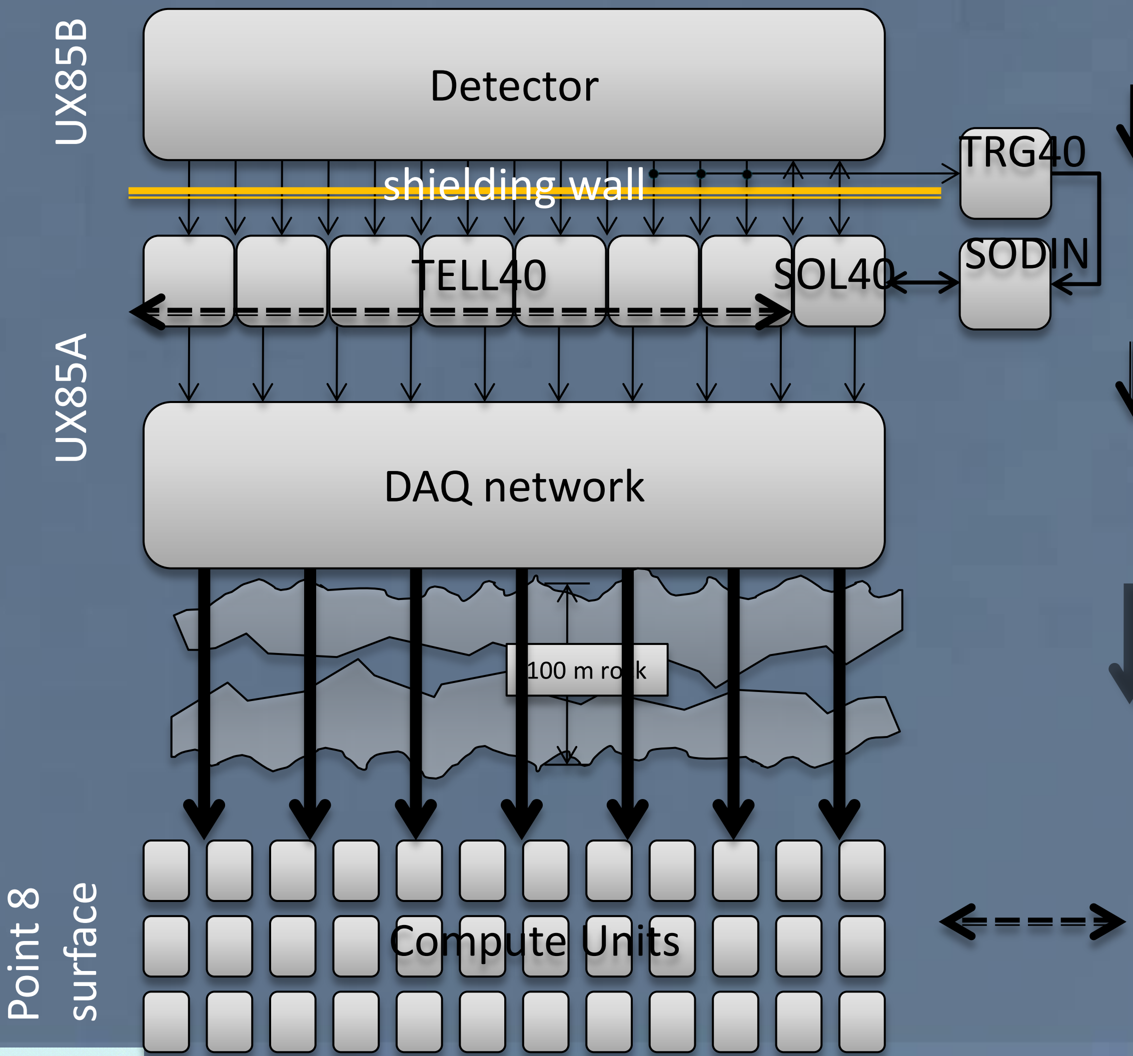
Readout System Review

Niko Neufeld

Electronics Upgrade Meeting, April 10th 2014

FTDR - Architecture



GBT: 4.8 Gbit/s (9000 x) DAQ
3.2 Gbit/s (2400 x)
TFC/ECS

Input into DAQ network
(10/40 Gigabit Ethernet
(1000 to 4000)

Output from DAQ network
into compute unit clusters
(100 Gbit Ethernet / EDR IB)
(400 links)

Bi-directional links for TFC/ECS to/from TELL40 for timing, back-pressure and LLT

Cost-optimised readout system

- Bring readout-boards, network and farm together in one location (containerized data-centre on SX)
  - run Versatile Links from detector to surface (~ 300 m)
- Move from AMC40 to PCIe40
  - Removes optical link from TELL40 to network
  - Free (and late) choice of network technologies
  - PC separates off event-builder network technology
  - Number of core-network ports can be halved

Readout System Review

- Held on Feb 25th, 9:00 – 16:00
- 4 reviewers
  - Christoph Schwick (PH/CMD – CMS DAQ)
  - Stefan Haas (PH/ESE)
  - Guido Haefeli (EPFL)
  - Jan Troska (PH/ESE)
- https://indico.cern.ch/event/297003/
- All documents can be found in EDMS https://edms.cern.ch/document/1357418/1

Review

- Organised by Renaud, Ken and myself
- Many thanks to the reviewers for stimulating, interesting discussions, their patience to read through a mountain of material and the timely preparation of the report
- Many thanks to the presenters – the feedback I got on the quality of the presentations was overwhelmingly positive

Questions to the reviewers

- Long-distance optical fibres:
  - Have we fully demonstrated the feasibility?
  - Are further tests recommended?
  - Are there further risks to be assessed?
- For both ATCA/AMC and PCIe:
  - Does the proposal comply with the global architecture?
  - Have we fully demonstrated the feasibility?
  - Are further tests recommended?
  - Are there further risks to be assessed?
  - Are the technology choices appropriate? (e.g. cost-effective, performance, obsolescence, long-term availability)
  - Can the system handle the target luminosity of 2x10^33?
  - Are the proposals scalable? (e.g. for future increases in luminosity)
- We propose a change of baseline from ATCA/AMC to PCIe. Do the reviewers support the choice of baseline (PCIe) and back-up (ATCA)?
- Comment on manpower/schedule
- Is documentation in good shape?

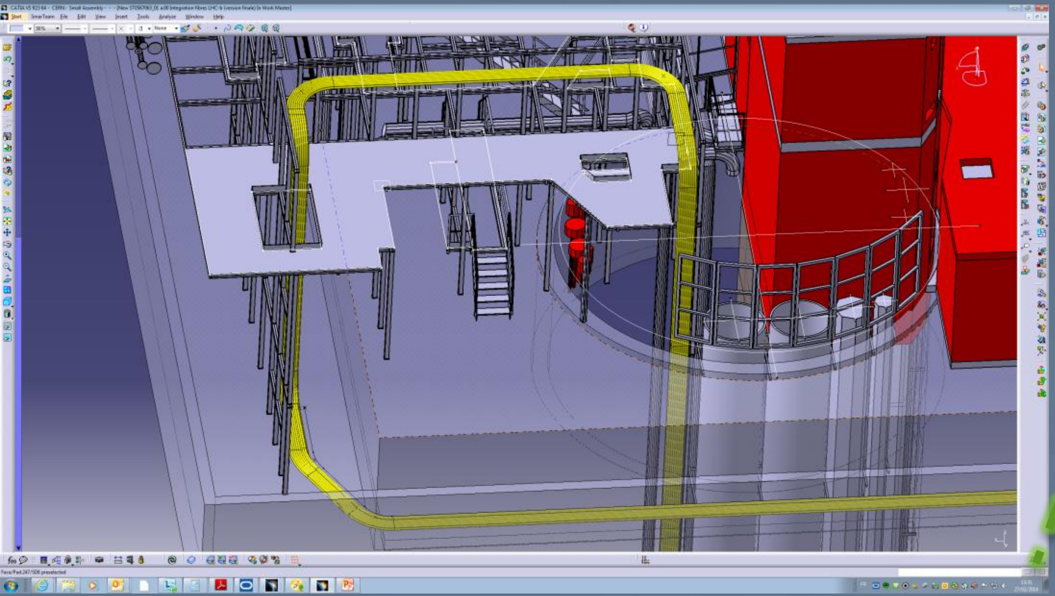
Optical links 1

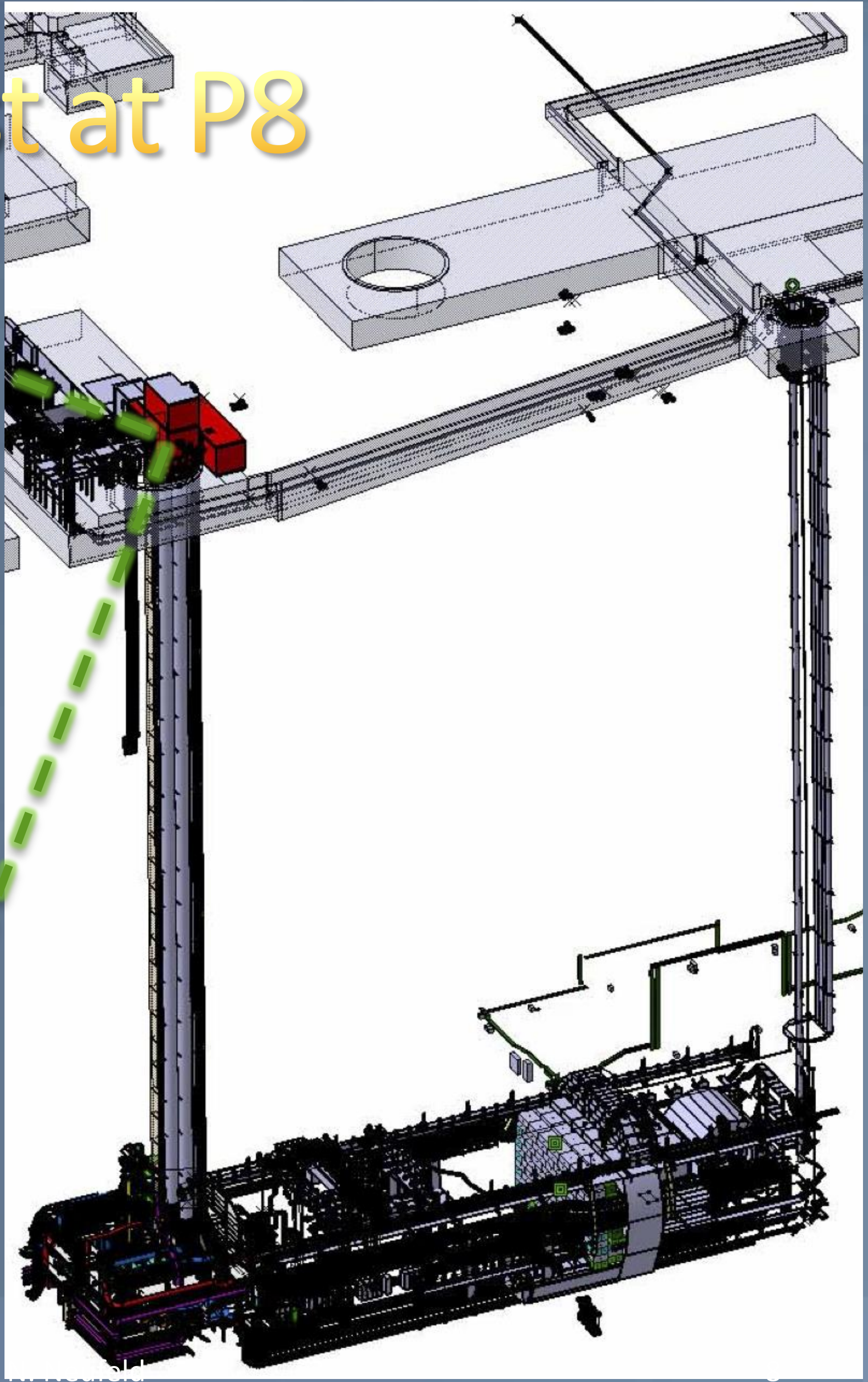
- OM3 vs OM4: *“[…] the review committee thinks that the choice of OM4 fibers would be preferable since it would allow future upgrades to higher bit rates in case this will become necessary. The final decision might depend on the cost of these fibers once they need to be purchased.”*

  → note that no intrinsic advantage of OM4 for current application is identified. Price difference for material currently 50% (1 → 1.5 MCHF)
- In situ tests: *“[…] recommend that the testing be repeated on a test span of both OM3 and OM4 fibre that are installed in the LHCb experimental area via the proposed fibre route. This would allow unforeseen effects to be observed.”*

  → see next page
- *“The radiation environment leaves a lot of margin for the majority of the optical links since the worst radiation level (80 Gy/10^12 neq/cm^2 for the tracker) is well below the Versatile Link Calorimeter grade qualification level.”*

Optical fibres 2 - test at P8


CATIA V5 R23 64 - CERN : Small Assembly - - - [New ST0567063_01 a.00 Integration Fibres LHC-b (version finale) (In Work Master)]
Start
SmarTeam
File
Edit
View
Insert
Tools
Analyze
Window
Help
30%
None



SD8 – try to get rid of 30 m loop

Thanks a lot to Laurent for info and pictures!

Test installation - planning

- Choose "worst" (== longest distance, most bends) point in cavern (in the bunker)
  - We will need some space!
- Planning:
  - Before June 2014 (cool-down start): install support structures in PM85 et US85
  - Before December 2014 (closure of cavern): install supports in surface gallery and UX85
  - Install prototypes of both solutions (blown fibres (from bottom to top and pre-connectorized cables (OM3 and OM4)
- Long-term tests starting from 2015 (with beam)

Optical fibres - 3

- We want to do more sampling of the Minipod transmitters and receivers
- Test-stand setup at CERN (by Rainer and Paolo)→ plan to test all AMC40 of the new batch before being handed out to SD (tests should not take more than 2 weeks total)
- Goal: get more information about inter-channel and inter-component variability

TELL40 implementation

- Basic question to reviewers: “We propose a change of baseline from ATCA/AMC to PCIe. Do the reviewers support the choice of baseline (PCIe) and back-up (ATCA)?”
- Basic answer:
  “Given the listed advantages the review committee endorses the choice of the PCIe based design as a baseline.”

However…

- *“a prototype should be produced and evaluated as soon as possible”*
    - → *see Jean-Pierre’s presentation*
- paraphrasing a bit: *care has to be taken in the PC environment: thermal stability, quality of power-supply and noise, especially when used for TFC*
- *“The PCIe demonstrator must show that the implementation of the optical components […] is mechanically feasible without restricting the options on [the] PC […]”*
  → we are in contact with PC vendors for all these issues. There is a choice of compatible 1U and 2U chassis
    - *GPU-cards have worse requirements than we do (size, power)*

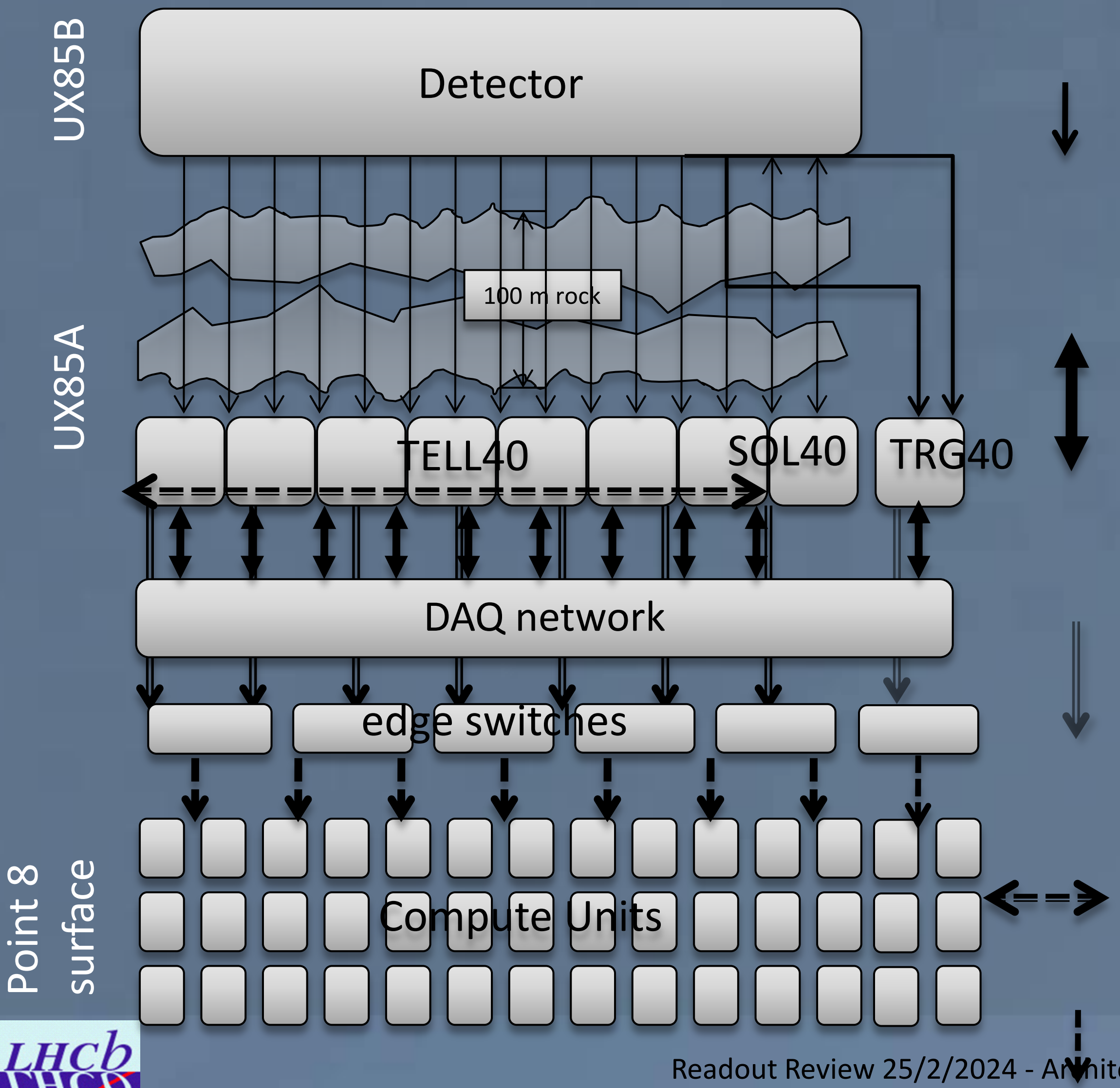
However 2)

- *"It is recommended that, even if it is considered unlikely by the experts, the collaboration addresses the problem which might occur if the PCI express bus (v3.0) in PCs becomes unavailable during the lifetime of the experiment"* →assume life-time == end of run 4. The "experts" are sure there will be no problem until end of run 3. Then if PCIe is replaced by something not backward compatible
  - will use PCs with a "legacy" slot (require only 1 per PC)
  - stock PCs and/or components (mainboards, memory modules, fans, power-supplies)
- If life-time > run 4, and no more PCIe, then redesign the card (FPGA cost will have come down enormously, optical components are pluggable and can be re-used or new ones will be qualified)

LLT

- *“In the reviewers' opinion, only the software option should be pursued if any. For this case it should be clarified in how far the resources in the FPGA of the PCIe board could be used to support the Software LLT with some pre-processing”*
  → will establish the feasibility of the software LLT in time so that hardware LLT could be still developed in case needed.

Evolution 2 - Architecture



GBT: 4.8 Gbit/s (9000 x) DAQ
3.2 Gbit/s (2400 x)
TFC/ECS running over 300 m

DAQ network >= 100 Gbit/s Ethernet / EDR InfiniBand / …/400 links)

AMC40 → PCIe40 + PC

Output from event-builder into edge switches (100 Gbit Ethernet / EDR IB) (<= 400 links)

Bi-directional links for TFC/ECS to/from TELL40 for timing, back-pressure and LLT

10G BaseT to farm-nodes

Features of upgraded LHCb DAQ 2)

- Push-protocol with centralized flow-control (throttle)
  - back-pressure from PC/DMA will activate throttle in PCIe40
- Event-building protocol
  - Lots of buffering in event PCs: use zero-copy (remote DMA)
  - 100 Gbit/s link matches input bandwidth → full-event-building for every bunch-crossing
  - Event-builder PCs act as gateways, network technology to farm-nodes can be different from event-building network
- LLT merged into event-building (“LLT as a co-processor”)
- Collateral benefits: more than 3x less FPGA resource usage for PCIe compared to 10 Gigabit Ethernet

# links 100 Gbit/s (from event-builder PCs)	400 - 500	# links 100G Ethernet (from event-builder PCs to edge switches)	400 - 500
Event-size (total – zero-suppressed)	~ 100 kB	# edge switches	~ 100
Event-building rate	40 MHz	# core-switches	2
# read-out boards	400 - 500	# farm-nodes	1000 (up to 4000)
output bw / read-out board	up to 100 Gbit/s (useable)	max. input bw / farm-node	10 – 40 Gbit/s

Conclusion

- *The review committee endorses the plan of the collaboration to continue with the PCIe based readout card as a baseline solution for the new Readout System. The committee notes that there are still risks involved and that a prototype should be produced as soon as possible in order to assess the critical technical issues mentioned above. It is further recommended to work out a resource loaded overall schedule of the project […]*
- *The tests performed on the long distance optical link are encouraging but need to be completed with further tests under realistic conditions.*

Conclusion

- The preparation of the review and the review itself have been very useful and productive
- We will now go on to create the PCIe40 prototype to address all issues raised
- We will do a extensive tests of the fibres in the final conditions
- We are ready to start working on the Trigger and Online TDR