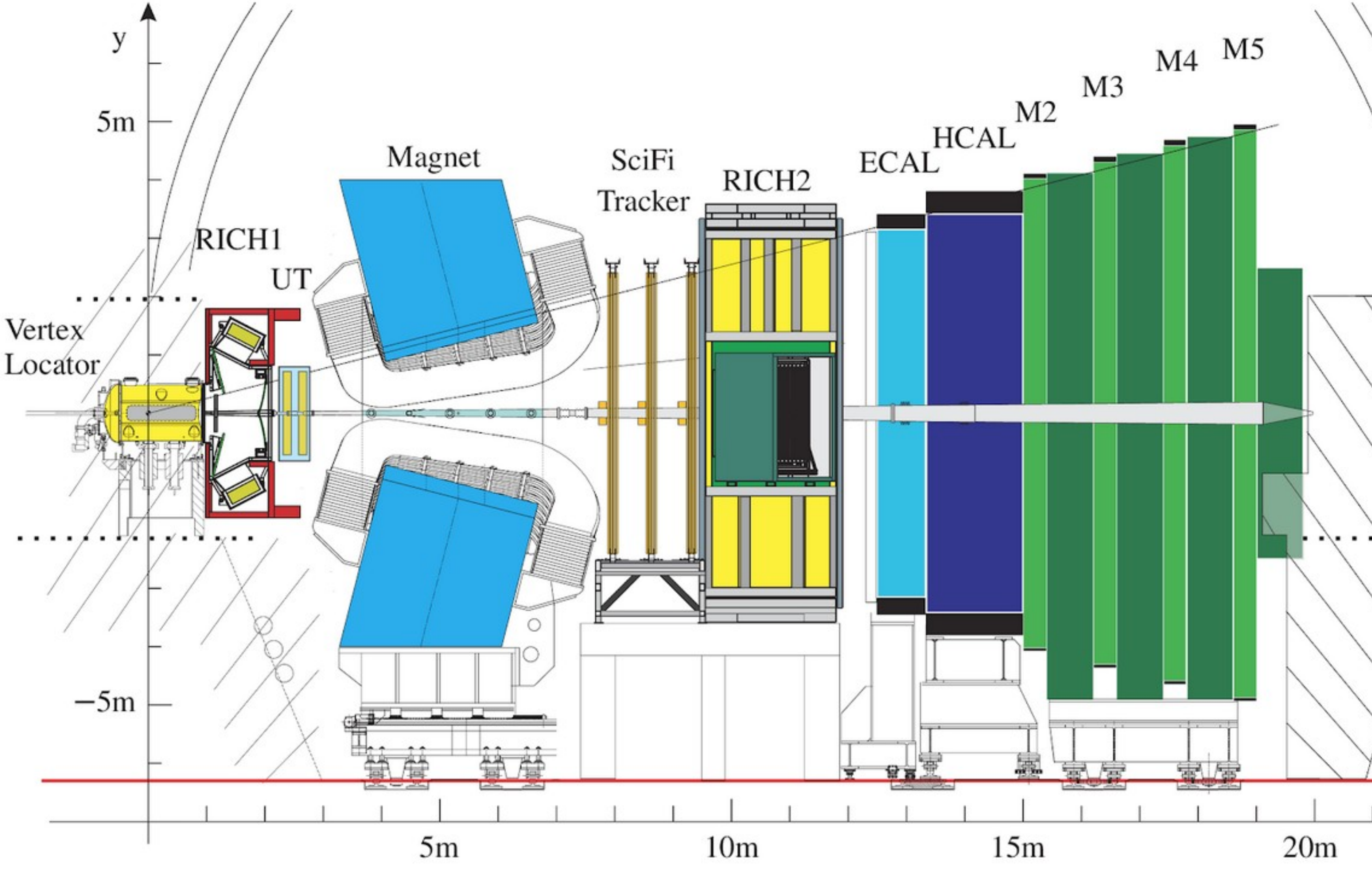
# LHCb DAQ & event filter in 2021-2024

Tommaso Colombo (CERN)
on behalf of the LHCb Onliners

Streaming Readout Workshop VII
Brookhaven National Laboratory
17 November 2020

# LHCb in 2021-2024

- Single-arm forward spectrometer at the LHC
- p-p bunch crossing rate: 30 MHz
- Luminosity: $2\times10^{33}$ cm$^{-2}$s$^{-1}$



# Trigger-less readout: why?

- With traditional calorimeter+muons trigger:

  Increase in luminosity

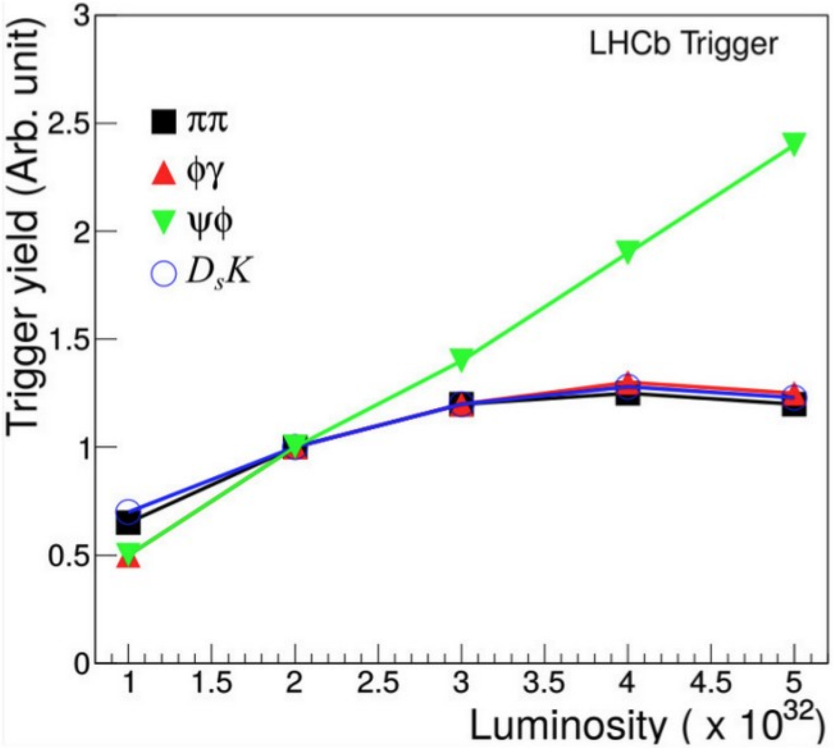
  ≠

  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

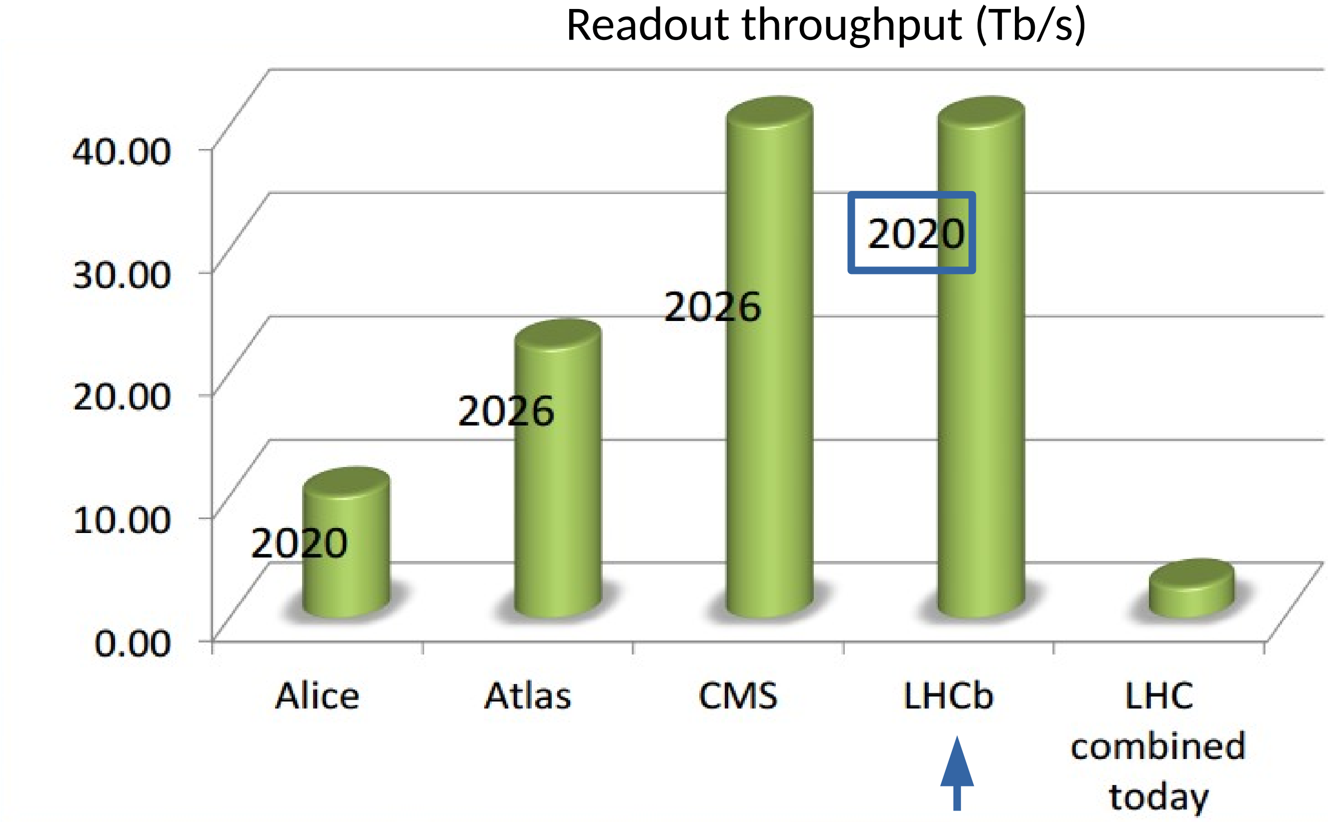
Low level trigger yield vs Luminosity ($cm^{-2}\ s^{-1}$) for a trigger rate of 1 MHz



# Trigger-less readout: how?

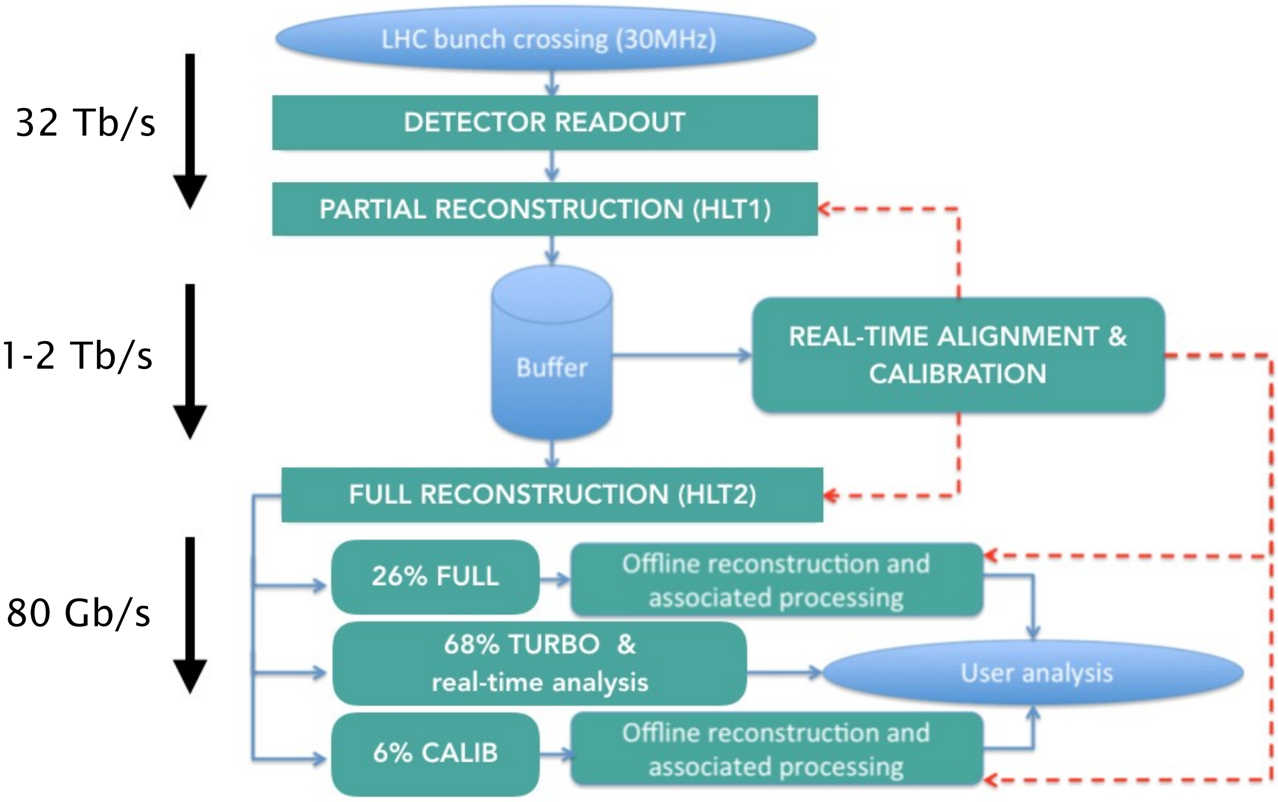
- Spectrometer geometry:
  fibres/cables are not "in the way"
- Relatively low radiation levels allow relaxed radiation-hardness requirements for FPGAs in many detector front-ends
- Zero-suppression on the detectors
- Total event size comparatively small (~100 kB)
- Bonus:
  software trigger can do online selection with offline-like reconstruction

# Trigger-less readout: when?

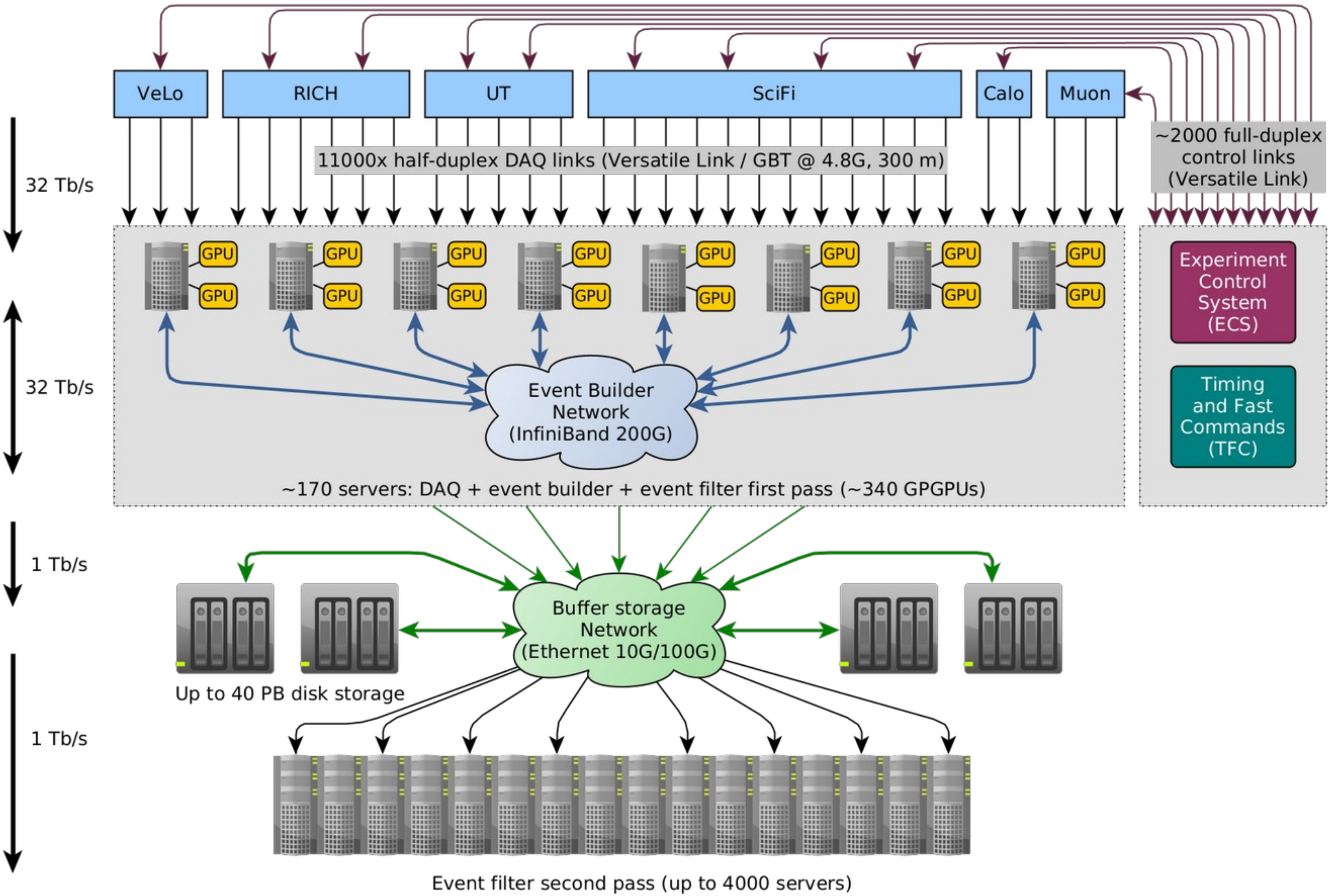


# Data-processing and event selection

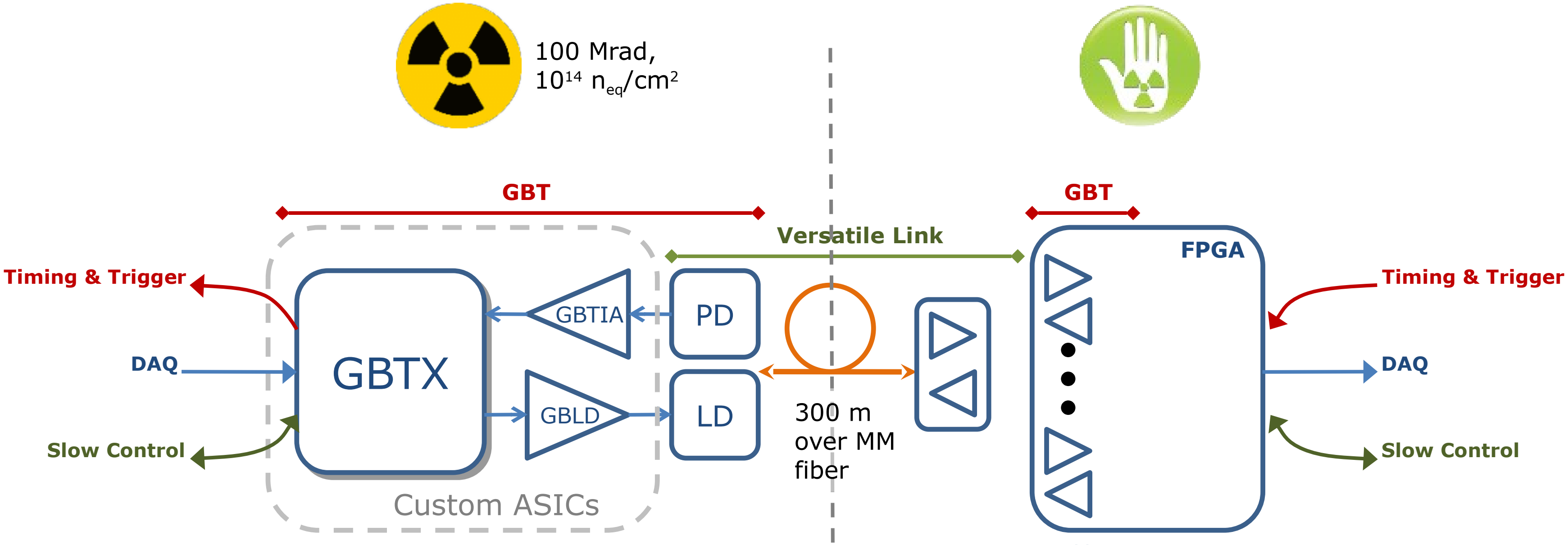
- Two stages of software filtering:
  1) "HLT1" on GPGPUs
  2) "HLT2" on CPUs
- Large storage buffer to decouple the two
- Calibration and alignment are performed "semi-live", while the data are buffered


32 Tb/s
1-2 Tb/s
80 Gb/s
LHC bunch crossing (30MHz)
DETECTOR READOUT
PARTIAL RECONSTRUCTION (HLT1)
Buffer
REAL-TIME ALIGNMENT & CALIBRATION
FULL RECONSTRUCTION (HLT2)
26% FULL
Offline reconstruction and associated processing
68% TURBO & real-time analysis
User analysis
6% CALIB
Offline reconstruction and associated processing

# System overview

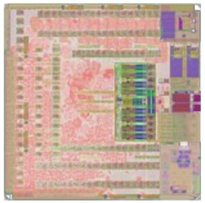

VeLo
RICH
UT
SciFi
Calo
Muon
11000x half-duplex DAQ links (Versatile Link / GBT @ 4.8G, 300 m)
~2000 full-duplex control links (Versatile Link)
32 Tb/s
GPU
Experiment Control System (ECS)
Timing and Fast Commands (TFC)
32 Tb/s
Event Builder Network (InfiniBand 200G)
~170 servers: DAQ + event builder + event filter first pass (~340 GPGPUs)
1 Tb/s
Buffer storage Network (Ethernet 10G/100G)
Up to 40 PB disk storage
1 Tb/s
Event filter second pass (up to 4000 servers)

# Front-end: GBT over Versatile Link



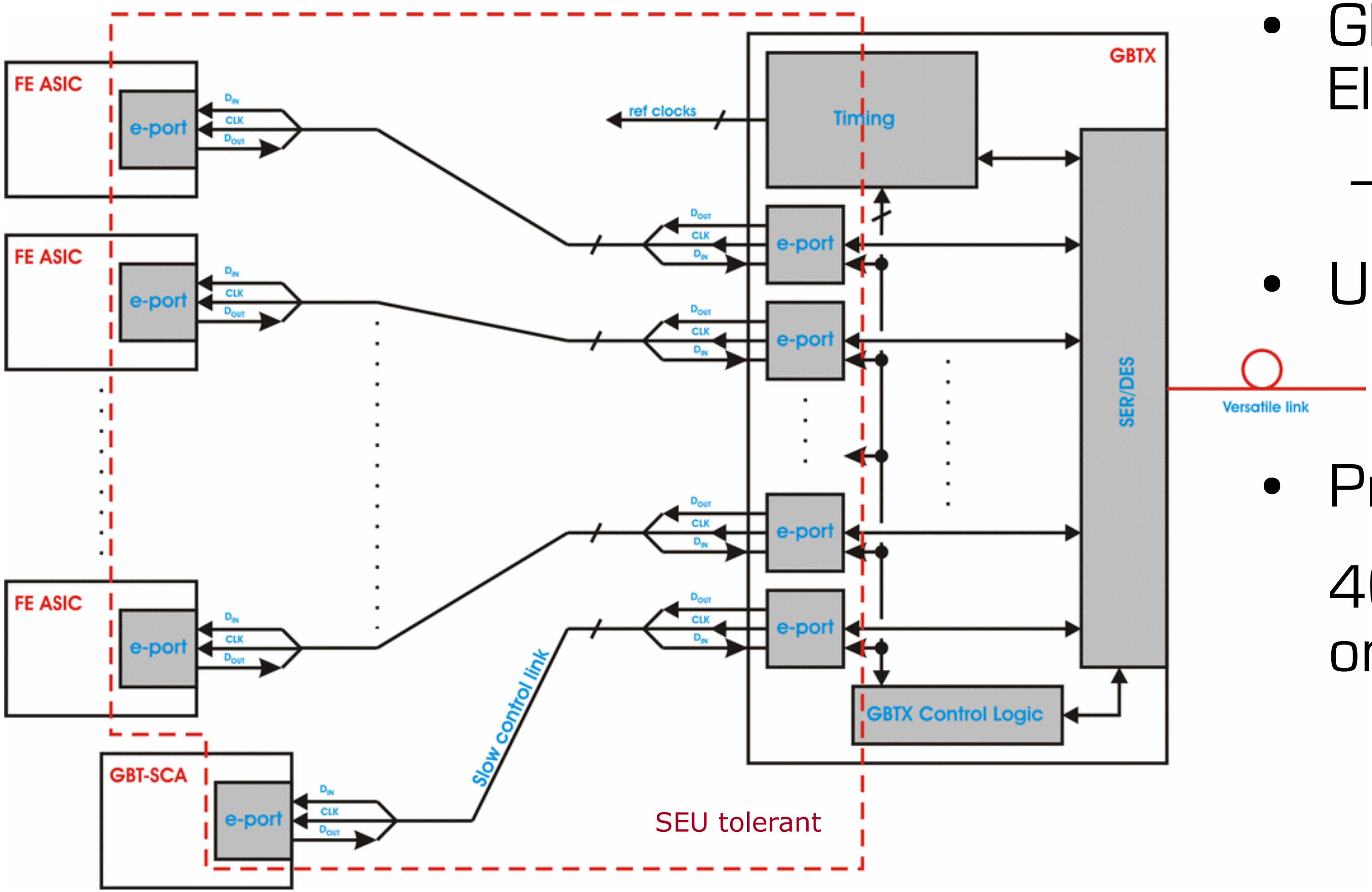
**On-Detector**
Custom Radiation-Hard Electronics

**Off-Detector**
Commercial Off-The-Shelf (COTS)



Credit:
P. Moreira
S. Baron
(CERN)

# Front-end: GBTx multiplexing



- GBT/Frontend interface: Electrical links (e-link)
  - Serial, bidirectional
- Up to 40 links per ASIC
- Programmable data rate:

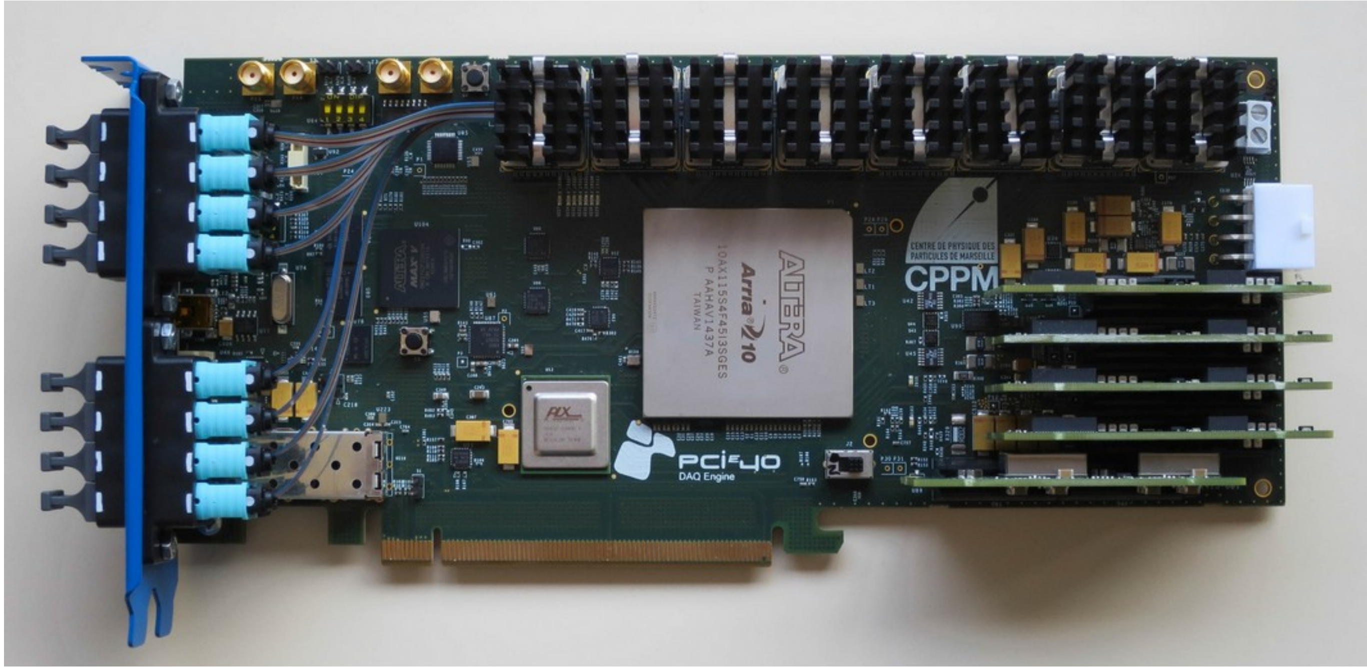
  40×80, 20×160, or 10×320 Mb/s

Credit:
P. Moreira
(CERN)

# Back-end: 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
- 16x PCIe 3.0



# One board, many firmware personalities

## 1 Readout Supervisor (SODIN)

- Reception and distribution of global 40 MHz timing
- Generation and distribution of synchronous and asynchronous commands
- Event type (physics, calibration, empty) generation

# One board, many firmware personalities

42 Interface Boards (SOL40)

- Distribution of the global timing to the front-ends
- Interface bridge between the control system and the front-ends

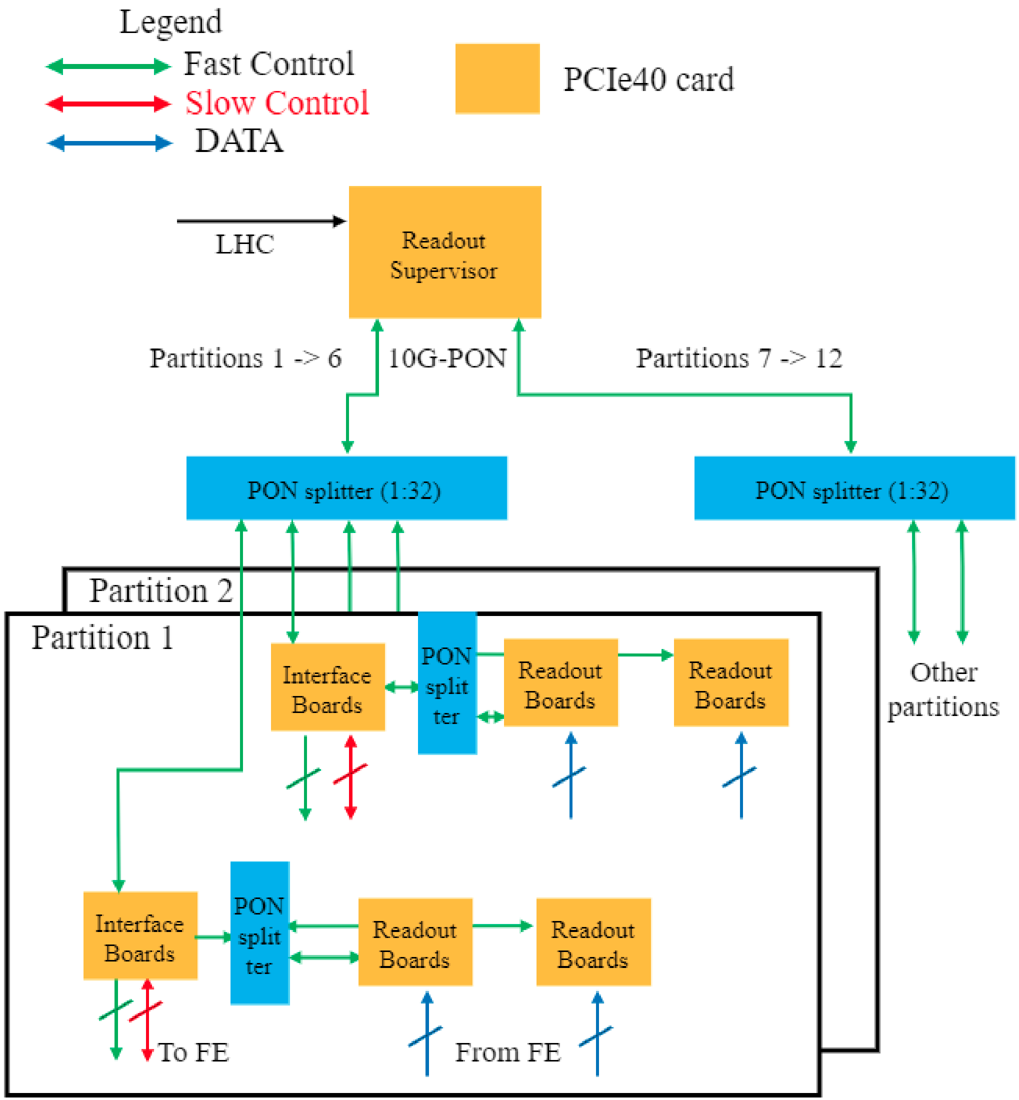
# One board, many firmware personalities

478 Readout Boards (TELL40)

- Data Acquisition
- First pre-processing of the data
- E.g.:
  - Re-ordering and separation on event boundaries of streaming data
  - Hit clustering

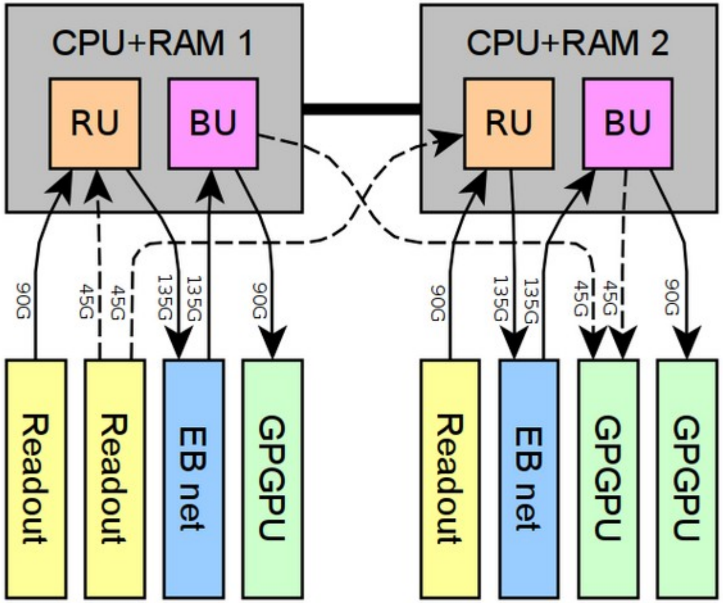
# Timing and Fast Commands

- Synchronously driving the Front-End electronics over GBT
- 10G-PON for efficient Back-End signal distribution and fixed phase clock recovery
- Partitioning for debugging and commissioning



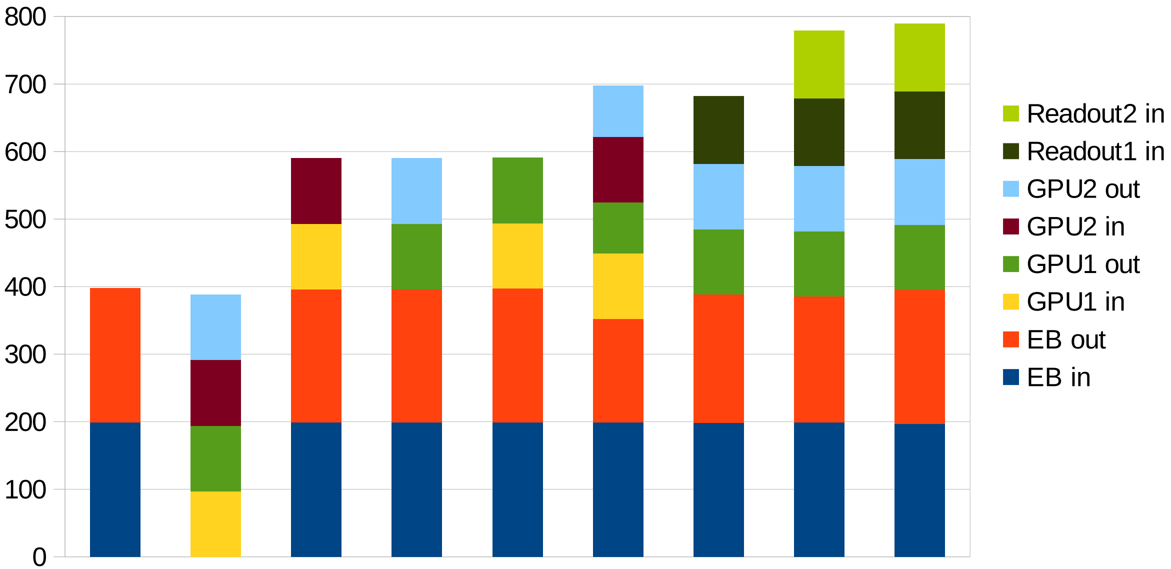
# Event builder server

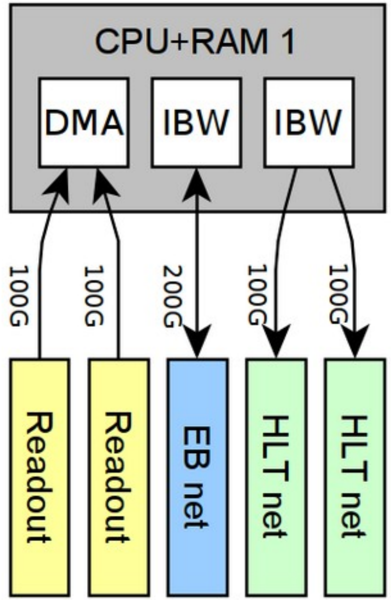
- 2 AMD EPYC 7002-series CPUs
  - PCIe 4.0
  - 8+8 DDR4 channels
- 3 readout boards
- 2 InfiniBand 200G NICs
- Up to 3 GPUs
- 512 GiB RAM (buffer to decouple EB and readout)



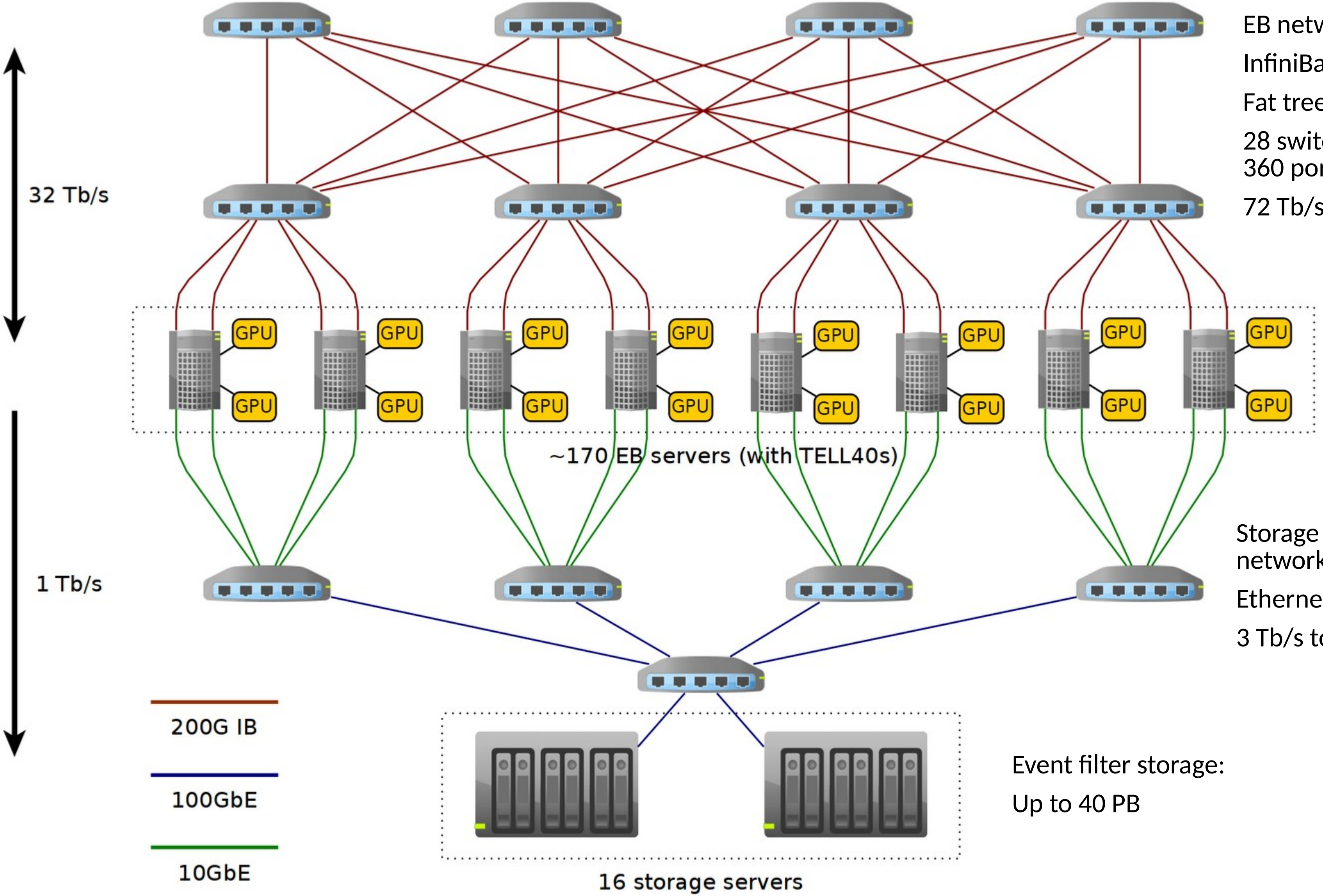
# Challenges for EB servers

Memory subsystem pushed to the limits! RDMA is crucial.


Total I/O throughput (Gb/s)



# Event builder networks



EB network:

InfiniBand 200G

Fat tree

28 switches
360 ports

72 Tb/s total

Storage / filter network:

Ethernet 10/100G

3 Tb/s total

Event filter storage:

Up to 40 PB

# Challenges for the EB network

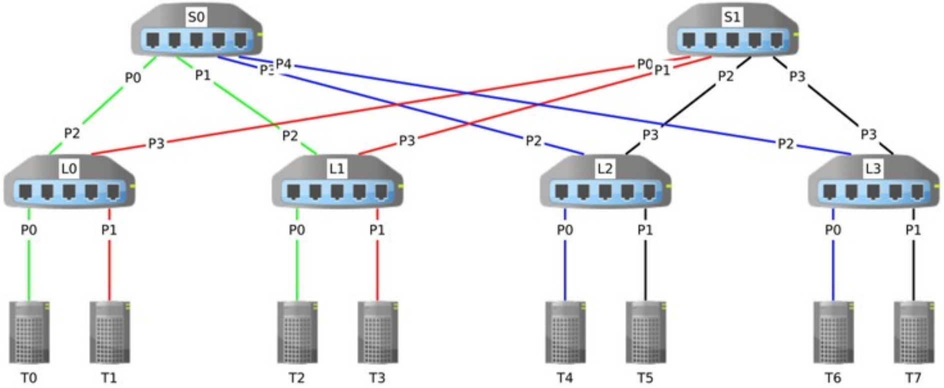
- Needs to collect data from 478 readout boards into a single "location"
- And hand it over to GPGPUs + CPUs for further processing
- Want high link-load (keeping costs low)
- Want to use some kind of remote DMA to reduce server-load
- Traffic is inherently congestion-inducing

→ Our solution: careful application-level traffic scheduling

→ Specialized routing algorithm for our network topology (fat tree)

# Event building, a.k.a. MPI_Alltoall

- Traffic pattern is *all-to-all gather*:
  For each event, one "builder" server receives fragments from all servers
- Schedule: linear shift
  - With N servers, the transfer of N events is divided into N phases
  - In every phase each server exchanges data with only one server
- If the start of a phase is synchronized, and the network is non-blocking
  → no link conflicts!



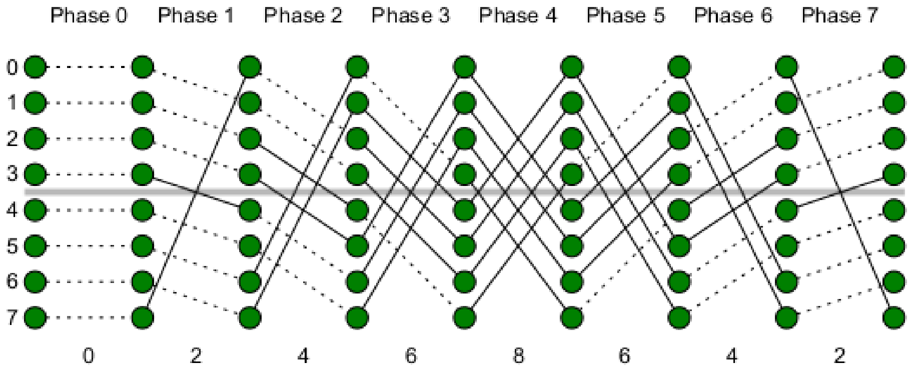
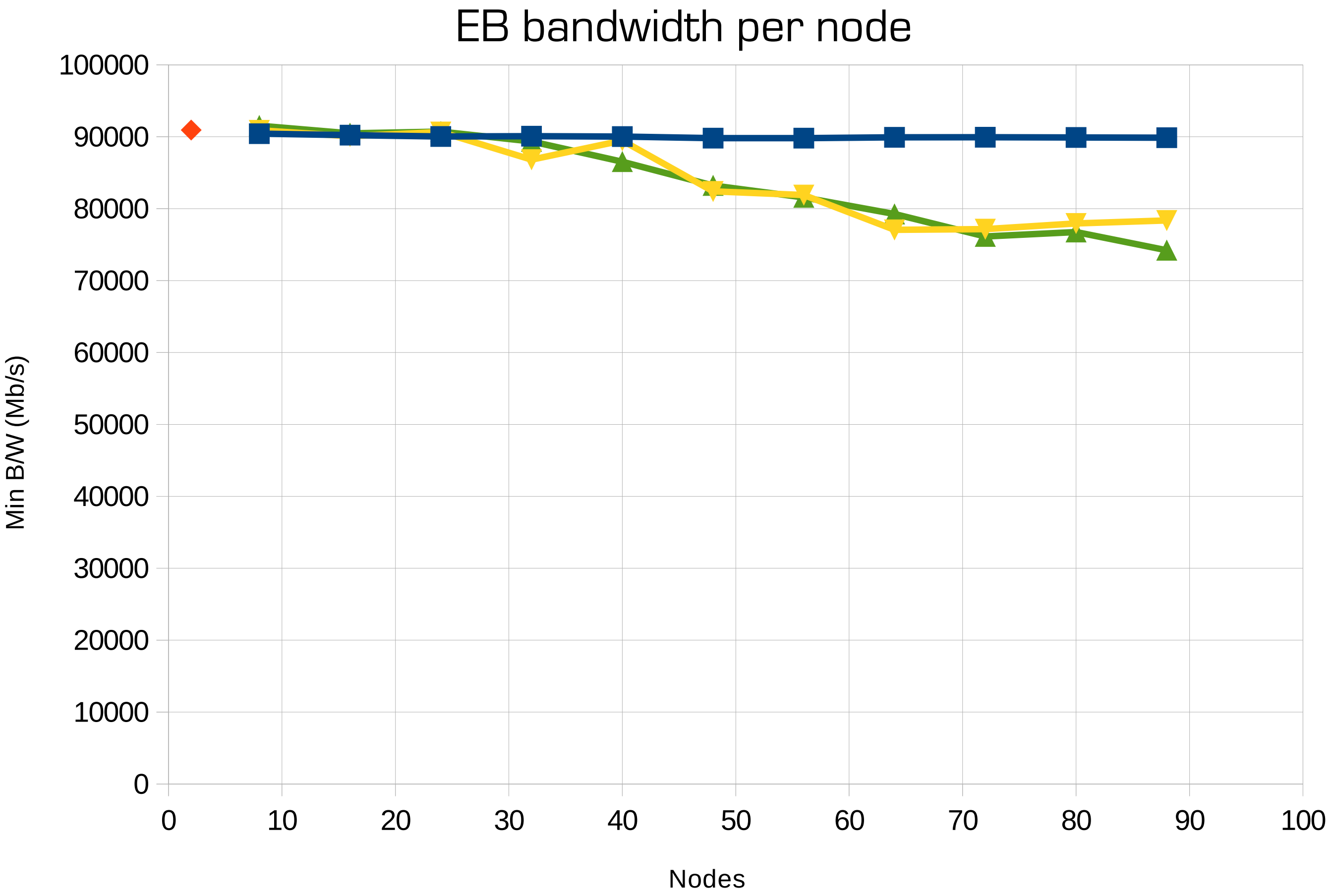


Image credit: B. Prisacari et al.

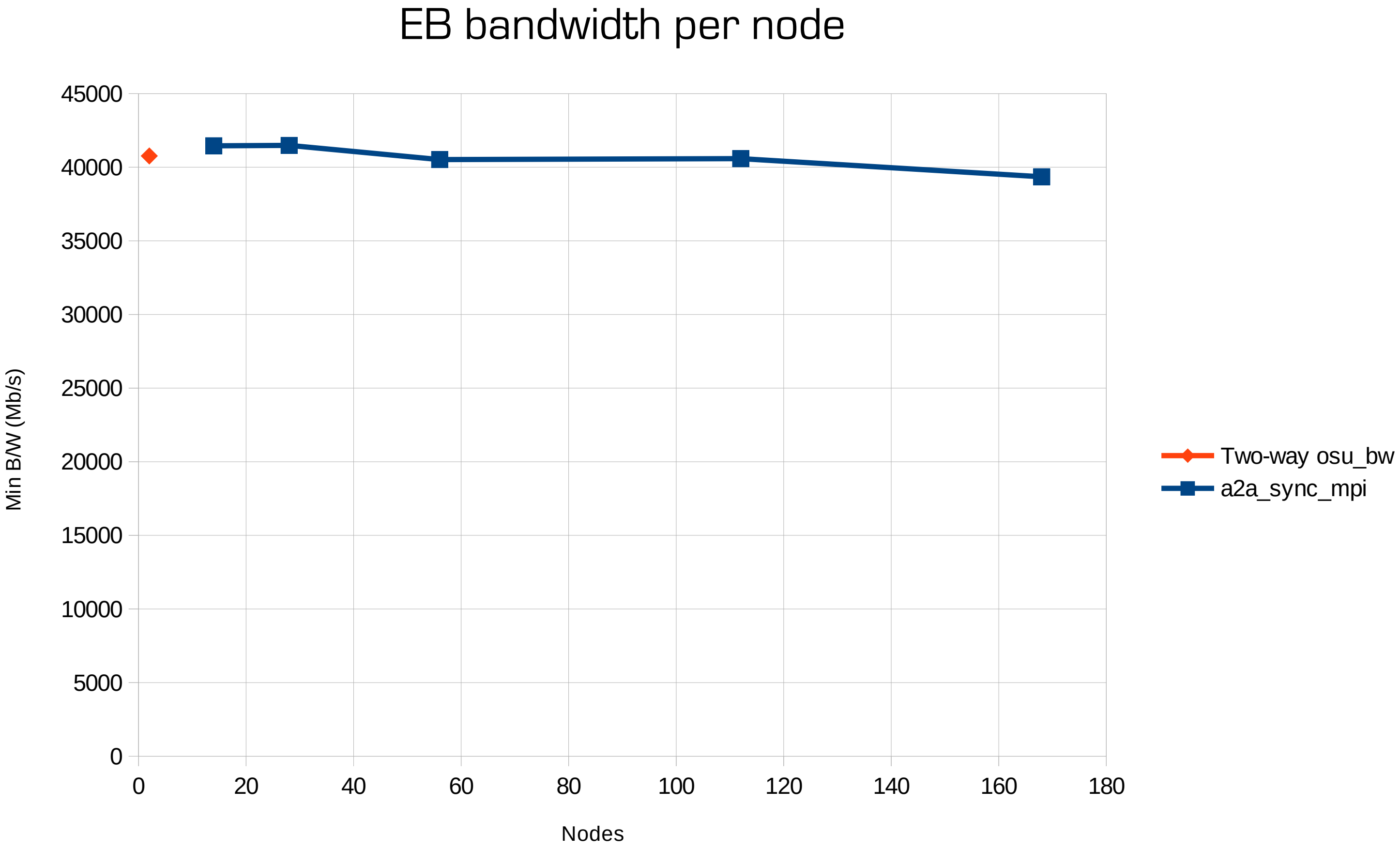
# Scalability on InfiniBand



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

With the right traffic shaping, almost perfect scalability!

# Scalability on InfiniBand



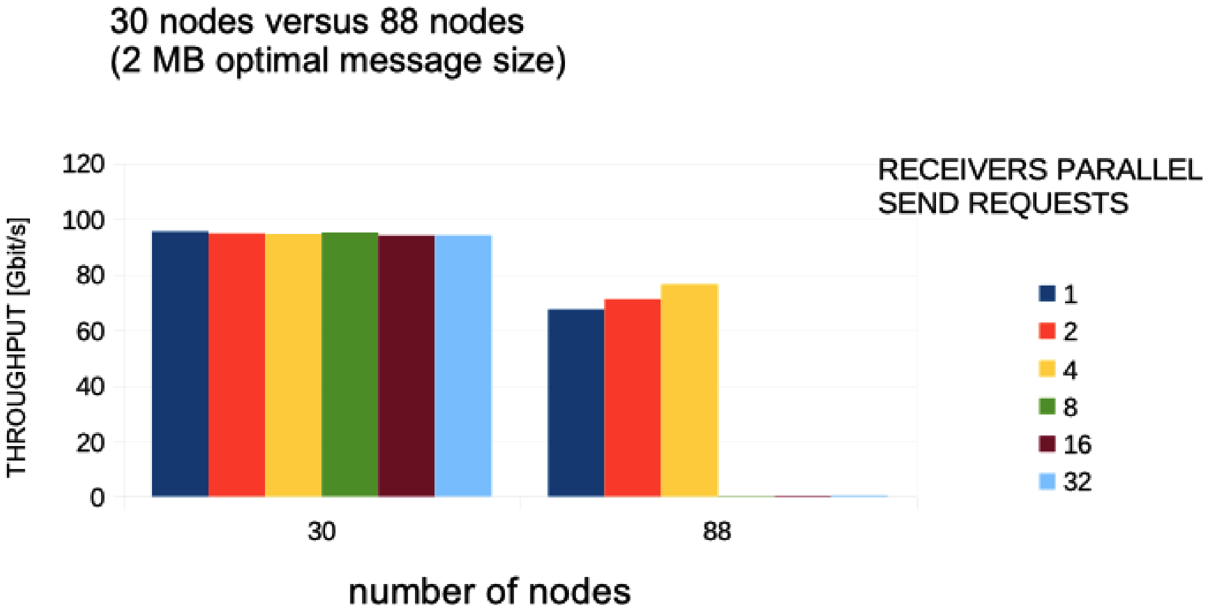
Tested on the CMS DAQ (InfiniBand 56G)

Very good scalability with almost 200 nodes

# Why InfiniBand?

- PCIe Gen4 allows using 200 Gbit/s connections:
  Lower cost, better scalability, but so far only effectively exist for IB!
- Remote DMA is crucial for EB server performance:
  - RDMA implementations do not like packet drops:
    either deep buffers or good flow control are needed.
  - Deep buffers @ 100G = expensive/non-existent RAM tech.
  - Many flow-control bugs found on available reference platforms.
- **Could never get access to a really big Ethernet test system:**
  Network congestion issues only appear at scale.
  For InfiniBand we have used super-computer sites.
- Lowest risk solution – within our budget – is the InfiniBand solution

# Scalability on Ethernet with deep buffers



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

# 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