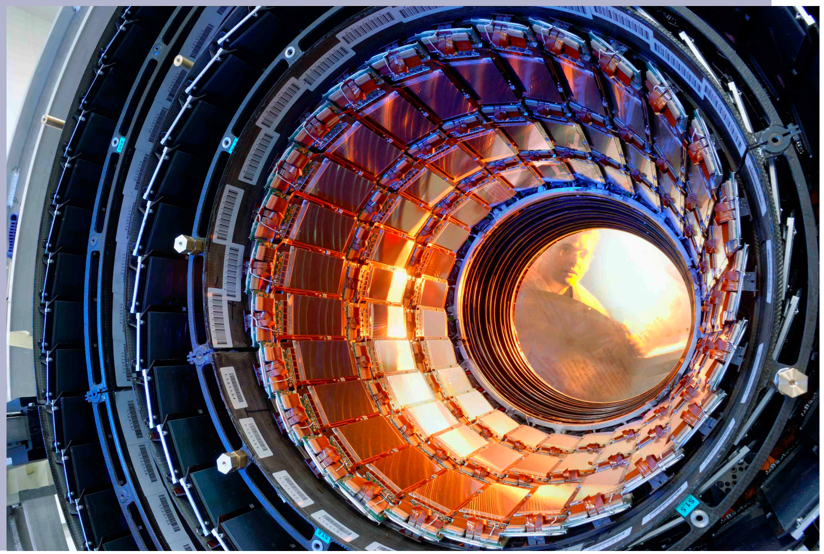
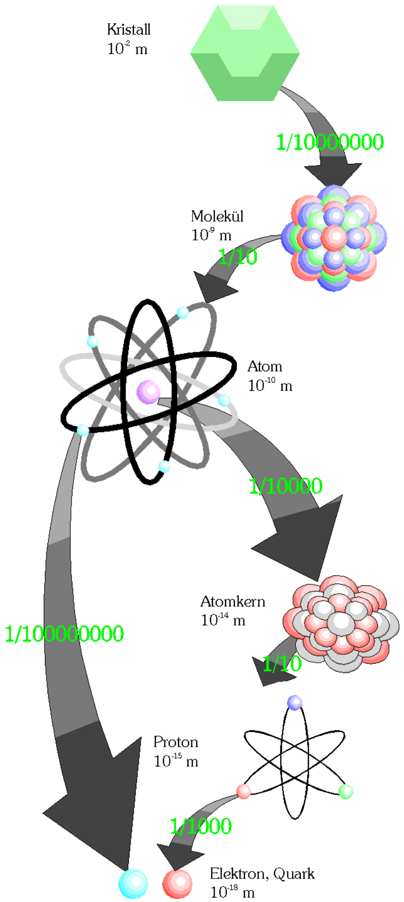
Data taking and analyzing at unprecedented scale

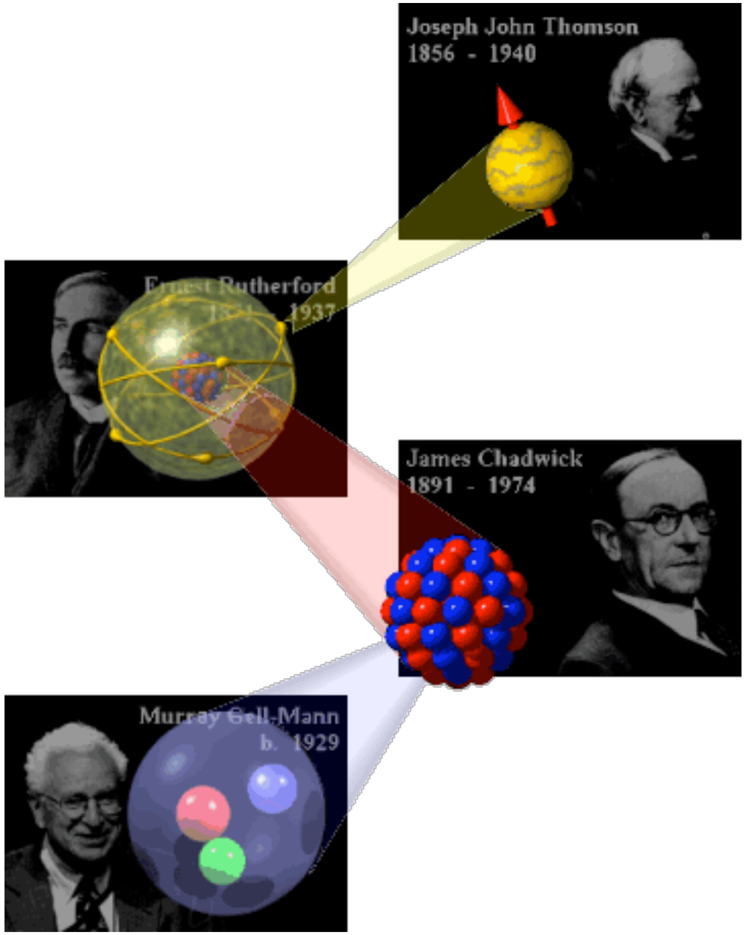
The example of CMS



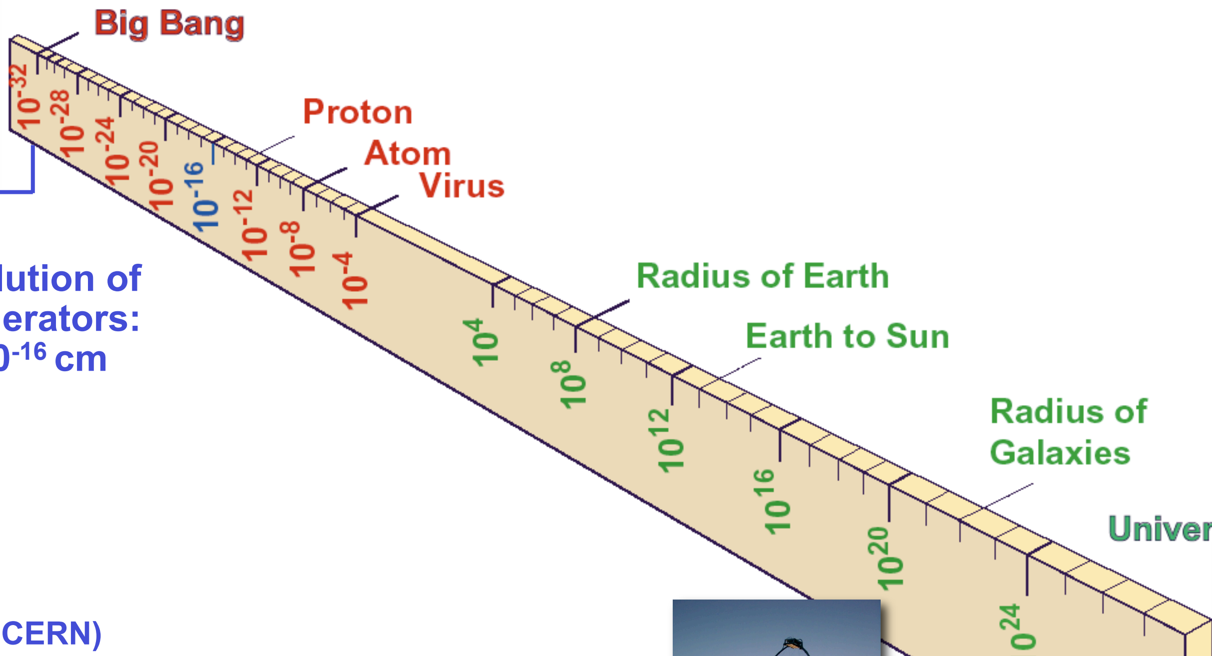
Dr. Marie-Christine Sawley
IPP-ETH Zurich
CERN Group

Research fundamentals



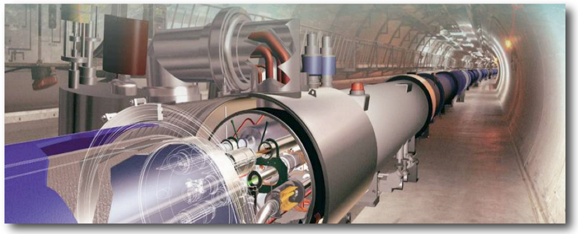


Working with different dimensions

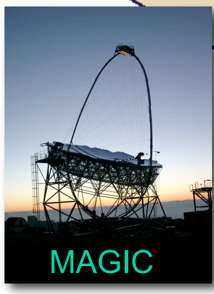


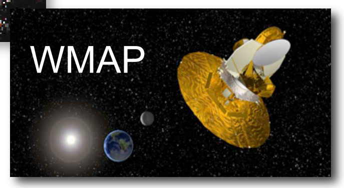
Resolution of Accelerators:
$< 10^{-16}$ cm

LEP (CERN)
1989-2000



LHC (CERN)
2008





CERN - "European Organization for Nuclear Research"

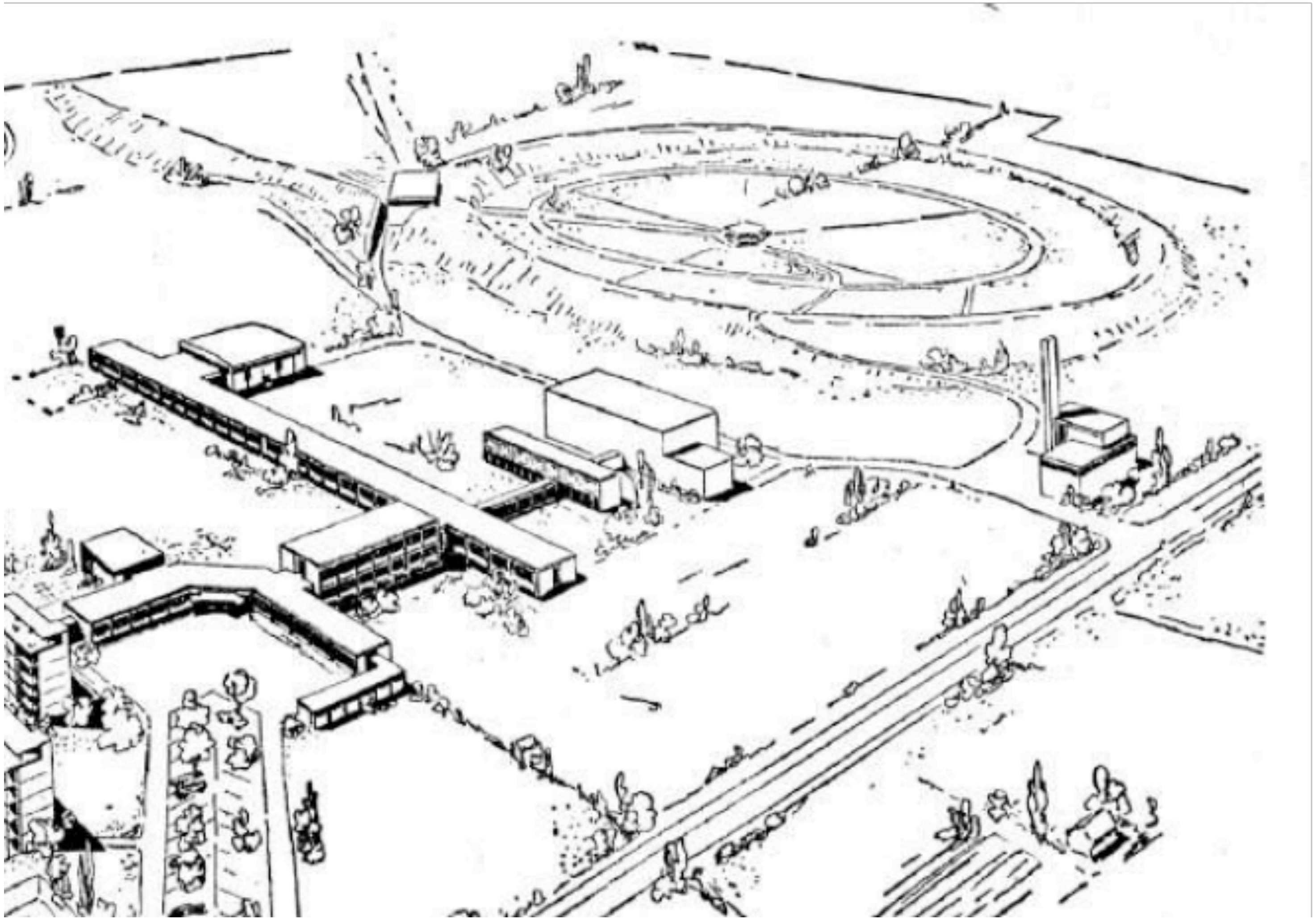
World largest lab in particle physics

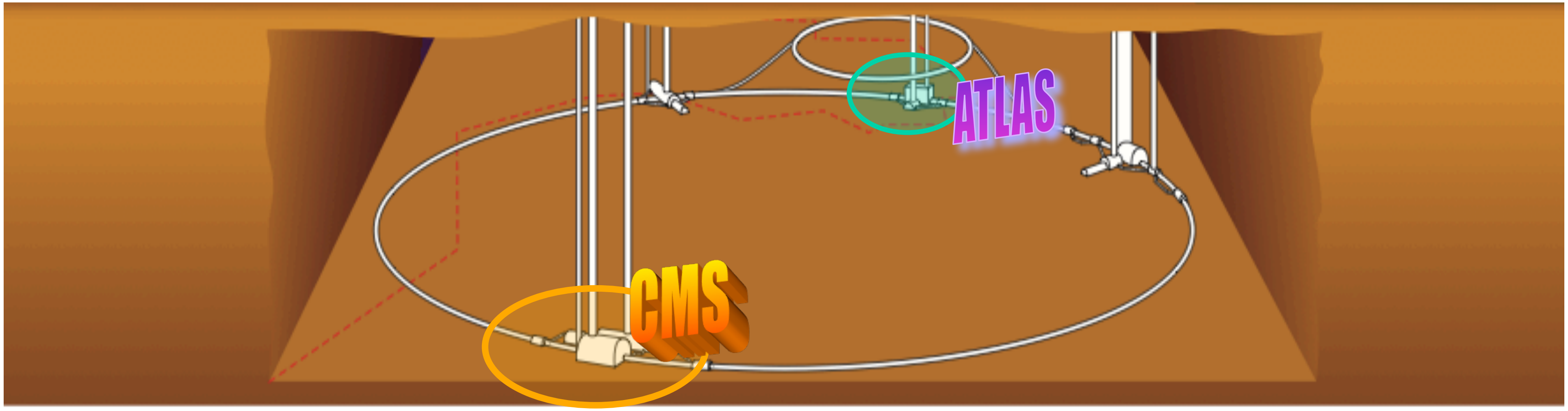
8000 scientists,
580 institutions,85 nationalities

20 members states

The largest accelerator ever built

Going back in time: "Journal de Geneve", 8 July 1954





Main experiments

ATLAS

2500+ members from 150 institutions in 37 countries. The ATLAS cavern could hold the nave of Notre Dame Cathedral.

CMS

3000+ members from 180 institutions in 39 countries. The CMS magnet is the largest solenoid ever built and contains almost twice as much iron as the Eiffel tower.

ALICE

1000+ members from 105 institutions in 30 countries. The ALICE Time Projection Chamber, a cylinder roughly 15 feet in diameter and 15 feet in length, has approximately 560,000 readout channels.

LHCb

650+ members from 48 institutions in 15 countries. The LHCb experiment searches for CP-violation, the asymmetry in the behavior of matter and antimatter.

Number of magnets

1,232 superconducting dipole magnets steer the beam around the ring. Each one is roughly 47 feet long and weighs around 35 tons.

Magnetic field

8.33 Tesla, or about 200,000 times the strength of the Earth's magnetic field

Super cold

The LHC will operate at 1.9 Kelvin, about 300 degrees Celsius below room temperature.

Superconducting

The total length of the superconducting wire for the LHC is roughly 155,000 miles, enough to go 6.8 times around the equator.

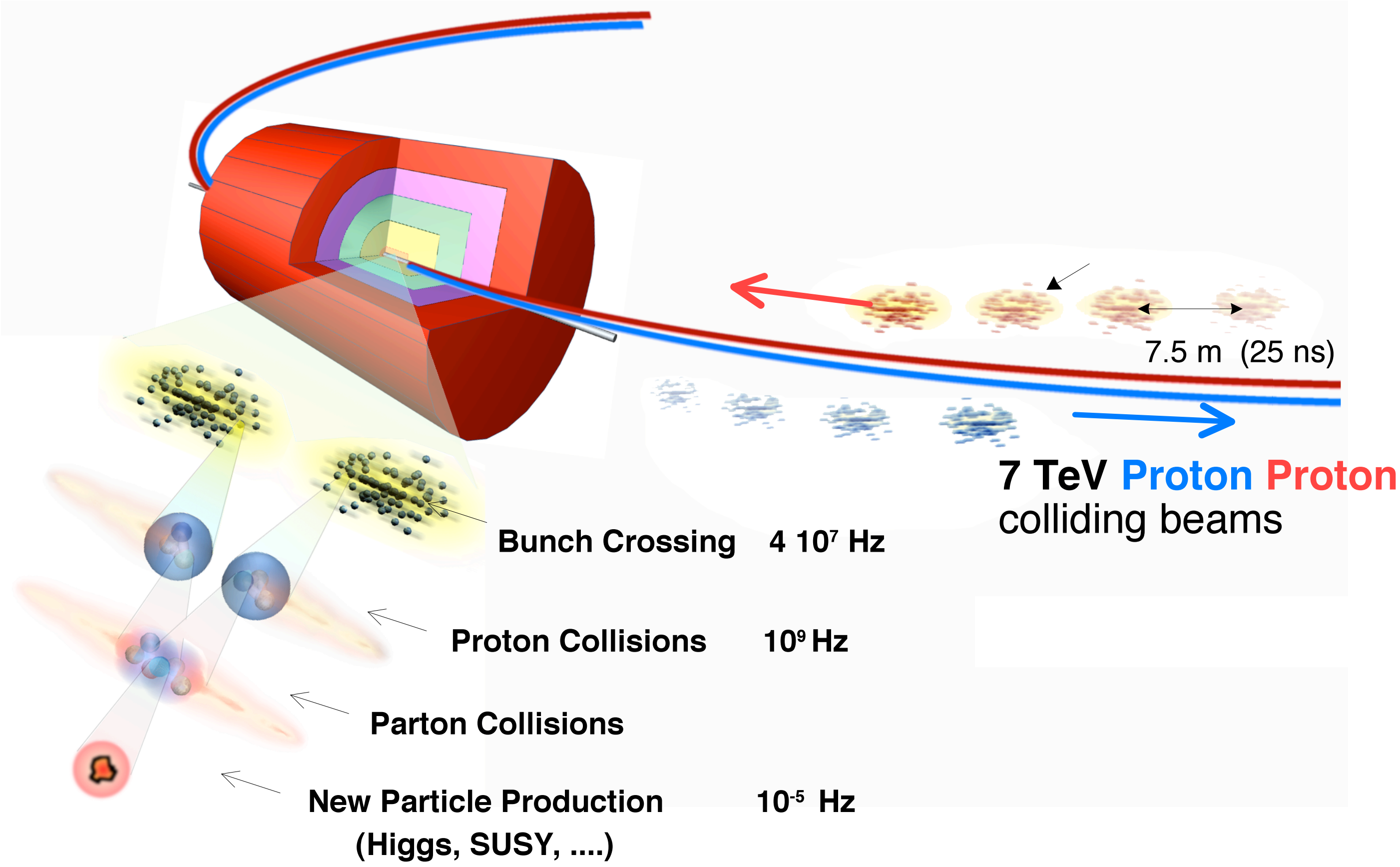
Super fast

The particles in the LHC will travel near the speed of light. Protons will travel around the 17-mile ring 11,000 times per second, colliding up to one billion times a second.

Super computing

LHC experiments will produce 15 petabytes–15 million gigabytes–of data every year, which has to be stored and made available to more than 7,000 scientists around the globe.

Collisions in the LHC and data recording



Selection of 1 event in 10,000,000,000,000

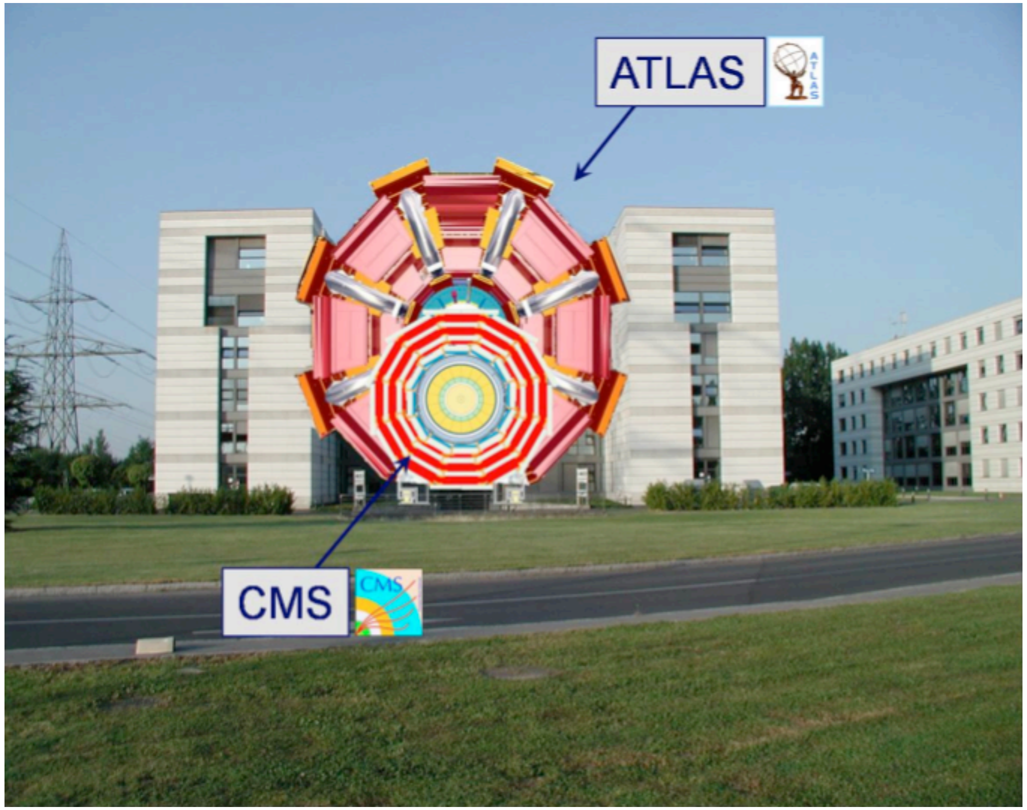
Differences between ATLAS and CMS?

CMS

12500 Tons

15 m high

21 m long



ATLAS

7000 Tons

25 m high

46 m long

Similar physics endeavour
Different technologies, methodologies, scientific teams and funding

Compact Muon Solenoid

a

Superconducting magnet

A cylindrical, superconducting magnet, about 40 feet long, 20 feet wide and weighing 220 tons that contains many of the CMS subsystems. This compact design led to the detector's name. Scientists need the magnet to bend the paths of charged particles, providing information on each particle's charge, mass, and speed.

b

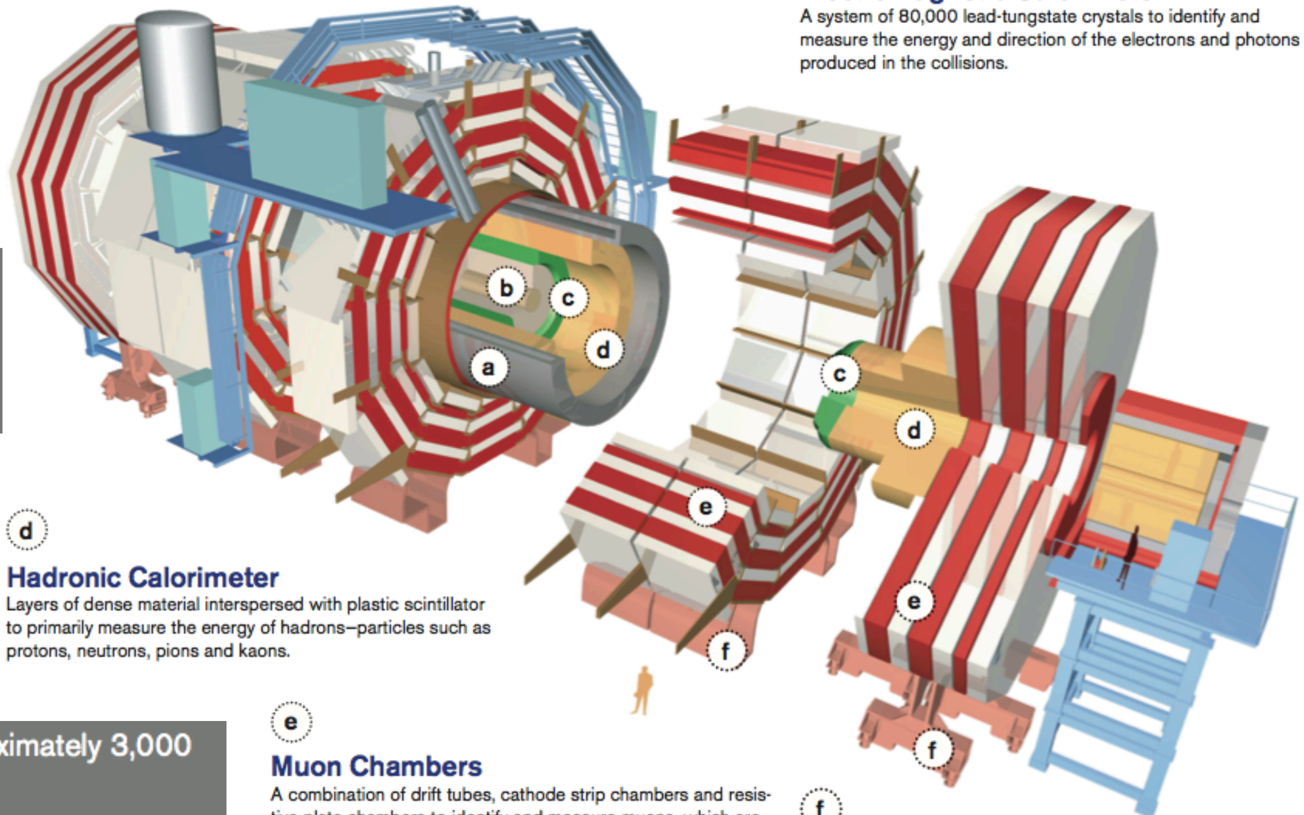
Tracker

The CMS tracker consists of 10 million silicon strips, 66 million silicon pixels and specialized electronics that can determine the exact coordinates of a particle track to within the width of a human hair.

c

Electromagnetic Calorimeter

A system of 80,000 lead-tungstate crystals to identify and measure the energy and direction of the electrons and photons produced in the collisions.



Total weight: **12,500 tons**

Overall diameter: **52 feet**

Overall length: **70 feet**

Number of detection elements: **100 million**

d

Hadronic Calorimeter

Layers of dense material interspersed with plastic scintillator to primarily measure the energy of hadrons–particles such as protons, neutrons, pions and kaons.

e

Muon Chambers

A combination of drift tubes, cathode strip chambers and resistive plate chambers to identify and measure muons, which are essentially heavier cousins of electrons.

f

Foundation

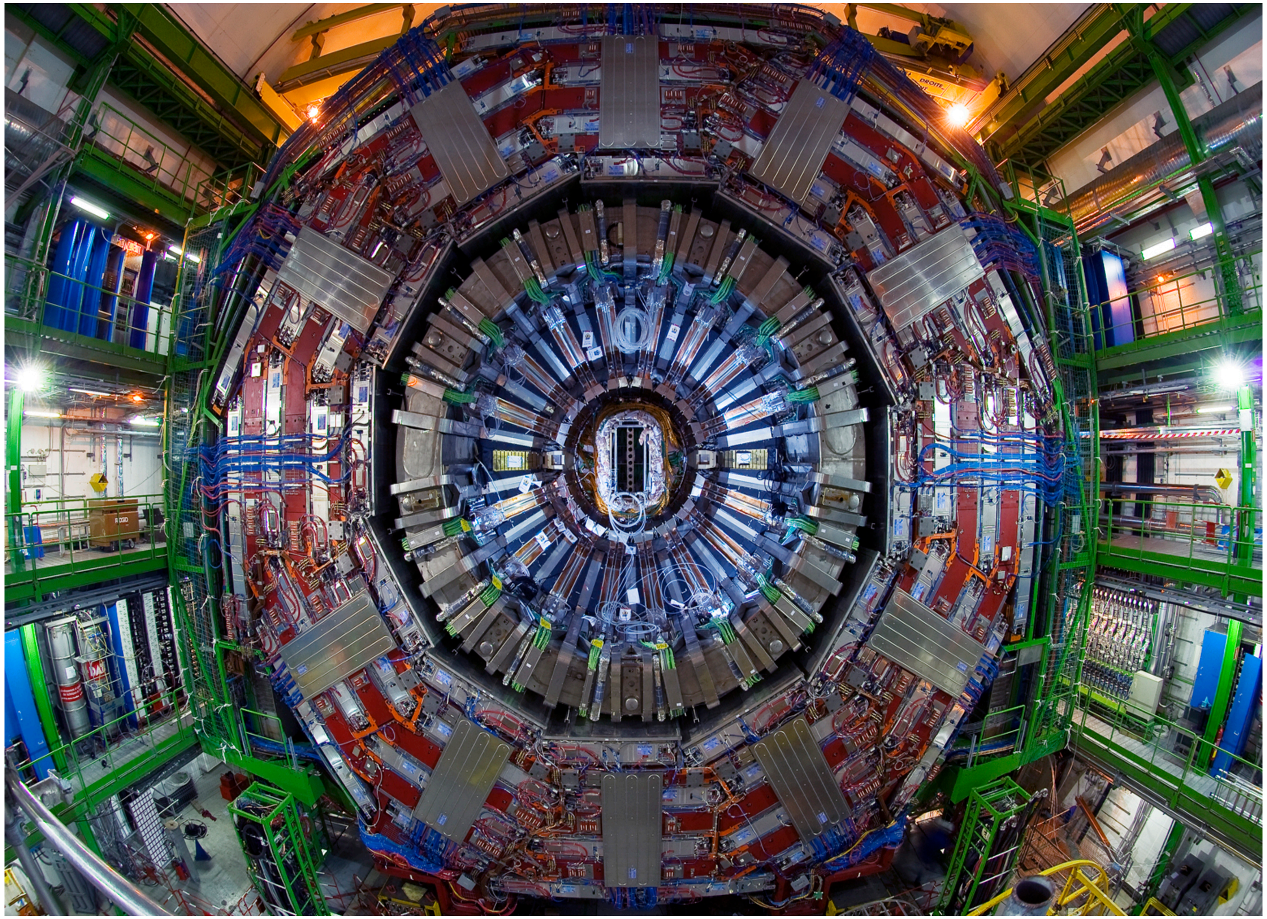
Massive feet made of steel that carry the weight of the entire detector with all its subsystems, a total of almost 12,500 tons.

Number of collaborating scientists: **Approximately 3,000**

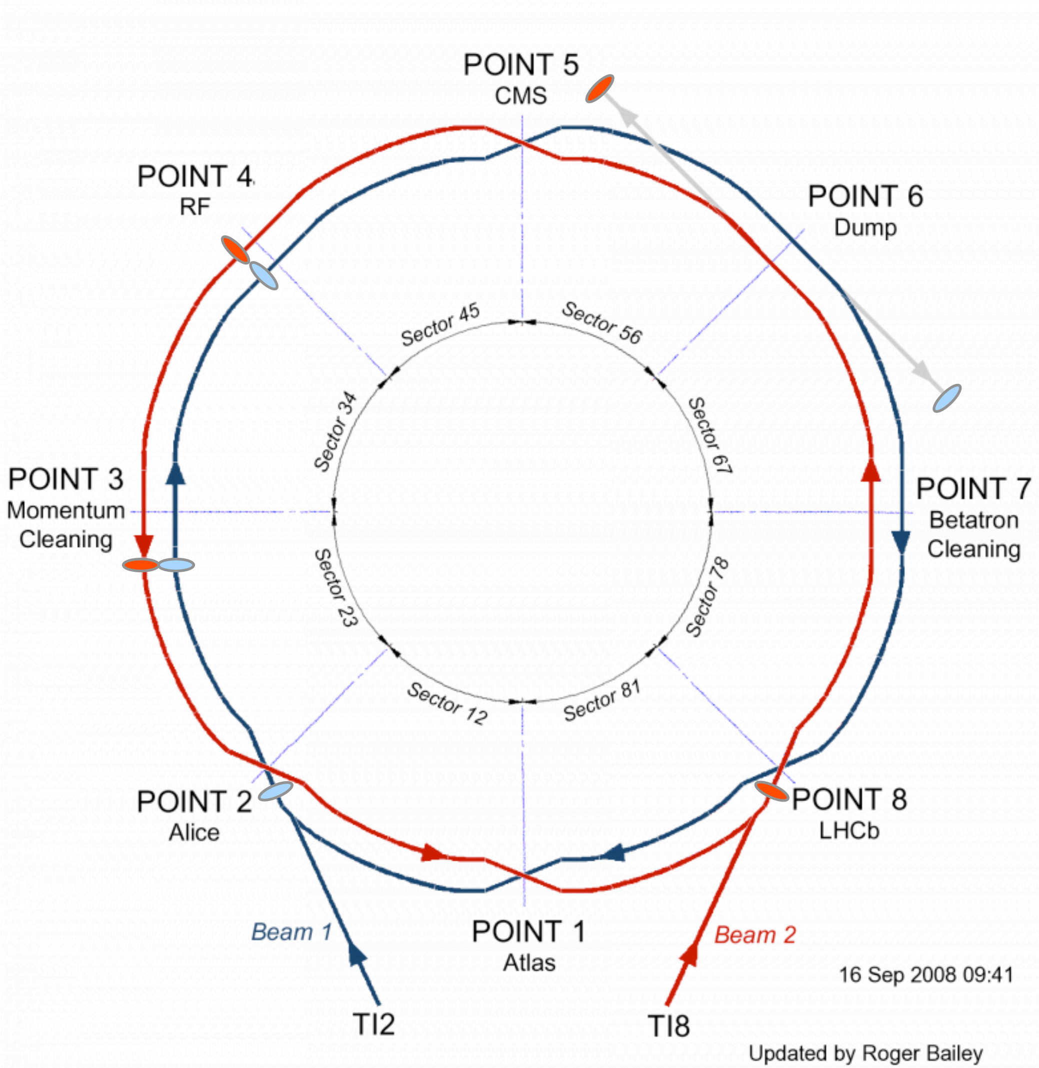
Number of collaborating countries: **39**

Location: **300 feet underground in Cessy, France**

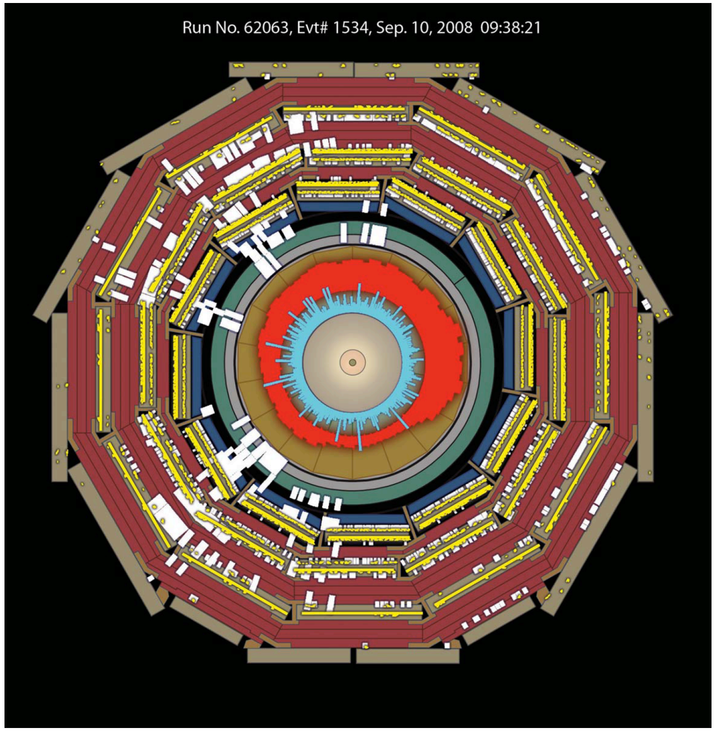
Fisheye View of the CMS Experiment



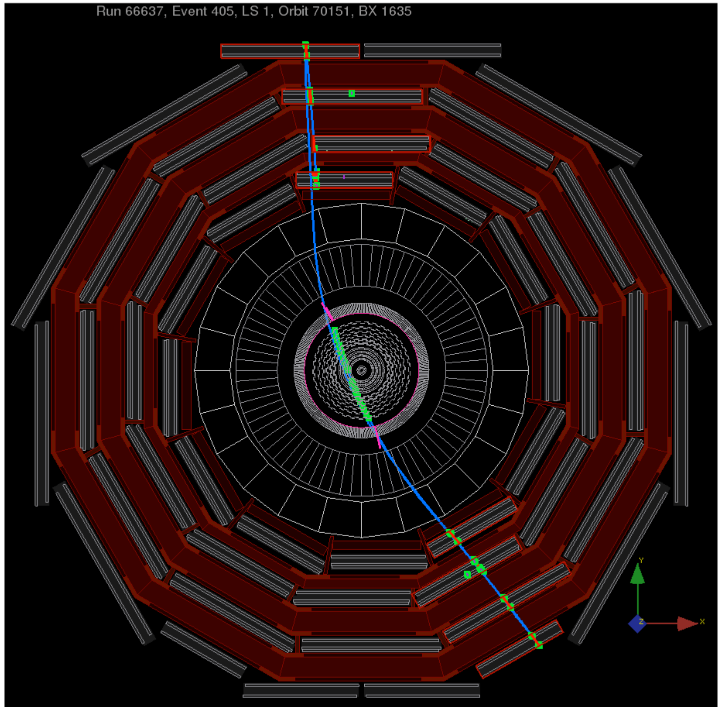
10 September 2008: first beam injection



...and first traces in CMS



1st beam event



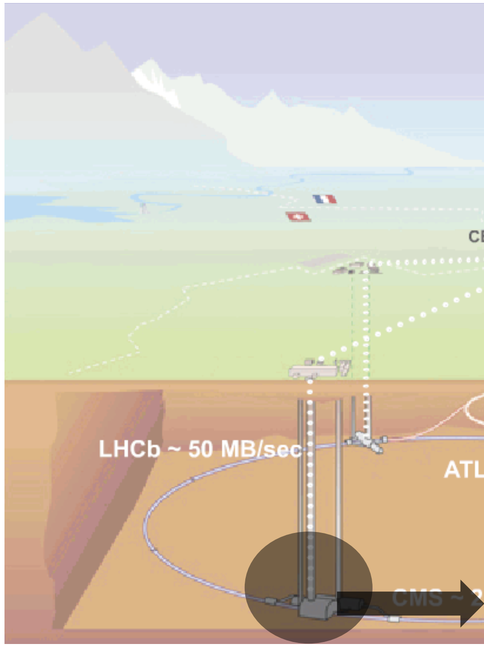
Cosmic muons

At the detector site: Online computing

High level trigger:

80 millions electronic channels
X 4 (each of them using 4 bytes)
X 40 millions (collision rate 40 MHz)
X 1/1000 (zero suppression)
X 1/100 000 (on line event filtering)
→ Around 10 Petabytes per year sent to CERN IT

- **DELL cluster (33 racks, dualcore Harpertown)**
- **230 TB disks acquisition system**



Core Computing

- From the HLT to delivery to the users, by going through CERN Tier-0, distributed Tier-1 and Tier-2
- Specific software development, user support, analysis tools
- Managing the data placement and distributed infrastructure during different phases of event reconstruction and MCPROD

The CMS Computing Project

- **Resource loaded WBS shows needs of 120 FTE/year**
- **Resources come from:**
  - The CMS collaborating institutes (human resources)
  - Dedicated funding for hardware and operations from science funding agencies
- **Relies on the WLCG and related projects:**
  - Providing and operating Grid computing services
- **Reviewed by:**
  - The CERN scientific review committee (LHCC)
  - individual funding agencies

Project	Activity	Task during FULL YEAR (*12.0 Months*)	Needed Mo or We	Needed FTE
Core Computing show 2 Managers	Computing Coordination www I show 2 Managers	L1 coordination show 2 Managers	6.00 Mo	0.50
		Computing resource planning and tracking show 2 Managers	5.40 Mo	0.45
		Computing resource planning and tracking show 2 Managers	5.40 Mo	0.45
	Processing and Data Access (PADA) www I show 2 Managers	Analysis / CRAB server validation show 2 Managers	18.00 Mo	1.50
		CMS Service Integration show 2 Managers	6.00 Mo	0.50
		Monitoring and Information Integration show 2 Managers	6.00 Mo	0.50
		Production Component Validation show 2 Managers	18.00 Mo	1.50
		Continuous Campaigns show 2 Managers	12.00 Mo	1.00
		L2 coordination show 2 Managers	6.00 Mo	0.50
	User Support www I show 2 Managers	Expert, Trouble-shooting, ticket tracking, CRAB support show 2 Managers	12.00 Mo	1.00
		User accounts and space administration show 2 Managers	3.60 Mo	0.30
		User Documentation Editor / Writer show 2 Managers	18.00 Mo	1.50
		L2 Coordination show 2 Managers	3.00 Mo	0.25
	Analysis operations www I show 2 Managers	L2 Coordination show 2 Managers	6.00 Mo	0.50
		Liaison to Physics show 2 Managers	3.00 Mo	0.25
		Physics group support for data placement and validation show 2 Managers	24.00 Mo	2.00
		CRAB server operations, debugging, validation, and support show 2 Managers	36.00 Mo	3.00
		User Support show 2 Managers	12.00 Mo	1.00
		Metrics and evaluation show 2 Managers	24.00 Mo	2.00
	Data Operations www I show 2 Managers	L2 coordination show 2 Managers	6.00 Mo	0.50
		Host Laboratory Processing (L3) show 2 Managers	12.00 Mo	1.00
		Distributed Re-Processing (L3) show 2 Managers	12.00 Mo	1.00
		Distributed Monte Carlo Production (L3) show 2 Managers	12.00 Mo	1.00
		Data Transfer and Integrity (L3) show 2 Managers	12.00 Mo	1.00
		Data Certification for physics (L3) show 2 Managers	12.00 Mo	1.00
		Data Operations show 2 Managers	120.00 Mo	10.00
	Facilities Operations www I show 2 Managers	L2 coordination show 2 Managers	6.00 Mo	0.50
		CMS VO management show 2 Managers	3.60 Mo	0.30
		Facilities operations at CERN show 2 Managers	30.00 Mo	2.50
		Distributed working fabric on Grid WMs show 2 Managers	6.00 Mo	0.50
		Distributed working fabric on DM (Storage/SRM) show 2 Managers	2.40 Mo	0.20
		Liaison with external projects (WLCG/EGEE/OSG, OPN, ?) show 2 Managers	15.60 Mo	1.30
		L3 coordination (T1/T2 coordination) show 2 Managers	12.00 Mo	1.00
		Site support show 2 Managers	45.60 Mo	3.80
		CMSSW deployment / verification show 2 Managers	6.00 Mo	0.50
		Site Quality Monitoring show 2 Managers	24.00 Mo	2.00
		CMS WebTools show 2 Managers	6.00 Mo	0.50
		Documentation, Training, Shift Organization show 2 Managers	6.00 Mo	0.50
	CMS Centers www I show 1 Manager	CMS Centre@CERN: hardware and system support show 1 Manager	3.00 Mo	0.25
		Common systems and tools for CMS Centres Worldwide show 1 Manager	12.00 Mo	1.00
		Documentation and User Support show 1 Manager	3.00 Mo	0.25

Project	Activity	Task during FULL YEAR (*12.0 Months*)	Needed Mo or We	Needed FTE
		L2 coordination	3.00 Mo	0.25
			3.60 Mo	0.30
			72.00 Mo	6.00
			218.40 Mo	18.20
			492.00 Mo	41.00
			10.00 Mo	0.83
			30.00 Mo	2.50
	Unspecified www I show 2 Managers	To be specified show 2 Managers	0.00 Mo	0
Σ	Σ	Σ (over 49 Tasks)	1420.60 Mo	118.38

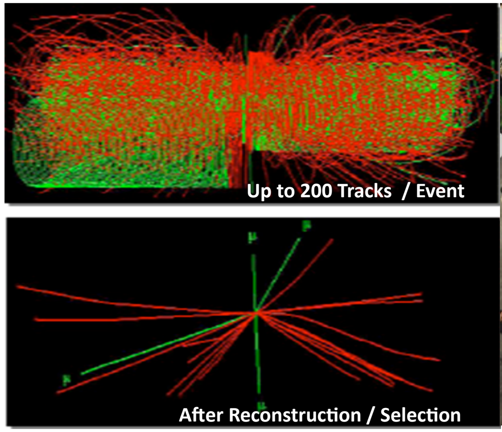
Σ (over 49 Tasks)

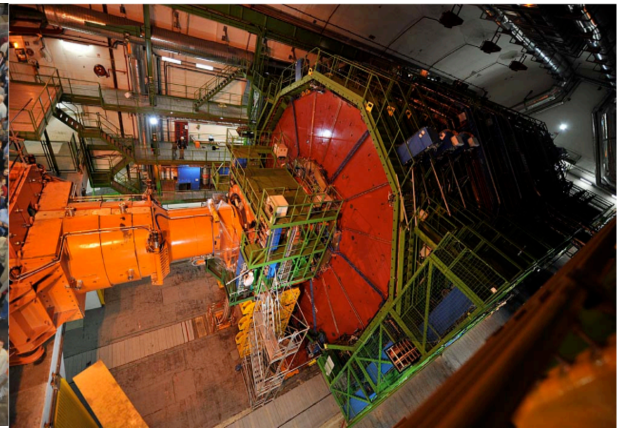
1420.60 Mo 118.38

The CMS computing project

- Provide Resources and Services to store/serve O(10) PB data/year

- Provide access to most interesting physics events to O(1500) CMS collaborators located in 200 institutions around the world





- Minimize constraints due to user localisation and resource variety
- Decentralize control and costs of computing infrastructure
- Team up with experts located at sites
- Share resources with other LHC experiments

→Find the answer on the Worldwide LCG GRID

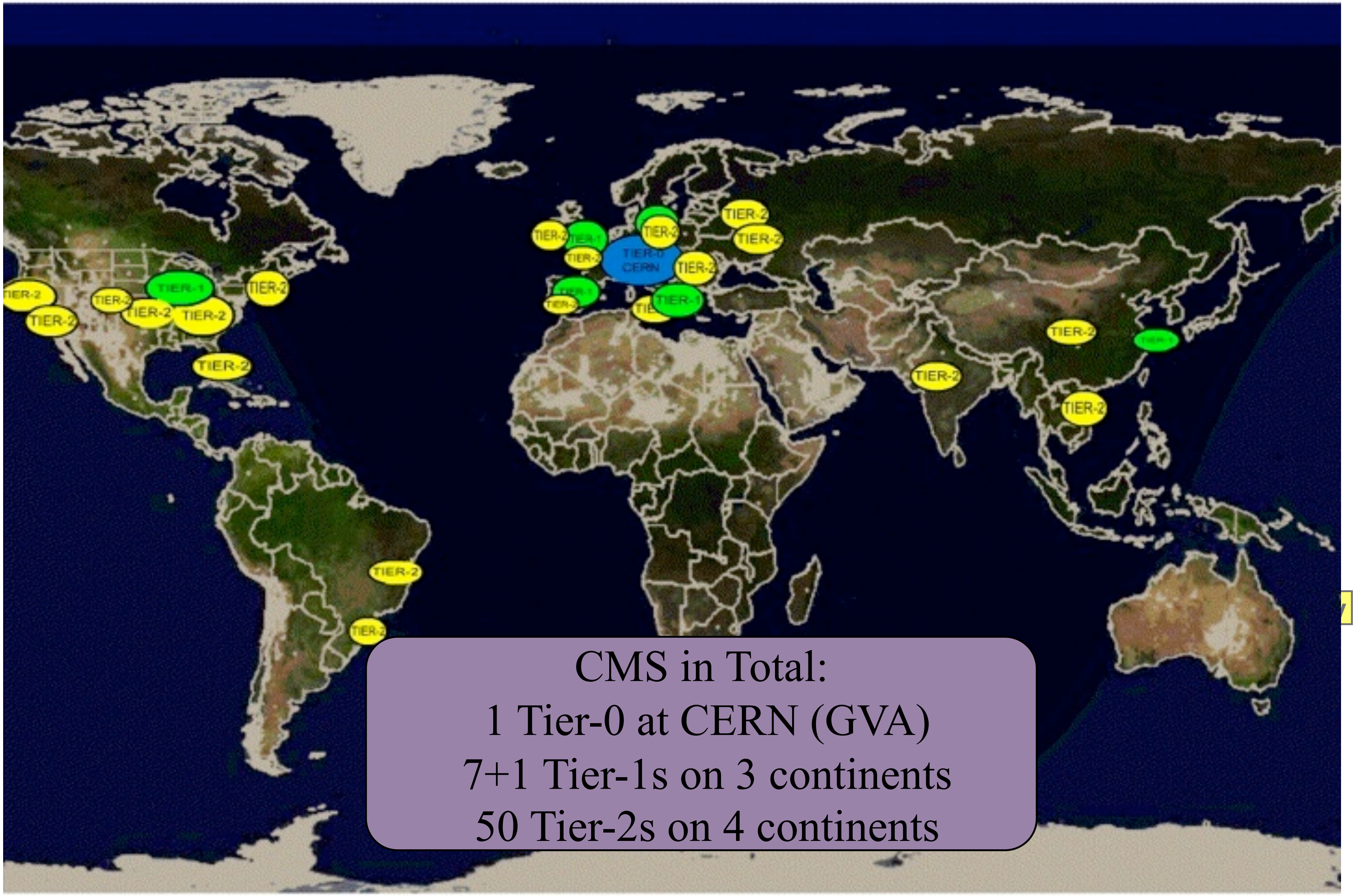
CMS Centers and Operations Shifts


CMS Remote Operations Centre at Fermilab

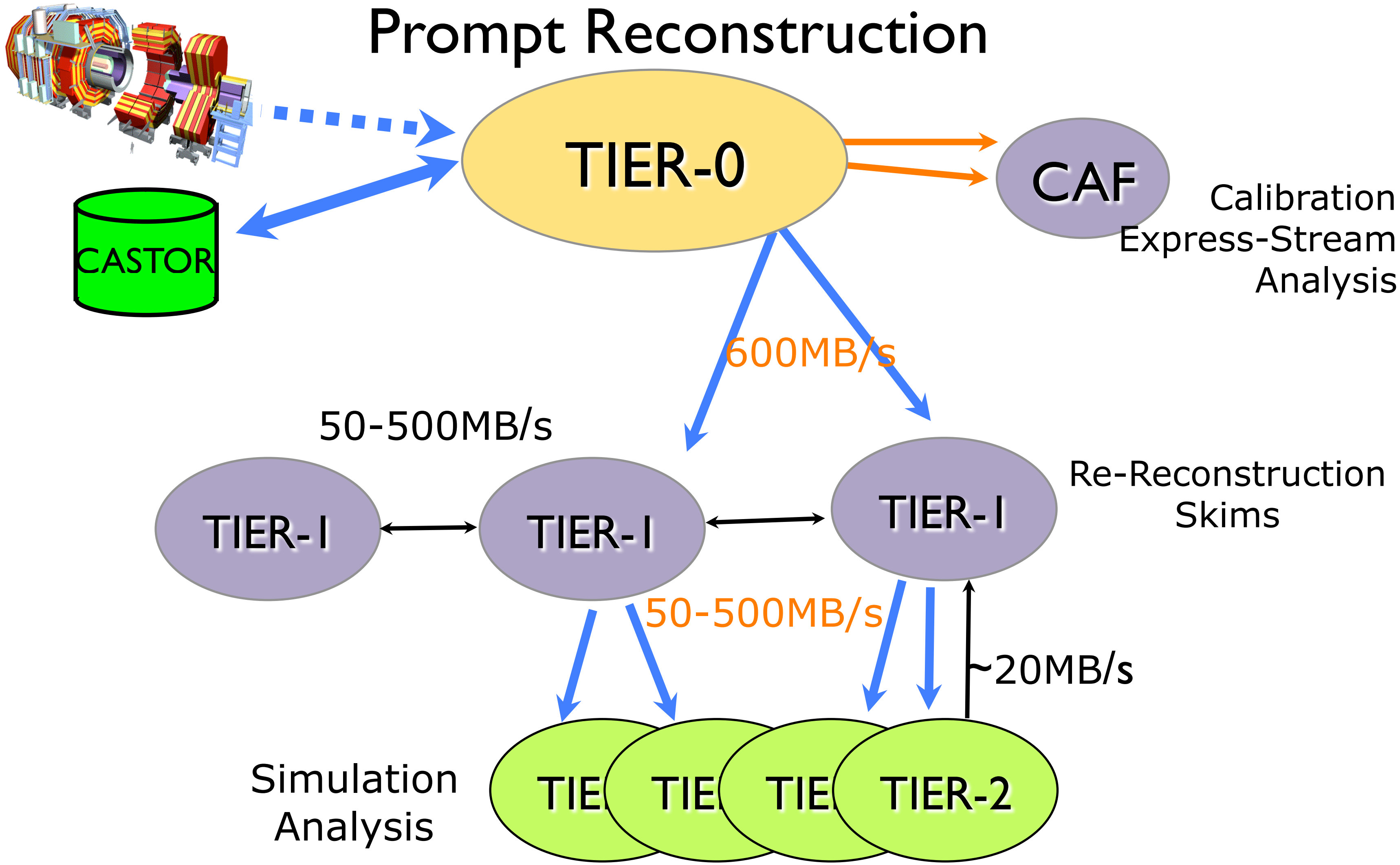

CMS Experiment Control Room


CMS Centre at CERN: monitoring, computing operations, analysis

- CMS running Computing shifts 24/7
- Encourage remote shifts
- Main task: monitor and alarm CMS sites & Computing Experts


CMS in Total:
1 Tier-0 at CERN (GVA)
7+1 Tier-1s on 3 continents
50 Tier-2s on 4 continents

Flow in the Computing Grid



Computing Resources: setting the scale

- Data Recording 2009-10 (Oct‘09-Oct‘10) 300 Hz / 2.2 x 10^9 events

Size&CPU per event

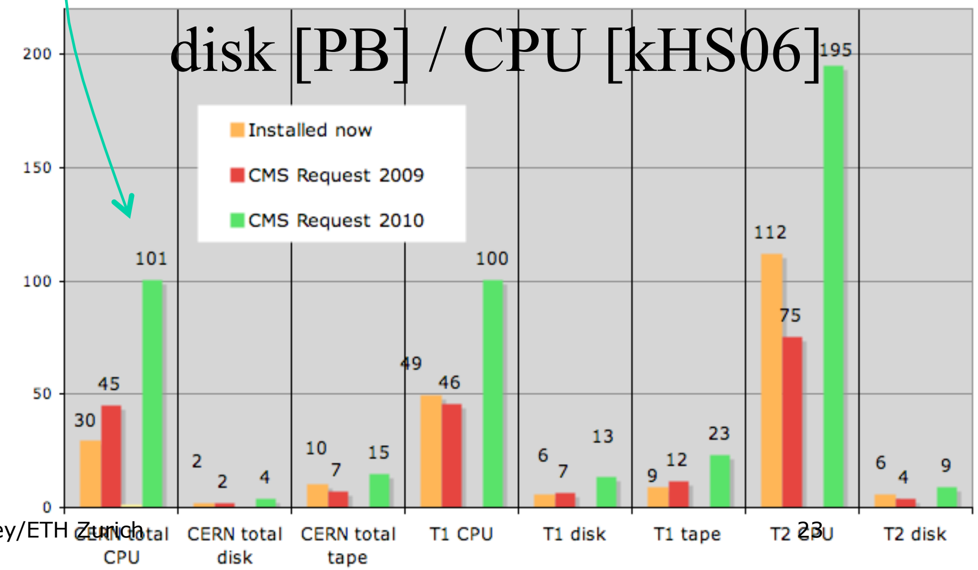
Data Tier	RAW	RECO	SIMRAW	SIMRECO	AOD
<Size> [MB]	1.5	0.5	2.0	0.5	0.1
CPU [HS06-sec]	-	100	1000		

- CMS datasets
  - -Higher level formatRECO, AOD)
  - - 1.5 times more Simulated

HEP-Spec2006 :

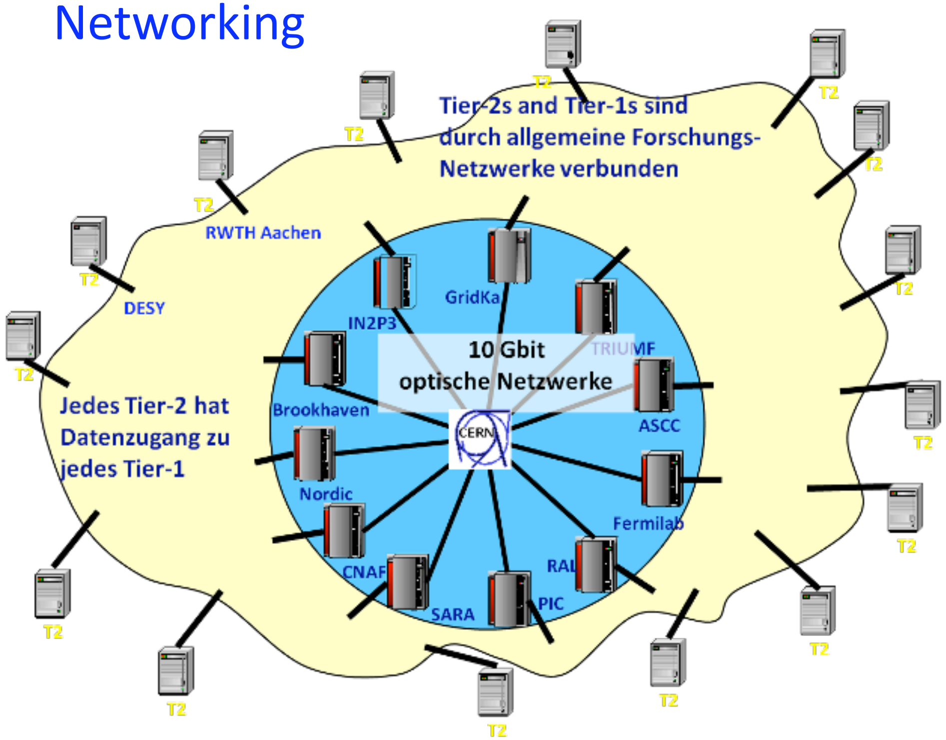
- ***Modern CPU ~ 8 HS06 / core***
- ***100 HS06-sec ~ 12.5 sec/event***
- ***100 kHS06 ~ 12,500 cores***

- During run 2009-10, CMS plans
  5 full re-Reconstructions, need :
  - - 400 kHS06 CPU (around 70‘000 computing cores)
  - - 26 PB disk
  - - 38 PB tape
- Resources ratio CERN / (T1+T2) :
  - - CPU : 25% , Disk : 15%



And not to forget the obvious...

Networking



GRID Middleware Services

- Storage Elements
- Computing Element
- Workload Management System
- Local File Catalog
- Information System
- Virtual Organisation Management Service
- Inter-operability between GRIDs EGEE, OSG, NorduGriD..

Site Specificities, e.g. Storage/Batch systems at CMS Tier-1s:

	FNAL	RAL	CCIN2P3	PIC	ASGC	INFN	FZK
Storage :	dCache/ Endstore	Castor	dCache/HPSS	dCache/ Endstore	Castor	Castor+ Storm	dCache/TSM
Batch :	Condor	Torque/Maui	Torque/Maui	BQS	Torque/Maui	LSF	PBSPro

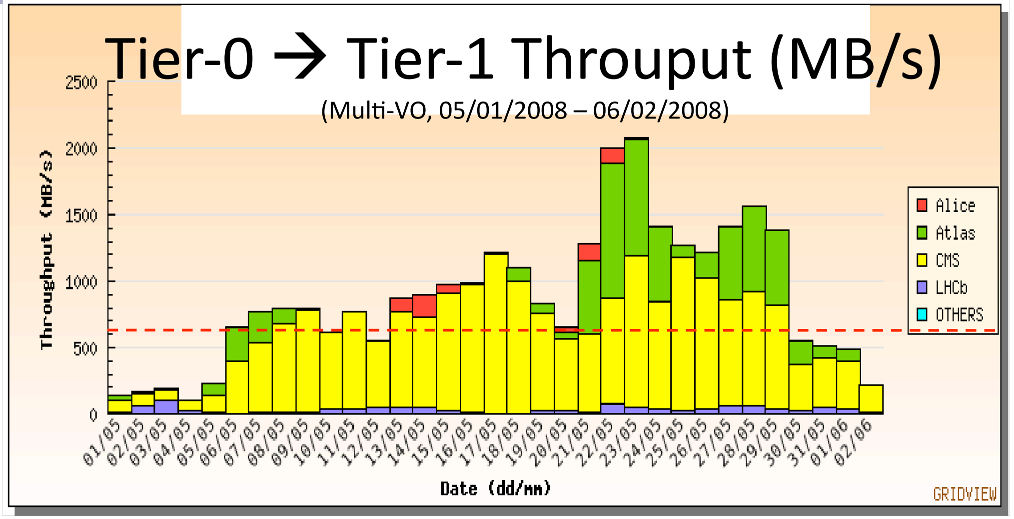
Data Reprocessing and Serving at T1s

CMS Data Distribution Model puts a heavy load on Tier-1s :

- **Disk and tape storage Capacity**
  - Custodial copy of Rec. Data (fraction) & archival of Simulated Data
  - Full set of Analysis Object Data - AOD (subset for 90% analyses)
  - Non-custodial copy of RECO/AOD encouraged
- **Processing Capacity**
  - Re-Reconstruction
  - Skimming of data to reduce the data size samples
- **Tape I/O bandwith**
  - Reading many times to Serve data to T2s for analysis
  - Writing for storage / Reading for Re-Reconstruction if not on disk
- **Full mesh (T1 – T2) strategy**
- **Pre-Staging strategy ?**
- **Other strong requirements**
  - 24/7 coverage and high (98%) availability (WLCG)
- **CMS Data Operations at Tier-1s**
  - Central operations or specialized workflows, no user access

Data Transfers challenges

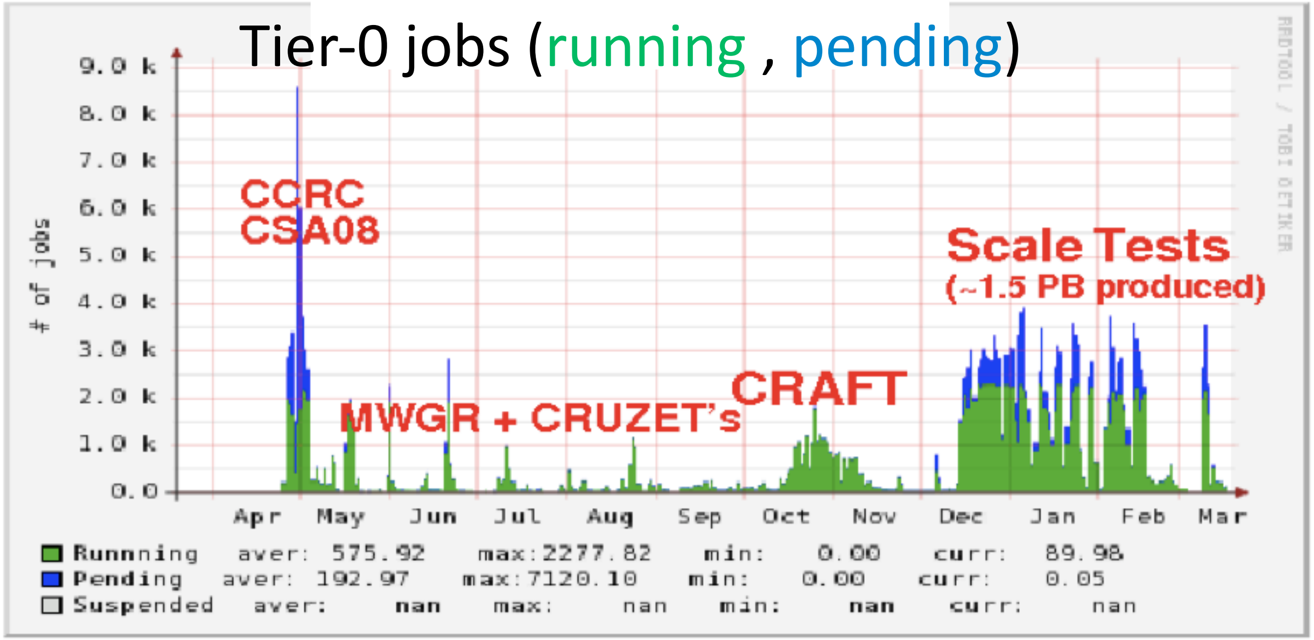
- **T0 – T1**
  - CMS regularly over design rate in 2008 multi-VO challenge

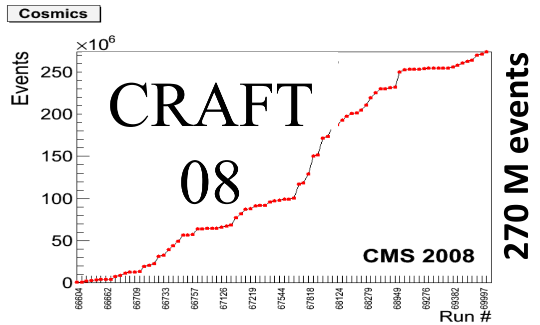


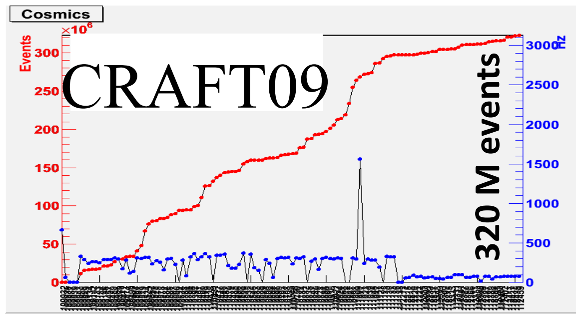
  - 2009-test : included tape writing step at T1s : observed T0→T1 transfer latency impacted by T1 tape system state (busy, overloaded, …)
- **T1 – T1: 2009 test**
  - simultaneous transfer 50TB AOD from 1 T1 to all T1s
    ➔ average 970MB/s (3 days), no big problems encountered
- **T1 – T2**
  - T1 Data Serving tests in during 2009-test

Primary Reconstruction tests at Tier-0

- CMS T0 operated from CMS Centre (CERN) and FNAL (Chicago)







- Despite lack of collision data, CMS able to commission T0 workflows, based on cosmic ray data taking at 300Hz :
  - Repacking – reformatting raw data, splitting into primary datasets
  - Prompt Reconstruction – first pass with a few days turnaround
  - Automated subscriptions to the Data Management system
  - Alignment and Calibration data skimming as input to the CAF

CMS Tier-2 Disk Space management

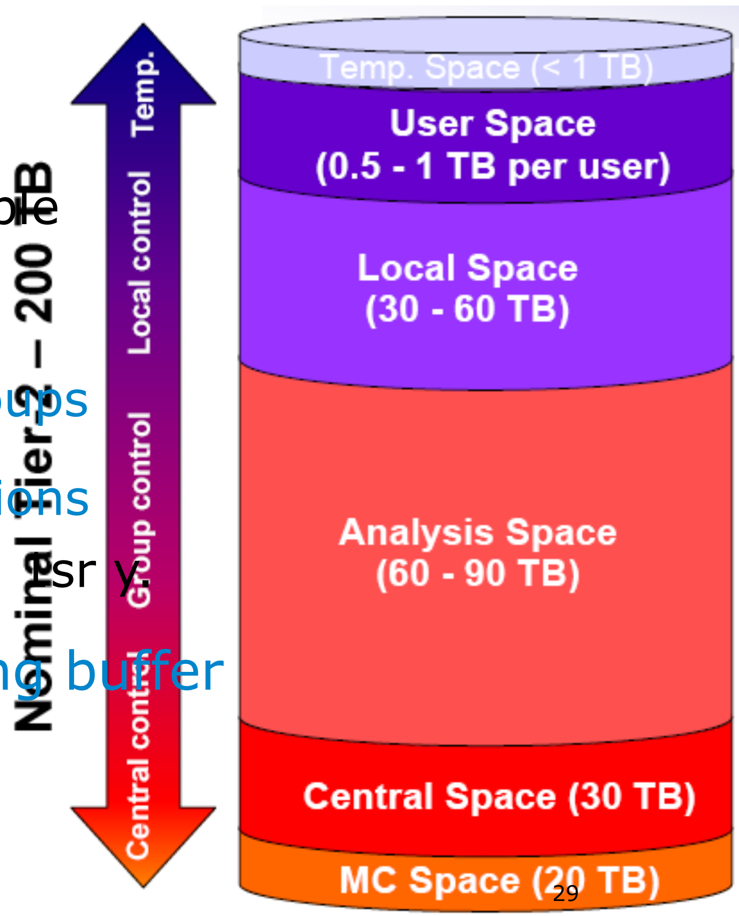
In CMS, jobs go to the data : distribute data broadly

CMS attempts to share management of the space across groups

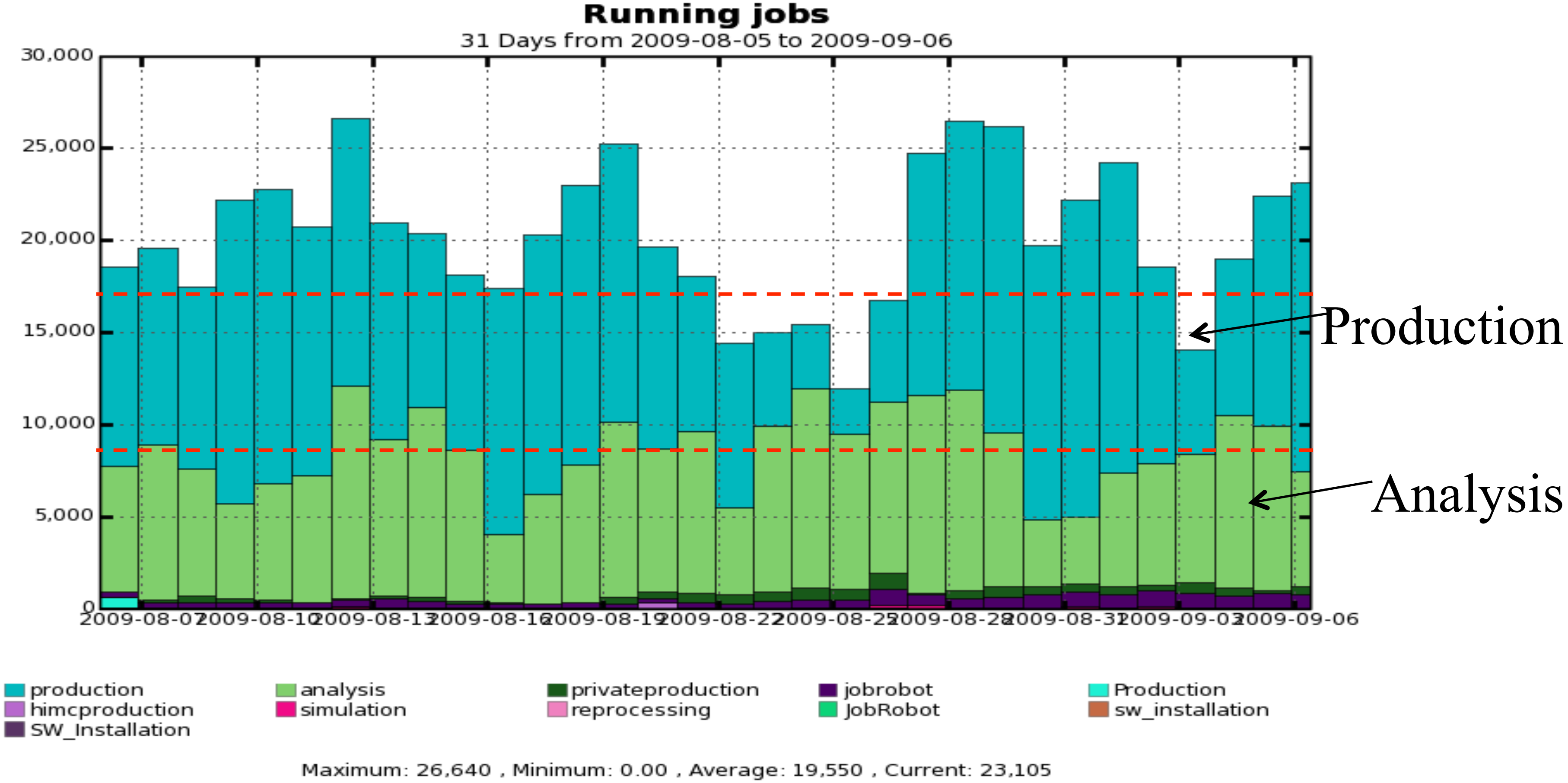
- Ensures people doing the work have some control

200TB of disk space at a nominal Tier-2

- 20 x 1TB is identified for storing local user produced files and making them grid accessible
- 30TB is identified for use by the local group
- 2-3 x 30 TB reserved to CMS PH Analysis groups
- 30 TB for centrally managed Analysis Operations expect to be able to host most RECO data in 1sr y.
- 20 TB of space for DataOps for MC staging buffer



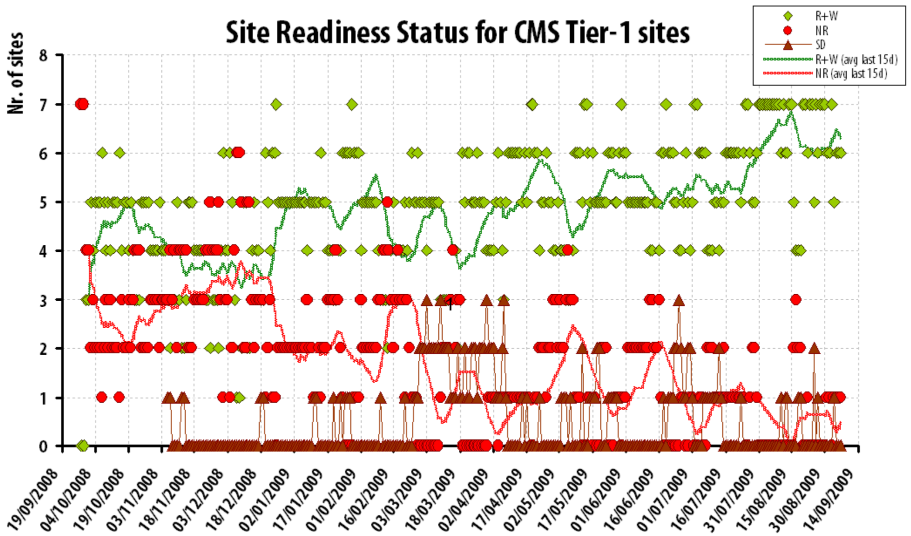
Job Slot utilization for Analysis

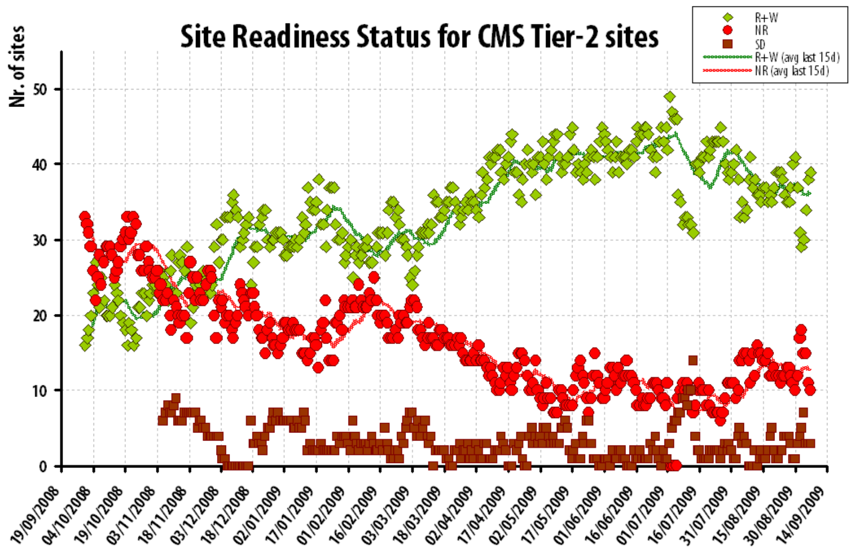


- Current CMS total CPU pledge at T2s : 17k jobs slots
- Analysis pledge : 50%
- Utilization in August was reasonnable
  - ➔ but need to go into sustained analysis mode

Site performance closely monitored

- **To measure global trends in the evolution of the reliability of sites**
  - Impressive results during the last year
- **Weekly reviews of the site readiness**
- **Production teams can better plan where to run productions**
- **Automatically map to production and analysis tools**





The CMS Site Commissioning team

Conclusions

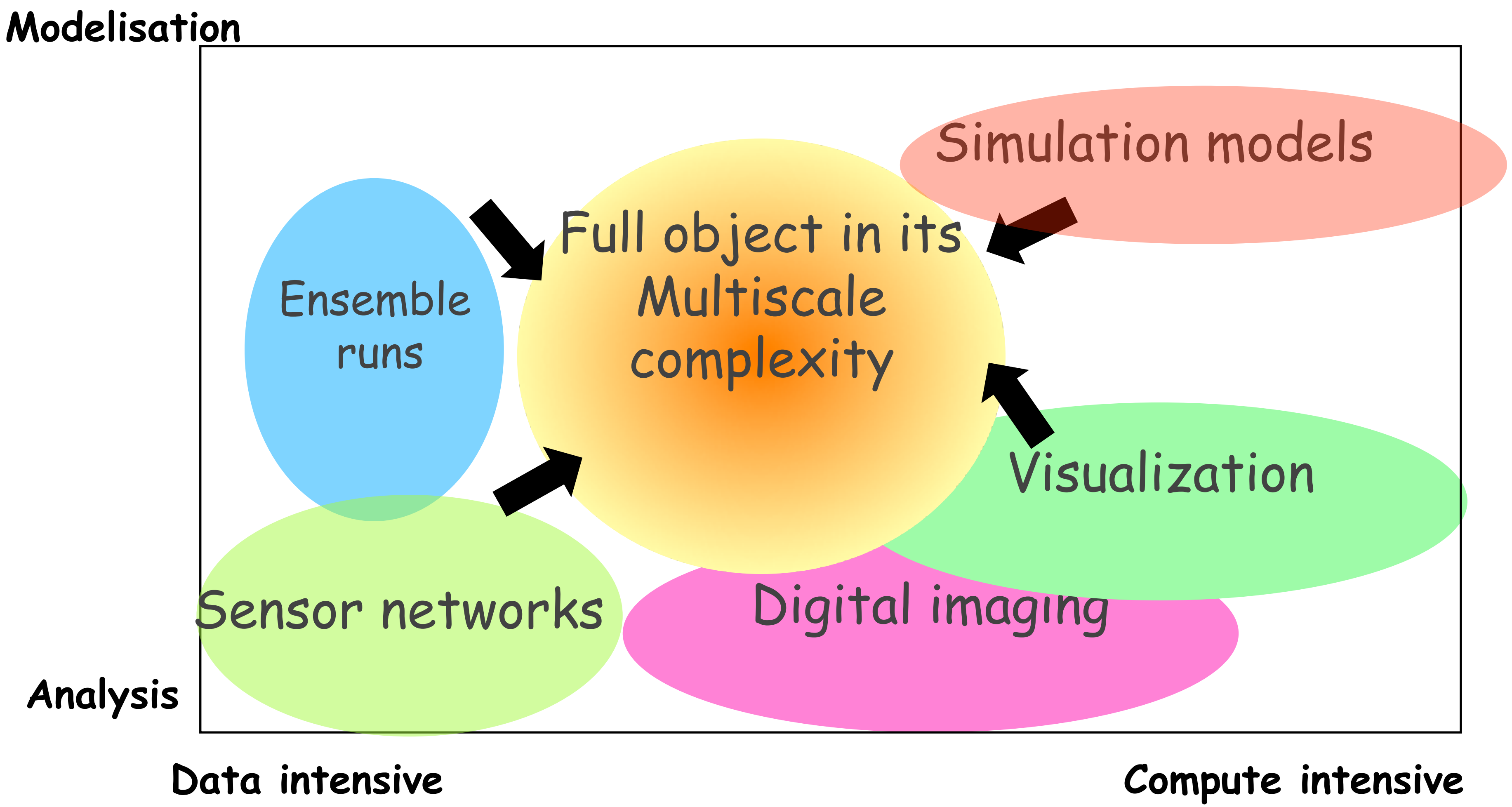
- **In last 5 years, CMS built a very large expertise in day to day GRID Computing Operations.**
- **Dedicated Data Challenges and Cosmic data taking very valuable, however not quite „the real thing“ which is :**
  - Sustained data processing
  - Strong demand on site readiness
  - High demand on data accessibility by colleagues physicists
- **Program until the LHC startup**
  - Tier-0 : repeat scale tests using simulated collision-like events
  - Tier-1 : STEP’09 tape and processing exercises where needed
  - Tier-2 : Support and improve distributed analysis efficiency
  - Review Critical Services coverage
  - Fine tune Computing Shifts procedures
  - Make sure (2010) resources pledges are available

Courtesy of M. Kasemann

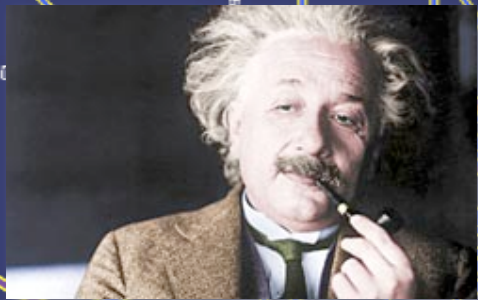
Something to learn for other communities?

- **"High Performance Computing"**
  - Balance between collecting, filtering, simulating, distributing and interpreting very large amount of data which comes into large bursts
  - The chain is as "HPC" as its weakest link
- **Challenges for data driven science**
  - On-line filtering of deluge of experimental or observational data
  - Repacking into high level objects, validation, quality control
  - Analyzing at fine granularity → accessibility, network capacity, data curation, heterogeneity of the systems
- **At the crossing point between experiments and simulation**
- **Going on step further**
  - Using data to enrich modelisation, simulation→ increase insight and knowledge
  - Integrating new data on the fly
- **Value stands in the data: simulated, experimental, observational, and in the knowledge it supports**

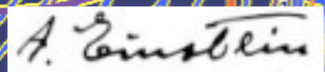
Cross fertilization: drawing the map



CMS



The most incomprehensible thing about the Universe is that it is comprehensible!



Special thanks to:

Matthias Kasemann
Felicitas Pauss
Klaus Freudenreich
Jim Virdee
CERN for photos and films

More information
www.cern.ch/cms