Samba and Ceph
Release the Kraken!

David Disseldorp
ddiss@samba.org
Agenda

- Ceph Overview
- State of Samba Integration
  - Performance
- Outlook
Ceph

• Distributed storage system
  – Scalable
  – Fault tolerant
  – Performant
  – Self-healing and self-managing
  – Runs on commodity hardware
  – Mature

• Various client access mechanisms
  – All layered atop a Reliable Autonomic Distributed Object Store (RADOS)
Ceph Architecture

**OBJECT**
- **RGW**
  A web services gateway for object storage, compatible with S3 and Swift

**BLOCK**
- **RBD**
  A reliable, fully-distributed block device with cloud platform integration

**FILE**
- **CEPHFS**
  A distributed file system with POSIX semantics and scale-out metadata management

**LIBRABDOS**
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

**RADOS**
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors
Components

- Object Storage Daemon (OSD)
  - Exposes underlying storage to clients
    - Objects with data and KV metadata
  - One per disk
    - Faster devices can be used for metadata / WAL
  - Handles data replication and recovery
- Monitor
  - Provide consensus on cluster state
## Ceph Placement

<table>
<thead>
<tr>
<th>POOL</th>
<th>Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10 11 10 01 01 01 01 10</td>
</tr>
<tr>
<td>B</td>
<td>01 10 11 10 10 01 01 01</td>
</tr>
<tr>
<td>C</td>
<td>10 01 10 11 01 11 10 01</td>
</tr>
<tr>
<td>D</td>
<td>01 10 01 01 11 10 01 01</td>
</tr>
</tbody>
</table>

**POOLS** (CONTAINING PGs)
Ceph Placement

CONTROLLED
REPLICATION
UNDER
SCALABLE
HASHING

OBJECT NAME → PG ID → [OSD.185, OSD.67]
Replication

- Client determines PG and corresponding OSDs
  - Sends object I/O to primary OSD
  - Writes acknowledged only after writing to all replicas
- Pools can be replicated or erasure coded
  - User-specified redundancy levels and failure domains
- Private OSD network used for replication traffic
Failure Handling

- Monitors and OSDs check state of other OSDs
  - Following outage, PG is assigned to a new node
  - Backfill from peers

- Periodic scrubbing of data and metadata
Ceph Placement

OBJECT NAME → PG ID → [OSD.185, OSD.31]
CephFS

- POSIX compatible clustered filesystem atop RADOS
- MDSes manage filesystem namespace
  - Active/Passive or Active/Active redundancy
- Linux kernel and user-space clients
- Snapshots
- Directory to pool mappings
CephFS

- Basic quotas
- Client caching
  - Fine grained
  - Capabilities granted and revoked by MDS
Samba Integration
Samba Gateway

Samba

Ceph VFS module

libcephfs

TDB

RADOS CLUSTER

data

metadata
Samba Ceph Integration

- CephFS module for Samba: vfs_ceph
  - Added in 2013 by Inktank
  - Maps SMB file and directory I/O to libcephfs API calls
- Static cephx credentials
  - Regardless of Samba authenticated user
  - User configurable via smb.conf
- POSIX ACLs
Samba Ceph Integration

- RADOS clustered mutex helper for CTDB
  - Removes recovery lock mount dependency
- Ceph librados service integration *(coming soon)*
  - Register service with manager daemon
Testing

- Ceph vstart
  - Deploy mock cluster from source
- Samba smbtorture
- cifs.ko fstests
Performance
Performance: Samba vs CephFS

- Preliminary results!
- Environment:
  - Ceph Version 12.2.2
  - Samba 4.6.9
    - Three Samba gateways
    - vfs_ceph
    - Non-overlapping share paths
  - Linux cifs.ko client
    - 4.4 kernel with many backports
    - SMB 3.0 mount
Hardware

• Ceph setup on 8 nodes
  – 5 OSD nodes – 24 cores – 128 GB RAM
  – 3 MON/MDS nodes – 24 cores – 128 GB RAM
  – 6 OSD daemons per node – Bluestore – SSD/NVME journals
• 10 client nodes
  – 16 cores – 16 GB RAM
• Network interconnect
  – Public network 10Gbit/s
  – Cluster network 100Gbit/s
CephFS aggregate IOPS over 10 clients

- 1nw 4k rw
- 1nw 1m rw
- 1nw 4m rw
- 4nw 4k rw
- 4nw 1m rw
- 4nw 4m rw
- 8nw 4k rw
- 8nw 1m rw
- 8nw 4m rw
- 16nw 4k rw
- 16nw 1m rw
- 16nw 4m rw

IOPS (Log Scale)
SMB3 (client caching enabled) aggregate IOPS over 10 clients
Challenges and Future
Challenges

- Cross-protocol client support
  - Coherent client caching
    - Map leases to CephFS `FILE` and `AUTH` capabilities
    - New libcephfs delegations API
  - Shared (NFS, CephFS) ACL model

- Unified authentication and user mapping
  - Use Kerberos / AD for Samba gateway and cephx
Challenges

- libcephfs asynchronous I/O

- Multichannel support
  - Experimental in upstream Samba
  - Not integrated with CTDB

- Automated deployment
Challenges

- Witness protocol
  - Continuous availability of SMB shares
  - Advertise Samba cluster state to clients
  - Transparent client failover
  - Load balancing
Samba: Future

- Ceph backed key-value store for Samba
- Replace or modify CTDB
  - Rocksdb?
  - Samba database API demanding
    - Multiple processes and writers
    - Record locking and transactions
References

- Samba: https://samba.org/
- CTDB: https://ctdb.samba.org/
- Samba Multichannel Blocker Bug: https://bugzilla.samba.org/show_bug.cgi?id=11897
- Greg Farnum: Intro to Ceph, The Distributed Storage System
- Placement diagrams: http://yauuu.me/ride-around-ceph-crush-map.html
Join Us at www.opensuse.org
License
This slide deck is licensed under the Creative Commons Attribution-ShareAlike 4.0 International license. It can be shared and adapted for any purpose (even commercially) as long as Attribution is given and any derivative work is distributed under the same license.

Details can be found at https://creativecommons.org/licenses/by-sa/4.0/

General Disclaimer
This document is not to be construed as a promise by any participating organisation to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. openSUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for openSUSE products remains at the sole discretion of openSUSE. Further, openSUSE reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All openSUSE marks referenced in this presentation are trademarks or registered trademarks of SUSE LLC, in the United States and other countries. All third-party trademarks are the property of their respective owners.

Credits
Template
Richard Brown
rbrown@opensuse.org

Design & Inspiration
openSUSE Design Team
http://opensuse.github.io/branding-guidelines/