# Linux Access to the Cloud and NAS reborn: Recent progress in SMB3.1.1

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 and various SMB3/CIFS based NAS appliances)
- Also wrote initial SMB2 kernel client prototype
- Member of the Samba team, coauthor of SNIA CIFS Technical Reference, former SNIA CIFS Working Group chair
- Principal Software Engineer, Azure Storage: Microsoft

### Outline

- General Linux File System Status Linux FS and VFS Activity
- What are the goals?
- What's New Key Feature Status
  - Kernel client
  - User space tools (cifs-utils)
  - Kernel server (new)
- Features expected soon
- Common Configuration Suggestions
- Testing Improvements

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

 A year ago Linux 5.0-rc4 "Shy Crocodile"

Last week:5.5 "Kleptomaniac Octopus"



# Discussions driving some of the FS development activity

- New mount API
- Improving support for Containers
- Unification of credential mgmt kernel keyring improved
- Better support for faster storage (NVME, RDMA)
- io\_uring and better async i/o
- "The Linux Copy Problem"
- Shift to Cloud (longer latencies, object & file coexisting)

### SMB on Linux reborn

- "Version 5.0 marks the reborn, new improved SMB3 Client For Linux"
- And it continues to improve rapidly, among the most active file systems on Linux for another year ...

# 2019 Linux FS/MM summit (in Puerto Rico in May)

Great group of talented developers



#### Samba Team meeting and testing in Redmond in the Fall



### Most Active Linux Filesystems this year

- 5338 kernel filesystem changesets last year (since 5.0-rc4 kernel) (flat)
  - FS activity: 6.2% of overall kernel changes (which are dominated by drivers) up slightly as % of activity
  - Kernel is huge (> 19 million lines of code, measured Saturday)
- There are many Linux file systems (>60), but six (and the VFS layer itself) drive 69% of activity (btrfs, xfs and cifs are the three most active)
  - File systems represent 4.8% of kernel source code (940KLOC) but among the most carefully watched areas
- cifs.ko (cifs/smb3 client) activity is strong
  - #3 most active of all fs with 375 changesets!
  - 54.9KLOC (not counting user space cifs-utils which are 11.6KLOC and samba tools which are larger)

# Linux File System Change Detail for past year (5.0-rc4 to now)

- •BTRFS 959 changesets (down slightly)
- •VFS (overall fs mapping layer and common functions) 958 (up significantly)
  •XFS 581 (up)
- •CIFS/SMB2/SMB3 client 375 (since 4.18 kernel activity has gone way up)
  •F2FS 271 (down)
- •NFS client 283 (down)
- •Others: EXT4 243(up), Ceph 164, AFS 135, GFS2 118, OCFS2 108 ...
- •NFS server 137 (flat). Linux NFS server **MUCH** smaller than CIFS or Samba
- •NB: Samba is as active as all Linux file systems put together (> 5000 changesets last year, over 1200 in the file server component alone!) broader in scope (by a lot) and also is user space not kernel. **98x larger than the NFS server in Linux! Samba now 3.4 million lines of code**

# Linux SMB3 Kernel Server Change Detail for past year

- Kernel server
  - Module: cifsd/ksmbd.ko
  - -Size: 24.5 KLOC
  - -2480 Changesets
- User space tools
  - -ksmbd-tools
  - -Size: 10.3 KLOC
  - -810 Changesets

FS Have a tough job! Responsible for > 200 of 850 syscalls. +12 since last year's SDC!

| Syscall name      | <b>Kernel Version introduced</b> |
|-------------------|----------------------------------|
| io_uring_enter    | 5.1                              |
| io_uring_register | 5.1                              |
| io_uring_setup    | 5.1                              |
| move_mount        | 5.2                              |
| open_tree         | 5.2                              |
| fsconfig          | 5.2                              |
| fsmount           | 5.2                              |
| fsopen            | 5.2                              |
| fspick            | 5.2                              |

#### What are the 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



# Examples of Great Progress we talked about last year!

- Reminders of some amazing progress ...
- A few examples

# Snapshots rock!

- Different mounts
- SS is Read-only
- Easy to use
- Example mount remote to cloud (Azure)



# rsize and wsize increase

### Previous default 1MB

 4MB gave 1 to 13% improved performance to Samba depending on network speed, 1% better for read.

#### Moved to 4MB in 4.20 kernel

And default block size moved to 1MB (helps cp)

# Examples of Great Progress more recently!

• This year ...

# 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 us
 3392 ?
             00:00:00 us
 3393 ?
             00:00:00 ksmbd
                          -tun0
             00:00:00 ksmbd-wlp0s20f3
 3394 ?
             00:00:00 ksmbd-enp0s31f6
 3395 ?
             00:00:00 ksmbd
 3396 ?
                          -lo
             00:00:00 ks
                          :48810
 3417 ?
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

- 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"

# Global Name Space – Much better! Thank you Paulo!

• Remember what DFS could do in Windows ... now the Linux client can handle failover and DFS entry caching too



# New security models: idsfromsid, modefromsid, cifsacl



# Multichannel (in 5.5 kernel)

- Thank you Aurelien!
- Expected to be a big performance win ...
- Working performance optimizations at plugfest this week



### Trace using multichannel w/current cifs.ko

ose

ose

Capture Analyze Statistics Telephony Wireless Tools File Edit View Go Help /mnt VFS: in q  $\langle \rangle \rangle \times K$  ||習 🛛 🗍 🎹 #1#1 0310 0310 X G ۲ te attr Sending Apply a display filter ... <Ctrl-/> C1002 hea alculate d 20 cred No. Time Source Destination Protoc Lengtl Info ery Info 1., 169.959., 192.168.2.110 192.168.2.101 TCP 66 52358 → 445 [ACK] Seg=1605 Ack=2096 Win=64128 Len=0 TSval=1600261290 TSecr=3919392865 ear cache 1... 169.977... 192.168.2.110 192.168.2.101 SMB2 168 Find Request SMB2\_FIND\_ID\_FULL\_DIRECTORY\_INFO Pattern: \* SMB2 143 Find Response, Error: STATUS\_NO\_MORE\_FILES 1... 169.978... 192.168.2.101 192.168.2.110 revalida 1... 169.978... 192.168.2.110 192.168.2.101 TCP 66 52360 → 445 [ACK] Seq=1599 Ack=1995 Win=64128 Len=0 TSval=1600261309 TSecr=3919392883 revalid VFS: le 1... 169.987... 192.168.2.110 192.168.2.101 SMB2 158 Close Request ES VES: 1... 169.988... 192.168.2.101 192.168.2.110 SMB2 194 Close Response ll path: 1... 169.989... 192.168.2.110 192.168.2.101 TCP 66 52354 → 445 [ACK] Seq=3966 Ack=6354 Win=64128 Len=0 TSval=1600261319 TSecr=3919392894 alculate 406 Create Request File: ;GetInfo Request FILE\_INFO/SMB2\_FILE\_ALL\_INFO 1... 170.587... 192.168.2.110 192.168.2.101 SMB2 n entrie itiate d 1... 170.595... 192.168.2.101 192.168.2.110 SMB2 454 Create Response File: [unknown];GetInfo Response und entr 1... 170.596... 192.168.2.110 192.168.2.101 TCP 66 52356 → 445 [ACK] Seq=2037 Ack=2513 Win=64128 Len=0 TSval=1600261926 TSecr=3919393501 try 2 fo 1... 170.600... 192.168.2.110 192.168.2.101 SMB2 174 GetInfo Request FILE\_INFO/SMB2\_FILE\_ALL\_INFO op throu 1... 170.601... 192.168.2.101 192.168.2.110 244 GetInfo Response SMB2 w entry 1... 170.601... 192.168.2.110 192.168.2.101 TCP 66 52358 → 445 [ACK] Seq=1713 Ack=2274 Win=64128 Len=0 TSval=1600261932 TSecr=3919393507 st entry FS VFS: 1... 170.603... 192.168.2.110 192.168.2.101 SMB2 158 Close Request FS VFS: 1... 170.605... 192.168.2.101 192.168.2.110 SMB2 194 Close Response lling fi 1... 170.606... 192.168.2.110 192.168.2.101 TCP 66 52360 → 445 [ACK] Seq=1691 Ack=2123 Win=64128 Len=0 TSval=1600261936 TSecr=3919393511 c: Mappi 2... 170.611... 192.168.2.110 192.168.2.101 SMB2 320 Create Request File: ;Find Request SMB2 FIND ID FULL DIRECTORY INFO Pattern: \* dex not uld not 2... 170.617... 192.168.2.101 192.168.2.110 SMB2 526 Create Response File: [unknown]; Find Response FS VFS: 2... 170.617... 192.168.2.110 192.168.2.101 TCP 66 52354 → 445 [ACK] Seq=4220 Ack=6814 Win=64128 Len=0 TSval=1600261948 TSecr=3919393522 dir inod 2... 170.634... 192.168.2.110 192.168.2.101 168 Find Request SMB2\_FIND\_ID\_FULL\_DIRECTORY\_INFO Pattern: \* SMB2 /FS: in SMB2 143 Find Response, Error: STATUS\_NO\_MORE\_FILES 2... 170.635... 192.168.2.101 192.168.2.110 ng privat 2... 170.635... 192.168.2.110 192.168.2.101 TCP 66 52356 → 445 [ACK] Seq=2139 Ack=2590 Win=64128 Len=0 TSval=1600261966 TSecr=3919393540 ng uncom 2... 170.643... 192.168.2.110 192.168.2.101 SMB2 158 Close Request dir free 2... 170.645... 192.168.2.101 192.168.2.110 SMB2 194 Close Response /FS: lea 2... 170.646... 192.168.2.110 192.168.2.101 TCP 66 52358 → 445 [ACK] Seq=1805 Ack=2402 Win=64128 Len=0 TSval=1600261977 TSecr=3919393551

#### Compounding helps a lot – thanks Ronnie!

Added in so far:

1) update timestamps on existing file: touch /mnt/file" goes from 6 request/resp pairs to 4

2) delete file "rm /mnt/file" from 5 to 2

3) make directory "mkdir /mnt/newdir" 6 to 3

4) remove directory "rmdir /mnt/newdir" 6 down to 2

5) rename goes from 9 request/response pairs to 5 ("mv /mnt/file /mnt/file1")

6) hardlink goes from 8 to only 3 (!) ("In /mnt/file1 /mnt/file2")

7) symlink with mfsymlinks enabled goes from 11 to 9 ("In -s /mnt/file1 /mnt/file3")

8) query file information "stat /mnt/file" goes from six roundtrips down to 2

9) And get/set xattr, and statfs and more

10) In 5.6 kernel Ronnie added support for compounding readdir (9 requests down to 7 for typical directory, more gain for larger directories)

# Sparse File Support (and other network fs can't do this)!

screen File Edit View Search Terminal Help [sahlberg@rawhide-2 cifs]\$ #Create a sparse file [sahlberg@rawhide-2 cifs]\$ sudo ./sparse-file.py /mnt/sparse Blocksize is 16384. Changing this to 64k as that is real block size on windows16 Needs fixing. 0...65536 131072...262144 327680...1048576 [sahlberg@rawhide-2 cifs]\$ #Check the FIEMAP [sahlberg@rawhide-2 cifs]\$ filefrag -v /mnt/sparse Filesystem type is: fe534d42 File size of /mnt/sparse is 1048576 (64 blocks of 16384 bytes) logical offset: physical offset: length: expected: flags: ext: 0: 4: 0.. 3: 0.. 3: 1: 15: 8.. 15: 8: 8.. 20.. 20.. 63: last,eof 2: 63: 44: /mnt/sparse: 1 extent found [sahlberg@rawhide-2 cifs]\$

⊗

[2 sahlberg@rawhide-2:/data/linux/fs/cifs] 0 sahlberg@rawhide-2:/ 1 sahlberg@

| - I   | <pre>Ioctl Response (0x0b)    StructureSize: 0x0031    unknown: 0000    Function: FSCTL_QUERY_ALLOCATED_RANGES (0x000940cf)    GUID handle File: sparse    In Data</pre>   | • |
|---|--|---|
|   | Offset: 0x00000070<br>Length: 16<br>▼ Range<br>File Offset: 0<br>Length: 9223372036854775807   |   |
|   | <ul> <li>Out Data         <ul> <li>Offset: 0x00000080</li> <li>Length: 48</li> <li>▼ Range                  <ul></ul></li></ul></li></ul>  |   |
|   | ▼ Range  |   |
|   | File Offset: 131072<br>Length: 131072<br>• Range<br>File Offset: 327680<br>Length: 720896  | Ŧ |
| 0000<br>0010<br>0020<br>0030<br>0040<br>0050<br>0050<br>0050<br>0050<br>0050<br>005 | 52       54       00       c1       f8       ef       52       54       00       f8       64       00       80       66       7d       7b       c0       a8       7c       c6       c0       a8       f8       f8 <td< td=""><td></td></td<> |   |

No.: 1094 · Time: 9.906657148 · Source: 192.168.124.198 · Destination: 192.168.124.203 · Protocol: SMB2 · Length: 246 · Info: loctl Response FSCTL\_QUERY\_ALLOCATED\_RANGES File: sparse

🗙 Close



#### Example test code using sparse files:

3

8

File Edit View Search Terminal Help

```
def fiogetbsz(f):
        return struct.unpack('I', fcntl.ioctl(f, FIGETBSZ, struct.pack('I', 0)))
[0]
def main():
        if len(sys.argv) != 2:
                usage()
        # test mapping: filefrag -v and hdparm --fibmap
        # ioctl(3, FIGETBSZ, 0x55c339b16070)
        # ioctl(3, FS IOC FIEMAP, {fm start=0, fm length=1844674407370955\
                 1615, fm flags=0, fm extent count=292}
        f = os.open(sys.argv[1], os.0 CREAT|os.0 RDWR)
        bs = fiogetbsz(f)
        print 'Blocksize is %d.' % bs
        if bs < 65536:
                print 'Changing this to 64k as that is real block size on window
s16 Needs fixing.
                bs = 65536
        pwrite(f, 'a' * 1024 * 1024, 0)
        smb2 set sparse(f, 1)
        smb2_set_zero data(f, bs, bs * 2)
        smb2 set zero data(f, bs * 4, bs * 5)
        buf = smb2 query allocated ranges(f, 0, 1024 * 1024)
        while len(buf):
                r = struct.unpack from('<2Q', buf, 0)</pre>
                print "%d...%d" % (r[0], r[0]+r[1])
                buf = buf[16:]
        os.close(f)
   name == " main ":
if
        main()
[sahlberg@rawhide-2 cifs]$
```

# Now 77 smb3 dynamic tracepoints (a year ago was 20 ...)!

#### root@smf-Thinkpad-P51:~# ls /sys/kernel/debug/tracing/events/cifs

#### enable filter

smb3 close err smb3 cmd done smb3 cmd enter smb3 cmd err smb3 credit timeout smb3 delete done smb3 delete enter smb3 delete err smb3 enter smb3 exit done smb3 exit err smb3 falloc done smb3 falloc enter smb3 falloc err smb3 flush err smb3 fsctl err

smb3 hardlink done smb3 hardlink enter smb3 hardlink err smb3 lease done smb3 lease err smb3 lock err smb3 mkdir done smb3 mkdir enter smb3 open done smb3 open enter smb3 open err smb3 partial send reconnect smb3 posix mkdir done smb3 posix mkdir enter smb3 posix mkdir err smb3 query dir done smb3 query dir enter

smb3\_query\_dir\_err smb3\_query\_info\_compound\_done smb3\_query\_info\_compound\_enter smb3\_query\_info\_compound\_err smb3\_query\_info\_enter smb3\_query\_info\_enter smb3\_query\_info\_err smb3\_read\_done smb3\_read\_enter smb3\_read\_err smb3\_reconnect smb3\_reconnect smb3\_rename\_done smb3\_rename\_enter smb3\_rename\_enter smb3\_rename\_err smb3\_rename\_err smb3\_rmdir\_done smb3\_rmdir\_enter

smb3 ses expired smb3 set eof done smb3 set eof enter smb3 set eof err smb3 set info compound done smb3 set info compound enter smb3 set info compound err smb3 set info err smb3 slow rsp smb3 tcon smb3 write done smb3 write enter smb3 write err smb3 zero done smb3 zero enter smb3 zero err

### **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



### An example

- On this laptop ... mounting to current Samba
  - Large writes 3x faster!
  - Large reads 2.5x faster
- Newly added SMB3.1.1 GCM support gives HUGE improvement in performance for large I/O

# Boot diskless systems via cifs.ko!

1 2 3 4 5 🔔 +(root) 192.168.30.85 — Konsole emacs@thor 🜓 60% 🕪 📋 📶 🛞 🐼 Mon Sep 23, 2 (root) 192.168.30.85 — Konsole . 🔿 leap:~ # uname -a Linux leap 5.3.0+ #21 SMP Mon Sep 23 13:51:55 -03 2019 x86\_64 x86\_64 x86\_64 GNU/Linux leap:~ # cat /proc/cmdline root=/dev/cifs rw ip=192.168.30.85::192.168.30.1:255.255.255.0::eth0:off cifsroot=//192.168.30.1/leap2,username=foo,password o,echo\_interval=30 nokaslr console=ttyS0 3 console=ttyS0 3 leap:~ # mount|grep cifs //192.168.30.1/leap2 on / type cifs (rw,relatime,vers=1.0,cache=strict,username=foo,uid=0,forceuid,gid=0,forcegid,addr=192.10 30.1, hard, unix, posixpaths, serverino, mapposix, cifsacl, acl, mfsymlinks, rsize=1048576, wsize=65536, bsize=1048576, echo\_interval=30 timeo=1) leap:~ # python -c 'print "hello world from SMB rootfs!!"' hello world from SMB rootfs!! leap:~ # mount //192.168.30.1/test /mnt/other-smb-share -o username=foo,password=foo,vers=3.1.1 leap:~ # mount|grep cifs //192.168.30.1/leap2 on / type cifs (rw,relatime,vers=1.0,cache=strict,username=foo,uid=0,forceuid,gid=0,forcegid,addr=192.10 30.1, hard, unix, posixpaths, serverino, mapposix, cifsacl, acl, mfsymlinks, rsize=1048576, wsize=65536, bsize=1048576, echo\_interval=30 timeo=1) //192.168.30.1/test on /mnt/other-smb-share type cifs (rw,relatime,vers=3.1.1,cache=strict,username=foo,uid=0,noforceuid,gid noforcegid,addr=192.168.30.1,file\_mode=0755,dir\_mode=0755,soft,nounix,serverino,mapposix,rsize=4194304,wsize=4194304,bsize=10 576,echo\_interval=60,actimeo=1) leap:~ # ls /mnt/other-smb-share/ bar foo leap:~ # cat /etc/os-release NAME="openSUSE Leap" VERSION="15.0" ID="opensuse-leap" ID\_LIKE="suse opensuse" VERSION ID="15.0" PRETTY\_NAME="openSUSE Leap 15.0" ANSI\_COLOR="0;32" CPE\_NAME="cpe:/o:opensuse:leap:15.0" BUG\_REPORT\_URL="https://bugs.opensuse.org" HOME\_URL="https://www.opensuse.org/" leap:~ #

### 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

### NetName context added

- In 5.3 kernel
- May help load balancers in the future (and debug tools too)

# Can now view detailed info on open files (not just "lsof" output)

#### Sample output from "cat /proc/fs/cifs/open\_files"

# Version:1
# Format:
# Format:
# <tree id> <persistent fid> <flags> <count> <pid> <uid> <filename> <mid
0x5 0x800000378 0x8000 1 7704 0 some-file 0x14
0xcb903c0c 0x84412e67 0x8000 1 7754 1001 rofile 0x1a6d
0xcb903c0c 0x9526b767 0x8000 1 7720 1000 file 0x1a5b
0xcb903c0c 0x9ce41a21 0x8000 1 7715 0 smallfile 0xd67</pre>

### RDMA – Performance Improved

- Thank you Long Li! Many fixes/improvements
- No longer CONFIG\_EXPERIMENTAL for smbdirect (RDMA) on Linux kernel client as of 5.3 kernel



# 4.20 (70 Changesets, December 23<sup>rd</sup>) cifs.ko internal module version 2.14

- RDMA and direct i/o performance improvements (add direct i/o to smb3 file ops)
- Much better compounding (create/delete/set/unlink/mkdir/rmdir etc.), huge perf improvements for metadata access
- Additional dynamic (ftrace) tracepoints
- Add /proc/fs/cifs/open\_files to allow easier debugging
- Slow response threshold is now configurable
- Requested rsize/wsize larger (4MB vs. 1MB)
- Query Info IOCTL passthrough (enables new "smb-info" tool to display useful metadata in much detail and also ACLs etc.), and allow ioctl on directories
- Many Bug Fixes (including for krb5 mounts to Azure, and fix for OFD locks, backup intent mounts)

# 5.0 (82 changesets) March 3<sup>rd</sup>, 2019 (cifs.ko internal module version 2.17)

- SMB3.1.1 requested by default (ie is now in default dialects list)
- DFS failover support added (can reconnect to alternate DFS target) for higher availability and

DFS referral caching now possible, cache updated regularly (Thank you Paulo)

- Support for reconnect if server IP address changes (coreq change in user space implemented in latest version of cifs-utils) (Thank you Paulo!)
- Performance improvement for get/set xattr (compounding support extended)
- Many Bug Fixes (24 important enough for stable) including for large file copy in cases where network connection is slow or interrupted, reconnect fixes, and fix for OFD lock support. The buildbot is really helping improve cifs.ko code quality!

# 5.1 (86 changesets) May 5<sup>th</sup> 2019 (cifs.ko internal module version 2.19)

- "fsctl passthrough" support improved: allows tools like cifs-utils to easily query any info available over SMB3 fsctl or query\_info
- New mount parm "handletimeout" to allow persistent/resilient handle behavior to be configurable
- Allow fallocate zero range to expand a file
- Improve perf: cache FILE\_ALL\_INFO for the shared root handle
- Improve perf: default inode block size reported as 1MB (NB: 4MB rsize/wsize)
- Cleanup mknod and special file handling (thank you Aurelien)
- Support guest mounts over smb3.1.1
- Add many dynamic trace points to ease debugging and perf analysis
- Bug fixes (23 important enough for stable). Adding even more tests to the buildbot really helped. Multiple fixes for 'crediting' (SMB3 flow control) thank you Pavel!

# 5.2 (64 changesets, so far) July 7 2019. cifs.ko version 2.20

- Bug fixes (11 important enough for stable)
- Improved perf: sparse file support now allows fiemap, SEEK\_HOLE and SEEK\_DATA (helps cp to Samba e.g.)
- Add support for fallocate ZERO\_RANGE
- Support "fsctl passthrough" for cases where send (write) data in SMB3 fsctl allows user space tools to do more!

# 5.3 (55 changesets) Sept 15<sup>th</sup>, 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 (76 changesets). Nov. 24<sup>th</sup>, 2019 Cifs 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.
- Can get keys for Wireshark encryption more easily via smbinfo <filename>

# 5.5 (61 changesets). January 26<sup>th</sup>, 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 – what is now merged (for April release) 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 Is from about 9 to 7) (Thank you Ronnie)

# 5.6 kernel – what is in progress (testing here at SMB3 Plugfest)

- Multichannel perf improvements
- Readdir improvements for modefromsid and cifsacl (so mode bits don't get overwritten by default mode in readdir)
- Change notify support
- Swap over SMB3

# Cifs-utils improvements

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

# Cifs-utils now even has a GUI!

#### secddesc-ui.py



# 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 /test Amount Used | Quota Limit | Warning Level | SID 70.0 kiB | 16.0 EiB | 16.0 EiB | S-1-5-32-544 27.0 kiB | 500.0 MiB | 450.0 MiB | S-1-5-21-3363399803-746912020-2622272238-1001 4.0 MiB | 16.0 EiB | 16.0 EiB | S-1-5-18 # ./smb2quota.py -c /test S-1-5-32-544,71680,18446744073709551615,18446744073709551615 S-1-5-21-3363399803-746912020-2622272238-1001,27648,524288000,471859200 S-1-5-18,4220928,18446744073709551615,18446744073709551615 # ./smb2guota.py -l /test SID:S-1-5-32-544 **Quota Used:71680** Quota Threshold Limit:NO WARNING\_THRESHOLD **Quota Limit:NO LIMIT** SID:S-1-5-21-3363399803-746912020-2622272238-1001 Quota Used:27648 Quota Threshold Limit:471859200 Quota Limit:524288000 SID:S-1-5-18 **Ouota Used:4220928** Quota Threshold Limit:NO WARNING THRESHOLD Quota Limit:NO LIMIT

# Common Configuration Options – Suggested use cases

- Frequently recommended
  - mfsymlinks
  - noperm
  - dir\_mode=, file\_mode=, uid=, gid=
- Sometimes recommended
  - cifsacl,idsfromsid or (now 5.6 and later) modefromsid
  - actime=
  - 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 127 groups of tests run over SMB3 (more than run over NFS)! And more being added every release

# Thanks to the buildbot – Best Releases Ever for SMB3!

- Prevents regressions
- Continues to improve quality





# Buildbot now has even more targets

| uildbot: Ho            | ome            | × +   |                           |   |   |   |
|------------------------|----------------|---|---------------------------|---|---|---|
| e c                    | Not secu       | re   smb3-test-rhel-75.southcentralus.cloudapp.azure.com/#/   |                           |   |   | ର୍ 🗯  |
| TESTING                | Ŧ              | CIFS TESTING Home   |                           |   |   |   |
| ion<br>iew<br>all View | fi<br>So       | Welcome to buildbot<br>0 builds running currently<br>20 recent builds                                   |                           |   |   |   |
| le View                | e<br>os        | cifs-testing  |                           | azure   |   | windows   |
|                        | <b>6</b><br>#1 | cifs-testing/313<br>build successful  | SUCCESS<br>3:36:13        | azure/234<br>build successful   | <b>SUCCESS</b><br>2:36:33                     | windows/56<br>build successful  |
|                        |                | cifs-testing/312<br>build successful  | <b>SUCCESS</b><br>4:37:56 | azure/233<br>build successful   | <b>SUCCESS</b><br>2:48:01                     | iraposix  |
|                        |                | cifs-testing/311<br>cancelled 'ssh fedora29.vm.test' (tailure) 'ssh fedora29.vm.test' (cancelled)       | CANCELLED<br>1:49:38      | azure/232<br>failed 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test' | FAILURE<br>' (failure) 'ssh fedora29.v3:23:31 | jraposix/15<br>failed 'ssh fedora29.ym.test' (failure) 'ssh fedora29.ym.test' (failure) 'ssh fedora29.    |
|                        |                | cifs-testing/310<br>build successful  | <b>SUCCESS</b><br>4:37:53 |   |   | jraposix/14<br>failed 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test' (failure) 'ssh fedora29.v   |
|                        |                | cifs-testing/309<br>failed 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test' (failure)            | <b>FAILURE</b><br>4:50:05 |   |   | jraposix/13<br>failed './start-samba.sh' (failure) 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test |
|                        |                | cifs-testing/308<br>failed 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test' (failure)            | 4:40:19                   |   |   |   |
|                        |                | cifs-testing/307<br>failed 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test' (failure) 'ssh fedor | FAILURE<br>a29.v4:42:25   | ksmbd   |   |   |
|                        |                | cifs-testing/306<br>build successful  | <b>SUCCESS</b><br>4:33:52 | ksmbd/5<br>build successful   | <b>SUCCESS</b><br>1:28:53                     |   |
|                        |                |   |                           | ksmbd/4<br>failed 'esh fadora 29 ym test _ ' (failura)                      | FAILURE                                       |   |

| ksmbd/5  | SUCCESS                           |
|--|-----------------------------------|
| build successful   | 1:28:53                           |
| ksmbd/4  | FAILURE                           |
| failed 'ssh fedora29.vm.test' (failure)                        | 1:49:36                           |
| ksmbd/3  | FAILURE                           |
| failed 'ssh fedora29.vm.test' (failure)                        | 2:06:47                           |
| ksmbd/2  | CANCELLED                         |
| cancelled 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.te | est' (failure) 'ssh fedora2:16:08 |
| ksmbd/1  | SUCCESS                           |
| build successful   | 30:00                             |

| a | posix   |
|---|---|
|   | raposix/15  |
| 1 | ailed 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm  |
| i | raposix/14  |
| 1 | ailed 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm  |
|   | raposix/13  |
| 1 | ailed './start-samba.sh' (failure) 'ssh fedora29.vm.test' (failure) 'ssh fedora29.vm.test |

2:45:36

### Thank you for your time

• Future is very bright!



# 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 <a href="http://www.sambaxp.org">http://www.sambaxp.org</a> and Microsoft channel 9 and of course SNIA ... <a href="http://www.snia.org/events/storage-developer">http://www.snia.org/events/storage-developer</a>
- And the code:
  - https://git.kernel.org/cgit/linux/kernel/git/torvalds/linux.git/tree/fs/cifs
  - 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