1st SIG-CISS meeting, Amsterdam, September 2017

Overview

Agenda: https://wiki.geant.org/display/CISS/1st+SIG-CISS+meeting

Participants: https://eventr.geant.org/events/2723 (~25)

Introduction to SIG-CISS, motivations and objectives

Speaker: Guido Aben, AARNET

The issue is that each NREN created their own interfaces around ~equivalent cloud middleware, thus users are unable to easily span multiple NRENs.

OpenStack Passport - generic cloud credits that can be spent on any company using OpenStack and supporting OpenStack Passport program.

RefStack - https://wiki.openstack.org/wiki/RefStack

Who we are together... How we organize...

Speaker: Francisco Bernabé, SURFsara

Discussion about OpenStack bodies.

Data services and management

Speaker: Maciej Brzezniak, PSNC

Challenges: increasing volume & performance, everything online, long-term storage, HPC storage, cloud VMs volumes, data analytics

PSNC uses Ceph-based services for backup, archival, repositories, big data

10 PB of data distributed over 2 sites (PSNC, SURFsara) using dCache and tape storage.

Open Research Cloud

Speaker: Simon Leinen, SWITCH

Slides: https://wiki.geant.org/download/attachments/88769009/ORC.pdf?version=1&modificationDate=1506339123422&api=v2

Organization web: http://www.openresearchcloud.org/

Working groups:

- Architecture Alignment WG
- Business Processes WG
- Data Management WG
- Identity WG
- Networking WG
- Security WG

Meeting in Boston attended mainly by Internet2 experts.

Cloud Services for Synchronization and Sharing (CS3)

Speaker: Jakub Moscicki, CERN

Next workshop: http://cs3.cyfronet.pl/

CERN resources: 200PB of tapes, 4,5B files on AFS, 1PB Ceph (and more I didn't catch)

Long advertisement for the Workshop on Cloud Services for Synchronisation and Sharing, 29 - 31 January 2018.

OpenCloudMesh (OCM) Project

Speaker: Peter Szegedi, GÉANT

Project members: GEANT, CERN and OwnCloud.

"Secure, open, frictionless file sharing everywhere"

Managing OS images

Speaker: Simon Leinen, SWITCH

SWITCH cloud offering - SWITCHengines (https://www.switch.ch/engines/), Virtual Private Cloud, SCALE-UP.

Pay-per-use service - they issue bills every 3 months.

Issues:

- users immediately mess up images by disabling security features - disable automatic upgrades, enable password authentication, etc
- SWITCH provides local "official" images with some specific setup (e.g. NTP, DNS, etc), but not too much value compared with official one provided by OS
- manual upgrades done once a month, others have bunch of bash scripts to do it automatically
- issues with old images: used by existing VMs, marked as non-public, they cannot be rebuilt any more (Ocata will solve this with flag - community), they have a lot of old official images lingering around
- issue with Ubuntu - Canonical sued commercial cloud provider for publishing their own Ubuntu images without consulting Canonical, thus all cloud providers providing their own Ubuntu images could be sued

Possible joint work:

- share the process of building/maintaining/testing images.

Cloudification of a Grid cluster

Speaker: Francisco Bernabé, SURFsara

Part of Life Science Cloud project.

OpenStack administration approaches:

- OpenStack SIG - SURFsara organized different staff/teams to follow individual components
- some follow OpenStack mailing lists - time consuming but keeps you updated better than documentation

Installation:

- Packstack/Devstack
- TripleO
- Manual operations are implemented with Ansible - available public playbooks where too complicated so they developed their own playbooks
- Still need to test how will the upgrade work

OpenStack separation

- VMs per component
- container per component

Upgrade procedure

- keystone, glance, nova, cinder, ..., horizon the last
- recommended to do it separated, leave keystone to run for a while
- there is an official upgrade order
- DON'T touch the database 😀

Delegated administration, reporting/show-back to institutions

Speaker: Simon Leinen, SWITCH

SWITCH cloud is missing proper institute level IaaS, central operators still have to grant access to users manually. They don't enable access to everyone authenticated. They are using some kind of custom developed vouchers which still require action by operator.

The main problem is that federated AAI doesn't cater for concept of projects in OpenStack. EGI is better in this area with VOs.

Group management is not solved by anyone in the room it seems.

No-one is using Ceilometer because it doesn't provide useful information and creates load on the system. Everyone is using their own custom solutions. People are using it for VM metering, switch off all other functions and keep data for limited period.

Billing component CloudKitty - https://wiki.openstack.org/wiki/CloudKitty.

SLA monitoring - only covering network backbone and AAI system.

Round table discussion

What kind of cloud services are being requested by institutes and are NRENs focusing on SaaS? (Tiziana)

- Nordunet - institutes request both IaaS and SaaS, but NREN has too limited resources to provide proper SaaS
- LToS service in EGI is successful because it is being operated by scientific community and not EGI or NGIs
- MAAS+JuJu for building services (GARR)

Kubernetes on OpenStack

Speaker: Saverio Proto, SWITCH

Practical demo of Kubernetes on OpenStack cloud.

Demonstration with 3 nodes nginx cluster with loadbalancer.

Persistent storage for MySQL database, Cinder + Ceph in background. Storage class (storage.k8s.io) can automatically create volumes in Cinder.

Glenna2 metaorchestration (ubernetes)

Speaker: Gurvinder Singh, UNINETT

Federated Kubernetes federation spreading over distributed clouds

- https://wiki.neic.no/wiki/Glenna2

- https://kubernetes.io/docs/concepts/cluster-administration/federation/
- GUI, CLI, API interfaces

Federated AAI (edugain) authentication

Advantage is using common Kubernetes API for accessing different clouds (institute/national IaaS, public clouds)

End User Portal

- AppStore
- applications maintained by researchers, university staff, NRENs, commercial providers
- first prototype was developed as part of project, now moving to pilot phase, not ready yet but, will be by the mid 2018. when the project ends
- all the code is in git at some place & available

Projects as groups of users which has quota, they have a pilot service for this.

Ansible

Speaker: Francisco Bernabé, SURFsara

Not popular on large clusters (1000s of nodes) because of SSH connections.

People have different setups:

- use Ansible for provisioning and cfengine is used later on for keeping the system consistent.
- use Puppet for most of things, Ansible just for single shot things or upgrades
- use Ansible to run Puppet agent, keep Puppet service off
- no-one uses Mcollective seriously for upgrades and/or orchestration.

Bare metal/VM provisioning:

- Foreman used for provisioning bare metal machines & VMs
- SURFsara uses Cobbler but looking at other solutions
- Spacewalk, some bad opinions about it 🙂
- Ansible for managing kickstarts.

OpenStack automation with MAAS and JuJu

Speaker: Fulvio Galeazzi, GARR

GARR infrastructure:

- 8500 CPU cores
- 10 PB of storage
- 40Gb/s network between sites
- Hardware: Dell Blades M1000e + Dell PowerVault MD3860f FC
- OpenStack & Ceph
- 3 sites from GARR and 2 external sites joining

MaaS (Metal as a Service) module for provisioning machines

JuJu

- configuring, managing, deploying and scaling workloads
- 200 publicly available charms (service description)
- GUI, CLI, API interfaces
- Ceph servers are not managed by JuJu but with Ansible

New sites can join the federation and contribute; procedure is to first join DMZ cluster where validation is performed.

Virtual Data Center

- cloud admins create projects with vdc admins and resource quota
- vdc admins manage individual users and child projects
- tweaked policy.json

Discussion about region, cells, availability zones

- SWITCH doesn't use cells and availability zones; cells are pushed by CERN

Comparison between Ansible & JuJu - Ansible simple & sequential, JuJu powerful...

Brief overview on the EOSC pilot

Speaker: Peter Szegedi, Christos Kanellopoulos, GÉANT

Trust, cost/benefits, representativity, open access.

EOSC - break silos such as EGI, EUDAT and migrate to common architecture EOSC through EOSC-hub project.

Swift cluster at SURFsara

Speaker: Ron Trompert, SURFsara

Eventually consistent - useful for unstructured data, not transactional data.

Very good for distributed setup.

SURFdrive, Nextcloud - Data Sharing as a Service.

Hardware setup is ~14 drives per machine.

Using 3 replicas, move to erasure coding in future as it is now supported over multiple regions.

Ceph in the GRNET cloud stack

Speaker: Nikos Kormpakis, GRNET

4 Ceph clusters - 2 prod, 2 testing

everything dualstack with public addresses

Use cases:

- block devices for VMs
- Pithos - OwnCloud-like storage service

rd0 cluster

- HP ProLiant DL380 DL with HP DS2600
- using RAID1 because they didn't trust Ceph - low performance, 50% space loss
- issue with Broadcom NICs flapping - at some point NIC would completely stop working
- replica=2
- had major, multi-day outage in 2016, never identified root cause

rd1 cluster

- Lenovo ThinkServer RD550 with SA120 arrays
- large number of threads with librbd, testing async messenger
- Jewel - problem with systemd/udev scripts

archipelago

- set of tools for managing volumes independently of storage backend
- issues with shm can cause data corruption
- Synnefo has hardcoded dependency
- starting from 2017. migrating to rbd

krbd vs librbd

- decided to use librbd - better performance, krbd uses page cache which caused crashes, admin socket available

Management with Puppet and Python Fabric script

Monitoring with icinga/checkmk (status), munin (debug), collectd/graphite (ceph), prometheus (network/disk), ELK

Recommendations: never use replica=2, keep up2date

CSC Ceph monitoring

Speaker: Pietari Hyvärinen, CSC

2 clusters Espoo (1PB+), Kajaani (3PB+)

Ceph monitor node→Collectd+Graphite+Grafana.

Lots of Grafana art for different situations - rebalancing, adding SSDs, outages during admin travel, restarting OSDs helps with lowering latency.

Useful links for setting up monitoring on slides.

Storage & data management

Speaker: Maciej Brzezniak, PSNC

Questions: Price/TB, Price/Performance, Price vs reliability

Server architectures

- 12HDD/1U - highest energy cost
- 36HDD/4U
- 72HDD/4U - lowest energy cost

Planning to test high density 4U server - QuantaPlex T21P-4U, 8 TB HDD, 35 HDD per board, 10 servers per rack, 6 PB in rack

Long term plan

- Main storage PSNC - 20 servers, 12 PB
- HPC centers - geo-replication, each with 10 servers with 4 PB
- local buffers

ScaleIO

- EMC product
- plan to use it for block I/O for high IOPS volumes ("golden" volumes)
- latency issues with high IOPS volumes with Ceph, heavy server load
- ScaleIO has less layers than Ceph
- description of ScaleIO deployment models
- default replication 2, can be increased
- plan to test alternative solution - Huawei's Fusion Storage

ScaleIO at RedIRIS

Speaker: Antonio Fuentes Bermejo, RedIRIS

Two clusters

- Madrid & Sevilla
- HP ProLiant SL4540 Gen 8 - "cheap" servers, ScaleIO handles HW problems well
- 10 Gb/s

Administration of ScaleIO with CLI, Dashboard is only for monitoring.

Issues

- no NFS support
- requires specific driver which is not available for all platforms (e.g. OpenStack, VMWare, RedHat are supported)

Modern data center (network) fabrics, L3/whitebox, SDN solutions

Speaker: Saverio Proto, SWITCH

Description of network architecture of SWITCHengines.

Quanta switch T5032-LY6 BMS x86

OS Cumulus Linux

- ONIE (Open Network Install Environment) Bootloader
- Foreman proxy - used for provisioning bare metal boxes, required additional Debian repositories on OS

- running Nagios for monitoring
- major upgrade required full reinstall

Managing switches with Puppet and Cumulus Linux community is mainly using Ansible.

AARNET uses the same solution, but vanilla, not so many extensions.

OpenStack vs OpenNebula

Discussion about user interfaces.

SURFsara - OpenNebula user interface demo, Horizon is used by admins & not exposed to users.

SWITCH - Django extension (quickstart) on top of Horizon which simplifies creation of VMs and network, developed by external company exclusively for SWITCH, lot of work required during upgrades (might be cheaper to open source the solution).

Research workflows

Document describing future work/taskforces needed, t-shirt offered as carrot.