The 40 MHz trigger-less DAQ system for the LHCb Upgrade

Antonio Falabella

INFN - CNAF (Bologna)

13^{th} Pisa Meeting on Advanced Detectors - 28th May, 2015
La Biodola, Isola d'Elba (Italy) 24 - 30 May, 2015

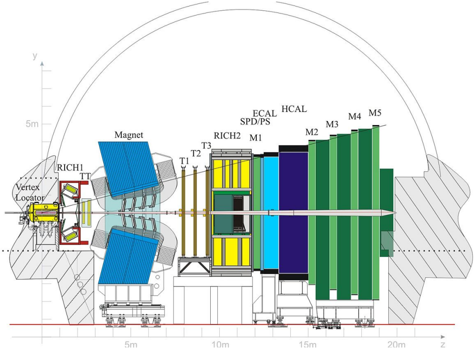
On behalf of the LHCb collaboration

Overview

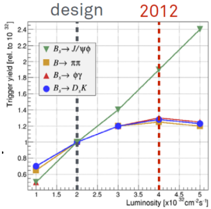
1. LHCb Experiment Upgrade
2. The DAQ for the Upgrade
3. Event Building performance evaluator
4. EB performance evaluator at the large scale
5. Summary

The LHCb Experiment

- LHCb is a heavy flavour physics experiment to study CP violation and rare decays of b and c hadrons with a high precision
- Run II instantaneous $L = 4 \times 10^{32}\mathrm{cm}^{-2}\mathrm{s}^{-1}$
- Integrated luminosity by the end of Run II $8\ \mathrm{fb}^{-1}$



- Run III scheduled 2020/2022
- LHC will run at the design $14\ \mathrm{TeV}$ energy
- Instanteneous luminosity will increase by a factor 5
- Upgrade DAQ to increase trigger efficiency



Trigger evolution for Run III

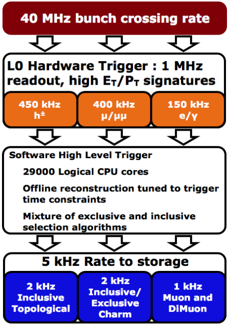
Trigger configuration for Run I and II:

- $L0$ hardware trigger reduces the rate from $40\ \mathrm{MHz}$ to $1.1\ \mathrm{MHz}$
- The software trigger further reduces the rate of data sent to storage

Trigger configuration for Run III:

- Software-only trigger with full event reconstruction
- Enhanced trigger efficiency

Run I



40 MHz bunch crossing rate

L0 Hardware Trigger : 1 MHz readout, high E_T/P_T signatures

450 kHz $h^{\pm}$ | 400 kHz $\mu/\mu\mu$ | 150 kHz e/γ

Software High Level Trigger

29000 Logical CPU cores

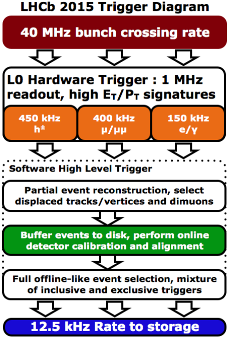
Offline reconstruction tuned to trigger time constraints

Mixture of exclusive and inclusive selection algorithms

5 kHz Rate to storage

2 kHz Inclusive Topological | 2 kHz Inclusive/ Exclusive Charm | 1 kHz Muon and DiMuon

Run II



LHCb 2015 Trigger Diagram

40 MHz bunch crossing rate

L0 Hardware Trigger : 1 MHz readout, high E_T/P_T signatures

450 kHz $h^{\pm}$ | 400 kHz $\mu/\mu\mu$ | 150 kHz e/γ

Software High Level Trigger

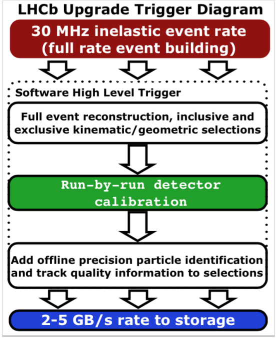
Partial event reconstruction, select displaced tracks/vertices and dimuons

Buffer events to disk, perform online detector calibration and alignment

Full offline-like event selection, mixture of inclusive and exclusive triggers

12.5 kHz Rate to storage

Run III



LHCb Upgrade Trigger Diagram

30 MHz inelastic event rate (full rate event building)

Software High Level Trigger

Full event reconstruction, inclusive and exclusive kinematic/geometric selections

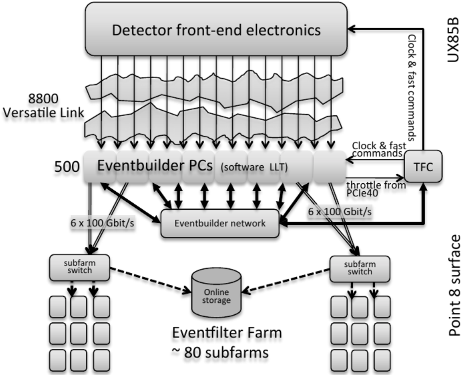
Run-by-run detector calibration

Add offline precision particle identification and track quality information to selections

2-5 GB/s rate to storage

DAQ implementation

- The idea for the DAQ Upgrade is to use a high-throughput network for the readout and the event building (EB)



- PCIe40 readout units push event fragments into builder units (~ 100 Gbit/s)
- ~ 500 EB nodes communicate at ~ 100 Gbit/s full-duplex (DAQ network)
- Output of EB to event filter farm for further processing

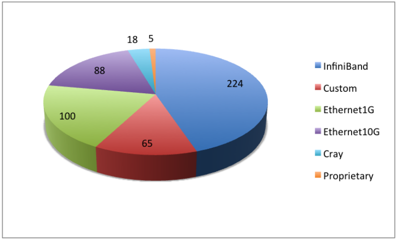
DAQ implementation

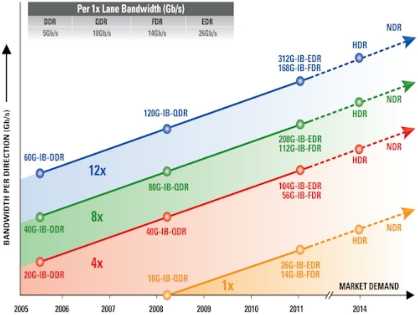
- The challenge is to handle $30\ \mathrm{Tbit/s}$ of aggregated traffic $(30\ \mathrm{MHz} \cdot 100\ \mathrm{KByte})$

Event rate	30 MHz
Mean nominal event size	100 KBytes
Readout board bandwidth	100 Gbit/s (16 lanes PCIe 3)
CPU cores	Up to 4000

- It can be done with commercial fabric technologies: InfiniBand, Ethernet
- Focus on InfiniBand in this talk

InfiniBand standard

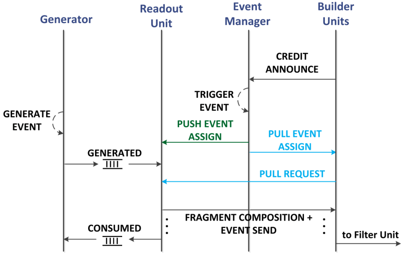




- InfiniBand standard is widely used in HPC computing
- High speed and cost effective
- Constant speed evolution
- Thorougly tested on different testbeds with an EB performance evaluator developed on purpose → *lhcb-daqpipe*

Event Building performance evaluator - *lhcb-daqpipe*

- Performance evaluator for the EB software: *lhcb-daqpipe* developed and tested with the collaboration of the lhcb-online group
- EB building blocks: Generator, Readout Unit (RU), Builder Unit (BU), Event Manager (Listener and Consumer)



- The generator emulates the PCIe40 output
- It writes metadata and data directly into RU memory
- The EM elects one node as the BU
- Each RU sends its fragment to the elected BU

- Performance measured on different test beds and with different InfiniBand cards

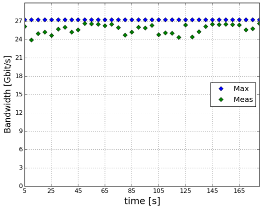
Event Building performance evaluator - *lhcb-daqpipe*

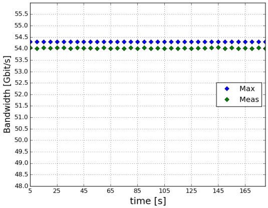
- *lhcb-daqpipe* allows to test both PULL and PUSH protocols
- It provides several transport layer implementations: IB verbs, TCP, UDP
- The processes on the nodes are spawned using MPI or by a synchronization mechanisms based on the ZeroMQ library

- We tested the EB software on test beds of increasing size:
  - At CNAF with 2 Intel Xeon server connected back-to-back
  - At Cern with 8 Intel Xeon cluster connected through an IB-switch
  - At the 512-node Galileo cluster at Cineca → next slides

QDR and FDR characterization

- We measured the point-to-point bandwidth for different InfiniBand HCAs with RDMA *write* semantics (similar results for *send* semantics)



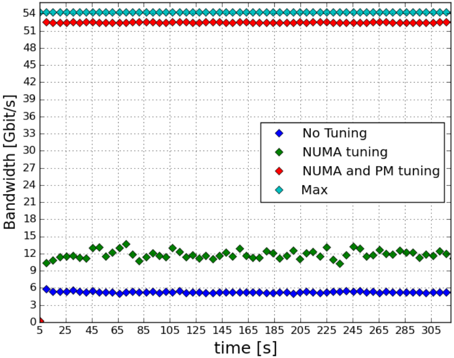


- QLogic: QLE7340, Single port $32\ \mathrm{Gbit/s}$ (QDR)
- Unidirectional throughput $27.2\ \mathrm{Gbit/s}$
- Galileo and Cern clusters

- Mellanox: MCB194A-FCAT, Dual port $56\mathrm{Gbit/s}$ (FDR)
- Unidirectional throughput $54.3\mathrm{bit/s}$ (per port)
- CNAF testbed

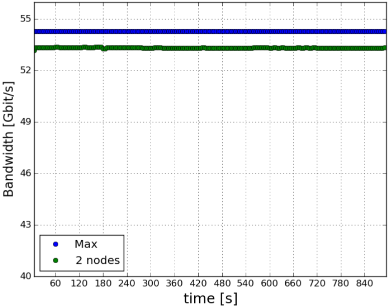
Tuning issues

- IO performance can be severely degraded without a proper tuning of the nodes
  - PCIe Gen3 x16 Lanes: any previous version of the PCI bus represents a bottleneck for the network traffic
  - Disable node interleaving in NUMA architectures
  - Disable Power Management and CPU frequency selection (PM and frequency switching are latency sources)



Event Builder performance

- Measured bandwidth @CNAF as seen by the builder units on two nodes equipped with Mellanox FDR (max bandwidth 54.3 Gbit/s considering the encoding)
- Duration of the tests: 15 minutes



- PM and node inteleaving disabled
- Bandwidth measured is on average 53.3 Gbit/s: **98%** of maximum allowed

EB performance evaluator on the large scale

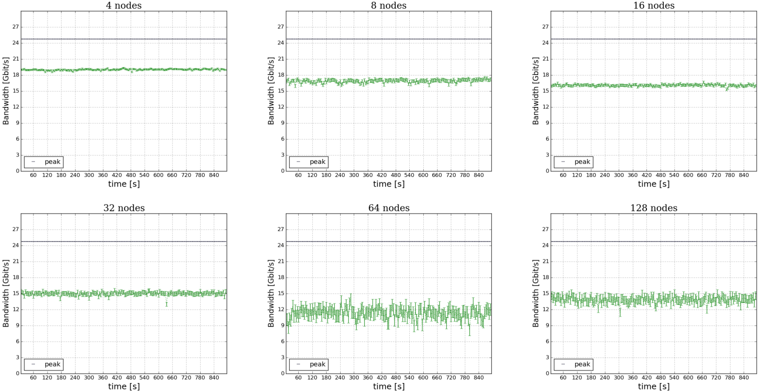
- Extensive test have been made on the CINECA Galileo TIER-1 cluster ▸ Link

Nodes	516
Processors	2 8-core Intel Haswell 2.40 GHz per node
RAM	128 GB/node, 8 GB/core
Network	InfiniBand with 4x QDR switches
MPI	OpenMPI v1.8.4

- The cluster size is similar to the LHCb Upgraded DAQ network

Event Builder performance

- Measured bandwidth as seen by the BU on an increasing number of nodes
- Blue: Average bidirectional bandwidth achievable $(24.7\ \mathrm{Gbit/s})$

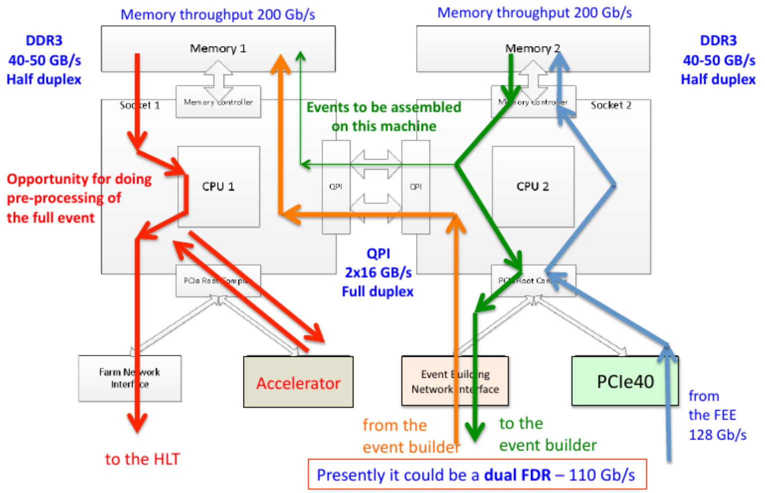


- The EB works properly up to a scale of **128 nodes**
- Few limitations to reach the maximum bandwidth:
    - cluster is in production so other processes are polluting the network traffic
    - **no control on power management and frequency switching**

Summary

- Software level trigger for the LHCb Upgrade is a challenging task → DAQ can be implemented with a InfiniBand-based network
- A performance evaluator has been developed in order to test the possible implementation choices
- I tested several InfiniBand HCAs on different test beds
- A control on the node interlieving and PM is needed to get the best performance
- Large scale tests have been performed showing that the EB prototype behave properly as the number of nodes increases
- Next developments:
  - Testing the EB performance evaluator with a higher number of nodes
  - Make it less sensible to PM issues

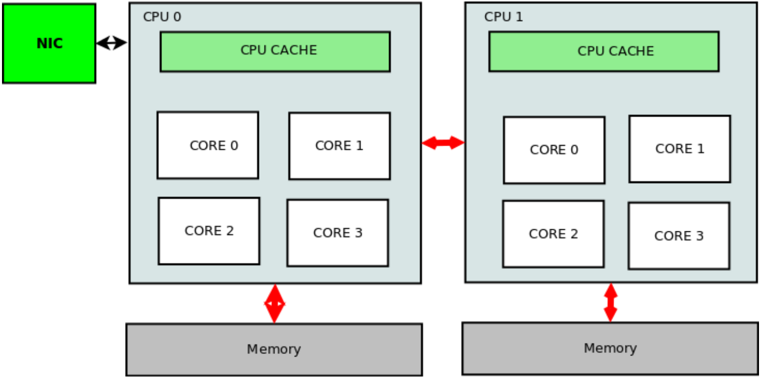
Backup



Presently it could be a **dual FDR** – 110 Gb/s

Numa - Non Uniform Memory Access

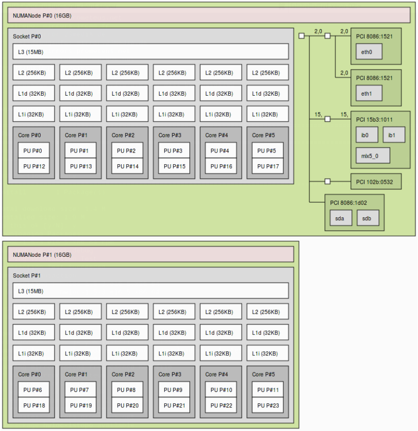
- Memory architecture for multiprocessors
- In the schema above each CPU (NUMA node) has its own local bank of memory
- A CPU has faster access to its local memory
- Access to non-local memory is a potential bottleneck for IO



- The NIC is connected to one of the CPU
- If CPU1 tries to access the NIC it will experience higher latency w.r.t. CPU0

Numa topology

lstopo, part of the *hwloc* package, produces CPU/Cache/Memory topology schemas



Topology of our machines

- Topology of the test bed machines consists of 2 NUMA nodes
- The FDR InfiniBand network interfaces ib0, ib1 and the ethernet interfaces are connected to the first NUMA node
- High network latency is experienced if the EB data fragments are sent by a process running on cores 6 to 11

Power management issues

- Power saving states (C-states) of CPUs reduce the power consumption but can be critical to performance
- C-0 corresponds to every CPU component turned on. C-states with higher values correspond to lower power consumption
- Switching back and forth among the various states will result in performance degradation
- Switching between CPU frequencies gives similar effects