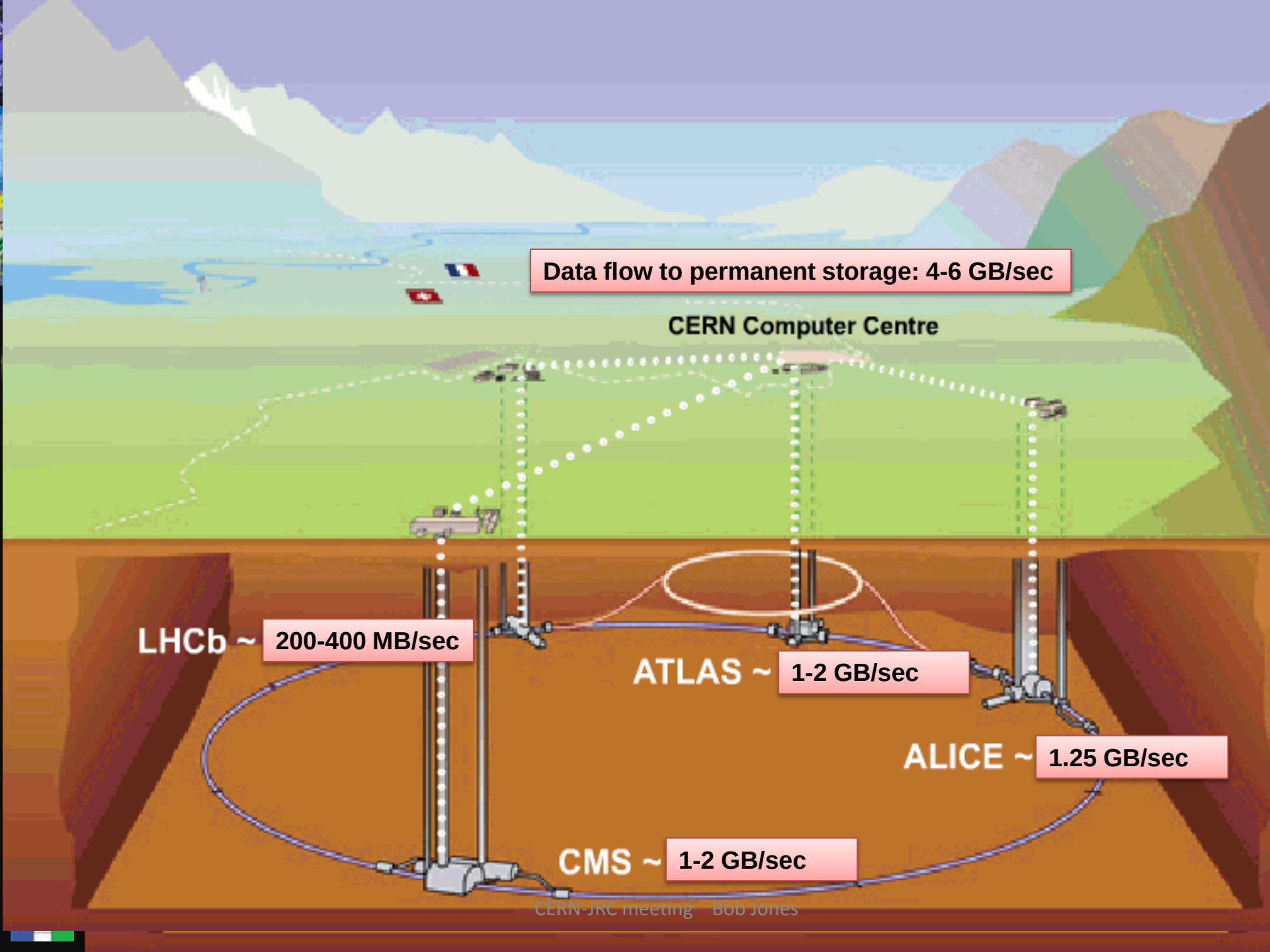
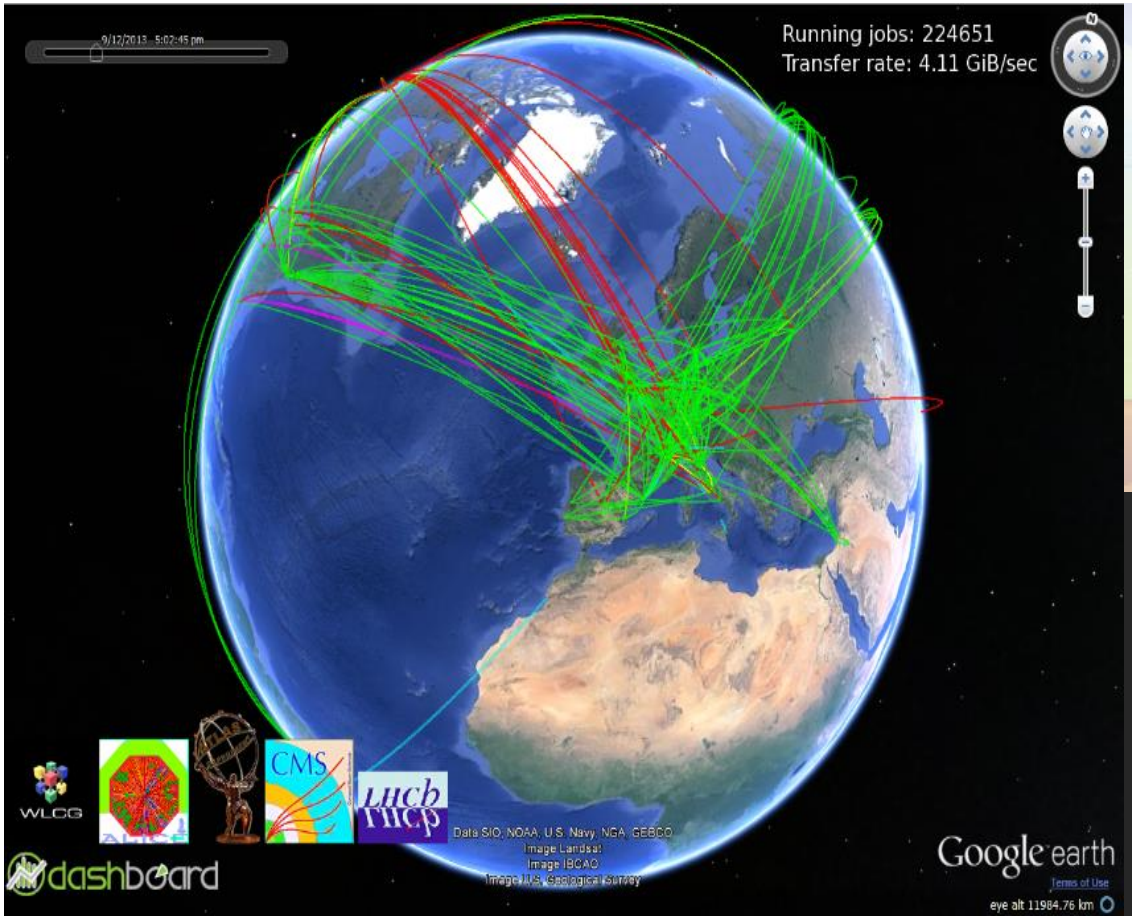

CERN

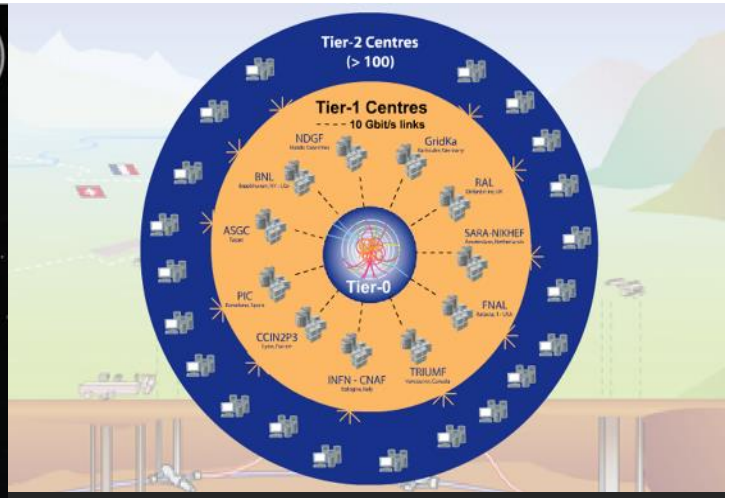
Big Science meets Big Data

Bob Jones
Head of CERN openlab


Data flow to permanent storage: 4-6 GB/sec
CERN Computer Centre
LHCb ~ 200-400 MB/sec
ATLAS ~ 1-2 GB/sec
ALICE ~ 1.25 GB/sec
CMS ~ 1-2 GB/sec

The Grid



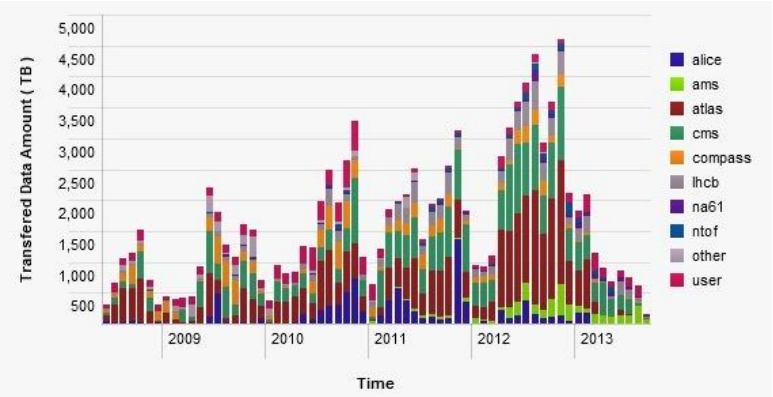


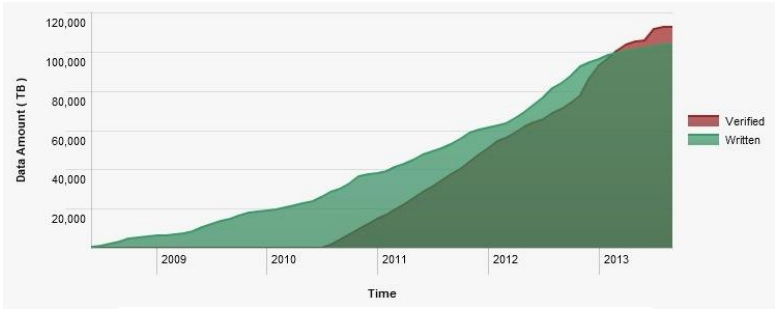
Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

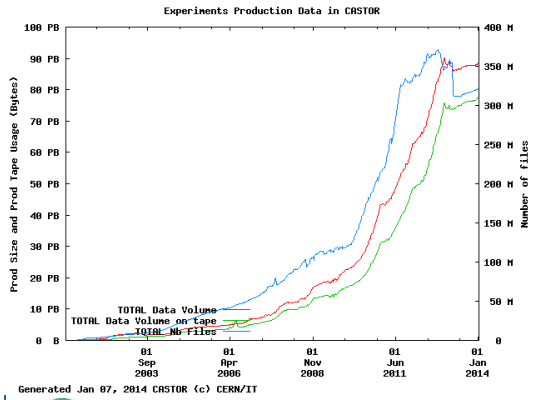
Tier-1 (12 centres)
- Permanent storag
- Re-processing
- Analysis

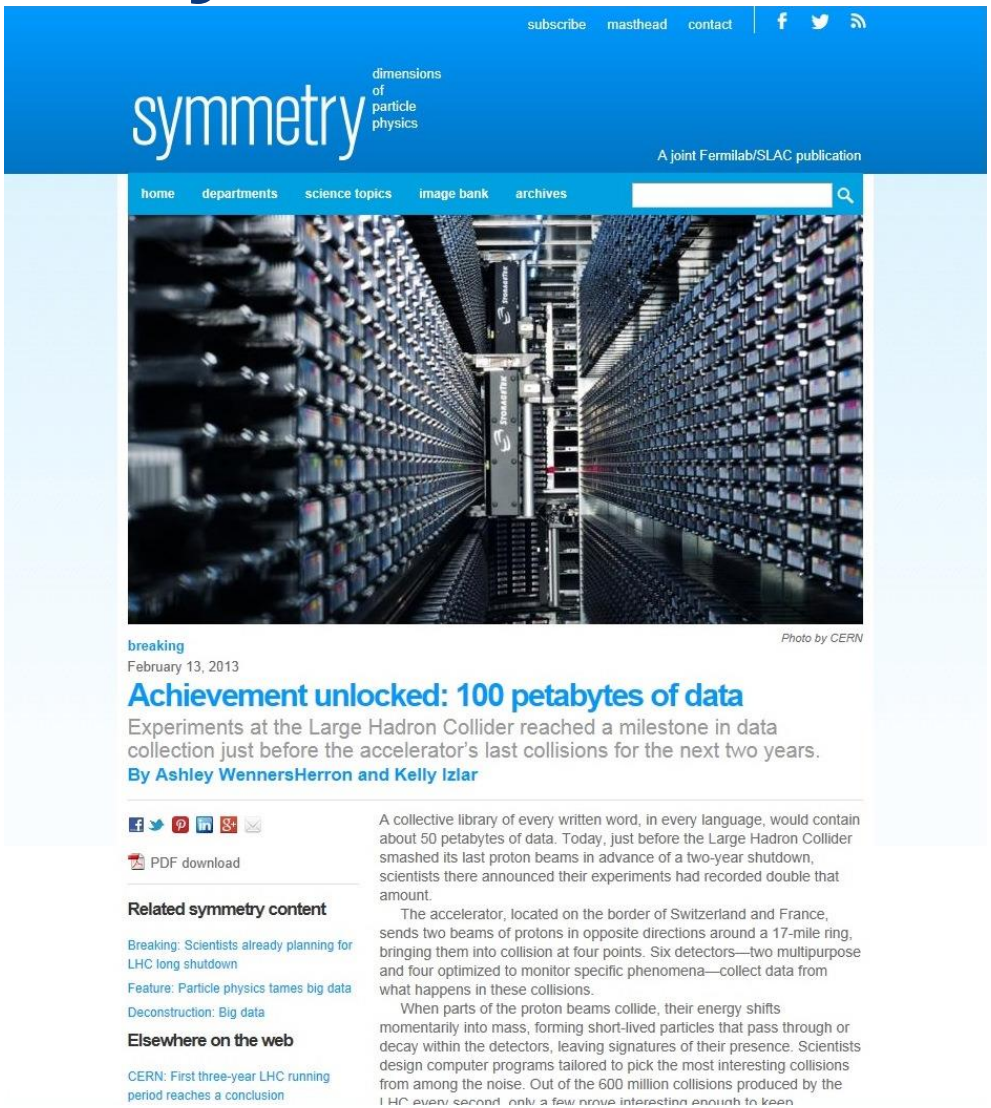
Tier-2 (~130 centres):
- Simulation
- End-user analysis

Managing 100 PBytes of data









subscribe masthead contact

symmetry dimensions of particle physics

A joint Fermilab/SLAC publication

home departments science topics image bank archives

Photo by CERN

breaking

February 13, 2013

Achievement unlocked: 100 petabytes of data

Experiments at the Large Hadron Collider reached a milestone in data collection just before the accelerator's last collisions for the next two years.

By Ashley WennersHerron and Kelly Izlar

PDF download

Related symmetry content

Breaking: Scientists already planning for LHC long shutdown

Feature: Particle physics tames big data

Deconstruction: Big data

Elsewhere on the web

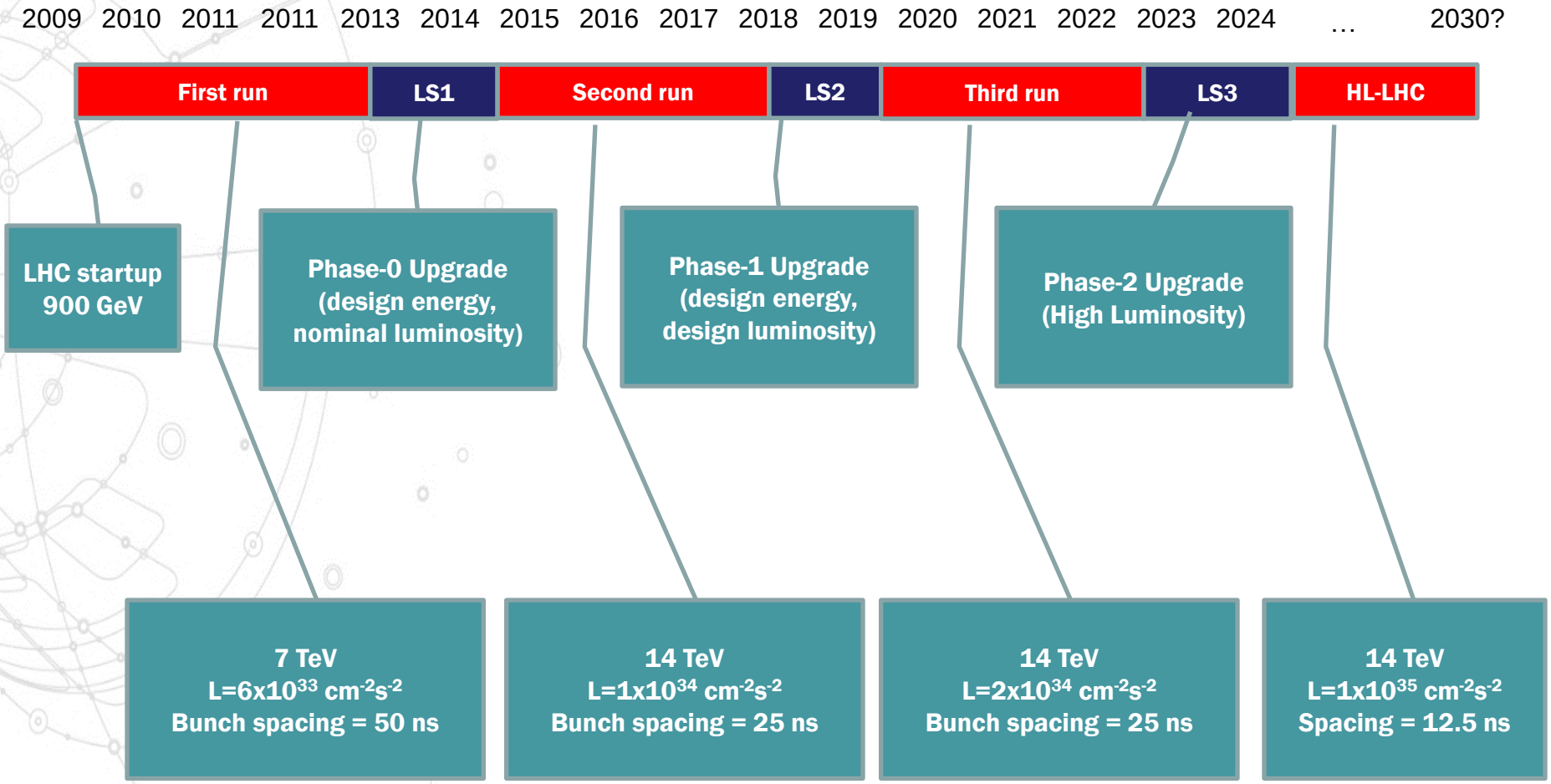
CERN: First three-year LHC running period reaches a conclusion

A collective library of every written word, in every language, would contain about 50 petabytes of data. Today, just before the Large Hadron Collider smashed its last proton beams in advance of a two-year shutdown, scientists there announced their experiments had recorded double that amount.

The accelerator, located on the border of Switzerland and France, sends two beams of protons in opposite directions around a 17-mile ring, bringing them into collision at four points. Six detectors—two multipurpose and four optimized to monitor specific phenomena—collect data from what happens in these collisions.

When parts of the proton beams collide, their energy shifts momentarily into mass, forming short-lived particles that pass through or decay within the detectors, leaving signatures of their presence. Scientists design computer programs tailored to pick the most interesting collisions from among the noise. Out of the 600 million collisions produced by the LHC every second, only a few prove interesting enough to keep

LHC Schedule



CERN openlab in a nutshell

- A science – industry partnership to drive R&D and innovation with over a decade of success
- Evaluate state-of-the-art technologies in a challenging environment and improve them
- Test in a research environment today what will be used in many business sectors tomorrow
- Train next generation of engineers/employees
- Disseminate results and outreach to new audiences

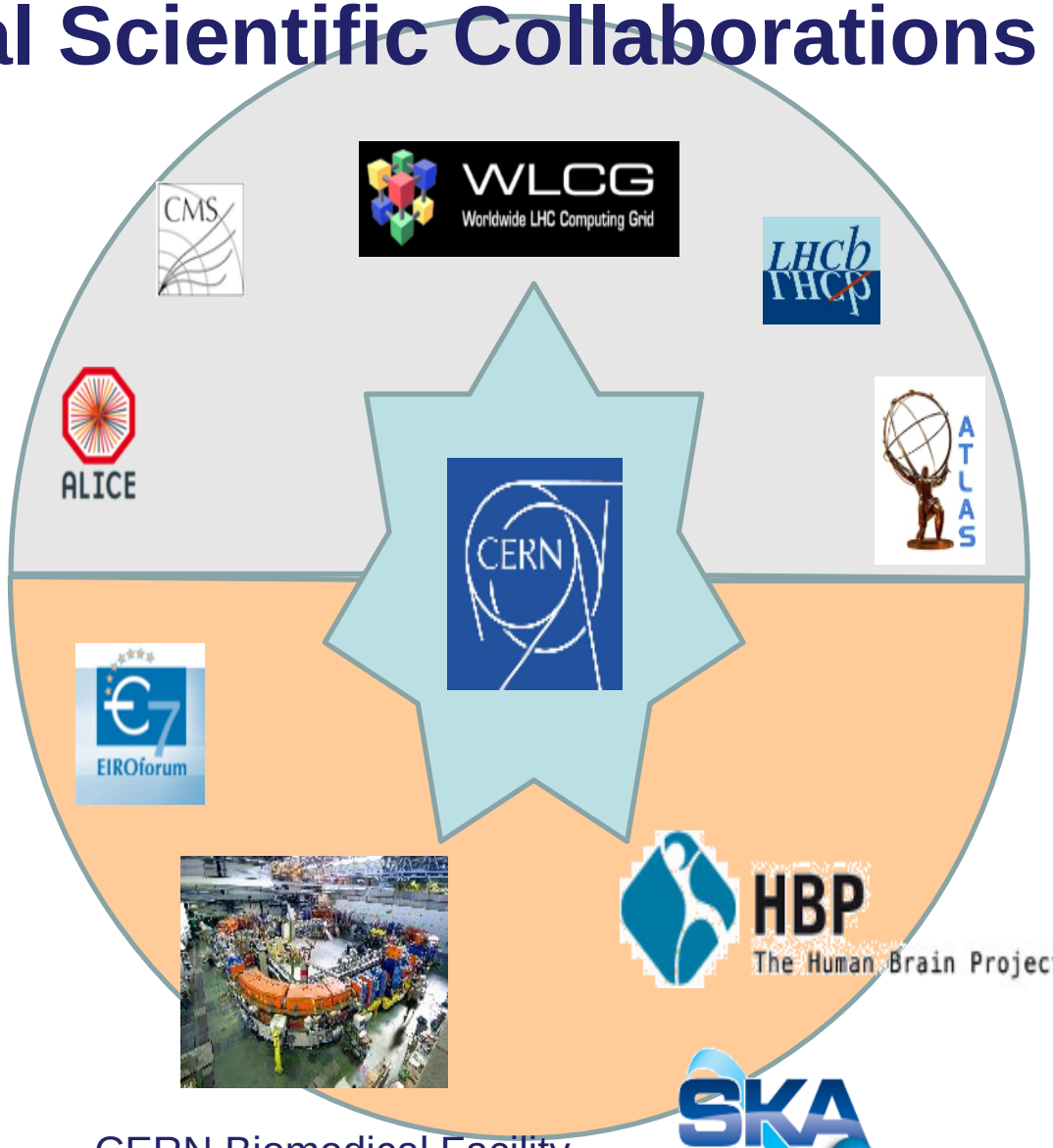
PARTNERS

CONTRIBUTOR

ASSOCIATE

International Scientific Collaborations

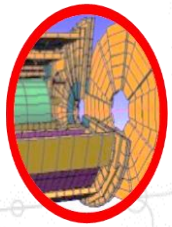
- Many scientific projects are global collaborations of 100s of partners
- Efficient computing and data infrastructures have become critical as the quantity, variety and rates of data generation keep increasing
- Funding does not scale in the same way
  - Optimization and sharing of resources
- Collaboration with commercial IT companies increasingly important
  - Requirements are not unique anymore



CERN Biomedical Facility

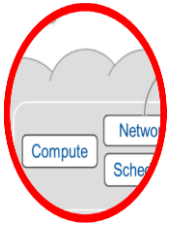
Challenges

Online triggers and DAQ



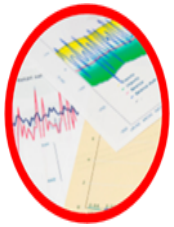
Offline simulation and processing

Data storage architectures



Resource management and provisioning

Networks and connectivity



Data analytics

White paper to be published in February

A European cloud computing partnership: big science teams up with big business

Helix Nebula – The Science Cloud: A catalyst for change in Europe
http://cds.cern.ch/record/1537032

HELIX NEBULA THESCIENCECLOUD

Strategic Plan

- Establish multi-tenant, multi-provider cloud infrastructure
- Identify and adopt policies for trust, security and privacy
- Create governance structure
- Define funding schemes

CERN

To support the computing capacity needs for the ATLAS experiment

EMBL

Setting up a new service to simplify analysis of large genomes, for a deeper insight into evolution and biodiversity

esa DLR cnes

To create an Earth Observation platform, focusing on earthquake and volcano research

PIC port d'informació científica

To improve the speed and quality of research for finding surrogate biomarkers based on brain images

Suppliers

Atos CGI CloudSigma CSA cloud security alliance DANTE www.dante.net egi indra interoute SAP sixsq SWITCH T-Systems terradue 2.0 the SERVER LABS the IT architects

Adopters

AWST Capgemini CONSULTING.TECHNOLOGY.OUTSOURCING cnrs ECMWF Ifremer OpenNebula.org The Open Source Toolkit for Cloud Computing orange Business Services THALES Telefónica Trust-IT Communicating ICT to Markets www.trust-itservices.com

ESFRI Cluster projects

ENVRI

Common Operations of Environmental Research Infrastructures

32 Research Infrastructures

CRISP

The Cluster of Research Infrastructures for Synergies in Physics

DASISH

DATA SERVICE INFRASTRUCTURE FOR THE SOCIAL SCIENCES AND HUMANITIES

Cross-Disciplinary Challenges

A matrix showing the interest in common topics for the four cluster initiatives

	CRISP	ENVRI	DASISH	BioMed
Data identity				
Semantic annotations and bridging				
Reference models				
Education & training				

zenodo

Research. Shared.

Search Communities Browse Upload Get started Sign in

Home / Publications / Realising the full potential of research data: common challenges in data management, sharing and integration across scientific disciplines

20 December 2013

Working paper Open access

Realising the full potential of research data: common challenges in data management, sharing and integration across scientific disciplines

Field,Laurence ; Suhr, Stephanie ; Ison, Jon ; Los, Wouter ; Wittenburg, Peter ; Broeder, Daan ; Hardisty, Alex ; Repo, Susanna ; Jenkinson, Andy

(show affliations)

Publication date:
20 December 2013

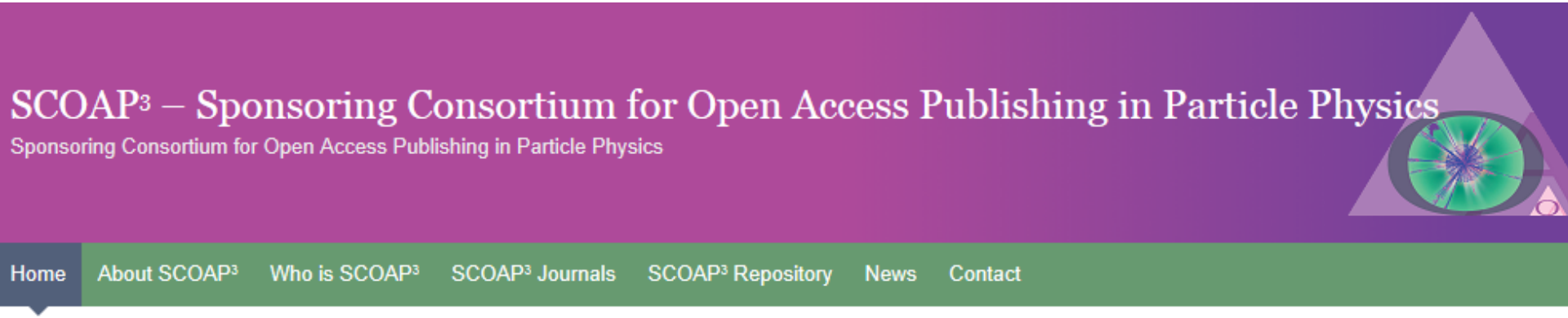
DOI:
10.5281/zenodo.7636

Collections:
Publications > Working papers
Communities
Open Access

License (for files):
Creative Commons Attribution

Uploaded by:
lfield (on 20 December 2013)

Open Access



http://scoap3.org/

Repository for Research Results

zenodo

Research. Shared.

Zenodo, is the child of the OpenAir initiative, a European portal for open access research

http://zenodo.org/

Zenodo is based on Invenio open-source technology developed by CERN and a growing developer community

Invenio is used by many organisations around the world

http://invenio-software.org/

- **Research. Shared.** – all research outputs from across all fields of science are welcome!
- **Citeable. Discoverable.** – uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- **Community Collections** – accept or reject uploads to your own community collections (e.g workshops, EU projects or your complete own digital repository).
- **Funding** – integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- **Flexible licensing** – because not everything is under Creative Commons.
- **Safe** – your research output is stored safely for the future in same cloud infrastructure as research data from CERN's Large Hadron Collider.
- **DropBox integration** – upload files straight from your DropBox.

Powered by:

INVENIO

OpenAIRE

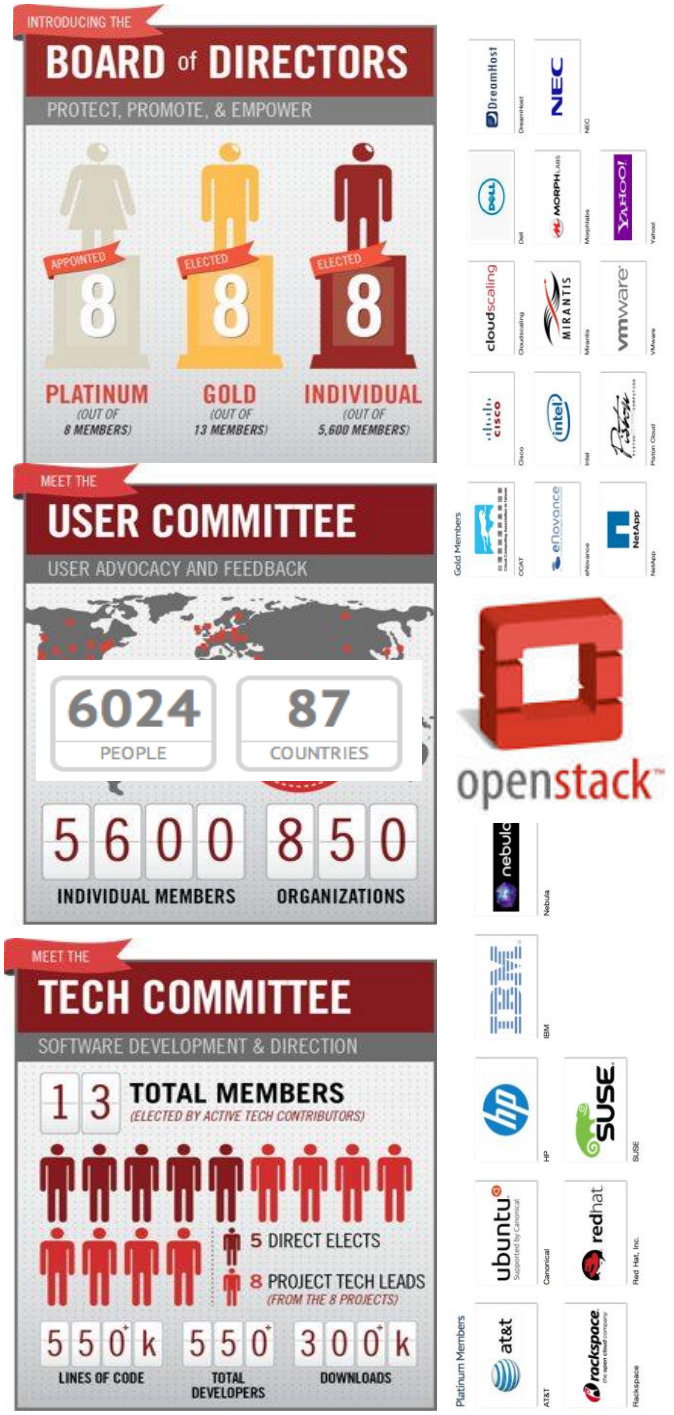
CERN

SEVENTH FRAMEWORK PROGRAMME

e-infrastructure

Cloud Computing

- CERN deploys a large scale Infrastructure-as-a-Service cloud
  - Currently ~50,000 cores across 2 data centres (Geneva & Budapest)
  - Expect to grow >150,000 cores by 2015
- Based on OpenStack
  - Adopt open source tools with sustainable communities used by other organisations
  - CERN has an seat on the management board and chairs the user committee
- Federation of clouds could give multi-site resource sharing



E-Infrastructure for the 21st Century

- The goal is to transform existing Distributed Computing Infrastructures (DCIs) based on a range of technologies into a *service-oriented platform* for the *global research community* that can be *sustained* through *innovative business models*

- Prepared by CERN on behalf of the EIROforum IT Working Group

DOI:10.5281/zenodo.7592

Summary

- CERN and the LHC program have been among the first to address “big data” challenges
- Solutions have been developed and important results obtained
- Now preparing for future needs in common with many scientific and business domains
- Need to exploit emerging technologies and share expertise with academia and commercial partners
- LHC schedule will ensure CERN stays at the bleeding edge, providing excellent opportunities to test ideas, technologies and organisational models ahead of the market

www.cern.ch