One for all!

CEPH and Openstack: A Dream Team

Udo Seidel

Agenda

- Openstack
- CEPH Storage
- Dream team: CEPH and Openstack
- Summary

Me :-)

- Teacher of mathematics and physics
- PhD in experimental physics
- Started with Linux in 1996
- Linux/UNIX trainer
- Solution engineer in HPC and CAx environment
- @Amadeus → Head of
  - Linux Strategy
  - Server Automation

My setup :-D

- Raspberry Pi2
- Fedora 21 with custom kernel
- HDMI2VGA
- Mini Bluetooth keyboard
- 10 Ah battery

Openstack

What?

- Infrastructure as a Service (IaaS)
- 'Open source' version of AWS
- New versions every 6 months
  - Current called Juno
  - Next called Kilo
- Managed by Openstack Foundation
- API, API, API!

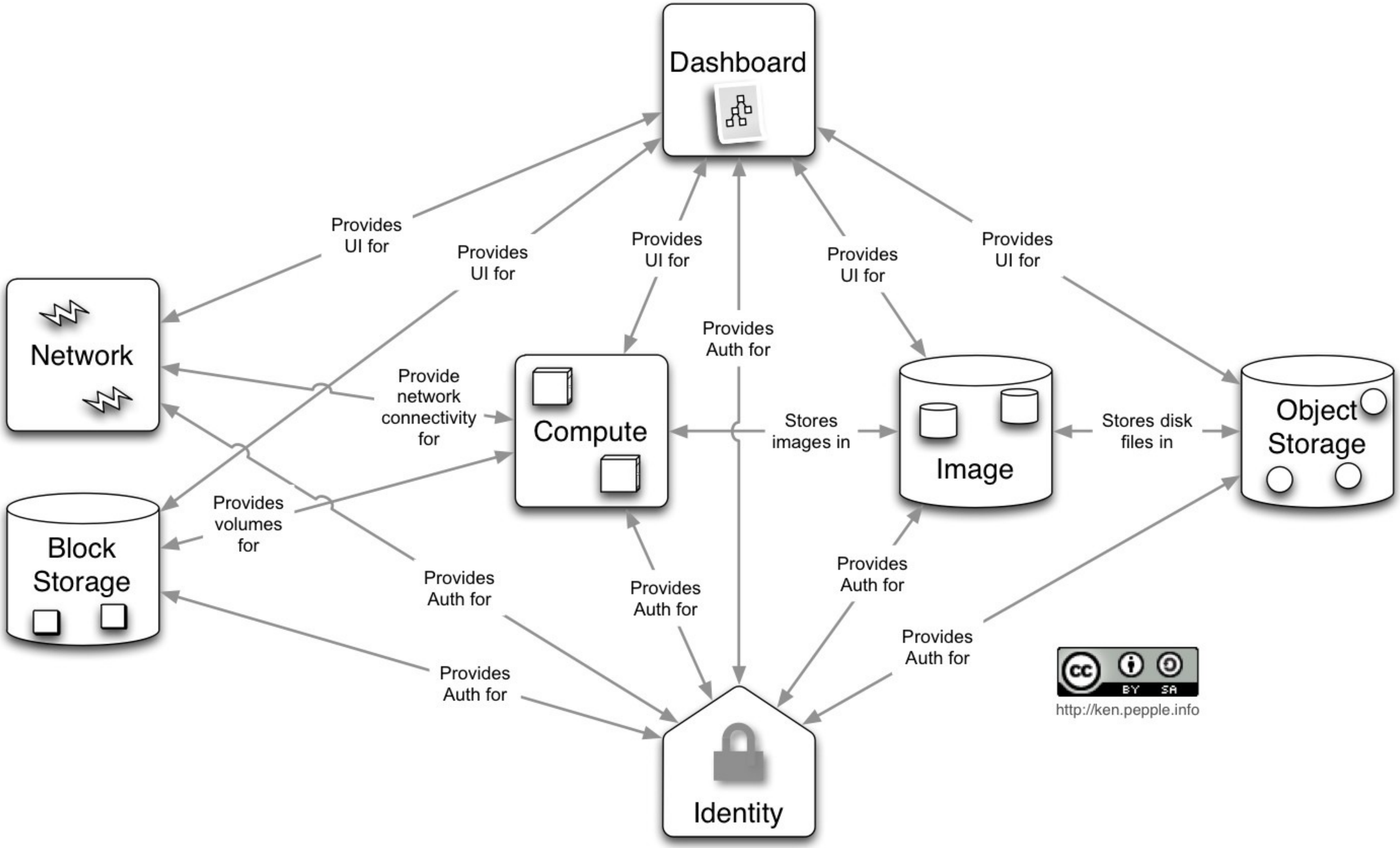
Openstack – High level

Network

Compute

Storage

Openstack architecture



Openstack Components

- Keystone – *identity*
- Glance - *image*
- Nova - *compute*
- Cinder - *block*
- Swift - *object*
- Neutron - *network*
- Horizon - *dashboard*

About Glance

- There since almost the beginning
- Image store
  - Server
  - Disk
- Several formats
- Different storage back-ends available

Behind Default Glance

- File Back-end
- Local or shared file system
- POSIX ?!?
- Scalability
- High availability

About Cinder

- Later than Glance
  - Part of Nova before
  - Separate since Folsom
- Block storage
- Different storage back-ends possible

Behind Default Cinder

- Logical Volume Manager
- 'Glance-like' challenges
  - Scalability
  - High availability

About Swift

- Since the beginning
- Replace Amazon S3
  - cloud storage
  - Scalable
  - Redundant
- Object store

Behind Swift

- RESTful API
- No POSIX like access
- No Block level access

Openstack Storage Questions

- Unification of storage types
- High availability
- Scalability
- Access/APIs
- Vendor (lock-in)

CEPH Storage

CEPH – what?

- Distributed storage system
- Started as part of PhD studies at UCSC
- Public announcement: 2006 at 7th OSDI
- File system: Linux kernel since 2.6.34
- Cephalopods

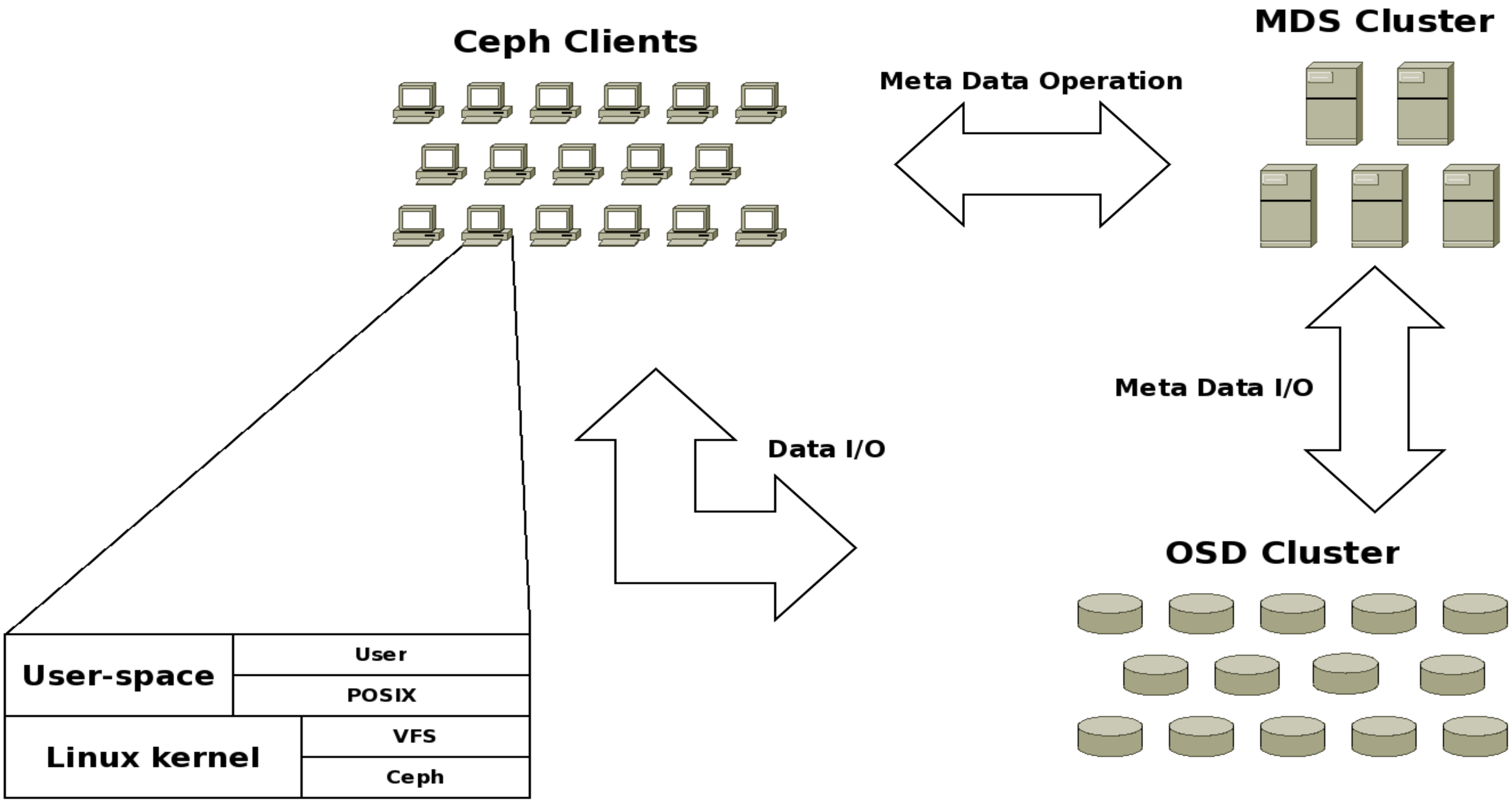
CEPH – Releases

- Like Linux Kernel
  - 'normal'
  - Long Term Support
- LTS
  - Since 2012
  - Firefly → 0.80.x
  - Giant → 0.87.x
  - Hammer → 0.93.x

CEPH – Commercial

- Past: Inktank Inc.
- Acquisition by Red Hat in 2014
- ICE – Inktank CEPH Enterprise
  - Server: RHEL/CentOS, Ubuntu
  - Client:
    - RHEL
    - S3 compatible application
    - ...
- SUSE Storage

CEPH – the full architecture



OSD failure approach

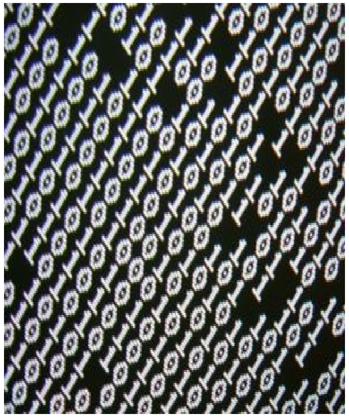
- Failure is normal
- Data distributed and replicated
- Dynamic OSD landscape

Data replication

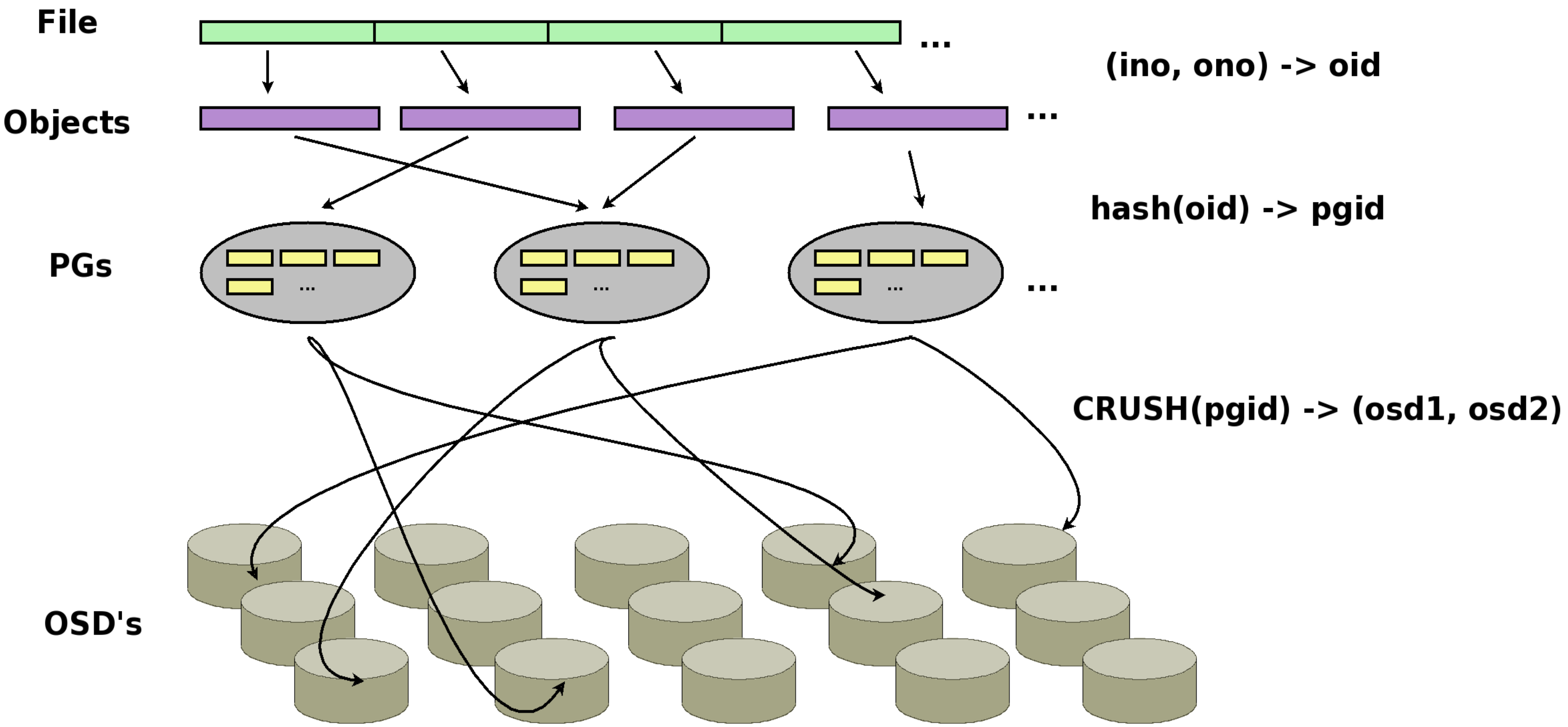
- N-way
  - Placement group
  - Failure domains
- Replication traffic
  - Within OSD network
  - Timing

Data distribution

- File stripped
- File pieces → Object IDs
- Object ID → Placement groups
- Placement groups → list of OSDs

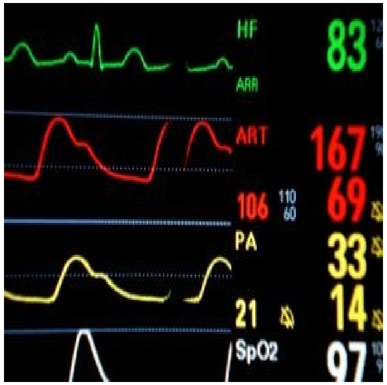


CRUSH


File
...
(ino, ono) -> oid
Objects
...
hash(oid) -> pgid
PGs
...
...
...
...
CRUSH(pgid) -> (osd1, osd2)
OSD's

CEPH cluster monitors

- CEPH components status
- First contact point
- Monitor cluster landscape



CEPH cluster map

- Objects
  - computers and containers
  - ID and weight
- Container → bucket
- Maps physical conditions
- Reflects data rules
- Known by all OSD’s

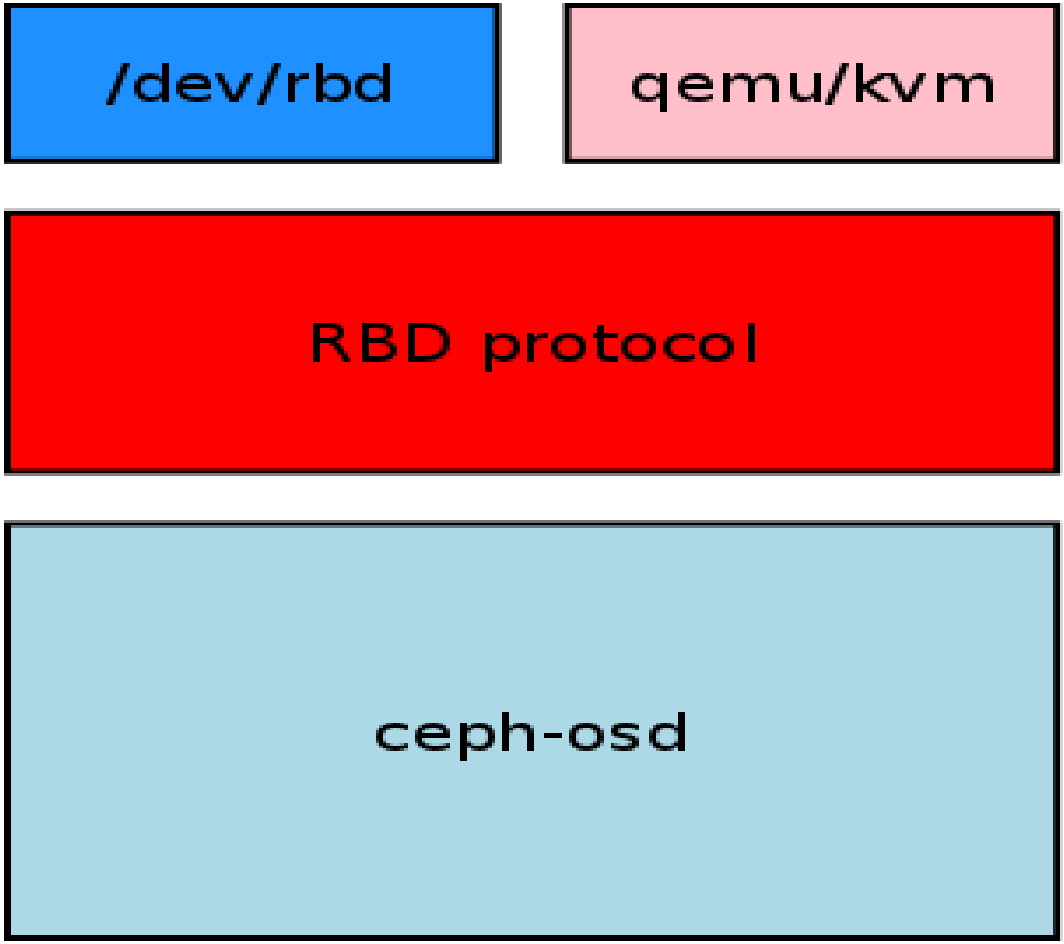
CEPH - RADOS

- Reliable Autonomic Distributed Object Storage
- OSD cluster access
  - Via *librados*
  - C, C++, Java, Python, Ruby, PHP
- ~~POSIX layer~~
- 'Visible' to all CEPH cluster members

CEPH Block Device

- Aka RADOS block device (RBD)
- Upstream since kernel 2.6.37
- RADOS storage exposed via
  - Simple block device
  - Interface library

The RADOS picture



CEPH Object Gateway

- Aka RADOS Gateway (RGW)
- RESTful API
  - Amazon S3
  - SWIFT APIs!!

- Proxy HTTP to RADOS
- Tested with *apache, nginx* and *lighthttpd*

CEPH File System

- Yes ..
- But …
- Skipped here!

CEPH Take Aways

- Scalable
- Flexible configuration
- No SPOF
- Built on commodity hardware
- Different interfaces
  - Language
  - Protocols

Dream Team
CEPH and Openstack

Remember: Openstack Storage

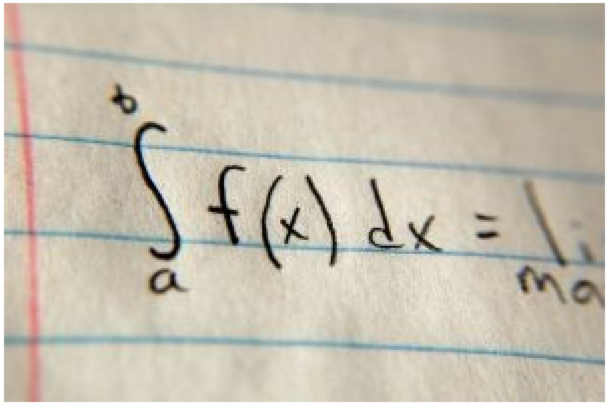
- Unification of storage types
- High availability
- Scalability
- Access/APIs
- Vendor (lock-in)

Why CEPH in the first place?

- One solution for different storage needs
- Full blown storage solution
  - Support
  - Operational model
  - Cloud'ish
- Separation of duties

Integration

- Focus: RADOS/RBD
- Two parts
  - Authentication
  - Technical access
- Both parties must be aware
- Independent for each of the storage components



Authentication

- CEPH part
  - Key rings
  - Configuration
  - For Glance and Cinder
- Openstack part
  - Glance and Cinder (and Nova)
  - Keystone
    - Only for Swift
    - Needs RGW

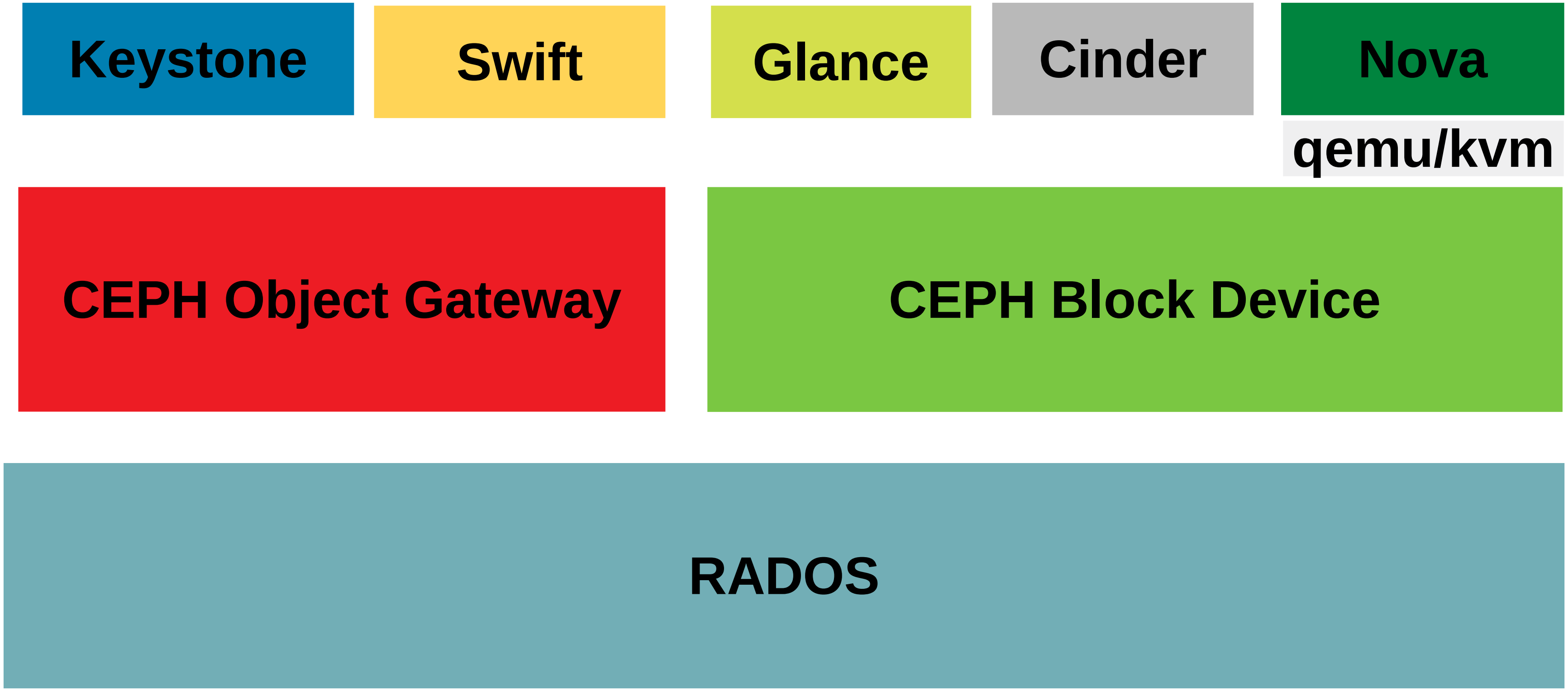
Access to RADOS/RBD I

- Via API/libraries
- ~~CEPHFS~~
- Easy for Glance/Cinder
  - CEPH keyring configuration
  - Update of *ceph.conf*
  - Update of API configuration
    - Cinder
    - Glance

Access to RADOS/RBD II

- Swift → more work
- ~~CEPHFS~~
- CEPH Object Gateway
  - Web server
  - RGW software
  - Keystone certificates
- Keystone authentication
  - Endlist configuration →RGW

Integration the full picture



Integration pitfalls

- CEPH versions not in sync
- Authentication
- CEPH Object Gateway setup
- Openstack version specifics

CEPH Openstack - Commercial

- RHEL Openstack Platform
- SUSE Openstack Cloud
- Mirantis Openstack
- Ubuntu Openstack

Why CEPH - reviewed

- Previous arguments still valid :-)
- High integration
- Modular usage
- No need for POSIX compatible interface
- Works even with other IaaS implementations

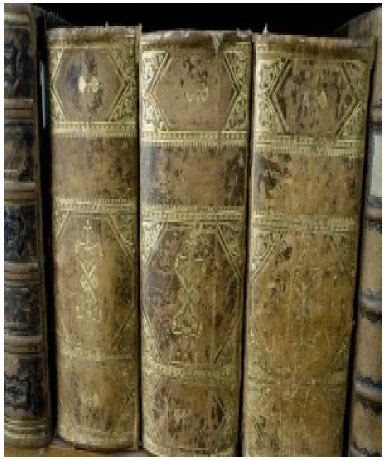
Summary

Take Aways

- Openstack storage challenges
- CEPH
  - Sophisticated storage engine
  - Mature
  - Can be used elsewhere
- CEPH + Openstack = <3

References

- http://ceph.com
- http://www.openstack.org



Thank you!

All for one!

CEPH and Openstack: A Dream Team

Udo Seidel

GUUG FFG 2015