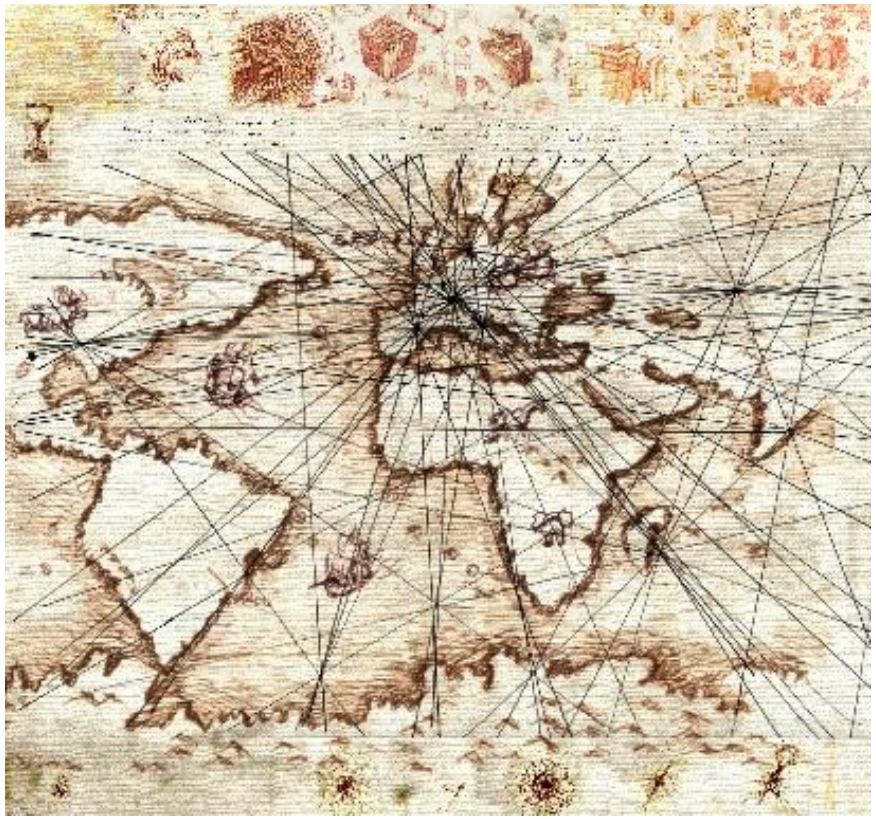
Large Datasets and High Energy Physics (HEP)



Peter Elmer, Princeton University
Frontiers in Diagnostic Technologies
29 Nov, 2011

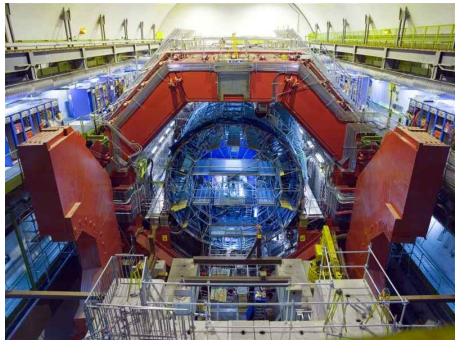
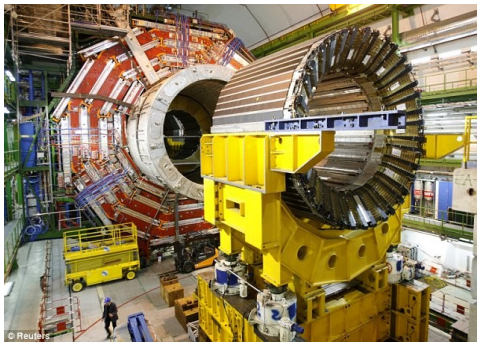
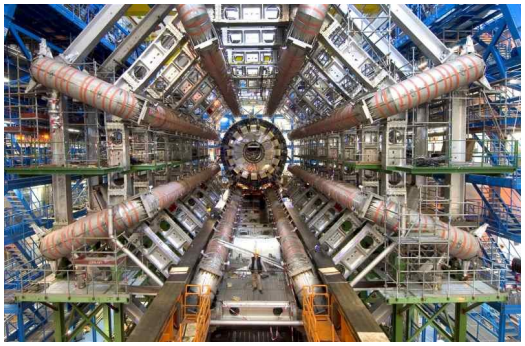
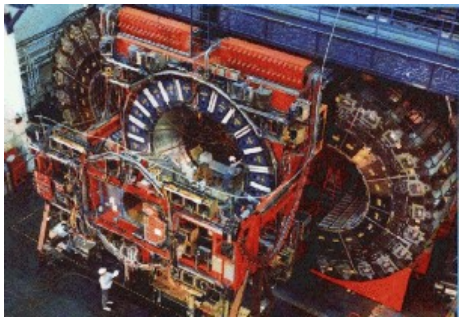
Overview

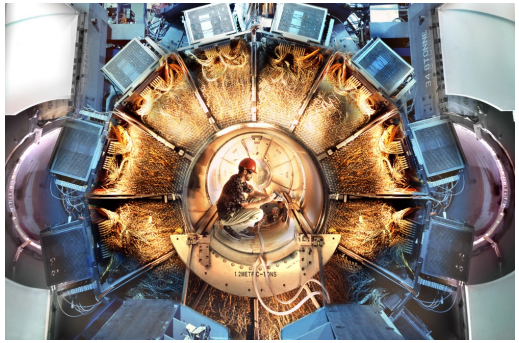
Abstract: High Energy Physics (HEP) has a long tradition of pushing scientific computing to its limits. In particular the use of very large datasets has always been required in order to search for the rare phenomena of interest in the field. This presentation will describe the manner in which large datasets are acquired, managed and analyzed in HEP, using the Large Hadron Collider (LHC) at CERN as a Concrete example.

High Energy Physics (HEP)

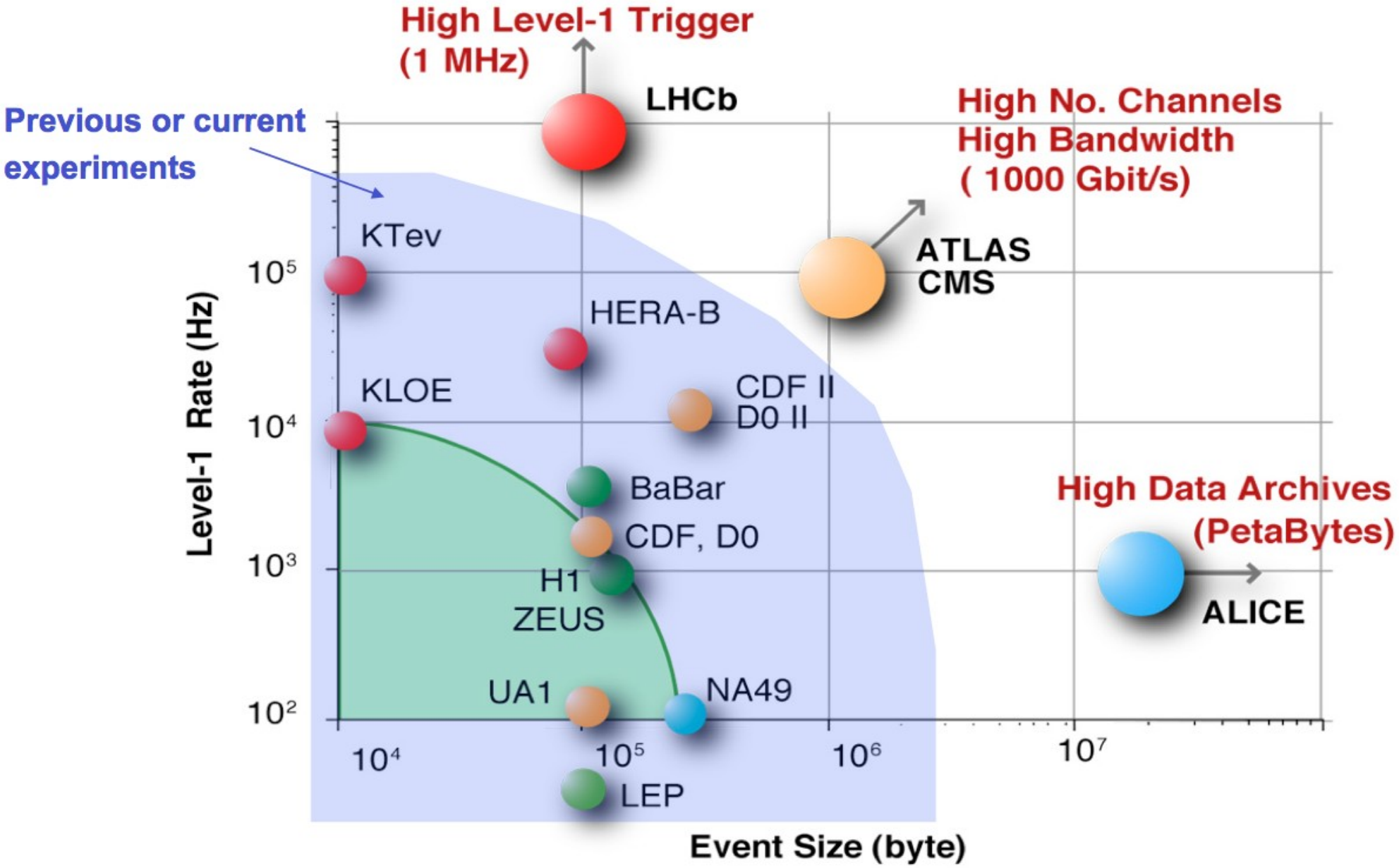
HEP generates (scientifically interesting) data by:

- observing the interaction of non-man-made (cosmic) particles with the earth or a target
- creating/accelerating a beam of some type of particle and directing them at a fixed target
- colliding two beams of accelerated particles with one another

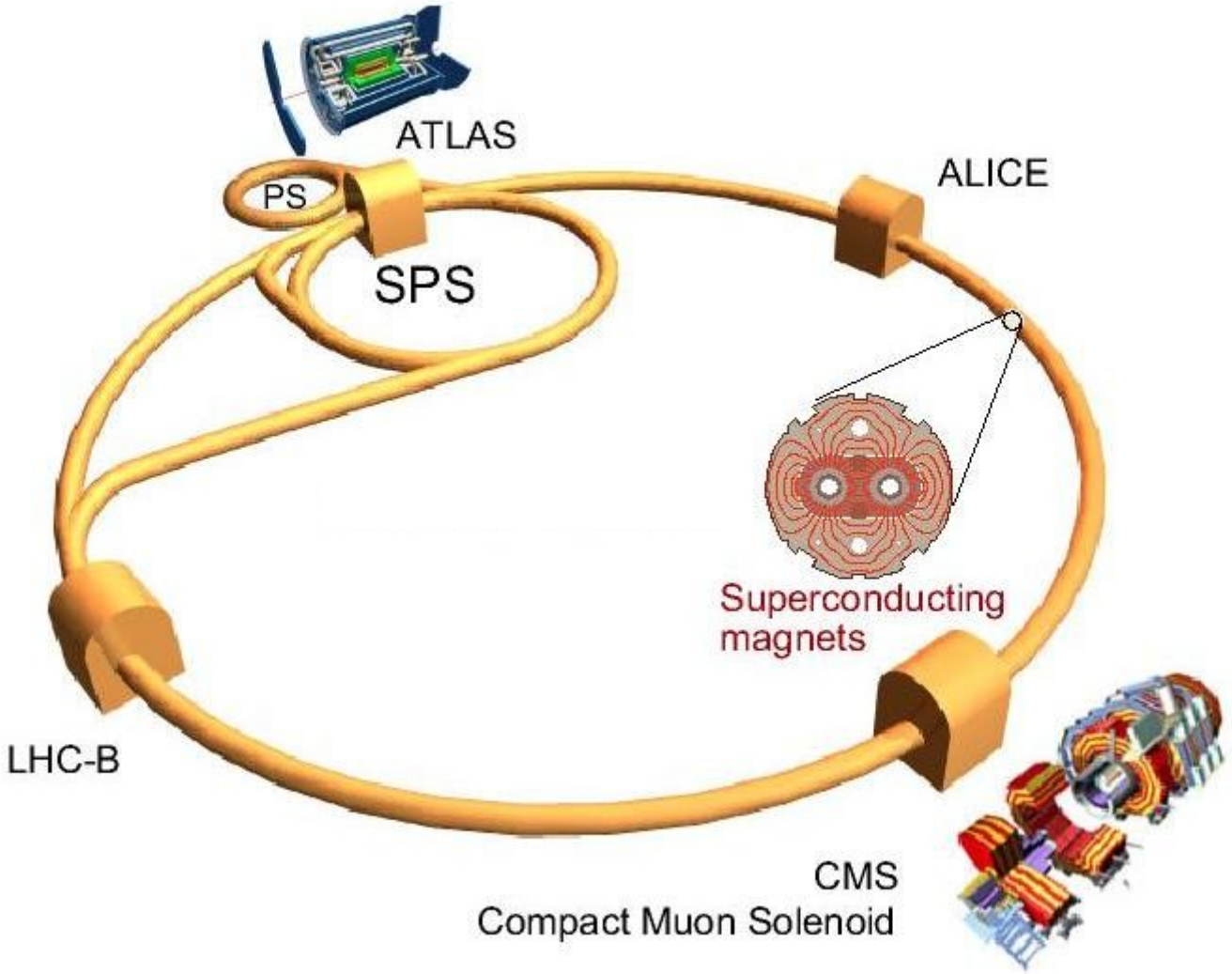


Ever larger datasets



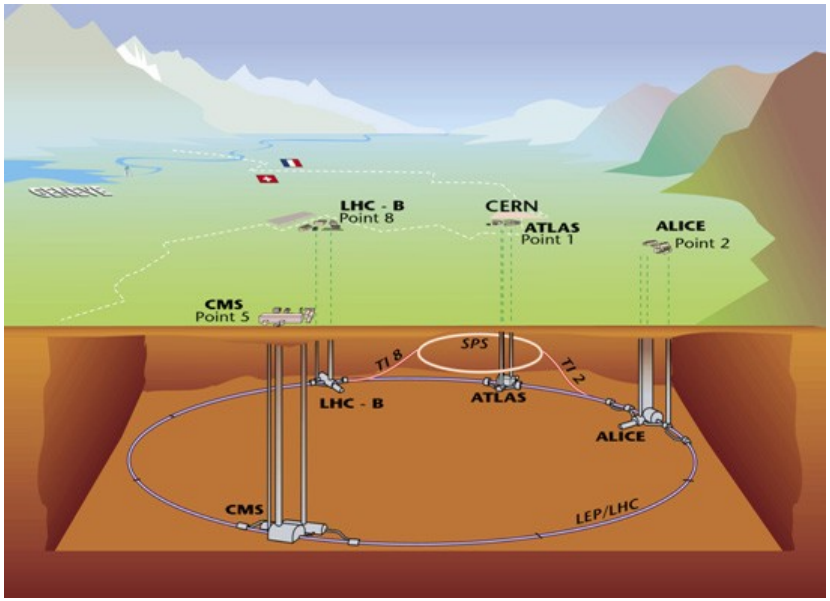
Large Hadron Collider (LHC)

Large Hadron Collider (LHC)

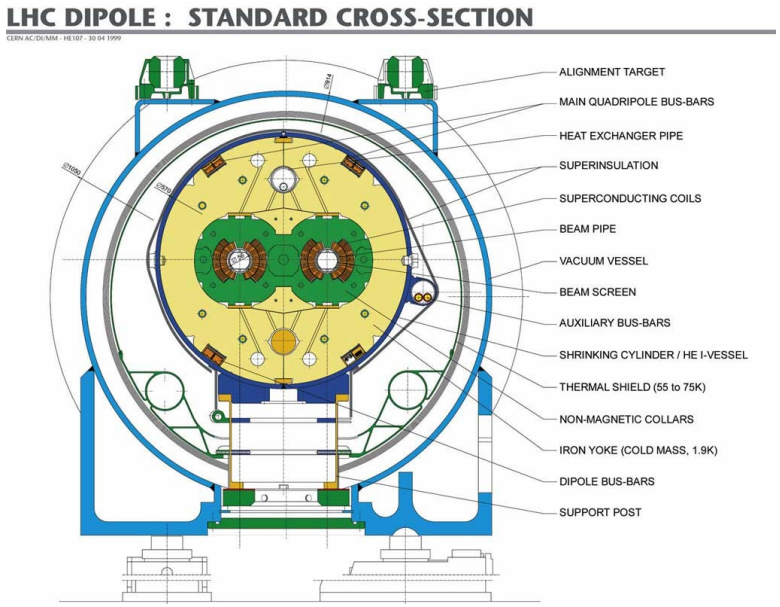


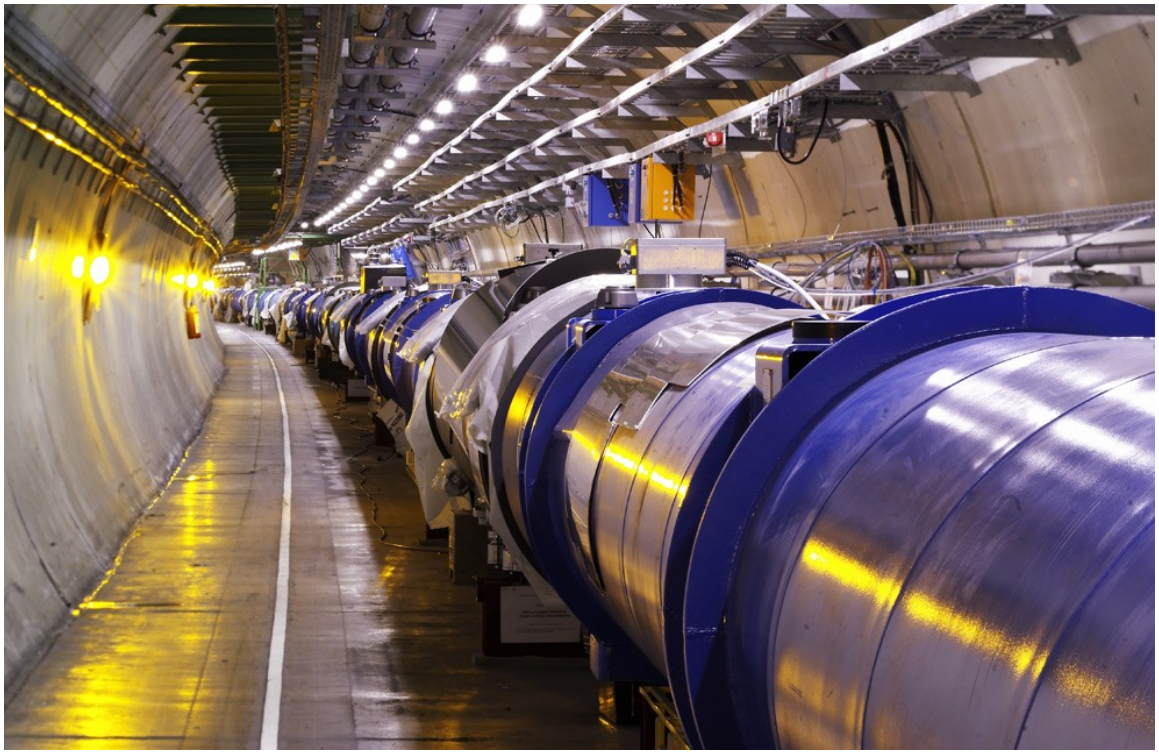
The LHC is a proton-proton collider at CERN in Geneva, Switzerland. It is a completely new accelerator being constructed in a 27km circular tunnel previously used for LEP.

Large Hadron Collider (LHC)



The 27km LHC (formerly LEP) tunnel is located 50-175m underground. Each of the 4 large experiments is located in a cavern at points around the ring. Counter-rotating proton bunches travel in a vacuum beampipe with 25ns (design) spacing, steered by 8.3T (design) dipole magnets.

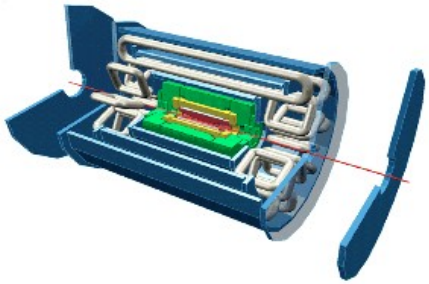
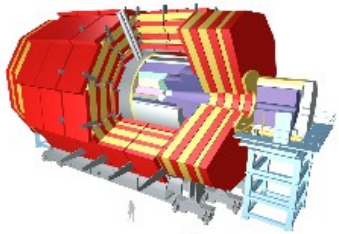
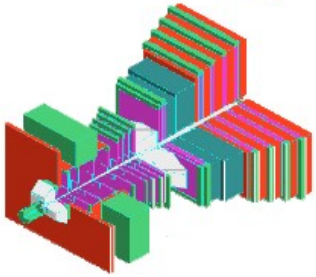
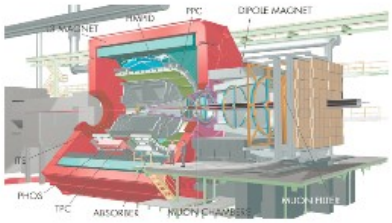




LHC Timeline

- 1994 – LHC approved
- 1998 – Civil engineering begins
- Sep 2008 – first protons circulated in ring, accident
- Nov 2009 – restart and first collisions (450GeV)
- Takes energy lead with respect to the Tevatron at Fermilab (USA)
- 2010-2011 running at 7TeV (half of design) with increasing "luminosity", will continue run through 2012 (or early 2013?)
- Shutdown during 2013-2014 for upgrades toward running at design energy, plans for further shutdowns/upgrades towards "high luminosity" running into the 2020's

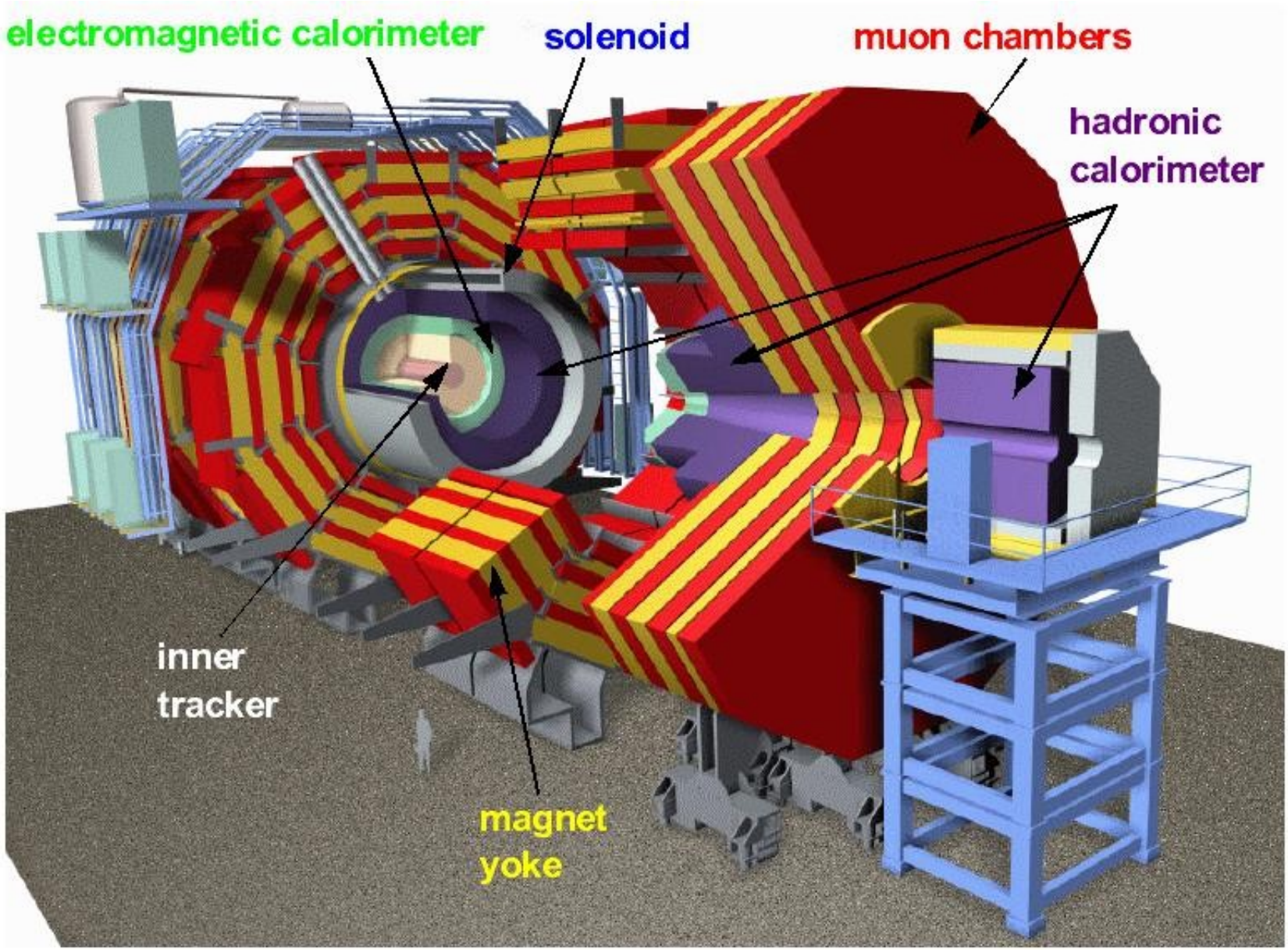
LHC Experiments

	No.Levels Trigger	Lvl 0,1,2 Rate (Hz)	Event Size (Byte)	Evt Build. Bandw.(GB/s)	High Level Trigger HLT Out MB/s (Event/s)
	3	LV-1 10^5 LV-2 $3x10^3$	1.5 MB	4.5	300 (200)
	2	LV-1 10^5	1.0 MB	100	300 (200) Pb-Pb 1500MB/s
	2	LV-0 10^6	30 kB	30	60 (2 kHz)
	4	Pb-Pb 500 p-p 10^3	50 MB 2 MB	25	1250 (100) 200 (100)

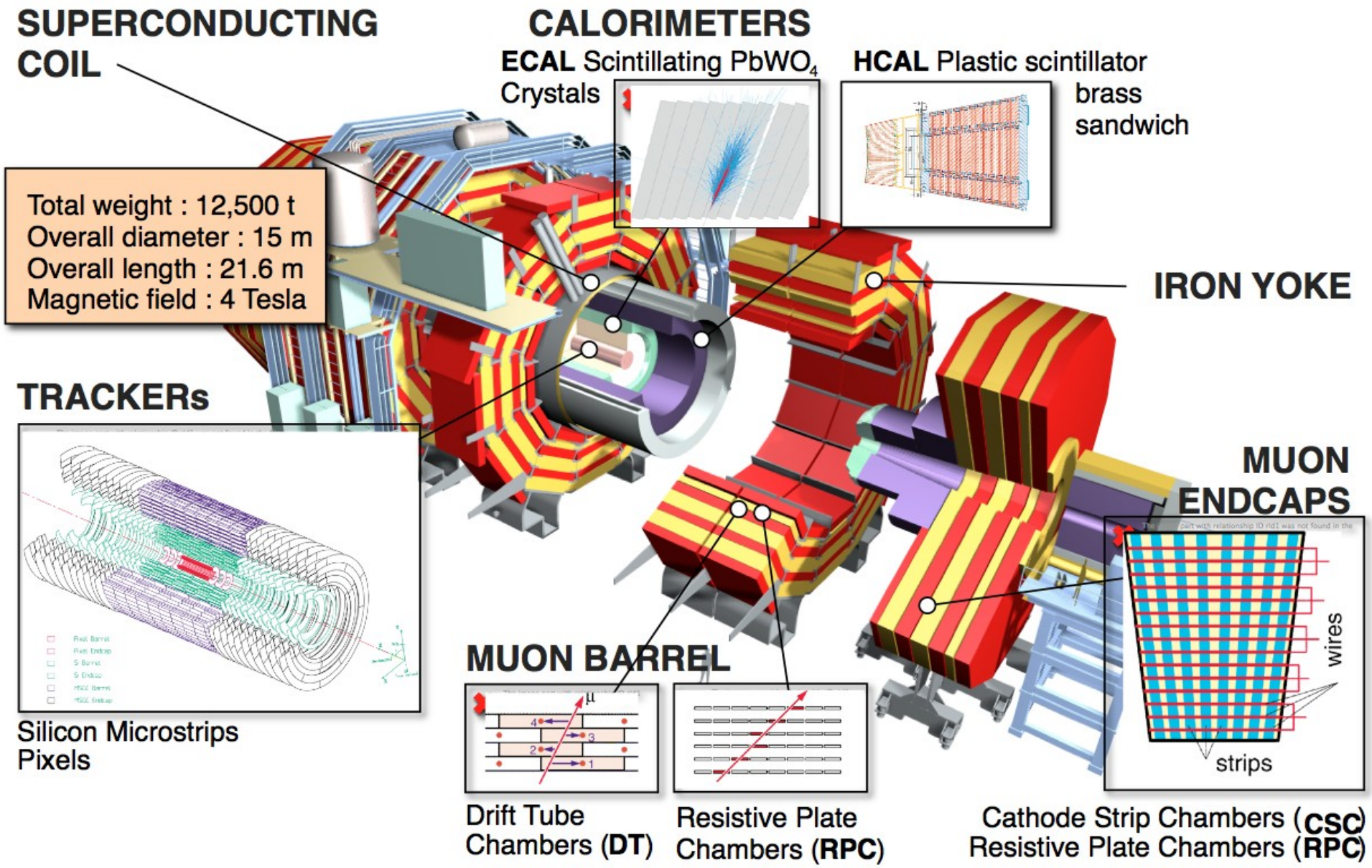
Compact Muon Solenoid (CMS)

- CMS is a general-purpose detector constructed to exploit the physics potential of the LHC collider being constructed at CERN. The physics goals include:
  - Searches for the Higgs particle and new particles
  - Search for deviations from the Standard Model
- ~3800+ collaborators, ~182 institutes, 39 countries
- It is one of 4 large experiments at the Large Hadron Collider (LHC)

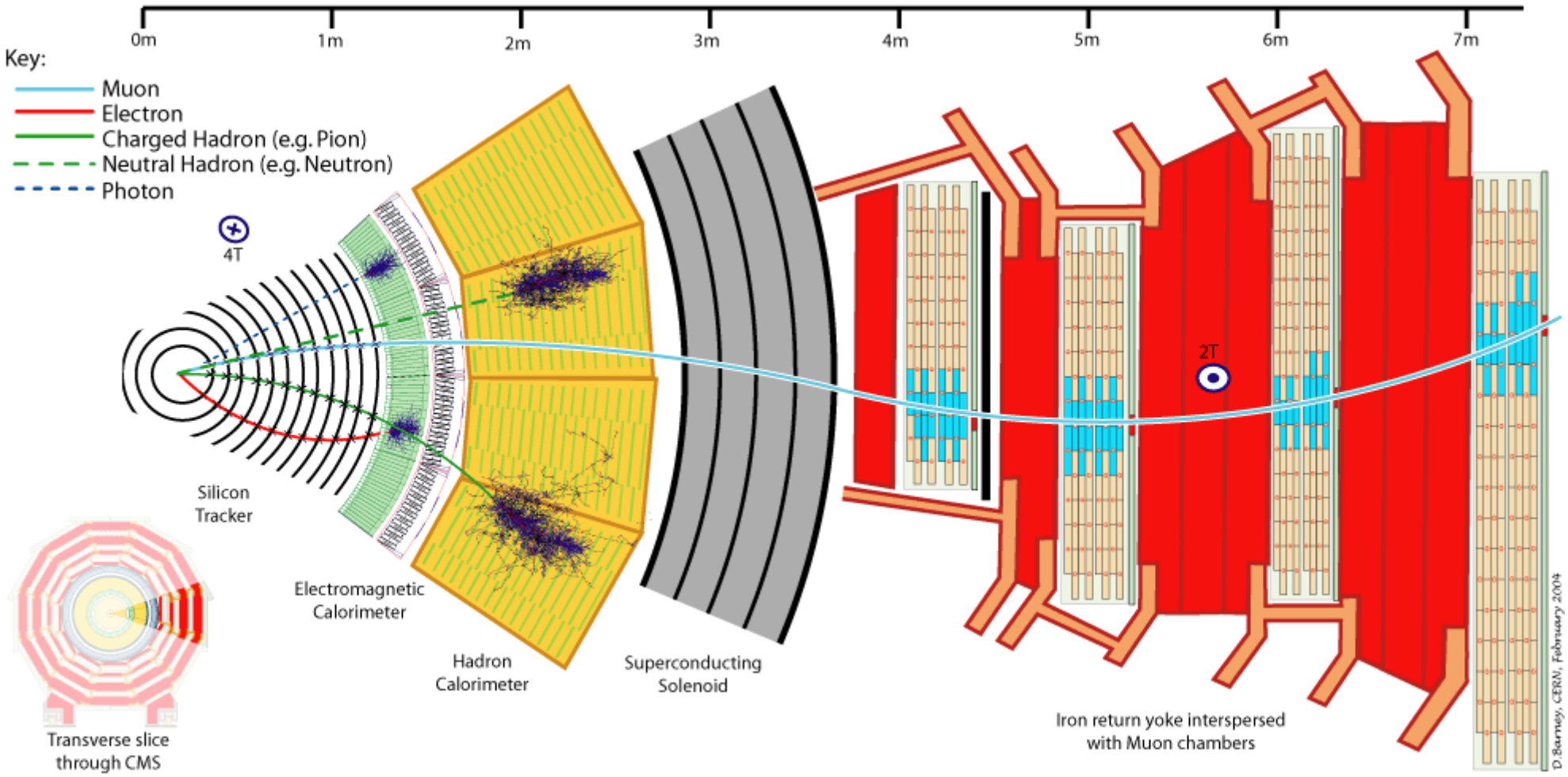
Compact Muon Solenoid (CMS)



Compact Muon Solenoid (CMS)

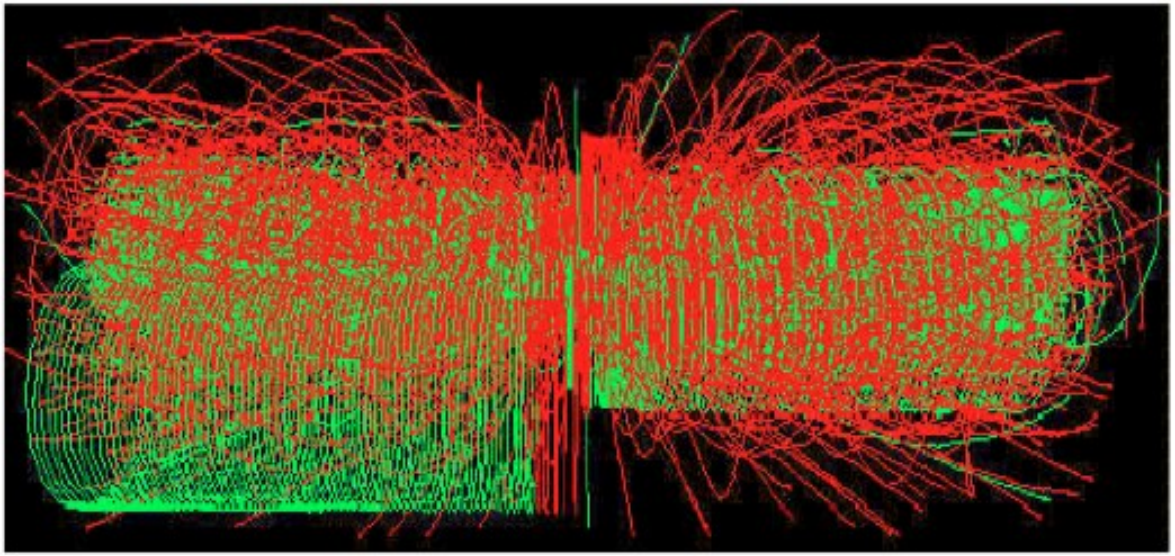


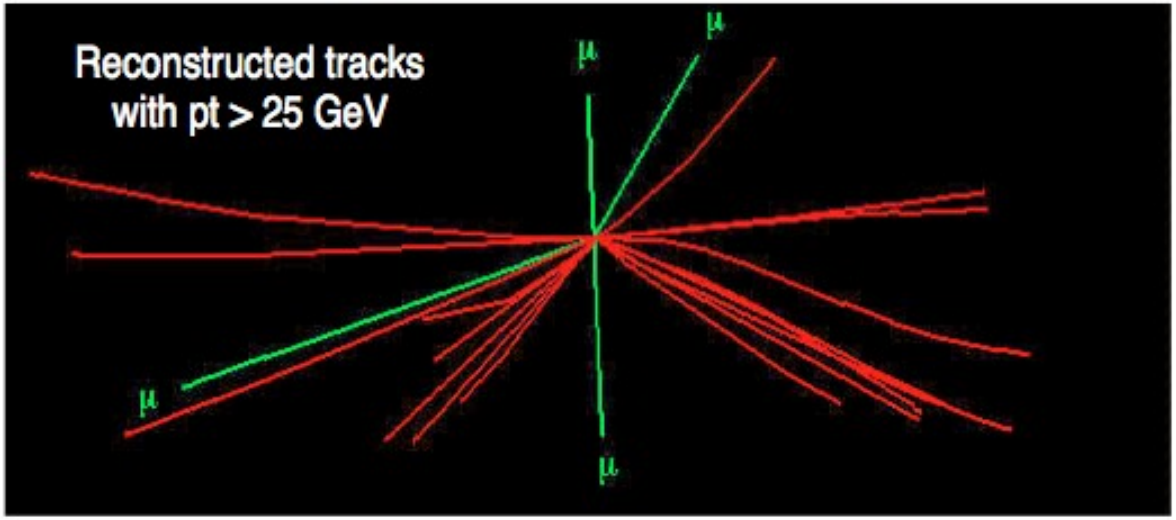
Particle Detection with CMS



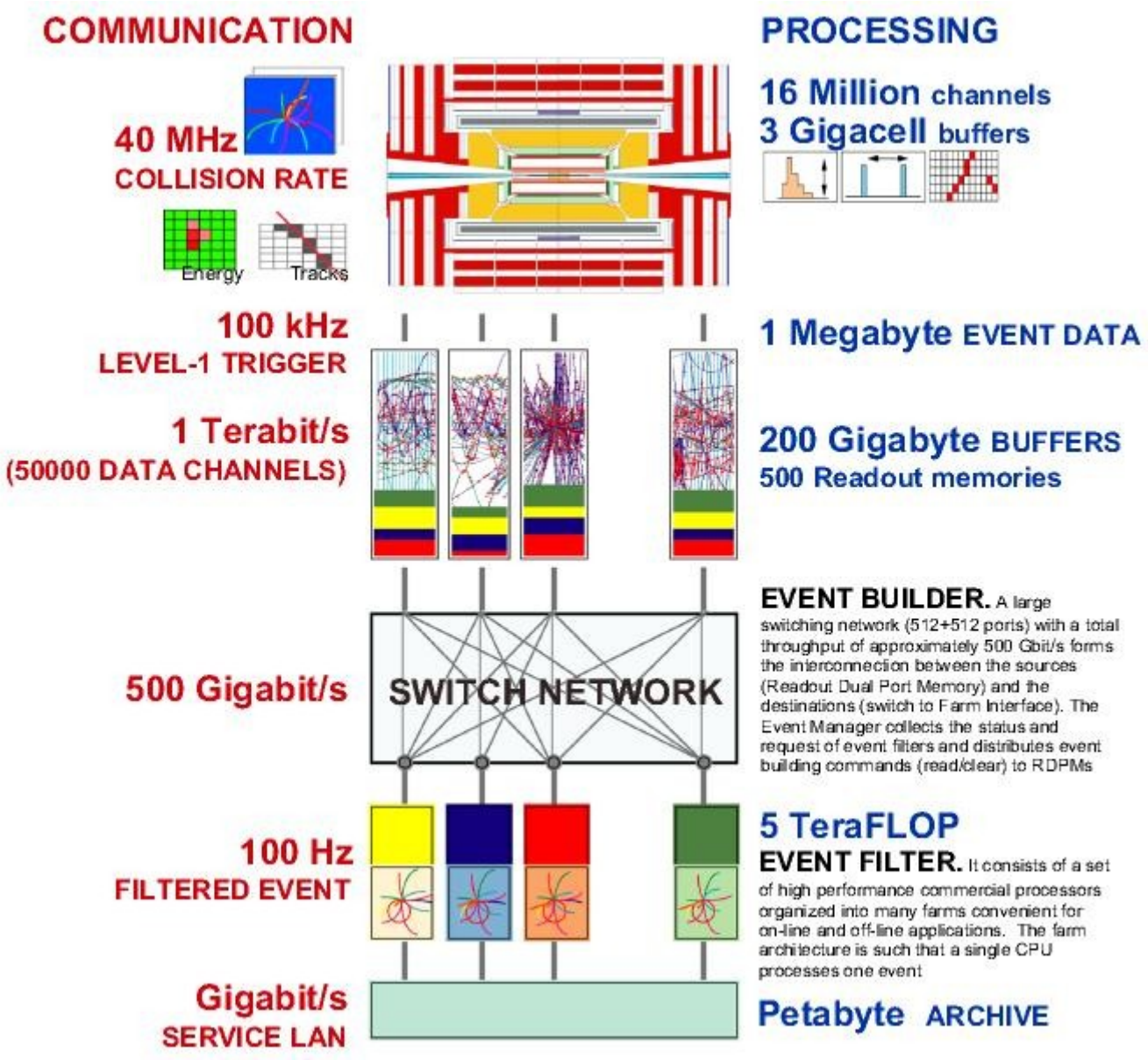
"Events"

$H \rightarrow Z^0Z^0 \rightarrow 4\mu$

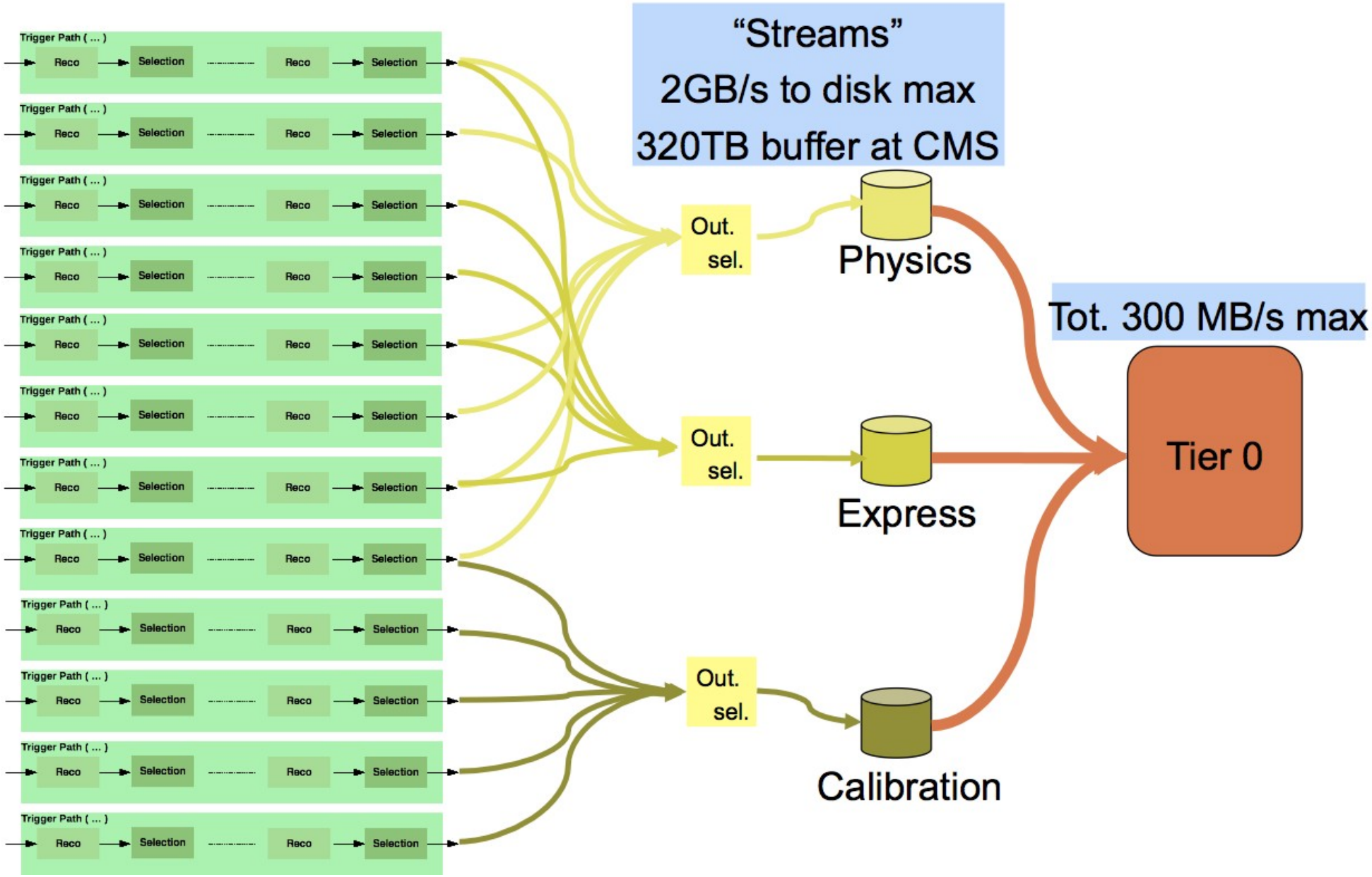




Event "Triggering"



HLT Streams (CMS)



Data formats

The data for each event is written in a simple flat file (not a database), typically a ROOT file and often as a TTree. (See previous talk on ROOT in this session by F.Rademakers.)

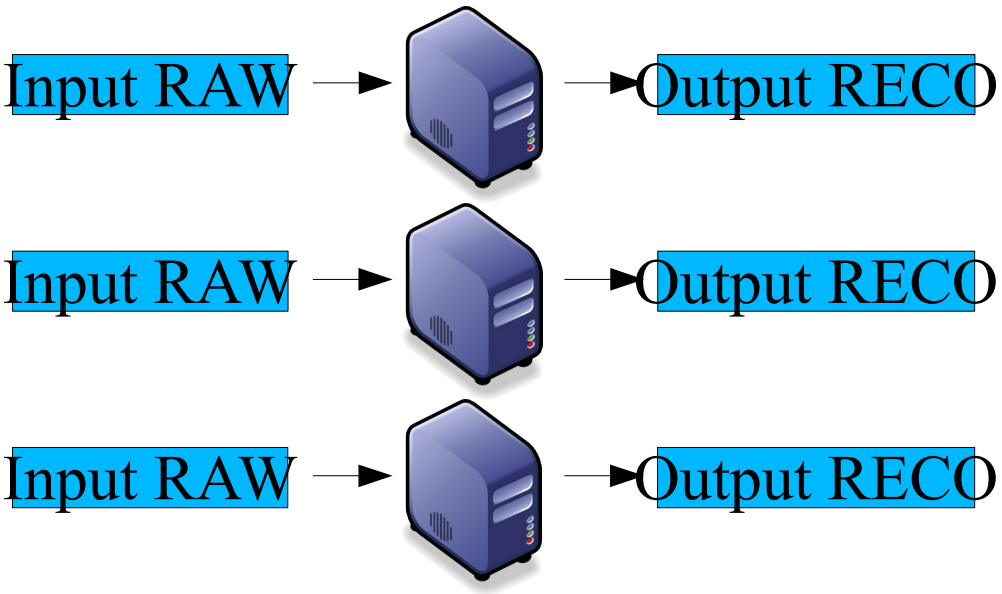
Event data is clustered together by the time it was taken and by the trigger decisions which led to the event selected (like events are clustered together in files). While not strictly required, filesizes of 1-2GB are common to simplify subsequent bookkeeping/access.

Event data (e.g. the RAW data from the detector/trigger and/or subsequent processings of it) constitutes the bulk of the “large datasets” in HEP colliding beam experiments. (i.e. this is the “tens of PB” at the LHC)

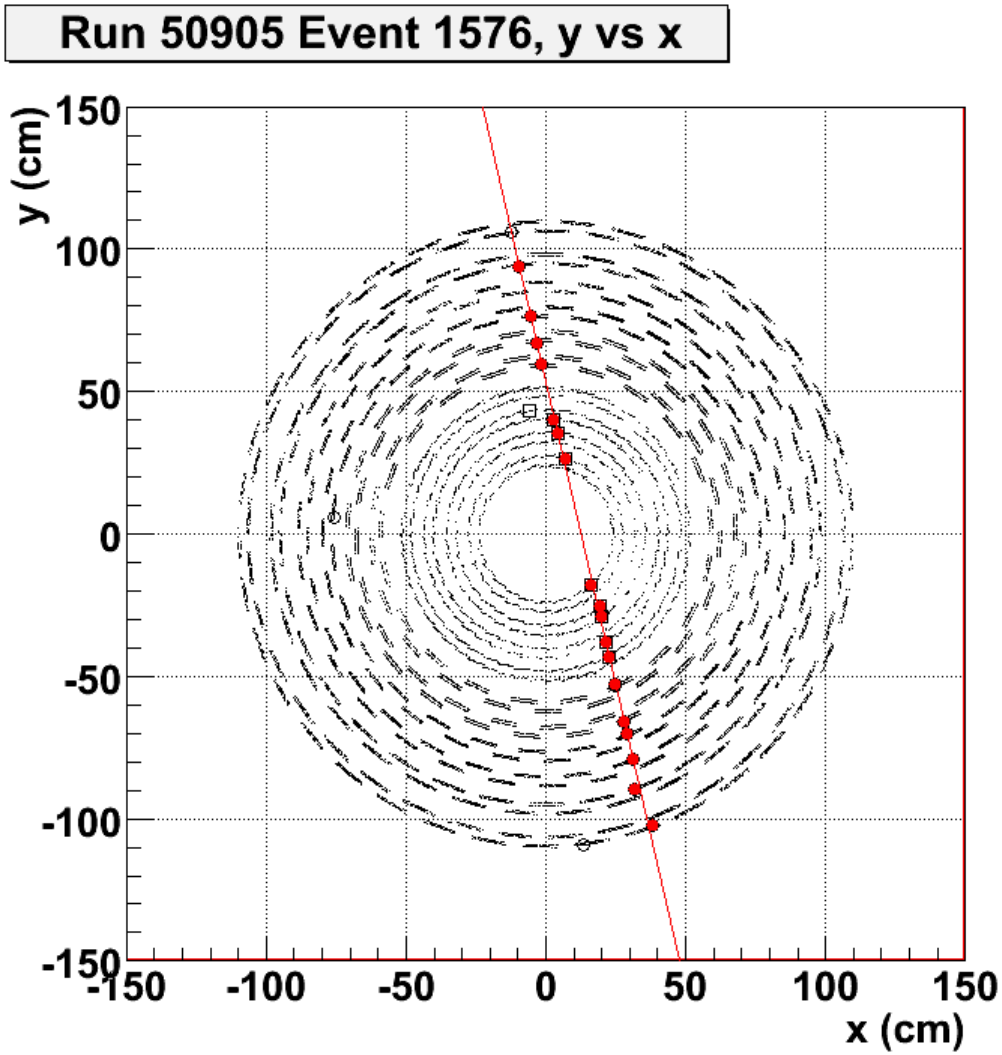
Embarrassingly parallel processing

Each event can be processed independently of all others, where "processed" means applying various pattern recognition algorithms, filtering, etc. We also simplify our problem by never modifying existing files, the problem maps naturally onto independent and parallel processes, each reading events from an input file and writing to an output file.

Our problem is embarrassingly parallel.

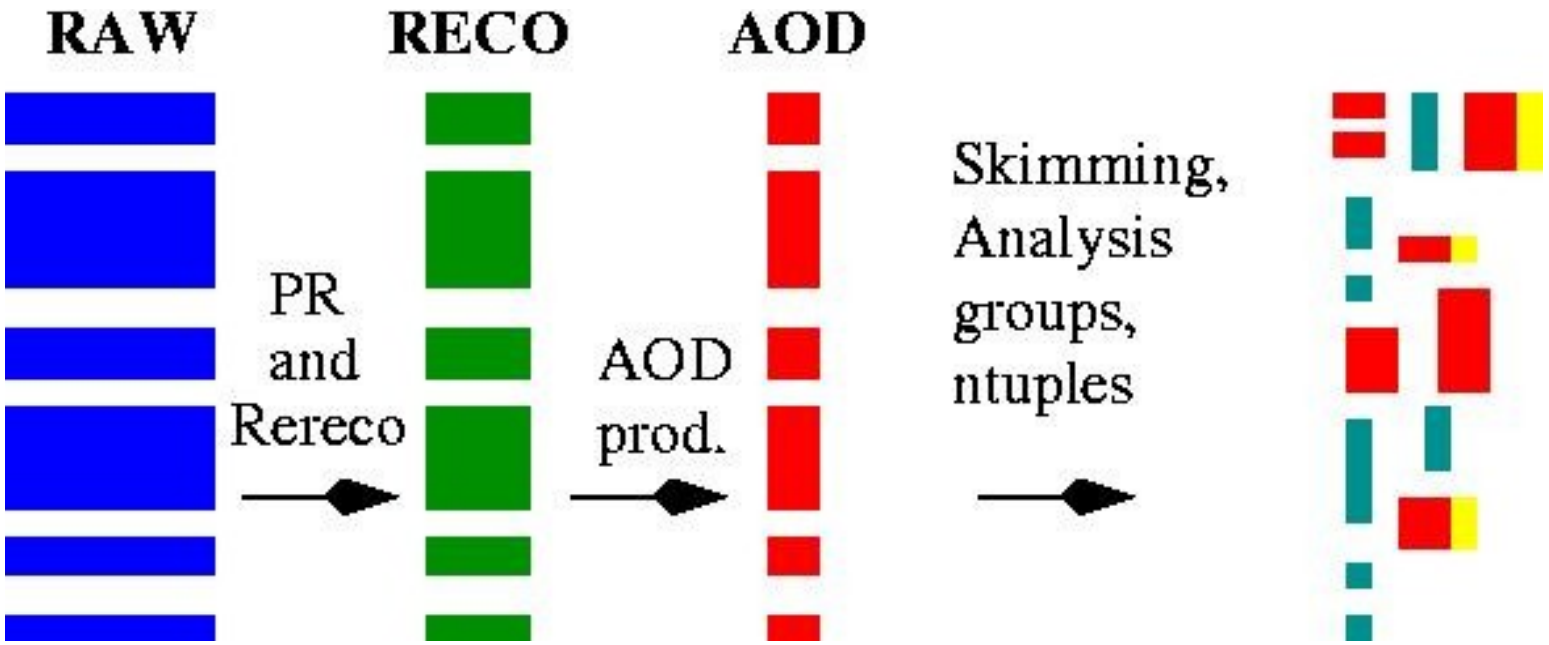


Event Reconstruction



Example cosmic rays in (no-field) tracker, simpler to understand

Data Reduction for analysis



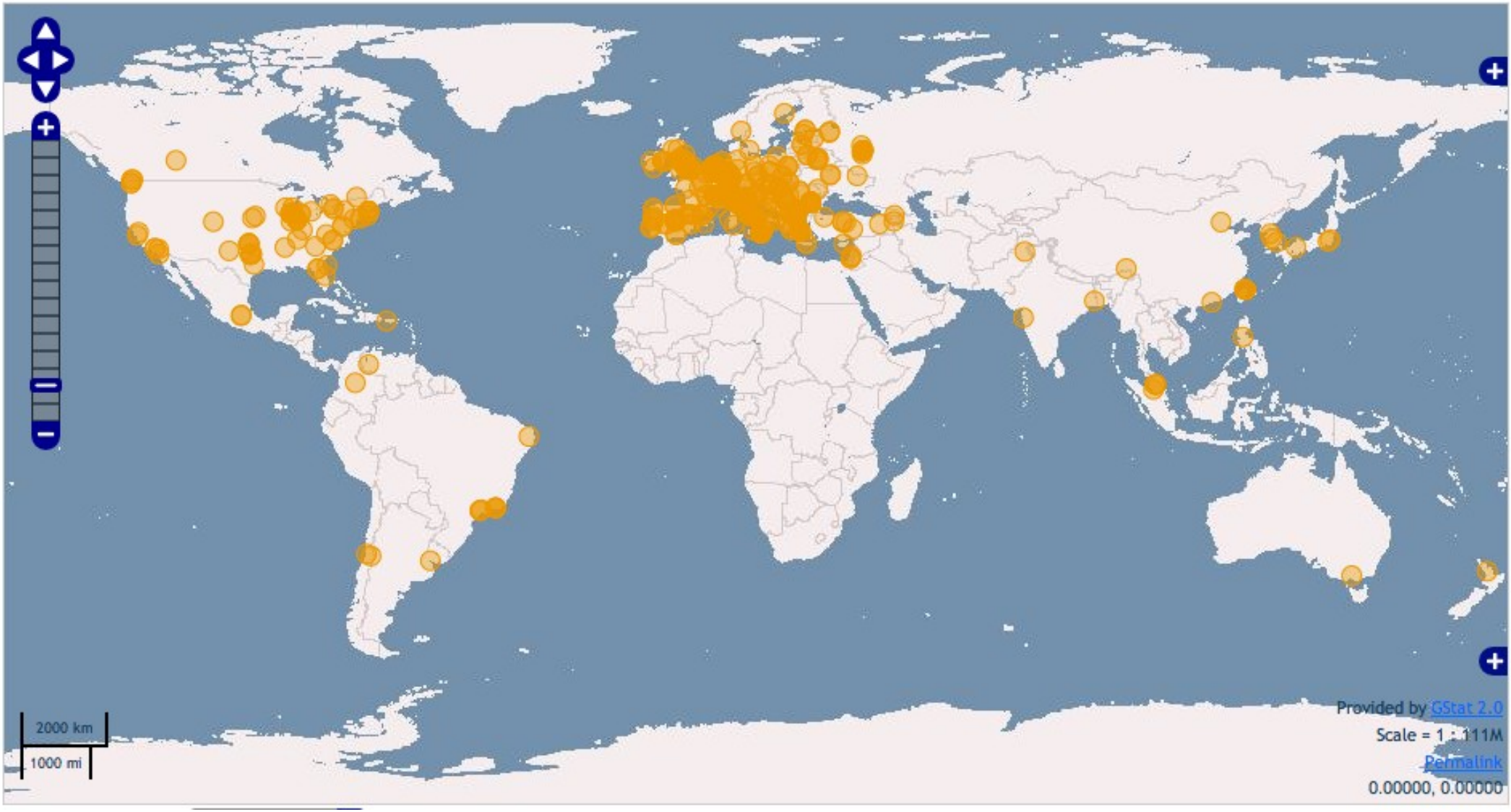
Organized centrally for the entire experiment

Done by analysis groups or Individuals, sometimes centrally

AOD = Analysis Object Data, a smaller subset of the reco data for each event

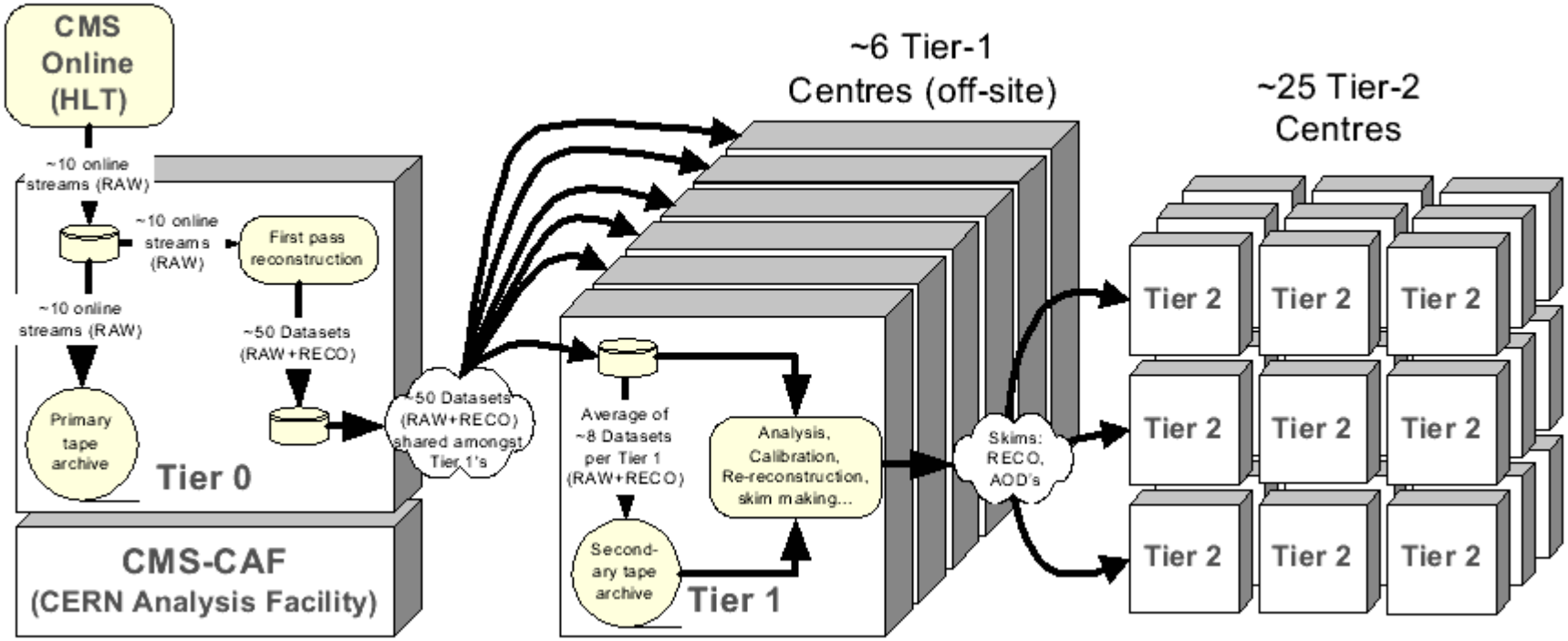
The data reduction chain above entails transforming the data, producing a Smaller (per event summary) and filtering out events not relevant for a given analysis. The goal is to produce a small enough dataset to permit fast iterative analysis (e.g. interactive with ROOT).

Worldwide Computer Centers



Worldwide LHC Computing Grid (WLCG)

Distributed Computing



Tier 1's: FNAL (USA), CNAF (Bologna), FZK (Karlsruhe), CCIn2p3 (Lyon), ASGC (Taiwan), PIC (Barcelona), RAL (UK)

- tape store, responsible for "custodial" data archiving, large scale organized processing/analysis

Tier 2's: often university sites, disk+cpu (no tape)

- responsible for Monte Carlo simulation and unscheduled "user" analysis

Example of computing scale

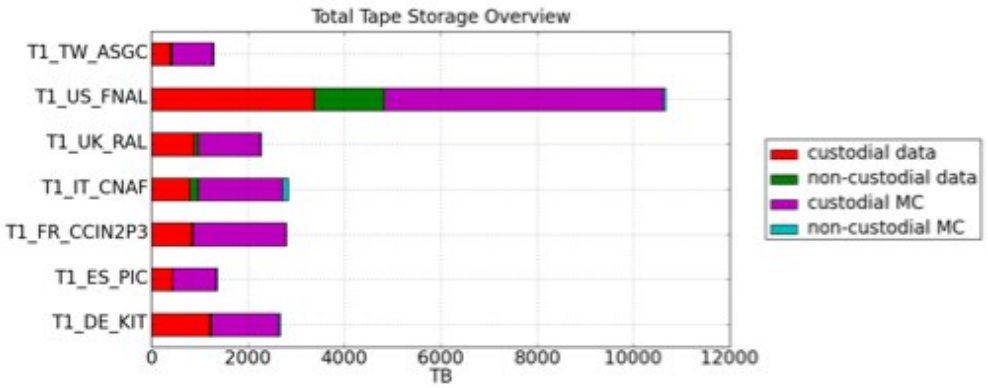
CMS resource request for 2011

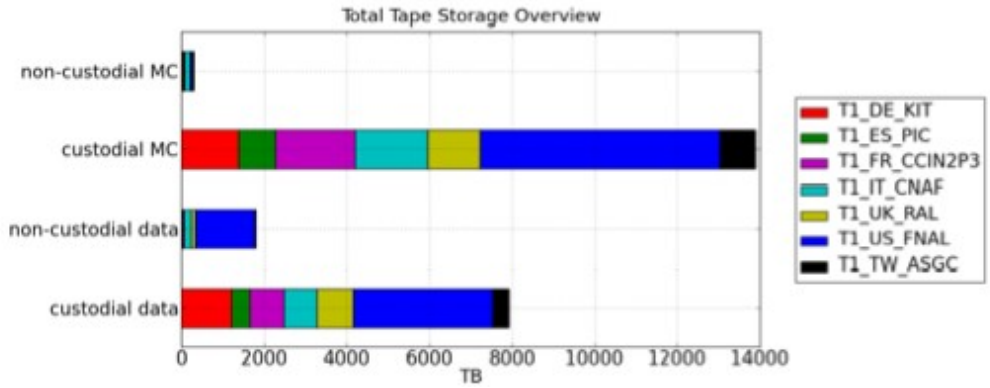
	CERN	Tier1's	Tier2's
Disk (PB)	4.6	16.2	20.2
Tape (PB)	21.6	44.4	-
CPU (kHS06)	109	130	315

Tape storage

Tape Storage Overview

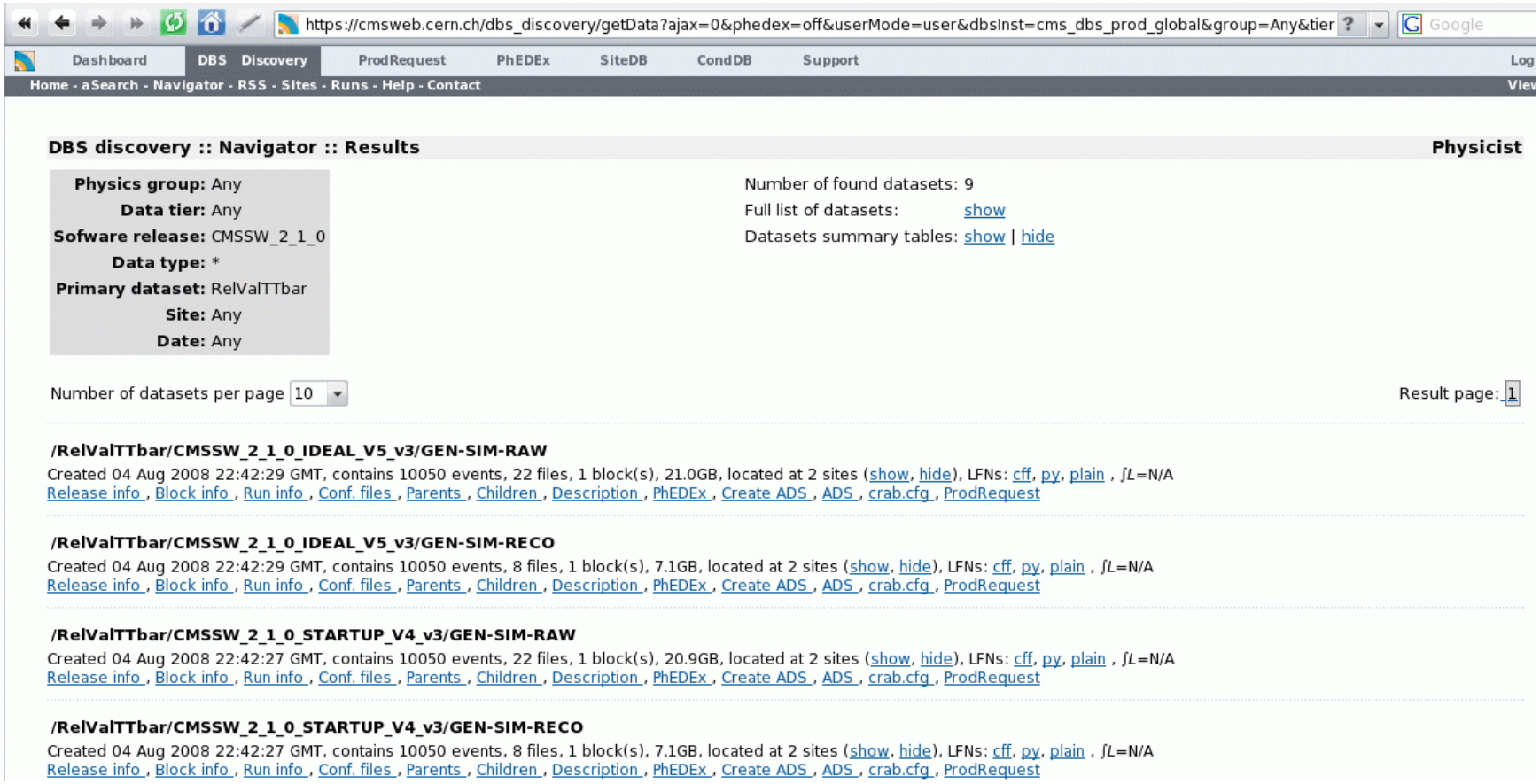
Current week Overview





~24PB of data stored in the Tier-1 computer centers, with additional replicas in the (disk-only, non-custodial) Tier-2's.

Dataset Bookkeeping System



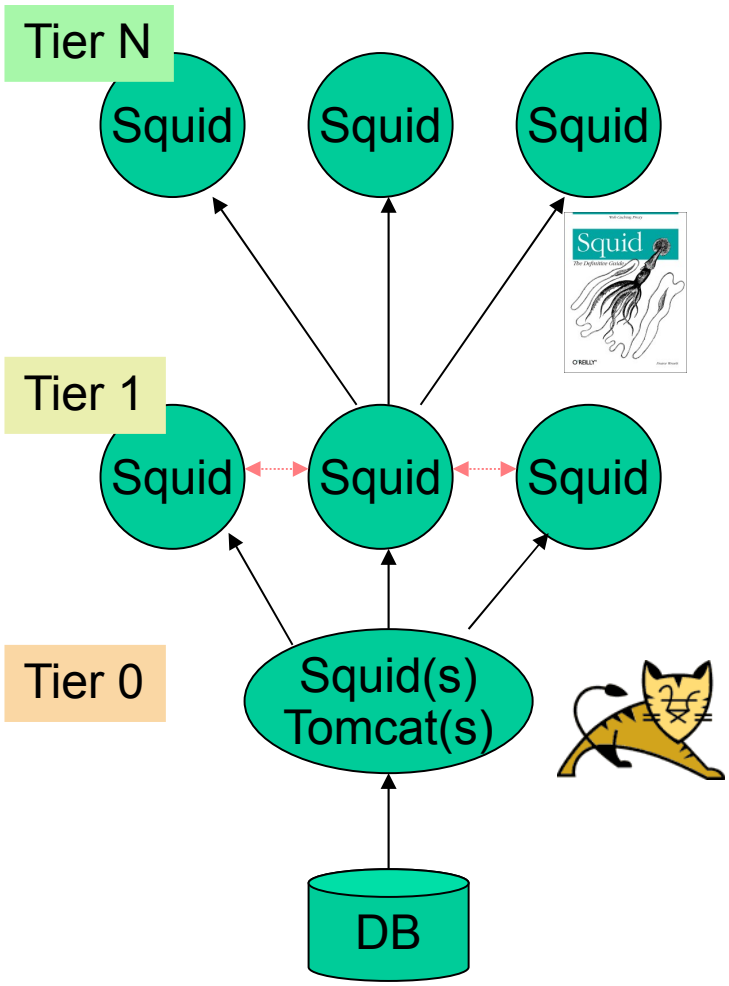
Sets of event data files are organized into “datasets” and a description of the dataset is stored in Oracle. The descriptions are accessible via a web interface and also via our workflow tools. Currently this is about 420GB of data.

Access to conditions data

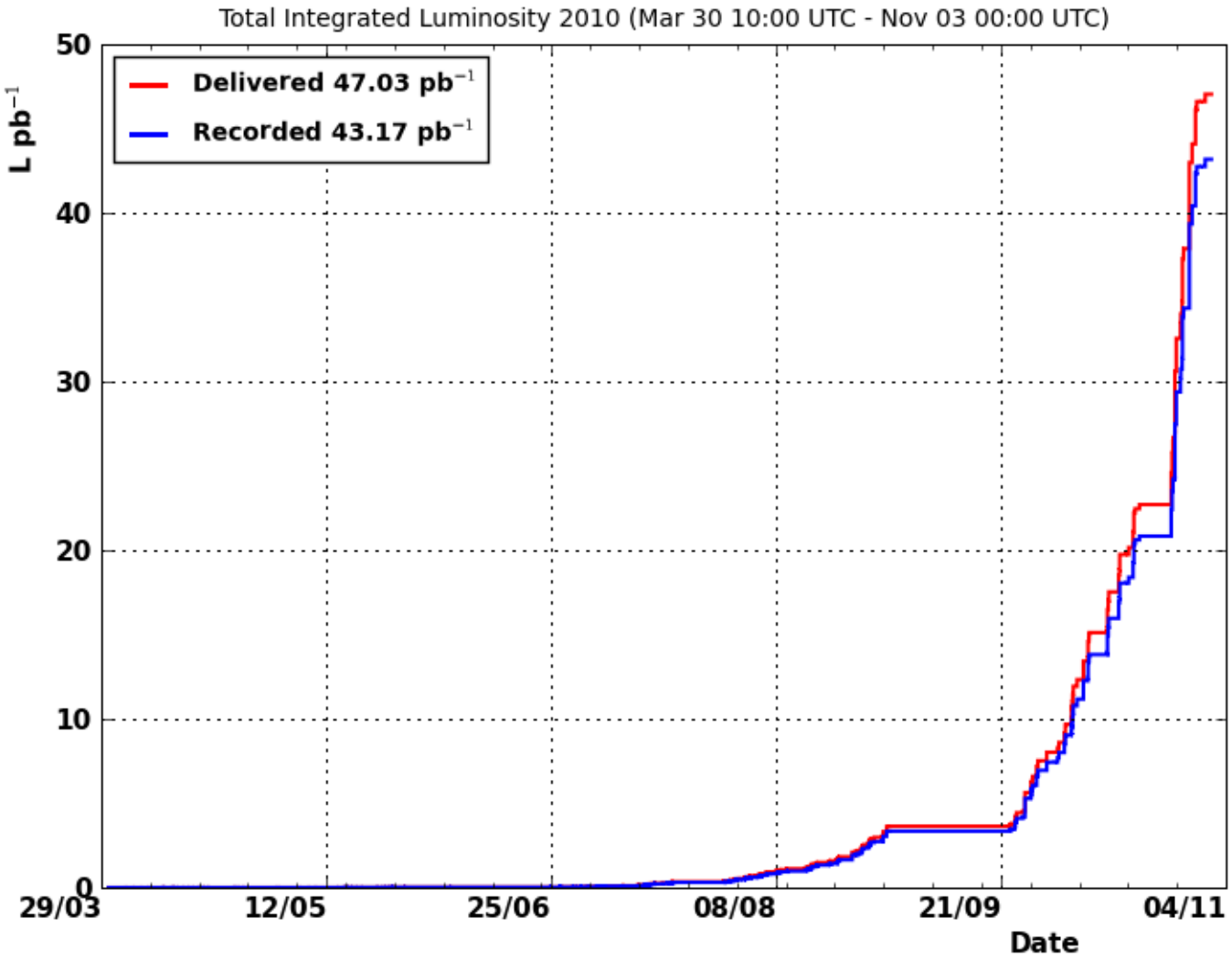
In addition to the PB's of "Event" data we do have a smaller amount of "calibrations" data which is needed at times during the processing of events. Each individual instance of our applications will need to read one or more calibrations.

This data is stored in an Oracle DB at CERN, where it is also created. For all other sites this data is read-only. For scalability we do not access directly the Oracle DB from individual batch nodes, but instead use a caching layer based on Squid/Tomcat.

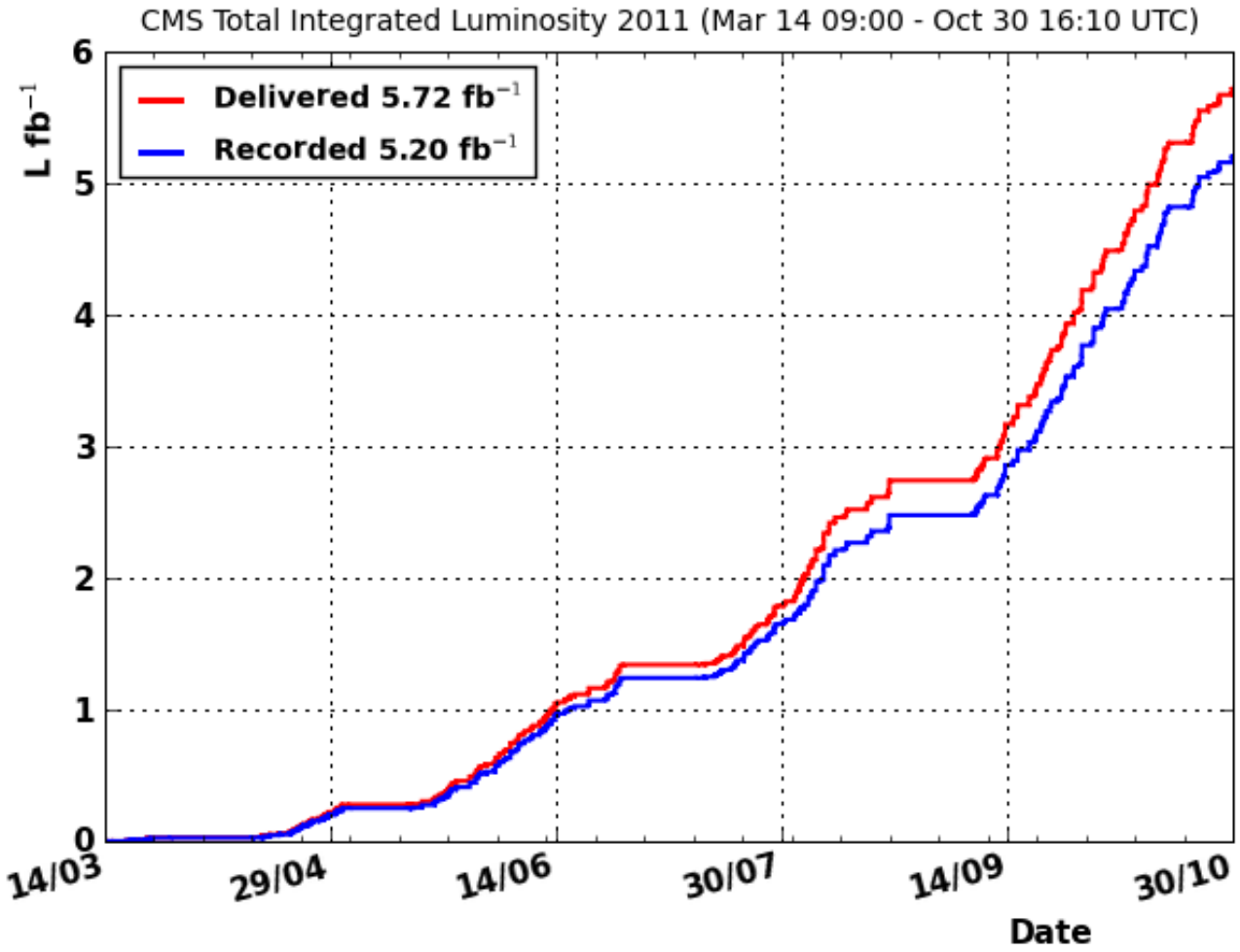
About 3.8TB of calibrations/conditions have been created in the CMS online and 400GB in the CMS offline.



Data accumulated in 2010



Data accumulated in 2011



Note 1k change in scale of y-axis relative to the previous 2010 plot.

Summary

- High Energy Physics experiments typically produce very large datasets in pursuit of rare phenomena
- The experiments at the Large Hadron Collider (LHC) at CERN in Geneva are the latest generation to push the envelope
- Using one of the LHC experiments as an example, I have described the typical methods used to produce highly scalable, embarrassingly parallel and geographically distributed access to these large datasets.