eNovance
Multi-Cloud Experts

OpenStack and Ceph
The winning pair

OPENSTACK SUMMIT ATLANTA | MAY 2014

WHOAMI

eNovance

Sébastien Han

- Cloud Architect
- Daily job focused on Ceph / OpenStack / Performance
- Blogger

Personal blog: http://www.sebastien-han.fr/blog/

Company blog: http://techs.enovance.com/

Ceph?

Let’s start with the bad news

Let's start with the bad news

eNovance

Once again COW clones didn't make it in time

- `libvirt_image_type=rbd` appeared in Havana
- In Icehouse, COW clones code went through feature freeze but was rejected
- Dmitry Borodaenko's branch contains the COW clones code: https://github.com/angdraug/nova/commits/rbd-ephemeral-clone
- Debian and Ubuntu packages already made available by eNovance in the official Debian mirrors for Sid and Jessie.
- For Wheezy, Precise and Trusty look at:

http://cloud.pkgs.enovance.com/{wheezy,precise,trusty}-icehouse

What’s new?

Icehouse additions

Icehouse additions

eNovance

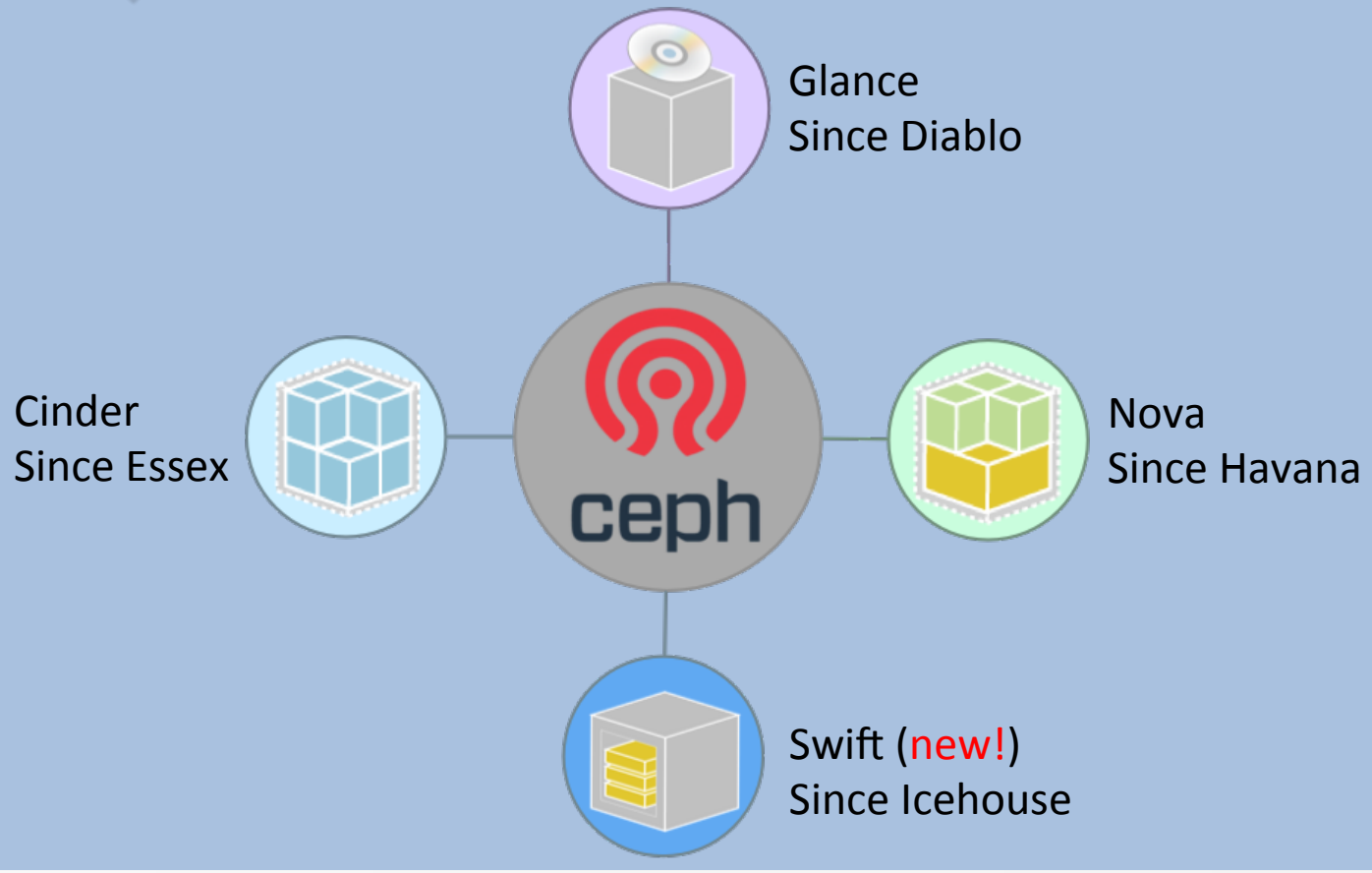
Icehouse is limited in terms of features but…

- Clone non-raw images in Glance RBD backend
  - Ceph doesn't support QCOW2 for hosting virtual machine disk
  - **Always** convert your images into RAW format before uploading them into Glance
  - Nova and Cinder automatically convert non-raw images on the fly
  - Useful when creating a volume from an image or while using Nova ephemeral
- Nova ephemeral backend dedicated pool and user
  - Prior Icehouse we had to use client.admin and it was a huge security hole
  - Fine grained authentication and access control
  - The hypervisor only accesses a specific pool with a right-limited user

Ceph in the OpenStack ecosystem

eNovance

Unify all the things!



- Continuous effort support
- Major feature during each release
- Swift was the missing piece
- Since Icehouse, we closed the loop

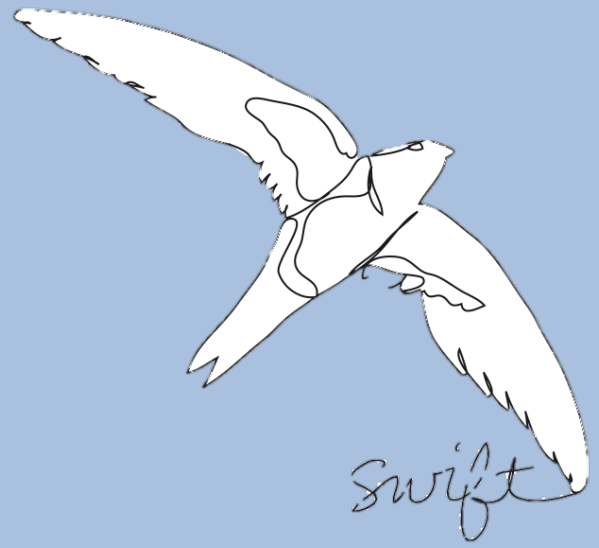
You can do everything with Ceph as a storage backend!

RADOS as a backend for Swift

eNovance

Getting into Swift

- Swift has a multi-backend functionality:
  - Local Storage
  - GlusterFS
  - Ceph RADOS
- You won't find it into Swift core (Swift's policy)
- Just like GlusterFS, you can get it from StackForge

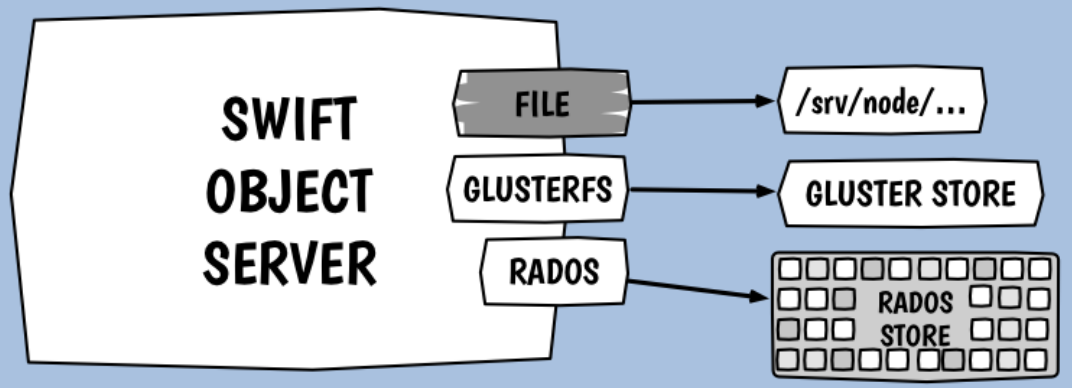


RADOS as a backend for Swift

eNovance

How does it work?

- Keep using the Swift API while taking advantage of Ceph
- API functions and middlewares are still usable
- Replication is handled by Ceph and not by the Swift object-server anymore
- Basically Swift is configured with a single replica



RADOS as a backend for Swift

eNovance

Comparaison table local storage VS Ceph

PROS	CONS
Re-use existing Ceph cluster	You need to know Ceph?
Distribution support and velocity	Performance
Erasure coding	
Atomic object store	
Single storage layer and flexibility with CRUSH	
One technology to maintain	

RADOS as a backend for Swift

eNovance

State of the implementation

- 100% unit tests coverage
- 100% functional tests coverage
- Production ready

Use cases:

1. SwiftCeph cluster where Ceph handles the replication (one location)
2. SwiftCeph cluster where Swift handles the replication (multiple locations)
3. Transition from Swift to Ceph

RADOS as a backend for Swift

eNovance

Little reminder

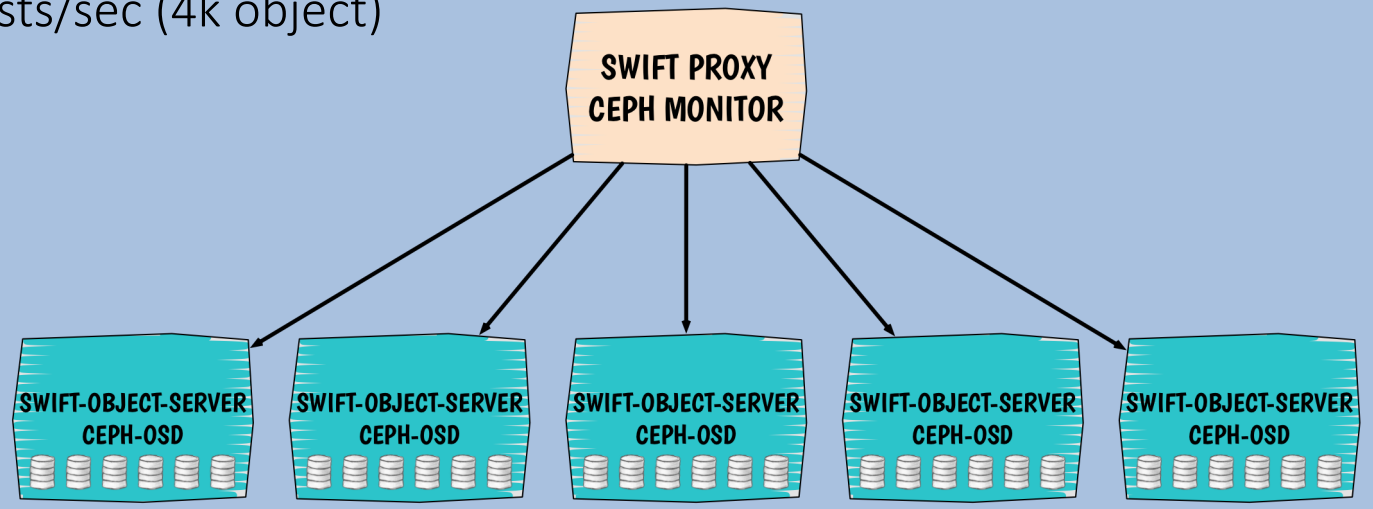
CHARASTERISTIC	SWIFT LOCAL STORAGE	CEPH STANDALONE
Atomic	NO	YES
Write method	Buffered IO	O_DIRECT
Object placement	Proxy	CRUSH
Acknowlegment (for 3 replicas)	Waits for 2 acks	Waits for all the acks

RADOS as a backend for Swift

eNovance

Benchmark platform and swift-proxy as a bottleneck

- The proxy wasn't able to deliver all the platform capability
- Not able to saturate the storage as well
- 400 PUT requests/sec (4k object)
- 500 GET requests/sec (4k object)



- Debian Wheezy
- Kernel 3.12
- Ceph 0.72.2
- 30 OSDs – 10K RPM
- 1 GB LACP network
- Tools: swift-bench
- Replica count: 3
- Swift temp auth
- Concurrency 32
- 10 000 PUTs & GETs

RADOS as a backend for Swift

eNovance

Introducing another benchmark tool

- This test sends requests directly to an object-server without a proxy in-between
- So we used Ceph with **a single** replica

WRITE METHOD	4K IOPS
NATIVE DISK	471
CEPH	294
SWIFT DEFAULT	810
SWIFT O_DIRECT	299

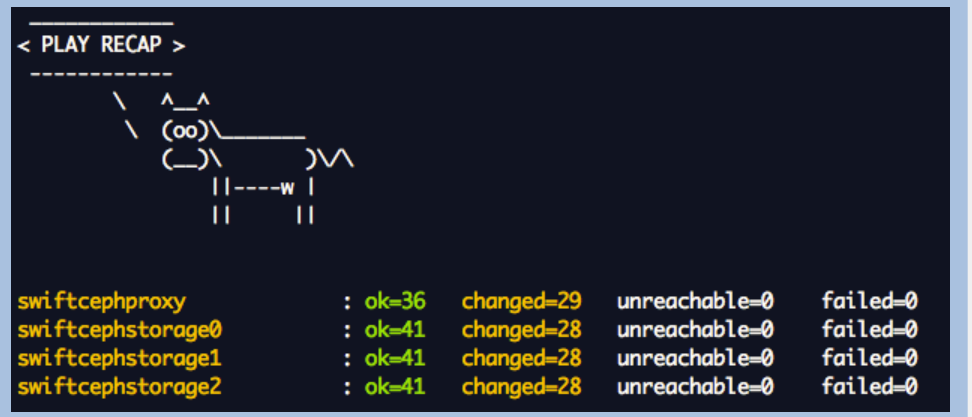
RADOS as a backend for Swift

eNovance

How can I test? Use Ansible and make the cows fly

- Ansible repo here: https://github.com/enovance/swiftceph-ansible
- It deploys:
  - Ceph monitor
  - Ceph OSDs
  - Swift proxy
  - Swift object servers

```
$ vagrant up
```

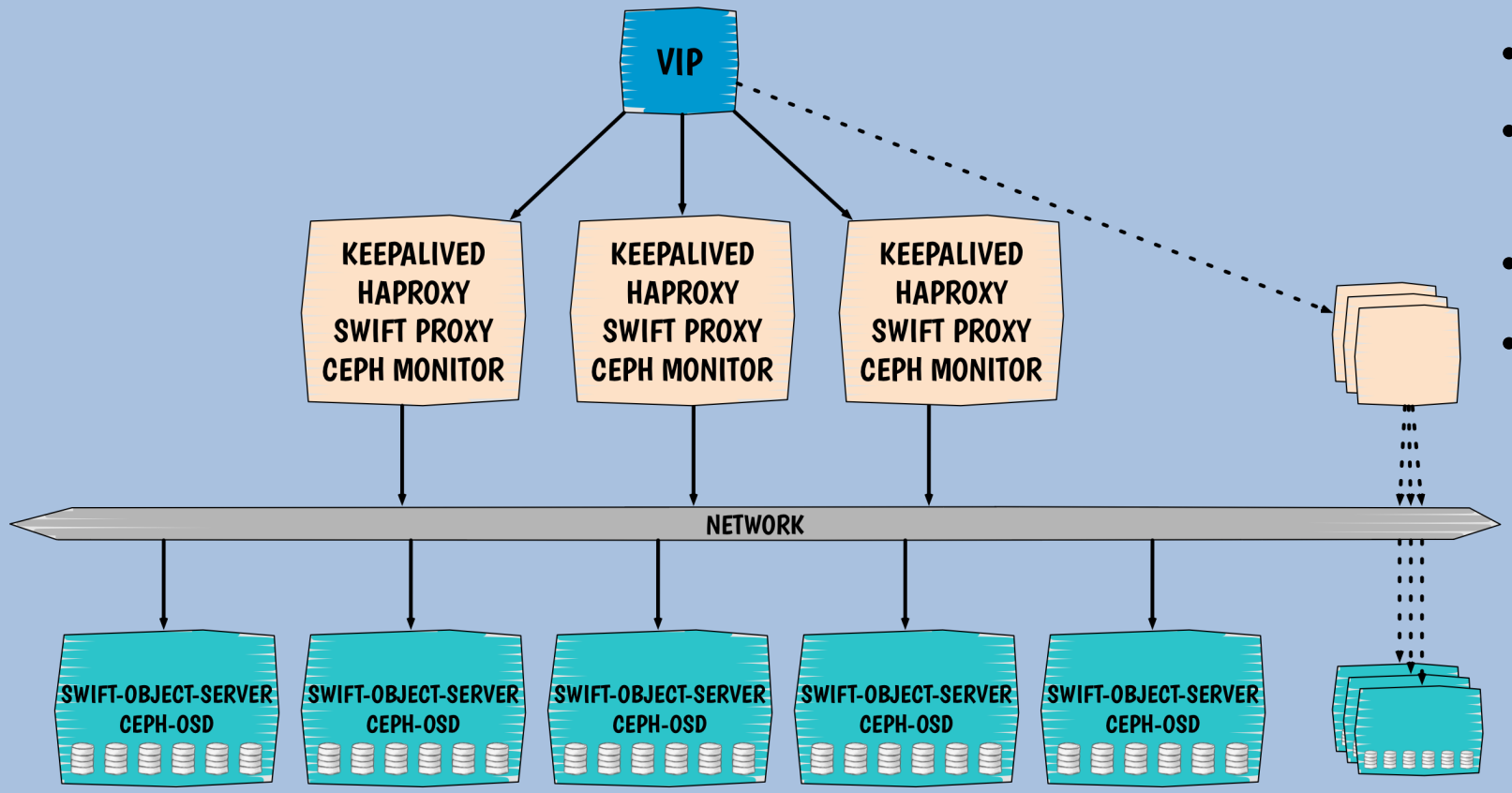


Standalone version of the RADOS code is *almost* available on StackForge in this mean time go to https://github.com/enovance/swift-ceph-backend

RADOS as a backend for Swift

eNovance

Architecture single datacenter

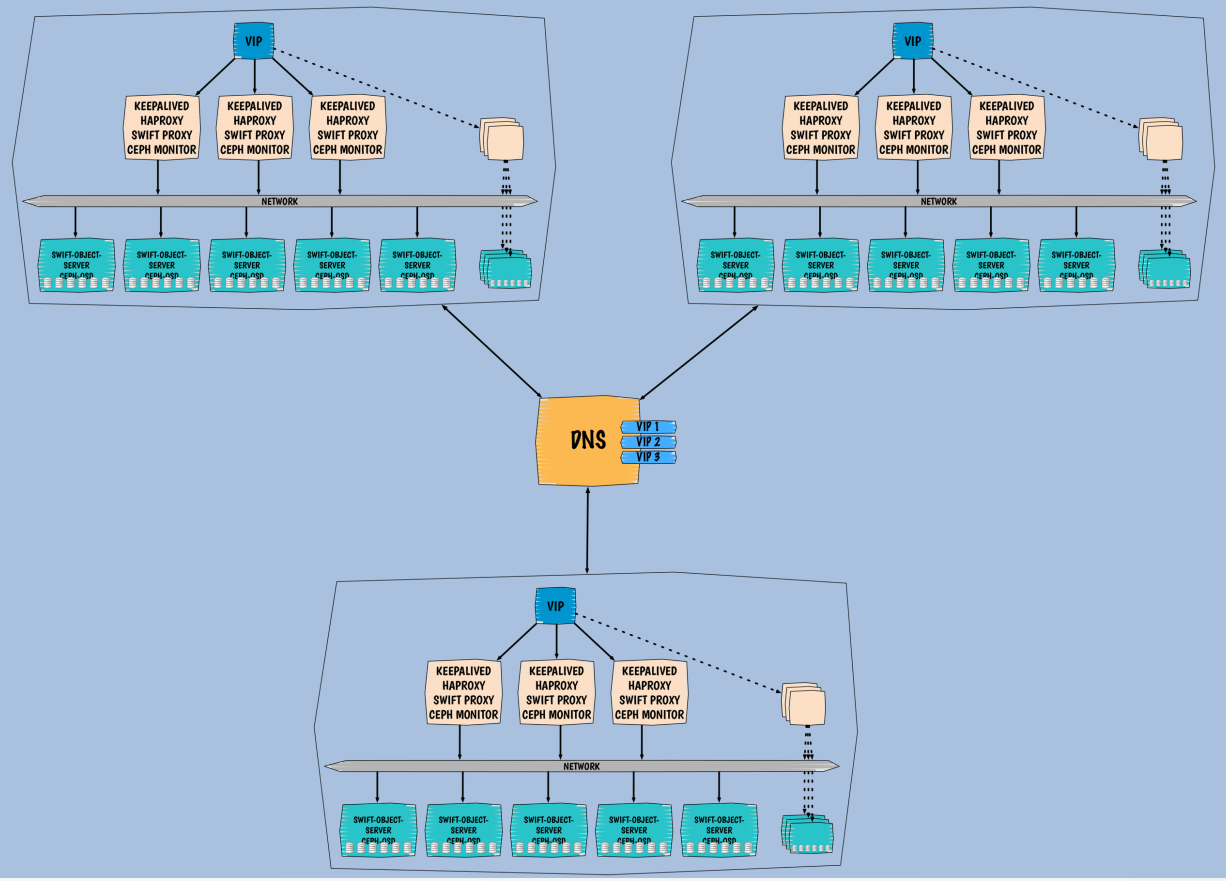


- Keepalived manages a VIP
- HAProxy loadbalances requests among swift-proxies
- Ceph handles the replication
- ceph-osd and object-server collocation (possible local hit)

RADOS as a backend for Swift

eNovance

Architecture multi-datacenter



- DNS magic (geoipdns/bind)
- Keepalived manages a VIP
- HAProxy loadbalances request among swift-proxies
- **Swift handles the replication**
- 3 distinct Ceph clusters
- 1 replica in Ceph stored 3 times by Swift
- Zones and affinities from Swift

RADOS as a backend for Swift

eNovance

Issues and caveats of Swift itself

- Swift Accounts and DBs still need to be replicated
  - /srv/node/sdb1/ needed
  - Setup Rsync
- Patch is under review to support multi-backend store
  - https://review.openstack.org/#/c/47713/
- Eventually Accounts and DBs will live into Ceph

DevStack Ceph

Oh lord!

DevStack Ceph

eNovance

Refactor DevStack and you'll get your patch merged

- Available here: https://review.openstack.org/#/c/65113/
- Ubuntu 14.04 ready
- It configures:
  - Glance
  - Cinder
  - Cinder backup

DevStack refactoring session this Friday at 4:50pm! (B 301)

Roadmap

Juno, here we are

Juno's expectations

eNovance

Let's be realistic

- Get COW clones into stable (Dmitry Borodaenko)
  - Validate features like live-migration and instance evacuation
- Use RBD snapshot instead of qemu-img (Vladik Romanovsky)
  - Efficient since we don't need to snapshot, get a flat file and upload it into Ceph
- DevStack Ceph (Sébastien Han)
  - Ease the adoption for developers
- Continuous integration system (Sébastien Han)
  - Having an infrastructure for testing RBD will help us to get patch easily merged
- Volume migration support with volume retype (Josh Durgin)
  - Move block from Ceph to other backend and the other way around

Merci !