SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

SMB3 POSIX Extensions Phase 2 ... what is next?

Steven French

Principal Software Engineer
Microsoft – Azure Storage

Legal Statement

This work represents the views of the author(s) and does not necessarily reflect the views of Microsoft

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

- Linux is a lot more than POSIX ...
- Why do these extensions matter?
- Implementation Status. What works today?
- Some details
- The Future ...
- How to handle Linux continuing to

Linux > POSIX

- Currently huge number of sys

(try “git grep SYSCALL_DEFIN

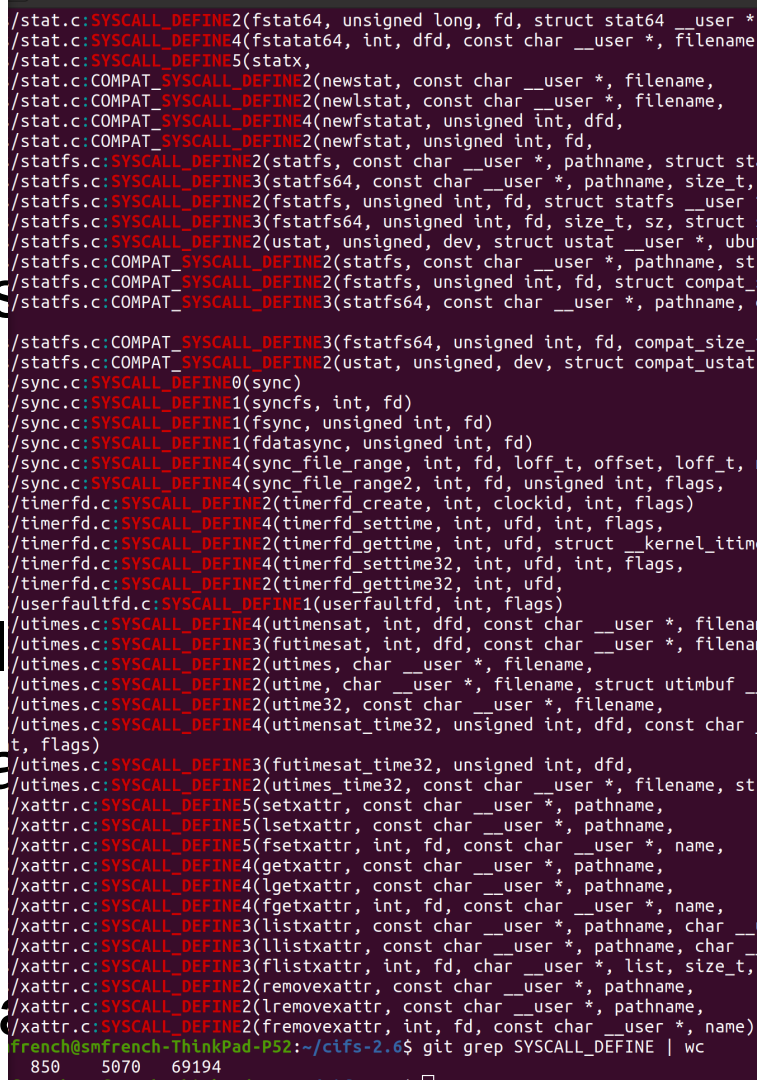
well over 850 and 500+ are

even documented “man syscall

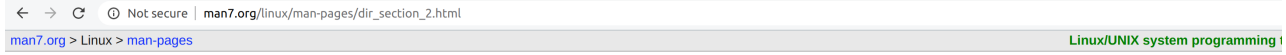
FS layer has 223). Verified toda

vs

- Only about 100 POSIX API c



513 syscalls with man pages!



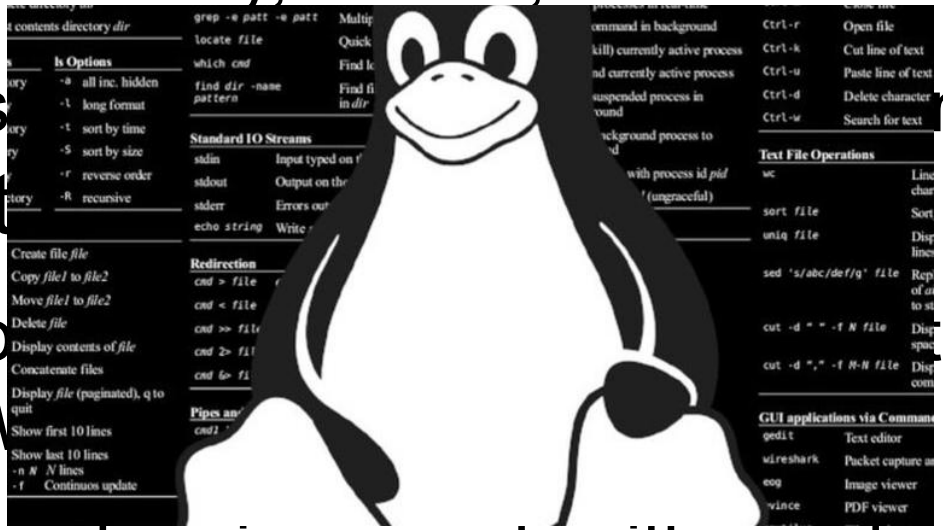
Linux man pages: section 2

accept(2)
accept4(2)
access(2)
acct(2)
add_key(2)
adjtimex(2)
afs_syscall(2)
alarm(2)
alloc_hugepages(2)
arch_prctl(2)
arm_fadvise(2)
arm_fadvise64_64(2)
arm_sync_file_range(2)
bdflush(2)
bind(2)
bpf(2)
break(2)
brk(2)
cacheflush(2)
capget(2)
capset(2)
chdir(2)
chmod(2)
chown(2)
chown32(2)
chroot(2)
clock_getres(2)
clock_gettime(2)
clock_nanosleep(2)
clock_settime(2)
clone(2)
clone2(2)
__clone2(2)
clone3(2)
close(2)
connect(2)
copy_file_range(2)
creat(2)
create_module(2)
delete_module(2)
dup(2)
dup2(2)
dup3(2)
epoll_create(2)
epoll_create1(2)
epoll_ctl(2)
ioctl_xfs_bulkstat(2)
ioctl_xfs_fsbulkstat(2)
ioctl_xfs_fscounts(2)
ioctl_xfs_fsgetxattr(2)
ioctl_xfs_fsgetxattra(2)
ioctl_xfs_fsinumbers(2)
ioctl_xfs_fsop_geometry(2)
ioctl_xfs_fssetxattr(2)
ioctl_xfs_getbmap(2)
ioctl_xfs_getbmapa(2)
ioctl_xfs_getbmapx(2)
ioctl_xfs_getresblks(2)
ioctl_xfs_goingdown(2)
ioctl_xfs_inumbers(2)
ioctl_xfs_scrub_metadata(2)
ioctl_xfs_setresblks(2)
io_destroy(2)
io_getevents(2)
ioperm(2)
iopl(2)
ioprio_get(2)
ioprio_set(2)
io_setup(2)
io_submit(2)
ipc(2)
isastream(2)
kcmp(2)
kexec_file_load(2)
kexec_load(2)
keyctl(2)
kill(2)
killpg(2)
lchown(2)
lchown32(2)
lgetxattr(2)
link(2)
linkat(2)
listen(2)
listxattr(2)
llistxattr(2)
llseek(2)
_llseek(2)
lock(2)
lookup_dcookie(2)
lremovexattr(2)
lseek(2)
rt_sigqueueinfo(2)
rt_sigreturn(2)
rt_sigsuspend(2)
rt_sigtimedwait(2)
rt_tgsigqueueinfo(2)
s390_guarded_storage(2)
s390_pci_mmio_read(2)
s390_pci_mmio_write(2)
s390_runtime_instr(2)
s390_sthyi(2)
sbrk(2)
sched_getaffinity(2)
sched_getattr(2)
sched_getparam(2)
sched_get_priority_max(2)
sched_get_priority_min(2)
sched_getscheduler(2)
sched_rr_get_interval(2)
sched_setaffinity(2)
sched_setattr(2)
sched_setparam(2)
sched_setscheduler(2)
sched_yield(2)
seccomp(2)
security(2)
select(2)
select_tut(2)
semctl(2)
semget(2)
semop(2)
semtimedop(2)
send(2)
sendfile(2)
sendfile64(2)
sendmmsg(2)
sendmsg(2)
sendto(2)
setcontext(2)
setdomainname(2)
setegid(2)
seteuid(2)
setfsgid(2)
setfsgid32(2)
setfsuid(2)
setfsuid32(2)
setgid(2)

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

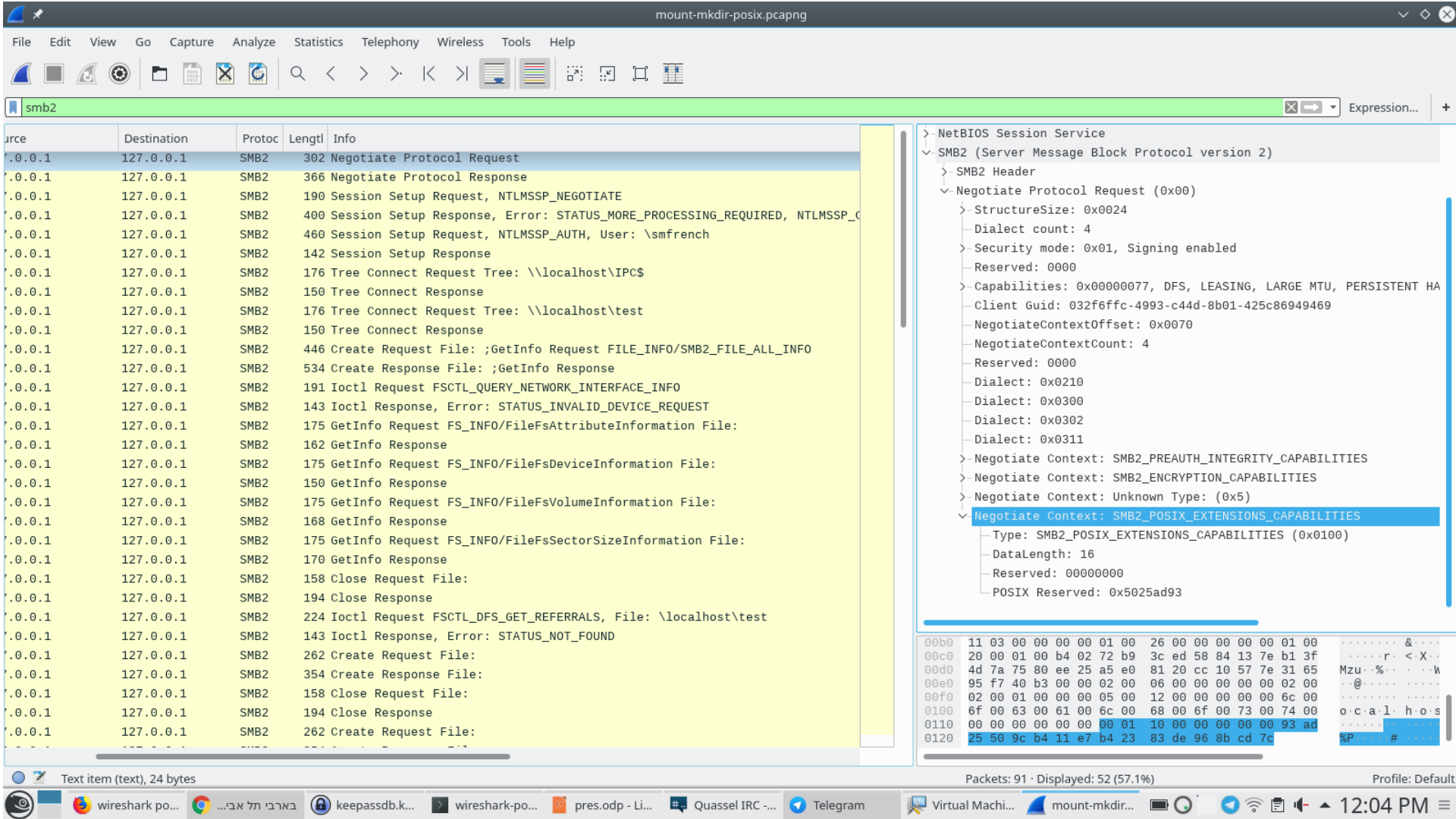
- **Network File systems matter**
  - these extensions to most popular network fs protoco
    (SMB3) are important
  - block devices struggle with file system tasks:
    locking, security, leases, consistent metadata
- Linux Apps mounts and
  continue t
- Improve co tomers have
  Linux and W

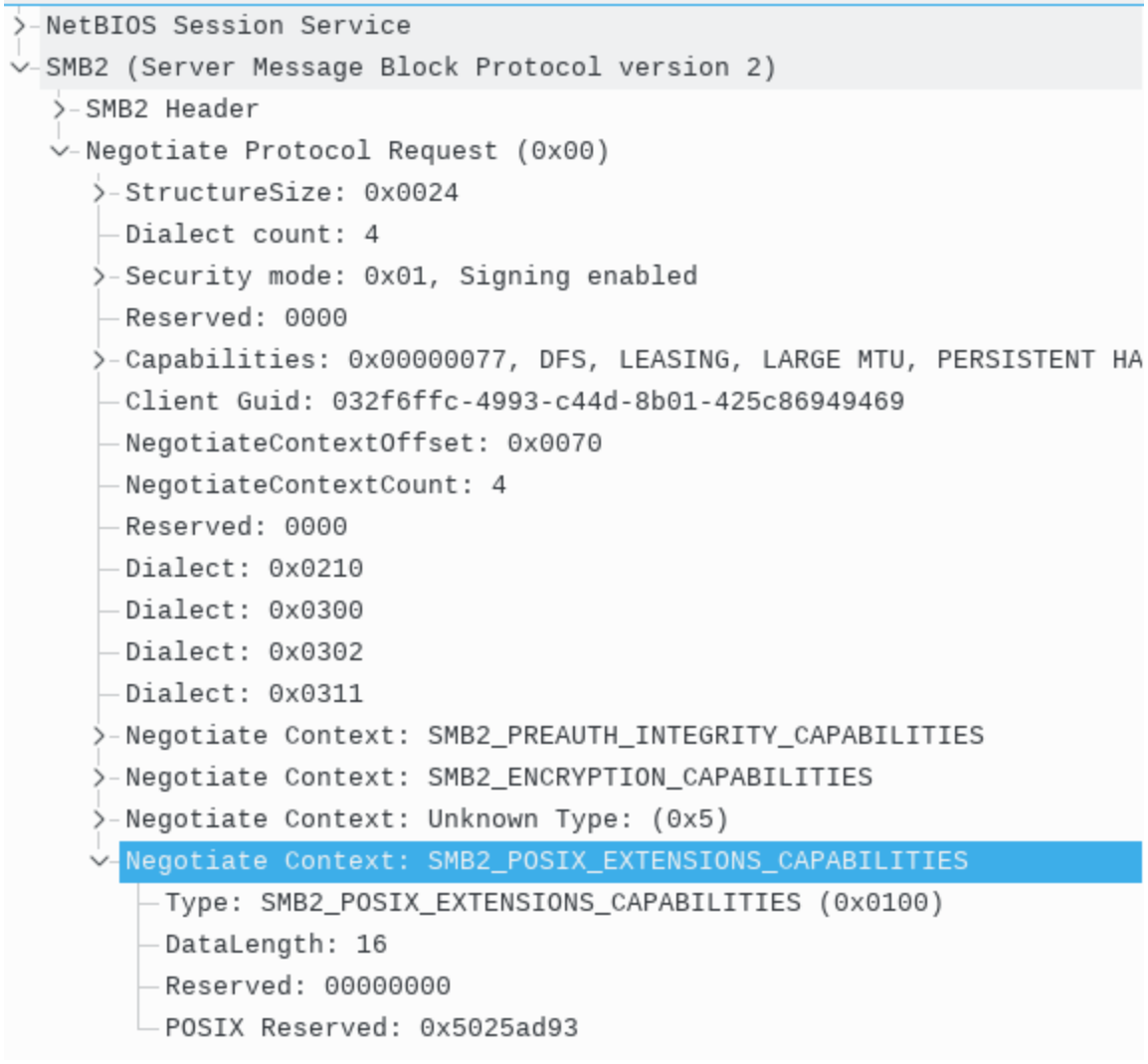


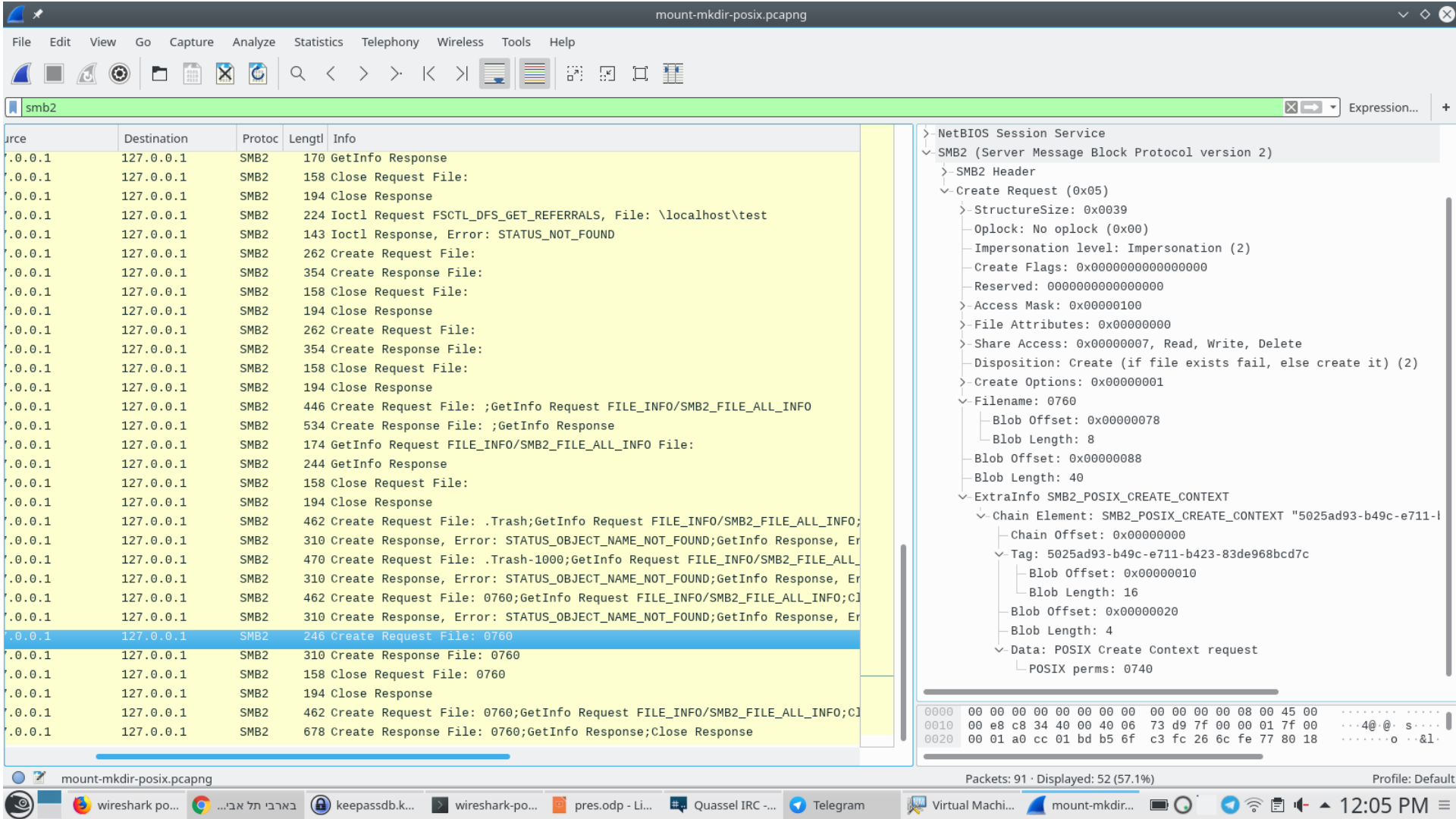
Make sure extensions work with most secure, most

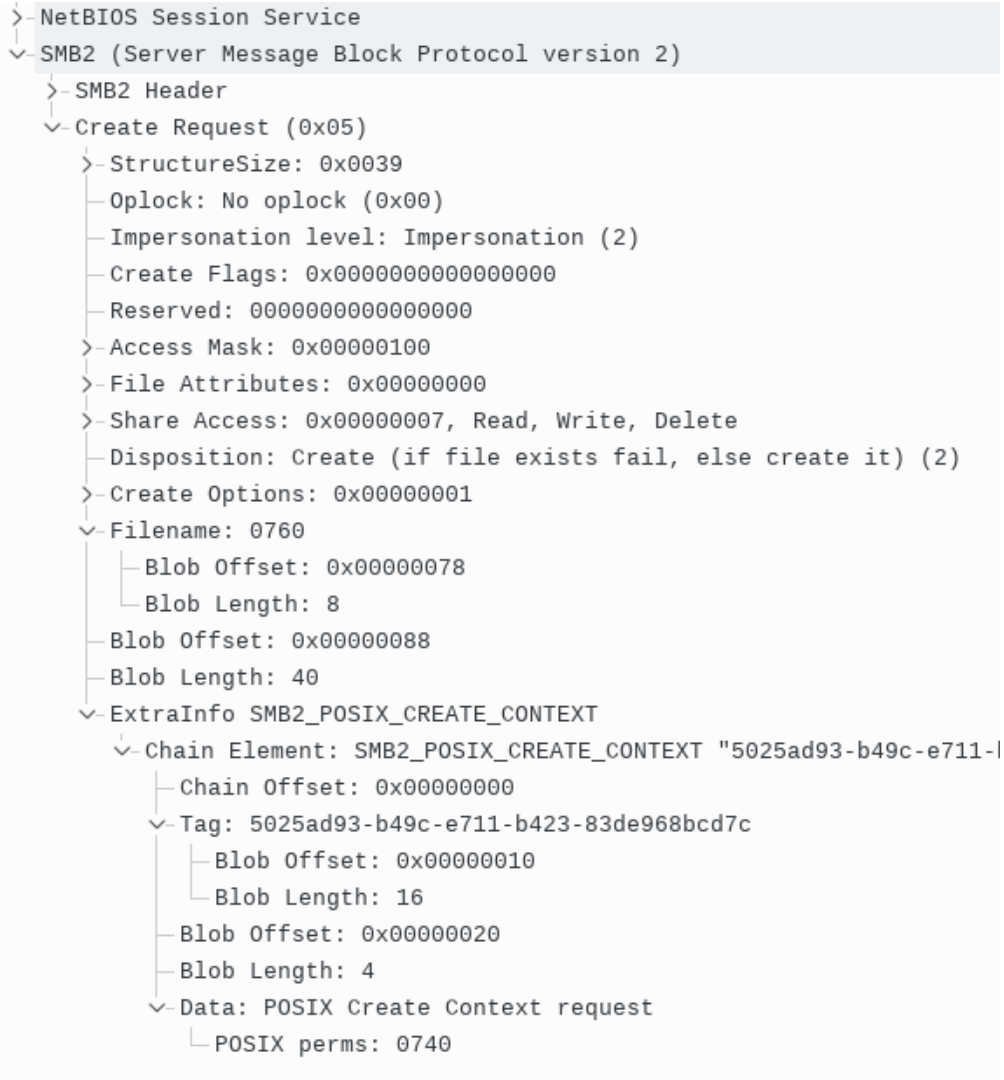
Quick Overview of Status

- Linux kernel client:
  - 5.1 kernel or later can be used. Enable with mount option “posix”
    - But … 5.8 kernel has last remaining piece on client needed to do broad testing: query info (stat) with new POSIX info level.
      - Query fs info (statfs), query info (stat), posix locking, mode bits, ownership, posix semantics on unlink and rename. All major features work
- Samba (experimental tree available, enable with smb.conf parm)
  - Server
    - All major features work. Merge delayed due to time consuming conflicts with other large charges. Special file handling (Sockets, FIFOs, char device handling) needs to be updated
  - Client tools (smbclient)

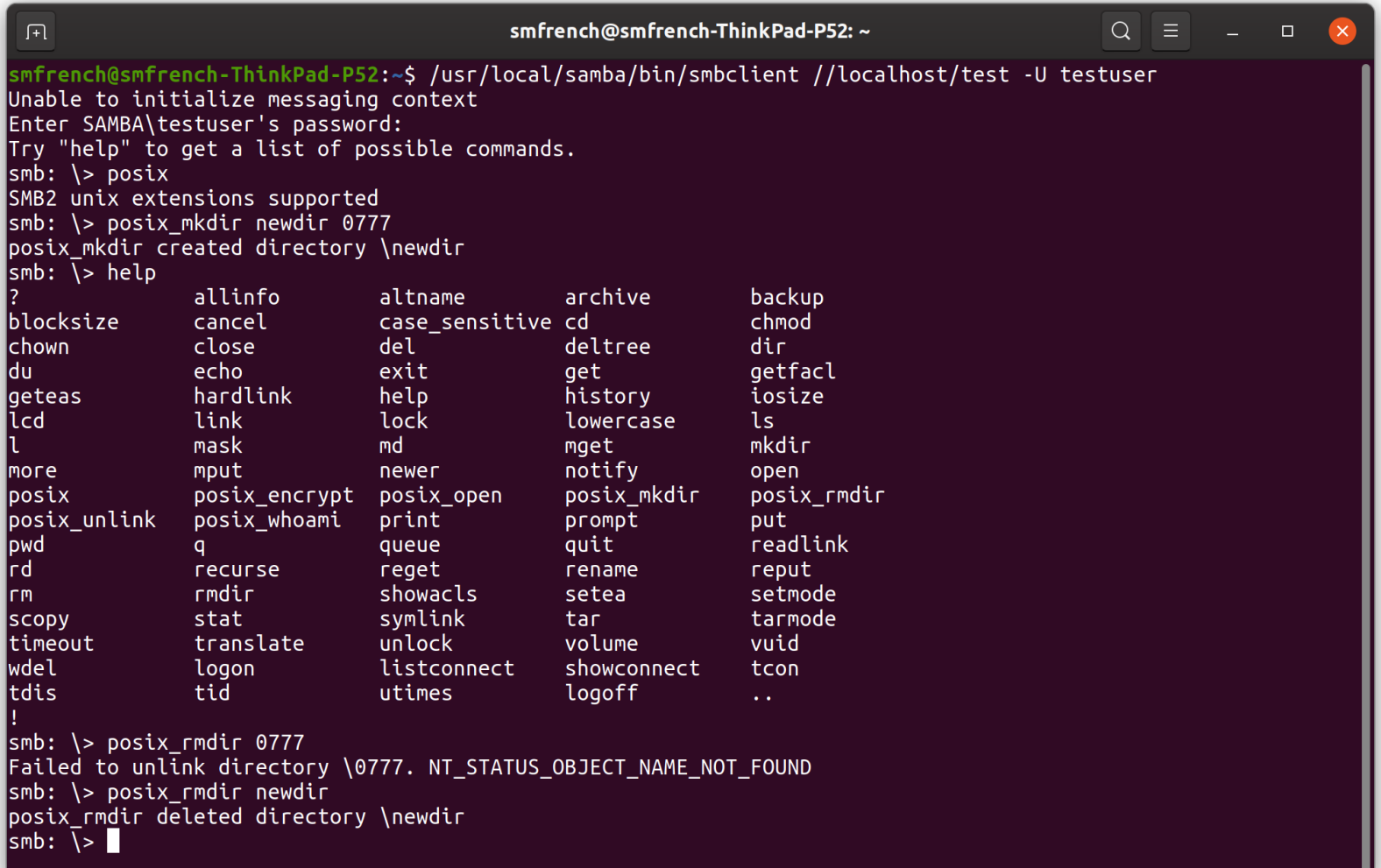

mount-mkdir-posix.pcapng
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
smb2
Expression...
urce Destination Protoc Lengtl Info
.0.0.1 127.0.0.1 SMB2 302 Negotiate Protocol Request
.0.0.1 127.0.0.1 SMB2 366 Negotiate Protocol Response
.0.0.1 127.0.0.1 SMB2 190 Session Setup Request, NTLMSSP_NEGOTIATE
.0.0.1 127.0.0.1 SMB2 400 Session Setup Response, Error: STATUS_MORE_PROCESSING_REQUIRED, NTLMSSP_C
.0.0.1 127.0.0.1 SMB2 460 Session Setup Request, NTLMSSP_AUTH, User: \smfrench
.0.0.1 127.0.0.1 SMB2 142 Session Setup Response
.0.0.1 127.0.0.1 SMB2 176 Tree Connect Request Tree: \\localhost\IPC$
.0.0.1 127.0.0.1 SMB2 150 Tree Connect Response
.0.0.1 127.0.0.1 SMB2 176 Tree Connect Request Tree: \\localhost\test
.0.0.1 127.0.0.1 SMB2 150 Tree Connect Response
.0.0.1 127.0.0.1 SMB2 446 Create Request File: ;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO
.0.0.1 127.0.0.1 SMB2 534 Create Response File: ;GetInfo Response
.0.0.1 127.0.0.1 SMB2 191 Ioctl Request FSCTL_QUERY_NETWORK_INTERFACE_INFO
.0.0.1 127.0.0.1 SMB2 143 Ioctl Response, Error: STATUS_INVALID_DEVICE_REQUEST
.0.0.1 127.0.0.1 SMB2 175 GetInfo Request FS_INFO/FileFsAttributeInformation File:
.0.0.1 127.0.0.1 SMB2 162 GetInfo Response
.0.0.1 127.0.0.1 SMB2 175 GetInfo Request FS_INFO/FileFsDeviceInformation File:
.0.0.1 127.0.0.1 SMB2 150 GetInfo Response
.0.0.1 127.0.0.1 SMB2 175 GetInfo Request FS_INFO/FileFsVolumeInformation File:
.0.0.1 127.0.0.1 SMB2 168 GetInfo Response
.0.0.1 127.0.0.1 SMB2 175 GetInfo Request FS_INFO/FileFsSectorSizeInformation File:
.0.0.1 127.0.0.1 SMB2 170 GetInfo Response
.0.0.1 127.0.0.1 SMB2 158 Close Request File:
.0.0.1 127.0.0.1 SMB2 194 Close Response
.0.0.1 127.0.0.1 SMB2 224 Ioctl Request FSCTL_DFS_GET_REFERRALS, File: \localhost\test
.0.0.1 127.0.0.1 SMB2 143 Ioctl Response, Error: STATUS_NOT_FOUND
.0.0.1 127.0.0.1 SMB2 262 Create Request File:
.0.0.1 127.0.0.1 SMB2 354 Create Response File:
.0.0.1 127.0.0.1 SMB2 158 Close Request File:
.0.0.1 127.0.0.1 SMB2 194 Close Response
.0.0.1 127.0.0.1 SMB2 262 Create Request File:
NetBIOS Session Service
SMB2 (Server Message Block Protocol version 2)
SMB2 Header
Negotiate Protocol Request (0x00)
StructureSize: 0x0024
Dialect count: 4
Security mode: 0x01, Signing enabled
Reserved: 0000
Capabilities: 0x00000077, DFS, LEASING, LARGE MTU, PERSISTENT HA
Client Guid: 032f6ffc-4993-c44d-8b01-425c86949469
NegotiateContextOffset: 0x0070
NegotiateContextCount: 4
Reserved: 0000
Dialect: 0x0210
Dialect: 0x0300
Dialect: 0x0302
Dialect: 0x0311
Negotiate Context: SMB2_PREAUTH_INTEGRITY_CAPABILITIES
Negotiate Context: SMB2_ENCRYPTION_CAPABILITIES
Negotiate Context: Unknown Type: (0x5)
Negotiate Context: SMB2_POSIX_EXTENSIONS_CAPABILITIES
Type: SMB2_POSIX_EXTENSIONS_CAPABILITIES (0x0100)
DataLength: 16
Reserved: 00000000
POSIX Reserved: 0x5025ad93
00b0 11 03 00 00 00 00 01 00 26 00 00 00 00 00 01 00 &
00c0 20 00 01 00 b4 02 72 b9 3c ed 58 84 13 7e b1 3f r <X ~ ?
00d0 4d 7a 75 80 ee 25 a5 e0 81 20 cc 10 57 7e 31 65 Mzu % W~1e
00e0 95 f7 40 b3 00 00 02 00 06 00 00 00 00 00 02 00 @
00f0 02 00 01 00 00 00 05 00 12 00 00 00 00 00 6c 00 l
0100 6f 00 63 00 61 00 6c 00 68 00 6f 00 73 00 74 00 o c a l h o s t
0110 00 00 00 00 00 00 00 01 10 00 00 00 00 00 93 ad
0120 25 50 9c b4 11 e7 b4 23 83 de 96 8b cd 7c %P #
Text item (text), 24 bytes
Packets: 91 · Displayed: 52 (57.1%)
Profile: Default
wireshark po...
keepassdb.k...
wireshark-po...
pres.odp - Li...
Quassel IRC -...
Telegram
Virtual Machi...
mount-mkdir...
12:04 PM


NetBIOS Session Service
SMB2 (Server Message Block Protocol version 2)
SMB2 Header
Negotiate Protocol Request (0x00)
StructureSize: 0x0024
Dialect count: 4
Security mode: 0x01, Signing enabled
Reserved: 0000
Capabilities: 0x00000077, DFS, LEASING, LARGE MTU, PERSISTENT HA
Client Guid: 032f6ffc-4993-c44d-8b01-425c86949469
NegotiateContextOffset: 0x0070
NegotiateContextCount: 4
Reserved: 0000
Dialect: 0x0210
Dialect: 0x0300
Dialect: 0x0302
Dialect: 0x0311
Negotiate Context: SMB2_PREAUTH_INTEGRITY_CAPABILITIES
Negotiate Context: SMB2_ENCRYPTION_CAPABILITIES
Negotiate Context: Unknown Type: (0x5)
Negotiate Context: SMB2_POSIX_EXTENSIONS_CAPABILITIES
Type: SMB2_POSIX_EXTENSIONS_CAPABILITIES (0x0100)
DataLength: 16
Reserved: 00000000
POSIX Reserved: 0x5025ad93


mount-mkdir-posix.pcapng
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
smb2
Expression...
urce Destination Protoc Lengtl Info
.0.0.1 127.0.0.1 SMB2 170 GetInfo Response
.0.0.1 127.0.0.1 SMB2 158 Close Request File:
.0.0.1 127.0.0.1 SMB2 194 Close Response
.0.0.1 127.0.0.1 SMB2 224 Ioctl Request FSCTL_DFS_GET_REFERRALS, File: \localhost\test
.0.0.1 127.0.0.1 SMB2 143 Ioctl Response, Error: STATUS_NOT_FOUND
.0.0.1 127.0.0.1 SMB2 262 Create Request File:
.0.0.1 127.0.0.1 SMB2 354 Create Response File:
.0.0.1 127.0.0.1 SMB2 158 Close Request File:
.0.0.1 127.0.0.1 SMB2 194 Close Response
.0.0.1 127.0.0.1 SMB2 262 Create Request File:
.0.0.1 127.0.0.1 SMB2 354 Create Response File:
.0.0.1 127.0.0.1 SMB2 158 Close Request File:
.0.0.1 127.0.0.1 SMB2 194 Close Response
.0.0.1 127.0.0.1 SMB2 446 Create Request File: ;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO
.0.0.1 127.0.0.1 SMB2 534 Create Response File: ;GetInfo Response
.0.0.1 127.0.0.1 SMB2 174 GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO File:
.0.0.1 127.0.0.1 SMB2 244 GetInfo Response
.0.0.1 127.0.0.1 SMB2 158 Close Request File:
.0.0.1 127.0.0.1 SMB2 194 Close Response
.0.0.1 127.0.0.1 SMB2 462 Create Request File: .Trash;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;
.0.0.1 127.0.0.1 SMB2 310 Create Response, Error: STATUS_OBJECT_NAME_NOT_FOUND;GetInfo Response, Er
.0.0.1 127.0.0.1 SMB2 470 Create Request File: .Trash-1000;GetInfo Request FILE_INFO/SMB2_FILE_ALL_
.0.0.1 127.0.0.1 SMB2 310 Create Response, Error: STATUS_OBJECT_NAME_NOT_FOUND;GetInfo Response, Er
.0.0.1 127.0.0.1 SMB2 462 Create Request File: 0760;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Cl
.0.0.1 127.0.0.1 SMB2 310 Create Response, Error: STATUS_OBJECT_NAME_NOT_FOUND;GetInfo Response, Er
.0.0.1 127.0.0.1 SMB2 246 Create Request File: 0760
.0.0.1 127.0.0.1 SMB2 310 Create Response File: 0760
.0.0.1 127.0.0.1 SMB2 158 Close Request File: 0760
.0.0.1 127.0.0.1 SMB2 194 Close Response
.0.0.1 127.0.0.1 SMB2 462 Create Request File: 0760;GetInfo Request FILE_INFO/SMB2_FILE_ALL_INFO;Cl
.0.0.1 127.0.0.1 SMB2 678 Create Response File: 0760;GetInfo Response;Close Response
NetBIOS Session Service
SMB2 (Server Message Block Protocol version 2)
SMB2 Header
Create Request (0x05)
StructureSize: 0x0039
Oplock: No oplock (0x00)
Impersonation level: Impersonation (2)
Create Flags: 0x0000000000000000
Reserved: 0000000000000000
Access Mask: 0x00000100
File Attributes: 0x00000000
Share Access: 0x00000007, Read, Write, Delete
Disposition: Create (if file exists fail, else create it) (2)
Create Options: 0x00000001
Filename: 0760
Blob Offset: 0x00000078
Blob Length: 8
Blob Offset: 0x00000088
Blob Length: 40
ExtraInfo SMB2_POSIX_CREATE_CONTEXT
Chain Element: SMB2_POSIX_CREATE_CONTEXT "5025ad93-b49c-e711-
Chain Offset: 0x00000000
Tag: 5025ad93-b49c-e711-b423-83de968bcd7c
Blob Offset: 0x00000010
Blob Length: 16
Blob Offset: 0x00000020
Blob Length: 4
Data: POSIX Create Context request
POSIX perms: 0740
0000 00 00 00 00 00 00 00 00 00 00 00 00 08 00 45 00
0010 00 e8 c8 34 40 00 40 06 73 d9 7f 00 00 01 7f 00
0020 00 01 a0 cc 01 bd b5 6f c3 fc 26 6c fe 77 80 18
mount-mkdir-posix.pcapng
Packets: 91 · Displayed: 52 (57.1%)
Profile: Default
wireshark po...
keepassdb.k...
wireshark-po...
pres.odp - Li...
Quassel IRC -...
Telegram
Virtual Machi...
mount-mkdir...
12:05 PM


NetBIOS Session Service
SMB2 (Server Message Block Protocol version 2)
SMB2 Header
Create Request (0x05)
StructureSize: 0x0039
Oplock: No oplock (0x00)
Impersonation level: Impersonation (2)
Create Flags: 0x0000000000000000
Reserved: 0000000000000000
Access Mask: 0x00000100
File Attributes: 0x00000000
Share Access: 0x00000007, Read, Write, Delete
Disposition: Create (if file exists fail, else create it) (2)
Create Options: 0x00000001
Filename: 0760
Blob Offset: 0x00000078
Blob Length: 8
Blob Offset: 0x00000088
Blob Length: 40
ExtraInfo SMB2_POSIX_CREATE_CONTEXT
Chain Element: SMB2_POSIX_CREATE_CONTEXT "5025ad93-b49c-e711-
Chain Offset: 0x00000000
Tag: 5025ad93-b49c-e711-b423-83de968bcd7c
Blob Offset: 0x00000010
Blob Length: 16
Blob Offset: 0x00000020
Blob Length: 4
Data: POSIX Create Context request
POSIX perms: 0740

Smbclient now has experimental support for SMB3.1.1 POSIX Extensions



Additional Examples of what works today from Linux kernel client

Why SMB3 for Linux?

- SMB3.1.1 (and related protocols) is the richest, most functional file protocol
- There are many Linux file systems (>60), but six (and the VFS layer itself) drive 75% of activity (btrfs, xfs, nfs and cifs are the most active). Kernel development is hard … reuse helpful
- cifs.ko (cifs/smb3 client) activity is strong
- The family of related protocols (including SMB3.1.1) has the most exhaustive set of documentation, test cases, implementations …

Why Not Other Protocols?

- SMB3.1.1 is easily extensible
- SMB3.1.1 works tightly with a set of protocols which can do more than any other file system protocol
- SMB3.1.1 has the best, most exhaustive set of testcases (not just smbtorture …)
- SMB3.1.1 and related protocols have more documentation (and documentation that has been tested and verified)
- SMB3.1.1 is proven across multiple client types, OS, architectures

(And don't forget, SAMBA rocks!)

- What about the Linux Kernel?
  - New API changes added about once a year to the VFS (minor global changes added more often, but not all could affect what we need to send on the wire in perfect world ...)
  - Need to quickly update protocol when not possible to do over SMB3
  - Need better interaction with key

What Next?

- Examine the xfstest skips (and failures) in much detail and add small incremental changes
  - “xfstests” is the standard Linux fs functional test suite and no one file system can pass all tests due to various fs optional features.
  - Some can be emulated some need new flags
- Where that is not possible, consider adding new POSIX extensions version (simply adding additional uuid to the POSIX negotiate context)

Examples from xfstest investigations

- Add support for renameat2 and rename exchange
- POSIX ACLs (can be emulated and there is pushback on implementing primitive POSIX ACLs)
- Support for additional chattr flags (“immutable” and “noatime” updates e.g.)
- fallocate –collapse-range
- Dedupe support
- Defragmentation support (may require VFS changes)

Examples from xfstest investigations

- Richacl support (tests 362 through 370) ??
- O_TMPFILE support (emulatable, but VFS changes would help)
- FITRIM support (may be emulatable)
- Quota support (may be emulatable already)
- Support for NFS export (nfs server on smb3 mounts)
- Case sensitive xattrs (EAs)
- SELinux support

Examples from xfstest investigations

- Support for online ‘label manipulation’ (see e.g. xfstest generic/492)
- Support for casefolding (“chattr +F”)

Quality Much Improved – Top Priority

- More xfstests pass (> 150 and growing) even without POSIX extensions, vast majority of the rest are skipped due to missing features or being inappropriate for network file systems
- Many potential issues pointed out by static analysis addressed
- Starting two years ago **The "Buildbot"** … reducing regressions. **VERY** exciting addition for CIT (thanks Ronnie, Aurelien and Paulo)
- POSIX Extensions (jra's tree) now a buildbot target for automated regression tests. Will continue to expanding test list...

Wireshark

- See Aurelien's dissector improvements
  - https://github.com/aaptel/wireshark/commits/smb3unix
  - And Pike sample test code
    - https://github.com/aaptel/pike/tree/smb3unix

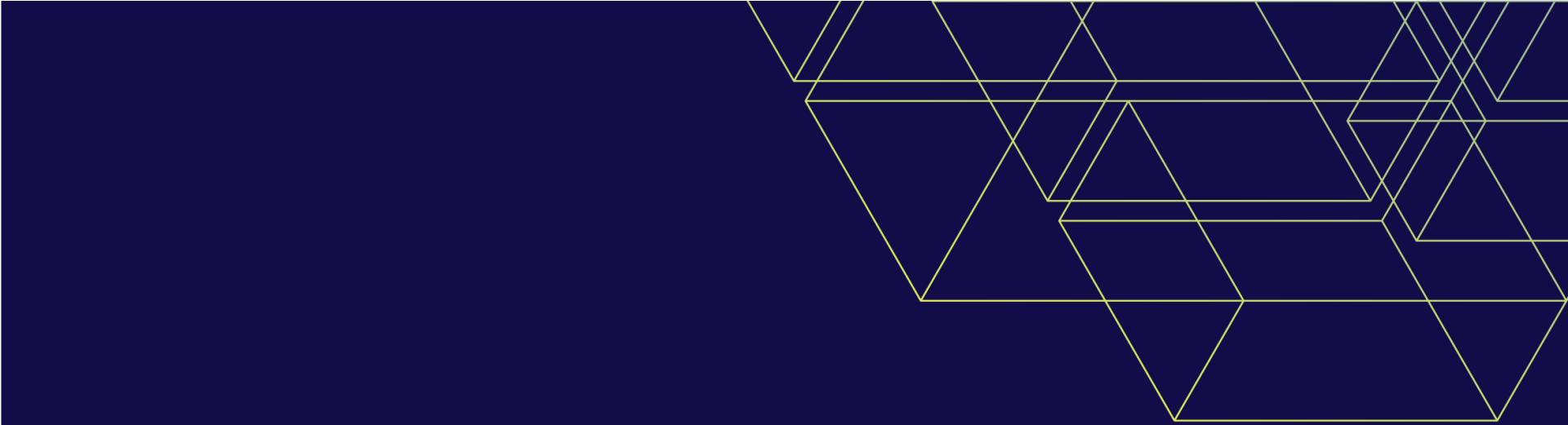
Next Steps

- Continue debugging test implementations (cifs.ko and JRAs Samba POSIX test branch). Current focus: enhancing smb3 client to better handle POSIX stat (getattr)
- Continue to add xfstests to the ‘jraposix’ test group in the buildbot (to regression test the client against Samba server with POSIX extensions)
- Extend ksmd support for POSIX extensions

Thank you for your time

- This is a very exciting time for ...

+ SMB 3



Please take a moment to rate this session.

Your feedback matters to us.