Ceph as (primary) storage for Apache CloudStack

Wido den Hollander <wido@42on.com>

Who am I?

- Wido den Hollander
  - Born (1986) and live in the Netherlands
  - Co-founder and owner of a webhosting company
    - Ceph and later CloudStack were adopted as technologies inside the company
  - Started 42on in September 2012
    - 42on is a professional services company for Ceph and the surrounding eco-system (like CloudStack)
  - Wrote various Ceph/RBD bindings and integrations:
    - PHP extension (phprados)
    - libvirt storage pool support
    - Apache CloudStack integration

Apache CloudStack

- Apache CloudStack is open source software designed to deploy and manage large networks of virtual machines, as a highly available, highly scalable Infrastructure as a Service (IaaS) cloud computing platform.
- Top-level Apache project since March 29th 2013
- Written in Java
- Hypervisor agnostic
  - RBD support only for KVM

Ceph

Ceph is a unified, open source distributed object store

Traditional vs Distributed

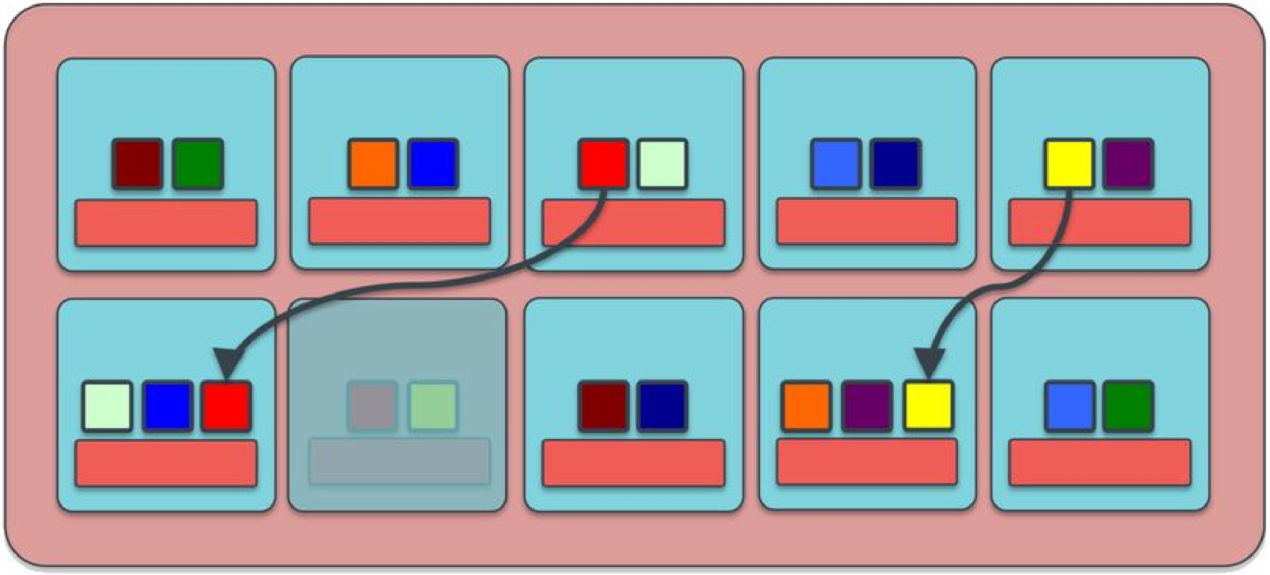
- Traditional storage systems don't scale that well
  - All have their limitations: Number of disks, shelfs, CPUs, network connections, etc
  - Scaling usually meant buying a second system
    - Migrating data requires service windows
- Ceph clusters can grow and shrink without service interruptions
- Ceph runs on commodity hardware
  - Just add more nodes to add capacity
  - Ceph fits in smaller budgets

Hardware failure is the rule

- As systems grow hardware failure becomes more frequent
  - A system with 1.000 nodes will see daily hardware issues
- Commodity hardware is cheaper, but less reliable. Ceph mitigates that.

Auto recovery

- Recovery when a OSD fails
- Data migration when the cluster expands or contracts



Block Devices

- Block devices are devices which move data in the form of blocks.
- Hard drives are block devices
- iSCSI presents SCSI block devices over IP
- Virtual Machines have block devices to boot from and store their data on
    - /dev/sda or /dev/vda is a block device in a virtual Linux machine

RBD: the RADOS Block Device

- Is a Block Device with special capabilities
  - Snapshotting
  - Cloning
- Ceph is a object store
  - Store billions of objects in pools
  - RADOS is the heart of Ceph
- RBD block devices are striped over RADOS objects
  - Default stripe size is 4MB
  - All objects are distributed over all available Object Store Daemons
- RBD is build on top of Ceph's object store and thus leverages from all the features Ceph has
- RBD is a driver inside Qemu/KVM

RBD: Object placement

- Ceph stores replicas of objects
  - The number of replicas can be configured
- With Ceph's 'crushmap' you can store replicas in different racks or on different machines
  - Provides higher availability when racks or machines fail
- Different pools can be created with their own data-placement rules

Storage in CloudStack

- Two types of storage
  - Primary Storage
    - Your instances run on this storage
  - Secondary Storage
    - Used for backup and template storage
- RBD has been implemented as Primary Storage

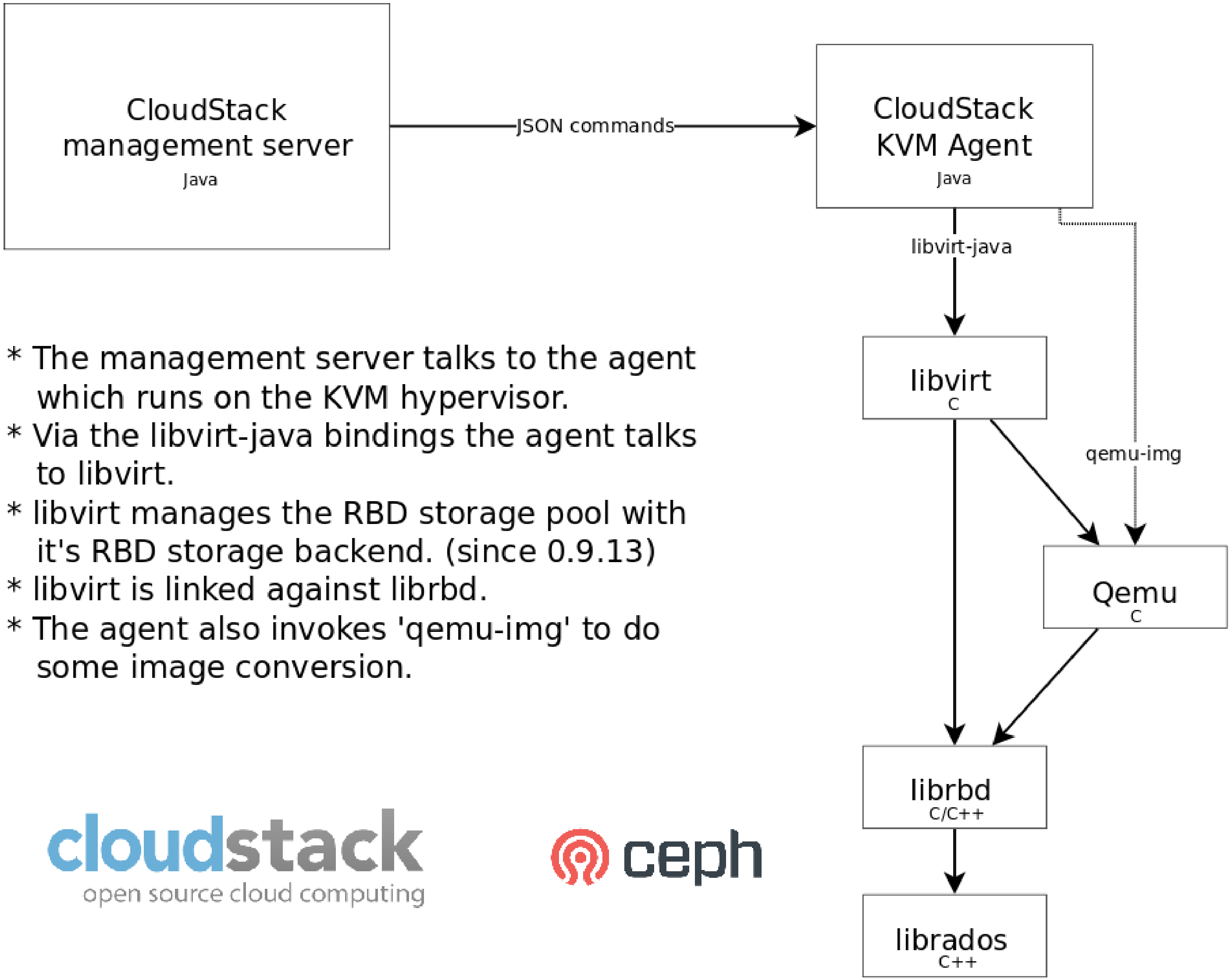
RBD for Primary Storage

- In 4.0 RBD support for Primary Storage for KVM was added
- Live migration is supported
- Ubuntu 12.04 is recommended for the hypervisors
  - Libvirt >= 0.9.13 has to be compiled manually
    - Enable RBD storage pool support
    - Ubuntu 13.04 has everything you need

RBD for Primary Storage

- In 4.0 RBD support for Primary Storage for KVM was added
  - No support for VMware or Xen, no ETA
- Live migration is supported
- No snapshot support
  - Current CloudStack code makes some assumptions which don't work with RBD
- NFS is still required for the System VMs

Primary storage flow (1/2)



- The management server talks to the agent which runs on the KVM hypervisor.
- Via the libvirt-java bindings the agent talks to libvirt.
- libvirt manages the RBD storage pool with it's RBD storage backend. (since 0.9.13)
- libvirt is linked against librbd.
- The agent also invokes 'qemu-img' to do some image conversion.

cloudstack
open source cloud computing

ceph

Primary storage flow (2/2)

- The management server never talks to the Ceph cluster.
- One management server can manage thousands of hypervisors
  - Management server can be clustered
- Multiple Ceph clusters or pools can be added to a CloudStack cluster

How to add Ceph storage

- Make sure you have a running Ceph cluster
- Add the RBD storage pool through the GUI
  - Infrastructure → Primary Storage
    - Tip: Add a tag 'rbd' to the storage pool
- Start creating instances
  - Require the tag 'rbd' in your disk offering
    - This makes sure that RBD image is created on your Ceph cluster

Future plans

- Implement snapshot and backup support
  - In 4.2 with new storage code
- Cloning (aka layering) support
  - One base/golden image for multiple Instances
- No more need for NFS for System VMs
  - Fixed in 4.2
- Ceph support for Secondary / Backup storage
  - Backup storage is new in 4.2
  - Ceph has a S3-compatible gateway
- 4.2 to be released in June this year

Resources

- CloudStack source code can be obtained from www.cloudstack.org
  - DEB and RPM packages are available
- Libvirt 0.9.13 or newer can be downloaded from libvirt.org
- Ceph can be downloaded from ceph.com
  - DEB and RPM packages are available
- Documentation on Ceph.com
  - http://ceph.com/docs/master/rbd/rbd-cloudstack/

Testing is needed!

- All the testing has been done in-house
- External feedback is very much appreciated
- Bugs can be reported in the Jira issue tracker
  - https://issues.apache.org/jira/

Thanks

- Find me on:
  - E-Mail: wido@42on.com
  - IRC: widodh @ Freenode / wido @ OFTC
  - Skype: widodh / contact42on
  - Twitter: widodh