

SWITCH cloud - S3

Version 01, 13. März 2024

Agenda

1. Introduction

2. Which S3 storage to use?

3. Not only Ceph

4. Where are we now?

5. Conclusion

Introduction

Switch

The Swiss NREN – 35 years.

Providing many services to our community.



Switch Cloud

The cloud built for the community.

"SWITCHEngines2":

- Openstack
- Block storage
- S3 storage
- Networking services



CURE – storage

Responsibilities:

- S3 clusters (~25PB raw)
- Block storage for Engines1
- Future S3 clusters



Which S3 to use?

- Workloads: a bit of everything
 - Backups: millions of small objects
 - Data science: very large files
 - Everything in the middle
- Our S3 clusters:
 - 2 * Ceph clusters
 - 2 * «off the shelf» S3 clusters
- Option 1: off the shelf S3
 - Bad experience => 40% of our total S3 capacity, 80% of our pain
 - Always significantly more expensive
- Option 2: opensource based
 - Ceph does the trick. Services on top are very limited -> gaps to fill
 - MinIO gateway not an option

Ref: talk of INFN/Alessandro Costantini

Switch_



Not only Ceph

Based on past experience:

- Manual maintenance can be very tedious => orchestrator needed
- A lot of configuration on the servers => Ansible
- Test environnements
- Configuration drift



Where are we now? (1/2)

Current status

- Deployment of our latest cluster:
 520 OSD ~ 2.5PB raw done in ~2h
- Ingress service (Haproxy/Keepalived) works
- Deploying/scraping clusters for testing from nothing to Ceph in ~16 mins => super powerful!

/ ubuntu@we:~/we/git/storage/osv3-cm > export CLUSTER_NAME=dev-bla
/ ubuntu@we:~/we/git/storage/osv3-cm > make bootstrap

Surprises

- Lack of IPv6 / dual stack support. In 2024. Really?
- Cephadm IS an orchestrator => need time to learn
- Working with Ansible can be surprisingly frustrating!

🖬 Admin 🖿 Monitoring 📰 Docs 📾 Produits 📾 Learning 🕋 Code 📾 Team 🚵 Tools 📾 Coms 📾 Support															
Switcl	Switch_Cloud @ Default - bernard.landon@switch.ch -														
Project API Access Compute		Pro	Project / Compute / Instances												
		•	Ins	stances											
Overview Instances							Instance ID = •			Filter A Launch Instance			nch Instance	Delete Instances More Actions	
	Images Key Pairs		Displaying 7 items		Image Name	IP Address	Flavor	Key Pair	Status		Availability Zone	Task	Power State	Age	Actions
	Serve	er Groups		dev-bla-lb-2	ubuntu-22.04	dev-bla-private 10.0.1.211, 2001:620:7:e00a:f816:3eff:fea8:254b dev-bla-public 192.168.10.172, 2001:620:7:e007::337	c004r008	dev-bla-ssh-key	Active	nî.	nova	None	Running	30 minutes	Create Snapshot 💌
`	Volumes Network	> >		dev-bla-lb-1	ubuntu-22.04	dev-bla-private 10.0.0.55, 2001:620:7:e00a:f816:3eff:fe7d:a7aa dev-bla-public 192.168.10.190, 2001:620:7:e007::211	c004r008	dev-bla-ssh-key	Active	шî	nova	None	Running	30 minutes	Create Snapshot 💌
Identity		>	D	dev-bla-node-5	ubuntu-22.04	10.0.2.238, 2001:620:7:e00a:f816:3eff:fe6b:934a	c004r008	dev-bla-ssh-key	Active	÷	nova	None	Running	30 minutes	Create Snapshot 👻
				dev-bla-node-4	ubuntu-22.04	10.0.2.98, 2001:620:7:e00a:f816:3eff:fec5:2faf	c004r008	dev-bla-ssh-key	Active	зг ^о	nova	None	Running	30 minutes	Create Snapshot 💌
				dev-bla-node-3	ubuntu-22.04	10.0.2.143, 2001:620:7:e00a:f816:3eff:fe16:7c47	c004r008	dev-bla-ssh-key	Active	mî (nova	None	Running	30 minutes	Create Snapshot
			D	dev-bla-node-2	ubuntu-22.04	10.0.3.209, 2001:620:7:e00a:f816:3eff:fed1:1058	c004r008	dev-bla-ssh-key	Active	۳î	nova	None	Running	30 minutes	Create Snapshot
				dev-bla-node-1	ubuntu-22.04	10.0.3.12, 2001:620:7:e00a:f816:3eff:fef0:ab07	c004r008	dev-bla-ssh-key	Active	mî (nova	None	Running	30 minutes	Create Snapshot
			Disp	laying 7 items											



Switch

Where are we now? (1/2)

Work in progress

- Workflow not yet fully set up
- Network setup not as we'd like. Maybe using a custom container service to deploy BGP?
- Monitoring

Main challenge

• Many workflows not easily supported (groups, ro, etc)



To wrap up

- We have an alpha release!
- We try to patch issues as we go...
- Solid foundations to spawn / maintain more clusters.
- Still a lot of work / a lot to learn !
- Not a promise but we hope to open-source our code! Any interest?

Switch

ngr/ce	phadm: discovei	ry servic	e (port 876	5) fails on	ipv6 only	clusters	#54285
⊱ Merged	adk3798 merged 1 commit into	ceph:main from t	thmour:main 🗘 on Nov	6, 2023			

Ingress service: make HAProxy to listen on IPv4 and IPv6 ... #55883

11 Open thegreenbear wants to merge 1 commit into ceph:main from thegreenbear:ingress-haproxy-bind-ipv4v6 🖓





Bernard Landon

Cloud Engineer CURE Team

bernard.landon@switch.ch EPFL Innovation Park Bâtiment I 1015 Lausanne

Switch

