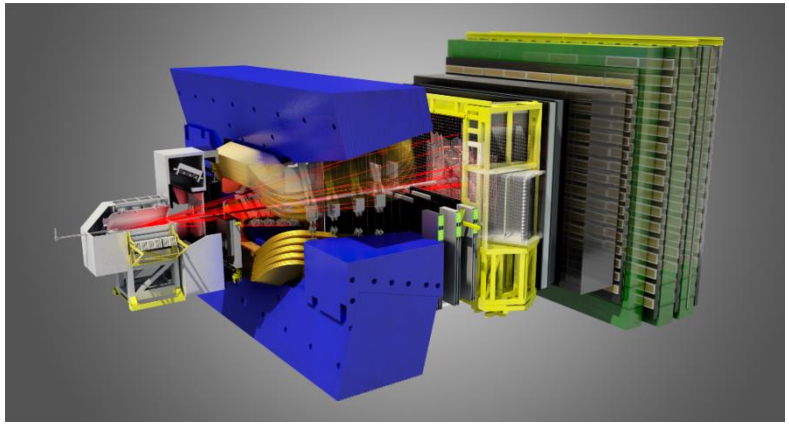
The LHCb triggerless readout system for LHC Run 3 and beyond

CepC workshop,
18–20 November 2019,
IHEP Beijing, China

Federico Alessio, CERN
on behalf of the LHCb Collaboration

Outline

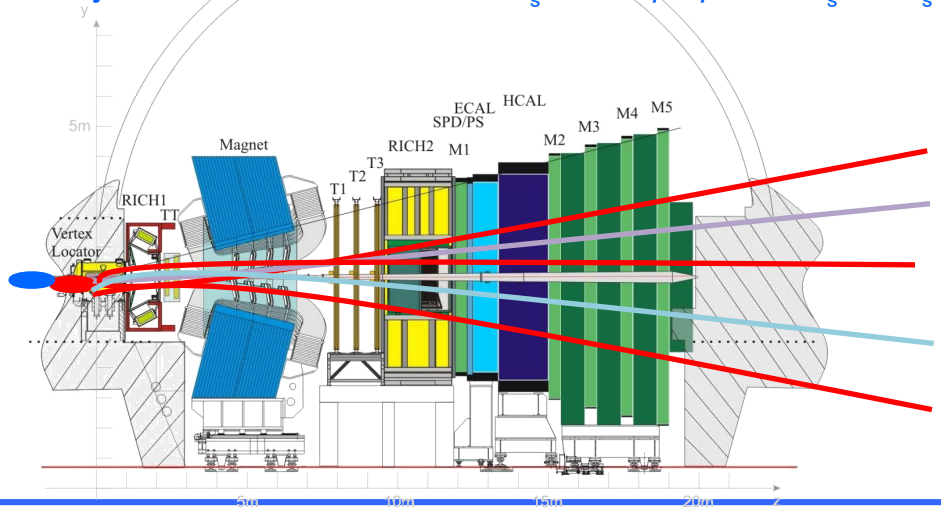
- Introduction to LHCb
- Motivations for an upgrade of the LHCb detector and detector upgrade
- Readout Architecture Upgrade:
  - Trigger-less FE electronics
  - Timing distribution
  - Readout boards
  - Event building
  - Automated run control
  - Monitoring system
  - Data center
- Checklist for your trigger-less readout system

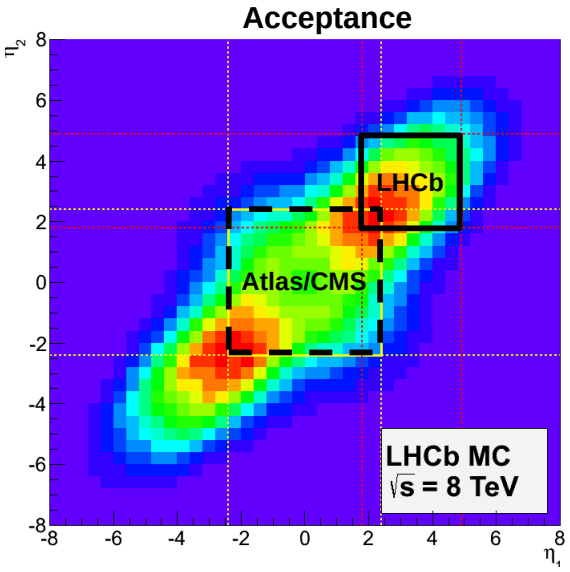


Old LHCb detector

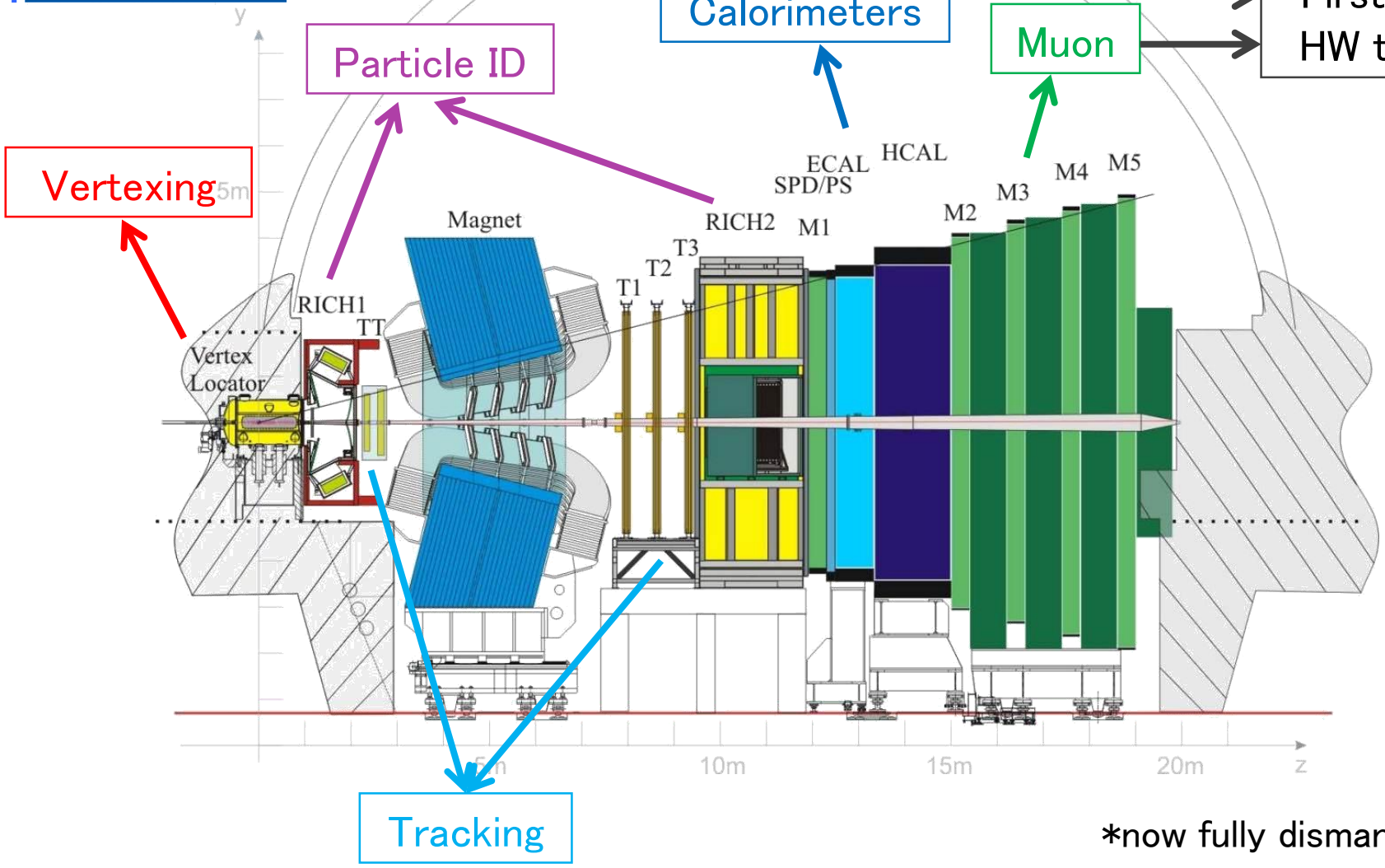
LHCb proved itself to be the Forward General-Purpose Detector at the LHC:

- forward arm spectrometer with unique coverage in pseudorapidity ($2 < \eta < 5$, 4% of solid angle)
- catching 40% of heavy quark production cross-section
- precision measurements in beauty and charm sectors
  - ✓ $\Delta p / p$ = 0.4% at 5 GeV/c to 0.6% at 100 GeV/c
  - ✓ impact parameter resolution 20 μm for high-pT tracks
  - ✓ decay time resolution 45 fs for $B_s \rightarrow J/\psi\ \phi$ and $B_s \rightarrow D_s\ \pi$



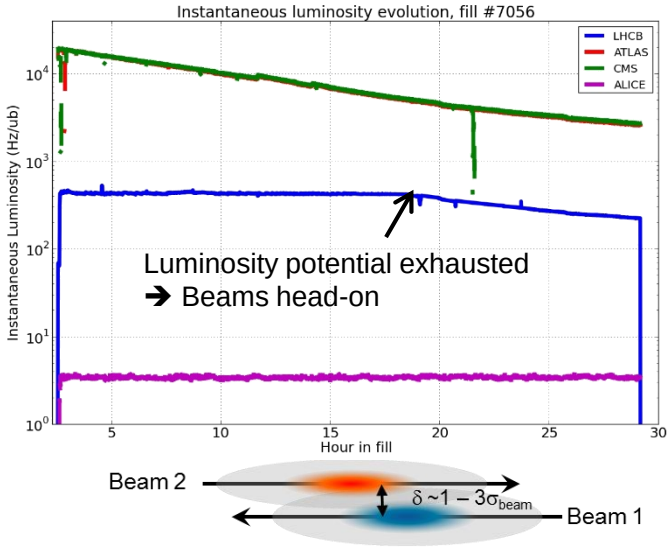


Old LHCb detector*



*now fully dismantled...

LHCb operation performance in Run1+2



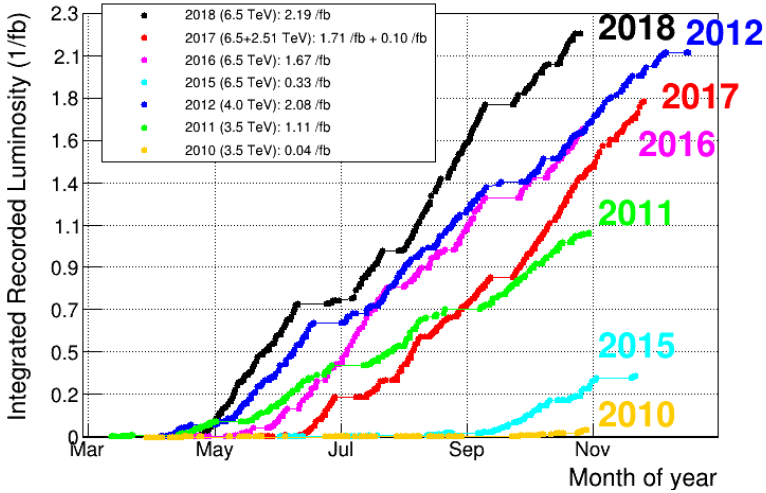
In Run1 and Run 2, LHCb had excellent performance :

- luminosity leveling at constant $4x10^{32}$ $cm^{-2}s^{-1}$ with a constant ~1.5 interactions per LHC crossing

  > 2x designed values!

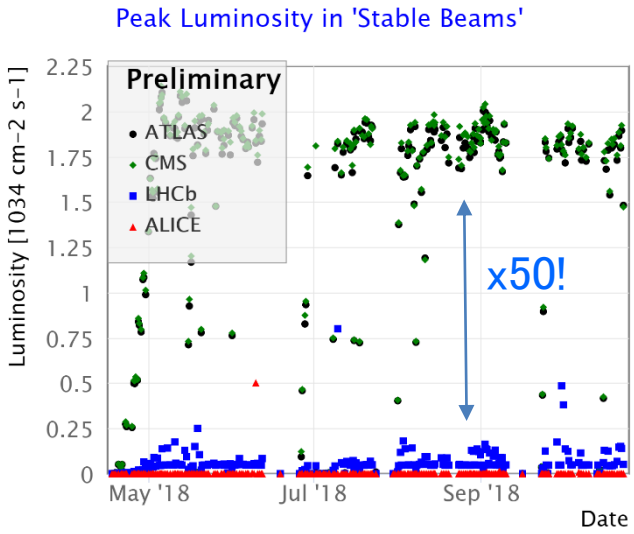
- >9 fb^{-1} data recorded with overall efficiency ~93%
- >99% detector channels working and operational
- >99% of online data good for offline analysis
- >96% efficiency for long tracks in track reco
- >90% ParticleID efficiencies

But this is not enough!
Flavor physics is all about statistics and precision measurement!



Why upgrading LHCb?

The amount of data and the physics yield from data recorded by the current LHCb experiment is limited by its detector.



While LHC accelerator will keep steadily increasing

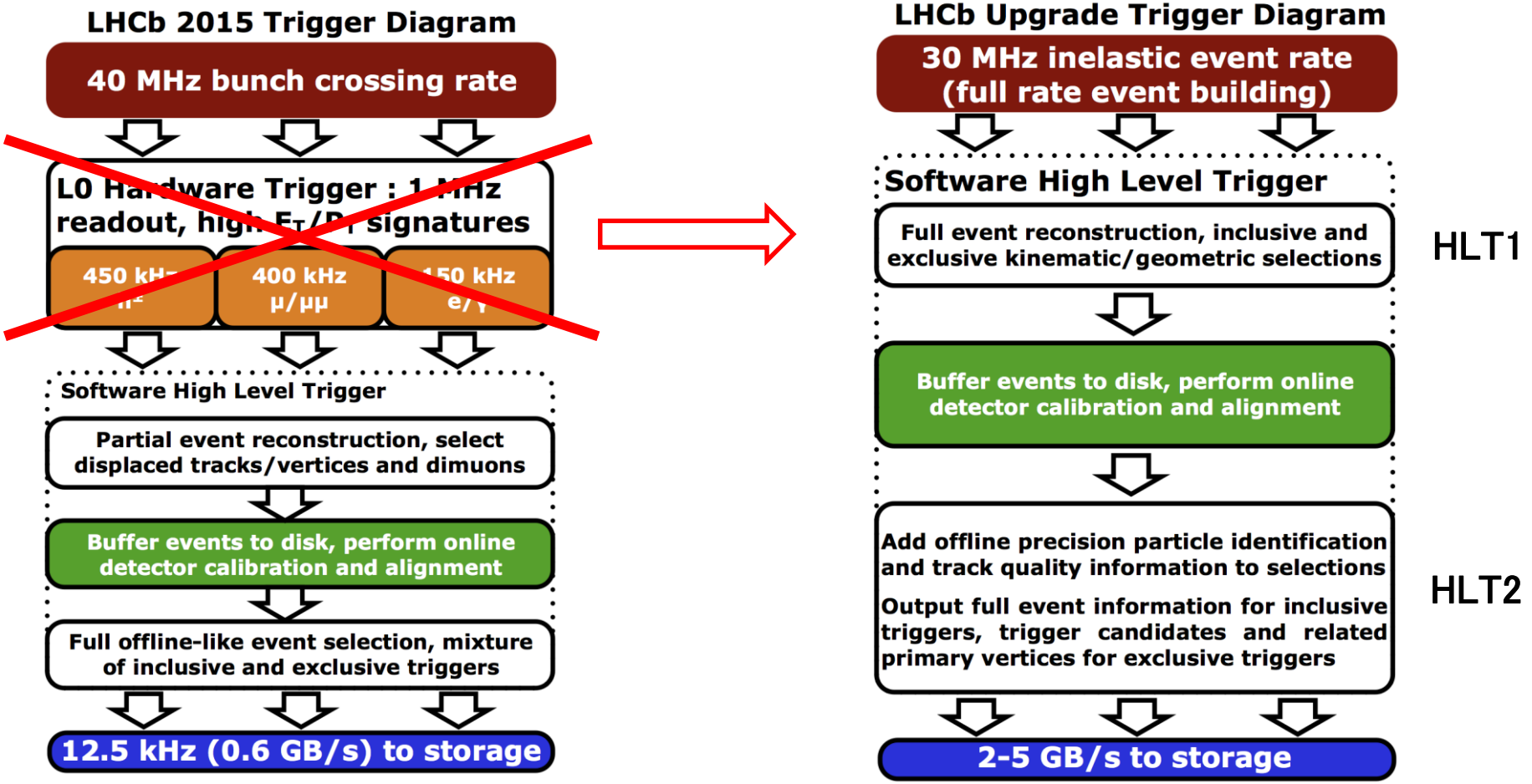
- energy / beam (3.5 → 4 → 6.5 TeV → ?)
- luminosity (peak $8x10^{33}$ → $2x10^{34}$ $cm^{-2}s^{-1}$ → ?)

LHCb will stay limited in terms of

- data bandwidth: limited to 1.1 MHz / 40 MHz max
- physics yields for hadronic channels at the hardware trigger
  - → Major limitations from harsher cuts applied on p_T and E_T at first level trigger
- detectors degradation at higher luminosities

Upgrade strategy: triggerless readout

→ remove the first-level hardware trigger!

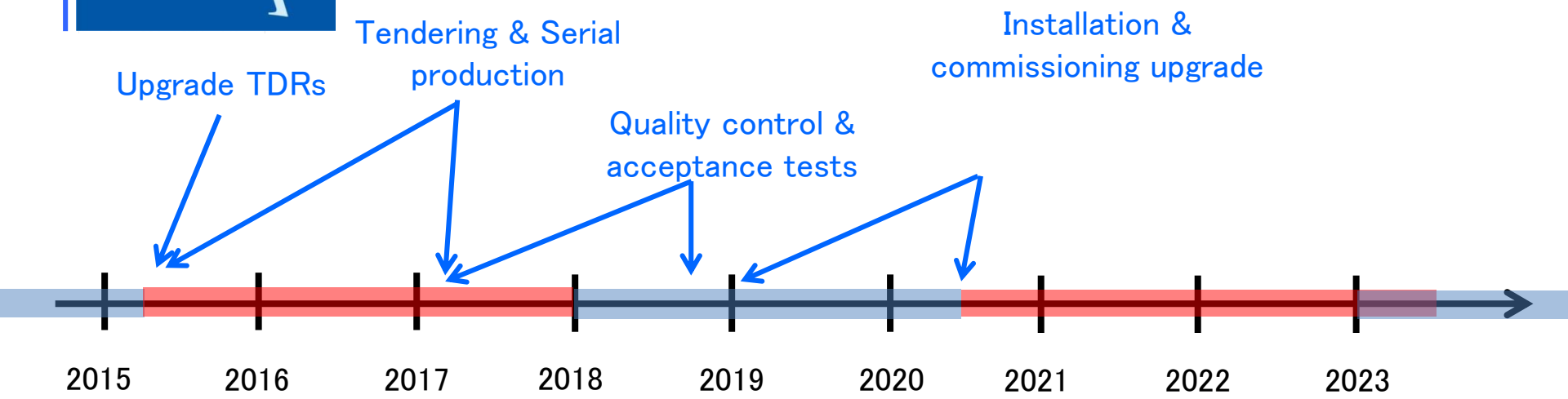


Implications of upgrade strategy

Removal of first-level hardware trigger implies

- read out every LHC bunch crossing
  - trigger-less Front-End electronics
  - multi-Tb/s readout network
- fully software flexible trigger
  - full event information available to improve trigger decision
  - maximize signal efficiencies at high events rate
  - online selections ~ identical to offline selections
- higher luminosities:
  - redesign (incompatible) sub-detectors for a peak luminosity of $2 \times 10^{33} cm^{-2} s^{-1}$
- more data by increasing bandwidth:
  - redesign readout architecture to record 40 MHz events

LHCb upgrade I plan



LHC LSI
Technical reviews & technology choices

LHC Run II
E = 13 TeV (pp)
@ 25 ns
+
$L_{peak}= 4x10^{32}cm^{-2}s^{-1}$
+
Pb^+Pb^+ and pPb^+

LHCb collected 9 fb^{-1} over 10 years

LHC LSII
LHCb Upgrade I

LHC Run III
E = 14 TeV @ 25 ns
+
$L_{peak}= 2x10^{33}cm^{-2}s^{-1}$
+
Pb^+Pb^+ and pPb^+

LHCb to collect > 30 fb^{-1} in the next 3 years

LHC LSIII
HL-LHC

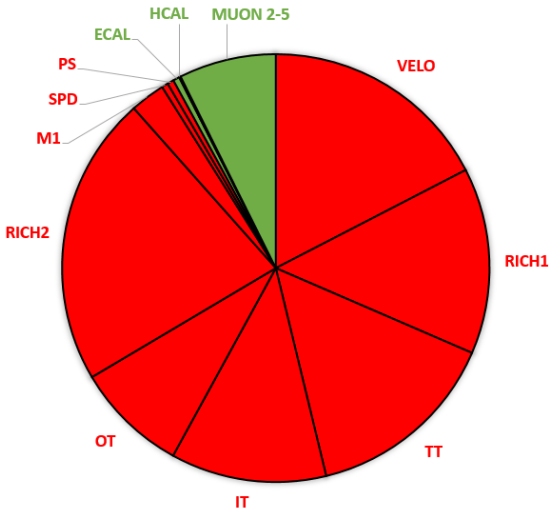
The upgrade of LHCb

Upgraded LHCb Detector

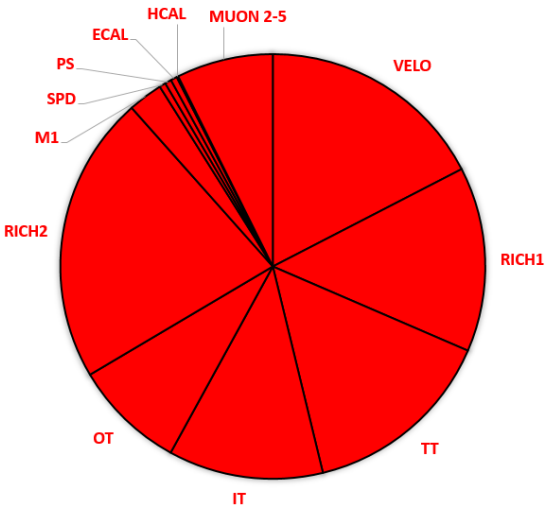
To be UPGRADED

To be kept

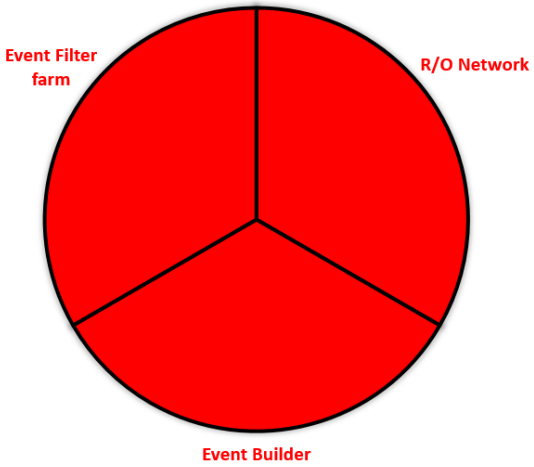
Detector Channels



R/O Electronics



DAQ



Major upgrade → it's a new detector all together.

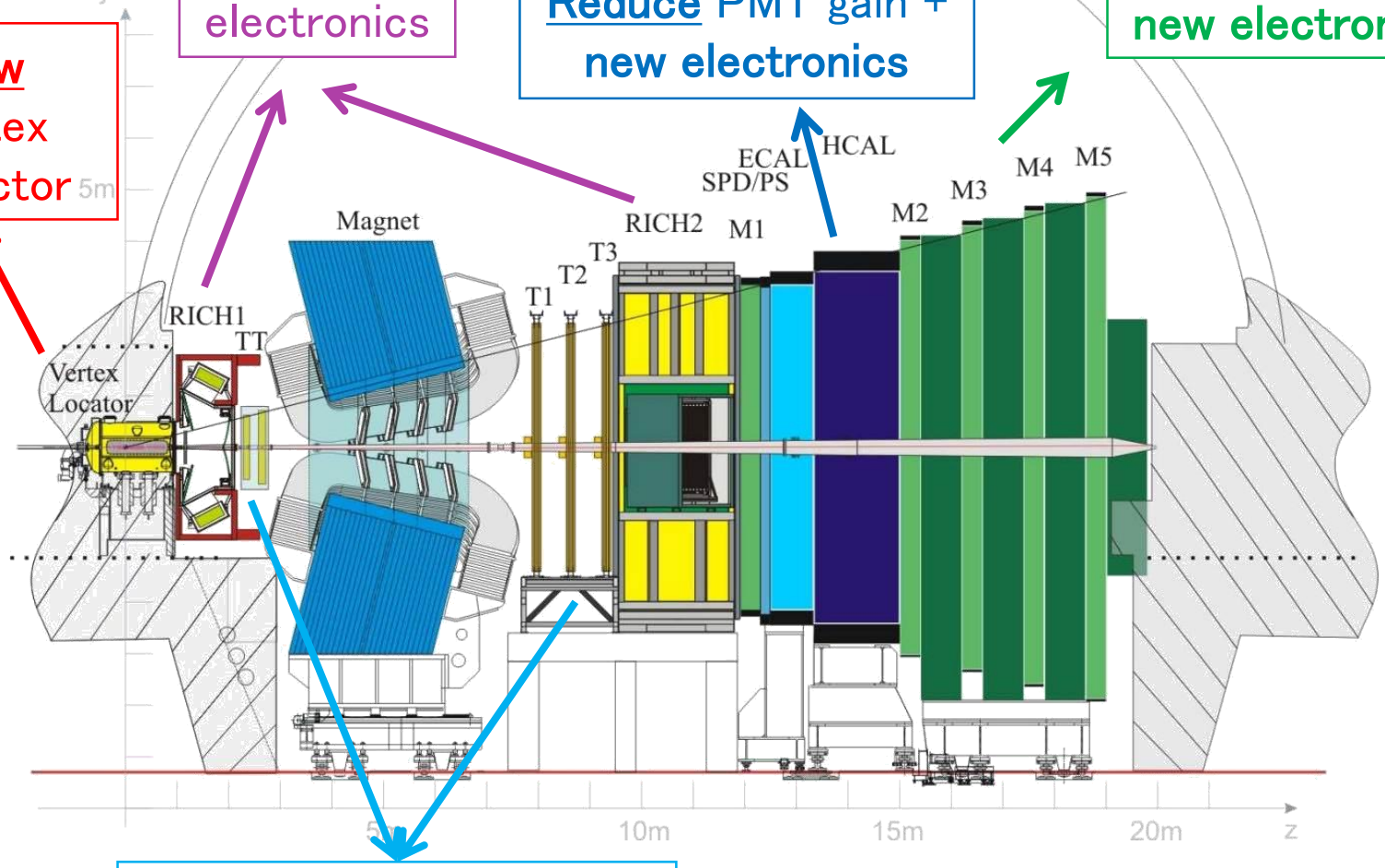
Upgraded LHCb detector

Particle ID
Replace
HPDs +
electronics

Calorimeters
Reduce PMT gain +
new electronics

Muon
new electronics

New
Vertex
Detector



New Tracking stations

Upgraded Readout Architecture

Full software trigger @ 40 MHz

Detector data from underground to surface

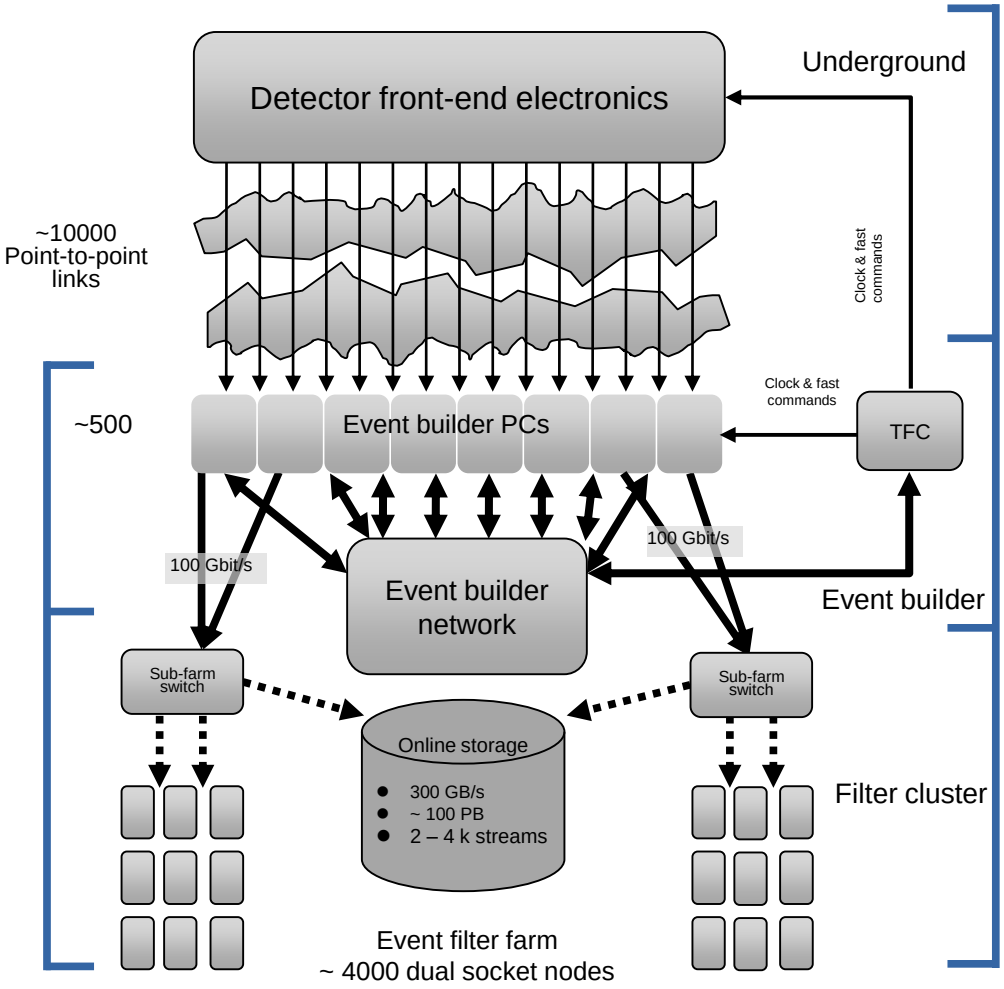
- ~ 300m OM3 optical fibers

Two separate networks:

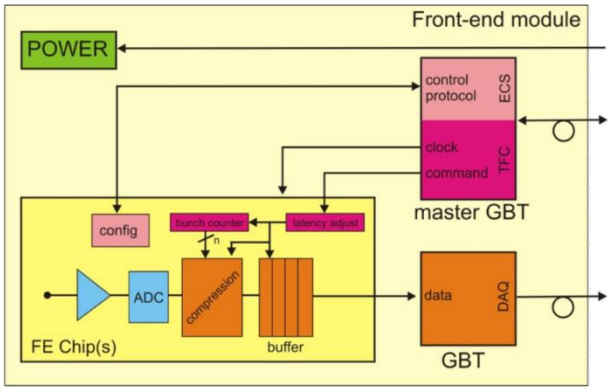
- Event Building
- Filter Farm

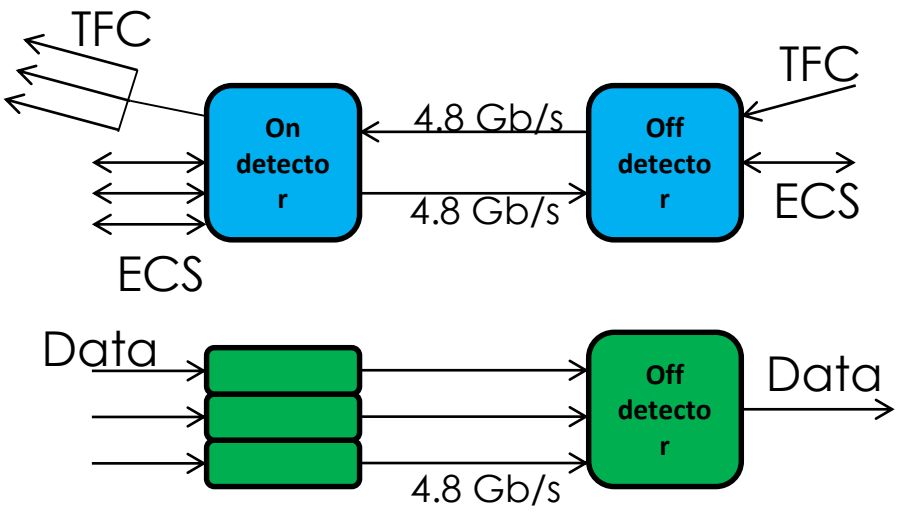
Network technologies:

- Mellanox InfiniBand HDR
- 100 / 200 Gb/s Ethernet



Trigger-less Front-Ends





- Need to compress (zero-suppress) data already at the FE to reduce data throughput
  - reduce # of links from ~80000 to ~12500 (20 MCHF to 3.1 MCHF)
- Separate network from data network: duplex for control, simplex for data
  - Use link bandwidth efficiently for data
    - → Pack data across data link continuously with elastic buffer before link
  - Compact links merging Timing, Fast (TFC) and Slow Control (ECS).
    - → Extensive usage of the CERN GBT and associated ASICs development
    - → Common tools and implementation for all sub-detectors → homogeneity

Data driven asynchronous readout and allow/account for variable (and large) latencies!

A common and generic hardware

LHCb developed a custom-made hardware readout card: PCIe40

- PCIe Gen 3 x8x8 pluggable card in commercial server
  - Validated up to ~90 Gb/s sustained
- 48 bidirectional or unidirectional high-speed links
  - Used at 5 Gb/s with custom protocol
  - MiniPod optics onboard
  - 12 links MPO connectors front panel

  + 2 dedicated SFP+/PON links for timing distribution
- Altera Arria X FPGA w/ embedded transceivers
  - Custom made protocol (CERN GBT)
- ~500 cards being produced
- Designed at CPPM, produced at FEDD, tested and validated at CERN

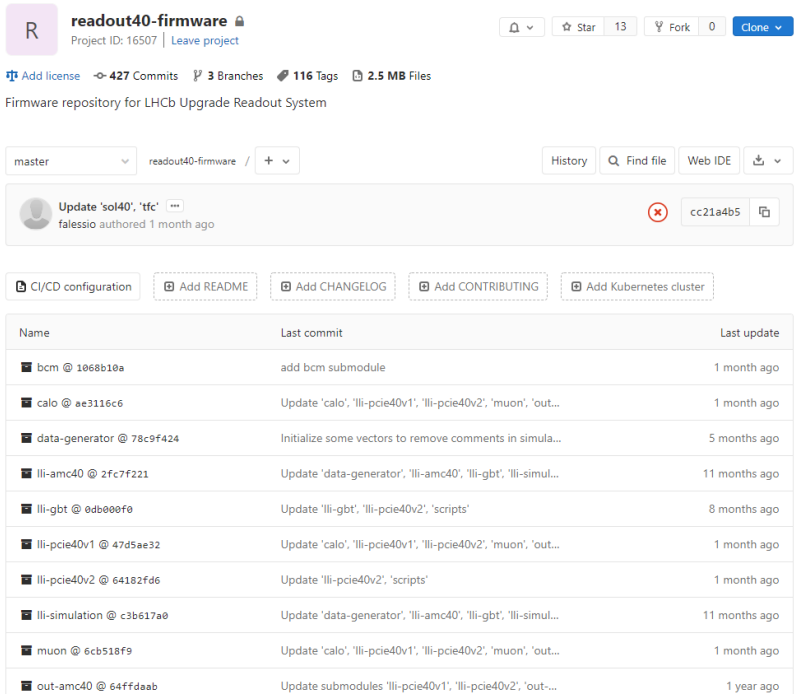
same hardware used for readout, supervision, controls

→ firmware defines the flavor of the card

Firmware framework

Common gitlab platform for code sharing, releases and CI/CD

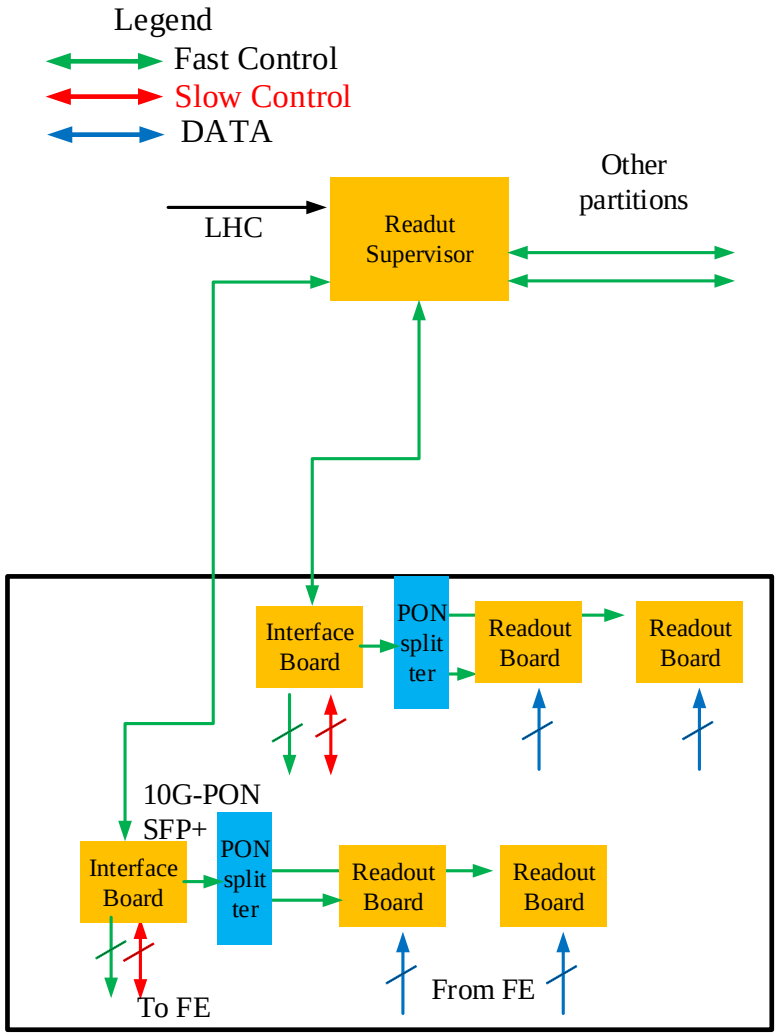
- Common platform to follow development
  - Make sure code doesn't break
- Continuous integration to spot mistakes incrementally
- Release management and notes
  - Versioning and tags
- Production of all firmware flavors automatically



- Inclusion of specific sub-detector firmware in a controlled and testable environment
  - Automatic checks on simulation
  - Automatic compilation reports

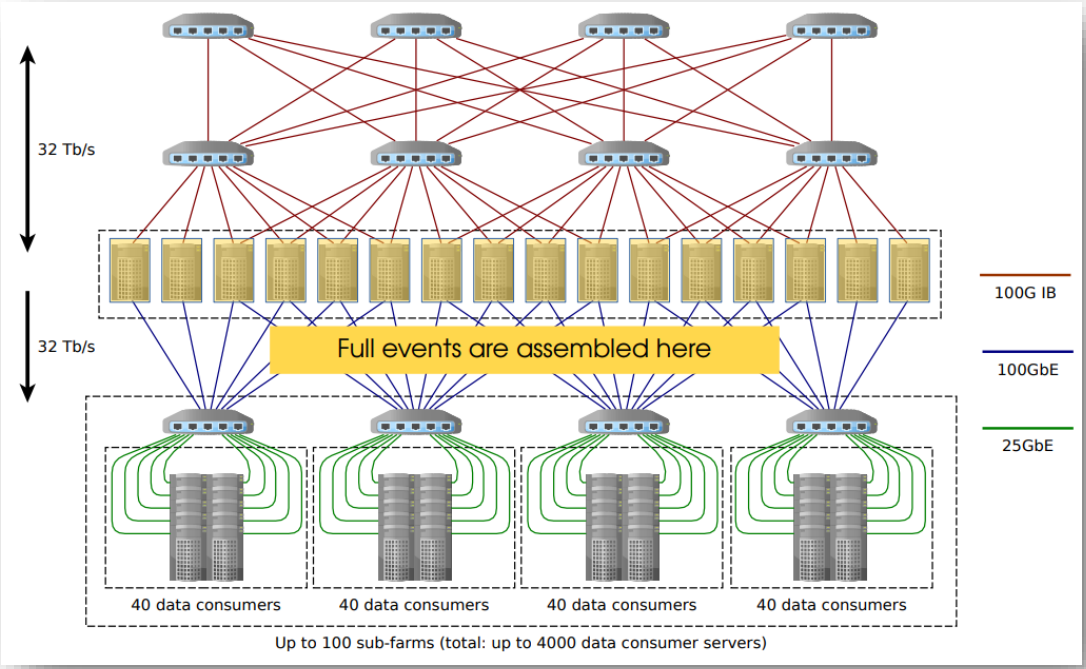
Timing and clock distribution

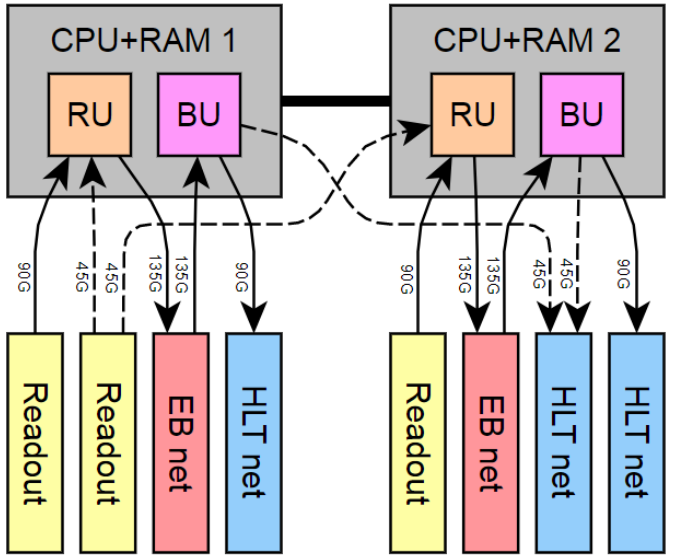
- LHC clock+orbit input to RS
- TFC from RS to IB
  - MiniPOD to classical SFP+ on SOL40
  - 8b10b protocol thru FPGA XCVR
  - fixed latency, fixed phase
- TFC from IB to RB is distributed through dedicated PON link
  - over PON-SFP+ and optical splitter
  - no fixed latency, no fixed phase
- TFC from IB to FE thru GBT-FE link
  - GBT protocol with fixed latency (v6)
  - Same 40 MHz edge as at the SODIN
- ECS to FE from IB thru GBT-FE link
  - Same optical network as TFC



Event building

- Dedicated event builder network
  - COTS server hosting the PCIe40 cards + NIC to EB
  - Fewer switch ports + more technology choice for network
  - Full events available already at the EB-nodes
  - Server with 4 slots may host hardware accelerator too and reduce output bandwidth
    - ✓ Selection of type of server ongoing





Server internal layout

Disk buffer is indeed a topic

Not easy to buffer 1 MHz of events out of fully HLT1 software trigger

- estimated event size 100 kB
  - One week at 50% machine LHC machine efficiency ~ 30 PB!
    → 100 kB * 1 MHz * 3.5 days = 30.2 PB

- Current studies ongoing in finding the right balance between CPU and storage
  - Given a fixed budget and emulating data taking in different conditions
  - Need to buy disks! (even just cutting to half the output to 500 kHz)

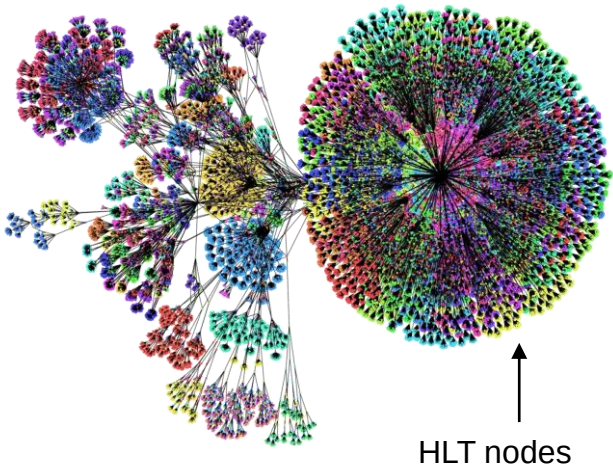
- Amount of disk is not everything, need to fold in also operation and feasibility
  - Failure rate?
  - Writing/reading speed?
  - Concurrent accesses between HLT1 (write) and HLT2 (read)?
  - What if HLT2 has to be stopped to update trigger and/ calibration and alignment procedures? For how long?

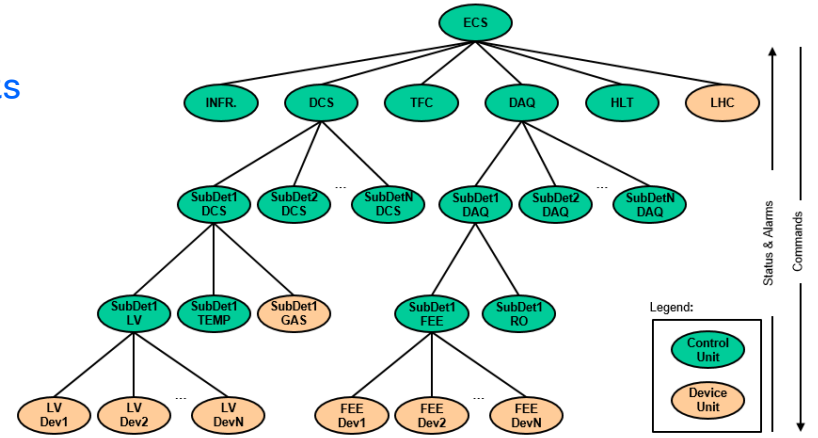
- Need to think real hard.

Control system

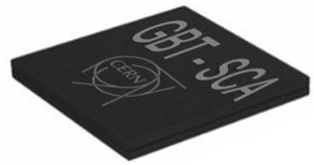
- LHCb has designed and implemented a coherent and homogeneous control system
- Controls the complete experiment
  - Run Control, DCS, LHC interface
- Operated by only 1 person
- Highly automated
- Based on WinCC OA 3.16 SCADA
  - Provides the UI, archiving, drivers, alarm-handling, etc
  - Allows for custom developments
  - CERN JCOP Framework: common sw components
- Control system modelled with an FSM tree
  - Dynamically configure the whole experiment according to needs/type of run
  - Allows for dynamic and independent partitioning
- FE controls integrated in tree as well

LHCb Control System Size
Picture: Courtesy of the CMS DCS Team

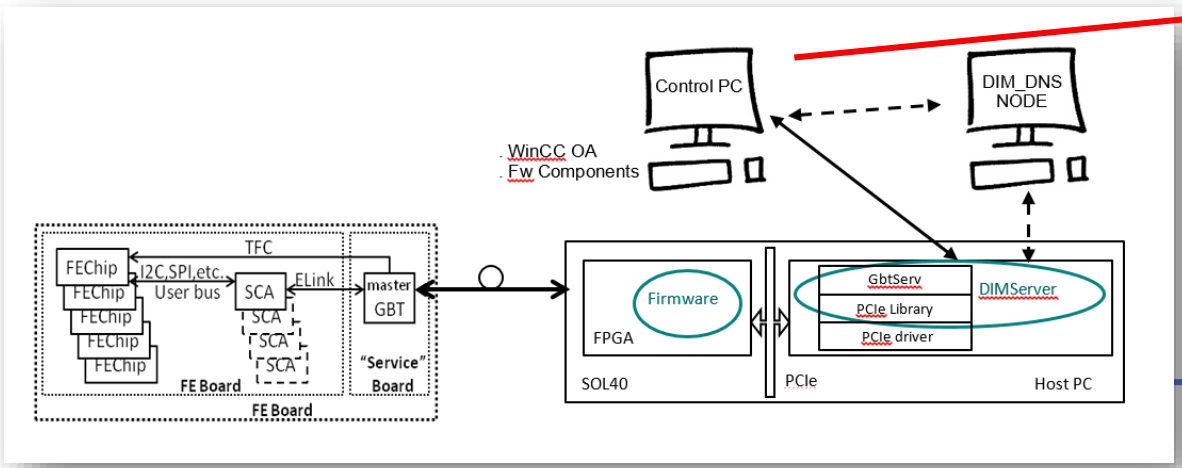


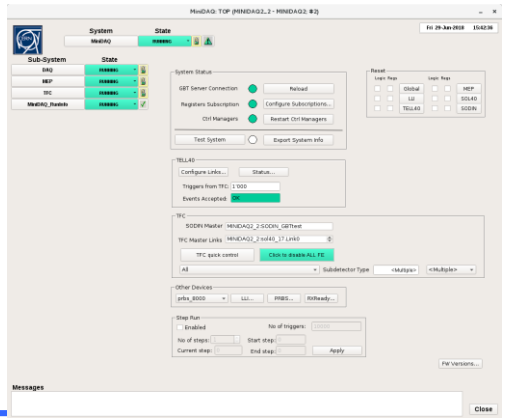


Control system: FE controls



- FE electronics connected and controlled through the PCIe40
  - Specific firmware implementation that provides the lower-level interfaces to the FE
  - Same hardware chipset (CERN GBT-SCA) implemented at the FE in all sub-detectors
    - ✓ "One implementation fits all"
  - Software server runs on same server as the PCIe40
    - ✓ Drives the firmware providing commands to the FE and receiving responses
    - ✓ Implements all different protocol communications
    - ✓ Complex action as single command (remote FPGA programming at FE)
    - ✓ Implements connection to the control system (using DIM)



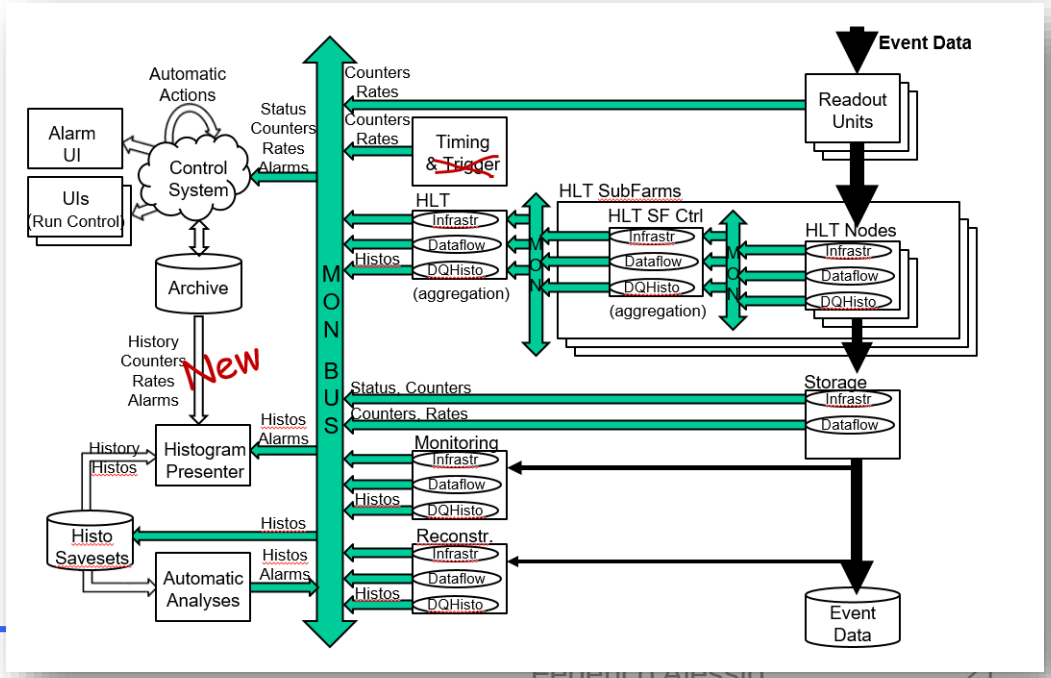


Monitoring system

- Monitoring system is crucial component to check detector & readout are well functioning
  - Has to tap on a continuous stream of data to monitor its quality
  - Has to have interfaces to all other running conditions + configurations
  - Has to aggregate data from multiple trigger nodes to obtain measured quantities
  - Has to archive its result, produce alarms for shifters and run control, produce reports

→ Common infrastructure
  - Interface to Control
  - Interface to Shifters

→ Specific "jobs" implementation
  - Sub-detectors
  - Run Control
  - Interfaces to DBs



Data center

- Modern data center infrastructure to host:
  - ~500 server nodes for readout cards + Event Builder + FARM nodes
  - Total of 134 racks that can host up 2000 servers
  - 2.1 MW power each module
  - 20 PB temporary storage
  - Save up on cooling, indirect fee-air cooling
  - Entire readout system in surface

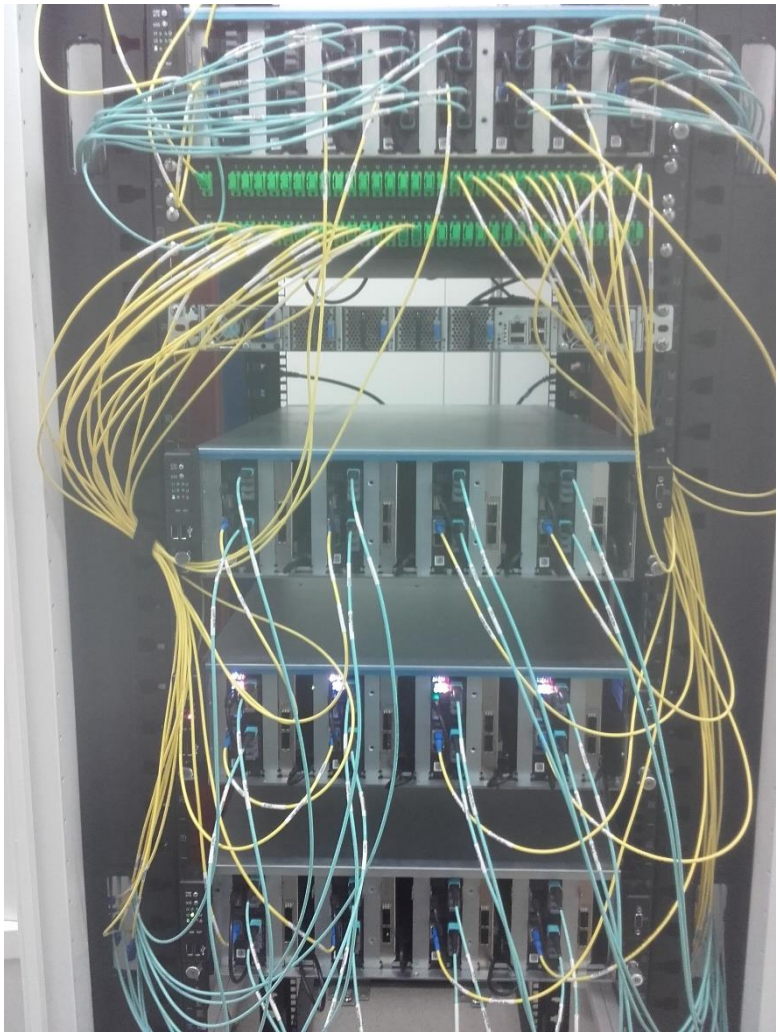
19000 fibers going underground…

Vertical slice as readout mock-up

Testing and integrating a slice of readout system

- Same as final configuration
- With fw and sw production releases
- FE event generators/emulators in FPGA
- Final timing distribution
- Event Building and final network
- Testbed for HLT and accelerators
- In the data center
- Integration with monitoring system
- With final production control system

Paramount to perform early commissioning (especially if installation schedule gets tight and start eating into contingency time…)



Conclusion & checklist

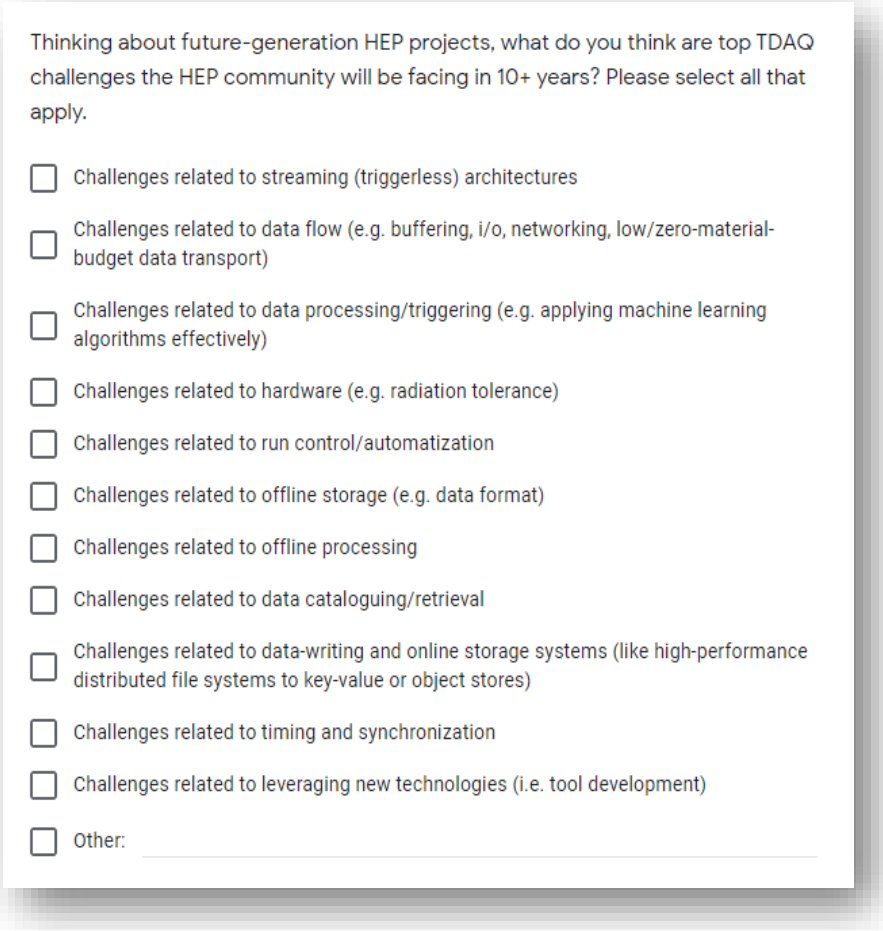
LHCb looking forward to exciting times with its newly upgraded readout system

→ At the forefront of future challenges in HEP experiments

→ Modern technologies and implementations

→ Trying to tick all the boxes…

Stay tuned for seeing it working! ☺



Thinking about future-generation HEP projects, what do you think are top TDAQ challenges the HEP community will be facing in 10+ years? Please select all that apply.

- [ ] Challenges related to streaming (triggerless) architectures
- [ ] Challenges related to data flow (e.g. buffering, i/o, networking, low/zero-material-budget data transport)
- [ ] Challenges related to data processing/triggering (e.g. applying machine learning algorithms effectively)
- [ ] Challenges related to hardware (e.g. radiation tolerance)
- [ ] Challenges related to run control/automatization
- [ ] Challenges related to offline storage (e.g. data format)
- [ ] Challenges related to offline processing
- [ ] Challenges related to data cataloguing/retrieval
- [ ] Challenges related to data-writing and online storage systems (like high-performance distributed file systems to key-value or object stores)
- [ ] Challenges related to timing and synchronization
- [ ] Challenges related to leveraging new technologies (i.e. tool development)
- [ ] Other:

https://docs.google.com/forms/d/e/1FAIpQLSebb0XcGhGA3wduXVdiqXyMIBBs0FDbpyfy0ORbGDzxG5R8ZA/viewform