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
- ~10,000 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

After upgrade (≥ Run 3)
- Target
  - Pb-Pb ≥ 10 nb⁻¹ → 9 x 10¹⁰ events
  - pp (@5.5 TeV) ≥ 6 pb⁻¹ → 1.4 x 10¹¹ events
  - Gain factor 100 in statistics
- Interaction rate 50 kHz (PbPb) → continuous triggerless read-out

3 TB/s data in Run 3

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 distribution 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
PCle40 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 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

- 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
BER $\ll 10^{-16}$

**Jitter**

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

![Jitter measurement over 48 links](image1)

**Prototype**

![Jitter measurement over 48 links](image2)

**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

Manufacturer

CERN

STEP A - PWR:
- Alim 12V (*)
- PC Win (*)
1 2 3 4
5 6 7 8 (*)
Read SN and load test bench
Configure (parallel):
- JTAG — load MAX10
- PC — load EEPROM UI9
- Visual check of switch positions
- Report
Functional (parallel):
- Check SN
- Check I, T and V
- Check Power ON / OFF sequence
- Report (optional)
Acceptance (parallel):
- 15 minutes
- Collect I, T and V
- Apply acceptance criteria based on statistical analysis
- Report

STEP B - P40:
- Alim 12V
- PC Win
- PROG ICE (*)
Configure (sequential):
- PROG — configure MMC
- Visual check of switch positions
- JTAG — configure USB Blaster

STEP C - MODULE
- Power ON PC-server
- Read SN for 8 modules
- Power OFF PC-server
- Load 8 modules in PC-server
- Power ON PC-server
Configure I (parallel):
- JTAG — configure Max5
- JTAG — load A10 FAM
- Check LED
- JTAG — load A10 FLASH
- Report
Power OFF / ON PC server
Configure II (parallel):
- PCIe — load EEPROM UI64
- Report
Functional (parallel):
- PCIe — check SN
- PCIe — ping all components
- PCIe — check I, T and V
- PCIe — check PLL
- PCIe — check DMA transfer
- Report

Acceptance (parallel):
- Mount AFFBP709
- Loop-back all optical I/O
- Collect I, T, V and transmission errors during 48h or 108h
- Apply acceptance criteria based on statistical analysis
- Report
- Dismount AFFBP709

Pack and store

(*) Hardware provided by CERN

ACES 2018

PCle40: A Common Readout Board for LHCb and ALICE
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
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

<table>
<thead>
<tr>
<th>obs</th>
<th>r_squared</th>
<th>alpha</th>
<th>cl-alpha</th>
<th>current</th>
<th>cl-current</th>
<th>p-value</th>
<th>sigma</th>
<th>cl-sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9V</td>
<td>0.004</td>
<td>-0.204</td>
<td>[-0.577, 0.369]</td>
<td>8.193</td>
<td>[8.117, 8.217]</td>
<td>0.0</td>
<td>0.115</td>
<td>[0.094, 0.137]</td>
</tr>
<tr>
<td>1.02Wcorr</td>
<td>0.073</td>
<td>0.026</td>
<td>[0.001, 0.051]</td>
<td>6.33</td>
<td>[6.326, 6.334]</td>
<td>0.475</td>
<td>0.015</td>
<td>[0.012, 0.028]</td>
</tr>
<tr>
<td>1.02Wcut</td>
<td>0.041</td>
<td>0.018</td>
<td>[-0.006, 0.041]</td>
<td>1.977</td>
<td>[1.975, 1.979]</td>
<td>0.502</td>
<td>0.008</td>
<td>[0.005, 0.007]</td>
</tr>
<tr>
<td>1.8V</td>
<td>0.019</td>
<td>0.021</td>
<td>[-0.021, 0.063]</td>
<td>7.011</td>
<td>[7.007, 7.014]</td>
<td>0.809</td>
<td>0.01</td>
<td>[0.008, 0.012]</td>
</tr>
<tr>
<td>1.8Va10</td>
<td>0.021</td>
<td>0.02</td>
<td>[-0.017, 0.057]</td>
<td>3.627</td>
<td>[3.625, 3.63]</td>
<td>0.729</td>
<td>0.009</td>
<td>[0.008, 0.011]</td>
</tr>
<tr>
<td>1.8Vcorr</td>
<td>0.004</td>
<td>0.01</td>
<td>[-0.031, 0.05]</td>
<td>1.458</td>
<td>[1.454, 1.462]</td>
<td>0.845</td>
<td>0.01</td>
<td>[0.008, 0.012]</td>
</tr>
<tr>
<td>2.5V</td>
<td>0.035</td>
<td>0.074</td>
<td>[-0.032, 0.179]</td>
<td>2.805</td>
<td>[2.8, 2.809]</td>
<td>0.003</td>
<td>0.025</td>
<td>[0.002, 0.029]</td>
</tr>
<tr>
<td>3.3V</td>
<td>0.001</td>
<td>-0.008</td>
<td>[-0.091, 0.074]</td>
<td>1.929</td>
<td>[1.923, 1.936]</td>
<td>0.881</td>
<td>0.02</td>
<td>[0.017, 0.024]</td>
</tr>
<tr>
<td>12V</td>
<td>0.018</td>
<td>0.069</td>
<td>[-0.069, 0.207]</td>
<td>2.883</td>
<td>[2.874, 2.892]</td>
<td>0.523</td>
<td>0.035</td>
<td>[0.029, 0.042]</td>
</tr>
<tr>
<td>12Vbias</td>
<td>0.017</td>
<td>-0.003</td>
<td>[-0.01, 0.004]</td>
<td>-0.021</td>
<td>[-0.021, -0.02]</td>
<td>0.032</td>
<td>0.002</td>
<td>[0.001, 0.002]</td>
</tr>
</tbody>
</table>
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

![Diagram of test setup](image-url)
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 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

Clock filtering for constant phase duplication of TFC over GBT.

Pre-programmed with 40 → 240 MHz

Filtered 240 MHz

240 MHz recovered clock from TFC

SMA

Clock as data input

Pre-programmed with 40 → 240 MHz

SMA

40 MHz

40 MHz

40 MHz

240 MHz

240 MHz

240 MHz

100 MHz

100 MHz

100 MHz

For test

GBT and TFC data stream = 4.8 Gbits/s

PCle = dual PCIe GEN 3 x 8

ACES 2018 PCIe40: A Common Readout Board for LHCb and ALICE 32
Thermal sensors locations

- U236(RS1)_LTC2990_U192
- (Bottom) LTC2990_U192
  ADDRESS I2C= 4Ch

- U89(RS2)_LTC2990_U13
- U91(RS1)_LTC2990_U13

- LTC2990_U13
  ADDRESS I2C= 4 Eh

- U235(RS2)_LTC2990_U192

- U92(RS1)_LTC2990_U14

- (Bottom) U90(RS2)_LTC2990_U14
  ADDRESS I2C= 4Fh

- LTC2990_U14
  ADDRESS I2C= 4Fh

- U16_MAX1619
  ADDRESS I2C= 18h
Thermal sensors locations

MP_LTC2990_U22
ADDRESS I2C = 4Ch

MP_U90(RS2)_LTC2990_U90

MP_U21(RS1)_LTC2990_U22

MP_U20(RS2)_LTC2990_U37

MP_LTC2990_U37
ADDRESS I2C = 4Dh

MP_U236(RS1)_LTC2990_U37

Référence
T° Ambiant
Eye diagrams
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?
- Final choice = Samtec
  ➔ 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