CERN

CERN Cloud Infrastructure Report

Arne Wiebalck
for the CERN Cloud Team

HEPiX Autumn Meeting
BNL, Upton, N.Y., U.S.
Oct 12, 2015

Numbers

What’s new

Operations

WIP

CERN Cloud Recap

- CERN Cloud Service one of the three major components in IT's AI project
  - Policy: Servers in CERN IT shall be virtual

- Based on OpenStack
  - Production service since July 2013
  - Performed three rolling upgrades since
  - In transition from Juno to Kilo
  - Nova, Glance, Keystone, Horizon, Cinder, Ceilometer, Heat

CERN Cloud Architecture (1)

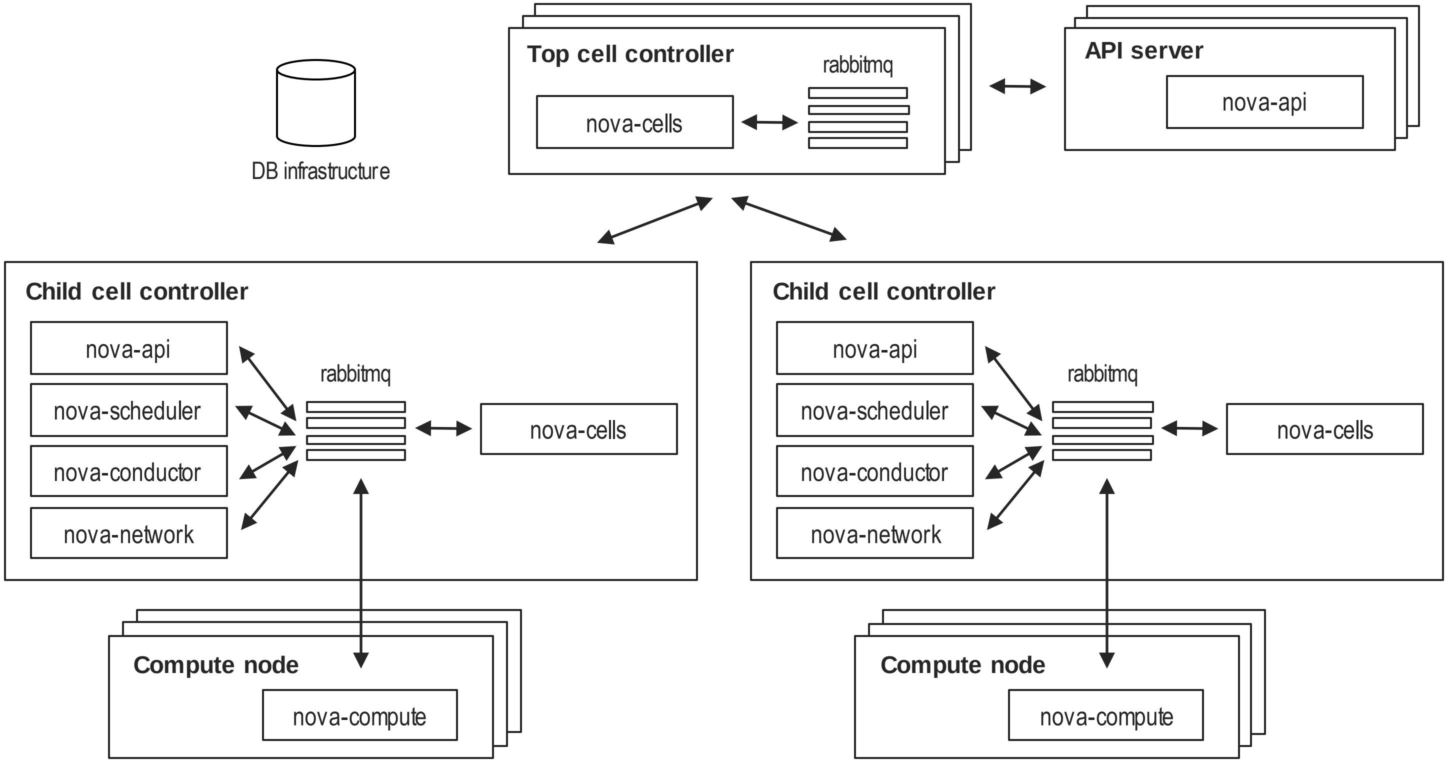
- Two data centers
  - 1 region (1 API), 26 cells
  - Cells map use cases
    hardware, hypervisor type, location, users, …

DANTE
100 GbE
Wigner RCP
T-Systems
100 GbE
CERN
Google earth

- Top cell on several physical nodes in HA
  - Clustered RabbitMQ with mirrored queues
  - API servers are VMs in various child cells

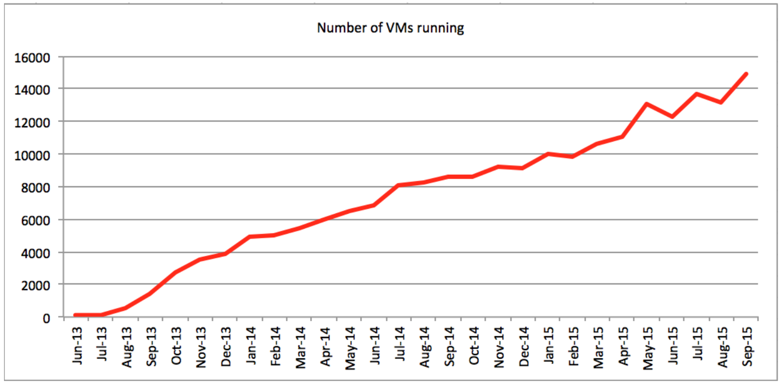
- Child cell controllers are OpenStack VMs
  - **One** controller per cell
  - Tradeoff between complexity and failure impact

CERN Cloud Architecture (2)

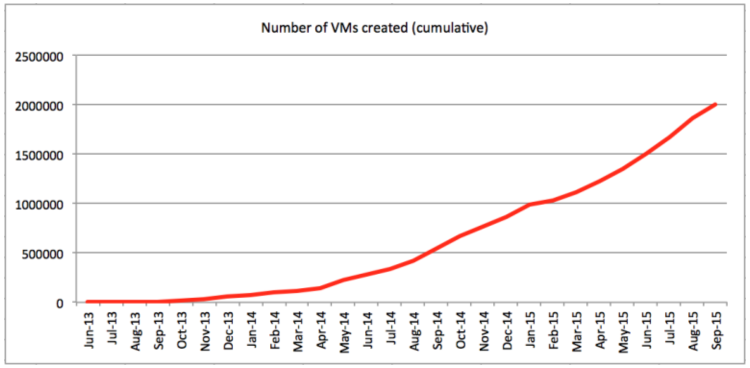


CERN Cloud in Numbers (1)

- 4’600 hypervisors in production (1y ago: 3000)
  - Majority qemu/kvm now on CC7 (~150 Hyper-V hosts) (SLC6)
  - ~2’000 HVs at Wigner in Hungary (batch, compute, services) (batch)
  - 250 HVs on critical power

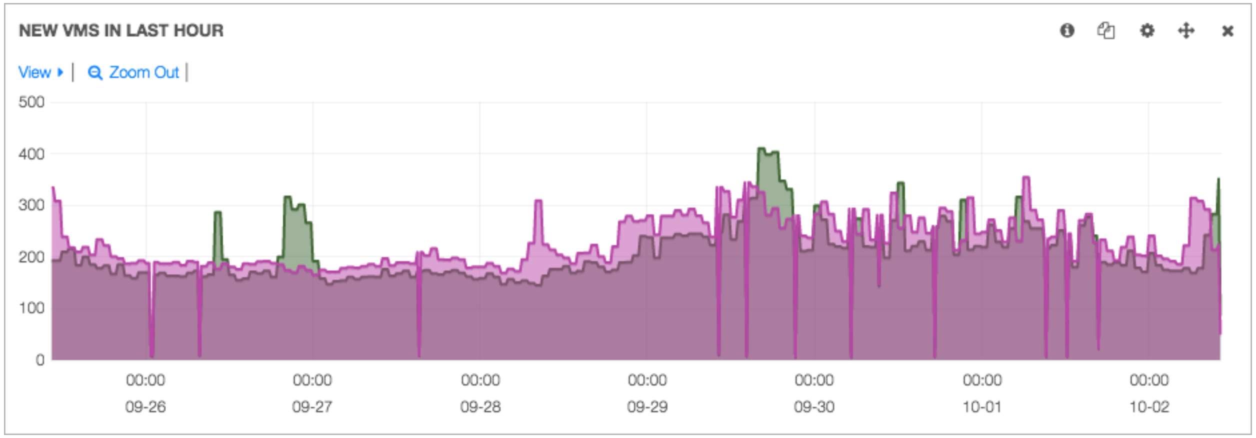


- 125k Cores (64k)
- 250 TB RAM (128TB)
- ~15’000 VMs (8’000)



- To be increased in 2016!
  - +65k cores until spring

CERN Cloud in Numbers (2)



Every 10s a VM gets created or deleted in our cloud!

- 2’000 images/snapshots (1’100)
  - Glance on Ceph
- 1’500 volumes (600)
  - Cinder on Ceph (& NetApp)

What’s new: Volume Types

- Extended list of available volume types
  - More performance, critical power, Windows, DR

Name	IOPS	Feature	Backend
standard	100	-	ceph
io1	500	QoS	ceph
cp1	100	critical power	ceph
cpio1	500	critical power	ceph
cp2	100	Windows	NetApp
wig-cp1	100	@Wigner	ceph (Wigner)
wig-cpio1	500	@Wigner	ceph (Wigner)

What’s new: Heat in production

- Orchestration of OpenStack resources through templates
  - Creation of a set of machines
  - Automatic, trigger-driven scaling

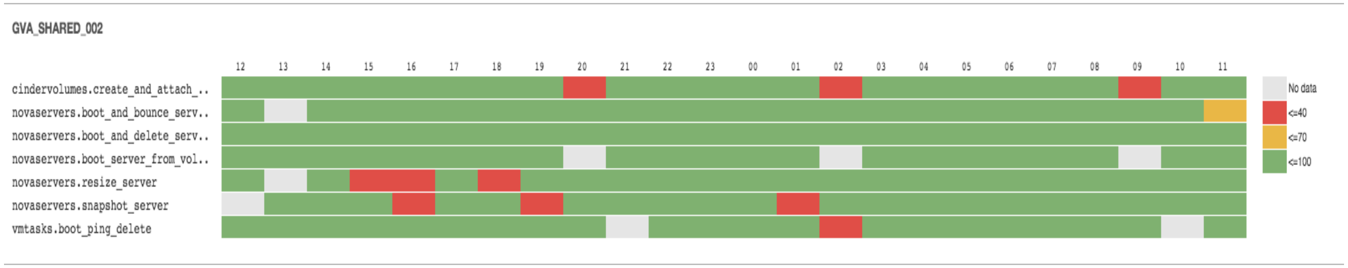
- In production (and already upgraded!)

- Templates & plugins that ease the CERN integration (SSO, Puppet)

- First users
  - IT Monitoring team creates ES servers via templates
  - CMS Tier-0 for maximum quota usage

What’s new: Rally in production

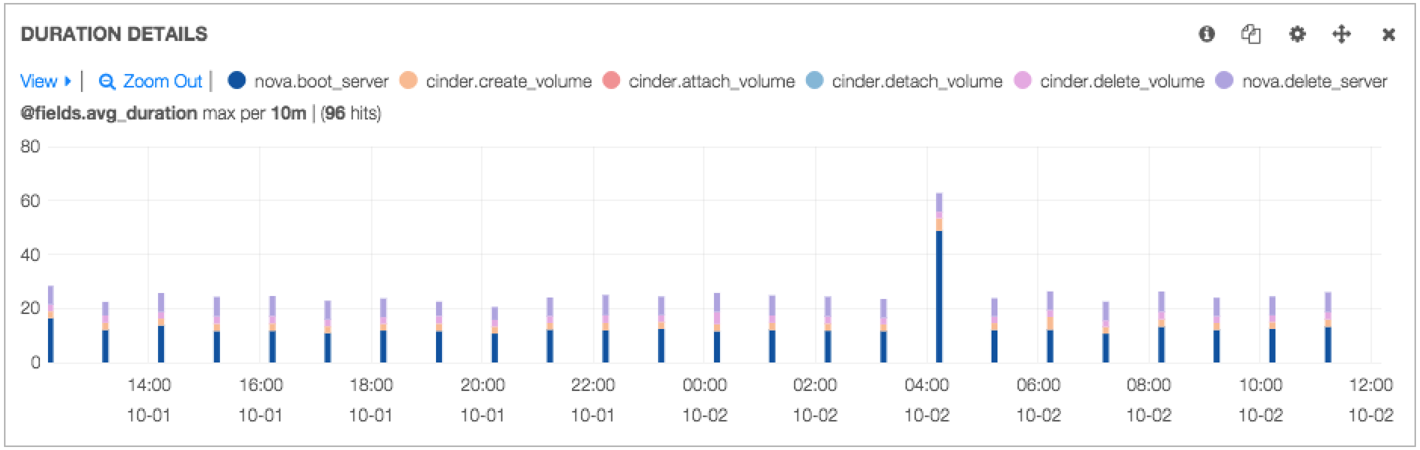
- Benchmarking & Verification for OpenStack



GVA_SHARED_002

	12	13	14	15	16	17	18	19	20	21	22	23	00	01	02	03	04	05	06	07	08	09	10	11
cindervolumes.create_and_attach_..	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
novaservers.boot_and_bounce_serv..	🟩	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨
novaservers.boot_and_delete_serv..	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
novaservers.boot_server_from_vol..	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩
novaservers.resize_server	🟩	⬜	🟩	🟥	🟥	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
novaservers.snapshot_server	⬜	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
vmtasks.boot_ping_delete	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩

⬜ No data
🟥 <=40
🟨 <=70
🟩 <=100

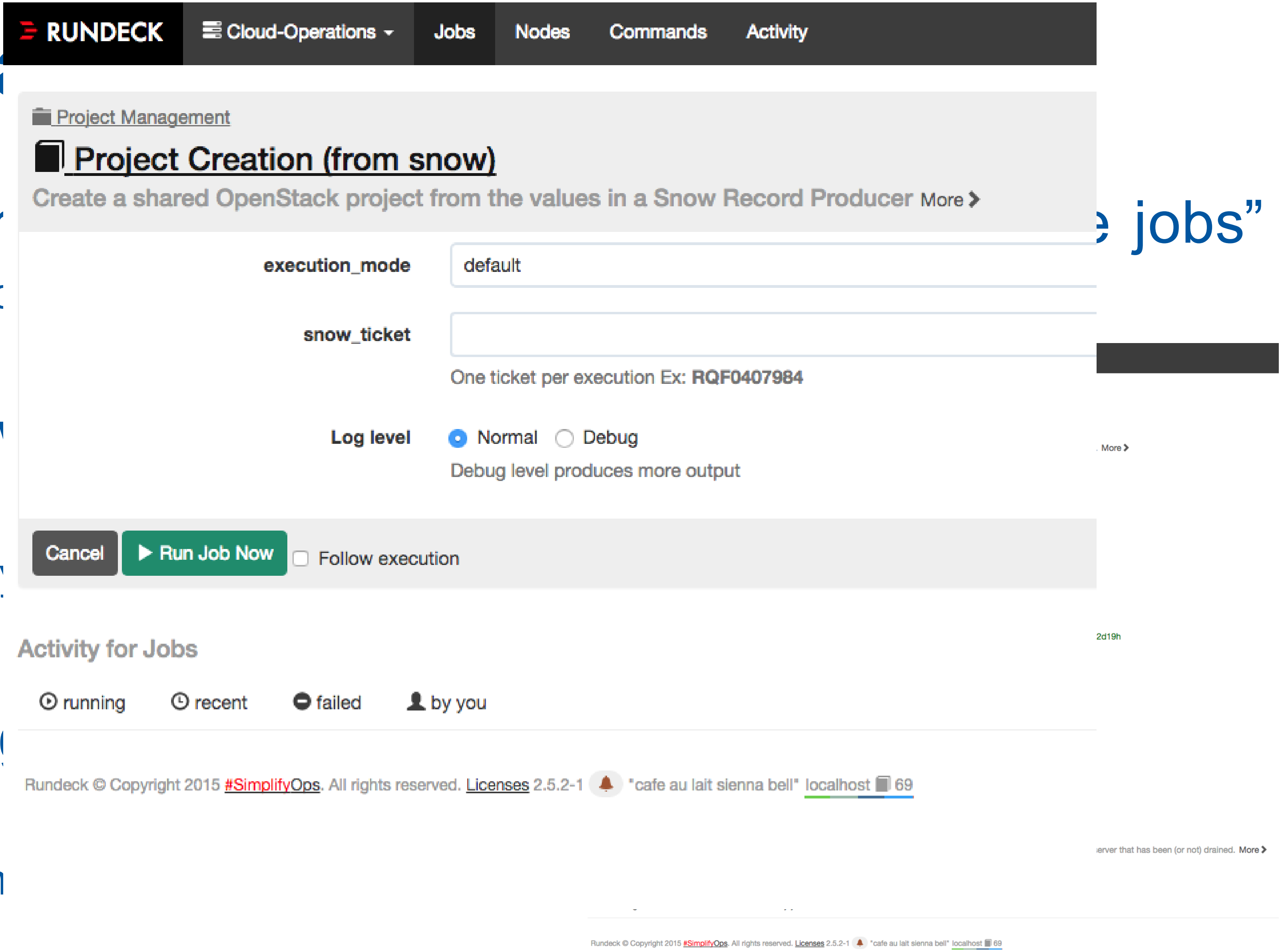


DURATION DETAILS

@fields.avg_duration max per 10m | (96 hits)

Wha

- “Tur e jobs”
  - Jo
- Allo of p
  - Sy
- Inte ope
  - In



Project Management

Project Creation (from snow)

Create a shared OpenStack project from the values in a Snow Record Producer More ›

execution_mode: default

snow_ticket:

One ticket per execution Ex: **RQF0407984**

Log level: ● Normal ○ Debug

Debug level produces more output

□ Follow execution

Activity for Jobs

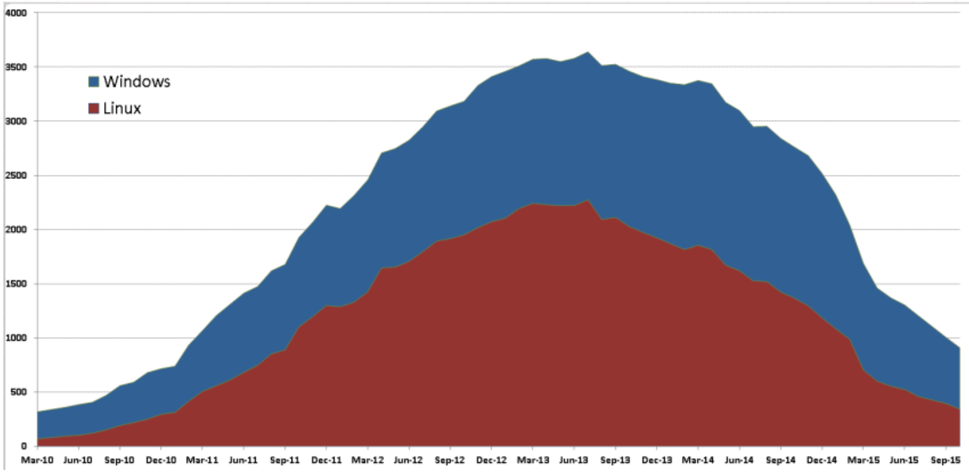
Rundeck © Copyright 2015 #SimplifyOps. All rights reserved. Licenses 2.5.2-1 "cafe au lait sienna bell" localhost 69

erver that has been (or not) drained. More ›

Rundeck © Copyright 2015 #SimplifyOps. All rights reserved. Licenses 2.5.2-1 "cafe au lait sienna bell" localhost 69

Operations: CVI Phase-out

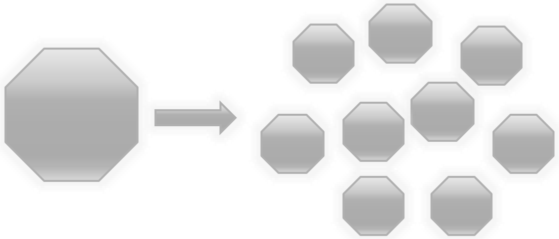
- Well underway
  - Creation blocked since summer 2014
  - 70% of CVI VMs gone



- Strategy
  - Delete/re-create where possible, migrate where necessary
  - In close collaboration with users
  - 400/650 machines migrated, physical hosts migrated as well

- Goal: less than 100 VMs by the end of the year

Operations: Cell split

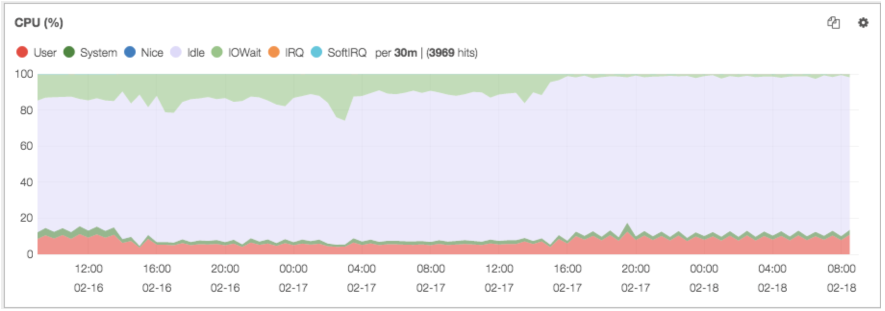


- The service's initial cell (cell01) contained around 1'000 compute nodes
  - KVM & Hyper-V, aggregates, different h/w, all AVZ, …
  - Mostly service nodes → important → HA control plane
  - Simply grew beyond all usual recommendations ☺
- We split that cell into 9 smaller cells … live!
  - New child cell controllers
  - Clone instance DB, remove all entries not in the new cell
  - Move compute nodes to new cell
  - Change instances' cell path in top level DB

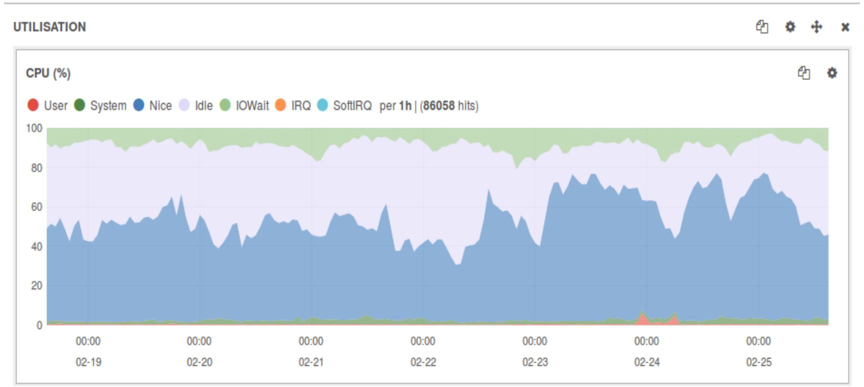
Operations: KVM Caching

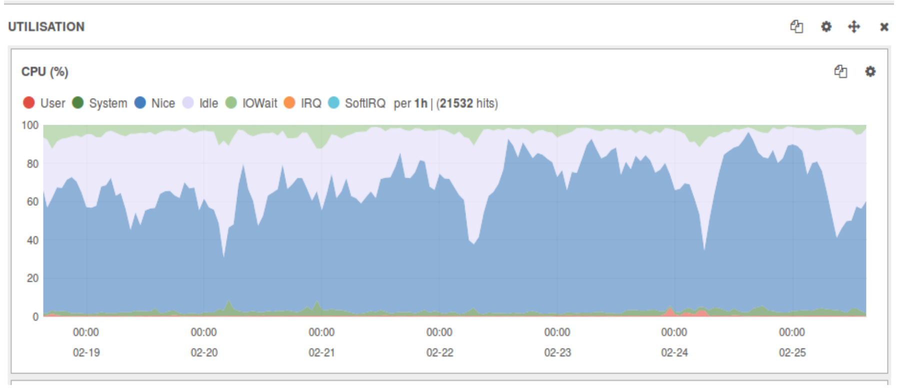
ATLAS SAM VM

(‘none’ to ‘write-back’)

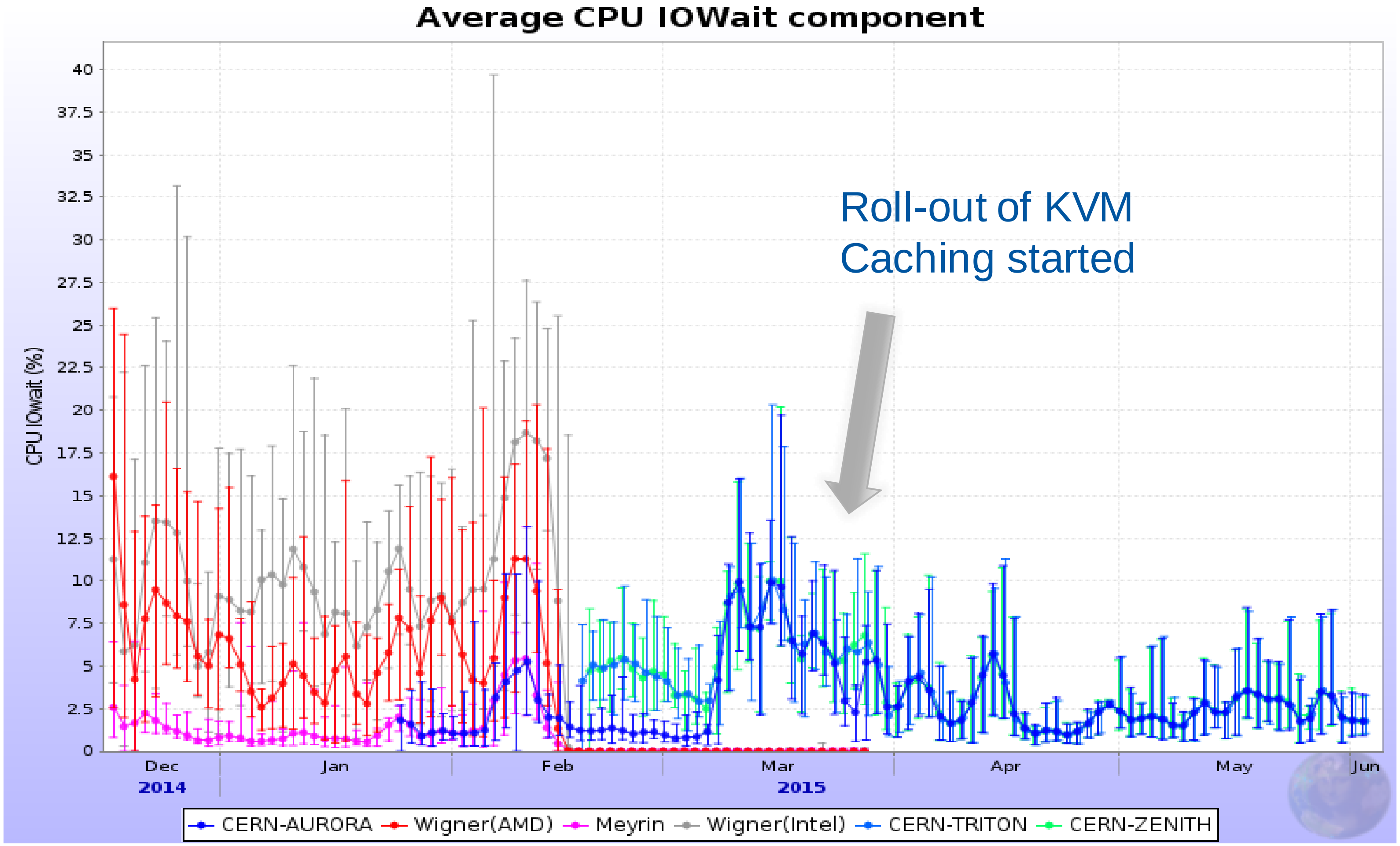


‘write-back’ rolled out on batch/compute



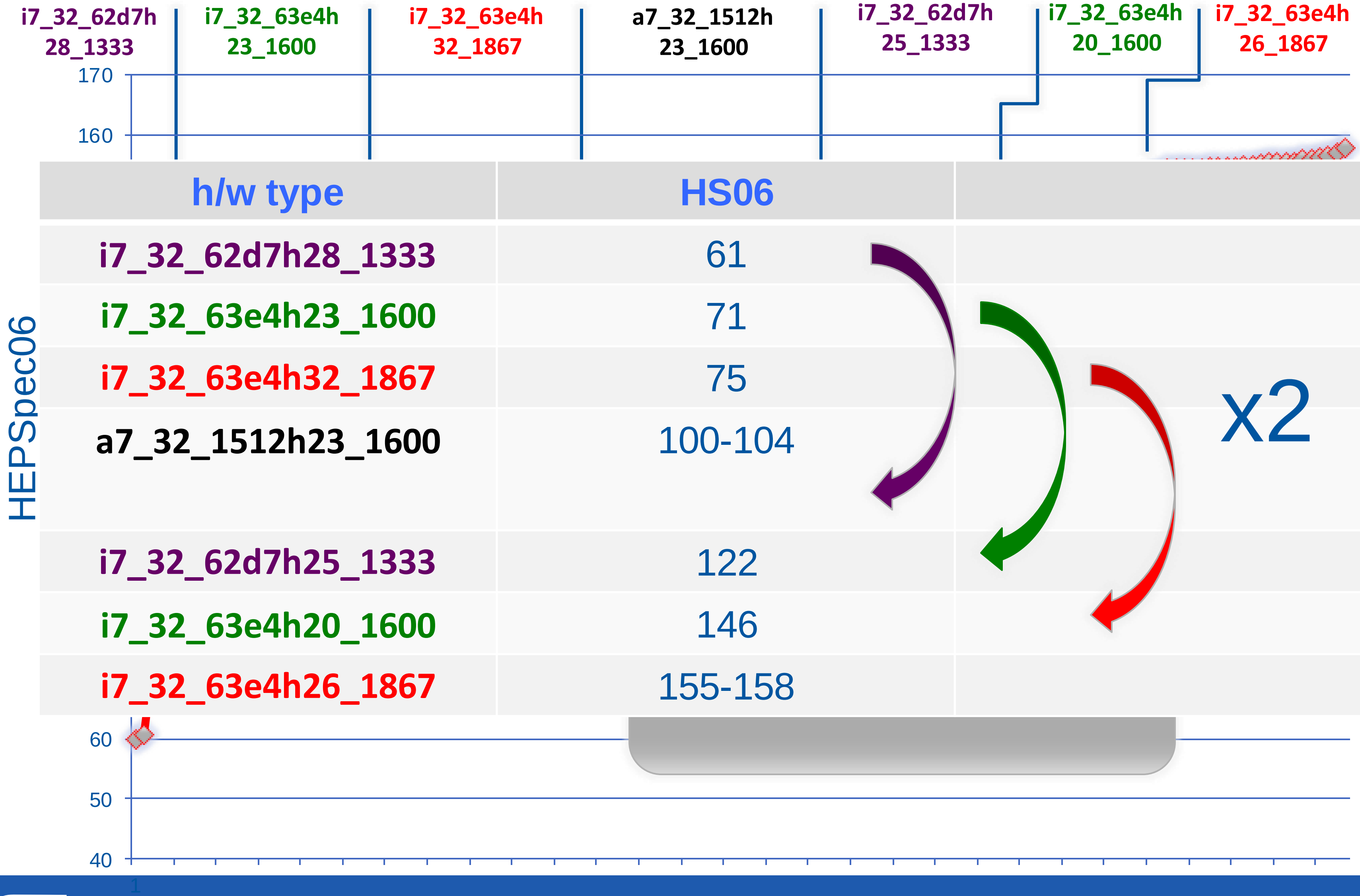


KVM Caching: ALICE



Roll-out of KVM Caching started

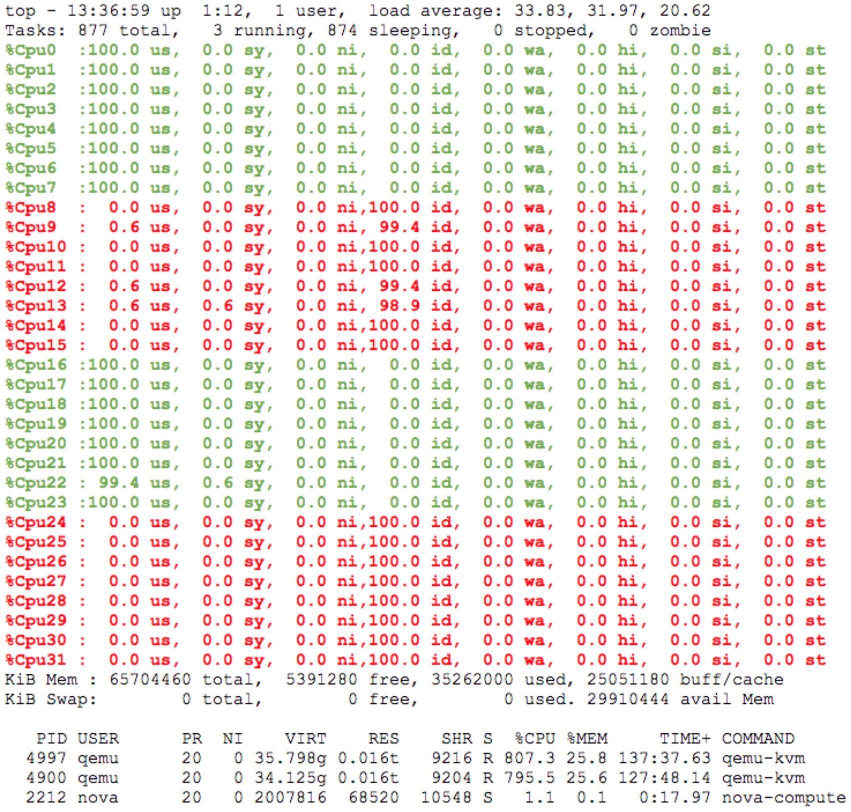
Operations: CPU Puzzle (1/2)



h/w type	HS06	
i7_32_62d7h28_1333	61	
i7_32_63e4h23_1600	71	
i7_32_63e4h32_1867	75	
a7_32_1512h23_1600	100-104	
i7_32_62d7h25_1333	122	
i7_32_63e4h20_1600	146	
i7_32_63e4h26_1867	155-158	

x2

Operations: CPU Puzzle (1/2)



```
top - 13:36:59 up  1:12,  1 user,  load average: 33.83, 31.97, 20.62
Tasks: 877 total,   3 running, 874 sleeping,   0 stopped,   0 zombie
%Cpu0  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu1  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu2  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu3  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu4  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu5  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu6  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu7  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu8  :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu9  :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu10 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu11 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu12 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu13 :  0.6 us,  0.6 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu14 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu15 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu16 :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu17 :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu18 :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu19 :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu20 :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu21 :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu22 : 99.4 us,  0.6 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu23 :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu24 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu25 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu26 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu27 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu28 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu29 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu30 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu31 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 65704460 total,  5391280 free, 35262000 used, 25051180 buff/cache
KiB Swap:        0 total,        0 free,        0 used. 29910444 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
 4997 qemu      20   0 35.798g 0.016t   9216 R 807.3 25.8 137:37.63 qemu-kvm
 4900 qemu      20   0 34.125g 0.016t   9204 R 795.5 25.6 127:48.14 qemu-kvm
 2212 nova      20   0 2007816  68520  10548 S   1.1  0.1   0:17.97 nova-compute
```

```
[root@hv001 ~]# virsh list --all
 Id    Name                           State
----------------------------------------------------
 3     instance-00002a86              running
 4     instance-00002a74              running

[root@hv001 ~]# virsh vcpupin instance-00002a86
VCPU: CPU Affinity
----------------------------------
   0: 0-7,16-23
   1: 0-7,16-23
   2: 0-7,16-23
   3: 0-7,16-23
   4: 0-7,16-23
   5: 0-7,16-23
   6: 0-7,16-23
…

[root@hv001 ~]# virsh vcpupin instance-00002a74
VCPU: CPU Affinity
----------------------------------
   0: 0-7,16-23
   1: 0-7,16-23
   2: 0-7,16-23
   3: 0-7,16-23
   4: 0-7,16-23
   5: 0-7,16-23
   6: 0-7,16-23
…
```

The VMs were pinned to the same NUMA nodes!

https://bugs.launchpad.net/nova/+bug/1461777 (fixed in Kilo)

WIP: Container integration

- Started to look into integration of containers with our OpenStack deployment
  - Initially triggered by the prospect of low performance overheads
  - LXC due to the lack of an upstream Docker driver
    (not suitable for general purpose)

- We’ve setup a test cell
  - Performance looks good
  - OpenStack patches for AFS & CVMFS done
  - AFS in containers: kernel access, multiple containers, tokens, …

- Started to look into OpenStack Magnum
  - Container orchestration via Docker or Kubernetes become first class OpenStack resources
  - More details probably already at next workshop

WIP: Life-cycle management

- Hardware in former cell01 will soon reach EOL
  - VMs are mostly pets and run services
  - Users would like to keep their VMs

- Service nodes left in SLC6 → CC7 migration
  - Juno on RDO RHEL6 was difficult, but Kilo?

- The service needs to support live-migration!
  - Not used in daily operations: resources & network constraints
  - “IP service bridging” (see Carles’ talk yesterday)
  - VMs booted from volume: unproblematic, fast
  - VMs on ephemeral disks: **block** live-migration seems to work (from SLC6 to CC7 out-of-box, from CC7 after qemu version update)
  - VMs with volumes: needs volume detach

- We need tools to do this at scale so that live-migration can be become part of our daily operations.

Summary

- The CERN OpenStack Cloud evolved into a rapidly growing but very stable service
  - Enabled the doubling of Tier-0 resources since 2012
  - Will enable significant growth 2016

- We moved some new OpenStack projects into production and have some more under evaluation

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

CERN