Event builder upgrades in LS1

Andrea Petrucci - CERN (PH/CMD)

ALICE, ATLAS, CMS & LHCb joint workshop on *DAQ@LHC*
12-14 March 2013, Château de Bossey, Switzerland

Acknowledgments

- Roberto Divia` (Alice)
- Niko Neufeld (LHCb)
- Wainer Vandelli (Atlas)

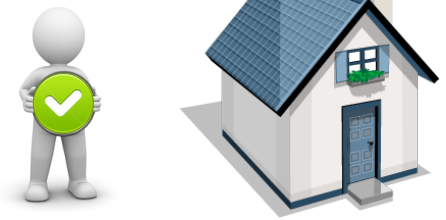
Outline

- Overview of event builders after LS1
- Event builders after LS1
  - LHCb
  - Alice and Atlas
  - CMS

Overview of event builders after LS1

Event builders after LS1 in a nutshell

LHCb



Ready

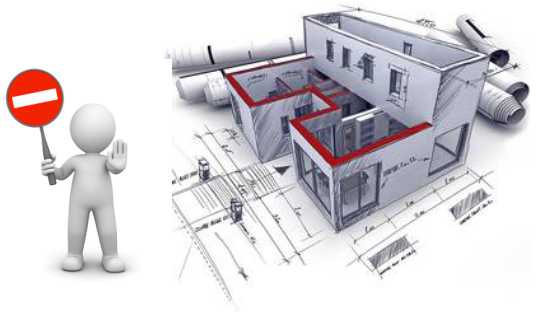
- For 2015 physics data

ALICE ATLAS EXPERIMENT

Maintenance

- Networks replacement (10 Gigabits Ethernet)

CMS



Re-factoring

- Use Infiniband and 10/40 GE for EB networks

LHCb: event builder upgrades in LS1

- No changes: ready for physics data

LHCb: Event builder upgrades in LS1

No changes in the event builder

- Keep the same PCs and networks
- Keep the same DAQ software

New requirements for HLT

- Extend the existing HLT farm due to the increase of luminosity
- Deferred event filtering see Markus Frank's talk

  "HLT infrastructure evolution in LS1"

Alice and Atlas: event builders upgrades in LS1

- Maintenance: networks replacement
  (10 Gigabits Ethernet)

Requirements for the event building after LS1

ALICE

- Two new detectors (CPV and DCal) no sensible impact on event rates and data rates
- Increase the readout bandwidth in the TRD detector
- Upgrade of the TPC (under evaluation) considerable impact on event rates and data rates

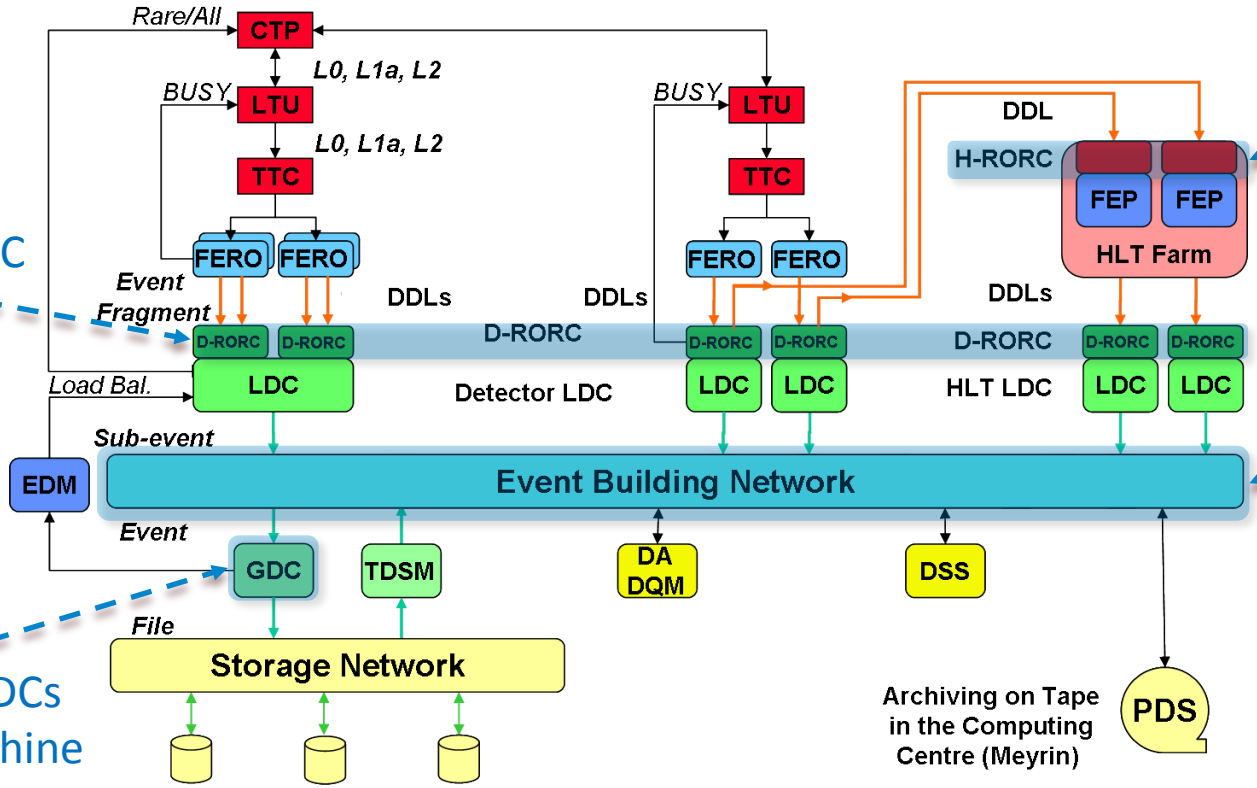
ATLAS

- L1 rate goes to 100 kHz (subject to various changes at the detector level)
- Depending on the LHC scenarios, in particular 50ns operation, peak event size can be >2MB
- Event building rate will be significantly higher
  - The trigger strategies for the new architecture are still being considered
  - The new ROS system and the new network will allow for more readout

Alice event builder changes (I)

- New D-RORC card to increase the readout bandwidth, see Filippo Costa's talk ("Future of the DAQ readout links")
- Planning to increase bandwidth between DAQ and HLT farm (new H-RORC) due to the increment of the readout bandwidth in the TRD (between 5 to 8 %) and other future sub-detectors upgrade
- Network replacement
  - EVB network replace Force10 with Brocade/HP (CERN IT contract)
  - Change network topology from point to point to tree. The tree topology allows to maximize Price/Performance
- Replacement cycle of PCs (new PC more powerful than current)
- Run more then one GDC per physical machine through virtualization (under evaluation)
  - Due to new machines trying to maximize the hardware resources usage in the GDCs with virtualization and the goal is to reduce the number of GDCs of factor between 4 and 10

Alice event building changes (II)



New H-RORC

New D-RORC

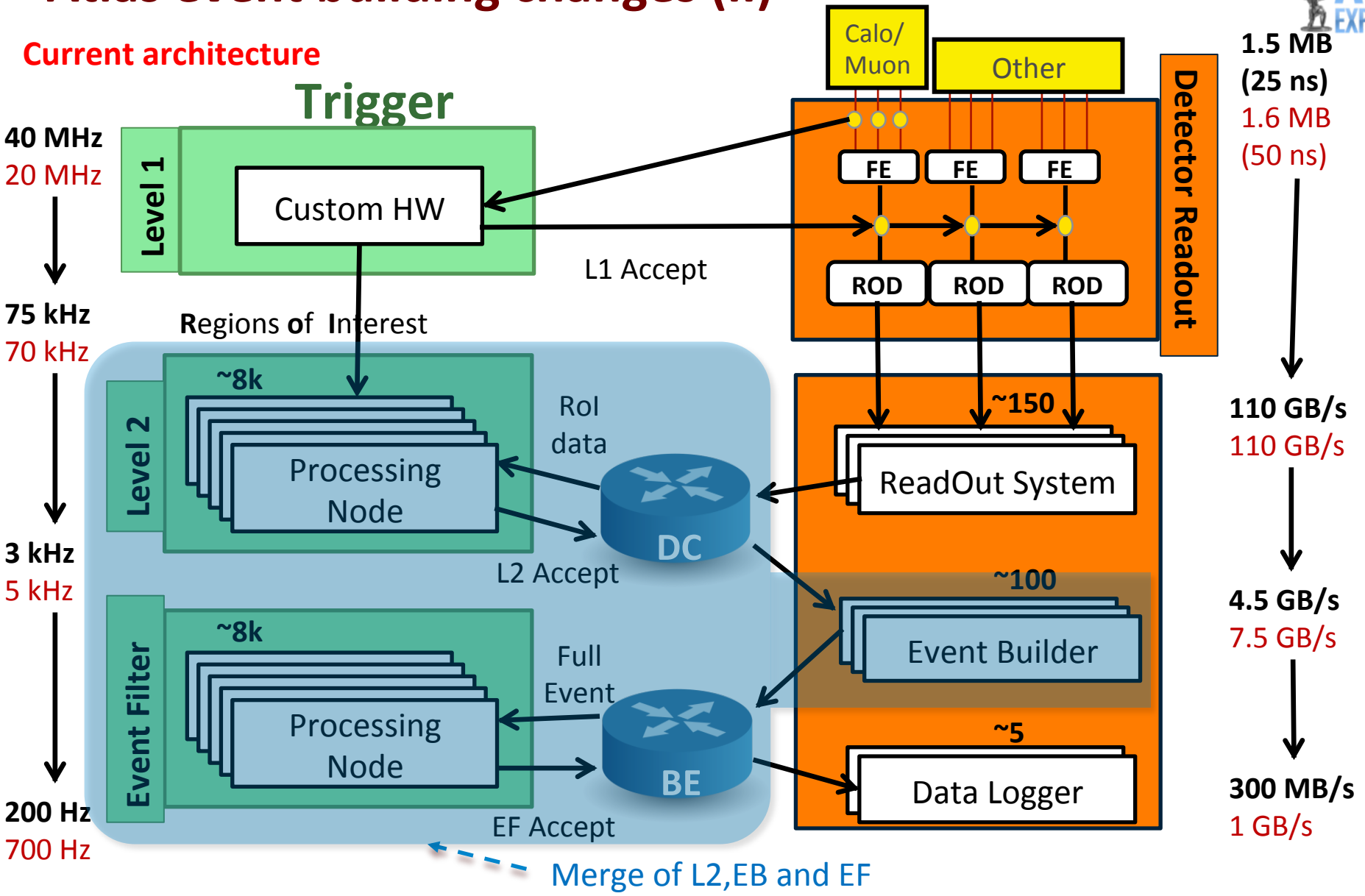
Replace EB network

Run more GDCs in same machine (virtualization)

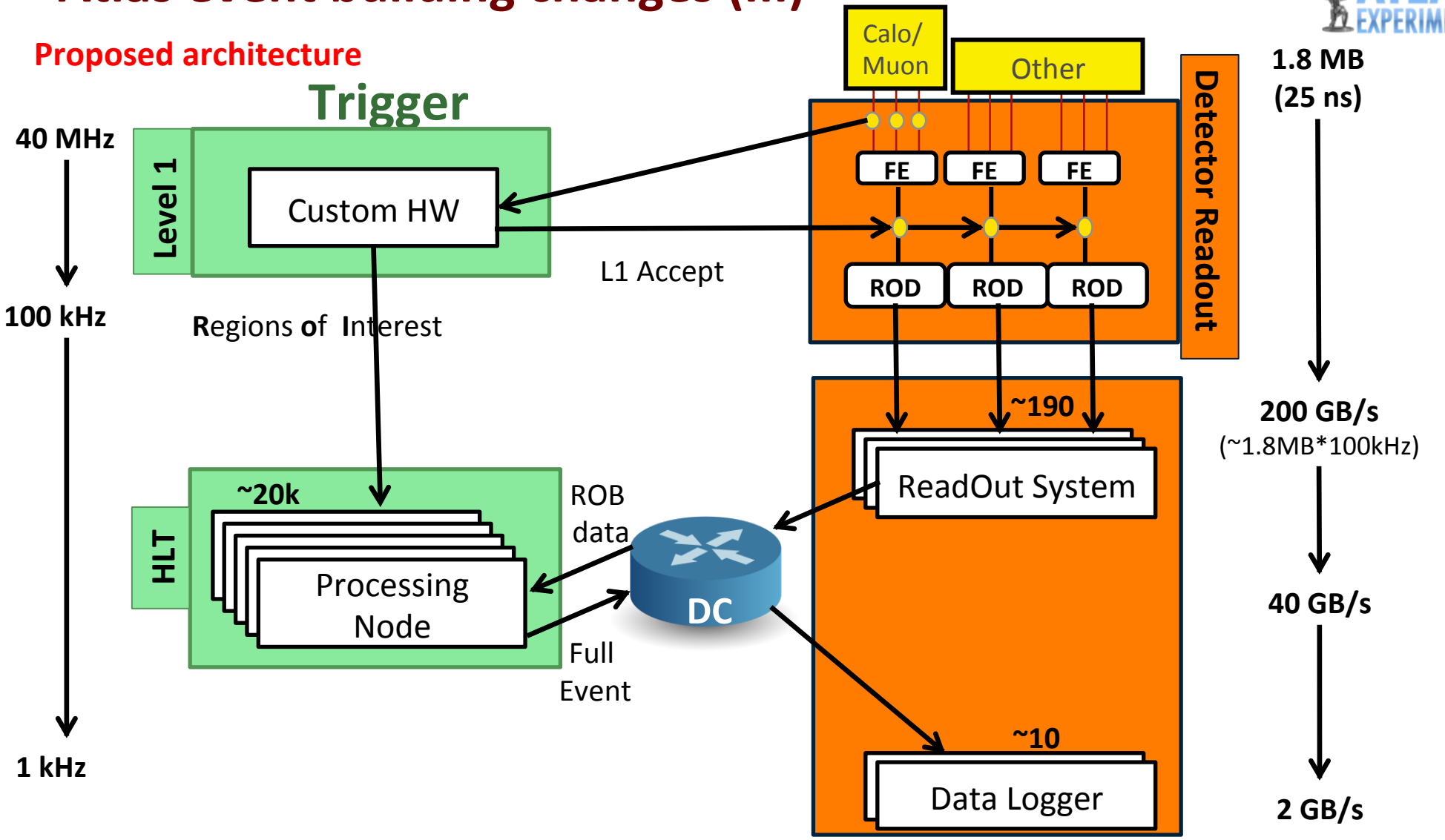
Atlas event building changes (I)

- Merge L2, EB and EF within a single system
  - L2 (processing & data collection based on ROIs), EB and EF (processing on the full event) functionalities in each HLT node
  - Automatic system to balance the CPU load (additional flexibility for HLT strategies)
  - Less connections to the ROS PCs
  - Implicit handling of HLT farm heterogeneity
- The event builder as a separate and well defined entity is disappearing
  - Incremental event building will be performance on each HLT node under the drive of the trigger code and/or based on data-flow optimization algorithms
- Changes in the DAQ software
  - Less DAQ applications
  - Look for common solutions & avoid code duplication
  - A single framework for HLTSV, DCM, HLTPU and SFO?
  - Developing a new message passing library based on asynchrouos I/O
  - Ease maintenance and scalability

Atlas event building changes (II)



Atlas event building changes (III)



CMS: event builder upgrades in LS1

- Re-factoring: use Infiniband and 10/40 GE for EB networks

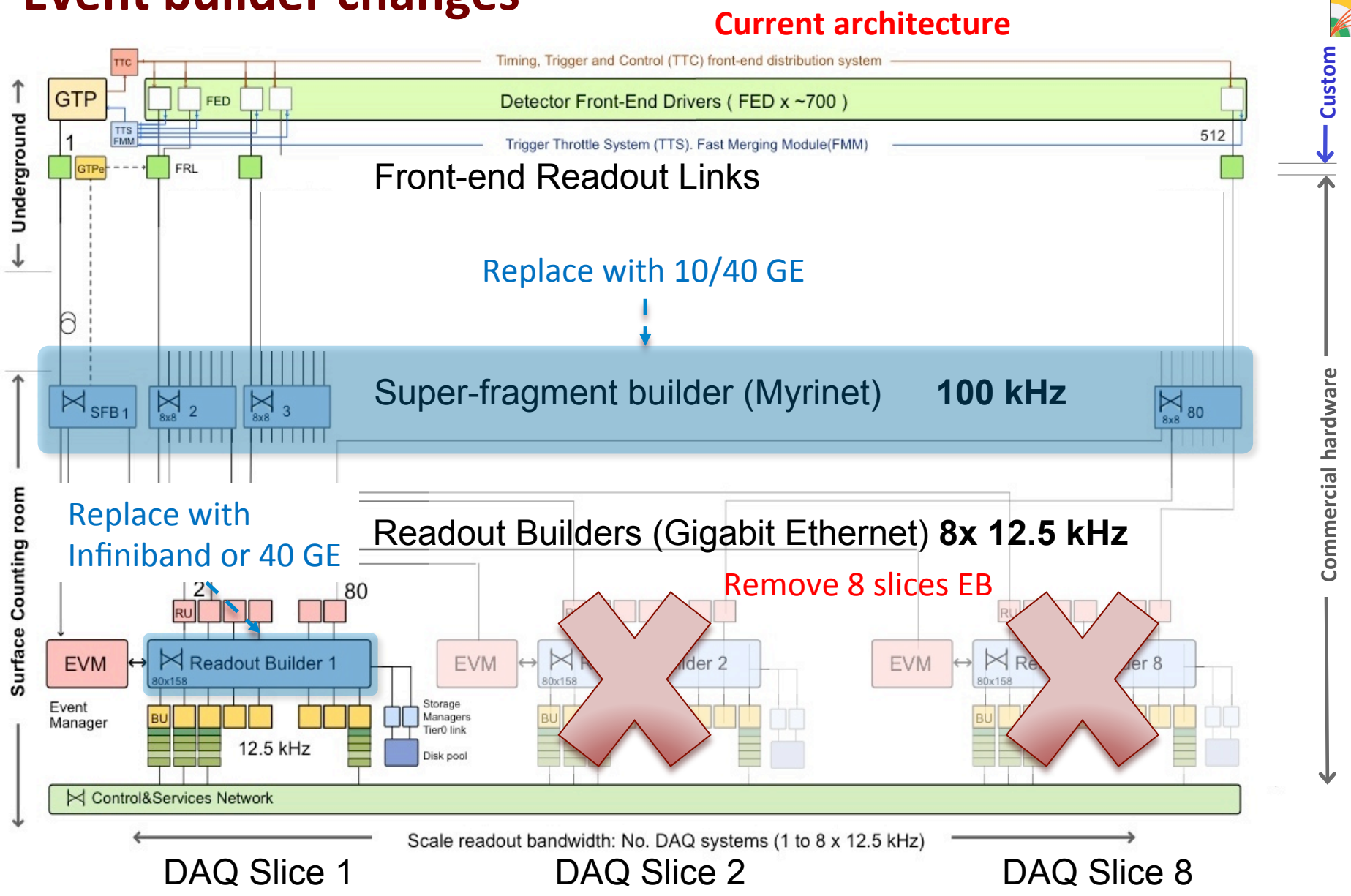
Requirements for the event building after LS1

- Accommodate sub-detectors
  - Some sub-detector new back-end electronics in µTCA standard with serial link to cDAQ
  - Some sub-detectors will be replaced which lead to higher data volumes
- Inputs (custom electronics)
  - About 500 2-4 Gbps “Legacy” FEDs
  - About 20-100 6-10 Gbps New FEDs
- Aging of existing hardware (PCs and Networks at least 5 years old)

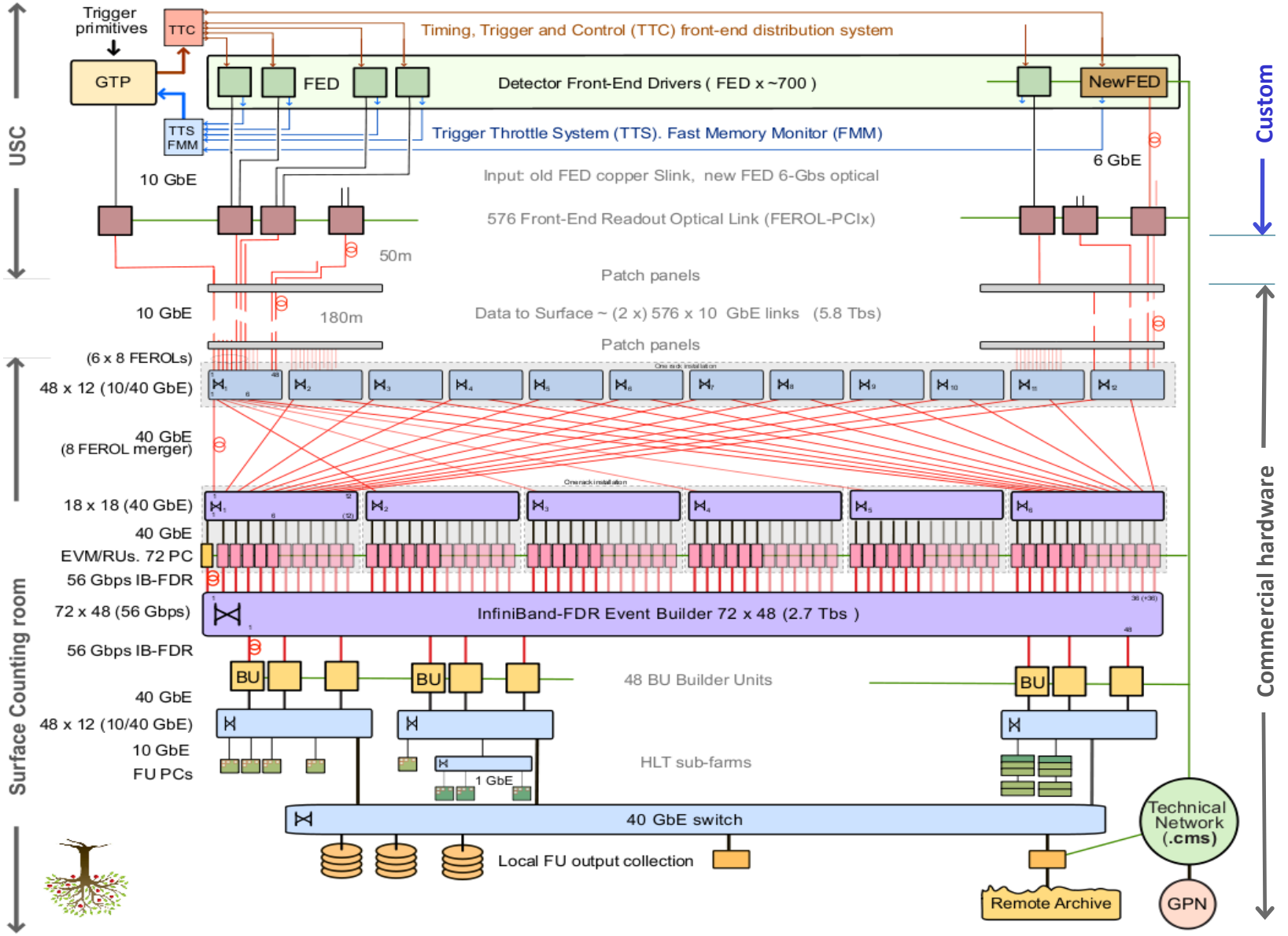
Event builder changes

- Re-implementation with up-to-date technology
  - Typically 10x less nodes with 10x more performance
- Considering 10/40 GE technologies to replace myrinet-based fedbuilder (2x 2 Gbps)
- Taking into account the 40 GE or Infiniband to change the event builder network (3x1GE)
- Single event builder (no need for 8 slices)
- New architecture between Event Builder and Filter Farm
  - see Markus Frank's talk "HLT infrastructure evolution in LS1"

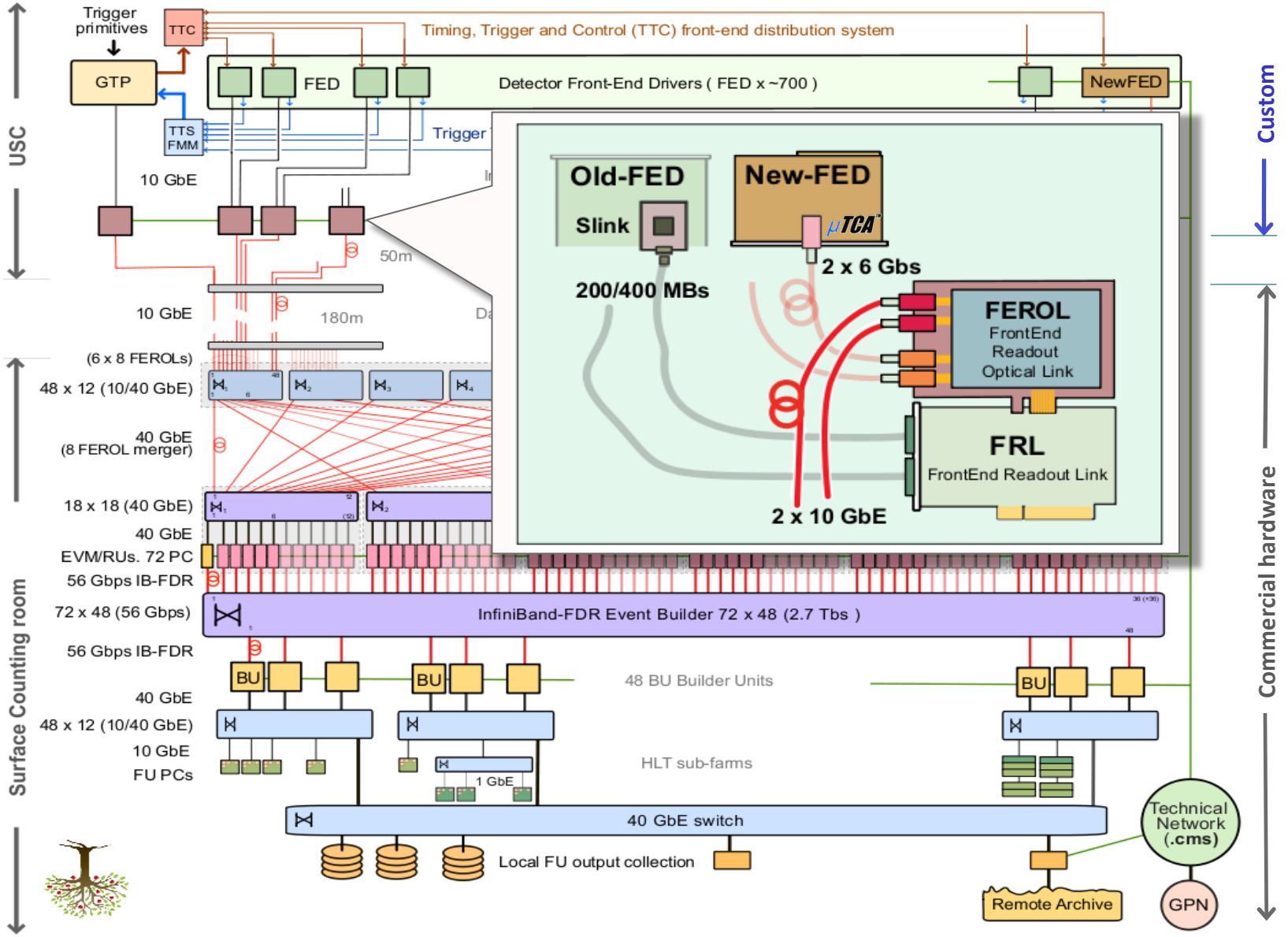
Event builder changes



Upgrade of CMS DAQ system (I)



Upgrade of CMS DAQ system (II)



FEROL aggregation into one RU

Aggregation n-to-1, for example

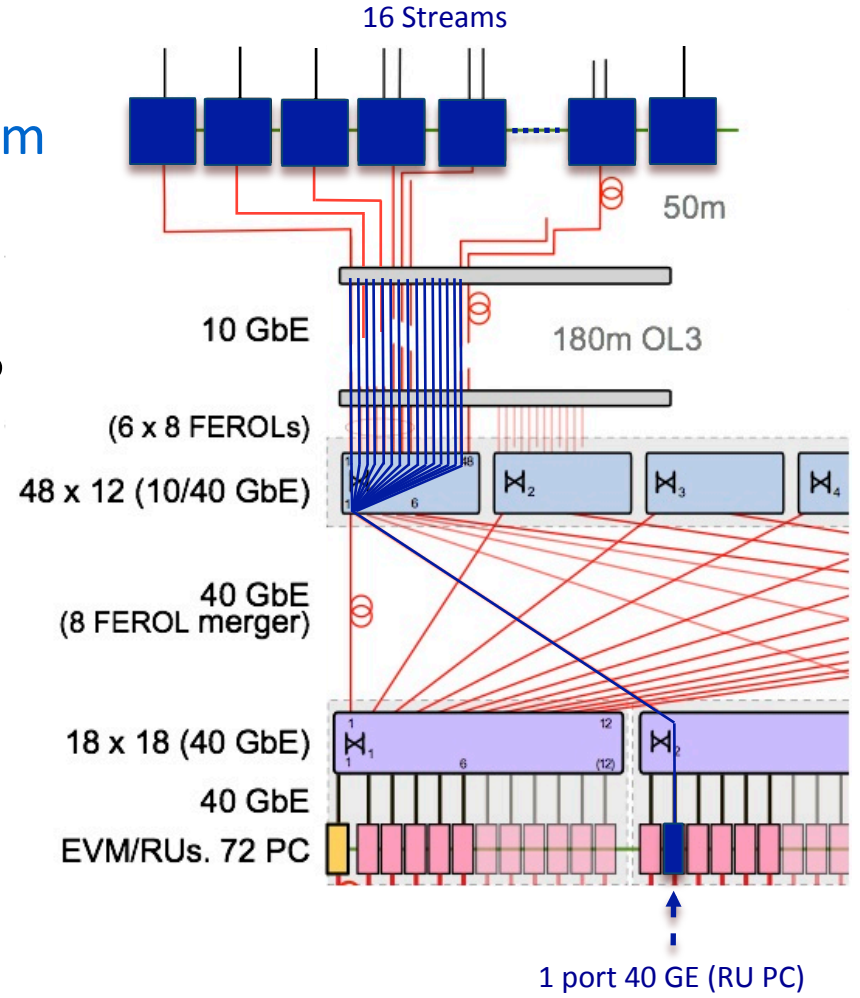
- 16 FEROLs each sending 2 Gbps TCP/IP stream over 10 GE link
- Concentrated in one 40 GE NIC into RU PC
- Reliability and congestion handled by TCP/IP

  Filippo Costa's talk "Future of the DAQ readout"

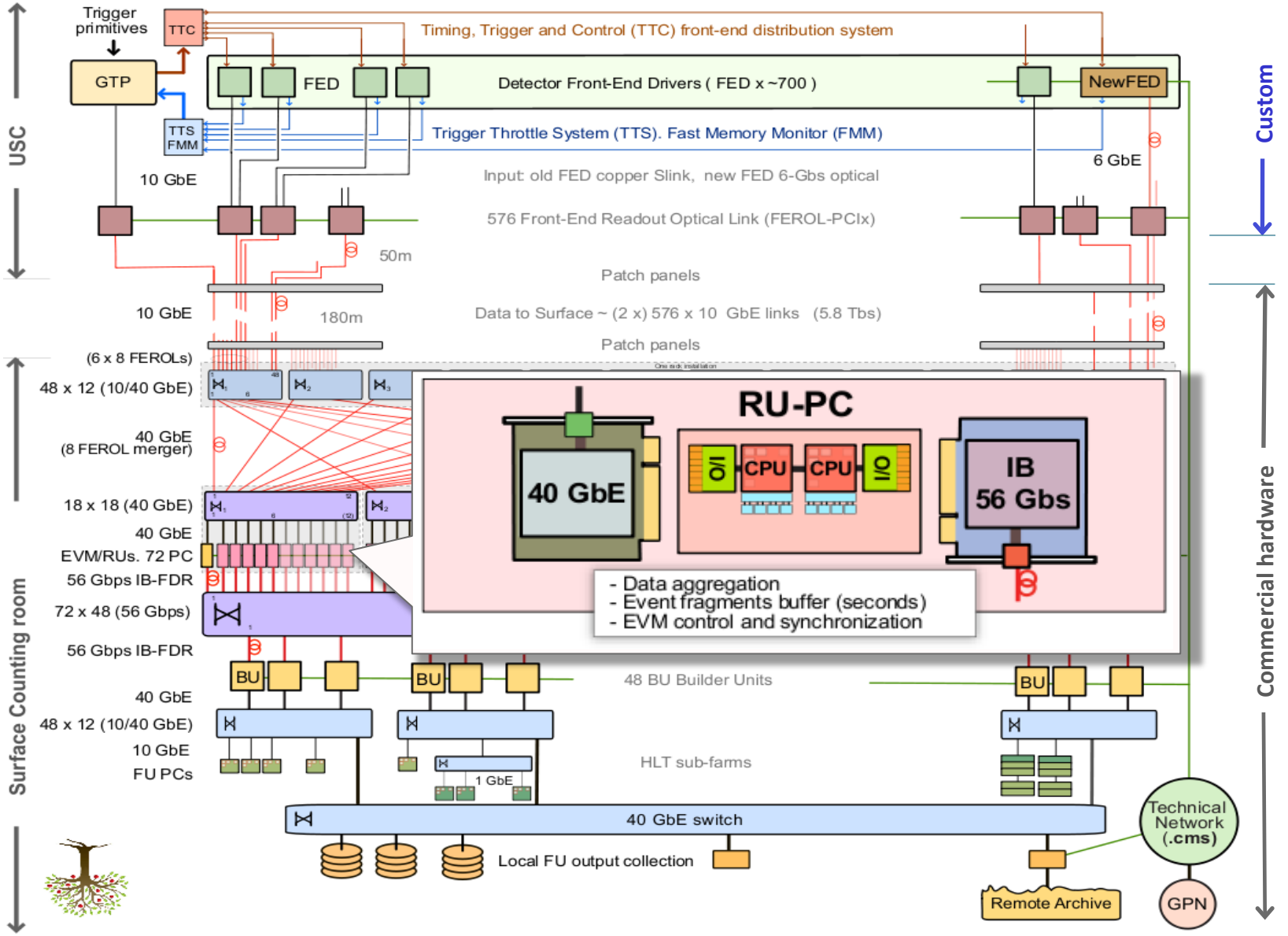
USC – SCX 180m

- with OM3 fibers up to 200 m
- 40 GE max. is 150 m - NOT feasible

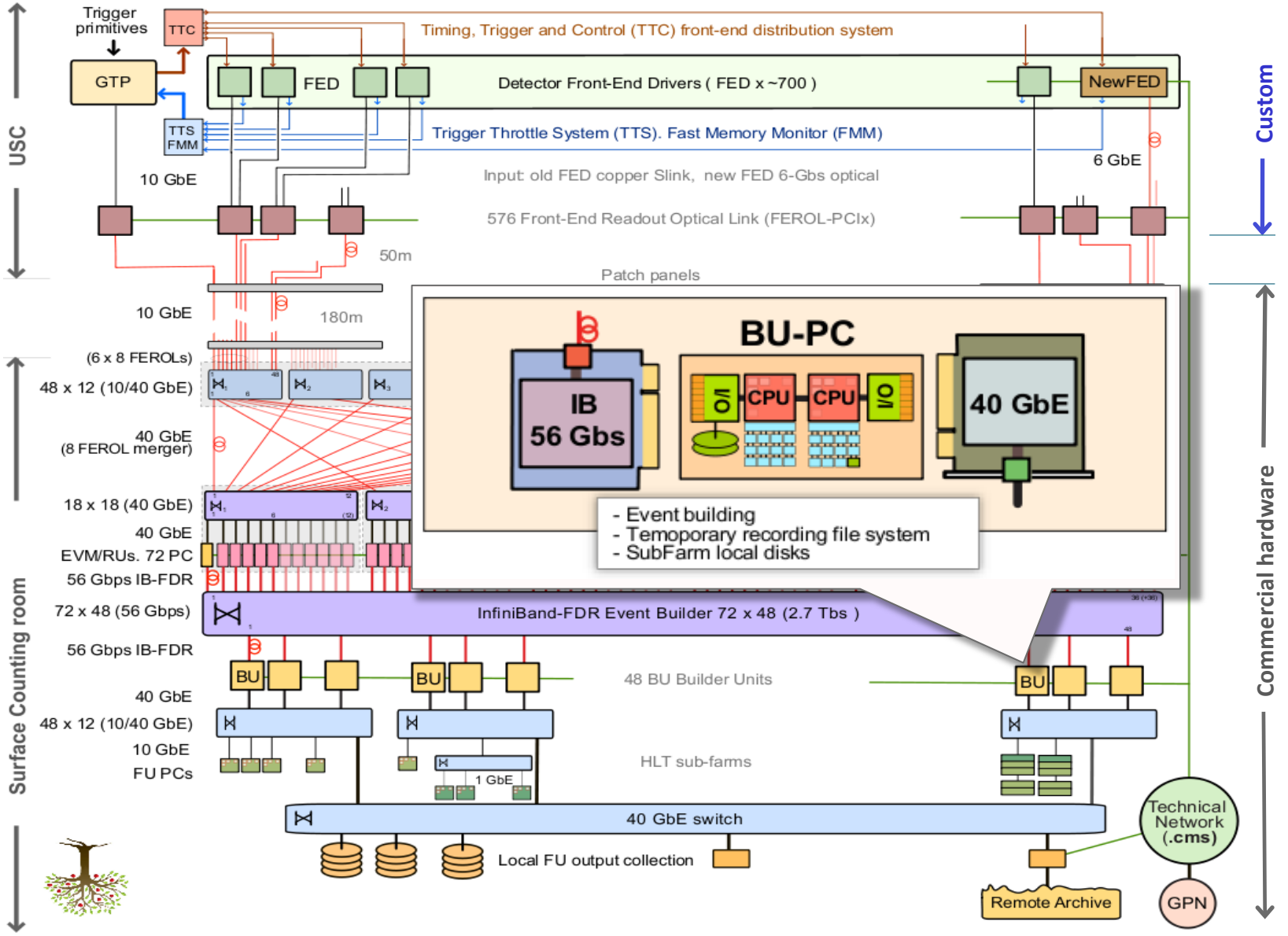
Network useful to re-configure when fault with optic, PC, etc



Upgrade of CMS DAQ system (III)



Upgrade of CMS DAQ system (IV)



Feasibility studies

- Networking technologies
- Event builder software
- Measurements

Networking technologies

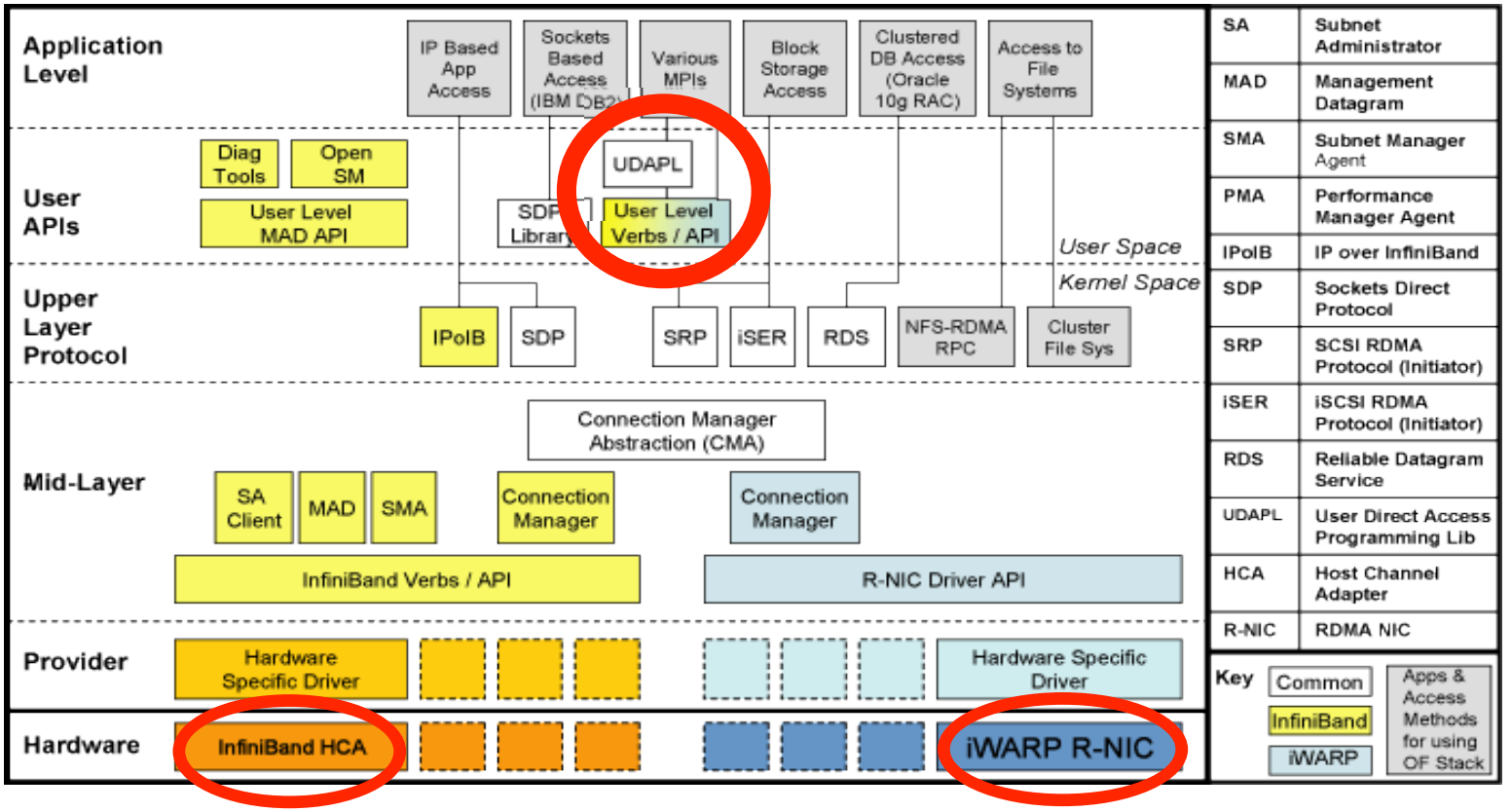
Our feasibility studies are focused in two network technologies

- **Ethernet**
  - 10/40 Gigabit Ethernet (different vendors)
  - iWARP (RDMA) – TCP/IP full offload (Chelsio T4 Unified Wire Adapters)
  - performance measurements using **TCP/IP** and **DAPL** (Direct Access Programming Library- OpenFabrics)
- **Infiniband**
  - 4x quad data rate (QDR)
    - 40 Gb/s - 8B/10B encoding -32 Gb/s data rate
  - 4x fourteen data rate (FDR)
    - 56 Gb/s – 64B/66B encoding – 54.54 Gb/s data rate
  - performance measurements using **DAPL** (Direct Access Programming Library- OpenFabrics) and IPoIB (IP over InfiniBand)

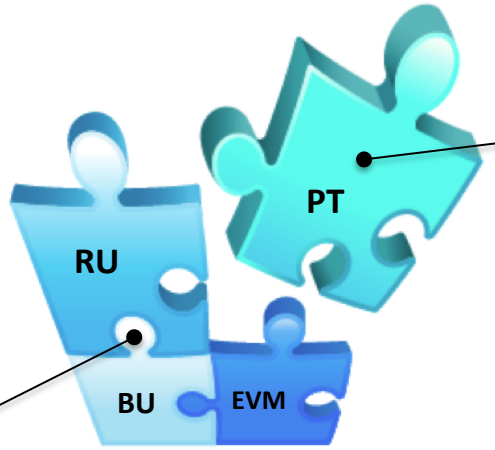
The OFED Stack (source: OpenFabrics Alliance)

A unified, cross-platform, transport-independent software stack for RDMA and kernel bypass

- http://www.openfabrics.org/



Event builder software



The pt*u*DAPL is a pluggable component to transparently access the DAT library in XDAQ – CMS online framework

Re-use and adapt same Event Builder software

Test setups

Setup 1 (LHCb)

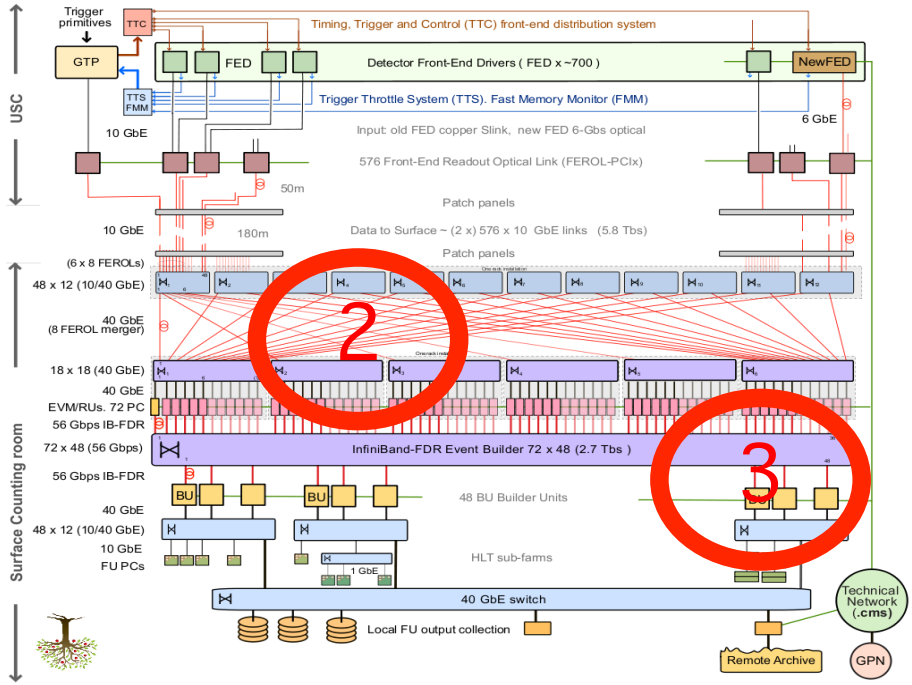
- Initial software development environment (ptuDAPL)
- first tests with Infiniband (QDR) and 10 GE (TCP, iWARP)

Setup 2 (FEROL test)

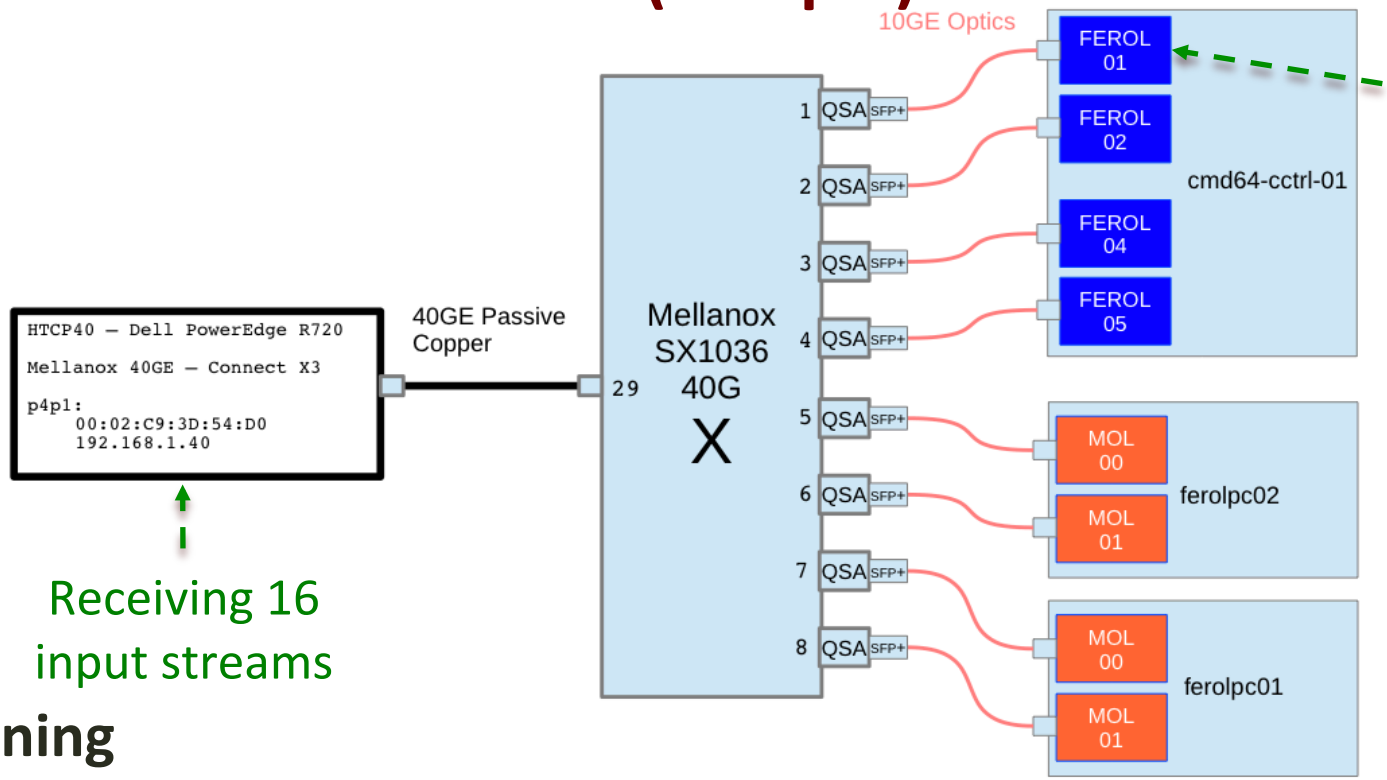
- Front-End Readout Optical link merger
- 16 Streams inputs to 1 x 40 GE

Setup 3 (Event builder)

- DAQ Event building
- Scalability
- N inputs to M outputs (IB or 40 GE)



FEROL Measurements (setup 2)



Tuning

- Cope with NUMA architecture, CPU and IRQ affinities in 10/40 GE

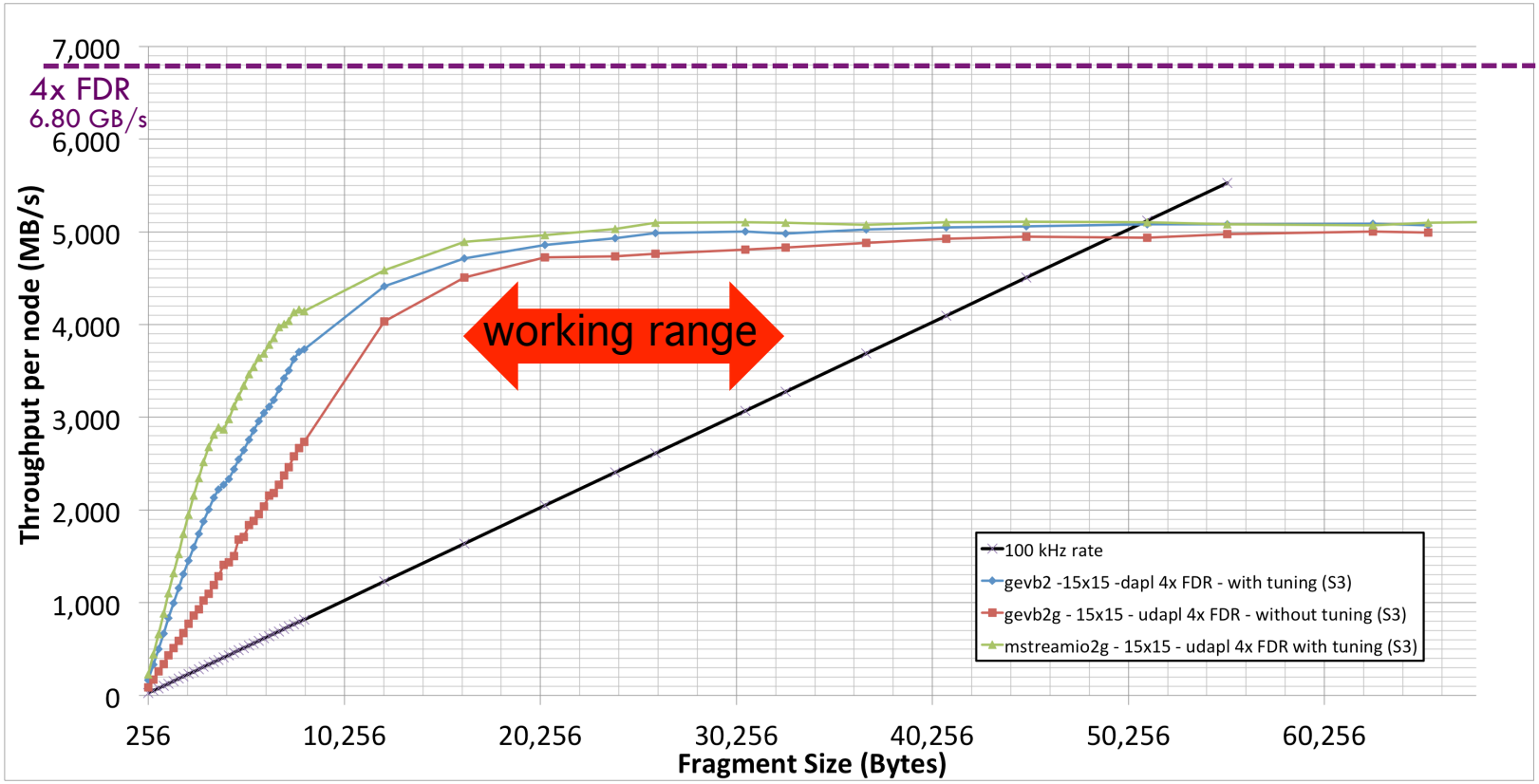
Performance

- 150 kHz with fragment size of 2kB (40 GE)
- 165 kHz with fragment size of 1kB (22 GE)

Stability

- No backpressure for 125 hours at 100 kHz with 2kB fragment size

Infiniband Measurements (setup 3)

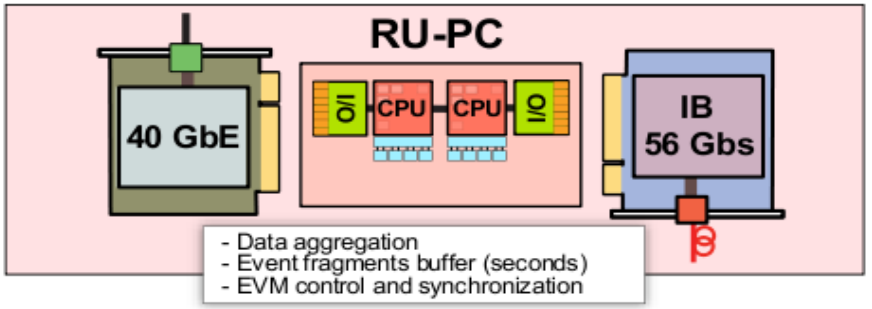


Fragment size - Bytes	100 kHz MB/s	mstreamio2g	gevb2g	gevb2g no tuning
16384	1638	4894	4715	4510
32768	3277	5098	4983	4829

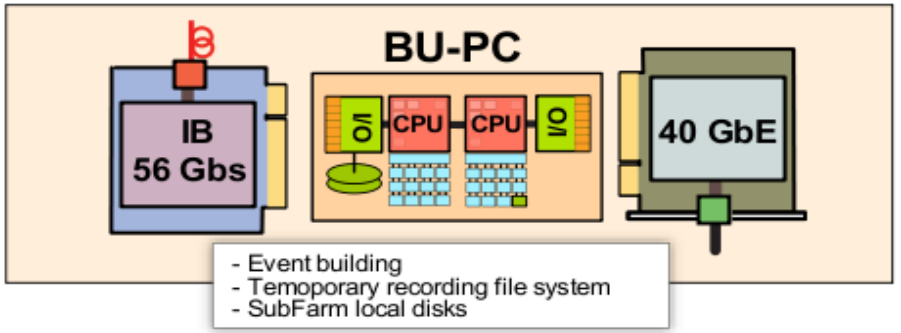
Detailed presentation about Experience with IB and Ethernet with Mellanox:
https://indico.cern.ch/getFile.py/access?contribId=14&sessionId=2&resId=0&materialId=slides&confId=218156

Pending Issues

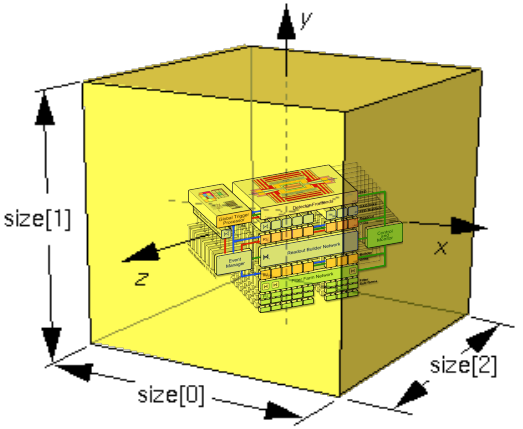
- Simultaneous input/output on RU



- Simultaneous input/output on BU



- Scaling of EVB from 15x15 to 72x48



Summary

Ready

- For 2015 physics data

Maintenance

- Networks replacement (10 Gigabits Ethernet)

Re-factoring

- Use Infiniband and 10/40 GE for EB networks

Thank you for your attention

- Are there any questions?

Backup slides

Setup 1 (LHCb)

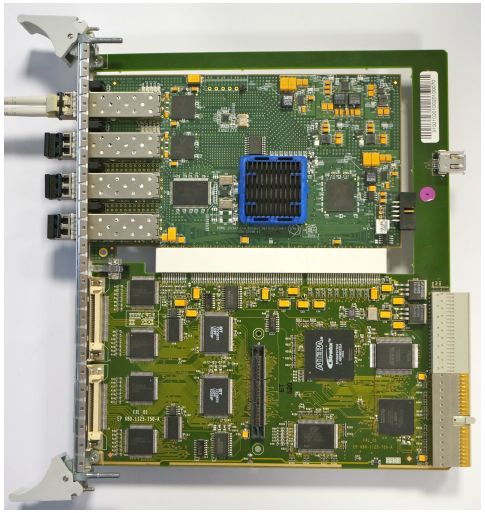
	Setup 1	
Nodes	8	
Type	DELL R710	
CPU	Xeon E5530 2x 4-core at 2.27 GHz	
Memory	3 GB	
Network	Ethernet	Infiniband
Adapter	Chelsio T420-CR 10GBASE-SFP RNIC (iWarp)	Qlogic HCA, qle7340 4x QDR PCIe
Switch	Voltaire Vantage 6048, 48 ports, 10 GbE	Qlogic 12300- BS01, 36 ports, 4x QDR

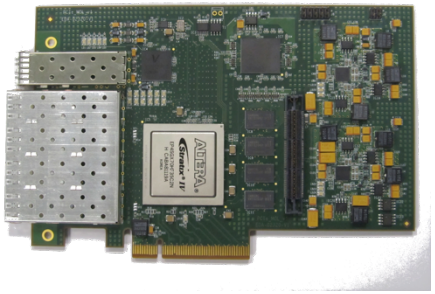
DELL R310/R620

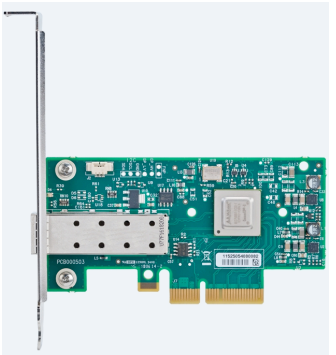
- Operating System: Scientific Linux CERN SLC release 5.3 (Boron)
- Linux version: 2.6.18-164.6.1.el5
- OFED version: OFED.1.5.2.x.x
- XDAQ version: release 11

Setup 2 (Hardware)

	Setup 2		
Nodes	4	4	1
Type	FEROLs	MOLs	DELL R620
CPU	-	-	Xeon E5-2670 2x 8-core at 2.6 GHz
Memory	-	-	32 GB
Network	10 GE		40 GE
Adapters	-		Mellanox - ConnectX-3 VPI MCX353A-FCBT
Switches	Mellanox 36 - QSFP40 GbE - MSX1036B-1SFR		







Setup 2 (Firmware/Software)

DELL R620

- Operating System: Scientific Linux CERN SLC release 6.2 beta (Carbon)
- Linux version: 2.6.32-220.2.1.el6.x86_64
- OFED version: OFED.1.5.3.3.1.0
- Ethernet driver: mlx4_en version 1.5.8.3
- XDAQ version: release 11
- TCP test: sock application http://www.icir.org/christian/sock.html

Mellanox - ConnectX-3 VPI

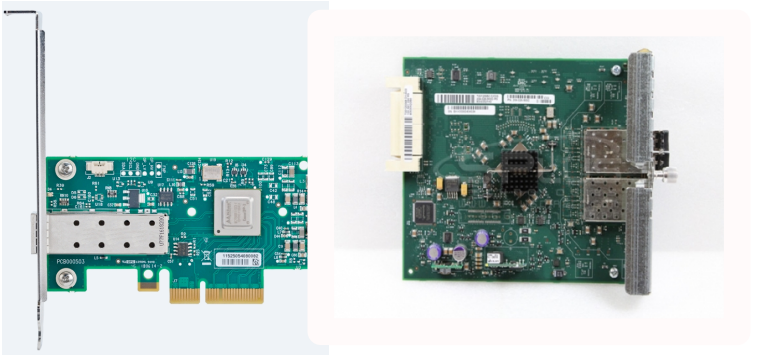
- Firmware version: 2.11.500

Mellanox 36 – MSX1036B-1SFR

- Firmware version: 9.1.6294
- Mellanox MLNX-OS™ version: 3.2.0506

Setup 3 (Hardware)

	Setup 3	
Nodes	32	
Type	DELL C6220	
CPU	Xeon E5-2670 2x 8-core at 2.6 GHz	
Memory	32 GB	
Network	IB FDR 4x/40 GE	
Adapters	Mellanox - ConnectX-3 VPI MCX353A-FCBT (# 4)	DELL mezzanine Mellanox FDR CX3 (# 24)
Switches	Mellanox 36 - QSFP FDR based Infiniband - MSX1036F-1SFR Mellanox 36 - QSFP40 GbE - MSX1036B-1SFR	



Setup 3 (Firmware/Software)

DELL C6220

- Operating System: Scientific Linux CERN SLC release 6.2 beta (Carbon)
- Linux version: 2.6.32-220.2.1.el6.x86_64
- OFED version: OFED.1.5.3.3.1.0
- Ethernet driver: mlx4_en version 1.5.8.3
- XDAQ version: release 11
- TCP test: sock application http://www.icir.org/christian/sock.html

Mellanox - ConnectX-3 VPI

- Firmware version: 2.11.500

DELL mezzanine Mellanox FDR CX3

- Firmware version: 2.10.4492

Mellanox 36 – MSX1036F-1SFR

- Firmware version: 9.1.3190
- Mellanox MLNX-OS™ version: 3.2.0300

Mellanox 36 – MSX1036B-1SFR

- Firmware version: 9.1.6294
- Mellanox MLNX-OS™ version: 3.2.0506