The CTDB Report

Martin Schwenke

<martin@meltin.net>

Samba Team · DDN

SambaXP 2026

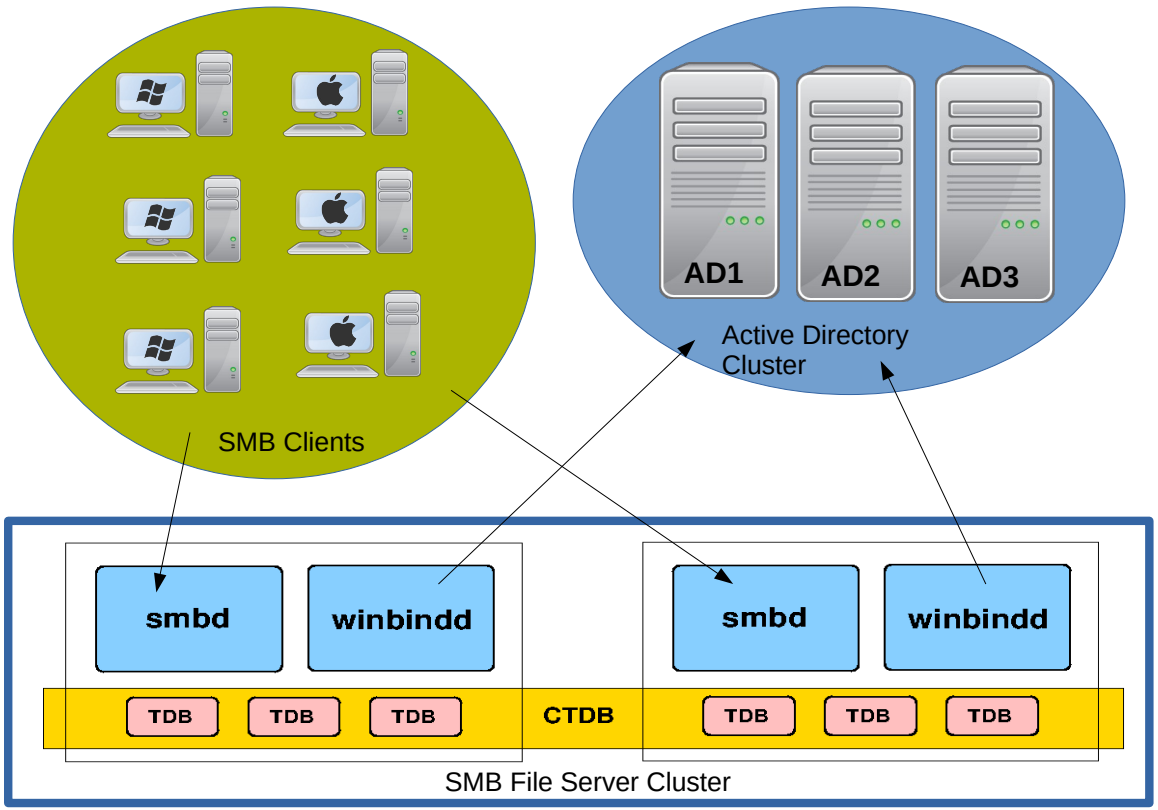
Overview

Audience

Everyone!

. . . a development-focused talk, but not just for developers

Clustered Samba



What is CTDB?

- Clustered database for Samba metadata
  - Distributed, volatile TDBs
  - Replicated, persistent TDBs
- Cluster-wide messaging transport
- Cluster management — leadership, membership
- Dynamic IP address failover
- Service management (`smbd`, `winbindd`, NFS, ...)

Progress

Commit Authors

90	Martin Schwenke
5	John Mulligan
3	Volker Lendecke
2	Peter Schwenke
2	Shachar Sharon
2	Stefan Metzmacher
1	Alexander Bokovoy
1	Andreas Schneider
106	

Commit Reviewers

42	Anoop C S
19	John Mulligan
18	Amitay Isaacs
12	Volker Lendecke
11	Martin Schwenke
9	Ralph Böhme
1	Alexander Bokovoy
112	

Commit Reviewers (all of Samba)

485	Anoop C S
307	Volker Lendecke
210	Ralph Böhme
168	Douglas Bagnall
162	Andreas Schneider
127	Stefan Metzmacher
114	Gary Lockyer
⋮	...
2090	

- Anoop rocks! Applause, please...
- There are many more reviewers, but not enough space here...

Main areas

- Failover on shutdown
- Load tunables from `/etc/ctdb/tunables.d/*.tunables`
- Limit `CTDB_SOCKET` usage: now only in `CTDB_TEST_MODE`
- Return event status for all scripts
- Fix some NFSv3 `SM_NOTIFY` issues
- JSON support for `ctdb status` output
- Replace numerous helper variables with `CTDB_TEST_HELPER_BINDIR`
- Support public IP addresses on interface `altname`'s

Requirement for failover on shutdown — NFS problem

- Client A has lock on file F held by NFS server on node X via public IP P
- Client B is blocked waiting for conflicting lock on F on node Y
- Shut down CTDB on node X:
  1. Release all public IP addresses
  2. Stop NFS server: lock on F is released on node X
  3. NFS server on node Y takes lock on F, grants lock to client B
  4. Another node notices X is down, starts failover
  5. P is assigned to node Z (could be Y!)
  6. Client A attempts to reclaim lock on F via NFS server on Z
  7. Lock reclaim fails...
  8. Application on A still thinks it has the lock
  9. Applications on A and B modify contended range in F
  10. :-(

Failover on shutdown — solution

- Need failover before shutdown completes to trigger NFS grace
- Also add configurable timeout between failover and continuing shutdown
- Ralph Böhme requested something similar 5 years ago for SMB durable file handles
- Note for later: adds a nested event loop, they're cool! ;-)

Load tunables from `/etc/ctdb/tunables.d/*.tunables`

- Historically `/etc/ctdb/tunables` (for example)
- As a vendor:
  - I want to put my settings in my own file...
  - ...while still allowing admins to add settings elsewhere
- Considered adding include files but:
  - Wildcards?
  - How deep?
  - Loop detection?
- Examples of `*.d/` directories without includes
  - `/etc/modprobe.d/`
  - `/etc/sysctl.d/`
- So: `/etc/ctdb/tunables.d/*.tunables`
- A safe, well-defined approach

Limit CTDB_SOCKET usage: now only in CTDB_TEST_MODE

- CTDB uses environment variables to set up test environment
- `chown()`, `chmod()` of socket reported as security issue
- `CTDB_SOCKET` not documented...
- So remove it!
- No, used in `clusteredmember` test environment
- Should use `fchown()`? Generally not supported on sockets...
- Limit `CTDB_SOCKET`: only use if `CTDB_TEST_MODE` is set
- Document that `CTDB_TEST_MODE` is only for testing...

Return event status for all scripts...

- Even those not run in latest iteration, due to prior script failure
- Requested by Ralph Böhme for CTDB Prometheus exporter
- Update some plumbling...
- Relevant API now returns results for all scripts
- `ctdb event status ...`
  - Unchanged
  - Only shows scripts from most recent iteration
  - Stops when timestamp in result list goes backwards

Fix some NFS SM_NOTIFY issues

- Regression introduced when `statd_callout.c` was split out
  - Wanted to add `sm-notify` forwarding...
  - `usage` message printed when processing `sm-notify`
  - Old `sm-notify` script used to "fail open"
    - Silently ignored unknown commands
  - So, add `sm-notify` no-op handling
  - Forwarding to come...
- On failover, notification generation would generate notifications from public IP addresses that hadn't moved
  - ...and expunge records from notification "database"
  - ...so, notification "database" records got lost
  - Only notify from public IP addresses that have moved
- Fixes by Peter Schwenke

JSON support for `ctdb status` output

- JSON output in addition to existing machine-readable format
- `ctdb status -j` or `ctdb status --json`
- Also `ctdb version -j` or `ctdb version --json`
  - Canary: does this version support JSON output?
- Gives error for commands not supporting `-j`/`--json`
- JSON comes to CTDB — just the beginning!
- Feature by John Mulligan with Shachar Sharon

Replace numerous helper variables with CTDB_TEST_HELPER_BINDIR

- CTDB uses environment variables to set up test environment
- CTDB has lots of helper programs
- 1 environment variable per helper!
- Not covered by CTDB_TEST_MODE
- Add CTDB_TEST_HELPER_BINDIR
  - One variable to rule them all...
  - ...and covered by CTDB_TEST_MODE
- Much simpler future
- More secure?

Support public IP addresses on interface `altnames`

- Useful for CTDB installation with centralised configuration
- Common `public_addresses` file for all nodes
- But interface names might be different across nodes
- `udev` rules?
- `ip link property add dev enp0s31f6 altname eth0`
  - Much easier for mortals!
- Works for all sorts of `ip ... dev eth0` commands...
- `ctdbd` change easy (use `if_nametoindex()`)...
- ...`releaseip` and `updateip` changes harder
- Can't post-match interface name when working with `altname`
- `$ ip -brief addr show to 192.168.1.123`
  `enp0s31f6 UP 192.168.1.123  ### no clue of altname`
- `$ ip -brief addr show to 192.168.1.123 dev eth0`
  `enp0s31f6 UP 192.168.1.123  ### eth0 matched!`

Miscellaneous

- use `server smb transports` (alternative to `smb ports`) and normalise transports to TCP ports (Stefan Metzmacher)
- Fix PCP PMDA build for newer versions (Alexander Bokovoy, Andreas Schneider)
- Fix a stuck cluster lock after delayed leader broadcast (Volker Lendecke)
- `updateip` fixes — constantly surprised this is used
- Fix CTDB startup with inconsistent cluster lock settings — fixes issue caused by recently added nested event loop (Volker Lendecke)
- Portability fixes: FreeBSD
- Script cleanups - `shfmt`, `shellcheck`, ...

Queue

Open Merge Requests

!4407 Avoid removing connections for released IP

- Some SMB multi-channel fixes in late 2023 broke TCP connection tracking for SMB
- Only noticed this year :-(
- Bug 15994

!4499 Add CTDB host monitoring

- Mostly useful for monitoring DNS servers
- Timeouts from DNS failures can cascade to:
  - AD discovery and connection
  - NFS client access checking by hostname
  - Any monitoring that depends on DNS
- Centralised monitoring is wonderful, but...
- What if `resolv.conf` updates are forgotten?

Coming soon?

ctdb-ipoib-arp ARP and Neighbour Advertisement (NA) for IPoIB

- Vinit Agnihotri wrote IPv4 ARP code years ago
- Not well integrated into the rest of the code
- Have reworked existing code and Vinit's patch
- How to construct the link-layer multicast address for IPv6 NA?

ctdb-leader-resignation Leader node should resign if not capable

- I have had this patch for 4 years
- Ralph Böhme noticed `ctdb stop` is slow
- Retried this, did not reliably improve speed
- Need to understand why and tweak...

Plans

Rewrite/future: transport

Next generation transport daemon

- ctdb-transportd: Written by Amitay Isaacs in 2019
- I fixed some bugs: it builds and runs...
- Noticed inter-node TCP connectivity is not yet implemented
- Need to write some tests for local connectivity
- Then add TCP connectivity
- Also tweak client code so new protocol can also be tunnelled through current `ctdbd`

Rewrite/future: JSON-RPC

The future is JSON?

- ctdb-queued-db: Queued DB implementation
- Extracted from my old ctdb-conntrack branch
- It queues adds/deletes, then dequeues to a persistent database every 0.1s (configurable)
- Tired of custom data formats and custom test harnesses!
- So, wrote some tests using JSON as the data format...
- I later discovered that I had nearly reinvented JSON-RPC
- JSON-RPC is simple, symmetric, elegant
- Wrote a `tevent_req`-based JSON-RPC handler computation
- Wrote test code for it...
- Rewrote my queued DB test code using it
- This could be protocol element 0!

Questions?