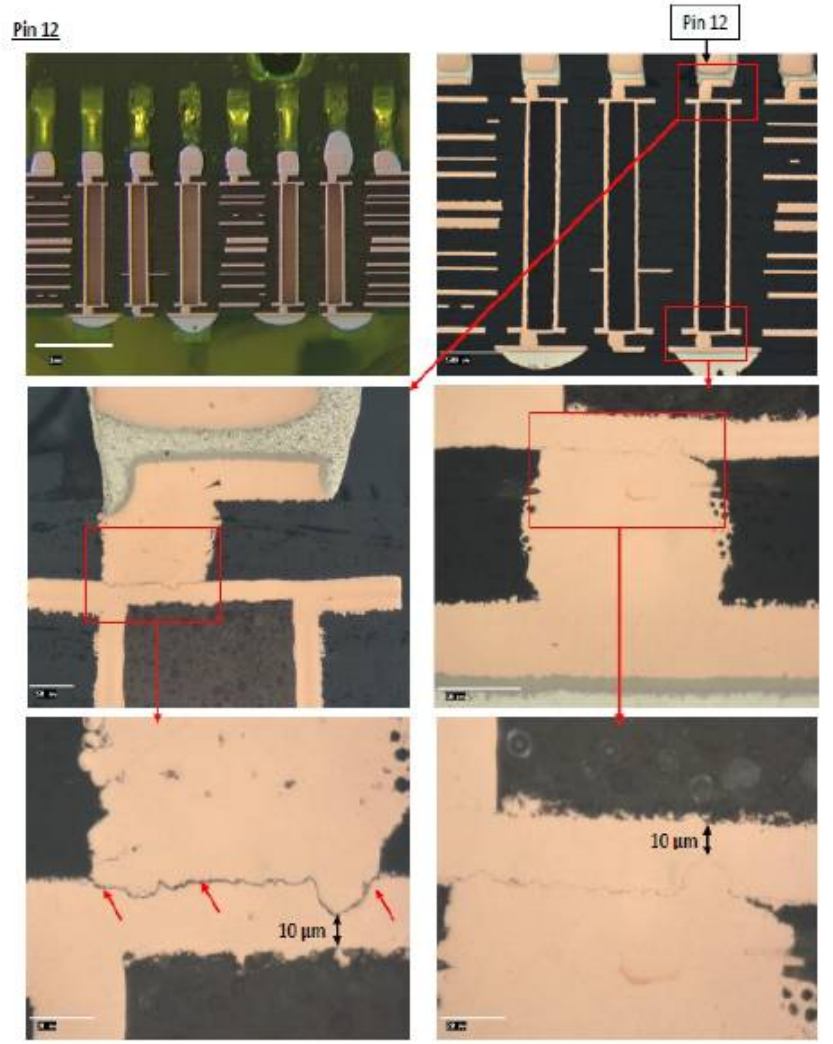


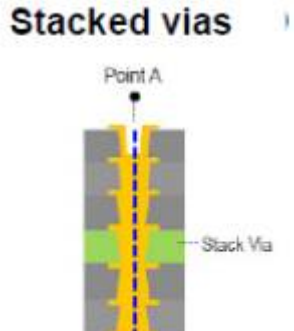
- Fir
  - Much easier to mount



The PCB episode

- First batch of 6 MiniDAQ2 almost failed. Three boards survived but would die soon.
- After a long investigation, the issue was localized on the PCB. It was due to micro-cracks in the so-called stacked vias.
- A new board with a PCB from a different manufacturer was delivered Feb 15, 2017.
- After an extensive campaign of tests we concluded that the board is fully functional.

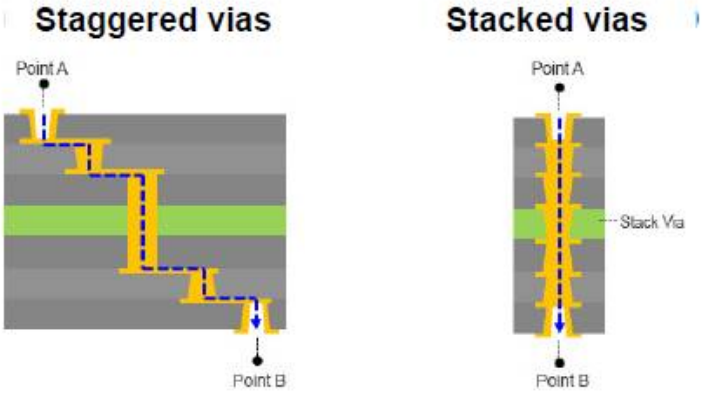




Routing

Use of staggered vias instead of stacked vias

- Slight degradation of signal integrity
- But more subcontractors able to manufacture the card



Stackup

- 14 layers
- 70µ thick planes for power
- HR408 high speed PCB
- More than 10000 vias among which 67% are microvias
- ~ 1750 components