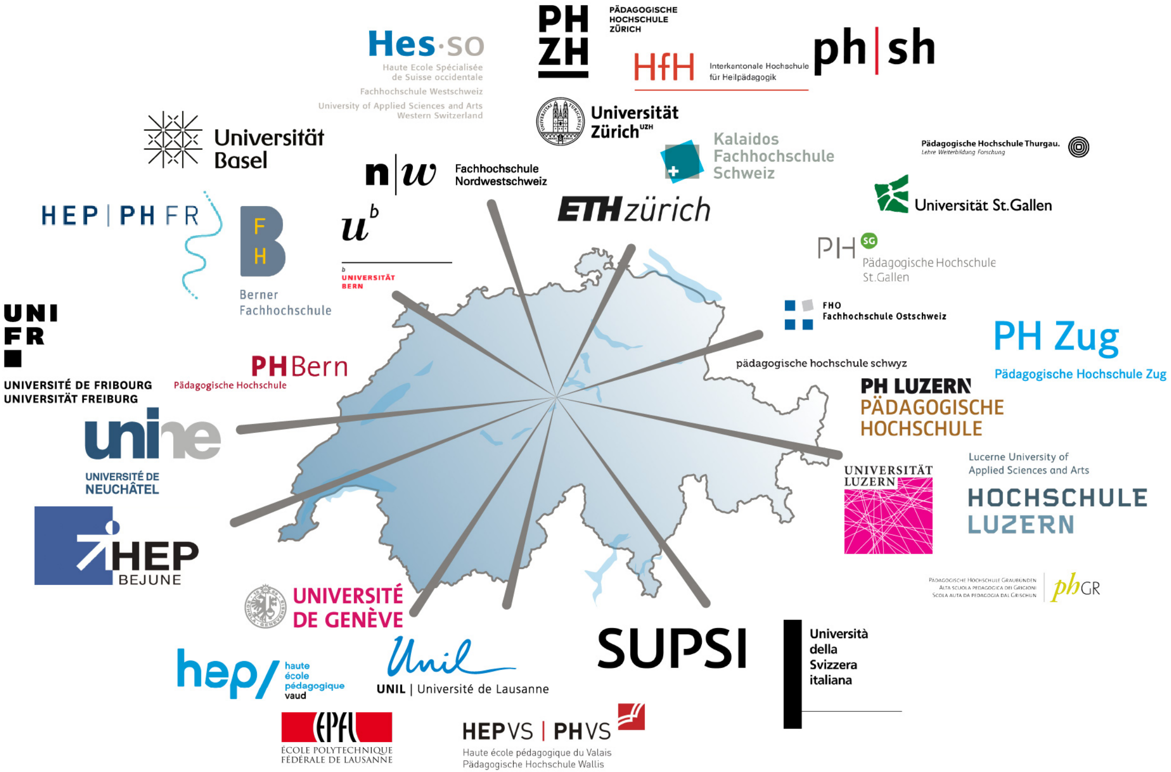
Cloud & IPv6 – schön schaurig

SWITCH

Jens-Christian Fischer
jens-christian.fischer@switch.ch
@jcfischer / @switchengines

IPv6 Forum, 2019-07-01

Academic community Switzerland


Hes·so
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts Western Switzerland
PH ZH
PÄDAGOGISCHE HOCHSCHULE ZÜRICH
HfH
Interkantonale Hochschule für Heilpädagogik
ph|sh
Universität Zürich UZH
Kalaidos Fachhochschule Schweiz
Pädagogische Hochschule Thurgau.
Lehre Weiterbildung Forschung
Universität Basel
n|w Fachhochschule Nordwestschweiz
ETH zürich
Universität St.Gallen
HEP|PH FR
F H
Berner Fachhochschule
u b
UNIVERSITÄT BERN
PH SG
Pädagogische Hochschule St.Gallen
FHO Fachhochschule Ostschweiz
PH Zug
Pädagogische Hochschule Zug
pädagogische hochschule schwyz
UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG
PH Bern
Pädagogische Hochschule
PH LUZERN
PÄDAGOGISCHE HOCHSCHULE
unine
UNIVERSITÉ DE NEUCHÂTEL
UNIVERSITÄT LUZERN
Lucerne University of Applied Sciences and Arts
HOCHSCHULE LUZERN
HEP BEJUNE
PADAGOGISCHE HOCHSCHULE GRAUBUNDEN
ALTA SCUOLA PEDAGOGICA DEI GRIGIONI
SCOLA AUTA DA PEDAGOGIA DAL GRISCHUN
phGR
UNIVERSITÉ DE GENÈVE
SUPSI
Università della Svizzera italiana
hep/ haute école pédagogique vaud
Unil
UNIL | Université de Lausanne
EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
HEPVS | PHVS
Haute école pédagogique du Valais
Pädagogische Hochschule Wallis

SWITCHengines

Customer tailored computing and storage performance for universities, research and teaching – further developed in the SCALE-UP project mandated by swissuniversities.

Customers

- Universities
- Research institutions
- eLearning Center
- University hospitals
- Spin-Offs

Services

- SWITCHengines (IaaS)
- Virtual Private Cloud (VPC)
- SCALE-UP (academic project)

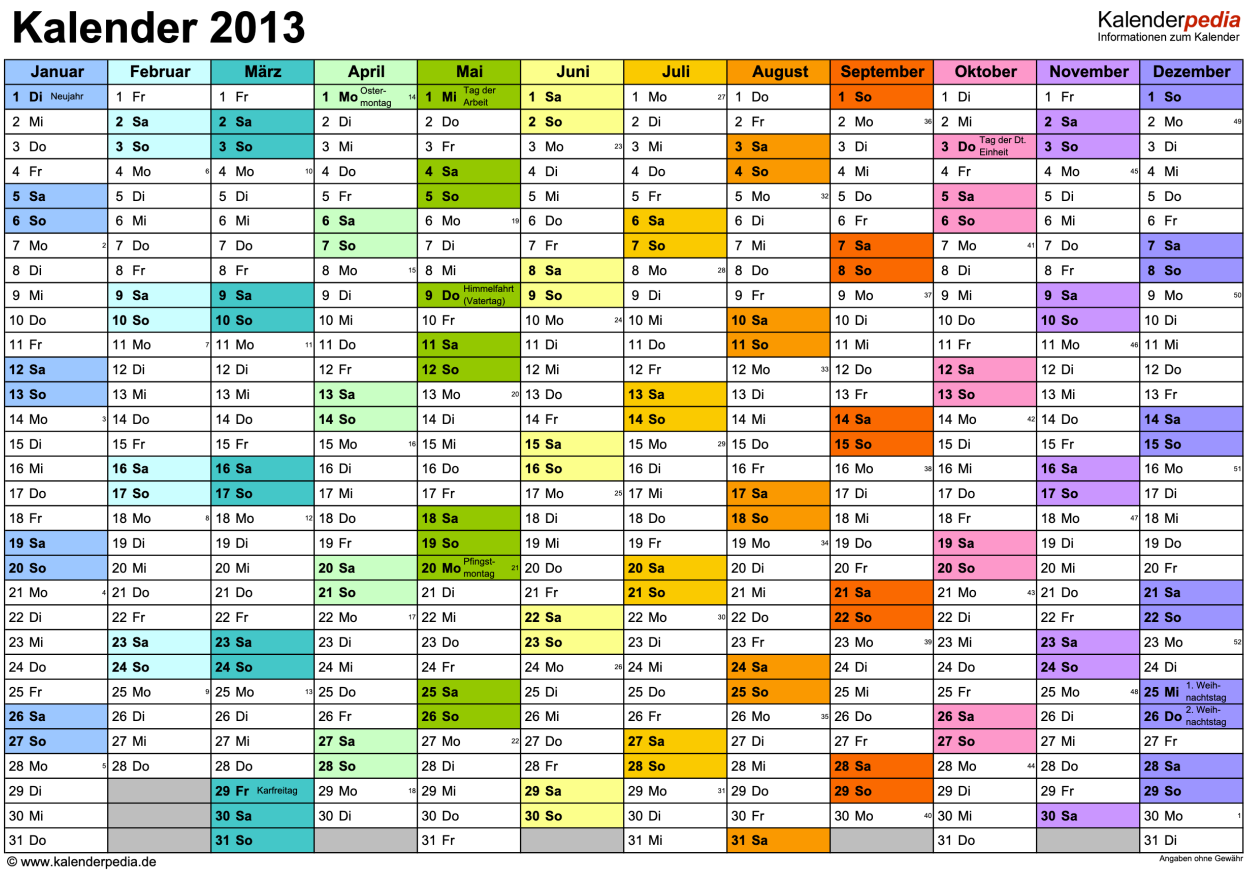
Your benefits

- Your data in Switzerland
- Integrated network and security
- Support for academic use cases
- Simple administration and billing
- Created together with you

Disclaimer

https://www.pexels.com/photo/ethernet-internet-network-router-2259948/

We are building a cloud



Kalender 2013

Kalender*pedia*
Informationen zum Kalender

Januar	Februar	März	April	Mai	Juni	Juli	August	September	Oktober	November	Dezember
1 Di Neujahr	1 Fr	1 Fr	**1 Mo** Ostermontag 14	**1 Mi** Tag der Arbeit	**1 Sa**	1 Mo 27	1 Do	**1 So**	1 Di	1 Fr	**1 So**
2 Mi	**2 Sa**	**2 Sa**	2 Di	2 Do	**2 So**	2 Di	2 Fr	2 Mo 36	2 Mi	**2 Sa**	2 Mo 49
3 Do	**3 So**	**3 So**	3 Mi	3 Fr	3 Mo 23	3 Mi	**3 Sa**	3 Di	**3 Do** Tag der Dt. Einheit	**3 So**	3 Di
4 Fr	4 Mo 6	4 Mo 10	4 Do	**4 Sa**	4 Di	4 Do	**4 So**	4 Mi	4 Fr	4 Mo 45	4 Mi
5 Sa	5 Di	5 Di	5 Fr	**5 So**	5 Mi	5 Fr	5 Mo 32	5 Do	**5 Sa**	5 Di	5 Do
6 So	6 Mi	6 Mi	**6 Sa**	6 Mo 19	6 Do	**6 Sa**	6 Di	6 Fr	**6 So**	6 Mi	6 Fr
7 Mo 2	7 Do	7 Do	**7 So**	7 Di	7 Fr	**7 So**	7 Mi	**7 Sa**	7 Mo 41	7 Do	**7 Sa**
8 Di	8 Fr	8 Fr	8 Mo 15	8 Mi	**8 Sa**	8 Mo 28	8 Do	**8 So**	8 Di	8 Fr	**8 So**
9 Mi	**9 Sa**	**9 Sa**	9 Di	**9 Do** Himmelfahrt (Vatertag)	**9 So**	9 Di	9 Fr	9 Mo 37	9 Mi	**9 Sa**	9 Mo 50
10 Do	**10 So**	**10 So**	10 Mi	10 Fr	10 Mo 24	10 Mi	**10 Sa**	10 Di	10 Do	**10 So**	10 Di
11 Fr	11 Mo 7	11 Mo 11	11 Do	**11 Sa**	11 Di	11 Do	**11 So**	11 Mi	11 Fr	11 Mo 46	11 Mi
12 Sa	12 Di	12 Di	12 Fr	**12 So**	12 Mi	12 Fr	12 Mo 33	12 Do	**12 Sa**	12 Di	12 Do
13 So	13 Mi	13 Mi	**13 Sa**	13 Mo 20	13 Do	**13 Sa**	13 Di	13 Fr	**13 So**	13 Mi	13 Fr
14 Mo 3	14 Do	14 Do	**14 So**	14 Di	14 Fr	**14 So**	14 Mi	**14 Sa**	14 Mo 42	14 Do	**14 Sa**
15 Di	15 Fr	15 Fr	15 Mo 16	15 Mi	**15 Sa**	15 Mo 29	15 Do	**15 So**	15 Di	15 Fr	**15 So**
16 Mi	**16 Sa**	**16 Sa**	16 Di	16 Do	**16 So**	16 Di	16 Fr	16 Mo 38	16 Mi	**16 Sa**	16 Mo 51
17 Do	**17 So**	**17 So**	17 Mi	17 Fr	17 Mo 25	17 Mi	**17 Sa**	17 Di	17 Do	**17 So**	17 Di
18 Fr	18 Mo 8	18 Mo 12	18 Do	**18 Sa**	18 Di	18 Do	**18 So**	18 Mi	18 Fr	18 Mo 47	18 Mi
19 Sa	19 Di	19 Di	19 Fr	**19 So**	19 Mi	19 Fr	19 Mo 34	19 Do	**19 Sa**	19 Di	19 Do
20 So	20 Mi	20 Mi	**20 Sa**	**20 Mo** Pfingstmontag 21	20 Do	**20 Sa**	20 Di	20 Fr	**20 So**	20 Mi	20 Fr
21 Mo 4	21 Do	21 Do	**21 So**	21 Di	21 Fr	**21 So**	21 Mi	**21 Sa**	21 Mo 43	21 Do	**21 Sa**
22 Di	22 Fr	22 Fr	22 Mo 17	22 Mi	**22 Sa**	22 Mo 30	22 Do	**22 So**	22 Di	22 Fr	**22 So**
23 Mi	**23 Sa**	**23 Sa**	23 Di	23 Do	**23 So**	23 Di	23 Fr	23 Mo 39	23 Mi	**23 Sa**	23 Mo 52
24 Do	**24 So**	**24 So**	24 Mi	24 Fr	24 Mo 26	24 Mi	**24 Sa**	24 Di	24 Do	**24 So**	24 Di
25 Fr	25 Mo 9	25 Mo 13	25 Do	**25 Sa**	25 Di	25 Do	**25 So**	25 Mi	25 Fr	25 Mo 48	**25 Mi** 1. Weihnachtstag
26 Sa	26 Di	26 Di	26 Fr	**26 So**	26 Mi	26 Fr	26 Mo 35	26 Do	**26 Sa**	26 Di	**26 Do** 2. Weihnachtstag
27 So	27 Mi	27 Mi	**27 Sa**	27 Mo 22	27 Do	**27 Sa**	27 Di	27 Fr	**27 So**	27 Mi	27 Fr
28 Mo 5	28 Do	28 Do	**28 So**	28 Di	28 Fr	**28 So**	28 Mi	**28 Sa**	28 Mo 44	28 Do	**28 Sa**
29 Di		**29 Fr** Karfreitag	29 Mo 18	29 Mi	**29 Sa**	29 Mo 31	29 Do	**29 So**	29 Di	29 Fr	**29 So**
30 Mi		**30 Sa**	30 Di	30 Do	**30 So**	30 Di	30 Fr	30 Mo 40	30 Mi	**30 Sa**	30 Mo 1
31 Do		**31 So**		31 Fr		31 Mi	**31 Sa**		31 Do		31 Di

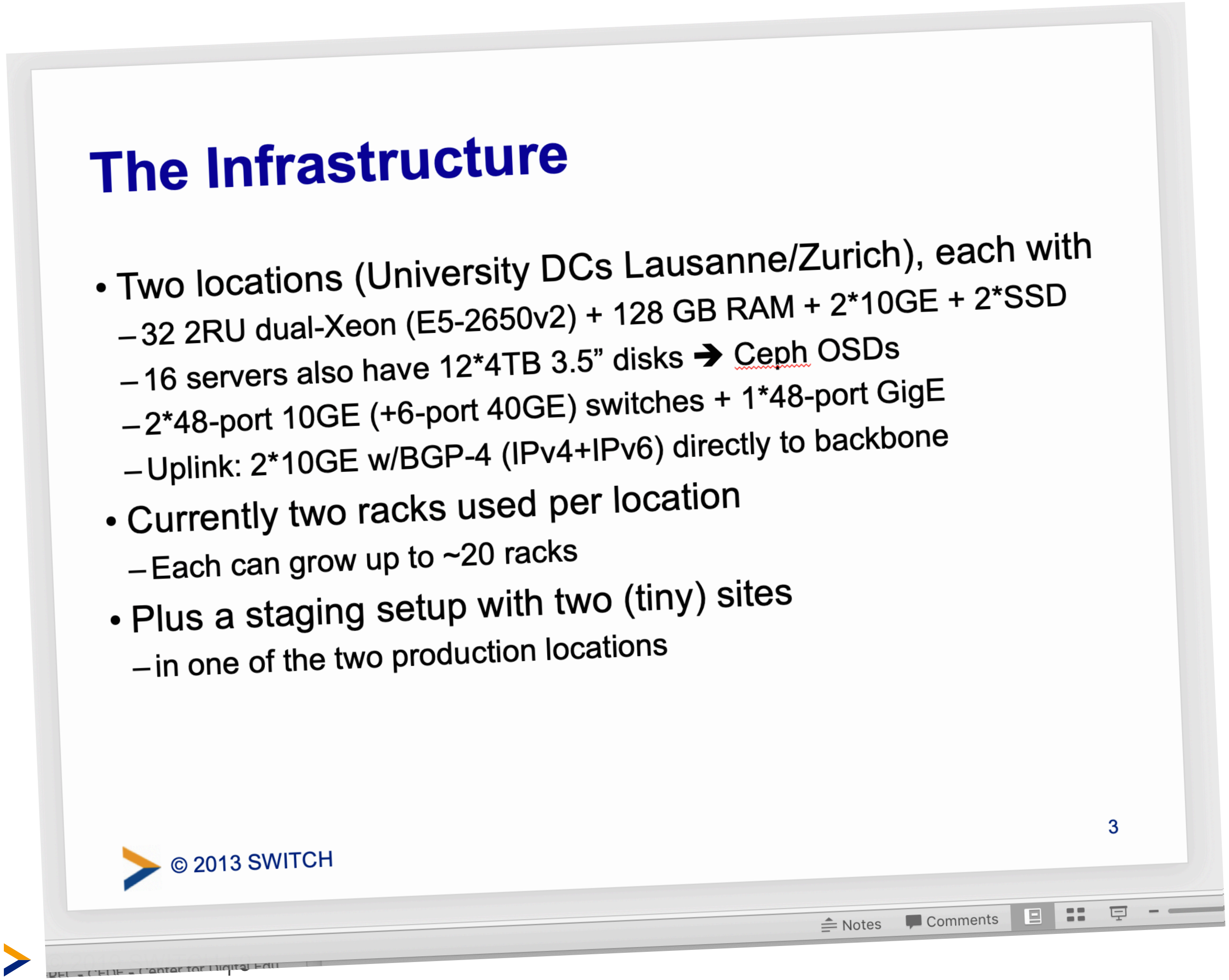
© www.kalenderpedia.de

Angaben ohne Gewähr

Software

openstack™
CLOUD SOFTWARE

But wai



The Infrastructure

- Two locations (University DCs Lausanne/Zurich), each with
  - 32 2RU dual-Xeon (E5-2650v2) + 128 GB RAM + 2*10GE + 2*SSD
  - 16 servers also have 12*4TB 3.5” disks ➔ Ceph OSDs
  - 2*48-port 10GE (+6-port 40GE) switches + 1*48-port GigE
  - Uplink: 2*10GE w/BGP-4 (IPv4+IPv6) directly to backbone
- Currently two racks used per location
  - Each can grow up to ~20 racks
- Plus a staging setup with two (tiny) sites
  - in one of the two production locations

© 2013 SWITCH

3

Storage

1280 TB raw

Compute

448 Cores

3.5 TB RAM

Datacenters in Zurich and Lausanne

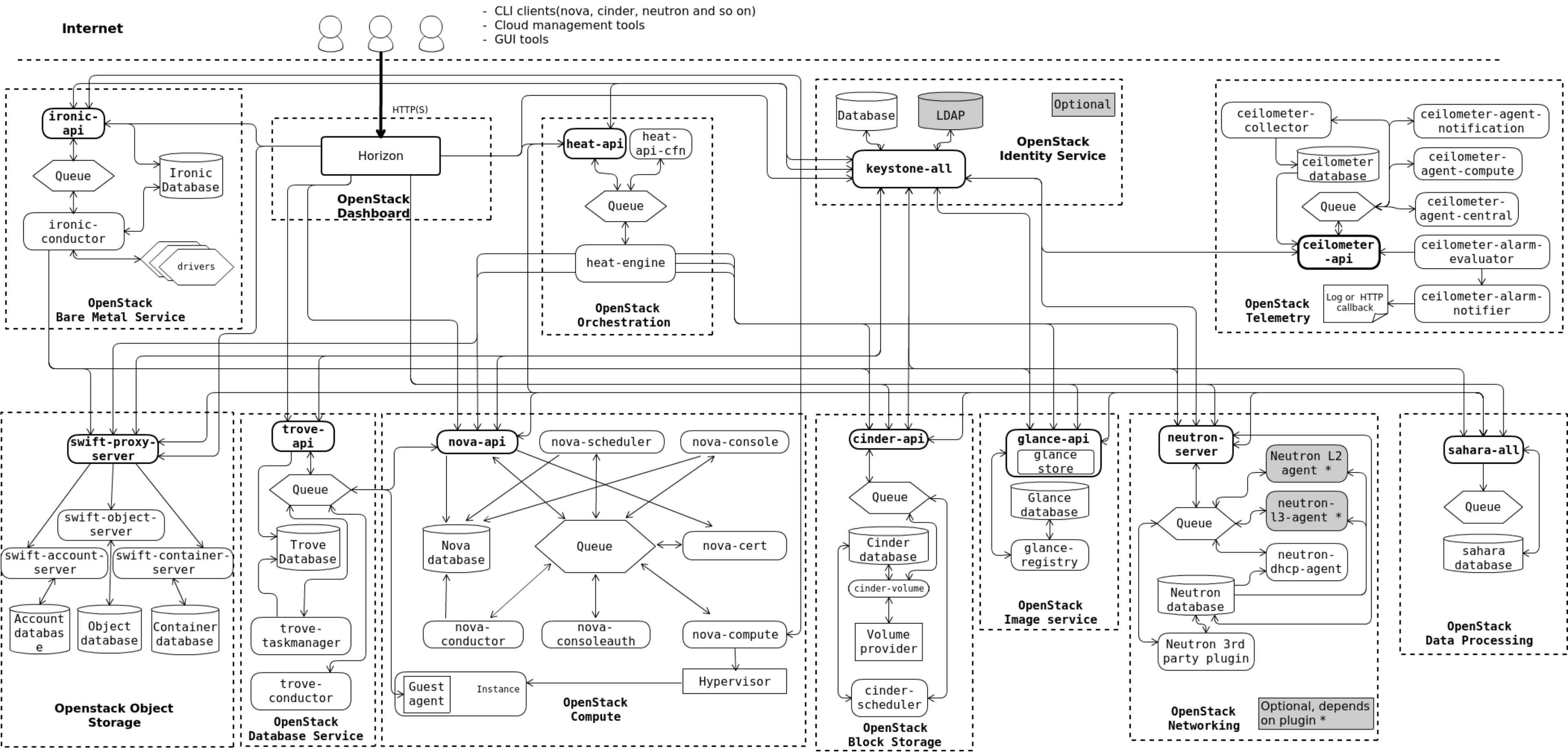
Zurich University of the Arts, Toni Areal

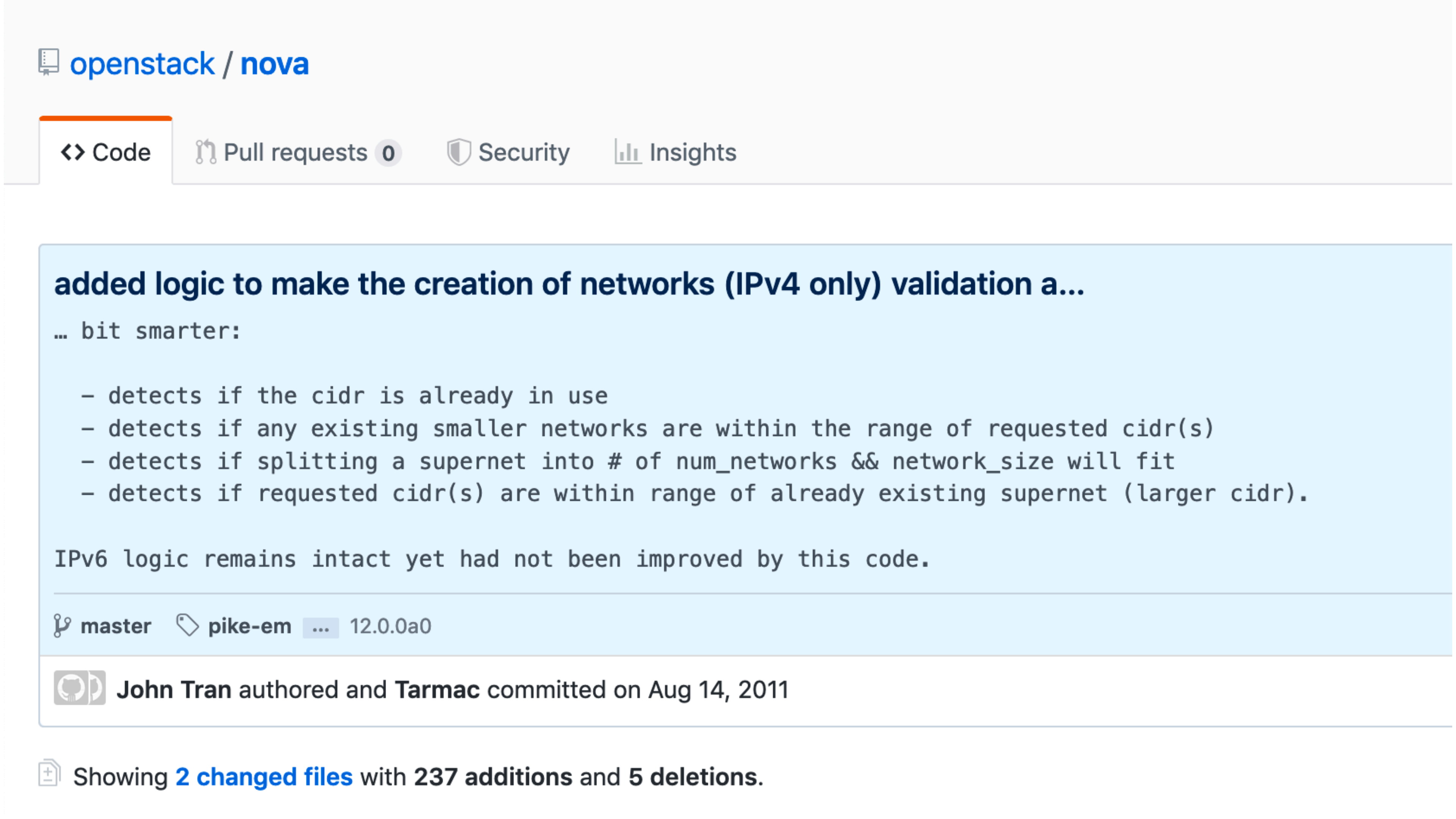
Université de Lausanne, Géopolis

Photo Fabrice Ducrest © UNIL

Now

Physical servers:	~ **400**
CPU cores:	**4076** (physical cores)
Memory:	~ **32** TB
Storage:	~ **10** PB (Ceph SATA) / ~ 1400 Disks ~ **100** TB (Ceph SSD) / 50 NVMe
GPU:	**8** Titan XP **16** T4 **34** P100
Network:	**26** Cumulus Linux 40 & 100Gbs switches Dual 10 Gbs; upgrading to 100 Gbs (Q2 2019) L2 tunnel to campus networks (VPC)


Internet
- CLI clients(nova, cinder, neutron and so on)
- Cloud management tools
- GUI tools
HTTP(S)
Horizon
OpenStack Dashboard
ironic-api
Queue
Ironic Database
ironic-conductor
drivers
OpenStack Bare Metal Service
heat-api
heat-api-cfn
Queue
heat-engine
OpenStack Orchestration
Database
LDAP
Optional
keystone-all
OpenStack Identity Service
ceilometer-collector
ceilometer database
Queue
ceilometer-api
ceilometer-agent-notification
ceilometer-agent-compute
ceilometer-agent-central
ceilometer-alarm-evaluator
ceilometer-alarm-notifier
Log or HTTP callback
OpenStack Telemetry
swift-proxy-server
swift-object-server
swift-account-server
swift-container-server
Account database
Object database
Container database
Openstack Object Storage
trove-api
Queue
Trove Database
trove-taskmanager
trove-conductor
OpenStack Database Service
nova-api
nova-scheduler
nova-console
Nova database
Queue
nova-cert
nova-conductor
nova-consoleauth
nova-compute
Hypervisor
Guest agent
Instance
OpenStack Compute
cinder-api
Queue
Cinder database
cinder-volume
Volume provider
cinder-scheduler
OpenStack Block Storage
glance-api
glance store
Glance database
glance-registry
OpenStack Image service
neutron-server
Neutron L2 agent *
neutron-l3-agent *
Queue
neutron-dhcp-agent
Neutron database
Neutron 3rd party plugin
OpenStack Networking
Optional, depends on plugin *
sahara-all
Queue
sahara database
OpenStack Data Processing

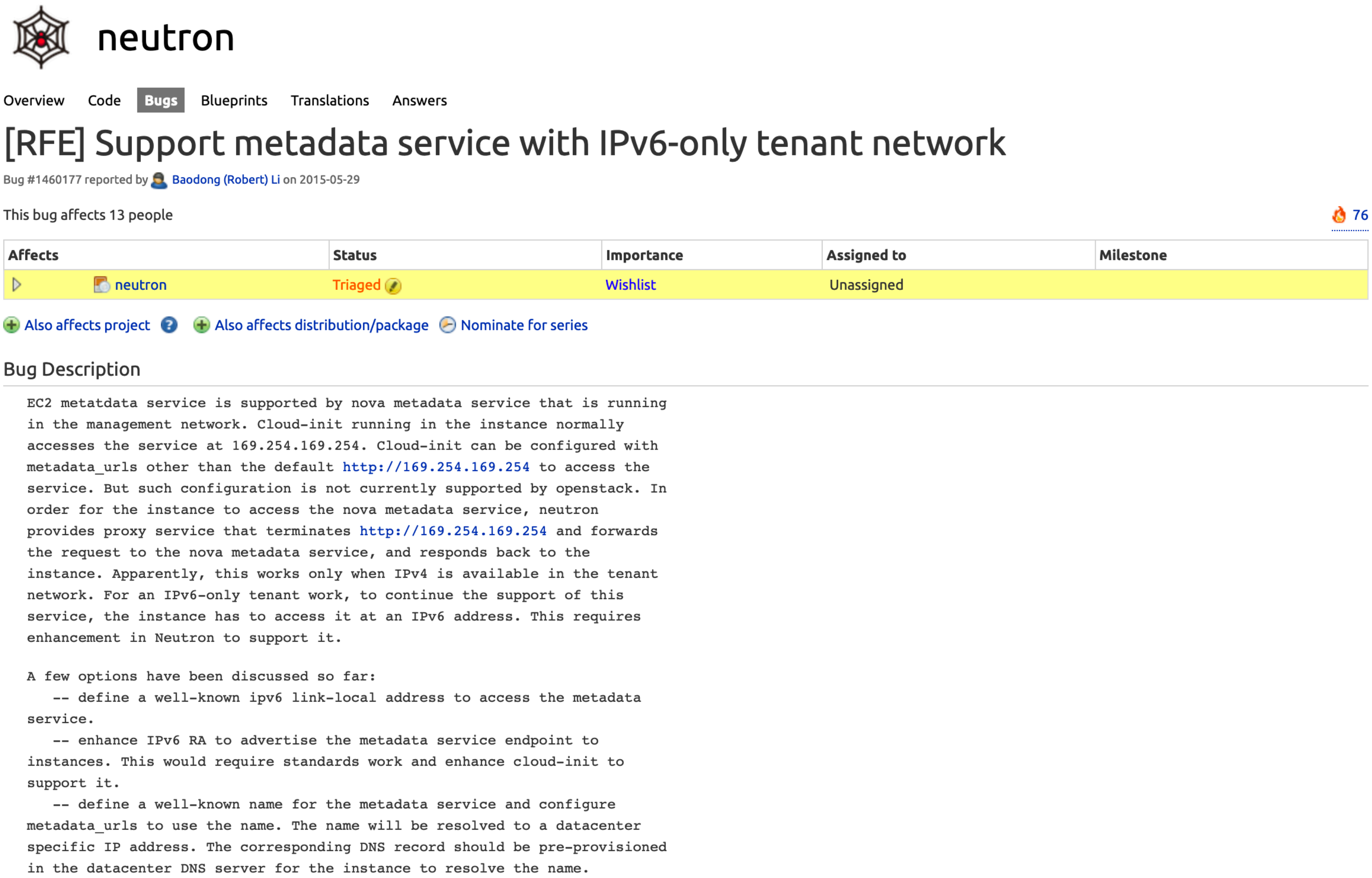


added logic to make the creation of networks (IPv4 only) validation a...

```
… bit smarter:

  - detects if the cidr is already in use
  - detects if any existing smaller networks are within the range of requested cidr(s)
  - detects if splitting a supernet into # of num_networks && network_size will fit
  - detects if requested cidr(s) are within range of already existing supernet (larger cidr).

IPv6 logic remains intact yet had not been improved by this code.
```

John Tran authored and **Tarmac** committed on Aug 14, 2011

Showing 2 changed files with **237 additions** and **5 deletions**.



neutron

Overview Code **Bugs** Blueprints Translations Answers

[RFE] Support metadata service with IPv6-only tenant network

Bug #1460177 reported by Baodong (Robert) Li on 2015-05-29

This bug affects 13 people

76

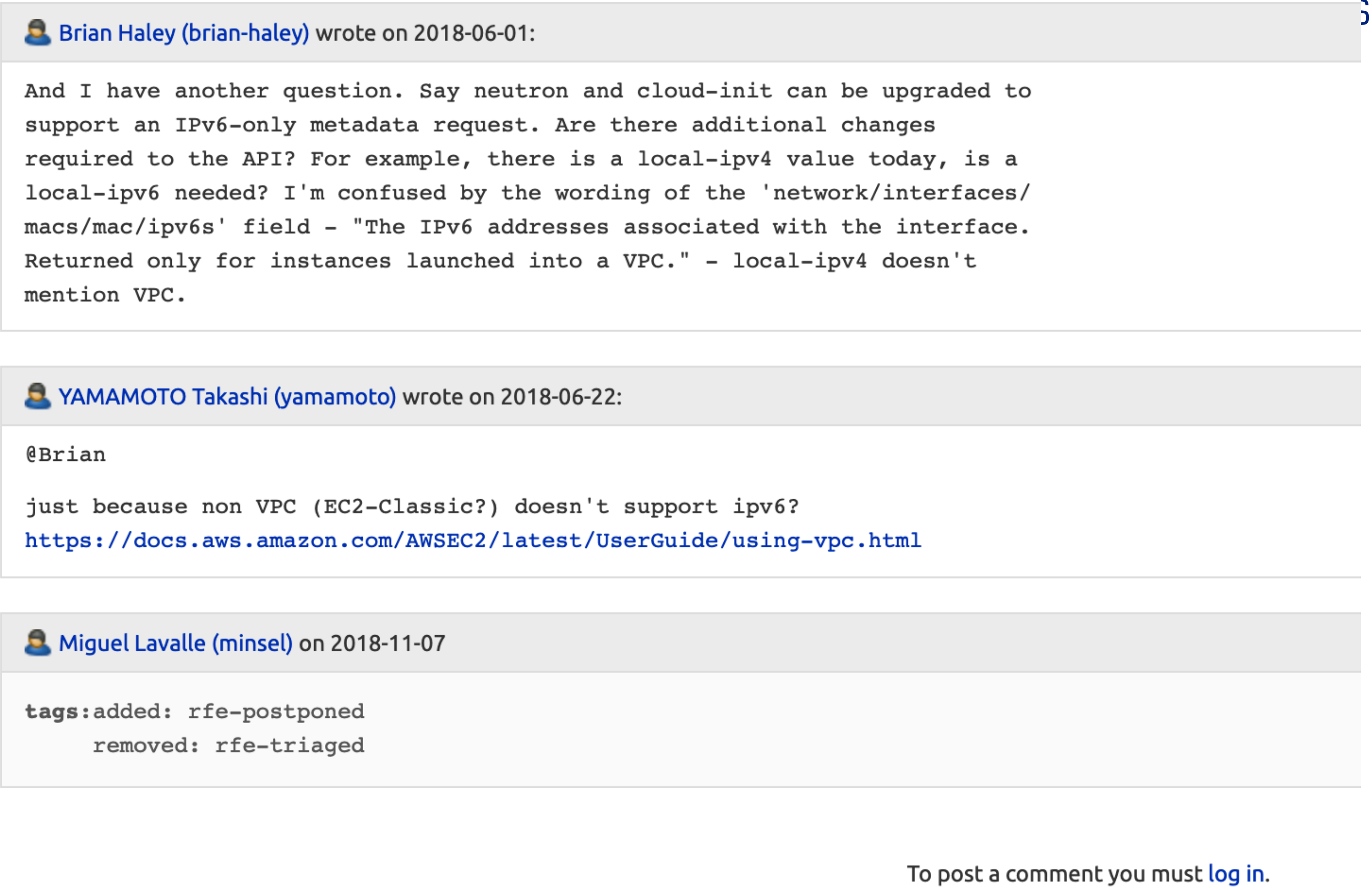
Affects	Status	Importance	Assigned to	Milestone
neutron	Triaged	Wishlist	Unassigned	

Also affects project ? Also affects distribution/package Nominate for series

Bug Description

EC2 metatdata service is supported by nova metadata service that is running in the management network. Cloud-init running in the instance normally accesses the service at 169.254.169.254. Cloud-init can be configured with metadata_urls other than the default http://169.254.169.254 to access the service. But such configuration is not currently supported by openstack. In order for the instance to access the nova metadata service, neutron provides proxy service that terminates http://169.254.169.254 and forwards the request to the nova metadata service, and responds back to the instance. Apparently, this works only when IPv4 is available in the tenant network. For an IPv6-only tenant work, to continue the support of this service, the instance has to access it at an IPv6 address. This requires enhancement in Neutron to support it.

A few options have been discussed so far:
-- define a well-known ipv6 link-local address to access the metadata service.
-- enhance IPv6 RA to advertise the metadata service endpoint to instances. This would require standards work and enhance cloud-init to support it.
-- define a well-known name for the metadata service and configure metadata_urls to use the name. The name will be resolved to a datacenter specific IP address. The corresponding DNS record should be pre-provisioned in the datacenter DNS server for the instance to resolve the name.



Brian Haley (brian-haley) wrote on 2018-06-01:

And I have another question. Say neutron and cloud-init can be upgraded to support an IPv6-only metadata request. Are there additional changes required to the API? For example, there is a local-ipv4 value today, is a local-ipv6 needed? I'm confused by the wording of the 'network/interfaces/macs/mac/ipv6s' field - "The IPv6 addresses associated with the interface. Returned only for instances launched into a VPC." - local-ipv4 doesn't mention VPC.

YAMAMOTO Takashi (yamamoto) wrote on 2018-06-22:

@Brian

just because non VPC (EC2-Classic?) doesn't support ipv6?
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-vpc.html

Miguel Lavalle (minsel) on 2018-11-07

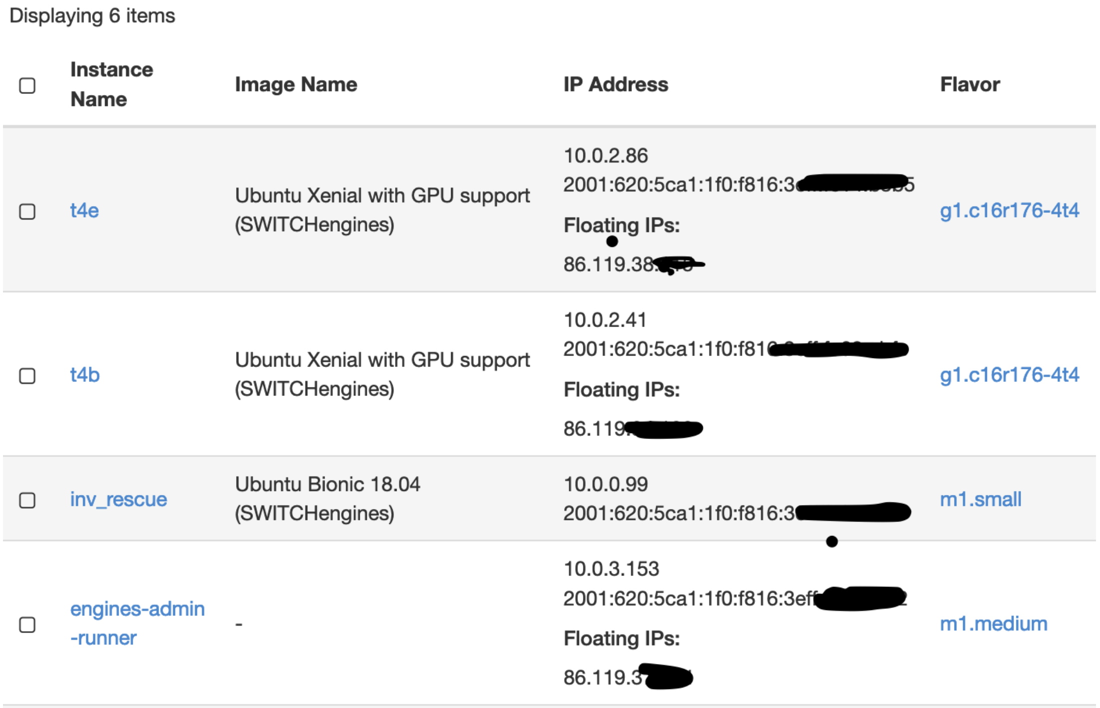
tags:added: rfe-postponed
removed: rfe-triaged

To post a comment you must log in.

Neutron Networking

- Took us 2 years to get a default IPv6 address to a newly started VM
- Still some manual work required for IPv6 routed internal private networks

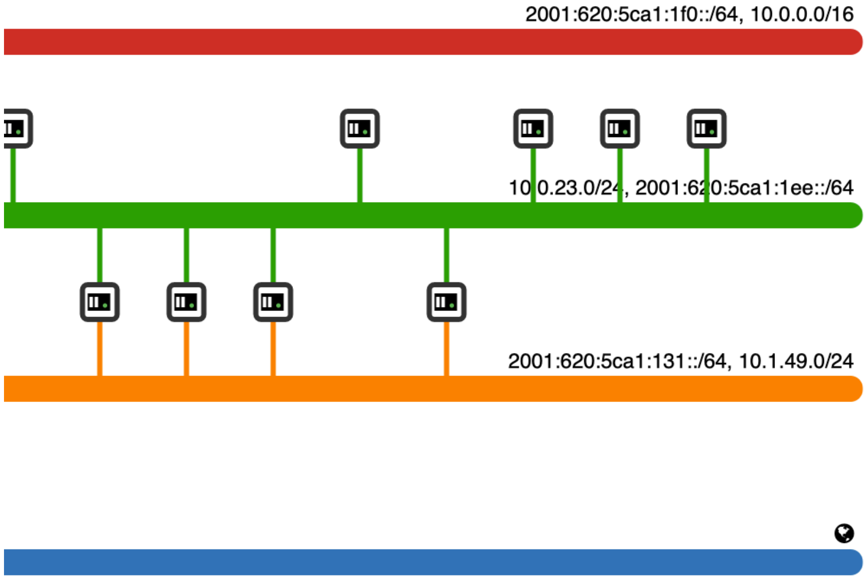
VMs with global routed IPv6



Displaying 6 items

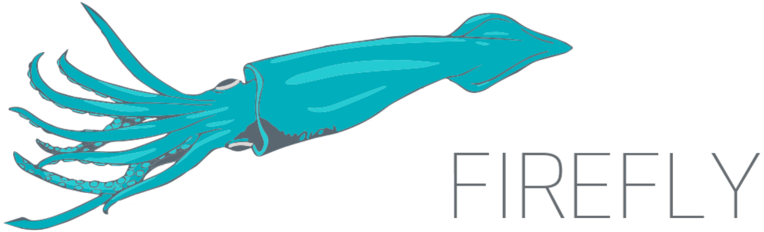
☐	Instance Name	Image Name	IP Address	Flavor
☐	t4e	Ubuntu Xenial with GPU support (SWITCHengines)	10.0.2.86 2001:620:5ca1:1f0:f816:3[redacted] **Floating IPs:** 86.119.38[redacted]	g1.c16r176-4t4
☐	t4b	Ubuntu Xenial with GPU support (SWITCHengines)	10.0.2.41 2001:620:5ca1:1f0:f81[redacted] **Floating IPs:** 86.119[redacted]	g1.c16r176-4t4
☐	inv_rescue	Ubuntu Bionic 18.04 (SWITCHengines)	10.0.0.99 2001:620:5ca1:1f0:f816:3[redacted]	m1.small
☐	engines-admin-runner	-	10.0.3.153 2001:620:5ca1:1f0:f816:3ef[redacted] **Floating IPs:** 86.119.3[redacted]	m1.medium

Complex SDN Setups


2001:620:5ca1:1f0::/64, 10.0.0.0/16
10.0.23.0/24, 2001:620:5ca1:1ee::/64
2001:620:5ca1:131::/64, 10.1.49.0/24

BOBTAIL

DUMPLING


FIREFLY

GIANT

HAMMER

JEWEL

ceph
LUMINOUS

NAUTILUS

Ceph – Software Defined Storage

- IPv6 from the beginning
- No *) problems

*) almost

Brocade L2 Switches

- Clear IPv6 neighbour caches (manually)
- Otherwise machines would loose connectivity

Which is bad in a storage cluster

Hard problems

- Random (huge) performance problems on random VMs (with big amount of IOPS)

Nothing seems to work

Attract “Management Attention”

Blame the running Bluestore migrations

Ask the people at CERN

Spend 2 Weeks trying to reproduce it

Start sniffing the network

Reproducibility

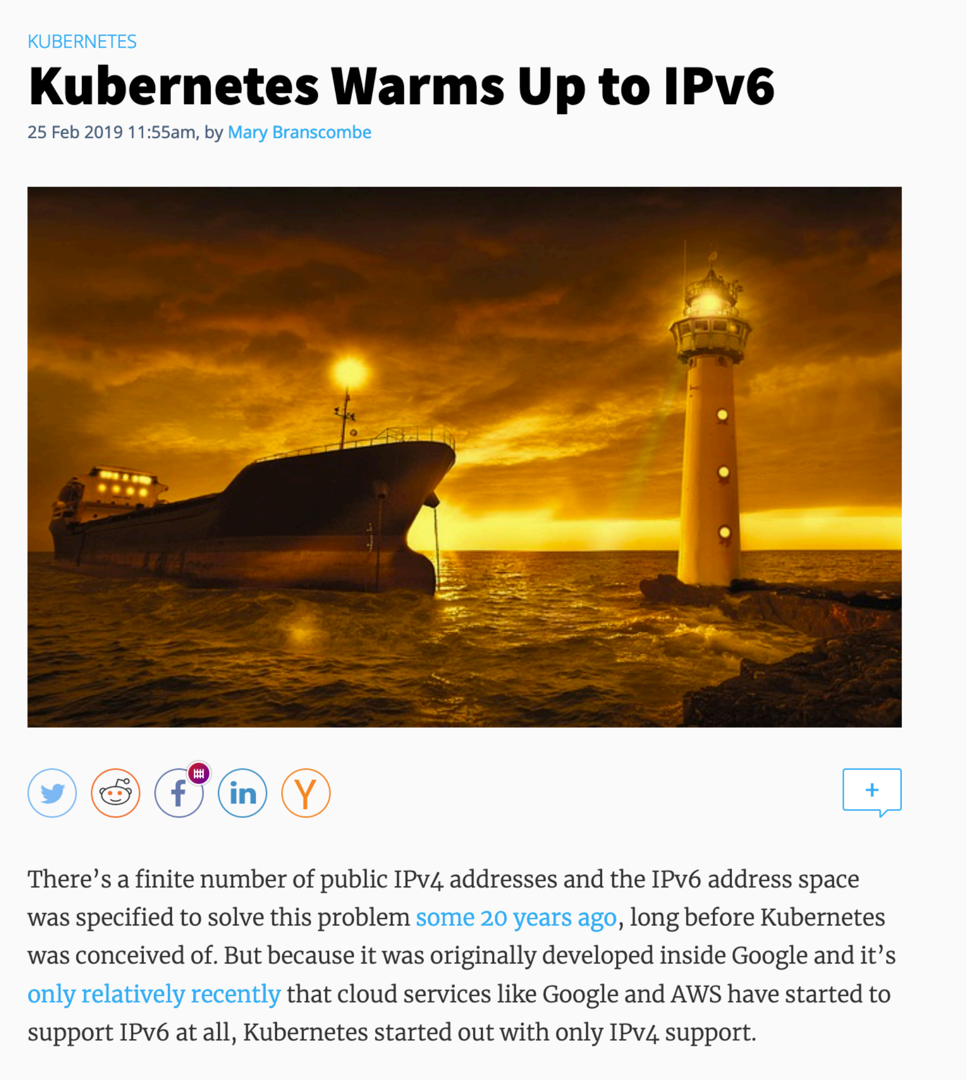
We found that the TCP traffic from the writing VM (the sender) to an OSD (the receiver) was limited to **one** 512-byte TCP segment **every 200 ms**

Install new Kernel

KVM / libvirt

- Talks IPv6 to Ceph Storage Cluster

Law of Constant Pain



KUBERNETES

Kubernetes Warms Up to IPv6

25 Feb 2019 11:55am, by Mary Branscombe

There's a finite number of public IPv4 addresses and the IPv6 address space was specified to solve this problem some 20 years ago, long before Kubernetes was conceived of. But because it was originally developed inside Google and it's only relatively recently that cloud services like Google and AWS have started to support IPv6 at all, Kubernetes started out with only IPv4 support.

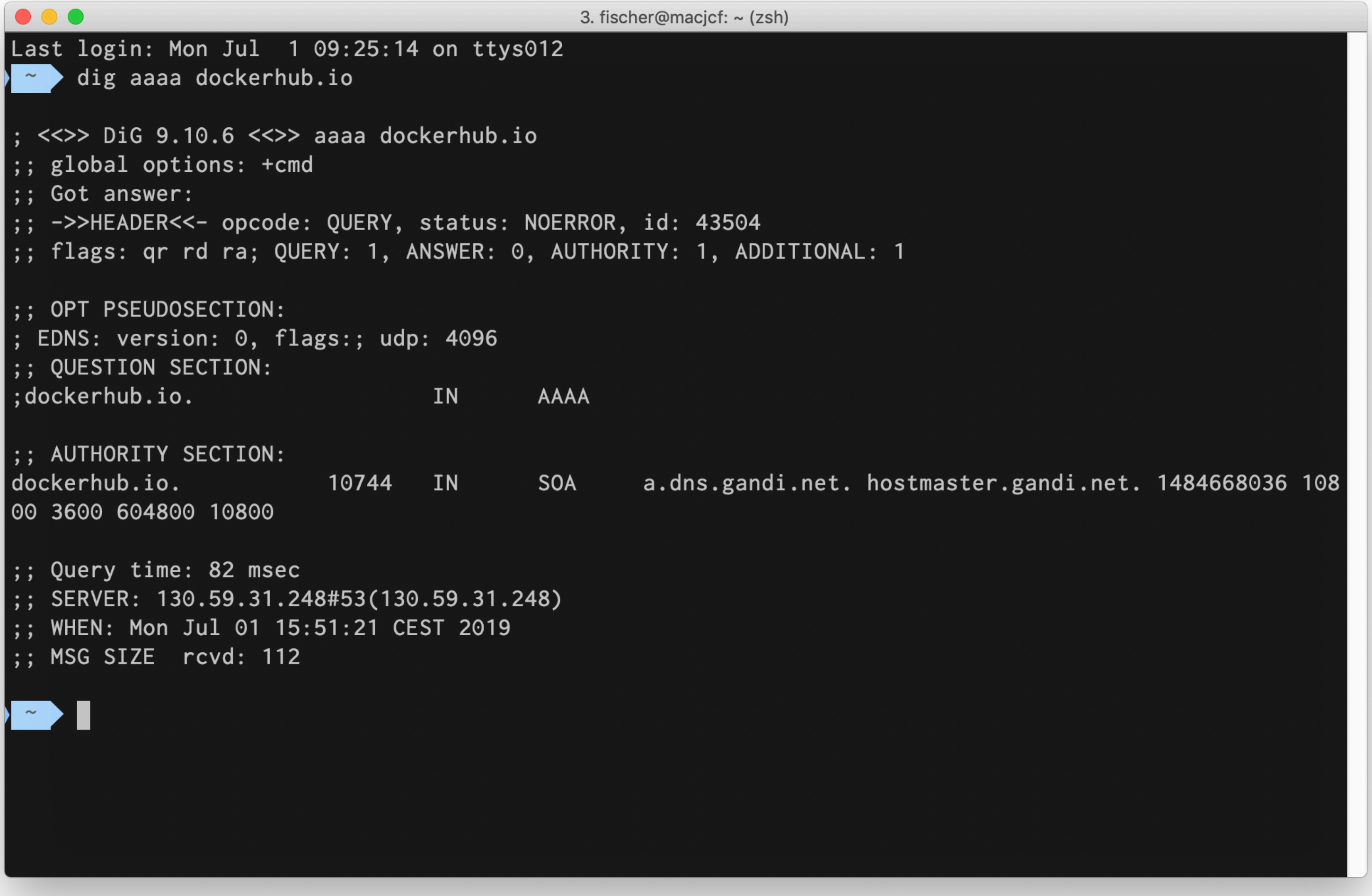
Networking in K8s

- Pods support IPv4 & IPv6 – it just works
- Internal K8s Services only work with IPv4 (even though it is claimed that IPv6 is supported)

=> Run everything in IPv4

The day we shut IPv4 down

- The day we shut of IPv4 outbound connectivity in one of our clusters
- And everything continued to work
- For hours and hours
- Until Kubernetes evicted a number of pods (Because k8s can be stupid)
- And tried to rebuild the images
- And we discovered:

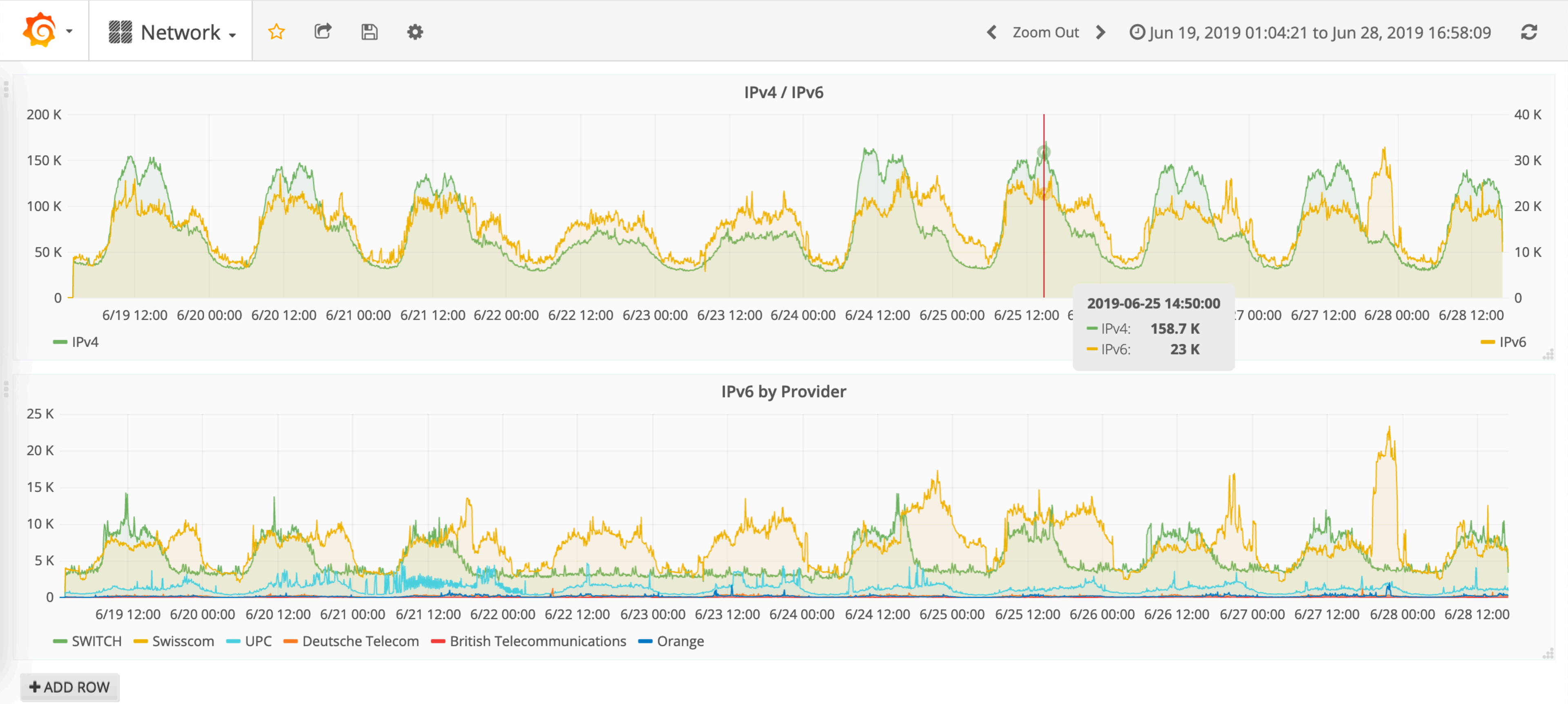


3. fischer@macjcf: ~ (zsh)

```
Last login: Mon Jul  1 09:25:14 on ttys012
 ~  dig aaaa dockerhub.io

; <<>> DiG 9.10.6 <<>> aaaa dockerhub.io
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 43504
;; flags: qr rd ra; QUERY: 1, ANSWER: 0, AUTHORITY: 1, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 4096
;; QUESTION SECTION:
;dockerhub.io.                  IN      AAAA

;; AUTHORITY SECTION:
dockerhub.io.           10744   IN      SOA     a.dns.gandi.net. hostmaster.gandi.net. 1484668036 108
00 3600 604800 10800

;; Query time: 82 msec
;; SERVER: 130.59.31.248#53(130.59.31.248)
;; WHEN: Mon Jul 01 15:51:21 CEST 2019
;; MSG SIZE  rcvd: 112

 ~  
```

What about the users

Enduser IPv6



Network

Zoom Out

Jun 19, 2019 01:04:21 to Jun 28, 2019 16:58:09

SWITCH

Working for a better digital world