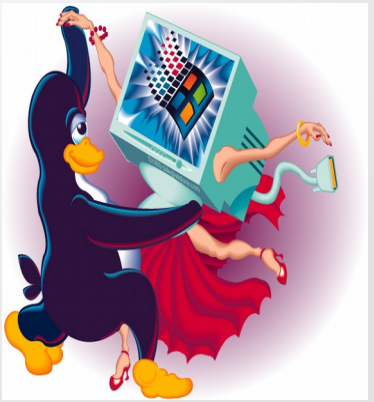
SMB3.1.1 and Beyond in the Linux Kernel: Providing Optimal File Access to Windows, Mac, Samba and Other File Servers

Steve French
Principal Systems Engineer – Primary Data



Legal Statement

- This work represents the views of the author(s) and does not necessarily reflect the views of Primary Data 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 and former SNIA CIFS Working Group chair
- Principal Systems Engineer, Protocols: Primary Data

Outline

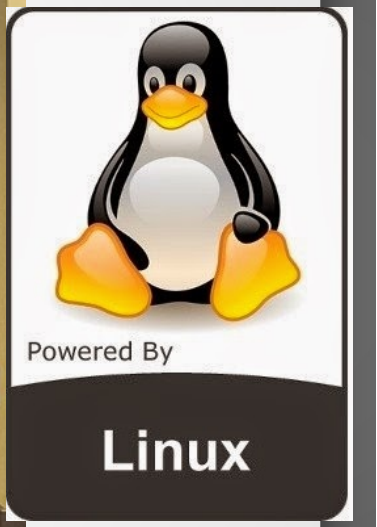
- File System Activity
- Key Feature Status
  - CopyOffload
  - Persistent/Resilient handles and HA
  - Fallocate
  - ACLs
  - Security Features/Encrypiton
  - Other optional SMB3 features
- Performance overview
- Three alternatives to POSIX compatibility
  - SMB3 with best effort POSIX emulation
  - CIFS POSIX Extensions (to Samba, cifs dialect only)
  - Mac AAPL context
  - SMB3 POSIX extensions
- Testing

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

- 12 months ago we had Linux version 4.2 ie “Hurr Durr I'm a Sheep”

Now we have 4.8-rc6 “Psychotic Stoned Sheep”

Working with great developers. See us here at 2016 Linux File System Summit in Raleigh



Some key features helping drive discussions and FS development activity ?

- Many of the high priority, evolving storage features are critical for NAS
  - Better support for NVMe
    - RDMA and low latency mechanisms to access VERY high speed storage
  - Faster network adapters (10Gb → 40Gb->100Gb ethernet … and cheaper RDMA)
  - RichACL
  - Xstat (extended stat)
  - Improved copy offload
  - Improved sparse file support (including for virtualization)
  - Shift of some workloads to object like access patterns

Most Active Linux Filesystems this year

- 4442 kernel filesystem changesets in last year (since 4.2 kernel)!
  - Linux kernel file system activity continuing strong (although down about 10% due to gradual maturing)
  - 5.4% of overall kernel changes (which are dominated by drivers) but fs is watched carefully
  - Kernel is now almost 15 million lines of source code (measured last week with sloccount tool)
- Six file systems (and the VFS) drive the majority of activity
  - File systems represent about 6% of the overall kernel source code (850,000 lines of code)
- cifs.ko (cifs/smb3 client) still among more active fs
  - Btrfs 797 changesets (increased)
  - VFS (overall fs mapping layer and common functions) 640
  - Xfs 453
  - Nfs client 459 (increased)
  - Ext4 243 (decreased)
  - CIFS/SMB2/SMB3 client 130
    - cifs.ko is 42,000 lines of kernel code (not counting user space helpers, and samba userspace tools)
  - Nfs server 144 (decreased)
- NB: Samba (cifs/smb2/smb3 server) is as active as the top 3 or 4 put together (1800 changesets) since it is broader in scope (by a lot) and also is in user space not in kernel

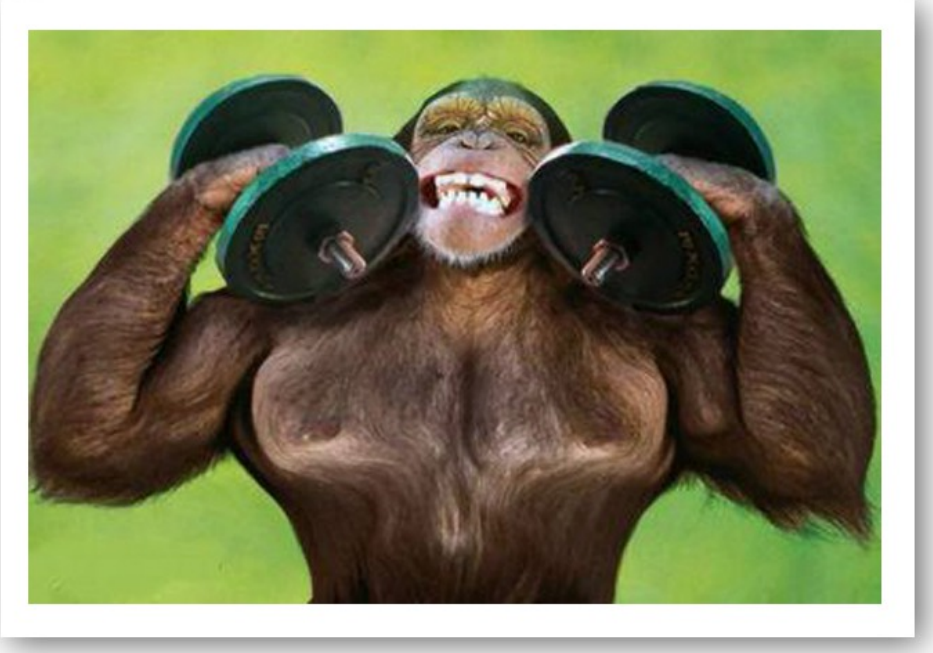
Fixes and Features by release

- Linux 4.2 (14 changesets)
  - SMB 3.11 (Windows 10) dialect support (improved security)
  - Faster copy offload (REFLINK, duplicate_extents) added for Windows Server 2016
- 4.3 (17 changesets)
  - Minor bug fixes (including Mac authentication issue when timestamps differ too much on server/client)
  - Add krb5 support for smb3
  - cifs.ko version updated to 2.08
  - Added ioctl to query detailed fs info on mounted share
- Linux 4.4 (17 changesets)
  - Allow copy offload across shares
  - Add resilient and persistent handle mount options and support for the (durable v2) create context

Fixes and Features (continued)

- Linux 4.5 (27 changesets)
  - Minor bug fixes
  - clone_file_range added to vfs, cifs support for clone_file_range
  - Allow O_DIRECT with cache=loose
  - Make echo interval tunable
  - (first phase of encryption support begun)
- Linux 4.6 (8 changesets)
  - Minor fixes
- Linux 4.7 (7 changes)
  - Fix badlock regression for guest mounts (mount with -o guest can fail to Samba servers when patched for badlock)
  - Cifs.ko version updated to 2.09
  - Minor fixes: including NetApp DFSpathname issue, Improved reconnection support and POSIX pathname and special character (trailing colon and space)
- 4.8 (18 changesets)
  - Allow mounts with prefixpath where top of share unaccessible
  - Fix for create when existing directory of same name
  - Misc minor fixes

Fixes and Features in progress



- Prefix path fixes
- Improved POSIX compatibility
- Improved reconnect and HA support
- Encrypted Share support
- ACLs and security improvements

Linux CIFS/SMB3 client bug status summary

High Level View of SMB3 Status

- SMB3 support is solid (and large file I/O FAST!), but lacks some optional advanced features (witness protocol integration e.g.) and a few basic features (ACL integration)
  - Metadata performance expected to be slower (need to add open/query compounding)
- SMB3 faster than CIFS (and sometimes NFS) for large file I/O
- SMB3 posix emulation is ok (use mount options “sfu” and “mfsymlinks”) but worse the cifs to Samba (and nfs)
- Can mount with SMB2.02, SMB2.1, SMB3, SMB3.02, 3.1.1
  - Specify vers=2.0 or vers=2.1 or 3.0 or 3.02 or 3.1.1 on mount

SMB3 Capabilities supported

- SMB2 CAP DFS
- SMB2 CAP LEASING
- SMB2 CAP LARGE_MTU
- SMB2 CAP PERSISTENT HANDLES
  - Client support added in Linux kernel 4.4
  - (NB: Samba server support is in progress)
- In progress
  - SMB2 CAP ENCRYPTION
- Unsupported capabilities
  - SMB2 CAP DIRECTORY LEASING
  - SMB2 CAP MULTI CHANNEL (not in client, though is supported in Samba server since Samba server version 4.4)

Copy Offload – big performance win

```
root@ubuntu:~# dd if=/dev/zero of=/mnt1/30M count=300 bs=100K
300+0 records in
300+0 records out
30720000 bytes (31 MB) copied, 0.445072 s, 69.0 MB/s
root@ubuntu:~# ls /mnt1
30M  3M  copy-of-3M  normal-non-ss-copy-of-3M  public
root@ubuntu:~# rm /mnt1/copy-of-3M
root@ubuntu:~# rm /mnt1/normal-non-ss-copy-of-3M
root@ubuntu:~# time cp /mnt1/3M /mnt1/normal-non-ss-copy-of-3M

real    0m0.068s
user    0m0.000s
sys     0m0.032s
root@ubuntu:~# time cp /mnt1/30M /mnt1/normal-non-ss-copy-of-30M

real    0m0.484s
user    0m0.000s
sys     0m0.351s
root@ubuntu:~# time cp --reflink /mnt1/3M /mnt1/ss-copy-of-3M

real    0m0.018s
user    0m0.000s
sys     0m0.007s
root@ubuntu:~# time cp --reflink /mnt1/30M /mnt1/ss-copy-of-30M

real    0m0.020s
user    0m0.000s
sys     0m0.010s
root@ubuntu:~#
```

DUPLICATE_EXTENTS is very efficient

```
    520 8.758876000 192.168.93.136   192.168.93.130  SMB2        342 Create Request File: ss-copy3-of-30M
    521 8.759457000 192.168.93.130   192.168.93.136  SMB2        334 Create Response File: ss-copy3-of-30M
    522 8.759611000 192.168.93.136   192.168.93.130  SMB2        175 GetInfo Request FILE_INFO/SMB2_FILE_INTERNAL_INFO File: ss-copy3-of-30M
    523 8.759911000 192.168.93.130   192.168.93.136  SMB2        150 GetInfo Response
    526 8.760144000 192.168.93.136   192.168.93.130  SMB2        191 Ioctl Request FILE_SYSTEM Function:0x0031 File: ss-copy3-of-30M
    527 8.760487000 192.168.93.130   192.168.93.136  SMB2        182 Ioctl Response FILE_SYSTEM Function:0x0031 File: ss-copy3-of-30M
                    192.168.93.136   192.168.93.130  SMB2        174 SetInfo Request FILE_INFO/SMB2_FILE_ENDOFFILE_INFO File: ss-copy3-of-30M
                    192.168.93.130   192.168.93.136  SMB2        136 SetInfo Response
    530 8.761086000 192.168.93.136   192.168.93.130  SMB2        230 Ioctl Request FILE_SYSTEM Function:0x00d1 File: ss-copy3-of-30M
    531 8.761481000 192.168.93.130   192.168.93.136  SMB2        182 Ioctl Response FILE_SYSTEM Function:0x00d1 File: ss-copy3-of-30M
    532 8.761610000 192.168.93.136   192.168.93.130  SMB2        158 Close Request File: ss-copy3-of-30M
    533 8.767873000 192.168.93.130   192.168.93.136  SMB2        194 Close Response
```

LibreOffice Impress

```
▶ Frame 530: 230 bytes on wire (1840 bits), 230 bytes captured (1840 bits) on interface 0
▶ Ethernet II, Src: Vmware_b4:dc:f2 (00:0c:29:b4:dc:f2), Dst: Vmware_84:48:c0 (00:0c:29:84:48:c0)
▶ Internet Protocol Version 4, Src: 192.168.93.136 (192.168.93.136), Dst: 192.168.93.130 (192.168.93.130)
▶ Transmission Control Protocol, Src Port: 41774 (41774), Dst Port: microsoft-ds (445), Seq: 1469, Ack: 1432, Len: 164
▶ NetBIOS Session Service
▼ SMB2 (Server Message Block Protocol version 2)
  ▶ SMB2 Header
  ▼ Ioctl Request (0x0b)
    ▶ StructureSize: 0x0039
    ▶ Function: Unknown (0x00098344)
    ▶ GUID handle File: ss-copy3-of-30M
      Max Ioctl In Size: 0
      Max Ioctl Out Size: 65280
    ▶ Flags: 0x00000001
```

Duplicate Extents vs CopyChunk for server side copy (to REFS)

```
root@ubuntu:~/xfstests-new/xfstests-dev# dd if=/dev/zero of=/mnt1/500M count=500 bs=1M
500+0 records in
500+0 records out
524288000 bytes (524 MB) copied, 17.212 s, 30.5 MB/s
root@ubuntu:~/xfstests-new/xfstests-dev# time cp /mnt1/500M /mnt1/normal-copy-500M

real    0m19.972s
user    0m0.004s
sys     0m0.289s
root@ubuntu:~/xfstests-new/xfstests-dev# ./src/cloner /mnt1/500M /mnt1/copy-chunk-500M
root@ubuntu:~/xfstests-new/xfstests-dev# time ./src/cloner /mnt1/500M /mnt1/copy-chunk-500M

real    0m0.531s
user    0m0.000s
sys     0m0.061s
root@ubuntu:~/xfstests-new/xfstests-dev# time ./src/cloner /mnt1/500M /mnt1/copy-chunk-500M-try2

real    0m18.513s
user    0m0.000s
sys     0m0.075s
root@ubuntu:~/xfstests-new/xfstests-dev# time cp --reflink /mnt1/500M /mnt1/reflink-copy-500M

real    0m0.034s
user    0m0.000s
sys     0m0.009s
root@ubuntu:~/xfstests-new/xfstests-dev#
```

Amazon

CopyChunk server (to NTFS) – times vary less new vs. existing target

```
root@ubuntu:~/xfstests-new/xfstests-dev/src# time dd if=/dev/zero of=/mnt1/200M
count=100 bs=2M
100+0 records in
100+0 records out
209715200 bytes (210 MB) copied, 5.3544 s, 39.2 MB/s

real    0m5.363s
user    0m0.000s
sys     0m4.643s
root@ubuntu:~/xfstests-new/xfstests-dev/src# mount -t cifs //192.168.93.142/publ
ic /mnt1 -o username=Administrator,vers=3.02,noperm,sfu^C
root@ubuntu:~/xfstests-new/xfstests-dev/src# ^C
root@ubuntu:~/xfstests-new/xfstests-dev/src# time ./cloner /mnt1/200M /mnt1/copy
chunk-of-200M

real    0m0.313s
user    0m0.000s
sys     0m0.032s
root@ubuntu:~/xfstests-new/xfstests-dev/src# time ./cloner /mnt1/200M /mnt1/copy
chunk-of-200M


real    0m0.250s
user    0m0.000s
sys     0m0.028s
root@ubuntu:~/xfstests-new/xfstests-dev/src#
root@ubuntu:~/xfstests-new/xfstests-dev/src# time ./cloner /mnt1/200M /mnt1/copy
chunk-of-200M-two

real    0m0.335s
user    0m0.000s
sys     0m0.029s
root@ubuntu:~/xfstests-new/xfstests-dev/src# time ./cloner /mnt1/200M /mnt1/copy
chunk-of-200M-two

real    0m0.240s
user    0m0.000s
sys     0m0.029s
```

Better HA: Persistent and Resilient Handles

- New mount options (and code to add corresponding create contexts etc.)
  - “resilienthandles”
  - “persistenthandles”
- Status of remaining items:
  - Add channel sequence number on reconnect
  - Improve server to server failover
    - Alternate DFS targets in DFS referrals
    - Witness protocol server or share redirection

fallocate

- We currently support
  - Simple fallocate
  - PUNCH_HOLE
  - ZERO_RANGE
  - KEEP_SIZE
- We have discussed ways to add support for the remaining two when the server supports duplicate extents (currently REFS on Windows 2016 is the only one that advertises “FS_SUPPORTS_BLOCK_REFCOUNTING” capability). We can add support for:
  - COLLAPSE_RANGE
  - INSERT_RANGE

SMB3 and ACLs

SMB3 Security Features

- SMB3.1.1 secure negotiate
- SMB3 Share Encryption

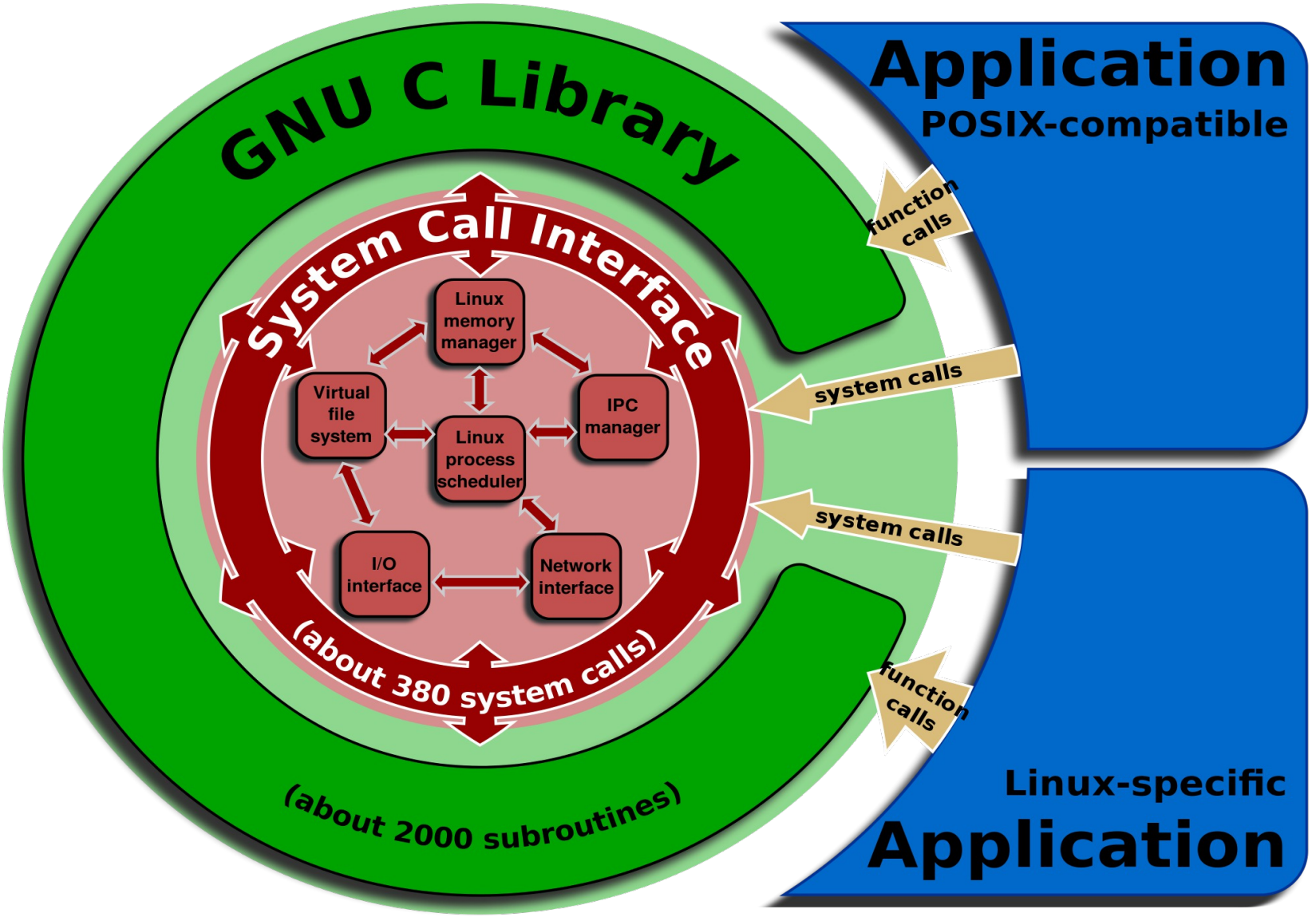
Other Optional features

- Xstat integration
  - Returns birth time and dos attributes in more standardized fashion (cifs has a private xattr for that, but few tools use it). Kernel patches exist, would help cifs a lot
- IOCTL to list alternate data streams
  - Querying data in alternate data streams (e.g. for backup) requires disabling posix pathnames (due to conflict with ":")
- Clustering, Witness protocol integration
- DFS reconnect to different DFS server
- See the performance slide, coming up ...
- Other suggestions ...

SMB3 and Performance

- Key Features
  - Compounding
  - Large file I/O
  - File Leases
    - Lease upgrades
  - Directory Leases
  - Copy Offload
  - Multi-Channel
    - And optional RDMA
  - Linux specific protocol optimizations

POSIX and Linux Compatibility



POSIX/Linux Compatibility Approach 1 Best Effort Emulation with SMB3

- *Implemented:*
  - *Hardlinks*
- *Emulated: (current cifs.ko SMB3 code)*
  - *POSIX Path Names:* Approximately 7 reserved characters not allowed in SMB3/NTFS etc. *(e.g. ? * \ : !)*
  - *Symlinks* (ala “mfsymlinks” Minshall-French symlinks, use “mfsymlinks” mount option)
  - *Pseudo-Files:* FIFOs, Pipes, Character Devices (ala “sfu” aka “Microsoft services for unix” use “sfu” mount option)
- *Partial:*
  - *Extended attribute flags* (lsattr/chattr) including compressed flag
  - *POSIX stat and statfs info*
  - *POSIX Byte Range Locks*
- *Not implemented, but emulatable with combination of SMB3 features and/or POSIX Extensions or even use of Apple AAPL create context*
  - *Xattrs* (Security/Trusted for SELinux, User xattrs for apps)
  - *POSIX Mode Bits*
  - *POSIX UID/GID ownership information*
  - *Case Sensitivity* in opening paths
- *Not solvable without additional extensions:*
  - *POSIX Delete (unlink) Behavior*

Approach 1: Enhance support for existing SMB3 features some servers already support

- Get mode from SMB3 ACL (or combination of that and SMB2_CREATE_QUERY_MAXIMAL_ACCESS_REQUEST create context)
- Recognize case sensitive volume at mount time and detect cases where server 'lies' about it
- Cleanup Microsoft “nfs symlink” code to better recognize this symlink (reparse point)
- Implement level 11 SMB2_QUERY_FS_INFO in Samba get “PhsyicalBytesPerSectorForPerformance” and map to statfs f_bsize
- Doesn't address posix byte range locking fully, nor does it always address case sensitive posix path names, nor conflict between streams (which have : separating the file and ADS name) and posix paths (which allow : in the name)

Approach 2 Use AAPL context on open

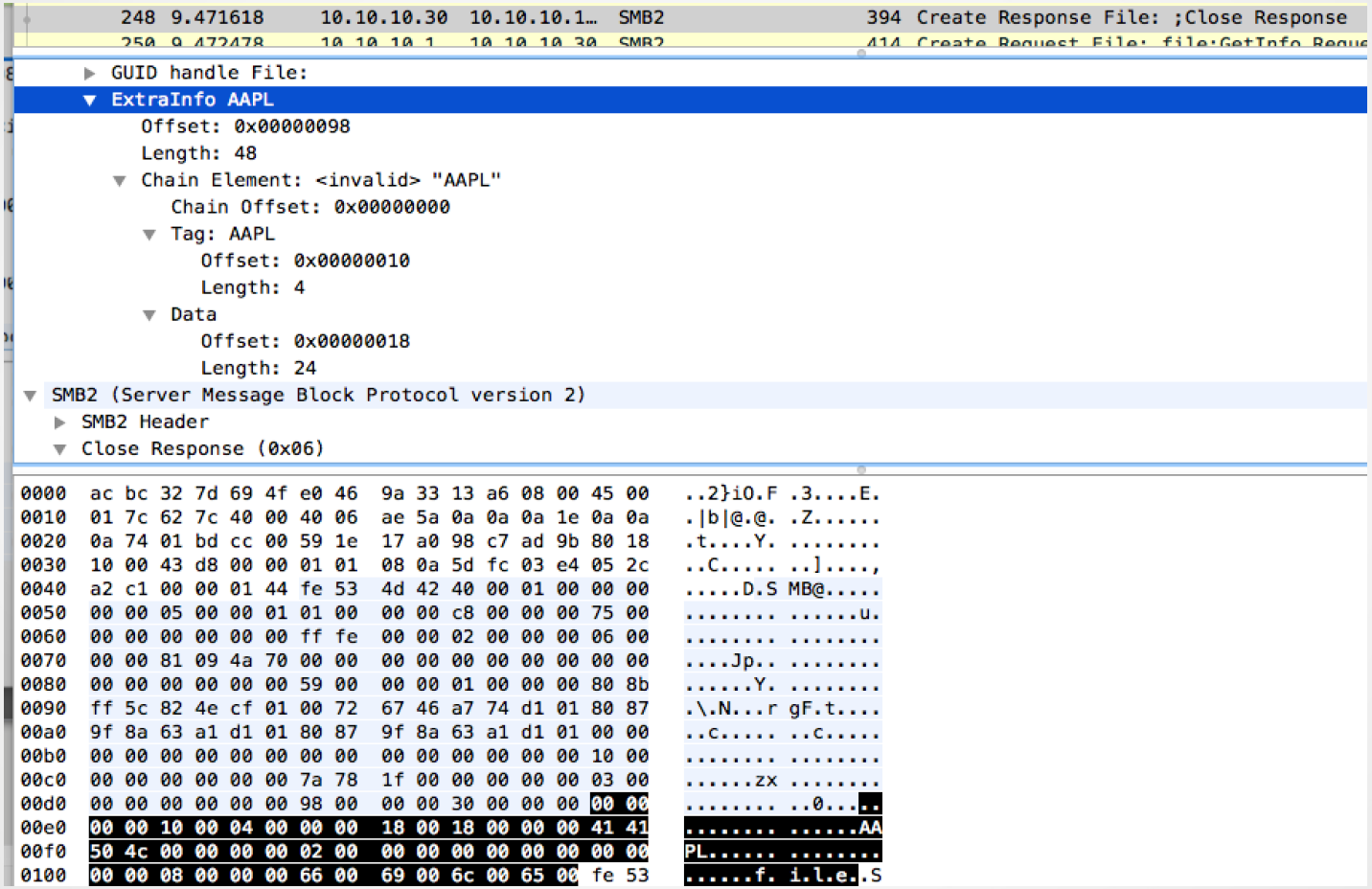
- Implement AAPL context
  - Improved Mac interop is another benefit
  - Samba even has a vfs_fruit module that adds other interesting features (spotlight integration e.g.)
- Subset of POSIX requirements can be solved
- kAAPL_SERVER_CAPS = 0x01,
  - kAAPL_SUPPORTS_READ_DIR_ATTR = 0x01,
  - kAAPL_SUPPORTS_OSX_COPYFILE = 0x02,
  - kAAPL_UNIX_BASED = 0x04
  - kAAPL_SUPPORTS_NFS_ACE = 0x08
- kAAPL_VOLUME_CAPS = 0x02,
  - kAAPL_SUPPORT_RESOLVE_ID = 0x01,
  - kAAPL_CASE_SENSITIVE = 0x02
- kAAPL_MODEL_INFO = 0x04 (pad, length, model string)

Approach 2 (continued) – Mac example

```
fset: 0x00000080
ngth: 40
ain Element: <invalid> "AAPL"
 Chain Offset: 0x00000000
 Tag: AAPL
    Offset: 0x00000010
    Length: 4
 Data
    Offset: 0x00000018
    Length: 16
er Message Block Protocol version 2)
```

```
)a 33 13 a6 ac bc  32 7d 69 4f 08 00 45 00   .F.3.... 2}iO..E.
f1 c2 40 00 40 06  1f 58 0a 0a 0a 74 0a 0a   .8..@.@. .X...t..
:c 00 01 bd 98 c7  ac 97 59 1e 17 a0 80 18   ........ ..Y.....
4e f9 00 00 01 01  08 0a 05 2c a2 c1 5d fc   ..N..... ...,..].
)0 00 01 00 fe 53  4d 42 40 00 01 00 00 00   .......S MB@.....
)5 00 00 01 00 00  00 00 a8 00 00 00 75 00   ........ ......u.
)0 00 00 00 ff fe  00 00 02 00 00 00 06 00   ........ ........
31 09 4a 70 00 00  00 00 00 00 00 00 00 00   ....Jp.. ........
```

Mac example (continued)



Approach 3 – POSIX Extensions for SMB3!

- See Jeremy's talk here and the working session on these extensions

SMB3 Performance Overview

Testing … testing … testing

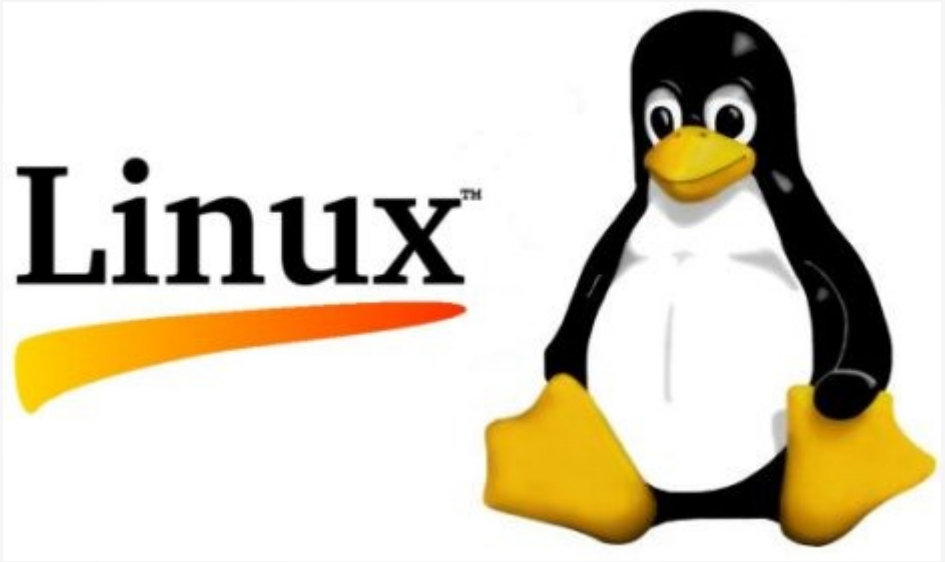
- Continue work on improving xfstest automation
- Can now use “scratch” mount with cifs.ko expanding the range of xfstests that can run against cifs or smb3 mounts
- Need to cleanup some bugs found by xfstest to remove 'noise' and make it easier to identify and fix any regressions early

XFSTEST details

- Continued improvements in automated testing is very important
- Surprising number work even to SMB3 without POSIX support
- Some tests fail due to lack of posix permissions (mode bits)
- Various tests fail due to falloc (missing features, and a bug)
- Failures due to other missing posix features
  - Advisory locking (e.g. test 131)
- Misc. failures and timestamp coherence client/server
  - Really hard to get mtime consistent on client/server in network file systems

- The Future of SMB3 and Linux is very bright
- Let's continue its improvement!

Thank you for your time



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
  - http://www.samba.org
  - 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
    - 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