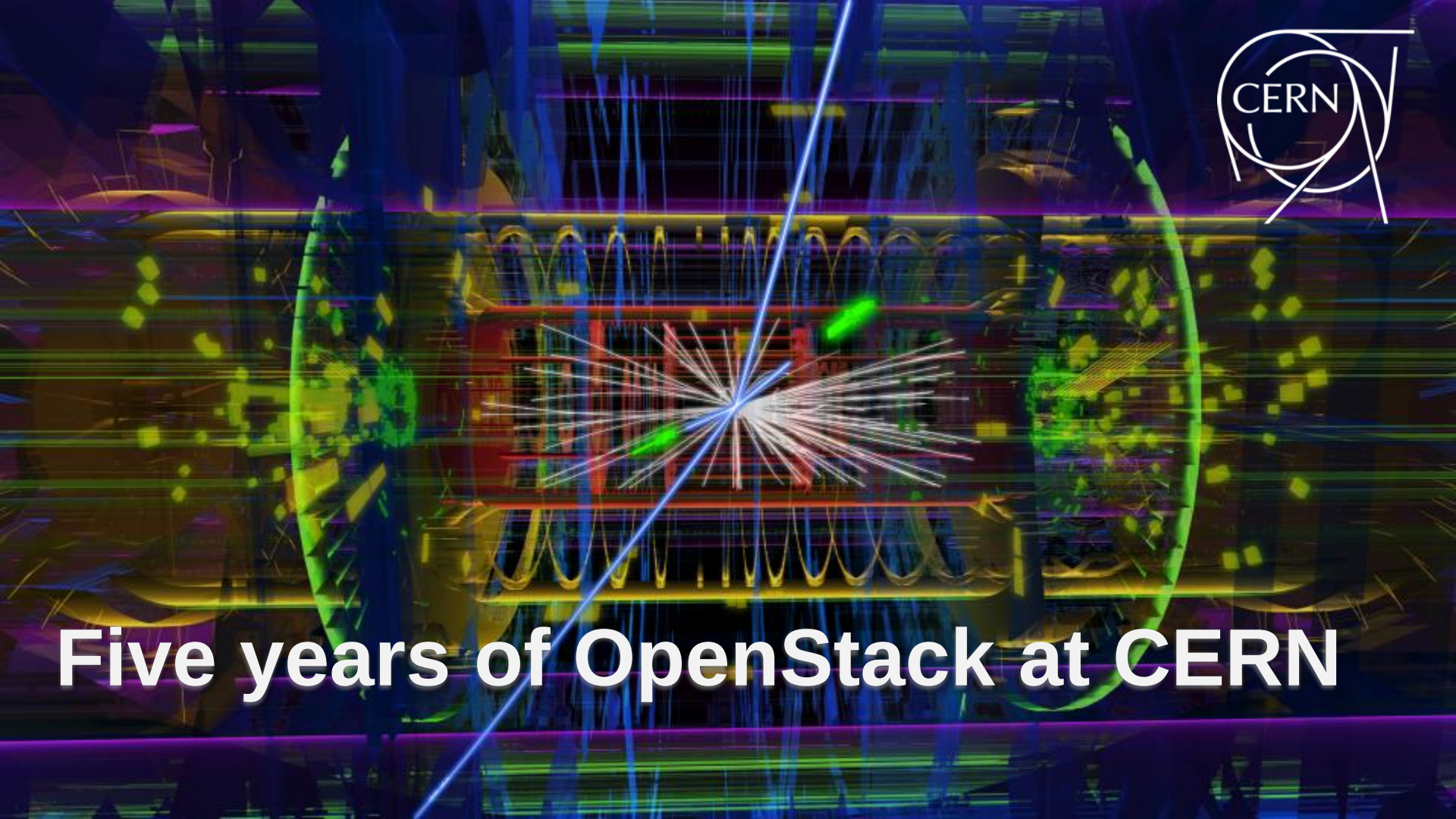


Five years of OpenStack at CERN

CERN: founded in 1954: 12 European States

"Science for Peace"

Today: 22 Member States

~ 2300 staff
~ 1400 other paid personnel
~ 12500 scientific users

Budget (2017) ~1000 MCHF

Member States: Austria, Belgium, Bulgaria, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Israel, Italy, Netherlands, Norway, Poland, Portugal, Romania, Slovak Republic, Spain, Sweden, Switzerland and United Kingdom

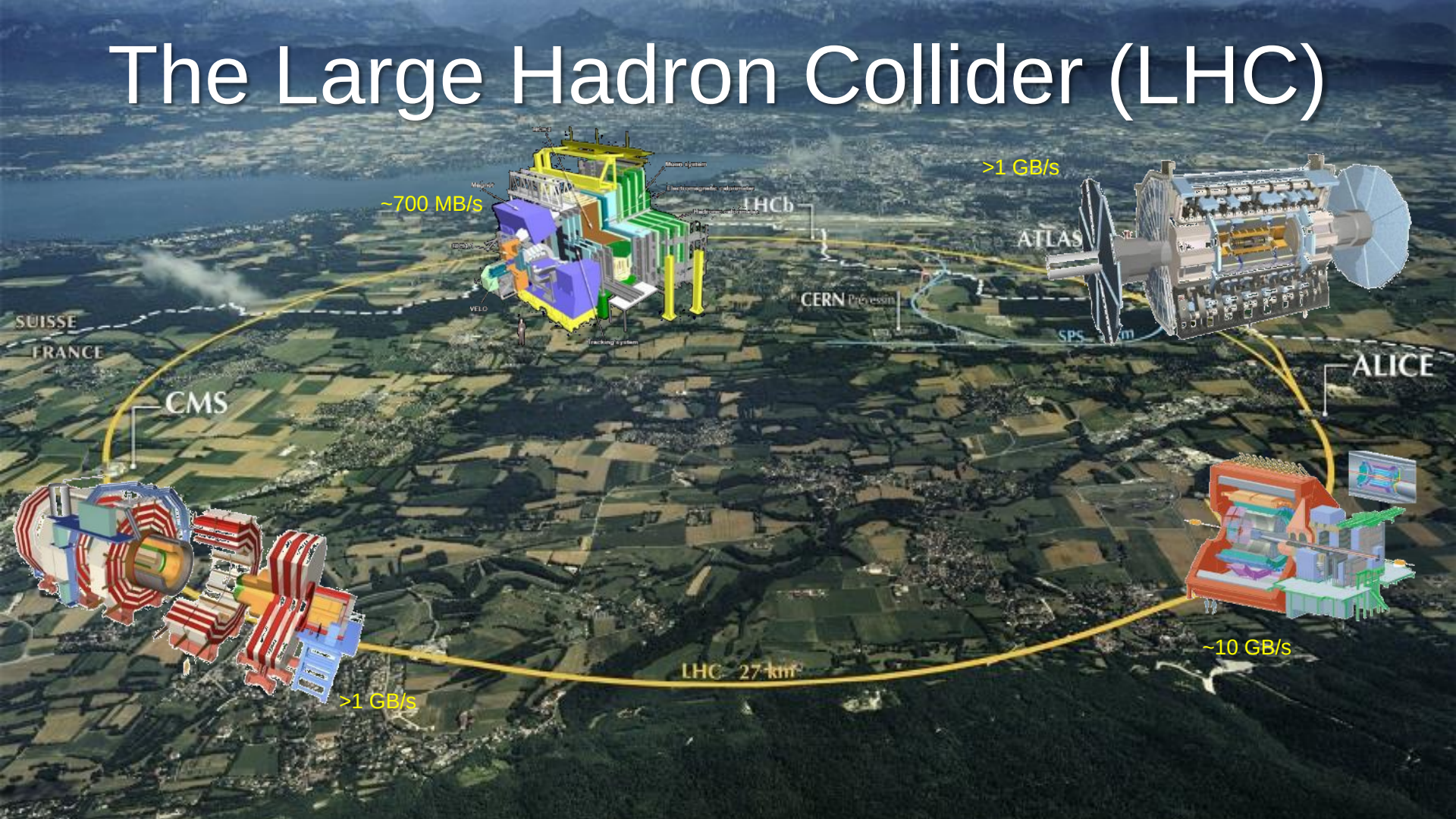
Associate Member States: Pakistan, India, Ukraine, Turkey

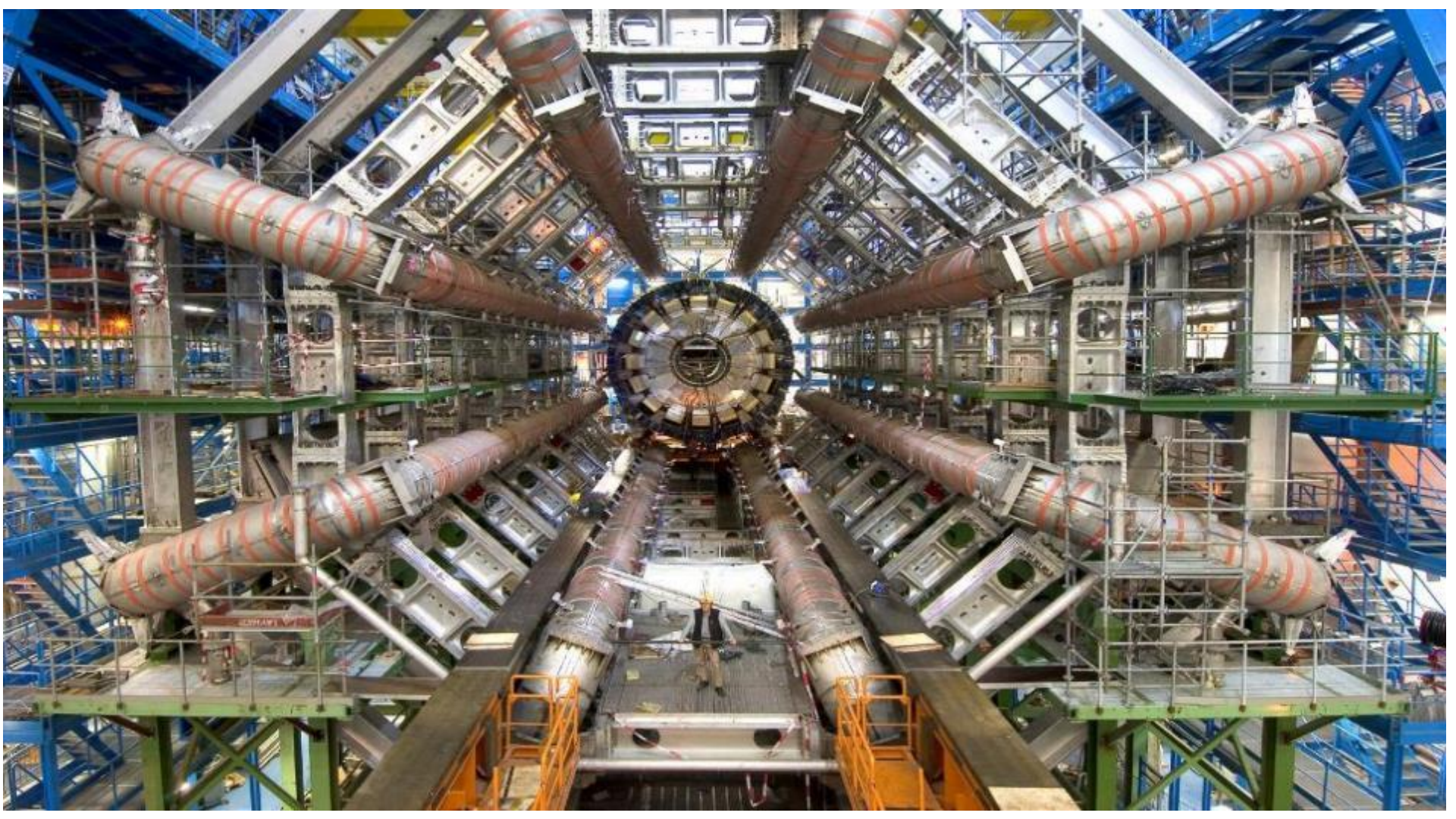
States in accession to Membership: Cyprus, Serbia

Applications for Membership or Associate Membership: Brazil, Croatia, Lithuania, Russia, Slovenia

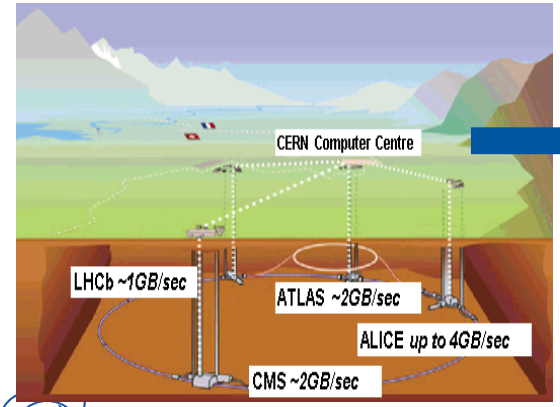
Observers to Council: India, Japan, Russia, United States of America; European Union, JINR and UNESCO

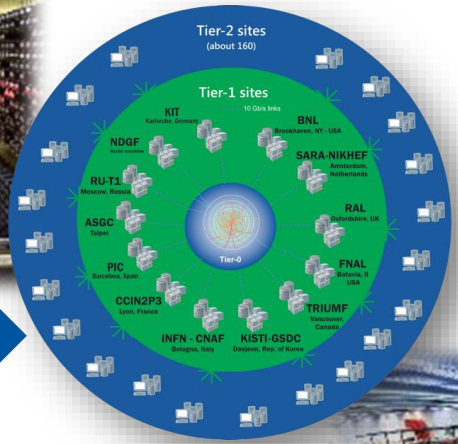
The Large Hadron Collider (LHC)





2016: 50 PB recorded on tape!





Data Centres by Numbers

Meyrin

Metric	Avg
Servers	10.9 K
Processors	20.4 K
Cores	161.2 K
Disks	60.7 K
Memory Modules	80.4 K
1GB NICs	16.4 K
10GB NICs	14.8 K

Wigner

Metric	Avg
Servers	3.5 K
Processors	7.0 K
Cores	56.0 K
Disks	29.7 K
Memory Modules	28.0 K
1GB NICs	6.6 K
10GB NICs	3.0 K

Network

Metric	Avg
Routers	233.0
Star Points	668.0
Switches	3.8 K
Wifi Points	2.0 K
UTP Outlets	75.5 K
Devices	309.7 K

Meyrin

Metric	Avg
Disk Space (TB)	148791
Total Memory (TB)	914

Wigner

Metric	Avg
Disk Space (TB)	97276
Total Memory (TB)	221

Tape Storage

Metric	Avg
Drives	104
Cartridges	25728
Used Space (TB)	195216
Free Space (TB)	34695

Managing all this became…

…very…

…very…

…very…

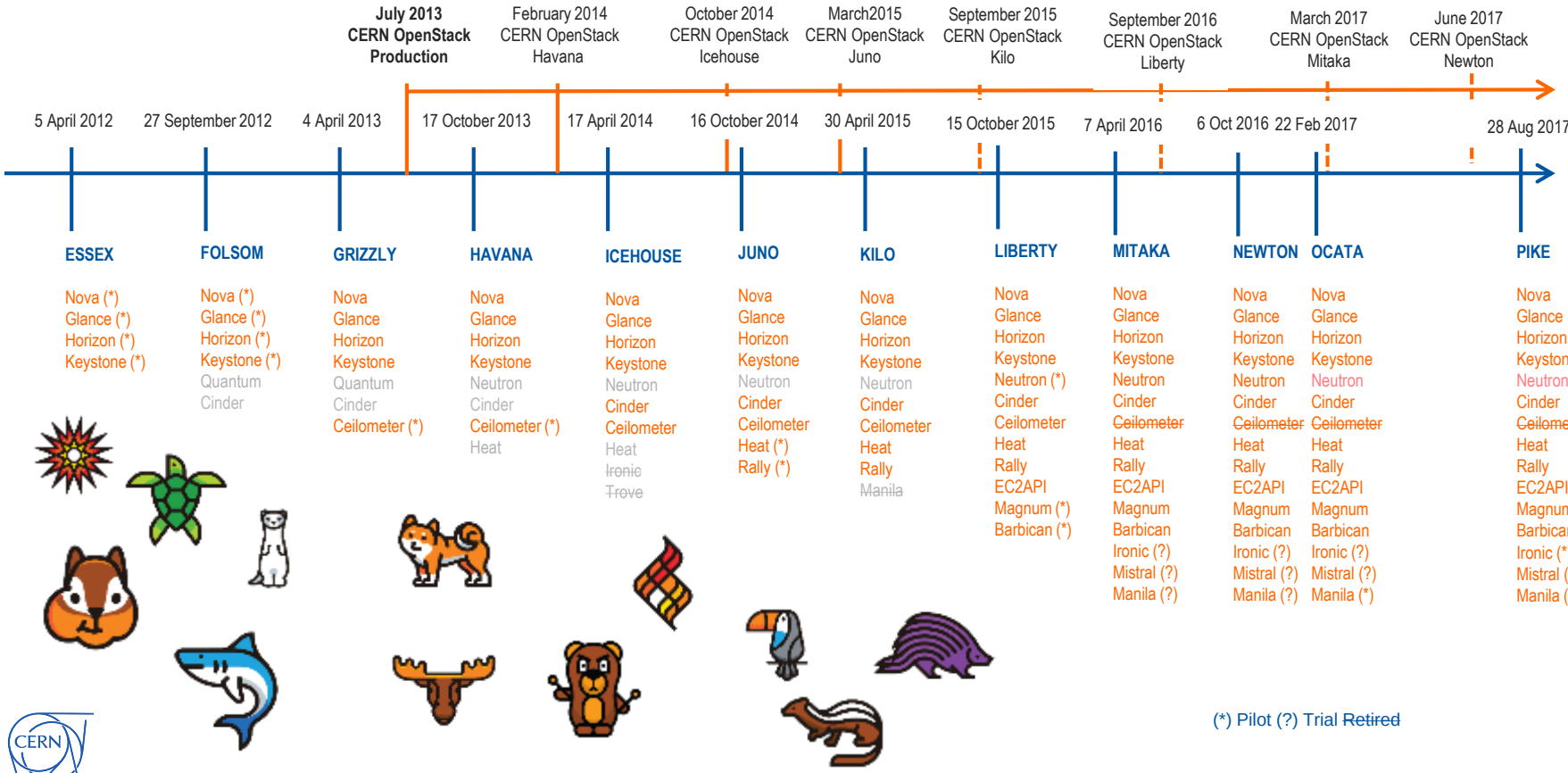
…tricky…

2012: Agile Infrastructure project

- Provisioning + Configuration + Monitoring
- Aim: virtualize all the machines
  - Unless really, really, really not possible
- Offer Cloud endpoints to users
- Scale horizontally
- Consolidate server provisioning
  - Yes, we use the private cloud for server consolidation usecases as well

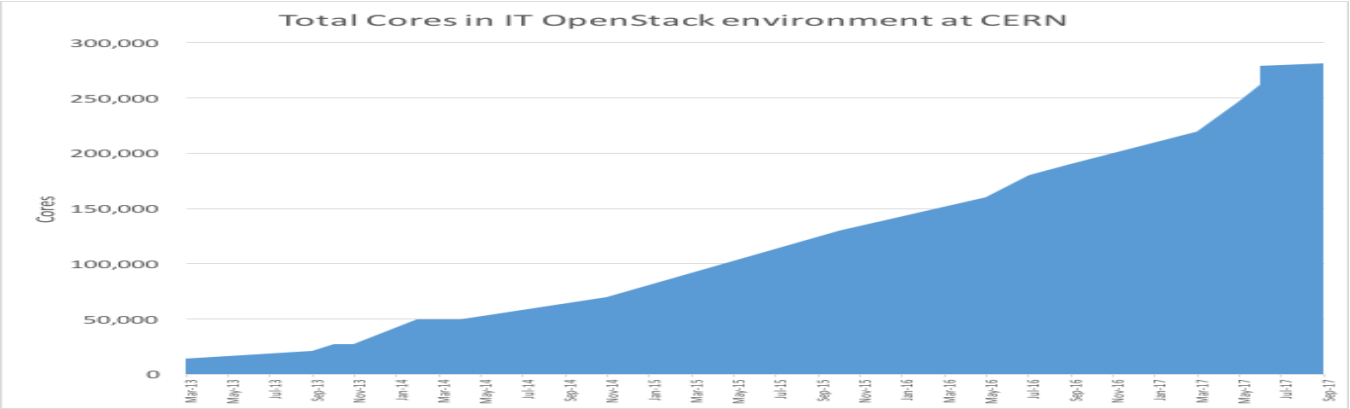
CERN Tool Chain

CERN OpenStack Service Timeline



(*) Pilot (?) Trial ~~Retired~~

CERN OpenStack cloud in numbers



Available	Used	Available	Used	Available	Used
279.1 K cores	262.1 K cores	766.1 TiB RAM	628.1 TiB RAM	13.5 PiB disk	8.5 PiB disk

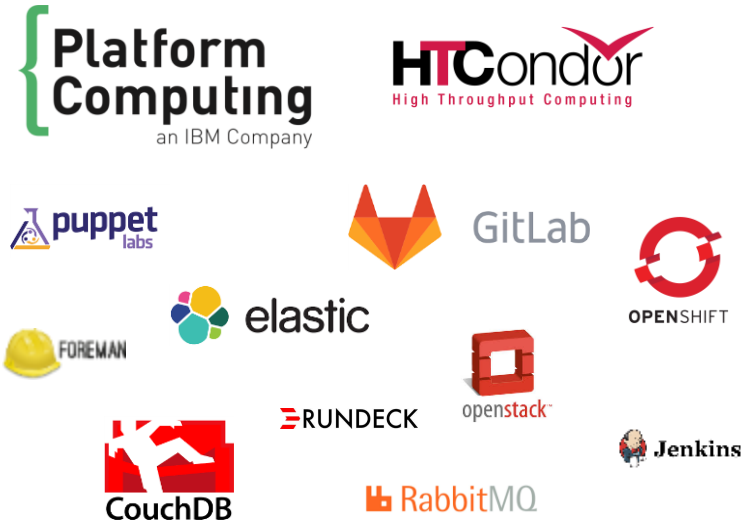
Openstack services stats

Users	Projects	VMs	Magnum clusters	Hypervisors	Images
3068	3812	33311	125	8415	3827

Volumes	Volume size	Fileshares	Fileshares size
4876	1.53 PiB	71	38.6 TiB

Rich Usage Spectrum …

- **Batch service**
  - Physics data analysis
- **IT Services**
  - Sometimes built on top of other virtualised services
- **Experiment services**
  - E.g. build machines
- **Engineering services**
  - E.g. micro-electronics/chip design
- **Infrastructure services**
  - E.g. hostel booking, car rental, …
- **Personal VMs**
  - Development



… rich requirement spectrum!

Scaling Nova

Top level cell

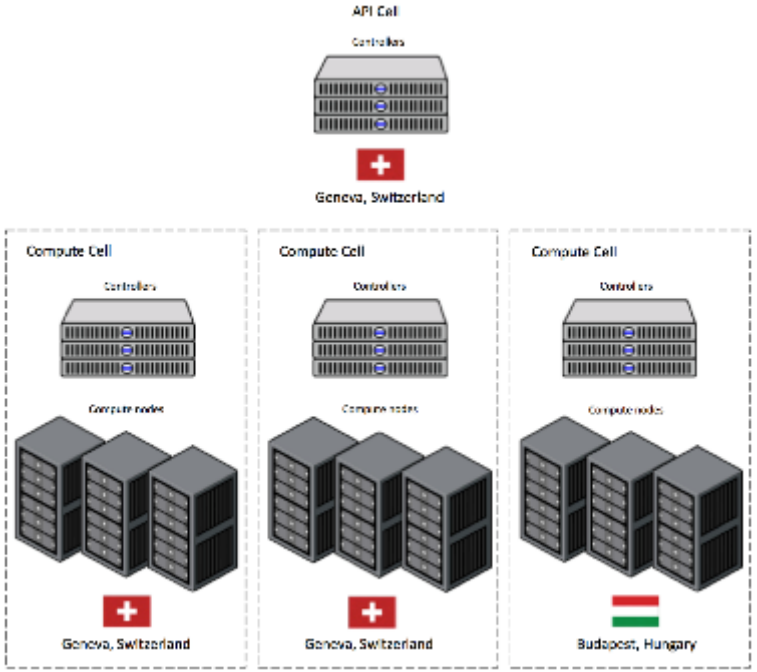
- Runs API service
- Top cell scheduler

>50 child cells run

- Compute nodes
- Scheduler
- Conductor

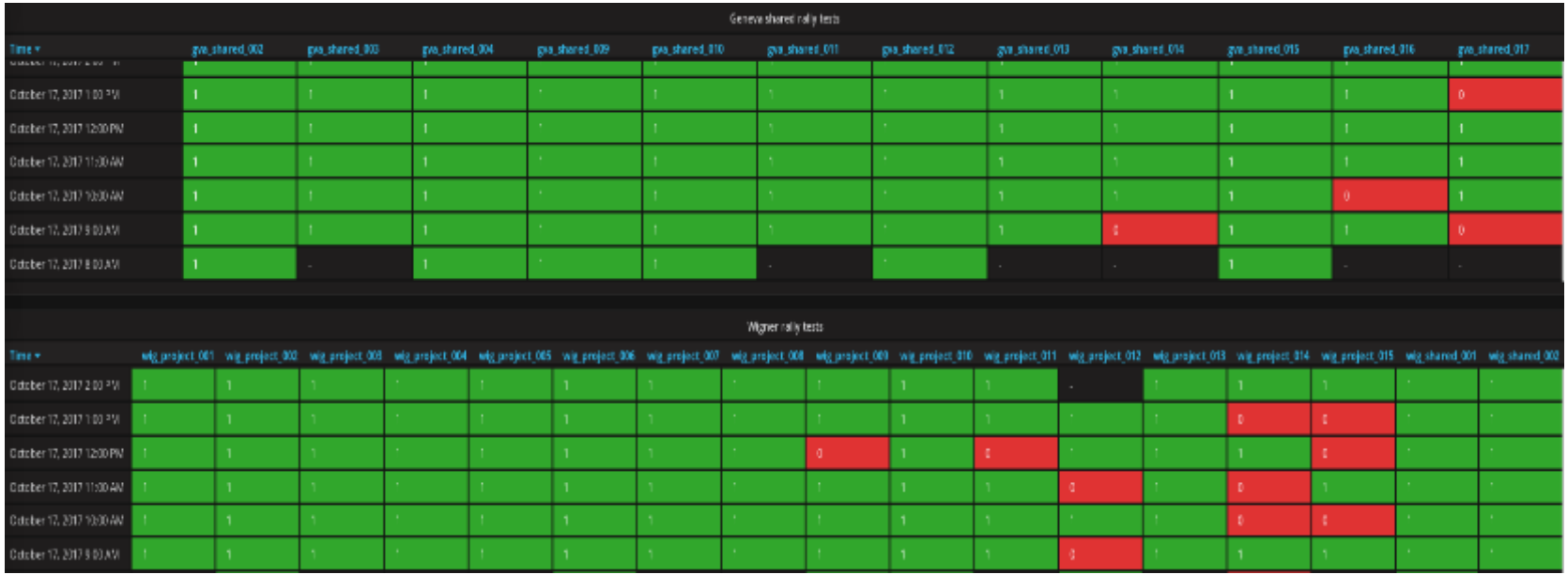
Cells v2 coming

- Default for all



Rally

RALLY
an OpenStack Community Project



Magnum

- Container Engine as a Service
  - Kubernetes, Docker, Mesos, DCOS…
  - 120 clusters, 700 nodes

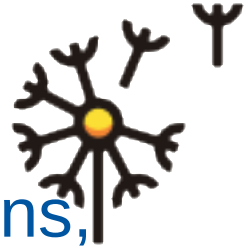
```
$ magnum cluster-create --name myswarmcluster --cluster-template swarm --node-count 100

$ magnum cluster-list
+------+----------------+------------+--------------+-----------------+
| uuid | name           | node_count | master_count | status          |
+------+----------------+------------+--------------+-----------------+
| .... | myswarmcluster | 100        | 1            | CREATE_COMPLETE |
+------+----------------+------------+--------------+-----------------+

$ $(magnum cluster-config myswarmcluster --dir magnum/myswarmcluster)

$ docker info / ps / ...
$ docker run --volume-driver cvmfs -v atlas.cern.ch:/cvmfs/atlas -it centos /bin/bash
[root@32f4cf39128d /]#
```

What's new? Mistral



- Workflow-as-a-Service used for multi-step actions, triggered by users or events
- Horizon dashboard for visualising results
- Examples
  - Expire personal resources after 6 months
  - Multi-step project creation
  - Scheduled snapshot of VMs
- Code at https://gitlab.cern.ch/cloud-infrastructure/mistral-workflows
- Some more complex cases coming in the pipeline

CERN

Automate provisioning

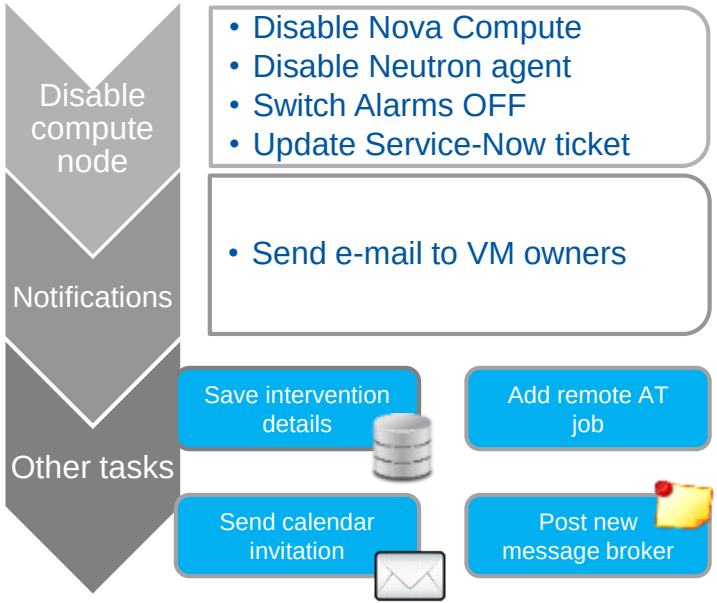
RUNDECK

Automate routine procedures

- Common place for workflows
- Clean web interface
- Scheduled jobs, cron-style
- Traceability and auditing
- Fine-grained access control
- …

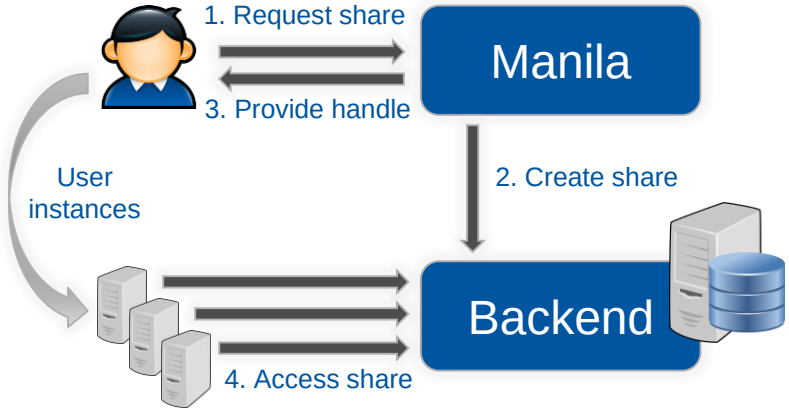
Procedures for

- OpenStack project creation
- OpenStack quota changes
- Notifications of VM owners
- Usage and health reports
- …



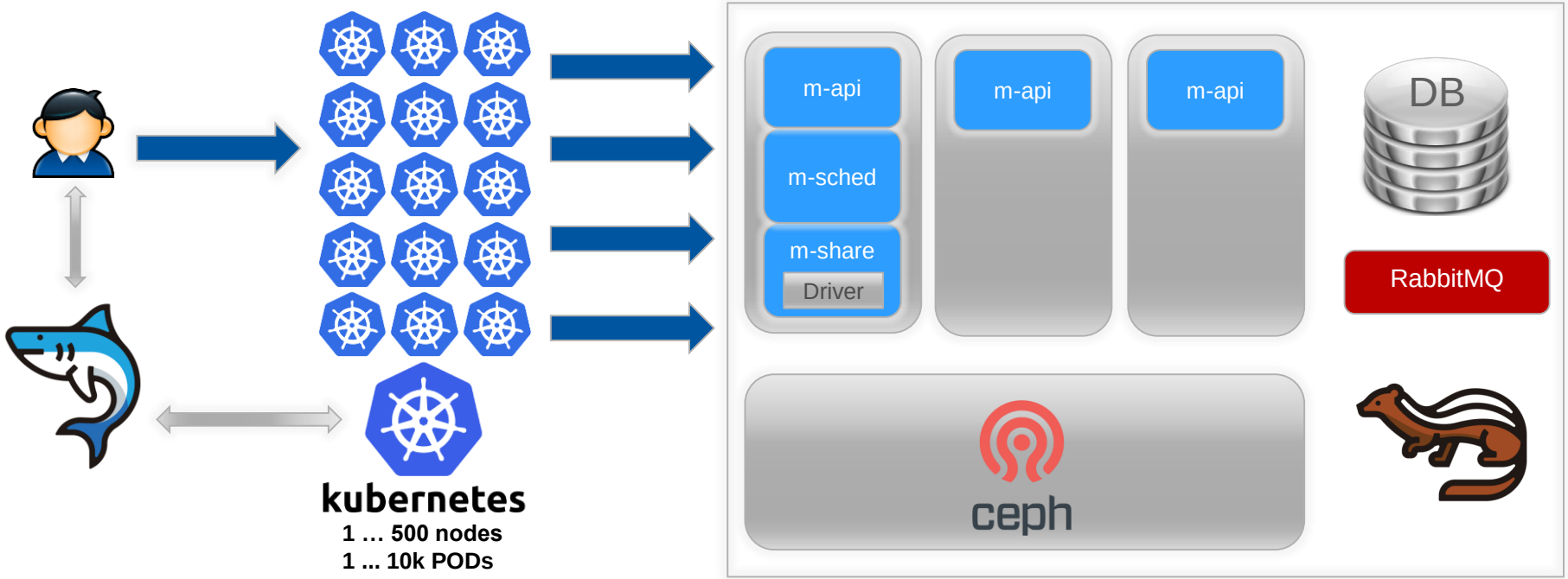
Manila: Overview

- File Share Project in OpenStack
  - Provisioning of shared file systems to VMs
  - ‘Cinder for file shares’
- APIs for tenants to request shares
  - Fulfilled by backend drivers
  - Accessed from instances
- Support for variety of NAS protocols
  - NFS, CIFS, MapR-FS, GlusterFS, **CephFS**, …
- Supports the notion of share types
  - Map features to backends



LHC Incident in April 2016

Manila testing: #fouinehammer



Operations areas going forward

- Further automate migrations
  - Around 5,000 VMs / year
  - First campaign in 2016 needed some additional scripting such as pausing very active VMs
  - Newton live migration includes most use cases
- Software Defined Networking
  - Nova network to Neutron migration to be completed
  - In addition to flat network in use currently
  - Introduce higher level functions such as LBaaS

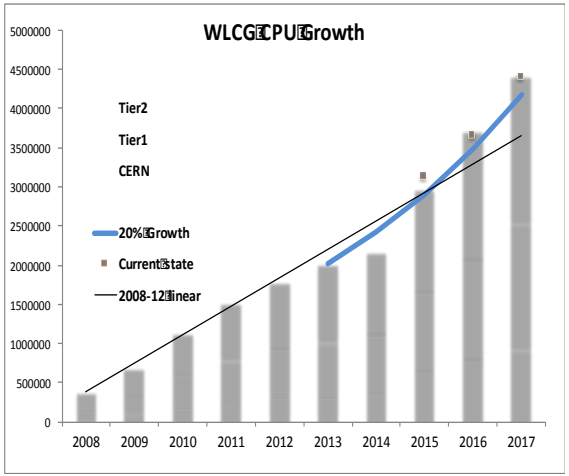
Development areas going forward

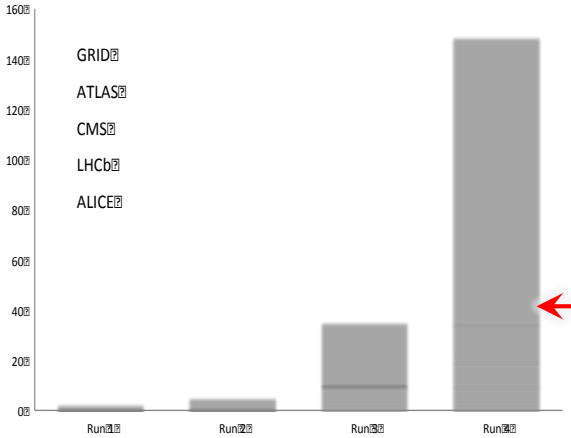
- Nova Pre-emptible VMs
- Nova Cells V2
- **M**agnum rolling upgrades

- Collaborations with Industry

Compute needs growing…

- With the needs of LHC computing in future years, efficient and flexible delivery of compute resources will be key
  - Computing needs in 2023 estimated at 60x the current capacity (HL-LHC)

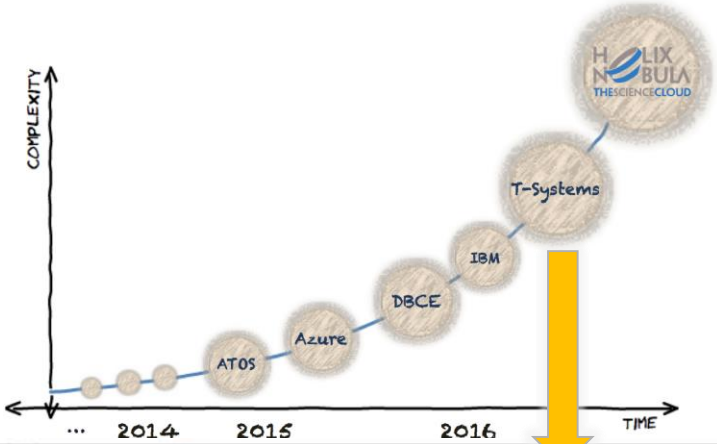


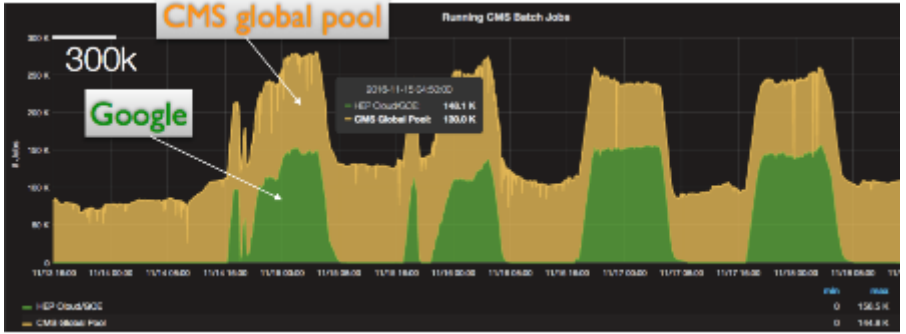


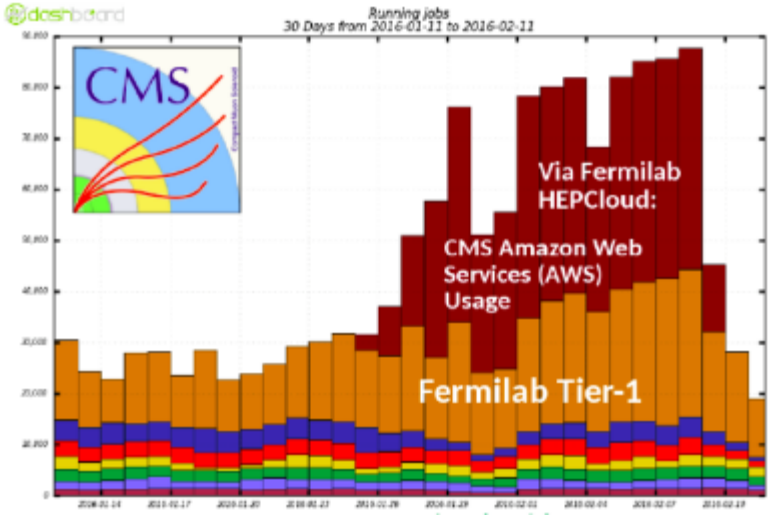
Compute: Growth > x60

What we think is affordable unless we do something differently

Commercial Clouds







Summary

- OpenStack has provided a strong base for scaling resources over the past 5 years
- Additional functionality on top of pure Infrastructure-as-a-Service is now coming to production
- Community and industry collaboration has been productive and inspirational for the CERN team
- Some big computing challenges up ahead…

CERN

Thank you!

Further Information

Technical details on the CERN cloud at
http://openstack-in-production.blogspot.fr

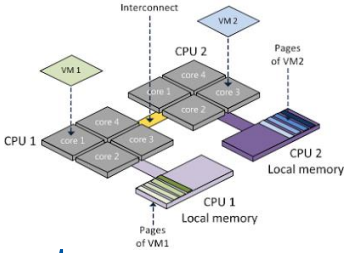
Custom CERN code is at https://github.com/cernops

Scientific Working Group at
https://wiki.openstack.org/wiki/Scientific_working_group

Helix Nebula details at http://www.helix-nebula.eu/

©CERN CC-BY-SA 4.0

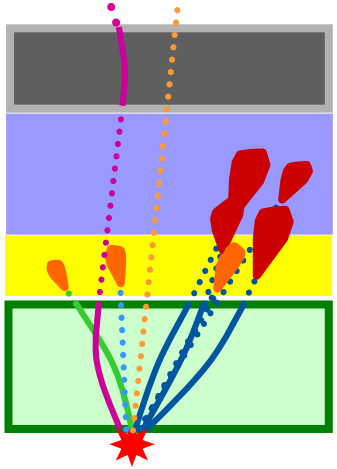
Tuning

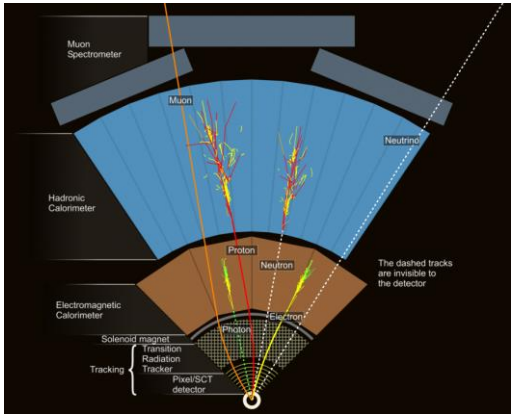


- Many hypervisors are configured for compute optimisation
  - CPU Passthrough so VM sees identical CPU
  - Extended Page Tables so memory page mapping is done in hardware
  - Core pinning so scheduler keeps the cores on the underlying physical cores
  - Huge pages to improve memory page cache utilisation
  - Flavors are set to be NUMA aware
- Improvements of up to 20% in performance
- Impact is that the VMs cannot be live migrated so service machines are not configured this way

Pick the interesting events

- 40 million per second
  - Fast, simple information
  - Hardware trigger in a few micro seconds
- 100 thousand per second
  - Fast algorithms in local computer farm
  - Software trigger in <1 second
- Few 100 per second
  - Recorded for study



Muon tracks

Energy deposits