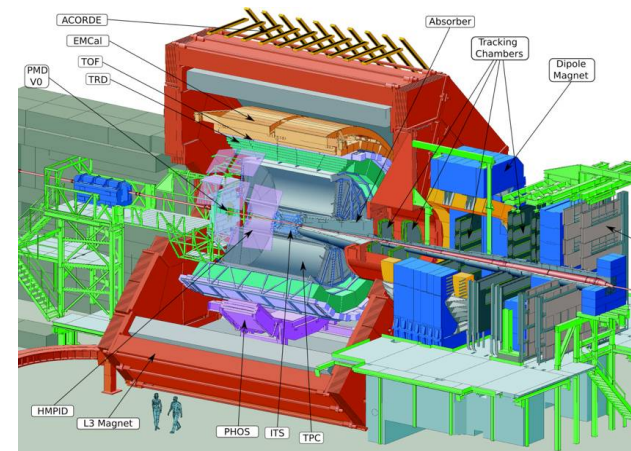
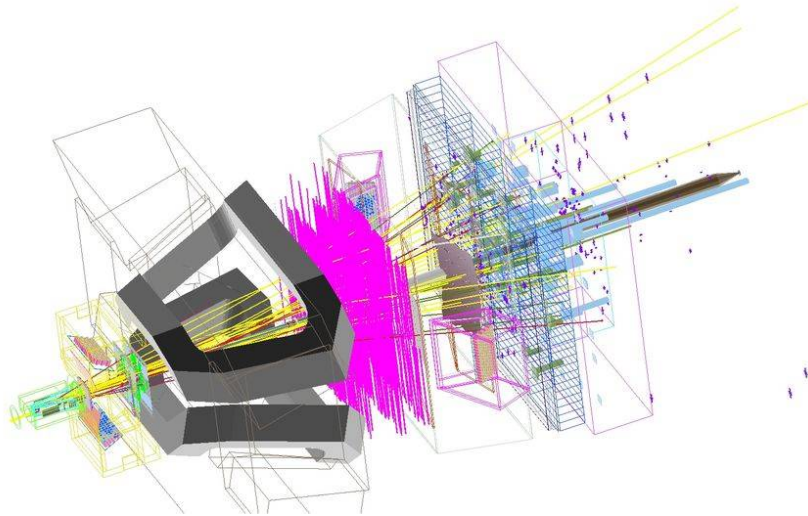




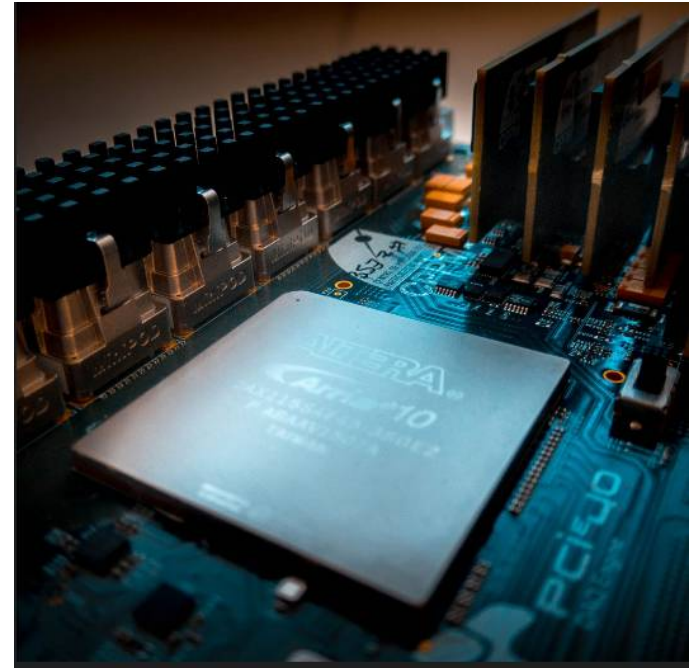
PCIe40: A Common Readout Board for LHCb and ALICE



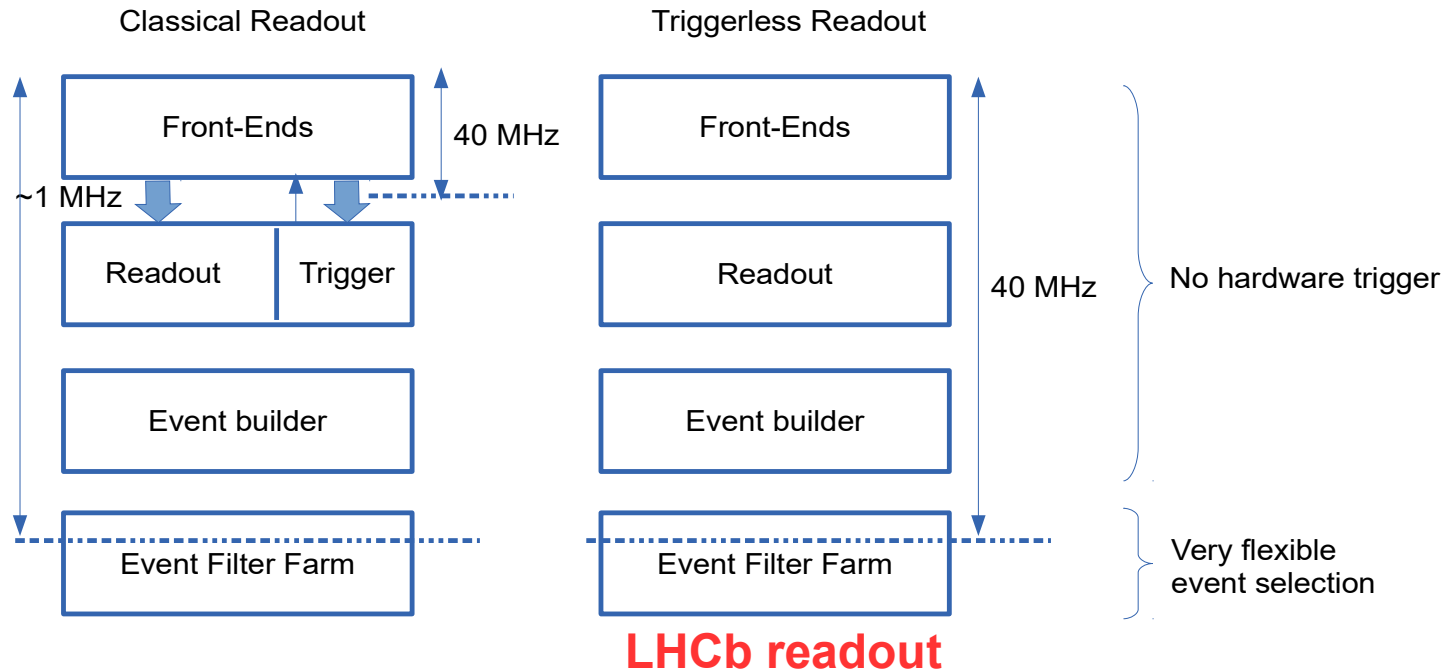
J.P. Cachemiche, on behalf of the LHCb collaboration

Outline

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



LHCb Upgrade key features

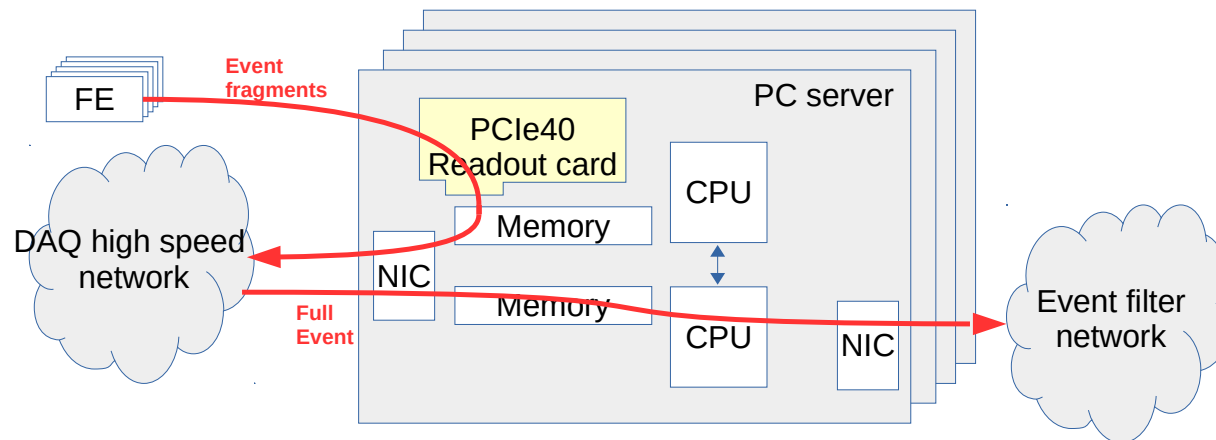


- LHCb uses a [triggerless readout](#)
- All event fragments routed at 40 MHz up to the farm

LHCb Upgrade key features

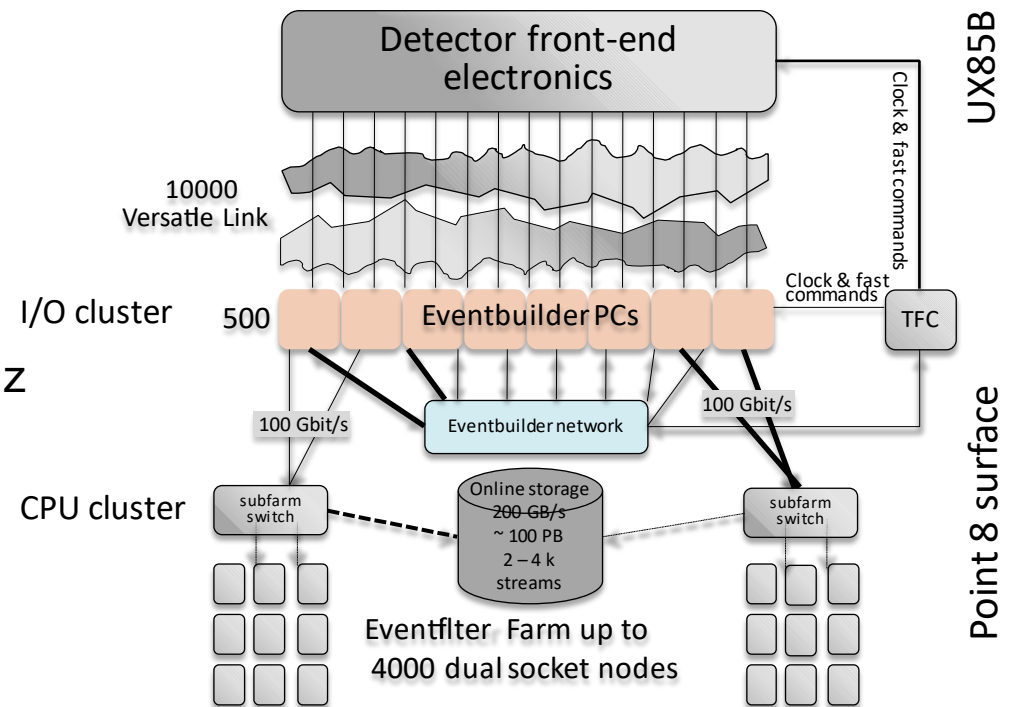
Principle

- 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
 - o 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

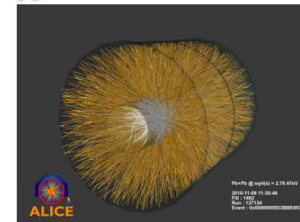
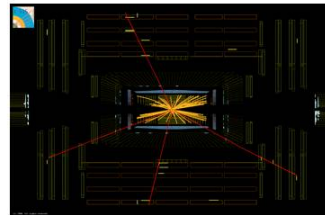


- **At present (Run1 & 2)**

- Interaction rate 8 kHz (Not all LHC bunches have collisions) → max. trigger rate < 3.5 kHz

- **Why low interaction rate?**

- Event topology too complex for simple electronics triggers



3 TB/s data in Run 3

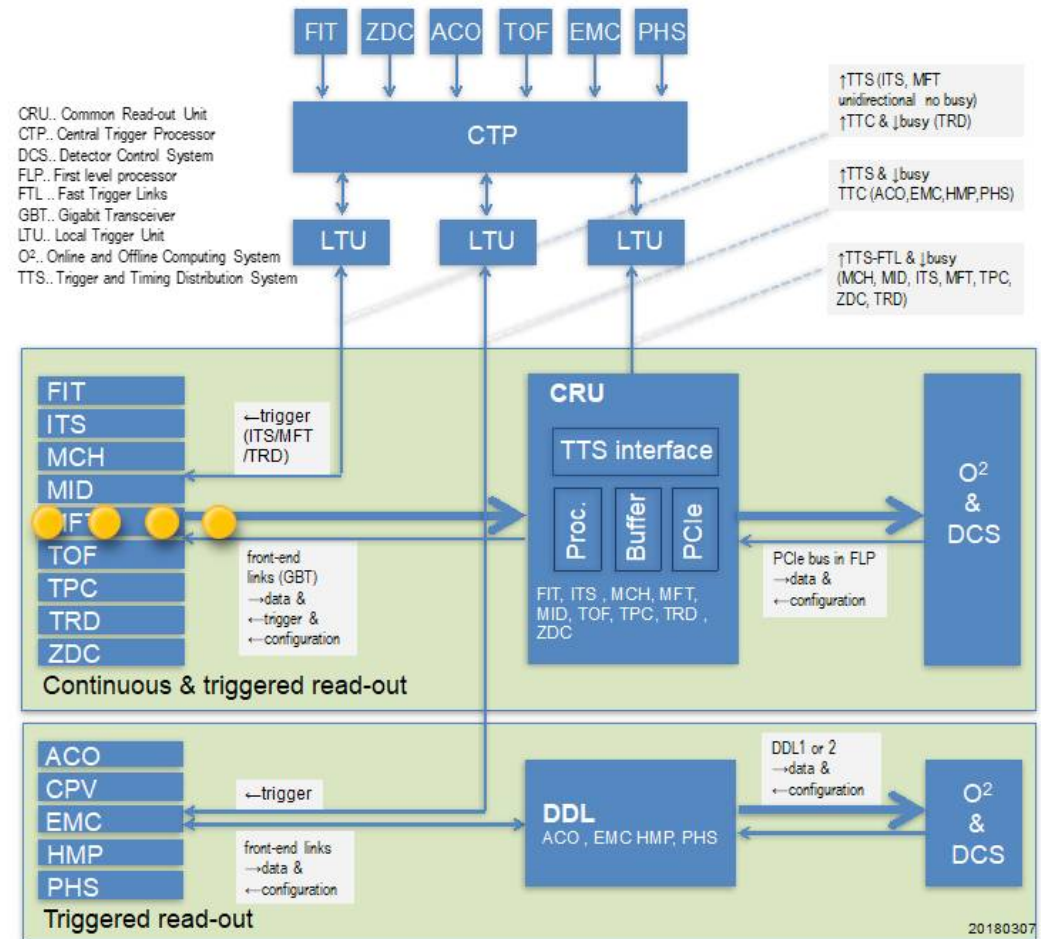
- **After upgrade (≥ Run 3)**

- Target
 - Pb-Pb $\geq 10 \text{ nb}^{-1}$ → 9×10^{10} events
 - pp (@5.5 TeV) $\geq 6 \text{ pb}^{-1}$ → 1.4×10^{11} events
 - Gain factor 100 in statistics
- Interaction rate 50 kHz (PbPb) → continuous triggerless read-out

Courtesy Alex Kluge

ALICE architecture

- Readout located on surface
 - o 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

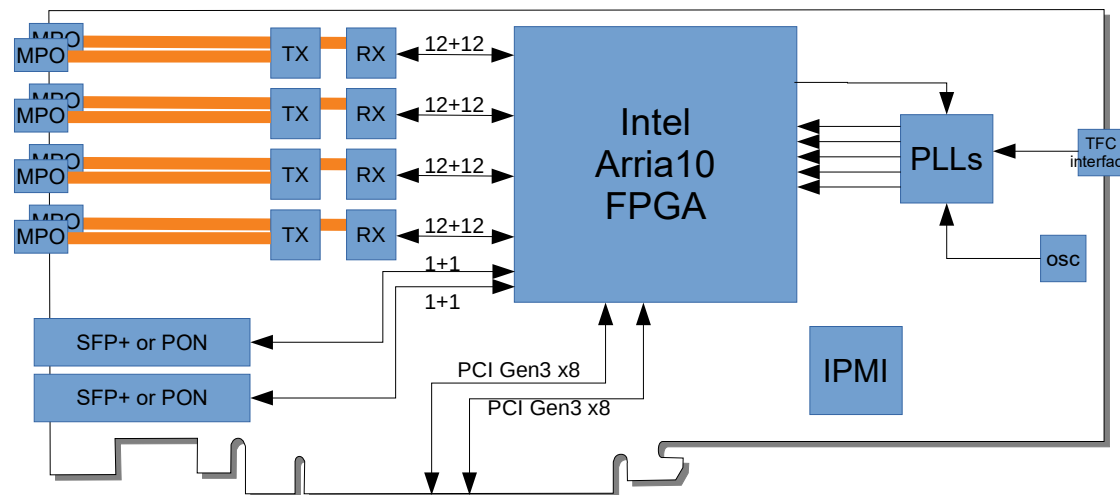


Courtesy Alex Kluge

The readout board : PCIe40

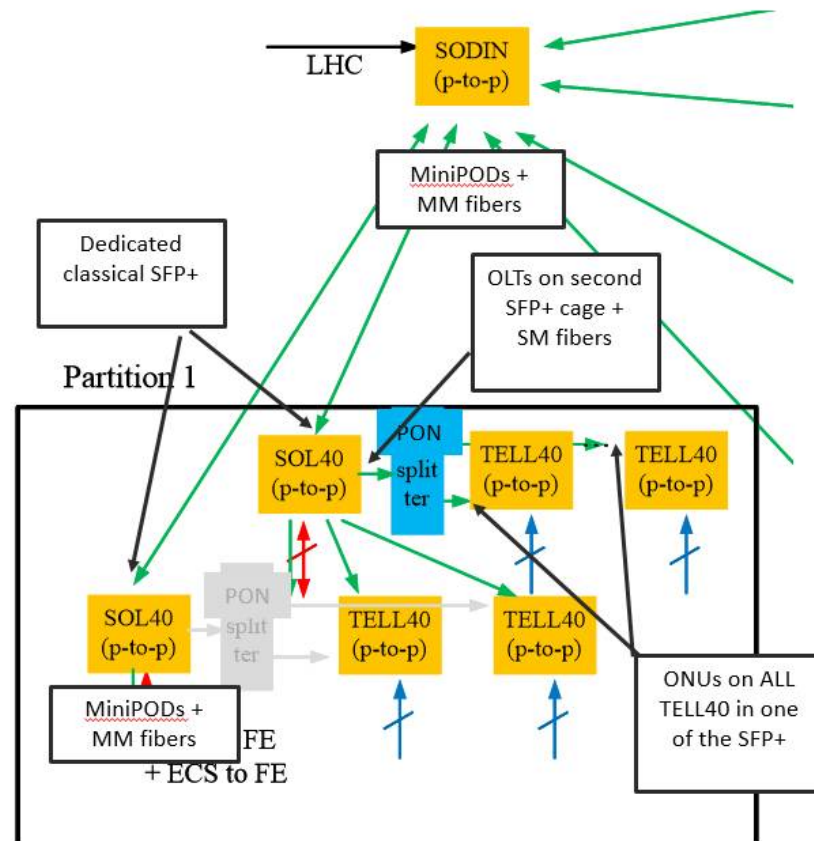
- Features :

- 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

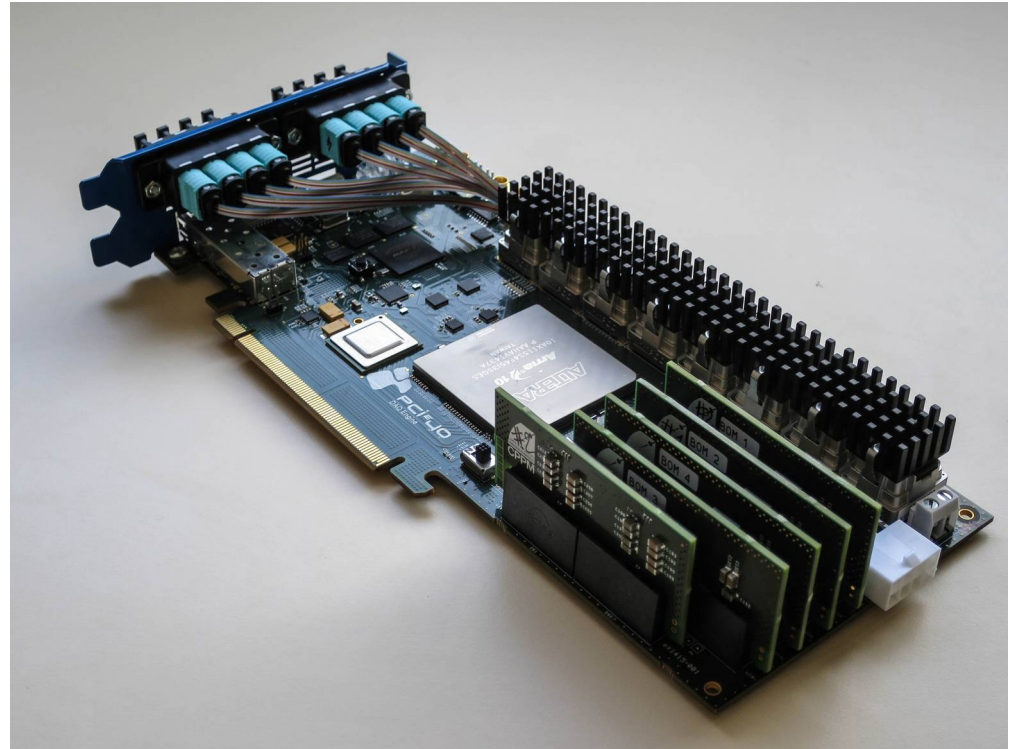
- Can be mapped over several functions by reprogramming the FPGA
- Different names for the same card in LHCb according to its programming :
 - o SODIN : Timing distribution and Fast Control
 - o SOL40 : Slow control
 - o TELL40 : Acquisition
- Minipods for interfaces with Front Ends
 - o GBT protocol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



Hardware design

PCle40 prototype

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



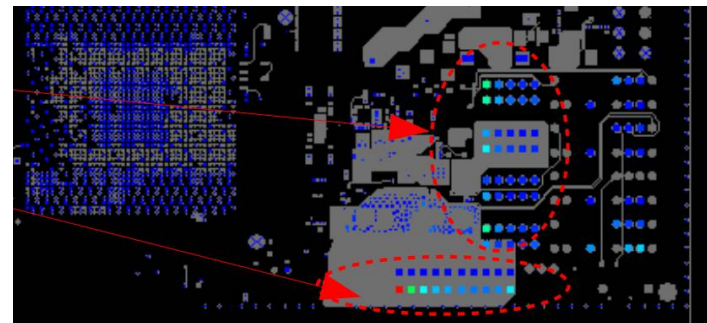
Preparing the final module

Power consumption of large FPGAs very high

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

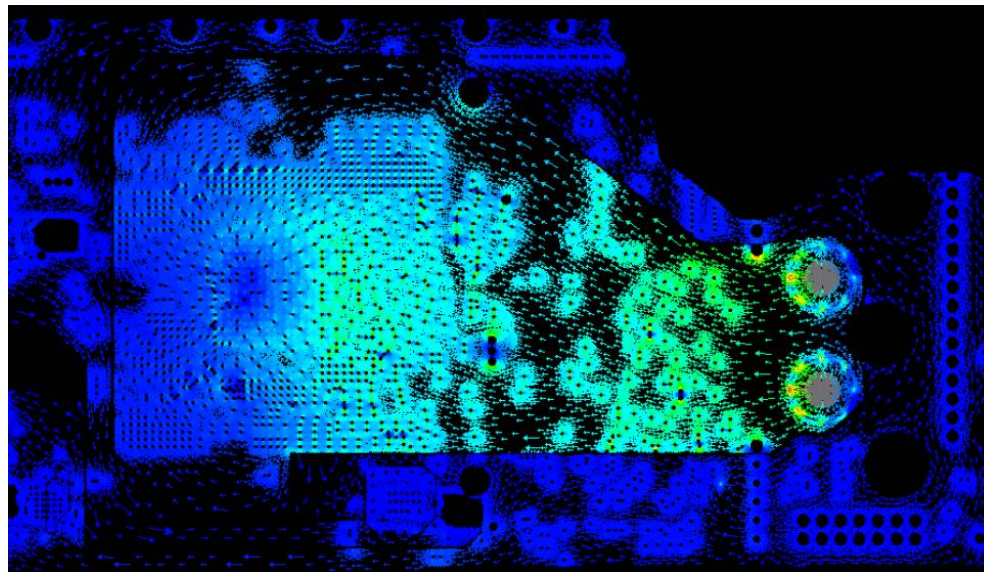
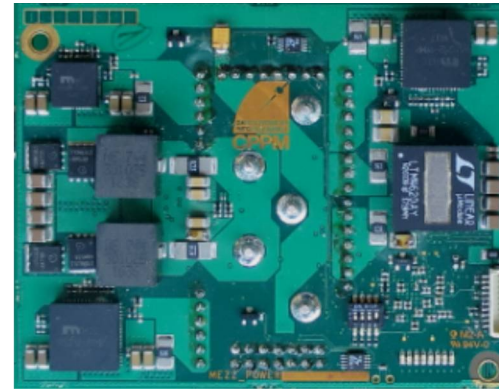
Refining of current flow simulations

- 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 symmetric

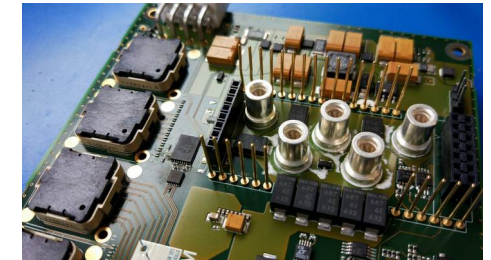


Preparing the final module

Replacement of the 5 vertical mezzanines by a single flat one



Current flow between mezzanine and FPGA with new design



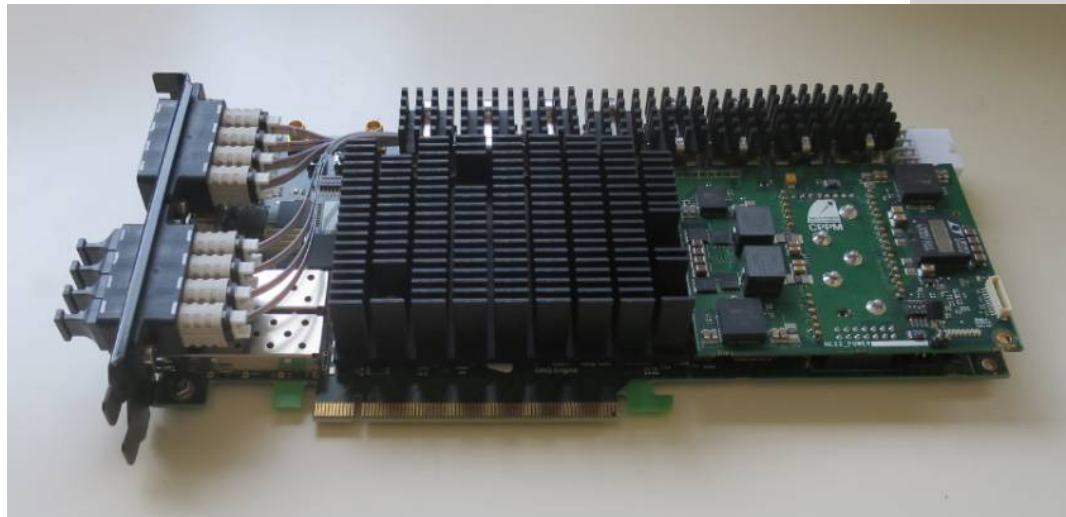
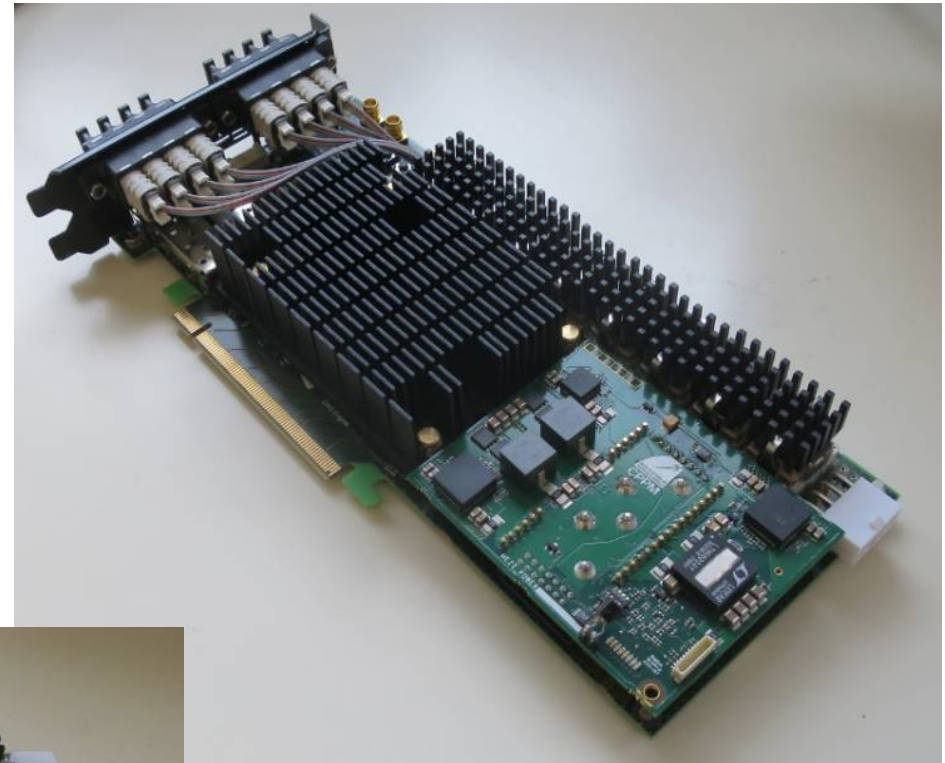
Optimizations

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 symmetrical 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 supply if temperature is too high
 - Serial flash for identifying modules during production

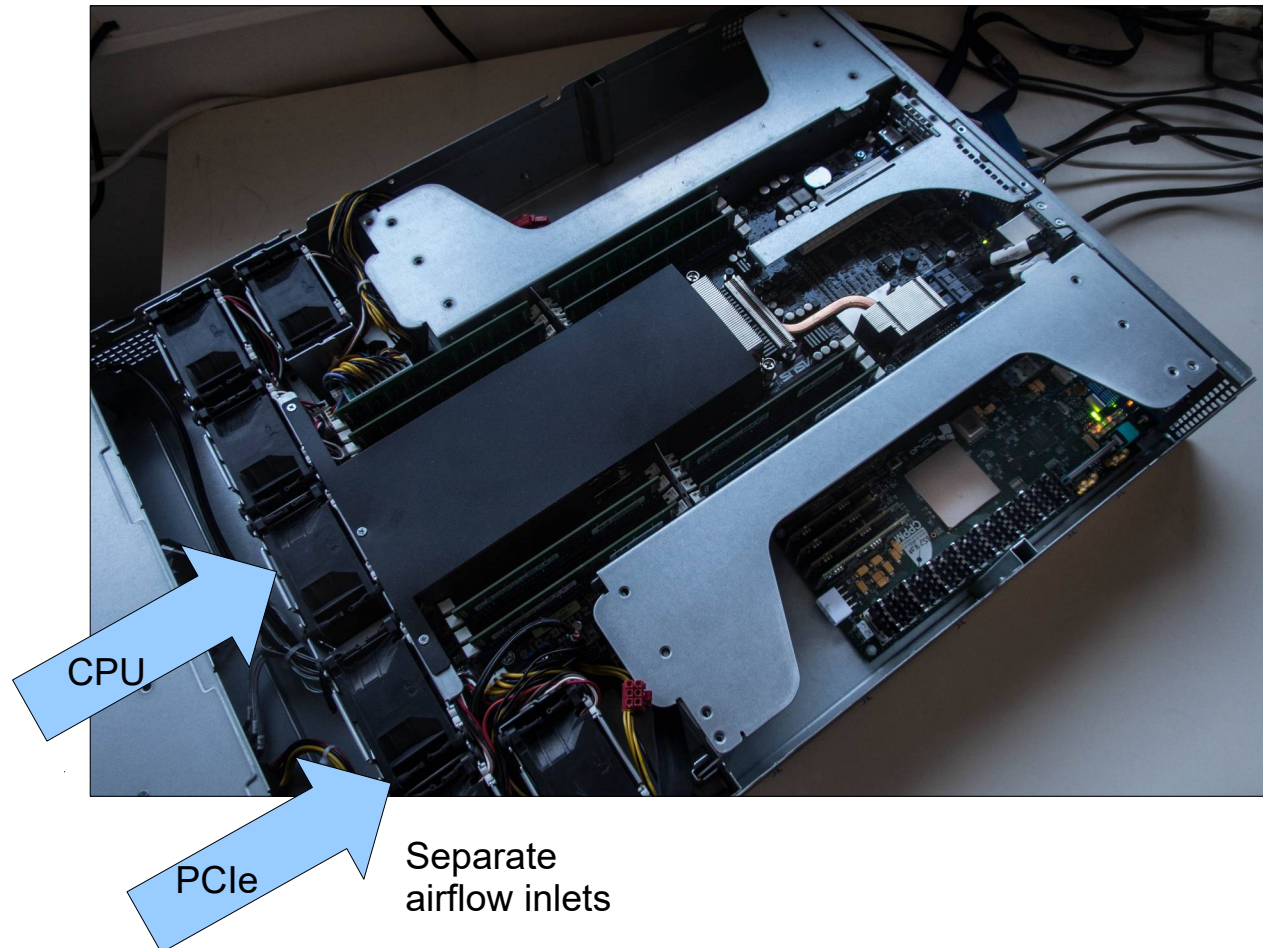
Final module

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



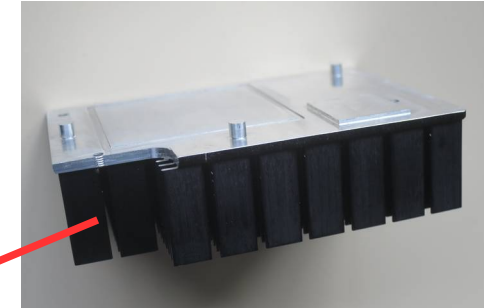
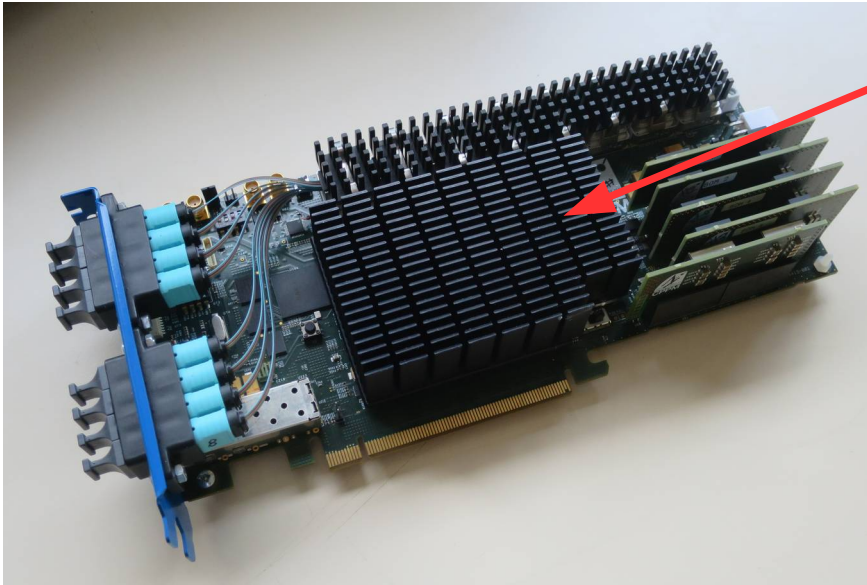
Cooling

- 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

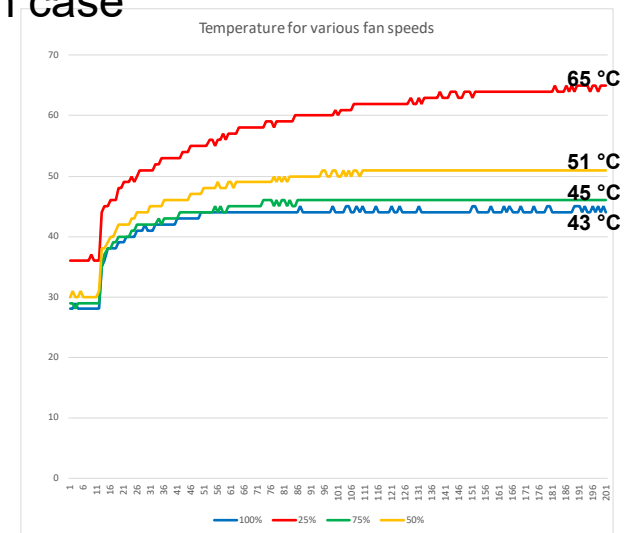
Power consumption and cooling

- 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



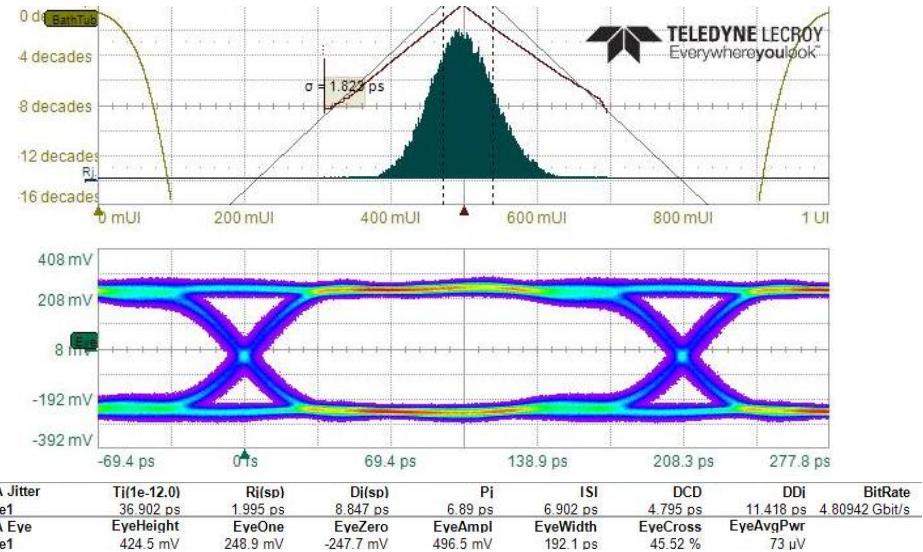
FPGA temperature for several fan speeds

Links measurements

BER $\ll 10^{-16}$

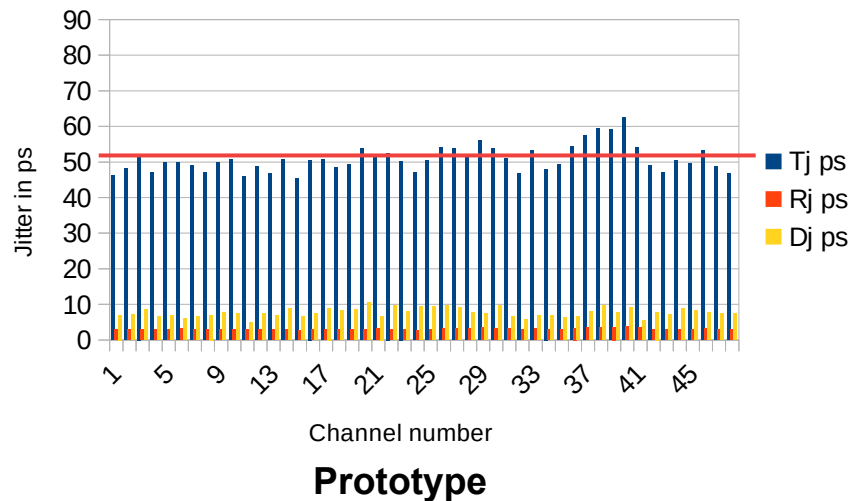
Jitter

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

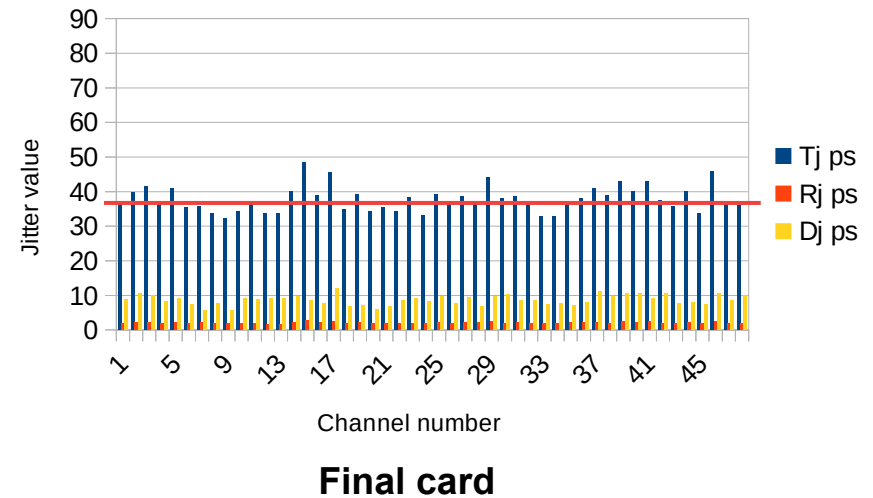


Measurements at reception stage
for a PRBS31 pattern
running at 4.8 Gbits/s

Jitter measurement over 48 links



Jitter measurement over 48 links



Production

Production

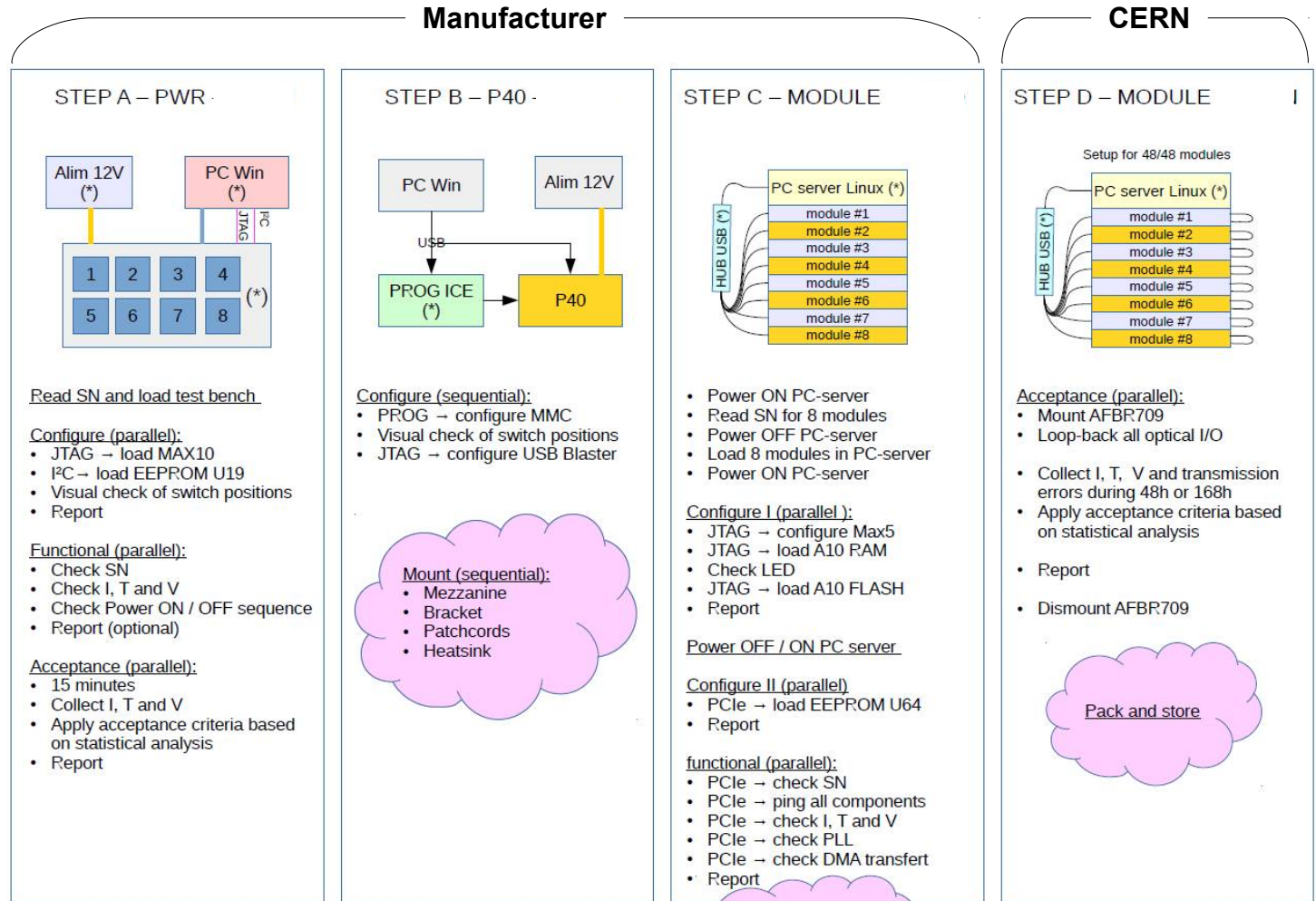
LHCb production started

- ~700 modules in 3 batches :
 - o Preseries of 24 cards
 - o First batch of 330 cards
 - o Second batch of 345 cards
- Schedule
 - o Preseries July 2018
 - o First batch November 2018
 - o Third batch April 2019

Alice should follow a similar route

Testing methodology

4 steps



(*) Hardware provided by CERN



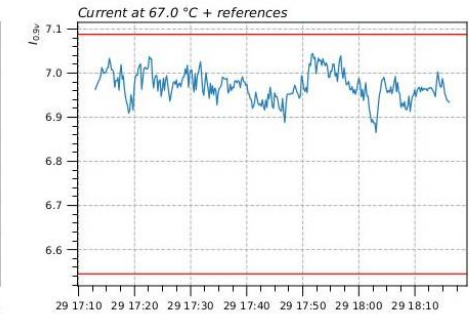
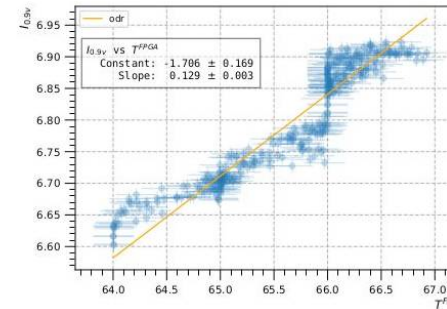
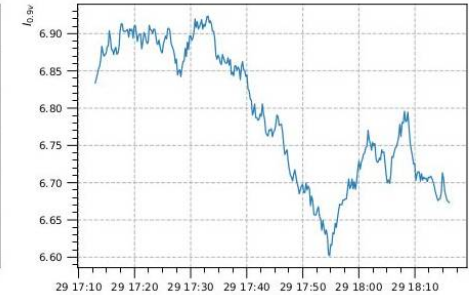
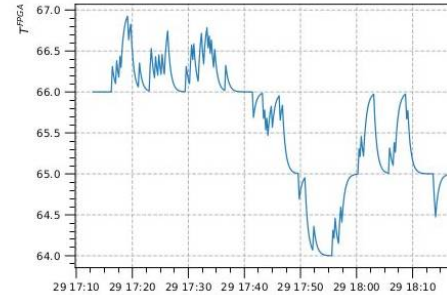
Acceptance tests

Run at CERN

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

$I_{0.9V}$ -- p40_tv20pr006 -- 0mhz -- 2018-03-29T16:57:00

Apr 18, 2018 16:11
Day 108 -- week 16

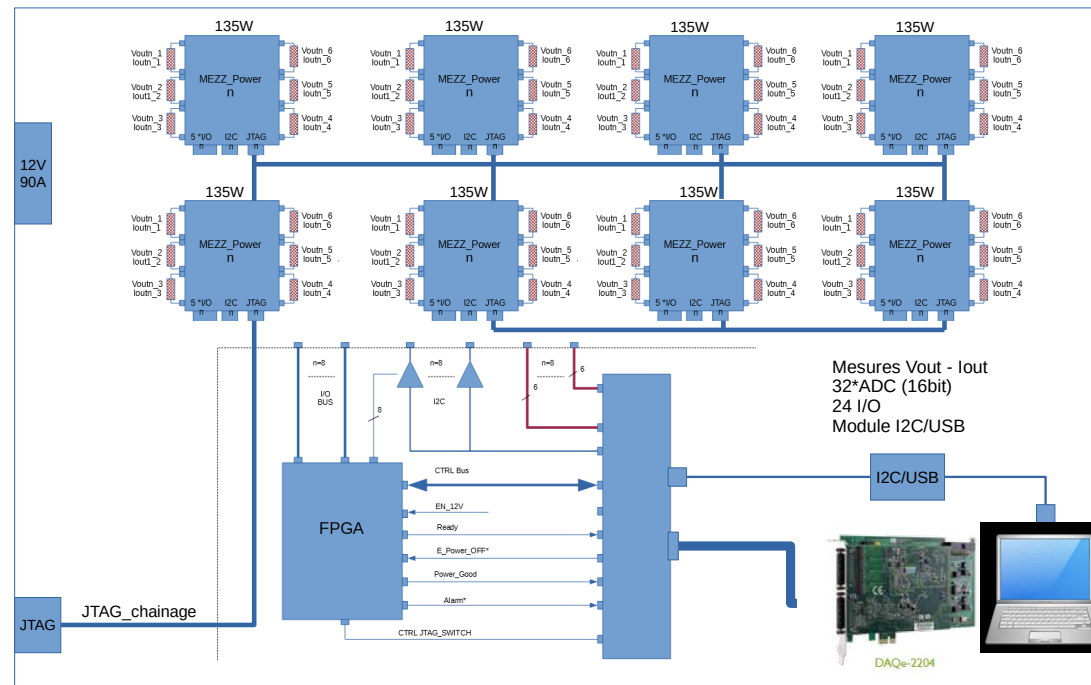


	obs	rsquared	alpha	cl-alpha	current	cl-current	p-value	sigma	cl-sigma
0	0.9V	0.004	-0.104	[-0.577, 0.369]	8.193	[8.17, 8.217]	0.0	0.115	[0.094, 0.137]
1	1.02Vccr	0.073	0.026	[0.001, 0.051]	6.33	[6.326, 6.334]	0.475	0.015	[0.012, 0.018]
2	1.02Vcct	0.041	0.018	[-0.006, 0.041]	1.977	[1.975, 1.979]	0.502	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
 - o Connected to commercial ADC card driven by a Windows PC
 - o Allows to test the cards at full load



Production setup for testing modules

Same approach for the full module

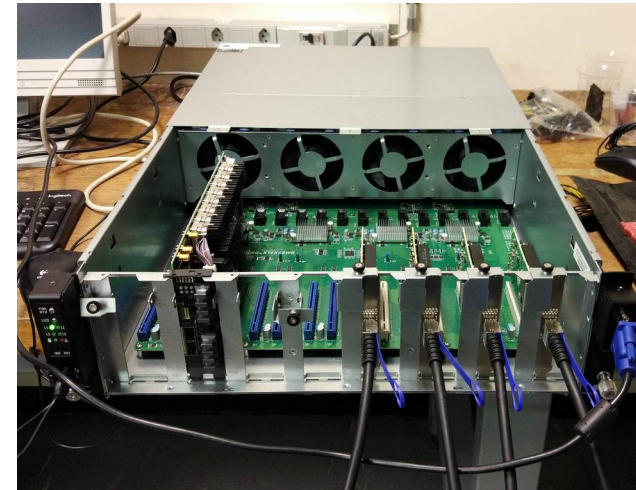
- PCIe crate expander or servers
- On going evaluation



Cubix crate expander



ASUS server

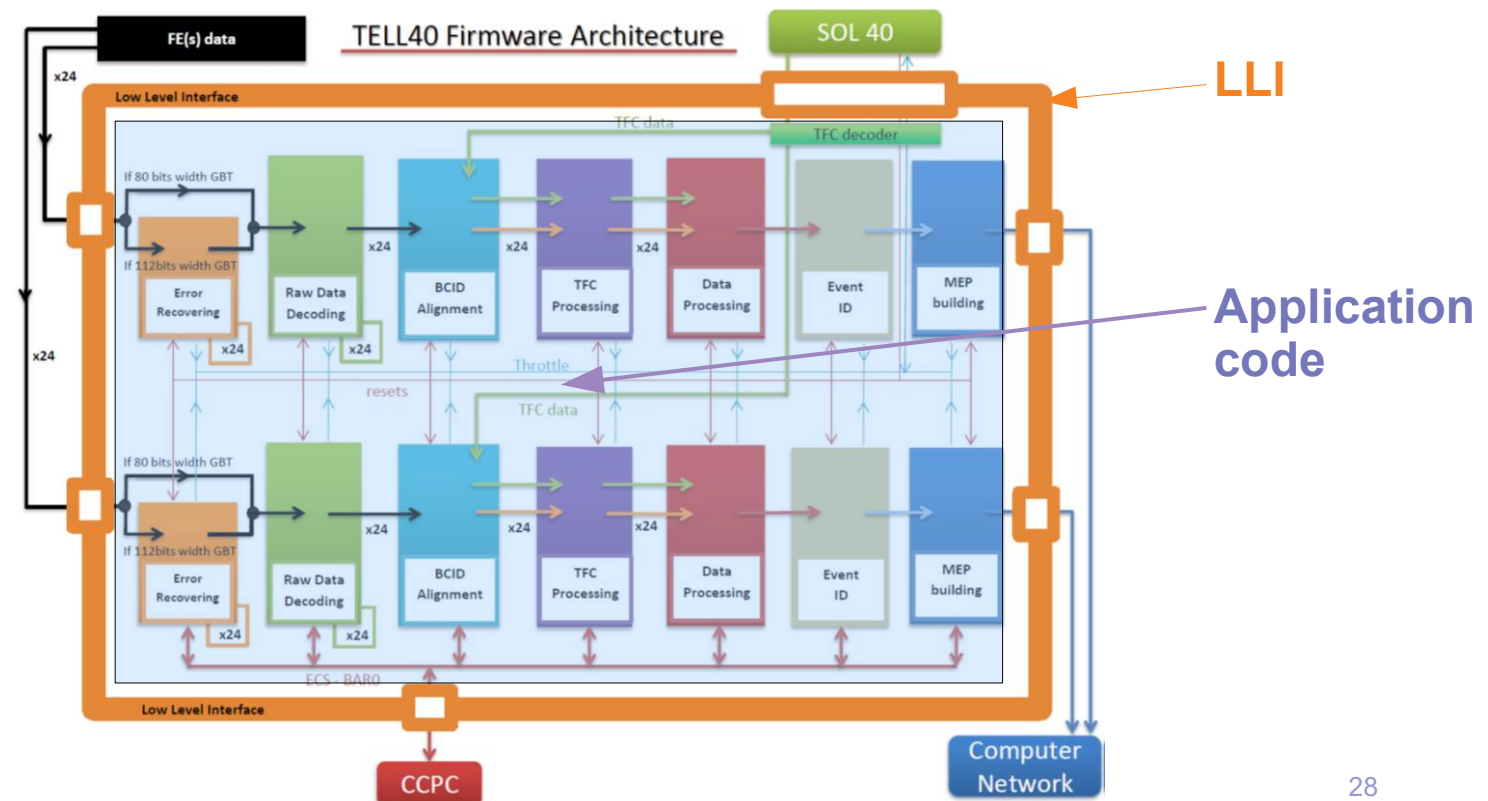


ASRock server

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

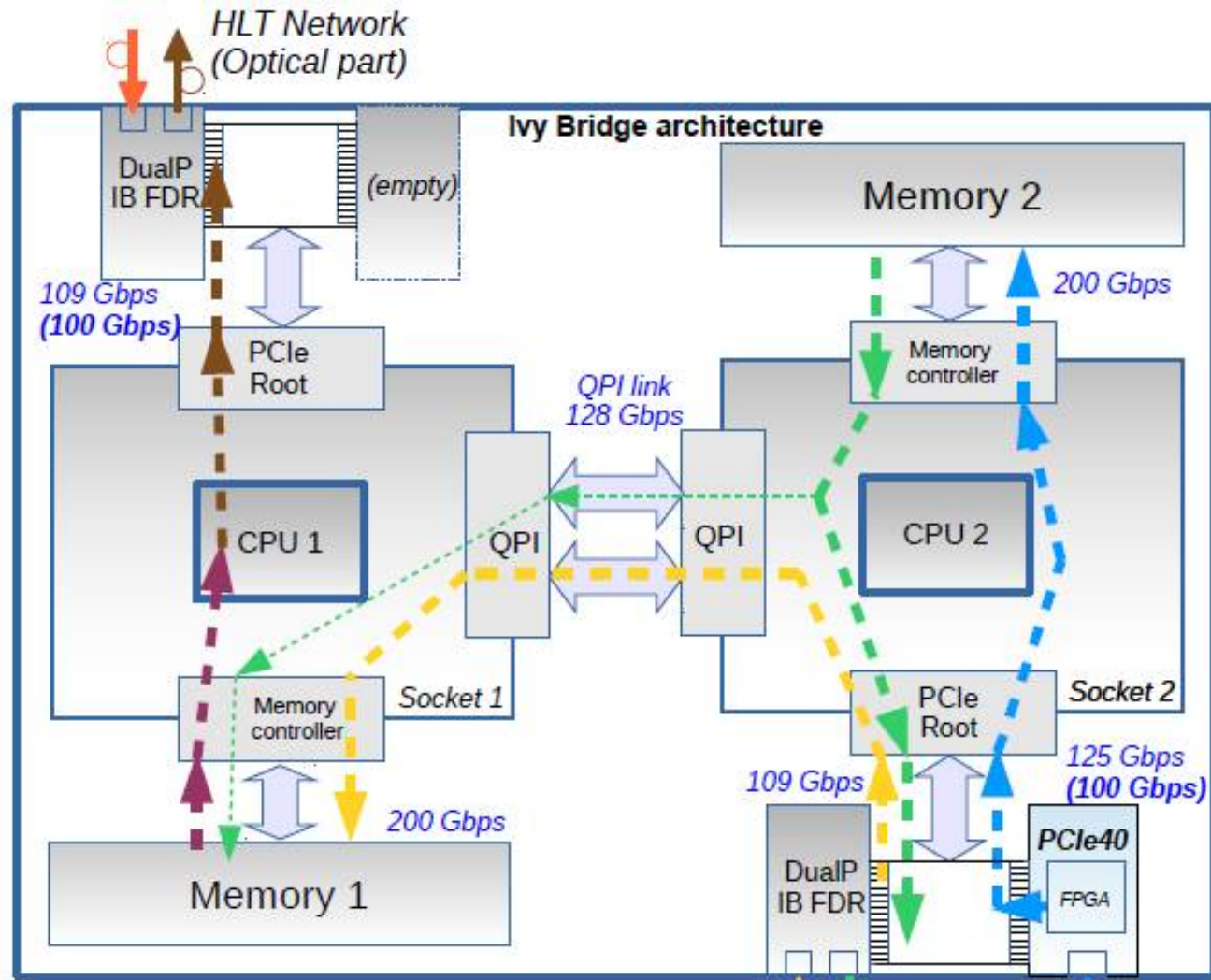


Conclusion

- Cards addressing 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 μ 3E 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

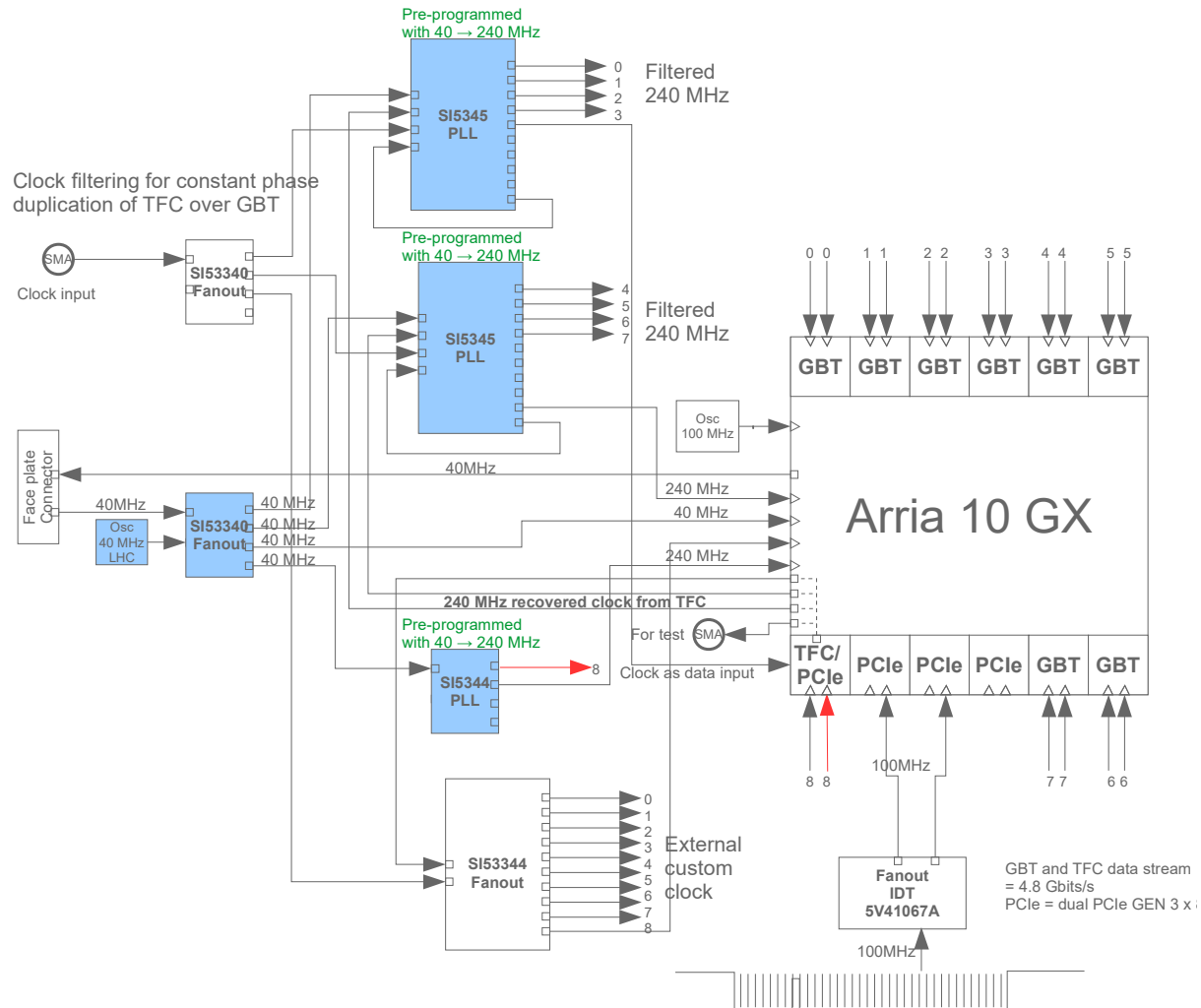


Thanks to Niko, Paolo, Rainer and online team

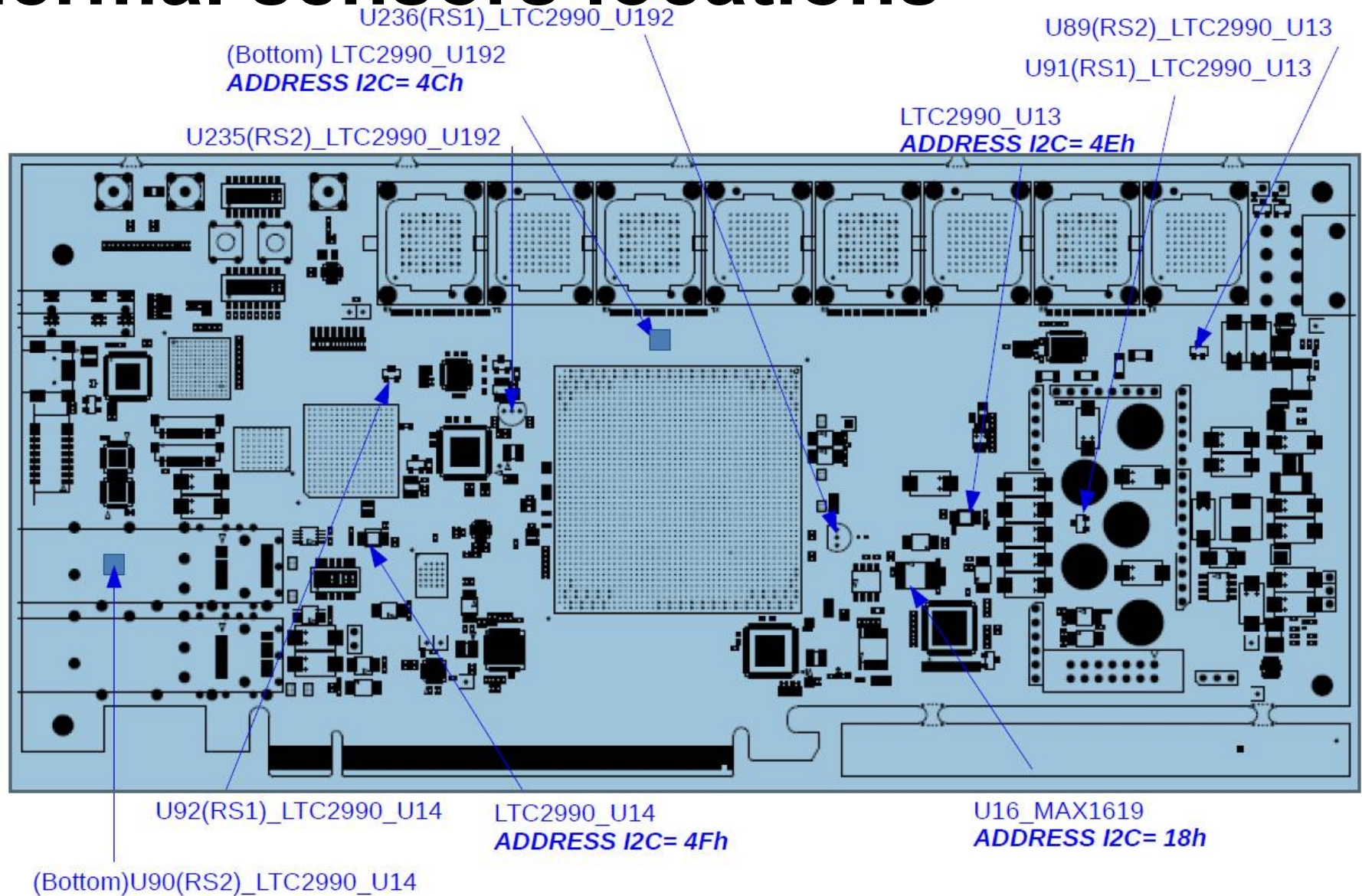
Event building Network (Optical part) Front-End (Optical part)

Clock distribution

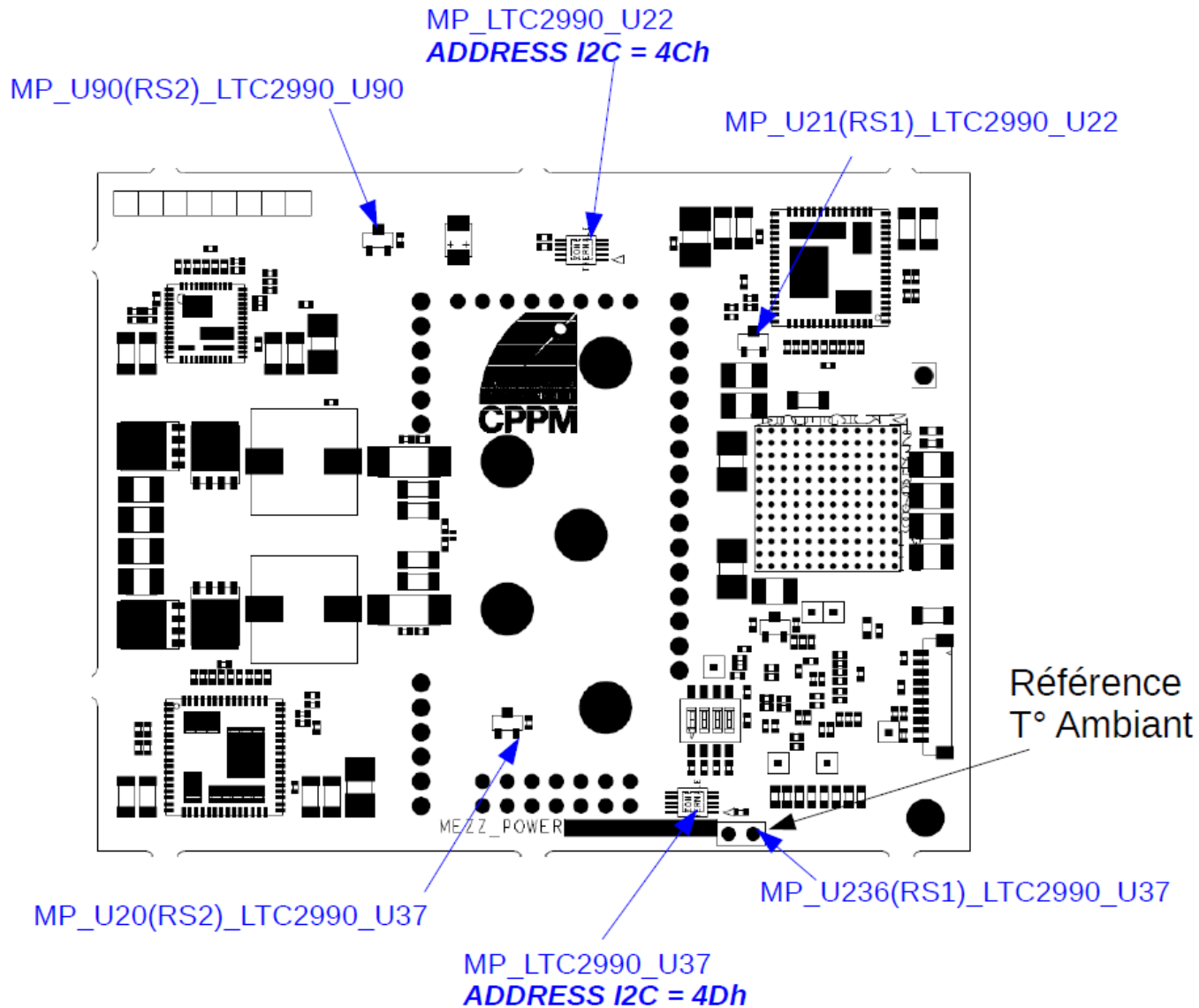
Clock Tree PCIe40V2



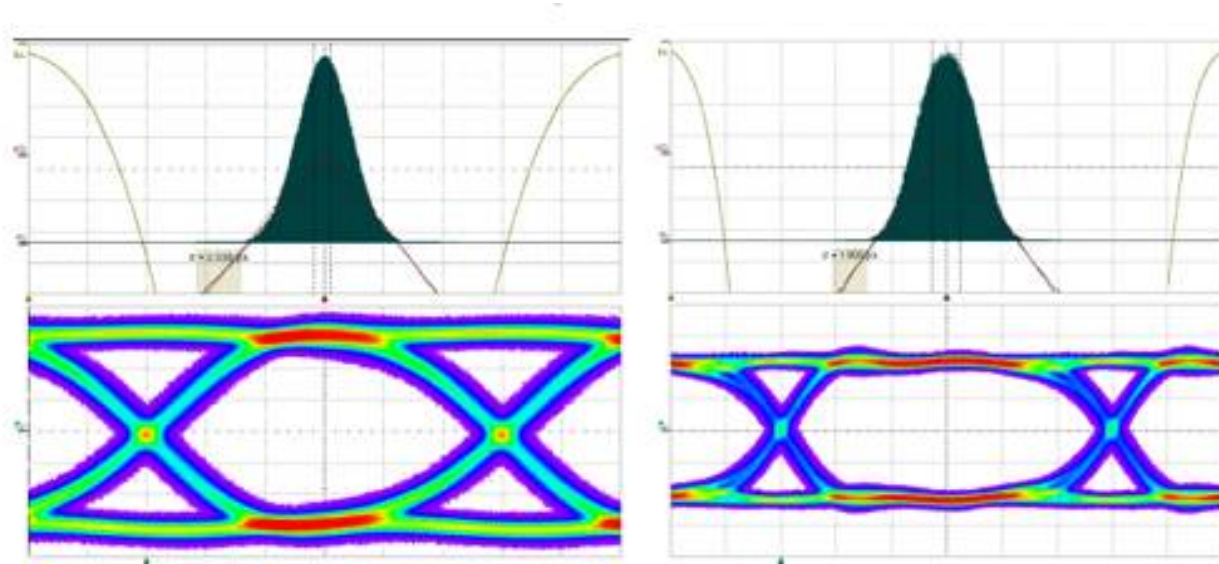
Thermal sensors locations



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



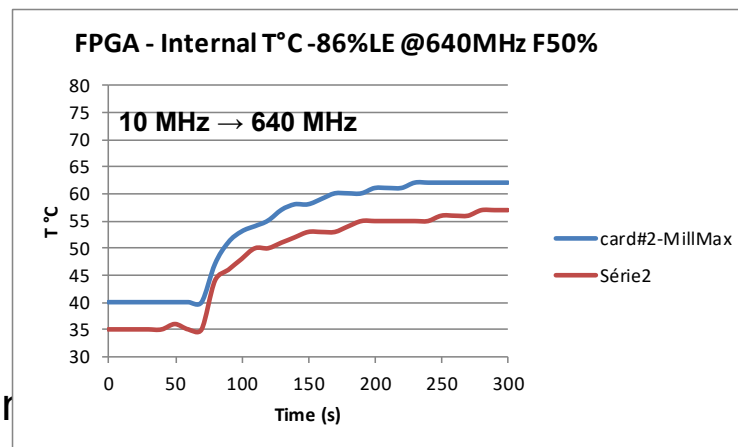
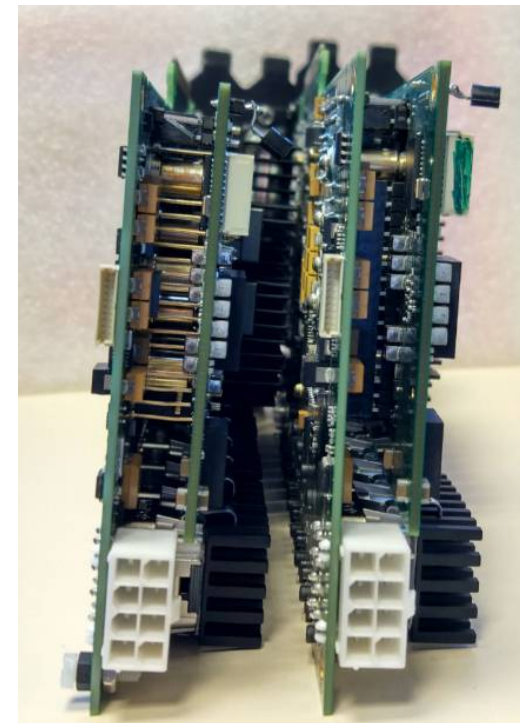
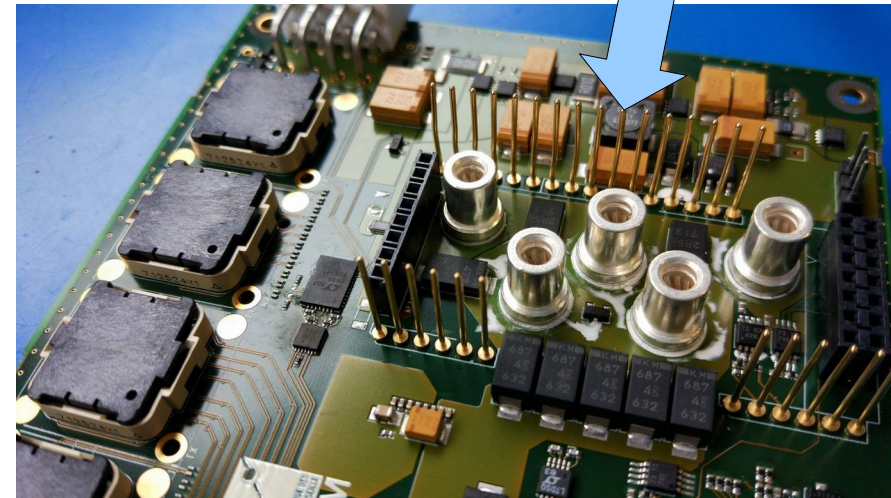
Mezzanine connector

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 ?

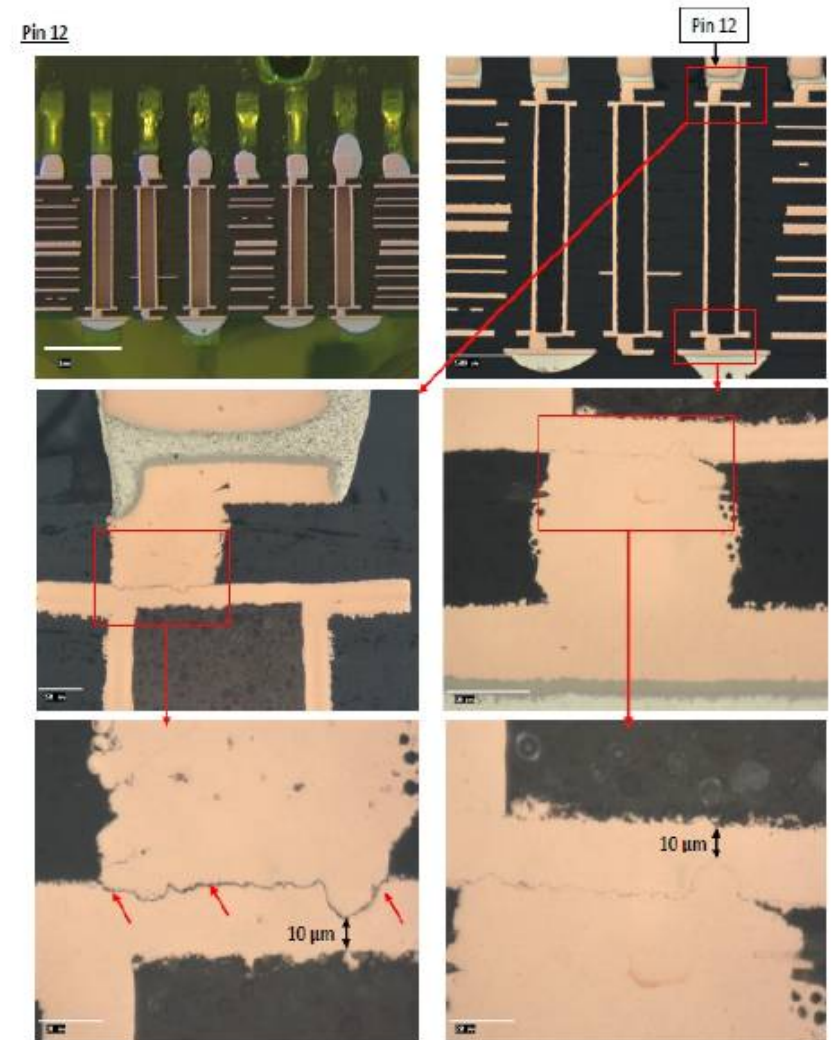


- Fir

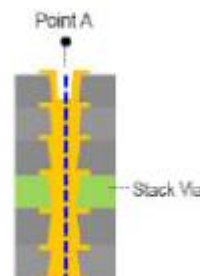
→ Much easier to mount

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.



Stacked vias

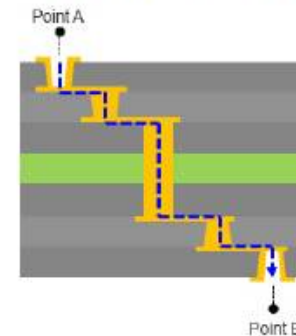


Routing

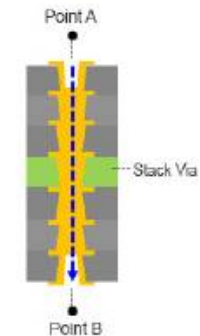
Use of staggered vias instead of stacked vias

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

Staggered vias



Stacked vias

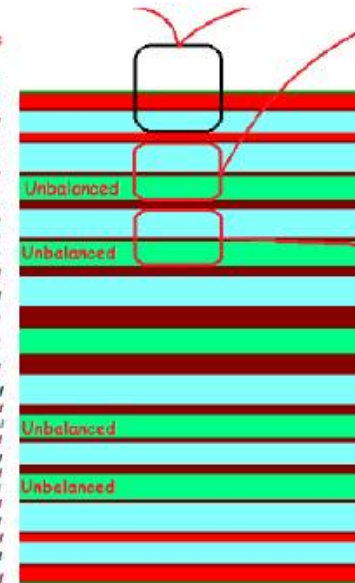


Stackup

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

PCB stackup & manufacturing requirements

SM above CU: 20μ +/-15μ registration 50μ
 CF: 12μ + plating = 55μ +/-10μ
 1 x 1086 RC 62% = 70μ +/-10μ
 CF: 12μ + plating = 40μ +/-10μ
 2 x 106 RC 70% = 90μ +/- 13μ
 CU: 17μ +/-5μ
 CCL: 1 x1086 RC 58% = 75μ +/- 13μ
 CU: 35μ +/-10μ
 2 x 106 RC 70% = 80μ +/- 13μ
 CU: 17μ +/-5μ
 CCL: 1 x1086 RC 58% = 75μ +/- 13μ
 CU: 35μ +/-10μ
 2 x 106 RC 70% = 80μ +/-13μ
 CU: 70μ +/-16μ
 CCL: 1 x1086 RC 58% = 75μ +/- 13μ
 CU: 70μ +/-16μ
 2 x 106 RC 70% = 80μ +/-13μ
 CU: 35μ +/-10μ
 CCL: 1 x1086 RC 58% = 75μ +/- 13μ
 CU: 17μ +/-5μ
 2 x 106 RC 70% = 85μ +/- 13μ
 CU: 35μ +/-10μ
 CCL: 1 x1086 RC 58% = 75μ +/- 13μ
 CU: 17μ +/-5μ
 2 x 106 RC 70% = 90μ +/- 13μ
 CF: 12μ + plating = 40μ +/-10μ
 1 x 1086 RC 62% = 70μ +/-10μ
 CF: 12μ + plating = 55μ +/-10μ
 SM above CU: 20μ +/-15μ registration 50μ



CAD Design rules & Copper balancing requirements

Bridge mini 100μ, Registration : QFN 50μ, BGA 50μ
 Top + W/Smimi 110μ + GND = Layout 60% = L1
 Full GND Smimi 100μ = Layout 90% = L2
 W/Smimi 90μ + GND = Layout 50% = L3
 Full GND Smimi 100μ = Layout 90% = L4
 Wmini 80μ, Smimi 90μ + GND = Layout 50% = L5
 VCC1/2 Smimi 120μ = layout 80% = L6
 VCC 0.9V + VCC 12V, Smimi 240μ = layout 80% = L7
 Full GND Smimi 240μ = Layout 90% = L8
 VCC5/6 Smimi 120μ = layout 80% = L9
 Wmini 80μ, Smimi 90μ + GND = Layout 60% = L10
 Full GND Smimi 100μ = Layout 90% = L11
 W/Smimi 90μ + GND = Layout 50% = L12
 Full GND Smimi 100μ = Layout 90% = L13
 Bottom + W/Smimi 110μ + GND = Layout 70% = L14
 Bridge mini 100μ, Registration : QFN 50μ, BGA 60μ