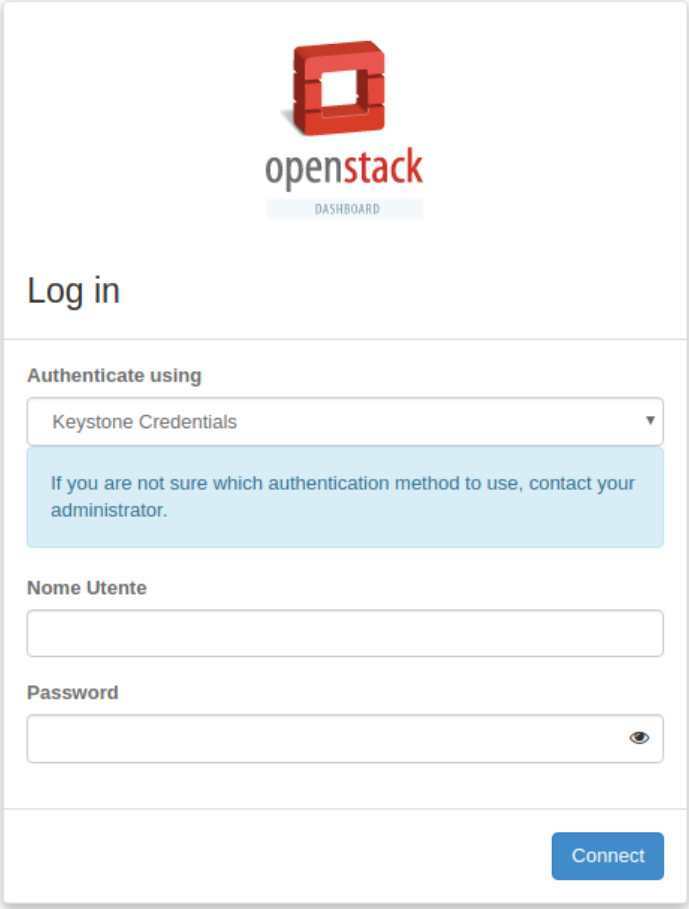
Cloud@CNAF Evolution

Diego Michelotto, Andrea Chierici, Alessandro Costantini, Cristina Duma

Outline

- Requirements
- Infrastructure
- Authn/Authz
- Problems and solutions
- Next steps
- A Cloud for INFN



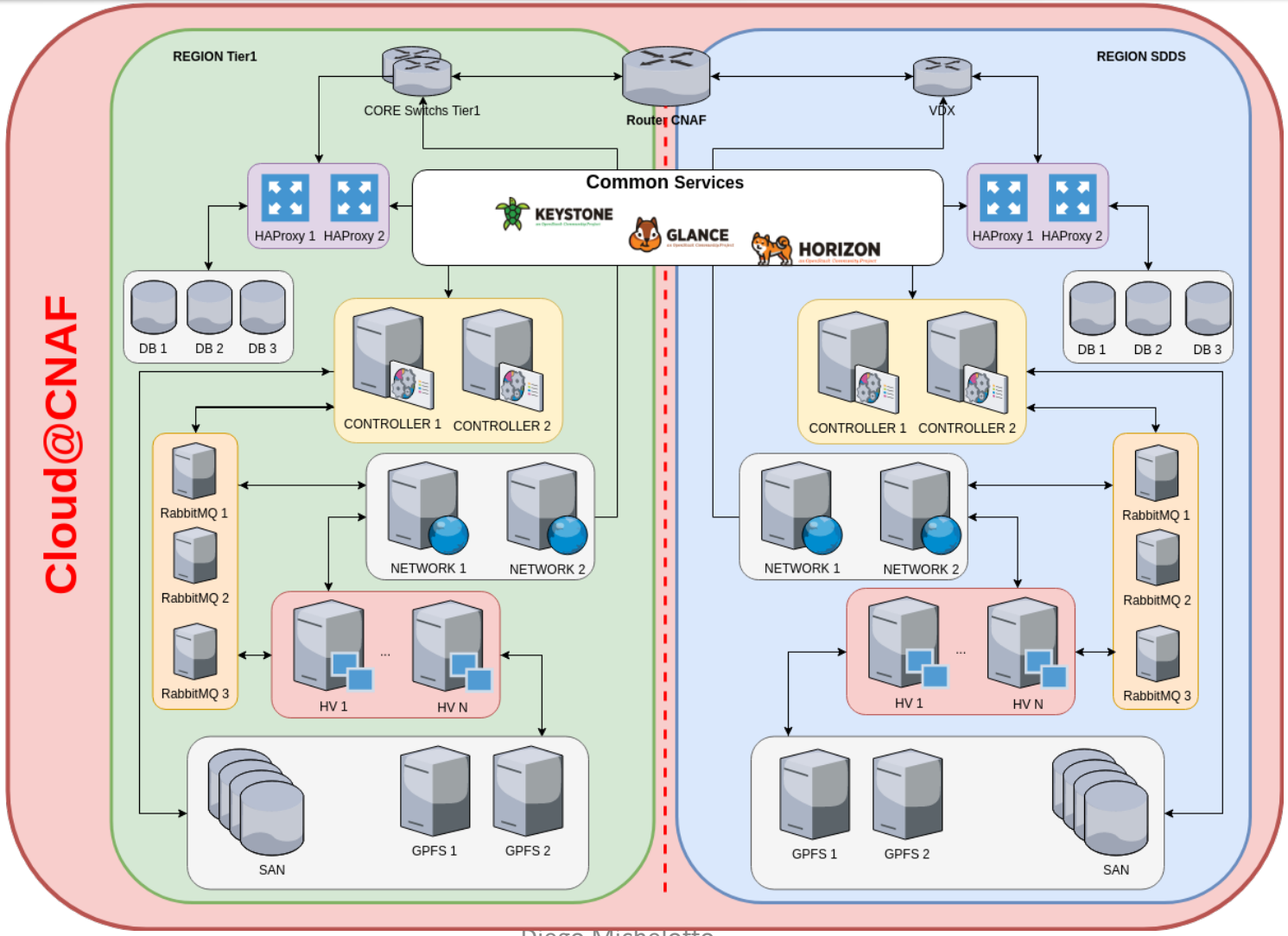
Cloud@CNAF before

- Based on (outdated) OpenStack[1] Mitaka release.
- Only one region managed by SDDS group.
- Used mainly by:
  - INFN experiments,
  - Developers and local users,
  - External, H2020 and regional projects.

Requirements

- Single management domain shared between SDDS and Tier-1 functional units.
- Single infrastructure for all CNAF use-cases:
  - R&D projects,
  - Developers,
  - WLCG and INFN experiments (Pledged experiments).
- Environment separation:
  - Tier-1 and SDDS regions,
  - Tier-1 data access,
  - LHCONE, LHCOPN networks,
  - H2020 Projects.

Infrastructure - Schema



Infrastructure - Services

- **Common core services**
  - Keystone
  - Glance
    - **GPFS backend**
  - Horizon

- **Per region support services**
  - 3 nodes RabbitMQ Cluster
  - 3 nodes Mysql Percona Multi Master Cluster
  - 2 Nodes HAProxy + Keepalived
    - Serve and manage all OpenStack services and DBs

- **Per region services**
  - Cinder
    - GPFS backend
  - Nova
    - GPFS backend as storage, Libvirt backend as virtualizator
  - Neutron
    - Linuxbridge, **VLAN**, External Network (/23)
  - Heat
- **Virtual vs Bare Metal**
  - Only openstack-nova-compute, neutron-linuxbridge, neutron-dhcp, neutron-l3-agent and neutron-metadata are on physical nodes (**Compute nodes and Network nodes**).
  - All other openstack **service are virtualized and replicated on different virtualization systems** (oVirt, VMWare).
  - DBs are on physical nodes on diffent racks with 15k SAS disks.

Infrastructure - Deployments

- **Rocky** (in production):
  - **SDDS region** ~ 1400 cores, 5TB RAM, 16TB Shared FS, 28TB local FS.
  - **Tier1 region** ~ 500 TB-N shared FS, ~5200 cores (to be added soon).

- **Testbed**:
  - 2 smaller production-like setup regions.
  - Necessary to test pre-production of new services, puppet classes and upgrades.
  - ISO 27001 testbed.

- **SGSI** separate instance

Infrastructure - SGSI

- CNAF got **ISO27001** certification to address strictly secure data handling requirements ("Sistema per la Gestione della Sicurezza dell'Informazione").
- A cloud deployment is going to be setup in June to host new experiments:
  - Harmony (genomics),
  - Alleanza Contro il Cancro (biomedic).
- **Separated** and **isolated** infrastructure.
- **Ceph** will be used as storage backend.

Authn/Authz

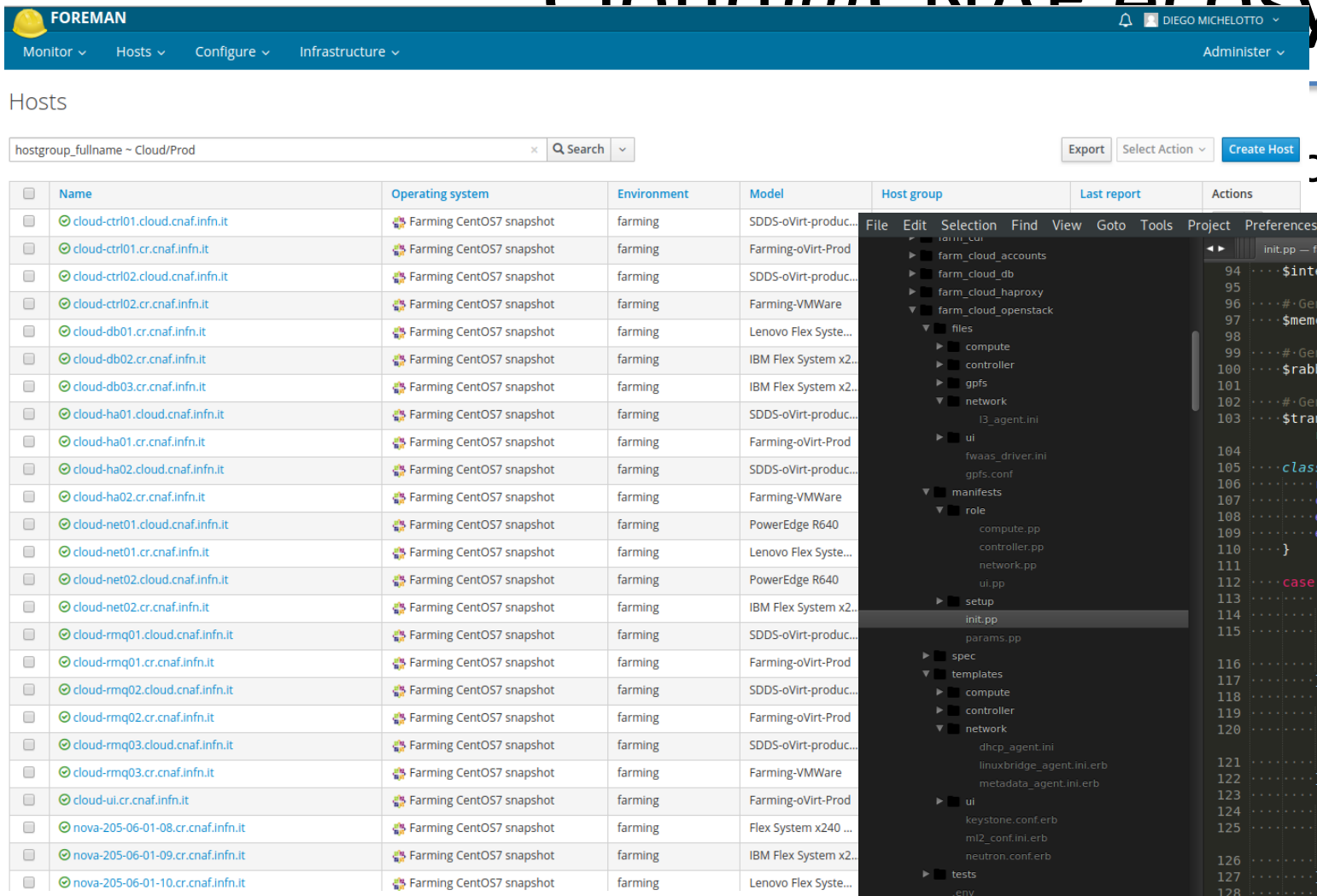
- Based on **OpenID-connect** provided by INDIGO-IAM[2]
  - Dedicated INDIGO-IAM service https://iam.cnaf.infn.it for CNAF, INFN AAI and EduGAIN users, **permissions managed** through IAM **groups**:
    - **No group** membership means **no cloud access**.
    - Users in **"cloud" group** can access cloud in the **shared project CNAF** with limited resources through **ephemeral user.**
    - Users in **"cloud/user" group** can access cloud in two **project: CNAF** one and **personal** one through **ephemeral user**.
    - Users in **"cloud/local" group** can access cloud through **keystone mapped user and projects**.
  - **Other INDIGO-IAM services** for R&D project like DODAS, eXtreme-DataCloud, DEEP Hybrid Data Cloud, etc. mapped on **ephemeral user with dedicated project.**

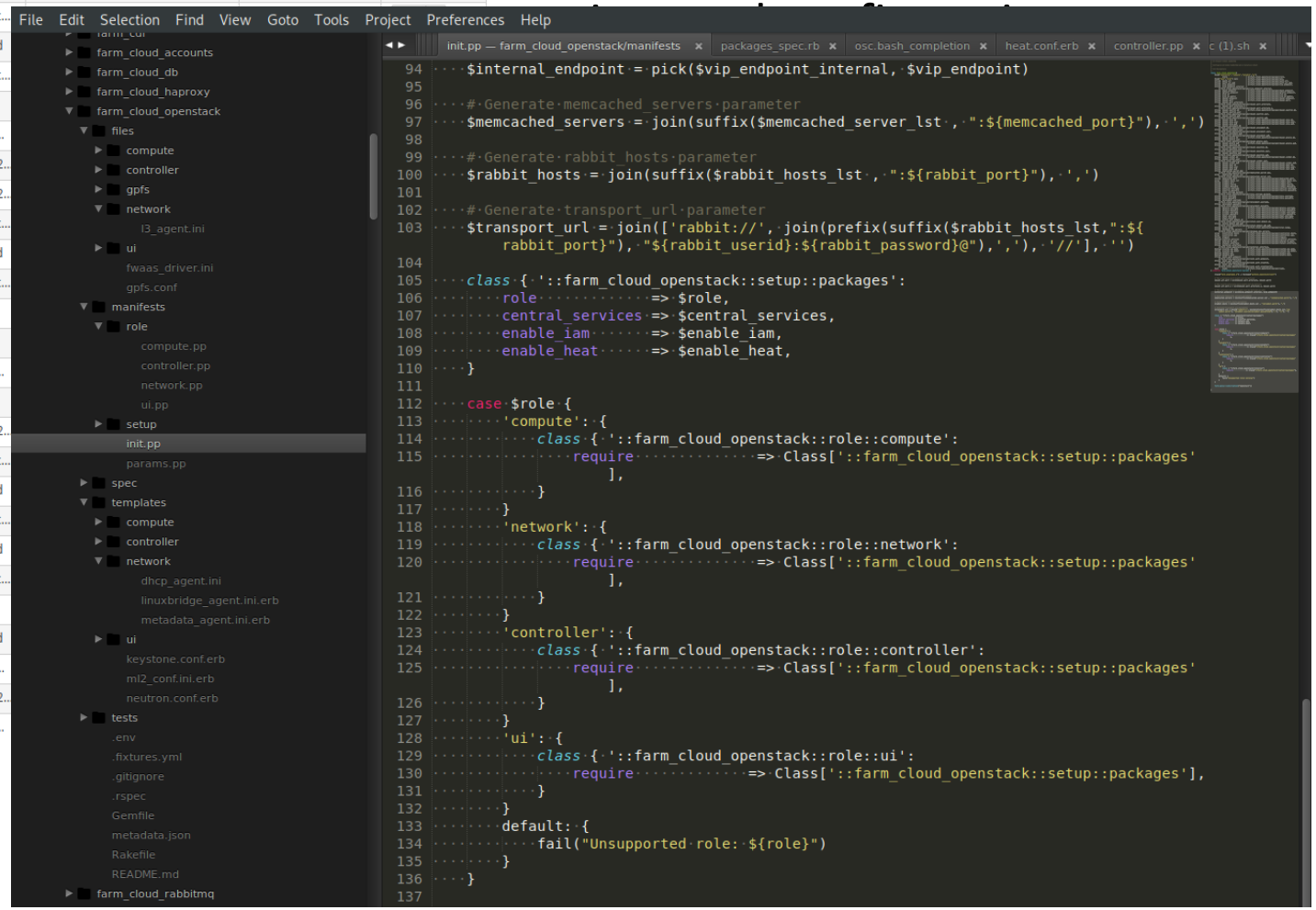
Cloud@CNAF ecosystem

- **Provisioning and configuration** managed via The Foreman[3] and Puppet[4].
  - Developed our own puppet classes for all the clusters, services and configurations.

Cloud@CNAF ecosystem

oreman[3] and Puppet[4].





Cloud@CNAF ecosystem

- **Provisioning and configuration** managed via The Foreman[3] and Puppet[4].
  - Developed our own puppet classes for all the clusters, services and configurations.
- All software repositories are locally **cloned** and **snapshotted**.
- Infrastructure **deployment** and **functionalities** are **tested** with Rally[5], smoke and stress test.

- **Provisi**
  - Deve
- All soft
- Infrastr
  and str

et[4].
ions.

smoke

Rally task results

- Task overview
- Input file
- ► Authenticate
- ► CinderVolumes
- ► GlanceImages
- ► KeystoneBasic
- ► NeutronNetworks
- ► NovaKeypair
- ► NovaServers
- ► Quotas

Task overview

Scenario ▲	Load duration (s)	Full duration (s)	Iterations	Runner	Errors	Hooks	Success (SLA)
Authenticate.keystone	0.482	2.688	1	constant	0	0	✔
CinderVolumes.create_and_attach_volume	24.119	40.619	1	constant	0	0	✔
CinderVolumes.create_and_delete_snapshot	5.990	24.749	1	constant	0	0	✔
CinderVolumes.create_and_delete_volume	6.080	14.943	1	constant	0	0	✔
CinderVolumes.create_and_delete_volume-2	8.156	15.075	1	constant	0	0	✔
CinderVolumes.create_and_delete_volume-3	8.148	16.293	1	constant	0	0	✔
CinderVolumes.create_and_extend_volume	10.512	17.821	1	constant	0	0	✔
CinderVolumes.create_and_list_snapshots	3.869	26.549	1	constant	0	0	✔
CinderVolumes.create_and_list_volume	5.388	15.000	1	constant	0	0	✔
CinderVolumes.create_and_list_volume-2	5.664	16.950	1	constant	0	0	✔
CinderVolumes.create_and_upload_volume_to_image	14.371	24.213	1	constant	0	0	✔
CinderVolumes.create_from_volume_and_delete_volume	10.490	30.689	1	constant	0	0	✔
GlanceImages.create_and_delete_image	2.822	7.266	1	constant	0	0	✔
GlanceImages.create_and_list_image	2.830	9.485	1	constant	0	0	✔
GlanceImages.list_images	0.170	3.525	1	constant	0	0	✔
KeystoneBasic.add_and_remove_user_role	1.347	12.692	1	constant	0	0	✔
KeystoneBasic.create_add_and_list_user_roles	1.296	13.044	1	constant	0	0	✔
KeystoneBasic.create_and_delete_role	0.676	10.338	1	constant	0	0	✔
KeystoneBasic.create_and_delete_service	1.231	10.219	1	constant	0	0	✔
KeystoneBasic.create_and_list_tenants	1.271	13.052	1	constant	0	0	✔
KeystoneBasic.create_update_and_delete_tenant	1.590	11.075	1	constant	0	0	✔
KeystoneBasic.get_entities	1.767	19.408	1	constant	0	0	✔
NeutronNetworks.create_and_delete_networks	2.785	11.706	1	constant	0	0	✔
NeutronNetworks.create_and_delete_ports	3.632	20.078	1	constant	0	0	✔
NeutronNetworks.create_and_delete_routers	10.733	29.603	1	constant	0	0	✔
NeutronNetworks.create_and_delete_subnets	3.899	17.180	1	constant	0	0	✔
NeutronNetworks.create_and_list_networks	1.747	16.396	1	constant	0	0	✔
NeutronNetworks.create_and_list_ports	2.577	18.899	1	constant	0	0	✔
NeutronNetworks.create_and_list_routers	8.257	35.458	1	constant	0	0	✔
NeutronNetworks.create_and_list_subnets	2.413	17.567	1	constant	0	0	✔
NeutronNetworks.create_and_update_networks	1.950	18.473	1	constant	0	0	✔
NeutronNetworks.create_and_update_ports	3.900	23.136	1	constant	0	0	✔
NeutronNetworks.create_and_update_routers	8.870	30.255	1	constant	0	0	✔
NeutronNetworks.create_and_update_subnets	3.682	24.011	1	constant	0	0	✔

Cloud@CNAF ecosystem

- **Provisioning and configuration** managed via The Foreman[3] and Puppet[4].
  - Developed our own puppet classes for all the clusters, services and configurations.
- All software repositories are locally **cloned** and **snapshotted**.
- Infrastructure **deployment** and **functionalities** are **tested** with Rally[5], smoke and stress test.
- Use of Rundeck[6] for **operations**.

RUNDECK | ops | Jobs | Nodes | Commands | Activity | dmichelotto | help

Jobs (24) Filter Expand All Collapse All

Job Actions

- Cloud
  - Infra
    - Configure IPMI R2
    - disable-mycls — Disable a Percona cluster member on the ha01/ha02 load balancers.
    - enable-mycls — Enable a Percona cluster member on the ha01/ha02 load balancers.
    - Install Compute 1 Network and Storage
    - Install Compute 1&2 Network, Storage GPFS [NO UPDATE & REBOOT]
    - Install Compute 2 GPFS
    - Install Compute 3 Complete installation
    - Install Compute LSD
  - Maintenance
    - start-compute-nodes — Start OpenStack Nova Compute services on all compute nodes
    - start-controller-nodes
    - start-network-nodes
    - stop-compute-nodes — Shutdown services on all OpenStack compute nodes
    - stop-compute-nodes-fe — Shutdown services on all OpenStack compute nodes in Ferrara
    - stop-controller-nodes
    - stop-controller-nodes-fe — Stop controller-nodes at FE
    - stop-network-nodes
    - Upgrade

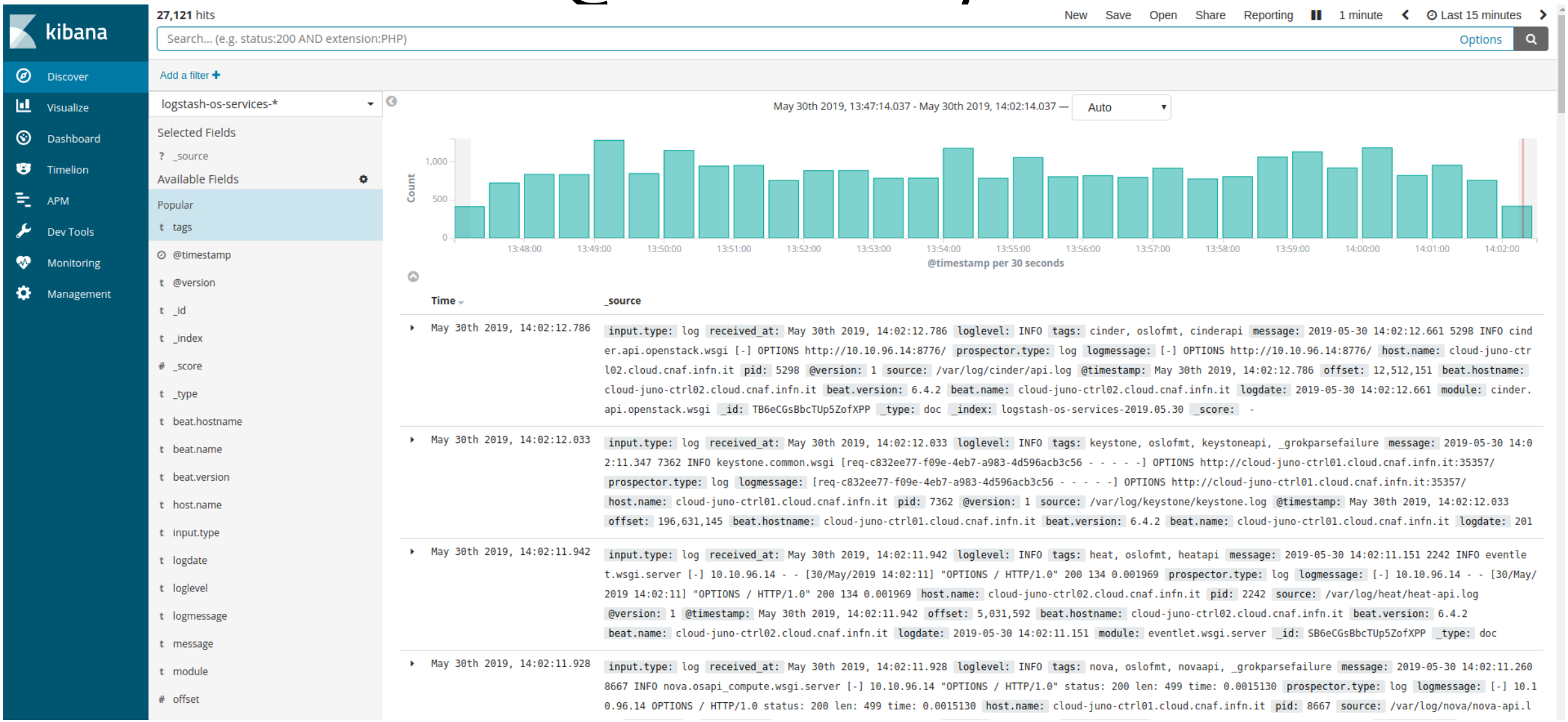
- Pr].
  –
- Al
- In ke
  ar
- U:

is

Cloud@CNAF ecosystem

- **Provisioning and configuration** managed via The Foreman[3] and Puppet[4].
  - Developed our own puppet classes for all the clusters, services and configurations.
- All software repositories are locally **cloned** and **snapshotted**.
- Infrastructure **deployment** and **functionalities** are **tested** with Rally[5], smoke and stress test.
- Use of Rundeck[6] for **operations**.
- Use of ELK[7] stack for **log collection, management and analysis** of the infrastructure.

Cloud@CNAF ecosystem



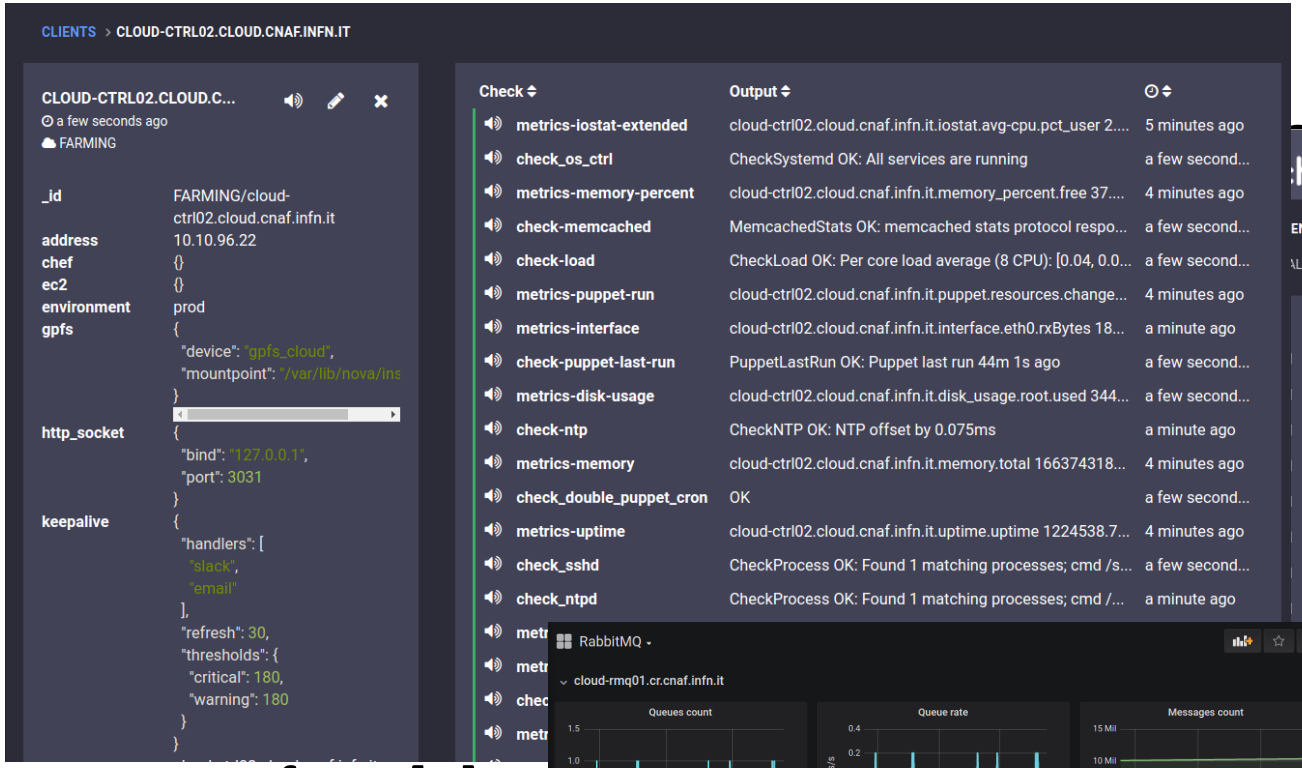
Cloud@CNAF ecosystem

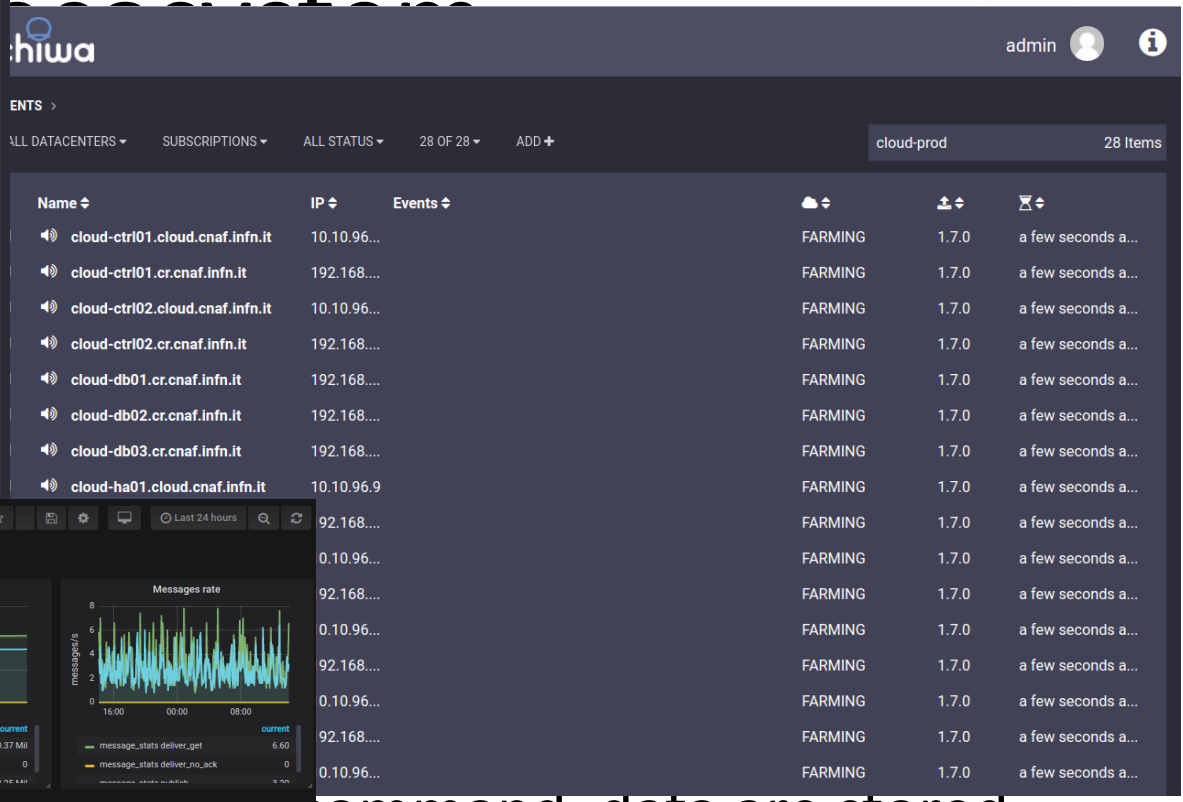
- **Provisioning and configuration** managed via The Foreman[3] and Puppet[4].
  - Developed our own puppet classes for all the clusters, services and configurations.
- All software repositories are locally **cloned** and **snapshotted**.
- Infrastructure **deployment** and **functionalities** are **tested** with Rally[5], smoke and stress test.
- Use of Rundeck[6] for **operations**.
- Use of ELK[7] stack for **log collection, management and analysis** of the infrastructure.
- Basic accounting made using `Openstack usage list` command, data are stored in InfluxDB[8] timeseries database and displayed with Grafana[9].

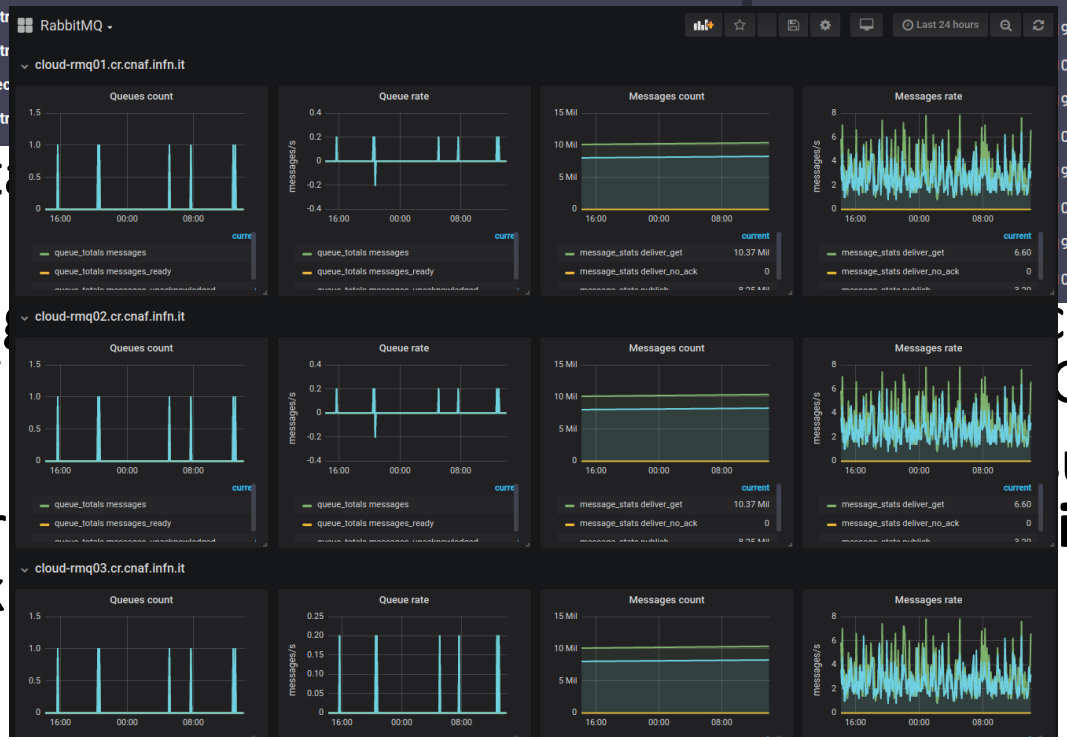
- **Pr** [4].
  - ns.
- Al
- In noke
  ar
- Us
- Us
  in
- Ba stored
  in

Cloud@CNAF ecosystem

- **Provisioning and configuration** managed via The Foreman[3] and Puppet[4].
  - Developed our own puppet classes for all the clusters, services and configurations.
- All software repositories are locally **cloned** and **snapshotted**.
- Infrastructure **deployment** and **functionalities** are **tested** with Rally[5], smoke and stress test.
- Use of Rundeck[6] for **operations**.
- Use of ELK[7] stack for **log collection, management and analysis** of the infrastructure.
- Basic accounting made using `Openstack usage list` command, data are stored in InfluxDB[8] timeseries database and displayed with Grafana[9].
- All services and performance are **monitored with Sensu**[10], is used InfluxDB and Grafana for data storing and displaying. Alert **notification** generate by Sensu is sent via **Slack**[11] and **e-mail**.







- Use of ELK[7] st
infrastructure.
- Basic accounting
in InfluxDB[8] ti
- All services and
and Grafana for
is sent via **Slack**

command, data are stored
Grafana[9].

u[10], is used InfluxDB
ication generate by Sensu

Cloud@CNAF security

- For the delegation of responsibility we align to what Harmony group will produce.
  - Delegate resp. to user with root access, whether internal or external
- For traceability and security reasons, users can only use images provided by Cloud@CNAF admins:
  - **Customized images** with rsyslog service enabled
  - **Injected ssh key** for root access used only in case of security incident.
- VMs external access through frontier firewall:
  - By default, CNAF outer perimeter **firewall blocks incoming access**.
  - Ports can be **opened upon express request**.

Problems and Solutions (1/2)

- Libvirt doesn't recognize GPFS as distributed file system.
  - Locally patched and tested for libvirt 4.5.
  - Patch submitted and **accepted** upstream for libvirt 5.0. https://bugzilla.redhat.com/show_bug.cgi?id=1679528
    - CentOS has not yet backported the patch but has taken it into consideration.

- Nova APIs fail when receive requests from HAProxy.
  - https://bugs.launchpad.net/nova/+bug/1728732
  - **Patch backported** with puppet ad-hoc class.

Problems and Solutions (2/2)

- CPU capabilities are different and can block live migrations. Two cases:
  - **Different vendor**: solved using **host aggregate**.
    - AMD vs Intel
  - **Same vendor** but different architecture, two sub-cases:
    - High number of nodes for each type, solved using **host aggregate.**
      - e.g. AMD G4 vs. AMD G5 or Intel Broadwell vs. Intel Skylake.
    - Low number of nodes for each type, solved configuring **CPU model** in nova.conf with the CPU baseline between different.

Next steps

- **Elastic partitioning of Farm**
  - Possibility to detach WNs from the production farm and assign them to cloud partition and vice versa.
- Fine tuning of virtual machines monitoring and accounting.
- Implement workflow management to monitor VM lifecycle.
- Improve logs parsing and analysis.
- GPU virtualization.

A Cloud for INFN

- We can now give access to Tier-1 resources both through standard grid approach and Cloud.
- Goal is to federate our infrastructure with other INFN clouds to implement a **unique** INFN cloud.
  - Federation mechanism: INFN-CC or IAM.
- (Possibly) complemented by Data-Lake like infrastructure for data.

Credits

- CNAF Network team:
  - Study, design and setup for Cloud@CNAF networks.
- CNAF Storage team:
  - GPFS setup for Cloud@CNAF.
- CNAF Software Development team:
  - Setup and integrations of IAM and ELK.

Thanks


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
спасибо
obrigado
merci
bedankt
dziękuję
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
grazie
arigatō
dakujem
мерси
감사합니다
ありがとう
tapadh leat
hvala
mauruuru
sagolun
taiku
takk
рахмат
Баярлалаа
спасибо
faafetai lava
kiitos
dankie
dhanyavad
köszönöm
gràcie
enkosi
bayarlalaa
nanni
nandri
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
akun
dankon
acio
sulpáy
chnorakaloutioun
gratias ago
gràcies
sobodi
dèkuji
mèsi
didi madloba
najis tuke
kam sah hamnida
তোমাকে ধন্যবাদ
rahmat
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
merce
trugarez
xièxie
ευχαριστώ
mersi
kia ora
barka
welalin
tack
spas

References

[1] Openstack: https://www.openstack.org/

[2] INDIGO-IAM: https://www.indigo-datacloud.eu/identity-and-access-management

[3] The Foreman: https://www.theforeman.org/

[4] Puppet: https://puppet.com/

[5] Rally: https://rally.readthedocs.io/en/latest/

[6] Rundeck: https://www.rundeck.com/open-source

[7] ELK: https://www.elastic.co/

[8] InfluxDB: https://www.influxdata.com/

[9] Grafana: https://grafana.com/

[10] Sensu: https://sensu.io/

[11] Slack: https://slack.com