# Large-scale DAQ tests for the LHCb upgrade

Antonio Falabella

INFN - CNAF (Bologna)

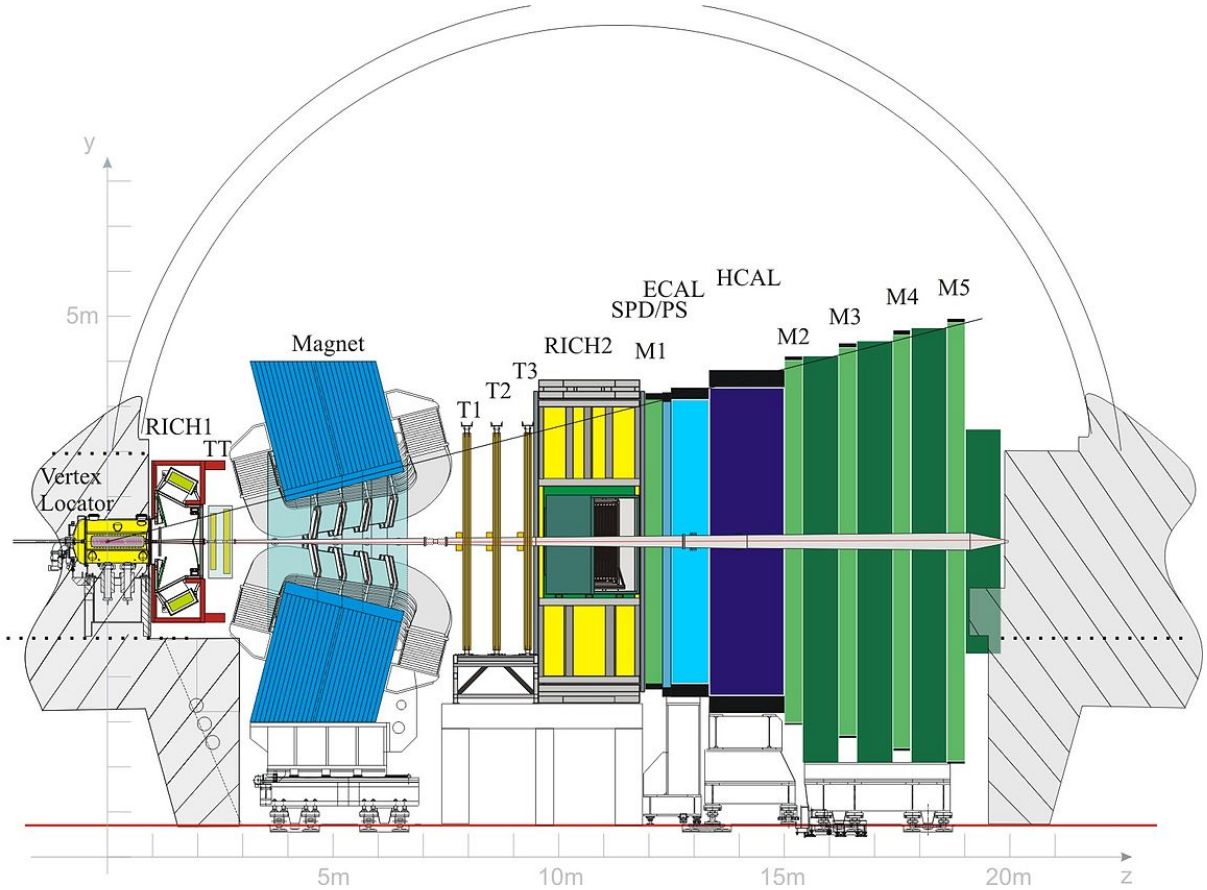
NPSS 20th *Real Time Conference*

5-10 June - Padova, Italy

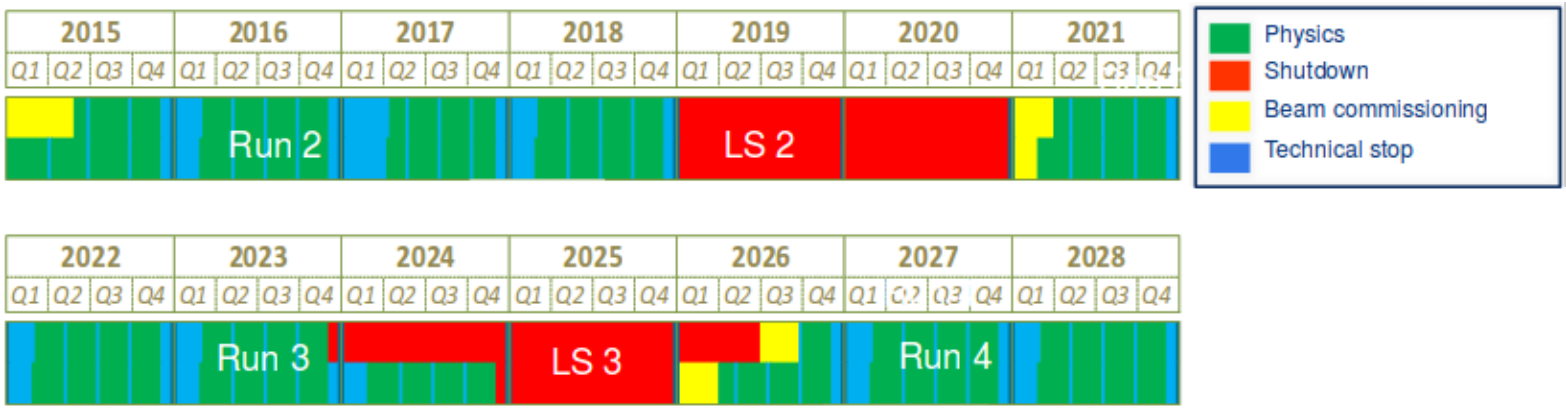
# Overview

# The LHCb experiment

- LHCb is a flavour physics experiment that aims to study CP violation parameters at very high precision and studying rare decays of $b$ and $c$ hadrons
- The LHCb detector provide high vertex resolution, tracking precision and particle discrimination capability



# Data taking schedule



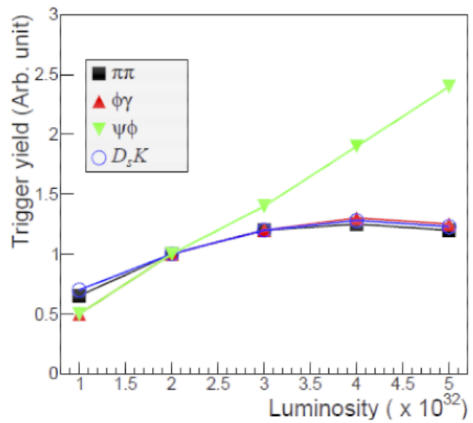

## Run II

- $1.5 \cdot 10^{34} cm^{-2} s^{-1}$
- Bunch separation $25ns$
- Pileup $\sim 40$

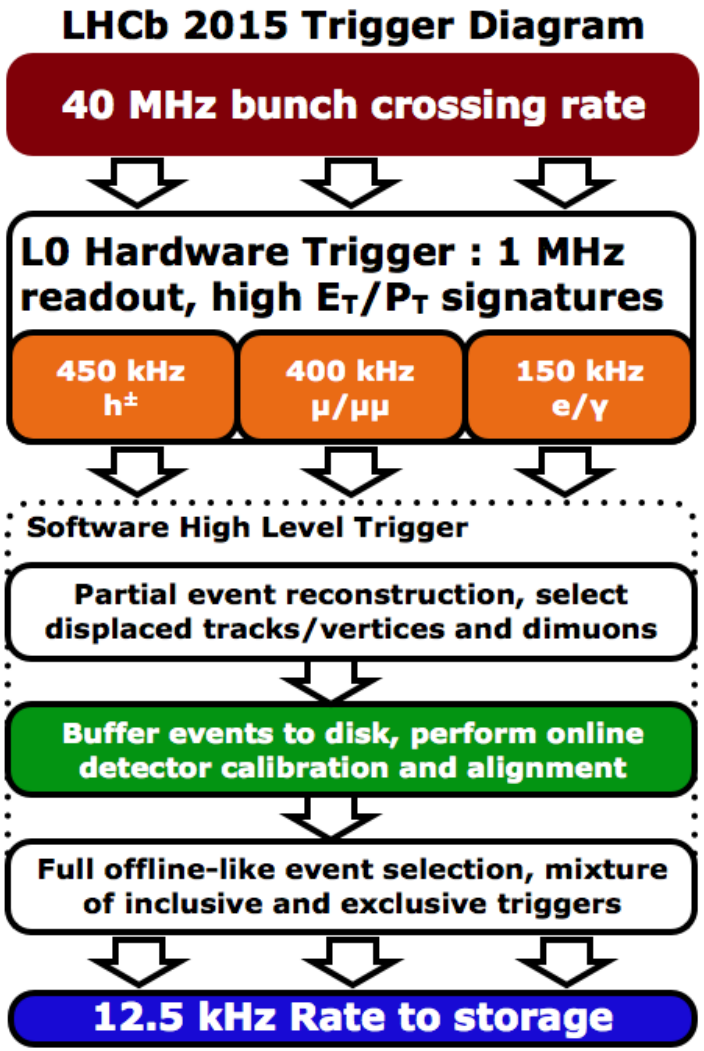
## Run III

- $2.2 \cdot 10^{34} cm^{-2} s^{-1}$
- Bunch separation $25ns$
- Pileup $\sim 60$

- The precision of several measurements will be limited by amount of collected data
- a possible solution $\rightarrow$ increase the instantaneous luminosity
- The output of many hadronic trigger algorithms saturates

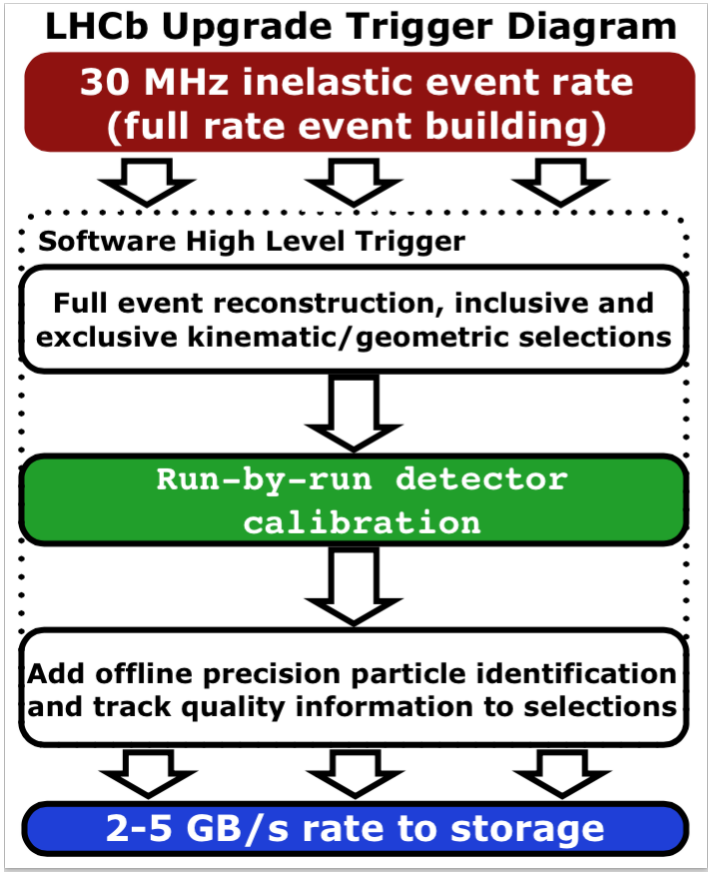


# Run II trigger configuration




- Run II trigger :
- L0 trigger : reduce the frequency from $40$ MHz to $1.1$ MHz
- A second stage software trigger reduce further the frequency to $12.5$ KHz

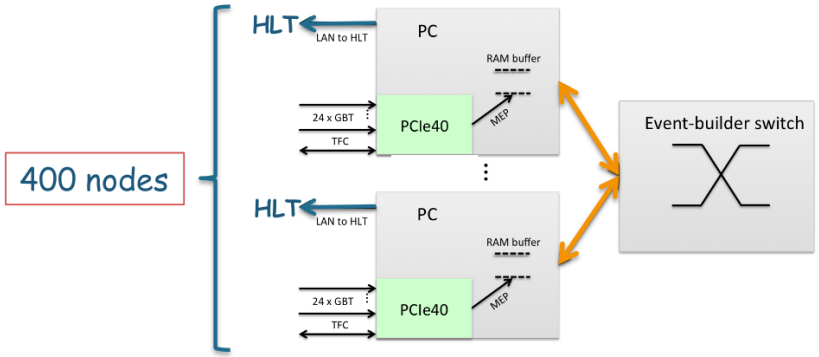
# Run III trigger configuration




- Run III trigger:
- No hardware trigger ($30$ MHz is the crossing frequency with non-empty events)
- Full software trigger
- Front-end electronic boards redesigned as long as the Event Builder

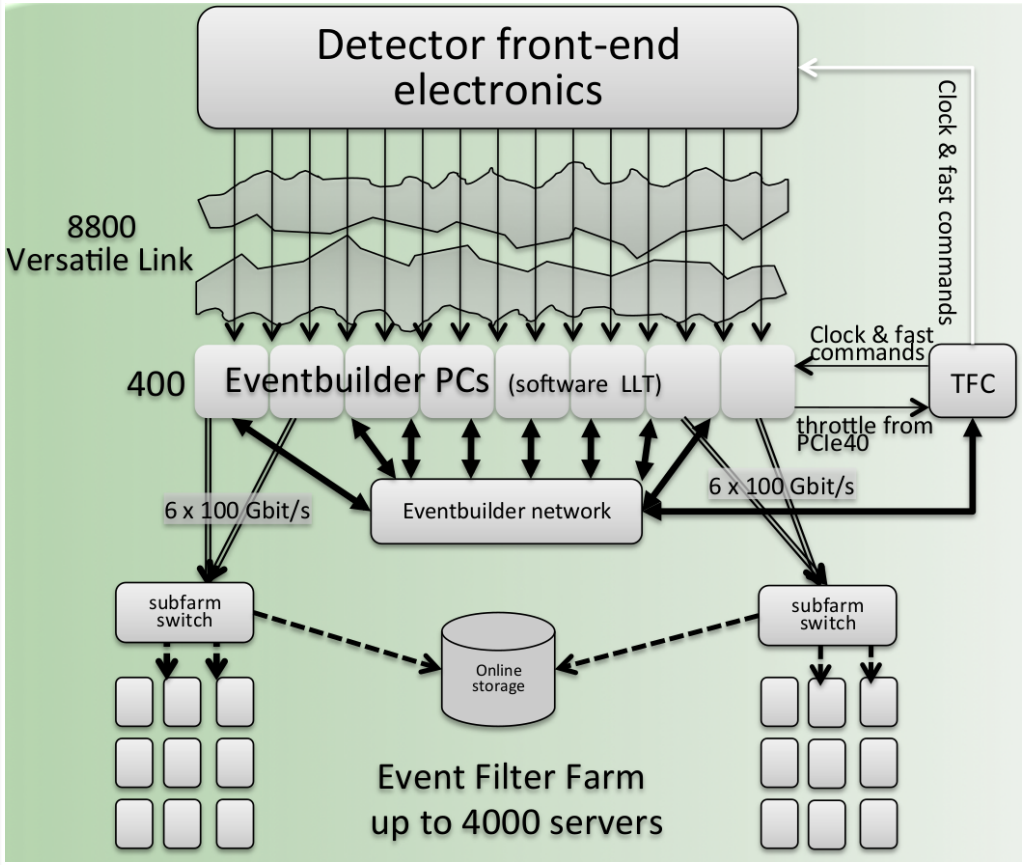
# Trigger-less data acquisition

- Detector read-out at $40\ \mathrm{MHz} \rightarrow$ event composition from subdetector fragments (*Event Building - EB*) at the same rate (*see also Paolo Durante's slides*)
- The EB operating at such an high frequency can be based on a high performance LAN




- Use PCIe v3 protocol to write the data fragments from the subdetectors directly to the EB node RAM (Multi Event Packet - MEP)

# Trigger-less data acquisition




- The EB LAN network size if of around $\sim 400$ nodes connected through an high-throughput network fabric
- Each node must be able to communicate at $\sim 100$ Gbit/s full-duplex
- The event can then be further processed before sending it to storage

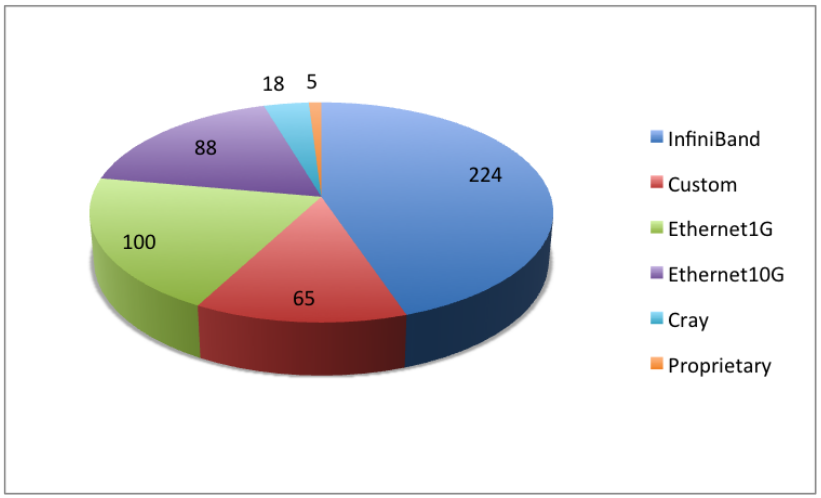
# DAQ implementation

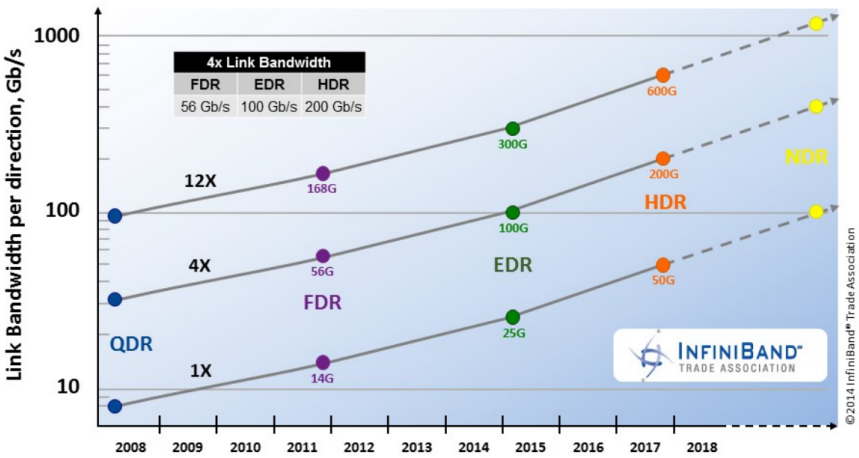
- The challenge is to handle an aggregated network traffic of $\sim 30$ Tbit/s (event frequency $\times$ event size: $30$ MHz $\cdot$ $100$ KByte)

| Event frequency | 30 MHz |
|---|---|
| Average event size | 100 KBytes |
| I/O of single node | 100 Gbit/s (full-duplex) |

- The EB LAN can be realized with off-the-shelf technologies
- InfiniBand, 100 Gigabit Ethernet, Omni-Path
- In this talk I will present the relevant results for the InfiniBand case

# InfiniBand standard

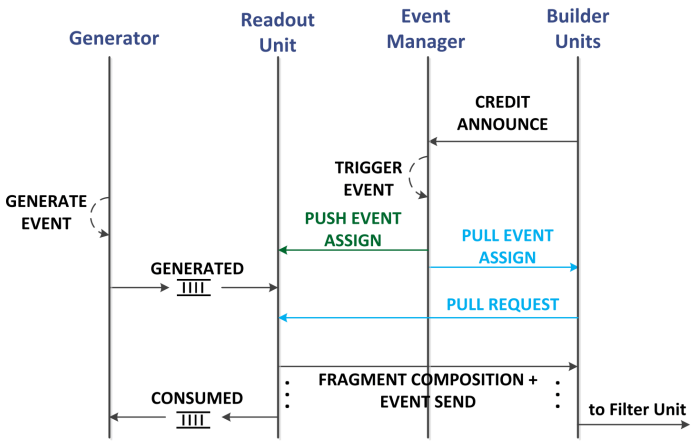



- The InfiniBand standard is widely used in data centers for SAN and HPC clusters
- InfiniBand provide high performance and low latency
- and a constant improvement

# EB prototype performance evaluator

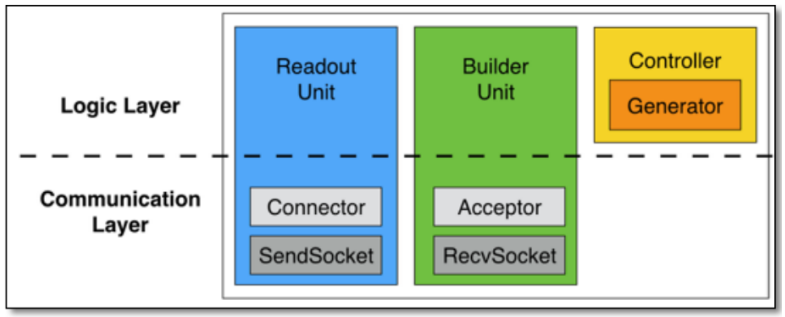
- We developed prototype evaluators to evaluate the performances of the EB in a collaboration between INFN and Cern
- We choose to evaluate both **performances** and **design choices**
- The Cern prototype is called LHCb-DAQPIPE
- It is logically divided into a Readout Unit (RU), Builder Unit (BU), Generator. These components run for every node of the EB
- The building processes is orchestrated by an Event Manager (EM)




- The generator emulates the PCIe40 output generating the subdetector fragments
- When a builder unit is idle it announce itself to the EM

# EB prototype performance evaluator

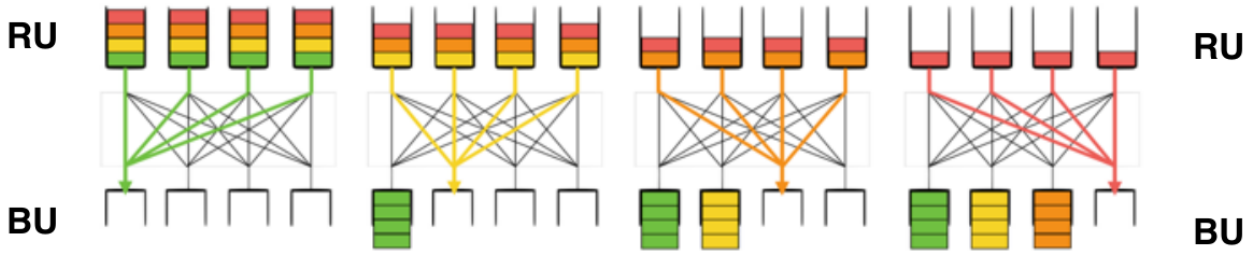
- The INFN prototype is called LSEB (*Large Scale Event Builder*)
- Simpler design: no EM is implemented



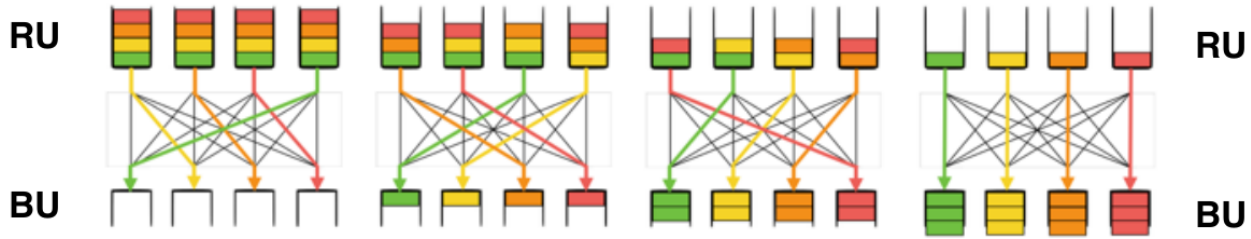

- The node performing the EB is selected in a round-robin fashion

- Both LHCb-DAQPIPE and LSEB are designed with a clean separation of the transport layer to allow for example to test different technologies

# EB prototype performance evaluator

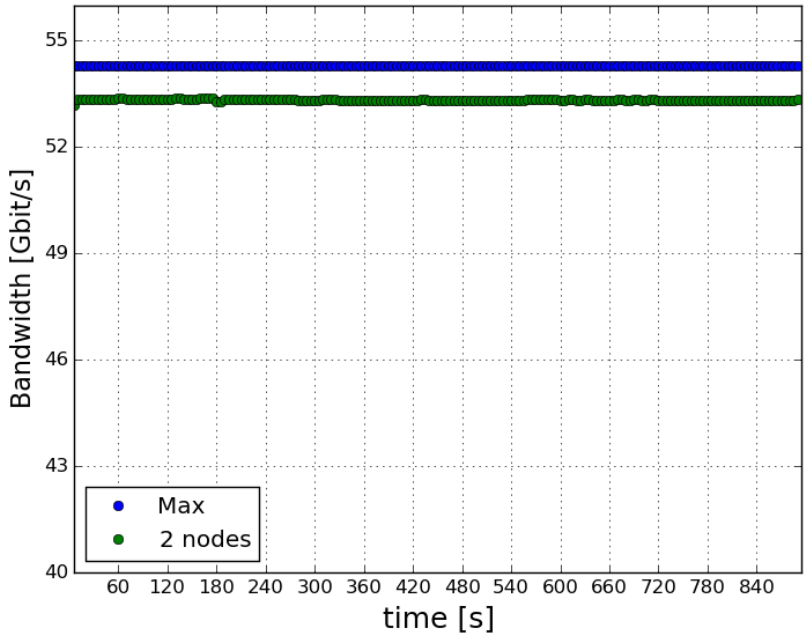



- to avoid possible network congestions the fragments are not send at the same moment, but the deliver is distributed over time



# EB performance on FDR

- We made preliminary tests on a simple two nodes test bed at INFN CNAF
- Two Xeon server equipped with Mellanox FDR (banda massima $54.3$ Gbit/s)




- Average value BU bandwidth over a 15 minutes test is $53.3$ Gbit/s: **98%** of the maximum allowed
- Similar results both for LHCb-DAQPIPE and LSEB

# EB software scalability

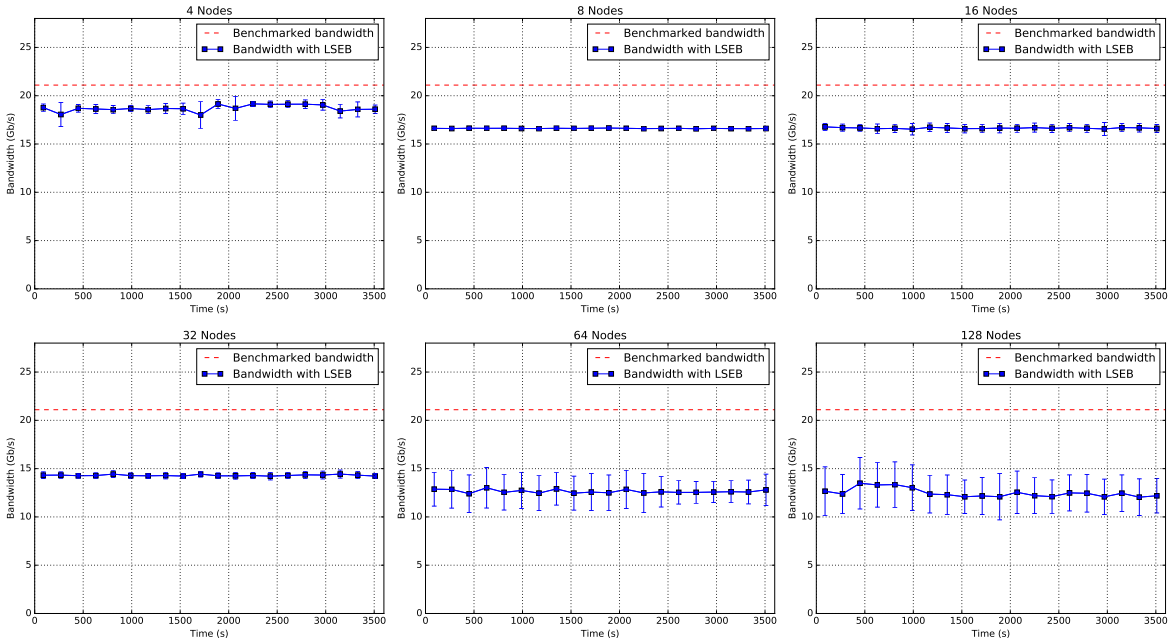
- We performed scalability tests using the Galileo cluster of the CINECA consortium ▸ Link

| | |
|---|---|
| Nodes | 516 |
| Processors | 2 8-core Intel Haswell $2.40\ \mathrm{GHz}$ per node |
| RAM | 128 GB/node, 8 GB/core |
| Network | InfiniBand QDR switch |

- The size of the cluster is similar to the LHCb EB LAN

# LSEB performance

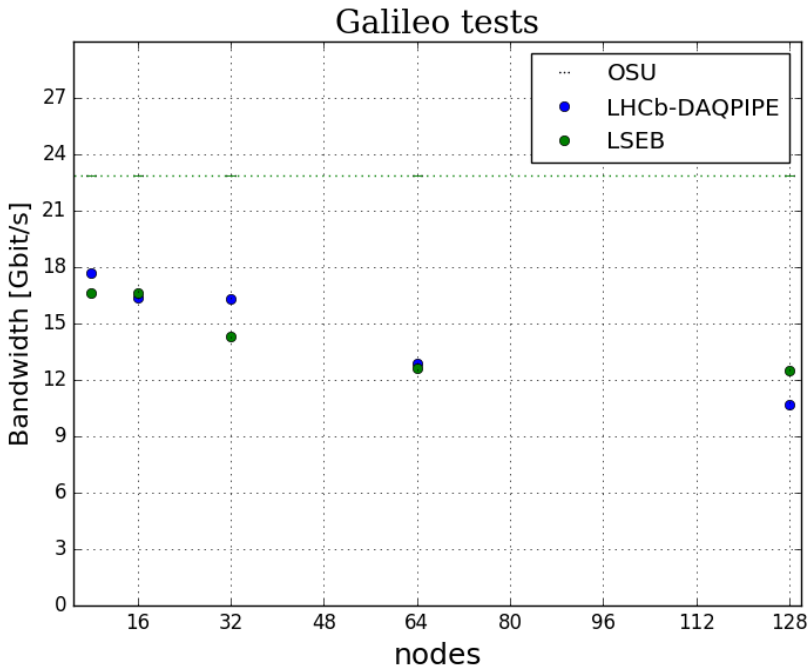
- Blue: Bandwidth of EB node (maximum $21.1$ Gbit/s)




- The EB work fine up to **128** nodi: $60\%$ of the maximum
- Some relevant limitation :
  - Not exclusive use of the cluster. Other processes polluting network usage
  - **No control on CPU power saving setting**

# Comparison between different implementations

- Comparison between LHCb-DAQPIPE and LSEB




- Performance very similar at the different sizes tested
- Both implementations are viable

# Towards higher size scalability tests

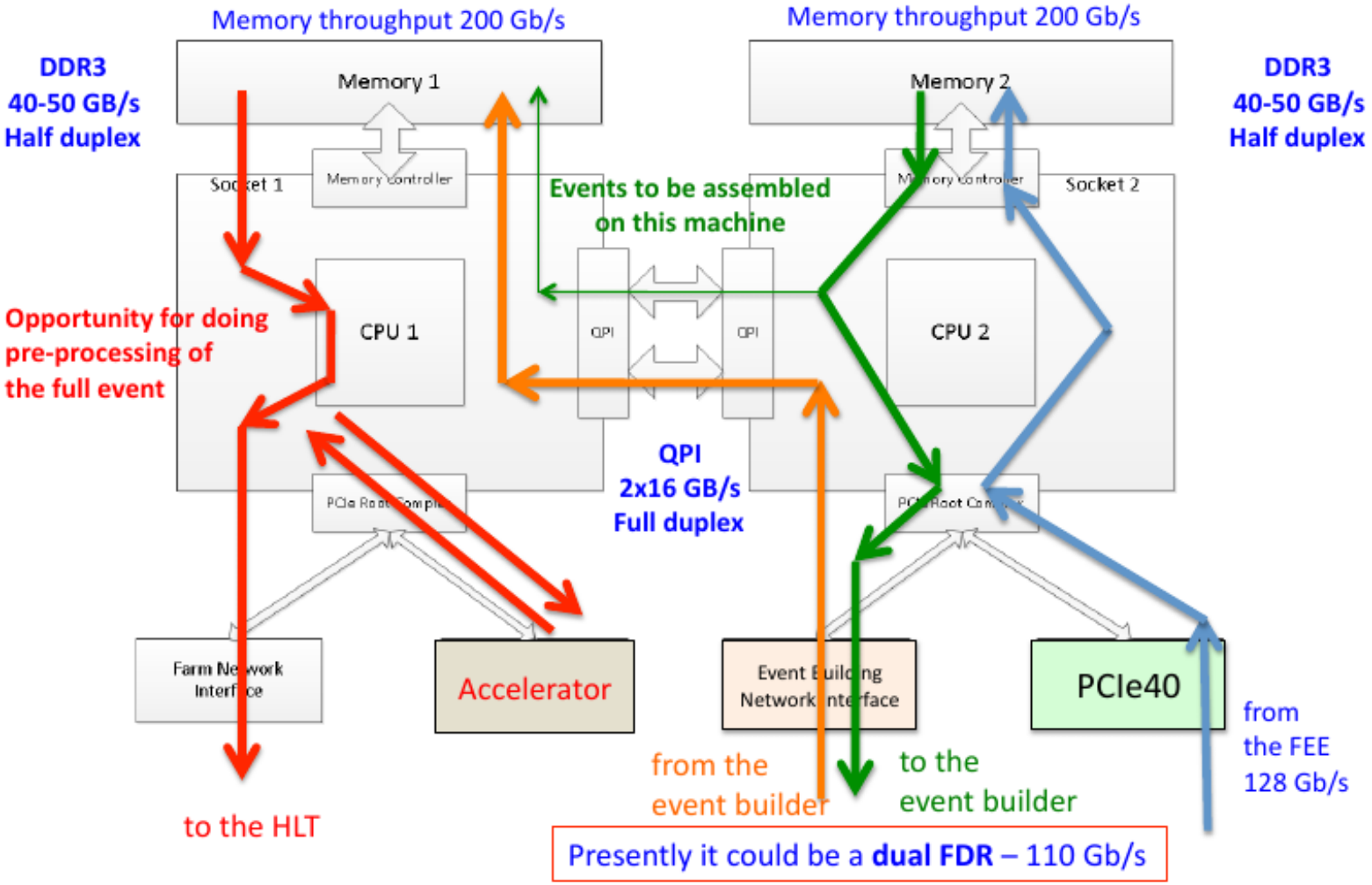
- We are in contact with CINECA to the test the EB software at their new Tier0 cluster (*Marconi*) Link

| Nodes | >1500 |
|---|---|
| Processors | Xeon Phi and Xeon E5-2600 v4 |
| Network | Intel OmniPath |

# Conclusions

- The LHCb experiment will be upgraded for the Run III data taking to face the challenge of read-out at $40\mathrm{MHz}$
- A software trigger solution has been designed requiring an high throughput EB LAN network
- We performed scalability tests of the two different software implementation of the EB software developed in collaboration between Cern and INFN
- We showed that both solutions work fine up 128 nodes on the Galileo cluster
- Further developments:
  - Test at bigger sizes at the CINECA Marconi cluster
  - Implement fault tolerance mechanisms

# Backup


Memory throughput 200 Gb/s
Memory throughput 200 Gb/s
DDR3
40-50 GB/s
Half duplex
DDR3
40-50 GB/s
Half duplex
Memory 1
Memory 2
Socket 1
Socket 2
Memory controller
Events to be assembled
on this machine
Opportunity for doing
pre-processing of
the full event
CPU 1
CPU 2
QPI
QPI
QPI
2x16 GB/s
Full duplex
Farm Network
Interface
Accelerator
Event Building
Network Interface
PCIe40
from
the FEE
128 Gb/s
to the HLT
from the
event builder
to the
event builder
Presently it could be a dual FDR – 110 Gb/s