CERN

6 years of CERN Cloud

From 0 to 300k cores

HEPIX - San Diego - 2019

Belmiro Moreira

on behalf of the CERN Cloud Team

belmiro.moreira@cern.ch @belmiromoreira

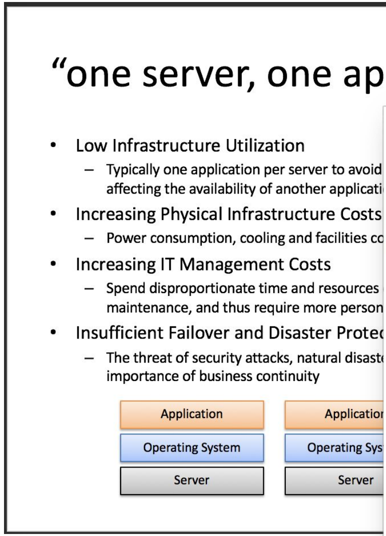
Outline

- Early Virtualization attempts
- Finding the right Cloud Orchestrator
- CERN OpenStack Cloud
- Some Infrastructure Highlights
- What’s next?

2009 - 2011

Virtualization and Server Consolidation

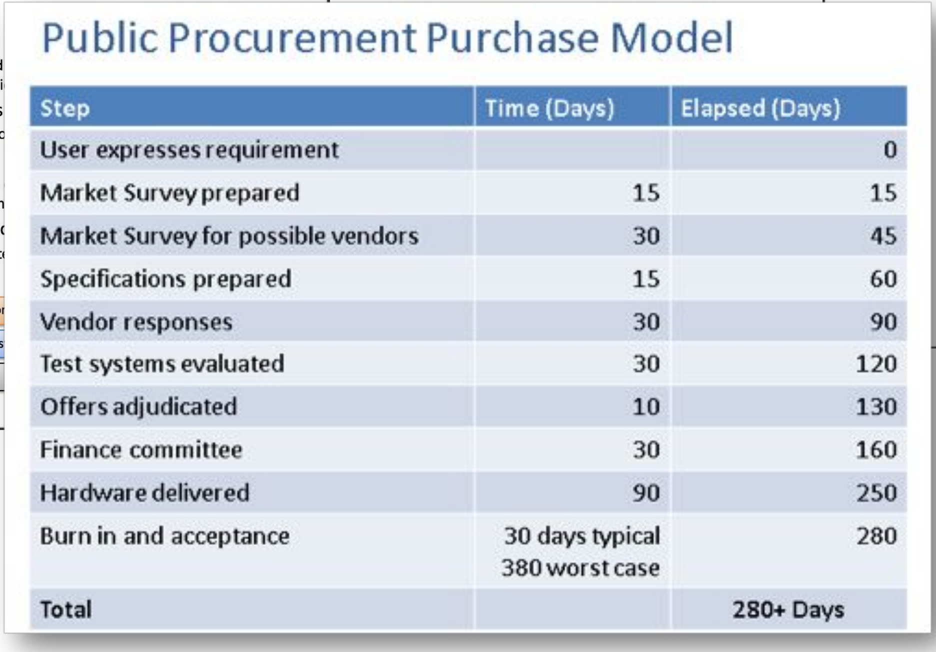
Why IaaS at CERN?



“one server, one application”

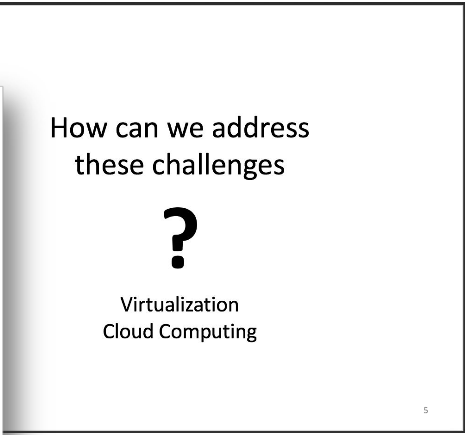
- Low Infrastructure Utilization
  - Typically one application per server to avoid affecting the availability of another applicati
- Increasing Physical Infrastructure Costs
  - Power consumption, cooling and facilities co
- Increasing IT Management Costs
  - Spend disproportionate time and resources maintenance, and thus require more person
- Insufficient Failover and Disaster Protec
  - The threat of security attacks, natural disast importance of business continuity

Application
Operating System
Server
Applicatio
Operating Sys
Server



Public Procurement Purchase Model

Step	Time (Days)	Elapsed (Days)
User expresses requirement		0
Market Survey prepared	15	15
Market Survey for possible vendors	30	45
Specifications prepared	15	60
Vendor responses	30	90
Test systems evaluated	30	120
Offers adjudicated	10	130
Finance committee	30	160
Hardware delivered	90	250
Burn in and acceptance	30 days typical 380 worst case	280
Total		280+ Days

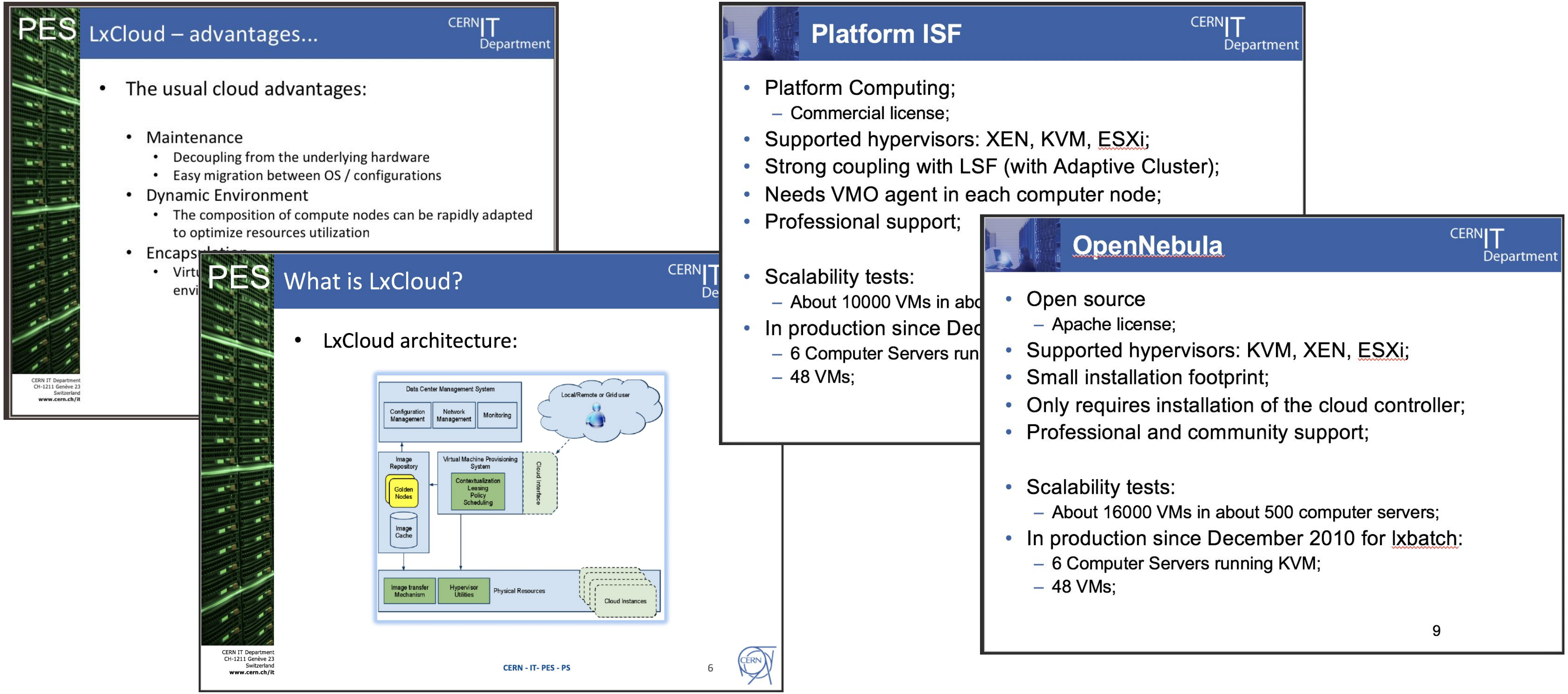


How can we address these challenges

?

Virtualization
Cloud Computing

LxCloud - Virtualize Batch Infrastructure



PES LxCloud – advantages...

CERN IT Department

- The usual cloud advantages:
  - Maintenance
    - Decoupling from the underlying hardware
    - Easy migration between OS / configurations
  - Dynamic Environment
    - The composition of compute nodes can be rapidly adapted to optimize resources utilization
  - Encaps
    - Virtu envi

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

PES What is LxCloud?

CERN IT De

- LxCloud architecture:

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN - IT- PES - PS

6

Platform ISF

CERN IT Department

- Platform Computing;
  - Commercial license;
- Supported hypervisors: XEN, KVM, ESXi;
- Strong coupling with LSF (with Adaptive Cluster);
- Needs VMO agent in each computer node;
- Professional support;
- Scalability tests:
  - About 10000 VMs in abo
- In production since Dec
  - 6 Computer Servers run
  - 48 VMs;

OpenNebula

CERN IT Department

- Open source
  - Apache license;
- Supported hypervisors: KVM, XEN, ESXi;
- Small installation footprint;
- Only requires installation of the cloud controller;
- Professional and community support;
- Scalability tests:
  - About 16000 VMs in about 500 computer servers;
- In production since December 2010 for lxbatch:
  - 6 Computer Servers running KVM;
  - 48 VMs;

9

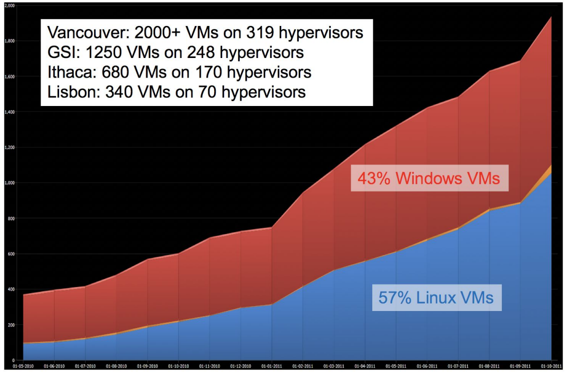
CVI - CERN Virtual Infrastructure

OIS **Reminder: what is CVI?** CERN IT Department

- The CERN Virtual Infrastructure custom virtual machines in the CERN computer center
  - These VMs have a long-term lifetime of months/years
- User kiosk for requesting a VM in less than 30 mins
- Based on Microsoft's System Center Virtual Machine Manager
  - Enterprise class centralized
  - Rich feature set:
    - Allows grouping of hypervisors administrative privileges
    - VM migration, High availability
    - Checkpointing
    - PowerShell Snap-In for admin

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

OIS **Spectacular growth continues!** CERN IT Department



Vancouver: 2000+ VMs on 319 hypervisors
GSI: 1250 VMs on 248 hypervisors
Ithaca: 680 VMs on 170 hypervisors
Lisbon: 340 VMs on 70 hypervisors

Number of Virtual Machines per Operating System

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CVI status update - 30

OIS **Updates since Spring 2011** CERN IT Department

- CVI 2.0 running stably
- Integration Components for RHEL6/SLC6
  - Released by Microsoft in July
  - Working fine
  - Much improved packaging wrt RHEL5
    - Provided as RPMs
    - Source code is snapshot of upstream kernel drivers
      - Still in staging area ☹

Service grows by ~100 VMs per month

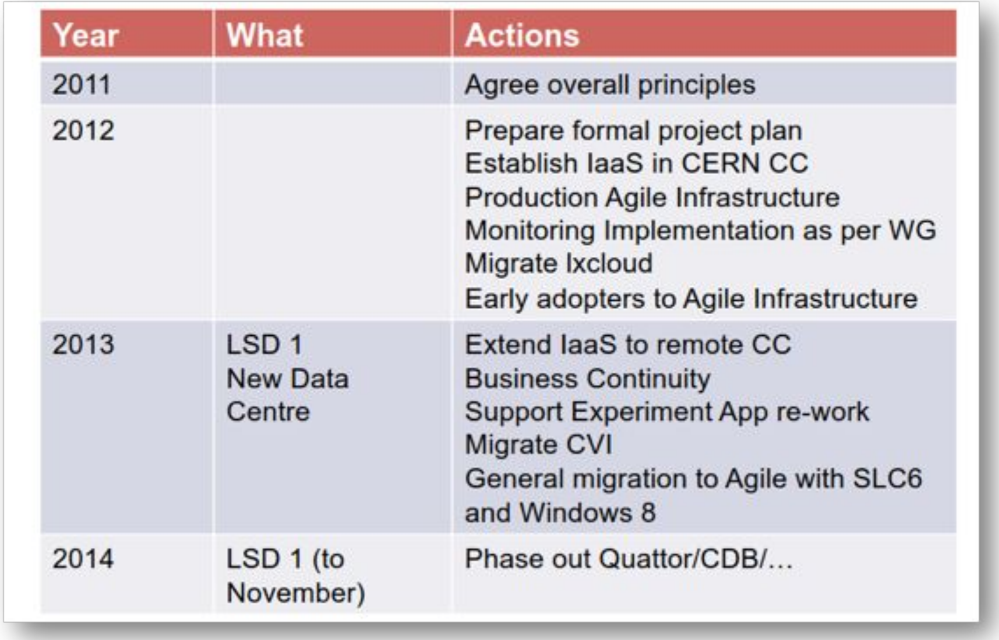
- New customers: IPv6 testing,3D rendering farm

CVI status update - 29

2011 - 2013

Cloud Prototype

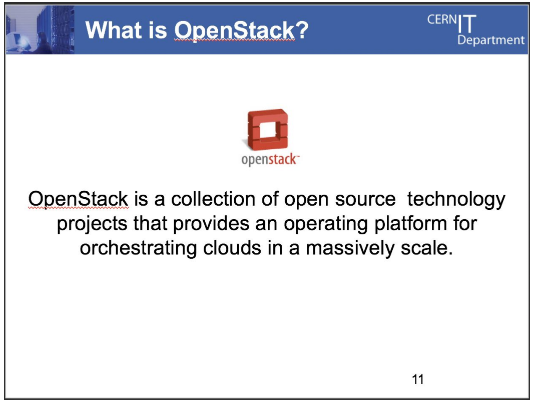
Agile Infrastructure Project



Year	What	Actions
2011		Agree overall principles
2012		Prepare formal project plan Establish IaaS in CERN CC Production Agile Infrastructure Monitoring Implementation as per WG Migrate lxcloud Early adopters to Agile Infrastructure
2013	LSD 1 New Data Centre	Extend IaaS to remote CC Business Continuity Support Experiment App re-work Migrate CVI General migration to Agile with SLC6 and Windows 8
2014	LSD 1 (to November)	Phase out Quattor/CDB/…

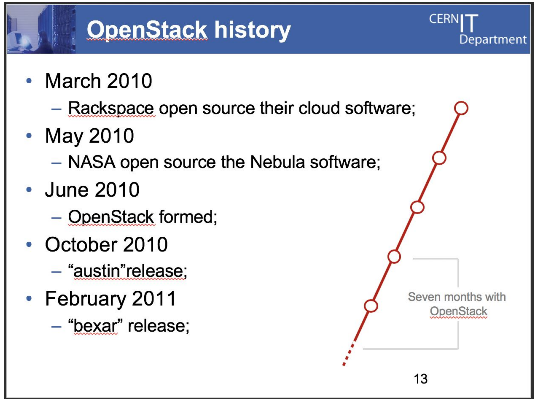
OpenStack at CERN - Early days

- Released on October 2010
- Created by Nasa and Rackspace
- Austin release (Nova, Swift)
- OpenSource (Apache 2.0)



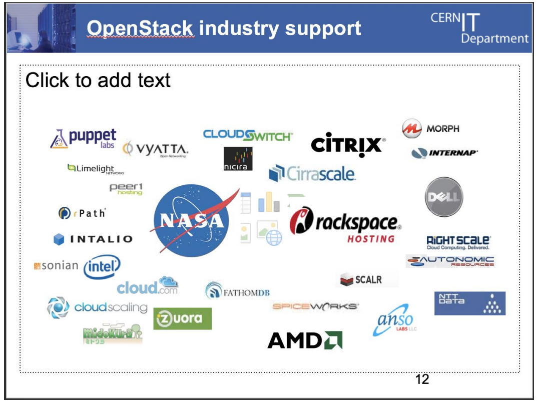
What is OpenStack?

OpenStack is a collection of open source technology projects that provides an operating platform for orchestrating clouds in a massively scale.



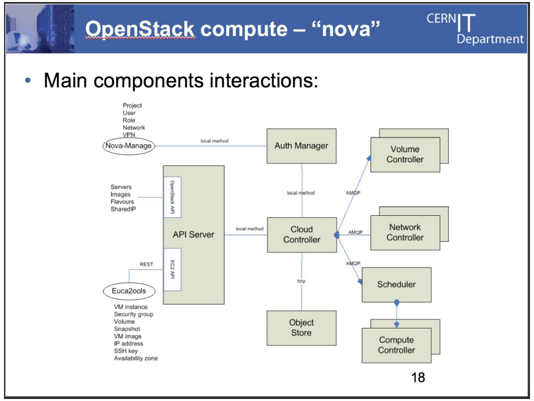
OpenStack history

- March 2010
  - Rackspace open source their cloud software;
- May 2010
  - NASA open source the Nebula software;
- June 2010
  - OpenStack formed;
- October 2010
  - “austin”release;
- February 2011
  - “bexar” release;



OpenStack industry support

Click to add text

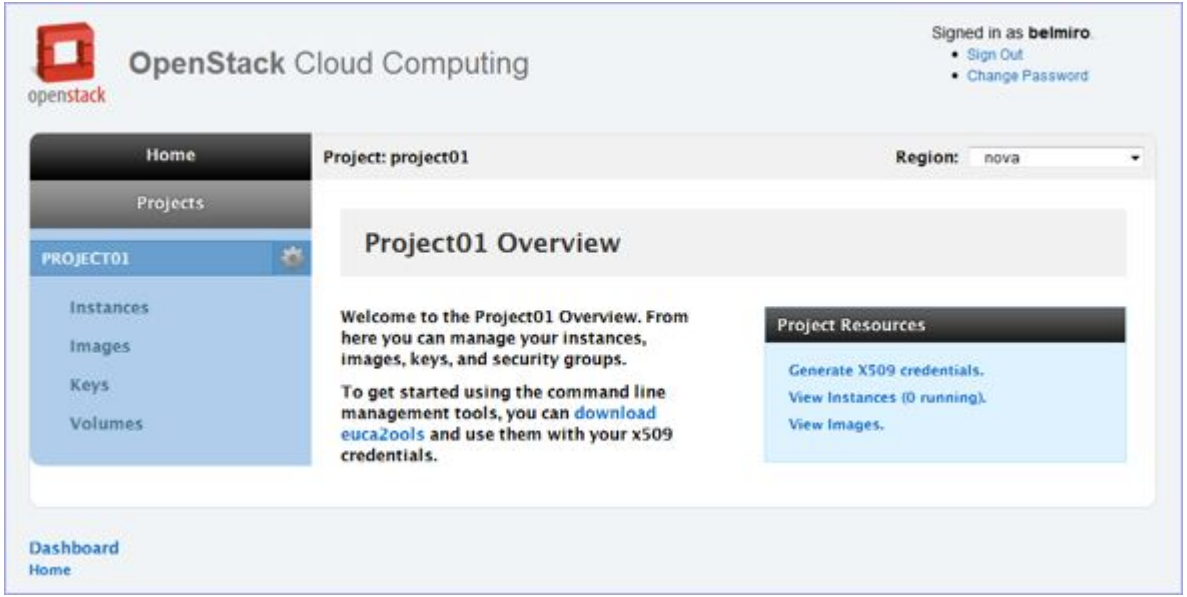


OpenStack compute – “nova”

- Main components interactions:

OpenStack at CERN - Early days

- First version of OpenStack Horizon



Welcome to the Project01 Overview. From here you can manage your instances, images, keys, and security groups.

To get started using the command line management tools, you can download euca2ools and use them with your x509 credentials.

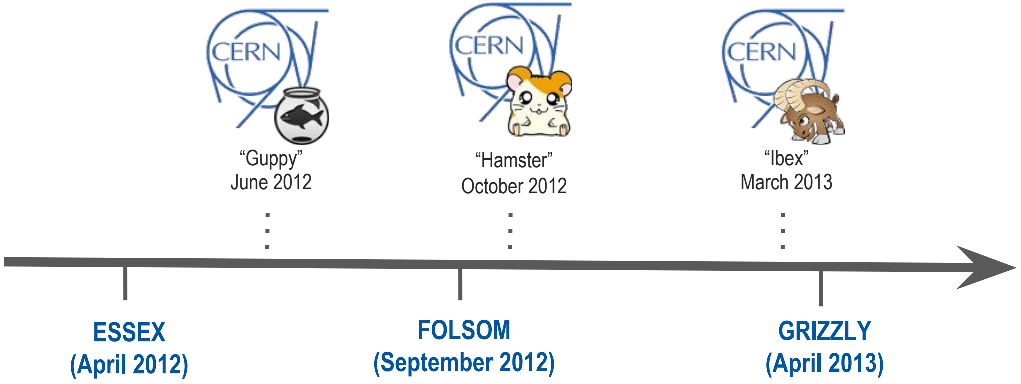
Generate X509 credentials.

View Instances (0 running).

View Images.

Prototyping CERN OpenStack Cloud

- Iterate Fast...
  - Build test infrastructures and open then to early adopters
  - Few hundred nodes available
  - 2 different virtualization technologies (KVM, Hyper-V)
  - Integration with other Agile Infrastructure projects (puppet, monitoring, ...)



"Guppy"
June 2012

"Hamster"
October 2012

"Ibex"
March 2013

ESSEX
(April 2012)

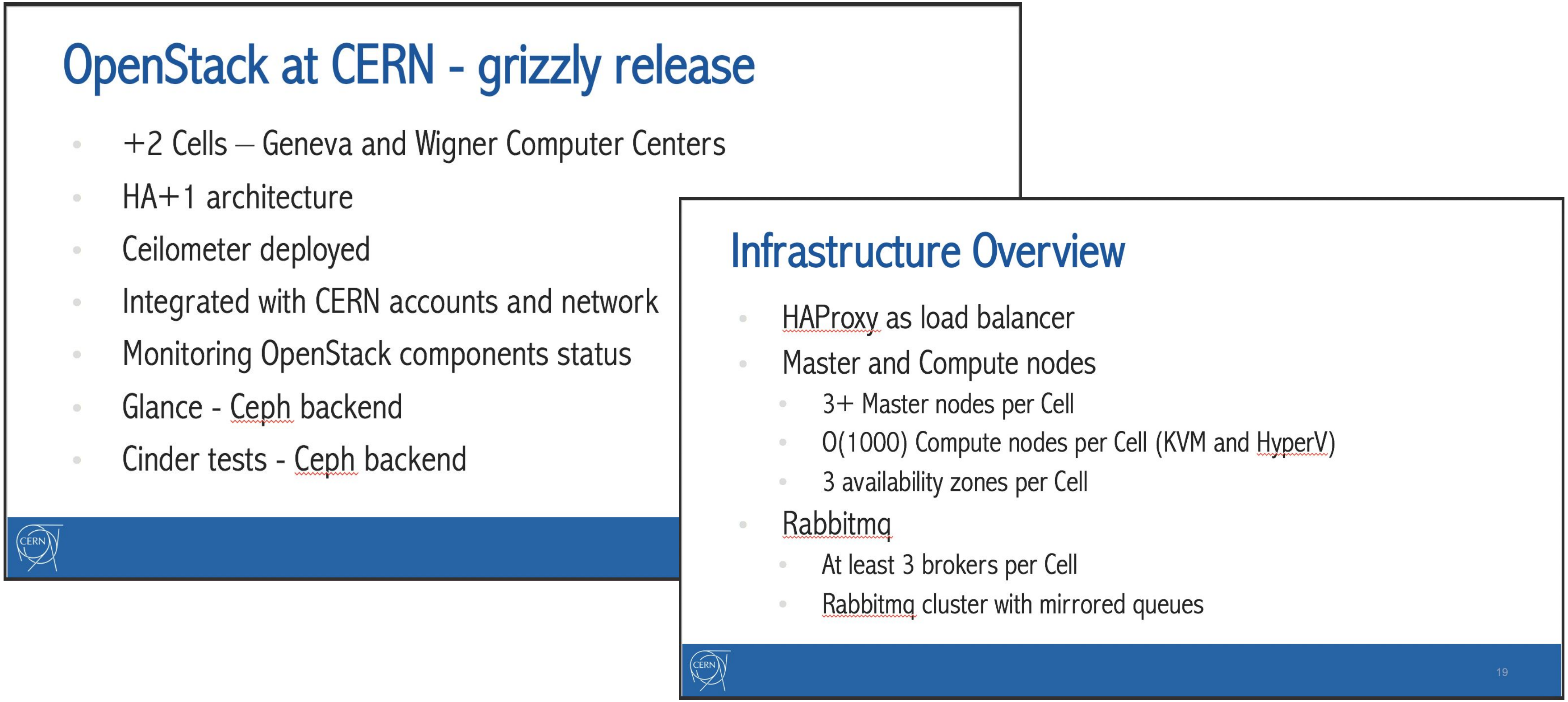
FOLSOM
(September 2012)

GRIZZLY
(April 2013)

2013 - 2019

From 0 to +300k cores

CERN OpenStack Cloud - 2013



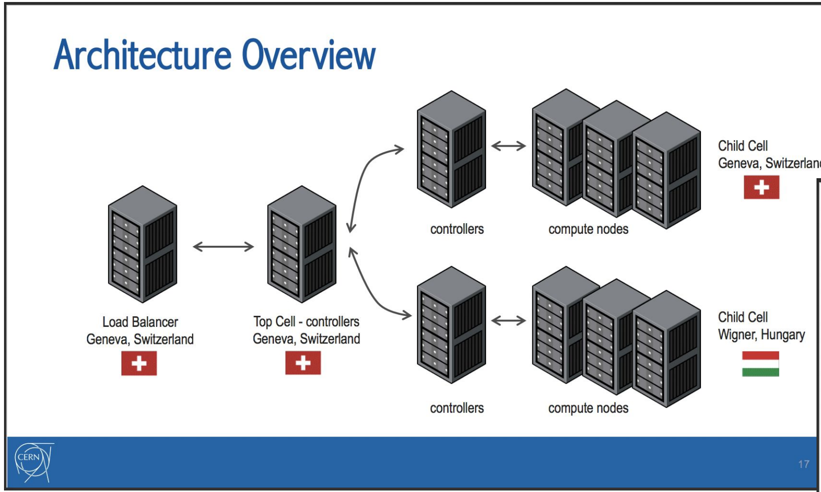
OpenStack at CERN - grizzly release

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

Infrastructure Overview

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

CERN OpenStack Cloud - 2013



Architecture Overview

Load Balancer
Geneva, Switzerland

Top Cell - controllers
Geneva, Switzerland

controllers

compute nodes

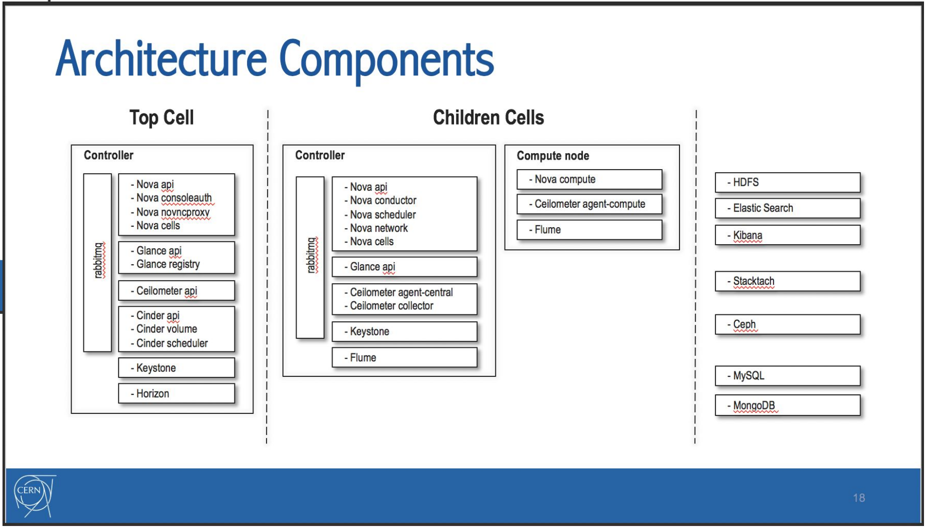
Child Cell
Geneva, Switzerland

controllers

compute nodes

Child Cell
Wigner, Hungary

17



Architecture Components

Top Cell

Controller

rabbitmq

- Nova api
- Nova consoleauth
- Nova novncproxy
- Nova cells

- Glance api
- Glance registry

- Ceilometer api

- Cinder api
- Cinder volume
- Cinder scheduler

- Keystone

- Horizon

Children Cells

Controller

rabbitmq

- Nova api
- Nova conductor
- Nova scheduler
- Nova network
- Nova cells

- Glance api

- Ceilometer agent-central
- Ceilometer collector

- Keystone

- Flume

Compute node

- Nova compute

- Ceilometer agent-compute

- Flume

- HDFS

- Elastic Search

- Kibana

- Stacktach

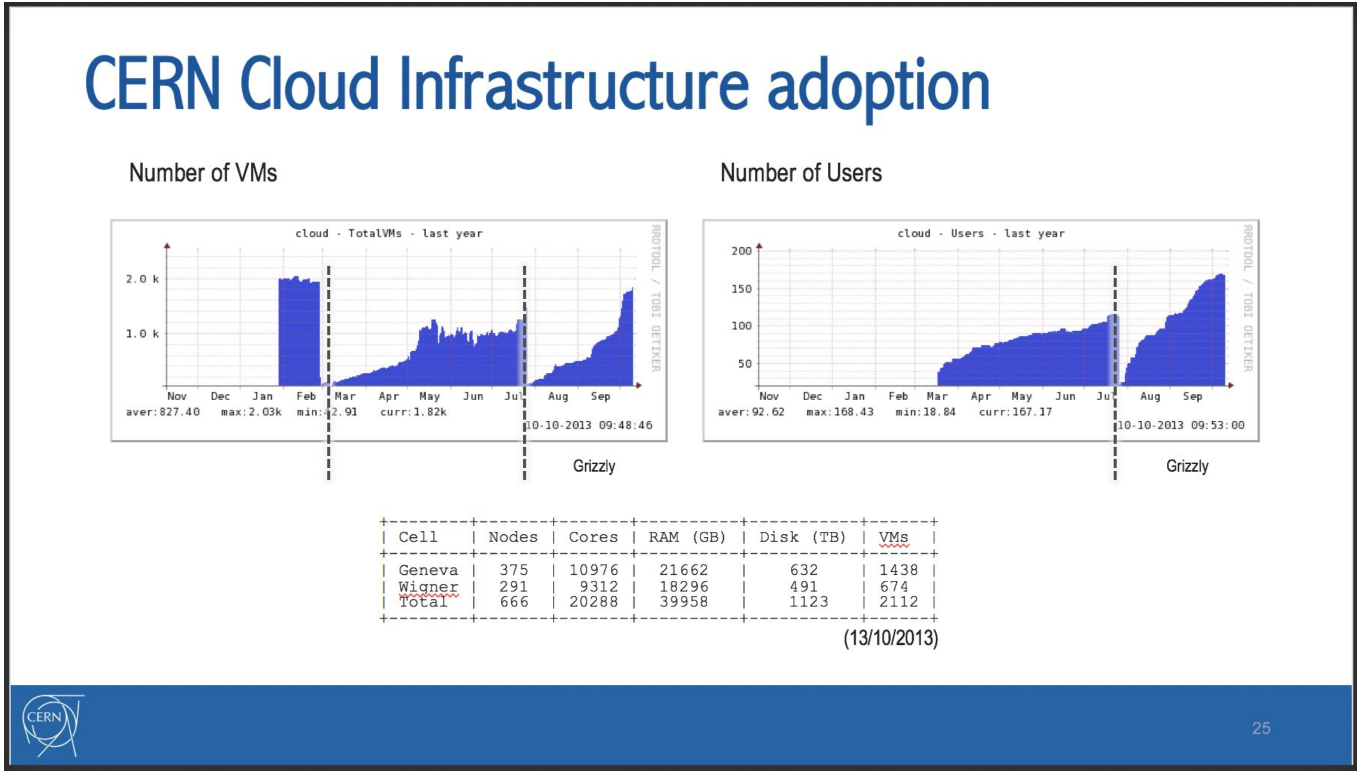
- Ceph

- MySQL

- MongoDB

18

CERN OpenStack Cloud - 2013



CERN Cloud Infrastructure adoption

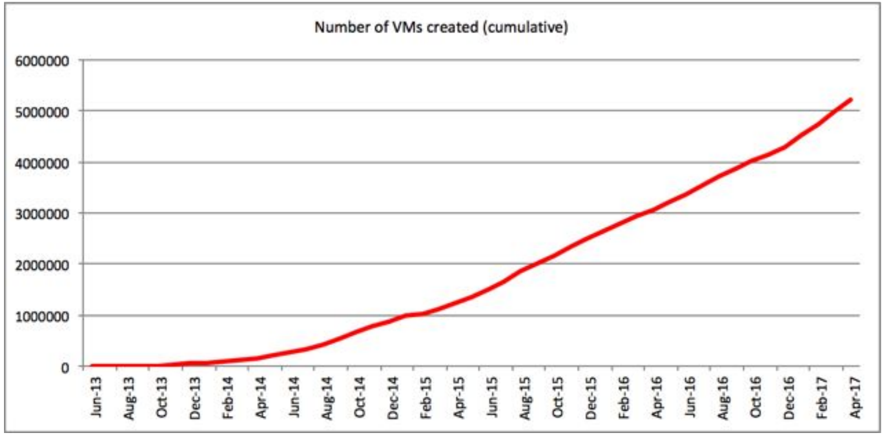
Number of VMs

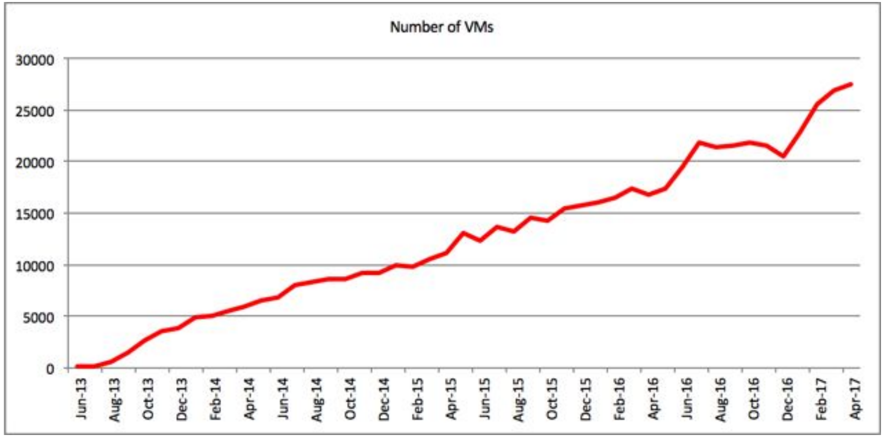
Number of Users

```
+---------+--------+--------+----------+-----------+------+
| Cell    | Nodes  | Cores  | RAM (GB) | Disk (TB) | VMs  |
+---------+--------+--------+----------+-----------+------+
| Geneva  |  375   | 10976  |  21662   |    632    | 1438 |
| Wigner  |  291   |  9312  |  18296   |    491    | 674  |
| Total   |  666   | 20288  |  39958   |   1123    | 2112 |
+---------+--------+--------+----------+-----------+------+
```

(13/10/2013)

CERN OpenStack Cloud - Growth





CERN OpenStack Cloud - Growth

- OpenStack projects available in the CERN Cloud over releases

Grizzly	Havana	Icehouse	Juno	Kilo	Liberty	Mitaka	Newton	Ocata	Pike	Queens	Rocky
Nova	Nova	Nova	Nova	Nova	Nova	Nova	Nova	Nova	Nova	Nova	Nova
Glance	Glance	Glance	Glance	Glance	Glance	Glance	Glance	Glance	Glance	Glance	Glance
Horizon	Horizon	Horizon	Horizon	Horizon	Horizon	Horizon	Horizon	Horizon	Horizon	Horizon	Horizon
Keystone	Keystone	Keystone	Keystone	Keystone	Keystone	Keystone	Keystone	Keystone	Keystone	Keystone	Keystone
Ceilometer *	Ceilometer *	Ceilometer	Ceilometer	Ceilometer	Ceilometer	~~Ceilometer~~	~~Ceilometer~~	~~Ceilometer~~	~~Ceilometer~~	~~Ceilometer~~	~~Ceilometer~~
		Cinder	Cinder	Cinder	Cinder	Cinder	Cinder	Cinder	Cinder	Cinder	Cinder
			Heat *	Heat	Heat	Heat	Heat	Heat	Heat	Heat	Heat
			Rally *	Rally	Rally	Rally	Rally	Rally	Rally	Rally	Rally
					EC2API	EC2API	EC2API	EC2API	EC2API	EC2API	EC2API
					Magnum *	Magnum	Magnum	Magnum	Magnum	Magnum	Magnum
					Barbican *	Barbican	Barbican	Barbican	Barbican	Barbican	Barbican
					Neutron *	Neutron	Neutron	Neutron	Neutron	Neutron	Neutron
						Ironic ?	Ironic ?	Ironic ?	Ironic *	Ironic	Ironic
						Mistral ?	Mistral ?	Mistral ?	Mistral *	Mistral	Mistral
						Manila ?	Manila ?	Manila *	Manila *	Manila	Manila
								Trove ?		Qinling ?	Qinling ?
								Murano ?		Watcher ?	Watcher ?

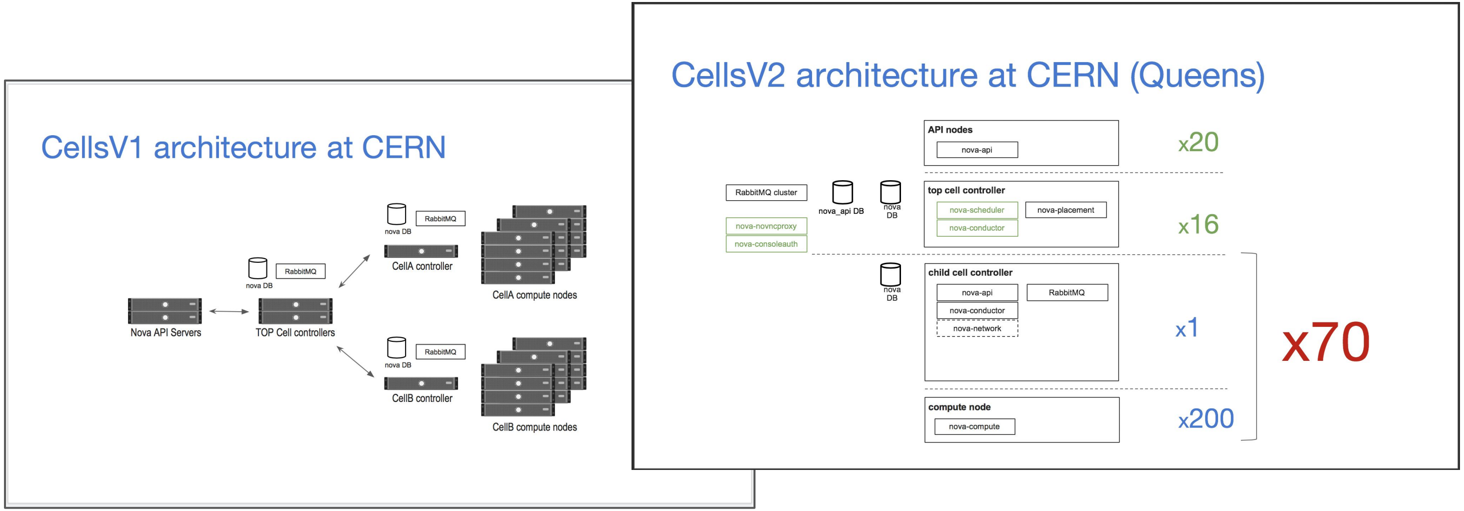
* - Pilot service

? - Trial service

Milestone Highlights

Nova - Cells

- Allows Nova to scale to thousands of compute nodes
- Biggest Nova Cells deployment
- Moved from 2 cells to +70 cells
- Upgrade from CellsV1 to CellsV2 in 2018

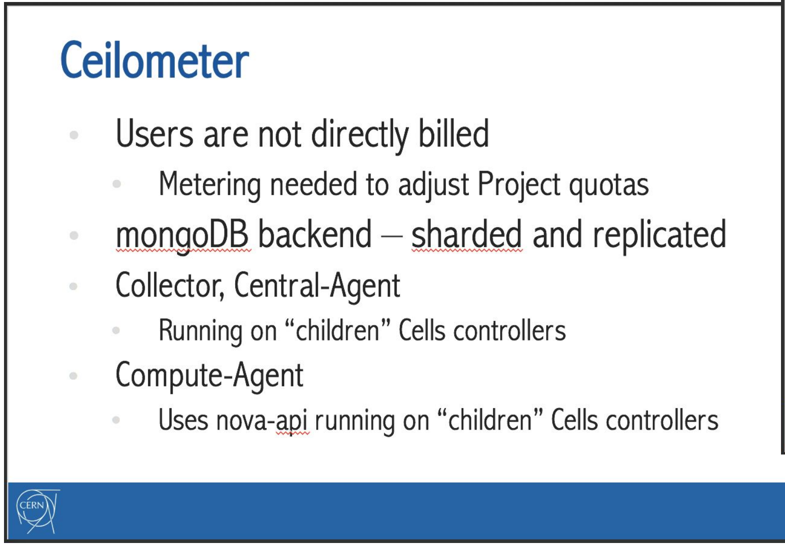


CellsV1 architecture at CERN

CellsV2 architecture at CERN (Queens)

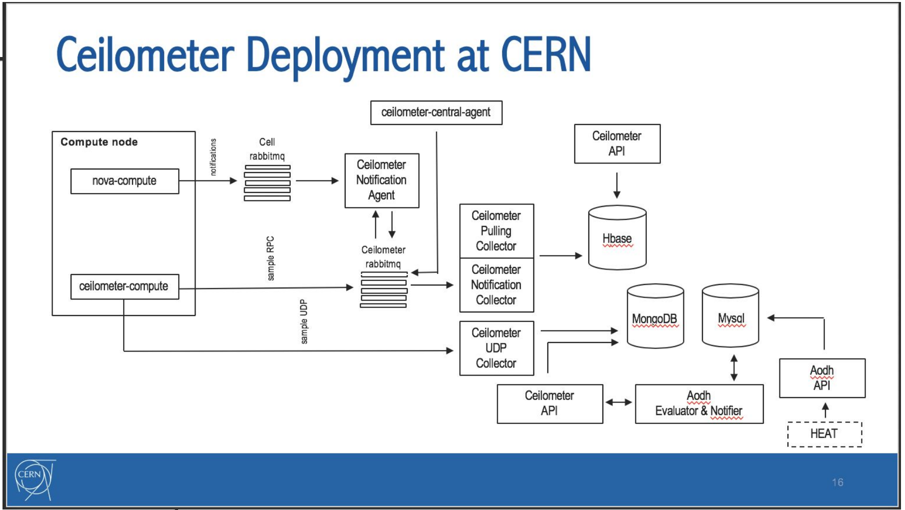
Ceilometer - The Rise & Fall

- OpenStack Ceilometer deployed
- Removed after run it for 3 years. Not scalable and difficult to retrieve data



Ceilometer

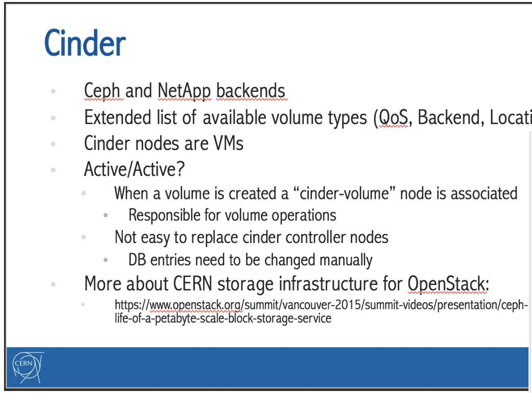
- Users are not directly billed
  - Metering needed to adjust Project quotas
- mongoDB backend – sharded and replicated
- Collector, Central-Agent
  - Running on “children” Cells controllers
- Compute-Agent
  - Uses nova-api running on “children” Cells controllers



Ceilometer Deployment at CERN

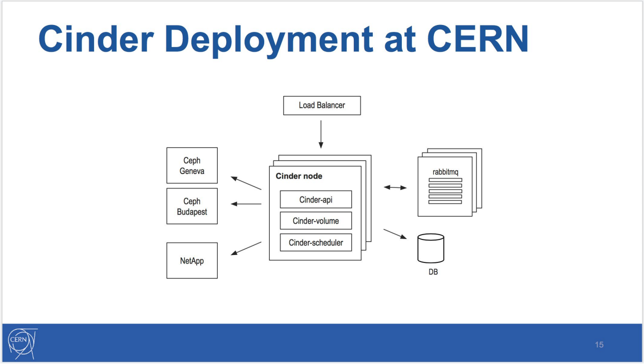
Storage - Cinder, Manila, S3

- OpenStack Cinder with Ceph backend (2014)
  - Several volume types available
- OpenStack Manila (Fileshare service). Backed by CephFS (2017)
- S3 available (end 2018)



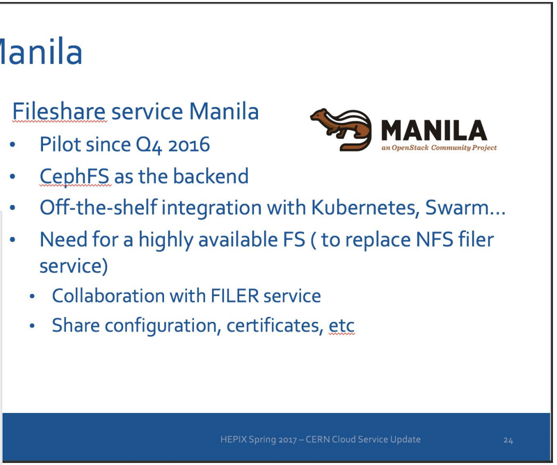
Cinder

- Ceph and NetApp backends
- Extended list of available volume types (QoS, Backend, Location)
- Cinder nodes are VMs
- Active/Active?
  - When a volume is created a "cinder-volume" node is associated
    - Responsible for volume operations
  - Not easy to replace cinder controller nodes
    - DB entries need to be changed manually
- More about CERN storage infrastructure for OpenStack:
  - https://www.openstack.org/summit/vancouver-2015/summit-videos/presentation/ceph-life-of-a-petabyte-scale-block-storage-service



Cinder Deployment at CERN

15



Manila

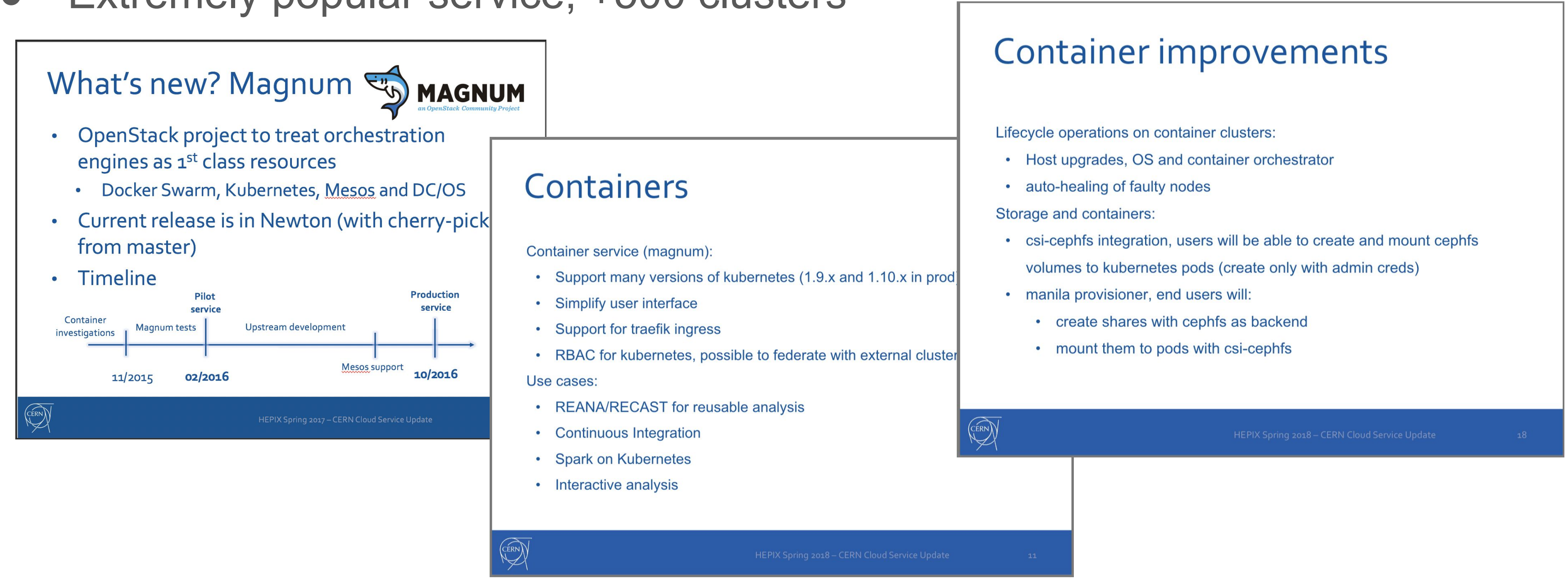
- Fileshare service Manila
  - Pilot since Q4 2016
  - CephFS as the backend
  - Off-the-shelf integration with Kubernetes, Swarm...
  - Need for a highly available FS (to replace NFS filer service)
    - Collaboration with FILER service
    - Share configuration, certificates, etc

MANILA an OpenStack Community Project

HEPIX Spring 2017 – CERN Cloud Service Update 24

Container Orchestration - Magnum

- OpenStack Magnum service available since 2016
- Extremely popular service, +500 clusters



What’s new? Magnum

MAGNUM
an OpenStack Community Project

- OpenStack project to treat orchestration engines as 1st class resources
  - Docker Swarm, Kubernetes, Mesos and DC/OS
- Current release is in Newton (with cherry-pick from master)
- Timeline

Container investigations
Magnum tests
Pilot service
Upstream development
Production service
Mesos support
11/2015
02/2016
10/2016

HEPIX Spring 2017 – CERN Cloud Service Update

Containers

Container service (magnum):

- Support many versions of kubernetes (1.9.x and 1.10.x in prod)
- Simplify user interface
- Support for traefik ingress
- RBAC for kubernetes, possible to federate with external cluster

Use cases:

- REANA/RECAST for reusable analysis
- Continuous Integration
- Spark on Kubernetes
- Interactive analysis

HEPIX Spring 2018 – CERN Cloud Service Update 11

Container improvements

Lifecycle operations on container clusters:

- Host upgrades, OS and container orchestrator
- auto-healing of faulty nodes

Storage and containers:

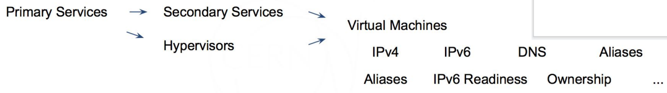
- csi-cephfs integration, users will be able to create and mount cephfs volumes to kubernetes pods (create only with admin creds)
- manila provisioner, end users will:
  - create shares with cephfs as backend
  - mount them to pods with csi-cephfs

HEPIX Spring 2018 – CERN Cloud Service Update 18

Networking - Nova-network to Neutron

LanDB

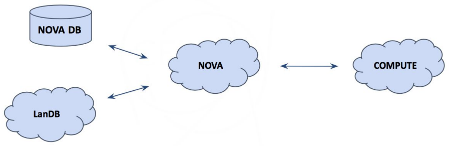
Source of Truth



- All devices must be present
- Used for different purposes
  - Security checks
  - DNS/DHCP Configuration
  - Switch/router configuration
  - Active Directory, …

Phase 1. Nova Network

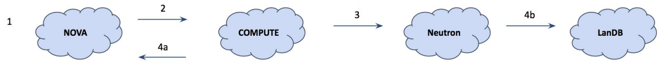
- Custom *NetworkManager*
- Late IP allocation - after scheduling to compute nodes
- Patching done directly in the Nova code



- Nova Network is being deprecated...
  - Quantum is the new thing… Neutron is the new thing...

Phase 2. Neutron

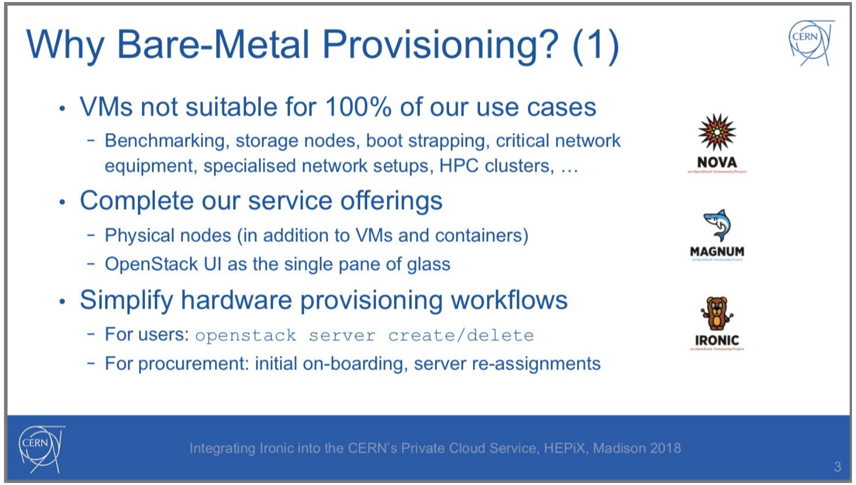
- Linuxbridge, Flat / Provider networks
- Better integration using ML2, **mechanism driver** and **extensions**
  - Quickly became possible to have it out of tree
  - Our extensions have a similar role to Neutron Segments
- Gradual enroll, cell by cell
- Vanilla upstream packages for Neutron, much smaller patch on Nova
- More split pieces, potential points of failure
  - Periodic consistency checks



https://gitlab.cern.ch/cloud-infrastructure/openstack-neutron-cern

Baremetal Provisioning - Ironic

- In production since 2018
- All new hardware is enrolled using Ironic. +1700 nodes managed by Ironic
- Existing hardware will be enrolled into Ironic during 2019

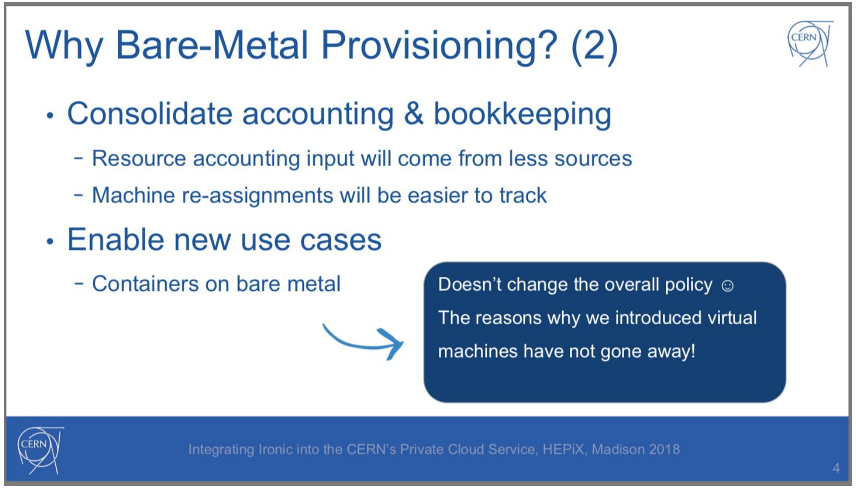


Why Bare-Metal Provisioning? (1)

- VMs not suitable for 100% of our use cases
  - Benchmarking, storage nodes, boot strapping, critical network equipment, specialised network setups, HPC clusters, …
- Complete our service offerings
  - Physical nodes (in addition to VMs and containers)
  - OpenStack UI as the single pane of glass
- Simplify hardware provisioning workflows
  - For users: `openstack server create/delete`
  - For procurement: initial on-boarding, server re-assignments

Integrating Ironic into the CERN's Private Cloud Service, HEPiX, Madison 2018

3



Why Bare-Metal Provisioning? (2)

- Consolidate accounting & bookkeeping
  - Resource accounting input will come from less sources
  - Machine re-assignments will be easier to track
- Enable new use cases
  - Containers on bare metal

Doesn't change the overall policy ☺

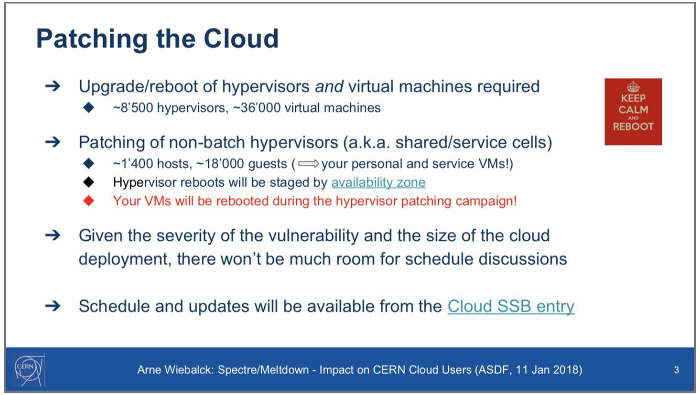
The reasons why we introduced virtual machines have not gone away!

Integrating Ironic into the CERN's Private Cloud Service, HEPiX, Madison 2018

4

Meltdown/Spectre/L1TF

- Reboot campaigns and performance impact

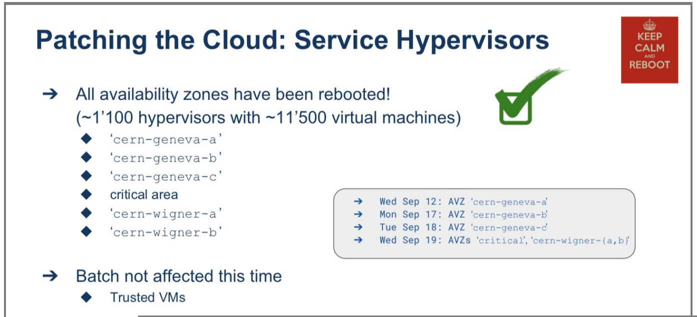


Patching the Cloud

KEEP CALM AND REBOOT

- ➔ Upgrade/reboot of hypervisors *and* virtual machines required
  - ◆ ~8’500 hypervisors, ~36’000 virtual machines
- ➔ Patching of non-batch hypervisors (a.k.a. shared/service cells)
  - ◆ ~1’400 hosts, ~18’000 guests (⟹your personal and service VMs!)
  - ◆ Hypervisor reboots will be staged by availability zone
  - ◆ Your VMs will be rebooted during the hypervisor patching campaign!
- ➔ Given the severity of the vulnerability and the size of the cloud deployment, there won’t be much room for schedule discussions
- ➔ Schedule and updates will be available from the Cloud SSB entry

Arne Wiebalck: Spectre/Meltdown - Impact on CERN Cloud Users (ASDF, 11 Jan 2018) 3



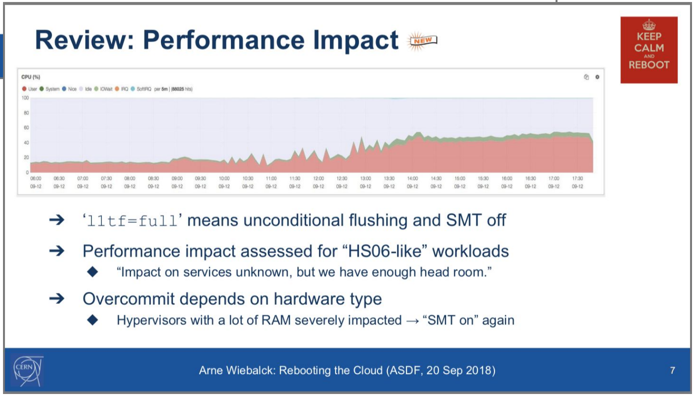
Patching the Cloud: Service Hypervisors

KEEP CALM AND REBOOT

- ➔ All availability zones have been rebooted!
  (~1’100 hypervisors with ~11’500 virtual machines)
  - ◆ ‘cern-geneva-a’
  - ◆ ‘cern-geneva-b’
  - ◆ ‘cern-geneva-c’
  - ◆ critical area
  - ◆ ‘cern-wigner-a’
  - ◆ ‘cern-wigner-b’

```
➔ Wed Sep 12: AVZ ‘cern-geneva-a’
➔ Mon Sep 17: AVZ ‘cern-geneva-b’
➔ Tue Sep 18: AVZ ‘cern-geneva-c’
➔ Wed Sep 19: AVZs ‘critical’,‘cern-wigner-{a,b}’
```

- ➔ Batch not affected this time
  - ◆ Trusted VMs



Review: Performance Impact NEW

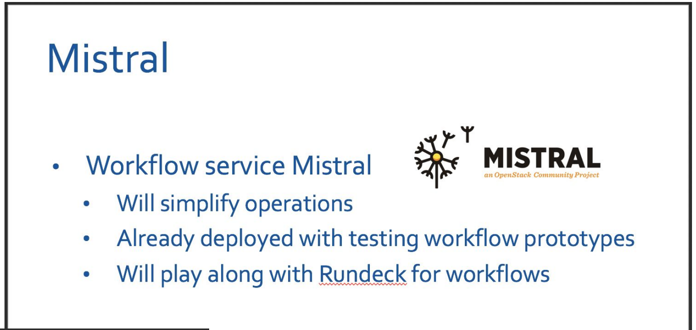
KEEP CALM AND REBOOT

- ➔ ‘`l1tf=full`’ means unconditional flushing and SMT off
- ➔ Performance impact assessed for “HS06-like” workloads
  - ◆ “Impact on services unknown, but we have enough head room.”
- ➔ Overcommit depends on hardware type
  - ◆ Hypervisors with a lot of RAM severely impacted → “SMT on” again

Arne Wiebalck: Rebooting the Cloud (ASDF, 20 Sep 2018) 7

Operations - Rundeck and Mistral

- OpenStack Mistral
- RunDeck

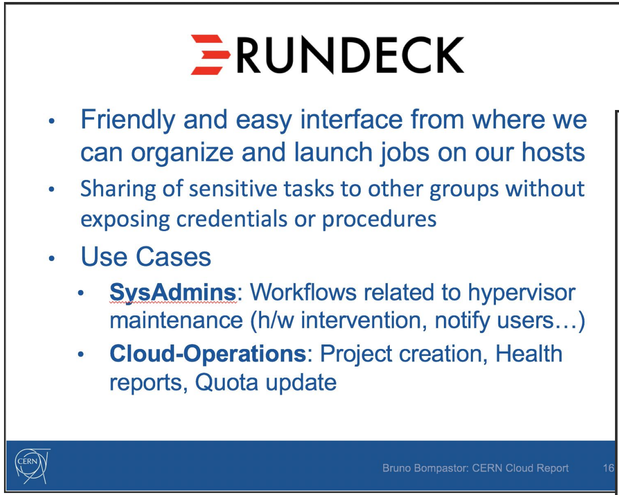


Mistral

- Workflow service Mistral
  - Will simplify operations
  - Already deployed with testing workflow prototypes
  - Will play along with Rundeck for workflows

MISTRAL
an OpenStack Community Project

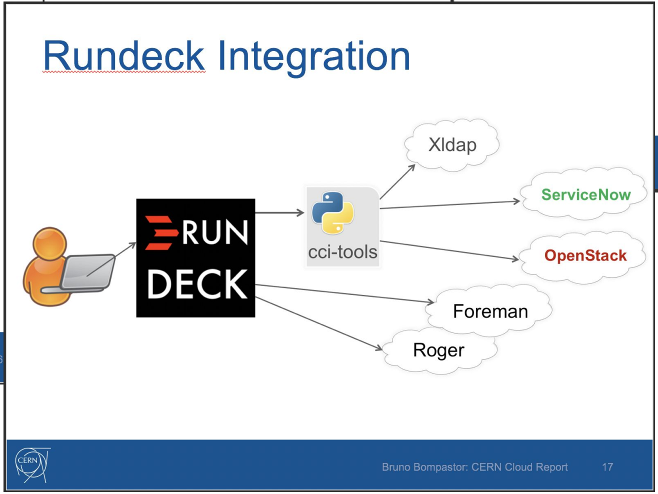
HEPIX Spring 2017 – CERN Cloud Service Update 23



RUNDECK

- Friendly and easy interface from where we can organize and launch jobs on our hosts
- Sharing of sensitive tasks to other groups without exposing credentials or procedures
- Use Cases
  - **SysAdmins**: Workflows related to hypervisor maintenance (h/w intervention, notify users…)
  - **Cloud-Operations**: Project creation, Health reports, Quota update

Bruno Bompastor: CERN Cloud Report 16



Rundeck Integration

Bruno Bompastor: CERN Cloud Report 17

Operations

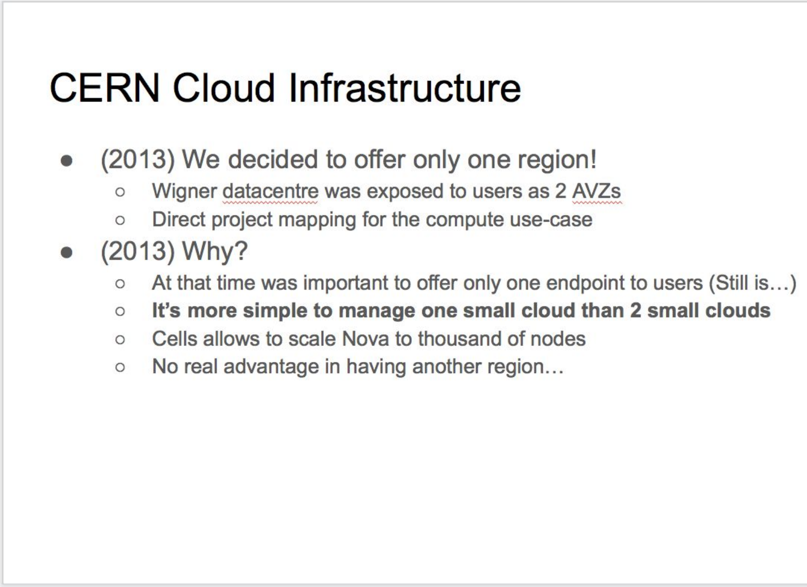
- Experience growing/managing the Infrastructure during the last 6 years
- Several upgrades during this journey
  - OpenStack release cycle is every 6 months!
  - SLC6 to CC7 upgrade
  - CC7 upgrades
- Supported for few years KVM and HyperV in the same infrastructure
  - Migrated CVI VMs to OpenStack HyperV and then to OpenStack KVM
- Security updates required reboot of all cloud
- Most user management operations are automated
  - project creation; quotas; ...
  - VM expiration

2019 - ...

What’s next?

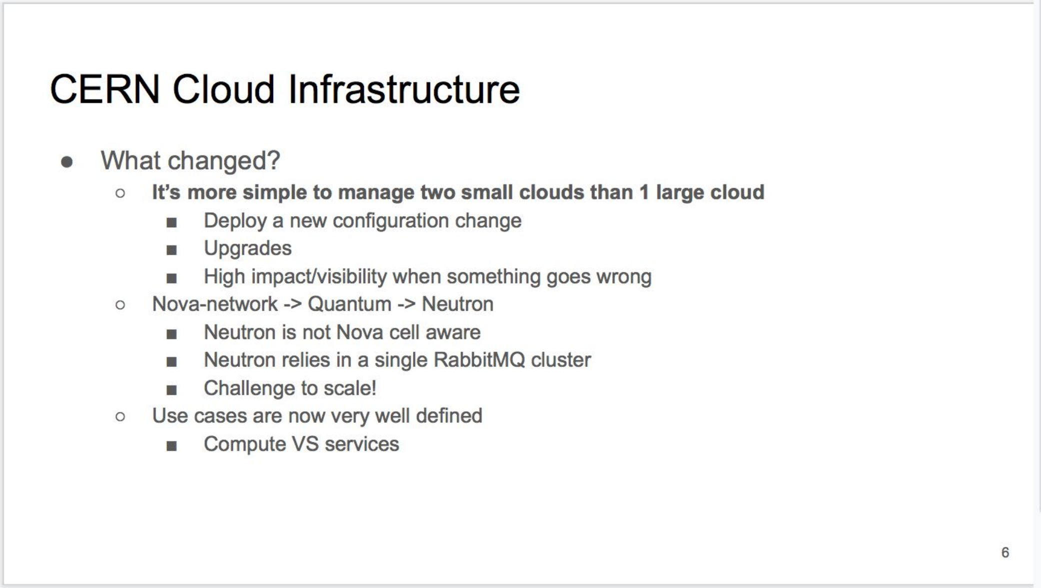
Splitting the Infrastructure into 2 Regions

https://techblog.web.cern.ch/techblog/post/region-split/



CERN Cloud Infrastructure

- (2013) We decided to offer only one region!
  - Wigner datacentre was exposed to users as 2 AVZs
  - Direct project mapping for the compute use-case
- (2013) Why?
  - At that time was important to offer only one endpoint to users (Still is…)
  - **It's more simple to manage one small cloud than 2 small clouds**
  - Cells allows to scale Nova to thousand of nodes
  - No real advantage in having another region…



CERN Cloud Infrastructure

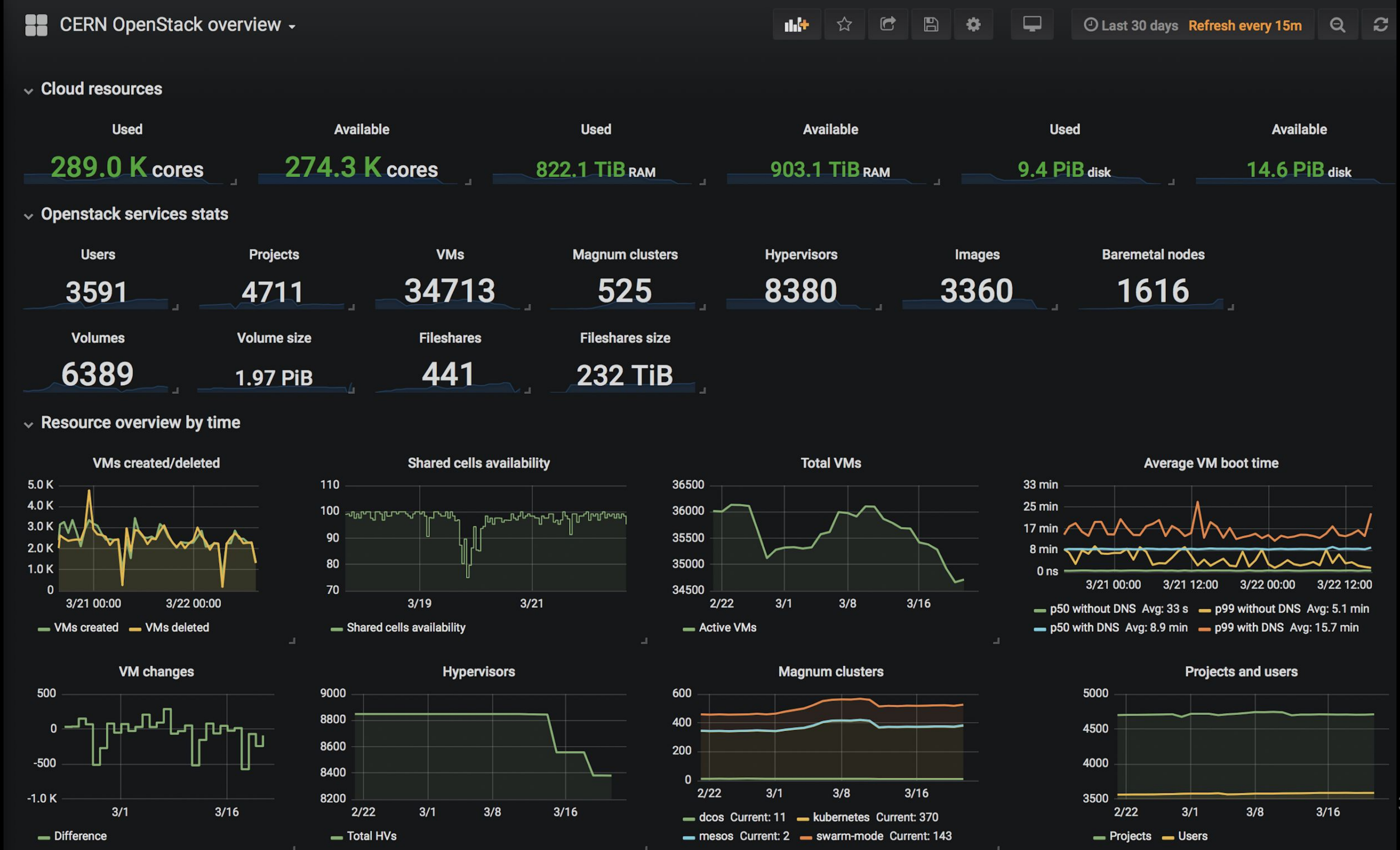
- What changed?
  - **It's more simple to manage two small clouds than 1 large cloud**
    - Deploy a new configuration change
    - Upgrades
    - High impact/visibility when something goes wrong
  - Nova-network -> Quantum -> Neutron
    - Neutron is not Nova cell aware
    - Neutron relies in a single RabbitMQ cluster
    - Challenge to scale!
  - Use cases are now very well defined
    - Compute VS services

Preemptible Instances

- Public Clouds
  - Based on different pricing/SLA considering resource availability
  - Reserved instances vs spot-market
- Private Clouds
  - Quotas are hard limits. Leads to a reduction in resource utilization
  - Preemptible instances
    - Projects that exhausted their quota can continue to create instances
      - Opportunistic workloads
      - Low SLA
- Preemptible Instances Workflow in OpenStack Nova
  - The creation of a non preemptible VM fails because there aren’t available resources
  - Instances that fail with “Nova Valid Host”, go to “PENDING” state instead of “ERROR”
  - The Reaper service is notified and it tries to free the requested resources
    - Rebuild the instance
    - Or change instance state to “ERROR”

Other Challenges

- Leveraging Container Orchestration to deploy OpenStack control plane
- Re-enroll existing physical resources into OpenStack Ironic
- Introduce SDN
- Dynamic resource provisioning based in Compute Nodes load



CERN OpenStack overview

Cloud resources

Used	Available	Used	Available	Used	Available
289.0 K cores	274.3 K cores	822.1 TiB RAM	903.1 TiB RAM	9.4 PiB disk	14.6 PiB disk

Openstack services stats

Users	Projects	VMs	Magnum clusters	Hypervisors	Images	Baremetal nodes
3591	4711	34713	525	8380	3360	1616

Volumes	Volume size	Fileshares	Fileshares size
6389	1.97 PiB	441	232 TiB

Resource overview by time

VMs created/deleted

Shared cells availability

Total VMs

Average VM boot time

p50 without DNS Avg: 33 s
p99 without DNS Avg: 5.1 min
p50 with DNS Avg: 8.9 min
p99 with DNS Avg: 15.7 min

VM changes

Hypervisors

Magnum clusters

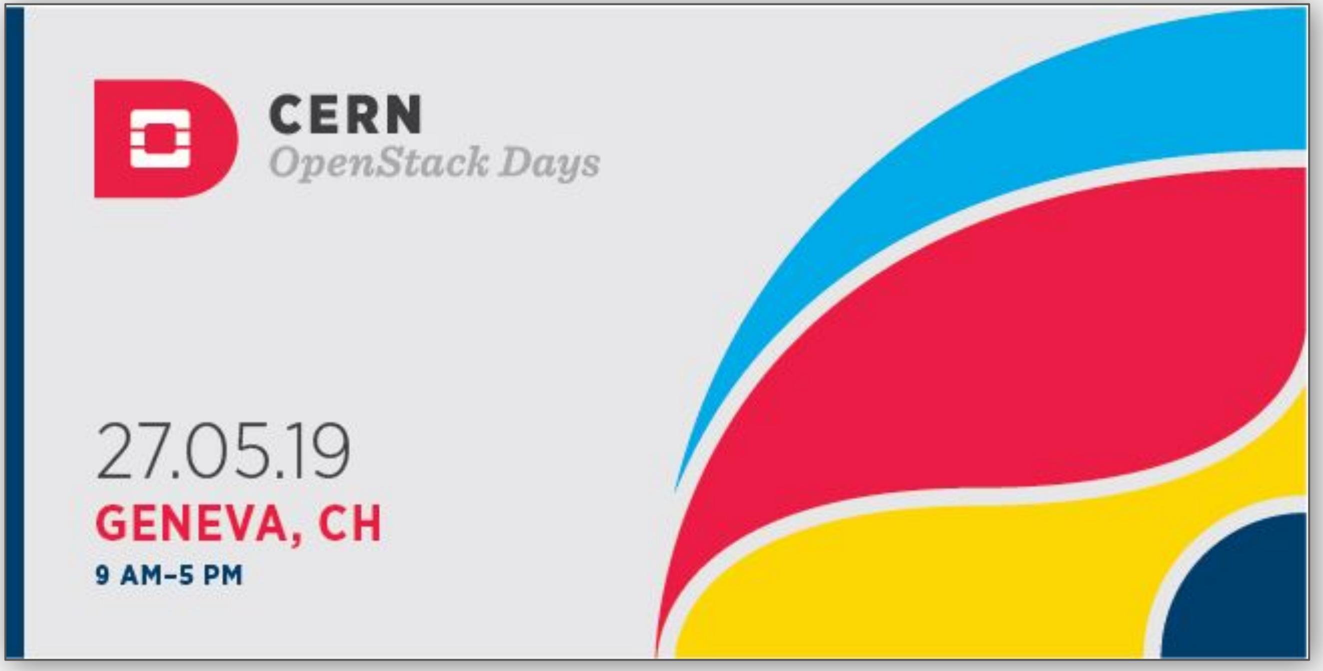
dcos Current: 11
kubernetes Current: 370
mesos Current: 2
swarm-mode Current: 143

Projects and users

Summary

- During the last 10 years, resource management and deployment model changed completely
  - From Virtualization and Server consolidation to a Cloud Infrastructure
  - From Baremetal to VMs, to managed Baremetal to Containers
- Continue to adapt the Infrastructure to the new technologies and requirements
  - Iterative approach to introduce new services, new functionality
  - Continue to explore new approaches to deploy/manage a large infrastructure
    - Control Plane managed by kubernetes
    - New regions
    - SDN

Is serverless the new model?



https://openstackdayscern.web.cern.ch

@belmiromoreira

CERN

www.cern.ch

Credits

- Used slides from several authors
  - Arne Wiebalck
  - Belmiro Moreira
  - Domenico Giordano
  - Jan Van Eldik
  - Luis Pigueiras
  - Ricardo Rocha
  - Spyridon Trigazis