Operating an OpenStack Cloud

Learning from building & operating SWITCHengines

SWITCH

Jens-Christian Fischer
jens-christian.fischer@switch.ch

SA7 T3, 26.11.2015

SWITCH

- Non Profit Foundation
- IT Services for Universities and Higher Education
  - Traditional Internet (SWITCHlan)
  - Federated Single Sign On (SWITCHaai)
  - Lecture Recording & Videos (SWITCHcast / SWITCHtube)
  - Many, many more

- A partner to Acacdemia
- Registry for .ch (since 1987)
- And Domain registration until 2015

© 2015 SWITCH

SWITCH Services

- Established services:
  - Network
  - Authentication and authorization infrastructure PKI, SWITCHham, SMS, videoconf
  - Security
  - Collaboration, e-learning services

- New(er) services:
  - Procurement
  - Cloud services

- Under development:
  - Swiss edu-ID, Information Security Management (ISMS)

© 2015 SWITCH

“Cloud”

- Infrastructure as a Service
  - SWITCHengines
- Software as a Service
  - SWITCHdrive – an Academic Sync & Share Service
  - SWITCHfilesender – overcome the 10 MB Mailbox limit
  - SWITCHcast – record and edit lectures
  - SWITCHtube – a secure video distribution tool

© 2015 SWITCH

Timeline for IaaS Cloud Offering

- 2012: Start of Project BCC – Building Cloud Competency
- 2013: Decision that SWITCH builds Swiss Academic Cloud services
- 2014: SCALE project. One of the deliverables – build an initial cloud infrastructure - 64 Servers purchased and installed
- 2015: Pilot usage with production load. End of SCALE project Start of SCALE-UP project
- 2016: SWITCHengines – an official SWITCH service
- 2017: End of SCALE-UP project

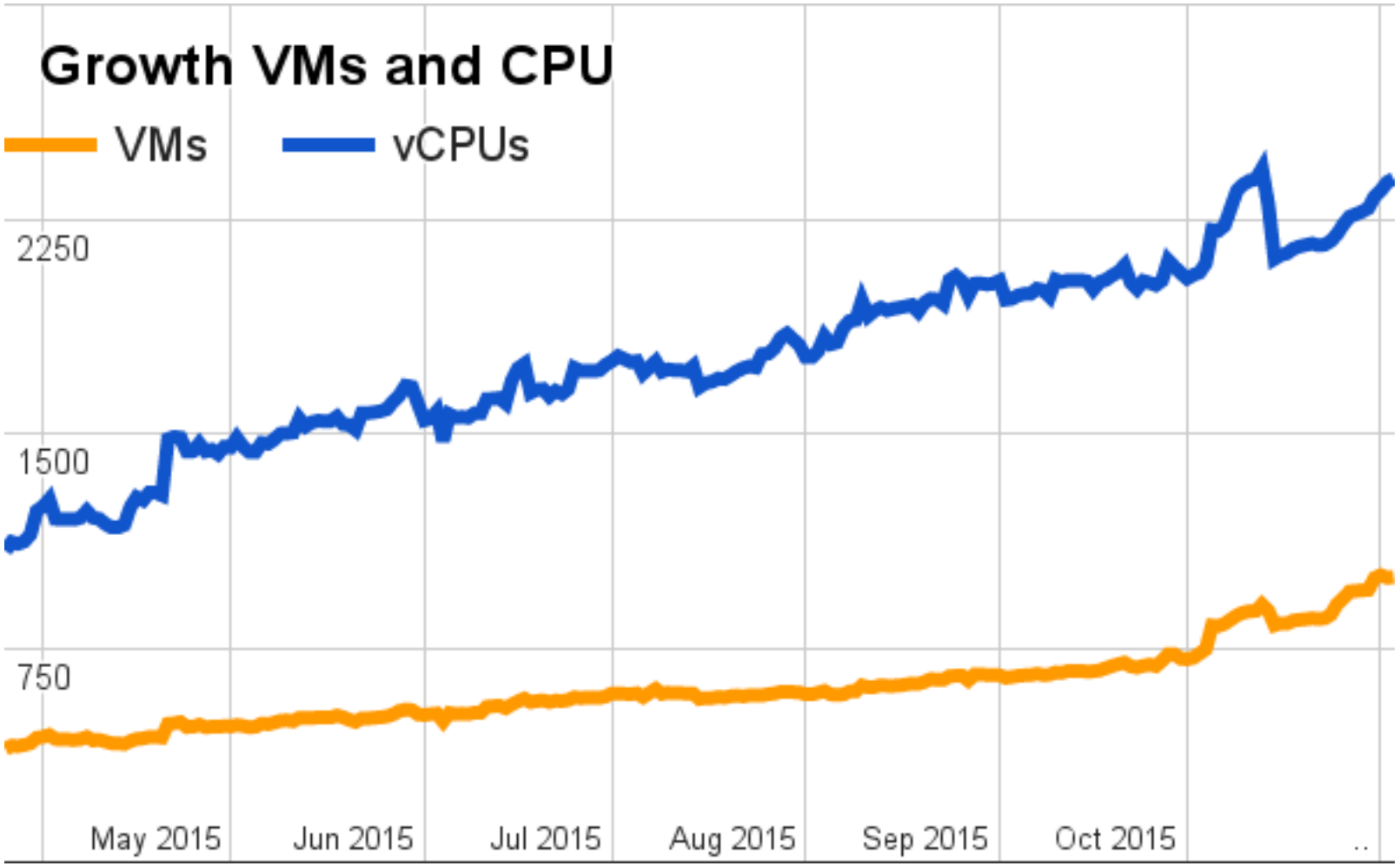
© 2015 SWITCH

Current Status

- SWITCHengines has been in “public beta” and “internal production” since December 2014
- Several SWITCH services run on it (SWITCHdrive, SWITCHengines, SWITCHfilesender, SWITCHtube)
- Over 630 individual users and around 40 research projects online
- Around 1000 VMs running on the cloud right now (2200 cores, 3700 GB RAM, 360 TB of storage allocated)
- Over 60’000 VMs have been created / decommissioned since the beginning

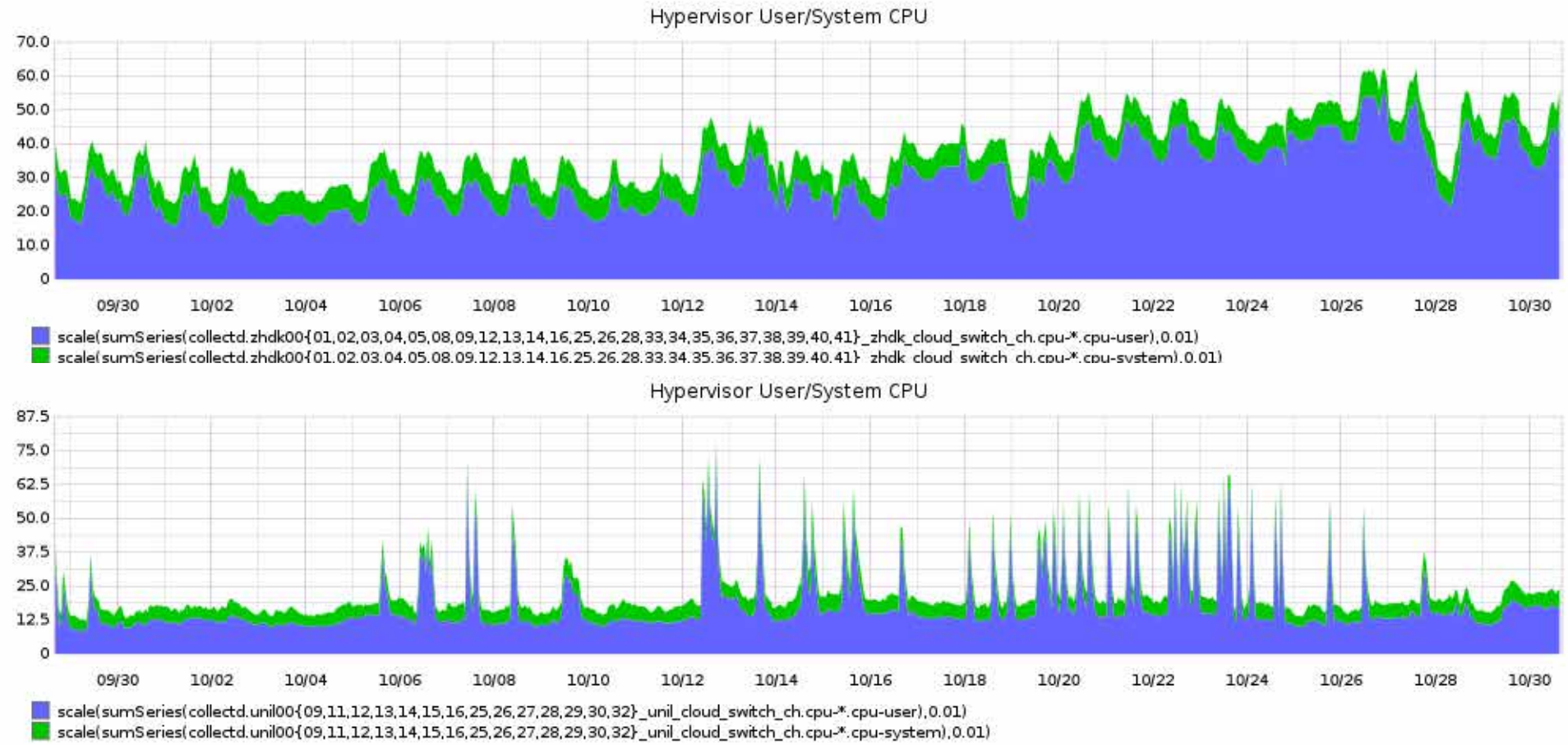
© 2015 SWITCH

Cloud Usage



© 2015 SWITCH

Cloud Usage



© 2015 SWITCH

Expansion under way

- Double the number of compute nodes (and increase the number of CPUs by 150% - Moore's Law at Work)
  - 2000 CPU cores
- Increase the storage capacity by 50%
  - 2.25 PB raw storage
- Online end of 2015

© 2015 SWITCH

Building

https://www.flickr.com/photos/criminalintent/5101528210

Hardware is easy

- Commercial Off the Shelf

- “Low Cost, Medium Performance” (but lots of it)

© 2015 SWITCH

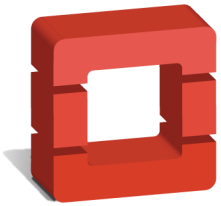
Building SWITCHengines

- 2 Geographic Locations
  - ZHdK
  - UNIL
- 32 Storage Nodes
  - 384 4TB SATA Disks (and a bunch of SSDs for Journals)
  - 1.5 PB RAW Storage
- 41 Compute Nodes
  - 1240 Cores
  - 9.2 TB RAM

Software

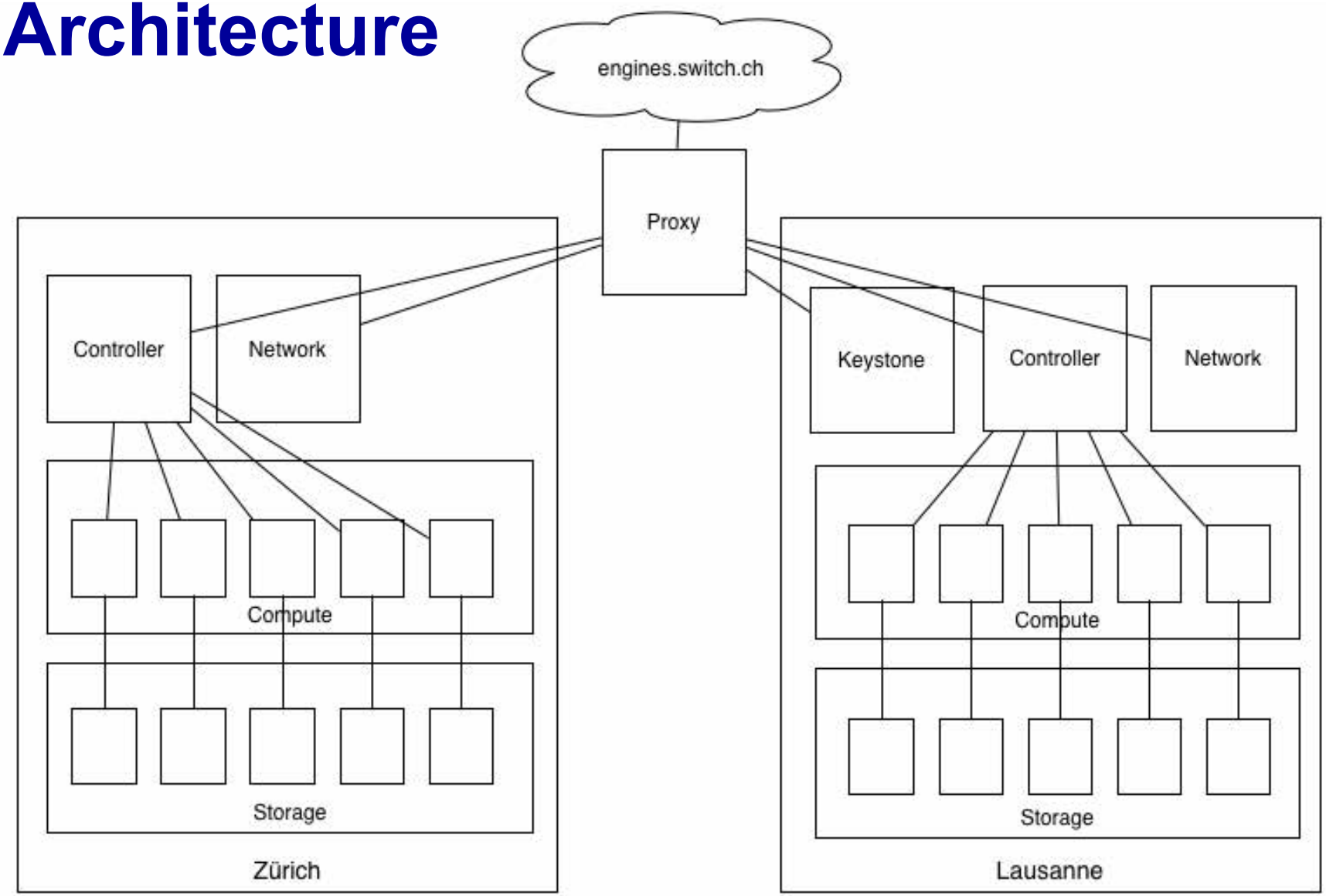
© 2015 SWITCH

But wait, there's more



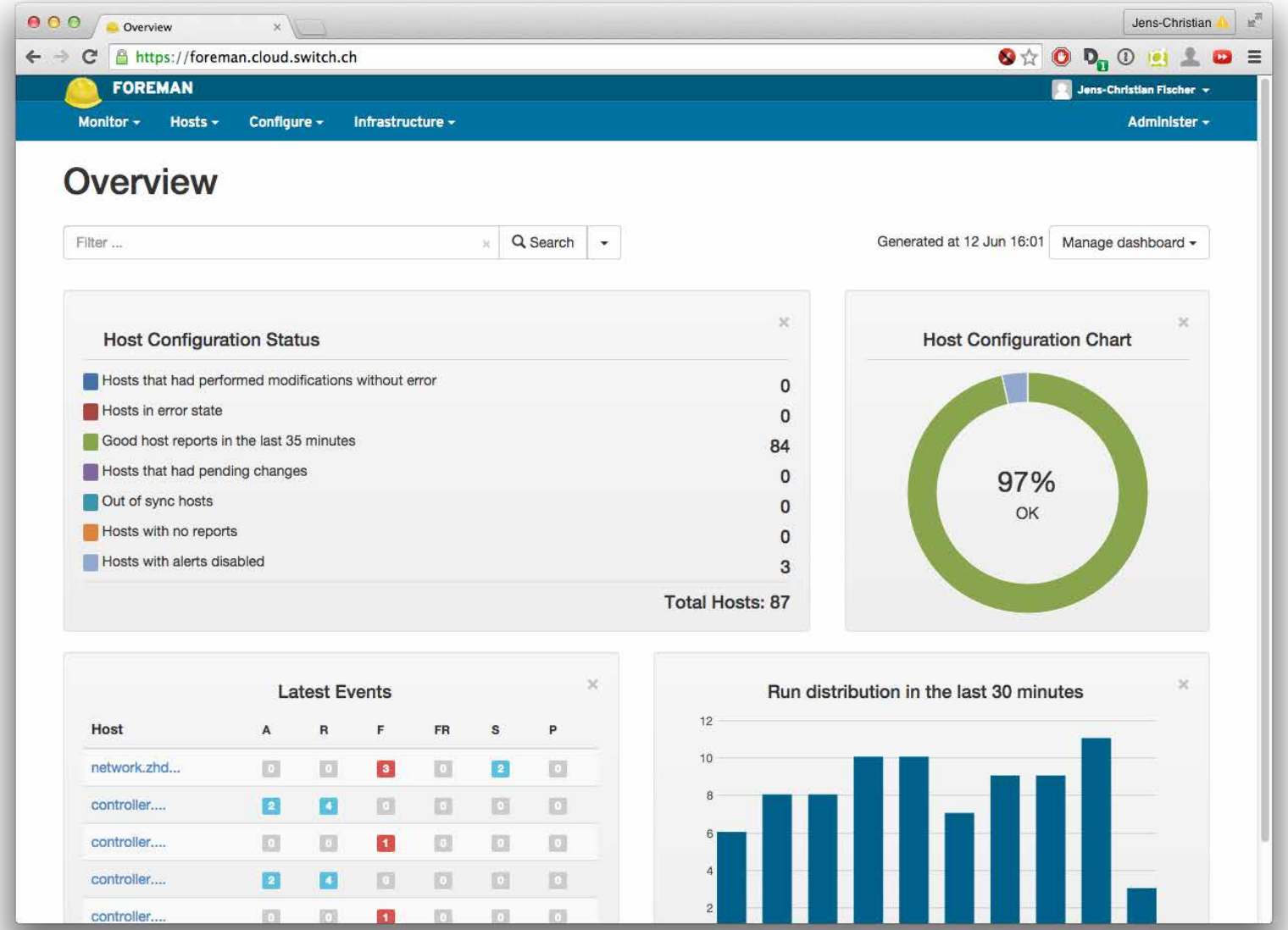
© 2015 SWITCH

Architecture



© 2015 SWITCH

Infrastructure - Foreman



© 2015 SWITCH

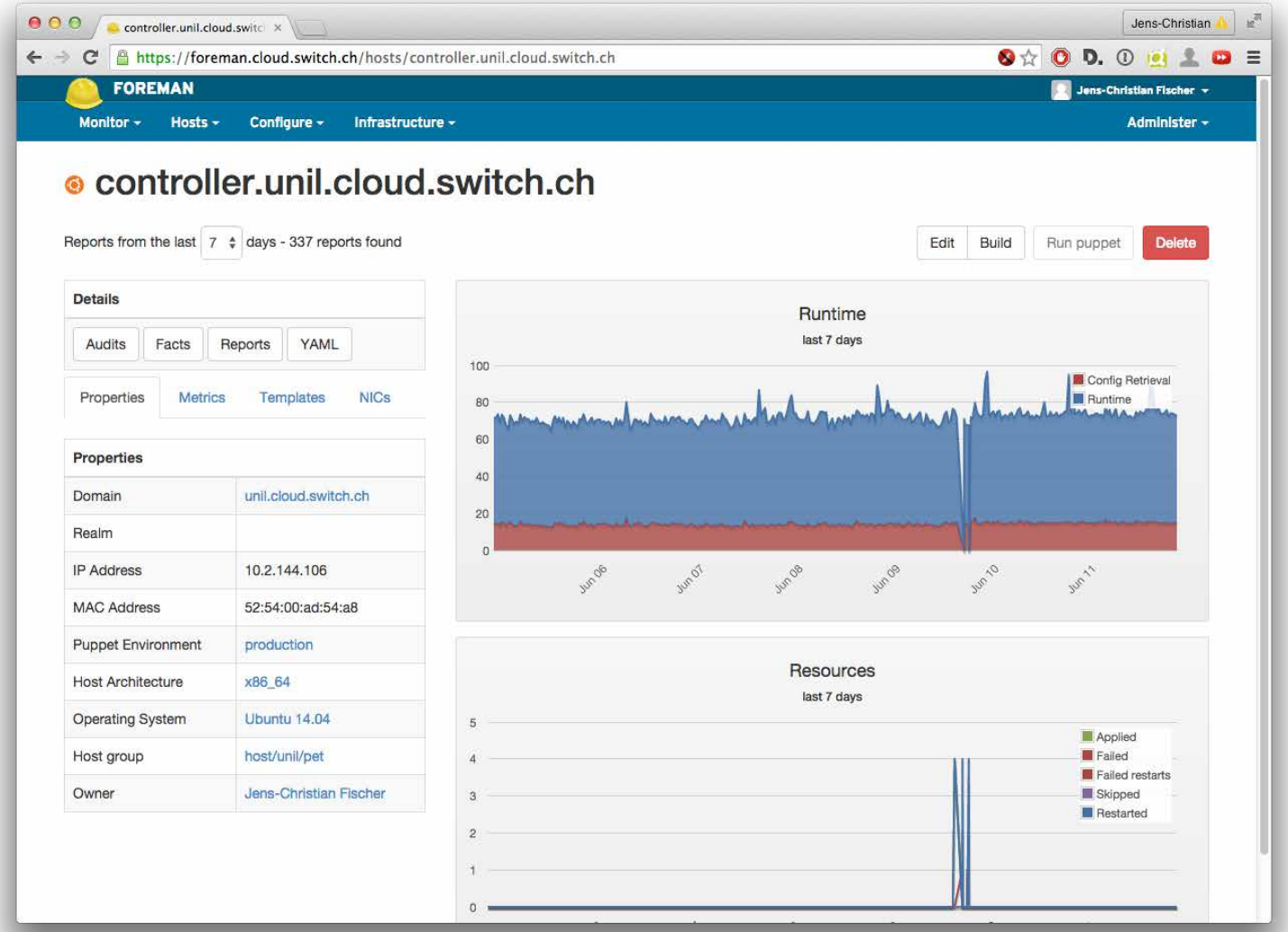
Ceph

- When 12 * 4 TB aren’t enough, go distributed storage systems

```
$ ceph -s
    cluster yyyyyyyyy-bdae-464b-a864-xxxxxxxxxxx
     health HEALTH_OK
     monmap e2: 3 mons at
{zhdk0009=[2001:xxx::1009]:6789/0,zhdk0013=[2001:yyy::1013]:6
789/0,zhdk0025=[2001:yyy::1025]:6789/0}, election epoch 612,
quorum 0,1,2 zhdk0009,zhdk0013,zhdk0025
     osdmap e44762: 125 osds: 125 up, 125 in
      pgmap v20730639: 3336 pgs, 17 pools, 85088 GB data,
23864 kobjects
            279 TB used, 174 TB / 454 TB avail
                3317 active+clean
                  19 active+clean+scrubbing+deep
  client io 10031 kB/s rd, 14072 kB/s wr, 1472 op/s
```

© 2015 SWITCH

Foreman – Server Definition



© 2015 SWITCH

Edit controller.unil.cloud.switch.ch

Unmanage host

Host | Puppet Classes | Interfaces | Operating System | Parameters | Additional Information

Included Classes

- aadm
- ceph
- nagios::openstack_cinder
- nagios::openstack_glance
- nagios::openstack_heat
- nagios::openstack_nova_head
- scale::controller
- switch::controller_mount_rbd
- nagios
- nagios::puppet_agent
- switch

Available Classes

Filter classes

- aadm
- + apache
- + apt
- + ceilometer
- + ceph
- + cinder
- + concat
- + epel
- + erlang
- + firewall
- + glance
- + heat
- + horizon
- + keystone
- + ldap
- + memcached
- + ntp
- + openstack
- + passenger
- + postgresql
- + proxy
- + qpid
- + quobyte
- + rabbitmq
- + radvd
- + rsync
- + ruby
- + rvm
- + scale
- + ssh
- + stdlib
- + swift

© 2015 SWITCH

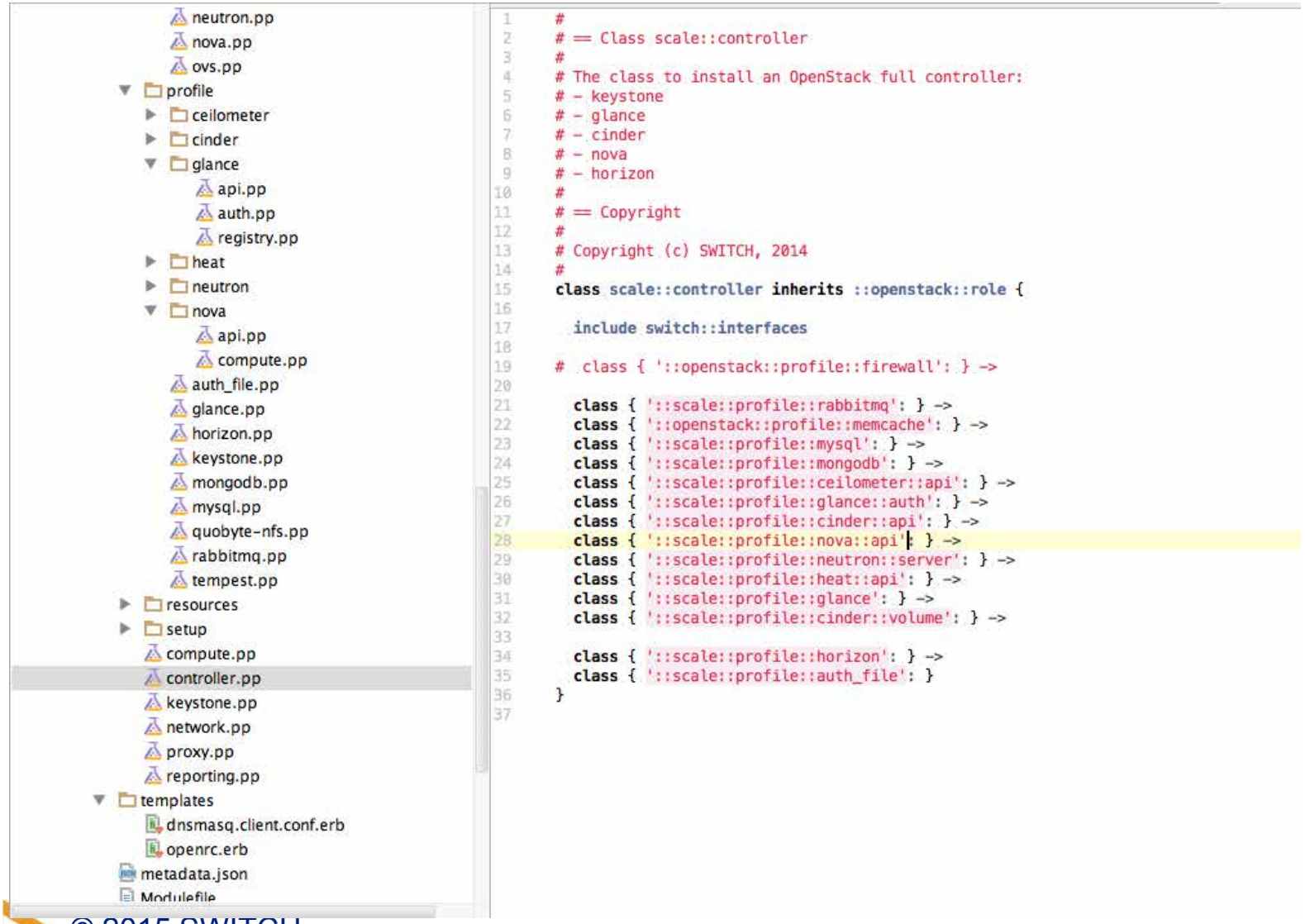
Puppet – Configuration Management

- http://puppetlabs.com

- https://github.com/puppetlabs/puppetlabs-openstack
- Modules for every component of OpenStack
- Generic Glue Modules that build the actual Cluster

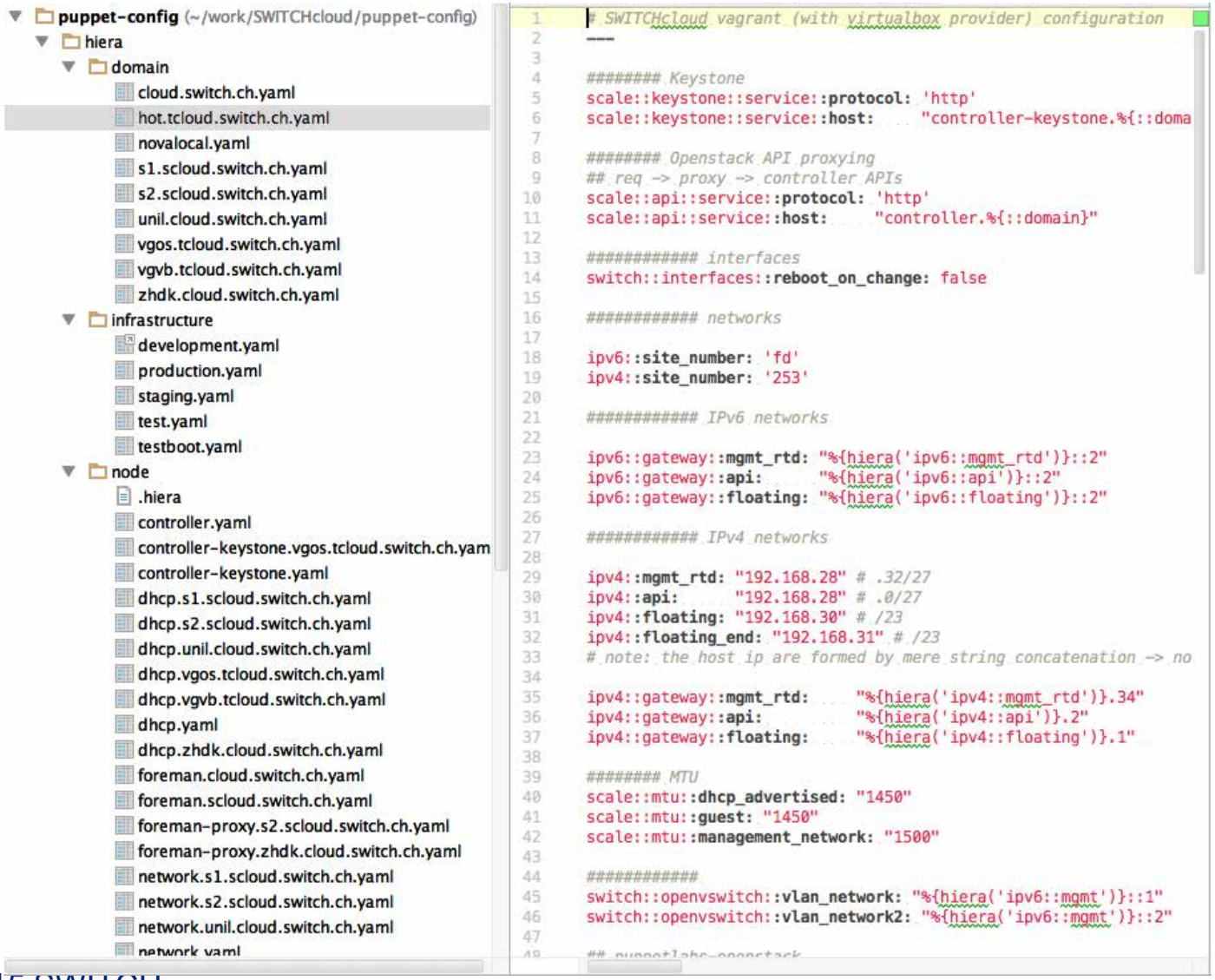
© 2015 SWITCH

Split between Code – Puppet



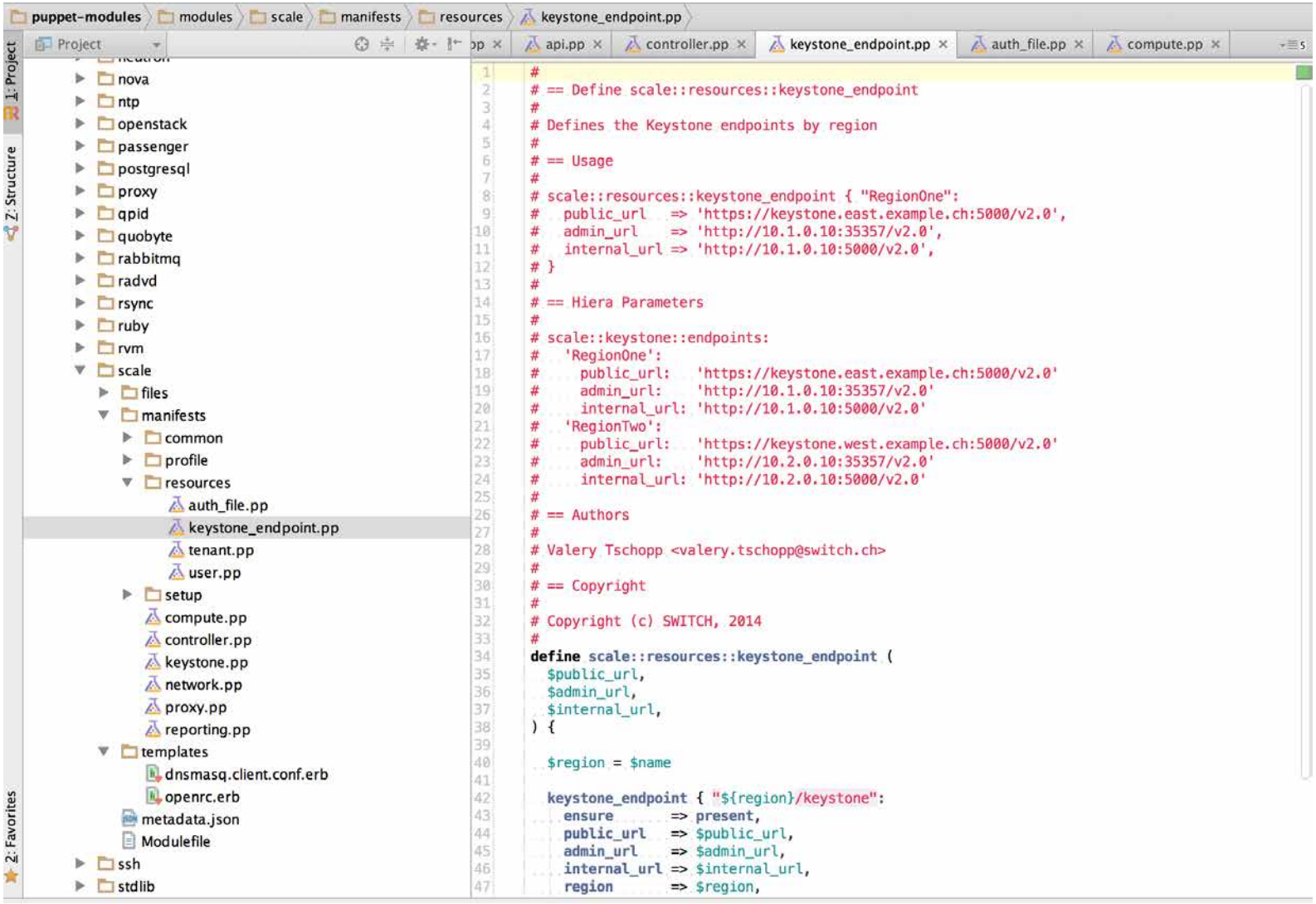
© 2015 SWITCH

Configuration - Hiera



© 2015 SWITCH

Puppet Glue



© 2015 SWITCH

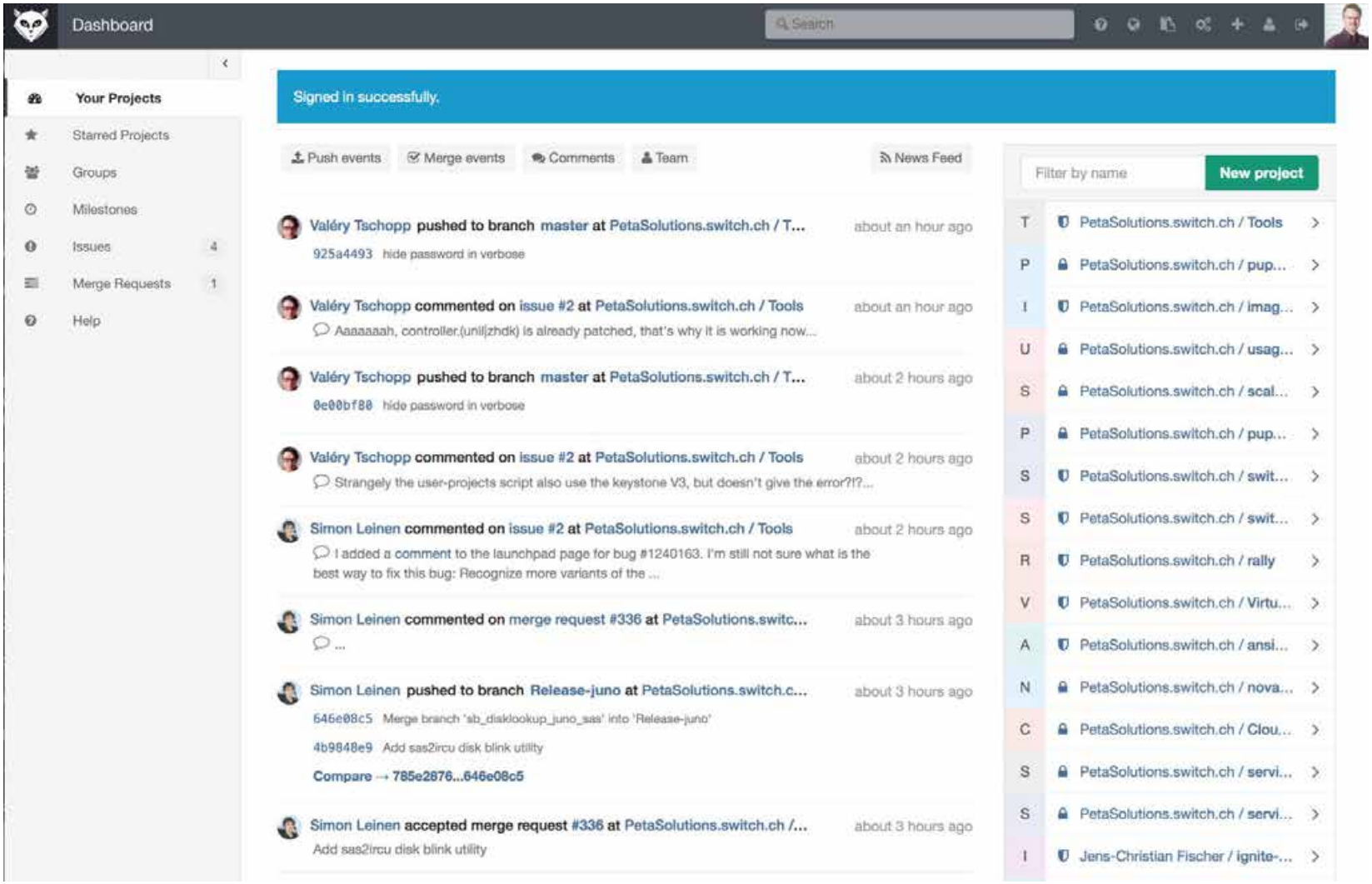
Orchestrating the Orchestrator

- Ansible - http://www.ansible.com/home

```
ansible-playbook disable_puppet.yml -i cloud --sudo --ask-
sudo-pass --extra-vars "hosts=production”

ansible-playbook prepare_release.yml -i cloud --sudo --ask-
sudo-pass --extra-vars "modules=v1.3 config=v1.3.1”

ansible-playbook test_release.yml -i cloud --sudo --ask-sudo-
pass --extra-vars "hosts=canary-vm-unil”

ansible-playbook do_release.yml -i cloud --sudo --ask-sudo-
pass --extra-vars "hosts=zhdk-all"
```

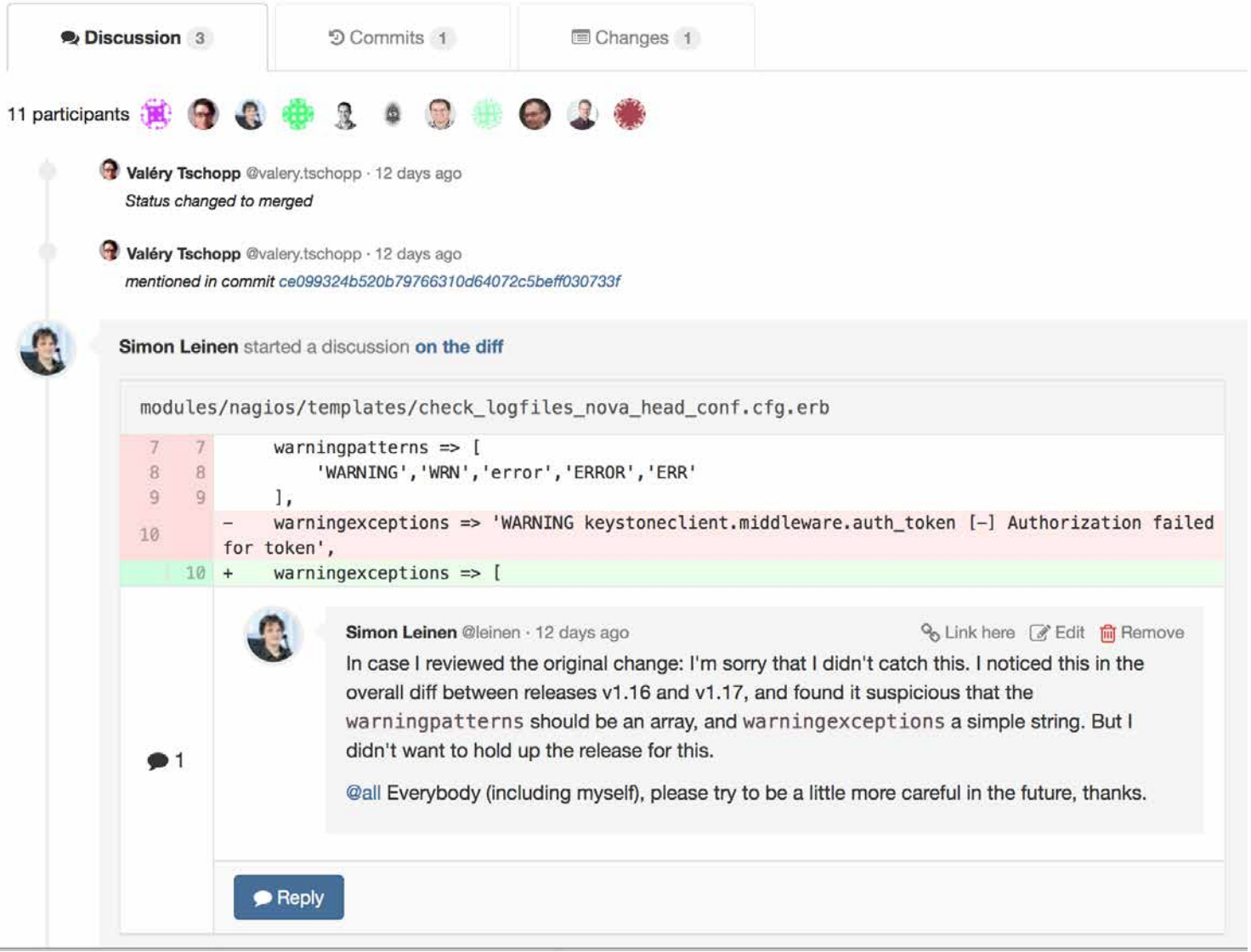
© 2015 SWITCH

Version? What version?



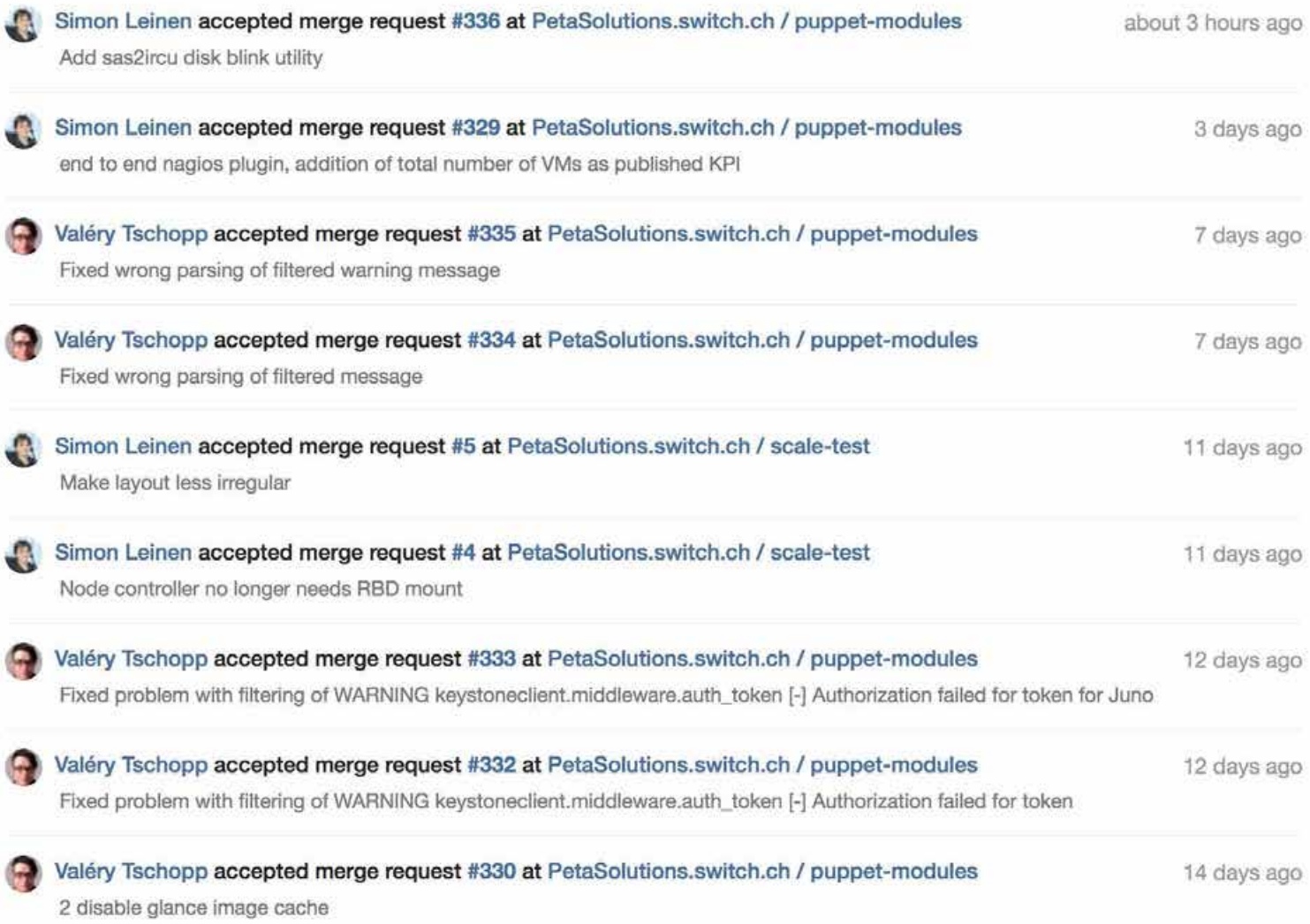
© 2015 SWITCH

Discussions



© 2015 SWITCH

Merge Requests



© 2015 SWITCH

Monitoring

monitor@switch.ch 15. Juni 2015 13:34
To: Jens-Christian Fischer Hide Details
Service: zhdk0024.zhdk.cloud.switch.ch/L_check_logfiles: WARNING, Type: PROBLEM

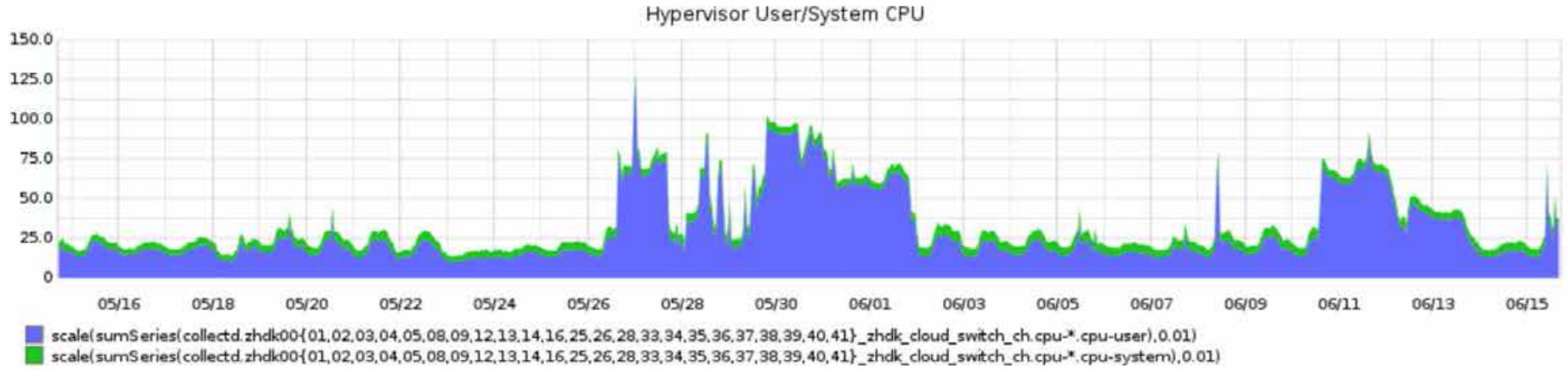
Host: zhdk0024.zhdk.cloud.switch.ch
Alias: zhdk0024.zhdk.cloud.switch.ch
Address: 2001:620:5ca1:100::1024
Service: L_check_logfiles
State: WARNING
Notificationtype: PROBLEM
Recipients:
Command: check_mk-mrpe
Output: WARN - WARNING - (2 warnings in check_logfiles.protocol-2015-06-15-13-34-26) - File=/var/log/ceph/ceph-osd.119.log Message=2015-06-15 13:34:15.900352 7fd45b7fe700 0 -- [2001:620:5ca1:100::1024]:6835/185927 submit_message osd_op_reply(923361 rbd_data.bd32471a66d894.00000000000001e2 [stat,set-alloc-hint object_size 8388608 write_size 8388608,write 15360004096] v4476210069454 uv10069454 ondisk = 0) v6 remote, [2001:620:5ca1:100::1009]:0/1089960, failed lossy con, dropping message 0x5810d00 ...
Perfdata: osd_logs83_lines=0;;;; osd_logs83_warnings=0;;;; osd_logs83_criticals=0;;;; osd_logs83_unknowns=0;;;; osd_logs41_lines=1;;;; osd_logs41_warnings=0;;;; osd_logs41_criticals=0;;;; osd_logs41_unknowns=0;;;; osd_logs19_lines=2;;;; osd_logs19_warnings=0;;;; osd_logs19_criticals=0;;;; osd_logs19_unknowns=0;;;; osd_logs70_lines=1;;;; osd_logs70_warnings=0;;;; osd_logs70_criticals=0;;;; osd_logs70_unknowns=0;;;; osd_logs75_lines=0;;;; osd_logs75_warnings=0;;;; osd_logs75_criticals=0;;;; osd_logs75_unknowns=0;;;; osd_logs38_lines=1;;;; osd_logs38_warnings=0;;;; osd_logs38_criticals=0;;;; osd_logs38_unknowns=0;;;; osd_logs101_lines=0;;;; osd_logs101_warnings=0;;;; osd_logs101_criticals=0;;;; osd_logs101_unknowns=0;;;; osd_logs119_lines=2;;;; osd_logs119_warnings=2;;;; osd_logs119_criticals=0;;;; osd_logs119_unknowns=0;;;; syslog_lines=1;;;; syslog_warnings=0;;;; syslog_criticals=0;;;; syslog_unknowns=0;;;; kern.log_lines=0;;;; kern.log_warnings=0;;;; kern.log_criticals=0;;;; kern.log_unknowns=0;;;; [sudo]

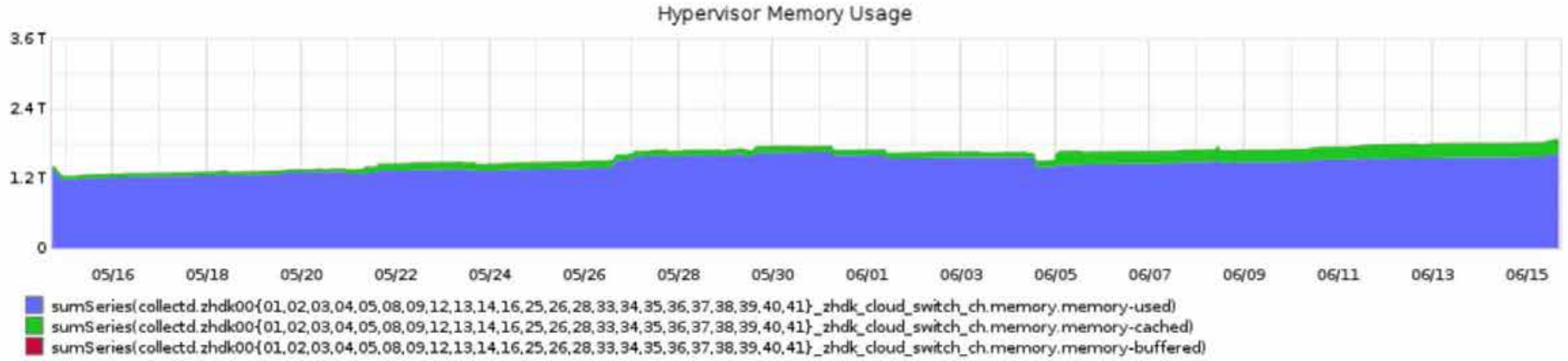
© 2015 SWITCH

Usage

SWITCHengines Load

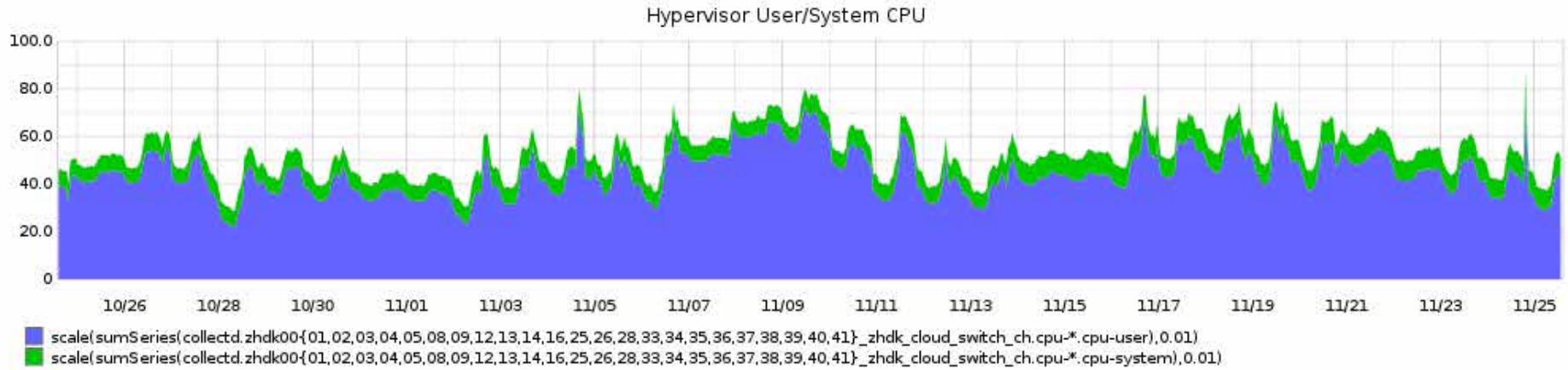
ZH Region

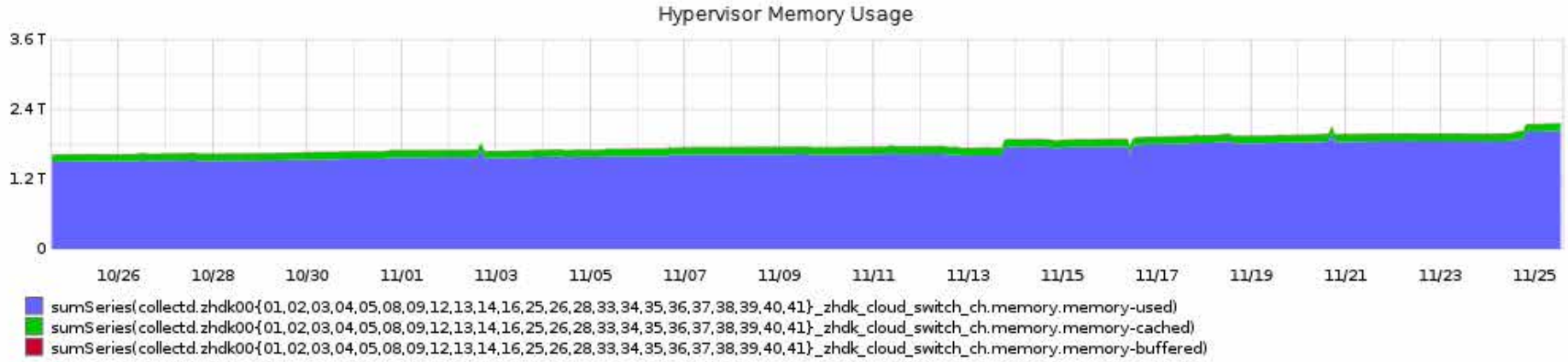




Usage (5 months later)

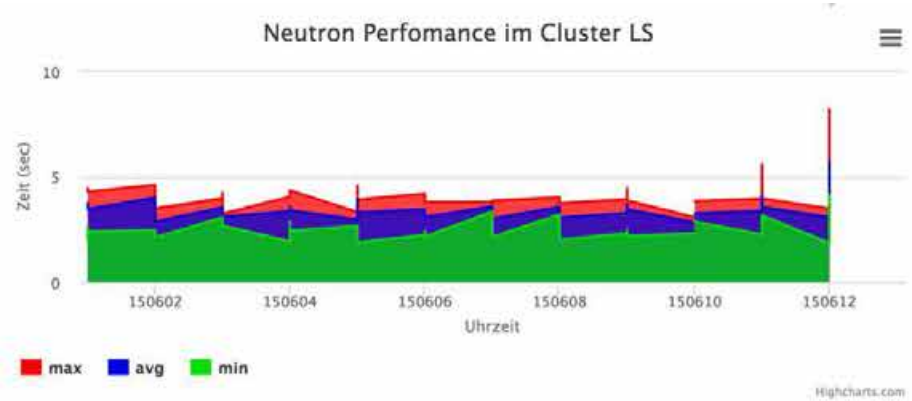
ZH Region

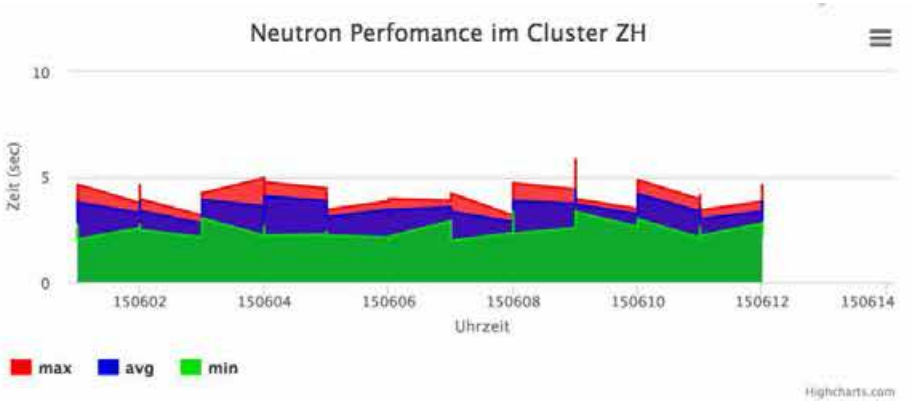


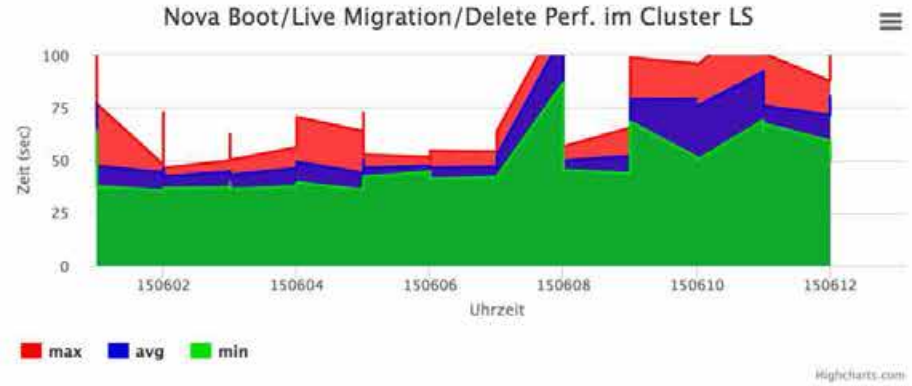


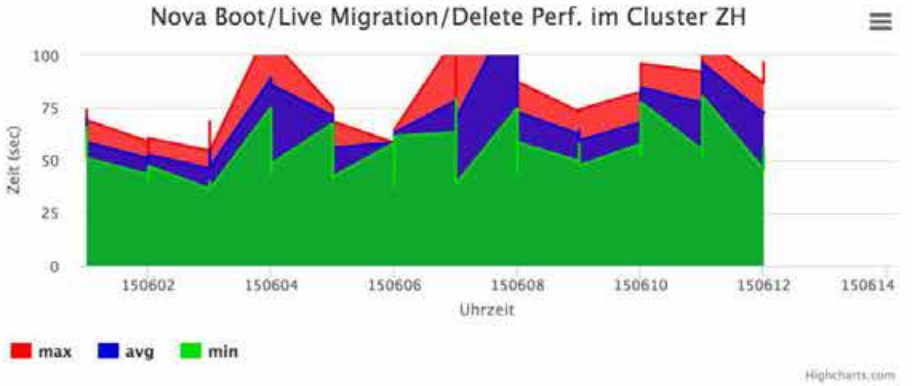
© 2015 SWITCH

Performance







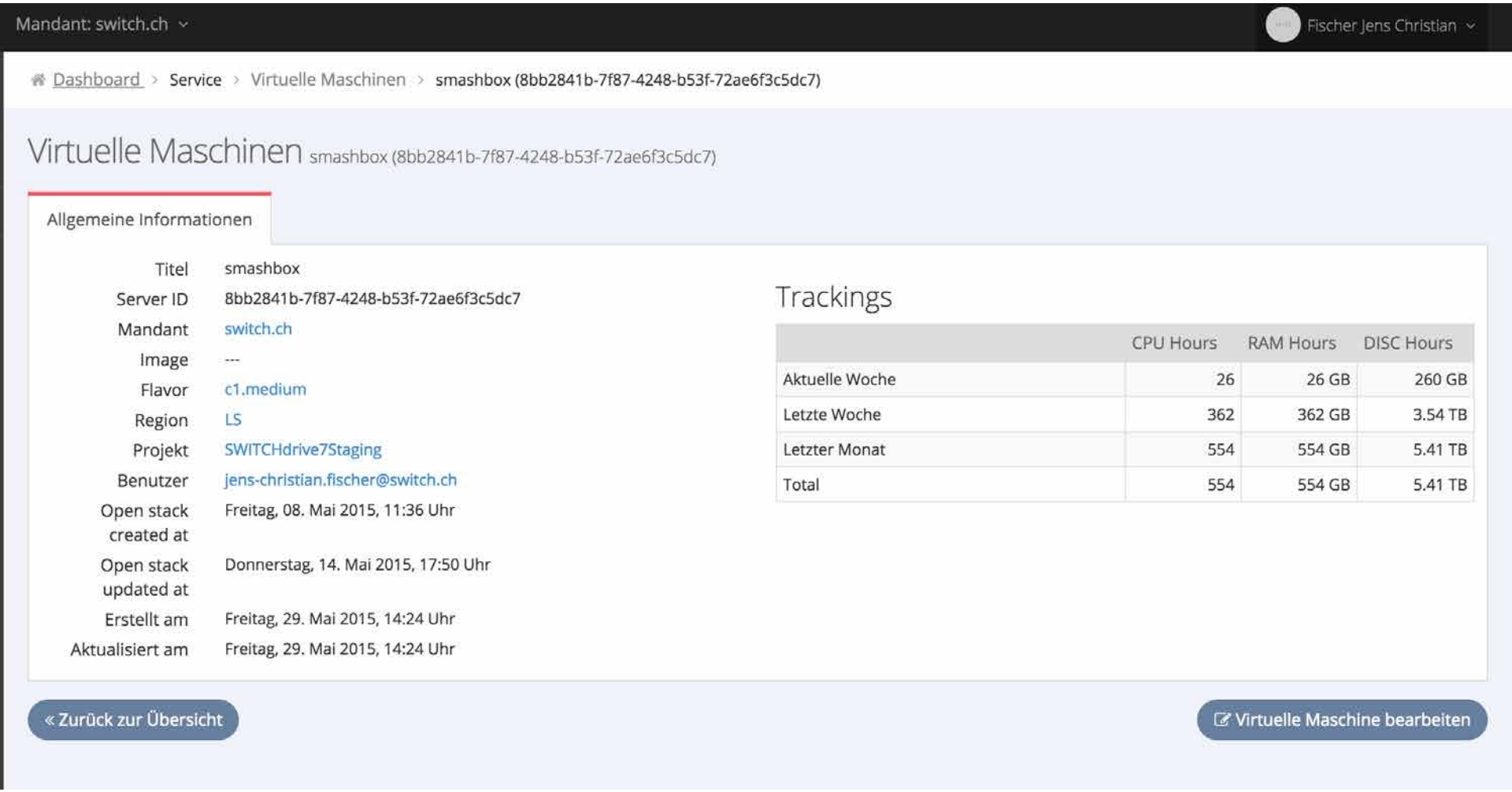


© 2015 SWITCH

Monitoring

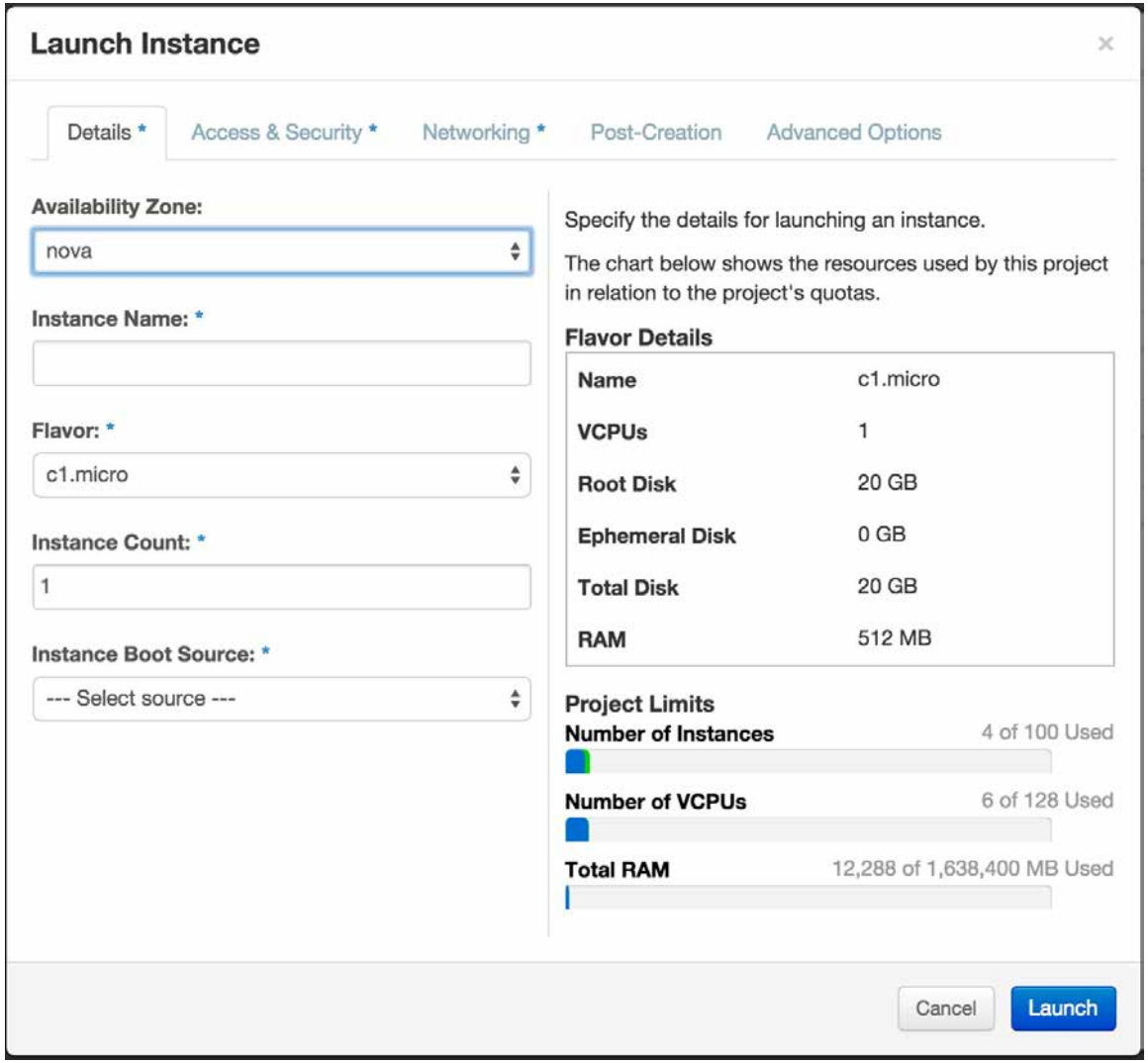
© 2015 SWITCH

Reporting



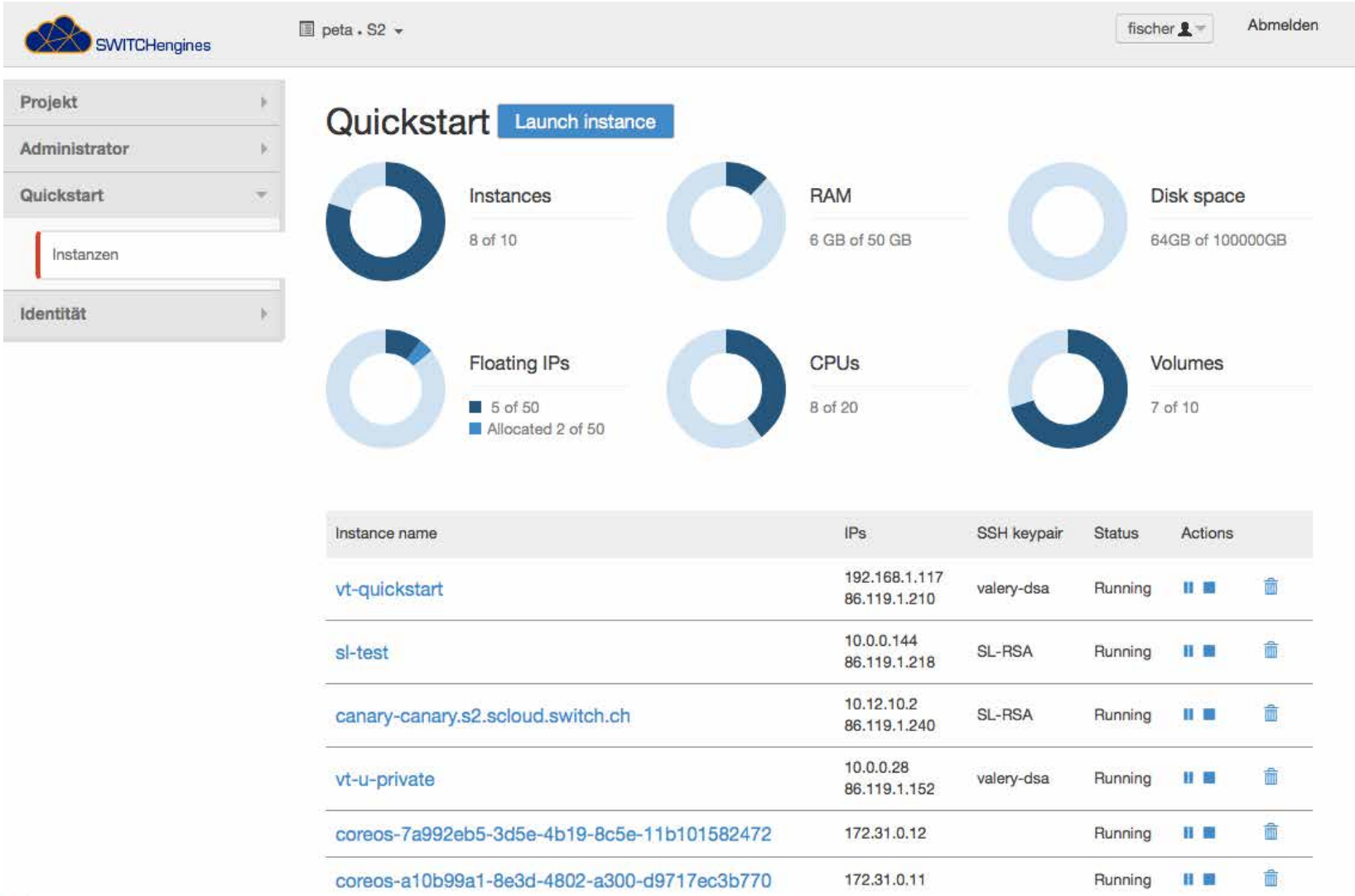
© 2015 SWITCH

User Interface



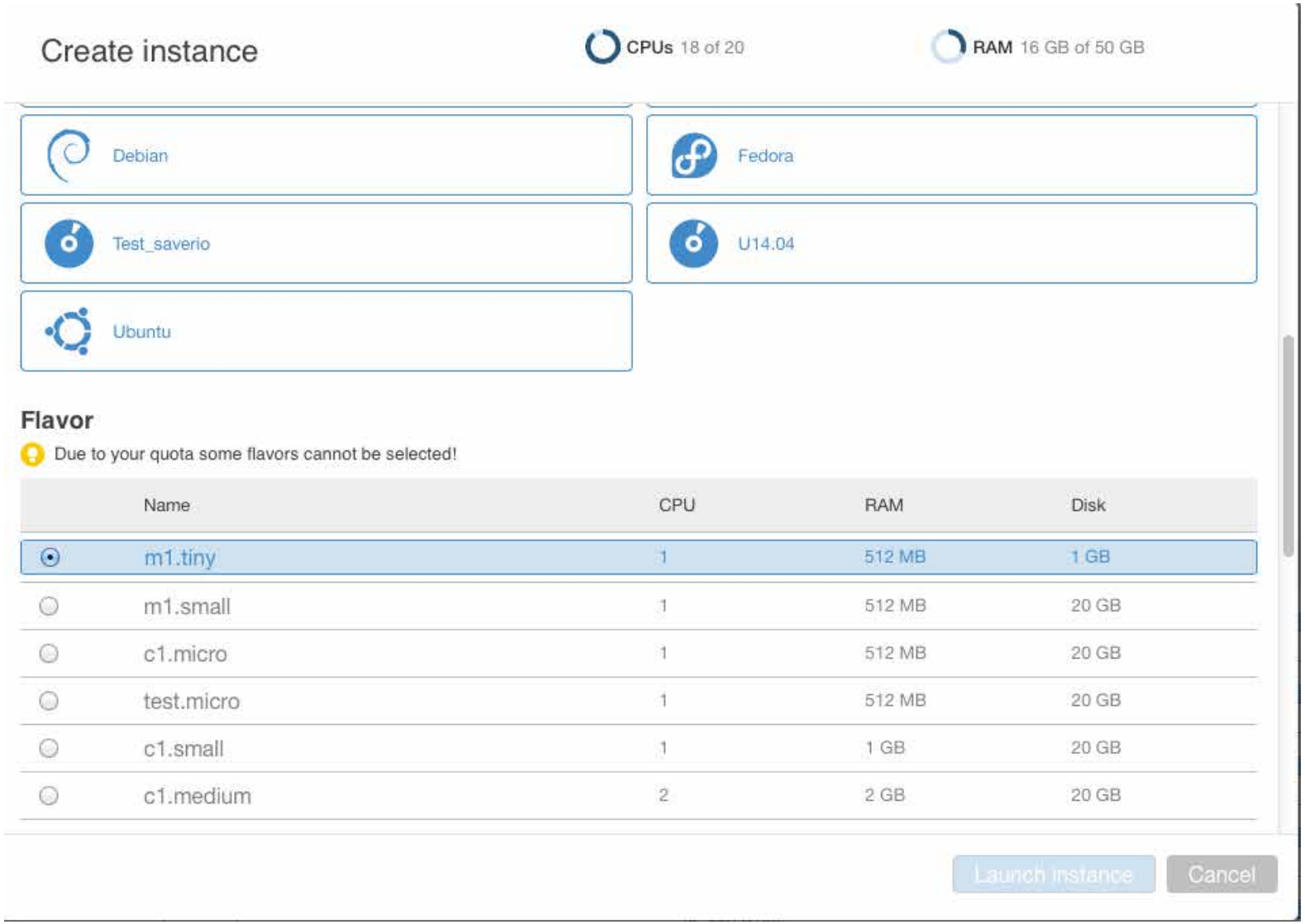
© 2015 SWITCH

New UI



© 2015 SWITCH

New UI



© 2015 SWITCH

OpenStack ++

- There is more to cloud than just cloud
- You need the tools
- You need the team

- And when you start to actually run the cloud, things get fun

© 2015 SWITCH

When you build it, they will come

- Network simulations
- Digital Humanities
- Aerial Photography of Lake shores (Leman / Baikal)
- CoreOS / Docker Container orchestration
- Classroom VMs
- ATLAS computation nodes
- Lotus Domino Servers
- Web servers from University IT departments
- TOR exit nodes
- …

© 2015 SWITCH

Joys of running a cloud

- User chooses “Public” network, her VM doesn’t launch
- Your network node takes down all VMs network connections
- A faulty disk causes major performance problems in a distributed, no SPOF storage system
- Your Performance test framework doesn’t delete default security rules => IPTables with 70’000 entries
- Your keystone service catalog entries get duplicated, all API access breaks because Keystone tokens are hashed differently
- …

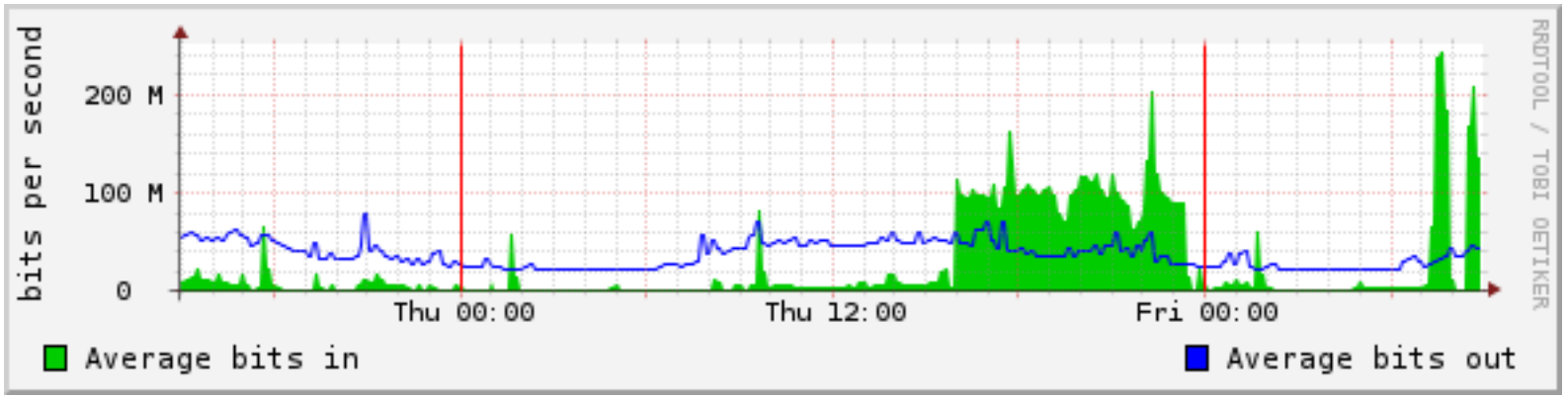
© 2015 SWITCH

Joys of running a cloud

- Deleting a Cinder volume takes down your whole Storage API services (bug in the Cinder/Ceph integration)
- Users enable password login to their VM (because: Easier)
- Your organisation changes the way the Firewall rules are distributed to the routers
- You need to manage access for 650 users and 50 groups/tenants
- Your users exhaust the number of floating (public) IP addresses you have assigned to the Cloud network
- …

© 2015 SWITCH

Watch the network traffic

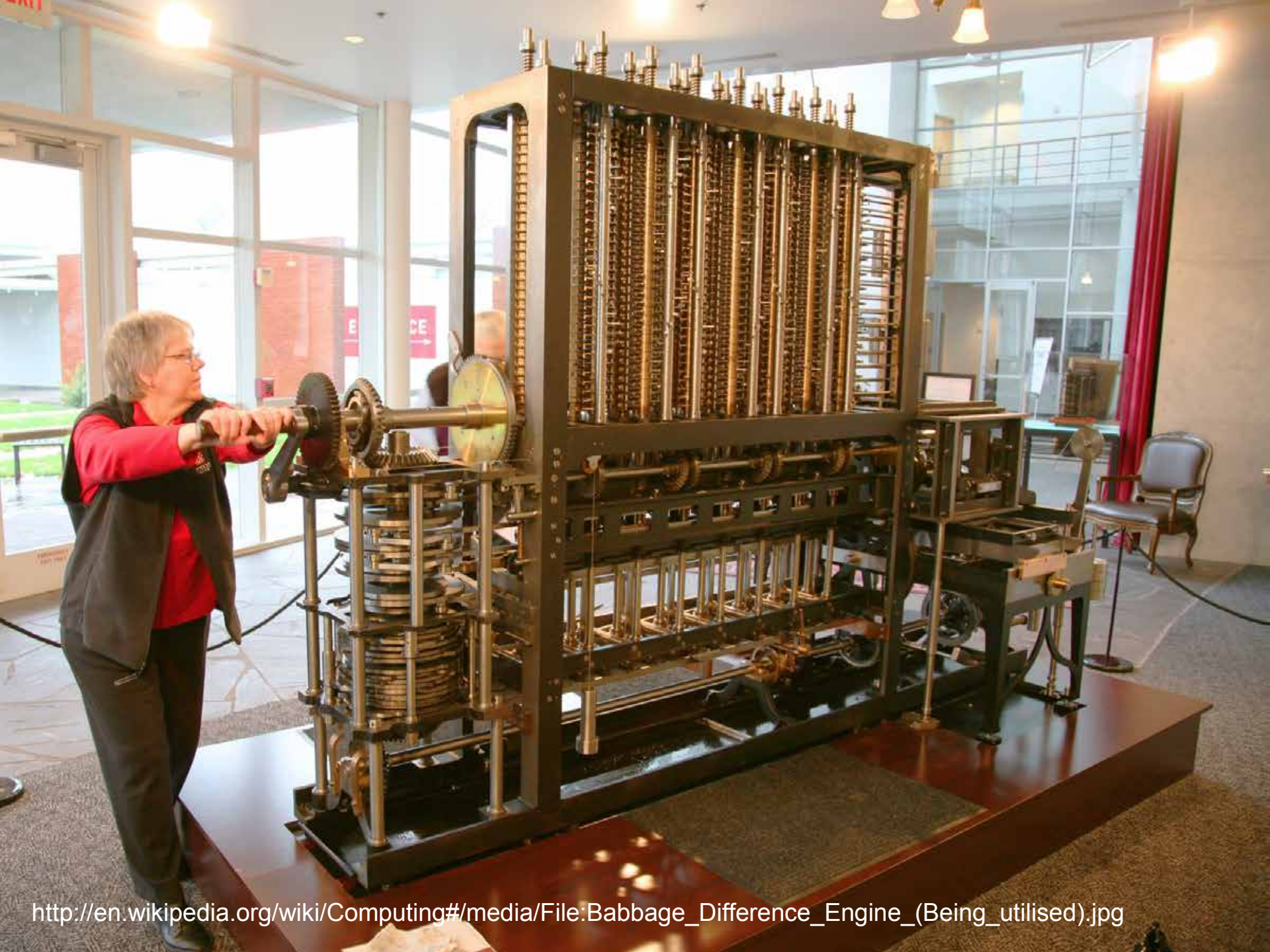


© 2015 SWITCH

Part of a DDOS attack

- Users VM was hacked
- (see above: User enables password login)

© 2015 SWITCH



http://en.wikipedia.org/wiki/Computing#/media/File:Babbage_Difference_Engine_(Being_utilised).jpg