ATLAS EXPERIMENT

Introduction to the **Big Data challenges at LHC**

Farida Fassi

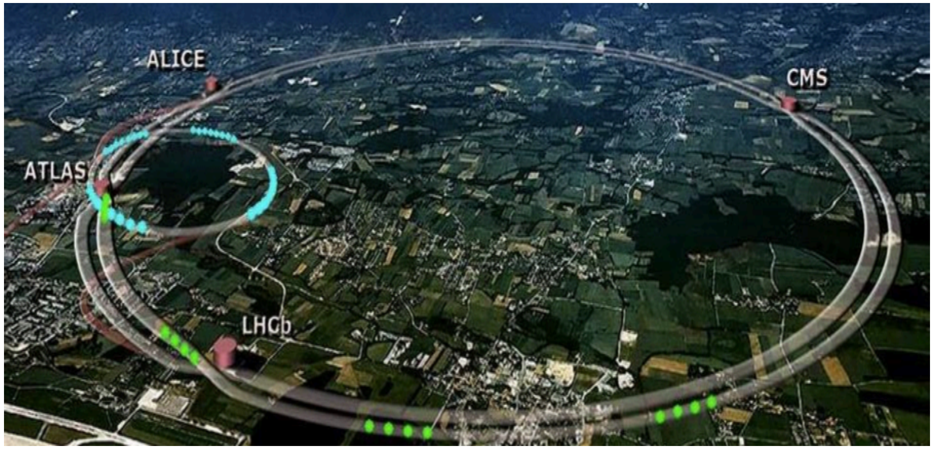
IFIC INSTITUT DE FÍSICA CORPUSCULAR

Mohammed V University in Rabat, Morocco

Why High Energy Physics e-Science?

- CERN is the European Organization founded in 1954 for Particle Physics Research in Geneva. Particle accelerators for High Energy Physics research.

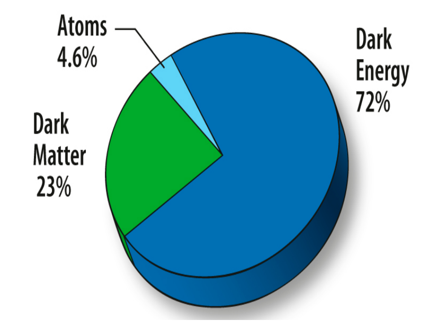
- Large Hadron Collider (LHC) is the largest and most powerful particle accelerator ever built.



- Data volumes at the LHC
  - up to 40 million collisions/second (MHz)
  - 1-1.5 MB/data per collision
  - 40 MHz * 1 MB = 40 TB/s

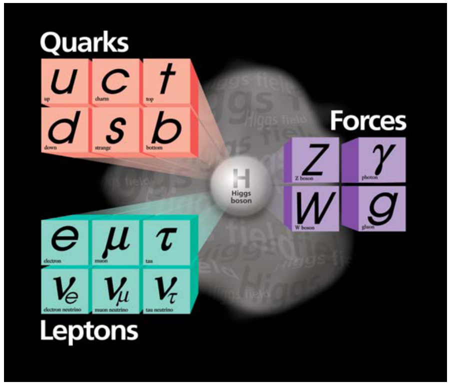
- High Energy Physics (HEP) uses immense data sets that require the computational Grids infrastructure deployed in the framework of the Worldwide LHC Computing Grid (WLCG).

Why LHC?
Fundamental questions of Particle Physics…

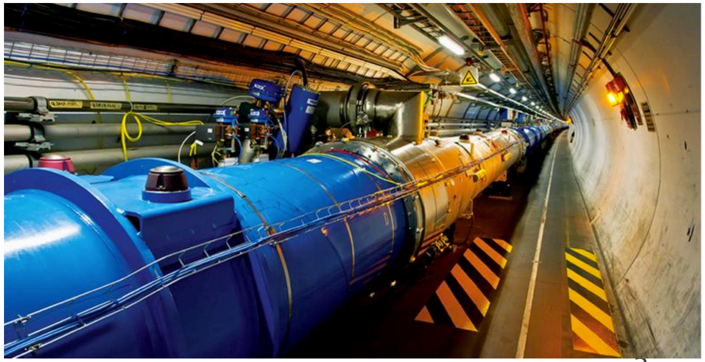


- Astrophysics/cosmological measurements show that only 4% of the matter is known.
- This tiny fraction of matter is well described by the "Stardard Model" (SM) of Particle Physics.

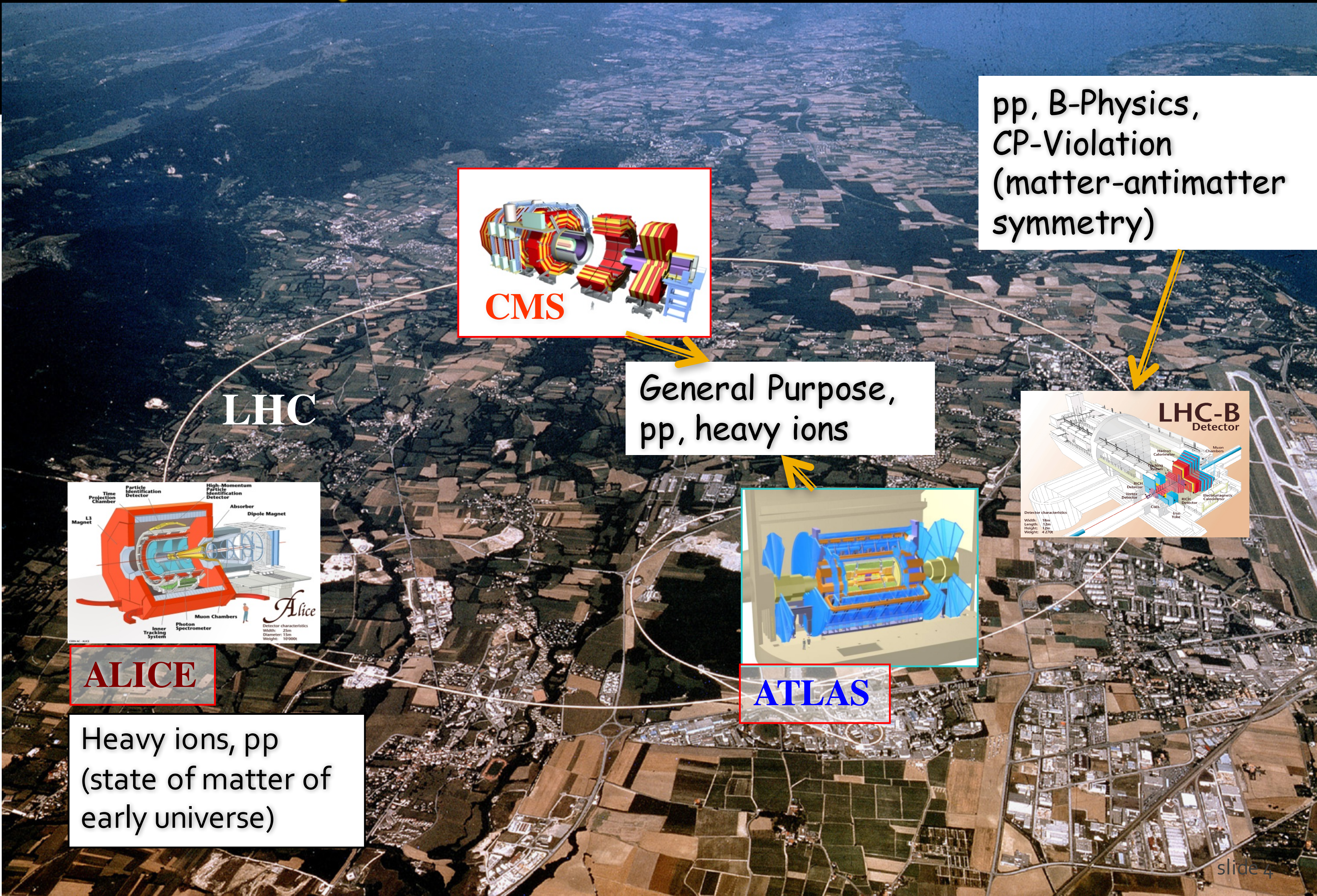
- LHC major goal is to provide an experimental verification
- of different theories within Particle Physics and HEP:
  - *What is the dark matter in the Universe?*
  - *Unification of fundamental forces?*
  - *Understanding space time matter versus antimatter.*
  - *How to explain that particles have mass?*
    - *Precise measurement of the Higgs boson.*



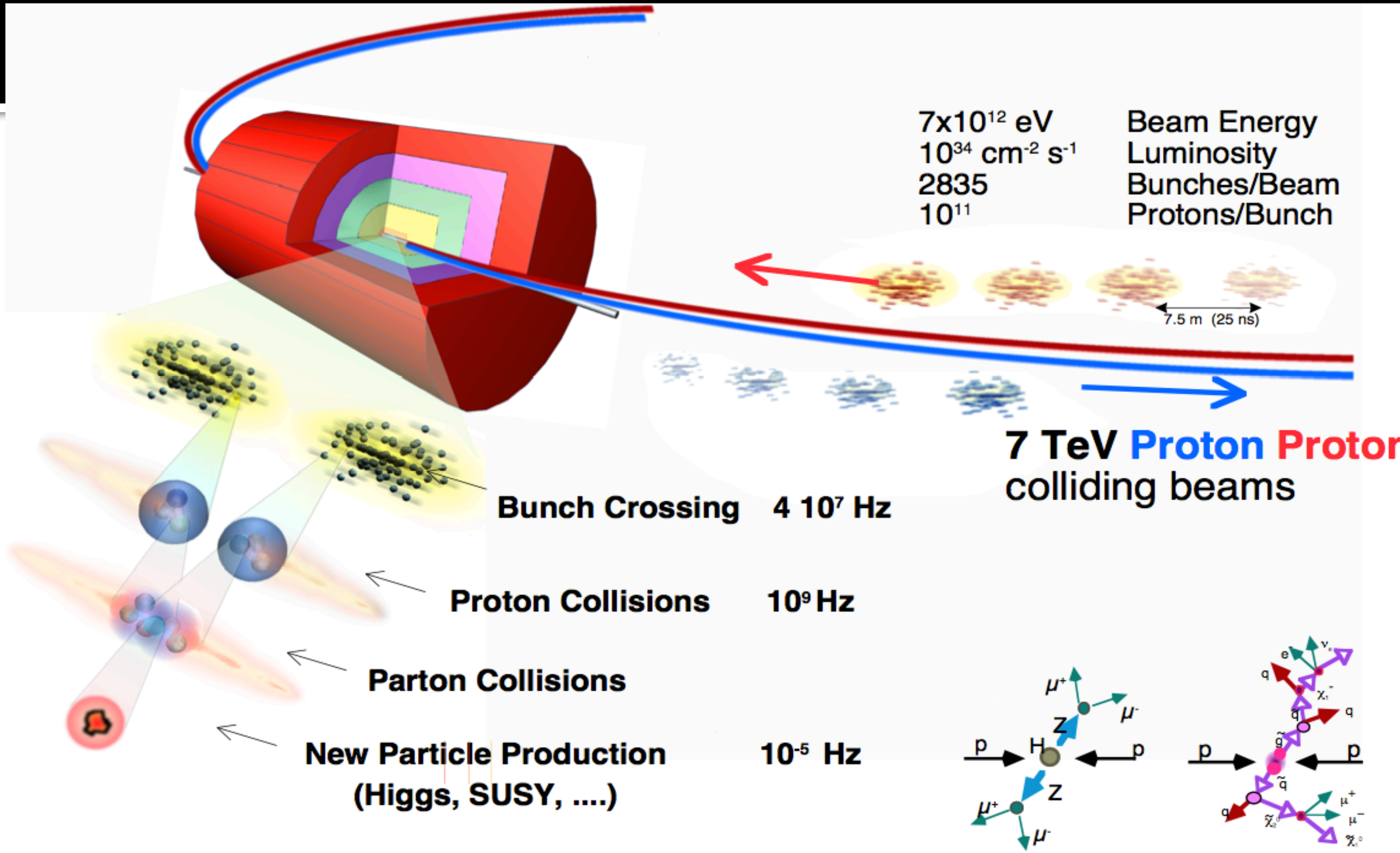
- The LHC is contained in a circular tunnel 27 km in circumference. LHC accelerates and collides beams of:
  - Protons and atomic nuclei



The Experiments at the LHC



Collisions at the LHC: summary

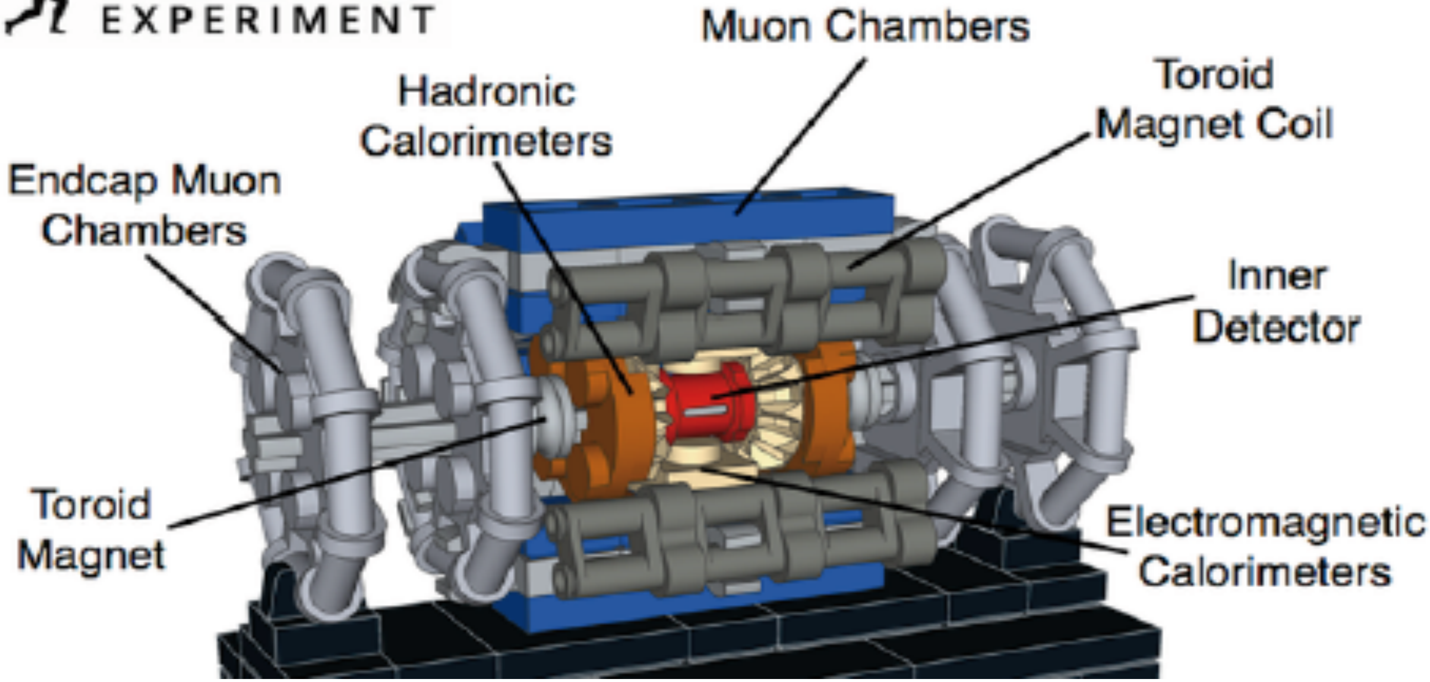


New physics rate ≈ .00001 Hz

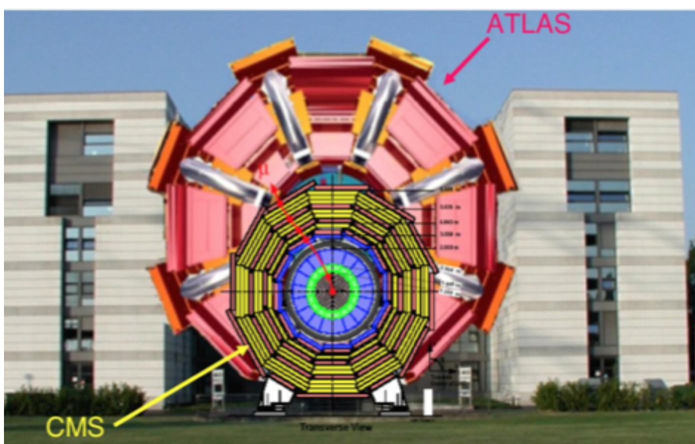
Event selection:
1 in 10,000,000,000,000

- LHC delivered billions of collision events to the experiments from proton-proton collisions in the Run-1 period (2009-2013)

ATLAS EXPERIMENT



Scale of ATLAS and CMS

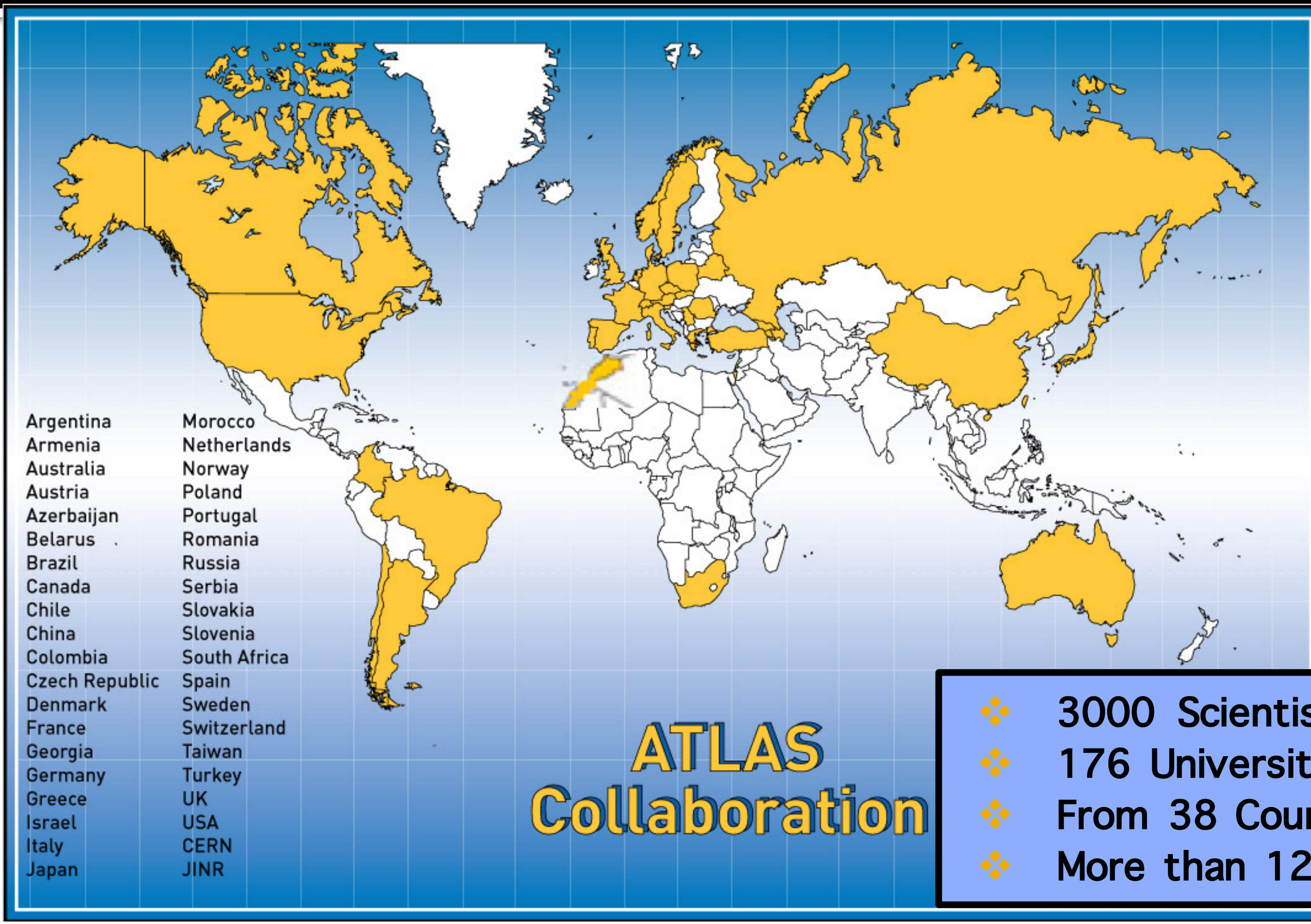


ATLAS (general purpose detector)

- Length: 44 m, diameter: 25 m
- Mass: ~7.0 ktons
- Two magnet fields:
  - Solenoid (ID): 2 T
  - Toroid (Muon System): 2-8 Tm

❖ ATLAS superimposed to a CERN 5-storey building

HEP e-Science more and more global



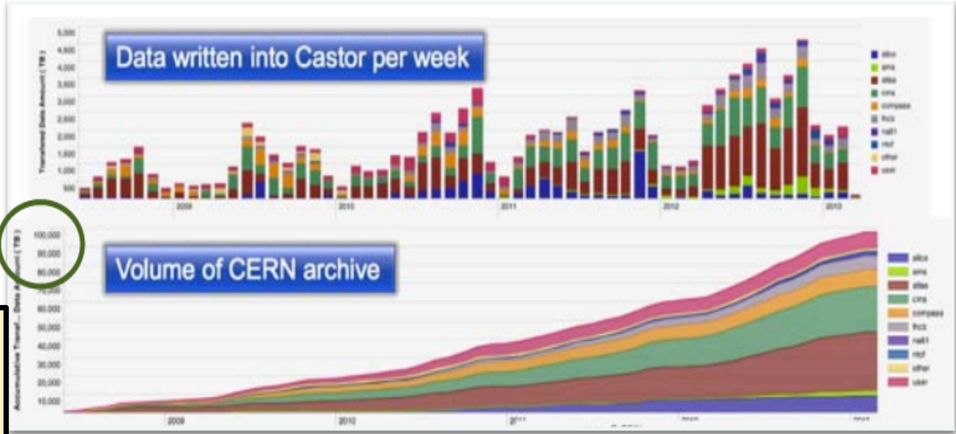
- 3000 Scientists
- 176 Universities and Labs
- From 38 Countries,
- More than 1200 students

LHC Big Data challenge…

- The LHC has already delivered billions of recorded collision events.

- Over 100 PB of data recorded.
- Several 100 PB more storage needed for data replication, simulation and Analysis derivation.

100PB



- Enormous challenge for the experiments for data collection, storage and processing

- What is this data?
- read-out of o(100M) detector channels
- 150 Million sensors deliver data
  - 40 Million times per second
- Raw data rate from LHC detector: 1PB/s
- This translates to Petabytes of data recorded world-wide (Grid)

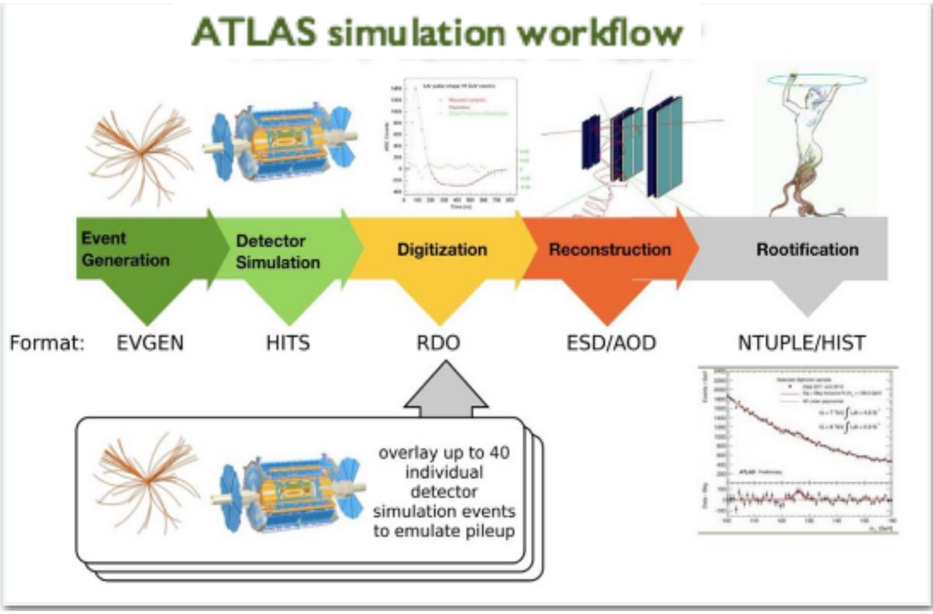
- Grid Computing is a critical tool to address the Big Data processing challenge and produce timely physics results…then success of LHC scientific program!!!

HEP Analysis Model: Detector Simulation

- The **Raw data** collected from the LHC is only part of the bigger data picture.

- MonteCarlo Simulation models the evolution of physics processes from collision to digital signals using knowledge from theory and test data.

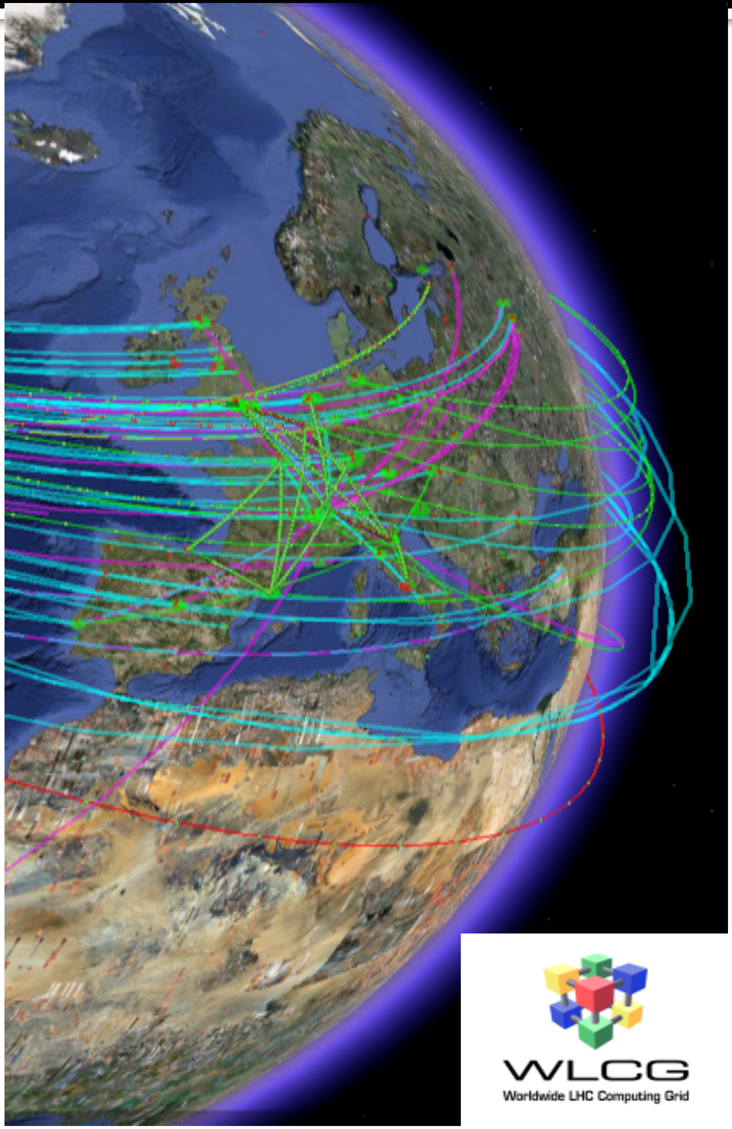
- Translate theoretical models into detector observations.
- Proper treatment of background estimation and sources of systematic errors.
- 10 billion events simulated by ATLAS to date



- Data-driven analysis compares (at statistical level) reconstructed events from real data with those predicted by simulation.

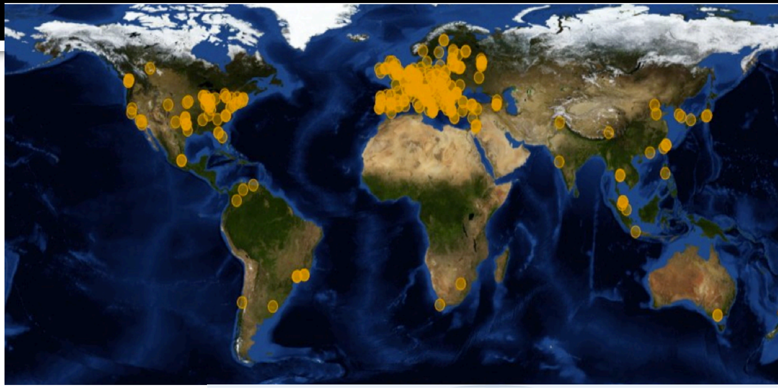
Comparable storage and processing requirements to Raw data

Data management



WLCG
Worldwide LHC Computing Grid

The Worldwide LHC Computing Grid (WLCG): Solution to LHC Big Data challenge…

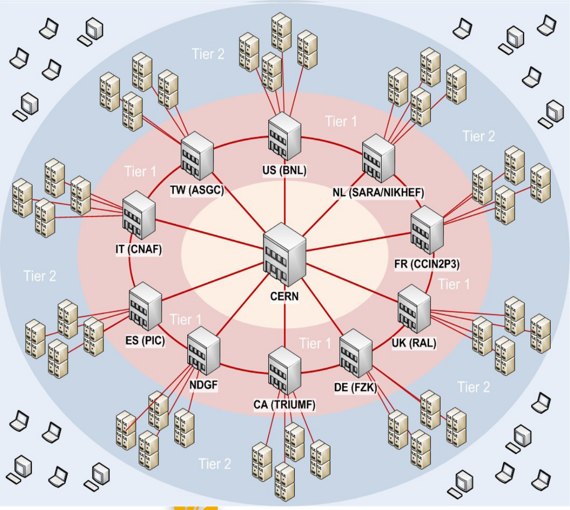


- **WLCG** is a global distributed computing Infrastructure, based on the Grid technologies.

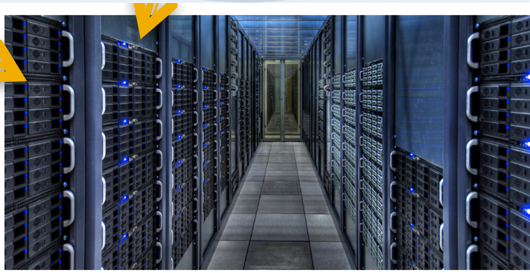
- WLCG provides seamless access to computing power and data storage capacity distributed over the globe.
- Computer centres worldwide arranged in a Tier structure.

Tier	Sites	Role	Example
0	1	Central Facility for data processing	CERN
1	12	Regional computing centres with high quality of service, 24/7 operations, large storage and compute resources	RAL
2	140	Computing centres within a regional cloud used primarily for data analysis and simulation	Edinburgh (ECDF)

ATLAS Tiered Computing mode



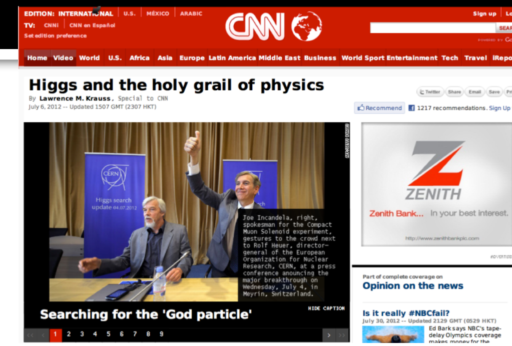
Tier-0 at CERN



❖ The main goal is to make use of the resources available and integrated into a single infrastructure accessible by all LHC, no matter where they are

From Big Data to Physics Discovery

- *Higgs boson is a major scientific discovery*



Global Effort → Global Success

Results today only possible due to extraordinary performance of accelerators – experiments – Grid computing

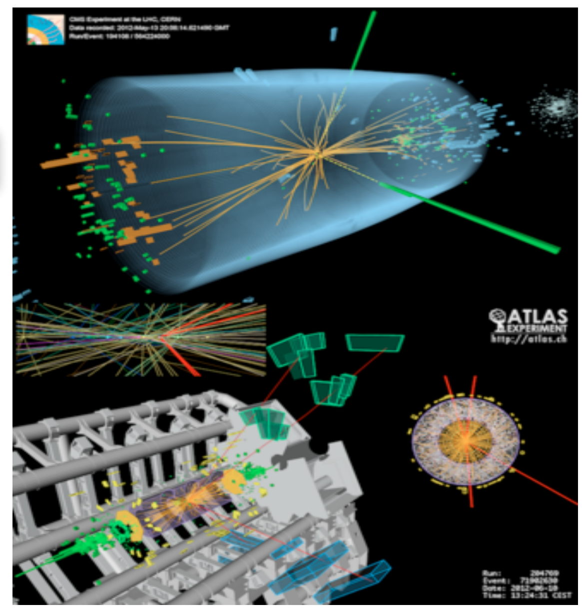
Observation of a new particle consistent with a Higgs Boson (but which one…?)

Historic Milestone but only the beginning

Global Implications for the future

CERN

R-D Heuer



Grid computing enables the rapid delivery of physics results

Distributed Data Processing system

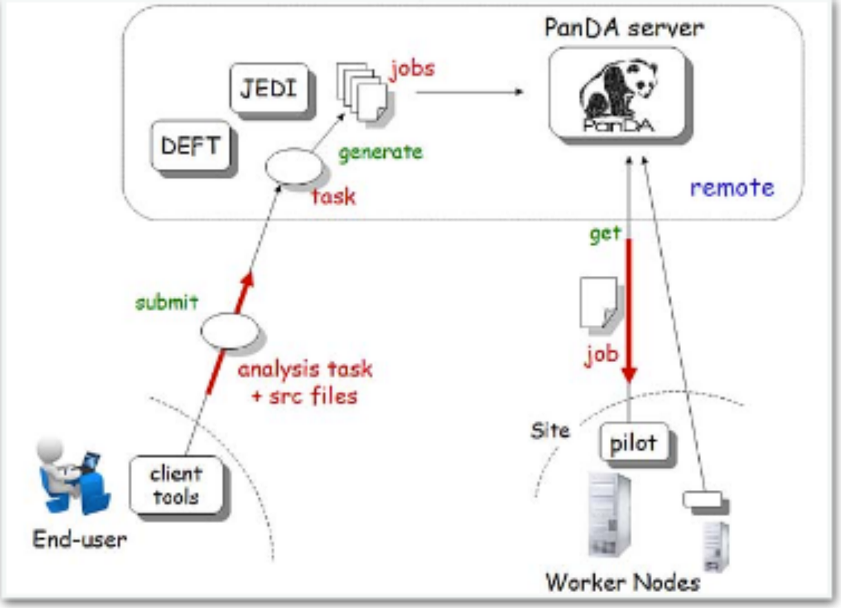
And,

Distributed Data Management system

Production and Distributed Analysis system

- Production and Distributed Analysis system (PanDA) is the Workload Management System (WMS) to run jobs on Grid.
  - PanDA is an unified system for Production and User Analysis capable of operating at LHC data processing scale.
- ATLAS Scientists use PanDA that makes distributed resources optimally accessible by all users.
- Flexibility in adapting to evolving hardware and network configurations.

- How PanDA works?
  - Uses a Pilot model to pull jobs from central queue once a suitable resource found.
- Pilot factories
  - continually submit jobs to available computing resources.
- The tasks are routed to sites based on the availability of relevant data and processing resources.



Data Management system

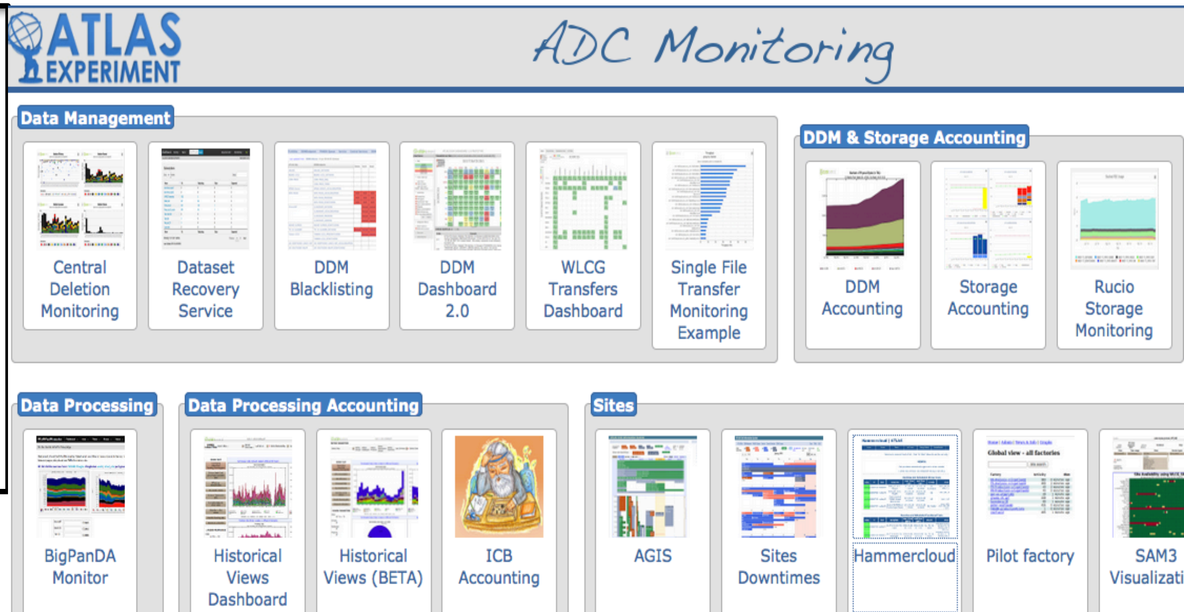
- Distributed Data Management (DDM) system designed to meet scalability, robustness, flexibility needed by ATLAS to manage the complete dataflow.
    - Rucio is the framework to manage all ATLAS data on the Grid.
        - Discover, transfer and delete data across all registered computing sites.
    - Rucio manages data replication and reduction within the lifetime range to Increase and/or reduce the number of copies based on data popularity.

ATLAS worldwide transfer volume per day

- Successful import/export data
- to Tier0+ Tiers1 sites
- ~100% efficiency

Monitoring & site validation

- The Monitoring system is critical for the success of all activities on the Grid.
- Monitoring is the key for synchronizing Distributed Operations.



HammerCloud | ATLAS

Home | Tests | Robot | Overviews | PanDA Dashb. | More HC... | Help | Administration

Welcome to HammerCloud-ATLAS.

Running and Scheduled AFT/PFT Tests

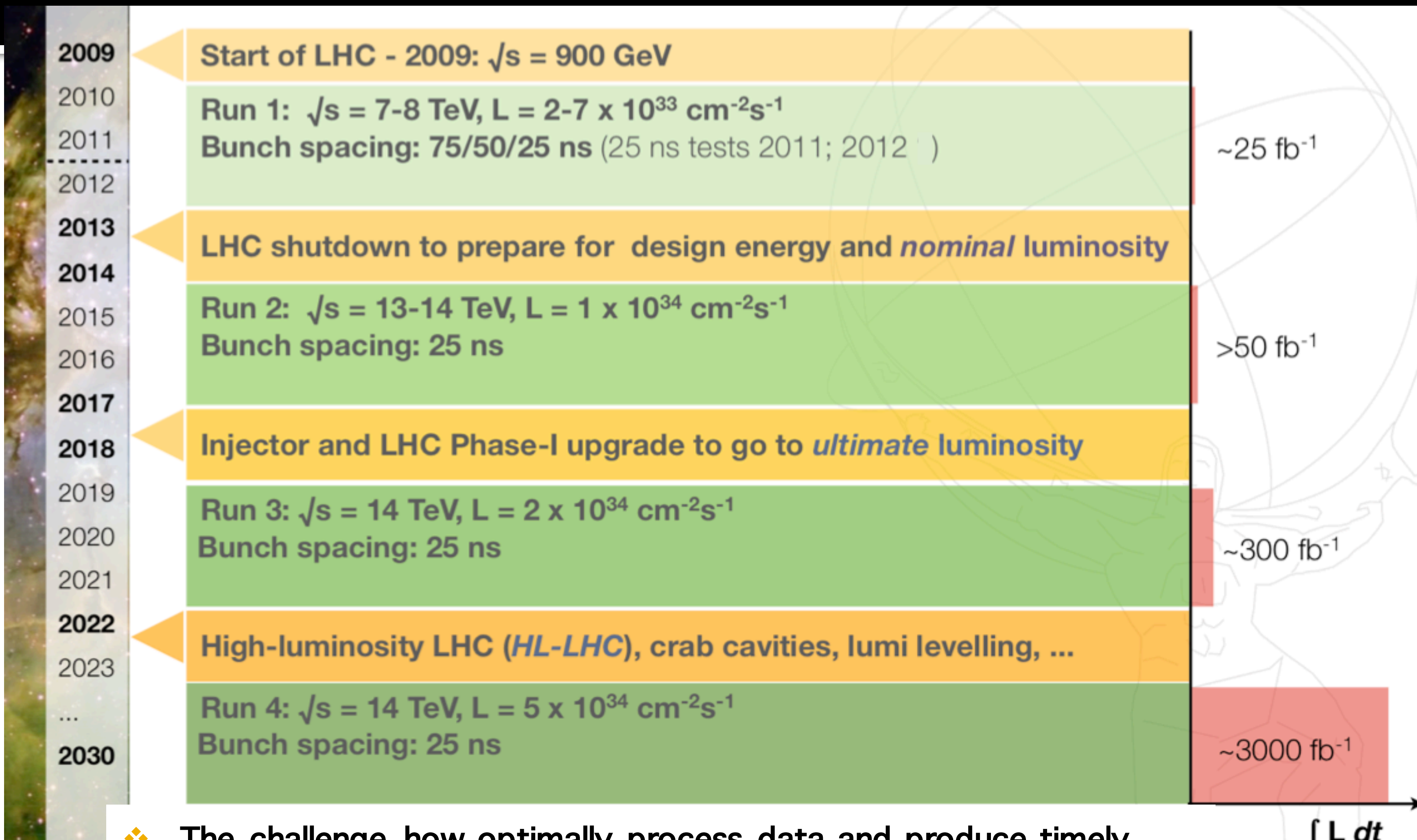
State	Id	Host	Template	Start (CET)	End (CET)	Sites	subm jobs	run jobs	comp jobs	fail jobs	tot jobs
running	20054631	it-hammercloud-submit-atlas-07	506: AFT Reco_trf 17.2.7.6 SMWZ_NTUP Panda	09/Apr, 0:16	09/Apr, 22:33	ANALY_TRIUMF, ANALY_AUSTRALIA, ANALY_SCINET, 200 more...	66	85	4790	309	5290
running	20054635	it-hammercloud-submit-atlas-07	508: AFT Reco_trf 17.2.7.6 data11 SMWZ_NTUP Panda	09/Apr, 2:36	10/Apr, 2:43	ANALY_TRIUMF, ANALY_AUSTRALIA, ANALY_SCINET, 200 more...	85	55	3816	197	4188
running	20054641	it-hammercloud-submit-atlas-02	489: PFT mc12 AtlasG4_trf 16.6.7.34	09/Apr, 11:52	10/Apr, 10:59	CA-VICTORIA-WESTGRID-T2, SFU-LCG2, TRIUMF, 174 more...	105	134	712	19	1009
running	20054643	it-hammercloud-submit-atlas-02	621: PFT mc14 Sim_tf 17.7.3.12	09/Apr, 12:12	10/Apr, 14:17	CA-VICTORIA-WESTGRID-T2, SFU-LCG2, TRIUMF, 174 more...	103	157	634	60	988
running	20054646	it-hammercloud-submit-atlas-07	571: PFT mc12 AtlasG4_trf 17.2.11.8	09/Apr, 12:44	10/Apr, 11:52	CA-VICTORIA-WESTGRID-T2, SFU-LCG2, TRIUMF, 174 more...	100	130	515	33	815
running	20054647	it-hammercloud-submit-atlas-07	505: AFT UA 17.2.7 Panda	09/Apr, 13:46	10/Apr, 14:05	ANALY_TRIUMF, ANALY_AUSTRALIA, ANALY_SCINET, 200 more...	92	35	610	29	809

- End-to-end testing tool for Grid sites.
- Sites validation before receiving activities.
- Stress tests mimicking user analysis are used for automatic exclusion of sites failing the tests from brokerage.
- Big improvement of success activities.

Distributed Analysis Support Team (DAST)

- DAST provides the first contact point to help thousant of Grid users.
  - DAST deals with all kind of the distributed analysis-related-issues.
  - An efficient user support is crucial to get physics results fast.
  - DAST plays a key role to solve these users-related-issues:
    - Panda-clients and Ganga,
    - ATLAS software, Physics Analysis Tools
    - Site service problems
    - DDM-clients, data access at sites and data replication
    - Monitoring system
- Two expert shifters on duty during working hours; one in the North American time zone and one in the European time zone, covering 16 hours/day.

The LHC Roadmap: the Challenge to Computing Repeats periodically…



- The challenge how optimally process data and produce timely Physics results that end up in a great success as Run-1.

Conclusions

- The LHC machine, the ATLAS experiment, the computing facilities behaved brilliantly during Run-1, and we have a major scientific discovery in our pocket; the Higgs Boson.

  - The LHC faces big computing challenges ahead to avoid constraining science output.
  - *Computing is crucial factor for the success Physics program of the LHC experiments.*
  - *Lots more excitement to come in LHC with Run-2 and beyond…*