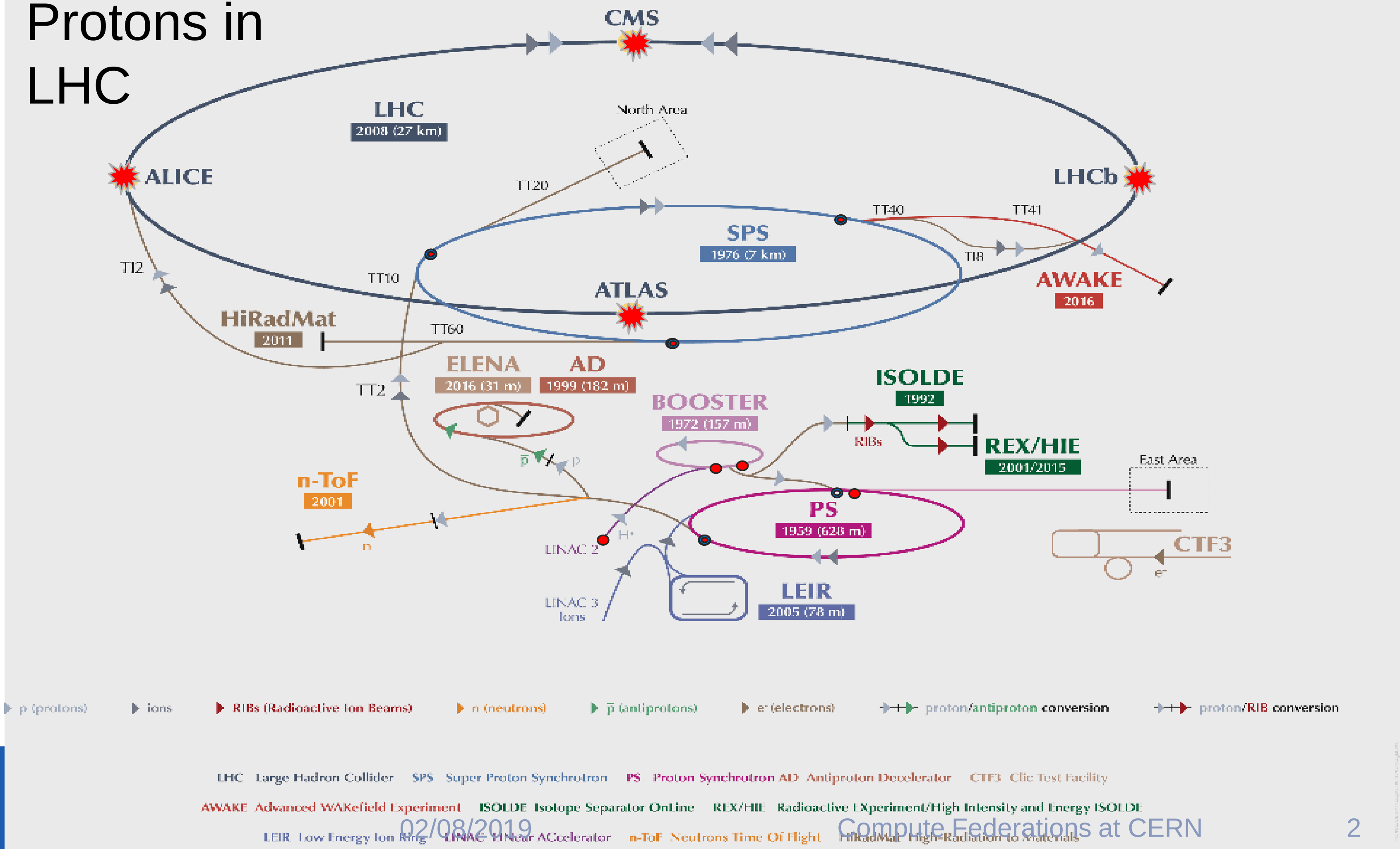
Compute Federation at CERN

Tim Bell,
CERN IT
tim.bell@cern.ch

Research Federation Visit
2nd August 2019

Protons in LHC



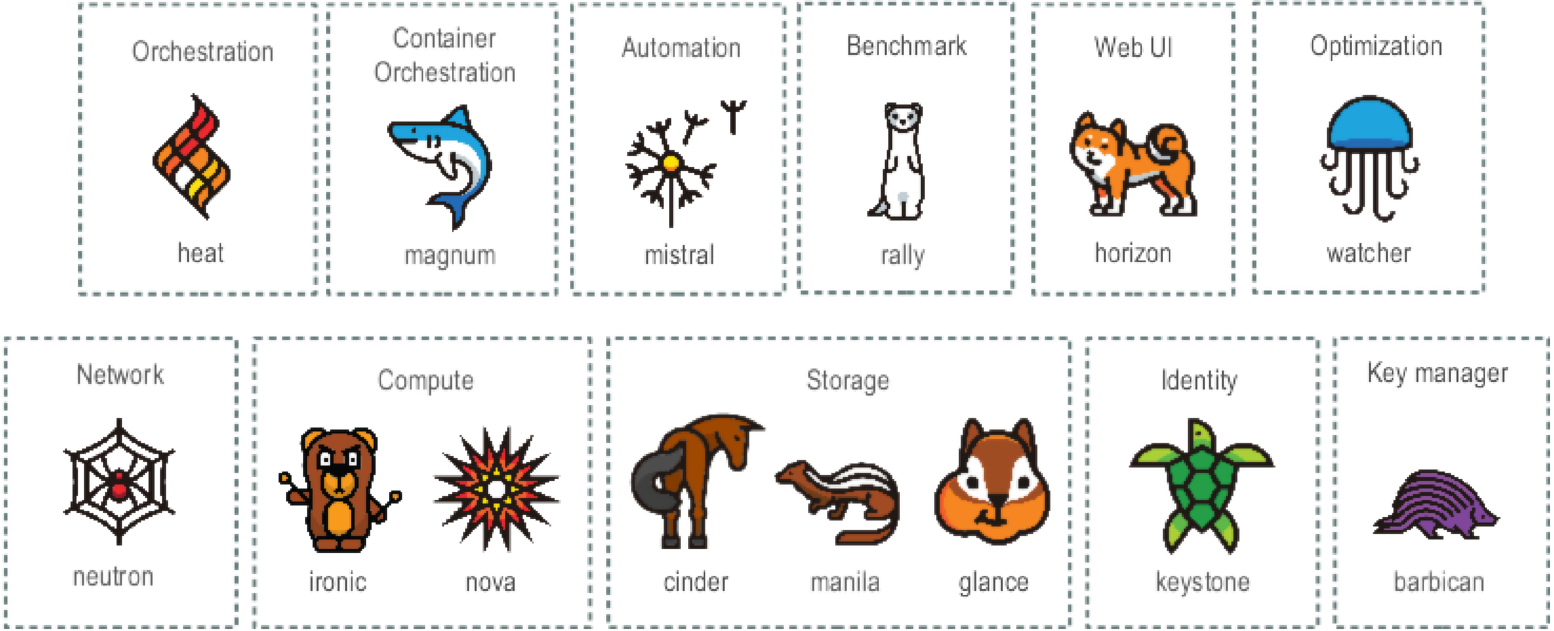
Our Approach: Tool Chain and DevOps

- CERN's requirements are no longer special

git
CentOS
kibana
puppet
FOREMAN
RDO
ceph
openstack
hadoop

- Small dedicated tools allowed for rapid validation & prototyping
- Adapted our processes, policies and work flows to the tools
- Join (and contribute to) existing communities

CERN Cloud Projects



IaaS+

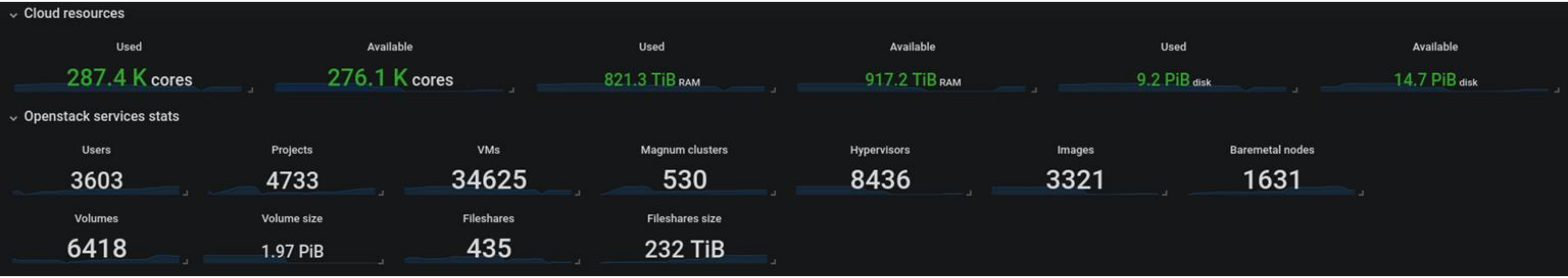
- Orchestration: heat
- Container Orchestration: magnum
- Automation: mistral
- Benchmark: rally
- Web UI: horizon
- Optimization: watcher

IaaS

- Network: neutron
- Compute: ironic, nova
- Storage: cinder, manila, glance
- Identity: keystone
- Key manager: barbican

CERN OpenStack Infrastructure

- Production since 2013
- Provides over 90% of CERN IT Compute resources
- Institution as well as physics services



Cloud resources

Used	Available	Used	Available	Used	Available
287.4 K cores	276.1 K cores	821.3 TiB RAM	917.2 TiB RAM	9.2 PiB disk	14.7 PiB disk

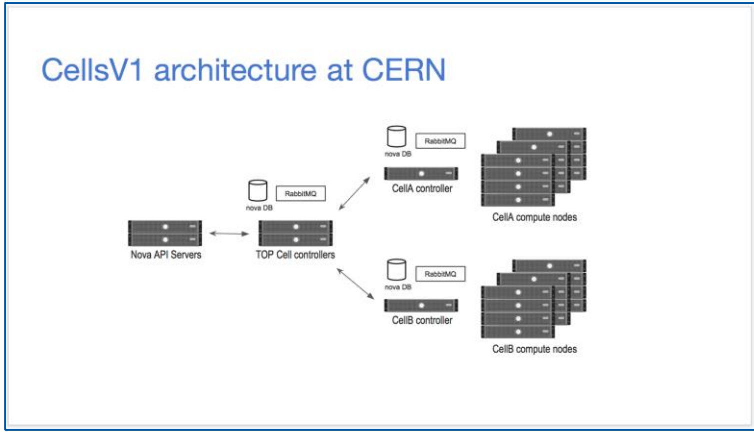
Openstack services stats

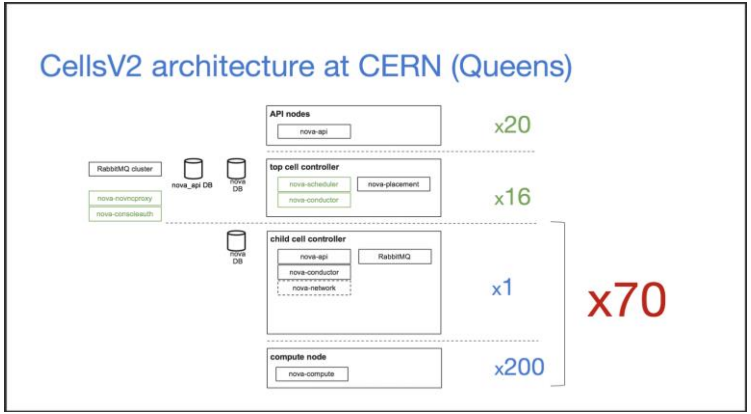
Users	Projects	VMs	Magnum clusters	Hypervisors	Images	Baremetal nodes
3603	4733	34625	530	8436	3321	1631

Volumes	Volume size	Fileshares	Fileshares size
6418	1.97 PiB	435	232 TiB

Nova – Cells

- Allows Nova to scale to thousands of compute nodes by federating multiple smaller units behind a single API endpoint
- Moved from 2 cells to +70 cells
- Cells V1 was developed with large clouds such as Rackspace and NECTaR but was not mainstream
- Upgrade from CellsV1 to CellsV2 in 2018, now all OpenStack clouds are Cells V2.

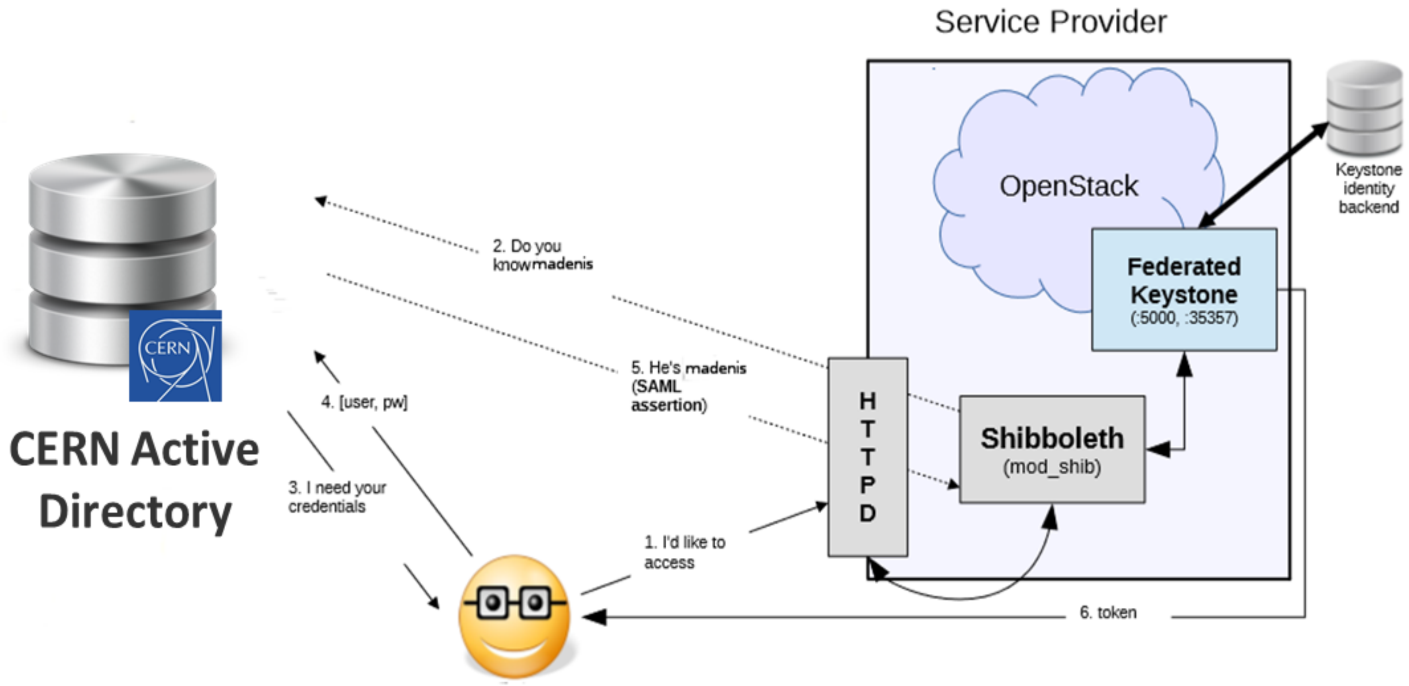

CellsV1 architecture at CERN


CellsV2 architecture at CERN (Queens)

OpenStack Identity Federation

- Authenticate to one cloud using another identity provider
- Started at OpenStack Hong Kong design summit (2013), production in Juno (2014)
- CERN 3 year fellow sponsored by Rackspace in CERN openlab
- Iterative design using open blueprints

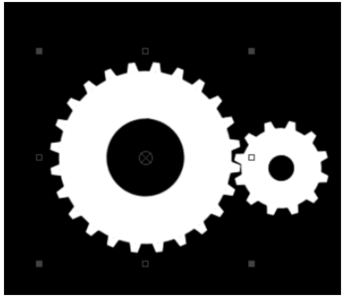
Identity Federation Implementation



Policy

Assertion

LOGIN: madenis
LANGUAGE: EN
DEPARTMENT: IT/OIS
FULLNAME: Marek Denis



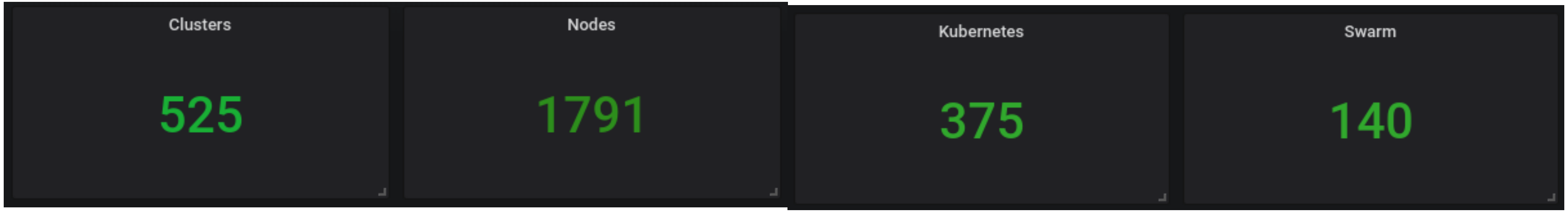
```
[
  { "local":
      [ { "user": { "name":   "{0}" } } ],
    "remote":
      [ { "type": "ADFS_LOGIN" } ]
   },
   {
    "local":
       [ { "group": { "id": “devs" } } ,
         {“group”: {ïd”:”openlab”} } ],
     "remote":
       [ { "type":"DEPARTMENT",
            "any_one_of": ["IT/OIS"] } ]
   }
]
```

Keystone credentials

```
{
   name: madenis
   groups: [
    “devs”,
    “openlab”
   ]
}
```

CERN Magnum Deployment

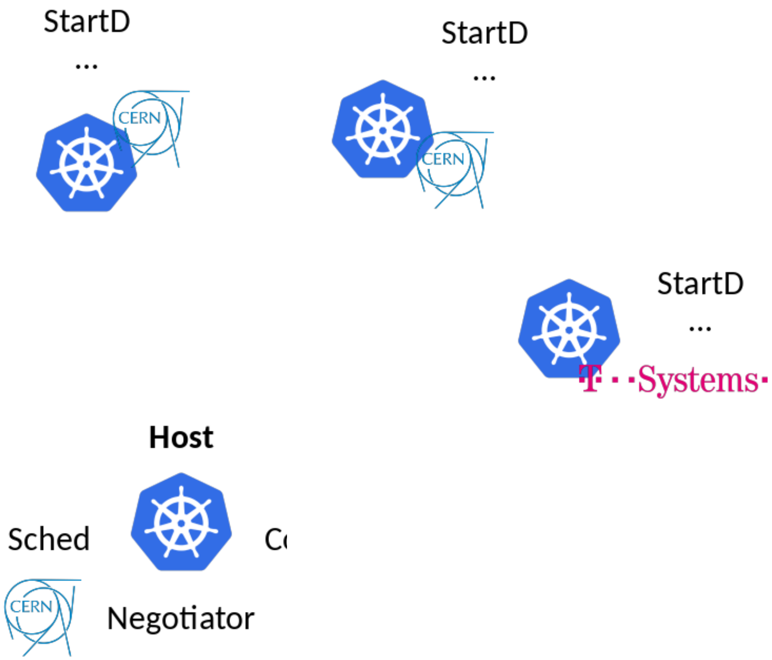
- One click Container Cluster as a Service (ala GKE)
  - Kubernetes, Docker Swarm or Mesos (on top of OpenStack VMs/bare metal)
  - In production since 2016
- Integrated into accounting, quota, CLI and Web
- Uses standard interfaces such as CSI to provide best-of-breed deployments
  - Monitoring, Storage (CVMFS/CephFS/EOS)



Clusters	Nodes	Kubernetes	Swarm
525	1791	375	140

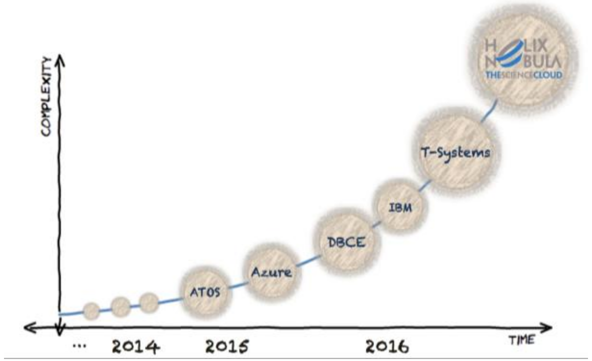
Federated Kubernetes across multiple clouds

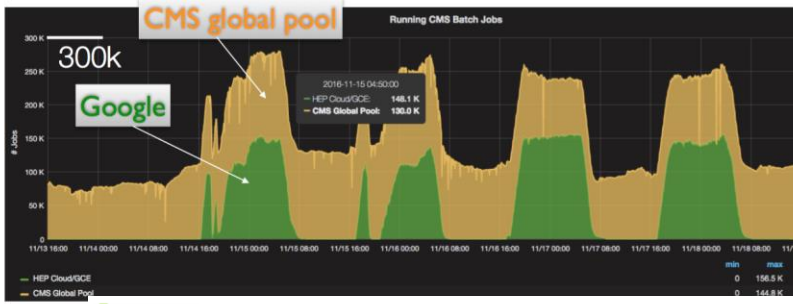
- Batch or other jobs on multiple clusters
- Transparent bursting to public clouds
- Same container images on premise and off
- Exploit public cloud services like Google's GKE and Amazon's EKS

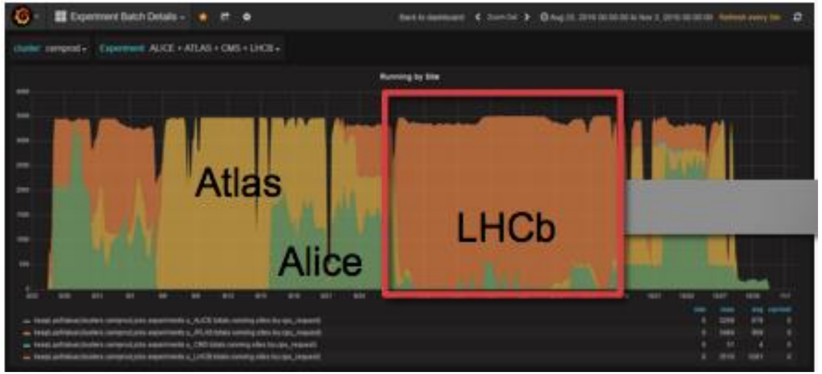


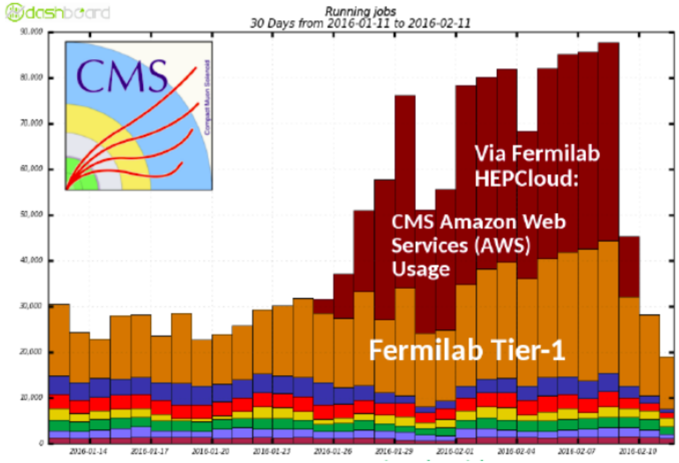
Credit: Ricardo Rocha, CERN Cloud

Commercial Clouds









Conclusions

- Compute federation is in development at multiple levels
  - Within a single cloud instance with Cells
  - Between multiple OpenStack clouds with Identity Federation
  - Between multiple diverse clouds with Kubernetes
- Many commonalities across sciences
  - Open collaboration through Special Interest Groups is effective to advance consistently
  - Easy sharing of enhancements through open source foundations such as CNCF and OpenStack

Further Information

- CERN blogs
  - https://techblog.web.cern.ch/techblog/
- Recent Talks at OpenStack summits
  - https://www.openstack.org/videos/search?search=cern
- Kubecon 2018, 2019
- Source code
  - https://github.com/cernops and https://github.com/openstack

CERN

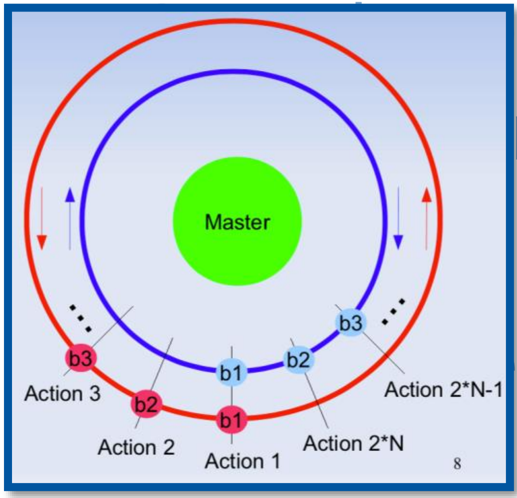
Backup Slides

HTC vs HPC

- CERN is mostly a high throughput computing lab:
  - File-based parallelism, massive batch system for data processing

- But we have several HPC use-cases:
  - Beam simulations, plasma physics, CFD, QCD, (ASICs)
  - Need full POSIX consistency, fast parallel IO

HTC vs HPC

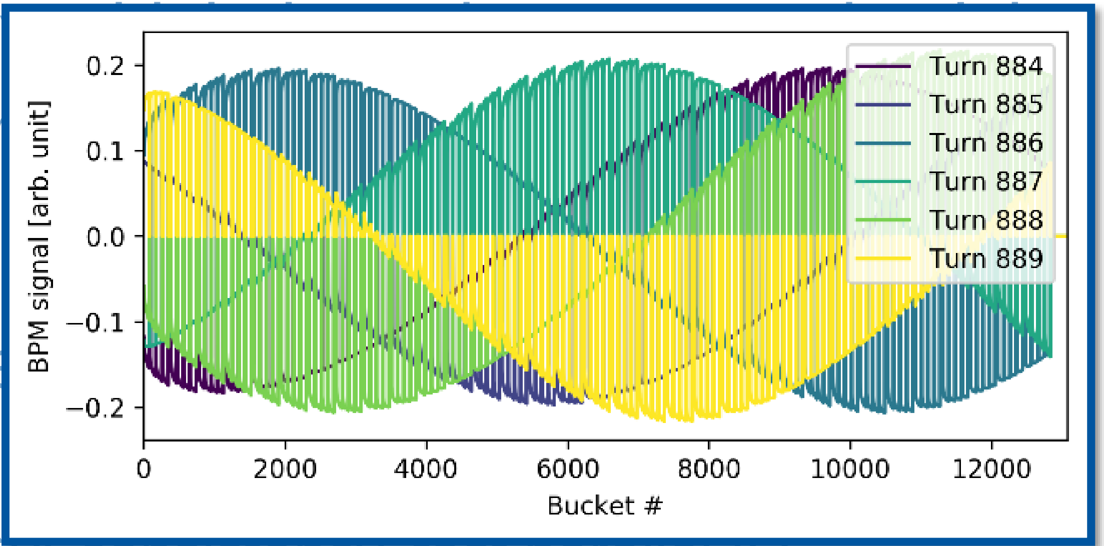


stl

par

se

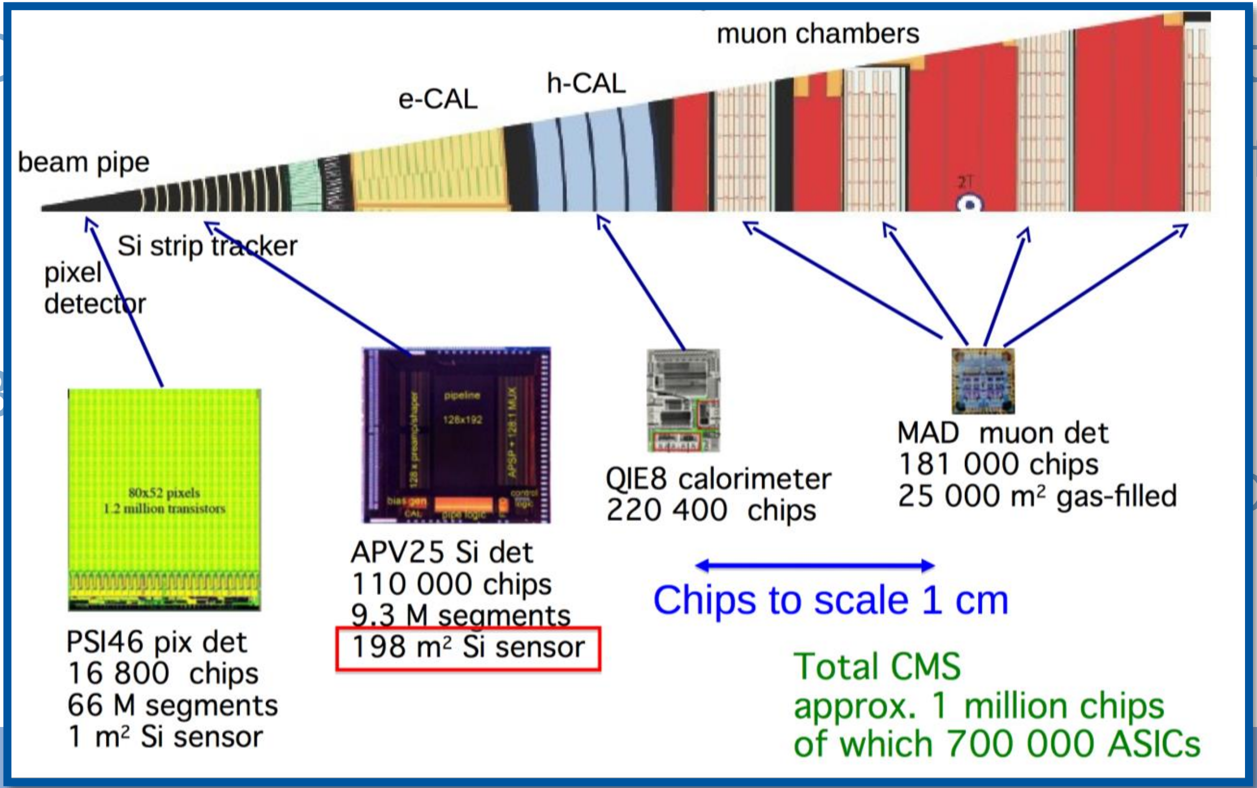
lati



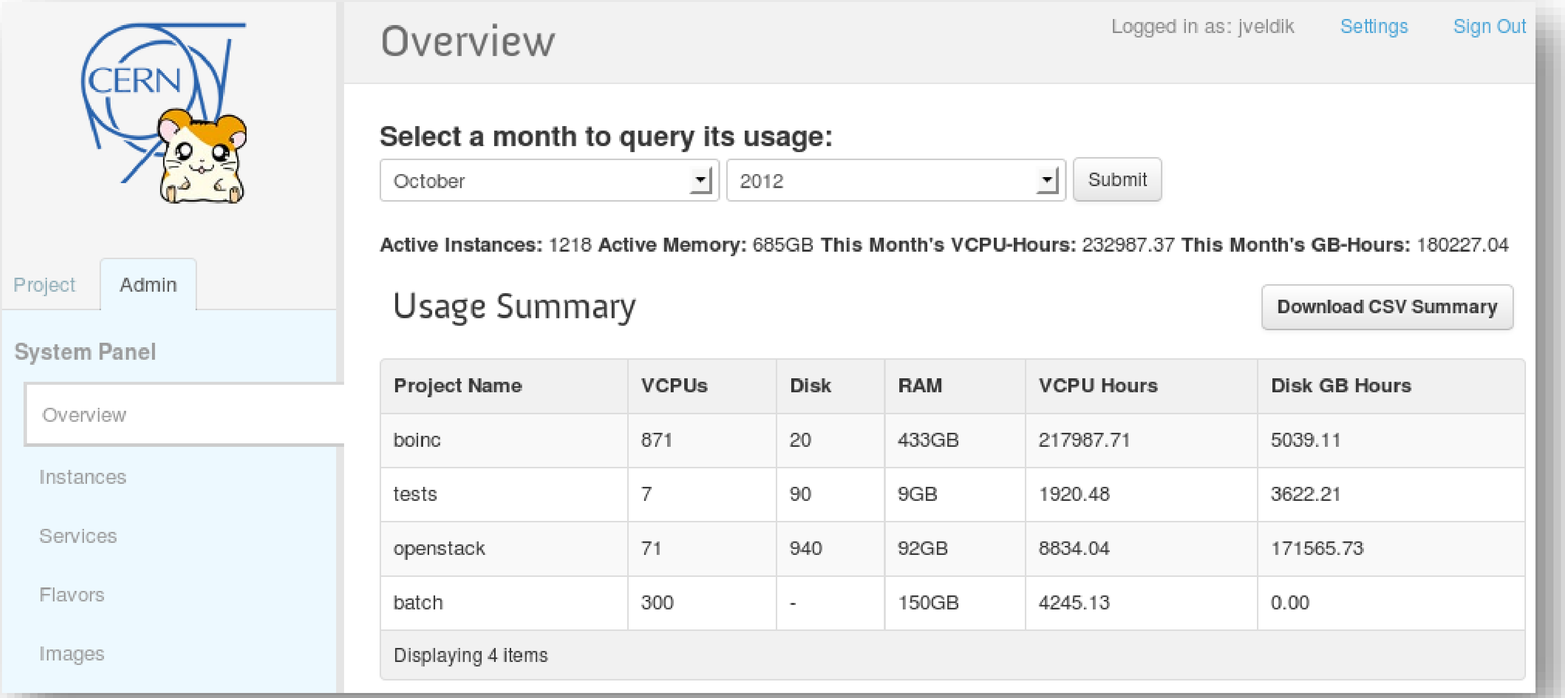
- Need full POSIX consistency, fast parallel IO

HTC vs HPC

- C ting lab:
  - for data
- B
  - , (ASICs)
  -



Early Prototypes



Overview

Select a month to query its usage:

October | 2012 | Submit

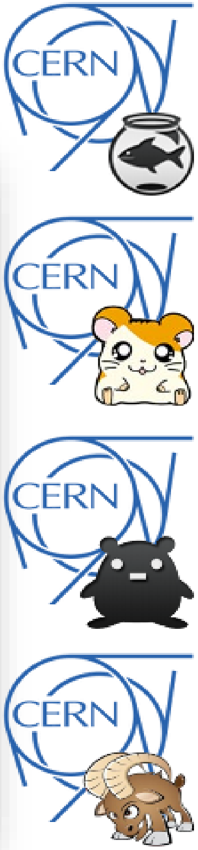
Active Instances: 1218 **Active Memory:** 685GB **This Month's VCPU-Hours:** 232987.37 **This Month's GB-Hours:** 180227.04

Usage Summary

Download CSV Summary

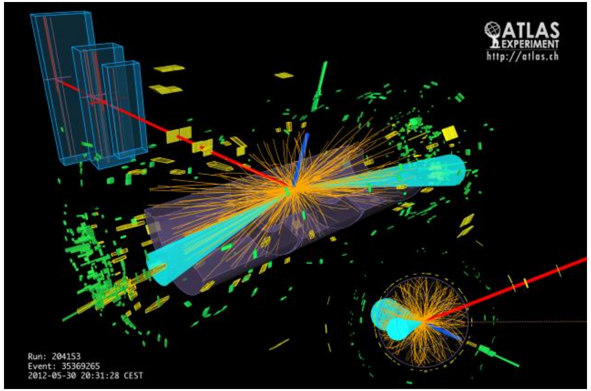
Project Name	VCPUs	Disk	RAM	VCPU Hours	Disk GB Hours
boinc	871	20	433GB	217987.71	5039.11
tests	7	90	9GB	1920.48	3622.21
openstack	71	940	92GB	8834.04	171565.73
batch	300	-	150GB	4245.13	0.00

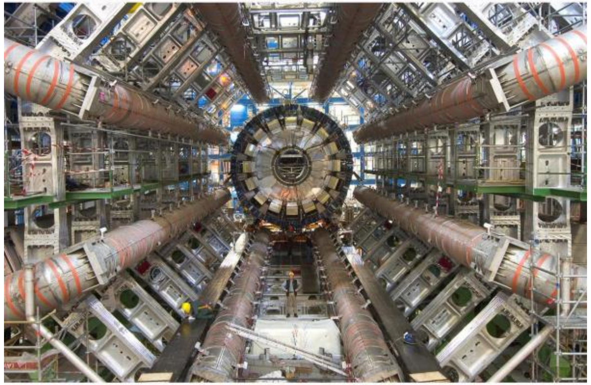
Displaying 4 items

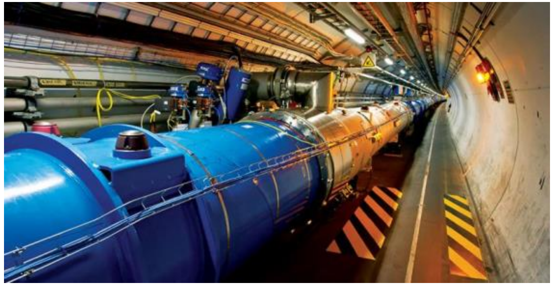


CERN Container Use Cases

- Batch Processing
- End user analysis / Jupyter Notebooks
- Machine Learning / TensorFlow / Keras
- Infrastructure Management
  - Data Movement, Web servers, PaaS …
- Continuous Integration / Deployment
- Run OpenStack :-)
- And many others





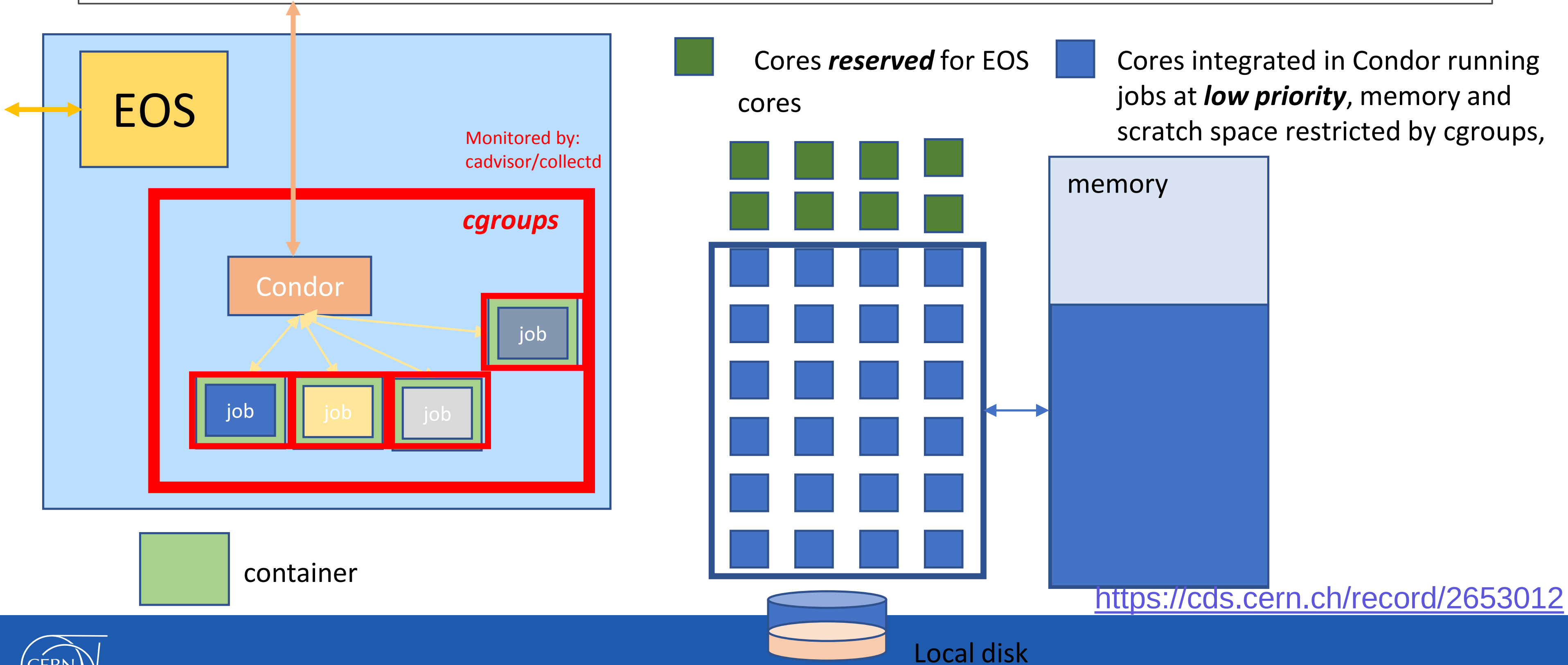


Credit: Ricardo Rocha, CERN Cloud

HTC own

- H
  t
- S
  c
  e
- N
  r
- N
  r

Batch on Storage Services - BEER



EOS

Monitored by:
cadvisor/collectd

cgroups

Condor

job

job

job

job

🟩 container

🟩 Cores ***reserved*** for EOS cores

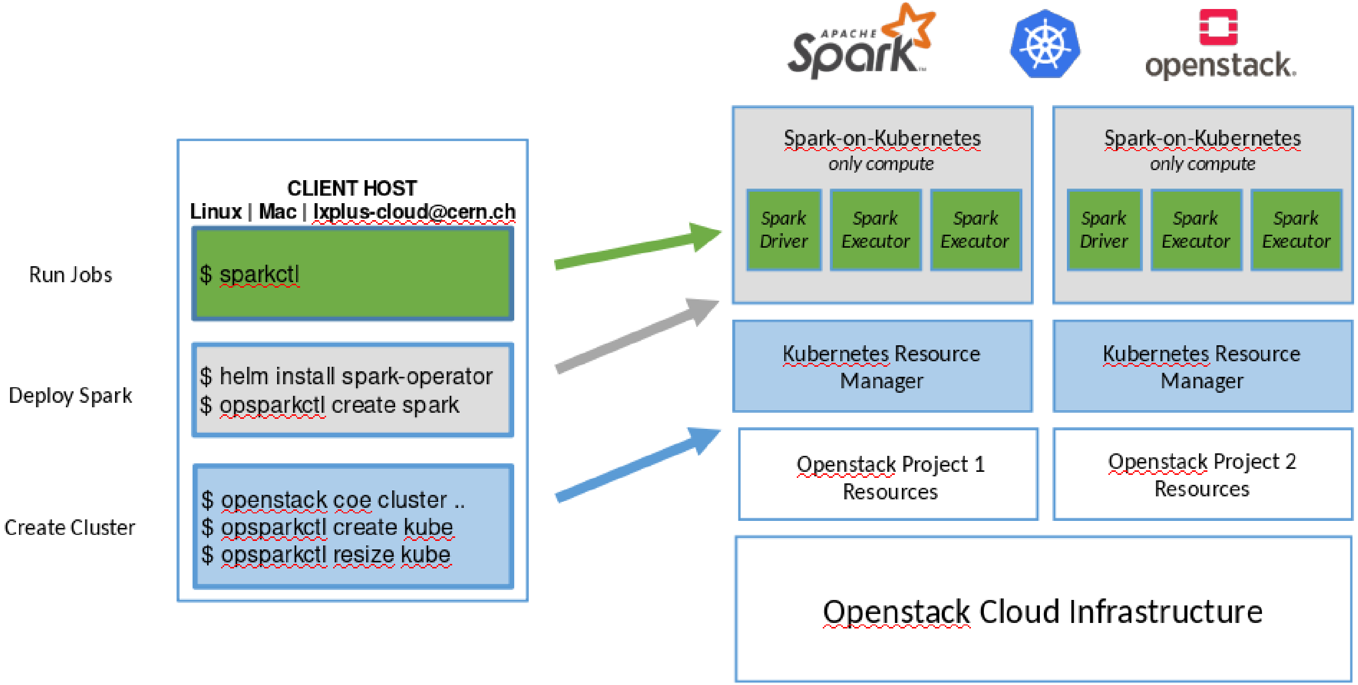
🟦 Cores integrated in Condor running jobs at ***low priority***, memory and scratch space restricted by cgroups,

memory

Local disk

https://cds.cern.ch/record/2653012

Use Case: Spark on K8s



Run Jobs

Deploy Spark

Create Cluster

CLIENT HOST
Linux | Mac | lxplus-cloud@cern.ch

```
$ sparkctl
```

```
$ helm install spark-operator
$ opsparkctl create spark
```

```
$ openstack coe cluster ..
$ opsparkctl create kube
$ opsparkctl resize kube
```

Spark-on-Kubernetes
only compute
Spark Driver | *Spark Executor* | *Spark Executor*

Spark-on-Kubernetes
only compute
Spark Driver | *Spark Executor* | *Spark Executor*

Kubernetes Resource Manager

Kubernetes Resource Manager

Openstack Project 1 Resources

Openstack Project 2 Resources

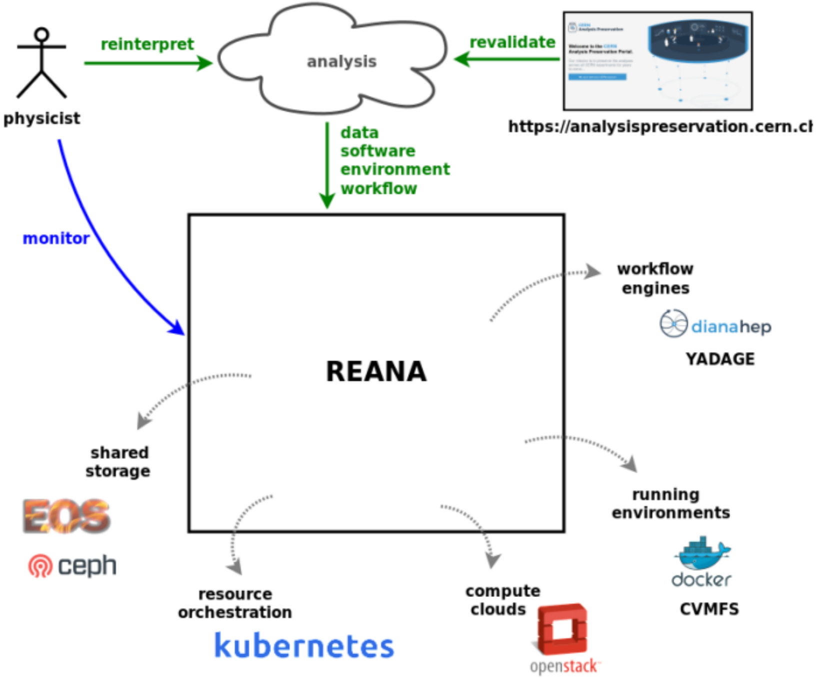
Openstack Cloud Infrastructure

Credit: CERN data analytics working group

Use case: REANA / RECAST

Reusable Analysis Platform

- Workflow Engine (Yadage)
- Each step a Kubernetes Job
- Integrated Monitoring & Logging
- Centralized Log Collection
- [”Rediscovering the Higgs”]() at Kubecon



Credit: CERN Invenio User Group Workshop

CERN Ceph Clusters		**Size**	**Version**
OpenStack Cinder/Glance	*Production*	5.5PB	jewel
	Satellite data centre (1000km away)	0.4PB	luminous
CephFS (HPC+Manila)	*Production*	0.8PB	luminous
	Manila testing cluster	0.4PB	luminous
	Hyperconverged HPC	0.4PB	luminous
CASTOR/XRootD	*Production*	4.2PB	luminous
	CERN Tape Archive	0.8PB	luminous
S3+SWIFT	*Production*	0.9PB	luminous

+5PB in the pipeline

What to consider when running a container service

- Design your network
  - By default, magnum creates a private network per cluster and assigns floating IPs to nodes
  - LBaaS for multi-master clusters
- Run a container registry
  - DockerHub is usually up but latency will always get you
  - Rebuild or mirror the containers used by magnum
- Provide self-service clusters -> Provide software
  - Upgrade magnum regularly, update its configuration regularly
  - Plan which container and glance images are available to users

Cluster Resize

Motivation: Remove specific nodes from the cluster (replace update cmd)

Forward compatibility with old cluster

Already in the upstream stable branch

ETA for release, 12 of April (upstream)

api-reference:

http://git.openstack.org/cgit/openstack/magnum/tree/api-ref/source/clusters.inc#n268

Thanks to Feilong Wang

Cluster Resize

```
$ openstack coe cluster resize --nodegroup kube-worker kube 3
Request to resize cluster kube has been accepted.

$ openstack coe cluster list
+--------------------------------------+------+---------+------------+--------------+
| uuid                                 | name | keypair | node_count | master_count |
+--------------------------------------+------+---------+------------+--------------+
| ed38e800-5884-4053-9b17-9f80995f1993 | kube | default |          3 |            1 |
+--------------------------------------+------+---------+------------+--------------+
…
--------------------+----------------+
 status             | health_status |
--------------------+----------------+
 UPDATE_IN_PROGRESS | HEALTHY       |
--------------------+----------------+

$ openstack coe cluster resize --nodegroup kube-worker \
  --nodes-to-remove 05b7b307-18fd-459a-a13a-a1923c2c840d kube 1
Request to resize cluster kube has been accepted.
```

Node Groups

```
$ openstack coe nodegroup list kube
+--------------------------------------+-------------+-----------+------------+--------+
| uuid                                 | name        | flavor_id | node_count | role   |
+--------------------------------------+-------------+-----------+------------+--------+
| 14ddaf00-9867-49ca-b10c-106c3656e4f1 | kube-master | m1.small  |          1 | master |
| 8a18cc5c-040d-4e67-aa4d-9aaf38241119 | kube-worker | m1.small  |          1 | worker |
+--------------------------------------+-------------+-----------+------------+--------+

$ openstack coe nodegroup show kube kube-master
+--------------------+--------------------------------------+
| Field              | Value                                |
+--------------------+--------------------------------------+
| name               | kube-master                          |
| cluster_id         | ed38e800-5884-4053-9b17-9f80995f1993 |
| flavor_id          | m1.small                             |
| node_addresses     | [u'172.24.4.120']                    |
| node_count         | 1                                    |
| role               | master                               |
| max_node_count     | None                                 |
| min_node_count     | 1                                    |
| is_default         | True                                 |
+--------------------+--------------------------------------+
```

Authentication to OpenStack Keystone

Use OpenStack tokens directly in kubectl

- Give *kubectl* access to users outside the cluster's OpenStack project
- A better (more secure) option than the current TLS certificates

```
$ openstack coe cluster create ... --labels keystone_auth_enabled=true

$ export OS_TOKEN=$(openstack token issue -c id -f value)

$ kubectl get pod
```

Cluster Metrics Monitoring (Prometheus)

Objectives

Provide an out-of-the-box solution for cluster, node and application metrics monitoring

Services Included

Metrics scraping and storage (Prometheus)
Data visualization (Grafana)
Alarms (Alertmanager)
Upstream Prometheus Operator Helm Chart

Slide Credit: Diogo Guerra, CERN Cloud

Cluster Upgrades

Upgrades of Kubernetes, Operating System, Add-ons

  Rolling in-place upgrade

  Rolling node-replacement

  Batch size for rolling upgrade

  https://storyboard.openstack.org/#!/story/2002210

More Add-ons

Ingress Controllers

- Traefik v1.7.x
  - Can be used with Neutron-lbaas/Octavia or HostNetwork
  - Octavia 1.13.2-alpha or newer

Node Problem Detector

- Customizable detectors for node health

Pod Security Policy

- Two modes, privileged or restricted by default

Magnum Deployment

- Clusters are described by *cluster templates*
- Shared/public templates for most common setups, customizable by users

```
$ openstack coe cluster template list
+------+---------------------------+
| uuid | name                      |
+------+---------------------------+
| .... | swarm                     |
| .... | swarm-ha                  |
| .... | kubernetes                |
| .... | kubernetes-ha             |
| .... | mesos                     |
| .... | mesos-ha                  |
| .... | dcos                      |
+------+---------------------------+
```

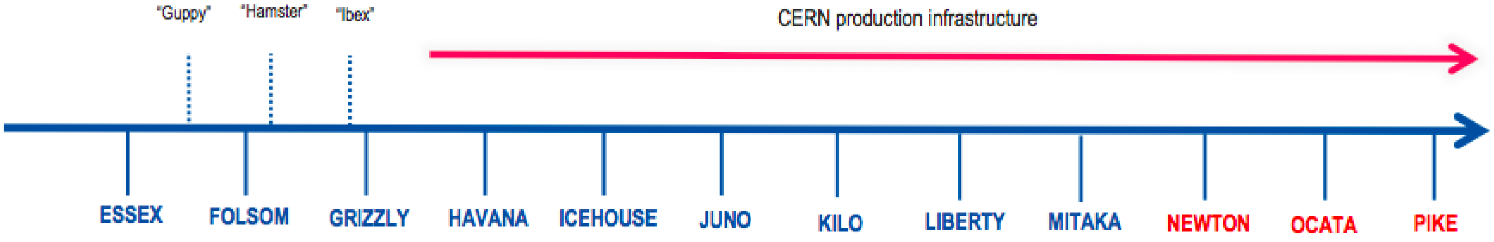
Magnum Deployment

- Clusters are described by *cluster templates*
- Shared/public templates for most common setups, customizable by users

```
$ openstack coe cluster create --name my-k8s --cluster-template kubernetes --node-count 100
             ~ 5 mins later
$ openstack coe cluster list
+------+------+---------+------------+--------------+--------------------+---------------+
| uuid | name | keypair | node_count | master_count | status             | health_status |
+------+------+---------+------------+--------------+--------------------+---------------+
| ...  | kube | default |          3 |            1 | UPDATE_IN_PROGRESS | HEALTHY       |
+------+------+---------+------------+--------------+--------------------+---------------+
$ $(openstack coe cluster config my-k8s --dir clusters/my-k8s --use-keystone)
$ OS_TOKEN=$(openstack token issue -c id -f value)
$ kubectl get ...
```

Resource Provisioning: IaaS

- Based on OpenStack
  - Collection of open source projects for cloud orchestration
  - Started by NASA and Rackspace in 2010
  - Grown into a global software community



NUMA roll-out

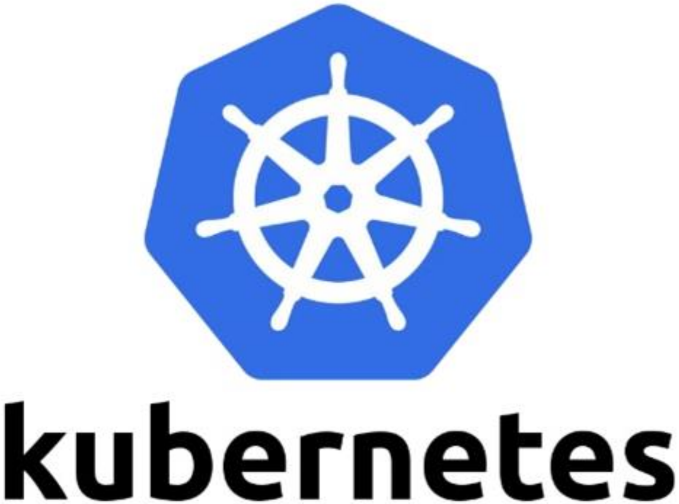
- Rolled out on ~2’000 batch hypervisors (~6’000 VMs)
    - HP allocation as boot parameter → reboot
    - VM NUMA awareness as flavor metadata → delete/recreate

- Cell-by-cell (~200 hosts):
    - Queue-reshuffle to minimize resource impact
    - Draining & deletion of batch VMs
    - Hypervisor reconfiguration (Puppet) & reboot
    - Recreation of batch VMs

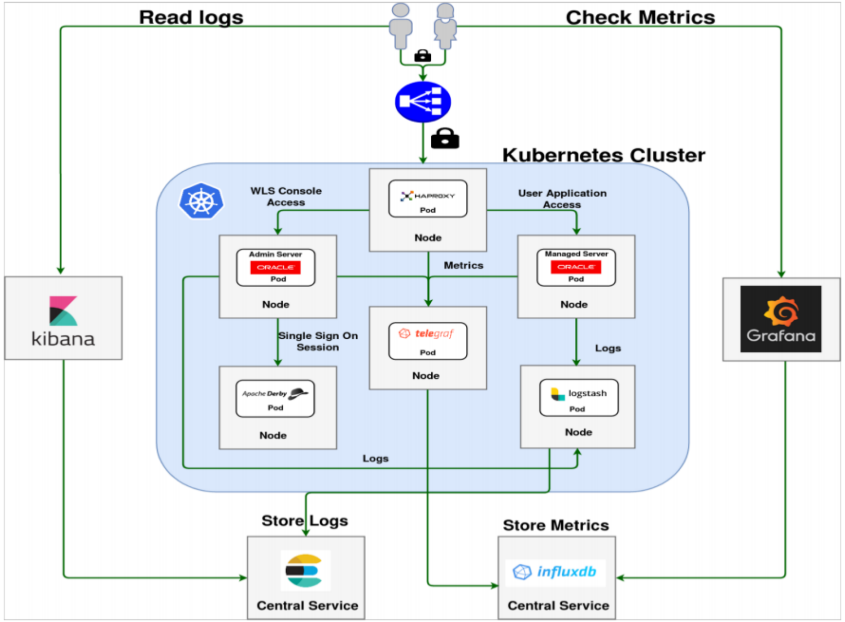
- Whole update took about 8 weeks
    - Organized between batch and cloud teams
    - No performance issue observed since

VM	Before	After
4x 8	8%	
2x 16	16%	
1x 24	20%	5%
1x 32	20%	3%

Container Orchestrators



Use case: Weblogic on kubernetes



- Many kubernetes clusters provisioned with OpenStack/Magnum
- Manila shares backed by cephfs for shared storage
- Central GitLab container registry
- Keystone Webhook for user AuthN

Credit: Antonio Nappi, CERN IT-DB

What is Magnum?

An OpenStack API service that allows creation of container clusters.

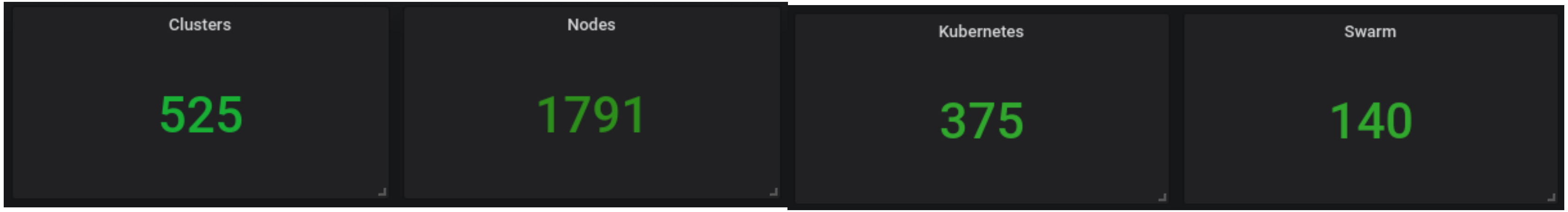
- Use your keystone credentials
- You choose your cluster type
  - Kubernetes
  - Docker Swarm
  - Mesos
- Single-tenant clusters
- Quickly create new clusters with advanced feature such as multi-master

MAGNUM

an OpenStack Community Project

CERN Magnum Deployment

- In production since 2016
- Running OpenStack Rocky release
- Working closely with upstream development
  - Slightly patched to adapt to the CERN network



Clusters
525

Nodes
1791

Kubernetes
375

Swarm
140

Magnum Cluster

A Magnum cluster is composed of:

- compute instances (virtual or physical)
- OpenStack Neutron networks
- security groups
- OpenStack Cinder for block volumes
- other resources (eg Load Balancer)
- OpenStack Heat to orchestrate the nodes

- Where your containers run
- Lifecycle operations
  - Scale up/down
  - Autoscale
  - Upgrade
  - Node heal/replace
- Self contained cluster with each own monitoring, data store, additional resources

Why use Magnum?

- Centrally managed self-service like GKE and AKS
  - Provide clusters to users with one-click deployment (or one API call)
  - Users don’t need to be system administrators
- Accounting comes for free if you use quotas in your projects
- Easy entrypoint to containers for new users
- Control your users’ deployments
  - OS
  - Monitoring

CERN Storage Integration

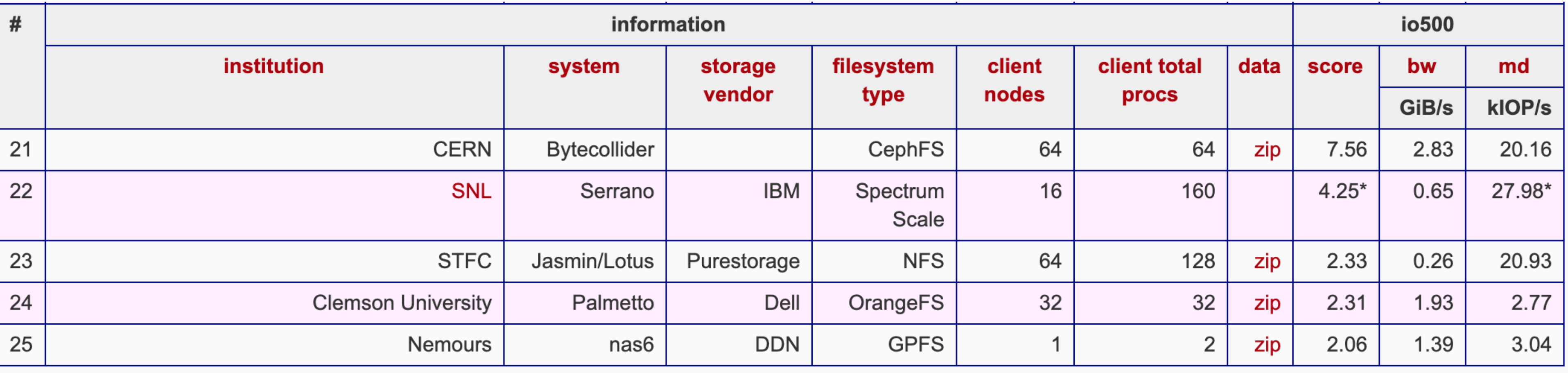
- CSI CephFS
  - Provides an interface between a CSI-enabled Container Orchestrator and the Ceph cluster
  - Provisions and mounts CephFS volumes
  - Supports both the kernel CephFS client and the CephFS FUSE driver
  - https://github.com/ceph/ceph-csi
- OpenStack Manila External Provisioner
  - Provisions new Manila shares, fetches existing ones
  - Maps them to Kubernetes PersistentVolume objects
  - Currently supports CephFS shares only (both in-tree CephFS plugin and csi-cephfs)
  - https://github.com/kubernetes/cloud-provider-openstack/tree/master/pkg/share/manila

Detailed results at https://techblog.web.cern.ch/techblog/post/container-storage-cephfs-scale-part3/

Credit: Robert Vasek, CERN Cloud

io500 – entered for 2018

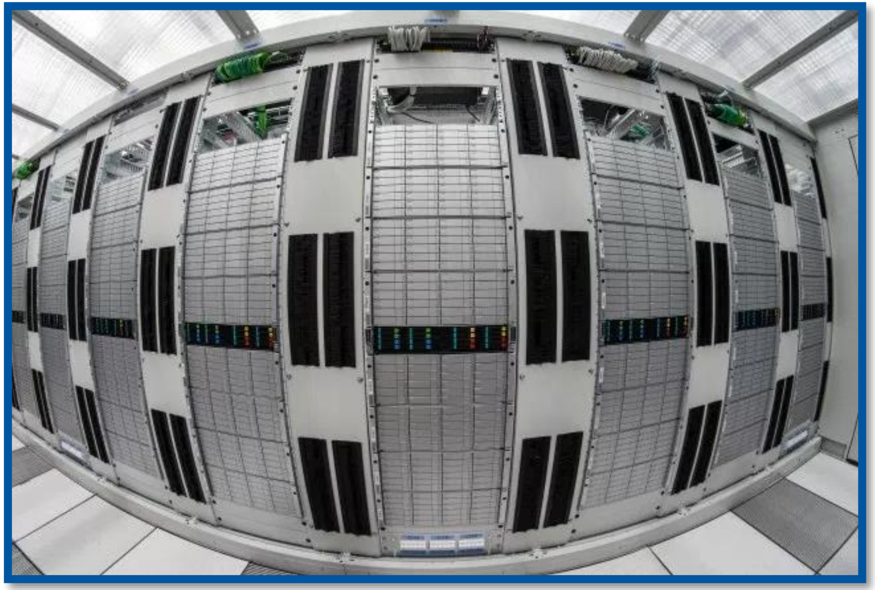
- Using CephFS on SSDs and Lazy IO, we made it onto the io500 list at #21 - https://www.vi4io.org/io500/start



#	information							io500		
	institution	**system**	**storage vendor**	**filesystem type**	**client nodes**	**client total procs**	**data**	**score**	**bw**	**md**
									GiB/s	**kIOP/s**
21	CERN	Bytecollider		CephFS	64	64	zip	7.56	2.83	20.16
22	SNL	Serrano	IBM	Spectrum Scale	16	160		4.25*	0.65	27.98*
23	STFC	Jasmin/Lotus	Purestorage	NFS	64	128	zip	2.33	0.26	20.93
24	Clemson University	Palmetto	Dell	OrangeFS	32	32	zip	2.31	1.93	2.77
25	Nemours	nas6	DDN	GPFS	1	2	zip	2.06	1.39	3.04

Bigbang Scale Tests

- *Bigbang* scale tests mutually benefit CERN & Ceph project
- *Bigbang I:* 30PB, 7200 OSDs, Ceph hammer. Several *osdmap* limitations
- *Bigbang II:* Similar size, Ceph jewel. Scalability limited by OSD/MON messaging. Motivated *ceph-mgr*
- *Bigbang III:* 65PB, 10800 OSDs



https://ceph.com/community/new-luminous-scalability/

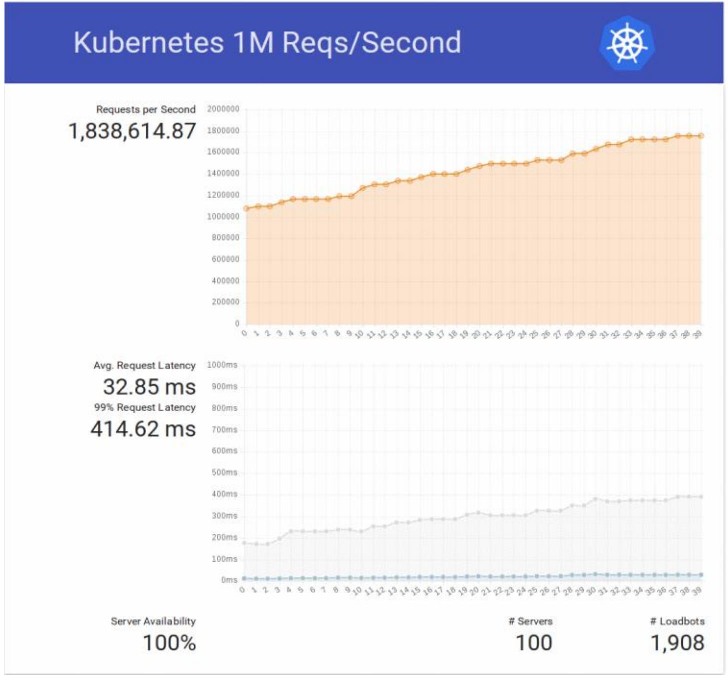
CERN Storage Integration

- CVMFS provides us with a massively scalable read-only file system
- Static content like compiled applications and conditions data
- Provides an interface between a CSI-enabled Container Orchestrator and the CERN application appliances
- https://github.com/cernops/cvmfs-csi/

Credit: Robert Vasek and Ricardo Rocha, CERN Cloud

First Attempt – 1M requests/Seq

- 200 Nodes
- Found multiple limits
  - Heat Orchestration scaling
  - Authentication caches
  - Volume deletion
  - Site services



Kubernetes 1M Reqs/Second

Requests per Second
1,838,614.87

Avg. Request Latency
32.85 ms

99% Request Latency
414.62 ms

Server Availability
100%

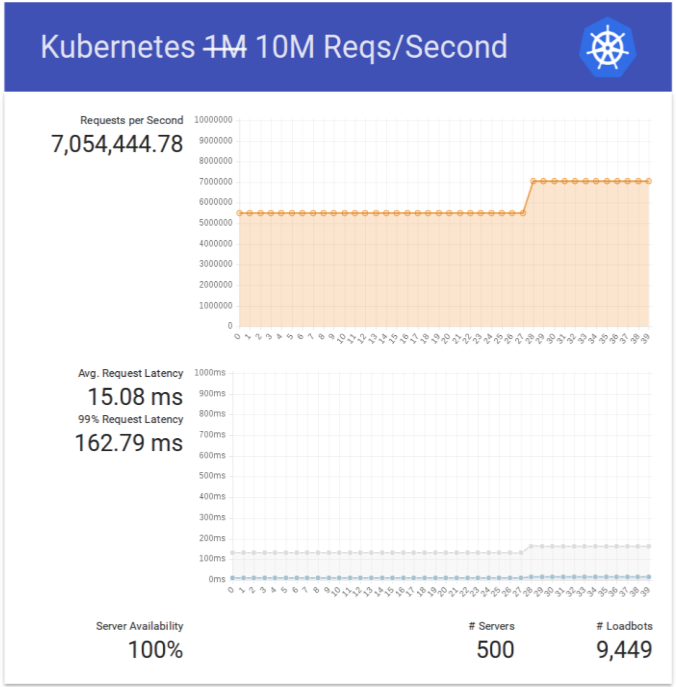
Servers
100

Loadbots
1,908

Second Attempt – 7M requests/Seq

- Fixes and scale to 1000 Nodes

Cluster Size (Nodes)	Concurrency	Deployment Time (min)
2	50	2.5
16	10	4
32	10	4
128	5	5.5
512	1	14
1000	1	23



Kubernetes ~~1M~~ 10M Reqs/Second

Requests per Second
7,054,444.78

Avg. Request Latency
15.08 ms

99% Request Latency
162.79 ms

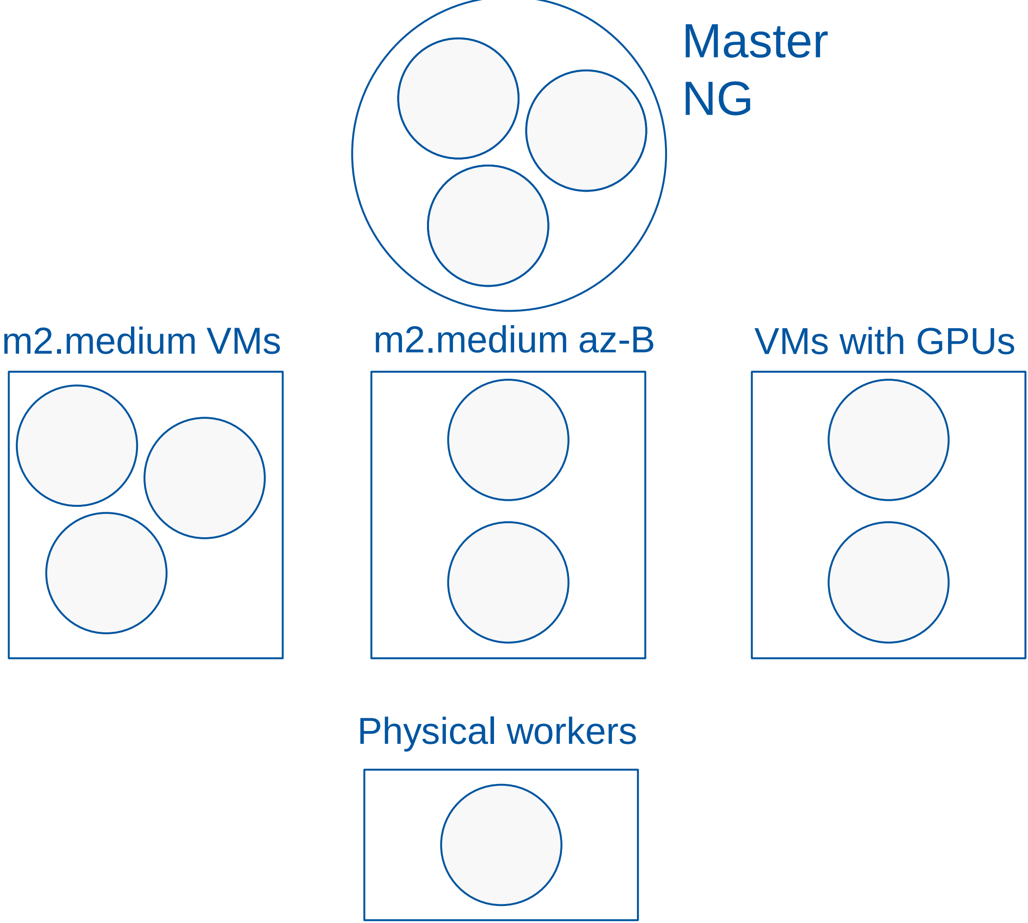
Server Availability
100%

Servers
500

Loadbots
9,449

Node Groups

- Define subclusters
- Vary Flavors
  - Small/Big VMs
  - Bare Metal
- Vary Zones
  - Improve redundancy



Auto Scaling

https://github.com/kubernetes/autoscaler

Optimize resource usage

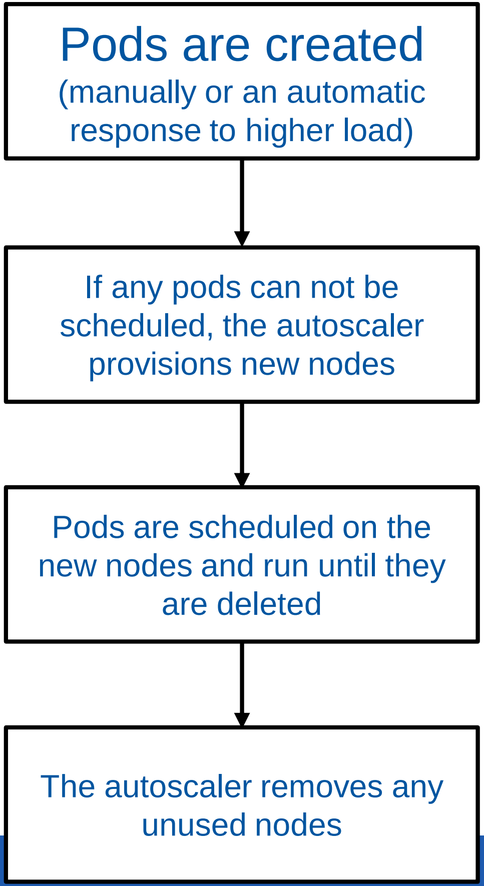
Dynamically resize the cluster based on current number of pods, and their required CPU / memory

New cloud provider for Magnum

Docs at autoscaler / cluster-autoscaler / cloudprovider / magnum

Merged PR: https://github.com/kubernetes/autoscaler/pull/1690

Slide Credit: Thomas Hartland, CERN Cloud



Pods are created
(manually or an automatic response to higher load)

If any pods can not be scheduled, the autoscaler provisions new nodes

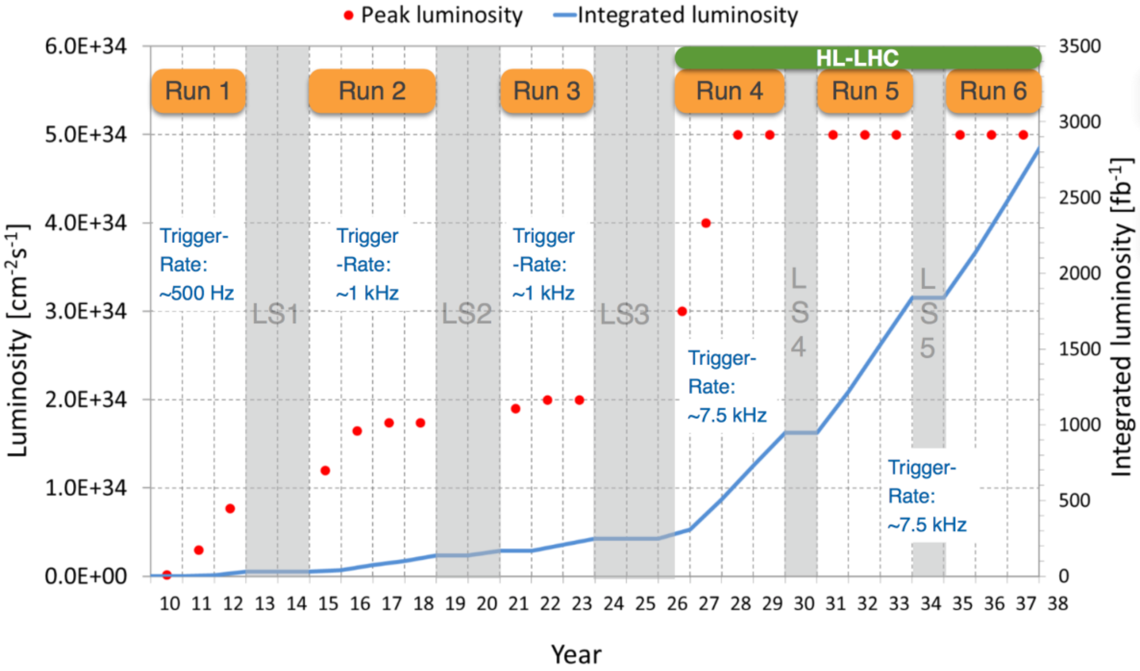
Pods are scheduled on the new nodes and run until they are deleted

The autoscaler removes any unused nodes

More enhancements rolling out

- Authentication using OpenStack Keystone
  - Native kubectl commands with cloud credentials
- Choice of Ingress controller
  - Nginx or Traefik
- Integrated monitoring with Prometheus
- Rolling cluster upgrades for Kubernetes, operating system and add-ons
- Integrated Node Problem Detector

LHC Schedule



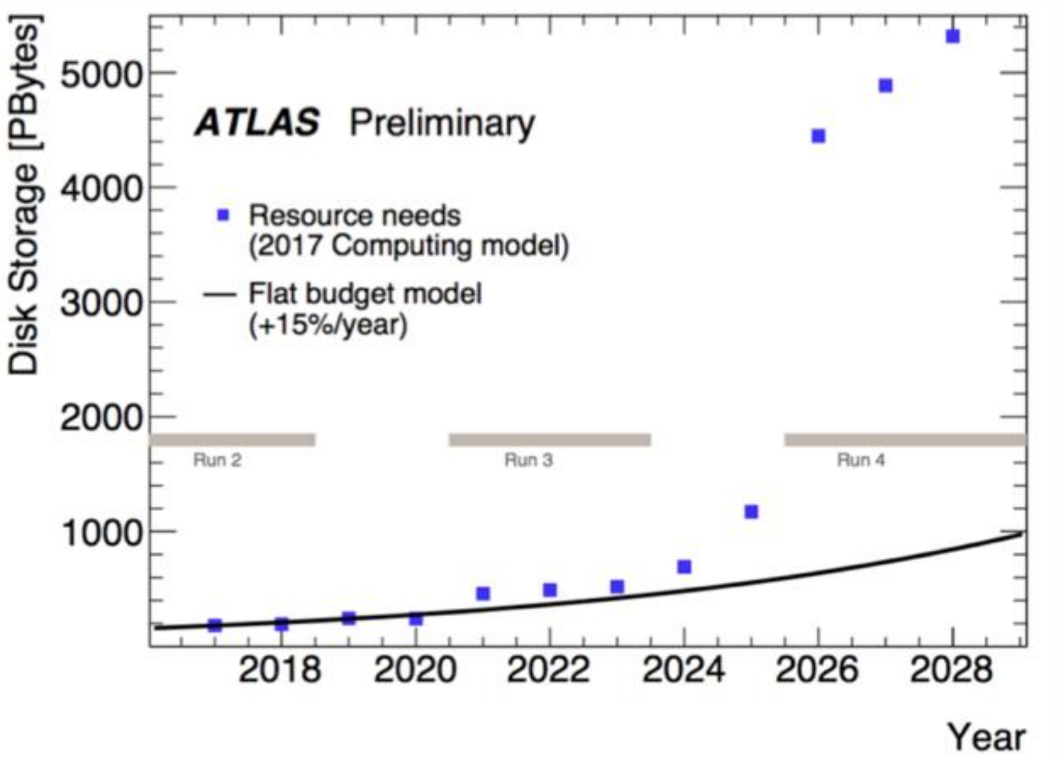
Run 3 — Alice, LHCb upgrades

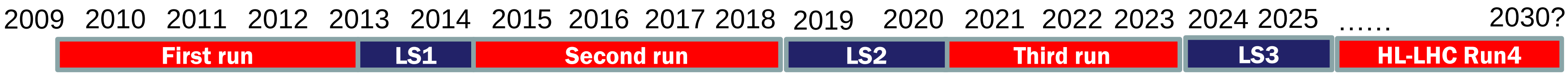
Run 4 — ATLAS, CMS upgrades

2026

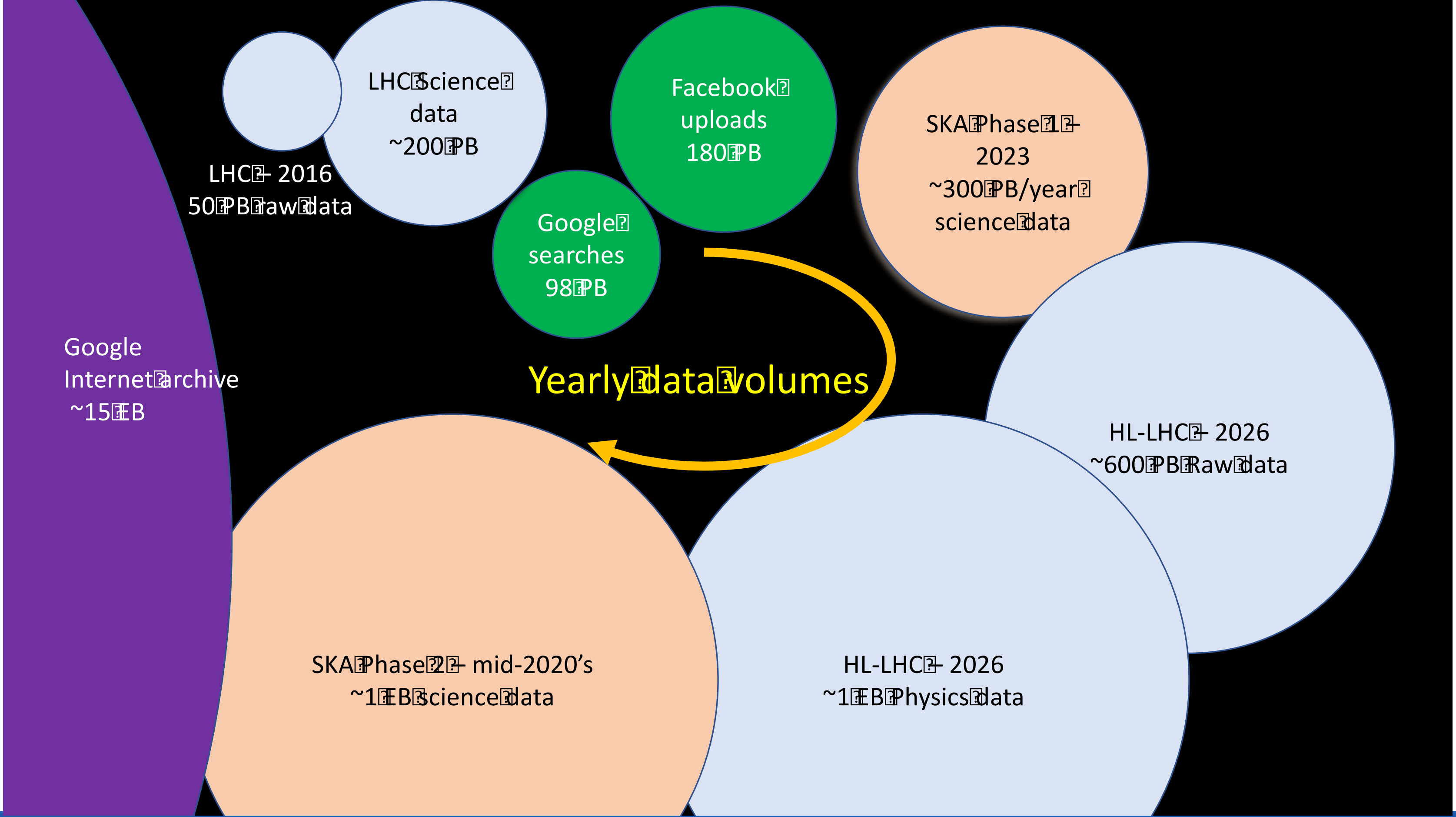
Raw data volume increases significantly for **High Luminosity LHC**

- Significant part of cost comes from global operations
- Even with technology increase of ~15%/year, we still have a big gap if we keep trying to do things with our current compute models




2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 …… 2030?

First run | **LS1** | **Second run** | **LS2** | **Third run** | **LS3** | **HL-LHC Run4**



LHC – 2016
50 PB raw data

LHC Science
data
~200 PB

Facebook
uploads
180 PB

Google
searches
98 PB

SKA Phase 1 –
2023
~300 PB/year
science data

Google
Internet archive
~15 EB

Yearly data volumes

HL-LHC – 2026
~600 PB Raw data

SKA Phase 2 – mid-2020's
~1 EB science data

HL-LHC – 2026
~1 EB Physics data

Radio

JIVE-VLBI

SKA

Visible light

ELT

ESO

EST

Gamma rays

CTA

Accelerator-based
Particle Physics

HL-LHC

CERN

Accelerator-based
Nuclear Physics

FAIR

Gravitational Waves

EGO-VIRGO

Cosmic-rays
Neutrinos

KM3NeT

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures

Future of particle physics ?

High Luminosity LHC until 2035

- Ten times more collisions than the original design

Studies in progress:
Compact Linear Collider (CLIC)

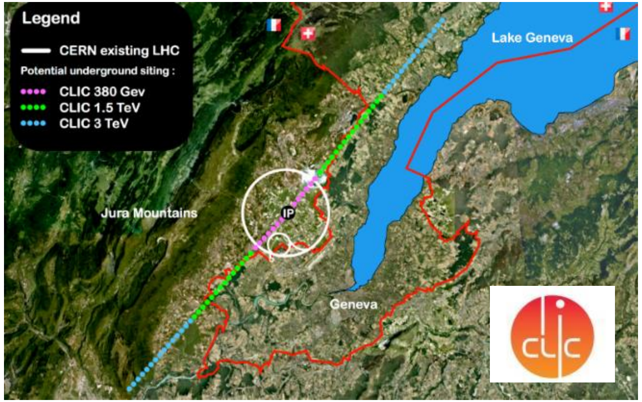
- Up to 50Km long
- Linear e^+e^- collider $\sqrt{s}$ up to 3 TeV

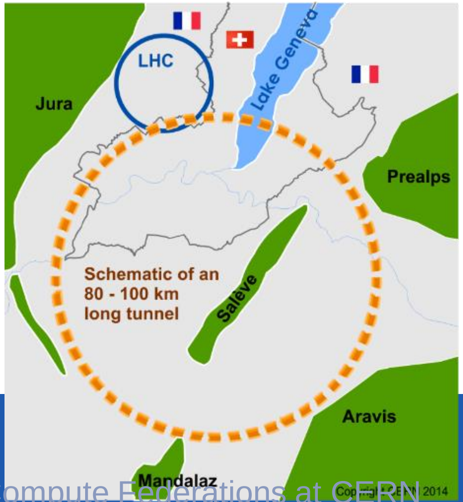
Future Circular Collider (FCC)

- ~100 Km circumference
- New technology magnets → 100 TeV pp collisions in 100km ring
- e^+e^- collider (FCC-ee) as 1st step?

European Strategy for Particle Physics

- Preparing next update in 2020


Legend
CERN existing LHC
Potential underground siting :
CLIC 380 Gev
CLIC 1.5 TeV
CLIC 3 TeV
Lake Geneva
Jura Mountains
IP
Geneva
CLIC


LHC
Lake Geneva
Jura
Prealps
Schematic of an
80 - 100 km
long tunnel
Salève
Aravis
Mandalaz
Copyright CERN 2014