LHCb Trigger & DAQ System Overview

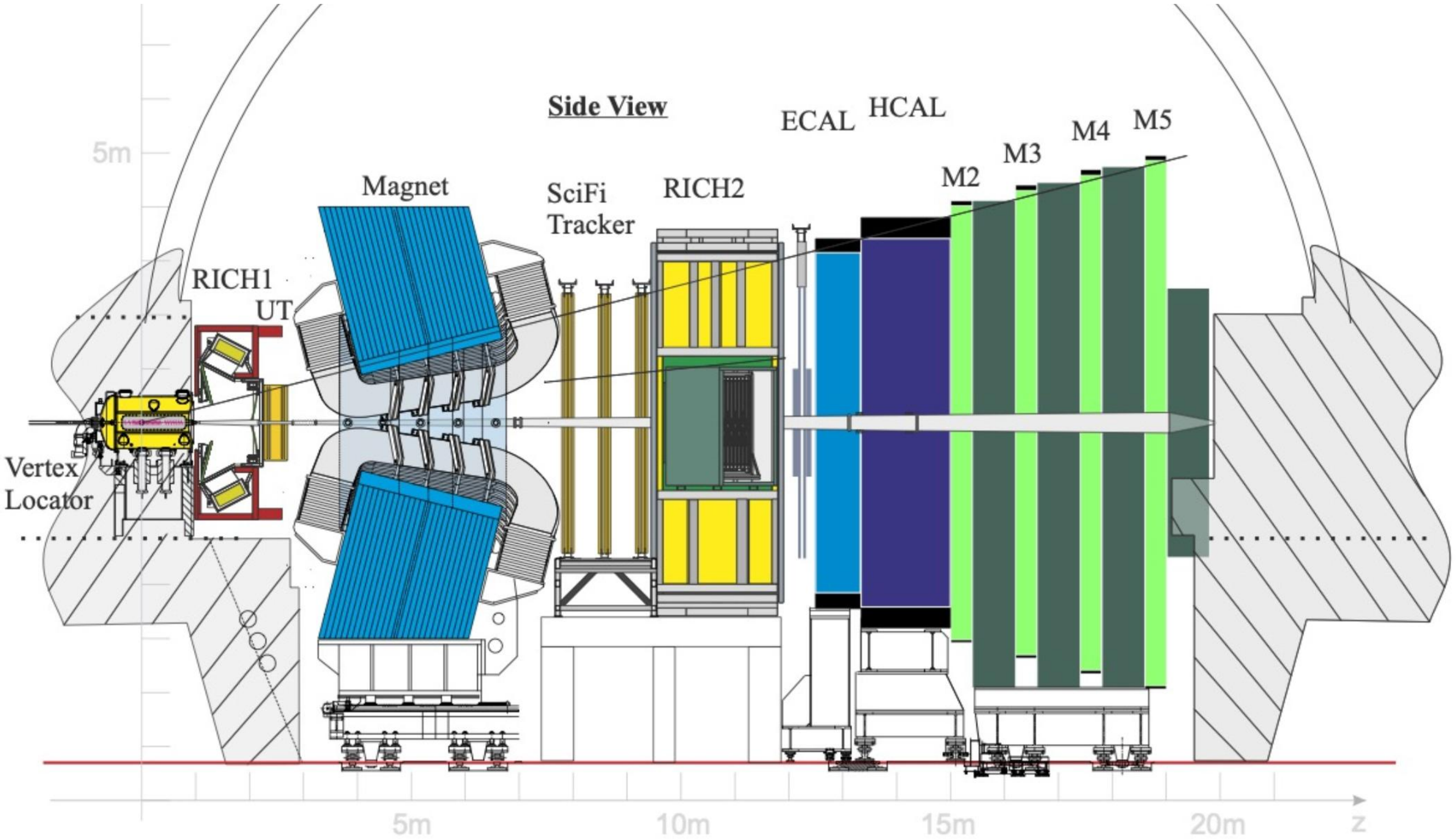
刘国明
华南师范大学

湖南湘潭，2025.7.4

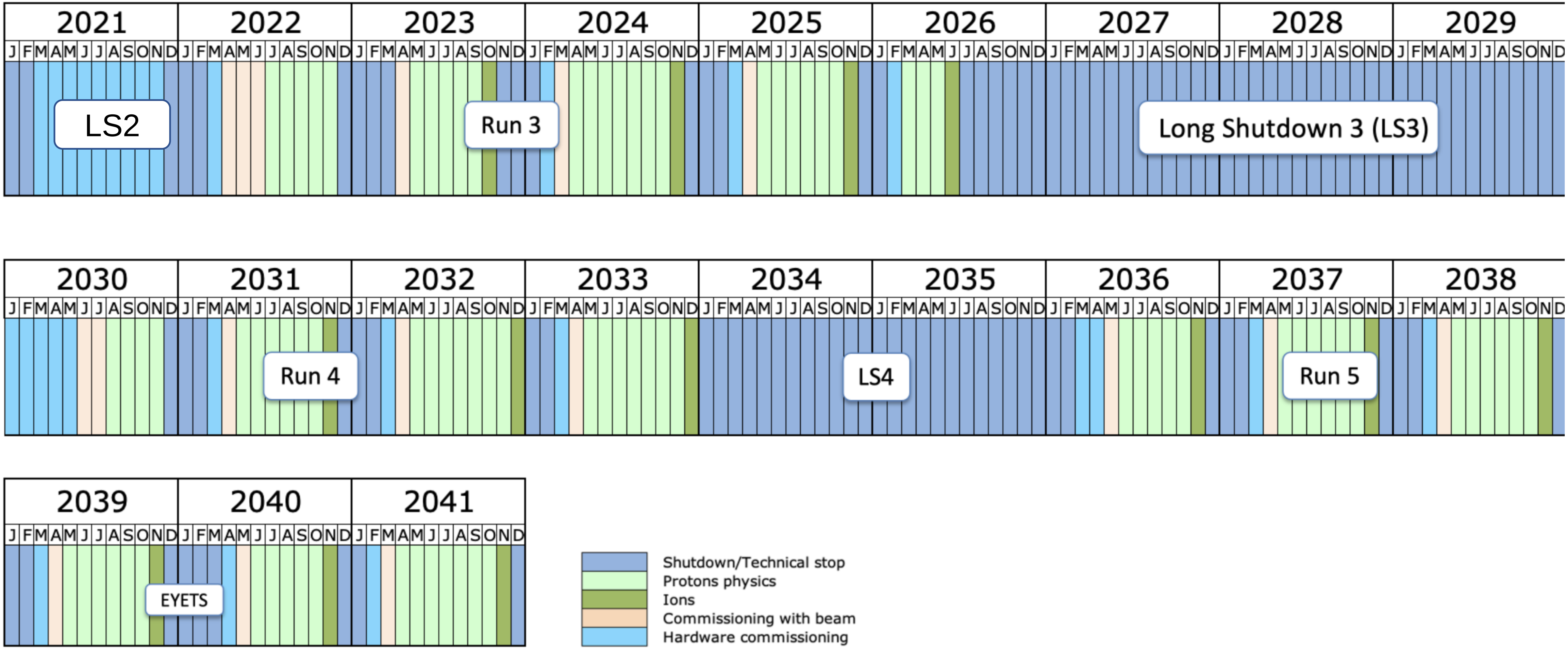
LHCb Detector in Run3

- Single-arm forward spectrometer at the LHC
- p-p bunch crossing rate: 30 MHz
- Luminosity:

  $2 \times 10^{33}\ \mathrm{cm^{-2}s^{-1}}$



LHC Schedule

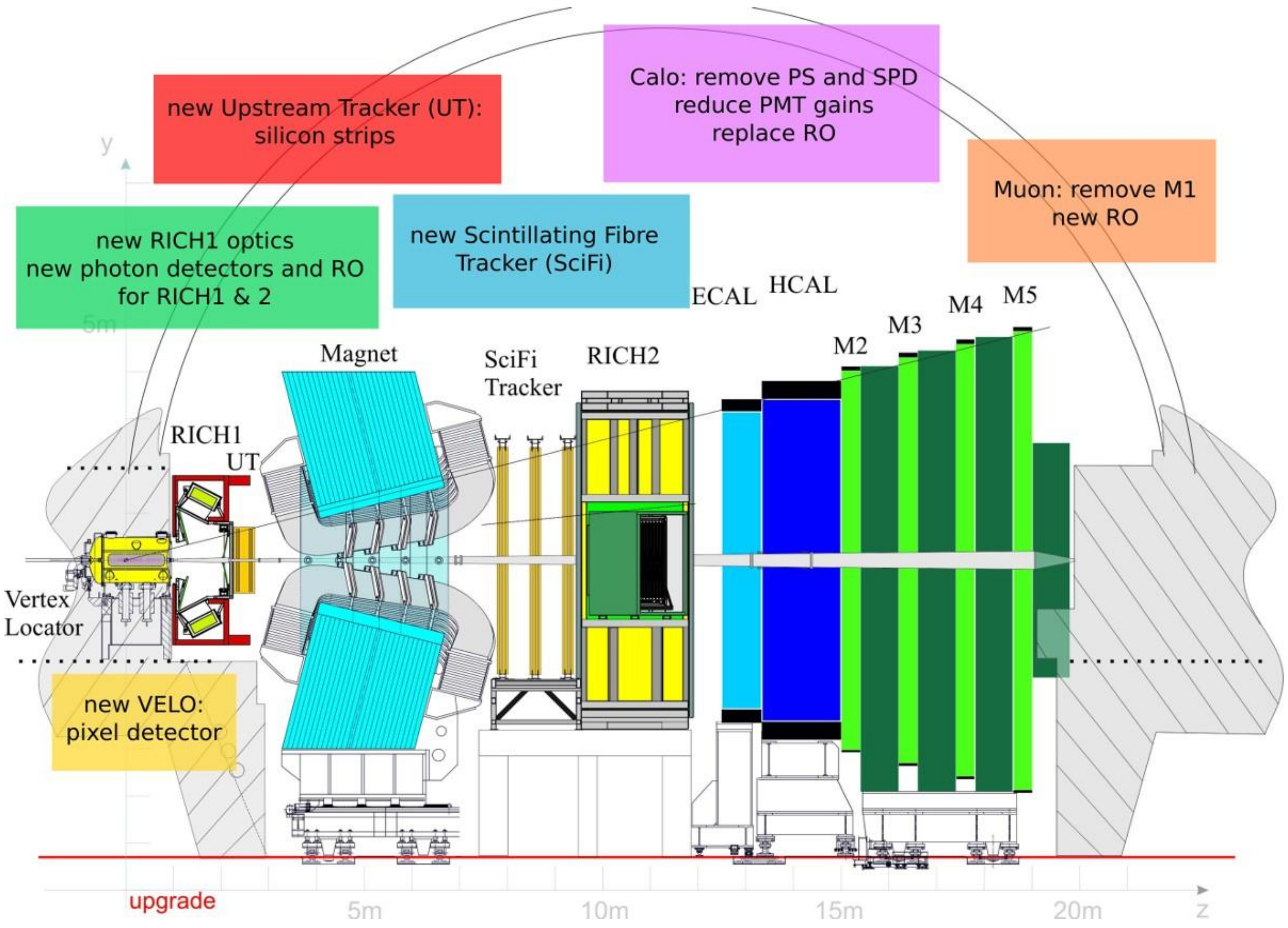


🟦 Shutdown/Technical stop
🟩 Protons physics
🟩 Ions
🟧 Commissioning with beam
🟦 Hardware commissioning

Last update: November 24

LHCb Upgrade I: trigger-less readout

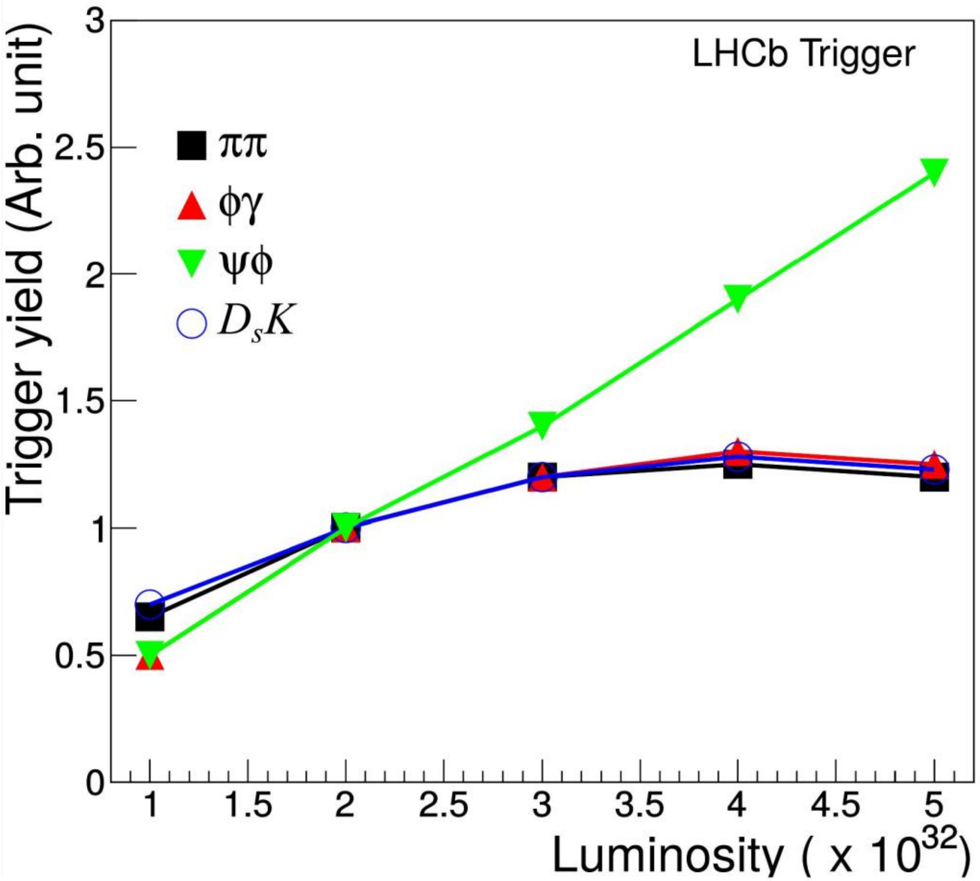
- Complete replacement of DAQ
- fully software trigger (HLT1 + HLT2)



Why trigger-less readout?

- Run 1 & 2：
  - Instantaneous luminosity:

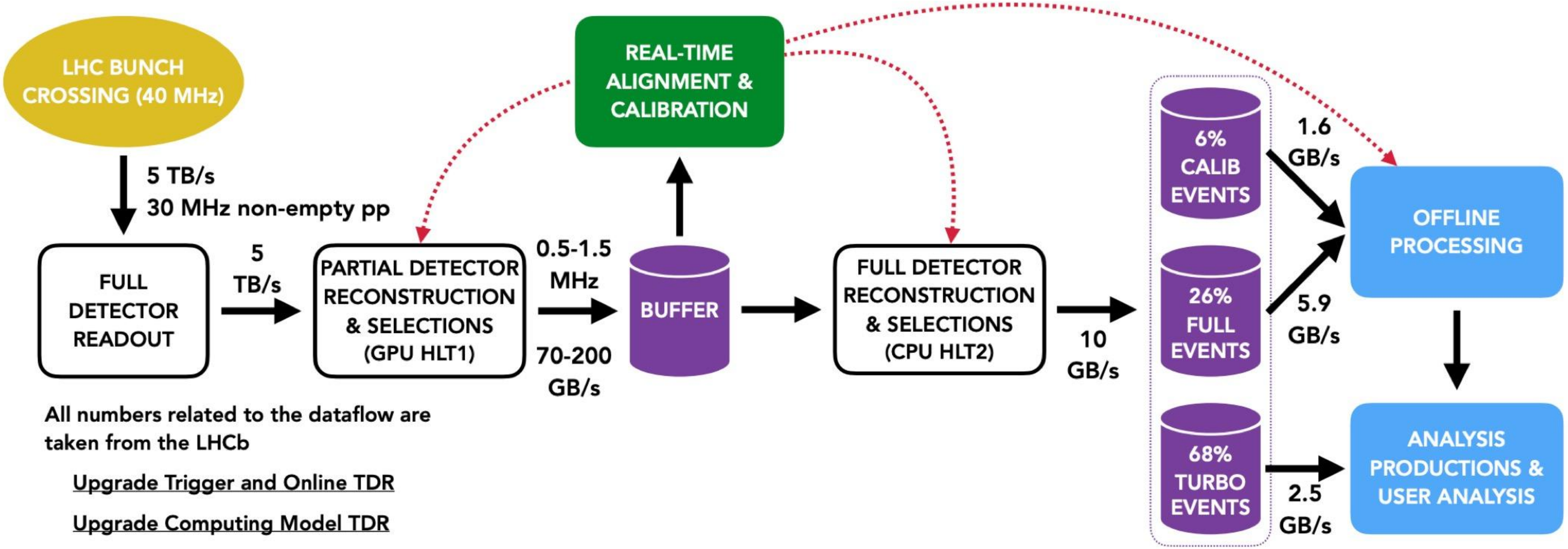
    $4 \times 10^{32}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}$
  - L0 trigger: hardware (40 →1 MHz)
    - using high Pt /Et signatures
    - 1 MHz limit saturates hadronic modes
- Solution: read full event at bunch-crossing rate
- high event rate but small event size



Dataflow in Run3

- **Two stages of software filtering:**
  - **HLT1 on GPGPUs**
  - **HLT2 on a CPU farm**

[LHCB-FIGURE-2020-016]



LHC BUNCH CROSSING (40 MHz)
5 TB/s
30 MHz non-empty pp
FULL DETECTOR READOUT
5 TB/s
PARTIAL DETECTOR RECONSTRUCTION & SELECTIONS (GPU HLT1)
0.5-1.5 MHz
70-200 GB/s
BUFFER
REAL-TIME ALIGNMENT & CALIBRATION
FULL DETECTOR RECONSTRUCTION & SELECTIONS (CPU HLT2)
10 GB/s
6% CALIB EVENTS
26% FULL EVENTS
68% TURBO EVENTS
1.6 GB/s
5.9 GB/s
2.5 GB/s
OFFLINE PROCESSING
ANALYSIS PRODUCTIONS & USER ANALYSIS

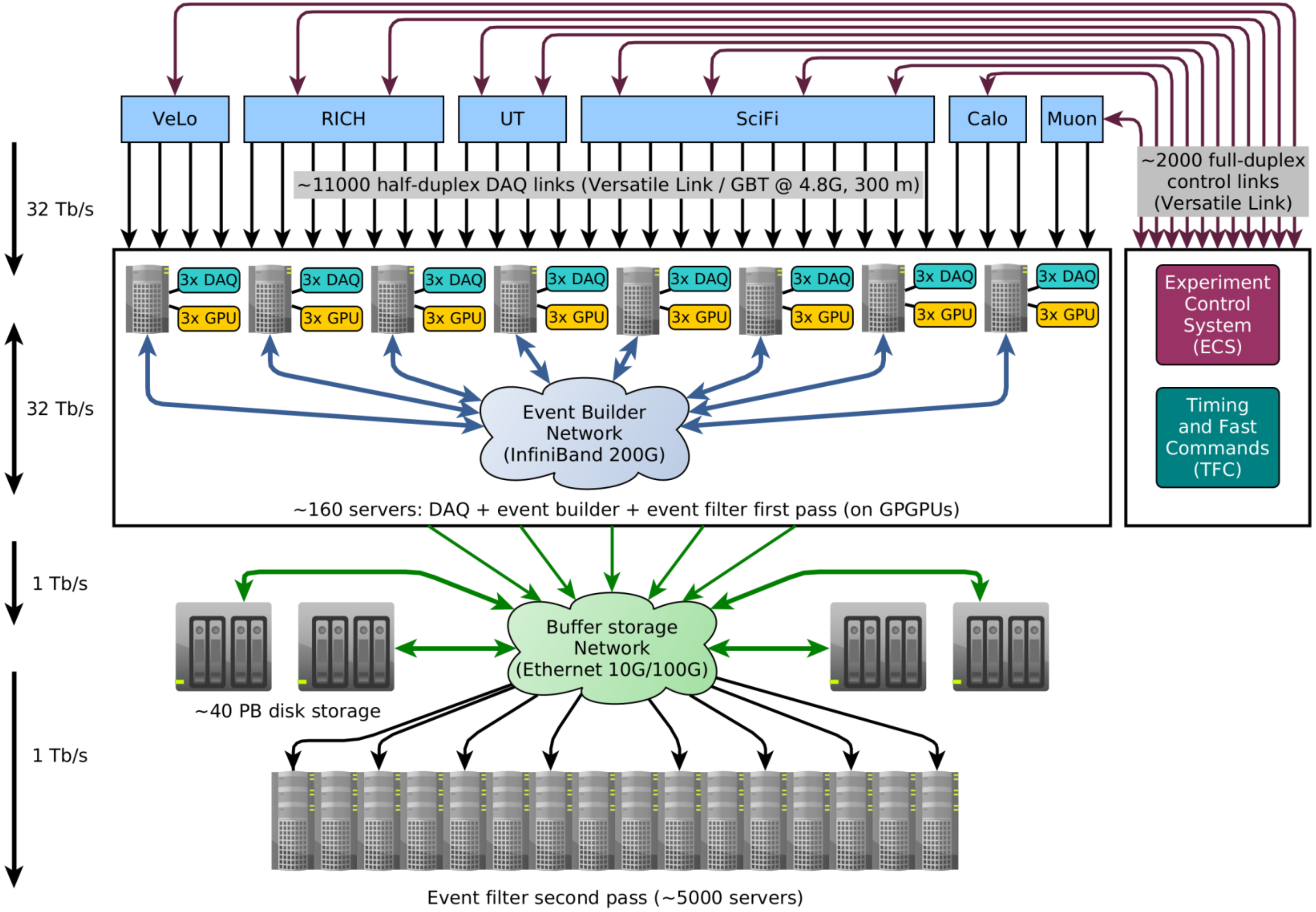
All numbers related to the dataflow are taken from the LHCb

Upgrade Trigger and Online TDR

Upgrade Computing Model TDR

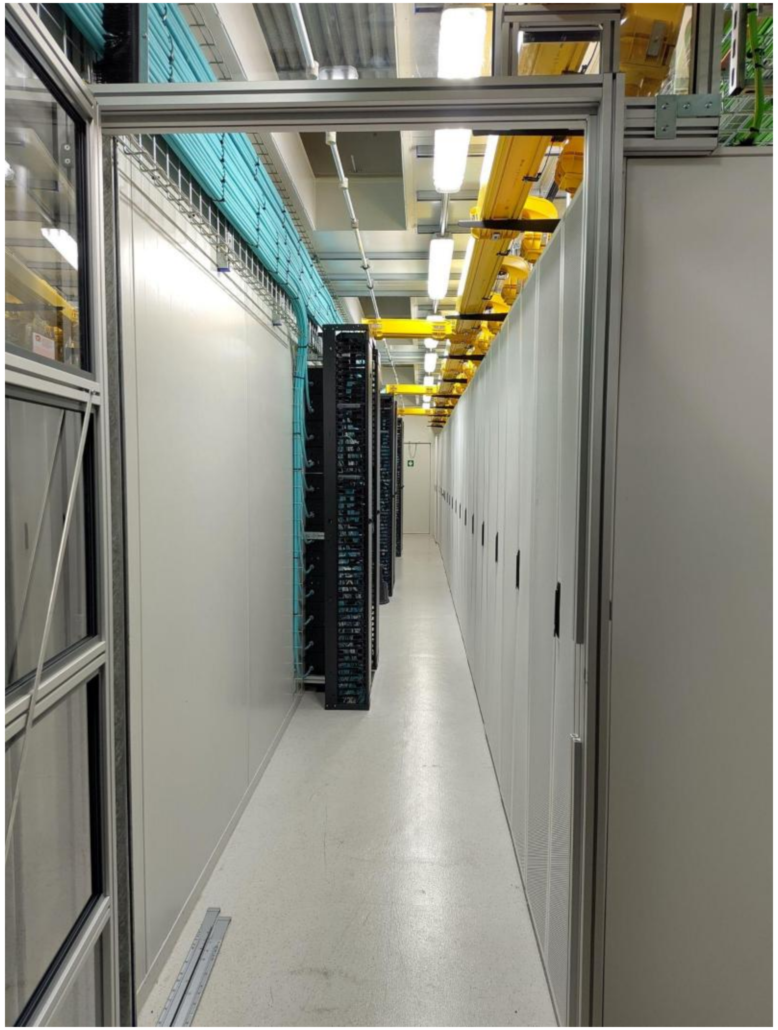
DAQ Architecture in Run3

- Event rate: 30 MHz non-empty bunch crossing
- Event size: ~ 100 kB
- Event Building (EB) **bandwidth: ~ 32 Tbit/s**
- PCIe40 readout boards
- **Dedicated** event builder network InfiniBand 200 Gbit/ s



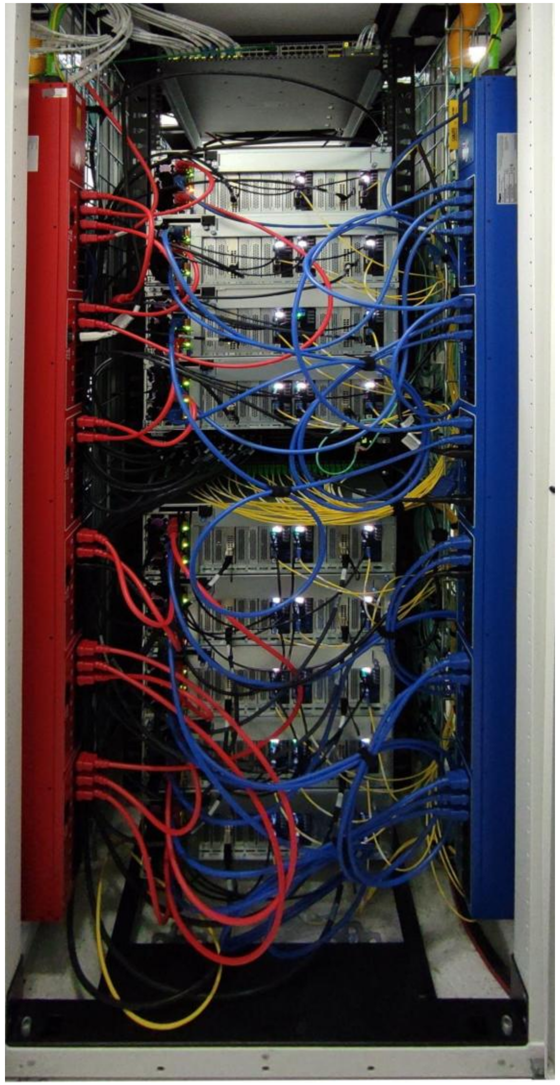
32 Tb/s
32 Tb/s
1 Tb/s
1 Tb/s
VeLo
RICH
UT
SciFi
Calo
Muon
~11000 half-duplex DAQ links (Versatile Link / GBT @ 4.8G, 300 m)
~2000 full-duplex control links (Versatile Link)
3x DAQ
3x GPU
Event Builder Network (InfiniBand 200G)
~160 servers: DAQ + event builder + event filter first pass (on GPGPUs)
Experiment Control System (ECS)
Timing and Fast Commands (TFC)
Buffer storage Network (Ethernet 10G/100G)
~40 PB disk storage
Event filter second pass (~5000 servers)

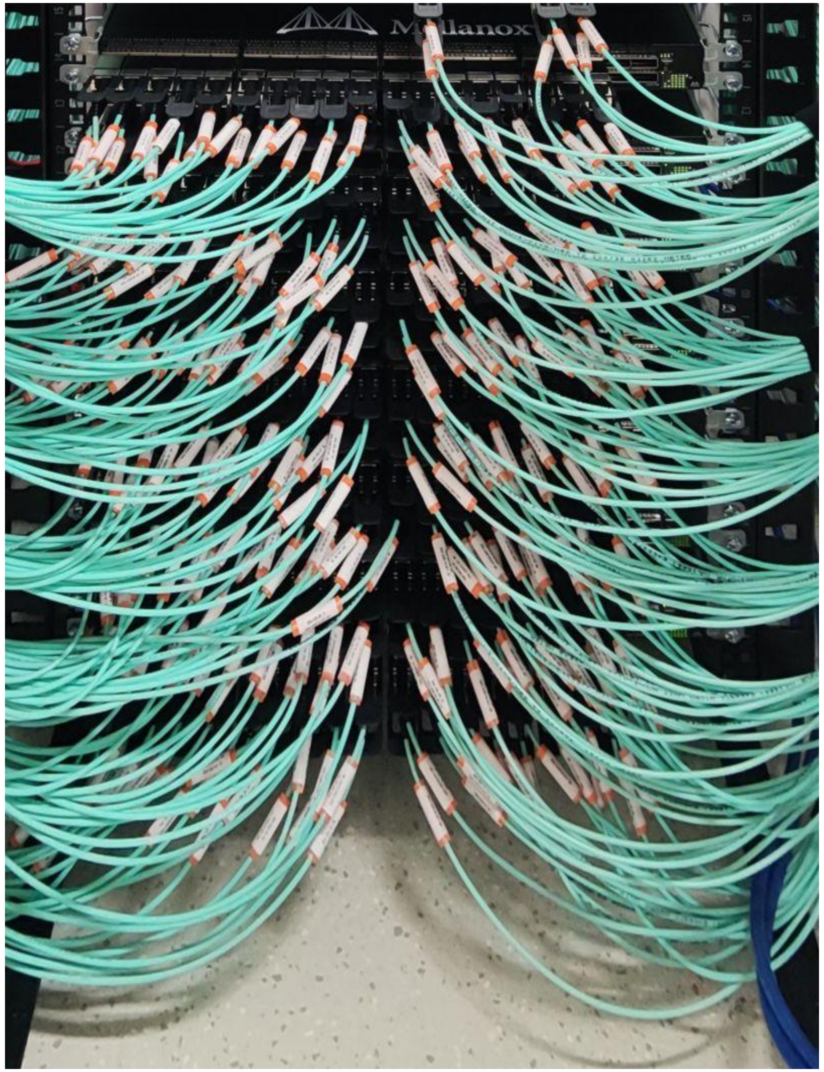
Even Builder Datacenter



Even Builder - Rack

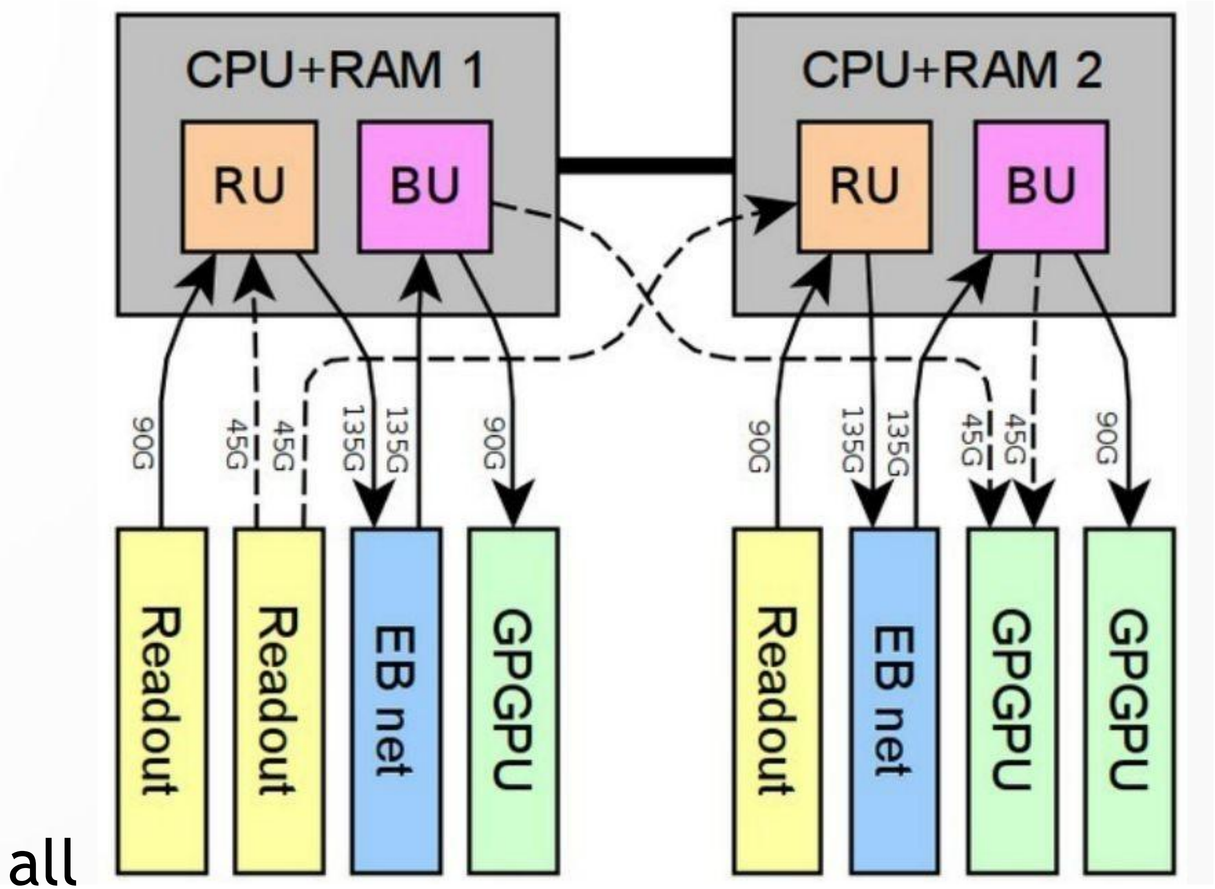
- 164 EB servers
- 24 racks: 18 EB, 2 control, 4 storage
- 28 40-port IB HDR switches: 18 leaf and 10 spine





Event Builder Server

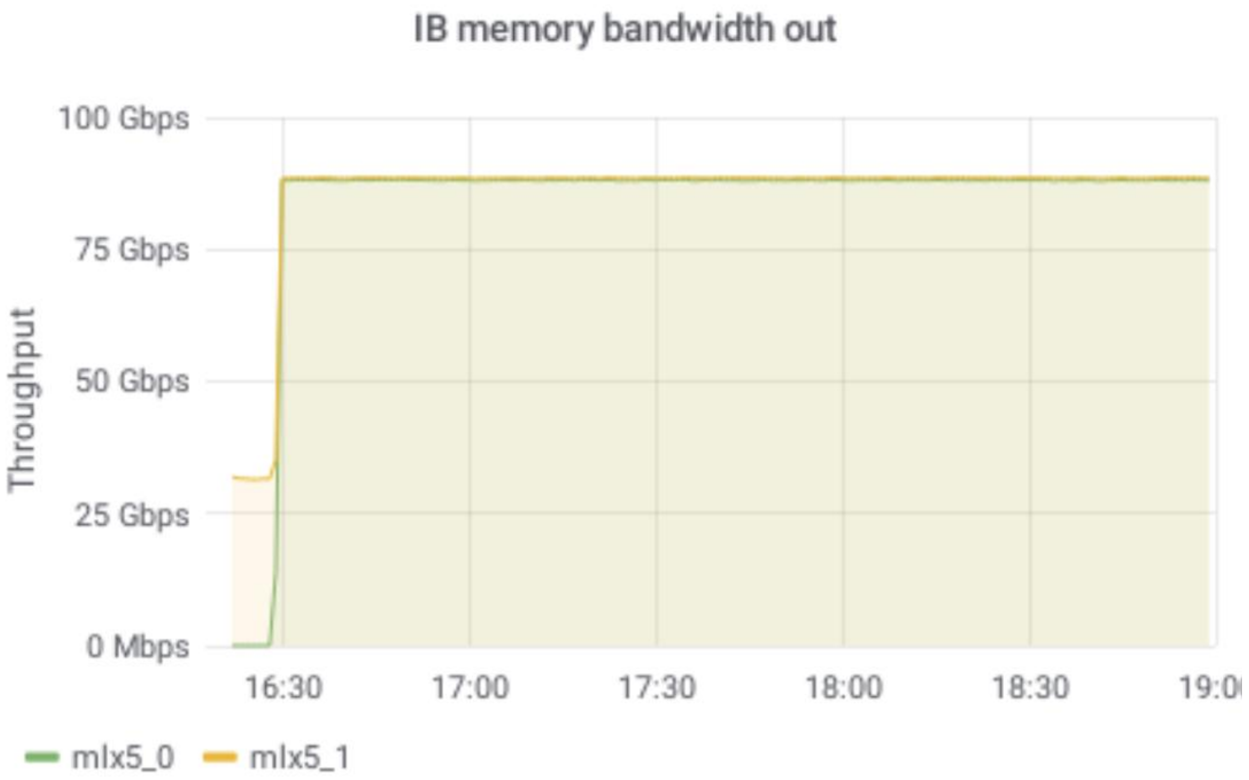
- **CPU:** 2 AMD EPYC 7502, 32 cores
- **RAM:** 512 GB DDR4
- 8x PCIe Gen4 x16 slots :
  - 1-3 GPGPU (RTX A5000)
  - 3 Readout Boards (TELL 40)
  - 2 InfiniBand HRD NICs (200 Gb/s)
- Every Readout Unit (RU) receives a fragment of the event
- Every Builder Unit (BU) has to gather all the fragments of the event

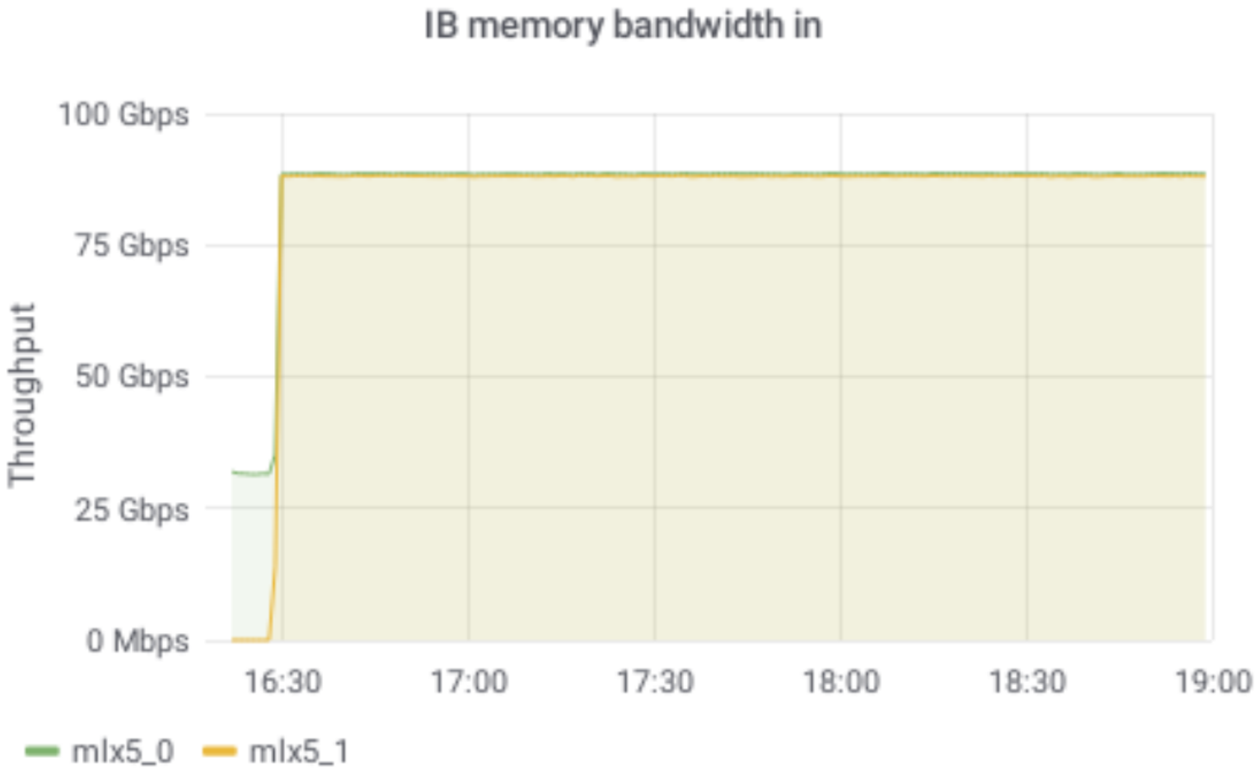


PCIe40 Readout Board

- Intel Arria10 FPGA
- PCIe Gen3 x16
- 48x10G capable transceiver on 8xMPO for up to 48 full-duplex Versatile Links
- First pre-processing of the data on board

Server Performance

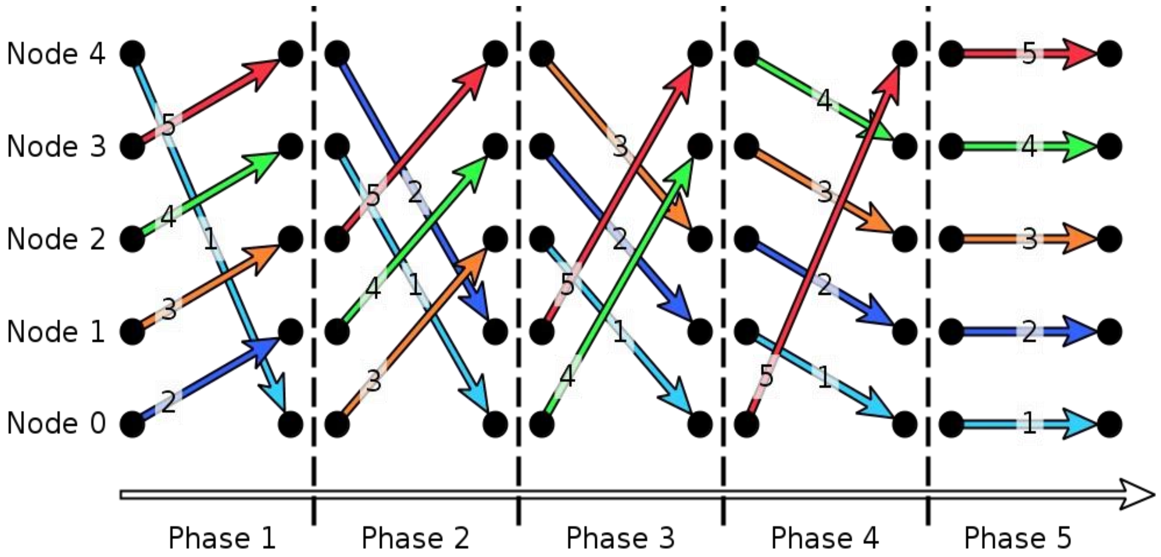


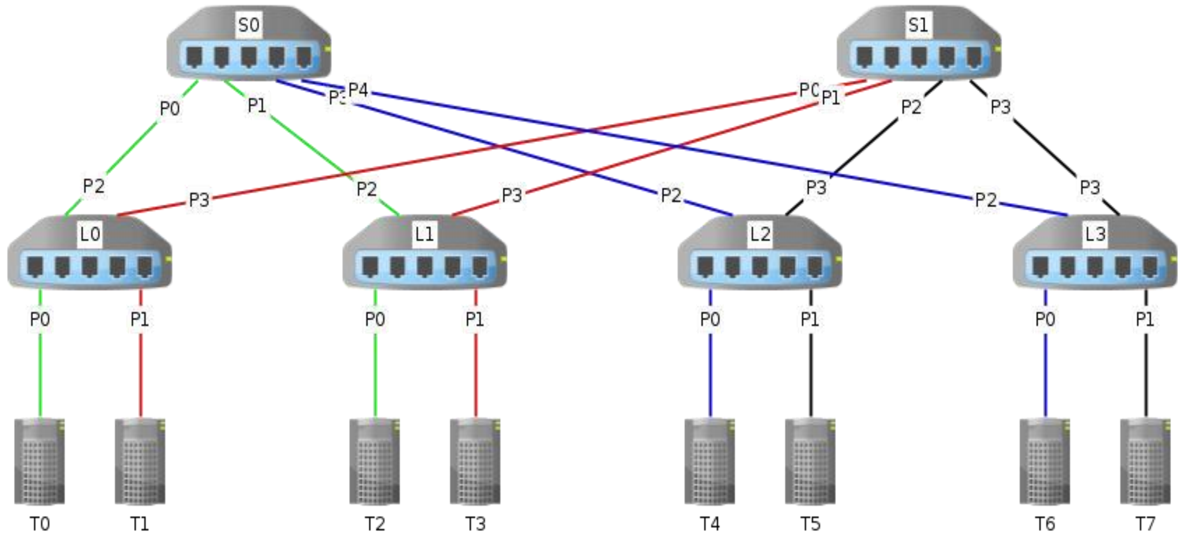


Server： 2X GPUs, 2X InfiniBand 100G cards and 2X TELL40 cards.

[LHCB-FIGURE-2019-009]

Event Builder: Traffic scheduling

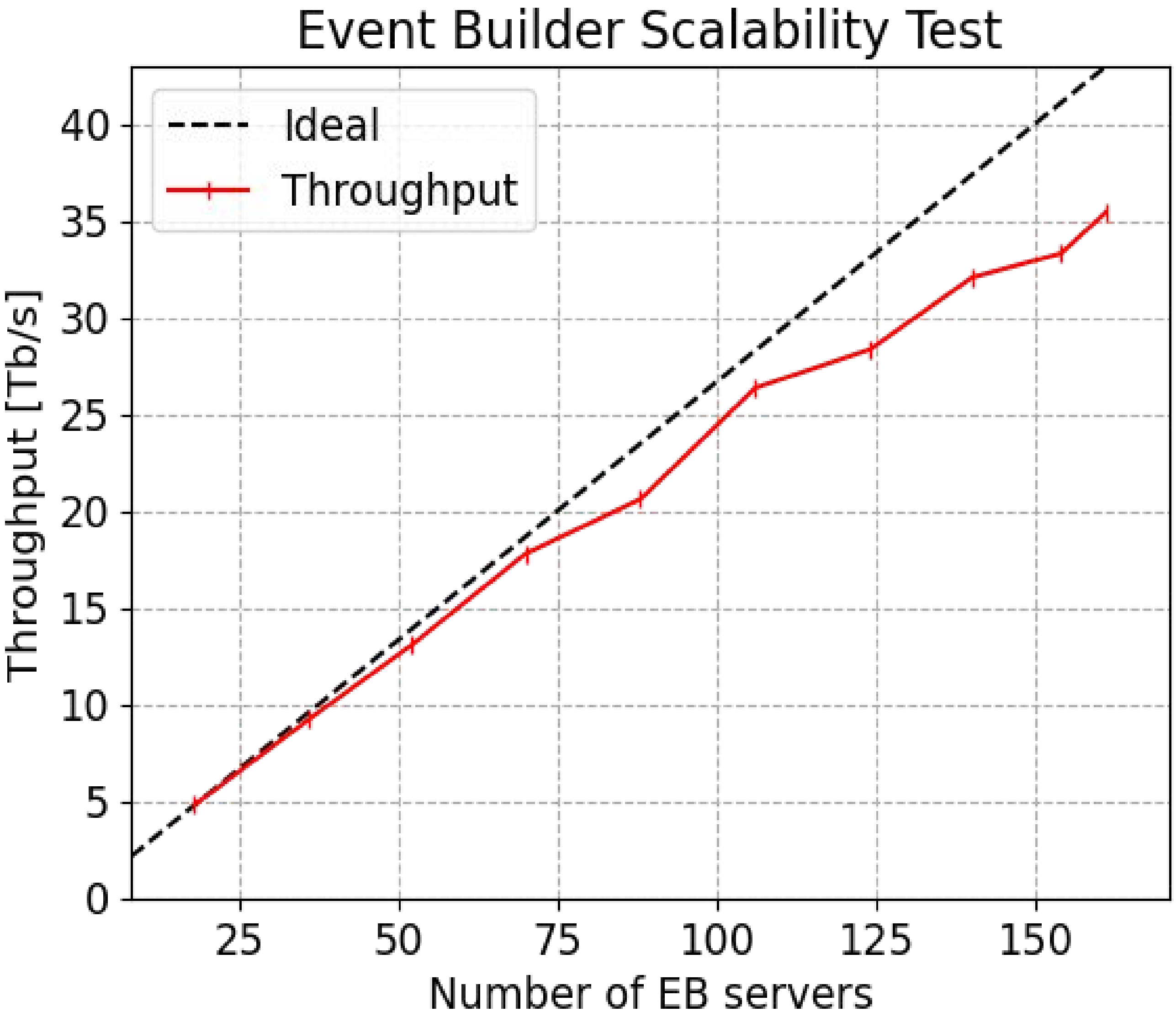




- The processing of N events is divided into N phases (N is the number of EB nodes)
- In every phase one RU sends data to one BU, and every BU receives data from one RU
- During phase n RU x sends data to BU (x + n)%N

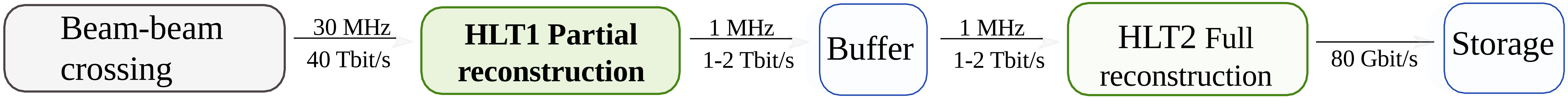
Congestion-free traffic on “selected networks” (e.g. fat-tree networks)

System scalability



[One year of LHCb triggerless DAQ]

Trigger in Run3



Beam-beam crossing

30 MHz
40 Tbit/s

HLT1 Partial reconstruction

1 MHz
1-2 Tbit/s

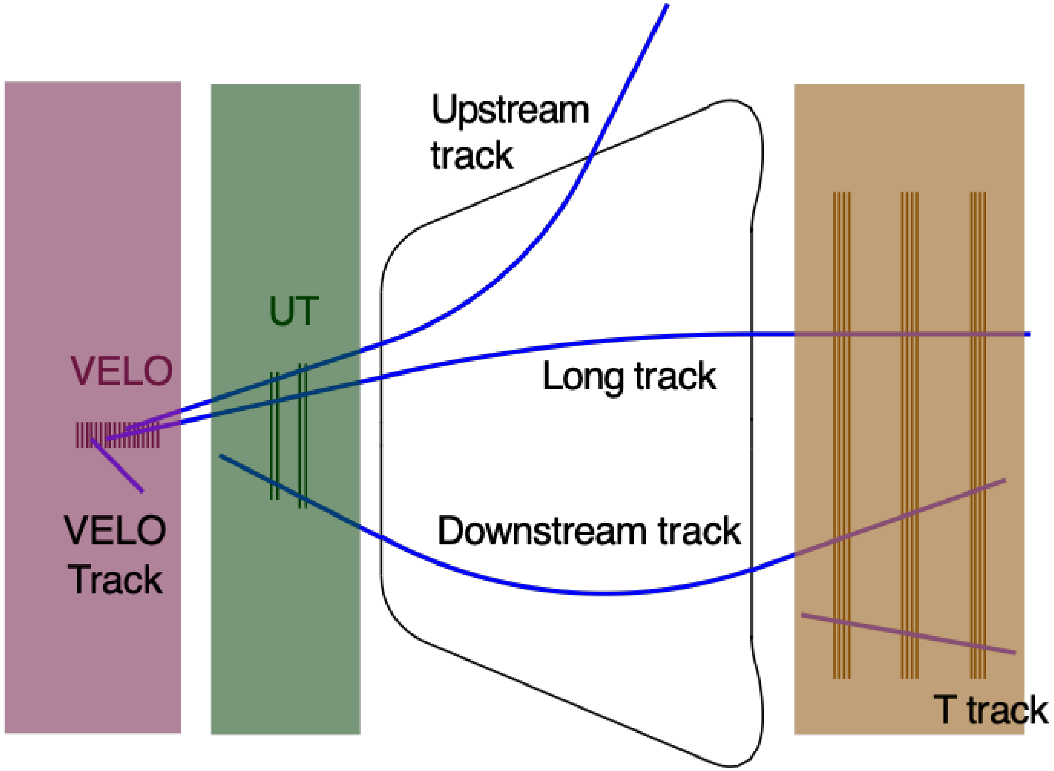
Buffer

1 MHz
1-2 Tbit/s

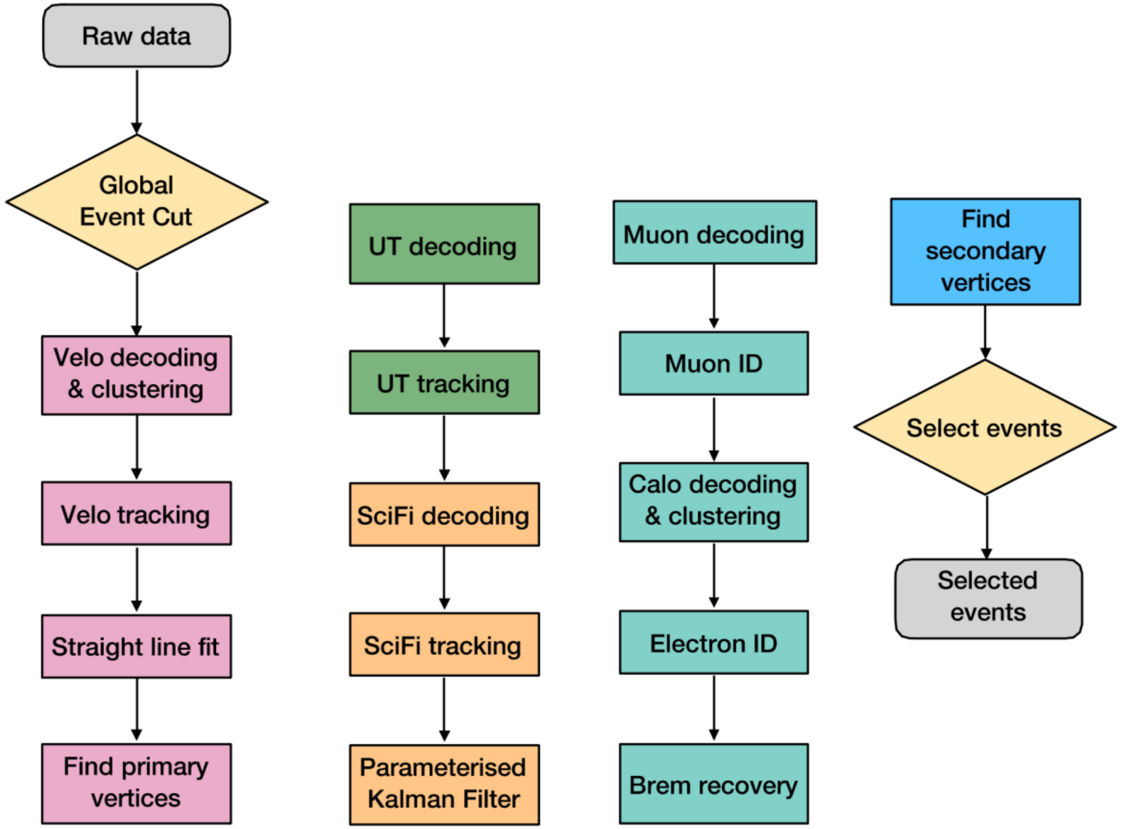
HLT2 Full reconstruction

80 Gbit/s

Storage



[Comput Softw Big Sci 4, 7 (2020)]



Raw data

Global Event Cut

Velo decoding & clustering

Velo tracking

Straight line fit

Find primary vertices

UT decoding

UT tracking

SciFi decoding

SciFi tracking

Parameterised Kalman Filter

Muon decoding

Muon ID

Calo decoding & clustering

Electron ID

Brem recovery

Find secondary vertices

Select events

Selected events

Allen project: HLT1 on GPU

- Framework for GPU-based execution of an algorithm sequence

https://gitlab.cern.ch/lhcb/Allen

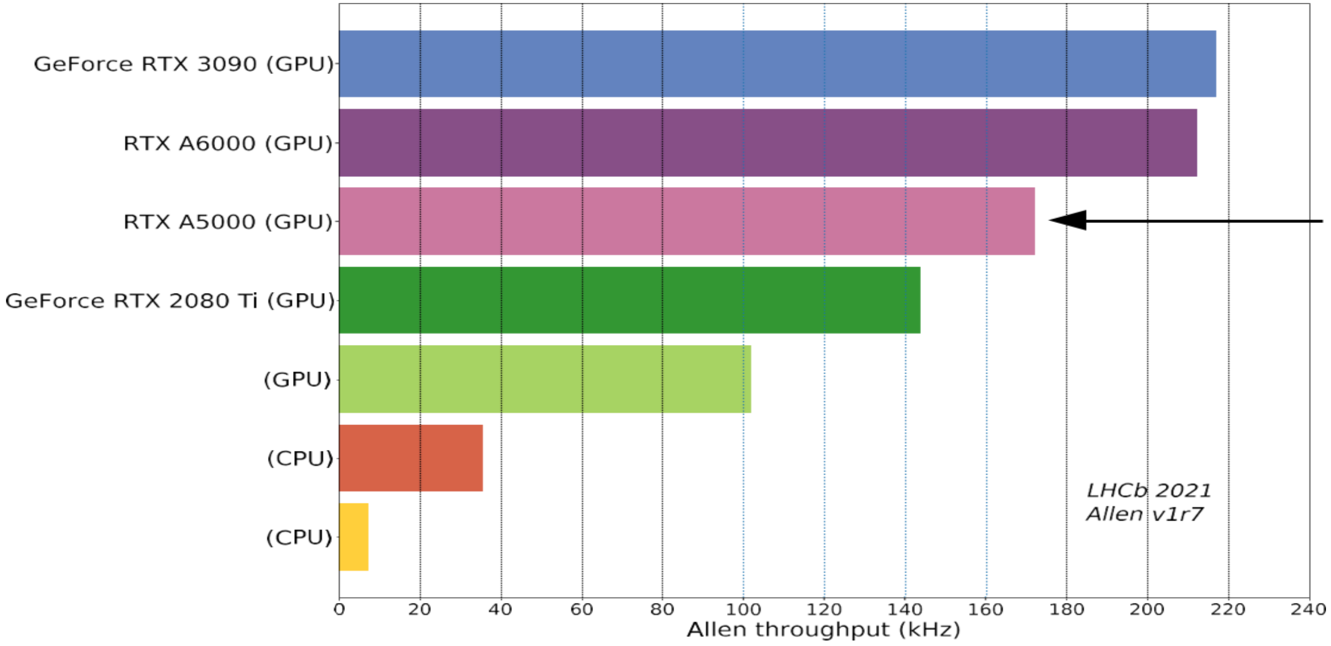
- Cross-architecture compatibility:

  CPU, NVidia GPU (CUDA), AMD GPU (HIP)

- Algorithm sequences defined in python, generated at runtime
- Only single precision is used
- Thousands of events are processed in parallel

HLT1 computing throughput

- Putting GPUs in EB free PCIe slots to reduce cost
- 30 MHz goal can be achieved with ~340 GPUs (maximum the Event Builder server can host is 500)
- Throughput scales well with theoretical TFLOPS of GPU card

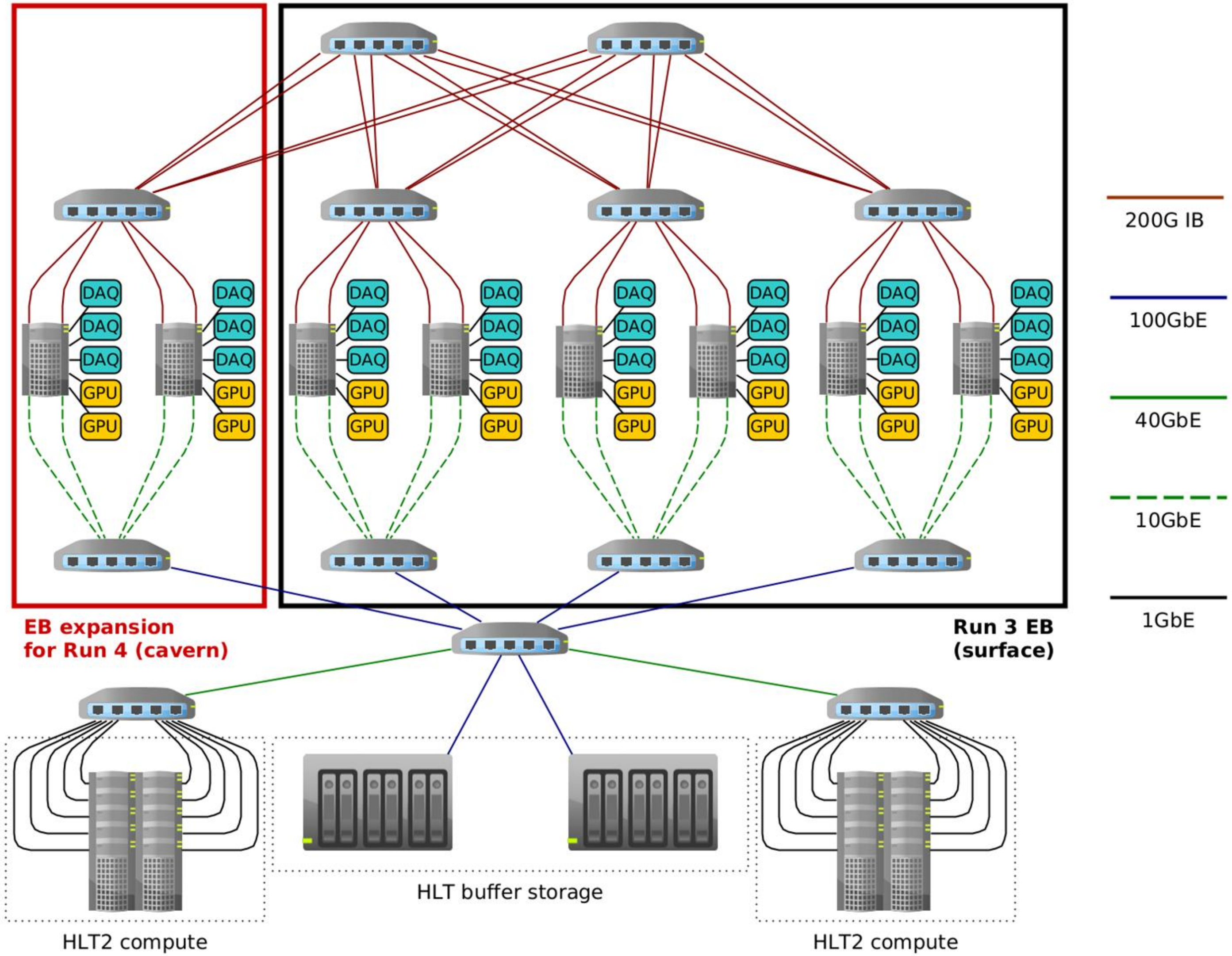


Chose RTX A5000 for the beginning of Run3

[LHCb-FIGURE-2020-014]

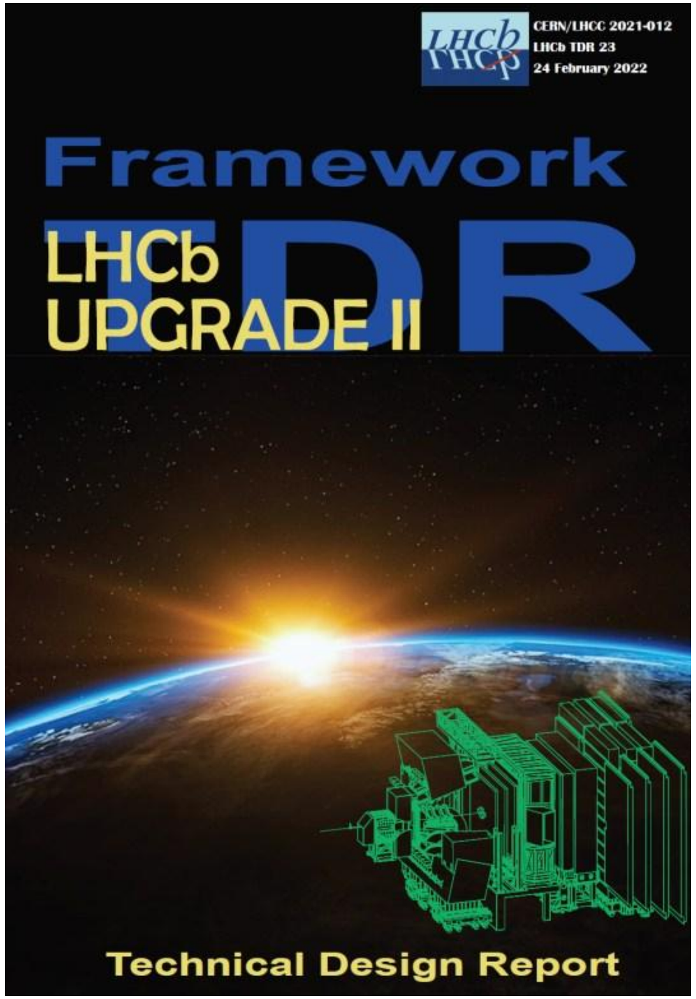
DAQ in Run4

- Target luminosity and overall trigger strategy unchanged
- Upgraded RICH will use lpGBT
- lpGBT cannot reach the surface data center
  - ➢ EB expansion in cavern



Upgrade II - Run5

- LHCb Trigger evolution: Run 5
  - Luminosity: $2 \cdot 10^{33} \rightarrow 1.5 \cdot 10^{34}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$
  - Pileup: 5 → 40
- Keep 40 MHz readout
  - But event size increase (pileup, upgraded detectors)
    → up to 200 Tbps DAQ
- More accelerated reconstruction in HLT
  - GPU in HLT2 reconstruction, clearly
  - Evaluating RETINA or other FPGA-based fast tracking in the whole detector upstream to HLT1
- big saving on buffer and HLT2 needs



CERN/LHCC 2021-012
LHCb TDR 23
24 February 2022

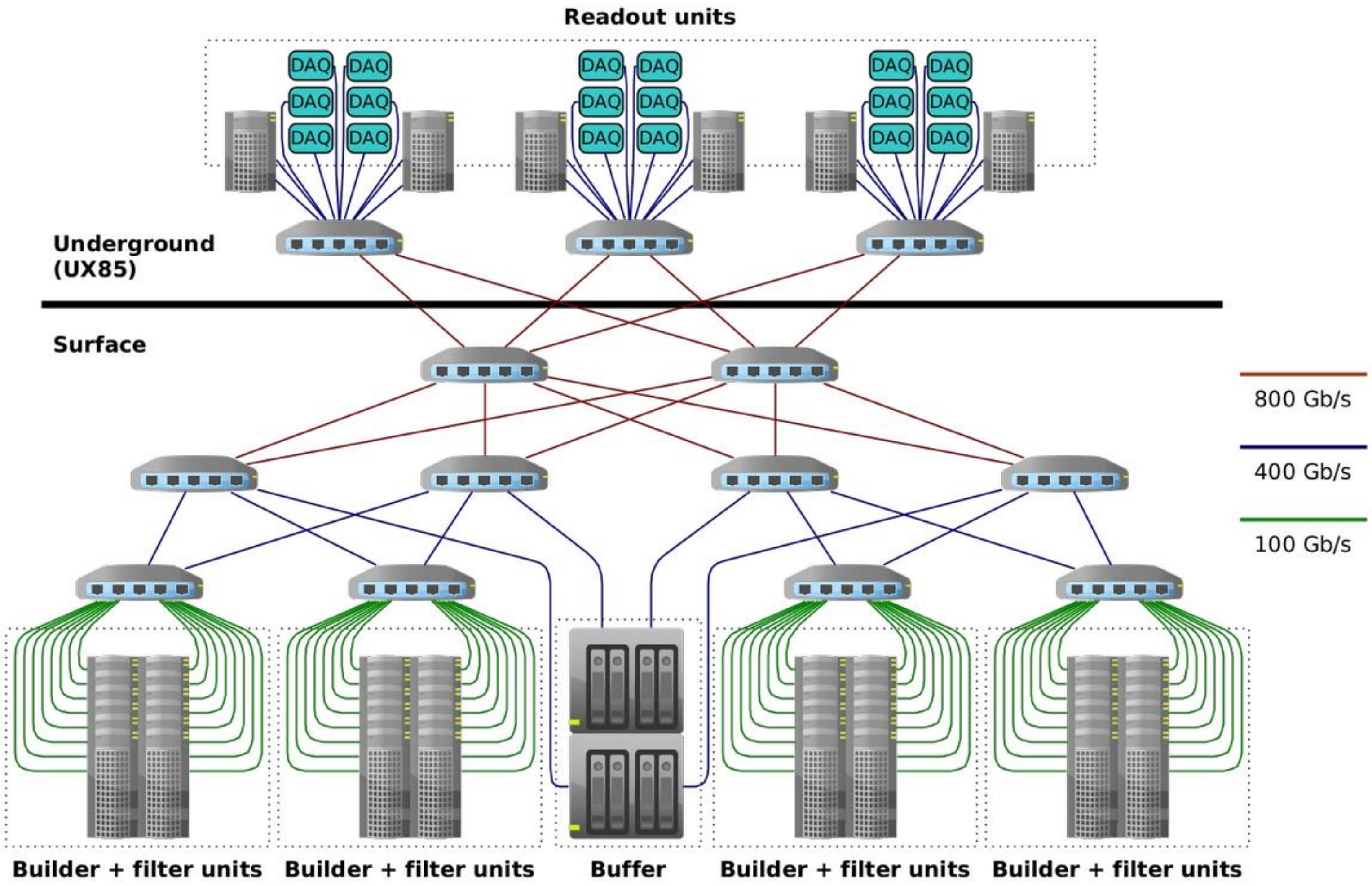
Framework TDR
LHCb UPGRADE II

Technical Design Report

DAQ in Run 5?

Network-attached DAQ:

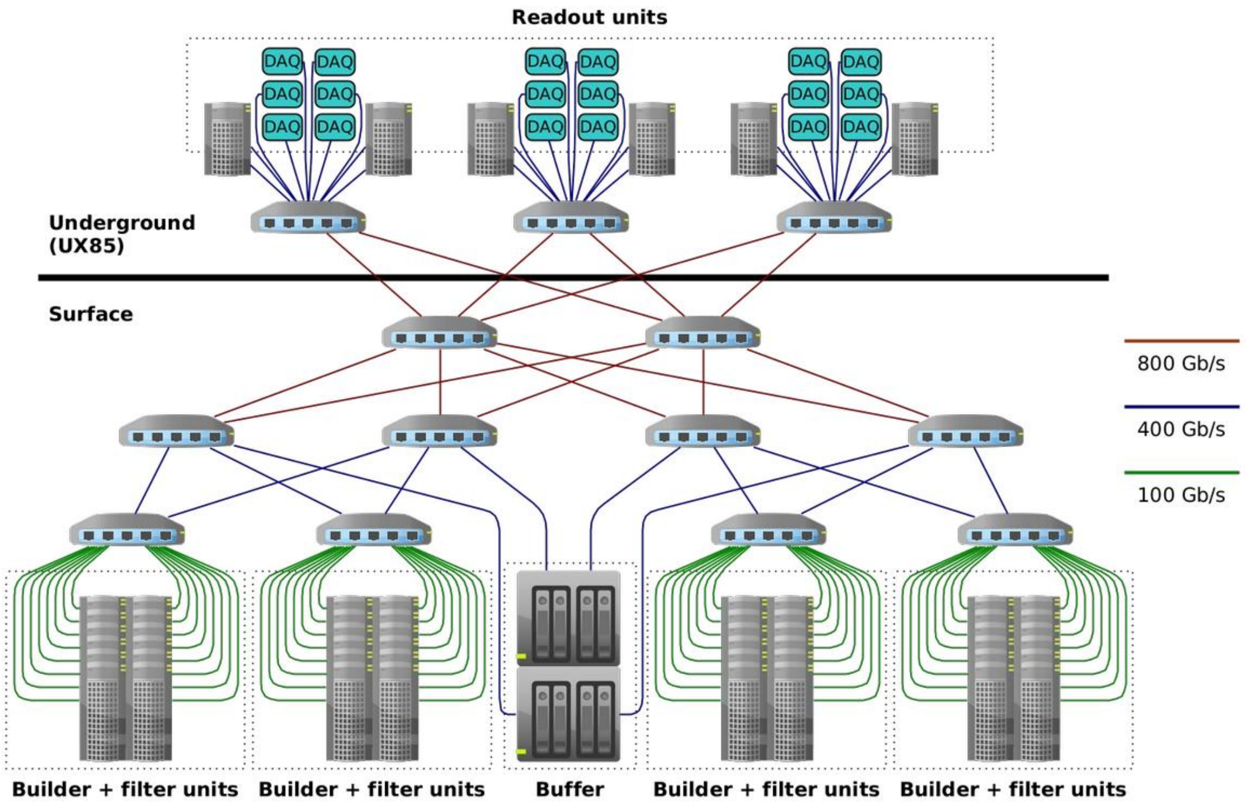
- DAQ boards are stand-alone entities
- attached to the network
- independent of the readout servers



DAQ in Run 5?

Flexible design adaptable to market offer in LS4:

- Compute can take any form factor, as long as it has a network input
- Some data aggregation to be done in network instead of FPGA
- Exploits “cheap” networking



DAQ in Run 5?

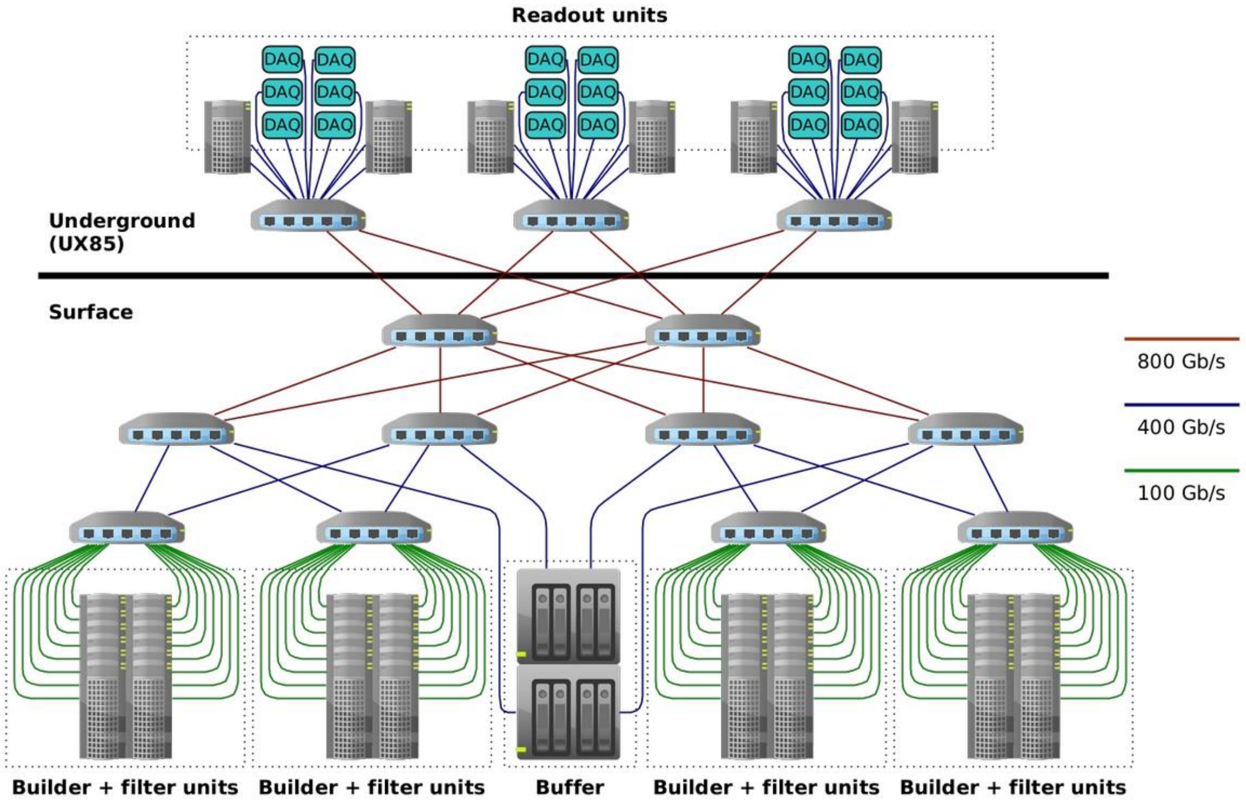
- Less resource efficient:

Network links are used mostly “half-duplex”

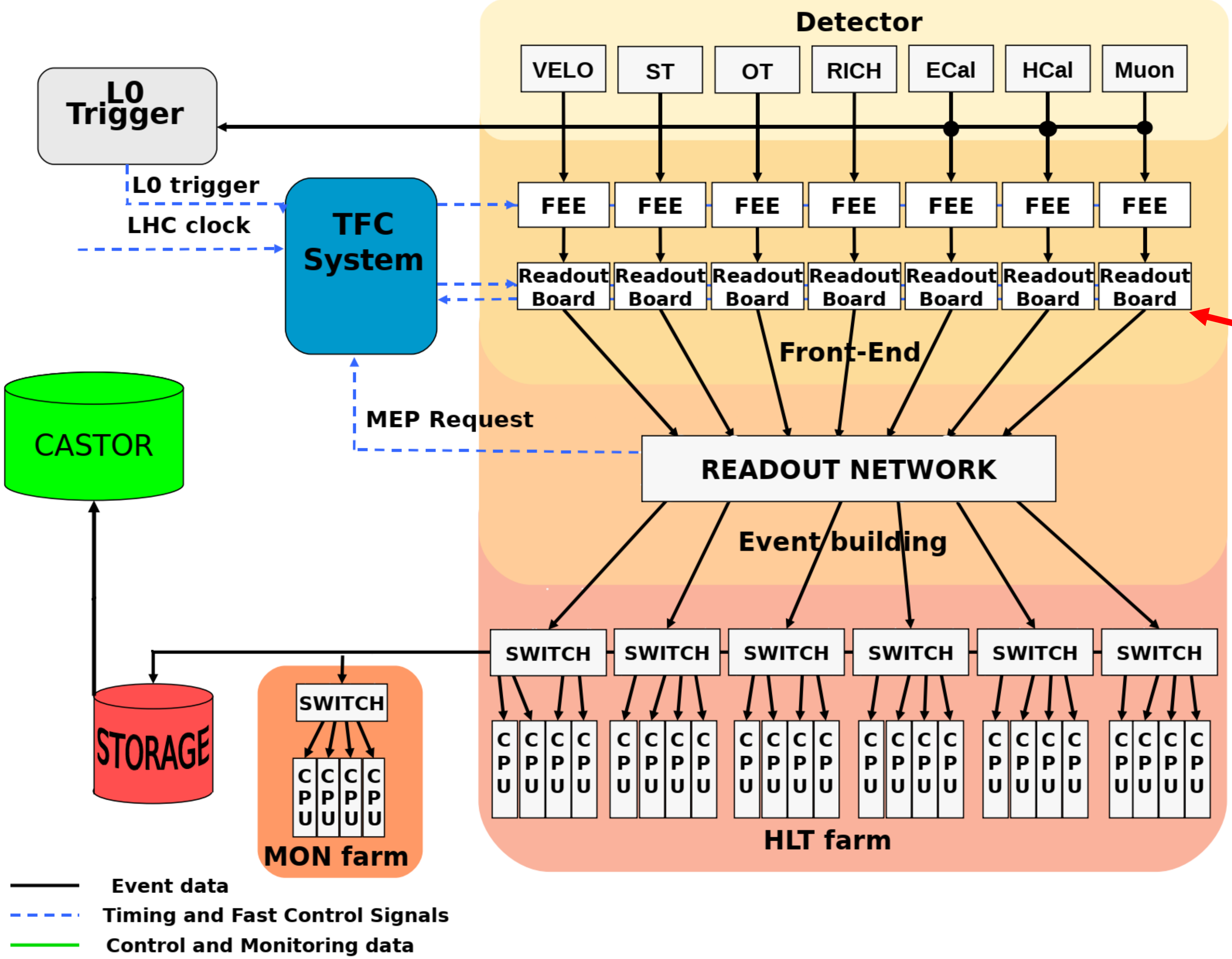
- Event building is “distributed” all over the compute servers

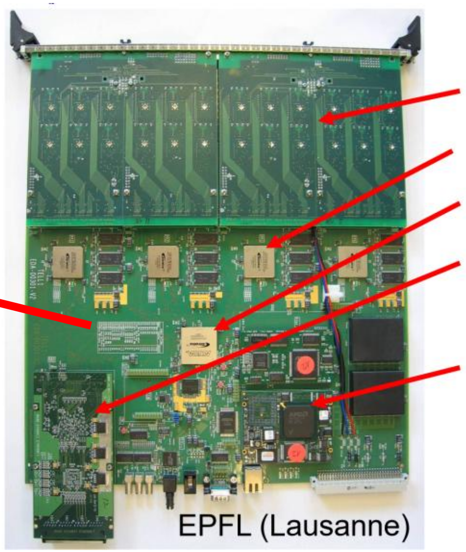
  - network optimization needed

- Might prevent using advanced RDMA on builder units



Back to Run 1&2?





Optical receiver card

Processing FPGAs

SyncLink FPGA

Quad Gigabit ethernet transmitter mezzanine

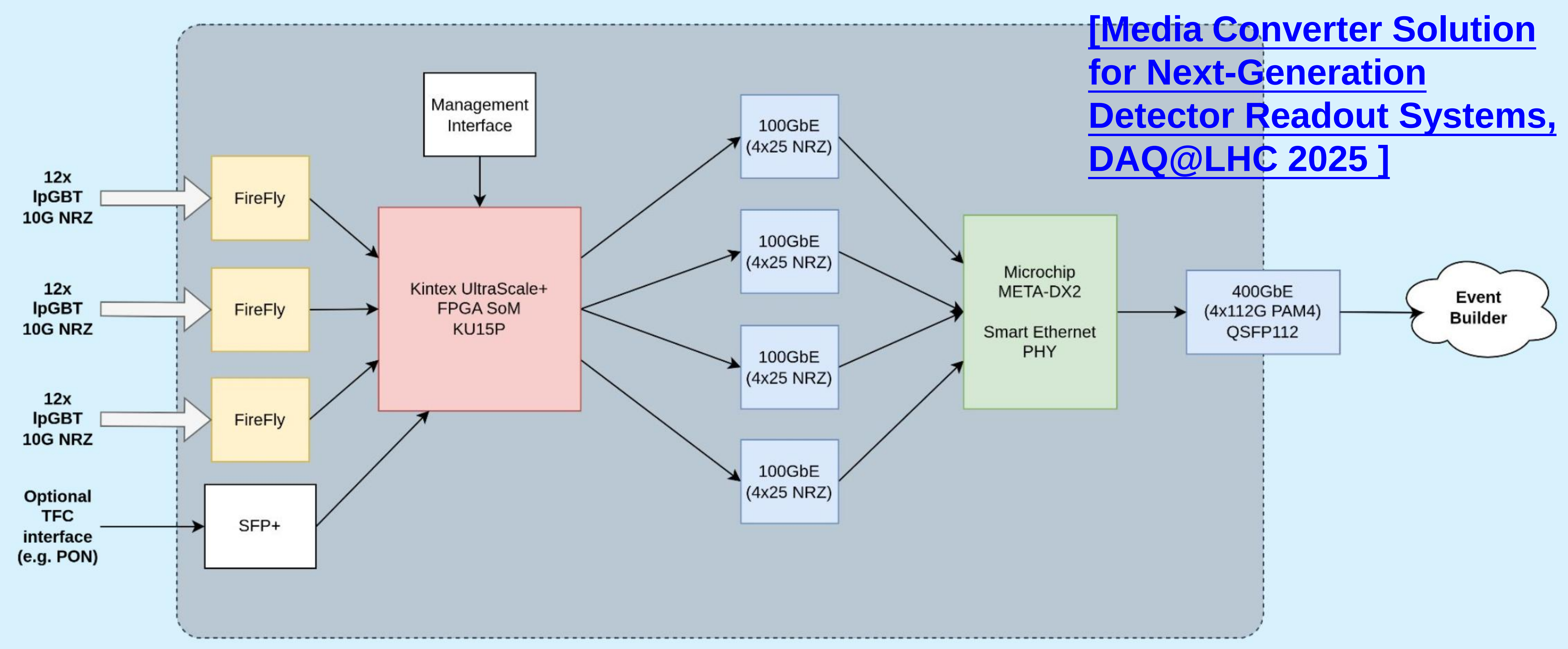
Credit Card PC

1 MHz readout to PC farm LHCb wide

LHCb DAQ architecture

DAQ Board in Run 5？

A proposed Solution: NetGBT



[Media Converter Solution for Next-Generation Detector Readout Systems, DAQ@LHC 2025]

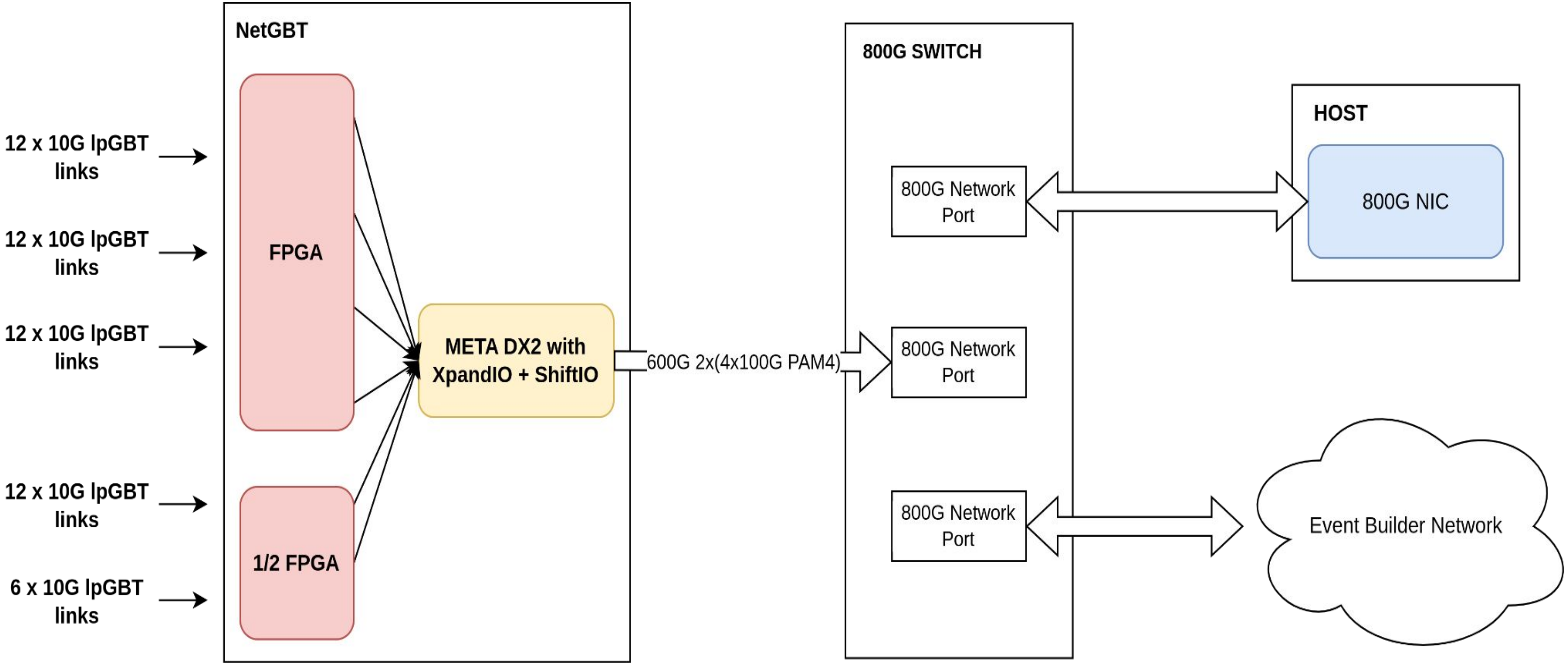
DAQ Board in Run 5？

A proposed Solution: NetGBT

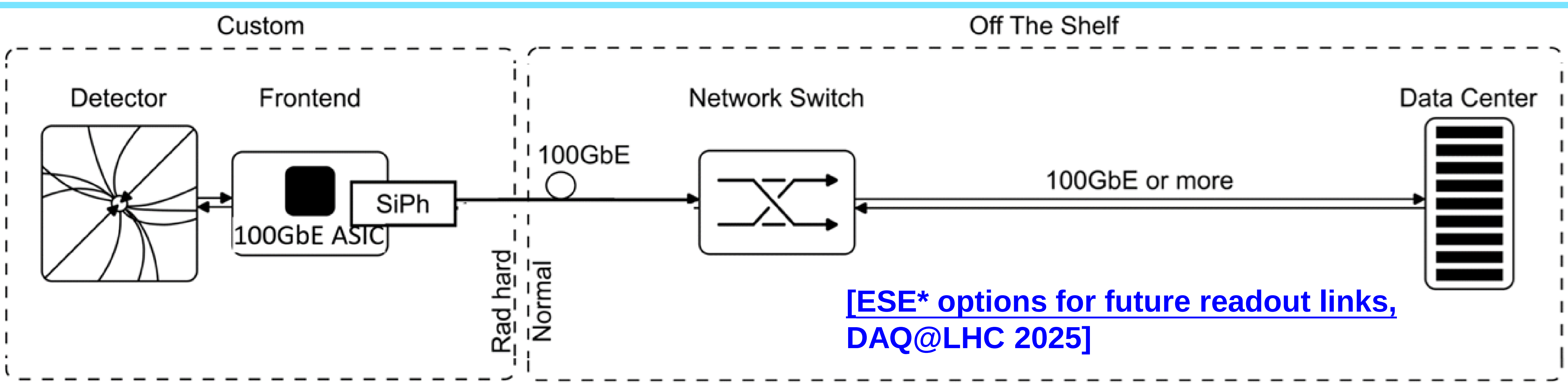
- Media converter based on a mid-end FPGA with many transceivers available.
- Interfacing with a number of lpGBT links and convert them into UDP/IP packets.
- A buffering layer should be added between the NetGBT network and the Event Builder network.
  - Dual Layer Switched Network
  - Host-Direct
  - Smart PHY Switched (SFP+, QSFP+ …)

DAQ Board in Run 5？

Buffering layer： Smart PHY



R&D on links: 100GbE@FE



[ESE* options for future readout links, DAQ@LHC 2025]

- No backend-approach: proof of concept validated
  - ➢ Unidirectionality
  - ➢ COTS compatibility
  - ➢ Ethernet frame compatible with SEU induced burst errors

Summary

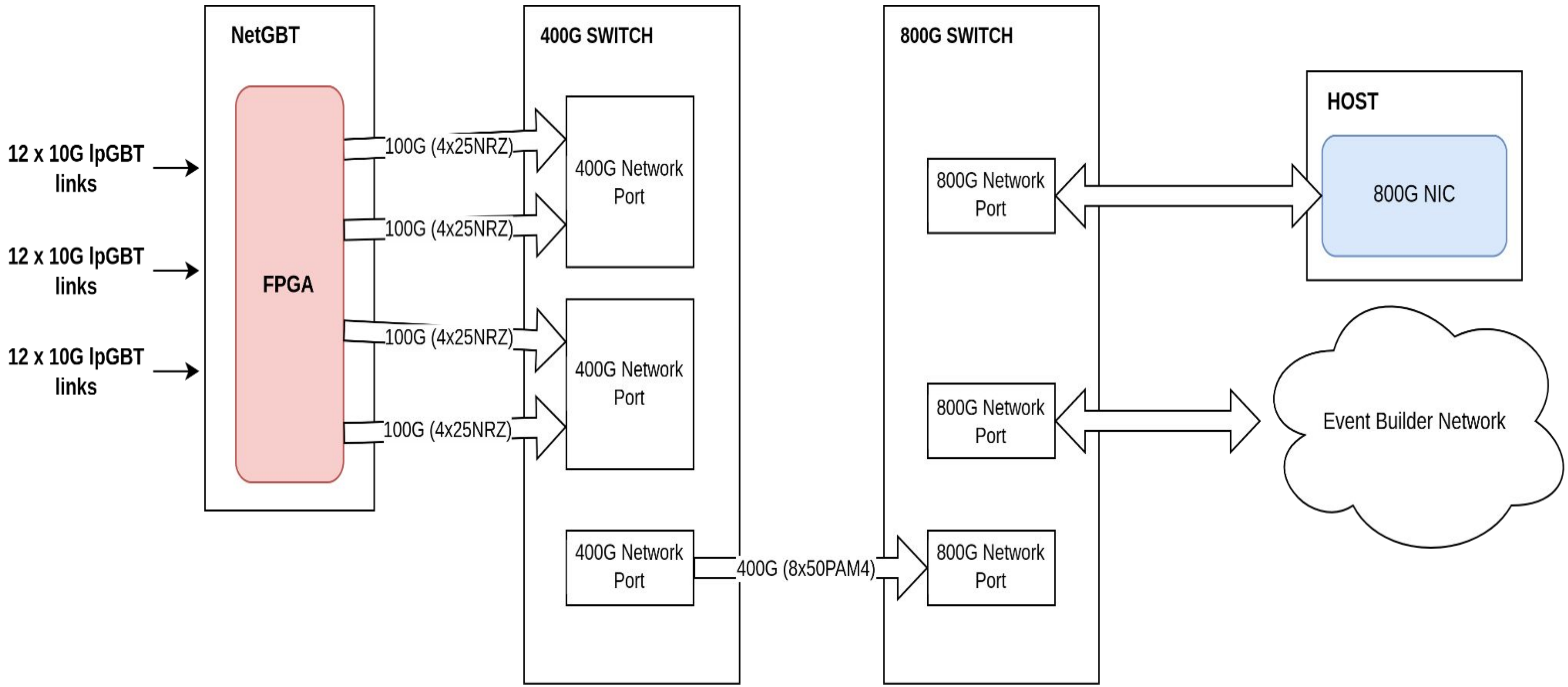
- The LHCb TDAQ system have been running successfully with the load of 32 Tb/s in Run 3
- Flexibility and scalability are essential
- Run 4 will keep overall TDAQ architecture unchanged hopefully
- R&D for Run 5 ongoing: follow the latest developments of HPC/AI, Ethernet all the way, …

Thanks for your attentions!

Backup Slides

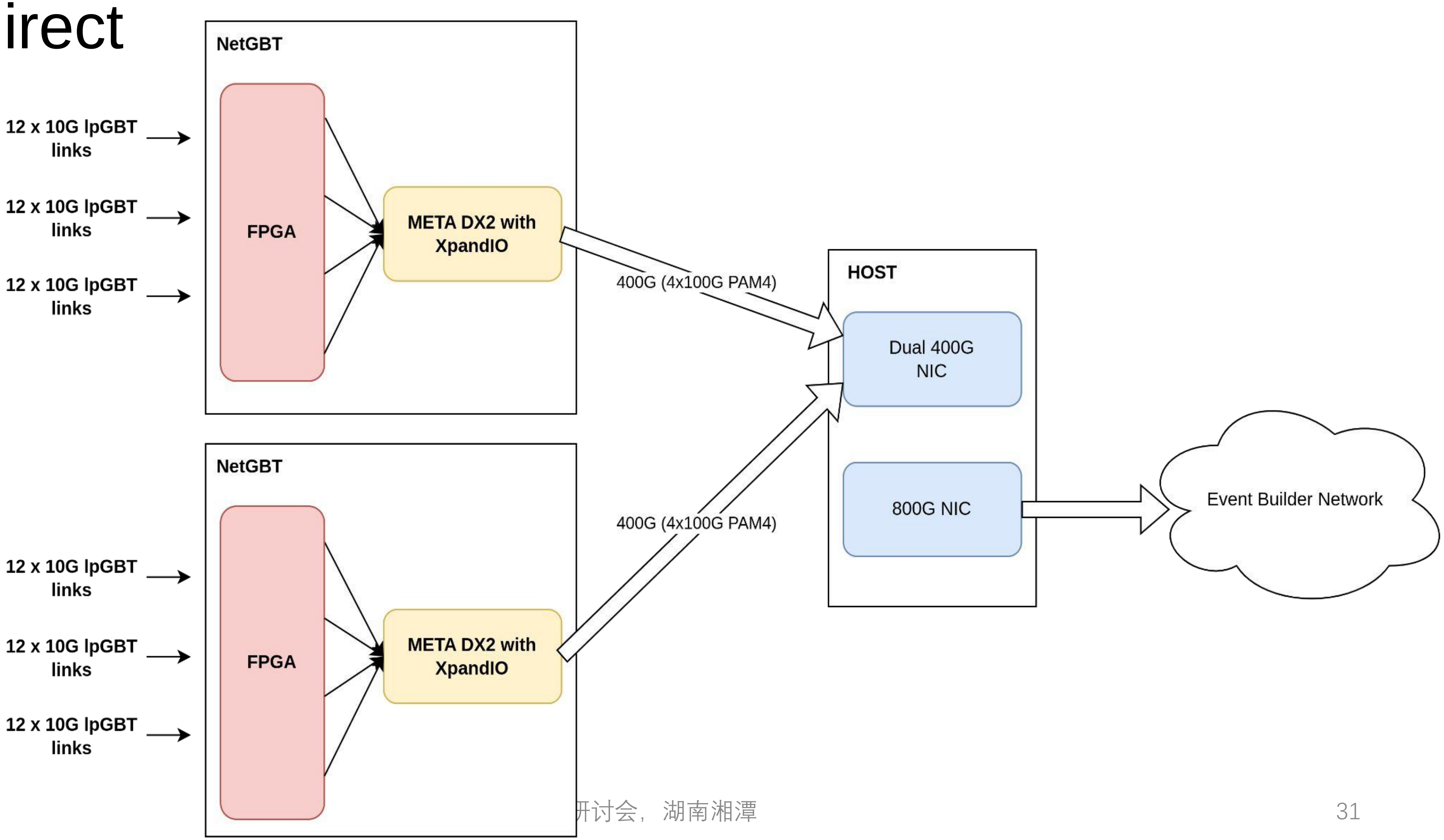
NetGBT

- Dual Layer Switched Network



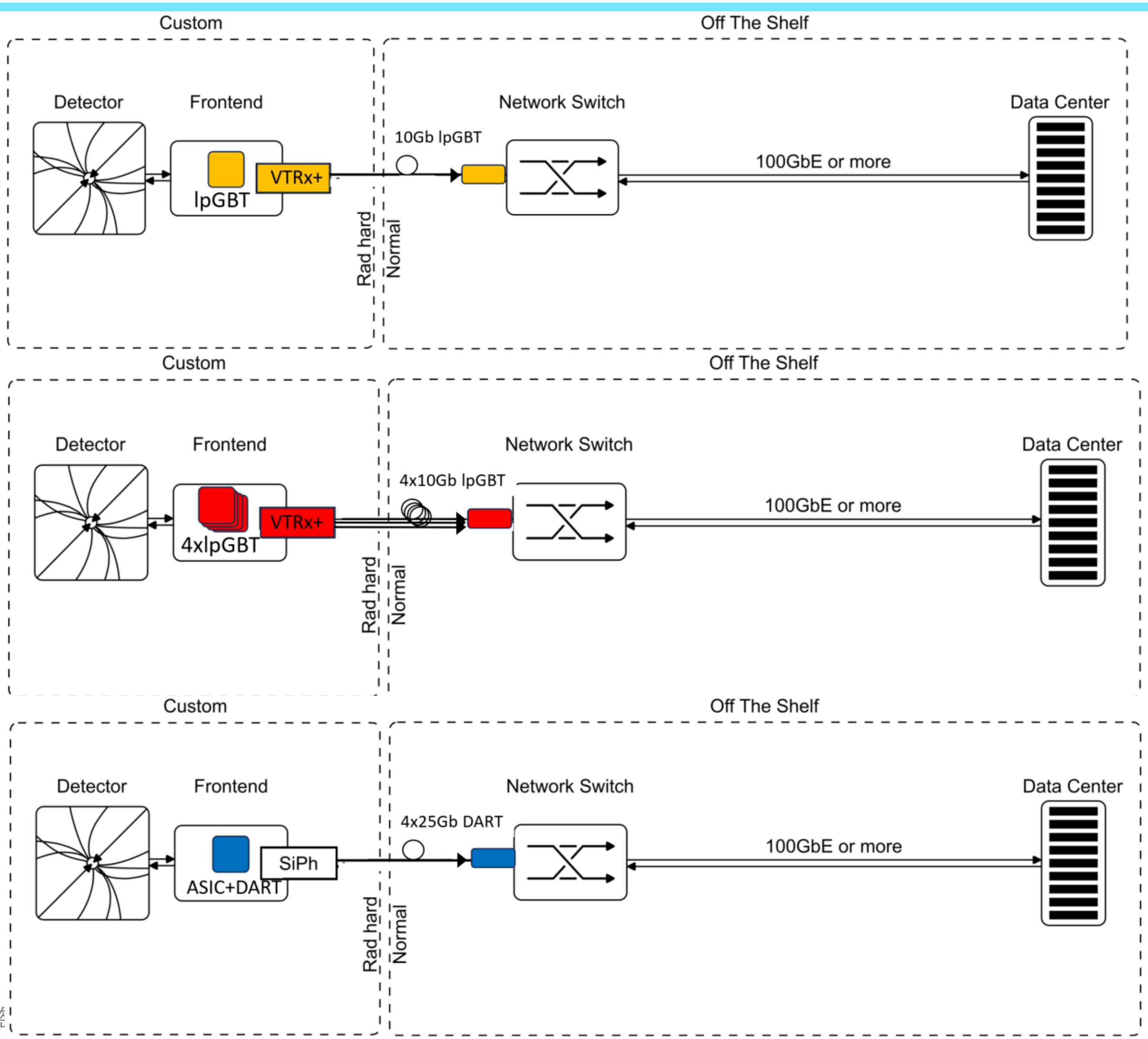
NetGBT

- Host-direct



R&D on links: 100GbE@FE

- Intermediate approach: Translator modules
  - SFP+ (10G lpGBT -> 10GbE) in test
  - QSFP (4x10G/VL+ -> 40GbE) module designed
  - CWDM (4x25G DART -> 100GbE) module : feasibility under study


Custom
Off The Shelf
Detector
Frontend
Network Switch
Data Center
10Gb lpGBT
lpGBT
VTRx+
100GbE or more
Rad hard
Normal
Custom
Off The Shelf
Detector
Frontend
Network Switch
Data Center
4x10Gb lpGBT
4xlpGBT
VTRx+
100GbE or more
Rad hard
Normal
Custom
Off The Shelf
Detector
Frontend
Network Switch
Data Center
4x25Gb DART
ASIC+DART
SiPh
100GbE or more
Rad hard
Normal