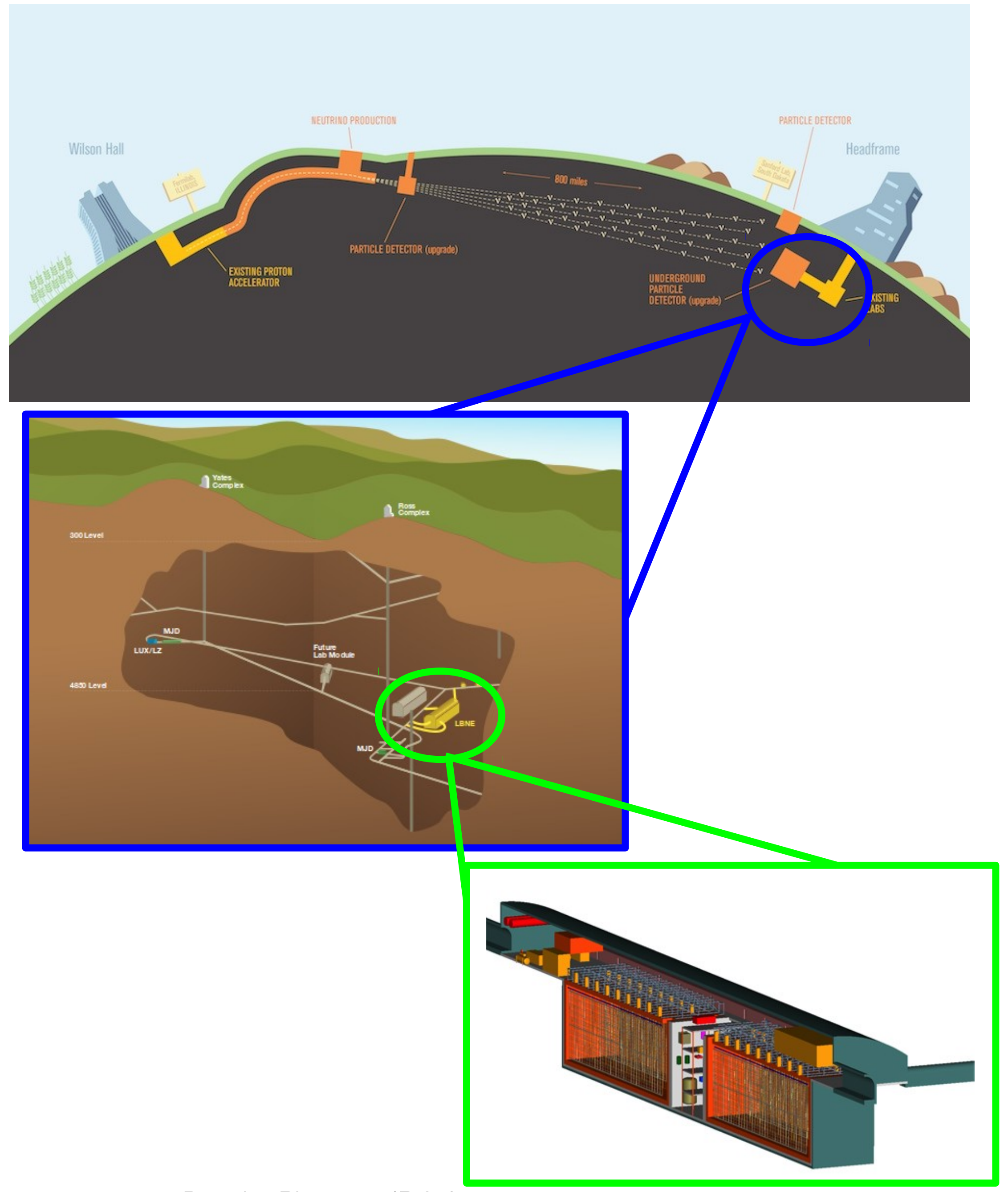
Python for Data Science 2

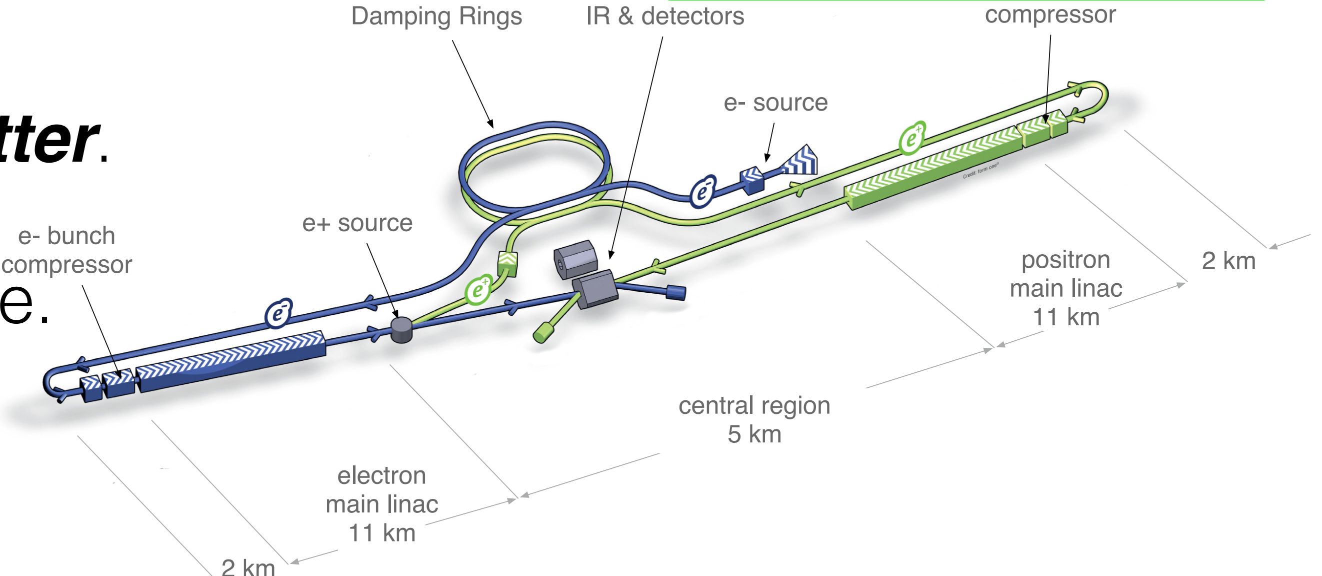
Lecture 12- Data from the Large Hadron Collider

Amir Farbin

Frontiers



- ***Energy Frontier***: ***Large Hadron Collider (LHC)*** at 13 TeV now, ***High Luminosity (HL)- LHC*** by 2025, perhaps 33 TeV LHC or 100 TeV Chinese machine in a couple of decades.
  - Having found Higgs, moving to studying the SM ***Higgs*** find new Higgses
  - Test ***naturalness***
  - Find ***Dark Matter*** (reasons to think related to naturalness)
- ***Intensity Frontier***:
  - ***B Factories***: upcoming SuperKEKB/SuperBelle
  - ***Neutrino Beam Experiments***:
    - Series of current and upcoming experiments: Nova, MicroBooNE, SBND, ICURUS
    - ***US's flagship experiment*** in next decade: ***Long Baseline Neutrino Facility (LBNF)/Deep Underground Neutrino Experiment (DUNE) at Intensity Frontier***
  - Measure properties of ***b-quarks*** and ***neutrinos*** (newly discovered mass)… search for ***matter/anti-matter asymmetry***.
  - Auxiliary Physics: Study ***Supernova***. Search for ***Proton Decay*** and ***Dark Matter***.
- ***Precision Frontier***: ***International Linear Collider (ILC)***, hopefully in next decade. Most energetic e^+e^- machine.
  - ***Precision studies*** of ***Higgs*** and hopefully ***new particles*** found at LHC.



The LHC

- pp collisions 7+7 TeV $\Rightarrow \sqrt{s}$ = 14 TeV

1232 super-conducting dipoles B=8.3 T

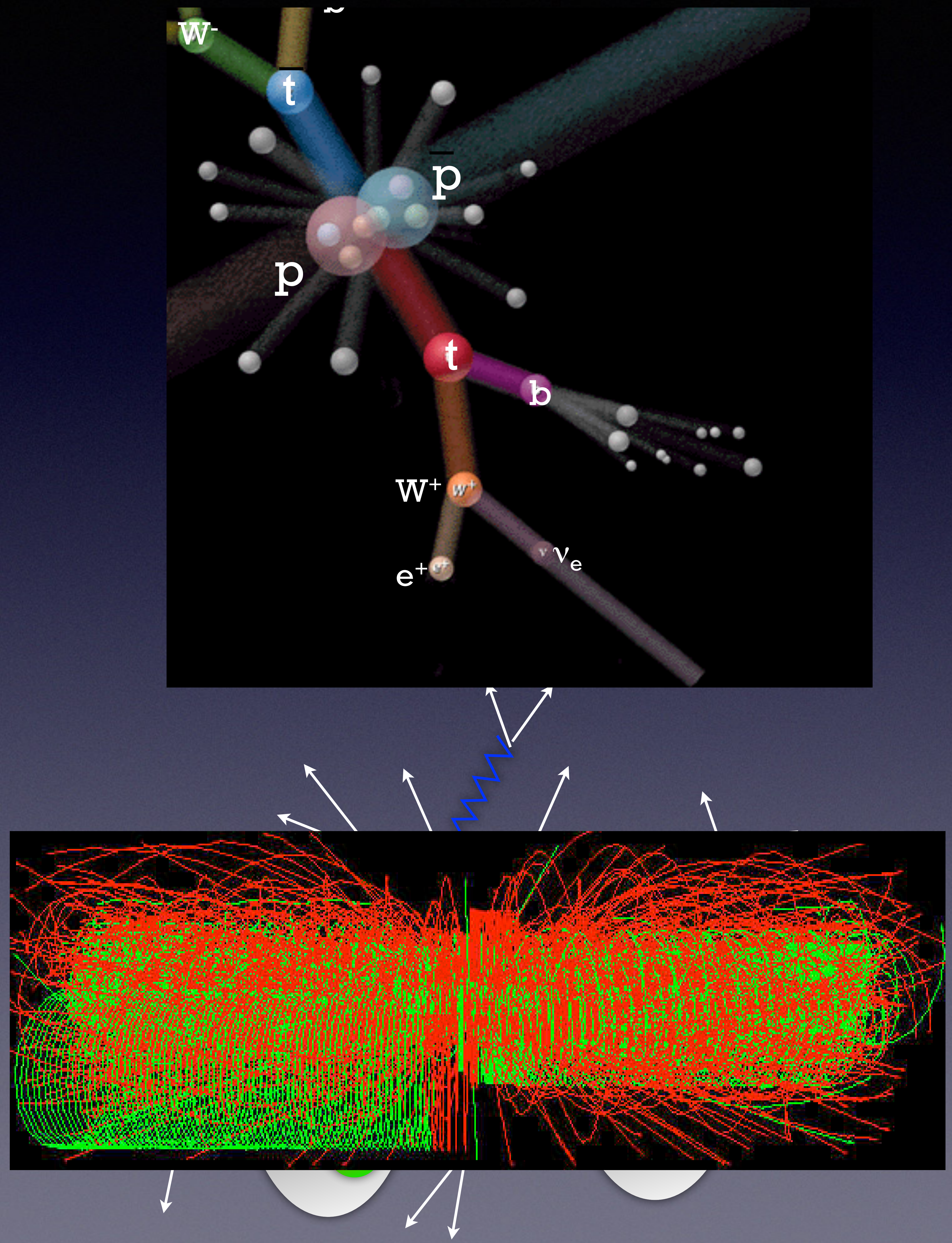
27 km

6400 other correctors

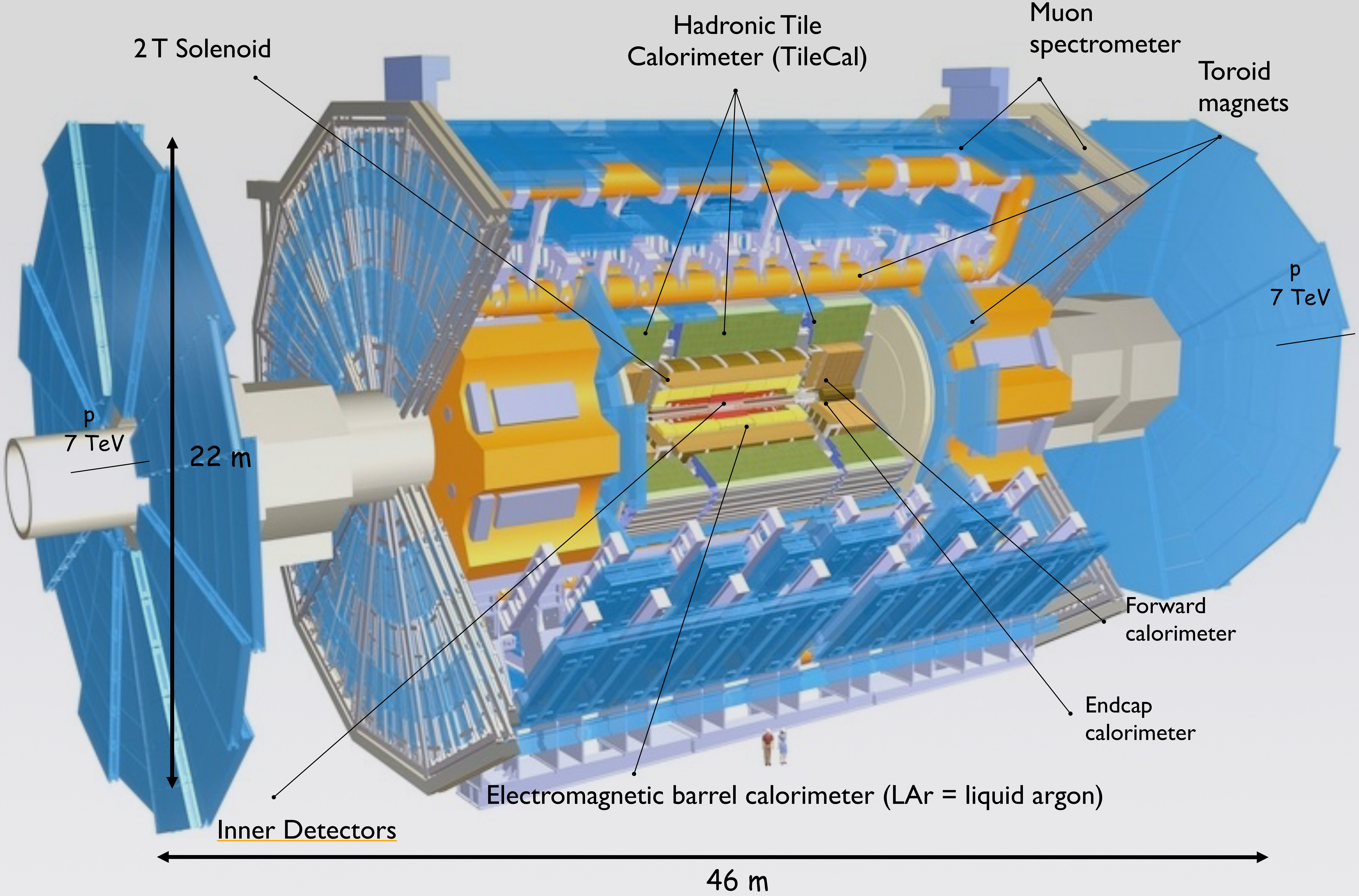
2835 Bunches, 10^{11} protons/bunch:

- L_{design} = 10^{34} cm^{-2} s^{-1}(after 2009)
- $L_{initial}$ $\leq$ few x 10^{33} cm^{-2} s^{-1} (until 2009)

LHC Environment

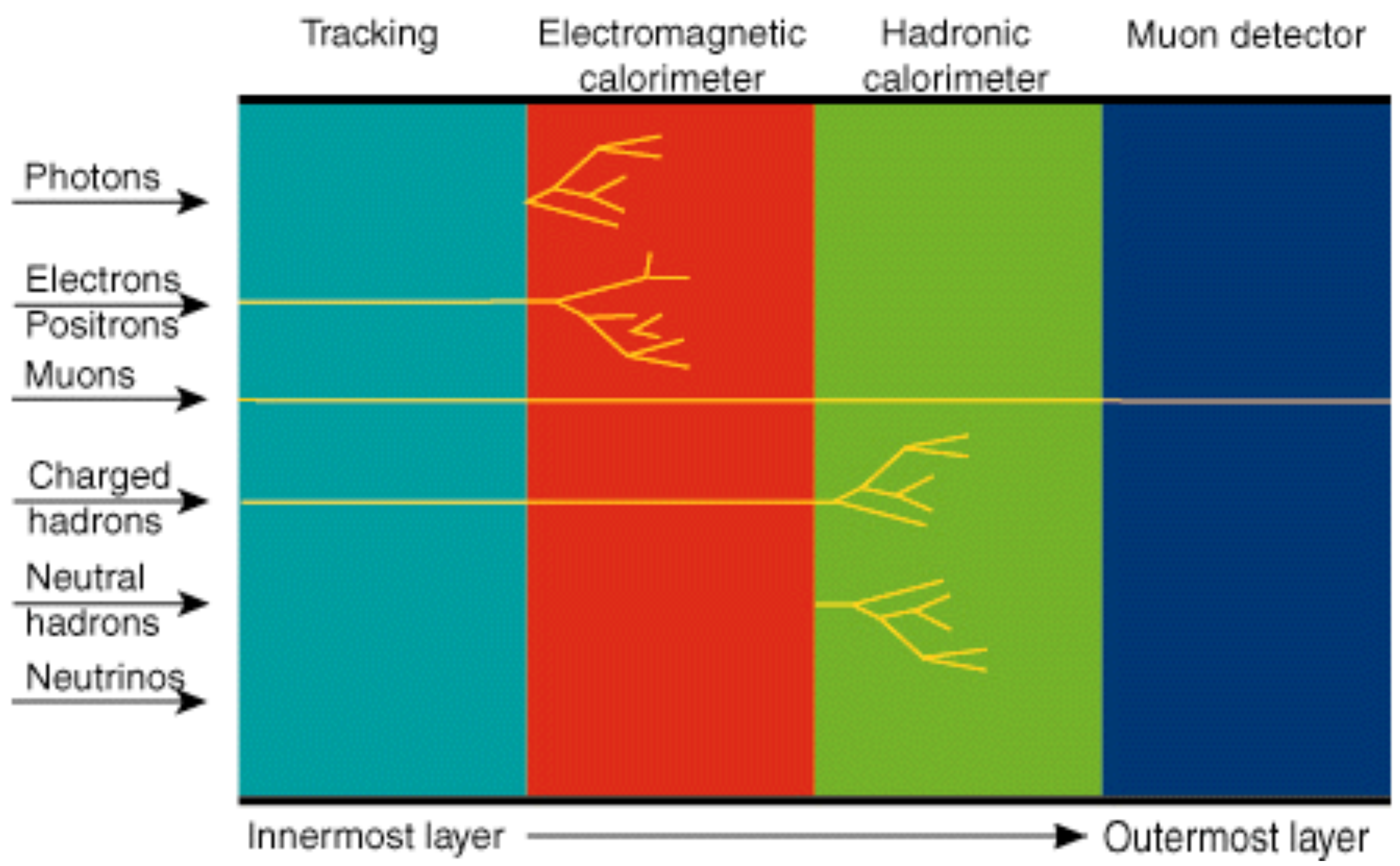


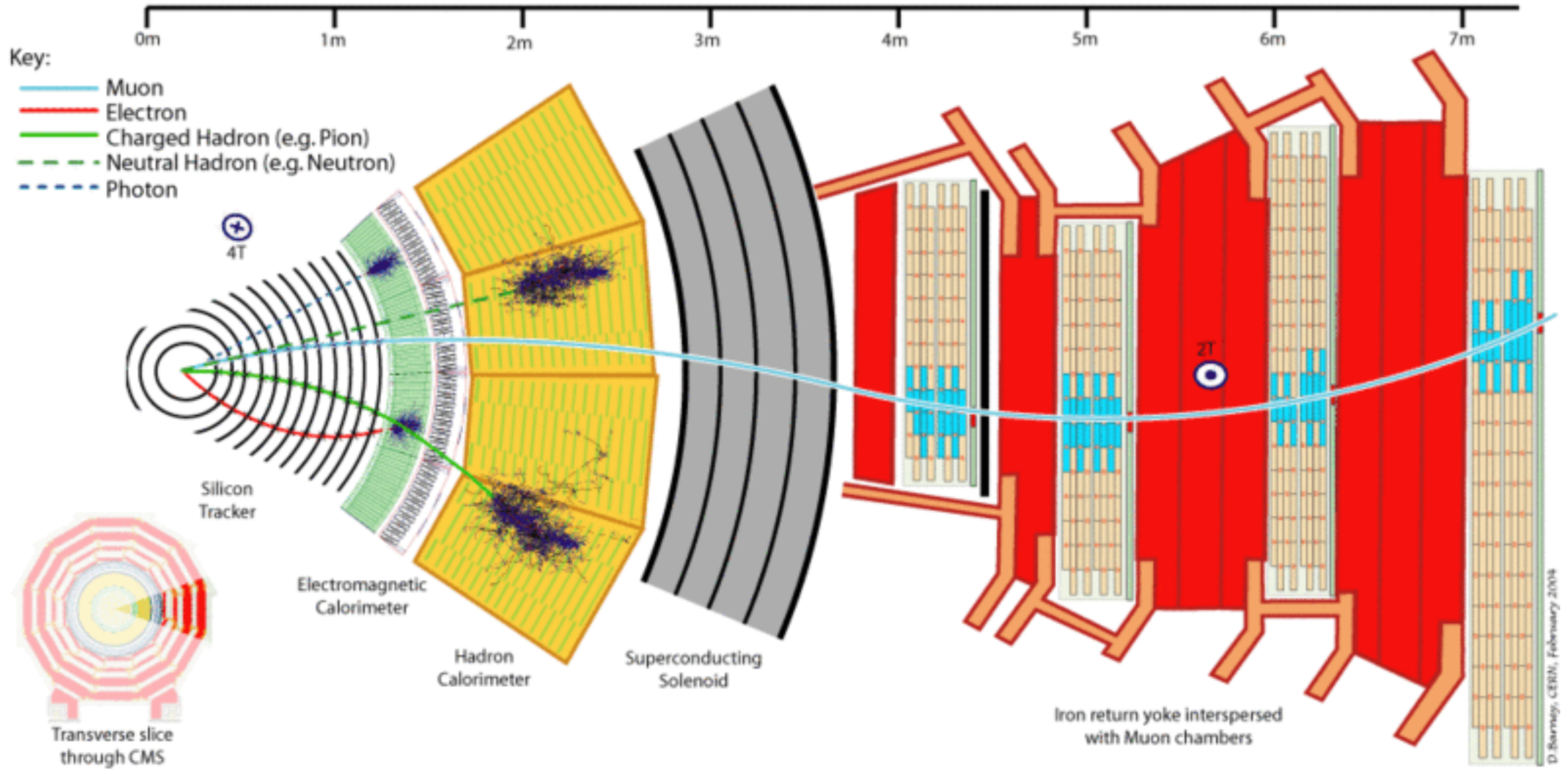
ATLAS

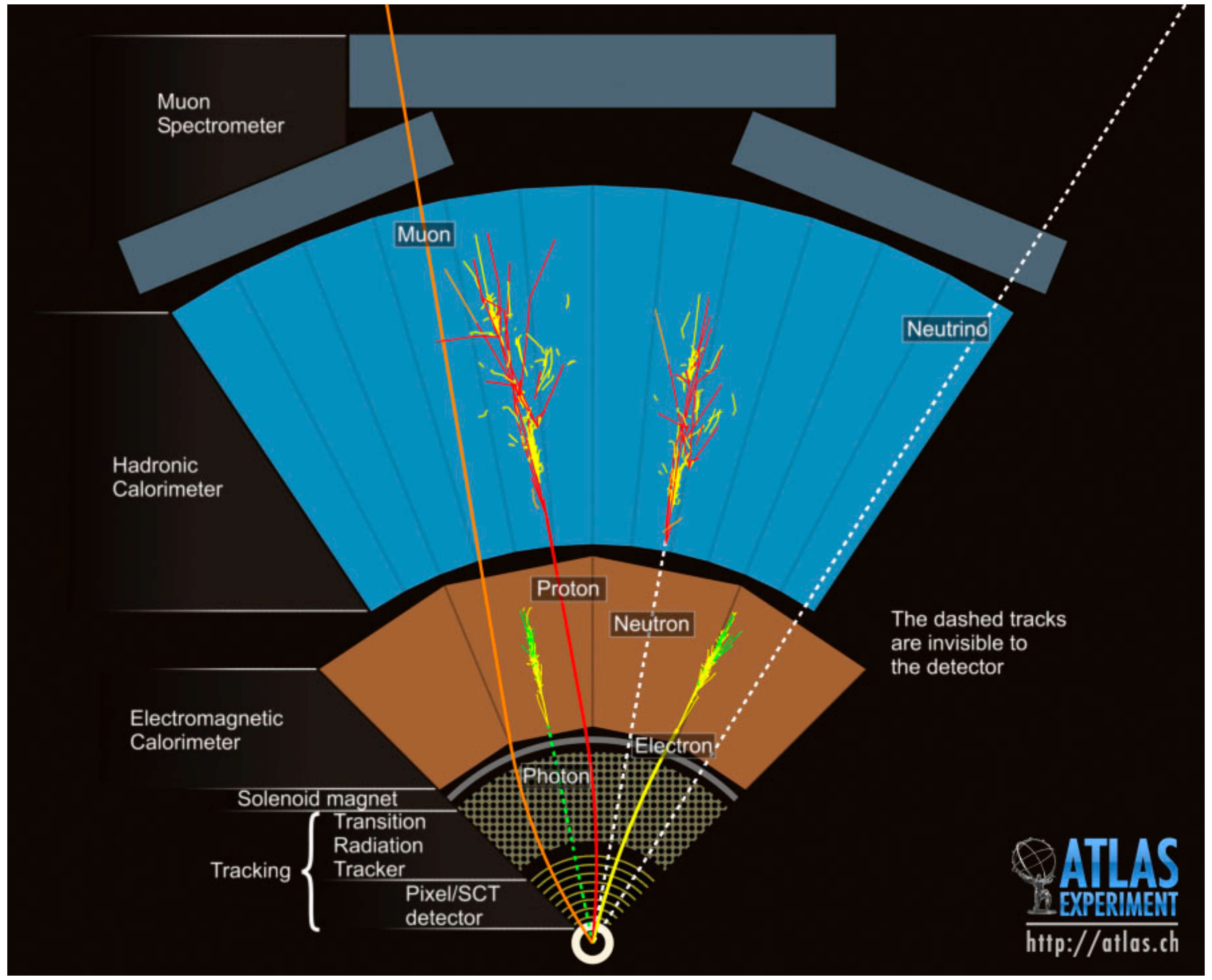


Total mass ~ 7000 tonnes, installed 92 m underground.

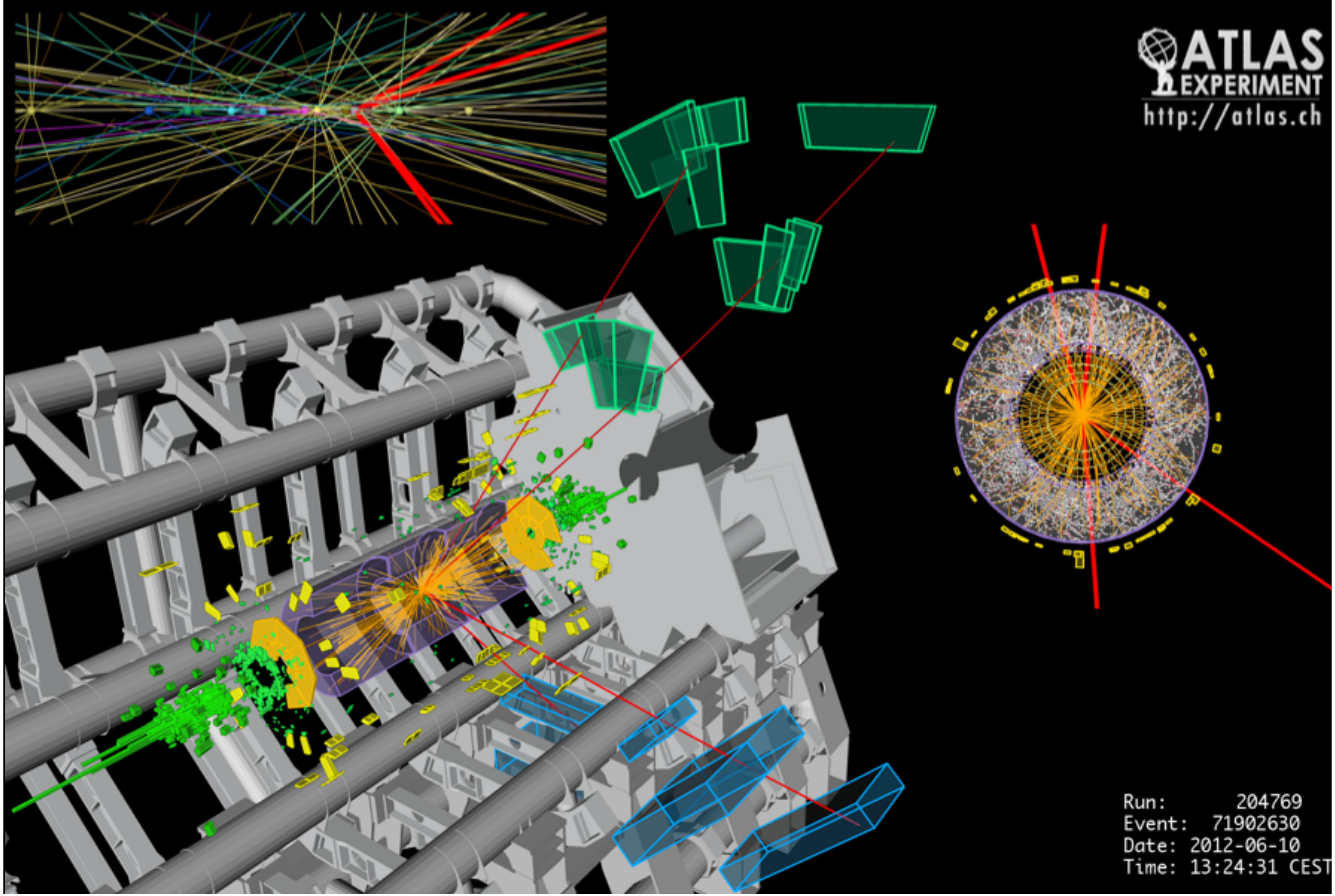
LHC/ILC detectors





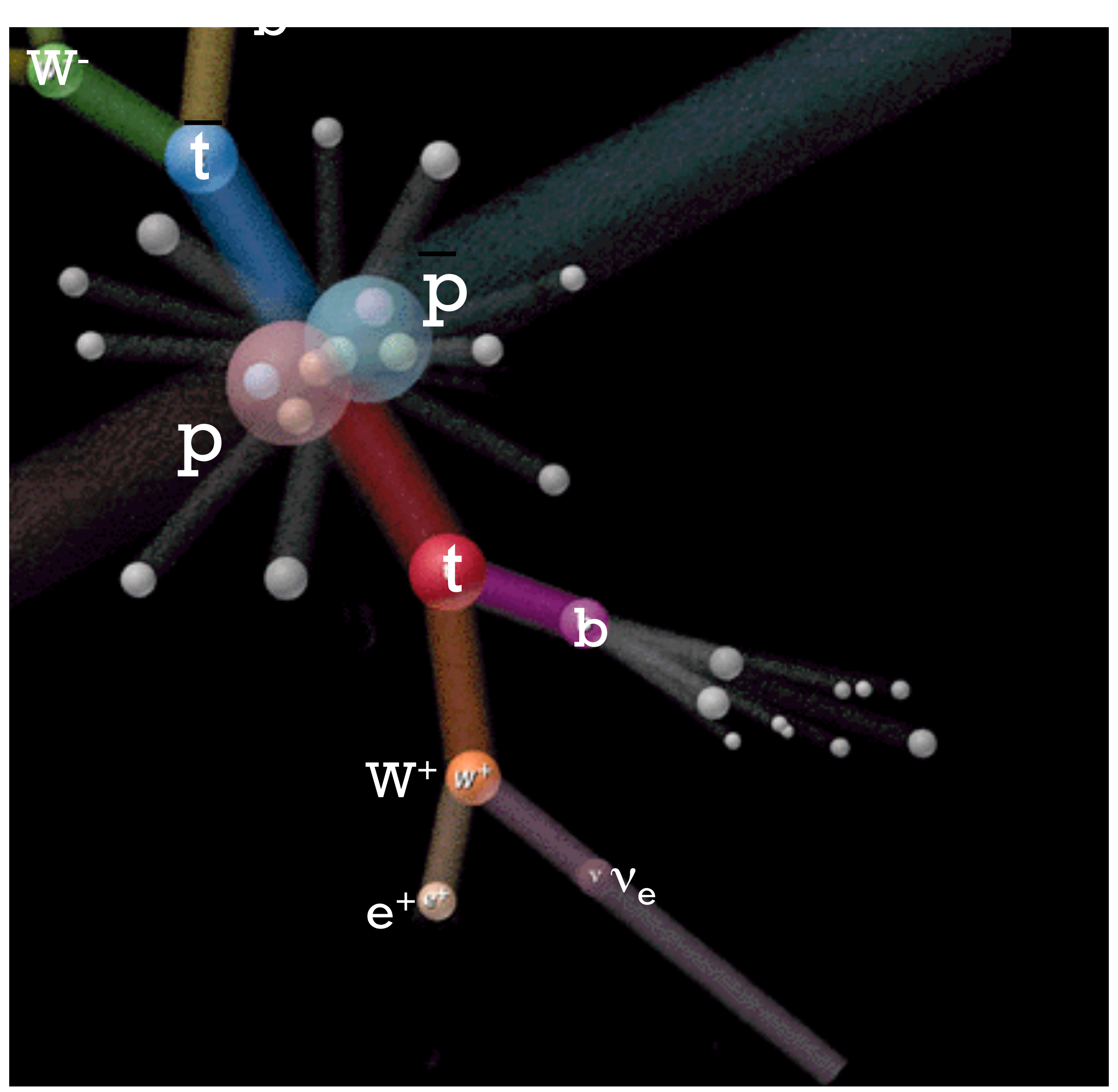


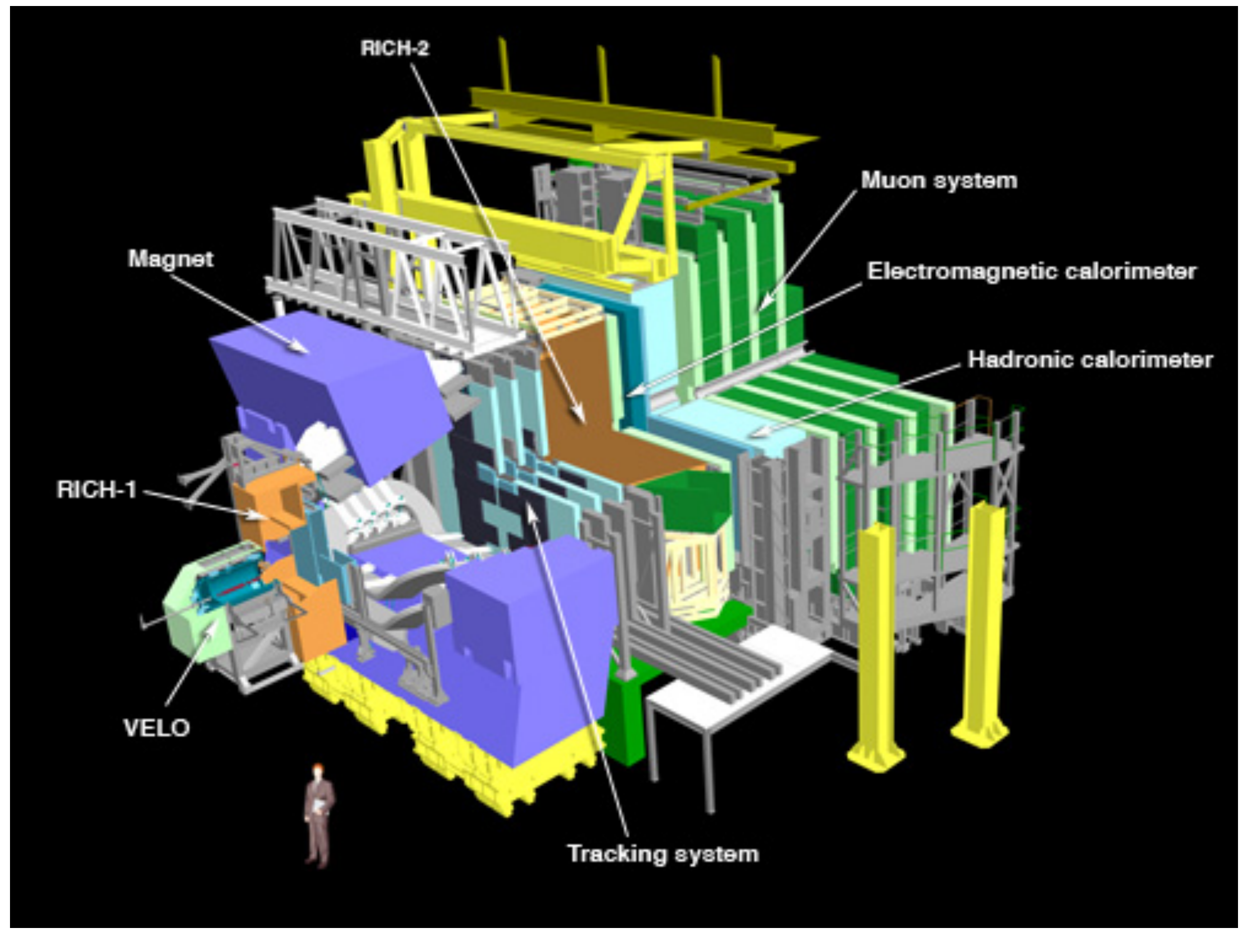
$$H \to ZZ \to 4l$$

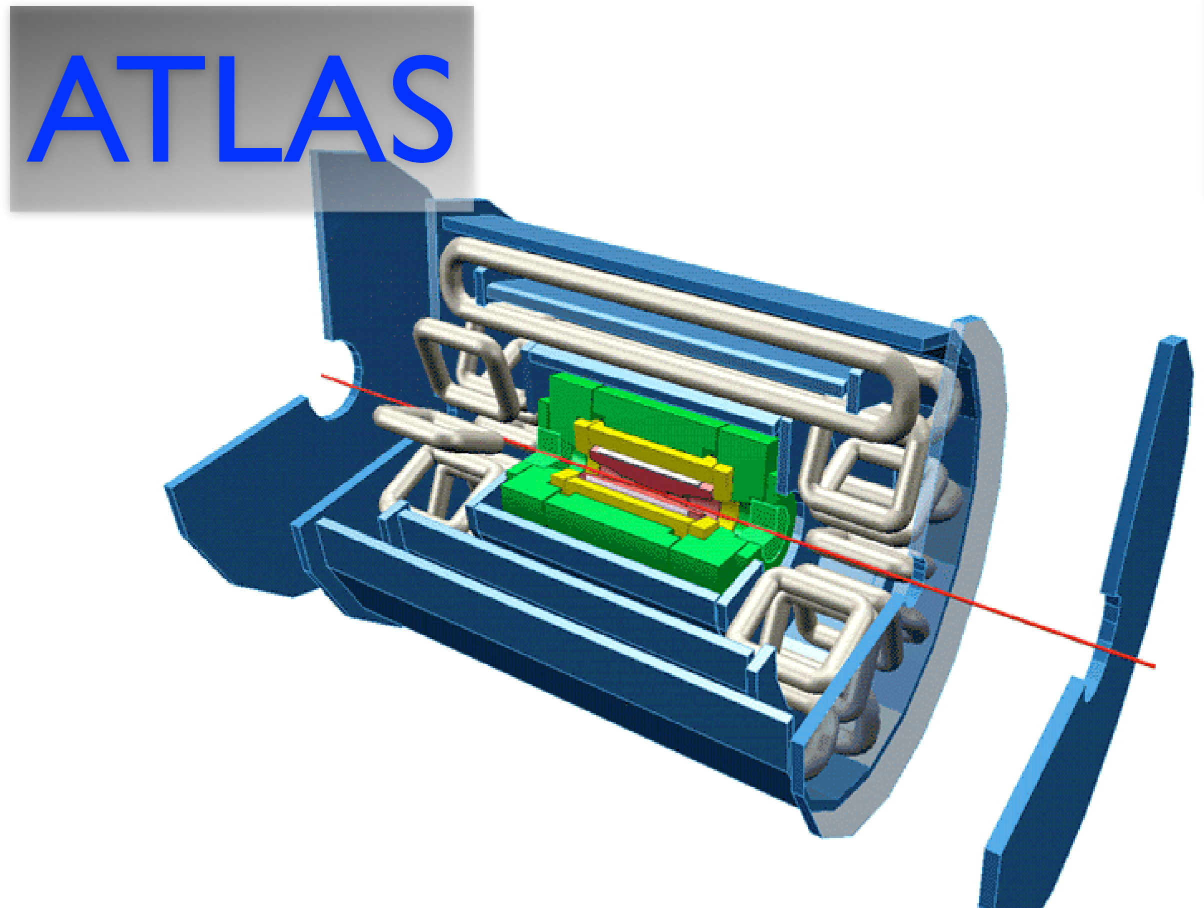

ATLAS
EXPERIMENT
http://atlas.ch
Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

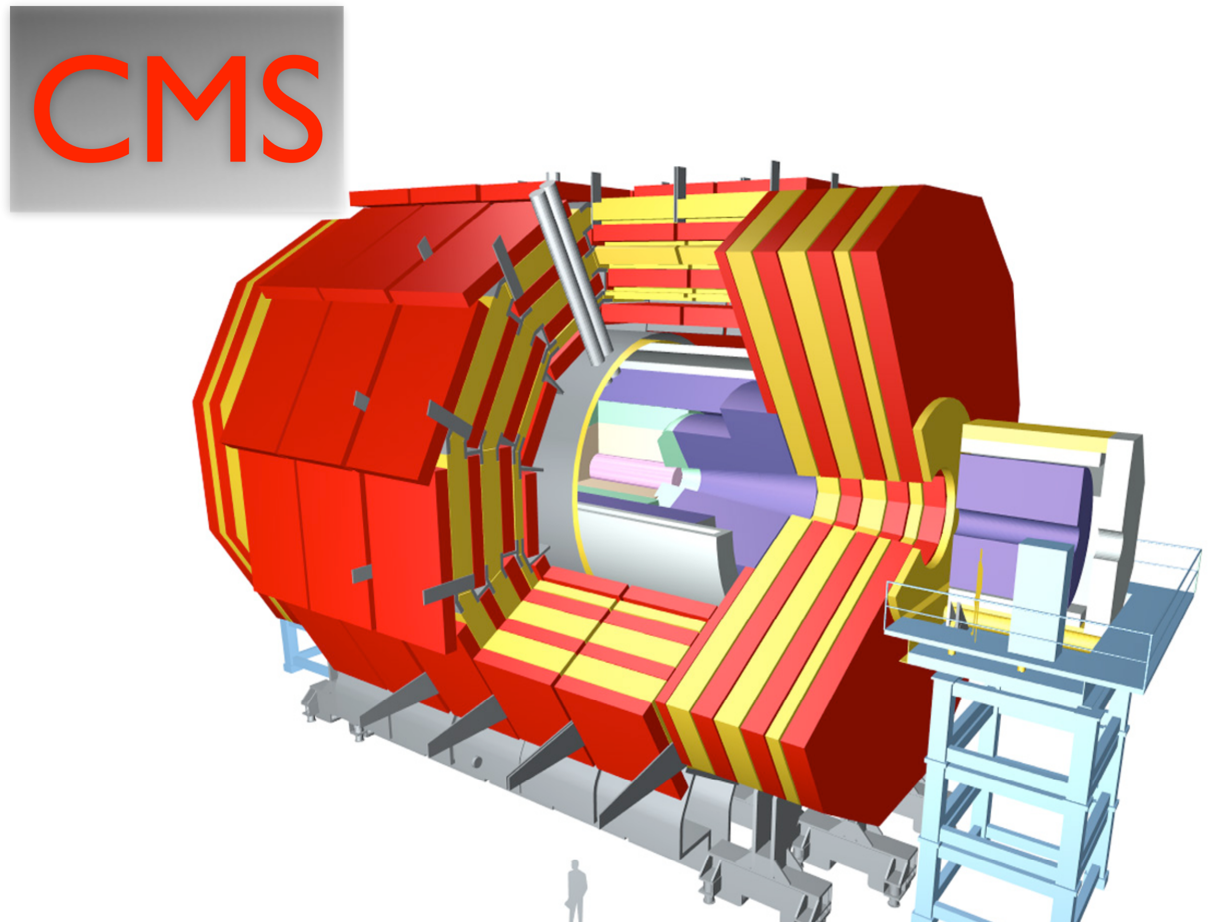
HEP Experiments

- 5 technical components to HEP experiment:
  - ***Accelerator***: e.g. LHC collisions creating quickly decaying heavy particles. Extremely high rate: 40 * O(50) Million collisions/sec.
  - ***Detector***: a big camera. ~ e.g. LHC 1.5 MB/event (60 TB/s)
    - Pictures of long-lived decay products of short lived heavy/ interesting particles.
    - Sub-detectors parts: Tracking, Calorimeters, Muon system, Particle ID (e.g. Cherenkov, Time of Flight)
  - ***DAQ/Trigger:*** Hardware/software
  - ***Software:*** Reconstruction (Raw data -> particle "features") / Analysis
  - ***Computing:*** GRID Monarch Model "Cloud" Computing/Data Management (software/hardware)





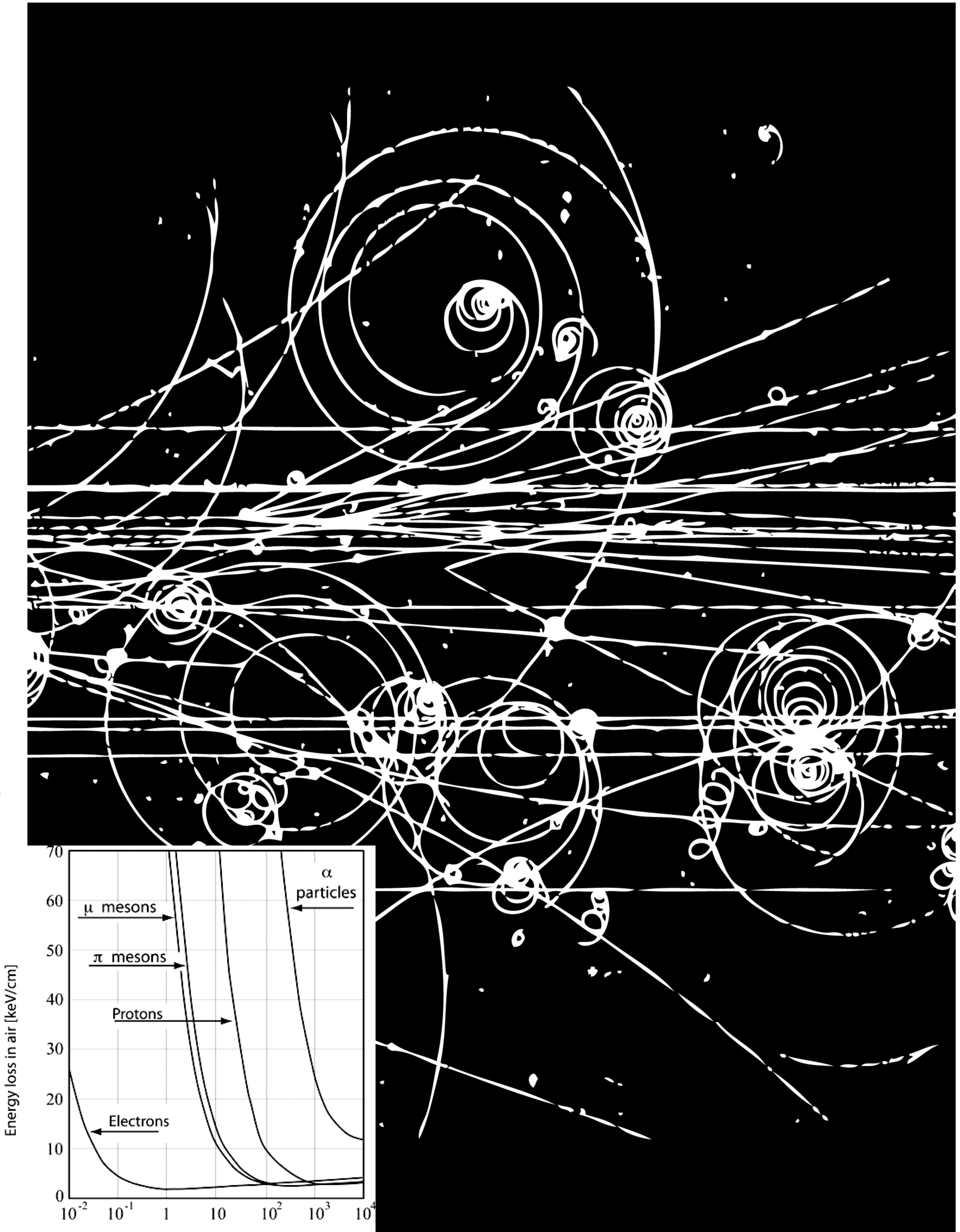




“Seeing”

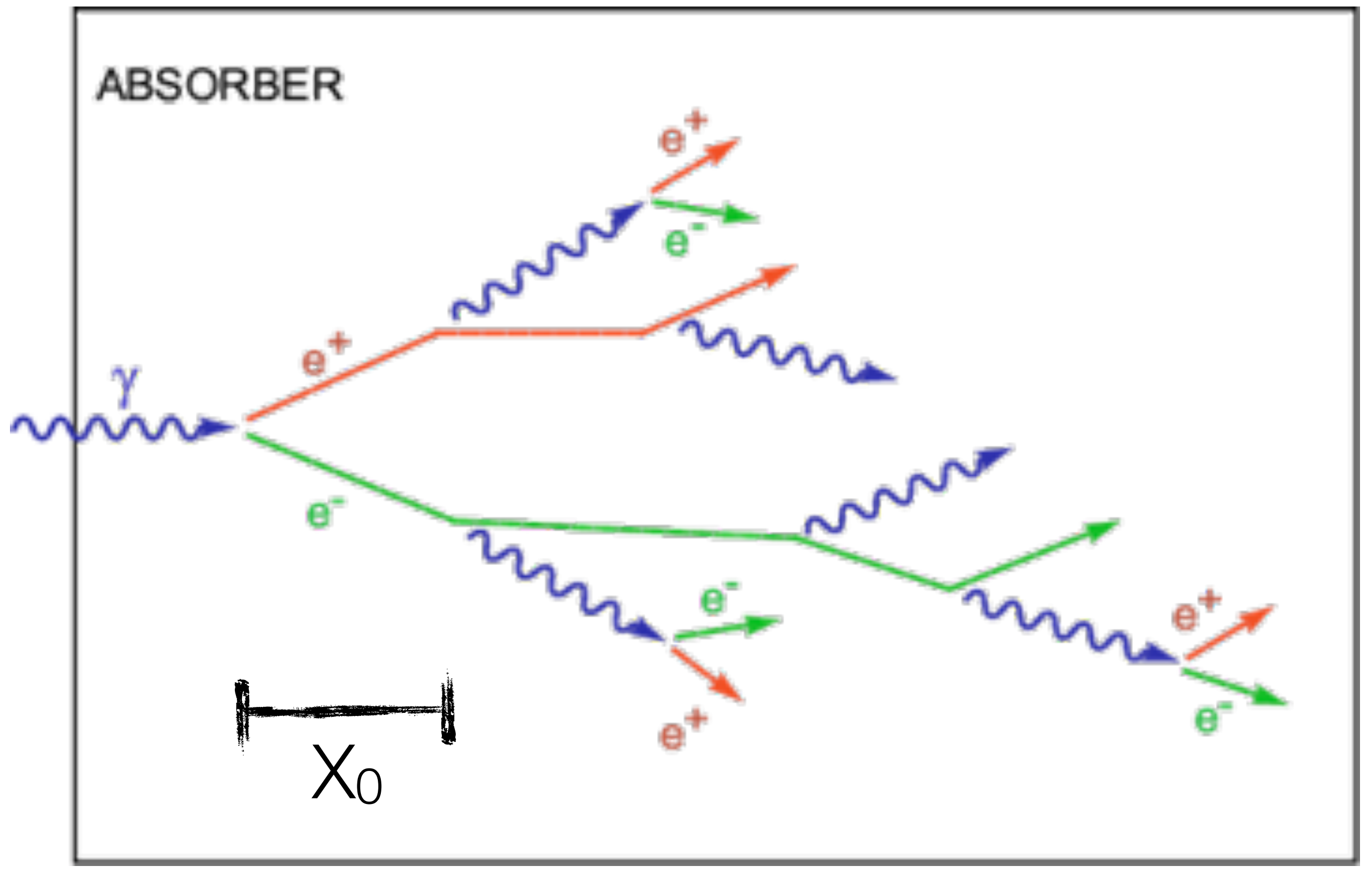
How do we “see” particles?

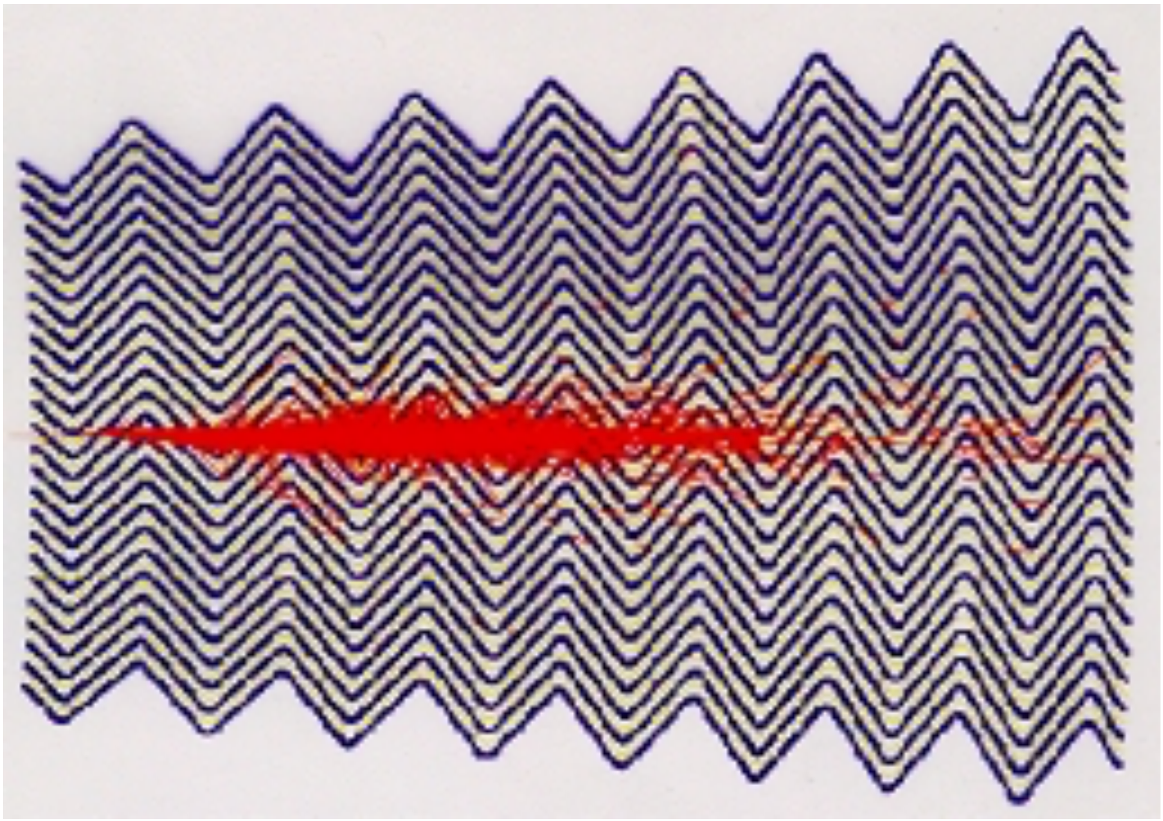
- ***Charged particles ionize media***
  - Image the ions.
    - In ***Magnetic Field*** the ***curvature*** of trajectory ***measures momentum***.
    - Momentum resolution degrades as less curvature: σ(p) ~ c p ⊕ d.
      - d due to multiple scattering.
  - Measure ***Energy Loss*** (~ # ions)
    - dE/dx = Energy Loss / Unit Length = f(m, v) = Bethe-Block Function
      - Identify the particle type
    - ***Stochastic process*** (Laudau)
  - Loose all energy ➝ range out.
    - Range characteristic of particle type.

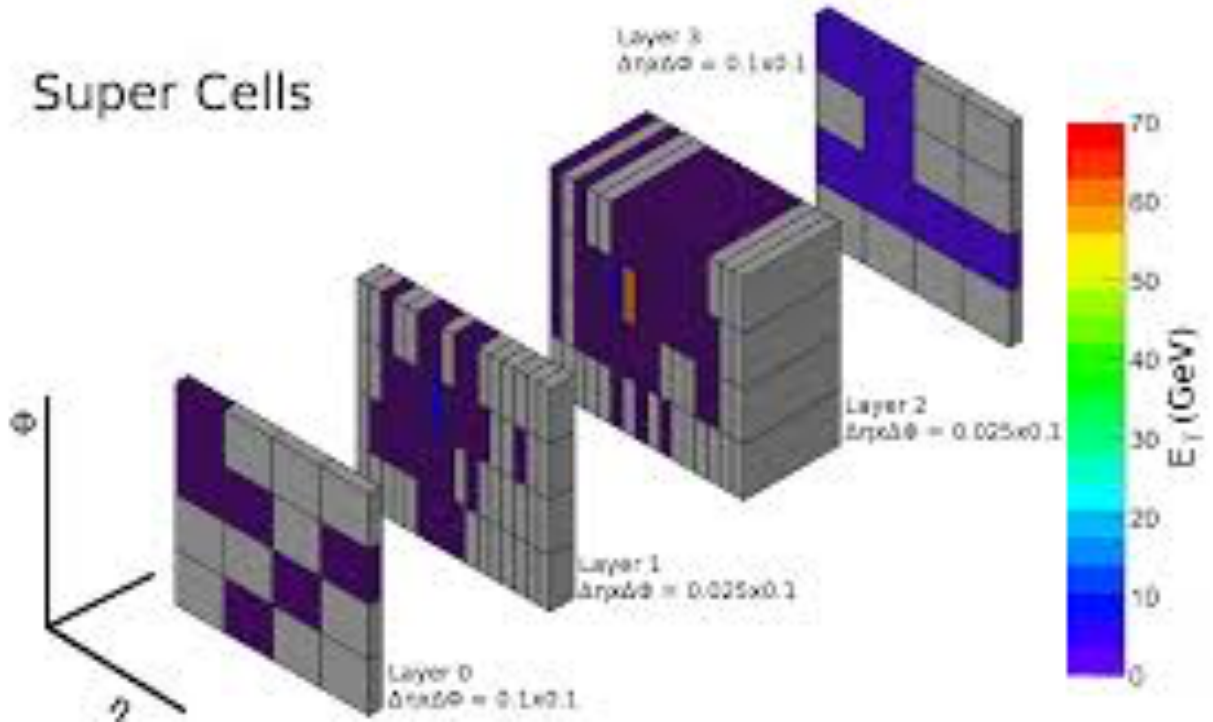


How do we "see" particles?

- Particles deposit their energy in a ***stochastic process*** know as ***"showering"***, secondary particles, that in turn also shower.
  - Number of secondary particles ~ Energy of initial particle.
  - Energy resolution improves with energy: $\sigma(E) / E = a/\sqrt{E} \oplus b/E \oplus c$.
    - a = sampling, b = noise, c = leakage.
  - Density and Shape of shower characteristic of type of particle.
- ***Electromagnetic calorimeter***: Low Z medium
  - ***Light particles***: electrons, photons, $\pi^0 \rightarrow \gamma\gamma$ interact with electrons in medium
- ***Hadronic calorimeters***: High Z medium
  - ***Heavy particles***: Hadrons (particles with quarks, e.g. charged pions/protons, neutrons, or jets of such particles)
    - Punch through low Z.
    - Produce secondaries through strong interactions with the nucleus in medium.
    - Unlike EM interactions, not all energy is observed.

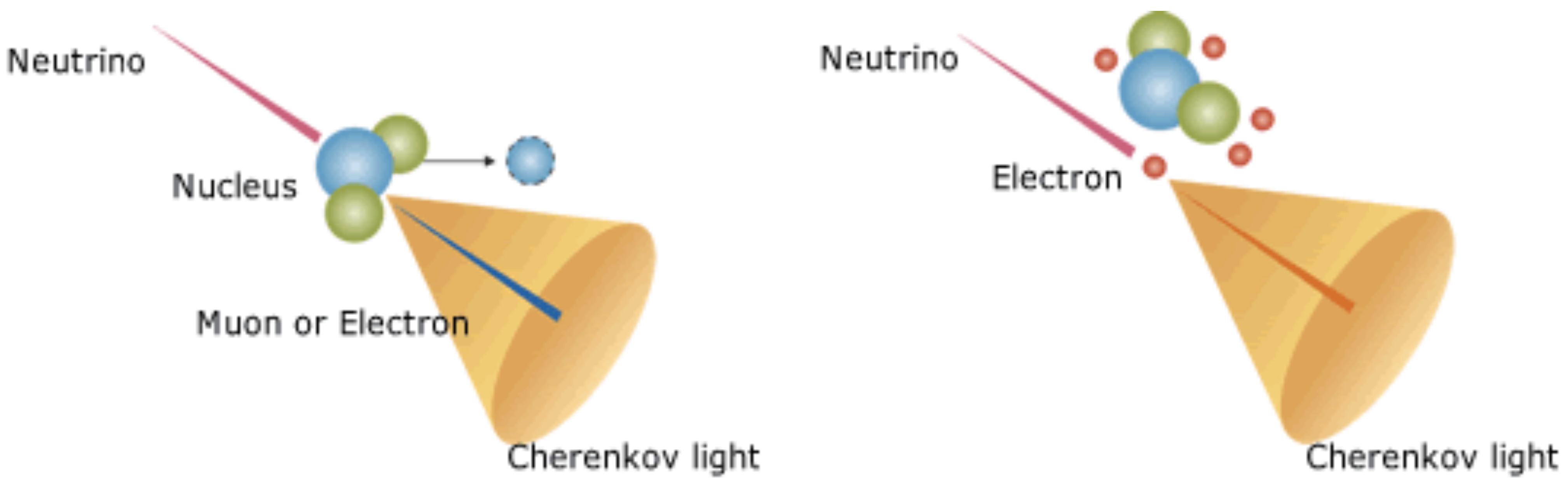




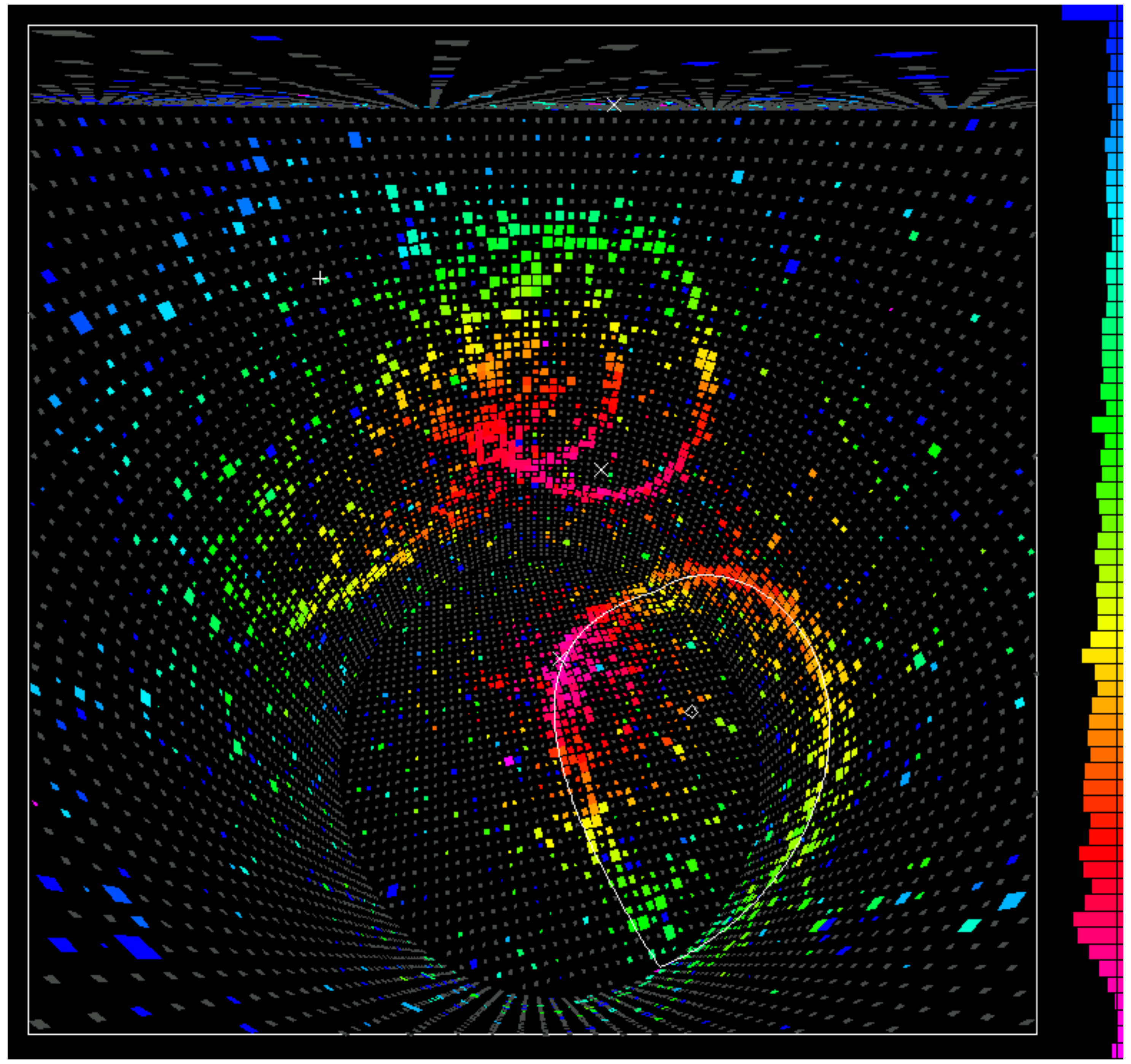


How do we "see" particles?

- Charged Particles traveling faster than speed of light in medium emit ***Cherenkov light*** (analogous to sonic boom).
  - Light emitted in cone, with angle function of speed and mass.
  - Depending on context, allow for particle identification and/or speed measurement.

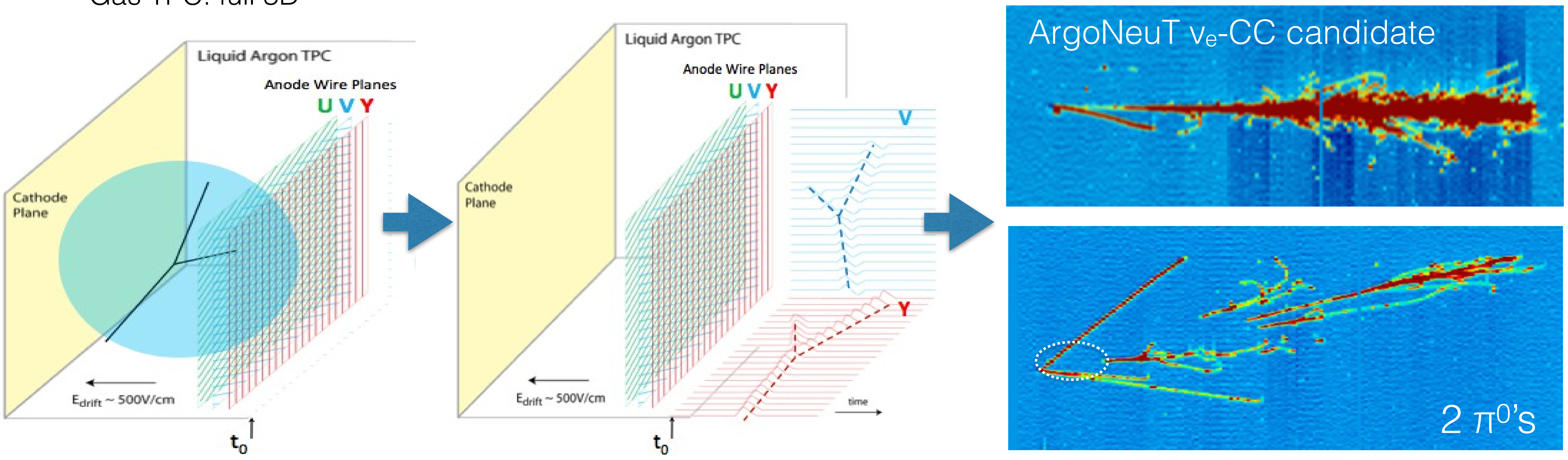


The generated charged particle emits the Cherenkov light.



Neutrino Detectors

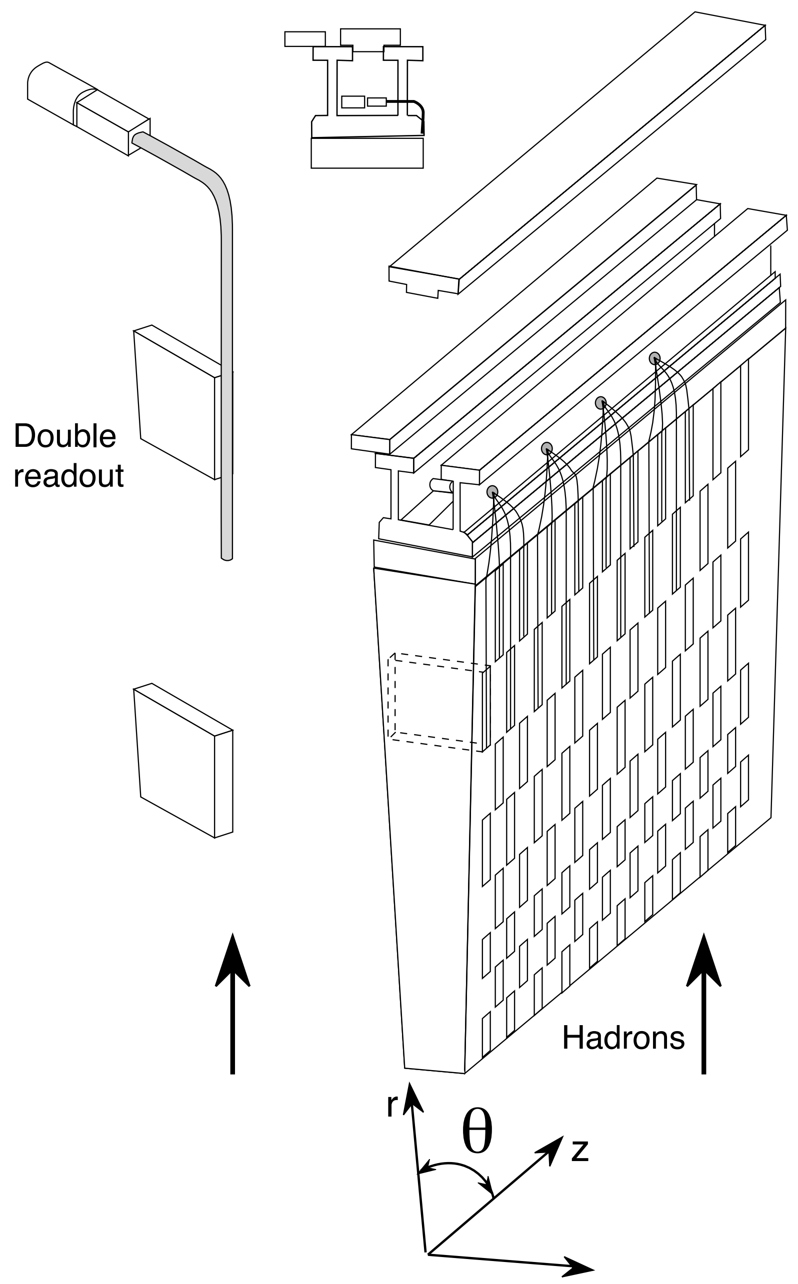
- ***Need large mass/volume*** to maximize chance of neutrino interaction.
- Technologies:
  - Water/Oil Cherenkov
  - Segmented Scintillators
  - ***Liquid Argon Time Projection Chamber: promises ~ 2x detection efficiency.***
    - ***Provides tracking, calorimetry, and ID all in same detector***.
    - Chosen technology for US's flagship LBNF/DUNE program.
    - Usually 2D read-out… 3D inferred.
  - Gas TPC: full 3D

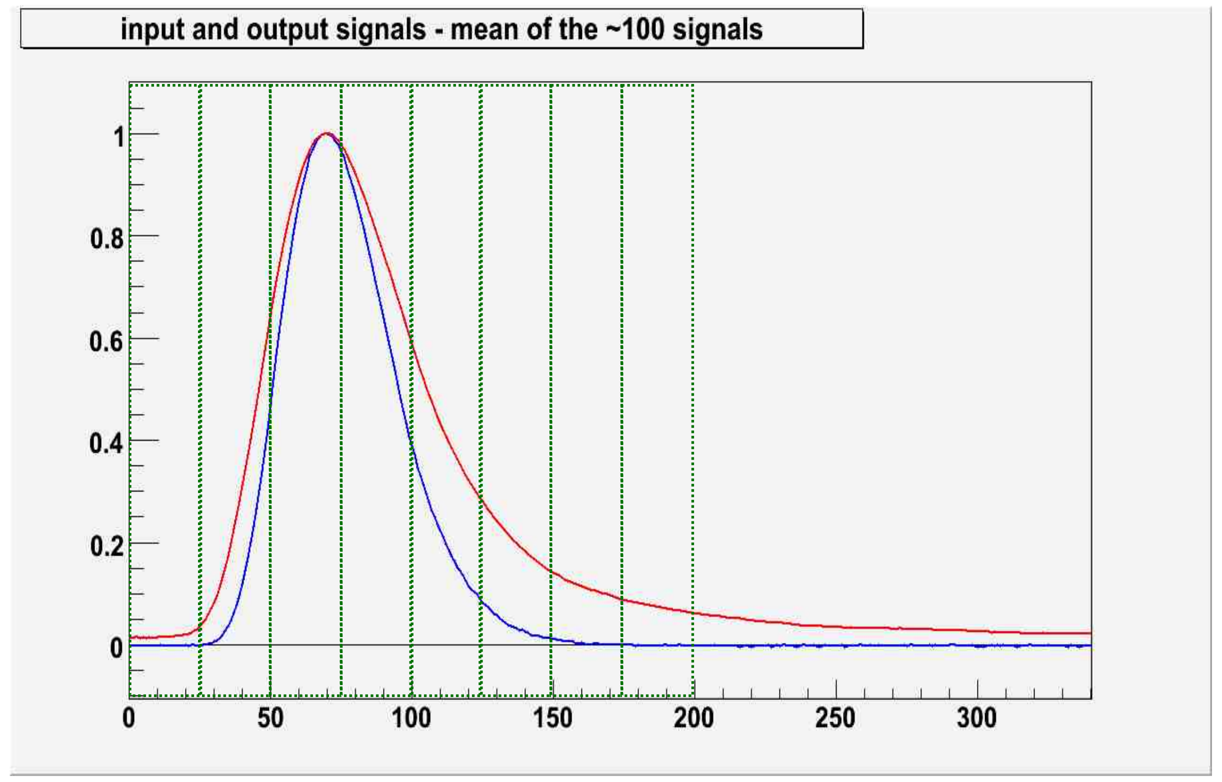


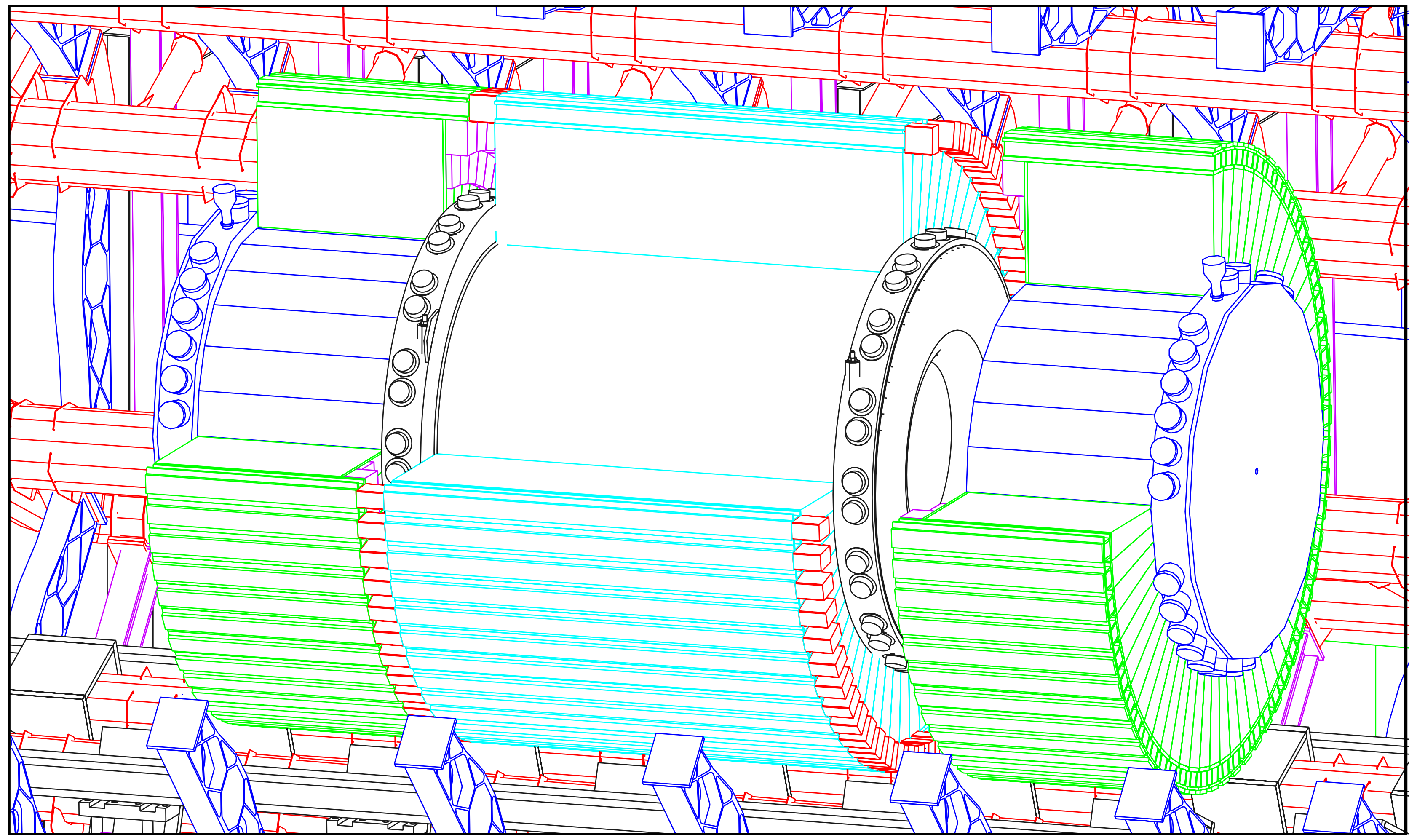
Data

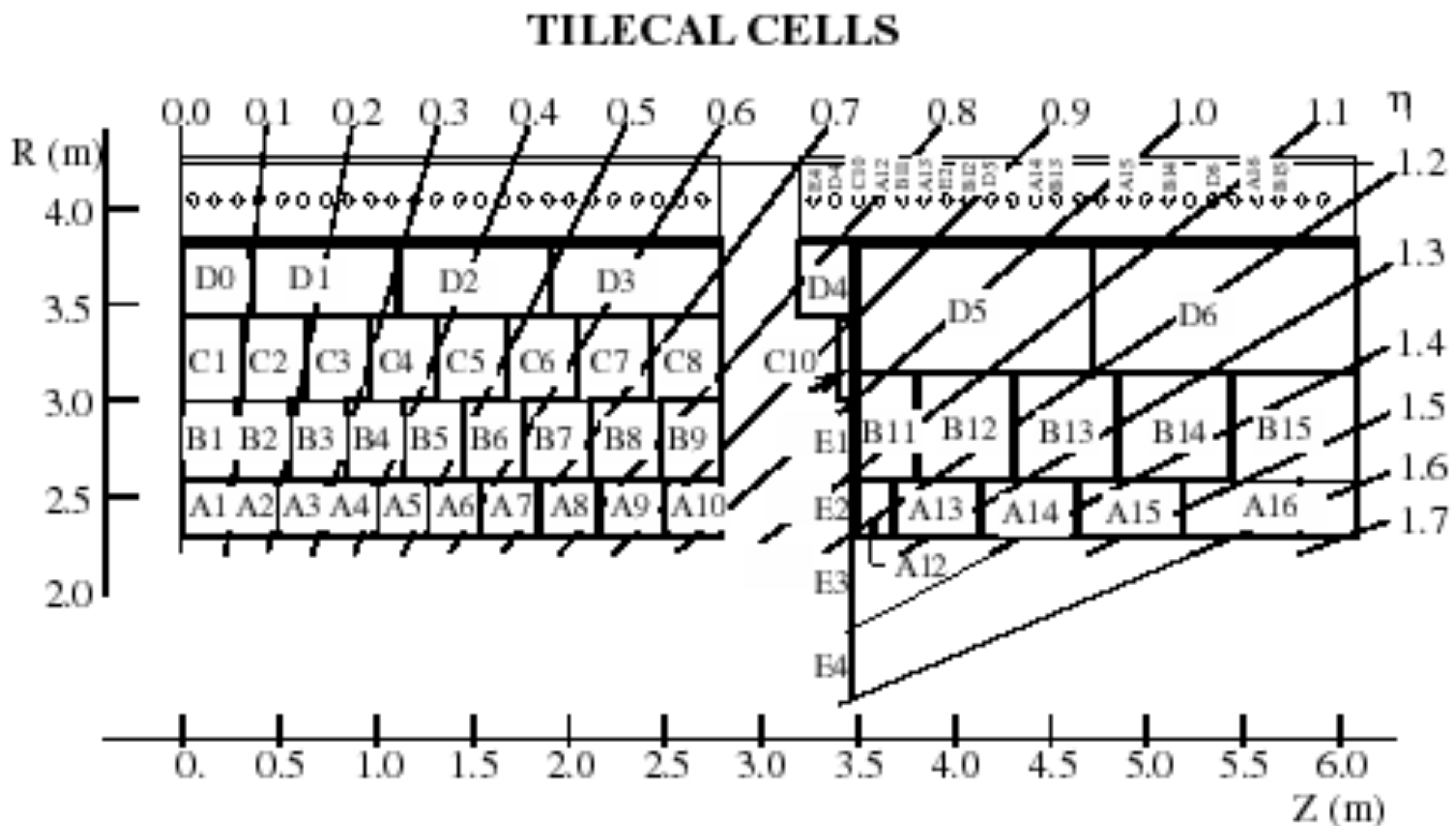
Example: Hadronic Calorimeter

- Steel and scintillating tiles, placed radially (opposed to longitudinal)
  - Segmented into cells. Projective geometry.
- Fibers are coupled radially to the tiles along the outside faces of each module (easier readout)
- A compact electronics "drawer" read-out is housed in the girder of each module
- Fibers read by Photomultiplier Tubes (PMTs)
- Projective Towers made of A,B,C,D layers



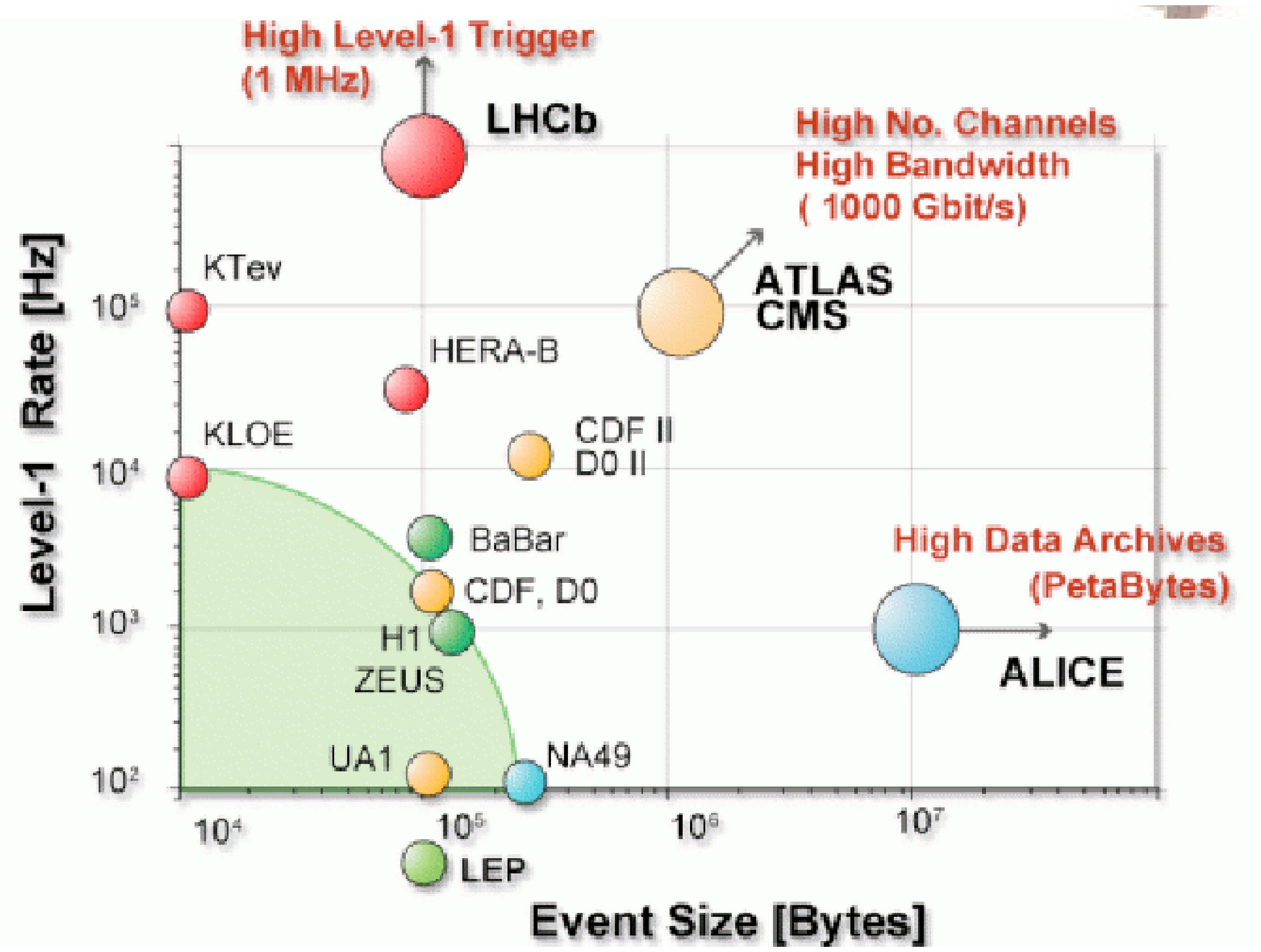






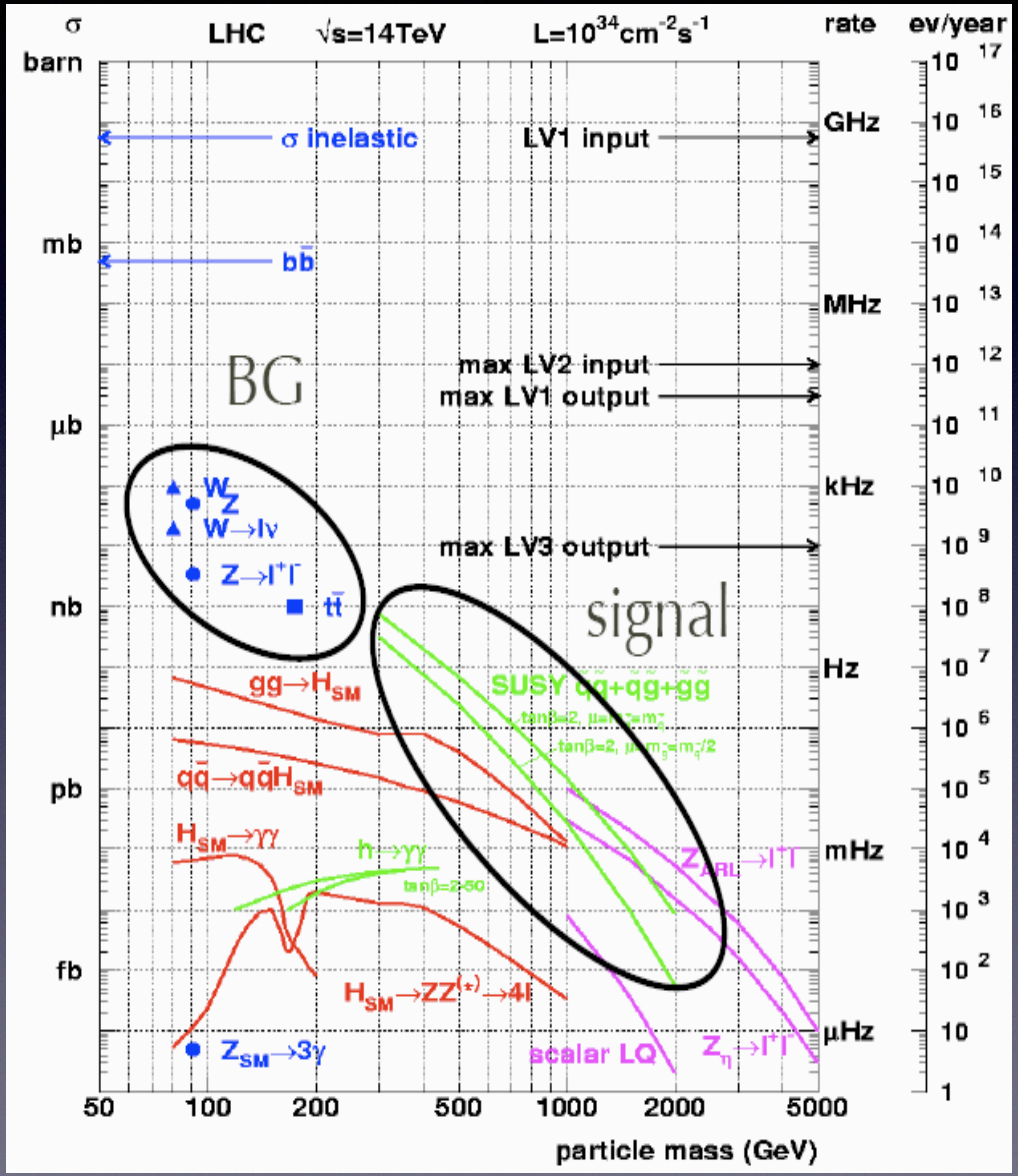
Back of the Envelope

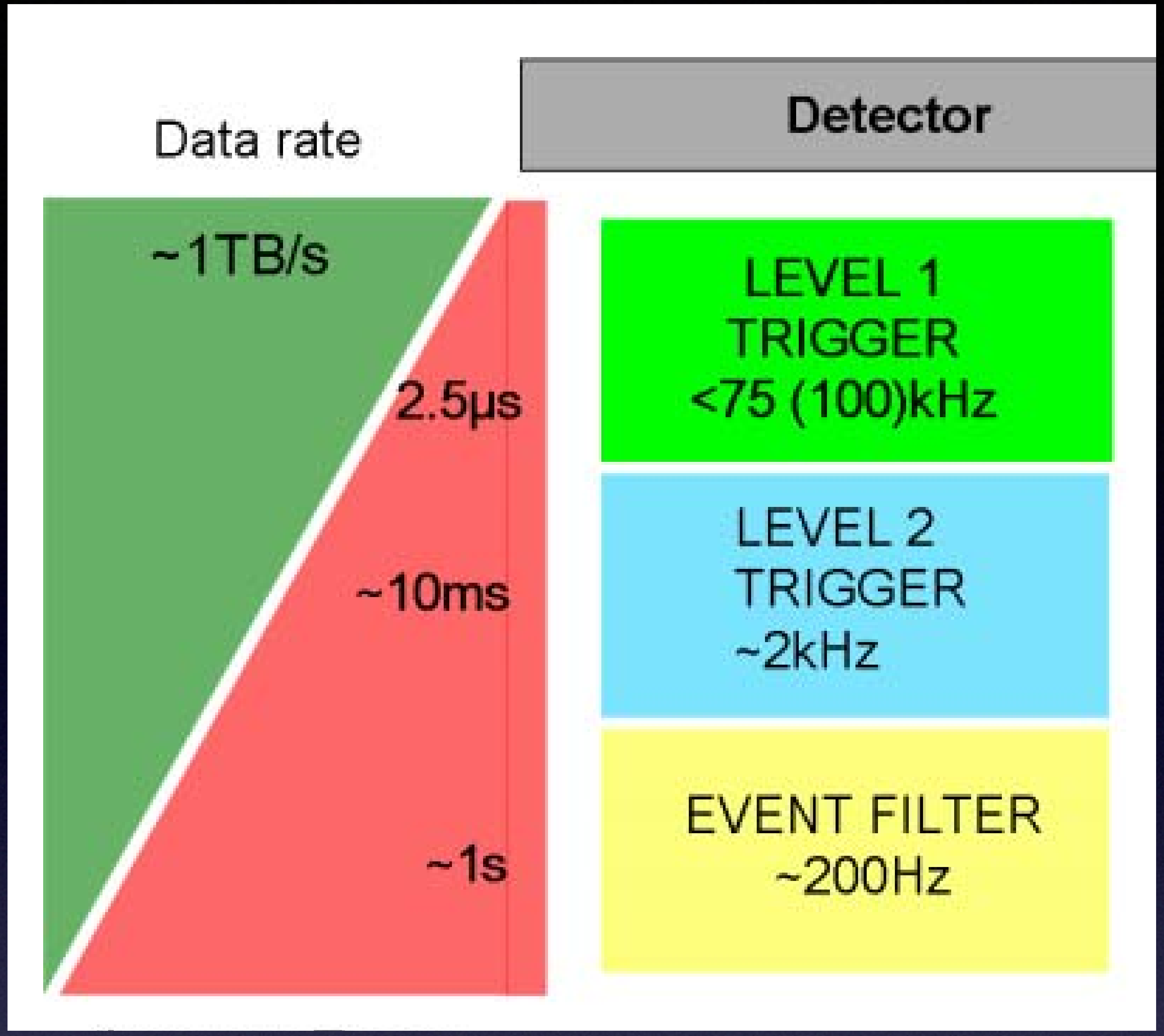
- LHC: 40 million collisions / sec at 1.5 MB/Event:
  - Network:
    - 60 million MB/s = 60,000 GB/s = 60 TB/s.
    - 10 Gb/s fiber = 1.25/s GB for 60,000 GB/s = 48,000 Fibers @ 1x1 mm^2 each = 12.3 cm radius.
    - 10 Gigabit ethernet standard in 2002... 1 Gigabit in 1998 (1.23 m radius)
    - Actually, the hardware to support this much data would would be huge.
  - Disk:
    - 60 TB/s = 20 * 3 TB HDs/s = about 175 million HDs / year.
  - Processing:
    - 1 CPU sec/event @ 40 Million events/s = 40 million CPUs

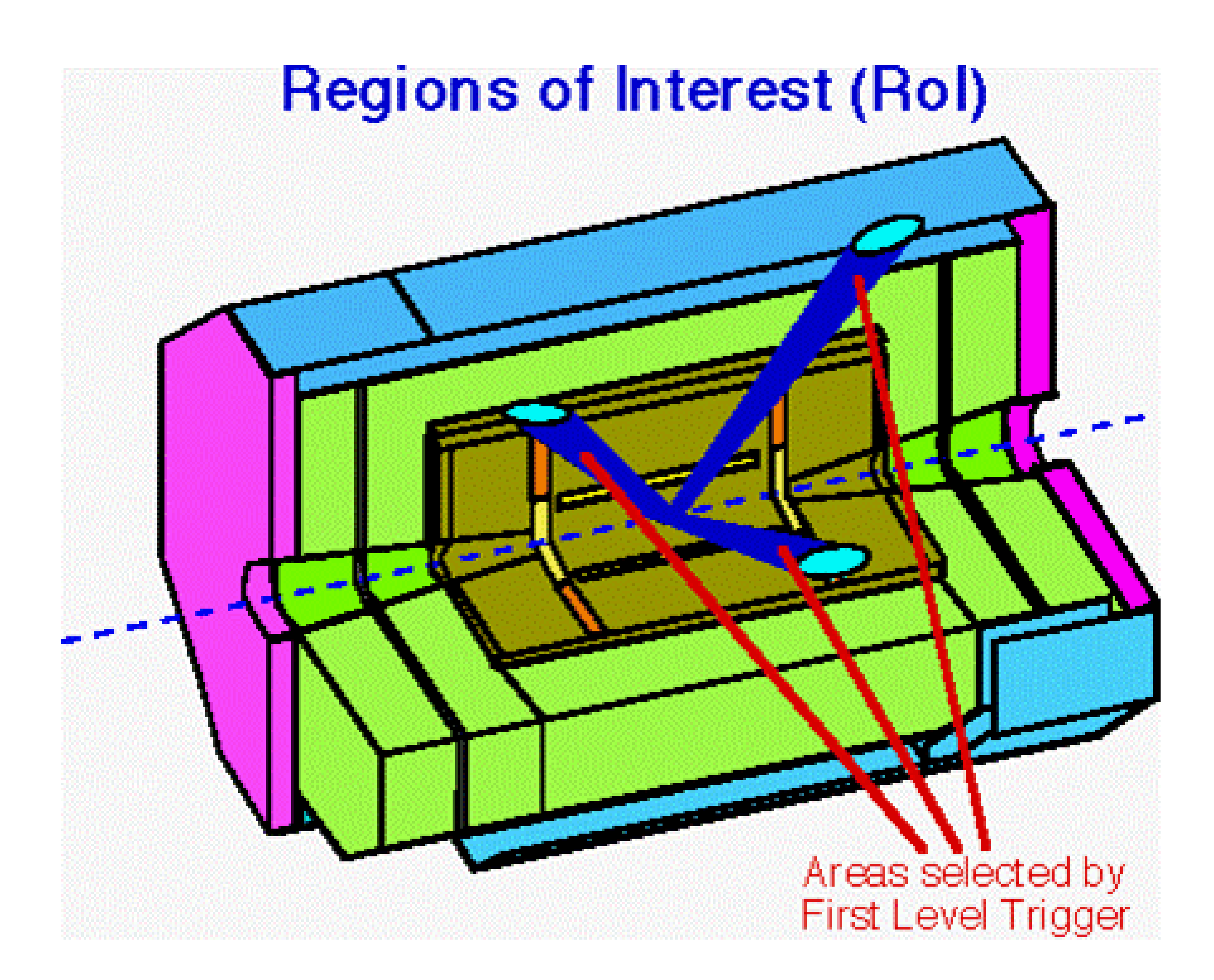


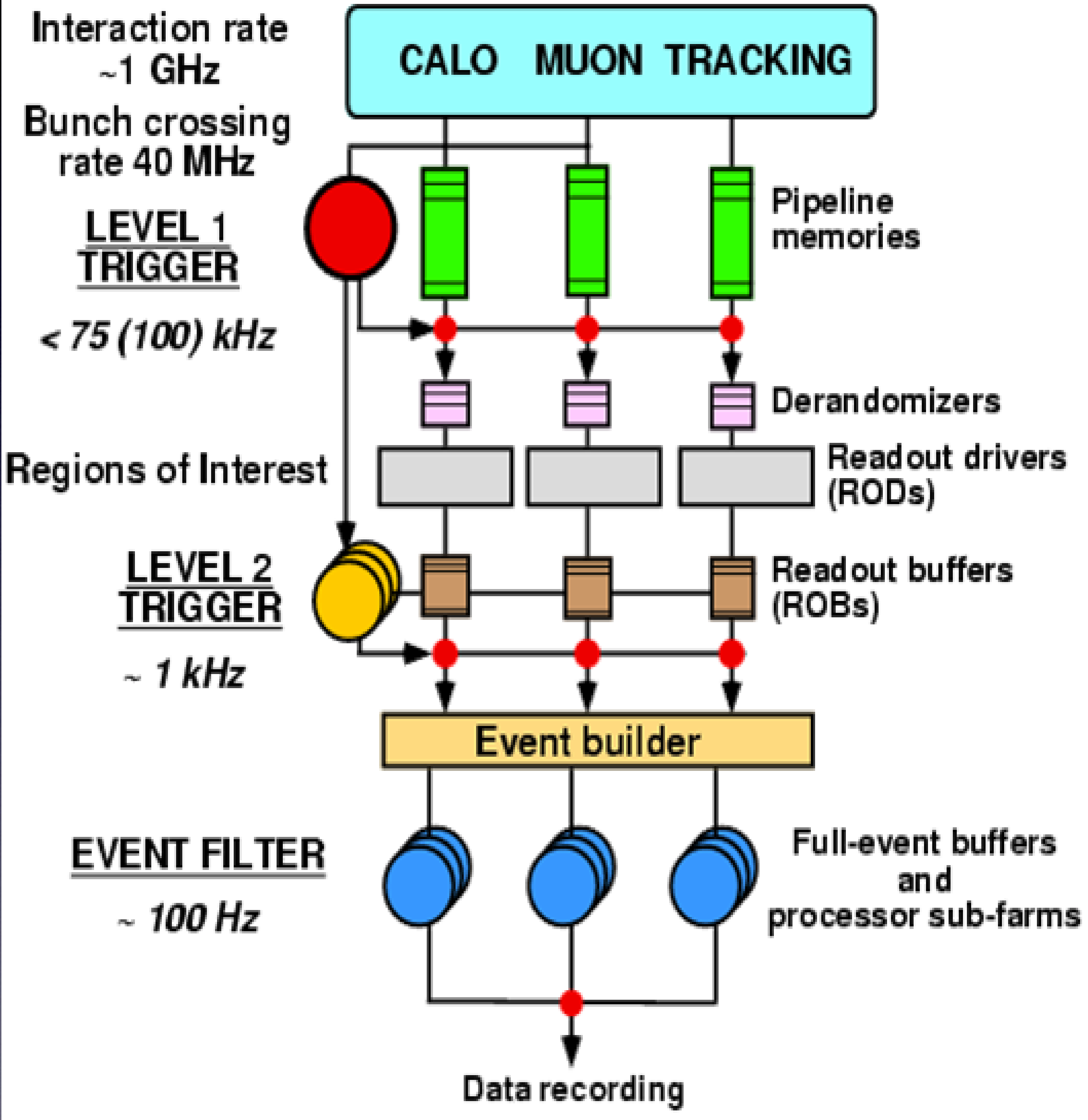
LHC Particle Factory

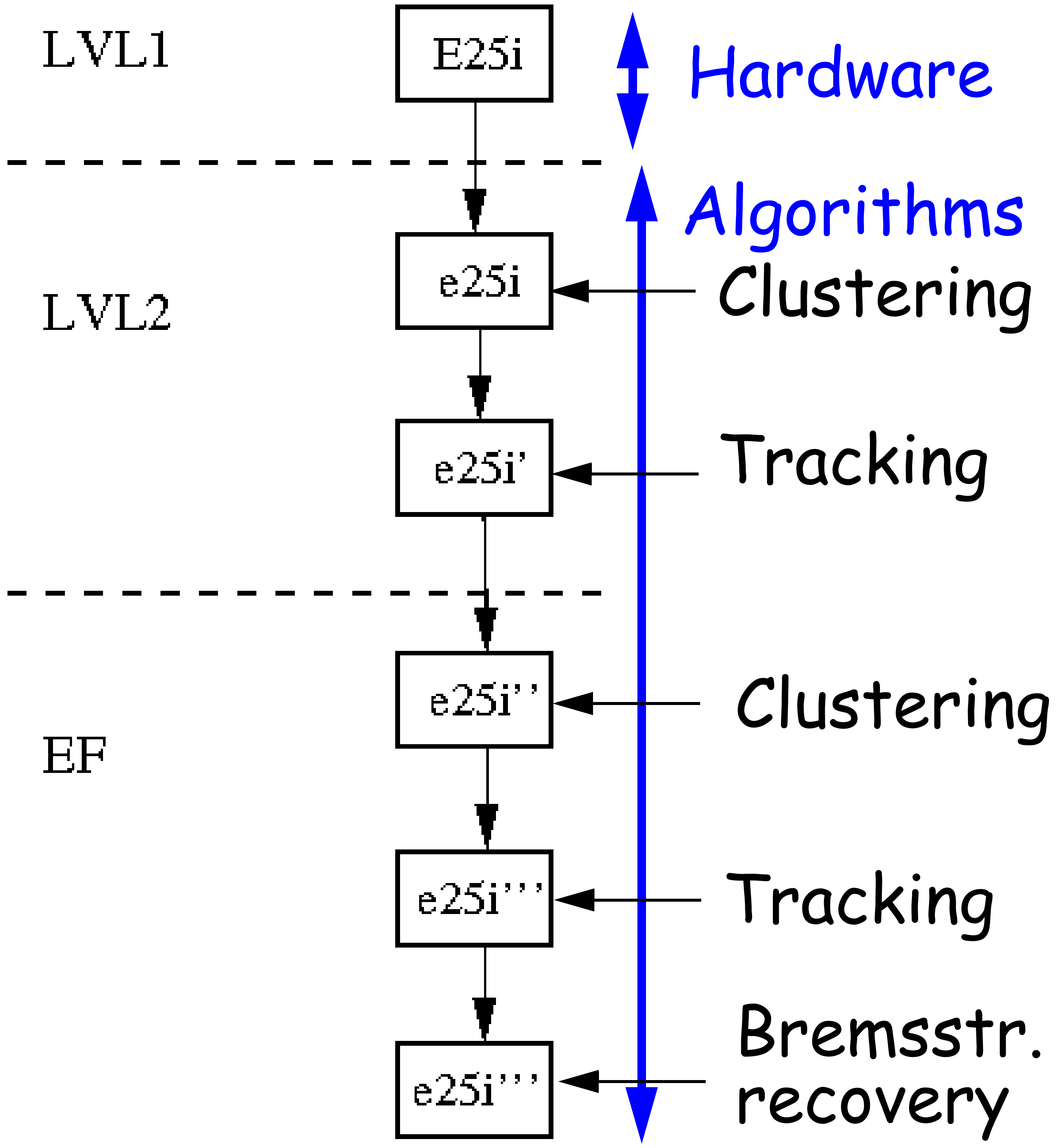
- At L=10^{34} cm^{-2} s^{-1}
  - $W \to l\nu$, $Z \to ll$ ~10^2 Hz
  - top at 10 Hz
  - Higgs at ~1 Hz
  - SUSY up to 10 Hz (depending on scale)
- Currently ~50 simultaneous interactions



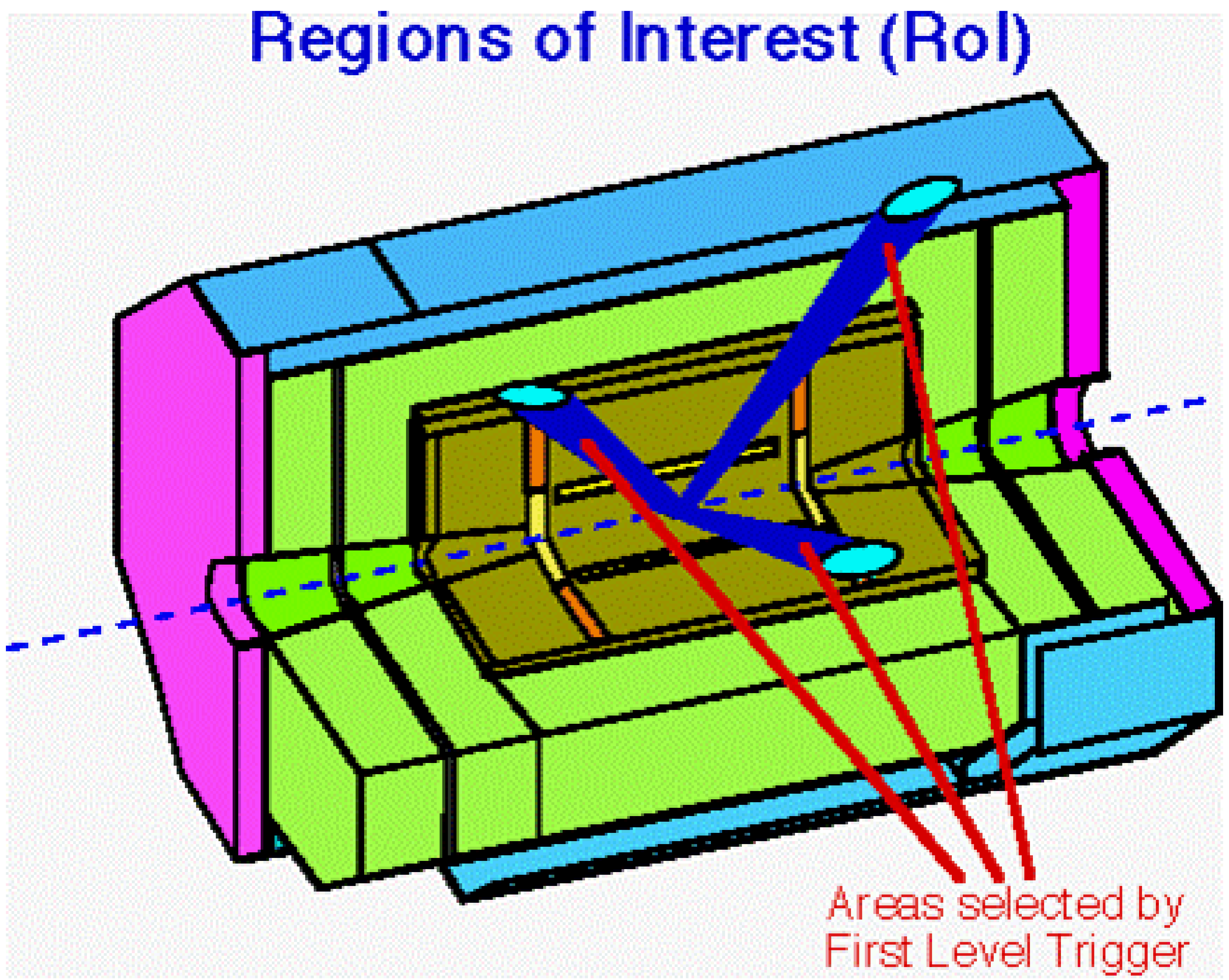

Data rate
Detector
~1TB/s
2.5μs
~10ms
~1s
LEVEL 1
TRIGGER
<75 (100)kHz
LEVEL 2
TRIGGER
~2kHz
EVENT FILTER
~200Hz


Regions of Interest (RoI)
Areas selected by
First Level Trigger

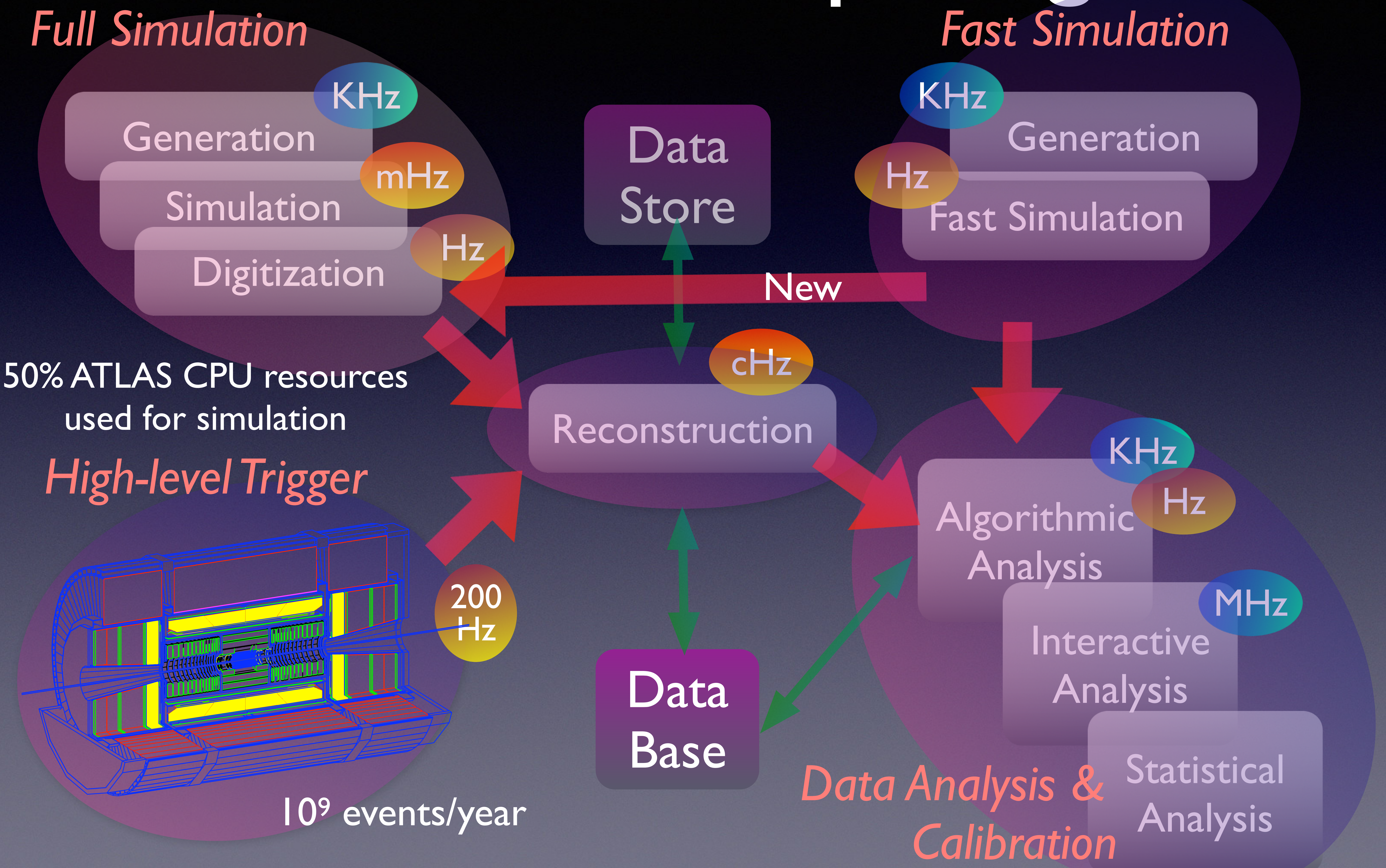

Interaction rate
~1 GHz
Bunch crossing
rate 40 MHz
CALO MUON TRACKING
LEVEL 1
TRIGGER
< 75 (100) kHz
Pipeline
memories
Derandomizers
Regions of Interest
Readout drivers
(RODs)
LEVEL 2
TRIGGER
~ 1 kHz
Readout buffers
(ROBs)
Event builder
EVENT FILTER
~ 100 Hz
Full-event buffers
and
processor sub-farms
Data recording



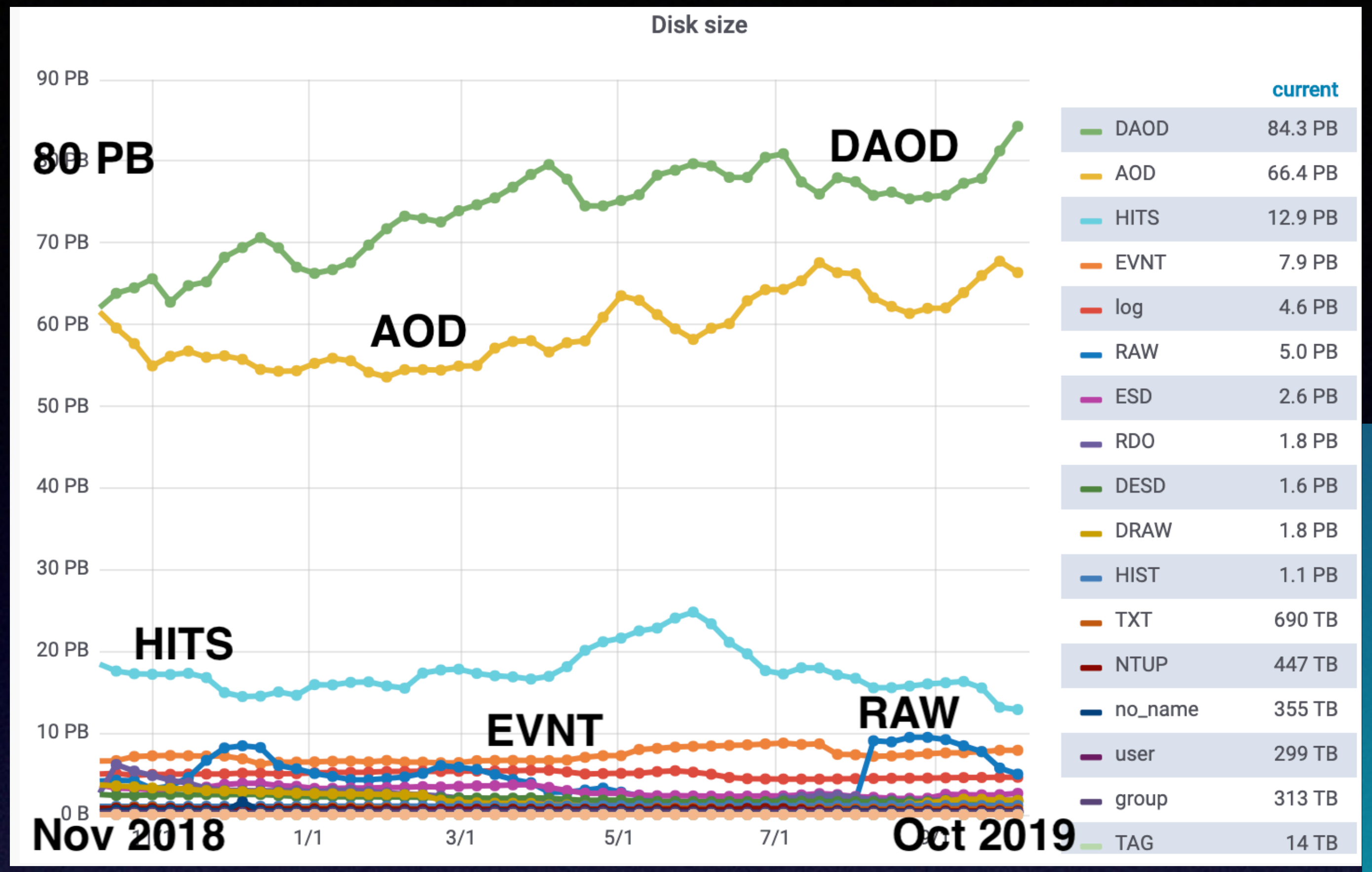
-> Chain Concept -> Fast Reject



ATLAS Computing



The Event Data Model



Raw Channels.
1.6 MB/event.

Raw Data Objects

Reconstruction Output.
Intended for calibration.
1000 KB/event.
Cells, Hits, Tracks,
Clusters, Electrons, Jets, ...

Event Summary Data

Data refinement

Analysis Object Data

Intended for Analysis.
~500 KB/event.
"Light-weight" Tracks,
Clusters, Electrons, Jets,
Electron Cells, Muon
HitOnTrack,...

Derived Physics Data

Intended for "interactive"
Analysis.
~10-50 KB/event.
What-ever is necessary for
a specific analysis/
calibration/study.

The Computing Model

- Resources Spread Around the GRID

•Interactive Analysis
•Plots, Fits, Toy MC, Studies, ...

Tier 3

DPD

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD
•Disk Access: AOD, fraction of ESD

Tier 2

AOD

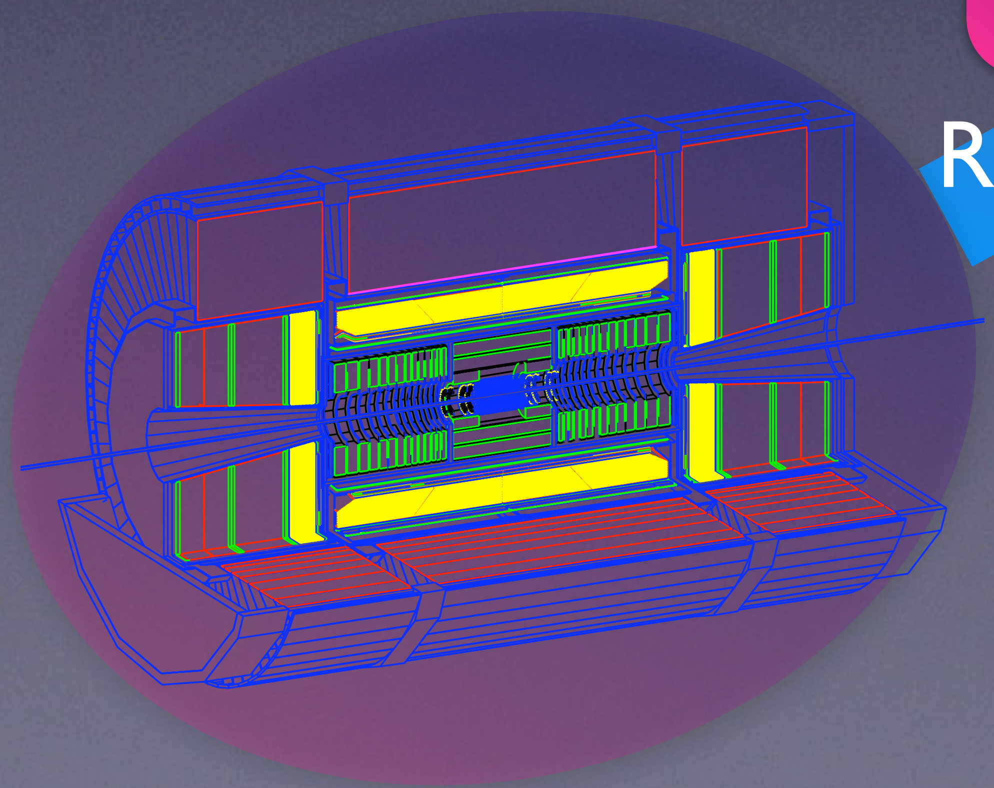
•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

Tier 1

RAW/
AOD/
ESD

Tier 0

•Production of simulated events.
•User Analysis
•Disk Store



RAW

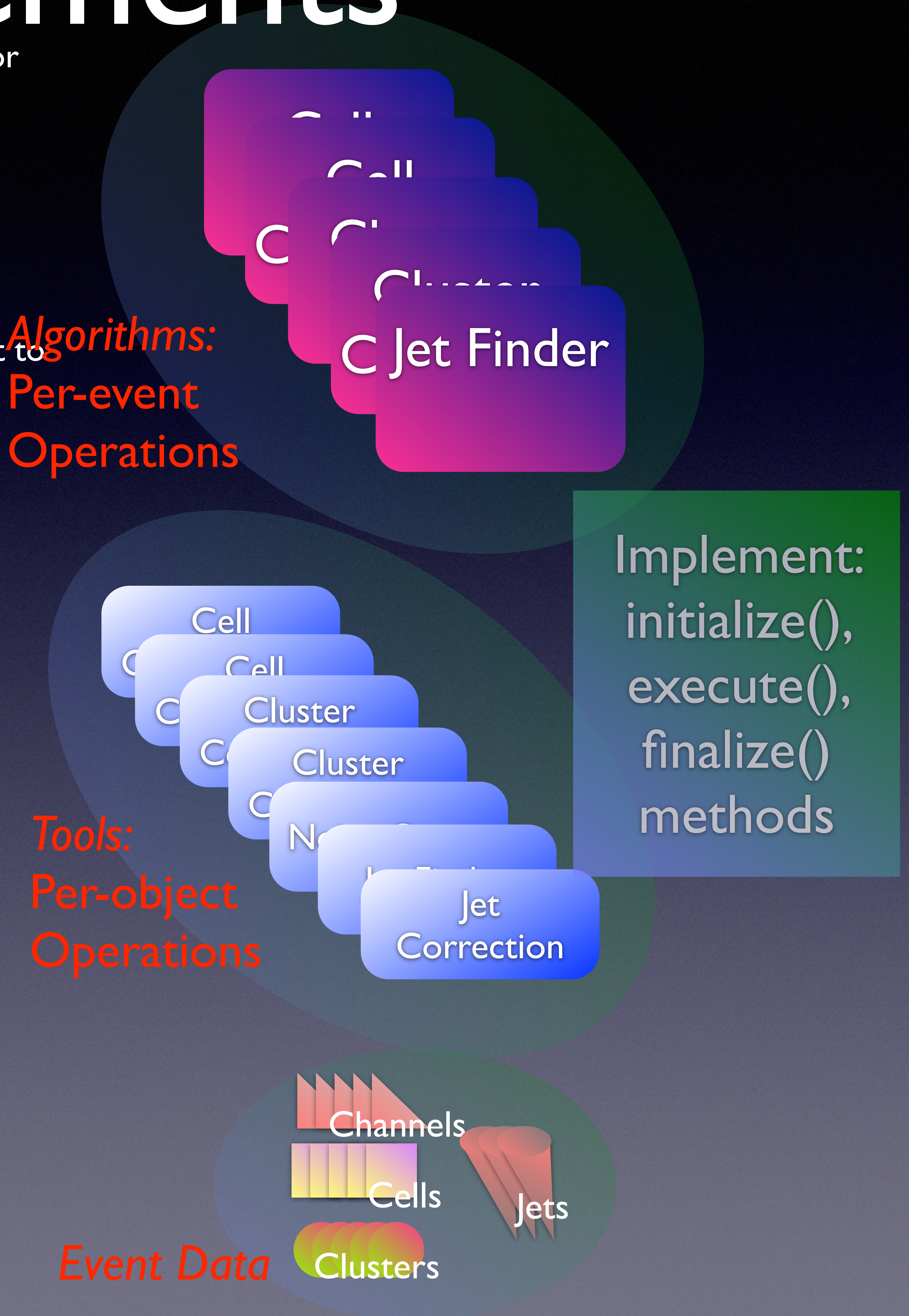
CERN
Analysis
Facility

•Primary purpose: calibrations
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

ATLAS Software Fundamentals

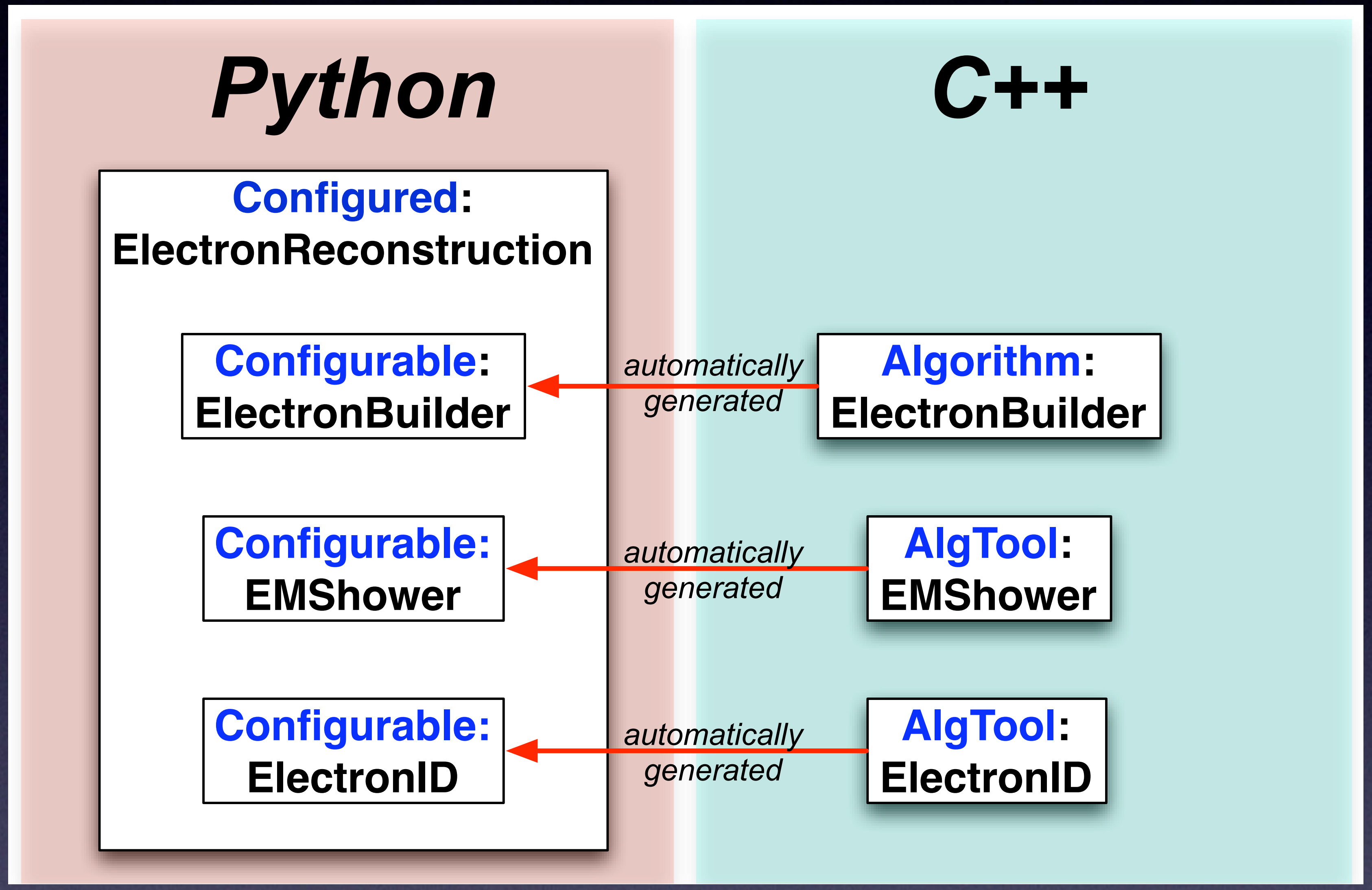
Framework Elements

- Athena is an extended version of LHCb's Gaudi framework used for high-level trigger, simulation/reconstruction, and analysis.
- Principles... separation of:
  - Data and algorithms
  - Transient (in memory) and Persistent (on disk) data (in contrast to CMS)
- Elements:
  - *Algorithms*- one execute per event, managed by framework.
  - *Tools*- multiple executes per event.
  - Event Data
  - Services
    - StoreGate- Transient Data Store- Mechanism for communication between Algorithms
    - Tool Service- Tool Factory
    - Interval of validity
    - Histogram Service
    - POOL- Persistency



Configuration

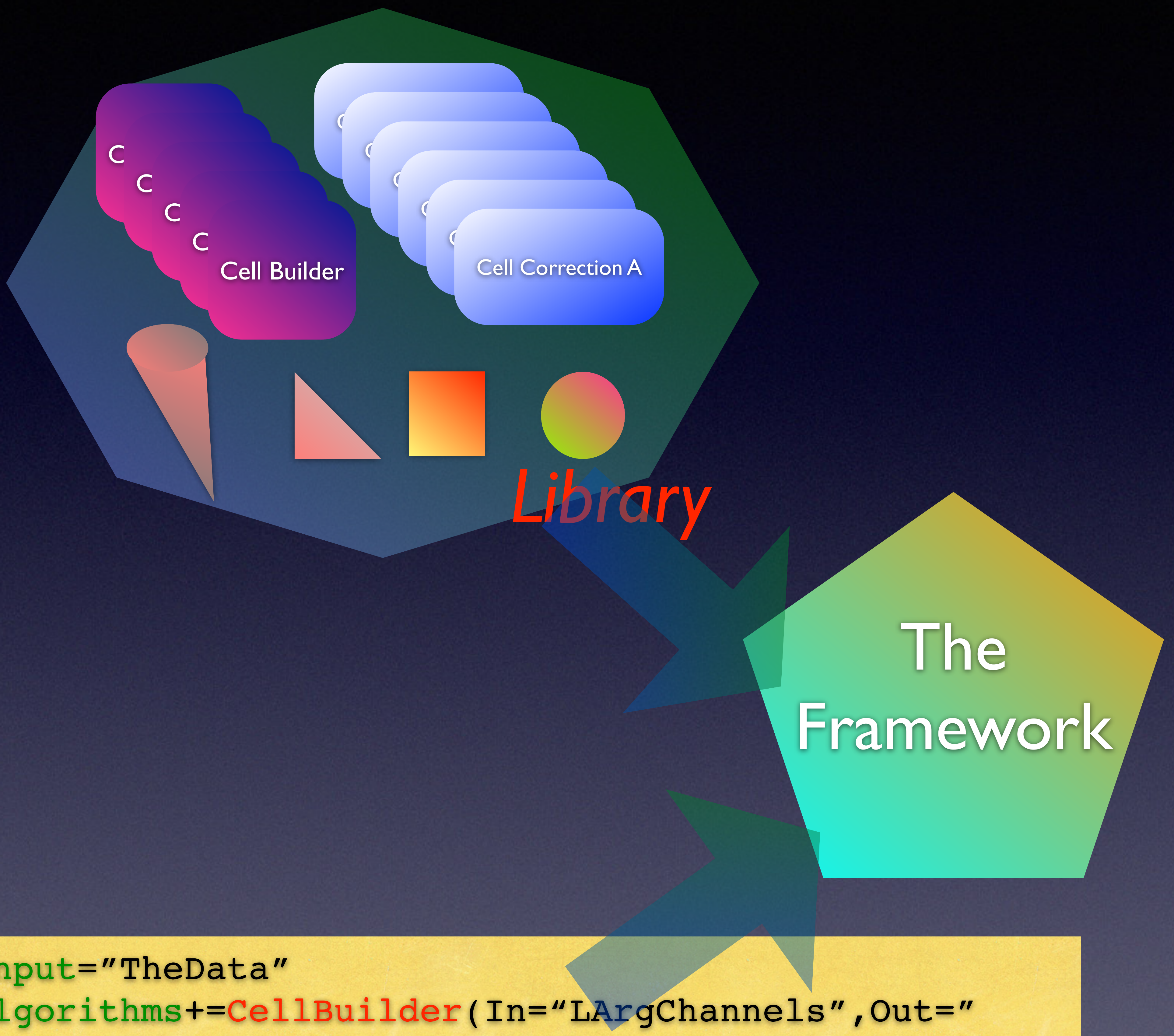
- Framework elements (eg Algorithms, Tools, Services) declare properties which can be set at runtime



- Application defined in python:
  - Load libraries
  - Instantiate tools/algs, configure properties
  - Define input/output

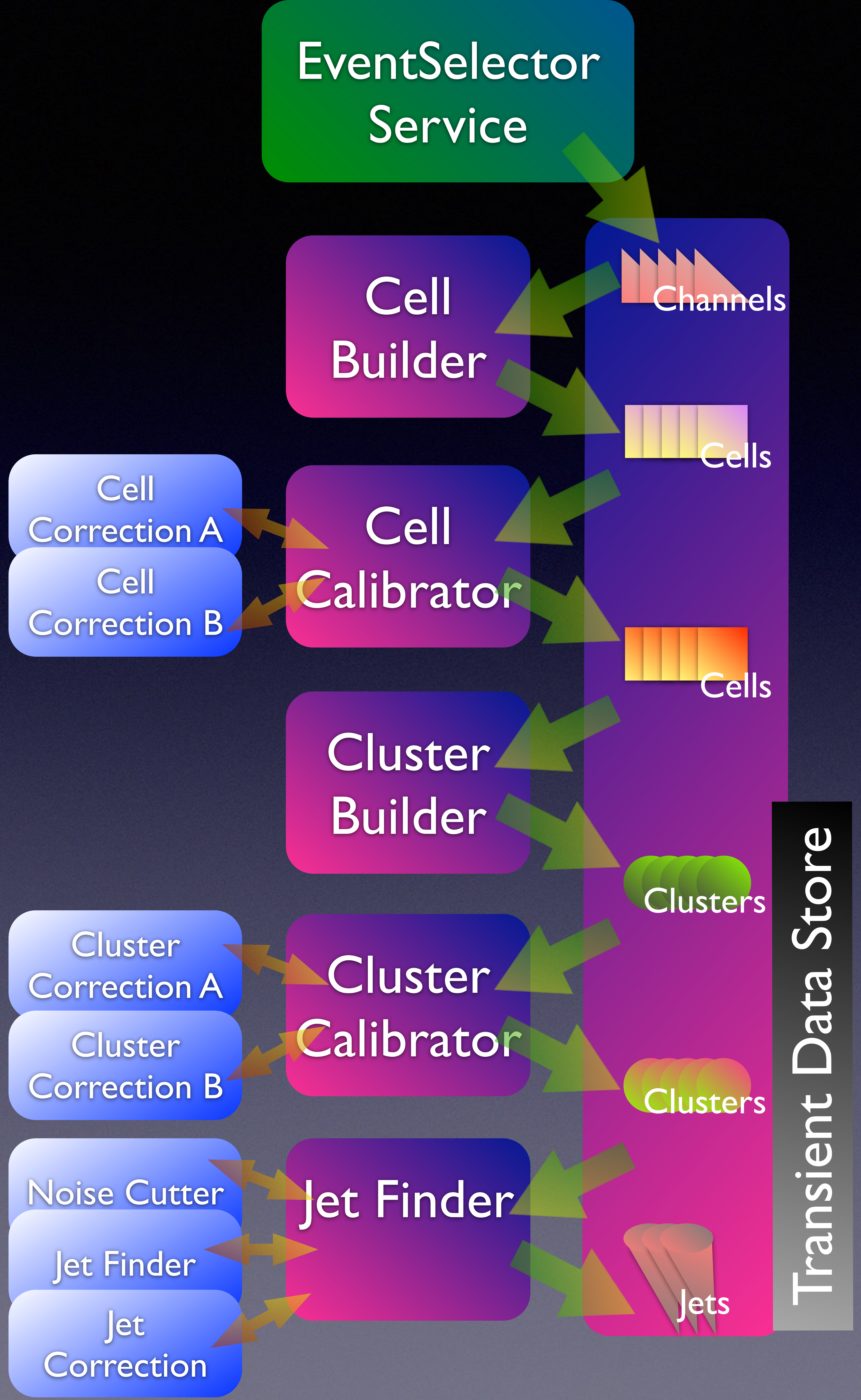
- Configurables:
  - Auto-generated python reflection of C++ components
  - Build configuration purely in python, persistify the configuration, build application later.
  - Build higher level abstractions in python

- Any application (eg reconstruction) is a specific configuration of a library of framework elements.




```
Input=”TheData”
Algorithms+=CellBuilder(In=“LArgChannels”,Out=”
Cells1”)
Algorithms+=CellCalibrator(In=“Cells1”,Out=”
Cells2”)
CellCalibrator+=CellCorrectionA()
CellCalibrator+=CellCorrectionB()
Algorithms+=ClusterBuilder(In=“Cells2”,Out=”
Clusters1”,MinEnergy=10*GeV)
....
```

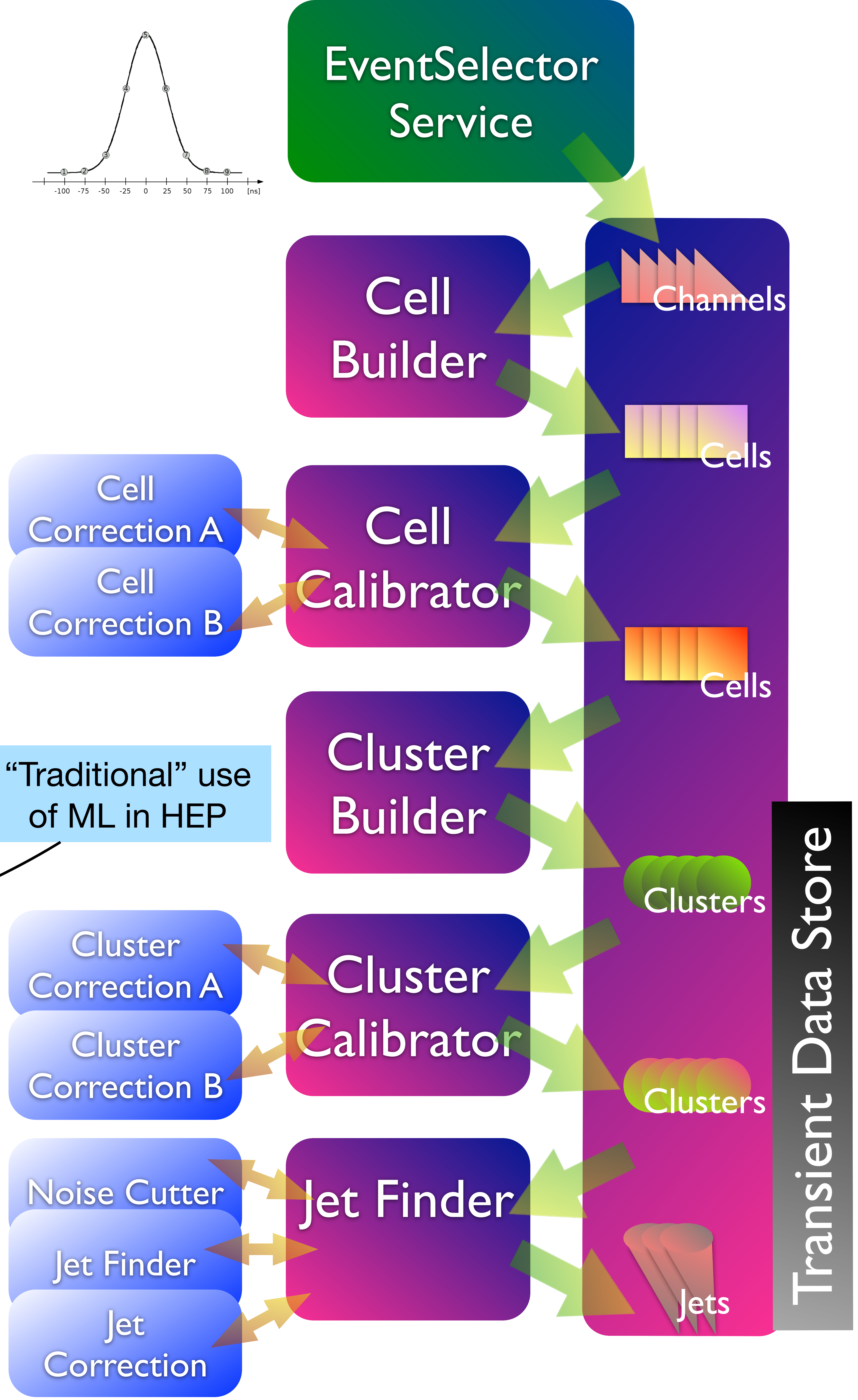
A Configuration



HEP Data

- The lowest-level (raw) data we generally have are the digitized outputs of detectors… e.g. voltages.
- Reconstruction is a series of sequential algorithms that construct features from outputs of the previous algorithm.
    - “raw” ⟶ “features”
- Highest level of Reconstruction output is usually particle candidates.
- Analysis, usually
    - choosing candidates ⟶ 4-vectors, separated by PID
    - 4-vectors ⟶ kinematic features (e.g. masses)
    - kinematic features ⟶ signal/background
    - statistical analysis ⟶ hypothesis test, limits, measurements
        - Background estimation
        - Lots of systematics

“Traditional” use of ML in HEP



HEP Data Analysis (Very Simplified)

- Lets say you are looking for the Higgs decaying to 2 photons.
  - Rare process. Large backgrounds
    - reducible: e.g. not really 2 photons, but look like it.
    - irreducible: e.g. really 2 photons, but not from Higgs.
- Data: 40M events/sec → 1000 events/sec stored by trigger → AOD format
- Simulation: Large samples of "signal" events +
- Derived Physics Data (DPD) (Reduce data size)
  - Select stored events with 2 energetic photons
  - Store only relevant information
- Compute features (invariant mass, angles, …)
- Select subset of events very likely to be Higgs (appropriate features) and not background.
  - Compute efficiency of selection.
- Use Statistical Fitting techniques to determine number of Higgs events in sample.
- Divide by efficiency → # of Higgs → 2 photon decays produced in LHC.