Event Building for the LHCb Upgrade

30 MHz Rate (32 Tb/s Aggregate Throughput)

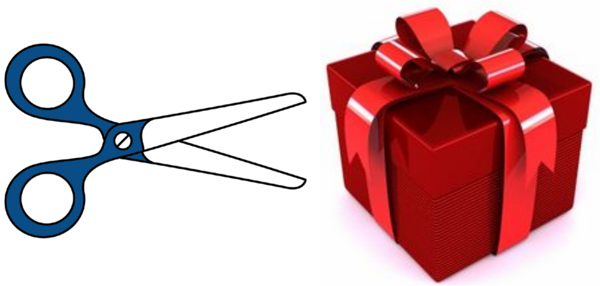
Domenico Galli

INFN Bologna and Università di Bologna

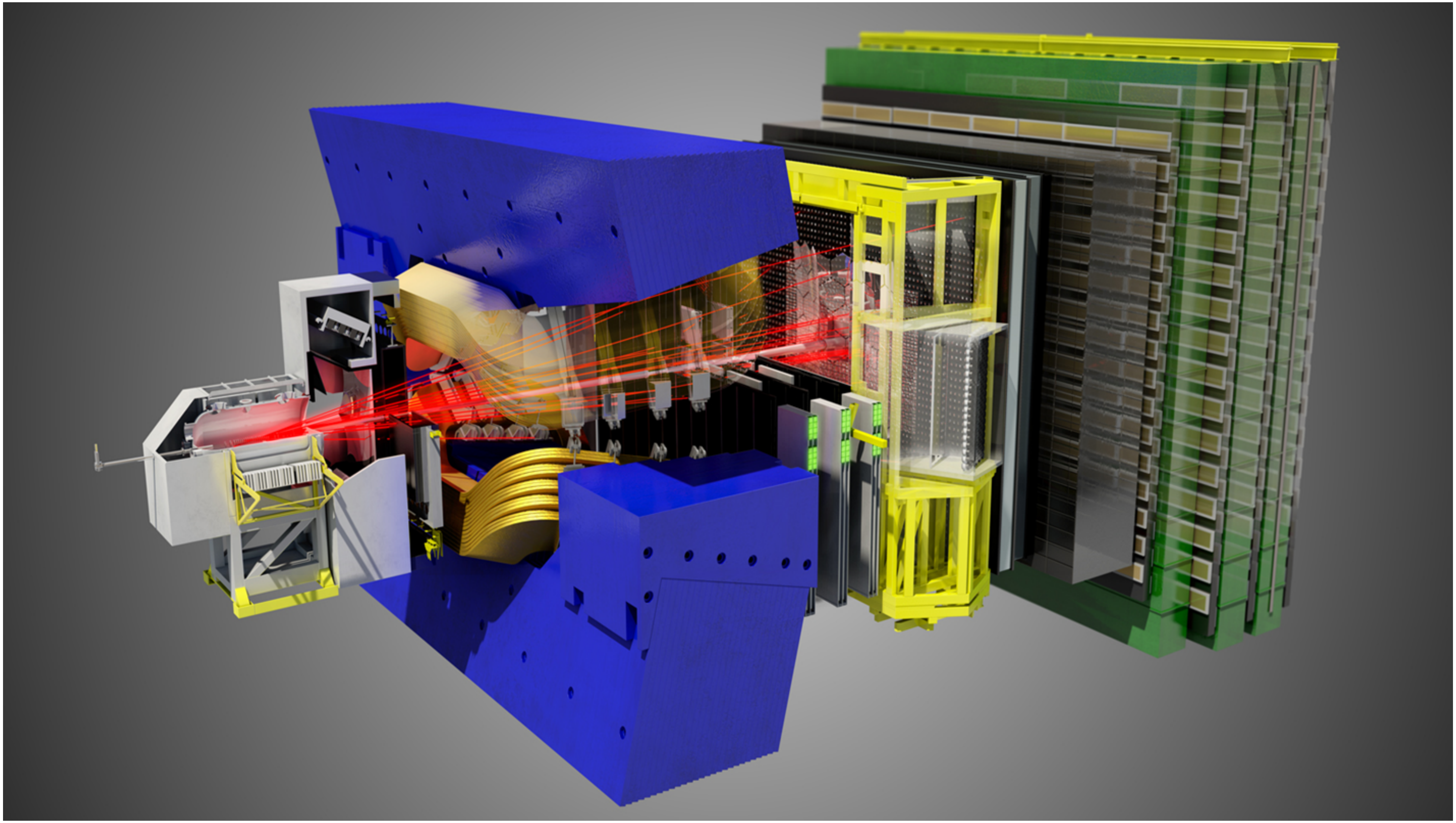
Workshop INFN CCR -- Frascati, 25-29 May 2015

The LHCb Experiment

- LHCb is a **high-precision** experiment devoted to the search for New Physics beyond the Standard Model:
  - By studying **CP violation** and **rare decays** in the b and c quark sectors;
  - Searching for deviations from the SM (New Physics) due to **virtual contributions** of new heavy particles in loop diagrams.
- Need for **upgrade**:
  - Deviations from SM are expected to be **small**;
  - Need to **increase** significantly the **precision** of the measurements and therefore the **statistics**.

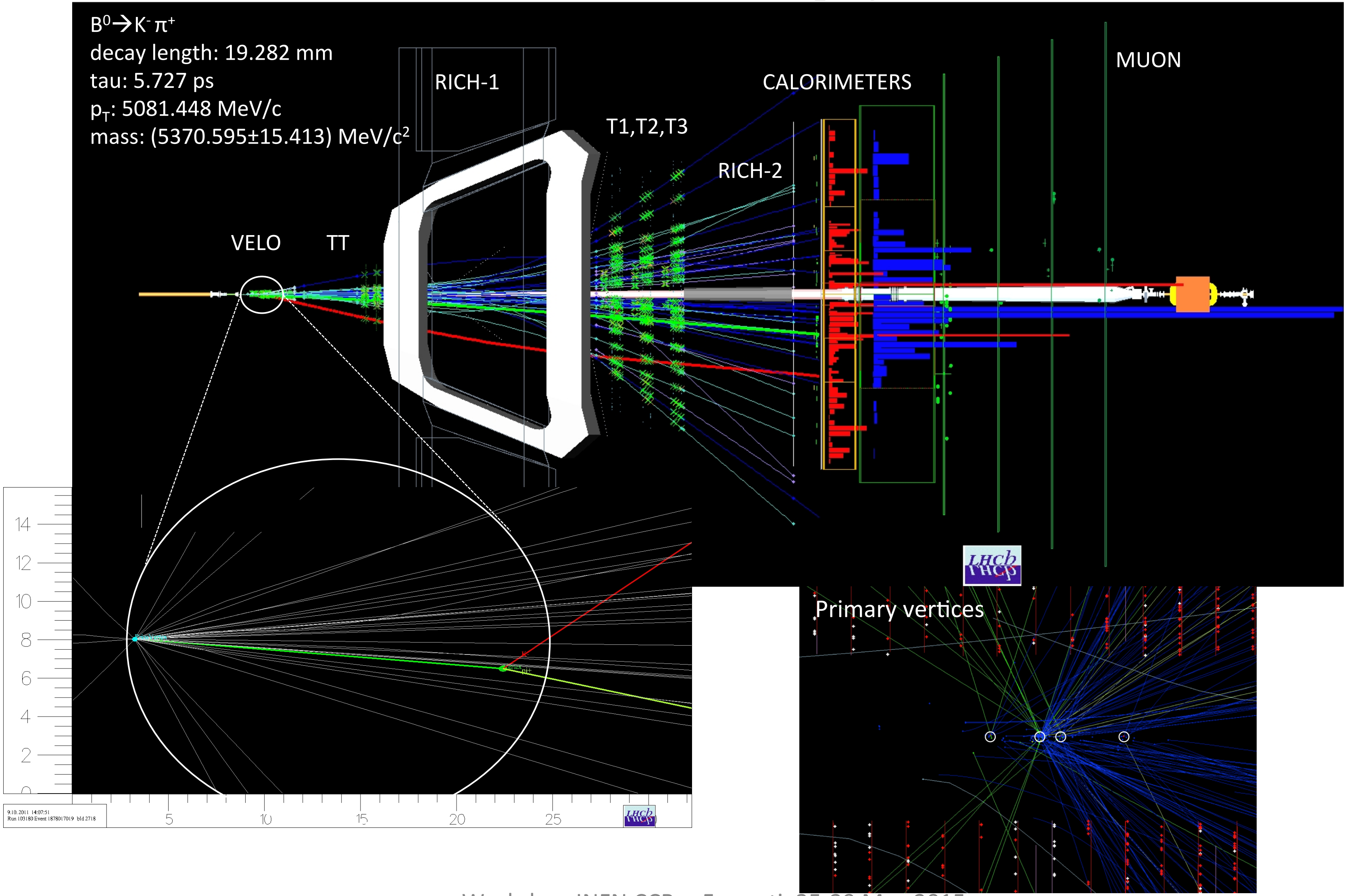


The LHCb Detector



LHCb Events

LHCb Event Display



$B^0 \rightarrow K^- \pi^+$

decay length: 19.282 mm

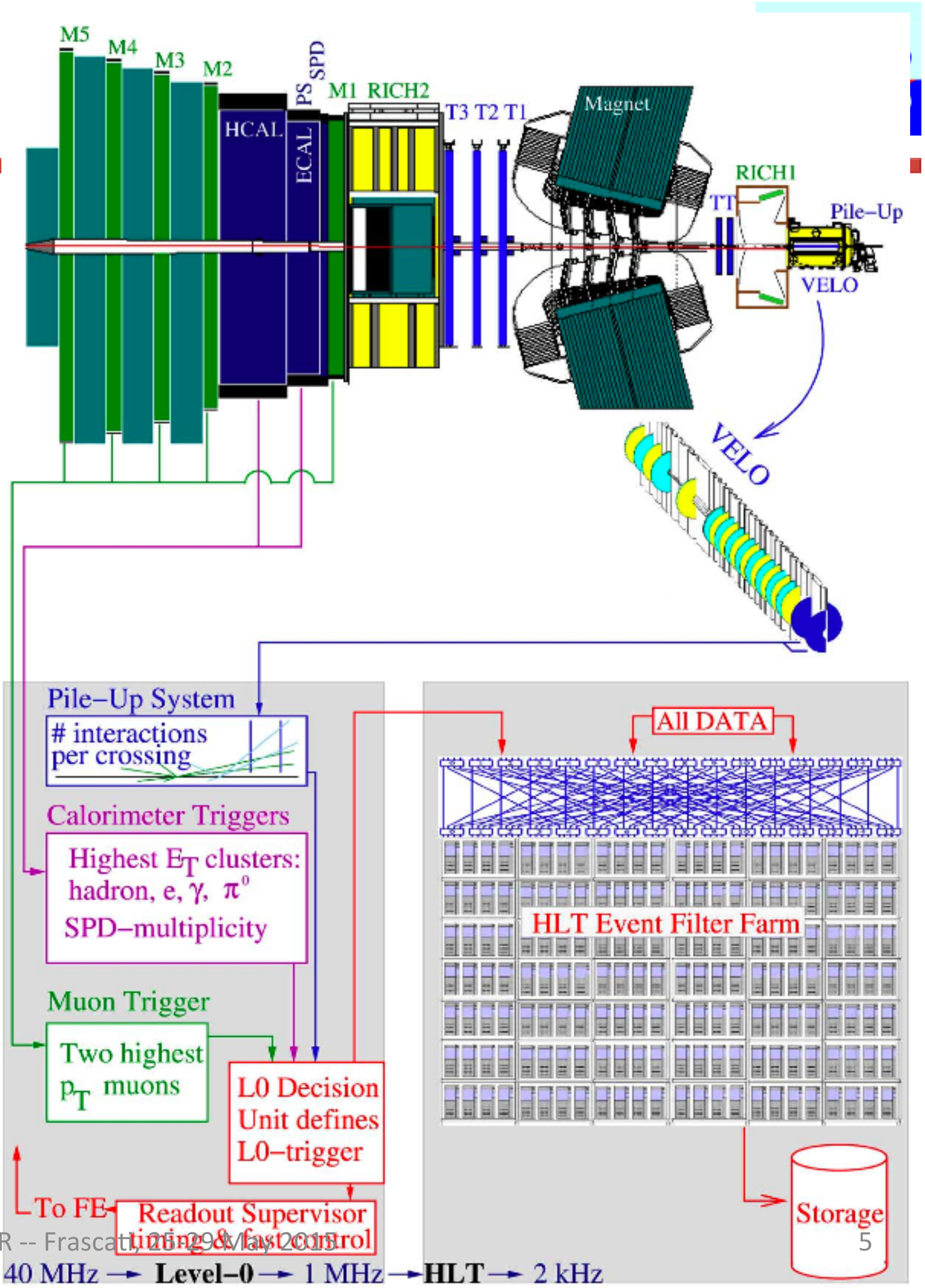
tau: 5.727 ps

p_T: 5081.448 MeV/c

mass: (5370.595±15.413) MeV/c^2

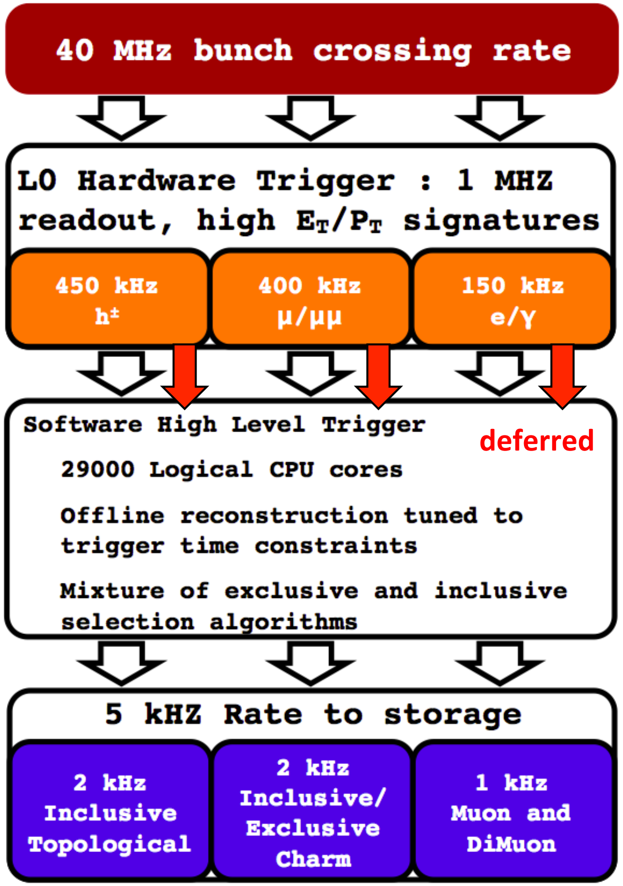
The Present LHCb Trigger

- 2 stages:
  - Level-0: synvhronous, hardware + FPGA; 40 MHz → 1 MHz.
  - HLT: software, PC farm: 1 MHz → 2 kHz.
- Front-End Electronics:
  - Interfaced to Read-out Network.
- Read-Out Network:
  - Gigabit Ethernet LAN.
  - Read-out @ 1.1 MHz.
  - Aggregate thoughput: 60 GiB/s.


M5
M4
M3
M2
SPD
PS
M1
RICH2
HCAL
ECAL
T3 T2 T1
Magnet
RICH1
TT
Pile–Up
VELO
VELO
Pile–Up System
# interactions per crossing
Calorimeter Triggers
Highest E_T clusters: hadron, e, γ, π^0
SPD–multiplicity
Muon Trigger
Two highest p_T muons
L0 Decision Unit defines L0–trigger
To FE
Readout Supervisor timing & fast control
All DATA
HLT Event Filter Farm
Storage
40 MHz → Level–0 → 1 MHz → HLT → 2 kHz

The Present LHCb Trigger (II)

- The **Level-0 trigger** based on the signals from ECAL, HCAL and MUON detectors read at 40 MHz, operates on custom electronics, with a **maximum output rate limited to 1.1 MHz**.
  - Fully pipelined, constant latency of about 4 μs.
  - Bandwidth to the HLT ~4 Tb/s.
  - High p_T muon (1.4 GeV) or di-muon.
  - High pT local cluster in HCAL (3.5 GeV) or ECAL (2.5 GeV).
  - 25% of the events are **deferred**: temporarily stored on disk and processed with the HLT farm during the inter-fills.
- **HLT** is a software trigger.
  - Reconstruct VELO tracks and primary vertices
  - Select events with at least one track matching p, p_T , impact parameter and track quality cuts.
  - At around **50 kHz** performs inclusive or exclusive selections of the events.
  - Full track reconstruction, without particle-identification.
  - Total accept rate to disk for offline analysis is **5 kHz**.



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

deferred

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

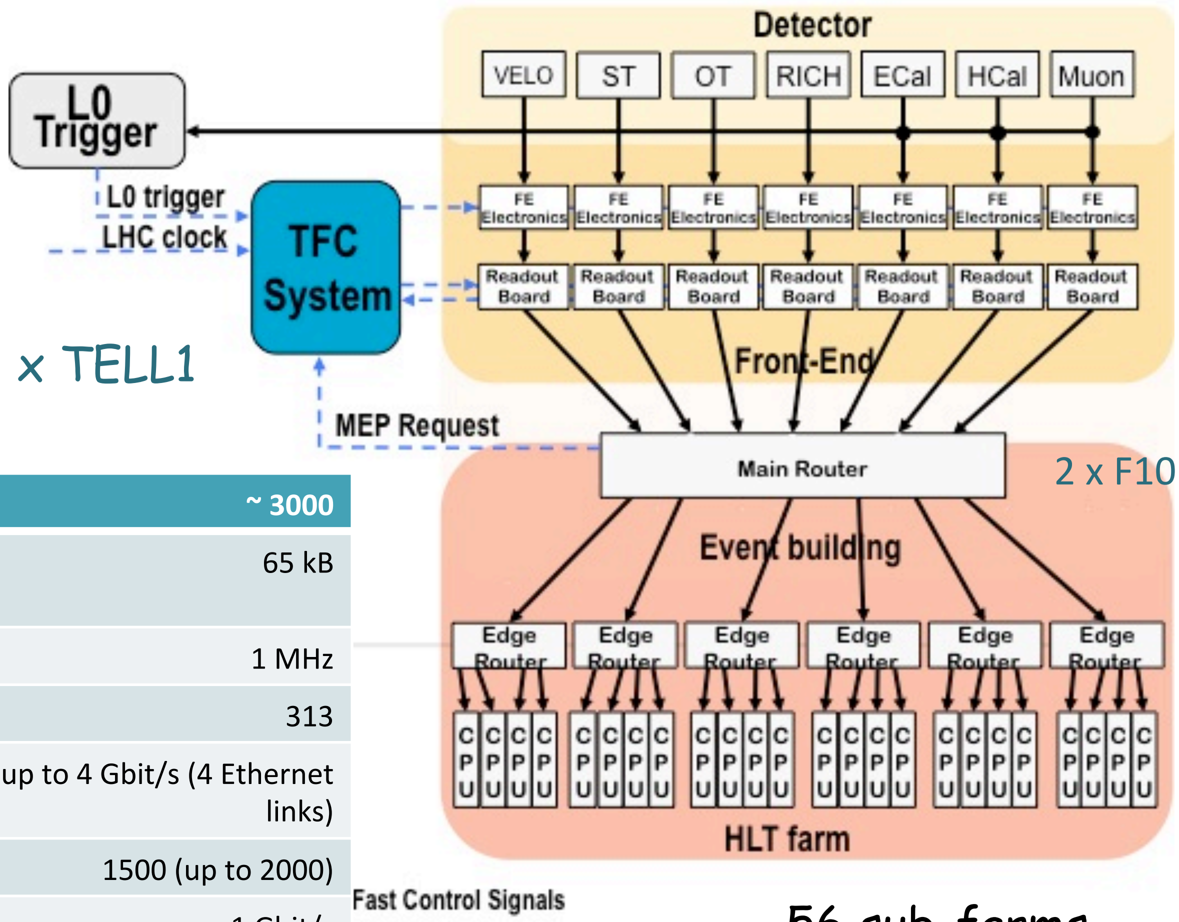
2 kHz Inclusive/ Exclusive Charm

1 kHz Muon and DiMuon

LHCb DAQ Today

Push-protocol with centralized flow-control

Readout boards: 313 x TELL1

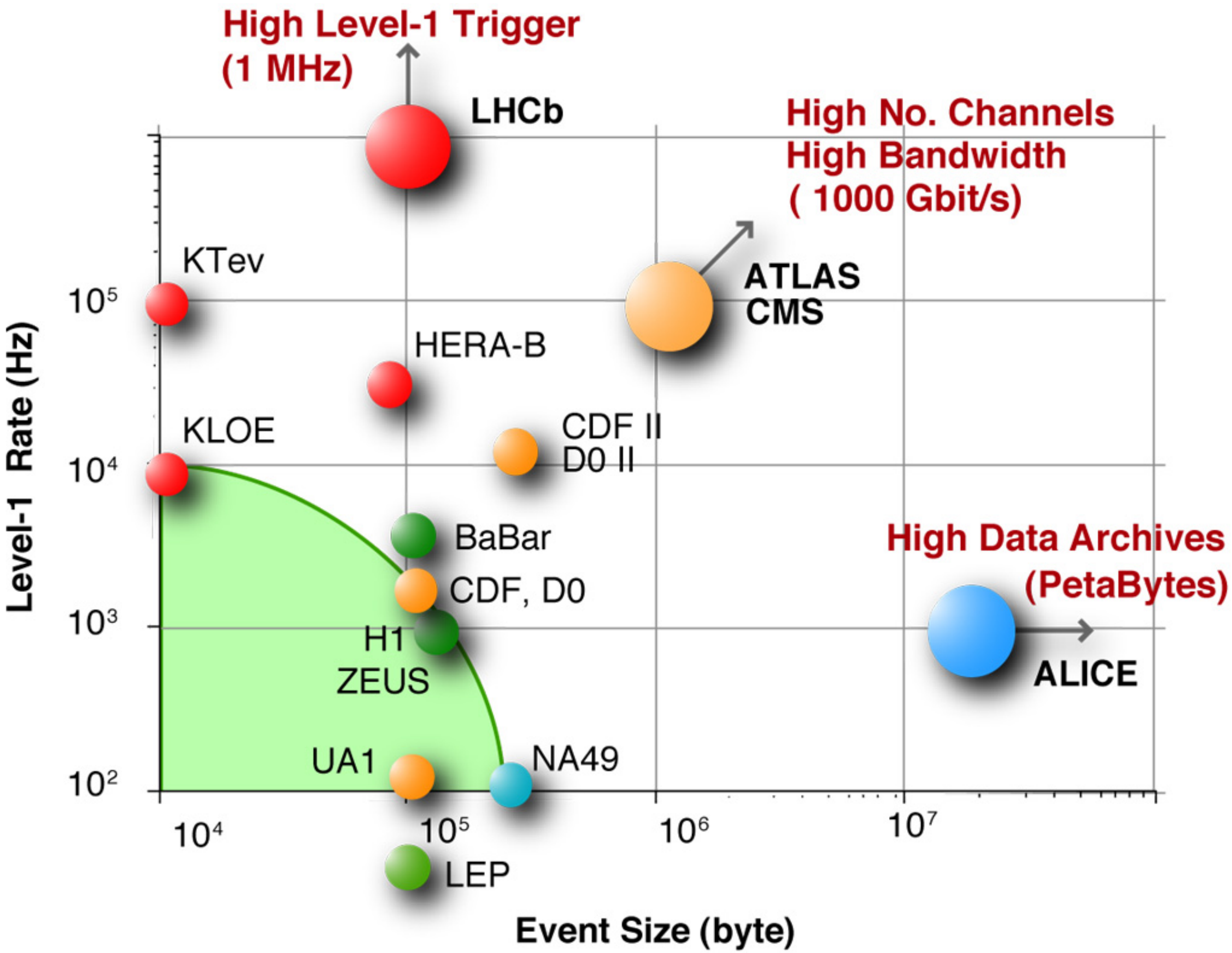


2 x F10 E1200i

56 sub-farms

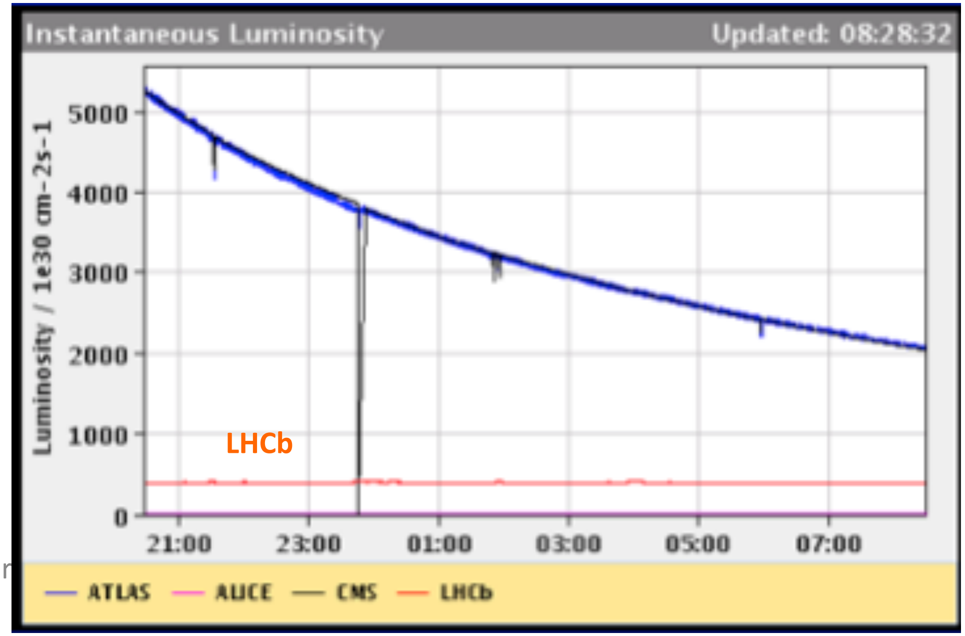
# links (UTP Cat 6)	~ 3000
Event-size (total – zero-suppressed)	65 kB
Read-out rate	1 MHz
# read-out boards	313
output bw / read-out board	up to 4 Gbit/s (4 Ethernet links)
# farm-nodes	1500 (up to 2000)
max. input bw / farm-node	1 Gbit/s
# core-routers	2
# edge routers	56

LHCb DAQ Today (II)



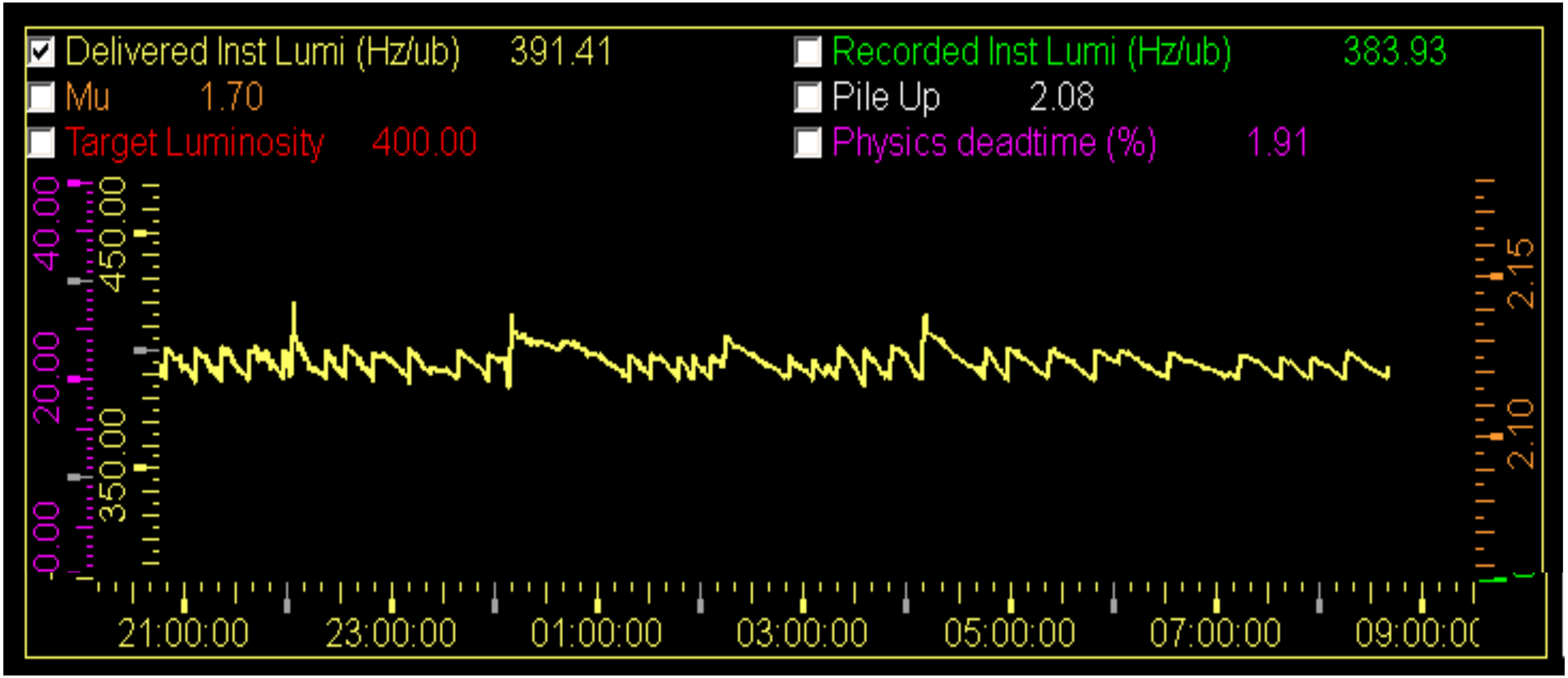
Luminosity and Event Multiplicity

- **Instantaneous luminosity** leveling at 4×10^{32} cm^{-2} s^{-1}, ±3% around the target value.
- LHCb was **designed** to operate with a **single collision per bunch crossing**, running at a instantaneous luminosity of 2×10^{32} cm^{-2} s^{-1} (assuming about **2700** circulating bunches):
  - At the time of design there were **worries** about possible **ambiguities** in assigning the B decay vertex to the proper primary vertex among many.
- Soon LHCb realized that running at **higher multiplicities** would have been **possible**. In 2012 we run at 4×10^{32} cm^{-2} s^{-1} with only **1262** colliding bunches:
  - 50 ns separation between bunches while the nominal 25 ns (will available by 2015).
  - **4 times more collisions** per crossing than planned in the design.
  - The average number of visible collisions per bunch crossing in 2012 raised up to $\mu > 2.5$.
  - $\mu \sim 5$ feasible but…



Luminosity and Event Multiplicity (II)

- At present conditions, if we **increase** the luminosity:
  - Trigger yield of hadronic events **saturates**;
  - The p_T **cut** should be **raised** to remain within the 1 MHz L0 output rate;
  - There would be **not** a real **gain**.



☑ Delivered Inst Lumi (Hz/ub) 391.41

□ Mu 1.70

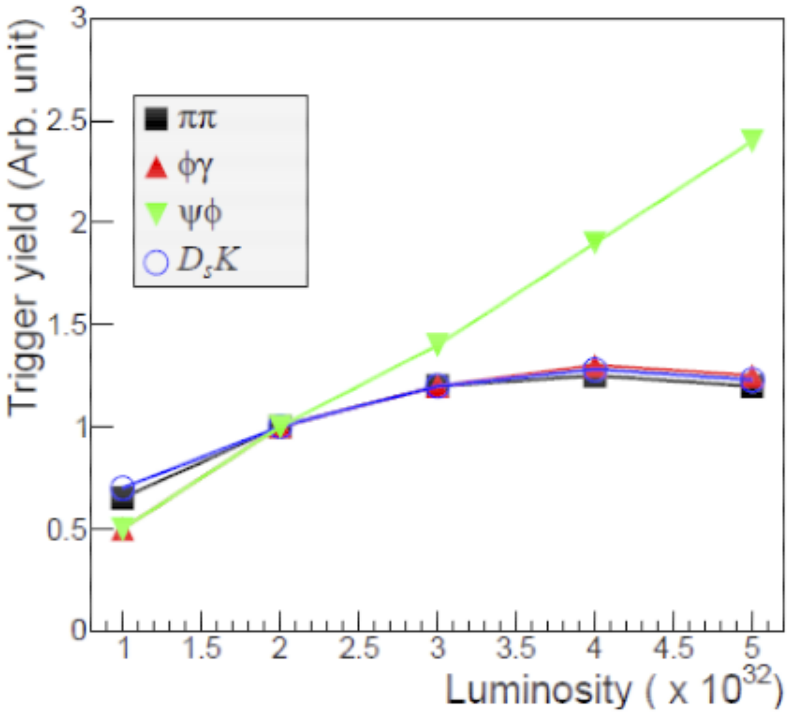
□ Target Luminosity 400.00

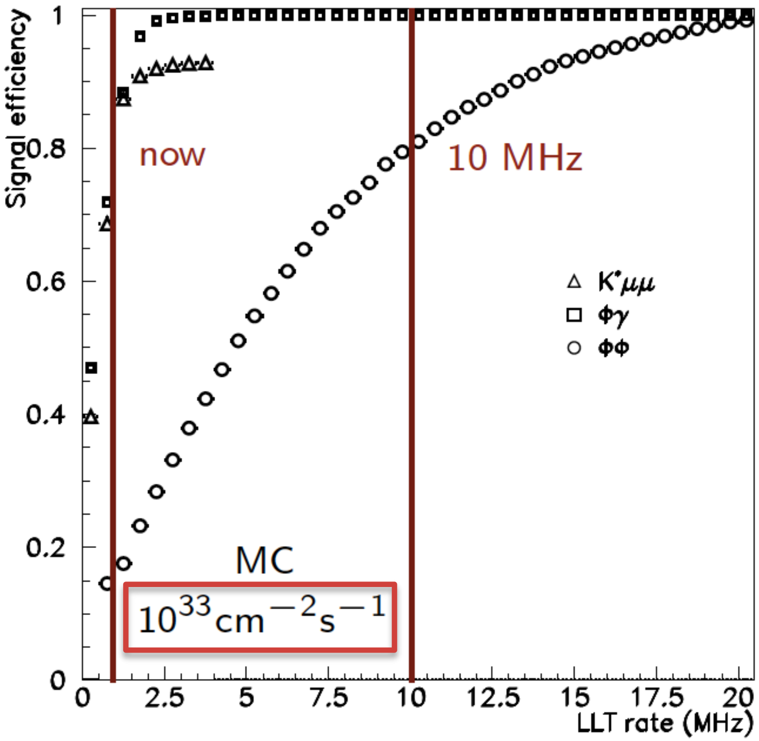
□ Recorded Inst Lumi (Hz/ub) 383.93

□ Pile Up 2.08

□ Physics deadtime (%) 1.91

The 1MHz L0 Rate Limitation

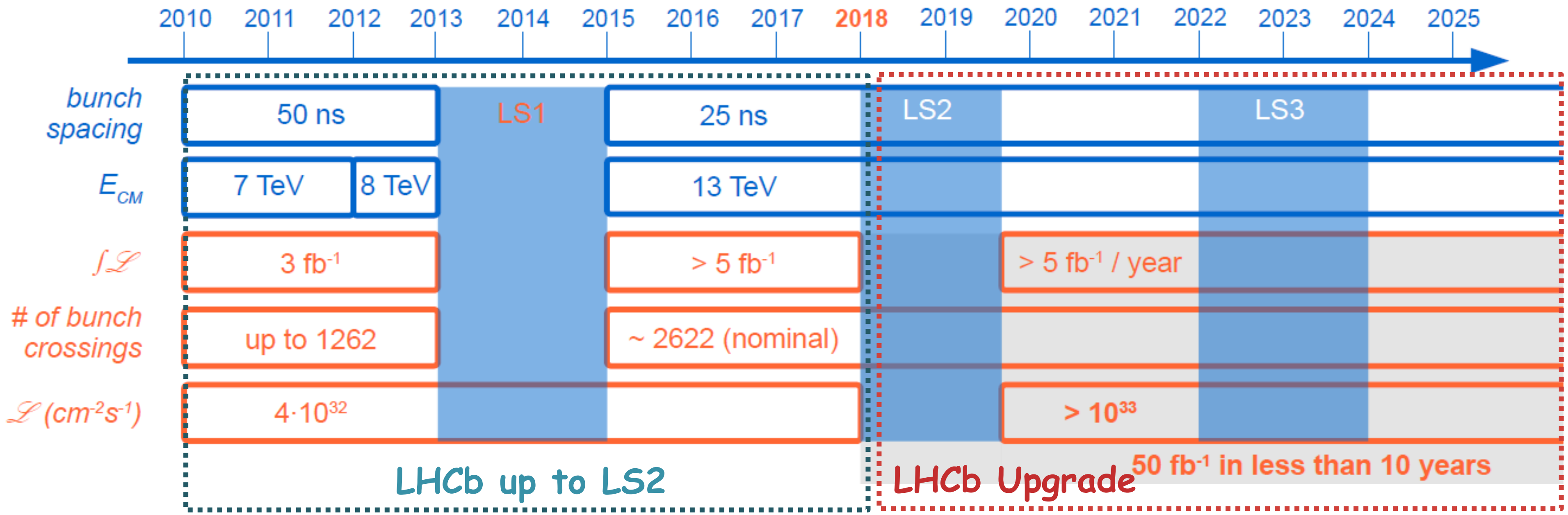




- Due to **the available bandwidth** and the **limited discrimination power** of the hadronic L0 trigger, LHCb experiences the **saturation of the trigger yield** on the **hadronic channels** around **4×10^{32} cm^{-2} s^{-1}**.
- **Increasing the first level trigger rate** considerably increases the efficiency on the hadronic channels.

The LHCb Upgrade - Timeline

- Shall take place during the Long Shutdown 2 (LS2)
  - In 2018.



2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025

bunch spacing: 50 ns | LS1 | 25 ns | LS2 | LS3

E_{CM}: 7 TeV | 8 TeV | 13 TeV

$\int\mathscr{L}$: 3 fb^{-1} | > 5 fb^{-1} | > 5 fb^{-1} / year

# of bunch crossings: up to 1262 | ~ 2622 (nominal)

$\mathscr{L}$ $(cm^{-2}s^{-1})$: $4\cdot10^{32}$ | $> 10^{33}$

LHCb up to LS2

LHCb Upgrade

50 fb^{-1} in less than 10 years

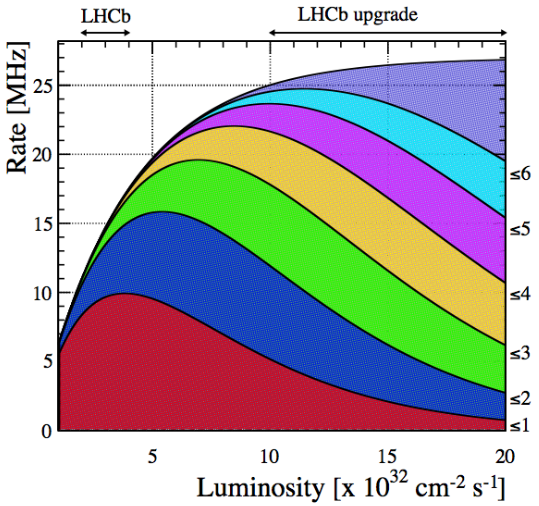
LHCb Upgrade: TDRs

- Letter of Intent for the LHCb Upgrade:
  - CERN-LHCC-2011-001 ; LHCC-I-018. - 2011.
- Framework TDR for the LHCb Upgrade: Technical Design Report:
  - CERN-LHCC-2012-007 ; LHCb-TDR-012. - 2012.
- LHCb VELO Upgrade Technical Design Report:
  - CERN-LHCC-2013-021 ; LHCB-TDR-013. - 2013.
- LHCb PID Upgrade Technical Design Report:
  - CERN-LHCC-2013-022 ; LHCB-TDR-014. – 2013.
- LHCb Tracker Upgrade Technical Design Report:
  - CERN-LHCC-2014-001; LHCB-TDR-015. – 2014
- LHCb Trigger and Online TDR:
  - CERN-LHCC-2014-016; LHCB-TDR-016. – 2014

The LHCb Upgrade

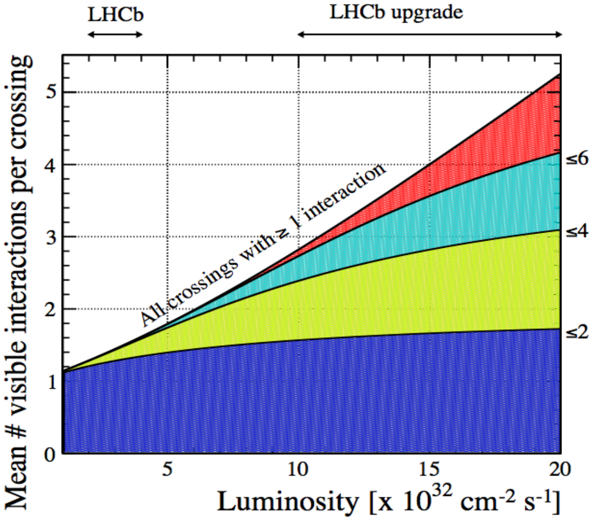
- **Readout** the whole detector at **40 MHz**.
- **Trigger-less data acquisition system**, running at 40 MHz (~30 MHz are non empty crossings):
  - Use **a (Software) Low Level Trigger** as a **throttle** mechanism, while progressively increasing the power of the event filter farm to run the HLT up to 40 MHz.
- We have foreseen to reach **20×10^{32} cm^{-2}s^{-1}** and therefore to prepare the sub-detectors on this purpose:
  - **pp interaction rate 27 MHz**.
  - At **20×10^{32} cm^{-2} s^{-1} pile up μ ≅ 5.2**
  - Increase the yield in the decays with muons by a **factor 5** and the yield of the hadronic channels by a **factor 10**.
- Collect **50 fb^{-1}** of data over ten years.
  - **8 fb^{-1}** is the integrated luminosity target, to reach by 2018 with the present detector;
  - **3.2 fb^{-1}** collected so far.

Running Conditions



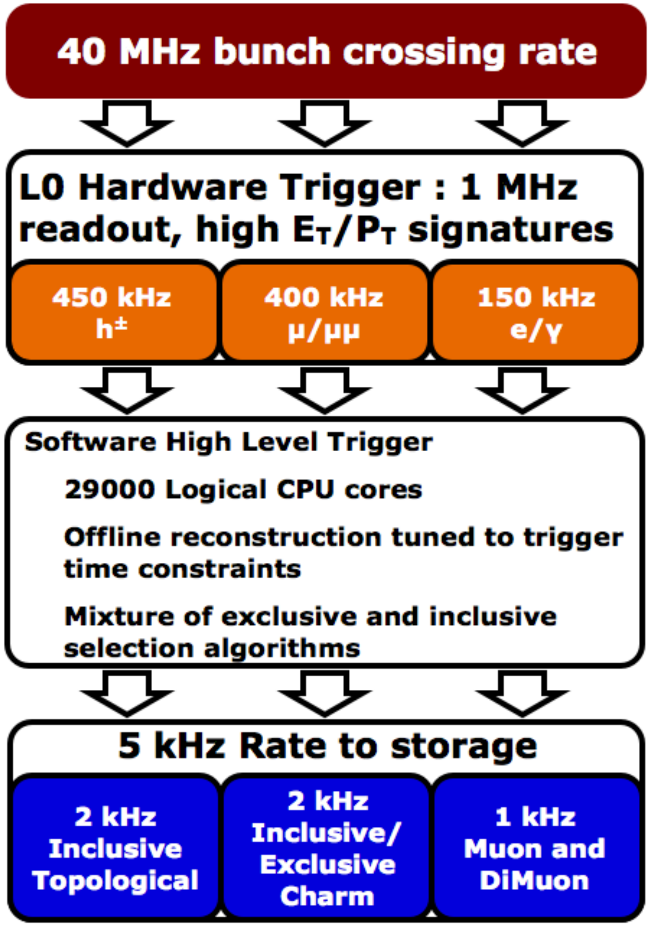
27 MHz

Mean visible interactions per crossing



μ =5.2

The LHCb Upgrade (II)



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

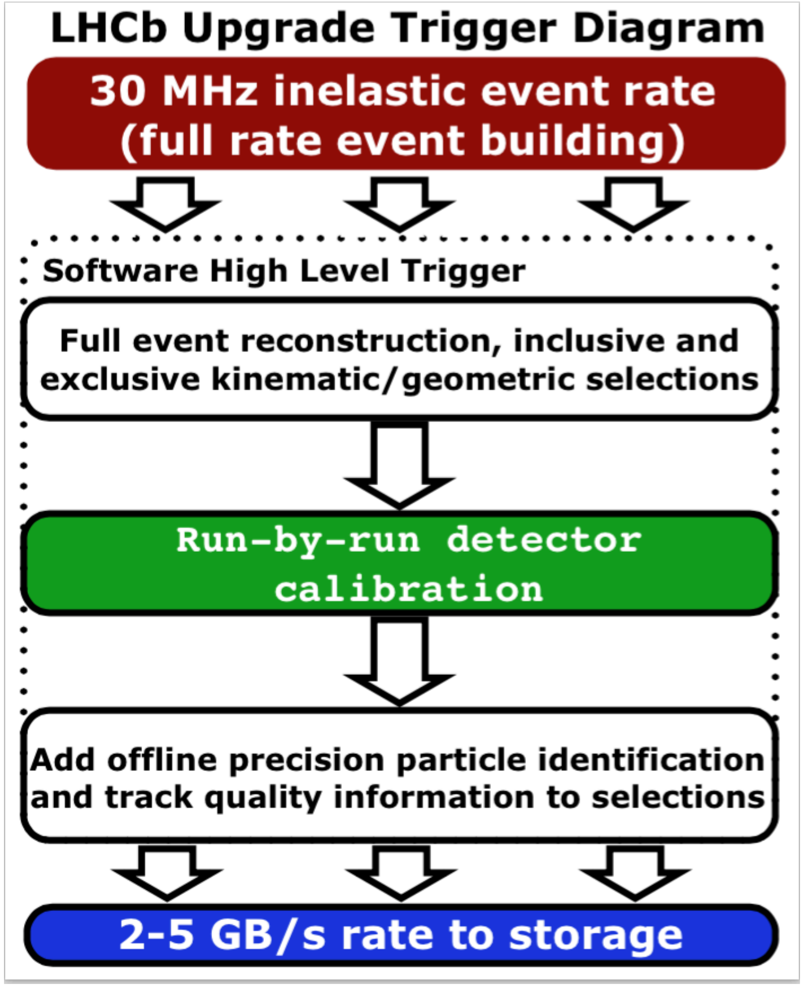
29000 Logical CPU cores

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



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

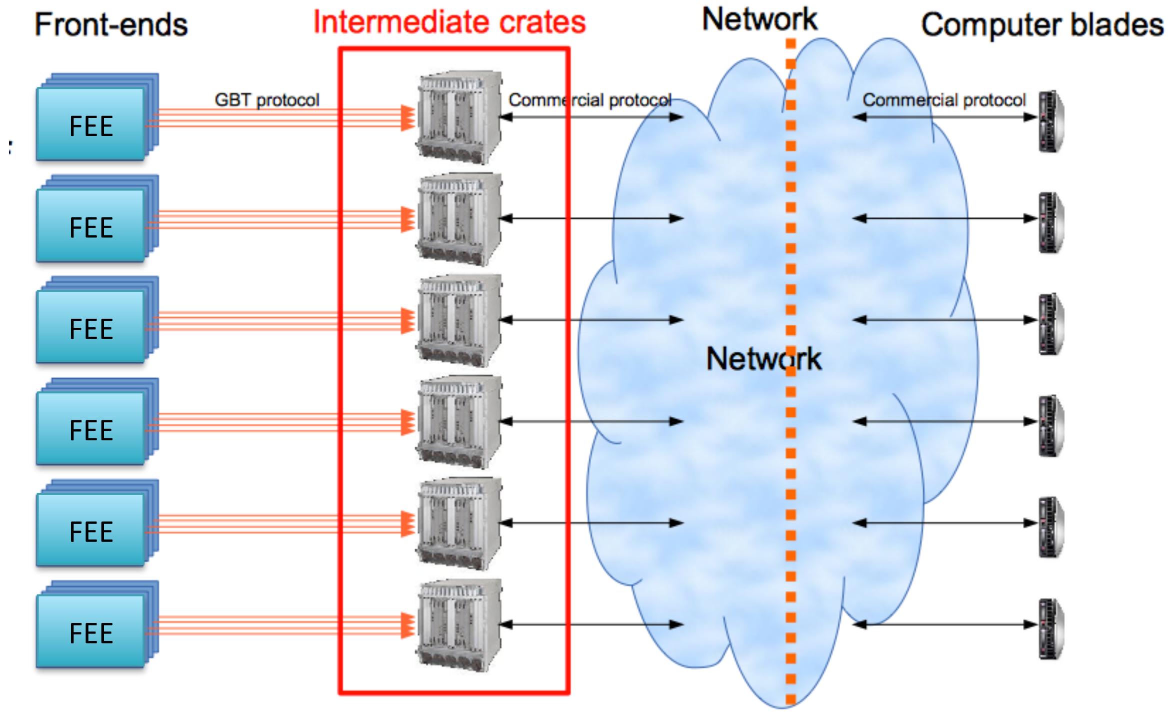
Run-by-run detector calibration

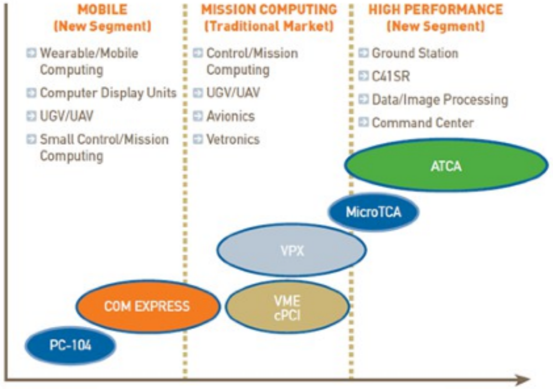
Add offline precision particle identification and track quality information to selections

2-5 GB/s rate to storage

20-50 kHz

LHCb DAQ Upgrade: First Idea





Performance comparison
(source Radisys)

Standard	Power consumption per slot
VPX 3U	75W
VPX 6U	150 W
μTCA	50 to 80 W
ATCA 10G	200W
ATCA 40G	400W

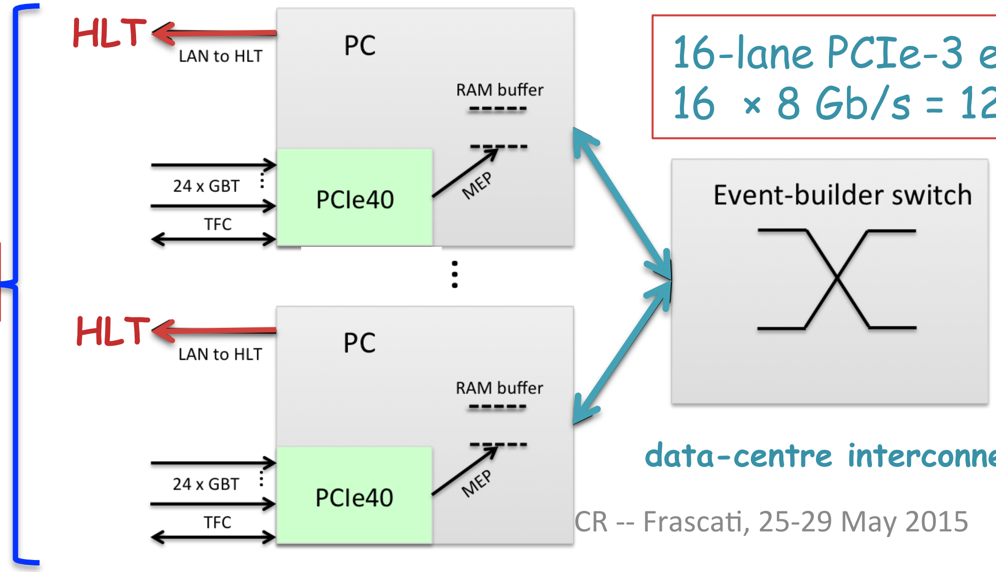
- **Intermediate layer** of electronics boards arranged in crates to decouple FEE and PC farm: for buffering and data format conversion.
- The optimal solution with this approach: **ATCA**, **μTCA** crates, **ATCA** carrier board hosting AMC standard mezzanine boards.
- AMC boards equipped with FPGAs to de-serialize the input streams and transmit event-fragments to the farm, using a standard network protocol, using **10 Gb Ethernet**.

DAQ Present View

- Use **PCIe Generation 3** as communication protocol to **inject data from the FEE directly into the event-builder PC**.
- A **much cheaper** event-builder network
  - **Data-centre interconnects** can be used on the PC:
  - Not realistically implementable on an FPGA (large software stack, lack of soft IP cores,…)
- Moreover PC provides: **huge memory for buffering**, **OS** and **libraries**.
- Up to date NIC and drivers available as pluggable modules.

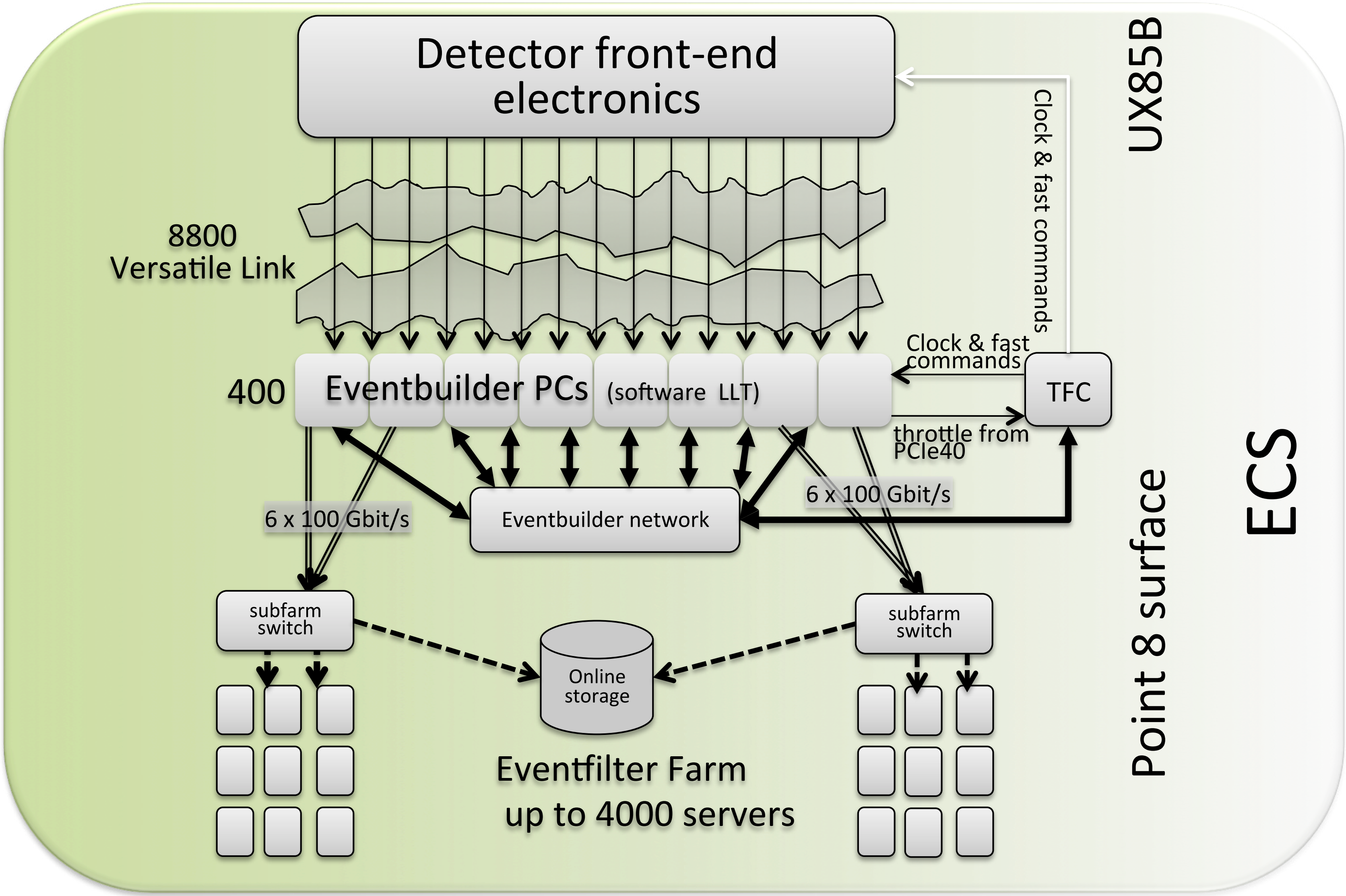
16-lane PCIe-3 edge-connector bandwidth:
16 × 8 Gb/s = 128 Gb/s = **16 GB/s**

400 nodes



data-centre interconnects

Online Architecture after LS2



PCI-e Gen 3 Tests

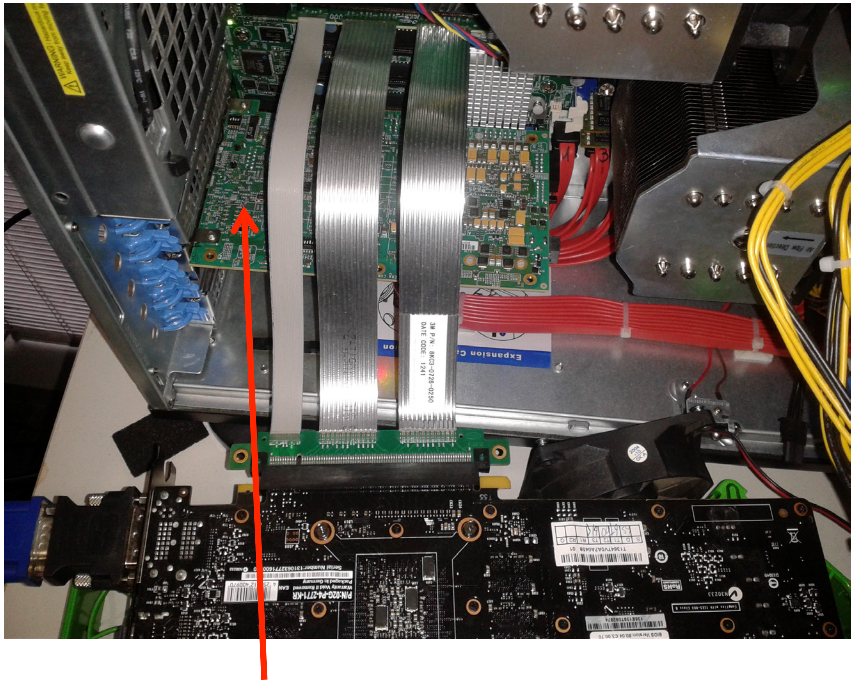
Electronics Front-End

→

Data-Centre Interconnect

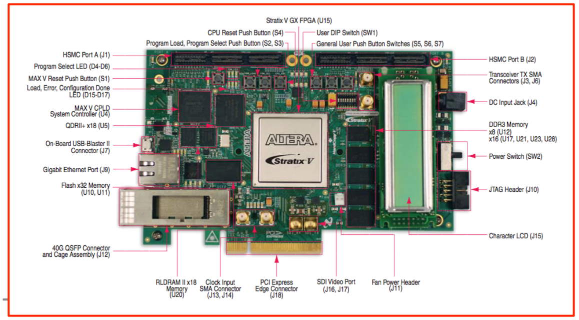
The PCIe-Gen3 DMA Test Setup

- ALTERA evaluation board, Stratix V GX FPGA



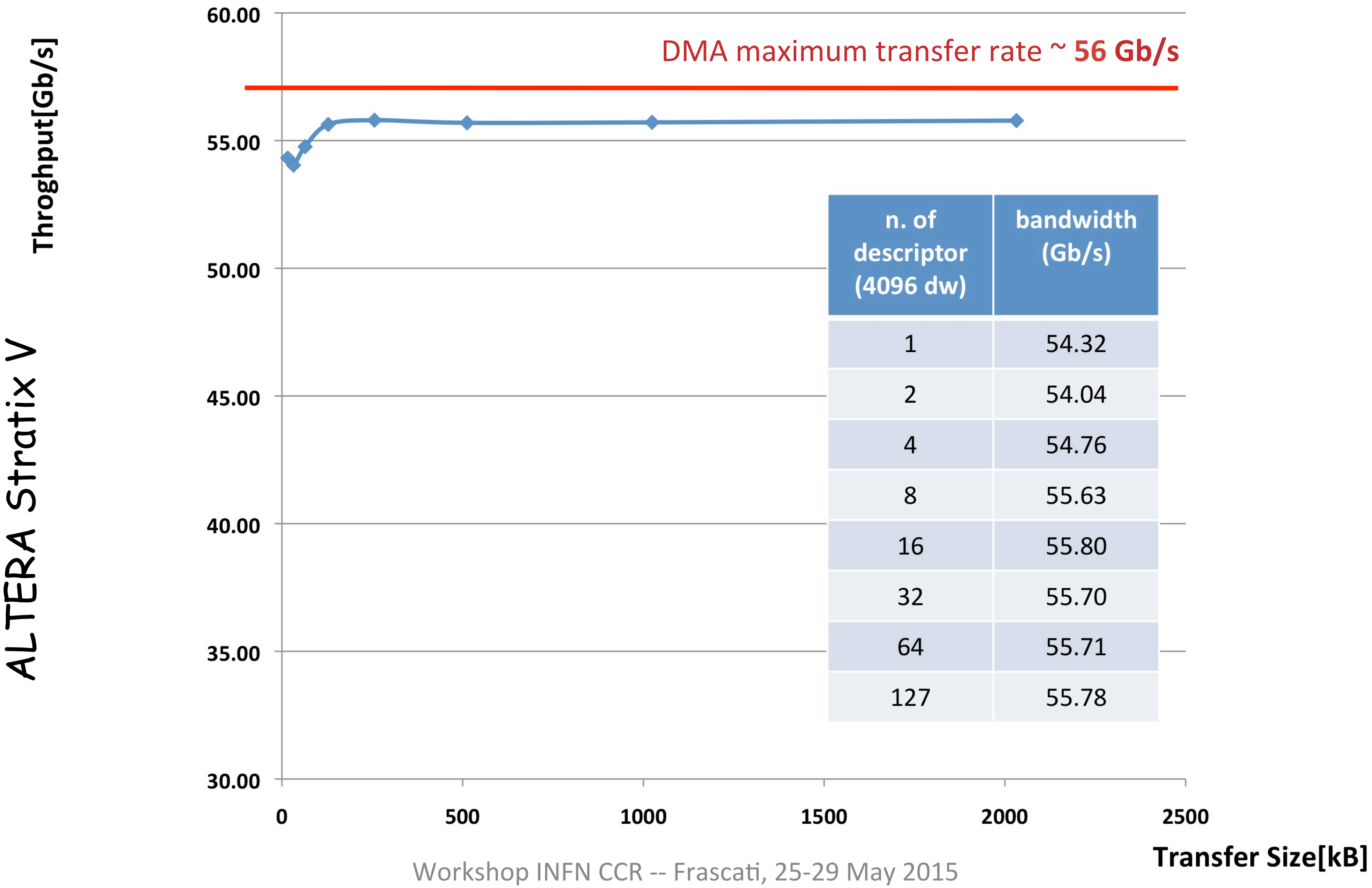
The FPGA provides 8-lane PCIe-3 hard IP blocks and DMA engines.

GPU used to test 16-lane PCIe-3 data transfer between the device and the host memory


Stratix V GX FPGA (U15)
CPU Reset Push Button (S4)
User DIP Switch (SW1)
Program Load, Program Select Push Button (S2, S3)
General User Push Button Switches (S5, S6, S7)
HSMC Port A (J1)
Program Select LED (D4-D6)
MAX V Reset Push Button (S1)
Load, Error, Configuration Done LED (D15-D17)
MAX V CPLD System Controller (U4)
QDRII+ x18 (U5)
On-Board USB-Blaster II Connector (J7)
Gigabit Ethernet Port (J9)
Flash x32 Memory (U10, U11)
40G QSFP Connector and Cage Assembly (J12)
HSMC Port B (J2)
Transceiver TX SMA Connectors (J3, J6)
DC Input Jack (J4)
DDR3 Memory x8 (U12) x16 (U17, U21, U23, U28)
Power Switch (SW2)
JTAG Header (J10)
Character LCD (J15)
RLDRAM II x18 Memory (U20)
Clock Input SMA Connector (J13, J14)
PCI Express Edge Connector (J18)
SDI Video Port (J16, J17)
Fan Power Header (J11)

DMA PCIe-Gen3 Effective Bandwidth

DMA over 8-lane
PCIe-3 hard IP blocks
ALTERA Stratix V



n. of descriptor (4096 dw)	bandwidth (Gb/s)
1	54.32
2	54.04
4	54.76
8	55.63
16	55.80
32	55.70
64	55.71
127	55.78

PCIe-Gen3 Based Readout

- A main FPGA manages the input streams and transmits data to the event-builder PC by using **DMA over PCIe Gen3**.

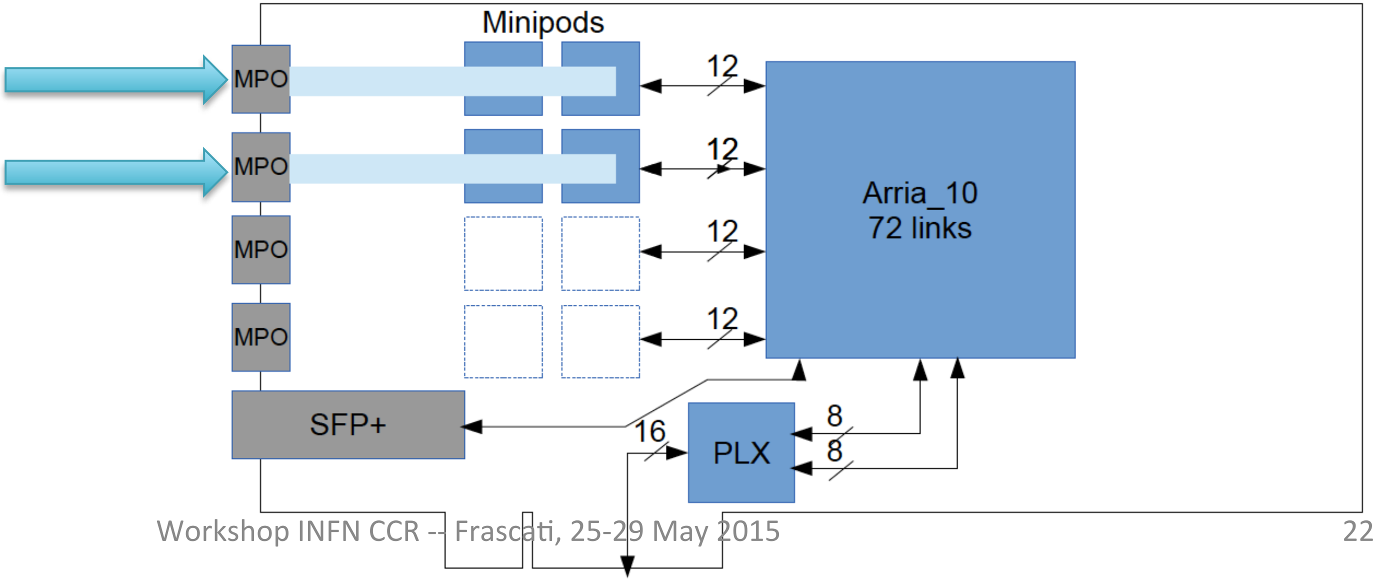
Nominal configuration:

1 bidir link for TFC

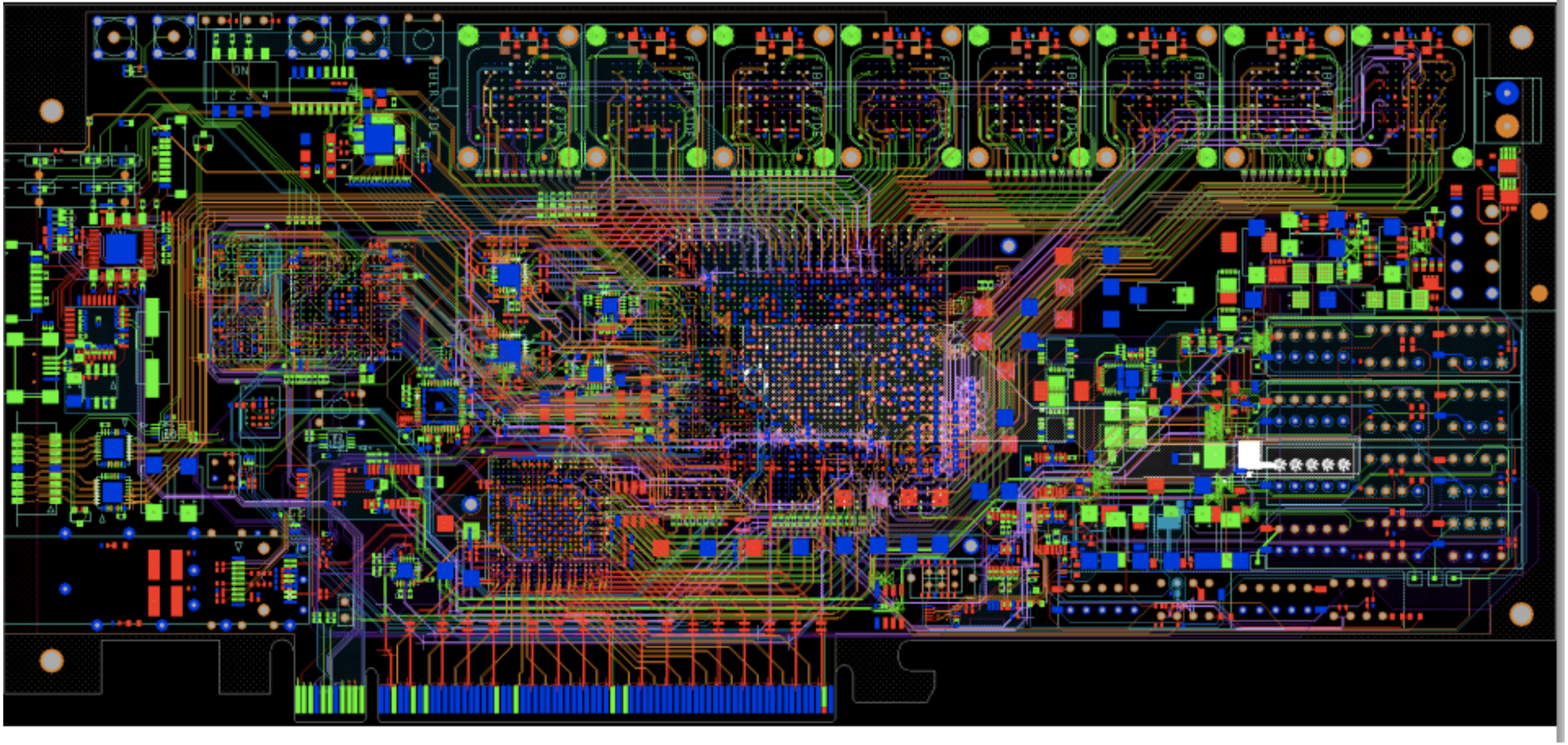
24 GBT inputs → limited by PCIe output bandwidth

- PCIe GEN3 x16 = 110 Gbits/s
- 24 GBT wide bus = 107 Gbits/s

Up to 48 bidir links available on board for low luminosity sub detectors → decrease the costs



PCIe layout



InfiniBand Tests

Event Builder Network

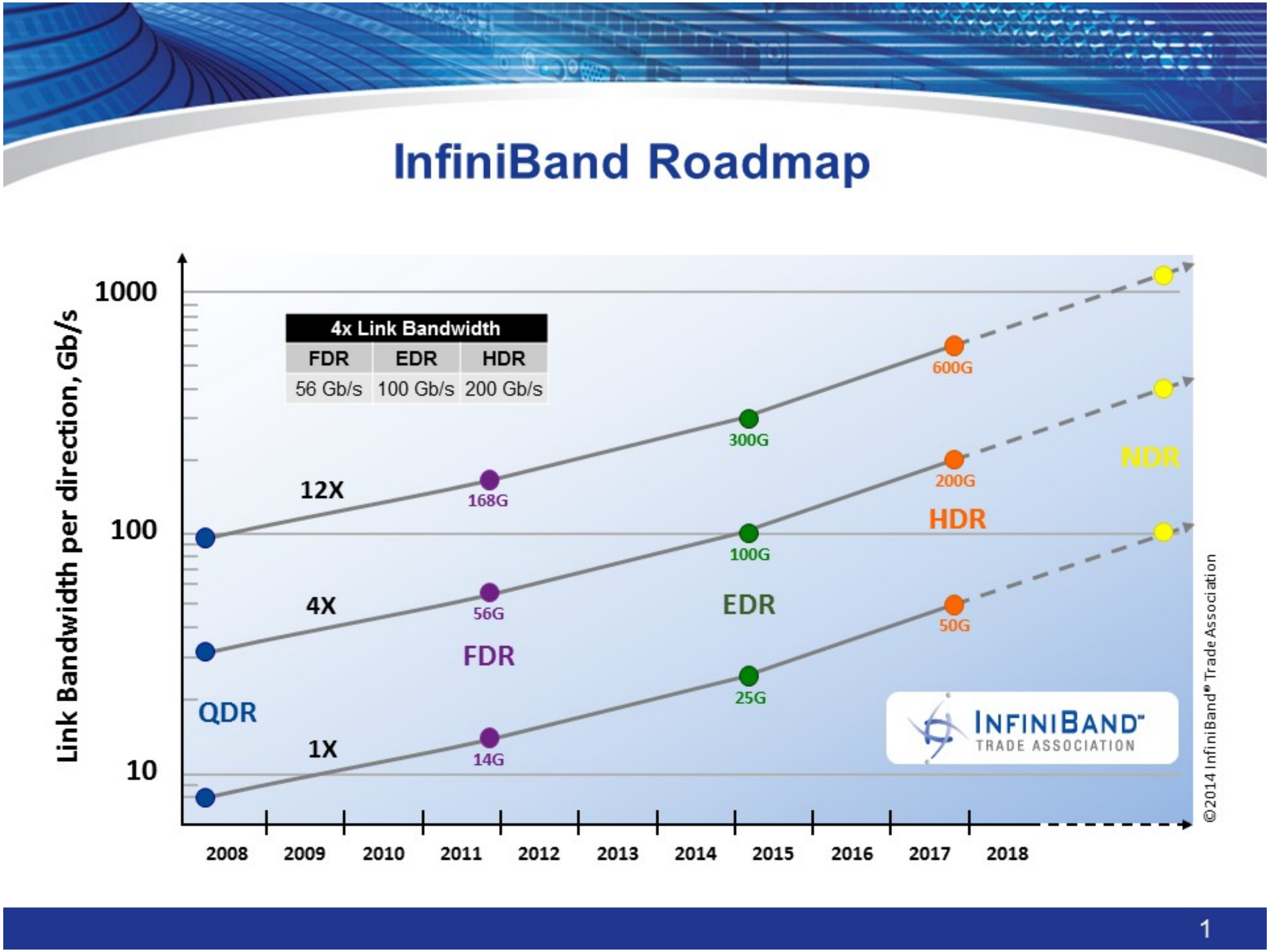
InfiniBand vs Ethernet

- Guaranteed delivery. Credit based flow control:
  - Ethernet: Best effort delivery. Any device may drop packets;
- Hardware based re-transmission:
  - Relies on TCP/IP to correct any errors;
- Dropped packets prevented by congestion management:
  - Subject to micro-bursts;
- Cut through design with late packet invalidation:
  - Store and forward. Cut-through usually limited to local cluster;
- RDMA baked into standard and proven by interoperability testing:
  - Standardization around compatible RDMA NICs only now starting;
  - Need same NICs are both ends;
- Trunking is built into the architecture:
  - Trunking is an add-on, multiple standards an extensions;

InfiniBand vs Ethernet (II)

- All links are used:
  - Spanning Tree creates idle links;
- Must use QoS when sharing with different applications:
  - Now adding congestion management for FCoE but standards still developing;
- Supports storage today;
- Green field design which applied lessons learnt from previous generation interconnects:
  - Carries legacy from it's origins as a CSMA/CD media;
- Legacy protocol support with IPoIB, SRP, vNICs and vHBAs;
- Provisioned port cost for 10 Gb Ethernet approx. **40% higher** than cost of 40 Gb/s InfiniBand.

EB network: Ib vs GbE



InfiniBand Roadmap

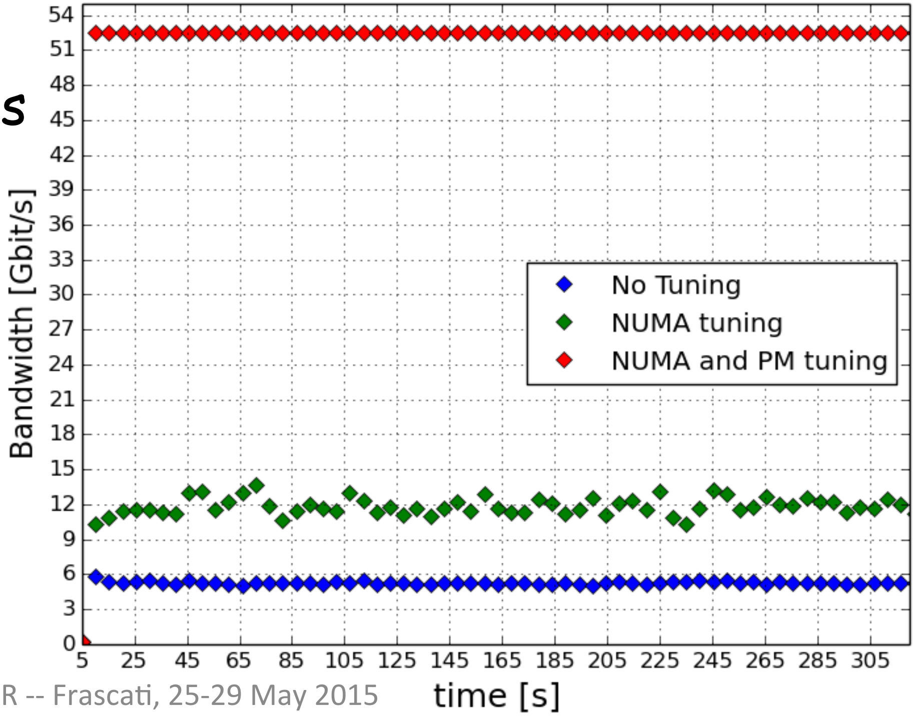
4x Link Bandwidth		
FDR	EDR	HDR
56 Gb/s	100 Gb/s	200 Gb/s

©2014 InfiniBand® Trade Association

1

IB Performance Test

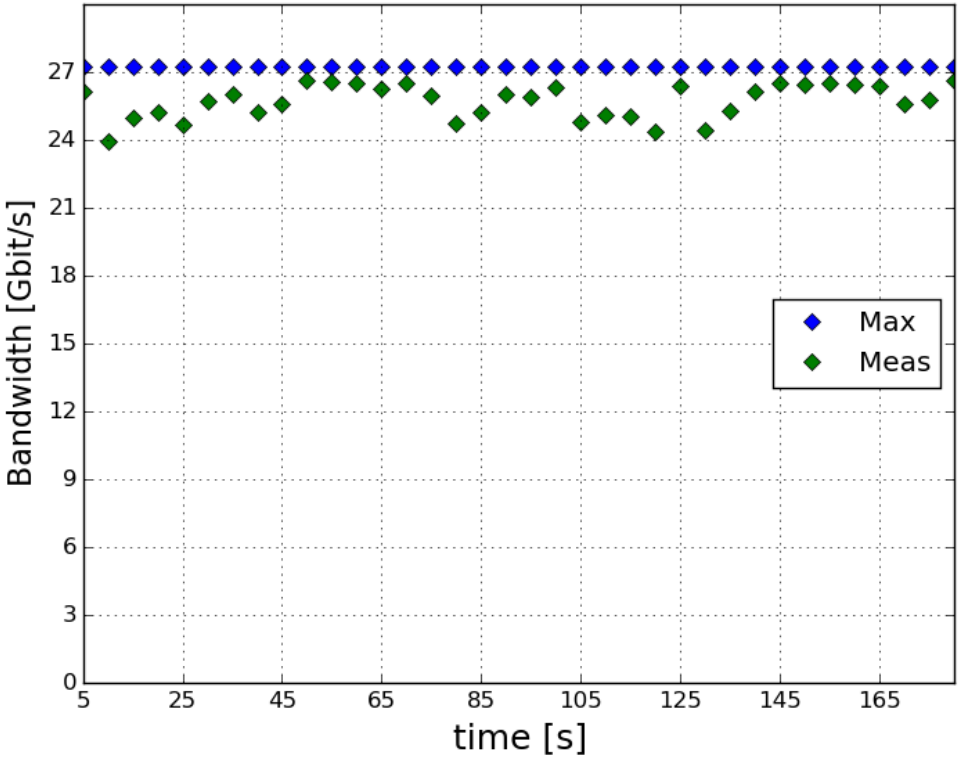
- Performances tests performed at **CNAF**.
- **PCIe Gen 3, 16 lanes** needed:
  - Any previous version of the PCI bus represents a bottleneck for the network traffic;
- Exploiting the best performances required some **tuning**:
  - Disable node interleaving and **bind processes** according to **NUMA** topology;
  - **Disable power saving** modes and CPU frequency selection:
    - **PM** and **frequency switching** are latency sources.



A. Falabella et al

IB Performance Test (II)

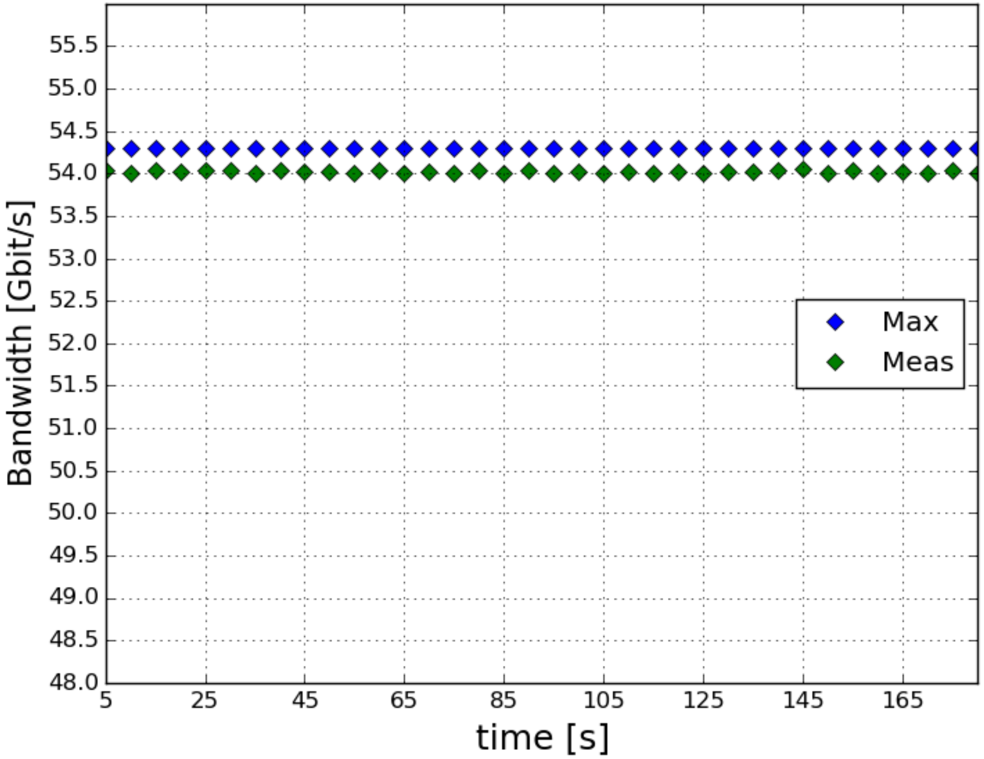
- Ib **QDR** (Quad Data Rate):
    - Point-to-point bandwidth with RDMA write semantic (similar results for send semantic);
    - QLogic : QLE7340, Single port **32 Gbit/s** (QDR);
    - Unidirectional throughput: **27.2 Gbit/s**;
    - Encoding 8b/10b.



A. Falabella et al

IB Performance Test (III)

- Ib **FDR** (Fourteen Data Rate):
  - Point-to-point bandwidth with RDMA write semantic (similar results for send semantic);
  - Mellanox : MCB194A-FCAT, Dual port, **56 Gbit/s** (FDR);
  - Unidirectional throughput: **54.3 Gbit/s** (per port);
  - Encoding 64b/66b.

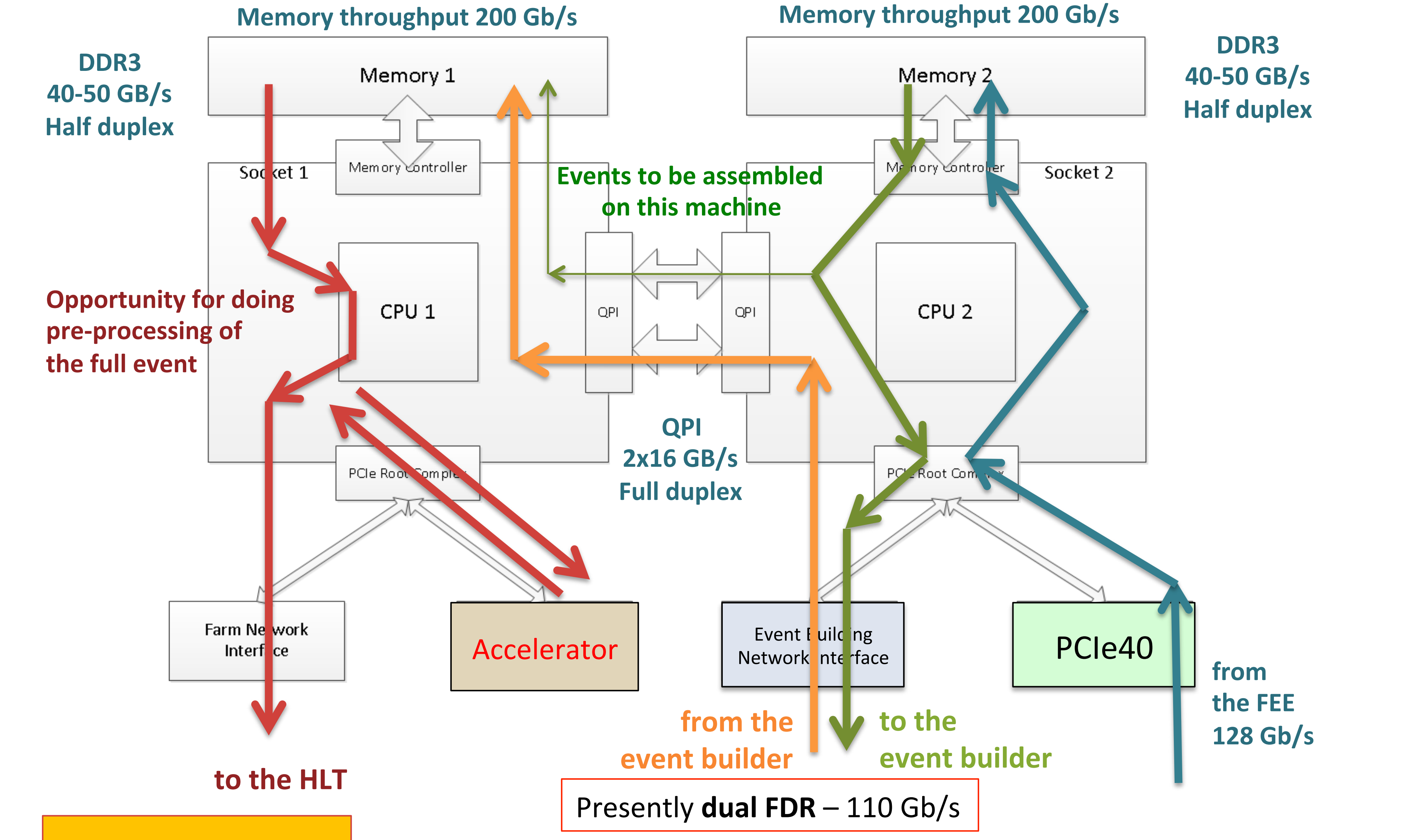


A. Falabella et al

Event Builder Tests

CPU NUMA Architectures
Event Builder Network

Event Builder Fluxes: 400 Gb/s



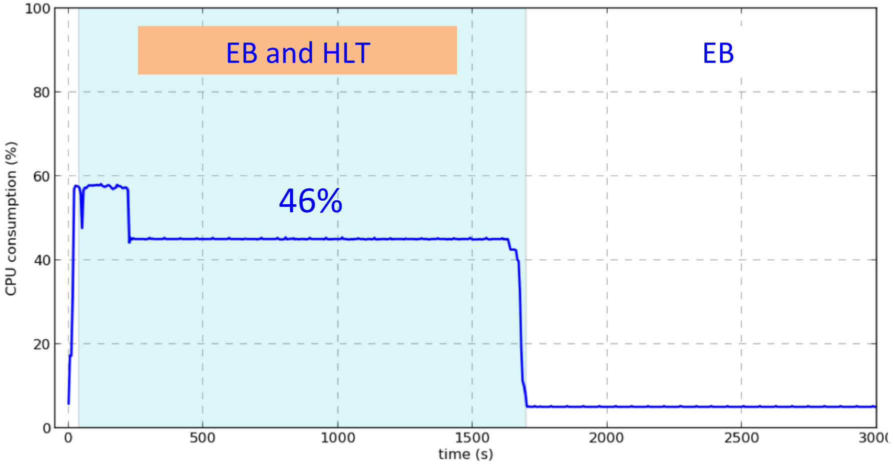
Presently **dual FDR** – 110 Gb/s

… it works!

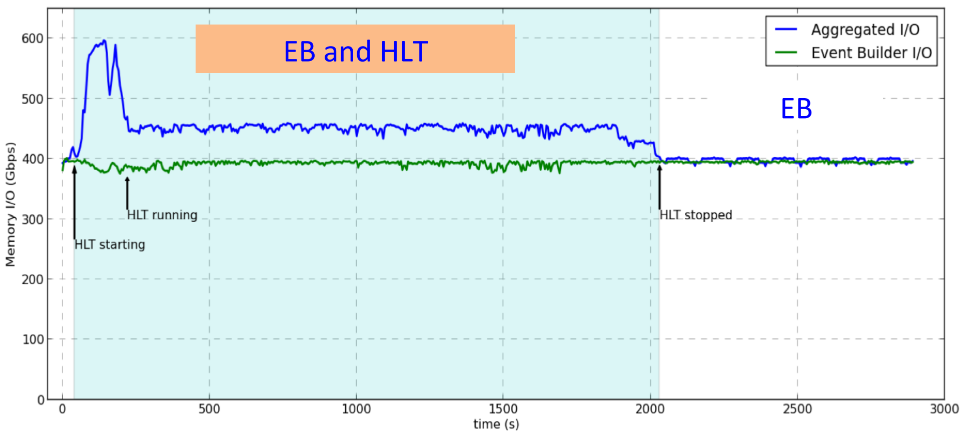
Event Builder CPU Performance

At about 400 Gb/s more than 80% of the CPU resources are free

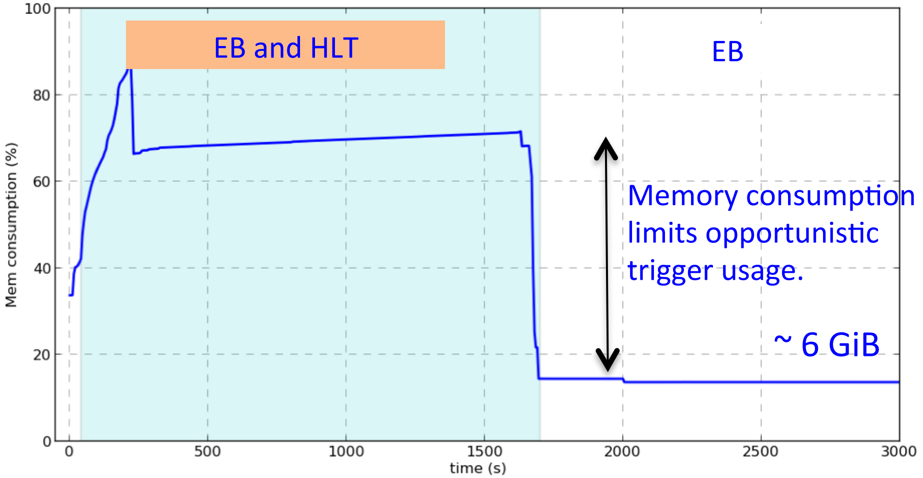
CPU consumption



Memory I/O bandwidth



Memory consumption



Memory consumption limits opportunistic trigger usage.

~ 6 GiB

- PC **sustains** the **event building at 100 Gb/s today**.
- The Event Builder performs **stably** at **400 Gb/s**
- Aggregated **CPU utilization** of EB application and trigger **46%**
- We currently observe **50% free resources for opportunistic triggering on EB nodes**: event builder execution requires about 6 logical core. Additional 18 instances of the HLT software running simultaneously.

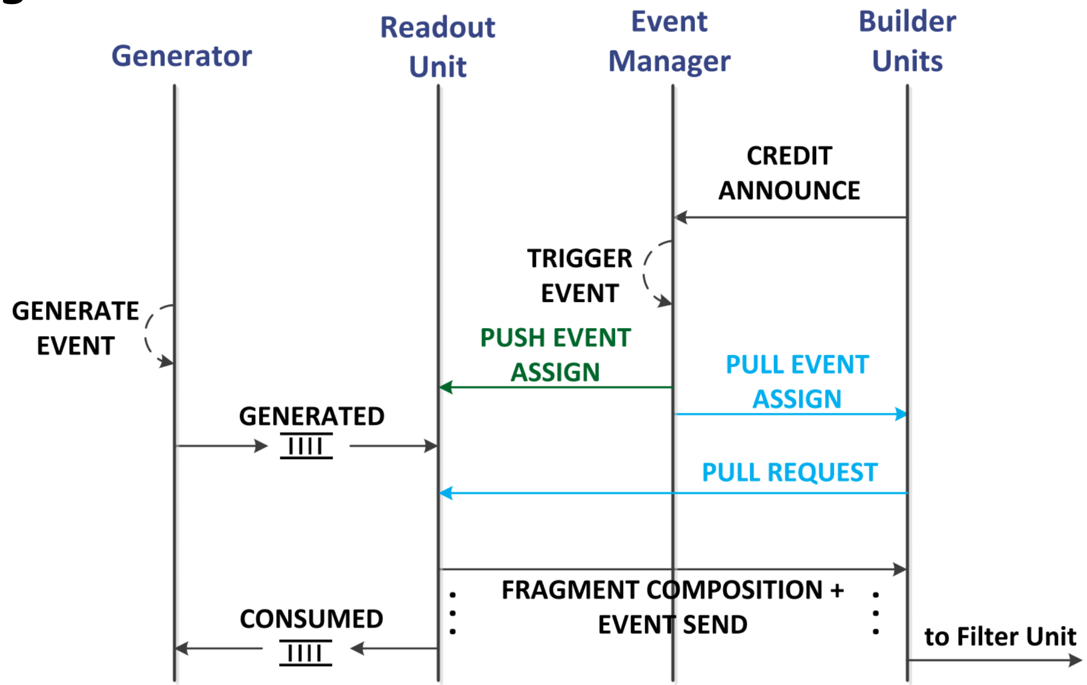
The CPUs used in the test are Intel E5-2670 v2 with a C610 chipset. The servers are equipped with 1866 MHz DDR3 memory in optimal configuration. Hyper-threading has been enabled.

Event Builder Performance

- LHCb-daqpipe software:
  - Allows to test both PULL and PUSH protocols;
  - It implements several trasport layer implementation: IB verbs, TCP, UDP;
- EB software tested 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;
  - On 128 nodes at the 512 nodes Galileo cluster at the Cineca.

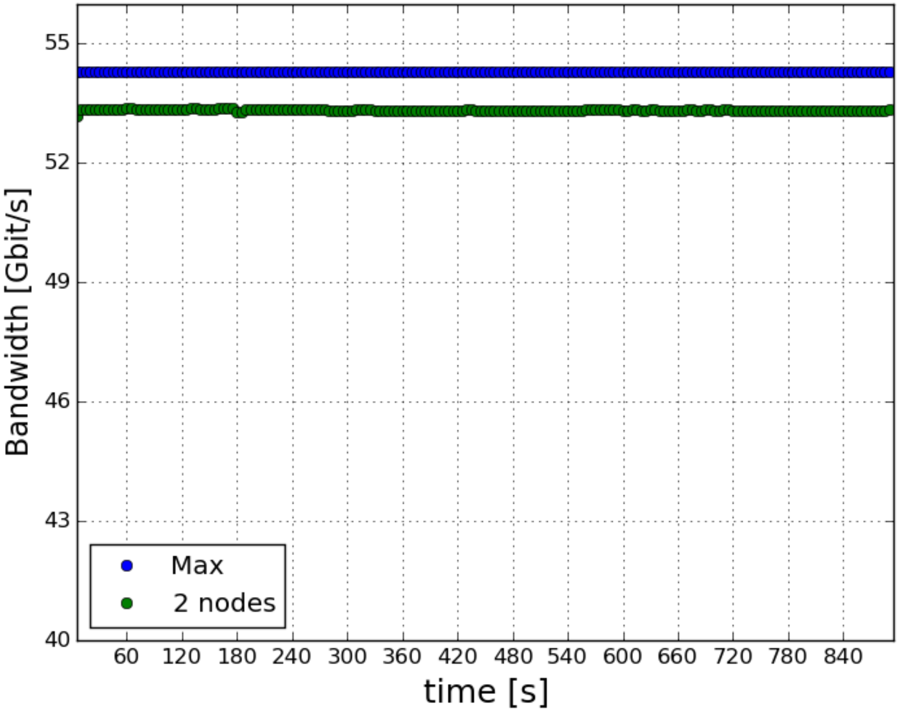
LHCb-daqpipe (II)

- LHCb DAQ Protocol Independent Performance Evaluator;
- LHCb-daqpipe building blocks:
  - 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.



EB Test on 2 Nodes

- Measured bandwidth 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 (average value reported).
- Bandwidth measured is on average **53.3 Gbit/s**:
  - **98%** of maximum allowed;
- PM disabled.

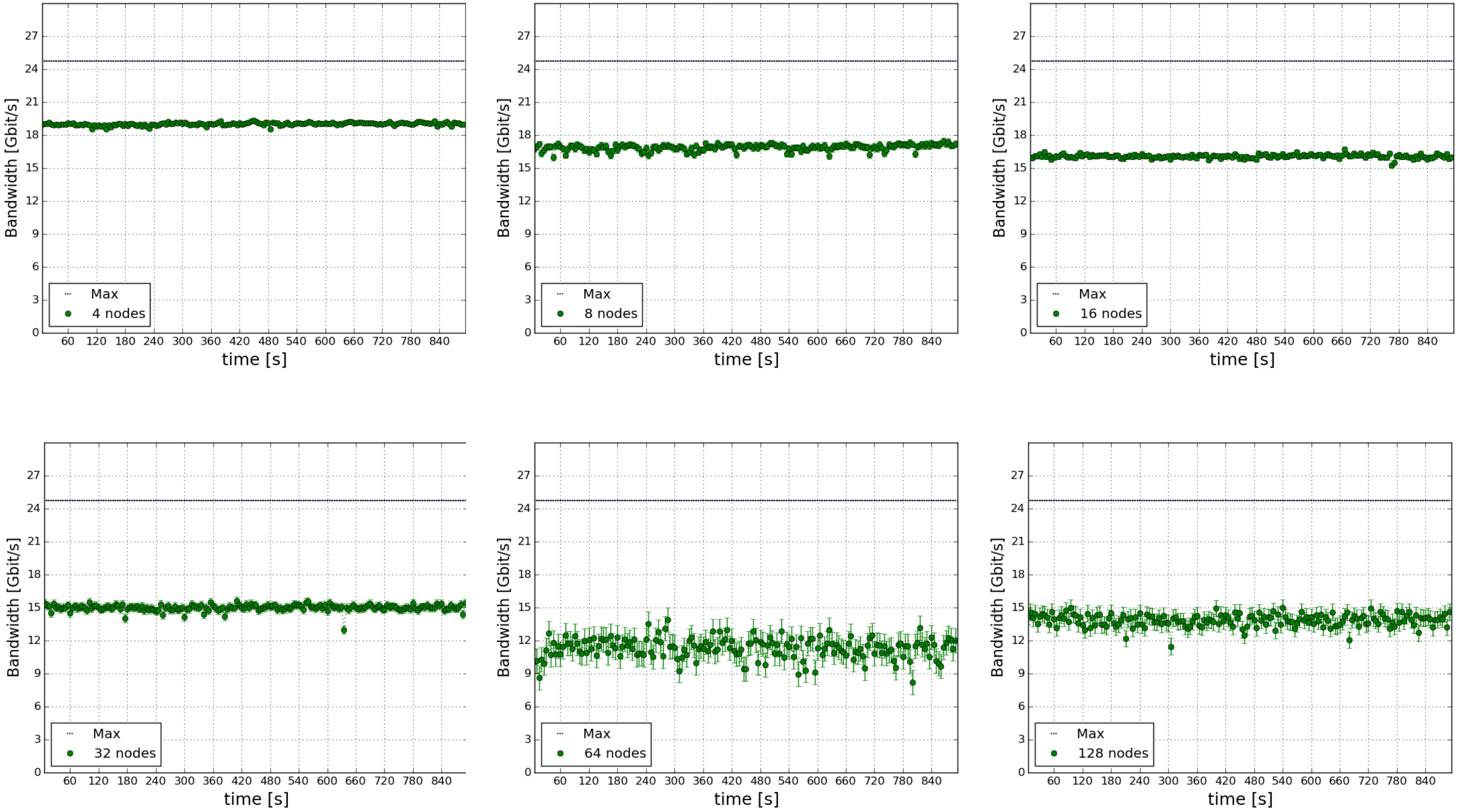


A. Falabella et al

EB Test on 128 Nodes

- Extensive test on the **CINECA Galileo TIER-1 cluster**.
  - **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.
- **Limitations**:
  - Cluster is in production:
    - **Other processes** are polluting the network traffic;
  - **No** control on **power management** and **frequency switching**;
- The fragment composition is performed **correctly** up to a scale of 128 nodes:
  - Maximum allowed for the cluster batch system.

EB Test on 128 Nodes (II)



A. Falabella et al

LHCb Upgrade: Software LLT

- **Throttle mechanism**, while progressively increasing the power of the EFF to run the HLT up to 40 MHz.
- The **LLT algorithms** can be executed in the event builder PC after the event building.
- Preliminary studies show that the LLT runs in **less than 1 ms**, if the CALO clusters are built in the FEE.
- Assuming 400 servers, 20 LLT processes running per PC, and a factor 8 for the CPU power from the Moore Law, the time budget available turns out to be **safely greater then 1 ms**:

$$\boxed{\frac{1}{40MHz} \times 400 \times 20 \times 8 \approx 3.2 \text{ ms}}$$

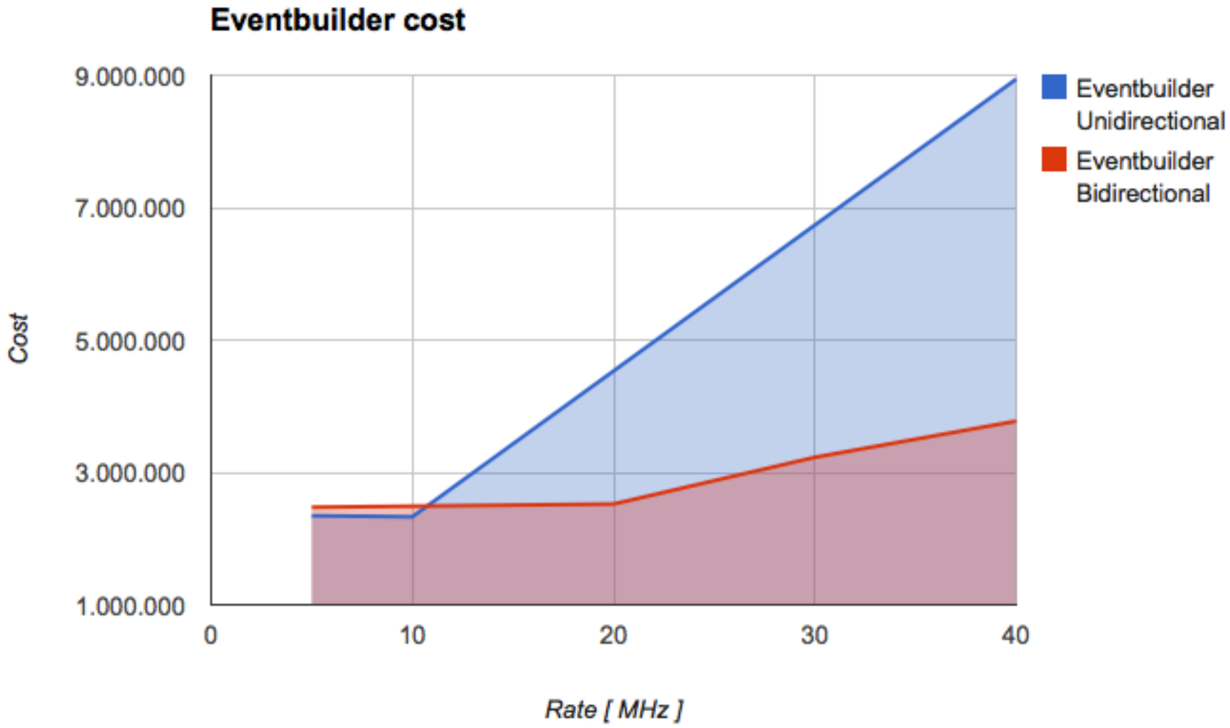
$$\boxed{\text{processing time budget} = \frac{1}{\text{event rate}} \times \text{nodes} \times \text{cores per node} \times \text{task per node}}$$

LHCb Upgrade: HLT Farm

- Trigger-less system at **40 MHz**:
  - A selective, efficient and adaptable software trigger;
- Average **event size**: **100 kB**;
- Expected **data flux**: **32 Tb/s**;
- Total **HLT trigger process latency**: **~15 ms**:
  - Tracking time budget (VELO + Tracking + PV searches): 50%
  - Tracking finds **99%** of offline tracks with p_T>500 MeV/c
- Number of **running trigger process** required: **4×10^5**;
- Number of **core/CPU** available in 2018: **~ 200**:
  - Intel tick-tock plan: 7 nm technology available by 2018-19, the number of core accordingly scales as
    $12 \times (32\ \text{nm}/\ 7\ \text{nm})^2 = 250$, equivalent 2010 cores.
- Number of **computing nodes** required: **~ 1000**.

Scaling and Cost

- **Unidirectional**: scaling the present LHCb architecture to 40 MHz, use of intermediate crates, ATCA and AMC board and cables, 10 and 40 GbEthernet. **Cost to operate at 40 MHz: 8.9 MCHF**.
  The cost due to the ATCA crate has not been included.
- **Bidirectional**: PCIe and InfiniBand proposed approach.
  Cost to operate at 40 MHz: 3.8 MCHF.



Involved Institutes

- INFN-Bologna: Umberto Marconi, Domenico Galli, Vincenzo Vagnoni, Stefano Perazzini et al.;
- Laboratorio di Elettronica INFN-Bologna: Ignazio Lax, Gabriele Balbi et al.;
- INFN-CNAF: Antonio Falabella, Francesco Giacomini, Matteo Manzali et al.;
- INFN-Padova: Marco Bellato, Gianmaria Collazuol et al.;
- CERN: Niko Neufeld, Daniel Hugo Cámpora Pérez, Guoming Liu, Adam Otto, Flavio Pisani, et al.;
- Altri…

Spare material

LHCb Upgrade: Consequences

- The **detector front-end electronics** has to be entirely **rebuilt**, because of the current readout speed is limited to 1 MHz.
  - Synchronous readout, no trigger.
  - No more buffering in the front-end electronics boards.
  - Zero suppression and data formatting before transmission to optimize the number of required links.
    - **Average event size 100 kB**
  - Three times the optical links as currently to get the required bandwidth, needed to transfer data from the front-end to the read-out boards at 40 MHz.
    - **GBT links simplex (DAQ) 9000, GBT duplex (ECS/TFC) 2400**
- **New HLT farm** and **network** to be built by exploiting new LAN technologies and powerful many-core processors.
- Rebuild the current sub-detectors equipped with embedded front-end chips.
  - Silicon strip detectors: VELO, TT, IT
  - RICH photo-detectors: front-end chip inside the HPD.
- Consolidate sub-detectors to let them stand the foreseen luminosity of $20.\times10^{32}$ cm^{-2} s^{-1}