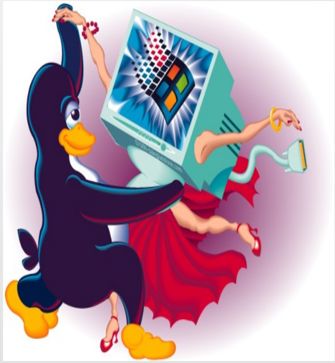
SambaXP 2024

Improving access to Remote Files from Linux: review of recent progress in the SMB3.1.1 client

Presented by Steve French

Principal Software Engineer

Microsoft Azure Storage



- This work represents the views of the author(s) and does not necessarily reflect the views of Microsoft Corporation
- Linux is a registered trademark of Linus Torvalds.
- Other company, product, and service names may be trademarks or service marks of others.

Who am I?

- Steve French smfrench@gmail.com
- Author and maintainer of Linux cifs vfs (for accessing Samba, Azure, Windows and various SMB3/CIFS based NAS appliances)
- Co-maintainer of the kernel server (ksmbd)
- Member of the Samba team (co-creator of the “net” utility)
- coauthor of SNIA CIFS Technical Reference, former SNIA CIFS Working Group chair
- Principal Software Engineer, Azure Storage: Microsoft

Outline

- Overview of Linux FS activity
- Recent ksmbd (server) improvements
- Recent client improvements
- Coming soon … what to look forward to
- Testing improvements

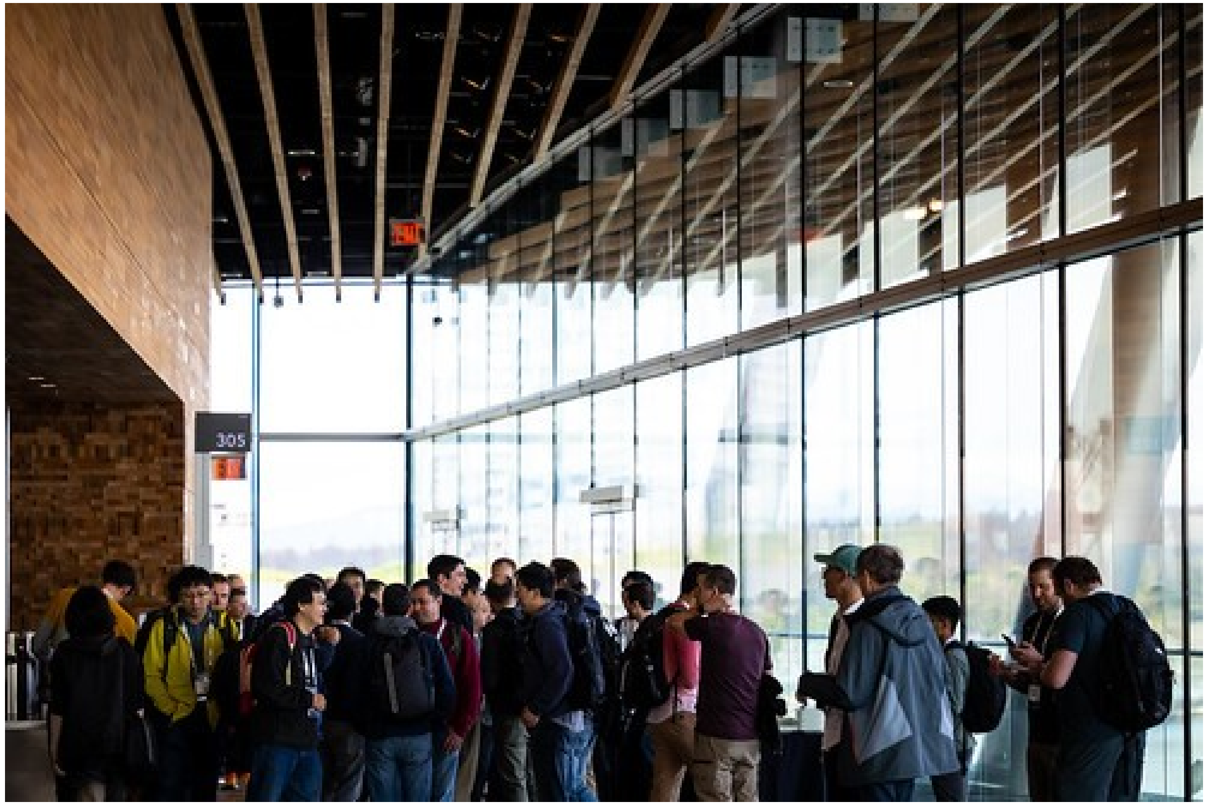
Linux Kernel: A year ago and now ...

- Now: 6.9-rc4 … then: 6.3-rc7

  But it has kept the same unusual name “Hurr durr I’ma ninja sloth”
- The kernel project is HUGE
- 87,735 changesets over that period
- 43,168 files changed
- 3,928,505 insertions (increased)
- 1,539,370 deletions

LSF/MM/eBPF summit

- Picture from May 2023 (looking forward to next, in Salt Lake City soon)
- Many excellent Linux storage developers working together

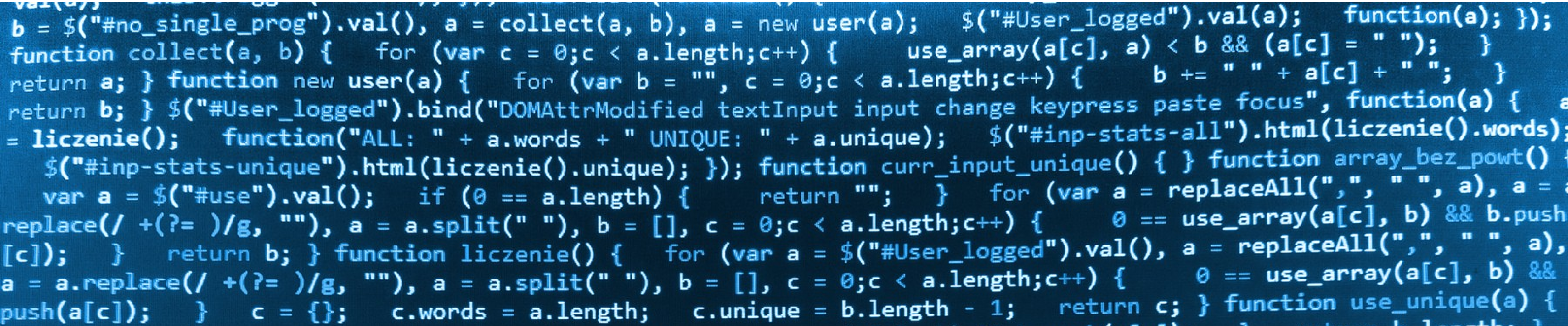


Some Linux FS topics of interest discussed recently

- Testing … testing … and more automated testing … (e.g. kdevops)
- Folios, netfs, iov_iter, variable size pages, and the redesign of page cache and offline (fscache), io_uring (async i/o improvements)
- Improvements to statx and fsinfo and to inotify/fanotify
- Idmapped mounts, fine grained timestamps
- Leveraging eBPF (not just dynamic tracing)
- Extending in kernel encryption: TLS handshake (for NFS) and QUIC (SMB3.1.1 and other)
- Shift to cloud
- Better support for faster storage (NVME) and net (RDMA/smbdirect)

Linux Filesystems Activity over past year (since 6.3-rc7)

- 8966 filesystems changesets (6.1% of total kernel changesets, carefully watched part of kernel, and FS activity is up 90%!)
- Linux kernel fs are 1.2 million lines of code (measured this week)
  - Some big advancements including
    - Bcachefs added, “ntfs classic” removed (in favor of NTFS3)
    - Ksmbd no longer “experimental”



```
b = $("#no_single_prog").val(), a = collect(a, b), a = new user(a);   $("#User_logged").val(a);   function(a); });
function collect(a, b) {   for (var c = 0;c < a.length;c++) {       use_array(a[c], a) < b && (a[c] = " ");   }
return a; } function new user(a) {   for (var b = "", c = 0;c < a.length;c++) {       b += " " + a[c] + " ";   }
return b; } $("#User_logged").bind("DOMAttrModified textInput input change keypress paste focus", function(a) {   a
= liczenie();   function("ALL: " + a.words + " UNIQUE: " + a.unique);   $("#inp-stats-all").html(liczenie().words);
  $("#inp-stats-unique").html(liczenie().unique); }); function curr_input_unique() { } function array_bez_powt() {
  var a = $("#use").val();   if (0 == a.length) {       return "";   }   for (var a = replaceAll(",", " ", a), a =
replace(/ +(?= )/g, ""), a = a.split(" "), b = [], c = 0;c < a.length;c++) {       0 == use_array(a[c], b) && b.push
[c]);   }   return b; } function liczenie() {   for (var a = $("#User_logged").val(), a = replaceAll(",", " ", a),
a = a.replace(/ +(?= )/g, ""), a = a.split(" "), b = [], c = 0;c < a.length;c++) {       0 == use_array(a[c], b) &&
push(a[c]);   }   c = {};   c.words = a.length;   c.unique = b.length - 1;   return c; } function use_unique(a) {
```

Most Active Linux Filesystems over the past year

- VFS (mapping layer) 567 changesets (activity up)
- The top filesystems and VFS dominate the activity
- Most active are bcachefs 3425 (!), BTRFS 1017 (down), XFS 609 (up a lot), ext4 402 (down)
- Most active non-local fs is SMB3.1.1 (cifs.ko) 333 (flat)
- Then NFSD (server) 276 (up) and NFS (client) 180 (down)
  - cifs.ko also had many more lines changed. It has been a VERY active year for cifs.ko
- Other: F2FS (272, down) GFS2 (179), ksmbd (153, flat), erofs (124), ceph (124), ntfs3 (116), ceph (105)

SMB3.1.1 Activity was strong this year

- cifs.ko activity was strong, 333 changesets
  - cifs is 63KLOC kernel code (not counting user space utilities)
- ksmbd activity flat
  - 26KLOC kernel code, 453 changesets since its introduction
- Samba server (userspace) is over 3.5 million lines of code (orders of magnitude bigger than the kernel smbd server or any of the NFS servers) and is even more active

Repeating our Goals for SMB3.1.1 on Linux

- Be the fastest, most secure general-purpose way to access file data, whether in the cloud or on premises or virtualized
  - Improve directory lease support
  - Keep improving compounding, multichannel
- Support more Linux/POSIX features – so apps don't know they run on SMB3 mounts (vs. local)
  - SMB3.1.1 POSIX extensions, new FSCTLs
  - Use xfstests to locate new features to emulate
- As Linux evolves, quickly add features to Linux kernel client and Samba and ksmbd
  - More test automation and keep adding more tests

Linux File API still growing (5 recently)

e.g.: “folios” and io_uring continue to improve as well, and lots of changes to internal APIs (netfs, fscache …)

Syscall name	Kernel Version introduced
fchmodat2	6.6-rc1
listmount (query children of mount)	6.8-rc1
statmount (query attributes of mnt)	6.8-rc1
And new fs ioctls:	
FS_IOC_GETUUID	6.9-rc1
FS_IOC_GETFSSYSFSPATH	6.9-rc1

One of the strengths of SMB3.1.1 is broad interop testing

- In-person plugfests are back!
- SMB3.1.1 plugfest restarted, colocated with SDC last two years
  - And will be again this year!
- Many exciting things being tested

Progress and Status update for Linux Kernel Server (ksmbd)

See Namjae Jeon (linkinjeon@kernel.org) talk, also at SambaXP

Improving the network stack: progress on QUIC and SMB3.1.1 for Linux

Very exciting progress in the experimental Linux kernel QUIC driver

See my earlier SambaXP talk

Recent improvements in the kernel client

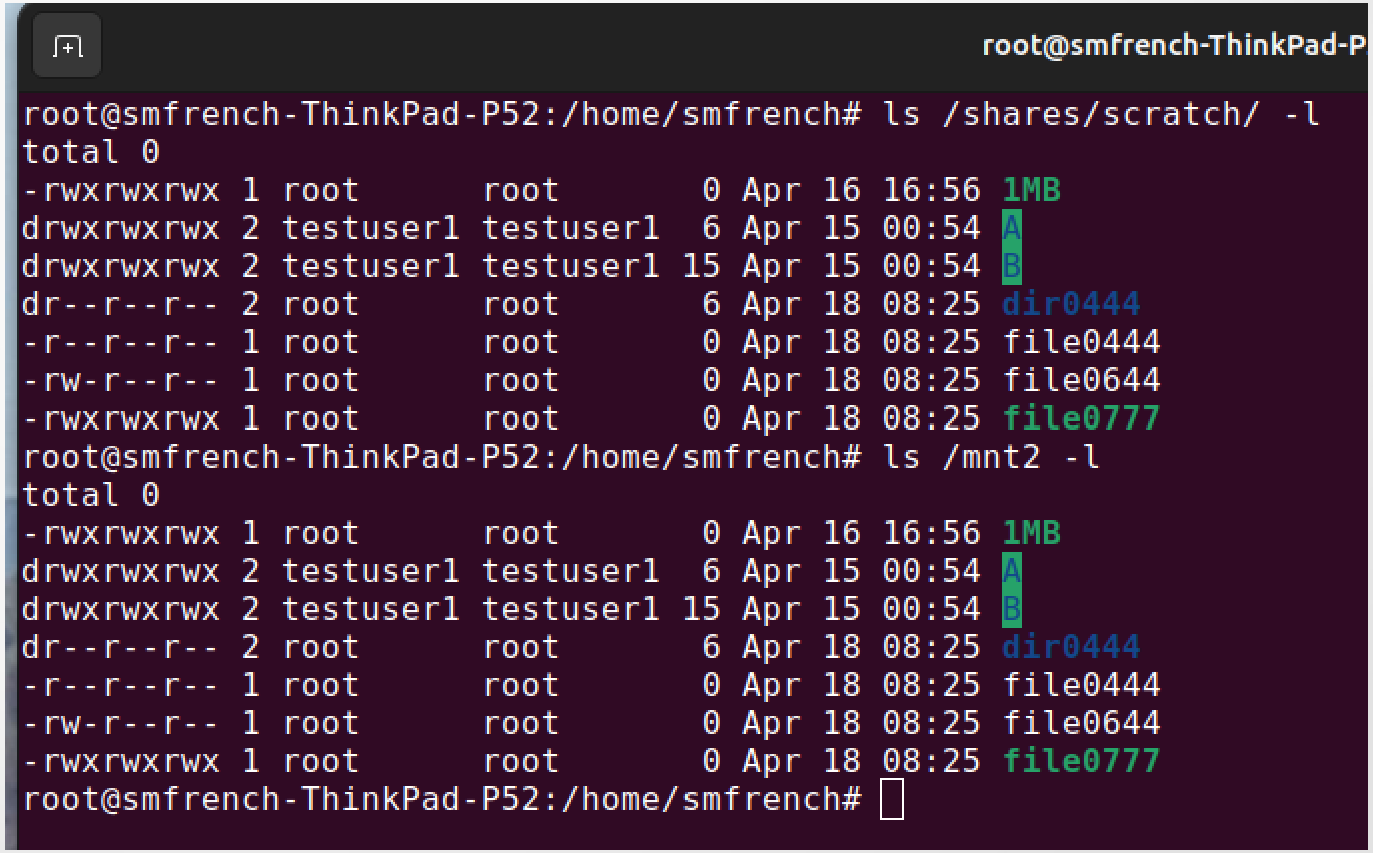
(cifs.ko)

DFS ie “The Global Namespace” - improvements

- DFS Interlink support (DFS link that points to another namespace)
- Sharing of DFS connections between mounts
- Reparse mount points - as cross fs boundaries on server - create
- submounts with noserverino
- Reconnection improved for DFS use cases
- Now handles cases where nested links have some bad targets (and can handle some scenarios which other clients can't)
- Thank you Paulo!

POSIX Extensions

- SMB3.1.1 POSIX/Linux extensions supported by ksmbd and Samba
- Details presented in earlier conferences
- Improved with recent enhancements for fifo, block, char, symlinks



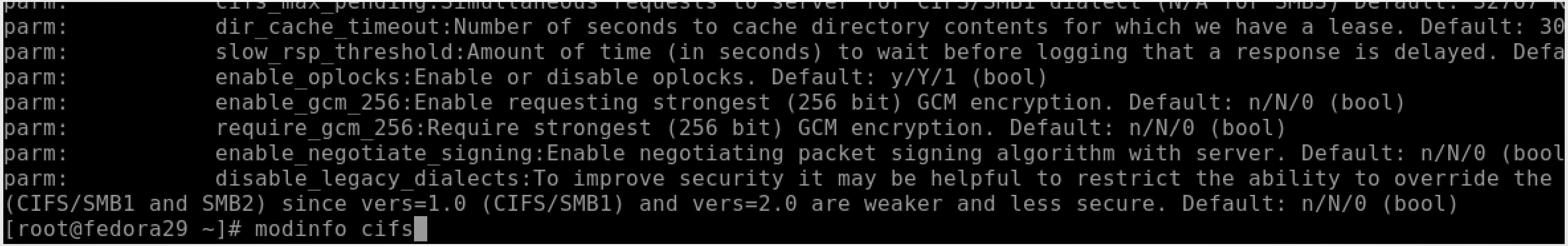
```
root@smfrench-ThinkPad-P52:/home/smfrench# ls /shares/scratch/ -l
total 0
-rwxrwxrwx 1 root       root        0 Apr 16 16:56 1MB
drwxrwxrwx 2 testuser1 testuser1  6 Apr 15 00:54 A
drwxrwxrwx 2 testuser1 testuser1 15 Apr 15 00:54 B
dr--r--r-- 2 root       root        6 Apr 18 08:25 dir0444
-r--r--r-- 1 root       root        0 Apr 18 08:25 file0444
-rw-r--r-- 1 root       root        0 Apr 18 08:25 file0644
-rwxrwxrwx 1 root       root        0 Apr 18 08:25 file0777
root@smfrench-ThinkPad-P52:/home/smfrench# ls /mnt2 -l
total 0
-rwxrwxrwx 1 root       root        0 Apr 16 16:56 1MB
drwxrwxrwx 2 testuser1 testuser1  6 Apr 15 00:54 A
drwxrwxrwx 2 testuser1 testuser1 15 Apr 15 00:54 B
dr--r--r-- 2 root       root        6 Apr 18 08:25 dir0444
-r--r--r-- 1 root       root        0 Apr 18 08:25 file0444
-rw-r--r-- 1 root       root        0 Apr 18 08:25 file0644
-rwxrwxrwx 1 root       root        0 Apr 18 08:25 file0777
root@smfrench-ThinkPad-P52:/home/smfrench# 
```

Multichannel improvements

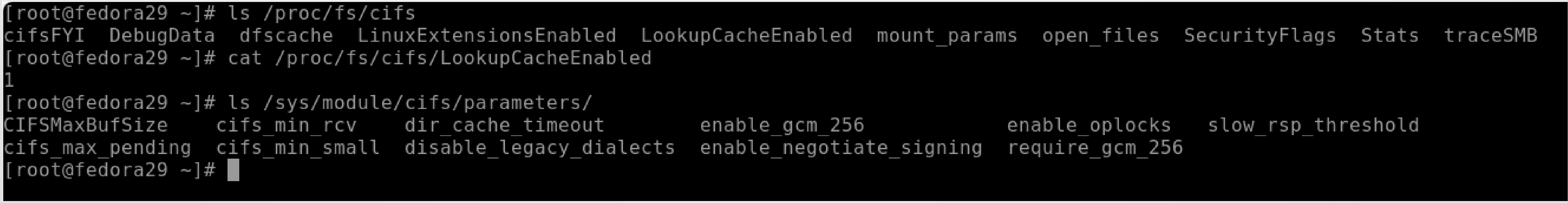
- Multiple Reconnect and Perf improvements including improved channel allocation for SMB3.1.1 requests (thank you Shyam Prasad)
- Soon will be enabled by default (when server supports multiple interfaces or RSS)

Directory and file caching improvements

- New module parm “dir_cache_timeout”
- Mount parms “max_cached_dirs=” and “handletimeout=”



```
parm:           cifs_max_pending:Simultaneous requests to server for CIFS/SMB1 dialect (N/A for SMB3) Default: 32767 N
parm:           dir_cache_timeout:Number of seconds to cache directory contents for which we have a lease. Default: 30
parm:           slow_rsp_threshold:Amount of time (in seconds) to wait before logging that a response is delayed. Defa
parm:           enable_oplocks:Enable or disable oplocks. Default: y/Y/1 (bool)
parm:           enable_gcm_256:Enable requesting strongest (256 bit) GCM encryption. Default: n/N/0 (bool)
parm:           require_gcm_256:Require strongest (256 bit) GCM encryption. Default: n/N/0 (bool)
parm:           enable_negotiate_signing:Enable negotiating packet signing algorithm with server. Default: n/N/0 (bool
parm:           disable_legacy_dialects:To improve security it may be helpful to restrict the ability to override the
(CIFS/SMB1 and SMB2) since vers=1.0 (CIFS/SMB1) and vers=2.0 are weaker and less secure. Default: n/N/0 (bool)
[root@fedora29 ~]# modinfo cifs
```



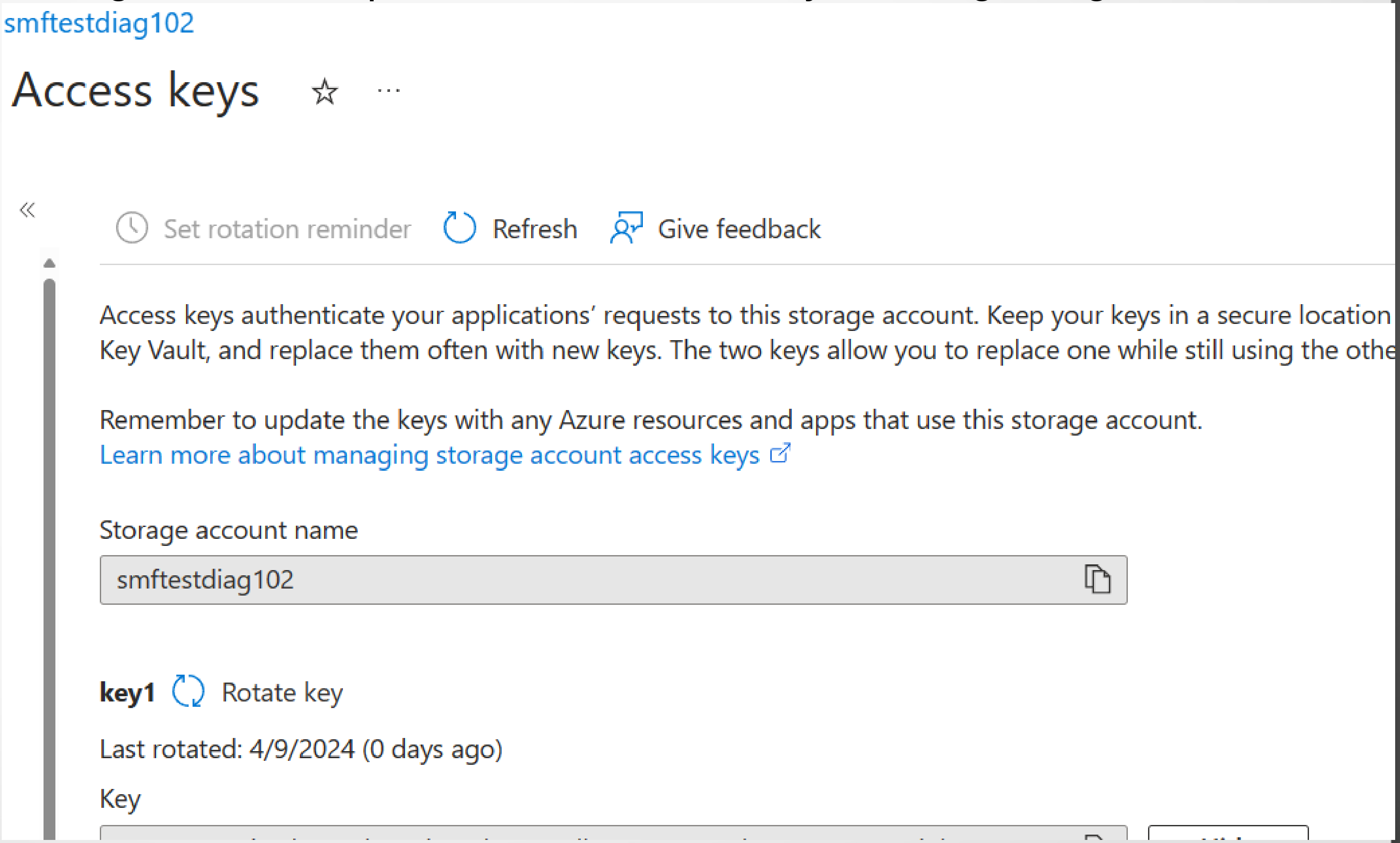
```
[root@fedora29 ~]# ls /proc/fs/cifs
cifsFYI  DebugData  dfscache  LinuxExtensionsEnabled  LookupCacheEnabled  mount_params  open_files  SecurityFlags  Stats  traceSMB
[root@fedora29 ~]# cat /proc/fs/cifs/LookupCacheEnabled
1
[root@fedora29 ~]# ls /sys/module/cifs/parameters/
CIFSMaxBufSize    cifs_min_rcv    dir_cache_timeout        enable_gcm_256            enable_oplocks   slow_rsp_threshold
cifs_max_pending  cifs_min_small  disable_legacy_dialects  enable_negotiate_signing  require_gcm_256
[root@fedora29 ~]# 
```

Improvements to special file support

- Much broader support for different types of special files (exported different ways by servers). Thank you Paulo!
- New mount parm “reparse=” allows you to choose whether reparse points (that encode special files like FIFOs, symlinks, block and char devices) should default to “wsl” format or the older Windows “nfs” server’s format.

password rotation (now can update on active mounts)

- Password rotation (key rotation) becoming common requirement due to security challenges e.g.



Access keys ☆ ···

Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location
Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the othe

Remember to update the keys with any Azure resources and apps that use this storage account.
Learn more about managing storage account access keys

Storage account name: smftestdiag102

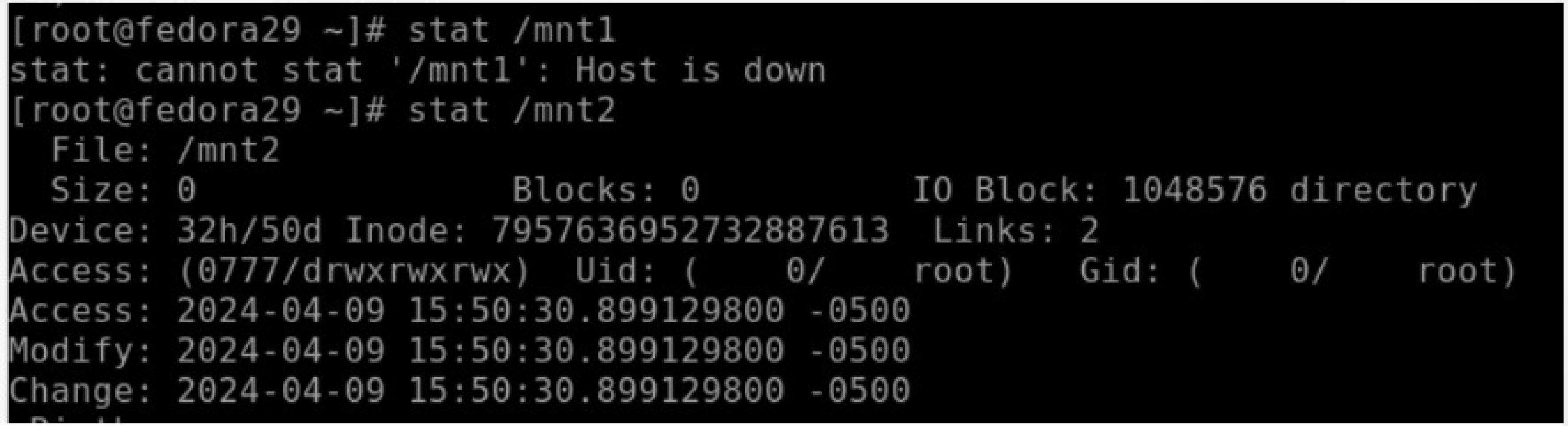
key1 Rotate key

Last rotated: 4/9/2024 (0 days ago)

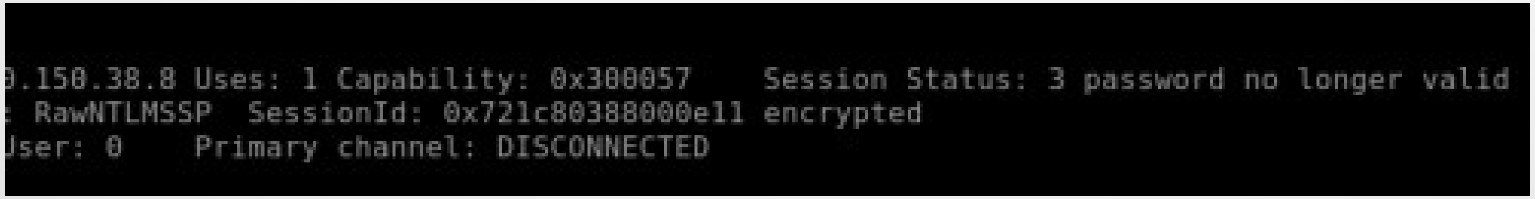
Key

password rotation (now can update on active mounts)

- If you had two mounts to the same server, one (“/mnt1”) with key 1 and one with key 2 (“/mnt2”), but then changed the first password (“rotate key”) but not the second then the first mount would be inaccessible (and before this change require “umount /mnt1” then “mount” again with new password) and DebugData would show DISCONNECTED (and password no longer valid)



```
[root@fedora29 ~]# stat /mnt1
stat: cannot stat '/mnt1': Host is down
[root@fedora29 ~]# stat /mnt2
  File: /mnt2
  Size: 0               Blocks: 0          IO Block: 1048576 directory
Device: 32h/50d Inode: 7957636952732887613  Links: 2
Access: (0777/drwxrwxrwx)  Uid: (    0/    root)   Gid: (    0/    root)
Access: 2024-04-09 15:50:30.899129800 -0500
Modify: 2024-04-09 15:50:30.899129800 -0500
Change: 2024-04-09 15:50:30.899129800 -0500
```

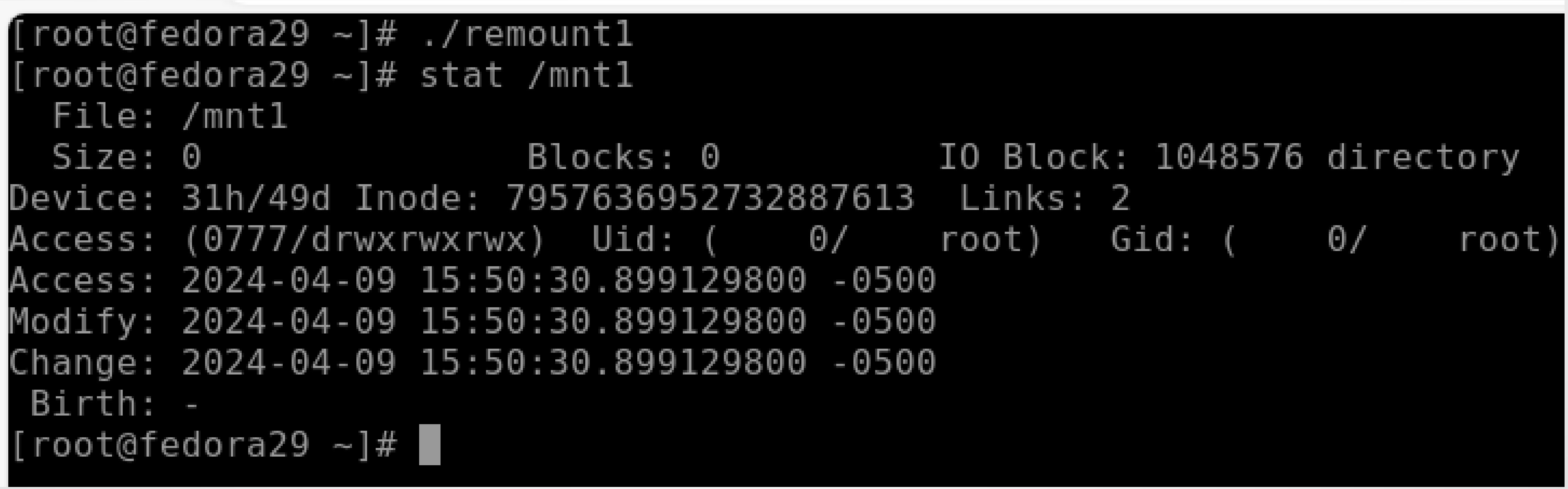


```
).150.38.8 Uses: 1 Capability: 0x300057    Session Status: 3 password no longer valid
: RawNTLMSSP  SessionId: 0x721c80388000e11 encrypted
Jser: 0    Primary channel: DISCONNECTED
```

password rotation (now can update on active mounts)

- Remount with new password (now in 6.9 kernel but also already backported to stable 6.8, 6.6 etc ...)

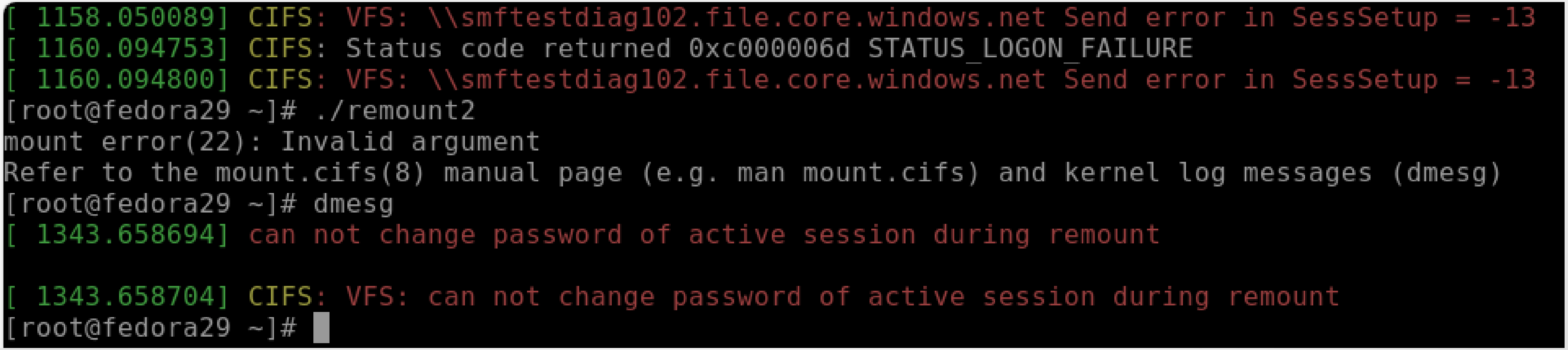
  (“mount -t cifs //server/share /mnt1 -o remount,password=newpassword”



```
[root@fedora29 ~]# ./remount1
[root@fedora29 ~]# stat /mnt1
  File: /mnt1
  Size: 0               Blocks: 0          IO Block: 1048576 directory
Device: 31h/49d Inode: 7957636952732887613  Links: 2
Access: (0777/drwxrwxrwx)  Uid: (    0/    root)   Gid: (    0/    root)
Access: 2024-04-09 15:50:30.899129800 -0500
Modify: 2024-04-09 15:50:30.899129800 -0500
Change: 2024-04-09 15:50:30.899129800 -0500
 Birth: -
[root@fedora29 ~]#
```

password rotation (now can update on active mounts)

- But remount of working mount (“/mnt2” in this example) with new (changed) password is not permitted since session not disconnected. Only allowed if server returned EACCESS or EKEYEXPIRED etc. and session is down, so we added a way to handle this with alt password (new mount parm “password2=”)
- “mount -t cifs //server/share /mnt2 -o remount,password2=newpassword”



```
[ 1158.050089] CIFS: VFS: \\smftestdiag102.file.core.windows.net Send error in SessSetup = -13
[ 1160.094753] CIFS: Status code returned 0xc000006d STATUS_LOGON_FAILURE
[ 1160.094800] CIFS: VFS: \\smftestdiag102.file.core.windows.net Send error in SessSetup = -13
[root@fedora29 ~]# ./remount2
mount error(22): Invalid argument
Refer to the mount.cifs(8) manual page (e.g. man mount.cifs) and kernel log messages (dmesg)
[root@fedora29 ~]# dmesg
[ 1343.658694] can not change password of active session during remount

[ 1343.658704] CIFS: VFS: can not change password of active session during remount
[root@fedora29 ~]# 
```

Features added for password rotation

- Until these recent changes:
  - Changing a password on a mount required “unmount” then “mount” again for Linux.
    - Which isn’t always practical if the app has open files on the mount or is hard to shutdown
- Added ability to change the password on remount (which doesn’t require apps to exit or files to close). Now can do:
    - mount –t cifs //srv/share /target –o remount, username=<user>,password=<newpassword>
    - Previously this would have returned “Invalid argument”
- Also added new mount option (also works on remount) “password2=“ so can have two passwords saved in session structure (and if reconnect fails with first one with access denied or key expired then is switched with “password2”
- Note that the “cifscreds” approach (used for cifs multiuser mounts in the non-krb5 case) which leverages the kernel keyring to save passwords instead of the session structure could not be used because the process doing the reconnect is not a child process of the mount process (or the process which launched cifscreds). The remount change and the “password2=“ approach use the normal session structure (in kernel memory but not in the kernel keyring) to store the password2 (where the “password” field is stored)
- The ability to remount with updated password is in Linux kernel now (6.9-rc1) and has been backported to 6.8 and 6.6 kernels (among others) already, and is expected to be broadly updated in distros over the coming months. The new mount option (password2) is in Linux starting with 6.9-rc4 but will be backported soon to some older kernels, and will request that it be picked up by other distros as well over the coming months.

Folios/Netfs/MM caching improvements

- Kernel multipage folios (supported by local fs and AFS so far, support pending for cifs.ko)
  - https://lwn.net/Articles/937239/
  - Also nice to see move to larger block sizes: https://lwn.net/Articles/945646/
- Merged so far (Thanks to David Howells!):
  - Move VM, pagecache & folios out of individual filesystems
  - Share code between 9p, afs, cifs, ceph
  - Filesystem just supplies read & write ops to/from iterators
  - Interleaving reads between local cache, servers

Folios/Netfs/MM caching improvements

- Currently being worked on:
  - David Howells has posted a large patch set for eval for 6.10 kernel
    - See the vfs.git vfs.netfs branch on kernel.org e.g. (in linux-next)
  - Better handling retries, and handling credit renegotiation
  - Support for write-through caching
  - Support for vectored writes
  - Support for content crypto (in devel)

6.3 kernel (April 23rd, 2023) (cifs module 2.42) very active release!

- Kernel idmapping improvements
- Improvements to use folios (better mm integration and cached writes)
- RDMA (smbdirect) improvements (thanks Metze and David)
- Many multichannel improvements (including using least loaded channel for sending I/O, and improvements for reconnect). Thanks Shyam!
- Various DFS fixes (thanks Paulo)
- Lower default deferred close timeout (from 5 to 1 second)

6.4 kernel (June 2023) (cifs.ko version: 2.43)

- Important deferred close (lease break corner case) fixes
- Reconnect and DFS fixes
- Important crediting improvements, and debugging improvements
- Perf improvement for large reads (rasize now defaults to 2x max rsize)
- At Linus’s suggestion source directories renamed
  - More intuitive
  - fs/cifs → fs/smb/client
  - fs/smbfs_common → fs/smb/common
  - fs/ksmbd → fs/smb/server

6.5 kernel (August 2023) (cifs.ko version: 2.44)

- Deferred close perf improvement (avoid unneeded lease break acks)
- Crediting (flow control) improvements to avoid low credit perf issue
- Reconnect and DFS fixes
- Fix null auth (sec=none) regression
- Allow dumping decryption keys (eg for reading network traces) via directory name, not just file.
- Directory caching improvement (‘laundromat’ thread added to clean up)
- Display client GUID and network namespace in /proc/fs/cifs/DebugData (can help with debugging containers e.g.)

6.6 kernel (October 2023) (cifs.ko version: 2.45)

- DFS (global namespace) fixes
- Improvements handling reparse points.
- Perf improvement for querying reparse point symlinks
- Reconnect improvement (write retry with channel sequence number)
- Add new mount parm "max_cached_dirs" to control how many directories are cached when server supports directory leases.
- Add new module parm /sys/module/cifs/parameters/dir_cache_timeout to control length of time a directory is cached with directory leases

6.7 kernel (Jan 2024) (cifs.ko version: 2.46)

- Reconnect and multichannel improvements
- Debugging improvements:
  - including client version in NTLM auth
  - Ioctl “CIFS_IOC_GET_TCON_INFO” to get sessid & tcon id of mnt
- Fallocate improvements for insert and zero range
- Fixes for metadata on server side symlinks (when reparse points)
- Very interesting OOB fixes found by Dr. Robert Morris with fuzzing

6.8 kernel (March 2024) (cifs.ko version: 2.47)

- netfs/folios (cached read/write) optimizations
- Compounding improvements
- Special file handling improvements (fifos, char/block, symlinks, and perf improvements for reparse points)
- Stats now show timestamp of when begun
- New mount option “retrans” (how often to retry on EAGAIN errors) and retry improvements
- Multichannel improvements

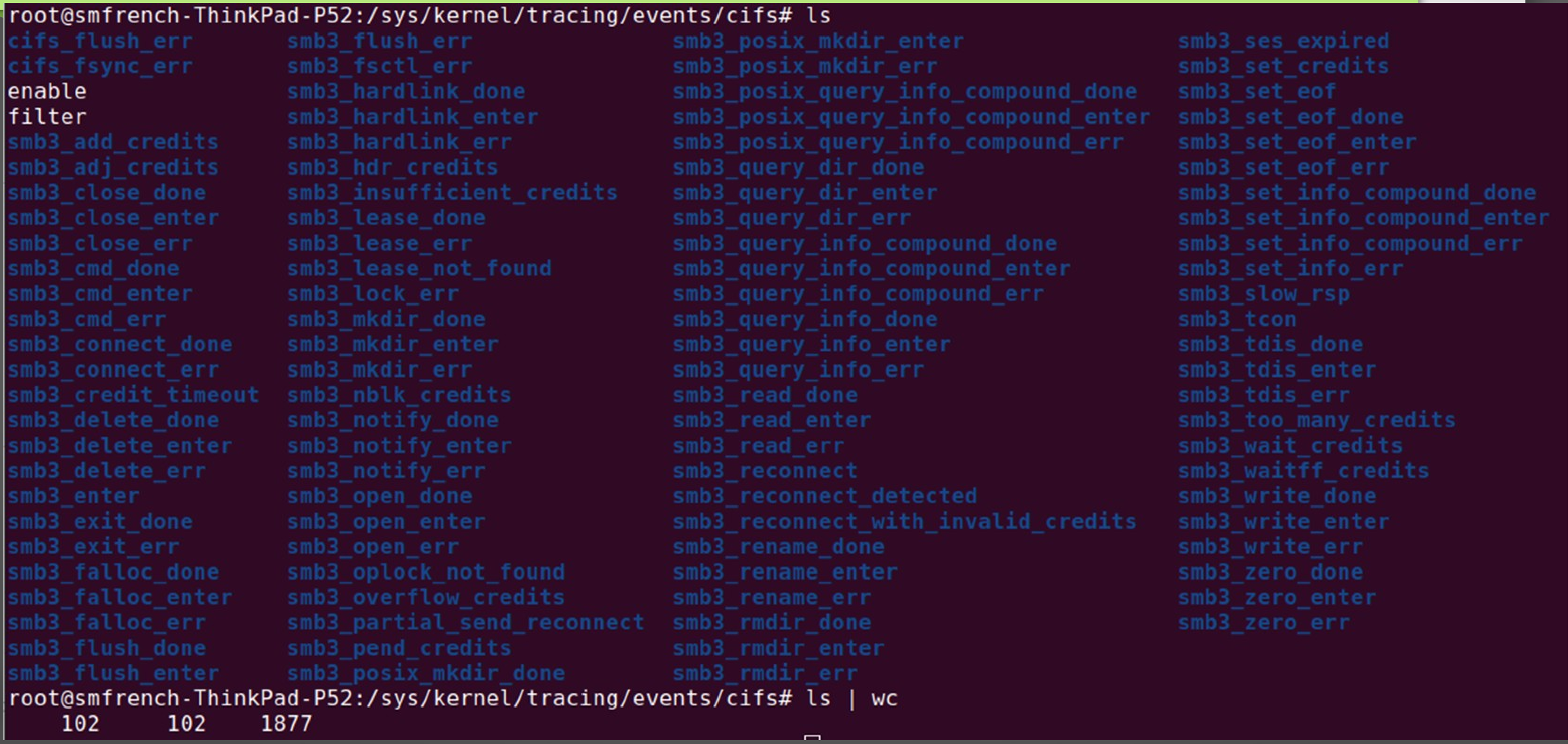
6.9 kernel (expected May 2024) (cifs.ko vers: 2.48)

- Allow updating password on remount (if session down due to password change on server, e.g. password rotation is getting common now)
- Allow alt password (“password2=”) on mount/remount to better handle key rotation e.g. plan for changed password rolled out over server(s)
- metadata caching improvements
- Add retry for some types of failed close operations
- Delete of open file improvements (do not defer close)
- Reparse mount option and add support for WSL reparse points
- Prep support for SMB3.1.1 compression over the wire (expect in 6.10)

Recent Debugging Improvements

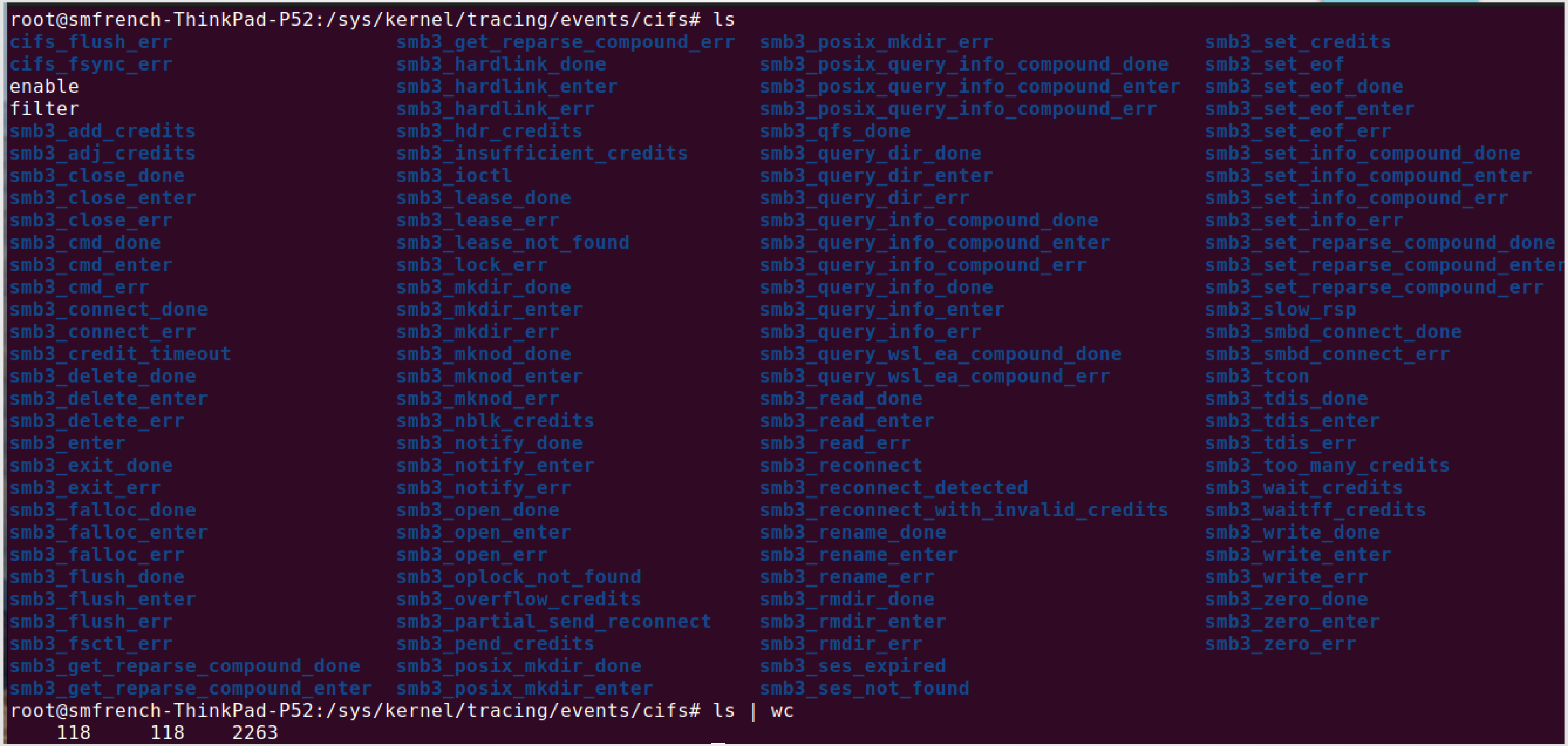
- More dynamic trace points added
- Ability to query the session id and tid for a mount
- Start time for stats visible (and reset when echo > /proc/fs/cifs/Stats)
- DebugData improvements:
  - Network namespace now shown for session (e.g. can help debug container reconnects)
  - Server capabilities now displayed
  - ClientGUID displayed
  - Unknown link (network) speed now shown properly

From 102 dynamic trace points last year at SambaXP ...



```
root@smfrench-ThinkPad-P52:/sys/kernel/tracing/events/cifs# ls
cifs_flush_err       smb3_flush_err               smb3_posix_mkdir_enter                smb3_ses_expired
cifs_fsync_err       smb3_fsctl_err               smb3_posix_mkdir_err                  smb3_set_credits
enable               smb3_hardlink_done           smb3_posix_query_info_compound_done   smb3_set_eof
filter               smb3_hardlink_enter          smb3_posix_query_info_compound_enter  smb3_set_eof_done
smb3_add_credits     smb3_hardlink_err            smb3_posix_query_info_compound_err    smb3_set_eof_enter
smb3_adj_credits     smb3_hdr_credits             smb3_query_dir_done                   smb3_set_eof_err
smb3_close_done      smb3_insufficient_credits    smb3_query_dir_enter                  smb3_set_info_compound_done
smb3_close_enter     smb3_lease_done              smb3_query_dir_err                    smb3_set_info_compound_enter
smb3_close_err       smb3_lease_err               smb3_query_info_compound_done         smb3_set_info_compound_err
smb3_cmd_done        smb3_lease_not_found         smb3_query_info_compound_enter        smb3_set_info_err
smb3_cmd_enter       smb3_lock_err                smb3_query_info_compound_err          smb3_slow_rsp
smb3_cmd_err         smb3_mkdir_done              smb3_query_info_done                  smb3_tcon
smb3_connect_done    smb3_mkdir_enter             smb3_query_info_enter                 smb3_tdis_done
smb3_connect_err     smb3_mkdir_err               smb3_query_info_err                   smb3_tdis_enter
smb3_credit_timeout  smb3_nblk_credits            smb3_read_done                        smb3_tdis_err
smb3_delete_done     smb3_notify_done             smb3_read_enter                       smb3_too_many_credits
smb3_delete_enter    smb3_notify_enter            smb3_read_err                         smb3_wait_credits
smb3_delete_err      smb3_notify_err              smb3_reconnect                        smb3_waitff_credits
smb3_enter           smb3_open_done               smb3_reconnect_detected               smb3_write_done
smb3_exit_done       smb3_open_enter              smb3_reconnect_with_invalid_credits   smb3_write_enter
smb3_exit_err        smb3_open_err                smb3_rename_done                      smb3_write_err
smb3_falloc_done     smb3_oplock_not_found        smb3_rename_enter                     smb3_zero_done
smb3_falloc_enter    smb3_overflow_credits        smb3_rename_err                       smb3_zero_enter
smb3_falloc_err      smb3_partial_send_reconnect  smb3_rmdir_done                       smb3_zero_err
smb3_flush_done      smb3_pend_credits            smb3_rmdir_enter
smb3_flush_enter     smb3_posix_mkdir_done        smb3_rmdir_err
root@smfrench-ThinkPad-P52:/sys/kernel/tracing/events/cifs# ls | wc
    102     102    1877
```

To 118 now (16 more eBPF trace points added)



```
root@smfrench-ThinkPad-P52:/sys/kernel/tracing/events/cifs# ls
cifs_flush_err                   smb3_get_reparse_compound_err  smb3_posix_mkdir_err                  smb3_set_credits
cifs_fsync_err                   smb3_hardlink_done             smb3_posix_query_info_compound_done   smb3_set_eof
enable                           smb3_hardlink_enter            smb3_posix_query_info_compound_enter  smb3_set_eof_done
filter                           smb3_hardlink_err              smb3_posix_query_info_compound_err    smb3_set_eof_enter
smb3_add_credits                 smb3_hdr_credits               smb3_qfs_done                         smb3_set_eof_err
smb3_adj_credits                 smb3_insufficient_credits      smb3_query_dir_done                   smb3_set_info_compound_done
smb3_close_done                  smb3_ioctl                     smb3_query_dir_enter                  smb3_set_info_compound_enter
smb3_close_enter                 smb3_lease_done                smb3_query_dir_err                    smb3_set_info_compound_err
smb3_close_err                   smb3_lease_err                 smb3_query_info_compound_done         smb3_set_info_err
smb3_cmd_done                    smb3_lease_not_found           smb3_query_info_compound_enter        smb3_set_reparse_compound_done
smb3_cmd_enter                   smb3_lock_err                  smb3_query_info_compound_err          smb3_set_reparse_compound_enter
smb3_cmd_err                     smb3_mkdir_done                smb3_query_info_done                  smb3_set_reparse_compound_err
smb3_connect_done                smb3_mkdir_enter               smb3_query_info_enter                 smb3_slow_rsp
smb3_connect_err                 smb3_mkdir_err                 smb3_query_info_err                   smb3_smbd_connect_done
smb3_credit_timeout              smb3_mknod_done                smb3_query_wsl_ea_compound_done       smb3_smbd_connect_err
smb3_delete_done                 smb3_mknod_enter               smb3_query_wsl_ea_compound_err        smb3_tcon
smb3_delete_enter                smb3_mknod_err                 smb3_read_done                        smb3_tdis_done
smb3_delete_err                  smb3_nblk_credits              smb3_read_enter                       smb3_tdis_enter
smb3_enter                       smb3_notify_done               smb3_read_err                         smb3_tdis_err
smb3_exit_done                   smb3_notify_enter              smb3_reconnect                        smb3_too_many_credits
smb3_exit_err                    smb3_notify_err                smb3_reconnect_detected               smb3_wait_credits
smb3_falloc_done                 smb3_open_done                 smb3_reconnect_with_invalid_credits   smb3_waitff_credits
smb3_falloc_enter                smb3_open_enter                smb3_rename_done                      smb3_write_done
smb3_falloc_err                  smb3_open_err                  smb3_rename_enter                     smb3_write_enter
smb3_flush_done                  smb3_oplock_not_found          smb3_rename_err                       smb3_write_err
smb3_flush_enter                 smb3_overflow_credits          smb3_rmdir_done                       smb3_zero_done
smb3_flush_err                   smb3_partial_send_reconnect    smb3_rmdir_enter                      smb3_zero_enter
smb3_fsctl_err                   smb3_pend_credits              smb3_rmdir_err                        smb3_zero_err
smb3_get_reparse_compound_done   smb3_posix_mkdir_done          smb3_ses_expired
smb3_get_reparse_compound_enter  smb3_posix_mkdir_enter         smb3_ses_not_found
root@smfrench-ThinkPad-P52:/sys/kernel/tracing/events/cifs# ls | wc
    118     118    2263
```

Planned and in-progress Improvements (To-dos)

(cifs.ko)

What improvements to expect in next few releases?

- Prototype of SMB3.1.1 over QUIC (new encrypted network transport)
- Reenabling support for swapfile over SMB3.1.1 mounts (was delayed due to MM/folios/netfs changes)
- Support for creating with O_TMPFILE
- More testing of SMB3.1.1 mounts to Samba with the POSIX extensions
- Support for new auth mechanisms (e.g. local KDC, IAKERB)
- Support for SMB3.1.1 compression (some of Enzo’s series has already been merged into mainline). See slides 48-55 of SNIA-SDC23-French-Advancing-Accessing-Remote-Files-SMB3_2.pdf
- Better use of parent lease key to avoid unneeded lease breaks

Anyone interested in the fastest transports ...?

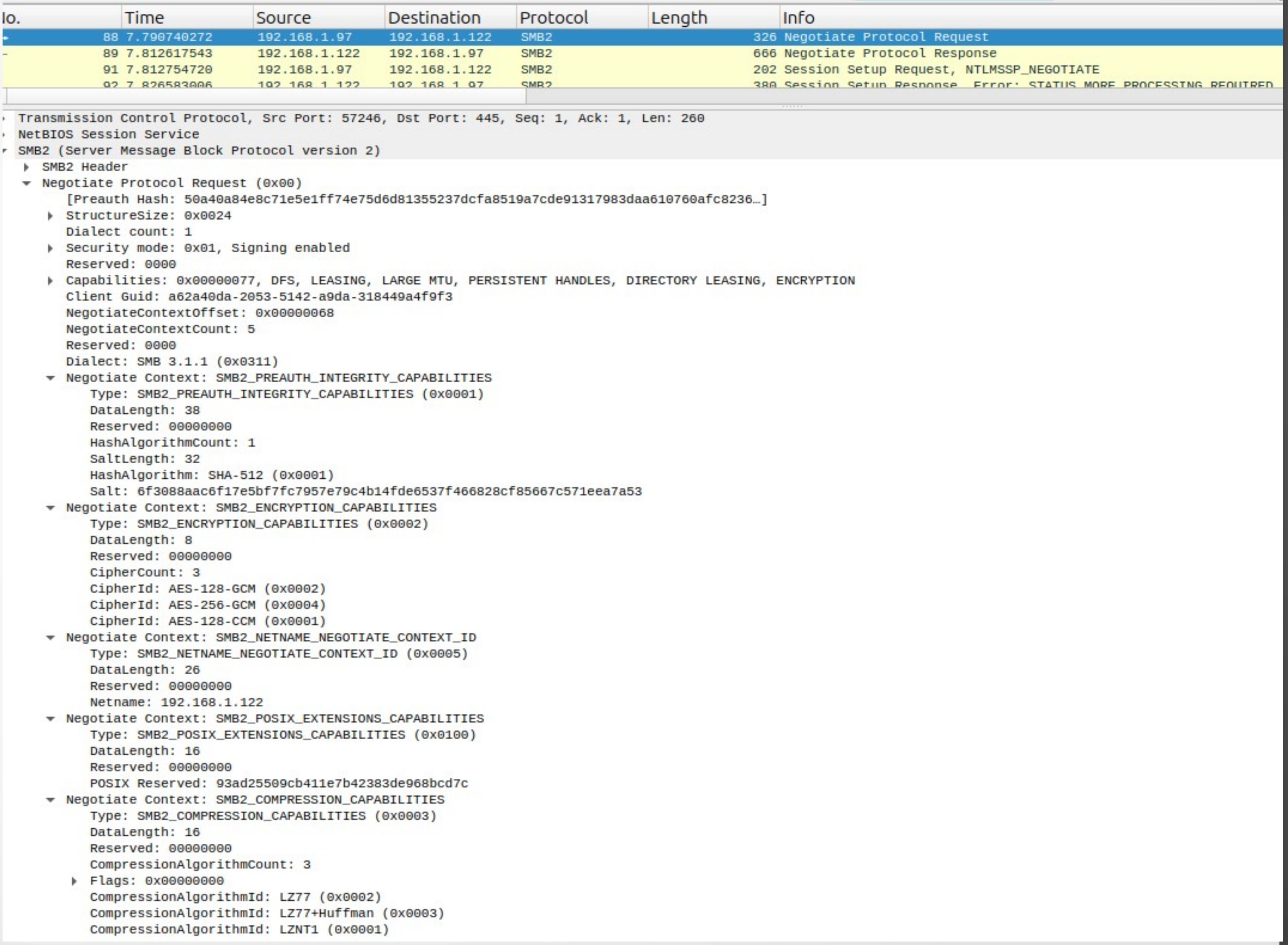
- Smbdirect (RDMA abstraction layer) is very powerful
- This could be available as a common code layer used by Samba, ksmbd server and cifs.ko client and other applications
- Tom Talpey and Metze had some great ideas (some described in my 2023 SDC presentation) on how to improve smbdirect (RDMA) for SMB3.1.1 mounts
- Also important is improving test automation for RDMA (smbdirect)
  - Ideas welcome

What about new Linux syscalls and Linux extensions

- Can consider new FSCTLs and infolevels to deal with the new Linux features, syscalls, fallocate flags etc.
- Xfstests help discover new Linux features that are missing

Work on SMB3.1.1 compression over network

- Can now negotiate compression algorithms with the server
- Will be big help in long latency networks
- Enzo is working on additional patches. See my SDC2023 presentation pages 48-52 for details

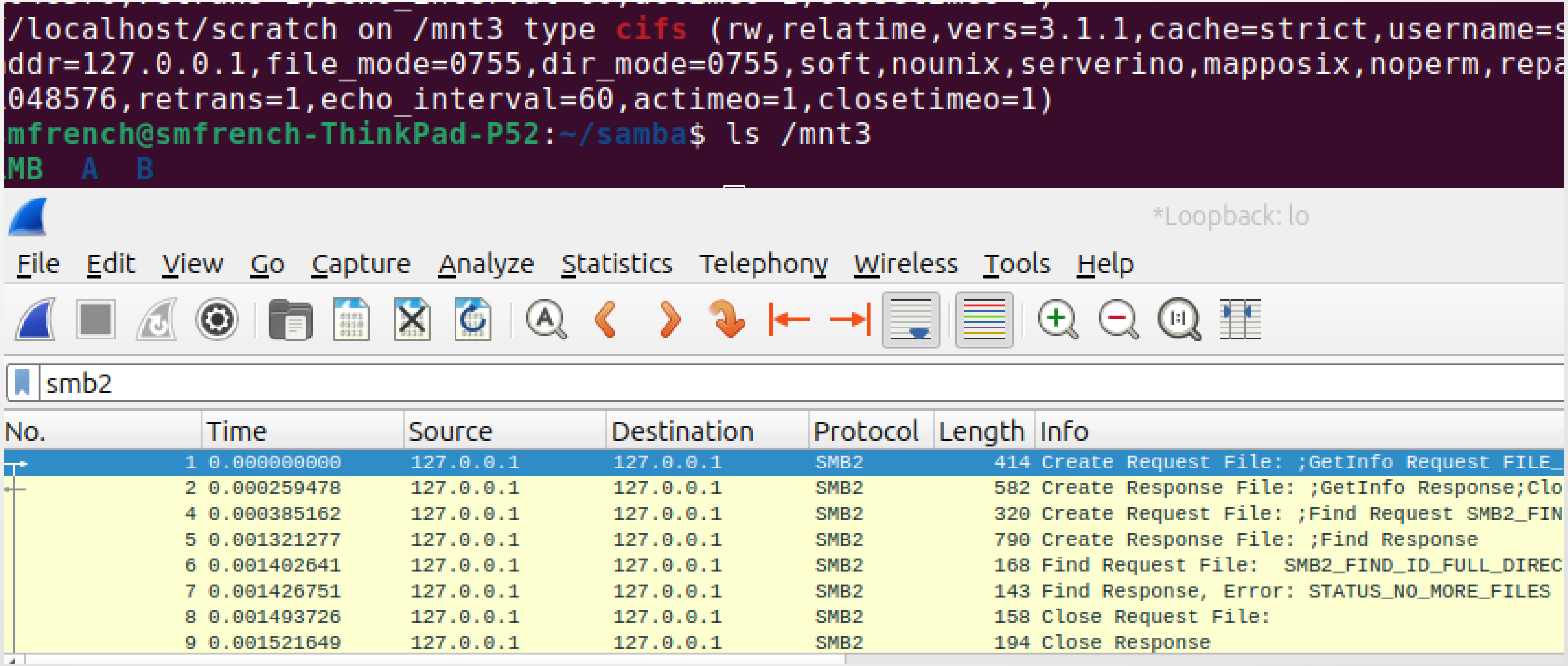


```
No.    Time            Source          Destination     Protocol   Length  Info
    88 7.790740272     192.168.1.97    192.168.1.122   SMB2          326  Negotiate Protocol Request
    89 7.812617543     192.168.1.122   192.168.1.97    SMB2          666  Negotiate Protocol Response
    91 7.812754720     192.168.1.97    192.168.1.122   SMB2          202  Session Setup Request, NTLMSSP_NEGOTIATE
    92 7.826583006     192.168.1.122   192.168.1.97    SMB2          380  Session Setup Response, Error: STATUS_MORE_PROCESSING_REQUIRED
```

```
▸ Transmission Control Protocol, Src Port: 57246, Dst Port: 445, Seq: 1, Ack: 1, Len: 260
▸ NetBIOS Session Service
▾ SMB2 (Server Message Block Protocol version 2)
  ▸ SMB2 Header
  ▾ Negotiate Protocol Request (0x00)
      [Preauth Hash: 50a40a84e8c71e5e1ff74e75d6d81355237dcfa8519a7cde91317983daa610760afc8236…]
    ▸ StructureSize: 0x0024
      Dialect count: 1
    ▸ Security mode: 0x01, Signing enabled
      Reserved: 0000
    ▸ Capabilities: 0x00000077, DFS, LEASING, LARGE MTU, PERSISTENT HANDLES, DIRECTORY LEASING, ENCRYPTION
      Client Guid: a62a40da-2053-5142-a9da-318449a4f9f3
      NegotiateContextOffset: 0x00000068
      NegotiateContextCount: 5
      Reserved: 0000
      Dialect: SMB 3.1.1 (0x0311)
    ▾ Negotiate Context: SMB2_PREAUTH_INTEGRITY_CAPABILITIES
        Type: SMB2_PREAUTH_INTEGRITY_CAPABILITIES (0x0001)
        DataLength: 38
        Reserved: 00000000
        HashAlgorithmCount: 1
        SaltLength: 32
        HashAlgorithm: SHA-512 (0x0001)
        Salt: 6f3088aac6f17e5bf7fc7957e79c4b14fde6537f466828cf85667c571eea7a53
    ▾ Negotiate Context: SMB2_ENCRYPTION_CAPABILITIES
        Type: SMB2_ENCRYPTION_CAPABILITIES (0x0002)
        DataLength: 8
        Reserved: 00000000
        CipherCount: 3
        CipherId: AES-128-GCM (0x0002)
        CipherId: AES-256-GCM (0x0004)
        CipherId: AES-128-CCM (0x0001)
    ▾ Negotiate Context: SMB2_NETNAME_NEGOTIATE_CONTEXT_ID
        Type: SMB2_NETNAME_NEGOTIATE_CONTEXT_ID (0x0005)
        DataLength: 26
        Reserved: 00000000
        Netname: 192.168.1.122
    ▾ Negotiate Context: SMB2_POSIX_EXTENSIONS_CAPABILITIES
        Type: SMB2_POSIX_EXTENSIONS_CAPABILITIES (0x0100)
        DataLength: 16
        Reserved: 00000000
        POSIX Reserved: 93ad25509cb411e7b42383de968bcd7c
    ▾ Negotiate Context: SMB2_COMPRESSION_CAPABILITIES
        Type: SMB2_COMPRESSION_CAPABILITIES (0x0003)
        DataLength: 16
        Reserved: 00000000
        CompressionAlgorithmCount: 3
      ▸ Flags: 0x00000000
        CompressionAlgorithmId: LZ77 (0x0002)
        CompressionAlgorithmId: LZ77+Huffman (0x0003)
        CompressionAlgorithmId: LZNT1 (0x0001)
```

Optimizing common cases: improve “ls” (readdir)

- Current behavior wastes a roundtrip for small directories
  - Extra Find Request (frames 6 & 7) could be compounded



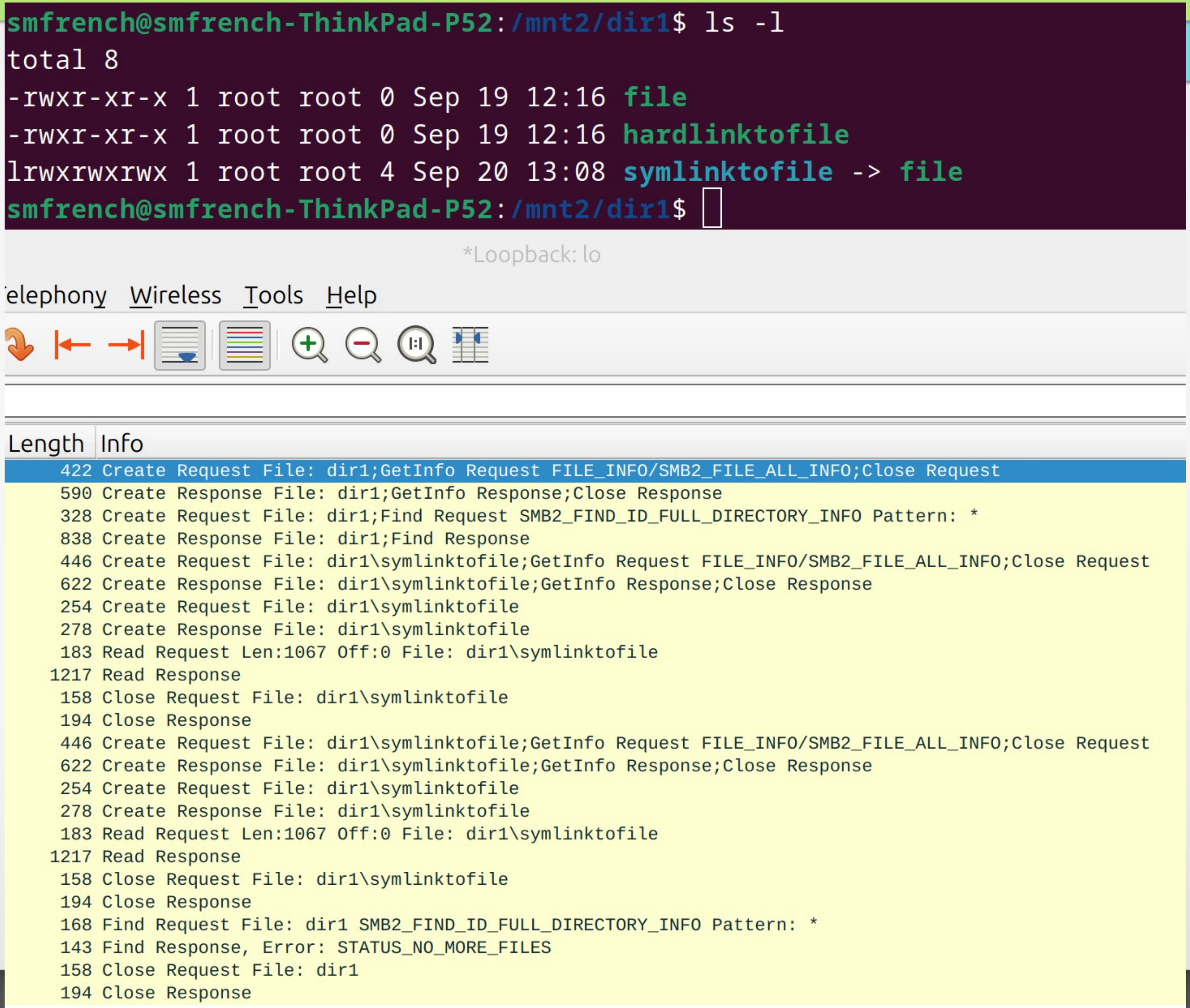
```
/localhost/scratch on /mnt3 type cifs (rw,relatime,vers=3.1.1,cache=strict,username=s
ddr=127.0.0.1,file_mode=0755,dir_mode=0755,soft,nounix,serverino,mapposix,noperm,repa
048576,retrans=1,echo_interval=60,actimeo=1,closetimeo=1)
mfrench@smfrench-ThinkPad-P52:~/samba$ ls /mnt3
MB  A  B
```

smb2

```
No.        Time          Source        Destination   Protocol  Length Info
         1 0.000000000   127.0.0.1     127.0.0.1     SMB2         414 Create Request File: ;GetInfo Request FILE_
         2 0.000259478   127.0.0.1     127.0.0.1     SMB2         582 Create Response File: ;GetInfo Response;Clo
         4 0.000385162   127.0.0.1     127.0.0.1     SMB2         320 Create Request File: ;Find Request SMB2_FIN
         5 0.001321277   127.0.0.1     127.0.0.1     SMB2         790 Create Response File: ;Find Response
         6 0.001402641   127.0.0.1     127.0.0.1     SMB2         168 Find Request File:  SMB2_FIND_ID_FULL_DIREC
         7 0.001426751   127.0.0.1     127.0.0.1     SMB2         143 Find Response, Error: STATUS_NO_MORE_FILES
         8 0.001493726   127.0.0.1     127.0.0.1     SMB2         158 Close Request File:
         9 0.001521649   127.0.0.1     127.0.0.1     SMB2         194 Close Response
```

Optimizing common cases: improve “ls -l” when symlinks

- Current behavior with “mfsymlinks” (note repeated read e.g.)

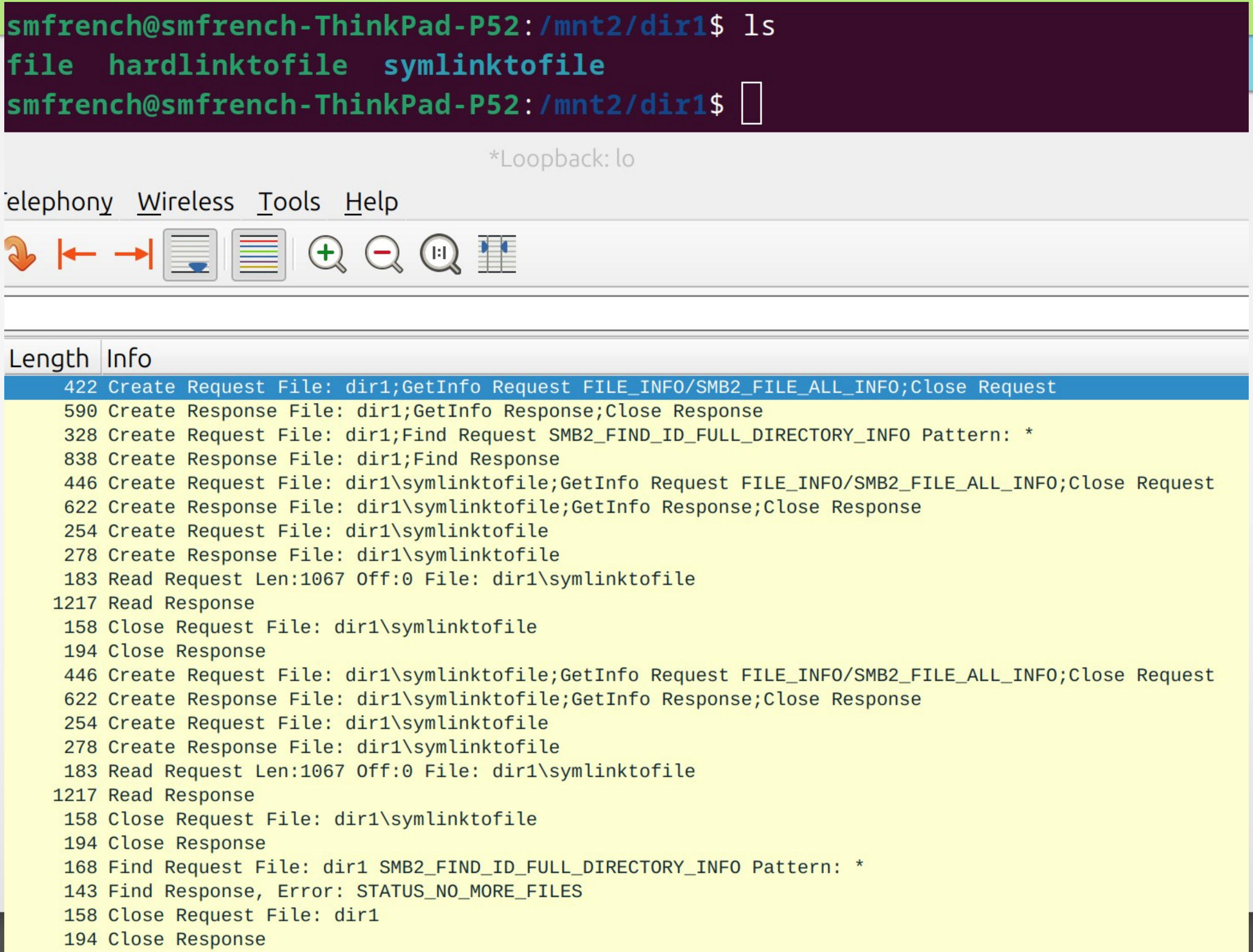


```
smfrench@smfrench-ThinkPad-P52:/mnt2/dir1$ ls -l
total 8
-rwxr-xr-x 1 root root 0 Sep 19 12:16 file
-rwxr-xr-x 1 root root 0 Sep 19 12:16 hardlinktofile
lrwxrwxrwx 1 root root 4 Sep 20 13:08 symlinktofile -> file
smfrench@smfrench-ThinkPad-P52:/mnt2/dir1$
```

```
Length Info
   422 Create Request File: dir1;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
   590 Create Response File: dir1;GetInfo Response;Close Response
   328 Create Request File: dir1;Find Request SMB2_FIND_ID_FULL_DIRECTORY_INFO Pattern: *
   838 Create Response File: dir1;Find Response
   446 Create Request File: dir1\symlinktofile;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
   622 Create Response File: dir1\symlinktofile;GetInfo Response;Close Response
   254 Create Request File: dir1\symlinktofile
   278 Create Response File: dir1\symlinktofile
   183 Read Request Len:1067 Off:0 File: dir1\symlinktofile
  1217 Read Response
   158 Close Request File: dir1\symlinktofile
   194 Close Response
   446 Create Request File: dir1\symlinktofile;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
   622 Create Response File: dir1\symlinktofile;GetInfo Response;Close Response
   254 Create Request File: dir1\symlinktofile
   278 Create Response File: dir1\symlinktofile
   183 Read Request Len:1067 Off:0 File: dir1\symlinktofile
  1217 Read Response
   158 Close Request File: dir1\symlinktofile
   194 Close Response
   168 Find Request File: dir1 SMB2_FIND_ID_FULL_DIRECTORY_INFO Pattern: *
   143 Find Response, Error: STATUS_NO_MORE_FILES
   158 Close Request File: dir1
   194 Close Response
```

Optimizing common cases: improve “ls” when symlinks

- Current behavior with “mfsymlinks”

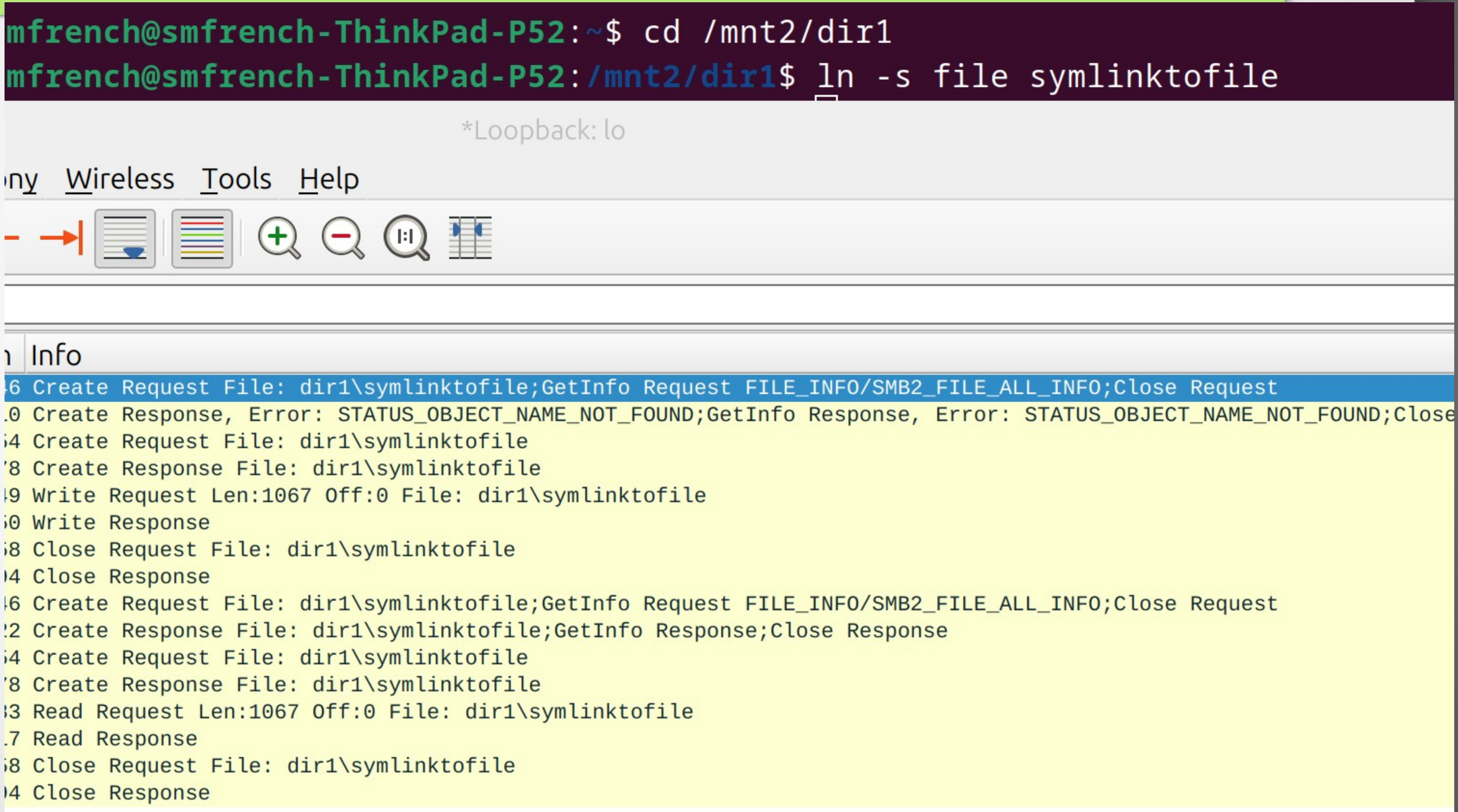


```
smfrench@smfrench-ThinkPad-P52:/mnt2/dir1$ ls
file  hardlinktofile  symlinktofile
smfrench@smfrench-ThinkPad-P52:/mnt2/dir1$ 
```

```
 422 Create Request File: dir1;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
 590 Create Response File: dir1;GetInfo Response;Close Response
 328 Create Request File: dir1;Find Request SMB2_FIND_ID_FULL_DIRECTORY_INFO Pattern: *
 838 Create Response File: dir1;Find Response
 446 Create Request File: dir1\symlinktofile;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
 622 Create Response File: dir1\symlinktofile;GetInfo Response;Close Response
 254 Create Request File: dir1\symlinktofile
 278 Create Response File: dir1\symlinktofile
 183 Read Request Len:1067 Off:0 File: dir1\symlinktofile
1217 Read Response
 158 Close Request File: dir1\symlinktofile
 194 Close Response
 446 Create Request File: dir1\symlinktofile;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
 622 Create Response File: dir1\symlinktofile;GetInfo Response;Close Response
 254 Create Request File: dir1\symlinktofile
 278 Create Response File: dir1\symlinktofile
 183 Read Request Len:1067 Off:0 File: dir1\symlinktofile
1217 Read Response
 158 Close Request File: dir1\symlinktofile
 194 Close Response
 168 Find Request File: dir1 SMB2_FIND_ID_FULL_DIRECTORY_INFO Pattern: *
 143 Find Response, Error: STATUS_NO_MORE_FILES
 158 Close Request File: dir1
 194 Close Response
```

Optimizing common cases: improve creating symlinks

- Current behavior with “mfsymlinks”



```
mfrench@smfrench-ThinkPad-P52:~$ cd /mnt2/dir1
mfrench@smfrench-ThinkPad-P52:/mnt2/dir1$ ln -s file symlinktofile
```

```
n  Info
6  Create Request File: dir1\symlinktofile;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
0  Create Response, Error: STATUS_OBJECT_NAME_NOT_FOUND;GetInfo Response, Error: STATUS_OBJECT_NAME_NOT_FOUND;Close
4  Create Request File: dir1\symlinktofile
8  Create Response File: dir1\symlinktofile
9  Write Request Len:1067 Off:0 File: dir1\symlinktofile
0  Write Response
8  Close Request File: dir1\symlinktofile
4  Close Response
6  Create Request File: dir1\symlinktofile;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Close Request
2  Create Response File: dir1\symlinktofile;GetInfo Response;Close Response
4  Create Request File: dir1\symlinktofile
8  Create Response File: dir1\symlinktofile
3  Read Request Len:1067 Off:0 File: dir1\symlinktofile
7  Read Response
8  Close Request File: dir1\symlinktofile
4  Close Response
```

Recent improvements in the user space tools

(cifs-utils)

Cifs-utils improvements

- Various important man page updates (e.g. new mount parms
- Various minor fixes
- New feature to use CLDAP to find closest site (thanks to David Voit)
  - Being reviewed now
- Additional debugging tools are being considered, as well as additional tools for change notify and making updating passwords easier

Testing Improvements

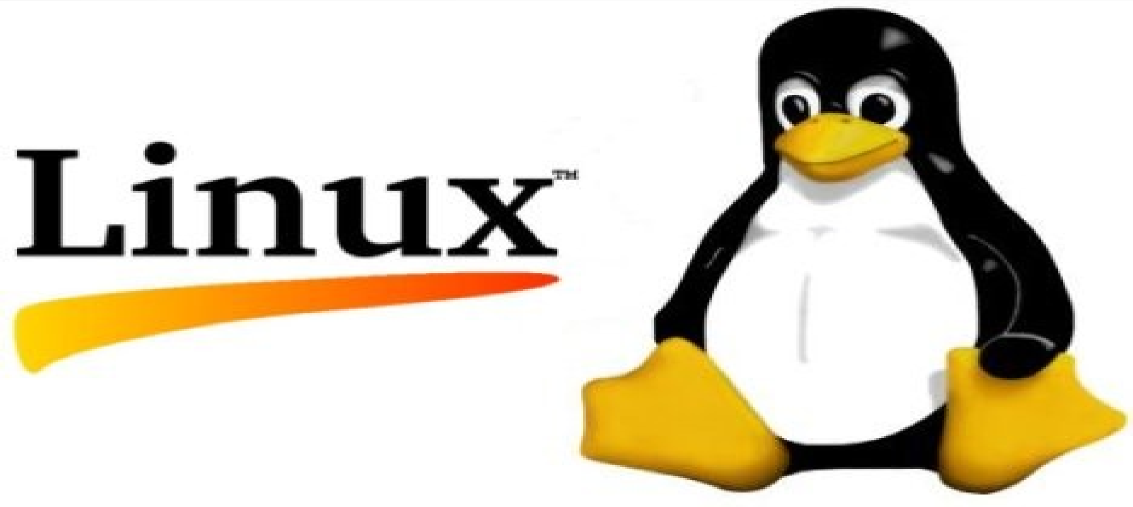
Test … test … test …

Additional tests are encouraged (generic or smb specific)

- Xfstests are the standard Linux filesystem functional tests
- Over last 9 months added 46 to the main “cifs-testing” regression testing group (288 tests run on every checkin from this group)
- Various server specific groups have added even more
- Azure SMB3.1.1 multichannel: added 20 more tests, now 153 tests
  - e.g. http://smb311-linux-testing.southcentralus.cloudapp.azure.com/#/builders/5/builds/91
- Ksmbd (Linux kernel server target) added 77 more tests, now includes 221 tests
  - e.g. http://smb311-linux-testing.southcentralus.cloudapp.azure.com/#/builders/10/builds/52
- Detailed wiki pages on wiki.samba.org go through how to setup xfstests with cifs.ko, and what features need to be added to enable more tests (tests that currently skip or fail so aren’t run in the ‘buildbot’)

Thank you for your time

- Future is very bright!



+ ***SMB3***

Additional Resources to Explore for SMB3 and Linux

https://msdn.microsoft.com/en-us/library/gg685446.aspx

- In particular MS-SMB2.pdf at https://msdn.microsoft.com/en-us/library/cc246482.aspx

https://wiki.samba.org/index.php/Xfstesting-cifs and test results

- http://smb311-linux-testing.southcentralus.cloudapp.azure.com/#/

Linux CIFS client https://wiki.samba.org/index.php/LinuxCIFS

Samba-technical mailing list

And various presentations at http://www.sambaxp.org and Microsoft Learn (learn.microsoft.com) and of course SNIA … http://www.snia.org/events/storage-developer

And the code:

- https://git.kernel.org/cgit/linux/kernel/git/torvalds/linux.git/tree/fs/smb
- For pending changes, soon to go into upstream kernel see:
  - https://git.samba.org/?p=sfrench/cifs-2.6.git;a=shortlog;h=refs/heads/for-next
- Kernel server code: https://git.samba.org/ksmbd.git/?p=ksmbd.git (ksmbd-for-next branch)