SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

The Future of Accessing Files remotely from Linux: SMB3.1.1 client status update

Steve French
Principal Software Engineer
Azure Storage - Microsoft

Legal Statement

- 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, Windows, various SMB3/CIFS based NAS appliances and the Cloud (Azure)
- Member of the Samba team, coauthor of SNIA CIFS Technical Reference,former SNIA CIFS Working Group chair
- Principal Software Engineer, Azure Storage: Microsoft

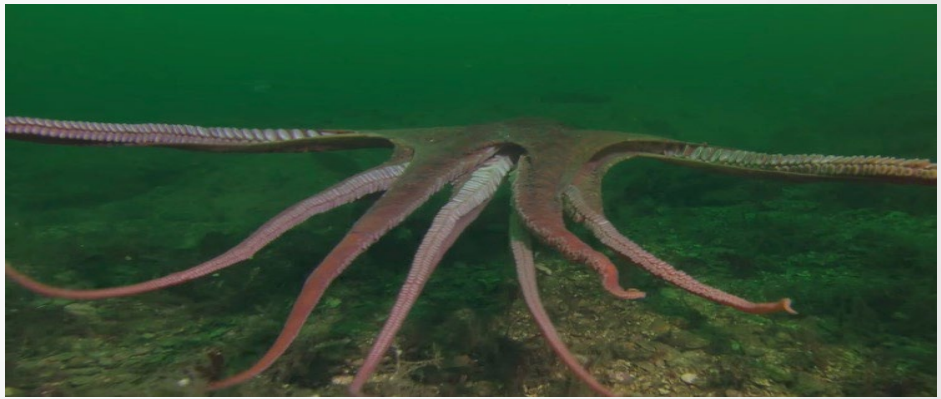
Outline

- Summary of Recent Linux VFS and FS Activity
- New Linux Kernel Server
- Recent Linux Client Improvements
- Expected Linux Client Features in near future
- Cifs-utils improvements
- Testing

A year ago … and now … kernel (including SMB3 client cifs.ko) improving

- A year ago Linux 5.3 “Bobtail Squid”

Now Linux 5.9-rc4: “Kleptomaniac Octopus”



Most Active Linux Filesystems this year

- 6345 kernel filesystem changesets last year (since Linux 5.3) (up)
  - FS activity: 6.9% of overall kernel changes, flat
  - Kernel is huge (> 20.5 million lines of code, measured 9/1/2020)
- There are many Linux file systems (>60), but six (and the VFS layer itself) drive ¾ of activity (btrfs, xfs, nfs and cifs are the most active)
  - File systems represent 4.7% of kernel source code (966KLOC) but among the most carefully watched areas
- cifs.ko (cifs/smb3 client) activity is strong

Linux File System Change Detail for past year (5.3 to now)

- VFS (overall fs mapping layer and common functions)

Linux != POSIX. Lots more syscalls and FS is responsible for > 200 of 850. +3 recently!

Syscall name	Kernel Version introduced
io_uring_ (various)	5.1
fsconfig, fsmount, fsopen, fspick, open_tree, move_mount	5.2
openat2	5.6
fsaccessat2	5.8
close_range	5.9

Discussions driving some of the FS development activity

- Rewrite of FSCACHE
- New mount API and other recent VFS changes
- Update to allow extended query fs information
- New notification mechanisms
- How to improve support for Containers
- Better support for faster storage (NVME, RDMA)
- io_uring and improved async i/o
- Shift to Cloud (longer latencies, object & file coexisting)

What about the server?

- Samba server is great (and huge, and full function)
- But now we also have a kernel server, ksmbd!

Linux SMB3 Kernel Server Update

- Lots of progress! Very exciting
- Kernel server
  - Module:

New Kernel Server “cifsd” arrives!

- ksmbd.ko and userspace helper utilities
- Thank you Namjae and team!
- See https://wiki.samba.org/index.php/Linux_Kernel_Server

```
root@smfrench-ThinkPad-P52:/home/smfrench/ksmbd-tools# mount -t cifs //localhost/test /mnt -o u
ame=testuser,password=testpass
root@smfrench-ThinkPad-P52:/home/smfrench/ksmbd-tools# ls /mnt
0740dir  1GB          fio-testfile.0.0  newfile0764  test-432
0750dir  310          fio-testfile.1.0  somefile     test-433
0754dir  314-dir      fio-testfile.2.0  syscalltest  timestamp-test.txt
0760     dbench       fio-testfile.3.0  test-430
0765dir  dir-no-posix fsx               test-431
root@smfrench-ThinkPad-P52:/home/smfrench/ksmbd-tools# ps -A | grep mbd
 3391 ?        00:00:00 usmbd
 3392 ?        00:00:00 usmbd
 3393 ?        00:00:00 ksmbd-tun0
 3394 ?        00:00:00 ksmbd-wlp0s20f3
 3395 ?        00:00:00 ksmbd-enp0s31f6
 3396 ?        00:00:00 ksmbd-lo
 3417 ?        00:00:00 ksmbd:48810
root@smfrench-ThinkPad-P52:/home/smfrench/ksmbd-tools# touch /mnt/newfile
root@smfrench-ThinkPad-P52:/home/smfrench/ksmbd-tools# mkdir /mnt/newdir
```

New in kernel server for SMB3 [continued]

- Great work by Namjae, Sergey and others
- See https://github.com/smfrench/smb3-kernel/tree/cifsd-for-next
- Still experimental
- Goal to send to linux-next soon if build verification run completes as expected
- Mirrored onto tree on github and samba.org (https://git.samba.org/?p=sfrench/cifsd.git)

New in kernel server for SMB3 [continued]

- Name of module: “ksmbd.ko”
- Name of source directory “cifsd” (to make it easier to find in the kernel fs directory, fs/cifsd will show up next to fs/cifs directory in the directory listing
- Name of daemons – begin with “ksmbd” to distinguish the “kernel” smb3 server from Samba (user space) whose processes are named “smbd”

Quality Much Improved(1)

- More improved xfstests pass 98 (+26)

```
generic/524 files ...  89s
generic/533 files ...  0s
Ran: cifs/001 generic/001 generic/002 generic/005 generic/006 generic/007
  generic/011 generic/013 generic/014 generic/020 generic/023 generic/024
  generic/028 generic/029 generic/030 generic/032 generic/033 generic/036
  generic/037 generic/069 generic/070 generic/074 generic/080 generic/084
  generic/086 generic/095 generic/098 generic/100 generic/109 generic/113
  generic/117 generic/124 generic/125 generic/129 generic/130 generic/132
  generic/133 generic/135 generic/141 generic/169 generic/198 generic/207
  generic/208 generic/210 generic/211 generic/212 generic/214 generic/215
  generic/221 generic/239 generic/245 generic/246 generic/247 generic/248
  generic/249 generic/257 generic/258 generic/286 generic/308 generic/309
  generic/310 generic/313 generic/315 generic/339 generic/340 generic/344
  generic/345 generic/346 generic/349 generic/350 generic/354 generic/360
  generic/377 generic/391 generic/393 generic/394 generic/406 generic/412
  generic/420 generic/422 generic/432 generic/433 generic/436 generic/437
  generic/438 generic/445 generic/446 generic/448 generic/451 generic/452
  generic/454 generic/460 generic/464 generic/465 generic/476 generic/504
  generic/524 generic/533
Passed all 98 tests
```

Quality Much Improved(2)

- Open source projects and commercial companies have begun to adopt ksmbd for their solution. (Mainly embedded targets)
  - DD-WRT (include in all firmware)
  - OpenWRT (include in Base version, optional in Normal version)
  - AXIS Network Camera(s3008)
- Many issues was fixed as ksmbd is distributed with their solutions
  - Compatibility issues with various smb clients(smart phone apps, smbclient)
  - Kernel oops or hang issues and leakages.
  - Potential issues found using static checker.
- Applied 463 patches Since SDC 2019.

Work in progress

- Add support for ACLs
  - Code implementation complete(storing ntacl to xattr).
  - Fixing the failure from smbtorture tests.
- Add support for Kerberos
  - Use the existing userspace kerb5 library
  - Require an auxiliary user-space daemon(ksmbd.gssd)
- OPEN_BACKUP_INTENT(TODO)
- SMB3 MULTI CHANNEL(TODO)

New git tree for upstream

- The upstream version of ksmbd with the following improvements is merged into smb3 kernel github tree(https://github.com/smfrench/smb3-kernel)
  - SMB1 code removal
  - Code cleanup(fixed the warnings from checkpatch.pl and sparse tool)
  - Fixed build error with the latest kernel source.
- It will the best way to integrate the testing (and upstreaming) of this into the linux kernel mainline.

What are the Linux SMB3.1.1 goals?

- Fastest, most secure general purpose way to access file data, whether in the cloud or on premises or virtualized
- Implement all reasonable Linux/POSIX features - so apps don’t know they run on SMB3 mounts (vs. local)
- As Linux evolves, and need for new features discovered, quickly add them to Linux kernel client and Samba

New Features

- Lots of Progress in the past year!

“modefromsid” mount option

- Useful for “nfs style” security where the client’s permission evaluation matters most
- Stored in ACE with ‘special SID’ unenforced by server
- Creating files with all 4096 mode combinations works

-rw--w-rwx 1 root root 14 May 13 00:25 407file

-rwsrwS--T 1 root root 0 May 13 00:26 4080file

-rwsrwS--t 1 root root 0 May 13 00:26 4081file

-rwsrwS-wT 1 root root 14 May 13 00:26 4082file

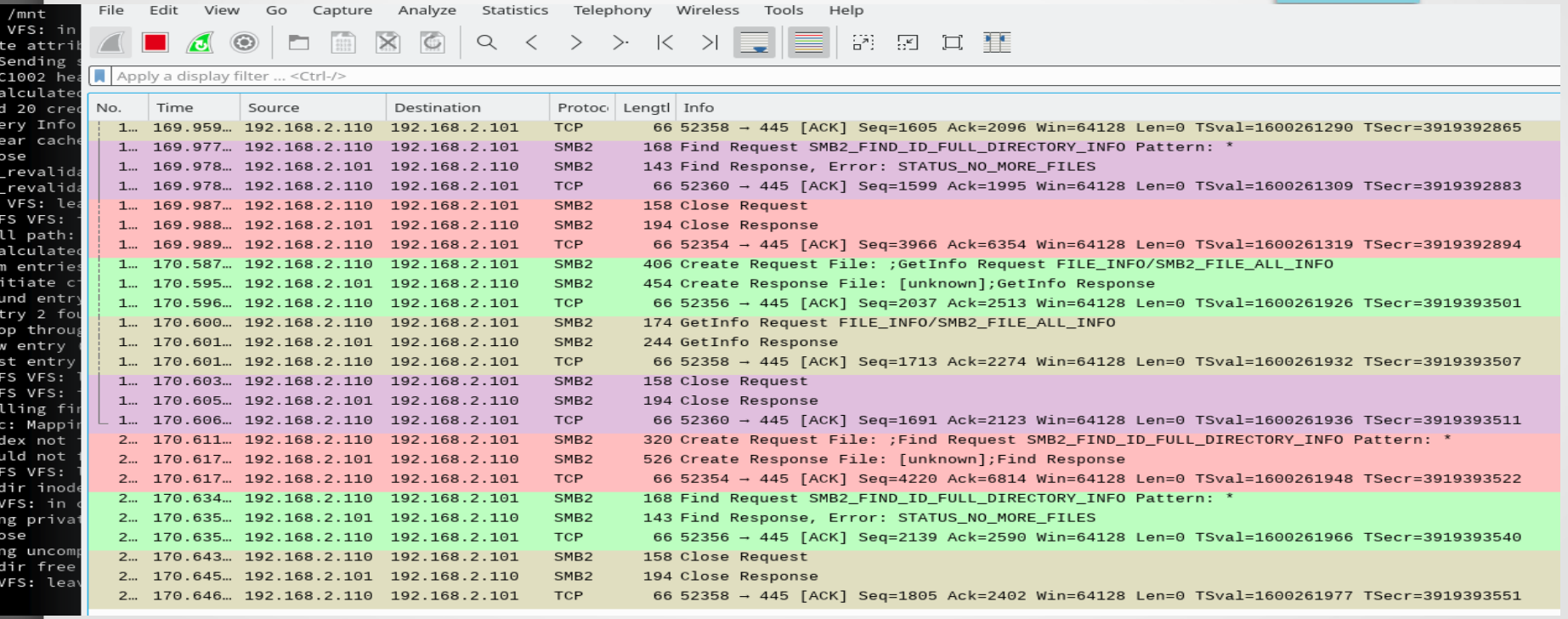
-rwsrwS-wt 1 root root 14 May 13 00:26 4083file

-rwsrwSr-T 1 root root 0 May 13

Multichannel added into Linux in 5.5 kernel

- Thank you Aurelien!
- Expected to be a big performance win …
- Big I/O performance improvement in 5.8 kernel (up to 5x faster in my testing)

Trace using multichannel w/current cifs.ko



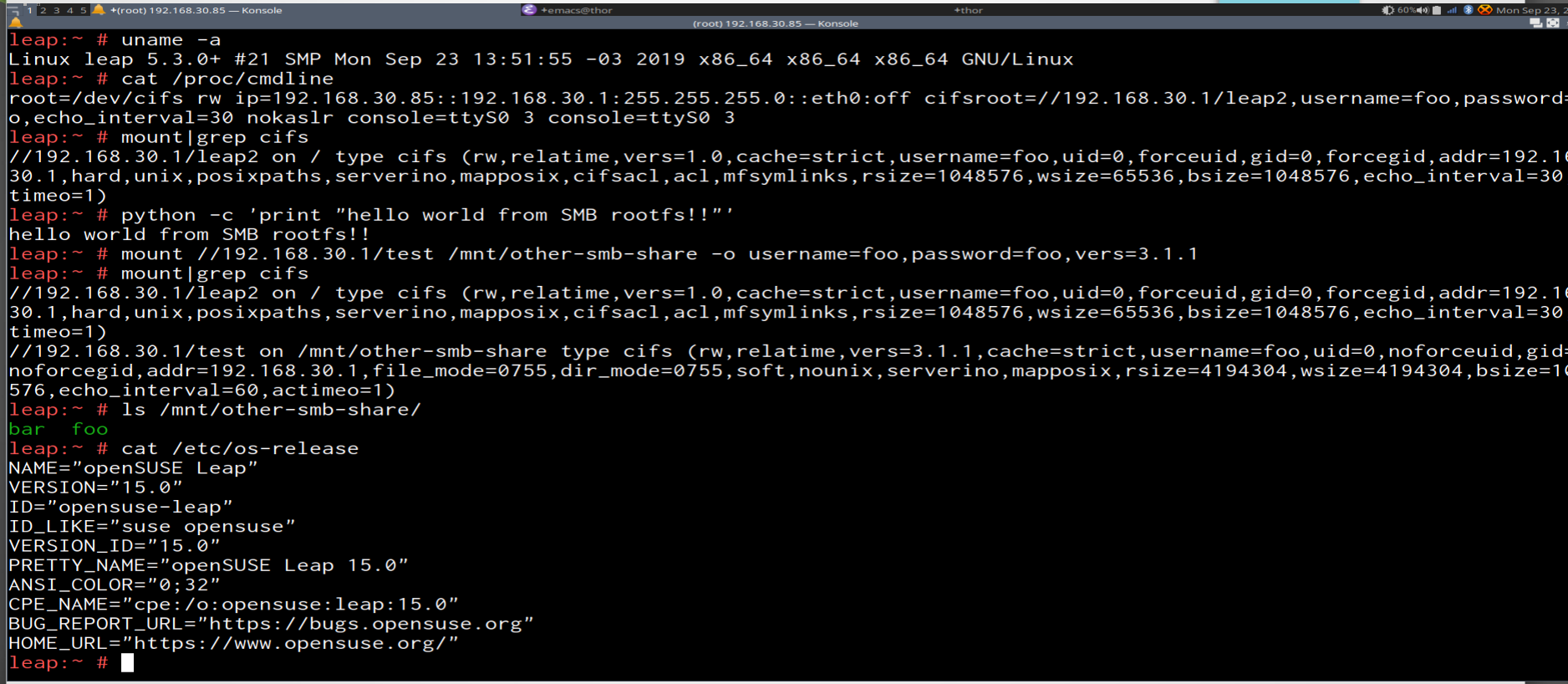
Now 82 smb3 dynamic tracepoints (adding more every year)

```
root@smfrench-ThinkPad-P52:~# ls /sys/kernel/debug/tracing/events/cifs
cifs_flush_err          smb3_lease_err                          smb3_read_enter
cifs_fsync_err          smb3_lock_err                           smb3_read_err
enable                  smb3_mkdir_done                         smb3_reconnect
filter                  smb3_mkdir_enter                        smb3_reconnect_with_invalid_credits
smb3_close_done         smb3_mkdir_err                          smb3_rename_done
smb3_close_enter        smb3_notify_done                        smb3_rename_enter
smb3_close_err          smb3_notify_enter                       smb3_rename_err
smb3_cmd_done           smb3_notify_err                         smb3_rmdir_done
smb3_cmd_enter          smb3_open_done                          smb3_rmdir_enter
smb3_cmd_err            smb3_open_enter                         smb3_rmdir_err
smb3_credit_timeout     smb3_open_err                           smb3_ses_expired
smb3_delete_done        smb3_partial_send_reconnect             smb3_set_eof_done
smb3_delete_enter       smb3_posix_mkdir_done                   smb3_set_eof_enter
smb3_delete_err         smb3_posix_mkdir_enter                  smb3_set_eof_err
smb3_enter              smb3_posix_mkdir_err                    smb3_set_info_compound_done
smb3_exit_done          smb3_posix_query_info_compound_done     smb3_set_info_compound_enter
smb3_exit_err           smb3_posix_query_info_compound_enter    smb3_set_info_compound_err
smb3_falloc_done        smb3_posix_query_info_compound_err      smb3_set_info_err
smb3_falloc_enter       smb3_query_dir_done                     smb3_slow_rsp
smb3_falloc_err         smb3_query_dir_enter                    smb3_tcon
smb3_flush_done         smb3_query_dir_err                      smb3_write_done
smb3_flush_enter        smb3_query_info_compound_done           smb3_write_enter
smb3_flush_err          smb3_query_info_compound_enter          smb3_write_err
smb3_fsctl_err          smb3_query_info_compound_err            smb3_zero_done
smb3_hardlink_done      smb3_query_info_done                    smb3_zero_enter
smb3_hardlink_enter     smb3_query_info_enter                   smb3_zero_err
smb3_hardlink_err       smb3_query_info_err
smb3_lease_done         smb3_read_done
```

GCM Fast

- Can more than double write perf! 80% for read
- Works with Windows, and with complementary recent changes to Samba server, mounts to Samba also benefit (a lot)
- In 5.3 kernel

Boot diskless systems via cifs.ko! Added in 5.5 kernel



Thank you Paulo!

- - Require ipconfig to set up network stack prior to mounting the SMB root filesystem:
  * E.g., "... ip=dhcp cifsroot=//localhost/share,..."

- - Current limitations:
  * no IPv6 support
  * default to insecure dialect SMB1 due to SMB1+UNIX extensions[1] (lack of SMB3+ POSIX extensions), although it can be changed through "cifsroot=" option. Fixes in progress for this to work with SMB3+
  * Init scripts that may fail due to unrecognized new cifsroot option

5.3 kernel, 55 changesets, Sept 15th, 2019. cifs internal module number 2.22

- Improve performance of open (cut network requests from 3 to 2), improves perf about 10%
- Improve encrypted read and write perf with the addition of GCM crypto (e.g. can more than double encrypted write performance and large reads MUCH faster as well)
- copy_file_range (fast server side copy) now supports cross share copy offload
- smbdirect (SMB3 over RDMA) no longer ‘experimental’ (thanks Long Li!)
- Send netname context on negotiate protocol (could help load balancers eg.)
- Can query symlinks stored as reparse points

5.4 kernel. 76 smb3 changesets. Nov. 24th, 2019. version 2.23

- Boot from cifs (root file system on cifs). Networking dependencies went in 5.5. Thank you Paulo from SuSE!
- mount parm “modefromsid” to allow setting mode bits in special ACE
- Allow decryption for large reads to be offloaded: new mount parm

  “esize=<min-offload-size>” to improve encrypted read performance via parallel decryption
- Allow disabling requesting leases for a mount (“nolease” mount parm)
- Add passthrough ioctl for SMB3 SetInfo. Thank you Ronnie from Redhat!
- Add new mount options for forced caching (“cache=ro” and “cache=singleclient”) and improved signing perf (“signloosely”)
- Display max requests in flight.

5.5. 61 changesets. January 26th, 2020 Cifs version 2.24

- Add support for flock
- SMB3 Multichannel support (Thank You Aurelien)
- Performance optimization query attributes on close (also is more correct for cases where timestamp update delayed to close time)
- Improvements to Boot from cifs (root file system on cifs) – network dependencies merged
- Readdir performance optimization (reparse points)

5.6 kernel March 2020 – 59 changesets, cifs.ko version 2.25

- “modefromsid” mount option much improved to set better ACL at file create time
- Add support for fallocate mode 0 for non-sparse files
- Allow setting owner info, DOS attributes and creation time from user space backup/restore tools (Thank you Boris Protopopov)
- Readdir performance optimization (add compouding support for readdir, cuts roundtrips for typical ls from about 9 to 7) (Thank you Ronnie)
- Readdir improvements for modefromsid and cifsacl (so mode bits don’t get overwritten by default mode in readdir)
- Add new ioctl for change notify (for user space tools to wait on directory change notifications)

5.7 kernel. 5/31/2020. 49 changesets, cifs.ko version 2.26

- Big perf improvement for signed connections (when multiple requests sent at same time)
- RDMA (smbdirect) improvements
- Swap over SMB3
- Support for POSIX readdir

5.8 kernel. 8/2/2020. 61 changesets cifs.ko version 2.28

- Big perf improvement for large I/O with multichannel (often > 4x faster)
- Support for “idsfromsid” (allowing alternate way of handling chown - mapping of POSIX uid/gid, owner information, into ‘special SID’)
- Support for POSIX queryinfo (All key parts of SMB3.1.1 POSIX extensions support complete)

What improvements to expect in the near future

- Even stronger encryption available: AES-GCM-256 for more demanding, most secure workloads
- Caching improvements
  - Extending directory leases beyond root directory
  - Use of handle leases to cache file data across close
- Continued optimization of network traffic, reducing roundtrips to continued improvements to use of ‘compounding’
- Multichannel reconnect improvements

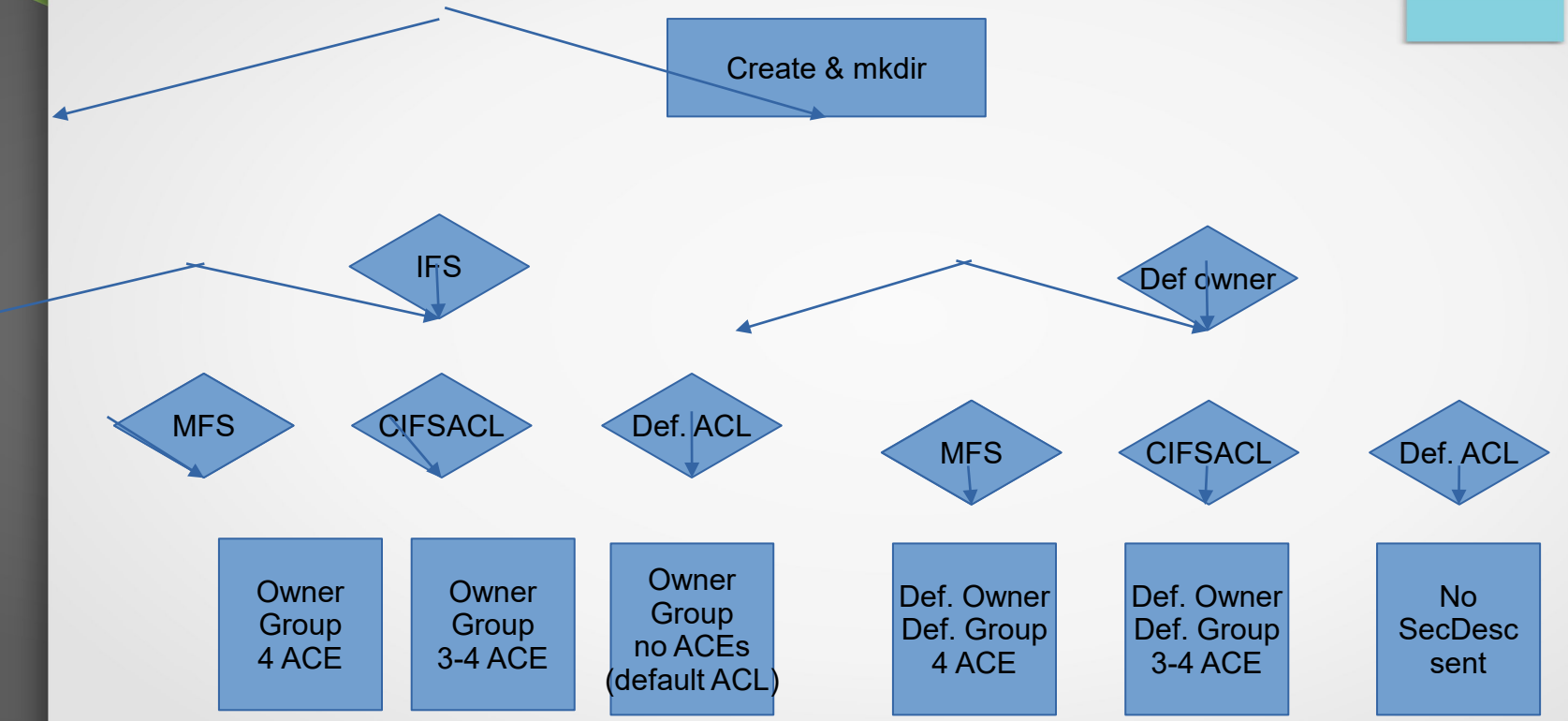
What about QUIC?

- It is not just about encryption and avoiding the “port 445 problem”
- QUIC has many performance features that can help as well
- Lack of kernel network driver for QUIC protocol is key issue, being discussed
  - Perhaps the opensourced cross platform ‘msquic’ github project could be used as a starting point
- Discussions continuing at SDC

What about Security Improvements?

- New client features being discussed
- Broaden the supported security scenarios
- Better SELinux integration with SMB3.1.1
- Improve the support for multiuser Kerberos mounts, winbind integration (e.g. for idmapping and ticket refresh – via cifs.upcall)
- Add support for ‘dummy mounts’ to ease cases where krb5 credentials aren’t available when mount is setup at boot
- Even stronger encryption (AES256)
- Solve the “port 445 problem”: add QUIC support (may be helpful for some non-encrypted cases in the future as well)
  - Need a QUIC kernel driver for Linux … would the open source project msquic be worth porting?

New security models: idsfromsid, modefromsid, cifsacl



Cifs-utils improvements

- Smbinfo rocks!
- Smbinfo rewritten in python
- Easy to extend
- New quota tools

cifs-utils

- With pass-through SMB3 fsctl and query-info (and set-info) now possible it is easy to write user space tools to get any interesting info from the server
- Would love more contributions!
- Recently added python to make it easier to contribute
- Look at smbinfo from cifs-utils for examples

Recent example of how these are used

- With pass-through ioctl can now get quota information
  - New userspace helper tool, smb2quota.py, to display quota information for Linux SMB client file system
  - Will be part of cifs-utils
  - Thank you Kenneth D'souza!
- Let’s add more!

Sample output from smb2quota

/smb2quota.py -t

Common Configuration Options – Suggested use cases

- Frequently recommended
  - mfsymlinks
  - noperm
  - dir_mode=, file_mode=, uid=, gid=
- Sometimes recommended
  - cifsacl,idsfromsid (5.8 or later) or modefromsid (5.6 or later)
  - actimeo=
  - sec=krb5
  - seal
  - sfu
  - hard
  - nostrictsync (and also cache=)

Testing … testing … testing

- The “buildbot” - automated regression testing! Thank you Paulo, Ronnie and Aurelien. See:

http://smb3-test-rhel-75.southcentralus.cloudapp.azure.com

- See xfstesting page in cifs wiki https://wiki.samba.org/index.php/Xfstesting-cifs
- Easy to setup, exclude file for slow tests or failing ones
- Huge improvement in XFSTEST – up to 180 groups of tests run over SMB3 (more than run over NFS)! And more being added every release (added > 50 this past year)

Thanks to the buildbot – Best Releases Ever for SMB3!

- Prevents regressions
- Continues to improve quality

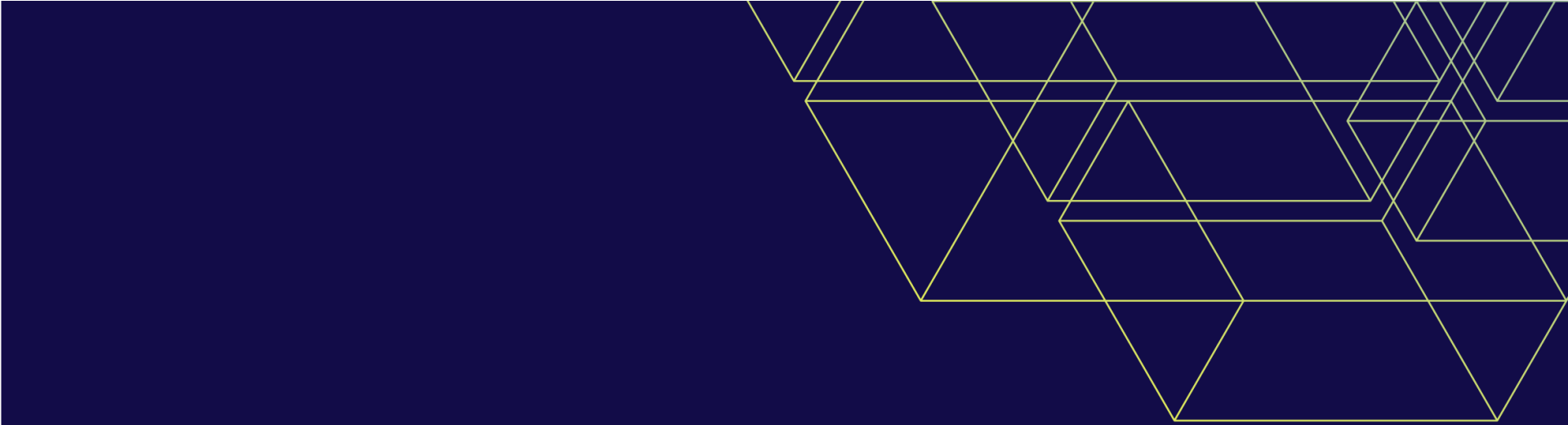
Thank you for your time

- Future is very bright!

+ **S M B 3**

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
- Linux CIFS client https://wiki.samba.org/index.php/LinuxCIFS
- Samba-technical mailing list and IRC channel
- And various presentations at http://www.sambaxp.org and Microsoft channel 9 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/cifs



Please take a moment to rate this session.

Your feedback matters to us.