SambaXP 2022

Accessing files from the smallest devices to the largest (and the cloud): Improvements to SMB3.1.1 and Linux

Presented by Steve French

Principal Software Engineer

Microsoft Azure Storage

- 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 ?

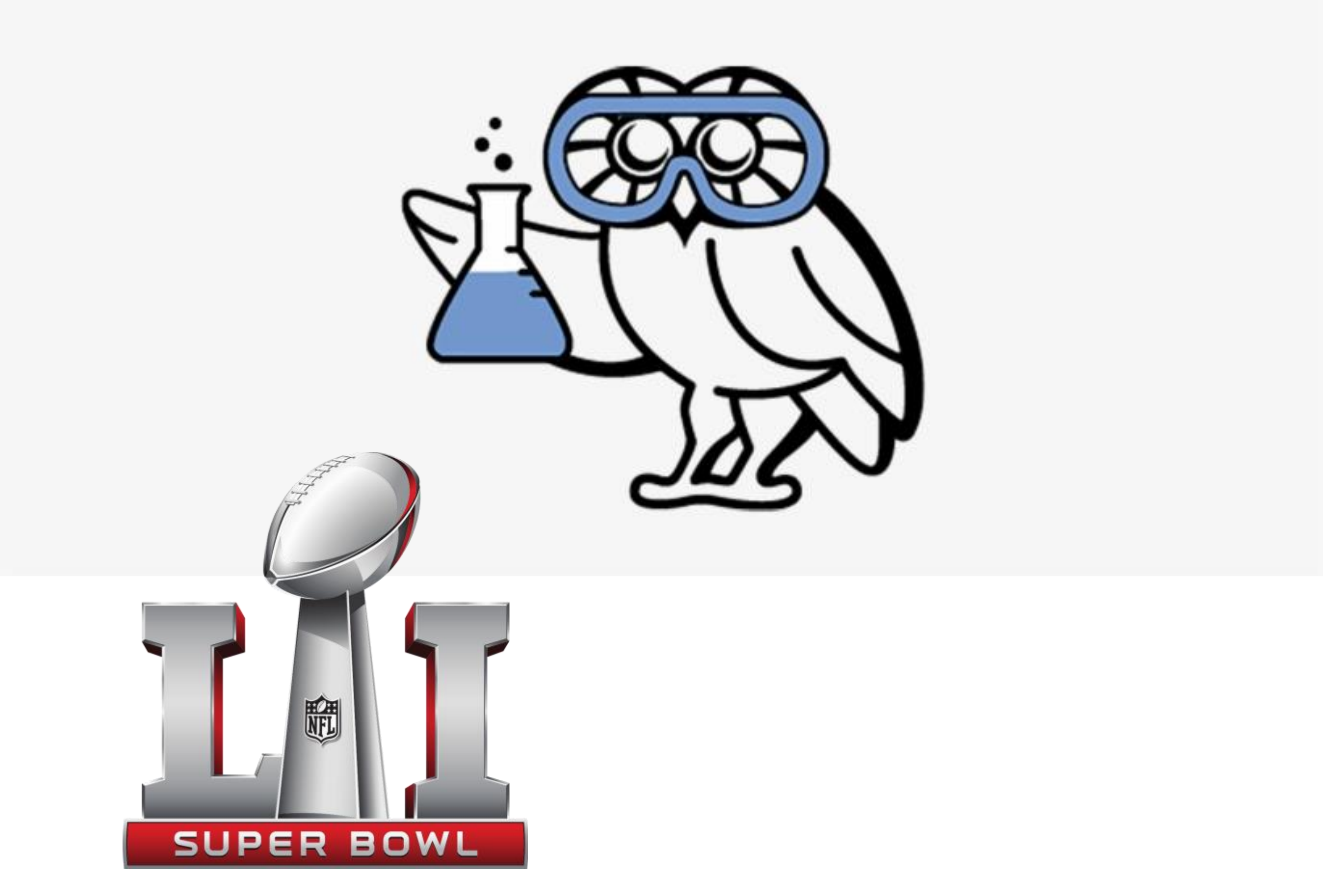
- Steve French smfrench@gmail.com
- Author and maintainer of Linux cifs vfs (for accessing Samba, Azure, Windows and various SMB3/CIFS based NAS appliances)
- Co-maintainer of the new kernel server (ksmbd)
- Also wrote initial SMB2 kernel client prototype
- Member of the Samba team
- coauthor of SNIA CIFS Technical Reference, former SNIA CIFS Working Group chair
- Principal Software Engineer, Azure Storage: Microsof

Outline

- Overview of Linux FS activity
- New Kernel server
- Recent client improvements
  - Kernel
  - Utilities
- Coming soon … what to look forward to
- Testing improvements

A year ago and now ...

- Now: 5.19-rc1 “Superb Owl”
- Then: 5.13-rc2 “Frozen Wasteland”



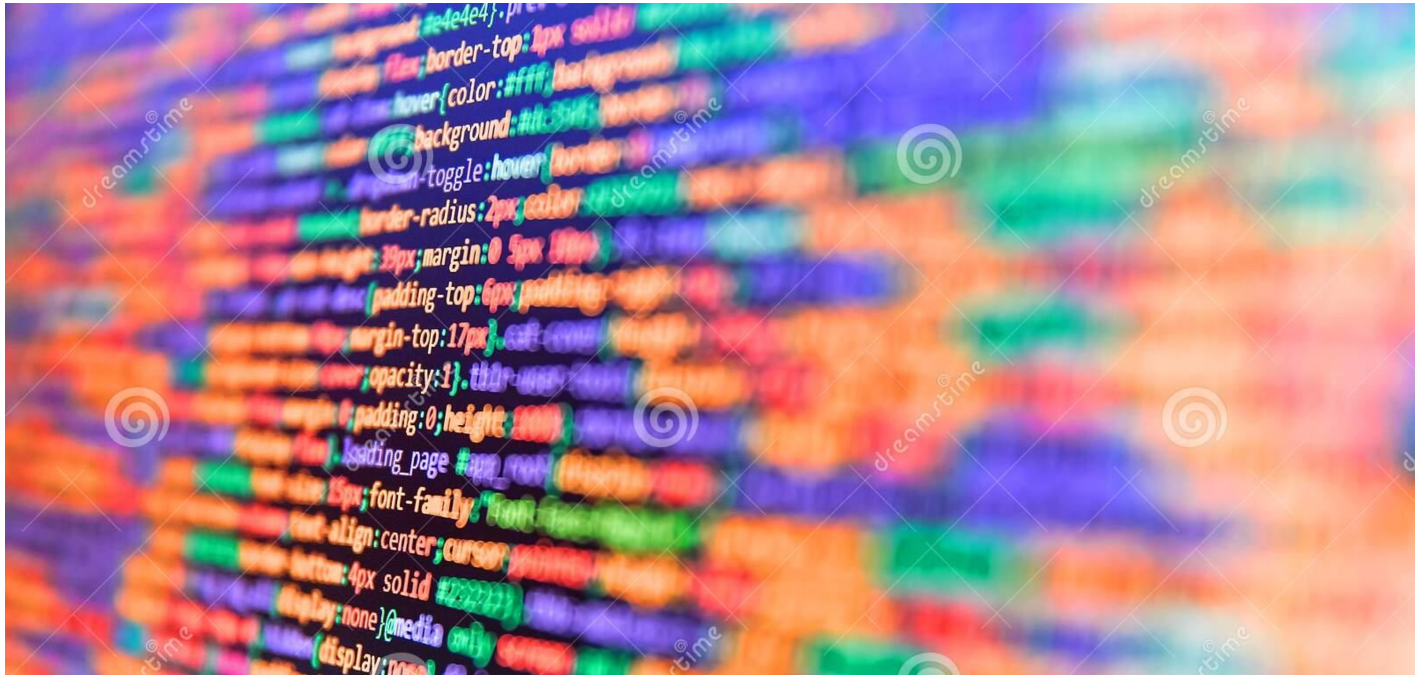
LSF/MM/eBPF summit recently concluded

Some Linux FS topics of interest from LSF and other recent discussions

- Folios, netfs and the redesign of page cache and offline (fscache)
- Improvements to statx and fsinfo and to inotify/fanotify
- Idmapped mounts
- Extending in kernel encryption: TLS handshake (for NFS) and QUIC (SMB3.1.1 and other)
- io_uring (async i/o improvements)
- Shift to cloud
- Better support for faster storage (NVME) and net (RDMA/smbdirect)

Linux Filesystems Activity over past year (since 5.13-rc2)

- 5320 filesystems changesets (6.6% of total kernel changesets, one of the most watched parts of the kernel, and FS Activity up slightly as percentage of kernel activity)
- Linux kernel fs are 1.07 million lines of code (measured this week)



Most Active Linux Filesystems over the past year

- VFS (mapping layer) 1022 changesets
- The top filesystems and VFS dominate the activity
- Most active are BTRFS 927 (almost 60KLOC changed!), XFS 519
- Then NFS (283) and SMB3.1.1 (cifs.ko) (273 changesets) clients
  - cifs.ko had almost twice as many lines changed. It has been a VERY active year for cifs.ko
- Other:
  - Ext4 (263 changesets), ksmbd (new, added in the 5.15 kernel) (189), nfsd (177), ceph (139), gfs2 (133), ntfs3 (new, added in 5.15) (123)

SMB3.1.1 Activity was strong this year

- cifs.ko activity was strong, 273 changesets
  - cifs is now 60KLOC kernel code (not counting user space utilities), 2% larger
- ksmbd activity was also strong
  - Introduced in the 5.15 kernel, 25KLOC kernel code, 189 changesets since its introduction
- Samba server (userspace) is over 3.5 million lines of code (orders of magnitude bigger than the kernel smbd server or any of the NFS servers) and is even more active

One of the strengths of SMB3.1.1 is broad interop testing

- Hopefully in-person plugfests resume soon – but a virtual SMB3.1.1 plugfest will be held this week

What are the goals?

- Repeating an older slide about goals of SMB3.1.1:
- 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 needs new features, quickly add to Linux kernel client and Samba and ksmbd

Progress and Status update for Linux Kernel Server (ksmbd)

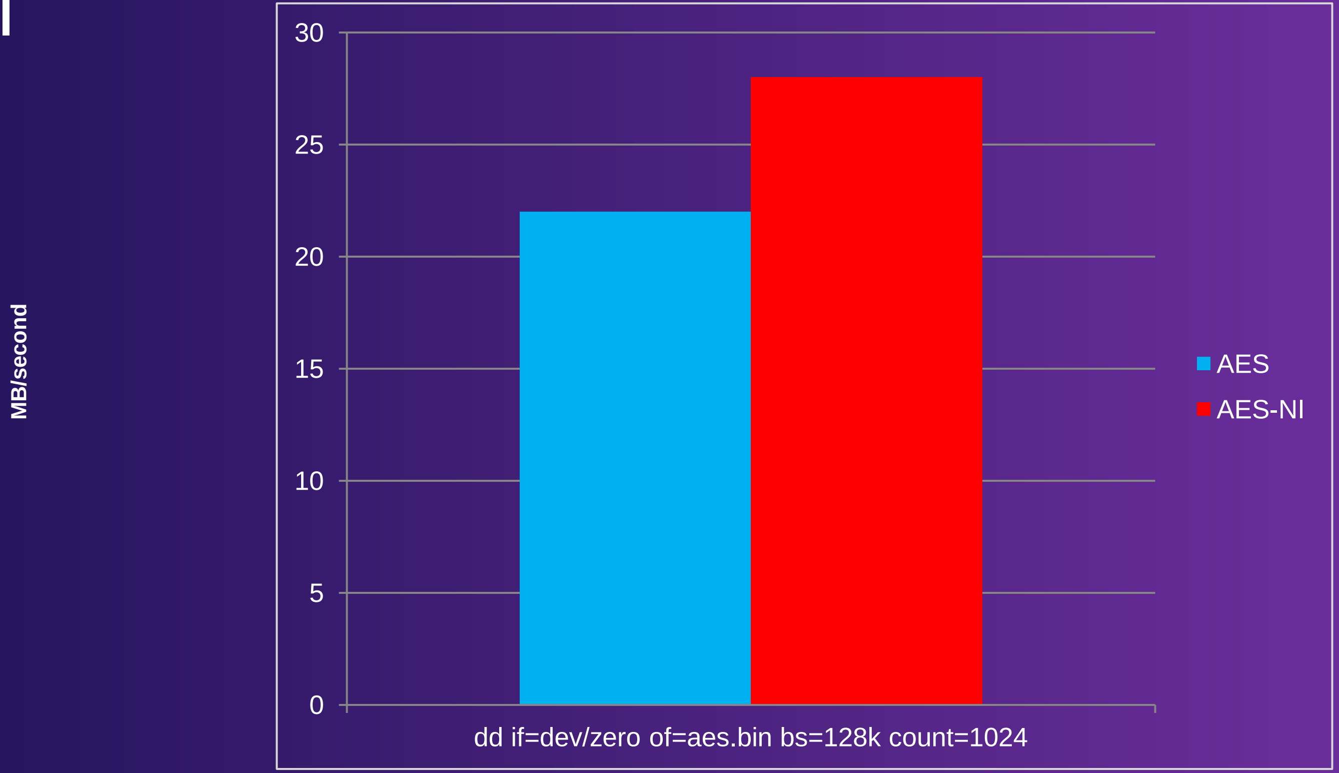
Provided by Namjae Jeon (linkinjeon@kernel.org)

ksmbd merged into mainline Linux in the fall (5.15)

- Getting reviews for 5 months since ksmbd v1 patch series went in Linux-next
- Many high profile developers reviewed, Thank you!
  - Multiple security issues were recently identified, and these fixes are the focus now
- Ksmbd is merged into linux-5.15-rc1 merge window
- To make module and directory name consistent: changed “cifsd” to “ksmbd”
- Later the cifs source directory will be renamed to smbfs_client to reduce confusion (and to avoid referencing old, deprecated, less secure protocol dialect ‘cifs.’ Modern clients and servers negotiate SMB3 or later, not old cifs)
  - Common code between client and server is now in “fs/smbfs_common” directory

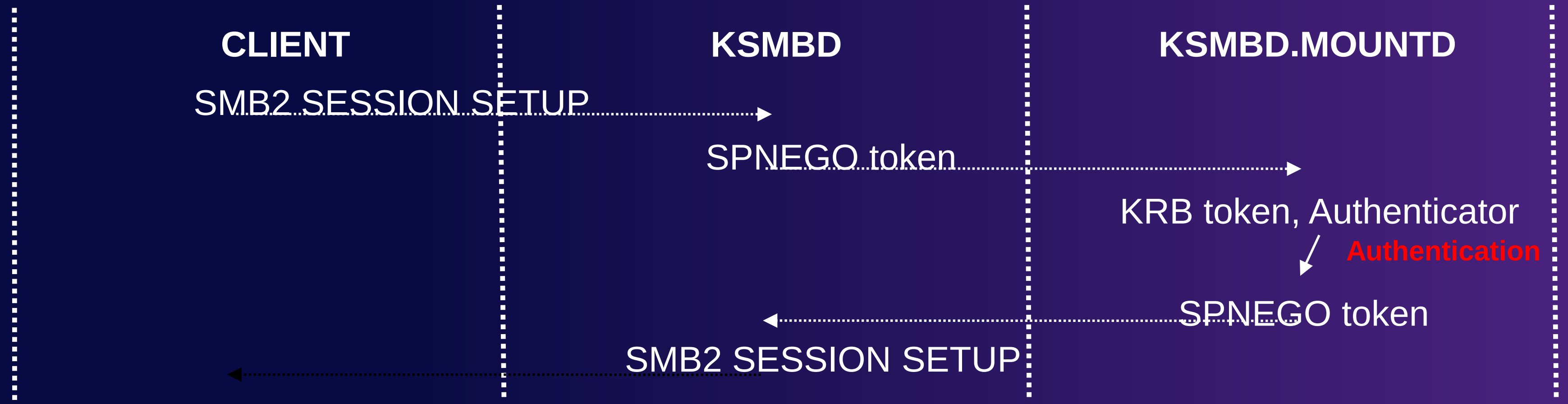
AES-256 encryption support

- ksmbd AES-256 CCM/GCM encryption support now available (strongest encryption)
- Ksmbd accelerated encryption(AES-GCM) performance using AES-NI support in kernel



Kerberos support

- Support authentication with Kerberos
- Ksmbd transmit Kerberos msg to ksmbd.mountd
- Ksmbd.mountd uses libkrb5 library

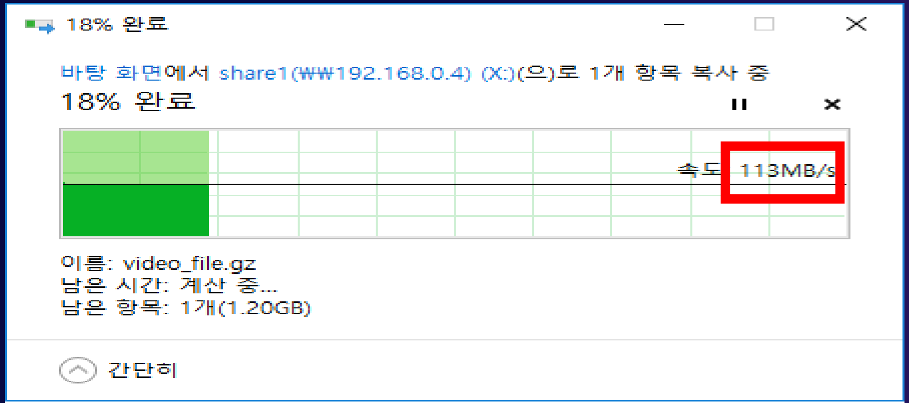


Duplicate extent support

- Ksmbd add support for FSCTL_DUPLICATE_EXTENT_TO_FILE
- This command can be used if share is in reflink support local fs (Linux client uses it for some fallocate related operations like insert range)
- Additional xfstests tests pass.
- Ksmbd doesn’t have to deal with VFS mapping(btrfs, etc.) layer like samba.

SMB3 multi-channel support

- SMB3 Multichannel feature greatly improves performance on Multi-port NIC or multiple NICs.
- Ksmbd kernel server started to support SMB3 multichannel.
- TODO Replay/retry features on channel failure.



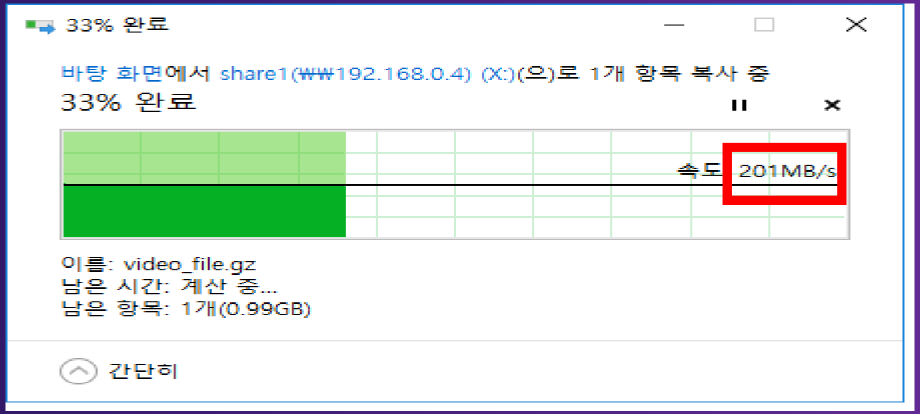
18% 완료

바탕 화면에서 share1(₩₩192.168.0.4) (X:)(으)로 1개 항목 복사 중

18% 완료

속도 113MB/s

이름: video_file.gz
남은 시간: 계산 중...
남은 항목: 1개(1.20GB)



33% 완료

바탕 화면에서 share1(₩₩192.168.0.4) (X:)(으)로 1개 항목 복사 중

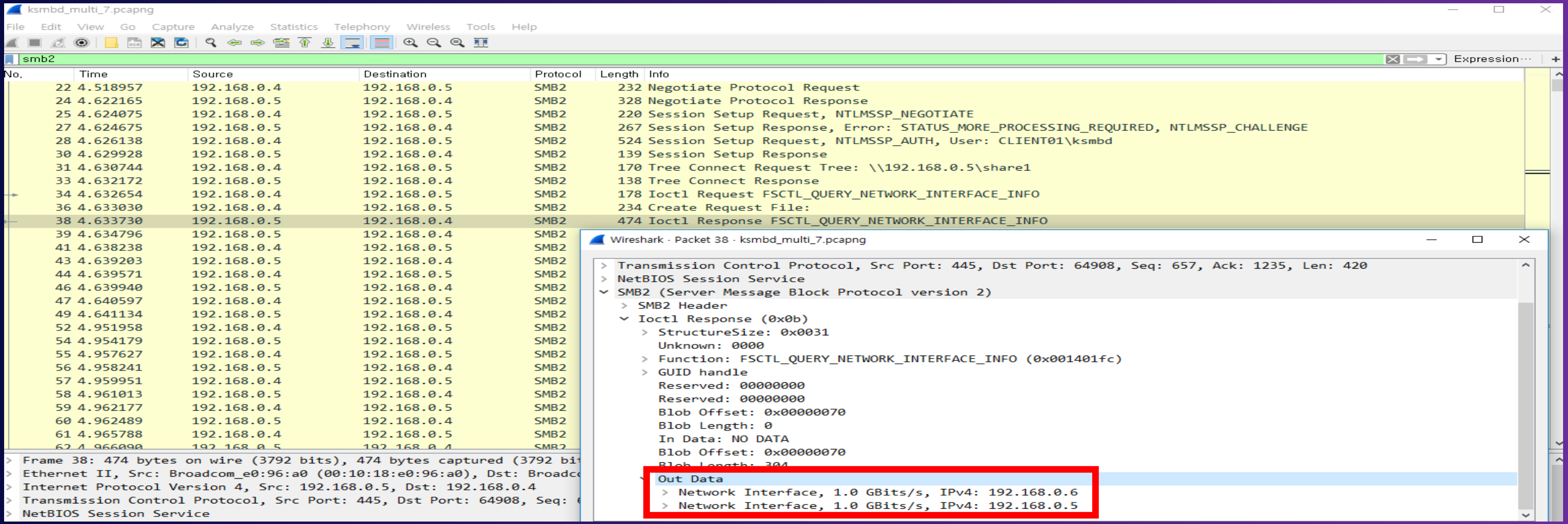
33% 완료

속도 201MB/s

이름: video_file.gz
남은 시간: 계산 중...
남은 항목: 1개(0.99GB)

SMB3 multi-channel support

- Send NICs information to client through FSCTL_QUERY_NETWORK_INTERFACE_INFO command



smb2

```
No.    Time        Source          Destination     Protocol  Length Info
    22 4.518957    192.168.0.4     192.168.0.5     SMB2         232 Negotiate Protocol Request
    24 4.622165    192.168.0.5     192.168.0.4     SMB2         328 Negotiate Protocol Response
    25 4.624075    192.168.0.4     192.168.0.5     SMB2         220 Session Setup Request, NTLMSSP_NEGOTIATE
    27 4.624675    192.168.0.5     192.168.0.4     SMB2         267 Session Setup Response, Error: STATUS_MORE_PROCESSING_REQUIRED, NTLMSSP_CHALLENGE
    28 4.626138    192.168.0.4     192.168.0.5     SMB2         524 Session Setup Request, NTLMSSP_AUTH, User: CLIENT01\ksmbd
    30 4.629928    192.168.0.5     192.168.0.4     SMB2         139 Session Setup Response
    31 4.630744    192.168.0.4     192.168.0.5     SMB2         170 Tree Connect Request Tree: \\192.168.0.5\share1
    33 4.632172    192.168.0.5     192.168.0.4     SMB2         138 Tree Connect Response
    34 4.632654    192.168.0.4     192.168.0.5     SMB2         178 Ioctl Request FSCTL_QUERY_NETWORK_INTERFACE_INFO
    36 4.633030    192.168.0.4     192.168.0.5     SMB2         234 Create Request File:
    38 4.633730    192.168.0.5     192.168.0.4     SMB2         474 Ioctl Response FSCTL_QUERY_NETWORK_INTERFACE_INFO
    39 4.634796    192.168.0.5     192.168.0.4     SMB2
    41 4.638238    192.168.0.4     192.168.0.5     SMB2
    43 4.639203    192.168.0.5     192.168.0.4     SMB2
    44 4.639571    192.168.0.4     192.168.0.5     SMB2
    46 4.639940    192.168.0.5     192.168.0.4     SMB2
    47 4.640597    192.168.0.4     192.168.0.5     SMB2
    49 4.641134    192.168.0.5     192.168.0.4     SMB2
    52 4.951958    192.168.0.4     192.168.0.5     SMB2
    54 4.954179    192.168.0.5     192.168.0.4     SMB2
    55 4.957627    192.168.0.4     192.168.0.5     SMB2
    56 4.958241    192.168.0.5     192.168.0.4     SMB2
    57 4.959951    192.168.0.4     192.168.0.5     SMB2
    58 4.961013    192.168.0.5     192.168.0.4     SMB2
    59 4.962177    192.168.0.4     192.168.0.5     SMB2
    60 4.962489    192.168.0.5     192.168.0.4     SMB2
    61 4.965788    192.168.0.4     192.168.0.5     SMB2
    62 4.966090    192.168.0.5     192.168.0.4     SMB2
```

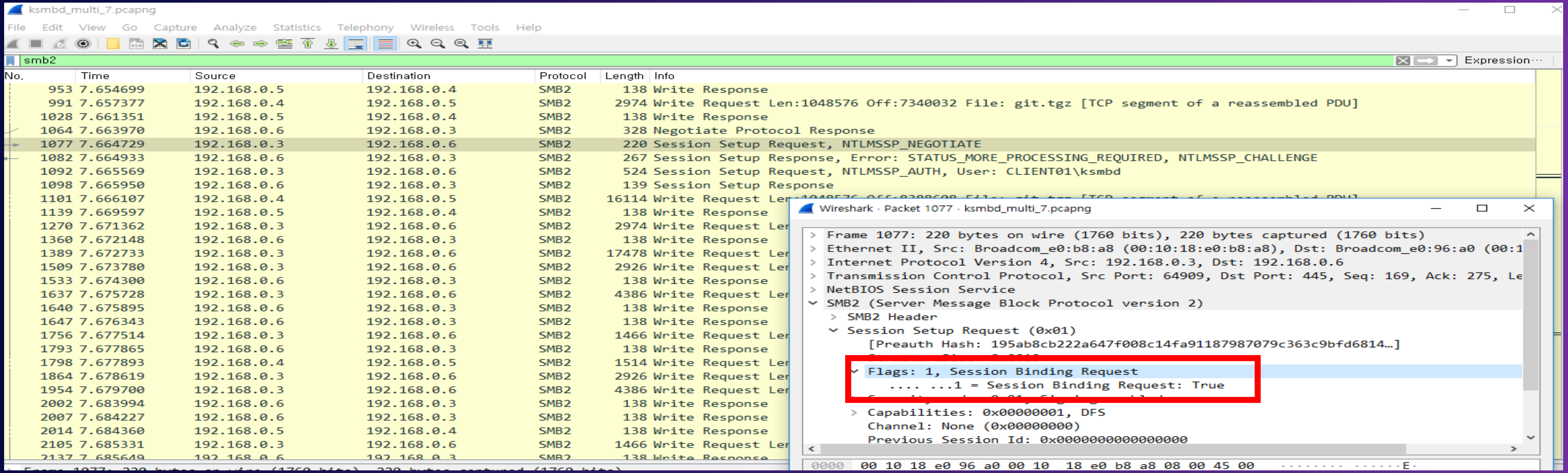
```
> Frame 38: 474 bytes on wire (3792 bits), 474 bytes captured (3792 bi
> Ethernet II, Src: Broadcom_e0:96:a0 (00:10:18:e0:96:a0), Dst: Broadc
> Internet Protocol Version 4, Src: 192.168.0.5, Dst: 192.168.0.4
> Transmission Control Protocol, Src Port: 445, Dst Port: 64908, Seq: 
> NetBIOS Session Service
```

Wireshark · Packet 38 · ksmbd_multi_7.pcapng

```
> Transmission Control Protocol, Src Port: 445, Dst Port: 64908, Seq: 657, Ack: 1235, Len: 420
> NetBIOS Session Service
v SMB2 (Server Message Block Protocol version 2)
  > SMB2 Header
  v Ioctl Response (0x0b)
    > StructureSize: 0x0031
      Unknown: 0000
    > Function: FSCTL_QUERY_NETWORK_INTERFACE_INFO (0x001401fc)
    > GUID handle
      Reserved: 00000000
      Reserved: 00000000
      Blob Offset: 0x00000070
      Blob Length: 0
      In Data: NO DATA
      Blob Offset: 0x00000070
      Blob Length: 304
    v Out Data
      > Network Interface, 1.0 GBits/s, IPv4: 192.168.0.6
      > Network Interface, 1.0 GBits/s, IPv4: 192.168.0.5
```

SMB3 multi-channel support

- Client send session binding request to ksmbd.



ksmbd_multi_7.pcapng
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
Expression…

Filter: smb2

```
No.   Time      Source       Destination  Protocol Length Info
  953 7.654699  192.168.0.5  192.168.0.4  SMB2       138 Write Response
  991 7.657377  192.168.0.4  192.168.0.5  SMB2      2974 Write Request Len:1048576 Off:7340032 File: git.tgz [TCP segment of a reassembled PDU]
 1028 7.661351  192.168.0.5  192.168.0.4  SMB2       138 Write Response
 1064 7.663970  192.168.0.6  192.168.0.3  SMB2       328 Negotiate Protocol Response
 1077 7.664729  192.168.0.3  192.168.0.6  SMB2       220 Session Setup Request, NTLMSSP_NEGOTIATE
 1082 7.664933  192.168.0.6  192.168.0.3  SMB2       267 Session Setup Response, Error: STATUS_MORE_PROCESSING_REQUIRED, NTLMSSP_CHALLENGE
 1092 7.665569  192.168.0.3  192.168.0.6  SMB2       524 Session Setup Request, NTLMSSP_AUTH, User: CLIENT01\ksmbd
 1098 7.665950  192.168.0.6  192.168.0.3  SMB2       139 Session Setup Response
 1101 7.666107  192.168.0.4  192.168.0.5  SMB2     16114 Write Request Len:1048576 Off:8388608 File: git.tgz [TCP segment of a reassembled PDU]
 1139 7.669597  192.168.0.5  192.168.0.4  SMB2       138 Write Response
 1270 7.671362  192.168.0.3  192.168.0.6  SMB2      2974 Write Request Len
 1360 7.672148  192.168.0.6  192.168.0.3  SMB2       138 Write Response
 1389 7.672733  192.168.0.3  192.168.0.6  SMB2     17478 Write Request Len
 1509 7.673780  192.168.0.3  192.168.0.6  SMB2      2926 Write Request Len
 1533 7.674300  192.168.0.6  192.168.0.3  SMB2       138 Write Response
 1637 7.675728  192.168.0.3  192.168.0.6  SMB2      4386 Write Request Len
 1640 7.675895  192.168.0.6  192.168.0.3  SMB2       138 Write Response
 1647 7.676343  192.168.0.6  192.168.0.3  SMB2       138 Write Response
 1756 7.677514  192.168.0.3  192.168.0.6  SMB2      1466 Write Request Len
 1793 7.677865  192.168.0.6  192.168.0.3  SMB2       138 Write Response
 1798 7.677893  192.168.0.4  192.168.0.5  SMB2      1514 Write Request Len
 1864 7.678619  192.168.0.3  192.168.0.6  SMB2      2926 Write Request Len
 1954 7.679700  192.168.0.3  192.168.0.6  SMB2      4386 Write Request Len
 2002 7.683994  192.168.0.6  192.168.0.3  SMB2       138 Write Response
 2007 7.684227  192.168.0.6  192.168.0.3  SMB2       138 Write Response
 2014 7.684360  192.168.0.5  192.168.0.4  SMB2       138 Write Response
 2105 7.685331  192.168.0.3  192.168.0.6  SMB2      1466 Write Request Len
 2137 7.685649  192.168.0.6  192.168.0.3  SMB2       138 Write Response
```

Frame 1077: 220 bytes on wire (1760 bits), 220 bytes captured (1760 bits)

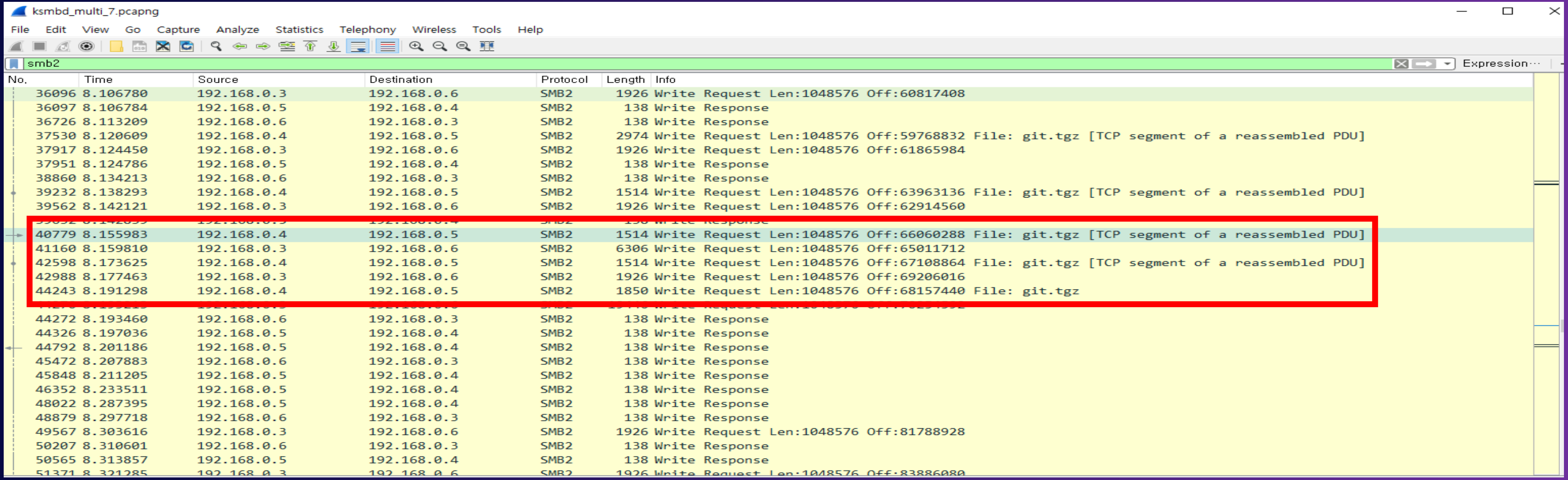
Wireshark · Packet 1077 · ksmbd_multi_7.pcapng

```
> Frame 1077: 220 bytes on wire (1760 bits), 220 bytes captured (1760 bits)
> Ethernet II, Src: Broadcom_e0:b8:a8 (00:10:18:e0:b8:a8), Dst: Broadcom_e0:96:a0 (00:1
> Internet Protocol Version 4, Src: 192.168.0.3, Dst: 192.168.0.6
> Transmission Control Protocol, Src Port: 64909, Dst Port: 445, Seq: 169, Ack: 275, Le
> NetBIOS Session Service
v SMB2 (Server Message Block Protocol version 2)
  > SMB2 Header
  v Session Setup Request (0x01)
      [Preauth Hash: 195ab8cb222a647f008c14fa91187987079c363c9bfd6814…]
    v Flags: 1, Session Binding Request
        .... ...1 = Session Binding Request: True
    > Capabilities: 0x00000001, DFS
      Channel: None (0x00000000)
      Previous Session Id: 0x0000000000000000

0000  00 10 18 e0 96 a0 00 10  18 e0 b8 a8 08 00 45 00   ........ ......E.
```

SMB3 multi-channel support

- Client send interleaved write requests to dual channels(192.168.0.3, 192.168.0.4)



```
ksmbd_multi_7.pcapng

smb2

No.   Time      Source       Destination  Protocol Length Info
36096 8.106780  192.168.0.3  192.168.0.6  SMB2     1926 Write Request Len:1048576 Off:60817408
36097 8.106784  192.168.0.5  192.168.0.4  SMB2      138 Write Response
36726 8.113209  192.168.0.6  192.168.0.3  SMB2      138 Write Response
37530 8.120609  192.168.0.4  192.168.0.5  SMB2     2974 Write Request Len:1048576 Off:59768832 File: git.tgz [TCP segment of a reassembled PDU]
37917 8.124450  192.168.0.3  192.168.0.6  SMB2     1926 Write Request Len:1048576 Off:61865984
37951 8.124786  192.168.0.5  192.168.0.4  SMB2      138 Write Response
38860 8.134213  192.168.0.6  192.168.0.3  SMB2      138 Write Response
39232 8.138293  192.168.0.4  192.168.0.5  SMB2     1514 Write Request Len:1048576 Off:63963136 File: git.tgz [TCP segment of a reassembled PDU]
39562 8.142121  192.168.0.3  192.168.0.6  SMB2     1926 Write Request Len:1048576 Off:62914560
39592 8.142839  192.168.0.5  192.168.0.4  SMB2      138 Write Response
40779 8.155983  192.168.0.4  192.168.0.5  SMB2     1514 Write Request Len:1048576 Off:66060288 File: git.tgz [TCP segment of a reassembled PDU]
41160 8.159810  192.168.0.3  192.168.0.6  SMB2     6306 Write Request Len:1048576 Off:65011712
42598 8.173625  192.168.0.4  192.168.0.5  SMB2     1514 Write Request Len:1048576 Off:67108864 File: git.tgz [TCP segment of a reassembled PDU]
42988 8.177463  192.168.0.3  192.168.0.6  SMB2     1926 Write Request Len:1048576 Off:69206016
44243 8.191298  192.168.0.4  192.168.0.5  SMB2     1850 Write Request Len:1048576 Off:68157440 File: git.tgz
44270 8.193219  192.168.0.3  192.168.0.6  SMB2    19446 Write Request Len:1048576 Off:70254592
44272 8.193460  192.168.0.6  192.168.0.3  SMB2      138 Write Response
44326 8.197036  192.168.0.5  192.168.0.4  SMB2      138 Write Response
44792 8.201186  192.168.0.5  192.168.0.4  SMB2      138 Write Response
45472 8.207883  192.168.0.6  192.168.0.3  SMB2      138 Write Response
45848 8.211205  192.168.0.5  192.168.0.4  SMB2      138 Write Response
46352 8.233511  192.168.0.5  192.168.0.4  SMB2      138 Write Response
48022 8.287395  192.168.0.5  192.168.0.4  SMB2      138 Write Response
48879 8.297718  192.168.0.6  192.168.0.3  SMB2      138 Write Response
49567 8.303616  192.168.0.3  192.168.0.6  SMB2     1926 Write Request Len:1048576 Off:81788928
50207 8.310601  192.168.0.6  192.168.0.3  SMB2      138 Write Response
50565 8.313857  192.168.0.5  192.168.0.4  SMB2      138 Write Response
51371 8.321285  192.168.0.3  192.168.0.6  SMB2     1926 Write Request Len:1048576 Off:83886080
```

Currently working features

- SMB Direct with windows client
  - Got test HW support from Chelsio (Bob Dugan)
  - Windows client connection success
  - Checking performance issue
  - Add interface to change SMBD parameter as per RDMA NICs
  - Credit management rework
- SMB2 directory leases
- SMB2 change notify
  - Considering using fanotify instead of inotify for SMB2_WATCH_TREE
  - Need to change fanotify codes as export symbol to call function by ksmbd.

ksmbd start to fully support smbdirect

- Handle large RDMA read/write size(bulk data) supported by SMB Direct multi-descriptors.(It supported single descriptor with 512KB size before)
  - 8MB RDMA read/write size by default.
  - Control read/write size through ksmbd configuration.(e.g. smbd io size = 16MB)
- Improve the compatibility with various RDMA types of NICs
  - Tested smb-direct working with iWARP(Chelsio, soft-iWARP), Infiniband(Mellanox NICs, Connectx3 ~ x5), ROCE(soft-ROCE).
- Auto-detection of RDMA NICs without configuration.
  - Server should send RDMA NIC info to client.
  - No need to specify RDMA NIC information to smb.conf.

Performance test environment

- Benchmark tool
  - Framtest (https://support.dvsus.com/hc/en-us/articles/212925466-How-to-use-frametest)
- Server

CPU: intel Silver 4114 x 2
DRAM: 512GB
NVMe SSD: Kioxia CM6 1.9T x 9(mdadm raid0 with XFS)
NIC: MCX516A-CCAT

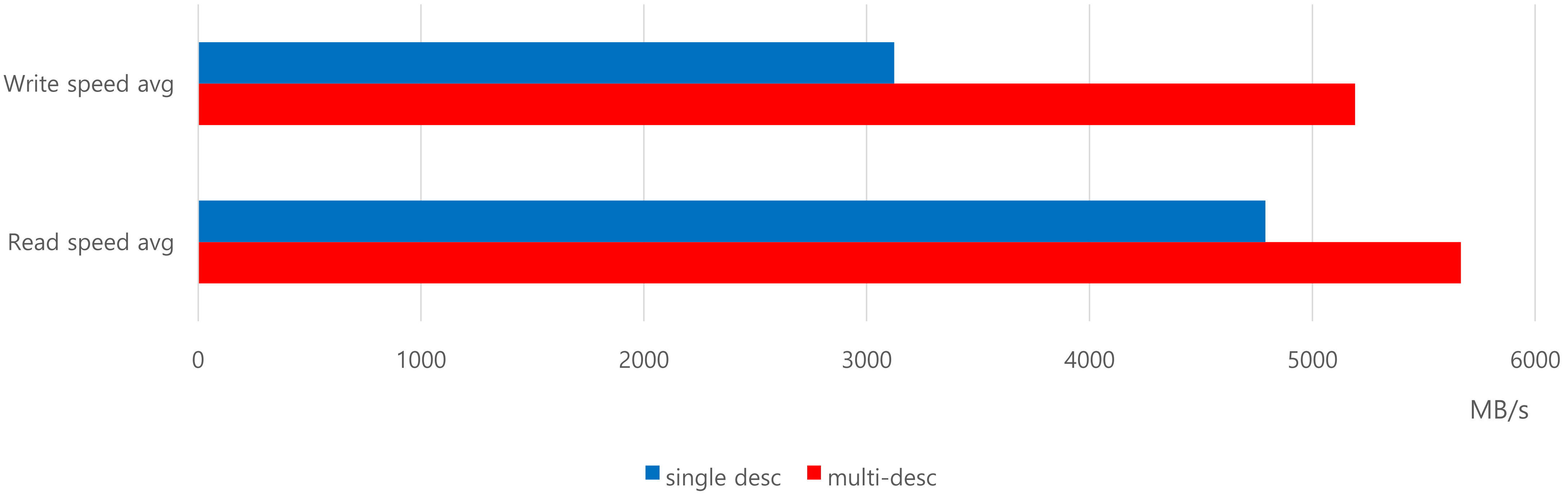
- Client

CPU: intel Silver 4215 x 2
DRAM: 64GB
NIC: MCX516A-CCAT

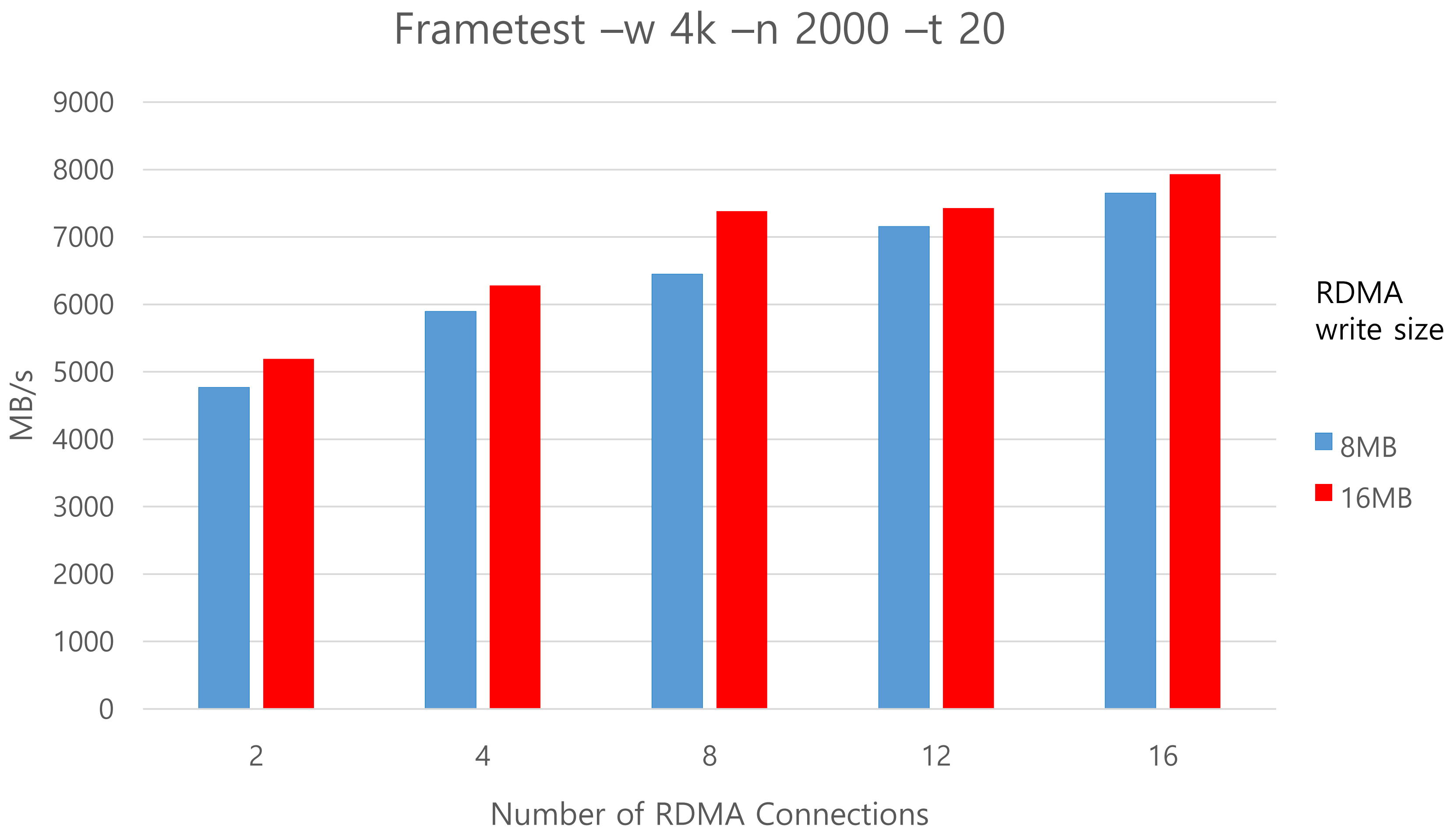
Performance comparison between single and multi-descriptor.

Frametest –w 4k –n 2000 –t 20

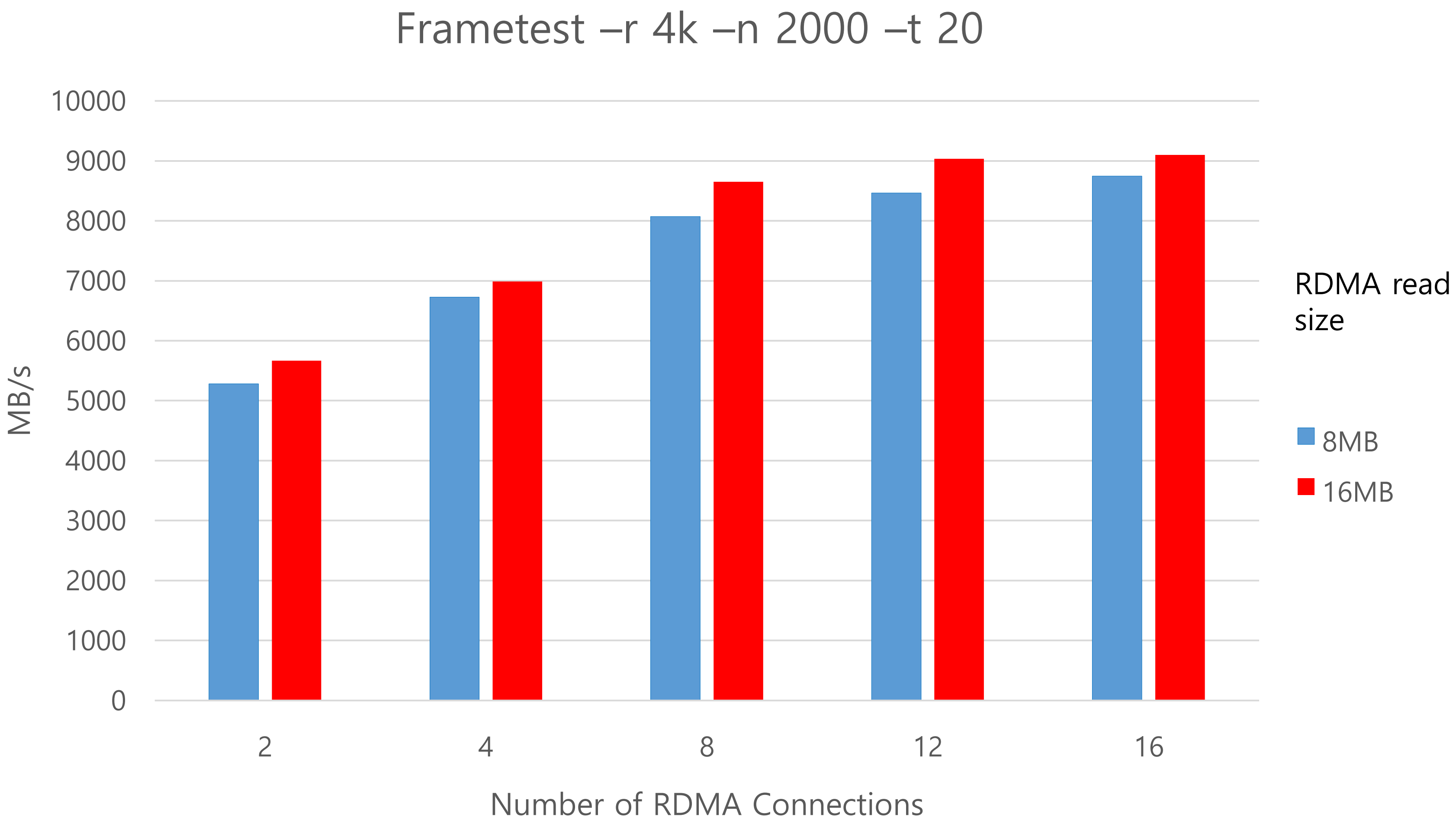
Frametest –r 4k –n 2000 –t 20



RDMA write performance per number of connections



RDMA read performance per number of connections



Linux Kernel Server, KSMBD (continued)

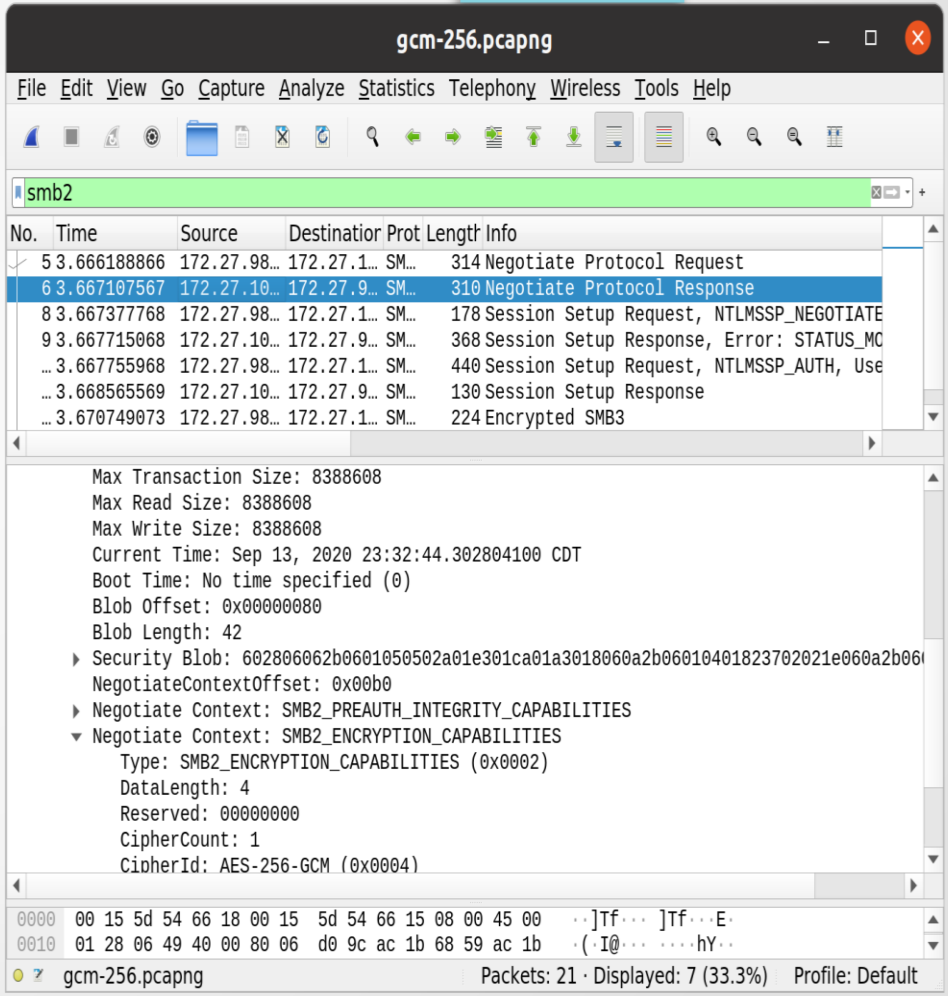
- If interested in contributing there are lots of cool features to work on, as well as improved integration with Samba (e.g. user space upcalls for additional features). The SMB3.1.1 family of protocols is huge!
- Roles: there are multiple developers helping Namjae (the maintainer). I am managing the git merges, ensuring additional functional testing is done regularly, and reviewing patches as requested by Namjae (my focus is largely on the client)
- Namjae would welcome additional help with code reviews, security auditing, testing and new features
- Very exciting time!

Recent improvements in the kernel client

(cifs.ko)

AES-GCM-256 (Strongest Encryption)

- Negotiated by default if server requires it (Azure, Windows, ksmbd etc. support it now)
- Client can require (force) if “require_gcm_256” module parm set
- See trace of Linux AES-GCM-256 mount to Windows (with “require_gcm_256” set on client)


gcm-256.pcapng
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help

smb2

```
No. Time        Source       Destinatior Prot Length Info
  5 3.666188866 172.27.98… 172.27.1… SM…  314 Negotiate Protocol Request
  6 3.667107567 172.27.10… 172.27.9… SM…  310 Negotiate Protocol Response
  8 3.667377768 172.27.98… 172.27.1… SM…  178 Session Setup Request, NTLMSSP_NEGOTIATE
  9 3.667715068 172.27.10… 172.27.9… SM…  368 Session Setup Response, Error: STATUS_MC
  … 3.667755968 172.27.98… 172.27.1… SM…  440 Session Setup Request, NTLMSSP_AUTH, Use
  … 3.668565569 172.27.10… 172.27.9… SM…  130 Session Setup Response
  … 3.670749073 172.27.98… 172.27.1… SM…  224 Encrypted SMB3
```

```
    Max Transaction Size: 8388608
    Max Read Size: 8388608
    Max Write Size: 8388608
    Current Time: Sep 13, 2020 23:32:44.302804100 CDT
    Boot Time: No time specified (0)
    Blob Offset: 0x00000080
    Blob Length: 42
  ▸ Security Blob: 602806062b0601050502a01e301ca01a3018060a2b06010401823702021e060a2b06
    NegotiateContextOffset: 0x00b0
  ▸ Negotiate Context: SMB2_PREAUTH_INTEGRITY_CAPABILITIES
  ▾ Negotiate Context: SMB2_ENCRYPTION_CAPABILITIES
      Type: SMB2_ENCRYPTION_CAPABILITIES (0x0002)
      DataLength: 4
      Reserved: 00000000
      CipherCount: 1
      CipherId: AES-256-GCM (0x0004)
```

```
0000  00 15 5d 54 66 18 00 15  5d 54 66 15 08 00 45 00   ··]Tf··· ]Tf···E·
0010  01 28 06 49 40 00 80 06  d0 9c ac 1b 68 59 ac 1b   ·(·I@··· ····hY··
```

gcm-256.pcapng   Packets: 21 · Displayed: 7 (33.3%)   Profile: Default

What about Performance Improvements?

It rocks! Let’s take a simple example and copy 10GB from Azure server down to Linux client VM

“dd if=/mnt/10GB of=/dev/null bs=1M count=10K”

- Old defaults (3.0) 143MB/sec
- With 3.1.1 201MB/sec (41% faster)
- And go to 2 channels & set new parm “rasize” to 4MB

453MB/sec, More than 3x faster!!

- Lots of great perf improvements!

And another example (thank you Rohith!)

- Support added for handle leases (deferred close) in 5.13 kernel. Here are two simple example of the huge caching perf gains even copying to Samba localhost
- Create a 2GB file and read it back (read is 4x faster)

```
dd if=/dev/urandom of=2G bs=1M count=2K ;
dd if=2G bs=1M count=2K of=/dev/null
```

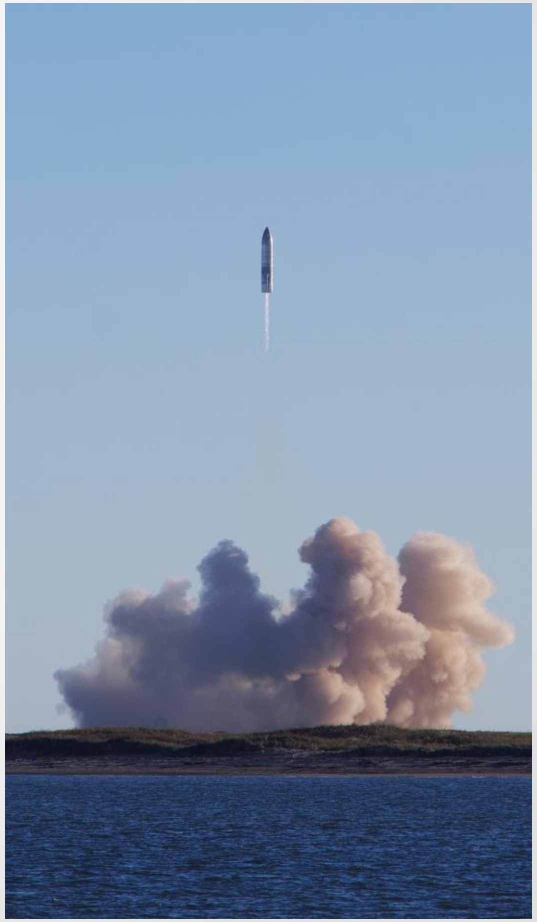
– Before:2.0 GiB copied, 0.583143 s, 3.7 GB/s

– Current: 2.0 GiB copied, 0.159237 s, 13.5 GB/s

- Read the same 4GB file twice (2nd time is 3x faster)
- dd if=4G of=/dev/null bs=1M count=4K ;
- dd if=4G of=/dev/null bs=1M count=4K

– Before: 4.0 GiB copied, 1.36794 s, 3.1 GB/s

– Current: 4.0 GiB copied, 0.441635 s, 9.7 GB/s

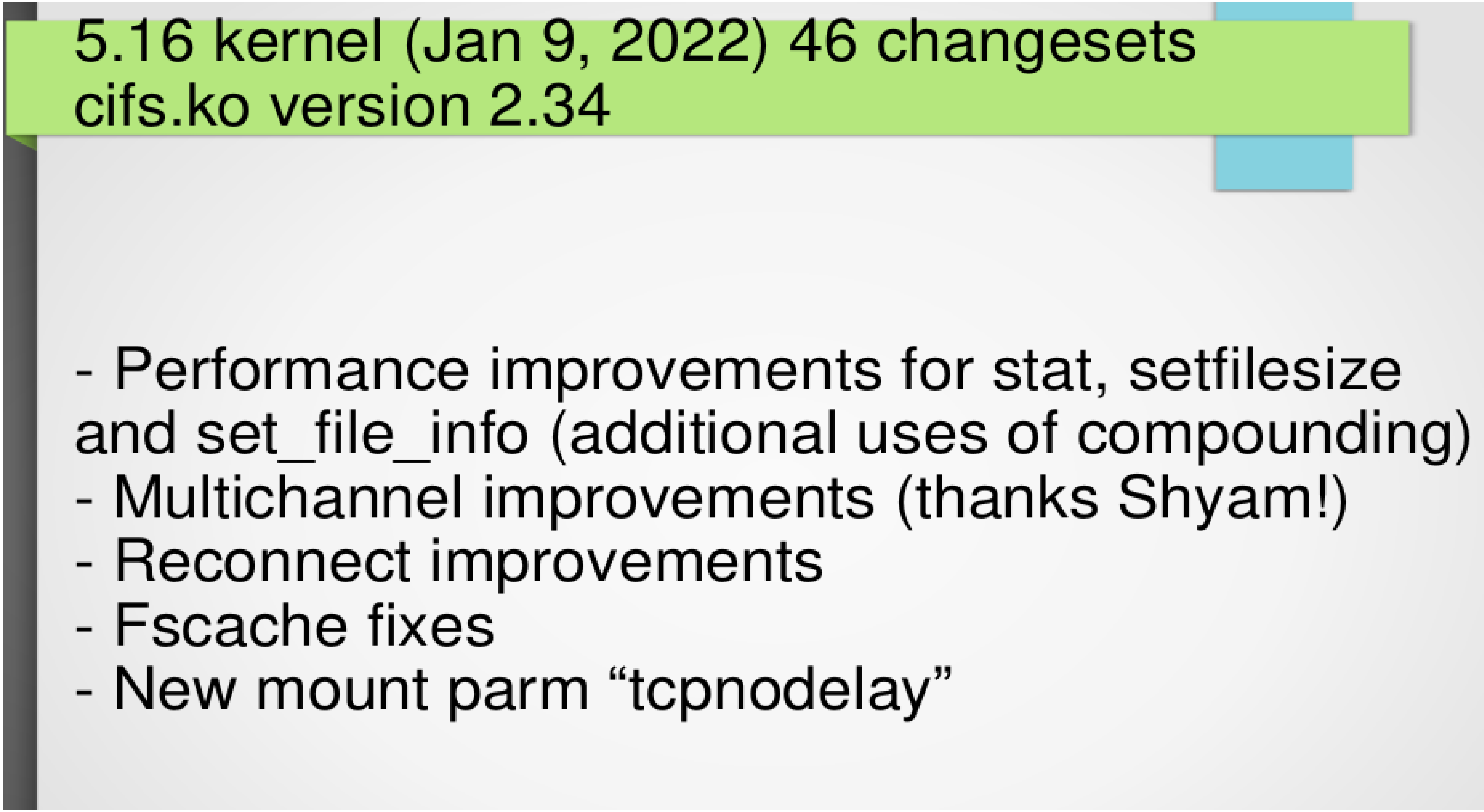


5.13 kernel (June 27th 2021)
66 changesets. cifs.ko version 2.32

- Huge performance boost for readahead in some configurations by setting new mount parameter ("rasize=") larger than rsize
- Add support for fcollapse and finsert (collapse and insert range calls)
- Add support for deferred close (handle leases), greatly improving performance of some workloads
- Improvements to directory caching of the root directory
- Strongest type of encryption (GCM256) is now sent by default in the list of allowed encryption algorithms (GCM128 preferred, then GCM256 then CCM128) and does not have to be enabled manually in module load time parameters
- Debugging of encrypted mounts improved (e.g. for multiuser mounts and also for GCM256)
- Add support for shutddown ioctl (useful to halt new activity to better allow emergency unmounts, and also required for some common testcases)
- Mount error handling improvements (see *"/proc/fs/cifs/mount_params")*

5.14 kernel (August 29th) 71 changesets, cifs.ko version 2.33

- Fallocate improvements (can now alloc smaller ranges up to 1MB). Thank you Ronnie!
- DFS reconnect improvements, and reconnect retry improvements. Thank you Paulo!
- Experimental support added for negotiating signing algorithm
- And 5.15 kernel
  - Important deferred close (handle lease) bug fixes
  - Support for weaker authentication (NTLMv1 and LANMAN) removed
  - (And experimental kernel server, ksmbd, merged)



5.16 kernel (Jan 9, 2022) 46 changesets cifs.ko version 2.34

- Performance improvements for stat, setfilesize and set_file_info (additional uses of compounding)
- Multichannel improvements (thanks Shyam!)
- Reconnect improvements
- Fscache fixes
- New mount parm “tcpnodelay”

5.17 kernel (March 20th), 51 changesets, cifs.ko ver 2.35

- Add support for new fscache (offline files caching mechanism)
- Send additional NTLMSSP info (including module and OS version) for improved debugging
- DFS and ACL fixes
- Restructuring of multichannel code

5.18 kernel (May 22nd), 40 changesets, cifs.ko ver 2.36

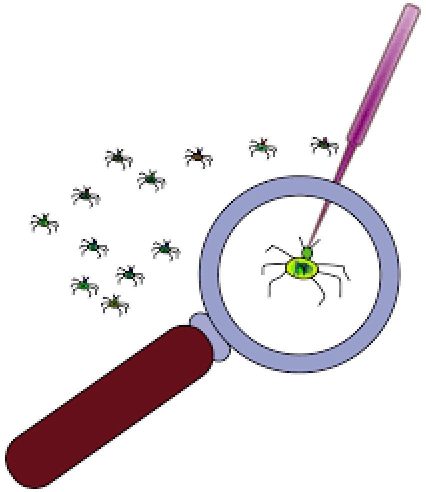
- Important performance improvement (reuse cached file handle for various common operations like stat and statfs if available), greatly reducing metadata operations (like open/close)
- Important fscache (offline file caching) and DFS improvements
- cross mount reflink now supported, which can dramatically improve copy performance from one share to another (on the same server) if they support duplicate extents.

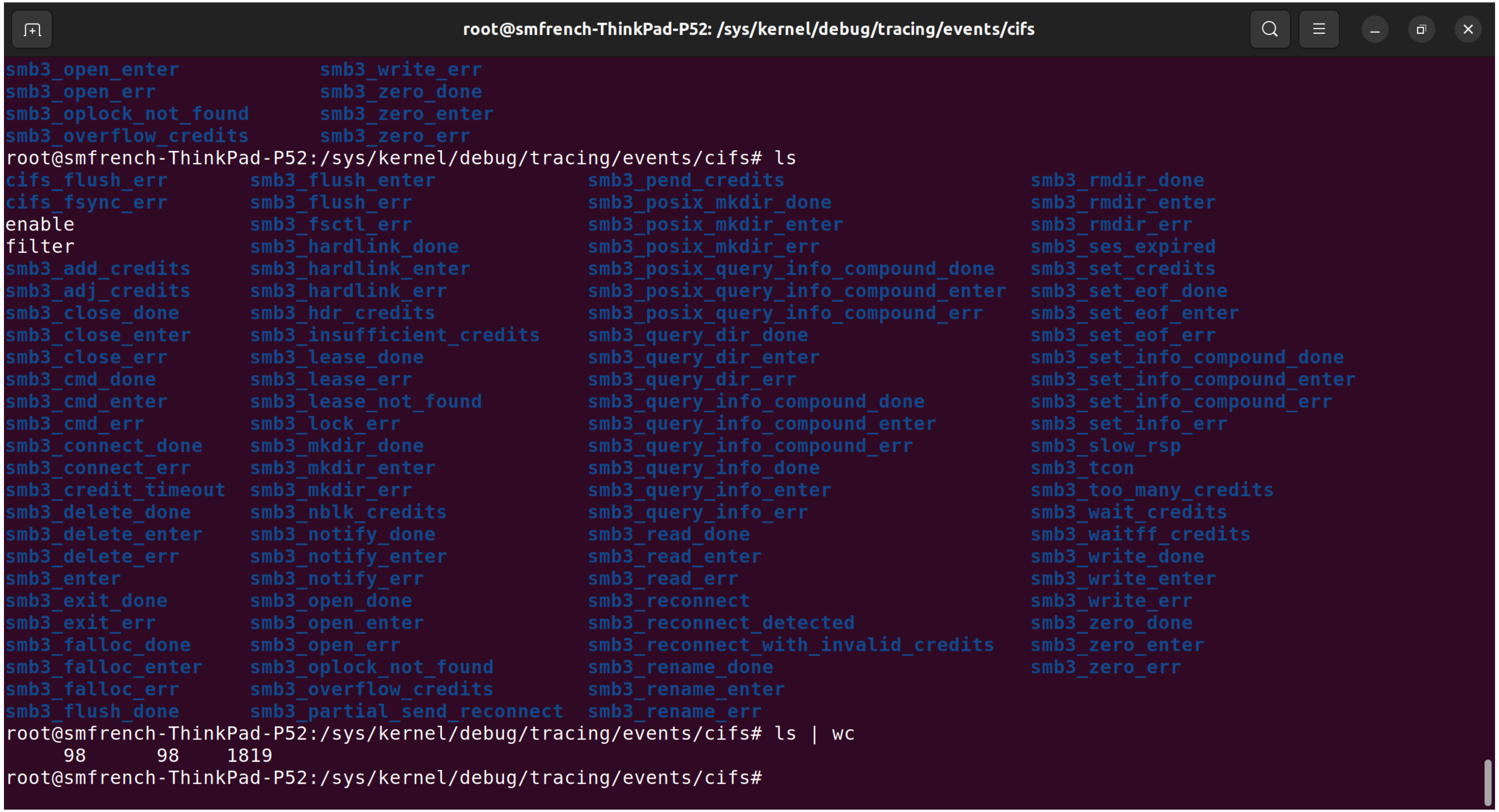
5.19-rc kernel (expected in early August)

- Important performance optimization for directory searches, now we cache the root directory content (to the many servers which support directory leases) reducing amount of network traffic for queries in the root directory
- Multichannel reconnect improvements (e.g. when address or interfaces change)
- RDMA (smbdirect) improvements

Tracing continues to improve ...

- Added more than 10 additional dynamic tracepoints





root@smfrench-ThinkPad-P52: /sys/kernel/debug/tracing/events/cifs

```
smb3_open_enter            smb3_write_err
smb3_open_err              smb3_zero_done
smb3_oplock_not_found      smb3_zero_enter
smb3_overflow_credits      smb3_zero_err
root@smfrench-ThinkPad-P52:/sys/kernel/debug/tracing/events/cifs# ls
cifs_flush_err       smb3_flush_enter             smb3_pend_credits                     smb3_rmdir_done
cifs_fsync_err       smb3_flush_err               smb3_posix_mkdir_done                 smb3_rmdir_enter
enable               smb3_fsctl_err               smb3_posix_mkdir_enter                smb3_rmdir_err
filter               smb3_hardlink_done           smb3_posix_mkdir_err                  smb3_ses_expired
smb3_add_credits     smb3_hardlink_enter          smb3_posix_query_info_compound_done   smb3_set_credits
smb3_adj_credits     smb3_hardlink_err            smb3_posix_query_info_compound_enter  smb3_set_eof_done
smb3_close_done      smb3_hdr_credits             smb3_posix_query_info_compound_err    smb3_set_eof_enter
smb3_close_enter     smb3_insufficient_credits    smb3_query_dir_done                   smb3_set_eof_err
smb3_close_err       smb3_lease_done              smb3_query_dir_enter                  smb3_set_info_compound_done
smb3_cmd_done        smb3_lease_err               smb3_query_dir_err                    smb3_set_info_compound_enter
smb3_cmd_enter       smb3_lease_not_found         smb3_query_info_compound_done         smb3_set_info_compound_err
smb3_cmd_err         smb3_lock_err                smb3_query_info_compound_enter        smb3_set_info_err
smb3_connect_done    smb3_mkdir_done              smb3_query_info_compound_err          smb3_slow_rsp
smb3_connect_err     smb3_mkdir_enter             smb3_query_info_done                  smb3_tcon
smb3_credit_timeout  smb3_mkdir_err               smb3_query_info_enter                 smb3_too_many_credits
smb3_delete_done     smb3_nblk_credits            smb3_query_info_err                   smb3_wait_credits
smb3_delete_enter    smb3_notify_done             smb3_read_done                        smb3_waitff_credits
smb3_delete_err      smb3_notify_enter            smb3_read_enter                       smb3_write_done
smb3_enter           smb3_notify_err              smb3_read_err                         smb3_write_enter
smb3_exit_done       smb3_open_done               smb3_reconnect                        smb3_write_err
smb3_exit_err        smb3_open_enter              smb3_reconnect_detected               smb3_zero_done
smb3_falloc_done     smb3_open_err                smb3_reconnect_with_invalid_credits   smb3_zero_enter
smb3_falloc_enter    smb3_oplock_not_found        smb3_rename_done                      smb3_zero_err
smb3_falloc_err      smb3_overflow_credits        smb3_rename_enter
smb3_flush_done      smb3_partial_send_reconnect  smb3_rename_err
root@smfrench-ThinkPad-P52:/sys/kernel/debug/tracing/events/cifs# ls | wc
     98      98    1819
root@smfrench-ThinkPad-P52:/sys/kernel/debug/tracing/events/cifs#
```

eBPF is amazing ...

- See Brendan Gregg’s website
- Also see e.g. https://wiki.samba.org/index.php/LinuxCIFS_troubleshooting
- Can be as simple to do as “trace-cmd record -e cifs”
  - And then “trace-cmd show” in another window
- Let us know if suggestions on other debugging tracepoints that would be helpful
- And don’t forget about *proc/fs/cifs/Stats, proc/fs/cifs/open_files and proc*/fs/cifs/DebugData ...

Recent improvements – cifs-utils

Userspace tools

Improved user space tools (cifs-utils)

- cifs-utils 6.14 released in Sept
  - Add commands to view Alternate Data Streams
  - setcifsacl improvements
  - mproved debugging (keydump)
- Cifs-utils 6.15 released in April
- More recently
  - Add support for gss-proxy (improving krb5 credential retrieval)
  - Misc. bug fixes

Speaker
Photo Will
Be Placed
Here

eBPF is amazing ...

- See Brendan Gregg’s website
- Also see e.g. https://wiki.samba.org/index.php/LinuxCIFS_troubleshooting
- Can be as simple to do as “trace-cmd record -e cifs”
  - And then “trace-cmd show” in another window
- Let us know if suggestions on other debugging tracepoints that would be helpful
- And don’t forget about *proc/fs/cifs/Stats, proc/fs/cifs/open_files and proc*/fs/cifs/DebugData ...

Coming soon ...

New features under development for SMB3.1.1 on Linux

What features can you expect in next few releases?

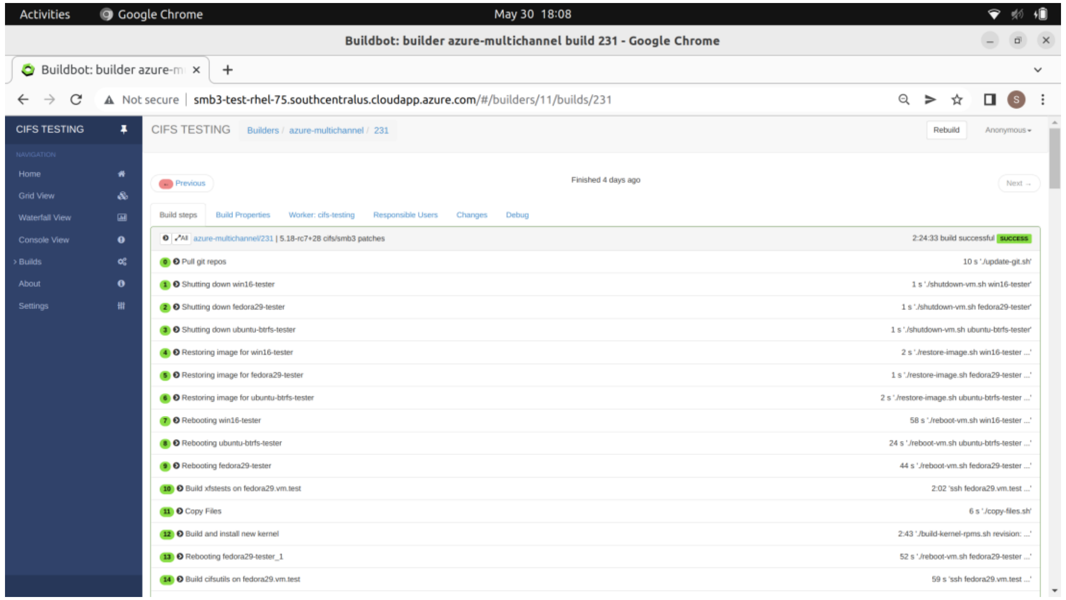
- Extending use of directory leases to greatly improve metadata caching (currently limited to contents of the root directory)
- SMB3.1.1 compression support (allow compressing network traffic based on the SMB3.1.1 compress mount parm)
- statx to return additional SMB3.1.1 attributes like “offline”
- Improvements to enable fanotify/inotify over SMB3.1.1 mounts (currently requires a private SMB3.1.1 specific ioctl)
- Prototype of SMB3.1.1 over QUIC (new encrypted network transport)
- More perf improvements for folios, cache, parallel i/o, multichannel
- More testing of the SMB3.1.1 POSIX extensions to ksmbd (and soon Samba server)

Testing Improvements

Section Subtitle

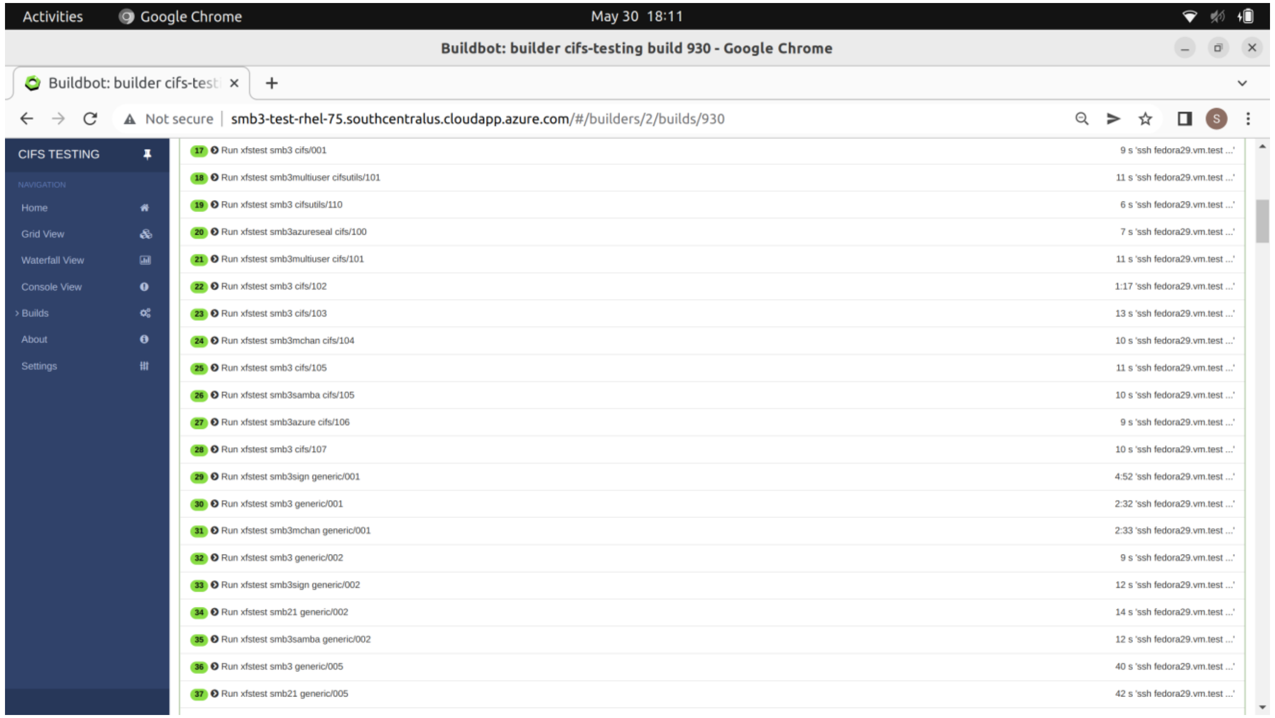
Automated testing has greatly improved

- Historically SMB3.1.1 plugfests multiple times a year have helped too
- The ‘buildbot’ continues to improve, more tests added, reducing regressions and improving quality
- Test groups for different server types and a general “cifs-testing” one



Finished 4 days ago

```
azure-multichannel/231 | 5.18-rc7+28 cifs/smb3 patches          2:24:33 build successful SUCCESS
0  Pull git repos                                  10 s './update-git.sh'
1  Shutting down win16-tester                      1 s './shutdown-vm.sh win16-tester'
2  Shutting down fedora29-tester                   1 s './shutdown-vm.sh fedora29-tester'
3  Shutting down ubuntu-btrfs-tester               1 s './shutdown-vm.sh ubuntu-btrfs-tester'
4  Restoring image for win16-tester                2 s './restore-image.sh win16-tester ...'
5  Restoring image for fedora29-tester             1 s './restore-image.sh fedora29-tester ...'
6  Restoring image for ubuntu-btrfs-tester         2 s './restore-image.sh ubuntu-btrfs-tester ...'
7  Rebooting win16-tester                          58 s './reboot-vm.sh win16-tester ...'
8  Rebooting ubuntu-btrfs-tester                   24 s './reboot-vm.sh ubuntu-btrfs-tester ...'
9  Rebooting fedora29-tester                       44 s './reboot-vm.sh fedora29-tester ...'
10 Build xfstests on fedora29.vm.test              2:02 'ssh fedora29.vm.test ...'
11 Copy Files                                      6 s './copy-files.sh'
12 Build and install new kernel                    2:43 './build-kernel-rpms.sh revision: ...'
13 Rebooting fedora29-tester_1                     52 s './reboot-vm.sh fedora29-tester ...'
14 Build cifsutils on fedora29.vm.test             59 s 'ssh fedora29.vm.test ...'
```


Activities
Google Chrome
May 30 18:11
Buildbot: builder cifs-testing build 930 - Google Chrome
Buildbot: builder cifs-testi
Not secure
smb3-test-rhel-75.southcentralus.cloudapp.azure.com/#/builders/2/builds/930
CIFS TESTING
NAVIGATION
Home
Grid View
Waterfall View
Console View
Builds
About
Settings

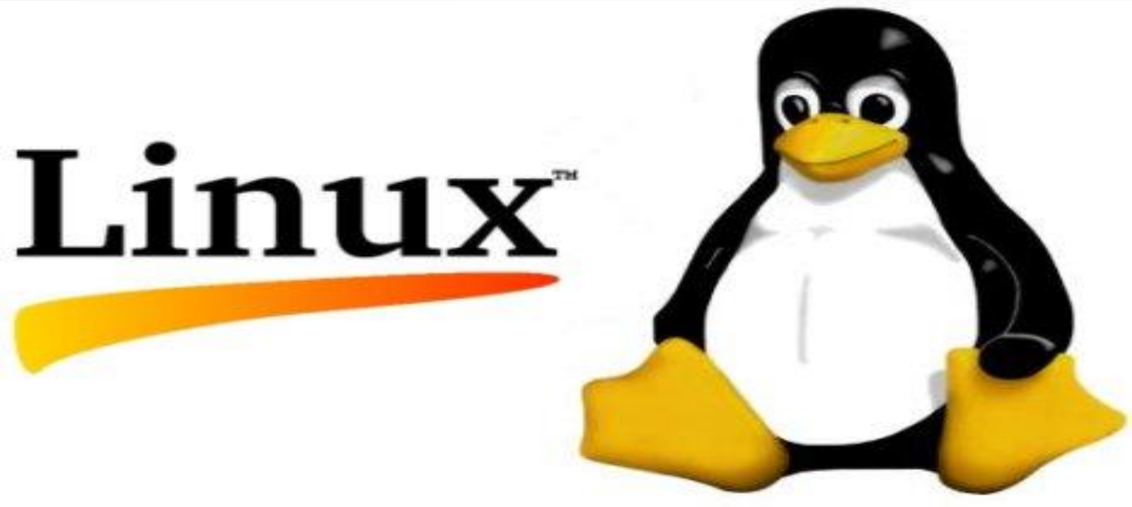
```
17 Run xfstest smb3 cifs/001                       9 s 'ssh fedora29.vm.test ...'
18 Run xfstest smb3multiuser cifsutils/101         11 s 'ssh fedora29.vm.test ...'
19 Run xfstest smb3 cifsutils/110                  6 s 'ssh fedora29.vm.test ...'
20 Run xfstest smb3azureseal cifs/100              7 s 'ssh fedora29.vm.test ...'
21 Run xfstest smb3multiuser cifs/101              11 s 'ssh fedora29.vm.test ...'
22 Run xfstest smb3 cifs/102                       1:17 'ssh fedora29.vm.test ...'
23 Run xfstest smb3 cifs/103                       13 s 'ssh fedora29.vm.test ...'
24 Run xfstest smb3mchan cifs/104                  10 s 'ssh fedora29.vm.test ...'
25 Run xfstest smb3 cifs/105                       11 s 'ssh fedora29.vm.test ...'
26 Run xfstest smb3samba cifs/105                  10 s 'ssh fedora29.vm.test ...'
27 Run xfstest smb3azure cifs/106                  9 s 'ssh fedora29.vm.test ...'
28 Run xfstest smb3 cifs/107                       10 s 'ssh fedora29.vm.test ...'
29 Run xfstest smb3sign generic/001                4:52 'ssh fedora29.vm.test ...'
30 Run xfstest smb3 generic/001                    2:32 'ssh fedora29.vm.test ...'
31 Run xfstest smb3mchan generic/001               2:33 'ssh fedora29.vm.test ...'
32 Run xfstest smb3 generic/002                    9 s 'ssh fedora29.vm.test ...'
33 Run xfstest smb3sign generic/002                12 s 'ssh fedora29.vm.test ...'
34 Run xfstest smb21 generic/002                   14 s 'ssh fedora29.vm.test ...'
35 Run xfstest smb3samba generic/002               12 s 'ssh fedora29.vm.test ...'
36 Run xfstest smb3 generic/005                    40 s 'ssh fedora29.vm.test ...'
37 Run xfstest smb21 generic/005                   42 s 'ssh fedora29.vm.test ...'
```

Additional tests are encouraged

- Xfstests are the standard Linux filesystem functional tests
- Added 21 to the main “cifs-testing” regression testing group (up to 245 tests run on every checkin from this group)
- Various server specific groups have added even more
  - Azure SMB3.1.1 multichannel: up 25% more tests, now includes 133 tests
  - Ksmbd (Linux kernel server target) up 15%, now includes 144 tests
- There are detailed wiki pages on wiki.samba.org going through how to setup xfstests with cifs.ko, and what features need to be added to enable more tests (tests that currently skip or fail so aren’t run in the ‘buildbot’)

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
  - 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
  - Kernel server code: https://git.samba.org/ksmbd.git/?p=ksmbd.git (ksmbd-for-next branch)