

Big Data and rare events: The boson in the hay-stack From Pbit/s to Pbyte/year

Niko Neufeld, CERN/PH-Department
niko.neufeld@cern.ch

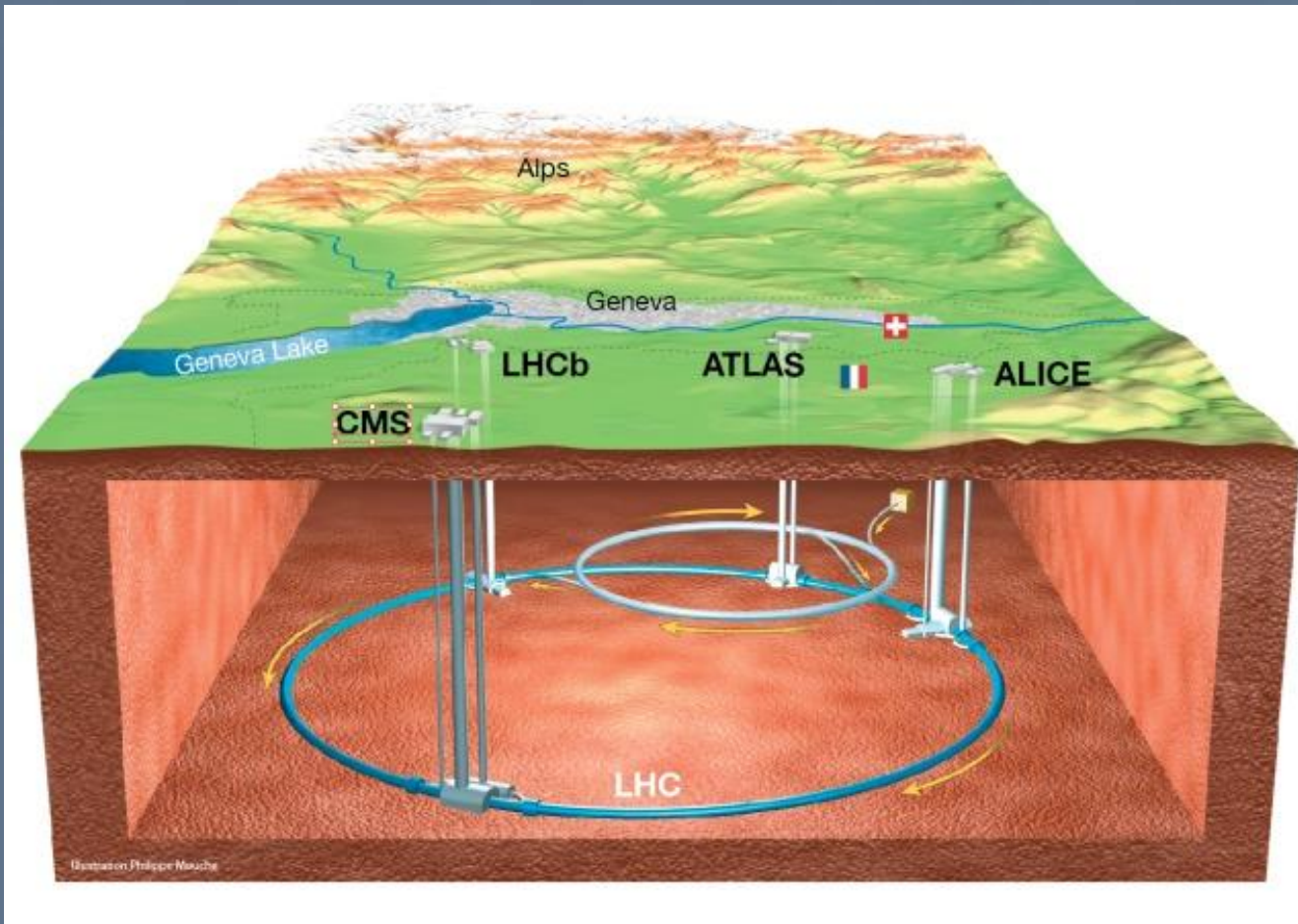
Disclaimer

- Trigger and DAQ are vast subjects covering a lot of physics and engineering
- Based entirely on personal bias I have selected a few topics
- While most of it will be only an overview at a few places we will go into some technical detail
- Some things will be only touched upon or left out altogether:
 - Derivation of the “physics” in the trigger
 - Mathematical treatment of DAQ/trigger problems (queuing theory)
 - DAQ of experiments outside HEP/LHC
 - Management of large networks and farms & High-speed mass storage
 - Control Systems

Acknowledgments

- Much of the material I have taken (and often modified) from colleagues such as O. Bähring, A. Hirstius, R. Schwemmer, W. Smith and various other presentations
- A missing indication of origin of a figure does not imply that it is originally mine

The Large Hadron Collider



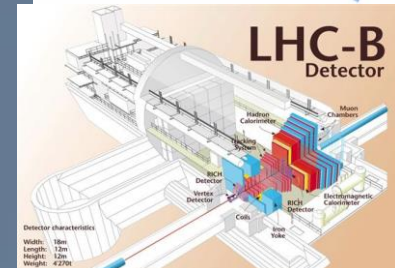
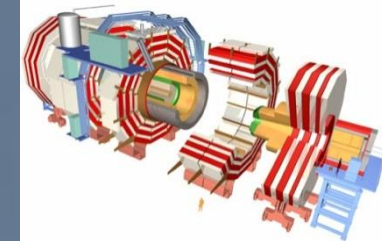
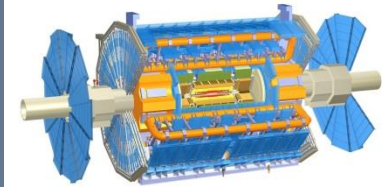
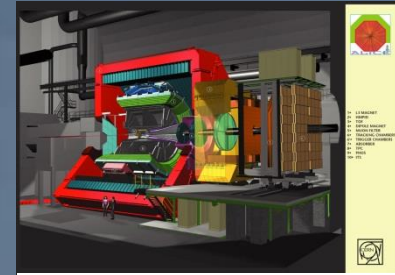
- 27 km
- Vacuum at 10^{-13} atm
- More than 9600 magnets
- Dipole magnets at $-271.3\text{ C} \rightarrow 0.8\text{ C}$ colder than outer space
- Energy in the beam corresponds to a TGV at 150 km/h
- Cost: 5 billion CHF
- 4 large experiments

So what do we do with all that?

The LHC Experiments today

- ALICE – “A Large Ion Collider Experiment”
 - Size: 26 m long, 16 m wide, 16m high; weight: 10000 t
 - 35 countries, 118 Institutes
 - Material costs: 110 MCHF
- ATLAS – “A Toroidal LHC ApparatuS”
 - Size: 46 m long, 25 m wide, 25 m high; weight: 7000 t
 - 38 countries, 174 institutes
 - Material costs: 540 MCHF
- CMS – “Compact Muon Solenoid”
 - Size: 22 m long, 15 m wide, 15 m high; weight: 12500 t
 - 40 countries, 172 institutes
 - Material costs: 500 MCHF
- LHCb – “LHC beauty” (b-quark is called “beauty” or “bottom” quark)
 - Size: 21 m long, 13 m wide, 10 m high; weight: 5600 t
 - 15 countries, 52 Institutes
 - Material costs: 75 MCHF
- Regular upgrades ... 2019/20 (Long Shutdown 2)

1 CHF ~ 1 USD



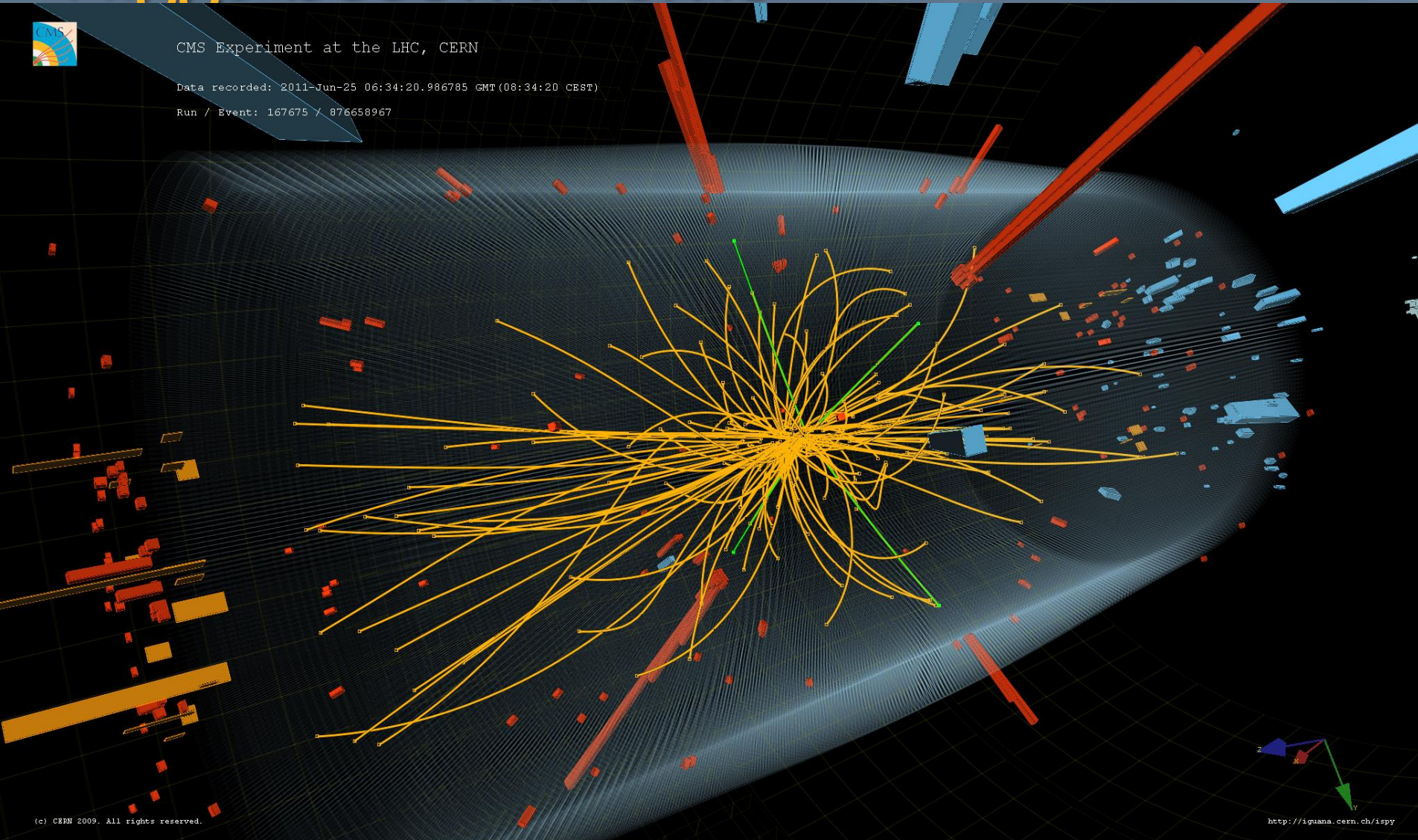
Higgs-boson in CMS



CMS Experiment at the LHC, CERN

Data recorded: 2011-Jun-25 06:34:20.986785 GMT (08:34:20 CEST)

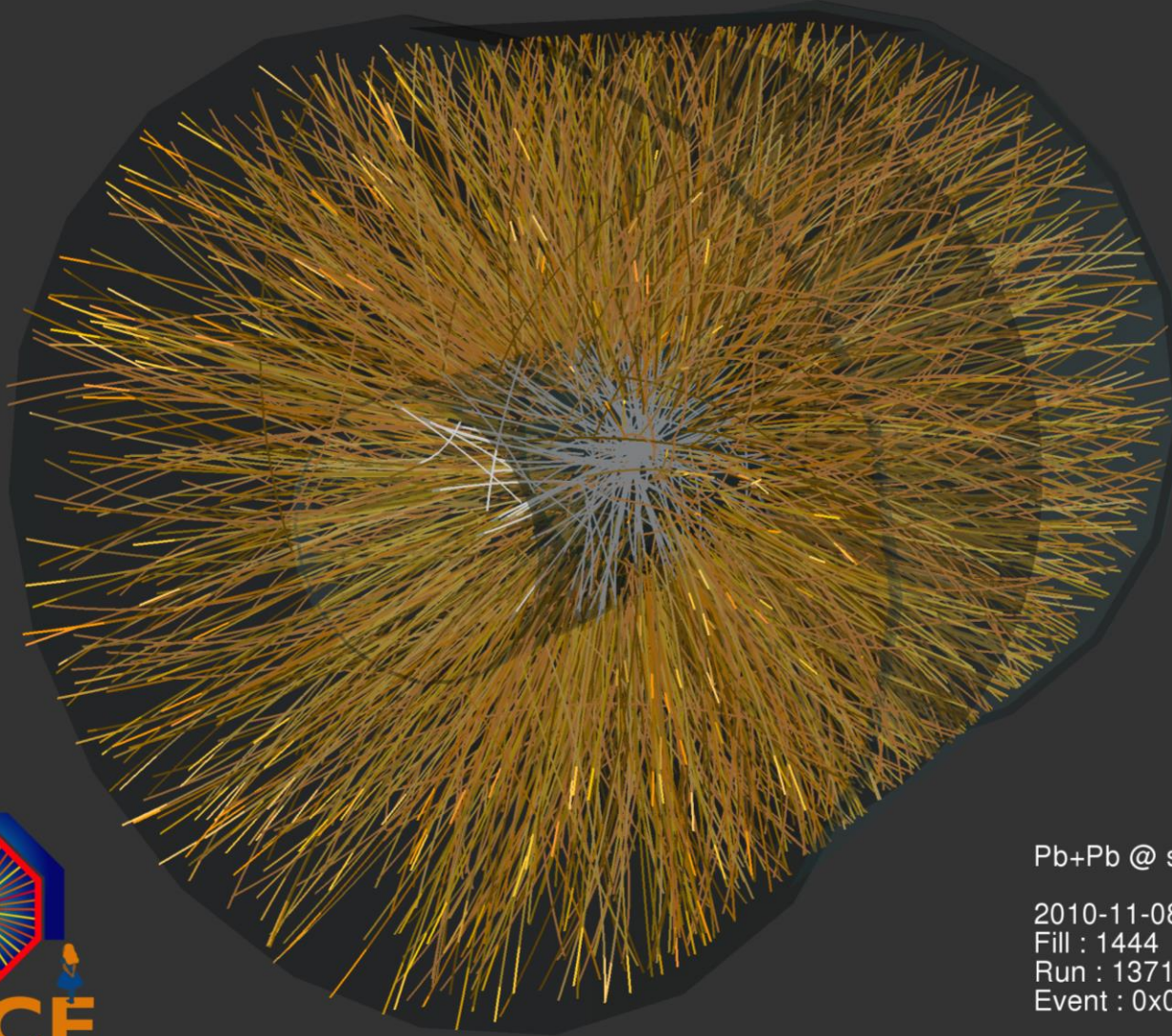
Run / Event: 167675 / 876658967



(c) CERN 2009. All rights reserved.

<http://lqana.cern.ch/ispy>

Lead meets lead in ALICE



Pb+Pb @ $\sqrt{s} = 2.76$ ATeV

2010-11-08 11:29:42

Fill : 1444

Run : 137124

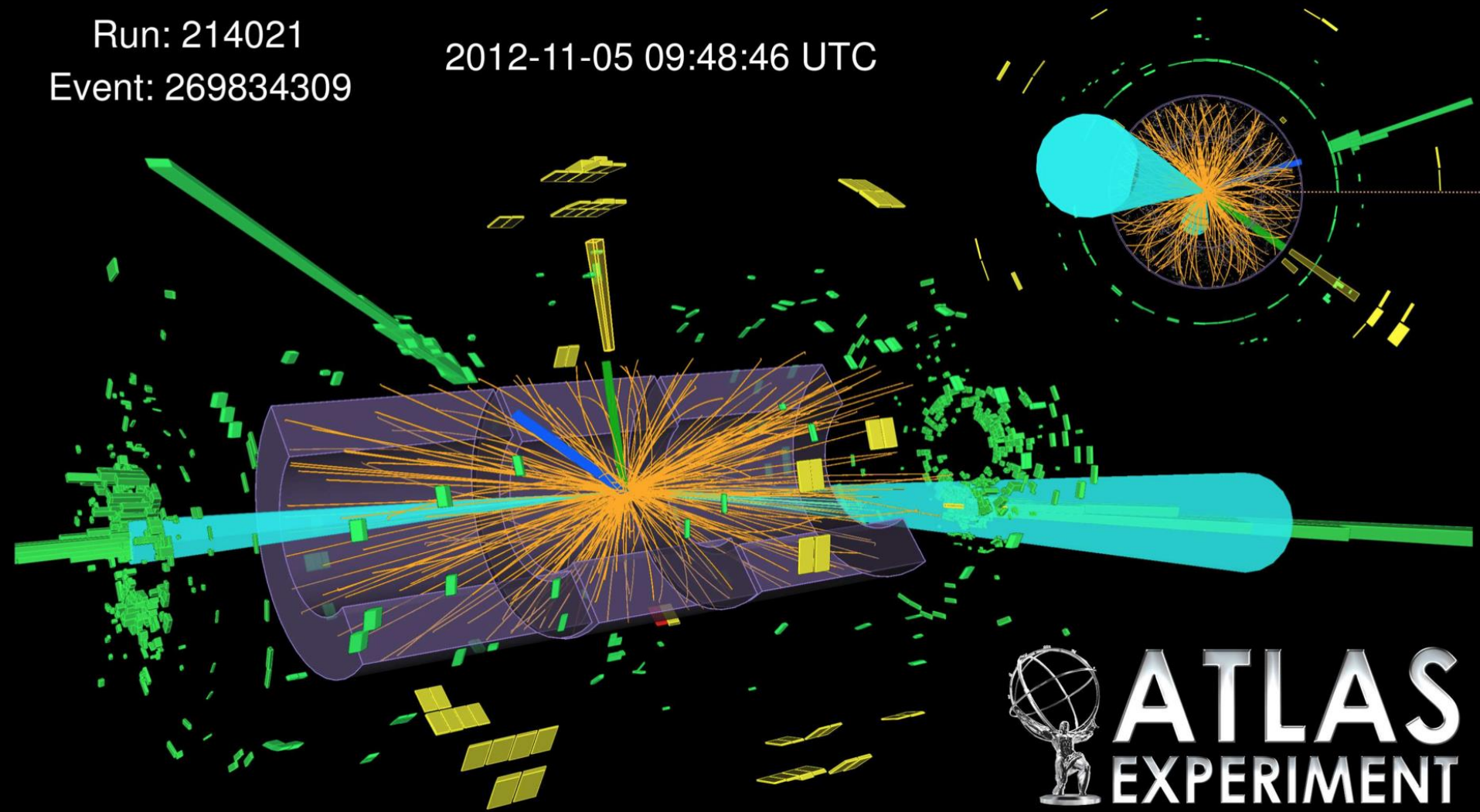
Event : 0x00000000271EC693

Mr. Higgs'es boson is also in ATLAS

Run: 214021

2012-11-05 09:48:46 UTC

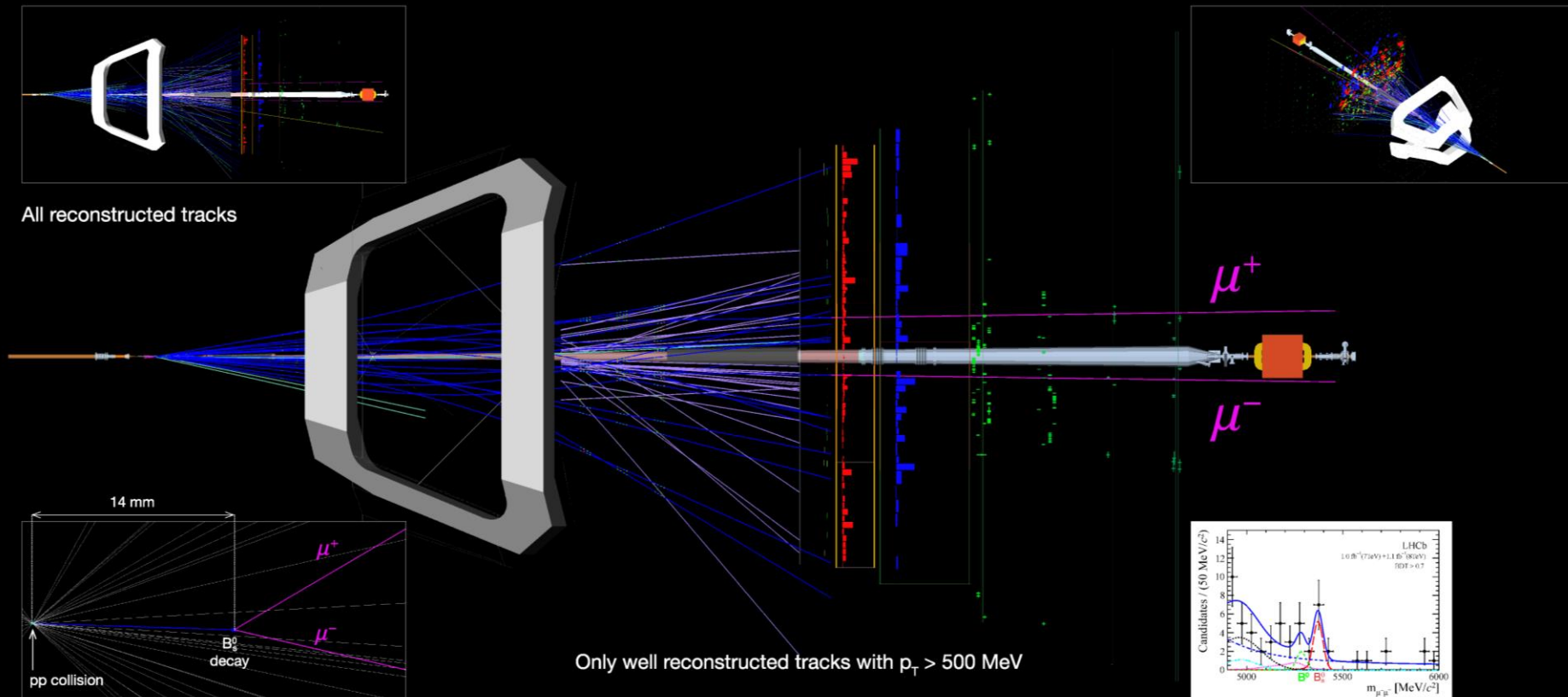
Event: 269834309



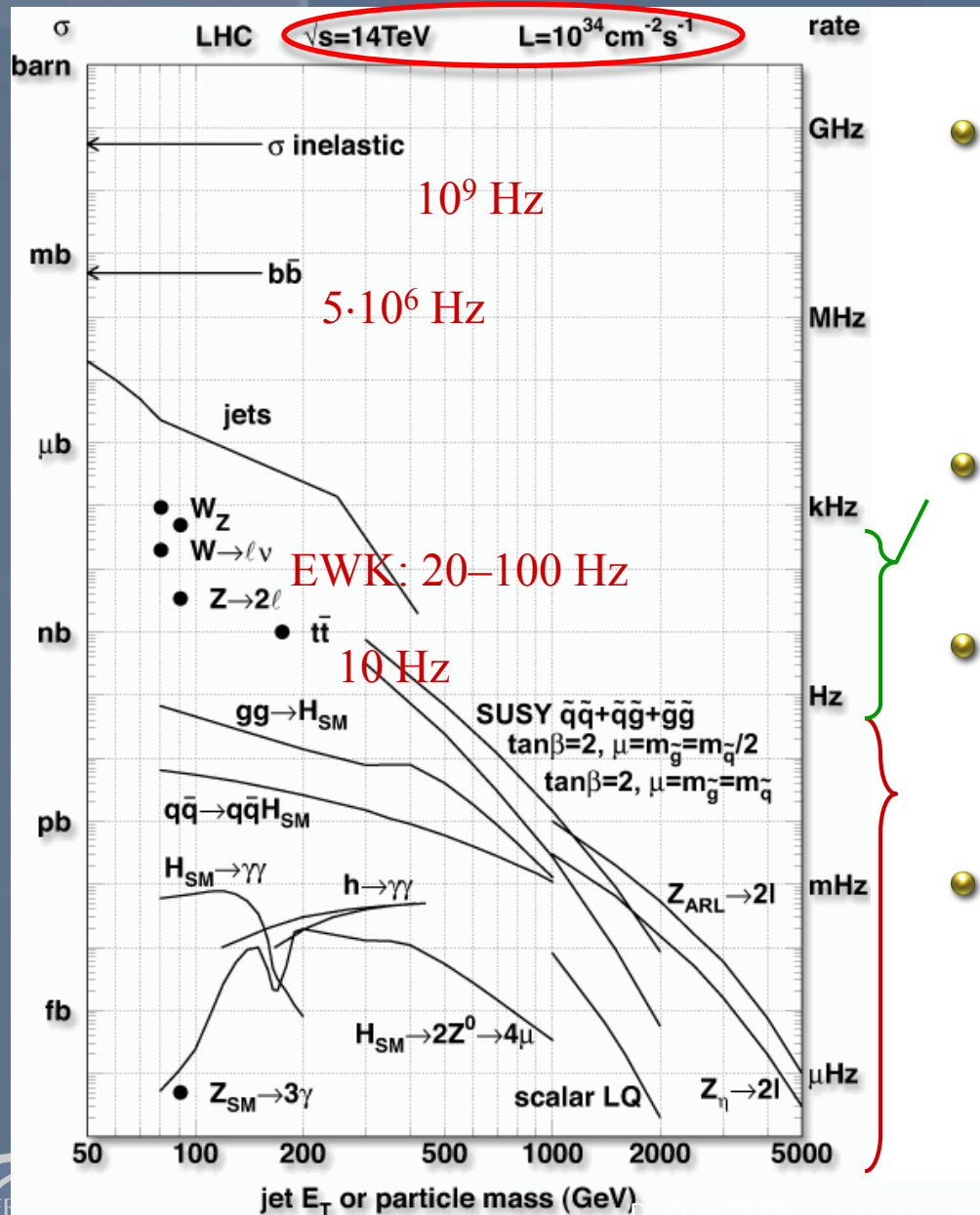
ATLAS
EXPERIMENT

An extremely rare event in LHCb

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



What's mother nature's menu?



A typical collision is “boring”

- Although we need also some of these “boring” data as cross-check, calibration tool and also some important “low-energy” physics

“Interesting” physics is about 6–8 orders of magnitude rarer: one in a million down to one in 100 millions

“Exciting” physics involving new particles/discoveries is ≥ 9 orders of magnitude below σ_{tot} : one in a billion or even more rare

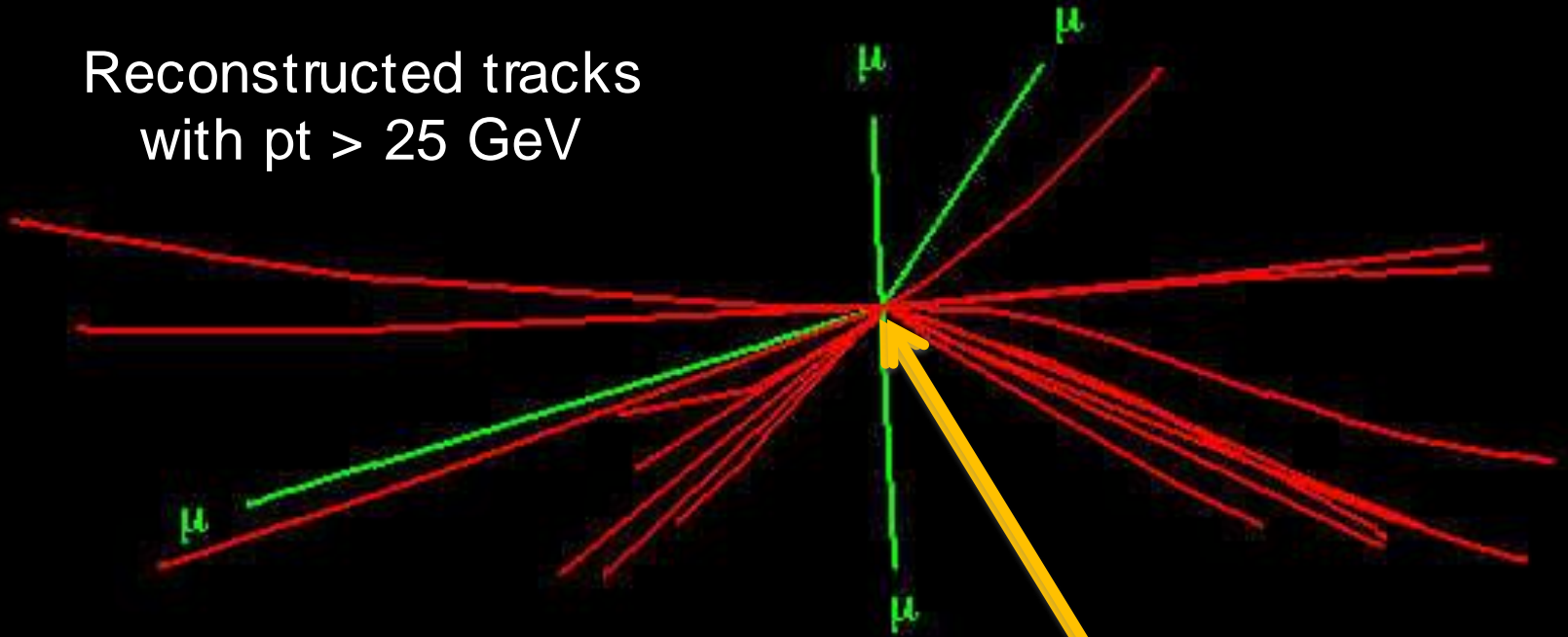
Need to efficiently identify these rare processes from the overwhelming background before reading out & storing the whole event

The boson in the hay-stack



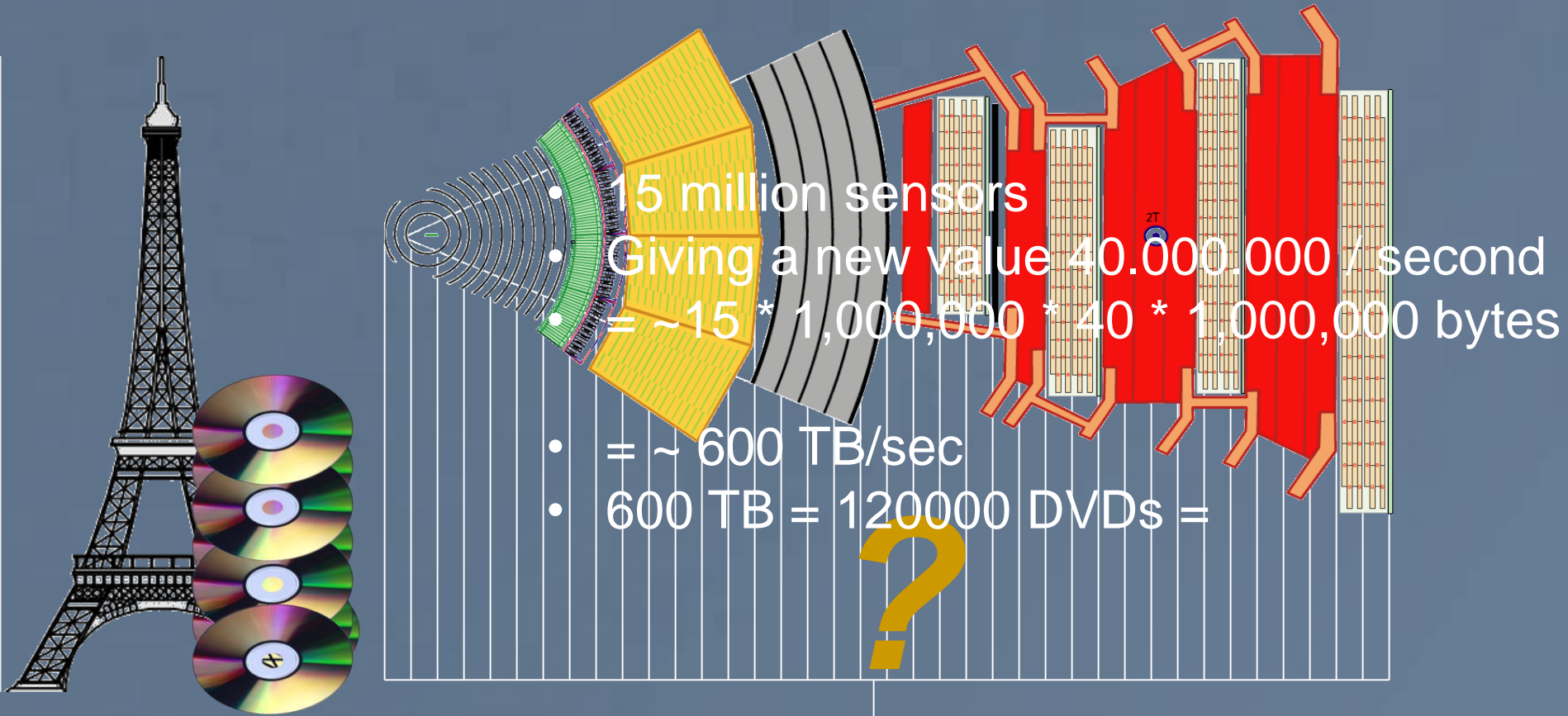
Simulation from CMS

Reconstructed tracks
with $p_t > 25$ GeV



This is what we find: looking for 1000 good bosons
We get this couple 0.001 times a day (s)

The “hay”: $O(10)$ million sensors



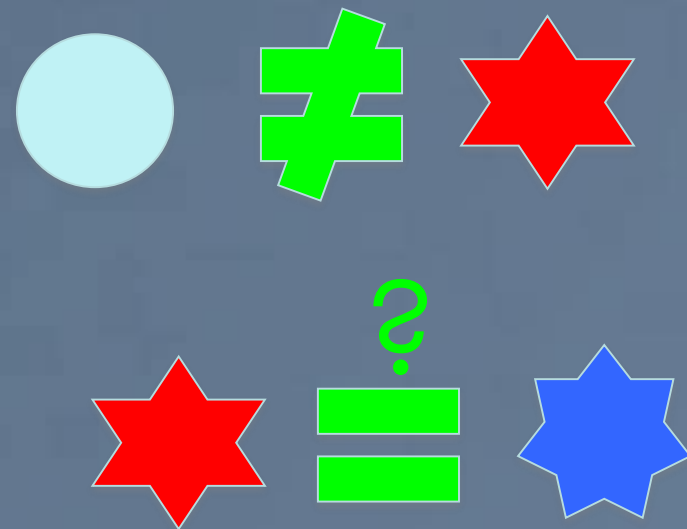
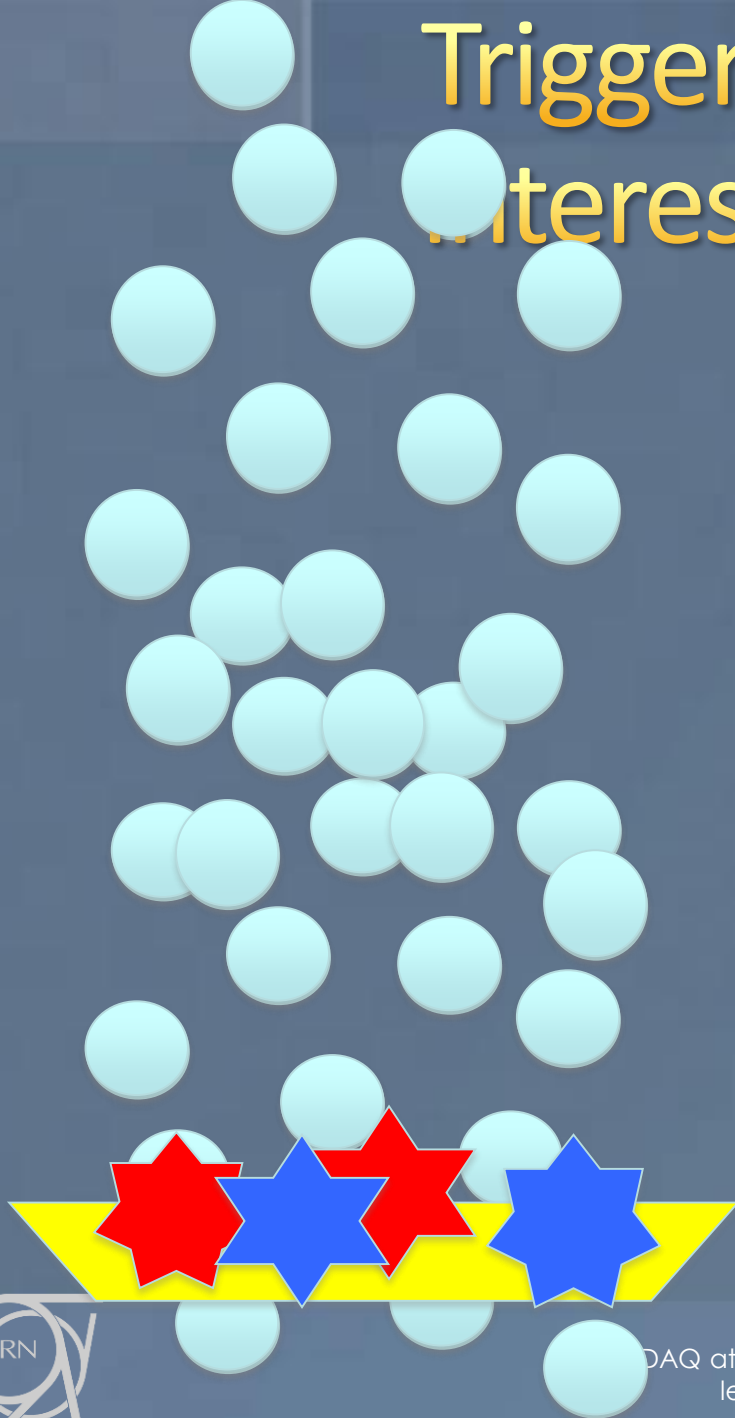
The diagram illustrates the LHC detector structure, showing the tunnel, the detector, and the data volume. On the left, the Eiffel Tower is shown next to a stack of four DVDs, representing the scale of the data volume. The detector structure is shown in the center, with a large yellow question mark indicating the challenge of handling the data. The text on the right lists the sensor count, data rate, and data volume.

- 15 million sensors
- Giving a new value 40.000.000 / second
- $= \sim 15 * 1,000,000 * 40 * 1,000,000$ bytes
- $= \sim 600$ TB/sec
- 600 TB $= 120000$ DVDs $=$

How do you sift through 600 Terabytes / s?

This means going through a 100 m high stack of DVDs

Triggering – selecting the interesting few



Filter 399 out of 400 collisions
Must keep the good = interesting ones

Data Rates

- Particle beams cross every 25 ns (40 MHz)
 - Up to 25 particle collisions per beam crossing
 - Up to 10^9 collisions per second
- Two event filter/trigger levels
 - Data processing starts at readout
 - Reducing 10^9 p-p collisions per second to ~ 1000 per second
- Raw data to be stored permanently: >25 PB/year

Physics Process	Events/s
Inelastic p-p scattering	10^8
b	10^6
$W \rightarrow e\nu ; W \rightarrow \mu\nu ; W \rightarrow \tau\nu$	20
$Z \rightarrow ee ; Z \rightarrow \mu\mu ; Z \rightarrow \tau\tau$	2
t	1
Higgs boson (all; $m_H = 120\text{GeV}$)	0.04
Higgs boson (simple signatures)	0.0003
Black Hole (certain properties)	0.0001

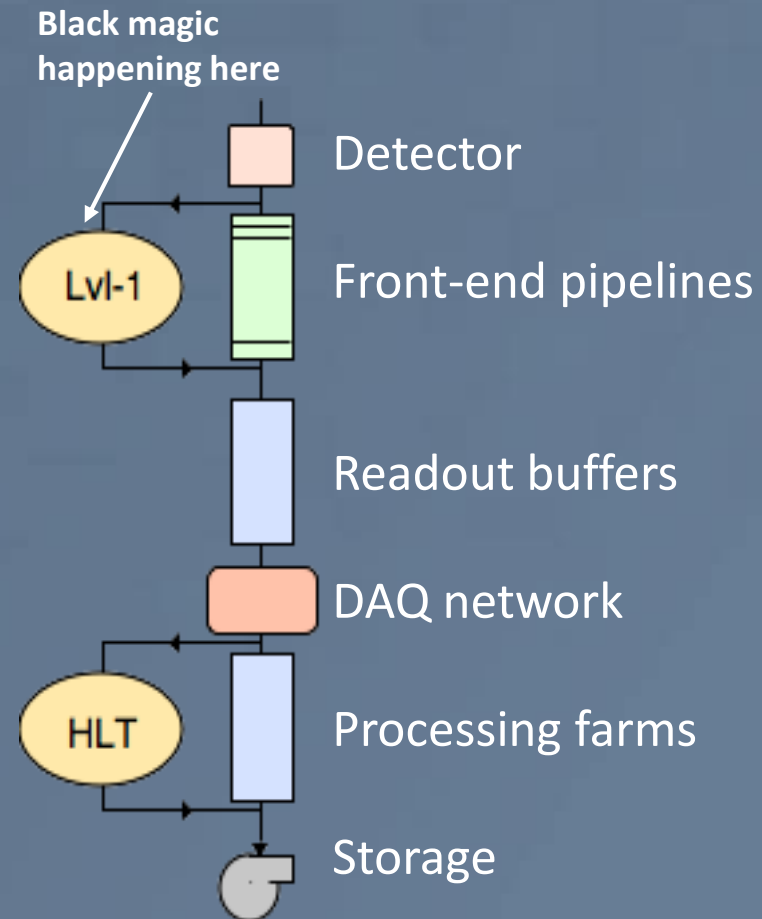
	Incoming data rate	Outgoing data rate	Reduction factor
Level1 Trigger (custom hardware)	40000000 s^{-1}	$10^5 - 10^6 \text{ s}^{-1}$	400-10,000
High Level Trigger (software on server farms)	$2000-1000000 \text{ s}^{-1}$	$1000 - 10000 \text{ s}^{-1}$	10-2000

Reducing the data step #1

The first level trigger

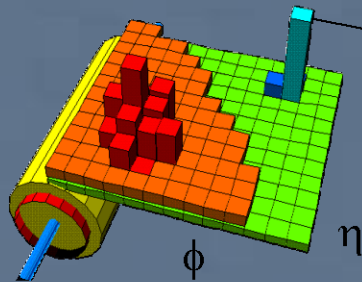
Trigger for LHC

- No (affordable) DAQ system could read out $O(10^7)$ channels at 40 MHz \rightarrow 400 TBit/s to read out – even assuming binary channels!
- What's worse: most of these millions of events per second are totally uninteresting: one Higgs event every 0.02 seconds
- A *first level trigger (Level-1, L1)* must somehow select the more interesting events and tell us which ones to deal with any further



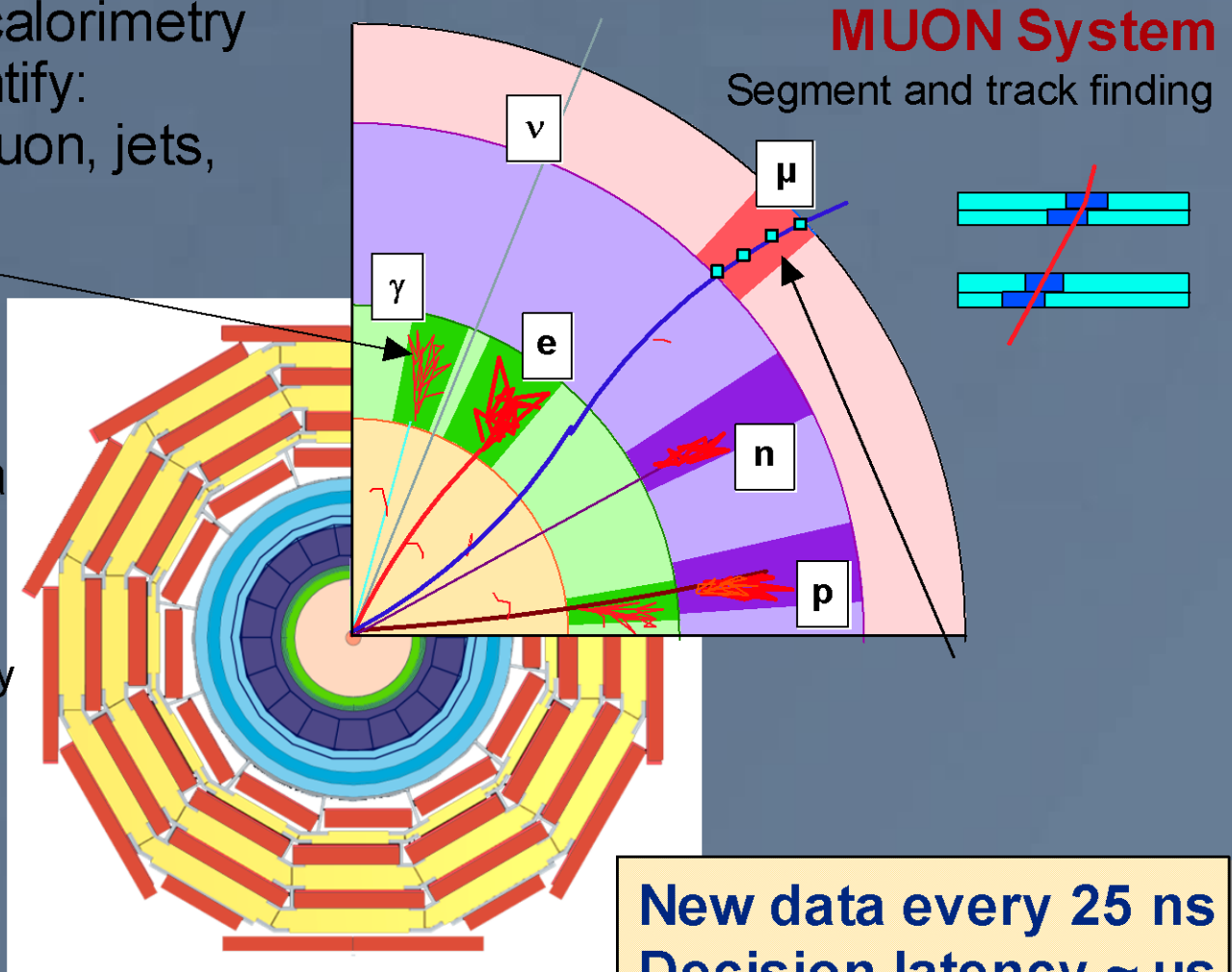
A small subset of the data to select events

Use prompt data (calorimetry and muons) to identify:
High p_t electron, muon, jets,
missing E_T



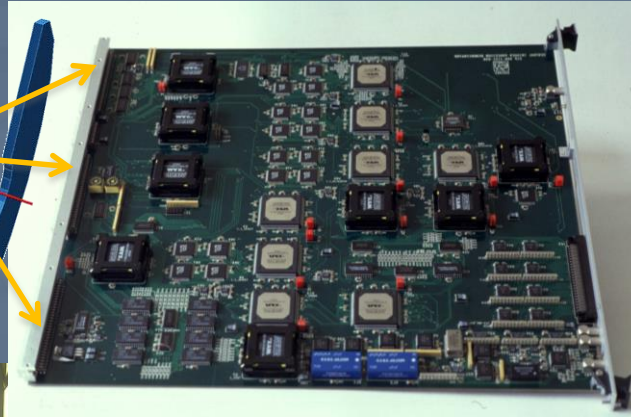
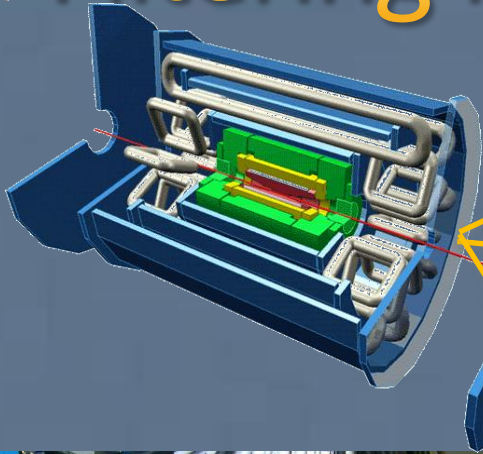
CALORIMETERS

Cluster finding and energy
deposition evaluation



New data every 25 ns
Decision latency $\sim \mu\text{s}$

Filtering in hardware



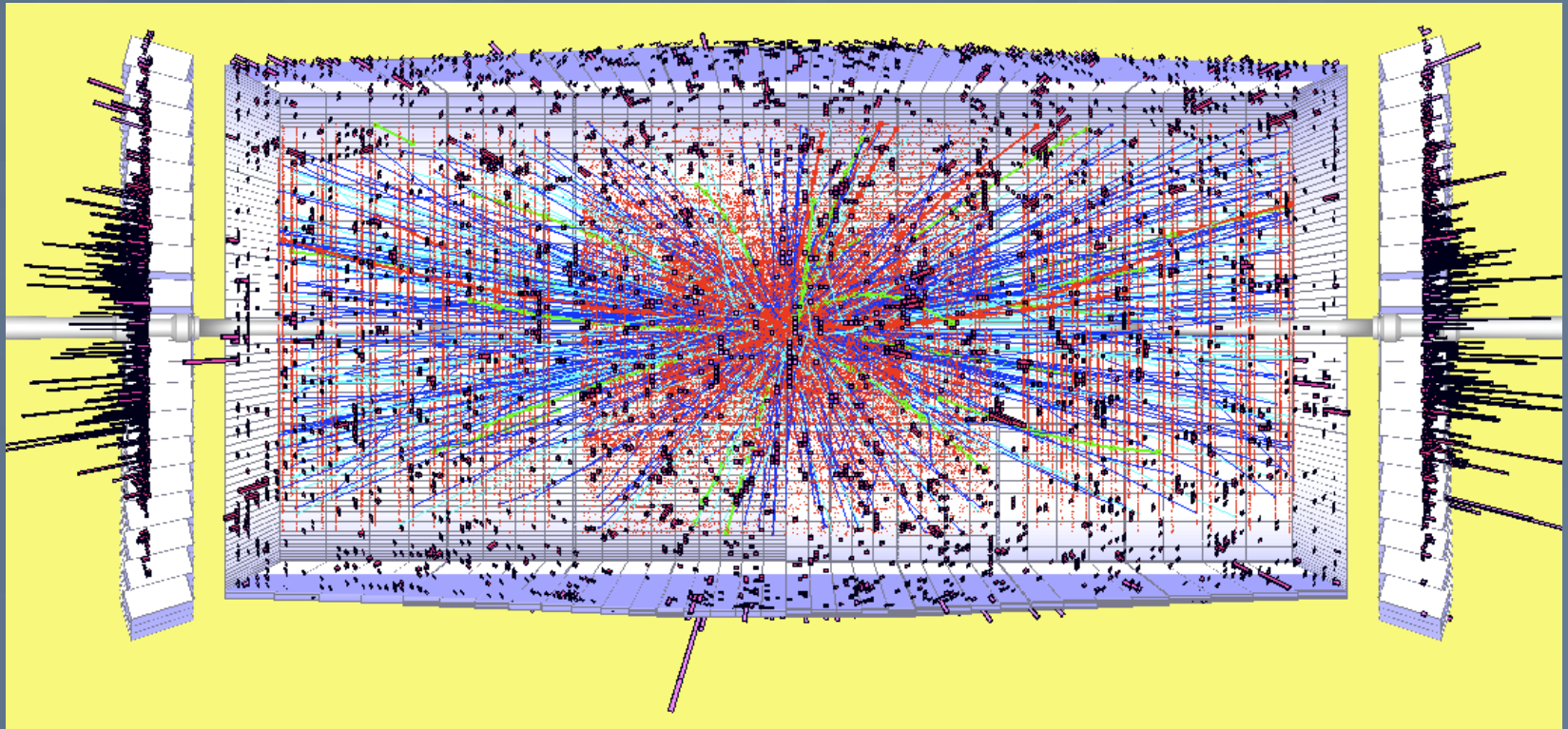
- Sophisticated electronics
- Hundreds of custom-built boards – process a small piece of the collision at enormous speeds (40 million times / second)
- They give a crude, but effective decision, based on simple criteria



Level 1 Trigger

- The Level 1 Triggers are implemented in hardware: FPGAs and ASICs → 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 signatures → working with partial information and with drastic simplifications has a price → interesting and valuable events are lost
- At higher rates we need more sophisticated systems and in particular full particle trajectories (momentum!)

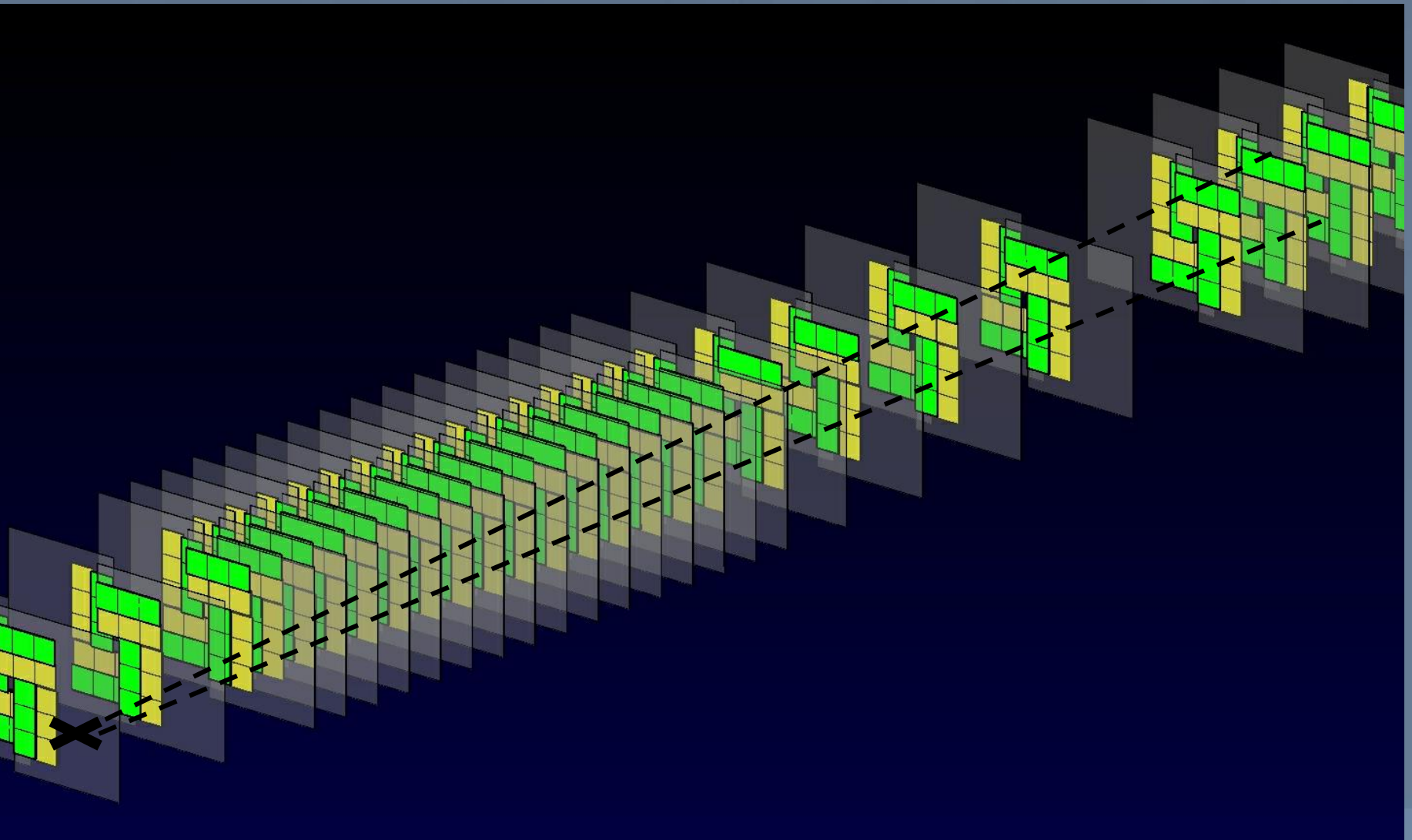
A Track-Trigger at 40 MHz for Run4



Goals:

- Find 1000s of particle trajectories in real-time (couple of micro-seconds)
- Improve sensitivity to interesting events

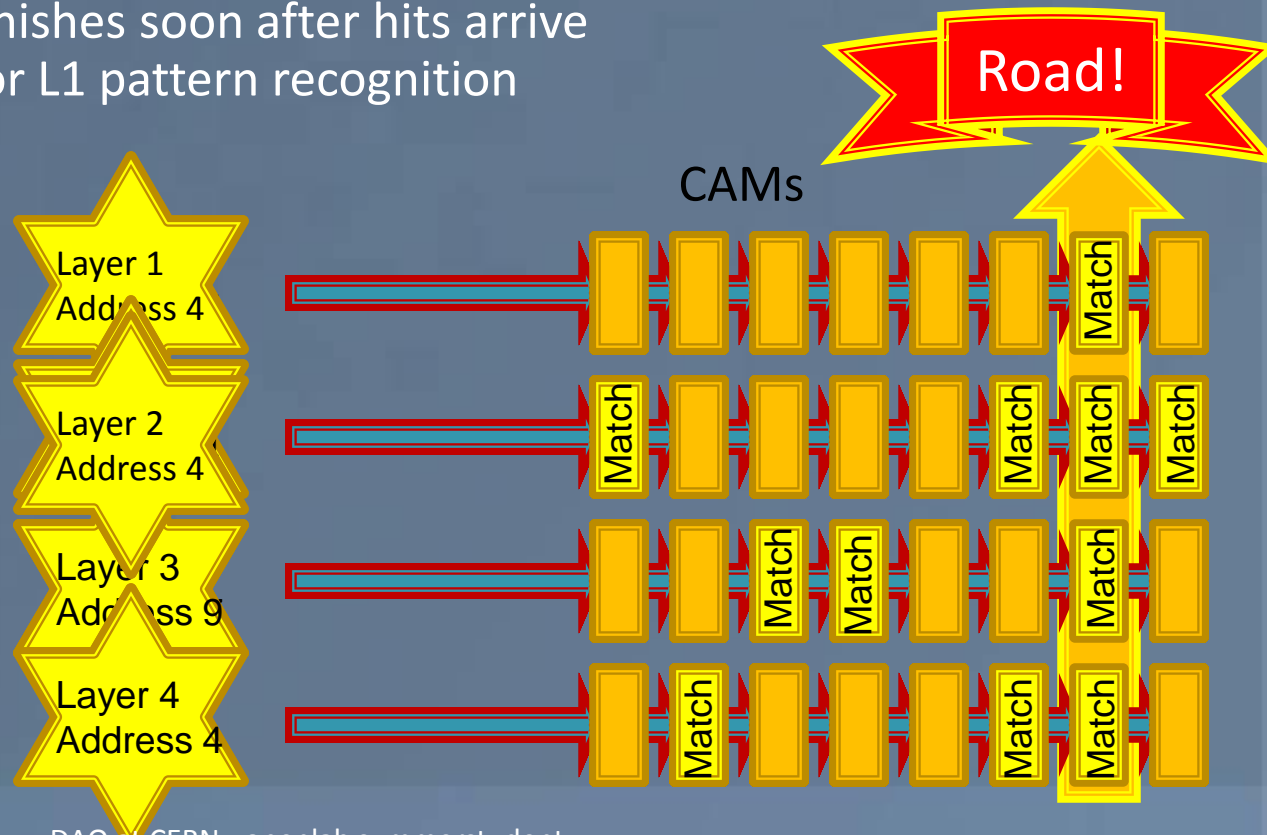
Pattern finding - tracks



Tracking Triggers: finding particle trajectories in hardware

Pattern Recognition Associative Memory (PRAM)

- Based on CAM cells to match and majority logic to associate hits in different detector layers to a set of pre-determined hit patterns
- Pattern recognition finishes soon after hits arrive
- Potential candidate for L1 pattern recognition
- However: Latency
- Challenges:
 - Increase pattern density by 2 orders of magnitude
 - Increase speed x 3
 - Same Power
 - Use 3D architecture: Vertically Integrated Pattern Recognition AM - VIPRAM



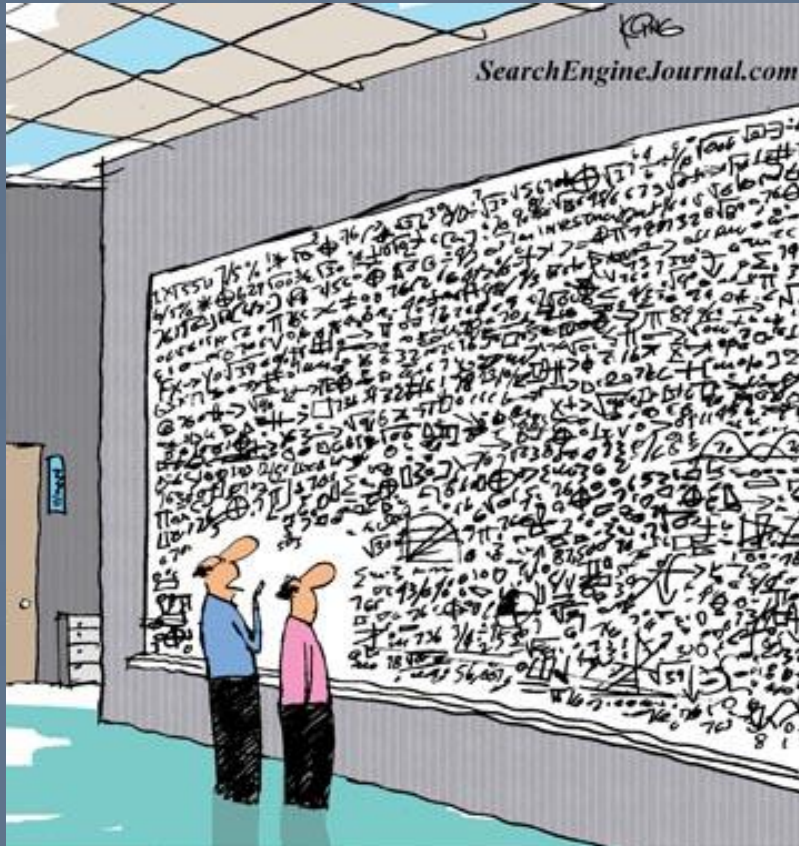
Level 1 challenge

- Can we do this **without custom hardware**?
- Maybe in GPGPUs / XeonPhis → studies ongoing in some (lower-rate) experiments
- We need low and – ideally – deterministic latency
- Need an efficient interface to detector-hardware: CPU/FPGA hybrid?
- Or forget about the whole L1 thing altogether and do everything in software → requires a lot of fast, low-power, radiation-hard low-cost links (remember the 600 TB/s)

Challenge #2

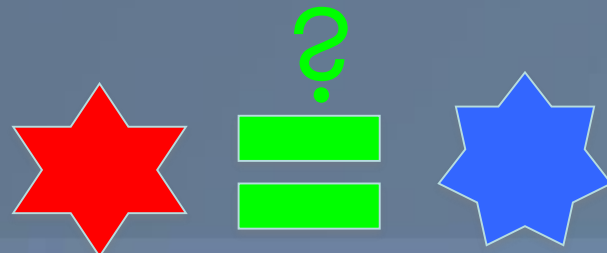
Data Acquisition and High Level Trigger

High Level Trigger



“And this, in simple terms, is how we find the Higgs Boson”

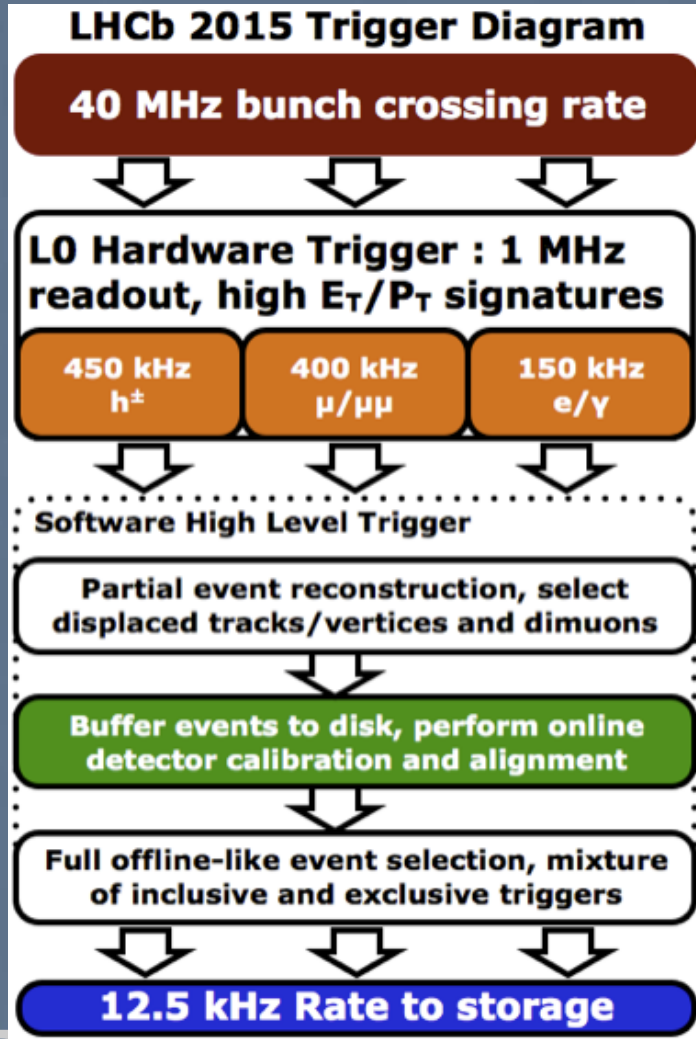
- Pack the knowledge of tens of thousands of physicists and decades of research into a huge sophisticated algorithm
- Several millions lines of code
- Takes (only!) a few 10 - 100 milliseconds *per collision*



High Level Trigger: Key Figures

- Existing code base: 5 MLOC of mostly C++
- Almost all algorithms are single-threaded (only few exceptions)
- Currently processing time on per event: 10 - 100 ms / process (hyper-thread)
- Currently between 100k and 1 million events per second are filtered online in each of the 4 experiments
- Uses all the physics knowledge available (if it's computationally affordable)

Things to be done in the HLT

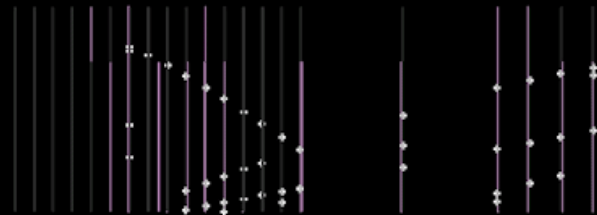
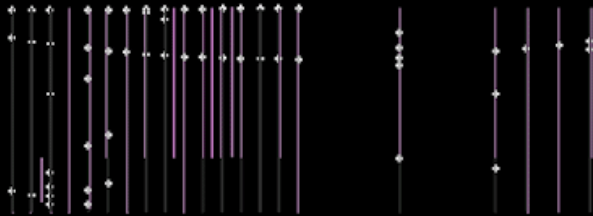


- Most importantly: reconstruct all charged particle trajectories
 - Find segments, connect them, re-fit to physical trajectory
- Associate the particles with the correct p-p interaction (multiple interactions in each crossing)
- Measure energy clusters (“jets”)
- Calculate decay chains and global event-properties
 - Total energy, missing energy, missing momentum, number of charged particles, etc...

Track finding & fitting

VELO RZ

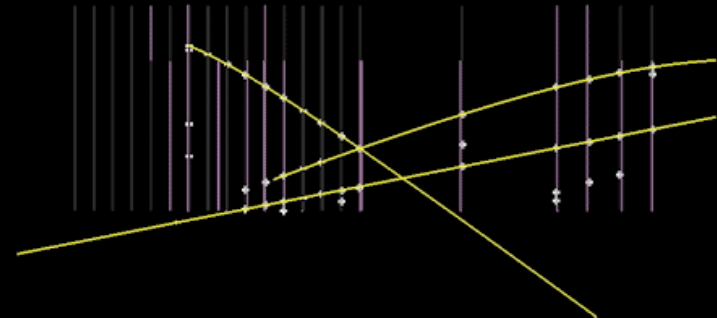
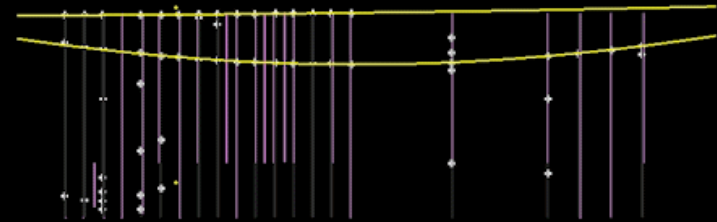
7.5.2009 4:15:49
Run 50436 Event 473



LHCb
LHCb

VELO RZ

7.5.2009 4:15:49
Run 50436 Event 473



LHCb
LHCb

- Can be much more complicated: lots of tracks / rings, curved / spiral trajectories, spurious measurements and various other imperfections

Reconstruction

Small part of a collision

```
0x01e84ce0: 0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84cf0: 0x01e8 0x87ec 0x01e8 0x85d8 0x7363 0x616e 0x0000 0x0000
0x01e84d00: 0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d10: 0x01e8 0x87e8 0x01e8 0x8618 0x7365 0x7400 0x0000 0x0000
0x01e84d20: 0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d30: 0x01e8 0x87a8 0x01e8 0x8658 0x7370 0x6c69 0x7400 0x0000
0x01e84d40: 0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d50: 0x01e8 0x8854 0x01e8 0x8698 0x7374 0x7269 0x6e67 0x0000
0x01e84d60: 0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d70: 0x01e8 0x875c 0x01e8 0x86d8 0x7375 0x6273 0x7400 0x0000
0x01e84d80: 0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d90: 0x01e8 0x87c0 0x01e8 0x8718 0x7377 0x6974 0x6368 0x0000
```

Address = which detector/sensor has read these data?

Data = what was detected?

Reconstruction = Conversion of electronic signals into physical „objects“

Track: $\phi = 0.23$, $\eta = 0.75$, $p_T = 2.3$ GeV/c

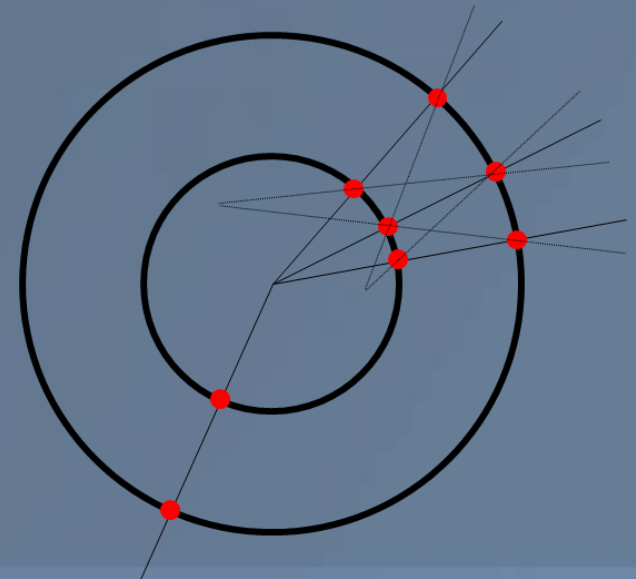
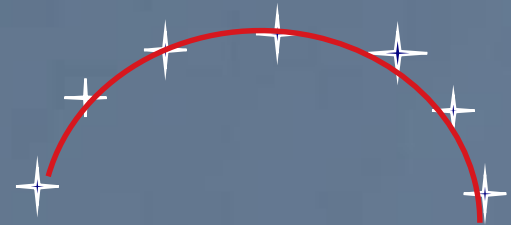
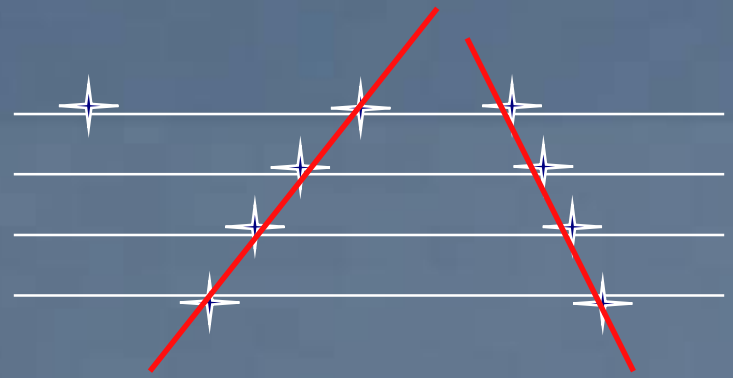
Probability to be a Pion: 70%, Kaon: 28%, Proton: 2%

Reconstruction steps

- Abstraction of electronic signals from the detector
 - Detector-element 1244 has measured a signal of 120 at time 1333096259.344245
 - 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)
- Particles frequently create signal in adjacent detector-elements
These signals are combined into “clusters”
- The clusters are combined into “tracks” (pattern-recognition)
- Tracks are combined to reconstruct collision points (“primary vertices” == where an original beam-proton has hit another beam-proton) and decay (“secondary” and “tertiary”)) vertices

Pattern recognition

- Recognition of tracks
 - Straight in the direction of the magnetic field
 - Deviation on a curved trajectory in a normal plane around the magnetic field
→ radius of curvature allows to determine the momentum
- Primary vertex
- Secondary vertices
 - Allows detection of particle decays



Pattern-recognition(2)



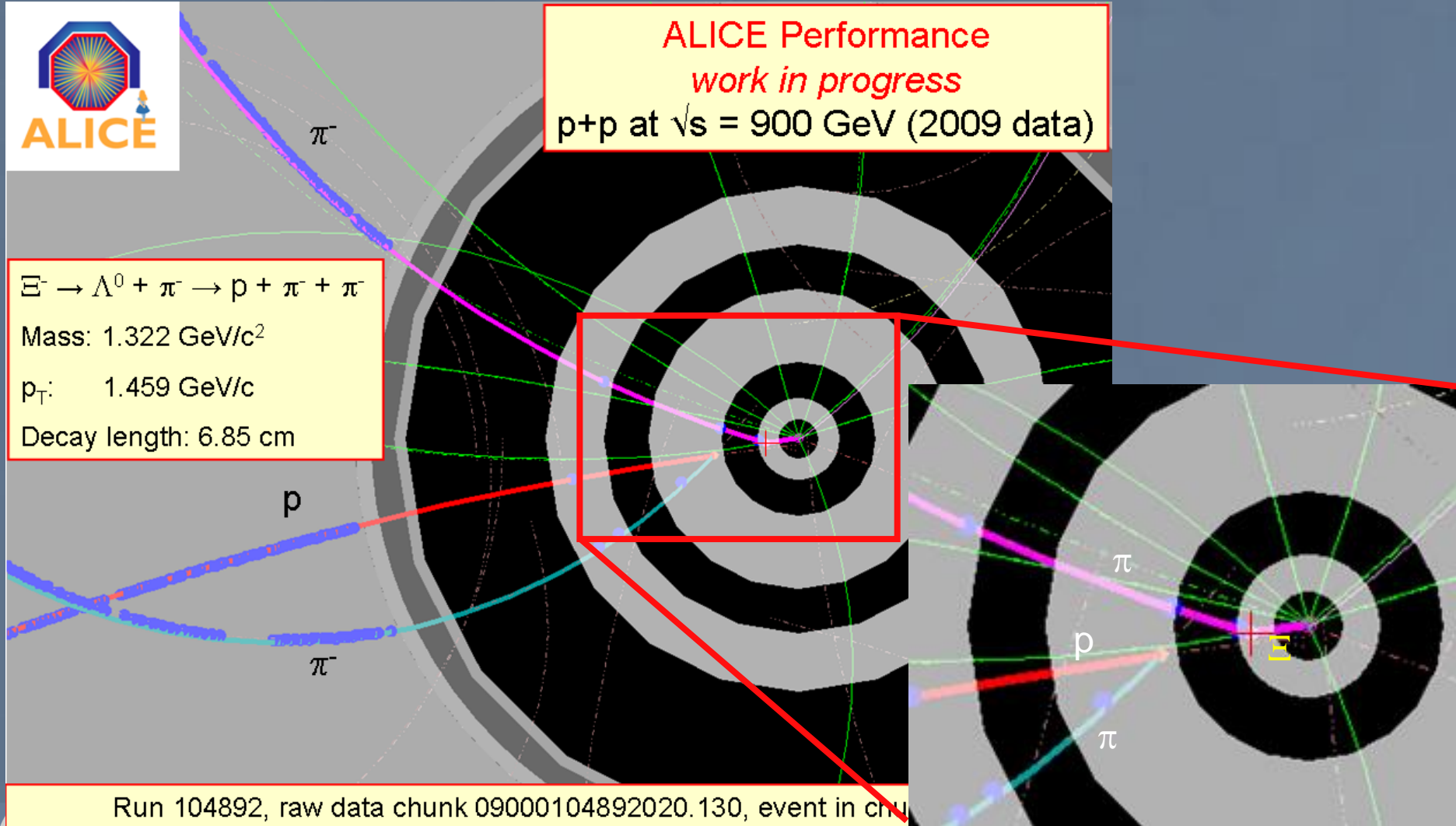
ALICE Performance
work in progress
p+p at $\sqrt{s} = 900$ GeV (2009 data)

$\Xi^- \rightarrow \Lambda^0 + \pi^- \rightarrow p + \pi^- + \pi^-$

Mass: $1.322 \text{ GeV}/c^2$

p_T : $1.459 \text{ GeV}/c$

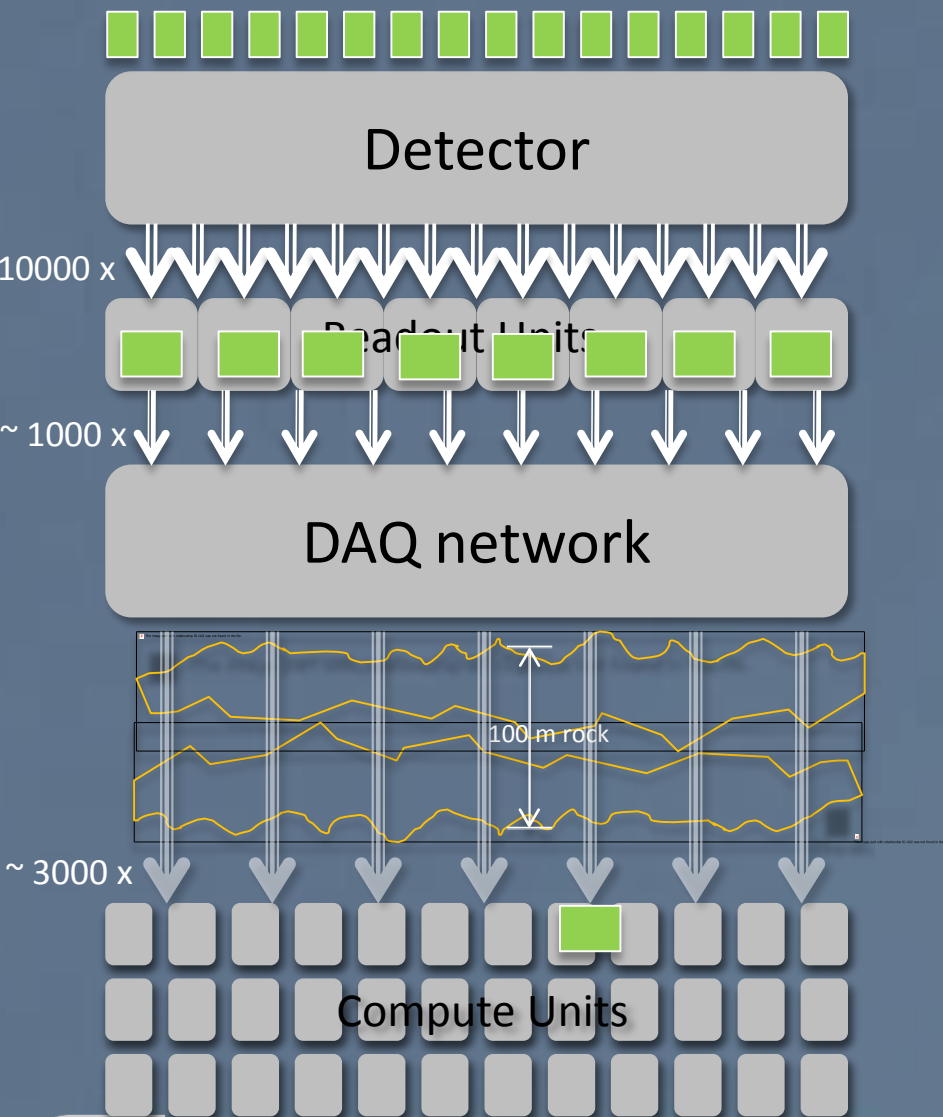
Decay length: 6.85 cm



Run 104892, raw data chunk 09000104892020.130, event in chunk

Data Acquisition & event-building – Or: How do we get the data to the HLT?

Data Acquisition (generic example)



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)

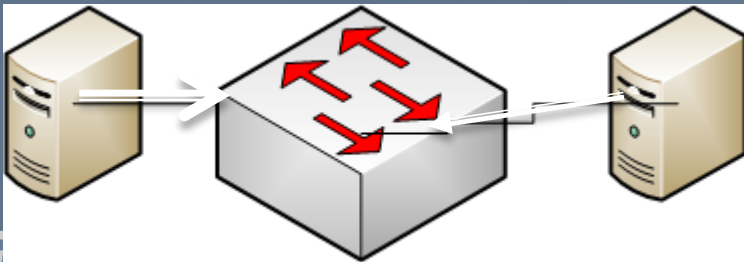
↓ GBT: custom radiation- hard link from the detector 3.2 Gbit/s

↓ DAQ (“event-building”) links – some LAN (Ethernet / InfiniBand / OmniPath)

↓ Links into compute-units: typically 10 Gbit/s (because filtering is currently compute-limited)

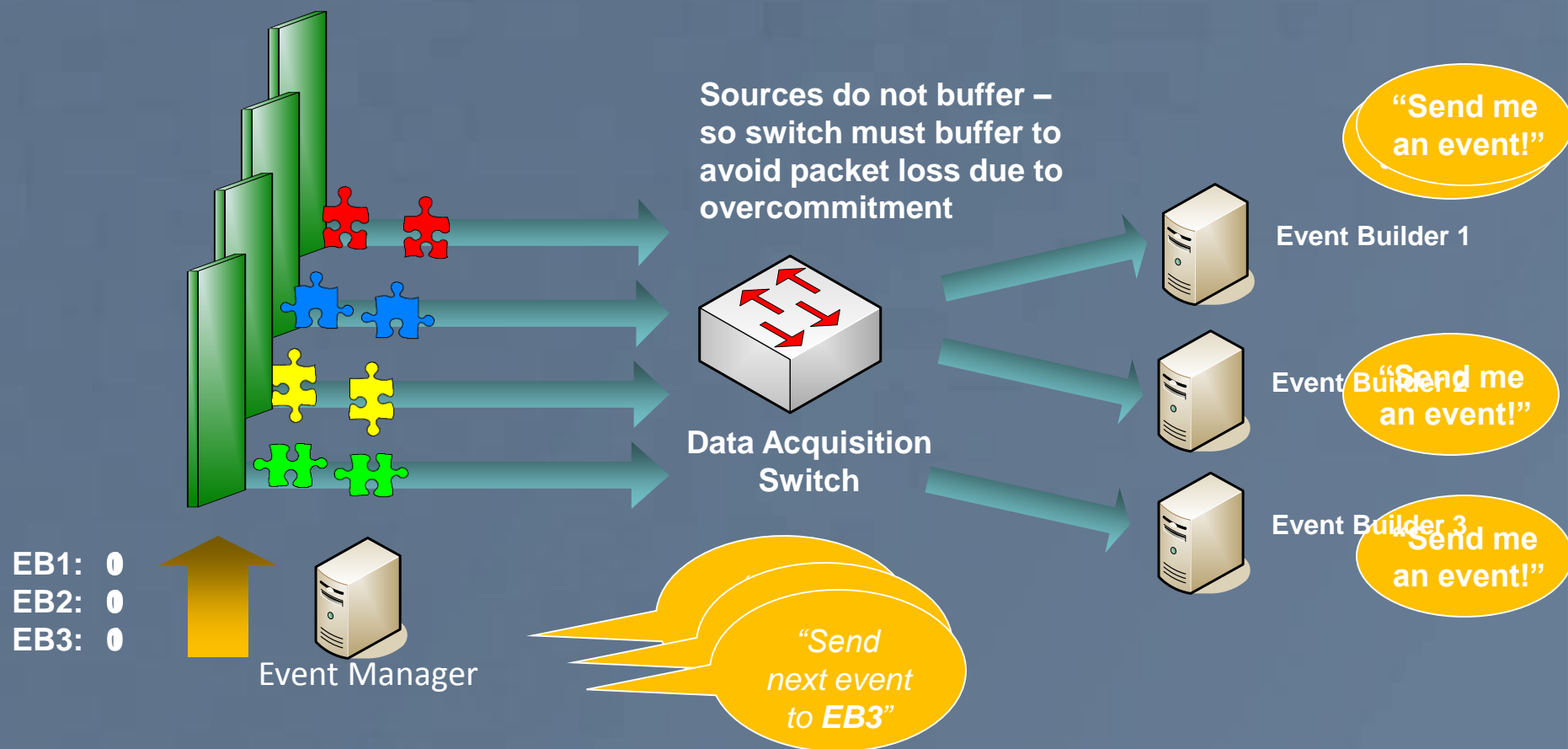
One network to rule them all?

- Ethernet, IEEE 802.3xx, has almost become synonymous with Local Area Networking
- Ethernet has many nice features: cheap, simple, cheap, etc...
- Ethernet does not:
 - guarantee delivery of messages
 - allow multiple network paths
 - provide quality of service or bandwidth assignment (albeit to a varying degree this is provided by many switches)
- Because of this raw Ethernet is rarely used, usually it serves as a transport medium for IP, UDP, TCP etc...



- Flow-control in standard Ethernet is only defined between immediate neighbors
- Sending station is free to throw away "x-offed" frames (and often does ☹)

Push-Based Event Building with store& forward switching and load-balancing

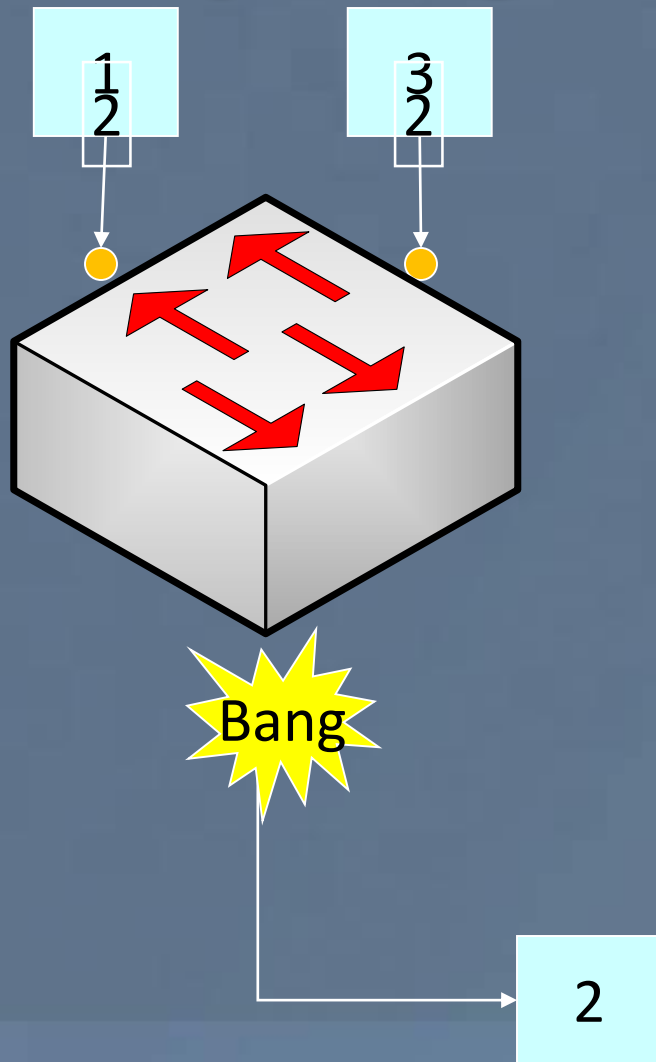


1 Event Builders notify Event Manager available capacity

2 Event Manager ensures that data are sent only to nodes with available capacity

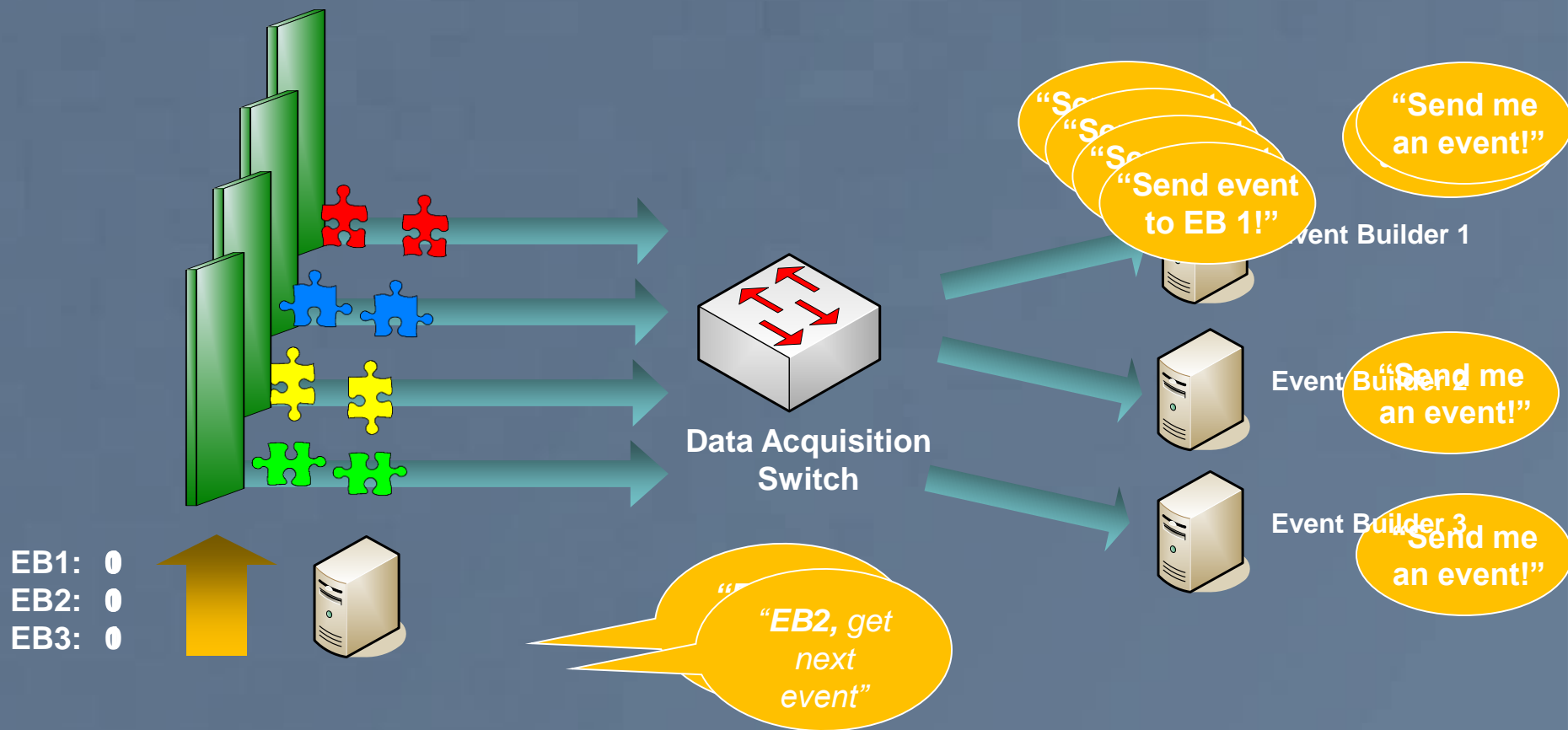
3 Readout system relies on feedback from Event Builders

Problem with push-based event building: Congestion



- "Bang" translates into random, uncontrolled packet-loss
- In Ethernet this is perfectly valid behavior and implemented by many (cheaper) devices
- Higher Level protocols are supposed to handle the packet loss due to *lack of buffering*
- This problem comes from *synchronized sources sending to the same destination at the same time*

Pull-Based Event Building



1 Event Builders notify Event Manager of available capacity

2 Event Manager elects event-builder node

3 Readout traffic is driven by Event Builders

Concrete example: CMS 2012 L-1 Trigger & DAQ

- Overall Trigger & DAQ Architecture: 2 Levels:
- Level-1 Trigger:
 - 25 ns input
 - 3.2 μ s latency

UXC↑

Interaction rate: 1 GHz

Bunch Crossing rate: 40 MHz

Level 1 Output: 100 kHz

Output to Storage: 400 Hz

Average Event Size: 1 MB

Data production 1 TB/day

Trigger/DAQ parameters

No.Levels

Level-0,1,2

Event

Readout

HLT Out

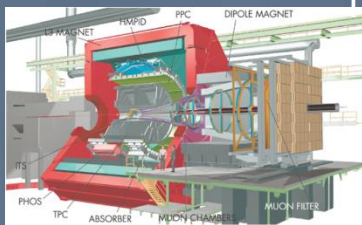
Trigger

Rate (Hz)

Size (Byte)

Bandw.(GB/s)

MB/s (Event/s)



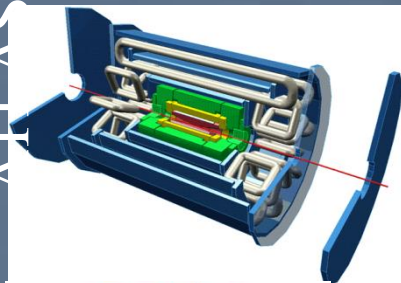
4

Pb-Pb **500**
p-p **10^3**

5×10^7
 2×10^6

25

1250 (10^2)
200 (10^2)



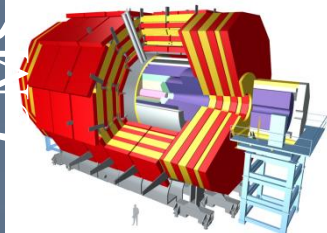
3

LV-1 **10^5**
LV-2 **3×10^3**

1.5×10^6

4.5

300 (2×10^2)



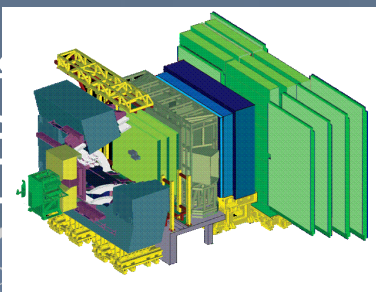
2

LV-1 **10^5**

10^6

100

~ 1000 (10^2)



2

LV-0 **10^6**

5×10^4

50

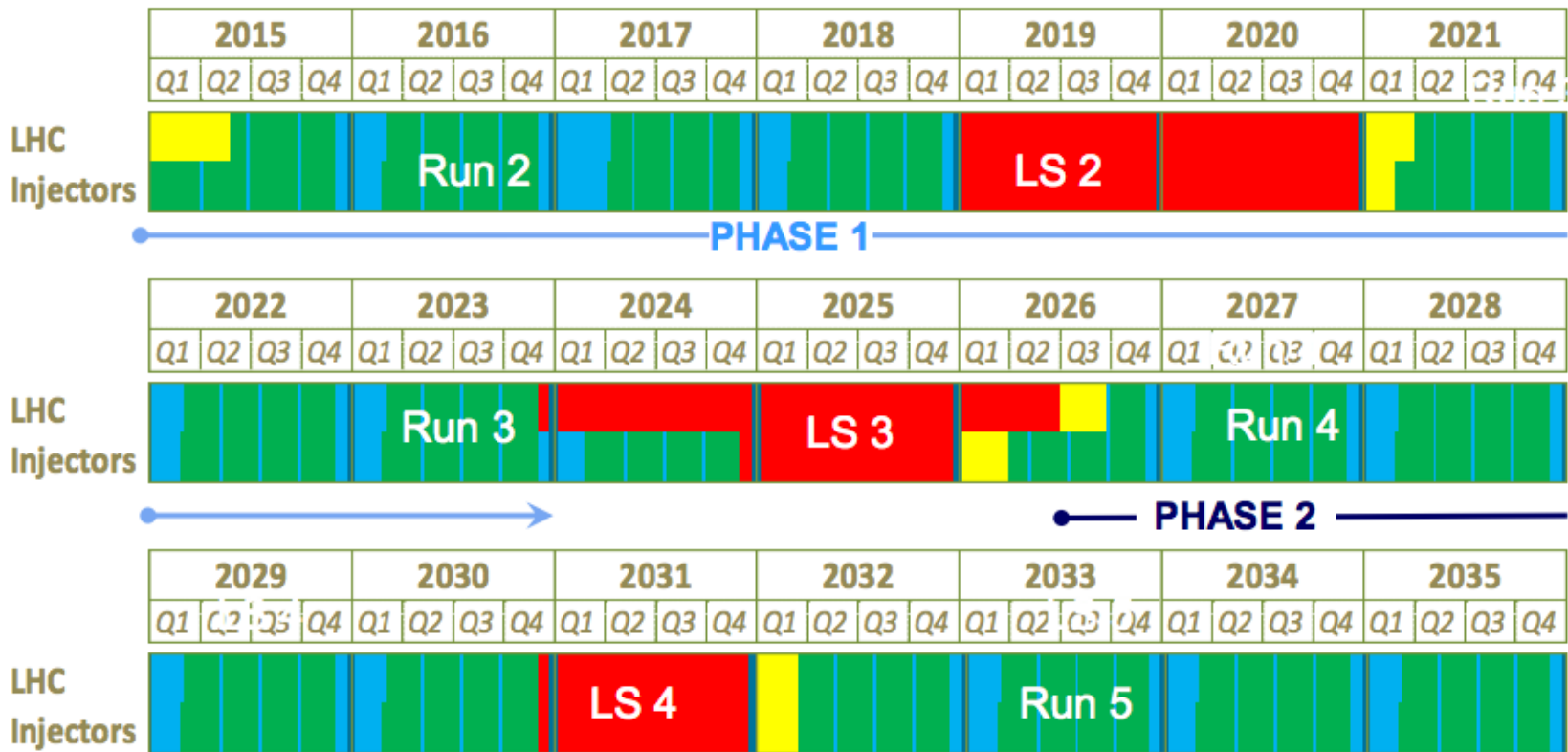
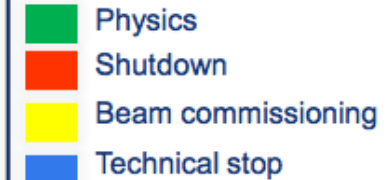
200 (4×10^3)



The LHC – long-term planning

LHC roadmap: according to MTP 2016-2020 V2

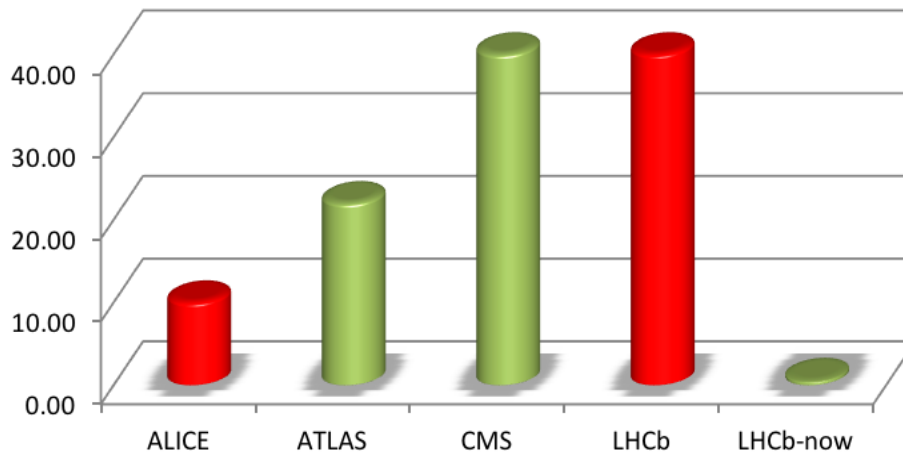
LS2 starting in 2019 => 24 months + 3 months BC
 LS3 LHC: starting in 2024 => 30 months + 3 months BC
 Injectors: in 2025 => 13 months + 3 months BC



Run3 upgrade

- Filter farm will need to handle:
 - **Event size** (~130 kB) (@ 30 MHz)
 - **Larger event rate:** 30MHz (+ 10 MHz empty crossings)
- New challenges for DAQ & High-Level Trigger

Data Network - Throughput



Future DAQs in numbers

	Event-size [kB]	Rate of events into HLT [kHz]	HLT bandwidth [Gb/s]	Year [CE]
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	4000	1000	32000	2022
LHCb	100	40000	32000	2019

40000 kHz == collision rate

→ *LHCb abandons Level 1 for an all-software trigger*

Design principles

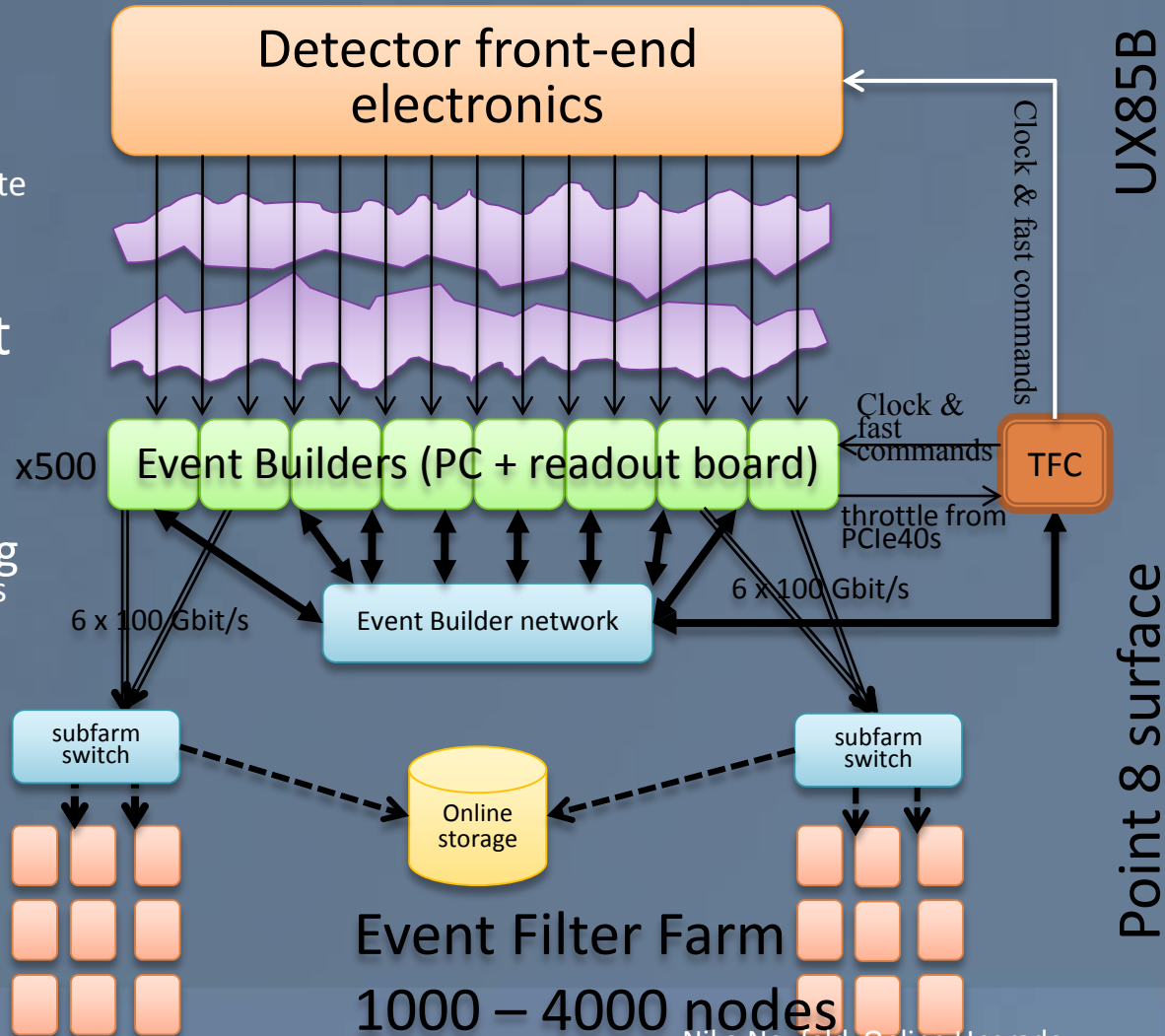
- Minimize number of expensive “core” network ports
- Use the most efficient technology for a given connection
 - different technologies should be able to co-exist (e.g. fast for building, slow for end-node)
 - keep distances short
- Exploit the economy of scale → try to do what everybody does (but smarter 😊)

DAQ challenge for Run3

- Transport multiple Terabit/s reliably and cost-effectively
- Integrate the network closely and efficiently with compute resources (be they classical CPU or “many-core”)
- Multiple network technologies should seamlessly co-exist in the same integrated fabric (“the right link for the right task”)

Run3 Online System

- Dimensioning the system:
 - ~10000 versatile links
 - ~500 readout nodes
 - ~40 MHz event-building rate
 - ~130 kB event size
- High bisection bandwidth in event builder network
 - ~40 Tb/s aggregate bandwidth
 - Use industry leading 100 Gbit/s LAN technologies
- Global configuration and control via ECS subsystem
- Global synchronization via TFC subsystem



Event Filter Farm
1000 – 4000 nodes

Summary

- The LHC experiments need to reduce 1 Pbit/s to ~ 25 PB/ year
- This is achieved with massive use of FPGAs, custom ASICs and x86 computing power
- Large, deep-buffer, local area networks are used to distribute data among the individual x86 servers
- The future will see massive increase of required *programmable* computing power and required networking bandwidth, much more data will be moved off detector