Trigger & DAQ at the LHC

Filtering data from 50 TB/s to 1 GB/s

Flavio Pisani

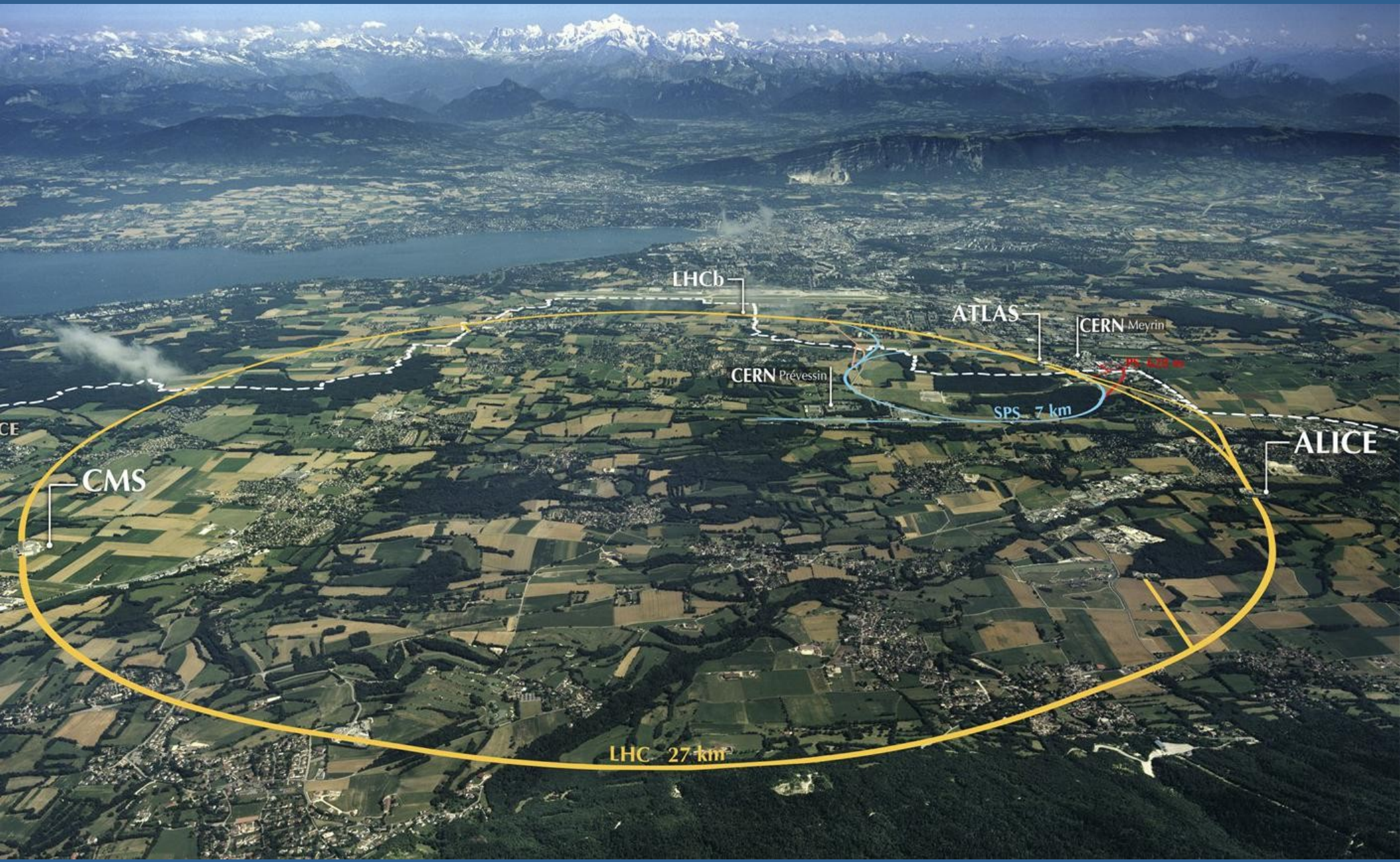
CERN EP/LBC

7 July 2022

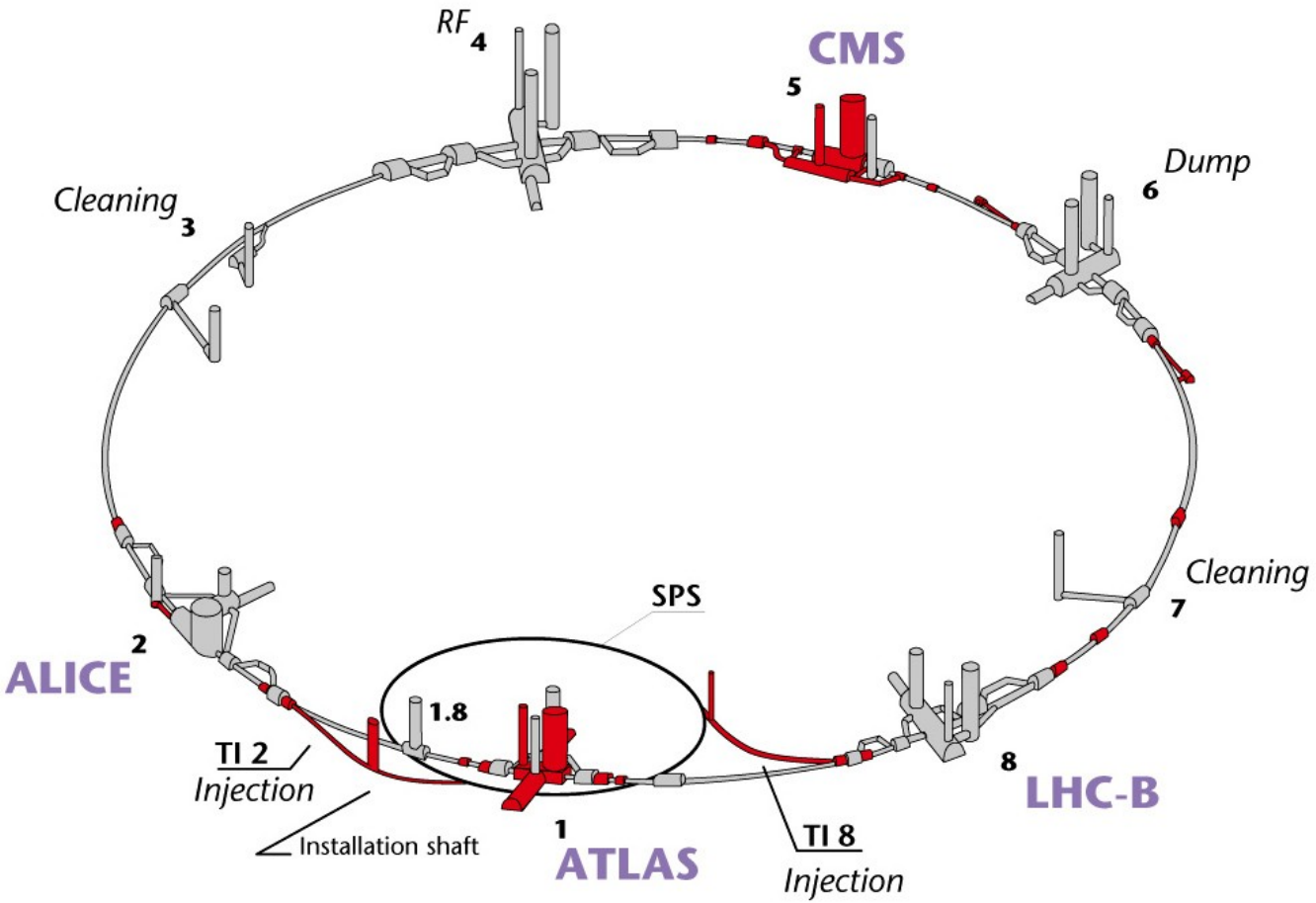
Intro & Acknowledgments

- The circle of slides:
  - Much of the material for this talk was originally prepared by N. Neufeld and T. Colombo for the 2019 edition
  - They in turn acknowledges that it much of his presentation was naturally based on the work of other colleagues
  - A missing indication of origin of a figure does not imply that it is originally mine
- Trigger and DAQ are vast subjects covering a lot of physics and engineering
- This talk is meant to give a high-level overview
- Some inevitable gross simplifications are to be expected
- Many topics are left out altogether

The Large Hadron Collider

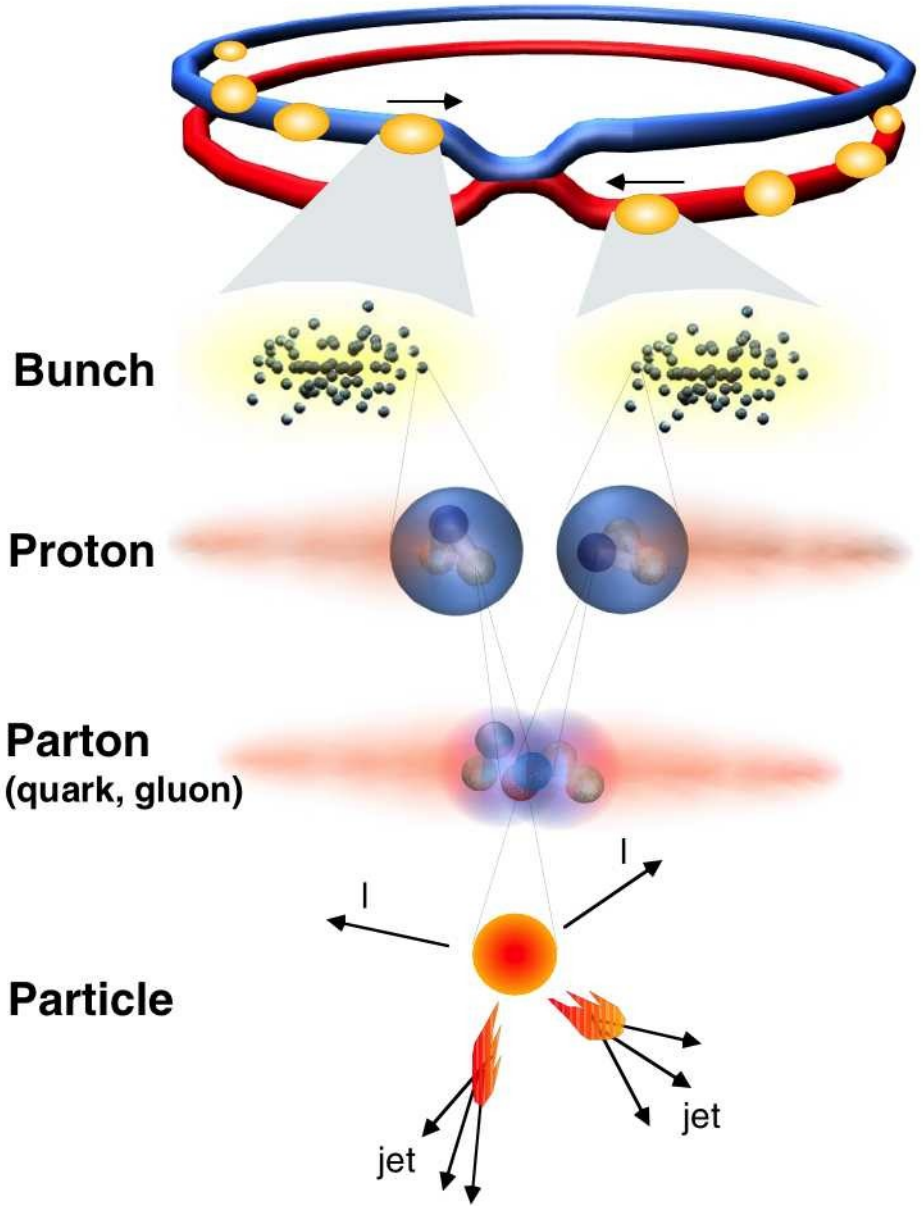


The Large Hadron Collider



- 27 km
- Vacuum at 10^{-13} atm
- >= 9600 magnets
- Magnetic field ~8 T
- Magnets at -271.3° C
- Energy in the beam corresponds to a car at the speed of sound
- 4 large experiments
- Cost: 5 billion CHF

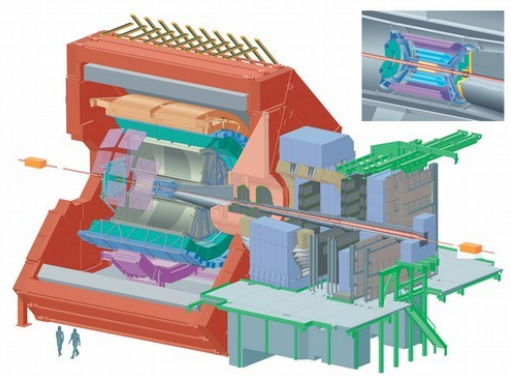
Colliding beams

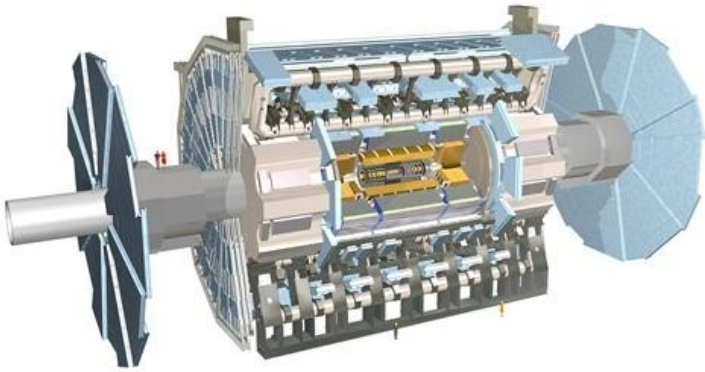


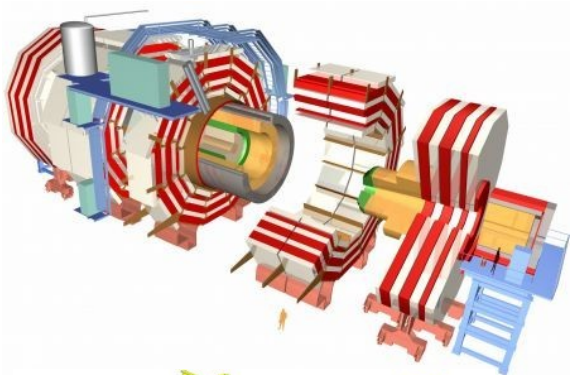
Proton-Proton	2835 bunch/beam
Protons/bunch	10^{11}
Beam energy	7 TeV (7×10^{12} eV)
Luminosity	10^{34} cm^{-2} s^{-1}
Crossing rate	40 MHz
Collisions ≈	10^{7} - 10^{9} Hz

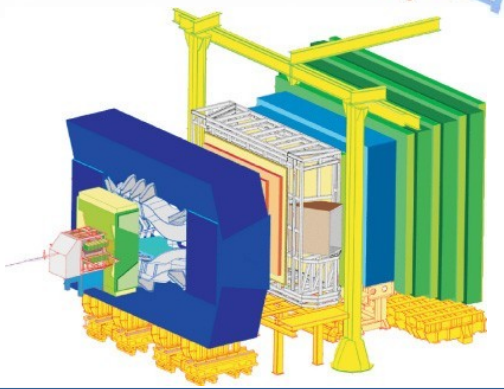
The LHC experiments

- ALICE – “A Large Ion Collider Experiment”
  - L x W x H: 26 x 16 x 16 m - Weight: 10000 t
  - 35 countries, 118 Institutes
  - Material costs: 110 MCHF
- ATLAS –“A Toroidal LHC ApparatuS”
  - L x W x H: 46 x 25 x 25 m - Weight: 7000 t
  - 38 countries, 174 institutes
  - Material costs: 540 MCHF
- CMS –“Compact Muon Solenoid”
  - L x W x H: 22 x 15 x 15 m - Weight: 12500 t
  - 40 countries, 172 institutes
  - Material costs: 500 MCHF
- LHCb – “LHC beauty” (b-quark is called “beauty” quark)
  - L x W x H: 21 x 13 x 10 m - Weight: 5600 t
  - 15 countries, 52 institutes
  - Material costs: 75 MCHF

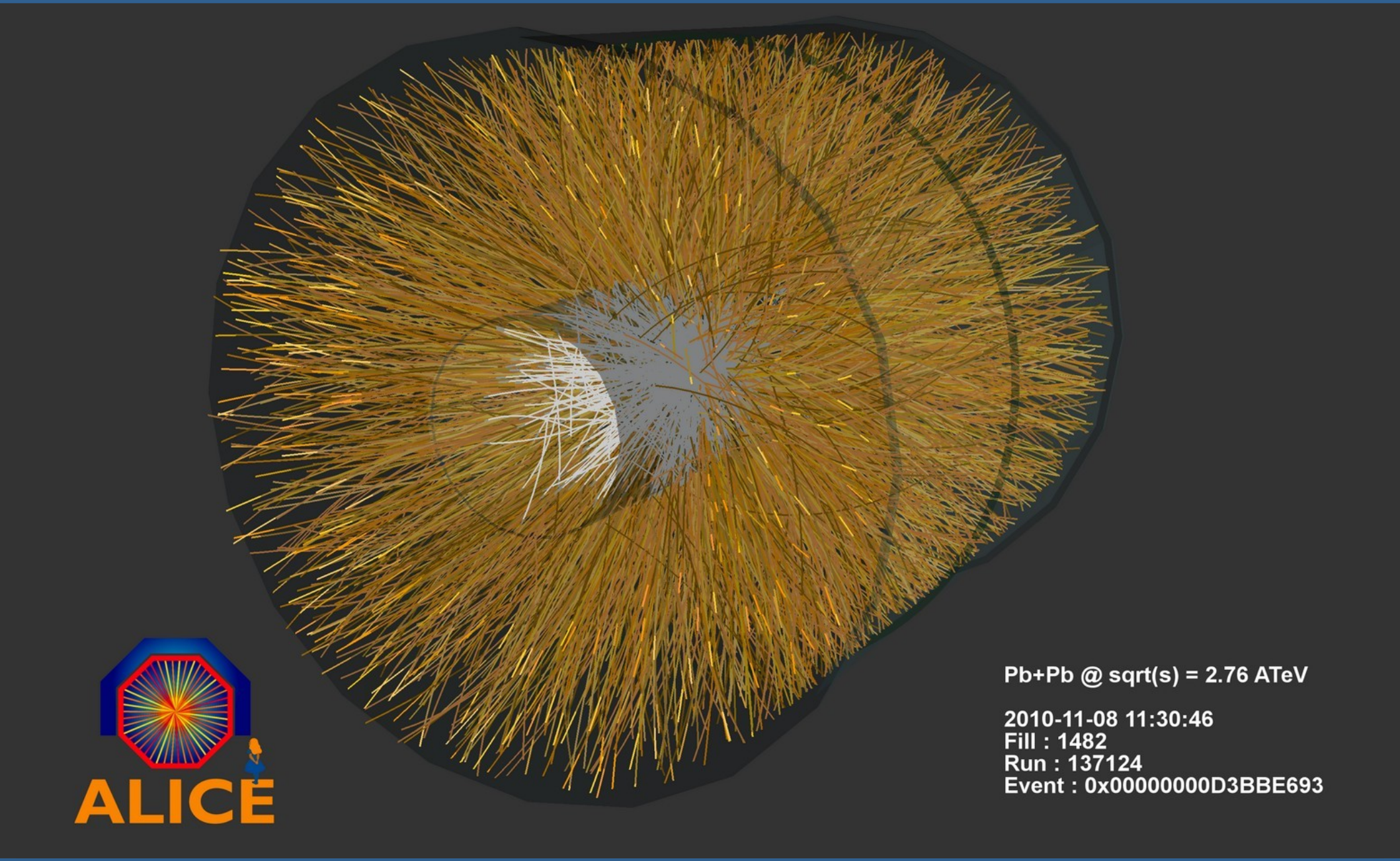




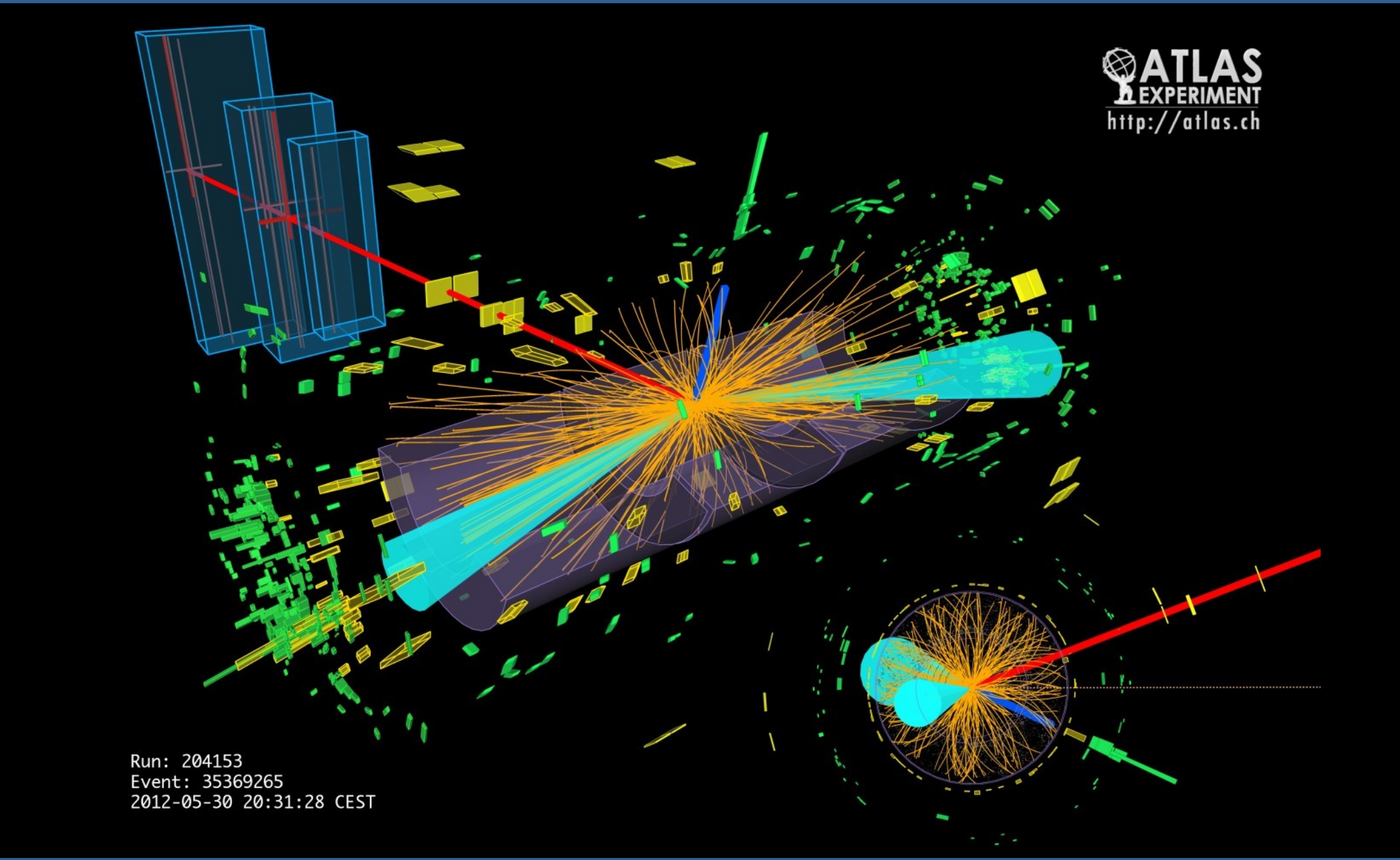




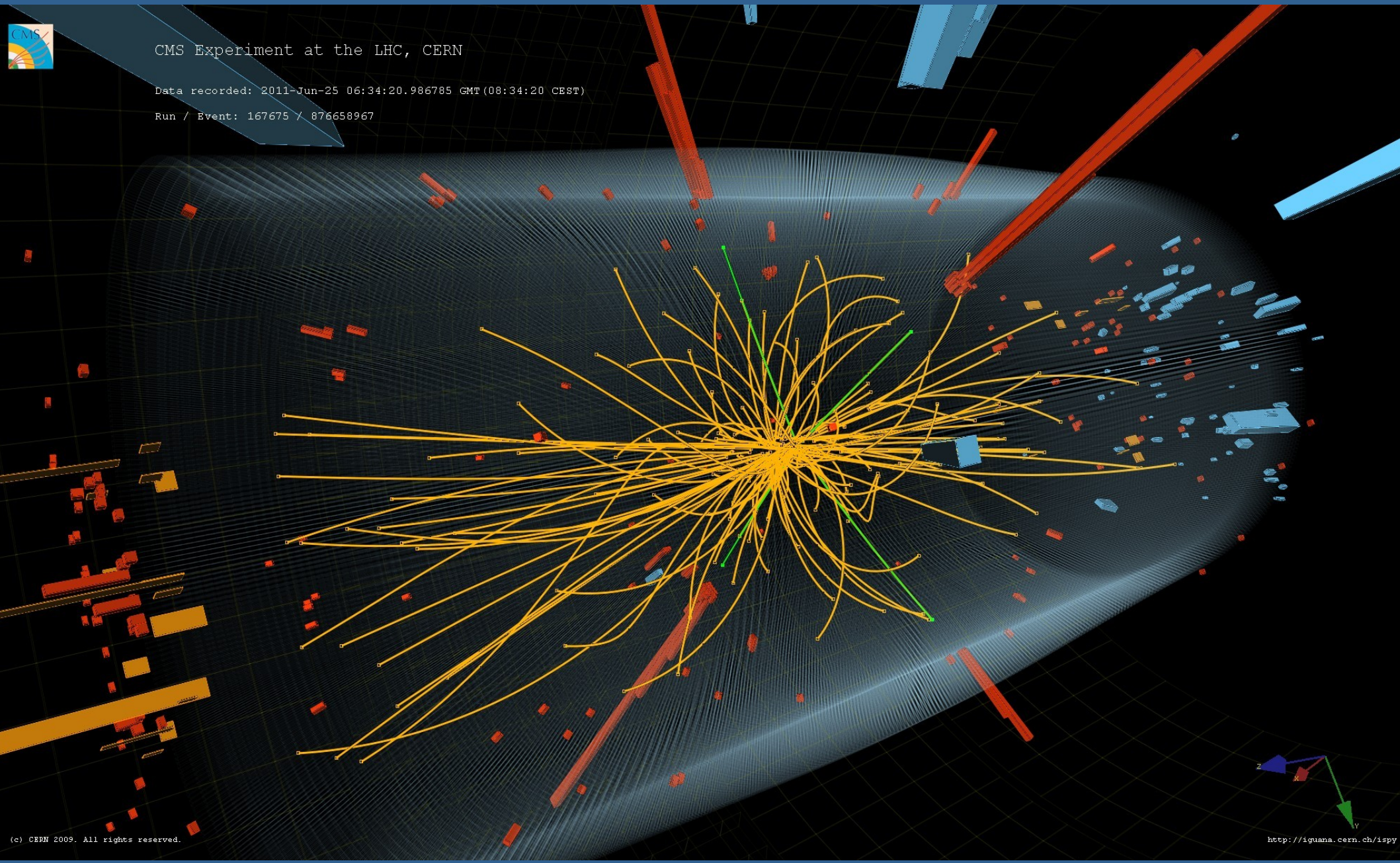
Lead-lead collision @ ALICE



Higgs boson @ ATLAS

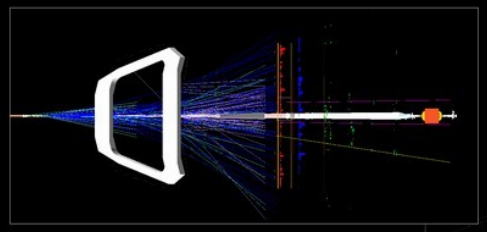


Another Higgs boson @ CMS

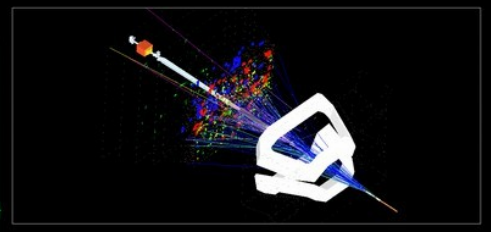


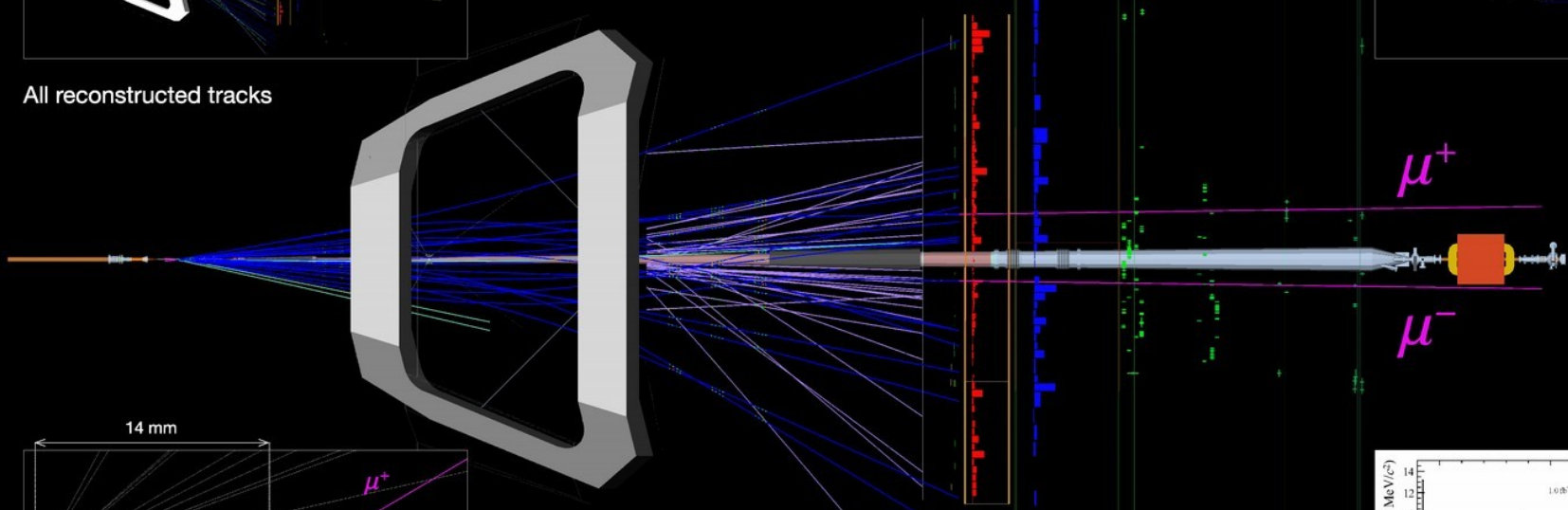
Rare B meson decay @ LHCb

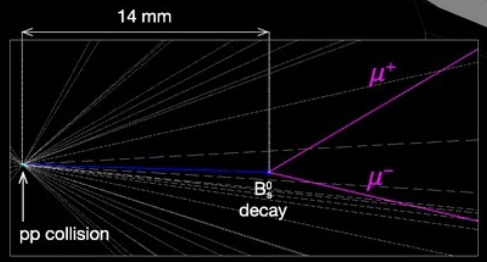
$$B_s^0 \longrightarrow \mu^+\mu^-$$



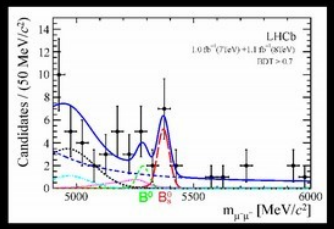
All reconstructed tracks





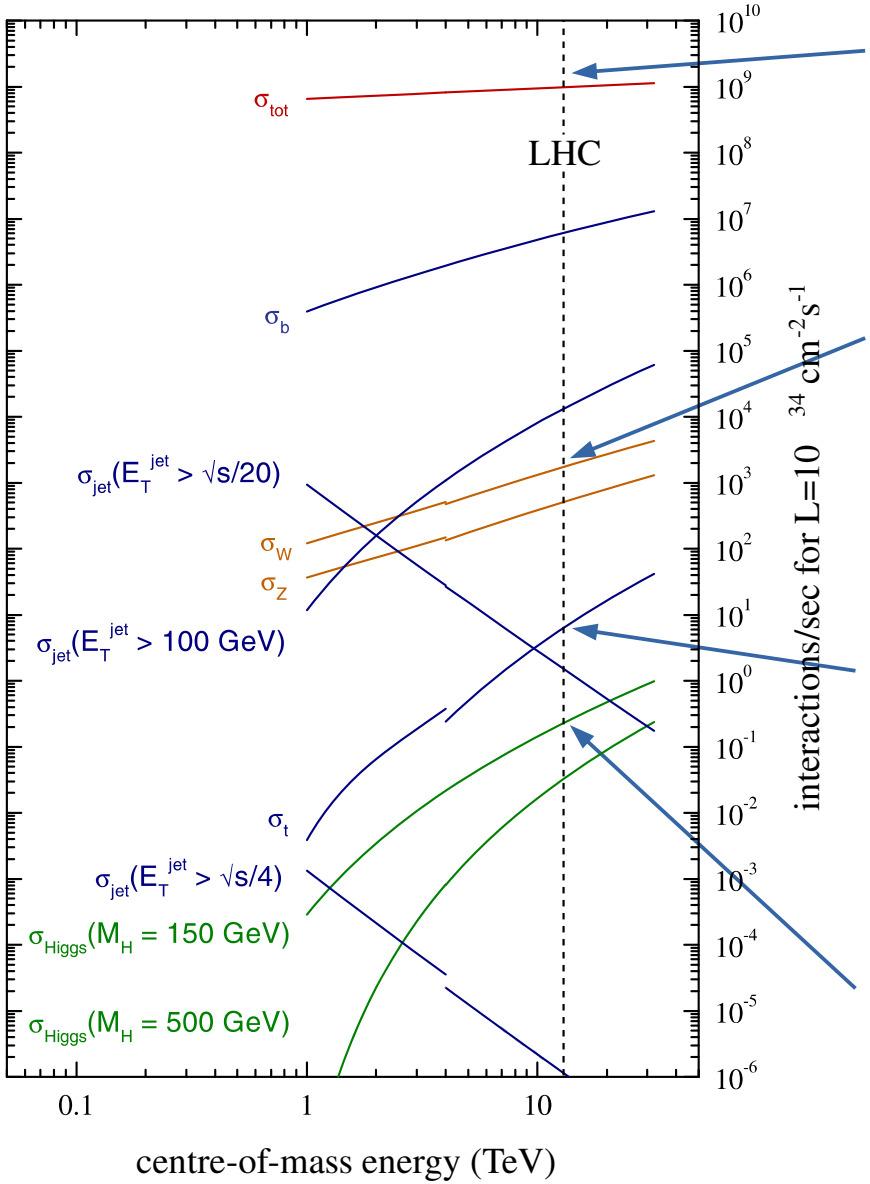


Only well reconstructed tracks with $p_T > 500$ MeV



31.05.2012 15:03

High-energy physics, or the art of finding needles in haystacks



The haystack (hard collisions): 1 GHz

W & Z bosons: 1 kHz

Discovered: 1983 @ SPPS

t quark: 10 Hz

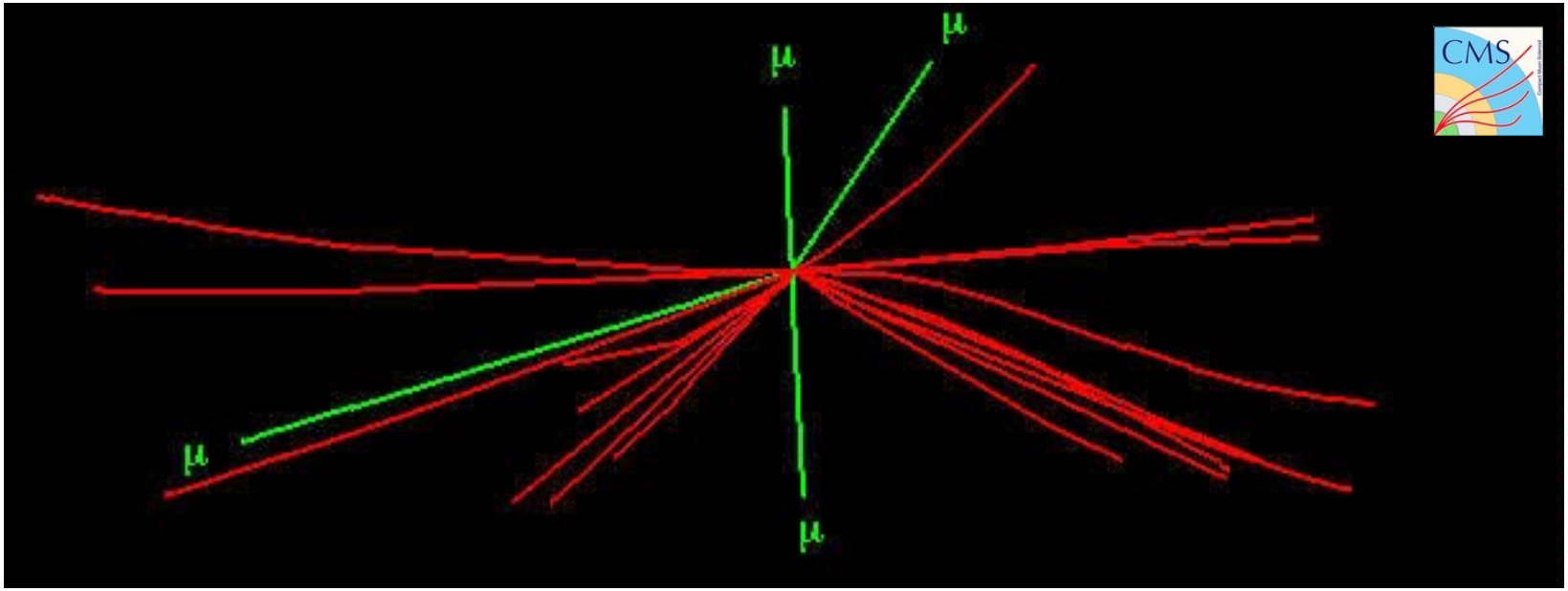
Discovered: 1995 @ Tevatron

Higgs boson 0.1 Hz

Discovered: 2012 @ LHC

- Contemporary high-energy physics focuses or ***rare*** processes
- The vast majority of the collisions is “boring”
- Interesting physics is ≥ 9 orders of magnitude rarer: one in a billion or more!
- Need to efficiently identify these interesting processes from the **before** reading out & storing the data

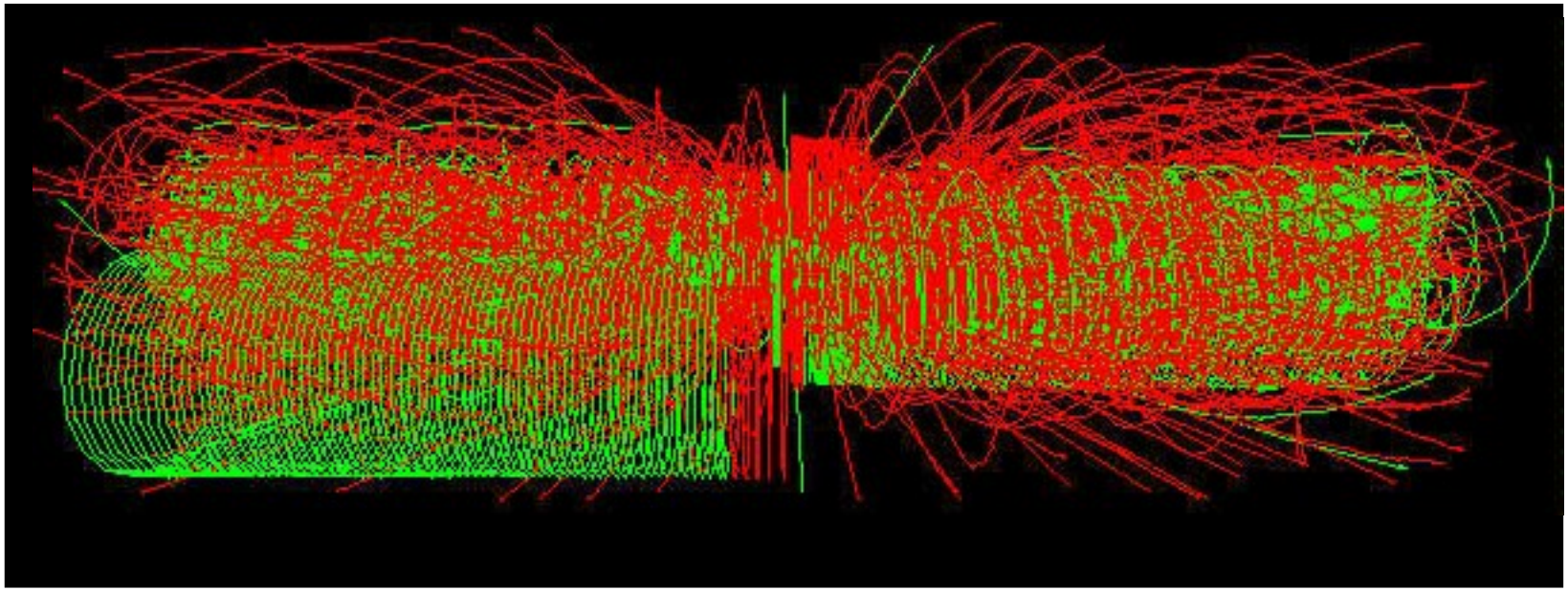
Finding the needle



This is what we're looking for:
a Higgs boson decaying in four easily identifiable muons

The LHC produces a few of these **per day**

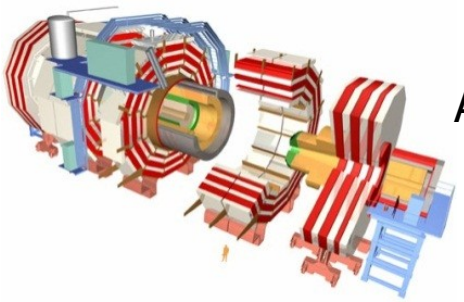
Finding the needle



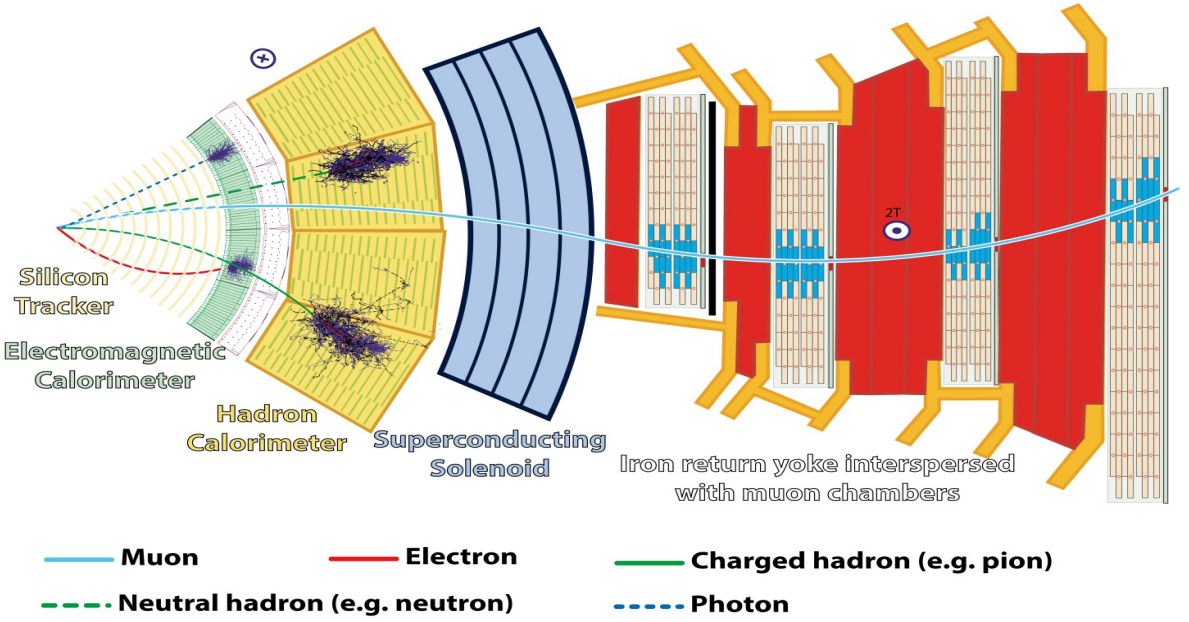
This is where it hides:
tens of other hard collisions producing 1000s of particles

The LHC makes 40 million of these per second!

LHC experiments: ≥ 10 million sensors



A slice of CMS



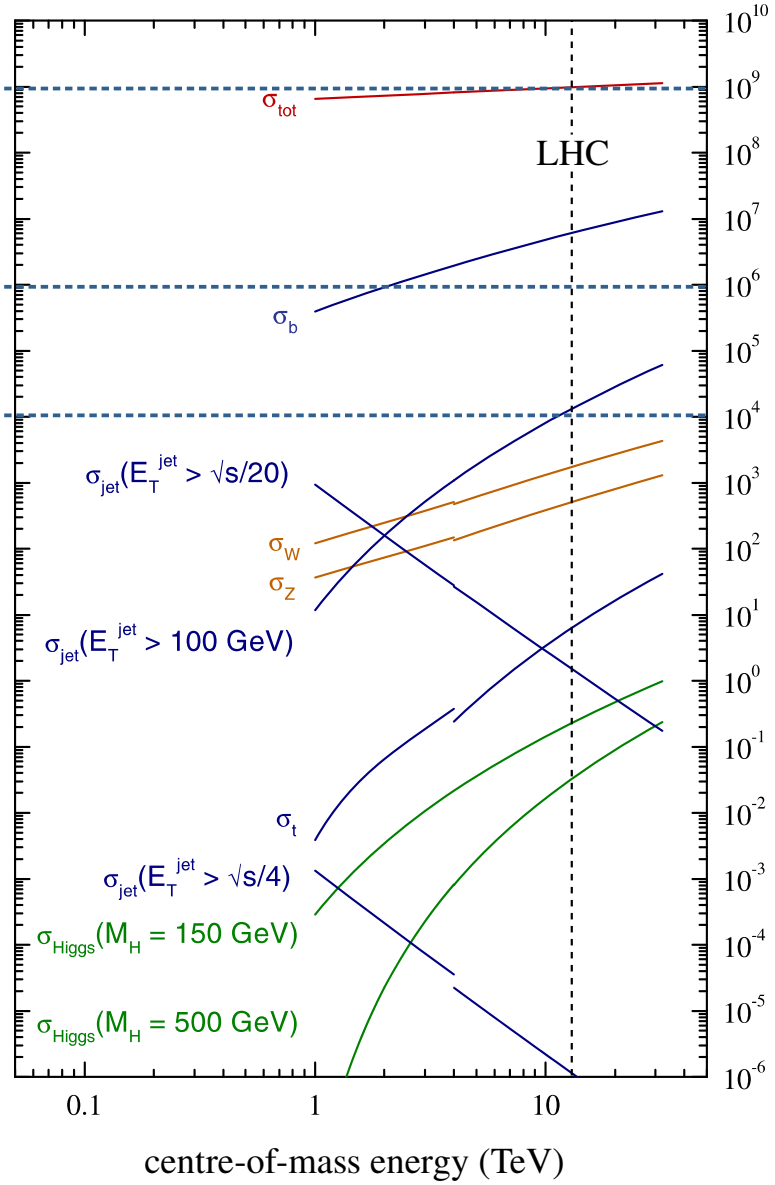
- Accurately measuring momentum (p) and energy (E) of the thousands of particles produced by the collisions requires **fine-grained** particle detectors
- High granularity → big data: ~1 MB per bunch crossing
- Theoretical output: 40 MHz x 1 MB = **40 TB/s**

40 TB/s, visualised

- 40 TB/s → **Forty 10 TB hard drives filled in ten seconds**
- **Saving all this data for later analysis is impossible**
  - Even if we had infinite money, we don't have infinite space (~5.4 m^3/h)
- Real-time data selection is necessary
- In HEP-speak: **triggering**
  - Acquire and save only interesting bunch crossings ("events")

Data selection

- Particle beams cross every 25 ns (40 MHz)
  - ~25 “hard” collisions per bunch crossing
  - Up to 10^9 collisions per second
- Typical trigger architecture
  - Two or more consecutive steps
  - Step 1 (“Level-1 trigger”):
    - Custom hardware
    - 40 MHz → 100-1000 kHz
    - Reduction factor: 40-400
  - Step 2 (“High-level trigger”)
    - Software
    - 100-1000 kHz → 1-10 kHz
    - Reduction factor: 100



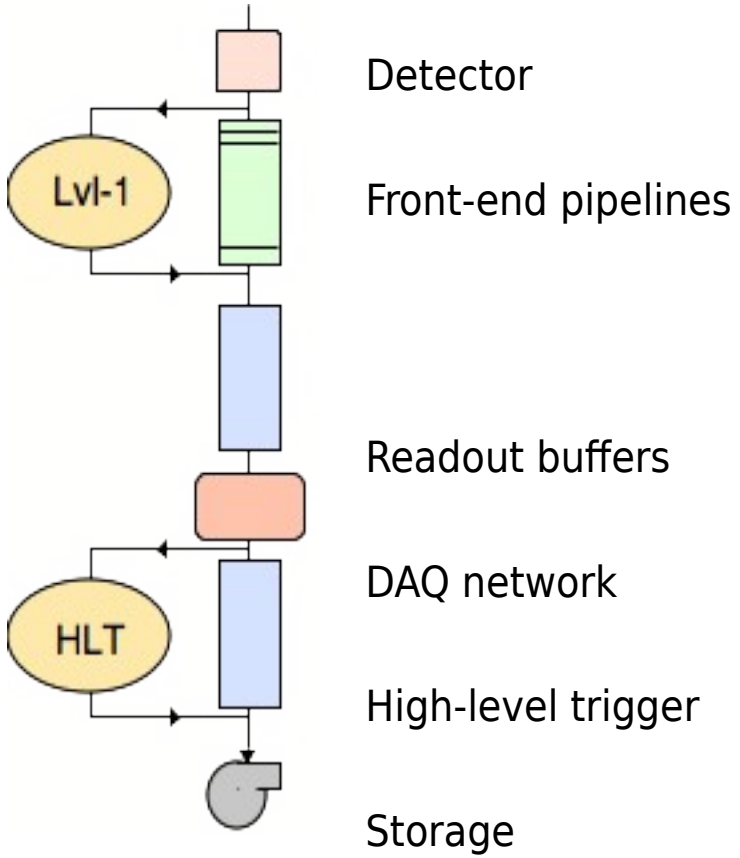
Step 1:
Level-1 Trigger

Level-1 trigger

- No (affordable) data acquisition (DAQ) system can read out tens of millions of sensors at 40 MHz
  - ATLAS & CMS would output ≥ 40 TB/s
  - This would require a lot of fast, low-power, low-cost links, but:
    - Radiation-hard optics are not there yet
    - Huge amounts of copper cables would “steal” valuable space from the detector itself
  - But…for “smaller” detectors and using clever data compression on the detector itself it is possible today!
- Remember: most of these millions of events per second are totally uninteresting: one Higgs event every 10 seconds
- The first data-selection step (Level-1 trigger) must somehow select the more interesting events without reading out all the sensors

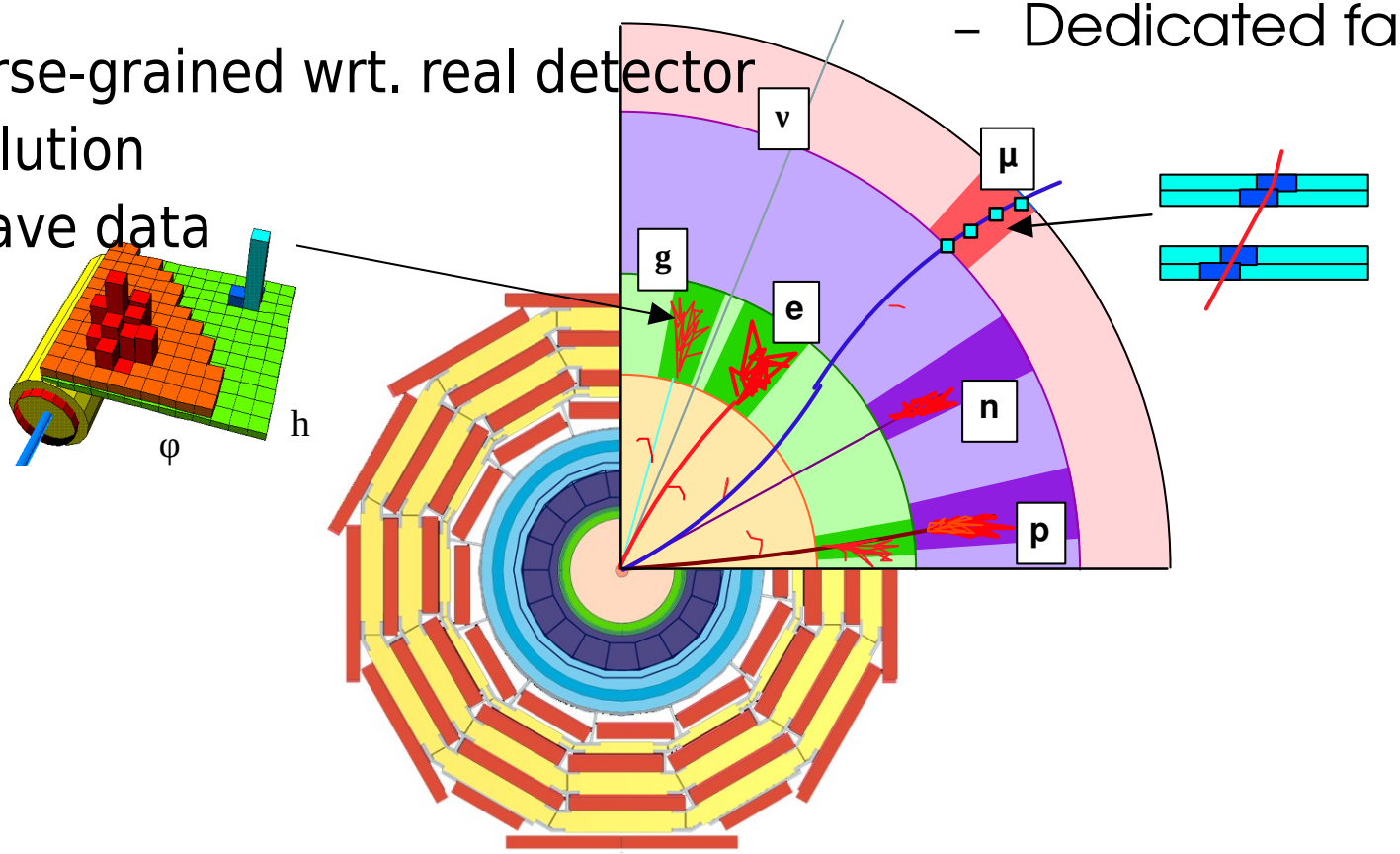
Selection based on a small subset of sensors

- Use data from fast sensors for Level-1 decision
  - Try to identify high-momentum and high-energy particles
- The remaining sensors have to buffer the data in their pipelines until a decision is made
  - The Level-1 trigger has hard latency constraints
  - Typical latency: a few μs
  - This requires custom electronics
- When the Level-1 accepts an event, it *triggers* the readout of the detector (hence the name)

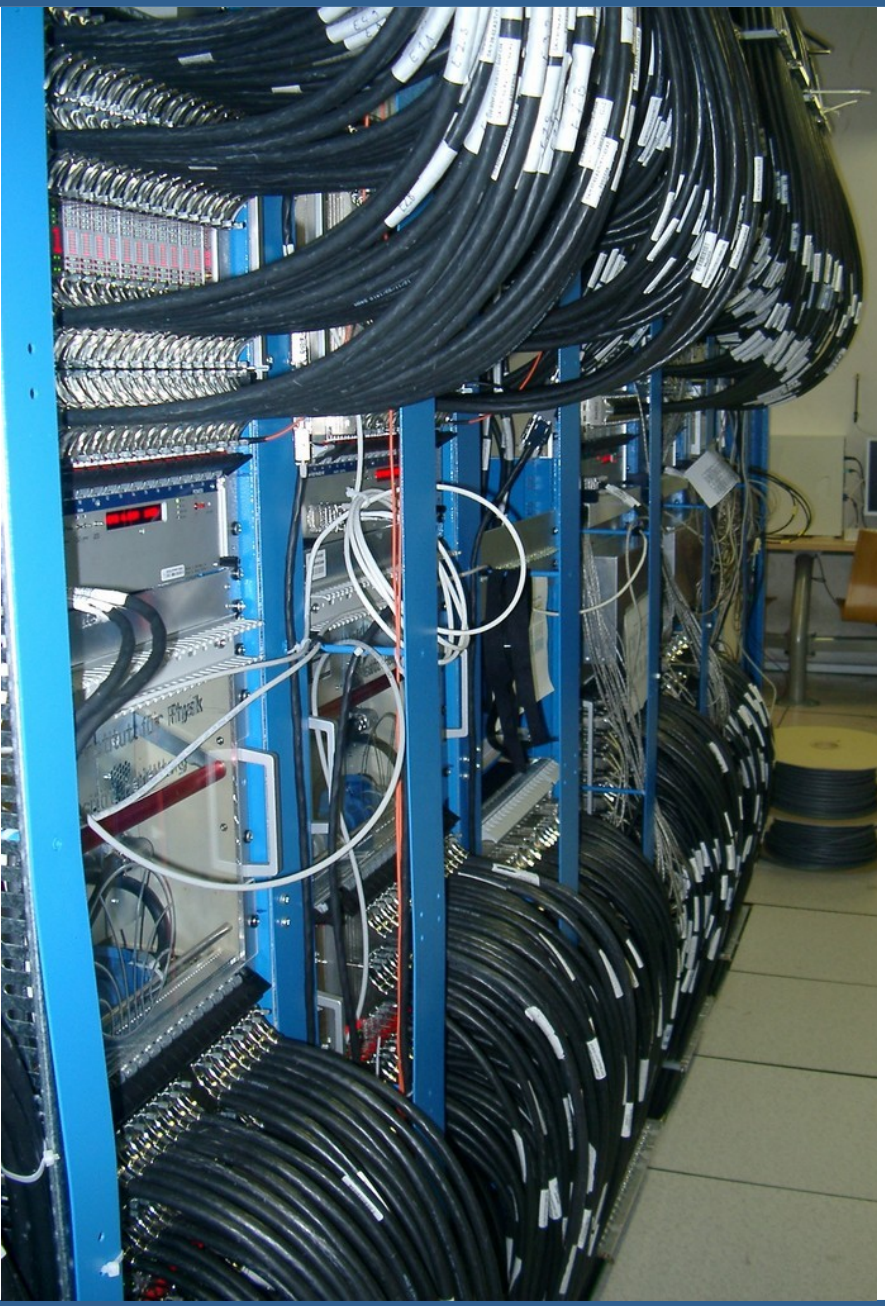


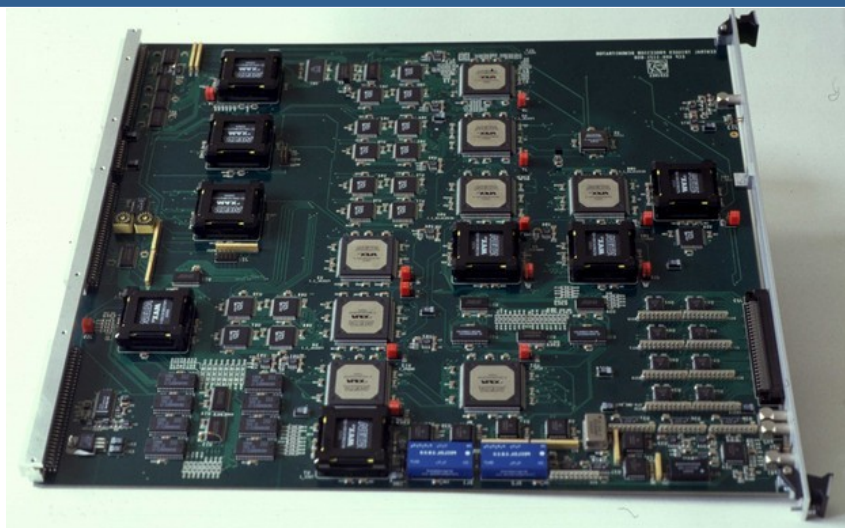
Fast local algorithms

- Calorimeters
  - Cluster finding
  - Energy deposition evaluation
  - Coarse-grained wrt. real detector resolution to save data
- Muon systems
  - Track finding
  - Momentum evaluation
  - Dedicated fast sensors



Custom electronics





- Sophisticated hardware (FPGAs, ASICs)
- Hundreds of custom-built boards process small pieces of the collision at enormous speeds
- They give a crude but effective decision, based on simple criteria

Summary

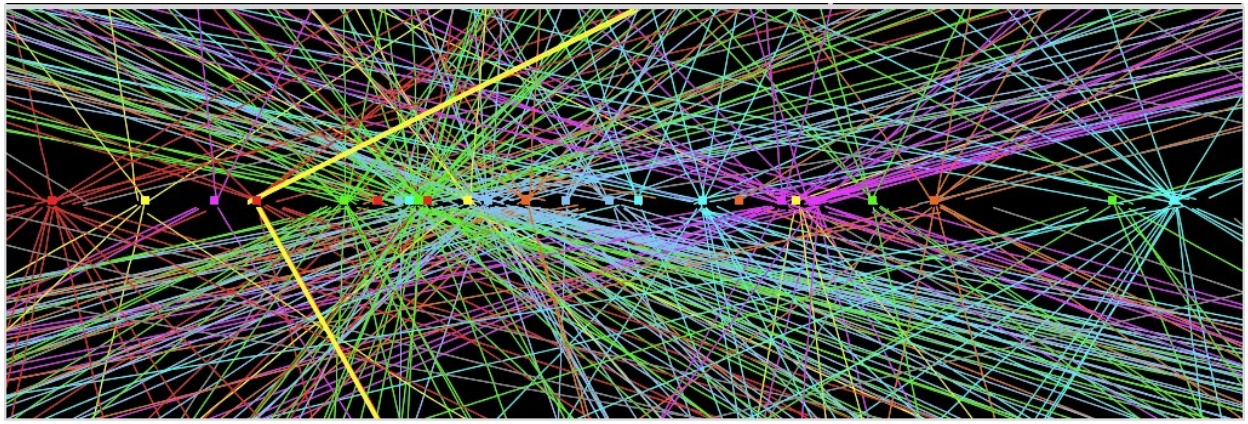
- The Level 1 Triggers are implemented in custom electronics:
  - difficult / expensive to upgrade or change
  - maintenance by experts only
- Decision time: ~ a small number of microseconds
  - The Level 1 Triggers are **hard real-time** systems
- They use “simple” hardware-friendly algorithms
  - Particle identification (high-energy particles & jets)
  - Local pattern recognition based on coarse-grained data
- Working with partial information and with drastic simplifications has a price:

  → Potentially interesting and valuable events are lost
- Future directions:
  - Eliminate / reduce hardware Level-1 (ALICE, LHCb)
  - Substantially upgrade Level-1 (ATLAS, CMS)

Step 2:
”High-level” trigger
(HLT)

HLT: Full “reconstruction” of the collisions

- Pack the knowledge of thousands of physicists and decades of research into a huge sophisticated algorithm
- **Reconstruct all charged particle trajectories**
  - Find segments, connect them, re-fit to physical trajectory
- Associate the particles with the correct p-p hard collision
  - Multiple interactions in each crossing

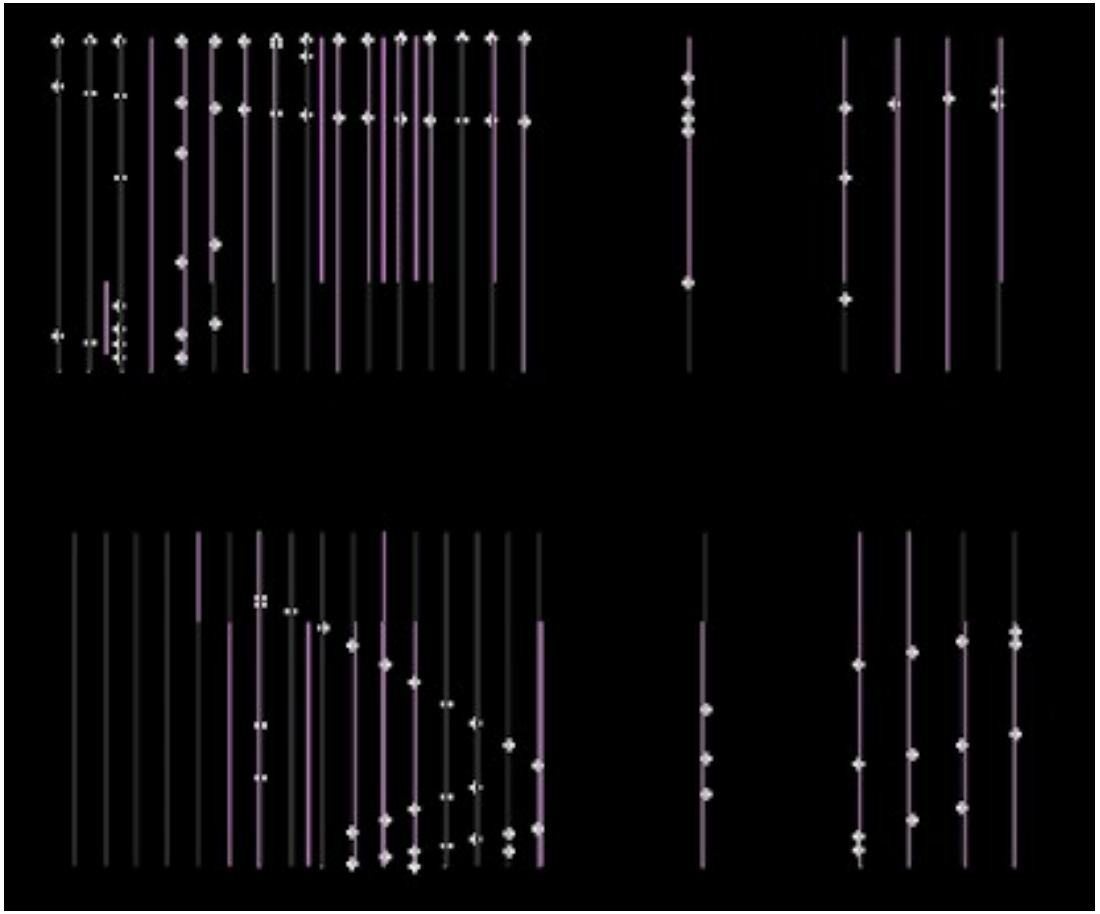


- Measure all of the energy depositions in the calorimeters with fine granularity
- Associate tracks and energy depositions

Data decoding

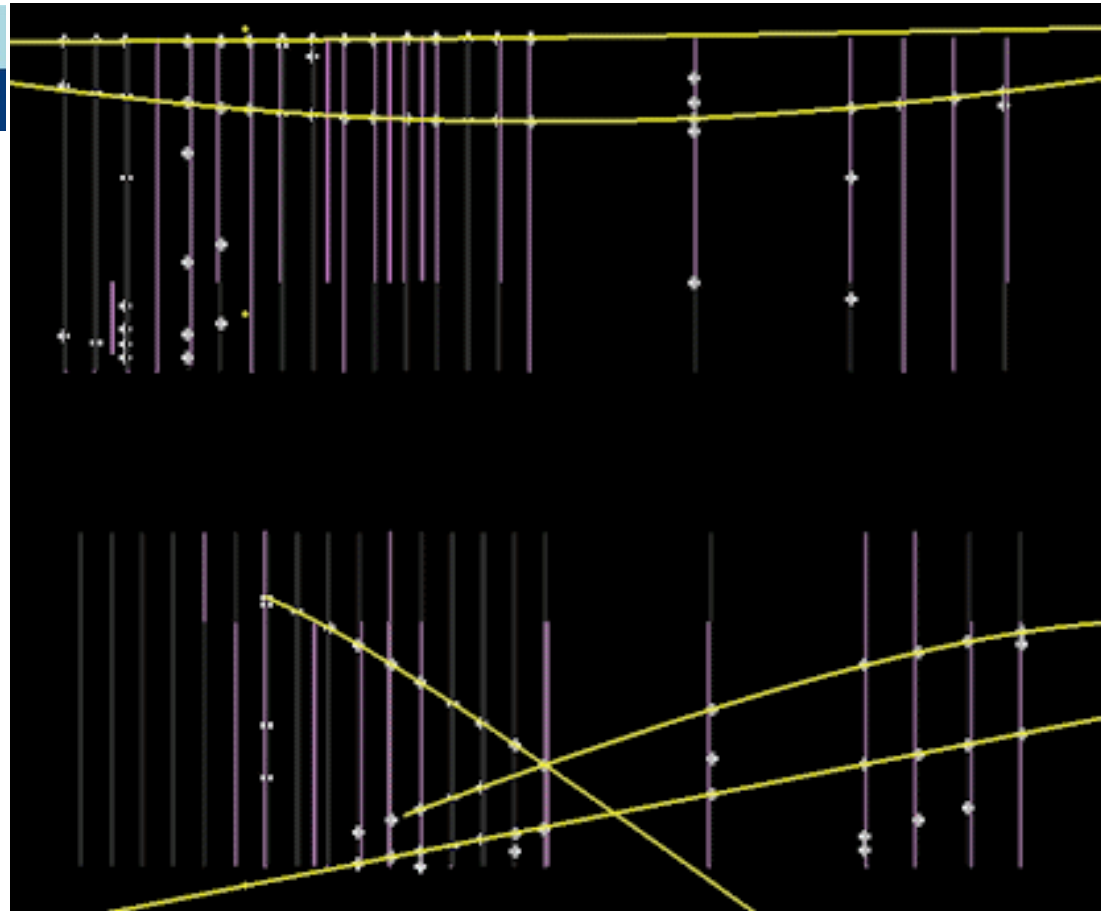
- What we get from the detector:
  - Sensor 1244 has measured a signal of 120 at time 1333096259.344245
- What we need to reconstruct what happened:
  - Signal at position x = 1.2 cm, y = 4.5 cm, z = 3.2 cm, deposited energy of 100 keV
- Requires precise information about location of the detector element (**alignment**) and of its signal sensitivity (**calibration**)

Track finding: pattern recognition



- Start with a collection of “hits” (footprints) on the various layers of the detector

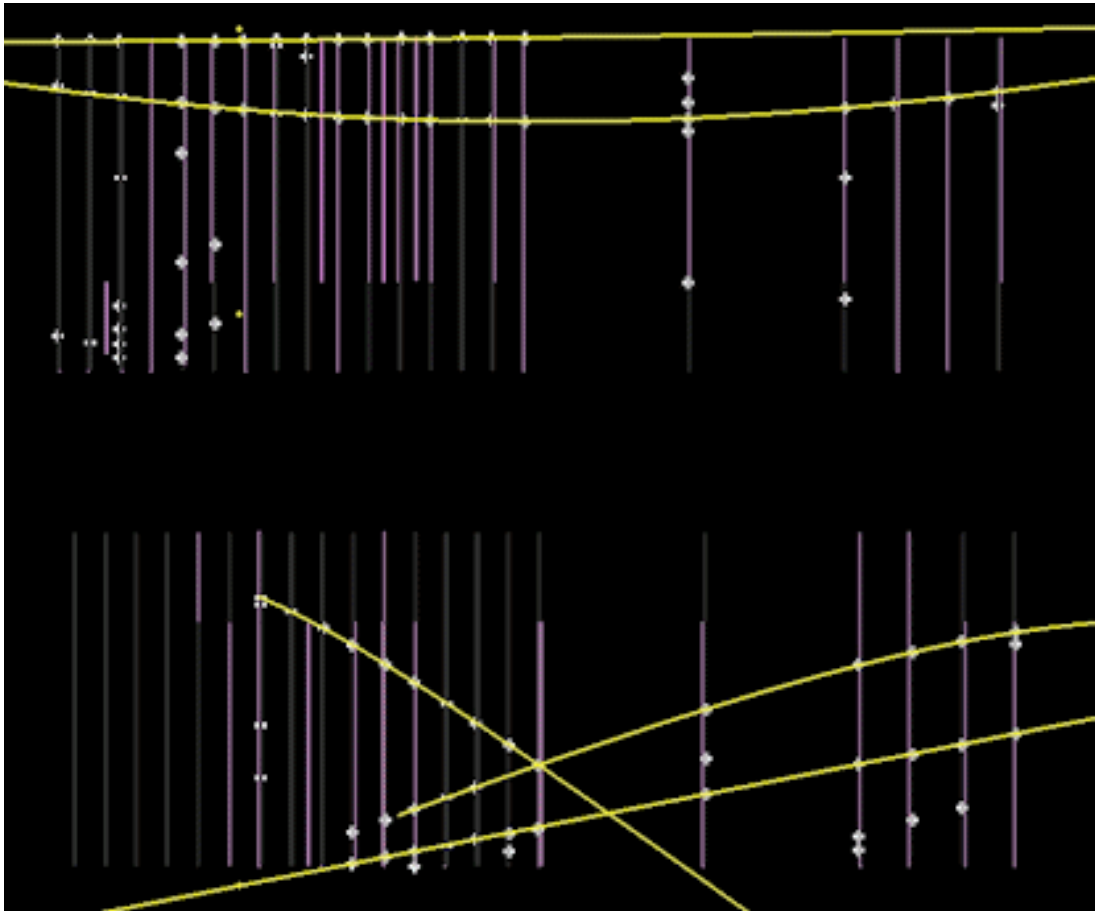
Track finding: pattern recognition



- Deviation on a curved trajectory in a normal plane around the magnetic field
- Radius of curvature allows to determine the particle's momentum

- Reconstruct the trajectory of the particle that left the footprints

Track finding: pattern recognition

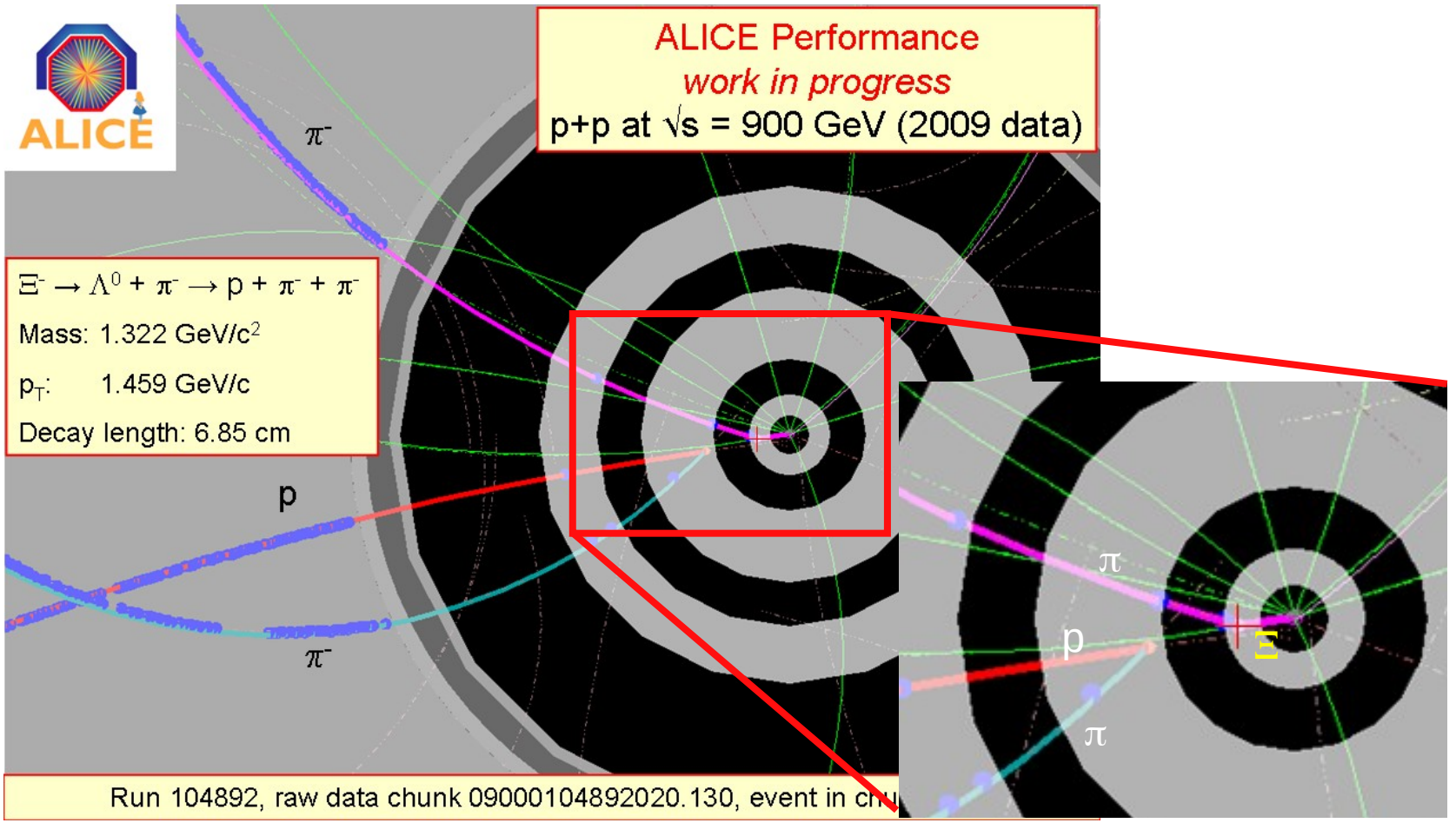


- Reconstruct the trajectory of the particle that left the footprints

Can become very complicated

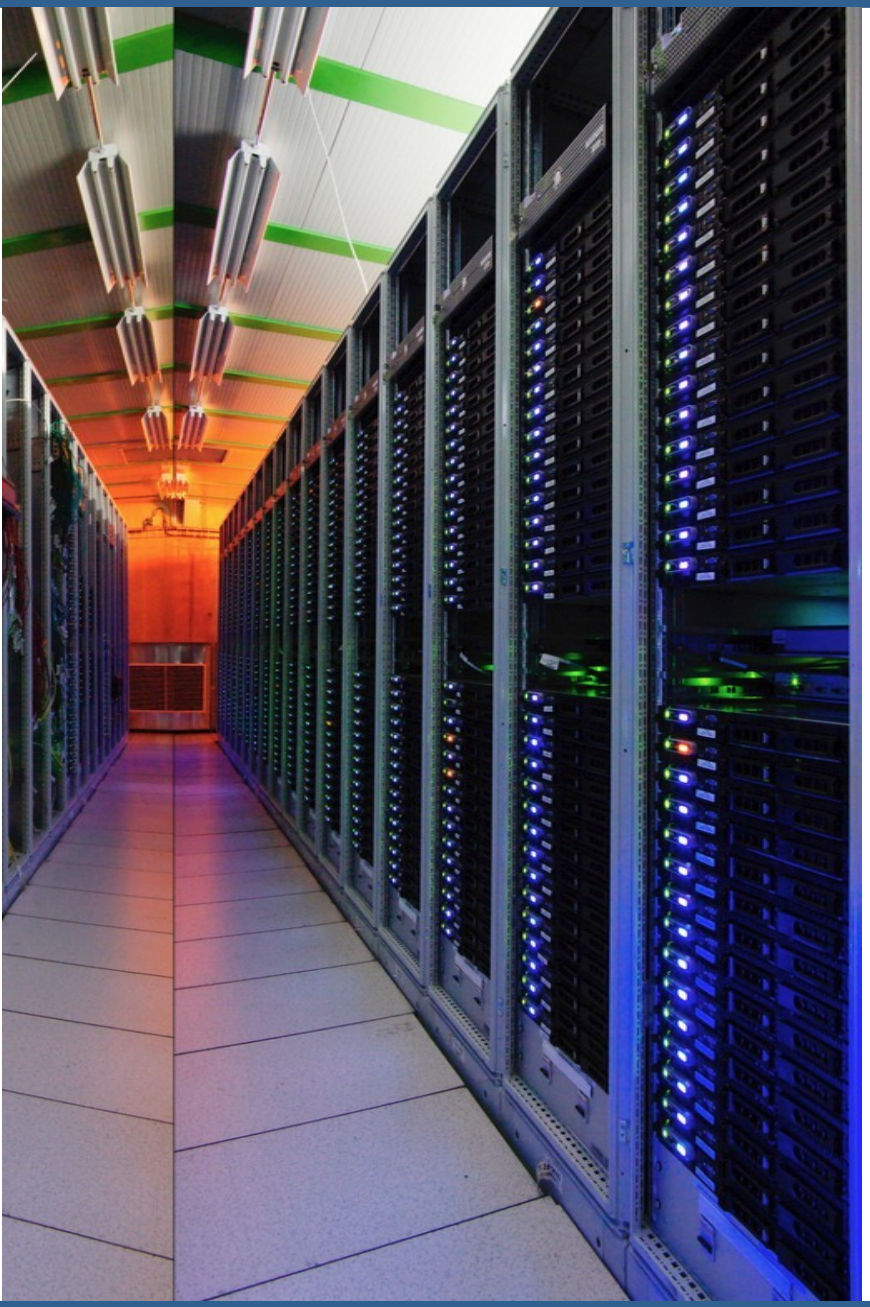
- Lots of tracks, rings, curved / spiral trajectories, spurious measurements and various other imperfections
- Problem scales like N^L where
  - N is the number of particles
  - L is the number of layers

Vertex finding



- Primary vertex:
  - Reconstruct where the collision occurred
- Secondary vertex:
  - Detect particle decays

Software on commodity hardware



- Commodity x86 servers
  - ~2000 per experiment
- Huge existing codebases
  - A few millions of LOC
  - Mostly single-threaded
  - Exploit the inherent parallelism of the data
- Current figures:
  - Input: 1000-100 kHz
  - Output: 10-1 kHz
  - Processing time per event: 10-100 ms

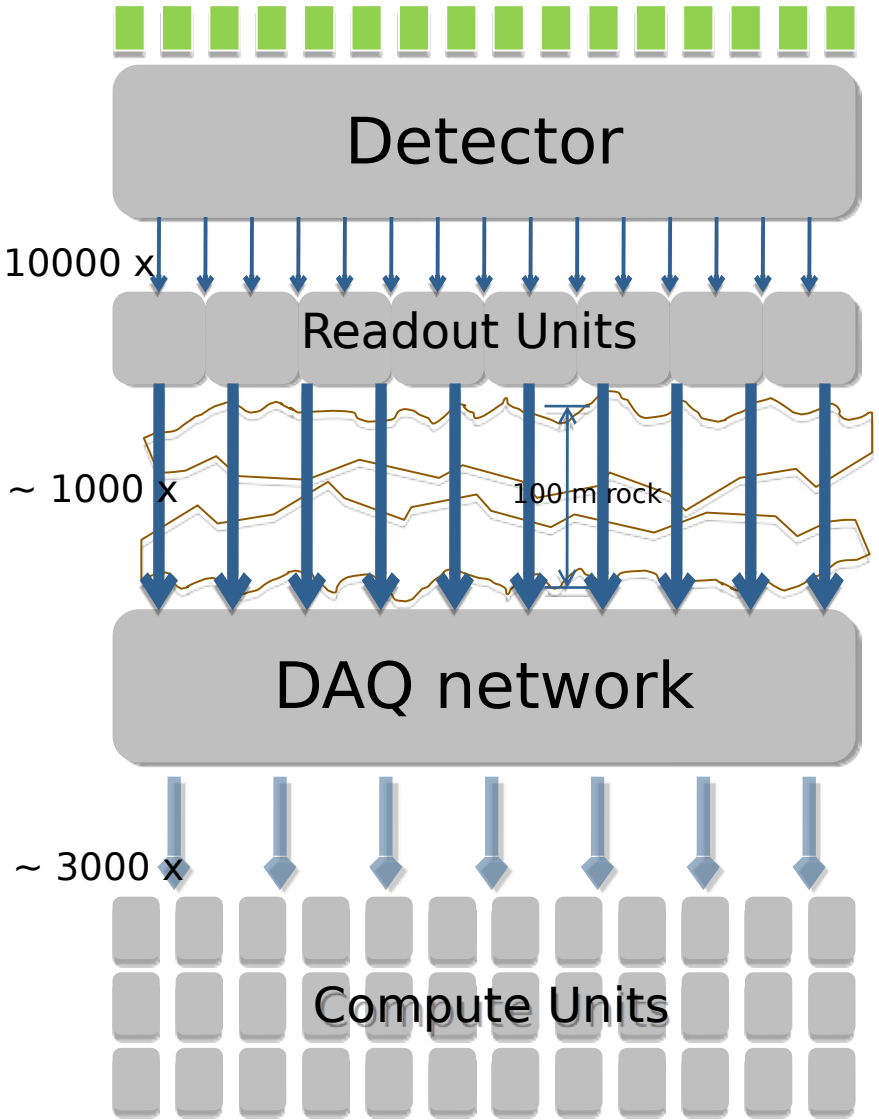
Summary: reconstruction steps

- Data decoding
  - Associate geometrical information and detector calibration to the data
- Clusterisation
  - Particles frequently create signal in adjacent sensors
  - These signals are combined into “clusters”
- The clusters are combined into “tracks” (pattern recognition)
  - Clusters of hits associated to the same particle are grouped together
- Tracks are combined to reconstruct vertices
- Events are selected according to the reconstructed physics objects
  - e.g. 4 muons with a transverse momentum > 20 GeV

Data Acquisition & Event Building,

or: how do we get the data to the HLT?

Data flow



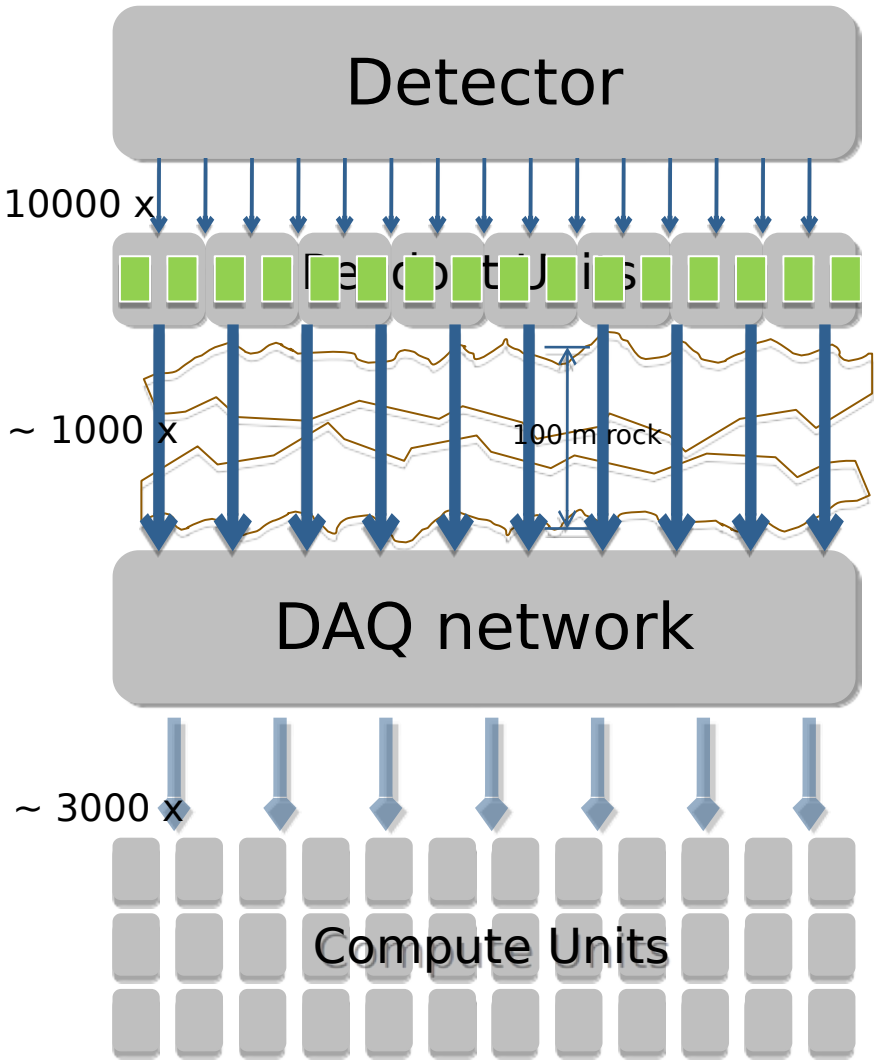
- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering
  (High Level Trigger – HLT)

Custom point-to-point radiation-hard link from the detector front-end

DAQ (event-building) links
Commodity LAN
(Ethernet, InfiniBand)

Links into compute units
Slower variant of DAQ links
(HLT is CPU limited)

Data flow



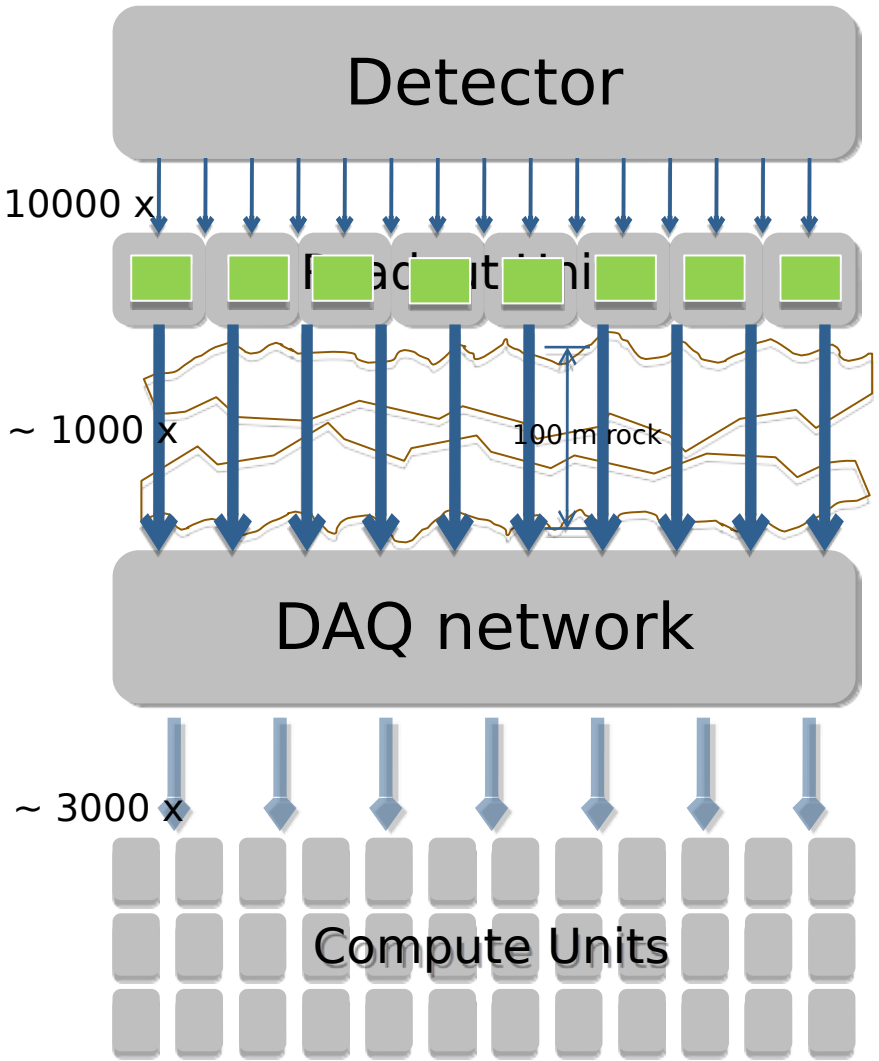
- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering
  (High Level Trigger – HLT)

Custom point-to-point radiation-hard link from the detector front-end

DAQ (event-building) links
Commodity LAN
(Ethernet, InfiniBand)

Links into compute units
Slower variant of DAQ links
(HLT is CPU limited)

Data flow



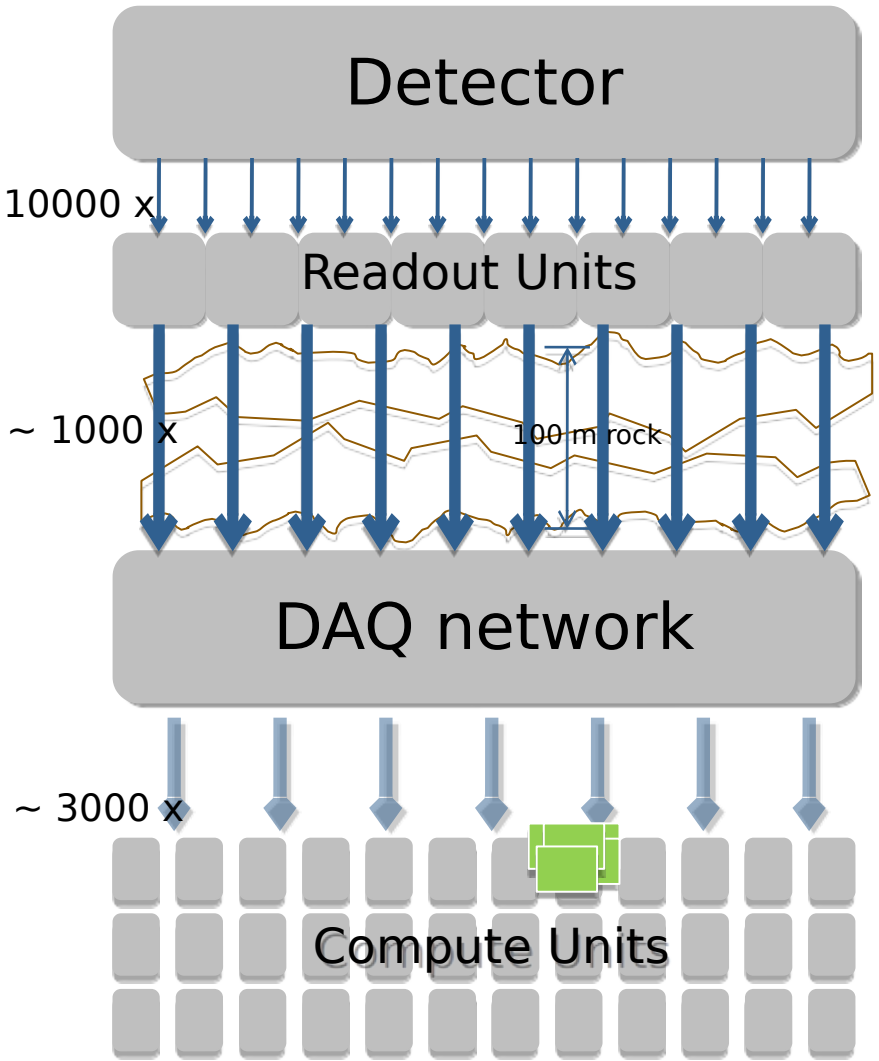
- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering
(High Level Trigger – HLT)

Custom point-to-point radiation-hard link from the detector front-end

DAQ (event-building) links
Commodity LAN
(Ethernet, InfiniBand)

Links into compute units
Slower variant of DAQ links
(HLT is CPU limited)

Data flow



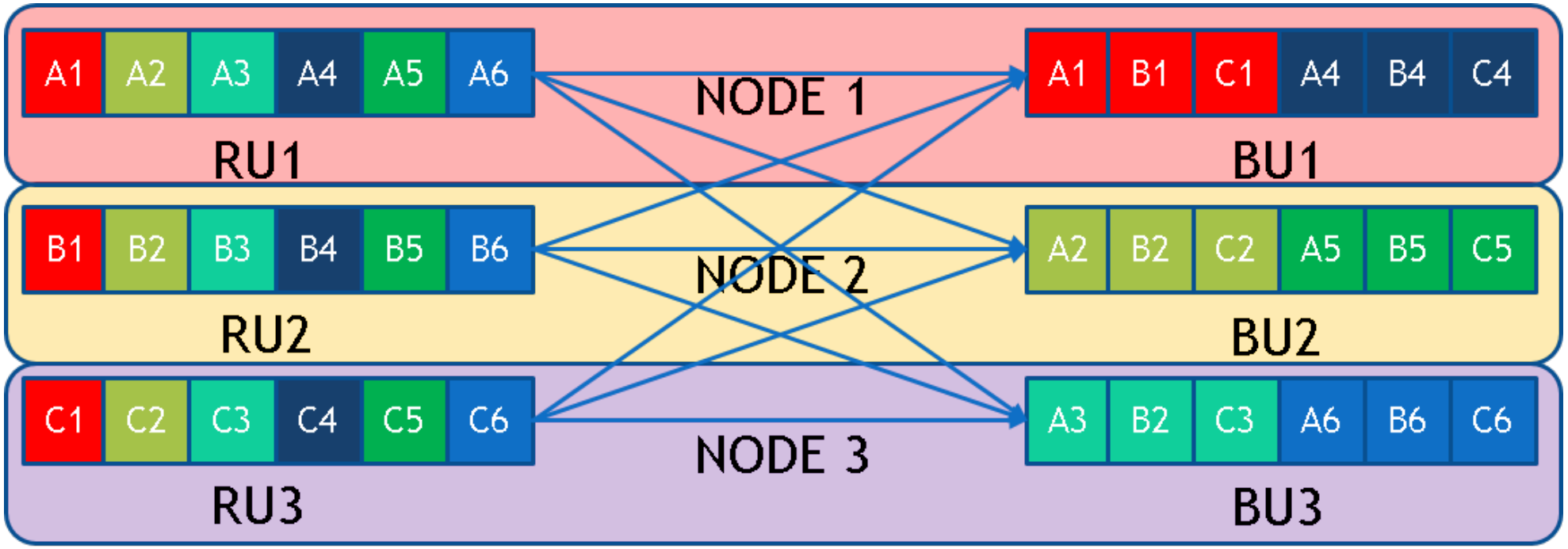
- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering
  (High Level Trigger – HLT)

Custom point-to-point radiation-hard link from the detector front-end

DAQ (event-building) links
Commodity LAN
(Ethernet, InfiniBand)

Links into compute units
Slower variant of DAQ links
(HLT is CPU limited)

Event Building in a nutshell



- RU (Readout Unit): it reads the data out from the detector
- BU (Builder Unit): it receives the full event data

DAQ networks

- The DAQ network aggregates data from different Readout Units

  → Many-to-one communication

- Data transfers are driven by the availability of the data from the detector

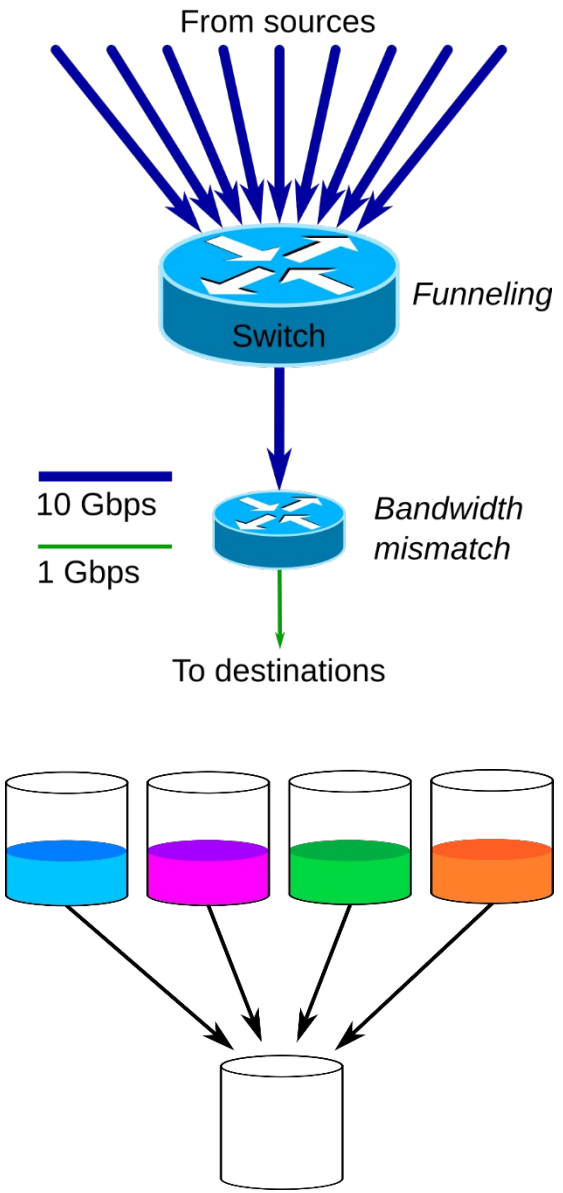
  → Synchronous, bursty traffic

- When many sources send synchronous microbursts of data to a destination

  → The network buffers are overflown

  → Congestion / packet loss

- Must be kept under control, otherwise: “Catastrophic throughput collapse”



DAQ networks

- The DAQ network aggregates data from different Readout Units

  → Many-to-one communication

- Data transfers are driven by the availability of the data from the detector

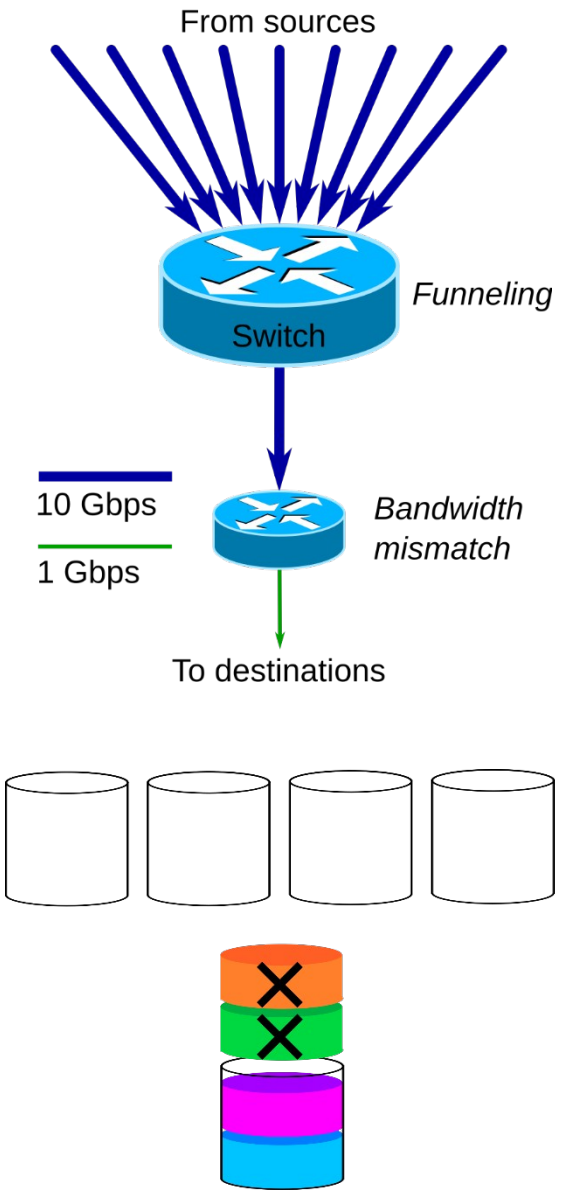
  → Synchronous, bursty traffic

- When many sources send synchronous microbursts of data to a destination

  → The network buffers are overflown

  → Congestion / packet loss

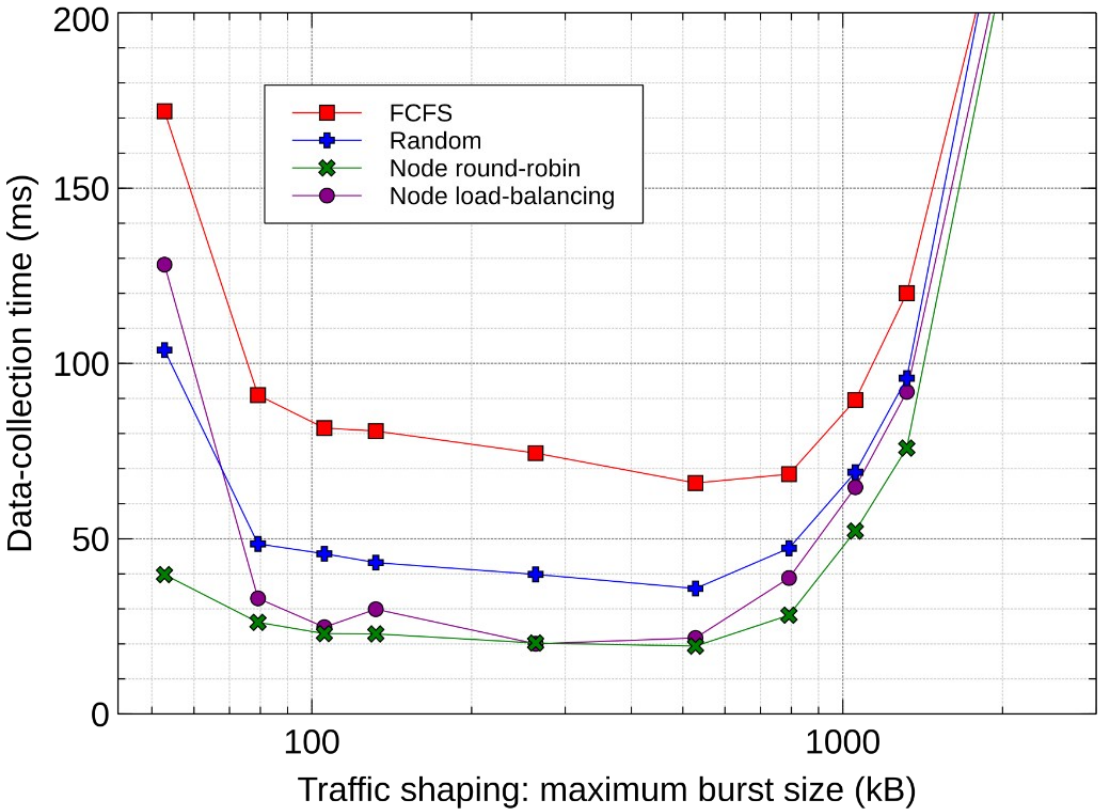
- Must be kept under control, otherwise: “Catastrophic throughput collapse”



Traffic shaping to the rescue

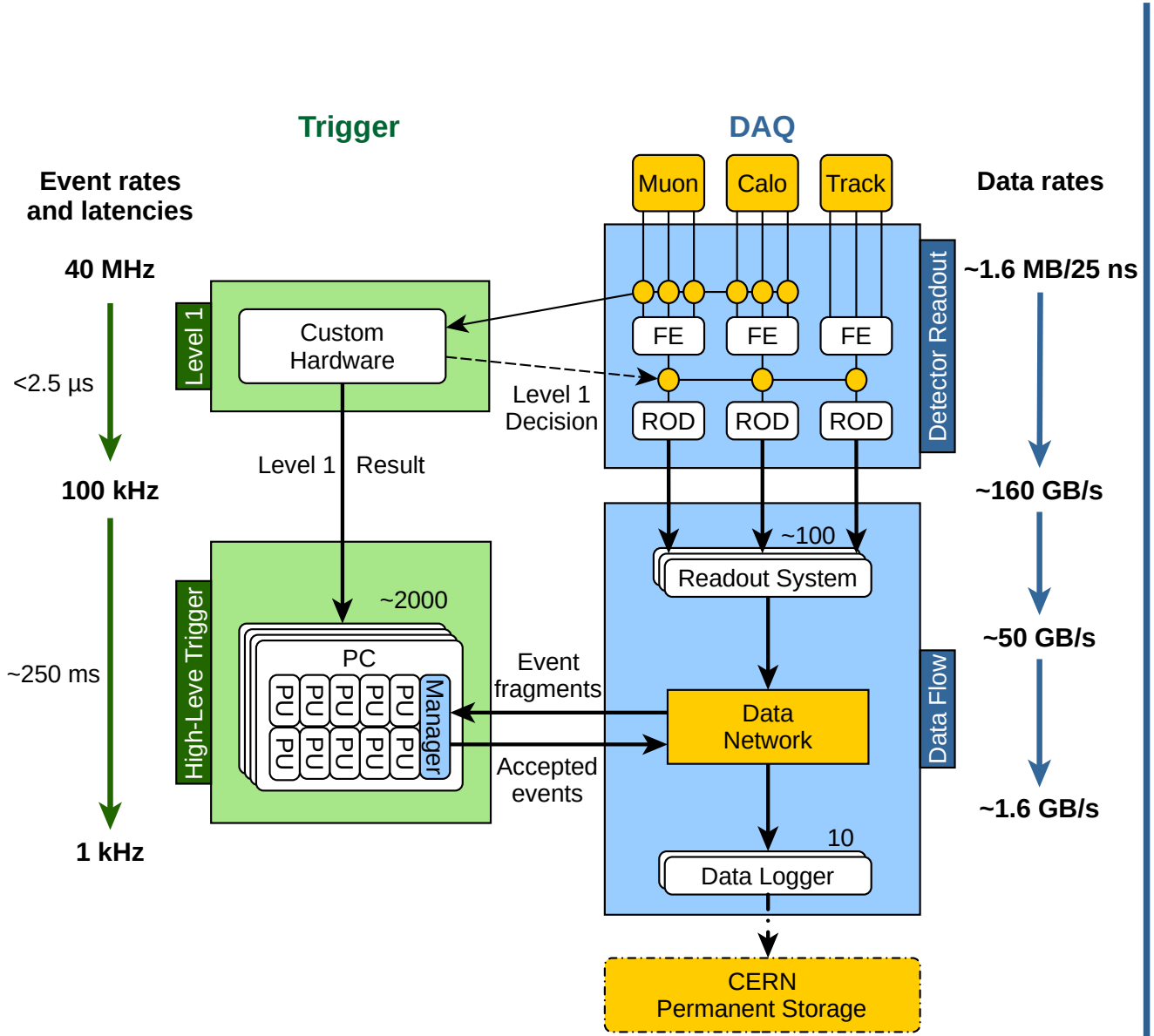
- In DAQ, we can precisely control how we use the network
- Traffic shaping:

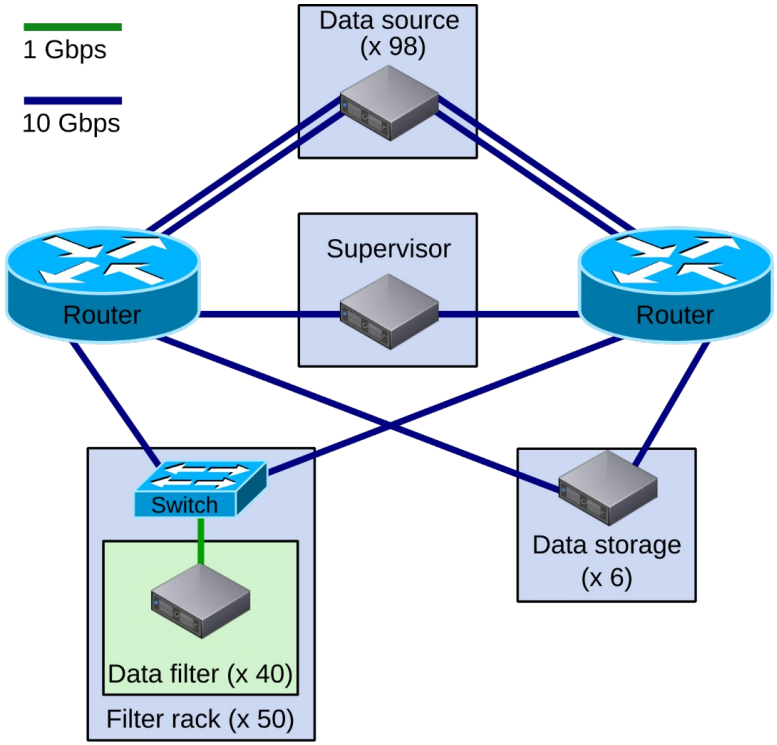
  Prevent too many synchronized sources from sending to the same destination at the same time
- Tuning needed:
  - Shaping too aggressive
    → bottleneck
  - Shaping too lax:
    → congestion



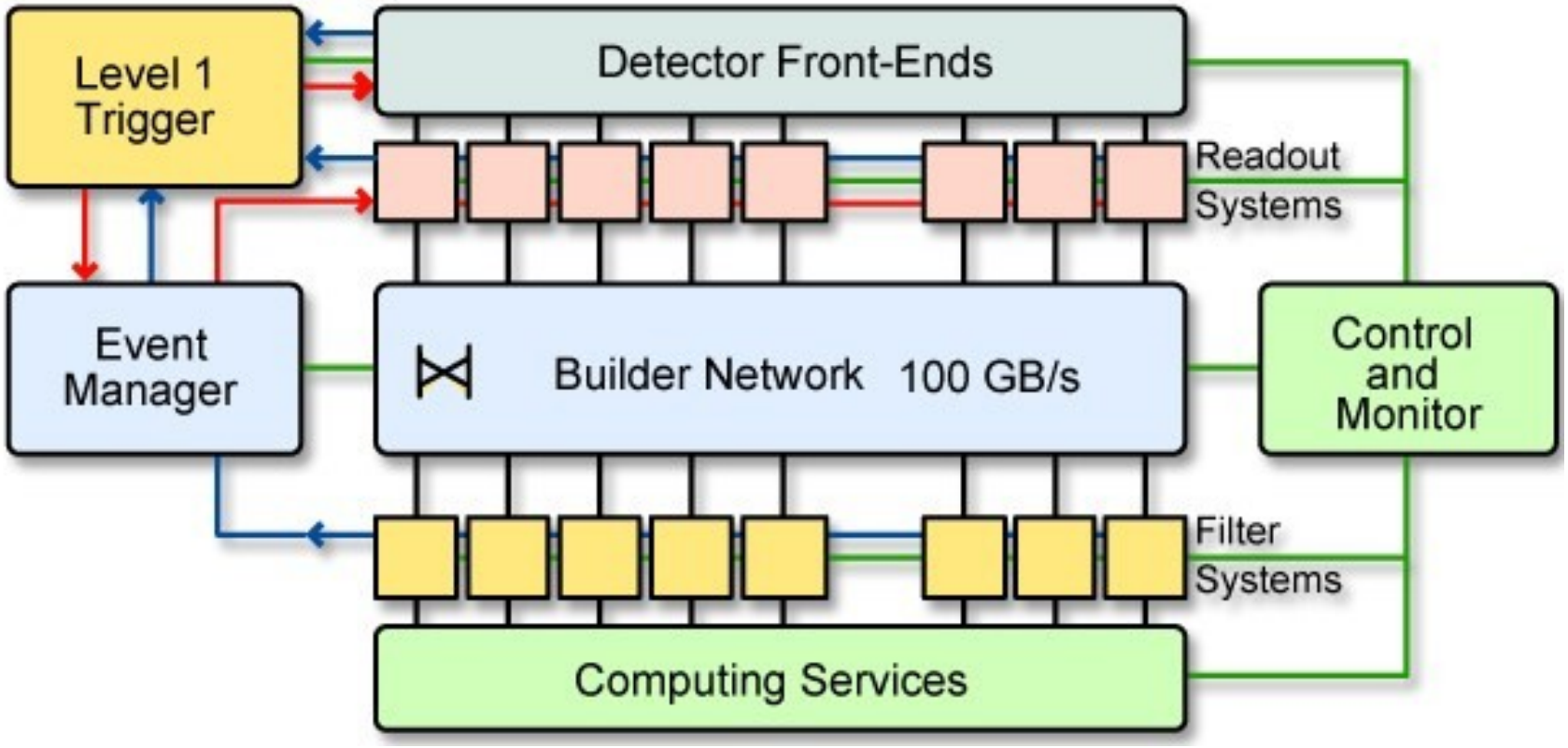
Putting it all together

ATLAS Trigger & DAQ in Run2





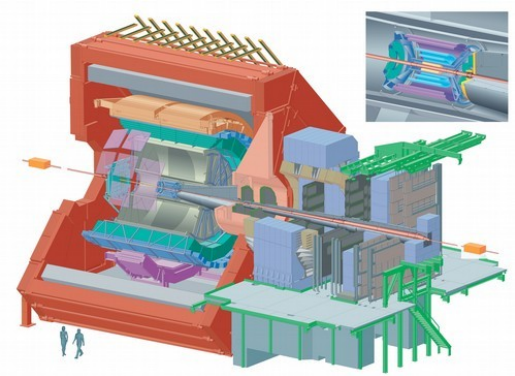
CMS Trigger & DAQ in Run2

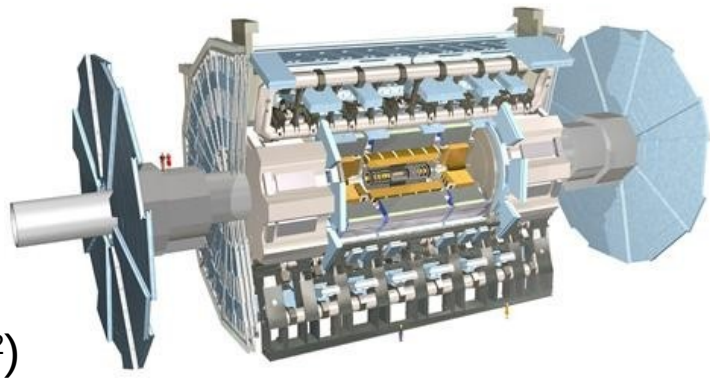


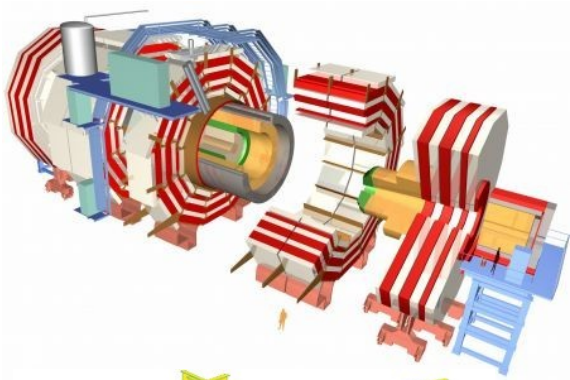
- Bunch Crossing rate: 40 MHz
- Level-1 Latency: 3.2 μs
- Level-1 Output: 100 kHz
- Output to Storage: 400 Hz
- Average Event Size: 1 MB
- Data production 1 TB/day

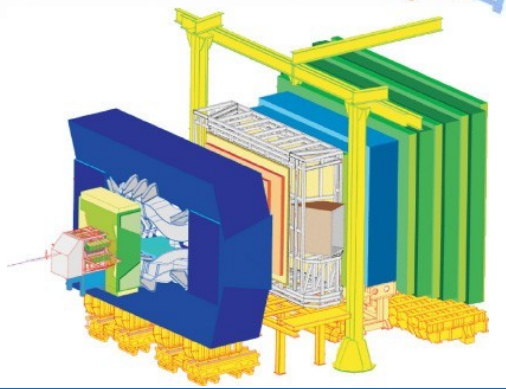
Trigger & DAQ datasheets

	Level-1 Rate (Hz)	Event Size (Byte)	Readout Bandw.(GB/s)	HLT Out MB/s (Event/s)
ALICE	**Pb-Pb 500** **p-p 10^3**	**$5x10^7$** **$2x10^6$**	**25**	**1250** (10^2) **200** (10^2)
ATLAS	**10^5**	**$1.5x10^6$**	**50**	**~1000** (10^2)
CMS	**10^5**	**10^6**	**100**	**~1000** (10^2)
LHCb	**10^6**	**$5x10^4$**	**50**	**700** ($1.2x10^4$)





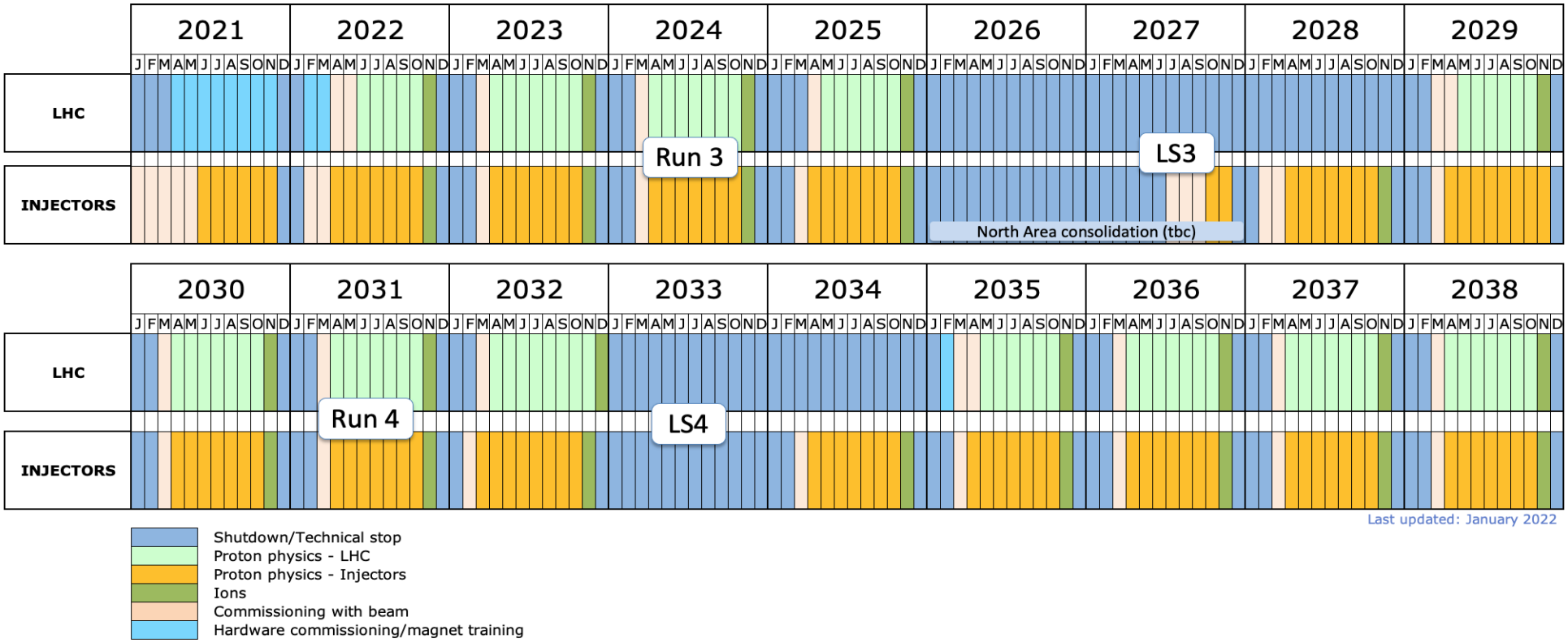




Harder, Better, Faster, Stronger

The upgrades

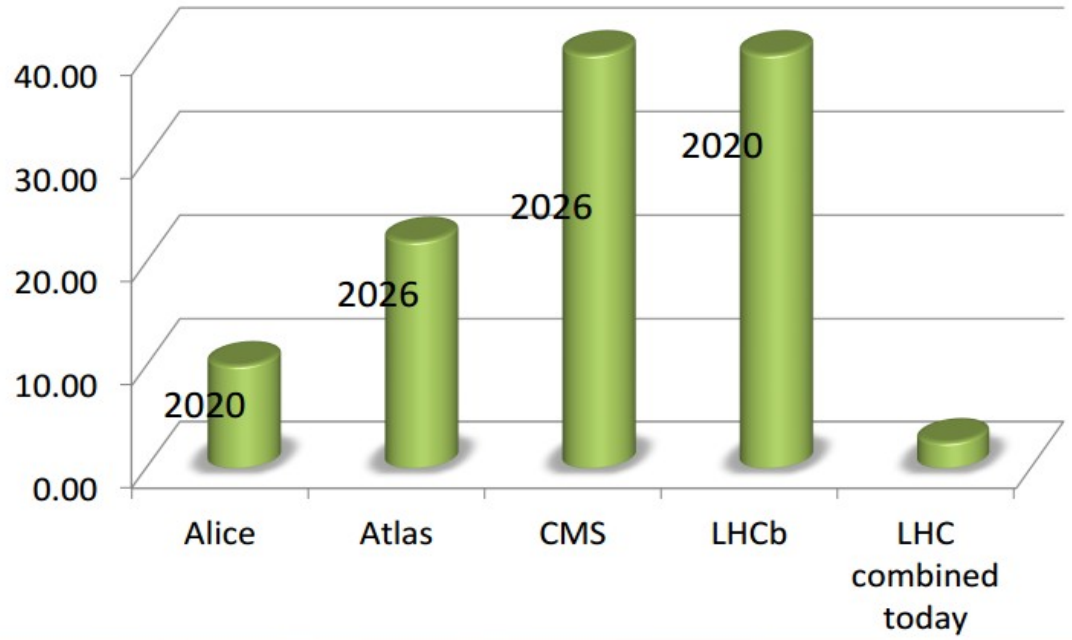
The LHC long-term planning



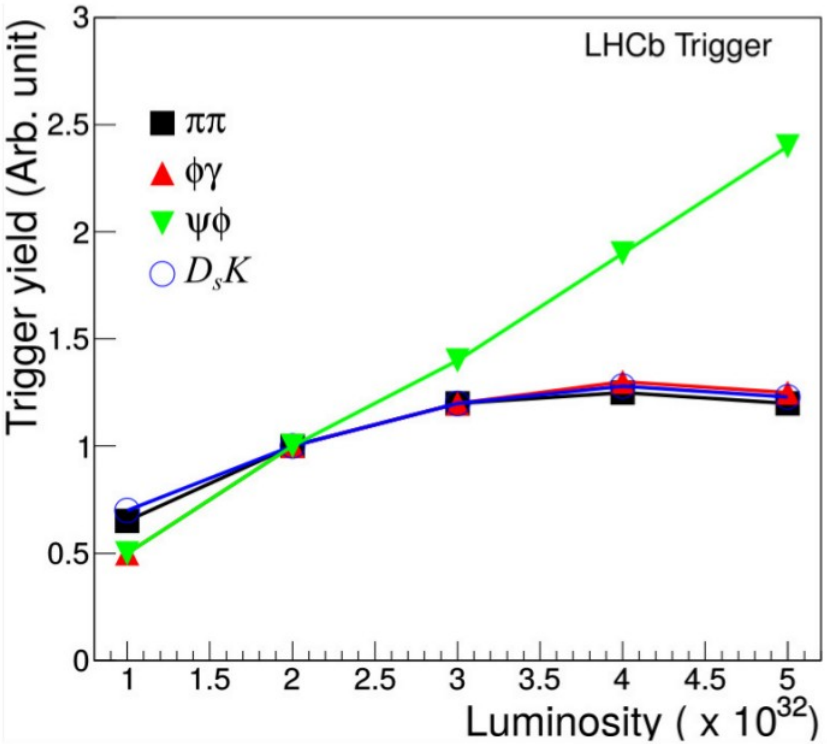
Experiment upgrade plans

	Event-size [kB]	Rate of events into HLT [kHz]	HLT bandwidth [Gb/s]	Year
ALICE	20000	50	8000	2021
ATLAS	4000	200	6400	2026
CMS	4000	1000	32000	2026
LHCb	100	40000	32000	2021

DAQ network throughput

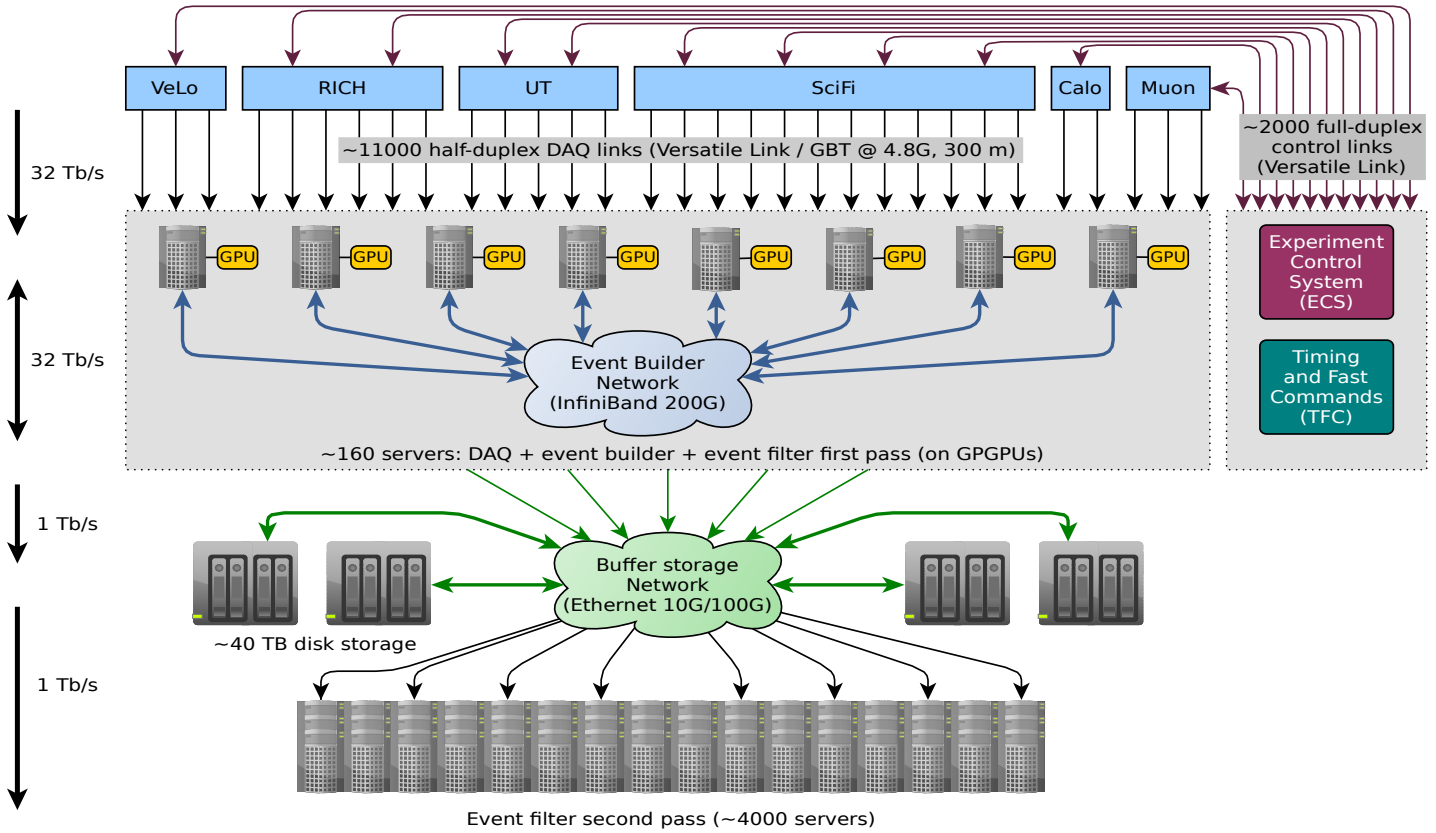


Why can LHCb read out all the collisions?



- Spectrometer geometry (fibres/cables are not "in the way")
- Relatively low radiation levels permit to relax the constraint on the FPGAs used for "middle" layer processing
- Smaller total event-size ~100 kB

DAQ challenge for Run3



- Multi-Terabit/s DAQ system
- Heterogeneous computing infrastructure (GPGPU + CPU)
- Different network technologies, the right link for the right task.

Summary

- The LHC experiments need to reduce ~50 TB/s to ~1 GB/s (or rather, 25 PB/ year)
- This is achieved with massive use of FPGAs, custom ASICs, x86 CPUs and GPGPUs
- Large commodity local area networks are used to distribute data among the individual computing elements
- The future will see massive increase of required programmable computing power and required networking bandwidth, much more data will be moved off detector

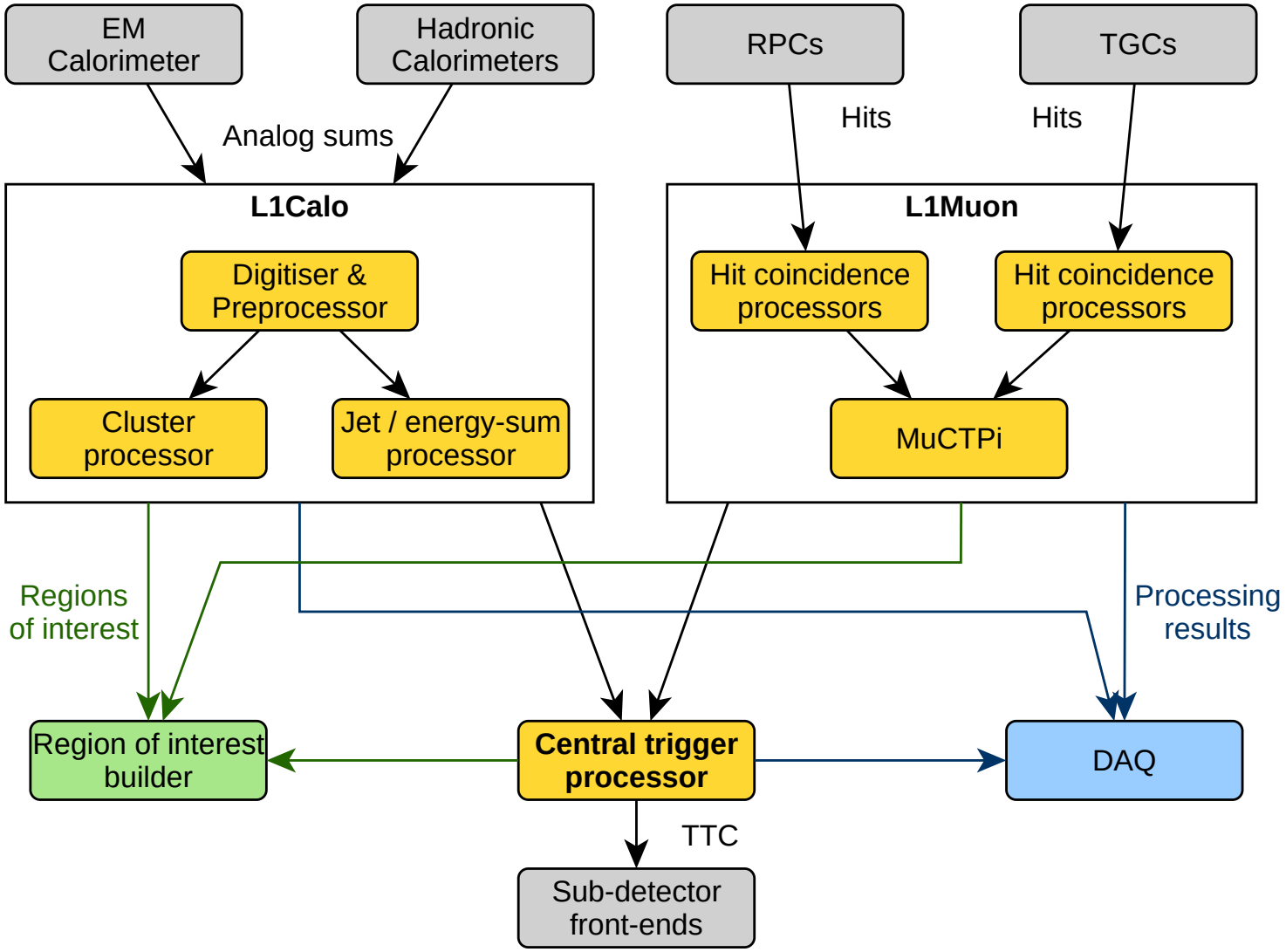
Thank you for your attention

Backup

Backup: Zero suppression

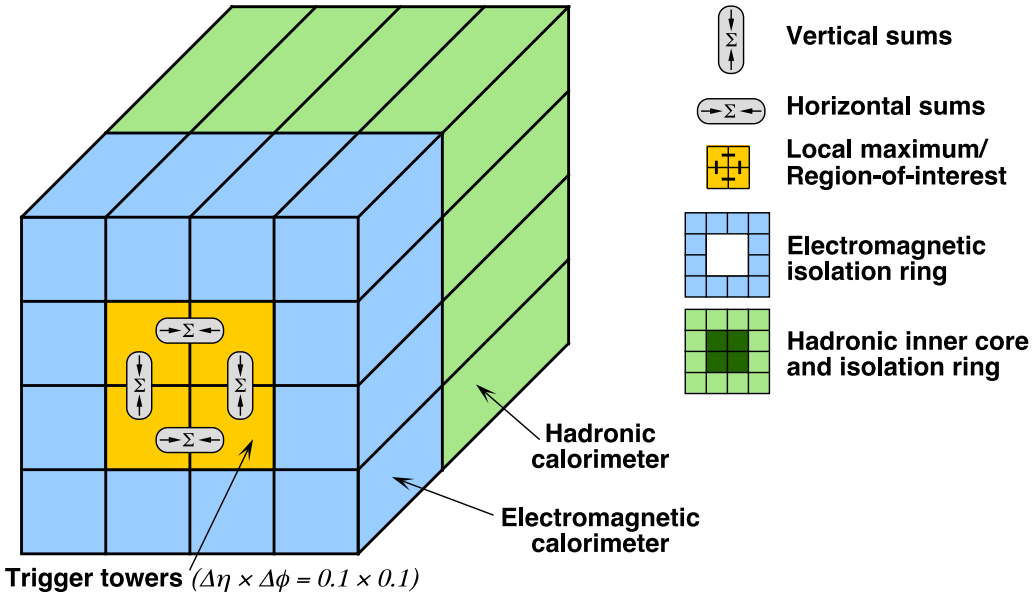
- Why spend bandwidth sending data that is zero for the majority of the time ?
- Perform zero-suppression and only send data with non-zero content
  - Identify the data with a channel number and/or a time-stamp
  - We do not want to loose information of interest so this must be done with great care taking into account pedestals, baseline variations, common mode, noise, etc.
  - Not worth it for occupancies above ~10%
- Alternative: data compression
  - Huffman encoding and alike, but needs power, silicon…
- TANSTAFL (There Aint No Such Thing As A Free Lunch)
  - Data rates fluctuates all the time and we have to fit this into links with a given bandwidth
  - Not any more event synchronous
  - Complicated buffer handling (overflows)
  - Before an experiment is built and running it is very difficult to give reliable estimates of data rates needed (background, new physics, etc.)

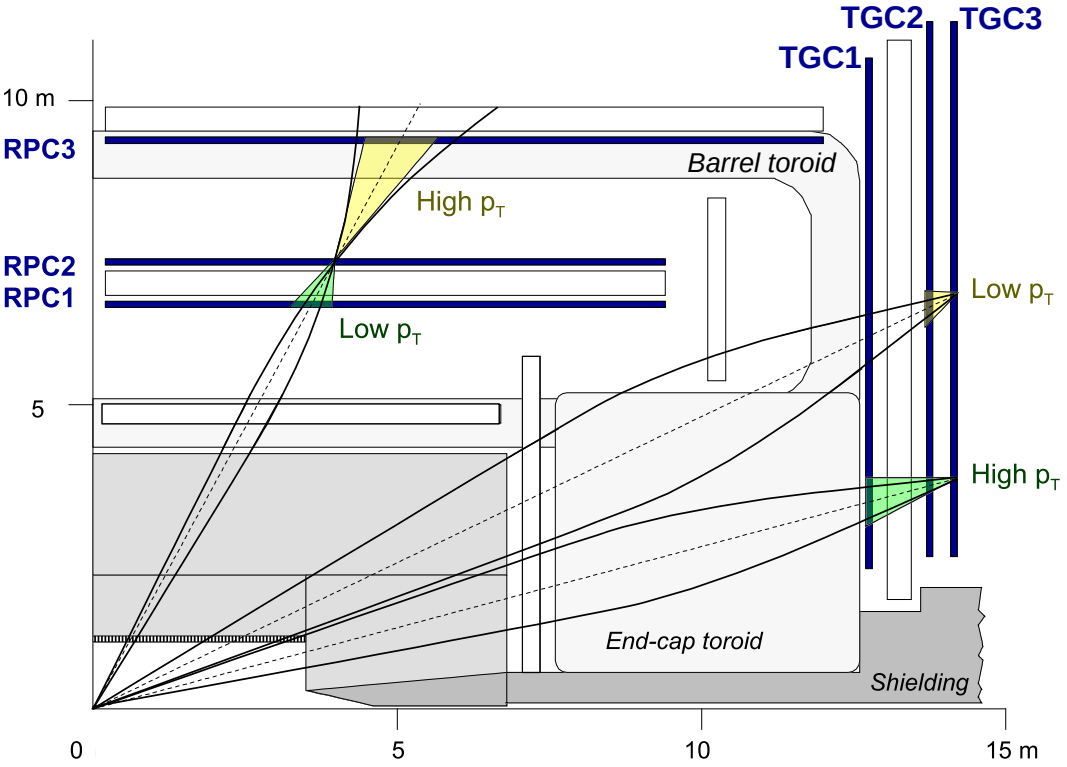
Backup: ATLAS Level-1



Backup: Level-1 algorithms

- Calorimeters
  - Cluster finding
  - Energy deposition evaluation





- Muon systems
  - Segment/track finding
  - Momentum evaluation

Backup: CMS Run2 DAQ details