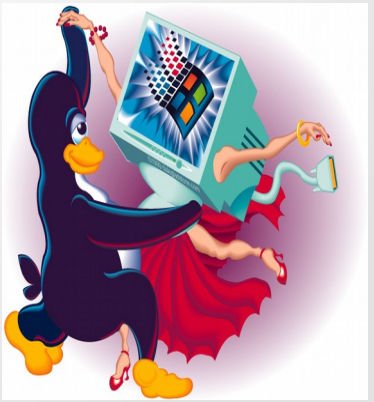
SMB3 and Beyond for Linux: State of Unix Extensions, as We Drive Toward Optimal POSIX Compatibility and Performance

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

- A year in review … general Linux file system status
- Key Feature Status
  - Goals
  - Completed Features
  - In progress features
  - Wish list
- What works?
  - File system test results much improved
  - What fails or is not supported
    - The solution: better POSIX compatibility through extensions
  - What about performance?
- For more information ...

Some key features helping drive discussions and FS development activity ?

- Many evolving general storage features are especially critical for NAS
  - Better support for NVMe
    - RDMA, low latency ways to access VERY high speed storage
  - Faster (and more) network interfaces
  - Security/crypto Improvements
    - And even RichACL (maybe someday … we can hope …)
  - Improved end-to-end reliability and failure handling
  - statx (extended stat)
  - Improved copy offload
  - Improved sparse file support (including for virtualization)
  - Shift to object like access patterns for more workloads

A year ago … and now … kernel (including cifs/smb3 client) improving

- 12 months ago we had Linux version 4.8-rc5 ie “Psychotic Stoned Sheep”

And then last week we got 4.14-“Fearless Coyote”

Working with great developers. Here we are at 2017 Linux File System Summit in Cambridge

Most Active Linux Filesystems this year

- 4560 kernel filesystem changesets in last year (since 4.8-rc5 kernel)!
  - Linux kernel file system activity continuing strong (up slightly)
  - FS activity accounts for 5.35% of overall kernel changes (dominated by drivers) but fs activity monitoredcarefully
  - Kernel is now > 17 million lines of source code (measured last week with sloccount tool) vs.
    - Last year kernel size was under 15 million lines of code!
- There are many Linux file systems, but seven (and the VFS layer itself) drive the majority of activity
  - File systems represent about 5.1% of the overall kernel source code (868,000 lines of code in Linux fs, up 2%)
- cifs.ko (cifs/smb3 client) among more active fs
  - VFS (overall fs mapping layer and common functions) 697 (up)
  - Btrfs 652 changesets (down a lot, but still most active individual fs by far)
  - XFS 534 (up a lot)
  - Nfs client 454 (flat)
  - Ext4 312 (up a lot)
  - CIFS/SMB2/SMB3 client 182 (up 40% over previous year)
    - cifs.ko is 44,244 lines of kernel code (not counting user space helpers, and samba userspace tools, it grew 5.2%)
  - Nfs server 142 (activity down slightly)
  - Ceph 131 (down)
- NB: Samba (cifs/smb2/smb3 server) is as active as the top 3 or 4 put together (thousands of changesets) since it is broader in scope (by a lot) and also is in user space not in kernel

What are the goals?

Did we mention the confusing names?

- CIFS is a dialect we are trying to deprecate but ...
- cifs.ko is the overall kernel module name for all smb dialects (with a mount helper “mount.cifs” and a type “mount -t cifs ...”) but …
  - SMB3 is the default dialect (and we want to discourage use of CIFS/SMB1 dialect)
- Is it time to break the module into multiple pieces including an smb3.ko so we aren't using a “cifs” module?

Dialect Upgrade … SMB3 default

- New behavior. SMB3 used by default so will see in dmesg:
  - “Kernel: No dialect specified on mount. Default has changed to a more secure dialect, SMB3 (vers=3.0), from CIFS (SMB1). To use the less secure SMB1 dialect to access old servers which do not support SMB3 specify vers=1.0 on mount. For somewhat newer servers such as Windows 7 try vers=2.1.”
- Next release will have multiprotocol
  - SMB3.02, SMB3, SMB2.1
  - SMB3.1.1 is almost done

Fixes and Features – big progress! remember list from last years SDC?

- ~~Prefix path fixes~~
- ~~Improved POSIX compatibility~~ (some work in progress e.g. SMB3 POSIX Extensions)
- ~~Return important SMB3 inode metadata via xattrs (create time, attributes, ADS names)~~
- ~~Improved reconnect and HA support~~
- ~~Encrypted Share support~~
- ~~ACLs and security improvements~~

What are most noticeable, most important improvements over last year?

- Dialect upgrade (no more insecure cifs default)
- MUCH improved SMB3 support

Fixes and Features by release

- Linux 4.2 (14 changesets)
  - Initial (minimal) 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\
  - mfsymlink support added for smb2/smb3 (symlink emulation, also used by Mac)
  - Misc minor fixes

Fixes and Features (continued)

- 4.9 (37 changesets)
  - Various reconnect improvements (e.g. send echo ASAP to reconnect smb session/tcon quicker after socket reconnect)
  - Uid/gid from special sid (new mount option “idsfromsid”)
  - Can override number of credits (new mount option “max_credits”)
  - Query file attributes or creation time via xattr
- 4.10 (17)
  - New snapshot mount parm (“snapshot”)
  - Misc bug fixes
- 4.11 (51 changesets)
  - SMB3 reconnect improvements (including better persistent & durable handles). Much higher reliability now when server crashes or failsover while I/o in flight or cached. Lots of corner cases fixed (Thank you Germano!)
  - Server side copy works much better: Clone file range (and “cp –reflink” command) now support more common “copychunk” copy offload style (had required ess common “duplicate extents” support). Thank you Sachin!
  - SMB3 DFS support (Thank you Aurelien!)
  - SMB3 Encryption support (Thank you Pavel!)

Fixes and Features (continued)

- 4.12 (36 changesets)
  - Posix smb3 name mapping improvements
  - Improved aio support
  - Add support for enumerating snapshots
  - Bug fixes
- 4.13 (27 changesets)
  - Change **default dialect to SMB3** from CIFS
  - Smb3 support for “cifsacl” mount option (and mode emulation)
  - Bug fixes
- 4.14 (10 changesets so far …)
  - Xattr enablement
  - Bug fixes (including reconnect improvements)
- Coming soon:
  - Multidialect support (SMB2.1 through SMB3.1.1)
  - RDMA !
  - POSIX Extensions for SMB3 (even if experimental)
  - SMB3.1.1 improvements

statx()

- After multiple years of technical discussion it is now merged into Linux kernel!
- VFS support and system call added In 4.11 kernel
  - CIFS enablement planned for 4.14 or 4.15
  - Can return creation time and a few new attributes (including e.g. 'compressed')
  - Extensible, more flags coming (for query more metadata that Samba and cifs.ko care about)
  - 'set' for statx also planned to be added into the vfs (then in cifs) but this first step is important

What about snapshots?

- Can enumerate snapshots e.g. with relatively new ioctl
  - CIFS_ENUMERATE_SNAPHSHOTS
  - SMB2.1 or later (in our implementation)
- New mount parameter
  - Snapshot=…
  - We will allow mounting of snapshots

What about SMB3 and ACLs?

```
@ubuntu:~/test1$ ls -la
otal 40
rwxrwxr-x   2 sfrench sfrench  4096 Sep 12 16:09 .
rwxr-xr-x 218 sfrench sfrench 28672 Sep 12 16:05 ..
r--r-----   1 sfrench sfrench     0 Sep 12 16:08 0440
rwx------   1 sfrench sfrench     0 Sep 12 16:08 0700
rwxrwxrwx   1 sfrench sfrench     0 Sep 12 16:08 0777
rw-r--r--   1 root    root        0 Sep 12 16:09 root-file
rwxr-----   1 sfrench sfrench  1362 Jul 12 21:32 test
@ubuntu:~/test1$ ls -la /mnt/test1
otal 1024
rwxrwxr-x 2 root root    0 Sep 12 16:09 .
rwxr-xr-x 2 root root    0 Sep 12 16:05 ..
r--r----- 1 root root    0 Sep 12 16:08 0440
rwx------ 1 root root    0 Sep 12 16:08 0700
rwxrwxrwx 1 root root    0 Sep 12 16:08 0777
rw-r--r-- 1 root root    0 Sep 12 16:09 root-file
rwxr----- 1 root root 1362 Jul 12 21:32 test
@ubuntu:~/test1$ ls -la /smb3-mnt-without-cifsacl/test1
otal 1024
rwxr-xr-x 2 root root    0 Sep 12 16:09 .
rwxr-xr-x 2 root root    0 Sep 12 16:05 ..
r-xr-xr-x 1 root root    0 Sep 12 16:08 0440
rwxr-xr-x 1 root root    0 Sep 12 16:08 0700
rwxr-xr-x 1 root root    0 Sep 12 16:08 0777
rwxr-xr-x 1 root root    0 Sep 12 16:09 root-file
```

Wish List (TODOs)

- Security
  - Finish up SMB3.1.1 secure negotiate
- Performance
- New function

What works?

- What does xfstest currently show as missing or broken?

XFSTest improvements

- Running earlier today noticed:
  - Xfstest has improved for SMB3 with recent cifs.ko patches:
    - e.g. Generic/029 and generic/030 tests succeed
- A few areas to dig into:
  - Xattrs: Generic/020 now runs (now that we have smb3 xattr support) but test fails

Fallocate (works, but minor TODOs)

- 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

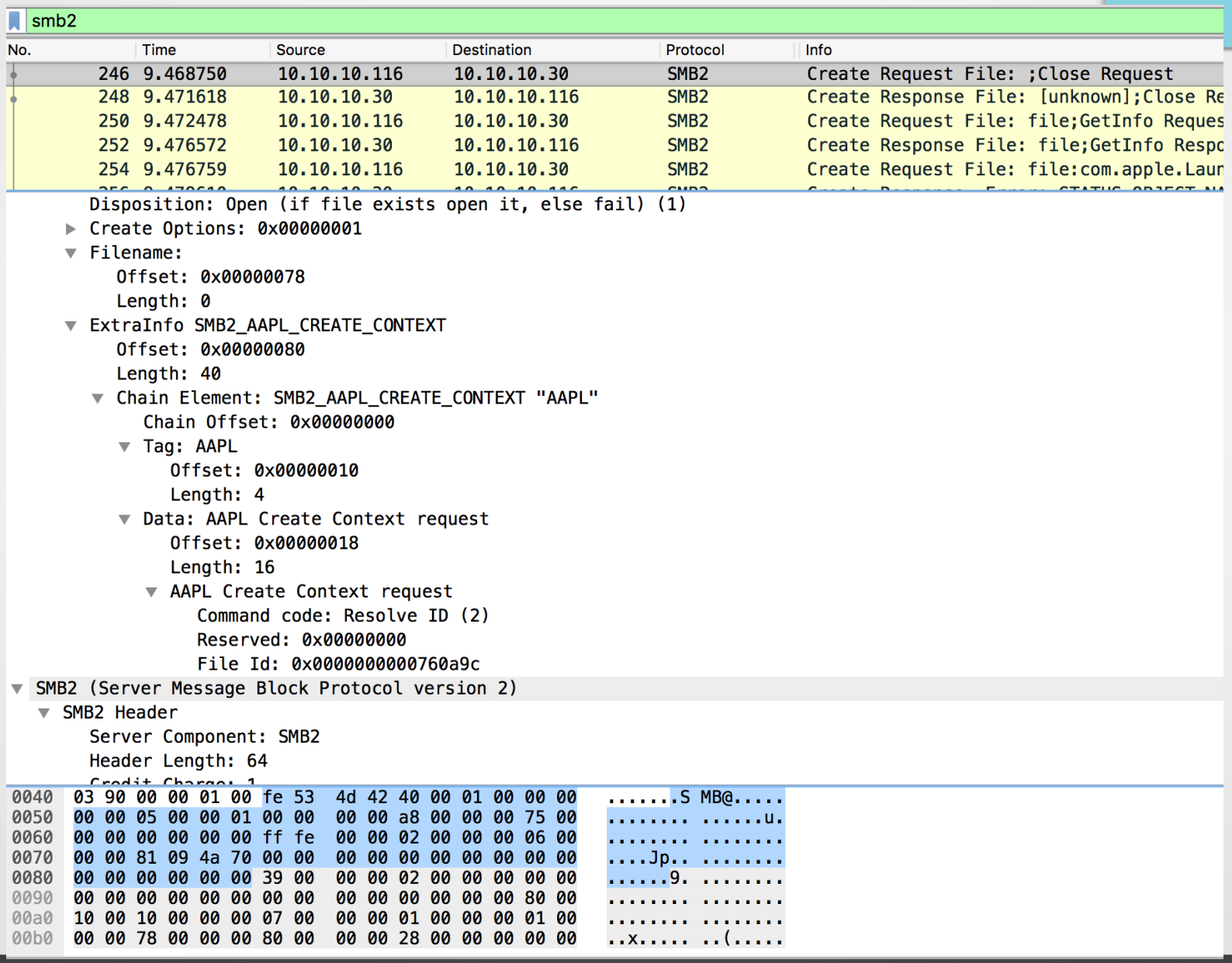
SMB3 POSIX Compatibility – what does it look like now

- For SMB3 can already handle (without extensions)
  - POSIX pathnames with reserved chars eg
    - “>” or “<” or “:” or “*” and even trailing “.” or space
  - Can create hardlinks
  - Can retrieve the mode
  - Can retrieve the owner if set using well known SID
  - Emulate rename/delete semantics as well as possible
- Can't handle
  - Case sensitive paths (only case preserving). Even need this to build kernel on smb3 mount
  - POSIX byte range locking
  - A few fields in statfs
  - Exact POSIX semantics of delete/rename (some access denied cases)

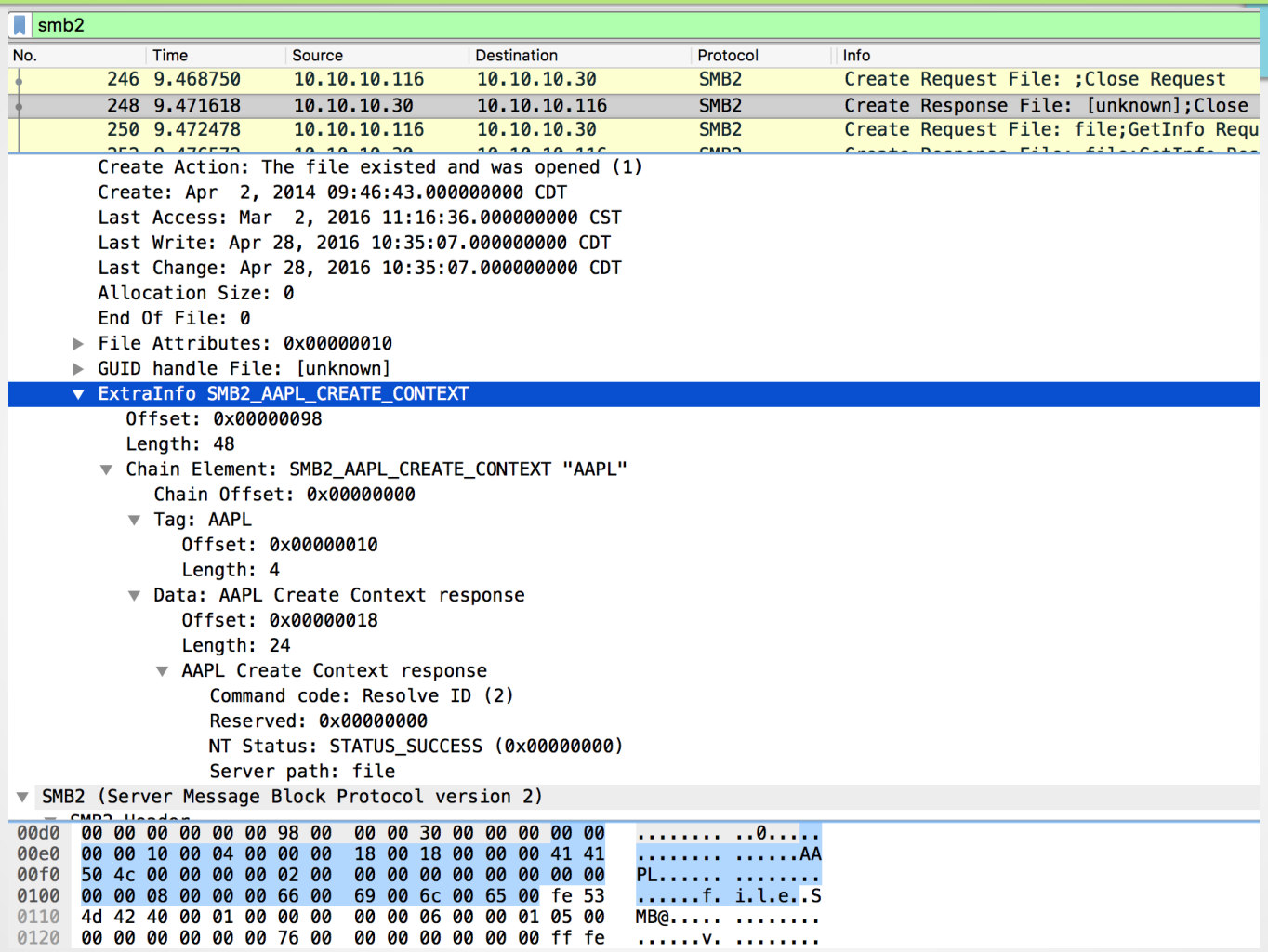
SMB3 POSIX Extensions?

- With SMB3 (new default dialect), what is the current POSIX emulation behavior with and without extensions?
- Principles of extensions (see talk with JRA and me tomorrow)
  - POSIX Create Context support is indicated by negotiate context
  - If server supports POSIX Create Context
    - Server will either accept or reject a create with a posix create context but won't ignore it (so we don't have file name collisions or unintended consquences)
    - POSIX Creates that succeed include
      - Posix delete/rename/byte range locks behavior
      - Case sensitivity
      - POSIX query info file, query dir
  - Mode (and owner) is inferred from ACL
  - Owner is SID, not uid
  - Symlinks are client followed, symlinks are opaque to server (so don't introduce security issues)
    - Mfsymlinks (and in future will optionally allow Windows NFS symlink reparse point)
  - SFM mapping is used for reserved POSIX characters

Note that Apple create context (AAPL) can be used for some of this



And the response:



SMB3 and Performance (focus areas **highlighted,** already supported areas normal text)

- Key Features
  - Async and vectored I/O
  - **Compounding (reduce number of roundtrips)**
  - Large file I/O
  - File Leases
    - **Lease upgrades**
  - **Directory Leases**
  - Copy Offload
  - **Multi-Channel**
    - **And optional RDMA**
  - Linux specific protocol optimizations

SMB3 RDMA Status

- Very exciting, preliminary performance results promising!
  - With 40Gbit adapter and Long Li's patches seeing
    - Using queue depth of 16, and 1MB I/O size
    - 88% network utilization on read (vs. 90% on Windows which can use multichannel more fully)
    - 70% on write
- Plan to begin merging soon
  - Low risk protocol definitions going in soon
- See Long Li's presentation from earlier today

Good year for SMB3

- Recap of highlights

Thank you for your time

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

For more information: 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