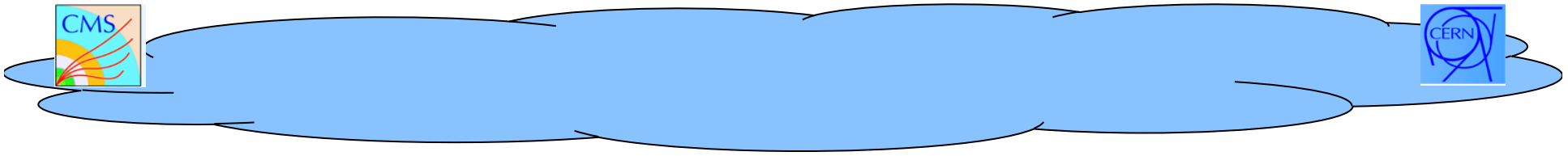


Overlay Opportunistic Clouds in CMS/ATLAS at CERN: The CMSooooooCloud in Detail

J.A. Coarasa

and

Wojciech Ozga

CERN, Geneva, Switzerland

OpenStack Summit, 15-18 April 2013,
Portland, USA

Outline

- Introduction
  - CERN
  - The Large Hadron Collider (LHC)
  - Alice, ATLAS, CMS, LHCb
    - The CMS experiment/cluster as an example (data path)
- The online cloud architectures
  - Requirements
  - Overlay in detail
    - Open vSwitch-ed
    - OpenStack infrastructure
- Onset, operation, outlook and conclusions

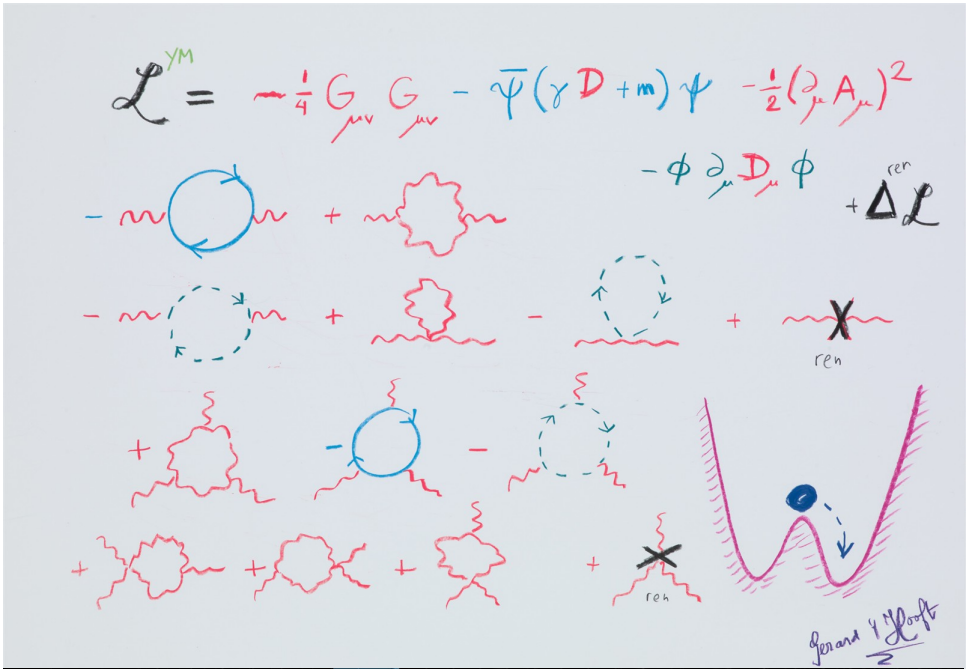
CERN

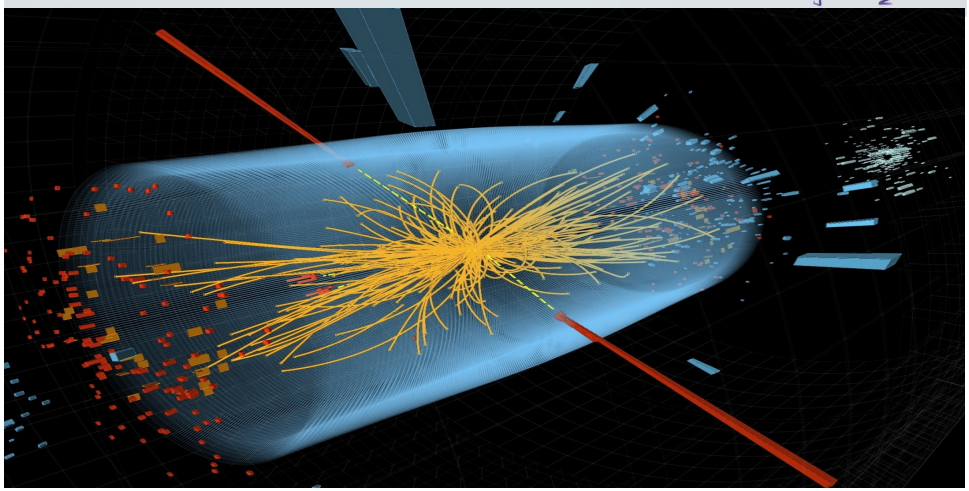
- European Organization for Nuclear Research
  (Conseil Européen pour la Recherche Nucléaire)
- An international organization founded in 1954
- The world's largest particle physics laboratory

http://home.web.cern.ch/about

CERN Mission

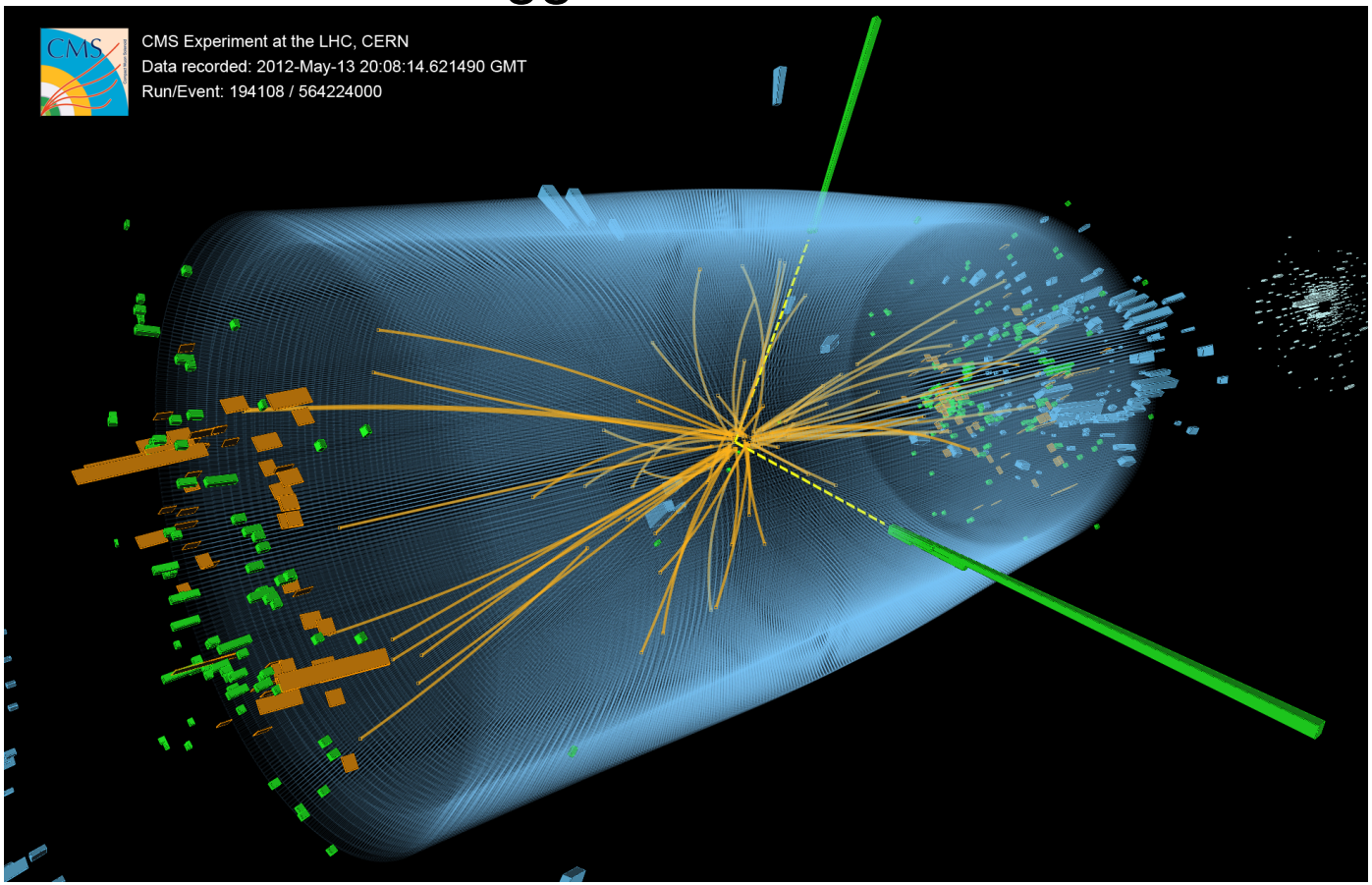
- Fundamental research.
  Answering fundamental questions:
  - What is the universe made of? How did it start?...
- Bringing nations together:
  - 20 European state members;
  - Users of more than 100 nations.
- Training Scientists and Engineers.





The ATLAS/CMS Achievement

Observation of a new particle with a mass of 125 GeV consistent with the Higgs boson of the Standard Model†

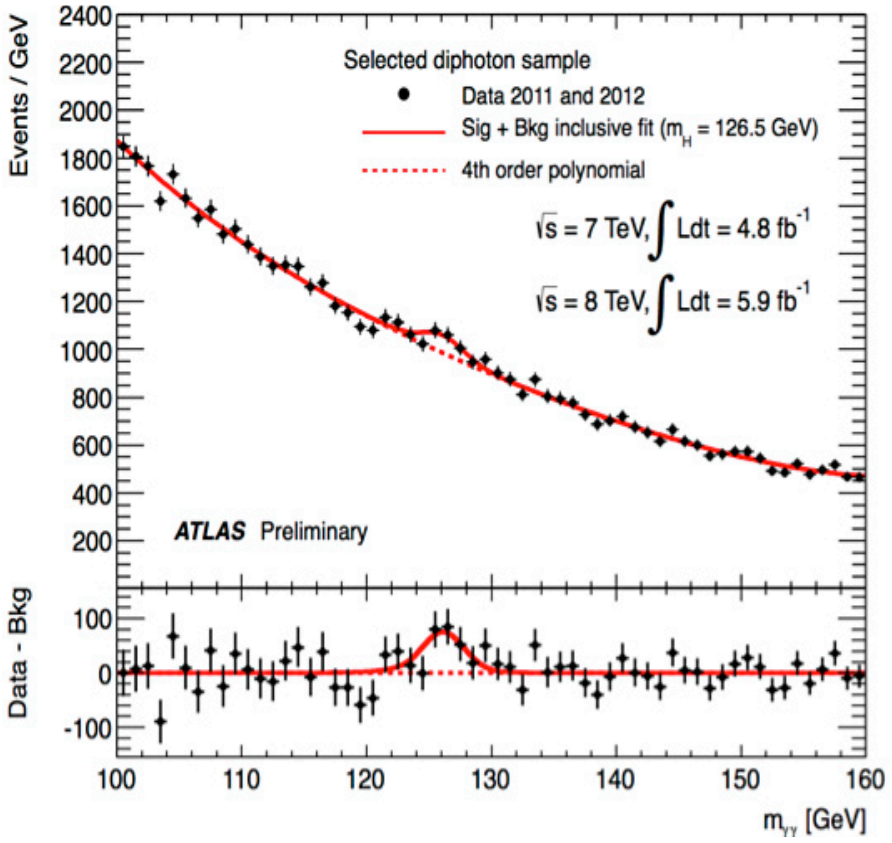


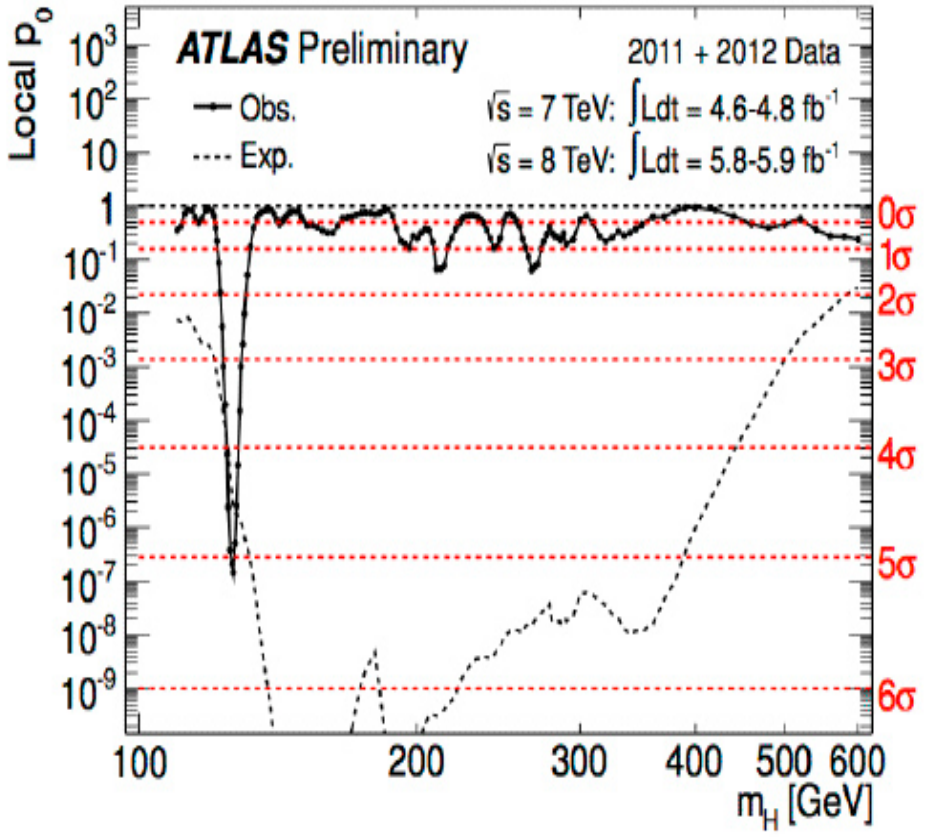
† http://press.web.cern.ch/press-releases/2012/07/cern-experiments-observe-particle-consistent-long-sought-higgs-boson

http://cms.web.cern.ch/news/observation-new-particle-mass-125-gev

The ATLAS/CMS Achievement

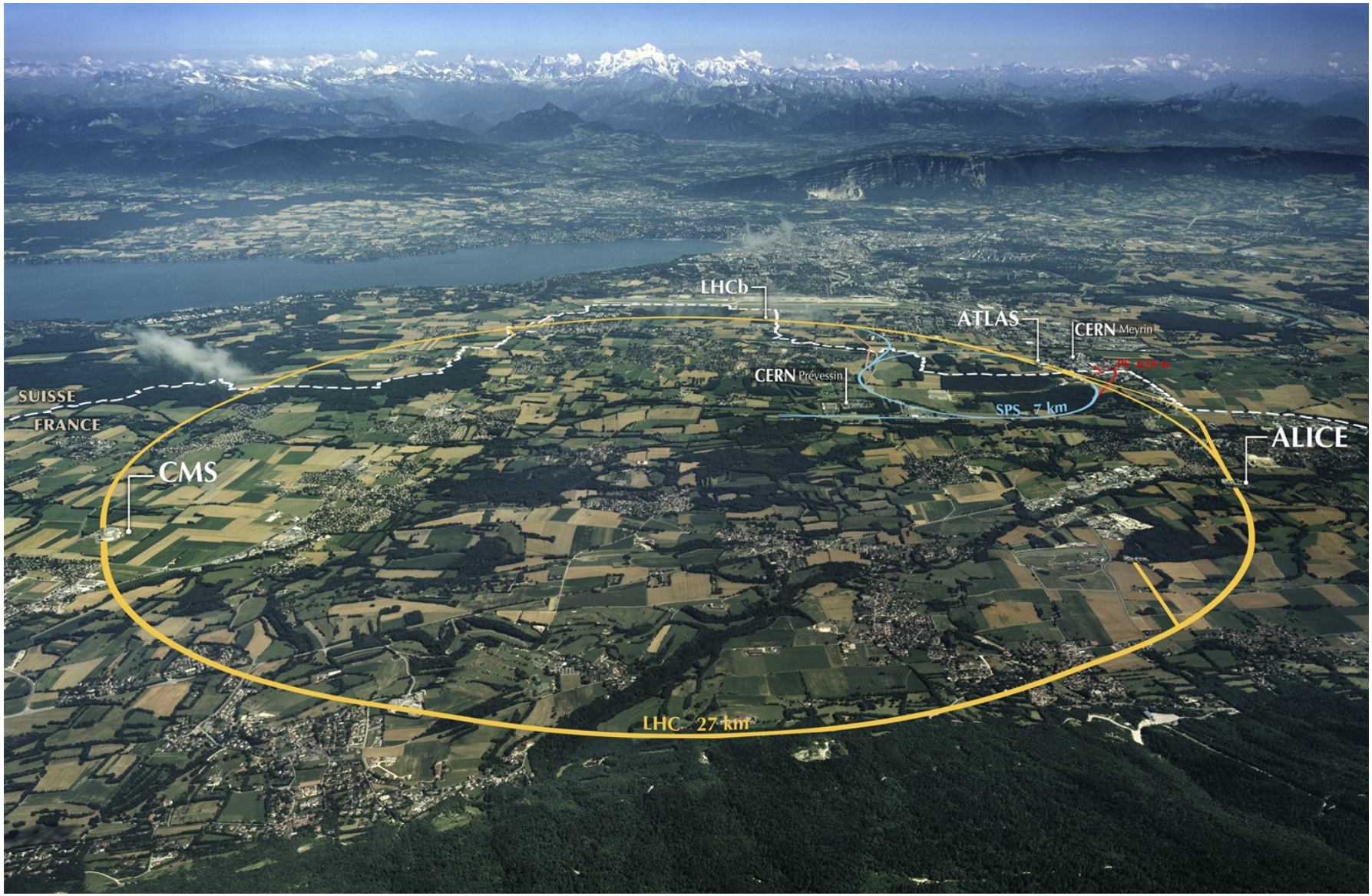
Observation of a new particle with a mass of 125 GeV consistent with the Higgs boson of the Standard Model†



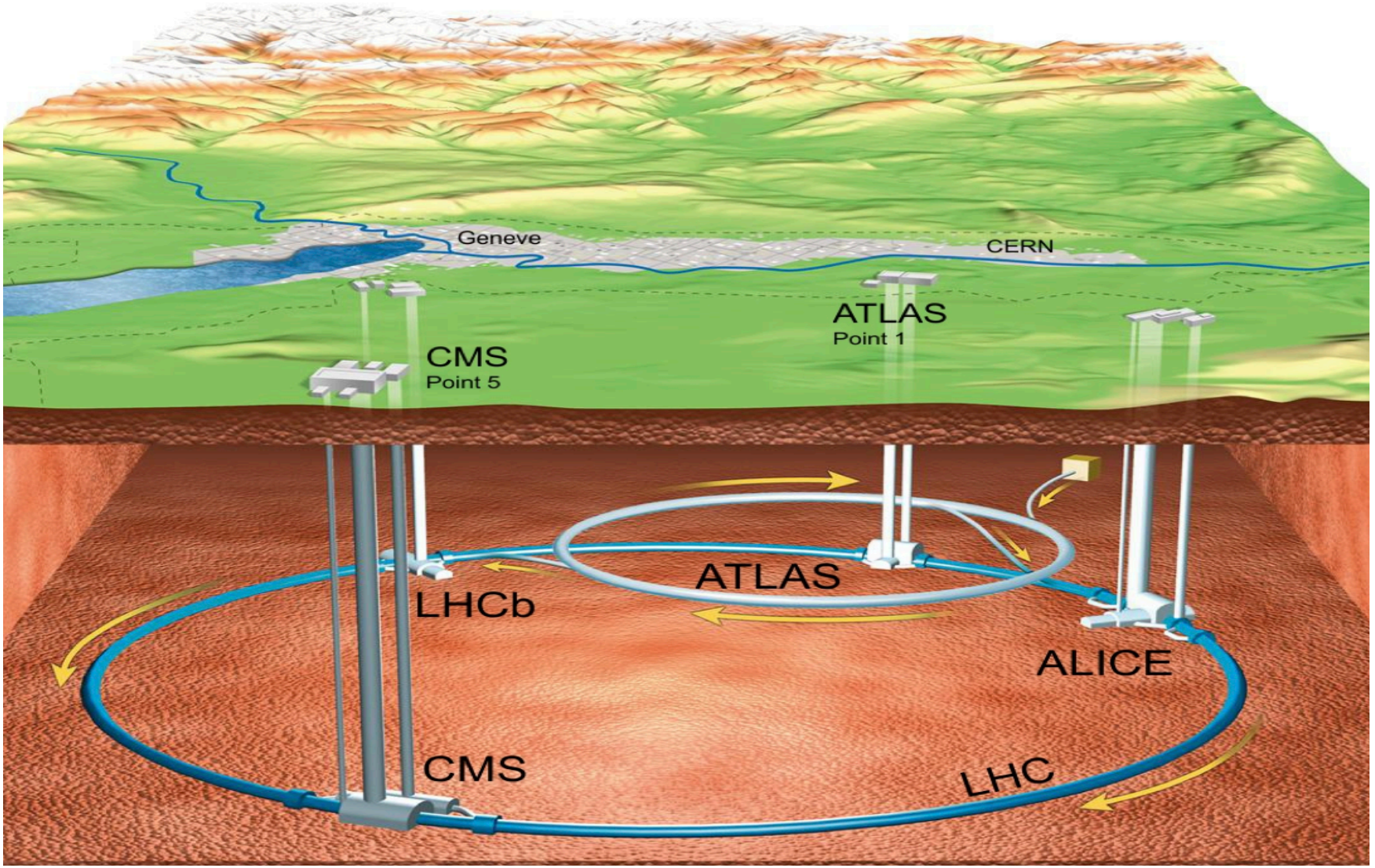


† http://press.web.cern.ch/press-releases/2012/07/cern-experiments-observe-particle-consistent-long-sought-higgs-boson
http://www.atlas.ch/news/2012/latest-results-from-higgs-search.html

The Large Hadron Collider (LHC)



The Large Hadron Collider (LHC)

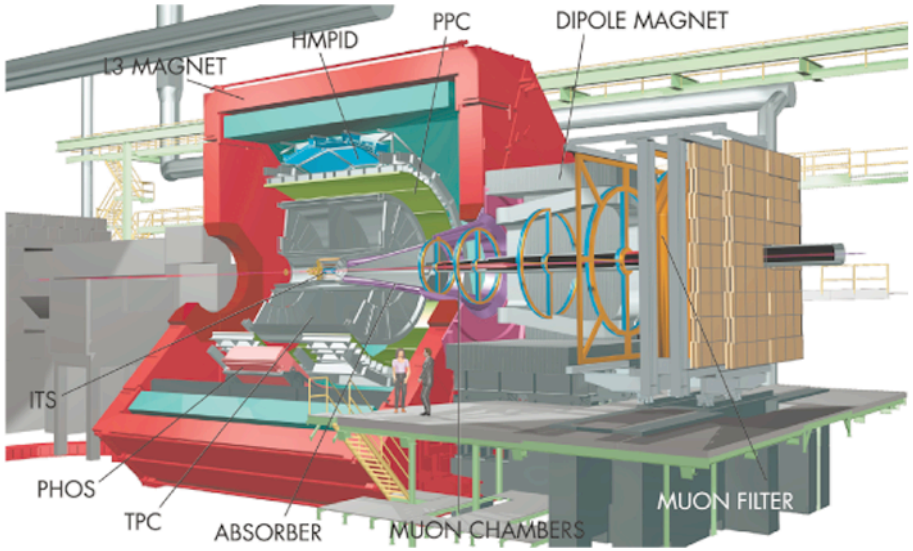


LHC Specific Experiments

ALICE

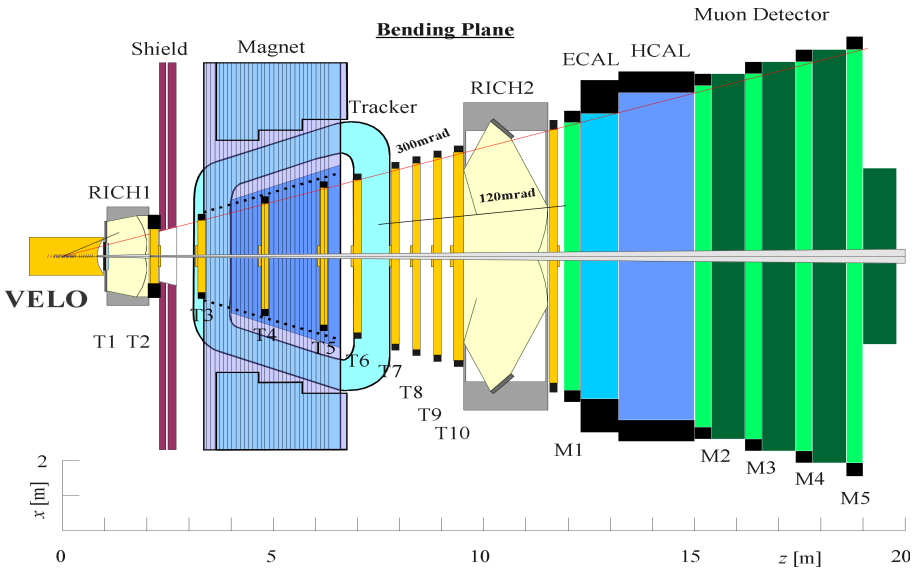
A Large Ion Collider Experiment

The ALICE Collaboration built a dedicated heavy-ion detector to study the physics of strongly interacting matter at extreme energy densities, where the formation of a new phase of matter, the quark-gluon plasma, is expected.



LHCb

Study of CP violation in B-meson decays at the LHC collider

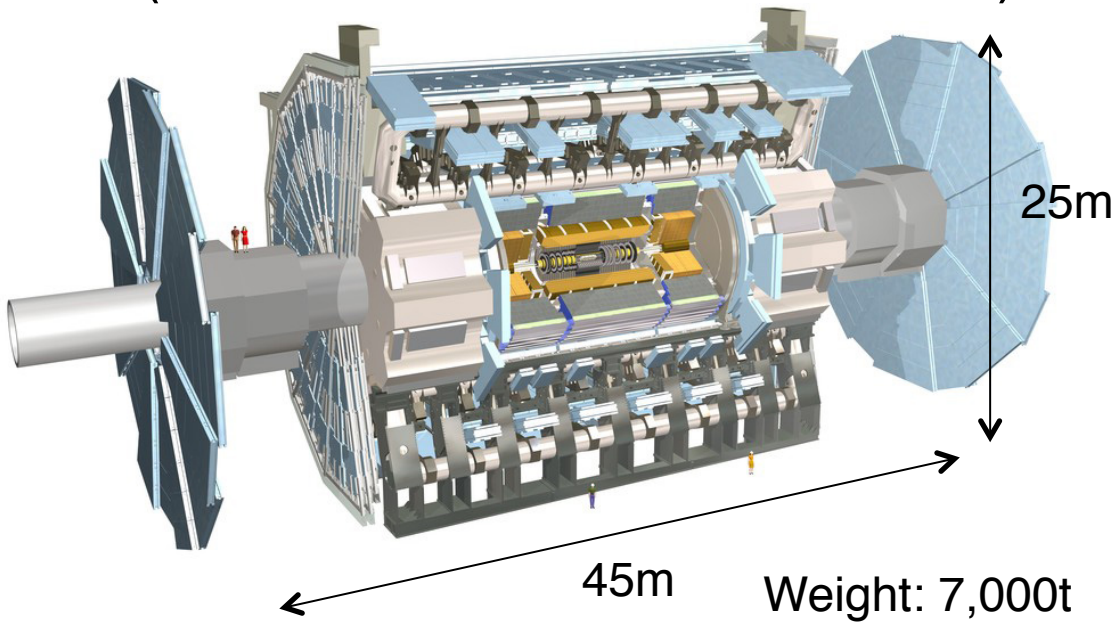


LHC Multipurpose Experiments

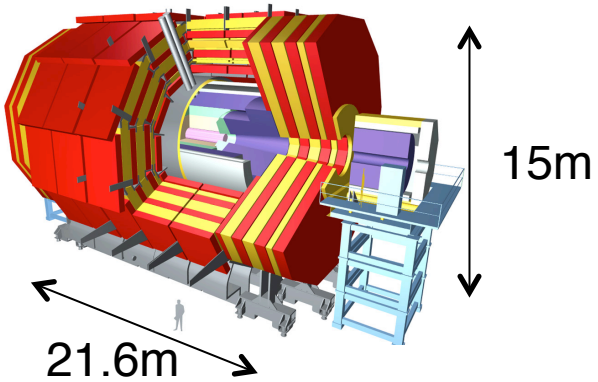
ATLAS
A Toroidal LHC ApparatuS

CMS
Compact Muon Solenoid

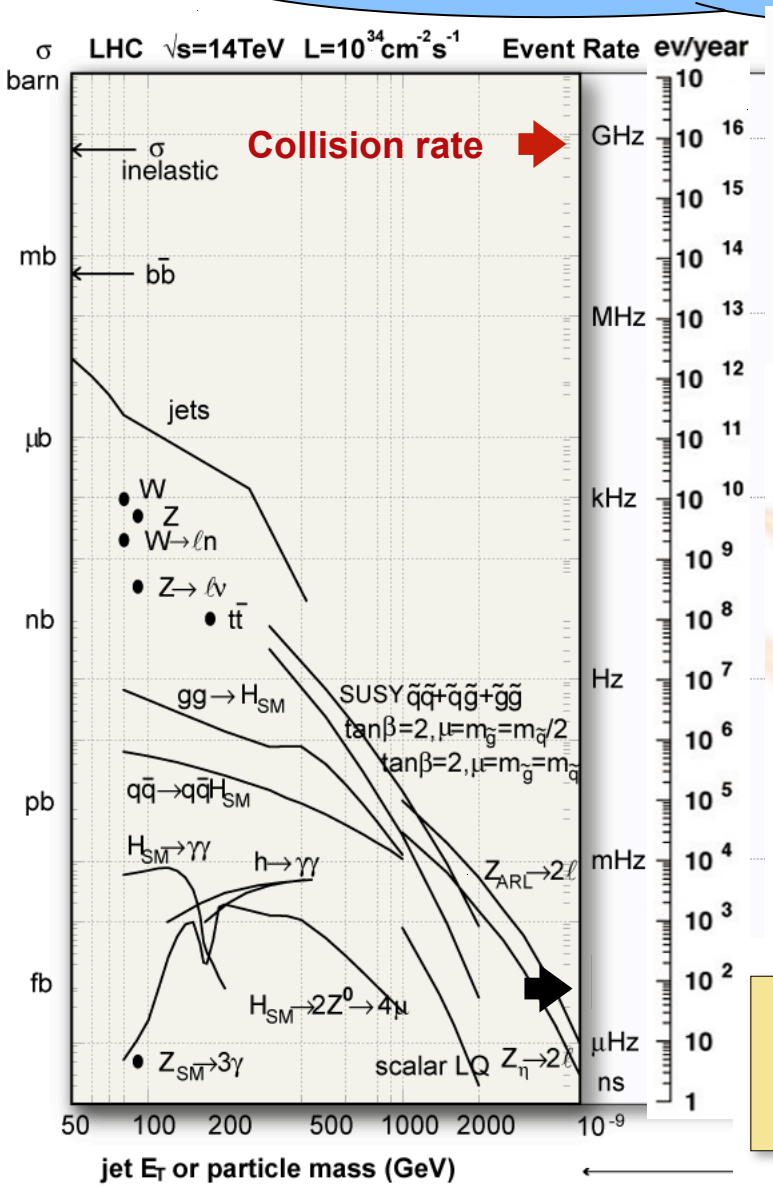
- They are multistorey-high digital cameras recording hundreds of images per second of debris from LHC particle collision.
- The collaborations have ~4000 active members each (~180 institutes, ~40 countries).

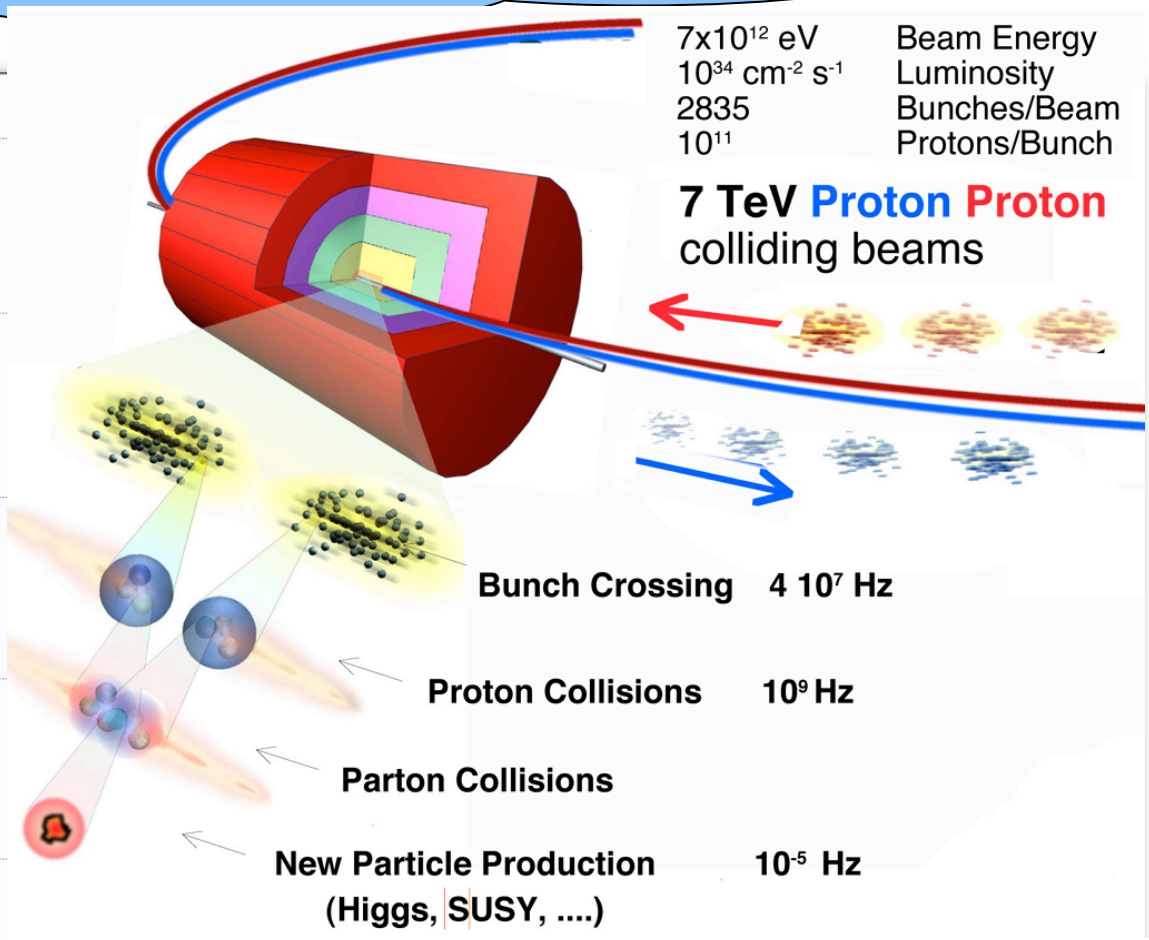


Magnetic Field: 3.8T



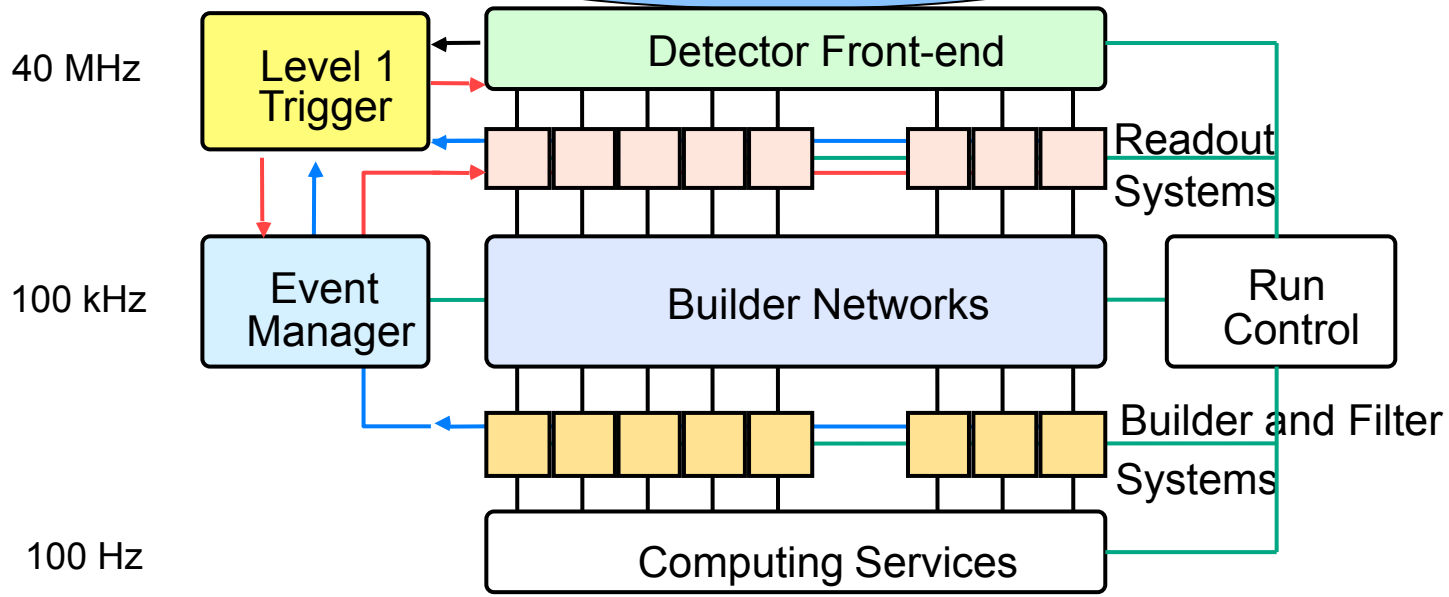
CMS: Collisions Overview





Event rates: 1 billion per second
Event selection: ~1/10^13

CMS: Data Origin, The DAQ



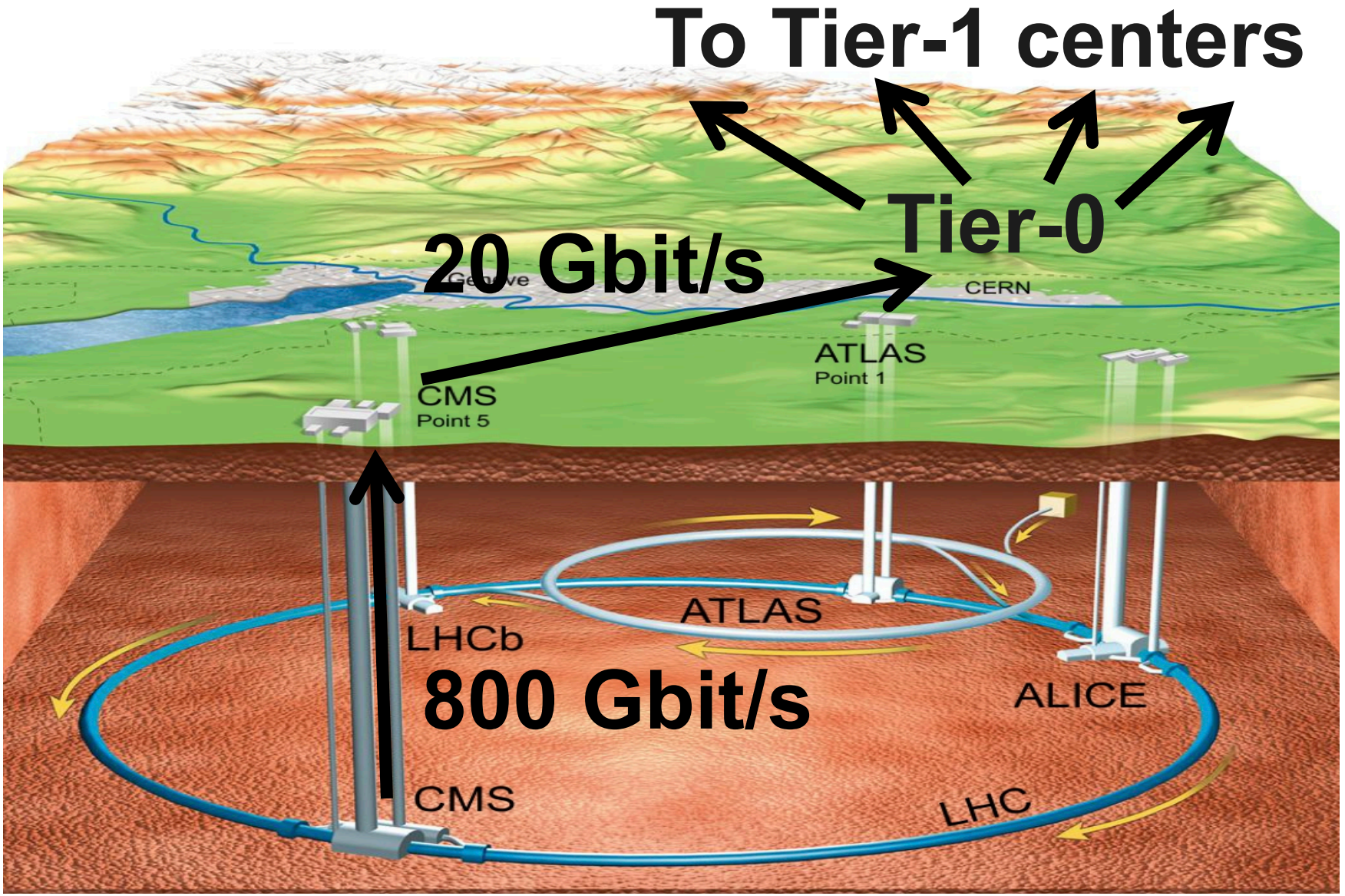
Large Data Volumes **(~100 Gbytes/s data flow, 20TB/day)**

- After Level 1 Trigger ~100 Gbytes/s (rate ~O(100) kHz) reach the event building **(2 stages, ~2000 computers).**
- High Level Trigger (HLT) filter cluster select 1 out 1000. Max. rate to tape: ~O(100) Hz

⇒ **The storage manager (stores and forwards) can sustain a 2GB/s traffic.**

⇒ **Up to ~300 Mbytes/s sustained forwarded to the CERN T0. (>20TB/day).**

CMS: The Data Path



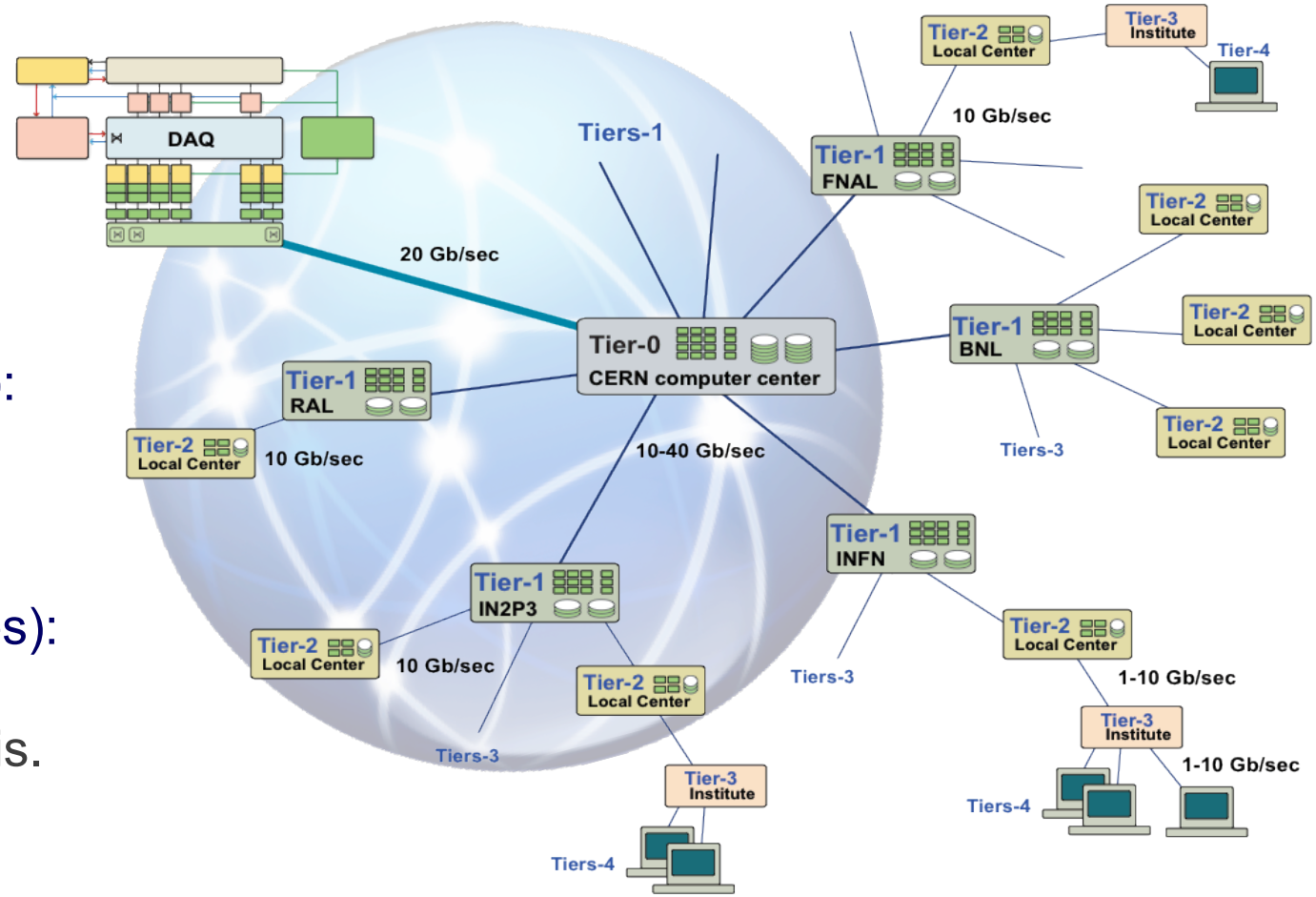
LHC Offline Computing. The GRID

The GRID. A distributed computing infrastructure (~150 kCores), uniting resources of HEP institutes around the world to provide seamless access to CPU and storage for the LHC experiments and other communities

Tier-0 (CERN):
- Recording;
- Reconstruction;
- Distribution.

Tier-1 (~10 centres):
- Storage;
- Reprocessing;
- Analysis.

Tier-2 (~200 centres):
- Simulation;
- End-user analysis.



The ATLAS/CMS Online Clusters

Large

- More than 3000 computers
- More than 100 switches (several thousand ports)

and complex clusters†

- Different kinds of hardware and ages
- Computers configured in more than 100 ways
- Set of segmented networks using sometimes VLANs
  - » 1 network per rack
  - » Up to 2 Additional networks in VLANs in some racks

designed for data taking

- shuffle data at 100GBytes/s and select and archive 20TBytes/day

†*The CMS online cluster: Setup, operation and maintenance of an evolving cluster.* **PoS ISGC2012 (2012) 023**.

The CMS Online Cluster: Details

More than 3000 computers mostly under *Scientific Linux CERN*:

High bandwidth networking

- 640 (4-core) as a 1st stage building, equipped with 2 Myrinet and 3 independent 1 Gbit Ethernet lines for data networking. (2560 cores);
- 1264 (720 (8-core) + 288 (12-core allowing HT)) + 256 (16-core allowing HT) as high level trigger computers with 1 or 2 Gbit Ethernet lines for data networking. (13312 cores);
- 16 (2-core) with access to 300 TBytes of FC storage, 4 Gbit Ethernet lines for data networking and 2 additional ones for networking to Tier 0;

The CMS Online Cluster: Details

More than 3000 computers mostly under *Scientific Linux CERN*:

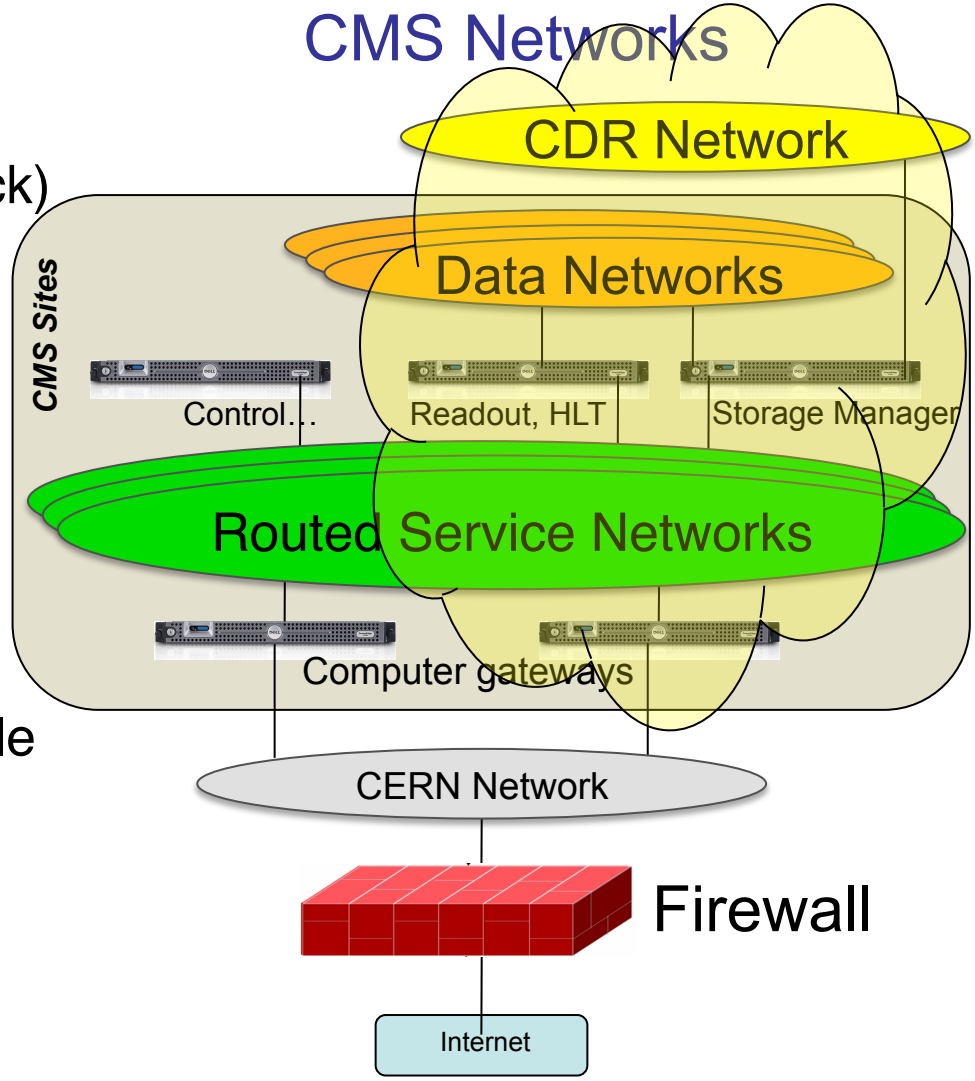
High bandwidth networking

- 640 (4-core) as a 1st stage building, equipped with 2 Myrinet and 3 independent 1 Gbit Ethernet lines for data networking. (2560 cores);
- 1264 (720 (8-core) + 288 (12-core allowing HT)) + 256 (16-core allowing HT) as high level trigger computers with 1 or 2 Gbit Ethernet lines for data networking. (13312 cores);
- 16 (2-core) with access to 300 TBytes of FC storage, 4 Gbit Ethernet lines for data networking and 2 additional ones for networking to Tier 0;
- More than 400 used by the subdetectors;
- 90 running Windows for Detector Control Systems;
- 12 computers as an ORACLE RAC;
- 12 computers as CMS control computers;
- 50 computers as desktop computers in the control rooms;
- 200 computers for commissioning, integration and testing;
- 15 computers as infrastructure and access servers;
- 250 active spare computers.

The CMS Online Cluster: Network Details

CMS Networks:

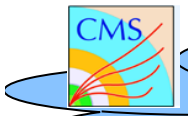
- Private Networks:
  - Service Networks (per rack) (~3000 1 Gbit ports);
  - Data Network (~4000 1Gbit ports)
    - Source routing on computers;
    - VLANs on switches.
  - Central Data Recording (CDR). Network to Tier 0;
  - Private networks for Oracle RAC;
  - Private networks for subdetector.
- Public CERN Network.



The ATLAS/CMS High Level Trigger (HLT) Clusters

	cluster	Nodes	Cores (HT on)/ node	Memory (Gbyte)/node	cores	Disk (Gbytes)
ATLAS	(1)	341	8	16	2728	72
	(2)	320	8 (16)	24	2560	225
	(3)	832	12 (24)	24	9984	451
	(1)+(2)+(3)	**1493**	**13312**	**33 TBytes**	**15272**	**277 Tbytes**
CMS	(1)	720	8	16	5760	72
	(2)	288	12 (24)	24	3456	225
	(3)	256	16 (32)	32	4096	451
	(1)+(2)+(3)	**1264**	**13312**	**26 TBytes**	**13312**	**227 TBytes**

- Three generations of hardware, some with limited local storage.



The ATLAS/CMS HLT Clusters versus Tier[0,1,2]

CPU in HEP-SPEC06†

	HLT farm	Tier0	Tier1	Tier2
sum	602k + ALICE	356k	603k	985k
ATLAS	197k	111k	260k	396k
CMS	195k	121k	150k	399k
ALICE		90k	101k	143k
LHCb	210k	34k	92k	47k

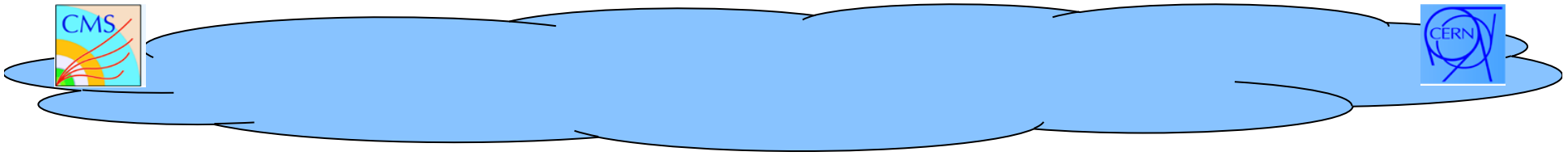
† http://w3.hepix.org/benchmarks/doku.php/

HEP-SPEC06 is based on the all_cpp benchmark subset (bset) of the widely used, industry standard SPEC® CPU2006 benchmark suite. This subset matches the percentage of floating point operations observed in batch jobs (~10%), and it scales perfectly with the experiment codes.

Opportunistic Usage: The Kick-off

The clusters can be used when not 100% in use:

- During the technical stops (~1 week every 10):
  - These timeslots already used by CMS during the past 8 months to set up the cloud infrastructure and test it;
- During the shutdown used to upgrade the accelerator (since Mid-February for more than a year);
- During inter-fill or when the cluster is under-used:
  - Already used while taking data for Heavy Ions this year by CMS;
  - This needs cautious testing and deeper integration;
  - Technically feasible but may arise concerns about putting in danger data taking.



The Online Cloud Architectures

- Requirements
- Overlay in detail
  - Open vSwitch-ed
  - OpenStack infrastructure

Requirements

- Opportunistic usage (the trigger idea! Use the resources!).
- No impact on data taking.
- Online groups retain full control of resources.
- Clear responsibility boundaries among online/offline groups.

⇒Virtualization/Cloud

Overlay on HLT nodes

Overlay! Not dedicated clouds. The HLT Nodes add software to be compute nodes.

- **Minimal changes** to convert HLT nodes in compute nodes participating in the cloud.

  ⇒not to have any impact on data taking

  Losing 1 min of data is wasting accelerator time (worth ~O(1000)CHF/min)

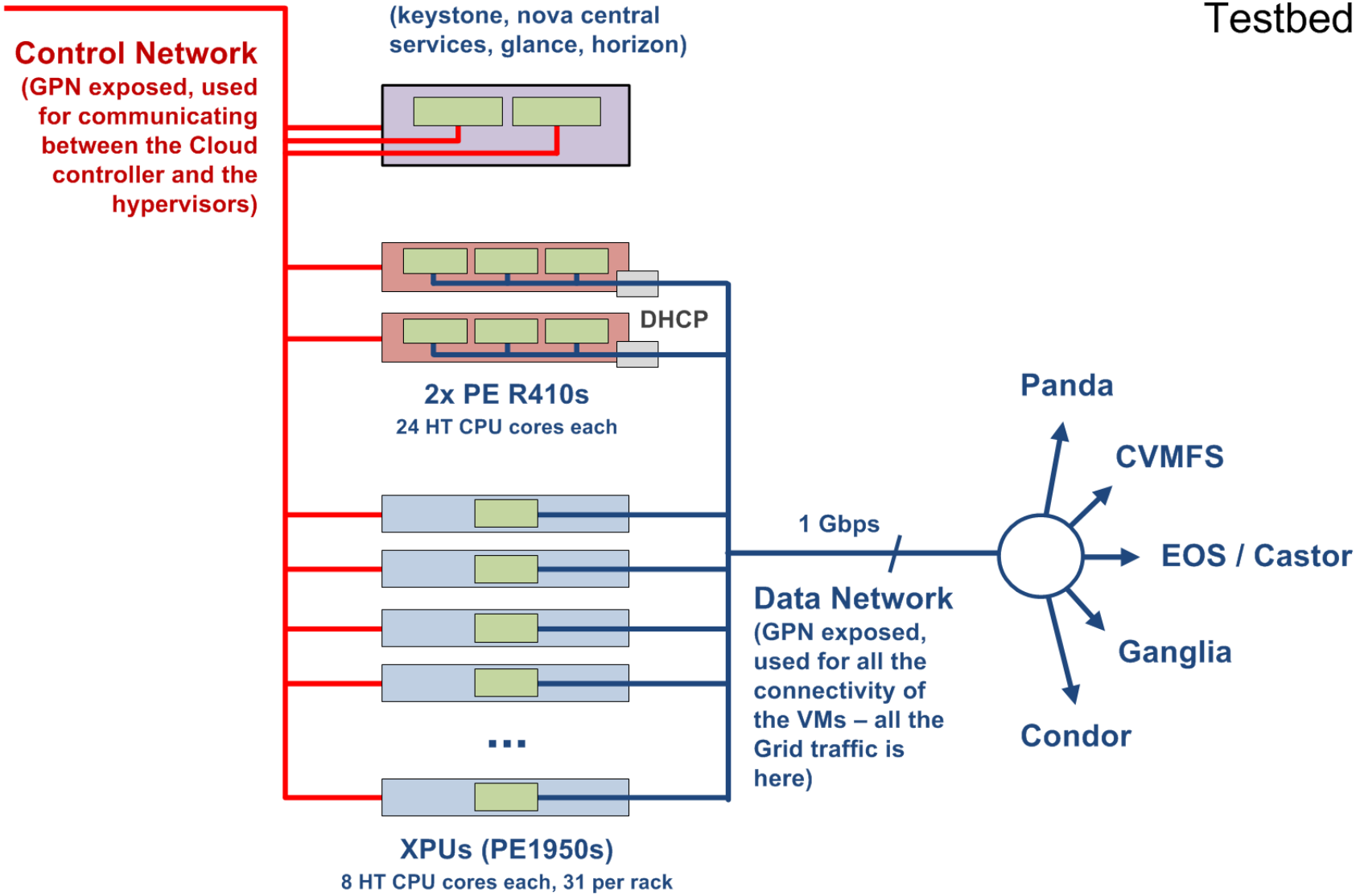
- Easy to quickly move resources between data taking and cloud usage.

Overlay on ATLAS HLT nodes

- Networking
  - Dedicated VLAN added with additional 1Gbit/s networking for the traffic;
  - ACLs to avoid traffic mixing.
- HLT hosts (netbooted)
  - Initial phase: different netbooted image for Cloud, including KVM and OpenStack;
  - Second phase: equal netbooted image for Cloud and data taking (including KVM and OpenStack).

The ATLAS HLT Cloud Testbed

Lab4: Cloud Testbed



Overlay on CMS HLT nodes

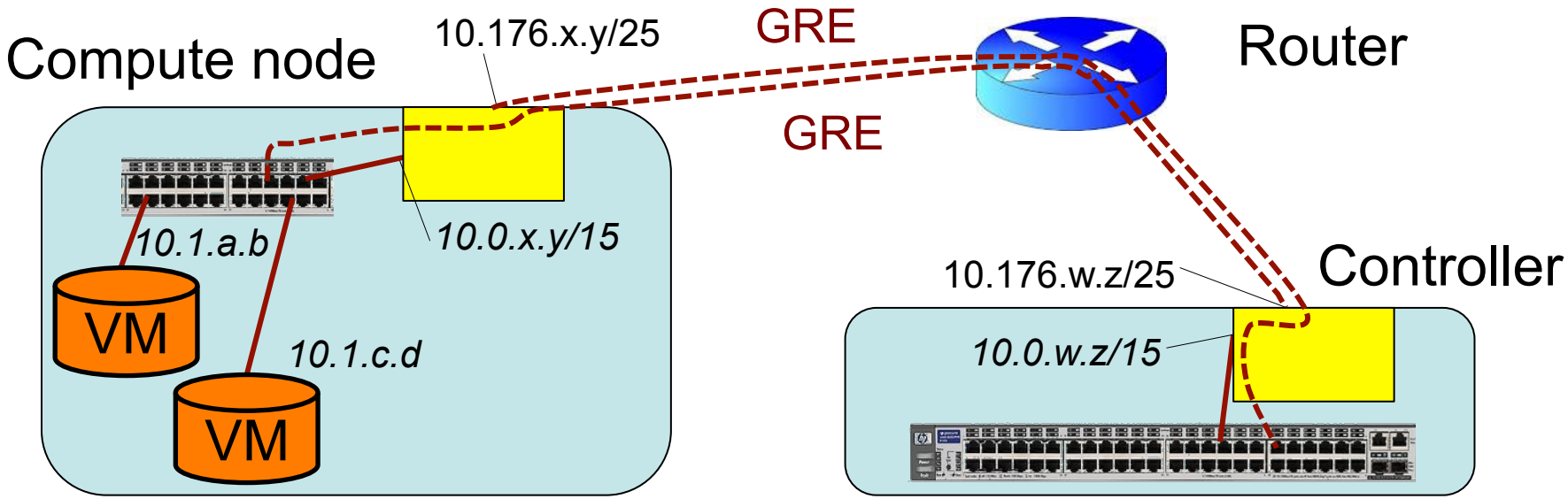
- Networking
  - A virtual switch added in the computer
- Software added
  - Libvirt, kvm
  - OpenStack†1 compute (Essex from EPEL-RHEL 6)
  - Open vSwitch†2 (version 1.7.0-1)
  - RabbitMQ and MySQL clients
  - Home made scripts
    - Configure all components and virtual network
    - Clean up leftovers if necessary
      - » VMs, image files, hooks to bridge…

†1 http://www.openstack.org

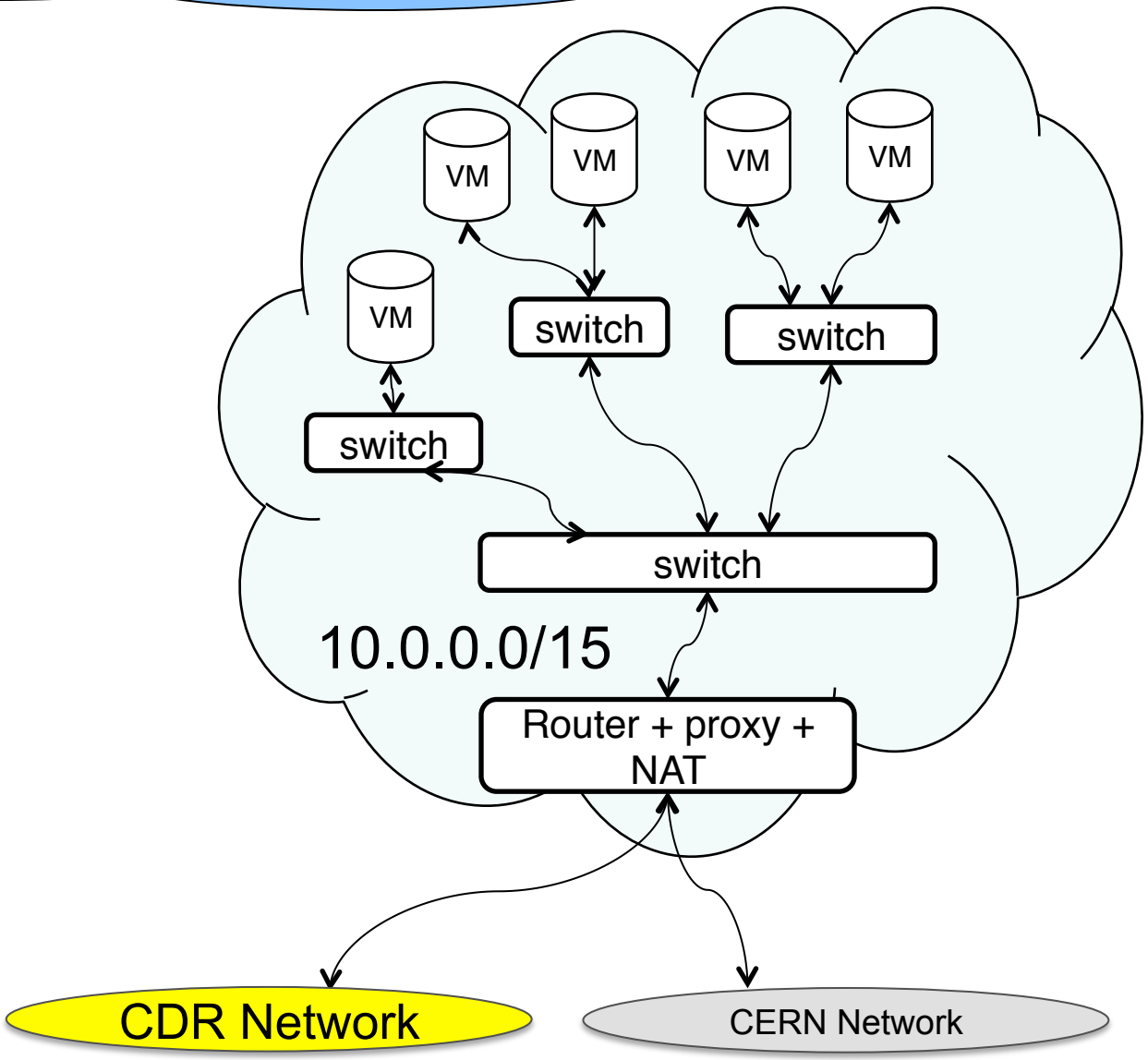
†2 http://openvswitch.org

The CMS Overlay/Virtual Network in Detail in the Proof of Concept Phase: Open vSwitch

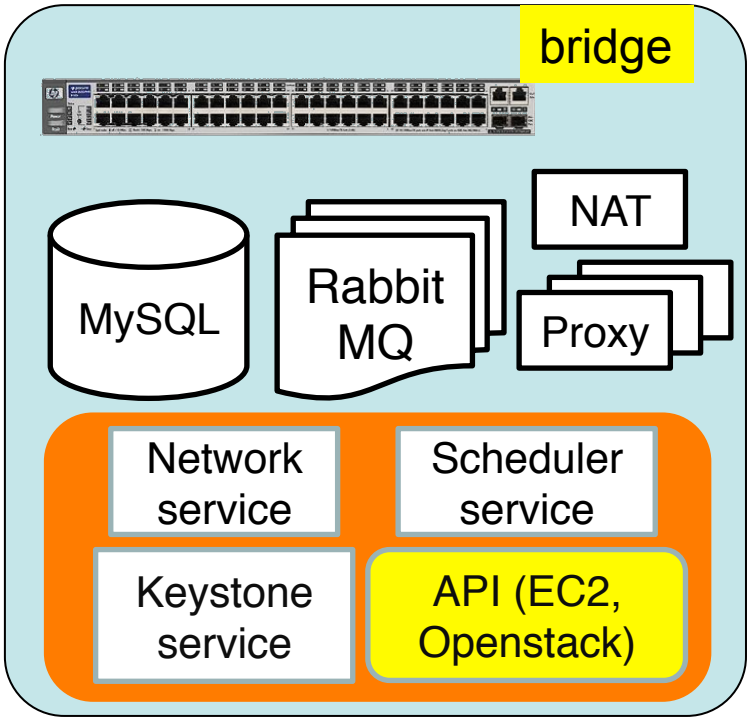
- Switches/Router NOT modified
- Compute nodes have a virtual switch
  - Where VMs hook up using 10.1.0.0/16 **(Flat network for VMs!)**
  - Where the compute node has a virtual interface (10.0.x.y/15) and traffic is routed to the control network (10.176.x.y/25)
  - And a port is connected to a central computer control network IP encapsulating with GRE
- A central computer acts as a big switch (potential bottleneck)
  - With *reciprocating GRE ports*

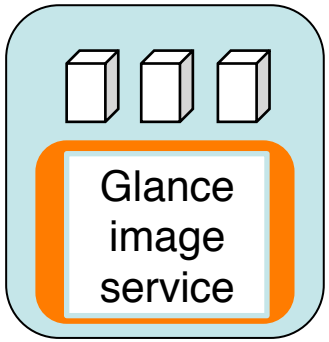


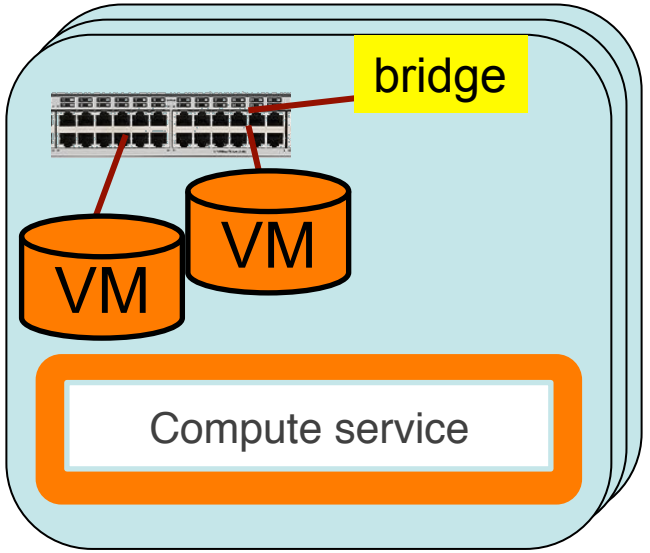
CMS cloud: the Flat Network



The CMS Cloud Controlling Layer: OpenStack Infrastructure







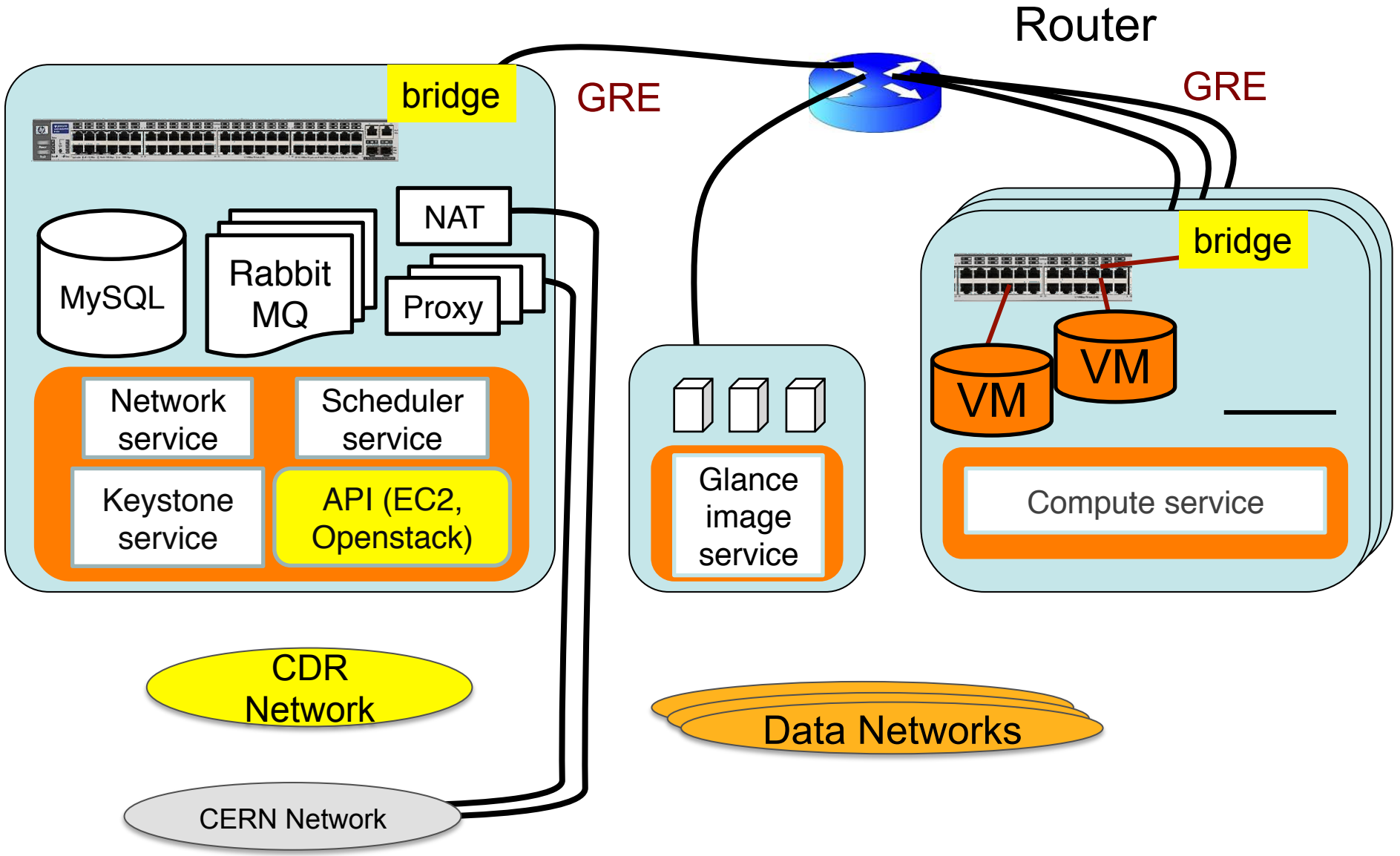
1xFat “controller” node
(Dell PowerEdge R610 48Gbytes, 8 CPU, 8x1Gbit Ethernet)

1xVM image store

To be behind squid proxies

1300xCompute nodes

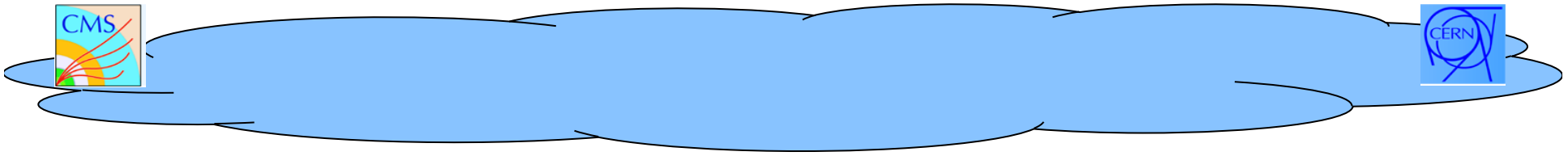
The CMS Cloud Controlling Layer: OpenStack Infrastructure



Operating Architecture for The GRID, our first client as IaaS

The HLT clusters aim to run jobs as a GRID site (simulation jobs in ATLAS, all except possibly user jobs in CMS)

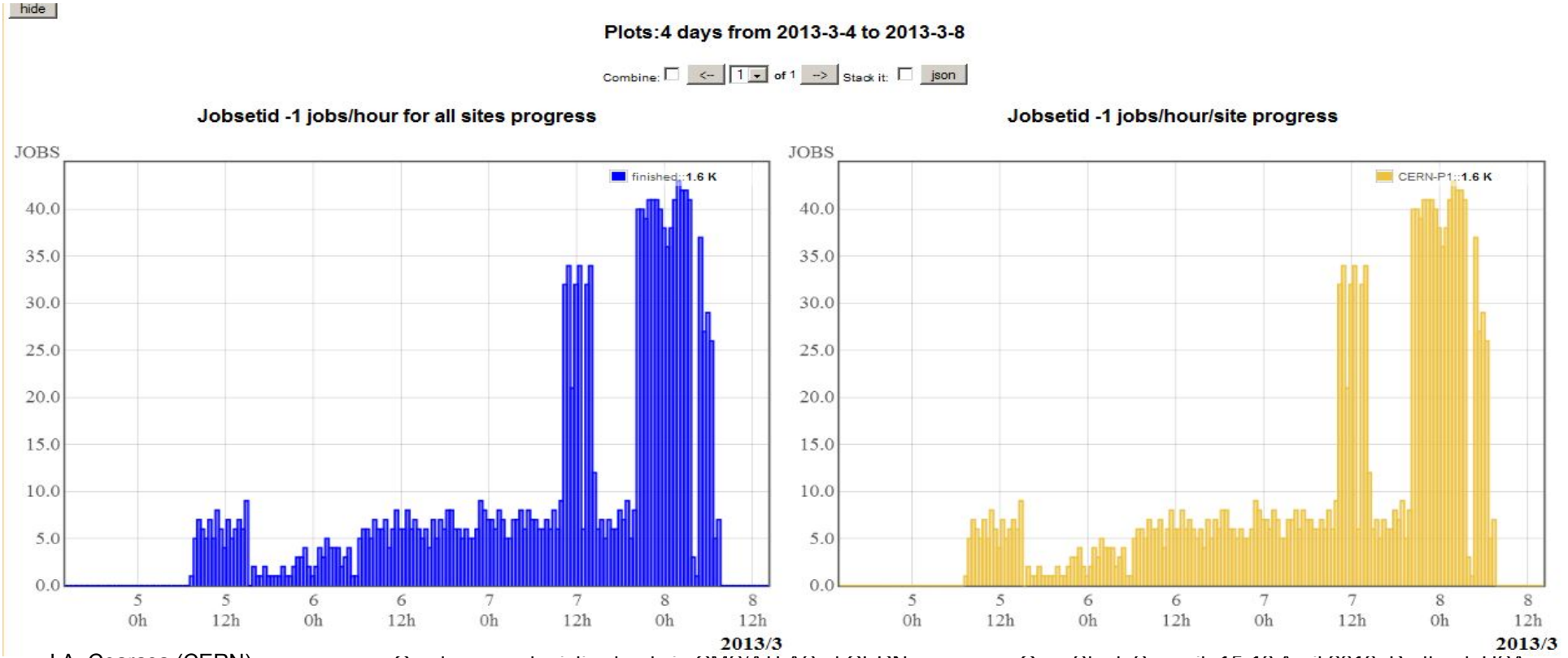
- A dedicated Factory in CERN IT instantiates VMs of the specific flavor from the specific VM image.
  - In CMS A dedicated VM image has been created/In ATLAS CernVM is used.
  - The factory uses condor to control the life of the VM through ec2 commands.
- CVMfs is used:
  - To get the proper Workflow;
  - To get the proper cmssw software.
- Frontier is used. The controller is a frontier server.
- Xrootd is being used to stage in/out files of the cluster
  - A patched version due to bugs in staging out.



Onset, Operation, Outlook and Conclusion

ATLAS HLT cloud onset and achievements

- End of March 2013: Setup of a test OpenStack cloud
- End of March: Running a small simulation production site on the test cloud.
- Production Site on HLT farm foreseen for July



Onset and Evolution of CMS online cloud: Growing on the Technical Stops

- July 2012: Deployment of first OpenStack infrastructure.
- 17-18/9/2012: HLT Cluster *cloudified*/migrated to SLC6.
  - Required cmssw tested on SLC6.
- 8-12/10/2012: First tests of big scale VMs deployment
  - We run the Folding@home project.
- Mid December 2012: First working image and revamped hardware for the controller/proxy.
  - We run the Folding@home project
  - We run the first cmssw workflows over Christmas.
- Cloud running since January 2013 when conditions permit (power/cooling) to integrate it as a GRID resource.
  - Also simultaneously when data taking on the heavy ions runs.

CMSoooooCloud Achievements

- Achievements
  - No impact on data taking.
    - Nor during the setup phase.
    - Neither during data taking on Heavy Ion runs.
  - Controlled ~1300 compute nodes (hypervisors).
  - Deployed to simultaneously run ~1000 VMs in a stable manner during more than 3 weeks.
  - Deployed ~250 VMs (newest cluster) in ~5 min if previously deployed (cached image in hypervisor).
  - Move resources to be used or not by the cloud in seconds.
  - Manual and automatic Workflows have been run:
    - Integration with GRID infrastructure on the way.

CMSooooCloud first Results: The last day of the test with ~1/2 cluster

Folding@home distributed computing

CMSoooCloud

Report generated on	12:43:13 November 06, 2012
Date of last work unit	2012-10-22 02:03:42
Active CPUs within 50 days	583
Team Id	222325
Grand Score	80978 (certificate)
Work Unit Count	768 (certificate)
Team Ranking (incl. aggregate)	18122 of 216623
Home Page	http://fah-web.stanford.edu/generic.html

Like Be the first of your friends to like this.

Team members

Rank (within team)	Donor	Score	WU
1	CMSoooCloud	80978	768

CMSooooCloud first Results: After Christmas

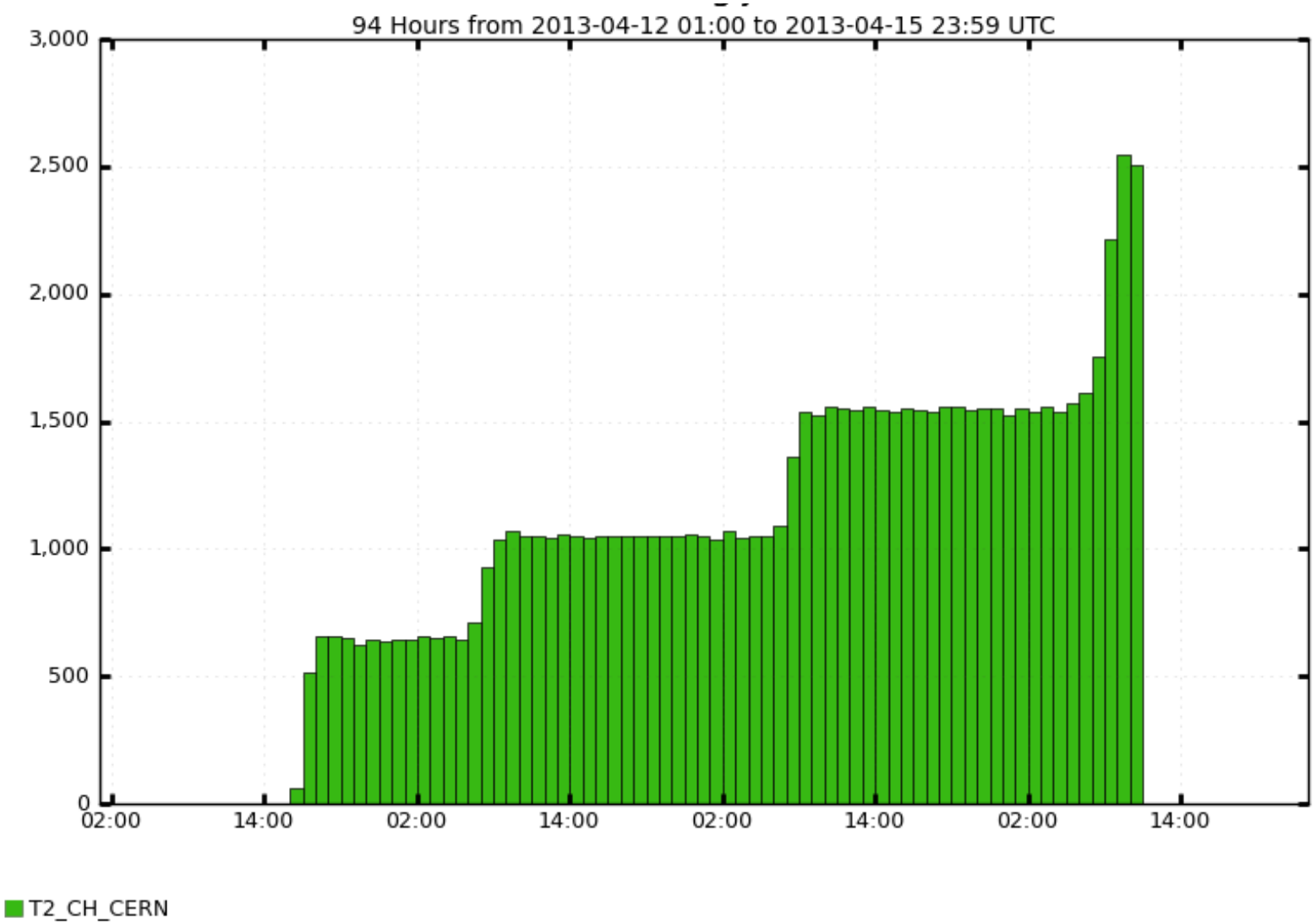
Folding@home distributed computing

CMSoooCloud

Report generated on	01:35:29 January 15, 2013
Date of last work unit	2013-01-14 02:08:44
Active CPUs within 50 days	3218
Team Id	222325
Grand Score	56984120 (certificate)
Work Unit Count	54625 (certificate)
Team Ranking (incl. aggregate)	320 of 217186
Home Page	http://fah-web.stanford.edu/generic.html
Fast Teampage URL	http://fah-web2.stanford.edu/teamstats/team222325.html

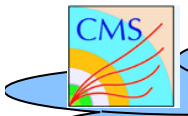
Like You like this.

CMSoooooCloud first Results: Running as a GRID site



Outlook of CMSoooCloud

- Reconfigure the network to use VLAN on data network in multi-host with hardware router mode.
- Reconfigure the proxys infrastructure to use the high bandwidth connection to CERN (up to 40 Gbit/s).
- Hide Glance server behind our proxys.
- Migrate to new version of OpenStack (Grizzly):
  - Use of Quantum possible;
  - High Availability features.
- Interoperate with CERN IT's OpenStack Cloud, as a connected cell/zone.



ATLAS HLT cloud: Conclusions

A testbed cloud has been deployed successfully.

Simulation jobs have been run on the testbed cloud.

An overlay Cloud layer is being deployed on the ATLAS online High Level Trigger cluster.

Conversations are taking place to define the policy of utilization.

CMSooooCloud: Conclusions

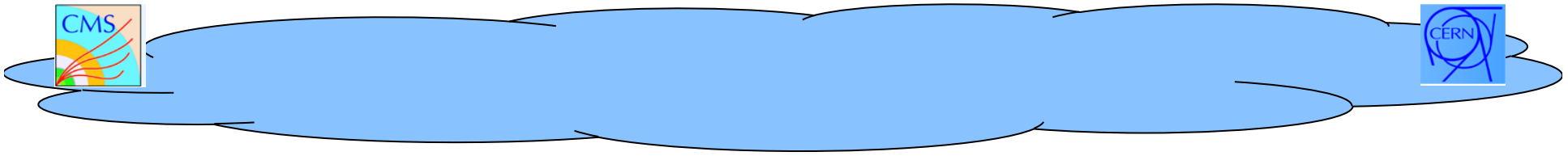
An overlay Cloud layer has been deployed on the CMS online High Level Trigger cluster with zero impact on data taking.

The man power dedicated to *cloudify* the HLT cluster has been low (~1.5 FTE or less for ~6 months) for the potential offered.

We are sharing the knowledge on how to deploy such an overlay layer to existing sites that may transition to cloud infrastructures (ATLAS online, Bologna, IFAE).

We are gaining experience on how to contextualize and deploy VMs on the cloud infrastructures, that are becoming commonplace, to run the GRID jobs.

We were able to run different kind of GRID jobs and non GRID jobs on our cloud.



Thank you. Questions?