Introduction to Ceph and Architectural Overview

Federico Lucifredi
Product Management Director, Ceph Storage
Boston, December 16th, 2015
CLOUD SERVICES

COMPUTE  NETWORK  STORAGE

the future of storage™
YOU TECHNOLOGY YOUR DATA
How Much Store Things All Human History?!

gaaaaaaaaaahhhhh!!!!!!

carving

writing

carving

paper

computers

distributed storage

cloud computing
GIANT SPENDY COMPUTER
“STORAGE APPLIANCE”
SUPPORT AND MAINTENANCE

34% of revenue (5.2 billion dollars)

PROPRIETARY SOFTWARE

1.1 billion in R&D Spent in a year

PROPRIETARY HARDWARE

1.6 million square feet of manufacturing space
THE CLOUD
philosophy

- OPEN SOURCE
- COMMUNITY-FOCUSED

design

- SCALABLE
- NO SINGLE POINT OF FAILURE
- SOFTWARE BASED
- SELF-MANAGING
8 years & 20,000 commits later...
CEPH STORAGE CLUSTER

A reliable, easy to manage, next-generation distributed object store that provides storage of unstructured data for applications

CEPH OBJECT GATEWAY
A powerful S3- and Swift-compatible gateway that brings the power of the Ceph Object Store to modern applications

CEPH BLOCK DEVICE
A distributed virtual block device that delivers high-performance, cost-effective storage for virtual machines and legacy applications

CEPH FILE SYSTEM
A distributed, scale-out filesystem with POSIX semantics that provides storage for a legacy and modern applications
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift

RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift

RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
OSDs:
- 10s to 10000s in a cluster
- One per disk
  - (or one per SSD, RAID group...)
- Serve stored objects to clients
- Intelligently peer to perform replication and recovery tasks

Monitors:
- Maintain cluster membership and state
- Provide consensus for distributed decision-making
- Small, odd number
- These do not serve stored objects to clients
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift

RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
LIBRADOS

- Provides direct access to RADOS for applications
- C, C++, Python, PHP, Java, Erlang
- Direct access to storage nodes
- No HTTP overhead
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift

RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

APP
HOST/VM
CLIENT

APP
RADOS Gateway:
- REST-based object storage proxy
- Uses RADOS to store objects
- API supports buckets, accounts
- Usage accounting for billing
- Compatible with S3 and Swift applications
LIBRADOSS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift

RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes
VIRTUALIZATION CONTAINER

LIBRBD

LIBRADO
HOST

KRBD (KERNEL MODULE)

LIBRADOS
RADOS Block Device:
• Storage of disk images in RADOS
• Decouples VMs from host
• Images are striped across the cluster (pool)
• Snapshots
• Copy-on-write clones
• Support in:
  • Mainline Linux Kernel (2.6.39+)
  • Qemu/KVM
  • OpenStack, CloudStack
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift

RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
Metadata Server

• Manages metadata for a POSIX-compliant shared filesystem
  • Directory hierarchy
  • File metadata (owner, timestamps, mode, etc.)

• Stores metadata in RADOS
• Does not serve file data to clients
• Only required for shared filesystem
What Makes Ceph Unique?
Part one: CRUSH
How Long Did It Take You To Find Your Keys This Morning?

azmeen, Flickr / CC BY 2.0
I Always Put My Keys on the Hook By the Door

vitamindave, Flickr / CC BY 2.0
HOW DO YOU FIND YOUR KEYS WHEN YOUR HOUSE IS INFINITELY BIG AND ALWAYS CHANGING?
The Answer: CRUSH!!!!!

pasukaru76, Flickr / CC SA 2.0
hash(object name) % num pg

CRUSH(pg, cluster state, rule set)
CRUSH

- Pseudo-random placement algorithm
  - Fast calculation, **no lookup**
  - Repeatable, deterministic
- Statistically uniform distribution
- Stable mapping
  - Limited data migration on change
- Rule-based configuration
  - Infrastructure topology aware
  - Adjustable replication
  - Weighting
What Makes Ceph Unique
Part two: thin provisioning
HOW DO YOU SPIN UP THOUSANDS OF VMs INSTANTLY AND EFFICIENTLY?
instant copy

144

0 0 0 0

= 144
CLIENT

write

write

write

write

= 148

144

4

= 148
CLIENT

read

read

read

144

4

= 148
What Makes Ceph Unique?
Part three: clustered metadata
<table>
<thead>
<tr>
<th>Permissions</th>
<th>Owner</th>
<th>Group</th>
<th>User</th>
<th>Size</th>
<th>Date</th>
<th>Time</th>
<th>File Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>64</td>
<td>Apr 26</td>
<td>17:54</td>
<td><code>libgssapi_krb5.so</code></td>
<td>-&gt; mit-krb5/libgssapi_krb5.so</td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>21</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgssapi_krb5.so.2</code></td>
<td>-&gt; libgssapi_krb5.so.2.2</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>216824</td>
<td>Jul 31</td>
<td>2012</td>
<td><code>libgssapi_krb5.so.2.2</code></td>
<td></td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>13</td>
<td>Apr 26</td>
<td>17:54</td>
<td><code>libgs.so.8</code></td>
<td>-&gt; libgs.so.8.71</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>17</td>
<td>root</td>
<td>root</td>
<td>9478048</td>
<td>Jan 25</td>
<td>2011</td>
<td><code>libgs.so.8.71</code></td>
<td></td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>21</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgssrpc.so</code></td>
<td>-&gt; mit-krb5/libgssrpc.so</td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>16</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgssrpc.so.4</code></td>
<td>-&gt; libgssrpc.so.4.1</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>115352</td>
<td>Jul 31</td>
<td>2012</td>
<td><code>libgssrpc.so.4.1</code></td>
<td></td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>21832</td>
<td>Sep 8</td>
<td>2010</td>
<td><code>libgthread-2.0.a</code></td>
<td></td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>972</td>
<td>Sep 8</td>
<td>2010</td>
<td><code>libgthread-2.0.a</code></td>
<td></td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>26</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgthread-2.0.so</code></td>
<td>-&gt; libgthread-2.0.so.0.2400.2</td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>26</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgthread-2.0.so.0</code></td>
<td>-&gt; libgthread-2.0.so.0.2400.2</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>17704</td>
<td>Sep 8</td>
<td>2010</td>
<td><code>libgthread-2.0.so.0.2400.2</code></td>
<td></td>
</tr>
<tr>
<td>drwxr-xr-x</td>
<td>2</td>
<td>root</td>
<td>root</td>
<td>4096</td>
<td>Apr 26</td>
<td>18:00</td>
<td><code>libgtk2.0-0</code></td>
<td></td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>49</td>
<td>root</td>
<td>root</td>
<td>9275282</td>
<td>Oct 14</td>
<td>2010</td>
<td><code>libgtk-x11-2.0.a</code></td>
<td></td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>49</td>
<td>root</td>
<td>root</td>
<td>981</td>
<td>Oct 14</td>
<td>2010</td>
<td><code>libgtk-x11-2.0.a</code></td>
<td></td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>26</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgtk-x11-2.0.so</code></td>
<td>-&gt; libgtk-x11-2.0.so.0.2000.1</td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>26</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgtk-x11-2.0.so.0</code></td>
<td>-&gt; libgtk-x11-2.0.so.0.2000.1</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>49</td>
<td>root</td>
<td>root</td>
<td>4319784</td>
<td>Oct 14</td>
<td>2010</td>
<td><code>libgtk-x11-2.0.so.0.2000.1</code></td>
<td></td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>15</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgvc.so</code></td>
<td>-&gt; libgvc.so.5.0.0</td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>15</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgvc.so.5</code></td>
<td>-&gt; libgvc.so.5.0.0</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>49</td>
<td>root</td>
<td>root</td>
<td>504424</td>
<td>Jul 5</td>
<td>2010</td>
<td><code>libgvc.so.5.0.0</code></td>
<td></td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>16</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgvc.so.5</code></td>
<td>-&gt; libgvc.so.5.0.0</td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>16</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libgvc.so.1</code></td>
<td>-&gt; libgvc.so.1.0.0</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>482856</td>
<td>Jul 5</td>
<td>2010</td>
<td><code>libgvc.so.1.0.0</code></td>
<td></td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>267948</td>
<td>Apr 13</td>
<td>2009</td>
<td><code>libHalf.a</code></td>
<td></td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>16</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libHalf.so</code></td>
<td>-&gt; libHalf.so.6.0.0</td>
</tr>
<tr>
<td>lrwxrwxrwx</td>
<td>root</td>
<td>root</td>
<td>root</td>
<td>16</td>
<td>Apr 26</td>
<td>17:55</td>
<td><code>libHalf.so.6</code></td>
<td>-&gt; libHalf.so.6.0.0</td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>269992</td>
<td>Apr 13</td>
<td>2009</td>
<td><code>libHalf.so.6.0.0</code></td>
<td></td>
</tr>
<tr>
<td>-rw-r--r--</td>
<td>50</td>
<td>root</td>
<td>root</td>
<td>52850</td>
<td>Nov 1</td>
<td>2009</td>
<td><code>libhistory.a</code></td>
<td></td>
</tr>
</tbody>
</table>
one tree

three metadata servers

??
DYNAMIC SUBTREE PARTITIONING
Getting Started With Ceph
Have a working cluster up quickly.

Read about the latest version of Ceph.
• The latest stuff is always at http://ceph.com/get

Deploy a test cluster using ceph-deploy.
• Read the quick-start guide at http://ceph.com/qsg

Deploy a test cluster on the AWS free-tier using Juju.
• Read the guide at http://ceph.com/juju

Read the rest of the docs!
• Find docs for the latest release at http://ceph.com/docs
Getting Involved With Ceph
Help build the best storage system around!

Most project discussion happens on the mailing list.
• Join or view archives at http://ceph.com/list

IRC is a great place to get help (or help others!)
• Find details and historical logs at http://ceph.com/irc

The tracker manages our bugs and feature requests.
• Register and start looking around at http://ceph.com/tracker

Doc updates and suggestions are always welcome.
• Learn how to contribute docs at http://ceph.com/docwriting
Ceph Hammer (v0.94.x)
Best Ceph ever.

1. Rados Performance enhancements: All Flash environments
2. Simplified RGW deployment
3. RGW Object Versioning and Bucket Sharding
4. RBD Mandatory Locking, Object Maps, Copy on Read
5. CephFS Snapshot improvements

and many more. See https://ceph.com/releases/v0-94-hammer-released/
Questions?

Federico Lucifredi
PM Director, Ceph

federico@redhat.com
@0xF2

redhat.com | ceph.com