smbd, quo vadis?

Ralph Böhme, Samba Team, SerNet

2022-06-02

Outline

Recap of recent `smbd` changes

Overview of what's up our sleeves

Work in progress

Prototypes

Design

Outro

SerNet

Recap of recent `smbd` changes

Symlink race security fixes (CVE-2021-20316)

- Convert VFS to be fully handle and *at based
  - eg ensure any stat metadata returned to clients is either from `fstat(fd, ...)` or `fstatat(dirfd, name, AT_SYMLINK_NOFOLLOW, ...)`
  - `...` but not from `stat()`
  - rinse and repeat for all other syscalls!

Fixed with the release of Samba 4.15

- Lessons learnded: dealing with symlinks safely is huge problem, cf talks from Jeremy and Volker
- Caused a performance regression, also cf Volker's talk

SerNet

New RPC infrastructure

```
   2502       1 systemd
3282458    2502  \_ smbd
3282460 3282458  |   \_ smbd-notifyd
3282461 3282458  |   \_ cleanupd
3289690    2502  \_ samba-dcerpcd
```

`samba-dcerpcd`

- RPC server now runs in its own process
- By default automagically started on demand
- smbd passes RPC blobs over a local pipe
- Allows running just the RPC server and using it with other SMB server implementation, like ksmbd
- See the 4.16 release notes for an excellent summary
- Kudos to Volker

SerNet

SMB1 on the way out

New configure option `--with-smb1-server`

- Default is to build with SMB1 support
- Pass `--without-smb1-server` to drop it
- Kudos to Jeremy and David Mulder

SerNet

Other noteworthy changes

- Merged s3/s4 RPC server:
  - Adds support for async RPC, needed for eg Witness and MS-PAR
- SMB3 Multichannel no longer experimental
- Updated Heimdal, adds Kerberos FAST support
- Certificate Auto Enrollment
- Trusted domains are no longer scanned in `winbindd`
- Support for Offline Domain Join (ODJ)

SerNet

Overview of what's up our sleeves

What's up our sleeves

Work in progress

- Fileserver Performance Regresssion from CVE-2021-20316
- SMB3 UNIX Extensions

Prototypes

- Splice support for iouring VFS module
- Clustering: locking.tdb splitup
- SMB3 Directory Leases
- SMB3 Persistant Handles
- Witness RPC Service
- SMB-Direct
- s4u2self

Design

- Fileserver performance: DOS attributes xattr
- ksmbd Integration
- NFS Interop: oplocks, sharemodes
- ctdb

SerNet

Work in progress

Fileserver Performance Regression

Symlink race security fixes (CVE-2021-20316) caused performance regression

- affected releases: 4.14 and newer
- most dominant effect on metadata heavy workloads
- Setting `wide links = yes` doesn't help, expensive checks are still done
- Worst effect likely on virtual filesytem like vfs_glusterfs and FUSE based ones

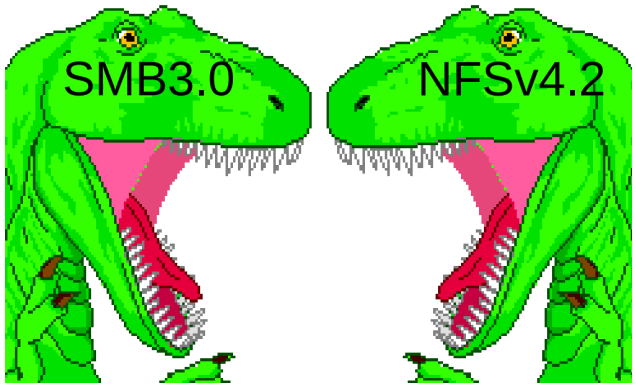
Version	Time
4.13 (baseline)	0m17.960s
4.15	0m19.968s
4.16	0m19.570s
master	0m17.677s

Table 1: `time rsync -a samba-4.16.1 /mnt/dir/`

- Some work being done, but as this is filesystem dependent I strongly encourage participation of vendors and distros
- For details please check out Volker's talk

SerNet

SMB3 UNIX Extensions



SMB3 UNIX Extensions:

- SMB3: great feature set, wide deployment
  - screaming performance with SMB-Direct
  - though no pNFS counterpart in SMB3
- SMB3 lacks UNIX extensions needed for Linux to Linux access.
- WIP branches from Jeremy and Volker
- Still agonizing over the best way to change low-level path processing `unix_convert()` function that would be responsible for catching symlinks in paths.
- Occasionally some team members kicks the WIP code around, commonly caused by work on other core smbd changes.

SerNet

SMB3 UNIX Extensions, cont.

Takeaway SMB3 UNIX Extensions:

- key feature to bring Samba into a good position to ...
  - conquer the datacenter
  - persistent storage for containers
  - cloud infrastructure
- low hanging fruit, rather weeks then months in development effort
- Little user, vendor or distro interest or funding, so expect slow progress

SerNet

Prototypes

Splice support for iouring VFS module

iouring: a new scalable AIO usespace API on Linux

- usecase: high performance fileserving
- Samba has VFS module that uses iouring
  - uses traditional socket to file fd buffer copying
- iouring can be used with `splice` and `sendmsg()`
  - avoids copying memory buffers
- Testing with a prototype over loopback got 35 Gbyte/s with 5 connections
- Currently the VFS module uses an unbound number of memory buffers that it preregisters with the kernel
  - Results in high memory usage under load
  - Need a way to limit the number and and free unused buffers

SerNet

Splice support for iouring VFS module, cont.

Takeaway:

- A must have feature for enterprise fileservers.
- Low hanging fruit, remaining work roughly a few weeks.
- Currently no user, vendor or distro funding.

SerNet

Clustering: locking.tdb splitup

`locking.tdb` **splitup:**

- Usecase: many clients open the same file on different Samba cluster nodes
  - 10000 client connect the same time in the morning and open `phonebook.exe`
- In the past this resulted in problems parsing and updating the record
  - Record contains an array of entries describing opens
  - Complete record was NDR encoded which turned out be slow for many entries
  - Got rewritten to only NDR encode the single array entry but hand marshall the array
  - Resulted in great speed up when opening many (thing thousands) handles on a single file

One problem remains on Clustered Samba:

- `locking.tdb` database record keeps bouncing across nodes as client on differing nodes request the record via ctdb

SerNet

Clustering: locking.tdb splitup, cont.

Solution:

- split up `locking.tdb` to store file open entries array record locally per node
- use additional record that stores a superset of the sharemodes
  - this record just has an array with an entry per node
  - record content becomes stable and content doesn't change
  - good candidate for ctdb read-only records

Takeaway:

- Big effort for a corner case
- But the resulting semantics may turn out to make a lot of sense
- Anyway, there's some notable vendor interest, work likely to continue in this area

SerNet

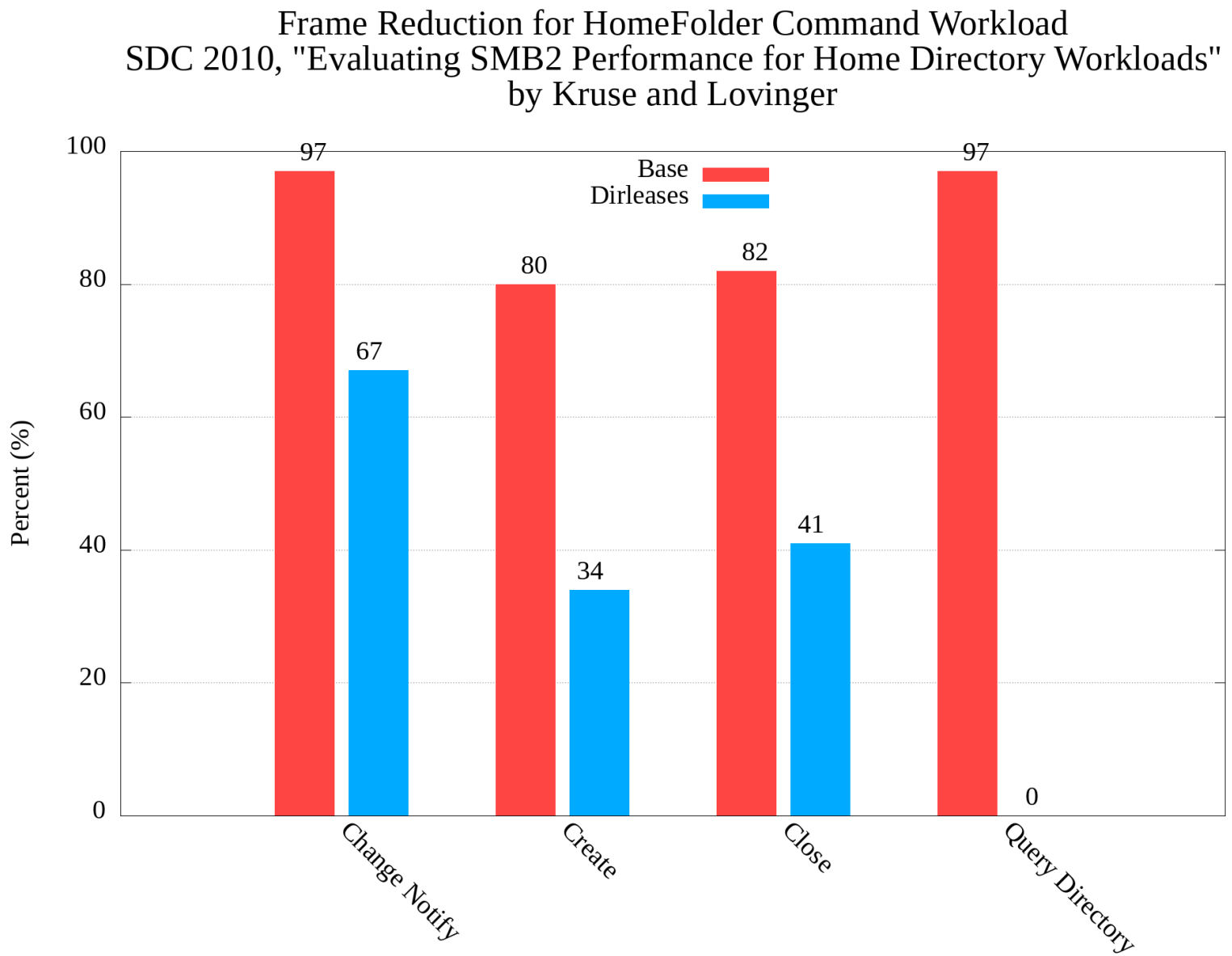
SMB3 Directory Leases

SMB3 Directory Leases:

- Allows a client to cache directory contents
- Supplements/replaces Change Notifications
- Reliable cache coherence
  - instead of best effort as with Change Notifications
- 100% packet reduction of `SMB2-QUERYDIR` packets for home folder workloads
- Effectiveness depends on workload, fewer changes result in better caching

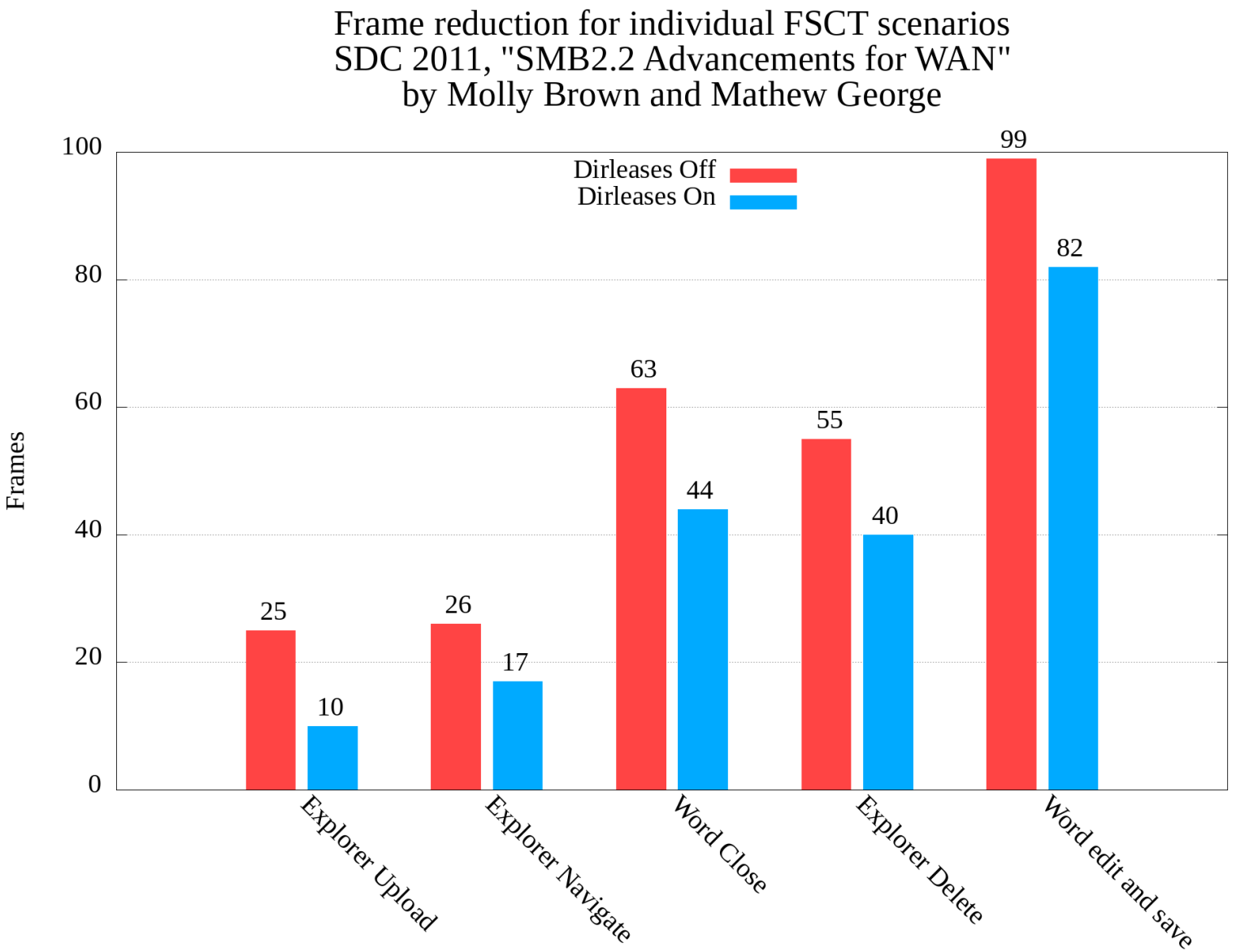
SerNet

SMB3 Directory Leases, HomeFolder Workload



SerNet

SMB3 Directory Leases, Misc. FSCT Scenarios



SerNet

SMB3 Directory Leases, Samba Status

Status:

- Basic functional prototype
- Not all operations that must trigger a cache/dirlease break are implemented yet
  - SetInfo levels `FSCTL_SET_ENCRYPTION`, `FileAllocationInformation`, `FileLinkInformation`, ...
- Passes Windows Protocol Testsuite
- Needs CI test coverage
- Completion depends on funding

SerNet

SMB3 Persistent Handles

Persistent Handles, key goals:

- SMB3 Persistent Handles part of Continuous Availability feature
  - Encompasses: SMB3 Multichannel, Witness RPC service, `SMB2_APP_INSTANCE_ID`, ...
- Server apps expect to be able to always access data on a continuously available share
  - Primary usecase: serving Hyper-V images over SMB3, MS-SQL
- Network or server failures are completely hidden from the application
  - Server guards access to file with Persistent Handles from disconnected clients
  - Filesystem client recovers disconnected handles and retries I/O
- Reliability on par with direct-attached storage / SAN

SerNet

SMB3 Persistent Handles, Samba Status

Samba Status:

- Basic prototype
  - implements SMB3 protocol changes, but that's the easy part
  - implements complex changes at our database layer
  - some CI tests
  - large patchset: some 140 patches
  - needs rebasing on current master (expect serious conflicts...)
  - still a lot to do...
  - cf my talk from SambaXP 2018
- Little interest from users, vendors or distros
- Needs funding, otherwise don't expect progress

SerNet

Witness RPC service

Witness?

- New RPC service to "witness" availability of other services, in particular SMB3
- Prompt and explicit notifications about failures in highly available systems
- Controlled way of dealing with reconnects instead of detecting failures due to timeouts
- Part of the larger Continous Availabilty feature
  - but not required for eg Persistent Handles

Status:

- Prototype from Günther Deschner
- Witness requires support for async RPC which we didn't have
  - prototype did some strange hacks
- Now that we have an async RPC server, this could move on
- Afaik currently noone working on it

SerNet

SMB-Direct

SMB-Direct usecase:

- File storage for the enterprise
- Minimal CPU utilisation and very low latency (server and client side)
- But keeps the traditional advantages of SMB file storage:
  - Ease of use, flexibility, choice of converged network, lower cost of networking infrastructure

SMB-Direct [MS-SMBD] is a simple transport layer

- Similar to TCP, over RDMA instead of IP
- Designed to serve SMB3 on top

What is RDMA?

- Remote Direct Memory Access
- Makes DMA possible over networks to remote peers
- Bypasses the operating system and its protocol stack
- Doesn't require any CPU interaction in order to do the transfer

SerNet

SMB-Direct, Samba Status

Status:

- Stefan Metzmacher started working on this as a side project
- No user/vendor/distro funding
- Large prototype consisting of several parts:
  1. New smbdirect.ko Linux kernel module
  2. Userspace API, smbd and smbclient are the consumers
  3. Changes to the smbd fileserver process using the API from 2
  4. Changes to the core SMB client library using the API from 2

Todo:

- Coordinate with Linux kernel developers on the right way to expose the UAPI
- Code is still immature, several weeks needed to get it to production quality
- ftrace tracepoints, standalone testsuite, Multichannel integration
- Integration into Samba's CI for automated tests
- For details see Stefan's SDC talk from 2018
- tldr: several months of development still needed
- Without funding expect little progress if at all

SerNet

Using S4U2Self in winbindd:

- winbindd provides group membership information of users
- commonly as nsswitch.conf provider for the system

Typically winbindd gets the Authorization Token via authentication

- Either via netr LogonSamLogon for NTLM
- Or via the "PAC Logon Info" element of the Kerberos service ticket

There're some situations when a service needs to impersonate a user locally:

- This can happen without getting an authentication for that user.
- SSH public-key authentication, sudo or nfs3 access are tyipical use cases.

Without S4U2SELF winbind tries to get the tokenGroups of the user via LDAP

- This doesn't always work
- The only reliable solution is S4U2Self

SerNet

S4U2SELF, Todo

The usage of S4U2Self with trusted domains/realms is complex:

- The example showed that a lot of transiting KDCs must be reached
- We should use site-aware KDCs (domain controllers) for all steps
- Currently winbindd prepares a custom krb5.conf
  - It fills in the KDC ip addresses for the default realm
  - But it's not possible to know all hops before calling krb5 functions
- It would be good if the Kerberos libraries would only do Kerberos
  - We can do (site-aware) DC lookups much more efficient.
  - It would be good to do the networking interaction on our own.
- MIT and Heimdal Kerberos libraries are missing features
- . . . the list goes on. . .
- See Stefan's talk from SambaXP 2020

SerNet

S4U2SELF, Key Takeaway

Key takeaway:

- Not needed for an SMB3 NAS box with Active Directory domain users
- Requires cooperation with Heimdal and MIT which has shown to be time consuming
- Summary: several months of complex coordination and development needed

SerNet

Design

Fileserver Performance cont.

Optimize DOS attributes storage in xattrs:

- Currently we always store one xattr per file/directory (`user.DOSATTRIB`)
- Reading that can be slow, especially on Fuse based filesystems

Previous solution: `smbd async dosmode`

- Parallelize xattr processing with a pthreadpool
- Accoding to customers the load is still too high

Nowadays we could use `stx_btime` and `STATX_ATTR_NODUMP` attribute

- Prerequisite: add support for `statx()`
- On the fly conversion of existing xattrs
- Would likely get rid of most xattrs with just ARCHIVE attribute and a creation date

SerNet

ksmbd Integration

ksmbd:

- Goal: low overhead, low footprint, performant fileserver
- Samba can achieve same performance by using iouring
- Current limitations: no domain support, no clustering, no VSS support, no Durable Handles, no Directory Leases and no Multi-Channel, no userspace pluggable VFS module support

Ideas:

- ksmbd can now make use of the new samba-dcerpcd for RPC
- Optionally use ksmbd instead of smbd

A lot of research and development needed on this end.

SerNet

Use case: file access by other means than SMB/Samba:

- NFS, ssh, FTP, ...
- for certain Windows features, Samba needs Linux kernel support

Oplocks / Leases:

- Samba can optionally use Linux kernel oplocks for Windows oplocks/leases
- Linux API: `fcntl(fd, F_SETLEASE, ...)`
- Limitations: shared oplocks are disabled, no upgrades and no lease keys
  - Linux will only grant shared oplock on files opened `O_RDONLY`, not `O_RDWR`
- Therefor only exclusive oplocks are granted by Samba if "kernel oplocks" are enabled
- leases are disabled as the semantics don't match at all with kernel oplocks

SMB/NFS Interop: cont.

Share modes:

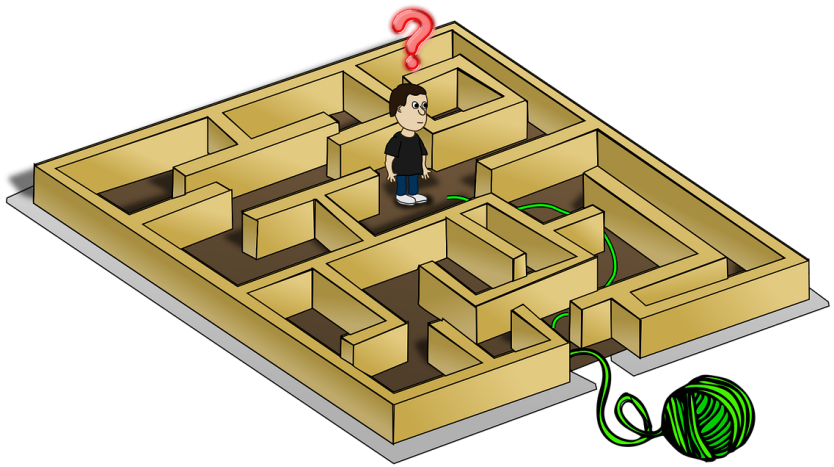
- Windows sharemodes are mandatory: `SHARE_[READ|WRITE|DELETE]`
- Sharemodes likely are Windows applications primary locking primitive
- Without kernel support for this, filesystem should be exported SMB only
- Linux kernel API `flock(fd, LOCK_MAND, ...)` is broken and will be removed from the kernel
- Samba just dropped support for using the API
  - VFS hook remains for filesystems that implement them via a filesystem specific hook (eg GPFS)

Takeaway:

- We need a big overhaul/redesign of the current kernel interfaces (`fcntl(fd, F_SETLEASE, ...)` and `flock(fd, LOCK_MAND, ...)`)
- A lot of research, brainstorming and coordination with the Linux kernel community is needed
- Without big vendors with deep ties into the Linux kernel community pushing this, this is not going to happen

SerNet

CTDB, which way to go?



The CTDB Report 2022 by Martin Schwenke, tomorrow, 4pm CEST

SerNet

Outro

Links

- Most code available at git.samba.org
- SDC 2018, Persistent Handles
- SDC 2017, SMB Direct Support within Samba and Linux
- Proposals
- Samba Roadmap

SerNet

Q&A

Thank you!
Questions?

Ralph Böhme
slow@samba.org
rb@sernet.de

SerNet