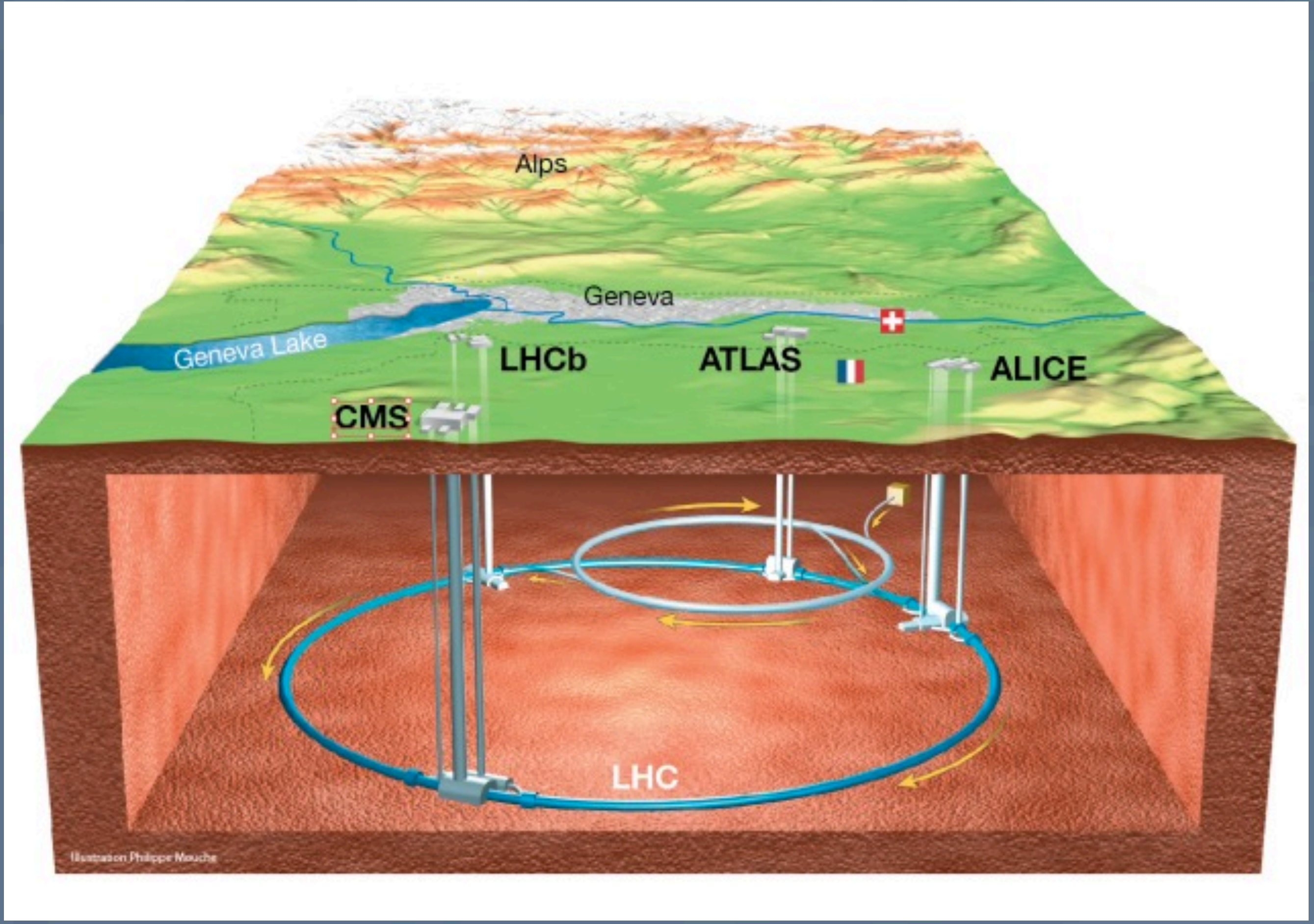
Big Data and rare events: The boson in the hay-stack

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

The Large Hadron Collider



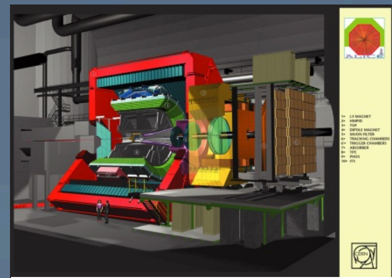
- 27 km
- Vacuum at 10^{-13} atm
- More than 9600 magnets
- Dipole magnets at -271.3 C → 0.8 C colder than outer space
- Energy in the beam corresponds to a TGV at 150 km/h
- Cost: 5 billion CHF
- 4 large experiments

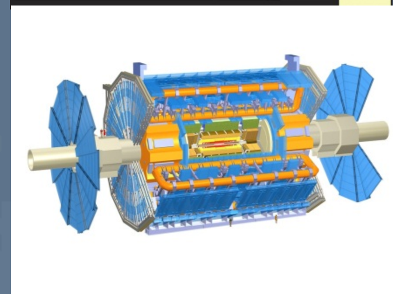
So what do we do with all that?

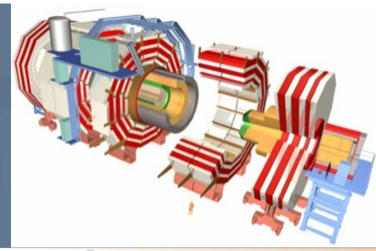
CERN

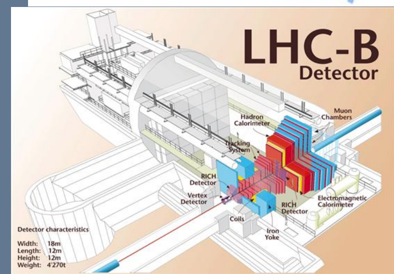
The LHC Experiments today

- 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: 4 6m 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 … first 2013/14 (Long Shutdown 1)



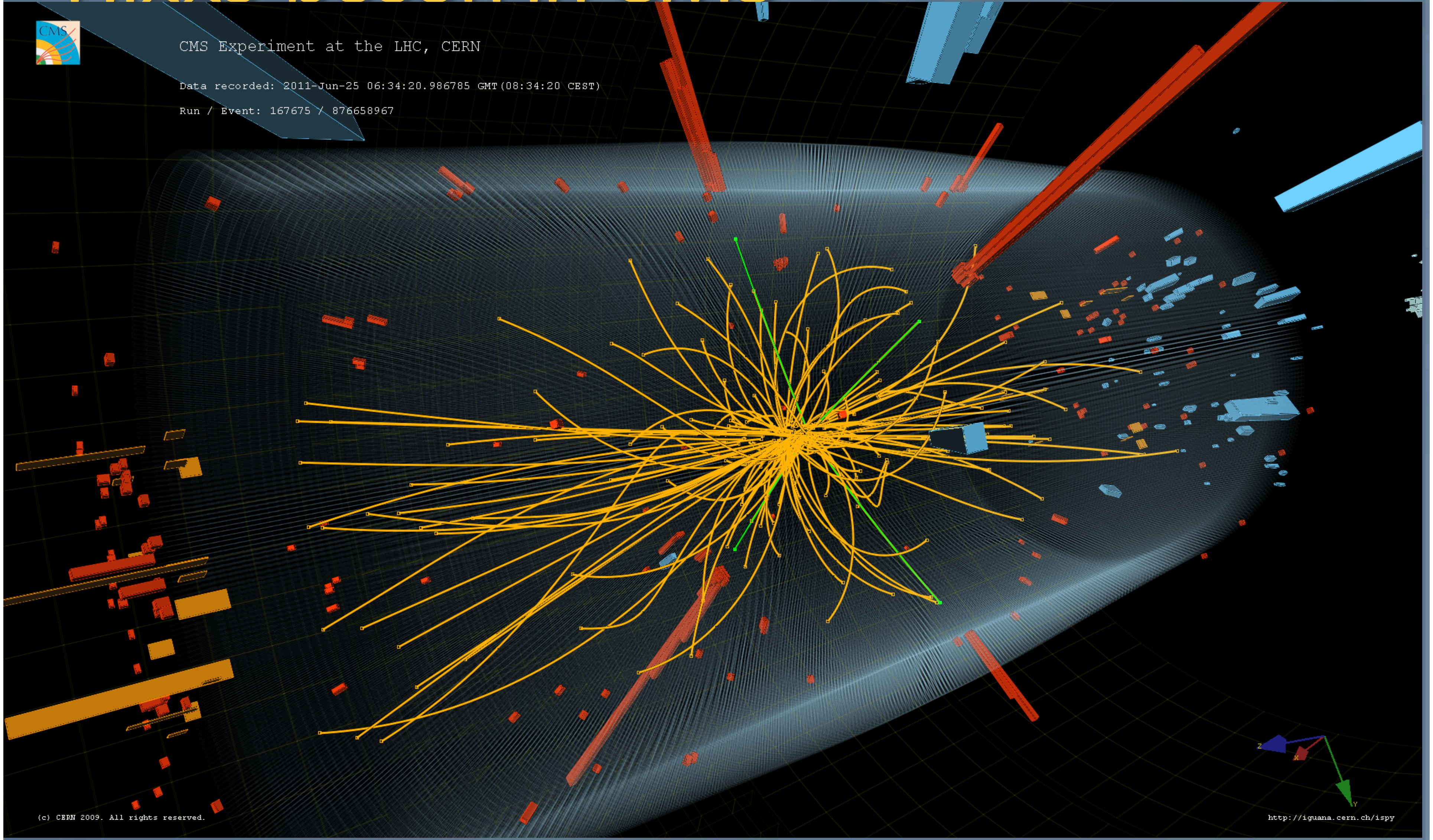




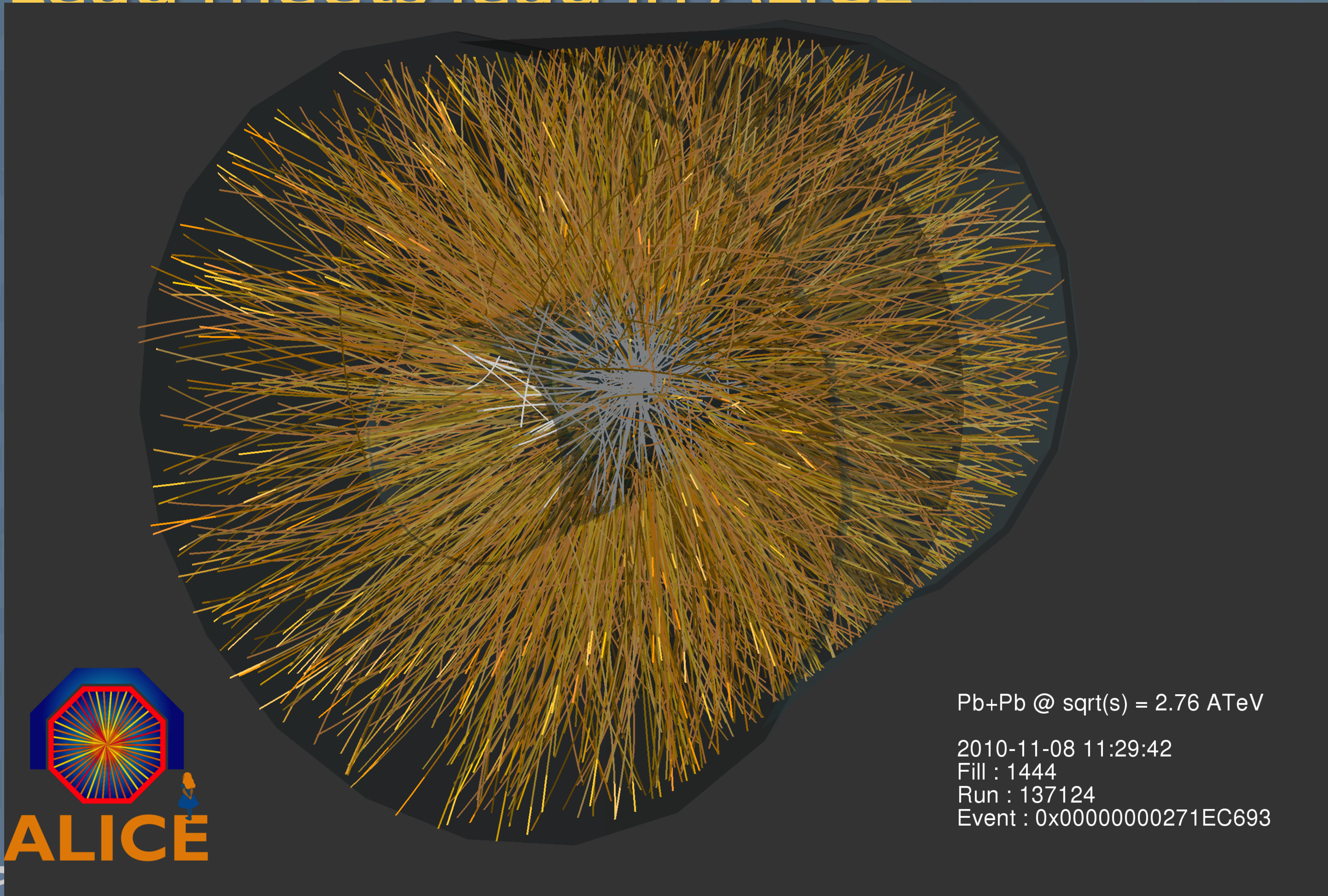


1 CHF ~ 1 USD

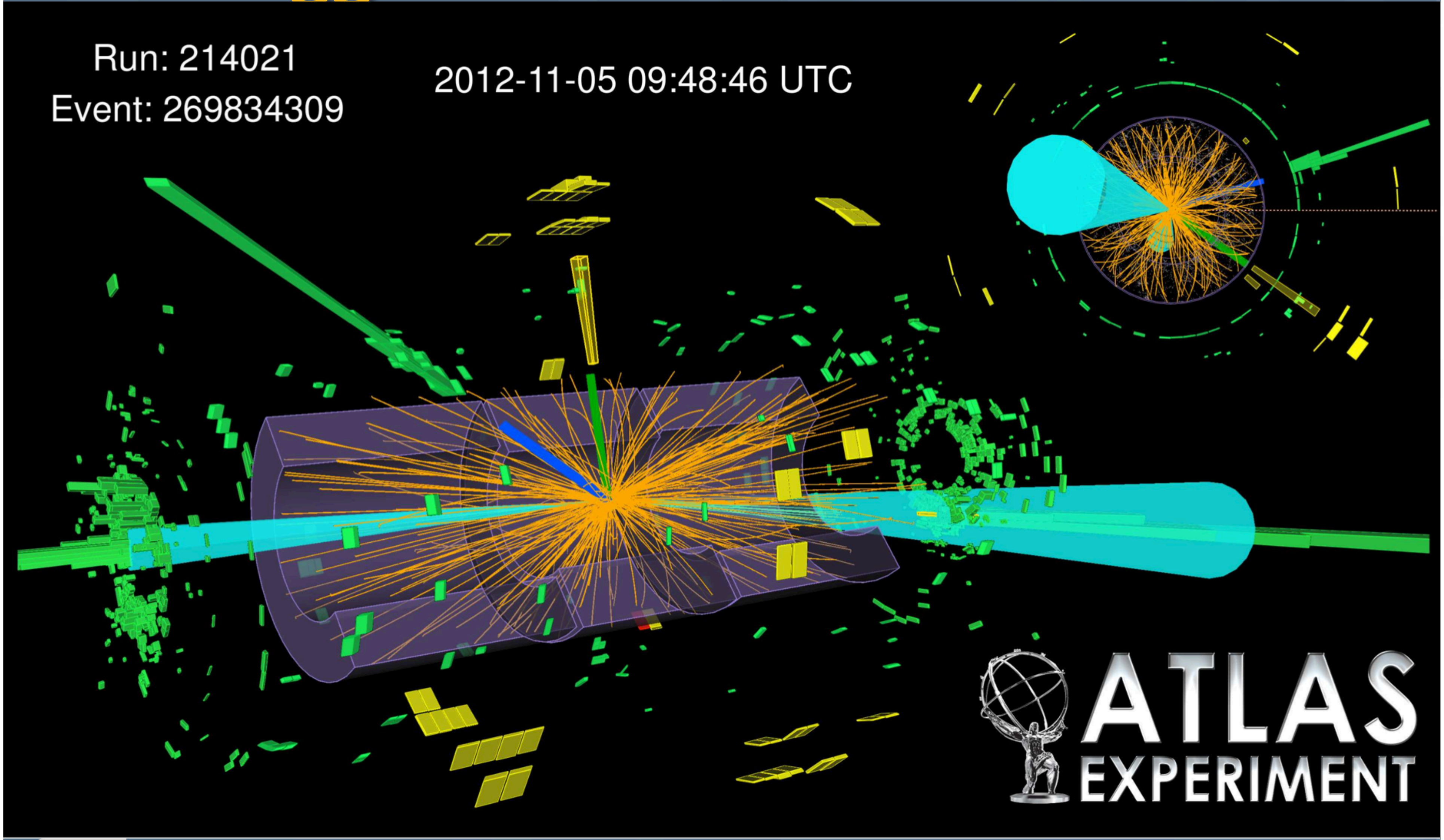
Higgs-boson in CMS



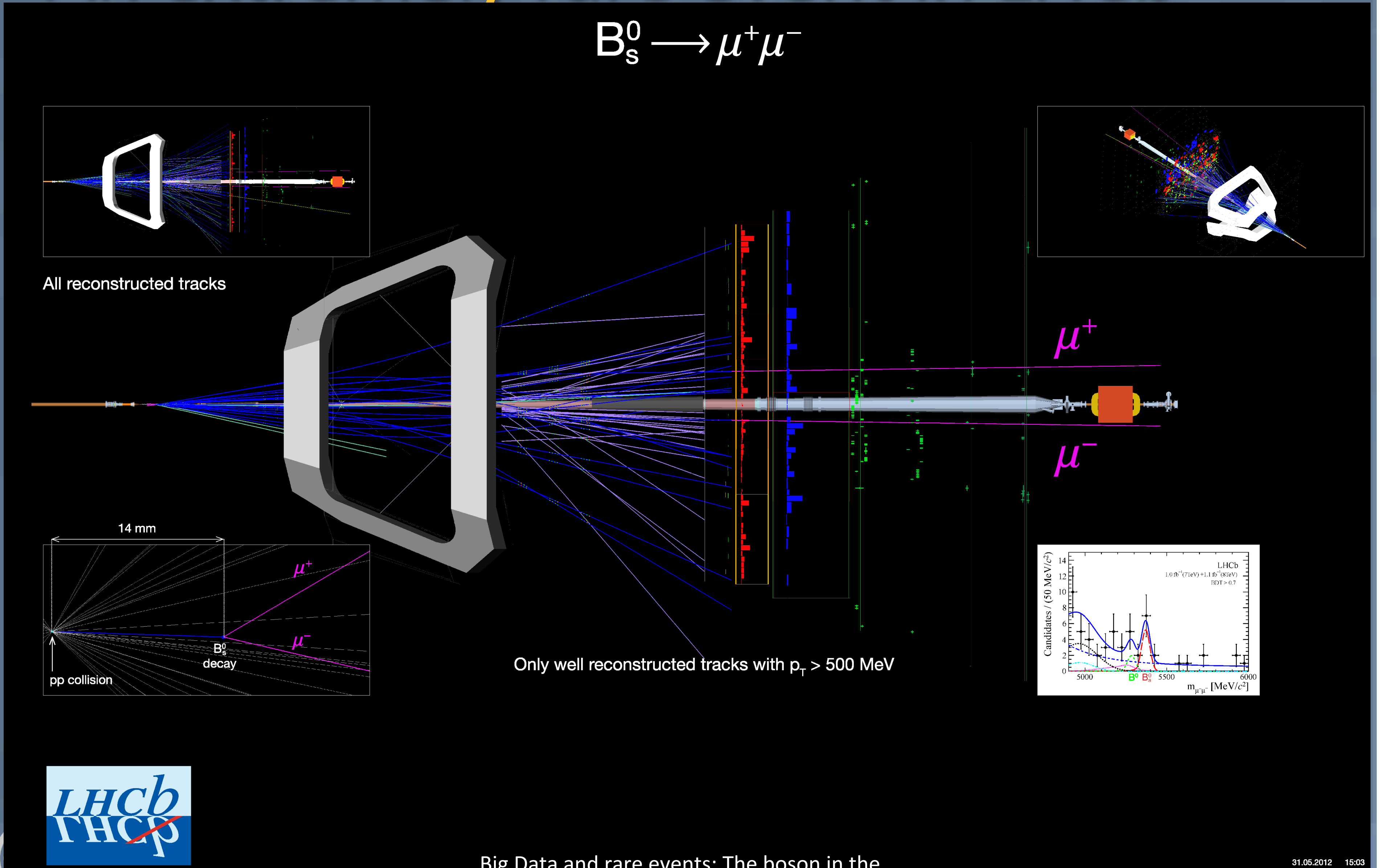
Lead meets lead in ALICE



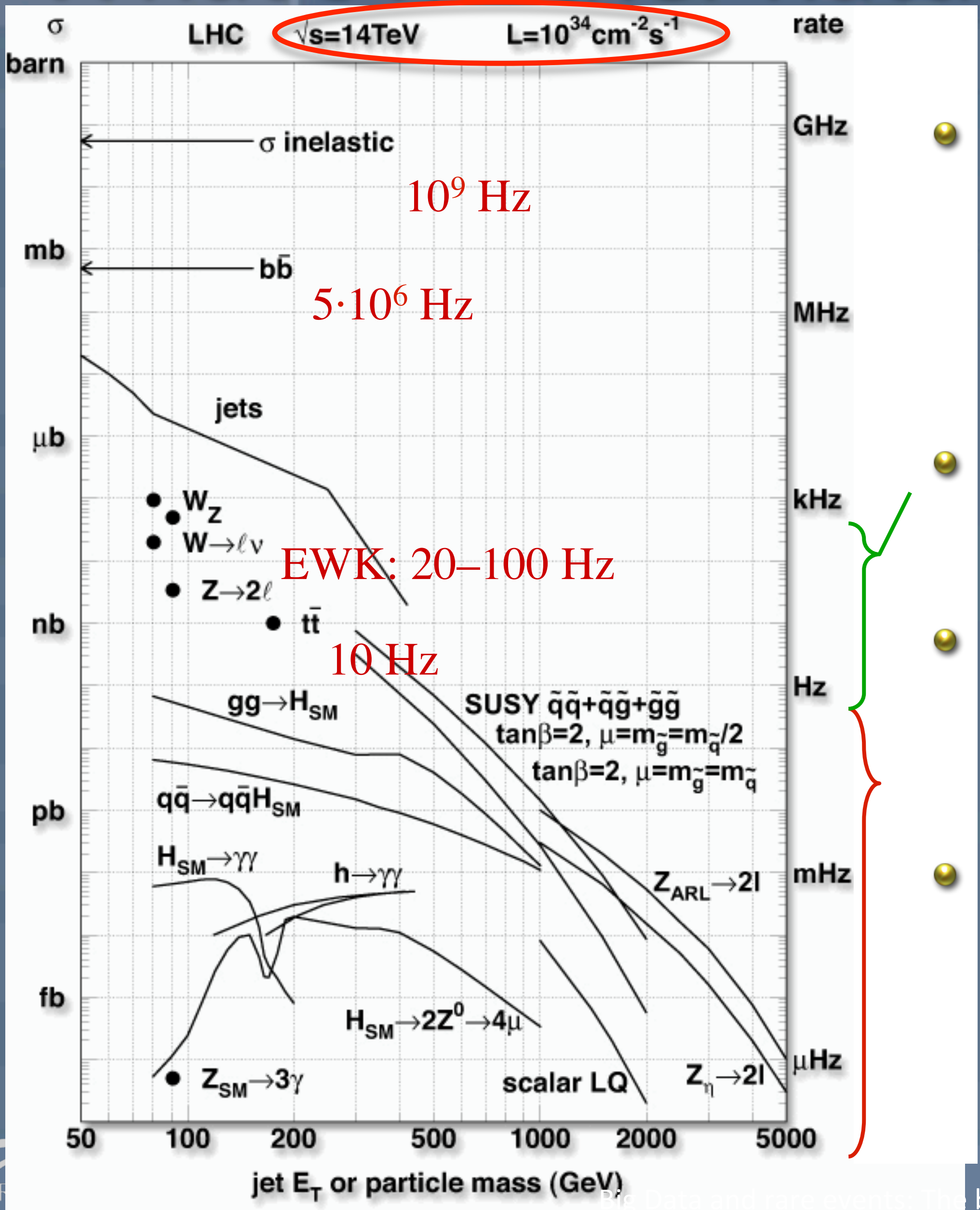
Mr. Higgs'es boson is also in ATLAS



An extremely rare event in LHCb



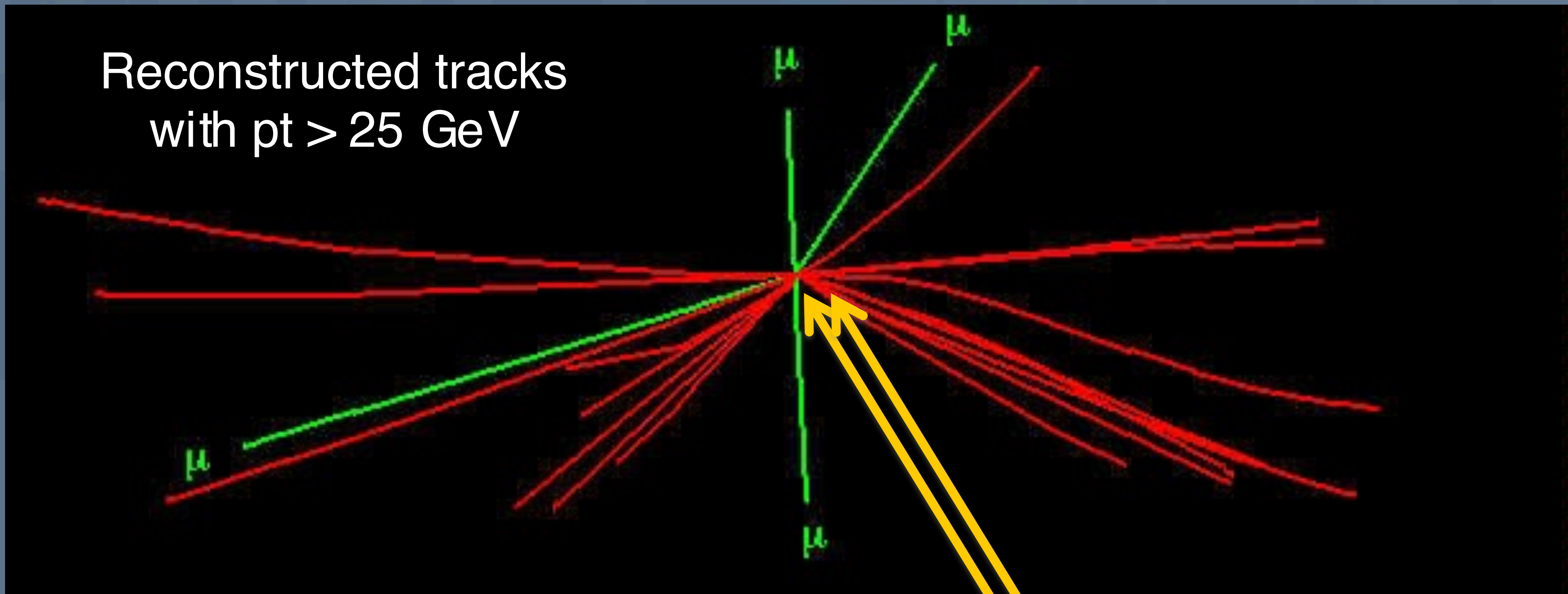
What's mother nature's menu?



- 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 $\sigma_{tot:}$: one in a billion or even mor rare
- Need to efficiently identify these rare processes from the overwhelming background <u>before</u> reading out & storing the whole event

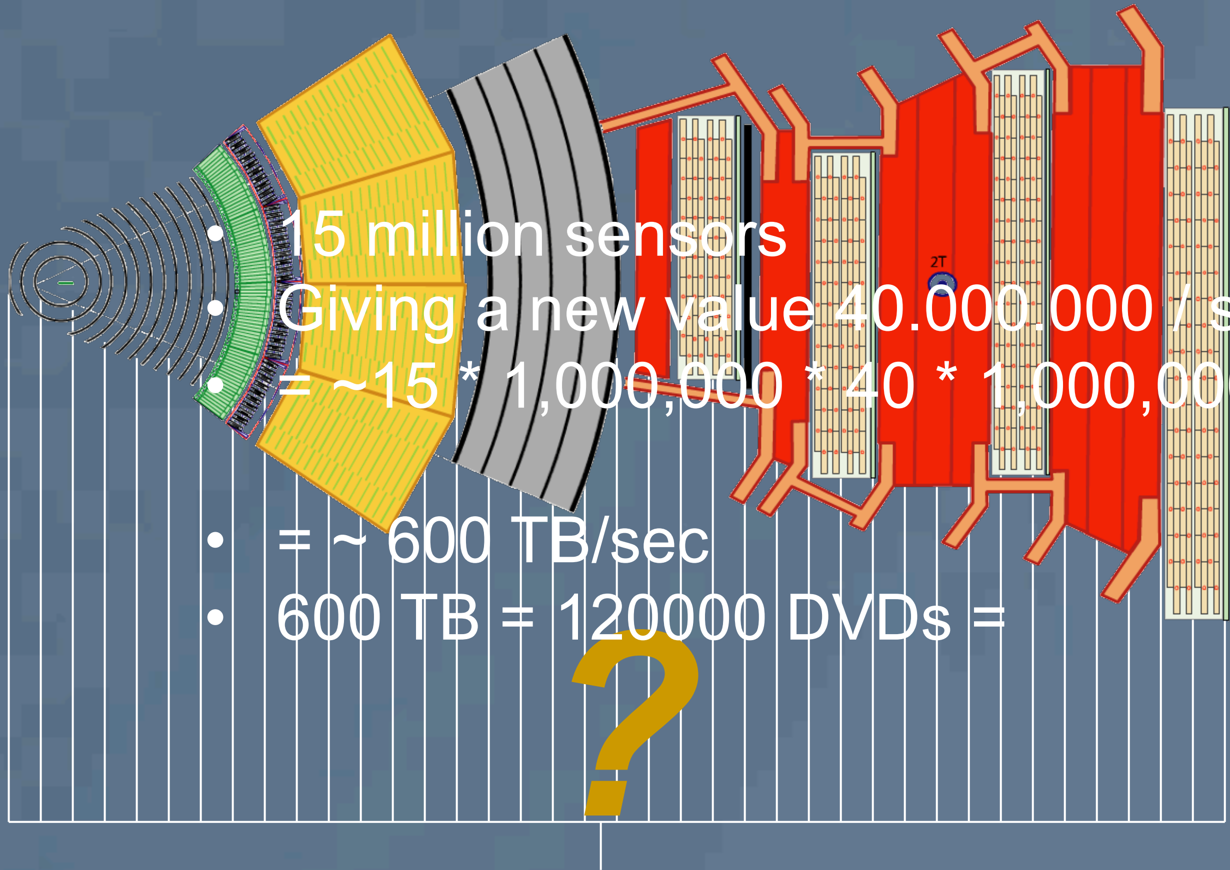
The needle in the hay-stack

Simulation from CMS



This is what we are looking for: 1 Higgs boson / find: 1000 particles
We hope to get this 20 times per **day** / get 40.000.000 times per second!

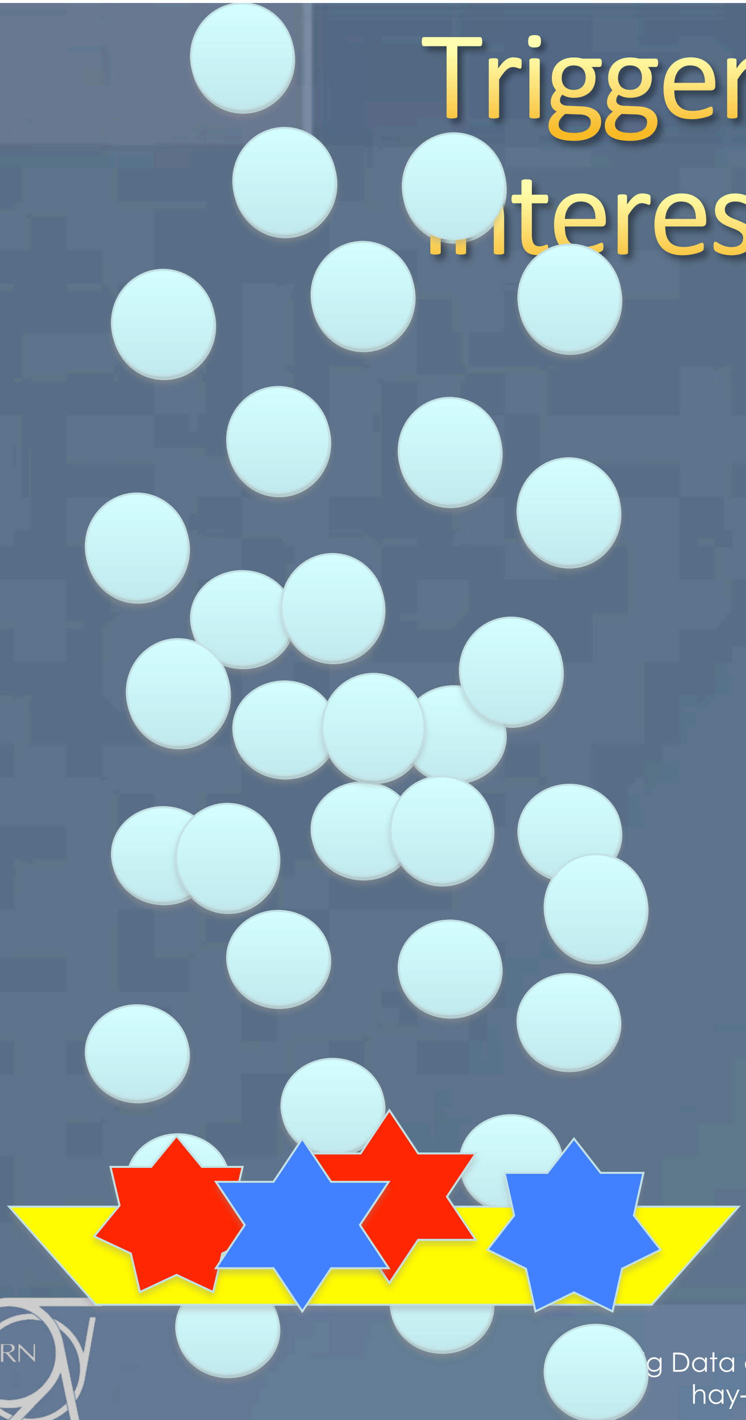
The hay: 15 million sensors

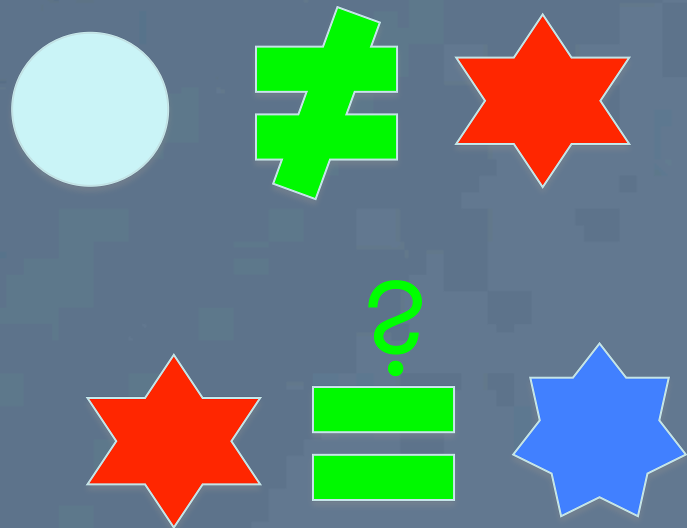


- 15 million sensors
- Giving a new value 40.000.000 / second
- = ~15 * 1,000,000 * 40 * 1,000,000 bytes
- = ~ 600 TB/sec
- 600 TB = 120000 DVDs =

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

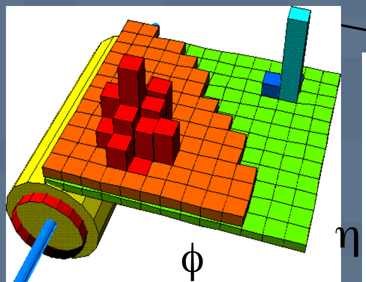
Physics Process	Events/s
Inelastic p-p scattering	10^8
b	10^6
$W \rightarrow e\upsilon$; $W \rightarrow \mu\upsilon$; $W \rightarrow \tau\upsilon$	20
$Z \rightarrow ee$; $Z \rightarrow \mu\mu$; $Z \rightarrow \tau\tau$	2
t	1
Higgs boson (all; m_H = 120GeV)	0.04
Higgs boson (simple signatures)	0.0003
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 s^{-1}	400-10,000
High Level Trigger (software on server farms)	2000-1000000 s^{-1}	1000 -10000 s^{-1}	10-2000

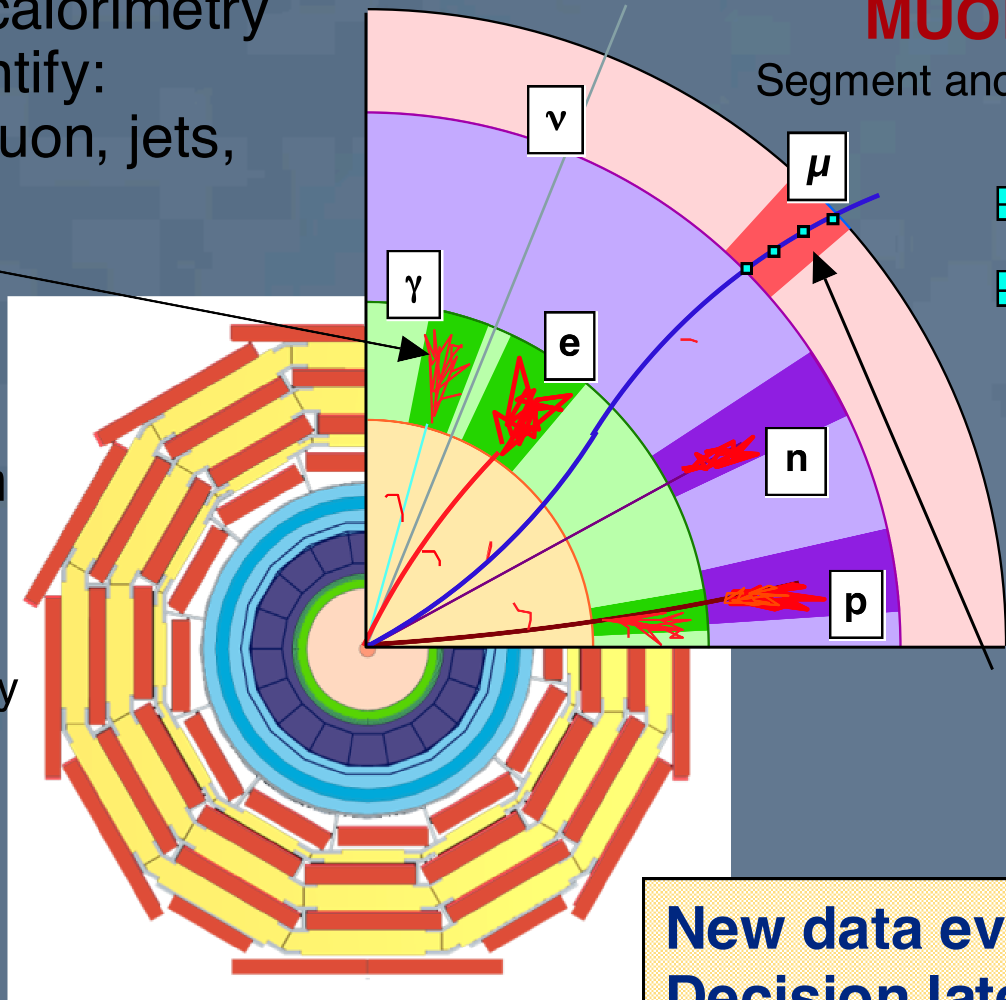
Challenge #1
The first level trigger

A small subset of the data to select events

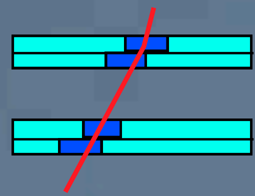
Use prompt data (calorimetry and muons) to identify:
High p_t electron, muon, jets, missing E_T



CALORIMETERs
Cluster finding and energy deposition evaluation

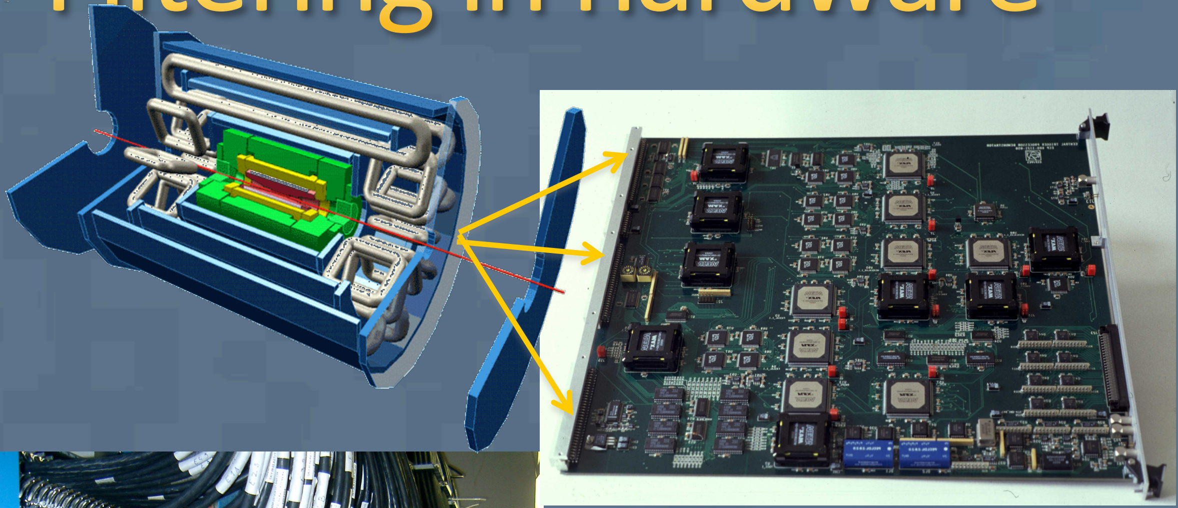


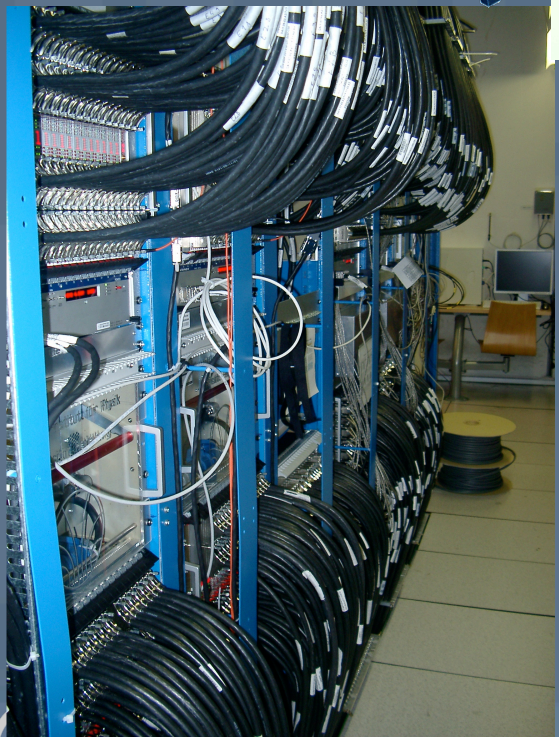
MUON System
Segment and track finding



New data every 25 ns
Decision latency ~ µs

Filtering in hardware



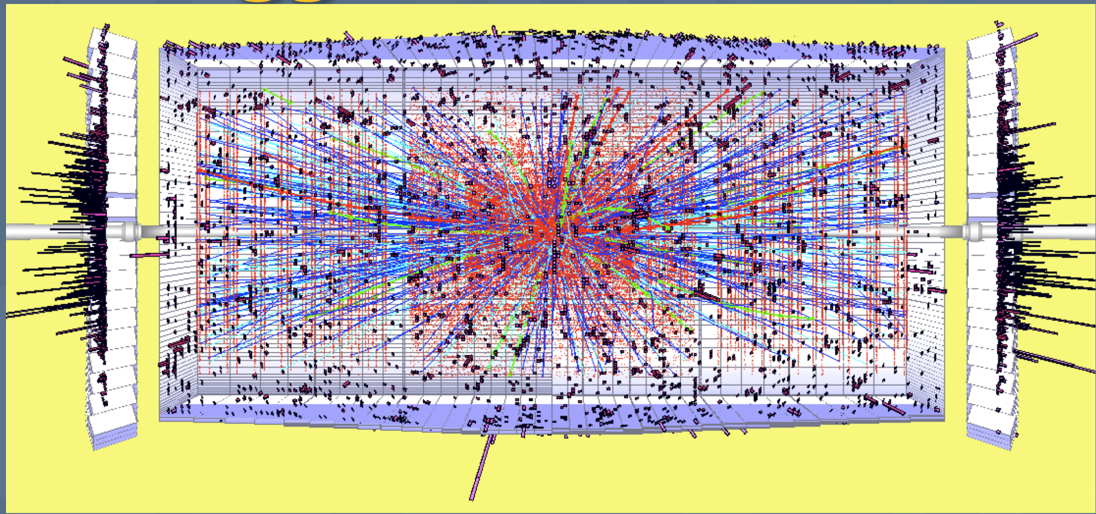


- Sophisticated electronics
- Hundreds of custom-built boards – process a small piece of the collision at enormous speeds (40 million times / second)
- They give a crude, but effective decision, based on simple criteria

Level 1 Trigger

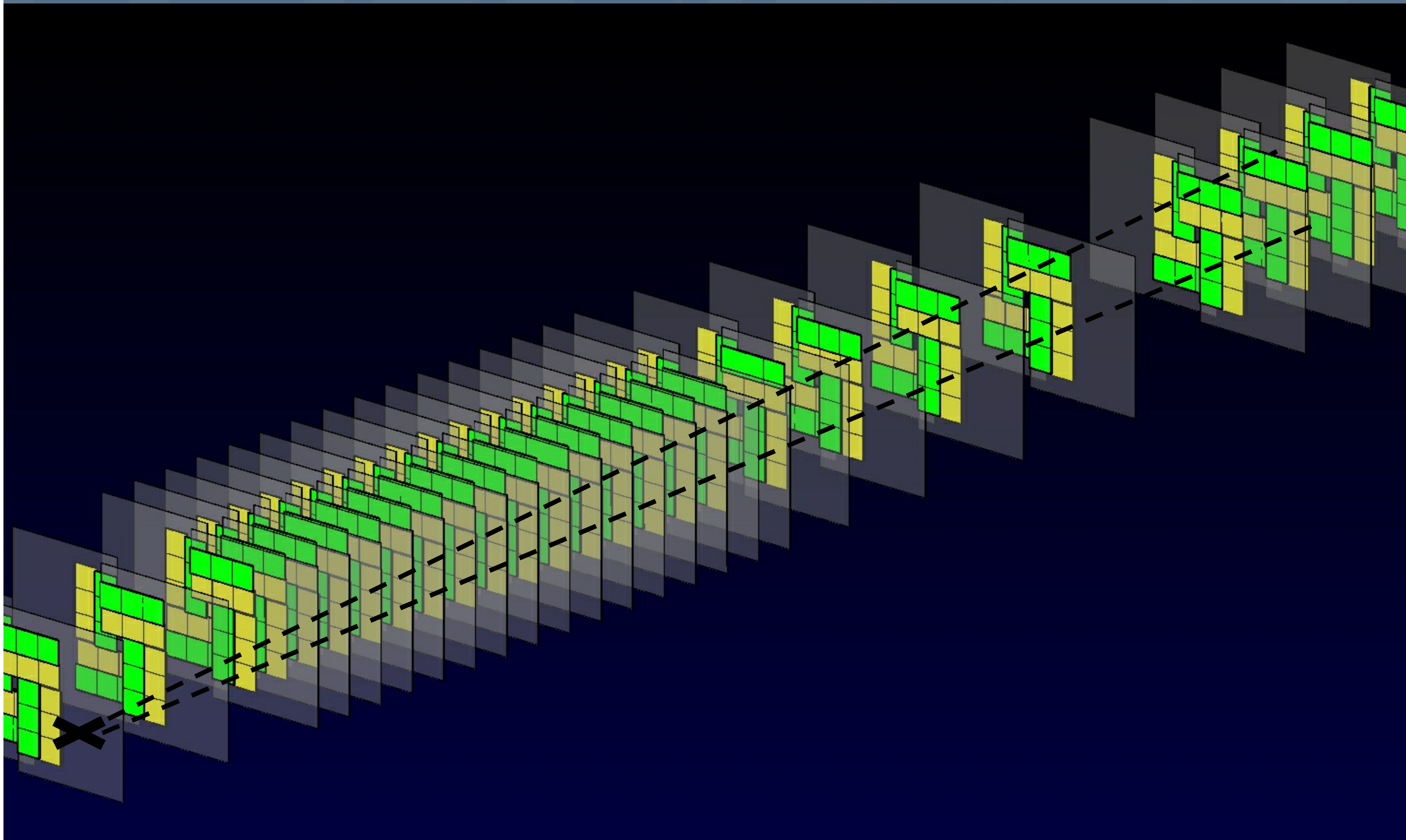
- The Level 1 Triggers are implemented in hardware: FPGAs and ASICs → difficult / expensive to upgrade or change, maintenance by experts only
- Decision time: ~ a small number of microseconds → The Level 1 Triggers are hard real-time systems
- They use "simple" hardware-friendly signatures → working with partial information and with drastic simplifications has a price → interesting and valuable events are lost

A Track-Trigger at 40 MHz 2020++

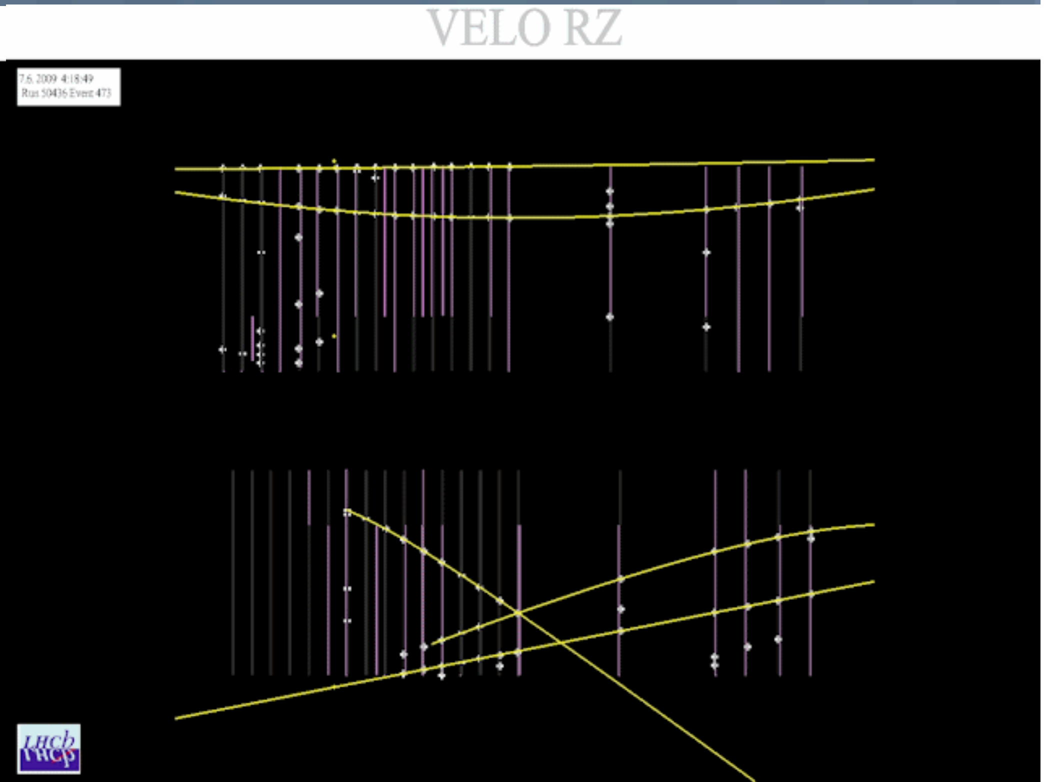


- Goals:
  - Find 1000s of particle trajectories in real-time (couple of micro-seconds)
  - Improve sensitivity to interesting events

Pattern finding - tracks



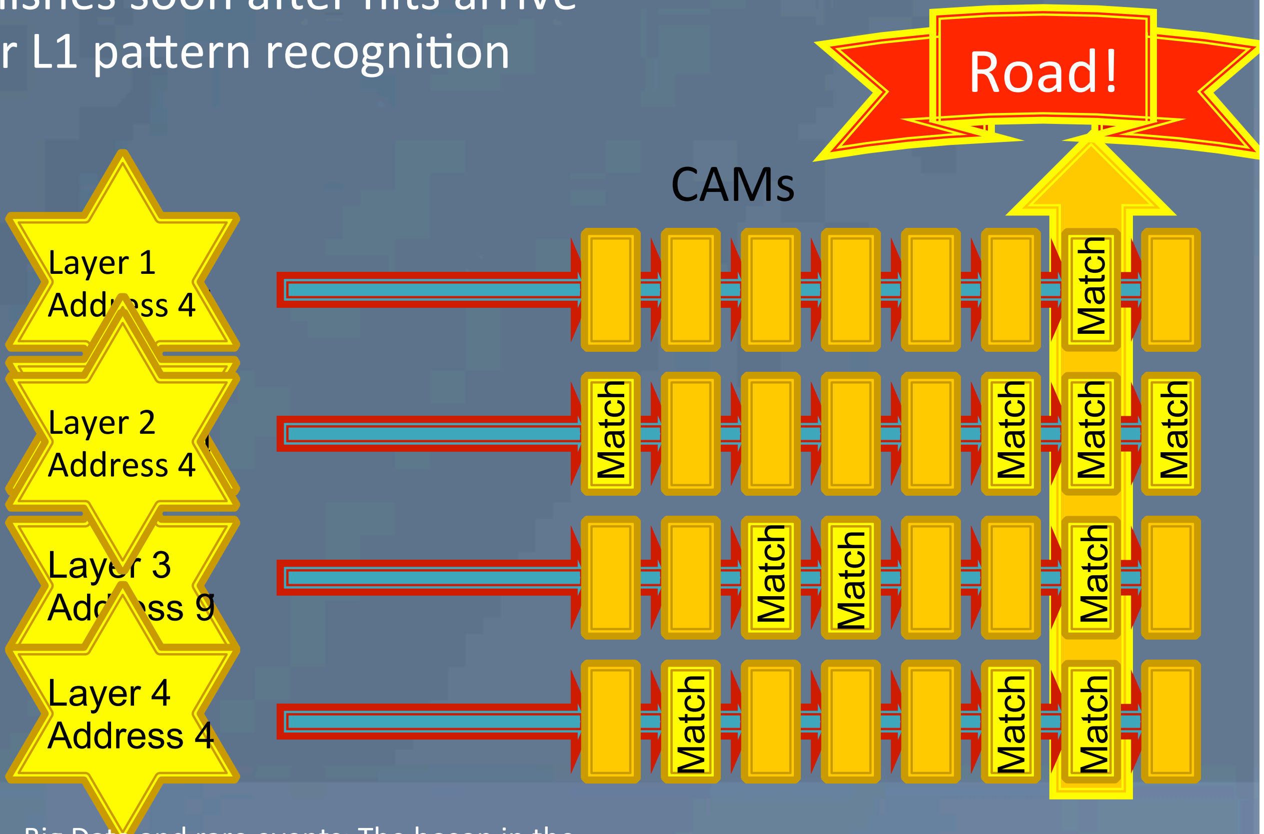
Same in 2 dimensions



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

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)

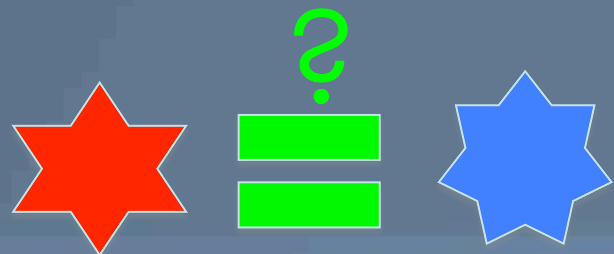
Challenge #2
Data Acquisition and High Level Trigger

CERN

High Level Trigger

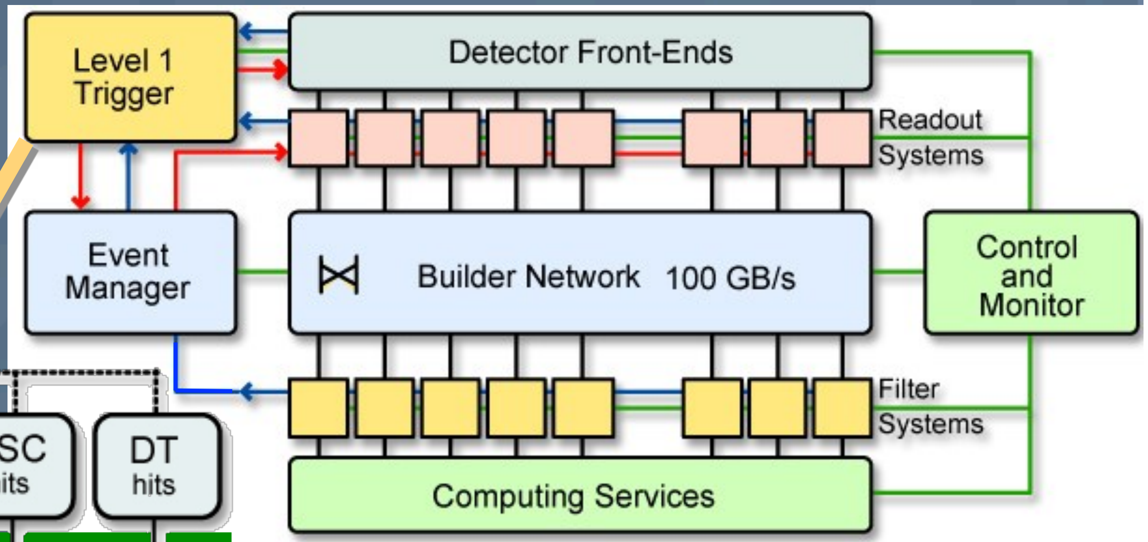
“And this, in simple terms, is how we find the Higgs Boson”

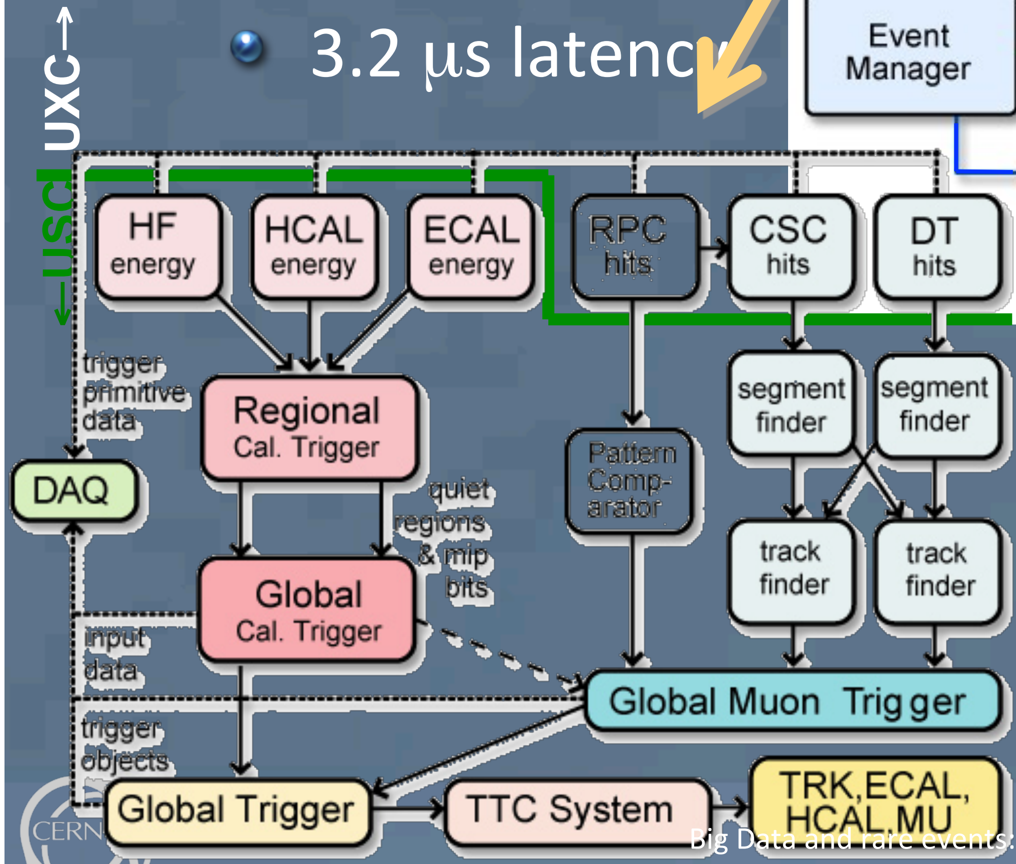
- Pack the knowledge of tens of thousands of physicists and decades of research into a huge sophisticated algorithm
- Several 100.000 lines of code
- Takes (only!) a few 10 - 100 milliseconds ***per collision***



CMS 2012 L-1 Trigger & DAQ

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





Interaction rate: 1 GHz

Bunch Crossing rate: 40 MHz

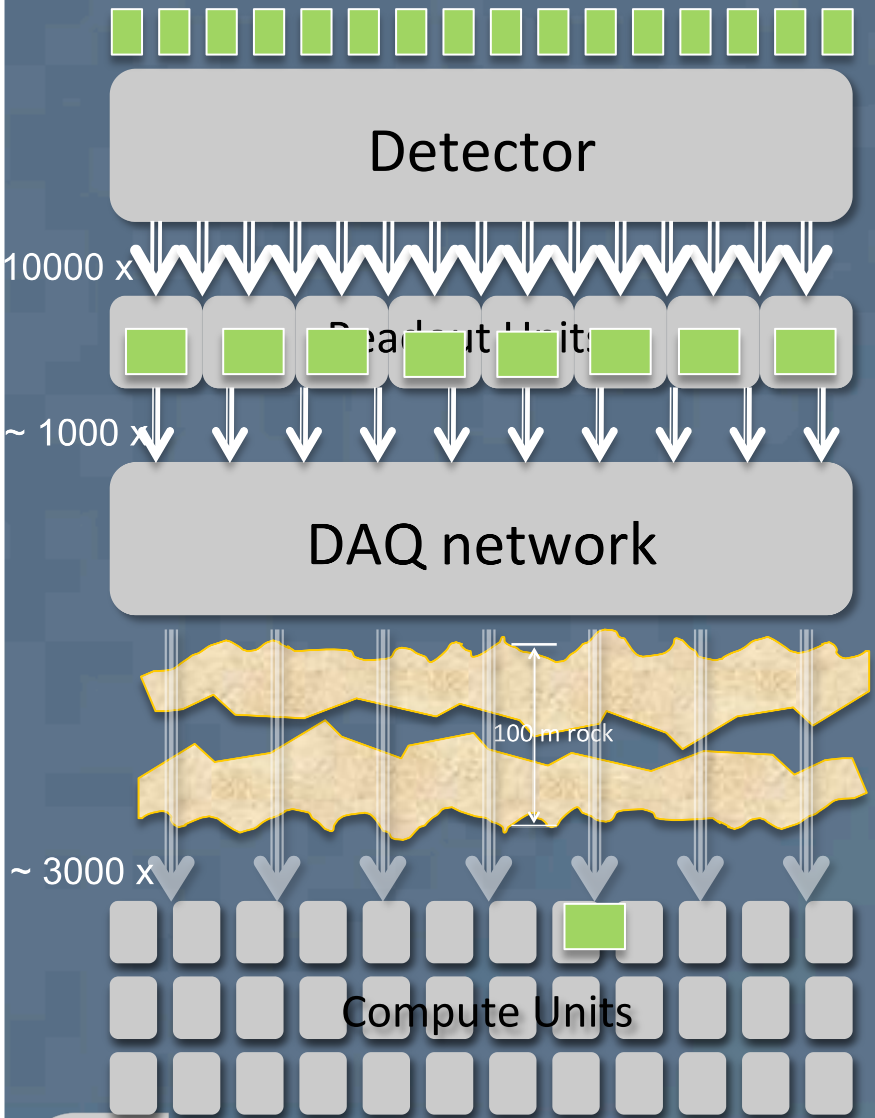
Level 1 Output: 100 kHz

Output to Storage: 400 Hz

Average Event Size: 1 MB

Data production 1 TB/day

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 (10/40/100 GbE / InfiniBand)

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

Future DAQs in numbers

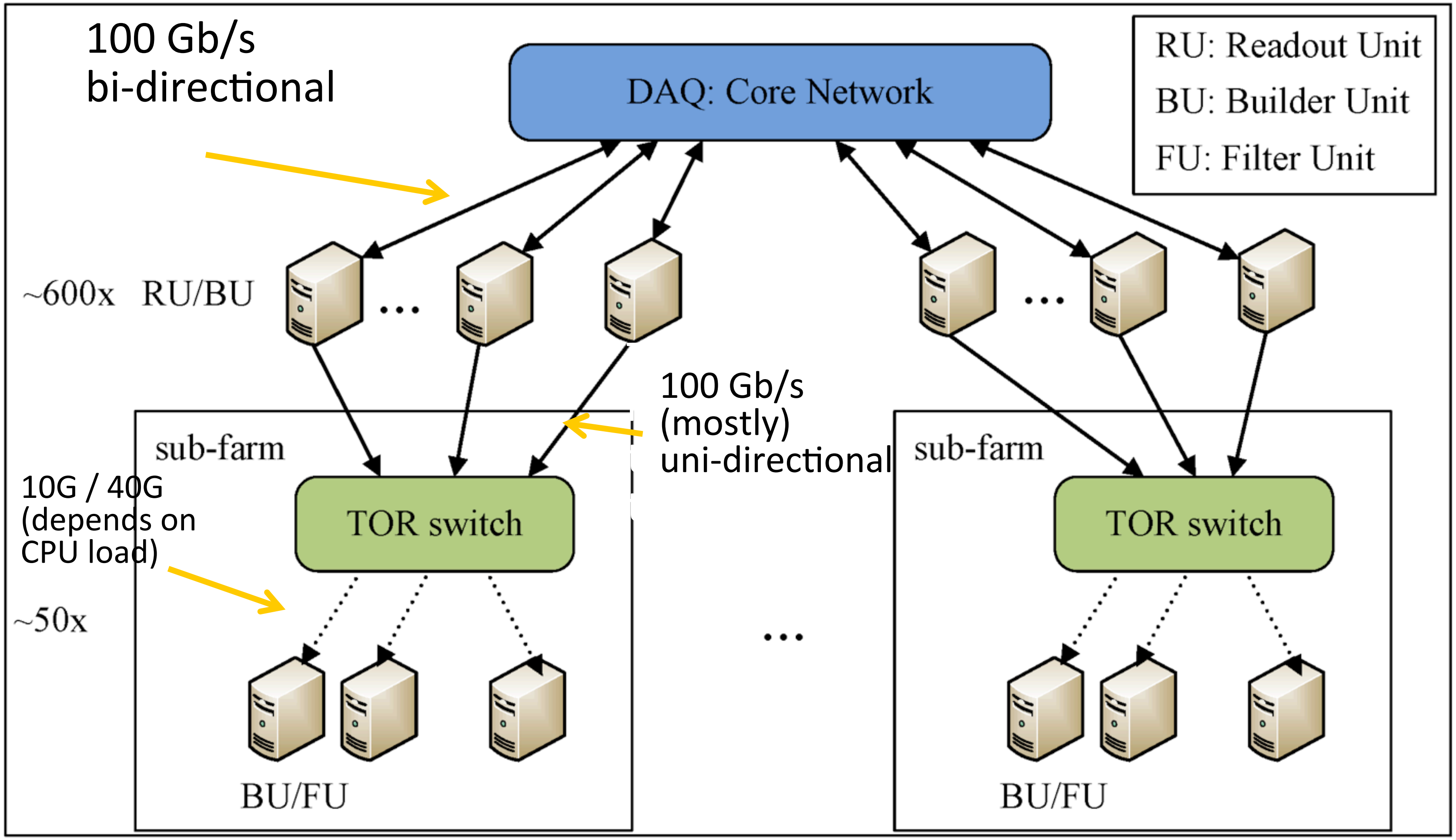
	Event-size [kB]	Rate of events into HLT [kHz]	HLT bandwidth [Gb/s]	Year [CE]
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	4000	1000	32000	2022
LHCb	100	40000	32000	2019

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 ☺)

A realistic DAQ / HLT for LHC



DAQ challenge

- 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")

Summary

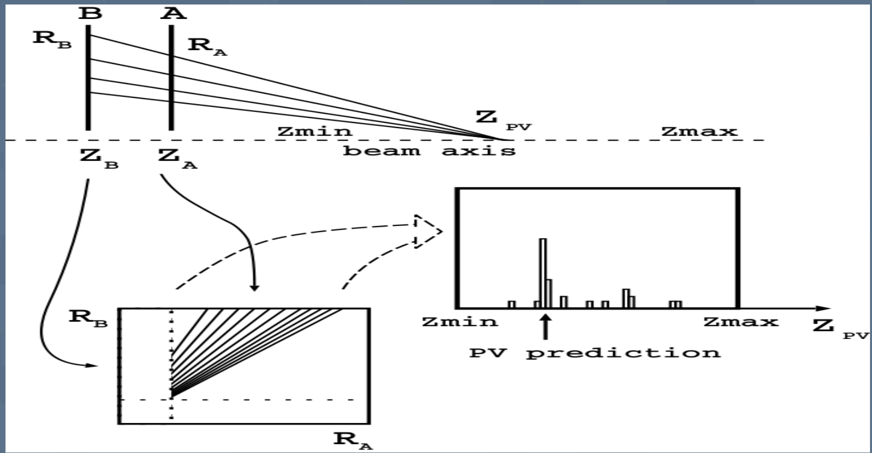
- The LHC experiments need to reduce 100 TB/s to ~ 25 PB/ year
- This is achieved with massive use of FPGAs, custom ASICs and x86 computing power
- Large, deep-buffer, local area networks are used to distribute data among the individual x86 servers
- The future will see massive increase of required *programmable* computing power, much more data will be moved off detector
    - Intense R&D ongoing on accelerators, non-X86, spatial computing and data-centre networking

Acknowledgements

- Many stimulating, fun discussions with the LHC Trigger-DAQ community and with a lot of smart people in CERN/IT (openlab) and industry are gratefully acknowledged
- Material has been adapted from A. Hirstius (now Intel), W. Smith (U. Wisconsin)

More material

Finding vertices in FPGAs

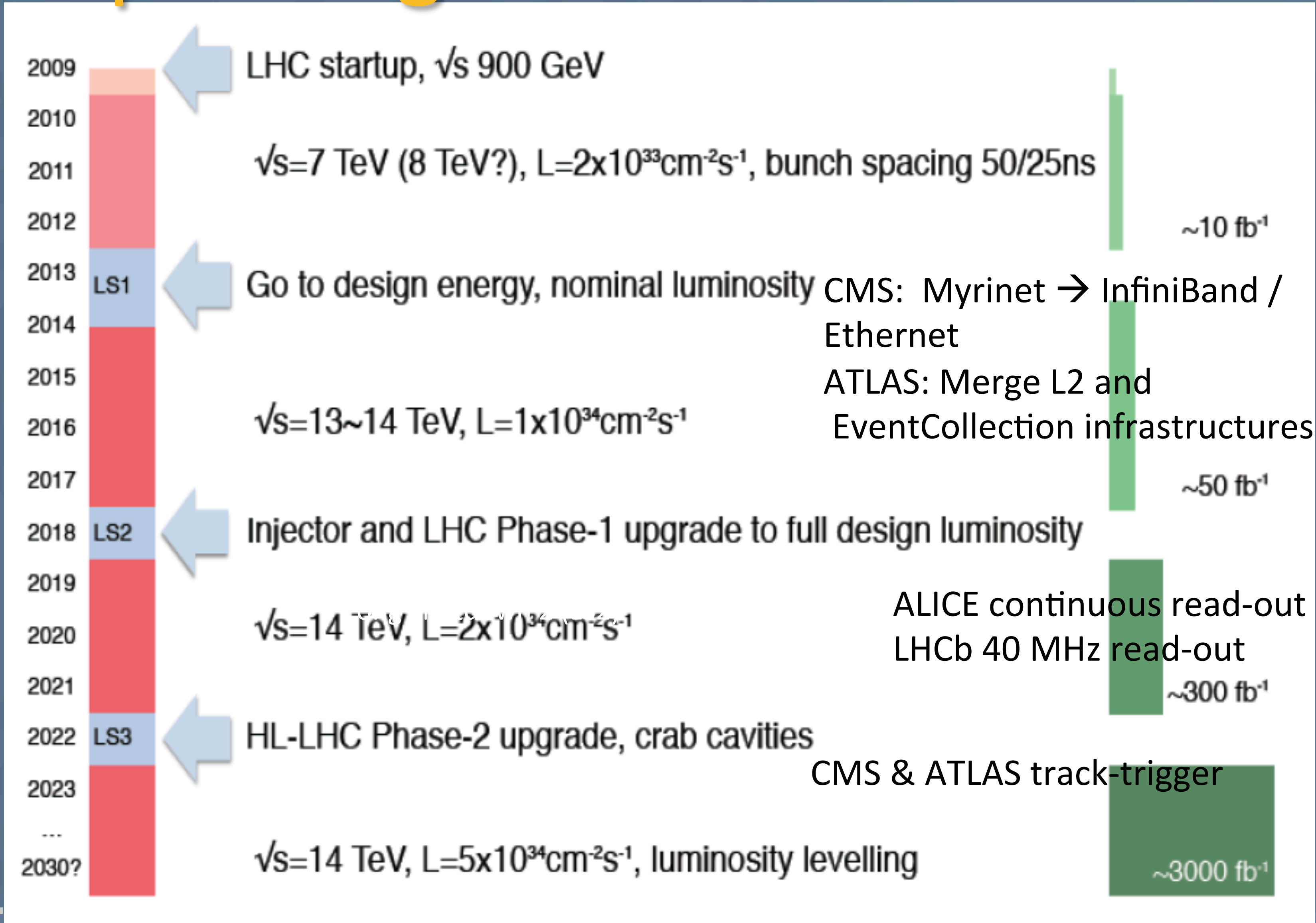


- Use r-coordinates of hits in Si-detector discs (detector geometry made for this task!)
- Find coincidences between hits on two discs
- Count & histogram

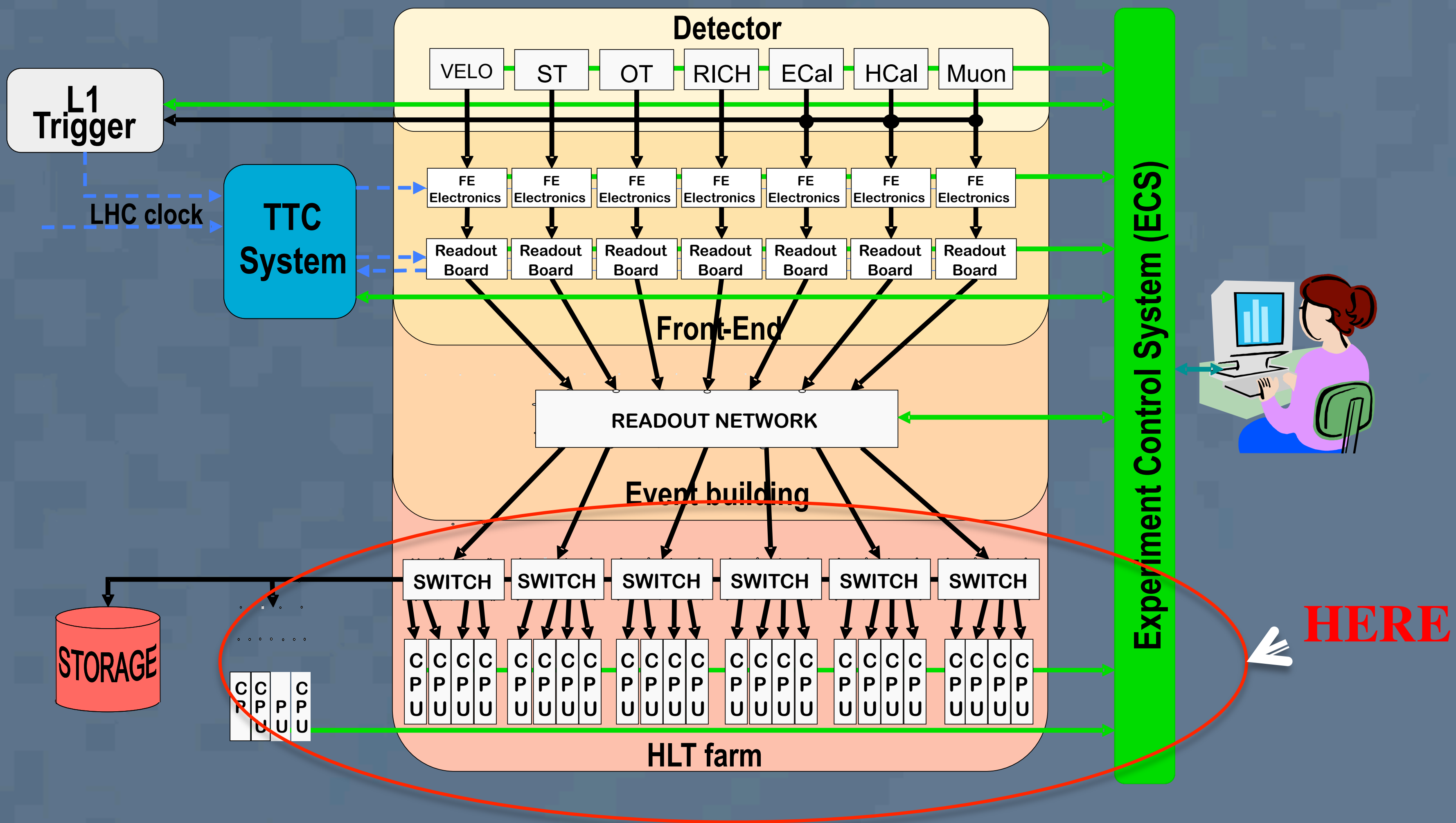
Moving on to Bigger Things…

- 15 million detector channels
- @ 40 MHz
- = ~15 * 1,000,000 * 40 * 1,000,000 bytes
- = ~ 600 TB/sec

LHC planning



The High Level Trigger is ...



The question is: How do we get the data in?

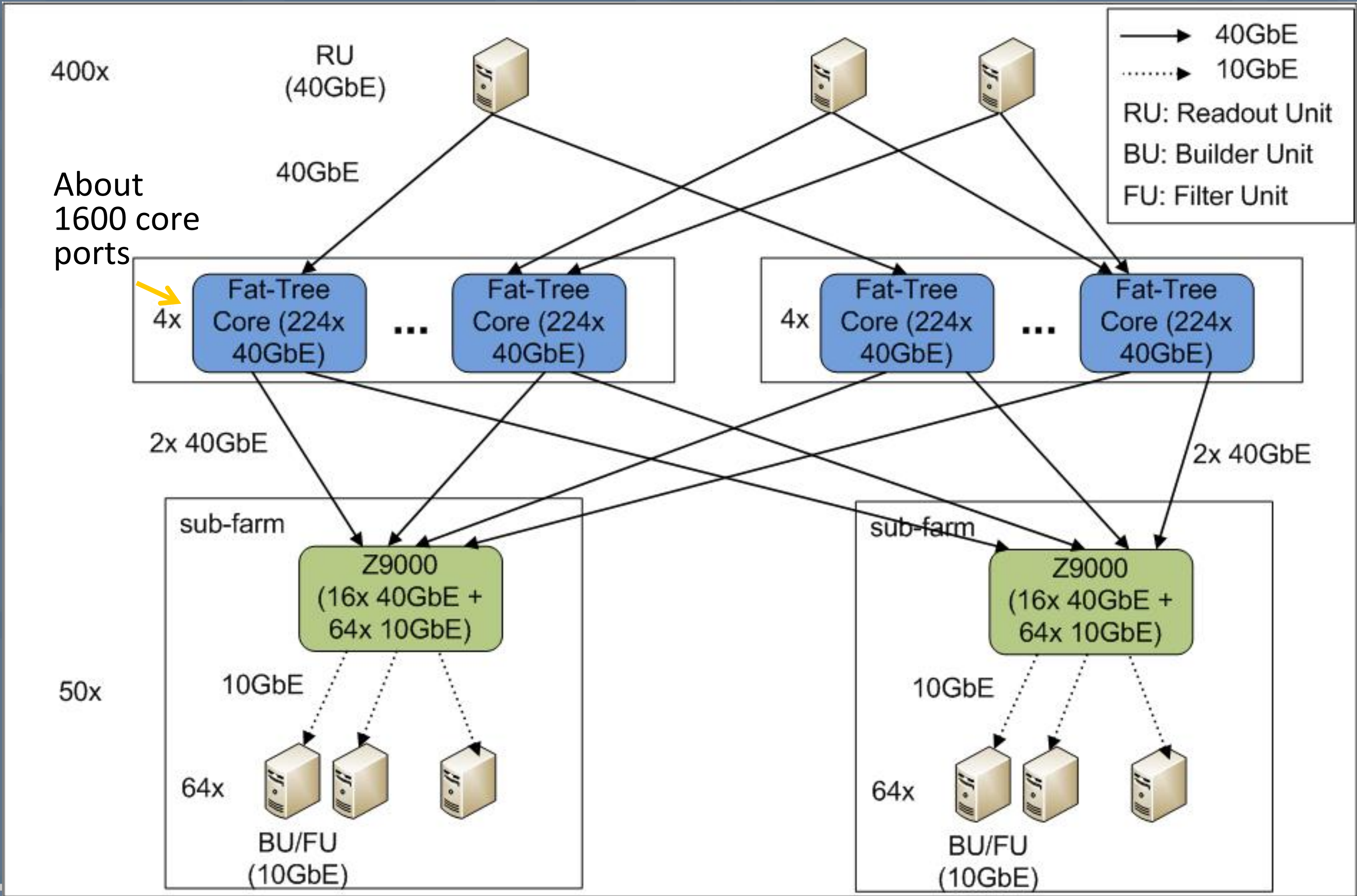
High Level Trigger: Key Figures

- Existing code base: 5 MLOC of mostly C++
- Almost all algorithms are single-threaded (only few exceptions)
- Currently processing time on a X5650 per event: several 10 ms / process (hyper-thread)
- Currently between 100k and 1 million events per second are filtered online in each of the 4 experiments

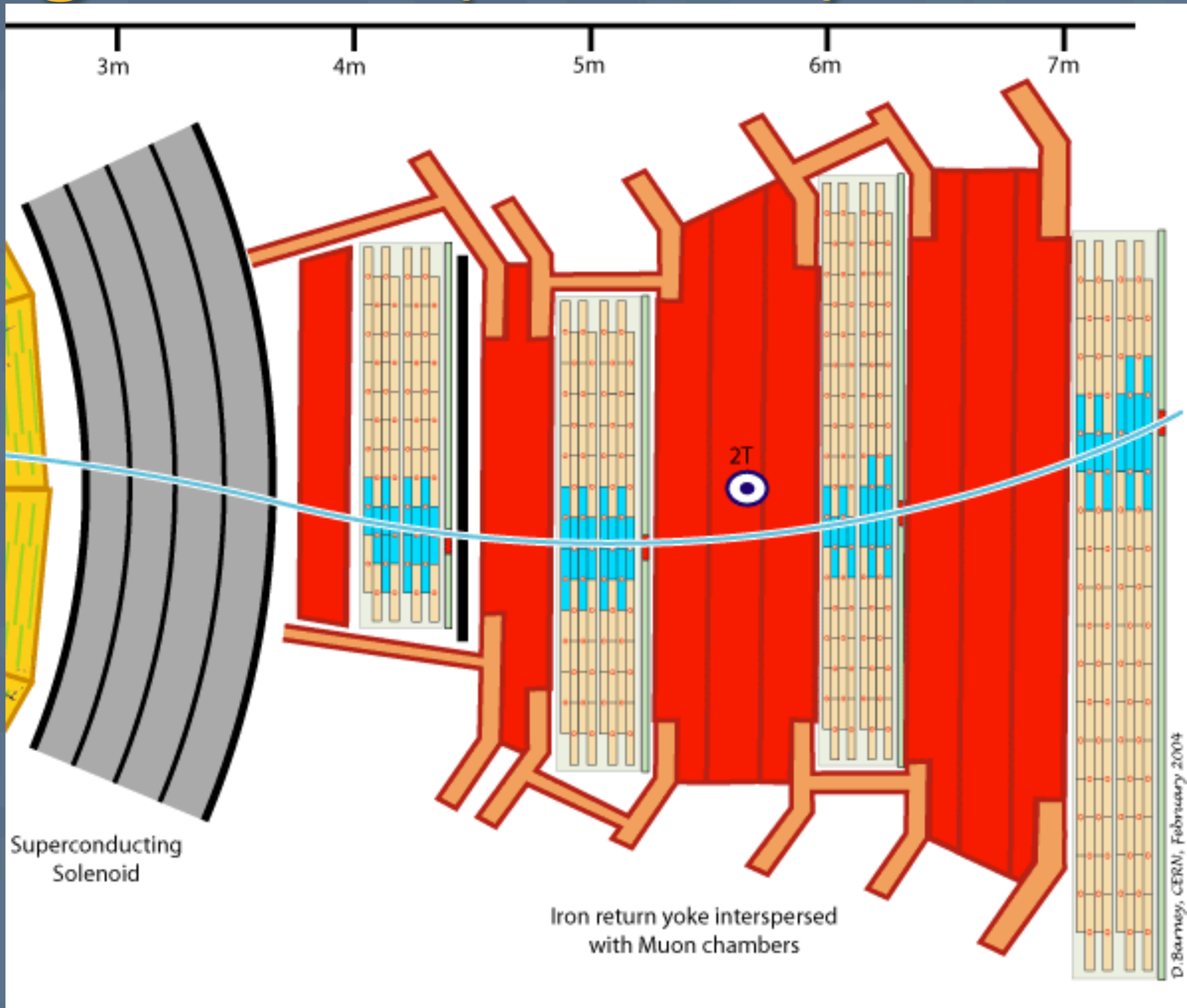
High Level Trigger compared to HPC

- Like HPC:
  - full ownership of the entire installation → can choose architecture and hardware components
  - single "client" / "customer"
  - have a high-bandwidth interconnect
- Unlike HPC:
  - many independent small tasks which execute quickly
    → no need for check-pointing (fast storage)
    → no need for low latency
  - data driven, i.e. when the LHC is **not** running (70% of the time) the farm is idle → interesting ways around this (deferal, "offline usage)
  - facility is very long-lived, growing incrementally

Classical fat-core event-builder



Finding Muons (2d view)



Calorimeter data