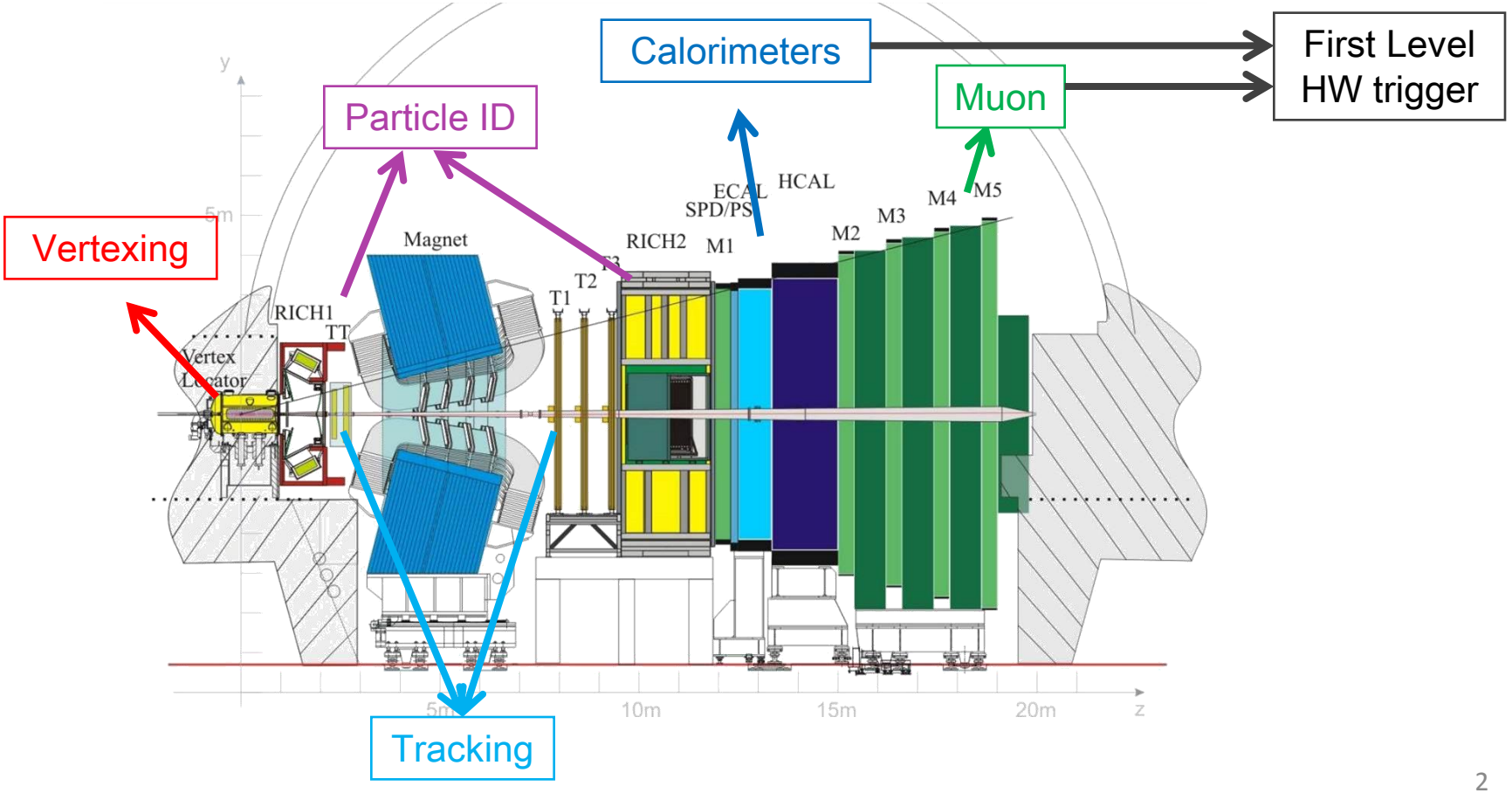
PCIe based readout for the LHCb upgrade

U. Marconi, INFN Bologna
On behalf of the LHCb Online Group
(Bologna-CERN-Padova)
CHEP2013, Amsterdam, 15th October 2013
Presented by Rainer Schwemmer

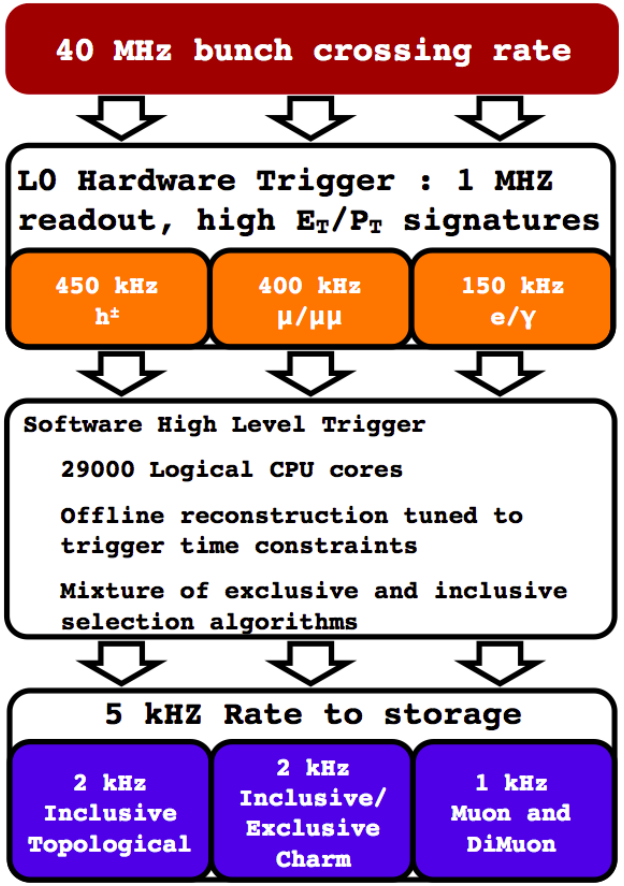
The current LHCb detector

LHCb is a forward detector at LHC, optimized for heavy quark physics. As a single harm spectrometer it offers a unique coverage in the pseudo-rapidity range **2 < η < 5**. Covering **4% of solid angle**, it allows catching **40% of the heavy quark production cross-section**.

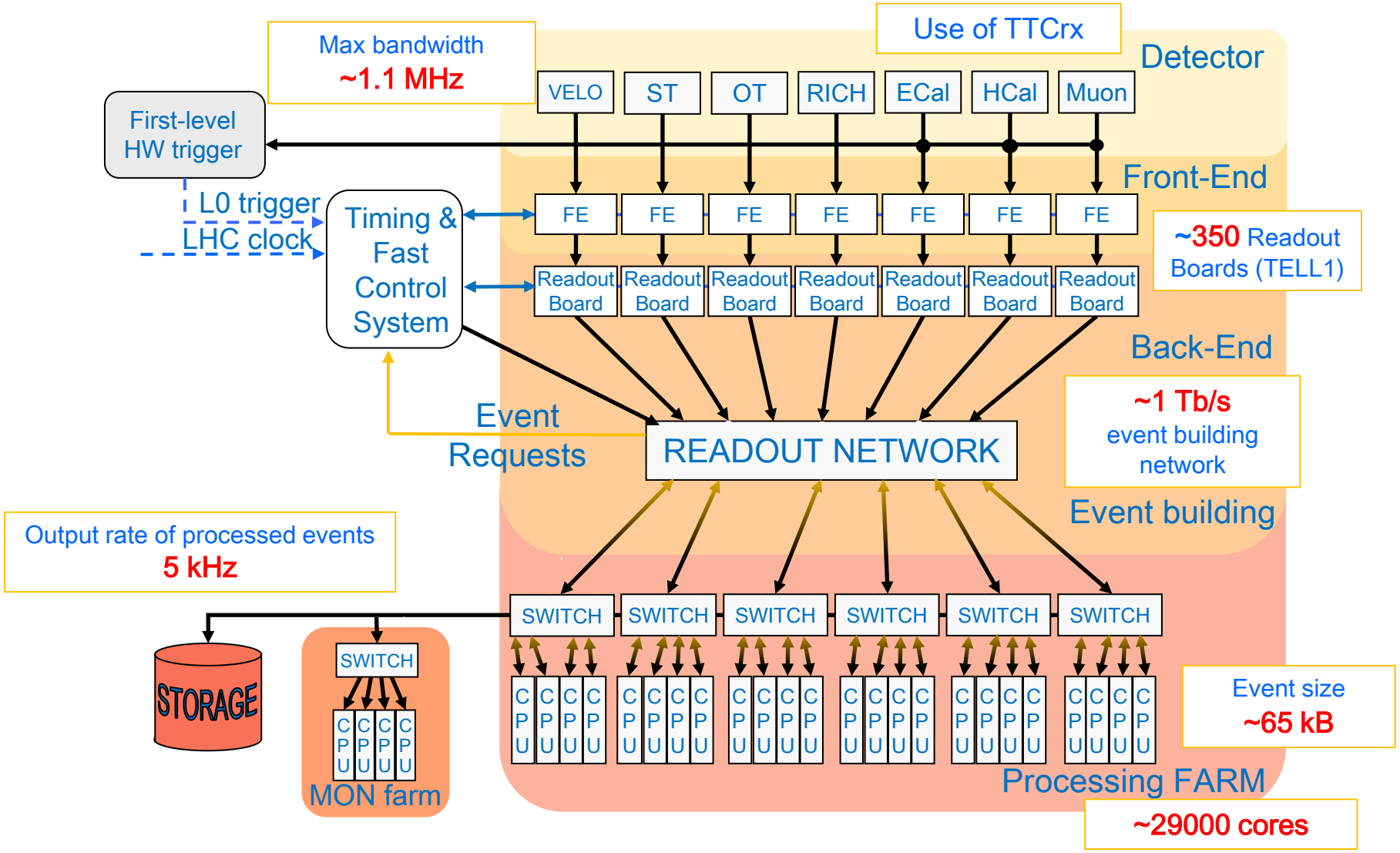


The current LHCb Trigger

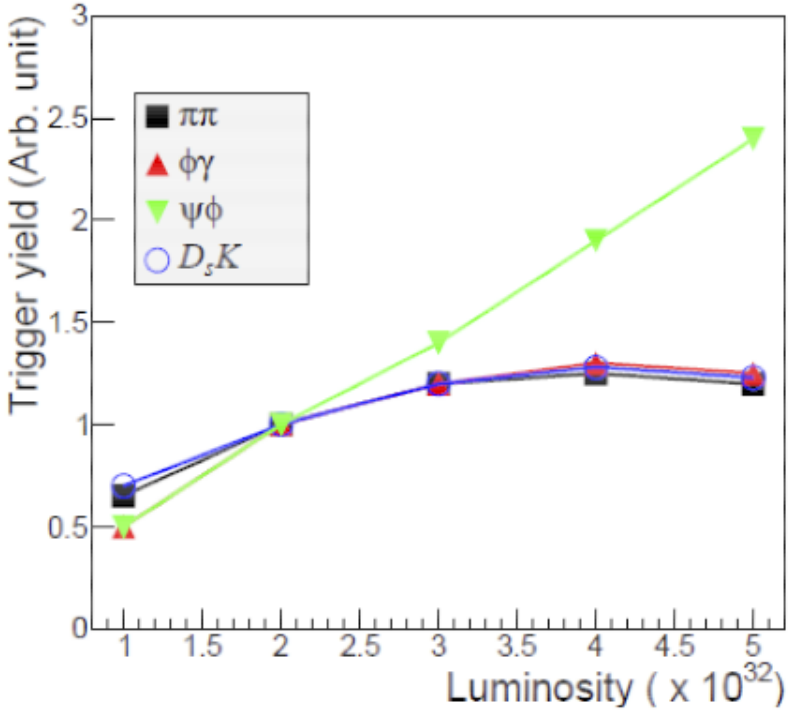
- The **Level-0 trigger** based on the signals from ECAL, HCAL and MUON detectors read at 40 MHz, operates on custom electronics, with a **maximum output rate limited to 1.1 MHz**.
  - Fully pipelined, constant latency of about 4 μs.
  - Bandwidth to HLT ~ 4 Tb/s, GOL serdes, optical links.
  - High p_T muon (1.4 GeV) or di-muon.
  - High p_T cluster in HCAL (3.5 GeV) or ECAL (2.5 GeV)
- **HLT1** is a software trigger
  - Reconstruct VELO tracks and primary vertices
  - Select events with at least one track matching **p**, **p_T** , **impact parameter** and **track quality cuts**.
  - Accept around **50 kHz**.
- **HLT2** performs inclusive or exclusive selections of the events.
  - Full track reconstruction, without particle-id.
  - 30% of the events are **deferred**: temporarily stored on disk and processed during the inter-fills.
  - Total accept rate to disk for offline analysis is around **5 kHz**.

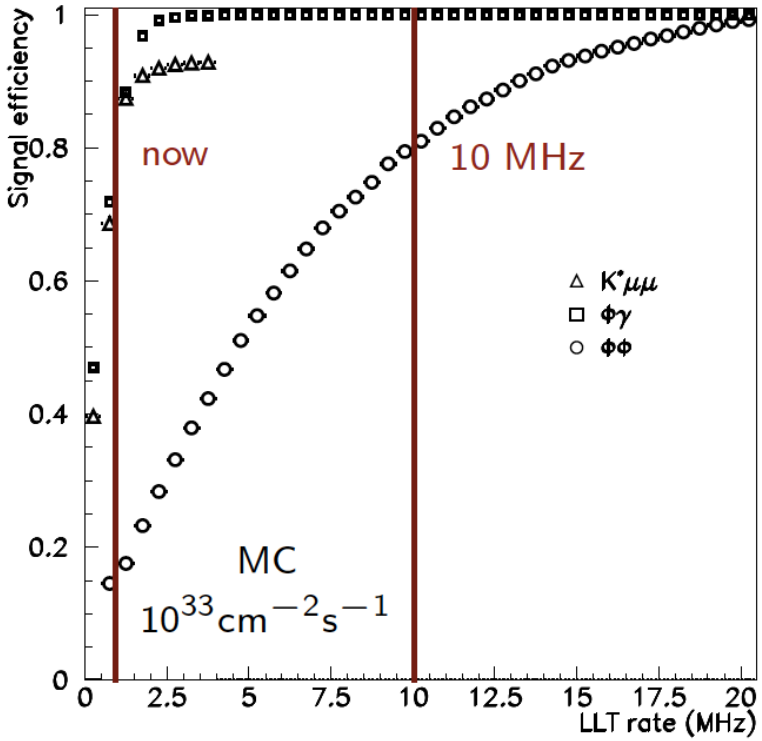


The current LHCb readout



The 1MHz trigger rate limitation



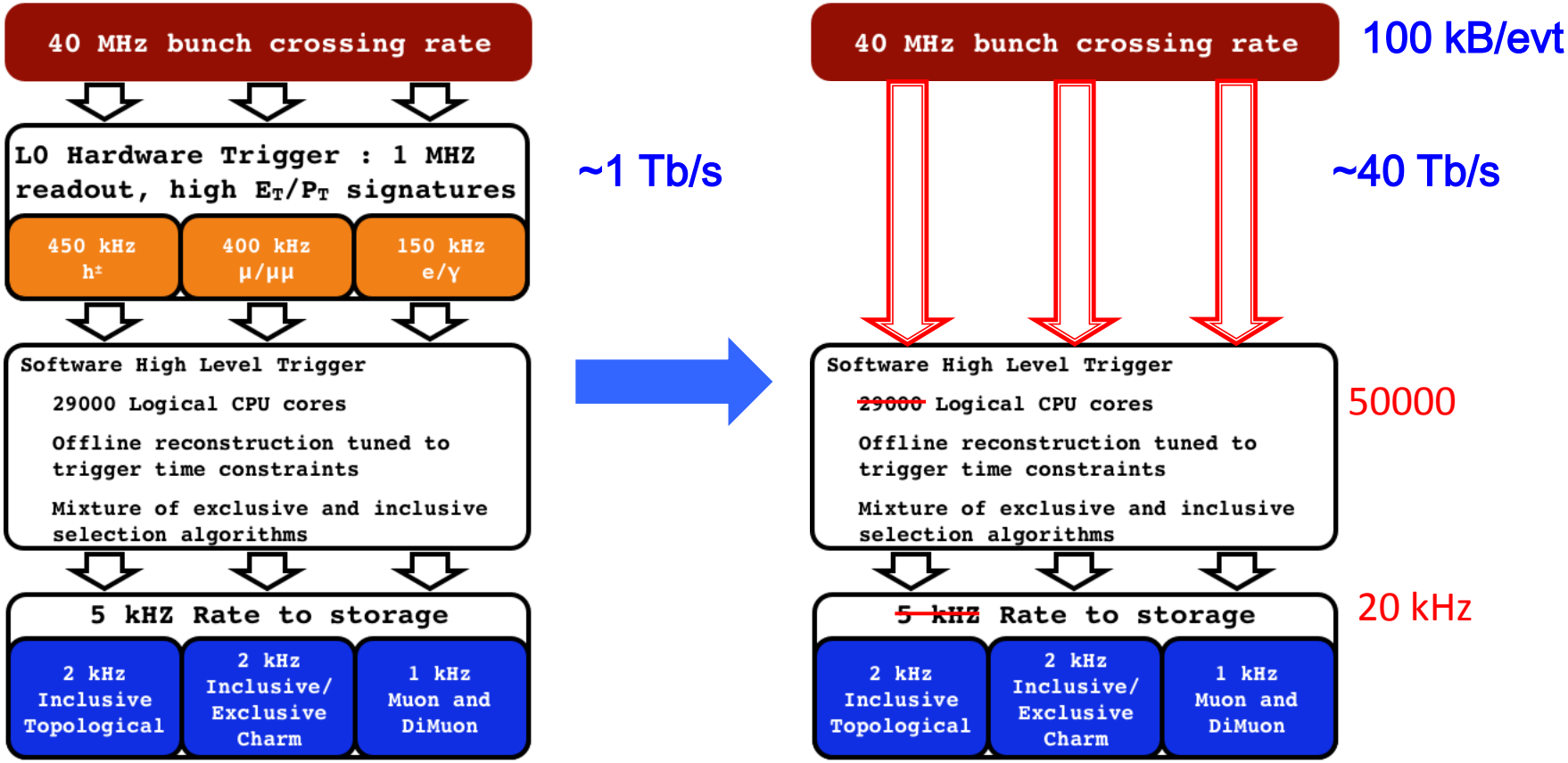


- Due to the available bandwidth and discrimination power of the hadronic L0 trigger LHCb has experienced the **saturation of the trigger yield on the hadronic channels**.
- Running at higher luminosity ($2.x10^{33}$ $cm^{-2}s^{-1}$ **)** and increasing the first level (hardware) trigger rate would considerably improve the signal detection efficiency, particularly on the hadronic channels.

"Framework TDR for the LHCb Upgrade: Technical Design Report"; The LHCb Collaboration,CERN/LHCC-2012-007, LHCb-TDR-12 - 2012

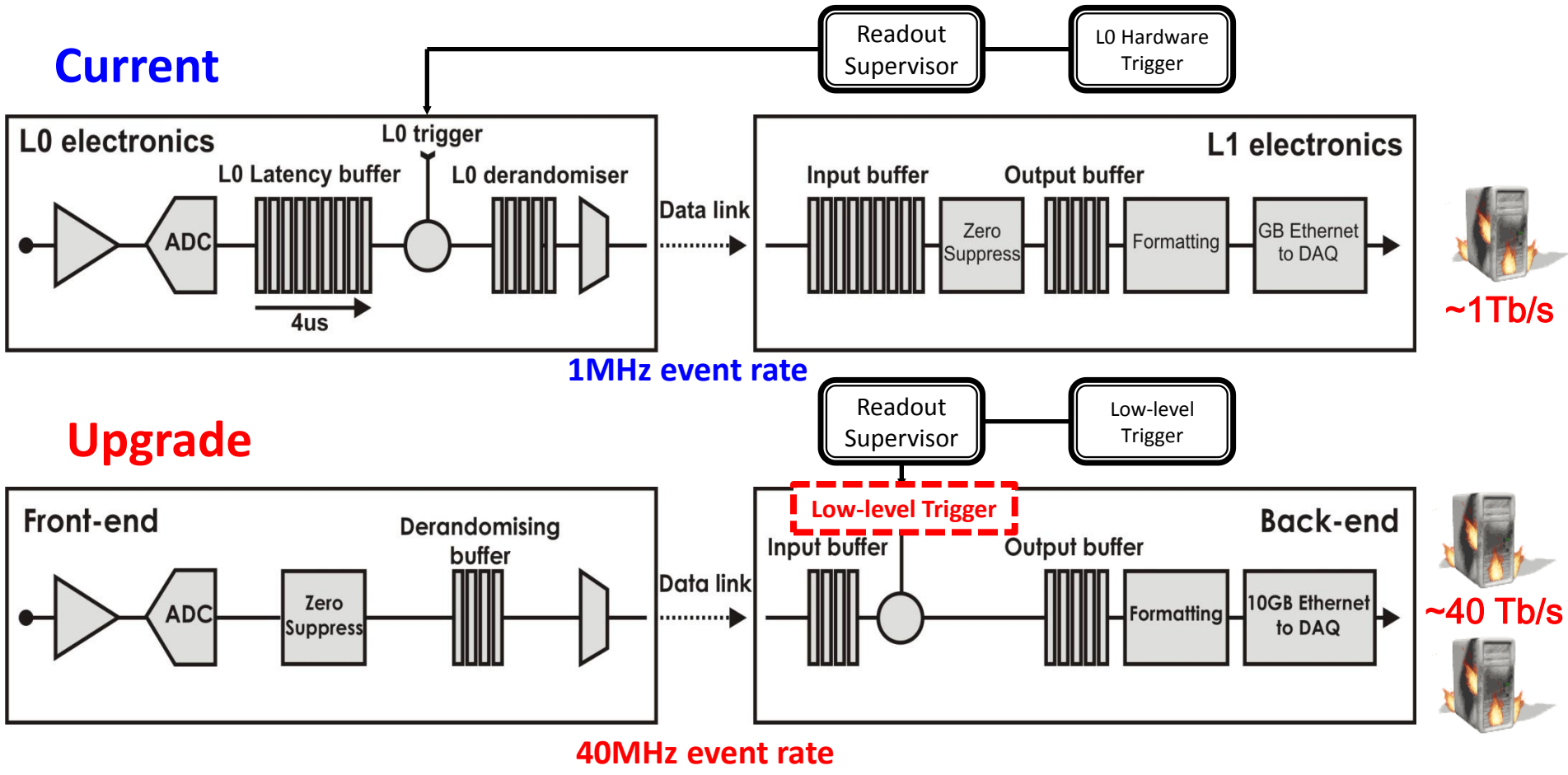
The LHCb upgrade strategy (I)

Event readout at the bunch crossing rate of 40 MHz.
Remove the first-level hardware trigger to fully exploit the HLT trigger capabilities to perform an efficient events selection.



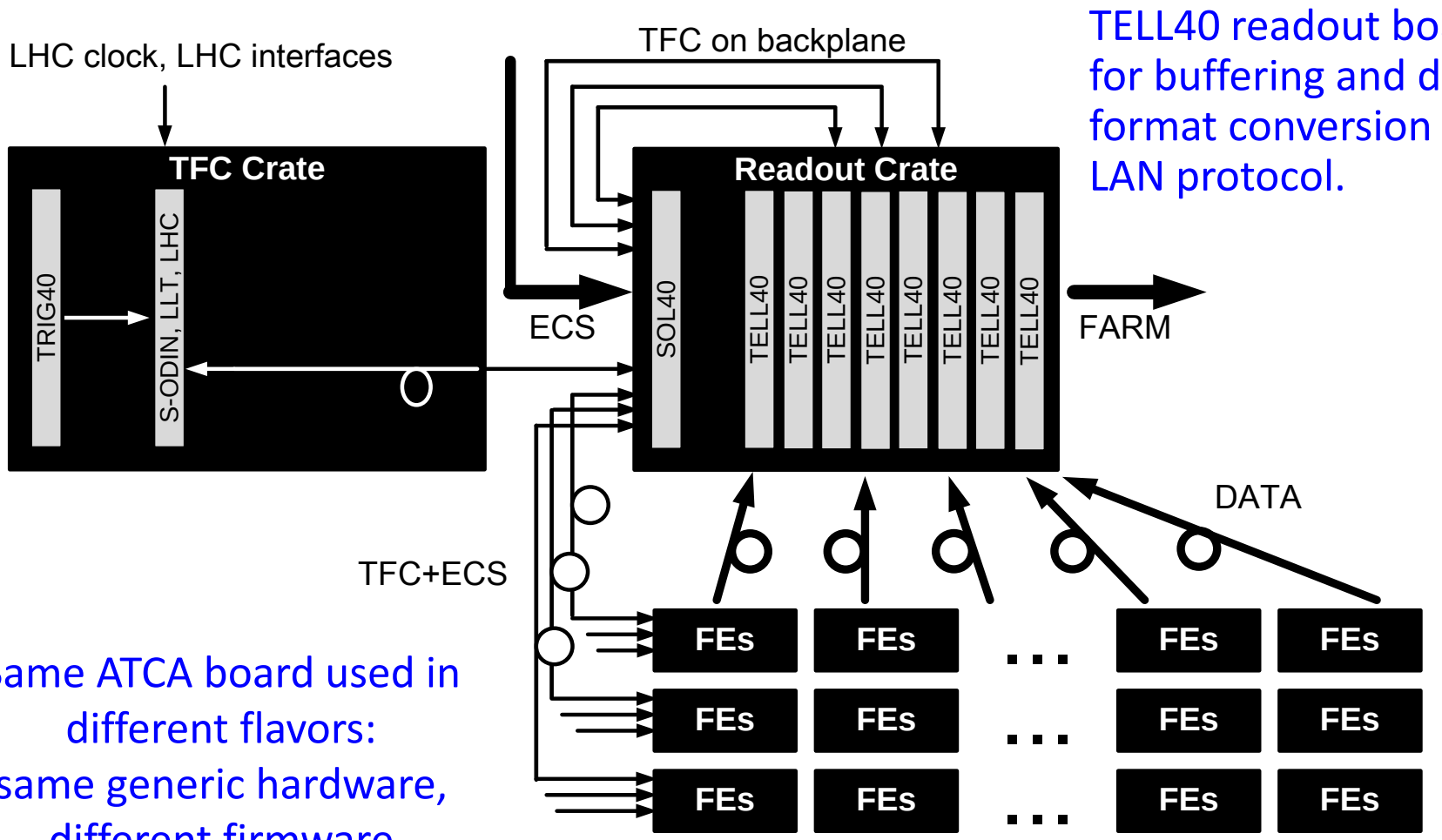
The LHCb upgrade strategy (II)

Accept all LHC bunch crossing: trigger-less Front-End electronics



Compress (zero-suppress) data already at the FE → 12000 links

The ATCA-based LHCb readout



TELL40 readout board for buffering and data format conversion to LAN protocol.

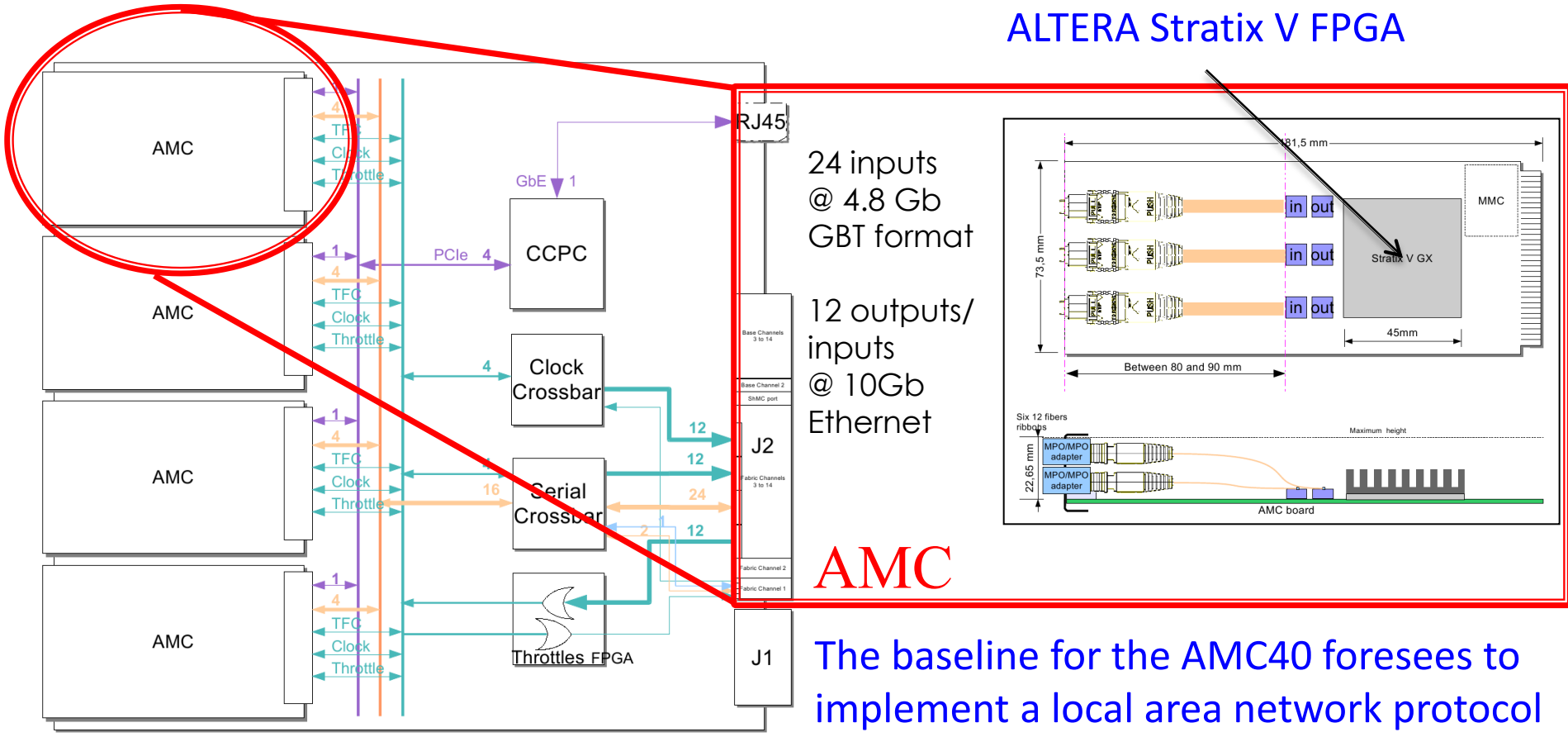
Same ATCA board used in different flavors: same generic hardware, different firmware

~12000 optical links from the FEE

The TELL40

Direct evolution of the current LHCb readout board TELL1

ALTERA Stratix V FPGA



The baseline for the AMC40 foresees to implement a local area network protocol (LAN) directly in the FPGA.

96 inputs in total @ 4.8 Gb → processing in FPGA → 48 x 10 GbE ports

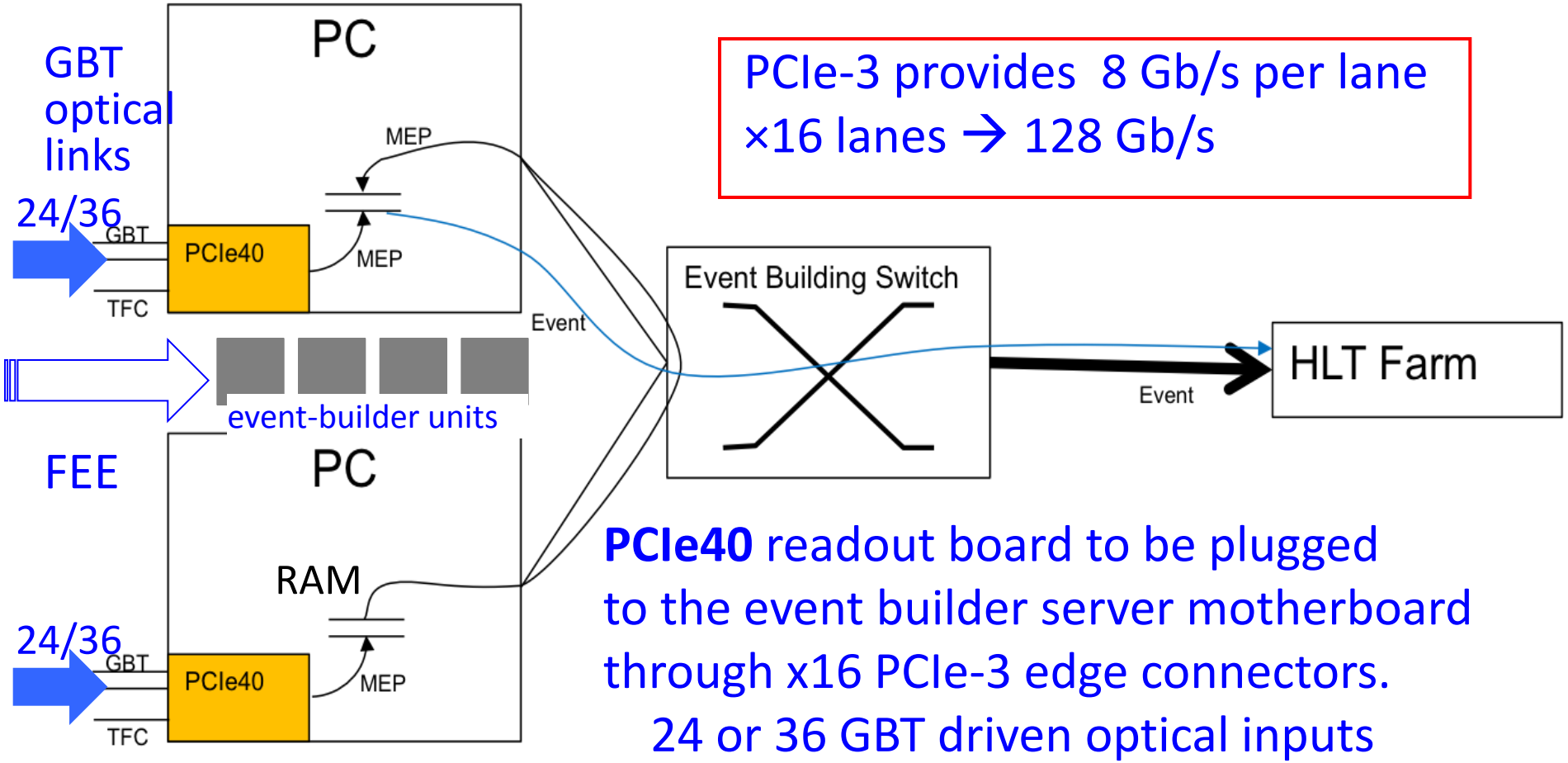
~0.5 Tb/s data aggregator board with 10 GbE to FARM

The PCIe alternative

- PCIe is the high-speed serial computer expansion bus standard: connects the host system-processor with integrated-peripherals (surface-mounted ICs) and add-on peripherals (expansion cards).
- The connection between two PCIe devices (link) is built up from a collection of one or more **lanes**: x1, x4, x8, x16
- PCIe Gen 3 **bandwidth** is 1 GB/s per lane.
  - Overhead of the 128b/130b encoding scheme ~1%.
    It was 20% with the 8b/10b Gen2.
- The FPGA implementation of a full network protocol (10 GbE or InfiniBand) is expensive. PCIe is a much simpler protocol to use to forward data to the event-builder.
  - 20% resource consumption to implement 10 GbEthernet
- FPGA vendors provides **PCIe-3 Express Hard IP blocks**
  - Resource savings, pre-verified, protocol-compliant, embedded memory buffers included, etc.

PCIe readout and event-building

24. × 4.8 Gb/s = 115.2 Gb/s or 36. × 3.2 = 115.2 Gb/s



PCIe-3 provides 8 Gb/s per lane
×16 lanes → 128 Gb/s

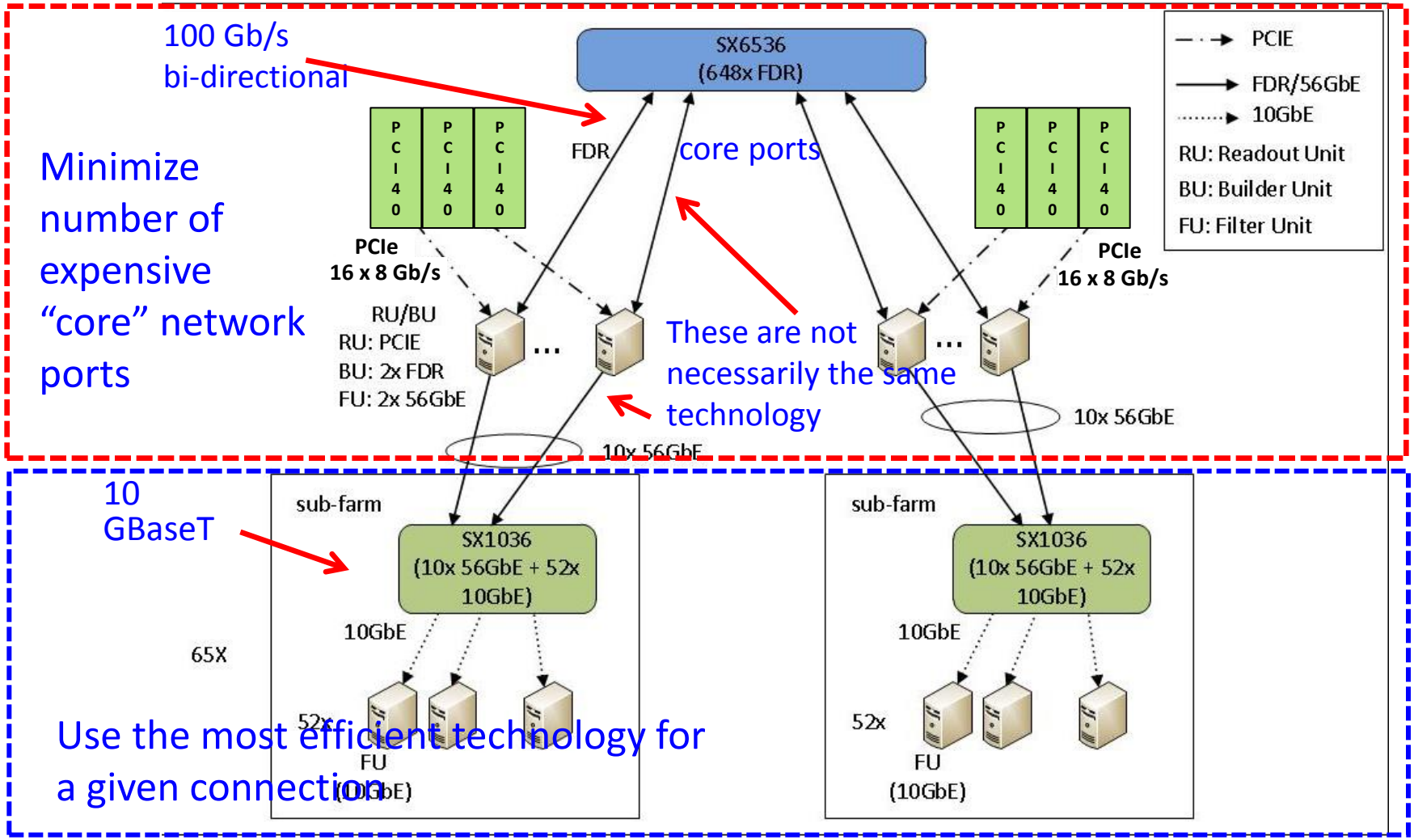
PCIe40 readout board to be plugged to the event builder server motherboard through x16 PCIe-3 edge connectors.

- 24 or 36 GBT driven optical inputs
- 24 input →500 event-builder servers
- 36 input →340 event-builder servers

From the FEE
~12000 GBT
optical links

The event-builder network

PCIe readout provides the max flexibility in the network technology



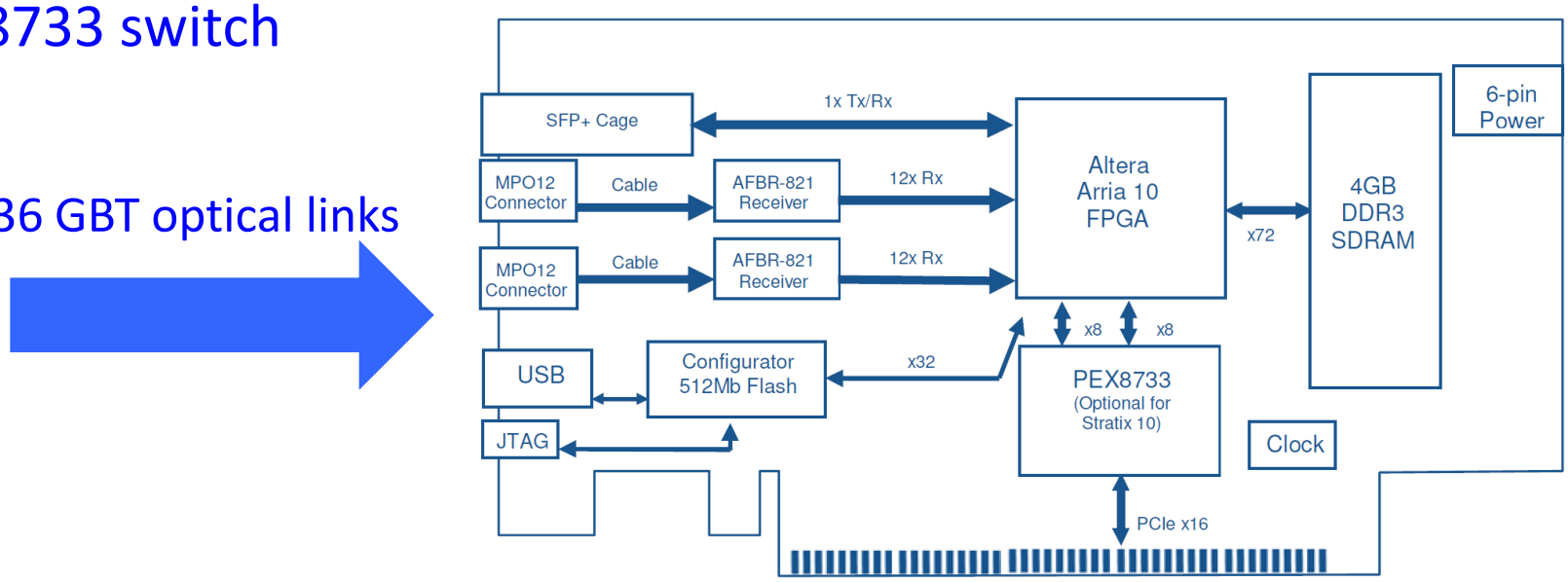
The event-filter (HLT)

"DAQ Architecture for the LHCb Upgrade". G. Liu and N. Neufeld. CHEP2013.

PCIe40

- Custom-made readout boards with fast optical links and powerful FPGAs
- Minimal variant – no external memories / no transmitters for GBT
- Power from 75W up to 450W (if needed, using GPU standard)
- PEX8733 switch

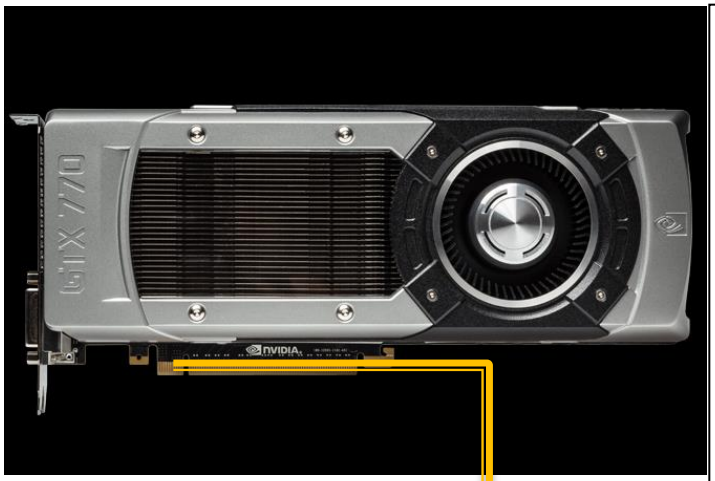
24/36 GBT optical links



- It can evolve to a newer version with 3 Avago transceivers and the Arria 10 FPGA.

PCIe-3 bandwidth tests using GPU

https://lbonupgrade.cern.ch/wiki/index.php/I/O_performance_of_PC_servers

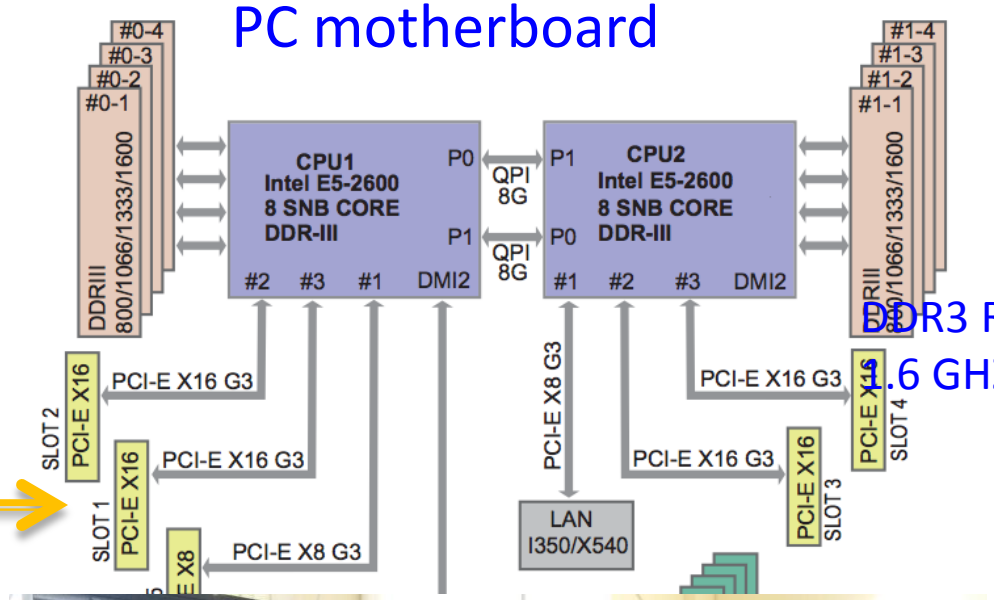


x16 PCIe-3
edge connector

The hardware used in our tests is the very first commercially available. Data transfer from the GPU internal memory to the CPU RAM.

CUDA test program:
Device-to-Host memory copy

PC motherboard



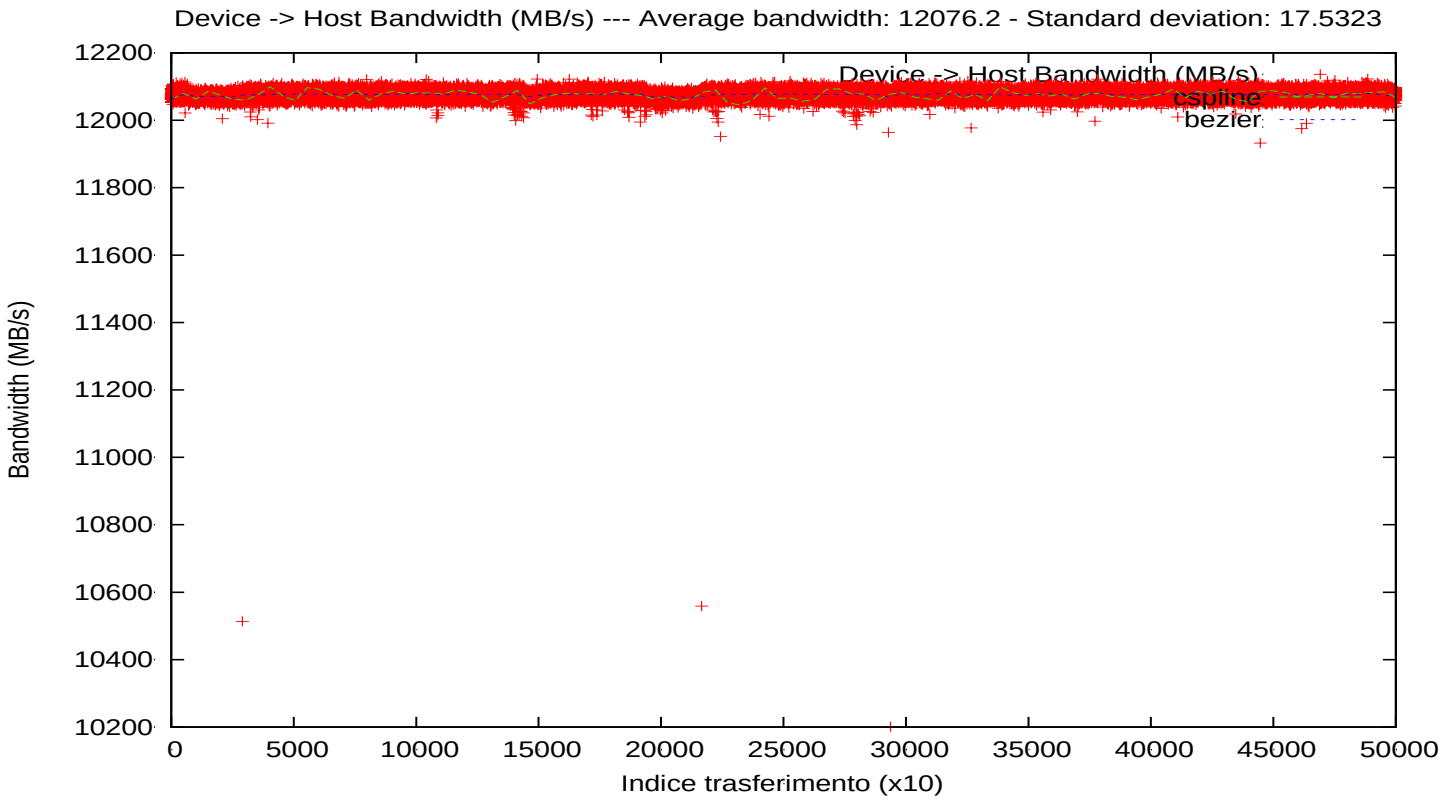
DDR3 RAM
1.6 GHZ, 64 bits

PCIe-3 bandwidth test results (I)

- GPU: NVIDIA GTX770 PCIe-3 x16
- Motherboard: Super X9DRD-iF
- Record size used for transfer: **32 MiB**
- **Device-to-host bandwidth: 12076 MiB/s = 101.3 Gb/s**

Small record size

Recorded throughput (in MiB/s).

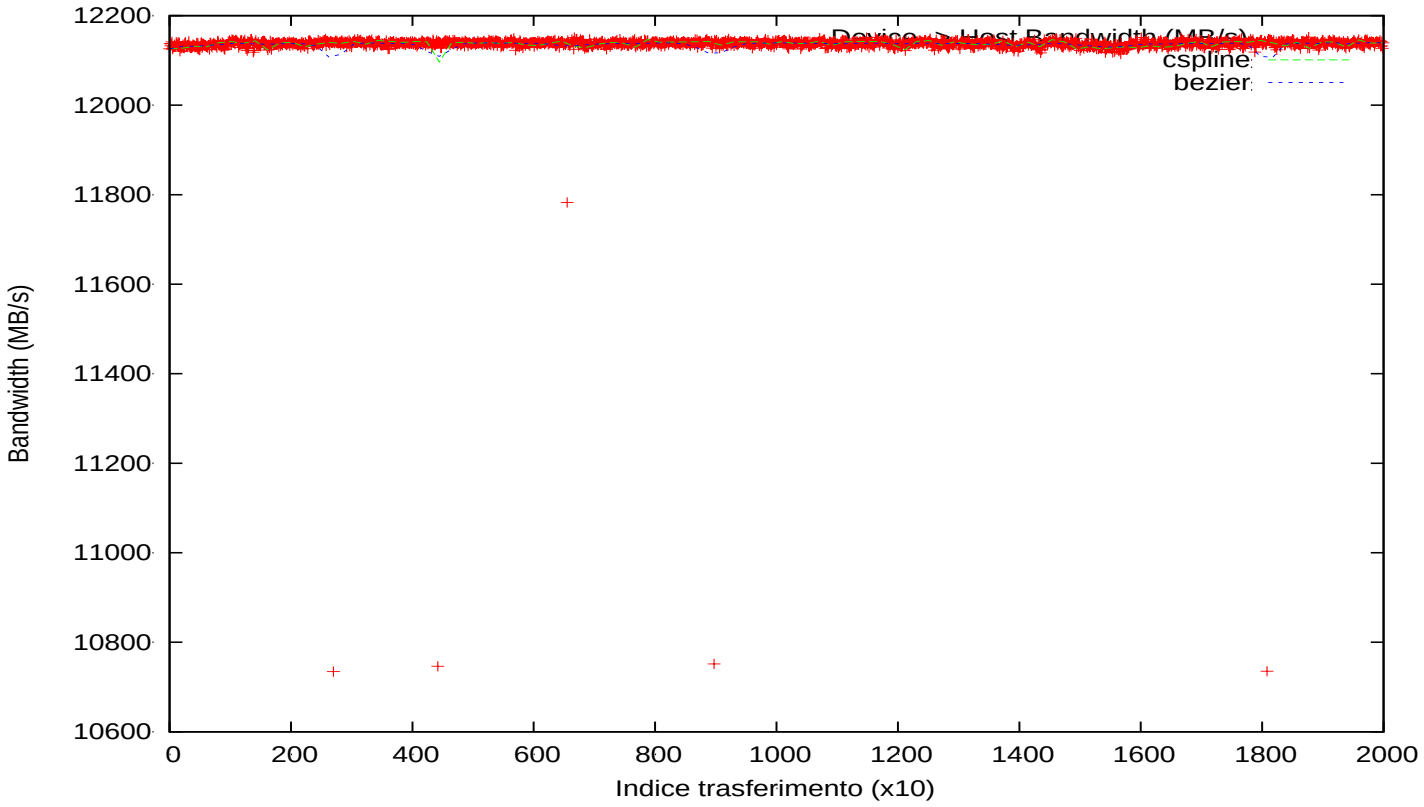


PCIe-3 bandwidth test results (II)

- GPU: NVIDIA GTX770 PCIe-3 x16
- Motherboard: Super X9DRD-iF
- Record size used for transfer: **1500 MiB**
- **Device-to-host bandwidth: 12135 MiB/s = 101.8 Gb/s**

Big record size

Recorded throughput (in MB/s).

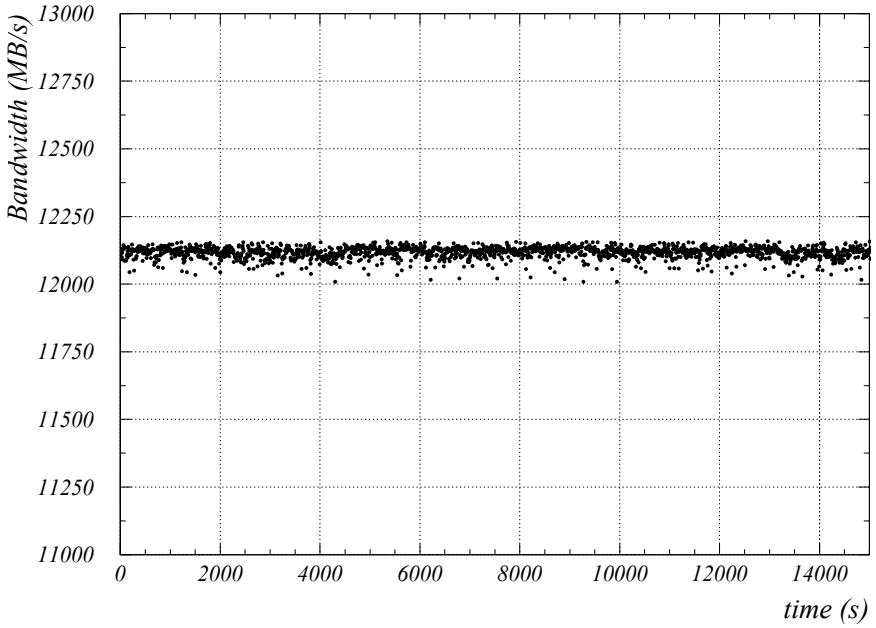


PCIe-3 bandwidth test results (III)

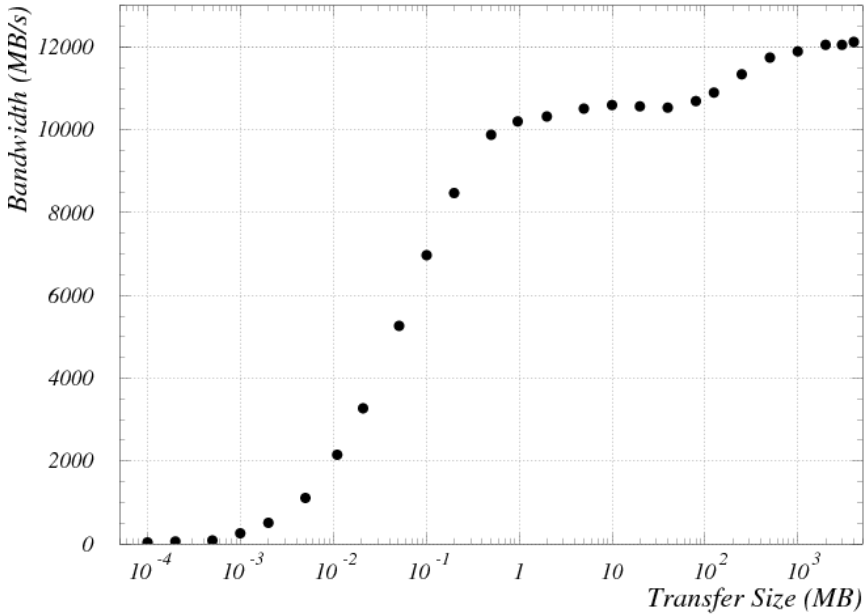
- GTX Titan (NVIDIA) PCIe-3 x16 GPU connected to the motherboard ASRock - Z77 Extreme - 4GPU
- Record size used: **2000 MB**
- **Device-to-host bandwidth: 12210 MiB/s = 102. 4 Gb/s**

Big record size
Different setup

Recorded throughput vs. the reporting records.



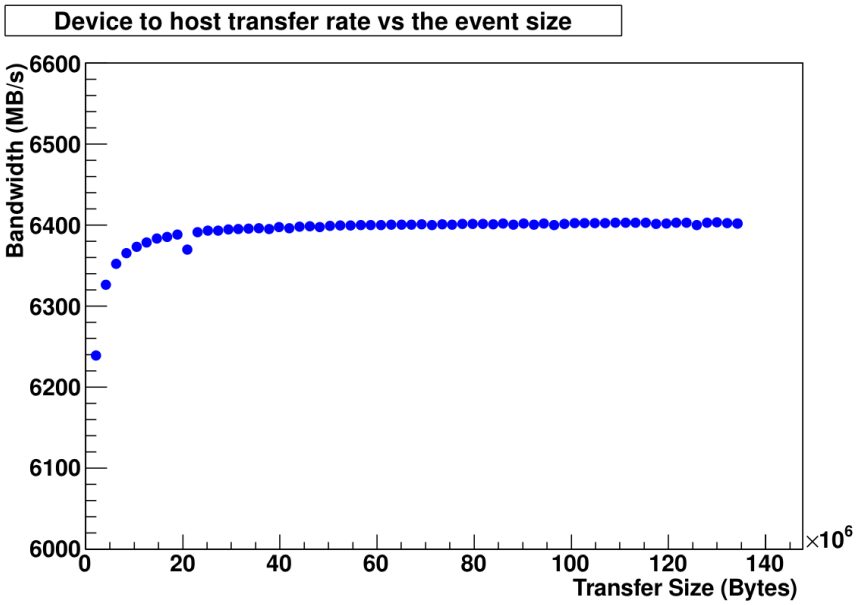
Recorded throughput (in MB/s) vs. the record size.



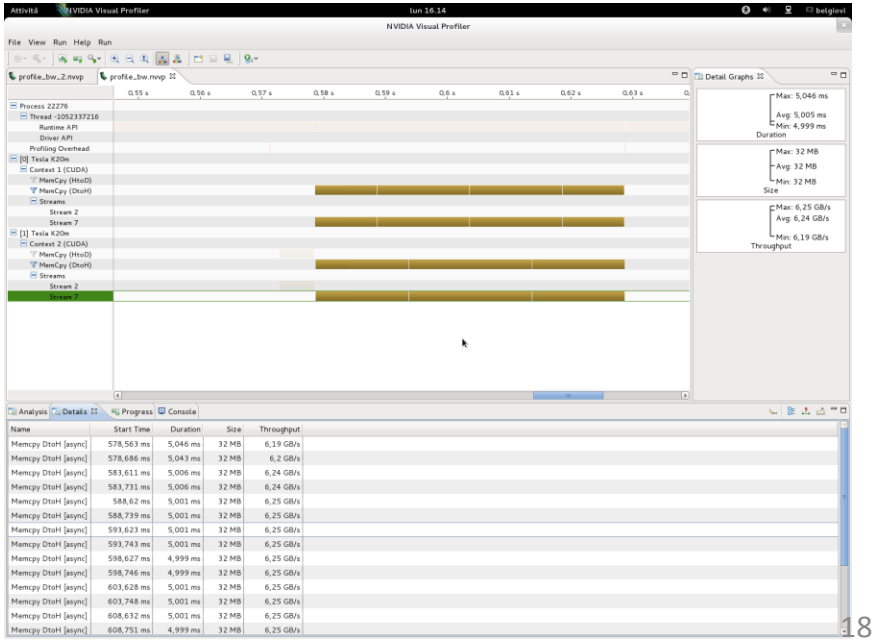
115.2 Gb/s or 24. × 4.8 Gb/s

PCIe-3 bandwidth test results (IV)

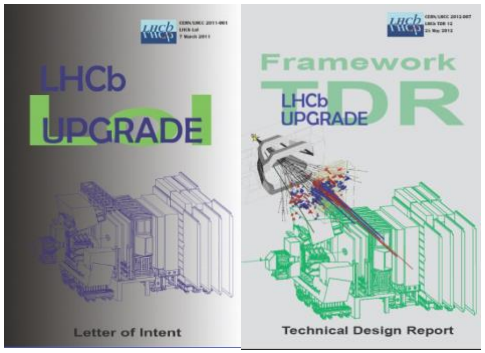
- Twofold synchronous data transfer to the server RAM using identical PCIe-2 GPUs connected to the PCIe-3 motherboard slots.
- GPU used for test: NVidia Tesla K20m x16 PCIe-2 connected to the motherboard Supermicro X9DRG-HF
- Bandwidth: **12787.9** MiB/s = 109 Gb/s, record size 33.5 MB



NVIDIA Profiler: synchronous threads



LHCb data taking plan



LHCb startup

LHCb collected ~3 fb^{-1}

LHCb to collect 5-7 fb^{-1}

Upgrade!

2010 2011 2012 2013 2014 2015 2016 2017 2018

LHC Run I

E = 7(2010-2011) & 8(2012) TeV (pp)
@ 50 ns
+
E = 2.76 TeV (Pb^+Pb^+)
+
E = 5 TeV (pPb^+)

LHC Long Shutdown I

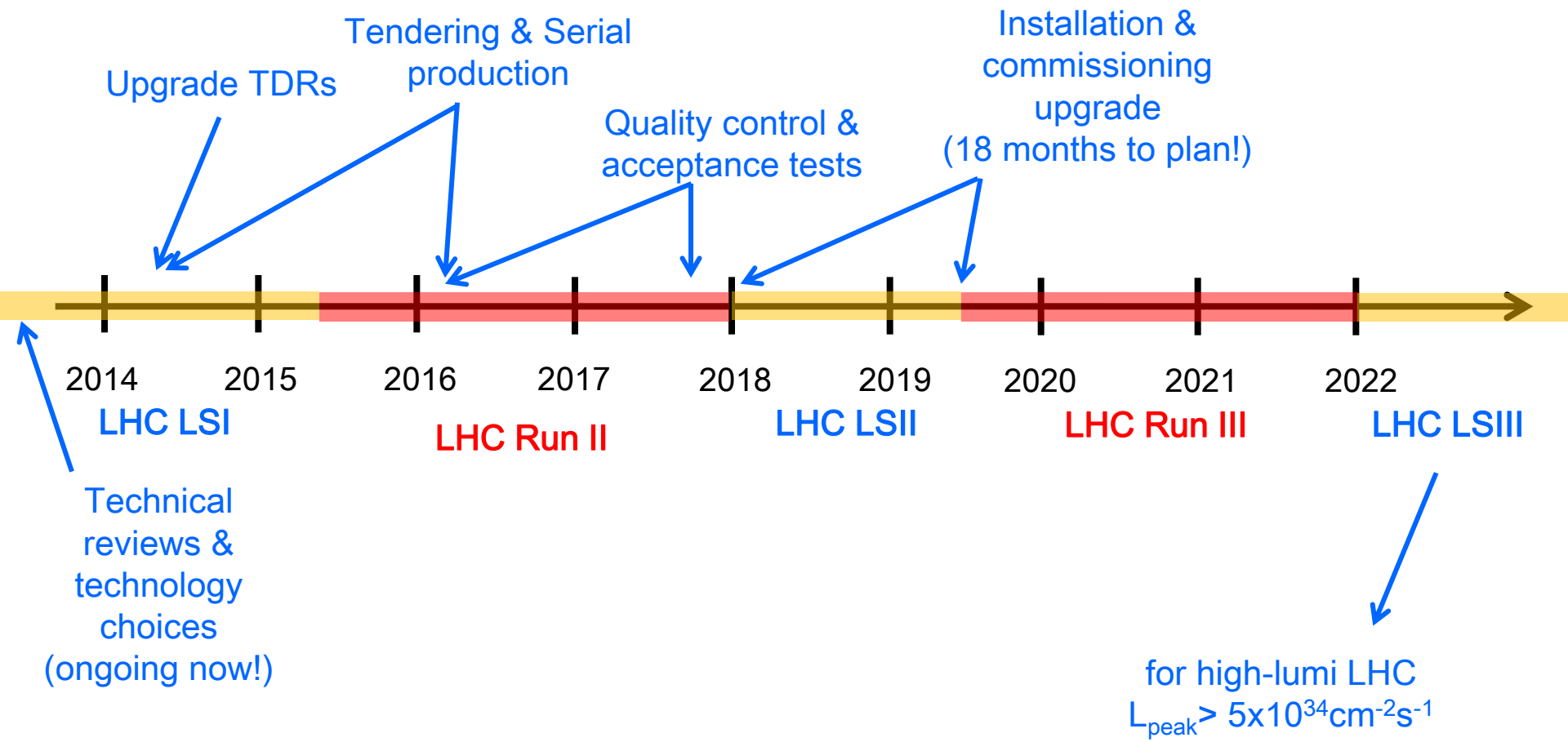
LHC splices repair
+ consolidation

LHC Run II

E = 13 TeV (pp)
@ 25 ns
+
L_{peak}= 2x10^{34}cm^{-2}s^{-1}
+
E = 2.76 TeV (Pb^+Pb^+)

Decision between the two solutions (ATCA vs PCIe) will be taken in a review in Q1 next year.

LHCb upgrade plan

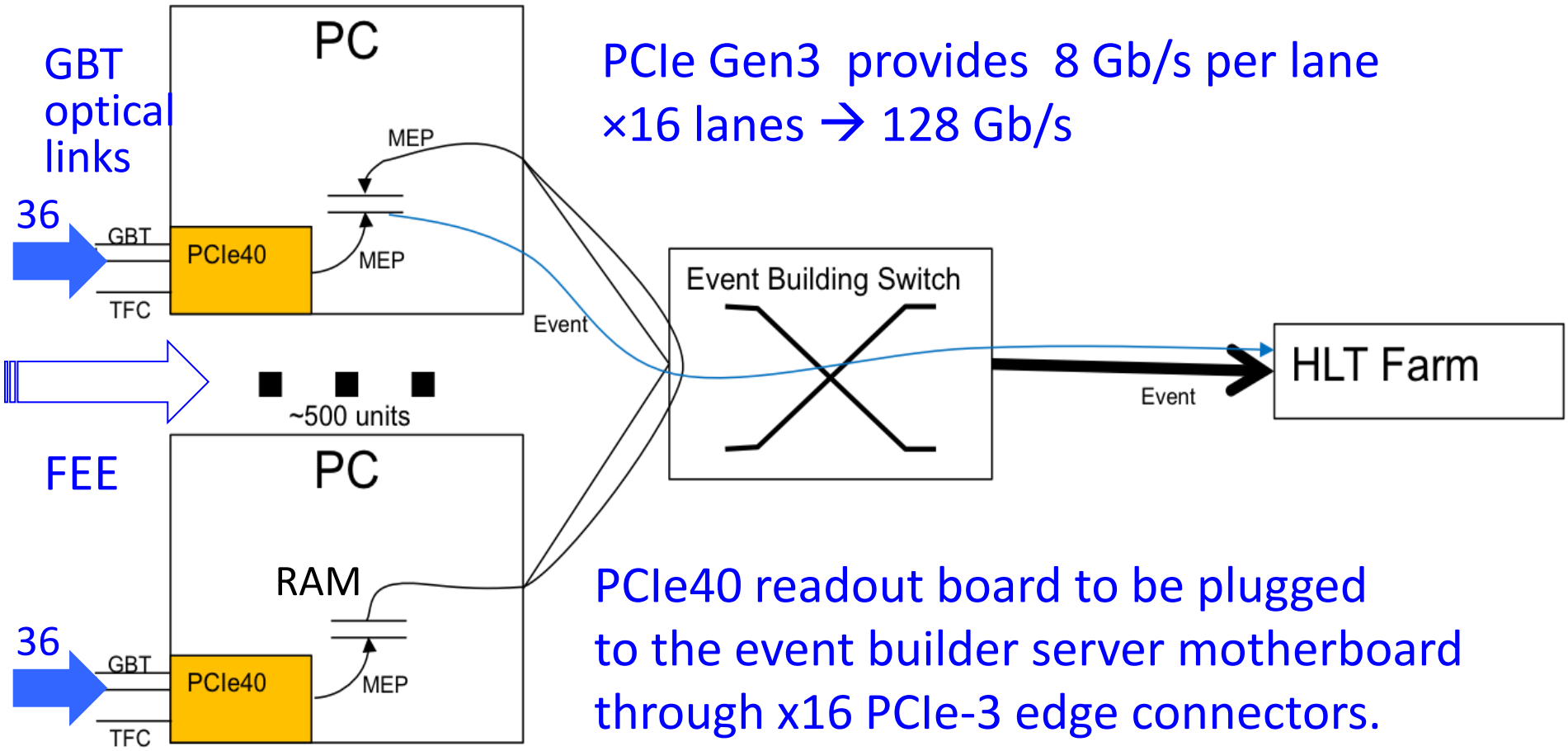


Conclusions

- PCIe is widely used to connects peripherals to host systems and has good survival chances in the future.
- A readout board for the LHCb upgrade looks very similar to a PCIe network card.
- PCIe-3 is a simple protocol and provides large bandwidth.
- PCIe provides maximum flexibility in later networking choice.
- GPUs prove that already today the necessary bandwidth is available.

PCIe readout and event-building

36. × 3.2 = 115.2 Gb/s



PCIe Gen3 provides 8 Gb/s per lane
×16 lanes → 128 Gb/s

PCIe40 readout board to be plugged to the event builder server motherboard through x16 PCIe-3 edge connectors.
~ 36 GBT driven optical inputs from the FEE
~ 500 event-builder server needed

From the FEE ~12000 GBT optical links