SHARE in Orlando 2015

SHARE
Educate · Network · Influence

High Availability and Clustering File Systems on Linux on z

Neale Ferguson

Sine Nomine Associates

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides **education**, professional **networking** and industry **influence.**

Copyright (c) 2015 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

Copyright (c) 2014 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

60 YEARS OF SHARE Influencing IT Since 1955

Agenda

- Sample Configuration
- GFS2
- GlusterFS
- DRBD
- Ceph

HA Sample Configuration

Components and Connectivity

Insert Custom Session QR if Desired.

Copyright (c) 2015 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

Copyright (c) 2014 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

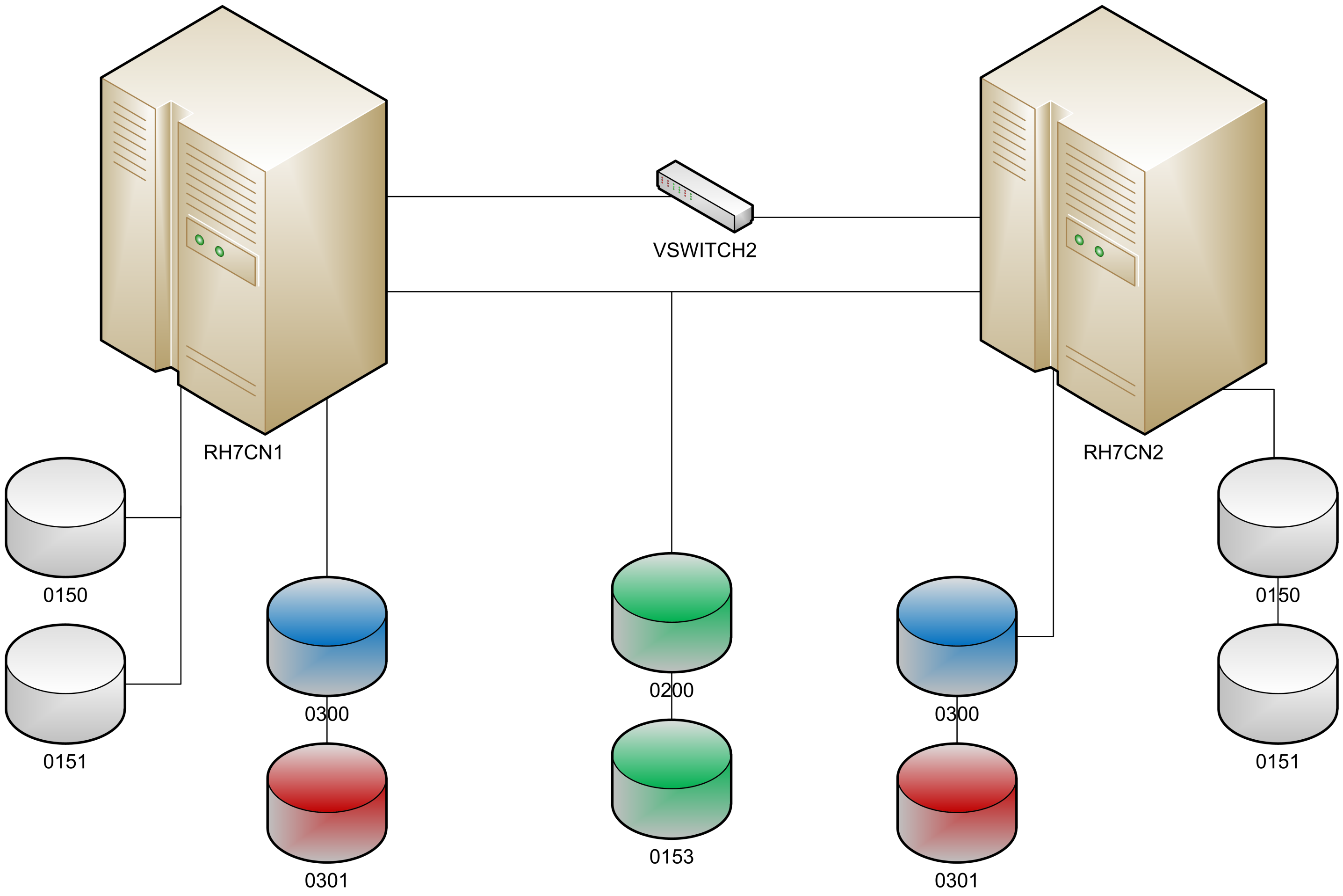
High Availability

- Eliminate Single Points of Failure
- Failover
- Simultaneous Read/Write
- Node failures invisible outside the cluster

High Availability

- Major Components
  - Cluster infrastructure — Provides fundamental functions for nodes to work together as a cluster
    - Configuration-file management, membership management, lock management, and fencing
  - High availability Service Management — Provides failover of services from one cluster node to another in case a node becomes inoperative
  - Cluster administration tools — Configuration and management tools for setting up, configuring, and managing the High Availability Implementation

Sample Configuration



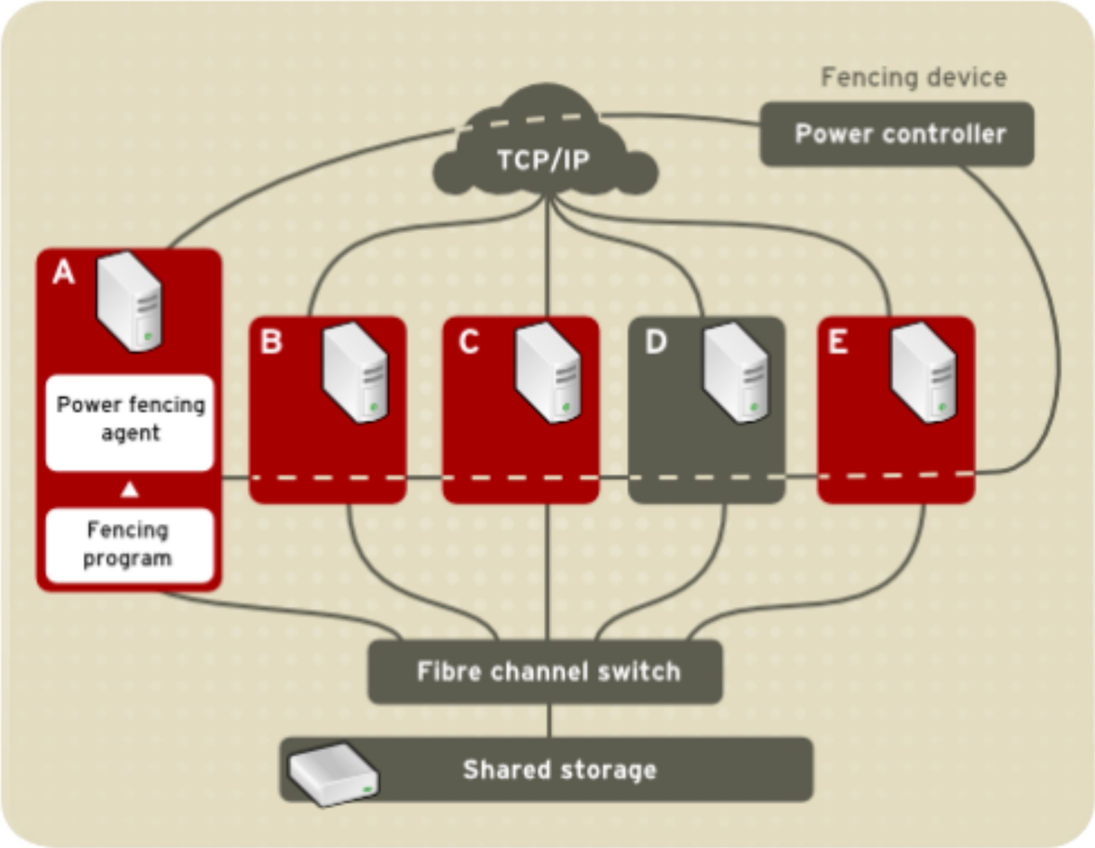
Sample Configuration

- High Availability Packages – RHEL7/CentOS7
  - pacemaker: HA manager
  - fence-agents: STONITH agents
- Firewall Issues
  - `firewall-cmd --permanent --add-service=high-availability`
  - `firewall-cmd --add-service=high-availability`

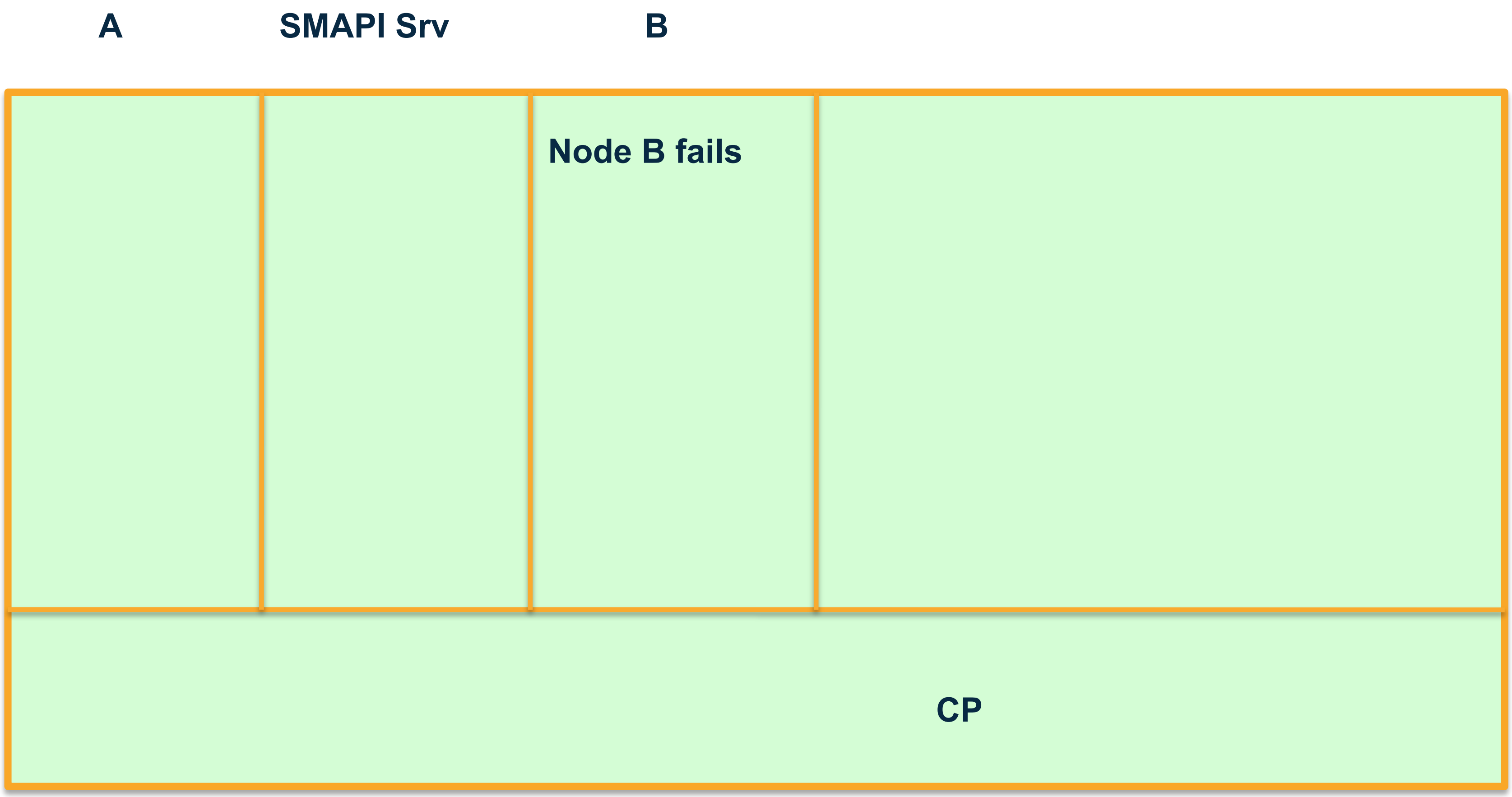
Fencing

- The disconnection of a node from the cluster's shared storage. Fencing cuts off I/O from shared storage, thus ensuring data integrity
- The cluster infrastructure performs fencing through the fence daemon: fenced
- Cluster determines that a node has failed and communicates to other cluster-infrastructure components that the node has failed
- fenced, when notified of the failure, fences the failed node

Power Fencing



z/VM Power Fencing



A

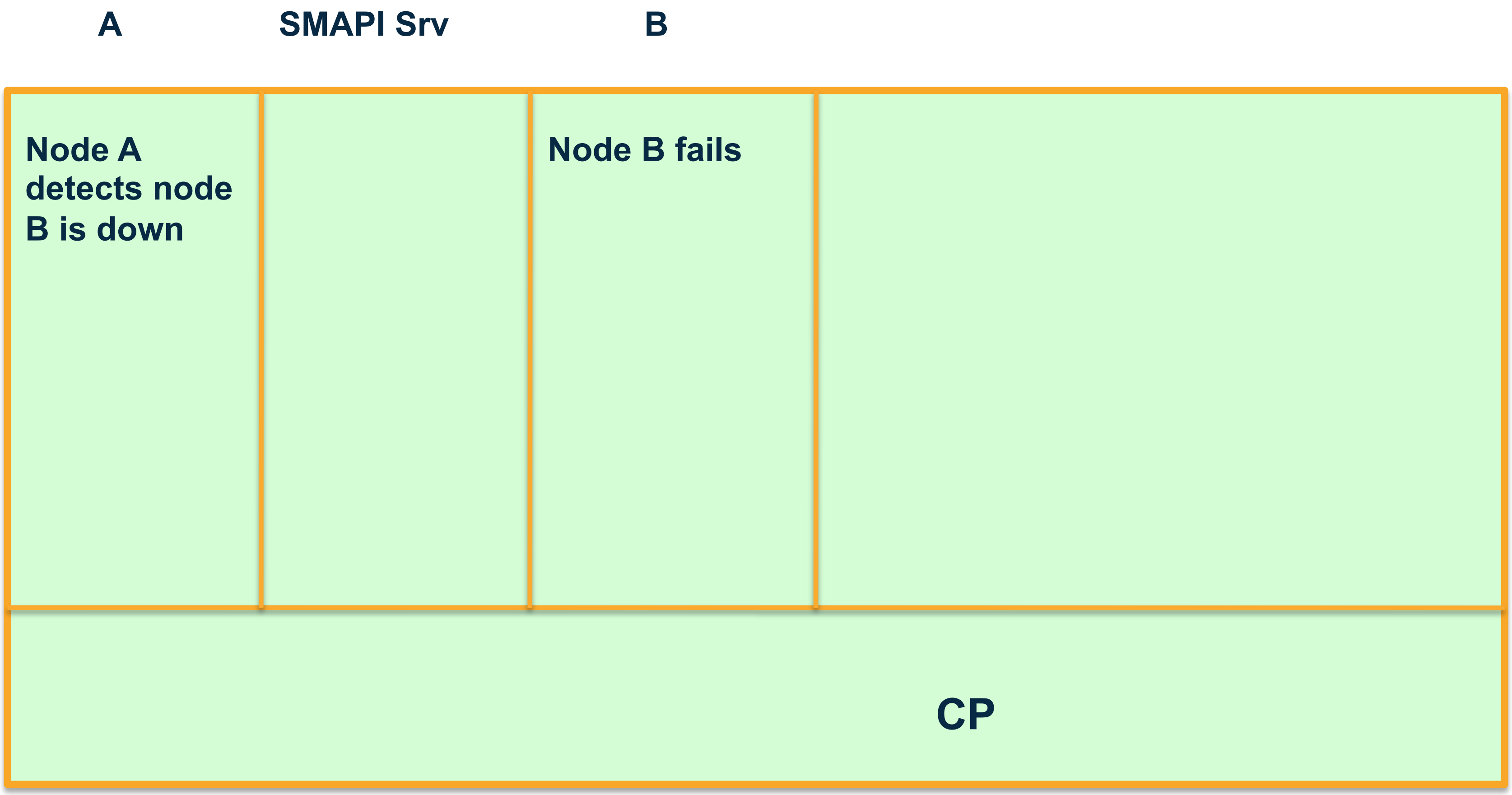
SMAPI Srv

B

Node B fails

CP

z/VM Power Fencing



A

SMAPI Srv

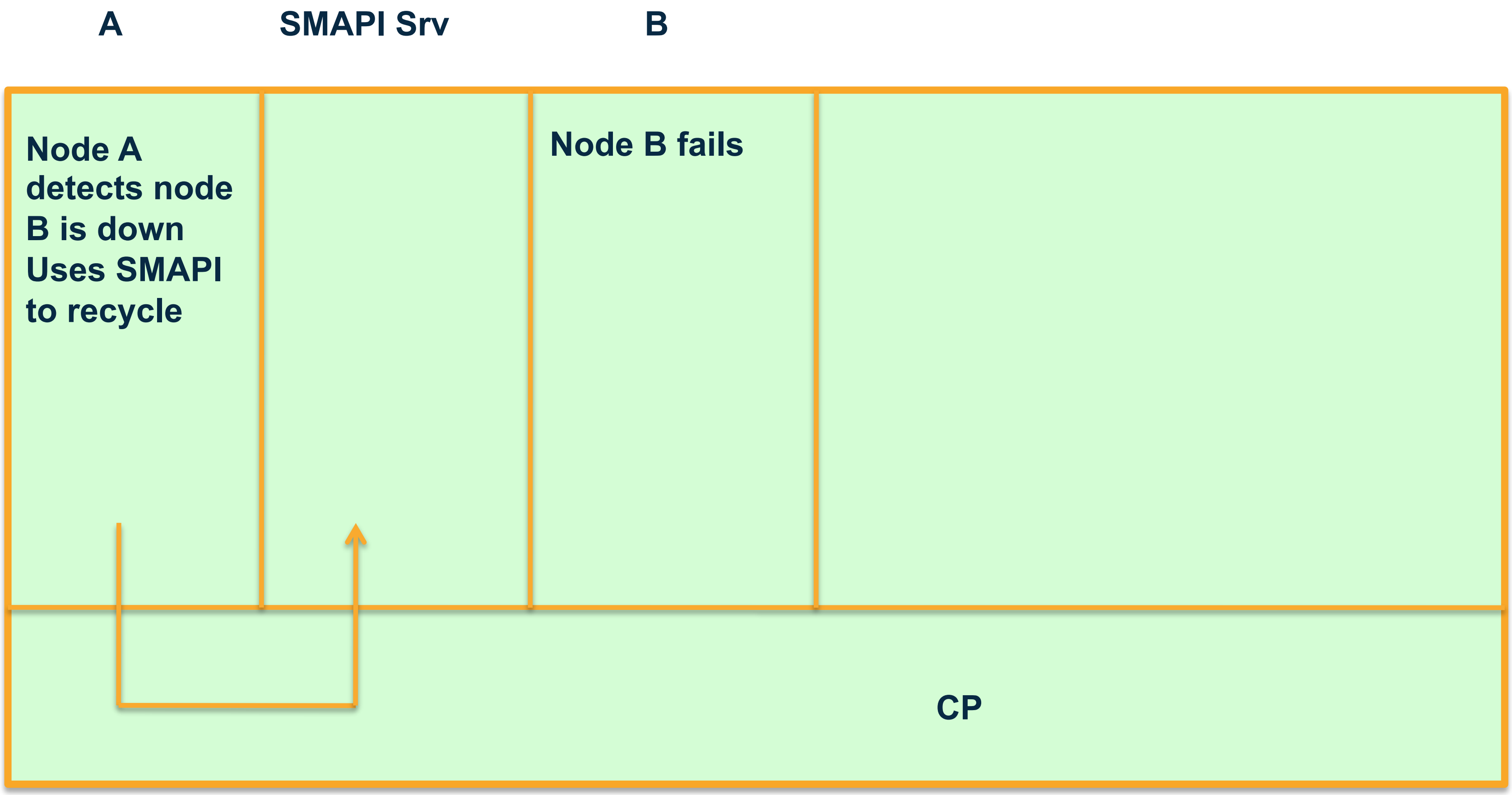
B

Node A detects node B is down

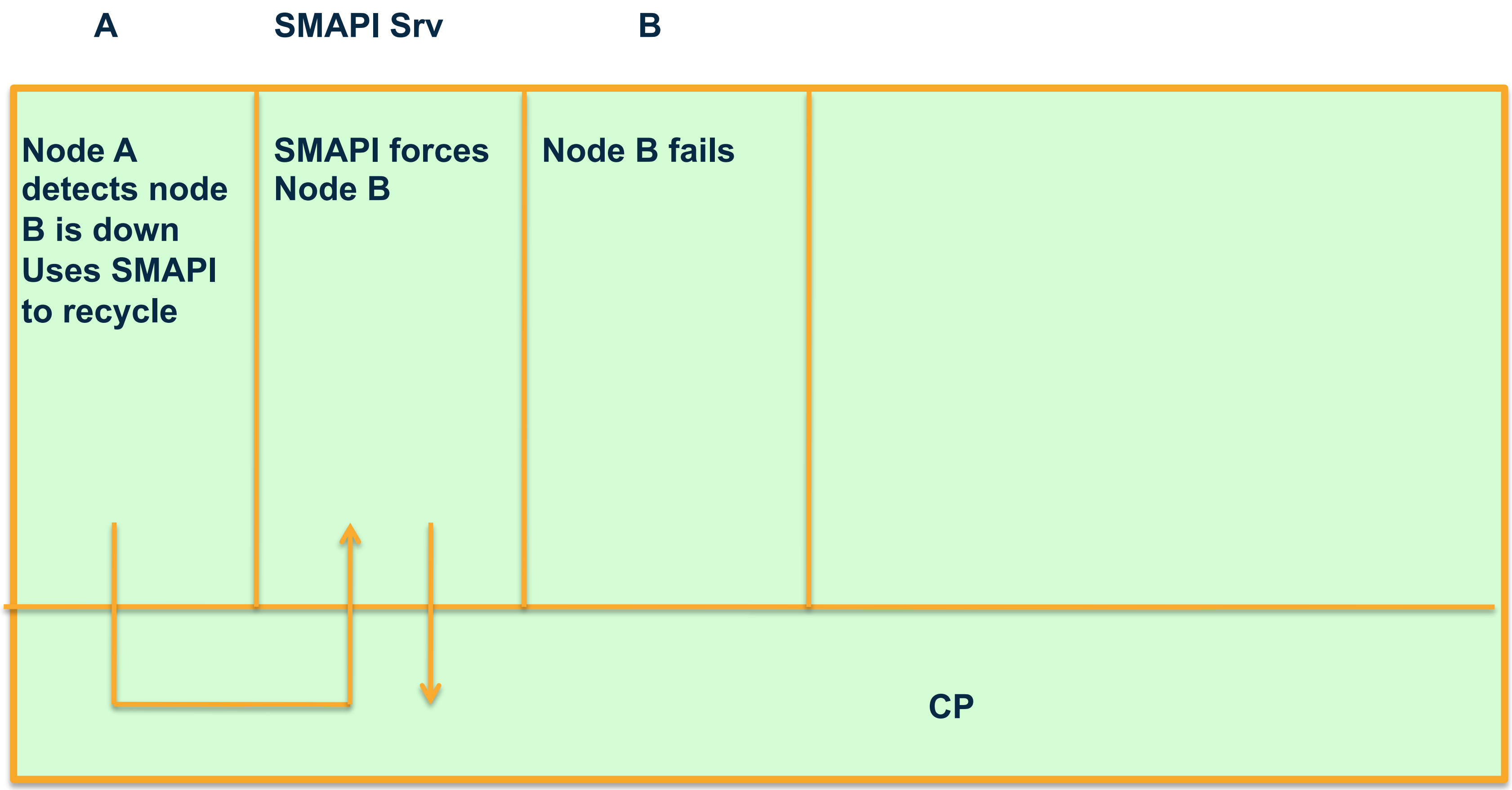
Node B fails

CP

z/VM Power Fencing



z/VM Power Fencing



A

SMAPI Srv

B

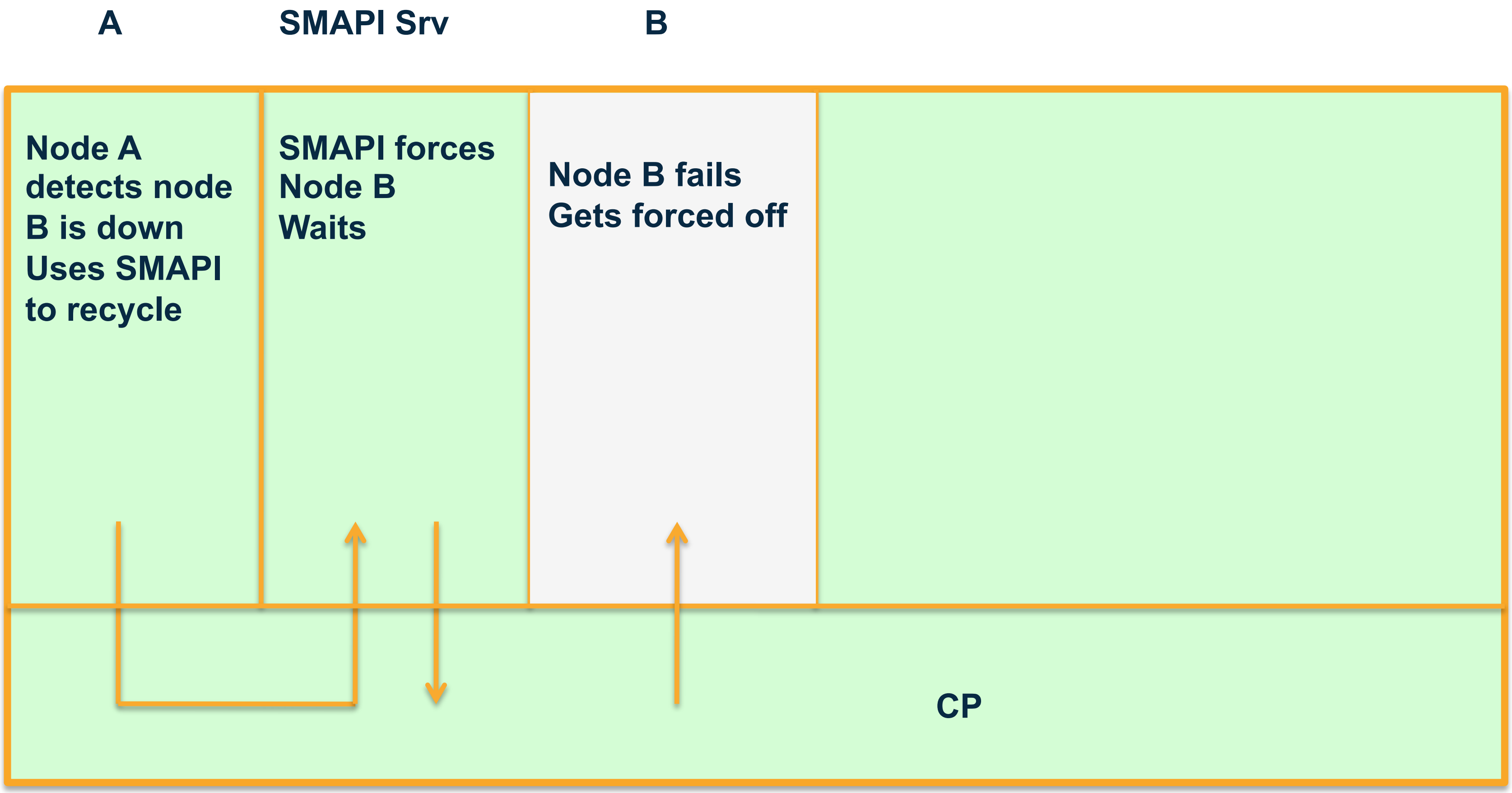
Node A detects node B is down Uses SMAPI to recycle

SMAPI forces Node B

Node B fails

CP

z/VM Power Fencing



A

SMAPI Srv

B

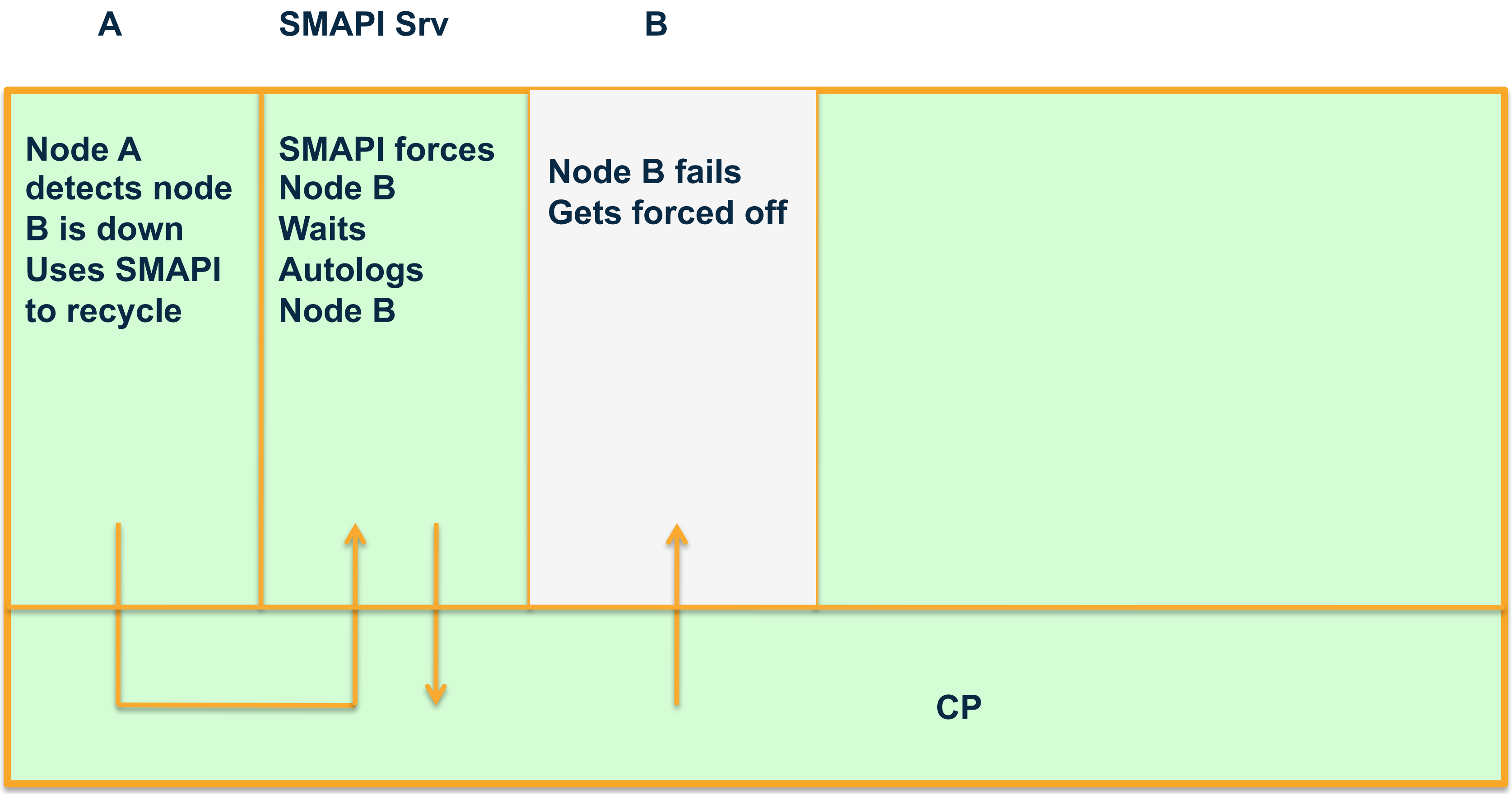
Node A detects node B is down Uses SMAPI to recycle

SMAPI forces Node B Waits

Node B fails Gets forced off

CP

z/VM Power Fencing



A

SMAPI Srv

B

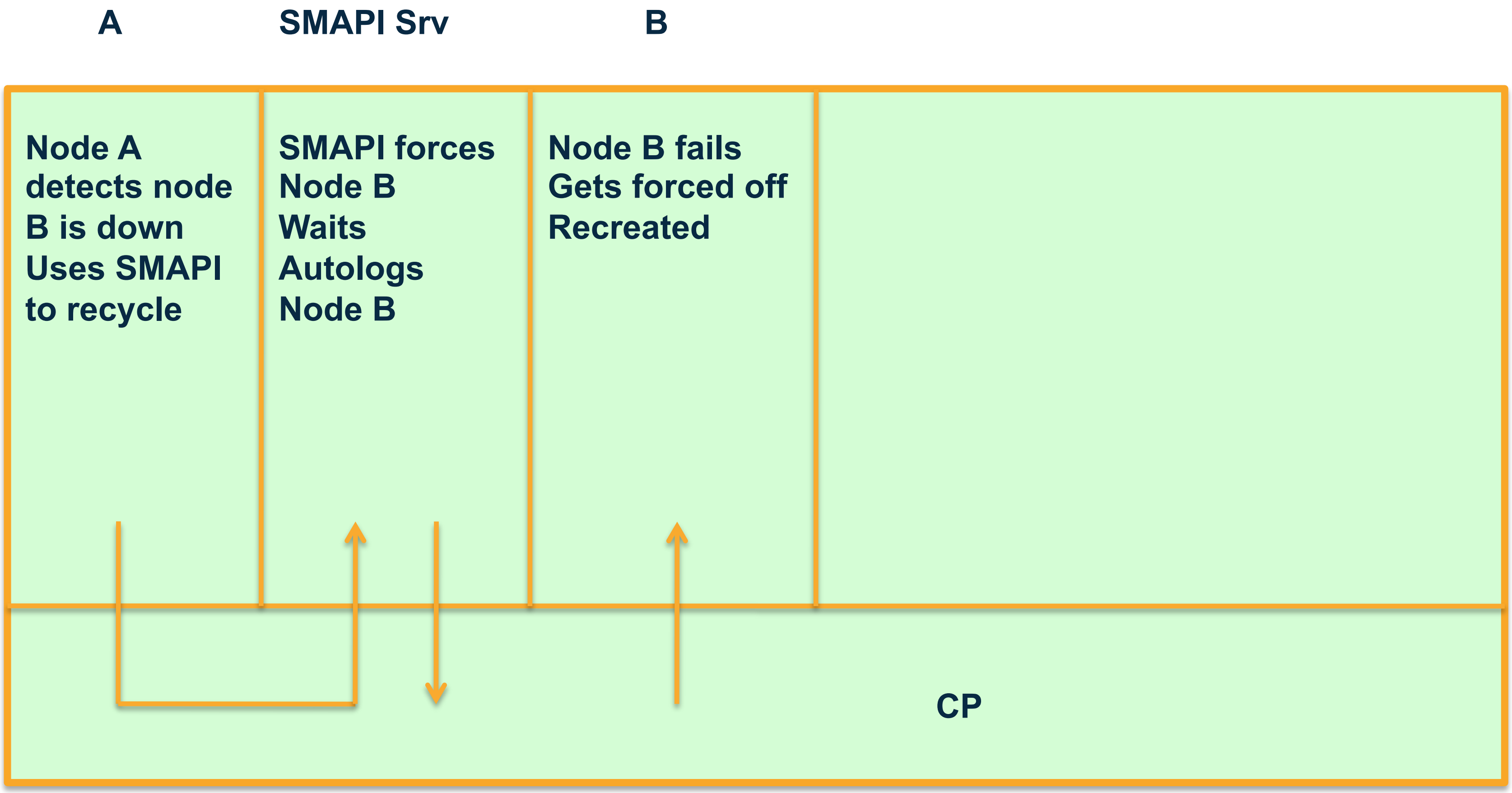
Node A detects node B is down Uses SMAPI to recycle

SMAPI forces Node B Waits Autologs Node B

Node B fails Gets forced off

CP

z/VM Power Fencing



A

SMAPI Srv

B

Node A detects node B is down
Uses SMAPI to recycle

SMAPI forces Node B
Waits
Autologs Node B

Node B fails
Gets forced off
Recreated

CP

z/VM Power Fencing

- Two choices of SMAPI-based fence devices
  - IUCV-based
  - TCP/IP
- Uses image_recycle API to fence a node
- Requires SMAPI configuration update to AUTHLIST:

```
Column 1             Column 66                Column 131
      |                    |                        |
      V                    V                        V
      XXXXXXXX             ALL                      IMAGE_OPERATIONS
```

Clustered File Systems

Configuration, Use Cases, and Requirements

Insert Custom Session QR if Desired.

Copyright (c) 2015 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

Copyright (c) 2014 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

GFS2

- Features
- Use cases
- Requirements

GFS2 - Features

- Shared disk file system for clustering
- Journal file system
- All nodes to have direct concurrent access to the same shared block storage
- May also be used as a local file system
- No client/server roles
- Uses Distributed Lock Manager (dlm) when clustered
- Requires some form of fencing to operate in clusters

GFS2 - Features

- Theoretically accommodates an 8EB filesystem
- Currently supports 100TB (64-bit mode)
- Almost always used with LVM
- Filesystems have unique names within a cluster
- Part of the Linux kernel source tree

GFS2 – Use Cases

- High Availaibility
  - MQSeries
  - Web Server
  - Oracle

GFS2 – Requirements

- Packages:
  - lvm2-cluster, gfs2-utils, **gfs2-kmod**
- SELinux
  - SELinux is highly recommended for security reasons in most situations, but should not be used with GFS2
  - SELinux stores information using extended attributes about every file system object
  - Reading, writing, and maintaining these extended attributes is possible but slows GFS2 down considerably
  - You should turn SELinux off on GFS2 file systems

GFS2 – Sample Configuration

```
USER COMMON NOLOG 8K 8K
MDISK 153 3390 * 3338 *
MINIOPT NOMDC
MDISK 200 3390 * 0500 *
MINIOPT NOMDC
```

```
USER RH7CN1 XXXXXXXX 768M 2G G
ACCOUNT 99999999 CLUSTER
MACHINE ESA
COMMAND SET VSWITCH VSWITCH2 GRANT &USERID
COMMAND COUPLE C600 TO SYSTEM VSWITCH2 IUCV
VSMREQIU
IPL CMS PARM AUTOCR
CONSOLE 0009 3215 T OPERATOR
SPOOL 00C 2540 READER *
SPOOL 00D 2540 PUNCH A
SPOOL 00E 1403 A
LINK MAINT 190 190 RR
LINK MAINT 19E 19E RR
LINK COMMON 153 153 MW
LINK COMMON 200 200 MW
NICDEF C600 TYPE QDIO DEVICES 3 MDOPT FORMAT
  MDISK 150 3390 * 3338 * M
  MDOPT FORMAT
  MDISK 151 3390 * 3338 * M
  MDOPT FORMAT
  MDISK 300 3390 * 200 * M
  MDOPT FORMAT
  MDISK 301 3390 * 200 * M
```

GFS2 – Sample Configuration

```
lvmconf --enable-cluster
pvcreate /dev/dasdd1
vgcreate –Ay –cy vg_cluster /dev/dasdd1
lvcreate –L 500m –n ha_lv vg_cluster
mkfs.gfs2 -j 2 -r 32 -t rh7cluster:vol1 /dev/mapper/vg_cluster-ha_lv
```

GlusterFS

- Features
- Use cases
- Requirements

GlusterFS - Features

- “GlusterFS is a powerful network/cluster filesystem written in user space which uses FUSE to hook itself with VFS layer. GlusterFS takes a layered approach to the file system, where features are added/removed as per the requirement. Though GlusterFS is a File System, it uses already tried and tested disk file systems like ext3, ext4, xfs, etc. to store the data. It can easily scale up to petabytes of storage which is available to user under a single mount point.”
- GPLv3 license

GlusterFS - Features

- File-based mirroring and replication
- File-based striping
- File-based load balancing
- Volume failover
- Scheduling and disk caching
- Storage quotas
- Volume snapshots with user serviceability

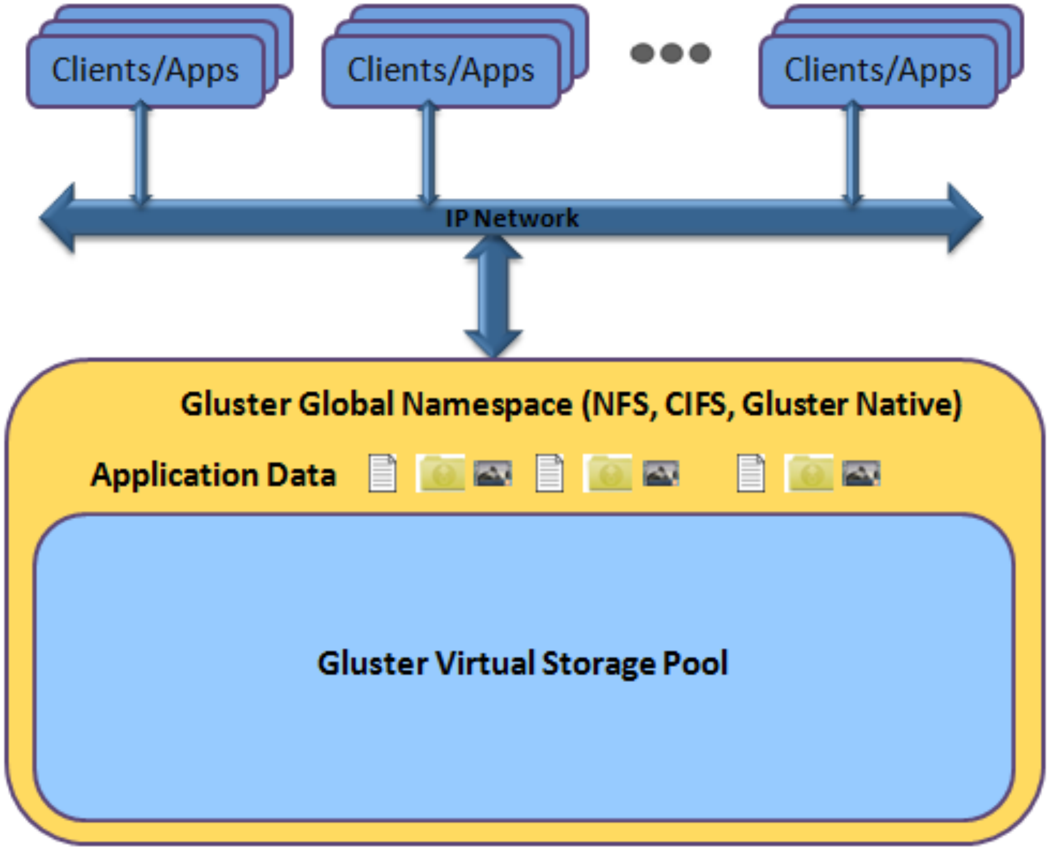
GlusterFS - Terms

- **Brick** - The storage filesystem that has been assigned to a volume
- **Client** - The machine which mounts the volume (this may also be a server)
- **Server** - The machine (virtual or bare metal) which hosts the actual filesystem in which data will be stored
- **Subvolume** - A brick after being processed by at least one translator
- **Volume** - The final share after it passes through all the translators
- **Translator** - Connects to one or more subvolumes, does something with them, and offers a subvolume connection

GlusterFS - Features

- **Server**: Exports a filesystem “as-is” leaving it to client to structure the data
- **Client**: Stateless, independent, consistent configurations
- Volumes can be added, deleted, or migrated dynamically enabling massive scaling

GlusterFS – Conceptual View


Clients/Apps
Clients/Apps
Clients/Apps
IP Network
Gluster Global Namespace (NFS, CIFS, Gluster Native)
Application Data
Gluster Virtual Storage Pool

Source: http://www.gluster.org/documentation/About_Gluster/

GlusterFS – Use Cases

- Red Hat targets three markets:
  - On site
  - Public cloud
  - Private cloud

GlusterFS – Requirements

- Packages:
  - `ntp glusterfs-server glusterfs-fuse glusterfs nfs-utils, selinux-policy-targeted`
- Extra daemons:
  - systemctl enable ntpd
  - systemctl start ntpd
  - timedatectl set-ntp 1
- Firewall rules

GlusterFS – Sample Configuration

- On both nodes:

```
dasdfmt –b 4096 –y /dev/dasdf
fdasd -a /dev/dasdf
pvcreate /dev/dasdf1
vgcreate –cn vg_gluster /dev/dasdf1
vgchange –ay vg_gluster
lvcreate -L120M -n gluster_vol vg_gluster
mkfs.xfs /dev/mapper/vg_gluster-gluster_vol
mount /dev/mapper/vg_gluster-gluster_vol /mnt/gluster/
gluster volume create vol1 replica 2 rh7cn1:/mnt/gluster/brick rh7cn2:/mnt/gluster/brick
gluster volume start vol1
mount -t glusterfs localhost:/vol1 /mnt/store
```

GlusterFS – Sample Configuration

- On one node:
  - echo “ABCDEF” >/mnt/store/test.file
- On both nodes:

  -
    ```
    ls -l /mnt/store
    total 1
    -rw-r--r--. 1 root root 7 Oct 7 15:26 test.file
    ```

  -
    ```
    cat /mnt/store/test.file
      ABCDEF
    ```

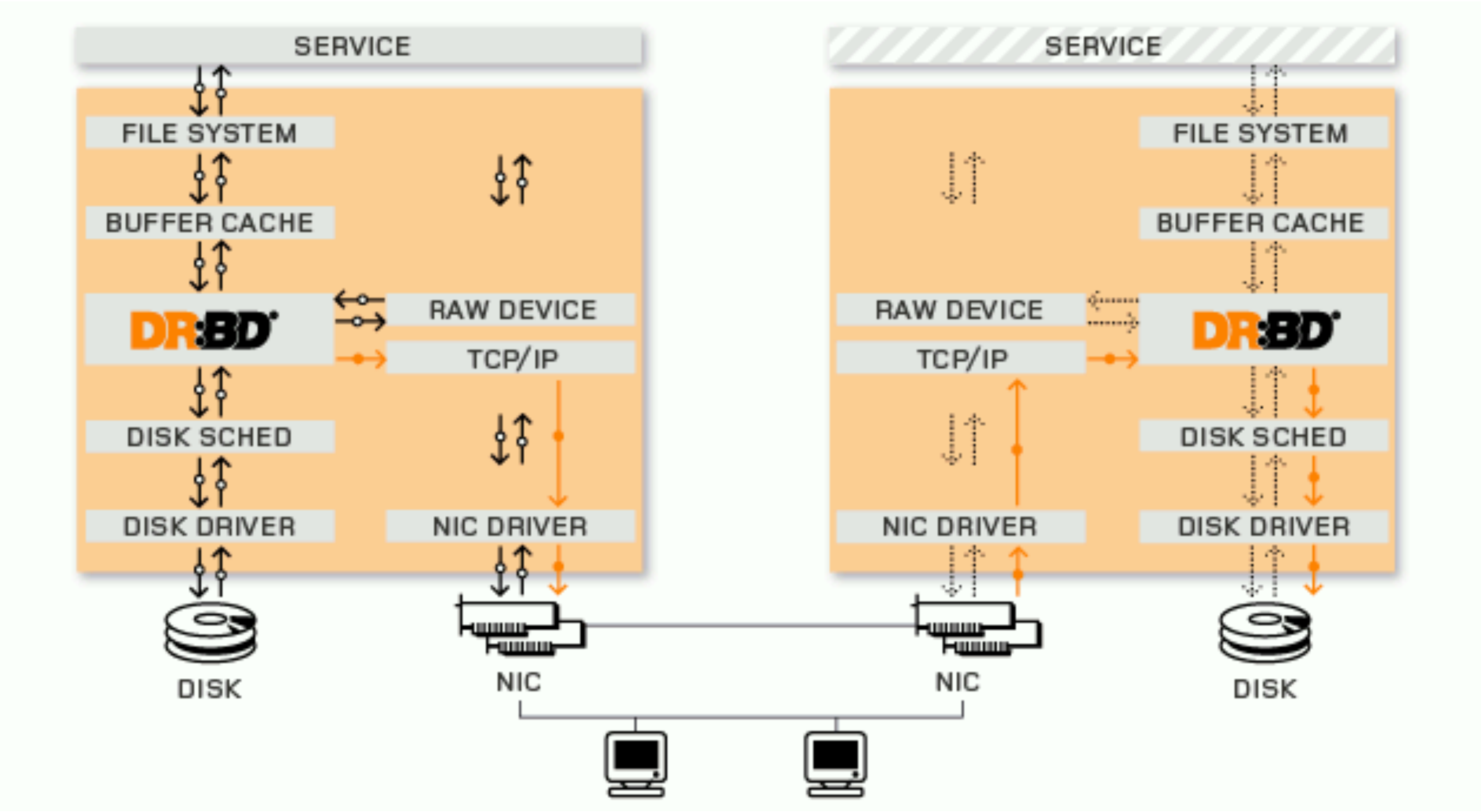
DRBD

- Features
- Use cases
- Requirements

DRBD - Features

- DRBD refers to block devices designed as a building block to form high availability (HA) clusters
- A whole block device is mirrored via the network
- DRBD often described as network based RAID-1
- Implemented as several userspace management applications
- Very loosely - PPRC/XDR-*lite*
- GPLv2 license

DRBD - Features



Source: http://drbd.linbit.com/

DRBD – Use Cases

- Consistency:
  - Design point is 100% consistency
  - Uptime is a secondary goal
  - Use with databases such as mysql
- Access:
  - Multiple primaries with multiple secondaries all interconnected
- Scale-out

DRBD – Requirements

- Packages:
  - drbd-kmod, drbd-utils, drbd-pacemaker
- Firewall:
  - Access to port 7789

```
<?xml version="1.0" encoding="utf-8"?>
<service>
  <short>Distributed Replicated Block Device</short>
  <description>This allows you to use the Distributed Replicated Block Device</description>
  <port protocol="tcp" port="7789"/>
</service>
```

  - firewall-cmd --permanent --add-service drbd

DRBD – Sample Configuration

```
resource disk1
{
       :
        syncer {
                rate 100M;
                verify-alg sha1;
        }

        on rh7cn1.devlab.sinenomine.net {
                device minor 1;
                disk /dev/disk/by-path/ccw-0.0.0301-part1;
                address 172.17.16.147:7789;
                meta-disk internal;
        }

        on rh7cn2.devlab.sinenomine.net {
                device minor 1;
                disk /dev/disk/by-path/ccw-0.0.0301-part1;
                address 172.17.16.148:7789;
                meta-disk internal;
        }
}
```

DRBD – Sample Configuration

- On both nodes
  - `drbdadm create-md disk1`
- On primary node
  - `/sbin/drbdadm -- --overwrite-data-of-peer primary disk1`
  - `mkfs.xfs /dev/drbd1`
  - `tail -20 /var/log/messages >/mnt/drbd/test.file`
  - `cat /mnt/drbd/test.file`

    ```
    Oct 23 13:44:42 rh7cn1 kernel: block drbd1: peer
    0000000000000004:0000000000000000:0000000000000000:0000000000000000 bits:35986 flags:0 Oct 23
    13:44:42 rh7cn1 kernel: block drbd1: uuid_compare()=2 by rule 30
    Oct 23 13:44:42 rh7cn1 kernel: block drbd1: Becoming sync source due to disk states.
    ```
  - `umount /mnt/drbd`
  - `drbdadm secondary disk1`
- On secondary node
  - `drbdadm primary disk1`
  - `mount /dev/drbd1 /mnt/drbd`
  - `ls -l /mnt/drbd`

    ```
    total 4
    -rw-r--r--. 1 root root 2076 Oct 23 13:55 test.file
    ```

HA & Clustered File Systems

Putting it all together

Insert Custom Session QR if Desired.

Copyright (c) 2015 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

Copyright (c) 2014 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

Node 1 & 2 – Cluster Status

```
ZVMPOWER  (stonith:fence_zvm):  Started rh7cn2.devlab.sinenomine.net
Clone Set: dlm-clone [dlm]
    Started: [ rh7cn2.devlab.sinenomine.net ]
    Stopped: [ rh7cn1.devlab.sinenomine.net ]
Resource Group: apachegroup
    VirtualIP   (ocf::heartbeat:IPaddr2):  Started rh7cn2.devlab.sinenomine.net
    Website     (ocf::heartbeat:apache):   Started rh7cn2.devlab.sinenomine.net
    httplvm     (ocf::heartbeat:LVM): Started rh7cn2.devlab.sinenomine.net
    http_fs     (ocf::heartbeat:Filesystem):    Started rh7cn2.devlab.sinenomine.net
Clone Set: clvmd-clone [clvmd]
    Started: [ rh7cn2.devlab.sinenomine.net, rh7cn1.devlab.sinenomine.net ]
Clone Set: clusterfs-clone [clusterfs]
    Started: [ rh7cn2.devlab.sinenomine.net, rh7cn1.devlab.sinenomine.net ]
Master/Slave Set: GDPSClone [GDPS]
    Masters: [ rh7cn2.devlab.sinenomine.net ]
    Slaves: [ rh7cn1.devlab.sinenomine.net ]
GDPSFS    (ocf::heartbeat:Filesystem):     Started rh7cn2.devlab.sinenomine.net
brick1    (ocf::heartbeat:Filesystem):     Stopped
brick2    (ocf::heartbeat:Filesystem):     Started rh7cn2.devlab.sinenomine.net
Clone Set: glusterd-clone [glusterd]
    Started: [ rh7cn2.devlab.sinenomine.net, rh7cn1.devlab.sinenomine.net ]
gvol1     (ocf::glusterfs:volume):   Started rh7cn1.devlab.sinenomine.net
gvol2     (ocf::glusterfs:volume):   Started rh7cn2.devlab.sinenomine.net
store1    (ocf::heartbeat:Filesystem):     Started rh7cn1.devlab.sinenomine.net
store2    (ocf::heartbeat:Filesystem):     Started rh7cn2.devlab.sinenomine.net
```

Node 1 & 2 – Active File Systems

```
Filesystem                         Type            Size  Used Avail Use% Mounted on
/dev/mapper/rhel_rh7cn2-root       xfs             3.7G  1.9G  1.8G  53% /
/dev/mapper/rhel_rh7cn2-boot       xfs             494M  104M  391M  21% /boot
/dev/mapper/vg_gluster-gluster_vol xfs             114M  6.3M  108M   6% /mnt/gluster
localhost:/vol1                    fuse.glusterfs  114M  6.3M  108M   6% /mnt/store
/dev/drbd1                         xfs             135M  7.3M  128M   6% /mnt/drbd
/dev/mapper/vg_cluster-ha_lv       gfs2            300M   35M  266M  12% /mnt/gfs2-demo
/dev/mapper/vg_apache-apache_lv    xfs             294M   16M  279M   6% /var/www
```

Clustered File Systems

A quick look at Ceph

Insert Custom Session QR if Desired.

Copyright (c) 2015 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

Copyright (c) 2014 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

Ceph

- Features
- Use cases
- Requirements

Ceph - Features

- Designed to present object, block, and file storage from a single distributed computer cluster
- Goals:
  - To be completely distributed without a single point of failure
  - Scalable to the exabyte level
- Fault tolerance via data replication

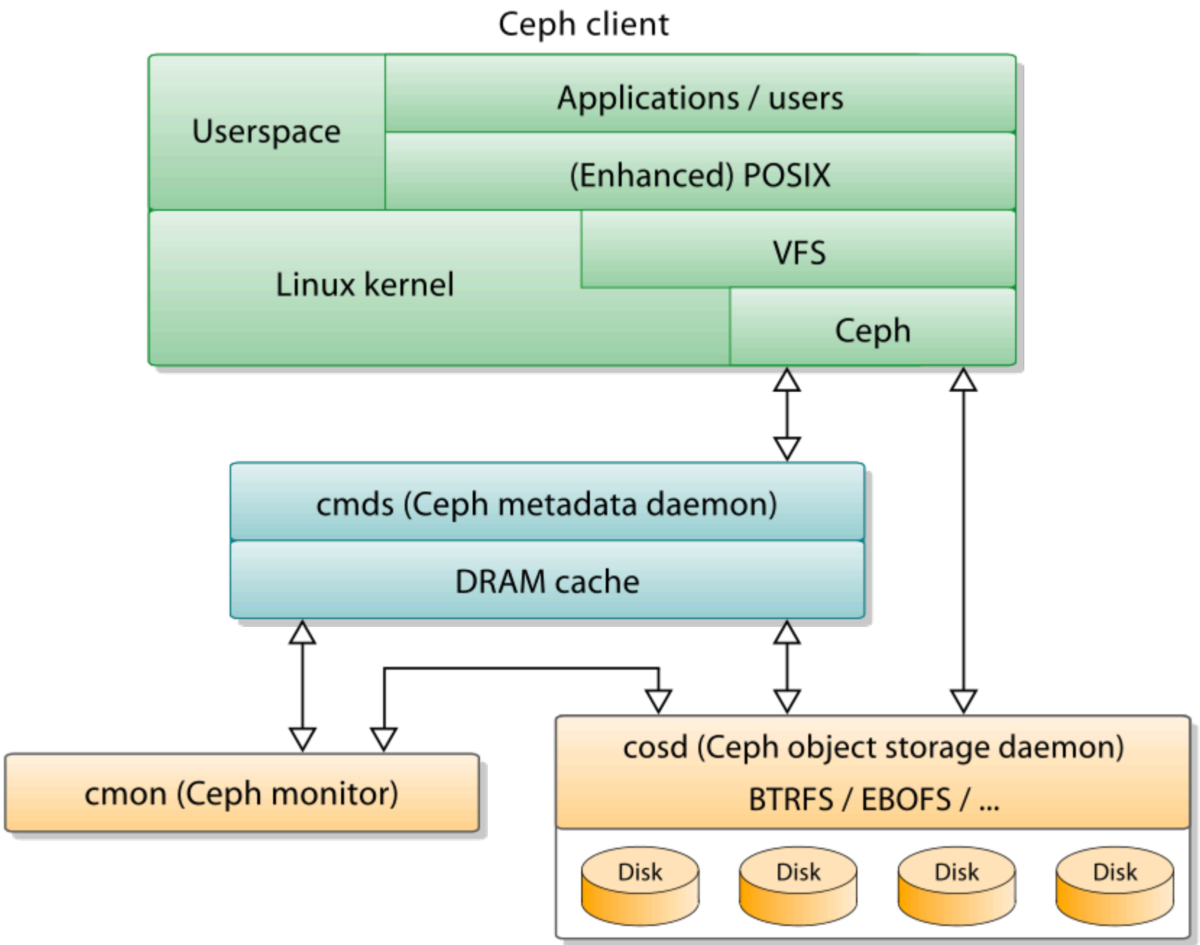
Ceph - Components

- Object servers
  - Stores data, handles data replication, recovery, backfilling, rebalancing, and provides some monitoring information
- Monitor
  - Maintains maps of the cluster state
- Metadata servers
  - Stores metadata enabling POSIX file system users to execute basic commands like `ls`, `find`, etc. without placing an enormous burden on the Ceph Storage Cluster
- Restful gateway
  - An object storage interface to provide applications with a RESTful gateway to Ceph Storage Clusters

Ceph - Components

- All of these are fully distributed which may run on the same set of servers
- A Ceph Storage Cluster requires:
  - At least one Ceph Monitor
  - At least two Ceph OSD Daemons
  - The Ceph Metadata Server is essential when running Ceph Filesystem clients
- OSDs may be added and removed at runtime

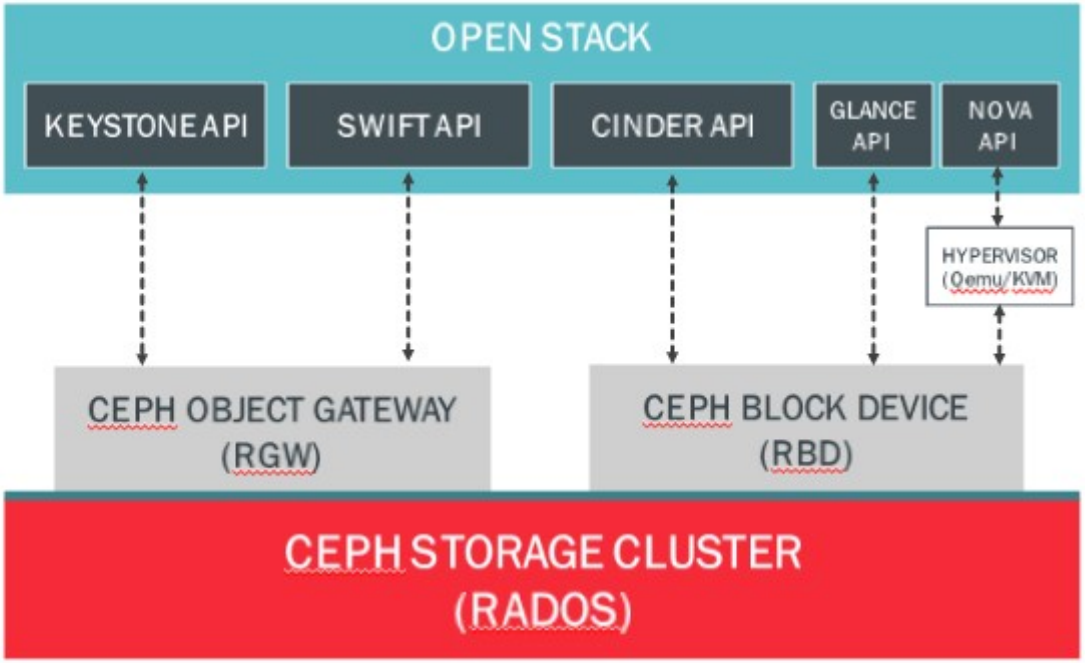
Ceph Architecture



Source: M. Tim Jones (2010-06-04). "Ceph: A Linux petabyte-scale distributed file system" (PDF). IBM

Ceph – Use Cases

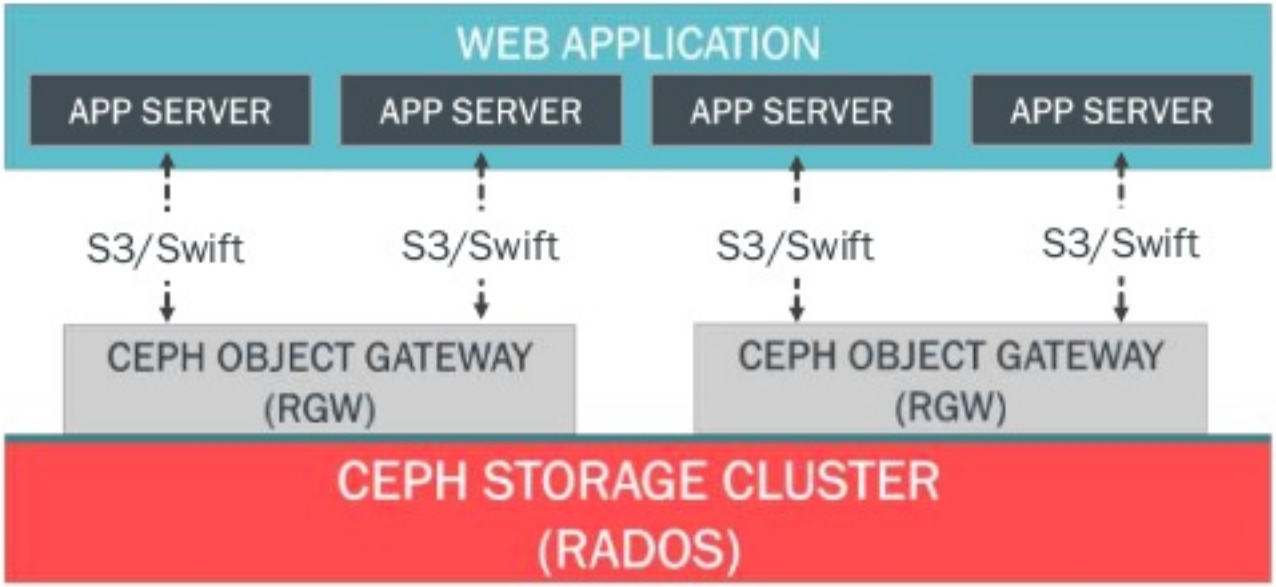
- Storage for an OpenStack-based cloud



Source: http://www.enterprisetech.com/2014/07/16/ceph-clustered-storage-dons-red-hat/

Ceph – Use Cases

- Cloud storage for Web Applications



Source: © Inktank 2013

Ceph – Requirements

- Lots of RPMs (a pig to build):
  - ceph: User space components
  - ceph-common: Utilities to mount and interact with a storage cluster
  - cephfs-java: Java libraries for the Ceph File System
  - ceph-fuse: FUSE based client
  - ceph-radosgw: Rados REST gateway
  - libcephfs1: File system client library
  - libcephfs_jni1: Java Native Interface library
  - librados2: RADOS distributed object store client library
  - python-ceph: Libraries for interacting with RADOS object storage
  - rbd-fuse: FUSE based client to map Ceph rbd images to files

Complete your session evaluations online at www.SHARE.org/Orlando-Eval