SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Joining the Cephalopods

Adding SMB Support to Ceph

Presented By

Günther Deschner
IBM, Samba Team

John Mulligan
IBM, Samba Team

Introductions

Günther Deschner

- Samba Team member since 2005
- Manager of Ceph SMB team in IBM, formerly at Red Hat working on Samba

John Mulligan

- Software Engineer at IBM (also formerly at Red Hat)
- Focused on Orchestration of storage systems
- Recently joined the Samba Team! (this year)

This talk covers our work in the Ceph, Samba, and other, upstream projects (not the IBM Ceph product)

Outline

Our goal was to introduce built-in management and integration of SMB protocol services for CephFS using Samba.

- Introduction to SMB on Ceph
- Adding SMB support to Ceph Orchestration
- Demonstration
- Ongoing performance improvements
- Future Plans

A One-Slide Introduction to Ceph

https://ceph.io - an open-source, distributed storage system

Protocols:

- RADOS - object based storage layer that all other layers build upon
- RBD - Ceph's block device protocol
- **CephFS** - Ceph's file system protocol
- RGW - Ceph's HTTP/Object protocol

Server Processes:

- MON - daemons that define the cluster and establish quorum
- OSD - daemons that expose storage to the cluster
- MDS - metadata server for the file system
- **MGR** - management daemon

Can’t you already use Samba with Ceph?

Yes, but:

- There are building blocks out there, not a cohesive platform
- Few community standards and best practices
- Many questions regarding performance
- Lack of maintenance around Open Source components (Samba VFS Module for Ceph)

Integration and Automation

Ceph already has built-in support for NFS

Our goal is to add a similar level of support and integration for SMB using components from the Samba project:

- Benefit from Ceph's orchestration platform (cephadm)
- Have consistent set of deployment methods
- Reuse Ceph's testing framework
- Provide workflows familiar to Ceph administrators

Reuse & Recycle

We've been involved with various Open Source Projects. Including ones aimed to Containerize Samba, originally for Kubernetes. We can directly reuse these projects with Ceph:

- ☑ Samba components:
  Ceph module, Ceph Snapshot module, CTDB rados mutex helper
- ☑ samba-containers project - Create container images with Samba and container tools
- ☑ sambacc - container tools wrapping Samba
- ☑ Samba Build - tools for building Samba packages from Git branches
- ☑ Testing platforms - Samba Integration Tests suite
- ? samba-operator for Kubernetes

SDC 24

cephadm: Ceph's Native Orchestration Layer

- A container orchestration system, dedicated to Ceph's needs
- Runs on host systems with Podman or Docker, systemd, and Python
- Server side component: a module within the Ceph MGR
- Per-node component: the cephadm "binary"
- Fewer constraints on what we can do compared to Kubernetes
  - Especially around networking
- Fewer pre-existing "building blocks" too
  - We have to build certain abstractions ourselves
  - No "guardrails", especially around networking ;-)

Teaching cephadm how to Samba

- Taking inspiration from the Samba on Kubernetes work:
  - We know how to initialize the configuration & run Samba inside containers
  - Needed to add “init containers” and “sidecar containers” feature to cephadm
  - Created a new “SMB Service” to deploy different sets of containers depending on the chosen configuration
- Flexible Features for Different Deployments:
  - Support for joining MS Active Directory, or “local” Users & Group configuration
  - Synchronizing the Samba configuration with Ceph’s configuration
  - Exporting Prometheus-style metrics
  - True clustering with Samba’s CTDB (with enhancements to CTDB)
  - IP Failover using CTDB

A Manager Module Made Manifest

- OK, so we can create servers, let's make it all manageable: provide a module, similar to Ceph's NFS module, for complete SMB configuration
- Commands/APIs to create, update & remove Clusters and Shares
- Imperative command set:
  - ceph smb cluster create …
  - ceph smb share create …
  - ceph smb share rm …
- A declarative version of the Commands/API based on "resources":
  - Cluster
  - Share
  - Users & Groups Config
  - Join Auth Information
- "Who orchestrates the orchestrators?"
  The SMB module ends up as the center of control

DEMO

I'll be showing you a brief tour of what's available so far Ceph's SMB support using the Ceph smb mgr module.

Demo Materials (to follow along at home)

- 5 Node Ceph Cluster (VM)
- AD DC (VM)
- Windows & Linux clients (VM)
- Ceph container builds from Ceph "main" branch

SDC 24

Samba and Ceph File Systems

- There are two ways to access CephFS from Samba:
  - Kernel mount CephFS with either FUSE or native driver
  - Userspace library via “vfs_ceph” in Samba
- Refreshing vfs_ceph:
  - Part of the Samba codebase, but was not keeping up with changes in Ceph
  - Use all new _at calls from libcephfs, adapting to the major Samba VFS change towards handle based system calls
  - Makes use of the libcephfs API
  - Revitalizing the module with our team’s efforts
  - Using new “low level” APIs from Ceph
  - Continuous integration tests exercising the module

Performance and Scale with vfs_ceph

- Samba architecture: smbd forks on a per-client basis
- Each client becomes an autonomous VFS client
  - SMB “impersonation” at a potentially high cost
  - Cheap for mounted filesystems
  - Potentially expensive when using library interface to connect storage
- Idea: add proxy layer
- libcephfs_proxy.so and libcephfsproxyd (https://github.com/ceph/ceph/pull/58376)
- Intercept each low-level API call
- Impersonation credentials per call via user credentials
  - Finally full support for supplementary groups
- Run-time configuration to enable proxy routing

Improving Case-Insensitivity on Ceph

- Huge percentage of performance drop due to case-folding operations
- Different options:
  - Samba VFS: get_real_filename() operation
  - Case sensitivity flag for cephfs
- Current design in Ceph: toggle file system case sensitivity
  https://tracker.ceph.com/issues/66373

Future Plans

- Improve CTDB scalability and performance
- Multiprotocol support (concurrent use of NFS and SMB)
- SMB3 transparent failover (Witness)
- Expanded support for different Active Directory integrations
- Service Co-location (multiple Samba containers on one cluster node)
- Further improvements for Samba's VFS ceph module:
  - Make use of new libcephfs features link async I/O calls and zero-copy

Thank You!

Questions?

Ceph Project:

- http://ceph.io
- https://github.com/ceph/ceph

Samba Project

- https://samba.org

Günther Deschner

- gd@samba.org
- gdeschne@redhat.com
- guenther.deschner@ibm.com

John Mulligan

- phlogistonjohn@asynchrono.us
- jmulligan@redhat.com
  jmulliga@ibm.com
- http://asynchrono.us

