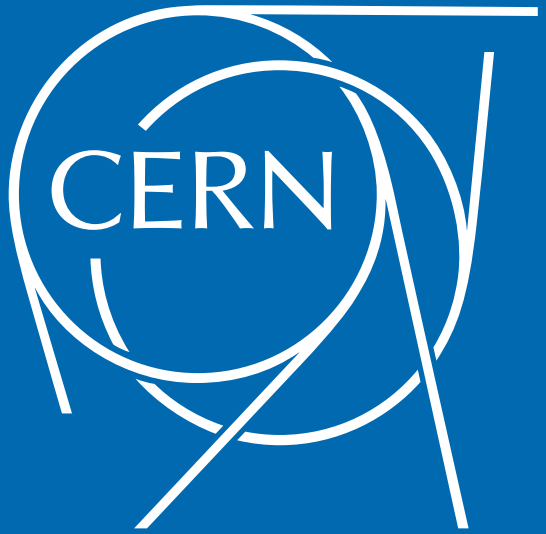

CERN

Deep Dive into the CERN Cloud Infrastructure

Openstack Design Summit – Hong Kong, 2013

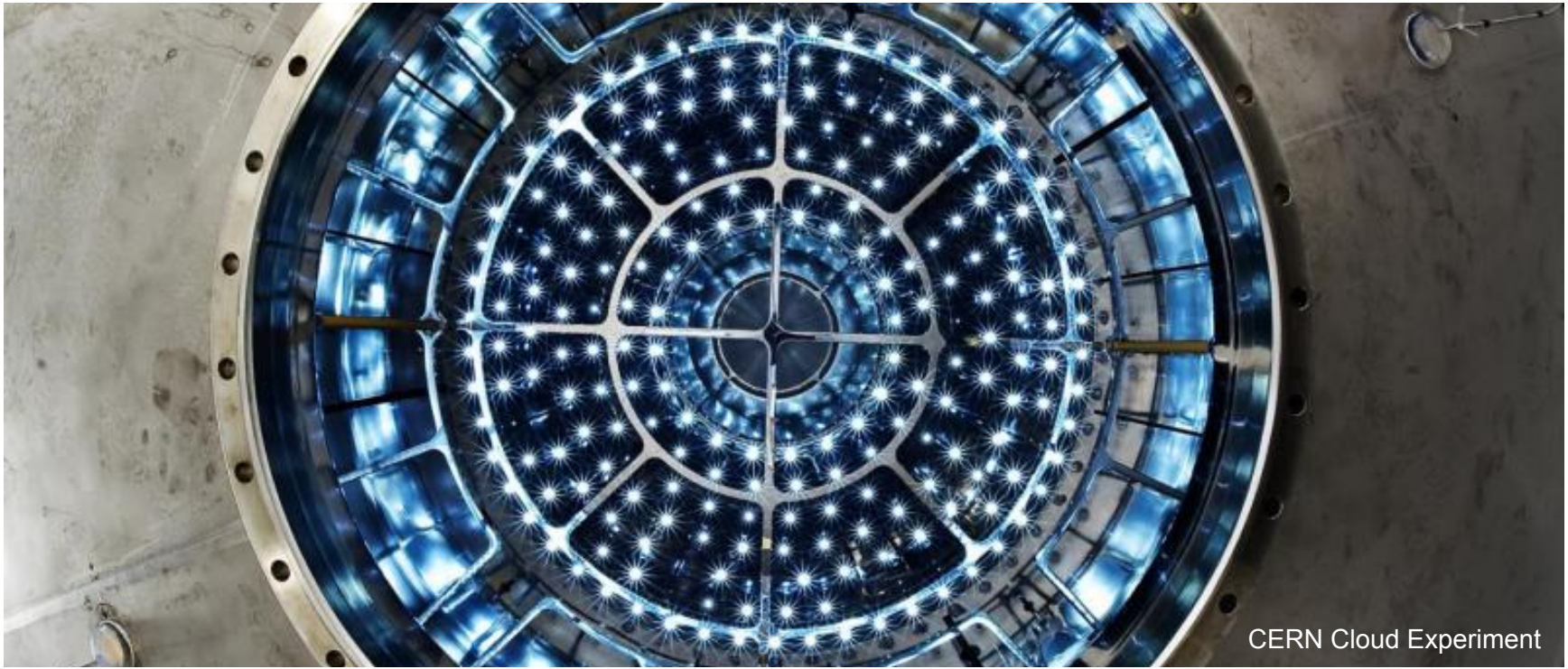
Belmiro Moreira

belmiro.moreira@cern.ch @belmiromoreira

What is CERN?

- Conseil Européen pour la Recherche Nucléaire – aka European Organization for Nuclear Research
- Founded in 1954 with an international treaty
  - 20 state members, other countries contribute to experiments
- Situated between Geneva and the Jura Mountains, straddling the Swiss-French border

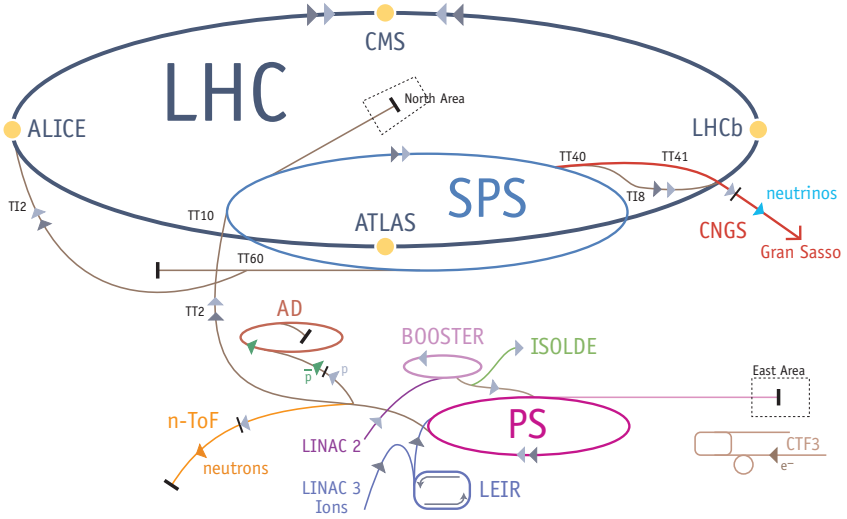
What is CERN?



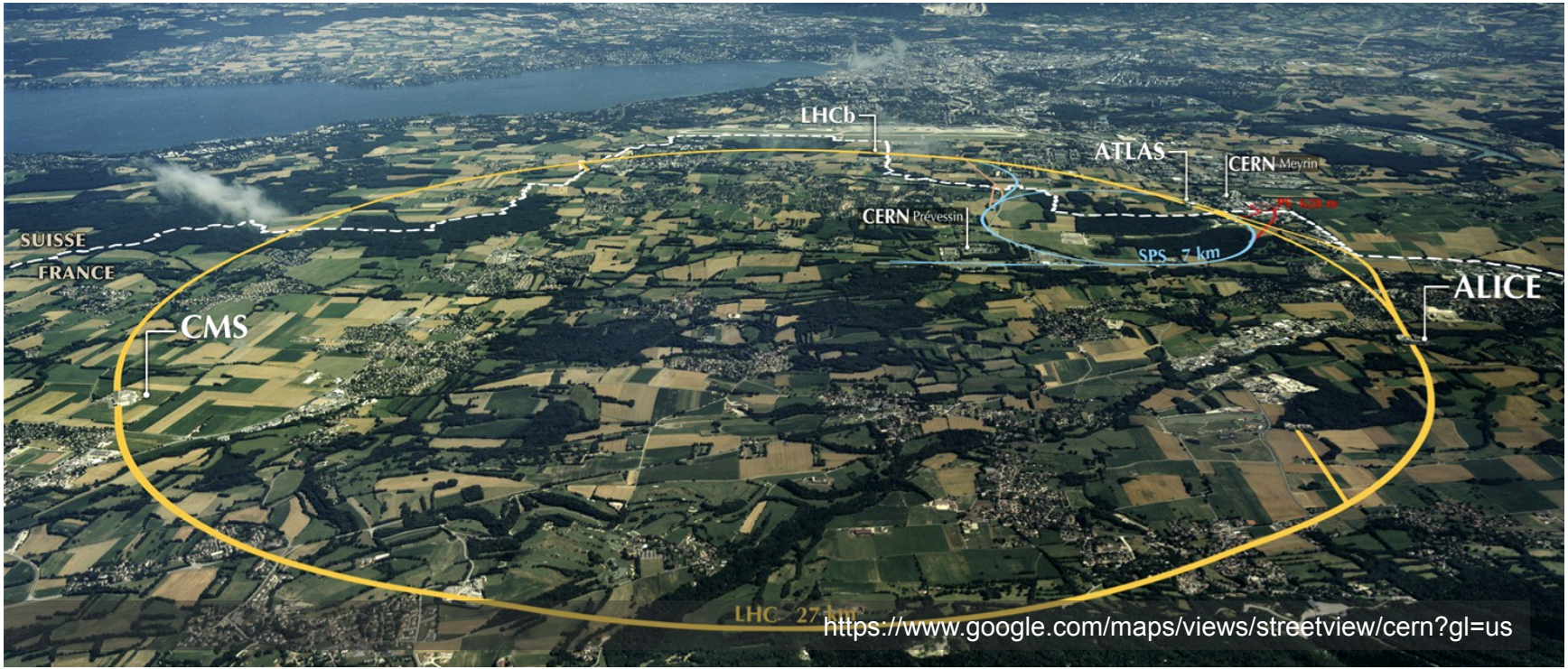
CERN Cloud Experiment

What is CERN?

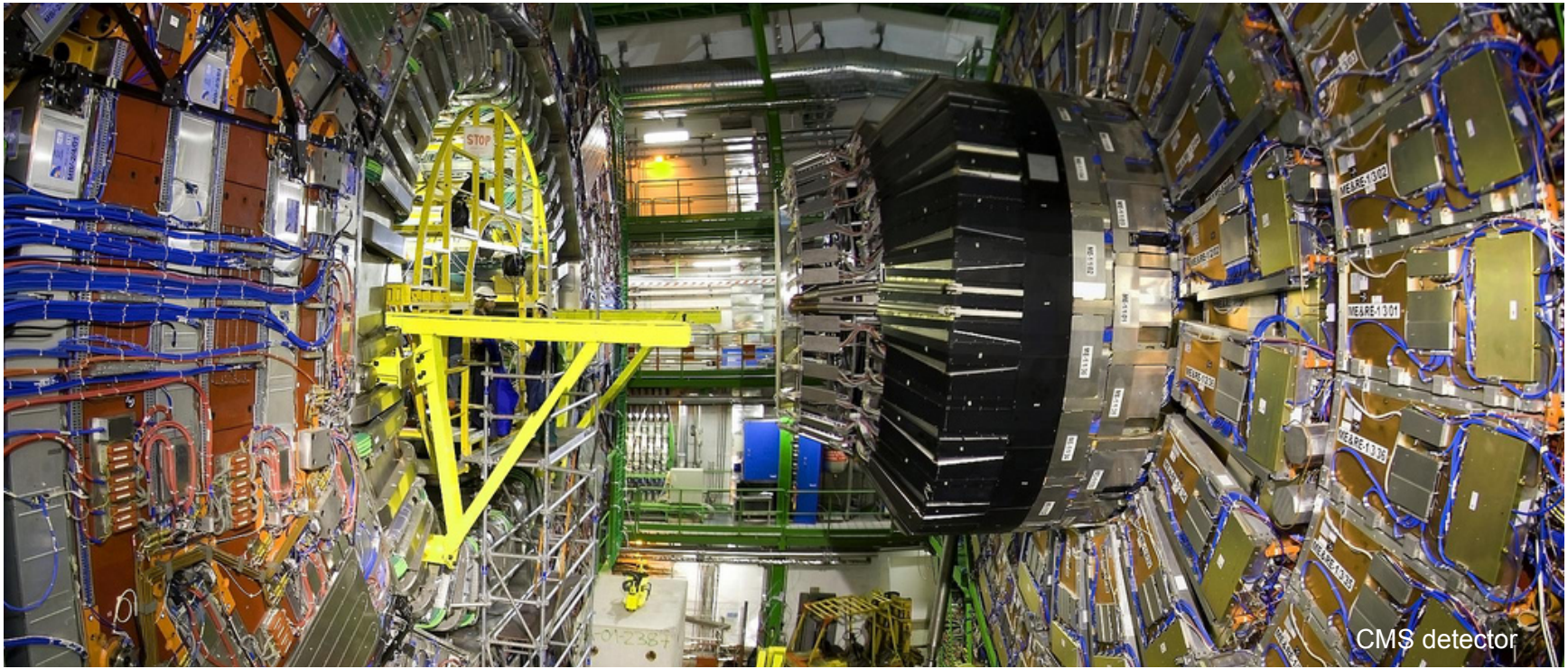
CERN provides particle accelerators and other infrastructure for high-energy physics research



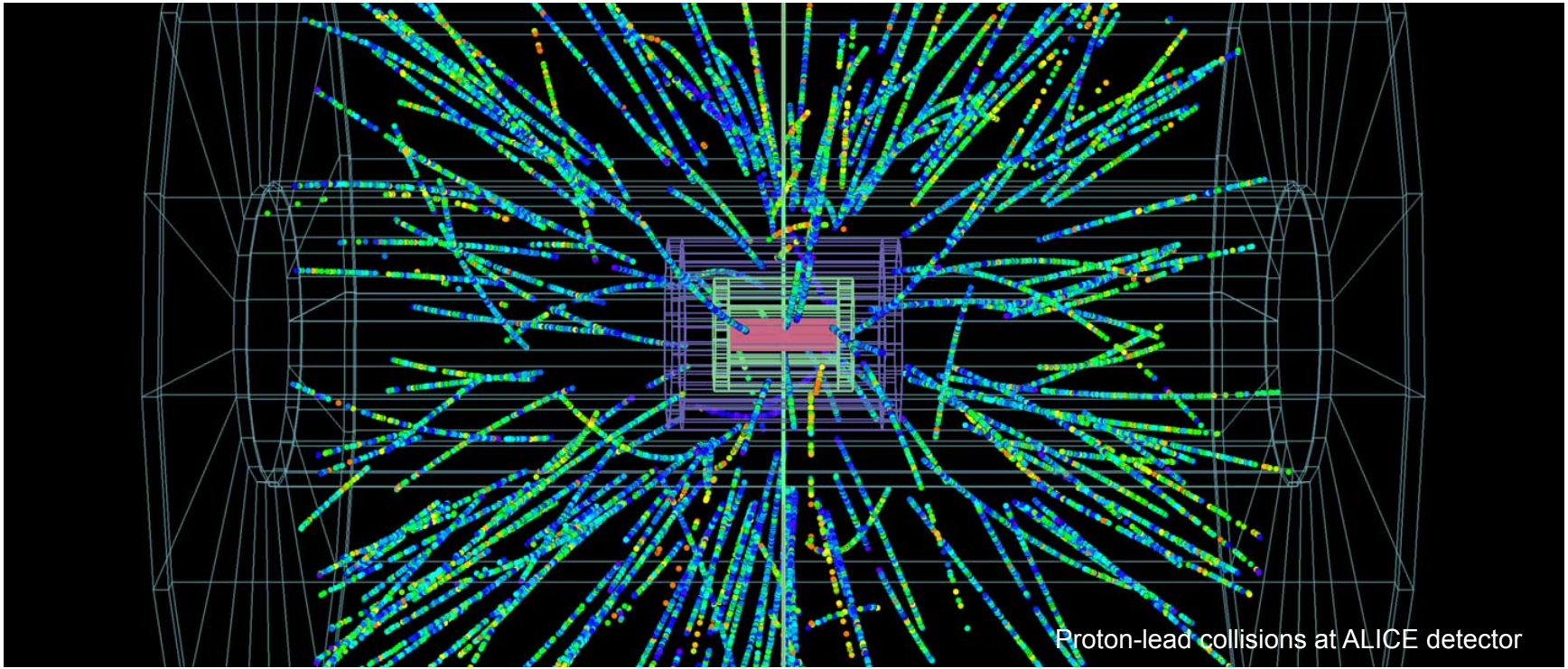
LHC - Large Hadron Collider



LHC and Experiments



LHC and Experiments

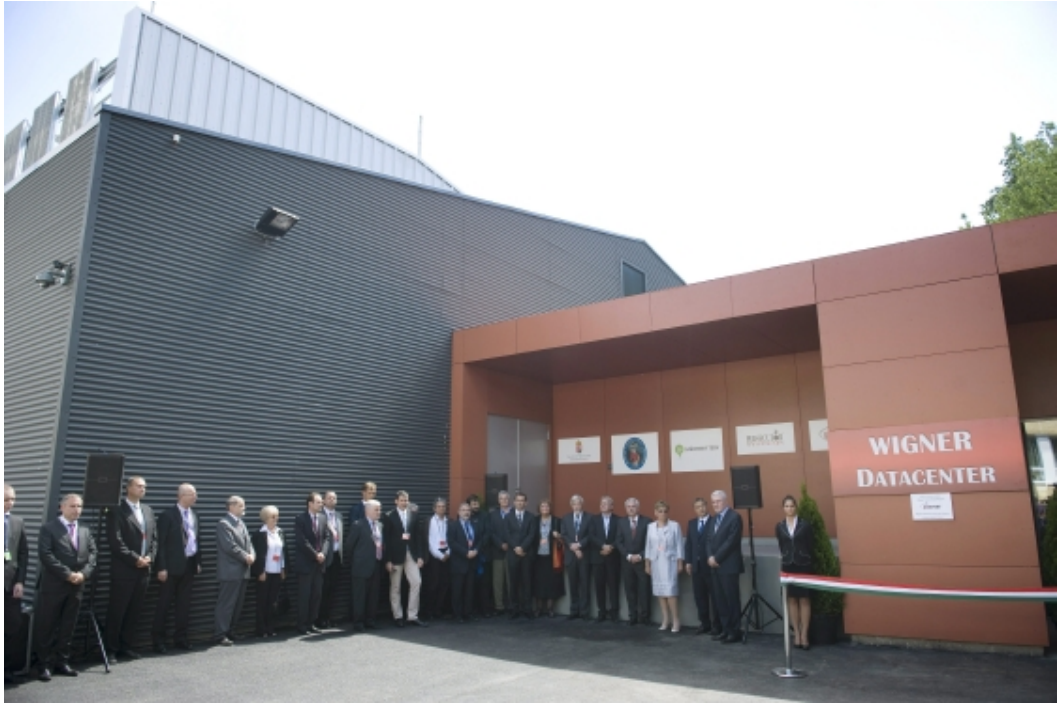


Proton-lead collisions at ALICE detector

CERN - Computer Center - Geneva, Switzerland

- 3.5 Mega Watts
- ~91000 cores
- ~120 PB HDD
- ~100 PB Tape
- ~310 TB Memory

CERN - Computer Center - Budapest, Hungary



- 2.5 Mega Watts
- ~20000 cores
- ~6 PB HDD

Computer Centers location

CERN IT Infrastructure in 2011

- ~10k servers
  - Dedicated compute, dedicated disk server, dedicated service nodes
  - Mostly running on real hardware
  - Server consolidation of some service nodes using Microsoft HyperV/ SCVMM
    - ~3400 VMs (~2000 Linux, ~1400 Windows)
  - Various other virtualization projects around
- Many diverse applications ("clusters")
  - Managed by different teams (CERN IT + experiment groups)

CERN IT Infrastructure challenges in 2011

- Expected new Computer Center in 2013
- Need to manage twice the servers
- No increase in staff numbers
- Increasing number of users / computing requirements
- Legacy tools - high maintenance and brittle

Why Build CERN Cloud

Improve operational efficiency

- Machine reception and testing
- Hardware interventions with long running programs
- Multiple operating system demand

Improve resource efficiency

- Exploit idle resources
- Highly variable load such as interactive or build machines

Improve responsiveness

- Self-service

Identify a new Tool Chain

- Identify the tools needed to build our Cloud Infrastructure
  - Configuration Manager tool
  - Cloud Manager tool
  - Monitoring tools
- Storage Solution

Strategy to deploy OpenStack

- Configuration infrastructure based on Puppet
- Community Puppet modules for OpenStack
- SLC6 Operating System
- EPEL/RDO - RPM Packages

Strategy to deploy OpenStack

- Deliver a production IaaS service though a series of time-based pre-production services of increasing functionality and Quality-of-Service
- Budapest Computer Center hardware deployed as OpenStack compute nodes
- Have an OpenStack production service in the Q2 of 2013

Pre-Production Infrastructure

Essex | Folsom

"Guppy"
June, 2012

- Deployed on Fedora 16
- Community OpenStack puppet modules
- Used for functionality tests
- Limited integration with CERN infrastructure

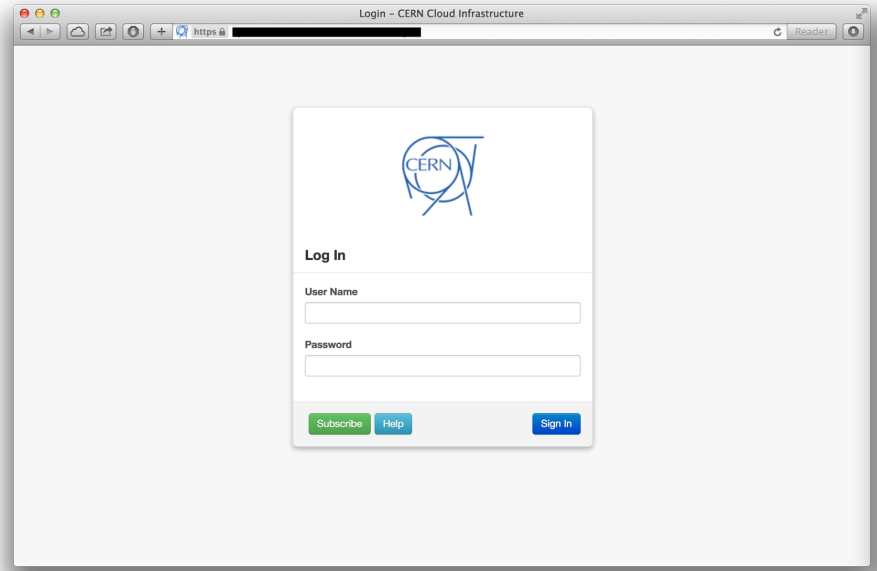
"Hamster"
October, 2012

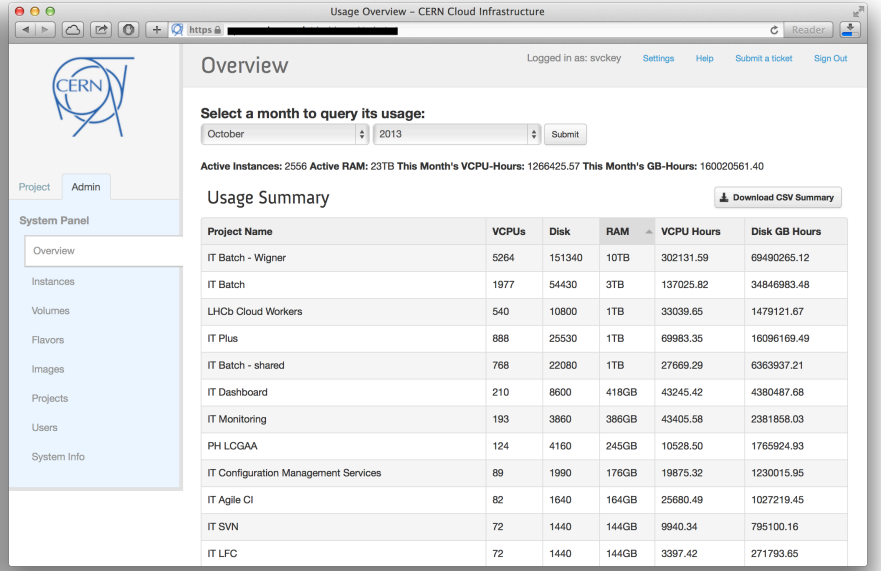
- Open to early adopters
- Deployed on SLC6 and Hyper-V
- CERN Network DB integration
- Keystone LDAP integration

"Ibex"
March, 2013

- Open to a wider community (ATLAS, CMS, LHCb, …)
- Some OpenStack services in HA
- ~14000 cores

OpenStack at CERN - grizzly release





Project Name	VCPUs	Disk	RAM	VCPU Hours	Disk GB Hours
IT Batch - Wigner	5264	151340	10TB	302131.59	69490265.12
IT Batch	1977	54430	3TB	137025.82	34846983.48
LHCb Cloud Workers	540	10800	1TB	33039.65	1479121.67
IT Plus	888	25530	1TB	69983.35	16096169.49
IT Batch - shared	768	22080	1TB	27669.29	6363937.21
IT Dashboard	210	8600	418GB	43245.42	4380487.68
IT Monitoring	193	3860	386GB	43405.58	2381858.03
PH LCGAA	124	4160	245GB	10528.50	1765924.93
IT Configuration Management Services	89	1990	176GB	19875.32	1230015.95
IT Agile CI	82	1640	164GB	25680.49	1027219.45
IT SVN	72	1440	144GB	9940.34	795100.16
IT LFC	72	1440	144GB	3397.42	271793.65

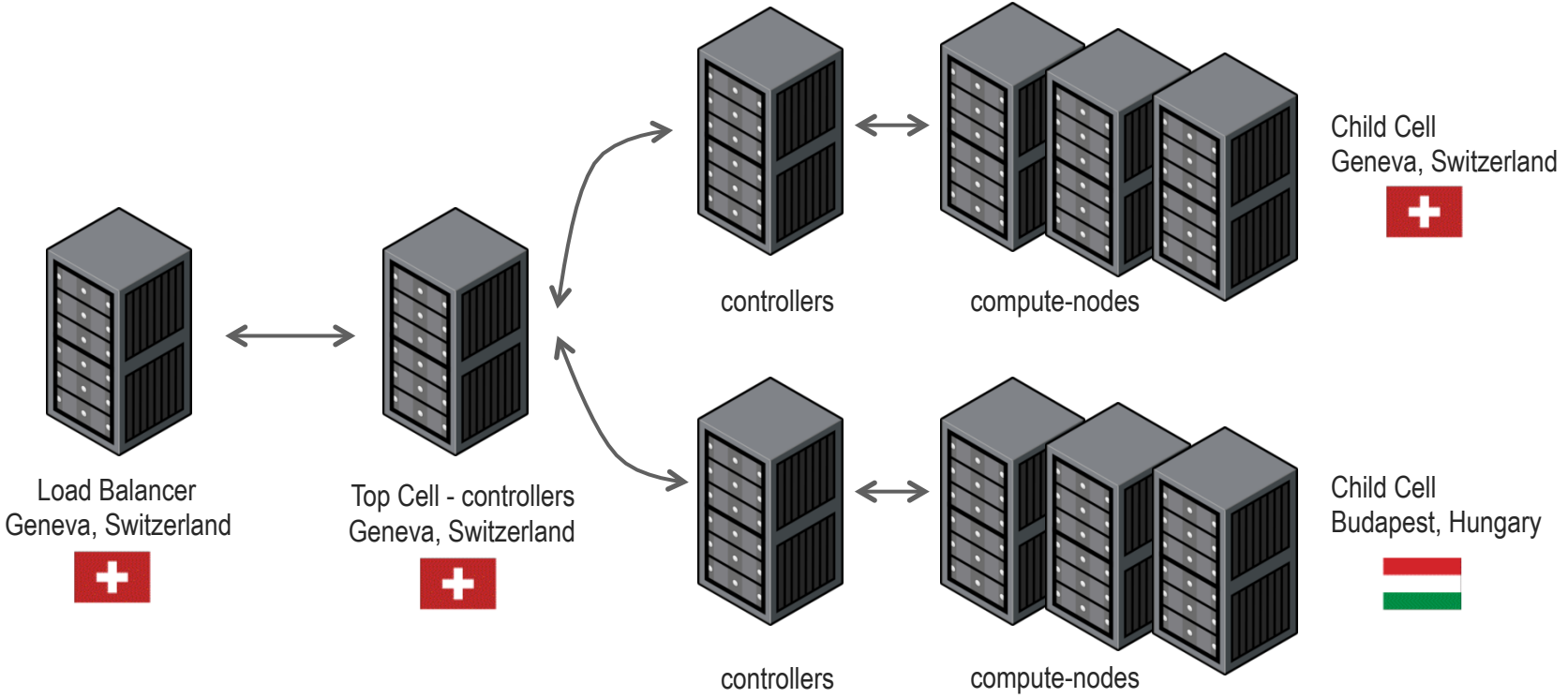
OpenStack at CERN - grizzly release

- +2 Children Cells – Geneva and Budapest Computer Centers
- HA+1 architecture
- Ceilometer deployed
- Integrated with CERN accounts and network infrastructure
- Monitoring OpenStack components status
- Glance - Ceph backend
- Cinder - Testing with Ceph backend

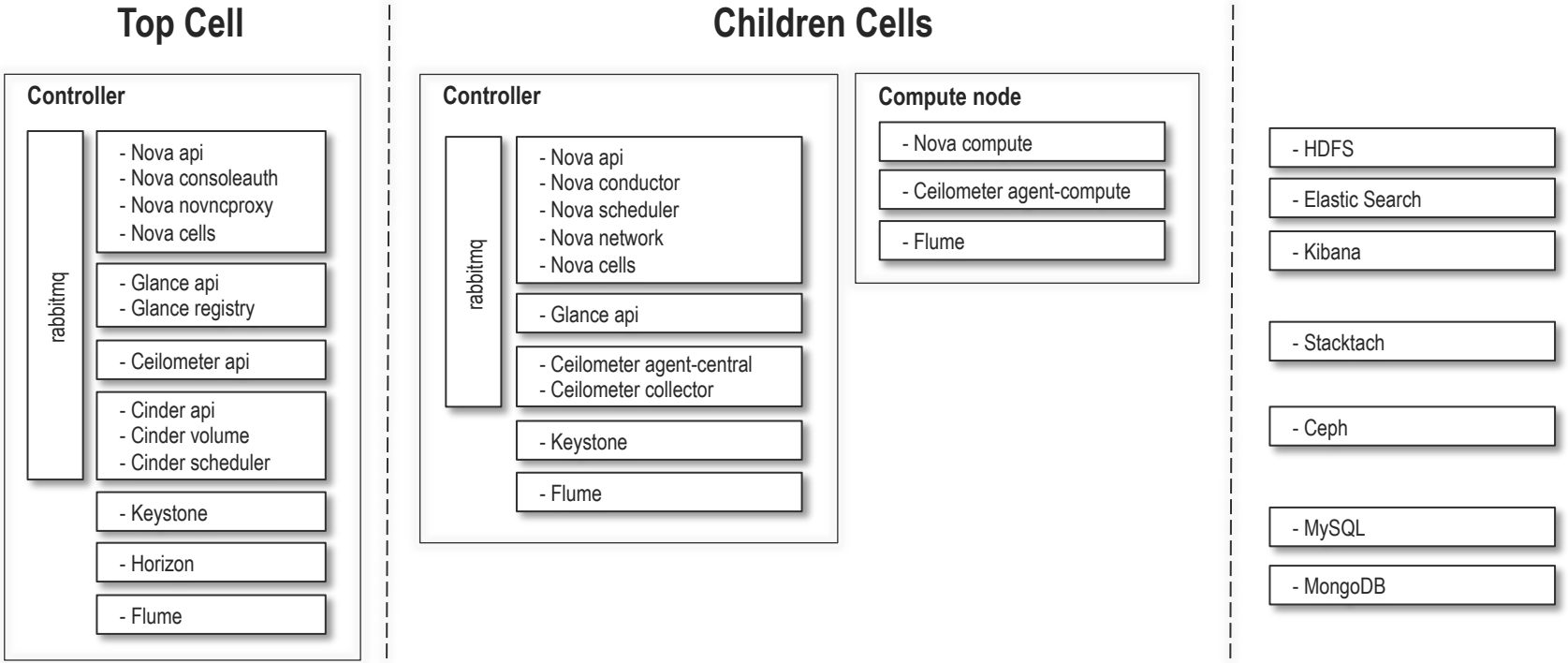
Infrastructure Overview

- Adding ~100 compute nodes every week
  - Geneva, Switzerland Cell
    - ~11000 cores
  - Budapest, Hungary Cell
    - ~10000 cores
- Today we have +2500 VMs
  - Several VMs have more than 8 cores

Architecture Overview



Architecture Components



Infrastructure Overview

- SLC6 and Microsoft Windows 2012
  - KVM and Microsoft HyperV
- All infrastructure “puppetized” (also, windows compute nodes!)
  - Using stackforge OpenStack puppet modules
  - Using CERN Foreman/Puppet configuration infrastructure
    - Master, Client architecture
    - Puppet managed VMs - share the same configuration infrastructure

Infrastructure Overview

- HAProxy as load balancer
- Master and Compute nodes
  - 3+ Master nodes per Cell
  - O(1000) Compute nodes per Child Cell (KVM and HyperV)
  - 3 availability zones per Cell
- Rabbitmq
  - At least 3 brokers per Cell
  - Rabbitmq cluster with mirrored queues

Infrastructure Overview

- MySql instance per Cell
  - MySql managed by CERN DB team
  - Running on top of Oracle CRS
  - active/slave configuration
  - NetApp storage backend
  - Backups every 6 hours

Nova Cells

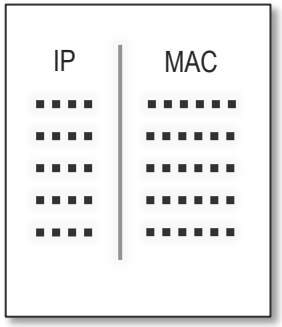
- Why Cells?
  - Scale transparently between different Computer Centers
- With cells we lost functionality
  - Security groups
  - Live migration
- "Parents" don't know about “children” compute
- Flavors not propagated to "children” cells

Nova Cells

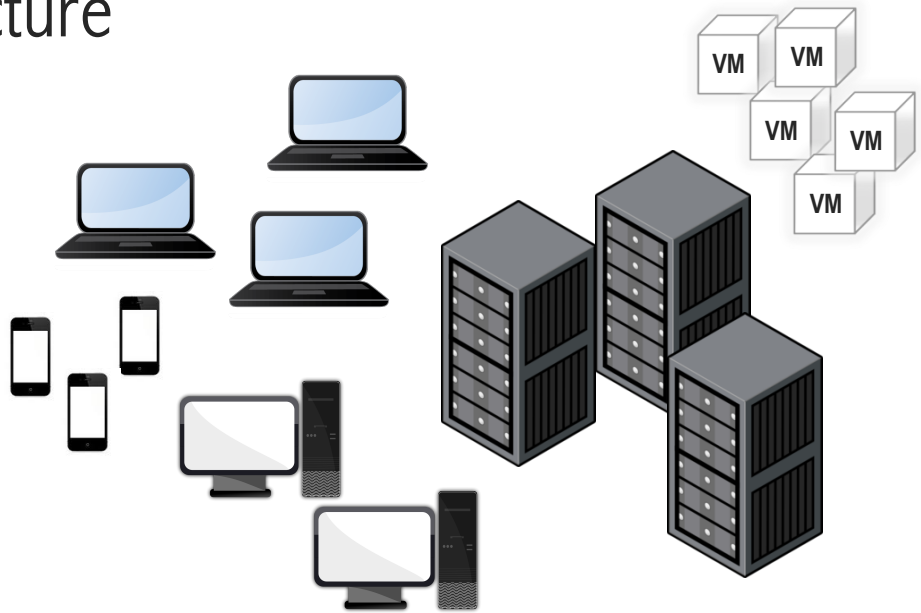
- Scheduling
  - Random cell selection on Grizzly
  - Implemented simple scheduler based on project
    - CERN Geneva only, CERN Wigner only, “both”
    - “both” selects the cell with more available free memory
- Cell/Cell communication doesn’t support multiple Rabbitmq servers
  - https://bugs.launchpad.net/nova/+bug/1178541

Nova Network

- CERN network infrastructure



CERN network DB



Nova Network

- Implemented a Nova Network CERN driver
  - Considers the “host” picked by nova-scheduler
  - MAC address selected from pre-registered addresses of “host” IP Service
  - Updates CERN network database address with instance hostname and responsible of the device
- Network constraints in some nova operations
  - Resize, Live-Migration

Nova Scheduler

- ImagePropertiesFilter
  - linux/windows hypervisors in the same infrastructure
- ProjectsToAggregateFilter
  - Projects need dedicated resources
  - Instances from defined projects are created in specific Aggregates
  - Aggregates can be shared by a set of projects
- Availability Zones
  - Implemented “default_schedule_zones”

Nova Conductor

- Reduces “dramatically” the number of DB connections
- Conductor “bottleneck”
  - Only 3+ processes for “all” DB requests
  - General “slowness” in the infrastructure
  - Fixed with backport
    - https://review.openstack.org/#/c/42342/

Nova Compute

- KVM and Hyper-V compute nodes share the same infrastructure
  - Hypervisor selection based on “Image” properties
- Hyper-V driver still lacks some functionality on Grizzly
  - Console access, metadata support with nova-network, resize support, ephemeral disk support, ceilometer metrics support

Keystone

- CERN's Active Directory infrastructure
  - Unified identity management across the site
  - +44000 users
  - +29000 groups
  - ~200 arrivals/departures per month
- Keystone integrated with CERN Active Directory
  - LDAP backend

Keystone

- CERN user subscribes the "cloud service”
  - Created "Personal Tenant" with limited quota
- Shared projects created by request
- Project life cycle
  - owner, member, admin – roles
  - "Personal project" disabled when user leaves
    - Delete resources (VMs, Volumes, Images, …)
  - User removed from "Shared Projects"

Ceilometer

- Users are not directly billed
  - Metering needed to adjust Project quotas
- mongoDB backend – sharded and replicated
- Collector, Central-Agent
  - Running on “children” Cells controllers
- Compute-Agent
  - Uses nova-api running on “children” Cells controllers

Glance

- Glance API
  - Using glance api v1
  - python-glanceclient doesn't support completely v2
- Glance Registry
  - With v1 we need to keep Glance Registry
    - Only runs in Top Cell behind the load balancer
- Glance backend
  - File Store (AFS)
  - Ceph

Glance

- Maintain small set of SLC5/6 images as default
  - Difficult to offer only the most updated set of images
  - Resize and Live Migration not available if image is deleted from Glance
- Users can upload images up to 25GB
  - Users don't pay storage!
  - Glance in Grizzly doesn't support quotas per Tenant!

Cinder

- Ceph backend
  - Still in evaluation
- SLC6 with qemu-kvm patched by Inktank to support RBD
- Cinder doesn't support cells in Grizzly
  - Fixed with backport:

    https://review.openstack.org/#/c/31561/

Ceph as Storage Backend

- 3 PB cluster available for Ceph
  - 48 OSDs servers
  - 5 Monitors servers
- Initial testing with FIO, libaio, bs 256k

```
fio --size=4g --bs=256k -numjobs=1 --direct=1 --rw=randrw
--ioengine=libaio --name=/mnt/vdb1/tmp4
```

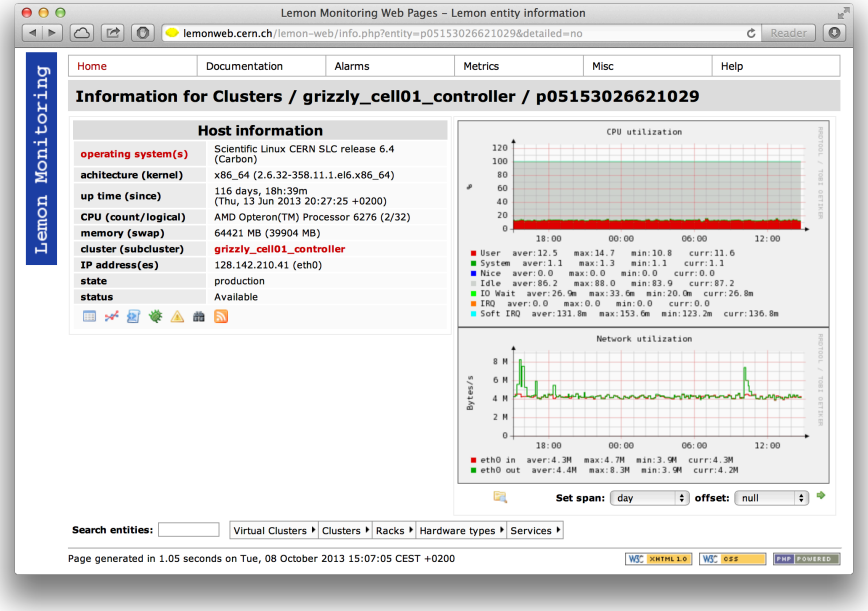
Rand RW	Rand R	Rand W
99 MB/s	103 MB/s	108 MB/s

Ceph as Storage Backend

- ulimits
  - With more than >1024 OSDs, we're getting various errors where clients cannot create enough processes
- authx for security (key lifecycle is a challenge as always)
- need librbd (from EPEL)

Monitoring - Lemon

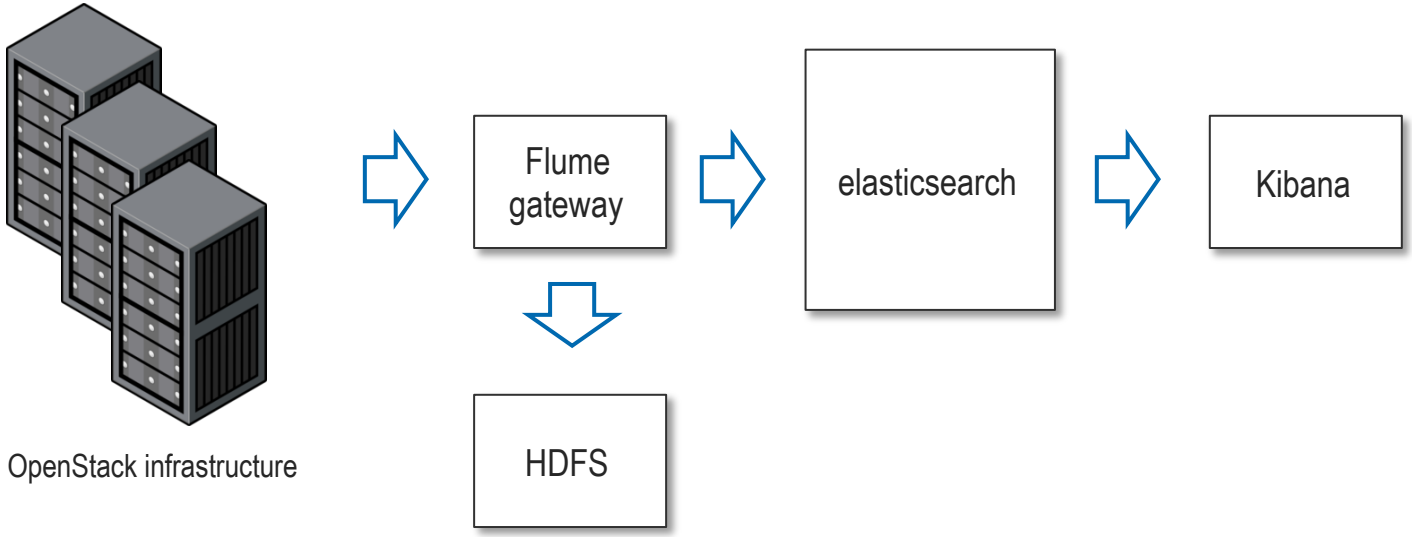
- Monitor “physical” and virtual “servers” with Lemon



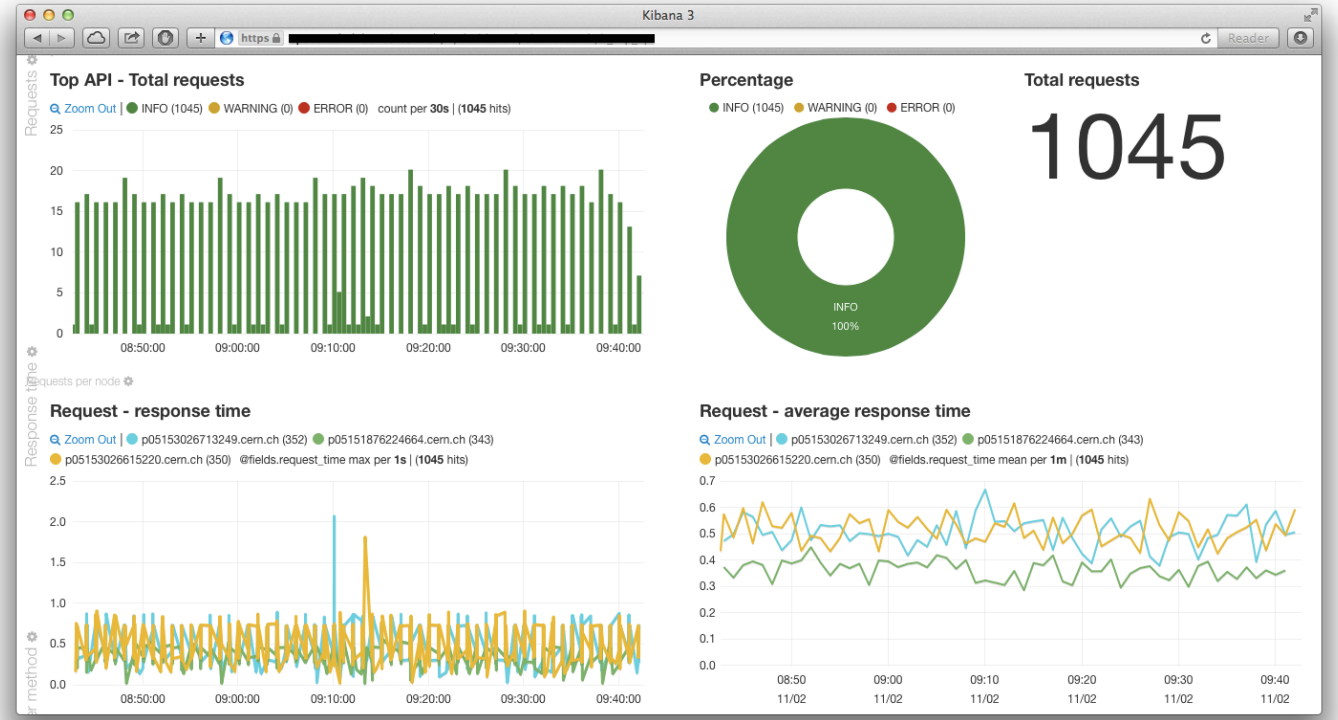
Monitoring - Flume, Elastic Search, Kibana

- How to monitor OpenStack status in all nodes?
  - ERRORs, WARNINGs – log visualization
  - identify in “real time” possible problems
  - preserve all logs for analytics
  - visualization of cloud infrastructure status
    - service managers
    - resource managers
    - users

Monitoring - Flume, Elastic Search, Kibana



Monitoring - Kibana



Monitoring - Kibana

Challenges

- Moving resources to the infrastructure
  - +100 compute nodes per week
  - 15000 servers – more than 300000 cores
- Migration from Grizzly to Havana
- Deploy Neutron
- Deploy Heat
- Kerberos, X.509 user certificate authentication
- Keystone Domains

belmiro.moreira@cern.ch
@belmiromoreira

www.cern.ch