Computing in High Energy Physics

John Apostolakis

SoFTware for Physics Group,

EP Dep, CERN

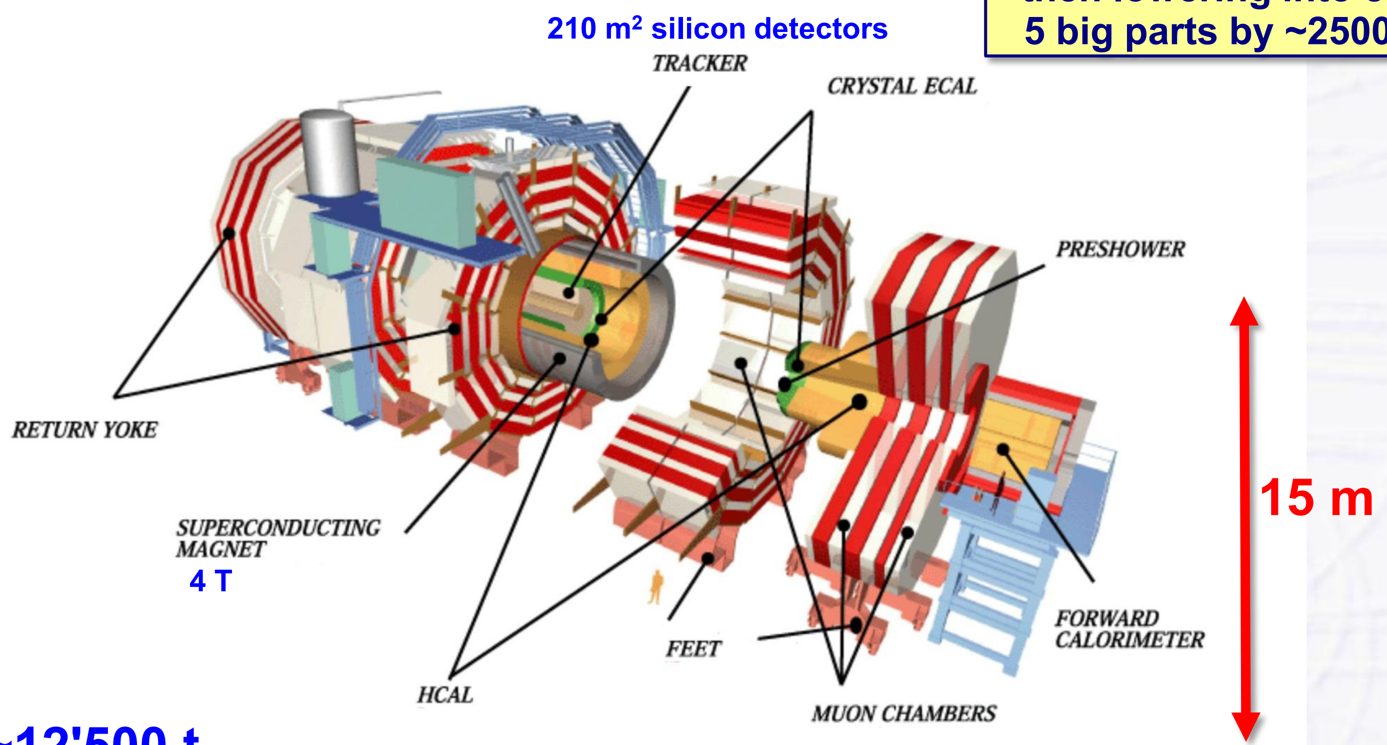
V1.0
2016.02.10

John.Apostolakis@cern.ch

Outline

- Uses of Computers
  - Data Acquisition – record
  - Reconstruction: Online, and off-line
  - Simulation
  - Data analysis
- Size of challenge
  - the GRID solution and its other applications

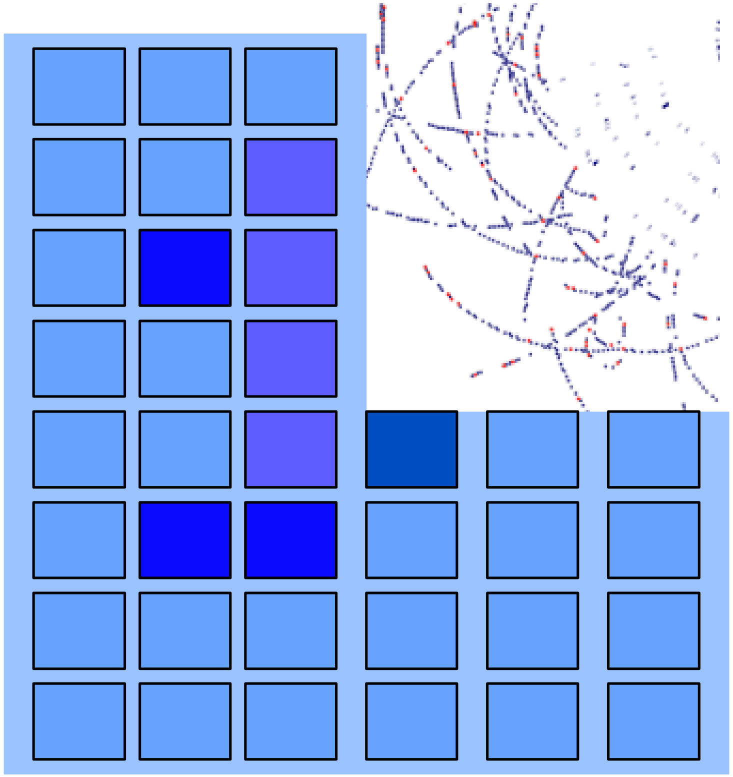
main assembly on surface, then lowering into cavern in 5 big parts by ~2500 t crane



~12'500 t

Data Acquisition (DAQ)

- Convert analog electronic signals into digital data
- Trigger – decision to record
  - Find interesting coll.
  - Assess – do they meet selection criteria

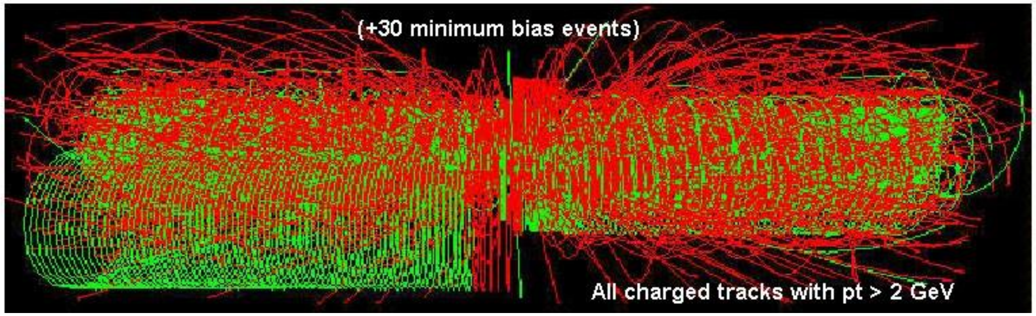


Reconstruction

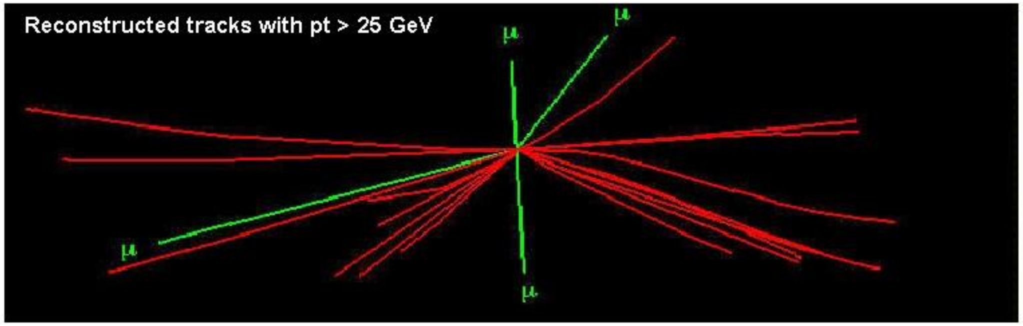
A lightning introduction

The Reconstruction challenge

Starting from this event



Looking for this "signature"



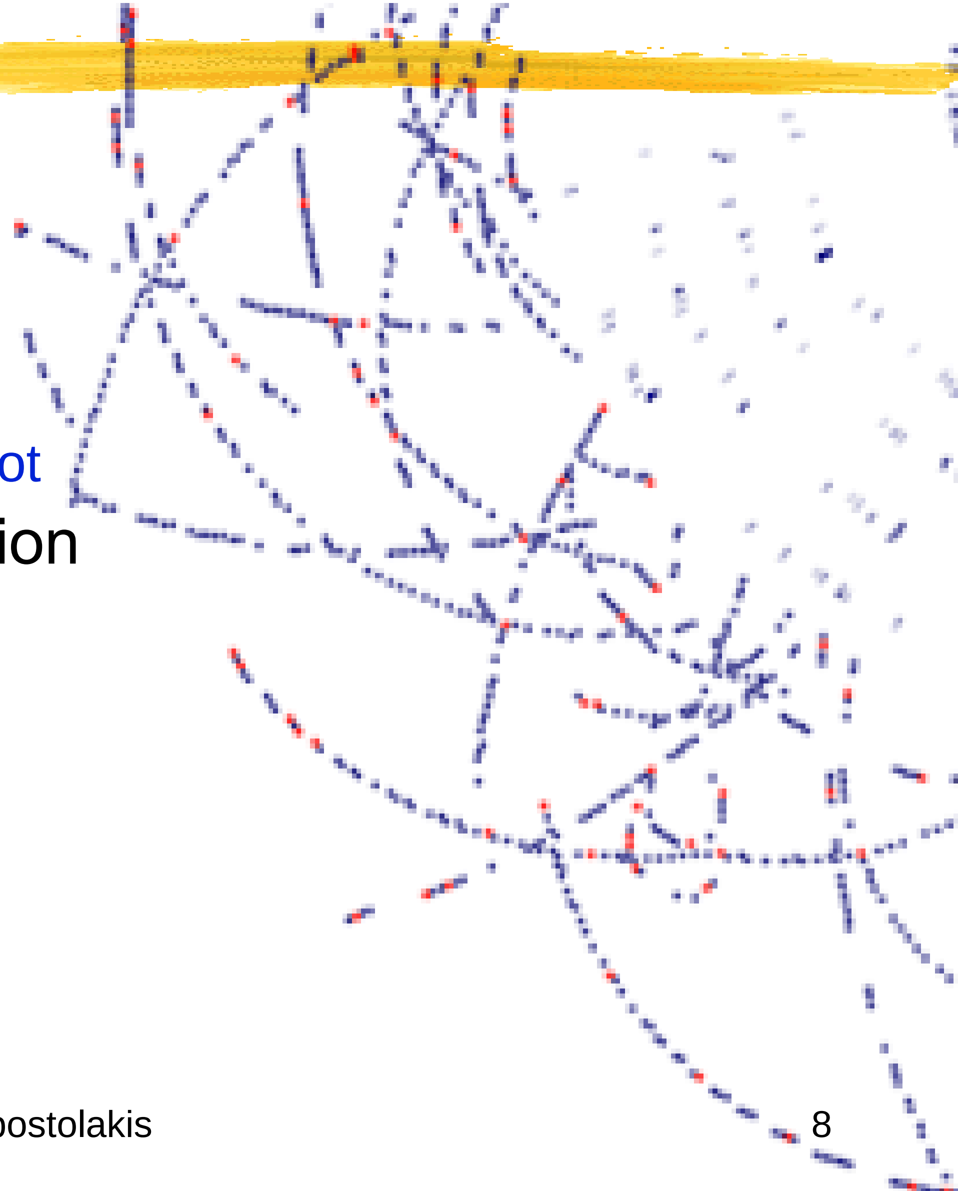
→ **Selectivity: 1 in 10^{13}**
(Like looking for a needle in 20 million haystacks)

Online and offline reconstruction

- Are collisions first-tagged really interesting enough to keep (given capacity constraints)?
  - Online reconstruction – seek to reconstruct 'as much as you can' quickly to enable decision
- Critical part of experiment – collisions which are not recorded are lost
- Later there is more time to reconstruct the contents of a collision – but this is also complex

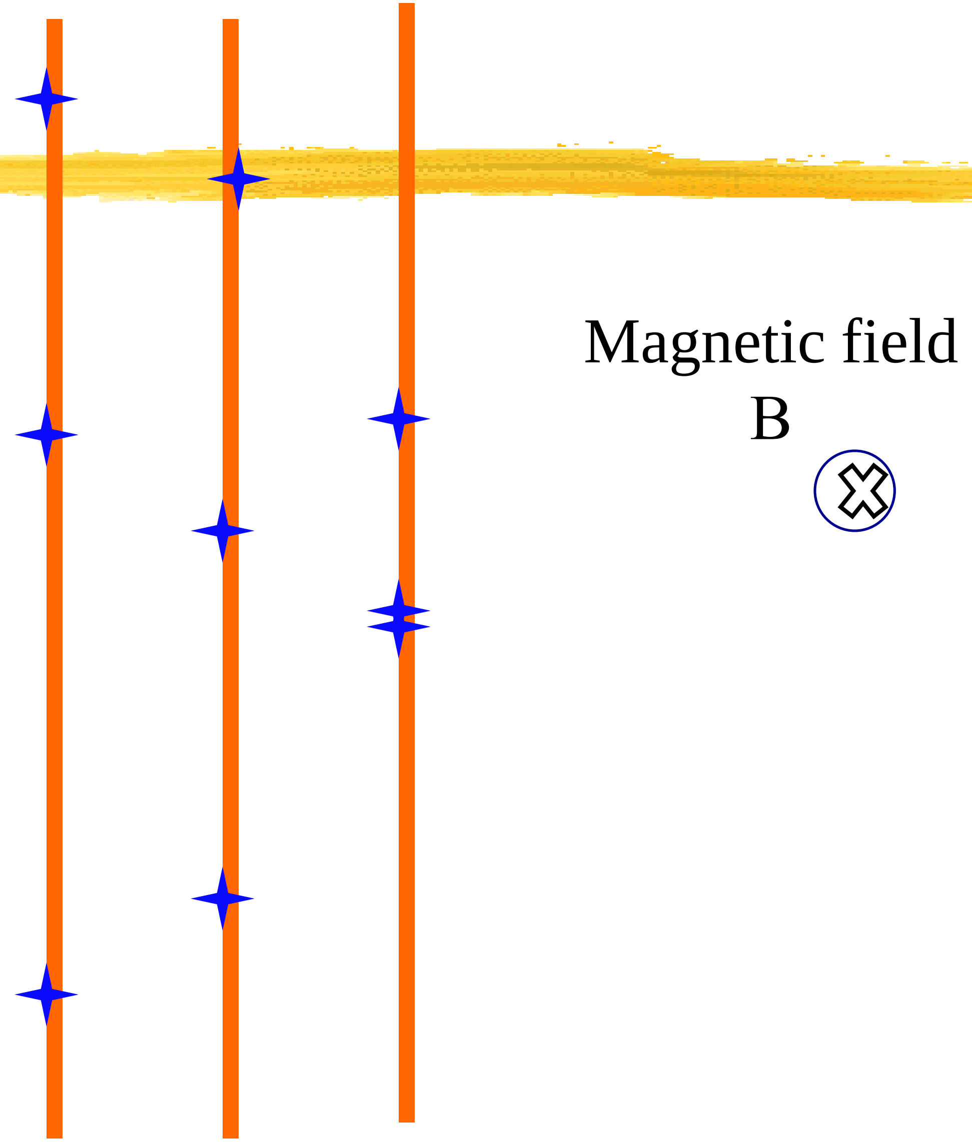
What is reconstruction

- Tracker hits form a puzzle
  - Which tracks created them?
- Each energy deposition is a clue
  - There are thousands of measurements in each snap-shot
- The experiment's reconstruction must obtain a solution!
  - In well measured magnetic field
  - Matches the traces to tracks



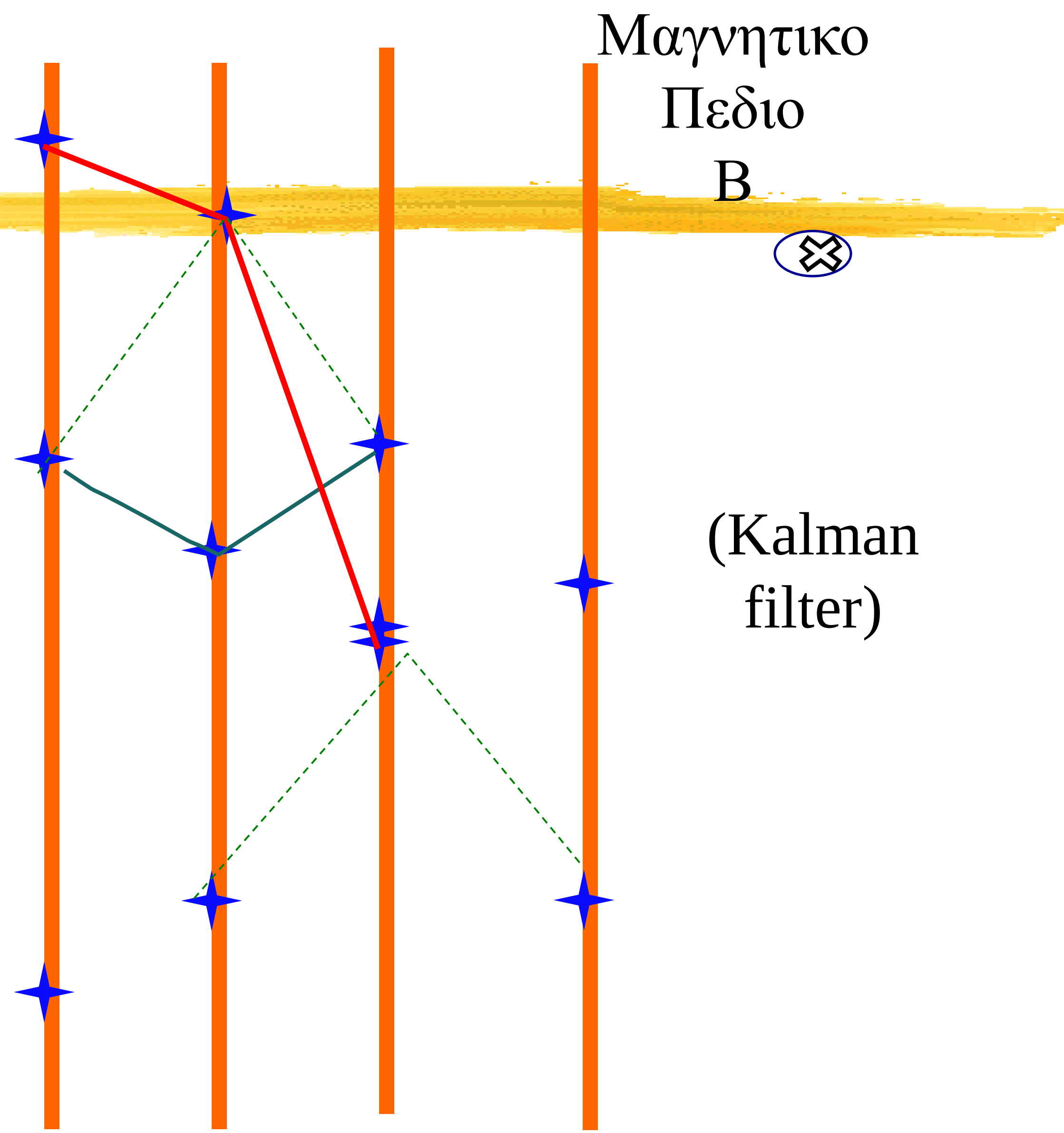
How it works – a simple example

- Start with the locations of the traces on first two planes



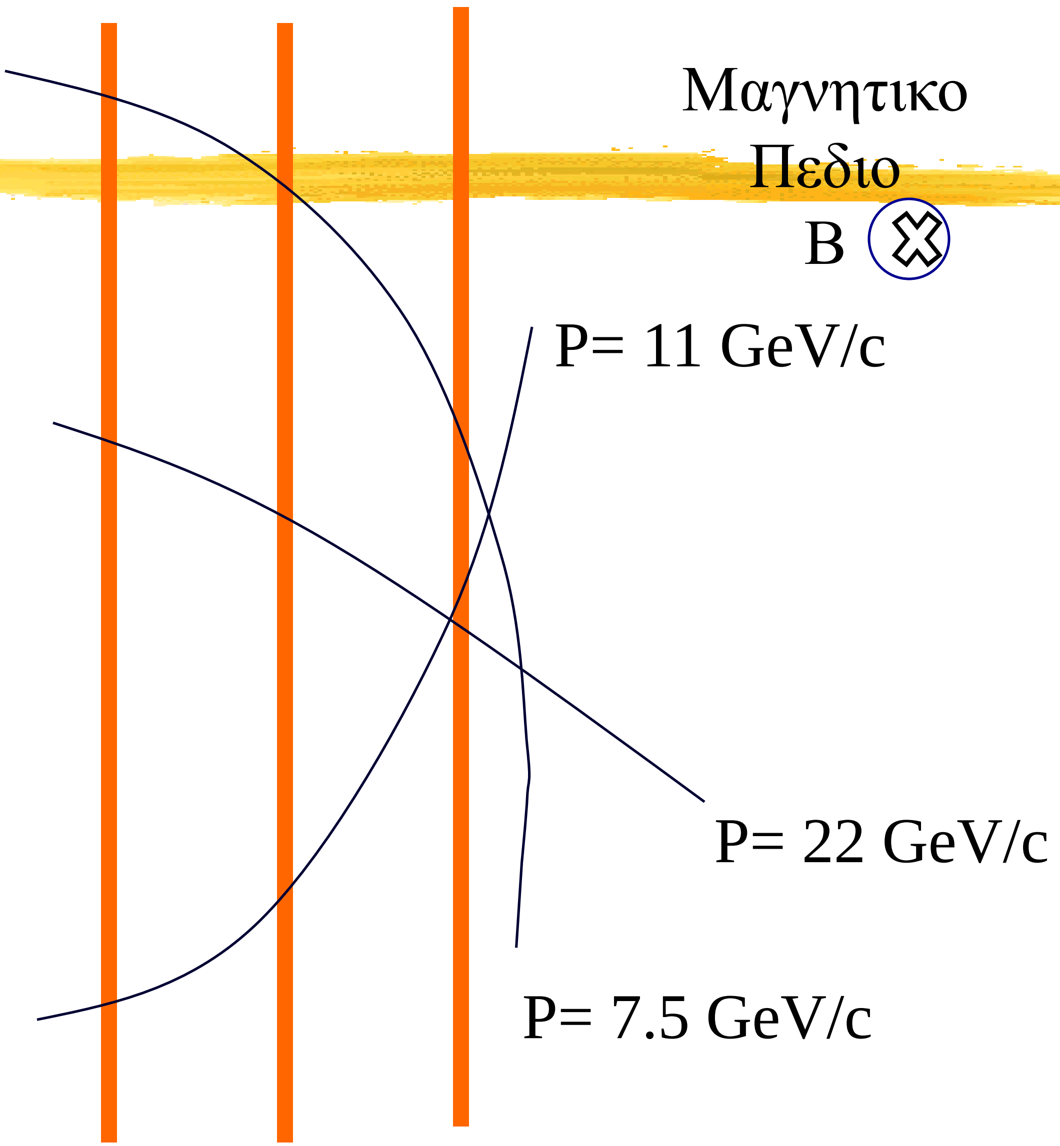
Reconstruction in practice

- Start with the locations of the traces on first two planes
- Try different combinations
    - Project to subsequent planes
    - Calculate differences between measured positions and 'predictions'



Reconstruction: result

- Start with the locations of the traces on first two planes
- Try different combinations
  - Project to subsequent planes
  - Calculate differences between measured positions and 'predictions'
- Finally the candidate tracks are identifed
  - else look 'quickly' for the straight(er) ones – high energy tracks



Simulation and Detectors

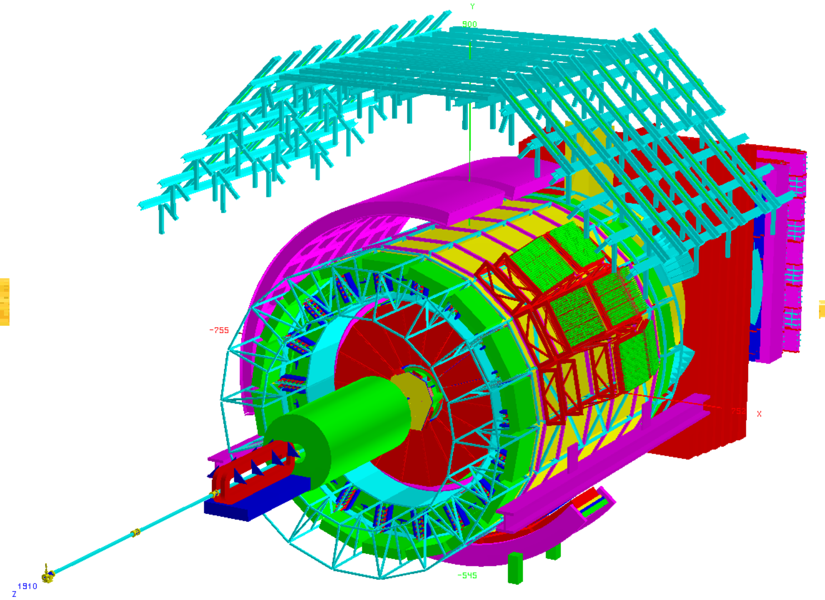
What is simulation ?

Why it exists ?

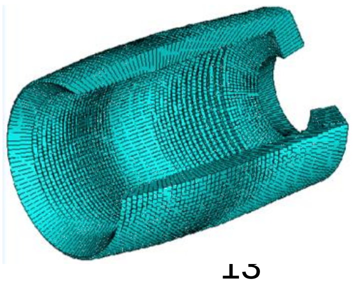
How is it done ?

Today's detectors

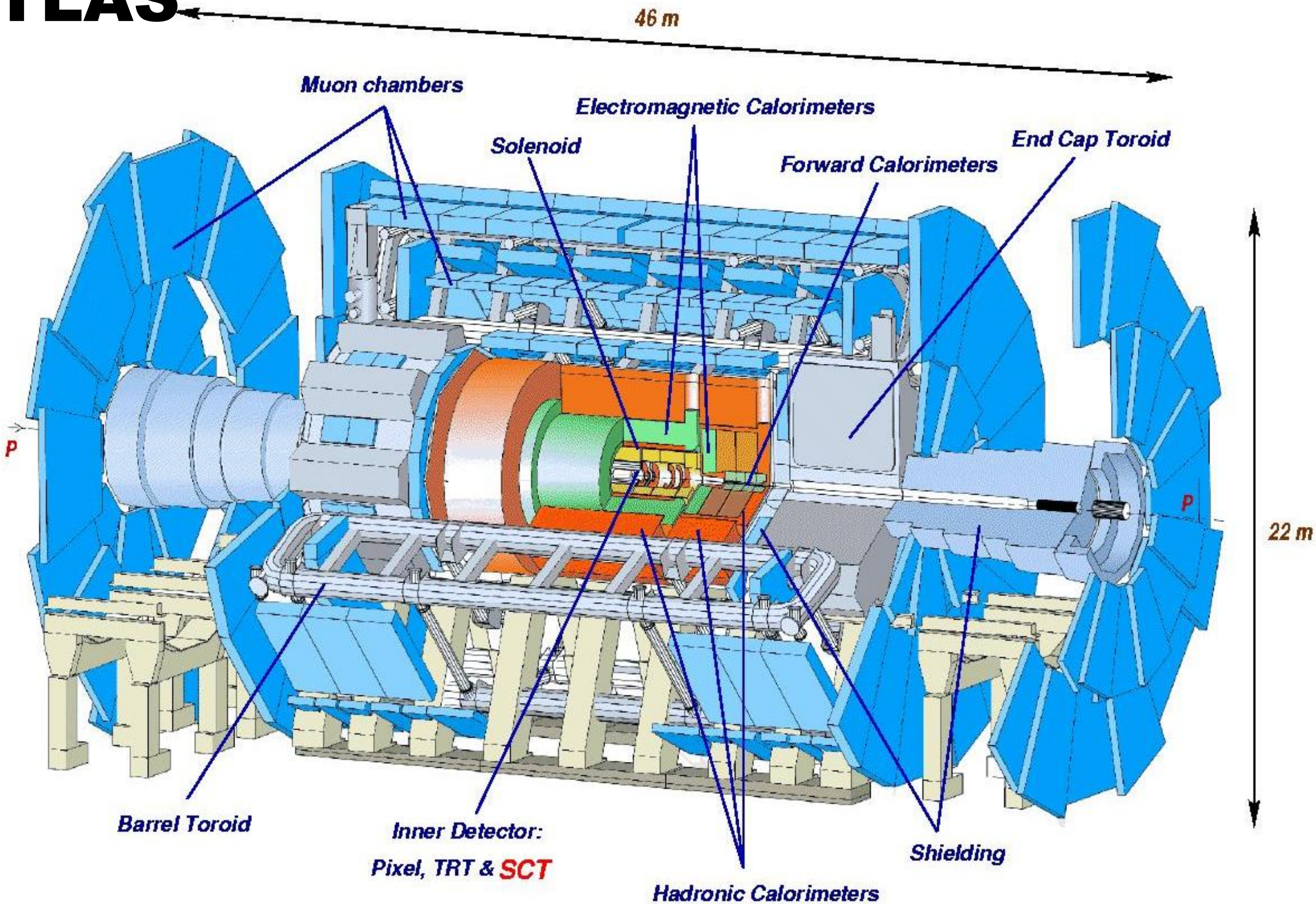
- Many different parts
  - Different capabilities
    - Measuring Location (trackers)
    - Measuring energy (calorimeters)
- Due to complexity
  - Different materials,
  - Most studies must use computers to create samples of tracker hits & energy deposition



ALICE Exp.



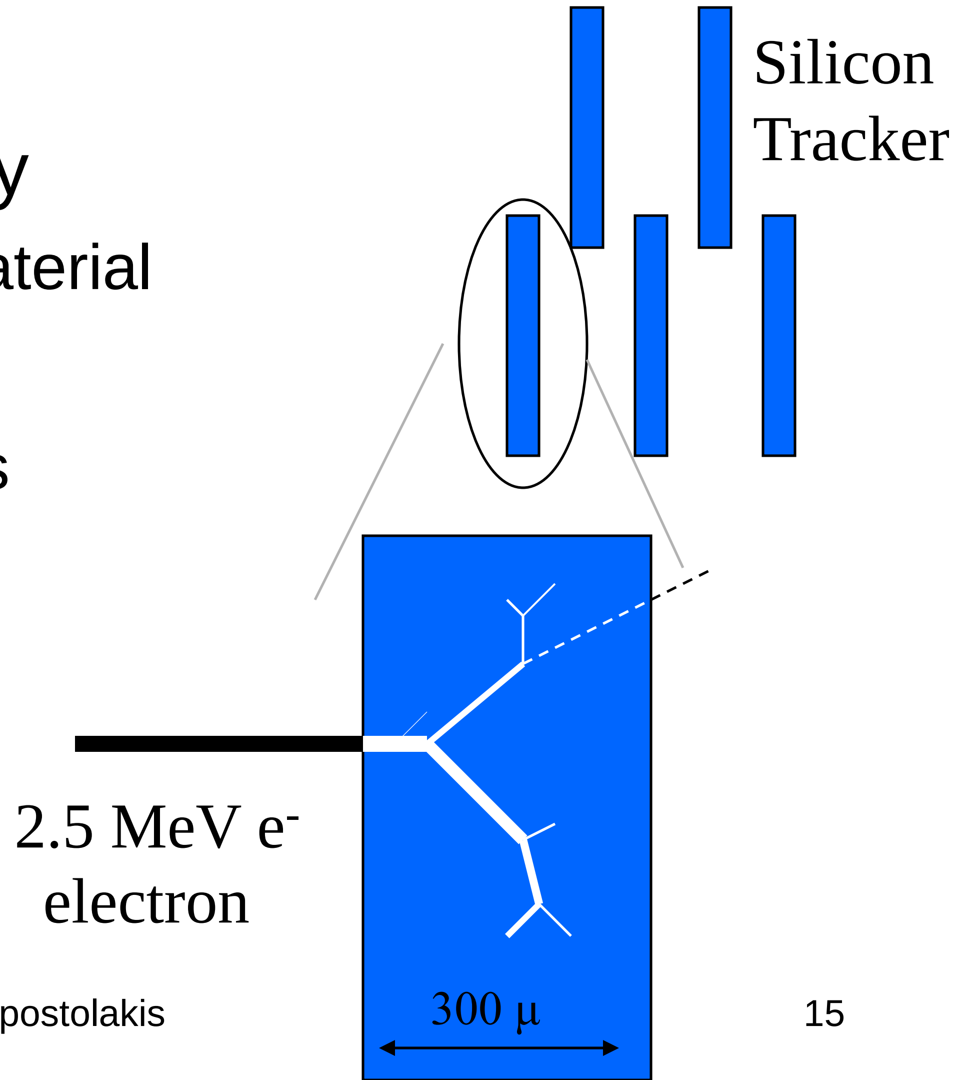
Today's detector Technologies: ATLAS



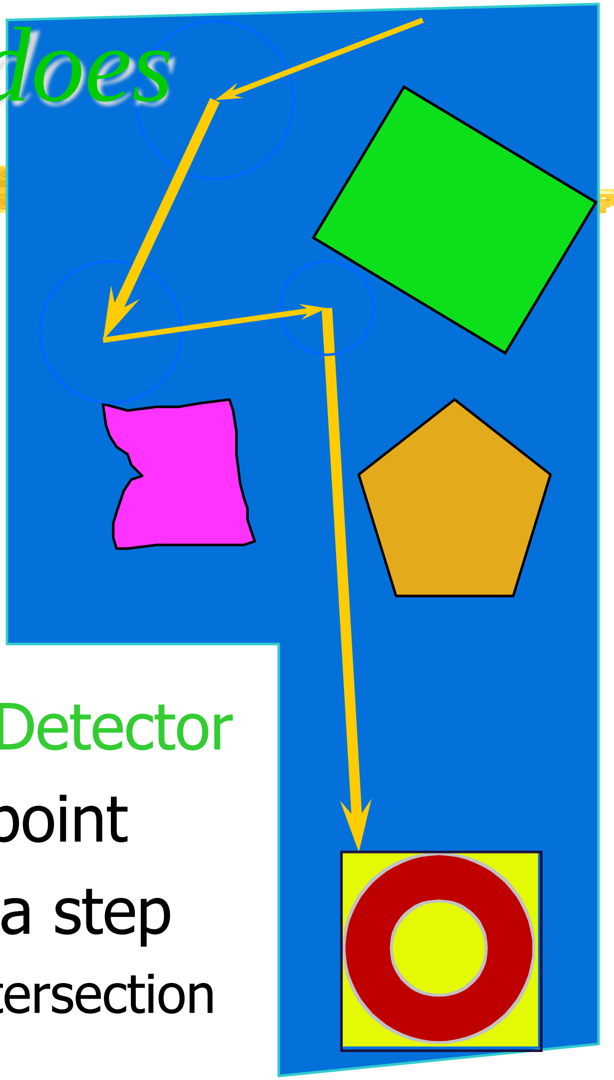
What is simulation ?

- We build <u>models</u>
  - Detector's Geometry
    - Shape, Location, Material
  - Physics interactions
    - All known processes
      - Electromagnetic
      - Nuclear (strong)
      - Weak (decay

$$\sigma_{total} = \Sigma\ \sigma_{per\text{-}interaction}$$



Geant4 geometry: what it does



Describes a Detector

- Hierarchy of volumes
- Many volumes repeat
  - Volume & sub-tree
- Up to millions of volumes for LHC era
- Import detectors from CAD systems

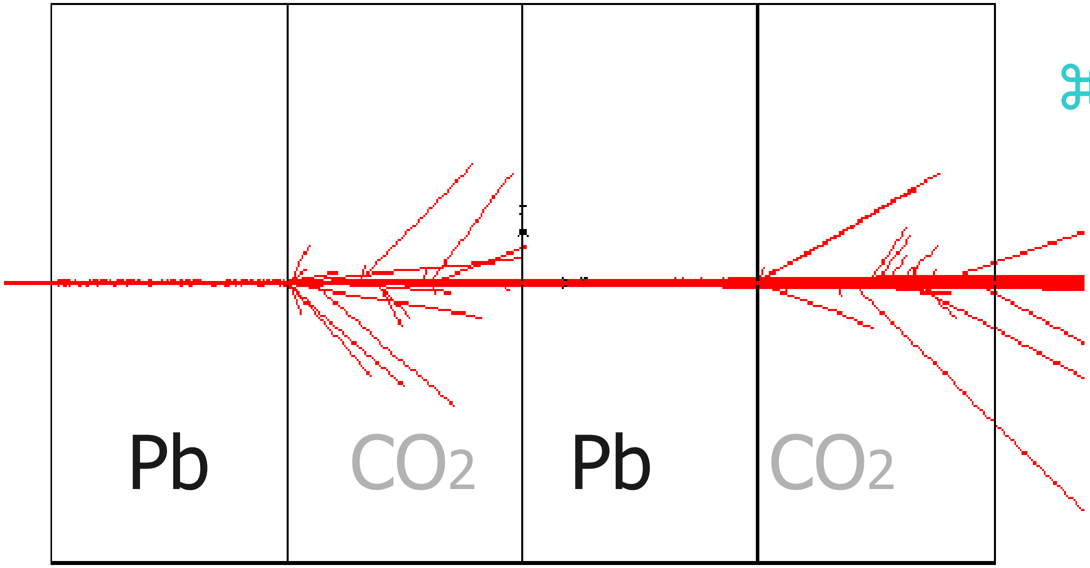
Navigates in Detector

- Locates a point
- Computes a step
  - Linear intersection

Physics processes

- Physics processes are modelled
- For example Electromagnetic processes include:
- Gammas:
  - Gamma-conversion, Compton scattering, Photo-electric effect
- Leptons(e, μ), charged hadrons, ions
  - Energy loss (Ionisation, Bremstrahlung) or PAI model energy loss, Multiple scattering, Transition radiation, Synchrotron radiation,
- Photons:
  - Cerenkov, Rayleigh, Reflection, Refraction, Absorption, Scintillation
- High energy muons and lepton-hadron interactions

A simple particle shower

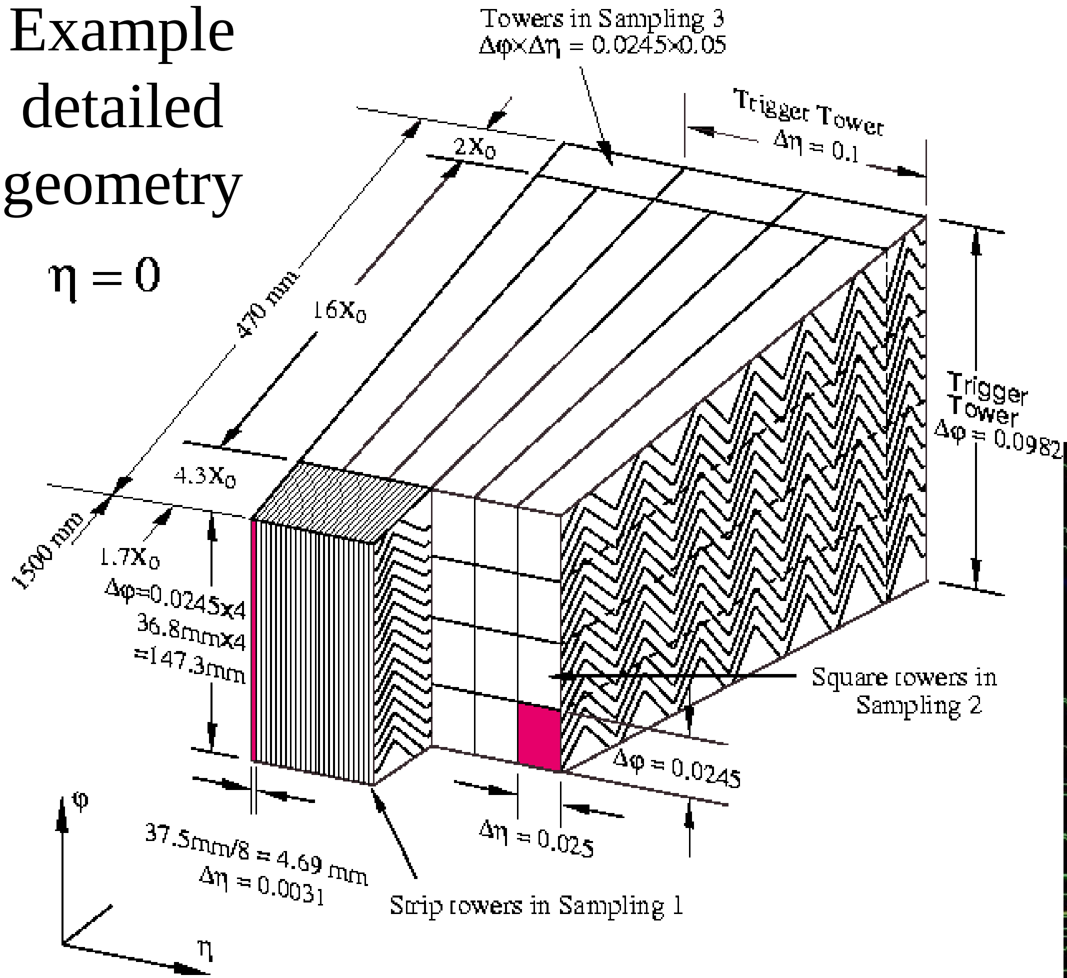


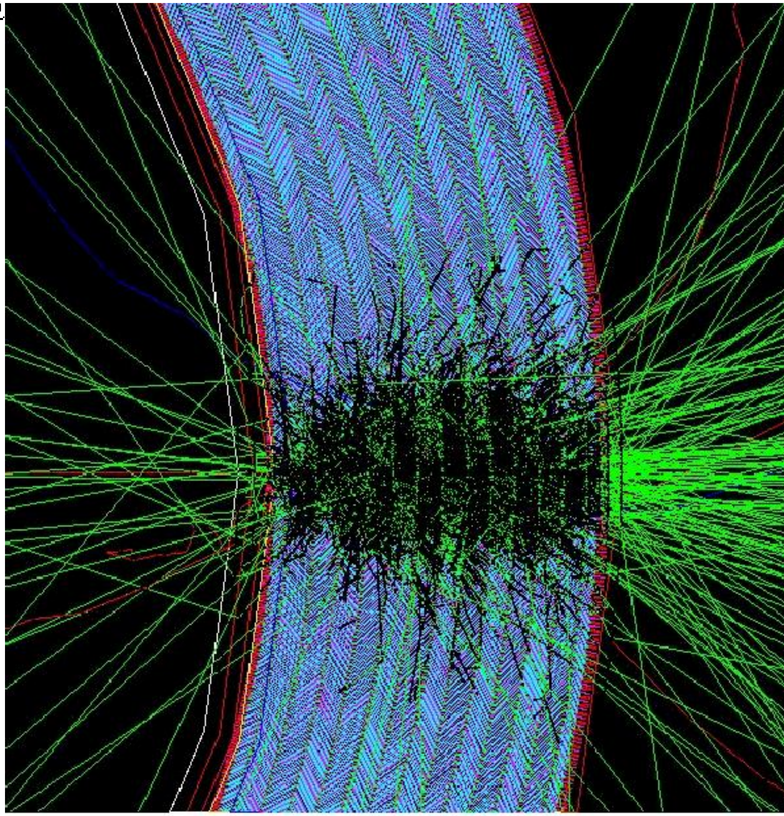
- In lead many secondary particles are produced
  - Most are contained
  - A few escape into CO2
- Energy deposition is measured in gas
  - Charged tracks ionise gas
  - Fewer new tracks produced

GEANT 3

Example detailed geometry

$\eta = 0$

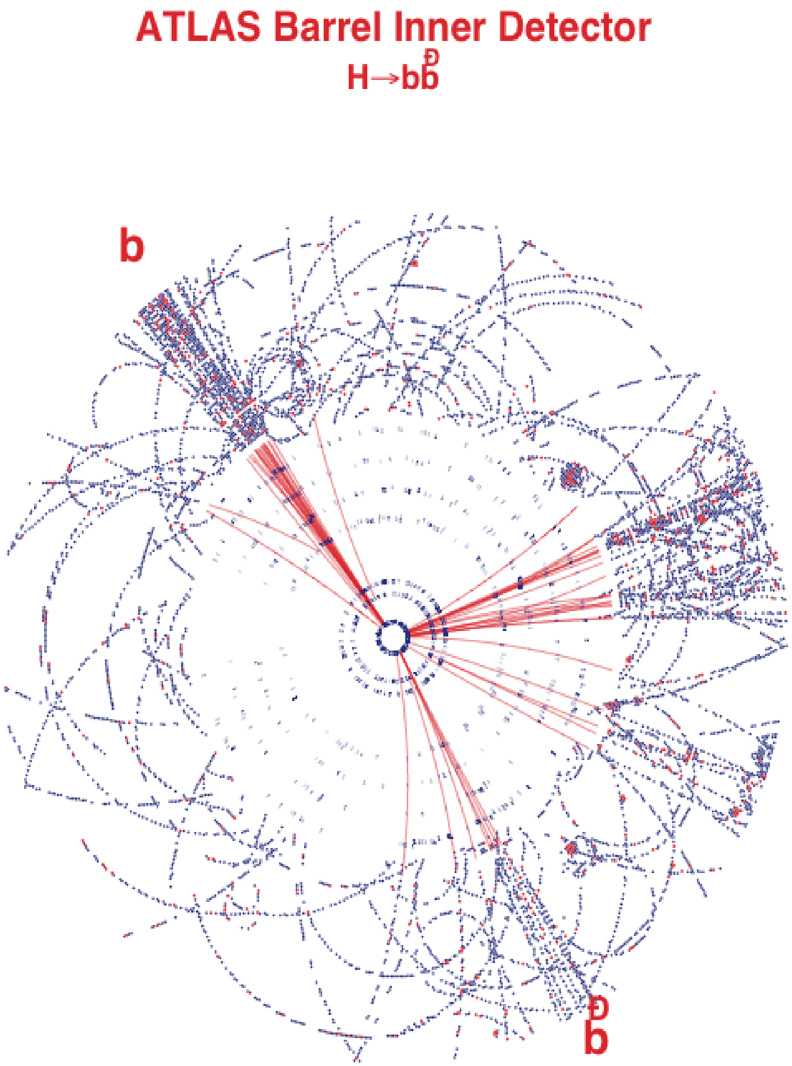




. Kordas "Geant4 for the ATLAS EM calo" — CALOR2000, Annecy, 12 October2000

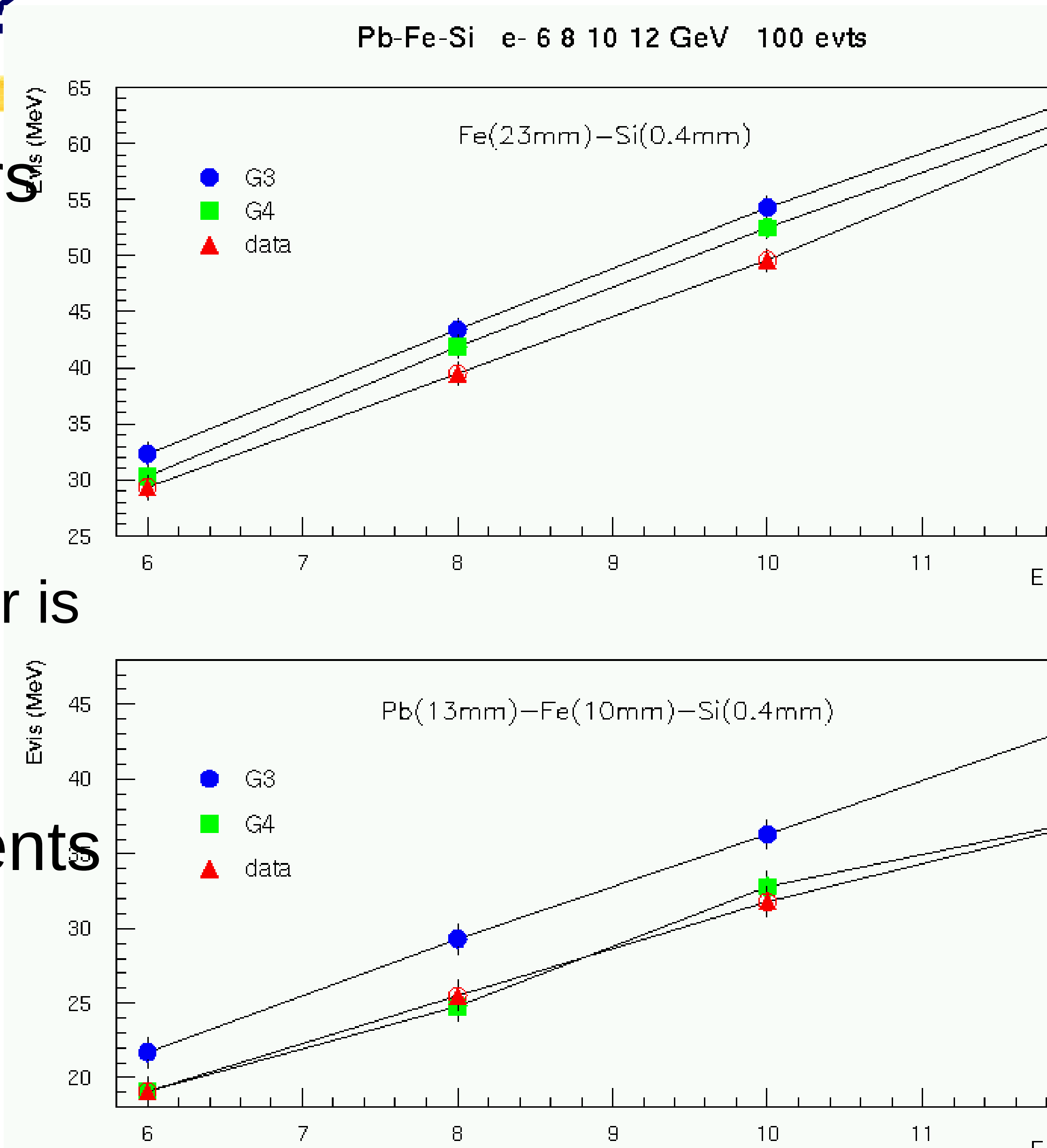
Atlas : Physics Signatures and Event Rates

- Beam crossing rate 40 MHz
- $\sigma_{inelastic}$ = 80 mb
  - In each beam crossing (rising each year, in 2012 ~ 25 interactions)
- Different physics 'targets'
  - Higgs Boson(s) (Discovery 2012)
  - Supersymmetric partner particles
  - Unexpected
  - Matter-antimatter differences (B mesons)
- Many examples of each channel are simulated

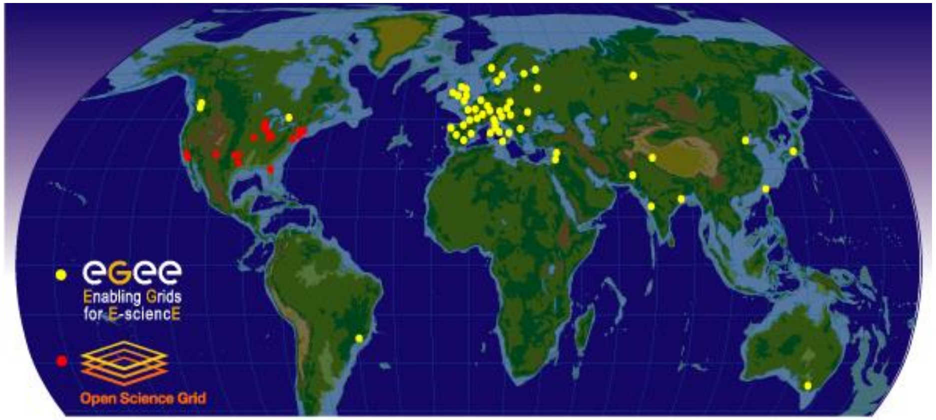


Why simulate ?

- To design detectors
  - Decise details
- To prepare the reconstruction
  - Before the detector is built and operates
- To understand events in the analysis



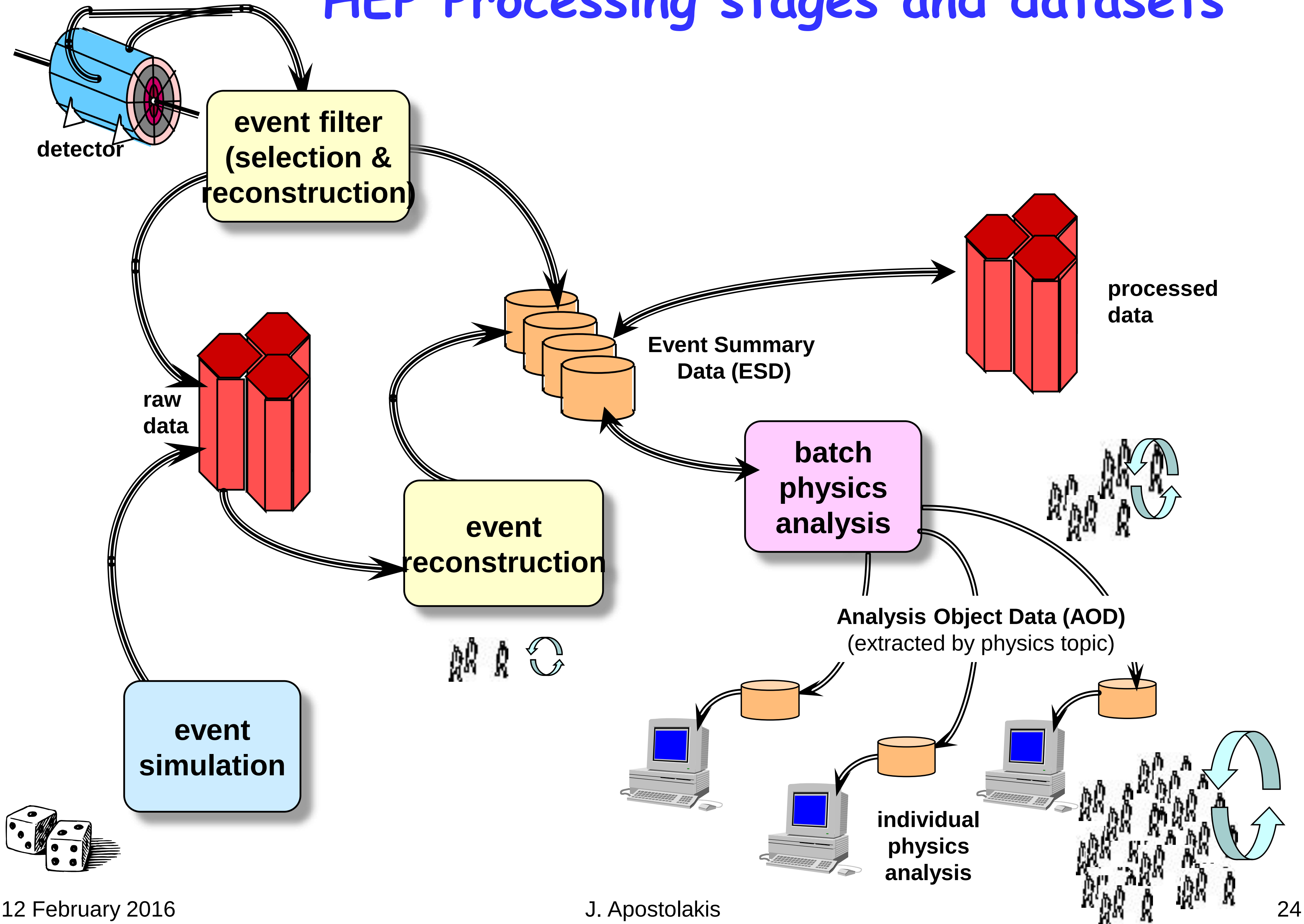
Data Analysis



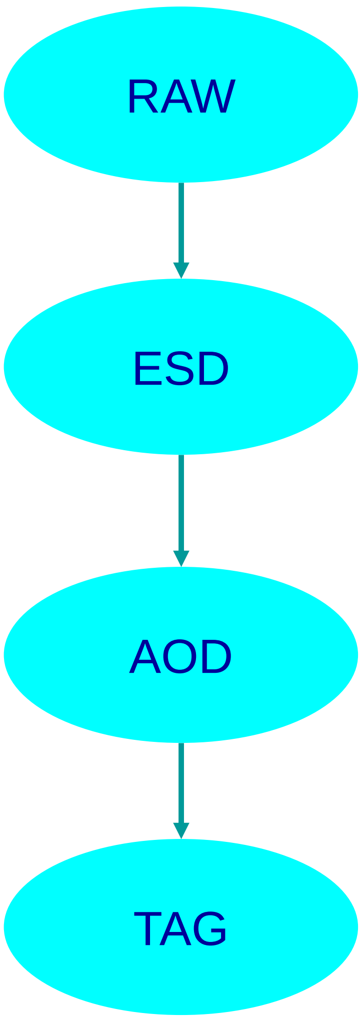
Data Analysis

- Uses the results of Reconstruction
  - the products are reconstructed tracks, Energy deposits (calorimeters)
  - Hierarchy of data from original (RAW), to summary (AOD)
- An experiment's physics teams use (large) pool of data
  - Not in one central location, but in multiple locations (cost, space of building, computers, disks) .... using the GRID
- The ROOT tool / framework used for analysis
  - https://root.cern.ch//
- Hypatia: a small part of analysis for a school setting
  - Introduction /Portal
    - http://hypatia.iasa.gr/en/index.html
  - http://indico.cern.ch/conferenceDisplay.py?confId=257353#2013-07-08

HEP Processing stages and datasets

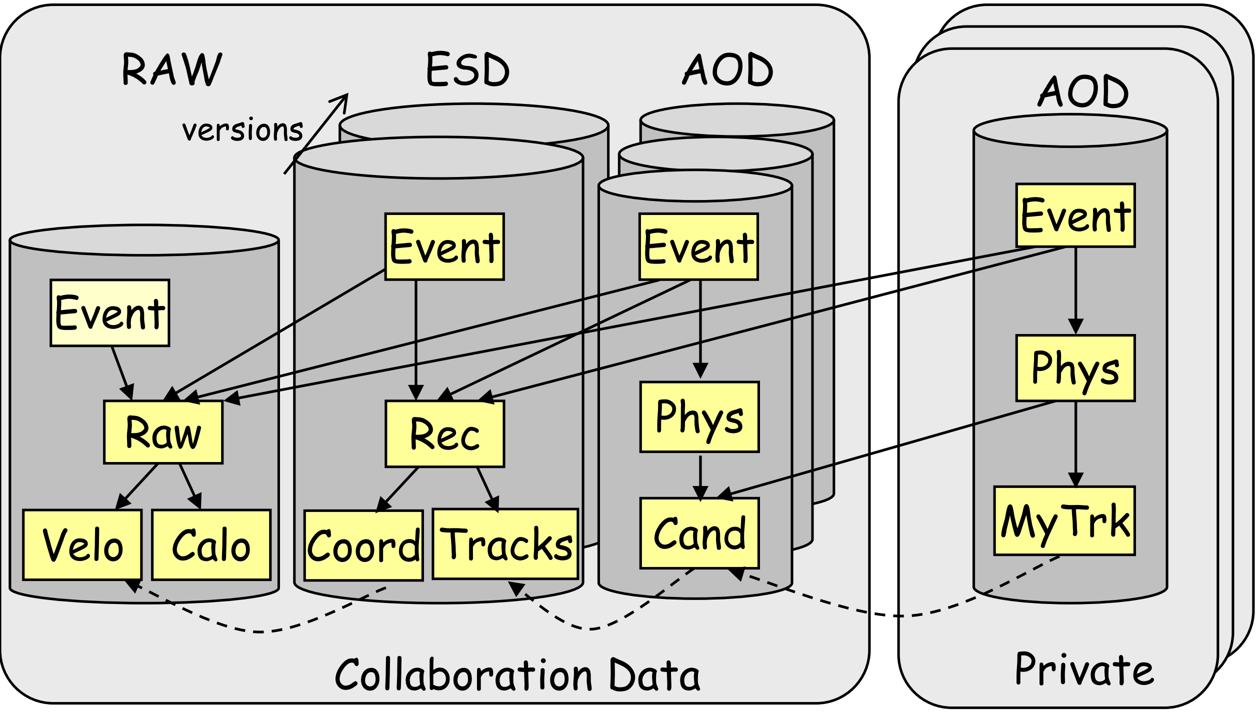


Data Hierarchy



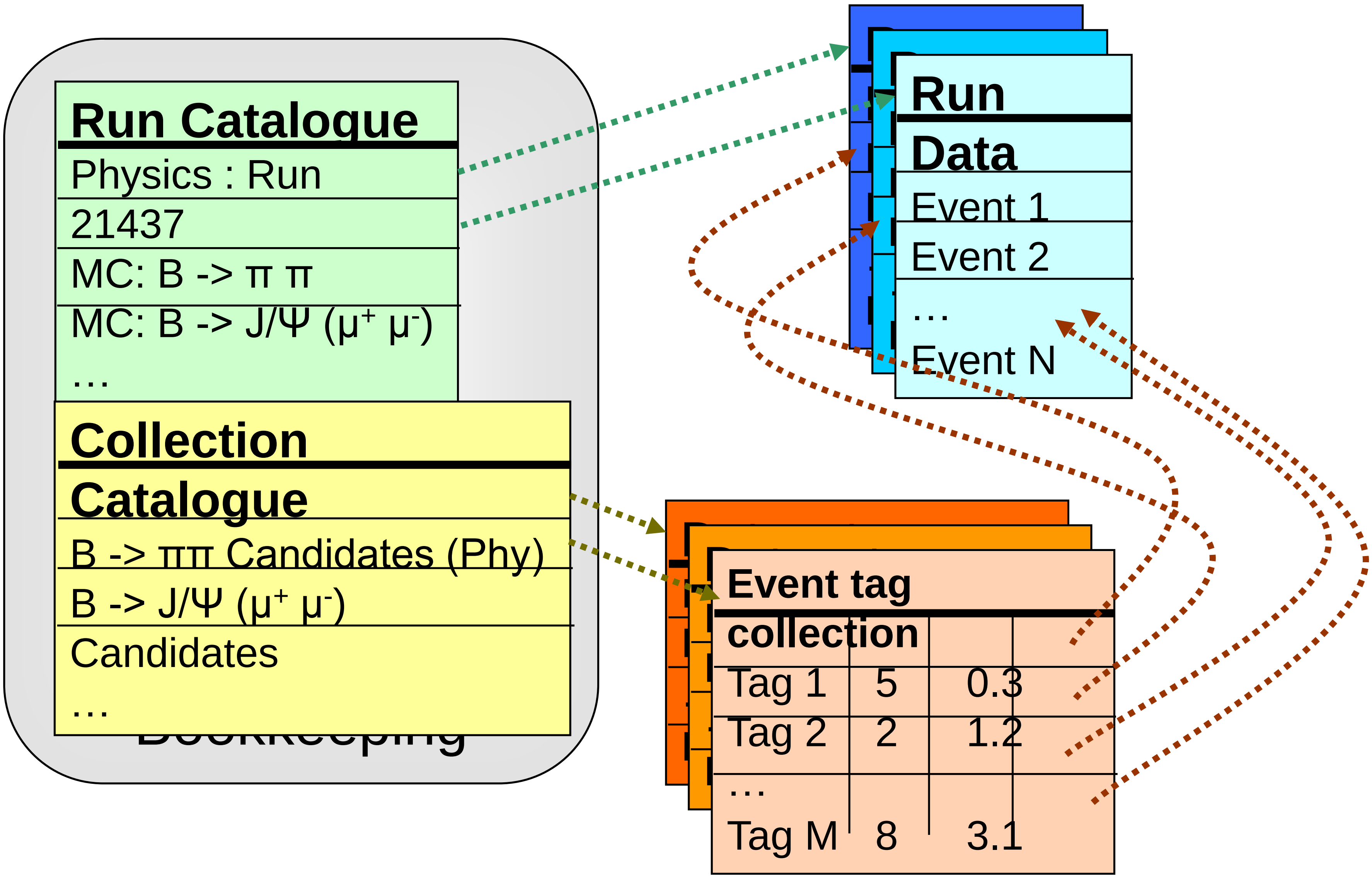
~2 MB/event	RAW	Raw data (DAQ)	Detector digitisation 10^9 events/yr * 2 MB =2 PB/yr
~100 kB/event	ESD	First reconstruction data	Pattern recognition information: Clusters, track candidates
~10 kB/event	AOD	Summary data for analysis	Physical information: Transverse momentum, Association of particles, jets, (best) id of particles,
~1 kB/event	TAG	Classification information	Relevant information for fast event selection

Event Data



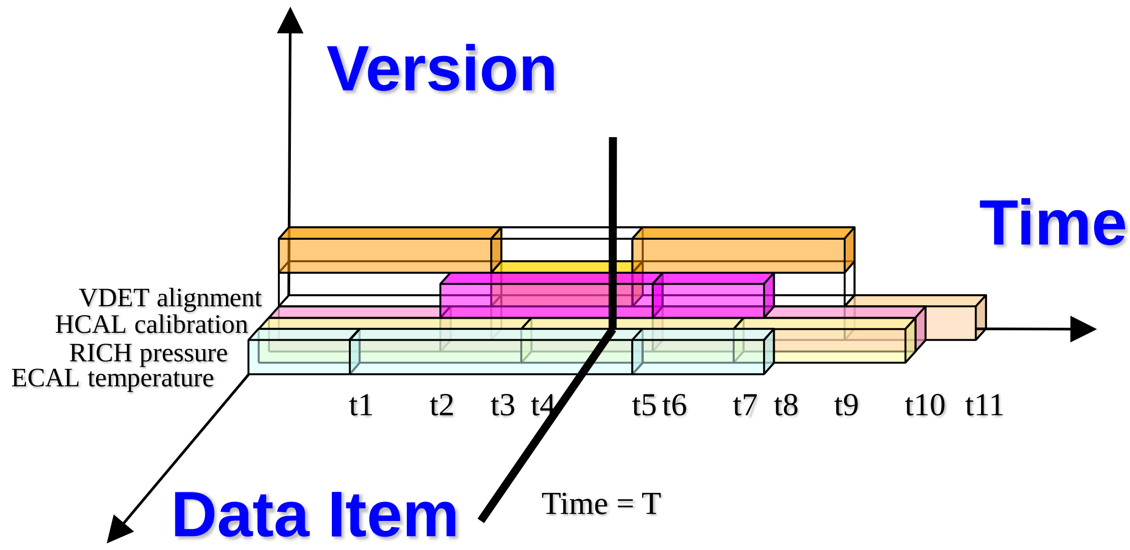
- Complex data models
  - ~500 structure types
- References to describe relationships between event objects
  - unidirectional
- Need to support transparent navigation
- Need ultimate resolution on selected events
  - need to run specialised algorithms
  - work interactively
- Not affordable if uncontrolled

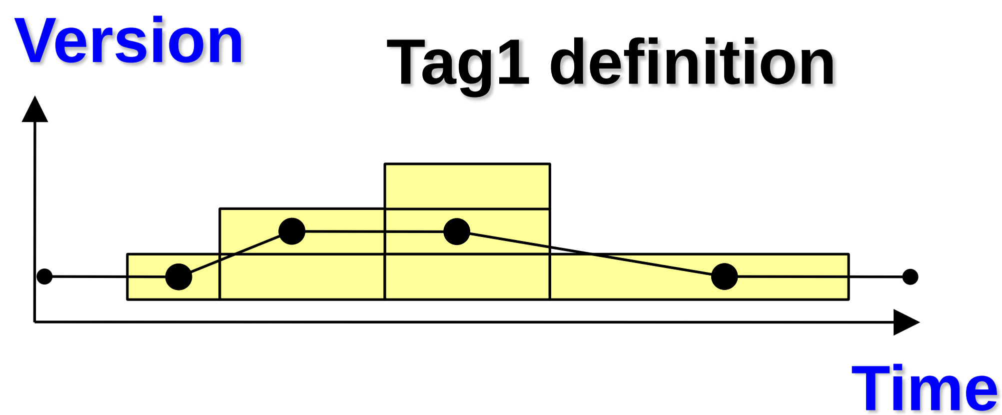
HEP Metadata - Event Collections



Detector Conditions Data

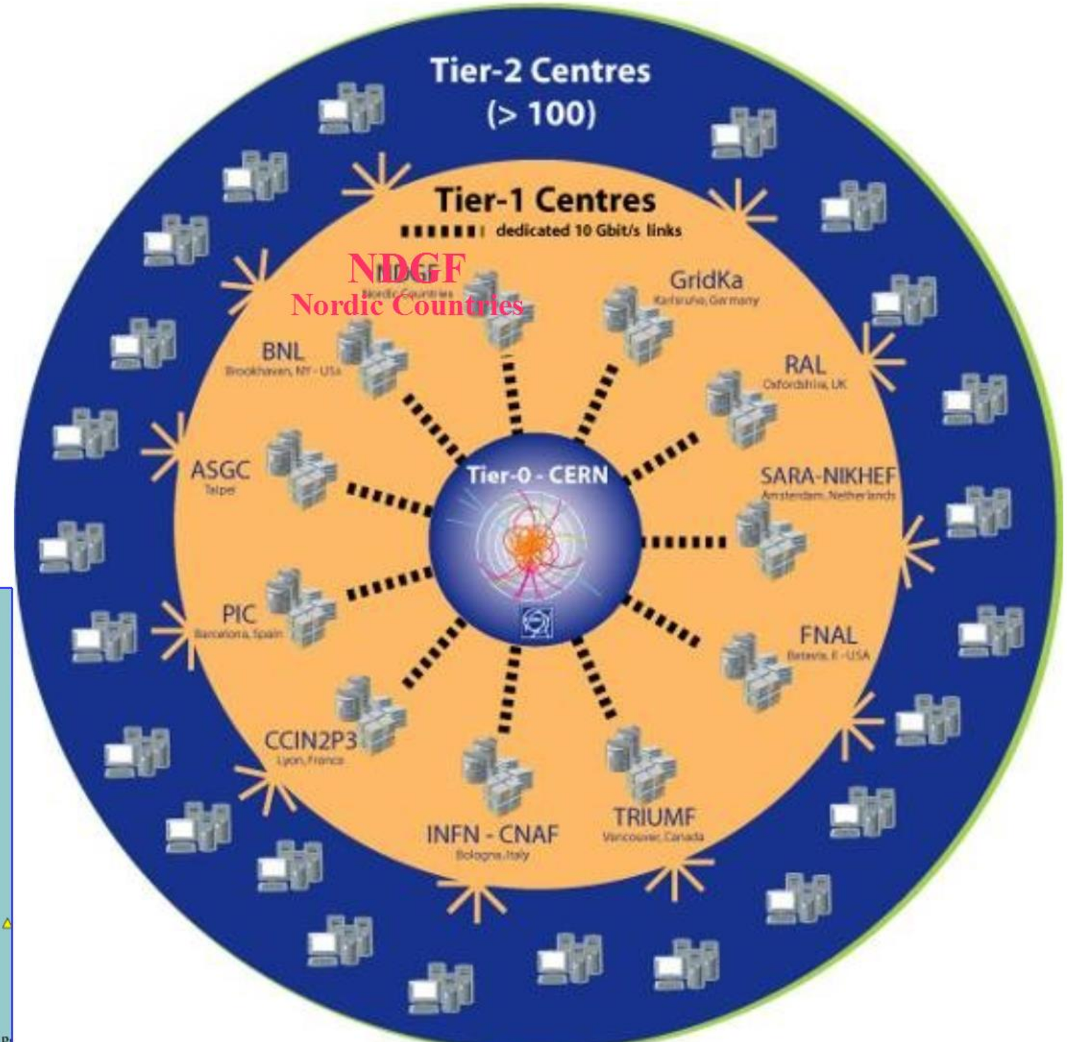
- Reflects changes in state of the detector with time
- Event Data cannot be reconstructed or analyzed without it
- Versioning
- Tagging
- Ability to extract slices of data required to run with job
- Long life-time





LHC Computing Grid project (LCG)

- **More than 170 computing centres**
- **12 large centres for primary data management: CERN (Tier-0) and eleven Tier-1s**
- **38 federations of smaller**



WLCG Collaboration

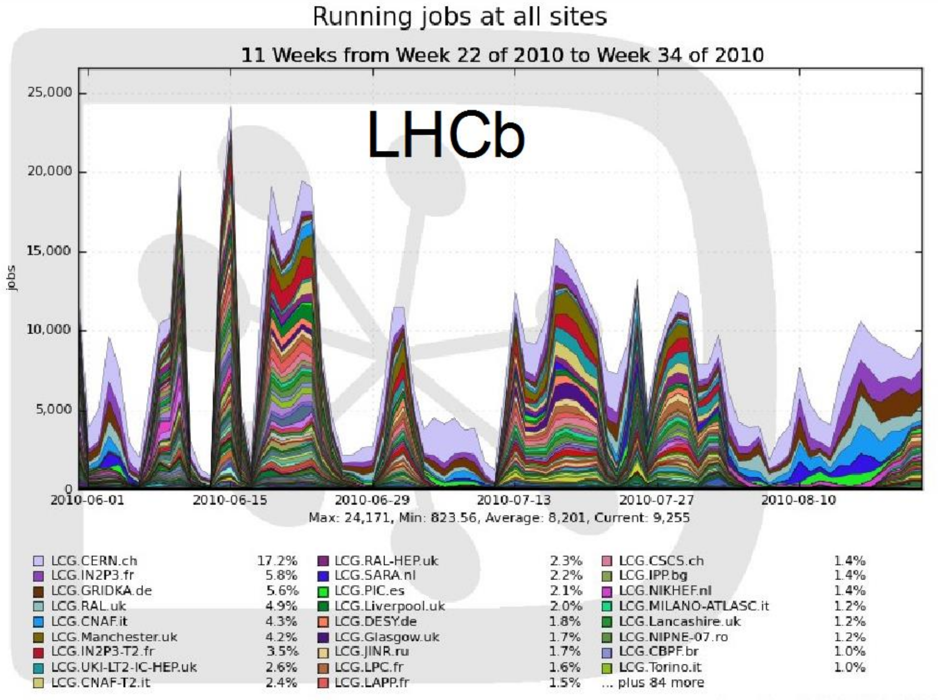
- **The Collaboration**
  - 4 LHC experiments
  - ~170 computing centres
  - 12 large centres (Tier-0, Tier-1)
  - 38 federations of smaller "Tier-2" centres
  - ~35 countries
- **Memorandum of Understanding**
  - Agreed in October 2005
- **Resources**
  - Focuses on the needs of the four LHC experiments
  - Commits resources
    - each October for the coming year
    - 5-year forward look
  - Agrees on standards and procedures
- **Relies on EGEE and OSG (and other regional efforts)**

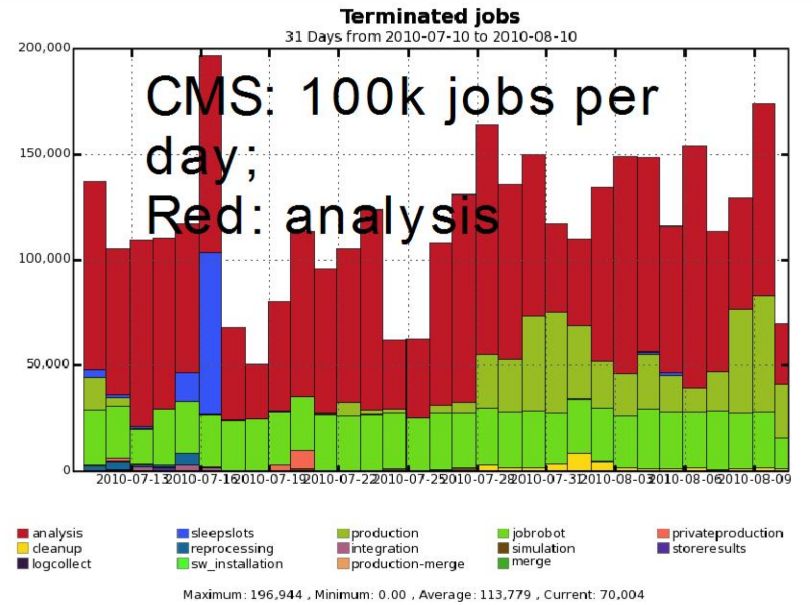
Open Science Grid

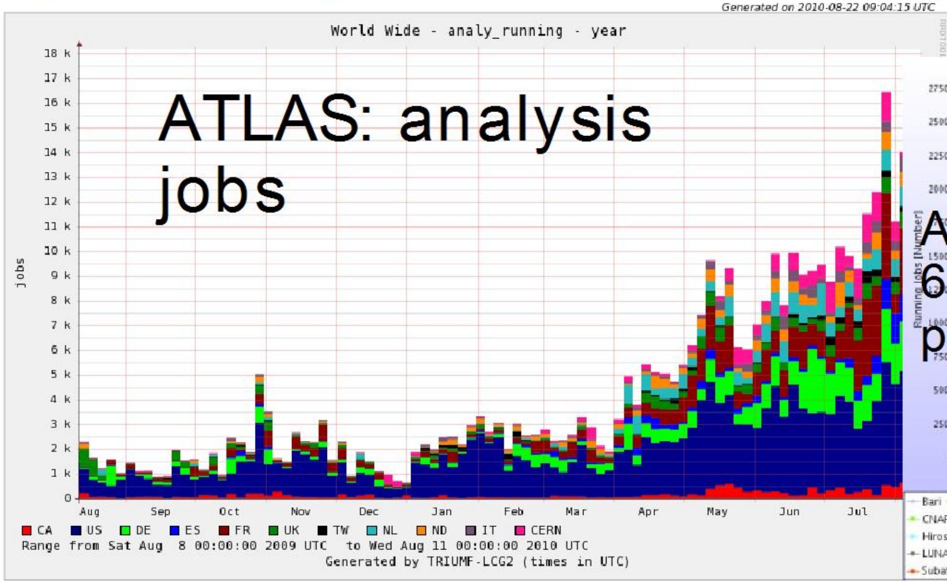
egee
Enabling Grids for E-science in Europe

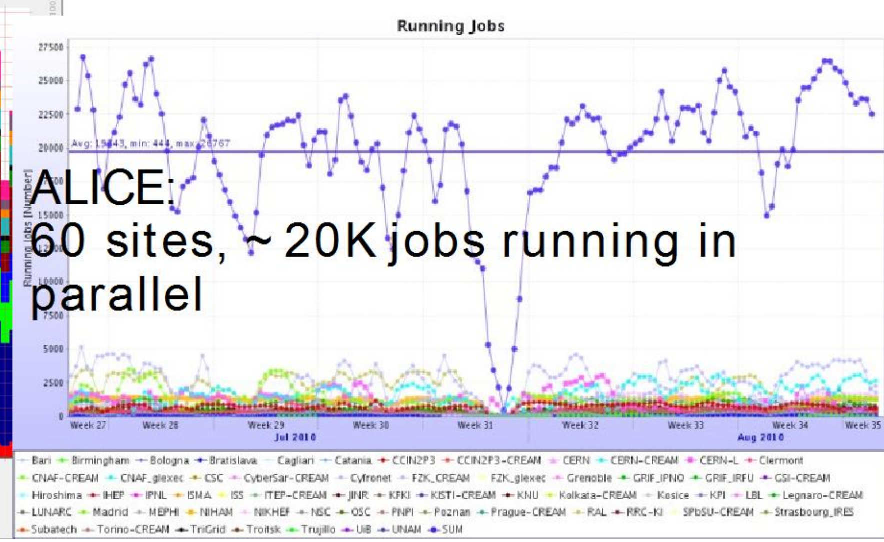
NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

Running jobs on LCG



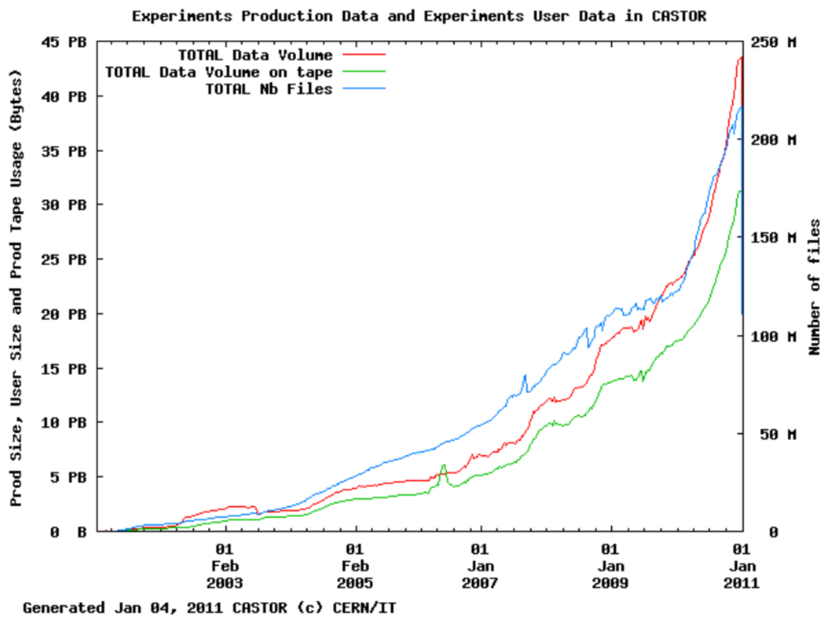




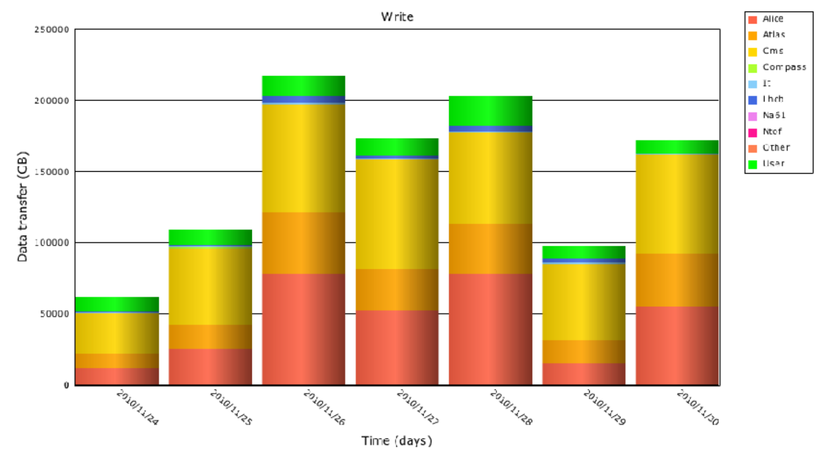


WLCG
Worldwide LHC Computing Grid
LCG

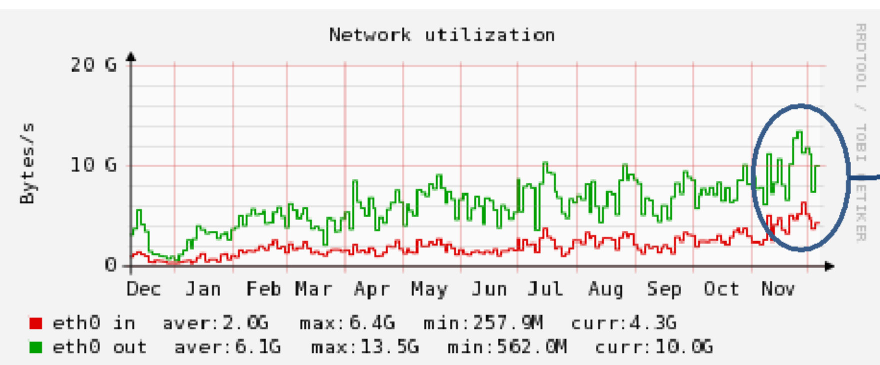
2010 Tier-0 Data Taking

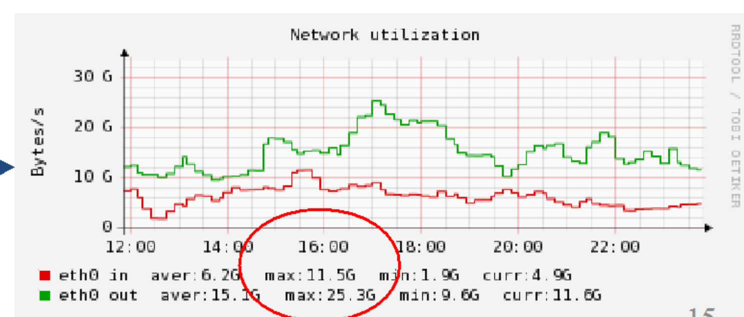


Stored ~ 15 PB in 2010 with peaks at 220 TB/day during Pb+Pb



Tier-0 Bandwidth
Average in: 2 GB/s with peaks at 11.5 GB/s
Average out: 6 GB/s with peaks at 25 GB/s



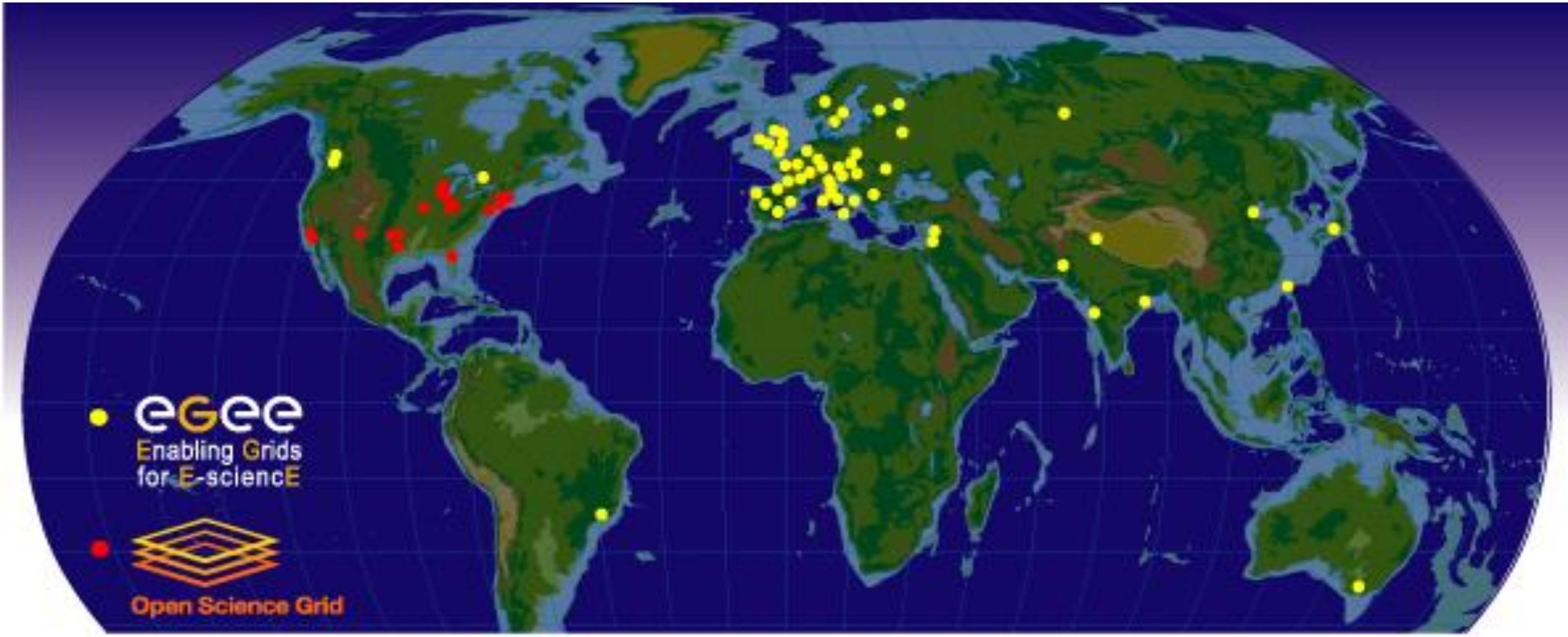


GRID vs Cloud

- **"Cloud computing" is gaining importance**
  - Web based solutions (http/https and RES)
  - Virtualization, upload machine images to remote sites
- **GRID has mainly a scientific user base**
  - Complex applications running across multiple sites, but works like a cluster batch system for the end user
  - Mainly suitable for parallel computing and massive data processing
- **Expect convergence in the future**
  - "Internal Cloud" at CERN
  - CernVM – virtual machine running e.g. at Amazon

LCG depends on two major science grid infrastructures ….

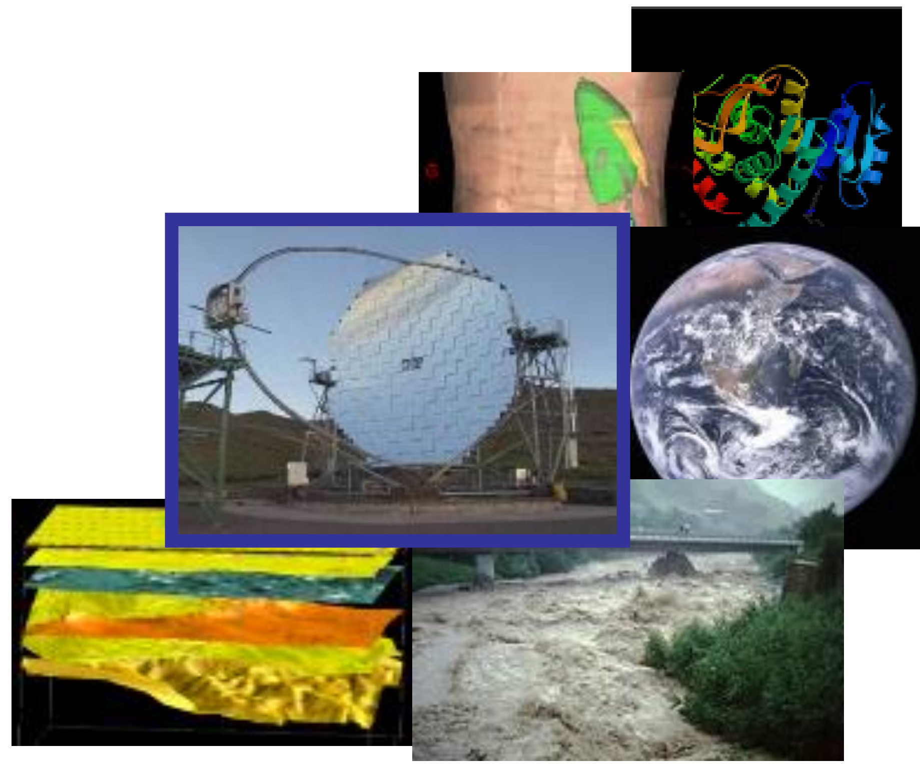
EGEE - Enabling Grids for E-Science
OSG - US Open Science Grid



A map of the worldwide LCG infrastructure operated by EGEE and OSG.

Applications

- **Many applications in different domains**
  - High Energy Physics (Pilot domain)
    - Experiments at CERN (LHC), DESY, Fermilab
  - Biomedical (Pilot domain)
    - Bioinformatics
    - Medical imaging
  - Earth Sciences
    - Geo-surveying
    - Solid Earth Physics
    - Hydrology, Climate
  - Computational Chemistry
  - Fusion
  - Astronomy
    - Cosmic Microwave Background
    - Gamma ray astronomy
  - Geology
  - Industrial Applications



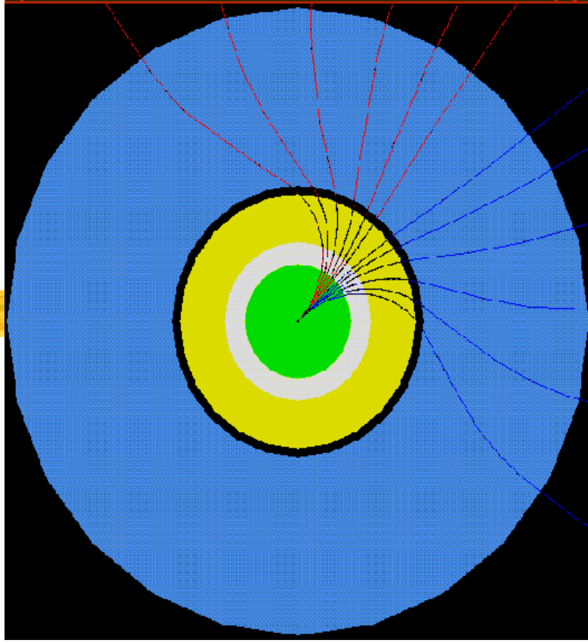
Backup

More on simulation

Applications beyond HEP

- Biomedical Bioinformatics
  - Medical imaging
- Earth Sciences
  - Geo-surveying
  - Solid Earth Physics
  - Hydrology, Climate
- Astronomy
  - Cosmic Microwave Background
  - Gamma ray astronomy
- Computational Chemistry
- Fusion
- Geology
- Industrial Applications

Propagating in a field



Charged particles follow paths that approximate their curved trajectories in an electromagnetic field.

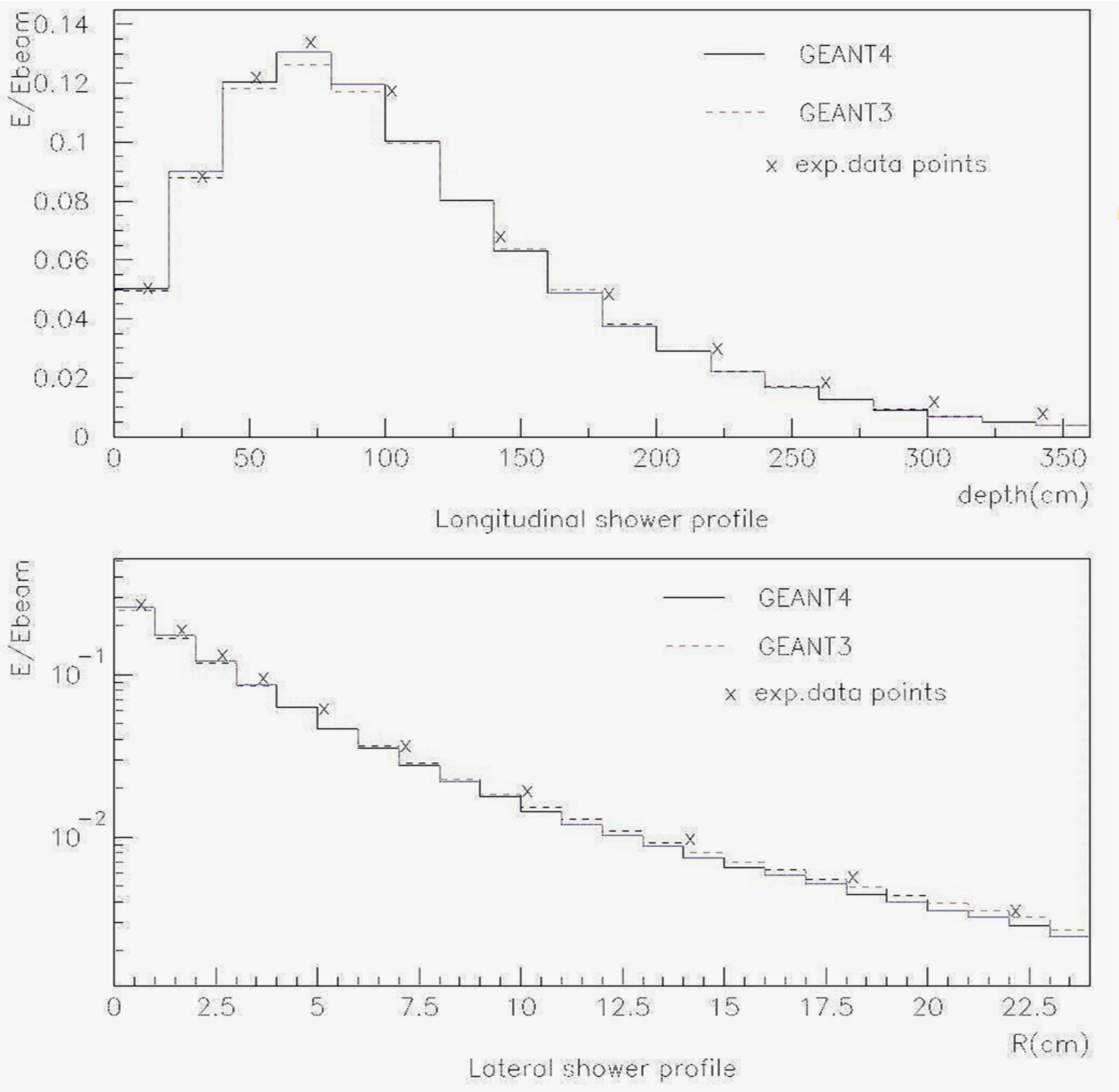
- It is possible to tailor
  - the accuracy of the splitting of the curve into linear segments,
  - the accuracy in intersecting each volume boundaries.
- These can be set now to different values for a single volume or for a hierarchy.

Shower profile

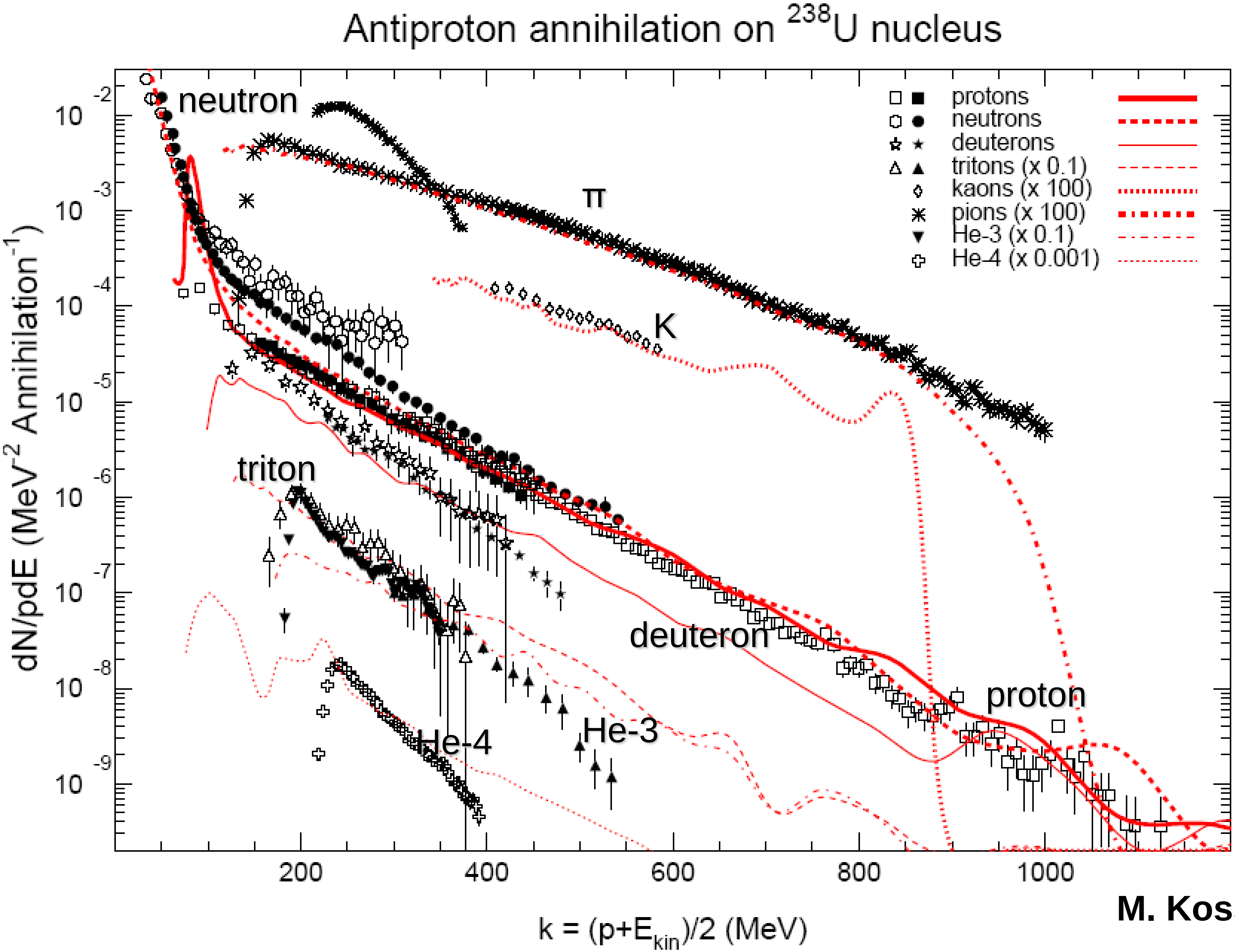
1 GeV electron in H_2O

G4, Data G3

- Good agreement seen with the data



Antiproton annihilation - CHIPS Model

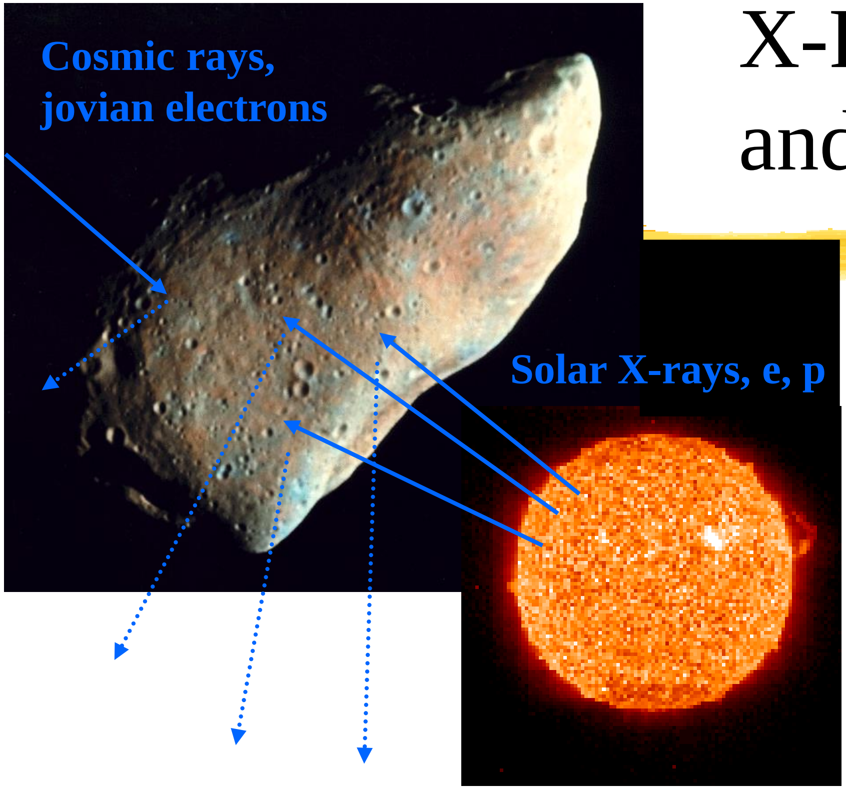


M. Kossov

Simulation 'packages'

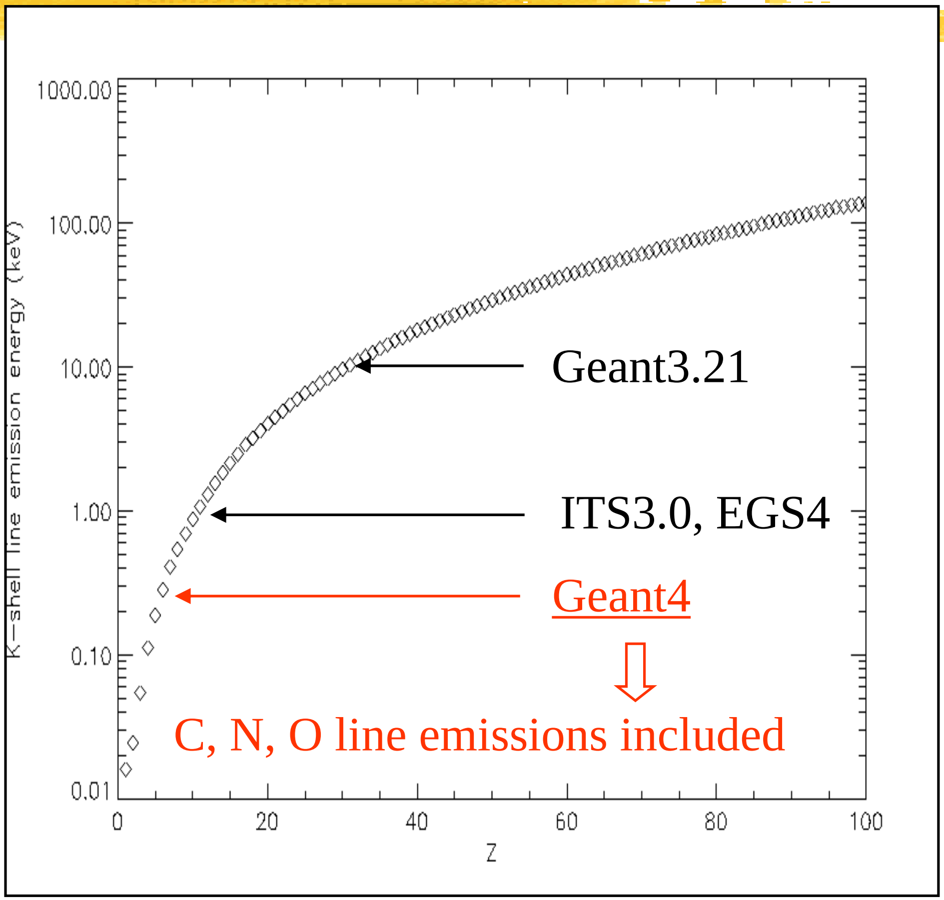
- Provides the means to simulate
  - the physical processes and
  - detector response of an experiment.
- As was realised by many in the past,
  - most of the parts needed can be common between experiments (eg physics, geometry blocks) .
- So it makes eminent sense to create and use a general purpose package
  - That includes the common parts,
  - And enables an experiment to describe those parts with are specific to it.

X-Ray Surveys of Asteroids and Moons

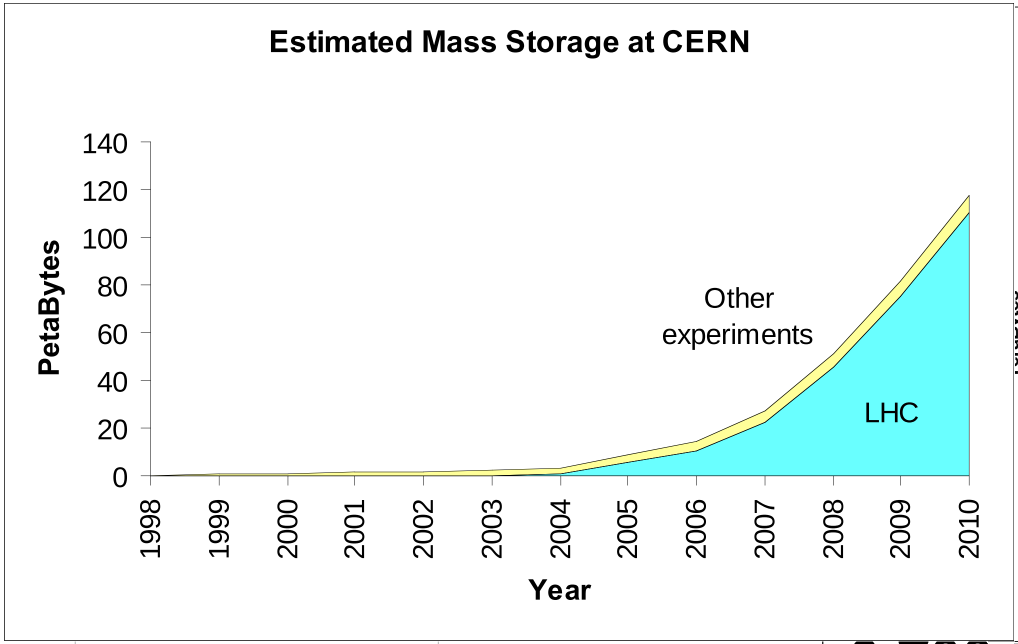


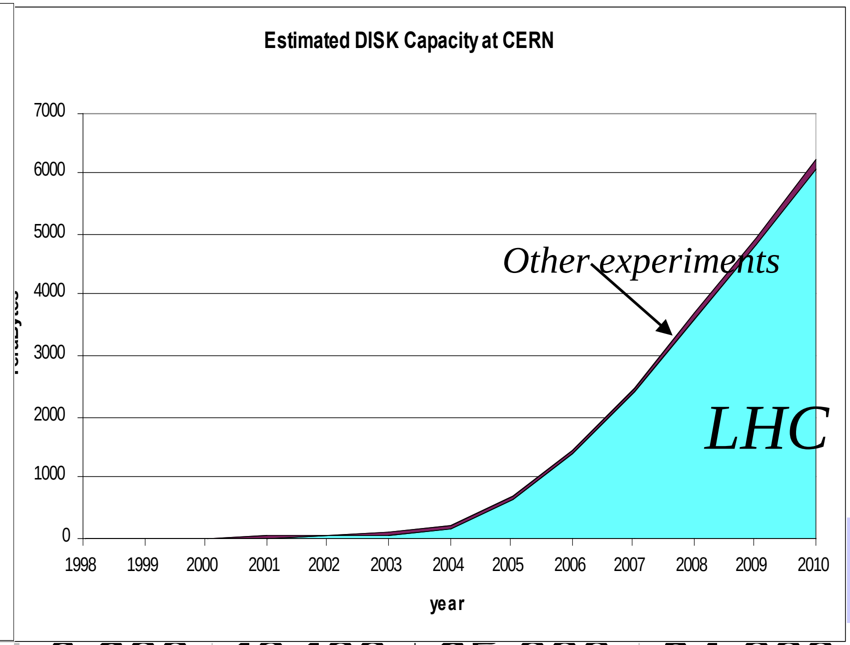
Courtesy SOHO EIT

Induced X-ray line emission:
indicator of target composition
(~100 μm surface layer)

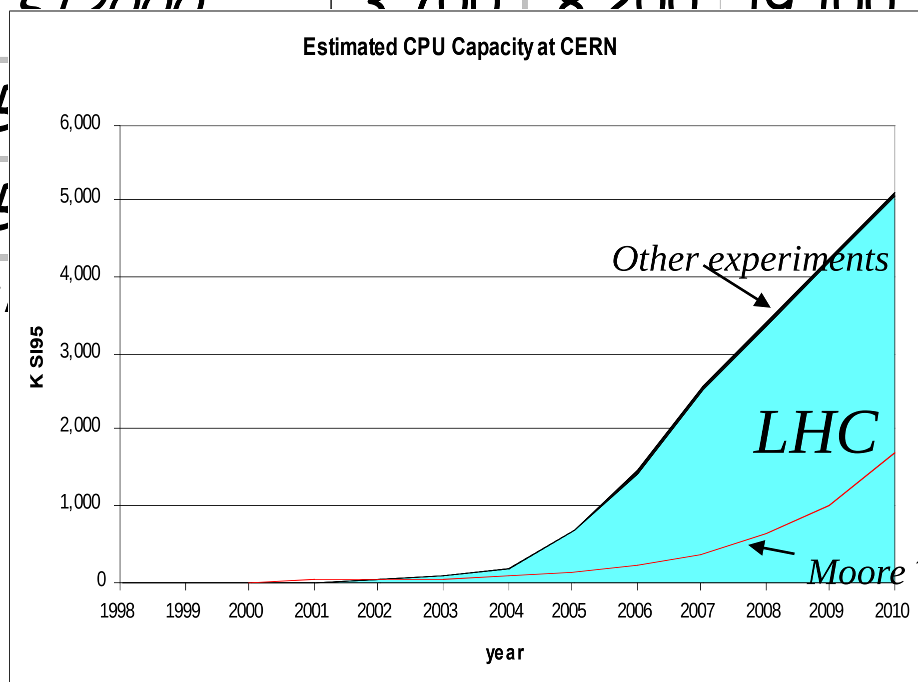


CERN Centre Capacity Requirements for all

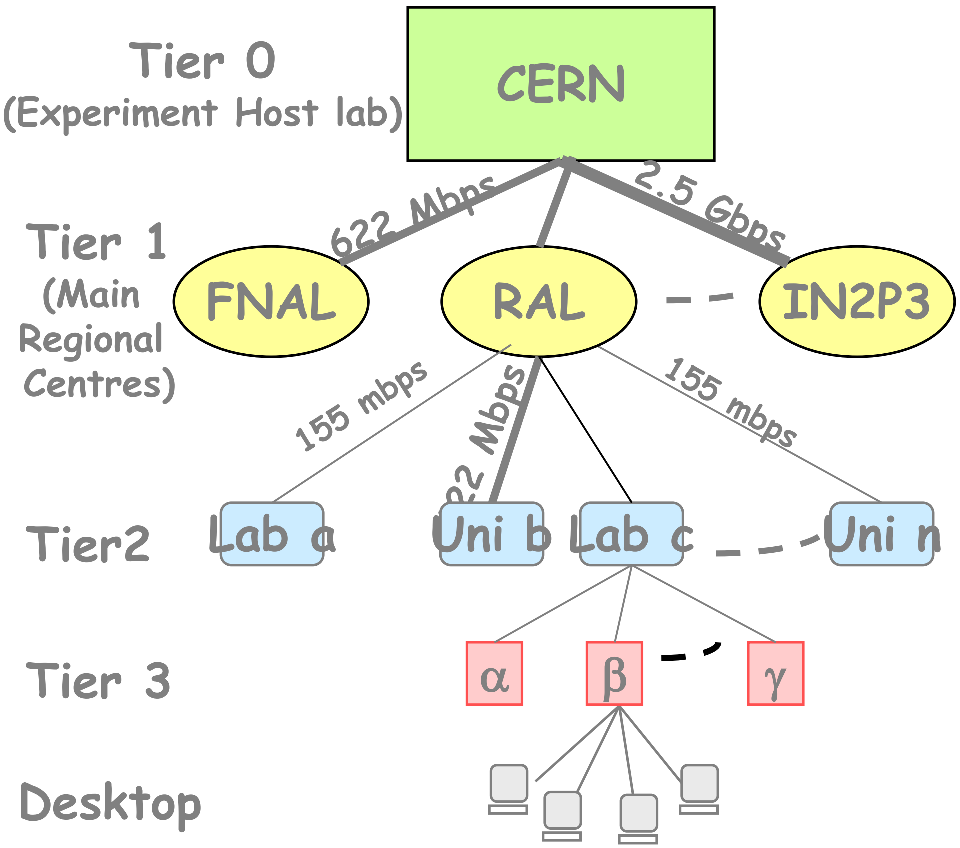




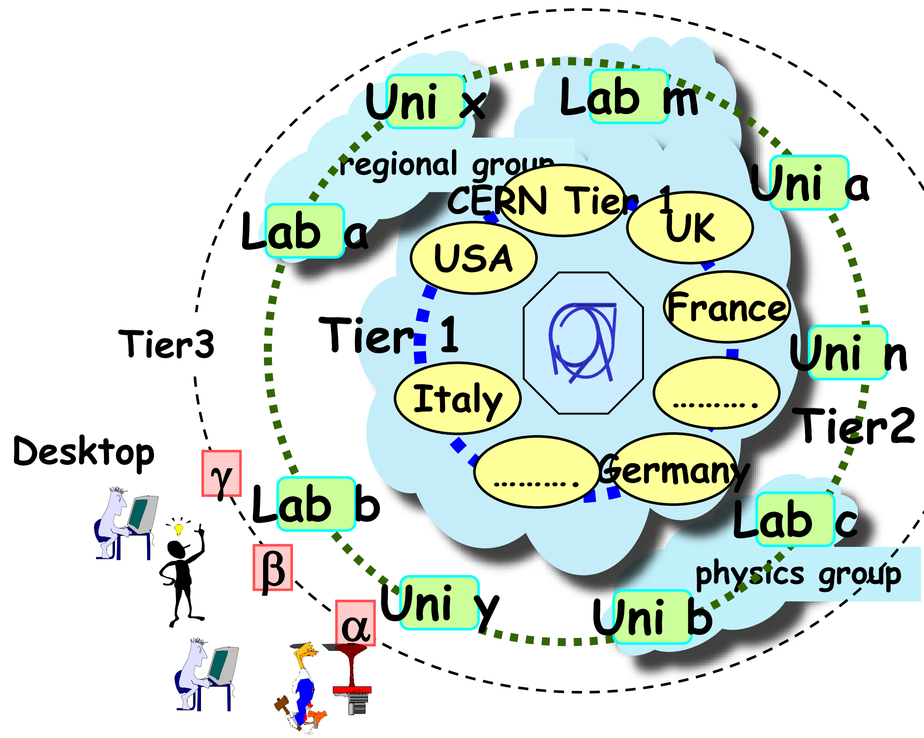
processing	K SI2000	3,700	8,200	19,100	25,000	34,000
disk	PB				5.0	6.7
tape media	PB				36	48
tape I/O	GB/s				3.9	3.9



A Multi-Tier Computing Model



Manager View



User View

Distributed Analysis - the real challenge

- Analysis will be performed with a mix of "official" experiment software and private user code
  - How can we make sure that the user code can execute and provide a correct result wherever it "lands"?
- Input datasets not necessarily known a-priori
- Possibly very sparse data access pattern when only a very few events match the query
- Large number of people submitting jobs concurrently and in an uncoordinated fashion resulting into a chaotic workload
- Wide range of user expertise
- Need for interactivity - requirements on system response time rather than throughput
- Ability to "suspend" an interactive session and resume it later, in a different location
- Need a continuous dialogue between developers and users