Minisymposium on Analysis and Representation of Large Data Sets

Visualization on large data volumes in Physics

Seeing is believing?

Madrid, 15 February 2012

Jesús Marco de Lucas

CSIC Research Professor @ **Instituto de Física de Cantabria**

marco@ifca.unican.es

(special thanks to E.Martinez, F.Castejon, L.Cabellos)

Outlook

Examples:

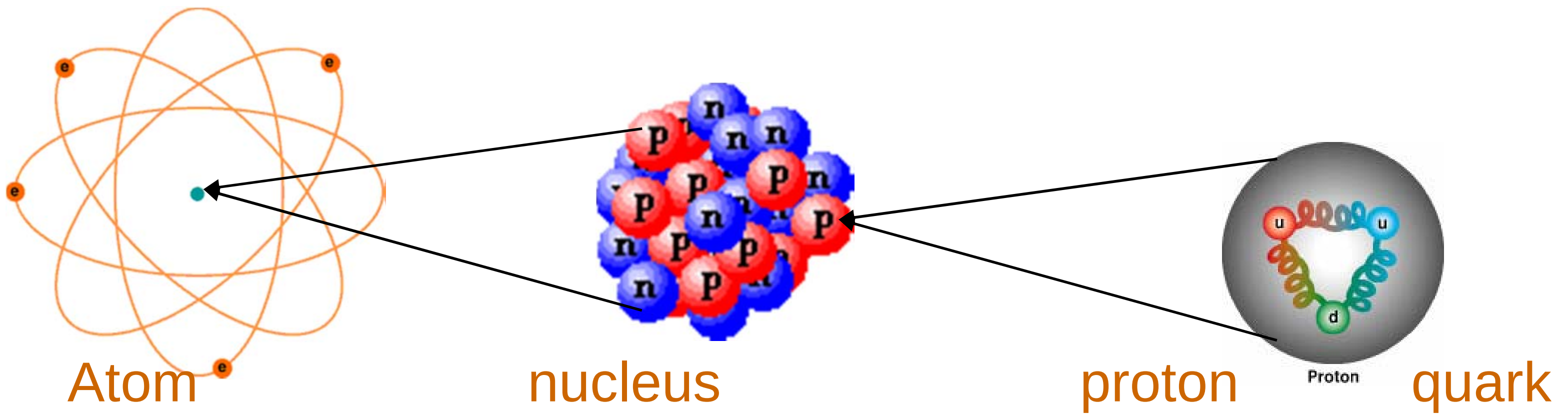
- Event Displays in Particle Physics
- The Cosmic Microwave Background Maps
- Visualization of Trajectories in a Fusion device

What else:

- Dataflow hardware and software
- Complex systems

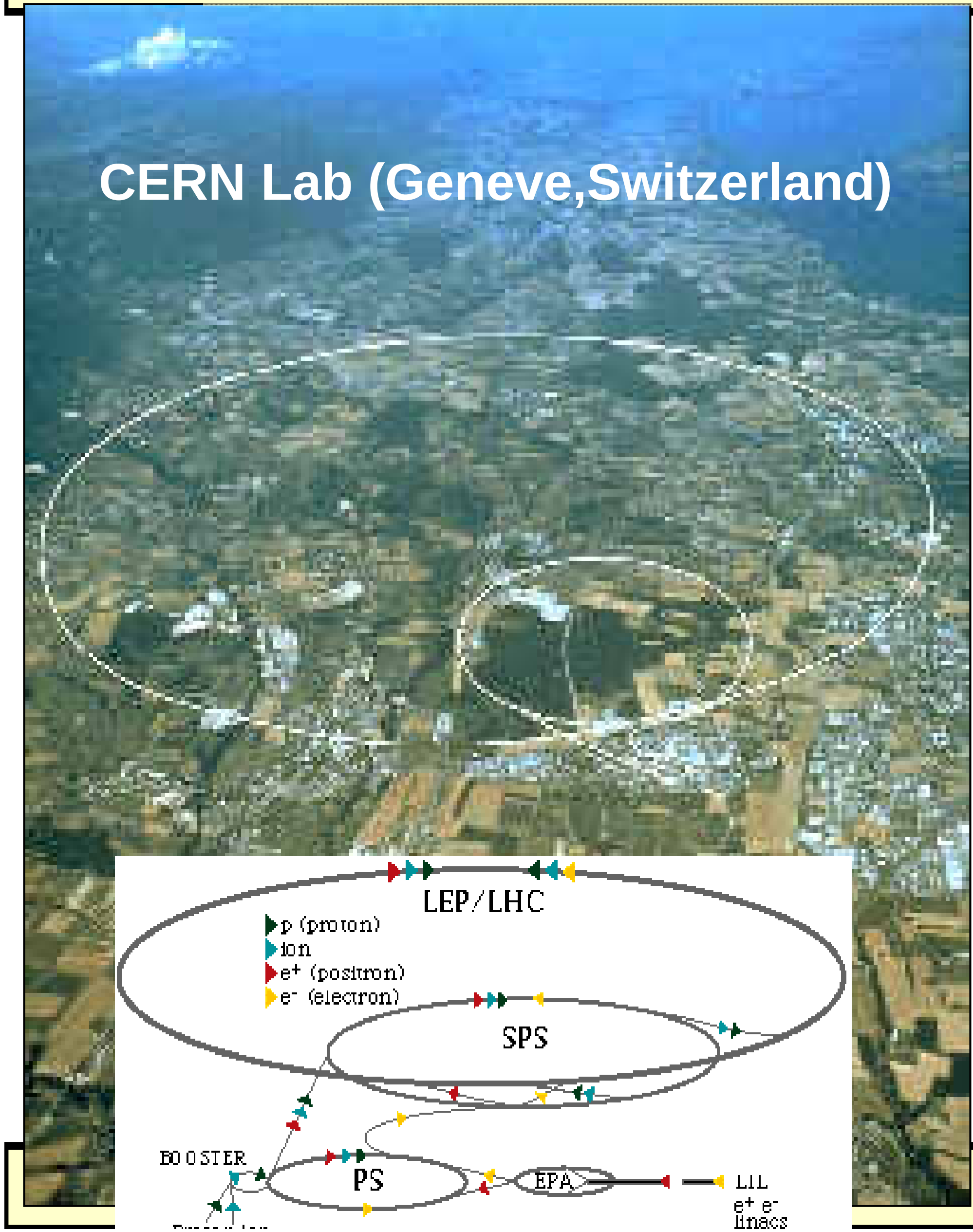
Searches in Particle Physics

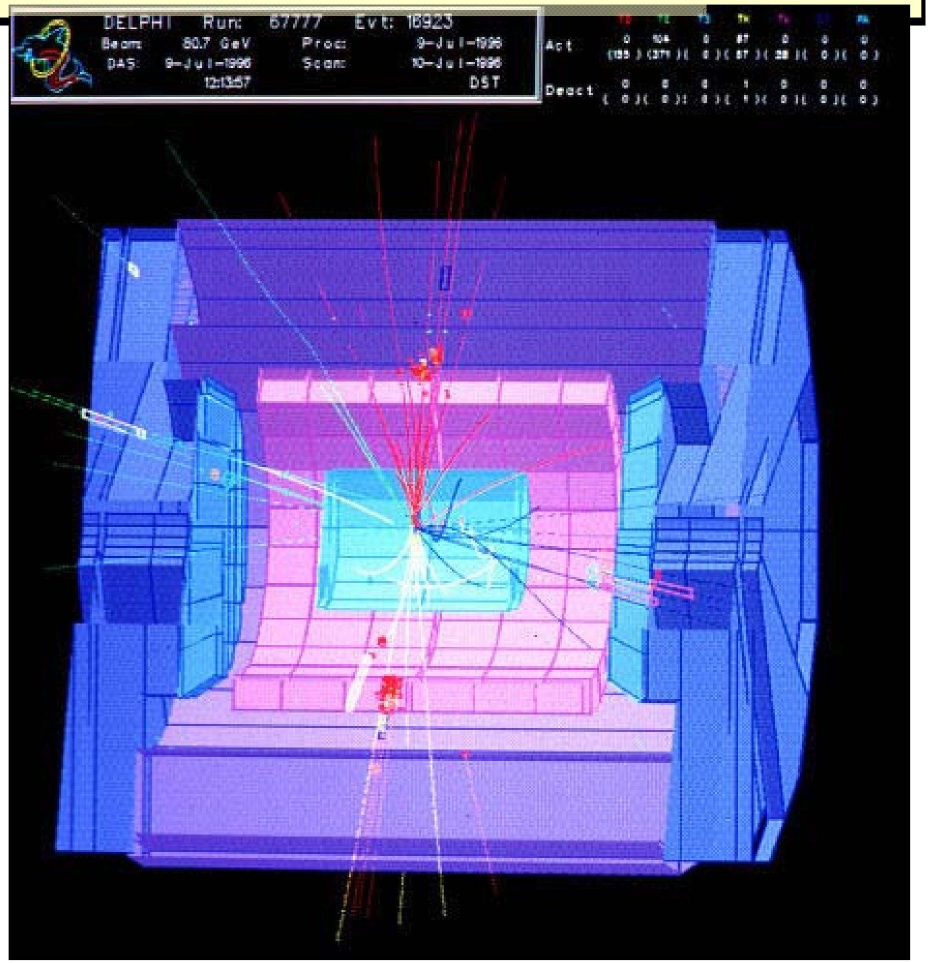
- Particle physics: studying the basic constituents of all matter around!



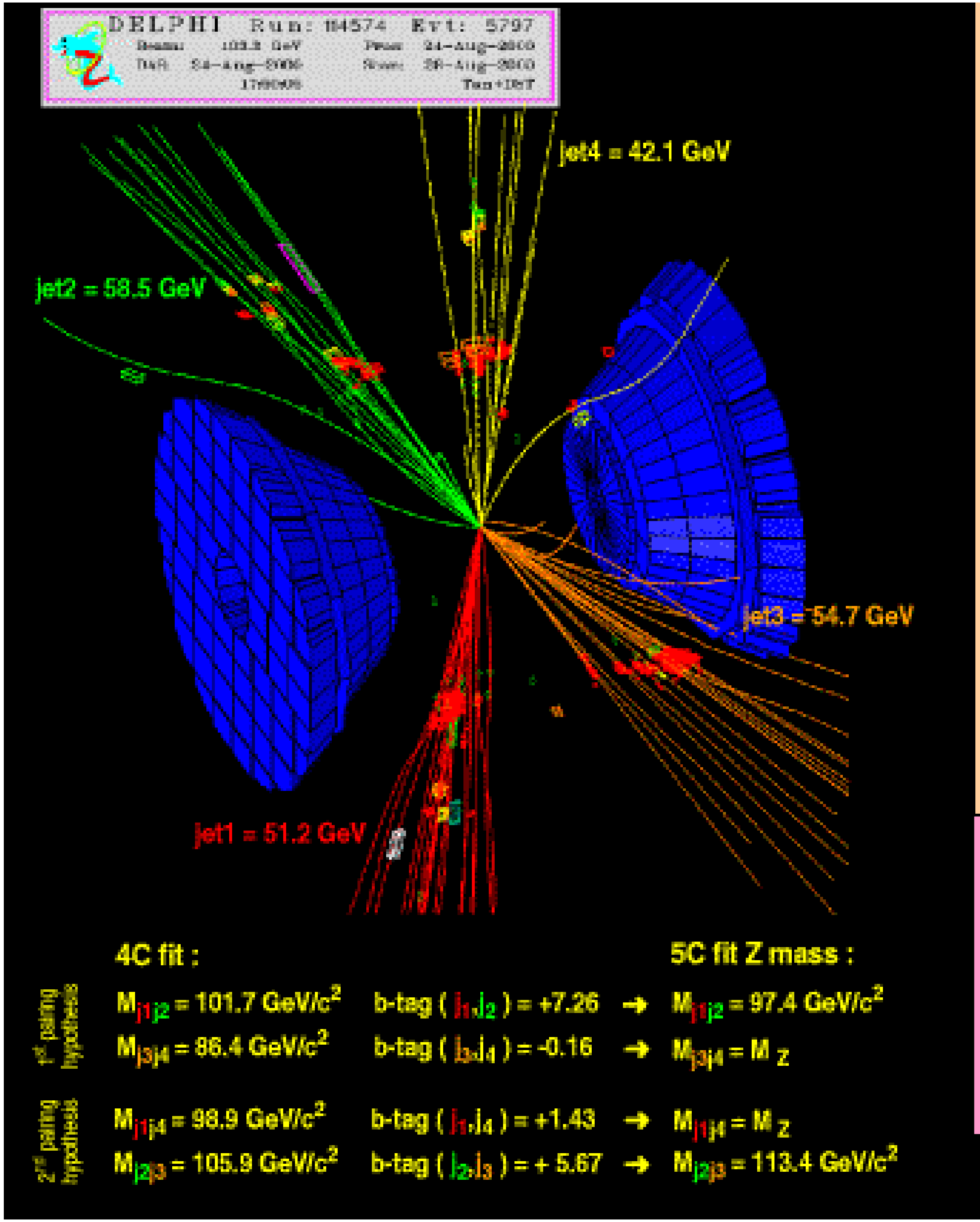
The origin of the *“mass”* of all particles is linked to a fundamental particle predicted but not yet discovered: **the Higgs boson**

Accelerators and detectors





LEP (ended in 2000)
Large Electron Positron Collider
E_{cm}= 200 GeV, e+e- collisions
Search for Higgs up to M=115 GeV

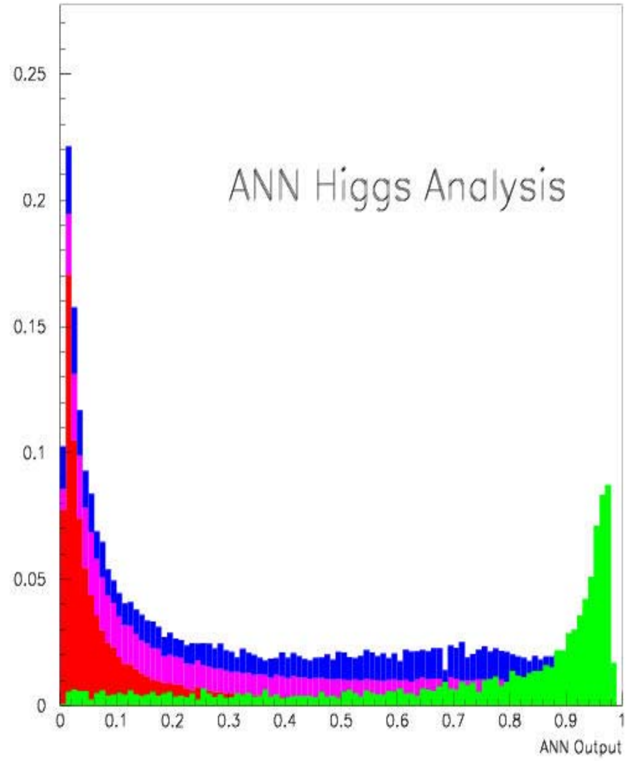


- Higgs Boson decays into two heavy particles:
  - B quarks
  - Z or W bosons
- Complex event characteristics:
  - Shape
  - Particles lifetime
  - Masses
- S/B ratio is extremely low:
  - LEP: 10 in 10^6
  - LHC: 10 in 10^9

NEURAL NETWORKS are used to optimize the search

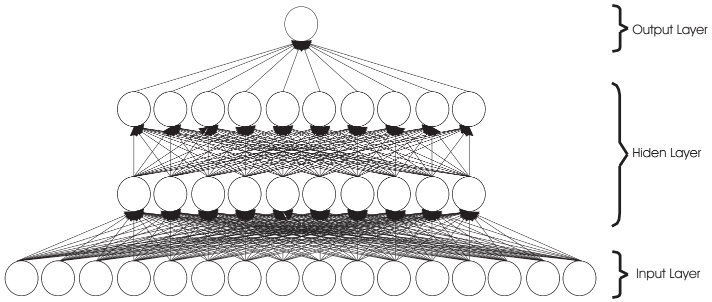
How did we select that "collision"?

- ANN: example of architecture 16-10-10-1
  - 16 input variables
  - 2 hidden layers with 10 nodes each
  - 1 output layer, 1=signal, 0=background
- Trained on MC sample
  - Higgs generated at a given mass value
  - All types of Background
  - 10x real data statistics
- Applied on real collected data to order in S/B the candidates to Higgs boson
- Training process:
  - Minimize classification "error"
  - Iterative process
  - No clear "best strategy"
- Computing Intensive: hours to days for each "try"

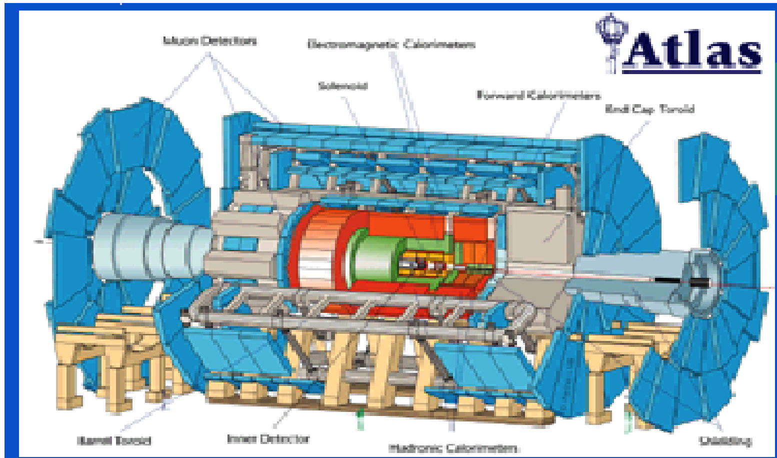


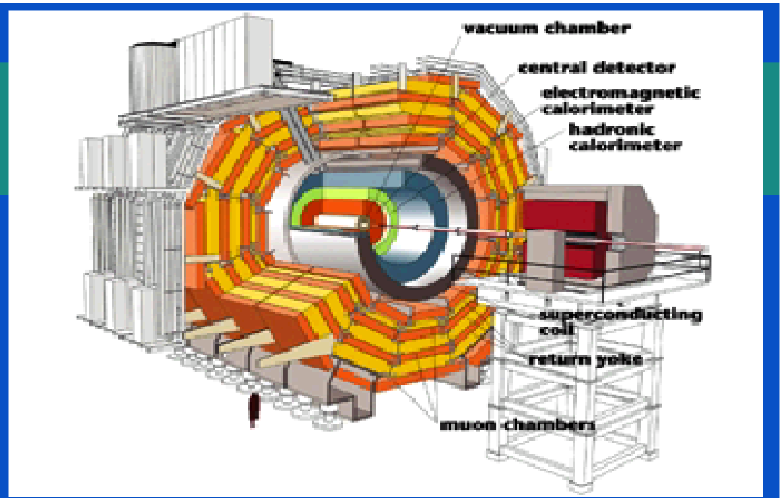
Distributed Analysis

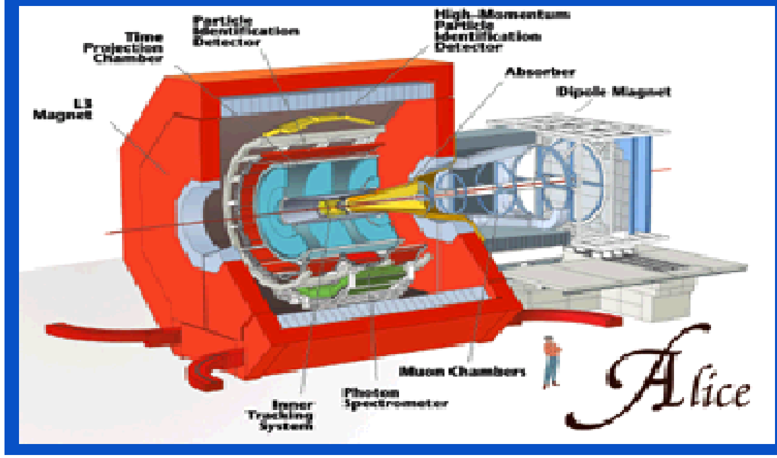
- Distributed Configuration:
  - Master node and N worker nodes.
    - Scan to filter events & select variables
    - ResultSet in XML, split according to N (number of slave nodes)
- Training procedure:
  - Master reads input parameters and sets the initial weights to random values.
  - The training data is distributed to the workers.
  - At each step:
    - The master sends the weights to the workers.
    - The workers compute the error and the gradient and return them to the master.
- This training procedure has been implemented using MPI and adapting the MLP-fit package.
- Conditions:
  - train an ANN with 650.000 simulated realistic LEP events, 20000 of them corresponding to signal events.
  - Use a 16-10-10-1 architecture (270 weights)
  - Need 1000 epochs training.
  - Similar sized samples for the test.
  - BFGS training method.

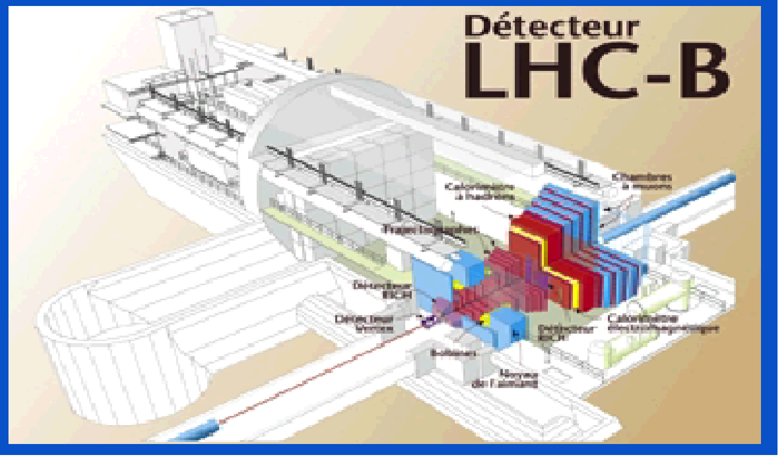


LHC Experiments

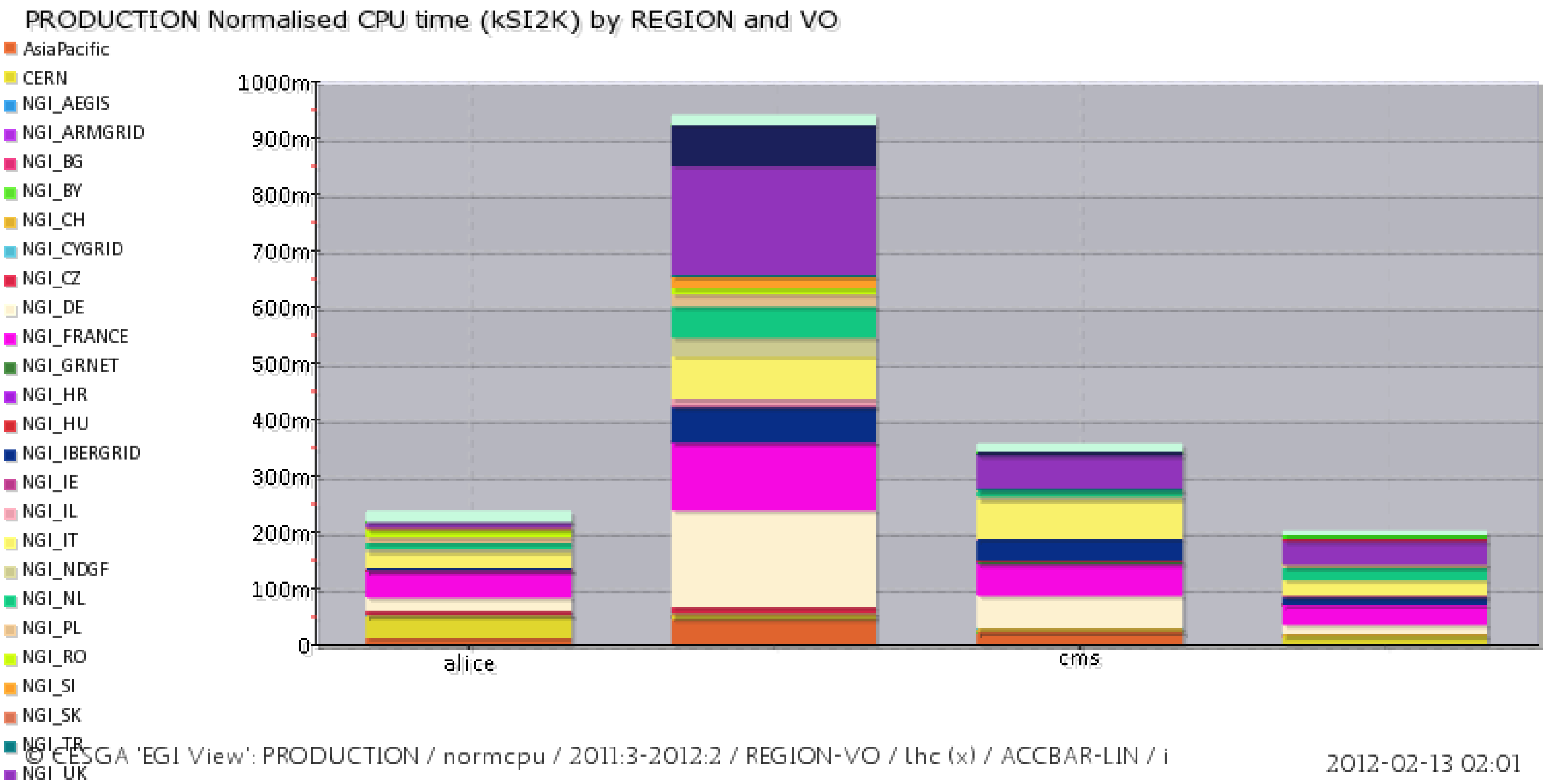






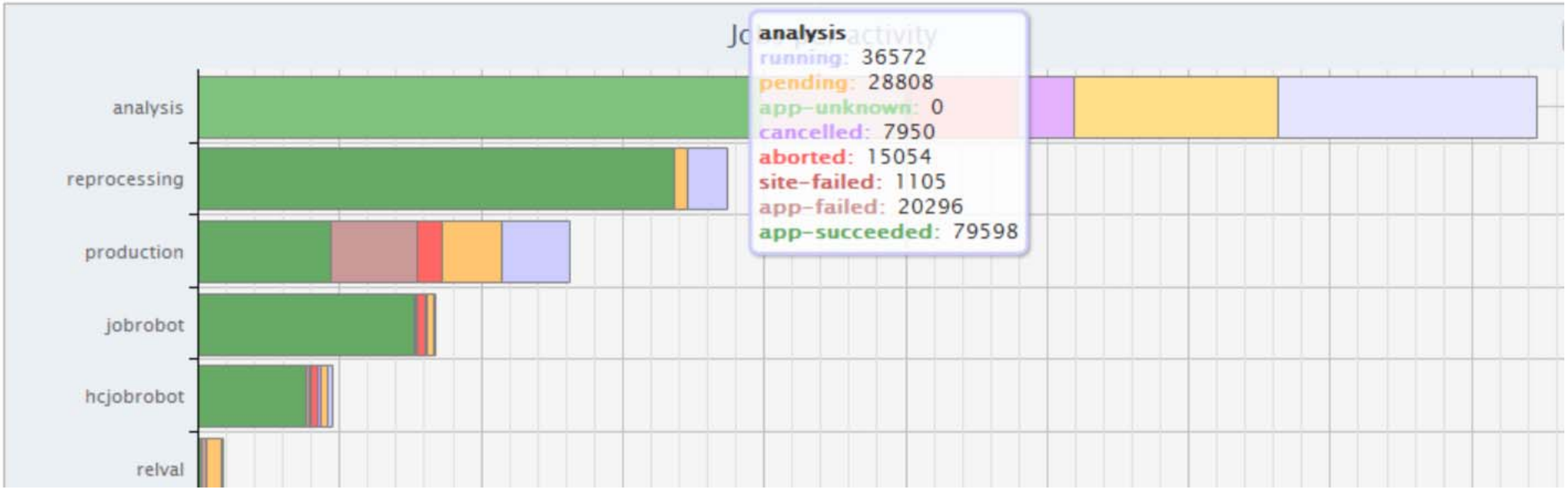


The GRID success!

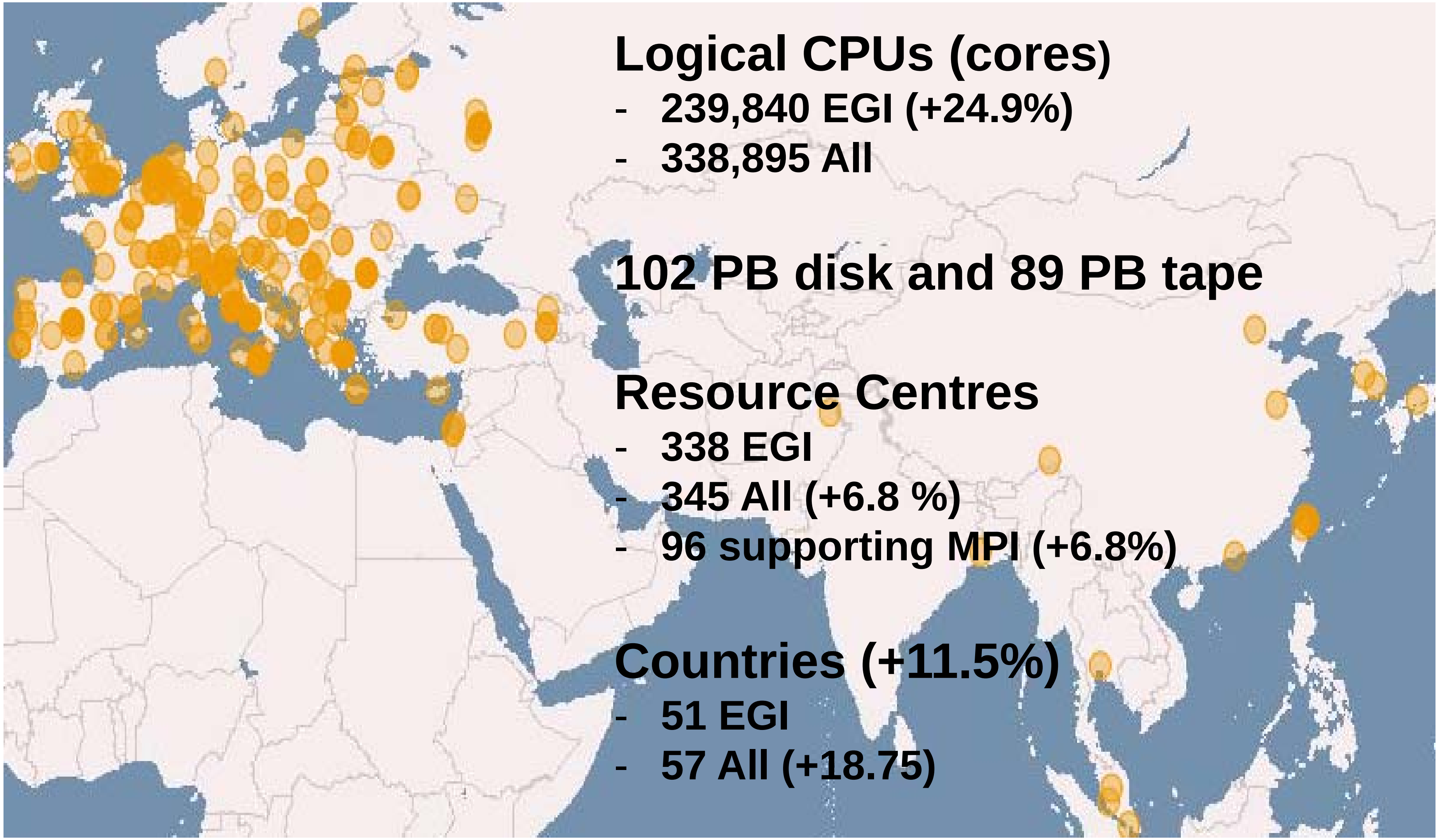


Processing...

Main Jobs Chart



European Grid Infrastructure
(2011)



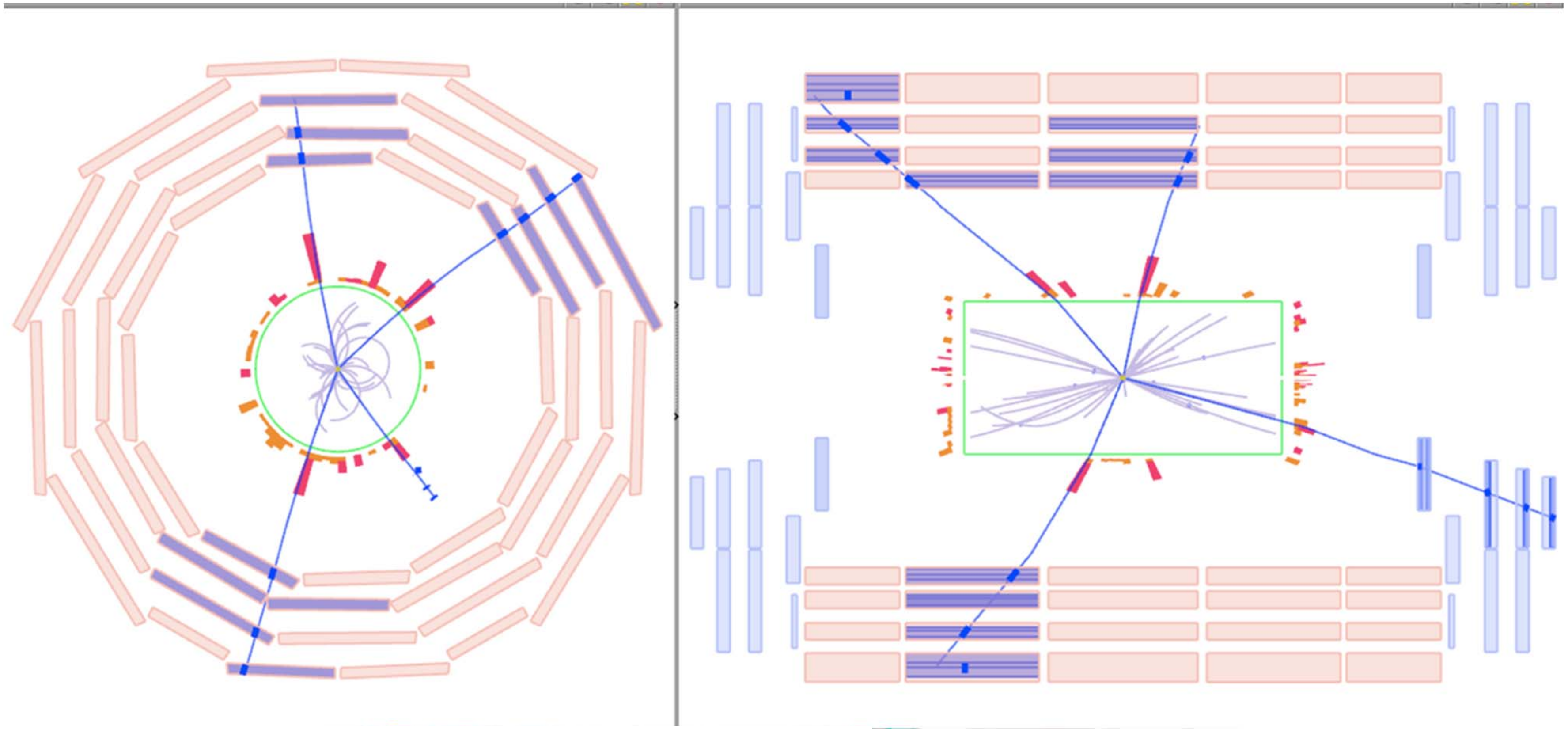
Logical CPUs (cores)
- 239,840 EGI (+24.9%)
- 338,895 All

102 PB disk and 89 PB tape

Resource Centres
- 338 EGI
- 345 All (+6.8 %)
- 96 supporting MPI (+6.8%)

Countries (+11.5%)
- 51 EGI
- 57 All (+18.75)

LHC collisions: a ZZ event

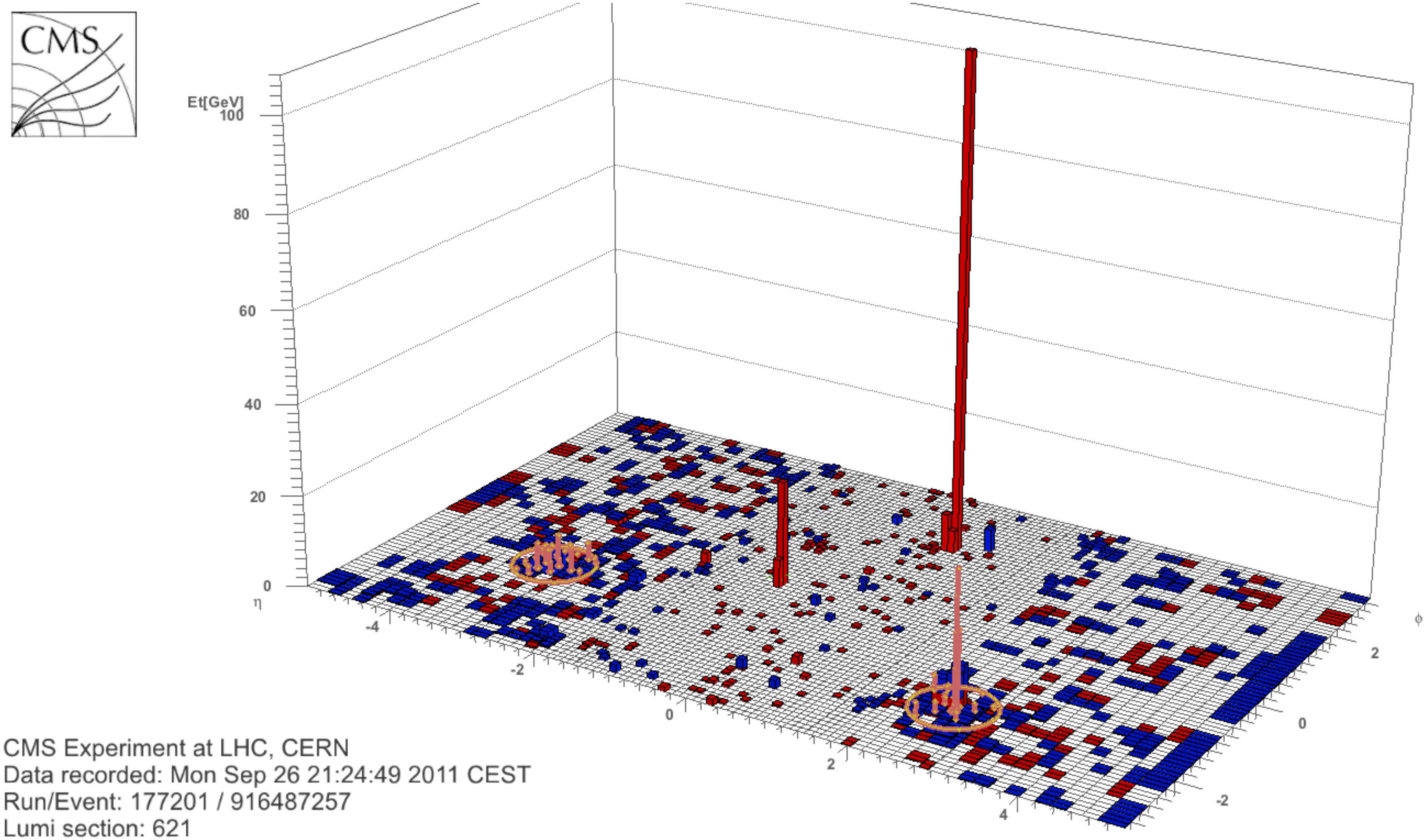


Invariant Masses

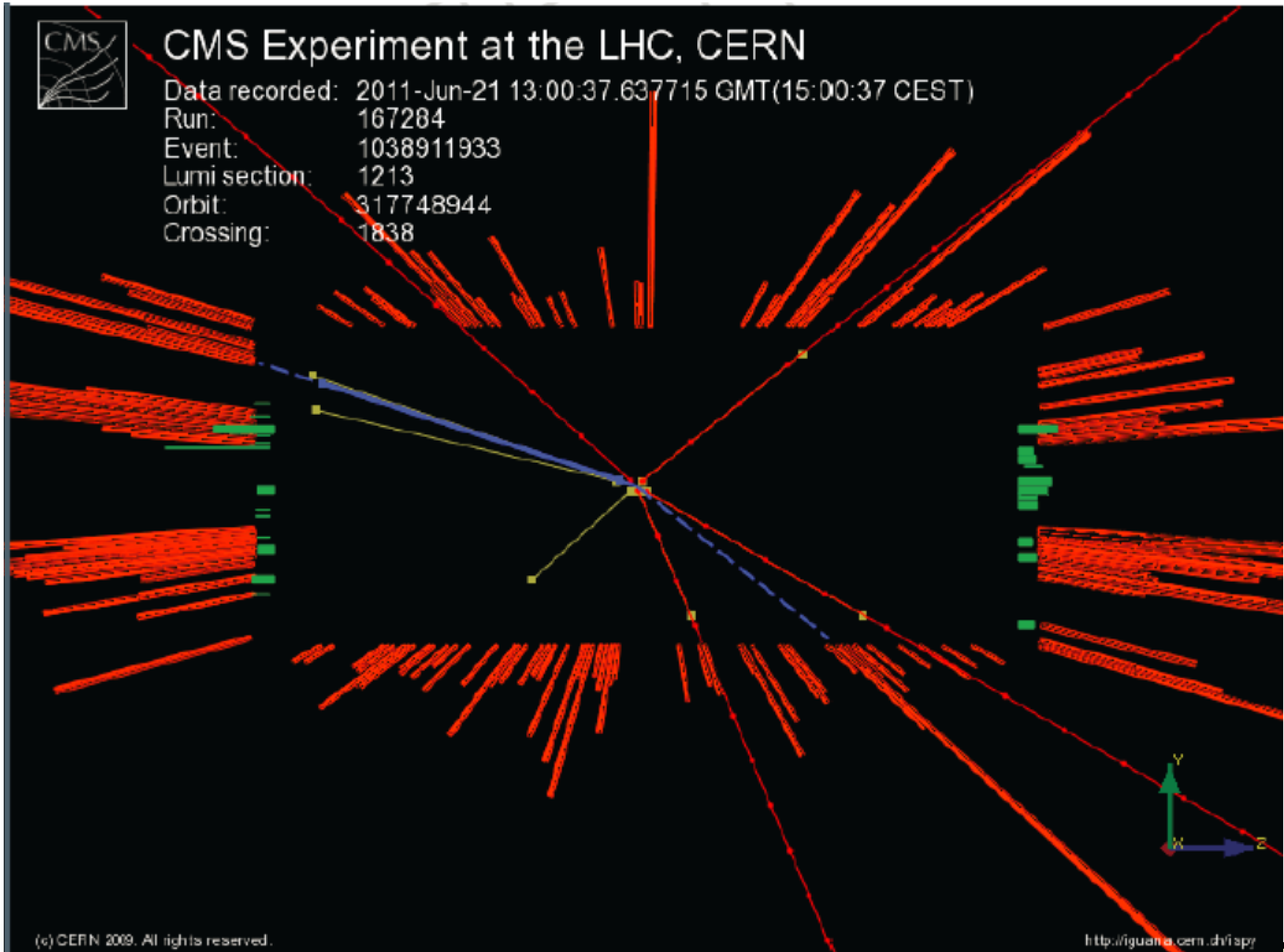
$\mu_0 + \mu_1$: 92.15 GeV (total(Z) p_T 26.5 GeV, ϕ -3.03),
$\mu_2 + \mu_3$: 92.24 GeV (total(Z) p_T 29.4 GeV, ϕ +.06),
$\mu_0 + \mu_2$: 70.12 GeV (total p_T 27 GeV),
$\mu_3 + \mu_1$: 83.1 GeV (total p_T 26.1 GeV).

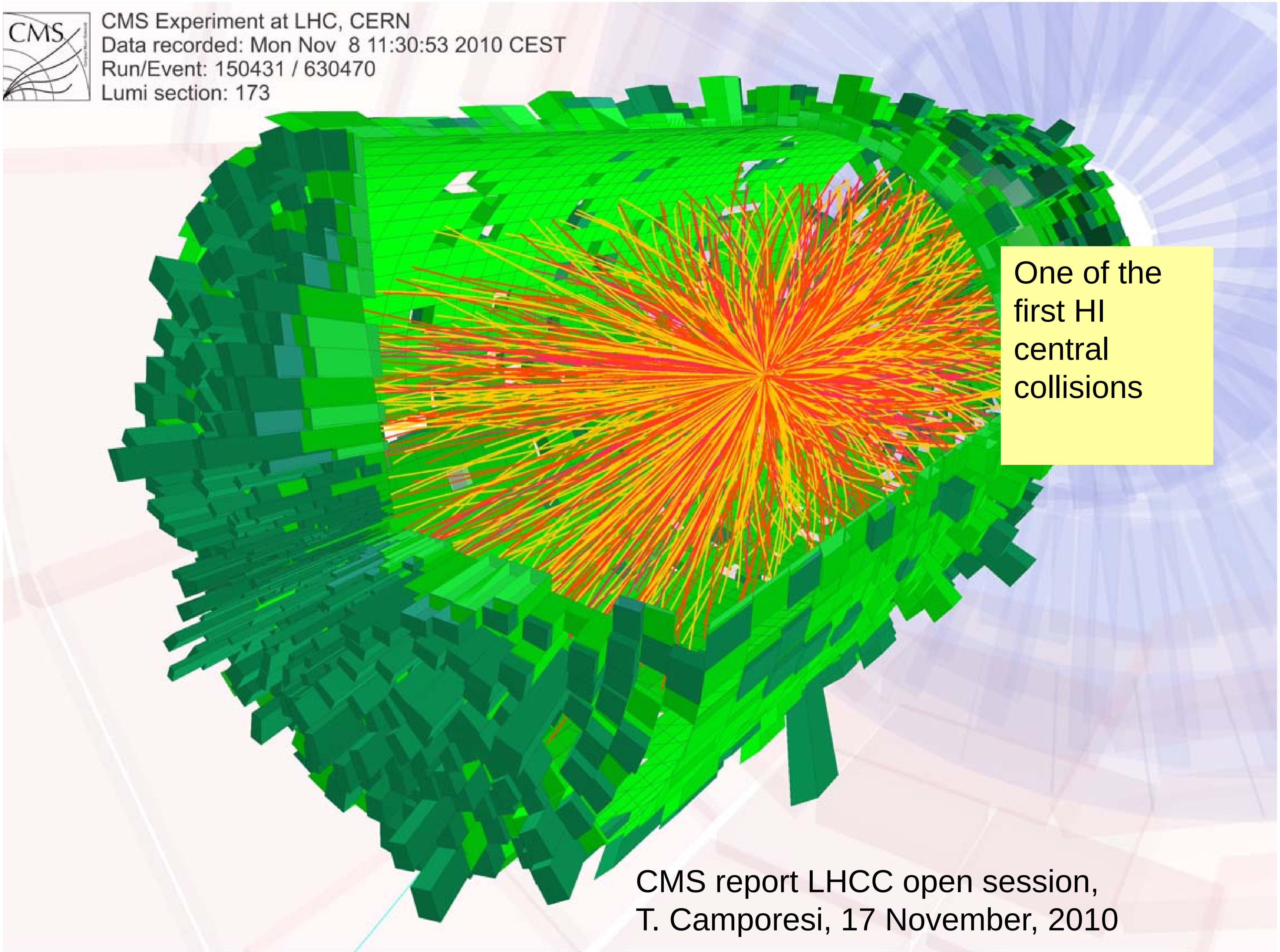
Invariant Mass of 4μ: 201 GeV

Higgs candidate decaying into $\gamma\gamma$



Higgs into 4 leptons





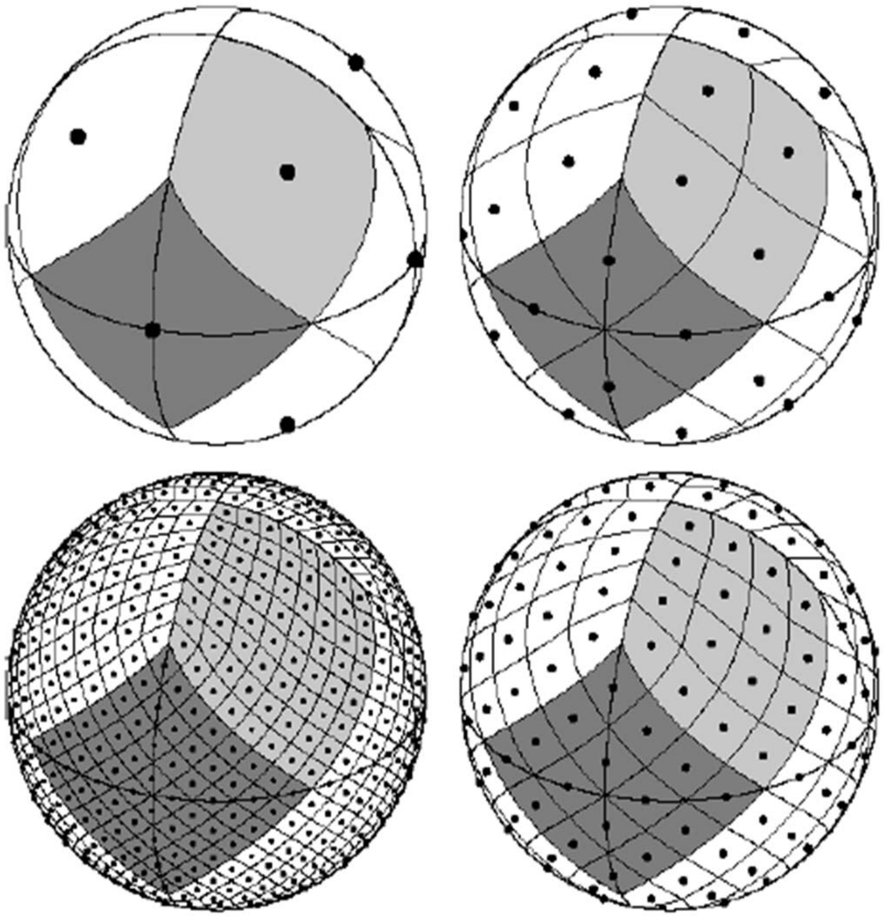
One of the first HI central collisions

CMS report LHCC open session,
T. Camporesi, 17 November, 2010

Symmetries in the Sky

- An example: spherical topology
  - Very relevant in Astronomy
  - But also in geophysics and in atomic and nuclear physics
- *"(There is) no known point set analog to uniform sampling in Euclidean space"*
- CMB Sky images
  - Topic: Cosmic Microwave Background (CMB) **anisotropies**
  - Data: multifrequency/high resolution temperature and polarization microwave sky observation
    - NASA WMAP mission
    - ESA PLANCK mission
  - O(200 Mbytes) images

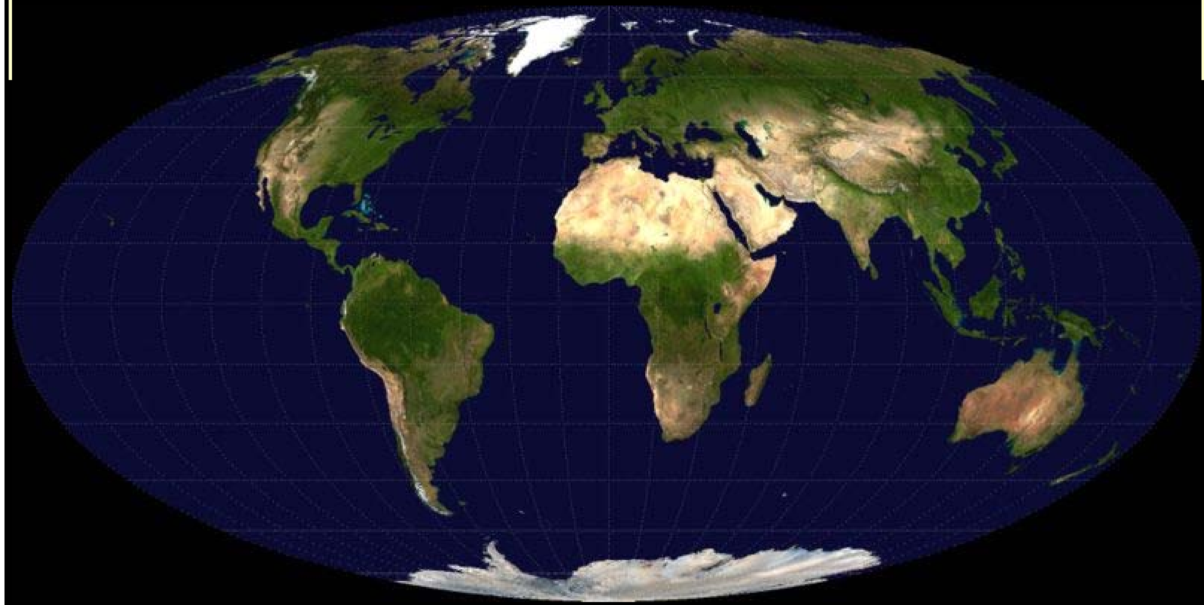
Discretization

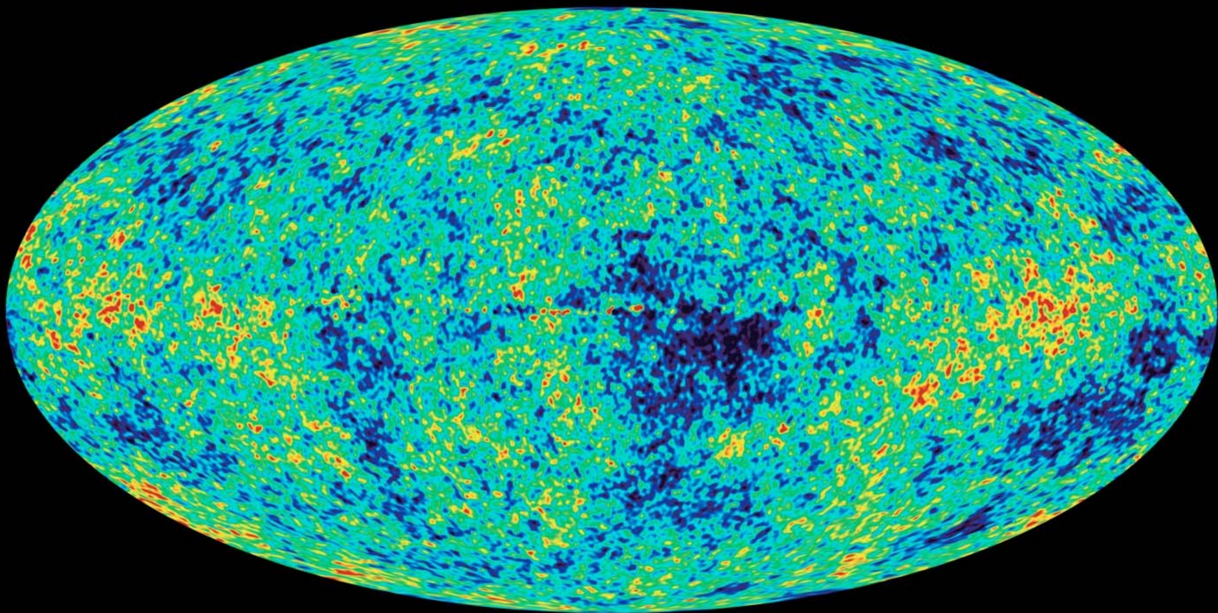


- HEALPIX
  - Hierarchical, Equal Area, isoLatitude
  - Systematic effects?
  - Few pixels per resolution element
- ~10^6 pixels x o(100) channels
- Objective:
  - Fourier analysis with spherical harmonics
  - Wavelet decomposition
  - Near neighbor searches

Sky maps

Proyección de Mollweide

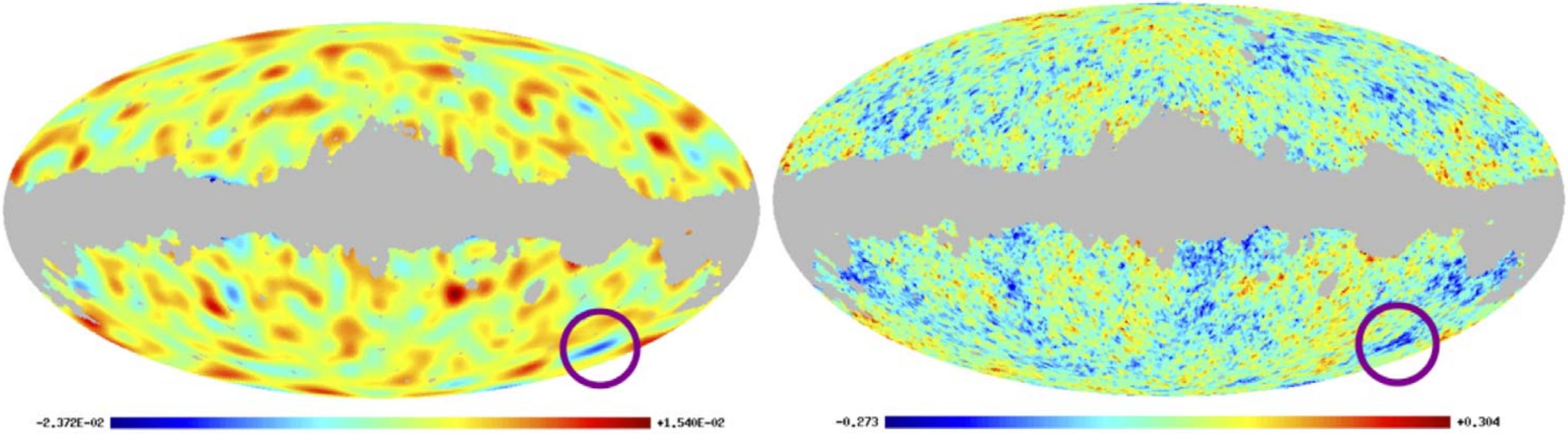




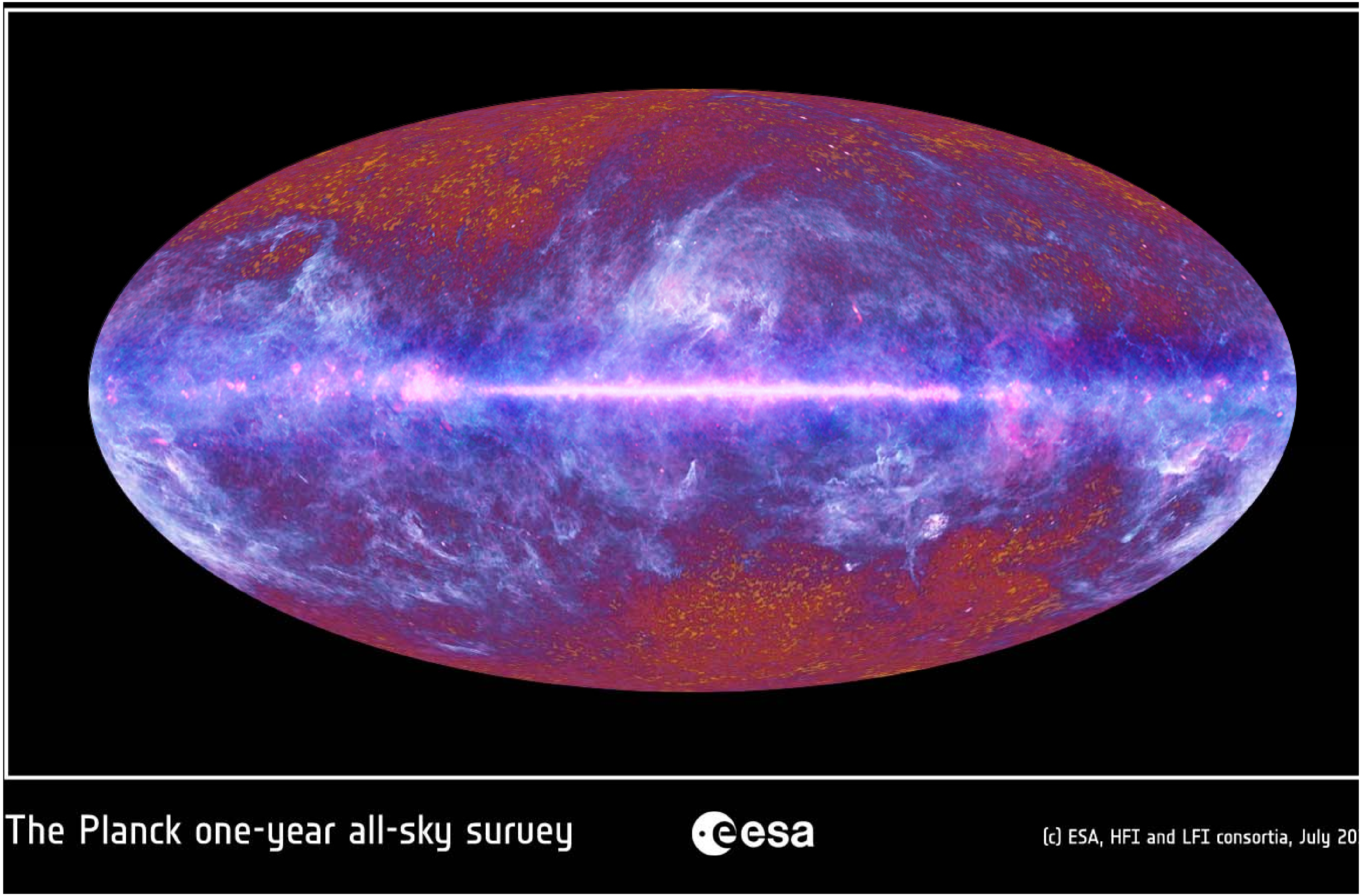
IFCA Instituto de Física de Cantabria

CSIC Consejo Superior de Investigaciones Científicas

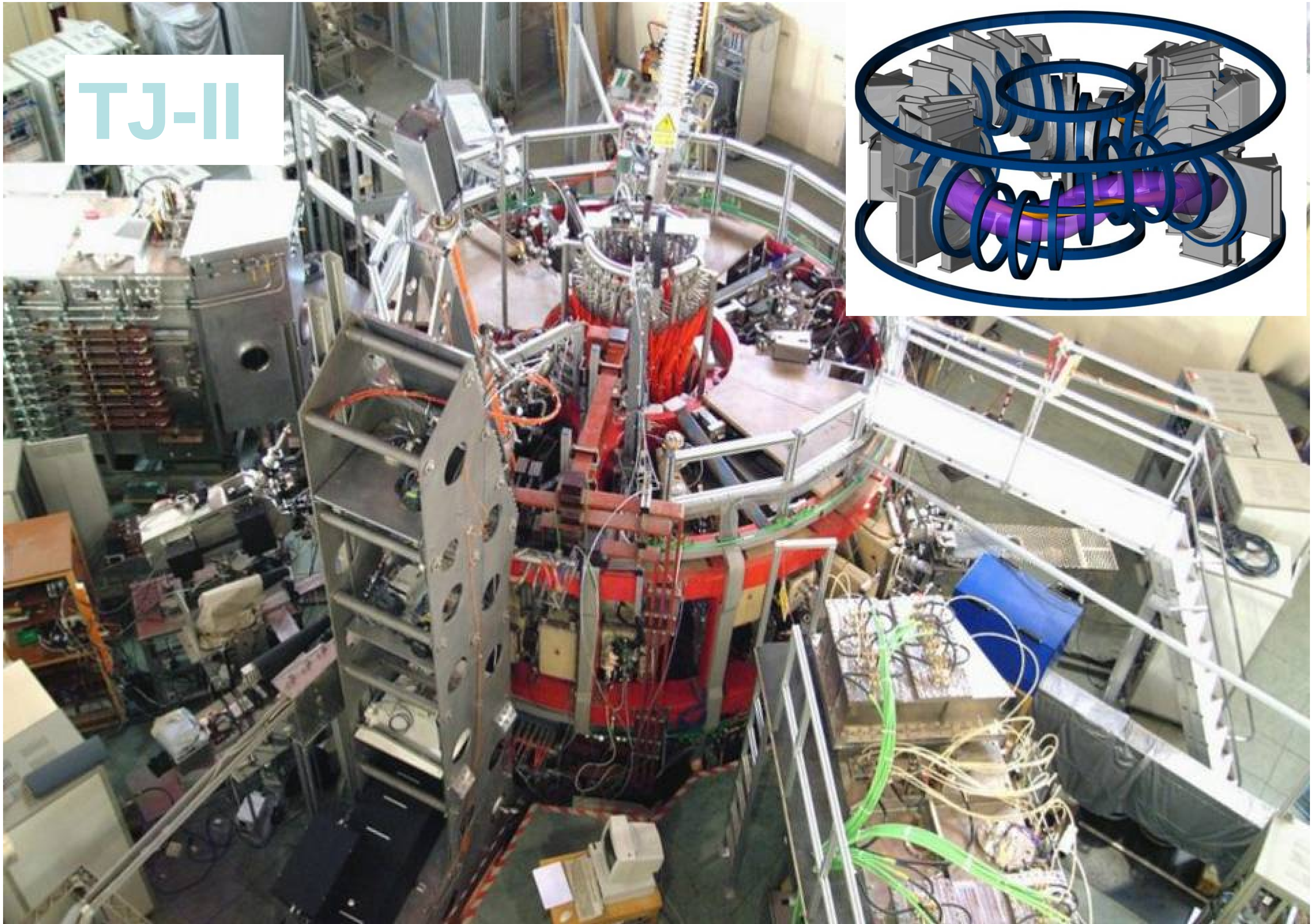
Analysis: non-gaussianity sources



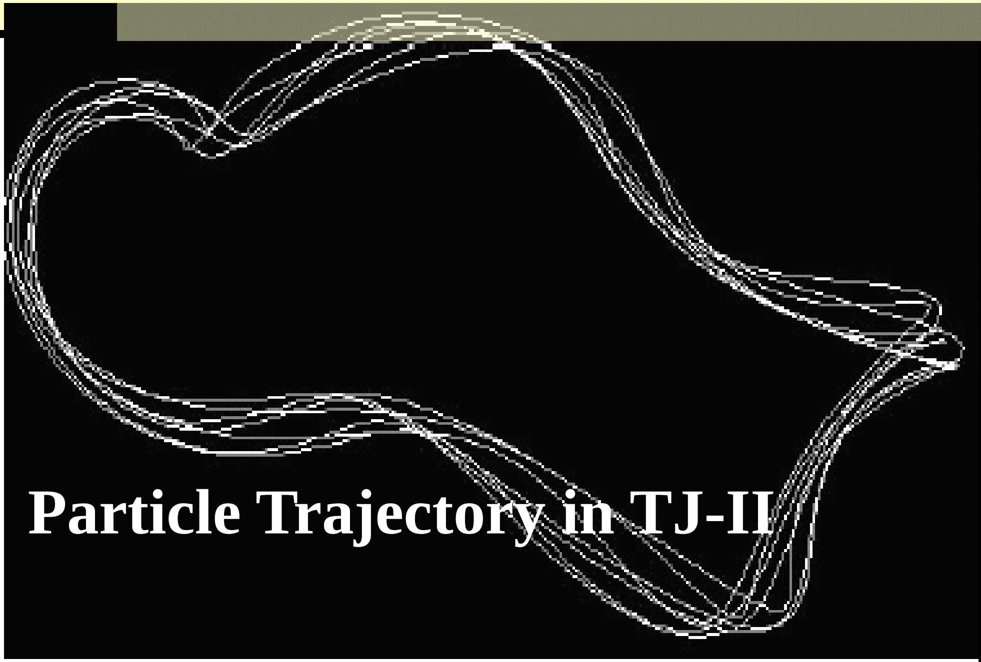
- Example of analysis on WMAP (P.Vielva, E.Martinez, et al. 2004)
- Skewness and the kurtosis of the SMHW (spherical Mexican hat wavelet) coefficients are calculated at different scales (ranging from a few arc minutes to tens of degrees).
- A non-Gaussian signal is detected at scales of the SMHW around 4^{O} (size in the sky of around 10^{O})

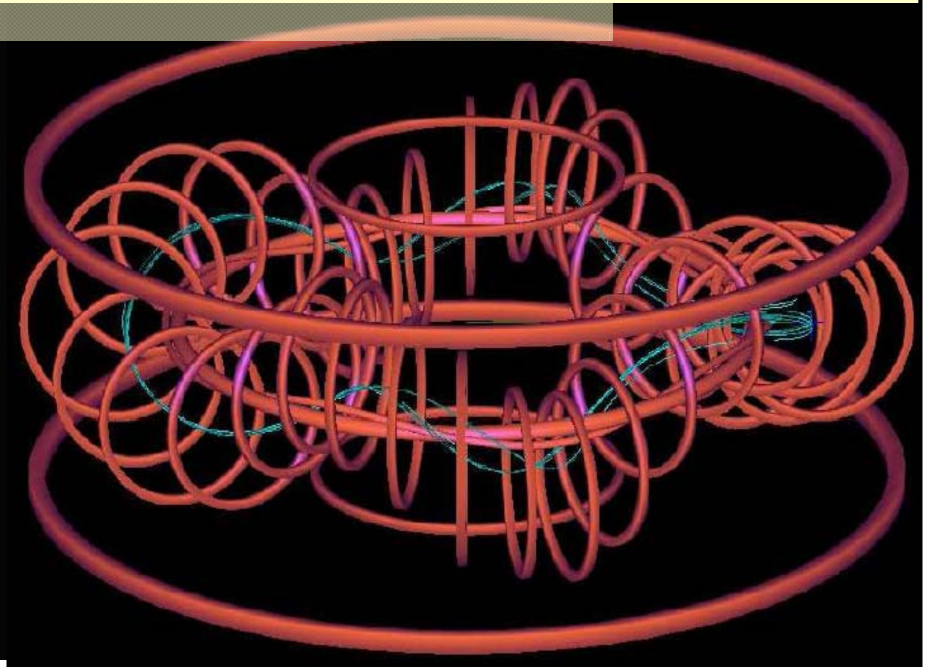

The Planck one-year all-sky survey
esa
(c) ESA, HFI and LFI consortia, July 20

TJ-II



ISDEP: The Collisional Transport





-ISDEP solves the distribution function of a population of test particles in a fusion device: Test particles interact with a static plasma background.

F. Castejon, et al., Plasma Physics and Controlled Fusion 49 (2007) 753

J. L. Velasco, F. Castejón and A. Tarancón. Physics of Plasmas 16 (2009) 052303

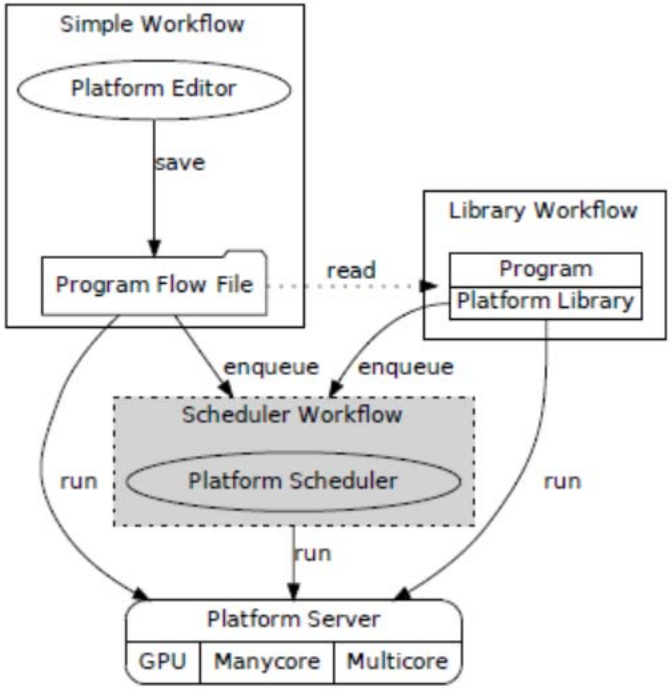
- **Scales perfectly in distributed computing platforms**
- **Interactive European Grid: visualization**

What else?

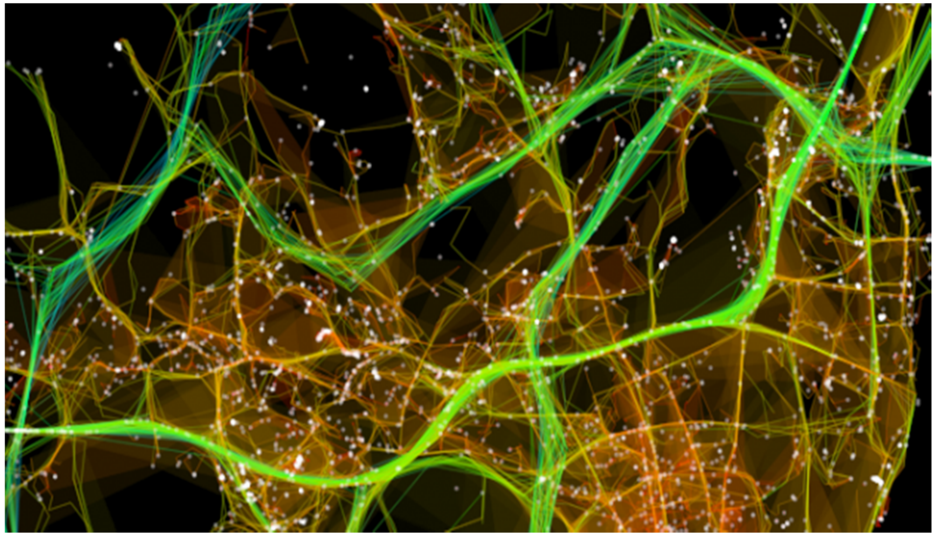
- Technology to handle large data volumes
  - Large storage is not a large problem
    - (1 Petabyte ~ 300 K€ + 10KW 3TB disks, 4 GPFS servers, 8G FC+10GbEth)
  - Fast processing "is" a problem:
    - Ex: SandyBridge processor + 2 SSD can handle up to 500MB/s
    - 100 GB in 1 node takes 200 s (base time)
    - 1 TB in 1 node will take 2000 s (base time)
    - While 10 TB in 10 nodes would take 200s
  - Distributed or Parallel storage?
  - Accommodate Interactive priority in the cluster/grid/cloud?
    - Dataflow oriented architecture

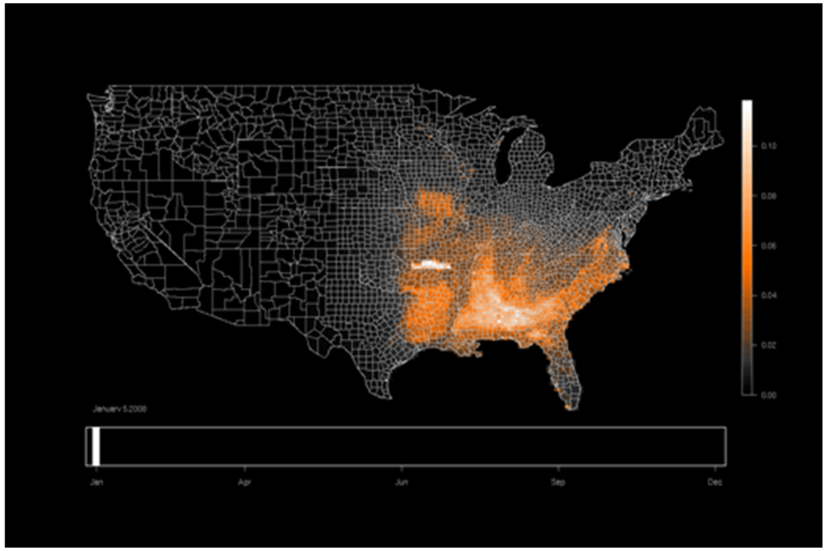
What else?

- Data flow orientation + hardware integration
  - New hardware, GPUs in particular, is very powerful
  - (Efficiently) Programming new hardware is a problem
  - Handling dataflows is not an easy task
    - Example: EUFORIA project: KEPLER: supercomputing + grid
- Working on a new model: SKEMA
  - Graphical interface
  - Handling of hardware using OpenCL
  - First simple examples:
    - FFT
    - Image Compression



What next?





- Global objective: pattern matching in a context