The PCIe-based readout system for the LHCb experiment

K. Arnaud, J. P. Cachemiche, P.-Y. Duval, F. Hachon, M. Jevaud, R. Le Gac, F. Rethore
Centre de Physique des Particules de Marseille
Outline

- Triggerless readout
- New readout architecture
- Current prototype PCIe40
- Overall architecture
General architecture

**5 main blocks:**
- Front End electronics
- Eventbuilder PCs
- TFC system
- Online network
- High Level Trigger PCs Farm

LHCb Readout system
Triggerless readout

**Does not mean there is no trigger**

- CPU processing power makes feasible computation of trigger by software:
  - Previous systems were able to process events at 1 Mhz
  - Software trigger implies to acquire data at 40 MHz and therefore « only » 40 times more processing power
  - Much more flexible
Initial architecture choice for LHCb

Triggerless: All events fragments are routed toward a single CPU blade

ATCA based
Memory issue

2 methods to make data converge toward a single CPU:
both require memory

- **Push**: requires expensive switches with memory inside
- **Pull**: requires memory on already very dense back-end boards
Major technological breakthrough

New architecture of Intel chips includes very large bandwidth data paths

- 40 lanes PCIe GEN3 at 8 Gbits/s
- Independent paths to memory

Ivy Bridge architecture
Major technological breakthrough

New architecture of Intel chips includes very large bandwidth data paths

- 40 lanes PCIe GEN3 at 8 Gbits/s
- Independent paths to memory
- Large memory available

Ivy Bridge architecture
Major technological breakthrough

New architecture of Intel chips includes very large bandwidth data paths

- 40 lanes PCIe GEN3 at 8 Gbits/s
- Independent paths to memory
- Large memory available
- Multicore CPUs
  - Powerful enough to handle both Event building and Software Trigger

Ivy Bridge architecture
New readout/trigger scheme

Move back-end card into already existing CPU blades
Advantages & Issue

😊 Advantages

- Large memory in the CPU → simpler acquisition board, cheaper switches
- Possibility to run LLT in the Event Building CPU blades
- No more intermediate crates
- Less optical links

😞 System life duration

- Average life of a PC = ~4y (up to 8y according to CERN statistics)
  ➔ What if PCIe slots not anymore supported by future GENx on PCs?
  Risk to redesign and remanufacture an expensive PCIe40 board.

  ➔ BUT:
  • GEN4 mecanically compatible with GEN3
  • GEN5 ? Probably the futur CPU mother board will be equipped with slots GEN3 or 4
Data path into the computer
Data path into the computer

Ivy Bridge architecture

Socket 1
PCle Root
CPU 1
Memory controller
Memory 1

Socket 2
PCle Root
CPU 2
Memory controller
Memory 2

QPI link 128 Gbps
QPI

PCIe40
FPGA

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

Data path into the computer
Data path into the computer

Ivy Bridge architecture

- PCIe Root
- Memory controller
- Memory 1
- CPU 1
- QPI link 128 Gbps
- Memory 2
- CPU 2
- PCIe Root
- Socket 1
- Socket 2
- PCIe40
- FPGA
- Multiple data flow paths (200 Gbps, 109 Gbps, 125 Gbps)

Event building Network (Optical part)
Front-End (Optical part)
Data path into the computer

Thanks to Niko, Paolo, Rainer and online team

Event building Network (Optical part)

Front-End (Optical part)
PCIE40 Synoptic view

Up to 48 optical inputs/outputs
Serial signal up to 10 Gbps

Bidirectional serial links
for the TFC

output bandwidth ~100 Gbps though PCIe Gen3 x16

► Generic board:
- Data Acquisition from the FE boards (@100Gbps)
- TFC supervisor
- TFC & ECS distribution to/from the FE boards
PCIE40  Challenging board

Design of:

➢ 114 high speed serial signals up to 10 Gbps including PCI Express GEN3x16.

➢ Reduce form factor to be compatible with most CPU blades

➢ Developed with the largest ALTERA FPGA currently available (Arria 10)
  ➢ Use of engineering sample (ES1 & ES2)
  ➢ Production chip still not released

➢ Power dissipation up to 160W
  ➢ 10 different voltages.
  ➢ Precise power sequencing required.
  ➢ Transport high current (50A/0.9V) to the FPGA core.
    ➢ Risk of PCB delamination by thermal effect
Simulation of current propagation on the ground shape of the power supply Mezzanine board – **CADENCE – SIGRITY suite**

Previous routing

Final routing

Very high density
Of current (150A/mm²)

Solution: two separated ground shape
PCIE40 Prototypes

2 prototypes: equipped with Arria10 ES1 & ES2
PCIE40 Flexible front panel optical configuration

8*MTP connectors of 12 fibres

2*MTP connectors of 48 fibres
Server integration

Board is operational in the server
Overall architecture

LHC clock is broadcasted to the Front-Ends

LHC clock is broadcasted to the Front-Ends

Front-Ends

Patch Panels

Patch Panels

300 m

TFC supervisor

PCle40

CPU

~12 or 5

Individual OL or PON (not decided yet)

Timing/Trigger distribution & Slow Control

PCle40

CPU

~50

LHC clock

Front-Ends
Overall architecture

TFC and acquisition on separate boards ➔ Better optimization

Same board used everywhere
Conclusion

Feasibility of a triggerless readout system with event building in the farm has been demonstrated.

Now the base line solution for LHCb experiment.

Two prototypes of PCIe40 boards are fully operational.
- Robustness tests (Temperature, cooling, BER, …) to be done in the next months.

20 Boards will be duplicated early 2016 to provide « Mini-daq » setups for the collaboration.

Full production foreseen 2016/2017.
Extra slides
Power supply to the FPGA CORE
Temperature Simulations

T° ambiant = 25°C
Flux 2m/s
Detection of stressed vias (0.9V – 60A 350 A/mm²!)
Debug status

Most of data paths and main features validated

<table>
<thead>
<tr>
<th>Hardware Functionalities</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td></td>
</tr>
<tr>
<td>Power supplies</td>
<td>✓</td>
</tr>
<tr>
<td>Power supply sequencing</td>
<td>✓</td>
</tr>
<tr>
<td>FPGAs programming</td>
<td></td>
</tr>
<tr>
<td>JTAG</td>
<td>✓</td>
</tr>
<tr>
<td>Integrated USB blaster</td>
<td>✓</td>
</tr>
<tr>
<td>CVP (through PCIe)</td>
<td>on-going</td>
</tr>
<tr>
<td>Flash programming</td>
<td></td>
</tr>
<tr>
<td>Serial</td>
<td>✓</td>
</tr>
<tr>
<td>Parallel</td>
<td>✓</td>
</tr>
<tr>
<td>FPGA peripherals</td>
<td></td>
</tr>
<tr>
<td>I2C, SPI busses</td>
<td>✓</td>
</tr>
<tr>
<td>Main clock PLL</td>
<td>✓</td>
</tr>
<tr>
<td>FPGA temperature monitoring</td>
<td>✓</td>
</tr>
<tr>
<td>Filtered clock PLLs</td>
<td>on-going</td>
</tr>
<tr>
<td>External temperature and current monitoring</td>
<td>on-going</td>
</tr>
<tr>
<td>Serial and optical links</td>
<td></td>
</tr>
<tr>
<td>Minipod optical links</td>
<td>✓</td>
</tr>
<tr>
<td>SFP+ optical link</td>
<td>✓</td>
</tr>
<tr>
<td>PLX PCIe bridge</td>
<td>✓</td>
</tr>
<tr>
<td>PCIe interface Gen2 (ES1)</td>
<td>✓</td>
</tr>
<tr>
<td>PCIe interface Gen3 (ES2)</td>
<td>✓</td>
</tr>
</tbody>
</table>

→ Validated end of August

Link at 10 Gbps

Most of data paths and main features validated

FPGA: Arria10

PCIe Bridge

PC Motherboard

FPGA: Arria10
Data rate computation

Number of PCIe40 needed for the data acquisition:
12000/24 (Optical links) = 500 boards

Single Event by PC :
12000*80bit = 960Kbit (1 collision)
(960Kbit x 40xE6 collisions /s)*2 ~ 77Gbps (data rate)

Data rate + Overhead ~100Gbps