Design and Commissioning of the first 32 Tbit/s event-builder

23rd IEEE Real Time Conference
03/08/2022

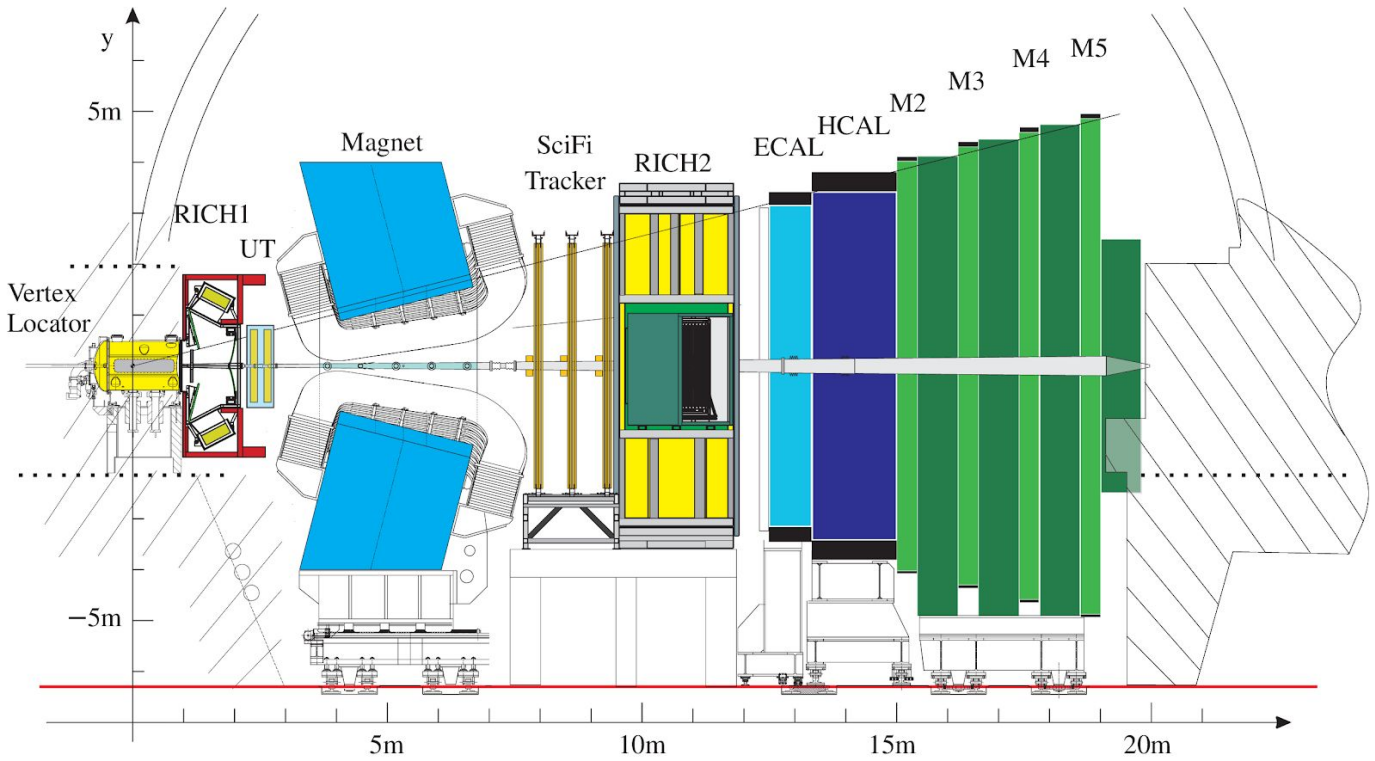
Flavio Pisani for the LHCb Online team

CERN

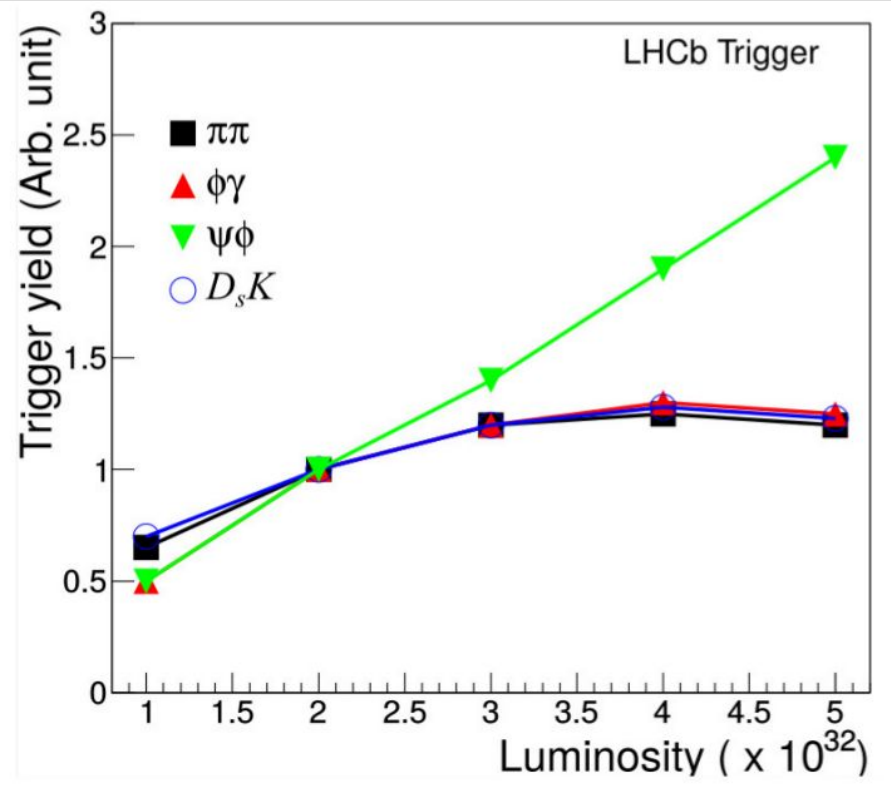
Outline

- Introduction to the LHCb experiment
- Moving towards full collision-rate readout
- EB design process
- EB hardware layout
- Standalone EB performance testing and optimization
- First full DAQ test

The LHCb experiment



Full collision-rate readout: why?



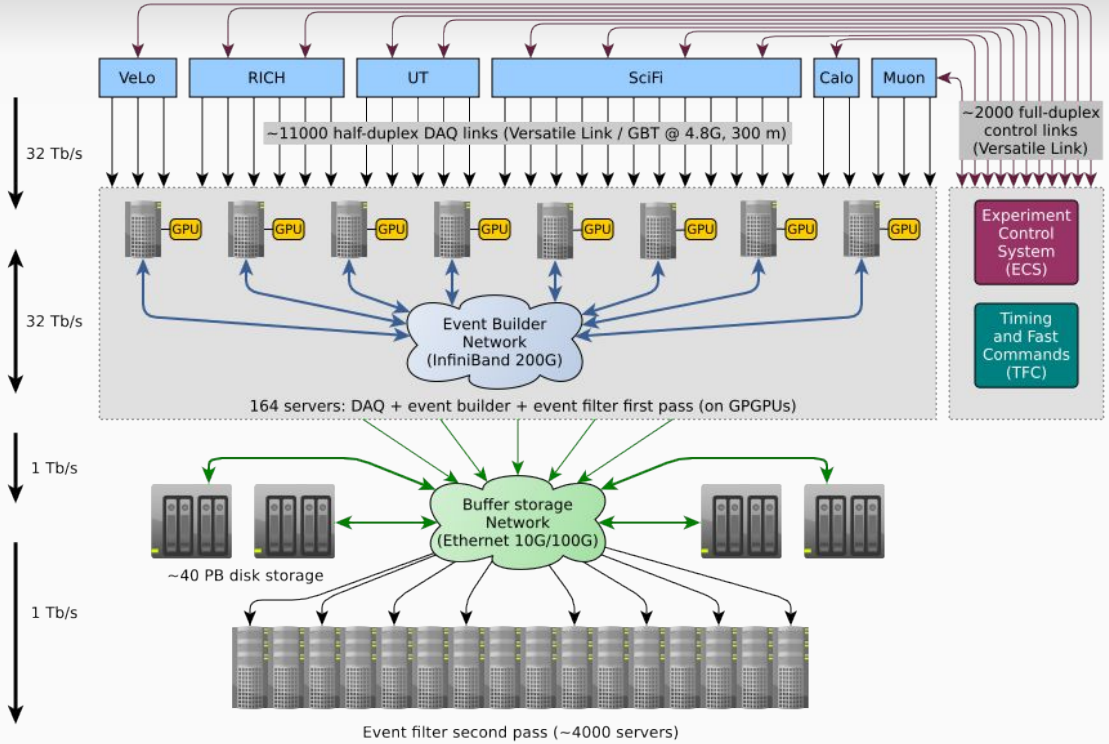
- The low-level trigger saturates in hadronic channels
- The instantaneous luminosity in Run 3 will go up to $2x10^{33}$ cm^{-2} s^{-1}

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

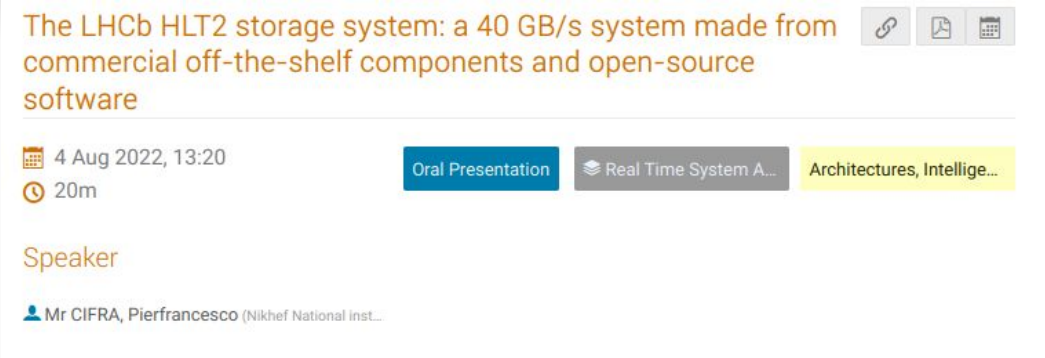
Full collision-rate readout: how?

- Spectrometer geometry: more space to route cable/fibers
- Zero-suppression on the detectors
- Relatively low radiation levels permit to relax the FPGA/ASIC constraints
- Comparatively small event-size (O 100 kB)
- Efficient and accurate software trigger that can perform online selection with offline-like quality

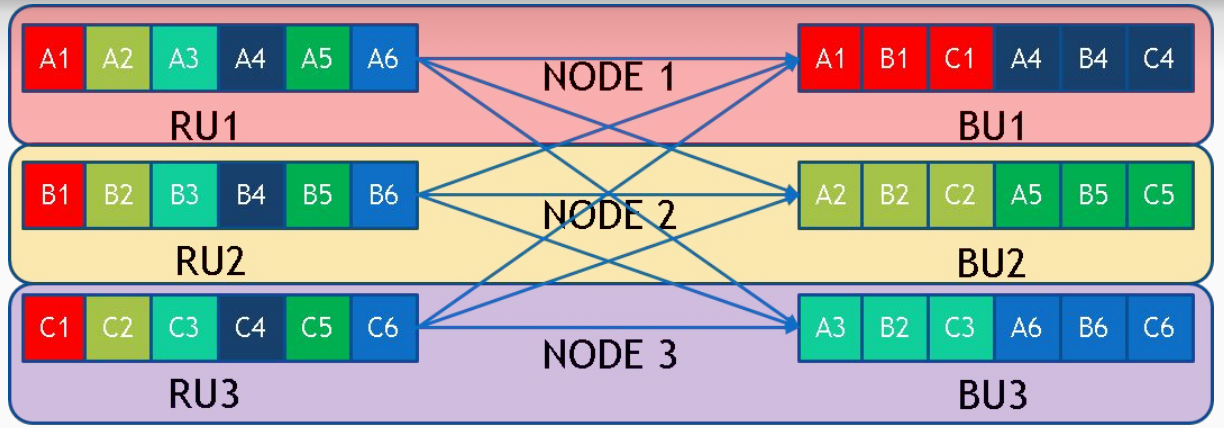
Online DAQ system overview



Do you want more details?



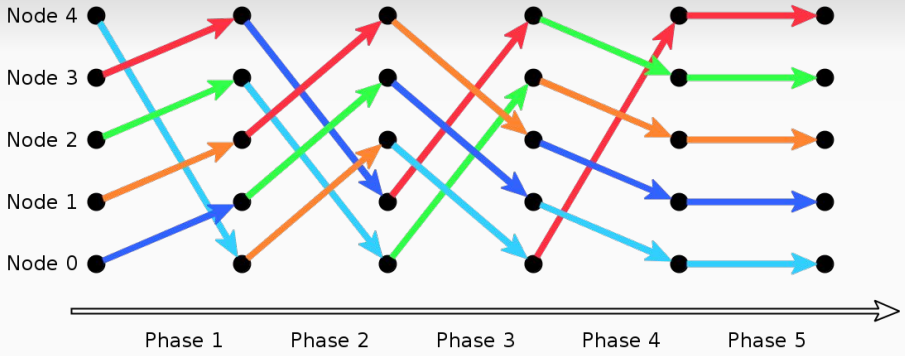
The Event Building process in a nutshell

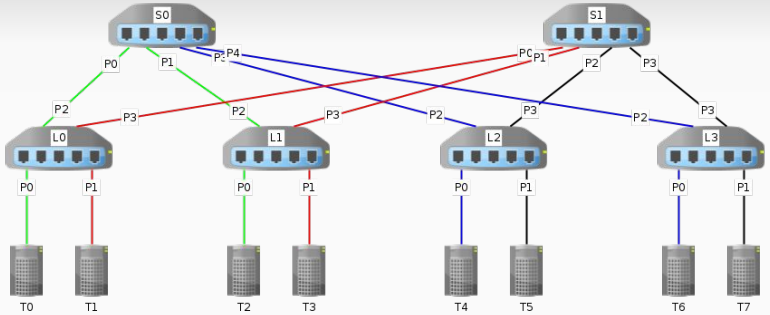


- Every event is divided into multiple fragments
- Every **Readout Unit (RU)** receives a fragment of the event
- Every **Builder Unit (BU)** has to gather all the fragments of the event

The many-to-one nature of the traffic generates network congestion

Traffic scheduling



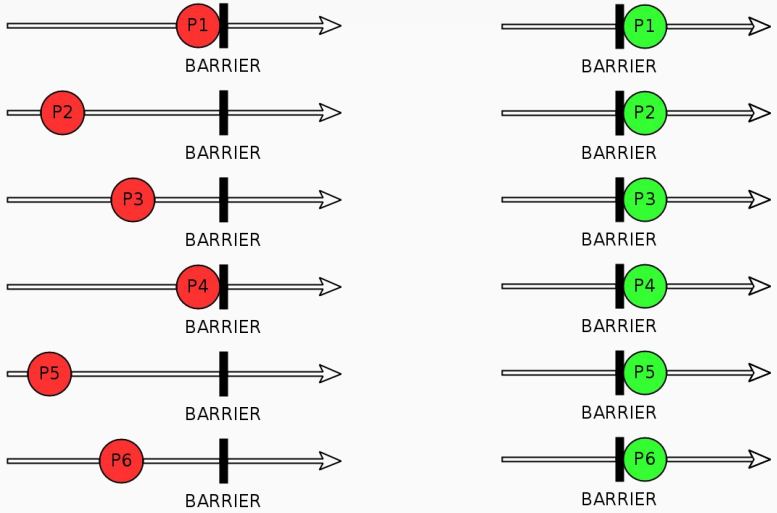


- The processing of **N** events is divided into **N** phases
- In every phase one RU sends data to one BU, and every BU receives data from one RU
- During phase **n** RU **x** sends data to BU (**x** + **n**)%**N**
- All the units switch synchronously from phase **n** to phase **n** + 1

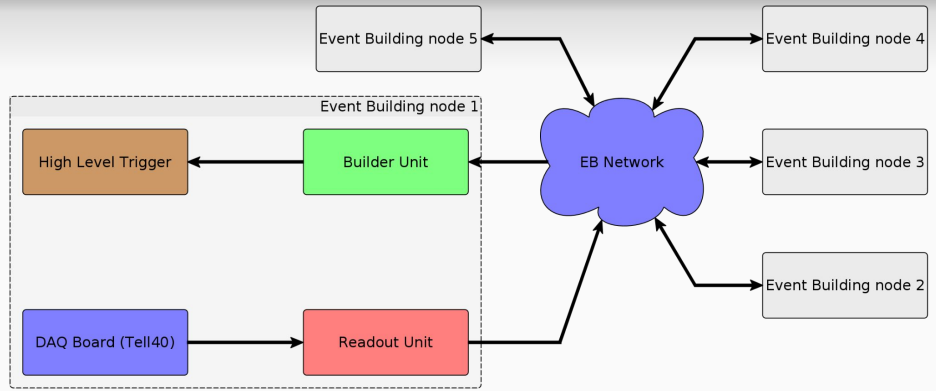
Congestion-free traffic on “selected networks” (e.g. fat-tree networks)

Barrier synchronization

- Strong synchronization imposed at every step of the scheduling
- In-band synchronization barrier
- Two step process:
  - The processes *report* they reached the barrier
  - The processes are *woken-up* when all have reached the barrier
- Centralized master/minion communication
- Distributed tree-like communication

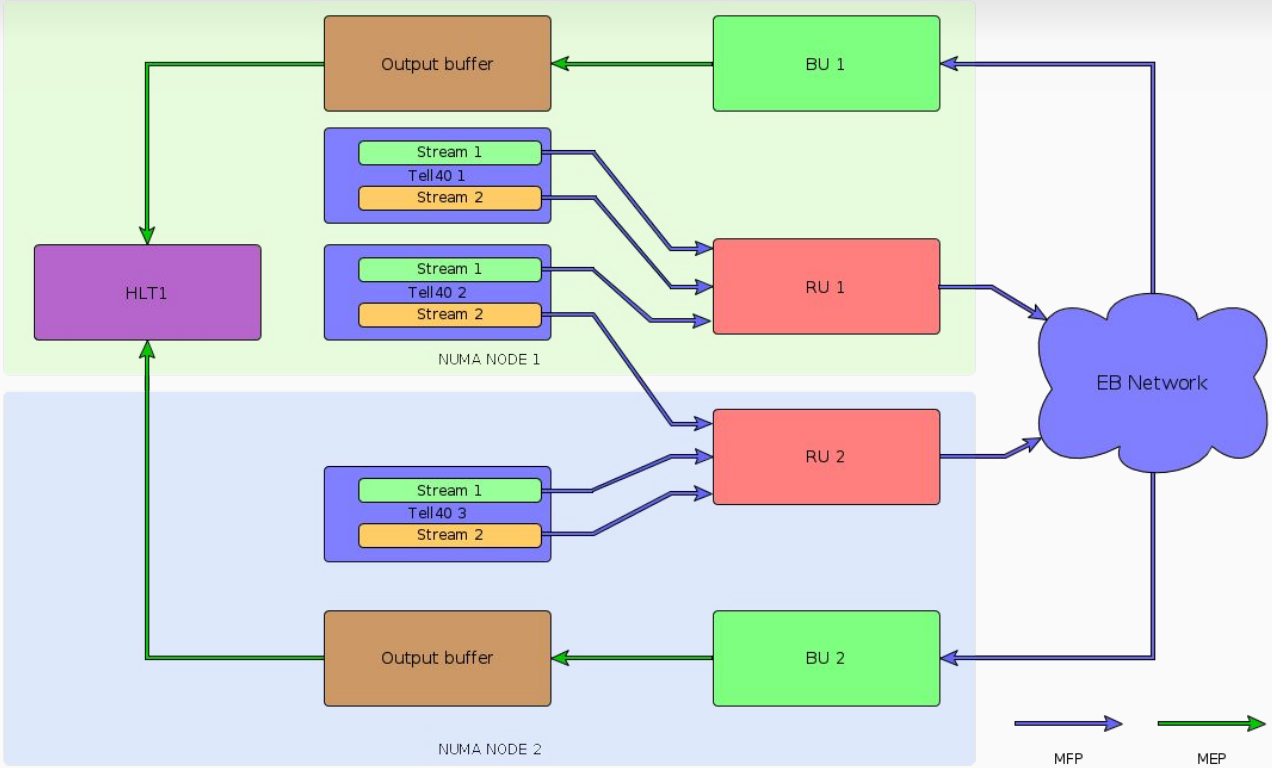


Software architecture

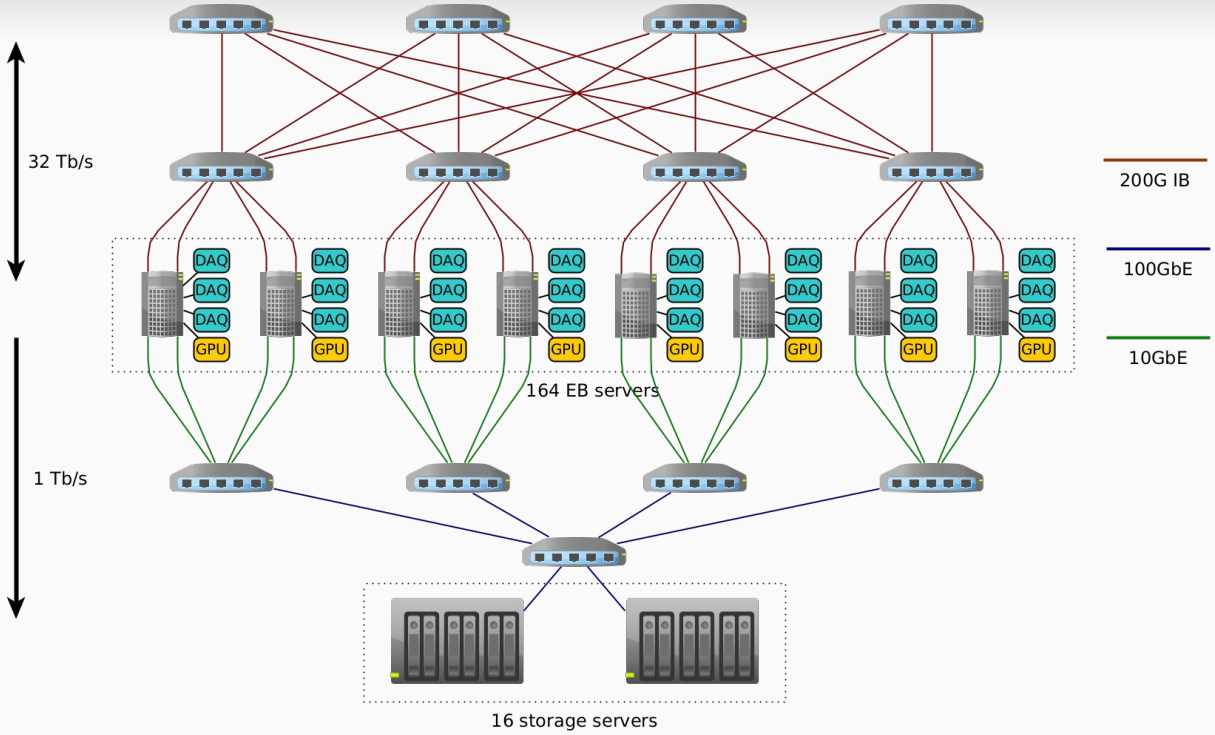


- Modular software architecture build in C++
- RU: it reads the data from the DAQ card and sends it over the EB network
- BU: it reads the data the EB network and it writes the built data into the HLT1 input buffer
- The scheduling synchronization is achieved using a barrier
- Dedicated low-level communication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

Event Builder server data flow

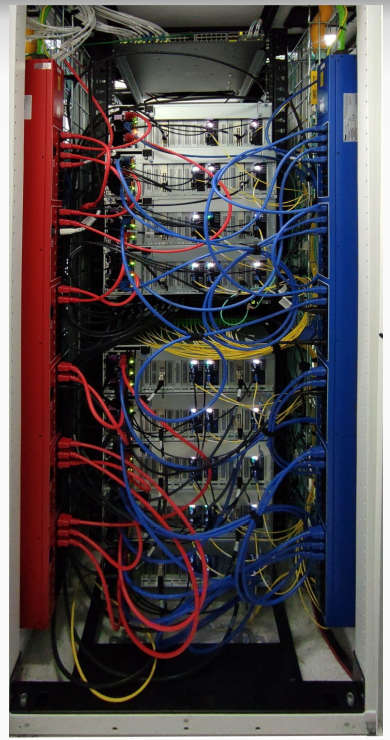


Network architecture



It exists!

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



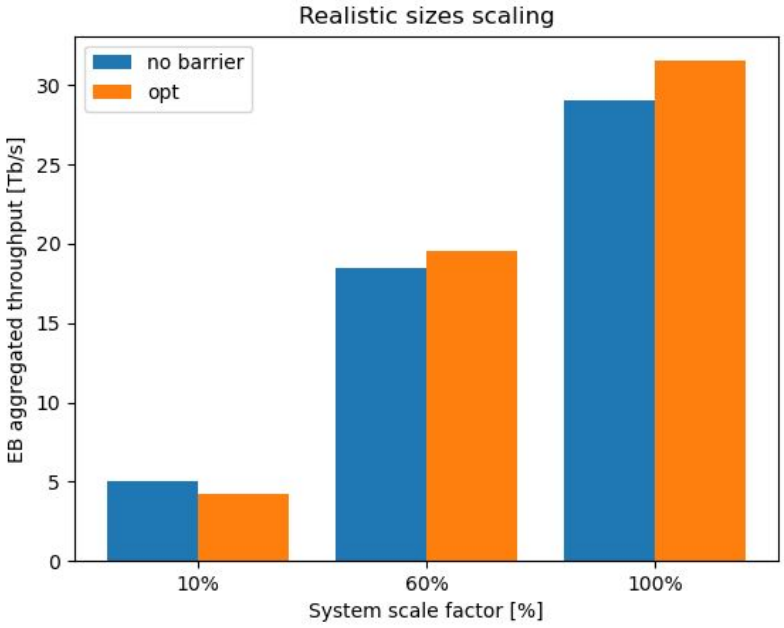
EB rack

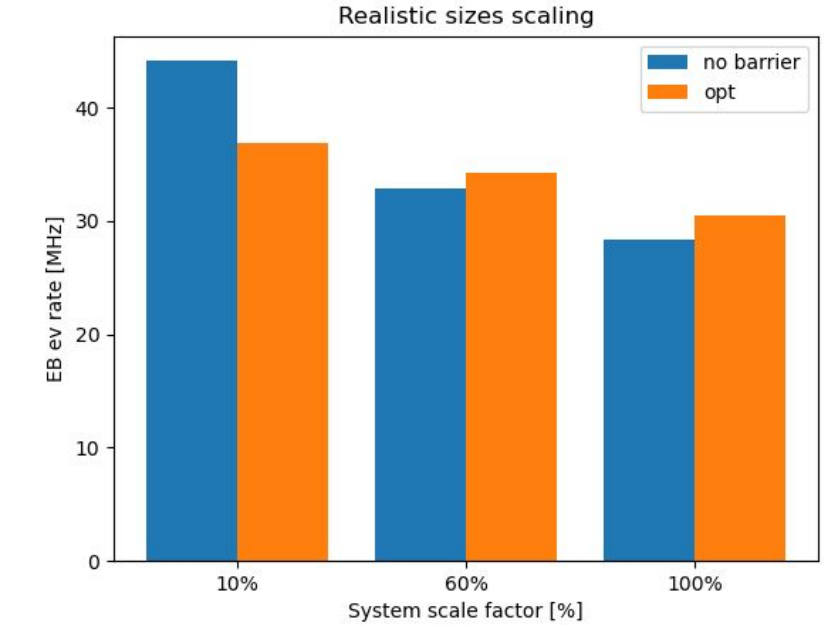
IB spine switches

Testing conditions

- Independent test of the EB in standalone mode
- Injected data using a CPU data generator
- Realistic event-size model based on MC simulations and inputs from the various sub-detectors
- Possibility to tune and optimize multiple parameters (buffer sizes, synchronization algorithm, ecc)

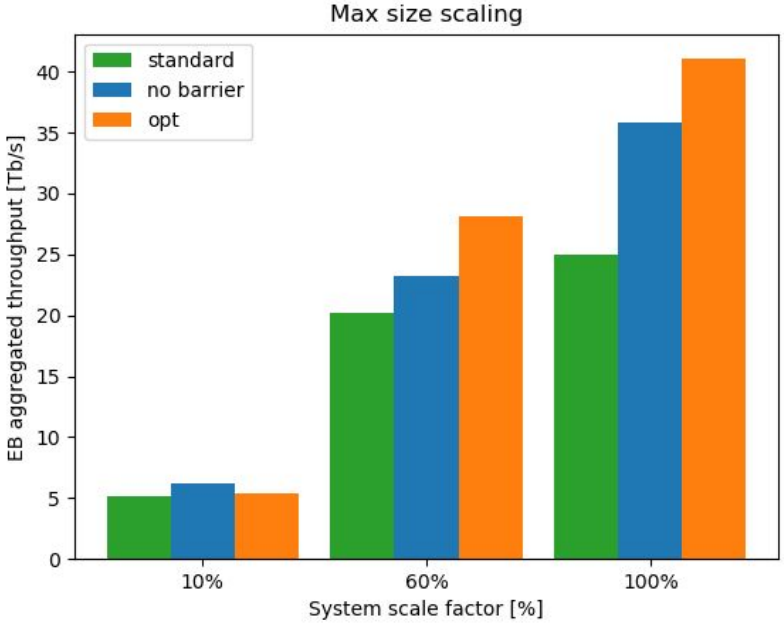
Performance testing

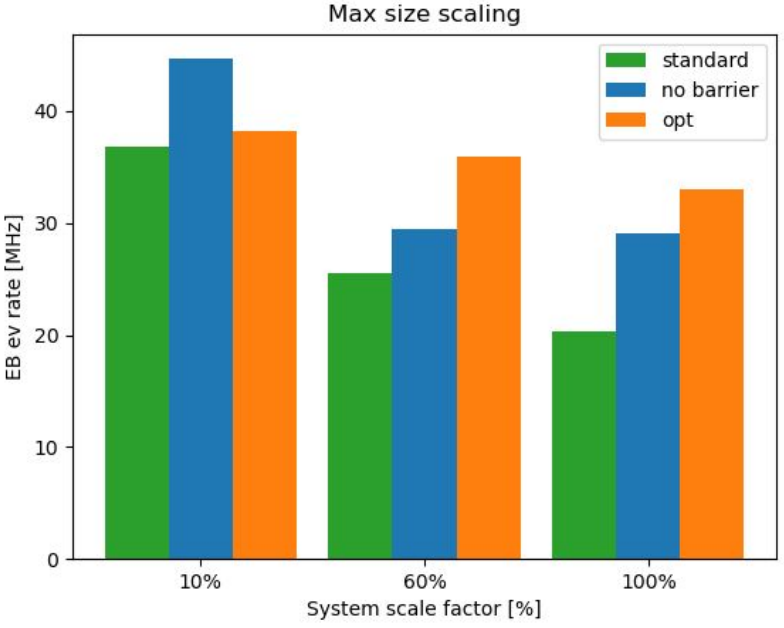




We need to keep strong synchronization in the scheduling

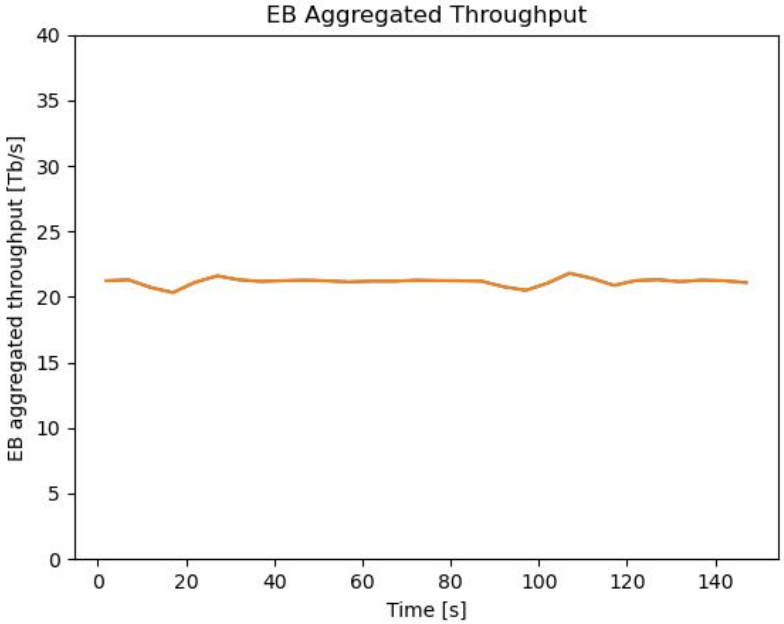
Performance testing

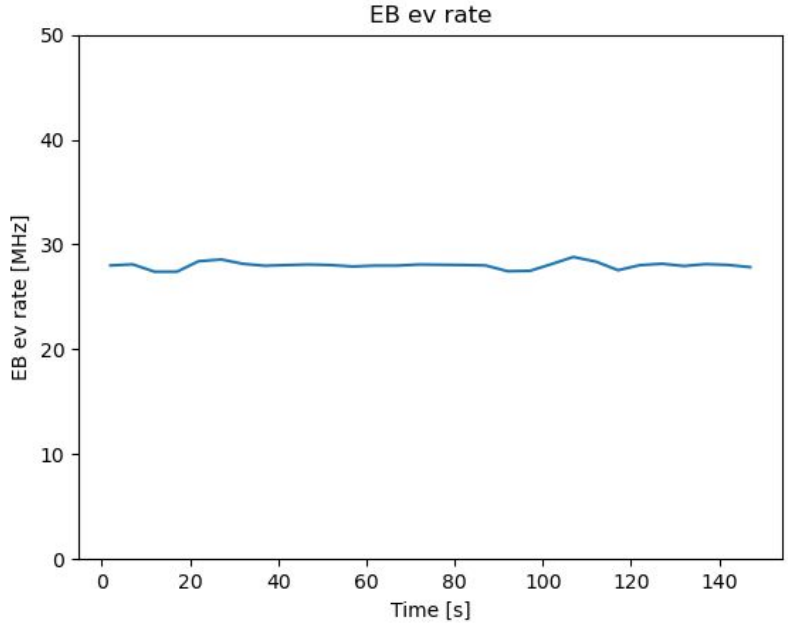




The optimization the the synchronization algorithm is crucial

Full system DAQ results





Conclusions

- The LHCb experiment has been upgraded to perform a full read-out at the bunch-crossing rate
- The EB has been designed around an off-the-shelf network technology (IB HDR)
- The EB software has been implemented and optimized to meet the performance requirements
- Testing with a software data generator show that the EB fulfills all the requirements
- Running the actual DAQ system the EB can forward the full beam-beam rate
- Optimizations are in progress both in the EB SW and on the Tell40 FW

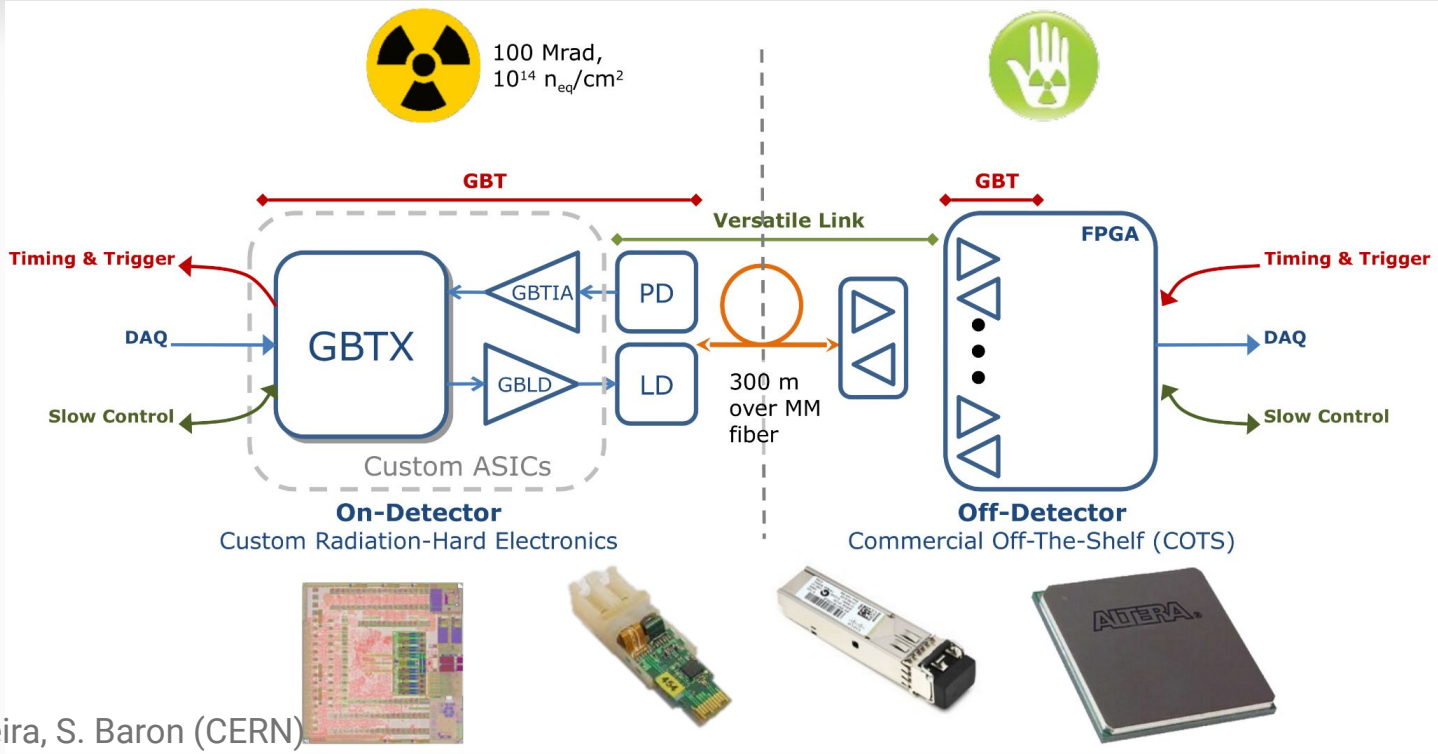
THANK YOU FOR YOUR ATTENTION

BACKUP

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

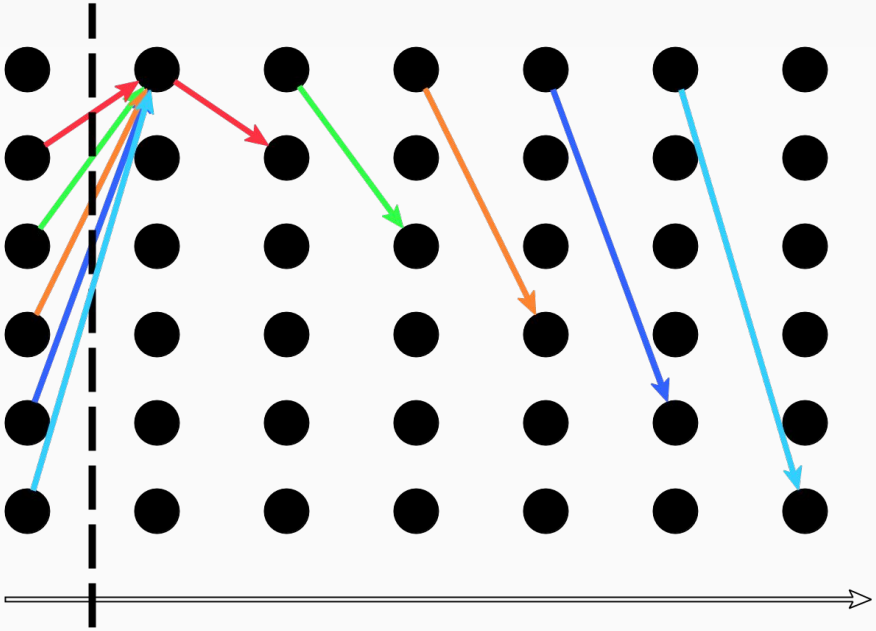


Credit: P. Moreira, S. Baron (CERN)

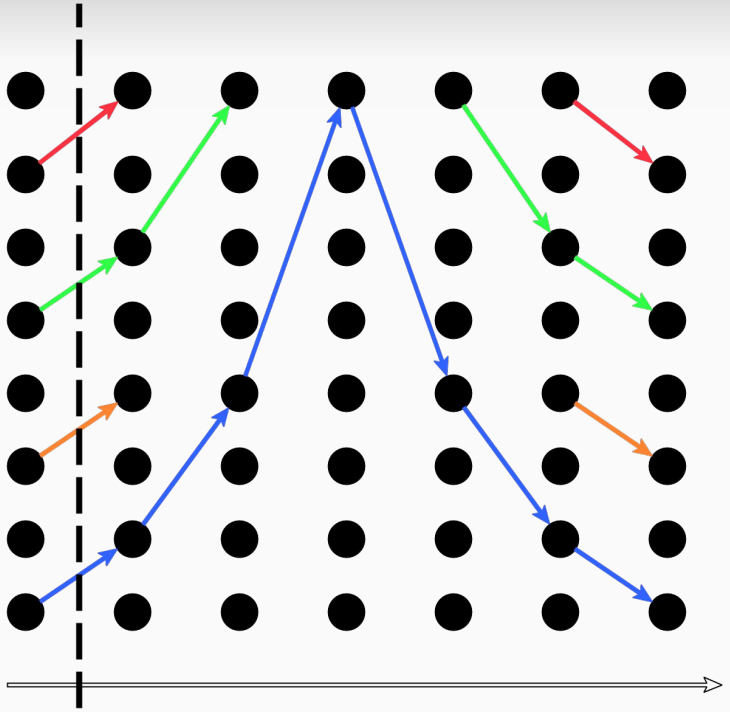
InfiniBuilder communication library

- IBVerbs custom communication library
- Designed to be an almost drop in replacement for MPI calls
- Support to communication collectives (Barrier, Gather, Scatter, ecc)
- Out-of-band setup over TCP/IP with process identification and assignment of unique IDs
- Efficient memory registration to optimize performance and warmup time

Barrier synchronization

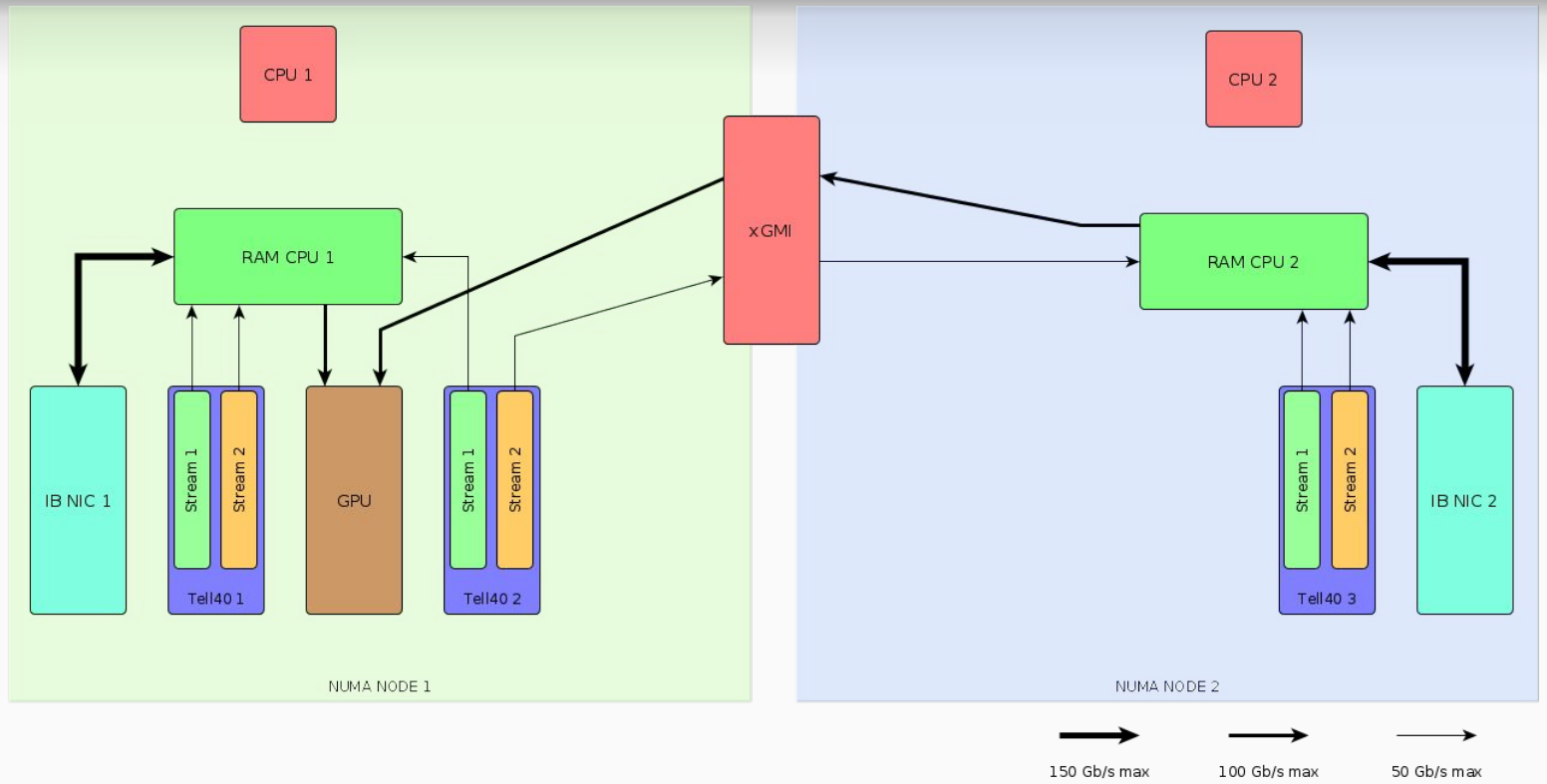


Centralized barrier

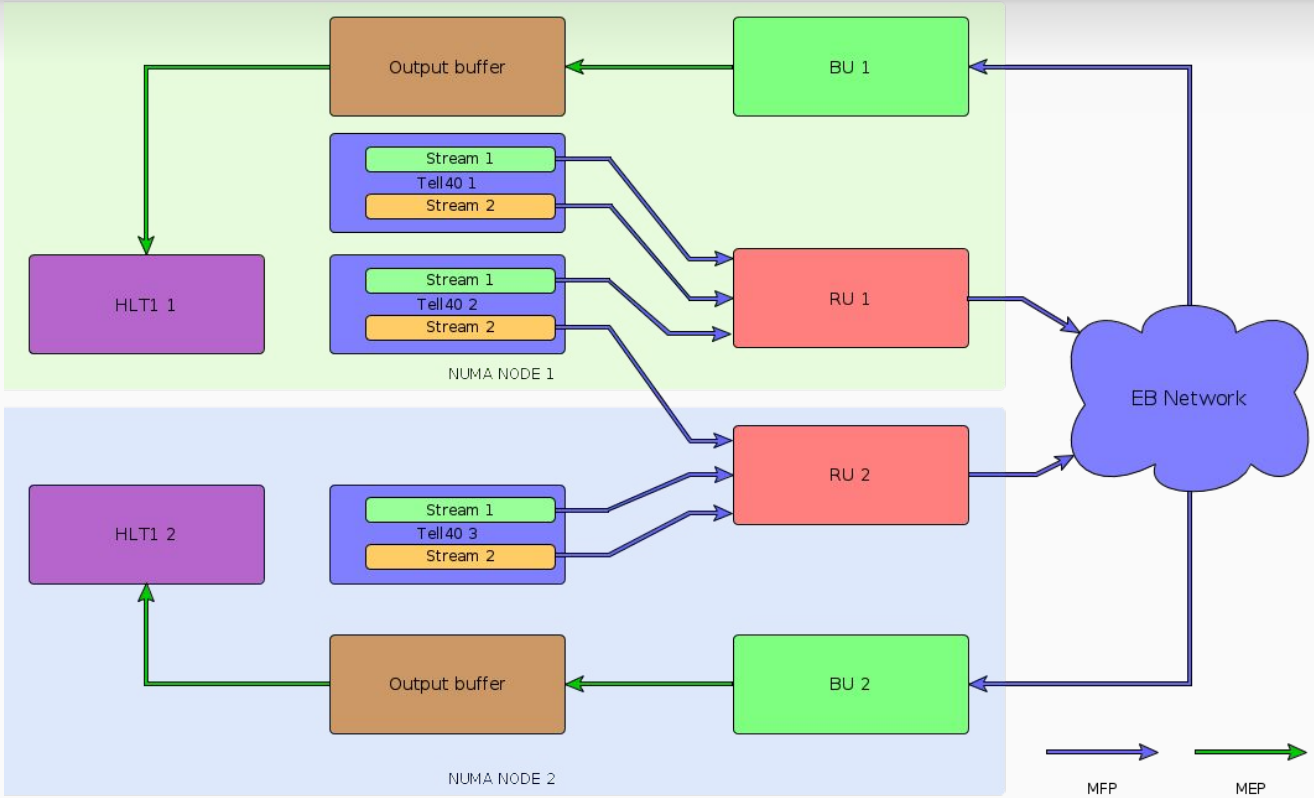


Distributed tree barrier

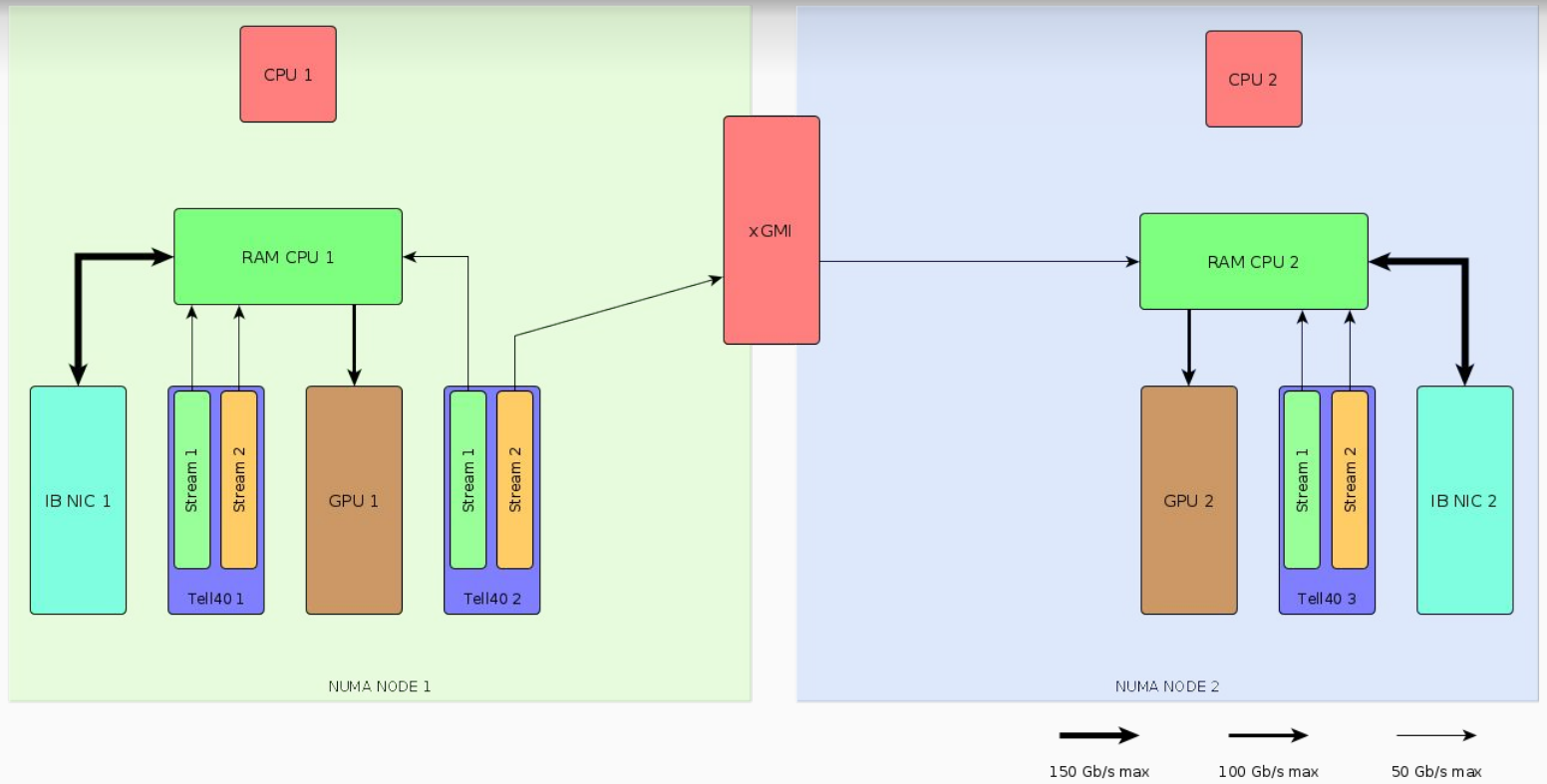
EB hardware layout



EB server data flow multi GPU option



EB hardware layout multi GPU option

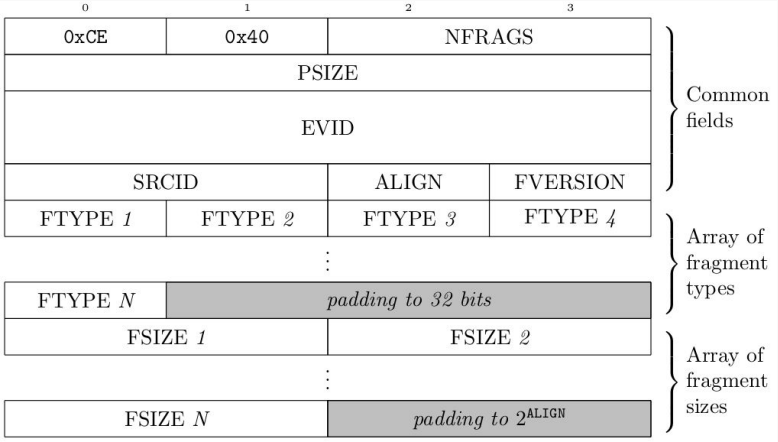


Event size model

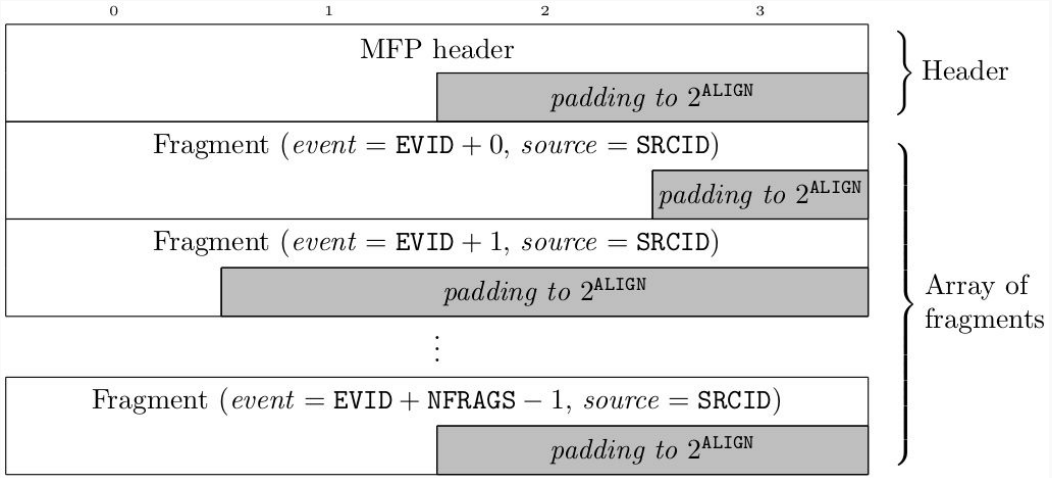
Sub-detector	fragment size [B]	#tell40 streams	event size [B]	event fraction	MEP size [GB]	MFP size [MB]	RU send size [MB]
Velo	156	104	16250	0.13	0.49	4.69	14.06
UT	100	200	20000	0.16	0.60	3.00	9.00
SCIFI	100	288	28800	0.23	0.86	3.00	9.00
Rich 1	166	132	22000	0.18	0.66	5.00	15.00
Rich 2	166	72	12000	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

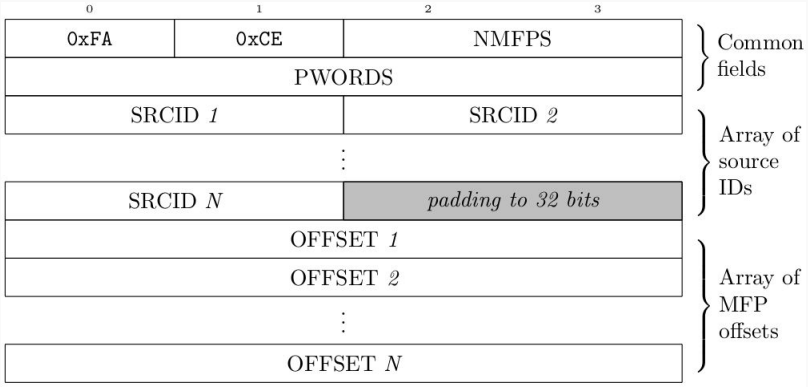


MFP



Multiple Event Packet (MEP)

MEP header



MEP