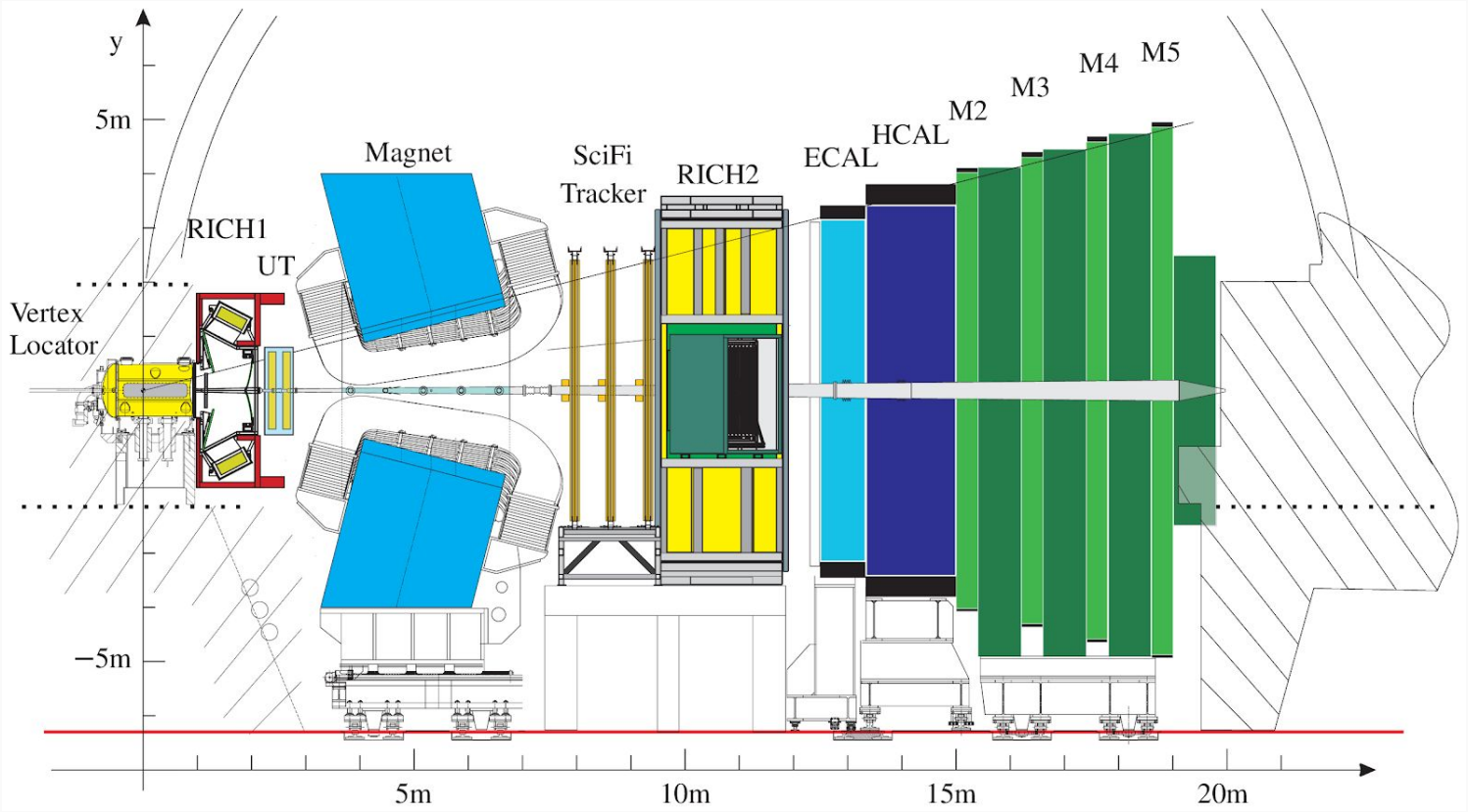
One year of LHCb triggerless DAQ: achievements and lessons learned

26th International Conference on Computing in High Energy & Nuclear Physics
08/05/2023

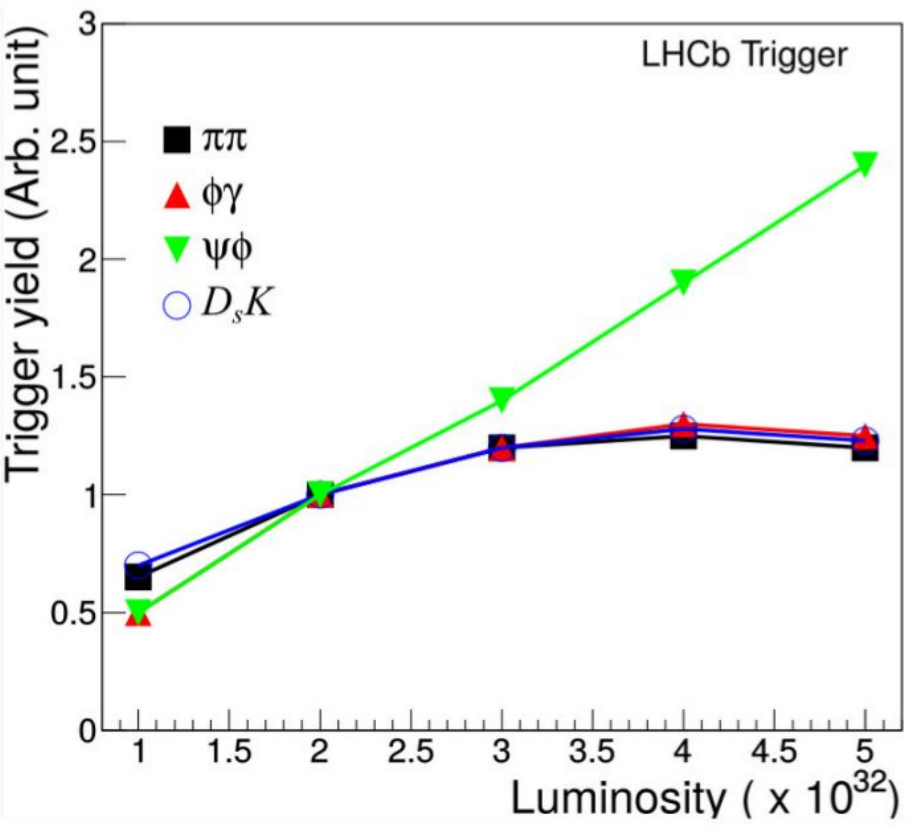
Flavio Pisani for the LHCb Online team

CERN

The LHCb experiment



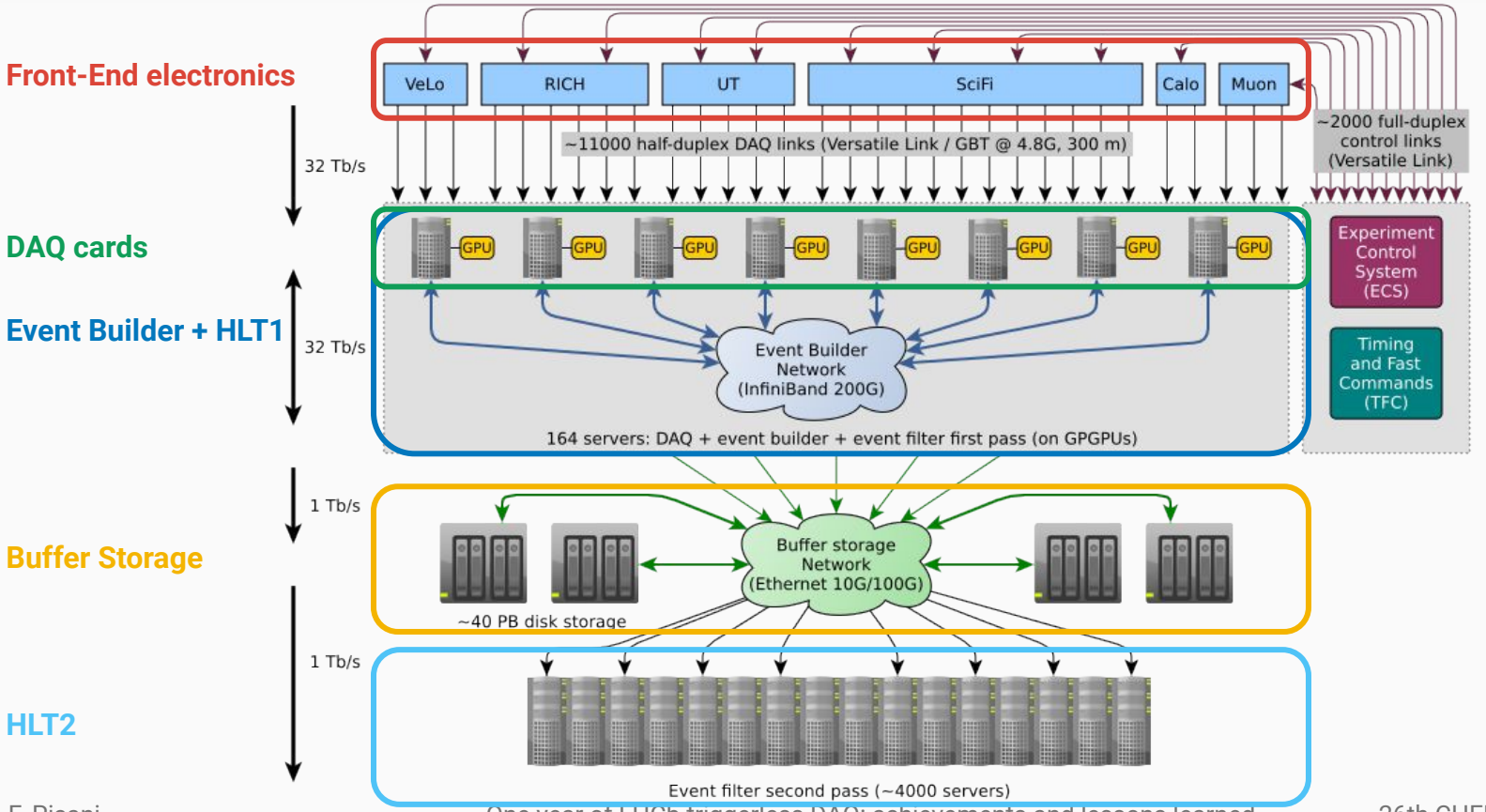
Full collision-rate readout: why?



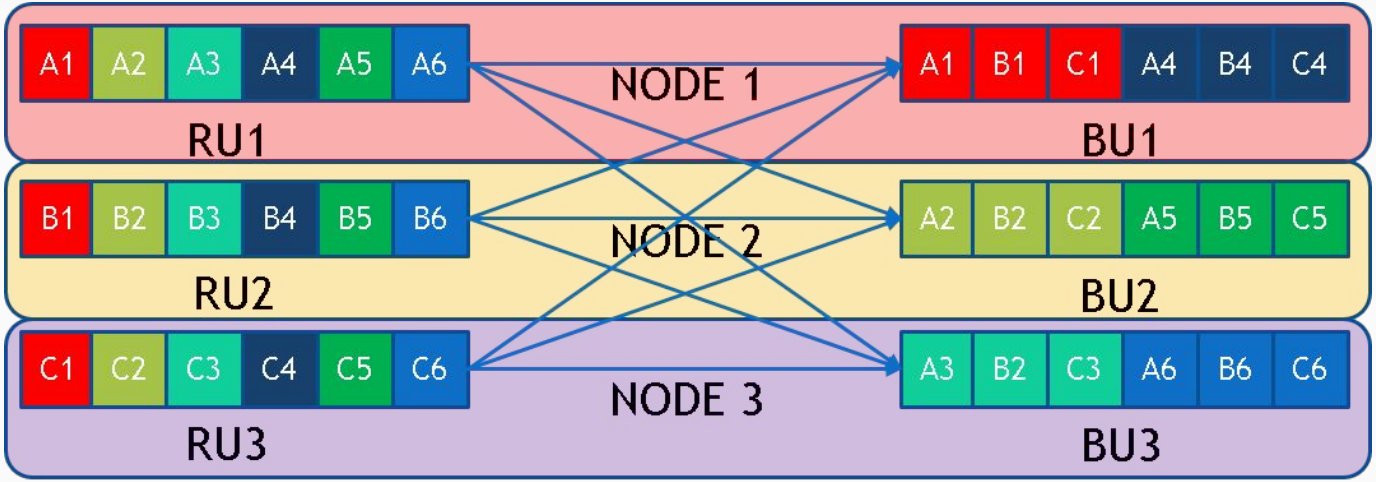
- The low-level trigger saturates in hadronic channels
- The instantaneous luminosity in Run 3 will go up to $2x10^{33}$ cm^{-2} s^{-1}

A substantial upgrade is needed to take advantage of the increased luminosity

Online DAQ system overview

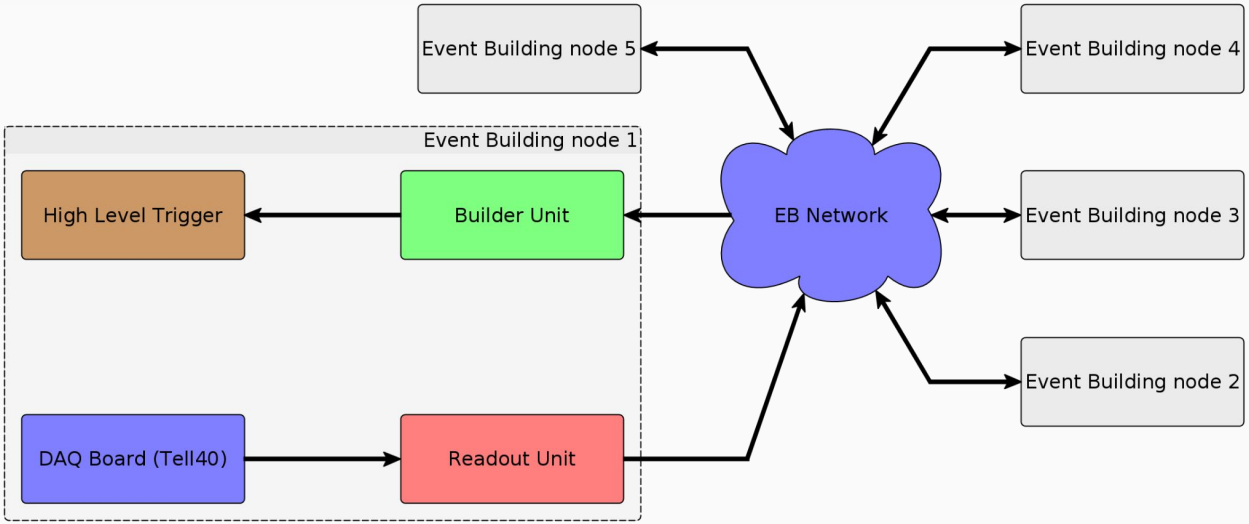


The Event Building process in a nutshell



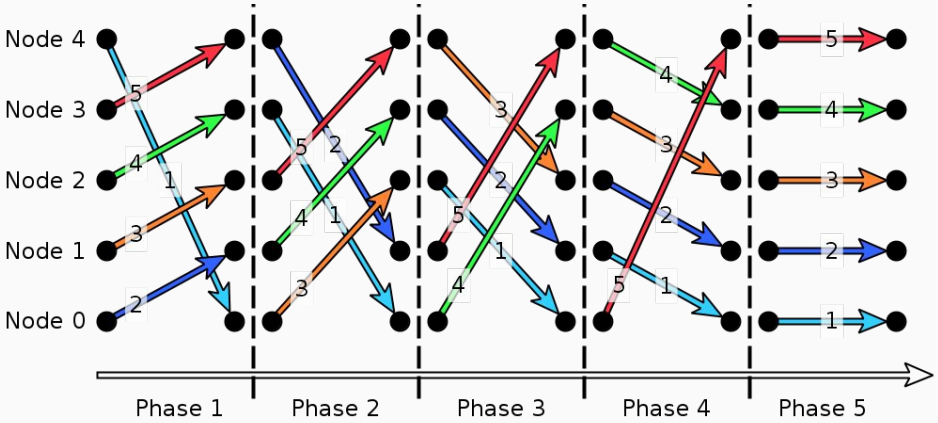
- Every event is divided into multiple fragments
- Every **Readout Unit (RU)** receives a fragment of the event
- Every **Builder Unit (BU)** has to gather all the fragments of the event

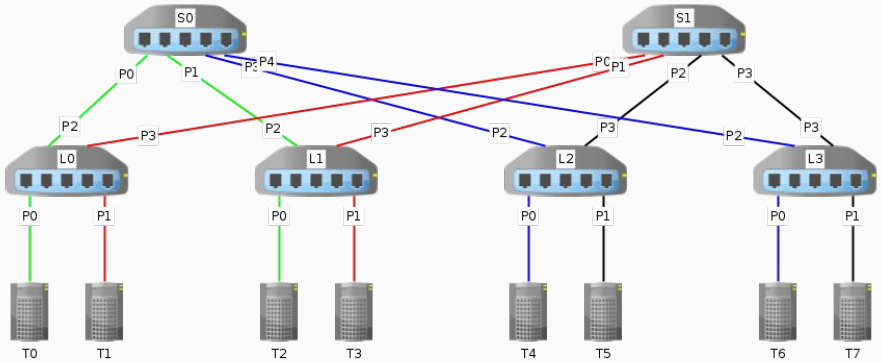
Event Builder: Software architecture



- Modular software architecture built in C++
- RU: it reads the data from the DAQ card and sends it over the EB network
- BU: it reads the data the EB network and it writes the built data into the HLT1 input buffer
- The scheduling synchronization is achieved using an in-band data barrier
- Dedicated low-level communication library
- Buffer-isolated critical sections to minimise slowdowns and deadtime

Event Builder: Traffic scheduling

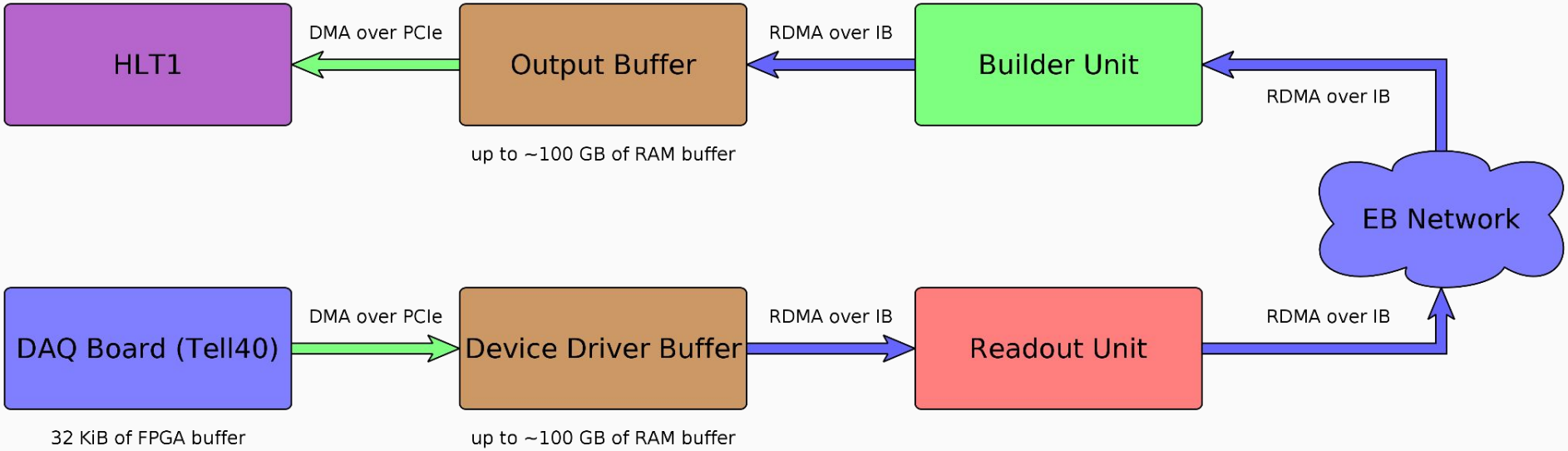




- The processing of **N** events is divided into **N** phases (**N** is the number of EB nodes
- In every phase one RU sends data to one BU, and every BU receives data from one RU
- During phase **n** RU **x** sends data to BU (**x** + **n**)%**N**
- All the units switch synchronously from phase **n** to phase **n** + 1

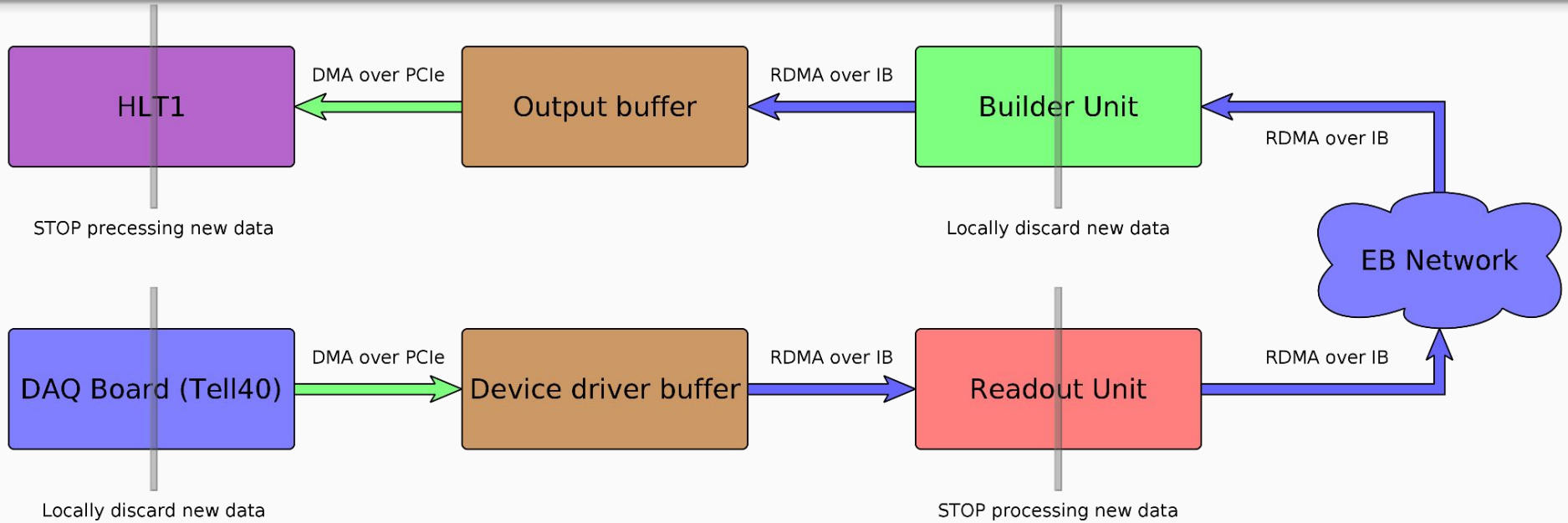
Congestion-free traffic on "selected networks" (e.g. fat-tree networks)

Latency and server flow



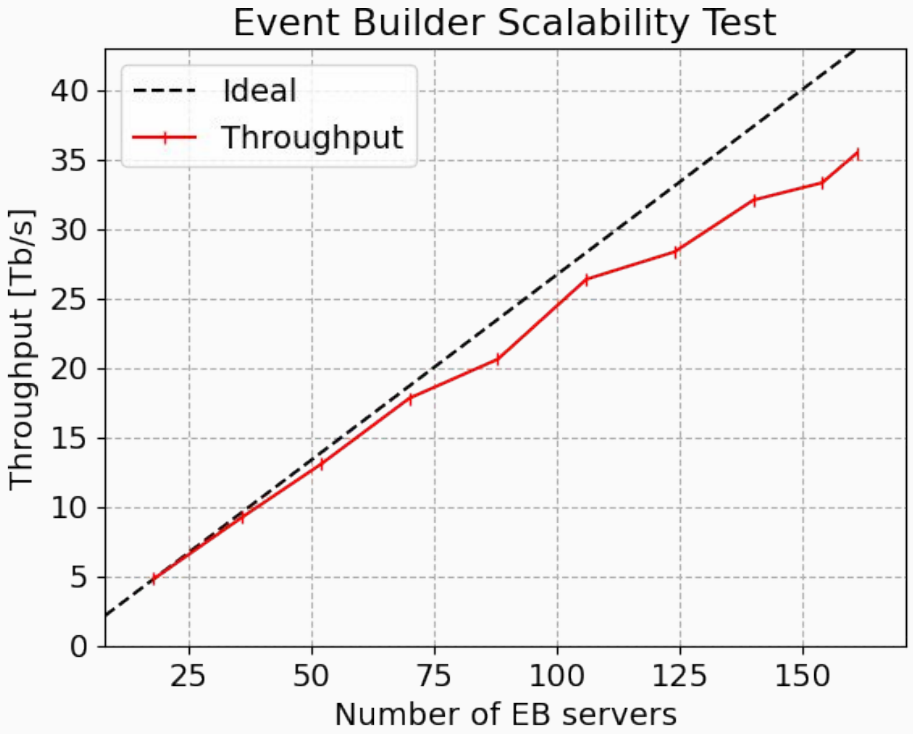
- The large amount of system memory available makes the system more robust against latency spikes.
- The DMA/RDMA architecture reduces the memory throughput required.

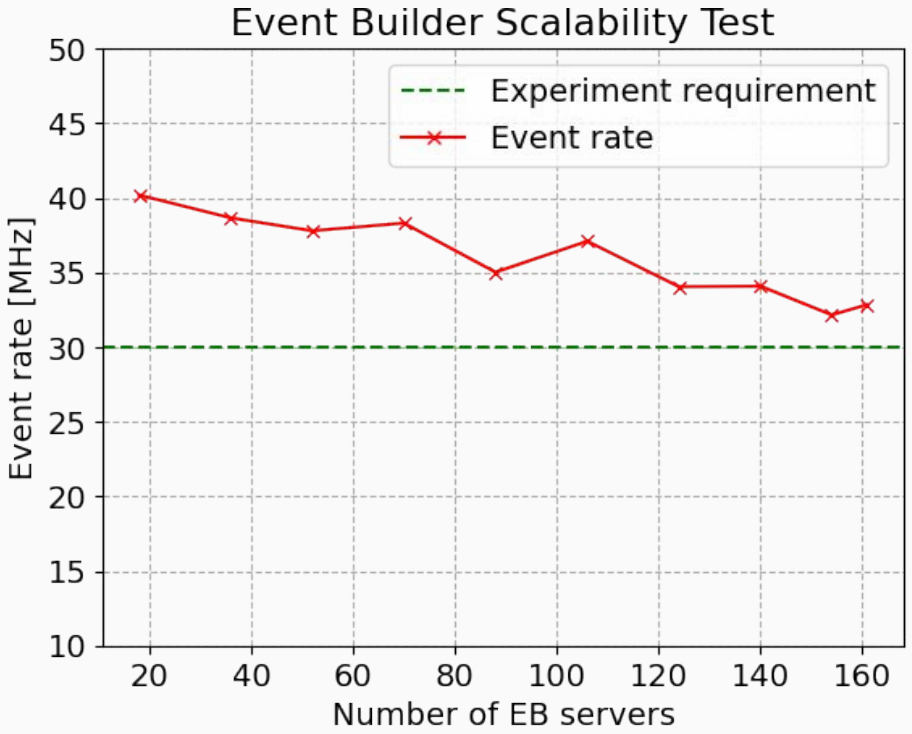
Backpressure and fault tolerance



- Buffers and discard policies allow reduce as much as possible the backpressure propagation
- In case of HW issues on a specific node it is possible to move only the RU/BU functionality to another different server

System scalability





Pros and Cons after the first year

- ✓ We removed the physics inefficiencies introduced by the HW trigger
- ✓ The system is resilient against network latency spikes
- ✓ We used off-the-shelf components to reduce cost
- ✓ The converged architecture significantly reduces costs

- ✘ The HW trigger provides non-physics functionality
- ✘ Special tuning to optimise the PCIe communications
- ✘ Highly converged architectures are less flexible and reliable

Conclusions

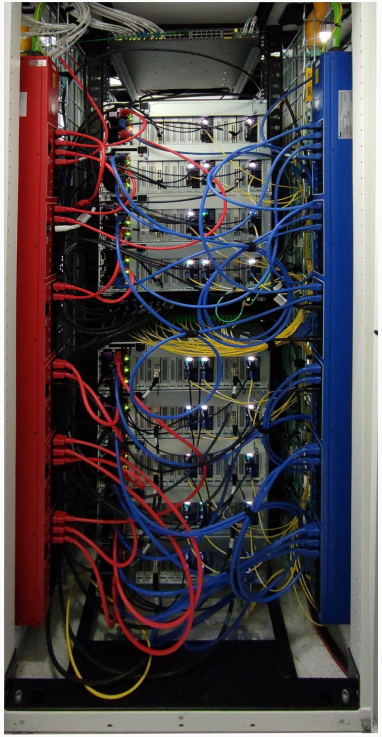
- The LHCb experiment has been upgraded to perform a full read-out at the bunch-crossing rate
- The design uses as much as possible off-the-shelf parts
- The system has been optimised to be less sensitive to latency spikes
- The DAQ system can sustain the load of 32 Tb/s
- The system has been successfully used for the first part of Run3

THANK YOU FOR YOUR ATTENTION

BACKUP

It exists!

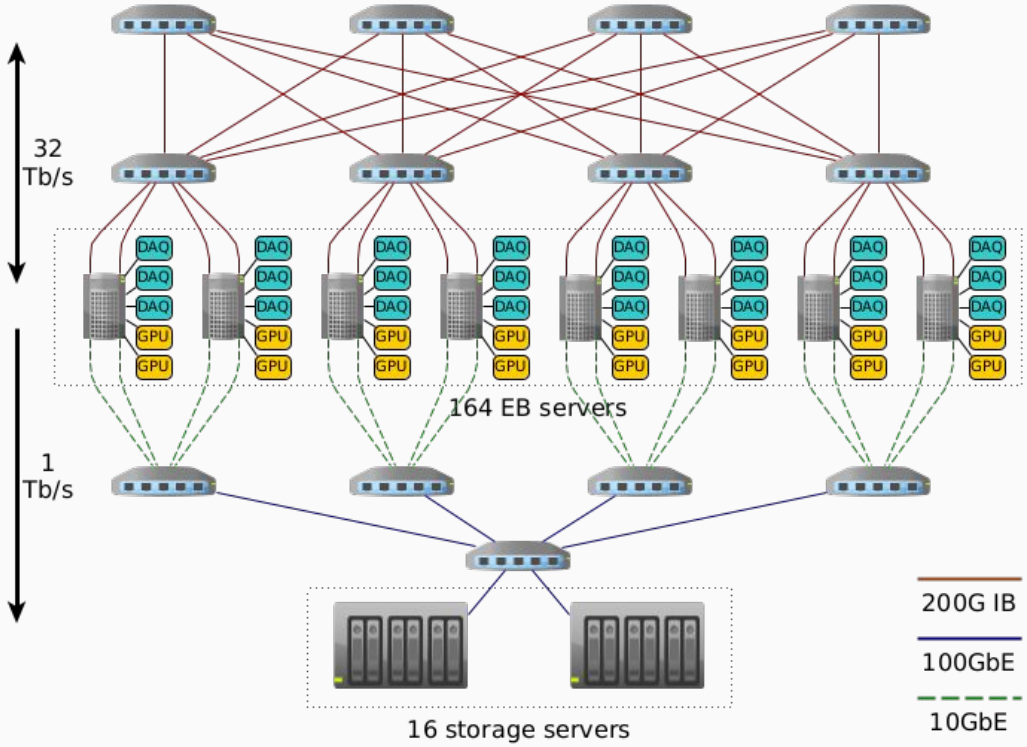
- 163 EB servers
- 24 racks: 18 EB, 2 control, 4 storage
- 28 40-port IB HDR switches: 18 leaf and 10 spine



EB rack

IB spine switches

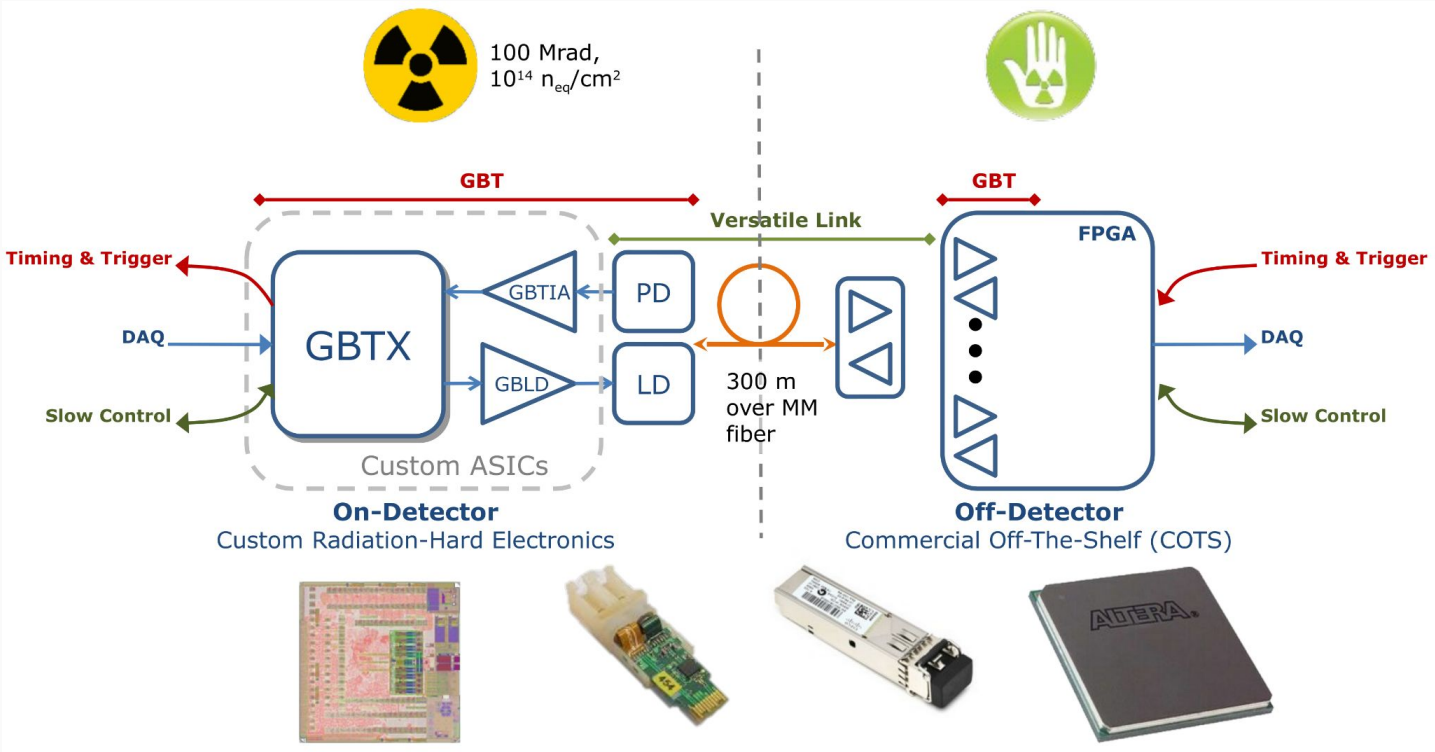
Network architecture



The PCIe40: a single custom-made FPGA board for DAQ and Control

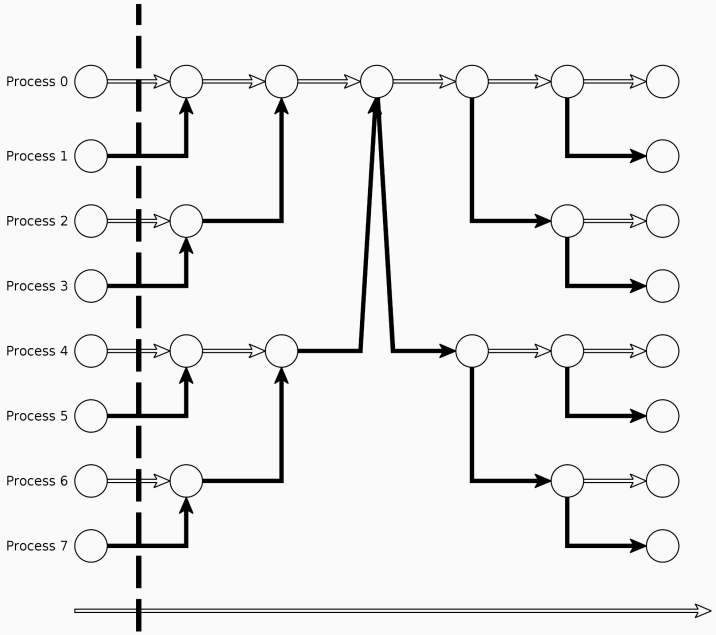
- Based on Intel Arria10
- 48x10G capable transceiver on 8xMPO for up to 48 full-duplex Versatile Links
- 2 dedicated 10G SFP+ for timing distribution
- 2x8 Gen3 PCIe
- Efficient and accurate software trigger that can perform online selection with offline-like quality
- One card multiple multiple FW personalities:
  - Readout Supervisor (SODIN)
  - Interface Board (SOL40)
  - DAQ card (TELL40)

Versatile link / GBT



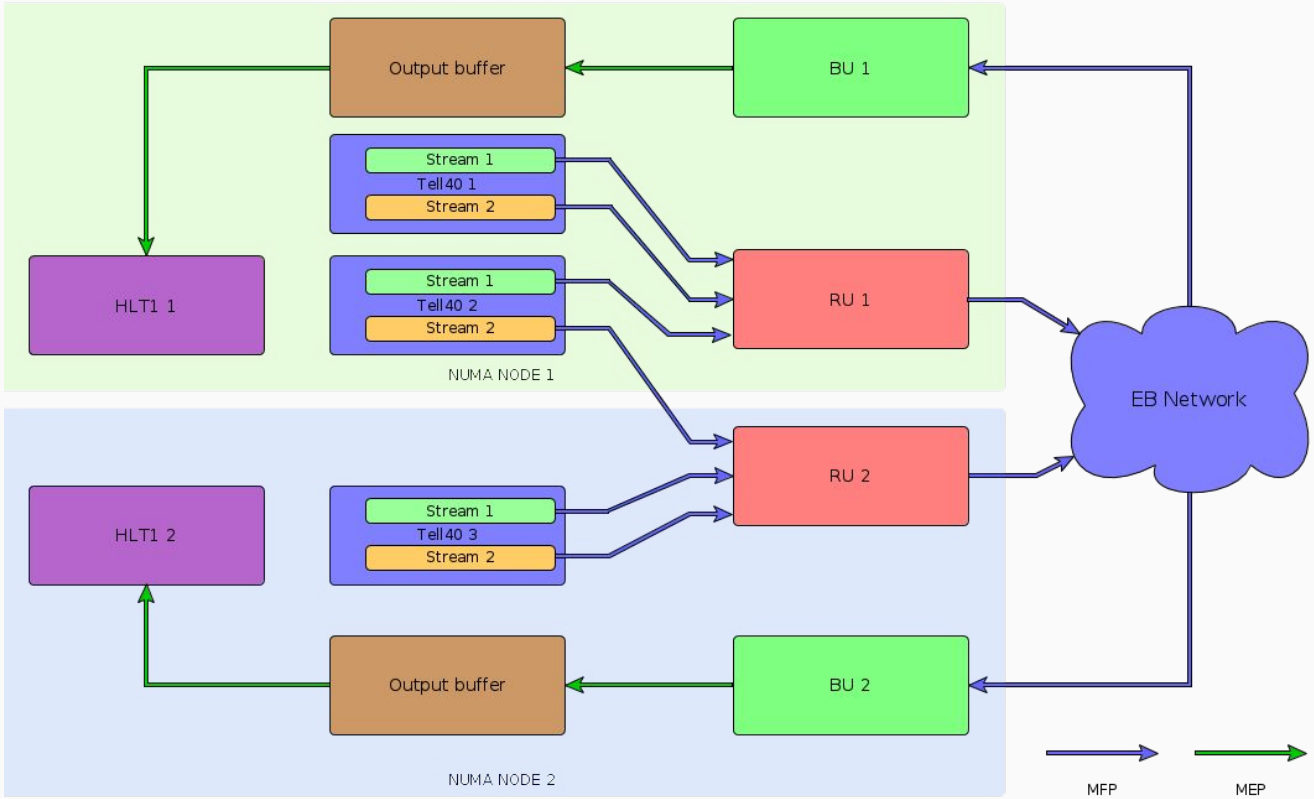
Credit: P. Moreira, S. Baron (CERN)

Barrier synchronization

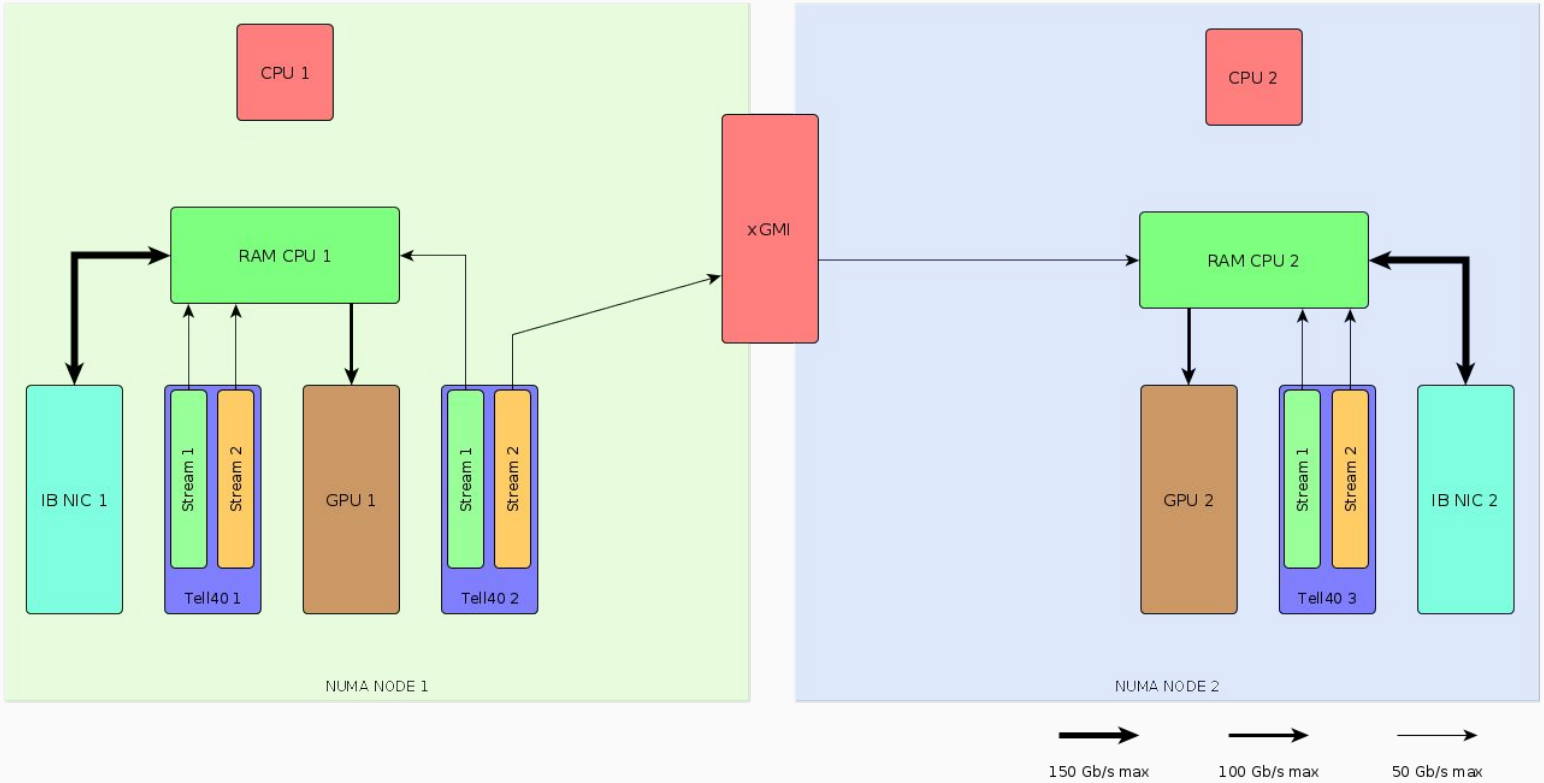


Distributed tree barrier

EB server data flow



EB hardware layout

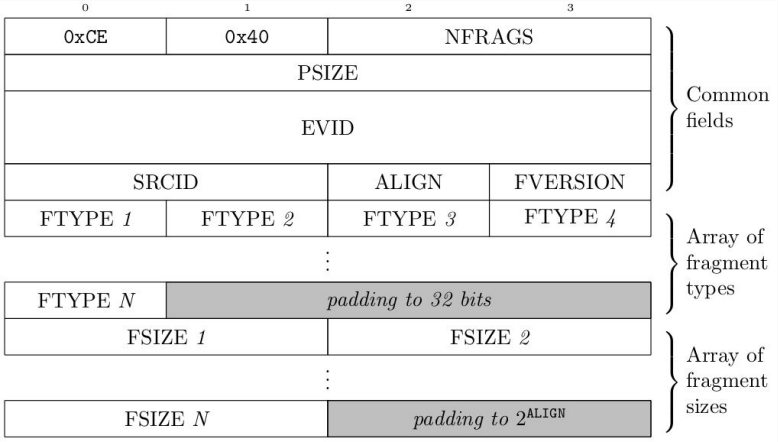


Event size model

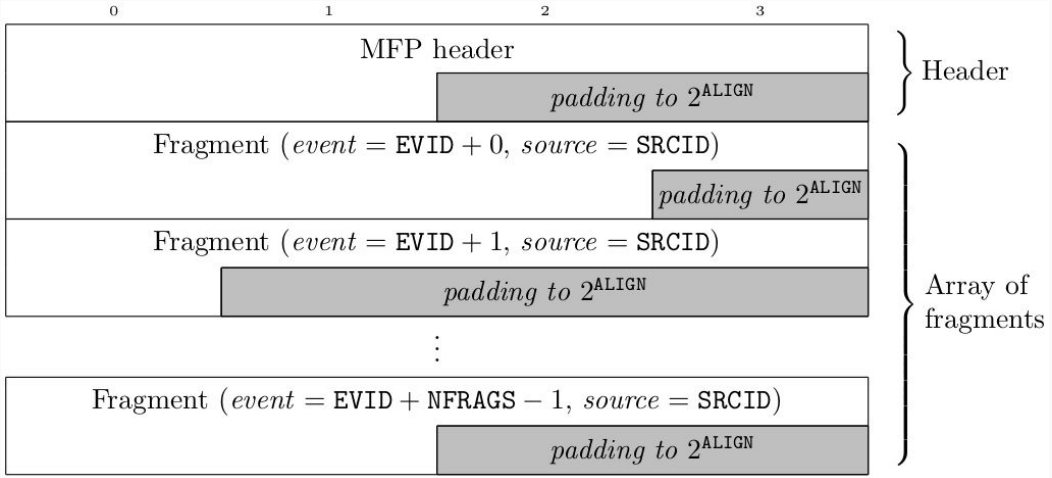
Sub-detector	fragment size [B]	#tell40 streams	event size [B]	event fraction	MEP size [GB]	MFP size [MB]	RU send size [MB]
Velo	156	104	16250	0.13	0.49	4.69	14.06
UT	100	200	20000	0.16	0.60	3.00	9.00
SCIFI	100	288	28800	0.23	0.86	3.00	9.00
Rich 1	166	132	22000	0.18	0.66	5.00	15.00
Rich 2	166	72	12000	0.10	0.36	5.00	15.00
Calo	156	104	16250	0.13	0.49	4.69	14.06
Muon	156	56	8750	0.07	0.26	4.69	14.06
Total	1000	956	124050	1	3.72	30.06	90.19

Multiple Fragment Packet (MFP)

MFP header

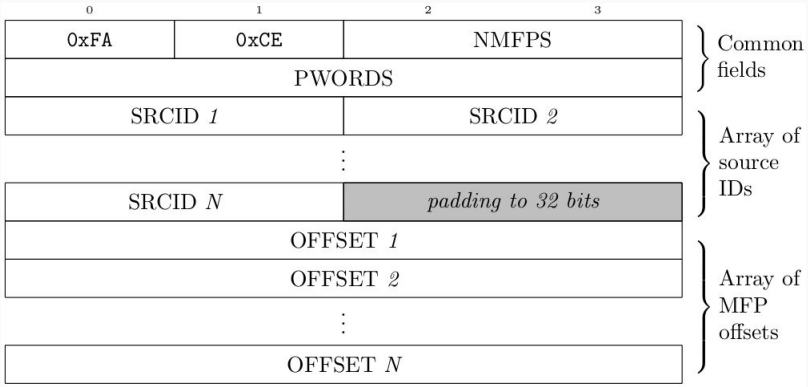


MFP



Multiple Event Packet (MEP)

MEP header



MEP