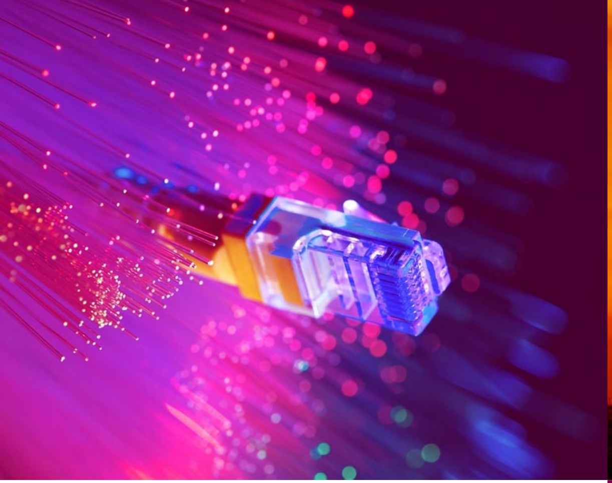
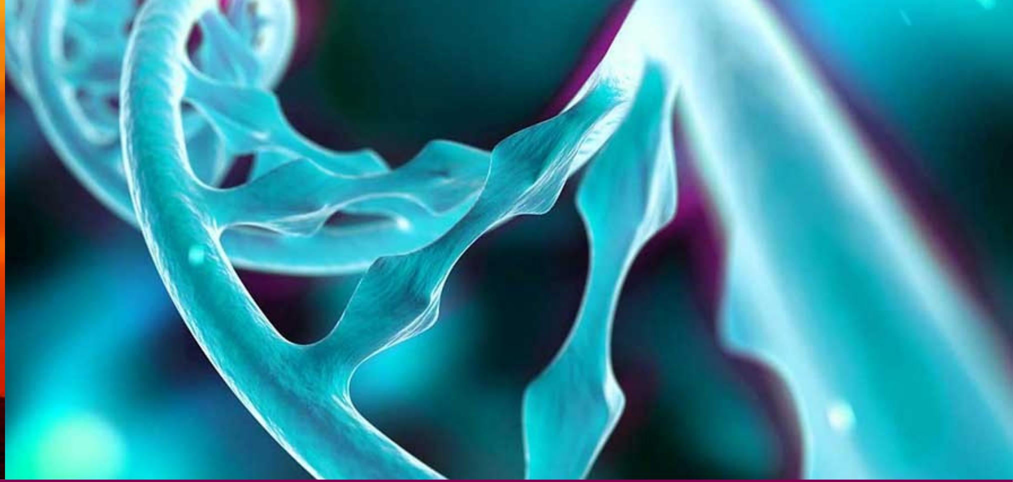




CSC

CSC Cloud update

SIG-CISS 13.3.2024 – Kalle Happonen

CSC – Finnish research, education, culture and public administration ICT knowledge center

CSC's Clouds

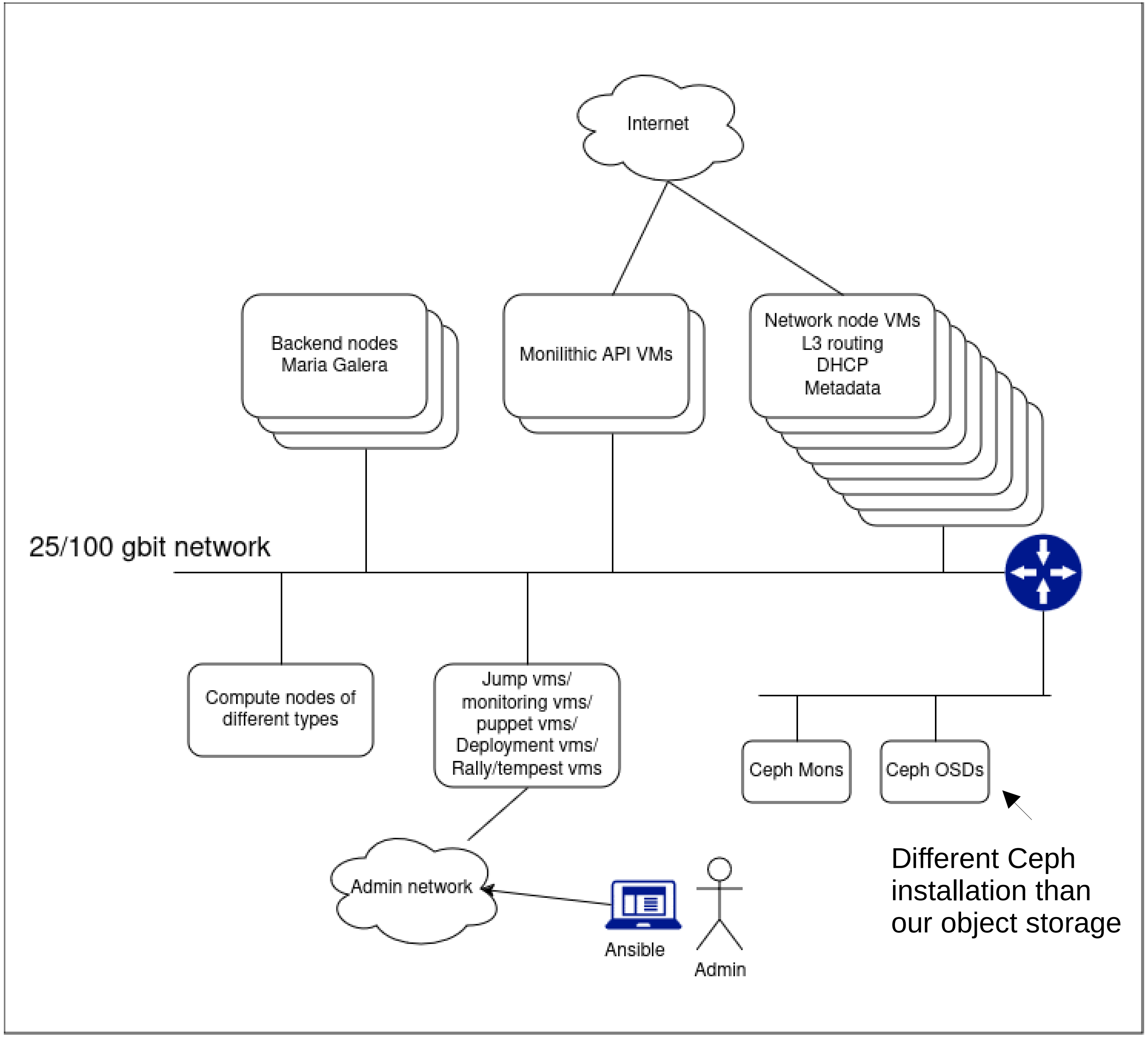
- 10th Anniversary for Production Cloud at CSC!
- 2 OpenStack services
  - Community cloud
  - Sensitive data cloud
- OKD/OpenShift cloud
- Object storage
- DBaaS
- (other cloud services also available but with partly different user base)

Cloud User base

- Mostly R&E (volume wise)
  - Ministry of Education pays for use – Free of charge for end users
- Internal use
  - These clouds are used for building our own services
- Capacity sales
  - These are also sold externally directly or via projects

OpenStacks

- OpenStacks in production since ‘14
- Tech:
  - ~600 compute nodes (Generic nodes, I/O intensive nodes, GPUs, HPC nodes)
  - ~7 PiB usable storage (Ceph)
  - ~1000 customers
- Catching up on updates
  - Now on Stein
  - Target Victoria in ‘24
- Good ol’
  - Ansible + Puppet + Monolithic API nodes (for now)
- Main problem: Paying back tech debt takes a while



Internet

Backend nodes
Maria Galera

Monilithic API VMs

Network node VMs
L3 routing
DHCP
Metadata

25/100 gbit network

Compute nodes of
different types

Jump vms/
monitoring vms/
puppet vms/
Deployment vms/
Rally/tempest vms

Ceph Mons

Ceph OSDs

Admin network

Ansible

Admin

Different Ceph
installation than
our object storage

OKD/OpenShift/K8S

- Runs on OpenStack
- Phasing out our OKD3 offering:
  - ~30 compute nodes (~2,4k vcores, 24 TB memory)
  - ~300+ customer projects, 1000+ namespaces
  - local image registry used size: 2.37 TB with Object Count: 657186.
- Just releasing our OKD4 offering
  - Main issue with RWX storage

New OKD4 load testing

- Scale was a problem with our OKD3 installation
- Pre-launch load testing on our scaled up OKD4 QA cluster
  - workload to simulate memory-intensive deployments
  - workload to simulate cpu-intensive deployments
  - workload to simulate the storage/volume-based pods.
  - a combination of the above
  - workloads only for the storage performance analysis
- Tried to do “wost-case” testing. Automated using ansible

Scale test findings

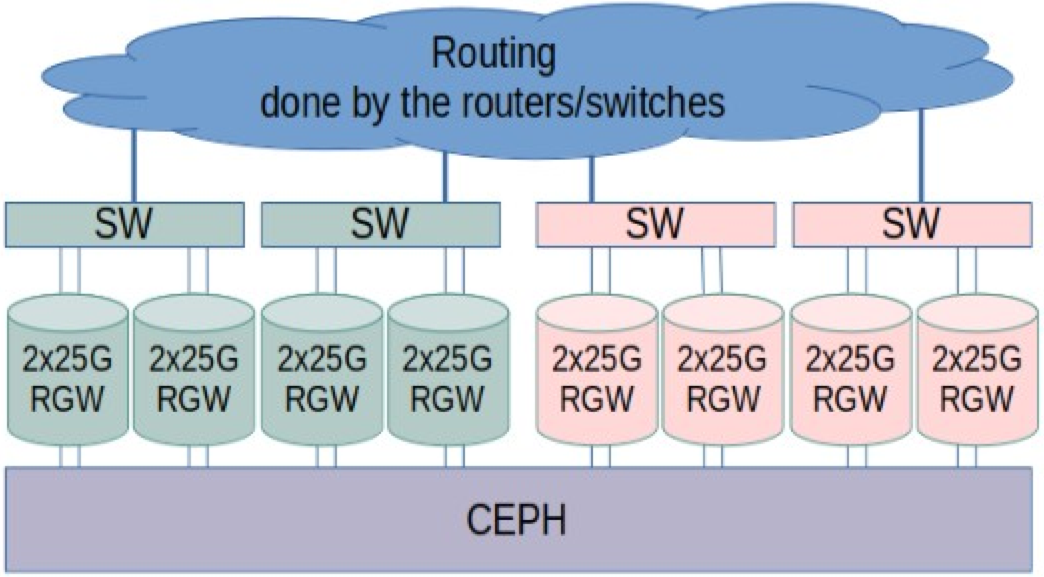
- API response time increased slightly – but not noticeably
- Application deployment worked normally all the time
- Using all of the available storage bandwidth mostly worked, but was noticeable on OKD and OpenStack.
  - OpenStack comment: possibly leaf switch uplinks got congested
- Limited amount of volumes can be attached to the OpenStack VMS
- CPU throttling was seen as expected to ensure fair use of resources
- Memory allocation seemed to work as expected

Object Storage

- Runs on Ceph
- ~17 PiB of usable storage
- Old version – Nautilus – upgrades coming soon
- Main data staging platform for R&E @ CSC
- Also widely used by other services at CSC for data storage

Allas

network architecture change

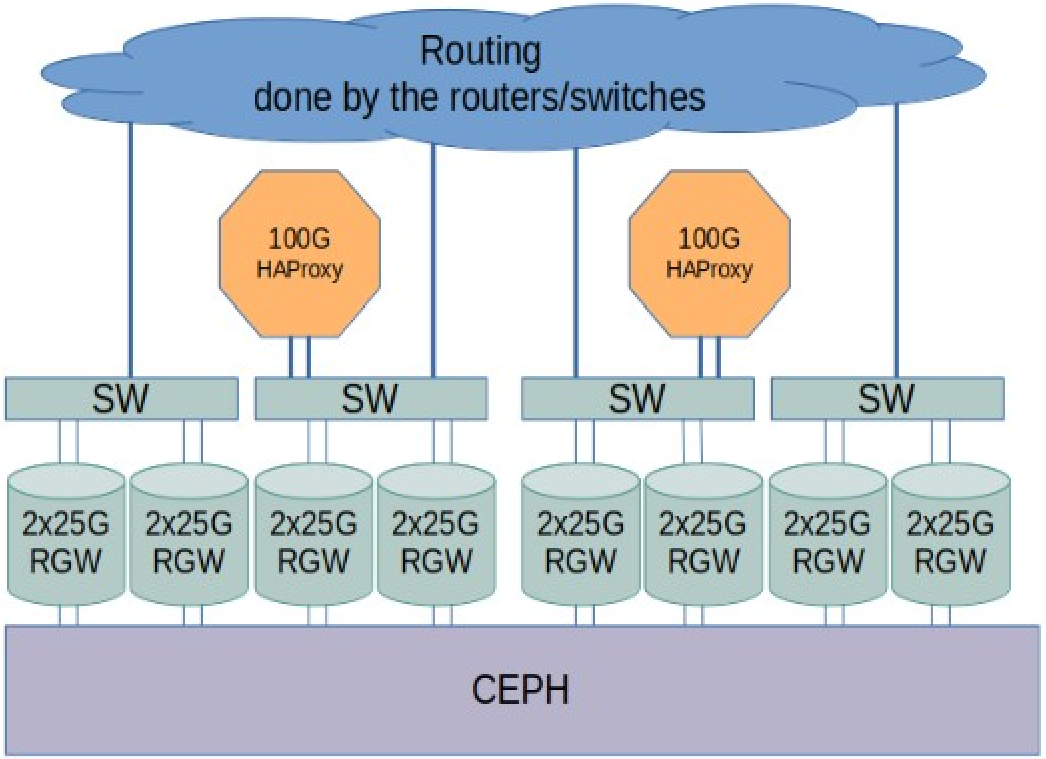


Current architecture

- Ceph as storage
- Each RGW node advertises
  one of the two service addresses
- Switches do routing and load balancing through ECMP
- 25G links all around

Allas

network architecture change

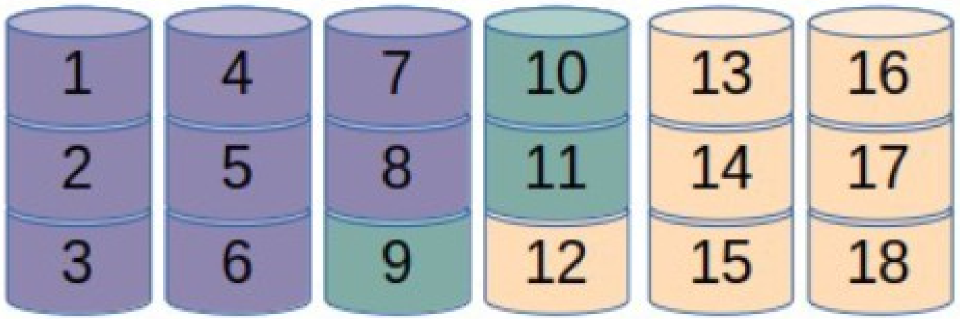


Future architecture

- Ceph as storage
- Both HAProxy nodes advertise both service addresses with inverted pairs of weights, switches route traffic through lowest weight
- HAProxies do the load balancing, all RGWs are in both HAP backend pools
- 100G links in and out of HAProxies
  25G links everywhere else

Allas

virtual rack split



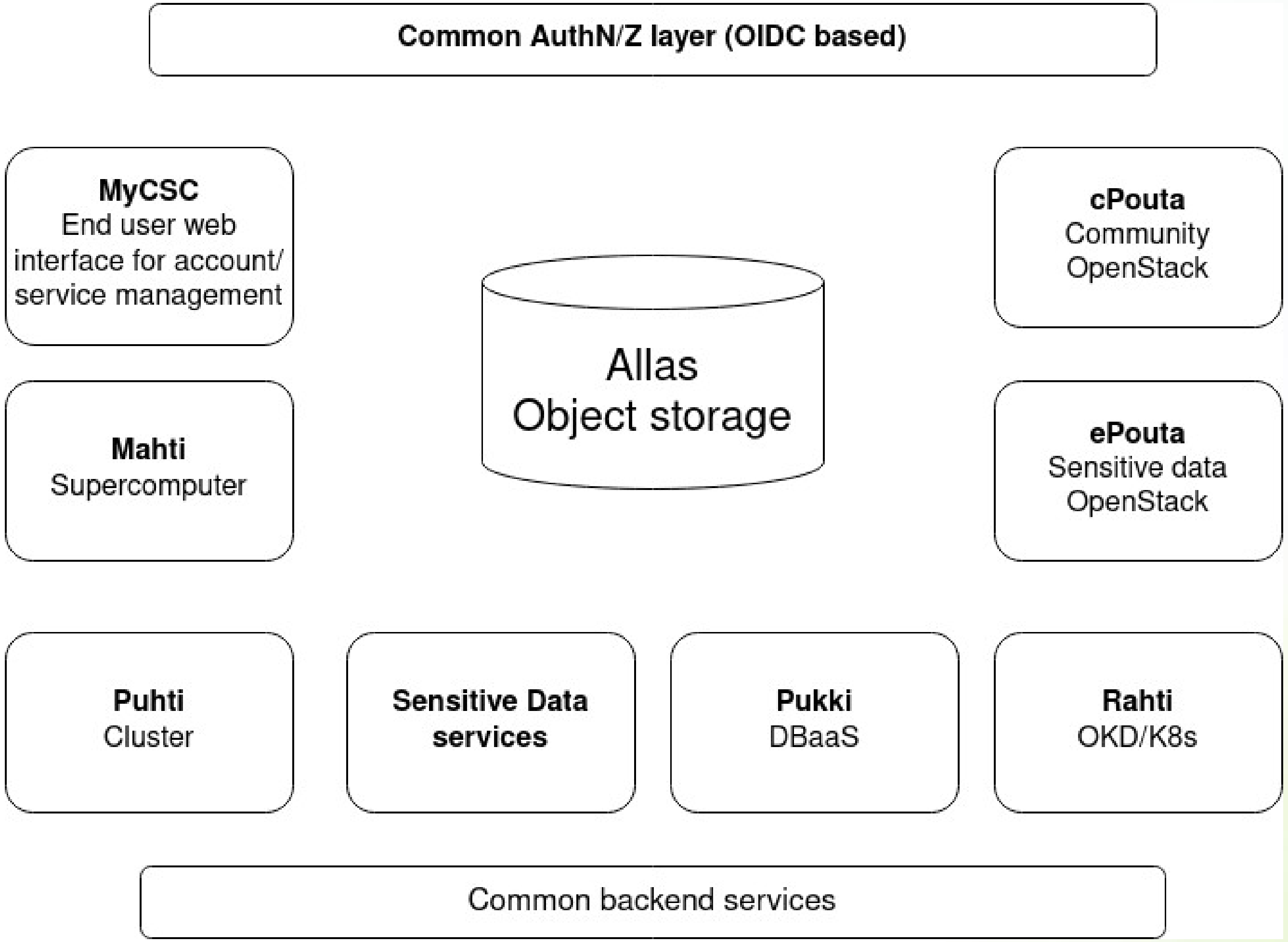
Split each rack to three virtual racks:

- Allas uses a 8+3 Erasure Coding scheme
  - 4 containers lost -> pg unavailable
- Now a CRUSH rule to spread among 11 different racks can be satisfied

DBaaS

- Just being released
- Based on OpenStack Trove – but not part of our other OpenStack platfoms
- However – runs on top of our OpenStack
  - Some internal code hacks to make this work – we should speak about this sometime in a conference
- What it offers
  - API + web interface to manage databases
  - Postgres supported – no HA capabilities yet
  - Easy lifecycle management of databases
  - Automatic daily backups

Tying it together (*)



Common AuthN/Z layer (OIDC based)

MyCSC
End user web interface for account/ service management

Mahti
Supercomputer

Allas
Object storage

cPouta
Community OpenStack

ePouta
Sensitive data OpenStack

Puhti
Cluster

Sensitive Data services

Pukki
DBaaS

Rahti
OKD/K8s

Common backend services

(*) This picture is incomplete, not all services shown

CSC

Click icon to add picture

https://www.facebook.com/CSCfi

https://twitter.com/CSCfi

https://www.youtube.com/c/CSCfi

https://www.linkedin.com/company/csc---it-center-for-science

Kuvat CSC:n arkisto ja Thinkstock