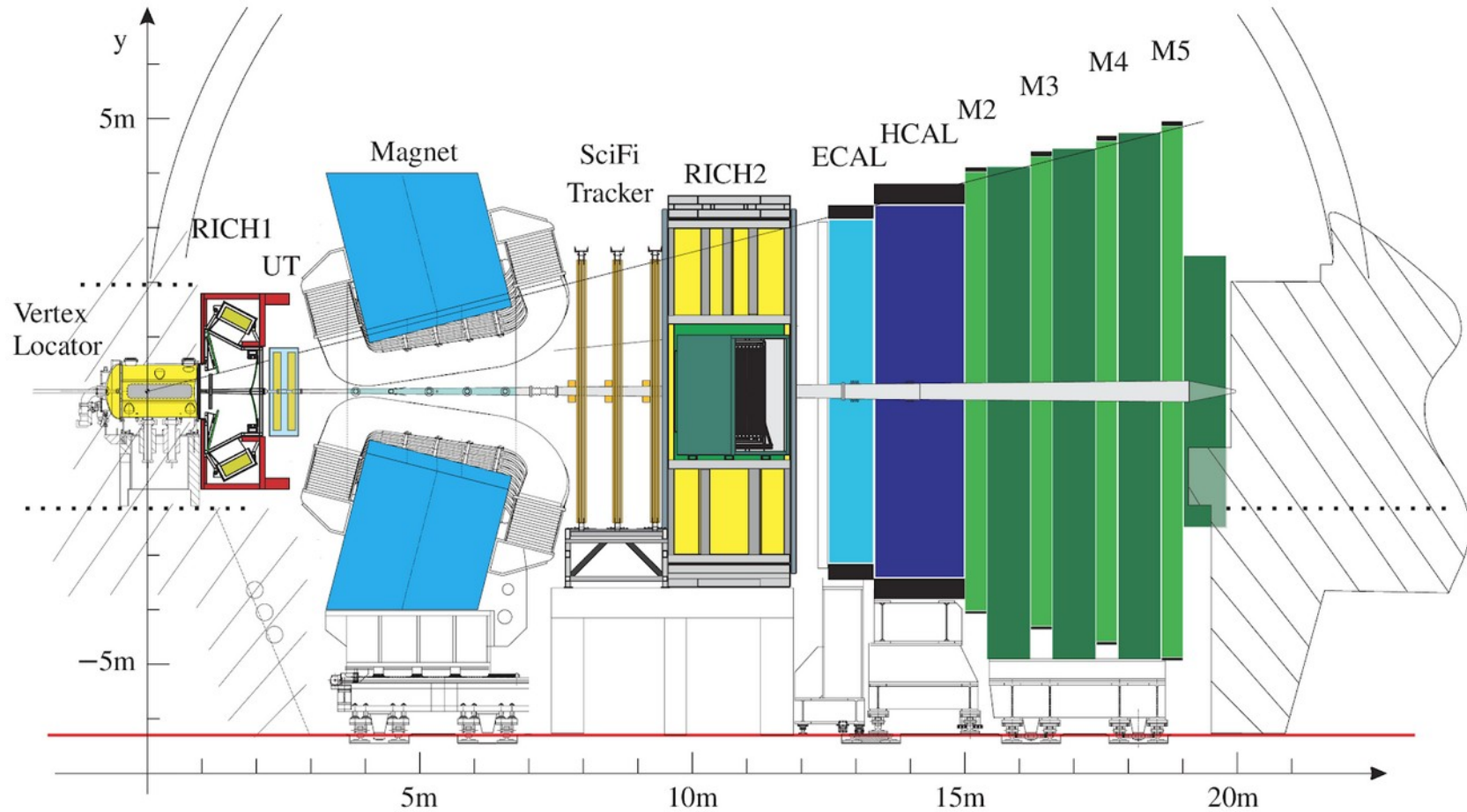


LHCb Run3 DAQ & event filter

Tommaso Colombo
on behalf of the LHCb Onliners

4th FCC Physics and Experiments Workshop
CERN
12 November 2020

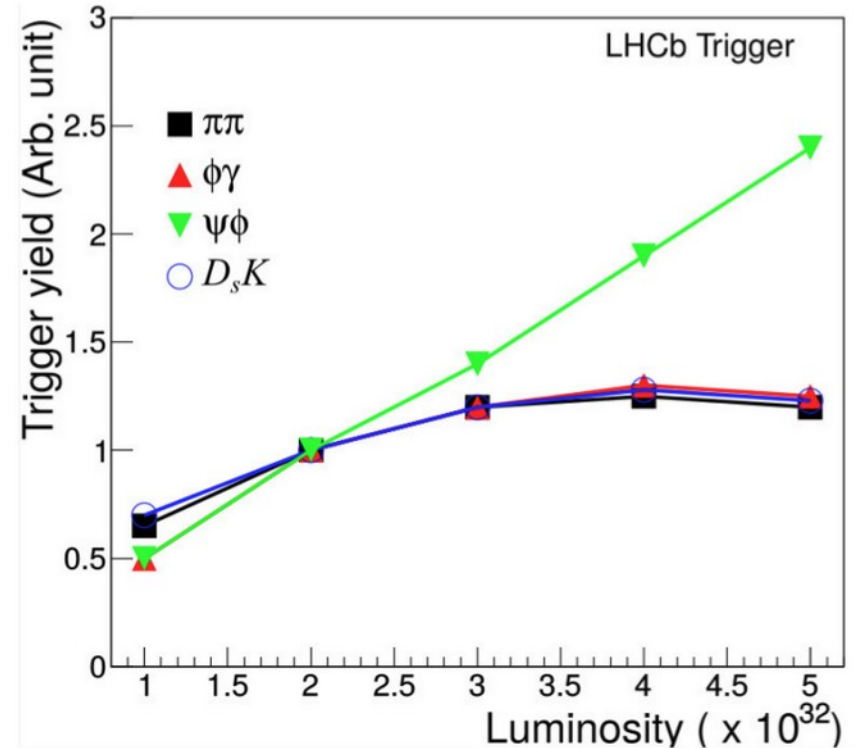
LHCb



Trigger-less readout: why?

- With traditional calorimeter+muons trigger:
Increase in luminosity
≠
increase in “interesting” events
- As luminosity grows, thresholds must be increased to keep rate constant
- Trigger inefficiency from higher thresholds is not compensated by higher lumi

Low level trigger yield vs Luminosity ($\text{cm}^{-2} \text{s}^{-1}$)
for a trigger rate of 1 MHz



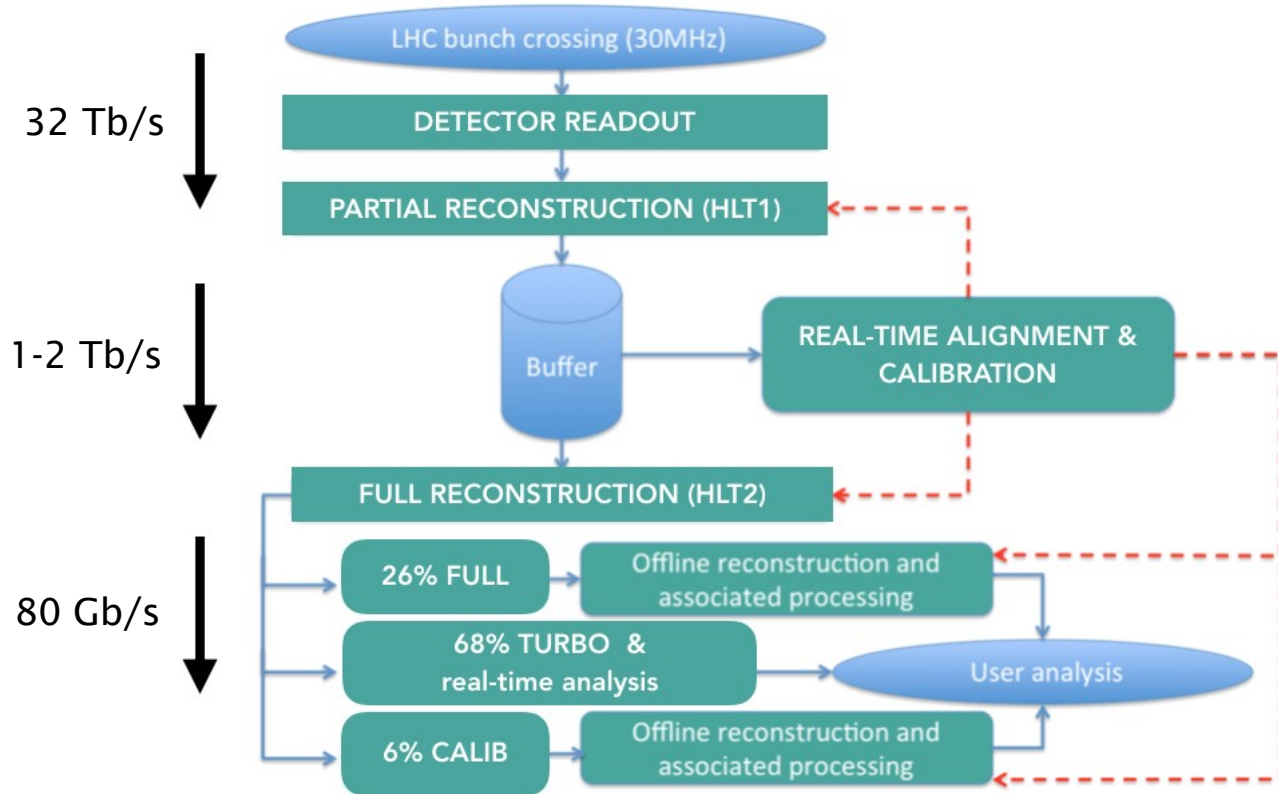
Trigger-less readout: how?

- Spectrometer geometry: fibres/cables are not "in the way"
- Relatively low radiation levels permit to relax the constraint on the FPGAs used for "middle" layer processing
- Zero-suppression on the detectors
- Total event-size comparatively small (~100 kB)
- Bonus: software trigger can do online selection with offline-like reconstruction

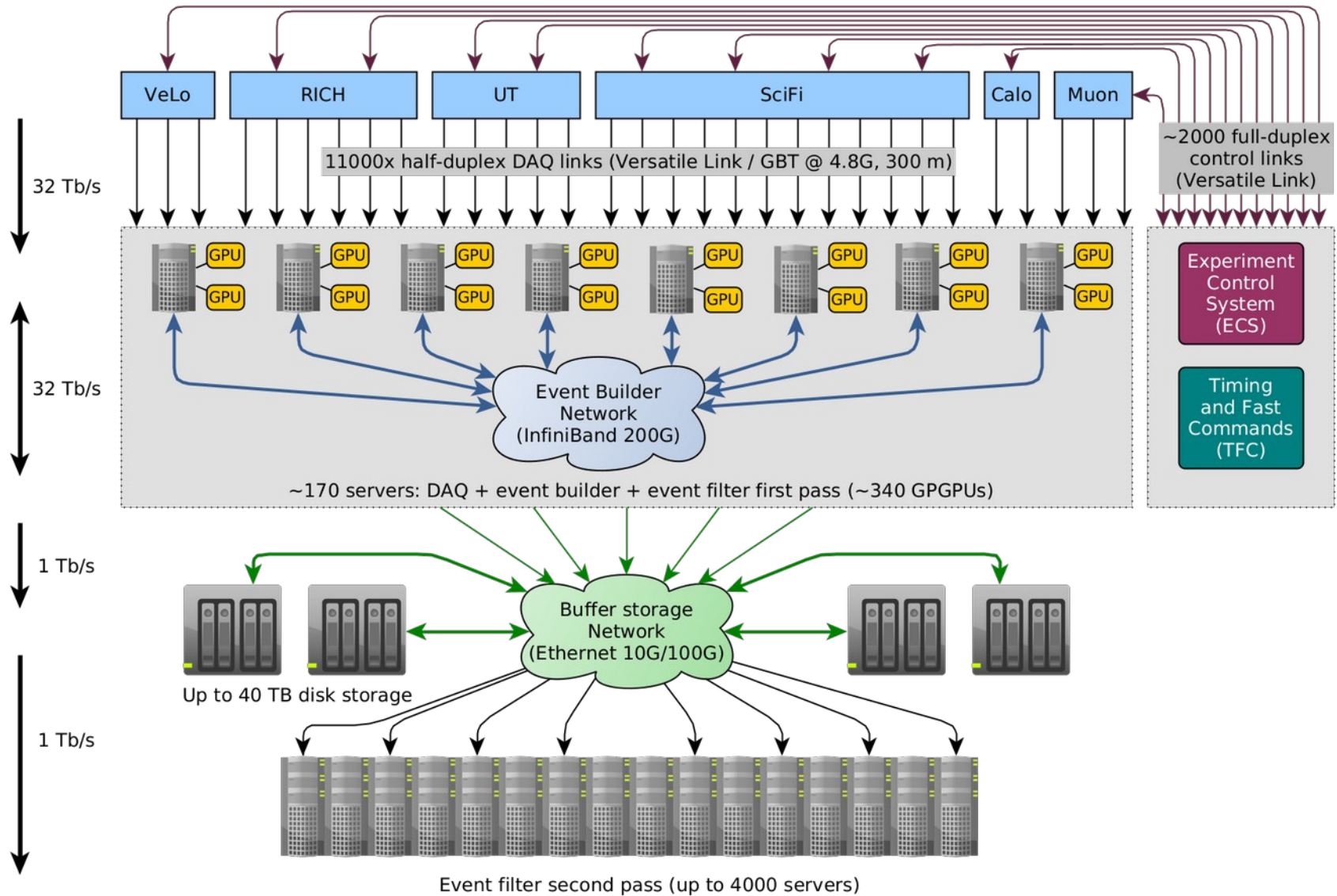


Data-processing and event selection

- Two stages of software filtering:
 - 1) "HLT1" on GPGPUs
 - 2) "HLT2" on CPUs
- Large storage buffer to decouple the two
- Calibration and alignment are performed "semi-live", while the data are buffered



System overview



The PCIe40

A single custom-made FPGA board for DAQ and Control

- Based on Intel Arria10
- 48x10G capable transceivers on 8xMPO for up to 48 full-duplex Versatile Links
- 2 dedicated 10G SFP+ for timing distribution
- 2x8 Gen3 PCIe



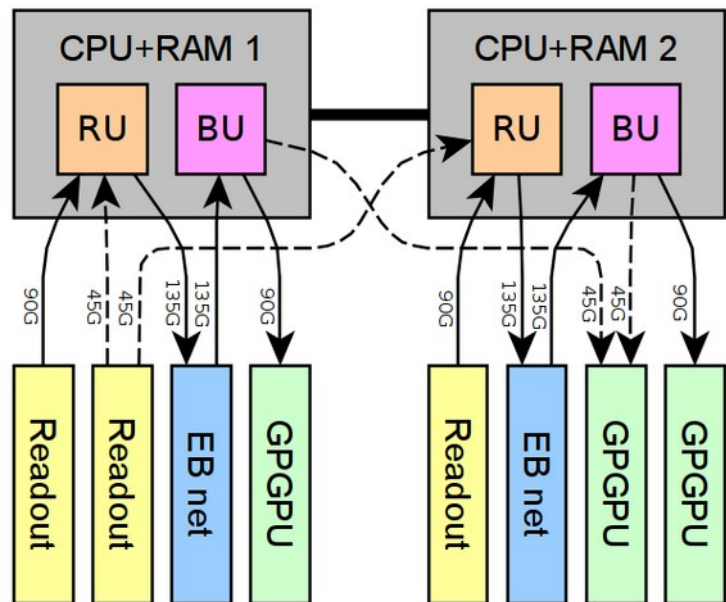
One board, many firmware personalities

- Readout Supervisor (SODIN):
 - Reception and distribution of global timing
 - Generation and distribution of synchronous and asynchronous commands
 - Generation of events veto, triggers and calibration events
- Interface Board (SOL40):
 - Distribution of the global timing to the front-ends
 - Interface bridge between the control system and the front-ends
- Readout Board (TELL40):
 - Acquisition and first pre-processing of the data

Event builder server



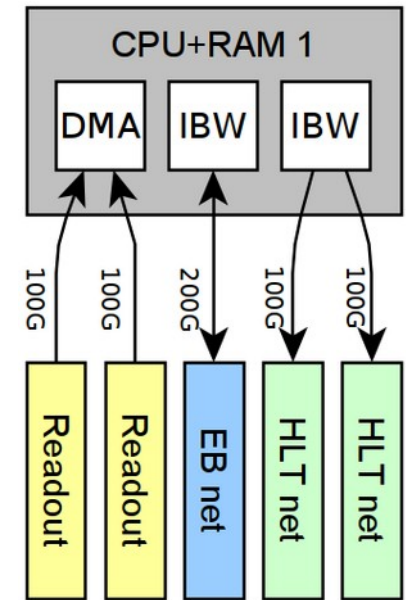
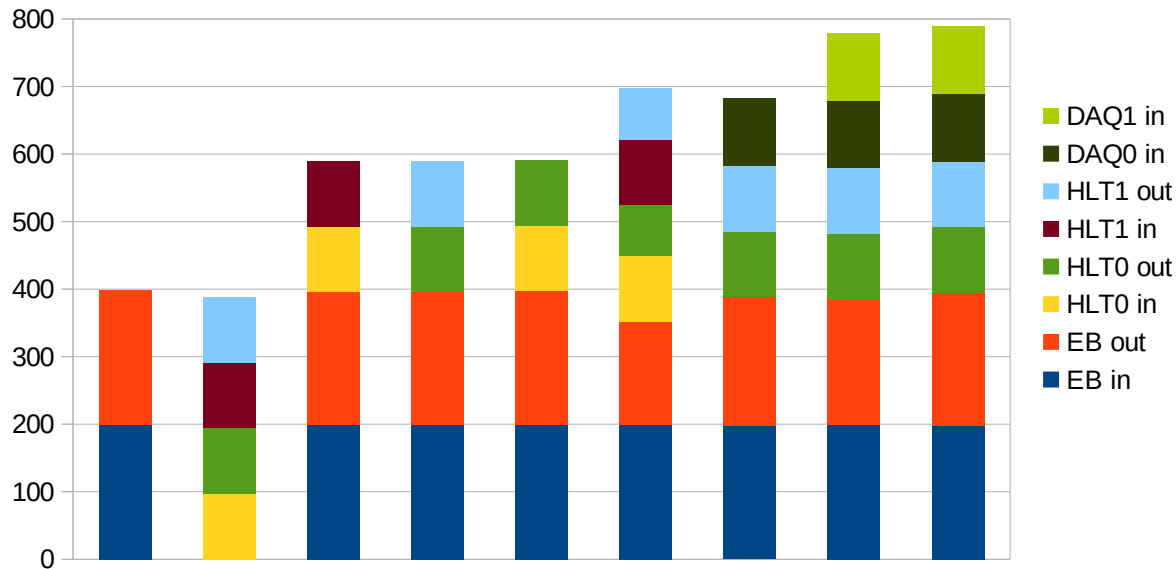
- Hosts:
 - 3 readout boards
 - 2 InfiniBand 200G NICs
 - Up to 3 GPUs
- Main system memory is used as EB buffer



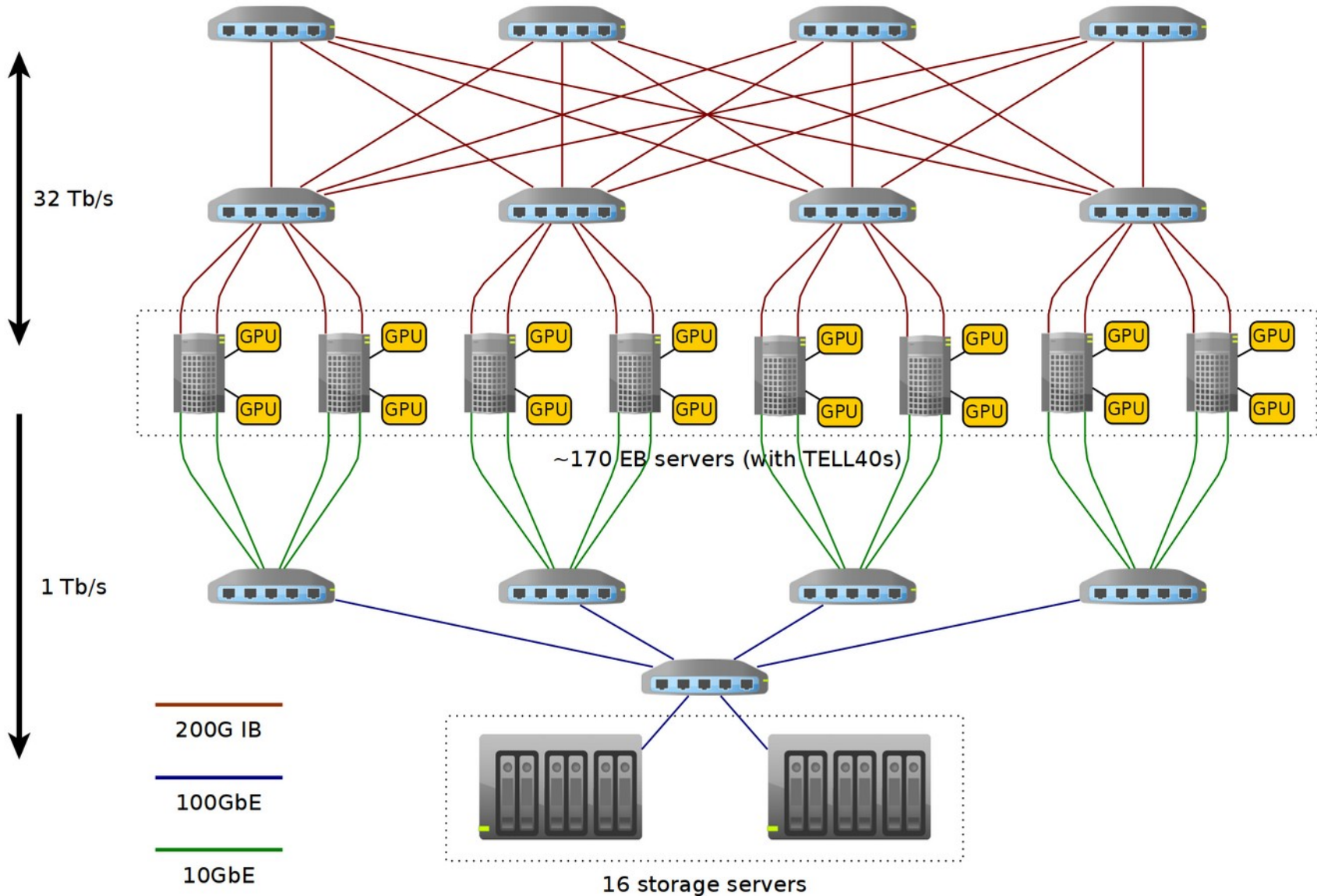
Challenges for EB servers

Memory subsystem pushed to the limits!

Total in/out throughput (Gb/s) with NPS=1, QPs=2, WrOrd=1



Event builder networks

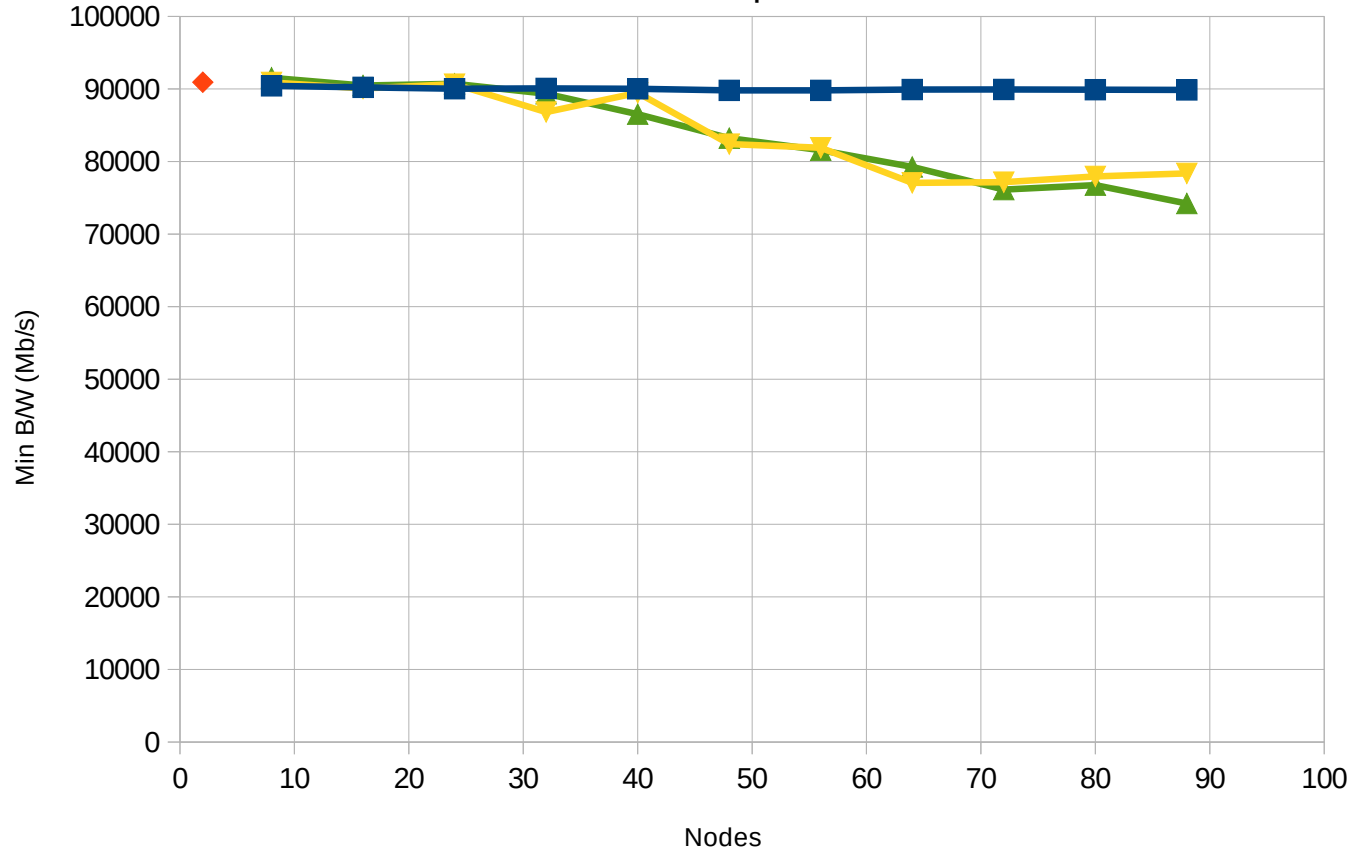


Challenges for the EB network

- Needs to collect data from 478 TELL40 FPGA boards into a single "location"
- And hand them over to GPGPUs + CPUs for further processing
- Want high link-load (keeping costs low)
- Want to use some kind of remote DMA to reduce server-load
- Traffic is inherently congestion inducing
 - Our solution: careful application-level traffic shaping (needs lossless network)
 - Specialized routing algorithm for our network topology (fat tree)

Scalability on InfiniBand

EB bandwidth per node



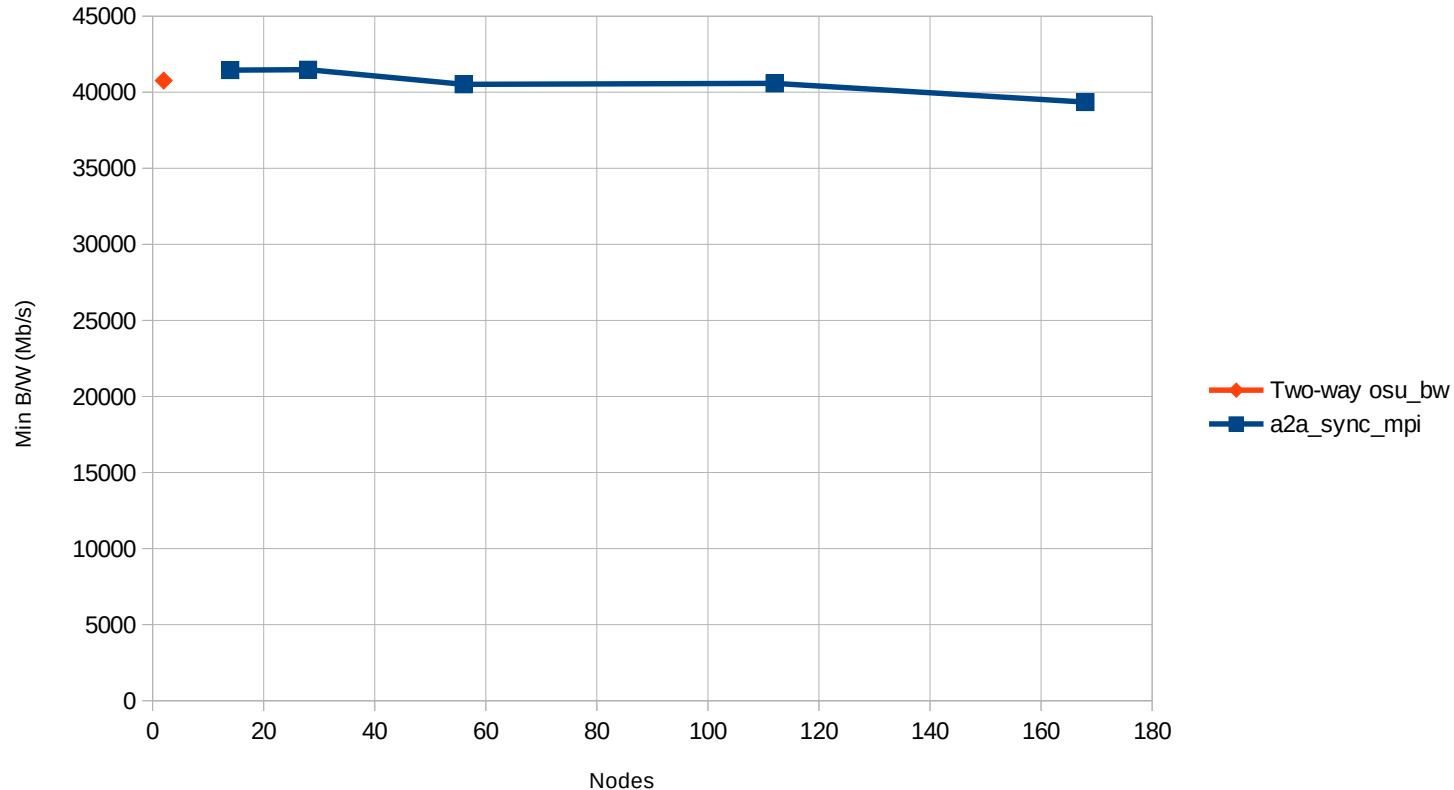
Tested at the
Goethe-HLR
HPC cluster
(InfiniBand 100G)

- Two-way osu_bw
- a2a_sync_mpi
- daqpipe / linear shift
- daqpipe / random

With the right
traffic shaping,
almost perfect
scalability!

Scalability on InfiniBand

EB bandwidth per node

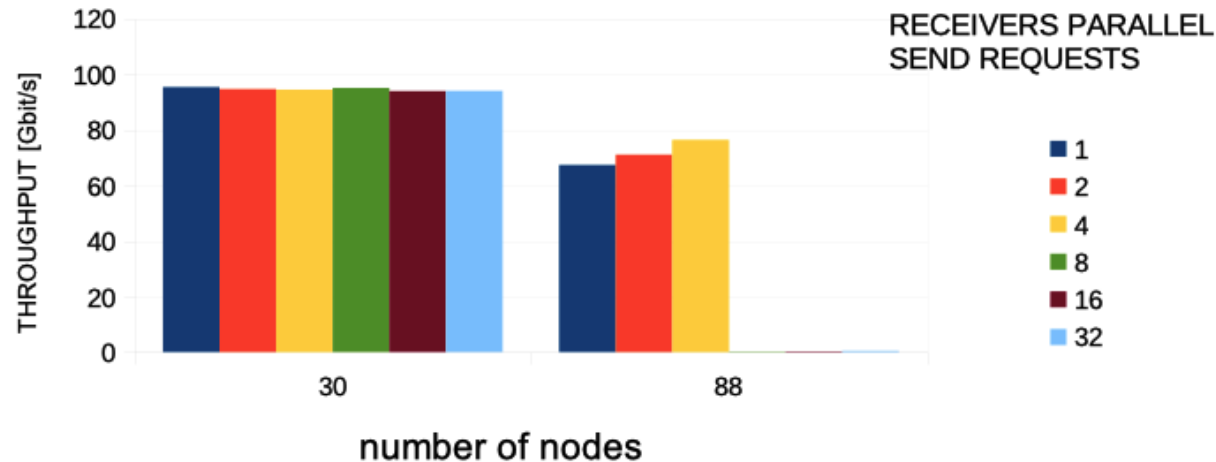


Tested on the
CMS DAQ
(InfiniBand 56G)

Very good
scalability with
almost 200
nodes

Scalability on Ethernet

30 nodes versus 88 nodes
(2 MB optimal message size)



- Deep buffers alone don't save us
- Hardware flow control from many Ethernet vendors is flakey

Why InfiniBand?

- PCIe Gen4 allows using 200 Gbit/s connections which save cost and help with scalability.

However 200 Gbit/s so far only effectively exists for InfiniBand!

- Ethernet flow-control could not be made to work properly on available reference platforms
- Ethernet remains – for us – affected by worrying / irritating scaling issues
- Probably most important: could never get access to a really big Ethernet test-system: need the full event-builder for testing. For InfiniBand we have used super-computer sites

→ Lowest risk solution – within our budget – is the InfiniBand solution

Summary

- LHCb can do and afford a full read-out at bunch-crossing rate
- Single stage synchronous readout built around GBT and a single flexible FPGA board
- Detector control uses the same FPGA boards as the timing distribution system
- AMD Rome (PCIe Gen4) based servers make compact, very-high-I/O event-builder, connected with 200 Gb/s InfiniBand
- Event-selection is entirely in software to maximize physics yield, increase the amount of data collected, flexibility and minimize cost
- The system is very well scalable, by up to 3 a factor without any substantial changes