

# Unlimited Fileserver with Samba CTDB and CephFS

## Who?

### **Robert Sander**

- Linux since 1995
- @gurubert

### **Heinlein Support GmbH**

- 20+ years experience and knowledge around Linux servers and E-Mail
- IT Consulting and 24/7 Linux-Support with > 30 employees
- ISP since 1992
  - jpberlin.de & mailbox.org

## Unlimited Fileserver - Motivation

- need to store „unstructured“ file data
- desktop environment
- archival system / cold storage
- no user will ever delete any file
- regulatory provisions to keep data for 10 years
- budget constraints
  
- storage space growth and reliability more important than speed

## Solutions

- NAS box
  - limited number of disk slots
  - second, third & fourth NAS box?
- SAN
  - block devices do not span multiple controllers
    - size limited
  - Fibre Channel
- Scale vertically has limits
- DRBD
  - block devices with clustered filesystem...
  - multiple filesystems for growth

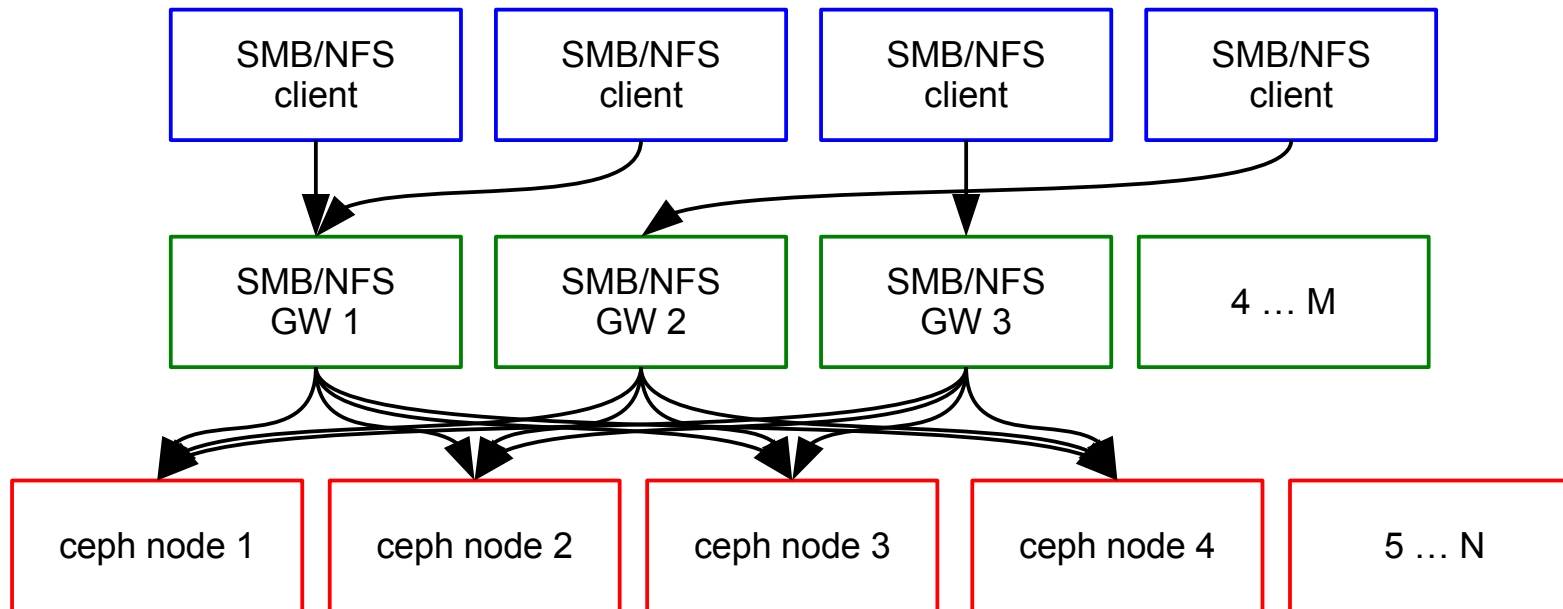
## Solution

- Scale horizontally
- GlusterFS
  - file based replication on the GlusterFS client
  - complicated volume setup
  - suitable for smaller installations
- Ceph to the rescue!
  - suitable for larger installations
  - needs at least 5 nodes to perform

## Concept

- multiple Samba gateways in front of Ceph cluster
- CephFS used for data + CTDB sync
- Samba config in CTDB "registry"
  
- Only file service covered
  - not for Samba AD servers
  - multiple AD servers form their own „cluster“
  
- CTDB can also manage NFS kernel service
  - and other service through event scripts

## Concept



## Ceph Setup

- A working cluster with CephFS

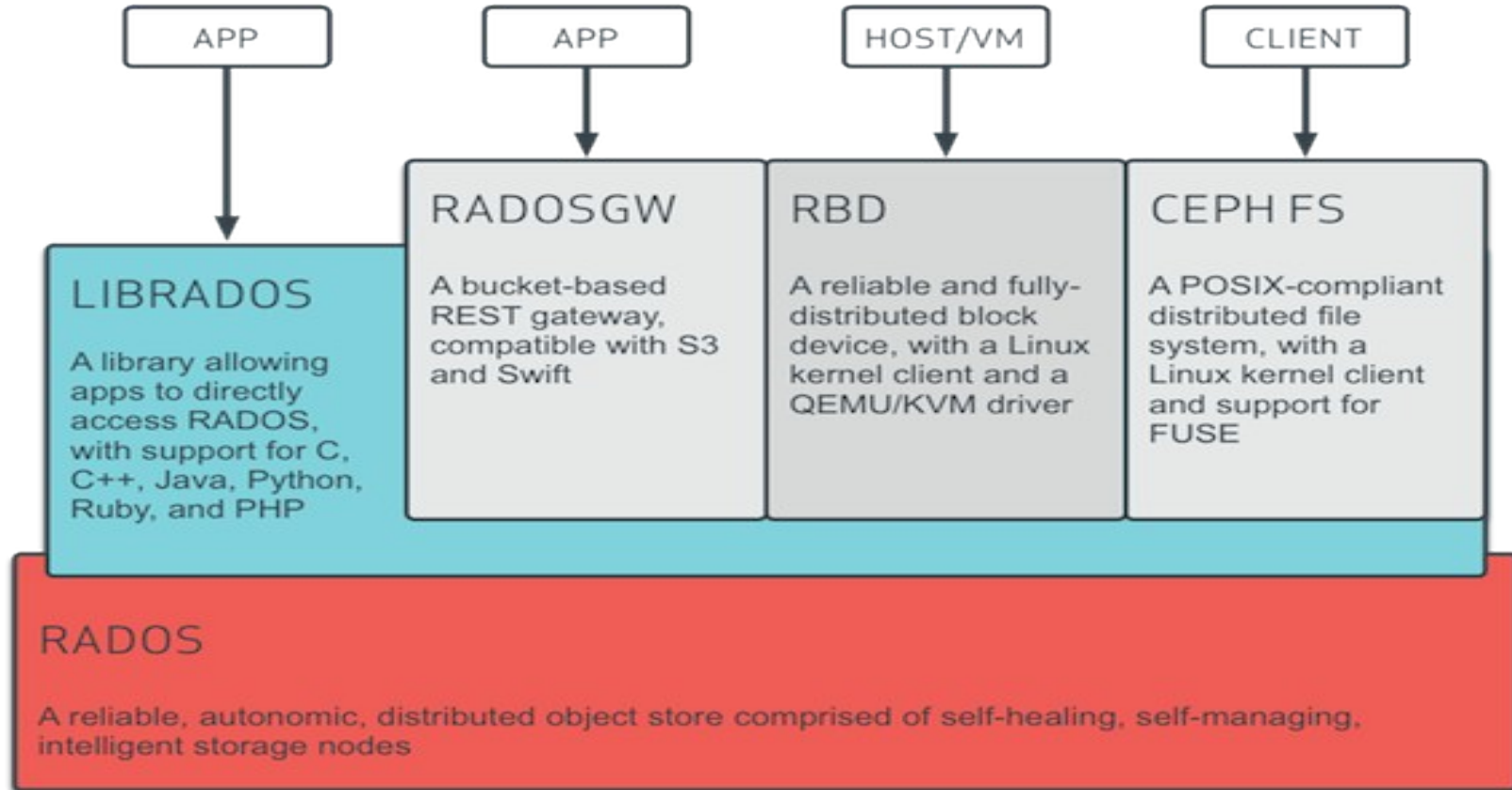


## What is Ceph?

- Ceph is an
  - open-source,
  - massively scalable,
  - software-defined storage system
- which provides
  - object,
  - block and
  - file system storage in a single platform.
- It runs on commodity hardware – saving you costs, giving you flexibility
- and because it's in the Linux kernel, it's easy to consume.

## Advantages

- Self-managing and self-healing
- No Single Point of Failure
- Super efficient placement algorithm
- No need for RAID
- Unified Storage
  - Only one system to scale
  - Easier capacity management
  - Adapt to changing demands
- Open & Extensible
  - S3, SWIFT APIs & RESTful management API
  - 100% open-source technology



## Object Storage

- The basis for the unified storage system
- Objects have
  - a name in a flat namespace
  - metadata attributes
  - payload data
- Distributed
- Replicated
- Parallel access
- APIs: S3, SWIFT, librados

## Block Storage: RBD

- Block device
  - attached direct to Linux host
  - qemu+rbd for virtual guests
- Data striped over multiple objects
  - fast parallel access
  - performs better than a single server
- Snapshots (copy on write)
- Thin provisioning

## File System: CephFS

- POSIX compliant
- Files mapped to objects
  - direct data path
- Ceph Metadata Server (MDS)
  - directories
  - ownership
  - access modes
- metadata itself stored as objects
- active / passive / ... setup
  - immediate switch-over

## Ceph Releases

- Mimic LTS 13.2.5 (March 2019)
    - first release June 2018
  - Luminous LTS 12.2.11 (January 2019)
    - first release August 2017
  - Kraken 11.2.1 (August 2017 EOL)
  - Jewel LTS 10.2.11 (July 2018)
    - first release April 2016
  - Hammer LTS 0.94.10 (August 2017 EOL)
    - first release Apr 2015
  - a stable release every 9 months
    - x.0.z - development releases, x.1.z - release candidates, x.2.z - stable/bugfix releases
- ← CephFS production ready

## CephFS configuration

- at least one Meta Data Server
  - `ceph-deploy mds create {host-name}`
  
- two pools needed: data & meta-data
  - `ceph osd pool create cephfs_data <pg_num>`
  - `ceph osd pool create cephfs_metadata <pg_num>`
  
  - meta-data pool on fast OSDs (SSD, NVMe) recommended
  
- `ceph fs new cephfs cephfs_metadata cephfs_data`



## Gateway Setup

- CephFS-Clients are SMB/NFS servers
- Samba CTDB
- Samba smbd
- NFS kernel server
- CentOS 7 issue with testparm in Samba 4.8.3:
  - <https://bugs.centos.org/view.php?id=15916>
  - Showstopper
- Debian/Ubuntu issues with /etc/ctdb scripts:
  - <https://bugs.launchpad.net/ubuntu/+source/ctdb/+bug/722201>
  - Patch for NFS service available

## CephFS mount

→ cephx key for cephfs access

- caps: [mds] allow rw
- caps: [mon] allow r
- caps: [osd] allow rw pool=cephfs\_data

→ /etc/ceph/ceph.conf

- [client]  
client\_acl\_type = posix\_acl  
fuse\_default\_permissions = false  
client reconnect stale = true

→ FUSE with /etc/ceph/ceph.client.clientid.keyring:

```
id=clientid /cephfs fuse.ceph _netdev 0 0
```

→ kernel client with mount.ceph:

```
mon1:6789,mon2:6789,mon3:6789:/ /cephfs ceph \  
name=clientid,secretfile=/etc/ceph/ceph.client.clientid.secret,_netdev 0 0
```

## Multiple CephFS filesystems

- experimental feature
- at least one MDS per filesystem
- two pools per filesystem (data & metadata)
- „no known bugs“
  
- mount option `mds_namespace`
  
- erasure coded CephFS possible for cold storage
  - `ceph osd pool set cephfs_data allow_ec_overwrites true`
  - only with BlueStore OSDs

## CTDB clustered trivial database

- provides a TDB that has
  - consistent data and
  - consistent lockingacross all nodes in a cluster.
- In case of node failures, CTDB will automatically recover and repair TDBs
- provides HA features such as
  - node monitoring,
  - node failover and
  - IP takeover.
- flexible with application specific management scripts

## CTDB configuration

- /etc/systemd/system/ctdb.service.d/override.conf
  - [Unit]  
After=cephfs.mount  
RequiresMountsFor=/cephfs
- /etc/ctdb/ctdbd.conf
  - CTDB\_RECOVERY\_LOCK=/cephfs/ctdb/lock  
CTDB\_NODES=/etc/ctdb/nodes  
CTDB\_PUBLIC\_ADDRESSES=/etc/ctdb/public\_addresses  
CTDB\_MANAGES\_SAMBA=yes  
CTDB\_MANAGES\_WINBIND=yes  
CTDB\_MANAGES\_NFS=yes  
CTDB\_MAX\_OPEN\_FILES=65536  
CTDB\_LOGGING=file:/var/log/log.ctdb

## ctdb configuration

- /etc/ctdb/nodes
  - list of host IPs
- /etc/ctdb/public\_addresses
  - list of service IPs
  - one for each host
- Service IPs in round-robin DNS
  - fileserver.example.com IN A 10.0.1.11  
IN A 10.0.1.12  
IN A 10.0.1.13
  - SMB clients pick a random IP at mount time
- CLI: `ctdb status`

## Samba configuration

- /etc/samba/smb.conf
  - [global]
    - clustering = yes
    - include = registry
  
- net conf
  - net conf addshare
  - net conf setparm
  - net conf list
  - net conf import
  
- share paths on CephFS mount

## Samba Ceph VFS

- userspace CephFS client for smbd
- Pro:
  - one Ceph client per SMB connection
  - better parallelism
- Con:
  - does not handle symbolic links in CephFS: [https://bugzilla.samba.org/show\\_bug.cgi?id=13918](https://bugzilla.samba.org/show_bug.cgi?id=13918)
  - ¿ Every smbd process is a single CephFS client ?
  
- [https://www.samba.org/samba/docs/4.10/man-html/vfs\\_ceph.8.html](https://www.samba.org/samba/docs/4.10/man-html/vfs_ceph.8.html)



## CTDB in Samba 4.10

- New configuration format
  - file `/etc/ctdb/ctdb.conf`
- Recovery Lock in RADOS object
  - no need for shared directory

## SMB3 witness protocol

- server can pro-actively tell client to connect to other IP
- can be used to drain cluster node before maintenance
- can be used to load balance clients between nodes
- active failover instead of client TCP reconnect
- implemented in future Samba CTDB release

## Kernel NFSv3 configuration

- /etc/sysconfig/nfs
  - NFS\_HOSTNAME="fileserver"
  - RPCNFSDARGS="-N 4"
  - RPCNFSDCOUNT=32
  - STATD\_PORT=595
  - STATD\_OUTGOING\_PORT=596
  - STATD\_HOSTNAME="\$NFS\_HOSTNAME"
  - STATD\_HA\_CALLOUT="/etc/ctdb/statd-callout"
  - GSS\_USE\_PROXY="yes"
  - MOUNTD\_PORT=597
  - RQUOTAD\_PORT=598
  - LOCKD\_UDPPORT=599
  - LOCKD\_TCPPOINT=599
- same port numbers on every host
- /etc/exports
  - /ceph/files \*(rw,fsid=1235,async,crossmnt,no\_subtree\_check)
  - identical on all gateway hosts, filesystem ID is a small integer != 0 or UUID

## Ganesha NFS

- userspace NFS server
- CephFS file system abstraction layer (FSAL)
- NFSv4.1+
- only one Ceph filesystem per Ganesha daemon
- HA only in active/passive mode
  - active/active may work
- Ganesha 2.7 will have clustering support
  - does not need CTDB
  - recovery via RADOS objects
- may replace kernel NFSd soon
  - but no POSIX ACLs over NFSv4

## SMB client configuration

- no special config needed
- Important: use RRDNS hostname as fileserver name
- /etc/fstab examples:
  - `//fileserver/data /mnt/smb cifs guest,icharset=utf8,_netdev 0 0`
  - `fileserver:/ceph/files /mnt/nfs nfs rw,vers=3,_netdev 0 0`

## Demo

## Summary

- Ceph and Samba fit
- Ceph and NFS fit
- endless storage capacity possible in one setup
  
- <https://github.com/HeinleinSupport/ceph-heinlein-tools/tree/master/ctdb>
  
- Disclaimer: "When not to Ceph" by Lars Marowsky-Brée  
<https://cephalocon2019.sched.com/event/M7jE/>  
<https://www.youtube.com/watch?v=vwBwISWvSPQ>