CephFS in production

ceph

42on.com

Who am I?

- Wido den Hollander (1986)
- Co-owner and CTO of a PCextreme B.V., a dutch hosting company
- Ceph trainer and consultant at 42on B.V.
- Part of the Ceph community since late 2009
  - Wrote the Apache CloudStack integration
  - libvirt RBD storage pool support
  - PHP and Java bindings for librados
- IPv6 fan :-)

What is 42on?

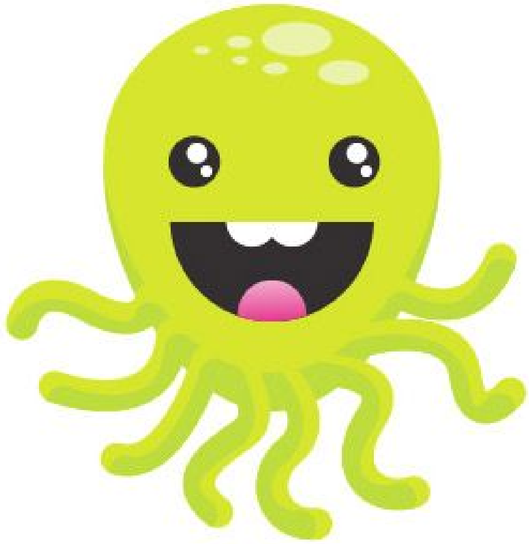
- Consultancy company focused on Ceph and it's Eco-system
- Founded in 2012
- Based in the Netherlands
- I'm the only employee
  - My consultancy company

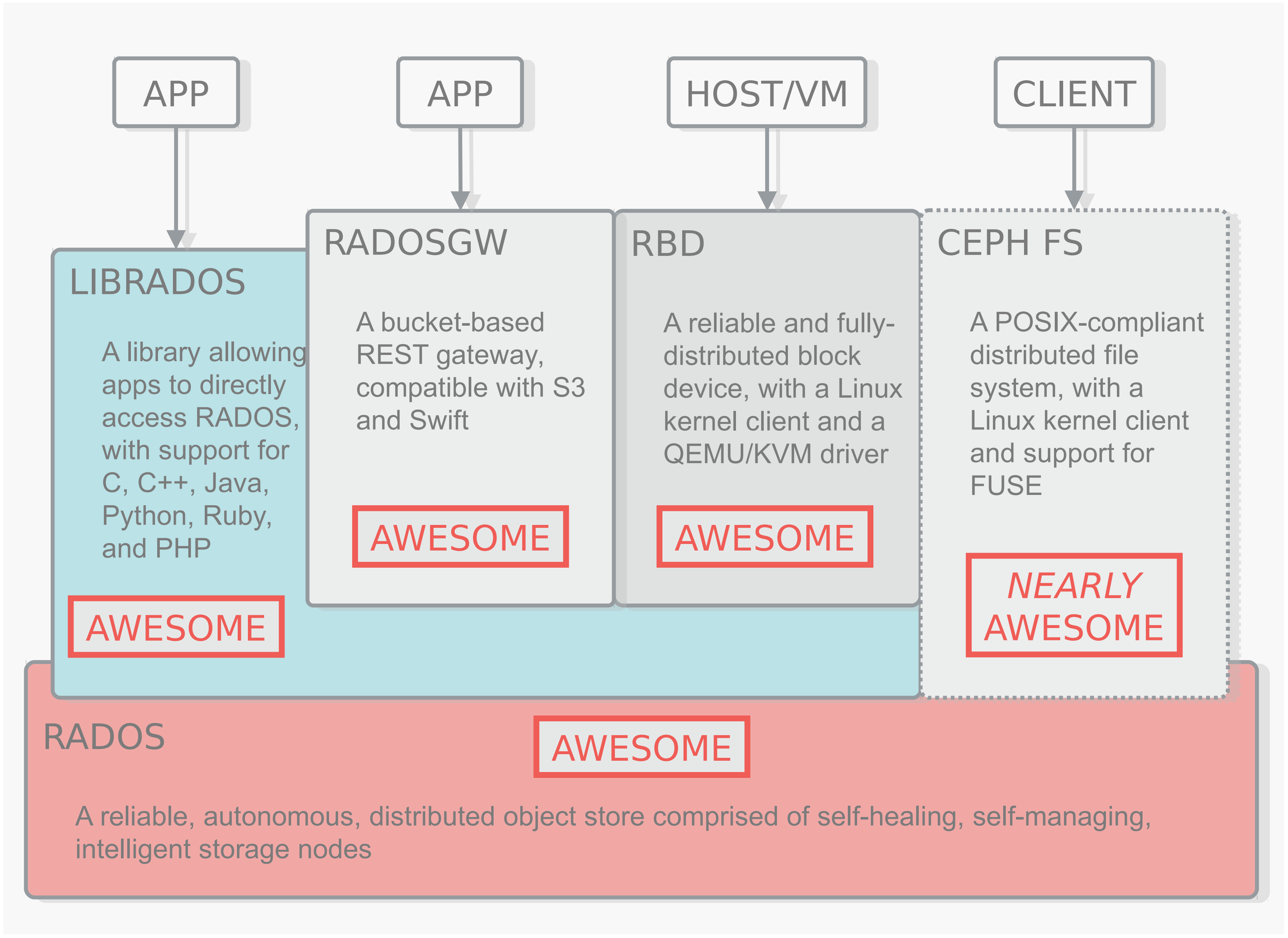
CephFS in production!?

CephFS in production!?

CephFS in production!?

No, it's more like this:





APP

APP

HOST/VM

CLIENT

LIBRADOS

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

AWESOME

RADOSGW

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

AWESOME

RBD

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

AWESOME

CEPH FS

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

NEARLY AWESOME

RADOS

AWESOME

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

Use Case

- Old Oracle ZFS system
  - 2PB in size
  - 6 weeks to migrate
    - Got the call just before Christmas 2014
- Roughly 500TB of data
- Filesystem/POSIX access is a requirement
  - NFS
  - Samba
  - CephFS

Dutch Forensic Institute

The Netherlands Forensic Institute (NFI) is one of the world's leading forensic laboratories.

From its state-of-the-art, purpose-built premises in The Hague, the Netherlands, the NFI provides products and services to a wide range of national and international clients.

http://www.forensicinstitute.nl/

Dutch Forensic Institute

Dutch Forensic Institute

I can't talk about **what** they store

But I can about **how** they store it!

Type of data

- Large files
  - tens, hundreds of thousands of GBs
  - needs to be stored safely (3x replication)
- Small files
  - >200k small files
  - usually temporary files
  - 'allowed' to loose them (2x replication)

Access to data

- Samba
  - Researchers on their Windows computer
- NFS / CephFS
  - Linux VMs with analytic software

The cluster: hardware

- 3 Monitors
- 2 Meta Data Servers (MDS)
  - Dual Xeon E5
  - 256GB memory
- 30 OSD machines
  - 24 1TB SAS drives
  - JBOD controller in WriteBack mode (with BBU)
    - No Journaling SSD
- Dell hardware
- 10Gbit networking
  - Cisco Nexus using 10Gbit Base-T (copper)

The cluster: software

- Ubuntu 14.04
  - XFS for FileStore filesystem
- Started with Ceph Giant (0.89.X)
  - December 2014 ~ January 2015
  - Now running Hammer (0.94.X)
- Ganesha for NFS access
- Samba for Windows access
  - Linux machines with CephFS and Samba re-export

The cluster: Ceph

- 720 OSDs
- ~800TB raw capacity
- 10 pools
  - Backing different directories in CephFS
  - 16k PGs in total
- 1 active MDS, 1 hot-standby

The cluster: Ceph

```
cluster de43a593-ca8e-4c87-b01c-f0d666206b0d
 health HEALTH_OK
 monmap e1: 3 mons at
{mon01=172.17.80.15:6789/0,mon02=172.17.80.16:6789/0,mon03=172.17.80.17:67
89/0}
        election epoch 284, quorum 0,1,2 mon01,mon02,mon03
 mdsmap e2506: 1/1/1 up {0=mds02=up:active}, 1 up:standby
 osdmap e43375: 720 osds: 720 up, 720 in
 pgmap v8241106: 16384 pgs, 10 pools, 83140 GB data, 26365 kobjects
          234 TB used, 548 TB / 782 TB avail
            16365 active+clean
            18 active+clean+scrubbing
            1 active+clean+scrubbing+deep
client io 79725 kB/s rd, 1057 MB/s wr, 3807 op/s
```

NFS with Ganesha

- Userspace NFS daemon
  - Talks directly to libcephfs
  - http://blog.widodh.nl/2014/12/nfs-ganesha-with-libcephfs-on-ubuntu-14-04/

```
EXPORT
{
    Export_ID = 1;
    Path = "/";
    Pseudo = "/";
    Access_Type = RW;
    NFS_Protocols = "3";
    Squash = No_Root_Squash;
    Transport_Protocols = TCP;
    SecType = "none";
    FSAL {
        Name = CEPH;
    }
}
```

CephFS attrs

File attributes instantly give insight in filesystem metadata
Lazy updated by MDS

```
getfattr -d -m ceph.dir.* /mnt/cephfs
getfattr: Removing leading '/' from absolute path names
# file: mnt/cephfs
ceph.dir.entries="2"
ceph.dir.files="0"
ceph.dir.rbytes="88890171665717"
ceph.dir.rctime="1430393746.09144842250"
ceph.dir.rentries="10335945"
ceph.dir.rfiles="9853255"
ceph.dir.rsubdirs="482690"
ceph.dir.subdirs="2"
```

CephFS file layouts

Using setattr ou can configure which data/directory is stored on which RADOS pool

```
$ touch file
$ getfattr -n ceph.file.layout file
# file: file
ceph.file.layout="stripe_unit=4194304 stripe_count=1 object_size=4194304
pool=cephfs_data"
```

CephFS data tiering

- In NFI's case directory structures are the same for each project
  - A project is a directory under a year
    - /mnt/cephfs/<year>/<project>
- Each project has a few directories
  - Directory for large files
    - Set to 64MB stripe size on own pool
  - TMP directory for small files
    - Set to directory with 2x replication and own pool
  - 'Regular' files for project
    - Stored with default file layouts

CephFS attrs

File attributes instantly give insight in filesystem metadata
Lazy updated by MDS

```
getfattr -d -m ceph.dir.* /mnt/cephfs
getfattr: Removing leading '/' from absolute path names
# file: mnt/cephfs
ceph.dir.entries="2"
ceph.dir.files="0"
ceph.dir.rbytes="88890171665717"
ceph.dir.rctime="1430393746.09144842250"
ceph.dir.rentries="10335945"
ceph.dir.rfiles="9853255"
ceph.dir.rsubdirs="482690"
ceph.dir.subdirs="2"
```

CephFS backups

- Directory structure is known
  - Year → Project
- CephFS attrs tell us when a directory or a child underneath was modified
  - `ceph.dir.rctime`
- Recursive walk over attrs to find changed directory
  - Start rsync when a modified directory is found

Performance

- Overall good performance which meets requirements
- 10Gbit/s Write/Read performance
  - Writes always faster then reads
  - 64MB striped files read much faster

Issues

0

No, that would be a lie!

Issues

- Mainly related to NFS and Samba re-exports
  - Stale files and/or directories
- MDS crashed a few times
  - Standby took over
  - These were fixed in Hammer
- **No** data loss or *corruption*

Future

- Upgrade to Jewel
  - Beginning of August
- Samba with VFS linked to libcephfs
- More clients to native CephFS
- Scale cluster when more old systems are migrated out
  - rsync data from Oracle ZFS system

Questions?

- Twitter: @widodh
- Skype: @widodh
- E-Mail: wido@42on.com
- Github: github.com/wido
- Blog: http://blog.widodh.nl/