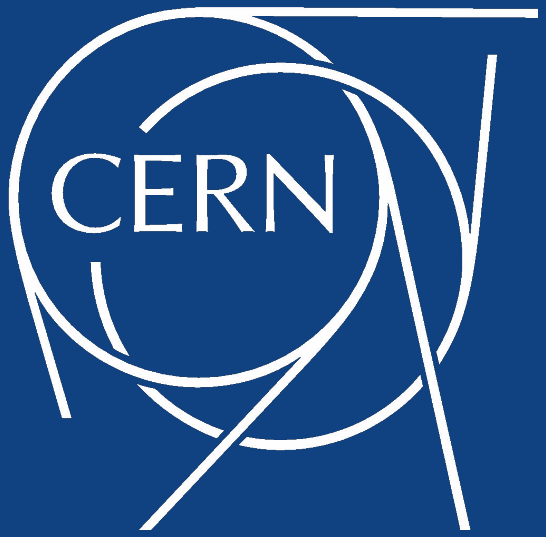

CERN

CERN Cloud Service Update

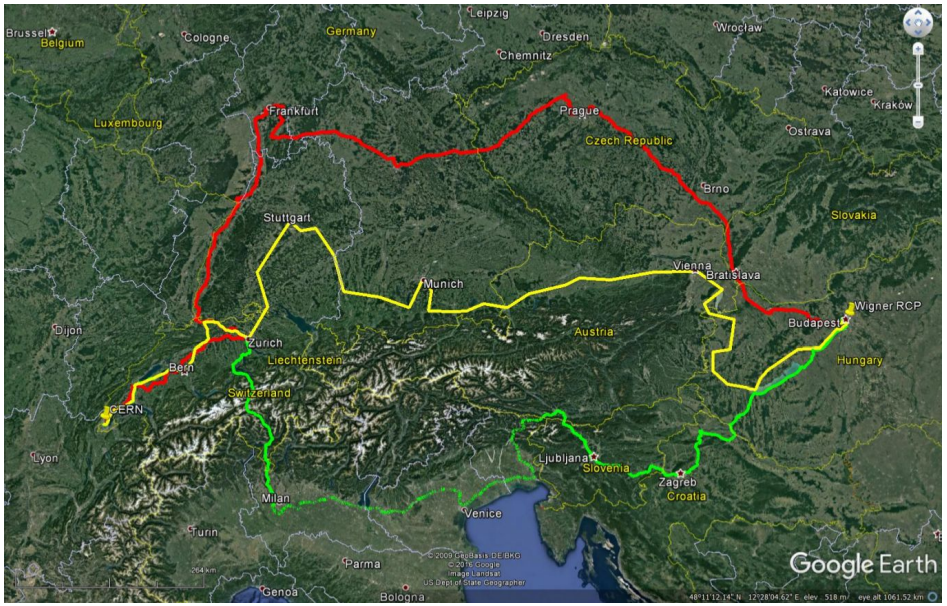
Spyros Trigazis
On behalf of the CERN Cloud Team

CERN Cloud Recap

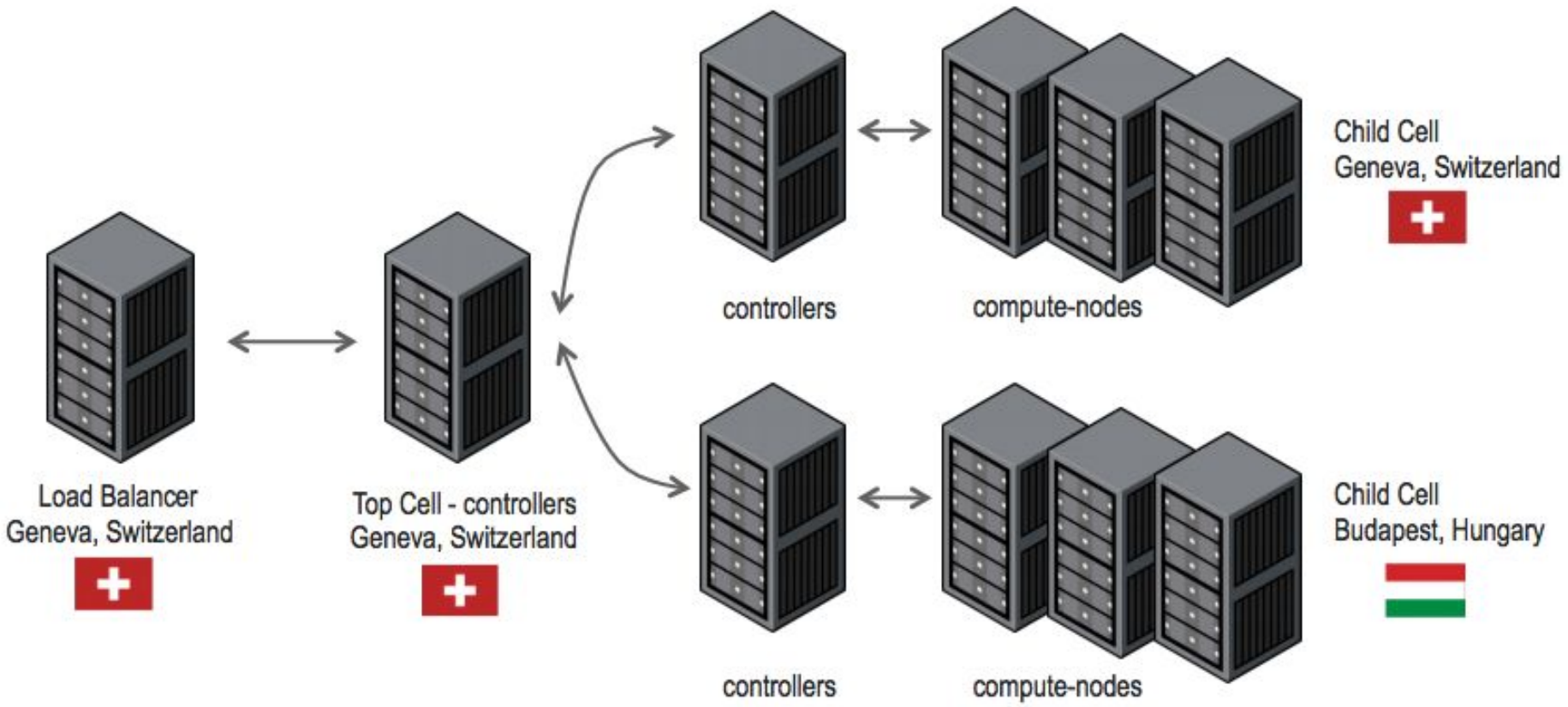
- Policy: servers in CERN IT shall be virtual
- Based on OpenStack
  - Production service since July 2013
  - Currently between Ocata/Queens, depending on the component (released in Feb 2017 and 2018 respectively)

CERN Cloud Architecture

- Composed by two data centers
  - 1 region (1 API) and 70 cells
  - Cells map different use cases
    - Shared: 5 AVZ
    - Project cells: special requirements (hardware, location)
    - Batch cells: optimized for batch workload (90% of HW)



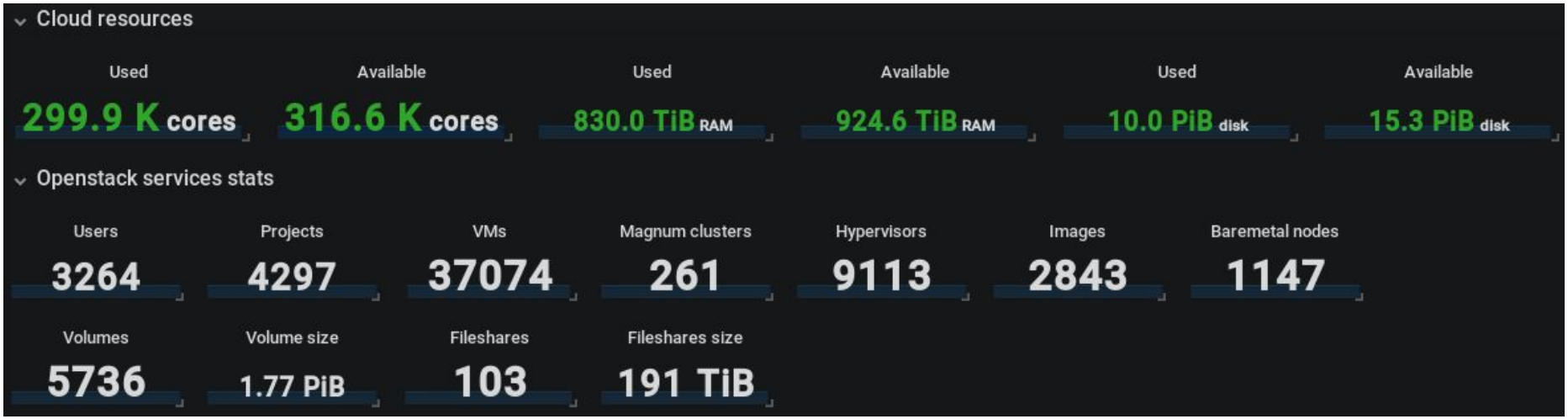
CERN Cloud Architecture



CERN Cloud in numbers

9000 hypervisors (7K 1yr ago)
320K cores (210k)
830 TB of RAM (350 TB)
261 container clusters (Magnum) (67)

5.7K volumes with 1.8 PB allocated (3.7K)
3.8K images/snapshots (3.8K)
102 fileshares with 18 TB allocated (27)



Operations and Activities

Compute & Network

- Upgraded to Queens, latest stable OpenStack release, Feb 2018
- Major migration from Cells V1 to Cells V2
  - Closer to upstream and vanilla OpenStack
  - Scale Placement component to accept load from 9k hypervisors [1]
- Reboot of the whole cloud to address Meltdown and Spectre [2]
- Standardize migration process from legacy nova-network to neutron [6]
- Available in 34 cells (10 1yr ago)
  - 4.7k hypervisors and 18k ports
  - scale message broker to 4.7k producers

Baremetal

Offer physical servers with OpenStack APIs

- In production since Q3 2017
- Consolidate accounting of resources
- All new deliveries are handed over with Ironic

Use Cases

- Cloud, all hypervisors are enrolled with Ironic and plan to migrate existing hypervisors
- HPC
- Windows infrastructure
- Databases
- Experiments

Storage

Block Storage service (cinder):

- Upgraded to the latest Queens release
- 30% of user VMs have volumes mounted for data
- support IOPS burst, significant improvement for windows VMs

Fileshare service (manila):

- In production since Q3 2017
- backend on cephfs
- to replace the filer service
- 3 backend clusters, production and two testing

Containers

Container service (magnum):

- Support many versions of kubernetes (1.9.x and 1.10.x in prod) and docker
- Simplify user interface
- Support for traefik ingress
- RBAC for kubernetes, possible to federate with external clusters [3]

Use cases:

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

Monitoring

- Rally is our heaviest user! Creates VMs/Volumes every hour

Workflows

- Introduce VM expiration
  - Implemented with OpenStack/Mistral
  - Personal VMs expire in 6 months
  - Users can prolong the expiration
- Management with Rundeck
  - 3 projects with different tasks and ACLs (admins, operators, management)
  - approval/creation/deletion of new projects
  - managing hypervisor interventions
  - service consistency checks
  - more than 100 tasks and 100K job executions

Future Plans

Pre-emptible Instances

Motivation:

- Better utilization of idle resources - elasticity [4]
- Fill hypervisors as much as possible

Work in Progress:

- Low SLA Virtual Machines that can be deleted at any time
- Introduce Reaper/Aardvark service to free up resources in two cases
  - The cloud is utilized more than a specified watermark eg 95%
  - There is no space for the standard instances

Software Defined Networks

Motivation:

- Provide Floating IPs and tenant networks
- Load Balancing and isolated setups

Investigation:

- Open vSwitch, Open Daylight, OVN, Tungsten (Open Contrail)

Work in Progress with IT Networking group:

- Deployed Tungsten in a new OpenStack region, open for few beta testers
  - Vanilla Neutron and Nova
- Better integration with our new hardware

GPUs and more Compute power

Offer Virtual Machine flavors with GPUs [5]

- Attach with PCI passthrough the GPU to the guest
- Aim to offer vGPUs (needs CentOS 7.5 and support from Nvidia)

Hyperconvergence

- Consolidate compute and storage node
- Hypervisors have a lot more space than the expected ratio to CPU and RAM
- Storage nodes are mostly CPU idle and do not consume RAM

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

Storage

Block Storage

- Support deferred deletion of volumes
  - Faster free-up of quotas
  - Possibility to undo deletion

Fileshare

- Offer snapshots of shares (needs Ceph Mimic, coming release)
- Investigate backups with Restic
- Investigate access to cephfs with nfs via nfs-ganesha

Summary

Since last year

- Container clusters and shared filesystems have more users
- Physical servers are offered with the same API
  - Operations and accounting have been improved

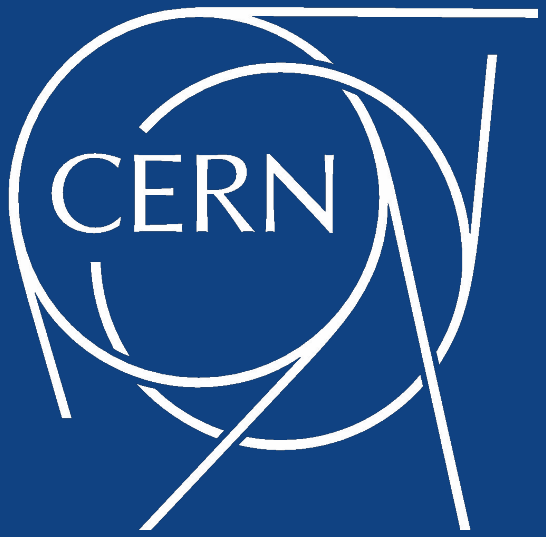
Next year we look forward to offer

- Software defined networks and GPUs
- More friendly interfaces

http://openstack-in-production.blogspot.com

References

1. https://www.openstack.org/summit/vancouver-2018/summit-schedule/events/20667/moving-from-cellsv1-to-cellsv2-at-cern
2. http://openstack-in-production.blogspot.ch/2018/01/keep-calm-and-reboot-patching-recent.html
3. https://www.youtube.com/watch?v=2PRGUOxL36M
4. http://openstack-in-production.blogspot.ch/2018/02/maximizing-resource-utilization-with.html
5. http://openstack-in-production.blogspot.ch/2018/05/introducing-gpus-to-cern-cloud.html
6. https://www.openstack.org/summit/vancouver-2018/summit-schedule/events/20767/evolution-of-openstack-networking-at-cern-nova-network-neutron-and-sdn


CERN