NFSv4ish Advanced Features

Ruben Gaspar

IMS

Agenda

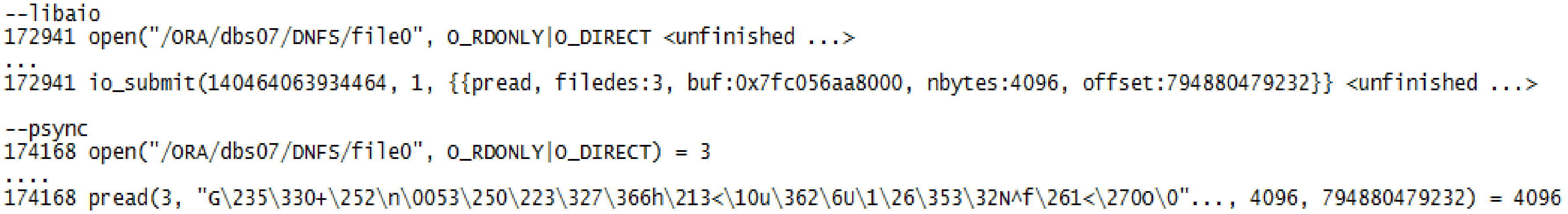
- O_DIRECT
- Delegations
- pNFS
- Kerberos + NFS

Please have a look to previous talk about this topic:

https://indico.cern.ch/event/505068/

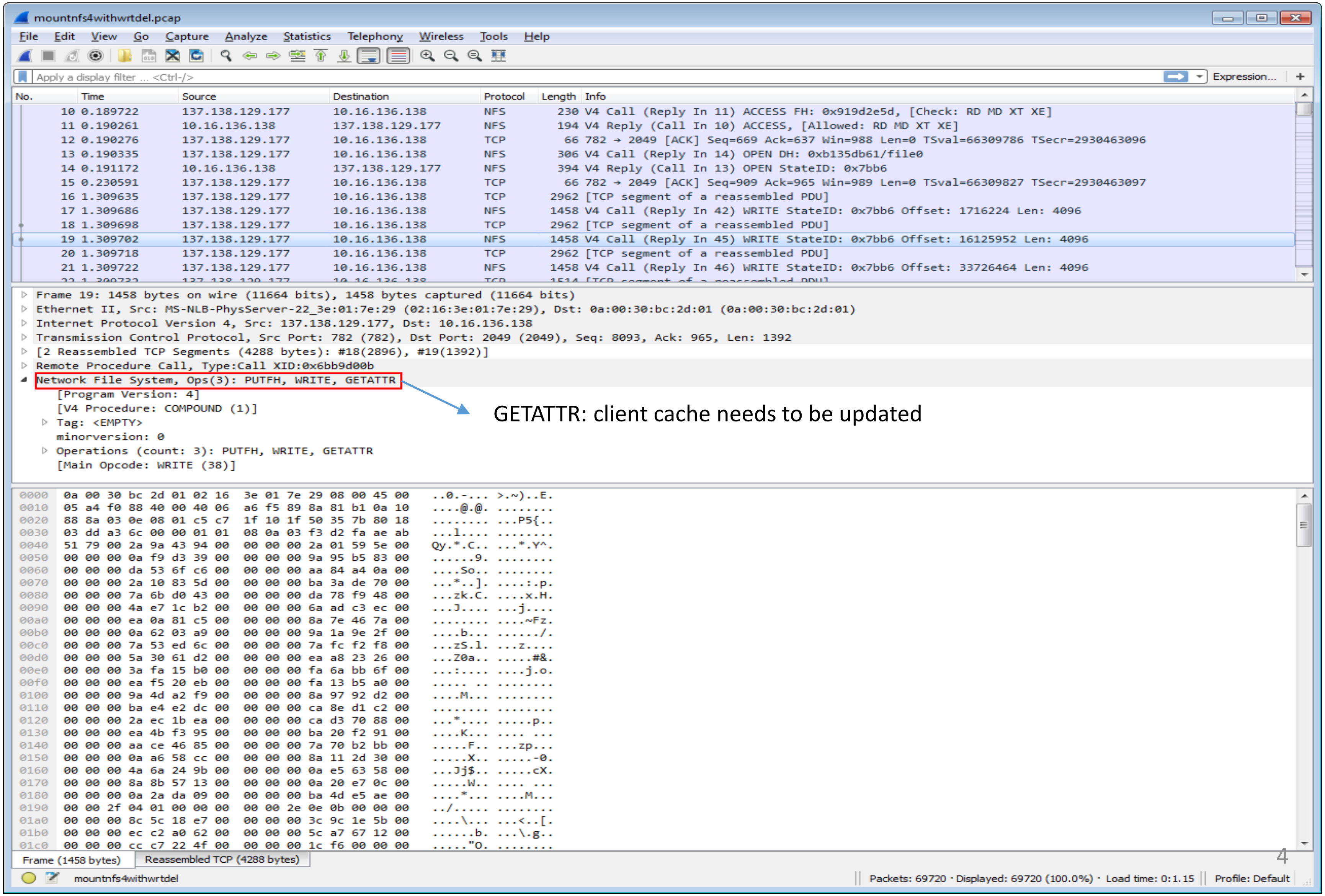
O_DIRECT

- Parallel access (e.g. clustering) relying on it
- From REHL 7.3 on, setting O_DIRECT will avoid delegation
- Tcpdump:
  - NFSv4: writes differed on a GETTR (to refresh client cache when directIO disabled)
  - NFSv4.1: metadata calls go to a metadata server, so GETATTR don’t appear



```
--libaio
172941 open("/ORA/dbs07/DNFS/file0", O_RDONLY|O_DIRECT <unfinished ...>
...
172941 io_submit(140464063934464, 1, {{pread, filedes:3, buf:0x7fc056aa8000, nbytes:4096, offset:794880479232}} <unfinished ...>

--psync
174168 open("/ORA/dbs07/DNFS/file0", O_RDONLY|O_DIRECT) = 3
....
174168 pread(3, "G\235\330+\252\n\0053\250\223\327\366h\213<\10u\362\6U\1\26\353\32N^f\261<\270o\0"..., 4096, 794880479232) = 4096
```

O_DIRECT (NFSv4, direct=0)


mountnfs4withwrtdel.pcap
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
Apply a display filter ... <Ctrl-/>
Expression...
+

```
No.   Time      Source           Destination      Protocol Length Info
   10 0.189722  137.138.129.177  10.16.136.138    NFS       230 V4 Call (Reply In 11) ACCESS FH: 0x919d2e5d, [Check: RD MD XT XE]
   11 0.190261  10.16.136.138    137.138.129.177  NFS       194 V4 Reply (Call In 10) ACCESS, [Allowed: RD MD XT XE]
   12 0.190276  137.138.129.177  10.16.136.138    TCP        66 782 → 2049 [ACK] Seq=669 Ack=637 Win=988 Len=0 TSval=66309786 TSecr=2930463096
   13 0.190335  137.138.129.177  10.16.136.138    NFS       306 V4 Call (Reply In 14) OPEN DH: 0xb135db61/file0
   14 0.191172  10.16.136.138    137.138.129.177  NFS       394 V4 Reply (Call In 13) OPEN StateID: 0x7bb6
   15 0.230591  137.138.129.177  10.16.136.138    TCP        66 782 → 2049 [ACK] Seq=909 Ack=965 Win=989 Len=0 TSval=66309827 TSecr=2930463097
   16 1.309635  137.138.129.177  10.16.136.138    TCP      2962 [TCP segment of a reassembled PDU]
   17 1.309686  137.138.129.177  10.16.136.138    NFS      1458 V4 Call (Reply In 42) WRITE StateID: 0x7bb6 Offset: 1716224 Len: 4096
   18 1.309698  137.138.129.177  10.16.136.138    TCP      2962 [TCP segment of a reassembled PDU]
   19 1.309702  137.138.129.177  10.16.136.138    NFS      1458 V4 Call (Reply In 45) WRITE StateID: 0x7bb6 Offset: 16125952 Len: 4096
   20 1.309718  137.138.129.177  10.16.136.138    TCP      2962 [TCP segment of a reassembled PDU]
   21 1.309722  137.138.129.177  10.16.136.138    NFS      1458 V4 Call (Reply In 46) WRITE StateID: 0x7bb6 Offset: 33726464 Len: 4096
   22 1.309732  137.138.129.177  10.16.136.138    TCP      1514 [TCP segment of a reassembled PDU]
```

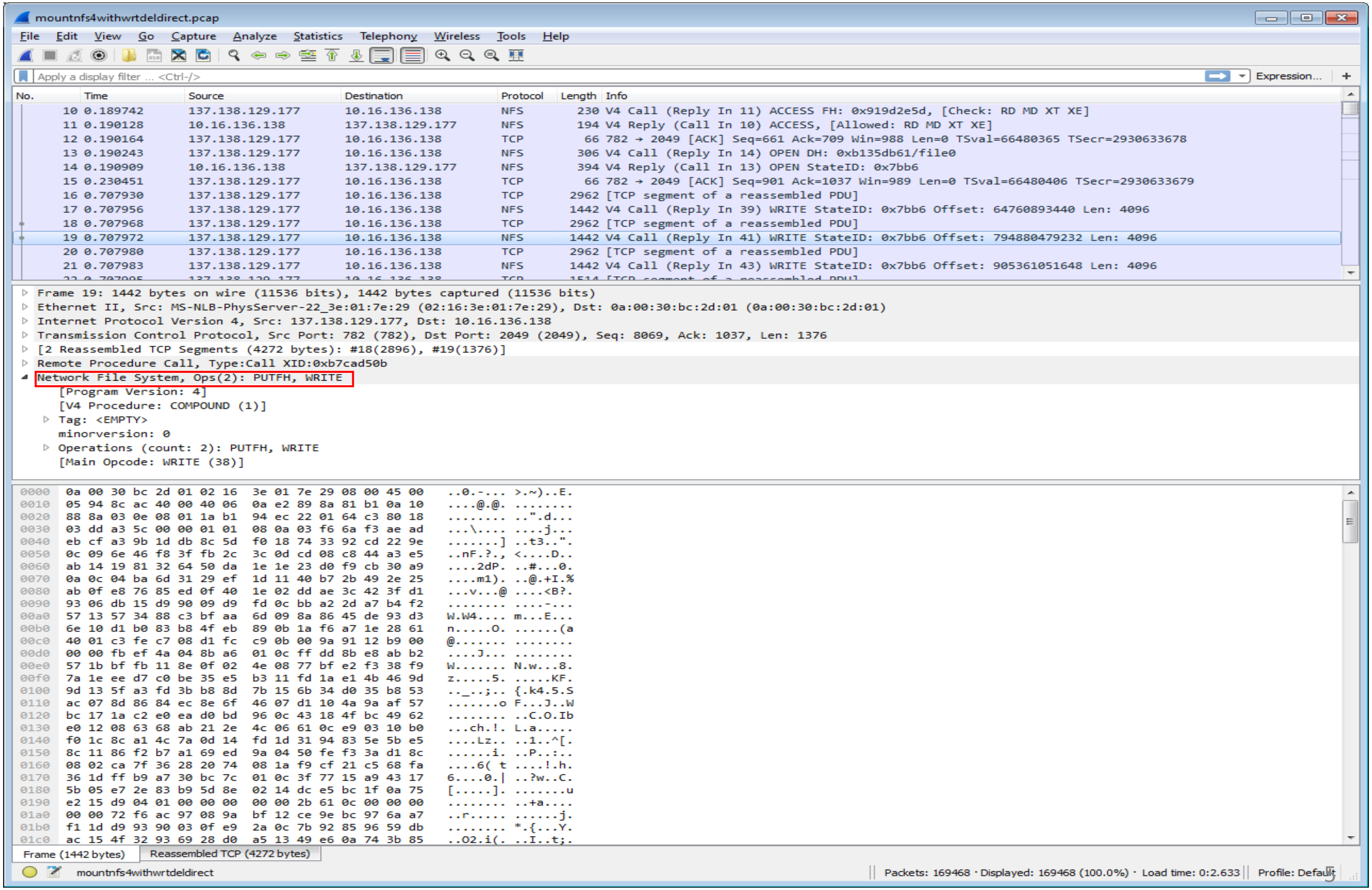
```
▷ Frame 19: 1458 bytes on wire (11664 bits), 1458 bytes captured (11664 bits)
▷ Ethernet II, Src: MS-NLB-PhysServer-22_3e:01:7e:29 (02:16:3e:01:7e:29), Dst: 0a:00:30:bc:2d:01 (0a:00:30:bc:2d:01)
▷ Internet Protocol Version 4, Src: 137.138.129.177, Dst: 10.16.136.138
▷ Transmission Control Protocol, Src Port: 782 (782), Dst Port: 2049 (2049), Seq: 8093, Ack: 965, Len: 1392
▷ [2 Reassembled TCP Segments (4288 bytes): #18(2896), #19(1392)]
▷ Remote Procedure Call, Type:Call XID:0x6bb9d00b
◢ Network File System, Ops(3): PUTFH, WRITE, GETATTR
    [Program Version: 4]
    [V4 Procedure: COMPOUND (1)]
  ▷ Tag: <EMPTY>
    minorversion: 0
  ▷ Operations (count: 3): PUTFH, WRITE, GETATTR
    [Main Opcode: WRITE (38)]
```

GETATTR: client cache needs to be updated

```
0000  0a 00 30 bc 2d 01 02 16  3e 01 7e 29 08 00 45 00   ..0.-... >.~)..E.
0010  05 a4 f0 88 40 00 40 06  a6 f5 89 8a 81 b1 0a 10   ....@.@. ........
0020  88 8a 03 0e 08 01 c5 c7  1f 10 1f 50 35 7b 80 18   ........ ...P5{..
0030  03 dd a3 6c 00 00 01 01  08 0a 03 f3 d2 fa ae ab   ...l.... ........
0040  51 79 00 2a 9a 43 94 00  00 00 00 2a 01 59 5e 00   Qy.*.C.. ...*.Y^.
0050  00 00 00 0a f9 d3 39 00  00 00 00 9a 95 b5 83 00   ......9. ........
0060  00 00 00 da 53 6f c6 00  00 00 00 aa 84 a4 0a 00   ....So.. ........
0070  00 00 00 2a 10 83 5d 00  00 00 00 ba 3a de 70 00   ...*..]. ....:.p.
0080  00 00 00 7a 6b d0 43 00  00 00 00 da 78 f9 48 00   ...zk.C. ....x.H.
0090  00 00 00 4a e7 1c b2 00  00 00 00 6a ad c3 ec 00   ...J.... ...j....
00a0  00 00 00 ea 0a 81 c5 00  00 00 00 8a 7e 46 7a 00   ........ ....~Fz.
00b0  00 00 00 0a 62 03 a9 00  00 00 00 9a 1a 9e 2f 00   ....b... ....../.
00c0  00 00 00 7a 53 ed 6c 00  00 00 00 7a fc f2 f8 00   ...zS.l. ...z....
00d0  00 00 00 5a 30 61 d2 00  00 00 00 ea a8 23 26 00   ...Z0a.. .....#&.
00e0  00 00 00 3a fa 15 b0 00  00 00 00 fa 6a bb 6f 00   ...:.... ....j.o.
00f0  00 00 00 ea f5 20 eb 00  00 00 00 fa 13 b5 a0 00   ..... .. ........
0100  00 00 00 9a 4d a2 f9 00  00 00 00 8a 97 92 d2 00   ....M... ........
0110  00 00 00 ba e4 e2 dc 00  00 00 00 ca 8e d1 c2 00   ........ ........
0120  00 00 00 2a ec 1b ea 00  00 00 00 ca d3 70 88 00   ...*.... .....p..
0130  00 00 00 ea 4b f3 95 00  00 00 00 ba 20 f2 91 00   ....K... .... ...
0140  00 00 00 aa ce 46 85 00  00 00 00 7a 70 b2 bb 00   .....F.. ...zp...
0150  00 00 00 0a a6 58 cc 00  00 00 00 8a 11 2d 30 00   .....X.. .....-0.
0160  00 00 00 4a 6a 24 9b 00  00 00 00 0a e5 63 58 00   ...Jj$.. .....cX.
0170  00 00 00 8a 8b 57 13 00  00 00 00 0a 20 e7 0c 00   .....W.. .... ...
0180  00 00 00 0a 2a da 09 00  00 00 00 ba 4d e5 ae 00   ....*... ....M...
0190  00 00 2f 04 01 00 00 00  00 00 2e 0e 0b 00 00 00   ../..... ........
01a0  00 00 00 8c 5c 18 e7 00  00 00 00 3c 9c 1e 5b 00   ....\... ...<..[.
01b0  00 00 00 ec c2 a0 62 00  00 00 00 5c a7 67 12 00   ......b. ...\.g..
01c0  00 00 00 cc c7 22 4f 00  00 00 00 1c f6 00 00 00   ....."O. ........
```

Frame (1458 bytes)
Reassembled TCP (4288 bytes)
mountnfs4withwrtdel
Packets: 69720 · Displayed: 69720 (100.0%) · Load time: 0:1.15
Profile: Default

O_DIRECT (NFSv4, direct=1)



```
No.  Time      Source           Destination      Protocol Length Info
  10 0.189742  137.138.129.177  10.16.136.138    NFS       230 V4 Call (Reply In 11) ACCESS FH: 0x919d2e5d, [Check: RD MD XT XE]
  11 0.190128  10.16.136.138    137.138.129.177  NFS       194 V4 Reply (Call In 10) ACCESS, [Allowed: RD MD XT XE]
  12 0.190164  137.138.129.177  10.16.136.138    TCP        66 782 → 2049 [ACK] Seq=661 Ack=709 Win=988 Len=0 TSval=66480365 TSecr=2930633678
  13 0.190243  137.138.129.177  10.16.136.138    NFS       306 V4 Call (Reply In 14) OPEN DH: 0xb135db61/file0
  14 0.190909  10.16.136.138    137.138.129.177  NFS       394 V4 Reply (Call In 13) OPEN StateID: 0x7bb6
  15 0.230451  137.138.129.177  10.16.136.138    TCP        66 782 → 2049 [ACK] Seq=901 Ack=1037 Win=989 Len=0 TSval=66480406 TSecr=2930633679
  16 0.707930  137.138.129.177  10.16.136.138    TCP      2962 [TCP segment of a reassembled PDU]
  17 0.707956  137.138.129.177  10.16.136.138    NFS      1442 V4 Call (Reply In 39) WRITE StateID: 0x7bb6 Offset: 64760893440 Len: 4096
  18 0.707968  137.138.129.177  10.16.136.138    TCP      2962 [TCP segment of a reassembled PDU]
  19 0.707972  137.138.129.177  10.16.136.138    NFS      1442 V4 Call (Reply In 41) WRITE StateID: 0x7bb6 Offset: 794880479232 Len: 4096
  20 0.707980  137.138.129.177  10.16.136.138    TCP      2962 [TCP segment of a reassembled PDU]
  21 0.707983  137.138.129.177  10.16.136.138    NFS      1442 V4 Call (Reply In 43) WRITE StateID: 0x7bb6 Offset: 905361051648 Len: 4096
```

```
▷ Frame 19: 1442 bytes on wire (11536 bits), 1442 bytes captured (11536 bits)
▷ Ethernet II, Src: MS-NLB-PhysServer-22_3e:01:7e:29 (02:16:3e:01:7e:29), Dst: 0a:00:30:bc:2d:01 (0a:00:30:bc:2d:01)
▷ Internet Protocol Version 4, Src: 137.138.129.177, Dst: 10.16.136.138
▷ Transmission Control Protocol, Src Port: 782 (782), Dst Port: 2049 (2049), Seq: 8069, Ack: 1037, Len: 1376
▷ [2 Reassembled TCP Segments (4272 bytes): #18(2896), #19(1376)]
▷ Remote Procedure Call, Type:Call XID:0xb7cad50b
◢ Network File System, Ops(2): PUTFH, WRITE
    [Program Version: 4]
    [V4 Procedure: COMPOUND (1)]
  ▷ Tag: <EMPTY>
    minorversion: 0
  ▷ Operations (count: 2): PUTFH, WRITE
    [Main Opcode: WRITE (38)]
```

```
0000  0a 00 30 bc 2d 01 02 16  3e 01 7e 29 08 00 45 00   ..0.-... >.~)..E.
0010  05 94 8c ac 40 00 40 06  0a e2 89 8a 81 b1 0a 10   ....@.@. ........
0020  88 8a 03 0e 08 01 1a b1  94 ec 22 01 64 c3 80 18   ........ ..".d...
0030  03 dd a3 5c 00 00 01 01  08 0a 03 f6 6a f3 ae ad   ...\.... ....j...
0040  eb cf a3 9b 1d db 8c 5d  f0 18 74 33 92 cd 22 9e   .......] ..t3..".
0050  0c 09 6e 46 f8 3f fb 2c  3c 0d cd 08 c8 44 a3 e5   ..nF.?., <....D..
0060  ab 14 19 81 32 64 50 da  1e 1e 23 d0 f9 cb 30 a9   ....2dP. ..#...0.
0070  0a 0c 04 ba 6d 31 29 ef  1d 11 40 b7 2b 49 2e 25   ....m1). ..@.+I.%
0080  ab 0f e8 76 85 ed 0f 40  1e 02 dd ae 3c 42 3f d1   ...v...@ ....<B?.
0090  93 06 db 15 d9 90 09 d9  fd 0c bb a2 2d a7 b4 f2   ........ ....-...
00a0  57 13 57 34 88 c3 bf aa  6d 09 8a 86 45 de 93 d3   W.W4.... m...E...
00b0  6e 10 d1 b0 83 b8 4f eb  89 0b 1a f6 a7 1e 28 61   n.....O. ......(a
00c0  40 01 c3 fe c7 08 d1 fc  c9 0b 00 9a 91 12 b9 00   @....... ........
00d0  00 00 fb ef 4a 04 8b a6  01 0c ff dd 8b e8 ab b2   ....J... ........
00e0  57 1b bf fb 11 8e 0f 02  4e 08 77 bf e2 f3 38 f9   W....... N.w...8.
00f0  7a 1e ee d7 c0 be 35 e5  b3 11 fd 1a e1 4b 46 9d   z.....5. .....KF.
0100  9d 13 5f a3 fd 3b b8 8d  7b 15 6b 34 d0 35 b8 53   .._..;.. {.k4.5.S
0110  ac 07 8d 86 84 ec 8e 6f  46 07 d1 10 4a 9a af 57   .......o F...J..W
0120  bc 17 1a c2 e0 ea d0 bd  96 0c 43 18 4f bc 49 62   ........ ..C.O.Ib
0130  e0 12 08 63 68 ab 21 2e  4c 06 61 0c e9 03 10 b0   ...ch.!. L.a.....
0140  f0 1c 8c a1 4c 7a 0d 14  fd 1d 31 94 83 5e 5b e5   ....Lz.. ..1..^[.
0150  8c 11 86 f2 b7 a1 69 ed  9a 04 50 fe f3 3a d1 8c   ......i. ..P..:..
0160  08 02 ca 7f 36 28 20 74  08 1a f9 cf 21 c5 68 fa   ....6( t ....!.h.
0170  36 1d ff b9 a7 30 bc 7c  01 0c 3f 77 15 a9 43 17   6....0.| ..?w..C.
0180  5b 05 e7 2e 83 b9 5d 8e  02 14 dc e5 bc 1f 0a 75   [.....]. .......u
0190  e2 15 d9 04 01 00 00 00  00 00 2b 61 0c 00 00 00   ........ ..+a....
01a0  00 00 72 f6 ac 97 08 9a  bf 12 ce 9e bc 97 6a a7   ..r..... ......j.
01b0  f1 1d d9 93 90 03 0f e9  2a 0c 7b 92 85 96 59 db   ........ *.{...Y.
01c0  ac 15 4f 32 93 69 28 d0  a5 13 49 e6 0a 74 3b 85   ..O2.i(. ..I..t;.
```

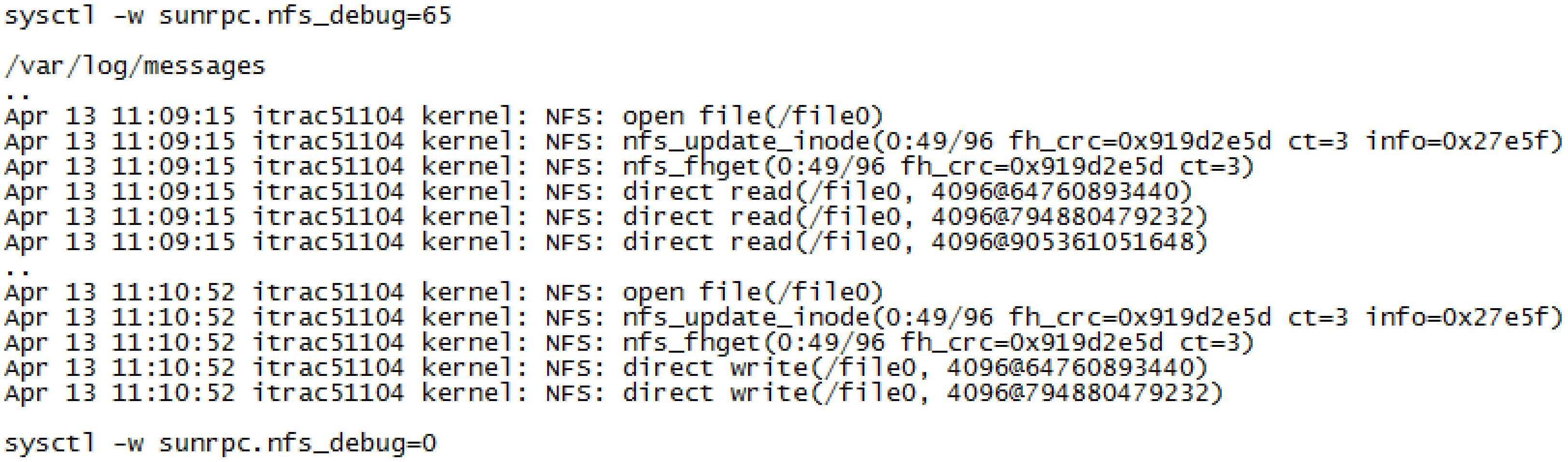
Frame (1442 bytes)
Reassembled TCP (4272 bytes)
mountnfs4withwrtdeldirect
Packets: 169468 · Displayed: 169468 (100.0%) · Load time: 0:2.633
Profile: Default

O_DIRECT (OS view)

- Enable extra NFS debugging (NetApp case 2006173081):

  ```
  sysctl -w sunrpc.nfs_debug=65
  ```

  - Logging at /var/log/messages
  - No difference among NFSv4 and NFSv4.1



```
sysctl -w sunrpc.nfs_debug=65

/var/log/messages
..
Apr 13 11:09:15 itrac51104 kernel: NFS: open file(/file0)
Apr 13 11:09:15 itrac51104 kernel: NFS: nfs_update_inode(0:49/96 fh_crc=0x919d2e5d ct=3 info=0x27e5f)
Apr 13 11:09:15 itrac51104 kernel: NFS: nfs_fhget(0:49/96 fh_crc=0x919d2e5d ct=3)
Apr 13 11:09:15 itrac51104 kernel: NFS: direct read(/file0, 4096@64760893440)
Apr 13 11:09:15 itrac51104 kernel: NFS: direct read(/file0, 4096@794880479232)
Apr 13 11:09:15 itrac51104 kernel: NFS: direct read(/file0, 4096@905361051648)
..
Apr 13 11:10:52 itrac51104 kernel: NFS: open file(/file0)
Apr 13 11:10:52 itrac51104 kernel: NFS: nfs_update_inode(0:49/96 fh_crc=0x919d2e5d ct=3 info=0x27e5f)
Apr 13 11:10:52 itrac51104 kernel: NFS: nfs_fhget(0:49/96 fh_crc=0x919d2e5d ct=3)
Apr 13 11:10:52 itrac51104 kernel: NFS: direct write(/file0, 4096@64760893440)
Apr 13 11:10:52 itrac51104 kernel: NFS: direct write(/file0, 4096@794880479232)

sysctl -w sunrpc.nfs_debug=0
```

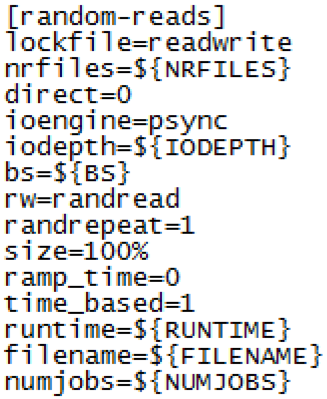
Delegations

Delegations

- Client establishes callback path information while SETCLIENTID verb, server checks with CB_NULL
- No extra callback path on RHEL 7 (https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/Storage_Administration_Guide/ch-nfs.html)

Delegations – test 1

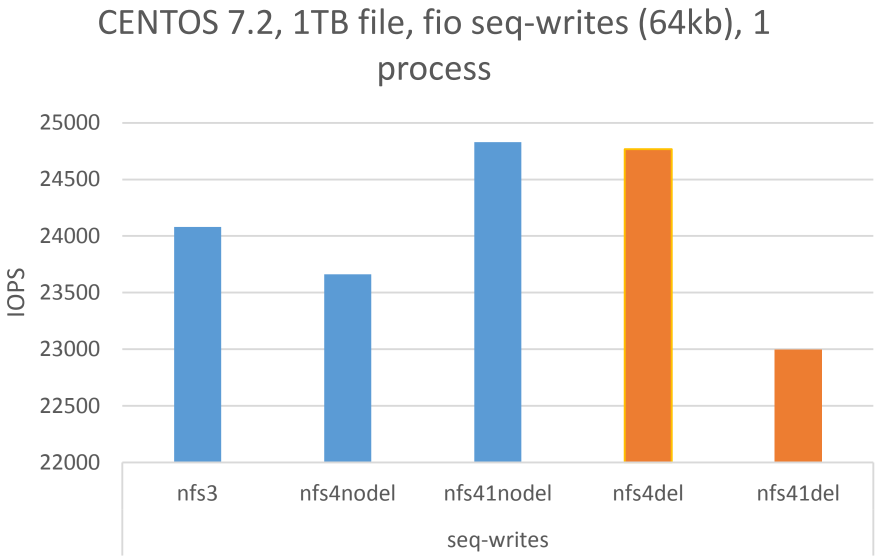
- Using fio-2.2.8 on a physical server CERNTOS 7.2

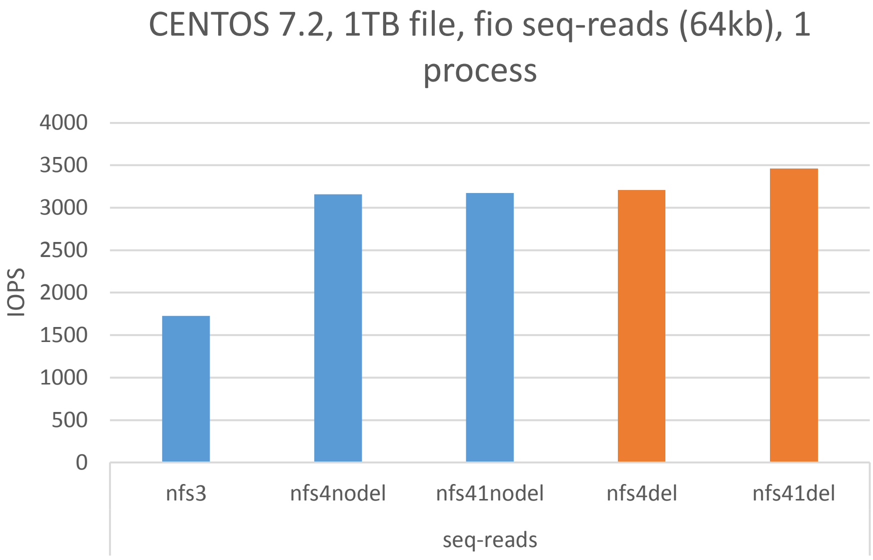


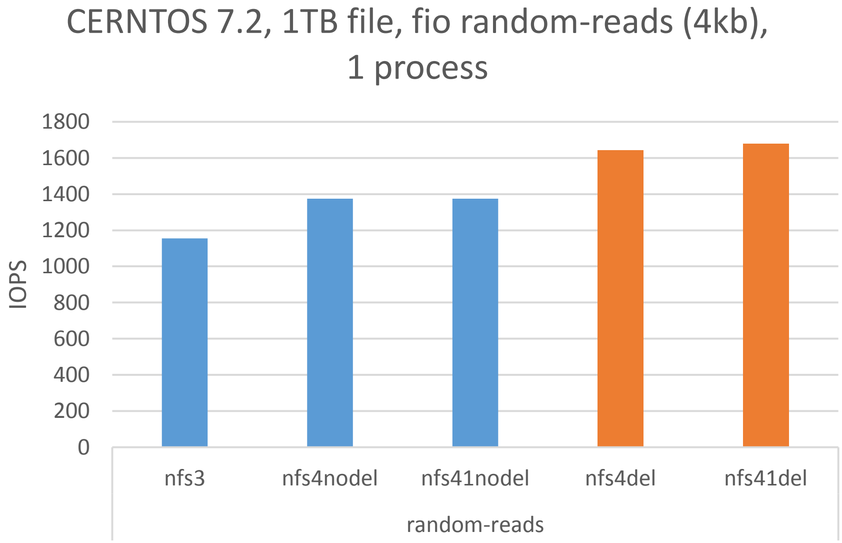
```
[random-reads]
lockfile=readwrite
nrfiles=${NRFILES}
direct=0
ioengine=psync
iodepth=${IODEPTH}
bs=${BS}
rw=randread
randrepeat=1
size=100%
ramp_time=0
time_based=1
runtime=${RUNTIME}
filename=${FILENAME}
numjobs=${NUMJOBS}
```

Changes with respect previous io tests

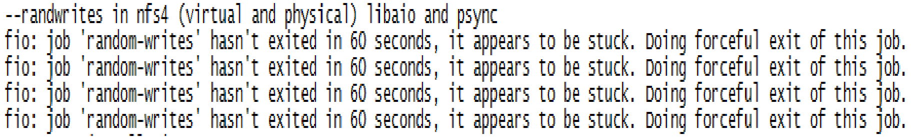
- fio loop 1000 times 10 secs IO


CENTOS 7.2, 1TB file, fio seq-writes (64kb), 1 process


CENTOS 7.2, 1TB file, fio seq-reads (64kb), 1 process


CERNTOS 7.2, 1TB file, fio random-reads (4kb), 1 process

Especially with NFSv4 random-writes tests don’t end properly either with sync or async IO. This is due to controller’s NVRAM. A job of 10secs may take some minutes to complete.



```
--randwrites in nfs4 (virtual and physical) libaio and psync
fio: job 'random-writes' hasn't exited in 60 seconds, it appears to be stuck. Doing forceful exit of this job.
fio: job 'random-writes' hasn't exited in 60 seconds, it appears to be stuck. Doing forceful exit of this job.
fio: job 'random-writes' hasn't exited in 60 seconds, it appears to be stuck. Doing forceful exit of this job.
fio: job 'random-writes' hasn't exited in 60 seconds, it appears to be stuck. Doing forceful exit of this job.
```

Delegations – test 2

- Based on article: http://cern.ch/go/xgl8
- Working on a CENTOS 7.2 server
  - 32 cores + 256GB RAM
- Python3 module developed to specifically test NFS delegations
  - https://gitlab.cern.ch/db/cerndb-infra-storage/tree/master/nfstestbench

```
python3 FileOps.py -h
usage: FileOps.py [-h] -f FILE_PREFIX -p POOLSIZE -c TOTALNUM [-i ITERACTIONS]
                  [-o] [-l] [-d] [-r | -w]

Delegation NFSv4 tests using multiple process program.

optional arguments:
  -h, --help        show this help message and exit
  -f FILE_PREFIX    Prefix value for file location
  -p POOLSIZE       Number of processes
  -c TOTALNUM       Number of files to work with
  -i ITERACTIONS    How many IO operations to do on a single file. Defaults to
                    10
  -o                if present we should use os.open, otherwise buffered IO.
  -l                Posix locking, otherwise no locking
  -d                direct IO, otherwise no direct IO
  -r                if present is a read
  -w                if present is a write
```

Delegations – test 2

- Create 10k files of 4KB
- Use Python multiprocessing module to distribute load on all cores
- Repeat a number of times the IO operation (R or W)
    - Use POSIX locks

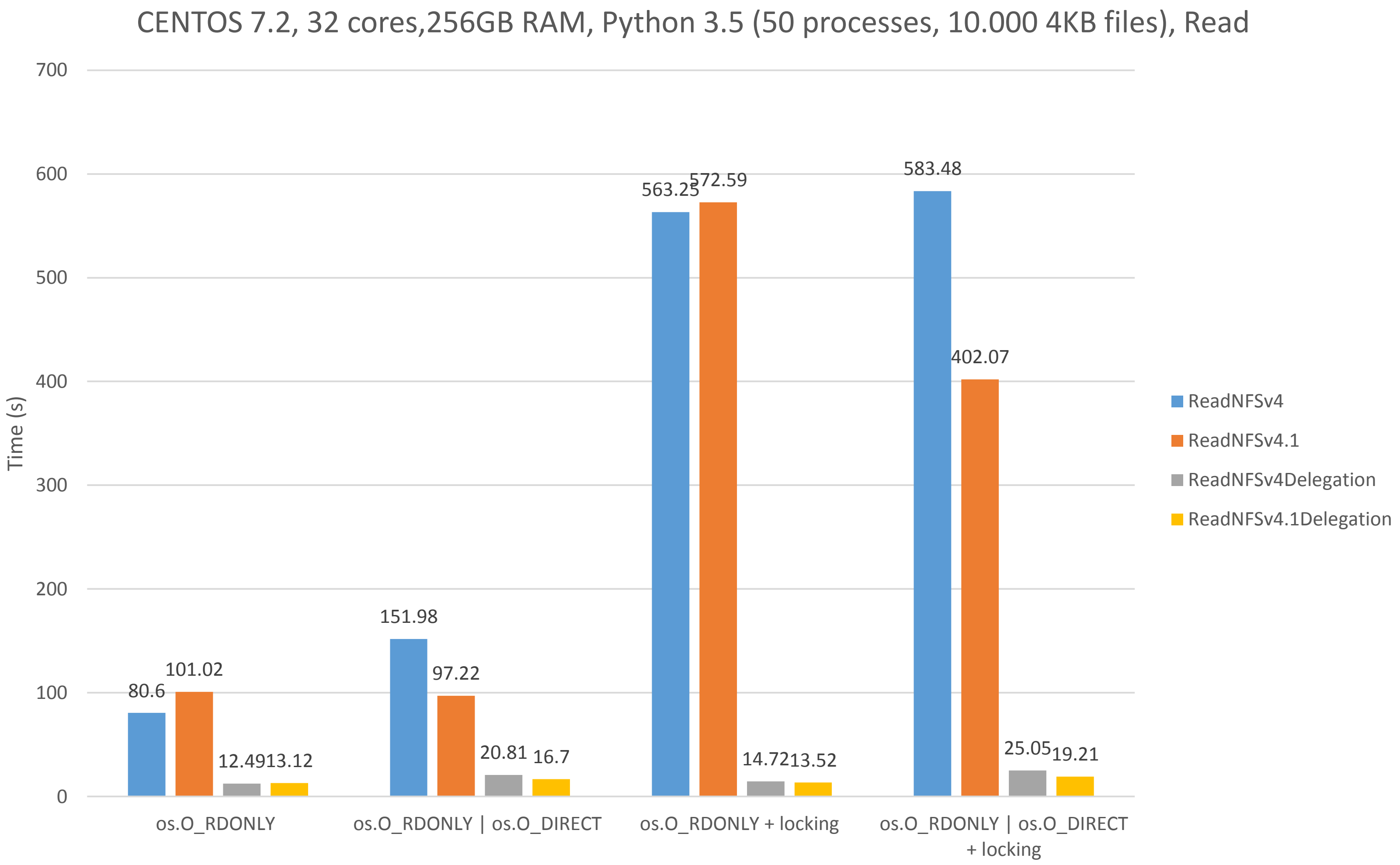
```
--Mount NFSv4 o NFSv4.1
mount -o rw,bg,hard,nointr,tcp,noatime,timeo=600,rsize=65536,wsize=65536,vers=4.1 -t nfs
dbnasc:/ORA/dbs07/DNFS /ORA/dbs07/DNFS
mount -o rw,bg,hard,nointr,tcp,noatime,timeo=600,rsize=65536,wsize=65536 -t nfs4
dbnasc:/ORA/dbs07/DNFS  /ORA/dbs07/DNFS

--create files
for i in `seq 1 10000`;do dd if=/dev/zero of=/ORA/dbs07/DNFS/file$i bs=4k count=1;done

--Run test
python3  FileOps.py  -f /ORA/dbs07/DNFS/file -p 50  -c 10000 -i 10 -o -w -d -l
Namespace(POOLSIZE=50, directio=True, file_prefix='/ORA/dbs07/DNFS/file', fine=True, isread=False,
iswrite=True, iteractions=10, locking=True, totalnum=10000)
Poolsize: 50, totalnum: 10000
45.07781410217285 run lasted (in seconds)
```

Delegations – test 2

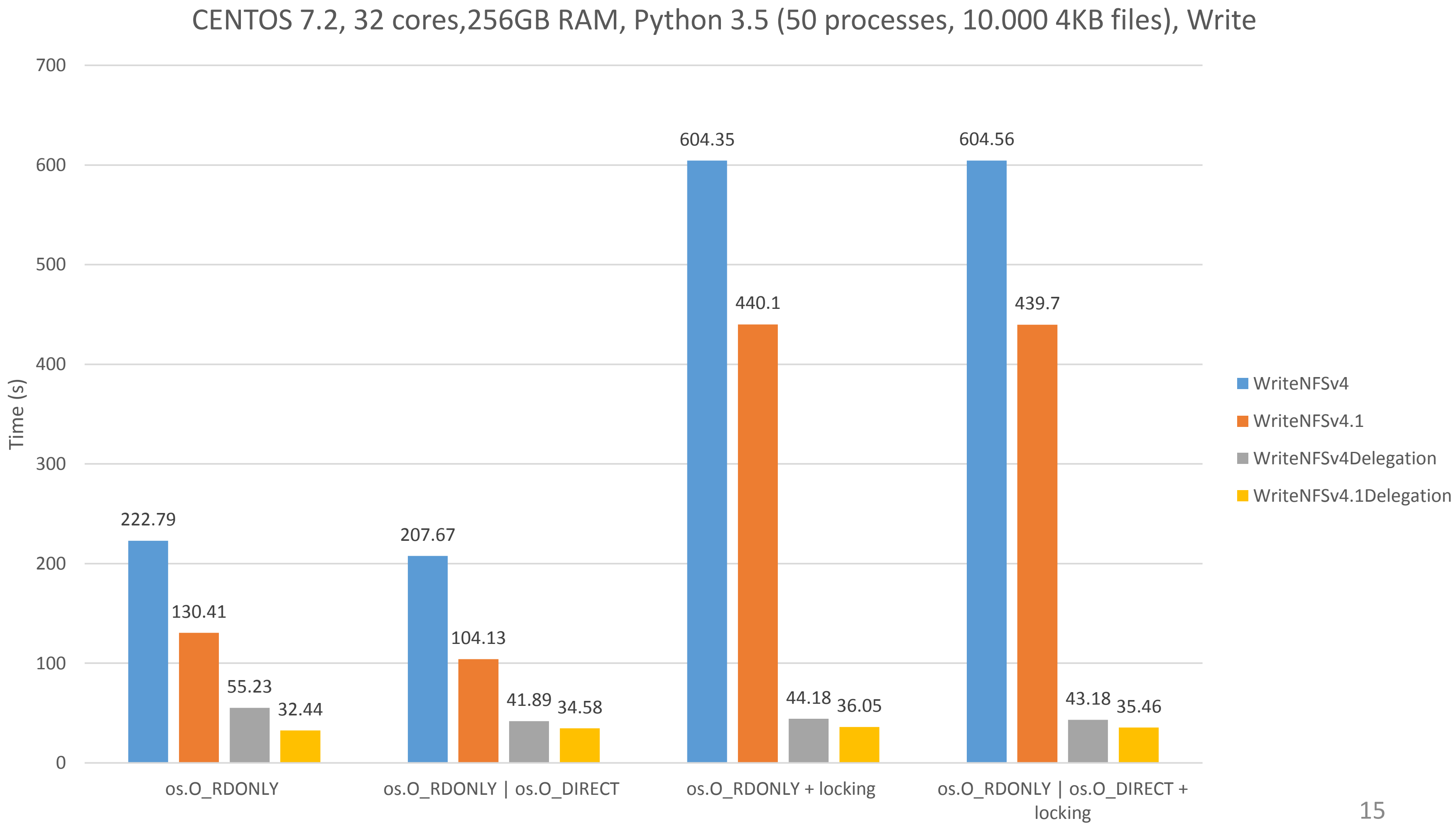
- Use client and server CLI to check right IO is on-going
  - collectl, mounstats, htop
  - smetrics (link to talk)
  - NFS server locks:

```
--nfs locks

sx50::*> vserver locks show -vserver vs2sx50 -volume dnfs07

Vserver: vs2sx50

Volume   Object Path               LIF         Protocol  Lock Type   Client

-------- ------------------------- ----------- --------- ----------- ----------

dnfs07   /ORA/dbs07/DNFS/file4501  vs2sx50_dbnasc501-cpub

                                               nfsv4.1   delegation  -

                Delegation Type: read

         /ORA/dbs07/DNFS/file4901  vs2sx50_dbnasc501-cpub

                                               nfsv4.1   delegation  -

                Delegation Type: read
```

Delegations - test 2



Delegations – test 2



Delegations: IO profiling

CENTOS 7.2 – ONTAP 8.3.1

READ (nfsstat)	**NFSv4**	**NFSv4_delg**	**NFSv4.1**	**NFSv4.1_delg**
OPEN	100000	10432	0	0
WRITE	0	0	0	0
READ	100000	99913	99998	99463
CLOSE	100000	10432	100000	20864
GETATTR	100050	480	190025	11824
LOCK	100000	480	100000	960
LOCKU	99990	480	100000	960

os.O_RDONLY | os.O_DIRECT + POSIX locks

CENTOS 7.2 – ONTAP 8.3.1

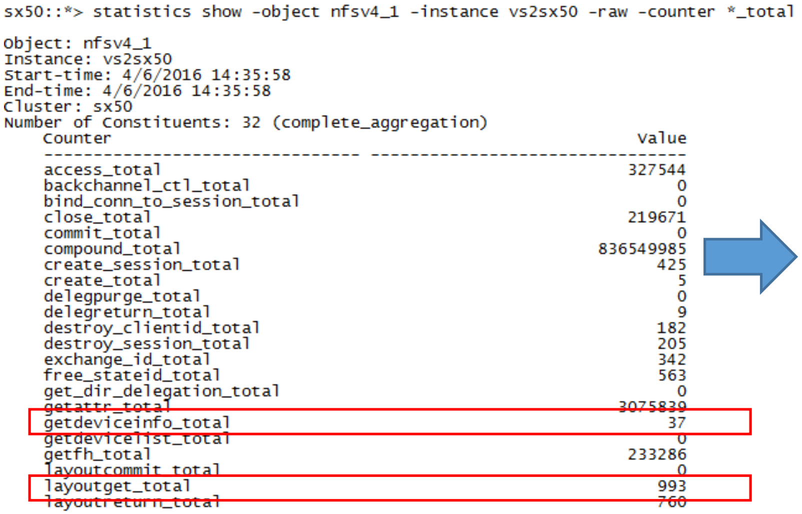
WRITE (nfsstat)	**NFSv4**	**NFSv4_delg**	**NFSv4.1**	**NFSv4.1_delg**
OPEN	100000	10000	0	0
WRITE	100000	99993	100000	99758
READ	0	0	0	0
CLOSE	100000	10000	100000	10000
GETATTR	100000	0	189917	0
LOCK	100000	0	100000	0
LOCKU	100000	0	99999	0

os.O_WRONLY | os.O_DIRECT + POSIX locks

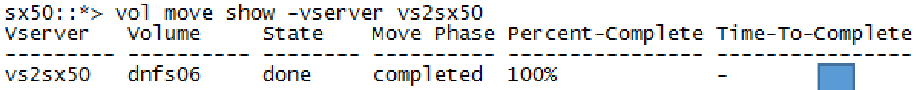
Pnfs

pNFS (ONTAP 8.3.1)

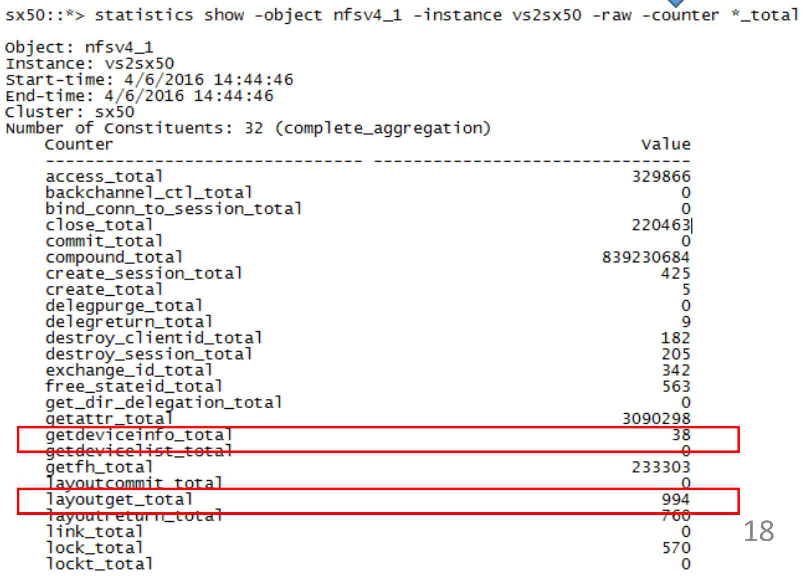
- It works with:
  - CENTOS 7.1 Openstack VM
  - Oracle 12.2.0.0.2 (beta2) and Kernel NFS



```
sx50::*> statistics show -object nfsv4_1 -instance vs2sx50 -raw -counter *_total

Object: nfsv4_1
Instance: vs2sx50
Start-time: 4/6/2016 14:35:58
End-time: 4/6/2016 14:35:58
Cluster: sx50
Number of Constituents: 32 (complete_aggregation)
    Counter                                                   Value
    ------------------------------- -------------------------------
    access_total                                             327544
    backchannel_ctl_total                                         0
    bind_conn_to_session_total                                    0
    close_total                                              219671
    commit_total                                                  0
    compound_total                                        836549985
    create_session_total                                        425
    create_total                                                  5
    delegpurge_total                                              0
    delegreturn_total                                             9
    destroy_clientid_total                                      182
    destroy_session_total                                       205
    exchange_id_total                                           342
    free_stateid_total                                          563
    get_dir_delegation_total                                      0
    getattr_total                                           3075839
    getdeviceinfo_total                                          37
    getdevicelist_total                                           0
    getfh_total                                              233286
    layoutcommit_total                                            0
    layoutget_total                                             993
    layoutreturn_total                                          760
```



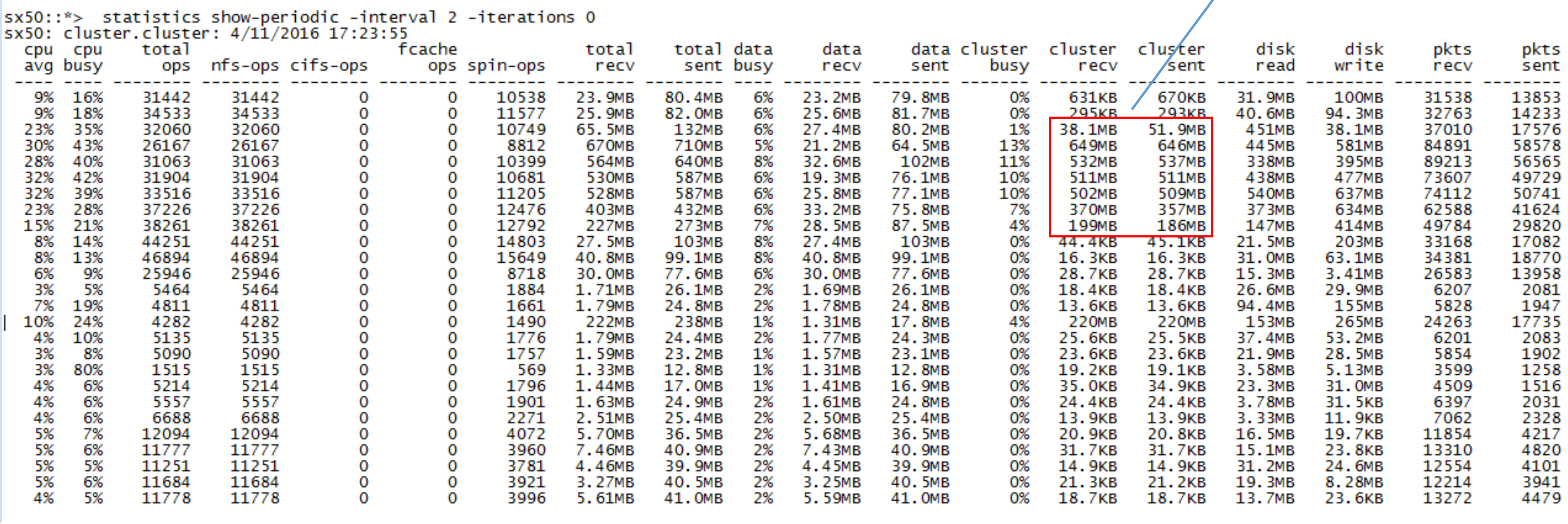
```
sx50::*> vol move show -vserver vs2sx50
Vserver   Volume     State    Move Phase Percent-Complete Time-To-Complete
--------- ---------- -------- ---------- ---------------- ----------------
vs2sx50   dnfs06     done     completed  100%             -
```



```
sx50::*> statistics show -object nfsv4_1 -instance vs2sx50 -raw -counter *_total

Object: nfsv4_1
Instance: vs2sx50
Start-time: 4/6/2016 14:44:46
End-time: 4/6/2016 14:44:46
Cluster: sx50
Number of Constituents: 32 (complete_aggregation)
    Counter                                                   Value
    ------------------------------- -------------------------------
    access_total                                             329866
    backchannel_ctl_total                                         0
    bind_conn_to_session_total                                    0
    close_total                                              220463
    commit_total                                                  0
    compound_total                                        839230684
    create_session_total                                        425
    create_total                                                  5
    delegpurge_total                                              0
    delegreturn_total                                             9
    destroy_clientid_total                                      182
    destroy_session_total                                       205
    exchange_id_total                                           342
    free_stateid_total                                          563
    get_dir_delegation_total                                      0
    getattr_total                                           3090298
    getdeviceinfo_total                                          38
    getdevicelist_total                                           0
    getfh_total                                              233303
    layoutcommit_total                                            0
    layoutget_total                                             994
    layoutreturn_total                                          760
    link_total                                                    0
    lock_total                                                  570
    lockt_total                                                   0
```

pNFS (ONTAP 8.3.1)

- It works with:
  - CENTOS 7.2 on a physical server
  - Oracle 12.2.0.0.2 (beta2) and Kernel NFS

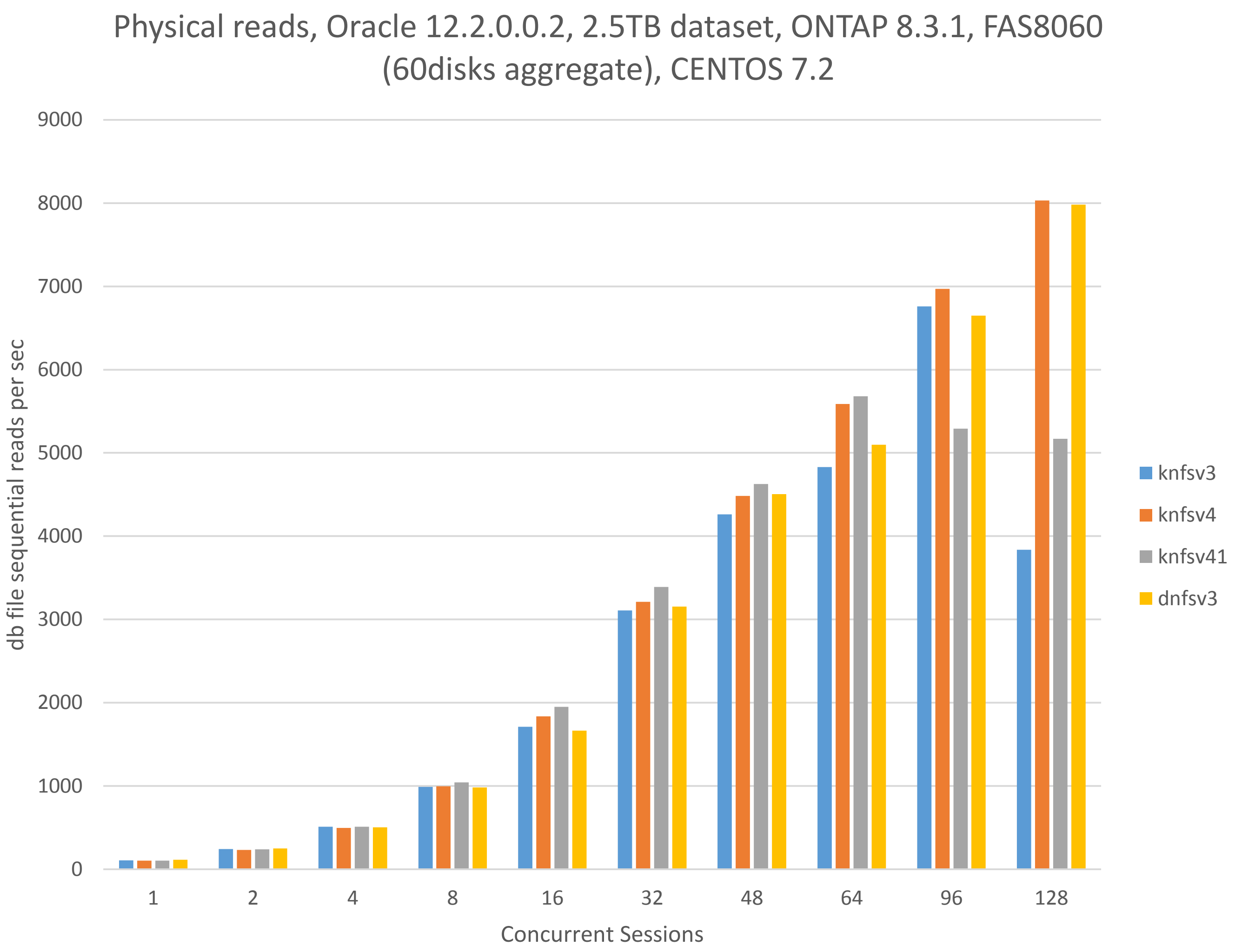


```
sx50::*>  statistics show-periodic -interval 2 -iterations 0
sx50: cluster.cluster: 4/11/2016 17:23:55
 cpu  cpu    total                   fcache             total    total data     data     data cluster  cluster  cluster     disk     disk     pkts     pkts
 avg busy      ops  nfs-ops cifs-ops      ops spin-ops     recv     sent busy     recv     sent    busy     recv     sent     read    write     recv     sent
---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- -------- ------- -------- -------- -------- -------- -------- --------
  9%  16%    31442    31442        0        0    10538   23.9MB   80.4MB   6%   23.2MB   79.8MB      0%    631KB    670KB   31.9MB    100MB    31538    13853
  9%  18%    34533    34533        0        0    11577   25.9MB   82.0MB   6%   25.6MB   81.7MB      0%    295KB    293KB   40.6MB   94.3MB    32763    14233
 23%  35%    32060    32060        0        0    10749   65.5MB    132MB   6%   27.4MB   80.2MB      1%   38.1MB   51.9MB    451MB   38.1MB    37010    17576
 30%  43%    26167    26167        0        0     8812    670MB    710MB   5%   21.2MB   64.5MB     13%    649MB    646MB    445MB    581MB    84891    58578
 28%  40%    31063    31063        0        0    10399    564MB    640MB   8%   32.6MB    102MB     11%    532MB    537MB    338MB    395MB    89213    56565
 32%  42%    31904    31904        0        0    10681    530MB    587MB   6%   19.3MB   76.1MB     10%    511MB    511MB    438MB    477MB    73607    49729
 32%  39%    33516    33516        0        0    11205    528MB    587MB   6%   25.8MB   77.1MB     10%    502MB    509MB    540MB    637MB    74112    50741
 23%  28%    37226    37226        0        0    12476    403MB    432MB   6%   33.2MB   75.8MB      7%    370MB    357MB    373MB    634MB    62588    41624
 15%  21%    38261    38261        0        0    12792    227MB    273MB   7%   28.5MB   87.5MB      4%    199MB    186MB    147MB    414MB    49784    29820
  8%  14%    44251    44251        0        0    14803   27.5MB    103MB   8%   27.4MB    103MB      0%   44.4KB   45.1KB   21.5MB    203MB    33168    17082
  8%  13%    46894    46894        0        0    15649   40.8MB   99.1MB   8%   40.8MB   99.1MB      0%   16.3KB   16.3KB   31.0MB   63.1MB    34381    18770
  6%   9%    25946    25946        0        0     8718   30.0MB   77.6MB   6%   30.0MB   77.6MB      0%   28.7KB   28.7KB   15.3MB   3.41MB    26583    13958
  3%   5%     5464     5464        0        0     1884   1.71MB   26.1MB   2%   1.69MB   26.1MB      0%   18.4KB   18.4KB   26.6MB   29.9MB     6207     2081
  7%  19%     4811     4811        0        0     1661   1.79MB   24.8MB   2%   1.78MB   24.8MB      0%   13.6KB   13.6KB   94.4MB    155MB     5828     1947
 10%  24%     4282     4282        0        0     1490    222MB    238MB   1%   1.31MB   17.8MB      4%    220MB    220MB    153MB    265MB    24263    17735
  4%  10%     5135     5135        0        0     1776   1.79MB   24.4MB   2%   1.77MB   24.3MB      0%   25.6KB   25.5KB   37.4MB   53.2MB     6201     2083
  3%   8%     5090     5090        0        0     1757   1.59MB   23.2MB   1%   1.57MB   23.1MB      0%   23.6KB   23.6KB   21.9MB   28.5MB     5854     1902
  3%  80%     1515     1515        0        0      569   1.33MB   12.8MB   1%   1.31MB   12.8MB      0%   19.2KB   19.1KB   3.58MB   5.13MB     3599     1258
  4%   6%     5214     5214        0        0     1796   1.44MB   17.0MB   1%   1.41MB   16.9MB      0%   35.0KB   34.9KB   23.3MB   31.0MB     4509     1516
  4%   6%     5557     5557        0        0     1901   1.63MB   24.9MB   2%   1.61MB   24.8MB      0%   24.4KB   24.4KB   3.78MB   31.5KB     6397     2031
  4%   6%     6688     6688        0        0     2271   2.51MB   25.4MB   2%   2.50MB   25.4MB      0%   13.9KB   13.9KB   3.33MB   11.9KB     7062     2328
  5%   7%    12094    12094        0        0     4072   5.70MB   36.5MB   2%   5.68MB   36.5MB      0%   20.9KB   20.8KB   16.5MB   19.7KB    11854     4217
  5%   6%    11777    11777        0        0     3960   7.46MB   40.9MB   2%   7.43MB   40.9MB      0%   31.7KB   31.7KB   15.1MB   23.8KB    13310     4820
  5%   5%    11251    11251        0        0     3781   4.46MB   39.9MB   2%   4.45MB   39.9MB      0%   14.9KB   14.9KB   31.2MB   24.6MB    12554     4101
  5%   6%    11684    11684        0        0     3921   3.27MB   40.5MB   2%   3.25MB   40.5MB      0%   21.3KB   21.2KB   19.3MB   8.28MB    12214     3941
  4%   5%    11778    11778        0        0     3996   5.61MB   41.0MB   2%   5.59MB   41.0MB      0%   18.7KB   18.7KB   13.7MB   23.6KB    13272     4479
```

vol move operation

pNFS and Oracle 12.2.0.0.2

- pNFS doesn’t work at all
- SR 3-12517180796
  - Bug 22261050 - dnfs_imc : ora-00600 : [ksfdcls5] - rms0 - abnormal instance termination
  - Bug 21477246 - ora-07445: core dump [kgnfswat()+62] [sigsegv]
- It should be corrected on latest Oracle 12.2 server.
  - To be checked! (Mid May new beta should be available)

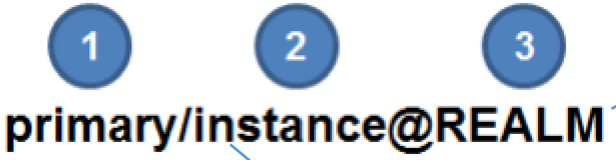


Regression on Oracle 12.2.0.0.2 with respect 12.2.0.0.1 on behalf dnfs IO:
pNFS and NFSv4.1 not working.
Needs to be followed up!

Kerberos

Kerberos + NFS

- TR-4073 (277 pages)
- Small intro to Kerberos



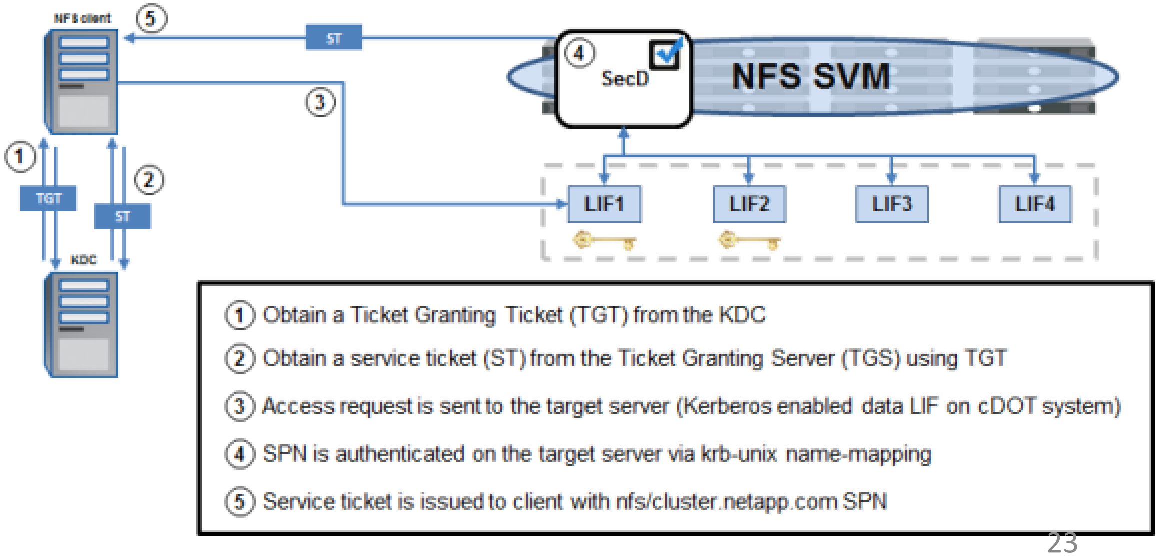
CERN.CH (DNS)

Service: FQDN *nfs/dbnasc501-d.cern.ch@CERN.CH*

Service: nfs

User e.g. joe

Figure 1) Kerberos workflow between client, KDC, and NFS server on NetApp storage.



1. Obtain a Ticket Granting Ticket (TGT) from the KDC
2. Obtain a service ticket (ST) from the Ticket Granting Server (TGS) using TGT
3. Access request is sent to the target server (Kerberos enabled data LIF on cDOT system)
4. SPN is authenticated on the target server via krb-unix name-mapping
5. Service ticket is issued to client with nfs/cluster.netapp.com SPN

NetApp TR-4073 drawing

Kerberos + NFS

- Some differences with respect our usual setup:
    - Create a Kerberos domain, enable encryption types (AES)
    - Enable resolution of user/service principals
        - LDAP (it looks not working)
        - Name mapping rules (it works)
    - Kerberos service SPN

```
--it doesnt work (AD privileges required)

sx50::*> kerberos interface modify -vserver vs3sx50 -lif vs3sx50_dbnasc501-dpub -kerberos enabled -spn nfs/dbnasc501-d.cern.ch@CERN.CH

--it works!

sx50::*> kerberos interface modify -vserver vs3sx50 -lif vs3sx50_dbnasc501-dpub -kerberos enabled -spn nfs/dbnasc501-d.cern.ch@CERN.CH  -keytab-uri http://web.cern.ch/dbnasc501-d.keytab
```

- Export policies

```
export-policy rule modify -vserver vs3sx50 -policyname kerberos -protocol nfs -rorule sys,krb5,krb5i..
```

Keytab dilemma

Yes, we can with Jarek Polok !

- On a computer object:

>> Dn: CN=dbnasc501-d,OU=CERN Linux Computers,DC=cern,DC=ch

5> objectClass: top; person; organizationalPerson; user; computer;

...

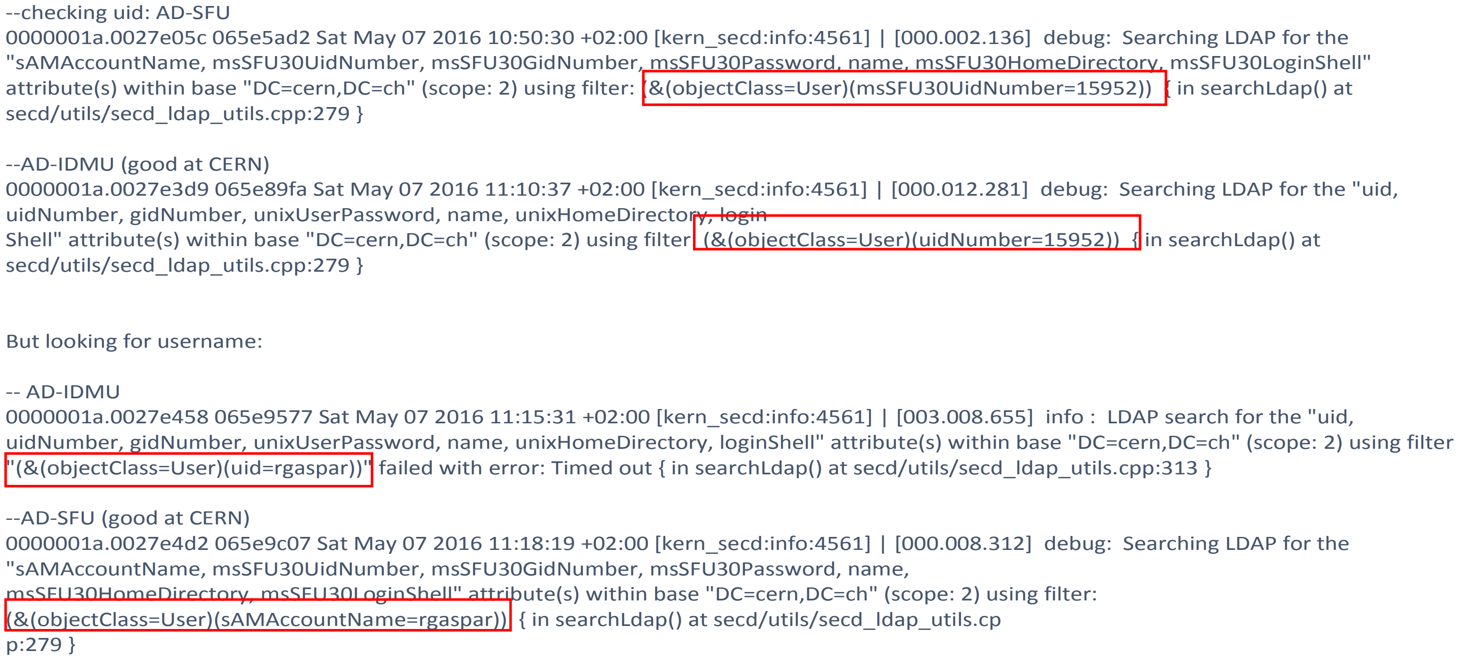
1> manager: CN=**service-db-systems**,OU=e-groups,OU=Workgroups,DC=cern,DC=ch;

- supportedEncryptionTypes requires A.D. admin privs

```
--version 0.9.10 at least!
cern-get-keytab --keytab dbnasc501-d.keytab --service nfs --alias dbnasc501-d.cern.ch --enctypes
'AES128_CTS_HMAC_SHA1|AES256_CTS_HMAC_SHA1' --debug --verbose

--We get
klist -tke dbnasc501-d.keytab
Keytab name: FILE:dbnasc501-d.keytab
KVNO Timestamp           Principal
---- ------------------- ------------------------------------------------------
  12 04/27/2016 15:37:32 dbnasc501-d$@CERN.CH (des-cbc-crc)
  12 04/27/2016 15:37:32 dbnasc501-d$@CERN.CH (des-cbc-md5)
  12 04/27/2016 15:37:32 dbnasc501-d$@CERN.CH (arcfour-hmac)
  12 04/27/2016 15:37:32 nfs/dbnasc501-d.cern.ch@CERN.CH (des-cbc-crc)
  12 04/27/2016 15:37:32 nfs/dbnasc501-d.cern.ch@CERN.CH (des-cbc-md5)
  12 04/27/2016 15:37:32 nfs/dbnasc501-d.cern.ch@CERN.CH (arcfour-hmac)

--But we want
klist -kte dbnasc501-d.keytab
Keytab name: FILE: dbnasc501-d.keytab
KVNO Timestamp           Principal
---- ------------------- ------------------------------------------------------
  15 04/28/2016 15:40:47 dbnasc501-d$@CERN.CH (aes128-cts-hmac-sha1-96)
  15 04/28/2016 15:40:47 dbnasc501-d$@CERN.CH (aes256-cts-hmac-sha1-96)
  15 04/28/2016 15:40:47 nfs/dbnasc501-d.cern.ch@CERN.CH (aes128-cts-hmac-sha1-96)
  15 04/28/2016 15:40:47 nfs/dbnasc501-d.cern.ch@CERN.CH (aes256-cts-hmac-sha1-96)
```

LDAP client dilemma

- Three possible variants in NetApp: AD-IDMU, AD-SFU,RFC-2307
- It depends on which property different schemas apply.
  - Name-mapping works! (local resolution)



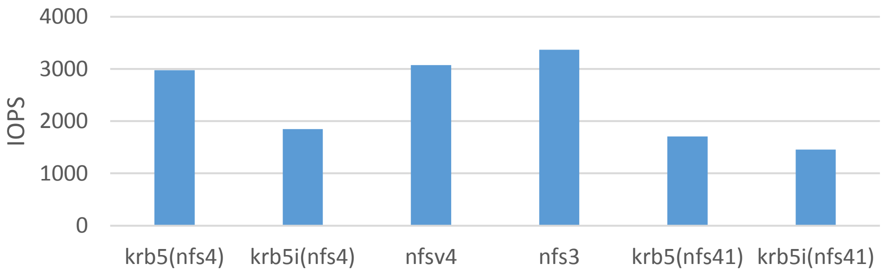
```
--checking uid: AD-SFU
0000001a.0027e05c 065e5ad2 Sat May 07 2016 10:50:30 +02:00 [kern_secd:info:4561] | [000.002.136]  debug:  Searching LDAP for the
"sAMAccountName, msSFU30UidNumber, msSFU30GidNumber, msSFU30Password, name, msSFU30HomeDirectory, msSFU30LoginShell"
attribute(s) within base "DC=cern,DC=ch" (scope: 2) using filter: (&(objectClass=User)(msSFU30UidNumber=15952)) { in searchLdap() at
secd/utils/secd_ldap_utils.cpp:279 }

--AD-IDMU (good at CERN)
0000001a.0027e3d9 065e89fa Sat May 07 2016 11:10:37 +02:00 [kern_secd:info:4561] | [000.012.281]  debug:  Searching LDAP for the "uid,
uidNumber, gidNumber, unixUserPassword, name, unixHomeDirectory, login
Shell" attribute(s) within base "DC=cern,DC=ch" (scope: 2) using filter (&(objectClass=User)(uidNumber=15952)) { in searchLdap() at
secd/utils/secd_ldap_utils.cpp:279 }


But looking for username:

-- AD-IDMU
0000001a.0027e458 065e9577 Sat May 07 2016 11:15:31 +02:00 [kern_secd:info:4561] | [003.008.655]  info :  LDAP search for the "uid,
uidNumber, gidNumber, unixUserPassword, name, unixHomeDirectory, loginShell" attribute(s) within base "DC=cern,DC=ch" (scope: 2) using filter
"(&(objectClass=User)(uid=rgaspar))" failed with error: Timed out { in searchLdap() at secd/utils/secd_ldap_utils.cpp:313 }

--AD-SFU (good at CERN)
0000001a.0027e4d2 065e9c07 Sat May 07 2016 11:18:19 +02:00 [kern_secd:info:4561] | [000.008.312]  debug:  Searching LDAP for the
"sAMAccountName, msSFU30UidNumber, msSFU30GidNumber, msSFU30Password, name,
msSFU30HomeDirectory, msSFU30LoginShell" attribute(s) within base "DC=cern,DC=ch" (scope: 2) using filter:
(&(objectClass=User)(sAMAccountName=rgaspar)) { in searchLdap() at secd/utils/secd_ldap_utils.cp
p:279 }
```

Kerberos experience

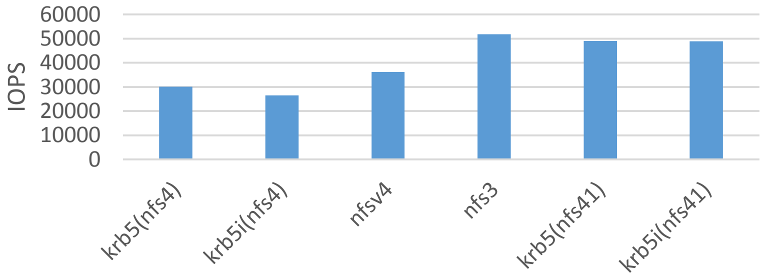
- Mount either on NFSv4 or NFSv4.1

```
mount dbnasc501-d:/ORA/dbs00/KERBEROS -t nfs4 -o
sec=krb5,rw,bg,hard,nointr,tcp,noatime,timeo=600,rsize=65536,wsize=65536 /ORA/dbs00/KERBEROS

mount dbnasc501-d:/ORA/dbs00/KERBEROS -t nfs4 -o
sec=krb5i,rw,bg,hard,nointr,tcp,noatime,timeo=600,rsize=65536,wsize=65536 /ORA/dbs00/KERBEROS
```

random_reads_4kb, 1 process, 1TB file, CENTOS 7.2 (physical server)



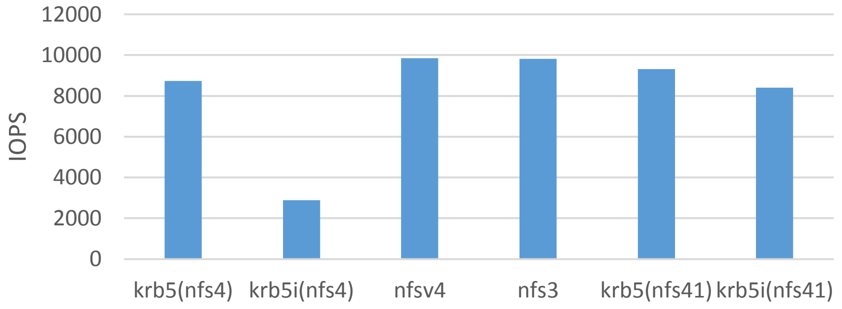
random_writes_4kb, 1 process, 1TB file, CENTOS 7.2 (physical server)



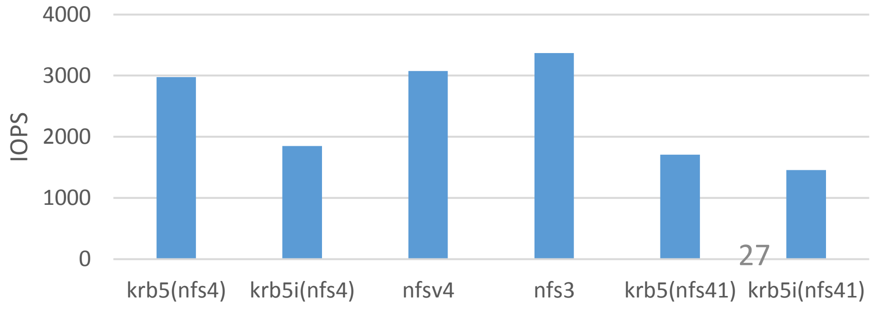
***libaio,nodirectIO,nolocking**

***FA8060, ONTAP 8.3.1, 60 disks aggregate**
CENTOS 7.2, 256GB RAM, 32 cores

sequential_writes_64kb, 1 process, 1TB file, CENTOS 7.2 (physical server)

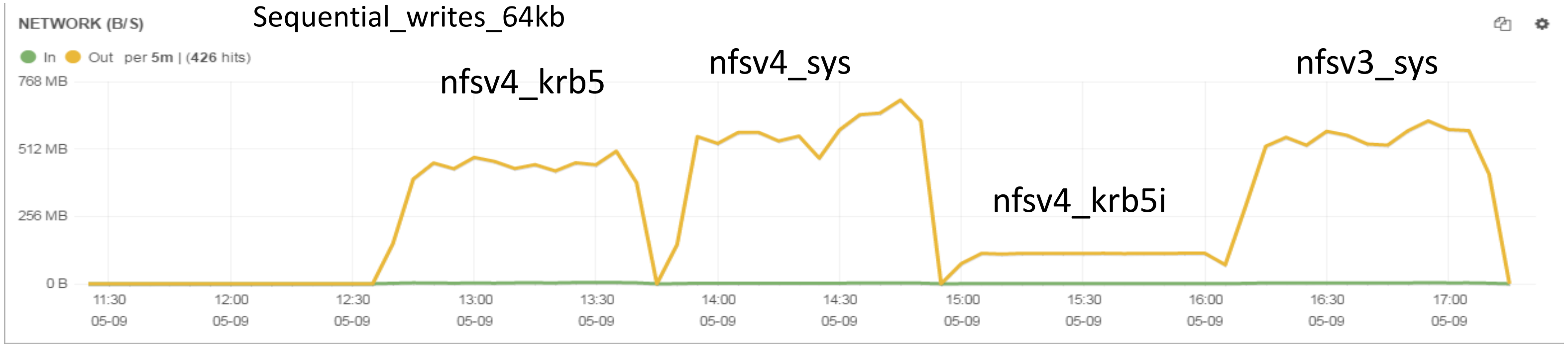


sequential_reads_64kb, 1 process, 1TB file, CENTOS 7.2 (physical server)



itrac51104:
CENTOS 7.2
32 cores, 256 GB RAM
10GbE

Sequential_writes_64kb

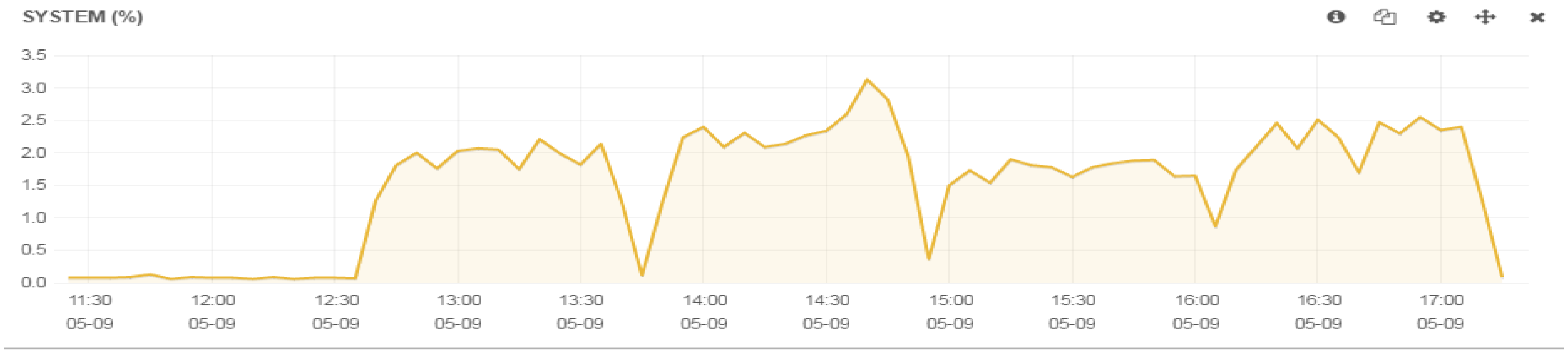


nfsv4_krb5

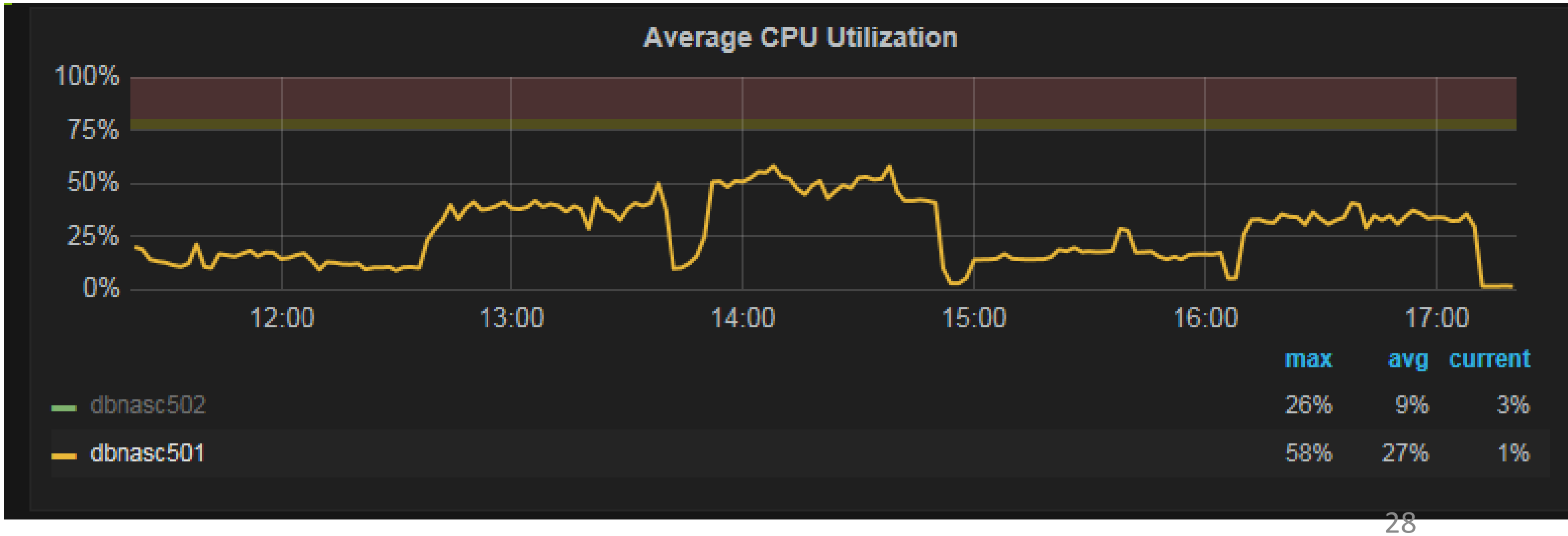
nfsv4_sys

nfsv4_krb5i

nfsv3_sys



dbnasc501 (FAS8060):
ONTAP 8.3.1
NVRAM: 8GB
16 cores, RAM: 64GB
(60 data disks in aggregate)
20GbE



	max	avg	current
dbnasc502	26%	9%	3%
dbnasc501	58%	27%	1%

Conclusions

- Delegations useful in particular environment, not DB traditional scenario but good may be for virtualisation, application server, etc. ones
  - Enough memory on the apps server
  - NetApp controller shouldn't be loaded
  - Linux kernel evolution, watch up!
- Pnfs operational at kernel level on CENTOS 7.2
  - To be tested on Oracle server 12.2.0.3
- Kerberos
  - It looks a great feature for CERN openwide storage related services (no export-policy but authentication)
    - Kind of CERNbox type
  - LDAP dilemma needs to be solved (NetApp case 2006265613)
    - 9th June update: it looks it works now!

```
vserver services name-service ldap client schema copy -schema AD-IDMU -new-schema-name AD-IDMU-mod -vserver vs3sx50
vserver services name-service ldap client schema modify -schema AD-IDMU-mod -uid-attribute sAMAccountName -vserver vs3sx50
```