Ceph Storage in OpenStack

Part 2

Jens-Christian Fischer
jens-christian.fischer@switch.ch
@jcfischer
@switchpeta

Ceph

Distributed, redundant storage

Open Source Software, commercial support

http://inktank.com

https://secure.flickr.com/photos/38118051@N04/3853449562

Ceph Design Goals

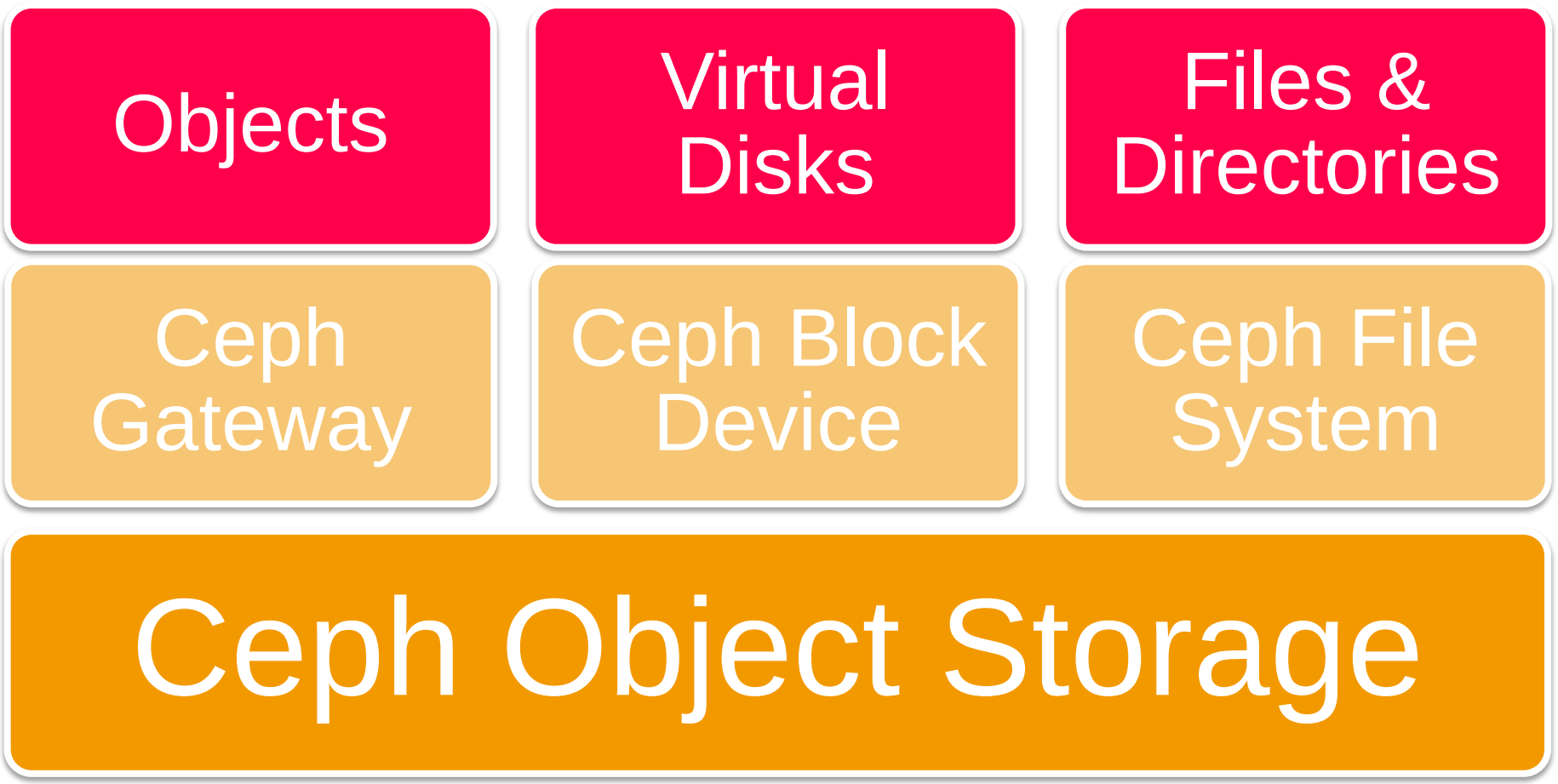
- Every component must scale
- There can be no single point of failure
- Software based, not an appliance
- Open Source
- Run on commodity hardware
- Everything must self-manage wherever possible

http://www.inktank.com/resource/end-of-raid-as-we-know-it-with-ceph-replication/

Different Storage Needs

- Object
  - Archival and backup storage
  - Primary data storage
  - S3 like storage
  - Web services and platforms
  - Application Development
- Block
  - SAN replacement
  - Virtual block devices, VM Images
- File
  - HPC
  - Posix compliant shared file system

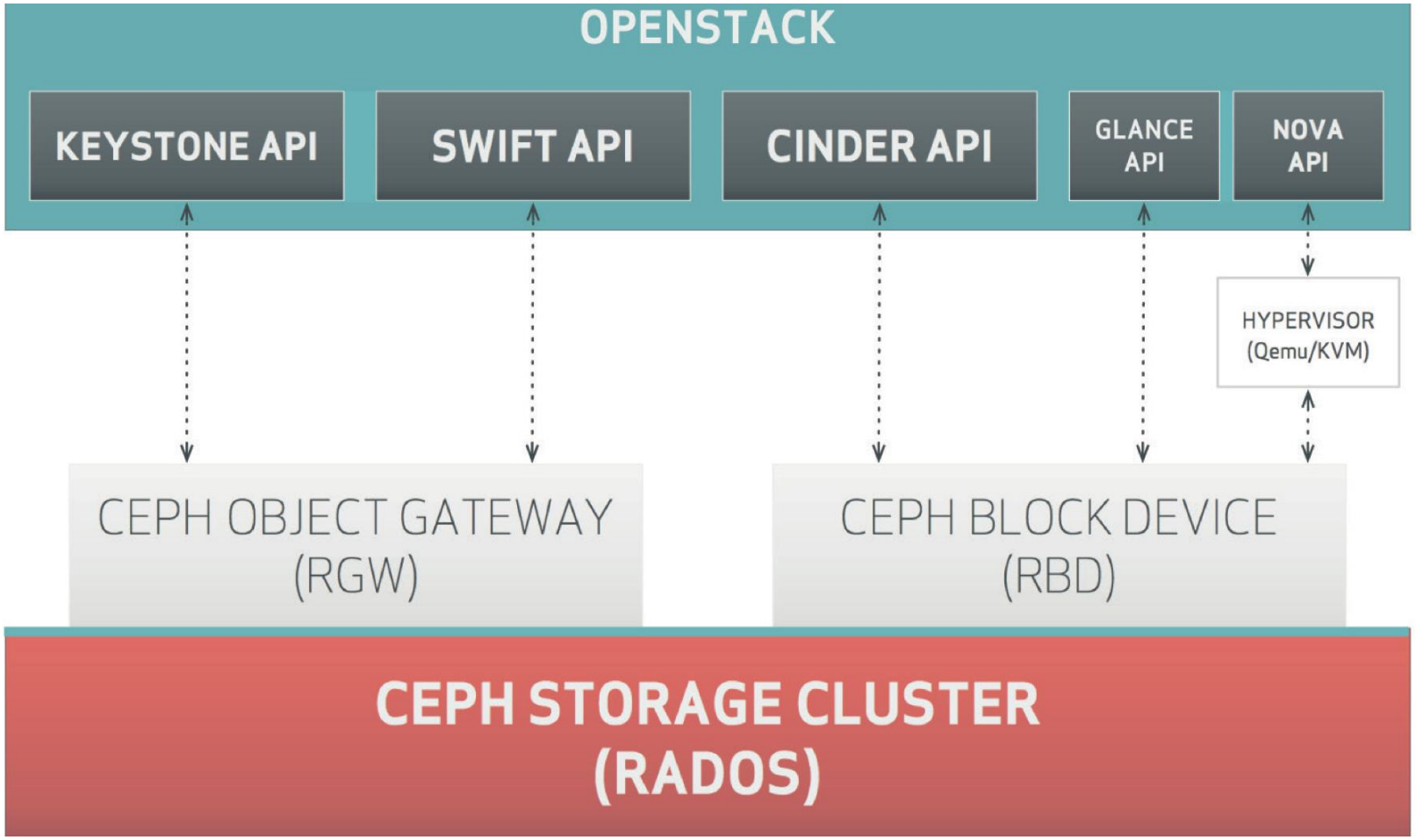
Ceph



Storage in OpenStack

- Glance
  - Image and volume snapshot storage (metadata, uses available storage for actual files)
- Cinder
  - Block Storage that is exposed as volumes to the virtual machines
- Swift
  - Object Storage (think S3)

Ceph as Storage in OpenStack



http://www.inktank.com/resource/complete-openstack-storage/

Ceph @ SWITCH

https://secure.flickr.com/photos/83275239@N00/7122323447

Cloud @ SWITCH

Serving the Swiss university and research community

4 major product releases planned

- “Academic Dropbox” early Q2 2014
- “Academic IaaS” mid 2014
- “Academic Storage as a Service” and
- “Academic Software Store” later in 2014

Built with OpenStack / Ceph

“Current” Preproduction Cluster

- 1 Controller Node
- 5 Compute Nodes (expected 8)
- Total 120 Cores
- 640 GB RAM
- 24 * 3 TB SATA Disks: ~72 TB Raw Storage

- OpenStack Havana Release
- Ceph Dumpling

First Planned Production Cluster

- Havana/Icehouse Release
- Ceph Emperor/Firefly

- 2 separate data centers
- Ceph cluster distributed as well (we'll see how that goes)

- Around 50 Hypervisors with 1000 cores
- Around 2 PB of raw storage

Storage numbers

- Dropbox: 50 – 250’000 users => 50 TB – 2.5 PB
- IaaS: 500 – 1000 VMs => 5 TB – 50 TB
- Storage as a Service: 100 TB – ?? PB

There’s a definitive need for scalable storage

OpenStack & Ceph

- Glance images in Ceph
- Cinder volumes in Ceph
- Ephemeral disks in Ceph

Thanks to the power of Copy on Write

- “Instant VM creation”
- “Instant volume creation”
- “Instant snapshots”

Ceph Support in Havana

- Almost there – basic support for Glance, Cinder, Nova
  - Edit config file, create pool and things work (unless you use CentOS)
- Not optimized: “Instant Copy” is really
  - download from Glance (Ceph) to disk
  - upload from disk to Cinder (Ceph)
- Patches available, active development, should be integrated in Icehouse

Object Storage

Use RadosGW (S3 compatible)

Current use cases:

- A4Mesh: Storage of hydrological scientific data
- SWITCH: Storage and Streaming of Video Data

Some weird problems with interruption of large streaming downloads

Shared Storage for VMs

Investigating NFS servers, backed either by RBD (Rados Block Device) or by Cinder Volumes

Not our favorite option, but currently a viable option.

Questions about scalability and performance

Block Devices

Cinder volumes

Boot from Volume

Nicely works for Live Migration

Very fast to spawn new volumes from snapshots

CephFS as shared instance storage

This page i

CephFS for shared file storage

Be careful about Linux kernel versions (3.12 is about right)

Works under light load

Be prepared for surprises

Or wait for another 9 months (according to word from Inktank)

Experience

- Ceph is extremely stable and has been very good to us
- Except for CephFS (which for the time being is being de-emphasized by Inktank)
- Software in rapid development – some functionality “in flux” – difficult to keep up. However: Gone through 2 major Ceph upgrades without downtime
- The Open{Source|Stack} problem: Documentation and experience reports strewn all over the Interwebs (in varying states of being wrong)

Would we do it again?

- Yes!

- Ceph is incredibly stable
  - unless you do stupid things to it
  - or use it in ways the developers tell you not to

- Responsive developers, fast turnaround on features

Nitty gritty

- http://ceph.com/docs/master/rbd/rbd-openstack/
- https://github.com/jdurgin/nova/tree/havana-ephemeral-rbd
- https://review.openstack.org/#/c/56527/
- http://techs.enovance.com/6424/back-from-the-summit-cephopenstack-integration